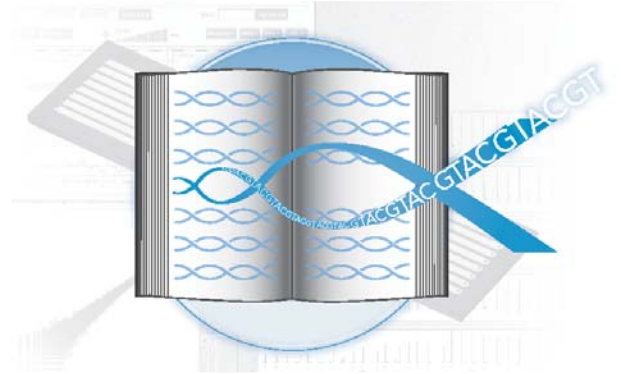


Genome Analyzer Pipeline Software User Guide

FOR RESEARCH ONLY





Notice

This publication and its contents are proprietary to Illumina, Inc., and are intended solely for the contractual use of its customers and for no other purpose than to operate the system described herein. This publication and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina, Inc.

For the proper operation of this system and/or all parts thereof, the instructions in this guide must be strictly and explicitly followed by experienced personnel. All of the contents of this guide must be fully read and understood prior to operating the system or any of the parts thereof.

FAILURE TO COMPLETELY READ AND FULLY UNDERSTAND AND FOLLOW ALL OF THE CONTENTS OF THIS GUIDE PRIOR TO OPERATING THIS SYSTEM, OR PARTS THEREOF, MAY RESULT IN DAMAGE TO THE EQUIPMENT, OR PARTS THEREOF, AND INJURY TO ANY PERSONS OPERATING THE SAME.

Illumina, Inc. does not assume any liability arising out of the application or use of any products, component parts, or software described herein. Illumina, Inc. further does not convey any license under its patent, trademark, copyright, or common-law rights nor the similar rights of others. Illumina, Inc. further reserves the right to make any changes in any processes, products, or parts thereof, described herein without notice. While every effort has been made to make this guide as complete and accurate as possible as of the publication date, no warranty or fitness is implied, nor does Illumina accept any liability for damages resulting from the information contained in this guide.

© 2008 Illumina, Inc. All rights reserved. **Illumina, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, and CSPro** are registered trademarks or trademarks of Illumina. All other brands and names contained herein are the property of their respective owners.

Revision History

Revision Letter	Date
A	January 2008

Table of Contents

Chapter 1	Overview	1
	Introduction	2
	Additional Information	2
	Genome Analyzer Pipeline Software Workflow	3
	Reporting Problems	4
	Technical Assistance	5
Chapter 2	Core Concepts	7
	Introduction	8
	Analysis Modules	8
	Understanding the Run Folder	10
	Run Folder Structure	11
	Images Folder	12
	Data Folder	12
	Run Folder Naming	13
	File Naming	14
	Parameters	14
	Paired Reads	14
	Calibration and Input Parameters	15
	Image Offsets	15
	Frequency Cross-Talk Matrix	16
	Phasing/Prephasing Estimates	17
	Sample Information	17
	Alignment Algorithms	18
Chapter 3	Running the Analysis	19
	Introduction	20
	Starting the Genome Analyzer Pipeline Software	20
	Running a Standard Analysis	21
	Parallelization Switch	21
	Nohup Command	21
	Command Line Options	22
	General Options	22
	GOAT Options	23
	GOAT and Bustard Options	23
	Paired Reads	24
	Makefile Targets	24

Chapter 4	Using GERALD	27
	Introduction	28
	GERALD Parameters	29
	ANALYSIS Variables.	29
	ANALYSIS Parameters	30
	Filtering Parameters.	31
	USE_BASES Option	31
	Lane-by-Lane Parameters	32
	FORCE Option.	33
	Rerunning the Analysis	33
	Contaminant Filtering	33
	GERALD Configuration File	34
	Lane-Specific Options	35
	Optional Parameters	35
	Paired-End Analysis Options	36
	Preparing the Reference Genome	37
	ELAND Alignments	39
	Missing Bases in ELAND	40
	Using ANALYSIS eland_tag	40
	Using ANALYSIS eland_extended	41
	Using ANALYSIS eland_pair	42
Chapter 5	Analysis Output	47
	Introduction	48
	Visual Analysis Summary	48
	Results Summary	48
	Cluster Intensity.	49
	Error Rates	50
	Text-Based Analysis Results	52
	Interpretation of Run Quality	54
	Summary.htm.	54
	IVC.htm	58
	All.htm and Error.htm	58
Chapter 6	Advanced Pipeline Usage	59
	Introduction	60
	Running Bustard as a Standalone Program	60
	Assigning a Control Lane.	60
	Running GERALD as a Standalone Program	61
	Additional "Make" Options.	61
	Running ELAND as a Standalone Program	62
	Compiling ELAND	62
	Command Line Syntax.	62
Appendix A	System Requirements and Software Installation	65
	Introduction	66
	System Requirements.	66
	Network Infrastructure	66
	Analysis Computer.	67

	Installation Prerequisites	69
	Setting Up Email Reporting	69
	Installing the Pipeline Software	71
	Compiling on Other Platforms	71
	Directory Setup	71
Appendix B	Output File Descriptions	73
	Introduction	74
	Output File Types	74
	Intensity Files	75
	Sequence Files	75
	Quality Score Files	76
	Efficiency	76
	Intermediate Output Data Files	77
	Output File Formats	80
	Parameters File Format	83
Appendix C	Using Parallelization	87
	Introduction	88
	“Make” Utilities	88
	Standard “Make”	88
	Distributed “Make”	88
	Customizing Parallelization	88
	Parallelization Limitations	91
	Memory Limitations	91




List of Figures

Figure 1	Three Steps of Data Analysis	2
Figure 2	Pipeline Modules	8
Figure 3	Run Folder Directory Structure	10
Figure 4	Frequency Cross-Talk Matrix and Phasing File Locations	16
Figure 5	Run Folder Structure and Output File Types	74

List of Tables

Table 1	Illumina Technical Support Contacts	5
Table 2	ANALYSIS Variables	29
Table 3	ANALYSIS Parameters	30
Table 4	USE_BASES Options	32
Table 5	Lane-by-Lane Parameters.	32
Table 6	GERALD Configuration File Parameters	34
Table 7	GERALD Configuration File Lane-Specific Options	35
Table 8	GERALD Configuration File Optional Parameters	35
Table 9	GERALD Configuration File Paired-End Analysis Options.	36
Table 10	Parameters for ANALYSIS eland_extended	42
Table 11	Parameters for ANALYSIS eland_pair	44
Table 12	Text-Based Analysis Results	52
Table 13	Example of Lane Results Summary	54
Table 14	Example of Expanded Lane Summary	54
Table 15	Data Volumes Per Experiment	66
Table 16	Intermediate Output File Descriptions	77
Table 17	Contaminant Filtering-Specific Files	79
Table 18	Final Output File Formats	80
Table 19	Intermediate Output File Formats	81



Chapter 1

Overview

Topics

- 2 Introduction
- 3 Genome Analyzer Pipeline Software Workflow
- 4 Reporting Problems
- 5 Technical Assistance

Introduction

The Genome Analyzer Pipeline Software (Pipeline) is a set of utilities designed to perform a complete offline data analysis of a sequencing run. It is supplied as source code and scripts.

Data analysis consists of three steps: image analysis, base calling, and sequence analysis.

- 1. Image analysis**—Uses the raw TIF files to locate clusters on the image, and outputs the cluster intensity, X,Y positions, and an estimate of the noise for each cluster. The output from image analysis provides the input for base calling.
- 2. Base calling**—Uses cluster intensities and noise estimate to output the sequence of bases read from each cluster, along with a confidence level for each base.
- 3. Sequence analysis**—Allows for alignment to a reference sequence, filtering of data based on predefined criteria, and visualization of the result.

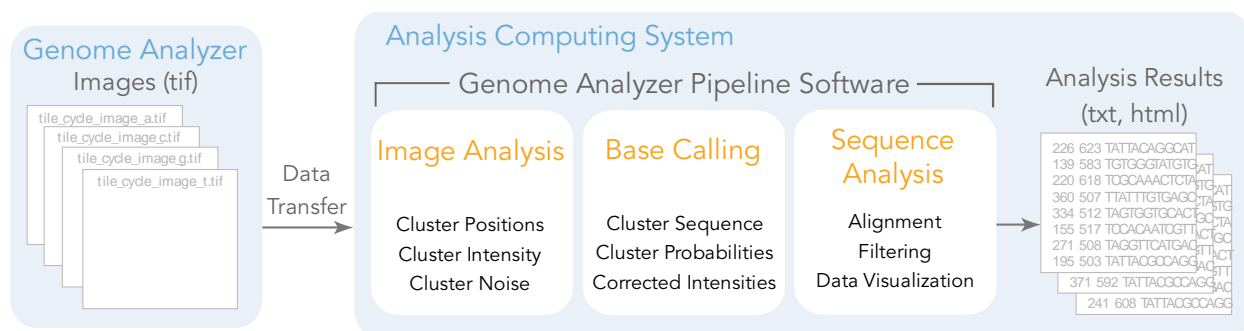


Figure 1 Three Steps of Data Analysis

The output data produced by the Genome Analyzer Pipeline Software are stored in flat text, tab-delimited files in a hierarchical folder structure called the Run Folder. The Run Folder includes all data folders generated from the Genome Analyzer and the data analysis structure. For a detailed description of the Run Folder structure, see *Understanding the Run Folder* on page 10.

The Pipeline requires a Linux system with specific processing and data storage capacity. For specific requirements, see *System Requirements* on page 66.

Additional Information

Additional information on the Genome Analyzer Pipeline Software can be found in the Pipeline/docs folder of your Pipeline software distribution.

Genome Analyzer Pipeline Software Workflow

The sequencing run image data are saved on the Genome Analyzer computer in a folder structure by cycle, lane, and tile number. The data are transferred to a network location for analysis after the sequencing run is complete or by mirroring the data to the storage location while the run progresses.

The following is an overview of the Pipeline workflow.

1. Install the Pipeline prerequisites on a suitable Linux system as described in *Installation Prerequisites* on page 69.
2. Install and compile the Pipeline using the “make” command.
For detailed information, see *System Requirements and Software Installation* on page 65.
3. Set up the “Instruments” directory for parameters files as described in *Directory Setup* on page 71.
4. Copy your run data to a location accessible to the Pipeline on your analysis computing system.
5. Create a configuration file that specifies what analysis should be done for each lane. The configuration file is described in *GERALD Configuration File* on page 34.
6. Change to the Run Folder location and generate Makefiles as described in *Starting the Genome Analyzer Pipeline Software* on page 20.
For detailed information on command line options, see *Command Line Options* on page 22.
7. Change into the newly created folder “Data/C1...Firecrest...” and start the analysis run as described in *Running a Standard Analysis* on page 21.
For parallelization computing requirements, see *Using Parallelization* on page 87.
For a description of the analysis output files, see *Text-Based Analysis Results* on page 52.

Reporting Problems

Contact Illumina Technical Support to report any issues with the Pipeline.

When reporting an issue, it helps to capture all the output and error messages produced by a run. This is done by redirecting the output using “nohup” or the facilities of a cluster management system. For an explanation of “nohup,” see *Running a Standard Analysis* on page 21.


It helps to attach the Makefile corresponding to the part of the Pipeline that is causing the problem. If there are GERALD-related issues, it helps to post the “config.txt” file found in the GERALD output folder. For problems relating to specific tiles or files, it is useful to send the output of “wc -l” and “ls -l” on these files.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 *Illumina Technical Support Contacts*

Contact	Number
Toll-free Customer Hotline (North America)	1-800-809-ILMN (1-800-809-4566)
International Customer Hotline	1-858-202-ILMN (1-858-202-4566)
Illumina Website	www.illumina.com
Email	techsupport@illumina.com



Chapter 2

Core Concepts

Topics

- 8 Introduction
- 8 Analysis Modules
- 10 Understanding the Run Folder
 - 11 Run Folder Structure
 - 13 Run Folder Naming
 - 14 File Naming
 - 14 Parameters
 - 14 Paired Reads
- 15 Calibration and Input Parameters
 - 15 Image Offsets
 - 16 Frequency Cross-Talk Matrix
 - 17 Phasing/Prephasing Estimates
 - 17 Sample Information
- 18 Alignment Algorithms

Introduction

Analysis modules perform the specific tasks of image analysis, base calling, and sequence alignment. During an analysis run, a defined folder structure is generated that captures the output of an instrument run in text files and parameters files. Parameters files contain calibration and input settings that optimize your analysis run and the alignment programs perform sequence analysis. This section describes these core concepts of the Genome Analyzer Pipeline Software.

Analysis Modules

The Pipeline is divided into modules that are managed by the “make” utility. The “make” utility is commonly used to build executables from source code and is designed to model dependency trees by specifying dependency rules for files. These dependencies are stored in a file called a Makefile. “Make” has a dual purpose within the Pipeline software:

- ▶ To build executables from source code
- ▶ To perform data analysis steps using the software

Each Pipeline module is a collection of Perl or Python scripts and C++ executables, and has its own Makefile associated with the analysis task. The script “goat_pipeline.py,” named after the General Oligo Analysis Tool (GOAT) calls the subscripts for three Pipeline modules: Firecrest, Bustard (“bustard.py”), and GERALD (“GERALD.pl”).

Any of the first two scripts can invoke the next script automatically, so there is no need to call more than one script for any given analysis run. Typically, the analysis begins with the image analysis script, “goat_pipeline.py.” However, if you need to reanalyze data, you can start with one of the other scripts and use different parameters.

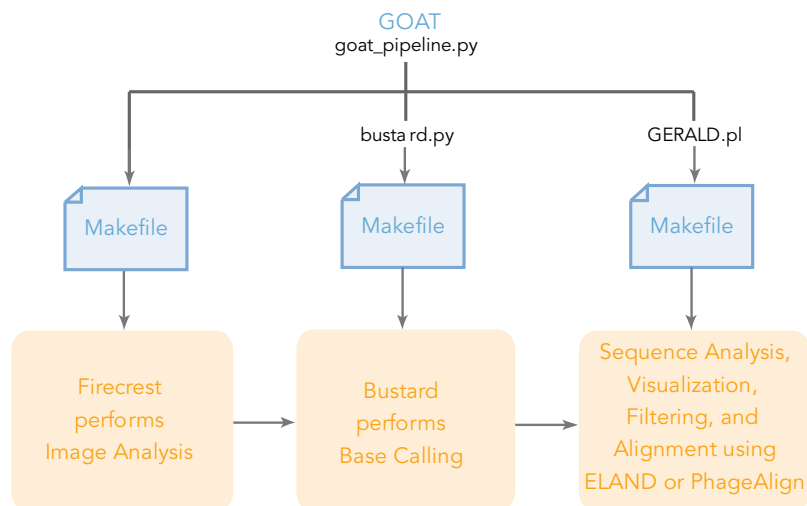


Figure 2 Pipeline Modules

- ▶ Firecrest is the module used for image analysis. Firecrest identifies cluster positions and extracts intensities. Through image filtering, it sharpens and enhances clusters, removes background noise, and detects clusters based on morphological features on the image. Firecrest also adjusts the scale and registration of an image.
- ▶ Bustard is the module used for base calling. Bustard deconvolves the signal from the clusters and applies correction for cross-talk, phasing, and prephasing.
 - Spectral cross-talk—The Genome Analyzer uses two lasers and four filters to detect four dyes attached to the four types of nucleotide, respectively. The frequency emission of these four dyes overlaps so that the four images are not independent. The frequency cross-talk is deconvolved using a frequency cross-talk matrix.
 - Phasing/Prephasing—Depending on the efficiency of the fluidics and the sequencing reactions, a small number of molecules in each cluster may run ahead (prephasing) or fall behind (phasing) of the current incorporation cycle. This effect is mitigated by applying corrections during the base calling step.
- ▶ **Generation of Recursive Analyses Linked by Dependency (GERALD)** is the module used for sequence alignment, data visualization, filtering, and alignment. The following two alignment programs work within the GERALD module:
 - **Efficient Large-Scale Alignment of Nucleotide Databases (ELAND)** is very fast and aligns for up to two errors from a reference for the first 32 bases. This algorithm is used for any reference larger than 100 Kbases.
 - PhageAlign does an exhaustive alignment (all possible alignments up to arbitrary edit distances), but is slow.

A run of the Pipeline is a two-stage process:

1. Generate the folders and Makefiles using one of the above scripts.
2. Start the Pipeline analysis by executing "make."

See *Starting the Genome Analyzer Pipeline Software* on page 20 for details.

In addition, the rest of this section describes the analysis and input parameters, and the command line options used in an analysis run.

Understanding the Run Folder

The Pipeline operates in a specific directory called the Run Folder where the images and analysis output files are saved by default in a hierarchical structure.

The following figure illustrates a typical Run Folder.

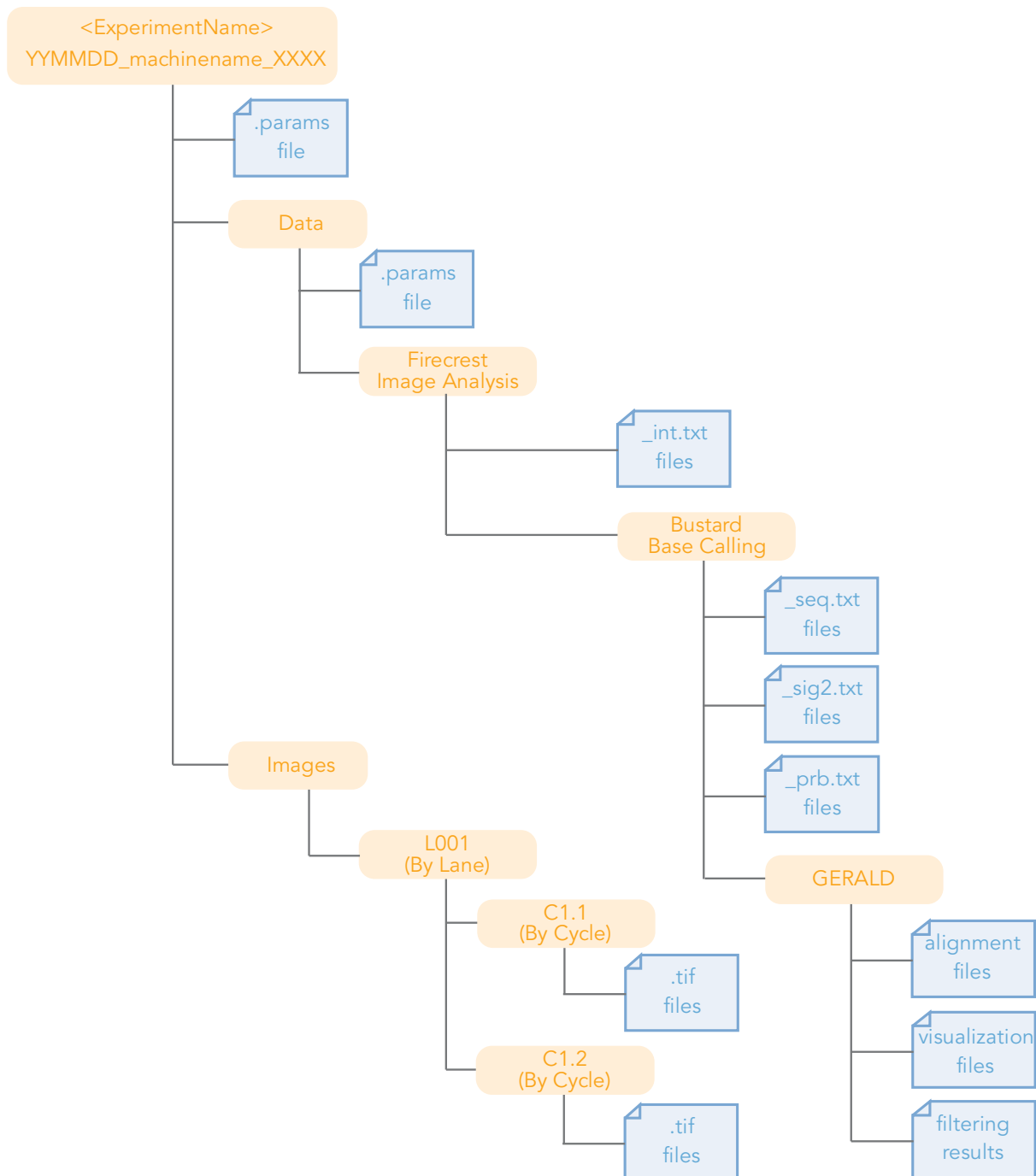


Figure 3 Run Folder Directory Structure

The standardized structure, file naming conventions, and file formats of the Run Folder allow for the following:

- ▶ A single point of data storage, logging, and analysis output during and after a run.
- ▶ Encoding sufficient information to trace the history of the data in the Run Folder back to the laboratory notebook without confusion between instruments, experiments, or sites.
- ▶ Standardized input and output enabling component software to operate without error, regardless of the instrument generating the data.
- ▶ Capturing and encoding enough information to independently reanalyze the data at any time, in such a way that existing extractions of sequence and related data are preserved, and parameters used during any point of the extraction process are captured and related to the subsequent output data.
- ▶ Subsequent analyses to be stored in the Run Folder.
- ▶ The software tools and other user software to implement and enforce these structures and standards.

Run Folder Structure

The Run Folder contains the Images folder and Data folder as illustrated in Figure 3. The Data folder contains Image Analysis folders and the Image Analysis folders contain Sequence folders.

- ▶ The Images folder holds the images from every tile for a given cycle of sequencing.
- ▶ The Data folder is created the first time analysis is initiated for a given experiment. Any analysis performed on the data is saved within the Data folder.

Each run of the main analysis modules creates a subdirectory in the Data folder of the Run Folder as follows:

- ▶ Each run of the image analysis software (Firecrest) creates a new image analysis output folder in the Data folder.
- ▶ Each run of the base calling software (Bustard) creates a new subdirectory in the image analysis subdirectory on which the base calls are based, resulting in a tree-like structure of analyses.
- ▶ Parameters and versions for any given analysis run are logged in the folder structure to make it possible to reconstruct any previous analysis run.

You can do multiple analyses of the data using different analysis parameters and the results will not be overwritten. The default naming convention consists of the number of cycles run, the version of the software used for the operation (Firecrest, Bustard), the date the analysis initiated, and the login of the user. If the user initiates a second analysis on the same day, a new folder structure is created and the results from the previous analysis are not overwritten.

Images Folder

The Images folder contains a subfolder for each lane that has been sequenced. The folders are named using the following convention where the lane number is padded to three digits:

<Sample-ID>_L<lane number>

If no sample-ID is known, only the lane number is used. For example, L001 contains the images taken in the first lane.

Each lane folder contains a subfolder for each cycle of sequencing. Each image-cycle subfolder contains four images for every tile, one for each of the four bases.

The Image folder naming follows the naming convention C<cycle number>.<version number>. Cycle number is indexed and represents the nth cycle. Version number allows for a cycle to be re-attained if the image acquisition were performed more than once, or the machine paused and a cycle repeated. For example, folders C1.1 and C1.2 would appear if images were acquired twice on the first cycle.

Within each image-cycle subfolder are four tif files for each tile. These files are named using the following convention:

<sample>_<lane>_<tile>_<base>.tif

In the example, s_1_67_g.tif, the "s" is the default sample-ID. Sample-IDs must not contain any underscores. Underscores are used as separators between the different identifiers of the filename to allow easy splitting by any software reading these filenames.

Data Folder

The Data folder contains a hierarchical structure that consists of the image analysis output folder, then the base calling output folder, and then the sequence alignment output folder.

A new subfolder is generated each time a set of images is processed by the image analysis module (Firecrest). These data are kept in one file per tile for raw intensities and use the extension _int.txt, and one file per tile for cluster noise and use the extension _nse.txt.

The Data folder contains a parameters file with multiple records corresponding to each subfolder which has been generated as a result of analyzing sets of images. The detailed information about the image analysis is stored one level above the corresponding data in the directory hierarchy. This allows a user to browse the different results of the image analysis without having to descend into the subfolders.

The parameters file explicitly records which cycle-image folders were used to generate the raw intensities and noise files, and any parameters used. It also records the name of the subfolder and the individual files within it. For a detailed description of the parameters file, see *Parameters* on page 14.

Image Analysis Folders

Each image analysis subfolder is named using the following convention:

```
C<first cycle>-<last-cycle>_<software><software-
version>_<date>_<user>
```

For example, C1-27_Firecrest1.8.20_31-07-2006_myuser.2 contains the second version of an analysis of cycles 1–27 performed using version 1.8.20 of the Firecrest software, run by the user “myuser” on the 31st of July 2006.

Base Calling Folders

Each image analysis folder may hold multiple sequence folders with the output of different runs of a base caller package. Each subfolder is named using the following convention:

```
<software><software-version>_<date>_<user>[.<version-number>]
```

For example, the folder name Bustard1.8.8_08-11-2005_myuser.3 represents the third run of the Bustard base caller on 8th of November 2005 by the user “myuser.”

Each image analysis folder also holds a parameters file that records any relevant information about the run of the base caller module.

Run Folder Naming

It is desirable to keep Experiment-Ids (or Sample-ID) and instrument names unique within any given enterprise. You should establish a convention under which each machine is able to allocate Run Folder names independently of other machines to avoid naming conflicts.

The top level Run Folder name is generated using three fields to identify the <ExperimentName>, separated by underscores. For example, YYMMDD_machinename_NNNN.

1. The first field is a six-digit number specifying the date of the run. The YYMMDD ordering ensures that a numerical sort of Run Folders places the names in chronological order.
2. The second field specifies the name of the sequencing machine. It may consist of any combination of upper or lower case letters, digits, or hyphens, but may **not** contain any other characters (especially not an underscore). It is assumed that the sequencing instrument is synonymous with the PC controlling it, and that the names assigned to the instruments are unique across the sequencing facility.
3. The third field is a four-digit counter specifying the experiment ID on that instrument. Each instrument should be capable of supplying a series of consecutively numbered experiment IDs (incremental unique index) from the onboard sample tracking database or a LIMS.

A Run Folder named 070108_instrument1_0147 indicates experiment number 147, run on instrument 1, on the 8th of Jan 2007. While the date and instrument name specify a unique Run Folder for any number of instruments, the addition of an experiment ID ensures both uniqueness and the ability to relate the contents of the Run Folder back to a laboratory notebook or LIMS.

Additional information is captured in the Run Folder name in fields separated by an underscore from the first three fields. For example, you may want to capture the flow cell number in the Run Folder name as follows:

```
YYMMDD_machinename_XXXX_FCYYY.
```

File Naming

Pipeline filenames have the following format:

```
<sample>_<lane>_[<tile>_] [<cycle>_] [<id>_] <type>.<filesuffix>
```

The individual components of the filename are:

Component	Description
<sample>	Alphanumeric string
<lane>	Single-digit number identifying a flow cell lane
<tile>	Four-digit number identifying a tile location in a flow cell lane
<cycle>	Two-digit number identifying a sequencing cycle
<id>	Single-digit number to distinguish files; for example, the different reads of a paired-end read
<type>	Alphabetical string identifying the type of content stored in the file
<filesuffix>	Suffix to identify the traditional file type

Example: s_1_0010_01_2_clu.txt is a valid filename.

Exceptions:

- ▶ For image (.tif) files, the <tile> location can have less than four digits.
- ▶ For image (.tif) files, the <tile> location may be replaced by two components identifying a row and column in the lane.

Parameters

The top level Run Folder, the Data Folder and subfolders, and the top level Image folder can all contain a parameters file. This read-only file is intended to contain any parameter data specific to the given level of information held in the folder.

For an example of the parameters file, see *Parameters File Format* on page 83.

Paired Reads

The simplest way to use paired-read data assumes that you have a single Run Folder containing the images for both reads, with a continuously incremented cycle count.

- ▶ For Genome Analyzer software SCS 1.0 and Pipeline version 0.3 and later, the Pipeline automatically knows where the second read starts.
- ▶ For older versions of Genome Analyzer instrument and analysis software, use the option `--new-read-cycle` to identify the start of the second read. For a description of the `--new-read-cycle` option, see *Command Line Options* on page 22.

An alternative way assumes that both reads of a pair are stored in two separate Run Folders. Specify both folders as arguments to "goat_pipeline.py." This generates output only in the first Run Folder and the second folder is not touched.

Calibration and Input Parameters

For an optimal analysis run, the Pipeline needs a number of calibration and input parameters. By default, the Pipeline auto-generates these parameters for each analysis.

Default offsets for runs on the same Genome Analyzer usually do not need to be changed. The Pipeline calculates these parameters automatically and uses them for the corresponding analysis steps.

For samples with biased-base compositions, as encountered in many tag-based or micro RNA applications, auto-calibration does not provide perfect results. For such samples, you need to dedicate one lane of the flow cell to a control sample and use the `--control-lane` command option to generate analysis parameters. For a detailed description, see *Command Line Options* on page 22.

Image Offsets

There are small pixel offsets among the four differently colored images taken of each tile. These are due to slightly different optical paths for each image. The Pipeline uses offsets to correct for this, and also corrects for linear rescaling of the image.

Each analysis run creates a file called `Data/default_offsets.txt` in the current Run Folder. The `Data/default_offsets.txt` file is used for subsequent analysis of the same run. If the file is located in `Instrument/<instrument>/default_offsets.txt`, the values in the file will be updated during the first run only. File locations are set using the `INSTRUMENT_DIR` variable, as described in *Directory Setup* on page 71.

The `default_offsets.txt` file contains four lines, corresponding to A, C, G, and T respectively, with four values each, using the A image as a reference. The following is an example of a typical `default_offsets.txt` file:

```
# Default offsets
0.00 0.00 0.00000 0.00000
-1.05 -1.62 -0.00017 0.00007
-1.20 -0.47 -0.00143 -0.00142
0.29 -0.92 -0.00159 -0.00142
```

The first two columns in a row correspond to the values of the X and Y offsets of the four images (in pixels).

The next two columns indicate scale factors applied to the image.

- ▶ A scale factor of 0 indicates that the image does not need to be rescaled.
- ▶ A scale factor of 0.001 for a 1000 x1000 pixel image indicates that images taken in the corresponding frequency channel tend to be one pixel larger than the reference channel.

Frequency Cross-Talk Matrix

The Genome Analyzer uses two different lasers to excite the dye attached to each nucleotide. The frequency emission of these four dyes overlaps, so the four images are not independent. As in Sanger sequencing, the frequency cross-talk has to be deconvolved using a frequency cross-talk matrix.

The frequency cross-talk is estimated during the analysis run and captured in a file called `s_matrix.txt`. The `s_matrix.txt` file is located in the Matrix folder as shown in Figure 4.

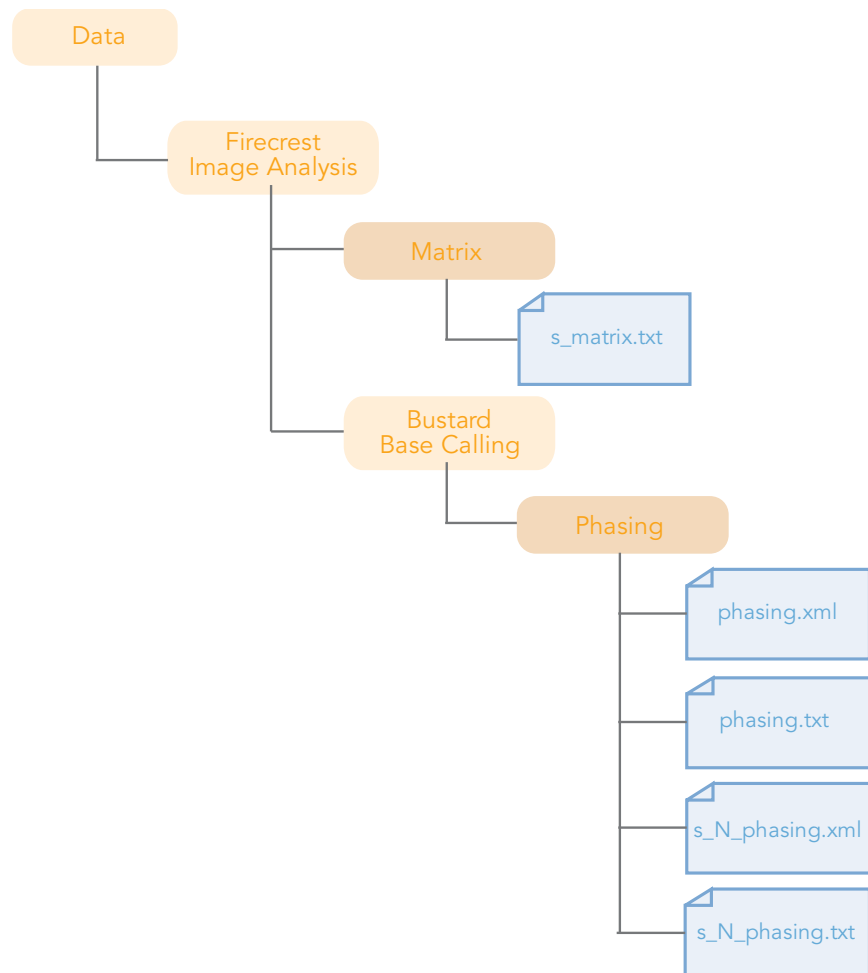


Figure 4 Frequency Cross-Talk Matrix and Phasing File Locations

The following is an example of a typical `s_matrix.txt` file:

```

# frequency response matrix definition
> C
> A
> T
> G
1.18  1.29  0.00  0.00
0.18  1.03  0.00  0.00
0.00  0.00  1.43  0.80
0.00  0.00  0.00  0.71
  
```

The lines starting with a greater than symbol (“>”) specify the order of the rows and columns in terms of the bases they represent.

The matrix elements show how the A, C, G, and T dyes/nucleotides (columns) cross-talk into the A, C, G, and T channels. A normal matrix should be diagonally dominant (diagonal elements tend to be the largest values) with the exception of the top-left and bottom-right corners (A/C and G/T cross-talk respectively). These are not as well-separated due to the fact that both corresponding dyes are excited by the same laser.

Phasing/Prephasing Estimates

Depending on the efficiency of the fluidics and the sequencing reactions, a small number of molecules in each cluster may run ahead (prephasing) or fall behind (phasing) the current incorporation cycle. This effect can be mitigated by applying corrections during the base calling step.

The phasing estimates are produced before a run of the base caller module and captured in a file called phasing.xml. The phasing.xml file is located in the Phasing folder as shown in Figure 4.

As the estimation uses statistical averaging over many clusters and sequences to estimate the correlation of signal between different cycles, the phasing estimates tend to be more accurate for tiles with larger numbers of clusters and a mixture of different sequences. Samples containing only a small number of different sequences do not produce reliable estimates.

Sample Information

Depending on the application, a reference genome may be supplied for the read sequences to be aligned against.

Alignment Algorithms

The Pipeline provides two alignment algorithms: PhageAlign and ELAND.

- ▶ PhageAlign performs an exhaustive alignment and always finds the best match but is very slow.
- ▶ Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) is very fast and used to match a large number of reads against the human genome with no more than two errors in the first 32 bases.

ELAND searches a set of large DNA files for a large number of short DNA reads allowing up to 2 errors per match. This description is based on the following definitions:

- ▶ **Large**—Human genome size and above, including small genomes
- ▶ **Large number**—At least eight million on a PC with 1 GB of RAM
- ▶ **Short**—32 bases or less

ELAND is much faster than PhageAlign but will only detect matches with two differences or fewer from your reads. This means that ELAND is less sensitive than PhageAlign, which will always find a best match (although possibly not a unique one) for your reads. Consider the following points when using ELAND:

- ▶ If your data is noisy, not all of it is going to align. If this happens to a significant proportion of your data, then it is possible that your data is too noisy to get good results.
- ▶ Error rates based on ELAND output underestimate the true error rate. Since reads with two or more errors in the first 32 bases do not get aligned, they do not contribute to the calculation.



Chapter 3

Running the Analysis

Topics

- 20 Introduction
- 20 Starting the Genome Analyzer Pipeline Software
- 21 Running a Standard Analysis
 - 21 Parallelization Switch
 - 21 Nohup Command
- 22 Command Line Options
 - 22 General Options
 - 23 GOAT Options
 - 23 GOAT and Bustard Options
 - 24 Paired Reads
 - 24 Makefile Targets

Introduction

This section describes the standard analysis run and command line options.

The standard invocation of the Pipeline assumes that you are performing image analysis, base calling, and sequencing alignment on a set of images in the Run Folder. It also assumes that the images are organized in a standard Run Folder directory structure as described in *Run Folder Structure* on page 11.

To successfully initiate image analysis, you need four images for each tile, for each cycle, and a parameters (.params) file in the Run Folder.

Starting the Genome Analyzer Pipeline Software

Although several different software programs are involved in an analysis run, a single command “goat_pipeline.py” will start the Pipeline automatically, and then trigger the launch of subsequent utilities.

This is the standard invocation of the Pipeline. Arguments contained in brackets [] are optional.

```
/path/Pipeline/Goat/goat_pipeline.py [--cycles=1-25|auto] [--tiles=s_1,s_2_0003,...]
  [--matrix=mymatrix.txt|auto|auto<n>] [--offsets=/path/default_offsets.txt|auto]
  [--phasing=0.01|auto|auto<n>] [--prephasing=0.01]
  [--directory=/path/C1-14_Firecrest1.8.20_01-08-2006_user] [--make]
  [--GERALD=/path/config.txt] [--control-lane=5]
  <run-folder-directory> [<run-folder-directory2>]
```

Some of the arguments above have sample values displayed. The only compulsory argument is the path to the Run Folder that is to be analyzed. The path can also point to any folder containing tiff images that are to be analyzed. Alternatively, you can provide a space-separated list of TIFF filenames.

Running a Standard Analysis

A standard analysis consists of calling the “goat_pipeline.py” script to generate an analysis directory using the “make” command, and then executing the “make” command.

Start a standard analysis run using the following command format:

```
Pipeline/Goat/goat_pipeline.py
  [--GERALD=<configfile>] [--make] <run-folder>
```

1. Type the following command to run a check on the Run Folder, report all detected folders and parameters files, and fill in any missing configuration options.

```
Pipeline/Goat/goat_pipeline.py
  --GERALD=/data/070813_ILMN-1_0217_FC1234/config.txt
  /data/070813_ILMN-1_0217_FC1234
```

Illumina recommends running this script before generating the Makefile to check for data integrity and consistency. It scans all the images folders and prints diagnostic output about the images and parameters files. No files or directories are modified on the data drive as a result of this command.

2. Add --make to the command listed above to create an analysis directory in the Run Folder. If you specify the --GERALD option, you will create the GERALD analysis folder and the corresponding Makefile.

```
Pipeline/Goat/goat_pipeline.py
  --GERALD=/data/070813_ILMN-1_0217_FC1234/config.txt
  --make /data/070813_ILMN-1_0217_FC1234
```

3. Change to the newly generated directory (for example, /data/070813_ILMN-1_0217_FC1234/Data/C1-26_Firecrest) and type the “make recursive” command. This command starts the actual analysis.

```
make recursive
```

The primary outputs are the sequences read with per-base quality values and, if alignment was performed, the alignments. These files can be found in the GERALD folder. The output files containing data statistics and histograms, used for quality control, can also be found in the GERALD folder.

A new output directory is created each time you rerun the analysis, so there is no need to remove any previous analysis files.

Parallelization Switch

If your system supports automatic load-sharing to multiple CPUs, you can parallelize the analysis run to <n> different processes by using the “make” utility parallelization switch.

```
make recursive -j n
```

For more information on parallelization, see *Using Parallelization* on page 87.

Nohup Command

You can use the Unix nohup command to redirect the standard output and keep the “make” process running even if your terminal is interrupted or if you log out. The standard output will be saved in a nohup.out file and stored in the location where you are executing the Makefile.

```
nohup make recursive -j n &
```

The optional “&” tells the system to run the analysis in the background, leaving you free to enter more commands.

Command Line Options

You can invoke the `goat_pipeline.py` and `bustard.py` scripts with a number of optional command line arguments.

General Options

Any of the following general options can be included in any order on a single command line.

--make

The `--make` command creates the analysis directory and a Makefile in the relevant analysis directory. You can start the analysis by changing to the directory and typing “make.” If this option is omitted, the Pipeline will not write any information to your Run Folder.

--new-read-cycle=<cycle>

Use this command to start a new read in a paired-end run. The calculation of the matrix correction and the application of the phasing correction will be reset at the specified cycle.

--GERALD=<config.txt>

Use this command to start the GERALD Makefile generator after the Bustard folder is created. You can specify multiple GERALD files by repeating the option with different configuration file names. For each GERALD configuration file specified, a separate GERALD subfolder is generated (under the same Bustard folder) with that configuration. For more information on the GERALD configuration file, see *GERALD Configuration File* on page 34.

--tiles=<tile>|<lane>[,<tile>|<lane>,...]

Use this command to select certain tiles for analysis. For example, `--tiles=s_1,s_3_0001,s_5_0002` selects all tiles in lane 1, selects only tile 1 in lane 3, and tile 2 in lane 5.

--cycles=<cycle>[-<cycle>[,<cycle>[-<cycle>...]]]:

Use this command to select certain cycles for analysis. For example, use `--cycles=3-31` to include only cycles 3 through 31 in the analysis.



If you skip cycles in the middle of a read, you cannot use ELAND to align the data.

Using the value “auto” tells the Pipeline to automatically select the lowest number of cycles present in any of the tiles and to make sure that all tiles have equal read lengths, regardless of the state of data acquisition/mirroring.

--compression=<method>

Use "--compression" to reduce the size of the Firecrest output. Allowed values are "none," "gzip" (the default), and "bzip2."

In the Pipeline v0.3, the intensity files are compressed by default. For previous versions, you must specify "--compression=none" on the command line.

GOAT Options

Use the following options with the "goat_pipeline.py" script.

--nobasecall

Use --nobasecall to skip the base calling step in the analysis.

--offsets=<filename>|autoldefault

Use --offsets=<filename> to specify a certain default offset file. If no offset file is specified, the Pipeline will create one in the Instruments folder.

**GOAT and Bustard
Options**

Use the following options with "goat_pipeline.py" and "bustard.py" scripts.

--control-lane=<n>

Use this command to select a lane <n> that is to be used to estimate phasing and matrix correction for all other lanes. This option is synonymous with --phasing=auto<n> --matrix=auto<n>. Control lanes are necessary for samples with skewed base compositions.

--matrix=<filename>|autolauto<n>|lane

Use the --matrix command to specify the frequency cross-talk matrix file, where filename refers to the path of the matrix file.

If no matrix is specified, or if you set the value to the default behavior "auto," the Pipeline auto-generates the matrix. A value of auto<n>, where <n> is a lane number between 1 and 8, is analogous to the --phasing=auto<n> option and allows the matrix estimation to be derived from only one lane.

--phasing=<x>|autolauto<n>

Use the --phasing command to apply a particular phasing correction. If you set the value to the default behavior "auto," the Pipeline auto-generates the phasing and prephasing values.

A value of auto<n>, where <n> is a lane number between 1 and 8, uses the automated phasing estimates from the corresponding lane. This is useful for samples with an uneven base composition (such as in gene expression), for which the current phasing estimator does not work reliably and phasing needs to be estimated from a single control lane.

You can specify a phasing value directly. For example, --phasing=0.01 indicates a phasing correction with a rate of 1% per cycle (1% of molecules in a cluster fall behind the other molecules). In this case, the option is normally combined with the --prephasing option.

--prephasing=<x>

Use the `--prephasing` command to apply a particular correction for prephasing. For example, using `--prephasing=0.01` sets a correction for prephasing with a prephasing rate of 1% per cycle.

The command `--prephasing=auto` is not recognized. Use `--phasing=auto` instead. By default the Pipeline autogenerates phasing estimates.

Paired Reads

The following additional variations on the “`goat_pipeline.py`” and “`bustard.py`” options are supported for paired reads.

--phasing=<read>:value, --phasing=<read>:<read>

Use this command to specify phasing options for one specific read of a pair.

The following example uses the default phasing option for read 1 but uses base phasing estimates from lane 5 for read 2:

```
--phasing=1:auto --phasing=2:auto5
```

The following example uses the phasing estimate for the second read and applies it to both read 1 and read 2:

```
--phasing=1:2
```

--matrix=<read>:value, --matrix=<read>:<read>

Use this command to specify matrix options for one specific read of a pair. This is analogous to the phasing options listed above.

Makefile Targets

Both “`goat_pipeline.py`” and “`bustard.py`” scripts generate Makefiles in the relevant image analysis and base caller directories that allow the complete analysis to be run by GNU Make. The Makefiles have the following advantages:

- ▶ Not all of the analysis needs to be run immediately.
- ▶ On a multiprocessor system or cluster, the analysis can easily be parallelized by specifying the “`-j`” option for “`make`.” For a description of parallelization, see *Using Parallelization* on page 87.
- ▶ In case of any failure or interruption during an analysis run, the run can easily be restarted at the last point.

The following optional targets are used with the “`make`” command.

all

All is the default Makefile target. It runs the complete analysis in the current directory (image analysis or base caller).

<sample>_<lane>

This target analyzes all tiles in a lane. For example, use `make s_1 s_2` to analyze lanes 1 and 2.

`<sample>_<lane>_<tileindex>`

This target analyzes a specific tile only. For example, use `make s_1_0007` to analyze lane 1, tile 7. This target is incompatible with auto-generated matrices and phasing estimates.

`_<tileindex>`

This target analyzes all tiles with the given index from any lane and is useful for analyzing randomly chosen subsets of a tile. For example, use `make _0020 _0040 _0060` to analyze tiles with indices 20, 40, 60 in all lanes.

This target is currently incompatible with auto-generated matrices and phasing estimates.

`clean`

This target removes all analysis output files. You would use “`make clean`” when you are low on disk space.



Using “`make clean`” removes all analysis results from the folder where the command is executed.

`recursive`

This target performs the analysis in the current directory and in all available subdirectories. Use this target to start a complete end-to-end analysis run from image analysis to sequence alignment using a single command.

The following example starts recursive full analysis:

```
make recursive
```

Specify the target by setting the `TARGET` environment variable. The following example removes all analysis results from ALL subfolders:

```
make recursive TARGET=clean
```

The following example performs a complete analysis for lanes 1 and 2. Multiple targets need to be enclosed in quotation marks.

```
make recursive TARGET="s_1 s_2"
```

`compress`

This target uses `gzip` to apply a loss-less compression to the output files after an analysis run. This significantly reduces the size of the analysis folders. Typically, the `Firecrest` and `Bustard` folders are reduced to 1/3 and 1/4 of their original size.

In the compressed state, no further analysis is possible. The folder must be uncompressed in order to reanalyze it.

`uncompress`

This target uncompresses a folder that has previously been compressed and returns it to its original state.

compress_images

This target uses bzip2 to compress the image data in the Images folder. This can take a significant amount of time, but reduces the size of the Images directory to about 60% of its original size.

In the compressed state, no further analysis is possible. The folder must be uncompressed in order to reanalyze it.

uncompress_images

This target uncompresses the Images folder that has previously been compressed and returns it to its original state.

Chapter 4

Using GERALD

Topics

- 28 Introduction
- 29 GERALD Parameters
 - 29 ANALYSIS Variables
 - 30 ANALYSIS Parameters
 - 31 Filtering Parameters
 - 31 USE_BASES Option
 - 32 Lane-by-Lane Parameters
 - 33 FORCE Option
 - 33 Rerunning the Analysis
 - 33 Contaminant Filtering
- 34 GERALD Configuration File
 - 35 Lane-Specific Options
 - 35 Optional Parameters
 - 36 Paired-End Analysis Options
- 37 Preparing the Reference Genome
- 39 ELAND Alignments
 - 40 Missing Bases in ELAND
 - 40 Using ANALYSIS eland_tag
 - 41 Using ANALYSIS eland_extended
 - 42 Using ANALYSIS eland_pair

Introduction

GERALD is the module that performs sequence alignments, visualization, produces statistics, and analysis output in a series of diagnostic QC plots and summary tables. These are presented in the form of html pages found in the GERALD output folder.

GERALD is usually run automatically as part of an overall Pipeline analysis but can also be run independently. For more information, see *Running GERALD as a Standalone Program* on page 61

As a result of running the "GERALD.pl" script, a new directory is created and named using the format GERALD_DD-MM-YYYY_user where the date is the current date and user is your computer login. If you want to rerun the analysis and change parameters, you can rerun GERALD with new parameters. A new directory will be created and no information will be overwritten.

GERALD uses multiple analysis parameters. Therefore, it is recommended to include the parameters in a configuration file and provide that file as input to GERALD.

You can define GERALD analysis parameters in the configuration file or in the command line. Command line arguments take precedence over parameters set in the configuration file. For a full description of analysis parameters and variables, see *GERALD Parameters* on page 29.

The following is an example of a GERALD invocation using a configuration file and command line arguments:

```
GERALD.pl config.txt --EXPT_DIR
/data/070813_ILMN-1_0217_FC1234/Data/C1-
27_Firecrest1.9.0_23-08-2007-user/Bustard1.9.0_23-
08-2007_user/
--FORCE --GENOME_DIR /data/Genomes --GENOME_FILE
BAC_plus_vector.fa
```

This section describes the GERALD parameters, analysis variables, configuration file options, and ELAND alignments.

GERALD Parameters

GERALD can be run in various analysis modes. Your analysis can be customized by specifying variables, parameters, and options.

ANALYSIS Variables Set the ANALYSIS variable to define the type of analysis you want to perform for each lane. The various analysis modes include default, sequence, eland, eland_extended, eland_tag, eland_pair, none, and monotemplate. You can mix and match analyses between lanes.

For all modes, except ANALYSIS none, you will get a sequence output file (s_N_sequence.txt) for each lane.

Table 2 ANALYSIS Variables

Variable	Alignment Program	Application	Description
ANALYSIS eland_extended	ELAND	Single reads	An improved version of ANALYSIS eland for analyzing single-read data. <ul style="list-style-type: none"> Better handling of > 32 base reads Each alignment is given a confidence value based on its base quality scores A single file of sorted alignments is produced for each lane For a detailed description, see <i>Using ANALYSIS eland_extended</i> on page 41.
ANALYSIS eland_pair	ELAND	Paired reads	Aligns paired-end reads against a target using ELAND alignments. A single-read alignment is done for each half of the pair, and then the best-scoring alignments are compared to find the best paired-read alignment. For a detailed description, see <i>Using ANALYSIS eland_pair</i> on page 42.
ANALYSIS eland_tag	ELAND	Gene Expression	Aligns reads to a non-redundant reference set of separate sequence tags and produces exact matches. For additional information, see <i>Using ANALYSIS eland_tag</i> on page 40.
ANALYSIS monotemplate	PhageAlign	Single reads	Aligns to a tag set using PhageAlign. Setting the parameter 6:ANALYSIS monotemplate performs monotemplate analysis for lane 6, where you can expect each read to be one of a small number (20 or less) of known template sequences. The reads are aligned using PhageAlign in tag mode, which treats each line of the reference sequence as a separate tag. No coverage plots will be produced since they are not relevant here. Some monotemplate-specific output is produced instead.
ANALYSIS sequence	None	Single reads Paired reads	Produces one file of sequence output per lane with no alignment. Setting the parameter 6:ANALYSIS sequence produces a file named s_6_sequence.txt. This file contains all sequences in a lane of a flow cell in an exportable format.

Table 2 ANALYSIS Variables (Continued)

Variable	Alignment Program	Application	Description
ANALYSIS sequence_pair	None	Paired reads	Produces two files of sequence output per lane, with no alignment. For example, s_1_1_sequence.txt and s_1_2_sequence.txt contain sequence output, one file for each half of the read pair.
ANALYSIS none	None	Any application	Omits the indicated lane from the analysis. Setting the parameter 8:ANALYSIS none ignores lane 8.
ANALYSIS default	PhageAlign	Single reads	Aligns each read against a reference sequence using PhageAlign. This mode is suitable only for small genome references.
ANALYSIS eland	ELAND	Single reads	Aligns each read against a large reference sequence using ELAND. Setting the parameter 6:ANALYSIS eland runs an ELAND whole-genome analysis for lane 6. You need to use ELAND if your reference sequence exceeds 1 MB in size. No coverage files will be generated. For more information on ELAND, see <i>ELAND Alignments</i> on page 39.
ANALYSIS expression	PhageAlign	Gene Expression	Aligns reads to a tag set using PhageAlign. This analysis mode is deprecated in favor of ANALYSIS eland_tag.

ANALYSIS Parameters

The content of the output file is determined by the following ANALYSIS parameters.

Table 3 ANALYSIS Parameters

Parameter	Description
USE_BASES	Use this parameter to identify bases to be used for alignment analysis. The USE_BASES string uses an asterisk (*) to indicate "fill up the read as far as possible with the preceding character." If USE_BASES all is set, all sequenced bases will show up in the analysis results. Otherwise, only cycles which have a Y at the corresponding position in the USE_BASES string will appear in the results. For a detailed description of USE_BASES syntax, see <i>USE_BASES Option</i> on page 31.
QF_PARAMS	Use this parameter if you want to use filtering different than the default filter. Set QF_PARAMS '(1==1)' to pass all of them. For information on default filtering, see <i>Filtering Parameters</i> on page 31.

Table 3 ANALYSIS Parameters (Continued)

Parameter	Description
SEQUENCE_FORMAT	<p>This parameter specifies what format to use for data export in the s_N_sequence.txt file. Allowed values are --fasta, --fastq, or --SCARF.</p> <ul style="list-style-type: none"> • fasta—This format is widely used but does not contain quality scores. • fastq—This format is an adaption of the fasta format that contains quality scores. However, the fastq format is not completely compatible with the fastq files currently in existence, which is read by various applications (for example, BioPerl). Because a larger dynamic range of quality scores is used, the quality scores are encoded in ASCII as 64+score, instead of the standard 32+score. This method is used to avoid running into non-printable characters. • SCARF (Solexa compact ASCII read format)—This easy-to-parse text based format, stores all the information for a single read in one line.
QUALITY_FORMAT	<p>Allowed values are --numeric and --symbolic.</p> <ul style="list-style-type: none"> • --numeric outputs the quality values as a space-separated string of numbers. • --symbolic outputs the quality values as a compact string of ASCII characters. Subtract 64 from the ASCII code to get the corresponding quality values. Under the current numbering scheme, quality values can theoretically be as low as -5, which has an ASCII encoding of 59=';'.

Filtering Parameters

GERALD uses filtering to remove low-quality base calls. Filters include CHASTITY, PURITY, and NEIGHBOUR:

- ▶ **CHASTITY**—The ratio of the brightest intensity over the sum of the brightest and second brightest intensities per base
- ▶ **PURITY**—The ratio of the brightest intensity over the sum of all of the four intensities per base
- ▶ **NEIGHBOUR**—Distance to the nearest cluster

Parameter thresholds filter out all clusters with a ratio less than or equal to 0.6 between the highest and the sum of the highest two intensities for the first 12 cycles.

USE_BASES Option

The USE_BASES option identifies which bases of a full read produced by a sequencing run should be used for the alignment analysis. A fully expanded USE_BASES value is a string with one character per sequencing cycle but more compact formats can be used as described in Table 4 on page 32. Each character in the string identifies whether the corresponding cycle should be aligned. The following notation is used:

- ▶ A lower-case “n” means ignore the cycle.

The first base could be ignored because it is part of the sequencing primer. The last base could be ignored because it is not corrected for prephasing and may have higher error rates.

It is important that the “n” is lower case. An upper case “N” signifies a deblock cycle, causing the Pipeline to try (and fail) to produce extra deblock-related output for that cycle.

- ▶ An upper-case “Y” means use the cycle for the alignment.
- ▶ A comma (,) denotes a read boundary used for multiple reads.
- ▶ An asterisk (*) means “fill up the read as far as possible with the preceding character.”
- ▶ A number means that the previous character is repeated that many times. Unspecified cycles are set to “n” by default.

The following table describes examples of USE_BASES options.

Table 4 USE_BASES Options

Option	Definition
USE_BASES nYYY	Ignore the first base and use bases 2–4.
USE_BASES Y30	Align the first 30 bases.
USE_BASES nY30	Ignore the first base and align the next 30 bases.
USE_BASES nY30n	Ignore the first base, align the next 30 bases, and ignore the last base.
USE_BASES nY*n	Ignore the first base, perform a single read, and ignore the last base. The length of read is automatically set to the number of sequencing cycles minus two.
USE_BASES nY*,nY*	Ignore the first base of each read and perform a paired read, resulting in the length of each read being set to the number of sequencing cycles associated with it minus one. The two reads do not need to be of the same length.
USE_BASES nY*	When used with ANALYSIS eland_pair, this is an abbreviation for USE_BASES nY*,nY*. When used with a single-read analysis mode, this means ignore the first base and perform a single-read.
USE_BASES all	Use all bases.

Lane-by-Lane Parameters

You can set the ANALYSIS parameter and other parameters on a lane-by-lane basis. You will need to do this for any parameters specific to the analysis of a particular lane.

Table 5 Lane-by-Lane Parameters

Option	Definition
6:GENOME_FILE mono1.txt	Specify the name of the file containing the reference sequence for lane 6.
67:GENOME_FILE mono1.txt	Specify the name of the file containing the reference sequence to use for lanes 6 and 7.

Table 5 Lane-by-Lane Parameters (Continued)

Option	Definition
<pre>3:QF_PARAMS='((NEIGHBOUR>=5.0)&&(PURITY>=0.7)&&((TILE!=4) (X_COORD>50)))'</pre>	<p>Set the quality filtering parameter, QF_PARAMS, on a lane-by-lane basis.</p> <p>For example, if the clusters near the left edge of tile 4 of lane 3 look questionable, use this command to set the filter.</p> <p>You can use any Boolean Perl expression as a filter parameter. The variable names are aliases to fields in a tab-separated text file.</p> <p>The best way to filter out individual tiles is to set BAD_TILES to be a list of the tiles you want to filter. See <i>Optional Parameters</i> on page 35 for an example of the BAD_TILES parameter.</p>

FORCE Option

The FORCE option creates GERALD directories and Makefiles. Without the FORCE option, GERALD will not create any directories and files and only operates in a diagnostic mode. You must specify this option to generate the GERALD analysis folder and subsequently run the analysis.

Rerunning the Analysis

The config.txt file used to generate an analysis is copied to the analysis folder so it can be used by GERALD if a reanalysis of the same data is required. To change parameters and rebuild the analysis, modify the configuration file and run the following command:

```
GERALD.pl config.txt --FORCE
```

By adding the OUT_DIR option, you can force GERALD to overwrite an existing Makefile. This way you can modify the analysis without directly editing the Makefile.

Contaminant Filtering

Contaminant filtering can be used with the variables ANALYSIS default and ANALYSIS eland. However, most of the time it is not needed.

GERALD attempts to filter contaminant sequences in a rigorous way by comparing the alignment of each read against the data versus the best alignment to the contaminant sequences. A one-sequence-per-line ASCII file is expected, with each sequence being at least READ_LENGTH bases in length.

To switch contaminant filtering on, specify the name of the file containing contaminant sequences in CONTAM_FILE. It is assumed that the CONTAM_FILE is located in GENOME_DIR. If it is not, specify the location in CONTAM_DIR.

GERALD Configuration File

This section describes a typical GERALD configuration file that uses the current features and parameters.

As part of the creation of the GERALD output folder, the GERALD configuration file specified (whether using the `--GERALD` option of `"goat_pipeline.py"` or directly as the argument to `"GERALD.pl"`) is copied to the GERALD output folder using the filename `"config.txt."` Some sites use standard configuration files, which may be stored in a central repository.

The GERALD configuration file specifies what analysis should be done for each lane, which GERALD translates into a Makefile. The Makefile specifies exactly what commands should be executed to carry out the requested analysis.

Table 6 GERALD Configuration File Parameters

Parameter	Definition
<code>EXPT_DIR</code> data/070813_ILMN-1_0217_FC1234/Data/C1-27_Firecrest1.9.0_23-08-2007-user/Bustard1.9.0_23-08-2007_user/	Provide the path to the experiment directory, if not specified on command line or auto-completed by <code>"goat_pipeline.py."</code>
<code>OUT_DIR</code> /home/user/Test4	Indicate the output directory, if other than a new GERALD folder inside of the <code>EXPT_DIR</code> folder.
<code>USE_BASES</code> nY*n	Ignore the first and last base of the read. The <code>USE_BASES</code> string contains a character for each cycle. <ul style="list-style-type: none"> • If the character is "Y," the cycle is used for alignment. • If the character is "n," the cycle is ignored. • Wild cards (*) are expanded to the full length of the read. For a detailed description of <code>USE_BASES</code> syntax, see <i>USE_BASES Option</i> on page 31.
<code>ELAND_GENOME</code> /home/user/Genomes/Eland/BAC_plus_vector/	Specify the genome reference for alignment with ELAND.
<code>GENOME_DIR</code> /home/user/Genomes	Specify where the genome file is located.
<code>ANALYSIS</code> eland	Align against a genomic sample and allow alignments to arbitrary positions of the provided reference.
<code>READ_LENGTH</code> 25	This parameter is no longer used with software version 0.3 and later. Specify the read length for the experiment. This is the read length used for the alignments, not the sequenced read length. Consequently, the value has to be less than or equal to the sequence length. It is useful to force the wild card expansion of <code>USE_BASES</code> to a predefined value.

Lane-Specific Options

The following table describes the lane-specific parameters in a GERALD configuration file.

Table 7 GERALD Configuration File Lane-Specific Options

Parameter	Definition
<code>7:USE_BASES nY20</code>	Align only 20 cycles for lane 7, starting with the second cycle.
<code>567:ANALYSIS sequence</code> <code>567:USE_BASES all</code>	Output sequence information for lanes 5, 6, and 7 only (no alignments are performed).
<code>8:ANALYSIS none</code>	Omit lane 8, which only contains primers.
<code>3:QF_PARAMS</code> <code>'((NEIGHBOUR>=3.0)&&(CHASTITY>=0.6)&&(X_COORD>50))'</code>	Filter parameters using the default (CHASTITY>=0.6). This example includes only those clusters with a separation of at least three pixels, with a CHASTITY filtering greater than or equal to 0.7, and an X coordinate greater than or equal to 50.

Optional Parameters

The following table describes the optional parameters in a GERALD configuration file.

Table 8 GERALD Configuration File Optional Parameters

Parameter	Definition
<code>EMAIL_LIST user@example.com</code> <code>user2@example.com</code> <code>EMAIL_SERVER mailserver</code> <code>EMAIL_DOMAIN example.com</code>	[Optional] Send notification to the user at end of an analysis run. For more information on email notification, see <i>Setting Up Email Reporting</i> on page 69.
<code>WEB_DIR_ROOT file://</code> <code>server.example.com/share</code>	[Optional] Include hyperlinks with a specific prefix to the Run Folder.
<code>BAD_TILES s_1_0001 s_2_0003</code>	Identify bad tiles. These tiles will be aligned but excluded from coverage.
<code>POST_RUN_COMMAND /yourPath/</code> <code>yourCommand yourArgs</code>	Allows user-defined scripts to be run after all GERALD targets have been built.

Paired-End Analysis Options

The following table describes the paired-end analysis options in a GERALD configuration file.

Table 9 GERALD Configuration File Paired-End Analysis Options

Parameter	Definition
<code>ANALYSIS eland_pair</code>	Use the paired-end alignment mode of ELAND to align paired reads against a target.
<code>USE_BASES Y*,nY*n</code>	Use all bases on the first read and ignore the first and last base of the second read.
<code>6:USE_BASES nY25</code>	Ignore the first base on both the first and second read; use 25 bases each and ignore any other bases.

For more information on USE_BASES syntax, see *USE_BASES Option* on page 31.

Preparing the Reference Genome

Several of the GERALD analysis modes (namely `eland`, `eland_extended`, `eland_pair`, and `eland_tag`) make use of the ELAND alignment program to align the reads produced by the Pipeline against a set of reference sequences. Before you can run an analysis that uses ELAND, you need to obtain the reference sequences you wish to align against in fasta format and convert or “squash” them into the format that ELAND can read. This is done by running a program `squashGenome` that is provided as part of the Pipeline installation.

The outcome of the squashing process is a folder containing a set of files that encode the reference sequences in a 2-bits-per-base binary format that is not human readable. Squashing only needs to be done once for each set of reference sequences you are interested in aligning against. For example, if you were doing some mouse and some human sequencing, you might create a folder containing a squashed version of the mouse genome and another folder containing a squashed version of the human genome (you probably do not want to squash both genomes into the same folder). Once created, a squashed folder can be copied between machines or placed on a shared drive, so as to be accessible from multiple machines. You specify the path of this folder as a parameter `ELAND_GENOME` when creating the configuration file for any analysis that involves ELAND.

See *Using ANALYSIS `eland_tag`* on page 40 for additional instructions on preparing a set of sequence tags for use as a reference sequence in `eland_tag` mode.

The fasta file format is very well known. Here is an example:

```
>chromosome:NCBI36:X:1:154913754:1
CTAACCTAACCTAACCTAACCTAACCTAACCTAACCTCTGAAAGTGG
ACCTATCAGCAG
GATGTGGGTGGGAGCAGATTAGAGAATAAAAGCAGACTGCCTGAGC
CAGCAGTGGCAACC
```

Please note that the names of the entries in any fasta files to be squashed cannot contain spaces. In `eland_pair` and `eland_extended`, the names of the entries cannot contain spaces, commas, or colons.

1. You must first create an empty folder for the squashed files to go into.
`mkdir path/myGenome`
2. Go to the location of the fasta format reference sequence files and enter the following command:
`<Pipeline>/Eland/squashGenome <path>/myGenome fastaFile1.fa [fastaFile2...]`
where `<Pipeline>` denotes the full path of the Pipeline installation location and `<path>` denotes the full path to the folder `myGenome` you created in step 1. This will cause files `fastaFile1.fa.2bpb` and `fastaFile1.fa.vld` to be created in folder `myGenome`.

Prior to Pipeline version 0.3, there was a restriction of a single entry per fasta file for the reference sequences. For Pipeline version 0.3 this restriction has been removed. If `fastaFile1` contained multiple entries, an additional file `fastaFile1.fa.idx` will be present in the folder `myGenome`.

For reasons of efficiency, ELAND thinks of the reference sequence as being in “blocks” of 16 MB, of which there can be at most 240. This limits the total length of DNA that ELAND can match against in a single run.

In a single ELAND run you can match against:

- ▶ One file of at most $240 \times 16 = 3824$ MB
- ▶ 239 files, each up to 16 MB in size
- ▶ Something in between, such as 24 files of up to 160 MB each. (The NCBI human genome will fit.)

Each file in the reference sequence must take up at least one block, so if you have a large number of short sequences to align against, you should place them in a single large file as individual fasta-format entries.

ELAND Alignments

Ensure the configuration file you use to run GERALD contains the following components:

- ▶ The path to your squashed genome files:

```
ELAND_GENOME /usr/local/share/eland/ncbi35
```

- ▶ The path to your list of repeats (optional):

```
ELAND_REPEAT /usr/local/share/eland/refs30_5
```

This can significantly speed up alignment against large targets.

- ▶ The analysis variable to run ELAND:

```
ANALYSIS eland
```

- ▶ Particular lanes that you want to analyze in the analysis variable:

```
34:ANALYSIS eland
```

This example indicates that lane 3 and 4 will be analyzed.



ELAND_GENOME refers to a directory, not a file. The usual GERALD variables GENOME_DIR and GENOME_FILE are not used for ELAND analysis. ELAND expects a different file format other than fasta.

You can only specify one ELAND_GENOME per lane.

After setting up the GERALD configuration file, you should be able to run “make.” The script `convertToFasta.pl` converts and concatenates all the reads into a single large fasta file which is then used as an input to ELAND. The script `convertFromELAND.pl` converts the results back into the by-tile PhageAlign format expected by the rest of the Pipeline.

ELAND operates on a lane-by-lane basis and uses up to 1 GB of memory. The Pipeline starts one ELAND job per lane. To prevent most computers from running out of memory, an artificial dependency in the GERALD Makefile prevents multiple instances of ELAND from running at the same time. You can remove this limitation by using the following option in the GERALD configuration file:

```
ELAND_MULTIPLE_INSTANCES 8
```

Be aware that this may use up to 8 GB of memory on your analysis computer. If insufficient memory is available, the analysis is likely to crash. Allowed values for this option are 1, 2, 4, and 8. A value of 1 indicates no lane parallelization and uses up to 1 GB of RAM, a value of 2 indicates two parallel jobs and uses up to 2 GB, etc.

Missing Bases in ELAND

Missing bases need to be specified as “N” characters and not “.” as in the sequence files. This conversion is managed automatically by GERALD but you need to be aware of it when running ELAND as a standalone program. For additional information on ELAND, see *Running ELAND as a Standalone Program* on page 62.

ELAND allows up to four external “N” characters at the beginning or end of the read. These bases are ignored.

ELAND allows up to two internal “N” characters, which are interpreted in one of two ways: “type D” for detection error and “type I” for insertion error.

- ▶ In a “type D” match, “N” indicates the base is there but not detected.

Read: ACNGT

Genome: ACCGT

- ▶ In a “type I” match, “N” indicates a base has been skipped.

Read: ACNGT

Genome: AC-GT

“Type D” and “type I” characters are given equal weight.

When lining up bases in your read with bases they align to in the reference, ignore any leading N characters and ignore any “type I” N characters because they are non-existent bases.

Most N characters are due to clusters wandering off the edge of the image for a cycle or two due to imperfect re-mapping of the tile position at different cycles. This produces a “type D” error. Otherwise, the Pipeline software will try to make a base call, even if the call is of low quality.

Using ANALYSIS eland_tag

For gene expression samples and other tag-counting applications, you can use ANALYSIS sequence to get purity-filtered sequences. These sequences are matched to the reference tag sets resulting in exact matches only. Using ANALYSIS eland_tag to align experimental reads to a reference set produces not only exact matches but also one or two mismatches.

In addition to the standard output files, ANALYSIS eland_tag creates a tag count file (s_N_tagcount.txt) for each lane that collapses and counts each distinct sequence. You can also specify GROUP_LANES. For example, GROUP_LANES 124 65 produces a file containing combined tag counts for each group of lanes in addition to a file for each lane.

ANALYSIS eland_tag uses ELAND to align to a non-redundant set of annotation tags. Illumina provides human and mouse annotation that consists of a non-redundant set of all possible GATC+16 or CATG+17 sequences in the genome and transcriptome, choosing the best annotation for each distinct sequence. You may also use publicly available annotation for SAGE tags or generate your own.

Squashing tag sets for eland

A separate fasta file should be prepared for each annotation tag set with one fasta header per file and each tag separated by Ns.

The following two examples show the beginning of a fasta file:

```
>mouseTranscrCanonicalCATG
AAAAAAAAAAAAAATCAC
NNNNNNNNNNNNNNNNNNNN
```

```

AAAAAAAAAAAACTTGA
NNNNNNNNNNNNNNNNNN
AAAAAAAAAACAGCAA
NNNNNNNNNNNNNNNNNN
AAAAAAAAAACATAAAT
NNNNNNNNNNNNNNNNNN
AAAAAAAAAAGAAAAAA
NNNNNNNNNNNNNNNNNN
AAAAAAAAAATGGCTAA
NNNNNNNNNNNNNNNNNN
AAAAAAAAAATGGGTCA
NNNNNNNNNNNNNNNNNN
AAAAAAAAATCCTTATGT
NNNNNNNNNNNNNNNNNN

```

```

>mouseMITO_CATG
CAAACCTCCATAGACCG
NNNNNNNNNNNNNNNNNN
TGTAATTTTACCTCTAA
NNNNNNNNNNNNNNNNNN
ACCAATGAACACTCTGA
NNNNNNNNNNNNNNNNNN
AAATCTTCTGGGTGTAG
NNNNNNNNNNNNNNNNNN
AACGGCTAAACGAGGGT
NNNNNNNNNNNNNNNNNN
CTAGTCCCTAATTAAGG
NNNNNNNNNNNNNNNNNN
AATATTTCAACAACAAA
NNNNNNNNNNNNNNNNNN
TTCCTAGTTGTTTATAG
NNNNNNNNNNNNNNNNNN
ACAAAAAATTGCTCCCC
NNNNNNNNNNNNNNNNNN

```

The fasta files are squashed as any other reference genome files. The first four bases (CATG or GATC) are stripped from the annotation tags because they are part of the sequencing primer.

ANALYSIS eland_tag aligns to one strand only. The annotation tags are non-symmetrical with a restriction site (GATC or CATG) on the left side only.

Using ANALYSIS eland_extended

ANALYSIS eland_extended is an improved version of the ANALYSIS eland mode. ANALYSIS eland can align reads longer than 32 bases but demands that the first 32 bases of the read have a unique best match in the genome. The position of this match is used as a “seed” to extend the match along the full length of the read. ANALYSIS eland_extended removes the uniqueness restriction by considering multiple 32 base matches to be considered and extended.

Configuring ANALYSIS eland_extended

There are two parameters that affect the output of the alignment, ELAND_SEED_LENGTH and ELAND_MAX_MATCHES. Both parameters can be specified lane-by-lane.

The following table describes the parameters for ANALYSIS eland_pair.

Table 10 Parameters for ANALYSIS eland_extended

Parameter	Description
ELAND_SEED_LENGTH	By default, the first 32 bases of the read are used as a “seed” alignment. Setting ELAND_SEED_LENGTH to 25, will use 25 bases for the initial seed alignment. This should increase the sensitivity since two errors per 25 bases is less stringent than two errors per 32 bases. A read is more likely to be repetitive at the 25 base level than at the 32 base level, so a decrease in ELAND_SEED_LENGTH should probably be used in conjunction with an increase in ELAND_MAX_MATCHES. Setting this to very low values will drastically slow down the alignment time and will probably result in a lot of poor confidence alignments.
ELAND_MAX_MATCHES	By default, ANALYSIS eland_extended will consider at most 12 alignments of each read. ELAND_MAX_MATCHES allows the maximum number of alignments considered per read to be varied between 1 and 255.

Both ANALYSIS eland_extended and ANALYSIS eland_pair share a common export file that contains all read, quality value, and alignment information for a lane of data.

- ▶ ANALYSIS eland_extended produces a single file per lane (s_N_export.txt).
- ▶ ANALYSIS eland_pair produces two files, one for each of the two reads (s_N_1_export.txt and s_N_2_export.txt).

For a detailed description of the export.txt files, see *Text-Based Analysis Results* on page 52 and *Output File Formats* on page 80.

Using ANALYSIS eland_pair

Based heavily on ANALYSIS eland_extended, ANALYSIS eland_pair allows the analysis of a paired-read run using ELAND alignments. As part of the analysis, it will:

- ▶ Remap all clusters across both runs to the clusters found in the first cycle of the first read.
- ▶ Reset the matrix and matrix calculation after the end of the first read.
- ▶ Generate sequence strings whose combined length is equal to the sum of the lengths of each individual read.

The following files are produced for read 1 and read 2, and have identical format and function to the corresponding single-read files:

- ▶ s_N_1_qraw.txt and s_N_2_qraw.txt
- ▶ s_N_1_eland_query.txt and s_N_2_eland_query.txt
- ▶ s_N_1_eland_multi.txt and s_N_2_eland_multi.txt
- ▶ s_N_1_frag.txt and s_N_2_frag.txt
- ▶ s_N_1_eland_extended.txt and s_N_2_eland_extended.txt

The script “pickBestPair.pl” compares `s_N_1_eland_extended.txt` and `s_N_2_eland_extended.txt`, along with their quality values, and produces following two files of alignments, which contain pairing information:

- ▶ `s_N_1_saf.txt` and `s_N_2_saf.txt`

These files are used to do quality value recalibration and generate the following files, which contain calibrated quality values:

- ▶ `s_N_1_qcal.txt` and `s_N_2_qcal.txt`

The script “pickBestPair.pl” is then re-run using the calibrated quality values to obtain two more files of alignments:

- ▶ `s_N_1_calsaf.txt` and `s_N_2_calsaf.txt`

Finally, these files are parsed into the following output files:

- ▶ `s_N_1_export.txt` and `s_N_2_export.txt`
- ▶ `s_N_1_sorted.txt` and `s_N_2_sorted.txt`

Another output file produced is `s_N_anomaly.txt`, which contains reads that do not align. For some applications, reads that do not align may be of interest, since amongst those that are due to read errors may be some that represent genuine differences between the sequenced DNA and the reference.

For a detailed description of the `export.txt` files, see *Text-Based Analysis Results* on page 52 and *Output File Formats* on page 80.

Configuring a Paired-Read Analysis

The alignments of the two reads that provide input to the pairing process may be varied by setting `ELAND_SEED_LENGTH` and `ELAND_MAX_MATCHES`. Both parameters may be set lane-by-lane, but the same values will apply to each of the two reads in a lane.

The paired-read analysis may be configured by passing options to `pickBestPair`. This is done by setting a parameter `PAIR_PARAMS` in the GERALD config file. For additional information, see *GERALD Configuration File* on page 34.

`PAIR_PARAMS` can be specified lane-by-lane. All of the options must be specified on a single line and space-separated, as in the following example:

```
8:PAIR_PARAMS --circular --min-percent-unique-pairs=30
```


The following table describes the parameters for `ANALYSIS eland_pair`.

Table 11 Parameters for ANALYSIS eland_pair

Parameter	Description
--circular	<p>This causes pickBestPair to treat each chromosome as circular and not linear, enabling it to detect valid pairings that “wrap around” when the two alignments are mapped onto the linear representation of the chromosome. [Optional]</p> <p>--circular=my_mitochondria_file.fa Treat alignments to my_mitochondria_file.fa as circular but other chromosomes as linear (as you might want to do when e.g. aligning to the whole human genome)</p> <p>--circular=chromosome1:100000,chromosome2:300000 Specify chromosomes to circularize and specify the size to “wrap around” (possibly of use when the chromosome size is uncertain)</p>
--min-percent-unique-pairs	<p>A unique pair is defined as a read pair such that its constituent reads can each be aligned to a unique position in the genome without needing to make use of the fact that they are paired.</p> <p>pickBestPair works in a two-pass fashion:</p> <ol style="list-style-type: none"> 1. On the first pass it looks for all clusters that pass the quality filter and have a unique alignment of each of their two reads, then uses this information to determine the nominal insert size distribution and the relative orientation of the two reads. 2. On a second pass this information is used to resolve repeats and other ambiguous cases. <p>The number of unique pairs, expressed as a percentage of the total number of clusters passing filters, must exceed a certain percentage. Otherwise, no pairing is attempted and the two reads are effectively treated as two sets of single reads.</p> <ul style="list-style-type: none"> • By default, this threshold is set to 30%. • For low quality data, a pairing can be forced by setting --min-percent-unique-pairs=5. • For some applications it may be useful to switch off the pairing completely. Set --min-percent-unique-pairs=101.
--min-percent-consistent-pairs	<p>Of the unique pairs, the vast majority should have the same orientation with respect to each other. If they don't, it is indicative of the following problems:</p> <ul style="list-style-type: none"> • Sample prep • Circularization is not switched on • A reference sequence is extremely diverged from the sample data <p>In such cases, no pairing is attempted and the two reads are effectively treated as two sets of single reads.</p> <p>By default, the threshold for this parameter is set to 70%.</p>
--min-paired-read-alignment-score	<p>For each cluster, all possible pairings of alignments between the two reads are compared. This is the score of the best one. Since we are considering the two reads as one, both reads in a cluster get the same paired-read alignment score.</p> <p>The alignment score is nominally on a Phred scale. However, it is probably not safe to assume the calibration is perfect. Nevertheless, it is a good discriminator between good and bad alignments. The score must exceed this threshold to go in the sorted.txt file.</p> <p>The default value is zero.</p>

Table 11 Parameters for ANALYSIS eland_pair (Continued)

Parameter	Description
--min-single-read-alignment-score	<p>Each read is given a single-read alignment score. This is identical to the alignment score from an eland_extended analysis. If a read has a zero paired-read alignment score, but a single-read alignment score that exceeds this threshold, its alignment will still go in the sorted.txt files.</p> <p>If the alignments of the two reads can not be paired (resulting in a zero paired score) and only one of the reads has an alignment exceeding --min-single-read-alignment-score, the read pair is treated as a singleton. The alignment of the shadow read is unreliable enough to be ignored. The default value is zero.</p>
--add-shadow-to-singleton-threshold	<p>If one read has a score exceeding --min-single-read-alignment-score but the other read either has no alignments or an alignment that does not exceed --min-single-read-alignment-score, then the non-aligning "shadow" read is added to the sorted.txt file with a zero alignment score, if the combined base quality of the shadow read (not alignment quality) exceeds this threshold.</p> <p>The default value of 1,000,000 indicates this feature is switched off.</p>



Chapter 5

Analysis Output

Topics

- 48 Introduction
- 48 Visual Analysis Summary
 - 48 Results Summary
 - 49 Cluster Intensity
 - 50 Error Rates
- 52 Text-Based Analysis Results
- 54 Interpretation of Run Quality
 - 54 Summary.htm
 - 58 IVC.htm
 - 58 All.htm and Error.htm

Introduction

The Pipeline produces various text-based files and visual output (png and htm formats) during an analysis run. This section will help you interpret the various files that appear in an analysis output directory.

Visual Analysis Summary

The results of an analysis are summarized as web pages that enable a large number of graphs to be viewed as thumbnail images. This section is intended to help you interpret the various graphs that appear in an analysis directory.

As the numbers of tiles and graphs have increased, it has become impractical to generate every possible graph for every tile. Therefore, the pages should be considered as a very basic view of the data.

Results Summary

For each Run Folder, a Summary.htm file is produced, which contains the results of your analysis run.

Summary.htm

Comprehensive results and performance measures are saved in Summary.htm. This is an entirely text-based file and it will load in reasonable time, even for a large experiment. It is located in the GERALD folder and contains links to all other web pages.

The following tables are included in a Summary.htm page:

- ▶ **Chip Summary** contains information about the instrument and the flow cell.
- ▶ **Chip Results Summary** contains the number of clusters before and after filtering and the yield in kilobases.
- ▶ **Lane Parameters Summary** contains information about the sample and the analysis performed in each lane, including the following:
 - Sample target
 - Sample type
 - Read length
 - Filter (For example, ((CHASTITY>=0.6))

Each of these items correspond to the GENOME_FILE/ELAND_GENOME, ANALYSIS, READ_LENGTH, and QF_PARAMS parameters in the GERALD configuration file. (READ_LENGTH does not have to be specified in Pipeline version 0.3.)

- ▶ **Lane Results Summary** contains basic data quality metrics for each lane, including the following:
 - Average number of clusters per tile
 - Intensity values
 - Percentage of clusters passing filtering

For low density flow cells (up to 6000 clusters), the percentage of clusters passing filtering should be greater than 70%.

For high density flow cells (20000–35000), the percentage of clusters passing filtering is usually between 40% and 50%. Clusters are randomly distributed across the flow cell surface and the number of overlapping clusters increases with the density.

- Alignment scores
- Average error rate per lane
- ▶ **Expanded Lane Summary** contains more detailed quality metrics for each lane, including the following:
 - Phasing percentages
 - Raw and filtered error rates
 - Raw and filtered number of perfect clusters
 - Filtered intensities, loss, and alignment
- ▶ **Pair Summary** contains two per-tile summary tables (one for each read) for lanes for which eland_pair was performed. These are preceded by a set of tables collectively entitled the Pair Summary. These provide statistics about the alignment outcomes of the two reads individually and as a pair, the latter including relative orientation and separation (insert size) of partner read alignments.



NOTE

Although the Summary.htm file is an HTML file, it is also a valid XML that is parseable by Perl's XML::Simple module. This means that you can mine the numbers in a Summary.htm file via a Perl script.

For information on interpreting results in the Summary.htm file, see *Interpretation of Run Quality* on page 54.

Cluster Intensity

Key web pages that illustrate cluster intensity are IVC.htm and All.htm.

IVC.htm

This file contains plots that display lane averages over all tiles in the lane. The plots displayed are All, Called, %Base_Calls, %All, and %Called.

- ▶ **All**—This is the lane average of the data displayed in All.htm. It plots each channel (A, C, G, T) separately as a different colored line. Means are calculated over all clusters, regardless of base calling. If all clusters are T, then channels A, C, and G will be at zero. If all bases are present in the sample at 25% rate and a well-balanced matrix is used for analysis, the graph will display all channels with similar intensities. If intensities are not similar, the results could indicate either poor cross-talk correction or poor absolute intensity balance between each channel.
- ▶ **Called**—This plot is similar to All, except means are calculated for each channel using clusters that the base caller has called in that channel. If all bases are present in the sample at 25% with pure signal (zero intensity in the non-called channels), the Called intensity will be four times that of All, as the intensities will only be averaged over 25% of the clusters. For impure clusters, the difference in intensity will be less than four times that of All.

The Called intensities are independent of base representation, so a well-balanced matrix will display all channels with similar intensities.

- ▶ **%Base_Calls**—The percentage of each base called as a function of cycle. Ideally, this should be constant for a genomic sample, reflecting the base representation of the sample. In practice, later cycles often show some bases more than others. As the signal decays, some bases may start to fall into the noise while other still rise above it. Matrix adjustments may help to optimize data.
- ▶ **%All** and **%Called**—Exactly the same as All and Called, but expressed as a percentage of the total intensities. These plots make it easier to see changes in relative intensities between channels as a function of cycle by removing any intensity decay.

For information on interpreting results in the IVC.htm file, see *Interpretation of Run Quality* on page 54.

All.htm

This file gives a tile-by-tile representation of the mean matrix-adjusted intensity of clusters plotted as a function of cycle. It plots each channel (A, C, G, T) separately as a different colored line. Means are calculated over all clusters, regardless of base calling.

If all clusters are T, channels A, C, and G will be at zero. If all bases are present in the sample at a rate of 25% and a well-balanced matrix is used for analysis, the graph will display all channels with similar intensities. If intensities are not similar, the results could indicate either poor cross-talk correction or poor absolute intensity balance among each channel.

A genome rich in GC content may not provide a balanced matrix for accurate cross-talk correction and absolute intensity balance.



NOTE

For large experiments (> 200 tiles per lane), All.htm, Perfect.htm, and Error.htm only show a subset of tiles. However, each file contains links to the full output results. For example, Error.htm links to FullError.htm. The full output files may take some time to open.

For information on interpreting results in the All.htm file, see *Interpretation of Run Quality* on page 54.

Error Rates

For all analysis modes except sequence, Perfect.htm and Error.htm are produced, which measure sequence error rates.

Perfect.htm

This graph shows the proportion of reads in a tile that have 0, 1, 2, 3, or 4 errors by the time they get to a given cycle.

Good data show a high proportion of reads with zero errors throughout the cycles.

Error.htm

This file shows a graph of error rates for each tile on a flow cell. The red bar shows the percentage of bases at each cycle that are wrong, as calculated based on alignment to the reference sequence. Issues such as focus or fluidics problems manifest themselves as spikes in the graph.

Good data is 1–1.5% or less for 25 aligned bases.

- ▶ PhageAlign allows any number of errors in an alignment and provides an accurate count of the error rate. However, it is too slow for aligning against target references larger than 2 MB.
- ▶ ELAND is capable of aligning against large genomes, such as human, in reasonable time. However, it allows only two errors per fragment. This means that error rates based on ELAND alignments are underestimated. Very poor quality data has more than two errors in the first 32 aligned bases and is excluded from the calculations.

Text-Based Analysis Results

The output files for each lane of a flow cell are named using the format `s_N_sequence.txt`, where N represents a specific lane of the flow cell. For paired-read analysis, there are two parallel output files, one for each read. The files are named using the format `s_N_R_sequence.txt`, where N represents a specific lane of the flow cell and R represents the read number. The files are found in the GERALD folder of a finished analysis run.

The output files for each tile are named using the format `s_N_TTTT_realign.txt`. For example, all files pertaining to tile 23 of lane 3 have names starting with `s_3_0023`.

The following table lists the files that contain the most meaningful data produced from your analysis run and the GERALD analysis mode that creates them. For descriptions of the GERALD analysis variables, see *ANALYSIS Variables* on page 29.

Table 12 Text-Based Analysis Results

GERALD Analysis Mode	Output File	Description
All modes except ANALYSIS none	<code>s_N_sequence.txt</code>	This file contains all sequences in a single lane of a flow cell in an exportable format. The content of this file is affected by the following parameters: <code>USE_BASES</code> , <code>QF_PARAMS</code> , <code>SEQUENCE_FORMAT</code> , <code>QUALITY_FORMAT</code> . For a description of each of these parameters, see <i>ANALYSIS Variables</i> on page 29.
ANALYSIS default ANALYSIS eland	<code>s_N_realign.txt</code>	This file contains filtered alignment information.
	<code>s_N_rescore.txt</code>	This file contains error rates for filtered data based on the alignments in the <code>rescore.txt</code> files. These are used to create the graphs in the <code>Error.htm</code> pages. Looking at the graphs in <code>Error.htm</code> will probably tell you what you need to know.
	<code>s_N_qreport.txt</code>	This file reports the accuracy of the base calling quality values, making use of the <code>_qraw.txt</code> files.
	<code>s_N_qcalreport.txt</code>	This file reports the accuracy of the recalibrated quality values, making use of the <code>_qcal.txt</code> files if they are present for the type of analysis you have specified. The format is identical to <code>s_N_qreport.txt</code> .
ANALYSIS eland_extended	<code>s_N_export.txt</code>	This file contains the results of alignment of all reads in the lane. The fields are tab separated to facilitate export to databases. This file has a line for every read, not just those that pass purity filtering. The last field on each line is a flag telling you whether or not the read passed the filter (1 or 0). For file formats, see <i>Output File Formats</i> on page 80.
	<code>s_N_sorted.txt</code>	This output file is similar to <code>s_N_export.txt</code> , except it contains only entries for reads which pass purity filtering and have a unique alignment in the reference. These are sorted by order of their alignment position, which is meant to facilitate the extraction of ranges of reads for purposes of visualization or SNP calling.

Table 12 Text-Based Analysis Results (Continued)

GERALD Analysis Mode	Output File	Description
ANALYSIS eland_pair	s_N_1_sequence.txt, s_N_2_sequence.txt	These parallel sets of files contain filtered sequences for each lane.
	s_N_1_export.txt s_N_2_export.txt	These parallel sets of files contain the results of alignment of all reads in the lane. The fields are tab separated to facilitate export to databases. Each file has a line for every read, not just those that pass purity filtering. The last field on each line is a flag telling you whether or not the read passed the filter (1 or 0). For information on file format, see <i>Output File Formats</i> on page 80.
	s_N_1_sorted.txt s_N_2_sorted.txt	These parallel sets of files are similar to s_N_1_export.txt and sN_2_export.txt, except they contain only entries for reads which pass purity filtering and have a unique alignment in the reference. These are sorted by order of their alignment position, which is meant to facilitate the extraction of ranges of reads for purposes of visualization or SNP calling.
	s_N_anomaly.txt	This file contains one line for each read for which the two halves of the read did not align with a nominal distance and orientation from each other. This is the file to mine for structural variation information.

For descriptions of file formats, see *Output File Formats* on page 80.

Numerous intermediate files are produced during an analysis run. For a description of these files, see *Intermediate Output Data Files* on page 77.

Interpretation of Run Quality

After the analysis of a run is complete, you will need to interpret the data in the report summary and various graphical outputs. This section describes a standard, systematic way to examine your data.

The starting point is to know what a standard run of acceptable quality looks like. This is something of a moving target and is dependent on individual instruments, instrument configuration, genomic sample type, type of analysis, chip preparation, and the current state of the art. Therefore, the numbers shown in this section are for example only.

Summary.htm

The Summary.htm file is the first file you should review after your analysis is complete.

The following are examples of two of the tables found in Summary.htm, Lane Results Summary and Expanded Lane Summary, each truncated to a single lane of information. For a brief description of the tables found in Summary.htm, see *Summary.htm* on page 48.

Table 13 Example of Lane Results Summary

Lane	Clusters	Av 1st Cycle Int	% intensity after 20 cycles	% PF Clusters	% Align (PF)	AV Alignment Score (PF)	% Error Rate (PF)
1	23621 +/- 407	1926 +/- 60	65.12 +/- 2.48	52.55 +/- 0.37	98.33 +/- 0.14	2855.55 +/- 90.70	6.71 +/- 0.63

Table 14 Example of Expanded Lane Summary

Lane Info		Phasing Info		Raw Data		Filtered Data						
Lane	Clusters	% Phasing	% Pre-phasing	% Error Rate (Raw)	Equiv Perfect Clusters (raw)	% re-tained	Cycle 2-4 Av Int	Cycle 2-10 Av % Loss	Cycle 10-20 Av % Loss	% Align (PF)	% Error Rate (PF)	Equiv Perfect Clusters (PF)
1	23621	0.9300	0.5800	11.17	12457	52.55	1728 +/- 43	2.31 +/- 0.24	181 +/- 0.15	98.33	6.71	9709

The key parameters that you should examine are listed in the following sections, along with conditions, possible causes for those conditions, and suggested actions to take to correct the condition.

Clusters

This column contains the average number of clusters per tile detected in the first cycle images. For 1 Gbases of data at 35 cycles, this value needs to be greater than 20,000.

Condition	Possible Cause	Suggested Action
Fewer clusters than expected:		Reanalyze with new default offsets. If the problem persists, ensure that the alignment config file contains "SIMILARITY" filtering. The use of "SIMILARITY" filtering will result in low numbers passing filters.
Few bright clusters on the flow cell	Problem with cluster formation	
Blurred images	Poor focus or dirty flow cell surface	
Lots of clusters visible	Cluster density or size is too great to distinguish individual objects	
More clusters than expected:		
Too many clusters on the flow cell	Problem with cluster formation	
Very large clusters	Double counting	

Average First Cycle Intensity

Generally, brighter is better, but this result is instrument and sample dependent. Ideally, this value should be greater than 1000. For some paired-end sample preparations, this value should be greater than 500.

Condition	Possible Cause
Low intensity	Problem with cluster formation or poor focus

Percentage of First Cycle Intensity Remaining After 20 Cycles of Sequencing

Generally, the higher, the better. Greater than 50% is acceptable, though it can be sample dependent.

Condition	Possible Cause	Suggested Action
Low value	A correct measure of rapid signal decay deduced from intensity plots	Check experiment fluidics or temperature control
	Problem with cycle 20 deduced from intensity plots.	Check fluidics and focus for this cycle
Exceptionally high value	Low first cycle intensity	Check first cycle focus

Percentage of Clusters Passing Filters

To remove the least reliable data from the analysis, the raw data can be filtered to remove any clusters that have “too much” intensity corresponding to bases other than the called base. By default, the purity of the signal from each cluster is examined over the first 12 cycles and $CHASTITY = \frac{\text{Highest_Intensity}}{\text{Highest_Intensity} + \text{Next_Highest_Intensity}}$ is calculated for each cycle. If $CHASTITY > 0.6$ for all 12 cycles, then the cluster passes the filters. Both $CHASTITY > 0.6$ and 12 cycles are essentially arbitrary, and are a compromise arrived at to remove most of the error prone data without throwing away too much of the good data.

The higher the value, the better. Ideally, a value of $> 70\%$ good, but this value is very dependent on cluster density. When there is above 20,000 clusters per tile, the percentage starts fall, since the major cause of an impure signal in the early cycles is the presence of another cluster within a few micrometers.

Condition	Possible Cause	Suggested Action
Very few clusters passing filter	<ul style="list-style-type: none"> Poor flow cell, perhaps unblocked DNA Faint clusters Out of focus Poor matrix A fluidics or sequencing failure Bubbles in individual tiles Too many clusters Large clusters High phasing or prephasing 	<p>Some of the causes may be at a single cycle. If the problem is isolated to these early cycles, it is possible that this filtering throws away very good data.</p> <p>Base calling errors may be limited to effected cycles, and as early cycles are fairly resistant to minor focus and fluidics problems, even the number of errors may be few. The filtering can always be set manually to some other values. Check before assuming all the data are poor.</p>

Percentage of Clusters Passing Filters that Align Uniquely to the Reference Genome

Optimal value depends on the genome sequenced and the read-length; the higher (up to 100% max), the better. For example, for 30-mers and human genome, the optimum is less than 80%.

This result is genome specific and dependent on the completeness of the reference. A failure to align could be due to repeat or missing regions, or due to indels where sample and reference do not match.

Condition	Possible Cause	Suggested Action
Much lower than expected when using ELAND	Fluidics or instrument problem	Look for an intensity dip in IVC plots. If there is a problem and it occurs after a sufficiently useful read-length, re-run ELAND analysis using only the “good” cycles before the instrument problem.
	Contamination from other genetic material resulting in an inability to align data	Align a few sample tiles with PhageAlign. Genomic contamination will show as early cycle error rates. If error rates remain fairly constant with cycle, then the “correct” genome has probably sequenced correctly. Non smooth error rate plots or IVC plots indicate the presence of specific tags or sequences.

Percentage Error Rate of Clusters Passing Filters

This value should be as low as possible, but it is very dependent on read-length. At 32 cycles, the error rate should be around 2%. Depending on the quality of the data, it will tend to rise at about this point. If there is a sudden rise beyond cycle 32, then it is likely that ELAND has effectively filtered out many clusters with more than two errors, thus suppressing the true error rate up to this point. The percentage aligning will also be low.

With PhageAlign analysis of control samples, error rates for 25 cycles should be < 1.5%.

Percentage of Phasing and Prephasing

Ideally, these values should be as low as possible. Satisfactory results can be obtained with up to 1% for each. Preferably, they should be closer to 0.5%

Condition	Possible Cause	Suggested Action
High phasing or prephasing	Reagent issue (reagents have deteriorated) Fluidics	Check for leaks or bubbles in images or early cycle discrepancies in intensity plots.
	Poor flow cell	Poor blocking can be evident as intensity in all channels from cycle 1.

Standard Deviations

Many values have standard deviations associated with them. This can be the first indication as to the uniformity of the flow cell. If standard deviations are high, then it indicates variability from tile to tile with a lane.

Condition	Possible Cause	Suggested Action
High standard deviations	Check poor tiles for: <ul style="list-style-type: none"> • Bubbles • Focus • Dirty flow cell surface 	Look at the tile-by-tile statistics that appear below the flow cell-wide summary.

After reviewing the tables in Summary.htm, examine the thumbnails, and the output files IVC.htm, All.htm, and Error.htm.

IVC.htm

For a detailed description of the plots found in the IVC.htm file, see *IVC.htm* on page 58.

Condition	Possible Cause
Intensity curves are not smooth	Cycle to cycle focus or fluidics problems
Called intensities are not equal ("% Called" may be +/- 5% out without major problems)	Poor fluidics or poorly blocked flow cell If from cycle 1, initial matrix estimate may also be in error

**All.htm and
Error.htm**

The results in both files should show consistency from tile to tile down a lane and from lane to lane, if the results are from the same sample.

Condition	Possible Cause
Tile variability	Bubbles Rapid focus fluctuations Dirty flow cell surface
Rising error rates (Rates will always rise eventually at high read-lengths)	Low intensity at start High decay rate High phasing or prephasing
High, but constant error rates from cycle 1	Genomic contamination



Chapter 6

Advanced Pipeline Usage

Topics

- 60 Introduction
- 60 Running Bustard as a Standalone Program
 - 60 Assigning a Control Lane
- 61 Running GERALD as a Standalone Program
 - 61 Additional "Make" Options
- 62 Running ELAND as a Standalone Program
 - 62 Compiling ELAND
 - 62 Command Line Syntax

Introduction

Bustard, GERALD, and ELAND may be run as standalone programs. This allows you to rerun your analysis using different parameter settings without running the rest of the Pipeline. You can rerun the base caller on a different subset of intensity files, perform alignments on the same base-called sequences, or rerun sequences against another genome.

Running Bustard as a Standalone Program

You can invoke the base calling script repeatedly without rerunning the image analysis. This lets you run the base caller on a different subset of cycles, tiles, and lanes with different parameter settings. The run is set up by a separate script called “bustard.py.”

The following example shows the various options you can specify with the base calling script:

```
/path/Pipeline/Goat/bustard.py
  [--cycles=1-25|auto] [--tiles=s_1,s_2_0003,...]
  [--matrix=mymatrix.txt|auto|auto<n>]
  [--phasing=0.01|auto|auto<n>] [--prephasing=0.01]
  [--make]
  [--GERALD=/path/config.txt] [--control-lane=<lane>]
  <Firecrest directory>
```

For example, the following command calls the base calling script and points to the image analysis directory:

```
/path/Pipeline/Goat/bustard.py
  /data/070813_ILMN-1_0217_FC1234/Data/C1-
  27_Firecrest1.9.0_23-08-2007-user
```

This will not generate any Makefiles and directories unless the “make” option has been specified.

Assigning a Control Lane

If you need to assign a control lane for more accurate matrix and phasing estimation, run base calling using the “bustard.py” script and use control-lane as an argument.

```
Pipeline/Goat/bustard.py --control-lane=4 --make /
  data/070813_ILMN-1_0217_FC1234/Data/C1-
  26_Firecrest*
```

Change to the newly generated Bustard folder and type the “make all” command.

```
make all
```

Running GERALD as a Standalone Program

You can run an analysis using GERALD without the rest of the Pipeline if you want to perform alignments with different parameters on the same base-called sequences.

GERALD uses a text-based configuration file containing all parameters required for alignment, visualization, and filtering. These parameters are the type of analysis to perform, which bases to used for alignment, and the reference files for a sequence alignment. The GERALD.pl script is used to generate the GERALD Makefile. The Makefile is executed using the “make” utility.

A typical invocation would be as follows:

```
Pipeline/Gerald/GERALD.pl gerald_config.txt
--EXPT_DIR path_to_bustard_folder --FORCE
```

The standard way to run GERALD is to set the parameters in a configuration file, create a Makefile, and start the analysis with the “make” command.

1. Edit the config.txt file as described in *GERALD Configuration File* on page 34.
2. Enter the following command to create a Makefile for sequence alignment. To generate a Makefile in GERALD, use FORCE instead of “make.”

```
GERALD.pl config.txt --FORCE
```

3. Change to the newly created GERALD folder under the “path_to_bustard_folder.” Type the “make” command for basic analysis.

You may prefer to use the parallelization option as follows:

```
make -j 3 all
```

The extent of the parallelization depends on the setup of your computer or computing cluster. For a description of parallelization, see *Using Parallelization* on page 87.

For more information on GERALD, see *Using GERALD* on page 27.

Additional “Make” Options

You may perform a partial build of your analysis. This feature may be useful for a sneak preview of your results, after which a full analysis may be built as described above. For example, to build all files for tile 12 of lane 3, use the following “make” option:

```
make TILE=s_3_0012
```

You may specify specific tiles to perform a partial build of your analysis. The following example will build tiles 0005, 0010, 0015, 0020, and 0025 from lanes 3 and 6:

```
make TILE=s_[36]_00[0-2][05]
```

This example specifies any tile for which the last digit is 0 or 5, the previous digit is 0, 1, or 2, and the previous two positions are 00.

Running ELAND as a Standalone Program

You can run ELAND without the rest of the Pipeline as a post-analysis step. Sequences that did not produce ELAND matches in the initial run, can be rerun against another genome.

You can run ELAND as a standalone program to align a file of up to 16 million fasta-formatted sequences of up to 32 bases in length against a squashed genome. However, running ELAND does not perform all of the various steps that are included during a GERALD run. For example:

- ▶ Quality value recalibration
- ▶ Extension of alignments beyond 32 bases
- ▶ Removal of sequences that fail signal purity filtering

If you require any or all of the above, it is best to create a modified config file to align to a different squashed genome, and rerun GERALD. For more information, see *GERALD Configuration File* on page 34.

Compiling ELAND

ELAND is compiled automatically as part of the Pipeline installation as described in *Installation Prerequisites* on page 69.

You can manually compile ELAND from the Pipeline/Eland directory using the “make” command. This compiles ELAND without compiling the rest of the Pipeline.

```
make -e eland
```

Command Line Syntax

Use the following command line syntax to run ELAND as a standalone program:

```
eland_executable queryFile squashedGenomeDir
[output_file].txt
[--multi[=N0[,N1,N2]] [repeatFile]
```

queryFile.txt

queryFile.txt is a file of query sequences. This must be either a multi-entry fasta format file or a one-sequence-per-line ASCII file. The length of each sequence must exceed the read length specified at compilation. Unspecified bases in the reads must be denoted by an “N.” IUPAC ambiguity codes are not handled.

squashedGenomeDir

squashedGenomeDir is the path to the directory of squashed genome files. For more information, see *Preparing the Reference Genome* on page 37.

[output_file].txt

The ELAND output file contains the initial output of the ELAND alignment program with one line of output per sequence. The name of the output file depends on the analysis you are running.

- ▶ ANALYSIS eland produces an output file named s_N_eland_results.txt.
- ▶ ANALYSIS eland_extended and ANALYSIS eland_pair are run with the --multi option, and produce an output file named s_N_eland_multi.txt.

For an explanation of intermediate output files, see *Intermediate Output Data Files* on page 77.

For a description of the output file format, see Table 19 on page 81.

--multi

If `--multi` is specified, ELAND will store and display multiple (10 by default) matches for each read.

- ▶ `--multi=20,40,80` will display at most 20 exact matches, 40 single-error matches, and 80 2-error matches.
- ▶ `--multi=20` will display at most 20 matches of any number of errors.

For a description of output file formats using the multi option, see Table 19 on page 81.

repeatFile.txt

You may want to specify a set of words that you know are highly repetitive in your target files at your read length of interest. You can then tell ELAND to ignore them, which greatly increases the speed of whole-human-genome alignments. There is no automatic way of generating a repeat file, but with a bit of Perl/shell scripting, it is straightforward to extract a list of repeats from the output of a few ELAND runs to improve the speed of future runs.

You can run the basic test harness script, "ELAND_test.pl" from the ELAND directory to verify correct operation.



Appendix A

System Requirements and Software Installation

Topics

- 66 Introduction
- 66 System Requirements
 - 66 Network Infrastructure
 - 67 Analysis Computer
- 69 Installation Prerequisites
 - 69 Setting Up Email Reporting
- 71 Installing the Pipeline Software
 - 71 Compiling on Other Platforms
 - 71 Directory Setup

Introduction

This section describes the Pipeline system requirements and the software installation instructions. It also describes how to set up your instrument directory.

System Requirements

Images are acquired and stored on the Genome Analyzer. They must be transferred to an external computer to be analyzed by the analysis software, which handles image processing, base calling, and sequence alignment. Based on an eight-lane flow cell with three columns and 110 rows per lane, each sequencing run generates approximately 1 TB of data during a full 2–3 day run. Paired-end runs generate approximately 2 TB of data over a 5–6 day run. However, about 70% of this is TIFF image data that can potentially be stored on tape after an analysis run is complete.

Depending on the application, single experiments run from 18–50 cycles. Paired-end experiments can double the number of cycles while gene expression experiments may use only 18-cycle protocols. Estimating required storage for individual runs depends on your application. The following table summarizes data volumes per experiment.

Table 15 Data Volumes Per Experiment

Cycles per Run	Run Time (hours)	Raw Data (TB)	Results Data (TB)
18	42.0	0.360	0.270
26	60.7	0.520	0.390
36	84.0	0.720	0.540
50	116.7	1.000	0.750
75	175.0	1.500	1.125
100	233.3	2.000	1.500

Network Infrastructure

These large data volumes mean that you will need:

1. A high-throughput ethernet connection (1 GB or more recommended) or other data transfer mechanism.
2. A suitably large holding area for the images and analysis output (1 TB per run). As there will almost certainly some overlap between copying, analysis, possible reanalysis, 2–3 TB is an absolute minimum.
3. You need to consider which parts of the data you want to backup and what infrastructure you want to provide for the backup. If you want to keep image data, then half a terabyte per run is required. The Pipeline provides the option to perform loss-less data compression.

Storage Configurations

You can configure your analysis server as either local storage or external network storage.

- ▶ Local server storage can be internal to the server, or Direct Attached Storage (DAS), which is a separate chassis attached to the server.
 - **Internal**—Simple but not scalable. Results data must be moved off to network storage at some point to make room for subsequent runs.
 - **DAS**—External chassis that is scalable since more than one DAS can be connected to the server. The server is an application server running the Pipeline and a file server providing access to results and receiving incoming raw data files.
- ▶ External network storage is either Network Attached Storage (NAS) or Storage Area Network (SAN). NAS and SAN are functionally equivalent, but SAN is larger, with higher performance, more connections, and more management options.
 - **NAS**—External chassis connected via an Ethernet to the server, instrument PC, and other clients on the network. NAS devices are scalable and highly optimized.
 - **SAN**—The most scalable with the highest performance. They have a very high bandwidth and support many simultaneous clients, but are complex to manage and significantly more expensive.

Server Configurations

You can use either a single multi-processor, multi-core computer running Linux, or a cluster of Linux servers with a head node. The Pipeline can take advantage of clustered and multi-processing servers.

- ▶ **Single multi-processor, multi-core server**—Simple but not scalable. It can only analyze data from one Genome Analyzer, or two depending on power and your turn-around requirements.
- ▶ **Linux Cluster**—Highly scalable and capable of running multiple jobs simultaneously. It requires one server as a management node and a minimum number of computational nodes to be as efficient as a standalone server. By adding computational nodes, the cluster can service more instruments.

Analysis Computer

The Pipeline may run on any Unix variant, if all of the prerequisites described in this section are met. However, Illumina does not support any platform other than Linux.

Illumina recommends and fully supports the following hardware configuration.

- ▶ High performance DL580 G4 server from Hewlett Packard
This system comes configured with Red Hat Linux and the full installation of the Genome Analyzer Pipeline Software.
- ▶ Single 4-way, dual-core server with Xeon 7140 class processors
- ▶ 32 GB fault-tolerant RAM
This is enough RAM to perform analysis tasks and file server tasks simultaneously. It uses high speed fault-tolerant hard drives for the operating system and applications.

▶ HP MSA20 Direct Attached Storage (DAS) unit

This capacity is intended to hold information from three runs, as follows:

- Last Processed Run—The results data from the last analyzed run are copied off to another storage server, where the run can be reviewed by the investigators and their staff. The raw image data is deleted.
- Currently Processed Run—The raw image data from the last completed instrument run are loaded and the Pipeline is performing analysis on that run.
- Next Run for Processing—The Genome Analyzer is copying the raw data from the current run up to the server.

As data volumes increase, the storage capacity can be scaled up by adding additional MSA20 DAS units.

On this type of hardware, you can expect to perform the image analysis and base calling for a full run in approximately one day. Sequence alignment takes additional time depending on which alignment program you are running; somewhere between a few hours (using our fast short-read whole-genome alignment program ELAND) and days (using more traditional alignment programs).

Pipeline parallelization is built around the multi-processor facilities of the “make” utility and scales very well to beyond eight nodes. Substantial speed increases are expected for parallelization across several hundred CPUs. For a detailed description, see *Using Parallelization* on page 87.

Installation Prerequisites

The following software is required to run the Genome Analyzer Pipeline Software:

- ▶ Perl 5.8 or later; install the XML::Simple module and its dependencies (<http://www.cpan.org>)
- ▶ Python 2.3 or later
- ▶ GNU make 3.78 or later (qmake from Sun Grid Engine (SGE) 6.0 has been reported to work)
- ▶ gnuplot 3.7 or later (4.0 is recommended)
- ▶ ImageMagick 5.4.7 or later
- ▶ Ghostscript
- ▶ SMTP server (for optional automated email run reports)
- ▶ zlib
- ▶ bzlib

For a compilation from source, the following additional software is required:

- ▶ gcc (including g++)
- ▶ Optimized FFT library (Only one of the following three FFT libraries are required, not all three)
 - FFTW 3.0.1 or greater (3.1 is recommended); GPLed. To download files, see <http://www.fftw.org>.
The single-precision version of FFTW is required (libfftw3f.a). This is produced by specifying the `--enable-single` option to the `./configure` procedure of FFTW as follows:

```
./configure --enable-single
make
make install
```
 - Intel Maths Kernel Library
 - IBM ESSL

If you are running the Linux distribution Red Hat, the required dependencies listed above are satisfied by the Red Hat packages `perl-*`, `python-*`, `make`, `autoconf`, `gnuplot`, `ImageMagick`, `ghostscript`, `zlib`, `zlib-devel`, `bzip2`, `bzip2-devel`, `libtiff-devel` and `gcc-*` as well as their respective prerequisites. The Perl XML::Simple module and `fftw3` need to be downloaded separately and installed from source.

Setting Up Email Reporting

The script `Gerald/runReport.pl` is called at the end of a run and sends you an email when a run successfully completes.

To use email notification, set up an SMTP server and set the following parameters in the GERALD configuration file. For additional information, see *GERALD Configuration File* on page 34.

1. Enter a space-separated list of the email addresses that should receive the run completion notification.

```
EMAIL_LIST your.name@domain.com that.name@domain.com
```

2. Indicate the path to the GERALD folder. The software assumes it can create a valid URL from the GERALD folder path by omitting the first two path elements and prepending WEB_DIR_ROOT.

```
WEB_DIR_ROOT http://server/SHARE
```

For example, if the path is /mnt/yourDrive/folder/folder/GERALD and WEB_DIR_ROOT is http://server/SHARE, the software will write the links as http://server/SHARE/folder/folder/GERALD/File.htm.

3. Identify your domain. Your SMTP server may refuse to accept emails from or send emails to addresses that do not end in @yourdomain.com.

```
EMAIL_DOMAIN yourdomain.com
```

4. Identify your IP address.

```
EMAIL_SERVER yourserver:2525
```

where yourserver is the name or IP address of a mail server that will accept SMTP email requests from you and 2525 is the port number of the SMTP service on that server.

Generally this will be 25. This is the default value if no port number is specified. The utility nmap, if installed, may help you identify which port on a server is hosting an SMTP service.

5. Test your email reporting by entering the following from the machine where you are running GERALD:

```
telnet yourserver yourPortNumber
```

If you don't get a friendly message, then email reporting will not work.

You can run runReport.pl directly in test mode by entering:

```
/runReport.pl --test yourserver:25 yourdomain.com  
anything your.name@yourdomain.com
```

You should receive a test email. If you do not, the transcript it generates should identify the problem.



The optional email reporting feature depends on how your SMTP servers are set up locally. Email reporting is not required to run the Pipeline to a successful completion.

Installing the Pipeline Software

To install the Pipeline, you obtain the source code and then compile the software. Compiling the software will first build all C++ code, and then copy the relevant executables into the directories GOAT and GERALD which contain the scripts and Makefile generators.



If you want to use the Intel Math Kernel Library as an FFT backend, compile the image analysis module Firecrest separately from the rest of the project. Specify the additional variable MKL to make, as in "make MKL=true" and set the MKL-specific paths in the Makefile to the appropriate locations on your system.

1. Go to the location where you want to install the Pipeline and type the following:

```
tar xvfz GAPipeline-version.tar.gz
```

where version is of the archive you have. You may have to adjust the path to the archive.

2. Change to the Pipeline directory and type:

```
make  
make install
```

Compiling on Other Platforms

Compiling the Pipeline with the current Makefiles works on all platforms, including many 32-bit and 64-bit Linux versions and Solaris. However, if your compilation does not succeed on a less commonly used platform (possibly 64-bit architectures or platforms other than Linux), you may have to make manual changes to the Makefiles. Compilation problems, may require you to adapt the platform-specific gcc-compiler flags. Because of the optimized FFT libraries, the Firecrest Makefile is particularly likely to be sensitive to platform-specific peculiarities.

Directory Setup

Create a directory called Instruments/<instrument_name> for each Genome Analyzer in the same directory as the Run Folder, where <instrument_name> is the hostname of the computer that is attached to the Genome Analyzer.

For example, the directory for the Run Folder /data/070813_ILMN-1_0217_FC1234 would be called /data/Instruments/ILMN-1/.

If this directory exists, the Pipeline will place a file called default_offsets.txt into this directory. The Pipeline automatically keeps this file up-to-date. For information on default_offsets.txt, see *Image Offsets* on page 15.

Use the environment variable INSTRUMENT_DIR, to override the default location of the Instruments directory:

```
export INSTRUMENT_DIR=/home/user/Instruments
```

If no instrument directory exists, the Pipeline will create one for you. If no default_offsets.txt file exists, the Pipeline will create one with offset values equal to zero.



Appendix B

Output File Descriptions

Topics

- 74 Introduction
- 74 Output File Types
 - 75 Intensity Files
 - 75 Sequence Files
 - 76 Quality Score Files
 - 76 Efficiency
- 77 Intermediate Output Data Files
- 80 Output File Formats
- 83 Parameters File Format

Introduction

This section describes the file types and file formats of the intermediate data output produced during an analysis run.

Output File Types

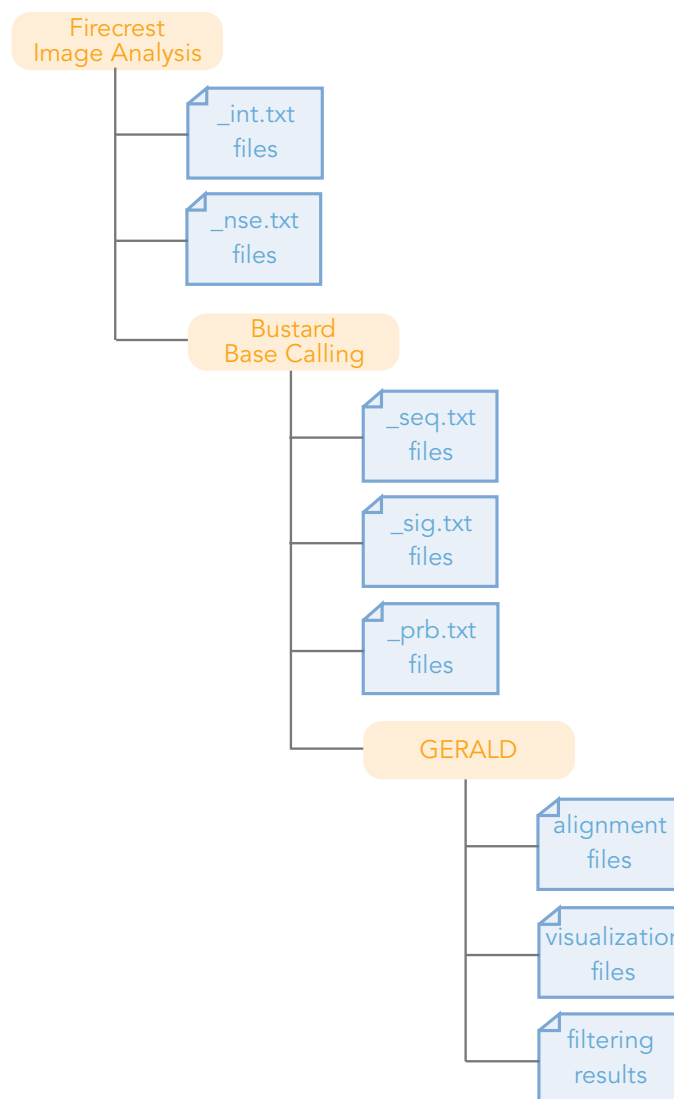


Figure 5 Run Folder Structure and Output File Types

Intensity Files

The prefix of the intensity filenames follows the prefix of the image filenames, but the tile position is padded to four digits. For example, s_1_0006_int.txt is the intensity file corresponding to the image files s_1_6_a.tif, s_1_6_c.tif, and so on.

Each intensity file has a set of data for remapped clusters on each line. Each row corresponds to the data from one cluster and each column is delimited by a space. Each row has a channel index, and a tile index in the first column, with the X offset and Y offset of the cluster in the second and third columns (all coordinates indexed from zero). These fields are tab-delimited. These values should be enough to uniquely identify any cluster for any run.

Following the coordinate fields are the data fields. The first value in a data field is the raw intensity for base A, the second is the raw intensity for base C, then G, and then T. These four values are separated by spaces and are followed by a tab to mark the beginning of the next four values. The next four values represent the corresponding intensities in the de-block scan (if a de-block scan has been performed), and then four intensities for the next cycle.

The number of fields should equal four coordinates plus four bases times the number of cycles of processed data, even if clusters don't yield data at the end of a set of cycles or part way through. In this case, the fields contain 0.0 and preserve delimiters.

The following is a sample line from an intensity file (_int.txt):

```
<Channel><TAB><Tile><TAB><X><TAB><Y><TAB><A int 1st
      cycle> <C int> <G int> <T int><TAB>\
<A int 2nd cycle>... <LF>
```

A second set of files with an identical layout, stores the estimates of the noise on the intensity estimates. These files end in _nse.txt.

Sequence Files

The data found in the sequence files (_seq.txt) located in the Bustard folder are raw sequences in the following condition:

- ▶ Trimming of any primer bases and splitting of a paired-read into two reads as specified by USE_BASES has not been applied.
- ▶ Signal purity filtering of low quality data has not been applied.
- ▶ There is one file per tile, resulting in over 1000 files in total.

Use the sequence.txt files in the GERALD folder for which all the above points have been applied.

The base calls are kept in one file per tile for the concomitant base calls, and use the extension _seq.txt. For a given intensity file, following base calling, we have a sequence file of the same name. For example, from an intensity file called s_1_0001_int.txt you would get a base-called file named s_1_0001_seq.txt.

Each sequence file has a sequence per row similar to the intensity files. Each row uses the same format as the intensity file, with the <lane>,<tile>,<X-offset>,<Y-offset> providing a unique key and a global co-ordinate for the sequence, and relating sequences to a cluster on the images. Following this format, the output is a string with one character for each base call in tab-delimited fields.

Another file holds the base caller confidence score that follows the format:

```
<channel><TAB><tile><TAB><X><TAB><Y><TAB><sequence><LF>
```

Quality Score Files

Each intensity and sequence file has an associated file containing the quality scores for the base calls. For example, the quality file for the sequence file s_1_0002_seq.txt is called s_1_0002_prb.txt. There are four quality scores corresponding to the four bases in the order of A, C, G, T, for each base call in the sequence file. These fields are separated by a space. Each set of four scores for one base call is separated from the next set by a tab.

For following example might be the scores for a sequence AGT:

```
30 -30 -30 -30<TAB>-22 20 -27 -30<TAB>-17 17 -30 -30
```

The scores are defined as $Q = -10 \cdot \log_{10}(p/(1-p))$, where p is the probability of a base call corresponding to the base in question. By definition, the four values of p add up to 1.

Quality scores are essential for almost any genetic application to be able to tell good bases from bad bases. For example, ELAND can use quality scores to break degenerate alignments, and quality scores are essential for calling SNPs.

Efficiency

To allow efficient handling by any software packages, there is one intensity and sequence file per tile. However, a single file can easily be created by simple concatenation of the individual files.

Intermediate Output Data Files

Intermediate output files are found in the GERALD folder and contain data used to build the more meaningful results files described in *Analysis Output* on page 47.

The files are named using one of the following formats:

- ▶ s_N_TTTT_name.txt, where N is the lane number, T is the tile number
- ▶ s_N_name.txt, where N is the lane number
- ▶ s_N_R_name.txt, where N is the lane number, R is the read number

Table 16 Intermediate Output File Descriptions

Output File	GERALD Analysis Mode	Description
s_N_TTTT_align.txt	ANALYSIS default ANALYSIS eland ANALYSIS monotemplate ANALYSIS eland_tag	Contains unfiltered first-pass alignments for a given tile.
s_N_TTTT_score.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	Contains error rate information from first pass alignments. Error rate information is contained in text form and indicates potential contaminants. If CONTAM_FILE is specified, sequences with a negative entry in the s_N_cdifff.txt file, such as likely contaminants, are ignored.
s_N_TTTT_prealign.txt	ANALYSIS default ANALYSIS eland	Contains the realignment of all sequences against the data, using the error rate information in s_N_TTTT_score.txt to refine the alignment by re-weighting each base at each cycle according to its confidence. If the lane in question were analyzed using ELAND, this file is just a copy of s_N_align.txt, because ELAND does not have the feature to weight the contribution of bases in an alignment.
s_N_TTTT_realign.txt	ANALYSIS default ANALYSIS eland	Consists of alignments in s_N_TTTT_prealign.txt, filtered to exclude alignments for those clusters that do not pass the quality criterion QF_PARAMS when applied to s_N_TTTT_qhg.txt. Even if contaminant filtering is switched on, the alignments here will not have been contaminant filtered. Use the s_N_TTTT_crediff.txt file and "qualityFilter.pl" script to retain non-contaminants only. <pre>cat s_N_TTTT_realign.txt qualityFilter.pl '(\$F[0]>0)' s_N_TTTT_crediff.txt</pre> Replace ">" with "<=" to retain contaminants only.
s_N_TTTT_rescore.txt	ANALYSIS default ANALYSIS eland	Contains the improved estimate of the error rate based on s_N_TTTT_realign.txt. If CONTAM_FILE is specified, the calculation ignores sequences with a negative entry in the s_N_TTTT_crediff.txt file, such as likely contaminants.
s_N_TTTT_rescore.png	ANALYSIS default ANALYSIS eland	This generated image is a viewable error rate graph drawn from the data in s_N_TTTT_rescore.txt. This image is used as a thumbnail in Error.htm as described in <i>Visual Analysis Summary</i> on page 48.

Table 16 Intermediate Output File Descriptions (Continued)

Output File	GERALD Analysis Mode	Description
s_N_TTTT_qalign.txt	ANALYSIS default ANALYSIS monotemplate	Contains the alignments using base quality values to weight the bases. This file is not produced if the alignments for the lane in question were generated from an ELAND analysis, as ELAND does not have the feature to weight bases by their quality values.
s_N_TTTT_qraw.txt	ANALYSIS default ANALYSIS eland ANALYSIS monotemplate	This file collates the scores found in the file s_N_prb.txt, which contains four scores for each base. The highest of these scores is the score pertaining to the called base. If ANALYSIS --symbolic is specified (default), the quality scores are encoded as ASCII characters.
s_N_qraw.txt	ANALYSIS eland_extended	If ANALYSIS --numeric is specified, these are encoded as space separated integers.
s_N_R_qraw.txt	ANALYSIS eland_pair	In both cases the file will only contain values for cycles that the Pipeline has been asked to include, such as those with a "Y" in the corresponding USE_BASES string. For detailed descriptions of USE_BASES, see <i>USE_BASES Option</i> on page 31.
s_N_TTTT_qcal.txt	ANALYSIS default ANALYSIS eland ANALYSIS monotemplate	Contains quality values for each base, recalibrated using a calibration table derived from the alignments.
s_N_qcal.txt	ANALYSIS eland_extended	
s_N_R_qcal.txt	ANALYSIS eland_pair	
s_N_eland_query.txt	ANALYSIS eland_extended	Contains all reads for lane N and are concatenated into a single fasta file to use as an ELAND query.
s_N_R_eland_query.txt	ANALYSIS eland_pair	
s_N_eland_result.txt	ANALYSIS eland_extended	Contains the initial output of the ELAND alignment program run in the "standard" single-match mode.
s_N_R_eland_result.txt	ANALYSIS eland_pair	
s_N_eland_multi.txt	ANALYSIS eland_extended	Contains the initial output of the ELAND alignment program run in multiple-match mode.
s_N_R_eland_multi.txt	ANALYSIS eland_pair	
s_N_frag.txt	ANALYSIS eland_extended	Contains the alignment positions (based on at most 32 bases) and does an alignment of the full read to each position. A numeral refers to a run of matching bases, while an upper case base or N refers to a base in the reference that differs from the read.
s_N_eland_extended.txt	ANALYSIS eland_extended	Contains the corrected alignment positions and the full alignment descriptions for >32 base reads. This file is not purity filtered.
s_N_R_eland_extended.txt	ANALYSIS eland_pair	
s_N_saf.txt	ANALYSIS eland_extended	Short Alignment Format (SAF) aims to describe the best alignment for each read. The raw quality values are used to pick the best alignment from the (potentially) multiple possibilities.
s_N_R_saf.txt	ANALYSIS eland_pair	The software aims to pick the pair of alignments that is most consistent with the statistical distribution of insert sizes.

Table 16 Intermediate Output File Descriptions (Continued)

Output File	GERALD Analysis Mode	Description
s_N_calsaf.txt	ANALYSIS eland_extended	These files are identical in format to s_N_saf.txt and s_N_R_saf.txt except the calibrated quality values are used to pick the best alignment.
s_N_R_calsaf.txt	ANALYSIS eland_pair	
s_N_qval.txt, s_N_qtable.txt	ANALYSIS default ANALYSIS eland	These are intermediate files produced during the generation of s_N_qcal.txt. Normally, they are deleted when the analysis is completed, but may be present in an analysis folder if the analysis was interrupted for any reason.

Table 17 Contaminant Filtering-Specific Files

Output File	GERALD Analysis Mode	Description
s_N_TTTT_calign.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	This file contains the first pass alignments of the sequences in the tile against the file of contaminant sequences specified in CONTAM_FILE. If CONTAM_FILE is not specified, this file is not produced.
s_N_TTTT_cdifff.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	This file is only produced if CONTAM_FILE is specified. It contains the difference in alignment scores of alignment to data versus alignment to contaminant file. If negative, the corresponding sequence aligns better to contaminant than to data.
s_N_TTTT_crealign.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	This file is only produced if CONTAM_FILE is specified. It contains realignments of the sequences in the tile against the file of contaminant sequences specified in CONTAM_FILE. The error rate information contained in s_N_score.txt refines the alignment by re-weighting each base at each cycle according to its confidence.
s_N_TTTT_cpredifff.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	This file is only produced if CONTAM_FILE is specified. It contains differences in alignment scores of realignment to data from s_N_prealign.txt versus realignment to contaminant file s_N_crealign.txt. If the entry is negative, the corresponding sequence aligns better to contaminant than to data. This data is analogous to the data in s_N_cdifff.txt.
s_N_TTTT_credifff.txt	ANALYSIS default ANALYSIS eland (if contaminant filtering is switched on)	This file is only produced if CONTAM_FILE is specified. It contains differences in alignment scores of realignment to data from s_N_prealign.txt versus alignment to contaminant file s_N_crealign.txt, and is filtered to exclude alignments for those clusters that do not pass the quality criterion QF_PARAMS when applied to s_N_qhg.txt. This is based on s_N_cpredifff.txt, filtered to have a line-to-line correspondence with the realignments in s_N_realign.txt.

Output File Formats

The sequences and base-specific quality scores are bundled by lane and come in several configurable text formats. The currently supported formats are fasta, fastq, and SCARF. For a description of each format, see *ANALYSIS Variables* on page 29.

Quality scores are stored as either symbolic ASCII values or numeric values. The parameters that set the configuration of the output format are described in *ANALYSIS Variables* on page 29.

Table 18 Final Output File Formats

Output File	Format
s_N_export.txt s_N_R_export.txt	<p>Not all fields are relevant to a single-read analysis.</p> <ol style="list-style-type: none"> 1. Machine (Parsed from Run Folder name) 2. Run Number (Parsed from Run Folder name) 3. Lane 4. Tile 5. X Coordinate of cluster 6. Y Coordinate of cluster 7. Index string (Blank for a non-indexed run) 8. Read number (1 or 2 for paired-read analysis, blank for a single-read analysis) 9. Read 10. Quality string—In symbolic ASCII format (ASCII character code = quality value + 64) by default (Set QUALITY_FORMAT --numeric in theGERALD config file for numeric values) 11. Match chromosome—Name of chromosome match OR code indicating why no match resulted 12. Match Contig—Gives the contig name if there is a match and the match chromosome is split into contigs (Blank if no match found) 13. Match Position—Always with respect to forward strand, numbering starts at 1 (Blank if no match found) 14. Match Strand—"F" for forward, "R" for reverse (Blank if no match found) 15. Match Descriptor—Concise description of alignment (Blank if no match found) <ul style="list-style-type: none"> • A numeral denotes a run of matching bases • A letter denotes substitution of a nucleotide: For a 35 base read, "35" denotes an exact match and "32C2" denotes substitution of a "C" at the 33rd position 16. Single-Read Alignment Score—Alignment score of a single-read match, or for a paired read, alignment score of a read if it were treated as a single read (Blank if no match found) 17. Paired-Read Alignment Score—Alignment score of a paired read and its partner, taken as a pair (Blank for single-read analysis) 18. Partner Chromosome—Name of the chromosome if the read is paired and its partner aligns to another chromosome (Blank for single-read analysis) 19. Partner Contig—Not blank if read is paired and its partner aligns to another chromosome and that partner is split into contigs (Blank for single-read analysis) 20. Partner Offset—If a partner of a paired read aligns to the same chromosome and contig, this number, added to the Match Position, gives the alignment position of the partner (Blank for single-read analysis) 21. Partner Strand—To which strand did the partner of the paired read align? "F" for forward, "R" for reverse (Blank if no match found, blank for single-read analysis) 22. Filtering—Did the read pass quality filtering? "Y" for yes, "N" for no

Table 18 Final Output File Formats (Continued)

Output File	Format
s_N_sequence.txt s_N_R_sequence.txt	Filtered output User-specified: fasta, fastq, scarf (one sequence per line, not identifier)
s_N_TTTT_realign.txt	Final quality-filtered sequence alignments Space-separated text values: <ol style="list-style-type: none"> 1. sequence 2. best score 3. number of hits at that score The following columns only appear if hits equal 1 (a single, unique match) <ol style="list-style-type: none"> 4. target:pos 5. strand 6. target sequence 7. next best score
s_N_rescore.txt	Estimate of the error rate based on s_N_TTTT_realign.txt Tabular text format, header data included

Table 19 Intermediate Output File Formats

Output File	Format
s_N_eland_results.txt s_N_R_eland_results.txt	Unfiltered ELAND alignment output Each line of the output file contains the following fields: <ol style="list-style-type: none"> 1. Sequence name (derived from file name and line number if format is not fasta) 2. Sequence 3. Type of match codes: <ul style="list-style-type: none"> • NM—No match found • QC—No matching done: QC failure (too many Ns) • RM—No matching done: repeat masked (may be seen if repeatFile.txt was specified) • U0—Best match found was a unique exact match • U1—Best match found was a unique 1-error match • U2—Best match found was a unique 2-error match • R0—Multiple exact matches found • R1—Multiple 1-error matches found, no exact matches • R2—Multiple 2-error matches found, no exact or 1-error matches 4. Number of exact matches found 5. Number of 1-error matches found 6. Number of 2-error matches found 7. The following fields are only used if a unique best match was found: 8. Genome file in which match was found 9. Position of match (bases in file are numbered starting at 1) 10. Direction of match (F=forward strand, R=reverse) 11. How N characters in read were interpreted (". "=not applicable, "D"=Deletion, "I"=Insertion) The following field is only used in the case of a unique inexact match: <ol style="list-style-type: none"> 12. Position and type of first substitution error (A numeral refers to a run of matching bases, an upper case base or N refers to a base in the reference that differs from the read. For example, 11A: after 11 matching bases, base 12 is A in the reference but not in the read)

Table 19 Intermediate Output File Formats (Continued)

Output File	Format
s_N_eland_multi.txt s_N_R_eland_multi.txt	<p>Each line of the output file contains the following fields:</p> <ol style="list-style-type: none"> 1. Sequence name 2. Sequence 3. Either NM, QC, RM (as described above) or the following: 4. x:y:z where x, y, and z are the number of exact, single-error, and 2-error matches found 5. Blank, if no matches found or if too many matches found, or the following: BAC_plus_vector.fa:163022R1,170128F2,E_coli.fa:3909847R1 This says there are two matches to BAC_plus_vector.fa: one in the reverse direction starting at position 160322 with one error, one in the forward direction starting at position 170128 with two errors. There is also a single-error match to E_coli.fa.
s_N_TTTT_align.txt	<p>Unfiltered first-pass alignments</p> <p>Each line of the output file contains the following fields:</p> <ol style="list-style-type: none"> 1. Sequence 2. Best score 3. Number of hits at that score <p>The following columns only appear if hits equal 1 (a single, unique match)</p> <ol style="list-style-type: none"> 4. Target:pos 5. Strand 6. Target sequence 7. Next best score
s_N_TTTT_prealign.txt	<p>Unfiltered second-pass alignments</p> <p>Each line of the output file contains the following fields:</p> <ol style="list-style-type: none"> 1. Sequence 2. Best score 3. Number of hits at that score <p>The following columns only appear if hits equal 1 (a single, unique match)</p> <ol style="list-style-type: none"> 4. Target:pos 5. Strand 6. Target sequence 7. Next best score

Parameters File Format

The top level Run Folder contains a parameters file, named <Run Folder Name>.params, and is written in the following format:

```
<experiment>
  <run>
  ...
</run>
<run>
  ...
</run>
</experiment>
```

For each restart of the instrument, a new run tag with corresponding parameter tags is added to the parameters file. For most experiments, there will only be one run.

The XML tags in the parameters file are self-explanatory. The following shows an example of a parameters file:

```
<experiment>
  <run>
    <instrument>slxa-b1</instrument>
  </run>
</experiment>
```

In the top level of the Data folder you will find the parameters file that records any information specific to the generation of the subfolders. This contains a tag-value list describing the cycle-image folders used to generate each folder of intensity and sequence files.

```
<?xml version="1.0"?>
<ImageAnalysis>
  <Run Name="C1-24_Firecrest1.9.0_30-07-2007_user">
    <Cycles First="1" Last="24" Number="24" />
    <ImageParameters>
      <AutoOffsetFlag>1</AutoOffsetFlag>
      <AutoSizeFlag>0</AutoSizeFlag>
      <DataOffsetFile>/data/070813_ILMN-1_0217_FC1234/
        Data/default_offsets.txt</DataOffsetFile>
      <Fwhm>2.700000</Fwhm>
      <InstrumentOffsetFile></InstrumentOffsetFile>
      <OffsetFile>/data/070813_ILMN-1_0217_FC1234/Data/
        default_offsets.txt</OffsetFile>
      <Offsets X="0.000000" Y="0.000000" />
      <Offsets X="0.790000" Y="-0.550000" />
      <Offsets X="-0.240000" Y="-0.140000" />
      <Offsets X="0.190000" Y="0.650000" />
      <RemappingDistance>1.500000</RemappingDistance>
      <SizeFile></SizeFile>
      <Threshold>4.000000</Threshold>
    </ImageParameters>
    <RunParameters>
      <AutoCycleFlag>0</AutoCycleFlag>
      <BasecallFlag>1</BasecallFlag>
      <Compression>gzip</Compression>
      <CompressionSuffix>.gz</CompressionSuffix>
      <Deblocked>0</Deblocked>
```

```

<DebugFlag>0</DebugFlag>
<ImagingReads Index="1">
  <FirstCycle>1</FirstCycle>
  <LastCycle>24</LastCycle>
  <RunFolder>/data/070813_ILMN-1_0217_FC1234</
  RunFolder>
</ImagingReads>
<Instrument>ILMN-1</Instrument>
<MakeFlag>1</MakeFlag>
<MaxCycle>-1</MaxCycle>
<MinCycle>-1</MinCycle>
<Reads Index="1">
  <FirstCycle>1</FirstCycle>
  <LastCycle>24</LastCycle>
  <RunFolder>/data/070813_ILMN-1_0217_FC1234</
  RunFolder>
</Reads>
<RunFolder>/data/070813_ILMN-1_0217_FC1234</
  RunFolder>
<Software Name="Firecrest" Version="1.9.0" />
<TileSelection>
  <Lane Index="8">
    <Sample>s</Sample>
    <Tile>10</Tile>
    <Tile>20</Tile>
    <Tile>30</Tile>
  </Lane>
</TileSelection>
<Time>
  <Start>30-07-07 12:50:45 BST</Start>
</Time>
<User Name="user" />
</Run>
<Run Name="Cl-24_Firecrest1.9.0_30-07-2007_user.2">
  ...
</Run>
</ImageAnalysis>

```

In each image analysis folder there is another parameters file containing the meta-information about the base caller runs.

```

<?xml version="1.0"?>
<BaseCallAnalysis>
  <Run Name="Bustard1.9.0_30-07-2007_user">
    <BaseCallParameters>
      <Matrix Path="">
        <AutoFlag>1</AutoFlag>
        <AutoLane>0</AutoLane>
        <Cycle>2</Cycle>
        <FirstCycle>1</FirstCycle>
        <LastCycle>24</LastCycle>
        <Read>1</Read>
      </Matrix>
      <MatrixElements />
      <Phasing Path="">
        <AutoFlag>1</AutoFlag>
        <AutoLane>0</AutoLane>
        <Cycle>1</Cycle>

```

```
<FirstCycle>1</FirstCycle>
<LastCycle>24</LastCycle>
<Read>1</Read>
</Phasing>
<PhasingRestarts />
</BaseCallParameters>
<Cycles First="1" Last="24" Number="24" />
<Input Path="C1-24_Firecrest1.9.0_30-07-
2007_user.2" />
<RunParameters>
  <AutoCycleFlag>0</AutoCycleFlag>
  <BasecallFlag>1</BasecallFlag>
  <Compression>gzip</Compression>
  <CompressionSuffix>.gz</CompressionSuffix>
  <Deblocked>0</Deblocked>
  <DebugFlag>0</DebugFlag>
  <ImagingReads Index="1">
    <FirstCycle>1</FirstCycle>
    <LastCycle>24</LastCycle>
    <RunFolder>/data/070813_ILMN-1_0217_FC1234</
    RunFolder>
  </ImagingReads>
  <Instrument>ILMN-1</Instrument>
  <MakeFlag>1</MakeFlag>
  <MaxCycle>-1</MaxCycle>
  <MinCycle>-1</MinCycle>
  <Reads Index="1">
    <FirstCycle>1</FirstCycle>
    <LastCycle>24</LastCycle>
    <RunFolder>/data/070813_ILMN-1_0217_FC1234</
    RunFolder>
  </Reads>
  <RunFolder>/data/070813_ILMN-1_0217_FC1234</
  RunFolder>
</RunParameters>
<Software Name="Bustard" Version="1.9.0" />
<TileSelection>
  <Lane Index="5">
    <Sample>s</Sample>
    <TileRange Max="5" Min="5" />
  </Lane>
</TileSelection>
<Time>
  <Start>30-07-07 18:01:50 BST</Start>
</Time>
<User Name="user" />
</Run>
</BaseCallAnalysis>
```




Appendix C

Using Parallelization

Topics

- 88 Introduction
- 88 “Make” Utilities
 - 88 Standard “Make”
 - 88 Customizing Parallelization
 - 88 Distributed “Make”
- 91 Parallelization Limitations
- 91 Memory Limitations

Introduction

One of the main considerations behind the current Pipeline architecture is the ability to use the parallelization facilities present on almost all SMP machines and on most Linux/Unix clusters. Parallelization is scalable and makes use of all available CPU power.

“Make” Utilities

Parallelization is built around the ability of the standard “make” utility to execute in parallel across multiple processes on the same computer. Since version 0.2.2, the Pipeline also provides a series of checkpoints and hooks that enables you to customize the parallelization for your computing setup. See *Customizing Parallelization* on page 88 for details.

Standard “Make”

The standard “make” utility has many limitations, but it is universally available and has a built-in parallelization switch (“-j”). For example, on a dual-processor, dual-core system, running “make -j 4” instead of “make,” executes the Pipeline run in parallel over four different processor cores, with an almost 4-fold decrease in analysis run time. On a 4-way SMP system, “-j 8” or more may be advisable.

Distributed “Make”

There are several distributed versions of “make” for cluster systems. Frequently used versions include “qmake” from Sun Grid Engine and “lsmake” from LSF.

To use “qmake,” a short wrapper script is required. See the grid engine documentation for details.

There are known issues with the use of “lsmake” that prevent parts of the Pipeline from running. Therefore, Illumina does not recommend using “lsmake” to run the Pipeline.



Distributed cluster computing may require significant system administration expertise. Illumina does not support external installations.

Customizing Parallelization

Many parts of the Pipeline are intrinsically parallelizable by lane or tile. However, some parts of the Pipeline cannot be parallelized completely. Pipeline v.0.2.2 and later, has a series of additional hooks and check-points for customization.

The Pipeline workflow is divided into the image analysis, base calling, alignment. You can divide it further into a series of steps with different levels of scalability where synchronization “barriers” cause the Pipeline to wait for each of the tasks within a step to finish before going to the next step.

You can parallelize the steps at the run level (no parallelization), the lane level (up to eight jobs in parallel), and the tile level (up to thousands of jobs in parallel). Each step is initiated by a "make" target. After completion of each of these steps, the Pipeline produces a file or a series of files at the lane/tile level, that determines whether all jobs belonging to the step have finished. Finally, hooks are provided upon completion of the step to issue user-defined external commands.

The Firecrest Makefile creates two files, lanes.txt and tiles.txt, containing a list of all lanes and tiles used in the run. This information is parsed and used to feed your own analysis scripts.

Example of Parallelization

Typing "make" in the Firecrest folder is equivalent to the following series of commands:

```
make default_offsets.txt
make s_1; make s_2; make s_3; make s_4; make s_5;
      make s_6; make s_7; make s_8
make all
```

This command addresses each lane sequentially. Using parallelization, you can run all eight commands on the second line in parallel, as long as you make sure that they all finish before the final "make all" is issued. There are several ways to parallelize these jobs. For example, you could send them to the queue of a batch system, or just use "ssh" or "rsh" to send them to a predetermined analysis computer.

In the following example, the second step is automatically started after the first step (make s_1;) as the external command, "cmdf1." The external command will be issued after completion of the first step.

```
make -j 2 default_offsets.txt cmdf1='make s_1;
      make s_2; make s_3; make s_4; \
make s_5; make s_6; make s_7; make s_8;' \
cmdf2='if [[ -e s_1_finished.txt && -e
      s_2_finished.txt && -e s_3_finished.txt \
&& -e s_4_finished.txt && -e s_5_finished.txt
&& -e s_6_finished.txt \
&& -e s_7_finished.txt && -e s_8_finished.txt ]]; then
      make all ; fi #'
```

This only makes sense if you parallelize the eight "make" commands instead of using "make s_1," as shown in the following example:

```
nohup ssh <mycomputenodel> make -j 4 s_1
```

—or—

```
bsub make s_1
```

After completing the eight "make" commands in the second step, the shell command "cmdf2" is run to check for the existence of all eight checkfiles. The next make command (make all) will be issued only after the completion of the first seven lanes.

```
if [[ -e s_1_finished.txt && -e s_2_finished.txt
&& -e s_3_finished.txt \
&& -e s_4_finished.txt && -e s_5_finished.txt
&& -e s_6_finished.txt \
&& -e s_7_finished.txt && -e s_8_finished.txt ]];
then make all ; fi #
```

The reason for the final comment symbol (#) at the end of the shell command above is that the Pipeline automatically supplies an argument to all commands issued at the lane level and is used as an identifier for the actual lane analyzed. In the example above, this argument is not used, and so it needs to be commented out.

**NOTE**

There is no need to declare the full shell command on the command line. You could put all of the shell commands into a shell script and call that script instead.

Image Analysis

This section lists the steps, corresponding make targets, checkfiles, and hooks for image analysis by the Firecrest module.

Parallelization Level	Run	Lane	Tile
Target	default_offsets.txt		
Check File	default_offsets.txt		
Hook	cmdf1		
Target		s_1	s_1_0001
Check File		s_1_finished.txt	(none)
Hook		cmdf2	(none)
Target	all		
Check File	finished.txt		
Hook	cmdf3		

Base Calling

This section lists the steps, corresponding make targets, checkfiles, and hooks for base calling by the Bustard module.

Parallelization Level	Run	Lane	Tile
Target		Phasing/ s_1_phasing.xml	Phasing/ s_1_0001_phasing.txt
Check File		Phasing/ s_1_phasing.xml	Phasing/ s_1_0001_phasing.txt
Hook		cmdb1	(none)
Target	Phasing/ phasing.xml		

Parallelization Level	Run	Lane	Tile
Check File	Phasing/ phasing.xml		
Hook	cmdb2		
Target		s_1	s_1_0001
Check File		s_1_finished.txt	s_1_0001_qhg.txt
Hook		cmdb3	(none)
Target	all		
Check File	finished.txt		
Hook	cmdb4		

Sequence Alignment

This section lists the steps, corresponding make targets, checkfiles and hooks for sequence alignment by the GERALD module.

Parallelization Level	Run	Lane
Target		s_1
Check File		s_1_finished.txt
Hook		(none)
Target	all	
Check File	finished.txt	
Hook	POST_RUN_COMMAND (Accessible from GERALD config file)	

Parallelization Limitations

The analysis works on a per-tile basis, so the maximum degree of parallelization achievable is equal to the total number of tiles scanned during the run. However, some parts of the Pipeline operate on a per-lane basis, and a few parts on a per-run basis, which means that scaling will cease to be linear at some stage for more than 8-way parallelization.

Memory Limitations

Most parts of the Pipeline have moderate memory requirements (<150 MB). ELAND uses up to 1 GB, which means that parallelization of ELAND is more likely to run into memory issues. Because many load-sharing systems do not take into account the memory used, ELAND is treated differently in the Pipeline. Its parallelization is artificially prevented by a non-essential "make" dependency. If you are certain that you cannot exhaust your available memory, you can use a special option to the GERALD configuration file (ELAND_MULTIPLE_INSTANCES 8) to remove this dependency. However, you are responsible for making sure that you have up to 8 GB of RAM at your disposal. For additional information, see *Using GERALD* on page 27.

Illumina, Inc.
9885 Towne Centre Drive
San Diego, CA 92121-1975
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com

