

# **Exome sequencing data analysis for diagnosing a genetic disease**

Galaxy Training! tutorial

# Tutorial presentation

- Exome sequencing data from a family trio
- Boy child affected by a disease : osteopetrosis
- Parents unaffected but consanguineous

**Goal : Identify the genetic variation responsible for the disease**

# Tutorial steps

1. Perform postprocessing from premapped reads
2. Variant calling
3. Variant annotation and reporting

# Tutorial steps

1. Perform postprocessing from premapped reads
2. Variant calling
3. Variant annotation and reporting

# Premapped reads

- Data characteristics for the trio :
  - Whole exome sequencing
  - Paired-end reads
- Steps already performed :
  - Quality control (fastq)
  - Read mapping (Human Hg19 assembly)
- Format available : bam format

# Premapped reads upload

The image shows the Galaxy France web interface. At the top left, the logo 'Galaxy France' is displayed. The top navigation bar includes links for 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and icons for a graduation cap, a bell, and a grid. A red box labeled '1' highlights the 'Shared Data' dropdown menu. This menu is open, showing a list of options: 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. A red box labeled '2' highlights the 'Data Libraries' option. Below the navigation bar, a light blue banner contains a maintenance notice: 'From the 4th to 7th of April, usegalaxy.fr will be shut down for maintenance'. On the left side, there is a 'Tools' panel with a search bar, an 'Upload Data' button, and links for 'Get Data', 'Send Data', and 'Collection Operations'. The main content area features a graphic of three vertical bars (blue, yellow, and red) and the text 'Welcome to usegalaxy.fr'.

# Premapped reads upload

   exclude restricted

Name	Description	Synopsis
ProteoRE	ProteoRE datasets	
covid-19		
<b>GTN - Material</b>	Galaxy Training Network Material	Galaxy Training Network Material. See ht ... <small>(more)</small>
workflow4metabolomics	Workflow4Metabolomics referenced histori ... <small>(more)</small>	https://workflow4metabolomics.org/refere ... <small>(more)</small>
Roscoff 2021	Data for Assembly and Annotation trainin ... <small>(more)</small>	









# Premapped reads upload

Libraries / GTN - Material

<input type="checkbox"/>	Name	Description
<input type="checkbox"/>	Assembly	DNA sequence data has become an indispen ... (more)
<input type="checkbox"/>	ChIP-Seq data analysis	ChIP-sequencing is a method used to anal ... (more)
<input type="checkbox"/>	Ecology	Learn to analyse Ecological data through ... (more)
<input type="checkbox"/>	Epigenetics	DNA methylation is an epigenetic mechani ... (more)
<input type="checkbox"/>	Genome Annotation	Genome annotation is a multi-level proce ... (more)
<input type="checkbox"/>	Imaging	Image analysis using Galaxy tools
<input type="checkbox"/>	Introduction to Galaxy Analyses	Galaxy is a scientific workflow, data in ... (more)
<input type="checkbox"/>	Metabolomics	Training material to analyse Mass spectr ... (more)
<input type="checkbox"/>	Metagenomics	Metagenomics is a discipline that enable ... (more)
<input type="checkbox"/>	New topic	Topic summary



# Premapped reads upload

	<input type="checkbox"/>	PAPAA PI3K_OG:Pancancer Aberrant Pathway Activity Analysis	Summary
	<input type="checkbox"/>	Proteomics	Training material for proteomics workflow ... (more)
	<input type="checkbox"/>	Refining Manual Genome Annotations with Apollo	We look at how to edit Genome Annotation ... (more)
	<input type="checkbox"/>	RNA interactome	RNA interactome data analysis
	<input type="checkbox"/>	Sequence analysis	Analyses of sequences
	<input type="checkbox"/>	Statistics and machine learning	Statistical Analyses for omics data and ... (more)
	<input type="checkbox"/>	The new topic	Summary
	<input type="checkbox"/>	Transcriptomics	Training material for all kinds of trans ... (more)
	<input type="checkbox"/>	User Interface and Features	A collection of microtutorials explainin ... (more)
	<input type="checkbox"/>	Variant Analysis	Exome sequencing means that all protein- ... (more)

# Premapped reads upload

Libraries / GTN - Material / Variant Analysis

<input type="checkbox"/>	Name	Description
<input type="checkbox"/>	Calling variants in diploid systems	
<input type="checkbox"/>	Calling variants in non-diploid systems	
<input type="checkbox"/>	DOI: 10.5281/zenodo.3960260	latest
<input type="checkbox"/>	Exome sequencing data analysis for diagnosing a genetic disease	
<input type="checkbox"/>	Identification of somatic and germline variants from tumor and normal sample pairs	
<input type="checkbox"/>	Mapping and molecular identification of phenotype-causing mutations	
<input type="checkbox"/>	Microbial Variant Calling	
<input type="checkbox"/>	Mutation calling, viral genome reconstruction and lineage/clade assignment from SARS-CoV-2 sequencing data	

# Premapped reads upload

Libraries / GTN - Material / Variant Analysis / Exome sequencing data analysis for diagnosing a genetic disease

<input type="checkbox"/>	Name	Description
<input type="checkbox"/>	DOI: 10.5281/zenodo.3054169	latest

# Premapped reads upload

Search **1** **Export to History** **Download** **Delete** **Details**  include deleted

Libraries / GTN - Material / Vari **as Datasets** **2** **as a Collection** **2** **Genomic data analysis for diagnosing a genetic disease / DOI: 10.5281/zenodo.3054169**

<input type="checkbox"/>	Name	Description	Type	Size
<input type="checkbox"/>	<a href="https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/hg19_chr8.fa.gz">https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/hg19_chr8.fa.gz</a>	uploaded fasta file	fasta	142.4 MB
<input checked="" type="checkbox"/>	<a href="https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam">https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam</a>	uploaded bam file	bam	336.9 MB
<input checked="" type="checkbox"/>	<a href="https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_mother.bam">https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_mother.bam</a>	uploaded bam file	bam	296.1 MB
<input checked="" type="checkbox"/>	<a href="https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_proband.bam">https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_proband.bam</a>	uploaded bam file	bam	391.6 MB
<input checked="" type="checkbox"/>	<a href="https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt">https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt</a>	uploaded tabular file	tabular	68 b

« < 1 > » 10 per page, 5 total

# Premapped reads upload

## Import into History

---





Select history:



1

or create new:

2




# Premapped reads upload

History    



search datasets  

**TP\_GTN\_WES\_disease**


4 shown

(empty)   



---

4: <https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt>   



---

3: [https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_proband.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_proband.bam)   

---

2: [https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_mother.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_mother.bam)   

---

1: [https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_father.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam)   

1

# Premapped reads upload

## Edit Dataset Attributes

☰ Attributes

⚙️ Convert

📄 Datatypes

👤 Permissions

### Name

[https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_father.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam)

### Info

uploaded bam file

### Annotation

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

### Database/Build

unspecified (?)

💾 Save

🔄 Auto-detect

# Premapped reads upload

Edit Dataset Attributes

Attributes Convert Datatypes Permissions

**1** **2 - Use self-explanatory names**

**Name**  
mapped\_reads\_father.bam

**Info**  
[https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_father.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam)  
uploaded **bam** file

**Annotation**

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

**3** **Database/Build**  
unspecified (?)  
hg19

**4**

Homo sapiens (hg19 with mtDNA replaced with rCRS) (Homo\_sapiens\_nuHg19\_mtrCRS)  
Human Feb. 2009 (GRCh37/hg19) (hg19)  
Homo sapiens (hg19 with mtDNA replaced with rCRS, and containing pUC18 and phiX174) (hg19\_CRS\_pUC18\_phiX174)  
**Human Feb. 2009 (GRCh37/hg19) (hg19)**  
GRCh37.p10 Sep. 2012 (GRCh37.p10/hg19Patch10) (hg19Patch10)  
GRCh37.p9 Jul. 2012 (GRCh37.p9/hg19Patch9) (hg19Patch9)  
GRCh37.p5 Jun. 2011 (GRCh37.p5/hg19Patch5) (hg19Patch5)



# Premapped reads upload

2: [https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_mother.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_mother.bam)   

1: mapped\_reads\_father.bam   

336.9 MB  
format: **bam**, database: **hg19**

[https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_father.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam)  
uploaded bam file




      

display at UCSC main test  
display at Ensembl Current  
display with IGV local  
display in IGB View

Binary bam alignments file

1








# Premapped reads upload


1: mapped\_reads\_father.bam   

336.9 MB

format: **bam**, database: **hg19**

```
https://zenodo.org/api/files/
dd4bcd95-4412-4ac0-
a7d2-23cf1c69e0bc
/mapped_reads_father.bam
uploaded bam file
```

1  




name

display at Ensembl [Current](#)

display with IGV [local](#)

display in IGB [View](#)

# Premapped reads upload

1: mapped\_reads\_father.b   








am



**father**

336.9 MB

format: **bam**, database: **hg19**

[https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped\\_reads\\_father.bam](https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam)  
uploaded bam file




**#father**  | Add Tags 



display at UCSC main test  
display at Ensembl Current  
display with IGV local  
display in IGB View




# Premapped reads upload



TP\_GTN\_WES\_disease




4 shown

68 b   

4: <https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt>   

3: mapped\_reads\_proband.bam     
proband

2: mapped\_reads\_mother.bam     
mother

1: mapped\_reads\_father.bam     
father

1

# Premapped reads upload

**Edit Dataset Attributes**

☰ Attributes ⚙️ Convert 📄 Datatypes 👤 Permissions

**1** **Name**  
Pedigree.txt

**2** **Info**  
<https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt>  
uploaded tabular file

**Annotation**

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

**Database/Build**  
unspecified (?)

**Number of comment lines**

**3** Save Auto-detect

# Premapped reads upload

**TP\_GTN\_WES\_disease**

4 shown

68 b ✓ ▶ 💬

---

**4: Pedigree.txt** 👁️ ✎ ✕

---

**3: mapped\_reads\_proband.bam** 👁️ ✎ ✕

proband

---

**2: mapped\_reads\_mother.bam** 👁️ ✎ ✕

mother

---

**1: mapped\_reads\_father.bam** 👁️ ✎ ✕

father

# Mapped reads postprocessing

## Warning :

- Depends on technology
- Depends on goal
- Depends on the pipeline used (steps, software, etc.)

## 1. Filter reads based on characteristics :

- Retain only forward and reverse reads mapped successfully to the reference
- Exclude possible contaminant DNA or sequencing artefact

## 2. Remove/Mark duplicate reads

- PCR-overamplification of genomic fragment during sequencing library preparation

# Mapped reads postprocessing - Filter reads

The screenshot shows a web interface for bioinformatics tools. At the top, there is a search bar containing the text 'filter bam output', which is highlighted with a red box and a red number '1'. Below the search bar are buttons for 'Upload Data' and 'Show Sections'. A list of tools is displayed, including 'Filter BAM datasets on a variety of attributes', 'Samtools view', 'Generate pileup', 'BAM filter', 'BAM-to-SAM', 'SAM-to-BAM', 'Merge BAM Files', 'Samtools split', 'Format MetaPhlan2', 'Convert, Merge, Randomize BAM', and 'Slice BAM'. The 'Filter SAM or BAM' tool is highlighted with a red box and a red number '2'. The tool description reads: 'Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region'.

Tools

1 filter bam output

Upload Data

Show Sections

Filter BAM datasets on a variety of attributes

Samtools view - reformat, filter, or subsample SAM, BAM or CRAM

Generate pileup from BAM dataset

BAM filter Removes reads from a BAM file based on criteria

BAM-to-SAM convert BAM to SAM

SAM-to-BAM convert SAM to BAM

Merge BAM Files merges BAM files together

Samtools split BAM dataset on readgroups

Format MetaPhlan2 output for Krona

Convert, Merge, Randomize BAM datasets and perform other transformations

Slice BAM by genomic regions

correctGCBias uses the output from computeGCBias to generate GC-corrected BAM/CRAM files

2 Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region



# Mapped reads processing - Filter reads

Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region (Galaxy Version 1.8+galaxy1)

SAM or BAM file to filter

**1**

**2 - Hold Ctrl key**

- 3: mapped\_reads\_proband.bam
- 2: mapped\_reads\_mother.bam
- 1: mapped\_reads\_father.bam

**3**

Header in output

Include header

Minimum MAPQ quality score

(-q)

Filter on bitwise flag

yes

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

# Mapped reads postprocessing - Filter reads

## Only output alignments with all of these flag bits set

Select/Unselect all

- Read is paired
- Read is mapped in a proper pair
- The read is unmapped
- The mate is unmapped
- Read is mapped to the reverse strand of the reference
- Mate is mapped to the reverse strand of the reference
- Read is the first in a pair
- Read is the second in a pair
- The alignment of this read is not primary
- The read fails platform/vendor quality checks
- The read is a PCR or optical duplicate
- Supplementary alignment

(-f)

## Skip alignments with any of these flag bits set

Select/Unselect all

- Read is paired
- Read is mapped in a proper pair
- The read is unmapped
- The mate is unmapped
- Read is mapped to the reverse strand of the reference
- Mate is mapped to the reverse strand of the reference
- Read is the first in a pair
- Read is the second in a pair
- The alignment of this read is not primary
- The read fails platform/vendor quality checks
- The read is a PCR or optical duplicate
- Supplementary alignment

1

## Select alignments from Library

(-l) Requires headers in the input SAM or BAM, otherwise no alignments will be output

## Select alignments from Read Group

(-r) Requires headers in the input SAM or BAM, otherwise no alignments will be output

## Output alignments overlapping the regions in the BED file

No bed dataset available.

(-L)

## Use inverse selection

No

Select the opposite of the listed chromosomes

## Select regions (only used when the input is in BAM format)

region should be presented in one of the following formats: `chr1`, `chr2:1,000` and `chr3:1000-2,000`

## Select the output format

BAM (-b)

## Email notification

Send an email notification when the job completes.

2

# Mapped reads postprocessing - Filter reads

**TP\_GTN\_WES\_disease**  
7 shown  
1 GB

7: Filter SAM or BAM, output SAM or BAM on data 3: bam  
proband

6: Filter SAM or BAM, output SAM or BAM on data 2: bam  
mother

5: Filter SAM or BAM, output SAM or BAM on data 1: bam  
father



**TP\_GTN\_WES\_disease**  
7 shown  
1 GB

7: filtered\_reads\_proband.bam  
proband

6: filtered\_reads\_mother.bam  
mother

5: filtered\_reads\_father.bam  
father

# Mapped reads postprocessing - Duplicate reads

1

Tools ☆ ☰

markdup ✕

📁 Upload Data

👁 Show Sections

**MarkDuplicatesWithMateCigar** examine aligned records in BAM datasets to locate duplicate molecules

**QualiMap BamQC**

**Map with BWA-MEM** - map medium and long reads (> 100 bp) against reference genome

**AddOrReplaceReadGroups** add or replaces read group information


**Map with BWA** - map short reads (< 100 bp) against reference genome

**FastqToSam** convert Fastq data into unaligned BAM






2

**MarkDuplicates** examine aligned records in BAM datasets to locate duplicate molecules

# Mapped reads postprocessing - Duplicate reads


 **MarkDuplicates** examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.3) ☆ 🔄 ▼

**Select SAM/BAM dataset or dataset collection**

   7: filtered\_reads\_proband.bam ▼  

If empty, upload or import a SAM/BAM dataset

**Comment**



You can provide multiple comments

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

No  
REMOVE\_DUPLICATES; default=False

**Assume the input file is already sorted** How can we know ?

Yes  
ASSUME\_SORTED; default=True

**The scoring strategy for choosing the non-duplicate among candidates**

SUM\_OF\_BASE\_QUALITIES ▼

DUPLICATE\_SCORING\_STRATEGY; default=SUM\_OF\_BASE\_QUALITIES

**Regular expression that can be used in unusual situations to parse non-standard read names in the incoming SAM/BAM dataset**

READ\_NAME\_REGEX; Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. See help below for more info; default="" (uses : separation)

# Mapped reads postprocessing - Duplicate reads

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ
@HD VN:1.3 SO:coordinate									
@SQ SN:chr8 LN:146364022									
@RG ID:001 SM:father PL:ILLUMINA									
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -t 8 -v 1 -R @RG\tID:001\tSM:father\tPL:ILLUMINA localref.fa /data/dnb02/galaxy_db/files/009/499/dataset_9499701.dat /data/dnb02/galaxy_db/files/009/499/datas									
DCW97JN1:309:C0C42ACXX:5:2202:19629:56029	163	chr8	11710	3	101M	=	11865	256	CCATGGCAGAGCTCCCTCCTCAGCACATGGGGAGCAGACAGGAAGT
DCW97JN1:309:C0C42ACXX:4:1206:10027:62829	163	chr8	11712	0	101M	=	11864	253	ATGGCAGAGCTCCCTCCTCAGCACATGGGGAGCAGACAGGAAGTTT
DCW97JN1:309:C0C42ACXX:4:1115:17796:60101	163	chr8	11712	15	101M	=	11869	253	ATGGCAGAGCTCCCTCCTCAGCACATGGGGAGCAGACAGGAAGTTT
DCW97JN1:309:C0C42ACXX:5:1216:6300:20909	99	chr8	11783	27	101M	=	11966	271	AGCCACGCTCCTCCAGGTCAGTCTTAAGACAACGAACTCTGGGC
DCW97JN1:309:C0C42ACXX:4:1206:10027:62829	83	chr8	11864	1	101M	=	11712	-253	AAGCCATGGTGCCCCACCCTCGGGTGGGTCCTGAGGAGAACAAAGC
DCW97JN1:309:C0C42ACXX:5:2202:19629:56029	83	chr8	11865	8	101M	=	11710	-256	AGCCATGGTGACCCACCCTCGGGTGGGTCCTGAGGAGAACAAAGCT
DCW97JN1:309:C0C42ACXX:4:1115:17796:60101	83	chr8	11869	15	96M5S	=	11712	-253	ATGGTGACCCACCCTCGGGTGGGTCCTGAGGAGAACAAAGCTCTGG
DCW97JN1:309:C0C42ACXX:5:1216:6300:20909	147	chr8	11966	27	13S88M	=	11783	-271	CCAGATCCCAAAACCCTGATCCCTACCCTGGATCCTAAGTCTGTCCCT
DCW97JN1:309:C0C42ACXX:5:2210:15831:85655	145	chr8	98822	0	52S35M14S	=	110566976	110468121	TTTTAAAAATTTAAAAAAAAAAAAATTTGGCCAAAAAAATTTATTTTTTT
DCW97JN1:309:C0C42ACXX:4:2209:3455:67435	161	chr8	98823	0	45S43M13S	=	39494954	39396232	CCCCAAAAAAATTTCCGGGTTTTGGGTTTTTCCACCCAAAAATTTT
DCW97JN1:309:C0C42ACXX:5:2305:4557:78030	2115	chr8	98823	0	58H34M9H	=	141889681	141790859	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAAATTT
DCW97JN1:309:C0C42ACXX:5:2111:10544:43299	2195	chr8	98824	0	43M58H	=	16979740	16880875	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAAATTTTTTTTTA

History ↺ + 🗄 ⚙

search datasets ? ✕

**TP\_GTN\_WES\_disease**

7 shown

1 GB ☑ 🗄 🗨

7: filtered\_reads\_proband.ba 👁 ✎ ✕

am proband

6: filtered\_reads\_mother.ba 👁 ✎ ✕

m mother

5: filtered\_reads\_father.ba 👁 ✎ ✕

m father

1

• `<instrument>:<run_number>:<flowcell_ID>:<lane>:<tile>:<x-pos>:<y-pos>`



SO tag :

- Sorting order of alignments
- Unknown, unsorted, queryname (QNAME) or coordinate (RNAME/POS)


# Mapped reads postprocessing - Duplicate reads

**MarkDuplicates** examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.3)

Select SAM/BAM dataset or dataset collection


1  

- 7: filtered\_reads\_proband.bam
- 6: filtered\_reads\_mother.bam
- 5: filtered\_reads\_father.bam
- 3: mapped\_reads\_proband.bam
- 2: mapped\_reads\_mother.bam
- 1: mapped\_reads\_father.bam

 This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

If empty, upload or import a SAM/BAM dataset

Comment



You can provide multiple comments

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

No

REMOVE\_DUPLICATES: default=False

**Assume the input file is already sorted**

Yes

ASSUME\_SORTED: default=True

**4 - Depends on goal and pipeline**

**5 - Use default**

The scoring strategy for choosing the non-duplicate among candidates

SUM\_OF\_BASE\_QUALITIES

DUPLICATE\_SCORING\_STRATEGY: default=SUM\_OF\_BASE\_QUALITIES

Regular expression that can be used in unusual situations to parse non-standard read names in the incoming SAM/BAM dataset

READ\_NAME\_REGEX; Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size. See help below for more info; default=" (uses : separation)

# Mapped reads postprocessing - Duplicate reads

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



OPTICAL\_DUPLICATE\_PIXEL\_DISTANCE; default=100

**Barcode Tag**

Barcode SAM tag. This tag can be utilized when you have data from an assay that includes Unique Molecular Indices. Typically 'RX'

**Select validation stringency**

Lenient

Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

**Email notification**



Send an email notification when the job completes.

✓ Execute

6



# Mapped reads postprocessing - Duplicate reads

```
## htsjdk.samtools.metrics.StringHeader
# MarkDuplicates TAGGING_POLICY=All INPUT=[filtered_reads_proband.bam] OUTPUT=/shared/ibfstor1/galaxy/jobs/001/460/1460354/outputs/galaxy_dataset_41a41de0-
f5f3-4b31-9ad7-223bed9aaba2.dat METRICS_FILE=/shared/ibfstor1/galaxy/jobs/001/460/1460354/outputs/galaxy_dataset_9510908c-1d36-475c-b100-0e1467fff83d.dat REMOVE_DUPLICATES=false
ASSUME_SORTED=true DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUALITIES OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 TMP_DIR=/shared/ibfstor1/galaxy/jobs/001/460/1460354/tmp] VERBOSITY=ERROR
QUIET=true VALIDATION_STRINGENCY=LENIENT MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=8000 SORTING_COLLECTION_SIZE_RATIO=0.25
TAG_DUPLICATE_SET_MEMBERS=false REMOVE_SEQUENCING_DUPLICATES=false CLEAR_DT=true ADD_PG_TAG_TO_READS=true PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=MarkDuplicates
READ_NAME_REGEX=<optimized capture of last three ':' separated fields as numeric values> MAX_OPTICAL_DUPLICATE_SET_SIZE=300000 COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000
CREATE_INDEX=false CREATE_MDS_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json USE_JDK_DEFLATER=false USE_JDK_INFLATER=false
## htsjdk.samtools.metrics.StringHeader
# Started on: Thu Mar 24 20:39:12 CET 2022

## METRICS CLASS picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS EXAMINED READ PAIRS EXAMINED SECONDARY_OR SUPPLEMENTARY_RDS UNMAPPED_READS UNPAIRED_READ_DUPLICATES READ_PAIR_DUPLICATES
READ_PAIR_OPTICAL_DUPLICATES PERCENT_DUPLICATION ESTIMATED_LIBRARY_SIZE
Unknown Library 0 2380197 1324 0 781643 244 0.328394 2777843
```

**Unmapped reads** (points to 0)

**Duplicates & Optical duplicates** (points to 781643 244)

**Percentage duplication** (points to 0.328394)

**Header** (points to READ\_PAIR\_DUPLICATES)

History

search datasets

TP\_GTN\_WES\_disease

13 shown

2.03 GB





13: MarkDuplicates on data 7: MarkDuplicates BAM output



12: MarkDuplicates on data 7: MarkDuplicate metrics

1




# Mapped reads postprocessing - Duplicate reads

QNAME	FLAG	RNAME	POS	MAPQ
@HD VN:1.5 SO:coordinate				
@SQ SN:chr8 LN:146364022				
@RG ID:001 SM:father PL:ILLUMINA				
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa mem -t 8 -v 1 -R @RG{TID:001 TSM:father PL:ILLUMINA localref.fa /data/dnb02/galaxy_db/files/009/499/dataset_9499701.dat /data/dnb02/galaxy_db/files/009/499/dataset_9499701.dat} @PG ID:MarkDuplicates VN:2.18.2-SNAPSHOT CL:MarkDuplicates TAGGING_POLICY=All INPUT=[filtered_reads_father_bam] OUTPUT=/shared/ibstor1/galaxy/jobs/001/460/1460352/outputs/galaxy_dataset_37efe38c				
DCW97JN1:309:C0C42ACXX:5:2202:19629:56029	163	chr8	11710	
DCW97JN1:309:C0C42ACXX:4:1206:10027:62829	163	chr8	11712	
DCW97JN1:309:C0C42ACXX:4:1115:17796:60101	163	chr8	11712	
DCW97JN1:309:C0C42ACXX:5:1216:6300:20909	99	chr8	11783	
DCW97JN1:309:C0C42ACXX:4:1206:10027:62829	83	chr8	11864	
DCW97JN1:309:C0C42ACXX:5:2202:19629:56029	83	chr8	11865	
DCW97JN1:309:C0C42ACXX:4:1115:17796:60101	83	chr8	11869	
DCW97JN1:309:C0C42ACXX:5:1216:6300:20909	147	chr8	11966	
DCW97JN1:309:C0C42ACXX:5:2210:15831:85655	145	chr8	98822	
DCW97JN1:309:C0C42ACXX:4:2209:3455:67435	161	chr8	98823	
DCW97JN1:309:C0C42ACXX:5:2305:4557:78030	2115	chr8	98823	
DCW97JN1:309:C0C42ACXX:5:2111:10544:43299	2195	chr8	98824	
DCW97JN1:309:C0C42ACXX:4:2211:6915:3569	99	chr8	115864	
DCW97JN1:309:C0C42ACXX:4:2206:12976:57510	99	chr8	115873	
DCW97JN1:309:C0C42ACXX:4:1313:14027:15986	1187	chr8	115884	
DCW97JN1:309:C0C42ACXX:5:1208:19040:61299	1187	chr8	115884	
DCW97JN1:309:C0C42ACXX:5:1312:19336:8504	163	chr8	115884	
DCW97JN1:309:C0C42ACXX:4:1108:20076:55158	99	chr8	115922	
DCW97JN1:309:C0C42ACXX:5:2206:1793:6208	99	chr8	115934	
DCW97JN1:309:C0C42ACXX:5:1207:18720:30262	163	chr8	115940	
DCW97JN1:309:C0C42ACXX:5:1102:15493:91613	1123	chr8	115945	
DCW97JN1:309:C0C42ACXX:5:1307:11684:10108	99	chr8	115945	
















History    

search datasets  

TP\_GTN\_WES\_disease

13 shown   

2.03 GB

- 13: MarkDuplicates on data 7: MarkDuplicates BAM output     
**proband**
- 12: MarkDuplicates on data 7: MarkDuplicate metrics     
**proband**
- 11: MarkDuplicates on data 6: MarkDuplicates BAM output     
**mother**
- 10: MarkDuplicates on data 6: MarkDuplicate metrics     
**mother**
- 9: MarkDuplicates on data 5: MarkDuplicates BAM output     
**father**

**1**

# Mapped reads postprocessing - Duplicate reads

## Decoding SAM flags

This utility makes it easy to identify what are the properties for a given combination of properties.

To decode a given SAM flag value, just enter the number

SAM Flag:

Toggle first in pair / second in pair

### Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties for those that you'd like to include. The flag value will be shown in the SAM

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment

## Decoding SAM flags

This utility makes it easy to identify what are the properties for a given combination of properties.

To decode a given SAM flag value, just enter the number

SAM Flag:

Toggle first in pair / second in pair

### Find SAM flag by property:




To find out what the SAM flag value would be for a given combination of properties for those that you'd like to include. The flag value will be shown in the SA
















- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment

# Mapped reads postprocessing - Duplicate reads

TP\_GTN\_WES\_disease

13 shown

2.03 GB   



13: markdup_proband.bam	  
<b>proband</b>	
12: markdup_proband_metrics	  
<b>proband</b>	
11: markdup_mother.bam	  
<b>mother</b>	
10: markdup_mother_metrics	  
<b>mother</b>	
9: markdup_father.bam	  
<b>father</b>	
8: markdup_father_metrics	  
<b>father</b>	


# Tutorial steps


1. Perform postprocessing from premapped reads
2. Variant calling
3. Variant annotation and reporting


# Variant calling

**1**

Tools  

freebayes 

 Upload Data

 Show Sections

**BamLeftAlign** indels in BAM datasets

**2** **FreeBayes** bayesian genetic variant detector

**Map with BWA-MEM** - map medium and long reads (> 100 bp) against reference genome

**SnpEff build:** database from Genbank or GFF record

**Map with BWA** - map short reads (< 100 bp) against reference genome

# Variant calling

FreeBayes bayesian genetic variant detector (Galaxy Version 1.3.6+galaxy0)



## Choose the source for the reference genome

Locally cached

### Run in batch mode?

- Run individually
- Merge output VCFs

1

Selecting individual mode will generate one VCF dataset for each input BAM dataset. Selecting the merge option will produce one VCF dataset for all input BAM datasets

### BAM or CRAM dataset(s)

2



13: markdup\_proband.bam  
11: markdup\_mother.bam  
9: markdup\_father.bam  
7: filtered\_reads\_proband.bam  
6: filtered\_reads\_mother.bam  
5: filtered\_reads\_father.bam



### Using reference genome

Human (Homo sapiens): hg19

3

# Variant calling

## Limit variant calling to a set of regions?

Do not limit

Sets --targets or --region options

## Read coverage

Use defaults

Sets --min-coverage, --limit-coverage, and --skip-coverage

## Choose parameter selection level

2. Simple diploid calling with filtering and coverage

Select how much control over the freebayes run you need

## Email notification



Send an email notification when the job completes.

✓ Execute

## Galaxy-specific options

Galaxy allows five levels of control over FreeBayes options, provided by the **Choose parameter selection level** menu option. These are:

1. *Simple diploid calling*: The simplest possible FreeBayes application. Equivalent to using FreeBayes with only a BAM input and no other parameter options.
2. *Simple diploid calling with filtering and coverage*: Same as #1 plus two additional options: -0 (standard filters: --min-mapping-quality 30 --min-base-quality 20 --min-supporting-allele-qsum 0 --genotype-variant-threshold 0) and --min-coverage.
3. *Frequency-based pooled calling*: This is equivalent to using FreeBayes with the following options: --haplotype-length 0 --min-alternate-count 1 --min-alternate-fraction 0 --pooled-continuous --report-monomorphic. This is the best choice for calling variants in mixtures such as viral, bacterial, or organellar genomes.
4. *Frequency-based pooled calling with filtering and coverage*: Same as #3 but adds -0 and --min-coverage like in #2.
5. *Complete list of all options*: Gives you full control by exposing all FreeBayes options as Galaxy parameters.



# Variant calling

## Dataset Information

Number	14
Name	FreeBayes on data 13, data 11, and data 9 (variants)
Created	Thursday Mar 24th 7:51:33 2022 UTC
Filesize	4.5 MB
Dbkey	hg19
Format	vcf
File contents	contents
History Content API ID	822eead7687ce5a1
History API ID	57e9be0d003985de
UUID	3ceb74fa-1ceb-44c5-91d0-d2eaf6ce9b09

## Tool Parameters

Input Parameter	Value
Choose the source for the reference genome	cached
Run in batch mode?	merge
BAM or CRAM dataset(s)	9 markdup_father.bam father 11 markdup_mother.bam mother 13 markdup_proband.bam proband
Using reference genome	hg19
Limit variant calling to a set of regions?	do_not_limit
Read coverage	do_not_set
Choose parameter selection level	simple_w_filters

The screenshot shows a web interface for dataset management. At the top, there is a search bar with the text "search datasets" and a search icon. Below the search bar, the dataset name "TP\_GTN\_WES\_disease" is displayed, along with "14 shown" and "2.03 GB". A list of datasets is shown, with the first one highlighted in green: "14: FreeBayes on data 13, data 11, and data 9 (variants)". This entry has three sub-entries: "father", "mother", and "proband", each with a small icon. Below the sub-entries, it says "8,376 lines, 62 comments" and "format: vcf, database: hg19". There are several icons below this text, including a trash can, a refresh icon, and a help icon. Below the dataset list, there is a section for "display at UCSC main" and "display at RVviewer main". A code block is shown with the following content: 

```
1. Chrom
##fileformat=VCFv4.2
##fileDate=20220324
##source=FreeBayes v1.3.6
##reference=/shared/bank/data.galaxypro
##contig=chr8,length=146364022
```

 Below the code block, there is a section for "13: markdup\_proband.ba" and "12: markdup\_proband\_me".

# Variant calling - VCF

Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
##fileformat=VCFv4.2							
##fileDate=20220324							
##source=freeBayes v1.3.6							
##reference=/shared/bank/data.galaxyproject.org/byhand/hg19/sam_index/hg19.fa							
##contig=<ID=chr8,length=146364022>							
##phasing=none							
##commandline="freebayes --region chr8:0..146364022 --bam b_0.bam --bam b_1.bam --bam b_2.bam --fasta-reference /shared/bank/data.galaxyproject.org/byhand/hg19/sam_index/hg19.fa --vcf ./vcf_output/part_ch							
##INFO=<D=NS,Number=1,Type=Integer,Description="Number of samples with data">							
##INFO=<D=DP,Number=1,Type=Integer,Description="Total read depth at the locus">							
##INFO=<D=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">							
##INFO=<D=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">							
##INFO=<D=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">							
##INFO=<D=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">							
##INFO=<D=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">							
##INFO=<D=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">							
##INFO=<D=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">							
##INFO=<D=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">							
##INFO=<D=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">							
##INFO=<D=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">							
##INFO=<D=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">							
##INFO=<D=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">							
##INFO=<D=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">							
##INFO=<D=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">							

History ↻ + □ ⚙

search datasets ? ✕

### TP\_GTN\_WES\_disease

14 shown

2.03 GB ☑ 🗨

14: FreeBayes on data 13, data 11, and data 9 (variants) 👁 ✎ ✕

1

father mother proband

8,376 lines, 62 comments

format: vcf, database: hg19

📄 🔗 🔍 🔄 🗨 ? 🗨

display at UCSC main test

display with IGV local

display at RViewer main

1. chrom

```
##fileformat=VCFv4.2
##fileDate=20220324
##source=freeBayes v1.3.6
##reference=/shared/bank/data.galaxyproj
##contig=<ID=chr8,length=146364022>
```

# Variant calling - VCF

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled  
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the  
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">  
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Number of observation for each allele">  
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">  
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">  
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">  
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">  
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
```

# Variant calling - VCF

## Mandatory columns

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr8	115956	.	A	T	9.09784e-07	.
chr8	116079	.	G	A	103.501	.
chr8	116701	.	A	G	3.98084e-05	.
chr8	116895	.	A	G	184.59	.
chr8	160552	.	G	A	1.00485	.
chr8	160608	.	A	C	722.504	.
chr8	160609	.	AAAAAATAAAAAATAAACATAAAAAATG	AAAAAATAAAAAATAAAAAATAAACATAAAAAATG	0.370623	.
chr8	160679	.	G	A	5.46006e-08	.
chr8	160719	.	C	T	9.28165e-15	.
chr8	160736	.	G	T	530.182	.
chr8	160760	.	C	G	237.975	.

# Variant calling - VCF

## Mandatory column

INFO

AB=0;ABP=0;AC=0;AF=0;AN=6;AO=4;CIGAR=1X;DP=51;DPB=51;DPRA=2.33333;EPP=11.6962;EPPR=36.6912;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=15.5049;PAIRED=1;PAI  
AB=0.276596;ABP=23.3852;AC=2;AF=0.333333;AN=6;AO=15;CIGAR=1X;DP=74;DPB=74;DPRA=0;EPP=20.5268;EPPR=29.8409;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=4.51  
AB=0.3125;ABP=7.89611;AC=1;AF=0.166667;AN=6;AO=14;CIGAR=1X;DP=240;DPB=240;DPRA=0;EPP=3.0103;EPPR=6.85361;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=11.6;F  
AB=0;ABP=0;AC=6;AF=1;AN=6;AO=6;CIGAR=1X;DP=6;DPB=6;DPRA=0;EPP=8.80089;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=3;NUMALT=1;ODDS=8.00168;PAIRED=1;PAIREDR=0;PAO=0;PI  
AB=0.25;ABP=9.52472;AC=2;AF=0.333333;AN=6;AO=3;CIGAR=1X;DP=19;DPB=19;DPRA=0.857143;EPP=3.73412;EPPR=3.55317;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=59.5;NS=3;NUMALT=1;ODDS=1  
AB=0.4375;ABP=5.72464;AC=3;AF=0.5;AN=6;AO=35;CIGAR=1X;DP=80;DPB=80;DPRA=0;EPP=48.239;EPPR=49.3833;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=7.50894;PAIRE  
AB=0.222222;ABP=15.074;AC=1;AF=0.166667;AN=6;AO=7;CIGAR=1M4I25M;DP=80;DPB=82.5385;DPRA=0;EPP=5.80219;EPPR=113.696;GTI=0;LEN=4;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;O  
AB=0.130584;ABP=347.946;AC=3;AF=0.5;AN=6;AO=38;CIGAR=1X;DP=291;DPB=291;DPRA=0;EPP=54.4399;EPPR=6.10873;GTI=2;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=18.194;F  
AB=0.101399;ABP=397.702;AC=2;AF=0.333333;AN=6;AO=29;CIGAR=1X;DP=441;DPB=441;DPRA=0.922581;EPP=3.68421;EPPR=65.7822;GTI=1;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;C  
AB=0.188995;ABP=354.186;AC=3;AF=0.5;AN=6;AO=79;CIGAR=1X;DP=418;DPB=418;DPRA=0;EPP=20.1897;EPPR=12.7531;GTI=1;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=34.1344;  
AB=0.124567;ABP=356.825;AC=2;AF=0.333333;AN=6;AO=37;CIGAR=1X;DP=382;DPB=382;DPRA=0;EPP=7.76406;EPPR=133.565;GTI=1;LEN=1;MEANALT=1.66667;MQM=60;MQMR=60;NS=3;NUMALT=1;OI  
AB=0.124031;ABP=161.393;AC=1;AF=0.166667;AN=6;AO=21;CIGAR=2X;DP=310;DPB=310;DPRA=0.962264;EPP=3.94093;EPPR=397.039;GTI=0;LEN=2;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;C

# Variant calling - VCF

FORMAT	proband	mother	father
GT:DP:AD:RO:QR:AO:QA:GL	0/0:30:27,3:27:891:3:92:0,-0.445657,-71.9117	0/0:12:11,1:11:353:1:33:0,-0.313225,-28.7828	0/0:9:9,0:9:286:0:0:0,-2.70927,-26.0508
GT:DP:AD:RO:QR:AO:QA:GL	0/1:24:16,8:16:644:8:260:-16.4945,0,-51.046	0/0:27:25,2:25:1021:2:64:0,-2.04915,-86.078	0/1:23:18,5:18:728:5:166:-8.34408,0,-58.9123
GT:DP:AD:RO:QR:AO:QA:GL	0/0:113:109,4:109:3436:4:144:0,-20.7059,-296.176	0/0:111:106,5:106:3382:5:193:0,-15.6745,-286.887	0/1:16:11,5:11:364:5:178:-11.5434,0,-28.2653
GT:DP:AD:RO:QR:AO:QA:GL	1/1:4:0,4:0:4:167:-15.4235,-1.20412,0	1/1:1:0,1:0:0:1:36:-3.59827,-0.30103,0	1/1:1:0,1:0:0:1:33:-3.29913,-0.30103,0
GT:DP:AD:RO:QR:AO:QA:GL	0/1:9:7,2:7:297:2:66:-3.55868,0,-24.3555	0/1:3:2,1:2:85:1:35:-2.59554,0,-7.15727	0/0:7:7,0:7:271:0:0:0,-2.10721,-24.7468
GT:DP:AD:RO:QR:AO:QA:GL	0/1:43:22,21:22:776:21:828:-61.8817,0,-57.2184	0/1:17:14,3:14:502:3:114:-5.51484,0,-40.3989	0/1:20:9,11:9:307:11:421:-32.2186,0,-21.9403
GT:DP:AD:RO:QR:AO:QA:GL	0/0:42:41,1:41:1422:1:34:0,-9.58258,-124.881	0/1:18:14,4:14:477:4:132:-6.46629,0,-37.5149	0/0:20:18,2:18:614:2:64:0,-0.00155201,-49.4499
GT:DP:AD:RO:QR:AO:QA:GL	0/1:133:118,15:118:3976:15:509:-6.09578,0,-318.014	0/1:59:49,10:49:1629:10:328:-12.0781,0,-129.133	0/1:99:86,13:86:2819:13:441:-10.2124,0,-224.147
GT:DP:AD:RO:QR:AO:QA:GL	0/1:185:166,19:166:6862:19:635:-1.7759,0,-561.244	0/1:101:91,10:91:3600:10:342:-0.707324,0,-293.6	0/0:155:154,0:154:6061:0:0:0,-46.3586,-544.867

**Genotypes  
format**

**Proband genotypes  
information**




**Mother genotypes  
information**

**Father genotypes  
information**




# Variant calling

**TP\_GTN\_WES\_disease**

14 shown




2.03 GB   

---

14: freebayes\_calling.vcf   




**father** **mother** **proband**

---

13: markup\_proband.ba  
m   

**proband**

---

12: markup\_proband\_me  
trics   

**proband**



# Tutorial steps


1. Perform postprocessing from premapped reads
2. Variant calling
3. Variant annotation and reporting





# Variant normalization

**1**

Tools  

bcftools norm 

 Upload Data

 Show Sections

**2**

**bcftools norm** Left-align and normalize indels; check if REF alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows

**bcftools merge** Merge multiple VCF/BCF files from non-overlapping sample sets to create one multi-sample file

**bcftools cnv** Call copy number variation from VCF B-allele frequency (BAF) and Log R Ratio intensity (LRR) values

# Variant normalization

**bcftools norm** Left-align and normalize indels; check if REF alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows (Galaxy Version 1.10) ☆ 🔗 ▼

## VCF/BCF Data

   14: freebayes\_calling.vcf **1**   

## Choose the source for the reference genome

Use a built-in genome ▼

**Reference genome** **2**

Human (Homo sapiens): hg19 ▼

## When any REF allele does not match the reference genome base **3**

ignore the problem (-w)

exclude the variant record from the output (-wx)

fix the variant record using the reference genome information (-ws)

exit with an error (-e)

Warnings about REF mismatches will be emitted to the standard error (stderr) stream, and it is recommended to check there for problems if you choose not to exit with an error immediately upon encountering a mismatch.

## Left-align and normalize indels? **4**

Yes

(--do-not-normalize)

# Variant normalization

1

Perform deduplication for the following types of variant records

- do not deduplicate any records
- snps
- indels
- both
- any

2

~multiallelics

split multiallelic sites into biallelic records (-)

split the following variant types

- SNPs
- indels
- both

[Restrict all operations to](#)



[Other Options](#)



3

output\_type

uncompressed VCF

Email notification






Send an email notification when the job completes.

4

Execute

# Variant normalization









15: bcftools norm on data   

14

[father](#) [mother](#) [proband](#)

8,737 lines, 65 comments  
format: **vcf**, database: **hg19**

Lines total/split/realigned/skipped:  
8376/287/1771/0

display at UCSC main test  
display with IGV local  
display at RViewer main

# Variant normalization - Alleles

14: freebayes\_calling.vcf   

**father** **mother** **proband**




8,376 lines, 62 comments  
format: **vcf**, database: **hg19**

display at UCSC main test  
display with IGV local  
display at RViewer main

**1. Chrom**







```
##fileformat=VCFv4.2
##fileDate=20220324
##source=freeBayes v1.3.6
##reference=/shared/bank/data.galaxypro:
##contig=<ID=chr8,length=146364022>
```

15: bcftools norm on data   

**father** **mother** **proband**

8,737 lines, 65 comments  
format: **vcf**, database: **hg19**

Lines total/split/realigned/skipped:  
8376/287/1771/0

display at UCSC main test  
display with IGV local  
display at RViewer main

160609 . AAAAAATAAAAAATAAACATAAAAAATG AAAATAAAAAATAAAAAATAAACATAAAAAATG

160609 . A AAAAT

163302 . CATATATG CATATG

163302 . CAT C

163366 . TAGAC CAGAG,TAGAG

163366 . TAGAC CAGAG

163370 . C G

# Variant normalization - Genotypes

## Initial file

```
163550 . AAGT GAGC,GAGT
```

```
1/2 169:0,61,108:0:0:61,108:2328,4362:-550.761,-359.801,-341.438,-191.22,0,-158.709
```

```
1/2 112:0,39,72:0:0:39,72:1461,2734:-343.835,-224.186,-212.446,-119.697,0,-98.023
```

```
1/1 112:0,112,0:0:0:112,0:4100,0:-368.767,-33.7154,0,-368.767,-33.7154,-368.767
```

## Normalized file

```
163550 . AAGT GAGC
```

```
163550 . A G
```

```
1/0 169:0,61:0:0:61:2328:-550.761,-359.801,-341.438
```

```
1/0 112:0,39:0:0:39:1461:-343.835,-224.186,-212.446
```

```
1/1 112:0,112:0:0:112:4100:-368.767,-33.7154,0
```

```
0/1 169:0,108:0:0:108:4362:-550.761,-191.22,-158.709
```

```
0/1 112:0,72:0:0:72:2734:-343.835,-119.697,-98.023
```

```
0/0 112:0,0:0:0:0:0:-368.767,-368.767,-368.767
```

# Variant filtering

## Only Homozygous reference

0/0:53:49,3:49:1823:3:103:0,-6.39174,-154.72	0/0:22:20,2:20:735:2:62:0,-0.733954,-60.5893	0/0:37:34,2:34:1262:2:67:0,-4.7389,-107.544
0/0:265:248,17:248:8589:17:592:0,-26.1668,-719.448	0/0:180:167,13:167:5745:13:447:0,-13.6264,-476.643	0/0:223:201,21:201:6904:21:716:0,-2.21506,-556.726
0/0:358:341,17:341:14409:17:568:0,-56.3297,-1243.15	0/0:250:237,13:237:9845:13:431:0,-36.1474,-845.732	0/0:260:238,22:238:9897:22:729:0,-12.3462,-823.558



## Only Homozygous alternate


1/1:105:0,105:0:0:105:3678:-331.212,-31.6082,0	1/1:47:1,46:1:37:46:1559:-136.894,-10.4506,0	1/1:61:0,61:0:0:61:2103:-189.536,-18.3628,0
--	--	---


**Do they bring some information in our case (proband affected)  
if we only consider genotypes?**


# Variant filtering

**1**

Tools  

bcftools view filter 

 Upload Data

 Show Sections

**2**

**bcftools view** VCF/BCF conversion, view, subset and filter VCF/BCF files

**bcftools filter** Apply fixed-threshold filters



# Variant filtering

bcftools view VCF/BCF conversion, view, subset and filter VCF/BCF files (Galaxy Version 1.10)

VCF/BCF Data

15: freebayes\_calling\_norm.vcf 1

Restrict to 2

Apply filters

Skip sites where FILTER column does not contain any of the strings listed (e.g. "PASS,") (--apply\_filters)

Regions

Do not restrict to Regions

Targets

Do not restrict to Targets

Include 3


Select sites for which the expression is true (--include)

Exclude

Exclude sites for which the expression is true (--exclude)

- Metrics (INFO, FORMAT)
- Boolean expressions : AND (&), OR (|), NOT (!), etc.
- Operators : Less (<), Less or equal (<=), Equal (=), Different (!=), etc.


# Variant filtering

 **bcftools view** VCF/BCF conversion, view, subset and filter VCF/BCF files (Galaxy Version 1.10)



## VCF/BCF Data

   15: freebayes\_calling\_norm.vcf   

Restrict to 

### Apply filters

Skip sites where FILTER column does not contain any of the strings listed (e.g. "PASS,") (--apply\_filters)

### Regions

Do not restrict to Regions 

### Targets

Do not restrict to Targets 

### Include

AF>0 & AF<1


Select sites for which the expression is true (--include)


### Exclude

Exclude sites for which the expression is true (--exclude)

# Variant filtering

[Subset Options](#) 

[Filter Options](#) 

[Output Options](#) 

**output\_type**

uncompressed VCF

1

**Email notification**






Send an email notification when the job completes.

✓ Execute

2

# Variant filtering







16: bcftools view on data   

15

[father](#) [mother](#) [proband](#)

6,468 lines, 67 comments  
format: **vcf**, database: **hg19**

[W::vcf\_parse\_format] Extreme  
FORMAT/AO value encountered and  
set to missing at chr8:6875540



       


display at UCSC main test  
display with IGV local  
display at RViewer main


**1. Chrom**


```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filte
##fileDate=20220324
##source=freeBayes v1.3.6
##reference=/shared/bank/data.galaxyproj
```

# Variant annotation

Tools  

1  

 Upload Data

 Show Sections

**SnpEff eff:** annotate variants for SARS-CoV-2




**SnpEff download:** download a pre-built database

**SnpEff databases:** list available databases







**SnpEff build:** database from Genbank or GFF record

2 **SnpEff eff:** annotate variants

# Variant annotation

 SnpEff eff: annotate variants (Galaxy Version 4.3+T.galaxy1)  


Sequence changes (SNPs, MNPs, InDels)

   16: freebayes\_calling\_norm\_filtered.vcf   

Input format

VCF 

Output format

VCF (only if input is VCF) 

Create CSV report, useful for downstream analysis (-csvStats)

No

Genome source


Locally installed snpEff database 

! Please provide a value for this option.

Genome

No options available 

# Variant annotation

 SnpEff eff: annotate variants (Galaxy Version 4.3+T.galaxy1) ☆ ▼

Sequence changes (SNPs, MNPs, InDels)

📁 ⬆️ 📁

Input format

Output format

Create CSV report, useful for downstream analysis (-csvStats)

No

**Genome source** ▼

**1**

Snpff Genome Version Name (e.g. GRCh38.86)

**2**

[https://sourceforge.net/projects/snpeff/files/databases/v4\\_3/](https://sourceforge.net/projects/snpeff/files/databases/v4_3/)

# Variant annotation

## Upstream / Downstream length

5000 bases

(-ud)

## Set size for splice sites (donor and acceptor) in bases

2 bases

(-ss)

## spliceRegion Settings

Use Defaults



# Variant annotation

## Annotation options

Select/Unselect all

- Use 'EFF' field compatible with older versions (instead of 'ANN')
- Use Classic Effect names and amino acid variant annotations (NON\_SYNONYMOUS\_CODING vs missense\_variant and G180R vs p.Gly180Arg/c.538G>C)
- Override classic and use Sequence Ontology terms for effects (missense\_variant vs NON\_SYNONYMOUS\_CODING)
- Override classic and use HGVS annotations for amino acid annotations (p.Gly180Arg/c.538G>C vs G180R)
- Old notation style notation: E.g. 'c.G123T' instead of 'c.123G>T' and 'X' instead of '\*'
- Use one letter Amino acid codes in HGVS notation. E.g. p.R47G instead of p.Arg47Gly
- Use transcript ID in HGVS notation. E.g. ENST00000252100:c.914C>G instead of c.914C>G
- Do not shift variants according to HGVS notation (most 3prime end)
- Do not add HGVS annotations
- Only use canonical transcripts
- Only use protein coding transcripts
- Use gene ID instead of gene name (VCF output)
- Disable IUB code expansion in input variants
- Add OICR tag in VCF file
- Add loss of function (LOF) and nonsense mediated decay (NMD) tags
- Do not add LOF and NMD annotations
- Disable motif annotations
- Disable NextProt annotations
- Disable interaction annotations
- Perform 'cancer' comparisons (somatic vs. germline)

# Variant annotation

## Use custom interval file for annotation

No bed dataset available.

(-interval)

## Only use the transcripts in this file

Nothing selected

Format is one transcript ID per line

## Filter output

Select/Unselect all

- 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

## Filter out specific Effects

No

# Variant annotation

## Chromosomal position

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

## Text to prepend to chromosome name

By default SnpEff simplifies all chromosome names. For instance 'chr1' is just '1'. You can prepend any string you want to the chromosome name (-chr)

## Produce Summary Stats



Yes

1

(-noStats)

## Suppress reporting usage statistics to server



Yes

(-noLog)

## Email notification



Send an email notification when the job completes.

✓ Execute

2

# Variant annotation - Content

## SnpEff: Variant analysis

### Contents

- [Summary](#)
- [Variant rate by chromosome](#)
- [Variants by type](#)
- [Number of variants by impact](#)
- [Number of variants by functional class](#)
- [Number of variants by effect](#)
- [Quality histogram](#)
- [InDel length histogram](#)
- [Base variant table](#)
- [Transition vs transversions \(ts/tv\)](#)
- [Allele frequency](#)
- [Allele Count](#)
- [Codon change table](#)
- [Amino acid change table](#)
- [Chromosome variants plots](#)
- [Details by gene](#)



2.04 GB

1

18: SnpEff eff: on data 16 - HTML stats	  
17: SnpEff eff: on data 16	  
16: freebayes_calling_nor m_filtered.vcf	  
father mother proband	
15: freebayes_calling_nor m.vcf	  
father mother proband	
14: freebayes_calling.vcf	  
father mother proband	
13: markdup_proband.ba	  

# Variant annotation - Summary

## Summary

<b>Genome</b>	hg19
<b>Date</b>	2022-03-25 11:34
<b>SnpEff version</b>	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
<b>Command line arguments</b>	SnpEff -i vcf -o vcf -stats /shared/ibfstor1/galaxy/jobs/001/469/1469180/outputs/galaxy_dataset_c7e86a06-3ffe-4324-9794-c54ffaf3b4c8.dat hg19 /shared/ibfstor1/galaxy/datasets/002/674/dataset_2674023.dat
<b>Warnings</b>	1,293
<b>Errors</b>	0
<b>Number of lines (input file)</b>	6,468
<b>Number of variants (before filter)</b>	6,468
<b>Number of not variants (i.e. reference equals alternative)</b>	0
<b>Number of variants processed (i.e. after filter and non-variants)</b>	6,468
<b>Number of known variants (i.e. non-empty ID)</b>	0 ( 0% )
<b>Number of multi-allelic VCF entries (i.e. more than two alleles)</b>	0
<b>Number of effects</b>	18,335
<b>Genome total length</b>	3,137,161,265
<b>Genome effective length</b>	146,364,022
<b>Variant rate</b>	1 variant every 22,628 bases

# Variant annotation - Variants details

## Variants rate details

Chromosome	Length	Variants	Variants rate
8	146,364,022	6,468	22,628
<b>Total</b>	<b>146,364,022</b>	<b>6,468</b>	<b>22,628</b>

## Number variants by type

Type	Total
SNP	5,101
MNP	132
INS	423
DEL	739
MIXED	73
INV	0
DUP	0
BND	0
INTERVAL	0
<b>Total</b>	<b>6,468</b>

## Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	322	1.756%
LOW	1,371	7.478%
MODERATE	807	4.401%
MODIFIER	15,835	86.365%

## Number of effects by functional class

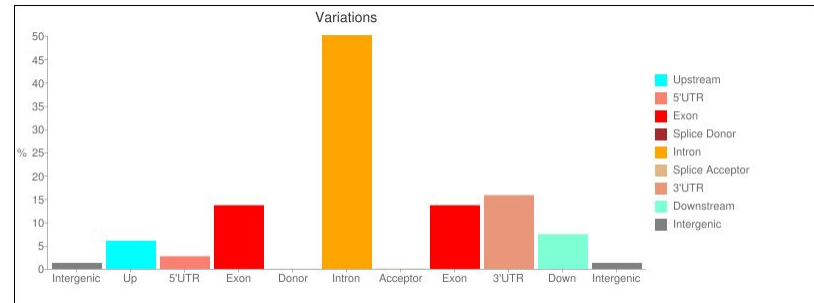
Type (alphabetical order)	Count	Percent
MISSENSE	743	45.667%
NONSENSE	4	0.246%
SILENT	880	54.087%

Missense / Silent ratio: 0.8443

# Variant annotation - Variants details

Type	Count	Percent
<b>Type (alphabetical order)</b>	<b>Count</b>	<b>Percent</b>
3_prime_UTR_variant	2,907	15.538%
5_prime_UTR_premature_start_codon_gain_variant	57	0.305%
5_prime_UTR_variant	440	2.352%
conservative_inframe_deletion	2	0.011%
conservative_inframe_insertion	4	0.021%
disruptive_inframe_deletion	5	0.027%
downstream_gene_variant	1,368	7.312%
frameshift_variant	7	0.037%
intergenic_region	236	1.261%
intragenic_variant	1	0.005%
intron_variant	9,544	51.013%
missense_variant	766	4.094%
non_coding_transcript_exon_variant	565	3.02%
non_coding_transcript_variant	2	0.011%
protein_protein_contact	6	0.032%
sequence_feature	135	0.722%
splice_acceptor_variant	13	0.069%
splice_donor_variant	3	0.016%
splice_region_variant	358	1.914%
start_lost	2	0.011%
stop_gained	7	0.037%
stop_lost	3	0.016%
stop_retained_variant	1	0.005%
structural_interaction_variant	284	1.518%
synonymous_variant	883	4.72%
upstream_gene_variant	1,110	5.933%

Type (alphabetical order)	Count	Percent
<b>DOWNSTREAM</b>	<b>1,368</b>	<b>7.461%</b>
<b>EXON</b>	<b>2,507</b>	<b>13.673%</b>
<b>INTERGENIC</b>	<b>236</b>	<b>1.287%</b>
<b>INTRON</b>	<b>9,209</b>	<b>50.226%</b>
<b>SPICE_SITE_ACCEPTOR</b>	<b>11</b>	<b>0.06%</b>
<b>SPICE_SITE_DONOR</b>	<b>3</b>	<b>0.016%</b>
<b>SPICE_SITE_REGION</b>	<b>349</b>	<b>1.903%</b>
<b>TRANSCRIPT</b>	<b>138</b>	<b>0.753%</b>
<b>UPSTREAM</b>	<b>1,110</b>	<b>6.054%</b>
<b>UTR_3_PRIME</b>	<b>2,907</b>	<b>15.855%</b>
<b>UTR_5_PRIME</b>	<b>497</b>	<b>2.711%</b>

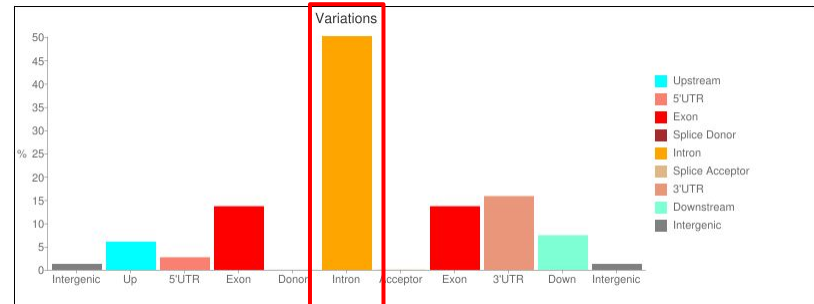




# Variant annotation - Variants details

Type	Count	Percent
<b>Type (alphabetical order)</b>	<b>Count</b>	<b>Percent</b>
3_prime_UTR_variant	2,907	15.538%
5_prime_UTR_premature_start_codon_gain_variant	57	0.305%
5_prime_UTR_variant	440	2.352%
conservative_inframe_deletion	2	0.011%
conservative_inframe_insertion	4	0.021%
disruptive_inframe_deletion	5	0.027%
downstream_gene_variant	1,368	7.312%
frameshift_variant	7	0.037%
intergenic_region	236	1.261%
intra-genic variant	1	0.005%
<b>intron_variant</b>	<b>9,544</b>	<b>51.013%</b>
missense_variant	766	4.094%
non_coding_transcript_exon_variant	565	3.02%
non_coding_transcript_variant	2	0.011%
protein_protein_contact	6	0.032%
sequence_feature	135	0.722%
splice_acceptor_variant	13	0.069%
splice_donor_variant	3	0.016%
splice_region_variant	358	1.914%
start_lost	2	0.011%
stop_gained	7	0.037%
stop_lost	3	0.016%
stop_retained_variant	1	0.005%
structural_interaction_variant	284	1.518%
synonymous_variant	883	4.72%
upstream_gene_variant	1,110	5.933%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,368	7.461%
EXON	2,507	13.673%
INTERGENIC	236	1.287%
<b>INTRON</b>	<b>9,209</b>	<b>50.226%</b>
SPLICE_SITE_ACCEPTOR	11	0.06%
SPLICE_SITE_DONOR	3	0.016%
SPLICE_SITE_REGION	349	1.903%
TRANSCRIPT	138	0.753%
UPSTREAM	1,110	6.054%
UTR_3_PRIME	2,907	15.855%
UTR_5_PRIME	497	2.711%



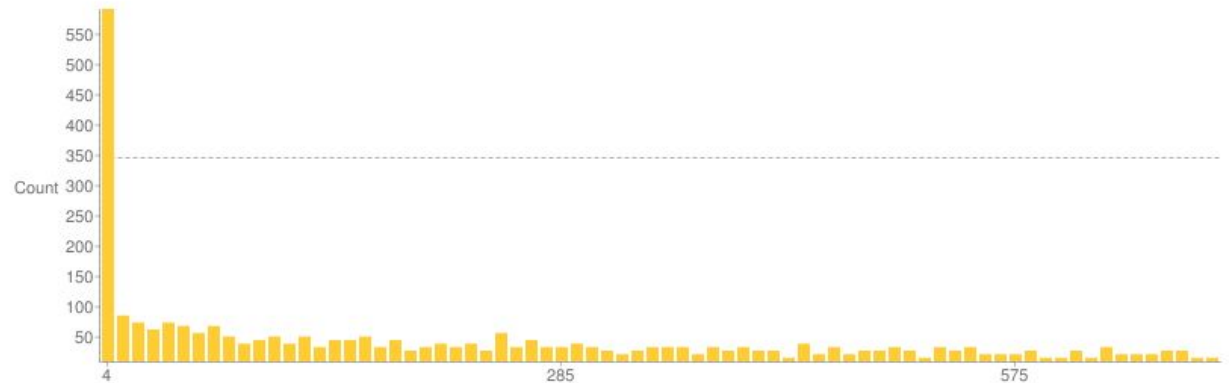


# Variant annotation - Variants quality

Quality:

Min 0  
Max 57,898  
Mean 1,449.862  
Median 691  
Standard deviation 2,384.312

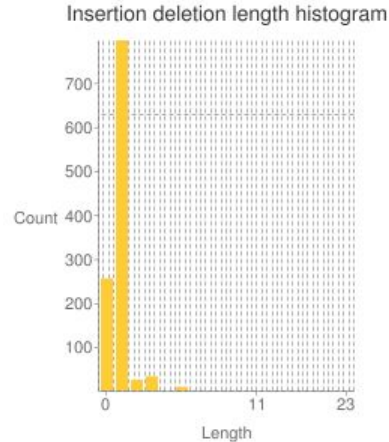
Values 0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37, :  
Count 456,23,14,22,14,14,6,12,16,14,14,14,9,7,7,11,5,8,12,13,9,8,12,10,10,8,4,3,9,3,6,7,8,7,8,6,8,6,9,10,12,10



# Variant annotation - Insertions/Deletions

## Insertions and deletions length:

<b>Min</b>	0
<b>Max</b>	23
<b>Mean</b>	1.104
<b>Median</b>	1
<b>Standard deviation</b>	1.693
<b>Values</b>	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 15, 17, 20, 21, 23
<b>Count</b>	259, 797, 31, 35, 7, 11, 5, 4, 2, 1, 3, 2, 1, 1, 1, 1, 1



# Variant annotation - Transitions/Transversions

## Base changes (SNPs)

	A	C	G	T
A	0	207	762	163
C	253	0	233	885
G	1,014	255	0	219
T	140	763	207	0

## Ts/Tv (transitions / transversions)

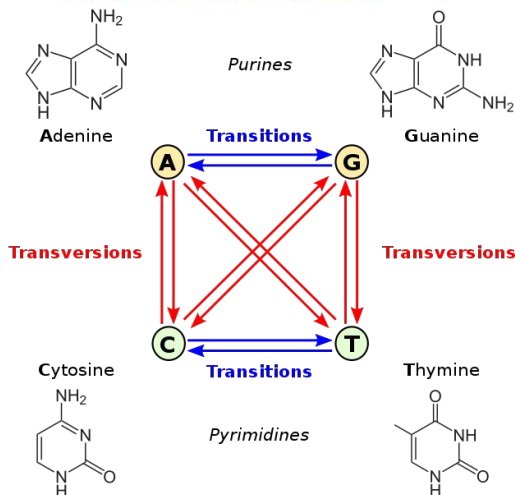
**Note:** Only SNPs are used for this statistic.

**Note:** This Ts/Tv ratio is a 'raw' ratio (ratio of observed events).

<b>Transitions</b>	8,638
<b>Transversions</b>	4,186
<b>Ts/Tv ratio</b>	2.0635

## All variants:

Sample	,	proband,	mother,	father,	Total
Transitions	,	2917,	2793,	2928,	8638
Transversions	,	1437,	1322,	1427,	4186
Ts/Tv	,	2.030,	2.113,	2.052,	2.064



Sequencing Type	# of Variants*	Tv/Tv Ratio
WGS	~4.4M	2.0-2.1
WES	~41k	3.0-3.3

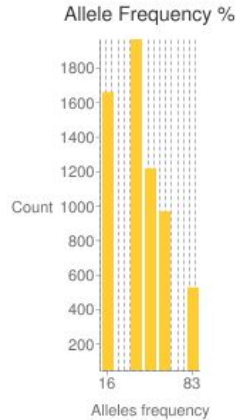
\*for a single sample

<https://en.wikipedia.org/wiki/Transversion>

<https://gatk.broadinstitute.org/hc/en-us/articles/360035531572-Evaluating-the-quality-of-a-germline-short-variant-callset>

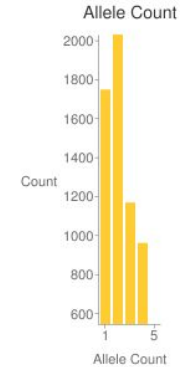
# Variant annotation - Allele details

## Allele frequency



<b>Min</b>	16
<b>Max</b>	83
<b>Mean</b>	41.217
<b>Median</b>	33
<b>Standard deviation</b>	21.155
<b>Values</b>	16,25,33,50,66,75,83
<b>Count</b>	1665,53,1965,1229,968,45,543

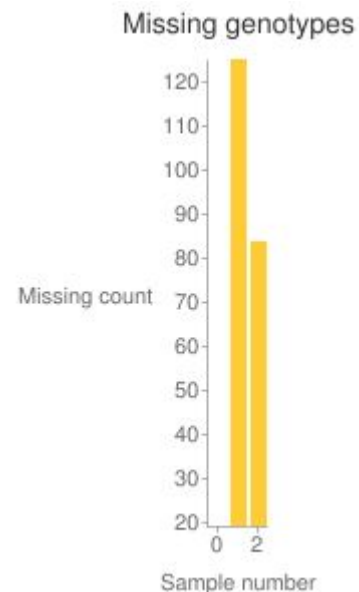
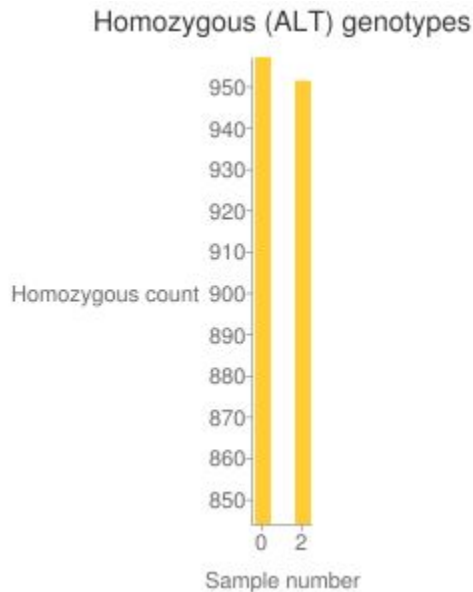
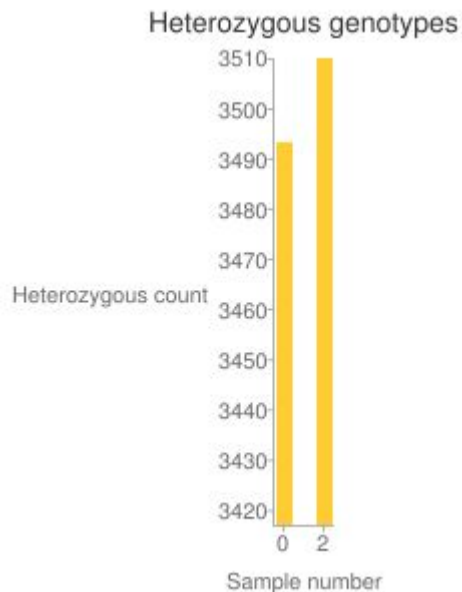
## Allele Count



<b>Min</b>	1
<b>Max</b>	5
<b>Mean</b>	2.462
<b>Median</b>	2
<b>Standard deviation</b>	1.262
<b>Values</b>	1,2,3,4,5
<b>Count</b>	1751,2029,1177,968,543

# Variant annotation - Genotypes details

## Hom/Het per sample



```
Sample_names , proband, mother, father
Reference , 1998, 2082, 1922
Het , 3494, 3417, 3510
Hom , 957, 844, 952
Missing , 19, 125, 84
```

# Variant annotation - Codon changes

## Codon changes

How to read this table:

- Rows are reference codons and columns are changed codons. E.g. Row 'AAA' column 'TAA' indicates how many 'AAA' codons have been replaced by 'TAA' codons.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	-	AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT	AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT	CAA	CAC	CAG
-	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;
AAA	1	&nbsp;	5	8	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	2	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;
AAC	2	3	&nbsp;	1	28	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	13	&nbsp;	3	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;

# Variant annotation - Amino acid changes

## Amino acid changes

How to read this table:

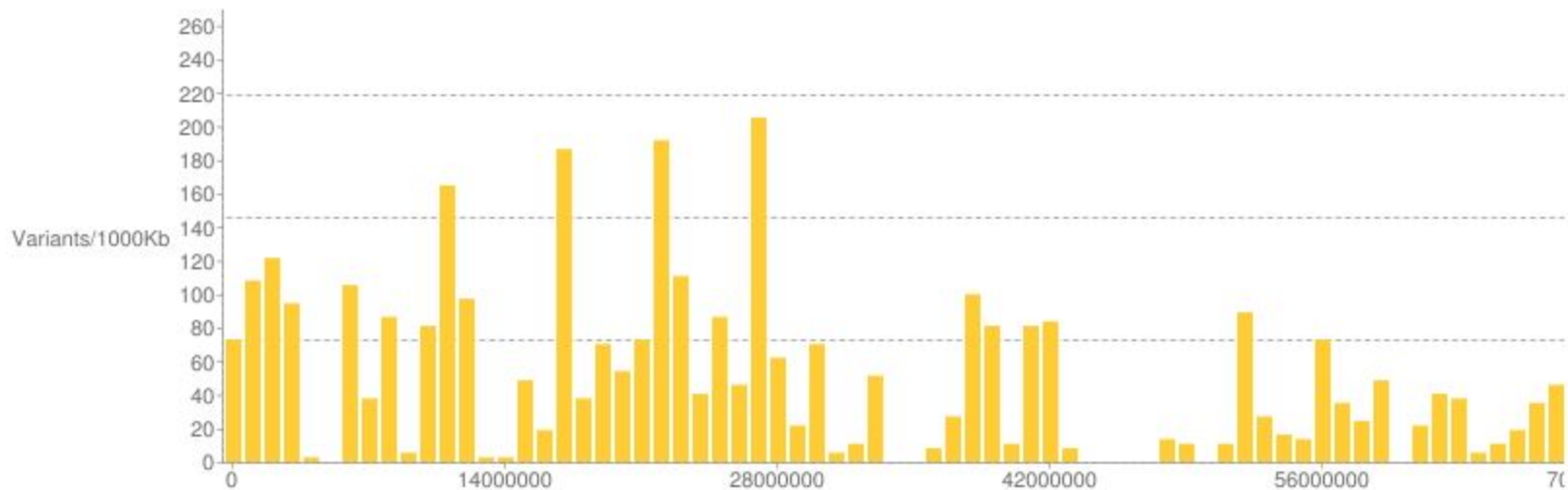
- Rows are reference amino acids and columns are changed amino acids. E.g. Row 'A' column 'E' indicates how many 'A' amino acids have been replaced by 'E' amino acids.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	*	-	?	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W
*	<b>1</b>	1	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	2	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;
-	&nbsp;	<b>&amp;nbsp;</b>	1	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;
?	&nbsp;	&nbsp;	<b>&amp;nbsp;</b>	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;
A	&nbsp;	1	&nbsp;	<b>166</b>	&nbsp;	1	1	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	6	23	33	&nbsp;
C	&nbsp;	3	&nbsp;	&nbsp;	<b>9</b>	&nbsp;	&nbsp;	&nbsp;	3	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	&nbsp;	5	&nbsp;	&nbsp;	&nbsp;	&nbsp;

# Variant annotation - Chromosomes details

**Variants by chromosome**

Var







# Variant annotation - ANN field

```
##SnEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"  
##SnEffCmd="SnEff -i vcf -o vcf -stats /shared/ibfstor1/galaxy/jobs/001/469/1469180/outputs/galaxy_dataset_c7e86a06-3ffe-4324-9794-c54ffaf3b4c8.dat hg19 /shared/ibfstor1/galaxy/datasets/002/674/dataset_  
##INFO=ID=ANN,Number=,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | c  
##INFO=ID=LOF,Number=,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected"  
##INFO=ID=NMD,Number=,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected">
```

'Allele | Annotation | Annotation\_Impact | Gene\_Name | Gene\_ID | Feature\_Type | Feature\_ID | Transcript\_BioType | Rank | HGVS.c | HGVS.p |  
cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO' ">

1

18: SnEff eff: on data 16  
- HTML stats

father mother proband



# Variant annotation - Examples

## Synonymous

```
ANN=G|synonymous_variant|LOW|OR4F21|OR4F21|transcript|NM_001005504.1|protein_coding|1/1|c.324T>C|p.Gly108Gly|324/939|324/939|108/312||
```

## Missense

```
ANN=G|missense_variant|MODERATE|FBXO25|FBXO25|transcript|NM_183421.1|protein_coding|3/11|c.138C>G|p.Ile46Met|404/2441|138/1104|46/367||
```

## Intronic

```
ANN=G|intron_variant|MODIFIER|FBXO25|FBXO25|transcript|NM_183421.1|protein_coding|1/10|c.-7-166C>G|||||
```

# Variant reporting - Pedigree

**Individual**

**Family ID**

**Father ID**

**Mother ID**

FAM	father	0	0	1	1
FAM	mother	0	0	2	1
FAM	proband	father	mother	1	2

4: Pedigree.txt



1

**Sex (1: male; 2: female)**

**Status (1: control; 2: case)**

# Variant reporting - Database creation

**1** gemini load

Upload Data

Show Sections

**2** **GEMINI load** Loading a VCF file into GEMINI

# Variant reporting - Database creation

 **GEMINI load** Loading a VCF file into GEMINI (Galaxy Version 0.20.1+galaxy2)  

## VCF dataset to be loaded in the GEMINI database

   17: freebayes\_calling\_norm\_filtered\_annotated.vcf   

Only build 37 (aka hg19) of the human genome is supported.

## The variants in this input are

annotated with snpEff 

GEMINI can parse and use annotations generated with either snpEff (both 'EFF'- and 'ANN'-style annotations are supported) or VEP. You can also load unannotated variants, but most of GEMINI's functionality will not be available or not be very useful without annotations. (-t)

## This input comes with genotype calls for its samples

Yes

This is usually the case, but some published datasets, like some 1000G VCFs, are missing genotype information. (--no-genotypes)

## Choose a gemini annotation source

GEMINI annotations w/ GERP & CADD (2022-03-23 snapshot) 

## Sample and family information in PED format

   4: Pedigree.txt   

The pedigree dataset is optional, but several GEMINI tools require the relationship between samples (i.e., the family structure) and/or the sample phenotype to be defined. The PED format is a simple tabular format (see the tool help below for details). If you choose to not provide sample information now, but later find that you need it for your analysis, you can also add it to an existing GEMINI database by using the GEMINI amend tool. (-p)

# Variant reporting - Database creation

## Load the following optional content into the database

Select/Unselect all

- GERP scores
- CADD scores (non-commercial use only; see licensing note below)
- Gene tables
- Sample genotypes
- Genotype likelihoods (sample PLs)
- only variants that passed all filters
- variant INFO field

5

The preselected defaults should be ok for most use cases (feel free to enable CADD scores for non-commercial use). If you are not interested in certain annotations, you can speed up database creation and decrease the resulting database size slightly by not loading them into the database. Note: GERP and CADD scores are optional parts of the annotation source and can only be loaded if available.

## Email notification



Send an email notification when the job completes.

Execute

6

# Variant reporting - Database creation

## Dataset Information

Number	19
Name	GEMINI load on data 4 and data 17
Created	Friday Mar 25th 2:37:11 2022 UTC
Filesize	190.8 MB
Dbkey	hg19
Format	gemini.sqlite
File contents	contents
History Content API ID	319b4d6eefbba9f5
History API ID	57e9be0d003985de
UUID	f41f617b-fc1c-4840-9ee4-cf206a5c4555

## Tool Parameters

Input Parameter	Value
VCF dataset to be loaded in the GEMINI database	17 freebayes_calling_norm_filtered_annotated.vcf father mother proband
The variants in this input are	annotated with snpEff
This input comes with genotype calls for its samples	True
Choose a gemini annotation source	2022-03-23
Sample and family information in PED format	4 Pedigree.txt
Load the following optional content into the database	GERP scores CADD scores (non-commercial use only; see licensing note below) Gene tables Sample genotypes variant INFO field

## Job Outputs

Tool Outputs	Dataset
--------------	---------

The screenshot shows a Galaxy job output window for the tool 'TP\_GTN\_WES\_disease'. The job title is '19: GEMINI load on data 4 and data 17'. Below the title, there are sample selection buttons for 'father', 'mother', and 'proband'. The output text shows the job progress: 'format: gemini.sqlite, database: hg19', 'Indexing', and 'Loading 6468 variants. Breaking into 12 chunks. Loading chunk 0. Loading chunk 1. Loading chunk 2. L'. At the bottom of the window, there is a red box around the 'G' icon in the toolbar, and a red arrow pointing to it from below. The version information at the bottom reads 'Gemini SQLite Database, version 0.20.1'.



# Variant reporting - Database content

Tools ☆ ☰

1  ✕

⬆️ Upload Data

👁️ Show Sections

**GEMINI query** Querying the GEMINI database

**GEMINI annotate** the variants in an existing GEMINI database with additional information

**GEMINI set\_somatic** Tag somatic mutations in a GEMINI database

**GEMINI amend** Amend an already loaded GEMINI database.

**GEMINI fusions** Identify somatic fusion genes from a GEMINI database

**GEMINI load** Loading a VCF file into GEMINI

2 **GEMINI database info** Retrieve information about tables, columns and annotation data stored in a GEMINI database

# Variant reporting - Database content


 **GEMINI database info** Retrieve information about tables, columns and annotation data stored in a GEMINI database (Galaxy Version 0.20.1)  

## GEMINI database

   19: GEMINI load on data 4 and data 17 

Only files with version 0.20.1 are accepted.

## Information to retrieve from the database

Names of database tables and their columns 

## Email notification



Send an email notification when the job completes.

 Execute 

# Variant reporting - Database content

table_name	column_name	type
variants	chrom	VARCHAR(20)
variants	start	INTEGER
variants	end	INTEGER
variants	vcf_id	TEXT
variants	variant_id	INTEGER
variants	anno_id	INTEGER
variants	ref	TEXT
variants	alt	TEXT
variants	qual	FLOAT
variants	filter	TEXT
variants	type	VARCHAR(20)
variants	sub_type	TEXT
variants	gts	BLOB
variants	gt_types	BLOB
variants	gt_phases	BLOB
variants	gt_depths	BLOB
variants	gt_ref_depths	BLOB
variants	gt_alt_depths	BLOB
variants	gt_alt_freqs	BLOB
variants	gt_quals	BLOB
variants	gt_copy_numbers	BLOB
variants	call_rate	FLOAT
variants	max_aaf_all	FLOAT
variants	in_dbsnp	BOOLEAN
variants	rs_ids	TEXT

variant_impacts	variant_id	INTEGER
variant_impacts	anno_id	INTEGER
variant_impacts	gene	VARCHAR(60)
variant_impacts	transcript	VARCHAR(60)
variant_impacts	is_exonic	BOOLEAN
variant_impacts	is_coding	BOOLEAN
variant_impacts	is_lof	BOOLEAN
variant_impacts	exon	TEXT
variant_impacts	codon_change	TEXT
variant_impacts	aa_change	TEXT
variant_impacts	aa_length	TEXT
variant_impacts	biotype	TEXT
variant_impacts	impact	VARCHAR(60)
variant_impacts	impact_so	TEXT
variant_impacts	impact_severity	VARCHAR(20)
variant_impacts	polyphen_pred	TEXT
variant_impacts	polyphen_score	FLOAT
variant_impacts	sift_pred	TEXT
variant_impacts	sift_score	FLOAT

# Variant reporting - Database content

samples	sample_id	INTEGER
samples	family_id	TEXT
samples	name	TEXT
samples	paternal_id	TEXT
samples	maternal_id	TEXT
samples	sex	TEXT
samples	phenotype	TEXT

gene_detailed	uid	INTEGER
gene_detailed	chrom	VARCHAR(60)
gene_detailed	gene	VARCHAR(60)
gene_detailed	is_hgnc	BOOLEAN
gene_detailed	ensembl_gene_id	TEXT
gene_detailed	transcript	VARCHAR(60)
gene_detailed	biotype	TEXT
gene_detailed	transcript_status	TEXT

gene_summary	uid	INTEGER
gene_summary	chrom	VARCHAR(60)
gene_summary	gene	VARCHAR(60)
gene_summary	is_hgnc	BOOLEAN
gene_summary	ensembl_gene_id	TEXT
gene_summary	hgnc_id	TEXT
gene_summary	transcript_min_start	INTEGER
gene_summary	transcript_max_end	INTEGER
gene_summary	strand	TEXT
gene_summary	synonym	TEXT

# Variant reporting - Querying

The screenshot shows the 'Tools' section of a software interface. At the top, there is a search bar containing the text 'gemini inheritance', which is highlighted with a red box and a red number '1'. Below the search bar are two buttons: 'Upload Data' and 'Show Sections'. A list of tools follows, each with a bold title and a description. The last tool in the list, 'GEMINI inheritance pattern based identification of candidate genes', is highlighted with a red box and a red number '2'.

**Tools** ☆ ☰

1 gemini inheritance ✕

⬆ Upload Data

👁 Show Sections

**GEMINI load** Loading a VCF file into GEMINI

**GEMINI query** Querying the GEMINI database

**GEMINI set\_somatic** Tag somatic mutations in a GEMINI database

**GEMINI amend** Amend an already loaded GEMINI database.

**GEMINI gene\_wise** Discover per-gene variant patterns across families

**GEMINI fusions** Identify somatic fusion genes from a GEMINI database

**GEMINI annotate** the variants in an existing GEMINI database with additional information

2 **GEMINI inheritance pattern** based identification of candidate genes

# Variant reporting - Querying

GEMINI inheritance pattern based identification of candidate genes (Galaxy Version 0.20.1)

**GEMINI database**

19: GEMINI load on data 4 and data 17

Only files with version 0.20.1 are accepted.

**Your assumption about the inheritance pattern of the phenotype of interest**

Autosomal recessive

Autosomal recessive

Autosomal dominant

X-linked recessive

X-linked dominant

Autosomal de-novo

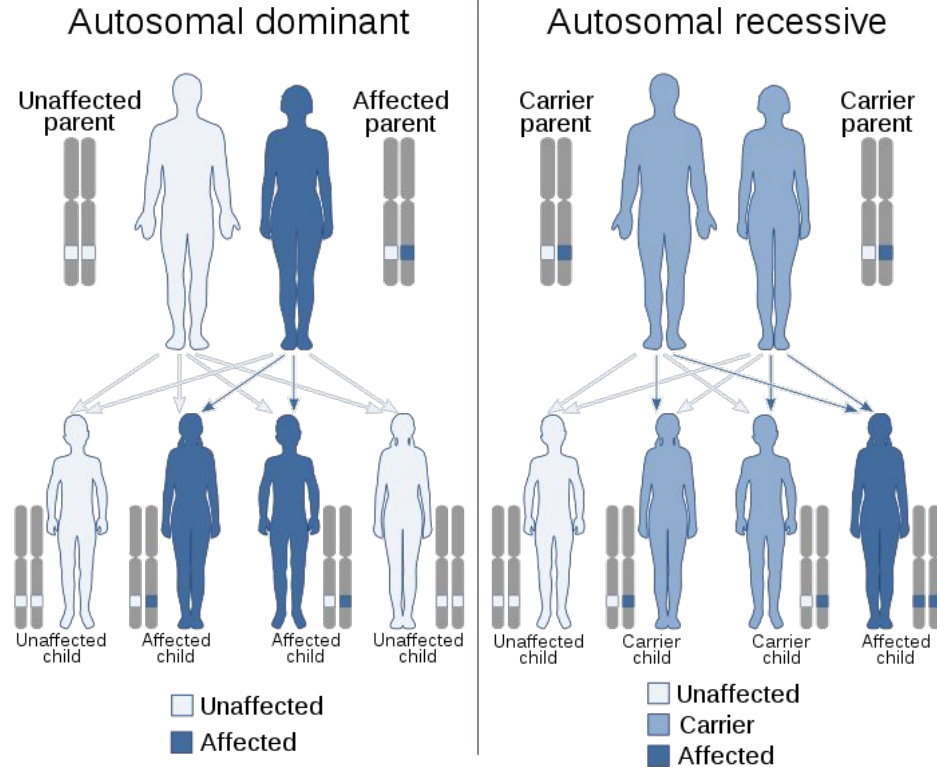
X-linked de-novo

Compound heterozygous

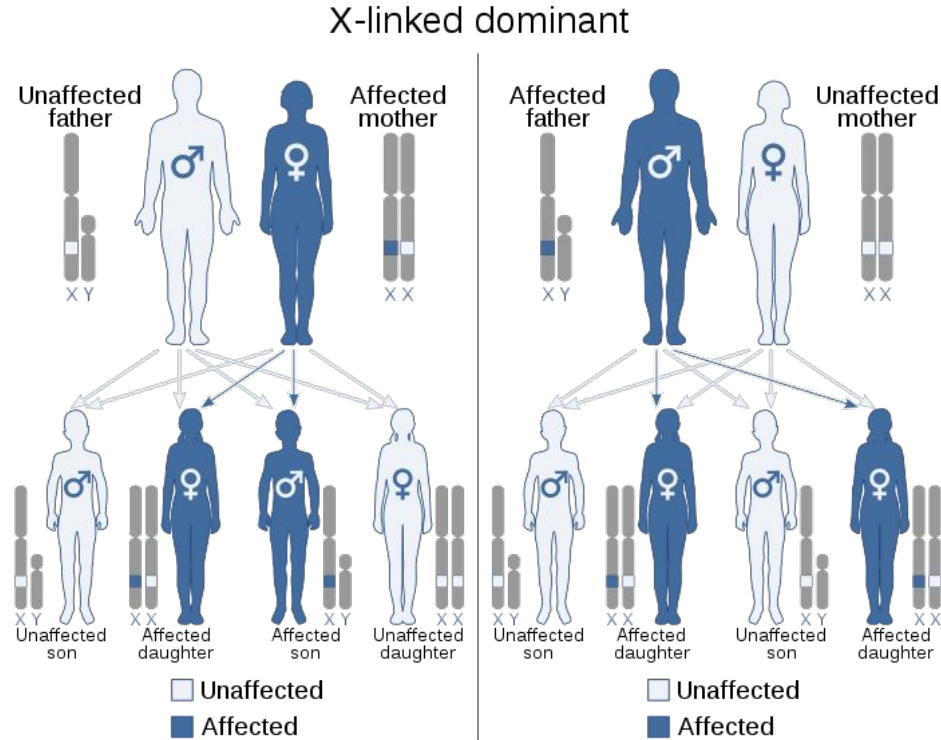
Violation of mendelian laws (LOH, plausible and implausible de-novo, uniparental disomy) samples. (--allow-unaffected)

**Which inheritance pattern to select ?**

# Variant reporting - Inheritance pattern



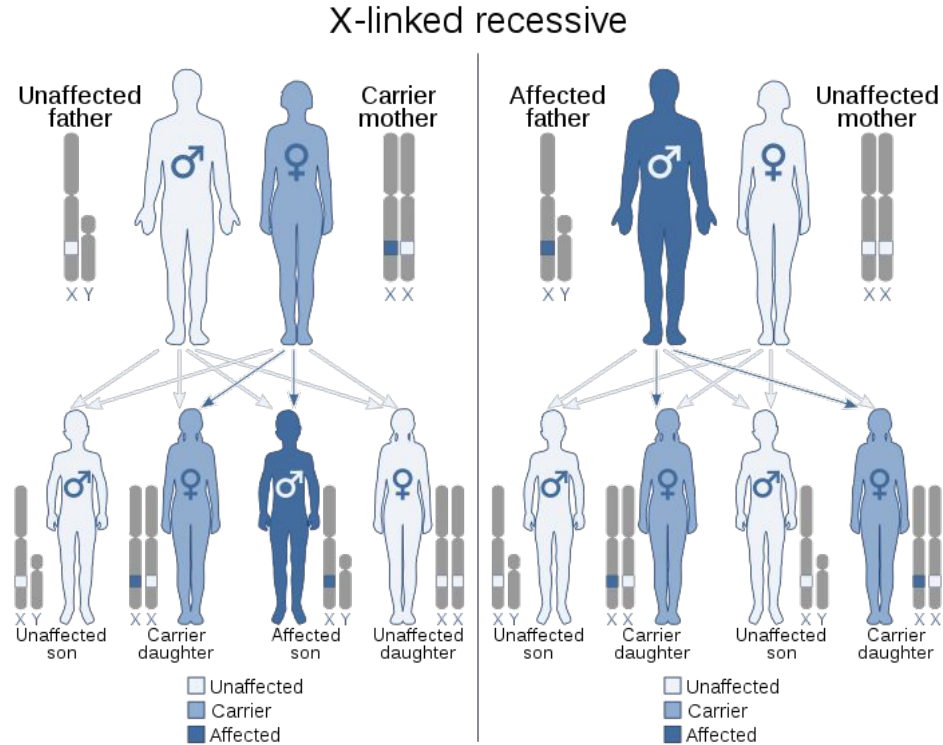
# Variant reporting - Inheritance pattern



Note: some X-linked dominant disorders are embryonic lethal in males, and most affect females less severely.



# Variant reporting - Inheritance pattern



Note: a few carriers may be mildly affected due to skewed X-inactivation.

# Variant reporting - Inheritance pattern

- Autosomal de-novo : mutation on autosomes (chr1-22), mutation not present in parents
- X-linked de-novo : mutation on the sex chromosome X, mutation not present in parents
- Compound heterozygous : 2 or more recessive alleles at a particular locus
- Violation of mendelian laws :
  - LOH : Loss of Heterozygosity, cross chromosomal event resulting in in loss of an entire gene and the surrounding chromosomal region
  - Plausible de-novo : parents are homozygous reference, offspring is heterozygous
  - Implausible de-novo : parents are homozygous reference, offspring is homozygous alternate
  - Uniparental disomy : one parent and the offspring are homozygous reference, the other parent is homozygous alternate OR one parent and the offspring are homozygous alternate and the other parent is homozygous reference

# Variant reporting - Inheritance pattern

- Autosomal recessive
- Autosomal dominant
- X-linked recessive
- X-linked dominant
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Violation of mendelian laws

# Variant reporting - Inheritance pattern

- Autosomal recessive
- Autosomal dominant
- X-linked recessive
- X-linked dominant
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

# Variant reporting - Inheritance pattern

- Autosomal recessive
- Autosomal dominant
- X-linked recessive
- X-linked dominant
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

# Variant reporting - Inheritance pattern

- Autosomal recessive
- Autosomal dominant
- X-linked recessive
- X-linked dominant
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**

# Variant reporting - Inheritance pattern

- Autosomal recessive
- ~~Autosomal dominant~~
- X-linked recessive
- ~~X-linked dominant~~
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**

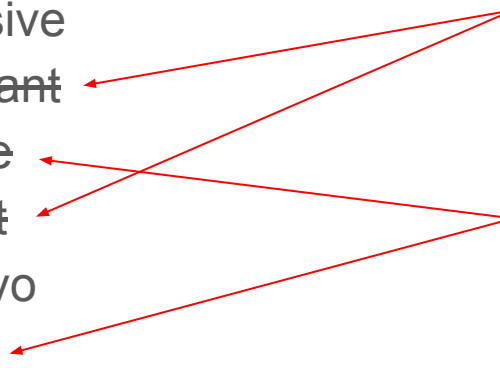
# Variant reporting - Inheritance pattern

- Autosomal recessive
- ~~Autosomal dominant~~
- ~~X-linked recessive~~
- ~~X-linked dominant~~
- Autosomal de-novo
- ~~X-linked de-novo~~
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**





# Variant reporting - Inheritance pattern

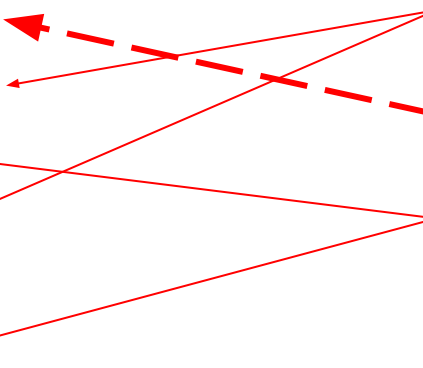
1

- Autosomal recessive
- ~~Autosomal dominant~~
- ~~X-linked recessive~~
- ~~X-linked dominant~~
- Autosomal de-novo
- ~~X-linked de-novo~~
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**



# Variant reporting - Inheritance pattern

1

- Autosomal recessive
- ~~Autosomal dominant~~
- ~~X-linked recessive~~
- ~~X-linked dominant~~

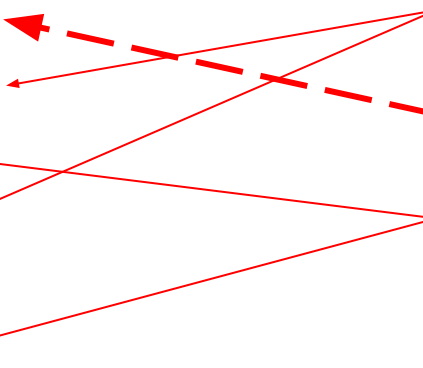
2

- Autosomal de-novo
- ~~X-linked de-novo~~
- Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**



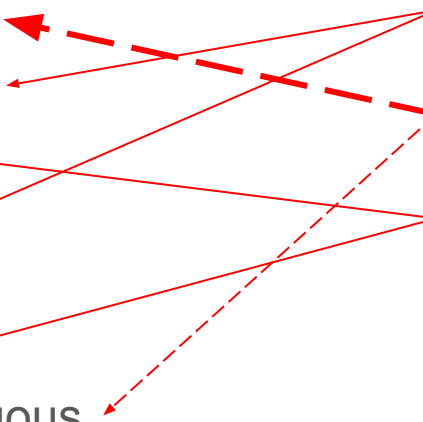
# Variant reporting - Inheritance pattern

- 1 ● Autosomal recessive
- ~~Autosomal dominant~~
- ~~X-linked recessive~~
- ~~X-linked dominant~~
- 2 ● Autosomal de-novo
- ~~X-linked de-novo~~
- 3 ● Compound heterozygous
- Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**



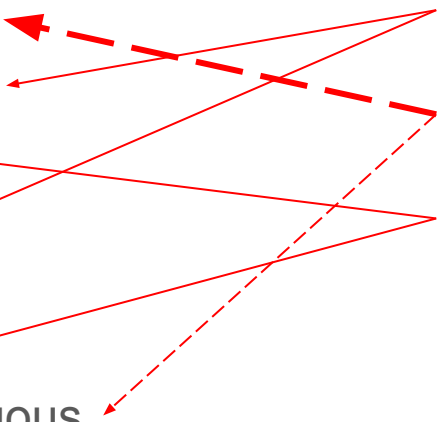
# Variant reporting - Inheritance pattern

- 1 ● Autosomal recessive
- ~~Autosomal dominant~~
- ~~X-linked recessive~~
- ~~X-linked dominant~~
- 2 ● Autosomal de-novo
- ~~X-linked de-novo~~
- 3 ● Compound heterozygous
- 4 ● Violation of mendelian laws

**Parents are unaffected**

**Parents are consanguineous**

**Chromosome 8**



# Variant reporting - Querying

 GEMINI inheritance pattern based identification of candidate genes (Galaxy Version 0.20.1)  

GEMINI database


   19: GEMINI load on data 4 and data 17   

Only files with version 0.20.1 are accepted.

Your assumption about the inheritance pattern of the phenotype of interest

Autosomal recessive 

Additional constraints on variants

 Insert Additional constraints on variants

Additional constraints on variants

1: Additional constraints on variants 

Additional constraints expressed in SQL syntax

impact\_severity != 'LOW'

Constraints defined here will become the WHERE clause of the SQL query issued to the GEMINI database. E.g. alt='G' or impact\_severity = 'HIGH'. (--filter)

# Variant reporting - Querying

## Include hits with less convincing inheritance patterns

No

The exact consequence of this setting depends on the type of inheritance pattern you are looking for (see the tool help below). (--lenient)

## Report candidates shared by unaffected samples

No

Activating this option will enable the reporting of variants as candidate causative even if they are shared by unaffected samples in the family tree. The default will only report variants that are unique to affected samples. (--allow-unaffected)

## Family-wise criteria for variant selection

### Minimum number of families with a candidate variant for a gene to be reported


This is the number of families required to have a variant fitting the inheritance model in the same gene in order for the gene and its variants to be reported. For example, we may only be interested in candidates where at least 4 families have a variant (with a fitting inheritance pattern) in that gene. (--min-kindreds)

### List of families to restrict the analysis to (comma-separated)

Leave empty for an analysis including all families (--families)

### Specify additional criteria to exclude families on a per-variant basis

# Variant reporting - Querying

Output - included information 

**Set of columns to include in the variant report table**

5

Custom (report user-specified columns)

The tool reports key information about the inheritance pattern detection for each candidate variant found. It can precede each such row with additional columns, listing information about the variant taken from the variants table of the GEMINI database. Here, you can control which subset of the variants table columns should be added to the output.

**Choose columns to include in the report**

Select/Unselect all

- gene
- chrom
- start
- end
- ref
- alt
- impact
- impact\_severity

alternative allele frequency (max\_aaf\_all)

6

(--columns)

**Additional columns (comma-separated)**

chrom,start,ref,alt,impact,gene,clinvar\_sig,clinvar\_disease\_name,clinvar\_gene\_phenotype,rs\_ids

7

Column must be specified by the exact name they have in the GEMINI database, e.g., is\_exonic or num\_hom\_alt, but, for genotype columns, GEMINI wildcard syntax is supported. The order of columns in the list is maintained in the output.

# Variant reporting - Querying

## Additional columns (comma-separated)

chrom,start,ref,alt,impact,gene,clinvar\_sig,clinvar\_disease\_name,clinvar\_gene\_phenotype,rs\_ids

Column must be specified by the exact name they have in the GEMINI database, e.g., is\_exonic or num\_hom\_alt, but, for genotype columns, GEMINI wildcard syntax is supported. The order of columns in the list is maintained in the output.

## Email notification



Send an email notification when the job completes.





✓ Execute



8






# Variant reporting - Results

max_aaf_all	chrom	start	ref	alt	impact	gene	clinvar_sig	clinvar_disease_name
0.6831	chr8	2048830	A	G	missense_variant	MYOM2	None	None
0.6716	chr8	6479041	C	T	missense_variant	MCPH1	benign	Primary_autosomal_recessive_microcephaly_1 not_specified Primary_Microcephaly
0.93555555555556	chr8	6681255	A	C	splice_region_variant	XKR5	None	None
-1.0	chr8	11666217	GTCCCAC	G	conservative_inframe_deletion	FDFT1	None	None
0.7798	chr8	12878806	T	G	missense_variant	KIAA1456	None	None
0.8221	chr8	12879098	G	A	missense_variant	KIAA1456	None	None
0.8221	chr8	12879538	A	G	missense_variant	KIAA1456	None	None
0.8313	chr8	17434640	G	C	splice_region_variant	PDGFRL	None	None
0.847026781661	chr8	17743019	G	A	missense_variant	FGL1	None	None
-1.0	chr8	17796381	AC	GT	missense_variant	PCM1	None	None
0.842472840145	chr8	17814914	A	G	missense_variant	PCM1	None	None




**History**    

search datasets  

**TP\_GTN\_WES\_disease**

21 shown   

2.23 GB

**21: GEMINI autosomal\_recessive pattern on data 19**   

**1**

**father** **mother** **proband**

# Variant reporting - Results

clinvar\_gene\_phenotype

None

primary\_microcephaly\&x2c\_recessive|primary\_autosomal\_recessive\_microcephaly\_1

None

None

None

None

None

carcinoma\_of\_colon

# Variant reporting - Results

rs_ids	variant_id	family_id	family_members	family_genotypes	samples	family_count
rs968381	228	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	G/G,A/G,A/G	proband	1
rs1057090	462	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	T/T,C/T,C/T	proband	1
rs9772979	490	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	C/C,A/C,A/C	proband	1
rs71711801	862	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	G/G,GTCCCAC/G,GTCCCAC/G	proband	1
rs3739310	936	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	G/G,T/G,T/G	proband	1
rs545589847,rs502882	939	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	A/A,G/A,G/A	proband	1

# Variant reporting - Results

**Most likely variant  
candidate for child's  
disease ?**

# Variant reporting - Results

max_aaf_all	chrom	start	ref	alt	impact	gene	clinvar_sig	clinvar_disease_name
3.24886289799e-05	chr8	86385979	G	A	stop_gained	CA2	None	None

clinvar\_gene\_phenotype

carbonic\_anhydrase\_ii\_variant|osteopetrosis\_with\_renal\_tubular\_acidosis

rs_ids	variant_id	family_id	family_members	family_genotypes
None	3883	FAM	proband(proband;affected;male),mother(mother;unaffected;female),father(father;unaffected;male)	A/A,G/A,G/A