

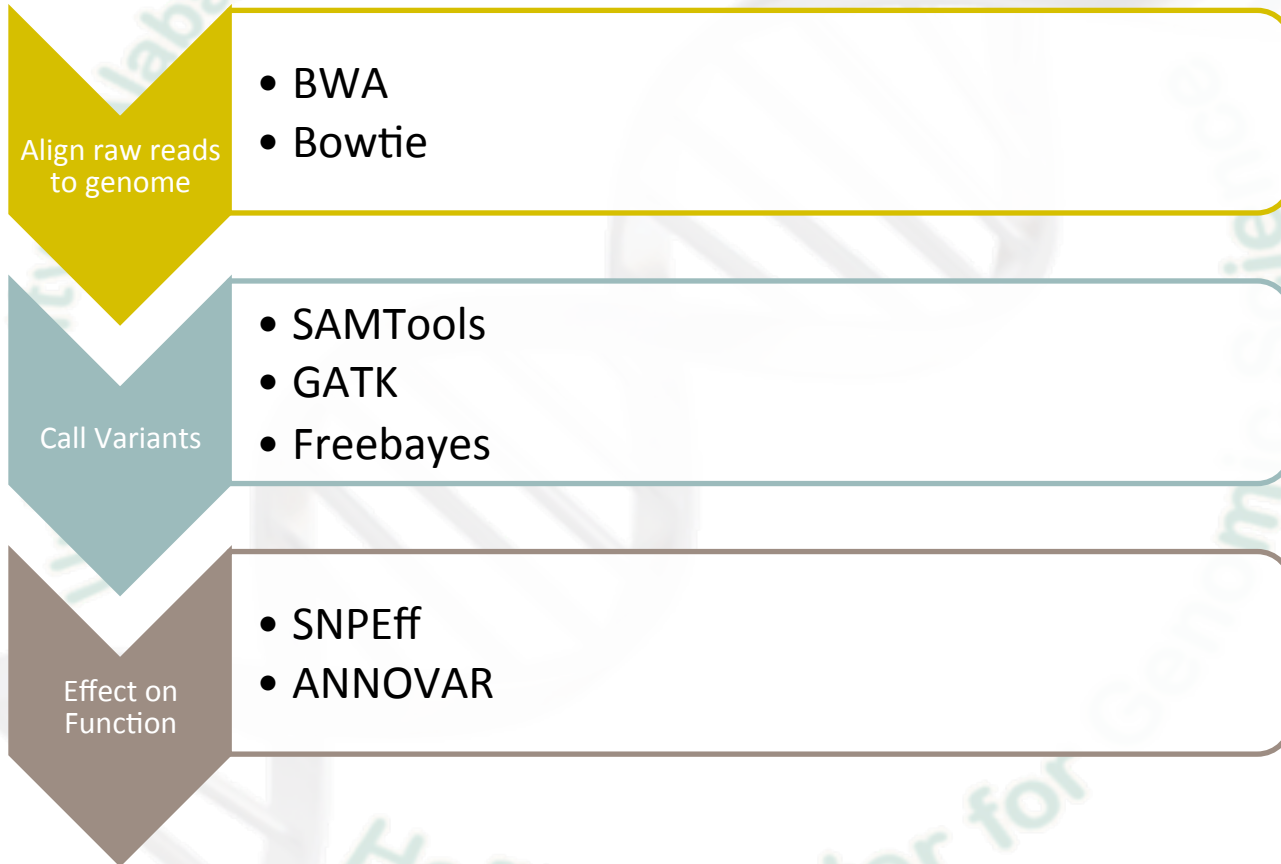


Variant Discovery using DNA-Seq

David Crossman, Ph.D.

UAB Heflin Center for Genomic Science

Whole Genome/Exome (DNA-Seq) analysis pipeline



Upload/Import Data

Tools 1

- Get Data 2**
 - Upload File from your computer
 - UCSC Main table browser
 - UCSC Test table browser
 - UCSC Archaea table browser
 - BX main browser
 - Get Microbial Data
 - BioMart Central server
 - BioMart Test server
 - CBI Rice Mart rice mart
 - GrameneMart Central server
 - modENCODE fly server
 - Flymine server
 - Flymine test server
 - modENCODE modMine server
 - Ratmine server
 - YeastMine server
 - metabolicMine server
 - modENCODE worm server
 - WormBase server
 - Wormbase test server
 - EuPathDB server
 - EncodeDB at NHGRI
 - EpiGRAPH server
 - EpiGRAPH test server
 - HbVar Human Hemoglobin Variants and Thalassemias

Upload File (version 1.1.3)

File Format:

Auto-detect

3a

Which format? See help below

File:

No file chosen

3b-1

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

3b-2

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input type="checkbox"/> MF2_R1.fastqsanger	33.2 Mb	07/19/2012 07:26:42 AM
<input type="checkbox"/> MF2_R2.fastqsanger	33.2 Mb	07/19/2012 07:26:45 AM
<input type="checkbox"/> MF3_R1.fastqsanger	17.1 Mb	07/19/2012 07:26:47 AM
<input type="checkbox"/> MF3_R2.fastqsanger	17.1 Mb	07/19/2012 07:26:48 AM
<input type="checkbox"/> Treeshrew67 GeneScaffold_800_4487.gtf	17.3 Kb	07/19/2012 07:26:48 AM
<input type="checkbox"/> GeneScaffold_800_4487.fasta	251.2 Kb	07/19/2012 07:26:48 AM

3b-3

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at galaxy.uabgrid.uab.edu using your Galaxy credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

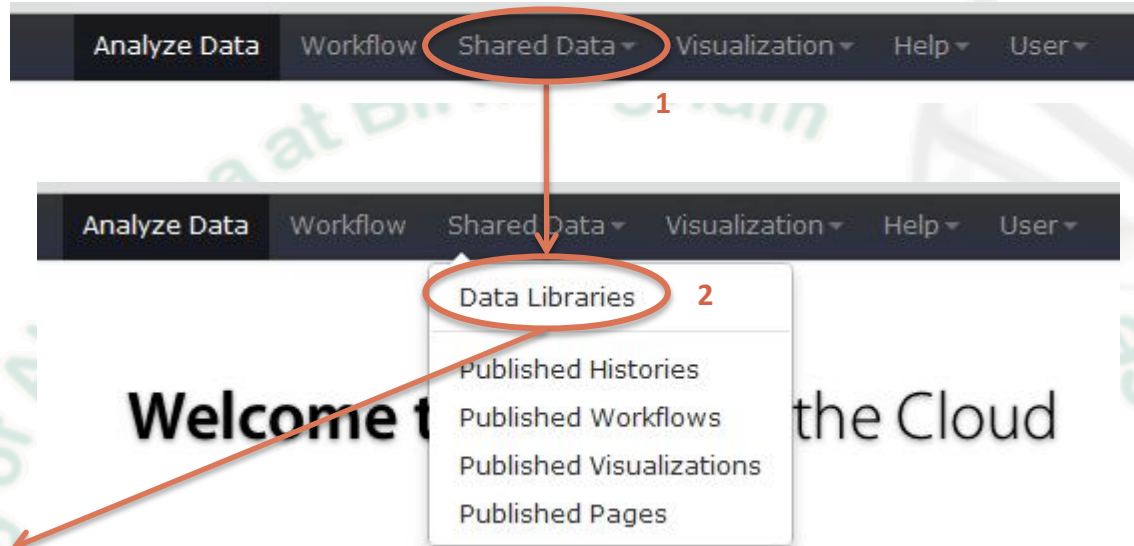
Click to Search or Select

3c

3d

1. Click "Get Data"
2. Click "Upload File"
3. Boxes to be aware of:
 - a) File Format
 - b) File to be uploaded:
 - 1) File from computer
 - 2) URL/text
 - 3) FTP
 - c) Genome
4. Click "Execute"

Shared Data



4 Data Library "GBS720"

human mitochondria DNA-Seq example

<input type="checkbox"/> Name	4a	Message	Data type	Date uploaded	File size
<input type="checkbox"/>	Proband_chr21_1.fastq	Consists of only chromosome 21.	fastqsanger	2015-01-08	142.3 MB
<input type="checkbox"/>	Proband_chr21_2.fastq	Consists of only chromosome 21.	fastqsanger	2015-01-08	142.3 MB

For selected datasets: 4b

1. Click on "Shared Data" (located on top toolbar)
2. Drop down box appears; click on "Data Libraries"
3. Under "Data Library Name" look for "GBS720." Click on it.
4. Will see this Data Library.
 - a) Put checkmark besides Name
 - b) Click Go

Quality Control of raw fastq reads

Tools

- NGS: QC and manipulation **1**
- FASTQC: FASTQ/SAM/BAM
- Fastqc: Fastqc QC using FastQC from Babraham **2**
- ILLUMINA FASTQ
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column
- ROCHE-454 DATA
- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

3a Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:

2: Proband_chr21_2.fastq

Title for the output file - to remind you what the job was for:

FastQC

Contaminant list:

Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

3b Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:

1: Proband_chr21_1.fastq *

Title for the output file - to remind you what the job was for:

Proband R1 FastQC *

Contaminant list:

Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

4

1. Click on "NGS: QC and manipulation"
2. Click on "Fastqc: Fastqc QC"
3. Select options:
 - a) This is what the window looks like when first opened
 - b) Choose fastq file and give it a useful name
4. Click "Execute"
5. Do the exact same thing for the other fastq file

FastQC Output Report

History [refresh] [settings]

GBS720 test
285.1 MB [edit] [delete]

4: Proband R2 FastQC data 2.html [eye] [edit] [delete]

3: Proband R1 FastQC data 1.html [eye] [edit] [delete]

2: Proband chr21 2.fastq [eye] [edit] [delete]

1: Proband chr21 1.fastq [eye] [edit] [delete]

Proband_chr21_1.fastq FastQC Report
FastQC Report
Thu 8 Jan 2015
Proband_chr21_1.fastq

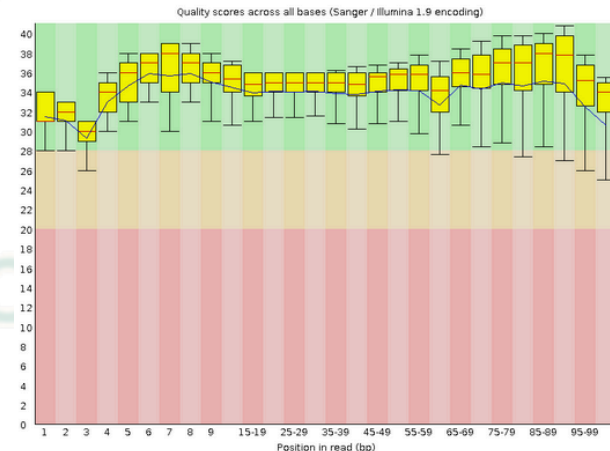
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ⚠ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ⚠ Kmer Content

Basic Statistics

Measure	Value
Filename	Proband_chr21_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	595614
Filtered Sequences	0
Sequence length	30-101
%GC	48

Per base sequence quality




If you think the data needs to be trimmed, then you can use the FastX-Toolkit under “NGS: QC and manipulation.”

BWA aligner

Map with BWA for Illumina (version 1.2.3)


Will you select a reference genome from your history or use a built-in index?:

Use a built-in index 


Select a reference genome:

Human (Homo sapiens): hg19 Full  **3a**

Is this library mate-paired?:


Paired-end  **3b**

Forward FASTQ file:

1: Proband_chr21_1.fastq  **3c**

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

Reverse FASTQ file:

2: Proband_chr21_2.fastq  **3d**

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:

Commonly Used  *

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

Suppress the header in the output SAM file:

BWA produces SAM with several lines of header information

Execute **4**

NGS: Mapping **1**

- [Lastz paired reads](#) map short paired reads against reference sequence
- [Lastz](#) map short reads against reference sequence
- [Map with Bowtie for SOLiD](#)
- [Map with Bowtie for Illumina](#) **2**
- [Map with BWA for Illumina](#)
- [Map with BWA for SOLiD](#)
- [Megablast](#) compare short reads against htgs, nt, and wgs databases
- [Parse blast XML output](#)
- [Map with PerM](#) for SOLiD and Illumina
- [Re-align with SRMA](#)
- [Map with Mosaik](#)

1. Click on “NGS: Mapping”
2. Click on “Map with BWA for Illumina”
3. Select options:
 - a) Choose built-in index “hg19 Full”
 - b) Choose Paired-end
 - c) Choose the F fastq file (has _1 in filename)
 - d) Choose the R fastq file (has _2 in filename)
4. Click “Execute”

* Will use Commonly Used setting right now, but you may need to go back and modify these in future experiments

flagstat on original alignment

NGS: SAM Tools

1

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Generate VCF with mpileup piped through bcftools view from BAM dataset(s)
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

flagstat (version 1.0.0)

BAM File to Convert: 3

13: (as bam) father Map with BWA for Illumina on data 2 and data 1: mapped reads

Execute

4

Flagstat is used to calculate stats on an alignment file

1. Click on “NGS: SAM Tools”
2. Click on “flagstat”
3. Select SAM/BAM file.
4. Click “Execute.”

2

Filter SAM

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Generate VCF with mpileup piped through bcftools view from BAM dataset(s)
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

Filter SAM (version 1.0.0)

Select dataset to filter:

5: Map with BWA for Illumina on data 2 and data 1: mapped reads

Flags

Flag 1 3c

Type:

Read is paired

Set the states for this flag:

- No
 Yes

Remove Flag 1

Flag 2 3e

Type:

Read is mapped in a proper pair

Set the states for this flag:

- No
 Yes

Remove Flag 2

Add new Flag

Execute 4

- Click on "NGS: SAM Tools"
- Click on "Filter SAM"
- Select options:
 - Select dataset to filter
 - Click "Add New Flag"
 - Flag 1: Read is paired – Yes
 - Click "Add New Flag"
 - Flag 2: Read is mapped in a proper pair – Yes
- Click "Execute"

Picard Tools (remove duplicates)

ERROR when running this tool. Will need to fix it, but it is important to remove duplicates. The problem is that the tool is removing the known locations and giving them assignments that they didn't align in the first place.

WORKAROUND for now is to just skip this tool for demo purposes. ³

NGS: Picard (beta) ¹

- FASTQ to BAM creates an unaligned BAM file
- SAM to FASTQ creates a FASTQ file
- BAM Index Statistics
- SAM/BAM Alignment Summary Metrics
- SAM/BAM GC Bias Metrics
- Estimate Library Complexity
- Insertion size metrics for PAIRED data
- SAM/BAM Hybrid Selection Metrics for targeted resequencing data
- Add or Replace Groups
- Reorder SAM/BAM
- Replace SAM/BAM Header
- Paired Read Mate Fixer for paired data
- Mark Duplicate reads ²

Mark Duplicate reads (version 1.56.0)

SAM/BAM dataset to mark duplicates in:

11: Child Filter SAM on data 9

If empty, upload or import a SAM/BAM dataset.

Title for the output file:

Child Dupes Marked ⁴

Use this remind you what the job was for

Remove duplicates from output file:



If true do not write duplicates to the output file instead of writing them with appropriate flags set.

Assume reads are already ordered:



If true assume input data are already sorted (most Galaxy SAM/BAM should be).

Regular expression that can be used to parse read names in the incoming SAM file:

[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*

Names are parsed to extract: tile/region, x coordinate and y coordinate, to estimate optical duplication rate

The maximum offset between two duplicate clusters in order to consider them optical duplicates.:

100

e.g. 5-10 pixels. Later Illumina software versions multiply pixel values by 10, in which case 50-100.

Execute ⁵

1. Click on "NGS: Picard (beta)"
2. Click on "Mark Duplicate reads"
3. Select dataset to mark duplicates
4. Give file a name
5. Click "Execute"

SAM-to-BAM

NGS: SAM Tools ¹

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format ²
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Generate VCF with mpileup piped through bcftools view from BAM dataset(s)
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

SAM-to-BAM (version 1.1.2)

Choose the source for the reference list:

Locally cached 

SAM File to Convert: **3a**

6: Filter SAM on data 5 

Execute ⁴

1. Click on "NGS: SAM Tools"
2. Click on "SAM-to-BAM"
3. Select options:
 - a) Select SAM file to convert
4. Click "Execute"

Picard Tools (add groups)

ERROR: Tool doesn't work correctly. Ignore for now.

NGS: Picard (beta) 1

- [FASTQ to BAM](#) creates an unaligned BAM file
- [SAM to FASTQ](#) creates a FASTQ file
- [BAM Index Statistics](#)
- [SAM/BAM Alignment Summary Metrics](#)
- [SAM/BAM GC Bias Metrics](#)
- [Estimate Library Complexity](#)
- [Insertion size metrics](#) for PAIRED data
- [SAM/BAM Hybrid Selection Metrics](#) for targeted resequencing data
- [Add or Replace Groups](#) 2
- [Reorder SAM/BAM](#)
- [Replace SAM/BAM Header](#)
- [Paired Read Mate Fixer](#) for paired data
- [Mark Duplicate reads](#)

Add or Replace Groups (version 1.56.0)

SAM/BAM dataset to add or replace read groups in:

13: Child SAM-to-BAM on data 11: converted BAM 3

If empty, upload or import a SAM/BAM dataset.

Read group ID (ID tag):

child 4a

The most important read group tag. Galaxy will use a value of '1' if nothing provided.

Read group sample name (SM tag):

child 4b

Read group library (LB tag):

child 4c

Read group platform (PL tag):

illumina

illumina, solid, 454, pacbio, helicos

Read group platform unit:

bc 4d

like run barcode, etc.

Specify additional (optional) arguments:

Use pre-set defaults

Allows you to set RGCN and RGDS.

Output bam instead of sam:

Uncheck for sam output

Execute 5

1. Click on "NGS: Picard (beta)"
2. Click on "Add or Replace Groups"
3. Select BAM file to add groups
4. Provide names for these various read groups
 - a) ID
 - b) Sample Name
 - c) Library
 - d) Platform
 - e) Platform Unit
5. Click "Execute"

flagstat on filtered alignment

NGS: SAM Tools

1

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Generate VCF with mpileup piped through bcftools view from BAM dataset(s)
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

flagstat (version 1.0.0)

BAM File to Convert: 3

13: (as bam) father Map with BWA for Illumina on data 2 and data 1: mapped reads

Execute

4

Flagstat is used to calculate stats on an alignment file

1. Click on “NGS: SAM Tools”
2. Click on “flagstat”
3. Select SAM/BAM file.
4. Click “Execute.”

Call Variants with SAM Tools

NGS: SAM Tools

1

- [Filter SAM](#) on bitwise flag values
- [Convert SAM](#) to interval
- [SAM-to-BAM](#) converts SAM format to BAM format
- [BAM-to-SAM](#) converts BAM format to SAM format
- [Merge BAM Files](#) merges BAM files together
- [Generate pileup](#) from BAM dataset
- [Generate VCF with mpileup piped through bcftools view](#) from BAM dataset(s)
- [Filter pileup](#) on coverage and SNPs
- [Pileup-to-Interval](#) condenses pileup format into ranges of bases
- [flagstat](#) provides simple stats on BAM files

2

Generate VCF with mpileup piped through bcftools view (version 0.1)

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index 3a

Select the BAM file to generate the pileup file for:

7: SAM-to-BAM on data 6: converted BAM 3b

Output format:

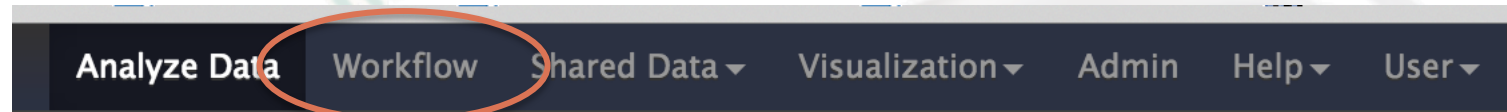
VCF 3c

Pileup output is provided for backward compatibility

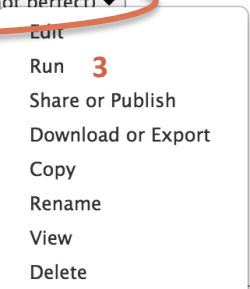
Execute 4

1. Click on "NGS: SAM Tools"
2. Click on "Generate VCF with mpileup piped through bcftools view"
3. Select options:
 - a) Choose "Use a built-in index"
 - b) Select BAMoutput file from previous slide
 - c) Keep output format as VCF
4. Click "Execute"

Common and de novo variants (not a perfect workflow)

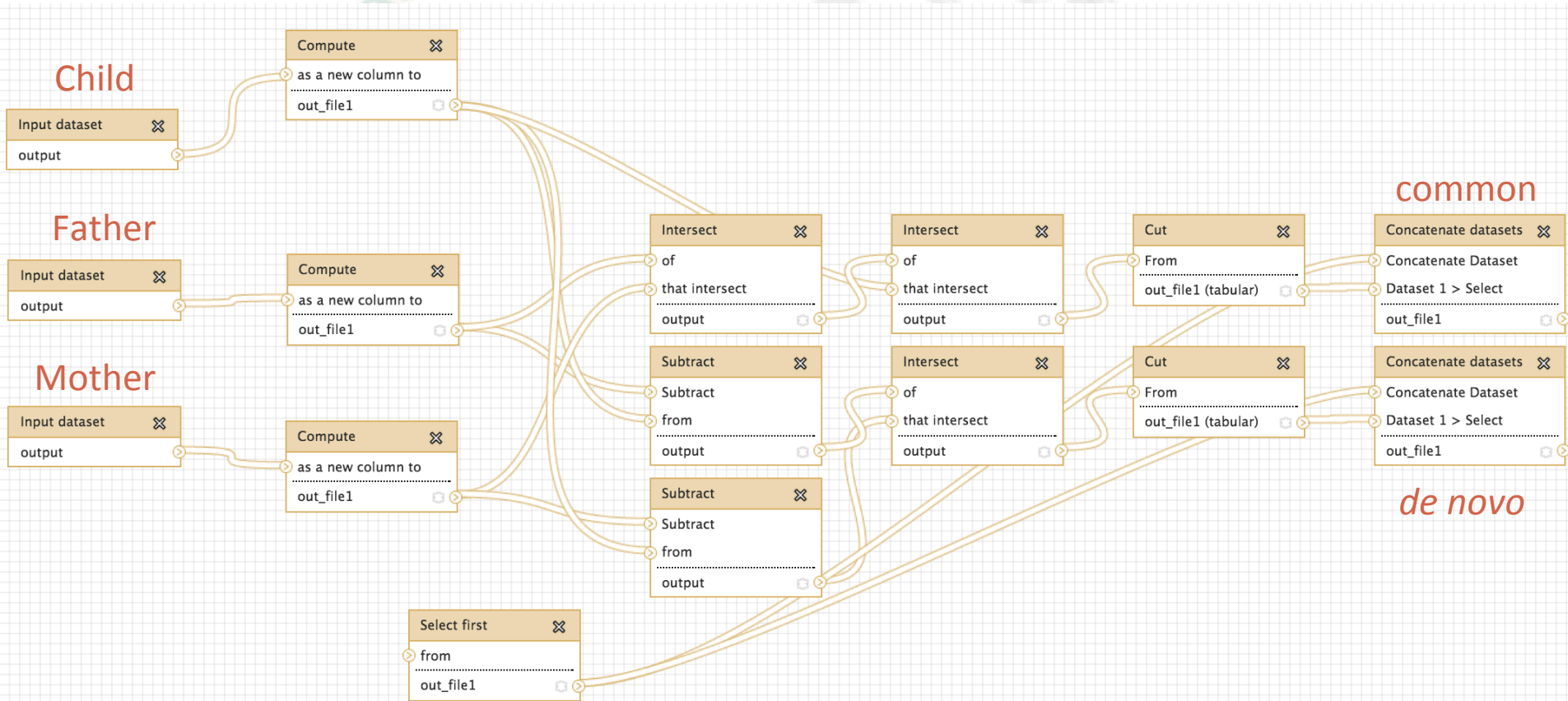


Name	# of Steps
Identifying common and de novo variants (not perfect) ▾	16
Contamination: SAM to taxonomy tree ▾	13
Contamination: TopHat BAM to taxonomy tree ▾	13
Workflow constructed from history 'RNA-Seq' ▾	19
Compute exon length with stats per chr ▾	3
Immersion course ▾	9
imported: GBS722 VACV-WR 125k - sample aenovo with BWASW to ref in hist ▾	8



1. Click on "Workflow"
2. Click on down-arrow of "Identifying common and de novo variants (not perfect)"
3. Click "Run"

Identifying common and de novo variants (not perfect) workflow



Common and de novo variants (not a perfect workflow)

Running workflow "Identifying common and de novo variants (not perfect)"

Step 1: Input dataset

Child vcf file

Input Child vcf Dataset 1

33: child Generate VCF with mpileup piped through bcftools view on data 24: mpileup to vcf
type to filter

Step 2: Input dataset

Father vcf

Input Father vcf Dataset 2

31: father Generate VCF with mpileup piped through bcftools view on data 22: mpileup to vcf
type to filter

Step 3: Input dataset

Mother vcf

Input Mother vcf Dataset 3

32: mother Generate VCF with mpileup piped through bcftools view on data 23: mpileup to vcf
type to filter

Step 4: Select first (version 1.0.0)

obtain vcf header lines

Select first




2




from

50: de novo SnpEff on data 46

1. Choose child vcf file
2. Choose father vcf file
3. Choose mother vcf file
4. Click "Run workflow" (at bottom of screen)

Common and de novo variants (not a perfect workflow)

46: de novo variants in child   
not in father or mother

45: common variants btw   
child-father-mother

These are the two final output files from this particular workflow

SNPEff (effect on function of variants)

SnpEff tools

1

- 2 [SnpEff Variant effect and annotation](#)
- [SnpEff Download](#) Download a new database
- [SnpSift Annotate](#) Annotate SNPs from dbSnp
- [SnpSift CaseControl](#) Count samples are in 'case' and 'control' groups.
- [SnpSift Filter](#) Filter variants using arbitrary expressions
- [SnpSift Intervals](#) Filter variants using intervals

SnpEff (version 1.0)

Sequence changes (SNPs, MNPs, InDels):

8: Generate VCF with mpileup piped through bcftools view on data 7: mp

3a

Input format:

VCF *

Output format:

Tabular *

Genome:

Homo_sapiens (hg19)

3b

Upstream / Downstream length:

5000 bases

Filter homozygous / heterozygous changes: *

- No filter (analyze everything)
- Analyze homozygous sequence changes only
- Analyze heterozygous sequence changes only

Filter sequence changes: *

- No filter (analyze everything)
- Analyze deletions only
- Analyze insertions only
- Only MNPs (multiple nucleotide polymorphisms)
- Only SNPs (single nucleotide polymorphisms)

Filter output: *

Select All Unselect All

- None
- Do not show DOWNSTREAM changes
- Do not show INTERGENIC changes
- Do not show INTRON changes
- Do not show UPSTREAM changes
- Do not show 5_PRIME_UTR or 3_PRIME_UTR changes

Chromosomal position: *

- Use default (based on input type)
- Force zero-based positions (both input and output)
- Force one-based positions (both input and output)

Execute

4

1. Click on "SnpEff tools"
2. Click on "SnpEff"
3. Select options:
 - a) Choose VCF output file from previous slide
 - b) Pick genome "Homo_sapiens (hg19)"
4. Click "Execute"

* Other options to be aware of!

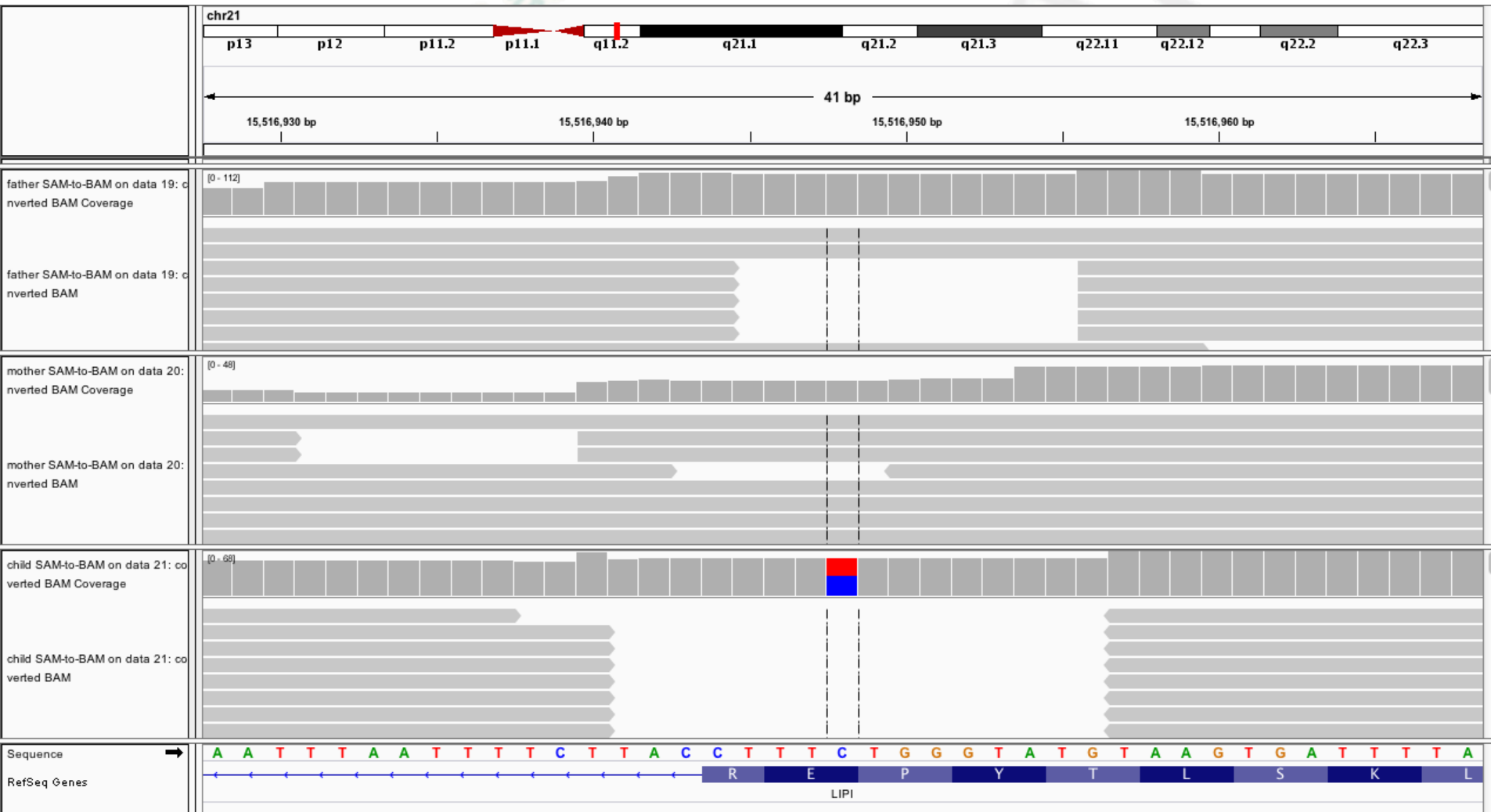
SNPEff output (cropped)

Chr	Pos	REF	ALT	Change_type	Homozygous	Quality	Cov.	Gene name	Transcript_ID	Effect	old_A/new_A	Old_codon/New_codon	Codon_Num(CDS)	Codon_Degeneracy	CDS_size
21	10910311	T	G	SNP	Hom	99	737	TPTE	NM_199259	NON_SYNONYMOUS_CODING	Y/S	tAt/tCt	464	1	1602
21	10970008	C	T	SNP	Hom	4.13	286	TPTE	NM_199259	SPLICE_SITE_DONOR					1602
21	11058226	G	C	SNP	Hom	21	882	BAGE2	NM_182482	NON_SYNONYMOUS_CODING	P/A	Cct/Gct	72	1	330
21	15481365	G	T	SNP	Hom	99	171	LIPI	NM_198996	NON_SYNONYMOUS_CODING	D/E	gaC/gaA	465	2	1446
21	15596772	T	G	SNP	Hom	99	63	RBM11	NM_144770	NON_SYNONYMOUS_CODING	L/V	Ttg/Gtg	116	2	846
21	15954528	G	A	SNP	Hom	99	405	SAMSN1	NM_001256370	NON_SYNONYMOUS_CODING	H/Y	Cac/Tac	64	1	1326
21	30339120	C	A	SNP	Hom	99	111	LTN1	NM_015565	NON_SYNONYMOUS_CODING	G/C	Ggc/Tgc	611	1	5439
21	31744127	A	T	SNP	Hom	99	110	KRTAP13-2	NM_181621	STOP_GAINED	C/*	tgT/tgA	135	2	528
21	34948686	*	+A	INS	Hom	99	74	SON	NM_138927	FRAME_SHIFT	-/?	-/A	2413		7281

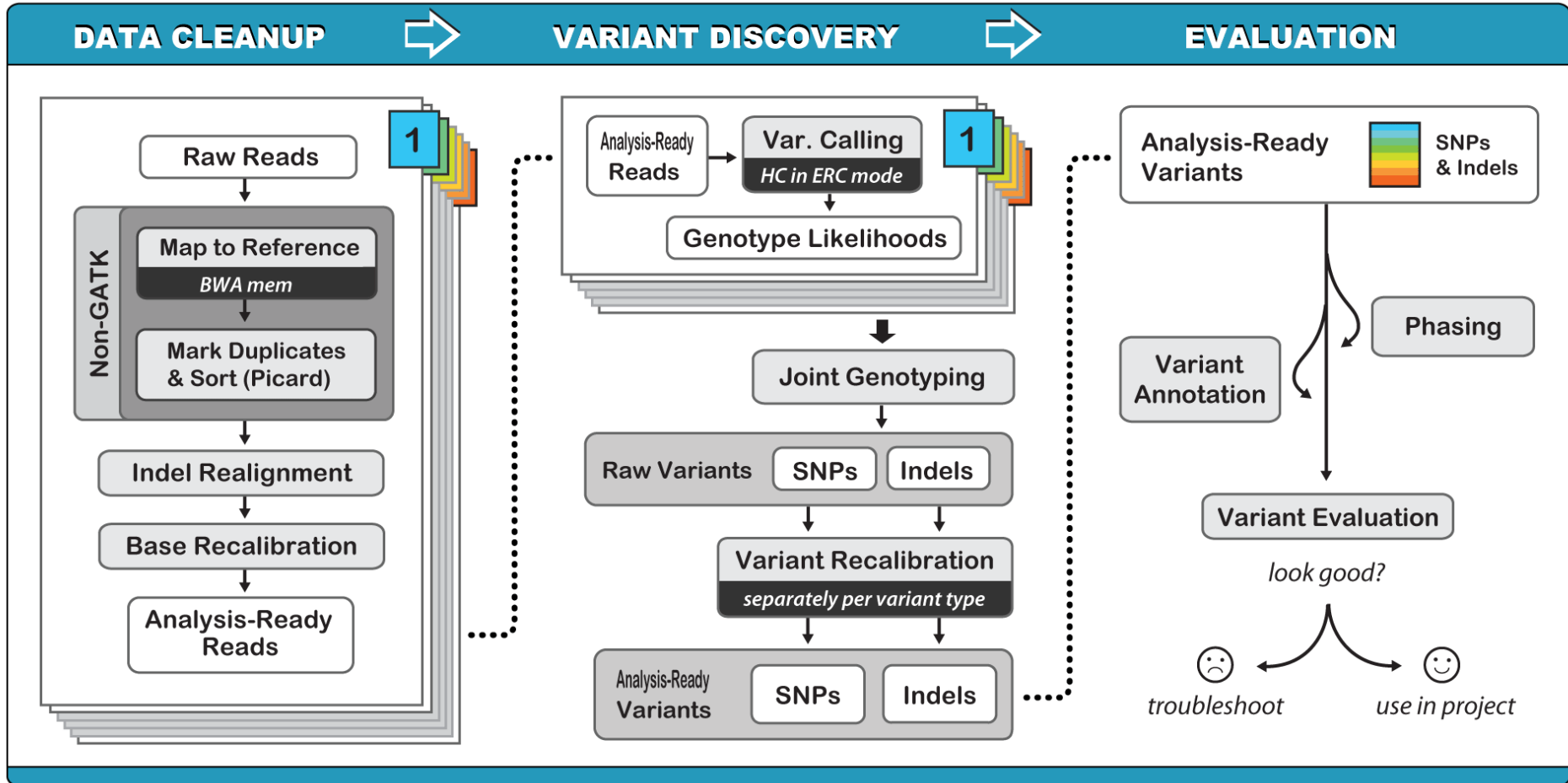
Tool	Link
CADD	http://cadd.gs.washington.edu/
Broad ExAC	http://exac.broadinstitute.org/
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation138/
Ensembl VEP	http://useast.ensembl.org/info/docs/tools/vep/index.html
MutationAssessor	http://mutationassessor.org/
MutationTaster	http://www.mutationtaster.org/
OMIM	http://www.ncbi.nlm.nih.gov/omim
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
GeneCards	http://www.genecards.org/
Wellcome Trust Sanger Institute Mouse Genomes Project	http://www.sanger.ac.uk/resources/mouse/genomes/

IGV

at Birmingham



GATK pipeline



GATK Best Practices

(<http://www.broadinstitute.org/gatk/>)

Best Practice Variant Detection with the GATK v4, for release 2.0

There are 18 comments on this article. To see them or add your own, read this post on the forum →

Introduction

1. The basic workflow

Our current best practice for making SNP and indel calls is divided into four sequential steps: initial mapping, refinement of the initial reads, multi-sample indel and SNP calling, and finally variant quality score recalibration. These steps are the same for targeted resequencing, whole exomes, deep whole genomes, and low-pass whole genomes. Example commands for each tool are available on the individual tool's wiki entry. [There is also a list of which resource files to use with which tool.](#)

Note that due to the specific attributes of a project the specific values used in each of the commands may need to be selected/modified by the analyst. Care should be taken by the analyst running our tools to understand what each parameter does and to evaluate which value best fits the data and project design.

2. Lane, Library, Sample, Cohort

There are four major organizational units for next-generation DNA sequencing processes that used throughout this documentation:

- **Lane:** The basic machine unit for sequencing. The lane reflects the basic independent run of an NGS machine. For Illumina machines, this is the physical sequencing lane.
- **Library:** A unit of DNA preparation that at some point is physically pooled together. Multiple lanes can be run from aliquots from the same library. The DNA library and its preparation is the natural unit that is being sequenced. For example, if the library has limited complexity, then many sequences are duplicated and will result in a high duplication rate across lanes.
- **Sample:** A single individual, such as human CEPH NA12878. Multiple libraries with different properties can be constructed from the original sample DNA source. Here we treat samples as independent individuals whose genome sequence we are attempting to determine. From this perspective, tumor / normal samples are different despite coming from the same individual.
- **Cohort:** A collection of samples being analyzed together. This organizational unit is the most subjective and depends intimately on the design goals of the sequencing project. For population discovery projects like the 1000 Genomes, the analysis cohort is the ~100 individual in each population. For exome projects with many samples (e.g., ESP with 800 EOMI samples) deeply sequenced we divide up the complete set of samples into cohorts of ~50 individuals for multi-sample analyses.

This document describes how to call variation within a single analysis cohort, comprised for one or many samples, each of one or many libraries that were sequenced on at least one lane of an NGS machine.

Note that many GATK commands can be run at the lane level, but will give better results seeing all of the data for a single sample, or even all of

GATK (beta) on PSU Galaxy

Basic Steps* (options are up to you):

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- [Depth of Coverage](#) on BAM files

- 6 ▪ [Print Reads](#) from BAM files

REALIGNMENT

- 3 ▪ [Realigner Target Creator](#) for use in local realignment

- 4 ▪ [Indel Realigner](#) - perform local realignment

BASE RECALIBRATION

- 5 ▪ [Count Covariates](#) on BAM files

- [Table Recalibration](#) on BAM files

- [Analyze Covariates](#) - draw plots

7

GENOTYPING

- [Unified Genotyper](#) SNP and indel caller

ANNOTATION

- [Variant Annotator](#)

FILTRATION

- [Variant Filtration](#) on VCF files

11

- [Select Variants](#) from VCF files

VARIANT QUALITY SCORE RECALIBRATION

8

- [Variant Recalibrator](#)

9

- [Apply Variant Recalibration](#)

VARIANT UTILITIES

- [Validate Variants](#)

- [Eval Variants](#)

10

- [Combine Variants](#)

1. BWA alignment
2. Mark duplicates (Picard)
3. Realigner Target Creator
4. Indel Realigner
5. Base Recalibrator (Count Covariates)
6. Print Reads
7. Unified Genotyper (new in Ver2 is Haplotype Caller) (SNPs and Indels done separately)
8. Variant Recalibrator (SNPs and Indels done separately)
9. Apply Recalibration (SNPs and Indels done separately)
10. Combine Variants
11. Select Variants
12. Compare/contrast variants
13. snpEFF

* This follows the **basic** pipeline shown 2 slides ago. Each project is different and may need additional tools to answer the biological question(s). Also, options for each tool will vary as well.

Freebayes

Human Genome Variation 1

- FreeBayes - Bayesian genetic variant detector 2
- SIFT predictions of functional sites
- g:Profiler tools for functional profiling of gene lists
- DAVID functional annotation for a list of genes
- CTD analysis of chemicals, diseases, or genes
- snpFreq significant SNPs in case-control data
- LD linkage disequilibrium and tag SNPs
- PASS significant transcription factor binding sites from CHIP data
- GPASS significant single-SNP associations in case-control studies
- BEAM significant single- and multi-locus SNP associations in case-control studies
- LPS LASSO-Patternsearch algorithm
- HVIS visualization of genomic data with the Hilbert curve

FreeBayes (version 0.0.3)

Choose the source for the reference list:

Locally cached ▾

Sample BAM files

Sample BAM file 1

BAM file:

5: (as bam) Map with BWA for Illumina on data 2 and data 1: mapped reads ▾

3a

Add new Sample BAM file

Using reference genome:

hg19 ▾

3b

Basic or Advanced options:

Basic ▾ *

Execute 4

- Click on "Human Genome Variation"
 - Click on "FreeBayes"
 - Select options:
 - Choose BWA output file
 - Pick genome "hg19"
 - Click "Execute"
- * Other options to be aware of!

References and web links

- Galaxy
 - PSU Public website: <https://usegalaxy.org/>
 - UAB: <https://www.uab.edu/galaxy>
- [Bowtie](#)
- [GATK](#)
- [SAMTools](#)
- [Picard Tools](#)
- [FreeBayes](#)
- [IGV](#)
- [SNPEff](#)
- [CADD](#)
- [Broad Exac](#)
- [dbSNP](#)
- [OMIM](#)
- [ClinVar](#)

University of Alabama at Birmingham

Heflin Center for Genomic Science

Thanks! Questions?

Contact info:

David K. Crossman, Ph.D.

Bioinformatics Director

Heflin Center for Genomic Science

University of Alabama at Birmingham

<http://www.heflingenetics.uab.edu>

dkcrossm@uab.edu