

USER GUIDE

applied
biosystems®
by *life* technologies™

LifeScope™ Genomic Analysis Software

Graphical User Interface

DATA ANALYSIS METHODS AND INTERPRETATION

Publication Part Number 4465696 Rev. A

Revision Date May 2011

life
technologies™

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Information in this document is subject to change without notice.

APPLIED BIOSYSTEMS DISCLAIMS ALL WARRANTIES WITH RESPECT TO THIS DOCUMENT, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO THOSE OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. TO THE FULLEST EXTENT ALLOWED BY LAW, IN NO EVENT SHALL APPLIED BIOSYSTEMS BE LIABLE, WHETHER IN CONTRACT, TORT, WARRANTY, OR UNDER ANY STATUTE OR ON ANY OTHER BASIS FOR SPECIAL, INCIDENTAL, INDIRECT, PUNITIVE, MULTIPLE OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH OR ARISING FROM THIS DOCUMENT, INCLUDING BUT NOT LIMITED TO THE USE THEREOF, WHETHER OR NOT FORESEEABLE AND WHETHER OR NOT APPLIED BIOSYSTEMS IS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

LIMITED USE LABEL LICENSE

No right to resell this product or any of its components is conveyed expressly, by implication, or by estoppel. For information on obtaining additional rights, please contact outlicensing@lifetech.com or Out Licensing, Life Technologies, 5791 Van Allen Way, Carlsbad, California 92008.

NOTICE TO PURCHASER: DISCLAIMER OF LICENSE

Purchase of this software product alone does not imply any license under any process, instrument or other apparatus, system, composition, reagent or kit rights under patent claims owned or otherwise controlled by Applied Biosystems, either expressly, or by estoppel.

TRADEMARKS

The trademarks mentioned herein are the property of Life Technologies Corporation or their respective owners. Windows and Internet Explorer are registered trademarks of Microsoft Corporation in the United States and other countries. Mozilla is a trademark of Mozilla Foundation Corporation. Sage is a trademark of Genzyme Corporation.

© 2011 Life Technologies Corporation. All rights reserved.

Part Number 4465696 Rev. A
May 2011

Contents

	About This Guide	17
	Purpose	17
	Prerequisites	17
CHAPTER 1	Introduction to LifeScope™ Genomic Analysis Software	19
	LifeScope™ Software Overview	19
	Features	19
	Data analyses	19
	LifeScope™ Software components	20
	Workflows	20
	Standard workflow	20
	Project workflow	21
	Analysis workflows	21
	Primary and secondary file types	26
	Input and output file formats	26
CHAPTER 2	Understand LifeScope™ Software	29
	Overview	29
	Terminology	30
	Projects and analyses	30
	Naming restrictions	30
	Reads Data	30
	Define input data	31
	Repositories	31
	Common scenarios	32
	Basic	32
	A sample sequenced on multiple lanes	32
	Data from multiple samples	32
CHAPTER 3	LifeScope™ Genomic Analysis Software Installation	33
	Introduction	33
	Prerequisites	33
	Hardware requirements	34
	System requirements	34
	Server requirements	34

	Client hardware and software requirements	35
	LifeScope™ Software Administration	35
	Installation workflow overview	36
	Copy data drive content	37
	Download LifeScope™ Software	37
	Install LifeScope™ Software	38
	Continue installing LifeScope™ Software	40
	Activate the LifeScope™ Software License Key	40
	Prerequisites	40
	Check the computer date	41
	Obtain the MAC address	41
	Find MAC Address with LifeScope™ Installer	43
	Activate License Key and obtain license file	44
	Apply the License File to LifeScope™ Software	45
	Check firewall restrictions	47
	BioScope™ Software users' PATH variable	47
	Download documentation and additional resources	47
	Integrative Genomics Viewer (IGV)	47
CHAPTER 4	Test LifeScope™ Genomic Analysis Software	49
	Introduction	49
	Workflow	50
	Verify the installation	50
	<i>(Optional)</i> Download and install demos files	51
	<i>(Optional)</i> Run demo analyses	51
	Check results	52
	<i>(Optional)</i> Download and install performance verification scripts	52
	<i>(Optional)</i> Run the performance verification test	52
	Monitor progress	53
	Check results	53
	Create user accounts	53
	Enable users	54
	Distribute the URL and user credentials	54
PART I	Overview	55
CHAPTER 5	LifeScope™ Genomic Analysis Software	
	Graphical User Interface	57
	Log in, log out	57
	Button labels and table titles	62
	Pop-up windows	62

PART II	Getting Started	63
CHAPTER 6	Create and Manage Projects	65
	Overview	65
	Workflow	65
	Create a project	66
	Prepare to create a project	66
	Log in to LifeScope™ Software	66
	Create a project	67
	Name and describe a project	67
	(Optional) Import data	67
	Add data to a project	68
	Choose data type	68
	Find data	68
	(Optional) Group unmapped data	71
	Group BAM data	72
	View project data	72
	Project overview	72
	Project status	72
	Read-sets	72
	BAM data	73
	Delete a project	73
CHAPTER 7	Perform an Analysis	75
	Workflow	76
	(Optional) Import references	77
	Import references	77
	Import your own references	78
	View import status	78
	Create an analysis	78
	Start analysis creation	78
	Choose data type	79
	Choose data	79
	Choose an analysis	79
	Choose references	81
	Reuse an analysis	81
	Reuse BAM data	82
	Edit an analysis	82
	Choose modules	83
	Set general parameters	84
	Set module parameters	84
	Review your analysis	85

	Review parameters	85
	Review data	85
	Run and monitor your analysis	85
	Start your analysis now	85
	Start your analysis later	85
	Stop an analysis	85
	View analysis status	85
	Delete an analysis	86
CHAPTER 8	View Analysis Results	87
	View analysis results	87
	View results now	87
	View results later	87
	View Results window	87
	Details	88
	Statistics	89
	Additional Files	89
	Logs	89
	Examples of view results	89
	Logs	91
	Failed analysis	91
	View failure details	92
PART III	Workflows	93
CHAPTER 9	Perform Targeted Resequencing and Enrichment Analyses ...	95
	Introduction to Targeted Resequencing analysis	95
	Targeted resequencing library types	96
	Enrichment statistics	96
	Targeted resequencing input files	97
	Enrichment statistics input files	97
	Targeted resequencing analysis modules	99
	Mapping	99
	Enrichment	99
	SNP Finding	99
	Small Indel	99
	Targeted resequencing parameters	100
	General	100
	Fragment Mapping	100
	Small indel	104
	Enrichment	108
	SNP Finding	110

(Optional) Annotation	112
Perform Targeted Resequencing analysis	112
(Optional) Import data	112
Log in to LifeScope™ Software	112
Create a project	112
Add data to the project	112
Create an analysis	113
Edit the analysis	113
Review and run the analysis	115
Checking Analysis Status	115
View analysis results	116
View results in a genome browser	116
Targeted resequencing output files	116
SmallIndel.Annotation	116
SNP.Finding.Annotation	117
SNP.Finding.targeted.frag	118
Enrichment	120
SAET	122
BAMStats	123
Mapping	124
Small Indel	125
CHAPTER 10	Perform Genomic Resequencing Analysis
	129
Introduction to genomic resequencing analysis	129
Genomic resequencing analysis library types	129
Genomic resequencing analysis input files	129
Genomic resequencing analysis modules	130
Mapping	130
Small Indel	130
SNP Finding	130
CNV	131
Genomic resequencing analysis parameters	131
General	131
Mapping	132
Small Indel	135
SNP Finding	140
Human CNV	142
(Optional) Annotation	144
Perform genomic resequencing analysis	144
(Optional) Import data	144
Log in to LifeScope™ Software	144

Create or select a project	144
Add data to a project	144
Create an analysis	145
Edit the analysis	145
Review and run the analysis	147
Checking Analysis Status	147
View analysis results	148
View results in a genome browser	148
Genomic resequencing analysis output files	148

CHAPTER 11 Perform Small RNA Analysis 149

Introduction to Small RNA analysis	149
Small RNA analysis library types	149
Small RNA analysis input files	149
Mapping input files	149
Coverage input files	151
Counts input files	151
Small RNA analysis modules	152
Small RNA Mapping	152
Coverage	153
Counts	153
Small RNA analysis parameters	153
General	153
Small RNA Mapping	153
Small RNA Counts	155
Coverage	155
Small RNA Filtered BAM Counts	156
Small RNA Filtered BAMStats	156
Perform Small RNA analysis	157
(Optional) Import data	157
Create or select a project	157
Add data to a project	157
Create an analysis	158
Edit the analysis	158
Review and run the analysis	160
Checking Analysis Status	160
View analysis results	161
View results in a genome browser	161
Small RNA output files	161
Mapping output files	161
Coverage output files	161
Counts output files	162

CHAPTER 12	Perform Whole Transcriptome Analysis	163
	Introduction to whole transcriptome analysis	163
	Whole transcriptome analysis library types	164
	Whole transcriptome analysis input files	164
	RNA-Seq reads	164
	Reference input files	165
	Annotations input files	165
	UCSC genome annotations	165
	ENSEMBL GTF files	166
	Whole transcriptome analysis modules	166
	Whole Transcriptome Fragment Mapping	167
	Whole Transcriptome Exon Sequence Extractor	167
	Coverage	167
	Whole Transcriptome Count Features	167
	Splice Finding	167
	Whole transcriptome analysis parameters	167
	General	167
	Whole Transcriptome Fragment Mapping	168
	Whole Transcriptome Exon Sequence Extractor	170
	Coverage	170
	Whole Transcriptome Count Features	171
	Splice Finding	172
	Perform whole transcriptome analysis	174
	(Optional) Import data	174
	Log in to LifeScope™ Software	174
	Create or select a project	174
	Add data to a project	174
	Create an analysis	175
	Edit the analysis	175
	Review and run the analysis	177
	Check Analysis Status	177
	View analysis results	178
	View results in a genome browser	178
	Whole transcriptome analysis output files	178
	Whole transcriptome analysis output file formats	178
	BAM files	178
	Alignment report	178
	WT filtering stats	179
CHAPTER 13	Perform Whole Transcriptome Mapping	181
	Introduction to whole transcriptome mapping	181
	Examples of running the whole transcriptome mapping module	181
	Reads input files	182

RNA-Seq reads	182
Annotations input files	182
UCSC genome annotations	183
Reference input files	183
Mapping parameters	184
Fragment Mapping	184
Paired-end	186
Map data	189
Mapping output files	190
Overview	190
BAM file differences	190
Mapping statistics parameters	191
Summary of mapping statistics output	191
Mapping statistics output files	192
Example of mapping statistics output	197
Whole transcriptome analysis output file formats	199
Alignment report	199
Whole transcriptome analysis filtering stats	200

CHAPTER 14 Perform MethylMiner™ Mapping 203

Introduction to MethylMiner™ mapping	203
MethylMiner™ mapping library types	204
MethylMiner™ mapping input files	204
MethylMiner™ mapping parameters	205
General parameters	205
Fragment mapping parameters	205
Perform MethylMiner™ mapping	207
Optional) Import data	207
Log in to LifeScope™ Software	207
Create or select a project	207
Add data to a project	208
Create an analysis	208
Edit the analysis	209
Review and run the analysis	209
Check Analysis Status	210
View MethylMiner™ mapping results	211
MethylMiner™ mapping output files	211
View output files in a genome browser	211
Further analysis of MethylMiner™ mapping results	212

CHAPTER 15	Perform ChIP-Seq Mapping	213
	Introduction to ChIP-Seq mapping	213
	ChIP-Seq library types	213
	ChIP-Seq input files	214
	ChIP-Seq parameters	214
	General	214
	Fragment mapping	214
	Perform ChIP-Seq mapping	216
	(Optional) Import data	216
	Log in to LifeScope™ Software	216
	Create or select a project	217
	Add data to a project	217
	Create an analysis	217
	Edit the analysis	218
	Review and run the analysis	219
	Check Analysis Status	219
	View ChIP-Seq mapping results	220
	ChIP-Seq output files	220
PART IV	Analysis Modules	223
CHAPTER 16	Perform Resequencing Mapping	225
	Introduction to resequencing mapping	225
	Input files	225
	Plan your input read-sets	226
	Legacy data	226
	Mapping parameters	226
	Fragment Mapping	226
	Paired-end	230
	Map data	233
	Mapping output files	233
	Mapping output files	234
	BAMStats	235
CHAPTER 17	Perform Human Copy Number Variation Analysis	237
	Introduction to Human Copy Number Variation analysis	237
	Human Copy Number Variation analysis parameters	237
	(Optional) Annotation	239
	Perform Human Copy Number Variation analysis	240
	View analysis status	240
	View Human CNV analysis output	240

	Details	241
	Statistics	241
	Additional files	242
	Logs	242
CHAPTER 18	Perform Inversion Analysis	243
	Introduction to inversion analysis	243
	Inversion analysis input files	243
	Inversion analysis parameters	244
	Main	244
	Perform inversion analysis	245
	View analysis status	246
	View inversion analysis output	246
	Details	246
	Statistics	246
	Additional files	246
	Logs	247
CHAPTER 19	Perform SNP Finding Analysis	249
	Introduction to SNP Finding analysis	249
	SNP analysis input files	249
	SNP Finding analysis parameters	250
	Main	250
	Advanced	251
	(Optional) Annotation	252
	Perform SNP Finding analysis	252
	View analysis status	253
	View SNP Finding analysis output	253
	Details	253
	Statistics	253
	Additional files	254
	Logs	254
CHAPTER 20	Perform Large Indel Analysis	255
	Introduction to large indel analysis	255
	Large indel analysis input files	255
	Large indel analysis parameters	256
	Perform large indel analysis	257
	View analysis status	258
	View large indel analysis output	258
	Details	258
	Statistics	258

	Additional files	258
	Logs	258
CHAPTER 21	Perform Small Indel Analysis	259
	Introduction to small indel analysis	259
	Small indel analysis input files	259
	Small indel analysis parameters	259
	Perform small indel analysis	264
	View analysis status	265
	View small indel analysis output	265
	Details	265
	Statistics	265
	Additional files	266
	Logs	267
CHAPTER 22	Add Genomic Annotations to Analysis Results	269
	Overview	269
	Memory requirement	270
	Input file handling	270
	Filters	270
	Statistics	271
	Workflows	272
	Annotation sources	272
	Annotation parameters	273
	Variant Statistics output file for SNPs	277
	Variant Statistics output file for small indels	279
	Variant Statistics output file for large indels	282
	Variant Statistics output file for CNVs	283
	Mutated Genes output file	285
	SNP Finding tab-delimited output file	286
	SNP Finding annotated tab-delimited output file	287
	SNP Finding filtered annotated tab-delimited output file	288
	About annotations and LifeScope™ Software modules	289
	Examples of annotation output files	290
	Additional files	290
	Logs	290

PART V Appendices 291

APPENDIX A File Format Descriptions and Data Uses 293

Introduction 293

XSQ file format 293

 XSQ file content overview 294

 XSQ file format properties 294

BAM headers in LifeScope™ Software 294

 Sequence dictionary (@SQ) 294

 Read group (@RG) 295

 Header (@HD) sort order 295

 XSQ metadata in BAM headers 296

Color-space attributes 300

Pairing information in a BAM file 300

 Calculation of tag names 300

 Proper pairs 300

 Single read mapping quality 301

Hard clipping of incomplete extensions 301

Visualize BAM output 302

 Integrative Genomics View (IGV) 302

 UC Santa Cruz (UCSC) genome browser 303

BED file format 303

BEDGRAPH file format 304

GFF3 file format 305

Reference file data overview 305

 Contig multi-fasta file 305

 Single contig FASTA file 305

 GTF file 305

 Reference sequence data validation 306

 Concatenation 306

 Select a reference file 306

Read-set file format 306

VCF file 308

APPENDIX B Legacy Formats 309

Introduction 309

GFF to BAM mapping 309

APPENDIX C	Run the SOLiD™ Accuracy Enhancement Tool	313
	Overview	313
	SAET usage guidelines	313
	SAET in analysis modules	314
	SAET input files	314
	SAET parameters	315
	SAET output files	316
APPENDIX D	Administration	317
	Introduction	318
	Log in	318
	Administrate users	320
	Search for users	320
	Add users	320
	Deactivate users	321
	Delete users	321
	Reactivate users	321
	Configure user accounts	321
	Configure licenses	321
	Configure a user's setup	321
	Configure a user's profile	322
	Read-set repository path: notify users	322
	Reset the password	322
	Troubleshooting	323
	Help	323
APPENDIX E	LifeScope™ Genomic Analysis Software v2	325
	END USER LICENSE AGREEMENT	325
	Glossary	337
	Documentation	351
	Related documentation	351
	Obtaining support	351
	Index	353

About This Guide

Purpose

This guide is designed to help you quickly perform next-generation sequencing analyses using LifeScope™ Software. These analyses support fragment, paired-end (PE) and long mate pair (LMP) library types of analyses. Specifically, the following workflows are supported:

- Whole genome resequencing
- Targeted resequencing
- Whole transcriptome resequencing
- Small RNA sequencing
- ChIP-Seq mapping
- MethylMiner™ mapping

Prerequisites

It is assumed that you have working knowledge of the:

- Linux® operating system
- Internet Protocol (IP) address of the LifeScope™ Software cluster.
- Linux environment and know how to:
 - Navigate to directories.
 - Edit and save files in a text editor.
 - Run Linux shell scripts.
 - Run basic Linux commands such as `chmod`, `ps`, `pwd`, `cd`, `echo`, `grep`, and other commands.



1

Introduction to LifeScope™ Genomic Analysis Software

LifeScope™ Software Overview

LifeScope™ Genomic Analysis Software is a modular data analysis bioinformatics tool for performing off-instrument secondary and tertiary analyses on sequence data generated by Life Technologies instruments. The resulting industry-standard files from LifeScope™ Software can be used with third-party visualization and analysis software tools.

After years of development on analysis tools for SOLiD™ System data, in response to customer feedback, Applied Biosystems LifeScope™ Genomic Analysis Software enables fast translation of next-generation data for biologically meaningful results. LifeScope™ Software matches the accuracy of the next generation 5500 Series SOLiD™ Sequencers with Exact Call Chemistry (ECC) and streamlines your data analysis.

Features

LifeScope™ Software is part of the LifeScope™ Genomic Analysis Solution. This informatics solution is comprised of genomics software combined with a specified hardware platform. LifeScope™ Software features:

- Seamless integration with the 5500 Series SOLiD™ Sequencer
- Performance-tuned algorithms for the 5500 Series and ECC Module
- Push-button workflows, intuitive user interface, and secure project management
- Optimized mapping and smaller file formats
- Annotated variant reports, numerous charts, and select visualization tools for simple data interpretation
- Graphically driven configuration of multistep analysis workflows
- Ability to save and reuse workflows
- Ability to resume a workflow without repeating completed analyses
- Secure project-based data management specific to your data analysis
- Projects can be stored and data reanalyzed

Data analyses

Complementing LifeScope™ Software features are the following types of data analyses:

- Whole genome sequencing
- Targeted resequencing
- Whole exome sequencing
- Whole transcriptome RNA sequencing
- Small RNA sequencing
- SNP detection
- Large and small indel detection
- Copy number variation detection

- Inversion detection
- Exon counting
- Junction splicing
- Fusion transcript detection

LifeScope™ Software components

LifeScope™ Software components include:

- The core software server, which is the main software architecture that maintains the interaction of the graphical user interface (GUI) and compute engine. The server works with high performance cluster schedulers to maximize the computation demands of the finely tuned data analysis algorithms.
- The research GUI, which enables scientists to perform mapping and predefined workflows from a desktop computer. The GUI offers auto-generated charts and plots for each analysis type module run within a workflow.
- The command-line user interface, which gives bioinformaticians the power to customize workflows and manipulate every parameter available in each analysis module using specific LifeScope command-line syntax.
- The Admin GUI, which is an administration tool for managing user accounts and licensing permissions.

Workflows

This section describes a standard workflows, a project workflow, and several analysis workflows.

Standard workflow

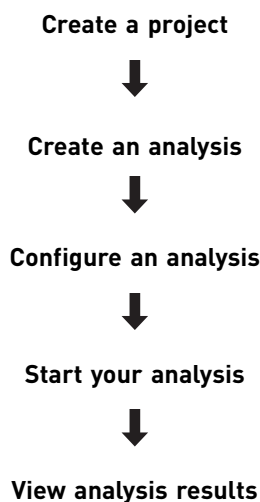
A standard workflow is a built-in series of commonly used analyses, which correspond to a common biological application. Using one of these workflows (data analyses) enables you to run an analysis with a minimum of required setup.

Standard workflows are provided for each of the data analyses listed on .

In LifeScope™ Software, running a standard workflow requires the following steps:

1. Create a project and analysis.
2. Identify your input data (read-set from your 5500 Series Sequencer instrument).
3. Identify the reference genome.
4. Choose the data analysis (see the list [on page 19](#)) to be executed on your data.
5. Start and monitor your analysis.
6. View the results of your analysis.

Project workflow



In *LifeScope™ Genomic Analysis Software Command Shell User Guide* (PN 4465697), these steps are covered in the following chapters:

- Chapter 5, *Understand The LifeScope™ Software Shell* – Introduces projects, analyses, and other concepts required for working with the LifeScope™ command shell.
- Chapter 6, *Run a Command Shell Analysis* – Describes the shell commands, including commands to create a project, create an analysis, configure an analysis, start an analysis run, and view analysis results.
- Chapter 7, *Run a Standard Workflow Analysis* – Describes LifeScope™ Software standard workflows, which provide built-in sequences of experiments corresponding to common biological applications, such as resequencing, targeted resequencing, small RNA, whole transcriptome, ChIP-Seq, and MethylMiner™ mapping.

Analysis workflows

In LifeScope™ Software, a workflow is a series of analysis steps, and each analysis step is dependent only on the steps that precede it. This section describes the flow of secondary and tertiary analyses. Primary analysis is done on the instrument.

Secondary analysis workflows

Secondary analysis always starts with mapping. The input files required by the mapping tool are XSQ (eXtensible SeQuence) files.

The results of mapping and pairing in secondary analysis are used as input for tertiary analysis tools. You can only run the inversion and large indel modules on the results from mapping paired data. The diBayes, CNV, and small indel modules can be executed on the output of mapping either fragment or paired data. The whole transcriptome analysis workflow is different from the resequencing workflows because it uses fragment data to perform its own mapping.

LifeScope™ Software secondary analyses include:

- Mapping

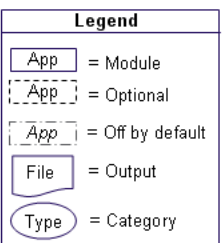
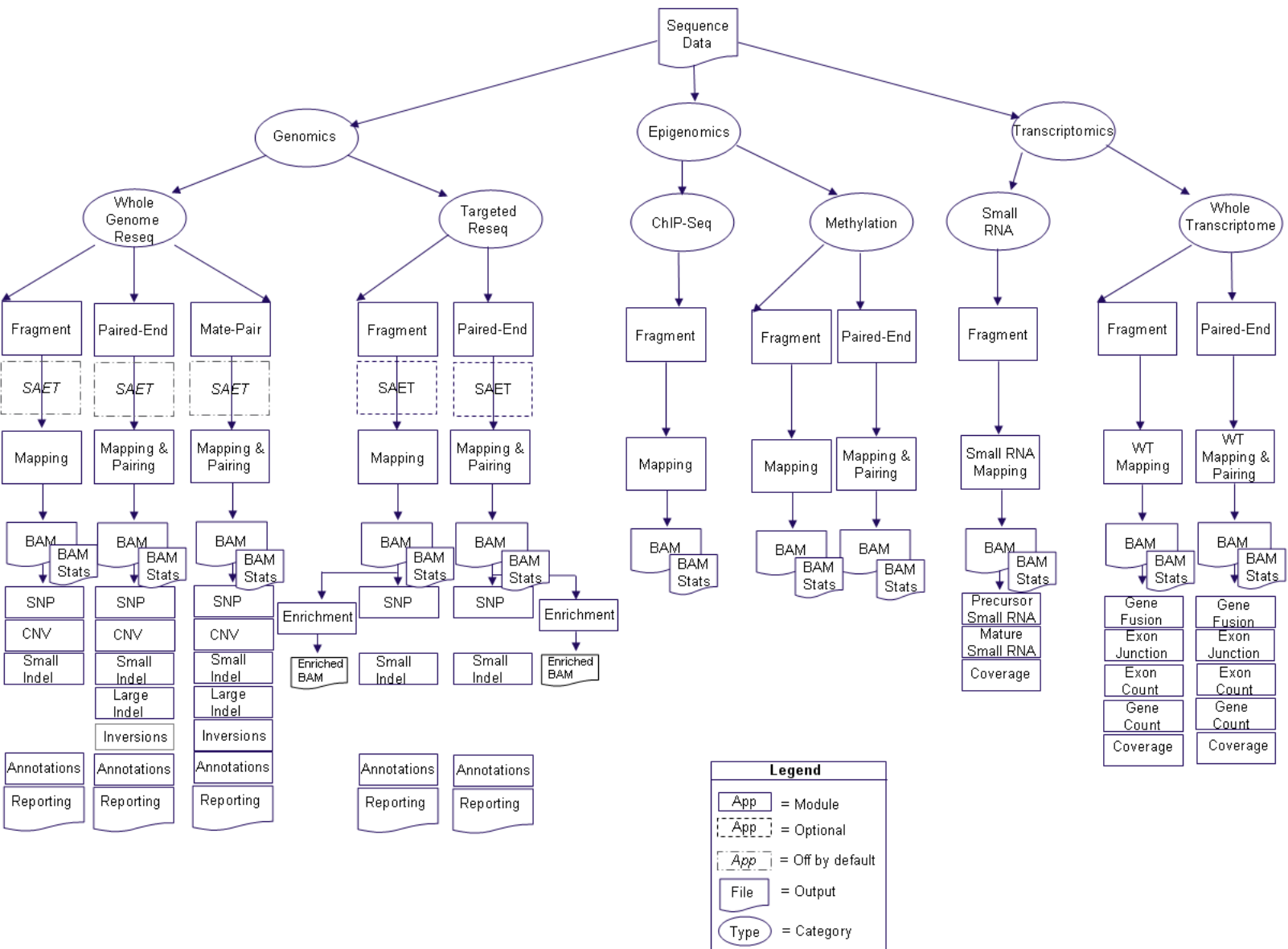
- Mapping statistics
- Reporting
- Whole Genome Resequencing
- Targeted Resequencing
- Whole Transcriptome
- Small RNA
- ChIP-Seq
- MethylMiner™

Tertiary analysis modules

The following table describes the data analysis modules used in workflows.

Analysis	Used in workflows...
Annotations	Whole Genome Resequencing, Targeted Resequencing
Human CNV	Whole Genome Resequencing, Targeted Resequencing
Inversions	Whole Genome Resequencing, Targeted Resequencing
Large indels	Whole Genome Resequencing, Targeted Resequencing
Small indels	Whole Genome Resequencing, Targeted Resequencing
SNPs	Whole Genome Resequencing, Targeted Resequencing
Small RNA count	Small RNA
Small RNA coverage	Small RNA
WT count	Whole Transcriptome
WT coverage	Whole Transcriptome
WT splice finder	Whole Transcriptome

The illustration [on page 23](#) shows the data analysis modules used in predetermined workflows.



The analysis workflows and the modules they execute are described in the following table.

Workflow	Analysis module	Library type	LifeScope™ Software modules involved
ChIP-Seq	ChIP-Seq	Fragment	Secondary: <ul style="list-style-type: none"> • SAET • Fragment mapping • Mapping statistics-
Genomic resequencing	Genomic Resequencing	Fragment	Secondary: <ul style="list-style-type: none"> • SAET • Mapping • Mapping statistics Tertiary: <ul style="list-style-type: none"> • SNP Finding • CNV • Small indels • Annotations
		Mate-pair	Secondary: <ul style="list-style-type: none"> • SAET • Paired mapping • Mapping statistics Tertiary: <ul style="list-style-type: none"> • SNP Finding • CNV • Inversions • Large indels • Small indels • Annotations
		Paired-end	Secondary: <ul style="list-style-type: none"> • SAET • Paired mapping • Mapping statistics Tertiary: <ul style="list-style-type: none"> • SNP Finding • Inversions • Large indels • CNV • Small indels • Annotations

Workflow	Analysis module	Library type	LifeScope™ Software modules involved
MethylMiner™	Methyl Miner	Fragment	Secondary: <ul style="list-style-type: none"> • SAET • Paired mapping • Mapping statistics
		Paired-end	Secondary: <ul style="list-style-type: none"> • SAET • Fragment mapping • Mapping statistics
Small RNA	Small RNA	Small RNA	Secondary: <ul style="list-style-type: none"> • Small RNA mapping
			Tertiary: <ul style="list-style-type: none"> • Small RNA count • Small RNA coverage
Targeted resequencing	Targeted Resequencing	Fragment	Secondary: <ul style="list-style-type: none"> • SAET • Fragment mapping • Mapping statistics
			Tertiary: <ul style="list-style-type: none"> • Enrichment • SNP Finding • Small indels • Annotations
		Paired-end	Secondary: <ul style="list-style-type: none"> • SAET • Paired mapping • Mapping statistics
			Tertiary: <ul style="list-style-type: none"> • Enrichment • SNP Finding • Small indels • Annotations

Workflow	Analysis module	Library type	LifeScope™ Software modules involved
Whole transcriptome	Whole Transcriptome	Fragment	Secondary: <ul style="list-style-type: none"> • WT splice junction extractor • WT fragment mapping • Mapping statistics
			Tertiary: <ul style="list-style-type: none"> • WT counts • WT coverage • Splice finder
		Paired-end	Secondary: <ul style="list-style-type: none"> • WT exon sequence extractor • WT splice junction extractor • WT paired-end mapping • Mapping statistics
			Tertiary: <ul style="list-style-type: none"> • Splice finder • WT counts • WT coverage

Primary and secondary file types

The following table lists the files used in secondary and tertiary analysis.

Analysis type	File type	File name extension	File content	Used in . . .
Primary	Raw reads file	*.xsq	Extensible sequence file	Secondary analysis
Secondary	BAM file	*.bam	Binary Alignment sequence Map (BAM), a generic file format used to store large numbers of nucleotide sequence alignments	Tertiary analysis

Input and output file formats

All analysis modules use specific input files and produce specific output files. The input and output file requirements vary, depending on the type of analysis that you want to perform. The following table provides the names of the input and output file types.

Analysis module	Input file types	Output file types
Mapping and pairing	*.xsq, *.csfasta, *.fasta, *.qv	*.bam
SNP Finding	*.bam	*.gff.3
CNV - singleSample		
Large indel -singleSample		

Analysis module	Input file types	Output file types
Large indel -pairedSample	—	—
Inversion	*.bam	*.gff.3, *.txt
Position error		position error
Whole transcriptome mapping	*.csfasta, *fasta, filter reference fasta, WT *.gtf reference	*.bam

Key

* = one or more

*.gff.3 = public, viewer-oriented *.gff v3

2

Understand LifeScope™ Software

This chapter covers:

■ Overview	29
■ Terminology	30
■ Common scenarios	32

Overview

This chapter describes the concepts and terminology required when working with LifeScope™ Genomic Analysis Software. These terms are explained in more detail in this chapter.

- **Project** – A collection of analyses based on scientific experiments or computational methodologies.
- **Module** – A module is a single step in an analysis workflow. For example, mapping, SAET, and CNV are modules.
- **Analysis** – A set of modules within an analysis workflow. An analysis can contain one or more modules.
- **Analysis workflow** – A pre-defined set of modules in a workflow based on common applications areas, such as targeted resequencing or small RNA analysis.
- **Read** – Sequencing data from a single bead with a single primer set.
- **Read-set** – Sequencing data associated with an indexed (barcoded) sample from one XSQ file.
- **Read-set repository** – An indexed database of eXtensible SeQuence (XSQ) files based on individual read-sets.
- **Group** – A collection of data that can be analyzed together. Grouping allows for aggregated data (for instance, data from different lanes or different runs) to be analyzed as one set of data.

Terminology

This section defines terminology and concepts used to describe LifeScope™ Software

Projects and analyses

Projects are containers for your analysis runs. You may use projects as you wish, to organize your analyses as is convenient and meaningful for you.

You create your projects at the `/projects` level. A project itself does not have a configuration or definition. A project is a way of organizing your LifeScope™ Software analysis runs.

Within each project, you create one or more analyses. Through the configuration of an analysis, you define the executable run.

The configuration of analysis includes these areas:

- The reference genome.
- The reads files.
- The type of analysis to be performed (which LifeScope™ Software modules perform the analysis). The analysis type can be any of the following:
 - A pre-defined workflow, which is a built-in sequence of commonly-used analyses, corresponding to a common biological application.
 - An analysis sequence you define, by either creating your own or customizing one of the LifeScope™ Software pre-defined workflows.

An example of a pre-defined workflow is the targeted resequencing fragment workflow. This workflow includes the following LifeScope™ Software modules:

- SAET
- Enrichment
- Mapping
- Mapping statistics (BAMStats)
- SNPs
- Small indels
- Annotations

Your projects are private and cannot be seen by other users or shared with other users.

Naming restrictions

Project and analysis names must conform to Linux filename conventions:

- Contain only alphanumeric and underscore characters.
- Do not contain spaces.
- Must be unique (among siblings).
- Do not begin with an underscore character (this is a LifeScope™ Software requirement, not a Linux rule).

Reads Data

The data from each lane of a 5500 Series SOLiD™ instrument results in a single XSQ output file (read). Different models of the 5500 support from 6 to 12 lanes, so typically multiple XSQ files are generated per sequencing run. You can direct your LifeScope™ Software runs to analyze data from specific barcodes in multiple lanes, or group the output of multiple lanes or multiple sequencing runs to be processed as a single specimen.

The type of data used in LifeScope™ Software is called a read-set, which is a group of reads belonging to one index (barcode) from one XSQ file. The smallest set of data you can specify in the shell is one index from one XSQ file.

You can optionally use groups to combine more than one index into an analysis. Reads that have been combined into a group are processed as one specimen, even if the read-sets come from different input XSQ files. Each group generates a combined set of output files.

In LifeScope™ Software, input data is added at the project level.

Define input data

Carefully plan your analysis input. The way you define your reads input affects the behavior of your analysis. Index IDs and group names give you flexibility in handling your input data. For example:

- Combine specific read-sets from one XSQ file or multiple XSQ files into one group to be analyzed together. All read-sets within a group must be of the same library type.
- Assign read-sets to different groups, to be analyzed separately within one LifeScope™ Software run. All read-sets within a run must be of the same library type.

Repositories

LifeScope™ Software repositories organize and persist your projects, your input data, and your reference data files.

The reads repository is a storage place in LifeScope™ Software for instrument data intended to be input data for LifeScope™ Software analyses. Data imported into the reads repository are in unmapped XSQ-format files.

Importing reads and references file into LifeScope™ Software repositories is optional, but preferred.

Files imported into the read and reference repositories are visible to all LifeScope™ Software users. However, each user has their own projects repository. The contents of your projects repository are private and cannot be seen by other users.

Note: In the current release, one repository cannot be split across multiple partitions or file systems (the various repositories can each be on a different partition or file system).

The reference repository

The reference repository is initially populated during the installation process with reference files for hg18 and hg19.

Common scenarios

This section describes how LifeScope™ Software handles common biologic analysis scenarios.

Basic

A basic scenario is to run a bioinformatics analysis, such as targeted resequencing, on a sequenced data. The steps to do this in LifeScope™ Software are:

- Create a project
- In that project, create an analysis.
- In that analysis:
 - Add the sequenced data file.
 - Specify the reference genome for this analysis.
 - Specify the type of analysis to be performed on the data.
 - Run the analysis.

A sample sequenced on multiple lanes

In this scenario, one sample is sequenced in multiple lanes on the sequencing instrument. The output data from the sequencing instrument is contained in multiple XSQ files.

The steps to do this in LifeScope™ shell are:

- Create a project
- In that project, create an analysis.
- In that analysis:
 - Add the XSQ data files. When you add the XSQ files, specify the same group name for each file.
 - Specify the reference genome for this analysis.
 - Specify the type of analysis or analyses to be performed on the data.
 - Run the analysis.

Data from multiple samples

This scenario performs, in one run, the same analysis on the sequenced data from multiple specimens. The data can be in one or multiple XSQ files.

LifeScope™ Software keeps the data from each specimen separate, and each specimen is analyzed separately. The steps to do this in LifeScope™ shell are:

- Create a project
- In that project, create an analysis.
- In that analysis:
 - Add the sequenced data file.
 - Specify different grouping for the data.
 - Specify the reference genome for this analysis.
 - Specify the type of analysis (such as mapping, SNPs, and CNV) to be performed on the data.
 - Run the analysis.

3

LifeScope™ Genomic Analysis Software Installation

This chapter covers:

■ Introduction.....	33
■ Prerequisites	33
■ LifeScope™ Software Administration	35
■ Installation workflow overview	36
■ Copy data drive content	37
■ Download LifeScope™ Software	37
■ Install LifeScope™ Software	38
■ Activate the LifeScope™ Software License Key	40
■ Apply the License File to LifeScope™ Software.....	45
■ Check firewall restrictions.....	47
■ BioScope™ Software users' PATH variable.....	47
■ Download documentation and additional resources	47

Introduction

This section describes the LifeScope™ Genomic Analysis Software installation procedure, and system software and hardware requirements. The section also includes high-level LifeScope™ Software installation instructions. For specific details, contact your Life Technologies account representative.

Prerequisites

Some prerequisite procedures in this chapter require that you:

- Know the Linux® operating system
- Know the Internet Protocol (IP) address of the LifeScope™ Software cluster.
- Have a login (ID) on the LifeScope™ Software cluster.
Portions of the install require root access.
- Know how to:
 - Navigate to directories in a Linux environment.
 - Edit and save files in a text editor.
 - Run Linux shell scripts.
 - Run basic Linux commands such as `chmod`, `ps`, `pwd`, `cd`, `echo`, `grep`, and other commands.

Hardware requirements

To successfully install and run the LifeScope application, we recommend that your system meets the following hardware requirements:

CPU Speed	2GHz Minimum
Cores	8 Minimum
Memory	24GB Minimum per compute node
Disk Space	500GB of local or shared storage per node
	200MB for LifeScope installation
	300GB for Reference file installation‡
	2GB for Examples installation
	200GB Approximate space for Stress Test data‡

‡ You can use any type of shared storage to meet this requirement.

System requirements

LifeScope™ Genomic Analysis Software requires a Linux® cluster with a TORQUE, Sun Grid Engine (SGE), or Load Sharing Facility (LSF) resource manager, and a scheduler that can support scheduling policies and dynamic job priorities.

Server requirements

Before you install LifeScope™ Genomic Analysis Software, be sure that your head node and clusters meet the following requirements:

- LifeScope™ Genomic Analysis Software supports only Redhat 4.7 (or later version) or CentOS 4.7 (or later version) distributions on 64-bit platforms.
- At least 2 GB of RAM per core
- PBS/TORQUE v2.3+, SGE v6.2+, or Platform LSF 7 Update 6

Note: If you use SGE, you must create a symmetric multiprocessing (SMP) parallel environment.

Note: If you use LSF, it is highly recommended that LifeScope™ Software jobs are set so that they cannot be preempted by LSF.

Before you install LifeScope™ Software on the head node, pre-install compatible versions of these software packages on *all* compute nodes:

- Perl v.5.8.5 (or later version)
- Python 2.3 (or later version)

GCC Compiler

Ensure that the GNU Compiler Collection (GCC) compiler v3.4.6 or v4.1.2 installed and configured. While LifeScope™ Software does not use the compiler itself, having the compiler installed and configured ensures that the necessary dynamically linked libraries are available.

Client hardware and software requirements

LifeScope™ Genomic Analysis Software supports:

- Windows XP SP3, CentOS 4.x, or Mac OS® X v10.5 or 10.6
- Internet Explorer® 6 or 7, or Firefox 3.0.1 or 3.5
- JRE 1.6
- 2 GB RAM
- 1024 x 768 display monitor or higher
- Internet Explorer® version 6 and later versions
- Mozilla® 3.0.1

LifeScope™ Software Administration

This section uses the acronyms in this list to describe installation roles and locations:

- **SA** – The Linux system administrator for the LifeScope™ Software cluster. The administrator's root access is required for user management and copying the data drive.
- **LSA** – The LifeScope™ Software administrator. The installation instructions recommend creating a new user to be the LSA. The LSA must run the installation script.
- **LSU** – LifeScope™ Software users. These users do not have admin permissions.
- **LSBF** – The LifeScope™ Software Binaries Folder. The LSBF is the installation directory.
- **LSDF** – The LifeScope™ Software Data Folder. The LSDF is a location for LifeScope™ Software repositories, resources, and other files.

To install or run LifeScope™ Software, one person must be designated as the master LifeScope™ Software user. The master user also becomes the default LifeScope™ Software Administrator (LSA). The LSA and regular LifeScope™ Software Users (LSUs) must share a common UNIX® user group.

We recommend that the LSA not share the same user account as the UNIX System Administrator (SA), because of a requirement that all LSUs be members of the LSA's primary UNIX user group. The SA's primary UNIX user group is usually `root`. Non-administrator users typically are not allowed to be members of the `root` user group.

We recommend creating a new Linux user account, `lifescope`, for the LSA. The primary UNIX user group for the `lifescope` user should be the same group as most LifeScope™ Software Users (for example, the group `users`).

After installation, LifeScope™ Software files and folders are owned by the LSA. All application jobs and processes are submitted by the LSA.

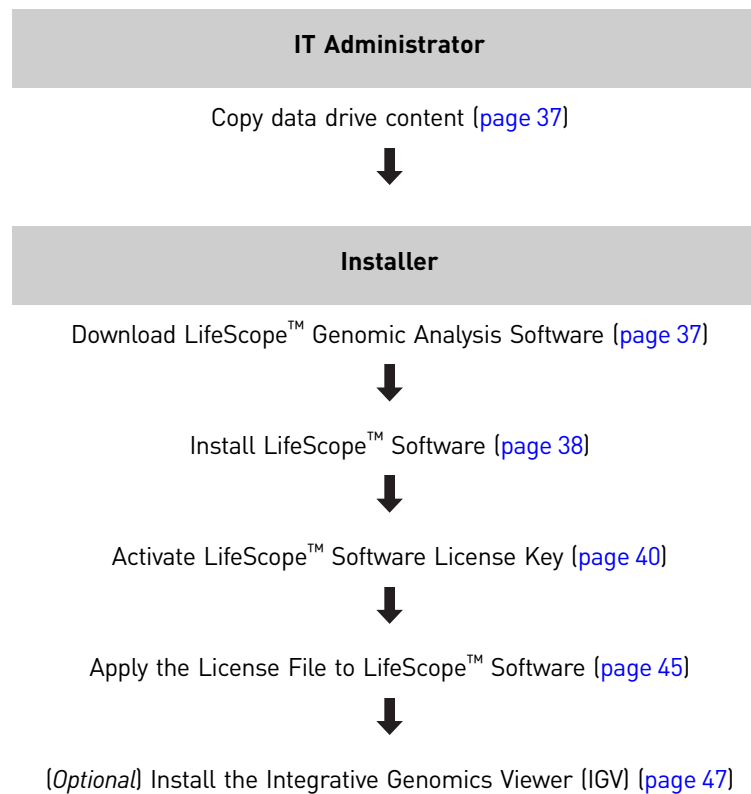
If you are not the LSA or if an LSA does not exist, please contact your System Administrator to have the master user and user group created. You must be logged in as the LSA to install LifeScope™ Software.

Choose two shared folders for holding the software and data in a suitably common location, accessible to all LifeScope™ Software users. Both locations should be accessible (mounted) from the compute nodes.

- Choose the LifeScope™ Software Binaries Folder (LSBF) to install the lifescape software. This folder should be writable to LSA and readable to LSUs. By default this location is `/opt/lifescape`. If the LSBF does not exist, create it with appropriate permissions.
- Choose the location of the LifeScope™ Software Data Folder (LSDF) to hold all LifeScope™ Software resources, repositories, pre-installation tar files, installer, examples, test data, and other files. This location should have read and write permissions for the LSA and all LSUs. This location also should be sufficiently large to hold the initial data and the results of future analysis. By default this location is `/share/lifescape`. If the LSDF does not exist, create it with appropriate permissions.

Make sure that you have the appropriate read and write permissions for the LSBF and LSDF folders, before proceeding with the installation.

Installation workflow overview



This procedure provides typical, general instructions for installing LifeScope™ Software. The actual procedures for your site might vary, depending on the configuration of the LifeScope™ Software cluster and the installation options you select.

Copy data drive content

(System Administrator) Copy the data drive content to the LifeScope™ Software Data Folder (LSDF) on the LifeScope™ Software server, and change owner:group to those of the LifeScope™ Software Administrator (LSA).

The data drive's file system is ext3, which typically is supported only on Linux.

The drive content includes the following:

- Files for use with hg18 and hg19 genomes:
 - Hash tables for the main mapping schemes
 - Mappability files for use with the CNV analysis module
 - Targeted regions files
- LifeScope™ Software reference repository files:
 - Hg18 and hg19 reference files and filter reference files
 - Genomic annotation files (dbSNP)
 - MiRNA database annotation files
- Data files for installation verification

The reference files downloaded in this step become the reference repository for your LifeScope™ Software. The location of the reference files (the `referenceData` directory) is required during installation.

Download LifeScope™ Software

1. (LSA) Visit the LifeScope™ Software project page:

<http://solidsoftwaretools.com/gf/project/lifescopy>

On the project page, you must:

- Enter your order number.
- Accept the EULA.

Download instructions are emailed to you.

2. Go to the download page, as directed by the emailed instructions.
3. Download LifeScope™ Software (.tar.gz file, for example, *LifeScope-2.0r8380-2011040333223344.tar.gz*) to the LifeScope™ Software Data Folder (LSDF).
4. Download the associated installer (`install.sh` file) to the LifeScope™ Software Data Folder (LSDF).

Install LifeScope™ Software

1. (LSA) Go to the downloaded folder LSDF and run the UNIX® command:
`chmod +x install.sh`
 to make the downloaded `install.sh` file executable.
2. Run the installer:
`./install.sh`
3. Installation options include:
 - 1) Install the LifeScope Genomic Analysis Software.
 - 2) Re-configure the LifeScope Software.
 - 3) Request a LifeScope Software license.
 - 4) Register a LifeScope Software license.
 - 5) Exit this application.
 To install LifeScope™ Software, type **1**.
4. At the prompt, type **y** to begin installation.
5. Select the type of installation:
 - 1) Standalone Workstation Server (default)
 - 2) Cluster
 - 3) Workstation w/Remote Submission
6. Enter the location where LifeScope™ Software is to be installed. The default directory is `/opt/lifescopy`.
 Accept the default or change the directory. Provide the location you have chosen for the LifeScope™ Software Binaries Folder (LSBF):
`/path/to/<LSBF>`
 If the directory you enter already exists, you are prompted “Do you still want to use this location?”
7. Enter the location for reference files. Provide the destination location of the data drive contents (from the [“Copy data drive content”](#) step on page 37):
`/path/to/<LSDF>/referenceData`
Note: LifeScope reference files are required for full functionality of LifeScope™ Software. The default directory is `/data/results/referenceData`.
8. Enter the location for reads. You may choose
`/path/to/<LSDF>/reads`
 The default directory is `/data/results/reads`.
9. Enter the location for projects. You may typically choose
`/path/to/<LSDF>/projects`
 The default directory is `/data/results/projects`.
10. Enter the location for Binary Alignment Map (BAM) files. You may choose
`/path/to/<LSDF>/bams`
 The default directory is `/data/results/bams`.

11. Enter the location for the scratch location for LifeScope. The default directory is `/scratch`.

Note: The scratch location is used as a temporary work space for submitted jobs. For best performance, cluster and remote server installations require a scratch location on the local file system of each *compute node* of the system, not on the file system of the *head node*.

12. Enter the IP address or Fully Qualified Domain Name (FQDN) for the LifeScope™ Software server, for example: `192.168.1.27`.

The address of the system where LifeScope™ Software is installed is required to successfully access and run the software. The system address is the IP address or Fully Qualified Domain Name (FQDN) of this system or head node.

Note: Do not use the name `localhost` or IP address `127.0.0.1`. The address should be the public-facing name or IP address of the LifeScope™ Software server. The LSUs access the software using this name or IP address. Typically, the address is the domain name system (DNS) name of the server.

Test the connection to the IP address or FQDN to verify a connection to the system.

Enter the Web URL port address to be used to access the LifeScope Server. Default is 9998.

Configure the license server

1. Enter the IP address or Fully Qualified Domain Name (FQDN) for the LifeScope License Server.
Test the connection to the IP address or FQDN to verify a connection to the system.
2. Enter the port address for the license server to be used with LifeScope™ Software.
3. Select an authentication realm to be used with LifeScope™ Software:
 - 1) LifeScope™ Software (default)
 - 2) Lightweight Directory Access Protocol (LDAP)
 - 3) Host

LifeScope – This authentication realm refers to the LifeScope™ Software application user database. If this realm is chosen, user passwords are maintained in the software database. Passwords are stored as one-way MD5 digests for security reasons. Users must be manually created in this realm. There is no relationship between the application users and Linux® host system users. User Account Management can be done by the LSA using the LifeScope™ Software Admin module (included in this installation).

LDAP – This realm refers to an LDAP-compliant authentication server (OpenLDAP, Active Directory, etc.). If this realm is chosen, user passwords are *not* stored in LifeScope™ Software. User credentials are authenticated against the configured LDAP server. Users need not be manually created in LifeScope™ Software with this realm (though they may be). A valid LDAP server address and LDAP Bind DN must be provided in the configuration. User Account Management can be done by the LSA using the LifeScope™ Software Admin module (included in this installation).

Host – This realm refers to the Linux host machine on which the LifeScope™ Software server is running. If this realm is chosen, user passwords are *not* stored in LifeScope™ Software. User credentials are authenticated against the local machine using SSH (therefore uses whatever authentication scheme SSH uses, for example, NIS, /etc/passwd). There is a one-to-one relationship between the application users and Linux host system users. Users need not be manually created in LifeScope with this realm.

Note: If you change the authentication realm after installation, you must stop and restart the LifeScope™ Software server.

Note: Changing a realm does not remove users from the previous realm. However, users might not be able to login in the new realm, unless they both exist in the new realm and are enabled in the new realm.

4. Select the cluster resource manager to be used by LifeScope™ Software:
 - 1) PBS/Torque
 - 2) SGE (default)
 - 3) LSF
5. Enter the cluster-resource-manager-job-submission-queue-name to be used by LifeScope, for example, *lifescopes*.
6. Enter the number of nodes available on the cluster. Default is 10.
7. Enter the number of cores for each compute node. Default is 8.
8. Enter the memory size (in GB) for each compute node. Default is 32.

IMPORTANT! Allow 2 GB for the compute nodes' OS. For example, for compute nodes with 24 GB, enter 22; for compute nodes with 48 GB, enter 46.

Continue installing LifeScope™ Software

1. At the prompt “Do you want to change any of these entries?” Default is n. To accept the entries, type **n**.
The selected LifeScope™ Software configurations are saved.
2. At the prompt, “Do you want to continue installing the LifeScope package?” type **y** to continue installation.

Activate the LifeScope™ Software License Key

(LSA) These instructions describe how to apply the License Key. The process includes (each of these steps is described below):

- Obtaining the MAC address of the server that has LifeScope™ Software.
- Activating your License Key and obtain a license file specific to your server.
- Applying the license file to activate your copy of LifeScope™ Software.

When applied to LifeScope™ Software, the core license file grants up to five users access to the software for one year.

Prerequisites

Before activating the LifeScope™ Software License Key, please ensure that:

- You have downloaded and installed LifeScope™ Software on computer hardware that meets the LifeScope™ Software minimum hardware requirements (please see www.lifetechnologies.com/LifeScope for more details).
- You are familiar with using command-line Linux®, and you have the necessary permissions to log on to the server that has LifeScope™ Software.
- You have a LifeScope™ Software License Key.

Internet connectivity

Direct internet connectivity for the host computer is highly recommended for the license activation process.

Check the computer date

Check that the computer date matches the current date and time. Each license has a start date. If the computer date is earlier than the license start date, access to the software is disabled until the license start date. If you need to reset the computer date, only reset it *forward*. **Do not** reset the computer date backward; doing so prevents access to LifeScope™ Software. Contact your system administrator (SA) if you need to change the date on the system.

IMPORTANT! Resetting your computer date backwards disables your license key and prevents the operation of LifeScope™ Software.

Obtain the MAC address

IMPORTANT! Run the Linux commands used in this procedure on the computer hardware that has LifeScope™ Software installed on it. If you want to run the LifeScope™ license server on a different machine, then you must obtain the MAC address for that machine.

IMPORTANT! The MAC address used for licensing LifeScope™ Software should be associated with a permanent Ethernet card on the host computer system. If there is a possibility that `eth0` is not a permanent Ethernet card, consult your system administrator to determine which Ethernet card should be used instead. If the Ethernet card associated with the LifeScope™ Software license is replaced on the host system, a new license file must be generated for your account.

To obtain the MAC address of the server that has LifeScope™ Software installed on it:

1. Run the Linux command

```
/sbin/ifconfig
```

A paragraph similar to the one shown [on page 42](#) appears:

```
[corona@foshtddev02 ~]$ /sbin/ifconfig
eth0      Link encap:Ethernet  HWaddr 00:1C:23:BB:E1:F3
          inet addr:10.1.1.1  Bcast:10.1.1.255  Mask:255.255.255.0
          inet6 addr: fe80::21c:23ff:febb:elf3/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:259842678 errors:0 dropped:0 overruns:0 frame:0
          TX packets:296616310 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:214032251167 (199.3 GiB)  TX bytes:339284293393 (315.9 GiB)
          Interrupt:169 Memory:f4000000-f4012800

eth1      Link encap:Ethernet  HWaddr 00:1C:23:BB:E1:F1
          inet addr:167.116.6.72  Bcast:167.116.6.255  Mask:255.255.255.0
          inet6 addr: fe80::21c:23ff:febb:elf1/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:25669852 errors:0 dropped:0 overruns:0 frame:0
          TX packets:82980132 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10648346642 (9.9 GiB)  TX bytes:117855364190 (109.7 GiB)
          Interrupt:169 Memory:f8000000-f8012800

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:11049203 errors:0 dropped:0 overruns:0 frame:0
          TX packets:11049203 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:6069766283 (5.6 GiB)  TX bytes:6069766283 (5.6 GiB)
```

2. In the paragraph labeled `eth0`, the first line should contain text that starts with `HWaddr`. The MAC address follows this text, and should look similar to `00:30:48:9F:72:76`

Find MAC Address with LifeScope™ Installer

Note: In this method, the MAC address is referred to as the “Computer ID.”

If you have difficulty finding the MAC address, the LifeScope™ Installer can find the the address for you. To use the LifeScope™ Installer:

1. From the software download directory, run
`install.sh`

The LifeScope™ Installer information shown below appears:

```
*****
This is the LifeScope Genomic Analysis Software installation and licensing program
*****

1) Install the LifeScope Genomic Analysis Software.

The following options require successful installation of LifeScope:
-----
2) Re-configure the LifeScope Software.
3) Request a LifeScope Software license.
4) Register a LifeScope Software license.

5) Exit this application.

Please select the operation you would like to perform (0/q to quit) [1]: 3

To request a LifeScope product license for operating the application:
Enter the following URL into a Web browser. This will present a form to
request a LifeScope product application license.
*****
* https://licensing.appliedbiosystems.com/activation/lifescopes *
*****

Enter the Computer ID listed below on the form. If this utility is being run on a
system other than the system running LifeScope, this Computer ID is invalid.
Stop and re-run this utility on the system where LifeScope is installed.
*****
* ComputerID: 00:1C:23:BB:E1:F3 *
*****

The License Key you must enter on the form is the license key
you were given with the LifeScope system.

When the form is completed and submitted, a LifeScope product license will be sent
to the email address specified on the form. If multiple people are to receive this key,
their email addresses should be entered in the "CC'ed" field of the form.

Once the LifeScope product license key is received, execute "Step 2" of this utility
to register the license with the License Server.
```

2. Choose option 3 (“Request a LifeScope product license.”) to display the information about product licensing, including the MAC address of the first Ethernet card.

Activate License Key and obtain license file

To activate the license and obtain the license file associated with your MAC address:

1. Visit the LifeScope™ License activation page, shown below, at <https://licensing.appliedbiosystems.com/activation/lifescopes>

LifeScope License Activation

Complete the fields below. Fields with an * are required.

*Computer ID

*License Key

*Email Address

Cc

*Country

*First Name

Middle Name

*Last Name

Organization

*Address (Line 1)

Address (Line 2)

*City

*State/Province/Territory

*Zip/Postal Code

*Phone Number

Remember Me (accept cookie)

[Contact support](#)

2. Enter the MAC address in the **Computer ID** field.
3. Enter the License Key included in the LifeScope™ 2 Genomic Analysis Software license instructions.
4. Enter the email address where your license file (.lic) should be sent.
5. Fill in the remaining information, then click **Submit** to activate your license.

Note: Your license begins the moment you activate your License Key.

The License File “Lifescopes.lic” is sent to you as an attachment in an email from “RightNow AB Technical Support <RightNow.ABTechnicalSupport@lifetech.com>”.

Apply the License File to LifeScope™ Software

(LSA) To apply the license to your copy of LifeScope™ Software:

1. Download the license file (.lic) to the system that has LifeScope™ Software has installed on it.

If you are using another computer for email, you can use tools such as `ftp` or `winscp` to transfer the .lic file to your LifeScope™ Software system. For more information on transferring files, contact your system administrator.

2. Move the .lic file to

`<lifescopy-installed-dir>/server/licenses/`

Your LifeScope™ Software is now licensed.

IMPORTANT! Keep a backup copy of your license file in a safe place.

3. To start the license server and confirm that the license has been correctly applied:

- a. Make sure LifeScope™ Software binaries are in the `PATH` environment variable. If the `PATH` variable is not set properly, set `PATH` using the Linux® command

```
export PATH=<lifescopy-installed-dir>/bin:$PATH
```

You can make this `PATH` setting permanent by modifying your local `.bashrc` file or the global `/etc/bashrc` files. Contact your system administrator for more information.

Note: BioScope™ Software users, see [page 47](#).

- b. Start the license server, using the Linux command

```
lscope-lmgrd.sh start
```

- c. Verify the license status and the available user licenses, shown [on page 46](#), using the command

```
lscope-lmgrd.sh status
```

```
[panakkj1@fospanakkj1d02 licenses]$ cd /share/apps/lifescopeserver/licenses/
[panakkj1@fospanakkj1d02 licenses]$ ls -l
total 8
-rwxrw-r-- 1 panakkj1 panakkj1 1144 May  5 17:50 LifeScope.lic
[panakkj1@fospanakkj1d02 licenses]$ echo $PATH
/share/apps/lifescopeserver/bin:/usr/lib64/qt-3.3/bin:/usr/kerberos/bin:/usr/local/bin:/usr/bin:/bin:/usr/X11R6/bin:/home/panakkj1/bin
[panakkj1@fospanakkj1d02 licenses]$ lscope-lmgrd.sh start
Started License Server

[panakkj1@fospanakkj1d02 licenses]$ lscope-lmgrd.sh status
lmutil - Copyright (c) 1989-2010 Flexera Software, Inc. All Rights Reserved.
Flexible License Manager status on Thu 5/5/2011 17:53

License server status: 27001@fospanakkj1d02.ads.invitrogen.net
  License file(s) on fospanakkj1d02.ads.invitrogen.net: /share/apps/LifeScope-2.0.r0-86630_20110412120300/server/licenses//LifeScope.lic:

fospanakkj1d02.ads.invitrogen.net: license server UP (MASTER) v11.9

Vendor daemon status (on fospanakkj1d02.ads.invitrogen.net):

  lifetech: UP v11.9
Feature usage info:

Users of LTC.BIOSCP.USERS: (Total of 5 licenses issued; Total of 0 licenses in use)
Users of LTC.BIOSCP.LAUNCH: (Total of 1 license issued; Total of 0 licenses in use)
Users of LTC.BIOSCP.CONCURRENT: (Total of 1 license issued; Total of 0 licenses in use)
```

- d. Check that you have activated the same number of licenses that you ordered by cross-verifying the number with the result of the status command.

You have completed the LifeScope™ licensing process.

After the license has been activated, refer to the *LifeScope™ Genomic Analysis Software Command Shell User Guide* (Part no. 4465696) administration appendix for instructions to make LifeScope™ Software available to users.

LifeScope™ Software installation is completed. However, you are recommended to do at least one installation verification step. These steps are optional but confirm that the software is functioning properly and that the LifeScope™ Software Data Folder (LSDF) is properly configured. The optional installation verification steps are:

- The demo analyses
- The performance verification tests

See [Chapter 4, “Test LifeScope™ Genomic Analysis Software”](#) on page 47 for instructions on optionally running example analysis workflows and optionally performing verification tests.

Check firewall restrictions

Ensure that there is no firewall restriction on the host running the LifeScope™ Software server. Consult your system administrator to check the firewall access for LifeScope™ Software users.

BioScope™ Software users' PATH variable

If you install LifeScope™ Software on a machine that also has BioScope™ Software installed and the BioScope™ Software is actively being used, you must follow these instructions:

1. LifeScope™ Software users must *not* use their shell profile script to update the PATH variable with the LifeScope™ Software bin directory. If the LifeScope™ Software bin directory is added to the PATH, BioScope™ Software cannot run correctly.
2. Run BioScope™ Software and LifeScope™ Software in different Linux® shell windows.

Set and export your PATH variable *every* time you open a Linux shell window to run LifeScope™ server or client software.

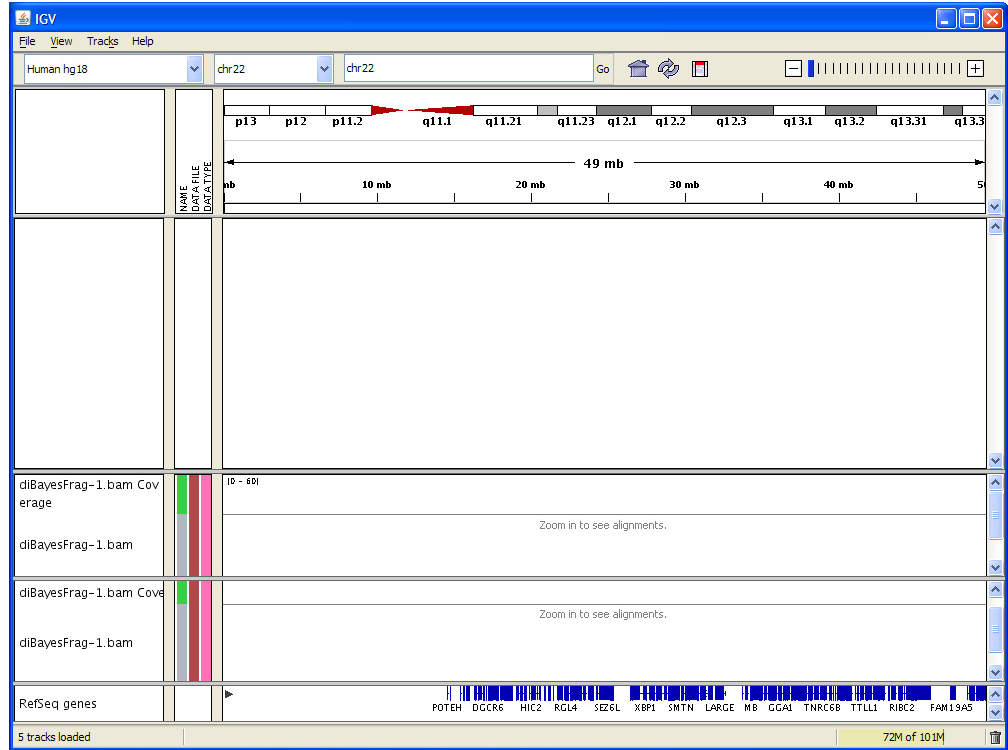
Download documentation and additional resources

To ensure you have the latest version of LifeScope™ Software, please visit www.lifetechnologies.com/LifeScope and look for the Download option. You also find user documentation and the latest information and technical resources. The full resources package includes pre-installation scripts, smaller data files, the installation wizard, and performance verification scripts. Additional links and online support resources are also available at this site.

Integrative Genomics Viewer (IGV)

You can optionally install the Integrative Genomics Viewer (IGV), shown [on page 48](#), available from the Broad Institute. The IGV is a visualization tool for interactive exploration of large, integrated datasets. It directly reads BAM files, which enables you to easily view and inspect alignments against the genome.

For more information about the IGV, go to www.broadinstitute.org/igv/.



4

Test LifeScope™ Genomic Analysis Software

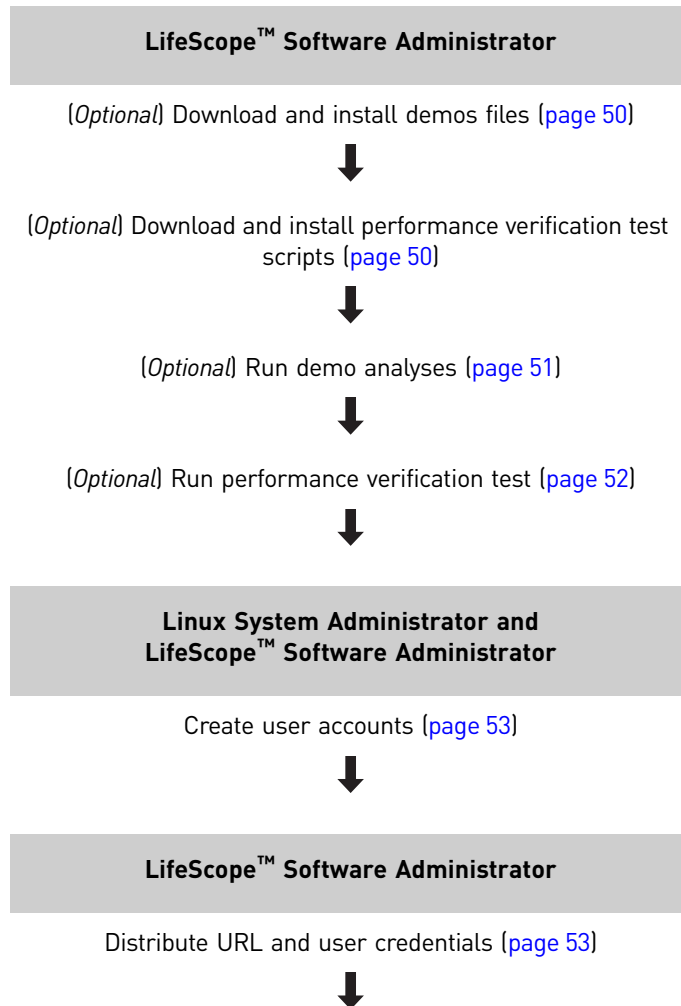
This chapter covers:

■ Introduction.....	49
■ Workflow.....	50
■ Verify the installation.....	50
■ (Optional) Download and install demos files.....	51
■ (Optional) Run demo analyses.....	51
■ (Optional) Download and install performance verification scripts.....	52
■ (Optional) Run the performance verification test.....	52
■ Create user accounts.....	53

Introduction

This chapter describes procedures and tests that the LifeScope™ Software administrators should perform after the software has been installed. LifeScope™ Software users may skip this chapter.

Workflow



Verify the installation

(LSA) Two mechanisms are provided to verify the installation:

- **Demos** – A set of individual analysis modules with small data sets. The demos take approximately 30 minutes or less to complete.
- **Performance verification** – A realistic computation scenario that takes approximately 24 hours to complete.

IMPORTANT! You are strongly recommended to run one or both of the verification procedures. The procedures are listed as optional because you are not required to run both of them. Run one (or both) of the optional verification procedures.

(Optional) Download and install demos files

(LSA) This step is required in order to run the optional sample analyses, “(Optional) Run demo analyses” on page 51.

Example scripts and data for demo LifeScope™ Software analyses are available from the following site:

http://solidsoftwaretools.com/gf/project/lscope_release

Included in this download are the required PLN and INI files for running individual LifeScope™ Software analysis modules, and scripts to run the analyses in the LifeScope™ Software command shell.

Follow these steps to install the demo files:

1. Download the `.tar.gz` file to the LifeScope™ Software Data Folder (LSDF).
2. Go to the LSDF directory and untar the `.tar.gz` file:

```
cd <LSDF>
tar -xzf exampleWorkflows.tar.gz
```

Note: The demo analyses are not recommended as the first experience for users new to LifeScope™ Software. The standard workflows are recommended for new users to learn LifeScope™ Software command-line analyses. Refer to the *LifeScope™ Genomic Analysis Software Command Shell User Guide* (PN 4465697) for the standard workflows.

(Optional) Run demo analyses

(LSA) Run demo analyses, which takes 30 minutes.

These instructions require the following (in addition to the software installation):

- The default reference repository, which includes hg18 and hg19 assemblies. The reference repository is created during the “Copy data drive content” on page 37 and “Install LifeScope™ Software” on page 38 procedures.
- The demos files and sample data installed during the “(Optional) Download and install demos files” on page 51.

Follow these steps to execute the demos in one run:

1. Go to the demos directory:

```
cd <LSDF>/examples
```
2. Execute this command in the Linux shell:

```
./runall.sh
```

This script runs each demo one at a time.

The `runall.sh` script by default executes the run under the `lifescope` user account. To use a different user name, use the `-u` and `-w` options:

```
./runall.sh -u username -w password
```

Check results

The results files are generated in the projects repository, at a location which is set during installation. Its default location is at `/share/lifescopelife/projects`. Within the projects repository directory, the demos results are generated at the following directories:

```
examples/lifescopelife/analysis_name/outputs/sample_name
```

The string `examples` is the project name used by all demos. The string `lifescopelife` is the LSA user name. `analysis_name` is the name of the analysis module used by a demo, and is for example, `cnv`, `dibayes`, `enrichment`, etc. Each `analysis_name` folder has an `outputs` subfolder. The final folder, `sample_name`, is determined by the sample in the input data.

In its results directory, each of the demos analyses creates an output file named `summary.log`. Check for a success message in the last statement of the log.

The results files are overwritten on subsequent runs executed by the same user.

(Optional) Download and install performance verification scripts

(LSA) This step is required in order to run the optional performance verification tests, “(Optional) Run the performance verification test” on page 52.

The performance verification test data file are in the LSDF directory, from the “Copy data drive content” step on page 37.

The performance verification test scripts are available from the following site:

http://solidsoftwaretools.com/gf/project/lscope_release/performanceVerificationTestScripts.tar.gz

Follow these steps to install the performance verification test scripts:

1. Download the file `performanceVerificationTestScripts.tar.gz` to the LifeScope™ Software Data Folder (LSDF).
2. Go to the LSDF directory and untar the `.tar.gz` file:


```
cd <LSDF>
tar -xzf performanceVerificationTestScripts.tar.gz
```

(Optional) Run the performance verification test

(LSA) Run a performance verification test, which takes approximately 24 hours.

The performance verification data files are installed during the “Copy data drive content” step on page 37. The performance verification test scripts are installed during the “(Optional) Download and install performance verification scripts” step on page 52.

Follow these steps to execute the performance verification test:

1. Follow the instructions in the *LifeScope™ Genomic Analysis Software Command Shell User Guide* (PN 4465697) to set the LifeScope™ Software environment.
2. Confirm that the performance verification test data is installed at this directory:

```
<LSDF>/performanceVerificationData
```

3. Go to the performance verification test scripts directory:

```
cd <LSDF>/performanceVerificationTestScripts
```

4. Execute this command in the Linux shell:

```
./runAll.sh -u username -w password
```

where *username* and *password* must be for the LifeScope™ Software administrator or for a valid LifeScope™ Software user account.

Monitor progress

Monitor the progress of your run by checking the output log files and also through job status commands such as `qstat`.

Check results

The results files are generated in the projects repository, at a location which is set during installation. Its default location is at `<LSDF>/projects`. Within the projects repository directory, the performance verification results are generated at the following results directories:

```
username/performanceVerification/analysis_name/outputs
```

Replace *username* with the user name used to run `runAll.sh`. The default is `lifescope`, the LSA account. The string `performanceVerification` is the project name used by the performance verification test. The analysis names used by `runAll.sh` are `genomicResequencing.20x`, `genomicResequencing.2x`, and `wholeTranscriptomePE`. The `outputs` subfolder name is fixed.

The performance verification tests creates a log file in the *analysis_name* directory. Check for a success message in the last line of each log file.

The results files are overwritten on subsequent runs executed by the same user.

Create user accounts

(LSA or SA) Create user accounts.

For instructions in this section that require using the LifeScope™ Software Admin Portal, refer to the administration appendix in the *LifeScope™ Genomic Analysis Software Command Shell User Guide* (Part no. 4465696).

The method for user account creation depends on the authentication realm for your installation:

- **LifeScope realm** – (LSA) The LifeScope™ Software administrator creates LifeScope™ Software user accounts using LifeScope™ Software. The LifeScope™ Software Admin Portal is recommended for user administration.
- **LDAP realm** – (SA and LSA) The Linux system administrator creates user accounts on the LDAP server. The LifeScope™ Software administrator must enable each user in the LifeScope™ Software Admin Portal.
- **Host realm** – (SA) The Linux system administrator creates Linux user accounts on the local operating system.

Enable users

(LSA) This step is not required if your installation uses the concurrent licensing scheme. (With concurrent licensing, all users in the authentication realm are automatically enabled for LifeScope™ Software.)

If your installation uses a named licensing scheme, the LSA must use the Admin Portal to identify the users for the named license.

Distribute the URL and user credentials

(LSA) The LifeScope™ Software URL follows this pattern:

`http://<fqdn of host>:<port number>/LifeScope.html`

For example, for an installation on host `orange.intranet.company.com`, using the default port 9998 for the LifeScope™ Software server, the URL is:

`http://orange.intranet.company.com:9998/LifeScope.html`

Notify LifeScope™ Software users (LSUs) of the LifeScope™ Software URL.

If new user accounts are created for LifeScope™ Software users, notify them of the username and password credentials required to access the LifeScope™ Software.

Ensure that there is no firewall restriction on the host running the LifeScope™ Software server. Consult your system administrator to check the firewall access for LifeScope™ Software users.

PART I

Overview

5

LifeScope™ Genomic Analysis Software Graphical User Interface

This appendix describes the components of the Applied Biosystems LifeScope™ Genomic Analysis Software graphical user interface (GUI).

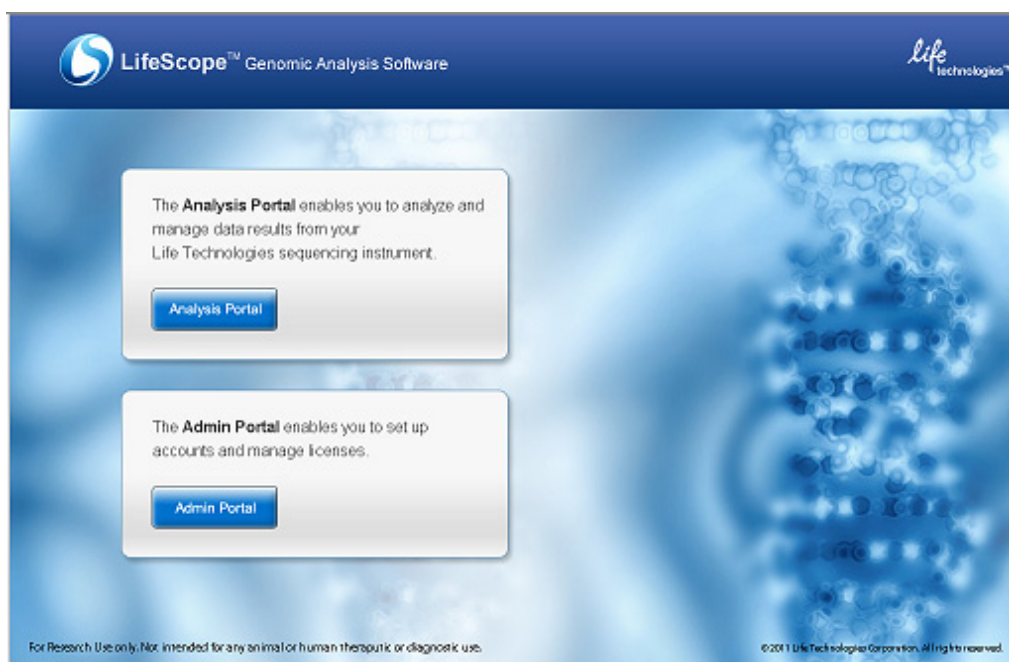
Log in, log out

Note: For a description of the graphical user interface, see [Chapter 5, “LifeScope™ Genomic Analysis Software Graphical User Interface”](#) on page 57.

1. To start LifeScope™ Software, go to:

`http://<IP address>:<port number>/LifeScope.html`

where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.



- To run LifeScope™ Genomic Analysis Software, click **Analysis Portal**,
 - (Administrator) To run the LifeScope™ Software Admin tool, click **Admin Portal**.
2. In the Welcome window, type your username and password, then click **Login**.
If you have forgotten your password, click **Forgot Password**, then follow the prompts. If you need assistance, click **Login Help**.

Welcome to
LifeScope

Genomic Analysis
Made Easy

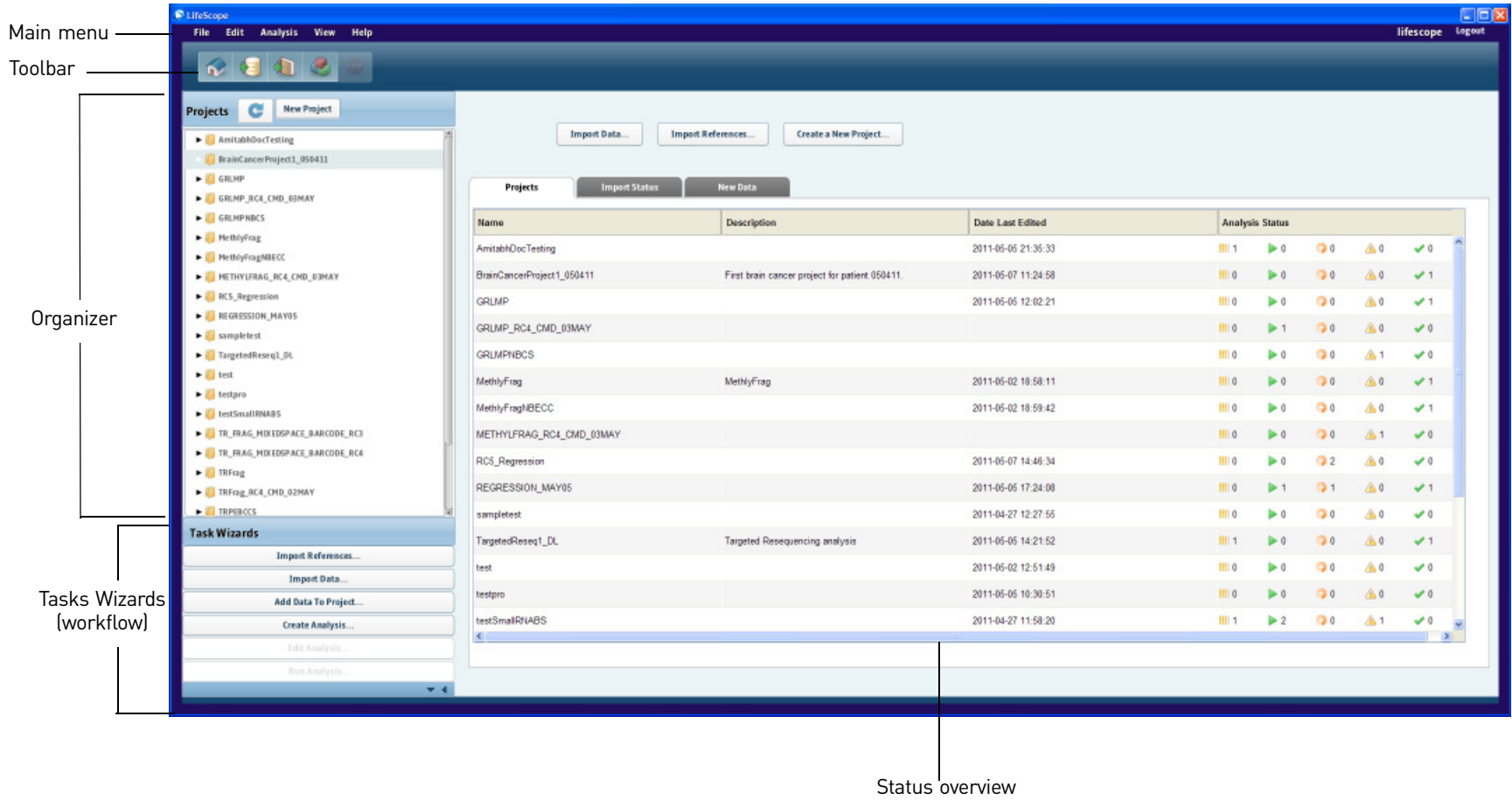
Login

* Username:






* Password:

[Login Help](#) [Forgot My Password](#) [Login](#)






After you log in, the home view, shown [on page 59](#), appears. The view shows a toolbar, a left panel that has a Projects and Task Wizards section, and a right panel that shows the status overview of a project and analysis results.



The following table describes the areas, components, and purpose of components in the LifeScope™ Software GUI.

Area	Components	Purpose
Main menu with action buttons	File	Create a new project, delete a project, add data, import data, import references, and exit LifeScope™ Software.
	Edit	Copy, paste, and delete text.
	Analysis	Create, edit, delete, reuse, run, and stop an analysis.
	View	View a project overview, project status, parameters, reads, and BAM data.
	Help	Open the user guide or tutorial; view information about LifeScope™ Software; and access Life Technologies Corporation on the Web.
Toolbar with navigation and action buttons		Return to the LifeScope™ Software home view.
		Import data into the read repository.
		Import references into the read repository.
		Create an analysis. Faded (inactive) until a project is selected in the organizer.
		Stop running an analysis. Faded (inactive until a project is running).

Area	Components	Purpose
Projects (left panel)		Refresh the list of projects to see project status.
	New Project	Create a project.
	Organizer	
		Show contents. Click to show the contents of a folder, analysis, or module.
		Hide contents. Click to hide the contents of a folder, analysis, or module.
		Project name. Click once to show data and analyses, and the project overview and status.
		Data. Click once to show the Read-sets and BAM Data in the status overview pane.
		Analysis, created and ready to review, edit, and run. Click once to show the analysis overview, status, and parameters.
		Analysis running.
		Analysis completed.
	 bac_analysis2	Analysis failed.
		Analysis module. Click to show the analysis results and the analysis overview, status, and parameters.
		Analysis module completed.
	 diBayes	Red text indicates that analysis module failed.
	Results. Click to show analysis results in the status overview (described in this table).	

Area	Components	Purpose
Tasks Wizards buttons (left panel)	Import References (into an analysis)	Utilities for performing tasks, the order of the buttons guides you in following the workflow. Clicking a button shows steps, which are highlighted to show progress through a workflow. Add Data To Project is faded (inactive) until a project is selected. Create Analysis is faded until a project is created. Edit Analysis is faded until an analysis is selected. Run Analysis is faded until an analysis is selected.
	Import Data (into an analysis)	
	Add Data To Project	
	Create Analysis	
	Edit Analysis	
	Run Analysis	
Status overview (right panel)	Tabs, data entry fields.	Import data and references; create a new project. Show analysis status information, task results, and project details. Tabs include data entry fields, buttons, and interactive tables. Placing your mouse cursor over an icon displays brief descriptions of the icons.
		Analysis ready to run.
		Analysis not ready to run/or stopped
		Analysis running.
		Analysis finished successfully
		Analysis failed.

Button labels and table titles

Button labels and table titles reflect the data type. For example, when you group XSQ data, the data is listed in a table titled “Available Data in Project.” When you group BAM data, the table is titled “BAM Data in Project.”

Pop-up windows

Small pop-up windows can be hidden by other, larger windows that are open. If you are unable to perform an action in a pop-up window, move, minimize, or close windows that you do not need open.

PART II

Getting Started

6

Create and Manage Projects

This chapter covers:

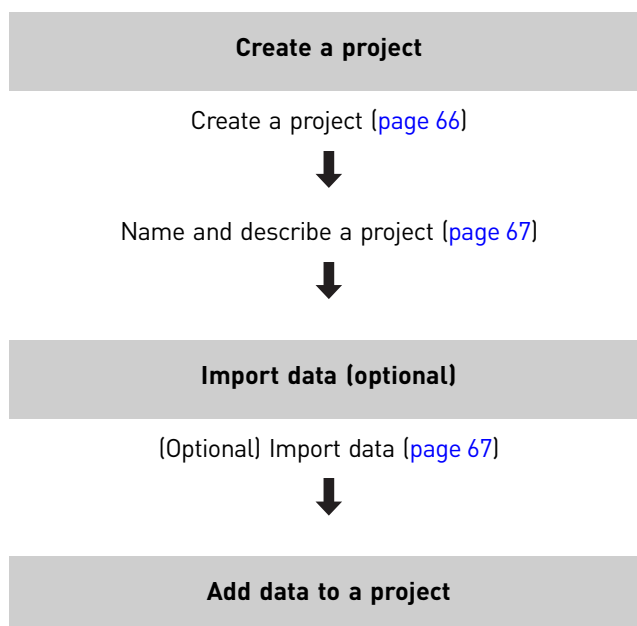
■ Overview	65
■ Workflow	65
■ Create a project	66
■ (Optional) Import data	67
■ Add data to a project	68
■ View project data	72
■ Delete a project	73

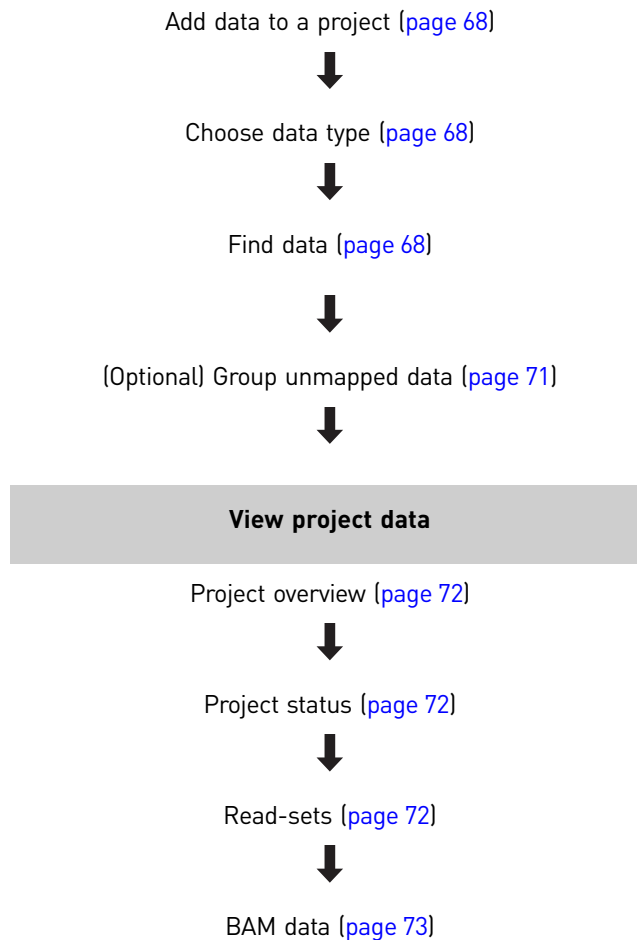
Overview

The first step in using LifeScope™ Genomic Analysis Software to perform analysis of next generation sequencing data is to create a project. Creating a project involves importing data and adding data.

Workflow

This following table illustrates a standard workflow for creating a project.





Create a project

Prepare to create a project

Determine your input and reference files

Before you create a project, you must know:

- The read and references files required by your analyses.
- The regions of interest file, for targeted resequencing analyses.
- Any non-default module parameter values.

Log in to LifeScope™ Software

1. If you are not already logged in, navigate to LifeScope™ Software at `http://<IP address>:<port number>/LifeScope.html` where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown on page 59).

Create a project

1. In the home view, there are three ways to create a project:

- Click **File** in the top menu, then select **New Project**, or
- Select the **New Project** button in the Projects panel (see [page 59](#)), or
- In the home view, click **Create a New Project** button in the status overview panel.

To exit the Enter Project Name view, click any option in the top menu, any project in the Projects organizer, or any button in the Task Wizards section.

Name and describe a project

In the Overview tab (see [page 59](#)), provide the following details about a project:

- Name of the project, for example, "Cancer_Study_3."
A project name can contain only letters, numbers, and the "_" symbol. Other special characters are not allowed.
- A description of the project, for example, "Clinical trial: pancreatic cancer. Transcript analysis experiment. Time-course samples covering treatment period."

After you have named and described your project, click **Create New Project**.

Your project appears in the Projects list. If necessary, scroll down the list of projects to see your newly created project. Now you are ready to import references and data that you can use in your project.


(Optional) Import data

Import data into the reads repository. The reads repository is a storage place in LifeScope™ Software for instrument data intended to be input data for LifeScope™ Software analyses. The XSQ, CSFASTA/QUAL data can be imported into the read repository. Data files that are not already in the XSQ file format to be converted to that format.

Files imported into the read and reference repositories are visible to all LifeScope™ Software users. However, each user has his or her own projects repository. The contents of a user's projects repository are private and cannot be seen by other users.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you import are distinct.


Import data from a desired location, such as your instrument computer or storage system, into the read-set repository. The type of data is called a read-set, which is a group of reads belonging to one barcode from one XSQ file.

1. Import data by using one of the following methods:
 - In the home view (see [page 59](#)), click the **Import Data** button (see [page 60](#)).
 - In the main menu (see [page 59](#)), click **File**, then **Import Data**.
 - In the toolbar (see [page 59](#)), click the **Import Data** button .
 - In the Task Wizards panel (see [page 59](#)), click the **Import Data** button.

The Import Data window appears.

2. In the Choose Data Type view of the Import Data window, select the type of data to be imported:
 - Raw unmapped (XSQ, CSFASTA/QUAL), or
 - Already-Mapped (BAM)

then click **Next**.

3. In the Specify Data Files view, click the  button next to the Data File 1 data entry field to open the repository file browser.
4. In the pop-up window, navigate to the data file you want to import, select the file, then click **OK**.

Note: You can copy the data, move the data to a location on the LifeScope™ Software server or make a symbolic link to the original data. These options are available in the upper-right corner of the Import Data · Specify Data Files view.

To add another data file, click that button.

To refrain from importing data, click **Cancel**.

To import the data and close the window, click **Finish**.

Note: After you import a file, it will not appear in the reads repository until the file has been completely converted.

5. After the data has been converted, click **Add Data**.

Add data to a project

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

Choose data type

In the Choose Data Type view of the Add Data to Project window, select the type of data to add to your project:

- Raw unmapped (XSQ) data, or
- Already-Mapped (BAM) data

then click **Next** to find data. If you do not want to add data, click **Cancel** to close the Add Data to Project window.

Find data

If you chose raw, unmapped (XSQ) data, there are two tabs with read repository filters you can use to find data:

Basic	Use this filter to search for data and view basic details about found data.
Advanced	Use this filter to refine searches and view more specific details about found data.

Within each tab are data entry fields, buttons, and a table of reads (sequencing data). Use the arrow buttons or page number button at the bottom of the table to navigate the table.

Find raw, unmapped data with the Basic filter

Find data by specifying details about a read.

1. Specify any or all of the following read details by:

- Typing the XSQ ID.
- Typing the barcode.
- Typing the run start time and end time.
- Selecting the library type.

To clear the search results, click **Clear Filters**.

2. Click **Filter** to see search results in the read repository table.

3. Click the checkboxes in the left column to select data, then click **Next** to group data.

Read Repository Filter

Basic | **Advanced** | [Clear Filters](#)

XSQ ID: Barcode:

Run Start Time From: To: Library Type:

<input type="checkbox"/>	XSQ ID	Barcode	Application	Data Space	Lane #	Run Start Time	Run End Time	Library Type	Read Length
<input type="checkbox"/>	diBayesFrag.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-24 13:47:00.0	2011-02-24 13:47:00.0	Fragment	
<input type="checkbox"/>	case2001_MatePair.xsq	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	MatePair	
<input type="checkbox"/>	diBayesPE.xsq	DefaultLibrary	Resequencing	Color Space	255	2011-02-01 10:30:15.0	2011-02-01 10:30:15.0	PairedEnd	
<input type="checkbox"/>	helen_Test1111.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-09 17:00:00.0	2011-02-09 17:00:00.0	Fragment	
<input type="checkbox"/>	helen_Test2222.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-09 17:00:00.0	2011-02-09 17:00:00.0	Fragment	
<input type="checkbox"/>	test_S1_F31301452711...	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	Fragment	
<input type="checkbox"/>	helen_Test22.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-09 17:00:00.0	2011-02-09 17:00:00.0	Fragment	
<input type="checkbox"/>	test_S1_F31301463425...	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	MatePair	
<input type="checkbox"/>	solid0054_20110102_P...	DefaultLibrary	Targeted Resequencing	Color Space	255	2011-02-12 09:34:30.0	2011-02-12 09:34:30.0	Fragment	
<input type="checkbox"/>	WT_chr17_PE.xsq	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	PairedEnd	
<input type="checkbox"/>	chr6reads.xsq	DefaultLibrary	NONE	Color Space	0	2011-01-13 15:34:36.0	2011-01-13 15:34:36.0	MatePair	50,0
<input type="checkbox"/>	test_S1_F31301064200...	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	Fragment	
<input type="checkbox"/>	test_S1_F31300885555...	DefaultLibrary	Unknown	Color Space	255	2011-01-11 20:10:20.0	2011-01-11 20:10:20.0	Fragment	
<input type="checkbox"/>	helen_Test111.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-09 17:00:00.0	2011-02-09 17:00:00.0	Fragment	
<input type="checkbox"/>	helen_Test222.xsq	DefaultLibrary	Unknown	Color Space	255	2011-02-09 17:00:00.0	2011-02-09 17:00:00.0	Fragment	

1 of 2

Filter

Reads

Find raw, unmapped data with the Advanced filters

The Advanced tab includes an AND filter and an OR filter. To narrow your search for a specific read:

1. Click the Edit AND Filter button.

Click the Edit OR Filter button.

In both filters, you can refine the display of found information by specifying the table column, relation, and value.

- To add a specification in the Edit AND Filter, click the **&** button.
- To add a specification in the Edit OR Filter, click the **%** button.
- To remove a specification in both filters, click the **X** button.

Click **Filter** to see search results in the read repository table.

2. Select the found data that you want to add, then, click **Next** to optionally group data. If you do not want to add data, click **Cancel** to exit the Add Data window.

Find BAM data

If you chose Already-Mapped (BAM) data, do the following procedure to find data:

1. In the Find Data BAM view, click  to open the Browse to BAM File window.
2. In the BAM File Folders column, select a folder to see its contents in the BAM Files column. Select a BAM file, then click **OK**.

The file appears in the BAM File data entry field.

Click **Add BAM to Project**.

The BAM data is added in the project.

(Optional) Group unmapped data

You can optionally group unmapped (XSQ) read-sets data to add to a project. In the Add Data · Group Data view, the available data in the project is shown in the Read-Sets in Project table.

Note: The Add Group to Project button is faded (inactive) until you select a read-set. The Add Read-Set to Selected Group button is faded until you select a group.

To group data:

1. Select the checkbox of the data to create a group of read-sets, then click the **Add Group to Project** button.
2. In the Group Read-sets window, type a group name, then click **OK**.

A project name can contain only letters, numbers, and the “_” symbol. Other special characters are not allowed.

A table with the newly created group appears in the Groups in Project section. Below the table are Edit and Delete buttons:

- Use the Edit button to change the name of a group. You can also delete a read-set from a group, unless the group has only one read-set. A group must contain at least one read-set.
- Use the Delete button to delete a group.

To select a group for a project, click its checkbox in the Groups in Project table.

To save the group, click **Finish**. To exit the Add Data to Project window without saving the group, click **Cancel**. To proceed, click **Add Analysis** and follow the procedure “Choose data” on page 79.

Group BAM data

In LifeScope™ Software, an imported .bam file is treated as a group of data. You cannot create a group of multiple, imported .bam files for tertiary analysis.

To group multiple .bam files and use the group for tertiary analysis:

1. Create a group of multiple .xsq files and run an analysis.
After secondary analysis (mapping), the group of .xsq files is created as a .bam file.
2. Use the created .bam file for tertiary analysis.


View project data

You can view a project overview, status, parameters, read-sets, and BAM data. To view project data:

1. Select a project in the Projects organizer (shown).

There are two ways to view project details. Select a project in the Projects organizer (shown on page 59), then:

- Click **View** in the top menu, then select the details that you want to view, or
- Click the Overview and Status tabs.

To view the Read-Sets and BAM Data tabs, click  on the subfolder under the project name in the Projects organizer.

Project overview

In a project overview, you can view and edit the name of a project, its description, and the persons to be alerted about the project: To show the project overview, either:

- Click **View** in the main menu, then **Project Overview**, or
- Select a project in the Projects list to display the Overview tab in the status overview pane.

To save changes that you make in the Overview tab, click **Apply Changes**. To cancel your changes, click **Revert to Last Saved Settings**.

Project status

To show the status of a project, select a project in the Projects list, then either:

- Click **View** in the main menu, then **Project Status**, or
- Click the **Status** tab to switch from the Overview tab in the status overview pane.

Read-sets

To show the read-sets in a project, select a project in the Projects list, then either:

- Click **View** in the main menu, then **Project Read-Sets**, or
- Select a project in the Projects list, click **Data**, then (if necessary) the **Read-Sets** tab in the status overview pane.

In the Data, Read-Sets view, click the **Create Analysis of Data** button below the Read-Sets tab to create an analysis from the listed read-sets.

BAM data

To show the BAM data in a project, select a project in the Projects list, then either:

- Click **View** in the main menu, then **Project BAM Data**, or
- Select a project in the Projects list, click **Data**, then (if necessary) the **BAM Data** tab in the status overview pane.

In the Data, BAM Data view, you can create an analysis from the listed BAM data by clicking the **Create Analysis of Data** button below the BAM Data tab.

Delete a project

To delete a project, click **File** in the top menu, then select **Delete Project**.

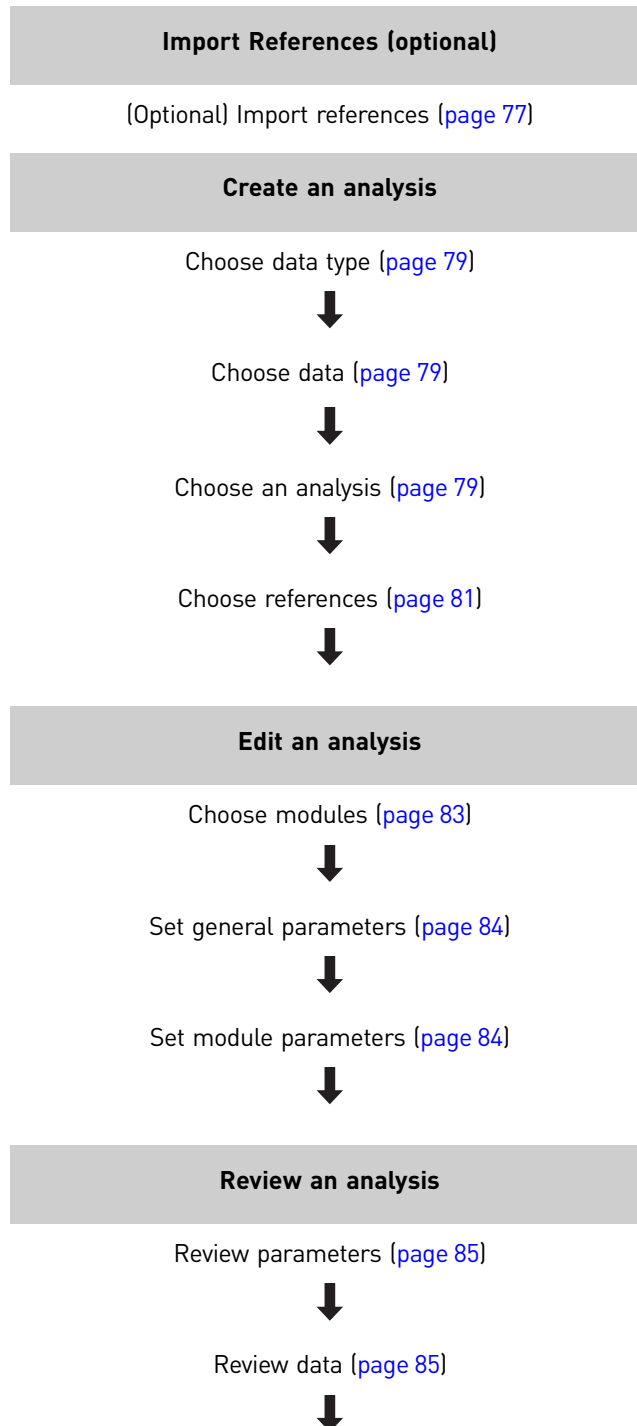
7

Perform an Analysis

This chapter covers:

■ Workflow	76
■ (Optional) Import references	77
■ Create an analysis	78
■ Reuse an analysis	81
■ Edit an analysis	82
■ Review your analysis	85
■ Run and monitor your analysis	85

Workflow



Run an analysisStart your analysis now ([page 85](#))Stop an analysis ([page 85](#))

(Optional) Import references

Import genomic repository reference files and folders from another directory into the reads repository. The reads repository is a storage place in LifeScope™ Software for instrument data intended to be input data for LifeScope™ Software analyses. Data imported into the reads repository are in unmapped eXtensible SeQuence (XSQ)-format files.


For more information about repositories, see “[Repositories](#)” on [page 31](#).

Import references

1. Access the Import References window by using one of the following methods:
 - In the home view (see [page 59](#)), click the **Import References** button (see [page 60](#)).
 - In the main menu (see [page 59](#)), click **File**, then **Import References**.
 - In the toolbar (see [page 59](#)), click the **Import References** button.
 - In the Task Wizards panel (see [page 59](#)), click the **Import References** button.

The Import References window appears.

2. Import a reference file and/or repository folder.

Click the  buttons to open file browsers and navigate to a specific reference file and/or repository folder, for example:

```
/data/analysis/referenceData/lifetech/hg19/human_filter_reference.fasta
```

```
/data/analysis/referenceData
```

Click **OK**.

To add more reference files, click **Another Reference File**.

3. To import the references, click **Close**, or to close the Import References window without importing references, click **Cancel**.

The location of reference repository is set by the LifeScope™ Software system administrator. If you are unable to locate reference files, please contact the system administrator.

You can check the progress of a file being imported by clicking the **Import Status** tab in the home view (shown [on page 59](#)). A COMPLETED message in the File Transfer Status column indicates that the imported file has been added to the reads repository.

Import your own references

You can also import your own references. Before importing your own references, validate them to prevent analysis errors or failure. Follow the procedure [“Validate the reference file” on page 30](#).

We recommend that you keep your original reference file and write a validated version of the reference file.

For more information about the reference repository, refer to the

View import status

To view the import status of files, click the Import Status tab.

Files that were successfully imported are indicated as COMPLETED in the File Transfer Status column.

Files that were not imported are indicated as FAILED! in the Conversion Status column. Failed imports cannot be removed in the LifeScope™ Software graphical user interface, but can be removed in the LifeScope™ Software command shell. For instructions on how to remove failed imports, refer to the *LifeScope Genomic Analysis Software Command Shell User Guide* (PN 4465697).

Create an analysis

Creating an analysis involves choosing a data type, data, an analysis type, and references. Analysis types are predefined workflows. You can create a new analysis or reuse an existing analysis to create an analysis. For instructions on reusing an analysis, see .

The following table lists the types of analysis, analysis modules, and references to chapters in this user guide.

Analysis Type	Includes	Refer to . . .
Genomic Resequencing	SAET, mapping, pairing, base translation, SNP finding, small indel, CNV, large indel, inversion, annotation	“Perform Genomic Resequencing Analysis” on page 129
Whole Transcriptome	Mapping, pairing, counts, coverage, splice finder exons, SASR gene fusions	“Perform Whole Transcriptome Analysis” on page 163
Targeted Resequencing	SAET, mapping, pairing, base translation, enrichment, SNP finding, small indel, annotation	“Perform Targeted Resequencing and Enrichment Analyses” on page 95
Small RNA	Mapping, counts	“Perform Small RNA Analysis” on page 149
ChIP-Seq	Mapping	“Perform ChIP-Seq Mapping” on page 213
MethylMiner™	Mapping	“Perform MethylMiner™ Mapping” on page 203

Start analysis creation

There are two ways to create an analysis:

- Select a project from the Projects list, click **Analysis** in the toolbar (see [page 59](#)), then click **Create**.
- In the Task Wizards section, click **Create Analysis**.

The Create Analysis · Choose Data Type view appears in the status overview pane (see [page 59](#)).


Choose data type

Choose unmapped (XSQ) or mapped (BAM) data, then click **Next** to choose data.

A table appears in the status overview pane. If you chose XSQ data, you will see an Available Data in Project table. If you chose BAM data, you will see a BAM Data in Project table.

BAM data

When you choose BAM data:

1. In the Add Data To Project window, click the  button, next to the BAM File data entry field, to open the file browser.
2. In the Browse to BAM File window, navigate to /data/results, select the BAM file that you want to add to your project, then click **OK**.
3. Click **Add BAM to Project**.

The BAM file appears in the BAM Data in Project area.

To delete a data file, click the **X** button. However, if there is only one imported data file, you cannot delete it.

If you decide to not add BAM data and instead add XSQ data to your project, click **Back**.

To refrain from adding the BAM data, click **Cancel**.

To add the BAM data and close the window, click **Finish**.

To proceed, click **Add Analysis**.

Choose data

In the Available Data in Project table, click the checkboxes of the data for analysis, then click **Next** to choose an analysis.

Choose an analysis


Note: If you added BAM data to your project, do not select the ChIP and Methyl Miner analysis modules.

In the Create Analysis · Choose Analysis view:

1. Select either **Create New Analysis** or **Reuse Old Analysis**.
2. Define the analysis with a name and description, and select an analysis type. The following table describes the analysis types.

Analysis Type	Description
Targeted Resequencing	Analysis method designed for the analysis of target-enriched sequencing data. Targeted Resequencing supports mapping, SNPs, and small indels analyses. See Chapter 9, “Perform Targeted Resequencing and Enrichment Analyses” on page 95 .
Genomic Resequencing	A workflow for identifying individual variants and structural variations in a genomic sample.

Analysis Type	Description
ChIP-Seq	A combined assay and sequencing technique for identifying and characterizing elements in protein-DNA interactions. See Chapter 15, "Perform ChIP-Seq Mapping" on page 213 ,
Whole Transcriptome	Analysis method in LifeScope™ Software that aligns to a reference genome. Using the mapping results, WTA counts the number of tags aligned with exons. See Chapter 12, "Perform Whole Transcriptome Analysis" on page 163 .
Methyl Miner	A system for the enrichment of methylated sequences from genomic DNA, that, with the use of SOLiD™ System sequencing, allows for focused evaluation of methylation patterns in genome-wide studies. See Chapter 14, "Perform MethylMiner™ Mapping" on page 203 .

To see a list of modules for an analysis type, place your cursor over a  button. The following table describes the analysis modules.

Analysis Module	Description	Used in Analysis Type
Coverage	This module calculates read coverage per position.	Small RNA Whole Transcriptome
Enrichment	This module creates statistics that provide a means to assess enrichment platform performance by: <ul style="list-style-type: none"> Looking at variations in coverage, both across all targets and on a per-target basis. Addressing the uniformity and completeness of coverage within the target region. Calculating the degree of enrichment. 	Targeted Resequencing
Exon count	This module counts and annotates the number of exons that are expressed in a gene region.	Whole Transcriptome
Exon junction	This module discovers splice junctions (introns from pre-messenger-RNAs to generate mature messenger-RNAs) and annotates the type and evidence for each discovered junction.	
Gene count	This module counts the number of reads that align within genomic features.	
Gene fusion	This module uses mapped reads to find gene fusions.	
Mapping: fragment	This module generates BAM files from gapped alignments and ungapped alignments.	ChIP-Seq Genomic Resequencing MethylMiner Targeted Resequencing Whole Transcriptome
Mapping: paired-end	This module matches reads from the F3 and F5-P2 files of a paired-end mapping run.	Genomic Resequencing MethylMiner Targeted Resequencing Whole Transcriptome

Analysis Module	Description	Used in Analysis Type
Mapping: mate-pair	Mate-pair mapping is similar to fragment mapping. Match files (which contain alignment information) are provided to pairing stage.	Genomic Resequencing
Mapping: fragment smRNA	This module maps fragmented small RNA (also known as microRNA or miRNA) reads.	Small RNA
Mature small RNA	This module generates tag counts for mature miRNA sequences.	
Precursor small RNA	This module generates tag counts for precursor miRNA sequences.	
SAET	This module increases mappability of reads to a reference genome.	(Optional) Genomic Resequencing Targeted Resequencing
Small Indels	This module detects gap alignments.	Genomic Resequencing Targeted Resequencing
SNP Finding	This module uses the diBayes algorithm to find Single Nucleotide Polymorphisms (SNPs).	Genomic Resequencing Small RNA Targeted Resequencing

3. Click **Next** to choose references.

The Create Analysis · Choose References view appears.


Choose references

Note: The procedure for choosing references is the same for XSQ and BAM data types.

1. Click the Select Reference button to browse for reference.
2. In the pop-up window, select a folder in the Genomic Ref. Folders pane, select a reference file, then click **OK**.
The referenced file appears in the Reference column.
3. To exit Choose References without saving your references, click **Cancel**. To save your changes and exit Choose References, click **Finish**. Otherwise, click **Edit** and follow the procedure “[Edit an analysis](#)” on page 82.

Reuse an analysis


There are two methods you can reuse an analysis to create an analysis:

- **Method 1:** Select a completed analysis , click **Analysis** in the top menu, and then click **Reuse**, or
- **Method 2:** Select a project, then click **Create Analysis** in the Task Wizards section (shown [page 59](#)).

Method 1

1. In the Create Analysis · Choose Analysis view, enter a new analysis name, then click **Next** to choose references, described [on page 81](#).
2. Click **Edit**, then follow the procedure “[Edit an analysis](#)” on page 82.

Method 2

1. In the Create Analysis · Choose Data Type view, follow the procedure “[Choose data type](#)” on page 79.
2. Follow the procedure “[Choose data](#)” on page 79.
3. In the Create Analysis · Choose Analysis view, click **Reuse Old Analysis**, then click the  button to select an old analysis.
4. In the Select Old Analysis window, select an analysis, then click **OK**.
5. In the Choose Analysis view, click **Next** and follow the procedure “[Choose references](#)” on page 81.
6. Click **Edit**, then follow the procedure “[Edit an analysis](#)” on page 82.

Reuse BAM data

You cannot reuse an analysis to run only BAM data.

To run an analysis with only BAM data:

1. Create an analysis, as described [on page 78](#).
2. Choose BAM data, as described [on page 79](#).
3. Edit the analysis, as described in the following section.

Edit an analysis

Note: To edit Advanced, SAET, and BamStats parameters, refer to the *Graphical User Interface* (4465697).

Editing an analysis involves choosing modules, setting general parameters, and setting module parameters.

Note: To see a brief description of the settings, place your cursor over a  button.

Choose modules

In the Choose Modules view, you can specify pre-processing of raw reads, the method for mapping data during secondary analysis, and the modules for tertiary analysis.

Secondary Analysis

During secondary analysis, information from reads in XSQ-formatted input files and a FASTA-format reference file are combined to generate alignment information in an intermediate file format known as a Binary Alignment sequence Map (BAM) file.

The following table shows the type of analysis that is done during genomic resequencing and whole transcriptome analyses.

Analysis	Mapping
Genomic resequencing	Fragment mapping
Whole transcriptome	<ul style="list-style-type: none"> Whole Transcriptome Exon Sequencing Extractor Whole Transcriptome Splice Junction Extractor Whole Transcriptome Splice Junction Extractor

Map data

Accept the default setting or click **Customize** to not include the BAMStats mapping module.

If raw reads need to be pre-processed with the SOLiD Accuracy Enhancer Tool (SAET), click the **Pre-Process** checkbox.

Tertiary Analysis

Note: The procedure for choosing modules is the same for XSQ and BAM data types.

In the Tertiary Analysis Modules section, select tertiary analysis modules for targeted resequencing:

- Click a module name, then click the > button to insert it into the Include list. To include all available modules, click **All>>**.
- To generate annotated output for a module, click its Annotate checkbox.

Modules	Description	Workflow
CNV	Detects copy number variations in a data sample that is mapped to the reference sequence.	Perform Human Copy Number Variation Analysis
SNP Finding	Takes the color-space reads, the quality values, the reference sequence, and error information on each SOLiD™ System slide as its input, and calls Single Nucleotide Polymorphisms (SNPs). For more information, refer to Chapter 19, “Perform SNP Finding Analysis” on page 249	Perform SNP Finding Analysis
Enrichment	Enrichment statistics provide a means to assess enrichment platform performance by looking at variations in coverage, both across all targets and on a per-target basis.	Perform Targeted Resequencing and Enrichment Analyses
Inversion	Calls inversions based on library size.	Perform Inversion Analysis

Large indel	Identifies deviations in clone insert size.	“Perform Large Indel Analysis”
Small Indel	Takes in a BAM file or a set of BAM files from Fragment, Paired-End, and Long Mate Pair libraries using the LifeScope™ Software alignment produced by LifeScope Software.	Perform Targeted Resequencing and Enrichment Analyses
Small RNA	Used to analyze high throughput small RNA sequencing raw data. The small RNA mapping module maps small RNA reads using the mapping program mapReads. The small RNA whole transcriptome coverage module calculates read coverage per position. The small RNA counts module generates tag counts for precursor and mature miRNA sequences.	“Perform Small RNA Analysis”
SNP Finding	Used to call Single Nucleotide Polymorphisms (SNPs) from mapped and processed SOLiD™ System color-space reads.	“Perform SNP Finding Analysis”

After you have chosen modules and customized mapping, click **Next** to set the general parameters.

Set general parameters

Note: To revert parameters to their default settings, click **Reset to Defaults**.

Set the parameters for files locations by clicking on the ... button and navigating to the locations.

After you set general parameters, you can optionally set module parameters. To skip setting module parameters and proceed to reviewing your analysis, click **Review**. Otherwise, click **Next** to set module parameters.

Set module parameters

Review the parameters for each module. If necessary, change the parameters.

To reject your edits, click **Cancel**. To save your edits and close the Edit Analysis window, click **Finish**. Otherwise, click **Review** to proceed.

Review your analysis

Review the parameters and data in your analysis.

Review parameters Review the parameters for modules before starting an analysis run. You can show and hide the parameters for each module by clicking on the ► next to each module.

To change a parameter, click **Edit**.


Review data In the Data tab, review the read-only data in the Targeted Resequencing Analyses to be Run table.

Run and monitor your analysis

You can immediately run your analysis or run it later.

Start your analysis now After you review your analysis and if you are satisfied with the parameters and data, click **Start Analysis**. The Run Analysis window closes.

Start your analysis later

1. Click your project in the Projects organizer (shown on page 59) to show the analysis module, select the analysis , then either:
 - Click **Run Analysis** in the Task Wizards section, or
 - Click **Analysis** in the top menu, then **Run**.

In the Run Analysis window, you can review the parameters for analysis modules and data that will be analyzed. When you are ready, click **Start Analysis**. The Run Analysis window closes.

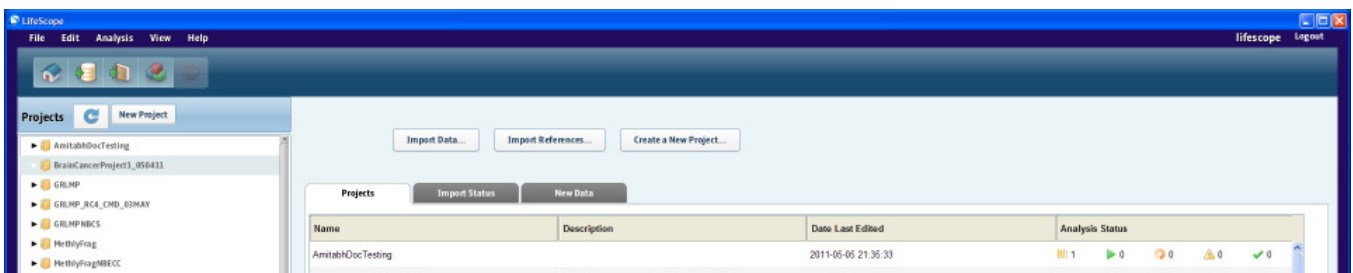
After the analysis starts, the home view (shown on page 59) appears. The progress of the secondary and tertiary analyses is shown in the Status tab .

Stop an analysis To stop the run of an analysis, click **Analysis** in the toolbar, then click **Stop**.

View analysis status

There are two ways to check the status of an analysis run:

- In the status overview Projects tab (shown on page 59). For a description of status icons, see the table on page 61.



- In the Status tab for a specific analysis.
1. In the Projects list, click the project to check the status of the analysis run.
 2. Click the **Status** tab in the status overview.
The Progress columns show the percentage of completion for the secondary and tertiary analyses.

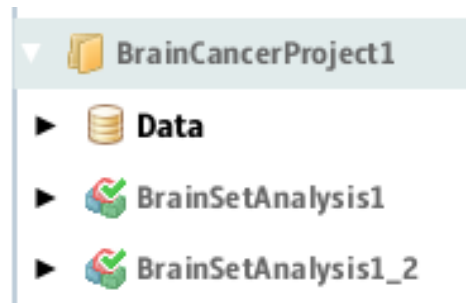
ProstateCancerProje... • ProstateCancerAnaly...

Overview Status Parameters

Analysis Runs

Name	XSQ ID	Analysis	Secondary Analysis	Secondary Progress	Tertiary Analysis	Tertiary Progress
solid0054_20110102_PE_LFD...	solid0054_20110102_PE_LFD...	ProstateCancerAnalysis1	[SAET, Mapping, BamStats]	<div style="width: 33%;"></div> 33%	[Enrichment, diBayes, SmallIn...	0%

When analysis is complete, a green check mark appears on the analysis name in the Projects organizer (shown [on page 59](#)).



Delete an analysis

To delete an analysis, select the analysis in the Projects list, click **Analysis**, then **Delete**.

8

View Analysis Results

This chapter covers:

- View analysis results 87
- View results now 87
- View results later 87
- View Results window 87
- Failed analysis. 91

View analysis results

You can view analysis results immediately after mapping has been completed or view results at a later time. The duration for an analysis run depends on the complexity of the analysis.



Note: If multiple View Results windows are open, actions such as file downloads and file exports will be slowed. To improve performance, close windows that you do not need open.

View results now

If an analysis run is brief, click **View Results** in the Secondary Progress and Tertiary Progress columns in the Status tab.

View results later

To view analysis results at a later time:

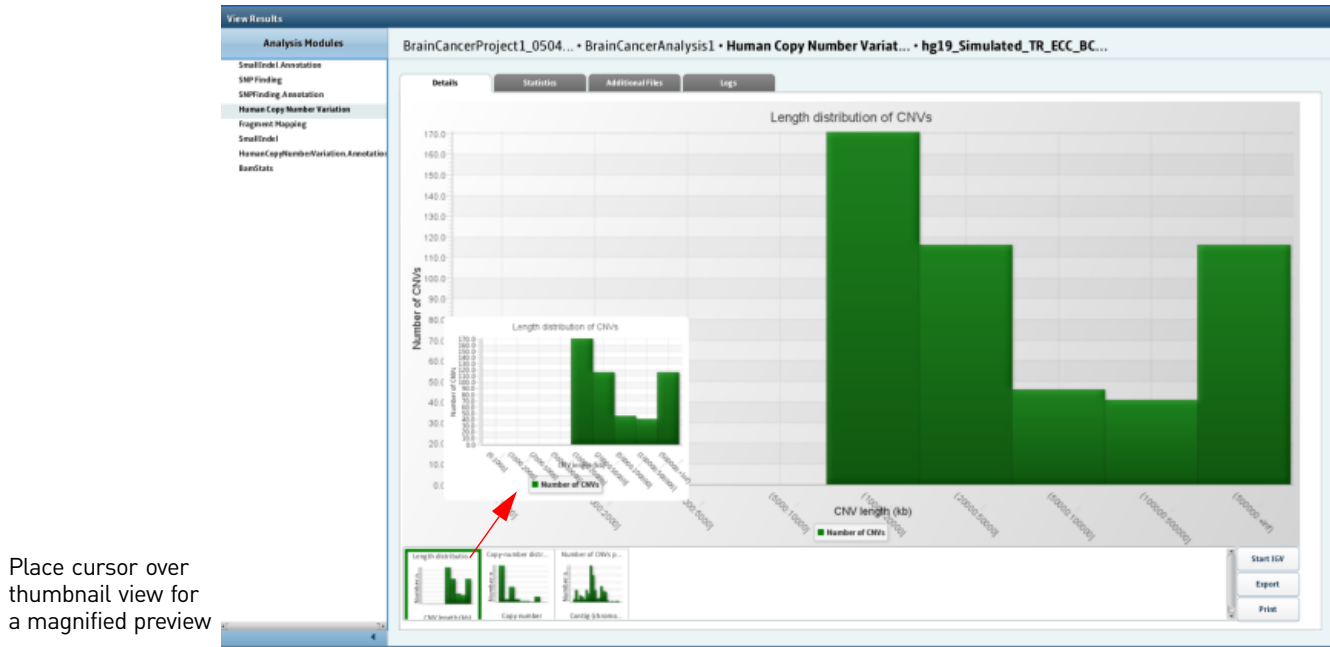
1. Select your project in the Projects organizer (shown [on page 59](#)).
2. Select the completed analysis , then either:
 - Click the Results icon  in the Projects organizer, or
 - Click the **Status** tab in the status overview (shown [on page 59](#)).
3. Click the **View Results** button in a Progress column to open the View Results window (shown [on page 88](#)).

BrainCancerProject1... • BrainCancerAnalysis1 • Results						
Data-Sets with Results in Analysis						
Name	XSQ ID	Analysis	Secondary Analysis	Secondary Progress	Tertiary Analysis	Tertiary Progress
hg19_Simulated_TR_ECC_BC1...	hg19_Simulated_TR_ECC_BC16...	BrainCancerAnalysis1	[Fragment Mapping, BamStats]	View Results	[SNP Finding, Human Copy Nu...]	View Results

View Results window

The left pane of the View Results window shows successful analysis modules. Annotated analysis modules that fail will not appear in the left pane of the window.

The right pane of the View Results window shows the analysis results, which include details, statistics, additional files, and logs.



Details

The Details tab consists of a canvas that shows analysis results in the form of bar charts and pie charts, as shown above. Thumbnail views of charts are below the canvas. Buttons include Start IGV, Export, and Print.

Placing your cursor over a thumbnail view opens a magnified preview. Clicking on a thumbnail view of a chart displays it on the canvas.

You can change the colors of bars or wedges in a chart by right-clicking on those areas and selecting a color in the Change Color pop-up window.

Start IGV

You can view BAM files and GFF3 files, described , with the Integrative Genomics Viewer (IGV) browser. In addition to a BAM file, you need a reference genome. For more information about the IGV, refer to [“Integrative Genomics View \(IGV\)” on page 302](#). For instructions on using the IGV, go to www.broadinstitute.org/igv/.

Export

You can export charts as a comma-separated values (.csv), .jpeg, .gif, or .png file.

To export details:

1. Click **Export**.
2. In the Export Results window, choose to export only the displayed, magnified view of data or all data in the Details tab.
3. Select an export format.
4. Click **Browser** to navigate to a destination for the exported data.

5. Click **OK** to export the data.

Print

Click this button to print analysis details.

Statistics

The Statistics tab shows a LifeScope report of statistics that you can view, export, and print. You cannot edit the statistics.

Additional Files

The Additional Files tab shows a list of result files that you can download and view with the IGV browser.

Note: A large file can take several minutes to download.

Download additional files

To download a file:

1. Select the file in the Additional Files table, then click **Download**.
2. In the Confirm window, click **OK**.

After a file has been downloaded, the message “Download Done” appears.

To view a file, click **Add to IGV**.

Logs

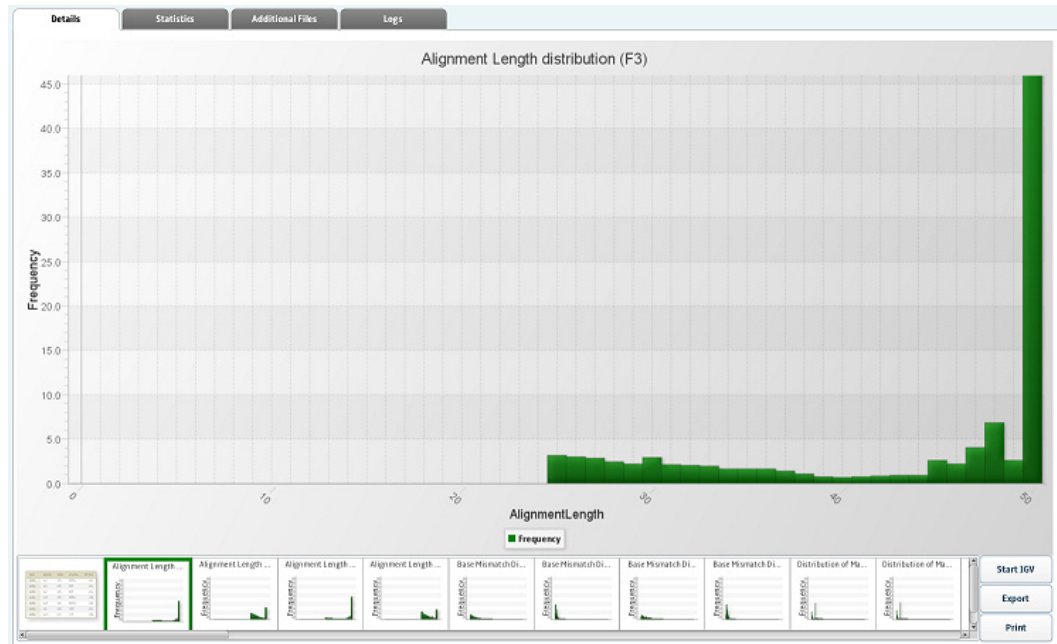
The Logs tab shows a list of log files that you can download. To download a log file, select the file in the Log Files table, then click **Download**. In the

Note: A large file can take several minutes to download. After a file has been downloaded, the message Download Done appears.

Examples of view results

This section shows examples of results for BAMStats.

Details



Statistics

There are not statistics for BAMStats.

Additional files

Additional files for small indel analysis output include comma-separated values (.csv), text (.txt), and wiggle (.wig) files.

The following table is an example of additional files for BAMStats results.

Directory	Files	File Type
bamstats/WTPE	WTPE_statstuple.txt	.txt
bamstats/WTPE/ K_201001207_PE_BC_MAGNUM1_WT_FC1_BC.s owmi-1-Idx_BC1-1/Misc	K_201001207_PE_BC_MAGNUM1_WT_FC1_BC.sowmi -1-Idx_BC1-1.bam_chr19.POS.wig	.wig
	K_201001207_PE_BC_MAGNUM1_WT_FC1_BC.sowmi -1-Idx_BC1-1.Mismatches.by.ReadPair.Type.csv	.ungap...
	BrainCancerAnalysis1.gff3	.csv

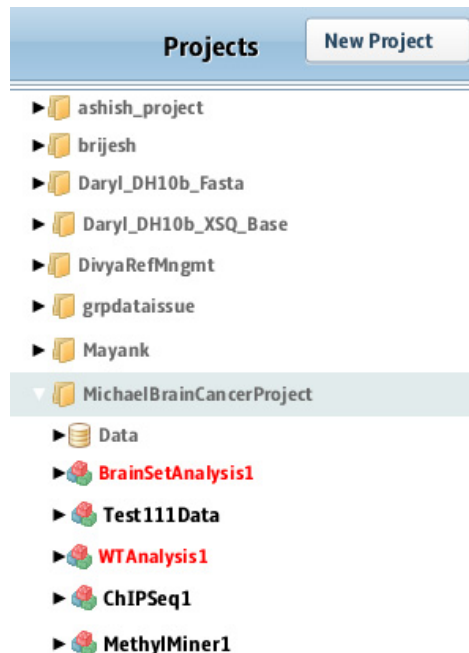
Logs


The following table is an example of logs for BAMStats results.

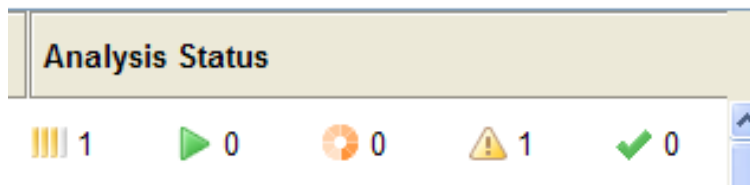
Directory	Files	File Type
bamstats	secondary-hg19-bamstats.20110429133358916.log	.log
	secondary-hg19-bamstats.main.20110429133401974.log	

Failed analysis

In the Projects organizer, a failed analysis is indicated by red text, as shown:



In the status overview for all projects, a failed analysis is indicated by the fault icon  in the Analysis Status column, as shown:



In the Status overview of a specific project, a failed analysis is indicated in red, as shown:

Analysis Runs			
Name	XSQ ID	Analysis	Progress
test_S1_F31298403116375_DefaultLibrary	test_S1_F31298403116375.xsq	a2	Fail !!

View failure details To find out the reason an analysis failed:

1. Click **Fail** in the Progress column to view the results of the analysis run.
2. In the View Results window, click a row in the Log Files table, then click **Download**.
3. In the Select Directory browser, navigate to a destination for the log file, then click **OK**.
4. Navigate to the log file and open it to see the details of the run.

PART III
Workflows

9

Perform Targeted Resequencing and Enrichment Analyses

This chapter covers:

- Introduction to Targeted Resequencing analysis..... 95
- Targeted resequencing library types 96
- Targeted resequencing input files 97
- Enrichment statistics input files 97
- Targeted resequencing analysis modules 99
- Targeted resequencing parameters 100
- Perform Targeted Resequencing analysis..... 112
- View analysis results 116
- Targeted resequencing output files 116

Introduction to Targeted Resequencing analysis

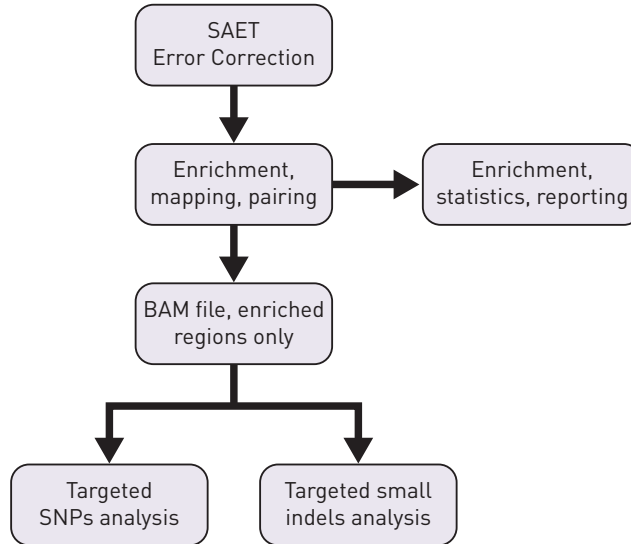
Targeted resequencing is a set of analyses designed for target-enriched sequencing data. Given a set of coordinates representing a region of interest in a genome, a library may be enriched for reads within those regions prior to sequencing.

The SAET module is enabled by default for enriched runs, as the smaller target size makes spectral error correction viable, even if it may not be practical for unenriched sequencing runs on a large reference sequence.

Targeted resequencing mapping produces a filtered BAM file for variant calling. The variant calling modules SNP Finding and Small Indel are run on the filtered data sets instead of the entire read population, and limit SNP and indel calls to the target region.

The input list of regions of interest (ROI) is provided by the user. Supported ROI input files include third-party files of baits, probes, or tiles. The input ROI files can be in BED format or in a text file as a list of “chr:start-end” values. The Targeted Resequencing mapping module sorts the input ROI file by chromosome and start position, and makes the resulting file available as output.

The following illustration shows the components of Targeted Resequencing analyses.



Targeted resequencing library types

The Targeted Resequencing analyses support data from these library types:

- Fragment
- Paired-End
- Mate-Pair
Mate-pair libraries are accepted, but pairing information is not considered for mate-pair libraries.
- Libraries generated with up to 96 barcodes
Individual sequencing and enrichment reports are provided for each barcoded library.

Targeted Resequencing output files are compatible with third-party genome browsers, including the IGV browser.

The Targeted Resequencing Map analysis type provides analysis of sequencing runs that have been run through targeted enrichment.

Enrichment statistics

Targeted resequencing support optional enrichment statistics, which provide a means to assess enrichment platform performance by:

- Looking at variations in coverage, both across all targets and on a per-target basis.
- Addressing the uniformity and completeness of coverage within the target region.
- Calculating the degree of enrichment.

Enrichment statistics scenarios

The enrichment statistics module supports these scenarios:

- Genome class:
 - Human
 - Non-human multi-chromosome
 - Single chromosome
- Library type:
 - 2X50
 - 2X25
 - fragment
- Experiment type:
 - Single sample, single run
 - Multiple runs, multiple samples

Targeted resequencing input files

Input File	File Extension	Analysis Module
Binary Alignment Map (BAM)	*.bam	Small Indel, SNP Finding, CNV
(Optional) Color space FASTA	*.csfasta	Mapping
eXtensible SeQuence (XSQ)	*.xsq	Small Indel, SNP Finding, CNV
FAST-All (FASTA)	*.fa, *.fasta	Mapping
(Optional) Quality Value	*qual, *qv	Mapping, pairing

Enrichment statistics input files

This section describes the input files required for by the enrichment statistics module.

The following table lists the two input files for the enrichment statistics module.

File	File format	Associated parameter	Description
target regions	BED, GFF v3, or text	enrichment.target.file	A list of target regions, either as regions of interest (ROI) or as target lists provided by vendors of enrichment platforms (such as baits, probes, or tiles files).
aligned reads	BAM	enrichment.input.file	Alignments generated by the sequencing run after being matched to the entire reference genome.

The input target regions file

The target regions file may be either in BED format or a list of chr:start-end elements. An example of partial content for a target region file is shown below.

```
chr13 31787000 31788297
chr13 31788404 31788957
chr13 31789616 31789695
chr13 31790390 31790601
chr13 31790852 31791495
chr13 31791644 31791981
```

The following information applies to the input target region file:

- Target region coordinates must be specified using the same reference used in mapping.
- Target regions should be sorted by chromosome/position. If not sorted, they will be sorted as the first step in the enrichment statistics module.
- After the first 3 required fields per line, any additional fields are ignored.
- The file can optionally be a list of target regions in the “chr:start-end” format.

An example of this format is: chr1:1234-4321

The first field, “chrom”, must have identifiers that match the sequence names used in the SAM file. For example, for the hg18 reference:

- Use “chr1”, not “1”.
- Use “chrX”, not “23” or “chr23”.

Note: A standard BED file uses zero-based indexing and a half-open interval (meaning the specified range is up to, but not including, the last position).

The input aligned reads file

This file is a BAM-formatted file of aligned reads generated by the sequencing run after being matched to the entire reference genome.

The following information applies to the input aligned reads file:

- The BAM file must have a header listing sequence identifiers.
The header is mandatory in the SAM specification for the binary BAM files.
- These sequence identifiers must be consistent with those used in the target list.
- The BAM file must be sorted by coordinate. The SO annotation in the header is not examined. All alignments within a chromosome must be contiguous. For sequential records, start coordinates within a chromosome must be equal or increasing.
- The BAM file may be from a fragment, paired-end, or mate-pair library.
All reads are treated independently; no special handling is performed based on library type.
- Reads from paired libraries will tend to have improved mapping scores based on unique paired hits, and will be more likely to pass uniqueness filtering than unpaired reads.

Targeted resequencing analysis modules

Targeted resequencing provides a mechanism to restrict resequencing analysis to specific regions of interest within the genome being studied. Targeted resequencing analysis includes the following modules:

- Mapping
- Enrichment
- SNP Finding
- Small indels

Mapping

The map data resequencing workflow uses eXtensible SeQuence (.xsq) files to create a .bam file.

The Targeted Resequencing mapping module uses eXtensible SeQuence (.xsq) files to create a *.bam file. Targeted Resequencing mapping also requires a regions of interest (ROI) file as input. The ROI file can be in BED, text, or tiles format.

Enrichment

Enrichment statistics provide a means to assess enrichment platform performance by looking at variations in coverage, both across all targets and on a per-target basis.

SNP Finding

The SNP Finding package is the module used to call Single Nucleotide Polymorphisms (SNPs) from mapped and processed SOLiD™ System reads. The module takes the reads, the quality values, the reference sequence, and error information on each SOLiD™ System slide as its input, and calls SNPs.

The module creates three results files:

- A list of SNPs.
- (Optional) A consensus *.fasta file with the same number of bases as the reference sequence.
- (Optional) A list of all covered positions.
- (Optional) A collection of annotated files.

For information about using the module to run a Targeted Resequencing experiment, see [Chapter 9, “Perform Targeted Resequencing and Enrichment Analyses” on page 95](#).

Small Indel

When an indel occurs in a sequence, and that sequence is measured using color-space, the color-space sequence has a gap the same size as the indel. The color-space sequence also leaves a signature that can indicate whether there is a measurement error within the gap. The Small Indel tool targets the processing of indel evidences found in the pairing step during secondary offline data analysis.

For information about using the module to run a Targeted Resequencing experiment, see [Chapter 9, “Perform Targeted Resequencing and Enrichment Analyses” on page 95](#).

Targeted resequencing parameters

Parameters for Targeted Resequencing analysis include:

- General parameters
- **Secondary analysis:** Fragment Mapping (including optional BAMStats and SOLiD Accuracy Enhancement Tool [SAET])
- **Tertiary analysis:** Small Indel, Enrichment, and SNP Finding.

General

General parameters include:

Parameter	Default value	Description
analysis.assembly.name	—	The name of the genome assembly used in current analysis. Examples: hg18, hg19.
annotation.dbsnp.file	—	The path to the file used to annotate SNPs and small InDels. LifeTech-provided files are dbSNP_b130.tab(hg18) and 00-All.vcf(hg19).
analysis.regions.file	—	The path to the file containing genomic regions used in targeted resequencing selection, such as a .bed format file containing exome targets.
annotation.gtf.file	—	The file containing gene and exon annotations corresponding to the genome assembly used in the analysis.

Use the  button to open the File Chooser and search for input files.

Fragment Mapping

There are four categories of fragment mapping parameters: Main, Advanced, SAET, and BAMStats.

Main

Parameter	Default value	Description
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	0	Replace resulting base-calls with a quality value less than the specified value.
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: <ul style="list-style-type: none"> True: Create. False: Do not create.
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–1,000.
Primary output filter type	Primary_only	
Reference weight	15	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.
Second map gapped algorithm type	GLOBAL	Do indel finding, and control the behavior of indel finding. Allowed values: <ul style="list-style-type: none"> • NONE: Turn off indel finding. • GLOBAL: Mapping reports global alignment up to one indel. • LOCAL: Mapping reports local alignment up to one indel.

SAET

The SAET tab is accessible only if you select SAET pre-processing when you choose modules.

Parameter	Default value	Description
Genome length	1000000	Expected length of sequenced (or enriched) DNA region. For example, 4,600,000 for the E.Coli 4.6 MB genome or 30,000,000 for the entire Human Transcriptome. Allowed values: Integers ≥ 200 .
On target ratio	0.5	The expected ratio of reads that come from the targeted region. Allowed values: Floats 0.0–1.0.
Update quality values	True	Update quality value of modified calls. Allowed values: True: Update the QV of modified calls. False: Do not update the qv for modified calls.
Trusted quality value	25	Correction is applied to calls with a quality value below the value of this parameter. Allowed values: Integers ≥ 1 .
Support votes	2	The minimum number of k-mer votes required make a correction. Allowed values: Integers ≥ 1 .
Trusted frequency	0	The lowest multiplicity of the seed to be included in the spectrum. (If set to 0, then the value is computed internally.) Allowed values: Integers ≥ 0 .
Maximum corrections per read	0	Maximum number of allowed corrections per read. (If set to 0, then the value is set to $\lceil \text{readLength}/8 \rceil$). Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: Integers 0–9.
Number of recursive runs	1	The error correction step is repeated the provided number of times. Reduce if over-corrections are observed, or increase if undercorrections are observed. Allowed values: 1, 2, or 3.
Position of error inflation point	0	Position in the read at which the error rate inflates, for instance, 35–40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers.

Parameter	Default value	Description
Disable random sampling for large data	False	Disables random sampling in spectrum building. If set to 0, then for large datasets (coverage > 300x), a subset of reads is used in spectrum building. Allowed values: <ul style="list-style-type: none"> • True: Disables random sampling in spectrum building. • False: Do not disable random sampling in spectrum building. true?
K-mer size	0	Size of k-mer (>5) used in spectrum construction and error correction. (If set to 0, then the value is computed internally.) Allowed values: Integers 0–28.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	`\${analysis.output.dir}/ fragment.mapping`	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	`\${task.output.dir}`	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10,000.
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 0–100,000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers 1–100,000
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0–1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.

Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers 1-100,000.
--	-----	--

Small indel

Categories for the Small Indel analysis module include Advanced and (if you selected Annotation for Small Indel output) Annotation. There are no Main parameters.

Advanced

There are five categories of Advanced parameters: General Options, Pileup, Mapping Quality Filtering, Heuristic Filtering, and Indel Size Filtering.

General Options

Parameters	Default value	Description
Detail level	0	For BAM file inputs, the level of detail in output: <ul style="list-style-type: none"> 0: Keeps only position information about the anchor read and no information for the ungapped alignment. 1-8: Keeps only some of the alignment's anchor alignment but none of the ungapped alignment. 9: Is most detailed, but also the slowest.
Zygosity profile name	max-mapping	Zygosity profile name. <ul style="list-style-type: none"> classic: Profile for classically (full read) mapped reads. max-mapping: Profile for seed and extend ungapped alignments. max-mapping-v2: Reserved for future use. gap-align-only, and no-calls: Force all zygosity calls to be homozygous calls.
Genomic region	-	Names a specific genomic region to be selected from the BAM file. Only full chromosomes are guaranteed not to alter results. Specifying partial chromosomes is allowed but may result in the loss of indels near the edges of that region.
Display base QVs	False	Display base QV scores in the GFF file. Allowed values: <p>True: Display the FASTQ base QV scores for all of the reads used for each indel in the GFF file. FASTQ strings contain semi-colons, so adding these strings may produce a GFF file that is not compatible with certain applications.</p> <p>False: Do not display QV scores in the output file.</p>
Number of alignments per pileup	1000	For pileups with more than this number of reads, set the expected number of alignments per pileup. Values much higher than 1000 may result in a significant increase in computational time. Allowed values: Integers ≥ 0 .

Parameters	Default value	Description
Random seed	94404	The random seed value used to determine which pseudo random set of reads to use when there are greater than 1000 reads in a pileup. The random number generator used is the Mersenne Twister MT19937 algorithm. Allowed values: Integers ≥ 0 .

Pileup

Parameters	Default value	Description
Min num evid	2	Minimum number of evidences required for an indel call. This parameter does not have an upper limit, but a value higher than the average coverage level in most cases causes a significant reduction in sensitivity.
Max num evid	-1	Maximum number of evidences. Use -1 for no maximum. Setting this value to some multiple of the average coverage could remove indels found in abnormally high coverage areas.
ConsGroup	1	Indel grouping method. Allowed values: 1: Conservative grouping of indels with 5bp max between consecutive evidences. 2: Lax grouping. Groups indels that are at maximum the higher of 15 or 7 times the indel size. 9: No grouping. Makes every indel evidence a separate pileup.

Mapping Quality Filtering

Parameters	Default value	Description
Max reported alignments	-1	Only uses those alignments where the NH field (the number of reported alignments; from the BAM record) is this value or lower. A value of -1 is to have no upper limit. The range where this is effective depends on the input BAM file's range of values of the NH tag.
Min mapping quality	8	Keeps only reads that have this or higher pairing qualities. For paired tags, mapping quality is for the pair (pairing quality), and for fragment, it is the single tag's map quality. Reads that are lower than this value are filtered out. Allowed values: Integers 0-100.

Parameters	Default value	Description
Min best mapping quality	10	For a particular indel called with a set of reads, at least one pairing quality in this set must be higher than this value. Allows for supporting evidences to have a lower mapping quality threshold than the best read.
Min anchor mapping quality	-1	Minimum mapping quality for a non-indel (anchor) tag. Effective only for paired reads, for the number of anchors queried as defined by <code>small.indel.detail.level</code> .
Ungapped BAM flag filter	ProperPair	For ungapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Gapped BAM flag filter	Primary	For gapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Edge length deletions	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0 .
Edge length insertions	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0

Heuristic Filtering

Parameters	Default value	Description
Perform filtering	True	Whether or not to perform filtering in each pileup. Allowed values: <ul style="list-style-type: none"> • True: Perform filtering on pileups. • False: Do not perform filtering on pileups. Parameters that change the makeup of pileups, such as Min num evid are still active.
Indel size distribution allowed	can-cluster	Indel sizes in a pileup are allowed to have certain indel size distributions. Allowed values: <ul style="list-style-type: none"> • similar-size: 75% of the reads of a pileup must have exactly the same size. • similar-size-any-large-deletions: Any pileups with at least 2 large deletion alignments, the other pileups must have similar sizes. • can-cluster: Allowed if at least one cluster of any indel size is found. • can-cluster-any-large-deletions: Any pileups with at least 2 large deletion alignments; other pileups must be able to cluster (will have indels with two more reads with larger deletions, even if they don't form good clusters). • any: Can have any size distribution.
Remove singletons	True	Remove the singletons that occur when different alignment methods are combined based on identical bead ids and read sequence. Allowed values: <ul style="list-style-type: none"> • True: Remove singletons. • False: Do not remove the singletons.
Alignment compatibility filter	1	Alignment compatibility level. Checks color space compatibility around the gap. Allowed values: <ul style="list-style-type: none"> • 0: No alignment compatibility filtering. • 1: The small indel module determines whether the data contains base-space or color-space sequence. • 2: Force the use of base-space sequence, if present. • 3: Force the use of color-space sequence, if present.
Max coverage ratio	12	Maximum allowed value for the ratio of number of reference alignments allowed by the <code>ungapped.bam.flag.filter</code> over the number of non-redundant indel variant reads selected with the <code>gapped.bam.flag.filter</code> . Use -1 for no limit (no coverage ratio filtering). Allowed values: Integers.

Parameters	Default value	Description
Max nonreds 4Fit	2	Maximum number of non-redundant reads where read position filtering is applied. Allowed values: Integers.
Min from end pos	9.1	Minimum average number of base pairs from the end of the read required of the pileup, when there are at most a certain number of reads defined by Max nonreds 4Fit . Allowed values: Floats.

Indel Size Filtering

Parameters	Default value	Description
Min insertions size	0	Minimum insertion size to include. Allowed values: Integers.
Min deletions size	0	Minimum deletion size to include. Allowed values: Integers.
Max insertions size	1000000000	Maximum insertion size to include. Allowed values: Integers.
Max deletions size	1000000000	Maximum deletion size to include. Allowed values: Integers.

(Optional) Annotation

You can optionally annotate the mapped output of Small Indel analysis. For descriptions of the Annotation parameters, see “Annotation” [on page 112](#).

Enrichment

There are two categories of enrichment parameters: Main and Advanced.

Main

Parameter name	Default value	Description
Report directory	<code>\${task.output.dir}</code>	Directory where all reports will be created.
Output directory	<code>\${task.output.dir}</code>	Directory where output BAMs will be created.
Target file	<code>\${analysis.regions.file}</code>	File specifying target regions or regions of interest to be analyzed. It may be BED, GFF (v.3), or a list of chr:start-end elements. Multiple target files are allowed.

Advanced

There are two categories of Advanced parameters: general and Reports.

Parameter name	Default value	Description
Extend bases	0	The number of bases on either side of the target region in which alignments may be captured. The target sequence is extended by this number of bp on either side. Used by both target capture and target coverage statistics. Allowed values: Integers ≥ 0 .
Minimum mapping score	8	Minimum mapping quality value (MAPQ) allowed for aligned reads. Reads below this threshold are not used. Allowed values: Integers $\geq 0-80$.
Minimum target overlap	0.0001	The fraction of an alignment that must be overlapped by a target in order to be classified as on target. "0.50" is interpreted as 50%, and implies that the midpoint of the read must fall within target region. Allowed values: $0.0001 < X \leq 1.0$
Minimum target overlap reverse	0.0001	
Reports		
Summary report	True	Create or do not create the summary statistics file. Allowed values: <ul style="list-style-type: none"> • True: Create the summary statistics file. • False: Do not create the summary statistics file.
Target coverage stats		Output or do not output per-target coverage statistics (min, max, mean) in tabular format. Allowed values: <ul style="list-style-type: none"> • True: Create per-target coverage statistics. • False: Do not create per-target coverage statistics.

Parameter name	Default value	Description
Coverage frequency	False	Create or do not create the per-target coverage frequency histogram. Allowed values: <ul style="list-style-type: none"> • True: Create the per-target coverage frequency histogram. • False: Do not create the per-target coverage frequency histogram.
Coverage bedgraph		Create a BEDGRAPH format coverage file for on-target reads. Allowed values: true, false, 1, 0 <ul style="list-style-type: none"> • True: Create a BEDGRAPH format coverage file for on-target reads. • False: Do not create a this file.
Genome coverage frequency		Create a per-chromosome coverage frequency histogram. Allowed values: <ul style="list-style-type: none"> • True: Output a per-chromosome coverage frequency histogram. • False: Do not create this histogram.

SNP Finding

There are two categories of SNP Finding parameters: Main and Advanced.

Main

Parameter name	Default value	Description
Call stringency	medium	Call stringency.
Skip high coverage positions (Het)	1	Skip high coverage positions (Het)
Minimum mapping quality value	8	Minimum mapping quality value.

Advanced

There are six categories of Advanced parameters: general, Read filter, General position filter, Heterozygous position filter, Homozygous position filter, and Output file processing.

Parameter name	Default value	Description
Detect adjacent SNPs	0	Detect adjacent SNPs.
Polymorphism rate	0.001	Polymorphism rate.
Read filter		
Include reads with unmapped mate	0	Include reads with unmapped mate.
Exclude reads with indels	True	Exclude reads with indels.
Require only uniquely mapped reads	0	Require only uniquely mapped reads

Parameter name	Default value	Description
Ignore reads with a higher mismatch count to alignment length ratio	1.0	Ignore reads with a higher mismatch count to alignment length ratio.
Ignore reads with a lower alignment length to read length ratio	1.0	Ignore reads with a lower alignment length to read length ratio.
General position filter		
Require alleles to be present in both strands	False	Require alleles to be present in both strands
Minimum base quality value for a position	14	Minimum base quality value for a position.
Minimum base quality value of the non-reference allele of a position	14	Minimum base quality value of the non-reference allele of a position.
Heterozygous position filter		
Minimum allele ratio (Het)	0.15	Minimum allele ratio (Het).
Minimum coverage (Het)	2	Minimum coverage (Het).
Minimum unique start position (Het)	2	Minimum unique start position (Het).
Minimum non-reference color QV (Het)	7	Minimum non-reference color QV (Het).
Minimum non-reference base QV (Het)	14	Minimum non-reference base QV (Het).
Minimum ratio of valid reads (Het)	0.65	Minimum ratio of valid reads (Het).
Minimum valid tricolor counts (Het)	2	Minimum valid tricolor counts (Het).
Homozygous position filter		
Minimum coverage (Hom)	1	Minimum coverage (Hom).
Minimum count of the non-reference allele (Hom)	2	Minimum count of the non-reference allele (Hom).
Minimum average non-reference base QV (hom)	14	Minimum average non-reference base QV (Hom).
Minimum average non-reference color QV (hom)	7	Minimum average non-reference color QV (Hom)
Minimum unique start position of the non-reference allele (Hom)	2	Minimum unique start position of the non-reference allele (Hom).
Output file processing		
Output fasta file	True	Output or do not output the FASTA file. <ul style="list-style-type: none"> • True: Output the FASTA file. • False: Do not output the FASTA file.
Output consensus file	True	Output or do not output the consensus file. <ul style="list-style-type: none"> • True: Output the consensus file. • False: Do not output the consensus file.

Parameter name	Default value	Description
Compress the consensus file	False	Compress or do not compress the consensus file. <ul style="list-style-type: none"> • True: Compress the consensus file. • False: Do not compress the consensus file.

(Optional) Annotation

You can optionally annotate the mapped output of Targeted Resequencing analysis. For descriptions of the Annotation parameters, see the following section [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform Targeted Resequencing analysis

(Optional) Import data

You can optionally import data to be analyzed. For instructions on how to import data, see “Import Data” on page 67.

Log in to LifeScope™ Software

1. Navigate to LifeScope™ Software at `http://<IP address>:<port number>/LifeScope.html` where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown on page 59).

Create a project

1. In the home view, click **Create a New Project**.
2. If you create a project:
 - a. Type a name and description in the Enter Project Name view.

Note: The name cannot have spaces or special characters.
 - b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown on page 59), click **Add Data to Project** to choose a data type.

Add data to the project

Note: Add data from the read repository to a project by choosing a data type and finding to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the reads you want to map and analyze. If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).
3. (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.

To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.

To remove a data file from a group, click the checkbox of the data file, then click **Delete**.

4. Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

1. In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.
2. In the Choose Analysis view, enter a name for the analysis. You can optionally describe your project.

If you are re-using an Old Analysis, select **Reuse Old Analysis** and select the name of the analysis you want to use.

Note: The name cannot have spaces or special characters.

3. Select **Targeted Resequencing**, then click **Next**.

4. In the Data To Be Analyzed table, click **Select the reference for the reads** to open the repository file browser.
5. In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:

data ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta

Select the file, then click **OK**.

The file name appears in the Reference column.

Note: To change the reference file, click **Select the reference for the reads**, then select another reference file.

6. After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from choosing references and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

1. In the Edit Analysis window, accept the default settings for mapping and pre-processing data in the Secondary Analysis section.
If you want to exclude BAMStats, click **Customize**. In the Customize Mapping window, uncheck BamStats, then click **OK**.
If you do not want to pre-process data, uncheck the box.
2. In the Tertiary Analysis section, accept the included Small Indel SNP Finding, CNV, and Small Indel modules, and Annotation settings and click **Next** to set module parameters. To include all modules, click the **All>>** button.

To exclude a module, select it in the Include column, then click the < button. To exclude all modules, click the **All<<** button.

To skip setting module parameters and review the analysis, click **Review**.


Set module parameters

This section describes the procedures for setting general parameters and parameters for the mapping, SNP Finding, and Small Indel modules. For descriptions of module parameters, see [“Targeted resequencing parameters” on page 100](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a  button.

Set general parameters

1. Enter an analysis assembly name.
2. Set the annotation.dbsnp, analysis.regions.file, and annotation.gtf.file parameters. Click the  buttons to open the File Chooser, navigate to reference files, and choose files.
Note: Verify that the file path to the annotation.gtf file is correct. An incorrect file path will result in a failed analysis.
3. If you want to accept the default parameters for all modules, click **Review**. If you want to edit module parameters, click **Next**.

Set Fragment Mapping parameters

1. There are four categories of mapping parameters: Main, Advanced, SAET, and BAMStats. Accept the default settings or click the tabs to edit the settings.
2. Click **Next** to edit Enrichment parameters.

Set Small Indel parameters

Categories of Small Indel parameters include Main, Advanced, and (if you selected Annotation for Small Indel output) Annotation. Accept the default settings or edit the settings.

Set Enrichment parameters

1. There are two categories of enrichment parameters: Main and Advanced. Accept the default settings or edit the settings.
2. Click **Next** to edit SM{ parameters.

Set SNP Finding parameters

1. Categories of SNP Finding parameters include Main, Advanced, and (if you selected Annotation for SNP Finding output) Annotation. Accept the default settings or edit the settings.
2. Click **Next** to edit Small Indel parameters.


If you are ready to start the analysis, click **Review** to proceed.

If you want to erase your edits, click **Cancel** to close the Edit Analysis window.

If you want to start the analysis later, click **Finish** to save your edits and close the Edit Analysis window.

Review and run the analysis

Note: For a description of whole transcriptome parameters, see [“Targeted resequencing parameters” on page 100](#).

The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  , next to parameter categories to show or hide the parameters.

1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**.
 The Run Analysis window closes.


Checking Analysis Status

1. In the Projects list, click the project to check the status of the analysis run.
 The Progress columns show the percentage of completion for the secondary and tertiary analyses.

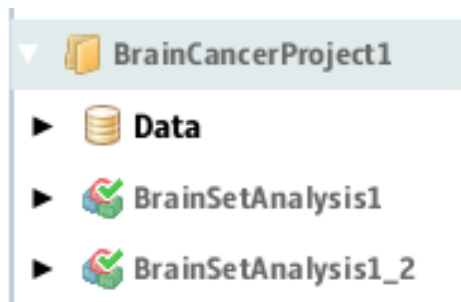
ProstateCancerProje... • ProstateCancerAnaly...

Overview Status Parameters

Analysis Runs



Name	XSQ ID	Analysis	Secondary Analysis	Secondary Progress	Tertiary Analysis	Tertiary Progress
solid0054_20110102_PE_LFD...	solid0054_20110102_PE_LFD...	ProstateCancerAnalysis1	[SAET, Mapping, BamStats]	 33%	[Enrichment, diBayes, SmallIn...	0%

When analysis is complete, a green check mark appears on the analysis name in the Projects organizer (shown [on page 59](#)).



View analysis results

To view analysis results:

1. In the Projects organizer, click the project to show its contents.
2. To view analysis results, click either the:
 - Results icon  for an overview of the analysis, then Click **View Results** in the **Secondary Progress** and **Tertiary Progress** columns to open the View Results window.
 - Completed analysis icon  for results of a specific analysis. Click the Status tab, then **View Results** to open the window.
3. In the View Results window, click each analysis module for analysis details.
4. Click each analysis module to see statistics, additional files, and logs.

The following table shows the results that are available for each analysis module.

Analysis Module	Details	Statistics	Additional Files	Logs
SmallIndel.Annotation	No	No	Yes	Yes
SNP.Finding.Annotation	No	No	Yes	Yes
SNP.Finding.targeted.frag	Yes	Yes	Yes	Yes
Enrichment	No	Yes	Yes	Yes
SAET	No	No	No	Yes
BAMStats	Yes	No	Yes	Yes
Mapping	No	No	Yes	Yes
SmallIndel	Yes	Yes	Yes	Yes

View results in a genome browser

Output files generated by Targeted Resequencing runs are compatible with third-party browser such as the Integrative Genomics Viewer (IGV) available from the Broad Institute and the UCSC Genome Browser. See [“Pairing information in a BAM file” on page 300](#) for information on using LifeScope™ Software output files with genome browsers.

For more information on viewing analysis results, refer to the chapter [Chapter 8, “View Analysis Results” on page 87](#).

Targeted resequencing output files

SmallIndel.Annotation

Details

There are no details for the SmallIndel.Annotation analysis module.

Statistics

There are no statistics for the SmallIndel.Annotation analysis module.

Additional files

The following table is an example of additional files for the SmallIndel.Annotation analysis module.

Directory	Files	File Type
SmallIndel.Annotation/ solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	README.txt	.txt

Logs

The following table is an example of logs for the SmallIndel.Annotation analysis module.

Directory	Files	File Type
SmallIndel.Annotation	tertiary-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary-SmallIndel.Annotation.20110421220403452.log	.log

SNP.Finding.Annotation

Details

There are no details for the SNP.Finding.Annotation analysis module.

Statistics

There are no statistics for the SNP.Finding.Annotation analysis module.

Additional files

The following table is an example of additional files for the SNP.Finding.Annotation analysis module.

Directory	Files	File Type
SNP.Finding.Annotation/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	README.txt	.txt

Logs

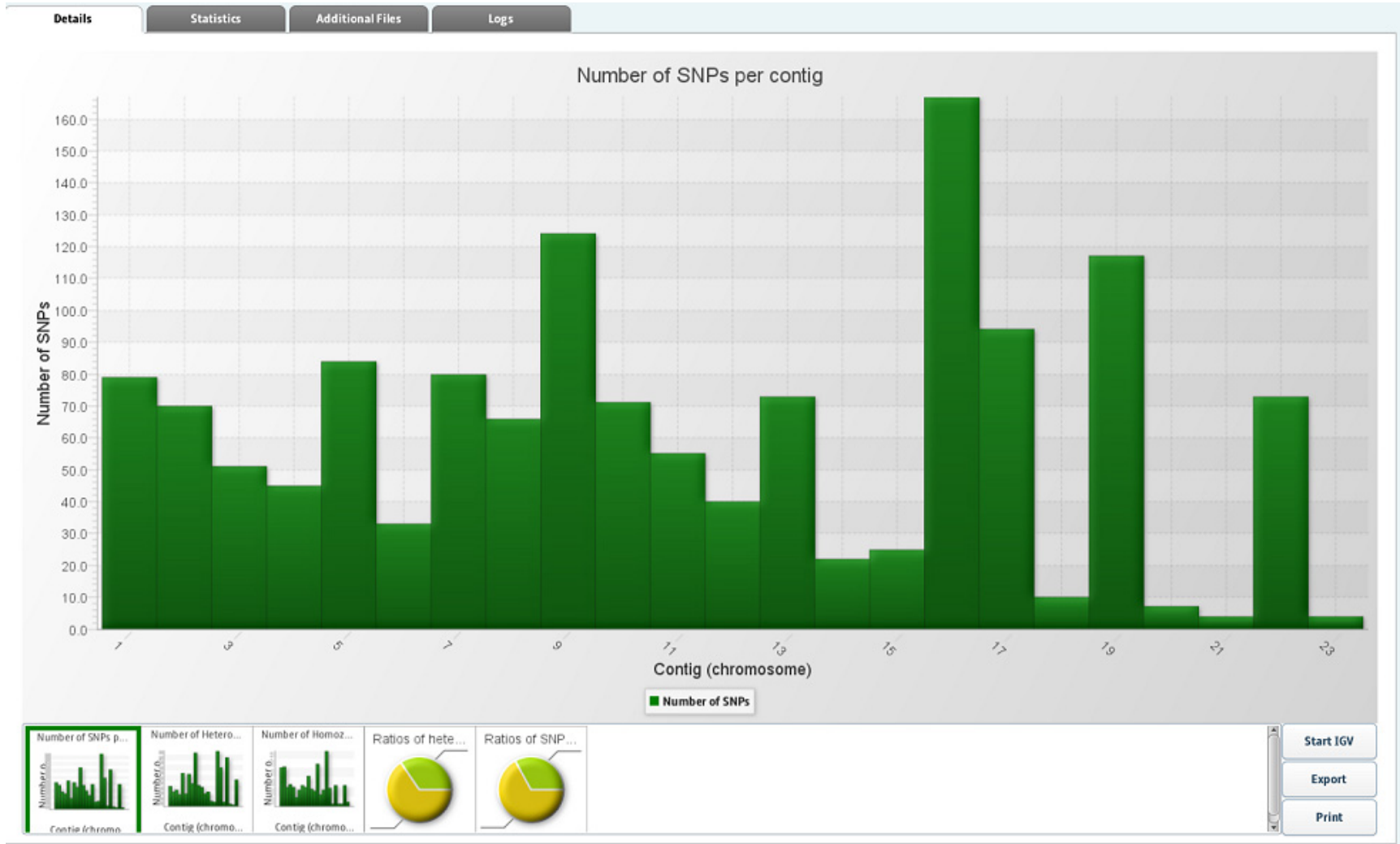
The following table is an example of SmallIndel.Annotation logs.

Directory	Files	File Type
SNP.Finding.Annotation	tertiary-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary-SNP.Finding.Annotation.20110421220103477.log	.log

SNP.Finding. targeted. frag

Details

The following illustration is an example of details for the SNP.Finding.targeted.frag analysis module.



Statistics

The following illustration is an example of statistics for the SNP.Finding.targeted.frag analysis module.

```

*****
LIFESCOPE REPORT
*****
Input File:      /panasas/lifescopetesttemp1/results/projects/corona/Darryl_Apr21/Analysis2/outputs/dibayes.targeted.
frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2_SNP.gff3
Date:   Apr 21, 2011 10:02:31 PM
Annotation file, dbSNP: /panasas/lifescopetesttemp1/results/projects/corona/Darryl_Apr21/Analysis2/outputs/dibayes.targeted.
frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2_SNP.gff3
Annotation file, Genes and Exons, GTF:/panasas/lifescopetesttemp1/results/projects/corona/Darryl_Apr21/Analysis2/outputs/dibayes.targeted.
frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2_SNP.gff3

*****
Statistics Overview
*****

-- Basic Statistics -----
Number of variants                1394
Number of heterozygous variants   913
Number of homozygous variants     481

-- Variant-Specific Statistics (SNP) -----
Number of transition SNPs         935
Number of transition heterozygous SNPs 603
Number of transition homozygous SNPs 332
Number of transversion SNPs      459
Number of transversion heterozygous SNPs 310
Number of transversion homozygous SNPs 149
Transition:Transversion ratio     2.037 : 1.000

-- dbSNP Annotation Statistics -----
Number of SNPs in dbSNP          1210
Number of heterozygous SNPs in dbSNP 731
Number of homozygous SNPs in dbSNP 479
dbSNP concordance                 86.80%
dbSNP heterozygous concordance    80.07%
dbSNP homozygous concordance     99.58%

-- Gene and Exon Statistics -----
Number (percent) of variants overlapping exon 453 ( 32.50%)
Number (percent) of heterozygous variants overlapping exon 288 ( 31.54%)

```

Additional files

The following table is an example of additional files for the SNP.Finding.targeted.frag analysis module.

Directory	Files	File Type
SNP.Finding.targeted.frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/contig2	Analysis2_SNP.gff3	.gff3
SNP.Finding.targeted.frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/contig19	Analysis2_Consensus_Calls.txt	.txt
SNP.Finding.targeted.frag/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/contig12	Analysis2_Consensus_Basespace.fasta	.fasta

Logs

The following table is an example of logs for the SmallIndel.Annotation analysis module.

Directory	Files	File Type
SNP.Finding.targeted.frag	tertiary-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary-SNP.Finding.targeted.frag.3.persistHeader.20110421214007357.log	.log

Enrichment

Details

There are no details for the enrichment analysis module.

Statistics

The following illustration is an example of statistics for the enrichment analysis module.

```

Run      Reads On      Percent On      Reads Off      Percent Off      Enrichment Fold
/panasas/lifescop20testtempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/enrichment/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F
3.colorRemake_DefaultLibrary/Analysis2.ontarget.bam      71681629      89.2239%      8657401      10.7761%      1347.43

Reads In Targets:      71681629      89.2239%
Reads Off Targets: 8657401      10.7761%
Ratio of Percent on/off Target: 8.27981
Total Target BP: 2039803
Total Genome Size: 3080436051
Ratio of target to genome: 0.00066218
Enrichment fold relative to target size: 1347.43

# Targets Not Covered      Target Bases Not Covered      Percent of Target Bases Not Covered      Percent of Target Covered >= 1X      Percent of
Target Covered >= 5X      Percent of Target Covered >= 10X      Percent of Target Covered >= 20X      Average Depth of Target Coverage
1      13067      0.64%      99.36%      98.94%      98.62%      98.22%      2451.59

Number of target regions with no coverage: 1
Percent of target bp not covered: 0.64% (13067 bp)
Percent of target bp covered at >= 1X: 99.36%
Percent of target bp covered at >= 5X: 98.94%
Percent of target bp covered at >= 10X: 98.62%
Percent of target bp covered at >= 20X: 98.22%
Maximum Depth of Coverage within targets: 51306
Average Depth of Coverage within targets: 2451.59
    
```

Additional files

The following table is an example of additional files for the enrichment analysis module.

Directory	Files	File Type
enrichment/solid0054_20110102_PE_ILFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	Analysis2.ontarget.bam.bai	.bai
	Analysis2_RainDance4K_targetHG18_withChr.validated.bed	.bed
	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary_target_statistics.txt	.txt
	Analysis2.ontarget.bam	.bam

Logs

The following table is an example of logs for the enrichment analysis module.

Directory	Files	File Type
enrichment	tertiary-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary-enrichment.main.20110421212207477.log	.log

SAET

Details

There are no details for the SAET analysis module.

Statistics

There are no statistics for the SAET analysis module.

Additional files

There are no additional files for the SAET analysis module.

Logs

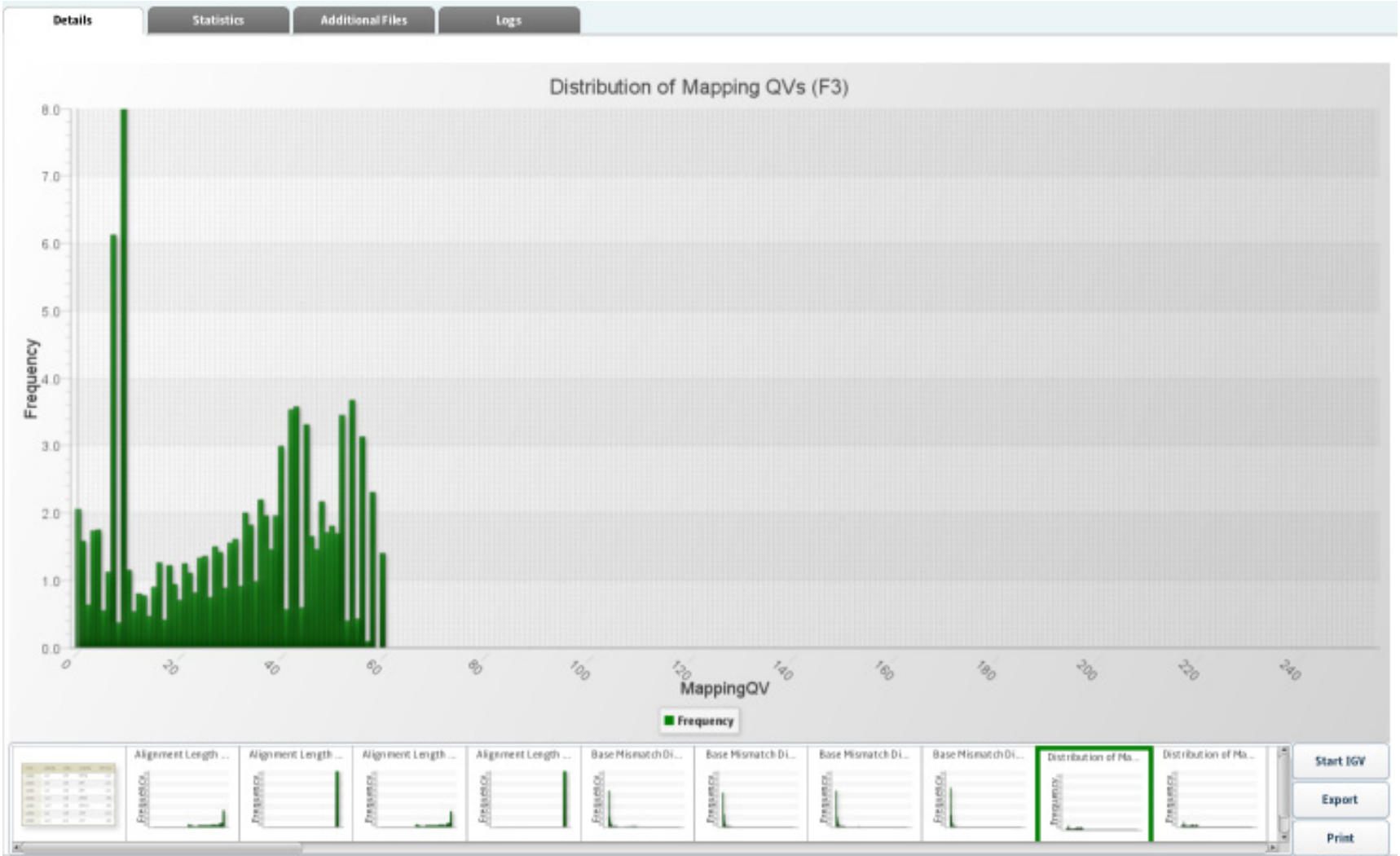
The following table is an example of logs for the SAET analysis module.

Directory	Files	File Type
Mapping.SAET	secondary-human_hg18-Mapping.SAET.main.20110421154741315.log	.log

BAMStats

Details

The following illustration is an example of BAMStats details.



Statistics

There are no statistics for the BAMStats analysis module.

Additional files

The following table is an example of BAMStats additional files.

Directory	Files	File Type
Mapping.BamStats/ solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/ solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1/Misc	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam_chrX.NEG.wig	.wig
	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.BaseQV.by.Position.csv	.csv
	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.Unique.Start.Positions.txt	.sql
	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam_chrX.NEG.wig	.wig

Logs

The following table is an example of logs for the BAMStats analysis module.

Directory	Files	File Type
Mapping.BAMStats	secondary-human_hg18-Mapping.BamStats.10.run.temp.10.20110421211327339.log	.log

Mapping

Details

There are no details for the Mapping analysis module.

Statistics

There are no statistics for the Mapping analysis module.

Additional files

The following table is an example of Mapping additional files.

Directory	Files	File Type
fragment.mapping/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam	.bam
	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam.bai	.bai

Logs

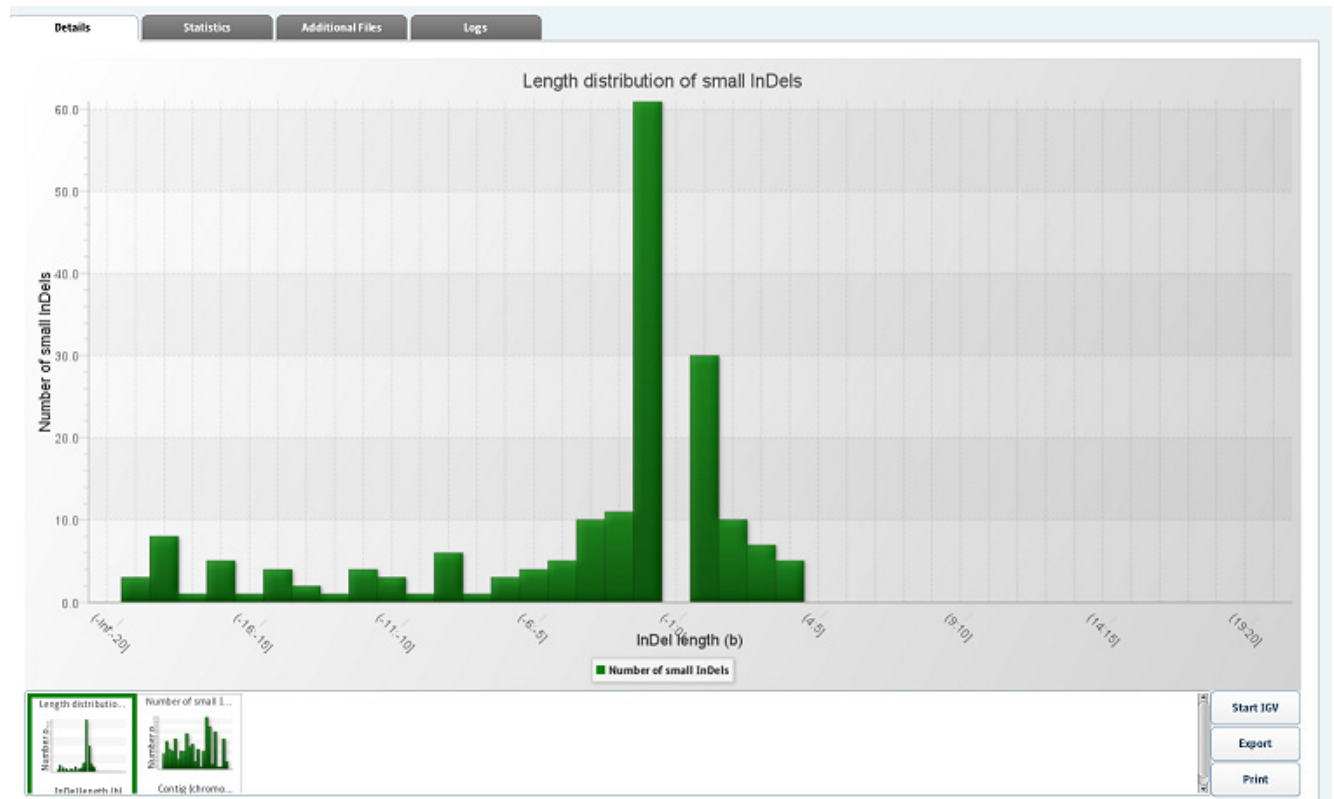
The following table is an example of logs for the BAMStats analysis module.

Directory	Files	File Type
fragment.mapping	secondary-human_hg18-fragment.mapping.3.run.20110421182159033.log	.log

Small Indel

Details

The following illustration is an example of Small Indel details.



Statistics

The following illustration is an example of Small Indel statistics.

```

*****
LIFESCOPE REPORT
*****
Input File:      /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.
indel/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2.gff3
Date:           Apr 21, 2011 10:05:52 PM
Annotation file, dbSNP: /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.
Annotation file, Genes and Exons, GTF: /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.

*****
Statistics Overview
*****

-- Basic Statistics -----
Number of variants                186
Number of heterozygous variants   127
Number of homozygous variants     59

-- Variant-Specific Statistics (InDel) -----
Distribution of InDel length

Length (bases)      Number
(-Inf:-20]          0
(-20:-19]           3
(-19:-18]           8
(-18:-17]           1
(-17:-16]           5
(-16:-15]           1
(-15:-14]           4
(-14:-13]           2
(-13:-12]           1
(-12:-11]           4
(-11:-10]           3
(-10:-9]            1
(-9:-8]             6
(-8:-7]             1
(-7:-6]             3
(-6:-5]             4
(-5:-4]             5
(-4:-3]            10
(-3:-2]            11
(-2:-1]            61

```

Additional files

The following table is an example of Small Indel additional files.

Directory	Files	File Type
small.indel/ solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.color Remake_DefaultLibrary	dbSnpDeployed_genes.bed	.bed
	dbSnpDeployed_genes.tab	.tab
	Analysis2.sql	.sql
	dbSnpDeployed_annotated.gff3	.gff3
	Analysis2.gff3	
	Analysis2.txt	.txt
	Analysis2.pas.sum	.sum
	Analysis2.align	.align
	Analysis2.ungapped	.ungapped
Analysis2.pas	.pas	

Logs

The following table is an example of Small Indel logs.

Directory	Files	File Type
small.indel	tertiary-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary-small.indel.20110421213917621.log	.log
	small.indel-solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary.20110421213922698.log	

Analysis Module	Additional Files	File Type
Small Indel Annotation	dbSnpDeployed_genes.bed	README.txt, .log
SNP.Finding Annotation	dbSnpDeployed_genes.tab	README.txt, .log
SNP.Finding Targeted Frag	Analysis2.sql	Charts, .stats, .gff3, .txt, fasta, .log
Enrichment	dbSnpDeployed_annotated.gff3	.stats, .bai, .bam, .bed, .txt, .log
SAET	Analysis2.gff3	.log
BAMStats	Analysis2.txt	Charts, .csv, .txt, .wig, .log
Mapping	Analysis2.pas.sum	.bai, .bam, .log
Small Indel	Analysis2.align	Charts, .stats, .align, .bed, .gff3, .pas, .sql, .sum, .tab, ungapped, .log
	Analysis2.ungapped	
	Analysis2.pas	

10

Perform Genomic Resequencing Analysis

This chapter covers:

■ Introduction to genomic resequencing analysis	129
■ Genomic resequencing analysis library types	129
■ Genomic resequencing analysis input files	129
■ Genomic resequencing analysis modules	130
■ Genomic resequencing analysis parameters	131
■ Perform genomic resequencing analysis	144
■ View analysis results	148
■ Genomic resequencing analysis output files	148

Introduction to genomic resequencing analysis

Genomic resequencing is a workflow for identifying individual variants and structural variations in a genomic sample.

Genomic resequencing output files are compatible with third-party genome browsers, including the Integrative Genome Viewer described [on page 47](#)).

Genomic resequencing analysis library types

Genomic resequencing analysis supports data from the following library types of data:

- Fragment
- Paired-end
- Mate-pair

Genomic resequencing analysis input files

Input File	File Extension	Analysis Module
Binary Alignment Map (BAM)	*.bam	Small Indel, SNP Finding, CNV
(Optional) Color space FASTA	*.csfasta	Mapping
eXtensible SeQuence (XSQ)	*.xsq	Small Indel, SNP Finding, CNV
FAST-All (FASTA)	*.fa, *.fasta	Mapping
(Optional) Quality Value	*qual, *qv	Mapping, pairing

Genomic resequencing analysis modules

Genomic resequencing analysis includes the following modules:

Secondary Analysis Module	Used for Library Type
Mapping	Fragment
Mapping and pairing	Paired-end
	Mate-pair

Tertiary Analysis Module	Used for Library Type
Small Indel	Fragment, mapping, pairing
SNP Finding	
CNV	

Mapping

The map data resequencing workflow uses eXtensible SeQuence (.xsq) files to create a .bam file.

Small Indel

When an indel occurs in a sequence, and that sequence is measured using color-space, the color-space sequence has a gap the same size as the indel. The color-space sequence also leaves a signature that can indicate whether there is a measurement error within the gap. The Small Indel tool targets the processing of indel evidences found in the pairing step during secondary offline data analysis.

You can optionally annotate input files. Annotation adds new attributes to the General Feature Format (GFF) entries in an input file, information from publicly available sources about the variants in an input file, features intersecting the variants in an input file, or any biological function potentially changed by the variants.

For information about using the Small Indel module to run a genomic resequencing analysis, see [Chapter 21, “Perform Small Indel Analysis” on page 259](#).

SNP Finding

The SNP Finding module is used to call Single Nucleotide Polymorphisms (SNPs) from mapped and processed SOLiD™ System color-space reads. The module takes the color-space reads, the quality values, the reference sequence, and error information on each SOLiD™ System slide as its input, and calls SNPs.

The module creates three results files:

- A list of SNPs.
- (Optional) A consensus *.fasta file with the same number of bases as the reference sequence.
- (Optional) A list of all covered positions.
- (Optional) A collection of annotated files.

For information about using the module to run a Targeted Resequencing experiment, see [Chapter 19, “Perform SNP Finding Analysis” on page 249](#).

CNV

The Copy Number Variation (CNV) analysis module detects CNV in SOLiD™ System data that originates from a single human sample. Slide(s) from this sample must be mapped to the hg18 reference to facilitate correct normalization.

For information about using the module, see [Chapter 17, “Perform Human Copy Number Variation Analysis”](#) on page 237.

Genomic resequencing analysis parameters

Parameters for genomic resequencing analysis include:

- General parameters
- **Secondary analysis:** Mapping (including optional BAMStats and SOLiD Accuracy Enhancement Tool [SAET])
- **Tertiary analysis:** Small Indel, SNP Finding, CNV

General

General parameters include:

Parameter	Default value	Description
analysis.assembly.name	unknown	The name of the genome assembly used in current analysis. Examples are hg18 and hg19.
annotation.dbsnp.file	n/a	The path to the file used to annotate SNPs and small InDels. LifeTech-provided files are dbSNP_b130.tab(hg18) and 00-All.vcf(hg19).
analysis.mappability.dir	n/a	The path to the directory containing binary mappability files used in CNV module.
analysis.regions.file	n/a	The path to the file containing genomic regions used in targeted resequencing selection, such as a .bed format file containing exome targets.
annotation.gtf.file	n/a	The path to the file containing gene and exon annotations corresponding to the genome assembly used in the analysis.

Use the  button to open the File Chooser and search for input files.

Mapping

Categories of parameters for mapping include Main, Advanced, (if you selected SAET pre-processing) SAET, and BAMStats.

Main

Parameter	Default value	Description
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	10	Replace resulting base-calls with a quality value less than the specified value. Allowed values: Integers 0–1,000
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: True: Create. False: Do not create.
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–1,000.
Primary output filter type	Primary_only	
Reference weight	8	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.

Parameter	Default value	Description
Second map gapped algorithm type	GLOBAL	Do indel finding, and control the behavior of indel finding. Allowed values: <ul style="list-style-type: none"> • GLOBAL: Mapping reports global alignment up to one indel. • LOCAL: Mapping reports local alignment up to one indel.

SAET

The SAET tab is accessible only if you select SAET pre-processing when you choose modules.

Parameter	Default value	Description
Genome length	2,800,000,000	Expected length of sequenced (or enriched) DNA region. For example, 4,600,000 for the E.Coli 4.6 MB genome or 30,000,000 for the entire Human Transcriptome. Allowed values: Integers ≥ 200 .
On target ratio	0.5	The expected ratio of reads that come from the targeted region. Allowed values: Floats 0.0–1.0.
Update quality values	True	Update quality value of modified calls. Allowed values: True: Update the QV of modified calls. False: Do not update the qv for modified calls.
Trusted quality value	25	Correction is applied to calls with a quality value below the value of this parameter. Allowed values: Integers ≥ 1 .
Support votes	2	The minimum number of k-mer votes required make a correction. Allowed values: Integers ≥ 1 .
Trusted frequency	0	The lowest multiplicity of the seed to be included in the spectrum. (If set to 0, then the value is computed internally.) Allowed values: Integers ≥ 0 .
Maximum corrections per read	0	Maximum number of allowed corrections per read. (If set to 0, then the value is set to $\lceil \text{readLength}/8 \rceil$). Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: Integers 0–9.

Parameter	Default value	Description
Number of recursive runs	1	The error correction step is repeated the provided number of times. Reduce if over-corrections are observed, or increase if undercorrections are observed. Allowed values: 1, 2, or 3.
Position of error inflation point	0	Position in the read at which the error rate inflates, for instance, 35-40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers.
Disable random sampling for large data	False	Disables random sampling in spectrum building. If set to 0, then for large datasets (coverage > 300x), a subset of reads is used in spectrum building. Allowed values: <ul style="list-style-type: none"> • True: Disables random sampling in spectrum building. • False: Do not disable random sampling in spectrum building. true?
K-mer size	0	Size of k-mer (>5) used in spectrum construction and error correction. (If set to 0, then the value is computed internally.) Allowed values: Integers 0-28.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	<code>\${analysis.output.dir}/bam</code>	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	<code>\${task.output.dir}</code>	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum Coverage	10000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 1 - 10,000.
Maximum insert size	100000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 1 - 100000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers 1 - 100000

Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 1 – 100000
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1–100000.

Small Indel

There are two categories of parameters for the Small Indel analysis module: Advanced and Annotation. There are no Main parameters.

Advanced

There are five categories of Advanced parameters: General Options, Pileup, Mapping Quality Filtering, Heuristic Filtering, Indel Size Filtering.

General Options

Parameters	Default value	Description
Detail level	3	For BAM file inputs, the level of detail in output: <ul style="list-style-type: none"> • 0: Keeps only position information about the anchor read and no information for the ungapped alignment. • 1-8: Keeps only some of the alignment's anchor alignment but none of the ungapped alignment. • 9: Is most detailed, but also the slowest. Allowed values: Integers 0 - 10
Zygoty profile name	max-mapping	Zygoty profile name. <ul style="list-style-type: none"> • classic Run classic mapping. • max-mapping: Run max mapping. • max-mapping-v2: Run max mapping version 2. • gap-align-only: • no-calls:
Genomic region	blank	Names a specific genomic region to be selected from the BAM file. Only full chromosomes are guaranteed not to alter results. Specifying partial chromosomes is allowed but may result in the loss of indels near the edges of that region.

Parameters	Default value	Description
Display base QVs	False	Display base QV scores in the GFF file. Allowed values: True: Display the FASTQ base QV scores for all of the reads used for each indel in the GFF file. FASTQ strings contain semi-colons, so adding these strings may produce a GFF file that is not compatible with certain applications. False: Do not display QV scores in the output file.
Number of alignments per pileup	1000	For pileups with more than this number of reads, set the expected number of alignments per pileup. Allowed values: Integers ≥ 0 .
Random seed	94404	The random seed value used to determine which pseudo random set of reads to use when there are greater than 1000 reads in a pileup. The random number generator used is the Mersenne Twister MT19937 algorithm. Allowed values: Integers ≥ 0 .

Pileup

Parameters	Default value	Description
Min num evid	2	Minimum number of evidences required for an indel call.
Max num evid	-1	Maximum number of evidences, use -1 for no maximum.
ConsGroup	1	Indel grouping method. Allowed values: 0: Conservative grouping of indels with 5bp max between consecutive evidences. 1: Lax grouping. Groups indels that are at maximum the higher of 15 or 7 times the indel size. 9: No grouping. Makes every indel evidence a separate pileup.

Mapping Quality Filtering

Parameters	Default value	Description
Max reported alignments	-1	Only uses those alignments where the NH field (the number of reported alignments; from the BAM header) is this value or lower.

Parameters	Default value	Description
Min mapping quality	8	Keeps only reads that have this or higher pairing qualities. For paired tags, mapping quality is for the pair (pairing quality), and for fragment, it is the single tag's map quality. Reads that are lower than this value are filtered out.
Min best mapping quality	10	For a particular indel called with a set of reads, at least one pairing quality in this set must be higher than this value. Allows for supporting evidences to have a lower mapping quality threshold than the best read.
Min anchor mapping quality	-1	Minimum mapping quality for a non-indel (anchor) tag. Effective only for paired reads, for the number of anchors queried as defined by <code>small.indel.detail.level</code> .
Ungapped BAM flag filter	ProperPair Primary	For ungapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Gapped BAM flag filter	Primary	For gapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Edge length datelines	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0 .
Edge length insertions	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0

Heuristic Filtering

Parameters	Default value	Description
Perform filtering	True	Whether or not to perform filtering in each pileup. Allowed values: <ul style="list-style-type: none"> • True: Perform filtering on pileups. • False: Do not perform filtering on pileups. Parameters that change the makeup of pileups, such as Min num evid are still active.
Indel size distribution allowed	can-cluster	Indel sizes in a pileup are allowed to have certain indel size distributions. Allowed values: <ul style="list-style-type: none"> • similar-size: 75% of the reads of a pileup must have exactly the same size. • similar-size-any-large-deletions: Any pileups with at least 2 large deletion alignments, the other pileups must have similar sizes. • can-cluster: Allowed if at least one cluster of any indel size is found. • can-cluster-any-large-deletions: Any pileups with at least 2 large deletion alignments; other pileups must be able to cluster (will have indels with two more reads with larger deletions, even if they don't form good clusters). any : Can have any size distribution (same as small .indel.require.called.indel.size=false)
Remove singletons	True	Remove the singletons that occur when different alignment methods are combined based on identical bead ids and read sequence. Allowed values: <ul style="list-style-type: none"> • True: Remove singletons. • False: Do not remove the singletons.
Alignment compatibility filter	1	Alignment compatibility level. Checks color space compatibility around the gap. Allowed values: <ul style="list-style-type: none"> • 0: No alignment compatibility filtering. • 1: The small indel module determines whether the data contains base-space or color-space sequence. • 2: Force the use of base-space sequence, if present. • 3: Force the use of color-space sequence, if present.
Max coverage ratio	12	Maximum clipped coverage/# non-redundant support ratio. Use -1 for no limit (no coverage ratio filtering). Allowed values: Integers.

Parameters	Default value	Description
Max nonreds 4Fit	2	Maximum number of non-redundant reads where read position filtering is applied. Allowed values: Integers.
Min from end pos	9.1	Minimum average number of base pairs from the end of the read required of the pileup, when there are at most a certain number of reads defined by Max nonreds 4Fit . Allowed values: Floats.

Indel Size Filtering

Parameters	Default value	Description
Min insertions size	0	Minimum insertion size to include. Allowed values: Integers.
Min deletions size	0	Minimum deletion size to include. Allowed values: Integers.
Max insertions size	1000000000	Maximum insertion size to include. Allowed values: Integers.
Max deletions size	1000000000	Maximum deletion size to include. Allowed values: Integers.

(Optional) Annotation

You can optionally annotate the mapped output of Small Indel analysis. For descriptions of the Annotation parameters, see “Annotation” [on page 144](#).

SNP Finding

There are two categories of SNP Finding parameters: Main and Advanced.

Main

Parameter name	Default value	Description
Call stringency	medium	Call stringency. Highest, high, medium, low, lowest
Skip high coverage positions (Het)	True	Skip high coverage positions (Het)
Minimum mapping quality value	8	Minimum mapping quality value. Allowed values: Integers 0-100

Advanced

There are six categories of Advanced parameters: general, Read filter, General position filter, Heterozygous position filter, Homozygous position filter, and Output file processing.

Parameter name	Default value	Description
Detect adjacent SNPs	False	Detect adjacent SNPs.
Polymorphism rate	0.001	Polymorphism rate. Allowed values: 0.0000 - 1.0000
Read filter		
Include reads with unmapped mate	False	Include reads with unmapped mate.
Exclude reads with indels	True	Exclude reads with indels.
Require only uniquely mapped reads	False	Require only uniquely mapped reads
Ignore reads with a higher mismatch count to alignment length ratio	1	Ignore reads with a higher mismatch count to alignment length ratio. Allowed values: 0.0000 - 1.0000
Ignore reads with a lower alignment length to read length ratio	0	Ignore reads with a lower alignment length to read length ratio. Allowed values: 0.0000 - 1.0000
General position filter		
Require alleles to be present in both strands	False	Require alleles to be present in both strands
Minimum base quality value for a position	20	Minimum base quality value for a position. Allowed values: Integers 0-40
Minimum base quality value of the non-reference allele of a position	20	Minimum base quality value of the non-reference allele of a position. Allowed values: Integers 0-40
Heterozygous position filter		
Minimum allele ratio (Het)	0.15	Minimum allele ratio (Het). Allowed values: 0.0000 - 0.5000

Parameter name	Default value	Description
Minimum coverage (Het)	2	Minimum coverage (Het). Allowed values: Integers ≥ 1
Minimum unique start position (Het)	2	Minimum unique start position (Het). Allowed values: Integers ≥ 1
Minimum non-reference color QV (Het)	7	Minimum non-reference color QV (Het). Allowed values: Integers 0-62
Minimum non-reference base QV (Het)	20	Minimum non-reference base QV (Het). Allowed values: Integers 0-41
Minimum ratio of valid reads (Het)	0.65	Minimum ratio of valid reads (Het). Allowed values: 0.0000 - 1.0000
Minimum valid tricolor counts (Het)	2	Minimum valid tricolor counts (Het). Allowed values: Integers ≥ 1
Homozygous position filter		
Minimum coverage (Hom)	1	Minimum coverage (Hom). Allowed values: Integers ≥ 1
Minimum count of the non-reference allele (Hom)	2	Minimum count of the non-reference allele (Hom). Allowed values: Integers ≥ 1
Minimum average non-reference base QV (hom)	20	Minimum average non-reference base QV (Hom). Allowed values: Integers 0-62
Minimum average non-reference color QV (hom)	7	Minimum average non-reference color QV (Hom). Allowed values: Integers 0-41
Minimum unique start position of the non-reference allele (Hom)	2	Minimum unique start position of the non-reference allele (Hom). Allowed values: Integers ≥ 1
Output file processing		
Output fasta file	True	Output or do not output the FASTA file. <ul style="list-style-type: none"> • True: Output the FASTA file. • False: Do not output the FASTA file.
Output consensus file	True	Output or do not output the consensus file. <ul style="list-style-type: none"> • True: Output the consensus file. • False: Do not output the consensus file.
Compress the consensus file	False	Compress or do not compress the consensus file. <ul style="list-style-type: none"> • True: Compress the consensus file. • False: Do not compress the consensus file.

(Optional) Annotation

You can optionally annotate the mapped output of SNP Finding analysis. For descriptions of the SNP Finding analysis parameters, see “Annotation” on page 144.

Human CNV

Categories of Human CNV parameters include Advanced and (if you selected Annotation for CNV output) Annotation. There are no Main parameters.

Advanced

There are two categories of Advanced parameters: general and `cnv.stringency.setting`.

Parameter	Default Value	Description
Window size	5000	Size of the window block to be considered as a region. Allowed values: Integers ≥ 100 .
Trim distance	1000	Distance in kilo bases to be trimmed from the extreme ends of the chromosome arms. Allowed values: Integers 0-100000
Min quality	0	Minimum quality value of the alignments. Allowed values: Integers 0-99
Ploidy	2	General ploidy of the genome. Allowed values: Integers ≥ 1
Ploidy exception	None	List of all the contigs whose ploidy is different to the general ploidy of the genome. Entries in the list are in the format {contig id: ploidy of the contig}, and are separated by commas. Use the string "None" to indicate no entries.
Local normalization	False	Whether or not genome-wide normalization or local normalization should be performed. <ul style="list-style-type: none"> • True: Perform chromosome-arm local normalization. • False: Perform genome-wide normalization.
Write coverage	False	Create coverage output files <ul style="list-style-type: none"> • True: Create coverage output files, in WIG format. • False: Do not create.
Coverage window size	1000	Size of the window block to be considered as a region for writing coverage output. The mean coverage of all bases in each of these windows is output. Allowed values: Integers: 1 – 100000

`cnv.stringency.setting` parameters

Parameter	Default Value	Description
Deletions min mappability	50	Minimum mappability percentage for regions to be shown as copy number deletions. Allowed values: 0.000-<100.0000 If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 0.
Insertions min mappability	10	Minimum mappability percentage for the regions to be shown as copy number insertions. Allowed values: 0.000-<100.0000. If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 0.

Deletion min windows	2	<p>Minimum number of windows for the regions to be shown as copy number deletions. Allowed values: Integers ≥ 1</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 4. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.
Insertion min windows	2	<p>Minimum number of windows for the regions to be shown as copy number insertions. Allowed values: Integers ≥ 1</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 4. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.
Deletion max p-value	1.0	<p>Maximum p-value for regions to be shown as copy number deletions. Allowed values: $>0.000-1.0000$.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this to 0.25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.0.
Insertion max p-value	1.0	<p>Maximum p-value for regions to be shown as copy number insertions. Allowed values: $>0.000-1.0000$.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this to 0.25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.0.
Stringency setting	Medium	High, medium, low
Deletion max ratio	0.5	<p>Maximum ratio between the coverage of the region and the expected coverage, for a region to be called as CNV deletion.</p> <p>Allowed values: $>0.000-<1.0000$.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then this parameter is set to 0.25. • Stringency setting is set to <i>Low</i>, then this parameter is set to 0.7.
Insert min ratio	1.25	<p>Minimum ratio between the coverage of the region and the expected coverage, for a region to be called as CNVs insertion.</p> <p>Allowed values: Floats ≥ 1.0.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then this parameter is set to 1.75. • Stringency setting is set to <i>Low</i>, then this parameter is set to 1.25.
Gender	Male	Set the ploidy of all chromosomes for Human.
Mappability directory	<code>\${analysis.mappability.dir}</code>	Path to the directory of mappability files.

**(Optional)
Annotation**

You can optionally annotate the mapped output of the analysis. For descriptions of Annotation parameters, see [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform genomic resequencing analysis**(Optional) Import
data**

You can optionally import data to be analyzed. For instructions on how to import data, see “Import Data” [on page 67](#).

**Log in to
LifeScope™
Software**

1. Navigate to LifeScope™ Software at `http://<IP address>:<port number>/LifeScope.html` where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown [on page 59](#)).

**Create or select a
project**

1. In the home view (shown [on page 59](#)), either click **Create a New Project** (described [on page 66](#)) or select a project in the Projects organizer (shown [on page 59](#)).
2. If you create a project:
 - a. Type a name and description in the Enter Project Name view.
Note: The name cannot have spaces or special characters.
 - b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown [on page 59](#)), click **Add Data to Project** to choose a data type.

**Add data to a
project**

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the read-sets you want to map and analyze.
If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).

- (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.

To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.

To remove a data file from a group, click the checkbox of the data file, then click **Delete**.

- Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

- In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.

- In the Choose Analysis view, enter a name for the analysis. You can optionally describe your project.

If you are re-using an Old Analysis, select **Reuse Old Analysis** and select the name of the analysis you want to use.

Note: The name cannot have spaces or special characters.

- Select **Genomic Resequencing**, then click **Next** to choose references.

- In the Data To Be Analyzed table, click **Select the reference for the reads** to open the repository file browser.

- In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:

data ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta

Select the file, then click **OK**.

The file name appears in the Reference column.

Note: To change the reference file, click on a button in the Reference column, then select another reference file.

- After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from choosing references and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

- In the Edit Analysis window, accept the default settings for mapping and pre-processing data in the Secondary Analysis section.

If you want to exclude BAMStats, click **Customize**. In the Customize Mapping window, uncheck BAMStats, then click **OK**.

If you do not want to pre-process data, uncheck the box.

- In the Tertiary Analysis section, accept the included Small Indel SNP Finding, Human CNV, and Small Indel modules, and Annotation settings and click **Next** to set module parameters. To include all modules, click the **All>>** button.
To exclude a module, select it in the Include column, then click the < button. To exclude all modules, click the **All<<** button.
To skip setting module parameters and review the analysis, click **Review**.


Set module parameters

This section describes the procedures for setting general parameters and parameters for the mapping, Small Indel, SNP Finding, and CNV modules. For descriptions of module parameters, see [“Genomic resequencing analysis parameters” on page 131](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a  button.

Set general parameters

- Enter an analysis assembly name.
- Set the analysis.regions.file parameter.
Click the  buttons to open the File Chooser, navigate to reference files, and choose files.
- If you want to accept the default parameters for all modules, click **Review**. If you want to edit module parameters, click **Next**.

Set mapping parameters

- There are three categories of mapping parameters: Main, Advanced, and BAMStats. Accept the default settings or click the Main, Advanced, and BAMStats tabs to edit the settings.
- Click **Next** to edit Small Indel parameters.

Set Human CNV parameters

- There are two categories of SNP Finding parameters: Advanced, and Annotation. There are no Main parameters. Accept the default settings or edit the settings.
- After you have edited module parameters:
 - Click **Review** to proceed, or
 - Click **Cancel** to erase your edits and close the Edit Analysis window, or
 - Click **Finish** to save your edits and close the Edit Analysis window.

Set SNP Finding parameters


- There are three categories of SNP Finding parameters: Main, Advanced, and Annotation. Accept the default settings or edit the settings.
- Click **Next** to edit CNV parameters.

Set Small Indel parameters

1. There are three categories of Small Indel parameters: Main, Advanced, and Annotation. Accept the default settings or edit the settings.
2. Click **Next** to edit SNP Finding parameters.

Review and run the analysis

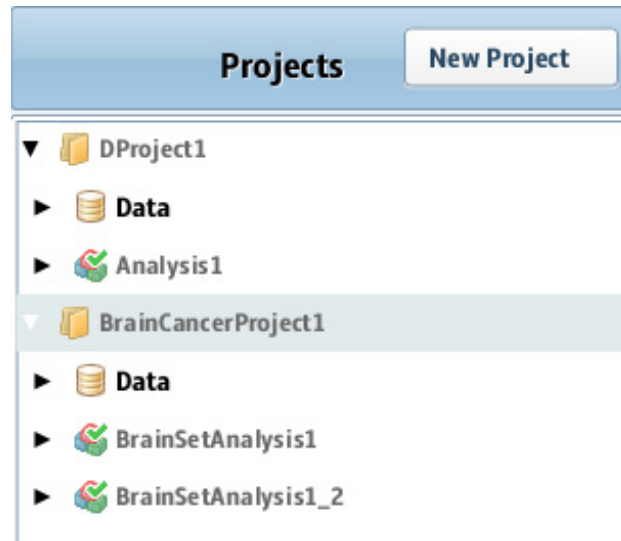
Note: For a description of genomic resequencing parameters, see “[Genomic resequencing analysis parameters](#)” on page 131.

The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  , next to parameter categories to show or hide the parameters.

1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**.
The window closes.

Checking Analysis Status

1. In the Projects organizer (shown [on page 59](#)), click the project to check the status of the analysis run.



2. Click the analysis name to show details about the analysis in the status overview (shown [on page 59](#)).
3. Click the Status tab to view the Progress column, which shows the percentage of completion for the Analysis.

LiverCancerProject1_042511

Overview Status

Analysis Runs

Name	XSQ ID	Analysis	Run Start Time	Secondary Analysis	Secondary Progress	Te
en_20100402_2_10103F...	helen_20100402_2_10103...	BrainSetAnalysis1	2011-02-09 17:00:00.0	[saet]	100%	

When analysis has been completed, a green check mark appears on the Analysis name.

4. Click the Analysis icon to see the overview of the results.
5. Click the **View Results in the Secondary Progress** box to open the View Results window.
6. Click each module step to see results in the tab sections.
7. Click each tab to view various results.
8. Close the View Results window to return to the Home view.

View analysis results

Refer to the chapter [Chapter 8, “View Analysis Results” on page 87](#) for instructions on how to review analysis results.

View results in a genome browser

Output files generated by Targeted Resequencing runs are compatible with third-party browser such as the Integrative Genomics Viewer (IGV) available from the Broad Institute and the UCSC Genome Browser. See [“Pairing information in a BAM file” on page 300](#) for information on using LifeScope™ Software output files with genome browsers.

Genomic resequencing analysis output files

See the following chapters for output files generated by a genomic resequencing analysis:

- [Chapter 17, “Perform Human Copy Number Variation Analysis” on page 237](#)
- [Chapter 19, “Perform SNP Finding Analysis” on page 249](#)
- [Chapter 21, “Perform Small Indel Analysis” on page 259](#)

11

Perform Small RNA Analysis

This chapter covers:

■ Introduction to Small RNA analysis	149
■ Small RNA analysis library types	149
■ Small RNA analysis input files	149
■ Small RNA analysis modules	152
■ Small RNA analysis parameters	153
■ Perform Small RNA analysis	157
■ View analysis results	161
■ Small RNA output files	161

Introduction to Small RNA analysis

LifeScope™ Genomic Analysis Software small RNA analysis modules are used to analyze high throughput small RNA sequencing raw data. These modules encapsulate a workflow to map reads to the target genome, to detect small RNAs in the genome, and to determine the expression of the detected small RNAs.

Small RNA analysis library types

Small RNA analysis supports the fragment library type.

Small RNA analysis input files

Small RNA analysis requires input for the Mapping, Coverage, and Counts modules:

Mapping input files

The small RNA mapping module requires the following input:

- One or more XSQ files containing sequencing data.
- A Genome Coordinates General Feature Format (GFF) file containing precursor sequences with genomic locations. This file is available at:
<http://www.mirbase.org/ftp>
- A single multi-fasta reference file for the filter mapping step.
- A single multi-fasta reference file for the genome mapping step.

The mapping modules accepts as input one or more input XSQ files.

RNA-Seq reads

The RNA-Seq reads used in the small RNA modules are different from the genomic reads. For RNA-Seq reads:

- Only transcribed sequences are measured by the system.
- Genome coverage is non-uniform due to variation in transcriptional intensity.
- A large subset of reads originate from uninteresting sequences such as ribosomal RNA (rRNA).
- A subset of the reads originates from splice junctions and cannot align contiguously on the genome.

Reads input specification

Plan your analysis input carefully. The way you define your reads input affects the behavior of your analysis. The following factors control how your input data is analyzed:

- **Index (barcode) IDs** – Using an index ID restricts your input to the reads data of one or more indices.
- **Grouping of reads** – Each group of reads is analyzed together as one specimen. The output data for a group is combined into one set of results files.
- **Multiple sample runs** – Unrelated reads can be processed together in one run of LifeScope™ Software, but analyzed separately as separate input data.

See [“Define input data” on page 31](#) for more information on designing how to add input data to your analysis.

Mapping one tag of paired data is not supported.

Legacy data

If the data you want to process with LifeScope™ Software is in CSFASTA and QUAL files, these files must be converted to the XSQ file format. The LifeScope™ Software graphical user interface automatically converts these files to the XSQ format before mapping data.

Coverage input files

The small RNA coverage module requires one or more BAM files containing mapped data. The following describe the alignment input accepted. This module accepts:

- Only fragment data for RNA libraries.
- Multiple BAM files as input.
- BAM files with different read lengths.
- Both color space and base space files, and a combination of color and base space data.

Counts input files

The small RNA counts module uses the following input:

- One or more BAM files containing mapped data.
- A Genome Coordinates GFF file containing precursor sequences with genomic locations.
- (Optional) A file containing mature form sequences with locations relative to precursor sequences. If the mature sequences file is not provided as input, the module compute counts per precursor sequence.

These files are described in the following sections: alignment files, precursor sequences, and mature form sequences.

BAM file metadata

The following table lists BAM file metadata used by the small RNA counts module.

Field name	Field in BAM header	Requirement
Library Type	The LT field in the @CO line for every @RG line. (Every @RG read group line has one @CO comment line.)	@RG lines must be present. Every @CO line must have an LT field, and the LT field must be set to the string "Fragment".
Sequence Name	The SN field in each @SQ header line.	Only contigs with an @SQ line with an SN field are analyzed. If a contig does not have an @SQ line in the header, the alignments in that contig are not counted.
Sort Order	The SO field in the HD header line.	The SO field must be present and must be set to the string "Chromosome".

Precursor sequences

This file is a Genome Coordinates GFF file containing precursor sequences with genomic locations. You must use the same file that was used during the small RNA mapping step. This file is available the following link:

<http://www.mirbase.org/ftp>

Mature form sequences

The mature form sequences file provides locations relative to precursor sequences. This tab-separated file is available the following link:

<ftp://mirbase.org/pub/mirbase/CURRENT/miRNA.xls>

Only this specific mature forms file is supported.

If the mature sequences file is not provided as input, the module compute counts per precursor sequence.

Small RNA analysis modules

Small RNA analysis includes the following modules:

Secondary Analysis Module	Used for Library Type
Small RNA Mapping	Fragment
BAMStats	

Tertiary Analysis Module	Used for Library Type
Small RNA Mapping	Fragment
Small RNA Filtered BAM Counts	
Small RNA Counts	
Coverage	

Small RNA Mapping

The small RNA mapping module maps small RNA (also known as microRNA or miRNA) reads using the mapping tool mapReads. This module maps the small RNA reads to three different references in three steps:

- In the first step, the given set of reads are mapped to filter sequences in order to eliminate the reads generated from irrelevant sources (such as tRNA, adaptor sequences, or others).
- In the second step, the remaining reads from the first step are mapped to the list of known small RNA precursor sequences (miRBase annotations) downloaded from Sanger's web site.
- In the third step, the unmapped reads from the second step are mapped to the genome reference sequence, in order to find novel small RNAs in the sample. The mapped reads outputs from the second and third steps are merged, and the merged file is the primary output of the small RNA mapping module.

The major components of this module are:

1. Filter reference mapping
2. miRBase reference mapping
3. Genome reference mapping
4. Merging of mapped reads output (MA files)

Parameters are provided to skip any of the three mapping steps.

Usual read lengths from small RNA sequencing are 35–50 bases, and the length of the miRNA fragments is only 18–28 bp. The reads contain both the miRNA sequence and a P2 adaptor sequence at the end. The mapReads phase does an extra step of extension onto adaptor, to correctly remove the adaptor sequence from the read and to report only small RNA sequences in the BAM file.

After the miRBase mapping step, the mapping output file (a MA file) has alignments with respect to the smaller miRBase reference locations. Before merging the MA files, the program converts these miRBase alignments to genomic reference locations, using the input GFF file (described in [“.gtf file format” on page 162](#)).

The mapping output files from miRBase mapping and genome mapping do not contain any common reads (common bead ids).

Coverage The small RNA coverage module calculates read coverage per position.

Counts The LifeScope™ Software small RNA counts module in calculates counts per precursor or mature sequence, where counts are the number of reads mapped. This module is similar to the WT counttags module, which computes counts per exon.

Small RNA analysis parameters

Parameters for small RNA analysis include:

- General parameters
- **Secondary analysis:** Small RNA Mapping, BAMStats
- **Tertiary analysis:** Small RNA Mapping, Small RNA Filtered BAM Counts, Small RNA Counts, and Coverage

General General parameters include:

Parameter name	Default value	Description
analysis.assembly.name	hg19	Name of the genome assembly used in current analysis. Examples are hg18 and hg19.
analysis.mirbase.precursor.file	/data/results/referenceData/lifetech/hg19/mirbase/hsa.gff	File containing precursor microRNA database annotations.
analysis.mirbase.mature.file	/data/results/referenceData/external/hg19/mirbase/miRNA.txt	File containing mature form microRNA database annotations.
analysis.filter.reference	/data/results/referenceData/lifetech/hg19/human_filter_reference.fasta	File containing a collection of sequences that are used to filter out commonly-occurring motifs and contaminating sequences prior to mapping.
analysis.sample.name	—	User-defined sample name.

Small RNA Mapping There are no Main parameters for the Small RNA Mapping module. Advanced parameters include:

Parameter name	Default value	Description
01 Mapping		

Parameter name	Default value	Description
Mapping in base space	False	Include base space data in the mapping. Allowed values: <ul style="list-style-type: none"> • True: Include base space data in mapping. • False: Do not include base space data in mapping.
Filter mapping	True	Turn filter mapping on or off. Allowed values: <ul style="list-style-type: none"> • True: Turn on filter mapping. • False: Turn off filter mapping.
miRBase mapping	True	Turn miRBase mapping on or off. Allowed values: <ul style="list-style-type: none"> • True: Turn on miRBase mapping. • False: Turn off miRBase mapping.
Genome mapping	True	Turn genome mapping on or off. Allowed values: <ul style="list-style-type: none"> • True: Turn on genome mapping. • False: Turn off genome mapping.
02 BAM Conversion		
Reference correction base filter QV threshold	10	Bases with a quality value below the value of this parameter are replaced with 'N'. Allowed values: Integers 0–255.
Reference weight	8	This parameter is used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. This parameter adds weight (in terms of a Phred score) to the signals that are compatible with the reference. Color combinations that result in a variant are considered compatible with the reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.
Create unmapped BAM files	False	Create or do not create BAM files with unmapped reads. Allowed values: <ul style="list-style-type: none"> • True: Create. • False: Do not create.
BAM generation mapping QV threshold	0	Provide control over the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–100.
BAM generation reference correction add color sequence	True	Add or do not add color sequence in BAM records. <ul style="list-style-type: none"> • True: Add color sequence. • False: Do not add color sequence.

Small RNA Counts There are no Main parameters for the Small RNA Counts module. Advanced parameters include:

Parameter name	Default value	Description
Counts min quality	2	The mapping quality value threshold for selecting the alignments. Alignments with a mapping quality value lower than this threshold are not written to output. Allowed values: Integers 0–100.
Counts per feature	True	Controls the aggregated count output. The primary key in the output GTF file is: {Feature, start, end}. Allowed values: <ul style="list-style-type: none"> • True: Reads of different starts and ends are counted separately for every feature and are output on separate lines. • False: The aggregated count of all the reads mapped to a feature are output on the same line.
Counts primary only	True	Use only primary alignments or use all the alignments from the input BAM file as input. Allowed values: <ul style="list-style-type: none"> • True: Consider only primary alignments (both gapped and ungapped). • False: Consider all alignments.
Coverage overflow limit	3	The maximum allowed offset on either side of an alignment or feature, when looking for overlap between an alignment and a feature. If this limit is set to 0, then the start and end coordinates of the alignment record and feature must exactly match. Allowed values: Integers 0–20.

Coverage There are no Main parameters for the Coverage module. Advanced Coverage parameters include:

Parameter name	Default value	Description
Coverage min quality	2	[Optional] The mapping quality value threshold for selecting the alignments. Alignments with a mapping quality value lower than this threshold are not written to output. Allowed values: Integers 0–100.
Coverage per chromosome	True	[Optional] Whether coverage output is generated as one file per every chromosome per strand, or as a single file with coverage of all chromosomes per strand. Allowed values: <ul style="list-style-type: none"> • True: Generate one file per every chromosome per strand. • False: Generate a single file with coverage of all chromosomes per strand.

Parameter name	Default value	Description
Coverage primary only	True	(Optional) Whether only primary alignments or all the alignments from the input BAM file are used as input. Allowed values: <ul style="list-style-type: none"> • True: Consider only primary alignments (both gapped and ungapped). • False: Consider all alignments.
Coverage min value	0	(Optional) The mapping quality threshold value for selecting alignments from the input BAM. In order to be included in the output, a position's coverage must be greater than or equal to this value. Allowed values: Integers ≥ 0 .

Small RNA Filtered BAM Counts There are no Main parameters for the Small RNA Filtered BAM Counts module. Advanced parameters include:

Parameter name	Default value	Description
Counts primary only	True	(Optional) Count only primary alignments from the BAM file or count all alignments from the input BAM file as input. True: Count only primary alignments (both gapped and ungapped). False: Consider all alignments.

Small RNA Filtered BAMStats There are two categories of Small RNA Filtered BAMStats parameters: Main and Advanced.

Main

Parameter name	Default value	Description
Maximum coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10,000.
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0–1.

Advanced

Parameter name	Default value	Description
Input directory for BAMStats	<code>\${analysis.output.dir}/fragment.mapping</code>	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.

Parameter name	Default value	Description
Output directory	`\${task.output.dir}`	The path to the output directory where BAMStats will write its chart (.cht) files.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers 1-100,000.

Perform Small RNA analysis

(Optional) Import data

You can optionally import data to be analyzed. For instructions on how to import data, see “Import Data” [on page 67](#).

Create or select a project

1. In the Home view (shown [on page 59](#)), either click **Create a New Project** (described [on page 66](#)) or select a project in the Projects organizer (shown [on page 59](#)).
2. If you create a project:
 - a. Type a name and description in the Enter Project Name view.
Note: The name cannot have spaces or special characters.
 - b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown [on page 59](#)), click **Add Data to Project** to choose a data type.

Add data to a project

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the read-sets you want to map and analyze.

If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).

- (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.

To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.

To remove a data file from a group, click the checkbox of the data file, then click **Delete**.

- Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

- In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.
- In the Choose Analysis view, enter a name for the analysis. You can optionally describe your project.

If you are re-using an Old Analysis, select **Reuse Old Analysis** and select the name of the analysis you want to use.

Note: The name cannot have spaces or special characters.

- Select **Small RNA**, then click **Next** to choose references.

- In the Data To Be Analyzed table, click **Select the reference for the reads** to open the repository file browser.
- In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:

data ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta

Select the file, then click **OK**.

The file name appears in the Reference column.

Note: To change the reference file, click on a button in the Select the reference for the reads column, then select another reference file.

- After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from choosing references and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

- In the Edit Analysis window, accept the Secondary Analysis default settings.

2. In the Tertiary Analysis section, you can include an analysis module by selecting it in the Available Modules column, then clicking the > button. To include all modules, click the **All>>** button.
To exclude a module, select it in the Include column, then click the < button. To exclude all modules, click the **All<<** button.
3. Click **Next** to set module parameters.

Set module parameters

This section describes the procedures for setting general parameters and parameters for mapping, see [“Genomic resequencing analysis parameters” on page 131](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a  button.

Set general parameters

1. Enter an analysis assembly name. and analysis sample name.
2. Set the parameters if necessary. The fields will be automatically populated if the structure of the reference directories is correctly set.
 - analysis.mirbase.precursor.file
 - analysis.mirbase.mature.file
 - analysis.filter.reference

Click the  buttons to browse directories and choose files.

3. If you want to accept the default parameters for all modules, click **Review**. If you want to edit module parameters, click **Next** to set Small RNA Mapping parameters.

Set Small RNA Mapping parameters

1. There are two categories of Small RNA Mapping parameters: Main and Advanced. Accept the default settings or click the Main and Advanced tabs to edit the settings.
2. Click **Next** to edit Small RNA Counts parameters.

Set Small RNA Counts parameters

1. Accept the default settings of the Advanced parameters or edit the settings. There are no Main parameters.
2. Click **Next** to edit Coverage parameters.

Set Coverage parameters

1. There are two categories of Coverage parameters: Main and Advanced. Accept the default settings or edit the settings.
2. Click **Next** to edit Small RNA Filter Mapping parameters.

Set Small RNA Filter Mapping parameters


1. There are two categories of Small RNA Filter Mapping parameters: Main and Advanced. Accept the default settings or click the Main and Advanced tabs to edit the settings.
2. Click **Next** to edit Small RNA Filtered BAMStats parameters.

Set Small RNA Filtered BAMStats parameters

1. There are two categories of Small RNA Filtered BAMStats parameters: Main and Advanced. Accept the default settings or click the Main and Advanced tabs to edit the settings.
2. After you have edited module parameters:
 - Click **Review** to proceed, or
 - Click **Cancel** to erase your edits and close the Edit Analysis window, or
 - Click **Finish** to save your edits and close the Edit Analysis window.

Review and run the analysis

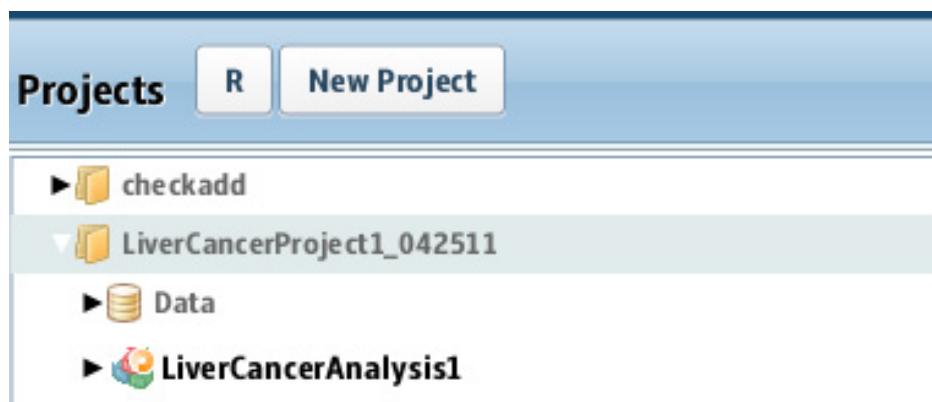
Note: For a description of genomic resequencing parameters, see [“Genomic resequencing analysis parameters” on page 131](#).

The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  , next to parameter categories to show or hide the parameters.

1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**.
The window closes.

Checking Analysis Status

1. In the Projects organizer (shown [on page 59](#)), click the project to check the status of the analysis run.



2. Click the analysis name to show details about the analysis in the status overview (shown [on page 59](#)).
3. Click the Status tab to view the Progress column, which shows the percentage of completion for the Analysis.

When analysis has been completed, a green check mark appears on the Analysis name.

4. Click the Analysis icon to see the overview of the results.
5. Click the **View Results in the Secondary Progress** box to open the View Results window.
6. Click each module step to see results in the tab sections.
7. Click each tab to view various results.
8. Close the View Results window to return to the Home view.

View analysis results

Refer to the chapter [Chapter 8, “View Analysis Results”](#) on page 87 for instructions on how to review analysis results.

View results in a genome browser

Output files generated by Targeted Resequencing runs are compatible with third-party browser such as the Integrative Genomics Viewer (IGV) available from the Broad Institute and the UCSC Genome Browser. See [“Pairing information in a BAM file”](#) on page 300 for information on using LifeScope™ Software output files with genome browsers.

Small RNA output files

Mapping output files

The output of Small RNA mapping is a BAM file.

Coverage output files

The small RNA coverage module’s output files are in the wiggle (*.wig) file format. Wiggle files are visualized in genomics browsers such as the Integrative Genomics Viewer (IGV). You can download the IGV browser from the Broad Institute web site:

www.broadinstitute.org/igv

The wiggle file specification is available from this site:

<http://genome.ucsc.edu/goldenPath/help/wiggle.html>

If the parameter `RNA.coverage.per.chromosome` is set to `false`, the small RNA coverage module generates two output files with coverage of all chromosomes per strand:

- `coverage_positive.wig`
- `coverage_negative.wig`

If the parameter `RNA.coverage.per.chromosome` is set to `true`, the small RNA coverage module generates one coverage file per chromosome per strand:

- `coverage_chr**_positive.wig`
- `coverage_chr**_negative.wig`

Counts output files The Small RNA counts module creates .gtf files containing the counts per feature, in precursor or mature form.

This is example content from the counts.gtf file:

```
chr1    mirBase hsa-miR-4251 3044539 3044599 0 - .
Precursor_ID "hsa-mir-4251"; RPM "0.00";
```

.gtf file format

The following table provides descriptions of the Small RNA Counts General Features Format (GFF) output file format.

Column name	Description	Examples
##	Header comment lines.	—
#	Header of the results.	—
seqname	The name of the sequence. Must be a chromosome or a scaffold.	chr1
source	The program that generated this feature.	mirBase
feature	The name of this type of feature.	start_codon stop_codon exon has-mir-* has-miR-*
start	The starting position of the feature in the sequence. The first base is numbered 1.	3044539
end	The ending position of the feature (inclusive).	3044599
score	The tag count for the feature.	0
strand	The strand containing the feature. Allowed values: <ul style="list-style-type: none"> • '+': The feature is on the 3' strand. • '-': The feature is on the 5' strand. • '.': Unknown or not required. 	-
frame	If the feature is a coding exon, frame is a number representing the reading frame of the first base. If the feature is not a coding exon, the value is '.'. Allowed values: <ul style="list-style-type: none"> • 0, 1, 2: The feature is a coding exon. • '.': The feature is not a coding exon. 	.
attributes	A list of attributes delimited by semi-colons. Each attribute is a type/value pair, with type and value separated by a single space character. Type is a string matching the pattern [A-Za-z1-9_]+. Value is a number or a double-quoted string.	Precursor_ID "hsa-mir-4251"; RPM "0.00";

Additional output includes "Reads Per Million RPM" for every feature defined as :

$$\text{RPM of a feature} = \frac{\text{\#count of feature} * \text{\#Total reads aligned to Sanger GFF file}}{1000000}$$

12

Perform Whole Transcriptome Analysis

This chapter covers:

■ Introduction to whole transcriptome analysis	163
■ Whole transcriptome analysis library types	164
■ Whole transcriptome analysis input files	164
■ Annotations input files	165
■ Reference input files	165
■ Whole transcriptome analysis modules	166
■ Whole transcriptome analysis parameters	167
■ Perform whole transcriptome analysis	174
■ View analysis results	178
■ Whole transcriptome analysis output files	178

Introduction to whole transcriptome analysis

This chapter has these purposes:

- To provide instructions for running whole transcriptome analysis single-read and paired-end modules using LifeScope™ Software.
- To list and describe configurable LifeScope™ Software parameters.

You can use the 5500 Series SOLiD™ Sequencer to sequence RNA prepared with RNA-Seq sample preparation kits. RNA sequencing produces high-depth, short-read sequencing data that can be used to measure RNA expression. Like microarray analysis, RNA-Seq measures expression intensity across many genomic features. Unlike microarray analysis, RNA-Seq can be used to identify novel transcriptome features in a sample.

Like other SOLiD™ applications, RNA-Seq produces short reads in the form of XSQ files. Using LifeScope™ Software, whole transcriptome analysis modules take these reads as input, and perform the following steps:

1. **Align reads** — The aligning reads step identifies the alignments between the reads and the reference genome sequence and reports the alignments as a BAM file. Whole transcriptome analysis mapping also makes use of gene annotations, which define exons, genes, and transcripts, in order to improve alignment.
2. **Count known exons** — The counting known exons step identifies the number of reads that align within genomic features.
3. **Calculate coverage** — The calculating coverage step reports the read coverage at each genomic position.

- 4. Find junctions** — The finding junctions step identifies splice junctions of various types, including fusion junctions from paired-end reads.

The 5500 Series SOLiD™ Sequencer™ System produces RNA-Seq reads in both single-read and paired-end configurations. While there is overlap in the analytical techniques, the analysis of single-read and paired-end data is performed with two distinct pipelines. The single-read pipeline consists of alignment, counting, and coverage steps. The paired-end pipeline extends single-read analysis with an alternative alignment step and an additional junction finding step.

Whole transcriptome analysis library types

Whole transcriptome analysis supports data from the following library types of data:

- Fragment
- Paired-end

Whole transcriptome analysis input files

The whole transcriptome analysis mapping module accepts as input one or more input XSQ files. The input reads can have different read lengths but they must be of the same library type. The XSQ files processed in a single analysis can not have different library types. In one mapping analysis, the input must be of only one library type, either fragment or paired-end.

Fragment mapping module supports mapping one tag of a paired-end read-set, but tertiary analysis does not run on BAM files generated from such processing.

RNA-Seq reads

The RNA-Seq reads used in whole transcriptome analysis are different from the genomic reads. For RNA-Seq reads:

- Only transcribed sequences are measured by the system.
- Genome coverage is non-uniform due to variation in transcriptional intensity.
- A large subset of reads originate from uninteresting sequences such as ribosomal RNA (rRNA).
- A subset of the reads originates from splice junctions and cannot align contiguously on the genome.

Multiple samples

LifeScope™ Software can perform alignment from multiple reads files (in XSQ format) for the same sample. If a sample is sequenced on the instrument in different runs or on different lanes, several XSQ files are generated for that sample. The reads from these XSQ files should be analyzed together. LifeScope™ Software facilitates the analysis by grouping the reads into read-sets.

Reference input files

WTA uses several files defining reference information: filter-reference, genome-reference and annotation files. The filter-reference file is a FASTA-formatted file defining uninteresting reference sequences such as Ribosomal RNA, tRNA, or vector sequences. To increase the speed of analysis, reads that align to sequences in the filter reference are excluded from downstream alignment steps. Many of the entries in the filter reference are species-specific. The filter reference file must be appropriate for the species being studied.

The principal output of alignment analysis is a set of alignments between the reads and the genome-reference in BAM format. The genome-reference file is a FASTA-formatted file defining sequences in the reference genome.

The annotation file is a GTF-formatted file defining known genes, transcripts, and exons in the genome reference. See “Annotations input files”.

Junction-reference and exon-reference files are generated internally by the software. These files are entirely derived from the GTF and genome-reference files. A user is not required to provide these references. The junction-reference file is a FASTA-formatted file with an entry for every potential intragenic splice-junction defined in the annotation file. The entry contains the position of the junction and the associated flanking sequence. The exon-reference is a FASTA-formatted file with an entry for every exon defined in the annotation file.

Annotations input files

While genomic resequencing relies only on alignment to a reference sequence, whole transcriptome analysis also makes use of gene annotations. Gene annotations define the exons, genes, and transcripts used to improve alignment.

Whole transcriptome analysis modules use genome annotation files in GTF format. Go to the following URLs to see the GTF format explained in detail:

genes.cse.wustl.edu/GTF2.html

genome.ucsc.edu/FAQ/FAQformat.html#format4

The GTF file must match the genome reference to ensure that the whole transcriptome analysis modules work correctly.

See the LifeScope™ Genomic Analysis Software Command Shell User Guide for more information.

UCSC genome annotations

Format UCSC Genome Browser Database annotations for WT analyses

The UCSC Genome Browser Database has genome annotations available for many assemblies at

hgdownload.cse.ucsc.edu/goldenPath/

The GTF-formatted annotations available for download are not properly normalized by gene ID. The required content is present for each assembly in the file export of the refGene database table `database/refGene/txt/gz`.

For example, annotation for human genome build 18 is available at:

hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz

Note: The GTF-formatted annotation is not in GTF format. You must convert the annotation before using it in whole transcriptome analysis.

Convert the refGene.txt.gz file

Run the script `bin/refgene2gff.sh` to convert the `refGene.txt.gz` file:

```
gunzip refGene.txt.gz
refgene2gff.sh -i refGene.txt -o refGene.gff
```

Genome annotations that are downloaded from the UCSC Genome Browser and converted by the annotation conversion script are optimal because they contain Human Genome Organization (HUGO)-style gene names. HUGO-style gene names allow interpretation when using a genome browser or reading reports.

The annotation conversion script works with the latest format of `refGene.txt` files. Assemblies, such as the rat genome, use an alternative format for the `refGene.txt` file. The `refgene2gff.sh` script does not convert alternative formats.

ENSEMBL GTF files Format ENSEMBL GTF files for whole transcriptome analysis pipelines

The ENSEMBL website ensembl.org/ has GTF-formatted genome annotations available for many popular assemblies. Unlike the GTF files directly downloadable from UCSC (see “Whole transcriptome analysis output file formats” on page 178), ENSEMBL GTF files are properly normalized by gene and transcript IDs. Visit the following site to download an ENSEMBL GTF file:

<http://www.ensembl.org/>

ENSEMBL GTF files use gene accession numbers instead of HUGO-style gene names. ENSEMBL GTF files also use unprefixed sequence identifiers, such as 1,2,3...X,Y,MT. The ENSEMBL GTF files are incompatible with genome reference FASTA files that have UCSC-style sequence IDs with the prefix “chr”, for example, chr1, chr2, chr3...chrX, chrY, chrM.

Reformat the ENSEMBL GTF file

To reformat the ENSEMBL GTF file to use UCSC-style gene IDs, run the script `reformat_ensembl_gtf.pl`:

```
reformat_ensembl_gtf.pl Homo_sapiens.GR
```

Whole transcriptome analysis modules

Whole transcriptome analysis includes the following modules:

Secondary Analysis Module	Used for Library Type
Mapping	Fragment, paired-end
Whole transcriptome splice junction extractor	Paired-end

Tertiary Analysis Module	Used for Library Type
Whole transcriptome exon sequence extractor	Paired-end
Coverage	Fragment, paired-end
Whole transcriptome count features	
Splice finding	

Whole Transcriptome Fragment Mapping

Whole transcriptome fragment mapping uses eXtensible SeQuence (.xsq) files to create a *.bam file.

Whole Transcriptome Exon Sequence Extractor

Extract sequences of exons to prepare a special reference for mapping and tertiary algorithms.

Coverage

The whole transcriptome coverage module calculates read coverage per position.

Whole Transcriptome Count Features

Filter, count, and normalize the reads that align on exons and genes, to generate easy-to-use expression output.

Splice Finding

Detect and quantify two types of evidences for splicing junctions, powered by the SASR (Suffix Array Single Read) aligner, to reconstruct transcript graphs and discover new fusion transcripts.

Whole transcriptome analysis parameters

Parameters for whole transcriptome analysis include:


- General parameters
- **Secondary analysis:** Mapping (including optional BAMStats)
- **Tertiary analysis:** Whole Transcriptome Fragment Mapping, Coverage, Whole Transcriptome Count Features, and Whole Transcriptome Exon Sequence Extractor, and Whole Transcriptome Splice Junction.

General

General parameters include:

Parameter	Default value	Description
analysis.assembly.name	—	The name of the genome assembly used in current analysis. Examples: hg18, hg19.
analysis.filter.reference	—	The file containing a collection of sequences that are used to filter out commonly-occurring motifs and contaminating sequences prior to mapping.

Parameter	Default value	Description
annotation.gtf.file	—	The file containing gene and exon annotations corresponding to the genome assembly used in the analysis.

Use the  button to open the File Chooser and search for input files.

Whole Transcriptome Fragment Mapping

There are two categories of mapping parameters: Main and Advanced.

Main

Parameter	Default value	Description
Enable filter mapping	True	Perform filter mapping. Allowed values: <ul style="list-style-type: none"> • true: Perform filter mapping. • false: Do not perform filter mapping.
Enable junction mapping	True	Perform junction mapping. Allowed values: <ul style="list-style-type: none"> • true: Perform junction mapping. • false: Do not perform junction mapping.
Map in base space	False	Map in base space. Set to true if input data has base space available. Allowed values: <ul style="list-style-type: none"> • true: Map in base space. • false: Map in color space. If only color space is available, then the module fails when base space mapping is turned on.
Add color sequence	True	Map in color space. Set to true if input data has color space available. Allowed values: <ul style="list-style-type: none"> • true: Map in base space. • false: Map in color space. If only color space is available, then the module fails when base space mapping is turned on.

Advanced

Parameter	Default value	Description
Min junction overhang	8	Alignments to junctions are not reported if they have fewer than this number of bases on aligned on either side of the splice. Allowed values: Integers 0 - 100.

Parameter	Default value	Description
Make unmapped BAM files	False	Create an additional BAM output file containing unmapped reads. Allowed values: <ul style="list-style-type: none"> true: Create a BAM output file containing the unmapped reads. false: Do not create the unmapped reads output file.
Mapping quality threshold	0	Minimum mapping quality for an alignment to be reported in a BAM file. Allowed values: Integers 0 - 255
Refcor base qv filter	2	Bases with a quality value below the value of this parameter are replaced with "N". Allowed values: Integers 0 - 255
Reference weight	8	This parameter is used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. This parameter adds weight (in terms of a Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0 - 100

Whole Transcriptome

Parameter	Default value	Description
Read length	75	Read length

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	\${analysis.output.dir}/wt.frag.map	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	\${task.output.dir}	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum Coverage	10000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 1 - 10,000.

Parameter	Default value	Description
Maximum insert size	100000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 1 - 100000.
Insert bin size	1000	Bin size for insert range distribution. Allowed values: Integers ≥ 1 .
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0, 1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> 0: Do not restrict coverage in WIG file format to only primary alignments. 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1 .

Whole Transcriptome Exon Sequence Extractor

There are no Whole Transcriptome Exon Sequence Extractor parameters to review or edit.

Coverage

You can optionally edit Advanced Coverage parameters.

Parameter	Default value	Description
Coverage min value	0	The mapping quality value threshold for selecting the alignments. Alignments with a mapping quality value lower than this threshold are not written to output. Allowed values: Integers 0 to 100
Coverage min quality	2	The mapping quality threshold value for selecting alignments from the input BAM. In order to be included in the output, a position's coverage must be greater than or equal to this value. Allowed values: Integers ≥ 0 .

Parameter	Default value	Description
Coverage per chromosome	True	Generate coverage output as one file per every chromosome per strand, or as a single file with coverage of all chromosomes per strand. Allowed values: <ul style="list-style-type: none"> true: Generate one file per every chromosome per strand. false: Generate a single file with coverage of all chromosomes per strand.
Coverage primary only	True	Use only primary alignments or all the alignments from the input BAM file as input. Allowed values: <ul style="list-style-type: none"> true: Consider only primary alignments (both gapped and ungapped). false: Consider all alignments.

Whole Transcriptome Count Features

There are two categories of Whole Transcriptome Count Features parameters: Main and Advanced.

Main parameters

Parameter	Default value	Description
Global		
*GTF file	\$(annotation.gtf.file)	The GTF file used to create the gene model.

Advanced parameters

Parameter	Default value	Description
Generate gene count	True	Generate gene count. Allowed values: <ul style="list-style-type: none"> true: Generate gene count. false: Do not generate gene count.
Counts primary	True	Use only primary alignments or use all the alignments from the input BAM file as input. Allowed values: <ul style="list-style-type: none"> true: Consider only primary alignments (both gapped and ungapped). false: Consider all alignments.
Mapping quality value	10	Minimum mapping quality value for the alignment. Allowed values: Integers 0–255.

Parameter	Default value	Description
Overflow limit	3	The maximum allowed offset on either side of an alignment or feature, when looking for overlap between an alignment and a feature. If this limit is set to 0, then the start and end coordinates of the alignment record and feature must match exactly. Allowed values: Integers 0–100.

Splice Finding You can optionally use the Splice Finding analysis module. There are two categories of Splice Finding parameters: Main and Advanced.

Main

Parameter	Default value	Description
Global		
*GTF file	\$(annotation.gitf.file)	The GTF file used to create the gene model.
General		
First read max read length	50	Maximum read length of the first read. Allowed values: 25–150
Second read max read length	25	Maximum read length of the second read. Allowed values: 25–150
Minimum single read evidence for junction	1	Minimum single read evidence for junction. Minimum value: 0 Maximum value: 1
Minimum paired end read evidence for junction	1	Minimum paired end read evidence for junction. Minimum value: 0 Maximum value: 1
Minimum combined evidence for junction	2	Minimum combined evidence for junction. Minimum value: 0 Maximum value: 1
Minimum single read evidence for alternative splicing	1	Minimum single read evidence for alternative splicing. Minimum value: 0 Maximum value: 1
Minimum paired-end read evidence for alternative splicing	1	Minimum paired-end read evidence for alternative splicing. Minimum value: 0 Maximum value: 1

Parameter	Default value	Description
Minimum combined evidence for alternative splicing	2	Minimum combined evidence for alternative splicing. Minimum value: 0 Maximum value: 1
Minimum split read evidence for fusion	2	Minimum split read evidence for fusion. Minimum value: 0 Maximum value: 1
Minimum paired end evidence for fusion	2	Minimum paired end evidence for fusion. Minimum value: 0 Maximum value: 1
Minimum combined evidence for fusion	4	Minimum combined evidence for fusion. Minimum value: 0 Maximum value: 1
Single Read		
Single read	1	Run single read fusion finding. Allowed values: <ul style="list-style-type: none"> • 0: Do not run single read fusion finding. • 1: Run single read fusion finding
Paired End		
Paired read	0	Run paired-end fusion finding. Allowed values: <ul style="list-style-type: none"> • 0: Do not run paired-end fusion finding. • 1: Run paired-end fusion finding.
Paired read min mapq	10	A read-pair that has mapping quality lower than this value is inadmissible as evidence. <ul style="list-style-type: none"> • Minimum value: 0 • Maximum value: 255
Output		
Output format	4	Output format parameter. Allowed values: <ul style="list-style-type: none"> • 1 = tabular output format • 2 = BED output format • 3 = BED, tabular, and SEQ output formats • 4 = all formats, including Circos With option 3, tabular, BED and also additional "seq" files are separately created. Seq files contain 50 base pairs of sequences from each end of a junction exon and are useful for validation. <ul style="list-style-type: none"> • Minimum value: 1 • Maximum value: 4

Perform whole transcriptome analysis

(Optional) Import data

You can optionally import data to be analyzed. For instructions on how to import data, see “Import Data” [on page 67](#).

Log in to LifeScope™ Software

1. Navigate to LifeScope™ Software at a url given to you by your network administrator.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown [on page 59](#)).

Create or select a project

1. In the home view (shown [on page 59](#)), either click **Create a New Project** (described [on page 66](#)) or select a project in the Projects organizer (shown [on page 59](#)).
2. If you create a project:
 - a. type a name and description in the Enter Project Name view.
Note: The name cannot have spaces or special characters.
 - b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown [on page 59](#)), click **Add Data to Project** to choose a data type.

Add data to a project

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the read-sets you want to map and analyze.
If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).
3. (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.
To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.

To remove a data file from a group, click the checkbox of the data file, then click **Delete**.

4. Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

1. In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.

2. In the Choose Analysis view, enter a name for the analysis. You can optionally describe your project.

Note: The name cannot have spaces or special characters.

Note: Note: If you are reusing an old analysis, select **Reuse Old Analysis** and select the name of your desire analysis.

3. Select **Whole Transcriptome** then click **Next** to choose references.



4. Click **Select the reference for the reads**.
5. In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:
 - data ▶ results ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta

Select the file, then click **OK**.

The file name appears in the Reference column.

Note: To change the reference file, click **Select the reference for the reads**.

6. After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from creating the analysis and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

1. In the Edit Analysis window, select the type of analysis modules to run during and after secondary analysis mapping.

If you want to exclude BAMStats, click **Customize**. In the Customize Mapping window, uncheck BAMStats, then click **OK**.

If you do not want to pre-process data, uncheck the box.

2. In the Tertiary Analysis section, you can include an analysis module by selecting it in the Available Modules column, then clicking the > button. To include all modules, click the **All>>** button.

To exclude a module, select it in the Include column, then click the < button. To exclude all modules, click the **All<<** button.

3. Click **Next** to set module parameters.


Set module parameters

This section describes the procedures for setting general parameters and parameters for the Whole Transcriptome Fragment Mapping, Coverage, Whole Transcriptome Count Features, and Whole Transcriptome Count Exon Sequence Extractor modules. For descriptions of module parameters, see [“Whole transcriptome analysis parameters” on page 167](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a  button.

Set general parameters

1. Enter an analysis assembly name.
2. Set the `analysis.fliter.reference` and `annotation.gtf.file` parameters.
Click the  buttons to open the File Chooser, navigate to reference files, and choose files.
Note: Verify that the file path to the `annotation.gtf` file is correct. An incorrect file path will result in a failed analysis.
3. If you want to accept the default parameters for all modules, click **Review**. If you want to edit module parameters, click **Next**.

Set Whole Transcriptome Fragment Mapping parameters

1. There are four categories of mapping parameters: Main, Advanced, Whole Transcriptome and BAMStats. Accept the default settings or click the tabs to edit the settings.
2. Click **Next** to edit Coverage parameters.

Set Coverage parameters

1. There is one category of Coverage parameters: Advanced. Accept the default settings or edit the settings.
2. Click **Next** to edit Whole Transcriptome Count Features parameters.

Set Whole Transcriptome Count Features parameters


1. There are two categories of Whole Transcriptome Count Features parameters: Main and Advanced. Accept the default settings or edit the settings.

There are no Whole Transcriptome Exon Sequence Extractor parameters.

2. If you are ready to start the analysis, click **Review** to proceed.
3. If you want to erase your edits, click **Cancel** to close the Edit Analysis window.
4. If you want to start the analysis later, click **Finish** to save your edits and close the Edit Analysis window.

Review and run the analysis

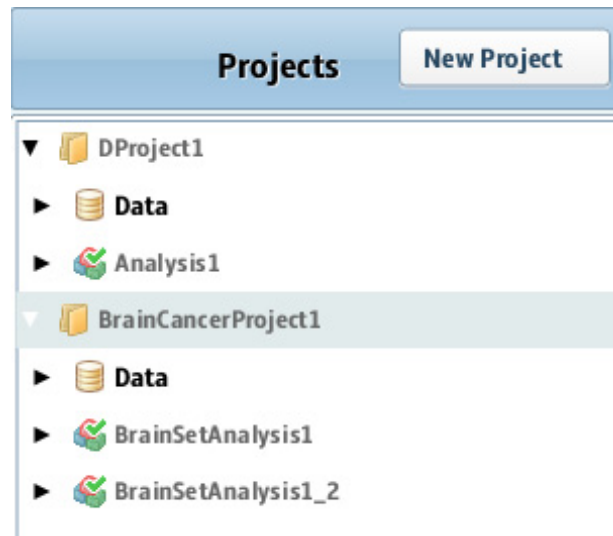
Note: For a description of whole transcriptome parameters, see “[Whole transcriptome analysis parameters](#)” on page 167.


The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  , next to parameter categories to show or hide the parameters.

1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**.
The Run Analysis window closes.

Check Analysis Status

1. In the Projects organizer (shown [on page 59](#)), click the project to check the status of the analysis run.



2. Click the analysis name to show and show details about the analysis in the status overview (shown [on page 59](#)).
To see analysis results in the status overview, click the Results icon  in the Projects organizer.
3. Click the Status tab to view the Progress column, which shows the percentage of completion for the analysis.
When analysis has been completed, a green check mark appears on the analysis module icon in the Projects organizer.
4. Click the Analysis icon to see the overview of the results. In the status overview, click the Status tab to see the progress of the analysis run.
5. Click **View Results** in a progress column to open the View Results window. In a progress column, a failed analysis is indicated by Fail!
6. Click each module to see results in the Overview and Status tabs.
7. Click each tab to view various results.
8. Close the View Results window to return to the Home view.

View analysis results

Refer to the chapter [Chapter 8, “View Analysis Results”](#) on page 87 for instructions on how to review analysis results.

View results in a genome browser

Output files generated by Targeted Resequencing runs are compatible with third-party browser such as the Integrative Genomics Viewer (IGV) available from the Broad Institute and the UCSC Genome Browser. See [“Pairing information in a BAM file”](#) on page 300 for information on using LifeScope™ Software output files with genome browsers.

Whole transcriptome analysis output files

Input file type(s)	Output file type
*.xsq	*.bam
*.bam	*.gff.3, *.gtf, *.txt, *.wig
*.bam	position error

Whole transcriptome analysis output file formats

BAM files

The whole transcriptome analysis paired-end pipeline produces BAM files that are identical to those produced by the resequencing pipeline samtools.sourceforge.net/SAM1.pdf. However, the BAM files from the whole transcriptome analysis single-read pipeline differ from BAM files produced elsewhere in LifeScope™ Software.

Alignment report

WT mapping runs produce an alignment report, the `alignmentReport.txt` file in the results directory.

Below is an example of the alignment report produced by the WT pipeline (the frequency lists have been truncated).

The 100.0% value reported for the “Reads mapped, not filtered” line should be interpreted as “Reads mapped, not filtered” comprising 100% of the category described in its section, not that the category “Reads mapped, not filtered” comprises 100% of the total reads.

```

-----
ALIGNMENT REPORT
-----

Counts:
Total reads:                102,837,343  (100.0%)
Reads mapped:               78,065,351  ( 75.9%)
Reads filtered:             12,029,269  ( 11.7%)
-----
Reads mapped, not filtered  68,780,034  (100.0%)
Reads with too many mappings (N >= 10):  5,897,833  (  8.6%)

```

```

Reads with number of mappings in proper range (N < 10):
62,882,201    ( 91.4%)
Reads uniquely aligned (score.clear.zone = 3):
53,417,152    ( 77.7%)
Reads uniquely aligned to junctions          4,816,207    (  7.0%)
Reads uniquely aligned to known junctions    4,773,948    (  6.9%)
    
```

ABSOLUTE FREQUENCIES

```

Alignment_Score    All_Read_Alignments_(N=_936)
Primary_Read_Alignments_(N=_935)
Unique_Read_Alignments_(N=_866)
0.0      0      0      0
1.0      0      0      0
2.0      0      0      0
36.0     0      0      0
37.0     87     87     87
38.0     0      0      0
39.0     0      0      0
47.0     0      0      0
48.0     0      0      0
49.0    849    848    779
50.0     0      0      0
Total    936    935    866
    
```

RELATIVE FREQUENCIES

```

Alignment_Score    All_Read_Alignments_(N=_936)
Primary_Read_Alignments_(N=_935)
Unique_Read_Alignments_(N=_866)
0.0      0.0%    0.0%    0.0%
1.0      0.0%    0.0%    0.0%
2.0      0.0%    0.0%    0.0%
36.0     0.0%    0.0%    0.0%
37.0     9.3%    9.3%    10.0%
38.0     0.0%    0.0%    0.0%
39.0     0.0%    0.0%    0.0%
40.0     0.0%    0.0%    0.0%
47.0     0.0%    0.0%    0.0%
48.0     0.0%    0.0%    0.0%
49.0    90.7%    90.7%    90.0%
50.0     0.0%    0.0%    0.0%
Total   100.0%  100.0%  100.0%
    
```

WT filtering stats

An example whole transcriptome human filter reference is provided under the `<examplesdir>/demos/WholeTranscriptome/references` folder in the optional LifeScope™ Software examples distribution. This example reference fasta includes contigs from barcode primers, human ribosomal RNAs, tRNAs, and other known targets for filtering. The filter reference may be expanded or removed with more csfasta reference records at the discretion of the user.

The stats for filtered reads are generated after a LifeScope™ Software run. First two lines of the stats file states the number of reads processed, and the number of reads mapped to the filtered reference followed by its percentage. Each following line reports a contig name and the count of reads that aligned to that contig. Paired-end filtering stats may be found in the Intermediate folder. The following is a truncated example of filtering stats output:

```

countOfReadsProcessed: 175,839,941
countOfReadsMapping:   18,783,220 (10.7%)
...
Barcode-047-3-end reverse    83
Barcode-048-3-end reverse   2,350
gi|124517659|ref|NR_003286.1| Homo sapiens 18S ribosomal RNA
(LOC100008588)    5,175,899
gi|142372596|ref|NR_003285.2| Homo sapiens 5.8S ribosomal RNA
(LOC100008587)    78,936
gi|124517661|ref|NR_003287.1| Homo sapiens 28S ribosomal RNA
(LOC100008589)    7,616,212
chr6.trna95-AlaAGC (58249908-58249836) Ala (AGC) 73 bp Sc:
42.26    5
chr6.trna25-AlaAGC (26859897-26859969) Ala (AGC) 73 bp Sc:
46.89    2
chr6.trna94-AlaAGC (58250620-58250548) Ala (AGC) 73 bp Sc:
54.62    4
chr6.trna160-AlaAGC (26881822-26881750) Ala (AGC) 73 bp Sc:
54.69    35

```

13

Perform Whole Transcriptome Mapping

This chapter contains:

■ Introduction to whole transcriptome mapping	181
■ Examples of running the whole transcriptome mapping module	181
■ Reads input files	182
■ Annotations input files	182
■ Reference input files	183
■ Mapping parameters	184
■ Map data	189
■ Mapping output files	190
■ Mapping statistics parameters	191
■ Summary of mapping statistics output	191
■ Mapping statistics output files	192
■ Whole transcriptome analysis output file formats	199

Introduction to whole transcriptome mapping

This chapter provides background information about secondary analysis (mapping and pairing, also known as aligning reads) and tertiary analysis (data visualization and analysis, including counting tags, determining coverage, and finding junctions). The information in this chapter includes descriptions of parameters you can configure to customize your analysis.

Examples of running the whole transcriptome mapping module

The following standard workflows provide examples of running the whole transcriptome analysis mapping module:

- Fragment whole transcriptome
- Pair-end whole transcriptome

Reads input files

The whole transcriptome analysis mapping module accepts as input one or more input XSQ files. The input reads can have different read lengths but they must be of the same library type. For example, paired-end read-sets of length 50 and 60 can be processed together. The XSQ files processed in a single analysis can not have different library types. In one mapping analysis, the input must be of only one library type, either fragment or paired-end.

Fragment mapping module supports mapping one tag of a paired-end read-set, but tertiary analysis does not run on BAM files generated from such processing.

If the data you want to process with LifeScope™ Software is in CSFASTA and QUAL files, these files must be converted to the XSQ file format. The LifeScope™ Software graphical user interface automatically converts these files to the XSQ format before mapping data.

RNA-Seq reads

The RNA-Seq reads used in whole transcriptome analysis are different from the genomic reads. For RNA-Seq reads:

- Only transcribed sequences are measured by the system.
- Genome coverage is non-uniform due to variation in transcriptional intensity.
- A large subset of reads originate from uninteresting sequences such as ribosomal RNA (rRNA).
- A subset of the reads originates from splice junctions and cannot align contiguously on the genome.

Multiple samples

LifeScope™ Software can perform alignment from multiple reads files (in XSQ format) for the same sample. If a sample is sequenced on the instrument in different runs or on different lanes, several XSQ files are generated for that sample. The reads from these XSQ files should be analyzed together. LifeScope™ Software facilitates the analysis by grouping the reads into read-sets.

Annotations input files

While genomic resequencing relies only on alignment to a reference sequence, whole transcriptome analysis also makes use of gene annotations. Gene annotations define the exons, genes, and transcripts used to improve alignment.

whole transcriptome analysis modules use genome annotation files in GTF format. Go to the following URLs to see the GTF format explained in detail:

genes.cse.wustl.edu/GTF2.html

genome.ucsc.edu/FAQ/FAQformat.html#format4

The GTF file must match the genome reference to ensure that the whole transcriptome analysis modules work correctly.

IMPORTANT! Make sure the GTF file is for the same genome assembly as the FASTA file, and that matching sequence identifiers are used. Gene and transcript identifiers in the GTF file must be properly normalized. Identifier normalization is a known issue in GTF files from the UCSC Genome Browser, which is a popular source of GTF-formatted annotation. The UCSC GTF files always report the same value for gene and transcript IDs.

Note: LifeScope™ Software includes a command line script (`refgene2gff.sh`) to transform a GTF file into the format required for use with whole transcriptome analysis pipelines. For additional information, see the *LifeScope™ Genomic Analysis Software Command Shell User Guide* (PN 4465697).

UCSC genome annotations

Format UCSC Genome Browser Database annotations for whole transcriptome analyses

The UCSC Genome Browser Database has genome annotations available for many assemblies at

hgdownload.cse.ucsc.edu/goldenPath/

The GTF-formatted annotations available for download are not properly normalized by gene ID. The required content is present for each assembly in the file export of the refGene database table `database/refGene/txt/gz`.

For example, annotation for human genome build 18 is available at:

hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz

Note: The GTF-formatted annotation is not in GTF format. You must convert the annotation before using it in whole transcriptome analysis.

Reference input files

WTA uses several files defining reference information: filter-reference, genome-reference and annotation files. The filter-reference file is a FASTA-formatted file defining uninteresting reference sequences such as Ribosomal RNA, tRNA, or vector sequences. To increase the speed of analysis, reads that align to sequences in the filter reference are excluded from downstream alignment steps. Many of the entries in the filter reference are species-specific. The filter reference file must be appropriate for the species being studied.

The principal output of alignment analysis is a set of alignments between the reads and the genome-reference in BAM format. The genome-reference file is a FASTA-formatted file defining sequences in the reference genome.

The annotation file is a GTF-formatted file defining known genes, transcripts, and exons in the genome reference. See [“Annotations input files” on page 182](#).

Junction-reference and exon-reference files are generated internally by the software. These files are entirely derived from the GTF and genome-reference files. A user is not required to provide these references. The junction-reference file is a FASTA-formatted file with an entry for every potential intragenic splice-junction defined in the annotation file. The entry contains the position of the junction and the associated flanking sequence. The exon-reference is a FASTA-formatted file with an entry for every exon defined in the annotation file.

Mapping parameters

There are three categories of mapping parameters:

Mapping Parameter	Used in analysis . . .
Fragment	ChIP-Seq
	Genomic resequencing
	MethylMiner™
	Targeted resequencing
	Whole transcriptome
Paired-end	Genomic resequencing
	MethylMiner™
	Targeted resequencing
	Whole transcriptome
Mate Pair	

Fragment Mapping

There are four categories of fragment mapping parameters: Main, Advanced, Whole Transcriptome, and BAMStats.

Main

Parameter	Default value	Description
Enable filter mapping	True	<ul style="list-style-type: none"> • True: Perform filter mapping. • False: Do not perform filter mapping.
Enable junction mapping	False	<ul style="list-style-type: none"> • True: Perform junction mapping. • False: Do not perform junction mapping.
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
Min junction overhang	8	Alignments to junctions are not reported if they have fewer than this number of bases unaligned on either side of the splice. Allowed values: Integers 0–100.
Make unmapped BAM files	0	Create BAM files containing unmapped reads. Values: <ul style="list-style-type: none"> • True: Create. • False: Do not create.
Mapping quality threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–255.
refcor base QV filter	10	Resulting base calls with quality less than this value will be replaced with "N." Allowed values: Integers 0–255.
Reference weight	8	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.

Whole Transcriptome

Parameter	Default value	Description
Read length	75	Position in the read at which the error rate inflates, for instance, 35–40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers 25–100.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	<code>\${analysis.output.dir}/fragment.mapping</code>	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	<code>\${task.output.dir}</code>	The path to the output directory where BAMStats will write its chart (.cht) files.

Maximum Coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10,000.
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 0–100000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers 1–100000
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0–1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers 1–100,000.

Paired-end

There are five categories of paired-end parameters: Main, Advanced, Whole Transcriptome Splice Junction Extractor, and BAMStats. There are no parameters for the Whole Transcriptome Exon Sequence Extractor analysis module.

Main

Parameter	Default value	Description
Enable filter mapping	True	<ul style="list-style-type: none"> • True: Perform filter mapping. • False: Do not perform filter mapping.
Enable junction mapping	True	<ul style="list-style-type: none"> • True: Perform junction mapping. • False: Do not perform junction mapping.
Enable exon mapping	True	<ul style="list-style-type: none"> • True: Perform exon mapping. • False: Do not perform exon mapping.
F3 rescue enabled	True	Attempt or do not attempt rescue of F3 reads. <ul style="list-style-type: none"> • True: Attempt rescue. • False: Do not attempt rescue.
F5 rescue enabled	True	Attempt or do not attempt rescue of F5 reads. <ul style="list-style-type: none"> • True: Attempt rescue. • False: Do not attempt rescue.

Parameter	Default value	Description
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Average insert size	120	The average the insert size. Used to compute the rescue distance: $\text{rescueDistance} = \text{avgInsertSize} + (3 * \text{stdInsertSize})$. Allowed values: Integers 0–100000.
Stdev insert size	60	Standard deviation of the insert size. Allowed values: Integers 1–100000.

Advanced

There are three categories of Advanced parameters: general, Rescue, and BAM Generation.

General

Parameter	Default value	Description
Min junction overhang	8	Alignments to junctions are not reported if they have fewer than this number of bases unaligned on either side of the splice. Allowed values: Integers 0–100.

Rescue

Parameter	Default value	Description
Max number of allowed mismatches for F3	8	Maximum number of mismatches allowed in a rescued F3 alignment. Allowed values: Integers 0–15.
Max number of allowed mismatches for F5	6	Maximum number of mismatches allowed in a rescued F5 alignment. Allowed values: Integers 0–10.

BAM Generation

Create unmapped BAM files	0	Create BAM files containing unmapped reads. Values: <ul style="list-style-type: none"> • True: Create. • False: Do not create.
Mapping quality threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–255.
refcor base QV filter	10	Resulting base calls with quality less than this value will be replaced with "N." Allowed values: Integers 0–255.
Reference weight	8	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.

Whole Transcriptome

Parameter	Default value	Description
Read length	75	Position in the read at which the error rate inflates, for instance, 35-40 for 50bp long reads. (If set to 0, then the value is equal to 0.8*readLength). Allowed values: Integers 25–100.


BAMStats

Parameter	Default value	Description
Input directory for BAMStats	`\${analysis.output.dir}/fragment.mapping`	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	`\${task.output.dir}`	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10000.

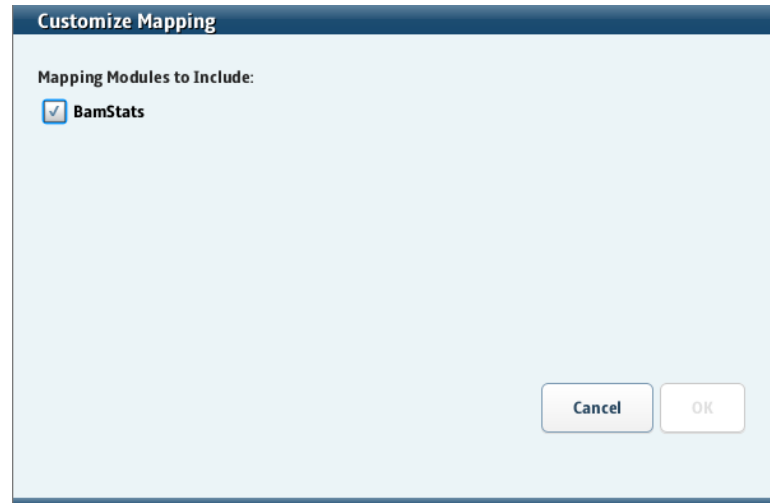
Parameter	Default value	Description
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 0–100000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers ≥ 1 .
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: <ul style="list-style-type: none">• 0: Do not combine data.• 1: Combine data.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none">• 0: Do not restrict coverage in WIG file format to only primary alignments.• 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1 .

Map data

Mapping data is automatic in LifeScope™ Software. To map data:

1. In the Projects organizer (shown [on page 59](#)), select a project.
2. Click an analysis icon , then click **Edit Analysis**.
3. In the Edit Analysis window, accept the default settings for mapping and pre-processing data in the Secondary Analysis section.

If you want to exclude BAMStats, click **Customize**. In the Customize Mapping window, uncheck BamStats, then click **OK**.



If you do not want to pre-process data, uncheck the box.

Mapping output files

Overview

The mapping module generates a BAM file containing alignments in coordinate order. For information about the BAM file, see [Appendix A, “File Format Descriptions and Data Uses”](#) on page 293.

The number of BAM files generated depends on the input XSQ files. One BAM file is generated per read-set (a read-set is defined as data from one barcode, in one XSQ file). If the input data is an XSQ file containing 96 barcodes, then the mapping module generates at least 96 output BAM files. Beads that are unclassified in any barcode are output into a separate additional BAM file.

The filenames for the output BAM files are created using information from the input XSQ file, including: file base name, file id, index name, and index id. The BAM files are named according to the following patterns:

- Non-indexed BAM files: *xsqname-fileID-1.bam*
- Indexed BAM files: *xsqname-fileID-idx_bcIndex-bcID.bam*

The fields in the filenames are:

- **xsqname** – The XSQ file base name, without the .xsq extension.
- **fileID** – The file ID for internal XSQ file tracking.
- **bcIndex** – The barcode index.
- **bcID** – The barcode identifier for internal barcode tracking.

The mapping output files are used by the BAMStats mapping statistics module and by the whole transcriptome tertiary analysis modules: whole transcriptome analysis counts, whole transcriptome analysis coverage, and whole transcriptome analysis splice junctions.

BAM file differences

This section describes two types of differences seen in WT BAM files, compared to other LifeScope™ Software BAM files.

Single-read optional fields

A BAM file produced by the WT paired-end pipeline is identical to that produced by the resequencing pipeline. However, the BAM format from the WT single-read pipeline differs from BAM files produced elsewhere in LifeScope™ Software. The single-read pipeline produces separate BAM files for filtered, unmapped, and mapped reads.

Mapping statistics parameters

For descriptions of mapping statistics parameters, see the BAMStats table [on page 188](#).

Summary of mapping statistics output

File formats

For every input BAM file, a set of statistics files are generated. These files are in CHT, CSV, TXT, and WIG formats. Each CHT file corresponds to one displayed chart. A CHT file specifies the type of chart, the displayed range of each axis, and the data points, without using external references.

The CHT file format is an internal file format based on the CSV file format, with addition header information included. CHT header information is the following:

```
# name:  
# type: scatter2d | pie | vbar | line  
# title:  
# xaxisname:  
# yaxisname:  
# xrange: <min>:<tickinterval>:<max>  
# yrange: <min>:<tickinterval>:<max>  
XAXISNAME, SERIES1NAME, SERIES2NAME, ...  
x1, y1.1, y1.2, ...  
x2, y2.1, y2.2, ...  
x3, y3.1, y3.2, ...
```

The wiggle format (.wig) is a public format typically used for coverage. Visit their site for more information:

hgdownload.cse.ucsc.edu/goldenPath/help/wiggle.html

A genome browser such as the Integrative Genomics Viewer (IGV) can be used to visualize the coverage. For information is available from their site:

www.broadinstitute.org/igv/

For a collection of input BAM files that belong to a sample, a set of cumulative statistics files are generated. The cumulative statistics files are also in CHT, CSV, TXT, and WIG formats. The cumulative statistics can be visualized in the LifeScope™ Software UI.

Summary information

A tab-separated summary file (*BAMfilename*-summary.tbl) is generated that summarizes key mapping quality statistics per input BAM file. This file contains one row for each input BAM file in the sample. The summary file is displayed in the LifeScope™ Software UI. See [“Summary file” on page 197](#).

Directory structure

The output of the BAMStats module has the following directory structure:

```
bamstats/
  <sampleName>/*.cht, *.tbl
  <sampleName>/<bam>/.*cht
  <sampleName>/<bam>/Misc/*.csv, *.txt, *.wig
  <sampleName>/Misc/*.csv, *.txt
```

The position error files and probe errors files are created in the BAM file directories.

Overview

The summary report contains a snapshot of statistics in all BAM files present in this sample. This report is displayed in the LifeScope™ Software UI.

```
<sampleName1>/BAMfilename-summary.tbl
```

Following directories and reports contain the cumulative statistics from all BAM files that belong to this sample. The Misc folder is not displayed in UI.

```
<sampleName1>/*.cht
<sampleName1>/Misc/*.csv, *.wig, *.txt
```

Following statistics are generated per BAM file in the mapping directory. These reports are not displayed in the UI.

```
<sampleName1>/<BAMfilename>/.*cht
<sampleName1>/<BAMfilename>/Misc/*.csv, *.txt, *.wig
```

In a sample, some of the BAM files possibly represent unhealthy DNA or RNA, causing the cumulative statistics to look poor. The summary tbl file is the unified location for examining the quality of data of all BAM files in the sample.

If the data for a particular BAM file is not as expected, look at the directory-level reports for that BAM file, for details.

Mapping statistics output files

This section describes the mapping statistics files generated by the BAMStats module. When statistics reports are separated by tag type, the report's file name includes the tag in the file name. The tags used in file names are:

Table 1 Tags in mapping statistics output file names

Library type	Tag
Fragment	F3

Table 1 Tags in mapping statistics output file names

Library type	Tag
Mate-pair	F3
	R3
Paired-end	F3
	F5-P2

The output files generated by BAMStats include the name of the BAM file. The file name pattern is:

BAMFileName-fileID-barcodeID.StatisticsReportName.tag.extension

This string is referred to by *prefix* in the following table, the mapping statistics output table. In the description of mapping statistics output file names in the following table, the string *tag* is used to refer to the tags in the following table.

Report	Description, axis information, filename
Alignment Length Distribution	
	<p>A bar plot giving the distribution of alignment lengths found in various tags used in alignment. The report is separated by tag type to provide visibility into the accuracy of individual tags.</p> <p>Only the primary alignment for each bead is considered in calculating the distribution.</p>
	<p>Y axis: Frequency X axis: Alignment length (from 0 to the maximum read length)</p>
	<p>Output file name: <i>prefix</i>.Alignment.Length.Distribution.<i>tag</i>.cht</p>
Alignment Length Distribution for Unique Alignments	
	<p>The same report as the Alignment Length Distribution, except that only tags that have a primary alignment are considered in calculating the distribution.</p> <p>By default BAM files contain primary alignments only. In this case, the reports AlignmentLengthDistribution and AlignmentLengthDistribution for Unique Alignments are identical. However, users can also print all alignments or secondary alignments in BAM file. In this case, the two distributions are different.</p> <p>Note: The title Unique Alignments is interpreted to mean primary alignments, for this report.</p>
	<p>Y axis: Frequency X axis: Alignment length (from 0 to the maximum read length)</p>
	<p>Output file name: <i>prefix</i>.Alignment.Length.Distribution.Unique.<i>tag</i>.cht</p>
Base Mismatch Distribution	
	<p>A bar plot giving the distribution of total number of mismatches found in various tags used in alignment. The bins are summed over all alignment lengths.</p> <p>Only the primary alignment for each bead is considered in calculating the distribution.</p>
	<p>Y axis: Frequency X axis: Alignment length (from 0 to the maximum mismatches allowed)</p>
	<p>Output file name: <i>prefix</i>.Mismatch.Distribution.<i>tag</i>cht</p>
Base Mismatch Distribution for Unique Alignments	

Report	Description, axis information, filename
	<p>The same report as the Mismatch Distribution, except that only tags that have a unique alignment are considered in calculating the distribution.</p> <p>Note: The title Unique Alignments is interpreted to mean primary alignments, for this report.</p> <p>Y axis: Frequency X axis: Alignment length (from 0 to the maximum mismatches allowed)</p> <p>Output file name: <i>prefix.Mismatch.Distribution.Unique.tag.cht</i></p>
Distribution of Alignment Length and Number of Mismatches in Tags	
	<p>A report providing a simultaneous picture of the alignment length and number of mismatches distributions in various tags used in alignment. Only primary alignments for each tag are considered in calculating this distribution. The bins range from 0 to the maximum read length, and from 0 to the maximum number of mismatches allowed.</p> <p>This report is located in the <code>Misc</code> folder, and is not displayed in the UI.</p> <p>Z axis: Frequency Y axis: Alignment length (from 0 to the maximum mismatches allowed) X axis: Number of mismatches (from 0 to the maximum mismatches allowed)</p> <p>Output file name: <i>prefix.AlignmentLength.Mismatch.tag.csv</i></p>
Base QV Distribution	
	<p>A bar plot providing a distribution of base quality values generated using the reference assisted error correction/base conversion algorithm. The base QV distributions are separated for each tag as quality of individual tags could be very different.</p> <p>Y axis: Frequency X axis: Base QV (from 0 to the maximum QV)</p> <p>Output file name: <i>prefix.BaseQV.tag.cht</i></p>
Base QVs by Position	
	<p>A report providing a distribution of base quality values by individual base positions. This report identifies if certain base positions (particularly towards the end of the read) have poor base quality values. All the tags (F3/R3/F5-P2) used in the alignment are combined into a single bin.</p> <p>Y axis: Base QV X axis: Base position (from 0 to read length)</p> <p>Output file name: <i>prefix.BaseQV.by.Position.csv</i></p>
Distribution of Mismatches by Base QV	
	<p>A 3D surface plot providing a distribution of errors (mismatches to reference) by base quality values bins. This report measures whether the base QVs generated are well calibrated to the probability of error in that particular base position.</p> <p>Y axis: Error rate X axis: Base QV (from 0 to maximum QV)</p> <p>Output file name: <i>prefix.Mismatches.By.BaseQV.tag.cht</i></p>
Distribution of Mismatches by Position	
	<p>A bar plot providing a distribution of errors (mismatches to reference) by position within the read. Only the primary alignments for each bead are used to generate this distribution.</p> <p>Y axis: Frequency X axis: Position (from 0 to maximum read length)</p> <p>Output file name: <i>prefix.Mismatches.By.Position.tag.cht</i></p>

Report	Description, axis information, filename
Distribution by Mapping QVs by Tag	
	A bar plot providing a distribution of mapping quality values for individual tags (F3/R3/F5-P2). Only the primary alignment for each bead is used in calculating this distribution.
	Y axis: Frequency X axis: Position (from 0 to maximum mapping QV)
	Output file name: <i>prefix.MappingQV.tag.cht</i>
Distribution by Pairing QVs	
	A bar plot providing a distribution of pairing quality values for individual tags (F3/R3/F5-P2). Only the primary alignment for each bead is used in calculating this distribution.
	Y axis: Frequency X axis: Pairing QV (from 0 to maximum pairing QV)
	Output file name: <i>prefix.PairingQVs.tag.cht</i>
Coverage Report	
	A line plot providing a distribution of coverage obtained after mapping/pairing. Only the primary alignment for each bead is used in this calculation.
	Y axis: Number of bases X axis: Coverage (from 0 to number of reads)
	Output file name: <i>prefix.Coverage.tag.cht</i>
Coverage Report by Chromosome (Contig) and Base Windows	
	A line plot providing a distribution of coverage within each reference window. The coverage is calculated within each window along a reference chromosome. The window size can theoretically be anywhere from a single base to the contig length. However the calculations of base level coverage are computationally expensive and less interpretable as the window size increases. Only the primary alignment for each bead is used in this calculation.
	Y axis: Coverage (from 0 to number of reads) X axis: Contig and window number
	Output file name: <i>prefix.Coverage.By.Chromosome.contig.n.tag.cht</i>
Coverage by Strand	
	A line plot providing the distribution of coverage within each reference window separated by reference strand (+/-). Only the primary alignment for each bead is used in this calculation.
	Y axis: Number of bases X axis: Coverage (from 0 to number of reads)
	Output file name: <i>prefix.Coverage.By.Strand.tag.cht</i>
Coverage files	

Report	Description, axis information, filename
	<p>Coverage reports in wiggle format. For each genome position, reports the number of reads that cover (map to or span) the position. Because reporting coverage for each position results in very large files, coverage is reported for bins, with each bin spanning a user-defined number of bases. For each bin, the mean coverage of all the positions in that bin is reported. The parameter <code>bamstats.wig.binsize</code> controls the size of the bins in this file.</p> <p>By default one coverage file is generated, per chromosome, per strand. If the parameter <code>bamstats.combined.report.both.strands</code> is set to true, then one file per chromosome is generated, combining coverage from both strands into one file per chromosome.</p> <p>Each coverage file includes a header as the first line. The header lines follow this pattern: <code>track type=wiggle_0 name=<chrname> description=<coverage from positive/negative/both strand></code> <code>visibility=full color=0,0,255 fixedStep chrom=<chrname> start=<startpos> step=<binsize> span=<binsize></code></p> <p>Output file name: <code>coverage_chrnn.POS.wig</code>, <code>coverage_chrnn.NEG.wig</code></p>
Insert Range Distribution	
	<p>A line plot providing the distribution of insert sizes for paired data (PE and LMP library types).</p> <p>Y axis: Frequency X axis: Insert Range bins</p> <p>Output file name: <code>prefix.Insert.Range.Distribution.cht</code></p>
Distribution of Read Pair Types	
	<p>For paired data, this report calculates the distribution of read pair types (AAA, AAB, C**, etc.).</p> <p>Y axis: Frequency X axis: Read pair type</p> <p>Output file name: <code>prefix.ReadPair.Type.cht</code></p>
Pairing Statistics	
	<p>A pie chart showing the percentages of the following F3, R3 (or F5) combinations, for paired data.:</p> <ul style="list-style-type: none"> • Mapped, Mapped • Mapped, Unmapped • Unmapped, Mapped • Unmapped, Unmapped • Mapped, Missing • Missing, Mapped • Unmapped, Missing • Missing, Unmapped <p>Output file name: <code>prefix.Pairing.Stats.cht</code></p>
Unique Start Position	

Report	Description, axis information, filename
	<p>A text report with the following statistics about the start position on the genome, based only on primary alignments. The report contains:</p> <ol style="list-style-type: none"> 1. The number of starting points in uniquely placed tags: Reports the positions in the reference with at least one uniquely placed alignment starting at that position. Unique here does not mean primary. Also reports the percentage of this number within the total number of positions in the reference. 2. The average number of uniquely mapped reads per starting point: Reports the total number of unique alignments divided by the number of starting points in uniquely placed tags. 3. An estimated number of starting points for all mapped tags: Reports the total number of primary alignments divided by the number of starting points in uniquely placed tags.
	Output file name: <i>prefix.Unique.Start.Positions.txt</i>

Example of mapping statistics output

This section provides example output of some mapping statistics output files.

Summary file

The summary file provides statistics for each BAM file in the sample. This list describes labels used in the summary file:

- **NumUnFilteredBeads** – The total number of beads, before any filtering on the instrument. This value is also the fragment count in the input XSQ file.
- **NumFilteredBeads** – The number of beads that pass filtering on the instrument.
- **NumMapped** – The number of reads with primary alignment.
- **% filtered that mapped** – The number of primary reads divided by the number of beads passing instrument filtering ($\text{NumMapped} / \text{NumFilteredBeads}$).
- **% total that mapped** – The number of primary reads divided by the number of total beads ($\text{NumMapped} / \text{NumUnFilteredBeads}$).

The number of beads filtered out in the instrument is found by subtracting the number of beads that pass filtering from the total number of beads:

$$\text{NumUnFilteredBeads} - \text{NumFilteredBeads}$$

The following is example contents for a summary file:

```
#title: BAMStats Summary
BamFileName, IsColorInBam, IsBaseInXSQ, IsECC, LibraryType, ReadLength, PredictedInsertSize, NumFilteredBeads, NumUnFilteredBeads, Tag1-NumMapped, Tag1- % total Mapped, Tag1- % filtered mapped, Tag2-NumMapped, Tag2- % total Mapped, Tag2- % filtered mapped, (Tag1-AlignmentLength;Min;Max;Avg;Median;StdDev), (Tag1-NumMismatches;Min;Max;Avg;Median;StdDev), (Tag1-MappingQV;Min;Max;Avg;Median;StdDev), (Tag1-BaseQV;Min;Max;Avg;Median;StdDev), (Tag2-AlignmentLength;Min;Max;Avg;Median;StdDev), (Tag2-NumMismatches;Min;Max;Avg;Median;StdDev), (Tag2-MappingQV;Min;Max;Avg;Median;StdDev), (Tag2-BaseQV;Min;Max;Avg;Median;StdDev), (Coverage;Min;Max;Avg;Median;StdDev)
```

Unique Start Position

The following is an example of the output of a Unique Start Position report:

```
#
Starting Points within Placed Tags
Number of Starting Points in Uniquely placed tag 109,068,047
(2.005740% of reference)
Average Number of Uniquely Mapped reads per Start Point
1.992770
Estimate number of starting point for all mapped tags
152,843,627 (2.810765% of reference)
```

Coverage files

Example contents for the file `coverage_chr1_positive.wig` are:

```
browser position chr1:1-200000000
browser hide all
browser pack refGene encodeRegions
# minimumMapq=25, minimumCoverage=1,
alignmentFilteringMode=PRIMARY, filterOrphanedMates=false,
track name="BAM Coverage positive strand" description="BAM
Coverage positive strand" visibility="full color 0,0,255"
priority=10 yLineMark=0 type=wiggle_0 yLineOnOff=on
variableStep chrom=chr1 span=1
336171
336181
336191
336201
336211
336221
336231
336241
336251
336261
336271
...
```

Example contents for the file `coverage_chr1_negative.wig` are:

```
browser position chr1:1-200000000
browser hide all
browser pack refGene encodeRegions
# minimumMapq=25, minimumCoverage=1,
alignmentFilteringMode=PRIMARY, filterOrphanedMates=false,
track name="BAM Coverage negative strand" description="BAM
Coverage negative strand" visibility="full color 0,0,255"
priority=10 yLineMark=0 type=wiggle_0 yLineOnOff=on
```

```
variableStep chrom=chr1 span=1
57432
57442
57452
57463
57473
57483
57493
57503
57513
57523
57533
...
```

Whole transcriptome analysis output file formats

Alignment report

Whole transcriptome analysis mapping runs produce an alignment report, the alignmentReport.txt file in the results directory.

Below is an example of the alignment report produced by the whole transcriptome analysis pipeline (the frequency lists have been truncated).

The 100.0% value reported for the “Reads mapped, not filtered” line should be interpreted as “Reads mapped, not filtered” comprising 100% of the category described in its section, not that the category “Reads mapped, not filtered” comprises 100% of the total reads.

```
-----
ALIGNMENT REPORT
-----

Counts:
Total reads:                102,837,343  (100.0%)
Reads mapped:               78,065,351  ( 75.9%)
Reads filtered:             12,029,269  ( 11.7%)
-----
Reads mapped, not filtered  68,780,034  (100.0%)
Reads with too many mappings (N >= 10):  5,897,833  (  8.6%)
Reads with number of mappings in proper range (N < 10):
62,882,201    ( 91.4%)
Reads uniquely aligned (score.clear.zone = 3):
53,417,152    ( 77.7%)
Reads uniquely aligned to junctions      4,816,207  (  7.0%)
Reads uniquely aligned to known junctions 4,773,948  (  6.9%)

ABSOLUTE FREQUENCIES
Alignment_Score    All_Read_Alignments_(N=_936)
Primary_Read_Alignments_(N=_935)
Unique_Read_Alignments_(N=_866)
0.0      0      0      0
```

1.0	0	0	0
2.0	0	0	0
36.0	0	0	0
37.0	87	87	87
38.0	0	0	0
39.0	0	0	0
47.0	0	0	0
48.0	0	0	0
49.0	849	848	779
50.0	0	0	0
Total	936	935	866

RELATIVE FREQUENCIES

Alignment_Score	All_Read_Alignments_(N=_936)		
Primary_Read_Alignments_(N=_935)	Unique_Read_Alignments_(N=_866)		
0.0	0.0%	0.0%	0.0%
1.0	0.0%	0.0%	0.0%
2.0	0.0%	0.0%	0.0%
36.0	0.0%	0.0%	0.0%
37.0	9.3%	9.3%	10.0%
38.0	0.0%	0.0%	0.0%
39.0	0.0%	0.0%	0.0%
40.0	0.0%	0.0%	0.0%
47.0	0.0%	0.0%	0.0%
48.0	0.0%	0.0%	0.0%
49.0	90.7%	90.7%	90.0%
50.0	0.0%	0.0%	0.0%
Total	100.0%	100.0%	100.0%

Whole transcriptome analysis filtering stats

An example whole transcriptome human filter reference is provided under the `<examplesdir>/demos/WholeTranscriptome/references` folder in the optional LifeScope™ Software examples distribution. This example reference fasta includes contigs from barcode primers, human ribosomal RNAs, tRNAs, and other known targets for filtering. The filter reference may be expanded or removed with more csfasta reference records at the discretion of the user.

The stats for filtered reads are generated after a LifeScope™ Software run. First two lines of the stats file states the number of reads processed, and the number of reads mapped to the filtered reference followed by its percentage. Each following line reports a contig name and the count of reads that aligned to that contig. Paired-end filtering stats may be found in the Intermediate folder. The following is a truncated example of filtering stats output:

```
countOfReadsProcessed: 175,839,941
countOfReadsMapping:   18,783,220 (10.7%)
...
Barcode-047-3-end reverse   83
Barcode-048-3-end reverse  2,350
gi|124517659|ref|NR_003286.1| Homo sapiens 18S ribosomal RNA
(LOC100008588)   5,175,899
gi|142372596|ref|NR_003285.2| Homo sapiens 5.8S ribosomal RNA
(LOC100008587)   78,936
```



```
gi|124517661|ref|NR_003287.1| Homo sapiens 28S ribosomal RNA  
(LOC100008589) 7,616,212  
chr6.trna95-AlaAGC (58249908-58249836) Ala (AGC) 73 bp Sc:  
42.26 5  
chr6.trna25-AlaAGC (26859897-26859969) Ala (AGC) 73 bp Sc:  
46.89 2  
chr6.trna94-AlaAGC (58250620-58250548) Ala (AGC) 73 bp Sc:  
54.62 4  
chr6.trna160-AlaAGC (26881822-26881750) Ala (AGC) 73 bp Sc:  
54.69 35
```


14

Perform MethylMiner™ Mapping

This chapter covers:

■ Introduction to MethylMiner™ mapping	203
■ MethylMiner™ mapping library types	204
■ MethylMiner™ mapping input files.	204
■ MethylMiner™ mapping parameters	205
■ Perform MethylMiner™ mapping	207
■ View MethylMiner™ mapping results	211
■ MethylMiner™ mapping output files	211
■ Further analysis of MethylMiner™ mapping results.	212

Introduction to MethylMiner™ mapping

DNA methylation is an epigenetic modification crucial for organism development and normal gene regulation. Life Technologies have introduced a versatile methyl-CpG binding protein-based system, the MethylMiner™ Kit, for the enrichment of methylated sequences from genomic DNA, that, with the use of SOLiD™ System sequencing, allows for focused evaluation of methylation patterns in genome-wide studies. The enriched reads can also be bisulfite-converted, which would additionally allow determination of methylation status of individual cytosines.

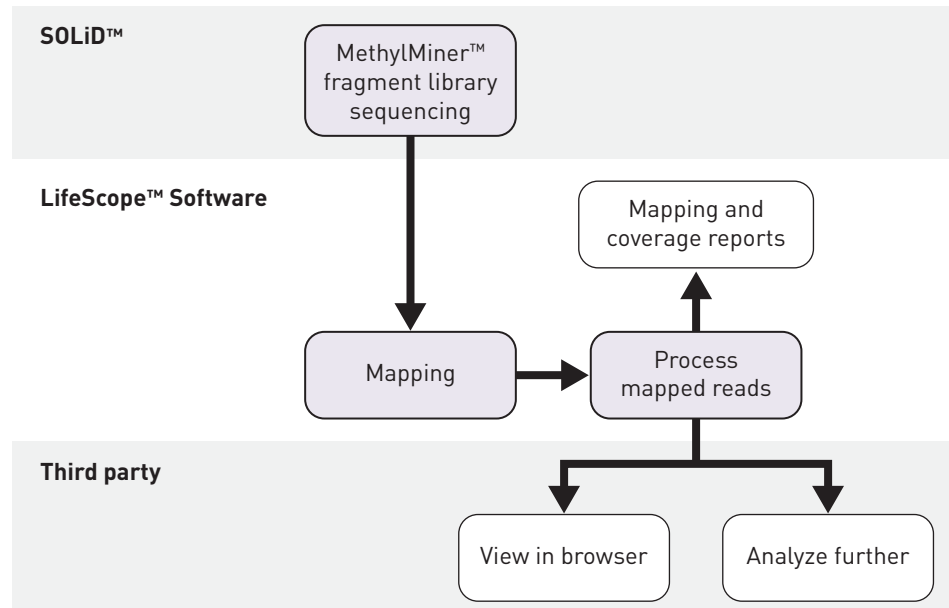
SOLiD System sequencing can also, of course, be used to study methylation in unenriched, whole-genome bisulfite-converted read data.

The purpose of LifeScope™ Genomic Analysis Software MethylMiner™ tools is to provide a data analysis workflow for mapping and analyzing MethylMiner™ enriched and unenriched fractions of genomic DNA as well as bisulfite-converted and unconverted reads sequenced on the SOLiD™ System.

The MethylMiner™ analysis workflow functionality currently includes:

- Mapping of unconverted and bisulfite-converted reads.
- Mapping statistics and statistics on read coverage and depth.
- Mapped reads output in BAM-format files.
- Visualization of mapped reads on publicly available genome browsers.

The following illustration shows the MethylMiner™ analysis workflow with SOLiD™ software and third-party browsers.



MethylMiner™ mapping library types

Genomic resequencing analysis supports data from the following library types:

- Fragment
- Paired-end

MethylMiner™ mapping input files


Input File	File Extension
(Optional) Color space FASTA	*.csfasta
FASTA	*.fa, *.fasta
(Optional) Quality Value	*qual, *qv
Extensible sequence	*.xsq

MethylMiner™ mapping parameters

General parameters

General parameters include:

Parameter	Default value	Description
analysis.assembly.name	—	Default: unknown
analysis.regions.file	—	Default: unknown

Use the  button to open the File Chooser and search for input files.

Fragment mapping parameters

There are three categories of parameters for mapping: Main, Advanced, and BAMStats

Main

Parameter	Default value	Description
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	0	Replace resulting base-calls with a quality value less than the specified value.
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: <ul style="list-style-type: none"> True: Create. False: Do not create.
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0 to 1,000.

Parameter	Default value	Description
Reference weight	15	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0 - 100.
Second map gapped algorithm type	GLOBAL	Do indel finding, and control the behavior of indel finding. Allowed values: <ul style="list-style-type: none"> • NONE: Turn off indel finding. • GLOBAL: Mapping reports global alignment up to one indel. • LOCAL: Mapping reports local alignment up to one indel.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	`\${analysis.output.dir}/bam	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	`\${task.output.dir}	The path to the output directory where BamStats will write its chart (.cht) files.
Maximum Coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 1 - 10,000.
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 1 - 100000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers >= 0-1.
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0-1.

Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1-100000

Perform MethylMiner™ mapping

Unconverted MethylMiner™ reads must be mapped to a regular (unconverted) reference genome sequence. Bisulfite-converted reads must be mapped to a pair of appropriately converted reference sequences (forward and reverse conversions), as recommended below.

Follow these recommendations for MethylMiner™ mapping:

- MethylMiner™ supports fragment and paired-end libraries. Do not use this module with mate-pair libraries.
- For mapping bisulfite reads, the converted reference sequence pairs below are recommended. Use one of the following:
 - Pair 1:
Reference with all non-CpG C's converted to T's
Reference with all non-CpG G's converted to A's
 - Pair 2:
Reference with all C's converted to T's
Reference with all G's converted to A's

Optional) Import data

You can optionally import data to be analyzed. For instructions on how to import data, see “Import Data” [on page 67](#).

Log in to LifeScope™ Software

1. Navigate to LifeScope™ Software at a designated url, given to you by your network administrator.
`http://<IP address>:<port number>/LifeScope.html`
where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown [on page 59](#)).

Create or select a project

1. In the home view (shown [on page 59](#)), either click **Create a New Project** (described [on page 66](#)) or select a project in the Projects organizer (shown [on page 59](#)).
2. If you create a project:
 - a. Type a name and description in the Enter Project Name view.

Note: The name cannot have spaces or special characters.

- b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown [on page 59](#)), click **Add Data to Project** to choose a data type.

Add data to a project

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the read-sets you want to map and analyze.
If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).
3. (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.
To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.
To remove a data file from a group, click the checkbox of the data file, then click **Delete**.
4. Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

1. In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.
2. In the Choose Analysis view, enter a name for the analysis and a description.
If you are re-using an Old Analysis, select **Reuse Old Analysis** and select the name of the analysis you want to use.
Note: The name cannot have spaces or special characters.
3. Select **MethylMiner™**, then click **Next** to choose references.

To see the type of data mapped by the MethylMiner™ module, place your cursor over the .

4. In the Data To Be Analyzed table, click **Select the reference for the reads** to open the repository file browser.
5. In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:
data ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta
Select the file, then click **OK**.
The file name appears in the Reference column.
Note: To change the reference file, click on **Select the reference for the reads**, then select another reference file.
6. After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from choosing references and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

In the Choose Modules view of the Edit Analysis window, accept the default settings for secondary and tertiary analyses, then click **Next** to set module parameters.

Set module parameters

This section describes the procedures for setting general parameters and fragment mapping parameters. For descriptions of module parameters, see [“MethylMiner™ mapping parameters” on page 205](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a  button.

Set general parameters


1. Enter an analysis assembly name, if necessary.
2. Click **Next** to set fragment mapping parameters.

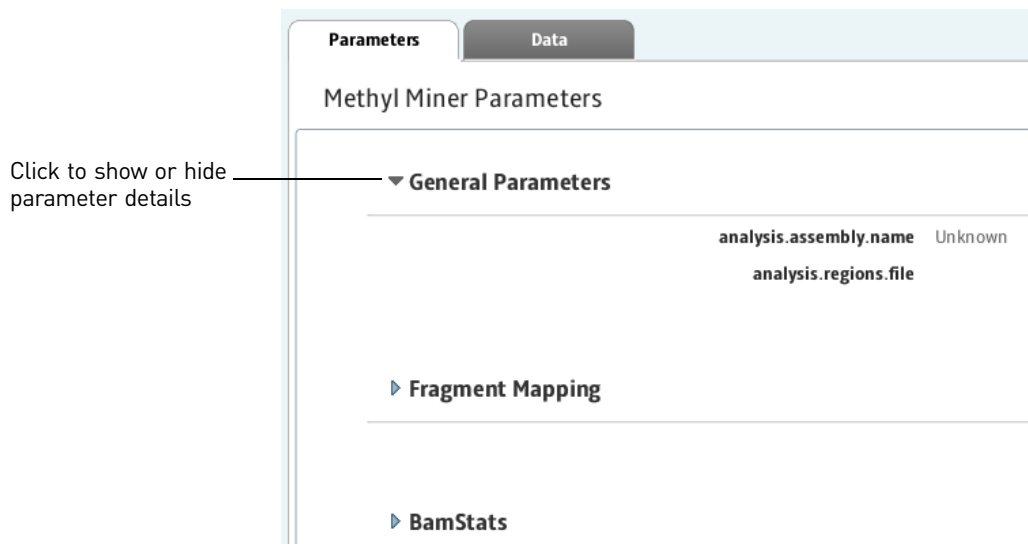
Set fragment mapping parameters

1. There are three categories of fragment mapping parameters: Main, Advanced, and BAMStats. Accept the default settings or click the Main, Advanced, and BAMStats tabs to edit the settings.
2. After you have edited an analysis:
 - Save your edits and proceed by clicking **Review**, or
 - Save your edits and close the window by clicking **Finish**, or
 - Erase your edits and close the Edit Analysis window by clicking **Cancel**.

Review and run the analysis

Note: For a description of MethylMiner™ parameters, see [“MethylMiner™ mapping parameters” on page 205](#).

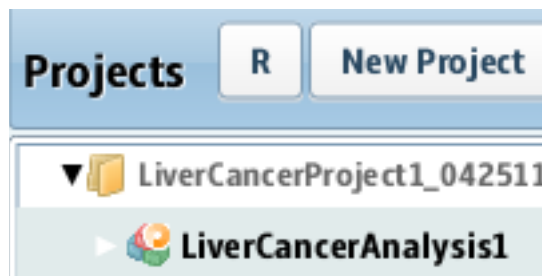
The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  next to the parameter categories to show or hide the parameters.



1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**. The window closes.

Check Analysis Status

1. In the Projects organizer (shown [on page 59](#)), click the project to check the status of data mapping.



2. Click the analysis name to show and show details about the analysis in the status overview (shown [on page 59](#)).
3. Click the Status tab to view the Progress column, which shows the percentage of completion for data mapping.

LiverCancerProject1... • LiverCancerAnalysis1

LiverCancerProject1... • LiverCancerAnalysis1						
Overview	Status	Parameters				
Analysis Runs						
Name	XSQ ID	Analysis	Secondary Analysis	Secondary Progress	Tertiary Analysis	Tertiary Progress
DH10B_Test4_None	DH10B_Test4.xsq	LiverCancerAnalysis1	[Fragment Mapping, BamS...	9%		

After MethylMiner™ mapping has been completed, you can view mapping results.

View MethylMiner™ mapping results

MethylMiner™ mapping results include additional files and log files.

LiverCancerProject1_042511 • LiverCancerAnalysis1 • Mapping • solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary

Details Statistics **Additional Files** Logs

Additional Files:

Directory	File	Size	Type
fragment.mapping/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam	646208 KB	bam
fragment.mapping/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam.bai	90 KB	bai

LiverCancerProject1_042511 • LiverCancerAnalysis1 • Mapping • solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary

Details Statistics Additional Files **Logs**

Log Files:

Directory	File	Size
fragment.mapping	refcor.secondary-Homo_sapiens.NCB136.54.dna.chr22-fragment.mapping.1.run.log	10 KB
fragment.mapping	refcor.secondary-Homo_sapiens.NCB136.54.dna.chr22-fragment.mapping.10.run.log	10 KB
fragment.mapping	refcor.secondary-Homo_sapiens.NCB136.54.dna.chr22-fragment.mapping.9.run.log	9 KB
fragment.mapping	secondary-Homo_sapiens.NCB136.54.dna.chr22-fragment.mapping.4.run.20110425193421974.log	28 KB
fragment.mapping	refcor.secondary-Homo_sapiens.NCB136.54.dna.chr22-fragment.mapping.2.run.log	10 KB

Refer to [Chapter 8, “View Analysis Results” on page 87](#) for instructions on how to review mapping results.

MethylMiner™ mapping output files

The output of MethylMiner™ mapping is BAM files.

The output mapping statistics reports generated by MethylMiner™ mapping include:

- The number and percent of all reads mapped and of reads mapped uniquely, in the file `mapping-stats.txt` and also in files ending in `F3.stats` and `F5-P2.stats`.
- Genome coverage tables for at least read depths 0X to 50X, in output files ending in `coverage-histogram-F3.txt` and `coverage-histogram-F5-P2.txt`.

View output files in a genome browser

Output files generated by mapping runs are compatible with third-party browser such as the Integrative Genomics Viewer (IGV) available from the Broad Institute and the UCSC Genome Browser.

Further analysis of MethylMiner™ mapping results

Besides viewing mapped reads in genome browsers, you can further analyze mapped reads outside of LifeScope™ Genomic Analysis Software with software available in the SOLiD™ development community or with other third-party tools.

MethylMiner™ unconverted mapped reads can be processed with peak-finding programs like MACS to identify genome regions of significant methylation.

Similarly, MethylMiner™ bisulfite-converted mapped reads can be processed with peak-finding programs to identify regions of significant methylation. These reads can also be processed at nucleotide resolution to report the methylation status of individual C bases, for bases covered at sufficient read depth.

15

Perform ChIP-Seq Mapping

This chapter covers:

■ Introduction to ChIP-Seq mapping	213
■ ChIP-Seq library types	213
■ ChIP-Seq input files	214
■ ChIP-Seq parameters	214
■ Perform ChIP-Seq mapping	216
■ View ChIP-Seq mapping results	220
■ ChIP-Seq output files	220

Introduction to ChIP-Seq mapping

LifeScope™ Genomic Analysis Software provides the ability to map data and create an output file type compatible with a variety of third-party Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data analysis tools. The ChIP-Seq application has publicly available analysis software that can be used with LifeScope™ Software output.

Examples of such software tools include:

- SAMTools and BEDTools for generic manipulation of SAM/BAM and BED files, respectively
- MACS, QuEST, and USeq for peak-finding
- browsers like the UCSC genome browser, or IGV for viewing mapped reads and peaks relative to other genomic annotations

The ChIP assay is a method for analyzing epigenetic modifications and genomic DNA sequences bound to specific regulatory proteins. ChIP-Seq is a combined assay and sequencing technique for identifying and characterizing elements in protein-DNA interactions. It typically examines transcription factors (TF) bound to DNA and finds DNA sequence motifs common to binding sites.

Using the MAGnify™ ChIP-Seq kit with the SOLiD™ sequencing system enables you to generate sequence read data from a ChIP-Seq experimental approach. LifeScope™ Software gives you the option to map the read data.

ChIP-Seq library types

ChIP-Seq mapping supports data from the fragment library type.

ChIP-Seq input files

Input File	File Extension
Color space FASTA (optional)	*.csfasta
FAST-All (FASTA)	*.fa, *.fasta
Quality Value (optional)	*qual, *qv
Extensible sequence	*.xsq


ChIP-Seq parameters

Parameters for ChIP-Seq mapping include general parameters and fragment mapping.

General

General parameters include:

Parameter	Default value	Description
analysis.assembly.name	hg19	The name of the genome assembly used in current analysis. Examples are hg18 and hg19.
analysis.regions.file	/data/results/referenceData/internal/hg19/targetedEnrichment	The path to the file containing genomic regions used in targeted resequencing selection, such as a .bed format file containing exome targets.

Use the  button to open the File Chooser and search for input files.

Fragment mapping

There are three categories of parameters for mapping: Main, Advanced, and BAMStats.

Main

Parameter	Default value	Description
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	0	Replace resulting base-calls with a quality value less than the specified value.
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: <ul style="list-style-type: none"> True: Create. False: Do not create.
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0 to 1,000.
Reference weight	15	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0 - 100.
Second map gapped algorithm type	GLOBAL	Do indel finding, and control the behavior of indel finding. Allowed values: <ul style="list-style-type: none"> • NONE: Turn off indel finding. • GLOBAL: Mapping reports global alignment up to one indel. • LOCAL: Mapping reports local alignment up to one indel.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	`\${analysis.output.dir}/fragment.mapping`	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	`\${task.output.dir}`	The path to the output directory where BamStats will write its chart (.cht) files.

Maximum Coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 1 - 10,000.
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 1 - 100000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers >= 0-1.
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0-1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1-100000

Mark Duplicates

Parameter	Default value	Description
Input directory with structured mapping output	<code>\${analysis.output.dir}/fragment.mapping</code>	Folder with one group per sequencing indexing run, and named [Group]/[File]-[number].bam.

Perform ChIP-Seq mapping

(Optional) Import data

You can optionally import data to be analyzed. For instructions on how to import data, see "Import Data" on page 67.

Log in to LifeScope™ Software

1. Navigate to LifeScope™ Software at `http://<IP address>:<port number>/LifeScope.html` where *IP address* is the address of the system or head node and *port number* is the number of the port used by the server.
2. In the Login screen, enter your username and password, then either click **Login** or press **Enter** to open the LifeScope™ Software home view (shown on page 59).

Create or select a project

1. In the home view (shown [on page 59](#)), either click **Create a New Project** (described [on page 66](#)) or select a project in the Projects organizer (shown [on page 66](#)).
2. If you create a project:
 - a. Type a name and description in the Enter Project Name view.
Note: The name cannot have spaces or special characters.
 - b. Click **Create New Project**.
 - c. In the Projects lists, select the new project.
3. In the Task Wizards section (shown [on page 59](#)), click **Add Data to Project** to choose a data type.

Add data to a project

Add data from the read repository to a project by choosing a data type and finding data. You can also optionally group data.

Note: Multiple files cannot have the same name. Make sure that the names of the files that you add are distinct.

Note: If the LifeScope™ Software administrator has changed the path to the read repository since your last login, restart LifeScope™ Software to update the repository to the new path.

1. In the Add Data to Project window, select **Raw unmapped (XSQ) data**, then click **Next** to find data.
2. In the Read Repository Filter table, select the read-sets you want to map and analyze.
If you want to group data, click **Next**. If you do not want to group data, skip [step 3](#) and proceed to [step 4](#).
3. (Optional) In the Read-sets in Project table, click the checkbox of the data files you want to group, then click **Add Group to Project**. The files appear in the Groups in Project table.
To rename a group, click the checkbox of the group in the Groups in Project table, then click **Edit**. In the Edit Group window, enter a new name.
To remove a data file from a group, click the checkbox of the data file, then click **Delete**.
4. Click **Add Analysis** to proceed, or
 - Click **Cancel** to refrain from adding data and close the Add Data to Project window, or
 - Click **Finish** to add the data and close the Add Data to Project window.

Create an analysis

1. In the Choose Data view of the Create Analysis window, select data files from the Available Data in Project table, then click **Next** to choose an analysis.
2. In the Choose Analysis view, enter a name for the analysis and a description.
If you are re-using an Old Analysis, select **Reuse Old Analysis** and select the name of the analysis you want to use.
Note: The name cannot have spaces or special characters.

3. Select **ChIP-Seq**, then click **Next** to choose references.

To see the type of data mapped by the ChIP-Seq module, place your cursor over the .

4. In the Data To Be Analyzed table, click **Select the reference for the reads** to open the repository file browser.
5. In the Browse for Reference File window, navigate to the location of the reference genome of your sample. Open the folder with your .fasta file, for example:
 data ▶ referenceData ▶ lifetech ▶ hg18 ▶ reference ▶ human_hg18.fasta
 Select the file, then click **OK**.

The file name appears in the Reference column.

Note: To change the reference file, click **Select the reference for the reads**, then select another reference file.

6. After you have created the analysis:
 - Click **Edit** to proceed, or
 - Click **Cancel** to refrain from choosing references and close the Create Analysis window, or
 - Click **Finish** to complete analysis creation and close the window.

Edit the analysis

In the Choose Modules view of the Edit Analysis window, accept the default settings for secondary and tertiary analyses, then click **Next** to set module parameters.

Set module parameters

This section describes the procedures for setting general parameters and fragment mapping parameters. For descriptions of module parameters, see [“ChIP-Seq parameters” on page 214](#).

You can restore the default settings of parameters by clicking the **Reset to Defaults** button.

To view descriptions of parameters, place your mouse cursor over a button.

Set general parameters

1. Enter an analysis assembly name.
2. Set the analysis.regions.file parameter.
Click the button to choose a file.
3. Click **Next** to set mapping parameters.


Set mapping parameters

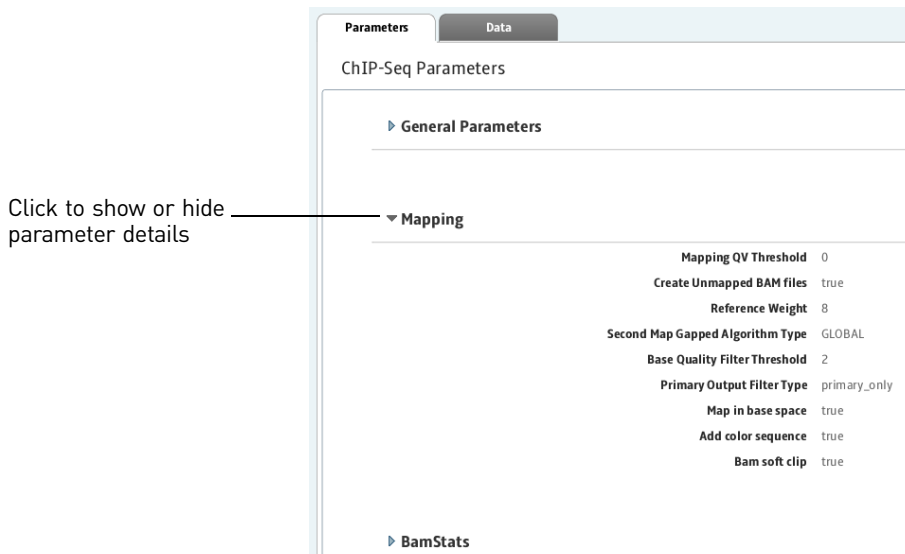
1. There are four categories of mapping parameters: Main, Advanced, BamStats, and Mark Duplicates. Accept the default settings or click the Main, Advanced, BAMStats, and Mark Duplicates tabs to edit the settings.

2. After you have edited an analysis:
 - Save your edits and proceed by clicking **Review**, or
 - Save your edits and close the window by clicking **Finish**, or
 - Erase your edits and close the Edit Analysis window by clicking **Cancel**.

Review and run the analysis

Note: For a description of ChIP-Seq parameters, see “ChIP-Seq parameters” on page 214.

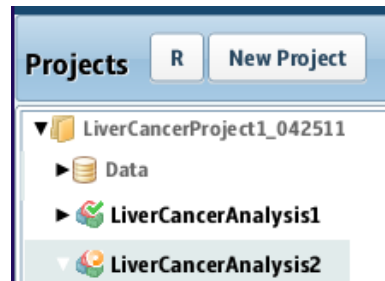
The Review Analysis view in the Run Analysis window includes two tabs: Parameters and Data. In the Parameters tab, click the  next to the parameter categories to show or hide the parameters.



1. Review the parameters. To edit the parameters, click **Edit**.
2. To review the data that will be analyzed, click the **Data** tab.
3. If you are ready to run the analysis, click **Start Analysis**.
The window closes.

Check Analysis Status

1. In the Projects organizer (shown on page 59), click the project to check the status of data mapping.



2. Click the analysis name to show and show details about the analysis in the status overview (shown on page 59).
3. Click the Status tab to view the Progress column, which shows the percentage of completion for data mapping.

LiverCancerProject1... • LiverCancerAnalysis2

Overview Status Parameters

Analysis Runs

Name	XSQ ID	Analysis	Secondary Analysis	Secondary Progress	Tertiary Analysis	Tertiary Progress
solid0054_20110102_PE_LFD...	solid0054_20110102_PE_LFD...	LiverCancerAnalysis2	[Mapping_BamStats]	13%		

After ChIP-Seq mapping has been completed, you can view mapping results.

View ChIP-Seq mapping results

ChIP-Seq mapping results include additional files and log files.

LiverCancerProject1_042511 • LiverCancerAnalysis2 • Mapping • solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary

Details Statistics Additional Files Logs

Additional Files:

Directory	File	Size	Type
fragment.mapping/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam.bai	90 KB	bai
fragment.mapping/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary	solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake-1.bam	545173 KB	bam

LiverCancerProject1_042511 • LiverCancerAnalysis2 • Mapping • solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary

Details Statistics Additional Files Logs

Log Files:

Directory	File	Size
fragment.mapping	secondary-Homo_sapiens_NCBH36.54.dna.chr22-fragment.mapping.main.20110426163010462.log	71 KB
fragment.mapping	secondary-Homo_sapiens_NCBH36.54.dna.chr22-fragment.mapping.10.run.20110426163028605.log	29 KB
fragment.mapping	secondary-Homo_sapiens_NCBH36.54.dna.chr22-fragment.mapping.11.gather-00.20110426164729990.log	10 KB
fragment.mapping	secondary-Homo_sapiens_NCBH36.54.dna.chr22-fragment.mapping.8.run.20110426163028444.log	28 KB

Refer to [Chapter 8, “View Analysis Results”](#) on page 87 for instructions on how to review mapping results.

ChIP-Seq output files

The outputs of ChIP-Seq mapping are a BAM file (.bam) and BAM file index (.bai). The .bai file is required to view the associated BAM file in some genomic browsers.

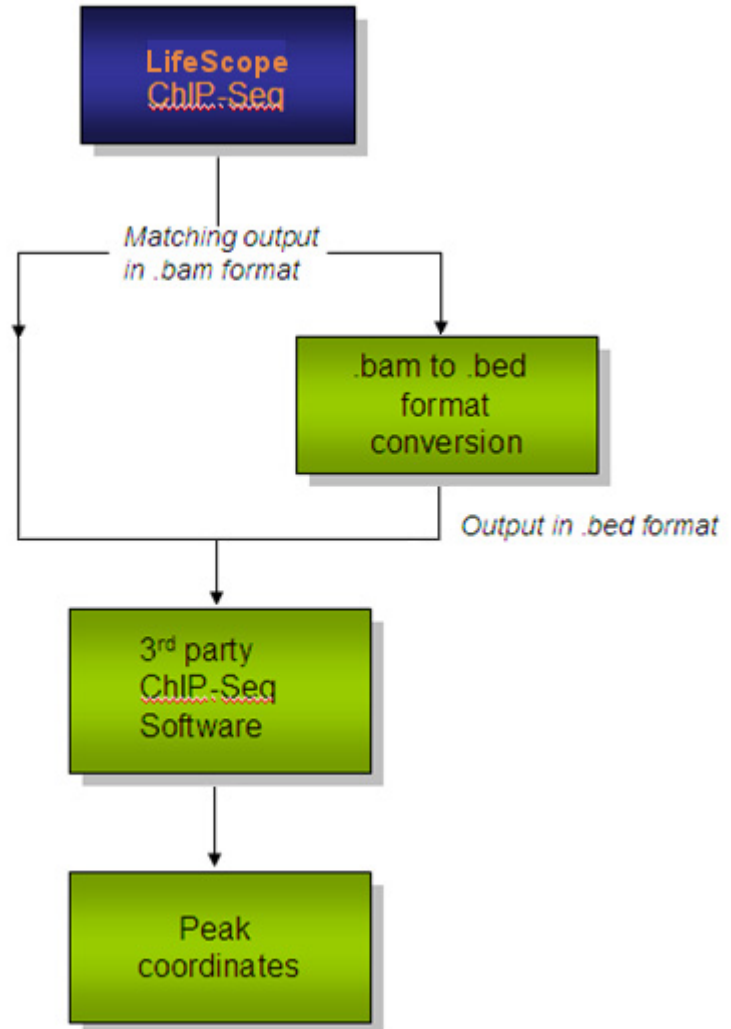
After ChIP-Seq mapping has been completed, the resulting BAM file can be used with compatible third-party commercial and academic ChIP-Seq analysis software tools (see the illustration [on page 221](#)).

Examples of such software tools include:

- SAMTools and BEDTools for generic manipulation of SAM/BAM and BED files, respectively
- MACS, QuEST, and USeq for peak-finding
- browsers like the UCSC genome browser, or IGV for viewing mapped reads and peaks relative to other genomic annotations

You can download a BAM-to-BED format converter from third-party tools sites, for example:

<http://sourceforge.net/projects/bedtools>



PART IV
Analysis Modules

16

Perform Resequencing Mapping

This chapter covers:

- Introduction to resequencing mapping. 225
- Input files. 225
- Mapping parameters 226
- Mapping output files 233

Introduction to resequencing mapping

This chapter describes how to perform resequencing mapping. The mapping analysis results files are required as input for tertiary analysis modules.

During secondary analysis LifeScope™ Genomic Analysis Software performs the following steps:

- Maps or aligns reads to a reference genome.
- Accepts one or more input XSQ (eXtensible SeQuence) files to generate one or more sorted BAM files containing aligned reads.
- For mate-pair or paired-end libraries, pairs alignments that have the same bead ID.
- *(Optional)* Generates mapping statistics.

The 5500 Series SOLiD™ Sequencer generates XSQ files, optionally with ECC (Exact Call Chemistry) reads. If an ECC primer round has been performed, the XSQ output also includes the sequence information in base space, in addition to color space. The ECC primer round is recommended during sequencing, for significant increases in sequencing accuracy.

XSQ is an extensible binary file format for storing sequence data, and supports multiple independent reads at the same position in a fragment. For information on the XSQ file format refer to this site:

<http://solidsoftwaretools.com/gf/project/xsq>

Input files

The mapping module accepts as input one or more input XSQ files. The input reads can have different read lengths but they must be of the same library type. For example, LMP read-sets of length 50 and 60 can be processed together. The XSQ files processed in a single analysis can not have different library types. In one mapping analysis, the input must be of only one library type, either Fragment, LMP, or PE.

Plan your input read-sets

Plan your analysis input carefully. The way you define your reads input affects the behavior of your analysis. The following factors control how your input data is analyzed:

- **Index (barcode) IDs** – Using an index ID restricts your input to the reads data of one or more indices.
- **Grouping of reads** – Each group of reads is analysed together as one specimen. The output data for a group is combined into one set of results files.
- **Multiple sample runs** – Unrelated reads can be processed together in one run of LifeScope™ Software, but analyzed separately as separate input data.

See [“Define input data” on page 31](#) for more information on designing adding input data to your analysis.

Mapping one tag of paired data is not supported.

Legacy data

If the data you want to process with LifeScope™ Software is in CSFASTA and QUAL files. LifeScope™ Software automatically converts these files to the XSQ format before mapping.

Mapping parameters

There are three categories of mapping parameters:

Mapping Parameter	Used in analysis . . .
Fragment	ChIP-Seq
	Genomic resequencing
	MethylMiner™
	Targeted resequencing
	Whole transcriptome
Paired-end	Genomic resequencing
	MethylMiner™
	Targeted resequencing
	Whole transcriptome
Mate Pair	Genomic resequencing

Fragment Mapping

There are four categories of fragment mapping parameters: Main, Advanced, SAET, and BAMStats.

Main

Parameter	Default value	Description
Add color sequence	True	Add color sequence to BAM records: Values: <ul style="list-style-type: none"> • True: Add color. • False: Do not add color.

Parameter	Default value	Description
Map in base space	False	Allow mapping in base space if input data has base space available. If only color space is available, then mapping will fail when base space mapping is turned on. Values: <ul style="list-style-type: none"> • True: Map in base space. • False: Map in color space.

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	0	Replace resulting base-calls with a quality value less than the specified value.
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: <p>True: Create.</p> <p>False: Do not create.</p>
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0–1,000.
Primary output filter type	Primary_only	
Reference weight	15	Used during base translation. In the read reconstruction process, multiple signals are combined to generate the final base call. Adds weight (in terms of Phred score) to the signals that are compatible with reference. Color combinations that result in a variant are considered compatible with reference. Additional weight helps to eliminate base errors caused by color error(s) during base translation. Allowed values: Integers 0–100.
Second map gapped algorithm type	GLOBAL	Do indel finding, and control the behavior of indel finding. Allowed values: <ul style="list-style-type: none"> • NONE: Turn off indel finding. • GLOBAL: Mapping reports global alignment up to one indel. • LOCAL: Mapping reports local alignment up to one indel.

SAET

The SAET tab is accessible only if you select SAET pre-processing when you choose modules.

Parameter	Default value	Description
Genome length	1000000	Expected length of sequenced (or enriched) DNA region. For example, 4,600,000 for the E.Coli 4.6 MB genome or 30,000,000 for the entire Human Transcriptome. Allowed values: Integers ≥ 200 .
On target ratio	0.5	The expected ratio of reads that come from the targeted region. Allowed values: Floats 0.0–1.0.
Update quality values	True	Update quality value of modified calls. Allowed values: True: Update the QV of modified calls. False: Do not update the qv for modified calls.
Trusted quality value	25	Correction is applied to calls with a quality value below the value of this parameter. Allowed values: Integers ≥ 1 .
Support votes	2	The minimum number of k-mer votes required make a correction. Allowed values: Integers ≥ 1 .
Trusted frequency	0	The lowest multiplicity of the seed to be included in the spectrum. (If set to 0, then the value is computed internally.) Allowed values: Integers ≥ 0 .
Maximum corrections per read	0	Maximum number of allowed corrections per read. (If set to 0, then the value is set to $\lceil \text{readLength}/8 \rceil$). Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: Integers 0–9.
Number of recursive runs	1	The error correction step is repeated the provided number of times. Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: 1, 2, or 3.
Position of error inflation point	0	Position in the read at which the error rate inflates, for instance, 35–40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers.

Parameter	Default value	Description
Disable random sampling for large data	False	Disables random sampling in spectrum building. If set to 0, then for large datasets (coverage > 300x), a subset of reads is used in spectrum building. Allowed values: <ul style="list-style-type: none"> • True: Disables random sampling in spectrum building. • False: Do not disable random sampling in spectrum building. true?
K-mer size	0	Size of k-mer (>5) used in spectrum construction and error correction. (If set to 0, then the value is computed internally.) Allowed values: Integers 0–28.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	\${analysis.output.dir}/ fragment.mapping	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	\${task.output.dir}	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum Coverage	10,000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10,000.
Maximum insert size	100,000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 0–100,000.
Insert bin size	100	Bin size for insert range distribution. Allowed values: Integers 1–100,000
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0–1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.

Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers 1-100,000.
--	-----	--

Paired-end

There are four categories of paired-end parameters: Main, Advanced, SAET, BAMStats, and Mark Duplicates

Main

Parameter	Default value	Description
Map in base space	False	Map in base space. Set to true if input data has base space available. Allowed values: <ul style="list-style-type: none"> • true: Map in base space. • false: Map in color space. If only color space is available, then the module fails when base space mapping is turned on.
Add color sequence	True	Map in color space. Set to true if input data has color space available. Allowed values: <ul style="list-style-type: none"> • true: Map in base space. • false: Map in color space. If only color space is available, then the module fails when base space mapping is turned on.
Minimum insert size estimate	0	0-100000
Maximum insert size estimate	20,000	0-100000

Advanced

Parameter	Default value	Description
BAM soft clip	False	Modify an unaligned portion of a read that needs to be presented in the BAM record. Values: <ul style="list-style-type: none"> • True: Soft clip. • False: Do not soft clip.
Base quality filter threshold	0	Replace resulting base-calls with a quality value less than the specified value.
Create unmapped BAM files	False	Create BAM files containing unmapped reads. Values: True: Create. False: Do not create.
Mapping QV/threshold	0	Provide control the contents written to the output BAM file depending on the quality value of the alignment. To preserve only high quality alignments, set this value to a positive integer. Allowed values: Integers 0-1,000.

SAET

Parameter	Default value	Description
Genome length	1000000	Expected length of sequenced (or enriched) DNA region. For example, 4,600,000 for the E.Coli 4.6 MB genome or 30,000,000 for the entire Human Transcriptome. Allowed values: Integers ≥ 200 .
On target ratio	0.5	The expected ratio of reads that come from the targeted region. Allowed values: Floats 0.0–1.0.
Update quality values	True	Update quality value of modified calls. Allowed values: True: Update the QV of modified calls. False: Do not update the qv for modified calls.
Trusted quality value	25	Correction is applied to calls with a quality value below the value of this parameter. Allowed values: Integers ≥ 1 .
Support votes	2	The minimum number of k-mer votes required make a correction. Allowed values: Integers ≥ 1 .
Trusted frequency	0	The lowest multiplicity of the seed to be included in the spectrum. (If set to 0, then the value is computed internally.) Allowed values: Integers ≥ 0 .
Maximum corrections per read	0	Maximum number of allowed corrections per read. (If set to 0, then the value is set to $\lceil \text{readLength}/8 \rceil$). Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: Integers 0–9.
Number of recursive runs	2	The error correction step is repeated the provided number of times. Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: 1, 2, or 3.
Position of error inflation point	0	Position in the read at which the error rate inflates, for instance, 35–40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers.

Parameter	Default value	Description
Disable random sampling for large data	False	Disables random sampling in spectrum building. If set to 0, then for large datasets (coverage > 300x), a subset of reads is used in spectrum building. Allowed values: <ul style="list-style-type: none"> • True: Disables random sampling in spectrum building. • False: Do not disable random sampling in spectrum building. true?
K-mer size	0	Size of k-mer (>5) used in spectrum construction and error correction. (If set to 0, then the value is computed internally.) Allowed values: Integers 0–28.

BAMStats

Parameter	Default value	Description
Input directory for BAMStats	\${analysis.output.dir}/ fragment.mapping	The input directory for BAMStats. There should be one directory per sample containing the BAM files for that sample.
Output directory	\${task.output.dir}	The path to the output directory where BAMStats will write its chart (.cht) files.
Maximum Coverage	10000	Defines the maximum coverage allowed for locations in the reference. Locations with coverage more than the maximum coverage value are ignored during coverage calculations. Allowed values: Integers 0–10,000.
Maximum insert size	100000	Defines the maximum insert size allowed for mate pair and paired-end libraries. Reads with an insert size greater than the maximum insert size value are ignored for the Insert Range Report calculations. Allowed values: Integers 0–100000.
Insert bin size	1000	Bin size for insert range distribution. Allowed values: Integers ≥ 1 .
Whether to combine data from both the strands for coverage in WIG format	0	Combine or do not combine data from both strands for coverage in WIG format. Allowed values: 0, 1.
Primary alignments only for coverage in WIG file format	1	Use only primary alignments for coverage in WIG file format. Allowed values: <ul style="list-style-type: none"> • 0: Do not restrict coverage in WIG file format to only primary alignments. • 1: Restrict coverage in WIG file format to only primary alignments.


Parameter	Default value	Description
Bin size for coverage in WIG file format	100	The bin size for coverage in WIG file format. Allowed values: Integers > 1.

Mark Duplicates

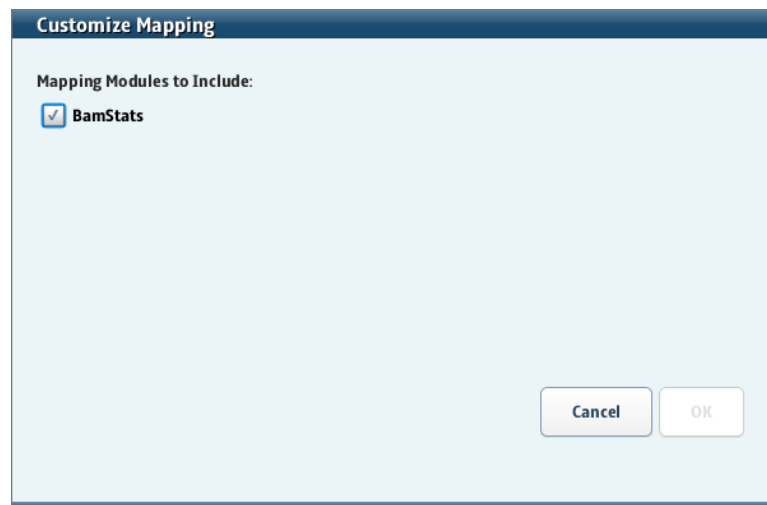
Parameter	Default value	Description
Input directory with structured mapping output	\${analysis.output.dir}/pair.mapping	Folder with one group per Sequencing Indexing run, and named [Group]/[File]-[number].bam

Map data

Mapping data is automatic in LifeScope™ Software. To map data:

1. In the Projects organizer (shown on page 59), select a project.
2. Click an analysis icon , then click **Edit Analysis**.
3. In the Edit Analysis window, accept the default settings for mapping and pre-processing data in the Secondary Analysis section.

If you want to exclude BAMStats, click **Customize**. In the Customize Mapping window, uncheck BAMStats, then click **OK**.



If you do not want to pre-process data, uncheck the box.

Mapping output files

For every input BAM file, a set of statistics files are generated. These files are in chart (.cht), comma-separated values (.csv), text (.txt), and wiggle (.wig) formats. Each CHT file corresponds to one displayed chart. A CHT file specifies the type of chart, the displayed range of each axis, and the data points, without using external references.

The .cht file format is an internal file format based on the .csv file format, with additional header information. This is an example of .cht header information:

```
# name:
# type: scatter2d | pie | vbar | line
# title:
# xaxisname:
# yaxisname:
# xrange: <min>:<tickinterval>:<max>
# yrange: <min>:<tickinterval>:<max>
XAXISNAME, SERIES1NAME, SERIES2NAME, ...
x1, y1.1, y1.2, ...
x2, y2.1, y2.2, ...
x3, y3.1, y3.2, ...
```

The wiggle format (.wig) is a public format typically used for coverage. Visit their site for more information:

hgdownload.cse.ucsc.edu/goldenPath/help/wiggle.html

A genome browser such as the Integrative Genomics Viewer (IGV) can be used to visualize the coverage. For information is available from their site:

www.broadinstitute.org/igv/

For a collection of input BAM files that belong to a sample, a set of cumulative statistics files are generated. The cumulative statistics files are also in CHT, CSV, TXT, and WIG formats. The cumulative statistics can be visualized in the LifeScope™ Software UI.

Mapping output files

Details

There are no details for the paired-end mapping analysis module.

Statistics

There are no statistics for the paired-end mapping analysis module.

Additional files

The following table is an example of additional files for the paired-end mapping analysis module.

Directory	Files	File Type
pair.mapping/WT_chr17_PE_None	WT_chr17_PE-1-1.bam.bai	.bai
	WT_chr17_PE-1-1.bam	.bam

Logs

The following table is an example of logs for the paired-end mapping analysis module.

Directory	Files	File Type
pair.mapping	secondary-hg18_validated-pair.mapping.2.gather-00.20110425225114531.log	.log
	refcor.secondary-hg18_validated-pair.mapping.1.run.log	

BAMStats

Details

The following illustration is an example of details for the BAMStats analysis module.



Statistics

There are no statistics for the BAMStats analysis module.

Additional files

The following table is an example of additional files for the BAMStats analysis module.

Directory	Files	File Type
PairEndMapping.BamStats/WT_chr17_PE_None/ WT_chr17_PE-1-1/Misc	WT_chr17_PE-1-1.bam_chr2.NEG.wig	.wig
PairEndMapping.BamStats/WT_chr17_PE_None/ Misc	WT_chr17_PE_None.Mismatches.by.ReadPair.Type.csv	.csv

Logs

The following table is an example of logs for the BAMStats analysis module.

Directory	Files	File Type
PairEndMapping.BamStats	secondary-hg18_validated- PairEndMapping.BamStats.1 .convertRef.20110425225121545.log	.log

17

Perform Human Copy Number Variation Analysis

This chapter covers:

- Introduction to Human Copy Number Variation analysis 237
- Human Copy Number Variation analysis parameters 237
- Perform Human Copy Number Variation analysis 240
- View Human CNV analysis output 240

Introduction to Human Copy Number Variation analysis

Human Copy Number Variation (Human CNV) analysis in LifeScope™ Genomic Analysis Software detects copy number variations in a data sample that is mapped to the human reference sequence hg18. LifeScope™ Software Human CNV analysis currently supports only humans, because normalization is species-specific.

By default, analysis does not call CNVs that are within 1 MBase of the centromeres and telomeres of the chromosomes. A centromere is a region of DNA typically found near the middle of a chromosome where two identical sister chromatids come in contact. The centromere is involved in cell division as the point of mitotic spindle. A telomere is a region of repetitive DNA at the end of a chromosome. The telomere protects the end of the chromosome from deterioration.

Human Copy Number Variation analysis parameters

Categories of CNV parameters include Advanced and (if you selected Annotation for CNV output) Annotation. There are no Main parameters.

Note: To revert parameters to their default settings, click **Reset to Defaults**.

Advanced

There are two categories of Advanced parameters: general and `cnv.stringency.setting`.

Parameter	Default Value	Description
Window size	5000	Size of the window block to be considered as a region. Allowed values: Integers ≥ 100 .
Trim distance	1000	Distance in kilo bases to be trimmed from the extreme ends of the chromosome arms. Allowed values: Integers 0-100000
Min quality	0	Minimum quality value of the alignments. Allowed values: Integers 0-99
Ploidy	2	General ploidy of the genome. Allowed values: Integers ≥ 1 .

Ploidy exception	None	List of all the contigs whose ploidy is different to the general ploidy of the genome. Entries in the list are in the format {contig id: ploidy of the contig}, and are separated by commas. Use the string "None" to indicate no entries.
Local normalization	False	Whether or not genome-wide normalization or local normalization should be performed. <ul style="list-style-type: none"> • True: Perform chromosome-arm local normalization. • False: Perform genome-wide normalization.
Write coverage	False	Create coverage output files <ul style="list-style-type: none"> • True: Create coverage output files, in WIG format. • False: Do not create.
Coverage window size	1000	Size of the window block to be considered as a region for writing coverage output. The mean coverage of all bases in each of these windows is output. Allowed values: Integers: 1–100000.

cnv.stringency.setting parameters

Parameter	Default Value	Description
Deletions min mappability	50	Minimum mappability percentage for regions to be shown as copy number deletions. Allowed values: 0.0000 - <100.0000 If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 0.
Insertions min mappability	10	Minimum mappability percentage for the regions to be shown as copy number insertions. Allowed values: 0.0000 - <100.0000. If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 0.
Deletion min windows	2	Minimum number of windows for the regions to be shown as copy number deletions. Allowed values: Integers ≥0. If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 4. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.
Insertion min windows	2	Minimum number of windows for the regions to be shown as copy number insertions. Allowed values: Integers ≥0. If this parameter is not specified, and <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this parameter to 4. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.

Deletion max p-value	1.0	<p>Maximum p-value for regions to be shown as copy number deletions. Allowed values: >0.0000 - 1.0000.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this to 0.25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.0.
Insertion max p-value	1.0	<p>Maximum p-value for regions to be shown as copy number insertions. Allowed values: >0.0000 - 1.0000.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then LifeScope™ Software sets this to 0.25. • Stringency setting is set to <i>Low</i>, then LifeScope™ Software sets this parameter to 1.0.
Stringency setting	Medium	<p>Allowed values:</p> <p>High: Recommend when a very low false positive tolerance is allowed</p> <p>Medium: Default values</p> <p>Low: Aggressive CNV calling</p> <p>Lower settings result in more CNV calls, but with more false positives.</p> <p>Higher settings result in fewer CNV calls, but with fewer false positives.</p>
Deletion max ratio	0.5	<p>Maximum ratio between the coverage of the region and the expected coverage, for a region to be called as CNV deletion.</p> <p>Allowed values: >0.000-<1.0000.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then this parameter is set to 0.25. • Stringency setting is set to <i>Low</i>, then this parameter is set to 0.7.
Insert min ratio	1.25	<p>Minimum ratio between the coverage of the region and the expected coverage, for a region to be called as CNVs insertion.</p> <p>Allowed values: Floats ≥1.0.</p> <p>If this parameter is not specified, and</p> <ul style="list-style-type: none"> • Stringency setting is set to <i>High</i>, then this parameter is set to 1.75. • Stringency setting is set to <i>Low</i>, then this parameter is set to 1.25.
Gender	Male	Set the ploidy of all chromosomes for Human.
Mappability directory	<code>\${analysis.mappability.dir}</code>	Path to the directory of mappability files.

**(Optional)
Annotation**

You can optionally annotate the mapped output of Human Copy Number Variation analysis. For descriptions of the Annotation parameters, see [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform Human Copy Number Variation analysis

The Human CNV analysis module detects CNV in SOLiD™ System data that originates from a single human sample. Slide(s) from this sample must be mapped to the hg18 reference to facilitate correct normalization.

Use the Human CNV module to perform tertiary analysis. To use the module:

1. Select a project in the Projects organizer (shown on [page 59](#)).
2. Create an analysis or edit an analysis that has not yet been run:

Create an analysis: Click either:

 - **Analysis** in the top menu, then **Create**, or
 - **Create Analysis** in the Task Wizards section (shown [page 59](#)).

Edit an analysis: Click **Edit Analysis** in the Task Wizards section.
3. In the Edit Analysis window, select **Human Copy Number Variation** in the Available Modules pane, click the > button to move it to the Include pane. You can optionally annotate the analysis output.
Click **Next** to set the general parameters.
4. Set the General Parameters:
 - analysis.assembly.name
 - annotation.dbsnp.file
 - analysis.regions.file
 - annotation.gtf.file
 Click **Next**.
5. Review and edit the **Pair End Mapping** parameters, described in “[Paired-end](#)” on [page 230](#), then click **Next**.
6. Review and edit the Human Copy Number Variation analysis parameters, described in “[Human Copy Number Variation analysis parameters](#)” on [page 237](#), then click **Next**.
7. If you selected Annotation in [step 3](#), review and edit the Annotation parameters, described in [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on [page 269](#).
8. Run the analysis.

View analysis status

In the Projects organizer (shown [on page 59](#)), click the **Human Copy Number Variation** analysis module, then click the **Status** tab in the overview section (shown [on page 59](#)).

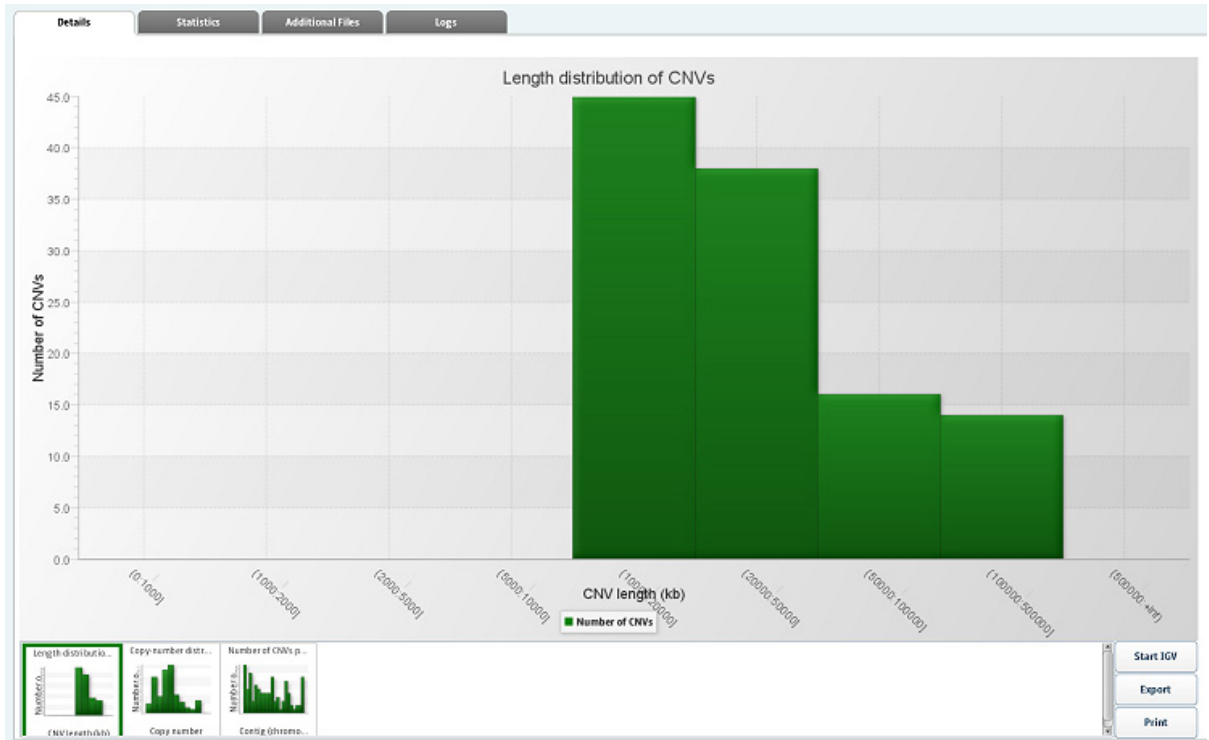
View Human CNV analysis output

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view details, statistics, additional files, and logs. For more information about viewing analysis results, see [Chapter 8, “View Analysis Results”](#) on [page 87](#).

Details

The Details tab shows analysis results in the form of bar charts and pie charts. The following illustration is an example of details for the Human CNV analysis module.



Statistics

The Statistics tab shows a LifeScope report of statistics that you can view, export, and print. You cannot edit the statistics. The following illustration is an example of the statistics in a Human CNV analysis.

```

*****
LIFESCOPE REPORT
*****
Input File:
/datapod1/sitest1/results/projects/lifescopel/GRIMP/GR_IMP_COLORS/SPACE_NONBARCODE/outputs/cnv/solid0049_20090724_HuReseqIMP_1_F3R3_None/outputC
NVs.gtf
Date: Apr 22, 2011 4:18:18 AM
Annotation file, dbSNP: /panasas/lifescopel/analysis/referenceData/lifetech/hg18/dbSNP/dbSNP_b130.tab
Annotation file, Genes and Exons, GTF: /panasas/lifescopel/analysis/referenceData/lifetech/hg18/refGene/refGene.hg18.20090513.gtf
*****
Statistics Overview
*****
-- Basic Statistics -----
Number of variants                               113
-- Variant-Specific Statistics (CNV) -----
Distribution of CNV copy number
Copy_Number      Number
0                 5
1                20
2                 9
3                24
4                27
5                10
6                 6
7                 3
8                 2
9                 7
>9                0
Distribution of CNV length
Length (bases)      Number
(0:1000]            0
(1000:2000]        0
(2000:5000]        0
(5000:10000]       0
(10000:20000]      45
(20000:50000]     38
(50000:100000]    16
    
```

Additional files

The following table is an example of additional files for the Human CNV analysis module.

Directory	Files	File Type
cnv/solid0049_20090724_HuReseqLMP_1_F3R3_None	dbSnpDeployed_annotated.gff3	.gff3
	dbSnpDeployed_genes.tab	.tab
	ALLCNVs.out	.out
	dbSnpDeployed_genes.bed	.bed

Logs

The following table is an example of logs for the Human CNV analysis module.

Directory	Files	File Type
cnv	cnv-solid0049_20090724_HuReseqLMP_1_F3R3_None.20110422034750570.log	.log

18

Perform Inversion Analysis

This chapter covers:

■ Introduction to inversion analysis	243
■ Inversion analysis input files	243
■ Inversion analysis parameters	244
■ Perform inversion analysis	245
■ View inversion analysis output	246

Introduction to inversion analysis

An inversion is defined by its two breakpoints. The numbers of mate-pairs supporting occurrence of the starting and ending inversion breakpoints are counted for each base pair. The genomic ranges corresponding to local peaks of these counts, if above a score threshold, are called as candidate breakpoint ranges. To define an inversion, its starting and ending breakpoints are paired up only if they are the reciprocal nearest neighbor of each other in the correct order. The score for the inversion is the harmonic mean of its two breakpoints. Each breakpoint range can be scanned for coverage of normal (AAA) mate pairs to identify a sub-range with the lowest normal mate pair coverage as the most probable location of a breakpoint, and to differentiate homozygous inversions from heterozygous ones. The supporting evidence of all inversions can be visually inspected using a genomic browser.

To better resolve inversion breakpoints, new references are constructed for the inversions identified and inverted with respect to flanking regions with redundant breakpoint ranges. SOLiD™ System reads are then matched and paired against the references, in which process reads split across boundaries of breakpoint ranges are parsed and clustered into candidate inversion breakpoints.

Inversion analysis input files

The inversion module takes the BAM file output of the mapping module as input. Because the inversion module specifically selects for pairs that are in the incorrect orientation, it is important to know the correct orientation. As a result, either the library type information in the LB field or the `library.type` parameter must be properly set.

Libraries in the input file or files must be either:

- All mate-pair
- All paired-end

When different library types are provided as input on the same run, the inversion module fails with an error.

Read lengths

Input BAM files may contain multiple read lengths.

Inversion analysis parameters

Note: To revert parameters to their default settings, click **Reset to Defaults**.

Review and edit the Main parameters for the Inversion analysis module. There are no Advanced parameters.

Main

Parameters	Default value	Description
Library type	Mate-pair	Specifies the library type. Allowed values: <ul style="list-style-type: none"> • matepair • pairedend
Calculate MP coverage	False	Whether to calculate normal mate-pair coverage around inversion breakpoints. Allowed values: <ul style="list-style-type: none"> • True: Use this calculation. • False: Do not use this calculation.
Number of chromosomes	25	The number of chromosomes used to distribute jobs on a cluster (one job per chromosome). Allowed values: Integers ≥ 0 .
ABX score	0	The ABX score. ABX is ABA/ABB/ABC, which are mates with both tags on the correct strands but in reverse order. Allowed values: ≥ 0.0000 – 1.000
Force update intermediate files	False	Whether to force updating of all intermediate files. Allowed values: <ul style="list-style-type: none"> • True: Force updating of all intermediate files. • False: Do not force updating of all intermediate files.
Downweight MP mismatches	False	Whether to down-weight mate-pairs with mismatches exponentially. Allowed values: <ul style="list-style-type: none"> • True: Down-weight mate-pairs with mismatches exponentially. • False: Do not down-weight mate-pairs with mismatches exponentially.
Map BXXX MP length	3000000	The maximal mapped length of BXX mate-pairs. Allowed values: Integers ≥ 0 .

Parameters	Default value	Description
Max inversion length	100000	The maximum mapped length of inversions. Allowed values: Integers ≥ 0 .
Min pairing quality	10	Minimum pairing quality threshold. Allowed values: Integers 0–100
Score run individually	True	Whether to score every run individually. Allowed values: <ul style="list-style-type: none"> • True: Score every run individually. • False: Do not score every run individually.
Breakpoint rescue	False	Whether to pair breakpoints with rescue. Allowed values: <ul style="list-style-type: none"> • True: Pair breakpoints with rescue. • False: Do not pair breakpoints with rescue.
Recover small inversions	False	Whether to recover small inversions. Allowed values: <ul style="list-style-type: none"> • True: Recover small inversions. • False: Do not recover small inversions.
Max length small inversions	0	The maximum mapped length of small inversions. Allowed values: Integers ≥ 0 .
Breakpoint score threshold	4	The break point score threshold. Allowed values: Integers 0–100.
Output score threshold	0	The output score threshold. Allowed values: Integers 0–100.
Breakpoint peak width	100	The break point peak width. Allowed values: Integers 0–100.

Perform inversion analysis

To perform inversion analysis:

1. Select a project in the Projects organizer (shown [on page 59](#)).
2. Create an analysis or edit an analysis that has not yet been run:
 - Create an analysis:** Click either:
 - **Analysis** in the top menu, then **Create**, or
 - **Create Analysis** in the Task Wizards section (shown [on page 59](#)).
 - Edit an analysis:** Click **Edit Analysis** in the Task Wizards section.
3. In the Edit Analysis window, select **Inversion** in the Available Modules pane, click the > button to move it to the Include pane. You can optionally annotate the analysis output.
Click **Next**.

- Set the general parameters, described in the following table:

Parameter	Description
analysis.assembly.name	The name of the genome assembly used in current analysis. Examples are hg18 and hg19.
annotation.dbsnp.file	The path to the file used to annotate SNPs and small InDels. LifeTech-provided files are dbSNP_b130.tab(hg18) and 00-All.vcf(hg19).
analysis.mappability.dir	The path to the directory containing binary mappability files used in CNV module.
analysis.regions.file	The path to the file containing genomic regions used in targeted resequencing selection, such as a .bed format file containing exome targets.
annotation.gtf.file	The path to the file containing gene and exon annotations corresponding to the genome assembly used in the analysis.

- Review and edit the mapping parameters, then click **Next**.
- Review and edit the Inversion analysis parameters, described in [“Inversion analysis parameters” on page 244](#), then click **Next**.
- If you selected Annotation in [step 3](#), review and edit the Annotation parameters, described in [Chapter 22, “Add Genomic Annotations to Analysis Results” on page 269](#).
- Run the analysis.

View analysis status

In the Projects organizer (shown [on page 59](#)), click the **Inversion** analysis module, then click the **Status** tab in the overview section (shown [on page 59](#)).

View inversion analysis output

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view details, statistics, additional files, and logs. For more information about viewing analysis results, see [Chapter 8, “View Analysis Results” on page 87](#).

Details

There are no details of the output for the Inversion analysis module.

Statistics

There are no statistics of the output for the Inversion analysis module.

Additional files

Additional files for large indel analysis output include count files (for example, 1, 2, 3), all and chromosome (chr) files, .gff and .gff3 files, .o5MR files, .orphan files, .txt files, and .w100 files.

The following table is an example of additional files for the Inversion analysis module.

Directory	Files	File Type
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None/starts/start.1	all 24	24

Directory	Files	File Type
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None/gff	chr14.gff	.gff
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None	inversions.s4.w100.100000.GFF3	.GFF3
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None/BXX	chr	chr
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None	pair.txt	.txt
inversion/solid0049_20090724_HuReseqLMP_1_F3R3_None	pair.orphan	orphan

Logs

The following table is an example of logs for the Inversion analysis module.

Directory	Files	File Type
inversion	inversion-solid0049_20090724_HuReseqLMP_1_F3R3_None.20110422034750573.log	.log
	tertiary-solid0049_20090724_HuReseqLMP_1_F3R3_None-inversion.20110422034715217.log	

19

Perform SNP Finding Analysis

This chapter covers:

- Introduction to SNP Finding analysis 249
- SNP analysis input files 249
- SNP Finding analysis parameters 250
- Perform SNP Finding analysis 252
- View SNP Finding analysis output 253

Introduction to SNP Finding analysis

The LifeScope™ Genomic Analysis Software SNP Finding analysis module uses the diBayes algorithm to find Single Nucleotide Polymorphisms (SNPs). The diBayes package performs independent SNP analysis at each position in the reference, using either a Bayesian or Frequentist algorithm.

The SNP Finding module is used to call SNPs from mapped and processed SOLiD™ System reads. The module takes the reads, quality values, the reference sequence, and error information on each SOLiD™ System slide as its input, and calls SNPs.

The SNP Finding module creates these results files:

- A list of SNPs.
- A quartile file with coverage as well as color and base quality value distribution.
- (Optional) A consensus FASTA file with the same number of bases as the reference sequence.
- (Optional) A consensus calls file with a list of all covered positions.
- (Optional) A collection of annotated files.

The consensus calls file and the quartile file are each generated as one file per contig. The list of SNPs and the consensus FASTA file are each generated both as one file per contig and also as a consolidated file for the entire run. See [“View SNP Finding analysis output” on page 253](#).

SNP analysis input files

The SNP input files include the reference file, the BAM file of the mapping module, and position error and probe error file output of the BAMStats mapping statistics module.

Input File	Description
Reference genome	<p>The sequence to which the reads are aligned (mapped). The required format for the reference file is the *.fasta format, for example, chr20.validated.fasta.</p> <p>Note: When you run the SNP Finding tool, you must use the same reference that was used for mapping.</p> <p>The reference sequence might have multiple chromosomes or contigs, and it might contain IUB codes at positions of known SNPs. A heterozygote is called with fewer reads providing evidence at these positions if the reference sequence contains IUB codes, because the prior probability of a heterozygote existing at this position is higher.</p>
F3 (R3/F5) position error file	<p>F3 (R3/F5) position error files are tab-delimited text files created during secondary analysis. The files record the frequencies of dicolor mismatches between reads and the reference at different positions in a read. For fragment runs, SNP Finding only requires an F3 position error file. Both F3 and R3(F5) position error files are created for mate-pair or paired-end runs, and both are required for running SNP Finding. Position error files are generated during mapping.</p>
F3 (R3/F5) probe error file	<p>F3 (R3/F5) probe error files are tab-delimited text files that record the frequencies of dicolor mismatches between the reads and the reference as a function of different 6-mer probes. LifeScope™ Software calculates the probe error files.</p> <p>Reads only have the F3 tag for fragment runs. Thus, SNP Finding requires only an F3 probe error file. Both F3 and R3/F5 probe error files are created for mate-pair (paired-end) runs, and both are required to run SNP Finding on paired data. Different SOLiD™ system runs of the same sample might generate different F3 (R3/F5) probe error files for each run, because of the random nature of probe errors.</p>
*.bam file	<p>The *.bam input file is generated by LifeScope™ Software at the end of mapping or pairing.</p>

SNP Finding analysis parameters

Certain parameters affect only the module's base-space or color-space algorithm. These parameters are noted in the table. Changes to a base-space parameter's setting have no effect if the module runs its color-space algorithm. Similarly, changes to a color-space parameter's setting have no effect if the module runs its base-space algorithm.

There are three categories of SNP Finding parameters: Main, Advanced, and Annotation.

Note: To revert parameters to their default settings, click **Reset to Defaults**.

Main

Parameter name	Default value	Description
Call stringency	medium	Call stringency.
Skip high coverage positions (Het)	1	Skip high coverage positions (Het)
Minimum mapping quality value	8	Minimum mapping quality value.

Advanced

There are six categories of Advanced parameters: general, Read filter, General position filter, Heterozygous position filter, Homozygous position filter, and Output file processing.

Parameter name	Default value	Description
Detect adjacent SNPs	0	Detect adjacent SNPs.
Polymorphism rate	0.001	Polymorphism rate.
Read filter		
Include reads with unmapped mate	0	Include reads with unmapped mate.
Exclude reads with indels	True	Exclude reads with indels.
Require only uniquely mapped reads	0	Require only uniquely mapped reads
Ignore reads with a higher mismatch count to alignment length ratio	1.0	Ignore reads with a higher mismatch count to alignment length ratio.
Ignore reads with a lower alignment length to read length ratio	1.0	Ignore reads with a lower alignment length to read length ratio.
General position filter		
Require alleles to be present in both strands	False	Require alleles to be present in both strands
Minimum base quality value for a position	14	Minimum base quality value for a position.
Minimum base quality value of the non-reference allele of a position	14	Minimum base quality value of the non-reference allele of a position.
Heterozygous position filter		
Minimum allele ratio (Het)	0.15	Minimum allele ratio (Het).
Minimum coverage (Het)	2	Minimum coverage (Het).
Minimum unique start position (Het)	2	Minimum unique start position (Het).
Minimum non-reference color QV (Het)	7	Minimum non-reference color QV (Het).
Minimum non-reference base QV (Het)	14	Minimum non-reference base QV (Het).
Minimum ratio of valid reads (Het)	0.65	Minimum ratio of valid reads (Het).
Minimum valid tricolor counts (Het)	2	Minimum valid tricolor counts (Het).
Homozygous position filter		
Minimum coverage (Hom)	1	Minimum coverage (Hom).
Minimum count of the non-reference allele (Hom)	2	Minimum count of the non-reference allele (Hom).
Minimum average non-reference base QV (hom)	14	Minimum average non-reference base QV (Hom).

Parameter name	Default value	Description
Minimum average non-reference color QV (hom)	7	Minimum average non-reference color QV (Hom)
Minimum unique start position of the non-reference allele (Hom)	2	Minimum unique start position of the non-reference allele (Hom).
Output file processing		
Output fasta file	True	Output or do not output the FASTA file. <ul style="list-style-type: none"> • True: Output the FASTA file. • False: Do not output the FASTA file.
Output consensus file	True	Output or do not output the consensus file. <ul style="list-style-type: none"> • True: Output the consensus file. • False: Do not output the consensus file.
Compress the consensus file	False	Compress or do not compress the consensus file. <ul style="list-style-type: none"> • True: Compress the consensus file. • False: Do not compress the consensus file.

(Optional) Annotation

You can optionally annotate the mapped output of SNP Finding analysis. For descriptions of the Annotation parameters, see [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform SNP Finding analysis

Use the SNP Finding module to perform tertiary analysis. To use the module:

1. Select a project in the Projects organizer (shown on [page 59](#)).
2. Create an analysis or edit an analysis that has not yet been run:
 - Create an analysis:** Click either:
 - **Analysis** in the top menu, then **Create**, or
 - **Create Analysis** in the Task Wizards section (shown [page 59](#)).
 - Edit an analysis:** Click **Edit Analysis** in the Task Wizards section.
3. In the Edit Analysis window, select **SNP Finding** in the Available Modules pane, click the > button to move it to the Include pane. You can optionally annotate the analysis output.

Click **Next** to set the general parameters.
4. Set the General Parameters:
 - analysis.assembly.name
 - annotation.dbsnp.file
 - analysis.regions.file
 - annotation.gtf.file

Click **Next**.
5. Review and edit all mapping parameters, described in [“SNP analysis input files”](#) on [page 249](#), then click **Next**.

6. Review and edit the SNP Finding analysis parameters, described in “SNP Finding analysis parameters” on page 250, then click **Next**.
7. If you selected Annotation in [step 3](#), review and edit the Annotation parameters, described in [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.
8. Run the analysis.

View analysis status

In the Projects organizer (shown [on page 59](#)), click the **SNP Finding** analysis module, then click the **Status** tab in the overview section (shown [on page 59](#)).

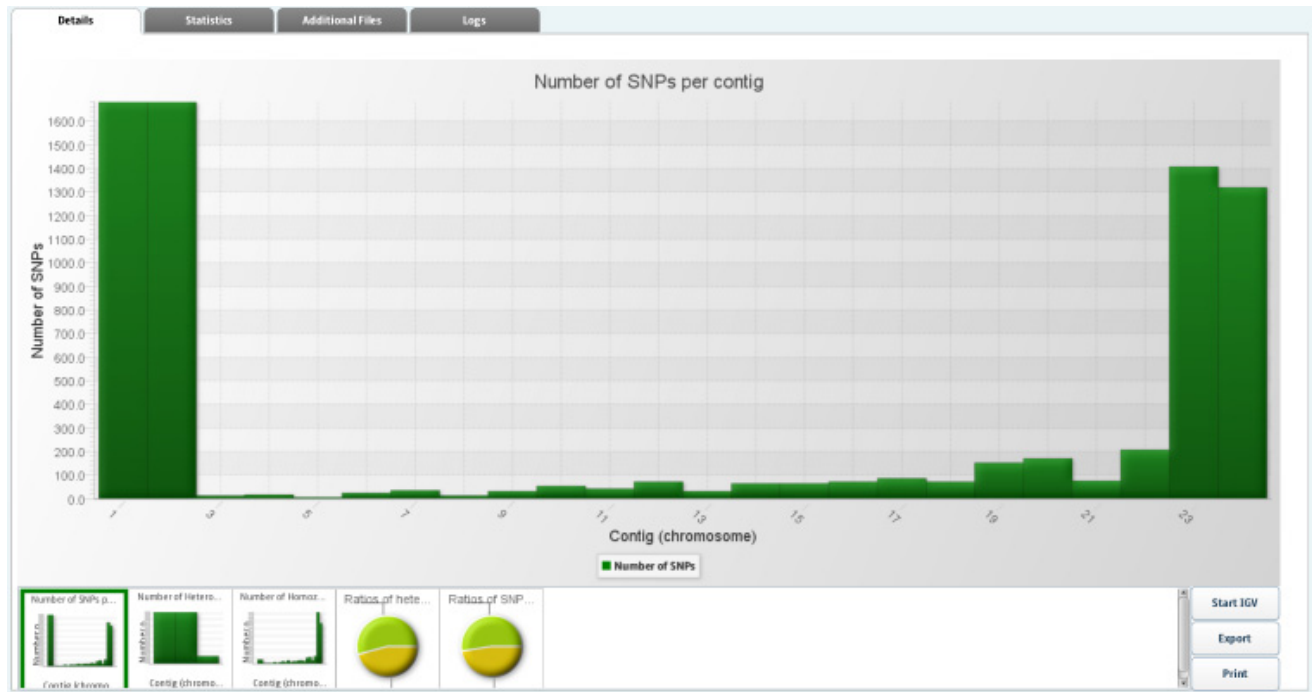
View SNP Finding analysis output

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view details, statistics, additional files, and logs. For more information about viewing analysis results, see [Chapter 19, “Perform SNP Finding Analysis”](#) on page 249.

Details

The Details tab shows analysis results in the form of bar charts and pie charts. The following illustration is an example of details for the SNP Finding analysis module.



Statistics

The Statistics tab shows a LifeScope report of statistics that you can view, export, and print. You cannot edit the statistics. The following illustration is an example of the statistics in a SNP Finding analysis.

```

*****
LIFESCOPE REPORT
*****
Input File:      /datapod1/sitest1/results/projects/lifescop/BrainCancerProject1_050411/BrainCancerAnalysis1/outputs/di
genome/hg19_simulated_TR_ECC_BC16_Frag50_320Mil_BC1/BrainCancerAnalysis1_SNP.gff3
Date:   May 4, 2011 9:55:13 PM
Annotation file, dbSNP: /panasas/lifescop20/analysis/referenceData/lifetech/hg19/dbSNP/00-All.vcf
Annotation file, Genes and Exons, GTF: /panasas/lifescop20/analysis/referenceData/lifetech/hg19/refGene/refGene.hg19.201
*****
Statistics Overview
*****
-- Basic Statistics -----
Number of variants                7388
Number of heterozygous variants   3375
Number of homozygous variants     4013
-- Variant-Specific Statistics (SNP) -----
Number of transition SNPs         3508
Number of transition heterozygous SNPs 1214
Number of transition homozygous SNPs 2294
Number of transversion SNPs      3880
Number of transversion heterozygous SNPs 2161
Number of transversion homozygous SNPs 1719
Transition:Transversion ratio     1.000 : 1.106
-- dbSNP Annotation Statistics -----
Number of          SNPs in dbSNP          140

```

Additional files Additional files for SNP Finding analysis output include .bed, .gff3, .fasta, and .txt files.

The following table is an example of additional files for the SNP Finding analysis module.

Directory	Files	File Type
dibayes.genome/ hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1/contig1	BrainCancerAnalysis1_Consensus_Calls.txt	.txt
dibayes.genome/ hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1/contig11	BrainCancerAnalysis1_Consensus_Basespace.fasta	.fasta
dibayes.genome/ hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1	BrainCancerAnalysis1_SNP_genes.bed	.bed
dibayes.genome/ hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1/contig17	BrainCancerAnalysis1_SNP.gff3	.gff3

Logs The following table is an example of logs for the SNP Finding analysis module.

Directory	Files	File Type
dibayes.genome	dibayes.genome-hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1.20110504200517354.log	.log

20

Perform Large Indel Analysis

This chapter covers:

- Introduction to large indel analysis 255
- Large indel analysis input files 255
- Large indel analysis parameters 256
- Perform large indel analysis 257
- View large indel analysis output 258

Introduction to large indel analysis

The large indel analysis module identifies deviations in clone insert size. These deviations indicate intrachromosomal structural variations compared to a reference genome. Insertions and deletions (indels) up to 100 kB are inferred by identifying positions in the genome in which the pairing distance between mapped mate-pairs is deviates significantly from what is expected at the given level of clone coverage.

The module creates a look-up table of standard deviations at each level of clone coverage. The table produces an asymptotic curve in which the minimum size of detectable indels at a given level of significance drops rapidly as the clone coverage increases. The look-up table is used to determine the significance of the deviation in average insert size at each position in the genome.

Regions of the genome that are significantly deviated are selected as candidate indels, and hierarchical clustering is used to segregate the clones into groups in which the difference in the sizes of all clones in a group is less than the specified range. Clusters with too few clones, specified by you, are removed and the candidates are assessed to determine if a homozygous or heterozygous population of deviated insert sizes remains. All clones deviated by ≥ 100 kB are discarded. Clones from various libraries with various insert sizes contribute to a single indel call by combining the probabilities associated with the clones from each library.

Large indel analysis input files

The Large Indel analysis module requires one or more Binary Alignment sequence Map (BAM) files containing mapped data and an hg18 or hg19 reference file as input.

Large indel analysis parameters

Note: To revert parameters to their default settings, click **Reset to Defaults**.

There are three categories for the Large Indel analysis module: Main, Advanced and (if you selected Annotation for Small Indel output) Annotation.

Main

Parameters	Default value	Description
Library type	Mate-pair	Specifies the library type. Allowed values: <ul style="list-style-type: none"> • matepair • pairedend
Max clone coverage	1000	The maximum physical (clone) coverage for analysis. Loci with clone coverage above this threshold are not analyzed. Allowed values: Integers ≥ 3 . You can use this parameter in combination with the parameter High coverage to reduce false positives in high density genomes, for example, bacteria.
Min coverage	3	The minimum physical (clone) coverage allowed. Loci with clone coverage below this threshold are not analyzed. Allowed values: Integers ≥ 1 .
Ploidy	2	General ploidy of the genome. Allowed values: Integers ≥ 1
Max insert size	100000	Maximum insert size, in base pairs.
Call stringency	medium	Specifies the large indel call stringency. Automates parameter adjustment to the desired stringency level. Allowed values: <ul style="list-style-type: none"> • highest: Recommend when a very low false positive tolerance is allowed. • high: Increased filtering. • medium: Default values. • low: Very aggressive. Lower settings results in more indel calls, but with more false positives. Higher settings result in fewer indel calls, but with fewer false positives.

Advanced

Parameters	Default value	Description
BAS file	—	The BAS file contains metadata information about the data source(s) and relevant mapping/pairing statistics. The BAS format is a pseudo-standard file format for storing BAM file metadata. For details see this site: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/README.bas
Min pairing quality	25 (for mate-pair) 10 (for paired-end)	Paired reads below this threshold are ignored. Allowed values: Integers 0–100.
p Value	1e-10	P-value threshold, that is, the raw probability of committing a Type 1 error incorrectly identifying a large indel. Allowed values: Floats 0.0–1.0.
High coverage	False	Eliminate clone coverage weighting. This option significantly reduces the number of false positives when analyzing very high coverage genomic data. (Very high coverage genomic data is typically greater than 1000x read coverage, and is common for bacterial genomes.) Allowed values: <ul style="list-style-type: none"> • True: Eliminate clone coverage weighting. The algorithm ignores coverage when calculating the structural significance of candidate indels. • False: No effect on the module.

(Optional) Annotation

You can optionally annotate the mapped output of Small Indel analysis. For descriptions of the Annotation parameters, see [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform large indel analysis

To perform large indel analysis:

1. Select a project in the Projects organizer (shown [on page 59](#)).
2. Create an analysis or edit an analysis that has not yet been run:
 - Create an analysis:** Click either:
 - **Analysis** in the top menu, then **Create**, or
 - **Create Analysis** in the Task Wizards section (shown [on page 59](#)).
 - Edit an analysis:** Click **Edit Analysis** in the Task Wizards section.
3. In the Edit Analysis window, select **Large Indel** in the Available Modules pane, click the > button to move it to the Include pane. You can optionally annotate the analysis output.

Click **Next** to set the general parameters.

4. Review and edit the general parameters, then click **Next**.
If you chose raw, unmapped data (XSQ), go to [step 5](#). If you chose already mapped data (BAM), skip to [step 6](#).
5. Review and edit the mapping parameters, then click **Next**.
6. Review and edit the Large Indel analysis parameters, described in “[Large indel analysis parameters](#)” on page 256, then click **Next**.
7. If you selected Annotation in [step 3](#), review and edit the Annotation parameters, described in [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.
8. Run the analysis.

View analysis status

In the Projects organizer (shown [on page 59](#)), click the **Small Indel** analysis module, then click the **Status** tab in the overview section (shown [on page 59](#)).

View large indel analysis output

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view details, statistics, additional files, and logs. For more information about viewing analysis results, see [Chapter 8, “View Analysis Results”](#) on page 87.

Details

There are no details of the output for large indel analysis.

Statistics

There are no statistics of the output for large indel analysis.

Additional files

Additional files for large indel analysis output include .bam and .bai files.

The following table is an example of additional files for the Large Indel analysis module.

Directory	Files	File Type
pair.mapping/Group_1	solid0064_20101210_MP_2X60_T3_1_F3R3.mixed-1-1.bam	.bam
	solid0064_20101210_MP_2X60_T3_1_F3R3.mixed-1-1.bam.bai	.bai

Logs

The following table is an example of logs for the Large Indel analysis module.

Directory	Files	File Type
pair.mapping	secondary-hg18-pair.mapping.2.run.20110502143247523.log	.log

21

Perform Small Indel Analysis

This chapter covers:

- Introduction to small indel analysis 259
- Small indel analysis input files 259
- Small indel analysis parameters 259
- Perform small indel analysis 264
- View small indel analysis output 265

Introduction to small indel analysis

Small indel analysis is a tertiary module in LifeScope™ Genomic Analysis Software. BAM file output from the mapping and/or pairing modules that contain gapped and ungapped alignments serve as direct inputs for small indel discovery. The gapped alignments in the BAM files are referred to as the indel evidence.

The LifeScope™ Software Small Indel module allows flexible processing of these indel evidences. LifeScope™ Software takes these gapped alignments and combines them (based on their proximity from each other) to form candidates. After these combined results are formed, small indel analysis allows for the combination of several runs together, and for filtering based on the average read position of the indel and number of non-redundant reads. The small indel module produces the candidates in a text format that is easy for you to analyze and that is also useful for importing into databases and in a public general feature format, version 3, (GFF3) format.

When an indel occurs in a sequence, and that sequence is measured using color-space, the color-space sequence has a gap the same size as the indel. The color-space sequence also leaves a signature that can indicate whether there is a measurement error within the gap, thereby reducing false-positive small indel calls.

Small indel analysis input files

The Small Indel analysis module requires one or more Binary Alignment sequence Map (BAM) files containing mapped data and an hg18 or hg19 reference file as input.

Small indel analysis parameters

Note: To revert parameters to their default settings, click **Reset to Defaults**.

Categories for the Small Indel analysis module include Advanced and (if you selected Annotation for Small Indel output) Annotation. There are no Main parameters.

Advanced

There are five categories of Advanced parameters: General Options, Pileup, Mapping Quality Filtering, Heuristic Filtering, and Indel Size Filtering.

General Options

Parameters	Default value	Description
Detail level	0	For BAM file inputs, the level of detail in output: <ul style="list-style-type: none"> • 0: Keeps only position information about the anchor read and no information for the ungapped alignment. • 1-8: Keeps only some of the alignment's anchor alignment but none of the ungapped alignment. • 9: Is most detailed, but also the slowest.
Zygoty profile name	max-mapping	Zygoty profile name. <ul style="list-style-type: none"> • classic Run classic mapping. • max-mapping: Run max mapping. • max-mapping-v2: Run max mapping version 2. • gap-align-only and no-calls: Force all zygoty calls to be homozygous calls.
Genomic region	-	Names a specific genomic region to be selected from the BAM file. Only full chromosomes are guaranteed not to alter results. Specifying partial chromosomes is allowed but may result in the loss of indels near the edges of that region. For example, chr1:2945-9659 causes a reduction of coverage for approximately a read length after position 2945.
Display base QVs	False	Display base QV scores in the GFF file. Allowed values: <p>True: Display the FASTQ base QV scores for all of the reads used for each indel in the GFF file. FASTQ strings contain semi-colons, so adding these strings may produce a GFF file that is not compatible with certain applications.</p> <p>False: Do not display QV scores in the output file.</p>
Number of alignments per pileup	1000	For pileups with more than this number of reads, set the expected number of alignments per pileup. Allowed values: Integers ≥ 0 .
Random seed	94404	The random seed value used to determine which pseudo random set of reads to use when there are greater than 1000 reads in a pileup. The random number generator used is the Mersenne Twister MT19937 algorithm. Allowed values: Integers ≥ 0 .

Pileup

Parameters	Default value	Description
Min num evid	2	Minimum number of evidences required for an indel call. A value higher than the average coverage level in most cases has a significant reduction in sensitivity. Allowed values: Integers.
Max num evid	-1	Maximum number of evidences. Allowed values: Integers. Use -1 for no maximum. Setting this value to some multiple of the average coverage could remove indels found in abnormally high coverage areas.
ConsGroup	1	Indel grouping method. Allowed values: 1: Conservative grouping of indels with 5bp max between consecutive evidences. 2: Lax grouping. Groups indels that are at maximum the higher of 15 or 7 times the indel size. 9: No grouping. Makes every indel evidence a separate pileup.

Mapping Quality Filtering

Parameters	Default value	Description
Max reported alignments	-1	Only uses those alignments where the NH field (the number of reported alignments, from the BAM record) is this value or lower. A value of -1 is to have no upper limit. The range where this is effective depends on the input BAM file's range of values of the NH tag.
Min mapping quality	8	Keeps only reads that have this or higher pairing qualities. For paired tags, mapping quality is for the pair (pairing quality), and for fragment, it is the single tag's map quality. Reads that are lower than this value are filtered out.
Min best mapping quality	10	For a particular indel called with a set of reads, at least one pairing quality in this set must be higher than this value. Allows for supporting evidences to have a lower mapping quality threshold than the best read.
Min anchor mapping quality	-1	Minimum mapping quality for a non-indel (anchor) tag. Effective only for paired reads, for the number of anchors queried as defined by <code>small.indel.detail.level</code> .

Parameters	Default value	Description
Ungapped BAM flag filter	ProperPair	For ungapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Gapped BAM flag filter	Primary	For gapped alignments, specifies the BAM flag properties that a read must have to be included. Allowed values: A comma-separated string of one or more of these values: <ul style="list-style-type: none"> • ProperPair • UniqueHit • NoOptDup • Primary • None None turns off all filters.
Edge length deletions	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0 .
Edge length insertions	0	Gap alignments that do not have this minimum length on either side of the indel will not be considered. Allowed values: Integers ≥ 0

Heuristic Filtering

Parameters	Default value	Description
Perform filtering	True	Whether or not to perform filtering in each pileup. Allowed values: <ul style="list-style-type: none"> • True: Perform filtering on pileups. • False: Do not perform filtering on pileups. Parameters that change the makeup of pileups, such as Min num evid are still active.

Parameters	Default value	Description
Indel size distribution allowed	can-cluster	<p>Indel sizes in a pileup are allowed to have certain indel size distributions.</p> <p>Allowed values:</p> <ul style="list-style-type: none"> • similar-size: 75% of the reads of a pileup must have exactly the same size. • similar-size-any-large-deletions: Any pileups with at least 2 large deletion alignments, the other pileups must have similar sizes. • can-cluster: Allowed if at least one cluster of any indel size is found. • can-cluster-any-large-deletions: Any pileups with at least 2 large deletion alignments; other pileups must be able to cluster (will have indels with two more reads with larger deletions, even if they don't form good clusters). • any: Can have any size distribution (same as <code>small.indel.require.called.indel.size=false</code>).
Remove singletons	True	<p>Remove the singletons that occur when different alignment methods are combined based on identical bead ids and read sequence. Allowed values:</p> <ul style="list-style-type: none"> • True: Remove singletons. • False: Do not remove the singletons.
Alignment compatibility filter	1	<p>Alignment compatibility level. Checks color space compatibility around the gap. Allowed values:</p> <ul style="list-style-type: none"> • 0: No alignment compatibility filtering. • 1: The small indel module determines whether the data contains base-space or color-space sequence. • 2: Force the use of base-space sequence, if present. • 3: Force the use of color-space sequence, if present.
Max coverage ratio	12	<p>Maximum clipped coverage/# non-redundant support ratio.</p> <p>Use -1 for no limit (no coverage ratio filtering).</p> <p>Allowed values: Integers.</p>
Max nonreds 4Fit	2	<p>Maximum number of non-redundant reads where read position filtering is applied.</p> <p>Allowed values: Integers.</p>
Min from end pos	9.1	<p>Minimum average number of base pairs from the end of the read required of the pileup, when there are at most a certain number of reads defined by Max nonreds 4Fit.</p> <p>Allowed values: Floats.</p>

Indel Size Filtering

Parameters	Default value	Description
Min insertions size	0	Minimum insertion size to include. Allowed values: Integers.
Min deletions size	0	Minimum deletion size to include. Allowed values: Integers.
Max insertions size	1000000000	Maximum insertion size to include. Allowed values: Integers.
Max deletions size	1000000000	Maximum deletion size to include. Allowed values: Integers.

(Optional) Annotation

You can optionally annotate the mapped output of Small Indel analysis. For descriptions of the Annotation parameters, see [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on page 269.

Perform small indel analysis

The Small Indel analysis module detects small indels in SOLiD™ System data that originates from a single human sample. Slide(s) from this sample must be mapped to the hg18 or hg19 reference to facilitate correct normalization. Use the Small Indel module to perform tertiary analysis.

To perform small indel analysis:

1. Select a project in the Projects organizer (shown [on page 59](#)).
2. Create an analysis or edit an analysis that has not yet been run:
 - Create an analysis:** Click either:
 - **Analysis** in the top menu, then **Create**, or
 - **Create Analysis** in the Task Wizards section (shown [on page 59](#)).
 - Edit an analysis:** Click **Edit Analysis** in the Task Wizards section.
3. In the Edit Analysis window, select **Small Indel** in the Available Modules pane, click the > button to move it to the Include pane. You can optionally annotate the analysis output.
Click **Next** to set the general parameters.
4. Review and edit the general parameters, then click **Next**.
If you chose raw, unmapped data (XSQ), go to [step 5](#). If you chose already mapped data (BAM), skip to [step 6](#).
5. Review and edit the mapping parameters, then click **Next**.
6. Review and edit the Small Indel analysis parameters, described in [“Small indel analysis parameters”](#) on page 259, then click **Next**.

7. If you selected Annotation in [step 3](#), review and edit the Annotation parameters, described in [Chapter 22, “Add Genomic Annotations to Analysis Results”](#) on [page 269](#).
8. Run the analysis.

View analysis status

In the Projects organizer (shown [on page 59](#)), click the **Small Indel** analysis module, then click the **Status** tab in the overview section (shown [on page 59](#)).

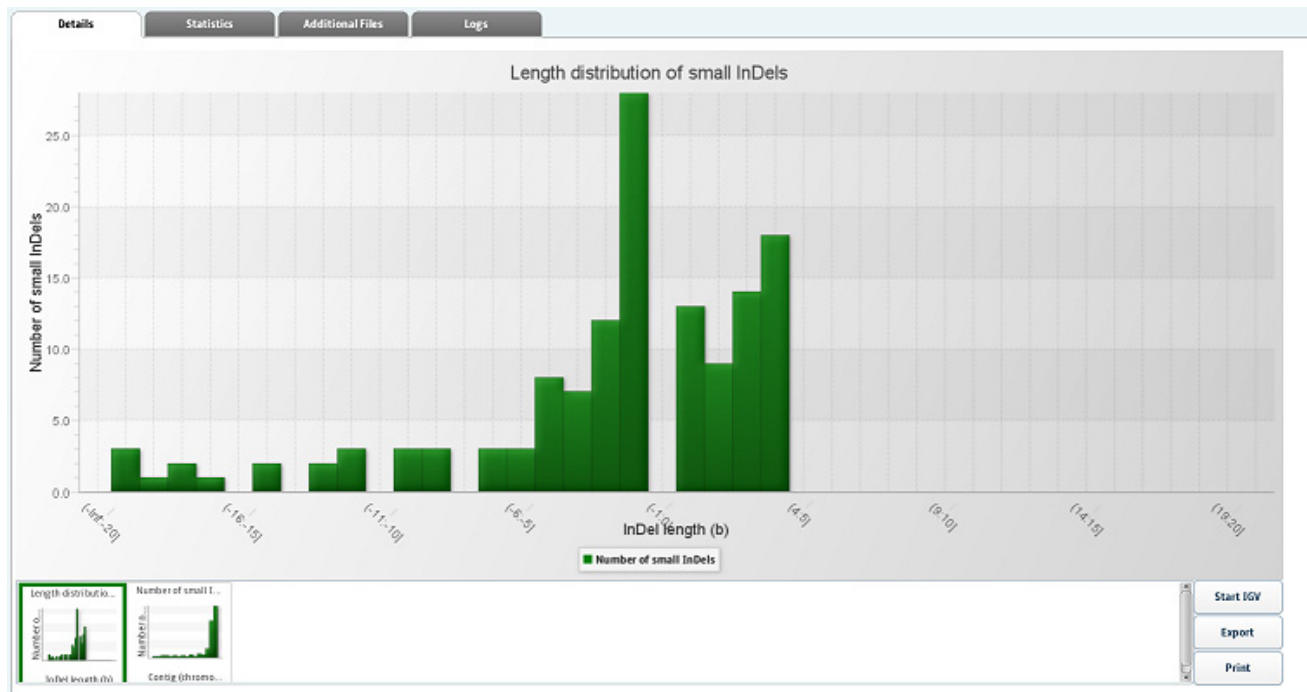
View small indel analysis output

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view details, statistics, additional files, and logs. For more information about viewing analysis results, see [“View analysis results”](#) on [page 87](#).

Details

The Details tab shows analysis results in the form of bar charts and pie charts. The illustration [on page 265](#) is an example of details for the Small Indel analysis module.



Statistics

The Statistics tab shows a LifeScope report of statistics that you can view, export, and print. You cannot edit the statistics. The following illustration is an example of the statistics in a small indel analysis.

```

*****
LIFESCOPE REPORT
*****
Input File:      /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.
indel/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2.gff3
Date:   Apr 21, 2011 10:05:52 PM
Annotation file, dbSNP: /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.
indel/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2.gff3
Annotation file, Genes and Exons, GTF: /panasas/lifescopetesttempl/results/projects/corona/Darryl_Apr21/Analysis2/outputs/small.
indel/solid0054_20110102_PE_LFD_RD_SetA_1_HuRef100_F3.colorRemake_DefaultLibrary/Analysis2.gff3
*****
Statistics Overview
*****
-- Basic Statistics -----
Number of variants                186
Number of heterozygous variants   127
Number of homozygous variants     59
-- Variant-Specific Statistics (InDel) -----
Distribution of InDel length
Length(bases)      Number
(-Inf:-20]         0
(-20:-19]          3
(-19:-18]          8
(-18:-17]          1
(-17:-16]          5
(-16:-15]          1
(-15:-14]          4
(-14:-13]          2
(-13:-12]          1
(-12:-11]          4
(-11:-10]          3
(-10:-9]           1
(-9:-8]            6
(-8:-7]            1
(-7:-6]            3
(-6:-5]            4
(-5:-4]            5
(-4:-3]            10
(-3:-2]            11
(-2:-1]            61

```

Additional files

Additional files for small indel analysis output include .align, .bed, .gff3, .pas, sql, sum, .tab, .txt, and .ungapped files.

The following table is an example of additional files for the Small Indel analysis module.

Directory	Files	File Type
small.indel/ hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_ BC1	BrainCancerAnalysis1.pas	.pas
	BrainCancerAnalysis1_genes.bed	.bed
	BrainCancerAnalysis1.ungapped	.ungap...
	BrainCancerAnalysis1.gff3	.gff3
	BrainCancerAnalysis1_annotated.gff3	
	BrainCancerAnalysis1.txt	.txt
	BrainCancerAnalysis1_genes.tab	.tab
	BrainCancerAnalysis1.align	.align
	BrainCancerAnalysis1.sql	.sql

Logs

The following table is an example of logs for the Small Indel analysis module.

Directory	Files	File Type
small.indel	small.indel-hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_BC1.20110504200216106.log	.log
	tertiary-hg19_Simulated_TR_ECC_BC16_Frag50_320Mil_BC1-small.indel.20110504200214855.log	

22

Add Genomic Annotations to Analysis Results

This chapter covers:

■ Overview	269
■ Annotation sources	272
■ Annotation parameters	273
■ About annotations and LifeScope™ Software modules	289
■ Examples of annotation output files	290

Overview

The annotations module is an optional post-processing step available with select LifeScope™ Genomic Analysis Software analysis modules. The annotations processing makes a copy of the analysis module's GFF file and adds new attributes to the GFF entries in the copied file. The new attributes are taken from information in publicly available sources about the variants in the input, about features intersecting the variants in the input file, or about any biological function potentially changed by the variants.

Annotations processing does not modify the original LifeScope™ Software analysis module's output file.

Several output files are potentially generated by the annotations processing, depending on the analysis modules involved. One annotations output file contains all of the content of the GFF file generated by the modules, with added attributes for annotations. Another annotations output file is a subset of the previous file, and it contains only those entries that pass filtering requirements that you specified. Annotation processing also produces output files containing only mutated genes, or only SNPs. Several statistics files are also produced.

Two sources of annotation are used:

- A UCSC or ENSEMBL GTF file, used to determine whether a variant overlaps a gene or exon.
- The National Center for Biotechnology Information (NCBI) dbSNP database, which contains information on SNPs and indels already found by other studies. We use the following fields of the of the dbSNP entries:
 - The reference SNP identifier (rsID)
 - The functional code
 - The corresponding locus ID (the gene)

If a conflict is found between the dbSNP data and the GTF data, both annotations are reported. The annotation module does not attempt to resolve these conflicts.

Annotation sources can be either the refGene GTF file with respect to the human hg18 reference build and the dbSNP 130 build or the human hg19 reference build and the dbSNP 132 build (both available with LifeScope™ Software), or a source that you provide. Only GTF files and dbSNP data can be used as sources. The data from these three tables in dbSNP is required:

- SNPChrPosOnRef
- SNPContigLoc
- SNPContigLocusId

Annotations processing supports these annotation types:

- DNA features: genes and protein-coding features
- Verified variants in existing database

Post-processing types include:

- Annotate genomic variants with the annotation types that apply
- Filter based on annotation
- Report a list of gene and protein-coding annotations involved the variants
- Report statistics

The annotations module is available with the following modules:

- Resequencing:
 - SNPs
 - Small indels
 - Large indels
 - Human CNV
 - Inversions
- Targeted resequencing:
 - SNPs
 - Small indels

Memory requirement

For hg18 or hg19 sources, 15gb is required.

Input file handling

Annotations processing uses the GFF output of an analysis modules, but does not modify the GFF file. Annotations are added to a copy of the GFF file.

Filters

Annotation processing optionally generates an output file containing only those entries that fulfill *all* the conditions you specify with annotation filtering options. Annotation filters are described in the following table.

Filtering option	Default
Report <i>only</i> variants in exons	Off
Report <i>only</i> variants in genes	Off
Report <i>only</i> variants that appear in dbSNP	Off
Report <i>only</i> variants that do not appear in dbSNP	Off
Report <i>all</i> variants, whether or not they appear in dbSNP	On

The following table shows the annotation filtering supported in select LifeScope™ Software modules.

Annotation type	SNP	CNV	Small indel	Large indel	Inversion
geneID	Yes	Yes	Yes	Yes	Yes
exonID	Yes	Yes	Yes	Yes	Yes
rsID	Yes	—	Yes	—	—
functionCode	Yes	—	—	—	—

Statistics

The following table lists the statistics included in the annotations output files and the LifeScope™ Genomic Analysis Software modules that support these annotations. See the following table for information about SNPs transitions and transversions.

Analysis type	Statistics
All supported modules (SNPs, CNVs, Indels)	Number of variants Number of variants per chromosome Number of heterozygous variants Number of homozygous variants Number of heterozygous SNPs per chromosome Number of homozygous SNPs per chromosome
SNPs	Number of heterozygous SNPs that are transitions, transversions Number of homozygous SNPs that are transitions, transversions (compared to the reference)
Indels	Indel variant length distribution (negative for deletion, positive for insertion)
CNVs	Copy number distribution CNV length distribution
Annotations from dbSNPs	Number of SNPs or indels in dbSNP Number of homozygous SNPs or indels in dbSNP Number of heterozygous SNPs or indels in dbSNP Overall dbSNP concordance (percentage of SNPs or indels in dbSNP) Heterozygous dbSNP concordance (the percentage of heterozygous SNPs or indels found in dbSNP) Homozygous dbSNP concordance (the percentage of homozygous SNPs or indels found in dbSNP)
Annotations from GTF file content	Number of variants in exons, and the percentage of exons that are variant Number of heterozygous variants in exons, and their percentage Number of homozygous variants in exons, and their percentage Number of variants in genes, and the percentage that are variant Number of heterozygous variants in genes, and their percentage Number of homozygous variants in genes, and their percentage

The following table describes how transitions and transversions are determined based on the reference allele and the SNP Finding output genotype.

SNP Finding genotype	Reference allele second allele	A	C	G	T	Other
A	A	0	TV	TS	TV	0
C	C	TV	0	TV	TS	0
G	G	TS	TV	0	TV	0
T	T	TV	TS	TV	0	0
M	A C	TV	TV	TV	TV	TV
R	A G	TS	TS	TS	TS	TS
W	A T	TV	TV	TV	TV	TV
S	C G	TV	TV	TV	TV	TV
Y	C T	TS	TS	TS	TS	TS
K	T G	TV	TV	TV	TV	TV
	Other	0	0	0	0	0

Workflows

Annotation is integrated into the resequencing and targeted resequencing standard workflows. The table shows the modules with integrated annotation support, in these workflows.

Workflow	Library type	SNP	CNV	Small indel	Large indel
Genomic resequencing	Fragment	Yes	Yes	Yes	—
Genomic resequencing	Mate-pair	Yes	Yes	Yes	Yes
Genomic resequencing	Paired-end	Yes	Yes	Yes	Yes
Targeted resequencing	Fragment	Yes	—	Yes	—
Targeted resequencing	Paired-end	Yes	—	Yes	—

Annotation sources

The table lists common annotation sources supported by LifeScope™ Software. Files required to support certain annotations are available with LifeScope™ Software. These files support the following:

- Annotations based on the hg18 and the dbSNP 130 build
- Annotations based on the hg19 and the dbSNP 132 build

Note: The dbSNP 132 file used is a VCF file containing 1000 Genome data.

Type of annotation	Location and notes
Variants in existing databases	NCBI Variation dbSNP resources: includes single nucleotide polymorphisms, microsatellites, and small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral polymorphisms and clinical mutations. (This text is taken from NCBI website http://www.ncbi.nlm.nih.gov/guide/variation .) http://www.ncbi.nlm.nih.gov/snp
Genes and protein-coding features Note: Supported by LifeScope™ Software only if the data is downloaded in GTF format.	RefGene data from UCSC: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz The European Bioinformatics Institute's Alternative Splicing Database Project: http://www.ebi.ac.uk/asd

Note: While every effort is made to provide correct web addresses, Applied Biosystems is not responsible for the changes to these addresses, nor for the data provided by these locations.

Annotation parameters

This section describes the annotation parameters used with LifeScope™ Software analyses. There are three categories of Annotation Statistics and Reporting options: General, Sources, and Filtering.

Annotation Statistics and Reporting General Options

Parameters	Default value	Description
Input GFF3 file	<code>\${analysis.output.dir}/small.indel/ \${analysis.sample.name}/ \${analysis.name}.gff3</code>	The input variant GFF file, generated by a LifeScope™ Software analysis module (see “ Show only variants in genes ” on page 275). Expected file extensions: .gff, .gff3. Example: /data//HuRef_chr22/DB_out/ HuRef_rmm13_SNP.gff3 Must be a valid path, and the file must be readable. This file may be the output of an earlier version of the software.
Output directory	<code>\${analysis.output.dir}/small.indel/ \${analysis.sample.name}</code>	Path to the output directory. ASF writes all output files to this directory.

Annotation Statistics and Reporting Sources Options

Parameters	Default value	Description
dBSNP concordance for dbSNP SNPs	False	Annotate variants with SNP entries from the dbSNP file. <ul style="list-style-type: none"> • True: Do not annotate variants with SNP entries from the dbSNP file. • False: Annotate variants with SNP entries from the dbSNP file.
dBSNP concordance for dbSNP indels	True	Annotate variants with indel entries from the dbSNP file. <ul style="list-style-type: none"> • True: Do not annotate variants with indel entries from the dbSNP file. • False: Annotate variants with indel entries from the dbSNP file.
dBSNP indel border slack	5	An indel is considered to be matched to a dbSNP indel if the two overlap or if the distance between them is less than or equal to the value of <code>annotation.indel.border.slack</code> . Allowed values: Integers ≥ 0

Annotation Statistics and Reporting Filtering Options

Parameters	Default value	Description
Show only coding variants	False	When set to true, filters the annotated output: restricts the Filtered Variant annotation output file to only the variants that overlap any exon. <ul style="list-style-type: none"> • True: Restrict output to variants that overlap any exon. • False: Do not filter the output for variants that overlap any exon. Note: If set to true, then the setting for the parameter Show only variants in genes is ignored.
Show only variants in genes [‡]		When True, filters the annotated output: restricts the Filtered Variant annotation output file to only the variants that overlap any gene, even if it does not overlap any exon. <ul style="list-style-type: none"> • True: Restrict output to variants that overlap any gene. • False: Do not filter the output for variants that overlap any gene. Note: If the parameter Show only coding variants is set to true, then the setting for this parameter is ignored.
Show only variants in dbSNP [‡]		When set to true, filters the annotated output: restricts the Filtered Variant annotation output file to only SNPs or small indels that exist in dbSNP. <ul style="list-style-type: none"> • True: Restrict output to SNPs or small indels variants that <i>appear</i> in dbSNP. • False: Do not filter the output for variants that <i>appear</i> in dbSNP.
Show only variant not in dbSNP [‡]		When set to true, filters the annotated output: restricts the Filtered Variant annotation output file to only SNPs or small indels that do not exist in dbSNP. <ul style="list-style-type: none"> • True: Restrict output to SNPs or small indels variants that <i>do not appear</i> in dbSNP. • False: Do not filter the output variants that <i>do not appear</i> in dbSNP.

[‡] Do not use for CNV analysis.

The following table describes function codes for refSNPs in gene features. This table is taken from the dbSNP website:

<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch5>

Functional class	Description	Database code
Locus region	Variation is within 2 kbp 5' or 500 bp 3' of a gene feature (on either strand), but the variation is not in the transcript for the gene. This class is indicated with an "L" in graphical summaries. <p>Note: As of build 127, function code 1 has been modified into a <i>two-digit code</i> that more precisely indicates the location of a SNP. The two-digit code has function code 1 as the first digit, which keeps the meaning as described above, and 3 or 5 as the second digit, which indicates whether the SNP is 3' or 5' to the region of interest. See function codes 13 and 15 in this table.</p>	1

Functional class	Description	Database code
Coding	Variation is in the coding region of the gene. This class is assigned if the allele-specific class is unknown. This class is indicated with a 'C' in graphical summaries. Note: This code was <i>retired</i> as of dbSNP build 127.	2
Coding-synon	The variation allele is synonymous with the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon makes no change to the amino acid specified by the reference sequence. A variation is a synonymous substitution if all alleles are classified as contig reference or coding-synon. This class is indicated with a 'C' in graphical summaries.	3
Coding-nonsynon	The variation allele is nonsynonymous for the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon changes the amino acid specified by the reference sequence. A variation is a nonsynonymous substitution if any alleles are classified as coding-nonsynon. This class is indicated with a "C" or "N" in graphical summaries. Note: As of build 128, function code 4 has been modified into a <i>two-digit code</i> that more precisely indicates the nonsynonymous nature of the SNP. The two-digit code has function code 4 as the first digit, which keeps its original meaning, and 1, 2, or 4 as the second digit, which indicates whether the SNP is nonsense, missense, or frameshift. See function codes 41, 42, and 44 in this table.	4
mRNA-UTR	The variation is in the transcript of a gene but not in the coding region of the transcript. This class is indicated by a "T" in graphical summaries. Note: As of build 127, function code 5 has been modified into a <i>two-digit code</i> that more precisely indicates the location of a SNP. The two digit code has function code 5 as the first digit, which keeps its original meaning, and 3 or 5 as the second digit, which indicates whether the SNP is 3' or 5' to the region of interest. See function codes 53 and 55 in this table.	5
Intron	The variation is in the intron of a gene but not in the first two or last two bases of the intron. This class is indicated by an "L" in graphical summaries.	6
Splice-site	The variation is in the first two or last two bases of the intron. This class is indicated by a "T" in graphical summaries. Note: As of build 127, function code 7 has been modified into a <i>two-digit code</i> that more precisely indicates the location of a SNP. The two-digit code has function codes 7 as the first digit, which keeps its original meaning, and 3 or 5 as the second digit, which indicates whether the SNP is 3' or 5' to the region of interest. See function codes 73 and 75 in this table.	7
Contig-reference	The variation allele is identical to the contig nucleotide. Typically, one allele of a variation is the same as the reference genome. The letter used to indicate the variation is a "C" or "N", depending on the state of the alternative allele for the variation.	8
Coding-exception	The variation is in the coding region of a gene, but the precise location cannot be resolved because of an error in the alignment of the exon. The class is indicated by a "C" in graphical summaries.	9
NearGene-3	Function Code 13, where: 1: locus region (see function code 1 in this table) 3: SNP is 3' to and 0.5 kbp away from gene	13

Functional class	Description	Database code
NearGene-5	Function Code 15, where: 1: locus region (see function code 1 in this table) 3: SNP is 5' to and 0.5 kbp away from gene	15
Coding-nonsynonymous sense	Function Code 41, where: 4: Coding-nonsynonymous (see function code 4 in this table) 1: Nonsense (changes to the Stop codon)	41
Coding-nonsynonymous missense	Function Code 42, where: 4: Coding-nonsynonymous (see function code 4 in this table) 2: missense (alters codon to make an altered amino acid in the protein product)	42
Coding-nonsynonymous frameshift	Function Code 44, where: 4: Coding-nonsynonymous (see function code 4 in this table) 4 (as the second digit): frameshift (alters codon to make an altered amino acid in protein product)	44
UTR-3	Function code 53, where: 5: UTR (untranslated region: see function code 5 in this table) 3: SNP located in the 3' untranslated region	53
UTR-5	Function code 55, where: 5: UTR (untranslated region: see function code 5 in this table) 5 (as the second digit): SNP located in the 5' untranslated region	55
Splice-3	Function code 73, where: 7: splice site (see function code 7 in this table) 3: 3' acceptor dinucleotide	73
Splice-5	Function code 75, where: 7: splice site (see function code 7 in this table) 5: 5' donor dinucleotide	75

Variant Statistics output file for SNPs

A Variant Statistics file is a text file containing the statistics about the variants in the input GFF file.

The following is an example of the content of a Variant Statistics file for SNPs:

```
*****
LIFESCOPE REPORT
*****

Input File:/data/swdev/minita_test/
solid0064_20101210_MP_2x60_T3_1/2.0/color_lmp/outputs/
SNP.Finding/color_lmp_SNP.gff3
Date:Apr 14, 2011 2:17:54 PM
Annotation file, dbSNP:/data/analysis/LifeScope_resources/
lifetech/hg18/dbSNP/dbSNP_b130.tab
```

Annotation file, Genes and Exons, GTF:/data/analysis/
LifeScope_resources//lifetech/hg18/GTF/refGene.20090513.gtf

Statistics Overview

-- Basic Statistics -----

Number of variants	3072881
Number of heterozygous variants	1639649
Number of homozygous variants	1433232

-- Variant-Specific Statistics (SNP) -----

Number of transition SNPs	2070418
Number of transition heterozygous SNPs	1103589
Number of transition homozygous SNPs	966829
Number of transversion SNPs	1002463
Number of transversion heterozygous SNPs	536060
Number of transversion homozygous SNPs	466403
Transition:Transversion ratio	2.065 : 1.000

-- dbSNP Annotation Statistics -----

Number of SNPs in dbSNP	2913081
Number of heterozygous SNPs in dbSNP	1506740
Number of homozygous SNPs in dbSNP	1406341
dbSNP concordance	94.80%
dbSNP heterozygous concordance	91.89%
dbSNP homozygous concordance	98.12%

-- Per-Chromosome Statistics -----

nVar: Number of variants
nHetVar: Number of heterozygous variants
nHomVar: Number of homozygous variants

contig	nVar	nHetVar	nHomVar
chr1	233502	125860	107642
chr2	247682	132717	114965
chr3	210137	115661	94476
chr4	225996	118326	107670
chr5	185137	102825	82312
chr6	191793	104138	87655
chr7	170500	94092	76408
chr8	160126	86553	73573
chr9	128205	70291	57914
chr10	157345	85408	71937
chr11	154271	80792	73479
chr12	139569	72915	66654
chr13	118842	60095	58747
chr14	93526	48742	44784

chr15	90141	46673	43468
chr16	95819	53400	42419
chr17	79447	44311	35136
chr18	85824	43117	42707
chr19	59142	31904	27238
chr20	65136	38340	26796
chr21	45656	26098	19558
chr22	36633	20355	16278
chrX	76085	18909	57176
chrY	22328	18127	4201
chrM	39	0	39

**Variant Statistics
output file for
small indels**

A Variant Statistics file is a text file containing the statistics about the variants in the input GFF file.

The following is an example of the content of a Variant Statistics file for small indels:

```
*****
LIFESCOPE REPORT
*****

Input File:/data/swdev/minita_test/
solid0064_20101210_MP_2x60_T3_1/2.0/color_lmp/outputs/
small.indel/color_lmp.gff3
Date:Apr 14, 2011 10:01:01 AM
Annotation file, dbSNP:/data/analysis/LifeScope_resources/
lifetech/hg18/dbSNP/dbSNP_b130.tab
Annotation file, Genes and Exons, GTF:/data/analysis/
LifeScope_resources//lifetech/hg18/GTF/refGene.20090513.gtf

*****
Statistics Overview
*****

-- Basic Statistics -----

Number of variants                296183
Number of heterozygous variants   194752
Number of homozygous variants    100934

-- Variant-Specific Statistics (InDel) -----
Distribution of InDel length

Length(bases)                    Number
(-Inf:-500]                      0
(-500:-490]                      14
(-490:-480]                      8
(-480:-470]                      13
(-470:-460]                      21
(-460:-450]                      14
(-450:-440]                      18
(-440:-430]                      17
(-430:-420]                      18
```

(-420:-410]	18
(-410:-400]	15
(-400:-390]	24
(-390:-380]	21
(-380:-370]	24
(-370:-360]	30
(-360:-350]	39
(-350:-340]	62
(-340:-330]	115
(-330:-320]	202
(-320:-310]	181
(-310:-300]	80
(-300:-290]	43
(-290:-280]	52
(-280:-270]	42
(-270:-260]	56
(-260:-250]	49
(-250:-240]	35
(-240:-230]	39
(-230:-220]	35
(-220:-210]	33
(-210:-200]	37
(-200:-190]	34
(-190:-180]	30
(-180:-170]	40
(-170:-160]	49
(-160:-150]	42
(-150:-140]	55
(-140:-130]	78
(-130:-120]	55
(-120:-110]	68
(-110:-100]	51
(-100:-90]	60
(-90:-80]	61
(-80:-70]	94
(-70:-60]	107
(-60:-50]	115
(-50:-40]	229
(-40:-30]	332
(-30:-20]	1064
(-20:-19]	220
(-19:-18]	392
(-18:-17]	268
(-17:-16]	568
(-16:-15]	436
(-15:-14]	624
(-14:-13]	541
(-13:-12]	1126
(-12:-11]	1009
(-11:-10]	1636
(-10:-9]	1276
(-9:-8]	2367
(-8:-7]	1566
(-7:-6]	3498

(-6:-5]	5458
(-5:-4]	15589
(-4:-3]	11836
(-3:-2]	22210
(-2:-1]	77896
(-1:0]	0
(0:1]	65663
(1:2]	19317
(2:3]	14044
(3:4]	25082
(4:5]	3749
(5:6]	2199
(6:7]	1011
(7:8]	1596
(8:9]	892
(9:10]	1169
(10:11]	705
(11:12]	1289
(12:13]	803
(13:14]	1264
(14:15]	3547
(15:16]	192
(16:17]	164
(17:18]	123
(18:19]	141
(19:20]	268
(20:30]	530
(30:+Inf)	0

-- dbSNP Annotation Statistics -----

Number of InDels in dbSNP	246619
Number of heterozygous InDels in dbSNP	150203
Number of homozygous InDels in dbSNP	96105
dbSNP concordance	83.27%
dbSNP heterozygous concordance	77.13%
dbSNP homozygous concordance	95.22%

-- Per-Chromosome Statistics -----

nVar: Number of variants
nHetVar: Number of heterozygous variants
nHomVar: Number of homozygous variants

contig	nVar	nHetVar	nHomVar
chr1	22753	15323	7385
chr2	24215	16211	7975
chr3	20327	13613	6686
chr4	22322	14798	7485
chr5	18224	12212	5989
chr6	19046	12807	6213
chr7	16887	11032	5824
chr8	14530	9625	4889

chr9	11539	7926	3601
chr10	14864	9836	5000
chr11	14951	9502	5428
chr12	14000	9180	4793
chr13	11885	7685	4184
chr14	9439	6105	3323
chr15	8945	5912	3022
chr16	7705	5304	2382
chr17	7849	5323	2508
chr18	8605	5530	3066
chr19	5866	3825	2026
chr20	6107	4120	1977
chr21	5067	2977	2072
chr22	3587	2301	1280
chrX	6351	2650	3675
chrY	1118	954	151
chrM	1	1	0

Variant Statistics output file for large indels

A Variant Statistics file is a text file containing the statistics about the variants in the input GFF file.

The following is an example of the content of a Variant Statistics file for large indels:

```
*****
LIFESCOPE REPORT
*****

Input File:/data/swdev/minita_test/
solid0064_20101210_MP_2x60_T3_1/2.0/color_lmp/outputs/
large.indel/large-indels.gff3
Date:Apr 14, 2011 10:00:05 AM
Annotation file, dbSNP:/data/analysis/LifeScope_resources/
lifetech/hg18/dbSNP/dbSNP_b130.tab
Annotation file, Genes and Exons, GTF:/data/analysis/
LifeScope_resources//lifetech/hg18/GTF/refGene.20090513.gtf

*****
Statistics Overview
*****

-- Basic Statistics -----

Number of variants                21
Number of heterozygous variants   2
Number of homozygous variants    19

-- Variant-Specific Statistics (InDel) -----
Distribution of InDel length

Length(bases)                    Number
(-Inf:-500000]                   0
(-500000:-100000]                0
(-100000:-50000]                 0
```

```
(-50000:-20000]          0
(-20000:-10000]         0
(-10000:-5000]          0
(-5000:-1000]           0
(-1000:-500]            0
(-500:-200]             7
(-200:-100]             5
(-100:0]                2
(0:100]                 2
(100:200]               4
(200:500]               1
(500:1000]              0
(1000:5000]             0
(5000:10000]            0
(10000:20000]           0
(20000:50000]           0
(50000:100000]          0
(100000:500000]         0
(500000:+Inf)           0
```

-- dbSNP Annotation Statistics -----

```
Number of                InDels in dbSNP          0
Number of heterozygous InDels in dbSNP          0
Number of homozygous InDels in dbSNP           0
dbSNP concordance                                0.00%
dbSNP heterozygous concordance                   0.00%
dbSNP homozygous concordance                     0.00%
```

-- Per-Chromosome Statistics -----

```
nVar: Number of variants
nHetVar: Number of heterozygous variants
nHomVar: Number of homozygous variants
```

contig	nVar	nHetVar	nHomVar
chr2	1	0	1
chr9	1	0	1
chr10	7	2	5
chr18	4	0	4
chr19	2	0	2
chrY	4	0	4
chrM	2	0	2

Variant Statistics output file for CNVs

A Variant Statistics file is a text file containing the statistics about the variants in the input GFF file.

The following is an example of the content of a Variant Statistics file for CNVs:

```
*****
LIFESCOPE REPORT
*****
```

```

Input File:/data/swdev/minita_test/
solid0064_20101210_MP_2x60_T3_1/2.0/color_lmp/outputs/cnv/
OutputCNVs.gff3
Date:Apr 14, 2011 9:58:12 AM
Annotation file, dbSNP:/data/analysis/LifeScope_resources/
lifetech/hg18/dbSNP/dbSNP_b130.tab
Annotation file, Genes and Exons, GTF:/data/analysis/
LifeScope_resources//lifetech/hg18/GTF/refGene.20090513.gtf

```

```

*****
Statistics Overview
*****

```

```

-- Basic Statistics -----

```

```

Number of variants                                     120

```

```

-- Variant-Specific Statistics (CNV) -----

```

```

Distribution of CNV copy number

```

Copy_Number	Number
0	5
1	20
2	23
3	20
4	22
5	8
6	9
7	1
8	2
9	10
>9	0

```

Distribution of CNV length

```

Length(bases)	Number
(0:1000]	0
(1000:2000]	0
(2000:5000]	0
(5000:10000]	0
(10000:20000]	41
(20000:50000]	54
(50000:100000]	11
(100000:500000]	10
(500000:+Inf)	4

```

-- Per-Chromosome Statistics -----

```

```

nVar: Number of variants

```

contig	nVar
chr1	7
chr2	5

chr3	11
chr4	4
chr5	7
chr6	11
chr7	4
chr8	5
chr9	3
chr10	3
chr11	8
chr12	4
chr13	3
chr14	2
chr15	4
chr16	2
chr17	4
chr19	4
chr20	1
chr22	2
chrX	17
chrY	9

Mutated Genes output file

The Mutated Genes output file contains the list of genes whose coding regions were modified by at least one of the variants in the input file. The Mutated Genes file is a text file in tab-separated format. The file contains an entry per row, and one gene per entry. The first 12 columns are the same as the columns in the BED file (*genes.bed*). Additional columns are added by the annotations processing.

The following table lists the information included in the Mutated Genes output file.

Field number	Field name	Description
1	Chromosome/Contig	The contig of the gene.
2	Start	The left-most coordinate of the gene.
3	End	The right-most coordinate of the gene.
4	GeneID	The gene identifier as provided by the GTF file.
5	Score	A score between 0 and 1000. The score is a normalized value for the number of coding variants such that all numbers scale between 0 and 1000. The score is equal to the number of coding variants for this entry divided by the maximum number of coding variants for any gene.
6	Strand	The strand of the gene.
7	Start	The left-most coordinate of the gene.
8	End	The right-most coordinate of the gene.
9	Variant color	The color depends on the type of the variant that overlaps this gene: Red: SNP Green: Indel (small indels, large indels) Cyan: CNV
10	Number of exons	The number of exons of the gene.

Field number	Field name	Description
11	Exon start position list	A comma-separated list of start-coordinates for exons in the gene. All coordinates are relative to the start position of the gene.
12	Exon length list	A comma-separated list of the lengths of the exons in the gene.
13	Total exon length	The sum the length of all exons.
14	Number of coding variants	The number of variants overlapping at least one exon.
15	Number of variants	The number of variants overlapping the gene.
16	Variant type	The type of the variant. (One type is listed per line.)
17	Variant start list	A comma-separated list for start coordinates of variants overlapping the gene. All coordinates are relative to the start position of the gene.
18	Variant length list	A comma-separated list of lengths of the variants overlapping the gene.

Potential uses of a Mutated Genes file

You can use a spreadsheet application or scripts to remove unwanted columns. First save a backup of your output file. The following are examples of uses for a Mutated Genes file:

- The first 12 columns provide information for visualization of modified genes and exons.
- Column 4, the GeneID, provides a list of modified genes. This column provides a gene list to focus on for enrichment analysis.
- A combination of the first 10 and the last 2 fields provides information useful for visualizing variants that overlap the gene.
- Fields 14 and 15 provide the number of variants that overlap exons and the number of variants that overlap the gene. With these columns you can filter by the number of variants overlapping the exons in the gene. With the GeneID in column 4, this information identifies genes to focus on for further analysis.

SNP Finding tab-delimited output file

The first eight columns of this output file are the same as the input GFF file. The ninth column in the input file is replaced by individual columns for each attribute, as shown in the table [on page 286](#). This output file is generated only on SNPs module runs.

Attribute	Description
Seqid	The string ID of the sequence to which the start and end coordinates refer.
Source	The source of the data.
Type	Sequence ontology derived type for this variation. For SNP Finding, this is always SNP.
Start	Start position of the SNP.
End	End position of the SNP.
Score	Calculated p-value of the SNP.
Strand	Not used.
Phase	Not used.

Attribute	Description
Genotype	Genotype in the form of IUB codes for bases observed in all the reads. The base of the reference sequence at the current position.
Coverage	The number of the reads that cover the current position.
refAlleleCounts	The number of reads of the reference allele at the current position.
refAlleleStarts	The number of different start positions of reads having the reference allele at the current position.
refAlleleMeanQV	The mean of quality values of all reference allele reads at the current position.
novelAlleleCount	The number of reads of the most abundant non-reference allele at the current position.
novelAlleleStarts	The number of different start positions of reads having the most abundant non-reference allele at the current position.
novelAlleleMeanQV	The mean of quality values of all novel allele reads.
mostAbundantAlleleDiColor2	The most abundant allele in the reads (not necessarily the reference allele) in dicolor encoding (for example, 00, 01, ... 32, 33), of 16 possible dicolors.
secondAbundantAlleleDiColor3	The second most abundant allele in the reads.
Het	Heterozygosity flag. Allowed values: 0,1 <ul style="list-style-type: none"> • 0: Homozygous SNP • 1: Heterozygous SNP

The following is a truncated example of SNP Finding tab-delimited output file content:

```
Seqid Source Type Start End Score Strand Phase genotype
reference coverage refAlleleCounts refAlleleStarts
refAlleleMeanQV novelAlleleCounts novelAlleleStarts
novelAlleleMeanQV mostAbundantAlleleDiColor2
secondAbundantAlleleDiColor3 het
chr1 SOLiD_SNP Finding SNP 10007 10007 0.806886 . . A
G 31 1 16 2 2 25 21 21 0
chr1 SOLiD_SNP Finding SNP 224472 224472 0.879526 . . C
T 3 1 1 26 2 2 24 10 10 0
chr1 SOLiD_SNP Finding SNP 553488 553488 0.0625 . . C
T 3 0 0 0 2 2 25 20 20 0
chr1 SOLiD_SNP Finding SNP 556955 556955 0.095335 . . Y
T 10 7 6 14 2 2 26 13 31 1
chr1 SOLiD_SNP Finding SNP 558326 558326 0.093729 . . R
A 8 6 6 26 2 2 22 03 21 1
chr1 SOLiD_SNP Finding SNP 558554 558554 0.0761 . . Y
T 1411 9 20 3 3 22 03 21 1
```

SNP Finding annotated tab-delimited output file

This output file is the same as the SNP Finding tab-delimited output file with four additional columns added. As with the SNP Finding tab-delimited output file, the first eight columns of this output file are the same as the input file. The ninth column in the input file is replaced by individual columns for each attribute, as shown in the following table. This output file is generated only on SNPs module runs.

Field	Description
geneID	The gene id as provided by the GTF file.

Field	Description
exonID	The exon id formed by concatenating the GeneID and the exon index number in the list of exons of the gene sorted in transcription order. Transcription order ties are broken listing by the shortest exon first (<GeneID>-<exon #>).
rsID	The dbSNP id of the SNP.
functionCode	The dbSNP functional code.

The following is a truncated example of SNP Finding annotated tab-delimited output file content:

```
Seqid Source Type Start End Score Strand Phase genotype
reference coverage refAlleleCounts refAlleleStarts
refAlleleMeanQV novelAlleleCounts novelAlleleStarts
novelAlleleMeanQV mostAbundantAlleleDiColor2
secondAbundantAlleleDiColor3 het geneID exonID rsID functionCode
chr1 SOLiD_SNP Finding SNP 10007 10007 0.806886 . . A
G 3 1 1 16 2 2 25 21 21 0 WASH5P
chr1 SOLiD_SNP Finding SNP 224472 224472 0.879526 . . C
T 3 1 1 26 2 2 24 10 10 0
chr1 SOLiD_SNP Finding SNP 553488 553488 0.0625 . . C
T 3 0 0 0 2 2 25 20 20 0
chr1 SOLiD_SNP Finding SNP 556955 556955 0.095335 . . Y
T 10 7 6 14 2 2 26 13 31 1 9326622
chr1 SOLiD_SNP Finding SNP 558326 558326 0.093729 . . R
A 8 6 6 26 2 2 22 03 21 1 2153587
chr1 SOLiD_SNP Finding SNP 558554 558554 0.0761 . . Y
T 14 11 9 20 3 3 22 03 21 1 8179256
chr1 SOLiD_SNP Finding SNP 658055 658055 0.0625 . . T
C 3 0 0 0 2 2 22 20 20 0
```

SNP Finding filtered annotated tab-delimited output file

This file is the same as the SNP Finding annotated tab-delimited output file, except that it contains only those entries that fulfill all the conditions set by the your annotation filtering settings (described in [“Annotation parameters” on page 273](#)).

This output file is generated only on SNPs module runs.

The following is a truncated example of SNP Finding filtered annotated tab-delimited output file content:

```
Seqid Source Type Start End Score Strand Phase genotype
reference coverage refAlleleCounts refAlleleStarts
refAlleleMeanQV novelAlleleCounts novelAlleleStarts
novelAlleleMeanQV mostAbundantAlleleDiColor2
secondAbundantAlleleDiColor3 het geneID exonID rsID functionCode
chr1 SOLiD_SNP Finding SNP 556955 556955 0.095335 . . Y
T 10 7 6 14 2 2 26 13 31 1 9326622
chr1 SOLiD_SNP Finding SNP 558326 558326 0.093729 . . R
A 8 6 6 26 2 2 22 03 21 1 2153587
chr1 SOLiD_SNP Finding SNP 558554 558554 0.0761 . . Y
T 14 11 9 20 3 3 22 03 21 1 8179256
chr1 SOLiD_SNP Finding SNP 708249 708249 0.003906 . . G
A 5 0 0 0 3 3 23 30 30 0 10900602
chr1 SOLiD_SNP Finding SNP 708418 708418 0.0625 . . C
T 2 0 0 0 2 2 15 13 13 0 10751453
```



```
chr1 SOLiD_SNP Finding SNP 710103 710103 0.003906 . . C
T 3 0 0 0 3 3 17 01 01 0 3121393
chr1 SOLiD_SNP Finding SNP 713754 713754 0.003906 . . C
G 4 0 0 0 3 3 21 33 33 0 2977670
```

About annotations and LifeScope™ Software modules

This section describes which annotation functionality is available with the various LifeScope™ Software modules.

The following table shows the LifeScope™ Software modules that support the annotation attributes.

Label	SNPs	CNV	Small indel	Large indel
geneID, exonID	Yes	Yes	Yes	Yes
rsID - SNPs	Yes	—	—	—
rsID - Indels	—	—	Yes	—
functionCode	Yes	—	Yes	—

The following table lists the LifeScope™ Software modules that support annotation filtering.

Label	SNPs	CNV	Small indel	Large indel
Only in exons	Yes	Yes	Yes	Yes
Only in genes	Yes	Yes	Yes	Yes
Only <i>not</i> in dbSNP	Yes	—	Yes	—
Only in dbSNP	Yes	—	Yes	—

The following table lists the LifeScope™ Software modules that support for the various annotation statistics. Heterozygous and homozygous statistics are included for applicable variants.

Label	SNPs	CNV	Small indel	Large indel
Number of variants (total and per chromosome)	Yes	Yes	Yes	Yes
Transitions and transversions	Yes	—	—	—
Variant length distribution	—	Yes	Yes	Yes
Copy number distribution	—	Yes	—	—
dbSNP concordance	Yes	—	Yes	—
Overlapping exons (number and percent)	Yes	Yes	Yes	Yes
Overlapping genes (number and percent)	Yes	Yes	Yes	Yes

Examples of annotation output files

If the analysis was successful, click **View Results** in the Progress column.

In the View Results window, you can view additional file and logs of annotation output. For more information about viewing analysis results, see [Chapter 8, “View Analysis Results”](#) on page 87.

Annotation output includes only text files (.txt) and log files (.log). There are neither details nor statistics for annotation output.

Additional files

The following table is an example of a text file listed in Additional Files.

Directory	Files	File Type
SmallIndel.Annotation/ hg19_Simulated_TR_CS_BC96_PE50X25_320Mil_BC55	README.txt	.txt

You can download the README.txt file, open it, and view the contents, for example:

```
Outputs generated at '/datapod1/sitest1/results/projects/lifescopes/TRPEBCCS/TRPEBCCS/outputs/small.indel/hg19_Simulated_TR_CS_BC96_PE50X25_320Mil_BC55'
```

Logs

The following table is an example of a log file listed in Log Files.

Directory	Files	File Type
Small Indel Annotation	tertiary-hg19_Simulated_TR_CS_BC96_PE50X25_320Mil_BC55-SmallIndel.Annotation.20110503204415163.log	.log

You can download the .log file, save and open it to view details about the analysis run.

PART V
Appendices



File Format Descriptions and Data Uses

This appendix covers:

■ Introduction.....	293
■ XSQ file format	293
■ BAM headers in LifeScope™ Software	294
■ Color-space attributes	300
■ Pairing information in a BAM file	300
■ Hard clipping of incomplete extensions	301
■ BED file format	303
■ BEDGRAPH file format.....	304
■ GFF3 file format	305
■ Reference file data overview.....	305
■ Read-set file format	306
■ VCF file.....	308

Introduction

This appendix describes file formats used with LifeScope™ Genomic Analysis Software 2.0. It also describes how data is used in files.

Before reading the section about the SOLiD™ System BAM file contents, you should be familiar with the general SAM specification and with the SAM specification field definitions. You can see the SAM specification at:

<http://samtools.sourceforge.net>

LifeScope™ Software secondary analysis (mapping and pairing) now produces a BAM file as the main alignment format. Mate-pair and paired-end analysis directly produces a BAM file, while a single file conversion is needed for fragment libraries. Depending on the output filter selected, unmapped and secondary alignments can be included.

LifeScope™ Software supports the XSQ sequence data file format introduced with the 5500 Series SOLiD™ Sequencer.

XSQ file format

XSQ (eXtensible SeQuence) is an extensible file format for storing sequence data. The XSQ format supports multiple independent reads at the same position in a fragment. XSQ is a binary format based on the open HDF format.

Sequencing run data are automatically exported from the 5500 Series SOLiD™ Sequencer in XSQ binary file format. If an ECC primer round has been performed, the XSQ output also includes the sequence information in base space, in addition to color space.

XSQ file content overview

5500 Series SOLiD™ Sequencer data

XSQ files generated by the 5500 Series SOLiD™ Sequencer contain:

- **Base-space data** – Reference-free base-space data from primary analysis.
- **Color-space data** – Referred to as 2+4 color, as follows:
 - The 2BE (Base Encoding) data is the data for first 5 primer rounds of regular 2BE probes. This data is in the same format as a data created by a SOLiD™ 4 System non-ECC run.
 - The 4BE data is the data for the 6th primer round, the optional ECC round.

Data for a maximum of one optional ECC primer round is supported in an XSQ file.

Paired-end libraries do not support ECC data. Fragment and mate-pair libraries support ECC data on both tags.

SOLiD™ 4 System data

SOLiD™ 4 System CSFASTA and QUAL files can be processed by LifeScope™ Software, but first must be converted to XSQ format. Converted XSQ files contain only:

- Color-space 2BE data
- Color for both tags (one tag for fragment and both tags for mate-pair and paired-end data).

XSQ file format properties

For a description of XSQ file format properties, see the XSQ file format specification, available on the SOLiD™ Software Community website:

<http://solidsoftwaretools.com/gf/project/xsq/>

BAM headers in LifeScope™ Software

The BAM file generates all of the header information required by the SAM format specification, including @HD, @SQ, and @RG lines.

To view the content of the BAM file header, use the following command:

```
samtools view -H <BAMfilename>
```

Sequence dictionary (@SQ)

Sequence header lines include the reference file URL, for example, `file:///share/reference/genomes/hg18.fa` in the optional UR field of the reference file. This value might become invalid if you relocate files.

Read group (@RG)

Read groups receive an arbitrary ID and sample name. The library field (LB) contains information that is important to downstream algorithms that use pairing information. A library name, which is specified by the tool parameter `library.name`, and the library type, are separated by a dash in the LB field. The library type is a structured value that details the nominal length of the two tags and the protocol used. as shown in the following syntax example:

```
l1 (x12) [F|MP|RR|RRBC]
```

In the syntax example, *l1* is the nominal length of the first read and *l2* is the nominal length of the second read. There will only be one number for fragment libraries. The library types correspond to fragment (*F*), mate-pair (*MP*), reverse read (*RR*), and reverse read-bar coded (*RRBC*).

Detecting structural variations, particularly large insertions and deletions, depends on the statistically likely range of pairing insert (PI) sizes. The pairing tool generates the information about PI sizes. The information has been used in legacy file formats to define the three-letter pairing “category”, specifically the third letter. The PI field in the read group captures the range of pairing insert sizes with a range of the form shown in the following example:

```
PI;low-high
```

In the PI example, *low* is the lower bound of the pairing range, and *high* is the upper bound.

Header (@HD) sort order

The LifeScope™ Software SNPs module requires that BAM headers of its input files contain an @HD line with a SO sort-order field, even if the BAM file is sorted properly. The required header line is:

```
@HD VN:1.0 GO:none. SO:coordinate
```

This header contains the information that the BAM file is sorted by genomic coordinates. BFAST mapping, for example, by default sorts the BAM file by coordinates, but does not update the header to reflect the sort order.

This section describes how to update the @HD SO field. If your BAM file is sorted properly but does not contain the SO sort-order field, follow these instructions before using a BAM file generated outside of LifeScope™ Software with the SNPs module:

1. Write the BAM file header to a file

```
samtools view -H input.bam > input_header.sam
```
2. Edit the BAM header file (`input_header.sam` created in the previous step). Insert the following as the first line of the header:

```
@HD VN:1.0 GO:none. SO:coordinate
```

Make sure that the fields are tab-separated.

3. Update your BAM file with this new header. This command requires `samtools-0.1.8` or later.

```
samtools reheader input_header.sam input.bam
```

These instructions are not required with BAM files generated by LifeScope™ Software mapping modules.

XSQ metadata in BAM headers

LifeScope™ Software secondary analysis modules write metadata contained in XSQ reads files to their mapping output files, as metadata in BAM header comment (@CO) lines. The new metadata is not in the BAM specification, but BAM files with this metadata conform to the BAM specification. The metadata table on [page 296](#) lists the XSQ file fields with both the new @CO field names and the existing BAM header fields that contain the XSQ metadata.

References to existing BAM header lines

Each @CO metadata line with metadata information references one of the following types of BAM header lines:

- **Program group** – @PG lines capture information about the program that generated the data originally. The sequencing instrument and primary analysis software are considered to be the initial program. An example @PG line is:

```
@PG ID:Mordor_201103161921_0_3 {...}
```
- **Read group** – @RG lines contain information on library type and read length. (See “[Read group \(@RG\)](#)” on [page 295](#) for more information.) An example @RG line is:

```
@RG ID:1 SM:Stooges LB:Larry DS:"75x35PE" PI:225 {...}
```
- **General header** – @HD lines are general header lines.

All the attributes on the @CO metadata line then apply to referenced line.

The following example shows how @CO lines containing BAM metadata information refer to read group and program group lines. Except for @CO HD lines, the second field of the @CO metadata lines tie its attributes to either @RG or @PG lines. In the example, the RG:1 fields tie the metadata in those @CO lines to the @RG read group with an ID of 1. The PG:Mordor_201012161921_0_3 field ties the metadata in that @CO line to the @PG program group with an ID of Mordor_201012161921_0_3.

```
@CO RG:1 IA:215 IS:5.232 IN:200 IM:250
@CO RG:1 TN:50 TX:75 TB:0 TC:1
@CO RG:1 UN:25 UX:35 TB:0 TC:1
@CO RG:1 BX:0 EC:0
@CO PG:Mordor_201012161921_0_3 AS:"primary5500 1.0.7"
@CO PG:Mordor_201012161921_0_3 \
    CL:"/usr/bin/primary5500 <with full arguments>"
@CO HD UF:"f47ac10b-58cc-4372-a567-0e02b2c3d479"
```

Metadata table

The following table lists the existing BAM header fields that contain the XSQ metadata, the new @CO metadata field names, and the XSQ file fields. See also “[Input read count fields](#)” on [page 298](#).

Field name	Public BAM 1.3 spec field	New field in @CO line	XSQ field name	Optional or mandatory
Reference	@SQ SP	—	Species { 'Homo sapiens' 'Mus musculus' 'other' ...}	Mandatory
Assembly	@SQ AS	—	Assembly { 'hg18' 'hg19' 'mm9' ...}	Mandatory
—	—	@CO HD UF	FileUUID	Optional

Field name	Public BAM 1.3 spec field	New field in @CO line	XSQ field name	Optional or mandatory
Program	@PG ID	—	InstrumentName_Date_FlowcellAssignment_LaneNumber	Mandatory
Analysis Software	—	@CO PG:x AS	AnalysisSoftware	Optional
InstrumentSerial	—	@CO PG:x PS	InstrumentSerial	Optional
InstrumentName	—	@CO PG:x PN	InstrumentName	Optional
InstrumentVendor	—	@CO PG:x PV	InstrumentVendor	Optional
InstrumentModel	—	@CO PG:x PM	InstrumentModel	Optional
ReadGroupID	@RG ID	—	LibraryName or LibraryName_IndexName (if present)	Mandatory
—	—	@CO RG:x IX	IndexName (if present)	Optional
—	—	@CO RG:x II	IndexID (if present)	Optional
Sequencing Center	@RG CN	—	SequencingCenter	
Description	@RG DS	—	l1(xl2)[F MP RR RRBC]	
Library Description	—	@CO RG:x LD	LibraryDetails.Description	—
Library Type	—	@CO RG:x LT	LibraryType { 'MatePair' 'PairedEnd' 'Fragment' }	Mandatory
Application Type	—	@CO RG:x AT	ApplicationType { 'Whole Genome Resequencing' 'Targeted Resequencing' 'Whole Transcriptome (Fragment)' 'Whole Transcriptome (PairedEnd)' 'Small RNA' 'ChIP-Seq' 'Methylation' }	Mandatory
—	@RG DT	—	RunStartDate	Optional
—	—	@CO RG:x DE	RunEndTime	Optional
Library	@RG LB	—	LibraryName	Mandatory
Predicted Insert Size	@RG PI	—	(LibraryInsertSizeMin + LibraryInsertSizeMax) / 2	Conditional
Calculated Average Insert Size	—	@CO RG:x IA	<calculated>	Mandatory
Calculated Insert Size Std Dev	—	@CO RG:x IS	<calculated>	Mandatory
Min Insert Size from User Input	—	@CO RG:x IN	LibraryInsertSizeMinimum	Conditional
Max Insert Size from User Input	—	@CO RG:x IM	LibraryInsertSizeMaximum	Conditional
Input Read Count Passing All Filtering Steps	—	@CO RG:x CU	The sum across all ImageUnits of <i>ImageUnitID.Fragments.NumFragmentsPassed</i> (passes all filtering steps)	Optional
Input Read Count Total	—	@CO RG:x CT	The sum across all ImageUnits of <i>ImageUnitID.FragmentCount</i> (filtered and unfiltered) Not used in data converted from pre-5500 systems. Mandatory for 5500 data.	Mandatory
Specimen		@CO RG:x SP	SampleIdentifier	Optional

Field name	Public BAM 1.3 spec field	New field in @CO line	XSQ field name	Optional or mandatory
Platform name	@RG PL	—	"SOLID"	Mandatory
Lane Id	@RG PU	—	FlowcellAssignment_LaneNumber	Optional
Sample (Pool)	@RG SM	—	SequencingSampleName	Optional
—	—	@CO RG:x SD	SequencingSampleDescription	Optional
Tag1 Was Base Present In XSQ	—	@CO RG:x BX	Tag1.IsBasePresent	Mandatory
Tag1 Min Read Len	—	@CO RG:x TN	Tag1.MinTrimmedReadLength	Mandatory
Tag1 Max Read Len	—	@CO RG:x TX	Tag1.NumColorCalls or Tag1.NumBaseCalls	Mandatory
Tag2 Was Base Present In XSQ	—	@CO RG:x BY	Tag2.IsBasePresent	Mandatory for MP, PE. Optional for fragment
Tag2 Min Read Len	—	@CO RG:x UN	Tag2.MinTrimmedReadLength	
Tag2 Max Read Len	—	@CO RG:x UX	Tag2.NumColorCalls or Tag2.NumBaseCalls	
ECC run	—	@CO RG:x EC	<whether this was an ECC run (inferred)>	Mandatory
ERCC	—	@CO RG:x ER	ERCC	Optional
—	—	@CO RG:x CO	Operator	Optional
—	—	@CO RG:x UU	LibraryIndexUUID	Optional
—	—	@CO RG:x PN	Application	Optional
—	—	@CO RG:x PJ	ProjectName	Optional
—	—	@CO RG:x SO	SampleOwner	Optional

Input read count fields

This section describes aspects of the Input Read Count fields from the viewpoint of the sequencing instrument and the viewpoint of secondary analysis.

The Input Read Counts fields are:

- **CU** – Input Read Count Passing All Filtering Steps, @CO RG:x CU
- **CT** – Input Read Count Total, @CO RG:x CT

The number of *total beads*, before any filtering on the instrument, is given by these fields:

- **NumUnfilteredBeads** – In mapping statistics output
- **FragmentCount** – In an XSQ file
- **CT** – In BAM header metadata

The number of *beads that pass filtering* on the instrument is given by these fields:

- **NumFilteredBeads** – In mapping statistics output
- **NumFragmentsPassed** – In an XSQ file
- **CU** – In BAM header metadata

The number of *beads filtered out* on the instrument is given by:

$$\text{NumUnfilteredBeads} - \text{NumFilteredBeads}$$

Example of @CO metadata in a BAM header

The following is an example of a BAM header with an @CO comment line containing metadata information. The @CO line applies to the @RG ID:12345678 line below it. (The @SQ lines for chromosomes 2–22 are deleted.)

```
@HD VN:1.0 GO:noneSO:coordinate
@SQ SN:chr1 LN:247249719 UR:file:/ref/hg18.fasta
    SP:Homosapiens AS:hg19
...
@SQ SN:chrX LN:154913754 UR:file:/ref/hg18.fasta
    SP:Homosapiens AS:hg19
@SQ SN:chrY LN:57772954 UR:file:/ref/hg18.fasta
    SP:Homosapiens AS:hg19
@SQ SN:chrM LN:16571 UR:file:/ref/hg18.fasta
    SP:Homosapiens AS:hg19
@PG ID:SOLiD_12321_1
@CO RG:12345678 LT:MatePair AT:WholeGenomeResequencing
    LN:<50X50MP> PI:1500 IA:1500 IS:300 IN:1200 IM:1800
    PU:1234 SM:sample BX:0 TN:50 TX:50 BY:1 UN:50 UX:60 EC:0
@RG ID:12345678 DS:50x50MP LB:MP PI:1500 PU:1234 SM:sample
```

@CO syntax and rules

This section describes the syntax rules and conventions for @CO metadata lines. The @CO metadata rules are:

- The second field of each metadata @CO line either notes that the line is a general header (HD), or refers to a single @RG ID or @PG ID.
- Each metadata @CO line contains one or more tab-separated attributes.
- For a metadata @CO line that refers to either a @RG or @PG line, all of the @CO attributes must apply to the same @RG or @PG.
- The beginning of metadata @CO lines must match the regular expression “^@CO\t[PR]G:” and contain no other non-metadata text. LifeScope™ Software ignores and does not change @CO lines that do not match the regular expression.
- The @CO PG:*id* line describing the instrument is required and must refer to a valid @PG ID:*id* program group line. This @CO line contains information about the primary analysis done on the instrument.

According to the BAM specification, the @HD line is optional. If present, the @HD line is required to be the first line of the header. The order of all other header lines is not defined. The following order of @CO metadata lines is preferred when possible (but not required):

- Place the @CO HD:UF comment immediately after the @HD line.
- Place @CO RG: and @CO PG: comment lines before @CO lines for ease of visibility.
- Place @CO RG: and @CO PG: comment lines before the @RG and @PG blocks that the comment lines refer to.

Color-space attributes

The SAM format specification includes the attribute tags CS, CQ and CM. All BAM files support color-space reads (see the following table).

Attribute tag	Description
CS	Color-space (CS) read. The CS field contains the original color-space read, which includes the primer base, in the orientation of the CSFASTA file. CS entries are not manipulated to be top-strand relative.
CQ	Color qualities. Color qualities are encoded according to the ascii-33 scheme used for the QUAL field. The orientation is the same as the orientation used for the CSFASTA file.
CM	The number of color-space mismatches.

Pairing information in a BAM file

The BAM file that is produced by the pairing tool supports both mate-pair and paired-end protocols using the standard SAM format fields, in particular the ISIZE and FLAG fields.

Calculation of tag names

The paired libraries use tag names to refer to members of the pair. The mate-pair libraries use F3 and R3 as the tag names. The paired-end libraries use F3 and F5 as the tag names. Use the FLAG field and information from the LB field of the read group to recapitulate tag names (see the following table).

FLAG bit	Library type	Tag name
0x0040 (first read in a pair)	MP	F3
0x0080 (second read in a pair)	MP	R3
0x0040	RR	F3
0x0080	RR	F5
0x040	F	F3

Proper pairs

Legacy file formats, such as *.mates, and GFF, described pairs using a three-letter category. Pairs in the AAA category correspond to the “proper pair” concept in the SAM format. The pairs reflect pairings that are not altered by a structural variation such as an inversion or deletion (see the illustration [on page 301](#)). The BAM file field values for proper pairs are different for mate-pair and paired-end libraries:

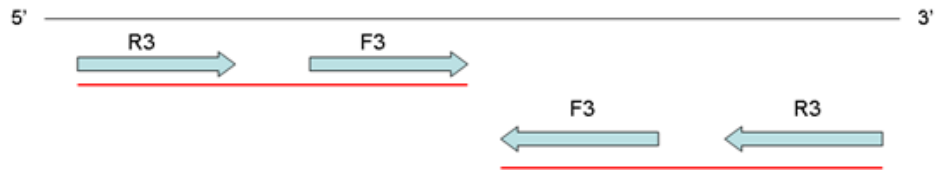
Mate-pair libraries

- Strand flag is equal for both mates (both 0 or both 1).
- ISIZE is between the lower and upper limit of the insert range.
- For forward strand hits R3 POS < F3 POS.
- For reverse strand hits F3 POS < R3 POS.

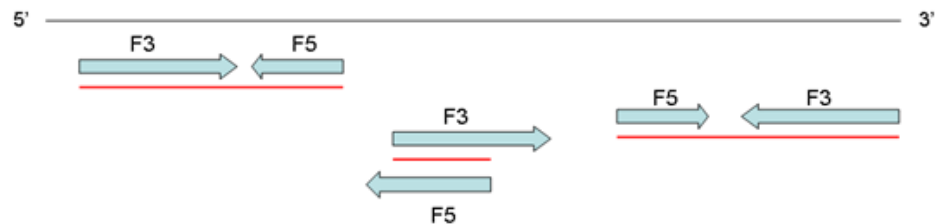
Paired-end

- Strand flag is opposite for the mates.
- ISIZE is between the lower and upper limit of the insert range. In the case of paired-end libraries, the ISIZE might be smaller than the sum of the alignment lengths.
- F3 POS < F5 POS if F3 is on the forward strand.
- F5 POS < F3 POS if F5 is on the forward strand.

Mate Pairs



Paired End



Single read mapping quality

As described in the SAM format specification, the MAPQ field for paired results contains a pairing quality value. Under some circumstances, it is valuable to include the original single-read alignment quality value. The original single-read alignment value is maintained in the SM:i attribute in BAM files.

Hard clipping of incomplete extensions

LifeScope™ Software mapping uses a seed-extend algorithm. The algorithm increases mapping throughput by matching a seed, usually 25 bp, and extending the alignment until mismatches drive down the alignment score. Many alignments do not completely cover the color-space read. Because the base-space sequence of color reads cannot be precisely known in the absence of alignment, incomplete extensions are represented as a hard-clip (H) operation in the BAM CIGAR string (see the following illustration).

T33232030301212311201322311232302131021221120112222

AAGGCCTCTGAACCCACTCAGGTA CT TAGCTGTAGATGGACATCAGTTAATTCGATGAC
 22221102112212013120323211322310211321210303023233T
 * * * _____

8H42M

The above illustration shows the read in normal orientation (see the top section) and aligned in reverse orientation to the reference top strand (see the middle section). The lines below the alignment show the extent of the seed (top horizontal line) and extension (bottom horizontal line) phases of mapping. The extension only results in 42 bases of alignment. The remaining portion of the color alignment has a number of mismatches that prevent extension. These are coded as hard-clipped regions. The CIGAR field in the BAM file is top-strand relative, so even though the hard clipping is on the end of the reversed color read, it is on the beginning of the CIGAR string.

Visualize BAM output

You can use third-party software visualization tools to view BAM files in a browser.

Integrative Genomics View (IGV)

The Integrative Genomics Viewer (IGV) available from the Broad Institute is a visualization tool for interactive exploration of large, integrated datasets. The IGV reads BAM files directly, which allows for easy viewing and inspection of alignments against the genome (see the illustration [on page 301](#)).

For more information, go to the following site:

www.broadinstitute.org/igv/

If you use IGV to visualize the BAM files, verify that the BAI file is present. The BAI file is the index that is built for BAM files and is a standard part of the public SAM specification. If the pairing and MaToBam tools do not automatically create the BAI file:

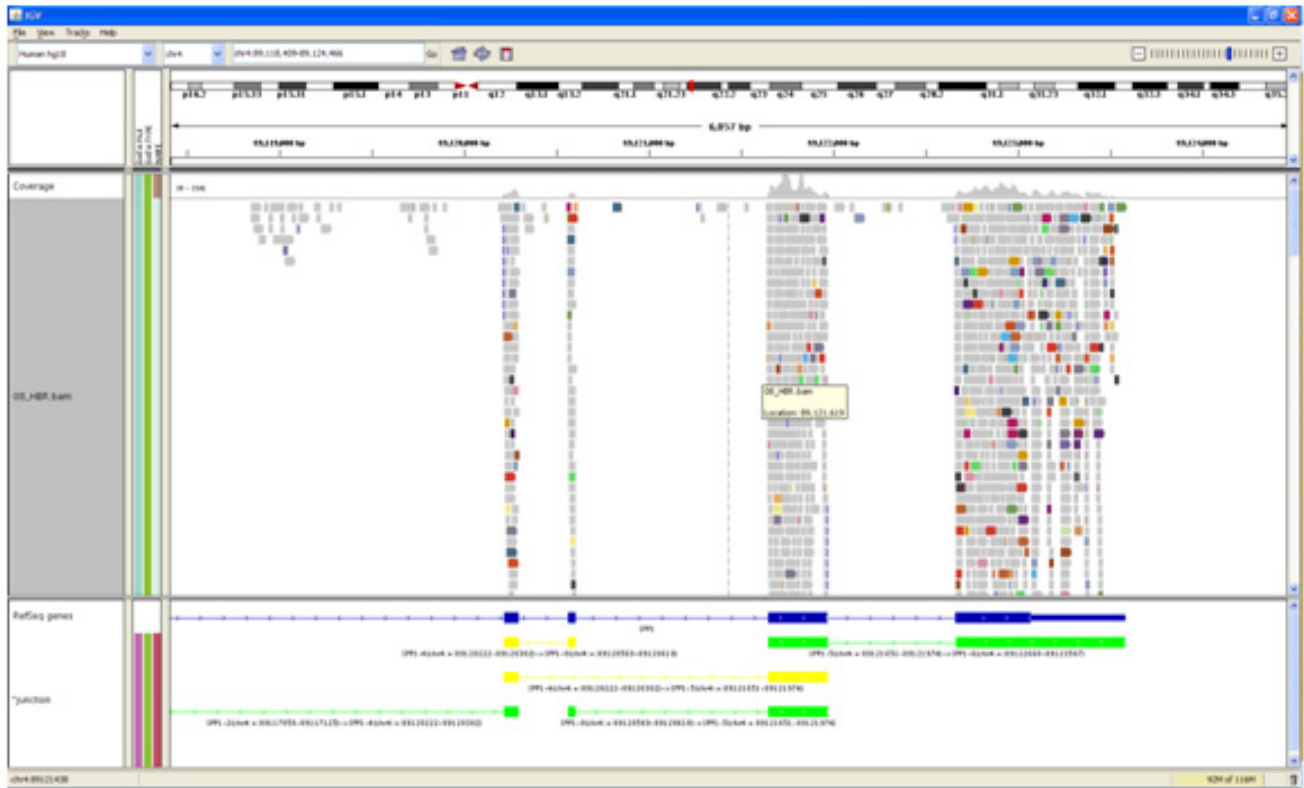
1. Log in to the LifeScope™ Software cluster.
2. At a command prompt, enter:

```
samtools index <bam file name>.bam
```

Indexing only works if the file is sorted in coordinate order. If the file is not sorted in coordinate order, at a command prompt, run the following command to sort the file in coordinate order:

```
samtools sort <unsorted bam name>.bam <sorted bam name>
```

The following illustration is an example of BAM file visualized in the IGV.



UC Santa Cruz (UCSC) genome browser

The UCSC Genome Browser serves as an interactive web-based microscope that allows researchers to view all 23 chromosomes of the human genome at any scale, from a full chromosome down to an individual nucleotide.

For more information, go to the Genome Browser website:

www.cbse.ucsc.edu/research/browser

BED file format

The Browser Extensible Display (BED) format was developed to extend the UCSC Genome Browser with user-defined tracks. BED is used to visualize the splice and fusion junctions in the UCSC Genome browser and in the IGV browser (see the illustration on page 304). For general documentation about the BED format, including information about all of the BED fields, go to:

genome.ucsc.edu/FAQ/FAQformat.html

For information about visualization software, see “Pairing information in a BAM file” on page 300.

Each line in the track defines a junction where chromStart is the smaller of the coordinates and chromEnd is the greater.

There are two blocks because a junction typically contains two exons. BlockSizes are the lengths of the exons. The block starts the beginning of the exons. When fusions on different strands or chromosomes, two lines are added to the output, with each line representing one chromosome. Different colors are used to color-code different types:

The illustration of the IGV on page 304 shows the Upstream Hypersensitive Region (UHR) gene region displayed with the Integrative Genomics Viewer (IGV) for positions 3,530,193 to 3,548,355 of Human Chr-1. The following sections describe the tracks in the illustration on page 304.

WIG (x2) tracks

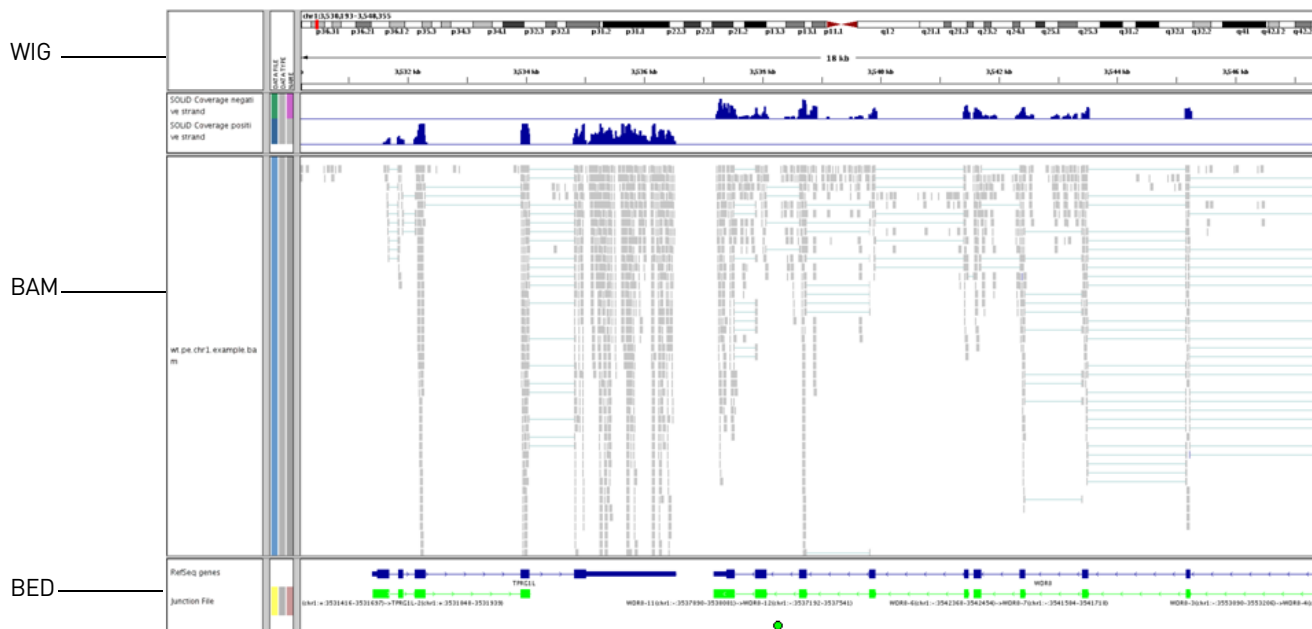
The top two tracks show the genomic coverage using the negative strand and positive strand generated by the Bam2Wig tool (Max: 100 coverage).

BAM track

The middle track shows the alignments from the BAM file. For display purposes, reads are filtered with MAPQ threshold of 45 (a stringent filter). Bases with quality value 5 to 20 are shaded.

BED track

The fifth track shows the junctions detected by the Junction Finder (BED file). As shown in the illustration on page 304, all junctions detected are “known” and so are shaded in green.



BEDGRAPH file format

The following table describes the BEDGRAPH format fields. These are tab-separated text fields.

Field name	Example content	Description
Track annotation headers	track type=bedGraph name="Coverage"	Annotations used for UCSC browser display purposes.

Field name	Example content	Description
Contig name	chr1	Contig name of the feature. This value comes from the sequence name (SN) in the BAM header. However, in order to be viewed in the UCSC browser, these values must be of the form chr1, chr2, etc.
Feature start	148	Start location of the feature. Zero-based, inclusive.
Feature end	150	End location of the feature. Zero-based, exclusive (half-open ranges).
Coverage	1	Depth of coverage over the given (start-end) range.

GFF3 file format

LifeScope™ Software creates the generic feature format, version 3, (GFF3) file as an output file containing the detected CNV Regions. You can visualize a CNVs.gff file in a browser such as the Integrative Genomics Viewer (IGV), which is available from the Broad Institute www.broadinstitute.org/igv/.

The file format for *.gff files is

`<cnv.output.prefix>_CNVs.gff3`

where `<cnv.output.prefix>` is the value defined for the output prefix.

Reference file data overview

Nearly all LifeScope™ Software tools use reference sequence data and, in some cases, the tools also use reference annotations and metadata.

Reference data takes a number of forms, which are described in the following sections.

Contig multi-fasta file

A single, multiple-entry FASTA file where each entry corresponds to a genomic contig or chromosome.

Single contig FASTA file

A single-entry FASTA file containing one contig or chromosome.

GTF file

A genome annotation file. This file describes gene models used by the whole transcriptome and small RNA modules. It is similar in form to the GTF files downloaded from the UCSC website, but includes some changes for WT or small RNA processing.

Genomic reference files downloaded from UCSC; for example, you can download hg19 from:

<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz>

The file can be used as a basis for the reference sequence data in LifeScope™ Software tools. A few small transformations are required to prepare the GTF files for use by LifeScope™ Software applications.

A GTF file is an ASCII text file composed of lines of text separated by line feed characters (UNIX-style new lines). Each line of text represents a comment or a tab delimited content about a genomic feature. A comment line always starts with a “#” character. GTF does not specify any format of structured content in comments. Comments containing parsable metadata should precede any lines containing genomic features. For more detail about the format, refer to the following site:

<http://genome.ucsc.edu/FAQ/FAQformat.html#format4>

Reference sequence data validation

The `reference_validation.pl` script provided with LifeScope™ Software checks and corrects common issues with FASTA sequence files that can cause errors in LifeScope™ Software runs. Examples of errors generated by an uncorrected GTF file can include extra spaces and mixed case.

Concatenation

The contig multi-FASTA reference file can be generated by concatenating the individual chromosome files together.

Select a reference file

Be sure to use the correct reference file type for your requirements. The identifiers in the FASTA definition line are carried through many tools, including the BAM file generated by mapping and pairing, and the GFF files typically written by tertiary analysis tools.

For some downstream viewing and integration tools, for example, the UCSC genome browser, the identifiers can be used to connect sequence references with other sources of annotation.

Read-set file format

An RRS (Read Range Specifier) file, also called a read-set file, is an internal file that defines which reads data is processed in a mapping run. LifeScope™ Software automatically creates the RRS file for your mapping analysis. This format is considered an internal file format, and the specification for this file format can change in future versions of LifeScope™ Software.

An RRS file allows a mapping analysis to process, in a single run:

- Data from a single XSQ reads file.
- Data from multiple XSQ reads files.
- Data for a range of barcodes (a subset of XSQ file data).
- Data for a range of panels within a barcode (also a subset of XSQ file data).

The RRS file is a tab-separated file with three columns which describe the input XSQ file, the sample name, and optionally a barcode range and panel range (see the following table in this section). The following rules apply to RRS files:

- One read-set (barcode) cannot be part of two samples.
- For a single analysis, mapping can be in color or base. If the mapping is in color, all XSQ files in the RRS file must have color available. If the mapping is in base, all XSQ files in the RRS file must have base available. Mapping fails if these restrictions are not met.
- For barcoded XSQ files, the RRS file can specify a range of barcodes. If some of specified barcodes are missing, mapping will log a warning for the missing barcode IDs.

- One read-set is specified per content line.
- At least one of the read-sets needs to be valid for mapping module to complete.

The requirements for RRS fields are described in the following table. The read-set file also accepts comments, which contain a '#' character at the beginning of the line.

Column name	Description
Sample name	<p>Multiple read-sets with the same sample name are treated as a single sample (and are processed together).</p> <p>Allowed values: A string of alpha-numeric characters up to 255 chars long. Spaces are not allowed.</p> <p>Examples: Human_101, Bob, Sample19</p> <p>(This sample name does not need to match any information provided on the instrument.)</p>
File id	<p>A positive integer that distinguishes between XSQ files according to these rules:</p> <ul style="list-style-type: none"> • Two different XSQ files must have different file IDs. • Two identical XSQ files must have same file ID. <p>(For read-set files created with the UI, LifeScope™ Software assigns this number.)</p>
read-set range URL	<p>A string specifying one or more XSQ data files and optionally also barcode and panel specifiers. The format is a URI scheme:</p> <p><i>fileName?start=barcodeID.panelstart&end=barcodeID.panelend</i></p> <p><i>fileName</i>: either a read in the reads repository or the absolute path to an XSQ file on the Linux file system (and that file must be a valid XSQ file for which the user has read permission).</p> <p><i>barcodeID</i>: integers from 1–96.</p> <p><i>panelstart</i>, <i>panelend</i>: integers from 1–4000. These fields correspond to panels on the sequencing instrument.</p> <p>The start= section must appear before the end= section. Start <i>barcodeID</i> value must be <= the end <i>barcodeID</i> value. The <i>panelstart</i> value must be <= the <i>panelend</i> value.</p>

Example RRS file

The following is a simple example of an RRS file which defines an entire XSQ file as the read-set. The XSQ file in this example has been imported into the reads repository, so an absolute path to the local file system is not required.

```
#SampleName FileID ReadSets
Huref 1 reads/solid0054_20110102_RD_HuRef100_F3.xsq
```

VCF file

The Variant Call Format, in a text format. The file contains meta-information lines, a header line, and data lines that each contain information about a position in the genome. This file format was developed as part of the 1000 Genomes project, and is described here:

<http://vcftools.sourceforge.net/specs.html>

Recent releases of dbSNP are available in as VCF files. An example is:

ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/v4.0/00-All.vcf.gz

B

Legacy Formats

This appendix covers:

- Introduction..... 309
- GFF to BAM mapping 309

Introduction

This appendix describes how header fields and alignment data in General Feature Format (GFF) files correspond to similar fields in the Binary Alignment sequence Map (BAM) file format.

GFF to BAM mapping

The following table shows the GFF file field name mapped to the corresponding BAM file field name for header fields.

GFF field	Corresponding BAM field	Comments
##reference-name	@SQ UR field	This header field contains the reference file name in the UR field.
##history	—	—
##color-code	—	Only a single color encoding is used.
##primer-base	—	This field is not needed. The primer base is stored with the color read attribute.
##max-num-mismatches	—	—
##max-read-length	@RG LB field	The read lengths in the file are determined from the library type information in the LB field of the read group.
##line-order	@HD SO field	The sort order field in the BAM header indicates the line order. Valid values are: <ul style="list-style-type: none"> • None • Coordinate • Qname

The following table shows the GFF file field name mapped to the corresponding BAM file field name for alignment data.

GFF field	Corresponding BAM field	Comments
seqid	RNAME	GFF files use an ordinal number to indicate the reference contig. The ordinal number could be translated to names by a header table. The BAM file uses the @SQ lines to perform a similar lookup.
source	@RG PL	The read group platform (@RG PL field) is used in a manner similar to the SOLiD™ GFF source.
type	—	All records are alignments.
start	POS	The 1-based, left-most, top-strand position of the alignment.
end	Calculated from CIGAR	The BAM file does not support an explicit end point. An explicit end point can be calculated from the CIGAR string.
score	—	A read quality is not stored in the BAM alignment record. Mapping and pairing quality is stored in the BAM alignment record.
strand	5th FLAG bit (query strand)	The fifth bit of the FLAG field indicates the strand of the alignment.
GFF attributes		
aID (bead id)	QNAME	The bead id.
at (adaptor type)	Calculated value	The adaptor type or primer set field can be calculated as described in “Calculation of tag names” on page 300 .
b (bases)	SEQ	Base space representation of the aligned read. Bases are all uppercase. Equal signs “=” are not used.
c (category)	2nd FLAG bit (proper pair) and calculation	The most commonly-used category value, AAA, corresponds to the proper pair bit in the FLAG field. Use POS, MPOS, and strand bits in the FLAG field to calculate other categories. You can find C** pairs (different contigs) by interrogating the mate reference (MRNM). The D** is indicated by the fourth FLAG bit.
g (color-space read)	CS attribute	The color read is stored in the CS attribute. However, unlike the 'g' GFF attribute, the color read is stored in an unaltered CSFASTA form that includes the last primer base.
mq (mq [mapping quality])	MAPQ (or SM for pairs)	The mapping quality is stored in the MAPQ field for fragment libraries and in the SM attribute for paired results.
o (offset)	CIGAR (hard-clipping)	The offset for an alignment can be computed by checking the number of hard clip (H) operations on the CIGAR string.
p (mappability ambiguity)	—	Use mapping quality for uniqueness determinations.
pq (pairing quality)	MAPQ (for pairs)	The pairing quality is stored in the MAPQ field for paired results.
q (color qualities)	CQ	The CQ attribute stores color quality. The values are not represented as integers, but are represented by ascii-33 encoding.
r (reference color at mismatches)	Calculate reference color at mismatches using the samtools fillmd command.	CIGAR strings do not distinguish between matches and mismatches. Calculate this information using the samtools fillmd command, though in a form that is different from the rb GFF attribute.
s (color annotations)	—	—

GFF field	Corresponding BAM field	Comments
<i>u (accumulated mismatch count)</i>	Per-alignment mismatches supported with CM attribute	The CM attribute specifies the number of color mismatches for the current alignment. You can use a combination of the number of color mismatches across alignments in the BAM file to reconstruct values like the 'u' attribute.



Run the SOLiD™ Accuracy Enhancement Tool

This chapter covers:

■ Overview	313
■ SAET usage guidelines	313
■ SAET in analysis modules	314
■ SAET input files	314
■ SAET parameters	315
■ SAET output files	316

Overview

The SOLiD™ Accuracy Enhancement Tool (SAET) improves color call accuracy prior to mapping. This method implements a proprietary spectral alignment error correction algorithm that uses quality values and properties of color space. Significantly improving its performance over conventional error correction methods, the use of SAET produces more accurate mapping with both an increased number of reads with zero miscalls and an increased number of mapped reads. The SNP calling on error-corrected data results in an increase in true positive calls and a decrease in the number of false positive calls.

SAET performance was tested on various datasets, including a large spectrum of genome sizes and complexities, coverages, read lengths, and experiment types. SAET shows similar performance on datasets sequenced from enriched regions or genomes of 1 kbp to 3 Gbp, with coverage from 10 to 4000x and a read length of 25 to 75 bp.

On average, SAET requires up to 5 GB RAM for each 1 Gbp of sequenced region. SAET can be applied for diploid organisms. SAET should not be used for rare variant detection datasets where the frequency of rare variant may be less than twice that of the color calling error rate.

For more information about SAET, see [Chapter 9, “Perform Targeted Resequencing and Enrichment Analyses”](#) on page 95.

SAET usage guidelines

SAET is recommended for use with targeted resequencing and is on by default in targeted resequencing standard workflows. SAET has not been validated with other experiment types and may have insufficient memory to run. If you do choose to run SAET with other experiment types, you can do the following:

- Run SAET before you run the resequencing or whole transcriptome mapping and pairing tools.

- Use the SAET output with resequencing and whole transcriptome mapping and pairing.
- Use updated quality values file for SNP calling.
- Avoid the generation of the updated quality values, if the downstream analysis modules in your workflow do not use quality values. Your SAET runtime is reduced approximately by half if you do not update quality values.

SAET in analysis modules

The following analysis modules apply SAET in targeted resequencing of the fragment and paired-end library types:

- SAET
- Mapping
- Enrichment
- SNPs
- Small indels
- Annotations

SAET corrects errors in reads before mapping. This correction significantly improves mapping quality and the quality of variant calls.

SAET input files

By default, SAET takes an input file of SOLiD™ system reads in the eXtensible SeQuence (XSQ) format, and an expected length of the sequenced region, such as the:

- Genome size for whole genome sequencing
- Size of enriched region in targeting resequencing
- Size of transcribed region in whole transcriptome sequencing

The required input file is in XSQ format, which is typically generated by 5500 Series SOLiD™ Sequencer. The file must contain reads in color-space that require correction. SAET constructs a spectrum of all k-mers from the set of all reads. Then each read is corrected individually if necessary.

In targeted resequencing, the SAET module accepts as an input the BED, GFF, or TXT file with segments of target regions.



SAET parameters

SAET is designed to reduce the error rate in the reads generated by the SOLiD™ System. Reducing the error rate increases the number of mapped reads for resequencing analyses, which can for example lead to an increase in true positive SNP-calls and a decrease in false positive SNP-calls.

The following table describes the SAET parameters.

Parameter	Default value	Description
Genome length	1000000	Expected length of sequenced (or enriched) DNA region. For example, 4,600,000 for the E.Coli 4.6 MB genome or 30,000,000 for the entire Human Transcriptome. Allowed values: Integers ≥ 200 .
On target ratio	0.5	The expected ratio of reads that come from the targeted region. Allowed values: Floats 0.0–1.0.
Update quality values	True	Update quality value of modified calls. Allowed values: True: Update the qv of modified calls. False: Do not update the qv for modified calls.
Trusted quality value	25	Correction is applied to calls with a quality value below the value of this parameter. Allowed values: Integers ≥ 1 .
Support votes	2	The minimum number of k-mer votes required make a correction. Allowed values: Integers ≥ 1 .
Trusted frequency	0	The lowest multiplicity of the seed to be included in the spectrum. (If set to 0, then the value is computed internally.) Allowed values: Integers ≥ 0 .
Maximum corrections per read	0	Maximum number of allowed corrections per read. (If set to 0, then the value is set to $\lceil \text{readLength}/8 \rceil$). Reduce if over-corrections are observed, or increase if under-corrections are observed. Allowed values: Integers 0–9.
Number of recursive runs	1	The error correction step is repeated the provided number of times. Reduce if over-corrections are observed, or increase if undercorrections are observed. Allowed values: 1, 2, or 3.

Parameter	Default value	Description
Position of error inflation point	0	Position in the read at which the error rate inflates, for instance, 35-40 for 50bp long reads. (If set to 0, then the value is equal to $0.8 * \text{readLength}$). Allowed values: Integers.
Disable random sampling for large data	False	Disables random sampling in spectrum building. If set to 0, then for large datasets (coverage > 300x), a subset of reads is used in spectrum building. Allowed values: <ul style="list-style-type: none"> • True: Disables random sampling in spectrum building. • False: Do not disable random sampling in spectrum building. true?
K-mer size	0	Size of k-mer (>5) used in spectrum construction and error correction. (If set to 0, then the value is computed internally.) Allowed values: Integers 0-28.

SAET output files

SAET writes the corrected reads and quality values into new XSQ files.

The SAET output files are used as input to mapping analyses.



Administration

This appendix covers:

■ Introduction.....	318
■ Log in	318
■ Administrate users	320
■ Search for users	320
■ Add users.....	320
■ Deactivate users	321
■ Configure user accounts	321
■ Reset the password.....	322
■ Troubleshooting	323
■ Help	323

Introduction

As an administrator, you can search for users, add users, deactivate user accounts, and configure licenses and user setups and profiles.

Log in

In the Admin window, enter your username and password, then click **Login** to open the Admin Portal shown on .

If you cannot remember your password, click the **Forgot Password** button, then follow the prompts.

If you need assistance, click **Login Help**.

Users

Filter User List

Last Name: First Name: Include Inactive

Login ID:

	Name	Login ID	Roles	Site	Phone
<input type="checkbox"/>	mayank	mayank	admin,user	ngp	354654
<input type="checkbox"/>	LifeScope	lifescope	admin		
<input type="checkbox"/>	ashish	ashish	admin,user	nagpur	997054

Select and...

Administrate users

Follow these steps to access and work within the Admin tool window:

1. If you are not already logged in as an administrator, do so.
Log in as “admin” to display the Admin tool window.
The left panel in the Admin tool window shows folders for Users and Licenses.
2. Click the arrow next to any folder to display a list of its contents.
3. Click any folder to open it and show detailed information about the contents in the right panel of the Admin window.
4. Click on a user name in the left panel to show a profile of the user in the right pane.

Search for users

There are two ways to search for users:

- Click **Users** in the top drop-down menu, then select **Search**.
- In the right (details) pane, a user’s Last Name, First Name, Login ID, or Email address in the data entry boxes, then click the **Filter** button. To include inactive users in your search, click that checkbox.

Add users

You can add a single user or you can add multiple users from the Lightweight Directory Access Protocol (LDAP).

To add a single user:

1. Click:
 - **Users** in the top menu, then select **Add User**, or
 - Click the **Add User** button in the Users section of the Admin tool.After users are added you can search for a specific user by entering the user’s last name, first name, login ID, or e-mail address, and then clicking the **Filter** button.
2. In the Add User window, provide the following information:
 - Name
Name can be input as “last name, first name” or as “first name<space>last name”.
 - To search for a user, click the **Browse** button.
 - Username
 - Password
 - Notes about the user
 - Select whether the user is an administrator or a biologist.
 - Email address
 - Phone number
 - Office
 - Site
3. Click **OK** to add the user.

To add multiple users:

1. Click **Users** in the top menu, then select **Add User** and repeat the steps explained in Step 2 [on page 320](#)
or
2. Click the **Add Users from LDAP** button in the Users section of the Admin tool, select the users to add, and click **Add**.

To add multiple users (LDAP only):

1. In the Add Users from LDAP window, select the users to add.
2. After you have selected users to add, click the **Add** button.

Deactivate users

To deactivate a user:

1. Click the checkbox next to the user's name in Users section of the Admin tool.
2. Select click the **X Deactivate Users** button.
3. In the Deactivate Users window, click **OK** to deactivate the user.

Delete users

Once added, users cannot be deleted; they can only be deactivated.

Reactivate users

To reactivate a user:

1. Click the checkbox next to the user's name in the left column list of users.
2. Click the **Active** box at the bottom of the User Profile page.
3. Click **Save Changes**.

Configure user accounts

This section describes how to configure licenses and a user's setup and profile.

Configure licenses

To configure a user's license for sites using named licenses:

1. Click **Configuration** in the top menu, then select **License**.
2. In the License section, click the checkbox in the License column for users who have software licenses. If a user's license has expired, click the License checkbox to erase the check mark.

To apply your changes, click **Save Changes**.

Configure a user's setup

To configure a user's setup:

1. Click **Configuration** in the top menu, then select **Setup**.
2. In the Edit Setup window, use the **Browse** buttons to search for the following repositories:

- Read-Sets
- References
- Projects

Note: If you change the path to the read-set repository, notify users to restart LifeScope™ Software so that the repository can be updated to the new path.

3. In the data entry box, edit the repository locations.
4. To edit network services, use the Browse buttons to search for the following services:
 - SMTP
 - LDAP
 - Licenses
5. In the data entry box, edit the network services.
6. To apply your changes, click **OK**.

Configure a user's profile

To configure a user's profile:

In the left panel of the Admin tool, click the **Users** folder, then click a user's name to view and edit the user's profile.

To apply your changes, click **Save Changes**.

Read-set repository path: notify users

If you change the path to the read-set repository, notify users to restart LifeScope™ Software so that the repository can be updated to the new path.

Reset the password

The `resetpwd.sh` script is provided as an emergency way to reset the administrator's password to its default value. To reset the admin password, follow these steps (at a Linux® prompt):

1. Stop the LifeScope™ Software server:


```
lscope-server.sh stop
```
2. Reset the admin password:


```
resetpwd.sh lifescape
```
3. Restart the LifeScope™ Software server.
4. In the LifeScope™ Software graphical user interface, change the admin password to a secure password, according to your local policy.

The use of the `resetpwd.sh` script is not recommended for any user account except the LifeScope™ Software administrator. Passwords reset for other users may become out of sync in the authenticating realm.

Troubleshooting

If power was lost, causing the LifeScope™ Software server to shut down, and the server generates the message “read-only db connection,” do the following procedure:

Note: This is a non-recoverable action.

1. Stop the server:
`lscope-server.sh stop`
2. Delete the locked user database file:
`rm /share/apps/lifescopeserver/UserDB/db*.lck`
3. Restart the LifeScope™ Software:
`lscope-server.sh start`

If the problem persists, or if the error message is different (indicating a corrupt database), do the following procedure to delete the user database (UserDB), which will remove all the current users in the system. Then recreate the users.

1. Stop the server:
`lscope-server.sh stop`
2. Delete the user database:
`rm -rf /share/apps/lifescopeserver/UserDB`
3. Restart the LifeScope™ Software:
`lscope-server.sh start`

Help

The Help option in the top menu contains instructions on how to use the Admin tool. Within the Help menu, select **User's Guide** for an Adobe Acrobat PDF copy of this document, **Tutorial** for step-by-step instructions, or **About** for the version of LifeScope™ Software and copyright information.





LifeScope™ Genomic Analysis Software v2

END USER LICENSE AGREEMENT

This is a legal agreement between you, the person or entity receiving software products and/or software support (“Licensee”), and Life Technologies Corporation, having offices at 5791 Van Allen Way, Carlsbad California 92008 USA (“Licensor”). This agreement is part of a package that includes one or more software products and certain electronic and/or written materials. This agreement covers your licensing of such software and/or purchase of support.

You must agree to the terms in this End User License Agreement (“EULA”) in order to access the software and/or receive support.

BY CLICKING YOUR ACCEPTANCE OF THIS EULA, OR BY INSTALLING OR USING THE SOFTWARE (defined below) OR ANY OTHER COMPONENT OF THE PACKAGE, YOU ACKNOWLEDGE THAT YOU HAVE READ ALL OF THE TERMS AND CONDITIONS OF THIS EULA, UNDERSTAND THEM, AND AGREE TO BE LEGALLY BOUND BY THEM. If you do not agree to the terms of this EULA, you may not install or use the Software, and may return it to Licensor for a refund or product credit. In addition to the restrictions imposed under this EULA, any other usage restrictions contained in the Order (defined below), Software installation instructions or release notes, and Support policies (defined below) shall apply to your use of the Software and receipt of Support.

As used in this EULA: “**Authorized Users**” means, collectively, the personnel authorized by you to use the Software for your benefit, provided you have both purchased a License (as defined below) and paid the corresponding license fees. Unless otherwise expressly allowed by this EULA, Authorized Users may include only your employees and agents having a need to know, and Authorized Users may not be entities or persons in the business of licensing or otherwise providing products or services competitive with the Software. “**Designated Site**” means your facilities or offices located at the postal address provided to Licensor for your billing and invoicing purposes, unless otherwise indicated in a license key provided to you. “**Software**” means the software product(s) accompanying this EULA and the content therein; including the associated data, user manuals, user documentation and application program interfaces (if any), and License Key(s) and File(s) provided, and any patches, updates, upgrades, improvements, enhancements, fixes and revised versions of any of the foregoing that may be provided to you from time to time, and any combination of the foregoing. “**License Key(s)**” means the alphanumeric code(s) provided to you by Licensor to enable you to obtain a License File(s). “**License File(s)**” means the file(s) provided by Licensor with which you activate your copy of the Software. “**Order**” means that part of a written or electronic document that identifies (1) the Software to be licensed to you, (2) the Authorized Scope (defined below), (3) any Support purchases, (4) the purchase price, and (5) location for delivery, and in each case as expressly agreed upon by Licensor. “**Affiliate**” means any entity Controlling,

Controlled by, or under common Control with the referenced entity, where the term “**Control**” means the possession, direct or indirect, of the power to direct or cause the direction of the management and policies of an entity, whether through the ownership of voting securities, by contract, or otherwise.

Your Payment Obligations

YOU AGREE TO PAY ALL AMOUNTS DUE OR INCURRED BY YOU, INCLUDING ANY LATE PAYMENT FEES, AS ARE SPECIFIED IN THIS EULA, IN THE ORDER, AND/OR ANY ASSOCIATED INVOICE. ALL FEES AND AMOUNTS DUE LICENSOR ARE EXCLUSIVE OF ALL TAXES, DUTIES SHIPPING FEES, AND SIMILAR AMOUNTS. IF ANY AUTHORITY IMPOSES A DUTY, TAX OR SIMILAR AMOUNT (OTHER THAN TAXES BASED ON LICENSOR’S INCOME), YOU AGREE TO PAY, OR TO PROMPTLY REIMBURSE LICENSOR FOR, ALL SUCH AMOUNTS. UNLESS OTHERWISE INDICATED, ALL INVOICES ARE PAYABLE THIRTY (30) DAYS FROM THE DATE OF INVOICE. OVERDUE AMOUNTS ARE SUBJECT TO A LATE PAYMENT CHARGE, AT THE LOWER RATE OF (i) ONE PERCENT (1%) PER MONTH, OR (ii) THE MAXIMUM RATE UNDER APPLICABLE LAW. YOU AGREE TO PROMPTLY PAY OR REIMBURSE LICENSOR FOR ALL COSTS AND EXPENSES, INCLUDING ALL REASONABLE ATTORNEYS’ FEES, RELATED TO BREACH OF YOUR OBLIGATIONS UNDER THIS EULA AND/OR LICENSOR’S ENFORCEMENT OF THIS EULA. ALL SHIPMENTS BY LICENSOR OR ITS DESIGNEE ARE FCA POINT OF SHIPMENT (INCOTERMS 2000).]

Acceptance

Except with respect to Software provided under a Trial License (defined below), you will be deemed to have accepted the Software unless you provide written notice of rejection within ten (10) days after receipt of the Software or the corresponding License Key(s) and File(s), if any (whichever event occurs first). Any such notice must state the reason for rejection, and you may only reject the Software if it fails to materially comply with its accompanying documentation. If you reject the Software, Licensor’s sole obligation and liability, and your sole and exclusive remedy, shall be for Licensor to use commercially reasonable efforts to deliver to you a replacement for the nonconforming Software, and if Licensor is not able to deliver a replacement for the Software, then Licensor will refund any license fees paid by you for the Software, and in the event of any such refund, the EULA shall terminate. Software provided under a Trial License shall be deemed accepted upon receipt.

Grant of Software License

Subject to the terms and conditions of this EULA, Licensor grants to you a non-exclusive, non-transferable license (“**License**”) for Authorized Users to use the Software and Support for your internal operations and internal data processing purposes within and up to the Authorized Scope described on the Order, and for which you have paid the applicable license and support fees and have registered the Software for use. Except as otherwise provided below, this EULA and the License granted hereunder shall be effective until terminated in accordance with Section 7 below. The term “**Authorized Scope**” means the following, and any other capacity, term/duration, or use restrictions indicated by Licensor on the license granted to you:

The “**Concurrent User License**” configuration is comprised of a defined number of simultaneous-use licenses shared among an unlimited number of computers that are on the network at the Designated Site. Each Concurrent User License has a term of one (1) year. For example, the “LifeScope™ Core Server Software C1” product enables simultaneous use of the Software on any of up to five (5) computers for one (1) year at

the Designated Site. In other words, at any given time, any combination of users from up to 5 computers on the network may access the Software at the Designated Site. Each Concurrent User must have a unique username and password to access the Software. Simultaneous use of the Software from additional computers above five (5) can be enabled through the purchase of the desired number of “Supplemental User C1” licenses and payment of the applicable license and support fees. You may not enable concurrent operation of the Software beyond the scope of the Concurrent User Licenses purchased. You may not network the Software for use at any other geographic location. Termination or expiration of the Concurrent User Licenses shall immediately terminate this EULA.

The “**Named User License**” configuration is comprised of a defined number of licenses shared among the same number of computers that are on the network at the Designated Site. Each Named User License has a term of one (1) year. For example, the “LifeScope™ Core Server Software N1” product enables simultaneous use of the Software on only five (5) specific computers for one (1) year at the Designated Site. In other words, at any given time, only users from the defined list of computers may access the software at the Designated Site. Each Named User must have a unique username and password to access the Software. Additional computers above five (5) may be added to the defined list of computers through the purchase of the desired number of “Supplemental User N1” licenses and payment of the applicable license and support fees. You may not enable operation of the Software beyond the scope of the Named User Licenses purchased. You may not network the Software for use at any other geographic location. Termination or expiration of the Named User Licenses shall immediately terminate this EULA.

The Software may be provided under a “**Trial License**” for your internal evaluation purposes only, pending your purchase of a commercial-use Software license and payment of the applicable license fees. Depending on the mode of delivery of the Trial License, you may (i) access the Software through an online account mechanism, or (ii) you may install and operate the Software on computer hardware located at the Designated Site. You may not network the Software for use at any other geographic location. A Trial License may not be used for any commercial purposes. A Trial License may automatically be converted into a commercial license (*e.g.*, with Authorized Scope converted into a Concurrent User License) upon payment of the applicable license fees to Licensor or issuance of the applicable License Key(s) and File(s), if any (whichever event occurs first). Upon such conversion, this EULA shall continue in full force and effect, subject to the restrictions applicable to the new Authorized Scope. Software provided under a Trial License is made available to you “AS IS”, AND LICENSOR MAKES NO REPRESENTATION OR WARRANTY REGARDING SUCH SOFTWARE.

RESTRICTIONS ON USE. You acknowledge that you are receiving LICENSED RIGHTS only. The Software may only be used internally, by your Authorized Users, with the License Key(s) and File(s) (if any) provided, for your copy(ies) of the Software. If any Software is provided on separate media (*e.g.*, a CD-ROM), you may make a single copy solely for your internal backup purposes. You shall not directly or indirectly: (i) sell, rent, lease, distribute, redistribute or transfer any of the Software or any rights in any of the Software, including without limitation through “charge back” or any other selling, reselling, distributing or redistributing within your organization of any usage capacity you have licensed, without the prior express written approval of Licensor, (ii) modify, translate, reverse engineer (except to the limited extent permitted by law), decompile, disassemble, attempt to discover the source code for, create derivative works based on, or sublicense any of the Software, (iii) use any Software for the benefit of any third parties (*e.g.*, in an ASP, outsourcing or service bureau relationship), or in

any way other than in its intended manner, without the prior express written approval of Licensor, (iv) remove any proprietary notice, labels, or marks on or in the Software, or (v) disable or circumvent any access control or related device, process or procedure established with respect to the Software, including the License Key(s) and File(s) (if any) or any other part thereof. Further, without Licensor's prior express written consent, you may not: (i) network the Software for use at any other geographic location; (ii) enable operation of concurrent "instances" of the Software beyond the scope of the license purchased; (iii) share accounts, e.g., pool multiple users through a single account, or (iv) "multiplex" or "pool" any Software, including through use of third party software products such as those made by Citrix Systems, Inc., or use any terminal applications/emulators to enable use of the software beyond the scope of the license purchased. If the Software design permits modification, then you may only use such modifications or new software programs for your internal purposes and otherwise consistent with the License and Authorized Scope. You are responsible for all use of the Software and for compliance with this EULA; any breach by you or any user of the Software shall be deemed to have been a breach by you. Licensor reserves all rights not expressly granted; no right or license is granted hereunder, express or implied or by way of estoppel, to any intellectual property rights other than as expressly set forth herein; and your purchase of a license to the Software does not by itself convey or imply the right to use the Software in combination with any other product(s). As between you and Licensor, Licensor retains all right, title, and interest in and to the Software, which rights include, but are not limited to, patent, copyright, moral, trademark, trade secret and all other intellectual property rights. You agree and acknowledge that you have been provided sufficient information such that you do not need to reverse engineer the Software in any way to permit other products or information to interoperate with the Software.

NO SEPARATION OF COMPONENTS. The Software is licensed as a single product. Some Licensor software products combine separately available components into a single product (e.g., a software suite product may be comprised of multiple component products). When licensed as a combination product, the component parts may not be separated for use independently of the combination product. You must first purchase a license to each component of the combination product before you may use it independently of the combination product.

ADDITIONAL PURCHASES. Purchase of additional or changed licenses or Authorized Scope is subject to availability and current pricing. Licensor may, from time to time, update the available Authorized Scope plans, and add or delete from available plans. To the extent you purchase an upgrade to your license by expanding the Authorized Scope under a changed or additional plan, Licensor will provide you with the additional license terms and conditions governing your use of the Software under such plans; all other terms and conditions of this EULA shall remain in effect.

TERM AND TERMINATION. Unless otherwise agreed, the term of this EULA shall continue until it expires or is terminated; however, if you are receiving a Trial License, the term shall expire thirty (30) days following the date you receive the Software. Termination or expiration of this EULA shall concurrently terminate all Licenses granted under this EULA. Licensor shall not refund any amounts paid by you hereunder in the event of expiration or termination of this EULA unless expressly provided in this EULA. Licensor may (i) terminate an Order and the Licenses to the Software and/or Support on that Order if you fail to pay any applicable fees due under that Order within fifteen (15) days after receipt of written notice of non-payment; and/or (ii) terminate this EULA (or any License) upon fifteen (15) days written notice if you breach this EULA and do not cure the breach within fifteen (15) days following receipt of written notice of breach. Immediately upon any termination or expiration of this

EULA, you agree to: (a) pay all amounts owed to Licensor; (b) un-install and cease use of the Software for which your rights have been terminated; (c) upon request, return to Licensor (or destroy) all copies of the Software and any other Confidential Information or proprietary materials in your possession for which your rights have been terminated; and (d) upon request, certify in writing your compliance with (b) and (c), above.

CONFIDENTIALITY. You agree to protect Licensor's Confidential Information with the same degree of care used to protect your own confidential information (but in no event less than a reasonable standard of care), and not to use or disclose any portion of such Confidential Information to third parties, except as expressly authorized in this EULA. You acknowledge that the Software, including its content, structure, organization and design constitute proprietary and valuable trade secrets (and other intellectual property rights) of Licensor and/or its licensors. The term "**Confidential Information**" means, collectively, non-public information that Licensor (and its licensors) provide and reasonably consider to be of a confidential, proprietary or trade secret nature, including but not limited to (i) the Software, (ii) Software License and Support prices, (iii) Software License Keys and Files, and (iv) confidential elements of the Software and Licensor's (and its licensors') technology and know-how, whether in tangible or intangible form, whether designated as confidential or not, and whether or not stored, compiled or memorialized physically, electronically, graphically, photographically, or in writing. Confidential Information does not include any information which you can demonstrate by credible evidence: (a) is, as of the time of its disclosure, or thereafter becomes part of the public domain through no fault of yours; (b) was rightfully known to you prior to the time of its disclosure, or to have been independently developed by you without use of Confidential Information; and/or (c) is subsequently learned from a third party not under a confidentiality obligation with respect to such Confidential Information. Confidential Information that is required to be disclosed by you pursuant to a duly authorized subpoena, court order, or government authority shall continue to be Confidential Information for all other purposes and you agree, prior to disclosing pursuant to a subpoena, court order, or government authority, to provide prompt written notice and assistance to Licensor prior to such disclosure, so that Licensor may seek a protective order or other appropriate remedy to protect against disclosure.

WARRANTY AND DISCLAIMER. Licensor warrants that (i) except with respect to Software provided under a Trial License (in respect of which no warranty is made, as described in Section Grant of Software License), for a period of twenty (20) days from the date of acceptance of the Software as described in Section Except with respect to Software provided under a Trial License (defined below), you will be deemed to have accepted the Software unless you provide written notice of rejection within ten (10) days after receipt of the Software or the corresponding License Key(s) and File(s), if any (whichever event occurs first). Any such notice must state the reason for rejection, and you may only reject the Software if it fails to materially comply with its accompanying documentation. If you reject the Software, Licensor's sole obligation and liability, and your sole and exclusive remedy, shall be for Licensor to use commercially reasonable efforts to deliver to you a replacement for the nonconforming Software, and if Licensor is not able to deliver a replacement for the Software, then Licensor will refund any license fees paid by you for the Software, and in the event of any such refund, the EULA shall terminate. Software provided under a Trial License shall be deemed accepted upon receipt., the Software will, under normal use and as unmodified, substantially perform the functions described in its accompanying documentation; and (ii) Licensor will perform Support services during the License term in a professional and workmanlike manner. No warranty is provided for uses beyond the Authorized Scope. THE FOREGOING EXPRESS WARRANTIES

REPLACE AND ARE IN LIEU OF ALL OTHER WARRANTIES AND REPRESENTATIONS BY LICENSOR, WHETHER EXPRESS, IMPLIED, OR STATUTORY, INCLUDING BUT NOT LIMITED TO ANY IMPLIED OR OTHER WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE, NON-INFRINGEMENT OR NON-MISAPPROPRIATION OF INTELLECTUAL PROPERTY RIGHTS OF A THIRD PARTY, CUSTOM, TRADE, QUIET ENJOYMENT, ACCURACY OF INFORMATIONAL CONTENT, OR SYSTEM INTEGRATION. NO WARRANTY IS MADE THAT ANY SOFTWARE WILL OPERATE IN AN ERROR FREE, UNINTERRUPTED OR COMPLETELY SECURE MANNER, IN COMBINATION WITH THIRD PARTY HARDWARE OR SOFTWARE PRODUCTS, OR THAT ALL DEFECTS CAN BE CORRECTED. YOU ACKNOWLEDGE THAT LICENSOR HAS NO CONTROL OVER THE SPECIFIC CONDITIONS UNDER WHICH YOU USE THE SOFTWARE. ACCORDINGLY, EXCEPT FOR THE FOREGOING EXPRESS WARRANTY, LICENSOR CANNOT AND DOES NOT WARRANT THE PERFORMANCE OF THE SOFTWARE OR ANY PARTICULAR RESULTS THAT MAY BE OBTAINED BY THE USE OF THE SOFTWARE. THE SOFTWARE AND SUPPORT DO NOT REPLACE YOUR OBLIGATION TO EXERCISE YOUR INDEPENDENT JUDGMENT IN USING THE SOFTWARE. The warranties made by Licensor may be voided by abuse or misuse of the Software and/or Support.

EXCLUSIVE REMEDY. Licensor's sole obligation and liability, and your sole and exclusive remedy under the warranties set forth in Section 9, shall be for Licensor to use commercially reasonable efforts to have the problem remedied, to re-perform Support services, to deliver to you a replacement for the defective Software, or to refund fees paid (in each case, as determined by Licensor and as applicable), provided that Licensor is notified in writing of all warranty problems during the applicable warranty period.

Third Party Software and Databases

This Software uses third-party software components from several sources. Portions of these software components are copyrighted and licensed by their respective owners. Various components require distribution of source code or if a URL is used to point the end-user to a source-code repository, and the source code is not available at such site, the distributor must, for a time determined by the license, offer to provide the source code. In such cases, please contact your Life Technologies representative. In addition, various licenses require that the end user receive a copy of the license. Such licenses may be found on-line as supporting files on the download page for the Software. In order to use this Software, the end-user must abide by the terms and conditions of these third-party licenses. You understand that third party products integrated into the Software or provided for use with the Software may be subject to additional terms and conditions and/or license agreements from the applicable third party vendor, which shall govern over conflicting terms of this EULA for purposes of your relationship with the third party vendor. You agree not to use any such third party product on a stand-alone basis independent of the Software, unless you have purchased the appropriate license from the third party vendor for use of such products. NOTWITHSTANDING ANYTHING TO THE CONTRARY IN THIS AGREEMENT, ALL THIRD PARTY SOFTWARE, DATABASES AND OTHER PROGRAMS AND SOFTWARE COMPONENTS ARE PROVIDED "AS IS" WITHOUT ANY WARRANTY WHATSOEVER FROM LICENSOR. ANY DATABASES OR OTHER INFORMATION PROVIDED BY LICENSOR ARE DESIGNED TO SUPPLEMENT OTHER SOURCES OF INFORMATION, ARE NOT INTENDED TO REPLACE YOUR PROFESSIONAL DISCRETION AND JUDGMENT AND

LICENSOR MAKES NO REPRESENTATIONS OR WARRANTIES REGARDING SUCH DATABASES OR INFORMATION, THEIR ACCURACY, COMPLETENESS OR OTHERWISE. Licensor agrees, upon request and as Licensor's sole liability and obligation, and for your convenience only, to have passed through to you (to the extent it may reasonably do so) any warranties and indemnifications provided by the applicable third party vendor of any third party products provided to you. To the extent any problem or liability arises from a third party product, you agree to seek recourse solely from the applicable third party vendor and not Licensor.

LICENSOR INDEMNIFICATION. Subject to the limitations set forth herein, Licensor agrees to defend you against any claims, actions, suits and proceedings brought against you by unaffiliated third parties arising from or related to a claim that the Software (other than any third party or open source components or elements) infringes upon such third party's copyrights, and Licensor agrees to pay all damages that a court finally awards to such third party, and all associated settlement amounts agreed to by Licensor in writing; provided that, Licensor receives from you (i) prompt written notice of the claim; (ii) all necessary assistance, information and authority necessary for Licensor to defend the claim and perform Licensor's obligations under this Section; and (iii) sole control of the defense of such claim and all associated settlement negotiations. If such a claim is made or appears likely to be made, you agree to permit Licensor to enable you to continue to use the affected Software, or to have the Software modified to make it non-infringing, or to have the Software replaced with a substantially functional equivalent. If Licensor determines that none of these options is reasonably available, then Licensor may terminate this EULA in whole or with respect to the affected Software product, and you may be entitled to a credit equal to the price paid for the affected product, less depreciation assuming a three (3) year useful life and straight-line depreciation. THIS SECTION STATES LICENSOR'S AND ITS AFFILIATES' ENTIRE OBLIGATION AND LIABILITY REGARDING INFRINGEMENT OF THIRD PARTY RIGHTS OF ANY KIND OR CLAIMS OF ANY SUCH INFRINGEMENT. Licensor will have no responsibility for (v) any use of any product after you have been notified to discontinue use because of a third party claim of infringement, (w) the alteration of the Software or the combination of the Software with third party materials, products or software, (x) use of the Software by any person or entity other than an Authorized User, (y) any misuse or unauthorized use of the Software, or (z) failure to use provided updated or modified Software to avoid a claim of infringement or misappropriation.

INDEMNIFICATION BY YOU. You agree to indemnify and defend Licensor, its licensors, and its affiliates, against any third party claims arising from or related to your use or misuse of the Software or any breach of the terms and conditions of this EULA, and you agree to pay all costs, losses, damages, and attorneys' fees that a court finally awards, and all associated settlements.

LIMITATION OF LIABILITY. EXCEPT TO THE EXTENT PROHIBITED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR'S OR ITS AFFILIATES' TOTAL, AGGREGATE LIABILITY ARISING FROM OR RELATED TO THIS AGREEMENT, THE SOFTWARE AND/OR SUPPORT (INCLUDING FOR NEGLIGENCE, STRICT LIABILITY, BREACH OF CONTRACT, MISREPRESENTATION, AND OTHER CONTRACT OR TORT CLAIMS), EXCEED THE AMOUNT OF YOUR DIRECT DAMAGES ACTUALLY INCURRED, UP TO THE AMOUNT OF FEES PAID TO LICENSOR UNDER THIS AGREEMENT FOR THE SOFTWARE PRODUCT OR SUPPORT PRODUCT THAT IS THE SUBJECT OF THE CLAIM UNDERLYING THE DAMAGES; OR, IN THE CASE OF SOFTWARE PROVIDED UNDER A TRIAL OR DEMONSTRATION LICENSE, ONE HUNDRED DOLLARS (\$100.00), WHICHEVER IS LESS.

EXCLUSION OF DAMAGES. EXCEPT TO THE EXTENT PROHIBITED BY APPLICABLE LAW, UNDER NO CIRCUMSTANCES SHALL LICENSOR, ITS AFFILIATES, OR ANY OF THEIR SUPPLIERS OR LICENSORS BE LIABLE HEREUNDER FOR ANY OF THE FOLLOWING: (I) THIRD PARTY CLAIMS, EXCEPT AS PROVIDED IN SECTION 12, (II) LOSS OR DAMAGE TO ANY SYSTEMS, RECORDS OR DATA, (III) DIRECT DAMAGES FOR BREACH OF WARRANTY (IN RESPECT OF WHICH ANY LIABILITY SHALL BE LIMITED TO RE-PERFORMANCE OR REFUND AS SPECIFIED IN SECTION Exclusive Remedy. Licensor's sole obligation and liability, and your sole and exclusive remedy under the warranties set forth in Section 9, shall be for Licensor to use commercially reasonable efforts to have the problem remedied, to re-perform Support services, to deliver to you a replacement for the defective Software, or to refund fees paid (in each case, as determined by Licensor and as applicable), provided that Licensor is notified in writing of all warranty problems during the applicable warranty period.), AND/OR (IV) INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL, PUNITIVE, RELIANCE, OR COVER DAMAGES (INCLUDING WITHOUT LIMITATION FOR LOST PROFITS, LOST SAVINGS AND DAMAGE TO ANY DATA OR SYSTEMS); EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES AND EVEN IF A LIMITED REMEDY SET FORTH HEREIN FAILS OF ITS ESSENTIAL PURPOSE. YOU ARE SOLELY RESPONSIBLE AND LIABLE FOR VERIFYING THE ACCURACY AND ADEQUACY OF ANY OUTPUT FROM THE SOFTWARE, AND FOR ANY RELIANCE THEREON.

THE FEES FOR THE SOFTWARE AND SUPPORT, THE REMEDIES SET FORTH IN THIS AGREEMENT, THE LIMITS ON LIABILITY SET FORTH IN SECTIONS Limitation of Liability. EXCEPT TO THE EXTENT PROHIBITED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR'S OR ITS AFFILIATES' TOTAL, AGGREGATE LIABILITY ARISING FROM OR RELATED TO THIS AGREEMENT, THE SOFTWARE AND/OR SUPPORT (INCLUDING FOR NEGLIGENCE, STRICT LIABILITY, BREACH OF CONTRACT, MISREPRESENTATION, AND OTHER CONTRACT OR TORT CLAIMS), EXCEED THE AMOUNT OF YOUR DIRECT DAMAGES ACTUALLY INCURRED, UP TO THE AMOUNT OF FEES PAID TO LICENSOR UNDER THIS AGREEMENT FOR THE SOFTWARE PRODUCT OR SUPPORT PRODUCT THAT IS THE SUBJECT OF THE CLAIM UNDERLYING THE DAMAGES; OR, IN THE CASE OF SOFTWARE PROVIDED UNDER A TRIAL OR DEMONSTRATION LICENSE, ONE HUNDRED DOLLARS (\$100.00), WHICHEVER IS LESS. AND Exclusion of Damages. EXCEPT TO THE EXTENT PROHIBITED BY APPLICABLE LAW, UNDER NO CIRCUMSTANCES SHALL LICENSOR, ITS AFFILIATES, OR ANY OF THEIR SUPPLIERS OR LICENSORS BE LIABLE HEREUNDER FOR ANY OF THE FOLLOWING: (I) THIRD PARTY CLAIMS, EXCEPT AS PROVIDED IN SECTION 12, (II) LOSS OR DAMAGE TO ANY SYSTEMS, RECORDS OR DATA, (III) DIRECT DAMAGES FOR BREACH OF WARRANTY (IN RESPECT OF WHICH ANY LIABILITY SHALL BE LIMITED TO RE-PERFORMANCE OR REFUND AS SPECIFIED IN SECTION Exclusive Remedy. Licensor's sole obligation and liability, and your sole and exclusive remedy under the warranties set forth in Section 9, shall be for Licensor to use commercially reasonable efforts to have the problem remedied, to re-perform Support services, to deliver to you a replacement for the defective Software, or to refund fees paid (in each case, as determined by Licensor and as applicable), provided that Licensor is notified in writing of all warranty problems during the applicable warranty period.), AND/OR (IV) INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL, PUNITIVE, RELIANCE, OR COVER DAMAGES (INCLUDING WITHOUT LIMITATION FOR LOST PROFITS, LOST SAVINGS AND DAMAGE TO ANY DATA OR SYSTEMS); EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES AND EVEN IF A LIMITED REMEDY SET FORTH HEREIN FAILS OF ITS ESSENTIAL PURPOSE. YOU

ARE SOLELY RESPONSIBLE AND LIABLE FOR VERIFYING THE ACCURACY AND ADEQUACY OF ANY OUTPUT FROM THE SOFTWARE, AND FOR ANY RELIANCE THEREON. AND THE OTHER PROVISIONS IN THIS AGREEMENT REFLECT THE ALLOCATION OF RISKS BETWEEN THE PARTIES. THIS SECTION IS AN ESSENTIAL ELEMENT OF THE BASIS OF THE BARGAIN BETWEEN THE PARTIES.

VERIFICATION. Licensor shall have the right to have on-site audits periodically conducted of your use of the Software. These audits are for license verification purposes only, will generally be conducted during regular business hours, and Licensor will use its reasonable efforts not to interfere unduly with your regular business activities. Licensor may also require you to accurately complete a self-audit questionnaire in a form Licensor may have provided. If an audit reveals unauthorized use, you must promptly purchase sufficient licenses and Authorized Scope to permit all usage disclosed. If material unlicensed use is found (*i.e.*, license shortage of 5% or more), you also shall reimburse Licensor for all costs incurred in connection with the verification, including without limitation reasonable attorneys' fees.

REGULATED USES. You acknowledge that the Software has not been cleared, approved, registered or otherwise qualified (collectively, "Approval") by Life Technologies Corporation with any regulatory agency for use in diagnostic or therapeutic procedures, or for any other use requiring compliance with any federal or state law regulating diagnostic or therapeutic products, blood products, medical devices or any similar product (hereafter collectively referred to as "federal or state drug laws"). The Software may not be used for any purpose that would require any such Approval unless proper Approval is obtained. You agree that if you elect to use the Software for a purpose that would subject you or the Software to the jurisdiction of any federal or state drug laws, you will be solely responsible for obtaining any required Approvals and otherwise ensuring that your use of the Software complies with such laws.

EUROPEAN COMMUNITY END USERS. If this Software is used within a country of the European Community, nothing in this Agreement shall be construed as restricting any rights available under the European Community Software Directive, O.J. Eur. Comm. (No. L. 122) 42 (1991).

LEGAL COMPLIANCE; RESTRICTED RIGHTS. The Software is provided solely for internal research and solely for lawful purposes and use. You shall be solely responsible for, and agree to comply with, all applicable laws, statutes, ordinances, and other governmental authority, however designated. Without limiting the foregoing, this EULA is expressly made subject to any United States government laws, regulations, orders or other restrictions regarding export from the United States and re-export from other jurisdictions of equipment, computer hardware, software, technical data and information or derivatives of such equipment, hardware, software or technical data and information. You agree to comply with all applicable export and re-export control laws and regulations in regard to products (including computer hardware, software, deliverables, technical data, source code, or any other technology, equipment, and/or derivatives of such hardware, software, deliverables, technical data, source code, equipment, or any other technology) received from Licensor. You further certify that you will not, directly or indirectly, without obtaining prior authorization from the competent government authorities as required by those laws and regulations: (1) sell, export, re-export, transfer, divert, or disclose technical data or dispose of any product or technology received from Licensor to any prohibited person, entity, or destination; or (2) use the product or technology for any use prohibited by the laws or regulations of the United States. You will reasonably cooperate with Licensor and will provide to Licensor promptly upon request any certificates or documents as are reasonably

requested to obtain approvals, consents, licenses and/or permits required for any payment or any export or import of products or services under this EULA, at Licensor's expense. Your breach of this provision shall constitute cause for immediate termination of this EULA. You agree to indemnify and hold harmless Licensor, its affiliates, and their respective officers, directors, employees and agents for your noncompliance with this Section. The Software is a "commercial item," as that term is defined in 48 C.F.R. 2.101, consisting of "commercial computer software" and "commercial computer software documentation," as such terms are used in 48 C.F.R. 12.212. Consistent with 48 C.F.R. 12.212 and 48 C.F.R. 227.7202-1 through 227.7202-4, all U.S. Government End Users acquire the Software with only those rights set forth herein.

GOVERNING LAW; SEVERABILITY. This EULA shall be governed in all respects by the laws of the State of California, USA, without regard to its conflicts of law rules or principles. Any dispute arising out of or related to this EULA shall be resolved only in the state or federal courts having subject matter jurisdiction in California. This Agreement shall not be governed by the United Nations Convention on Contracts for the International Sale of Goods. Each party hereby consents to the exclusive jurisdiction and venue of such courts. If any provision of this EULA is held to be illegal or unenforceable for any reason, then such provision shall be deemed to be restated so as to be enforceable to the maximum extent permissible under law; the remainder of this EULA shall remain in full force and effect.

SOFTWARE MAINTENANCE AND SUPPORT. Licensor offers certain software maintenance and technical support service programs documented in Licensor's then-current Support policies ("**Support**"). Subscription to Support shall be governed by this EULA and by the terms and conditions set forth in such policies. Licensor shall have no obligation to provide Support if (a) the Software is not used in accordance with the Documentation or Authorized Scope; (b) the Software was modified by you; (c) you have not implemented all upgrades that would otherwise correct the problem; or (d) the problem is caused by your misuse, negligence or other cause within your control. Licensor may change its Support policies and prices at any time. Licensor reserves the right to discontinue Support services for any Software where Licensor generally discontinues such services to all Licensees of such Software, in which case such discontinuation shall not automatically terminate this EULA and the License. If you terminate Support and then re-enroll, Licensor may charge you a reinstatement fee.

GENERAL. This EULA, including any Orders, Support policies, and associated Licensor invoices (all of which are incorporated herein), are collectively the parties' complete agreement regarding its subject matter, superseding any prior oral or written communications, representations or agreements. In the event that any prior oral or written communication is in direct conflict with the terms of this EULA, this EULA shall control. You understand and agree that, to the extent Licensor permits you to use a non-Licensor purchase order or other form to order Software and/or Support, Licensor does so solely for your convenience. Any terms in any such forms that purport to vary or are in addition to or inconsistent with any terms in this EULA or in the applicable Order shall be deemed to be void and of no effect. Amendments or changes to this EULA must be in mutually executed writings to be effective. Sections 1, 4, 5, and 7 through 22, inclusive, shall survive the termination or expiration of this EULA. The parties are independent contractors for all purposes under this EULA. Neither party shall be liable for any delay or failure due to force majeure and other causes beyond its reasonable control; provided that the foregoing shall not apply to any of your payment obligations. Any notices under this EULA to Licensor must be personally delivered or sent by certified or registered mail, return receipt requested, or

by nationally recognized overnight express courier, to the address specified herein or such other address as Licensor may specify in writing. Such notices will be effective upon receipt, which may be shown by confirmation of delivery. All such notices shall be sent to the attention of the Chief Legal Officer of Life Technologies Corporation (unless otherwise specified by Licensor). You may not assign or otherwise transfer this EULA or any License without Licensor's prior written consent. This EULA shall be binding upon and inure to the benefit of the parties' successors and permitted assigns. You agree, at Licensor's request and reasonable expense, to provide reasonable assistance and cooperation to Licensor and its designees, and to give testimony and execute documents and to take such further acts reasonably requested by the other to acquire, transfer, maintain, perfect, and enforce Licensor's intellectual property rights as described in this EULA. To the extent you fail to do so, you appoint Licensor's or its affiliates' officers as your attorney in fact to execute documents on your (and your personnel's), successors' and assigns' behalf for this limited purpose.

v1.0



Glossary

alignment	The process of mapping sequencing reads to a reference genome or sequence.
alignment browser	An interactive software in which to view alignments of sequencing reads with the reference genome or sequence.
alignment score	Matching score; an optimal alignment is an alignment giving the highest score of matches.
allele	One of two or more alternative nucleotide sequences at the same location on homologous chromosomes.
analysis	A data analysis run in LifeScope™ Software; may include secondary analysis on data from multiple sequencer runs.
analysis type	An <i>analysis</i> based on common genetic-analysis experiment types. LifeScope™ Software provides three: Genomic Resequencing Analysis, Targeted Resequencing Analysis, and Whole Transcriptome Analysis.
annotated gene	Within one of several reference databases, a gene sequence that has biological attributes attached that describe structure or function, such as coding regions or biochemical function.
annotation	Biological attributes or metadata that are attached to sequence data or files. Examples include: genes and protein-coding features, and verified variants.
BAM file	The compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. The BAM file is generated from the mapping step and has special attributes to support SOLiD™ data.
barcode	A short, unique sequence that is incorporated into a library that enables identification of the library during multiplex sequencing.
barcode group, barcode pool	The conceptual grouping used during data analysis if more than one barcode is used during libration preparation from a single biological specimen.
barcoded library	A library that has a unique barcode sequence incorporated that enables identification of the library during multiplex sequencing.
base-space analysis	Analysis of SOLiD sequencing data that has been converted from color space to a nucleotide sequence.

Bayesian	An algorithm that is used to evaluate the probability of the existence of a heterozygote or a non-reference homozygote when the coverage at the position is not high.
bead	The substrate to which the individual strands of DNA molecules in a SOLiD™ library are attached. The covalently attached strand serves as the template for SOLiD sequencing after the bead is immobilized on the flowchip surface.
BED file	BED format provides a flexible way to define the data lines that are displayed in an annotation track. The required BED fields are: <ul style="list-style-type: none"> • chrom - name of the chromosome • chromStart - starting position of the feature in the chromosome (the first base in a chromosome is numbered 0). • chromEnd - ending position of the feature in the chromosome
BEDGRAPH	In targeted resequencing, a text file in BEDGRAPH format that describes the depth of coverage within targets by on-target alignments.
bisulfite read	A sequence read of bisulfite-converted DNA, in which unmethylated cytosine residues have been converted to uracil. Used in methylation analysis of genomic DNA.
BLAST	In bioinformatics, Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences.
call stringency	The criteria for calling a nucleic acid species or genetic variant present in the biological sample.
calling threshold	The criteria for calling a nucleic acid species or genetic variant present in the biological sample.
ChIP-Seq	Sequence analysis of genomic regions associated with DNA-binding proteins.
classic mapping	An algorithm that searches for matches at the full read length, allowing up to a user-specified number of mismatches.
clear zone	The threshold to decide whether a read is mapped uniquely in the reference.
clipping, hard clipping	An operation that codes a portion of a color alignment with a number of mismatches that prevent seed or anchor extension.
clone	A single bead with a clonal population of templates for sequencing, generated during templated bead preparation/emulsion PCR.
color balance	The relative proportion of beads in a given sequencing cycle that are called as each of the four colors.
color call	For each SOLiD chemistry cycle, the color that is determined for each bead.

color space, color-space analysis	Color space data is a sequence of colors obtained from 2-base encoded probes in SOLiD System sequencing, representing a series of overlapping dinucleotides. Color space data is converted to a nucleotide sequence by aligning to a reference.
complexity	The size of the genome that is composed of unique sequences. With respect to SOLiD™ libraries, the number of independent molecules derived from the original DNA input, not from amplification.
consensus calls	A file that lists the SNPs called with the SNP Finding algorithm, and provides general information about each position.
contig	A nucleotide sequence that has been assembled from shorter, overlapping sequences.
core	A single core is equivalent to a processor on a traditional computer. Multiple cores increase the performance and efficiency of a computer that is running multiple programs.
counting	Expression analysis that focuses on relative or absolute quantification of RNA molecules in a biological specimen.
copy number variation (CNV)	A variation in the number of copies of DNA segments among individuals in a population. The length of the segment can vary, and it is typically more than a few base pairs. Some use a broader definition that includes any insertion, deletion, or duplication longer than a few base pairs.
CountTags	A tool to count the number of reads that align within genomic features.
counting	Sequence analysis that generates read or tag counts for annotated regions of a reference sequence.
coverage	<ul style="list-style-type: none"> • In genomic analysis, the number of aligned (or mapped) sequencing reads that span a position in the reference genome. • In RNA analysis, this term is sometimes used to describe the fraction of the reference sequence that has sequencing reads aligned (or mapped), at a certain calling threshold.
csfasta file	The SOLiD System produces color-space sequence reads in a fasta-format labeled as CSFASTA. These reads can be retained and analyzed in color space.
dbSNP	A Single Nucleotide Polymorphism database repository for single nucleotide polymorphisms and short insertion and deletion polymorphisms, hosted by the NCBI.
<i>de novo</i> sequencing	The initial generation of the primary genetic sequence of a particular organism, without use of a reference genome or sequence.
deletion	A gap in a nucleotide sequence with respect to a reference genome or sequence.
SNP Finding	Analysis module to identify SNPs from mapped and processed SOLiD System color space reads.

discordant	In large indel analysis, mate-paired reads that have been mapped to a reference genome, and whose inter-read distance deviates significantly from the expected insert size of the mate-paired library as a whole. Discordant pairs have insert sizes that deviate significantly from the expected value. Discordant pairs containing a putative deletion appear larger when mapped to the reference.
EBI	European Bioinformatics Institute.
enriched data	A targeted resequencing tool, sequencing data that has been filtered for the target regions of interest.
enrichment statistics	A targeted resequencing tool, the parameters used to assess variations in coverage across all enrichment targets and on a per-target basis.
ENSEMBL	A genome browser. www.ensembl.org .
error correction	(SAET) This step takes each read in the input file and attempts to correct it if any of its k-mers are not present in the spectrum. Use of multi-threaded versioning may be recommended for speed of process.
Error expectation metric (EEM)	A whole transcriptome pipeline, a parameter that is used in the calculation of the Junction Confidence Value metric, to determine the statistical significance of detected variant fusion transcripts .
evidence graph	A visual representation of the data structure of a candidate fusion transcript.
exon	In eukaryotic organisms, a segment of a gene that encodes part or all of a protein. Exons may be separated by introns that are spliced out of the primary transcript to produce a mature mRNA for translation into protein.
exon mapping	A whole transcriptome pipeline, a step in which reads are mapped to the set of exon sequences defined by the genome annotation.
F3 tag	Sequencing data derived from the P1 end of the template in the SOLiD™ templated bead, using forward ligation chemistry. The F3 tag is generated using the SOLiD™ FWD1 Seq. Primers or the SOLiD™ Small RNA Seq. Primers. See the <i>5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Guide</i> (Part no. 4465650) for an illustration.
F5 tag	Sequencing data derived from the P2 end of the template in the SOLiD™ templated bead, using reverse ligation chemistry. The F5 tag is generated using the SOLiD™ REV1 (DNA) Seq. Primers or the SOLiD™ REV1 (RNA) Seq. Primers. See the <i>5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Guide</i> (Part no. 4465650) for an illustration.
false positive	Sequence variants that are present in the output data from a sequencing run but are not present in the biological source.

filter	A reference sequence of interest that is used to select reads that align with or map to the reference sequence for further analysis.
filter mapping	Alignment of sequencing reads to a reference sequence of interest, to select reads for further analysis.
Frequentist	An algorithm for finding SNPs (alternative to SNP Finding algorithm).
function code (dbSNPs)	A code in the dbSNP database that indicates the consequence, if any, of a SNP to the transcript in which it is located.
fusion transcript	An RNA molecule that results from transcription of a gene fusion. See also gene fusion
gapped alignment	A read alignment to the reference sequence that indicates an insertion or deletion. The pairing algorithm searches for gapped alignments (indels) when one of the tags (F3/R3/F5-P2) maps to the reference genome and the other tag does not map to the genome within the insert-size range. Small indel analysis calls indels from a consensus of gapped alignments, using the BAM file as input.
gene fusion	A section of the genome that maps to an exon from one gene followed by an exon from another gene. It can occur as the result of a translocation, deletion, or chromosomal inversion. A gene fusion junction excludes exon-to-exon boundaries that arise from alternative splicing of a transcript.
genomic classification table	For mate-paired or paired-end reads, a code that describes the strandedness, distance, and orientation of the two reads.
genomic mapping	See mapping.
genomics	Global analysis of the genome to discern elements involved in regulation of gene activity or expression, with an emphasis on genetic variation such as single nucleotide polymorphisms, small and large insertions and deletions, and other structural variants such as translocations and inversions. Some use the term genomics to as an umbrella term that includes transcriptomics, epigenomics, and analysis of the genome.
group	A collection of read-sets that are required to be analyzed together, in groups that you define. Reads that are grouped together are treated as one specimen, even if the reads are in different *.xsq input files.
hemizygous	A small indels tool, a parameter that indicates that the reference allele and the small indel are both present.
hg18, hg19	Reference sequence assemblies of the human genome using the nomenclature used by the UCSC Genome Browser.
highly expressed junctions	A whole transcriptome pipeline, high-count sequencing reads that span an exon-exon junction.
homozygous	In diploid organisms, having two identical alleles in the corresponding genes.

HUGO-style gene names	Gene names following the Human Genome Organization style.
HuRef	A diploid human genome sequence of one individual (J. Craig Venter).
indel	A difference in sequence due to either an insertion or a deletion event; especially used when the evolutionary direction of the change is unspecified.
insert size	The physical size of the genomic DNA segments or RNA molecules represented in a SOLiD™ library. <ul style="list-style-type: none"> • Fragment libraries: the size of the sheared DNA fragments. • Mate-paired libraries: the length of the genomic DNA segment spanned by the corresponding mate-pair tags. • Whole transcriptome libraries: the size of the RNA fragments.
insertion	An insertion of nucleotide sequence with respect to the reference genome or sequence.
intron	The genomic sequence between two exons that is spliced out of a primary transcript prior to translation. See <i>exon</i> .
inversion	A segment of DNA that is in its native location but is in the reverse orientation.
junction	A place where two regions that are not contiguous in the genomic sequence are joined in a single sequenced region under consideration.
Junction Confidence Value (JCV) metric	Whole transcriptome pipeline, a statistical confidence metric to detect false positives and assign a confidence level to a detected junction.
junction mapping	Whole transcriptome pipeline, a mapping step that incorporates exon, gene, and transcript definitions in the genome annotation for that region.
junk seed	An initial alignment in a seed-and-extend algorithm that is not aligned with the correct sequence
large indel	An insertion or deletion, with respect to a reference sequence, of more than 200 bp.
LB field	A field in a BAM file that contains library type information.
library	A set of DNA or cDNA molecules prepared from the same biological specimen and prepared for sequencing on the SOLiD System.
local mapping	The initial alignment step that starts with locating short matches between a read and the reference sequence.
locus-spanning	Pairs of reads derived from mate pairs or paired-end reads that span a relatively large distance in the reference sequence.

Map Fragment data tool	Tool that maps sequence data from fragment libraries.
mappability	In the CNVs analysis module, the fraction of mappable bases in the candidate CNV region.
mapping	The process of aligning sequencing reads to a reference genome or sequence.
mate alignment	The process of mapping paired-end or mate-paired reads to the reference sequence, taking into account the proximity of the reads on a template clone.
mate-paired library	Library consisting of two DNA segments that reside a known distance apart in the genome, linked by an internal adaptor, and with P1 and P2 Adaptors ligated to the 5' and 3' ends of the template strand , respectively.
mates	Sequencing reads that are linked through their origin in a mate-paired library or a paired-end sequencing run.
mer	The number of nucleotides in a sequencing read.
merging (WTA)	Algorithm parameter in paired-end WTA analysis that enables the merging of genomic and junction mappings.
methylation analysis	The study of how methylation of nucleic acids is involved in DNA structure and control of gene expression.
module	A virtual piece of LifeScope™ Software, such as mapping, SNP Finding, or CNV.
NCBI	National Center for Biotechnology Information.
node	A node is a single computer that is connected to the network. A multi-node cluster has several nodes (computers).
normalization	In general, the process of comparing an experimental measurement to a reference measurement. In the CNV analysis module, the process of comparing the relative coverage of a region of interest to the global coverage, based on the human reference genome. In targeted resequencing, the process of adjusting the amplicon amounts after PCR enrichment and before library preparation, so that amplicons are represented equally in the library. In the whole transcriptome pipeline, the parameter RPKM is a normalized measure of gene expression.
paired-end sequencing	Sequencing runs that acquire sequence from each end of the insert in a DNA fragment or whole transcriptome library, using both forward and reverse reads.
pairing	An algorithm that occurs after mapping in the LifeScope™ Software workflow

pairing distances	Pairing distances (sometimes called insert sizes) for each pair are assigned during the mapping/pairing pipelines and subsequently used by the large indel tool to determine indel candidacy.
parallelization	A tertiary analysis tool by which you can split spectrum generation into multiple jobs, where each job generates a subspectrum from the subset of reads.
pipeline	Sets of program tools that are used in sequence for different data analysis goals, such as whole transcriptome analysis or methylation analysis.
ploidy	The number of chromosome sets in a cell.
polyploidy	The number of chromosome sets in a cell.
polymorphism	A genetic variant in a population of individuals that may or not may be associated with an observable (phenotypic) trait.
primer set	In the SOLiD System, the set of primers that are used sequentially to initiate ligation sequence chemistry.
project	A project is a container for your analysis runs. You create your projects. Within each project, you create one or more analyses. Each project is private to an individual LifeScope™ Software user.
properties file	
p-value	The statistical significance of an alignment score is frequently assessed by its p-value, which is the probability that this score or a higher one can occur simply by chance, given the probabilistic models for the sequences.
quality value	An empirically defined value based on a phred-like score equating to the confidence that the color called for that cycle is the correct one. In general, the brighter the bead, the greater the difference in signal between the primary and secondary colors, the higher the quality value (QV).
R3 tag	The R3 tag applies only to mate-paired libraries; sequencing data derived from the mate-pair tag closer to the P2 end of the SOLiD™ templated bead, using forward ligation chemistry. The R3 tag initiates in the IA sequence using the SOLiD™ FWD2 Seq. Primers. See the <i>5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Guide</i> (Part no. 4465650) for an illustration.
raw reads	A format in which the nucleotide sequence appears without headers or comments.
read, sequencing read	Sequencing data from a single bead with a single primer set.
read-set	A group of reads belonging to one barcode from one *.xsq file.
read-set group	A collection of similar read-sets designated by a user.

read-set repository	A storage place in LifeScope™ Software for instrument data intended to be input data for LifeScope™ Software analyses.
reference, reference genome, reference sequence	A sequence against which sequencing reads are aligned before further bioinformatics analysis.
reference file	A file containing the reference sequence information, usually in the *.fasta format.
RefSeq	A multi-organism database archive of DNA, RNA, and protein sequences, hosted by the NCBI.
repository	A virtual container with your reference files, reads files, and projects. Repositories include: <ul style="list-style-type: none"> • Projects, which contains your projects. • Reads (read-sets), which contains your *.xsq data files. • References, which contains the reference genomes used in your analyses.
rescue	An alignment method that is applied to read pairs that have at least one alignment, but no pair of alignments occurring within an expected range. Use the rescue method to find additional alignments.
rescue distance	The maximum distance x for a pair considered to be properly paired (SAM flag 0x2) is calculated by solving equation $\Phi((x-\mu)/\sigma)=x/L*p_0$, where μ is the mean, σ is the standard error of the insert size distribution, L is the length of the genome, p_0 is prior of anomalous pair and $\Phi()$ is the standard cumulative distribution function.
resequencing	The process of genomic sequencing in cases where a reference sequence already exists, and the new sequence is compared to the reference.
resource manager	Software for executing batch and interactive jobs on a cluster of networked computers.
RNA-Seq	Gene expression analysis using sequence-based approaches. RNA-Seq can include whole transcriptome analysis, small RNA analysis, and SOLiD SAGE™ analysis.
RPKM	The number of reads mapping to a transcript per kilobase of transcript length per million mappable reads. RPKM is used to set a threshold for calling a transcript or new RNA species or isoform “present.” 1 RPKM is equivalent to 20 reads mapping to a 1 kb transcript per 20×10^6 mappable reads.
SAET	SOLiD Accuracy Enhancer Tool for correcting spectral-alignment errors in raw data; reduces color-calling error rate without alignment to reference genome. SAET takes as input a file with SOLiD reads in .csfasta format and outputs corrected reads and quality values into new .csfasta and .qual files, respectively.

SAGE™ analysis	(Serial Analysis of Gene Expression) Nucleotide sequence analysis seeking to find specific gene expression information using short stretches of cDNA (also known as <i>tags</i>) from the 3' ends of RNA molecules. In the SOLiD System, the SAGE tag is 25–27 bp in length.
sample, barcoded sample	In the 5500 Series SOLiD™ ICS, the set of templated beads that will be sequenced in a single flowchip lane. A barcoded sample contains templated beads from up to 96 barcoded libraries.
SAMtools	SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format. SAMtools is hosted by SourceForge.net.
Satay plot	A cross-axis graph displaying the spectral purity and signal intensity of beads in a sequencing run that meet defined thresholds in each parameter (<i>on-axis beads</i>). <i>Best beads</i> meet the highest stringency thresholds; <i>good beads</i> meet less stringent thresholds.
secondary analysis	Sequence alignment and mapping to a reference sequence.
seed	A field: size of seed used in spectrum construction. Seeding: Using a heuristic method, finding homologous sequences, not by comparing either sequence in its entirety, but rather by locating short matches between two sequences.
seed-and-extend	Seed-and-extend algorithms map reads to a reference genome. Certain softwares can map reads with any number of differences or mismatches using the observation that for an r bp read to align to the reference with at most k differences, the alignment must have a region of length $s=r/k+1$ called a seed that exactly matches the reference.
sense strand	The strand of DNA with the same nucleotide sequence as that of the corresponding mRNA. Also called the <i>coding strand</i> .
sequence trimming	See trimming .
sequencing sample	A physical pooling of one or more library preparations before templated bead preparation. The sequencing sample can be used on one or more lanes on a flowchip.
single read	Forward-only sequence data that generates the F3 tag .
small indel	An insertion or deletion, with respect to a reference sequence, of less than 10 base pairs in length.
small RNA analysis	Global sequence analysis of the small RNA population of a cellular RNA sample; small RNA includes microRNAs (miRNAs), short interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), and repeat-associated siRNAs (rasiRNAs).
small RNA library	A SOLiD™ System-compatible library that is prepared from the small RNA fraction of a total RNA sample.
SNP	Single Nucleotide Polymorphisms (SNPs); single base pair variants in genomic DNA or the corresponding RNA transcript.

spectrum building	(SAET) When a temporary file with spectrum constructed from all reads generates in the output directory. This process is fast, resource less, and creates temporary swap files when memory usage is above 2GB.
splice junction	Exon-to-exon boundaries that arise from alternative splicing of a transcript. See also gene fusion .
splicing	The process whereby introns are removed from a primary mRNA, resulting in a mature mRNA that is ready for translation into protein.
strandedness	The polarity or orientation of a nucleic acid strand with respect to being sense or antisense. Libraries prepared using the SOLiD Total RNA-Seq Kit preserve the strandedness of the original RNA molecule such that F3 tag reads align to the sense strand and F5 tag reads align to the antisense strand.
stringency	Stringency is one criterion used to filter out junk reads in ReadBuilder stage. Specifically, in the SNP Finding tool, the stringency parameter (call.stringency) defines the criteria by which to report SNPs in analyzing genomic data. There are four empirical stringency settings (low, medium, high, highest).
tag	A tag is: <ul style="list-style-type: none"> • Sequencing data from a single bead with a single primer set; sometimes used interchangeably with read, sequencing read. • A length of DNA or cDNA to be sequenced; especially, a relatively short stretch of DNA or cDNA that is used to infer information about the longer native molecule from which it is derived, such as in mate-paired library sequencing and SAGE™ analysis, respectively.
tags: BC, F3, F5, R3	Sequencing data derived from specific locations on the SOLiD™ templated bead. See the <i>5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Guide</i> (Part no. 4465650) for an illustration.
targeted resequencing	Comparative sequence analysis of selected candidate genes or regions to discover genetic variants and mutations.
template strand	The strand of each DNA molecule that is covalently attached to SOLiD™ P1 beads during templated bead preparation and that serves as a template for SOLiD™ System sequencing.
tertiary analysis	Data analysis that takes place after mapping and alignment.
threading	Threading is a technique that tries to match a target sequence on a library of known three-dimensional structures by “threading” the target sequence over the known coordinates. In this manner, threading tries to predict the three-dimensional structure starting from a given protein sequence. It is sometimes successful when comparisons based on sequences or sequence profiles alone fail due to a too low sequence similarity.
transcript rescue distance	See rescue distance .

transcriptome	The compilation of all transcribed sequences from a genome, both coding and non-coding.
transition	A single nucleotide (point) mutation that changes a purine nucleotide to another purine nucleotide (for example, A to G), or a pyrimidine nucleotide to another pyrimidine nucleotide (for example T to C).
translocation	A mutation in which a chromosomal segment changes position, usually moving from one chromosome to a different, nonhomologous chromosome.
transversion	A single nucleotide (point) mutation that changes a purine nucleotide to a pyrimidine nucleotide (for example, A to C), or a pyrimidine nucleotide to a purine nucleotide.
trimming	Filtering noise algorithmically which removes 5' overhangs and shortens 3' overhangs, typically to 4–5 bases. Trimming reads requires both modifying the .csfasta reads file to replace the color to be trimmed with a dot ("."), and modifying the mapping parameters to account for the alignment change.
trusted frequency (SAET)	(SAET) Use this developer option (-trustfreq freq) to overwrite estimated frequency cutoff of trusted seeds. All seeds with frequency < "freq" are filtered out of spectrum.
trusted seed	(SAET) If globally computed cutoff for frequency of trusted seeds does not meet your purpose, e.g., it is too low and too many junk seeds are considered correct or it is too high and many correct but low frequency seeds are filtered out, then use -trustfreq option to overwrite estimated frequency cutoff.
ungapped alignment	See gapped alignment .
uniquely placed reads	A read that is mapped only once in a genome with a given number of mismatches.
variant	A difference in the nucleotide sequence of interest, with respect to the reference sequence.
visualization	Viewing mapped reads on any number of publicly available genome browsers.
weighting	A weighting scheme can outperform a simple binary scheme traditionally applied in genomic analysis, independent of the organism, and used to improve the quality of the data.
whole transcriptome analysis	Global sequence analysis of coding and non-coding RNA transcripts along their entire length.
whole transcriptome library	A SOLiD™ System-compatible library that is prepared from total or poly(A) RNA that enables sequence analysis of the transcripts along their entire length.
workflow	Configuration files that are used for common end-to-end analysis runs, including multiple modules . LifeScope™ Software runs these common workflows: resequencing, targeted resequencing, whole transcriptome, Methyl Miner, and ChIP-Seq.

WTA	See whole transcriptome analysis .
*.xsq	eXtensible SeQuence, an extensible file format for storing sequence data. A binary sequence output file generated by the instrument. This file contains primary analysis results for a single lane in a flowchip.
z-normalize	Parameters called moving statistics, such as the number of locus-spanning pairs, and the sample average insert size, are calculated from a subset of pairs at each genomic position, yielding the absolute insert size deviation between the sample and the population in units of standard deviation. This normalization step allows multiple libraries with variable insert sizes to be combined into one analysis.
zygosity	The combination of <i>alleles</i> at a site in a nucleotide sequence; for example, homozygous reference allele, homozygous non-reference allele, or heterozygous.

Documentation

Related documentation

For the latest documentation on LifeScope™ Genomic Analysis Software, go to:

http://www3.appliedbiosystems.com/AB_Home/Support/index.htm

Document	Part number	Description
<i>LifeScope Genomic Analysis Software User Guide, Command Shell</i>	4465697	Describes how to use the software, via the command shell interface, for secondary and tertiary data analysis.

Obtaining support

For the latest services and support information for all locations, go to:

www.appliedbiosystems.com

At the Applied Biosystems web site, you can:

- Access worldwide telephone and fax numbers to contact Applied Biosystems Technical Support and Sales facilities.
- Submit a question directly to Technical Support.
- Order Applied Biosystems user documents, SDSs, certificates of analysis, and other related documents.
- Download PDF documents.
- Obtain information about customer training.
- Download software updates and patches.



Numerics

2BE 294

4BE 294

A

alignment 337

color 301, 302

format 293

mate 343

nucleotide sequence 26

score 301

unmapped and secondary 293

viewing and inspection 47, 302

alignment browser 337

alignment score 337

alignment, gapped 341

allele 337

analysis 337

create 78

delete 86

edit 82

methylation 343

reuse 81

review 85

SAGE 346

secondary 346

start 85

analysis results, view 87

analysis type 337

annotated gene 337

annotation 269, 337

CNV output file 283

conflicts 269

dbSNP 130 build 270

dbSNP 131 build 270

dbSNP tables 270

filtering options 270

function codes 275

gene features 275

hg18 reference build 270

hg19 reference build 270

large indels output file 282

mutated genes output file 285

optional annotation sources 272

options 270

parameters table 273

small indels annotated output file 279

SNPs output file 277, 286, 287

supported attributes 289

transitions and transversions 272

variant overlapping gene or exon 269

variant statistics output file 277, 282, 283

variants appearing in dbSNP 270

variants in exons 270

variants in genes 270

annotation parameters 273

B

BAM 97, 129

BAM file 337

header requirement 294

pairing information 300

with samtools command 294

WT format differences 191

barcode 337

barcode group 337

barcode pool 337

barcoded library 337

barcoded sample 346

base-space analysis 337

Bayesian 338

bead 338

BED file 338

BED format 220, 303

BEDGRAPH 338

bedgraph 110

BFAST

BAM file headers 295

Binary Alignment Map 26

bisulfite read 338

bisulfite reads 207
 BLAST 338
 browser
 IGV 88, 148, 161, 178, 211
 UCSC Genome Browser 148, 161, 178, 211

C

call stringency 338
 calling threshold 338
 ChIP-Seq 213, 338
 analysis 203, 213
 analysis tools 213
 BAM-to-BED format converter 220
 ChIP-Seq analysis software tools 220
 ChIP-Seq Map Data 216
 chromosome files 306
 classic mapping 338
 clear zone 338
 clip, hard 301
 clipping 338
 clipping, hard 301
 clone 338
 CNV 131
 CNVs 237
 color balance 338
 color call 338
 color quality value 339
 color space 339
 color attribute tags 300
 color-space analysis 339
 command shell
 common analysis scenarios 32
 concepts 29, 30
 repositories 31
 terminology 29, 30
 complexity 339
 computer cluster 34
 concordant 339
 consensus calls 339
 contig 339
 copy number variation (CNV) 339
 counting 339
 CountTags 339
 coverage 339
 create an analysis 78
 CSFASTA 97, 129, 204, 214

csfasta 26, 97, 129, 204, 214, 300
 csfasta file 339

D

dbSNP 339
 dbSNPs 341
 de novo sequencing 339
 delete an analysis 86
 deletion 339
 discordant 340
 distance
 rescue 345
 documentation, related 351

E

EBI 340
 ECC data 294
 edit an analysis 82
 enriched data 340
 enrichment statistics 340
 aligned reads input 98
 input files 97
 target regions file 97
 ENSEMBL 340
 ENSEMBL GTF file 166
 reformat_ensembl_gtf.pl 166
 error correction 340
 See also SAET
 Error expectation metric (EEM) 340
 evidence graph 340
 exon 163, 340
 exon mapping 340

F

F3 tag 340
 F5 tag 340
 false positive 340
 file formats 26, 300
 file types 26
 filename restrictions 30
 filter 341
 filter mapping 341
 filter reference fasta 27
 find CNVs
 See also CNV

Frequentist [341](#)
 function code (dbSNPs) [341](#)
 fusion junction [164](#)
 fusion transcript [341](#)
 fusion, gene [341](#)

G

gapped alignment [341](#)
 gene
 gene annotations [163](#), [165](#), [182](#)
 gene fusion [341](#)
 general parameters
 small RNA [153](#)
 genome
 annotations [165](#), [183](#), [305](#)
 genome, reference [345](#)
 genomic browser
 IGV [116](#), [302](#)
 UCSC Genome Browser [116](#), [148](#), [161](#), [178](#), [211](#),
 [303](#)
 genomic classification table [341](#)
 genomic mapping [341](#)
 genomics [341](#)
 gff file format [162](#)
 group [341](#)
 GTF files [305](#)
 annotation source [269](#)

H

hard clipping [302](#), [338](#)
 hemizygous [341](#)
 hg18
 annotations based on hg18 [272](#)
 genomic reference download from UCSC [305](#)
 hg18 reference build [270](#)
 hg18, hg19 [341](#)
 hg19
 annotations based on hg19 [272](#)
 highly expressed junctions [341](#)
 homozygous [341](#)
 HUGO-style gene names [342](#)
 Human Copy Number Variation
 output files [305](#)
 HuRef [342](#)

I

IGV viewer [302](#), [303](#)
 IGV. *See* Integrative Genomics Viewer
 indel [342](#)
 large [342](#)
 small [346](#)
 insert size [342](#)
 insertion [342](#)
 Integrative Genomics Viewer [88](#), [116](#), [148](#), [161](#), [178](#),
 [211](#)
 inversion [243](#), [342](#)

J

JCV metric [342](#)
 junction [342](#)
 splice [347](#)
 Junction Confidence Value (JCV) metric [342](#)
 junction mapping [342](#)
 junk seed [342](#)

L

large indel [255](#), [342](#)
 LB field [342](#)
 legacy format translation
 @HD SO field [309](#)
 @RG LB field [309](#)
 @SQ UR field [309](#)
 ##color-code [309](#)
 ##history [309](#)
 ##line-order [309](#)
 ##max-num-mismatches [309](#)
 ##max-read-length [309](#)
 ##primer-base [309](#)
 ##reference-name [309](#)
 library
 mate-paired [343](#)
 library [342](#)
 library, small RNA [346](#)
 local mapping [342](#)
 locus-spanning [342](#)

M

Map Data
 MethylMiner™ Map Data [207](#)
 Map Fragment data tool [343](#)
 mappability [343](#)

mapping 343
 filter 341
 genomic 341
 junction 342
 local 342
 quality 310
 matching file, .ma 343
 mate alignment 343
 mate-pair 243
 mate-paired library 343
 mates 343
 mer 343
 merging (WTA) 343
 methylation analysis 343
 mismatches
 color space 300
 dicolor 250
 module 343

N

NCBI 269, 343
 node 343
 normalization 343

P

paired-end sequencing 343
 pairing 343
 distances 344
 quality 310
 parallelization 344
 parameters
 small RNA, general 153
 penalty 344
 pipeline 344
 ploidy 344
 polymorphism 344
 polyploidy 344
 primer set 344
 project 344
 properties file 344
 p-value 344

Q

qual 97, 129, 204, 214
 quality value 344
 QV 97, 129, 204, 214

qv 26

R

R3 tag 344
 raw reads 344
 read 344
 sibling 346
 single 346
 read-set 344
 read-set group 344
 read-set repository 345
 reference
 file 345
 genome 345
 sequence 345
 reference data 305
 reference fasta 26, 27, 97, 129, 204, 214
 reference file
 multi-fasta 306
 reference file types 306
 reference repository
 initial content 37
 reference SNP identifier 269
 reformat_ensembl_gtf.pl 166
 RefSeq 345
 refSNP identifier 269
 repository 345
 requirements 34
 rescue 345
 distance 345
 resequencing 345
 resource manager 34, 345
 reuse an analysis 81
 review an analysis 85
 RNA library, small 346
 RNA sequencing 163
 RNA-Seq 345
 RPKM 345
 rsID 269

S

SAET
 usage 313
 SAET. *See* SOLiD™ Accuracy Enhancer Tool 345
 SAGE analysis 346
 sample 346

- sample, sequencing 346
- SAMtools 346
- samtools
 - fillmd command 310
 - index command 302
 - sort command 302
 - view command 294
- Satay plot 346
- scheduler 34
- secondary analysis 346
- seed 301, 302, 346
 - junk 342
 - seed-extend 301
- seed-and-extend 346
- sense strand 346
- sequence
 - trimming 346
- sequence, reference 345
- sequencing
 - paired-end 343
- sequencing read 344
- sequencing sample 346
- Serial Analysis of Gene Expression 346
- sibling read 346
- Single Nucleotide Polymorphisms 346
- single read 346
- small indel 99, 130, 259, 346
 - detection 259
 - output file formats 246, 258, 265
- small RNA count 84
- Small RNA Counts
 - gff output file format 162
- small RNA counts 153
 - BAM metadata 151
 - input files 151
 - output files 162
 - precursor sequence 151
 - Reads Per Million RPM 162
- small RNA coverage 84, 153
 - input files 151
 - output files 161
 - RNA.coverage.per.chromosome parameter 161
- small RNA library 346
- small RNA mapping 84, 152
 - input files 149
 - output files 161
- small RNA parameters 153
- SNP 346

- SNP Finding 99, 130, 249, 339
- SNPs 99, 130
- SNPs parameters 250
- SOLiD™ Accuracy Enhancer Tool (SAET) 345
- spectrum building 347
- splice junction 347
- splicing 347
- standard workflows 24
- start an analysis 85
- strandedness 347
- stringency 347
- system requirements 34

T

- tags 347
 - F3 340
 - F5 340
 - R3 344
- Targeted Resequencing 99
 - barcode support 96
 - supported analyses 96
 - viewing results in a genome browser 116, 148, 161, 178
- TORQUE 34

U

- UCSC Genome Browser 116, 165, 183, 303

V

- Variant Call Format 308

W

- whole transcriptome 181
 - alignmentReport.txt 178, 199
 - analysis 163
 - whole transcriptome analysis 163
- whole transcriptome coverage 84, 167
- WTA merging 343

X

- XSQ 97, 129
- XSQ file format 293
- xsq files
 - xsq file format 293, 294

Z

zygosity [104](#), [135](#), [260](#)



Headquarters

5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1 760 603 7200 | Toll Free in USA 800 955 6288

For support visit www.appliedbiosystems.com/support

www.lifetechnologies.com

