Exome sequencing data analysis for diagnosing a genetic disease

Galaxy Training! tutorial

https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/exome-seq/tutorial.html

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

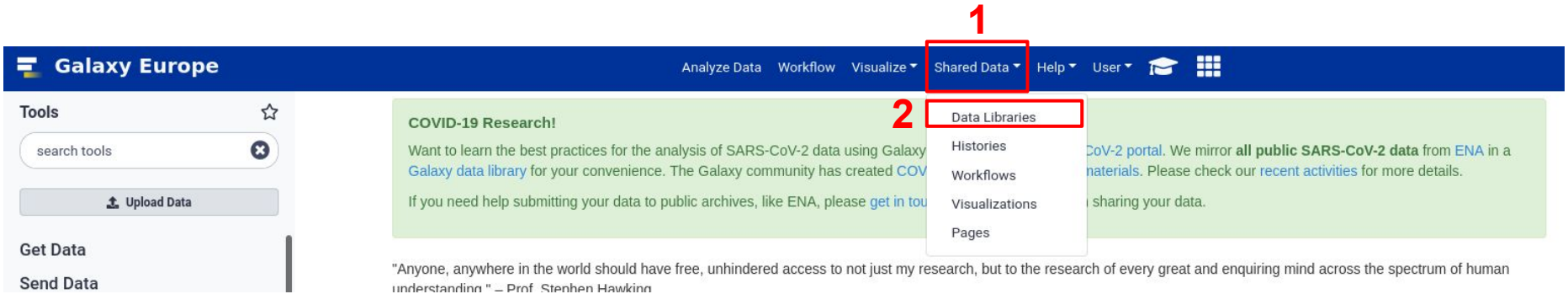
Premapped reads

- 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



Premapped reads upload

Galaxy Europe

Analyze Data | Workflow | Visualize | Shared Data | Help | User

Search

Name	Description	Synopsis
covid-19	Public data related to COVID-19	RAW sequence data & VCFs
Earth System Community Modeling	Input data used for running simulation w...	Earth System Community Modeling data
Galaxy courses	Data for Galaxy courses	
Genomes + annotations	Reference genomes and gene annotations	
GTN - Material	Galaxy Training Network Material	Galaxy Training Network Material. See ht...
PGP-UK Open Access Data	Open Access Genomic, Transcriptomic and ...	https://www.personalgenomes.org.uk

« 0 1 2 » 20 per page, 6 total

Premapped reads upload

Galaxy Europe

Analyze Data Workflow Visualize Shared Data Help User

Libraries / GTN - Material

Search | Export to History | Download | Delete | Details | include deleted

Name	Description
Assembly	
ChIP-Seq data analysis	
Ecology	
Epigenetics	
Genome Annotation	
Imaging	
Introduction to Galaxy Analyses	
Metabolomics	
Metagenomics	
Proteomics	
RNA interactome	RNA interactome data analysis
Sequence analysis	
Statistics and machine learning	
The new topic	Summary
Transcriptomics	

« ‹ 1 2 › » 15 per page, 18 total

Premapped reads upload

Galaxy Europe

Analyze Data | Workflow | Visualize | Shared Data | Help | User

Libraries / GTN - Material

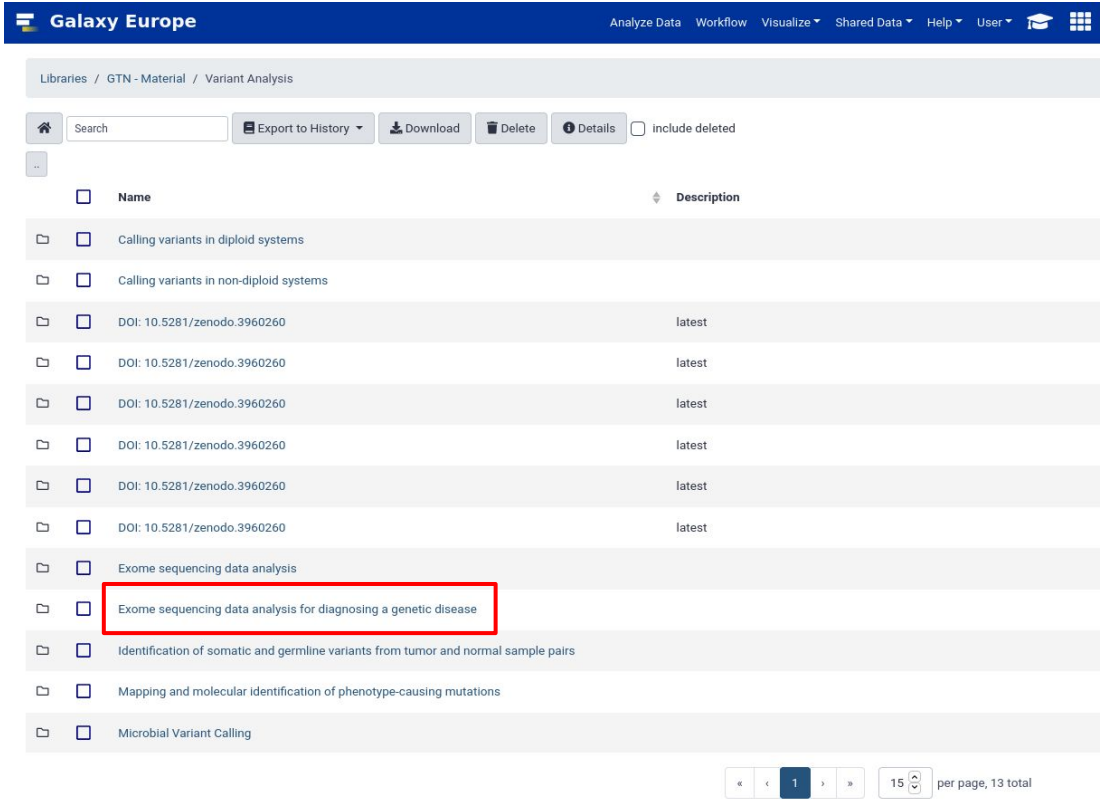
Search | Export to History | Download | Delete | Details | include deleted

..

	Name	Description	Type
	User Interface and Features		folder
	Variant Analysis		folder
	Visualisation		folder

« ‹ 1 2 › » 15 per page, 18 total

Premapped reads upload



Premapped reads upload

Galaxy Europe

Analyze Data Workflow Visualize Shared Data Help User

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

Search | Export to History | Download | Delete | Details | include deleted

..

	Name	Description	Type
	DOI: 10.5281/zenodo.3054169		folder

« ‹ 1 › » 15 per page, 1 total

Premapped reads upload

Libraries / GTN - Material / Variant Analysis / Exome sequencing data analysis for diagnosing a genetic disease / DOI: 10.5281/zenodo.3054169

2

Search | Export to History | Download | Delete | Details | include deleted

3

as Datasets

as a Collection

..

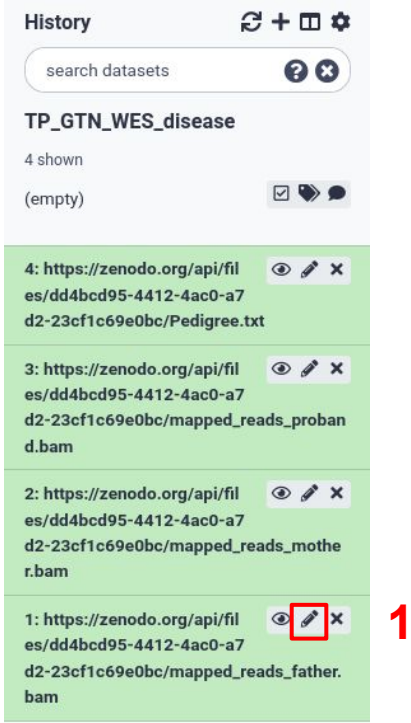
	Name	Description	Type	Size
	https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/hg19_chr8.fa.gz	uploaded fasta file	fasta	142.4 MB
☑	https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_father.bam	uploaded bam file	bam	336.9 MB
☑	https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_mother.bam	uploaded bam file	bam	296.1 MB
☑	https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/mapped_reads_proband.bam	uploaded bam file	bam	391.6 MB
☑	https://zenodo.org/api/files/dd4bcd95-4412-4ac0-a7d2-23cf1c69e0bc/Pedigree.txt	uploaded tabular file	tabular	68 b

1

« ‹ 1 › » 15 per page, 5 total

Premapped reads upload

Premapped reads upload



Premapped reads upload

Edit dataset attributes

Attributes | Convert | Datatypes | Permissions

Edit attributes

Auto-detect | Save

Name

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

----- Additional Species Are Below -----

Premapped reads upload

Edit dataset attributes

Attributes | Convert | Datatypes | Permissions

5

Edit attributes

2 - Use self-explanatory names

Auto-detect | Save

Name

mapped_reads_father.bam

1

Info

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

---- Additional Species Are Below ----

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)

GRCh37.p2 Aug. 2009 (GRCh37.p2/hg19Patch2) (hg19Patch2)

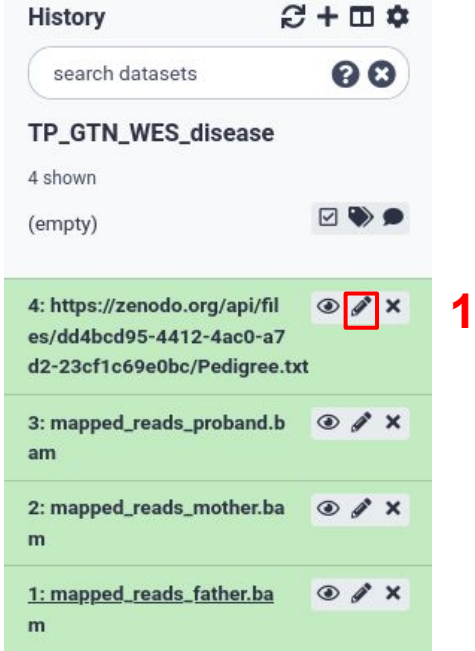
Hg19 lgg inversion Mar 2013 (1/hg19LggInv) (hg19LggInv)

Homo sapiens (hg19 with mtDNA replaced with rCRS) (Homo_sapiens_nuHg19_mtrCRS)

4

Human Feb. 2009 (GRCh37/hg19) (hg19)

Premapped reads upload



Premapped reads upload

Edit dataset attributes

Attributes | Convert | Datatypes | Permissions

Edit attributes

Auto-detect | Save

3

2

Name

Pedigree.txt

1

Info

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

Annotation

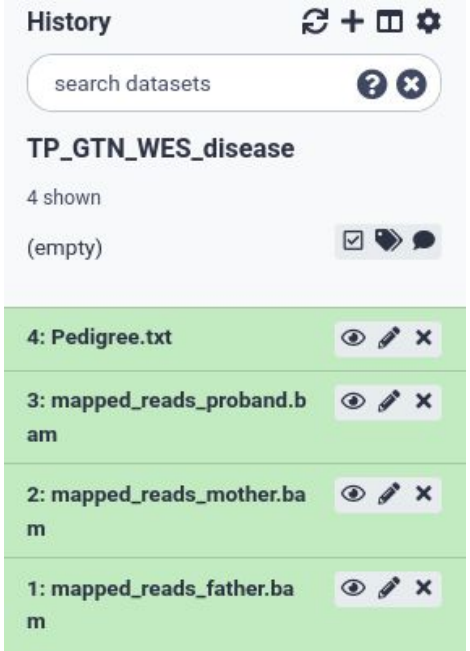
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build

----- Additional Species Are Below -----

Number of comment lines

Premapped reads upload



Mapped reads postprocessing

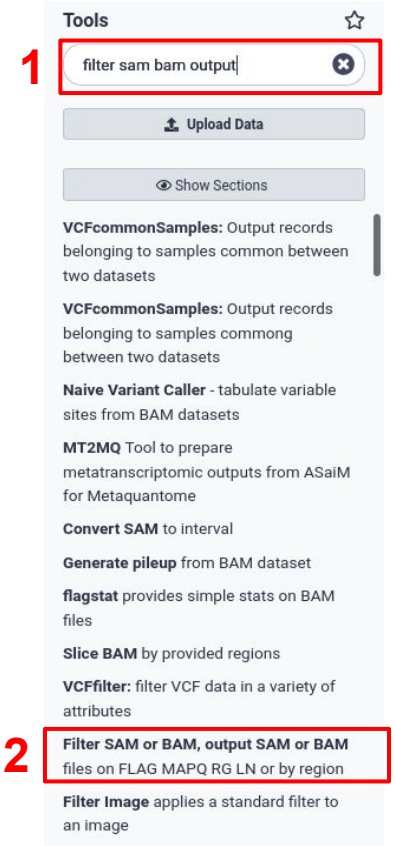
Warning :

- 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



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)

Favorite | Versions | Options

SAM or BAM file to filter

2 - Hold Ctrl key

1

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.

Header in output

Include header

Minimum MAPQ quality score

(-q)

Filter on bitwise flag

yes

3

Mapped reads postprocessing - Filter reads

Skip alignments with any of these flag bits set

Select/Unselect all

- [ ] Read is paired
- [ ] Read is mapped in a proper pair
- [x] The read is unmapped
- [x] 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

(-F)

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)

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

No

Send an email notification when the job completes.

2

Execute

Mapped reads postprocessing - Filter reads

Filter SAM or BAM, output SAM or BAM

Tool Parameters

Input Parameter	Value
SAM or BAM file to filter	• 1: mapped_reads_father.bam
Header in output	Include header
Minimum MAPQ quality score	Not available.
Filter on bitwise flag	yes
Only output alignments with all of these flag bits set	Nothing selected.
Skip alignments with any of these flag bits set	The read is unmapped The mate is unmapped
Select alignments from Library	Empty.
Select alignments from Read Group	Empty.
Output alignments overlapping the regions in the BED file	
Use inverse selection	False
Select regions (only used when the input is in BAM format)	Empty.
Select the output format	BAM (-b)

Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/devteam/samtool_filter2/samtool_filter2/1.8+galaxy1
Command Line	*empty*
Tool Standard Output	*empty*
Tool Standard Error	*empty*
Tool Exit Code:	0
Job API ID:	11ac94870d0bb33a16e2d37293526d54

1

History

search datasets

TP_GTN_WES_disease

7 shown

1 GB

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

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

3

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

337.2 MB

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

2

display at UCSC main

display at Ensembl Current

display with IGV local Human hg19

display in IGB View

display at bam.iobio bam.iobio.io

Binary bam alignments file

Mapped reads postprocessing - Filter reads

Edit dataset attributes

Attributes | Convert | Datatypes | Permissions

2

Edit attributes

Auto-detect | Save

1

Name

filtered_reads_father.bam

Info

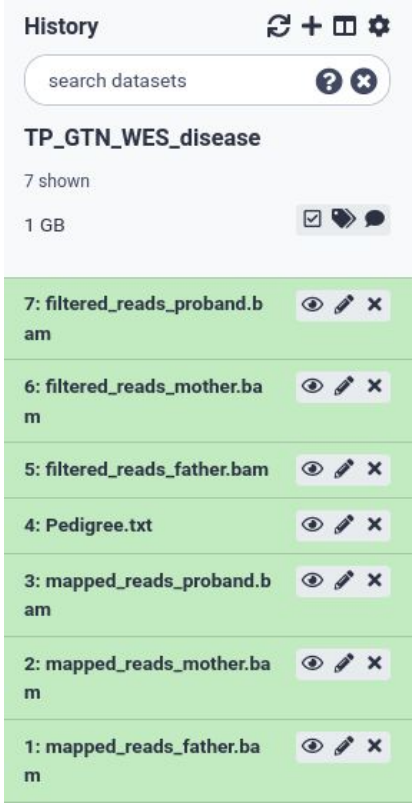
Annotation

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

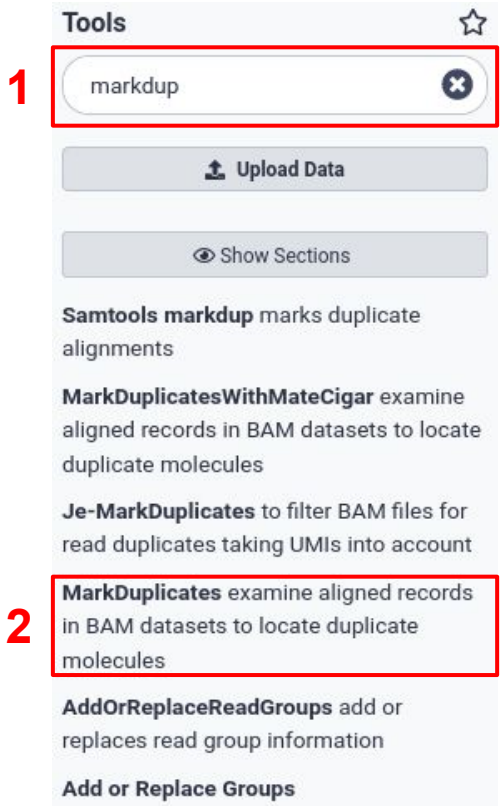
Database/Build

Human Feb. 2009 (GRCh37/hg19) (hg19)

Mapped reads postprocessing - Filter reads



Mapped reads postprocessing - Duplicate reads



Mapped reads postprocessing - Duplicate reads

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.2)

Favorite | Versions | Options

Select SAM/BAM dataset or dataset collection

7: filtered_reads_proband.bam

If empty, upload or import a SAM/BAM dataset

Comment

+ Insert 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

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

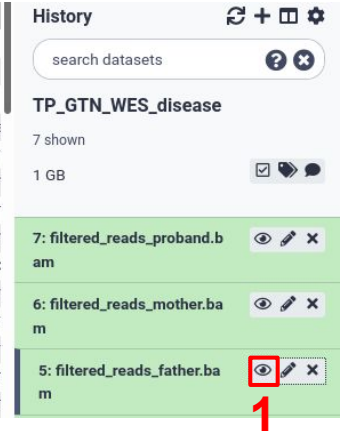
How can we know ?

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	AGCCACGTCTCCCCAGGTCAGTCTTAAGGACAACGAAACTCTGGGC
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	CCAGATCCCAAACCCTGATCCCTACCCTGGATCCTAAGTCTGTCCCT
DCW97JN1:309:C0C42ACXX:5:2210:15831:85655	145	chr8	98822	0	52S35M14S	=	110566976	110468121	TTTTAAAATTTAAAAAAAAAAAATTGGCCAAAAAAATTTTATTTTTTT
DCW97JN1:309:C0C42ACXX:4:2209:3455:67435	161	chr8	98823	0	45S43M13S	=	39494954	39396232	CCCCAAAAAAAATTTCGGGGTTTTGGGTTTTTTCCACCCAAAATTTT



Illumina read format :

- <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)

https://support.basespace.illumina.com/articles/descriptive/fastq-files/

Mapped reads postprocessing - Duplicate reads

MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.2)

Favorite | Versions | Options

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

2

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

+ Insert 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

4 - Depends on goal and pipeline

3

Assume the input file is already sorted

Yes

ASSUME_SORTED; default=True

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 duplicte clusters in order to consider them optical duplicates

100

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

No

Send an email notification when the job completes.

Execute

6

Mapped reads postprocessing - Duplicate reads

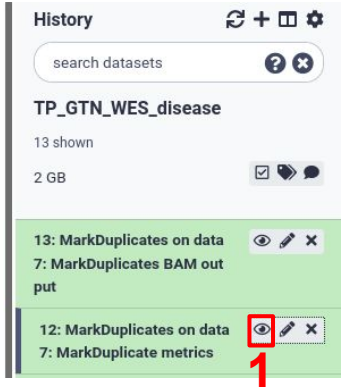
```
## htsjdk.samtools.metrics.StringHeader
# MarkDuplicates INPUT=[filtered_reads_proband_bam] OUTPUT=/data/dnb03/galaxy_db/job_working_directory/015/738/15738941/outputs/galaxy_dataset_206994b9-d2b3-44ef-82e2-
d697790e2d19.dat METRICS_FILE=/data/dnb03/galaxy_db/job_working_directory/015/738/15738941/outputs/galaxy_dataset_d56642f1-be33-4f35-963f-a0cdb9399ce0.dat REMOVE_DUPLICATES=false
ASSUME_SORTED=true DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUALITIES OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 TMP_DIR=[/data/dnb03/galaxy_db/job_working_directory/015/738/15738941/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 TAGGING_POLICY=DontTag 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_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json USE_JDK_DEFLATER=false USE_JDK_INFLATER=false
## htsjdk.samtools.metrics.StringHeader
# Started on: Sat Mar 20 17:35:44 CET 2021

## 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       0       781643  244     0.328394        2777843

## HISTOGRAM    java.lang.Double
BIN     VALUE
1.0     1.000065
2.0     1.424589
3.0     1.604798
4.0     1.681296
```

| Header

Unmapped reads

Duplicates & Optical duplicates

Percentage duplication



Mapped reads postprocessing - Duplicate reads

QNAME	FLAG	RNAME	POS
@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\tPL:ILLUMINA localref.fa /data/dnb02/galaxy_db/files/009/499/dataset_9499701.dat /data/dnb02/galaxy_db/files/009/499/data			
@PG ID:MarkDuplicates VN:2.18.2-SNAPSHOT CL:MarkDuplicates INPUT=[filtered_reads_father_bam] OUTPUT=/data/dnb03/galaxy_db/job_working_directory/015/738/15738939/outputs/galaxy_dataset_8ee3c20e-e7			
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

History

search datasets

TP_GTN_WES_disease

13 shown

2 GB

13: MarkDuplicates on data 7: MarkDuplicates BAM output

12: MarkDuplicates on data 7: MarkDuplicate metrics

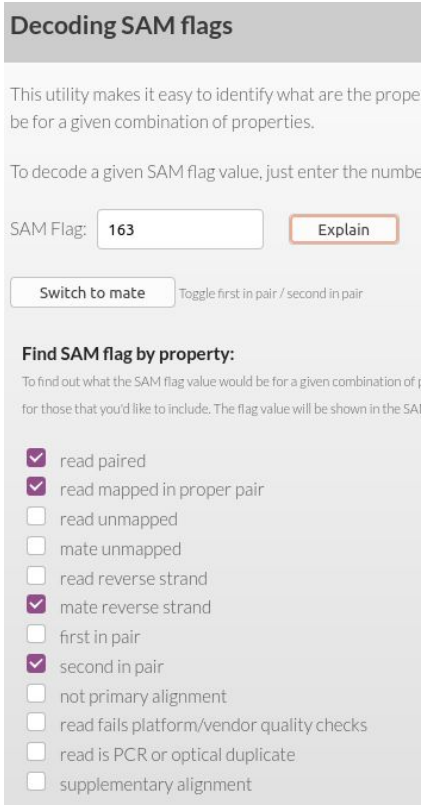
11: MarkDuplicates on data 6: MarkDuplicates BAM output

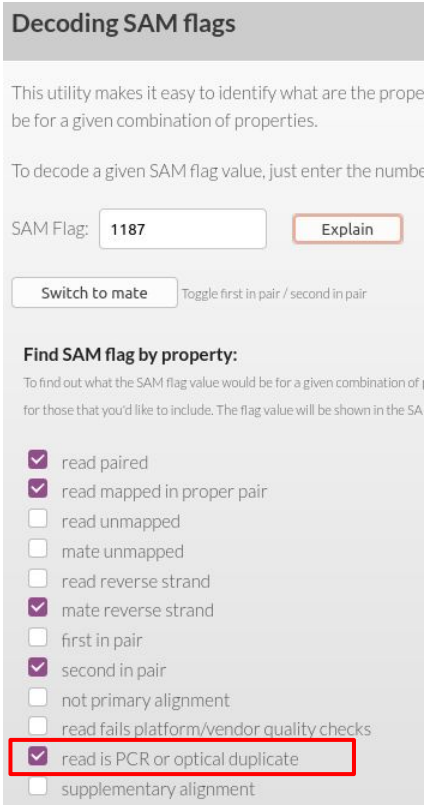
10: MarkDuplicates on data 6: MarkDuplicate metrics

9: MarkDuplicates on data 5: MarkDuplicates BAM output

1

Mapped reads postprocessing - Duplicate reads





http://broadinstitute.github.io/picard/explain-flags

Mapped reads postprocessing - Duplicate reads

Edit dataset attributes

Attributes updated.

Attributes | Convert | Datatypes | Permissions

Edit attributes

Auto-detect | Save

2

Name

markdup_proband.bam

1

Info

Picked up _JAVA_OPTIONS: -Xmx12G -Xms1G -Djava.io.tmpdir=/data/2/galaxy_db/tmp
17:35:44.460 INFO NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/usr/local/tools/_conda/envs/__picard@2.18.2/share/picard-2.18.2-0/picard.jar!/com/intel/g

Annotation

History

search datasets

TP_GTN_WES_disease

13 shown

2 GB

3

13: markdup_proband.bam

12: markdup_proband_metrics

11: markdup_mother.bam

10: markdup_mother_metrics

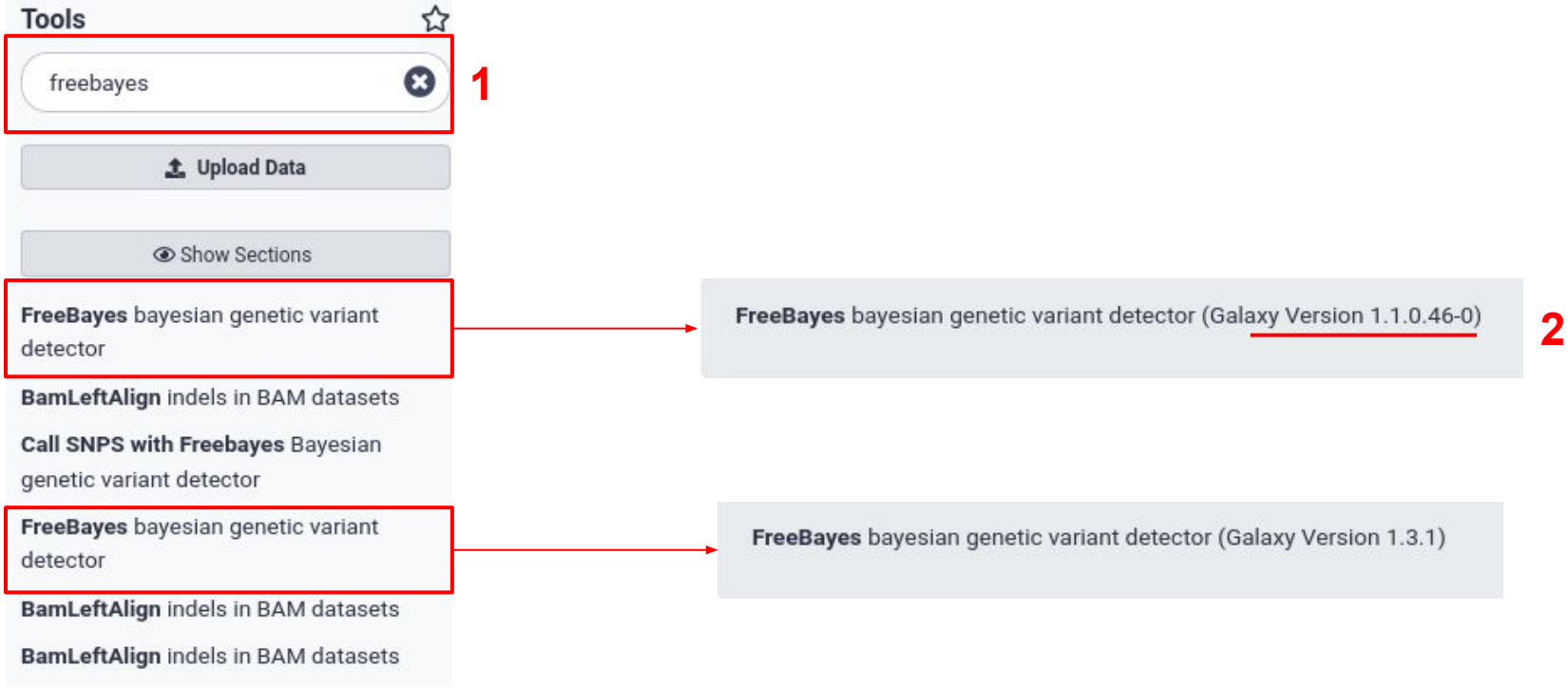
9: markdup_father.bam

8: markdup_father_metrics

Tutorial steps

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

Variant calling



Variant calling

FreeBayes bayesian genetic variant detector (Galaxy Version 1.1.0.46-0)

Favorite | Versions | Options

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 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

3

Human (Homo sapiens): hg19

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

1. Simple diploid calling

1

Select how much control over the freebayes run you need

Email notification

No

Send an email notification when the job completes.

Execute

2

Variant calling

FreeBayes

Tool Parameters

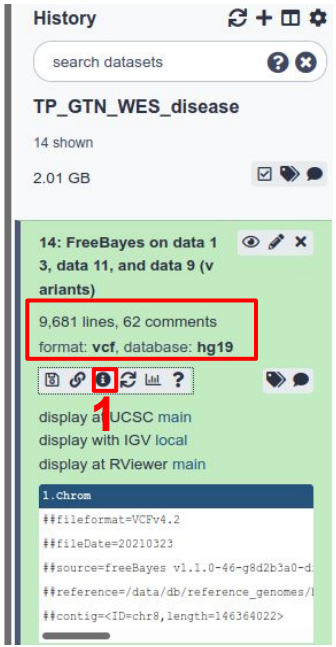
Input Parameter	Value
Choose the source for the reference genome	cached
Run in batch mode?	merge
BAM dataset(s)	• 9: markdup_father.bam • 11: markdup_mother.bam • 13: markdup_proband.bam
Using reference genome	hg19
Limit variant calling to a set of regions?	do_not_limit
Choose parameter selection level	simple

Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/devteam/freebayes/freebayes/1.1.0.46-0
Command Line	*empty*
Tool Standard Output	*empty*
Tool Standard Error	*empty*
Tool Exit Code:	0
Job API ID:	11ac94870d0bb33ac051a6b8d251e8ea

Dataset Storage

This dataset is stored in a Galaxy object store with id **files10**.



Variant calling - VCF

Chrom	Pos	ID	Ref	Alt	Qual	Filter	Info
##fileformat=VCFv4.2							
##fileDate=20210323							
##source=freeBayes v1.1.0-46-g8d2b3a0-dirty							
##reference=/data/db/reference_genomes/hg19/seq/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 /data/db/reference_genomes/hg19/seq/hg19.fa --vcf ./vcf_output/part_chr8:0..146364022.vcf"							
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">							
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">							
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype">							
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">							
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">							
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">							
##INFO=<ID=RO,Number=1,Type=Integer,Description="Count of full observations of the reference haplotype.">							
##INFO=<ID=AO,Number=A,Type=Integer,Description="Count of full observations of this alternate haplotype.">							
##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded fractionally">							
##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">							
##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">							
##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">							
##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">							
##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">							
##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">							

History

search datasets

TP_GTN_WES_disease

14 shown

2.01 GB

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

1

9,681 lines, 62 comments

format: **vcf**, database: **hg19**

display at UCSC main

display with IGV local

display at RViewer main

1.Chrom

```
##fileformat=VCFv4.2
##fileDate=20210323
##source=freeBayes v1.1.0-46-g8d2b3a0-d
##reference=/data/db/reference_genomes/
##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	11713	.	C	T	28.9882	.
chr8	11737	.	C	T	9.49649	.
chr8	11780	.	T	A	0.00632346	.
chr8	11793	.	C	G	0.00516241	.
chr8	11922	.	T	C	1.26638	.
chr8	11935	.	T	C	1.25509	.
chr8	11953	.	A	C	3.16611	.
chr8	116079	.	G	A	100.779	.
chr8	116701	.	A	G	8.91211e-09	.
chr8	116895	.	A	G	214.71	.
chr8	160552	.	G	A	3.87014	.
chr8	160608	.	A	C	722.504	.
chr8	160609	.	AA	AAAATA	0.0812872	.

Variant calling - VCF

Mandatory column

INFO
AB=0;ABP=0;AC=4;AF=1;AN=4;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=19.3333;MQMR=0;NS=2;NUMALT=1;ODDS=6.67245;PAIRED=1;PAIREDR=0
AB=0;ABP=0;AC=4;AF=0.666667;AN=6;AO=3;CIGAR=1X;DP=4;DPB=4;DPRA=1.5;EPP=3.73412;EPPR=5.18177;GTI=0;LEN=1;MEANALT=1;MQM=19.3333;MQMR=27;NS=3;NUMALT=1;ODDS=2.05301;PAIRE
AB=0;ABP=0;AC=2;AF=0.333333;AN=6;AO=2;CIGAR=1X;DP=4;DPB=4;DPRA=2;EPP=7.35324;EPPR=3.0103;GTI=0;LEN=1;MEANALT=1;MQM=9;MQMR=33.5;NS=3;NUMALT=1;ODDS=7.2724;PAIRED=1;PAI
AB=0.666667;ABP=3.73412;AC=1;AF=0.166667;AN=6;AO=2;CIGAR=1X;DP=5;DPB=5;DPRA=3;EPP=7.35324;EPPR=3.73412;GTI=0;LEN=1;MEANALT=1;MQM=9;MQMR=31.3333;NS=3;NUMALT=1;ODDS=7.4
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=9.52472;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=8;MQMR=0;NS=1;NUMALT=1;ODDS=1.08321;PAIRED=1;PAIREDR=0;PAO=(
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=9.52472;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=8;MQMR=0;NS=1;NUMALT=1;ODDS=1.09352;PAIRED=1;PAIREDR=0;PAO=(
AB=0;ABP=0;AC=4;AF=1;AN=4;AO=4;CIGAR=1X;DP=4;DPB=4;DPRA=0;EPP=5.18177;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=12.75;MQMR=0;NS=2;NUMALT=1;ODDS=0.0698374;PAIRED=1;PAIREDR=0
AB=0.28;ABP=24.0302;AC=2;AF=0.333333;AN=6;AO=17;CIGAR=1X;DP=78;DPB=78;DPRA=0;EPP=18.4661;EPPR=25.259;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=4.70393;P
AB=0.217391;ABP=18.9659;AC=1;AF=0.166667;AN=6;AO=15;CIGAR=1X;DP=267;DPB=267;DPRA=0;EPP=3.15506;EPPR=7.24817;GTI=0;LEN=1;MEANALT=2;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=2
AB=0;ABP=0;AC=6;AF=1;AN=6;AO=16;CIGAR=1X;DP=20;DPB=20;DPRA=0;EPP=3.55317;EPPR=5.18177;GTI=0;LEN=1;MEANALT=1.66667;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=10.518;PAIRED=1;P
AB=0.307692;ABP=7.18621;AC=2;AF=0.333333;AN=6;AO=5;CIGAR=1X;DP=22;DPB=22;DPRA=0;EPP=3.44459;EPPR=3.13803;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=59.5294;NS=3;NUMALT=1;ODDS

Variant calling - VCF

FORMAT	mother	father	proband
GT:DP:AD:RO:QR:AO:QA:GL	1/1:1:0,1:0:0:1:30:-2.95865,-0.30103,0	1/1:2:0,2:0:0:2:62:-1.69419,-0.60206,0	.
GT:DP:AD:RO:QR:AO:QA:GL	1/1:1:0,1:0:0:1:25:-2.48652,-0.30103,0	1/1:2:0,2:0:0:2:66:-1.70351,-0.60206,0	0/0:1:1,0:1:28:0:0:0,-0.30103,-2.44648
GT:DP:AD:RO:QR:AO:QA:GL	0/0:1:1,0:1:33:0:0:0,-0.30103,-3.22103	1/1:2:0,2:0:0:2:69:-1.70487,-0.60206,0	0/0:1:1,0:1:35:0:0:0,-0.30103,-2.63623
GT:DP:AD:RO:QR:AO:QA:GL	0/0:1:1,0:1:39:0:0:0,-0.30103,-3.64612	0/1:3:1,2:1:28:2:71:-0.802831,0,-1.54339	0/0:1:1,0:1:31:0:0:0,-0.30103,-2.55471
GT:DP:AD:RO:QR:AO:QA:GL	.	1/1:3:0,3:0:0:3:117:-2.23884,-0.90309,0	.
GT:DP:AD:RO:QR:AO:QA:GL	.	1/1:3:0,3:0:0:3:101:-2.23436,-0.90309,0	.
GT:DP:AD:RO:QR:AO:QA:GL	.	1/1:3:0,3:0:0:3:120:-2.23908,-0.90309,0	1/1:1:0,1:0:0:1:2:-0.199493,-0.30103,0
GT:DP:AD:RO:QR:AO:QA:GL	0/0:28:25,3:25:1021:3:66:0,-2.27016,-85.998	0/1:25:19,6:19:730:6:168:-7.87004,0,-58.47	0/1:25:17,8:17:646:8:260:-16.1934,0,-50.9025
GT:DP:AD:RO:QR:AO:QA:GL	0/0:123:114,6:114:3446:6:195:0,-18.3908,-292.603	0/1:23:18,5:18:378:5:178:-9.4362,0,-27.2972	0/0:121:116,4:116:3482:4:144:0,-22.901,-300.366
GT:DP:AD:RO:QR:AO:QA:GL	1/1:5:0,5:0:0:5:54:-4.96641,-1.50515,0	1/1:4:0,2:0:0:2:35:-3.2067,-0.60206,0	1/1:11:1,9:1:2:9:177:-15.8854,-2.8303,0

Genotypes format | **Mother genotypes information** | **Father genotypes information** | **Proband genotypes information**

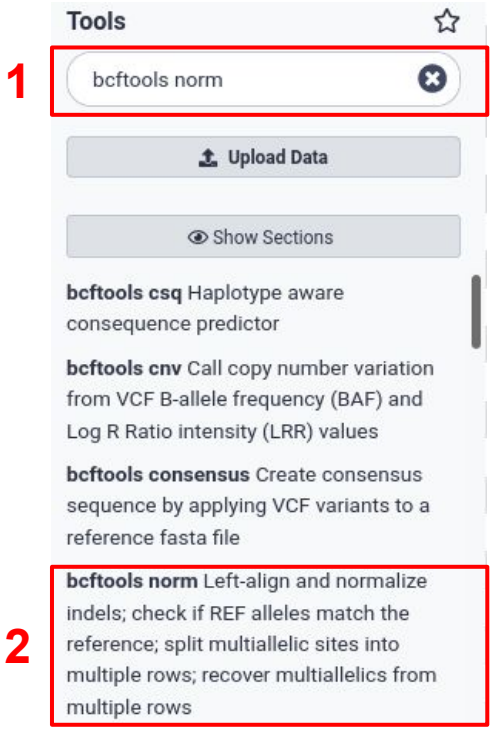
Tutorial steps

1. Perform postprocessing from premapped reads

2. Variant calling

3. Variant annotation and reporting

Variant normalization



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 1.9+galaxy1)

Favorite | Versions | Options

VCF/BCF Data

14: FreeBayes on data 13, data 11, and data 9 (variants) **1**

Choose the source for the reference genome

Use a built-in genome

Reference genome

Human (Homo sapiens): hg19 **2**

When any REF allele does not match the reference genome base

- (●) 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)

3

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?

[x] Yes **4**

Variant normalization

1

Perform deduplication for the folowing 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

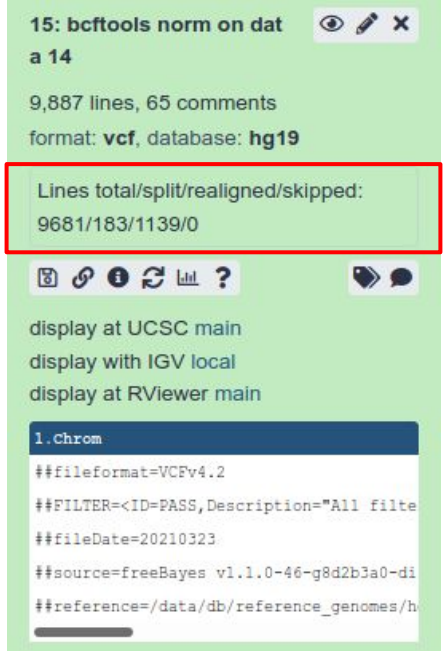
No

Send an email notification when the job completes.

4

✓ Execute

Variant normalization



Variant normalization - Initial file (alleles)

chr8	160552	.	G	A	3.87014	.	AB=0.307692;ABP=7.18621;AC=2;AF=0
chr8	160608	.	A	C	722.504	.	AB=0.4375;ABP=5.72464;AC=3;AF=0.5;
chr8	160609	.	AA	AAAATA	0.0812872	.	AB=0.210526;ABP=16.8392;AC=1;AF=0
chr8	160736	.	G	T	525.434	.	AB=0.191686;ABP=360.521;AC=3;AF=0
chr8	160826	.	C	T	6351.88	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=231
chr8	161062	.	C	T	2356.87	.	AB=0.632979;ABP=31.8863;AC=3;AF=0
chr8	161167	.	C	T	118.178	.	AB=0.179688;ABP=117.08;AC=3;AF=0.5
chr8	161176	.	T	C	111.716	.	AB=0.210526;ABP=85.9834;AC=3;AF=0
chr8	161240	.	A	G	168.568	.	AB=0.4;ABP=5.18177;AC=4;AF=0.6666
chr8	162973	.	T	C	4.89955	.	AB=0.25;ABP=7.35324;AC=1;AF=0.166
chr8	163226	.	T	C	8650.8	.	AB=0.493002;ABP=3.28384;AC=2;AF=0
chr8	163249	.	T	C	8433.27	.	AB=0.495327;ABP=3.13206;AC=2;AF=0
chr8	163302	.	CATATATG	CATATG	18744.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=663
chr8	163366	.	TAGAC	CAGAG,TAGAG	30471.4	.	AB=0.404658,0.58952;ABP=57.2529,50
chr8	163387	.	C	T	10073.7	.	AB=0.587644;ABP=49.4474;AC=2;AF=0
chr8	163419	.	GAA	AGT	6732.12	.	AB=0.277477;ABP=241.712;AC=4;AF=0

1

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

9,681 lines, 62 comments

format: **vcf**, database: **hg19**

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

```
1.Chrom
##fileformat=VCFv4.2
##fileDate=20210323
##source=freeBayes v1.1.0-46-g8d2b3a0-d
##reference=/data/db/reference_genomes/
##contig=<ID=chr8,length=146364022>
```

Variant normalization - Normalized file (alleles)

chr8	160552	.	G	A	3.87014	.	AB=0.307692;ABP=7.18621;AC=2;AF=0.333333;AN=6;AO=5;
chr8	160608	.	A	C	722.504	.	AB=0.4375;ABP=5.72464;AC=3;AF=0.5;AN=6;AO=35;CIGAR=
chr8	160609	.	A	AAAAT	0.0812872	.	AB=0.210526;ABP=16.8392;AC=1;AF=0.166667;AN=6;AO=7;
chr8	160736	.	G	T	525.434	.	AB=0.191686;ABP=360.521;AC=3;AF=0.5;AN=6;AO=83;CIGA
chr8	160826	.	C	T	6351.88	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=231;CIGAR=1X;DP=236;
chr8	161062	.	C	T	2356.87	.	AB=0.632979;ABP=31.8863;AC=3;AF=0.5;AN=6;AO=119;CIG
chr8	161167	.	C	T	118.178	.	AB=0.179688;ABP=117.08;AC=3;AF=0.5;AN=6;AO=23;CIGAR
chr8	161176	.	T	C	111.716	.	AB=0.210526;ABP=85.9834;AC=3;AF=0.5;AN=6;AO=24;CIGA
chr8	161240	.	A	G	168.568	.	AB=0.4;ABP=5.18177;AC=4;AF=0.666667;AN=6;AO=11;CIGA
chr8	162973	.	T	C	4.89955	.	AB=0.25;ABP=7.35324;AC=1;AF=0.166667;AN=6;AO=2;CIGA
chr8	163226	.	T	C	8650.8	.	AB=0.493002;ABP=3.28384;AC=2;AF=0.333333;AN=6;AO=31
chr8	163249	.	T	C	8433.27	.	AB=0.495327;ABP=3.13206;AC=2;AF=0.333333;AN=6;AO=31
chr8	163302	.	CAT	C	18744.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=663;CIGAR=1M2D5M;DP
chr8	163366	.	TAGAC	CAGAG	30471.4	.	AB=0.404658;ABP=57.2529;AC=4;AF=0.666667;AN=6;AO=55
chr8	163370	.	C	G	30471.4	.	AB=0.58952;ABP=50.8301;AC=2;AF=0.333333;AN=6;AO=424
chr8	163387	.	C	T	10073.7	.	AB=0.587644;ABP=49.4474;AC=2;AF=0.333333;AN=6;AO=40
chr8	163419	.	GAA	AGT	6732.12	.	AB=0.277477;ABP=241.712;AC=4;AF=0.666667;AN=6;AO=29

1

15: bcftools norm on data 14

9,887 lines, 65 comments

format: **vcf**, database: **hg19**

Lines total/split/realigned/skipped: 9681/183/1139/0

display at UCSC main

display with IGV local

display at RViewer main

1.chrom

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filt
##fileDate=20210323
##source=freeBayes v1.1.0-46-g8d2b3a0-d
##reference=/data/db/reference_genomes/
```

Variant normalization - Initial file (genotypes)

chr8	163302	.	CATATATG	CATATG	18744.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=
chr8	163366	.	TAGAC	CAGAG,TAGAG	30471.4	.	AB=0.404658,0.58952;ABP=57.2529,
chr8	163387	.	C	T	10073.7	.	AB=0.587644;ABP=49.4474;AC=2;AF
chr8	163419	.	GAA	AGT	6732.12	.	AB=0.277477;ABP=241.712;AC=4;AF
chr8	163432	.	A	G	9010.58	.	AB=0.346416;ABP=123.071;AC=4;AF
chr8	163438	.	C	T	21195.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=
chr8	163550	.	AAGT	GAGC,GAGT	12279.9	.	AB=0.346801,0.639731;ABP=63.555
chr8	163654	.	C	T	6081.47	.	AB=0.401487;ABP=25.6856;AC=4;AF
chr8	163784	.	C	G	4144.35	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=
chr8	169366	.	T	C	10888.4	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=
chr8	169403	.	G	A	0.000305373	.	AB=0.220183;ABP=77.1392;AC=1;AF
chr8	169476	.	C	G	404.305	.	AB=0.21875;ABP=112.941;AC=2;AF=
chr8	169483	.	T	G	341.779	.	AB=0.2;ABP=128.087;AC=2;AF=0.33
chr8	169641	.	A	G	62.9338	.	AB=0.75;ABP=5.18177;AC=3;AF=0.5
chr8	181859	.	G	A	24.1605	.	AB=0.75;ABP=5.18177;AC=1;AF=0.2

1

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

9,681 lines, 62 comments

format: **vcf**, database: **hg19**

display at UCSC main

display with IGV local

display at RViewer main

```
1.Chrom
##fileformat=VCFv4.2
##fileDate=20210323
##source=freeBayes v1.1.0-46-g8d2b3a0-d
##reference=/data/db/reference_genomes/
##contig=<ID=chr8,length=146364022>
```

1/2:120:0,40,77:0:0:40,77:1480,2744:-347.312,-223.79,-210.846,-124.443,0,-100.661

1/1:117:0,117,0:0:0:117,0:4126,0:-375.053,-38.2308,0,-377.701,-38.2308,-375.292

Mother

Father

Proband

1/2:177:0,63,113:0:0:63,113:2341,4406:-551.213,-362.959,-340.977,-192.495,0,-155.458

Variant normalization - Normalized file (genotypes)

chr8	163302	.	CAT	C	18744.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=663;CIGAR=1M2D5M;DP
chr8	163366	.	TAGAC	CAGAG	30471.4	.	AB=0.404658;ABP=57.2529;AC=4;AF=0.666667;AN=6;AO=55
chr8	163370	.	C	G	30471.4	.	AB=0.58952;ABP=50.8301;AC=2;AF=0.333333;AN=6;AO=424
chr8	163387	.	C	T	10073.7	.	AB=0.587644;ABP=49.4474;AC=2;AF=0.333333;AN=6;AO=40
chr8	163419	.	GAA	AGT	6732.12	.	AB=0.277477;ABP=241.712;AC=4;AF=0.666667;AN=6;AO=29
chr8	163432	.	A	G	9010.58	.	AB=0.346416;ABP=123.071;AC=4;AF=0.666667;AN=6;AO=36
chr8	163438	.	C	T	21195.2	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=730;CIGAR=1X;DP=732;D
chr8	163550	.	AAGT	GAGC	12279.9	.	AB=0.346801;ABP=63.5556;AC=4;AF=0.666667;AN=6;AO=22
chr8	163550	.	A	G	12279.9	.	AB=0.639731;ABP=53.3782;AC=2;AF=0.333333;AN=6;AO=19
chr8	163654	.	C	T	6081.47	.	AB=0.401487;ABP=25.6856;AC=4;AF=0.666667;AN=6;AO=24
chr8	163784	.	C	G	4144.35	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=134;CIGAR=1X;DP=135;D
chr8	169366	.	T	C	10888.4	.	AB=0;ABP=0;AC=6;AF=1;AN=6;AO=368;CIGAR=1X;DP=381;D
chr8	169403	.	G	A	0.000305373	.	AB=0.220183;ABP=77.1392;AC=1;AF=0.166667;AN=6;AO=37
chr8	169476	.	C	G	404.305	.	AB=0.21875;ABP=112.941;AC=2;AF=0.333333;AN=6;AO=36;C
chr8	169483	.	T	G	341.779	.	AB=0.2;ABP=128.087;AC=2;AF=0.333333;AN=6;AO=33;CIGA
chr8	169641	.	A	G	62.9338	.	AB=0.75;ABP=5.18177;AC=3;AF=0.5;AN=6;AO=4;CIGAR=1X;
chr8	181859	.	G	A	24.1605	.	AB=0.75;ABP=5.18177;AC=1;AF=0.25;AN=4;AO=3;CIGAR=1X

1

15: bcftools norm on data 14

9,887 lines, 65 comments

format: **vcf**, database: **hg19**

Lines total/split/realigned/skipped: 9681/183/1139/0

display at UCSC main

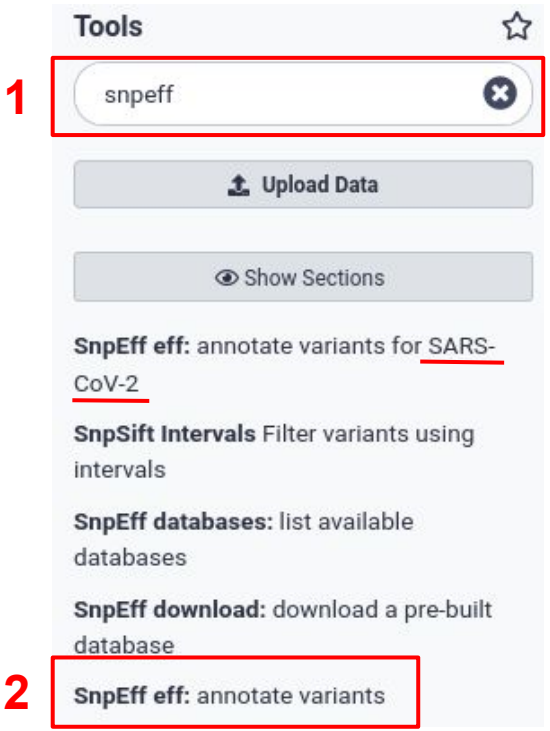
display with IGV local

display at RViewer main

```
1.Chrom
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filt
##fileDate=20210323
##source=freeBayes v1.1.0-46-g8d2b3a0-d
##reference=/data/db/reference_genomes/
```

Mother	Father	Proband
1/0:120:0,40:0:0:40:1480:-347.312,-223.79,-210.846	1/1:117:0,117:0:0:117:4126:-375.053,-38.2308,0	1/0:177:0,63:0:0:63:2341:-551.213,-362.959,-340.977
0/1:120:0,77:0:0:77:2744:-347.312,-124.443,-100.661	0/0:117:0,0:0:0:0:0:-375.053,-377.701,-375.292	0/1:177:0,113:0:0:113:4406:-551.213,-192.495,-155.458

Variant annotation



Variant annotation

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

Favorite | Versions | Options

Sequence changes (SNPs, MNPs, InDels)

15: bcftools norm on data 14

1

Input format

VCF

2

Output format

VCF (only if input is VCF)

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

No

Genome source

Locally installed snpEff database

Genome

3

Homo sapiens : hg19

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 Ontolgy 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

- (selected) 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

4

(-noStats)

Suppress reporting usage statistics to server

Yes

(-noLog)

Email notification

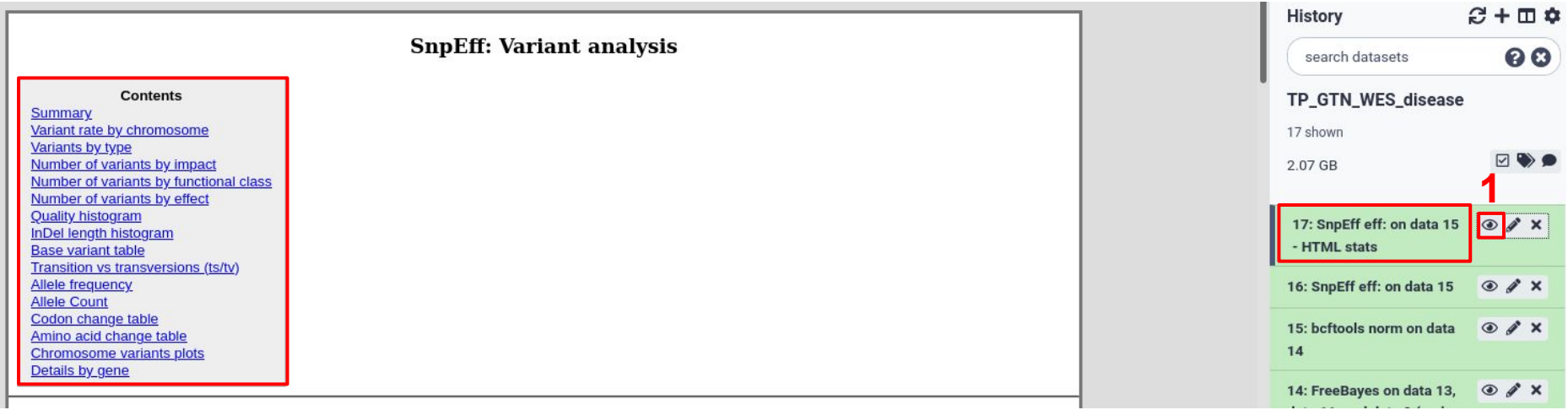
No

Send an email notification when the job completes.

Execute

5

Variant annotation - Content



Variant annotation - Summary

Summary

Genome	hg19
Date	2021-03-18 17:58
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/dnb03/galaxy_db/job_working_directory/015/796/15796878/outpu hg19 /data/dnb03/galaxy_db/files/2/9/3/dataset_293de6fc-1e55-4aec-a452-202e66cc94ca.dat
Warnings	1,061
Errors	0
Number of lines (input file)	9,887
Number of variants (before filter)	9,887
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	9,887
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of effects	28,940
Genome total length	3,137,161,265
Genome effective length	146,364,022
Variant rate	1 variant every 14,803 bases

Variant annotation - Variants details

Variants rate details

Chromosome	Length	Variants	Variants rate
8	146,364,022	9,887	14,803
Total	**146,364,022**	**9,887**	**14,803**

Number variants by type

Type	Total
SNP	8,668
MNP	240
INS	393
DEL	534
MIXED	52
INV	0
DUP	0
BND	0
INTERVAL	0
Total	**9,887**

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	436	1.507%
LOW	1,694	5.853%
MODERATE	1,403	4.848%
MODIFIER	25,407	87.792%

Number of effects by functional class

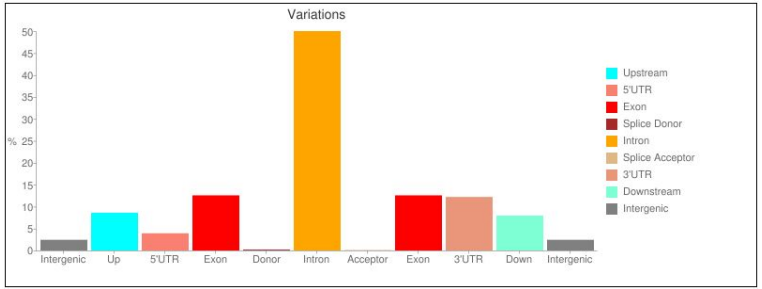
Type (alphabetical order)	Count	Percent
MISSENSE	1,320	54.054%
NONSENSE	9	0.369%
SILENT	1,113	45.577%

Missense / Silent ratio: 1.186

Variant annotation - Variants details

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	3,532	12.001%
5_prime_UTR_premature_start_codon_gain_variant	65	0.221%
5_prime_UTR_variant	1,070	3.636%
conservative_inframe_deletion	1	0.003%
conservative_inframe_insertion	9	0.031%
disruptive_inframe_deletion	2	0.007%
downstream_gene_variant	2,302	7.822%
frameshift_variant	17	0.058%
intergenic_region	704	2.392%
intron_variant	14,900	50.629%
missense_variant	1,352	4.594%
non_coding_transcript_exon_variant	836	2.841%
non_coding_transcript_variant	3	0.01%
sequence_feature	197	0.669%
splice_acceptor_variant	27	0.092%
splice_donor_variant	62	0.211%
splice_region_variant	409	1.39%
start_lost	3	0.01%
stop_gained	12	0.041%
stop_lost	10	0.034%
stop_retained_variant	1	0.003%
structural_interaction_variant	314	1.067%
synonymous_variant	1,114	3.785%
upstream_gene_variant	2,488	8.454%

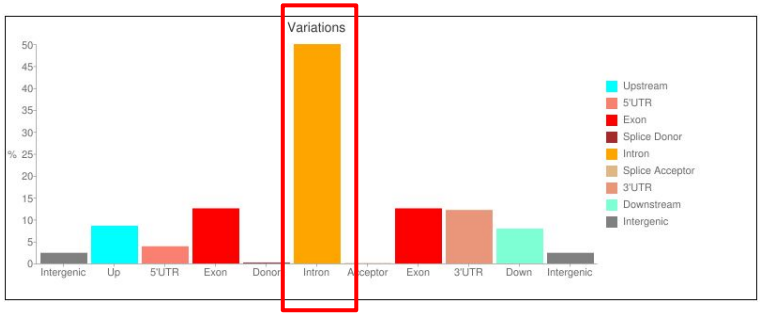
Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,302	7.954%
EXON	3,638	12.571%
INTERGENIC	704	2.433%
INTRON	14,479	50.031%
SPLICE_SITE_ACCEPTOR	27	0.093%
SPLICE_SITE_DONOR	57	0.197%
SPLICE_SITE_REGION	378	1.306%
TRANSCRIPT	200	0.691%
UPSTREAM	2,488	8.597%
UTR_3_PRIME	3,532	12.205%
UTR_5_PRIME	1,135	3.922%



Variant annotation - Variants details

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	3,532	12.001%
5_prime_UTR_premature_start_codon_gain_variant	65	0.221%
5_prime_UTR_variant	1,070	3.636%
conservative_inframe_deletion	1	0.003%
conservative_inframe_insertion	9	0.031%
disruptive_inframe_deletion	2	0.007%
downstream_gene_variant	2,302	7.822%
frameshift_variant	17	0.058%
intergenic_region	704	2.392%
intron_variant	14,900	50.629%
missense_variant	1,352	4.594%
non_coding_transcript_exon_variant	836	2.841%
non_coding_transcript_variant	3	0.01%
sequence_feature	197	0.669%
splice_acceptor_variant	27	0.092%
splice_donor_variant	62	0.211%
splice_region_variant	409	1.39%
start_lost	3	0.01%
stop_gained	12	0.041%
stop_lost	10	0.034%
stop_retained_variant	1	0.003%
structural_interaction_variant	314	1.067%
synonymous_variant	1,114	3.785%
upstream_gene_variant	2,488	8.454%

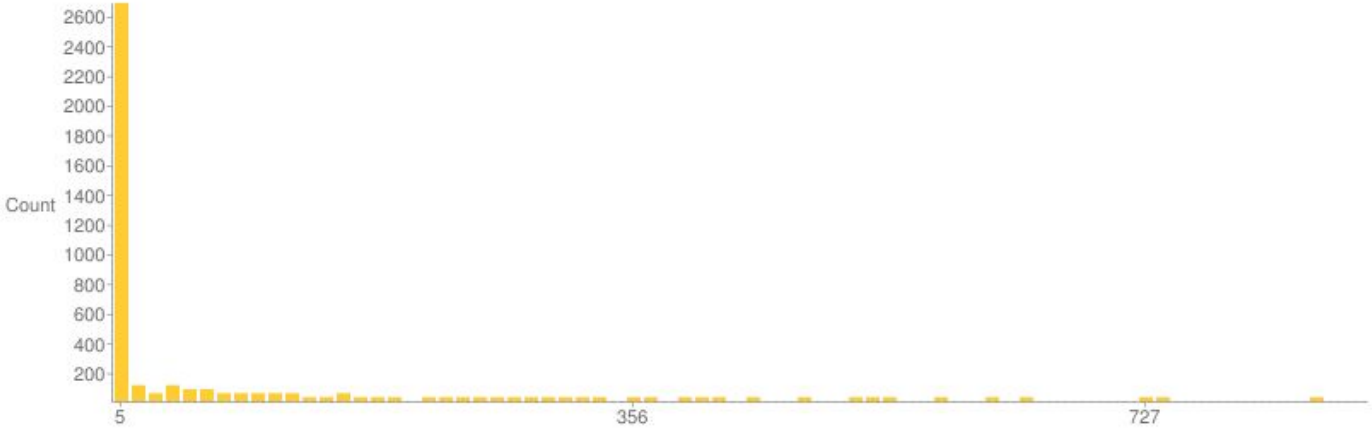
Type (alphabetical order)	Count	Percent
DOWNSTREAM	2,302	7.954%
EXON	3,638	12.571%
INTERGENIC	704	2.433%
INTRON	14,479	50.031%
SPLICE_SITE_ACCEPTOR	27	0.093%
SPLICE_SITE_DONOR	57	0.197%
SPLICE_SITE_REGION	378	1.306%
TRANSCRIPT	200	0.691%
UPSTREAM	2,488	8.597%
UTR_3_PRIME	3,532	12.205%
UTR_5_PRIME	1,135	3.922%



Variant annotation - Variants quality

Quality:

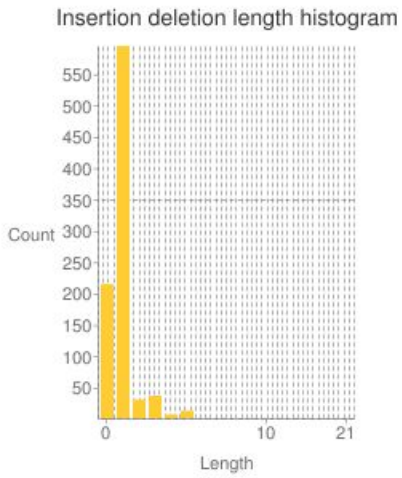
Min	0
Max	50,552
Mean	1,468.087
Median	458
Standard deviation	2,687.315
Values	0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41
Count	2485,35,27,29,20,24,11,14,7,19,7,12,10,13,12,15,14,12,10,13,10,10,14,7,9,7,9,5,12,11,9,9,6,5,6,7,9,8,9,13,14,13,10,1



Variant annotation - Insertions/Deletions

Insertions and deletions length:

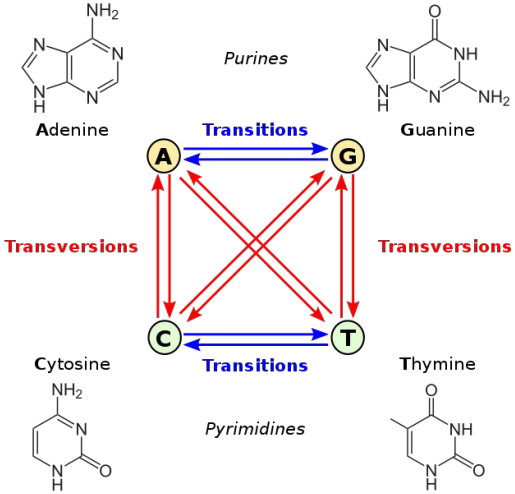
Min	0
Max	21
Mean	1.151
Median	1
Standard deviation	1.684
Values	0,1,2,3,4,5,6,7,8,9,10,11,12,17,20,21
Count	220,595,34,37,8,14,5,2,3,1,1,2,2,1,1,1



Variant annotation - Transitions/Transversions

Base changes (SNPs)

	A	C	G	T
A	0	962	1,204	270
C	330	0	403	1,182
G	1,260	415	0	304
T	248	1,230	860	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).

Transitions	13,739
Transversions	6,876
Ts/Tv ratio	1.9981

All variants:

```
Sample ,mother,father,proband,Total
Transitions ,4408,4605,4726,13739
Transversions ,2183,2318,2375,6876
Ts/Tv ,2.019,1.987,1.990,1.998
```

Only known variants (i.e. the ones having a non-empty ID field):

```
No results available (empty input?)
```

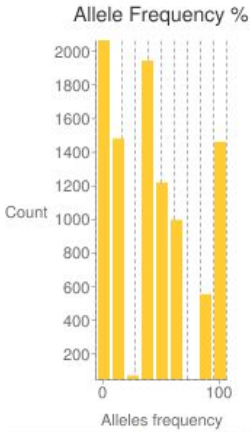
Sequencing Type	# of Variants*	TiTv Ratio
WGS	~4.4M	2.0-2.1
WES	~41k	3.0-3.3

*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

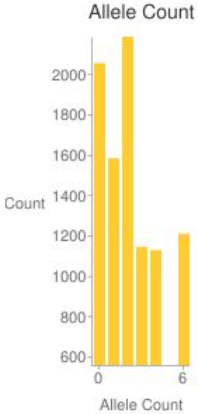
Variant annotation - Allele details

Allele frequency



Min	0
Max	100
Mean	41.984
Median	33
Standard deviation	33.984
Values	0,16,25,33,50,66,75,83,100
Count	2062,1484,72,1953,1230,1010,47,558,1471

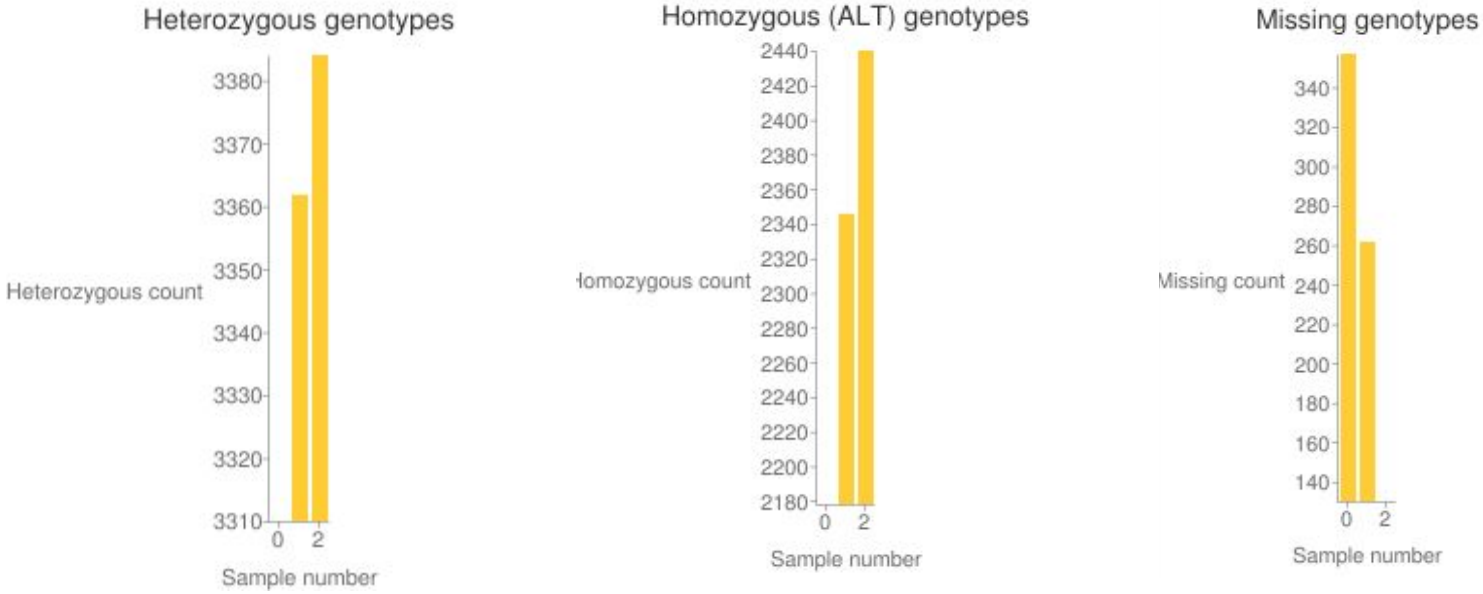
Allele Count



Min	0
Max	6
Mean	2.426
Median	2
Standard deviation	1.963
Values	0,1,2,3,4,5,6
Count	2062,1588,2184,1156,1130,558,1209

Variant annotation - Genotypes details

Hom/Het per sample




```
Sample_names , mother, father, proband
Reference , 4042, 3915, 3933
Het , 3310, 3362, 3384
Hom , 2178, 2348, 2440
Missing , 357, 262, 130
```

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	CAT	CCA	CCC
-					5						3												
AAA			8	13						2													
AAC	2			1	28		8				13		3		3								
AAG		6	1					5				5								1			
AAT		7	26										15								8		
ACA							5	32	1					13								1	
ACC			2						12		2				3								84
ACG						33	3		5							4							
ACT					6	1	15										5						
AGA		2									3		2										
AGC	2		12				5					1	16										
AGG				4						2			6										
AGT					6				6		20	6											

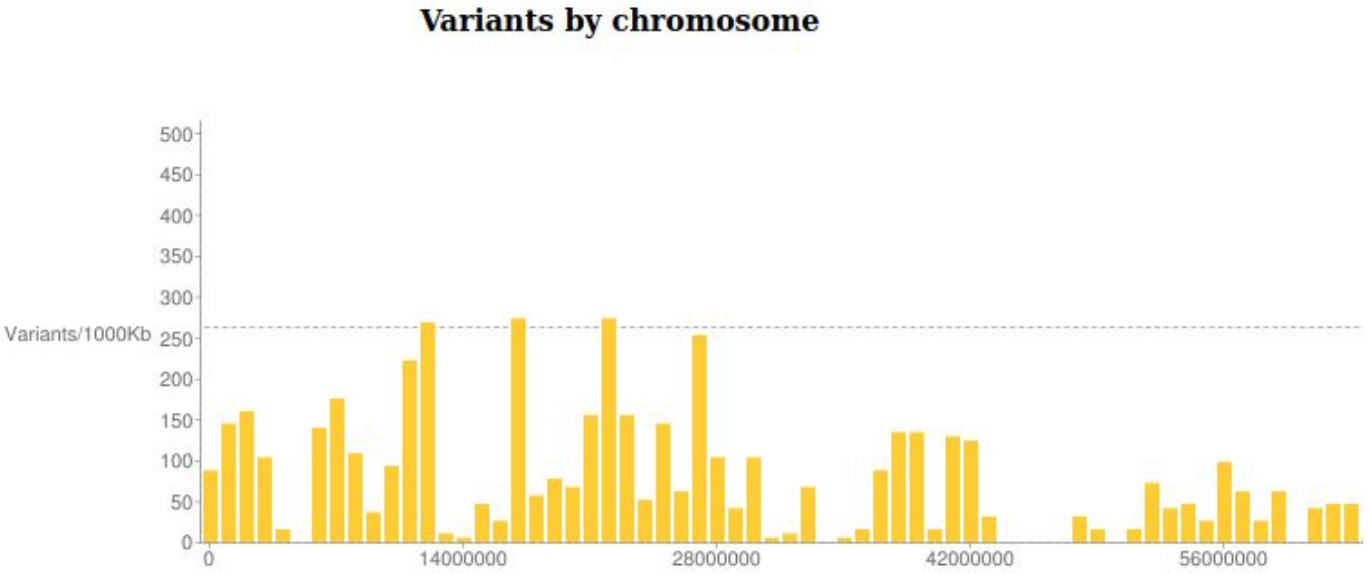
Variant annotation - Amino acid changes

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 '
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	Y
*	1	1			3				1				3				2						
-			12						3						5				3		5		
?																							
A		2		185		1	1		15							3			8	22	37		
C	4	3			9				10									17	2			2	3
D				21		47	9		21	7					7								2
E	1		1	5		4	29		66			11					8				2		
F					1			34			7		4								2		1
G					1	5	4		117									22	5		1		
H		1								33			2		1	29	2	29					
I						7					22		7	3						19	29		
K							8					19			9		1	7		5			
L								8					122	4		21	4	6	4		19		
M											9		2							26	9		
N		2				16				8	3	8			54				31	8			1
P		3		5						8			20			96	2	3	12	10			
Q							2			5		3				2	16	8					
R	4				15				18	21		6	4			5	32	49	12			16	
S	1	3		8	3	2		2	11						18	23		10	114	13			1
T				18							21		1	4	8	89			3	107			
V		1		13		2	2		169		18		8	23							51		
W	1								5									5	2				
Y	1				2	5		3		1						1			3				37

Variant annotation - Chromosomes details



Variant annotation - ANN field

```
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##SnpEffCmd="SnpEff -i vcf -o vcf -stats /data/dnb03/galaxy_db/job_working_directory/015/706/15706434/outputs/galaxy_dataset_61795ebc-d4e3-4436-a13b-2f65e9efa44e.dat hg19 /data/dnb03/galaxy_db/files/9/a
##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'">
```

1

'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' ">

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|OR4F21|OR4F21|transcript|NM_001005504.1|protein_coding|1/1|c.130T>C|p.Phe44Leu|130/939|130/939|44/312||

Intronic

ANN=G|intron_variant|MODIFIER|RPL23AP53|RPL23AP53|transcript|NR_003572.2|pseudogene|3/3|n.423-74G>C|||||

Variant reporting - Pedigree

Individual

Father ID

Family ID

Mother ID

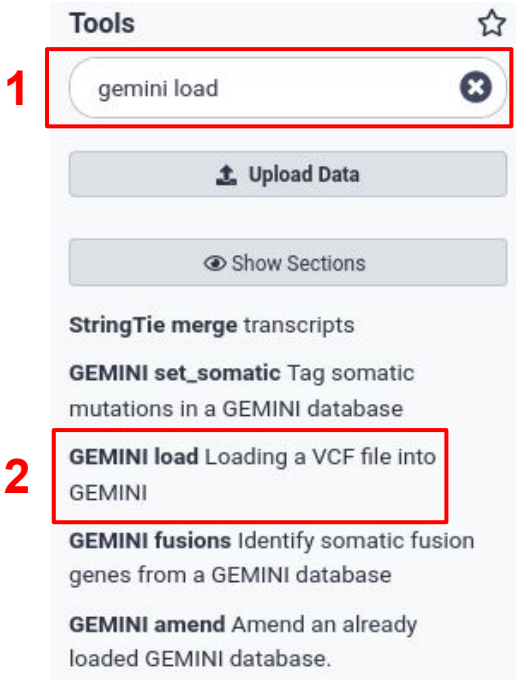
FAM	father	0	0	1	1
FAM	mother	0	0	2	1
FAM	proband	father	mother	1	2

Sex (1: male; 2: female)

Status (1: control; 2: case)

1

Variant reporting - Database creation



Variant reporting - Database creation

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

Favorite | Versions | Options

VCF dataset to be loaded in the GEMINI database

16: SnpEff eff: on data 15

1

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

The variants in this input are

annotated with snpEff

2

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 (2019-01-12 snapshot)

3

Sample and family information in PED format

4: Pedigree.txt

4

Variant reporting - Database creation

Load the following optional content into the database

- [ ] Select/Unselect all

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

No

Send an email notification when the job completes.

✔ Execute

6

Variant reporting - Database creation

GEMINI load

Tool Parameters

Input Parameter	Value
VCF dataset to be loaded in the GEMINI database	• 6: SnpEff eff: on data 15
The variants in this input are	annotated with snpEff
This input comes with genotype calls for its samples	True
Choose a gemini annotation source	2019-01-12
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 Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/iuc/gemini_load/gemini_load/0.20.1+galaxy2
Command Line	*empty*
Tool Standard Output	`Indexing /data/dnb03/galaxy_db/job_working_directory/015/797/15797044/working/input.vcf.gz with grabix. Loading 9887 variants. Break…` **(Click to expand)**
Tool Standard Error	`CADD scores are being loaded (to skip use:--skip-cadd). GERP per bp is being loaded (to skip use:--skip-gerp-bp). [W::hts_idx_load2]…` **(Click to expand)**
Tool Exit Code:	0
Job API ID:	11ac94870d0bb33a07467735207e9a77

History

search datasets

TP_GTN_WES_disease

18 shown

2.22 GB

18: GEMINI load on data 4 and data 16

207.1 MB

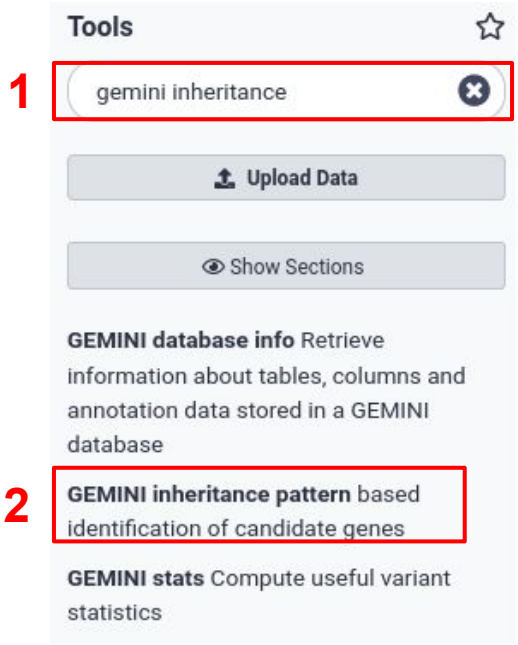
format: **gemini.sqlite**, database: **hg19**

Indexing /data/dnb03/galaxy_db/job_working_directory/015/797/15797044/working/input.vcf.gz with grabix.
Loading 9887 variants.
Breaking /data/dnb03/galaxy_db/job_working_directory/015/797/15797044/working/input.vcf.gz into 10 chunks.
Loading chunk 0.
Loa

Gemini SQLite Database, version 0.20.1

1

Variant reporting - Querying



Variant reporting - Querying

GEMINI database

18: GEMINI load on data 4 and data 16

Only files with version 0.20.1 are accepted.

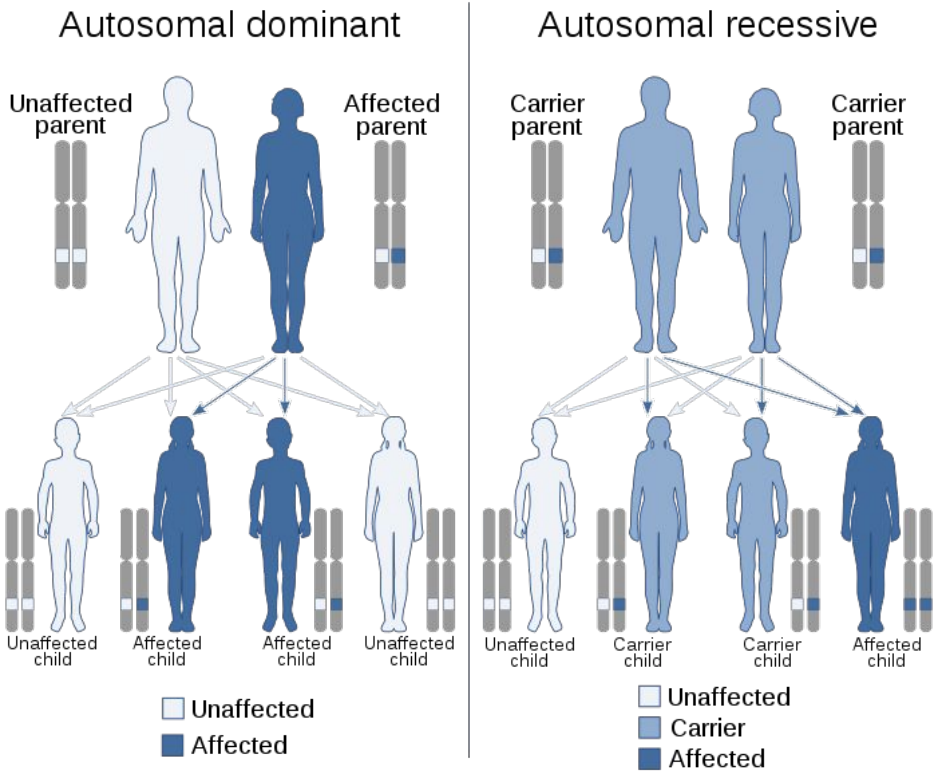
Your assumption about the inheritance pattern of the phenotype of interest

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)

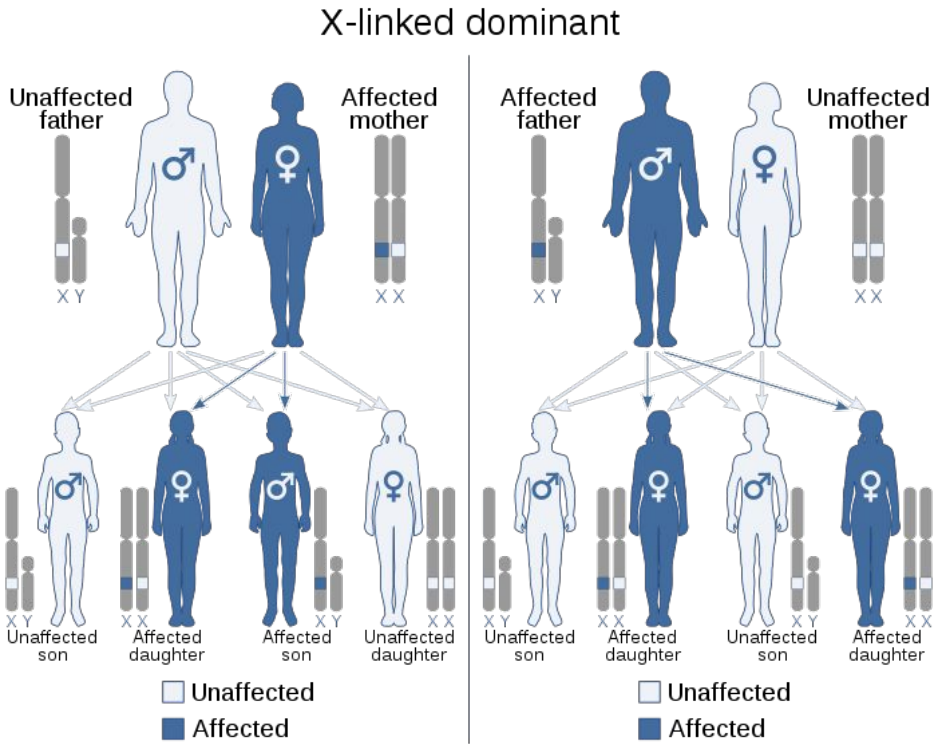
Which inheritance pattern to select ?

Variant reporting - Inheritance pattern



https://en.wikipedia.org

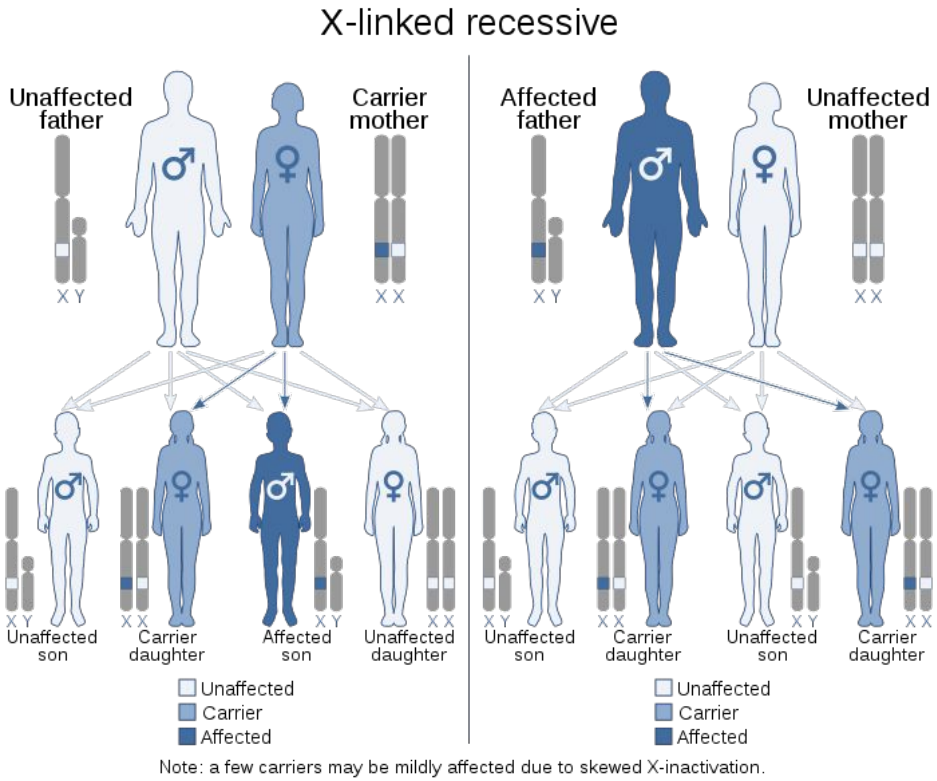
Variant reporting - Inheritance pattern



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

https://en.wikipedia.org

Variant reporting - Inheritance pattern



https://en.wikipedia.org

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 consiguineous

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 consiguineous

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 consiguineous

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 database

18: GEMINI load on data 4 and data 16

1

Only files with version 0.20.1 are accepted.

Your assumption about the inheritance pattern of the phenotype of interest

Autosomal recessive

2

Additional constraints on variants

+ Insert Additional constraints on variants

3

Additional constraints on variants

1: Additional constraints on variants

Additional constraints expressed in SQL syntax

impact_severity != 'LOW'

4

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

1

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

No, analyze all variants from all included families

Variant reporting - Querying

Output - included information

Set of columns to include in the variant report table

Custom (report user-specified columns) 5

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
- [x] 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

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

No

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_Microcep
0.935555555556	chr8	6681255	A	C	splice_region_variant	XKR5	None	None
-1.0	chr8	11666217	GTCCCAC	G	conservative_inframe_deletion	FDFT1	None	None
0.671189839572	chr8	12042879	T	C	splice_region_variant	FAM86B1	None	None
0.6916	chr8	12044200	A	G	splice_region_variant	FAM86B1	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

Variant reporting - Results

clinvar_gene_phenotype
None
primary_microcephaly\x2c_recessive|primary_autosomal_recessive_microcephaly_1
None
None
None
None
None
None
None
carcinoma_of_colon
None
None

Variant reporting - Results

rs_ids	variant_id	family_id	family_members	family_genotypes	samples	family_count
rs968381	293	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	A/G,A/G,G/G	proband	1
rs1057090	603	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	C/T,C/T,T/T	proband	1
rs9772979	638	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	A/C,A/C,C/C	proband	1
rs71711801	1238	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	GTCCCAC/G,GTCCCAC/G,G/G	proband	1
rs142379100	1376	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	T/C,T/C,C/C	proband	1
rs2684084	1381	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	A/G,A/G,G/G	proband	1
rs3739310	1500	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	T/G,T/G,G/G	proband	1
rs545589847,rs502882	1503	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	G/A,G/A,A/A	proband	1
rs608052	1506	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	A/G,A/G,G/G	proband	1
rs2705051	1770	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	G/C,G/C,C/C	proband	1
rs484373	1836	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	G/A,G/A,A/A	proband	1
rs754721723	1850	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	AC/GT,AC/GT,GT/GT	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	samples
None	5779	FAM	mother(mother;unaffected;female),father(father;unaffected;male),proband(proband;affected;male)	G/A,G/A,A/A	proband