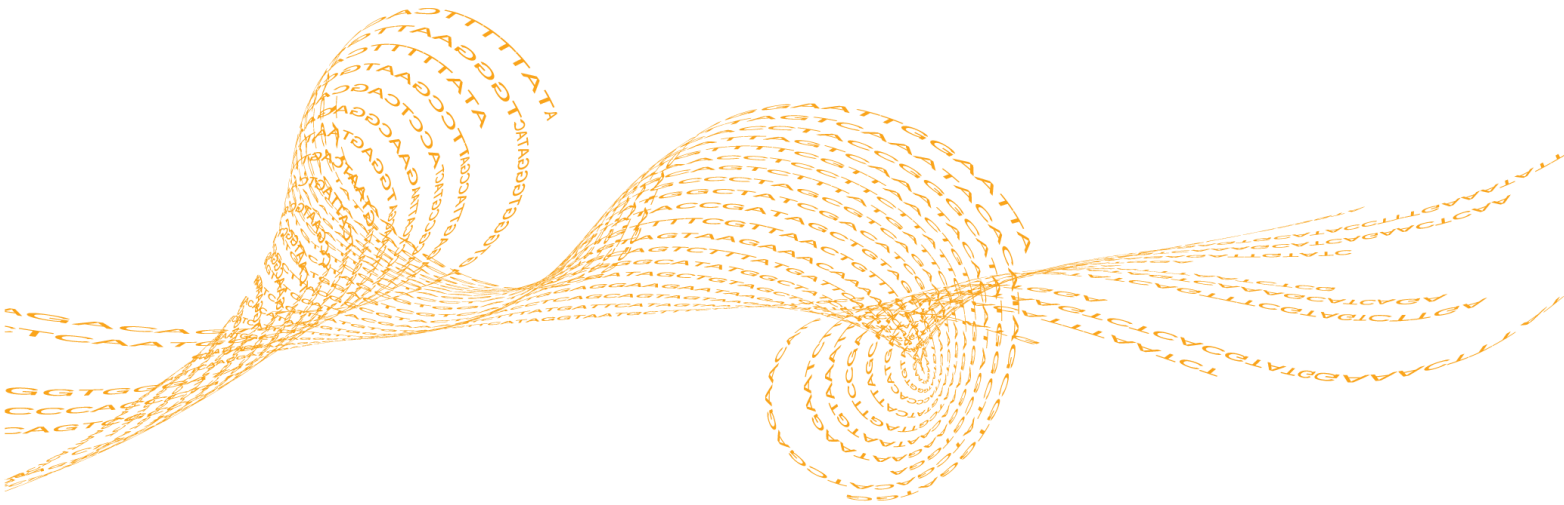
Isaac Enrichment App



ILLUMINA PROPRIETARY
15050960 Rev. A
February 2014

illumina®

Introduction

The BaseSpace® app Isaac Enrichment analyzes DNA that has been enriched for particular target sequences using Nextera® Rapid Capture. Alignment is performed with Isaac, and variant calling with the Isaac Variant Caller. Variant analysis is performed for just the target regions. Statistics reporting accumulates coverage and enrichment-specific statistics for each target as well as overall metrics.

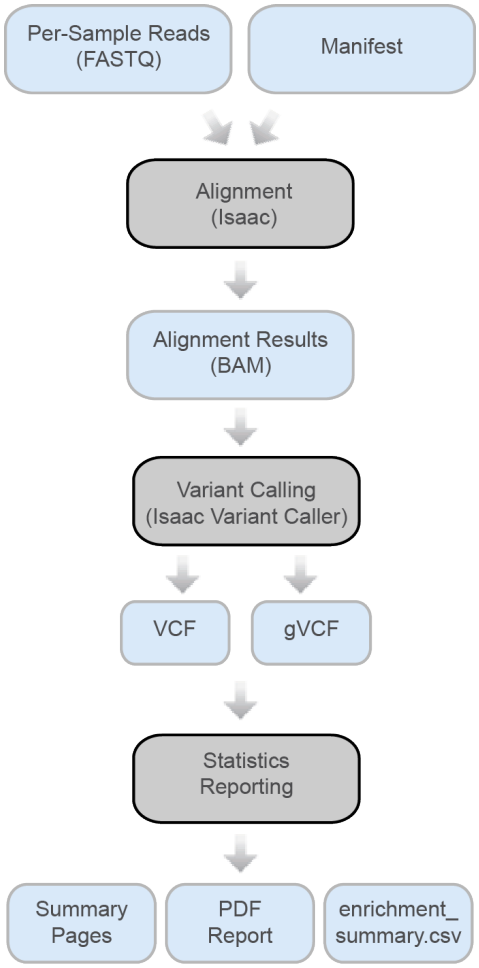
The main output files generated by the Isaac Enrichment workflow are:

- BAM files, containing the reads after alignment.
- VCF files, containing the variant calls.
- Genome VCF (.genome.vcf) files, describing the calls for all variant and non-variant sites in the genome.
- Annotated VCF file. This binary file can be loaded into VariantStudio for viewing; see www.illumina.com/clinical/clinical_informatics/illumina-variantstudio.ilmn.

In addition, there is a biological summary, QC summary, PDF report, and enrichment.csv file.

See *Isaac Enrichment Methods* on page 26 and *Isaac Enrichment Output* on page 7 for more information.

Figure 1 Isaac Enrichment App Workflow



Versions

The following module versions are used in the Isaac Enrichment app:

- Isaac: 01.13.10.21
- Isaac Variant Caller: 2.0.17
- Picard: 1.79
- Samtools: 0.1.18
- Tabix: 0.2.5 (r1005)
- IAS (Annotation Service) VEP 72.4

Current Limitations

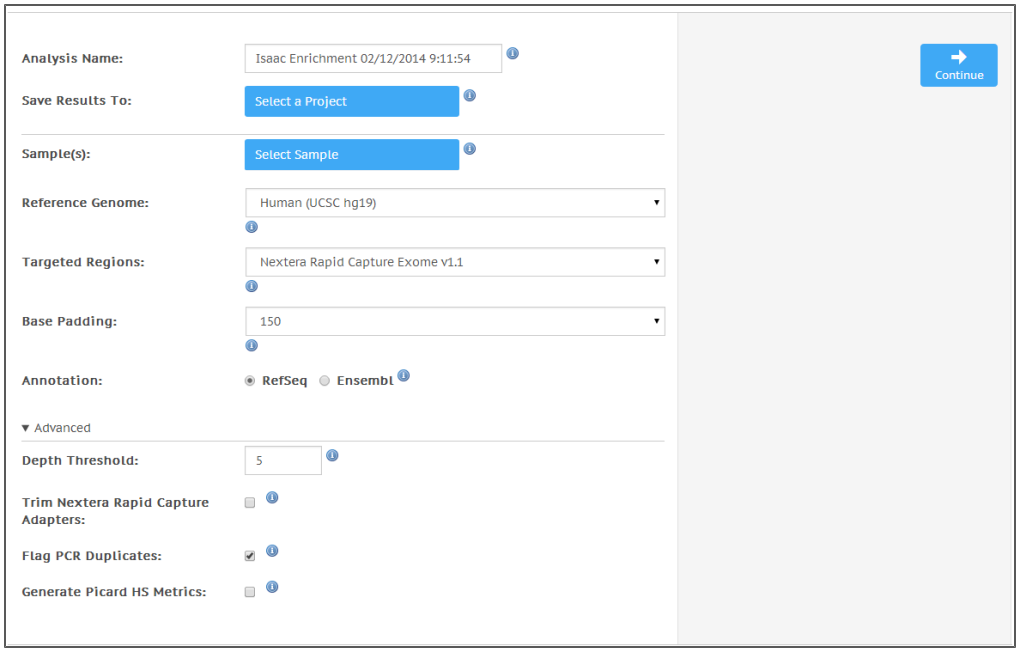
Before running the Isaac Enrichment app, be aware of the following limitations:

- hg19 reference only
- Read length of at least 32 bp
- Data set size smaller than 200 gigabases
- No minimum number of reads, but use a reasonable input size to get your required coverage

Running Isaac Enrichment

1. Navigate to the project or sample that you want to analyze.
2. Click the **Apps** button and select **Isaac Enrichment** from the drop-down list.
3. Read the End-User License Agreement and permissions, and click **Accept** if you agree.
4. Fill out the required fields in the Isaac Enrichment input form:
   a. **Analysis Name**: Provide the analysis name. Default name is the app name with the date and time the analysis was started.
   b. **Save Results To**: Select the project that stores the app results.
   c. **Sample(s)**: Browse to the sample you want to analyze, and select the checkbox. You can analyze multiple samples.
   d. **Reference Genome**: Select the reference genome. Currently, you can only use hg19.
   e. **Targeted Regions**: Select the targeted region of your enrichment.
   f. **Base Padding**: Select the padding you want. Padding defines the amount of sequence immediately upstream and downstream of the targeted regions that is also used in enrichment analysis.
   g. **Annotation**: Choose which gene and transcript annotation reference database to use.
5. If desired, fill out the advanced fields in the Isaac Enrichment input form:
   a. **Depth Threshold**: If the coverage depth at a location is less than the specified threshold, the Isaac Variant Caller filters the variant. Decreasing this value increases variant calling sensitivity, but raise the risk of false positives. The variant caller reports the variants, but filters them in the VCF files by adding a LowDP flag. Default value: 5.
   b. **Trim Nextera Rapid Capture Adapters**: If selected, Nextera Rapid Capture adapters are trimmed. Use this setting only if not already applied as a sample sheet setting.
   c. **Flag PCR Duplicates**: If selected, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as two clusters from a paired-end run where both clusters have the exact same alignment positions for each read. Optical duplicates are already filtered out during RTA processing.
   d. **Generate Picard HS Metrics**: If selected, Picard HS metrics are generated. See *Picard Metrics* on page 28 for more information.

Figure 2 Isaac Enrichment Input Form



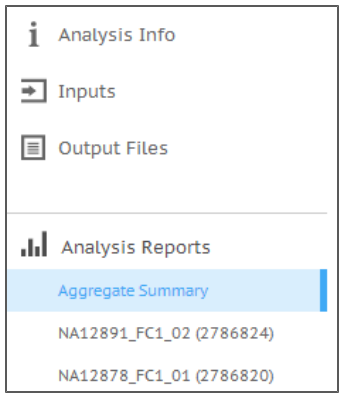
6 Click **Continue**.

The Isaac Enrichment app now starts analyzing your sample. When completed, the status of the app session is automatically updated, and you receive an email.

Isaac Enrichment Output

This chapter describes the Isaac Enrichment app output. To go to the results, click the **Projects** button, then the project, then the analysis.

Figure 3 Isaac Enrichment Output Navigation Bar



When the analysis is completed, you can access your output through the left navigation bar, which provides the following:

- **Analysis Info**: an overview of the app session settings. See *Analysis Info* on page 15 for a description.
- **Output Files**: access to the output files, organized by sample and app session. See*Isaac Enrichment Output Files* on page 17 for descriptions.
- **Inputs**: overview of input settings, see *Inputs Overview* on page 17
- **Aggregate Summary**: access to analysis metrics for the aggregate results. The Aggregate Summary is only displayed if multiple samples are analyzed. See *Aggregate Summary Page* on page 7.
- **Sample Pages**: access to analysis reports for each sample. See *Sample Summary Page* on page 9 for a description.

Aggregate Summary Page

The Isaac Enrichment app provides an overview of metrics for all samples combined on the Aggregate Summary page. You can view the histograms with metrics graphed by sample, or look at the numerical data by clicking Show tables with details.

NOTE
PCR duplicate reads are not removed from statistics. Results are not directly comparable to Picard HsMetrics.

Alignment Summary

The mean region coverage depth is plotted against sample, and a table is available with additional metrics. The metrics displayed are explained here:

Statistic	Definition
Mean coverage	The total number of targeted bases divided by the targeted region size.
Target coverage at 1X	Percentage targets with coverage greater than 1X.

Statistic	Definition
Target coverage at 10X	Percentage targets with coverage greater than 10X.
Target coverage at 20X	Percentage targets with coverage greater than 20X.
Target coverage at 50X	Percentage targets with coverage greater than 50X.

Enrichment Summary

Aligned bases are plotted against sample, and a table is available with additional metrics. The metrics displayed are explained here:

- Read Level

Statistic	Definition
Total aligned reads	The total number of reads passing filter that aligned.
Percent aligned reads	Percentage of reads passing filter that aligned.
Target aligned reads	Number of reads that aligned to the target.
Read enrichment	100*(Target aligned reads/Total aligned reads).
Padded target aligned reads	Number of reads that aligned to the padded target.
Padded read enrichment	100*(Padded target aligned reads/Total aligned reads).

- Base Level

Statistic	Definition
Total aligned bases	Total aligned bases.
Target aligned bases	Total aligned bases in the target region.
Base enrichment	100*(Total Aligned Bases in Targeted Regions/Total Aligned Bases).
Padded target aligned bases	Total aligned bases in the padded target region.
Padded base enrichment	100*(Total Aligned Bases in Padded Targeted Regions/Total Aligned Bases).

Variant Summary

- SNVs

  Number of SNVs passing are plotted against sample, and a table is available with additional metrics. The metrics displayed are explained here:

Statistic	Definition
Total Passing	Total number of Single Nucleotide Variants present in the data set passing the quality filters.

Statistic	Definition
Het/Hom Ratio	Number of heterozygous SNVs/Number of homozygous SNVs.
Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).
Percent found in dbSNP	100*(Number of SNVs in dbSNP/Number of SNVs).

- Indels
  Number of indels passing are plotted against sample, and a table is available with additional metrics. The metrics displayed are explained here:

Statistic	Definition
Total Passing	Total number of indels present in the data set passing the quality filters.
Het/Hom Ratio	Number of heterozygous indels/Number of homozygous indels.
Percent found in dbSNP	100*(Number of Indels in dbSNP/Number of Indels).

Sample Summary Page

The Isaac Enrichment app provides an overview of statistics per sample on the sample pages. you can also download the Enrichment Summary Report as PDF.

NOTE
PCR duplicate reads are not removed from statistics. Results are not directly comparable to Picard HsMetrics.

Sample Information

Statistic	Definition
Total PF Reads	The number of reads passing filter for the sample.
Percent Q30	The percentage of bases with a quality score of 30 or higher.
Percent Duplicate Paired Reads	Percentage of paired reads that have duplicates.
Fragment Length Median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.

Statistic	Definition
Fragment Length Standard Deviation	Standard deviation of the sequenced fragment length.

Enrichment Summary

Statistic	Definition
Total length of targeted reference	Total length of sequenced bases in the target region.
Padding size	The length of sequence immediately upstream and downstream of the enrichment targets that is included for a padded target.

Read Level Enrichment

Statistic	Definition
Total aligned reads	The total number of reads passing filter that aligned.
Percent aligned reads	Percentage of reads passing filter that aligned.
Target aligned reads	Number of reads that aligned to the target.
Read enrichment	100*(Target aligned reads/Total aligned reads).
Padded target aligned reads	Number of reads that aligned to the padded target.
Padded read enrichment	100*(Padded target aligned reads/Total aligned reads).

Base Level Enrichment

Statistic	Definition
Total aligned bases	Total aligned bases.
Target aligned bases	Total aligned bases in the target region.
Base enrichment	100*(Total Aligned Bases in Targeted Regions/Total Aligned Bases).
Padded target aligned bases	Total aligned bases in the padded target region.
Padded base enrichment	100*(Total Aligned Bases in Padded Targeted Regions/Total Aligned Bases).

Small Variants Summary

This table provides metrics about the number of SNVs, deletions, and insertions.

Statistic	Definition
Total Passing	Total number of variants present in the data set that pass the quality filters.
Percent Found in dbSNP	100*(Number of variants in dbSNP/Number of variants).
Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.
Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).

Variants by Sequence Context

Statistic	Definition
Number in Genes	The number of variants that fall into a gene.
Number in Exons	The number of variants that fall into an exon.
Number in Coding Regions	The number of variants that fall into a coding region.
Number in UTR Regions	The number of variants that fall into an untranslated region (UTR).
Number in Splice Site Regions	The number of variants that fall into a splice site region.

Variants by Consequence

Statistic	Definition
Frameshifts	The number of variants that cause a frameshift.
Non-synonymous	The number of variants that cause an amino acid change in a coding region.
Synonymous	The number of variants that are within a coding region, but do not cause an amino acid change.
Stop Gained	The number of variants that cause an additional stop codon.
Stop Lost	The number of variants that cause the loss of a stop codon.

Coverage Summary

Statistic	Definition
Mean coverage	The total number of targeted bases divided by the targeted region size.

Statistic	Definition
Uniformity of coverage (Pct > 0.2*mean):	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
Target coverage at 1X	Percentage targets with coverage greater than 1X.
Target coverage at 10X	Percentage targets with coverage greater than 10X.
Target coverage at 20X	Percentage targets with coverage greater than 20X.
Target coverage at 50X	Percentage targets with coverage greater than 50X.

Enrichment Summary Reports by Sample

The Isaac Enrichment app provides enrichment statistics PDF report for each sample.

NOTE
PCR duplicate reads are not removed from statistics. Results are not directly comparable to Picard HsMetrics.

Sample Information

Provides the sample ID and name, and the following metrics:

Statistic	Definition
Total PF Reads	The number of reads passing filter for the sample.
Percent Q30	The percentage of bases with a quality score of 30 or higher.
Median Read Length	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Adapters Trimmed	Whether adapter trimming was used.

Enrichment Summary

Provides the target manifest name, and the following metrics:

Statistic	Definition
Total length of targeted reference	Total length of sequenced bases in the target region.
Padding size	The length of sequence immediately upstream and downstream of the enrichment targets that is included for a padded target.

Read Level Enrichment

Statistic	Definition
Total aligned reads	The total number of reads passing filter that aligned.
Percent aligned reads	The percentage of reads passing filter that aligned.
Targeted aligned reads	Number of reads that aligned to the target.
Read enrichment	100*(Target aligned reads/Total aligned reads).
Padded target aligned reads	Number of reads that aligned to the padded target.
Padded read enrichment	100*(Padded target aligned reads/Total aligned reads).

Base Level Enrichment

Statistic	Definition
Total aligned bases	Total aligned bases.
Targeted aligned bases	Total aligned bases in the target region.
Base enrichment (not padded)	100*(Total Aligned Bases in Targeted Regions/Total Aligned Bases).
Padded target aligned bases	Total aligned bases in the padded target region.
Padded base enrichment	100*(Total Aligned Bases in Padded Targeted Regions/Total Aligned Bases).

Small Variants Summary

This table provides metrics about the number of SNVs, deletions, and insertions.

Statistic	Definition
Total Passing	Total number of variants present in the data set that pass the quality filters.
Percent Found in dbSNP	100*(Number of variants in dbSNP/Number of variants).
Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.
Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).

Variants by Sequence Context

Statistic	Definition
Number in Genes	The number of variants that fall into a gene.
Number in Exons	The number of variants that fall into an exon.
Number in Coding Regions	The number of variants that fall into a coding region.
Number in UTR Regions	The number of variants that fall into an untranslated region (UTR).
Number in Splice Site Regions	The number of variants that fall into a splice site region.

Variants by Consequence

Statistic	Definition
Frameshifts	The number of variants that cause a frameshift.
Non-synonymous	The number of variants that cause an amino acid change in a coding region.
Synonymous	The number of variants that are within a coding region, but do not cause an amino acid change.
Stop Gained	The number of variants that cause an additional stop codon.
Stop Lost	The number of variants that cause the loss of a stop codon.

Coverage Summary

Statistic	Definition
Mean region coverage depth	The total number of targeted bases divided by the targeted region size.
Uniformity of coverage (Pct > 0.2*mean):	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
Target coverage at 1X	Percentage targets with coverage greater than 1X.
Target coverage at 10X	Percentage targets with coverage greater than 10X.
Target coverage at 20X	Percentage targets with coverage greater than 20X.
Target coverage at 50X	Percentage targets with coverage greater than 50X.

In addition, the app provides two graphs:

- A Mean Coverage by Targeted Region graph that plots the mean coverage by the targeted region.

- A Targeted Regions Depth of Coverage graph that plots the number of targeted sequences by the depth of coverage.

Statistic	Definition
Depth of Sequencing Coverage	The coverage depth of a position in the genome refers to the number of sequenced bases that align to that position.
Number of Targeted Bases Covered at Depth	Number of targeted bases that have at least the indicated depth of coverage.
Total Targeted Bases Covered	Total bases aligning to the target regions that have at least the indicated depth of coverage.
Target Coverage	Percent of targeted bases that reach the indicated depth of coverage.

Fragment Length Summary

Statistic	Definition
Fragment length median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Minimum	Minimum length of the sequenced fragment.
Maximum	Maximum length of the sequenced fragment.
Standard Deviation	Standard deviation of the sequenced fragment length.

Gaps Summary

The app also provides a Targeted Regions Gap Length Distribution graph that plots the number of gaps on a log scale by the length of the gap in bases.

Duplicates Information

Statistic	Definition
Percent duplicate paired reads	Percentage of paired reads that have duplicates.

Analysis Details

Provides the analysis settings and software versions used.

Analysis Info

This app provides an overview of the analysis on the Analysis Info page.

A brief description of the metrics is below.

Row	Definition
Name	Name of the app session.
Application	App that generated this analysis.
Date started	Date the app session started.
Date completed	Date the app session completed.
Duration	Duration of analysis.
Session Type	The number of nodes used.
Size	Total size of all output files.
Status	Status of the app session.

Log Files

Clicking the **Log Files** link at the bottom of the Analysis Info page provides access to Isaac Enrichment app log files.

The following files log information to help follow data processing and debugging:

- **WorkflowLog.txt**: Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).
- **WorkflowError.txt**: Workflow standard error output (contains errors messages created while running the workflow).
- **Logging.zip**: Contains all detailed workflow log files for each step of the workflow.
- **IlluminaAppsService.log.copy**: Wrapper log file containing information about communication (get and post requests) between BaseSpace and AWS.
- **CompletedJobInfo.xml**: Contains information about the completed job.
- **SampleSheet.csv**: Sample sheet.

The following files contain additional information in case things (like mono) do not work as expected:

- **monoErr.txt**: Wrapper mono call error output (contains anything that WorkflowError.txt does not catch; in most cases empty, except one line).
- **monoOut.txt**: Wrapper mono call standard output (contains command calling the workflow and anything that WorkflowLog.txt does not catch).

NOTE
For explanation about mono, see www.mono-project.com.

Isaac Enrichment Status

For single samples, the status of the Isaac Enrichment app session can have the following values:

1 Preparing Run Data

2 Finished Preparing Run Data

3 Analysis Started

4 Alignment for Sample {SampleName}

5 Variant analysis for Sample {SampleName}

6 Statistics evaluation for Sample {SampleName}

7 Report generation for Sample {SampleName}

8 Analysis Completed for Sample {SampleName}

9 Finalizing Analysis Results for Sample {SampleName}

10 Finished Finalizing Analysis Results

For multiple samples, the status of the Isaac Enrichment app session can have the following values:

1 Preparing Run Data

2 Finished Preparing Run Data

3 Analysis Started

Begin loop over samples:

4 Analysis Started for Sample {currentSampleIndex} of {totalSampleCount}

5 Alignment for Sample {currentSampleIndex} of {totalSampleCount}

6 Variant analysis for Sample {currentSampleIndex} of {totalSampleCount}

7 Statistics evaluation for Sample {currentSampleIndex} of {totalSampleCount}

8 Report generation for Sample {currentSampleIndex} of {totalSampleCount}

9 Analysis Completed for Sample {currentSampleIndex} of {totalSampleCount}

10 Finalizing Analysis Results for Sample {currentSampleIndex} of {totalSampleCount}

When all samples are done:

11 Aggregate Analysis Started

12 Finalizing PDF Reports

13 Finalizing Aggregate Results

14 Finished Finalizing Analysis Results

Inputs Overview

The Isaac Enrichment app provides an overview of the input samples and settings that were specified when setting up the Isaac Enrichment run.

Isaac Enrichment Output Files

The Output Files page provides access to the output files. See the following pages for descriptions:

- *BAM Files* on page 17
- *VCF Files* on page 18
- *gVCF Files* on page 19
- *Enrichment_summary.csv* on page 19
- *Aggregate Summary Report* on page 22

BAM Files

The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128

Mb) produced by different sequencing platforms. SAM is a text format file that is human-readable. The Binary Alignment/Map (BAM) keeps the same information as SAM, but in a compressed, binary format that is only machine readable.

If you use an app in BaseSpace that uses BAM files as input, the app locates the file when launched. If using BAM files in other tools, download the file to use it in the external tool.

Go to samtools.sourceforge.net/SAM1.pdf to see the exact SAM specification.

VCF Files

VCF is a text file format that contains information about variants found at specific positions in a reference genome. The file format consists of meta-information lines, a header line, and then data lines. Each data line contains information about a single variant.

If you use an app in BaseSpace that uses VCF files as input, the app locates the file when launched. If using VCF files in other tools, download the file to use it in the external tool.

A detailed description of the VCF format is provided in the *BaseSpace User Guide*.

Isaac Enrichment VCF Entries

The VCF files for Isaac Enrichment can have the following entries in the FILTER, FORMAT, and INFO fields:

Table 1 VCF FILTER Entries

Entry	Description
IndelConflict	Locus is in region with conflicting indel calls
SiteConflict	Site genotype conflicts with proximal indel call, typically a heterozygous SNV call made inside of a heterozygous deletion
LowGQX	Locus GQX is less than 30 or not present
HighDPFRatio	The fraction of base calls filtered out at a site is greater than 0.4
HighSNVSB	SNV strand bias value (SNVSB) exceeds 10
HighDepth	Locus depth is greater than 3x the mean chromosome depth
OffTarget	Variant is not on target

Table 2 VCF FORMAT Entries

Entry	Description
GQX	Minimum of {Genotype quality assuming variant position,Genotype quality assuming non-variant position}
GT	Genotype
GQ	Genotype Quality
DP	Filtered base call depth used for site genotyping
DPF	Base calls filtered from input before site genotyping
AD	Allelic depths for the ref and alt alleles in the order listed. For indels, this value only includes reads that confidently support each allele (posterior probability 0.999 or higher that read contains indicated allele vs all other intersecting indel alleles)
DPI	Read depth associated with indel, taken from the site preceding the indel.

Table 3 VCF INFO Entries

Entry	Description
SNVSB	SNV site strand bias
SNVHPOL	SNV contextual homopolymer length
CIGAR	CIGAR alignment for each alternate indel allele
RU	Smallest repeating sequence unit extended or contracted in the indel allele relative to the reference. RUs longer than 20 bases are not reported.
REFREP	Number of times RU is repeated in reference.
IDREP	Number of times RU is repeated in indel allele.
END	End position of the region described in this record
BLOCKAVG_ min30p3a	Non-variant site block. All sites in a block are constrained to be non-variant, have the same filter value, and have all sample values in range [x,y], y <= max(x+3,(x*1.3)). All printed site block sample values are the minimum observed in the region spanned by the block

gVCF Files

This application also produces the Genome Variant Call Format file (gVCF). gVCF was developed to store sequencing information for both variant and non-variant positions, which is required for human clinical applications. gVCF is a set of conventions applied to the standard variant call format (VCF) 4.1 as documented by the 1000 Genomes Project. These conventions allow representation of genotype, annotation, and other information across all sites in the genome in a compact format. Typical human whole-genome sequencing results expressed in gVCF with annotation are less than 1 Gbyte, or about 1/100 the size of the BAM file used for variant calling. If you are performing targeted sequencing, gVCF is also an appropriate choice to represent and compress the results.

gVCF is a text file format, stored as a gzip compressed file (*.genome.vcf.gz). Compression is further achieved by joining contiguous non-variant regions with similar properties into single 'block' VCF records. To maximize the utility of gVCF, especially for high stringency applications, the properties of the compressed blocks are conservative. Block properties like depth and genotype quality reflect the minimum of any site in the block. The gVCF file can be indexed (creating a *.tbi file) and used with existing VCF tools such as tabix and IGV, making it convenient both for direct interpretation and as a starting point for tertiary analysis.

For more information, see sites.google.com/site/gvcftools/home/about-gvcf.

Enrichment_summary.csv

The Isaac Enrichment app produces an overview of statistics for each sample and the aggregate results in a comma-separated values (CSV) format: the *.enrichment_ summary.csv. These files are located in the results folder for each sample and the aggregate results.

A brief description of the metrics is below.

NOTE

PCR duplicate reads are not removed from statistics. Results are not directly comparable to Picard HsMetrics.

Statistic	Definition
Sample ID	IDs of samples reported on in the file.
Sample Name	Names of samples reported on in the file.
Run Folder	Run folders for samples reported on in the file.
Target manifest	The target manifest file used for analysis. This file specifies the targeted regions for the aligner and variant caller.
Reference Genome	Reference genome selected.
Padding size	The length of sequence immediately upstream and downstream of the enrichment targets that is included for a padded target.
Total length of targeted reference	Total length of sequenced bases in the target region.
Total PF reads	The number of reads passing filter for the sample.
Percent Q30	The percentage of bases with a quality score of 30 or higher.
Total aligned reads	The total number of reads passing filter that aligned.
Percent aligned reads	Percentage of reads passing filter that aligned.
Total aligned reads	The total number of reads passing filter that aligned.
Target aligned reads	Number of reads that aligned to the target.
Read enrichment	100*(Target aligned reads/Total aligned reads).
Padded target aligned reads	Number of reads that aligned to the padded target.
Padded read enrichment	100*(Padded target aligned reads/Total aligned reads).
Total aligned bases	Total aligned bases.
Target aligned bases	Total aligned bases in the target region.
Base enrichment	100*(Total Aligned Bases in Targeted Regions/Total Aligned Bases).
Padded target aligned bases	Total aligned bases in the padded target region.

Statistic	Definition
Padded base enrichment	100*(Total Aligned Bases in Padded Targeted Regions/Total Aligned Bases).
Percent duplicate paired reads	Percentage of paired reads that have duplicates.
Mean region coverage depth	The total number of targeted bases divided by the targeted region size.
Uniformity of coverage (Pct > 0.2*mean):	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
Target coverage at 1X	Percentage targets with coverage greater than 1X.
Target coverage at 10X	Percentage targets with coverage greater than 10X.
Target coverage at 20X	Percentage targets with coverage greater than 20X.
Target coverage at 50X	Percentage targets with coverage greater than 50X.
Fragment length median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Fragment length min	Minimum length of the sequenced fragment.
Fragment length max	Maximum length of the sequenced fragment.
Fragment length SD	Standard deviation of the sequenced fragment length.
SNVs, Indels, Insertions, Deletions	Total number of variants present in the data set that pass the quality filters.
SNVs, Indels, Insertions, Deletions (Percent found in dbSNP)	100*(Number of variants in dbSNP/Number of variants).
SNV Ts/Tv ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).
SNVs, Indels, Insertions, Deletions Het/Hom ratio	Number of heterozygous variants/Number of homozygous variants.

Statistic	Definition
SNVs, Insertions, Deletions in Genes	The number of variants that fall into a gene.
SNVs, Insertions, Deletions in Exons	The number of variants that fall into an exon.
SNVs, Insertions, Deletions in Coding Regions	The number of variants that fall into a coding region.
SNVs, Insertions, Deletions in UTR Region	The number of variants that fall into an untranslated region (UTR).
SNVs, Insertions, Deletions in Splice Site Region	The number of variants that fall into a splice site region.
Stop Gained SNVs, Insertions, Deletions	The number of variants that cause an additional stop codon.
Stop Lost SNVs, Insertions, Deletions	The number of variants that cause the loss of a stop codon.
Frameshift SNVs, Insertions, Deletions	The number of variants that cause a frameshift.
Non-synonymous SNVs, Insertions, Deletions	The number of variants that cause an amino acid change in a coding region.
Synonymous SNVs	The number of variants that are within a coding region, but do not cause an amino acid change.

Aggregate Summary Report

The Isaac Enrichment app provides an aggregate enrichment statistics PDF report for all samples combined on the Summary page.

A brief description of the report is below.

NOTE
PCR duplicate reads are not removed from statistics. Results are not directly comparable to Picard HsMetrics.

Sample Information

Defines the sample numbers and names in the report.

Enrichment Summary

Statistic	Definition
Total length of targeted reference	Total length of sequenced bases in the target region.
Padding size	The length of sequence immediately upstream and downstream of the enrichment targets that is included for a padded target.

Read Level Enrichment

Statistic	Definition
Total aligned reads	The total number of reads passing filter that aligned.
Targeted aligned reads	Number of reads that aligned to the target.
Read enrichment	100*(Target aligned reads/Total aligned reads).
Padded target aligned reads	Number of reads that aligned to the padded target.
Padded read enrichment	100*(Padded target aligned reads/Total aligned reads).

Base Level Enrichment

Statistic	Definition
Total aligned bases	Total aligned bases.
Targeted aligned bases	Total aligned bases in the target region.
Base enrichment	100*(Total Aligned Bases in Targeted Regions/Total Aligned Bases).
Padded target aligned bases	Total aligned bases in the padded target region.
Padded base enrichment	100*(Total Aligned Bases in Padded Targeted Regions/Total Aligned Bases).

The Base Enrichment histogram graphs the total aligned bases, padded targeted aligned bases, and padded base enrichment by sample.

SNV Summary

Statistic	Definition
SNVs	Total number of Single Nucleotide Variants present in the data set passing the quality filters.

Statistic	Definition
SNVs (Percent found in dbSNP)	100*(Number of SNVs in dbSNP/Number of SNVs).
SNV Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).
SNV Het/Hom Ratio	Number of heterozygous SNVs/Number of homozygous SNVs.

The SNVs histogram graphs the number of SNVs passing by sample.

Indel Summary

Statistic	Definition
Indels	Total number of indels present in the data set passing the quality filters.
Indels (Percent found in dbSNP)	100*(Number of Indels in dbSNP/Number of Indels).
Indel Het/Hom ratio	Number of heterozygous indels/Number of homozygous indels.

The Indels histogram graphs the number of indels passing by sample.

Coverage Summary

Statistic	Definition
Mean region coverage depth	The total number of targeted bases divided by the targeted region size.
Uniformity of coverage (Pct > 0.2*mean):	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
Target coverage at 1X	Percentage targets with coverage greater than 1X.
Target coverage at 10X	Percentage targets with coverage greater than 10X.
Target coverage at 20X	Percentage targets with coverage greater than 20X.
Target coverage at 50X	Percentage targets with coverage greater than 50X.

In addition, the app provides two graphs:

- A Mean Coverage and Uniformity histogram that plots the mean region coverage depth and uniformity of coverage by sample.
- A Targeted Regions Depth of Coverage histogram that plots the number of targeted sequences by the depth of coverage.

Statistic	Definition
Depth of Sequencing Coverage	The coverage depth of a position in the genome refers to the number of sequenced bases that align to that position.
Number of Targeted Bases Covered	Number of targeted bases that have at least the indicated depth of coverage.

Fragment Length Summary

Statistic	Definition
Fragment length median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Minimum	Minimum length of the sequenced fragment.
Maximum	Maximum length of the sequenced fragment.
Standard Deviation	Standard deviation of the sequenced fragment length.

The Fragment Length Medians histogram graphs the fragment length median by sample.

Duplicates Summary

Statistic	Definition
Percent duplicate paired reads	Percentage of paired reads that have duplicates.

The Percent Duplicate Paired Reads histogram graphs the percent duplicate paired reads by sample.

Isaac Enrichment Methods

This chapter describes the methods that are used in the Isaac Enrichment app.

Isaac Aligner

The Isaac aligner[1] aligns DNA sequencing data, single or paired-end, with read lengths and low error rates using the following steps:

- **Candidate mapping positions**—Identifies the complete set of relevant candidate mapping positions using a 32-mer seed-based search.
- **Mapping selection**—Selects the best mapping among all candidates.
- **Alignment score**—Determines alignment scores for the selected candidates based on a Bayesian model.
- **Alignment output**—Generates final output in a sorted duplicate-marked BAM file and summary file.

1 Come Raczy, Roman Petrovski, Christopher T. Saunders, Ilya Chorny, Semyon Kruglyak, Elliott H. Margulies, Han-Yu Chuang, Morten Källberg, Swathi A. Kumar, Arnold Liao, Kristina M. Little, Michael P. Strömberg and Stephen W. Tanner (2013) Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms. Bioinformatics 29(16):2041-3 bioinformatics.oxfordjournals.org/content/29/16/2041

Candidate Mapping

To align reads, the Isaac aligner first identifies a small but complete set of relevant candidate mapping positions. The Isaac aligner begins with a seed-based search using 32-mers as seeds. After the initial single-seed search, Isaac performs a multiseed search for only those reads that were not mapped unambiguously with a single seed.

Mapping Selection

Following a seed-based search, the Isaac aligner selects the best mapping among all the candidates. For paired-end data sets, all mappings where only one end is aligned (called orphan mappings) trigger a local search to find additional mapping candidates. These candidates (called shadow mappings) are defined through the expected minimum and maximum insert size. After optional trimming of low quality 3' ends and adapter sequences, the possible mapping positions of each fragment are compared. This step takes into account pair-end information (when available), possible gaps using a banded Smith-Waterman gap aligner, and possible shadows. The selection is based on the Smith-Waterman score and on the log-probability of each mapping.

Alignment Scores

The alignment scores of each read pair are based on a Bayesian model, where the probability of each mapping is inferred from the base qualities and the positions of the mismatches. The final mapping quality is the alignment score, truncated to 60 for scores above 60, and possibly corrected to known ambiguities in the reference as flagged in the seeds. Following alignment, reads are sorted. Further analysis is performed to identify duplicates and optionally to realign indels.

Alignment Output

After sorting the reads, the Isaac aligner generates compressed binary alignment output files, called BAM (*.bam) files, using the following process:

- **Marking duplicates**—Detection of duplicates is based on the location and observed length of each fragment. The Isaac aligner identifies and marks duplicates even when they appear on oversized fragments or chimeric fragments. Optical duplicates are already filtered out during RTA processing.
- **Realigning indels**—The Isaac aligner tracks previously detected indels, over a window large enough for the current read length, and applies the known indels to all reads with mismatches.
- **Generating BAM files**—The first step in BAM file generation is creation of the BAM record, which contains all required information except the name of the read. The Isaac aligner reads data from base call (BCL) files that were written during primary analysis on the sequencer to generate the read names. Data are then compressed into blocks of 64 kb or less to create the BAM file.

Isaac Variant Caller

The Isaac Variant Caller (the algorithm is also referred to as Starling2) identifies single nucleotide polymorphisms (SNPs) and small indels using the following steps:

- **Read filtering**—Filters out reads failing quality checks.
- **Indel calling**—Identifies a set of possible indel candidates and realigns all reads overlapping the candidates using a multiple sequence aligner.
- **SNP calling**—Computes the probability of each possible genotype given the aligned read data and a prior distribution of variation in the genome.
- **Indel genotypes**—Calls indel genotypes and assigns probabilities.
- **Variant call output**—Generates output in a VCF file and a compressed genome variant call (gVCF) file. See *VCF Files* on page 18 and *gVCF Files* on page 19 for details.

Indel Candidates

Input reads are filtered by removing any of the following:

- Reads that failed primary analysis quality checks.
- Reads marked as PCR duplicates.
- Paired-end reads not marked as a proper pair.
- Reads with a mapping quality less than 20.

Indel Calling

The variant caller proceeds with candidate indel discovery and generates alternate read alignments based on the candidate indels. As part of the realignment process, the variant caller selects a representative alignment to be used for site genotype calling and depth summarization by the SNP caller.

SNP Calling

The variant caller runs a series of filters on the set of filtered and realigned reads for SNP calling without affecting indel calls. First, any contiguous trailing sequence of N base calls is trimmed from the ends of reads. Using a mismatch density filter, reads having

an unexpectedly high number of disagreements with the reference are masked, as follows:

- The variant caller treats each insertion or deletion as a single mismatch.
- Base calls with more than two mismatches to the reference sequence within 20 bases of the call are ignored.
- If the call occurs within the first or last 20 bases of a read, the mismatch limit is applied to a 41-base window at the corresponding end of the read.
- The mismatch limit is applied to the entire read when the read length is 41 or shorter.

Indel Genotypes

The variant caller filters out all bases marked by the mismatch density filter and any N base calls that remain after the end-trimming step. These filtered base calls are not used for site-genotyping but appear in the filtered base call counts in the variant caller output for each site.

All remaining base calls are used for site-genotyping. The genotyping method heuristically adjusts the joint error probability that is calculated from multiple observations of the same allele on each strand of the genome. This correction accounts for the possibility of error dependencies.

This method treats the highest-quality base call from each allele and strand as an independent observation and leaves the associated base call quality scores unmodified. Quality scores for subsequent base calls for each allele and strand are then adjusted. This adjustment is done to increase the joint error probability of the given allele above the error expected from independent base call observations.

Variant Call Output

After the site and indel genotyping methods are complete, the variant caller applies a final set of heuristic filters to produce the final set of non-filtered calls in the output.

The output in the genome variant call (gVCF) file captures the genotype at each position and the probability that the consensus call differs from reference. This score is expressed as a Phred-scaled quality score.

Picard Metrics

Picard is a suite of tools in Java that work with next-generation sequencing data in BAM format. Isaac Enrichment uses the CalculateHsMetrics tool in Picard to compute a set of Hybrid Selection specific metrics from an aligned SAM or BAM file. If a reference sequence is provided, AT/GC dropout metrics are calculated. GC and mean coverage information for every target can also be computed.

For more information, see: picard.sourceforge.net/command-line-overview.shtml

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 4 Illumina General Contact Information

Illumina Website	www.illumina.com
Email	techsupport@illumina.com

Table 5 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Austria	0800.296575	Netherlands	0800.0223859
Belgium	0800.81102	Norway	800.16836
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000

Safety Data Sheets

Safety data sheets (SDSs) are available on the Illumina website at www.illumina.com/msds.

Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to www.illumina.com/support, select a product, then click **Documentation & Literature**.

Illumina
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com