JAX Big Genomic Data Skills Training - 2018 Mouse Exome Variant Discovery Module

Introduction

This data analysis module introduces next generation sequence analysis to students. The sequence provided comes from a Mouse Exome sequencing experiment at The Jackson Laboratory (JAX). For decades JAX has bred, cared for and distributed mouse models of human disease. With thousands of sub-strains of mice being bred, spontaneous mutations in the mouse genome occur with reasonable frequency across the plethora of JAX mouse breeding colonies. Additionally JAX has led several large scale mutagenesis projects to discover novel genes and/or novel gene functions. Over 5,000 spontaneous or induced mutant alleles with clinical phenotypes are cataloged in the Mouse Genomic Informatics database.(http://www.informatics.jax.org) Estimates suggest that ~1000 of these mutant alleles occur in coding sequence or within ~20 bp of intron/exon boundaries (Fairfield et al. Genome Biology 2011, 12:R86). Therefore exome sequencing can and has been used to uncover DNA sequence variants that lead to clinical variation relevant to human disease.

In the first decade of the 21st century scientists at JAX designed a sequence capture probe pool in order to enable mouse exome sequencing experiments. The exome capture probes cover 203,225 exonic regions in the mouse genome representing ~ 54.3 Mb of the C57BL6/J mouse genome (For specific details see Fairfield et al. 2001, above).

This training exercise is based upon exome sequencing of a mutant mouse in the JAX colonies.

Students are provided with the sorted exome sequences, paired end reads (both directions). JAX has sorted the reads to come from only Chromosome 1 in order to accelerate the computational steps. Genomic DNA was prepared from the mouse using standard techniques and Illumina paired end libraries were generated at JAX. Sequence read depth is ~18X, in other words each captured exonic region was sequenced ~ 18 times.

The goal of this exercise is identify the genomic DNA sequence variant and gene responsible for the phenotype in the mouse. The mouse is phenotypically different than wildtype C57BL6/J animals. The mouse of interest for this exercise is named “Leg dragger”; the mutation was spontaneous on the C57BL6/J strain. Homozygotes are slightly smaller than their unaffected littermates and lose most of the use of their rear legs such that they drag their rear legs and pull themselves along with their front legs to move. This phenotype can be detected as early as 2 weeks of age and is evident by 3 weeks of age. A few homozygotes nearing wean age are found to roll over and over in a struggle to right themselves. When raised by their tails they do not splay their legs outward but rather cross the front pair and the rear pair. Auditory brainstem response analysis of one homozygous animal at 18 days of age showed severe hearing loss and no others were tested. Heterozygotes appear normal and fertile but produce slightly fewer homozygotes than the 25% expected when intercrossed. The strain appears to provide a model for autosomal recessive **spastic paraplegia 30**.

The mouse model system has numerous, significant benefits. First and foremost one can breed the animals, determine heritability of traits and map traits to different parts of the mouse genome. In the case of this leg-dragger, mapping data does exist and the phenotype appears to be driven by a mutation on mouse chromosome 1.

For the purpose of this training exercise knowing that the mutation is very likely on Mouse chromosome 1 simplifies the data analysis approach. It is computationally intensive to map tens of millions of short sequence reads to the mouse genome. By knowing the chromosome of interest, this exercise can run more rapidly as the 'mapping reads to the reference' step can be targeted. A mouse chromosome 1 '*mm10.chr1.fa'* file will be specified so that the reference genome will be represented by just a single reference chromosome.

Specifics about how analyze data and target to a specific region of a genome are provided below in this document.

What do students need to know coming in (or cover before launching the analysis)?
Basic molecular biology- DNA, RNA, proteins. Transcription. Translation.
What is Next Generation or High Throughput DNA sequencing.
What does it mean to align sequence to a reference genome.
What is the difference between a variant, polymorphism, and mutation in genomic DNA, and what is a pathogenic mutation.

Bonus Knowledge for students
Introductory knowledge of bioinformatics resources would be useful including NCBI gene resources, OMIM.
Know why different genome sequence references exist, e.g. different 'releases' of the genome with different coordinates and annotations
Having experience with Ensembl genome database would be useful.

Knowledge (concept) goals
What is a fastq file, and what information does it hold (in groups of four lines)?
What QC steps are recommended for genomic sequencing
What is the difference between aligning directly to a genome or aligning to a target region.
How are duplicate short reads identified and removed to avoid bias.

Practical skills goals
Logging in to Galaxy
Uploading a working sequence set and a fasta file of a single human chromosome
Gaining familiarity with file types, fasta, fastq, bam, vcf, bed
Perform QC steps and possibly process the sequence files to remove issues
Aligning individual files to a single chromosome

Calling sequence variants from aligned short reads
Interpreting types of variants reported in a vcf file

Leg-Dragger Mouse Exome Variant Discovery Module

1. Getting Data

To Get Mouse Exome Data:

I. Data for the module can be downloaded from ftp://ftp.jax.org/encode/MouseExome, open as GUEST

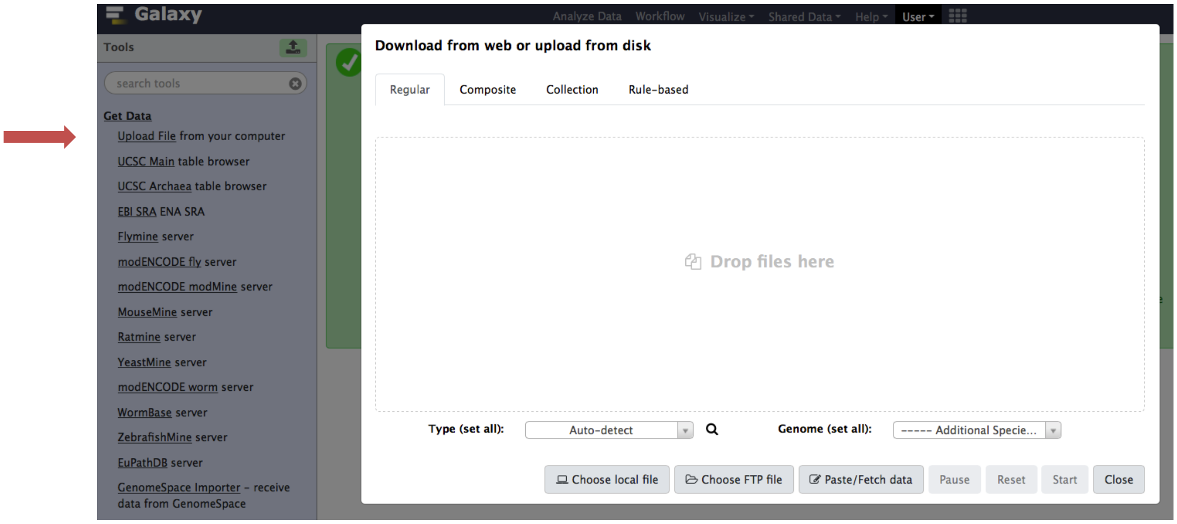
II. You should see a variety of files for your use:

SRR1783944_chr1_R1.fastq.gz This is *forward* reads from exome sequencing, file is filtered to include only Chr. 1 sequencing reads.)

SRR1783944_chr1_R2.fastq.gz This is *reverse* reads from exome sequencing

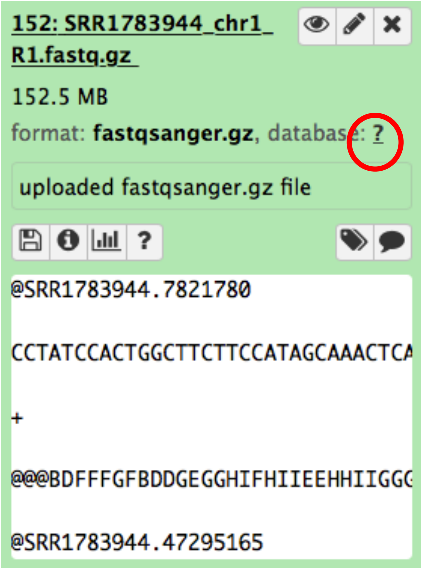
Mm10.chr1.fa this is a single fasta file for mouse chromosome 1, used for mapping reads as reference

III. Open up your Galaxy then click ‘Upload File’ under ‘Get Data’. Upload all the files by dragging the files into the upload window.



IV. **Note on formats**: In some cases raw data files come in different formats. Usually fastq, fastqsanger or others. Data is essentially the same but certain analysis tools expect certain formats. If necessary sequence files can be converted (For this practice, you don’t need to modify the datatype.):

a. Load data in to Galaxy
b. Highlight data file and click on **?** after database (red circle):



152: SRR1783944_chr1_R1.fastq.gz

152.5 MB

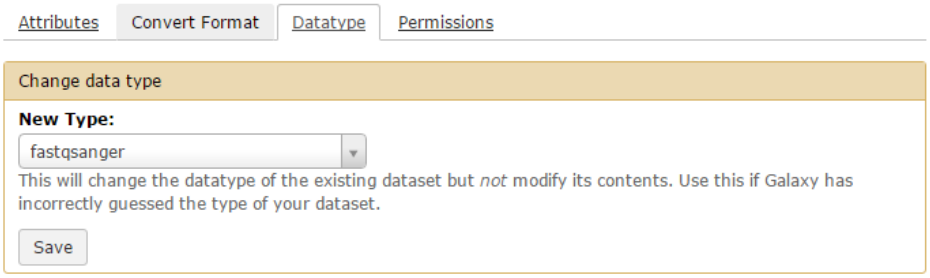
format: **fastqsanger.gz**, database: ?

uploaded fastqsanger.gz file

```
@SRR1783944.7821780

CCTATCCACTGGCTTCTTCCATAGCAAACTCA

+

@@@BDFFFGFBDDGEGGHIFHIIEEHHIIGGG

@SRR1783944.47295165
```

c. Go to main screen and highlight “datatype” tab and use pull down to select type you need and hit save.

d. You can also reach this screen via the ‘edit attributes’ pencil icon

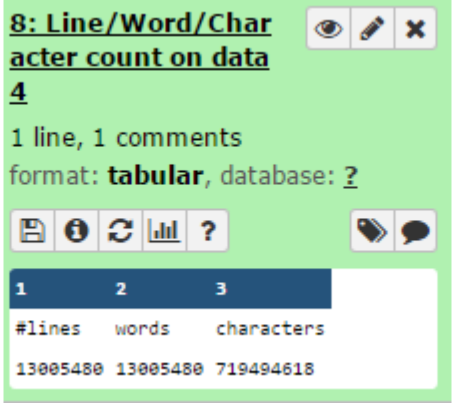


Change data type

New Type: fastqsanger

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

2. **Simple raw data visualization** Line/Word/Character tool count
   It may be valuable to have students run this very simple Galaxy tool. The tool simply tells you how many sequence reads are in your file; paired files should have equal numbers of reads. For instance after hitting the eye-ball symbol on the Line/Word/Character count tool students will see that the file includes, in this instance, over 13M reads (#lines):



8: Line/Word/Character count on data 4

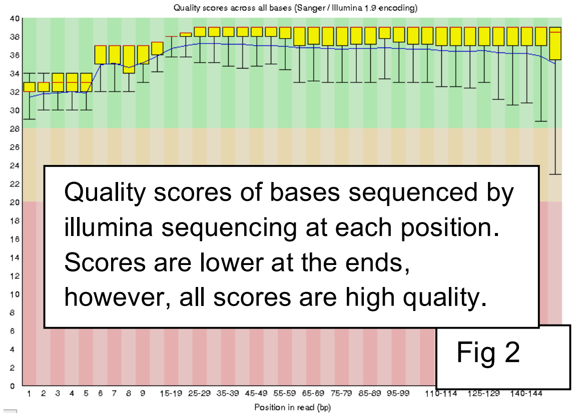
1 line, 1 comments

format: **tabular**, database: ?

```
1        2        3
#lines   words    characters
13005480 13005480 719494618
```

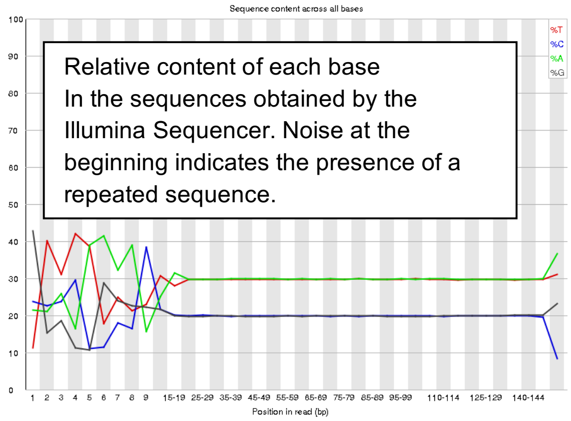
3. Performing Quality Control (QC) (fastqsanger)

The raw data is presented in a fastq file, which is specific for Illumina sequencing. This file is comprised not only of the nucleotide sequence, but also includes an ID number and quality score which is important for determining the integrity of the data obtained. A fastq file is obtained for both the forward and reverse reads (R1 and R2), and these typically range from 50-150 base pairs, in this module raw reads are 150bp. These files are stored separately and run through the FastQC tool on Galaxy in order perform quality control checks on raw sequence data. This tool is characterized by primarily the per base sequence quality (Fig.1), the per base sequence content (Fig.2), the adapter content (Fig.3) and the Kmer content (Fig.5). The per base sequence quality should be over 30 for it to be considered a high quality score for use. Quality scores tend to be lower near the beginning of the read and drop off near the end. The per base sequence content should be uniform, such that there are equal numbers of each base (~25%) over the whole read. The adapter content indicates the location and amount of the adapter sequence that is included in the read, which is important to note for trimming purposes. Finally, the Kmer content indicates sequences that are abnormally repeated.
Based on the FastQC, the reads should be trimmed using the *Trimmomatic* tool in Galaxy in order to remove any low quality portions of the reads that would affect alignment in subsequent steps.

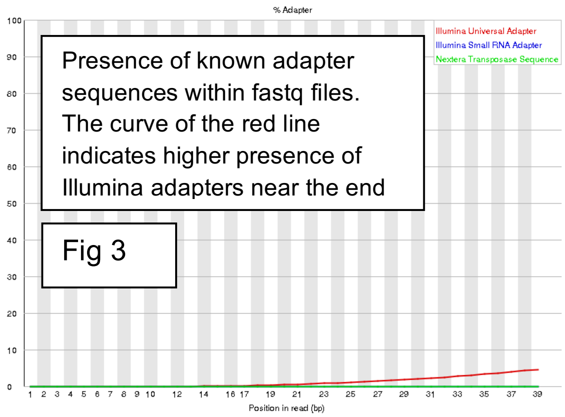


Quality scores of bases sequenced by illumina sequencing at each position. Scores are lower at the ends, however, all scores are high quality.

Fig 2

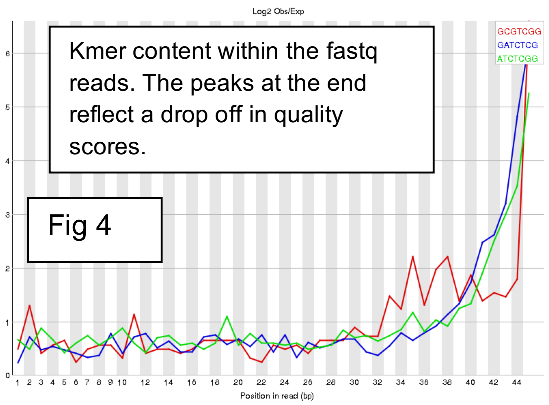


Relative content of each base
In the sequences obtained by the Illumina Sequencer. Noise at the beginning indicates the presence of a repeated sequence.



Presence of known adapter sequences within fastq files. The curve of the red line indicates higher presence of Illumina adapters near the end

Fig 3



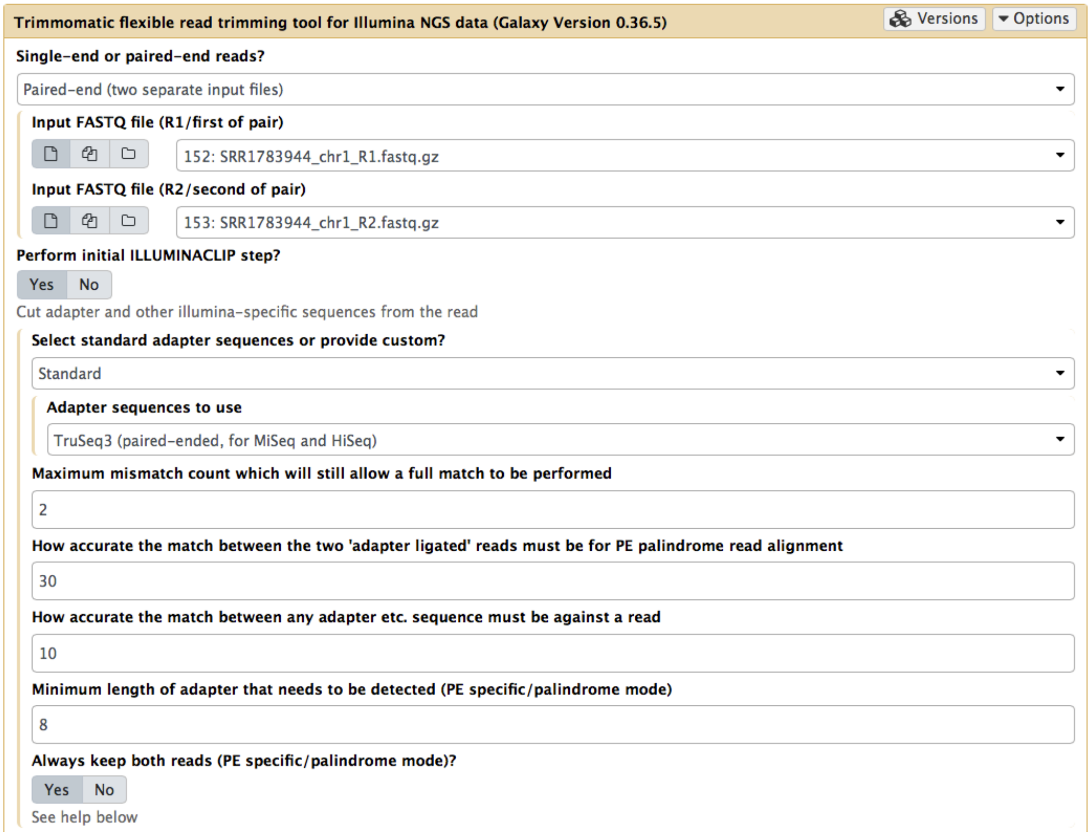
Kmer content within the fastq reads. The peaks at the end reflect a drop off in quality scores.

Fig 4

4. QC processing with trimmomatic. (Actually unnecessary with this very clean data)

Trimmomatic is one popular tool for removing systematic problems from NGS data. (You are welcome to use others that are available and that you prefer).

A suggested set of settings for our data is shown below:



Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.5)

Single-end or paired-end reads?

Paired-end (two separate input files)

Input FASTQ file (R1/first of pair)

152: SRR1783944_chr1_R1.fastq.gz

Input FASTQ file (R2/second of pair)

153: SRR1783944_chr1_R2.fastq.gz

Perform initial ILLUMINACLIP step?

[✓ Yes] No

Cut adapter and other illumina-specific sequences from the read

Select standard adapter sequences or provide custom?

Standard

Adapter sequences to use

TruSeq3 (paired-ended, for MiSeq and HiSeq)

Maximum mismatch count which will still allow a full match to be performed

2

How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment

30

How accurate the match between any adapter etc. sequence must be against a read

10

Minimum length of adapter that needs to be detected (PE specific/palindrome mode)

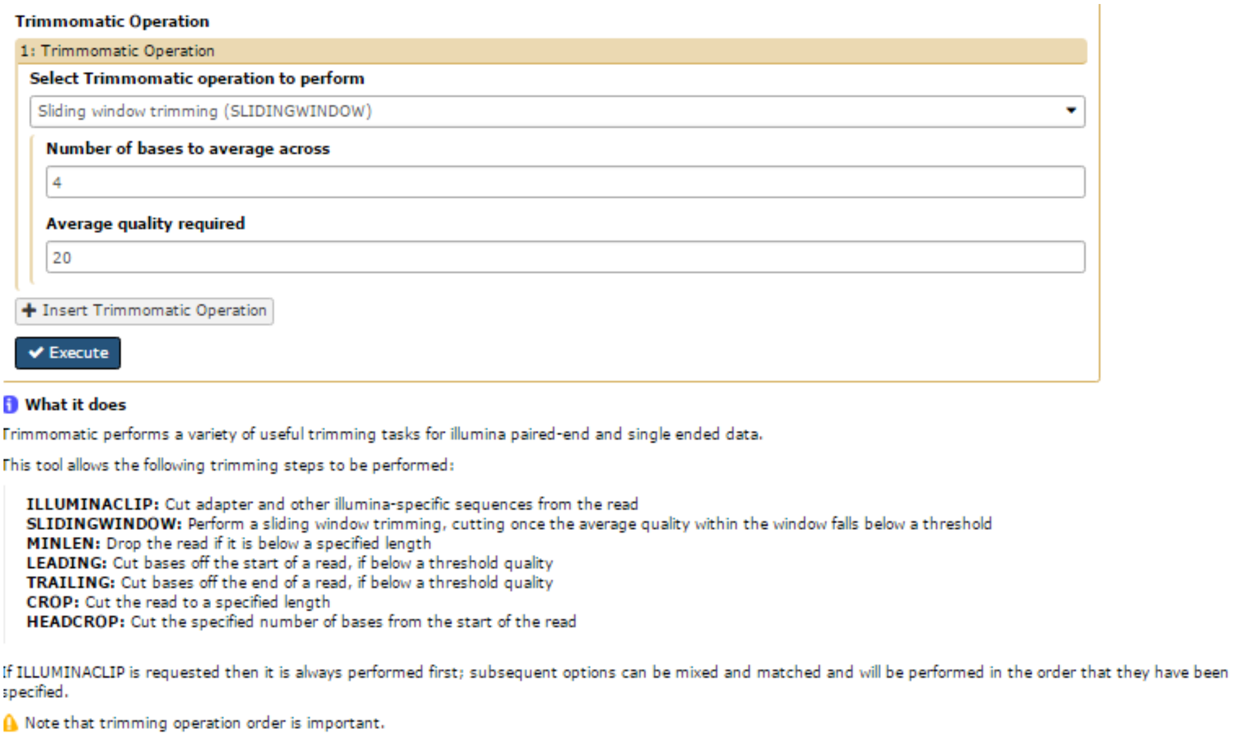
8

Always keep both reads (PE specific/palindrome mode)?

[✓ Yes] No

See help below

Continued:



Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

What it does

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.

This tool allows the following trimming steps to be performed:

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read
SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold
MINLEN: Drop the read if it is below a specified length
LEADING: Cut bases off the start of a read, if below a threshold quality
TRAILING: Cut bases off the end of a read, if below a threshold quality
CROP: Cut the read to a specified length
HEADCROP: Cut the specified number of bases from the start of the read

If ILLUMINACLIP is requested then it is always performed first; subsequent options can be mixed and matched and will be performed in the order that they have been specified.

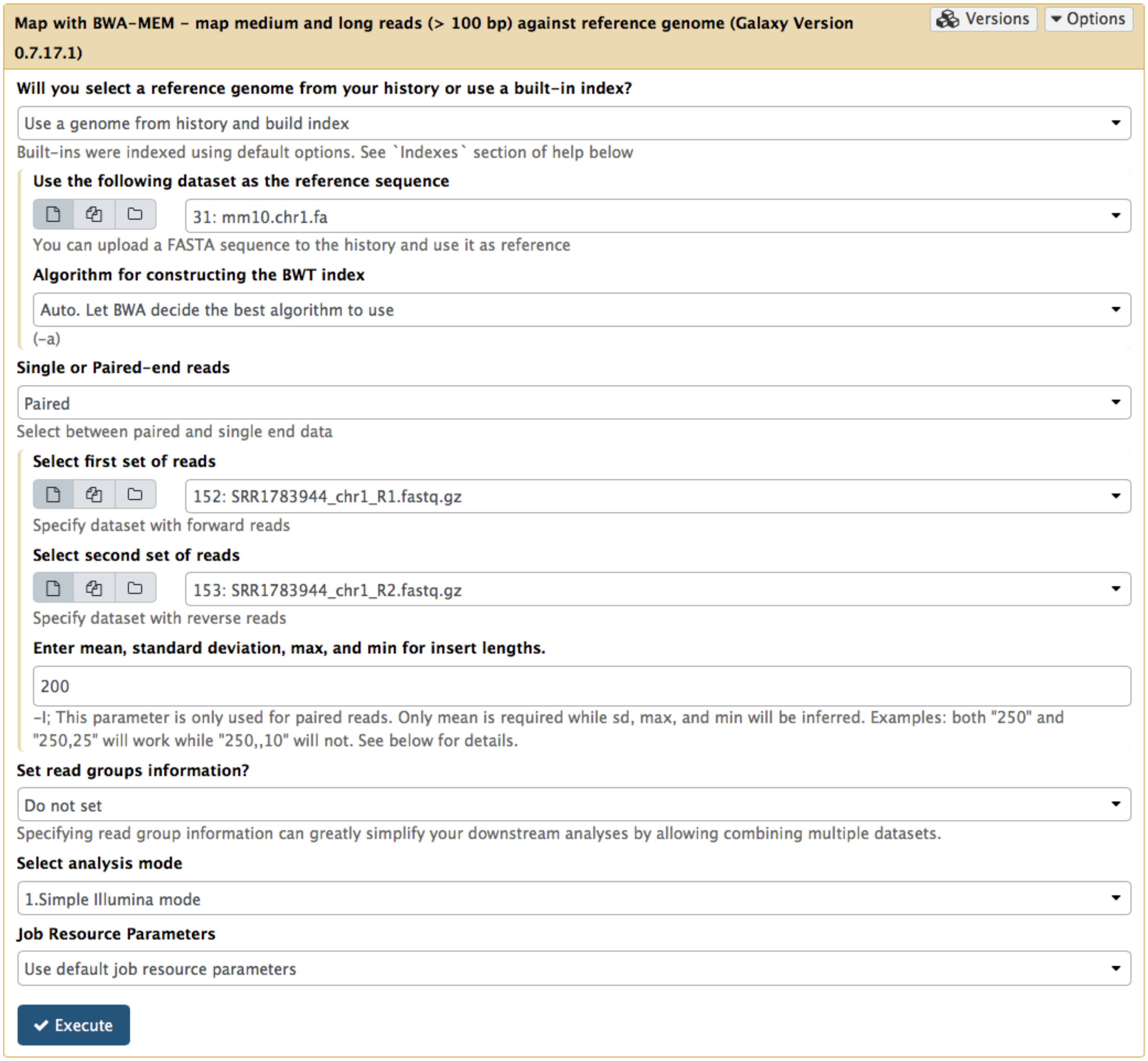
Note that trimming operation order is important.

5. **Examine the updated (newly generated) fastq file** with FASTQC, using the same logic as for step 4 above.

   Has Trimmomatic changed and/or improved following the QC processing?

6. **Alignment to Reference** using Map with BWA-MEM

   Aligning a sequence to a reference is a critical and time consuming step in the process. This is the step where the short sequences are aligned back to a reference genome, human, mouse yeast etc. It is vital that you know what reference genome you are aligning to, specifically what version or release of an annotated genome. In this module mapping will be to mouse chromosome 1. Genome version is very important: in this exercise we use Mouse Dec. 2011 GRCm38/mm10 or mm10. In the window below notice that you will select: '**Use the following dataset as the reference sequence'** then select 'mm10.chr1.fa'. This set will create a BAM file (Binary Alignment Mapping file). Analysis may take several hours (or days depending on activity on galaxy instance).



Map with BWA-MEM – map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1)

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

Use a genome from history and build index

Built-ins were indexed using default options. See `Indexes` section of help below

Use the following dataset as the reference sequence

31: mm10.chr1.fa

You can upload a FASTA sequence to the history and use it as reference

Algorithm for constructing the BWT index

Auto. Let BWA decide the best algorithm to use

(-a)

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

152: SRR1783944_chr1_R1.fastq.gz

Specify dataset with forward reads

Select second set of reads

153: SRR1783944_chr1_R2.fastq.gz

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

200

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

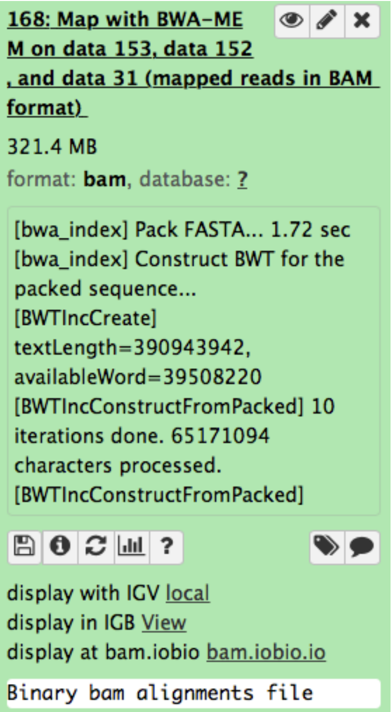
Select analysis mode

1.Simple Illumina mode

Job Resource Parameters

Use default job resource parameters

7. **Post BWA-MEM**
   At this point it is possible to visualize all the reads aligned against the reference genome (Mouse mm10 in this case). This can be done by using IGV which is linked to Galaxy.



168: Map with BWA-MEM on data 153, data 152, and data 31 (mapped reads in BAM format)

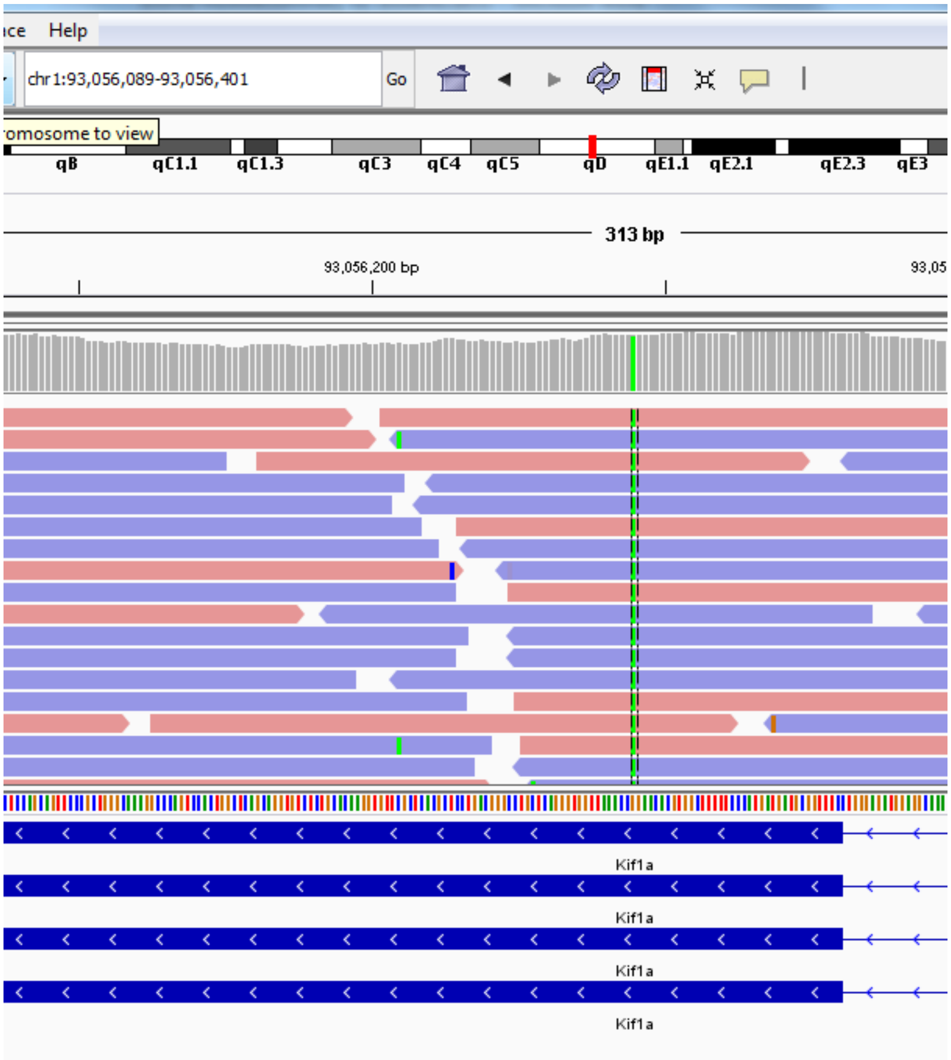
321.4 MB

format: **bam**, database: ?

```
[bwa_index] Pack FASTA... 1.72 sec
[bwa_index] Construct BWT for the
packed sequence...
[BWTIncCreate]
textLength=390943942,
availableWord=39508220
[BWTIncConstructFromPacked] 10
iterations done. 65171094
characters processed.
[BWTIncConstructFromPacked]
```

Binary bam alignments file

Use the *display with IGV*. Depending on your computer platform you may need to install IGV on your machine. It will load the genome reference.
Be patient you will need to wait as it builds the graphical interface with all the reads. Also the level of zoom-in is important. The display below is mouse data from chromosome 1 from this exome dataset.

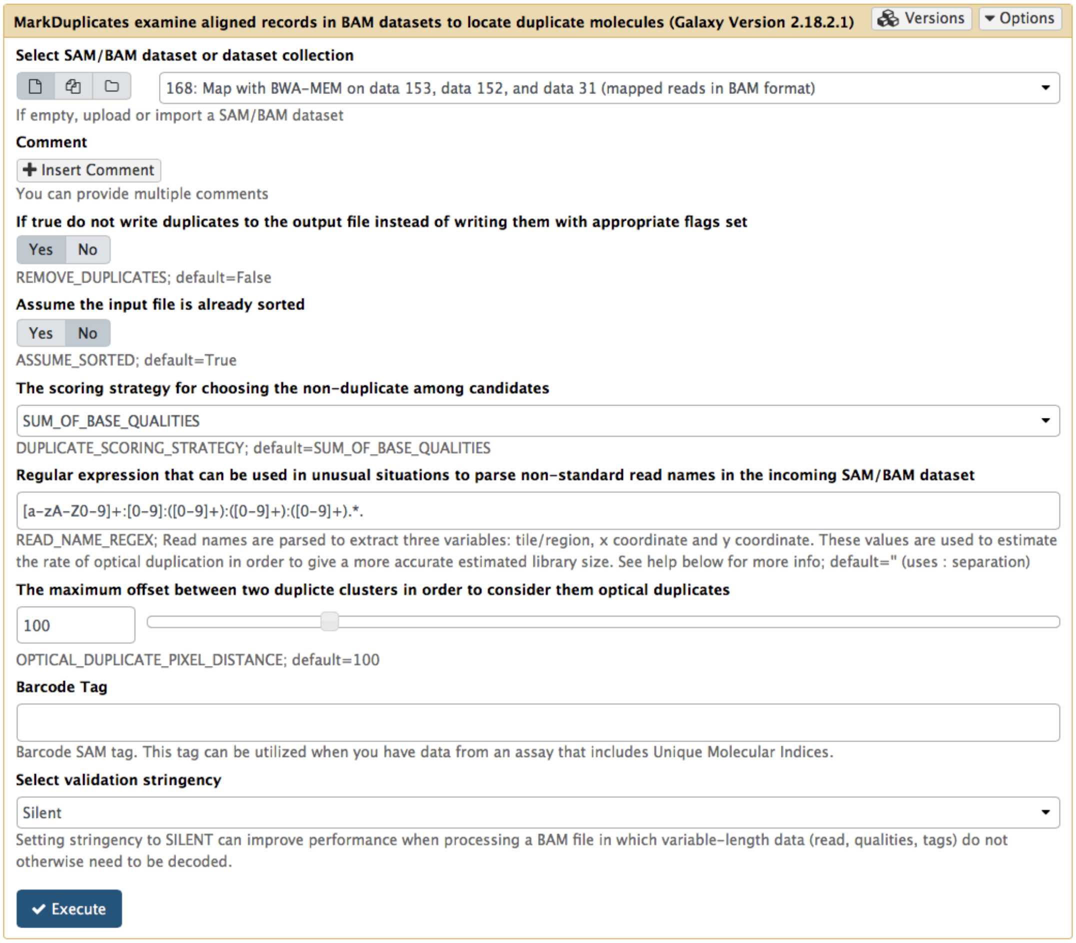


Note: IGV Visualization Tool

IGV can be used to visualize data with reference to a chromosome. The bottom track shows the genes on the chromosome including the introns and exons. Variants are visible in the reads shown above the gene. More information about each variant can be found by scrolling over the variant to see the cigar and phred scores. The total number of reads for each variant can be found including the number and percentage of reads for each type of base (A, T, C, G).

8. **Mark duplicates**

When looking for DNA sequence variants, using the Mark duplicates tool is important for weeding out duplicate (identical reads) that can introduce frequency and absolute number bias in variant calling. Duplicate short sequence reads start and end on the exact base in a BAM file and can be easily identified in your BAM outputs from BWA-mem in step 7. Duplicates will also have identical base scores.



MarkDuplicates examine aligned records in BAM datasets to locate duplicate molecules (Galaxy Version 2.18.2.1)

Select SAM/BAM dataset or dataset collection

168: Map with BWA-MEM on data 153, data 152, and data 31 (mapped reads in BAM format)

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

[✓ Yes] No

REMOVE_DUPLICATES; default=False

Assume the input file is already sorted

Yes [✓ No]

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

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

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)

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.

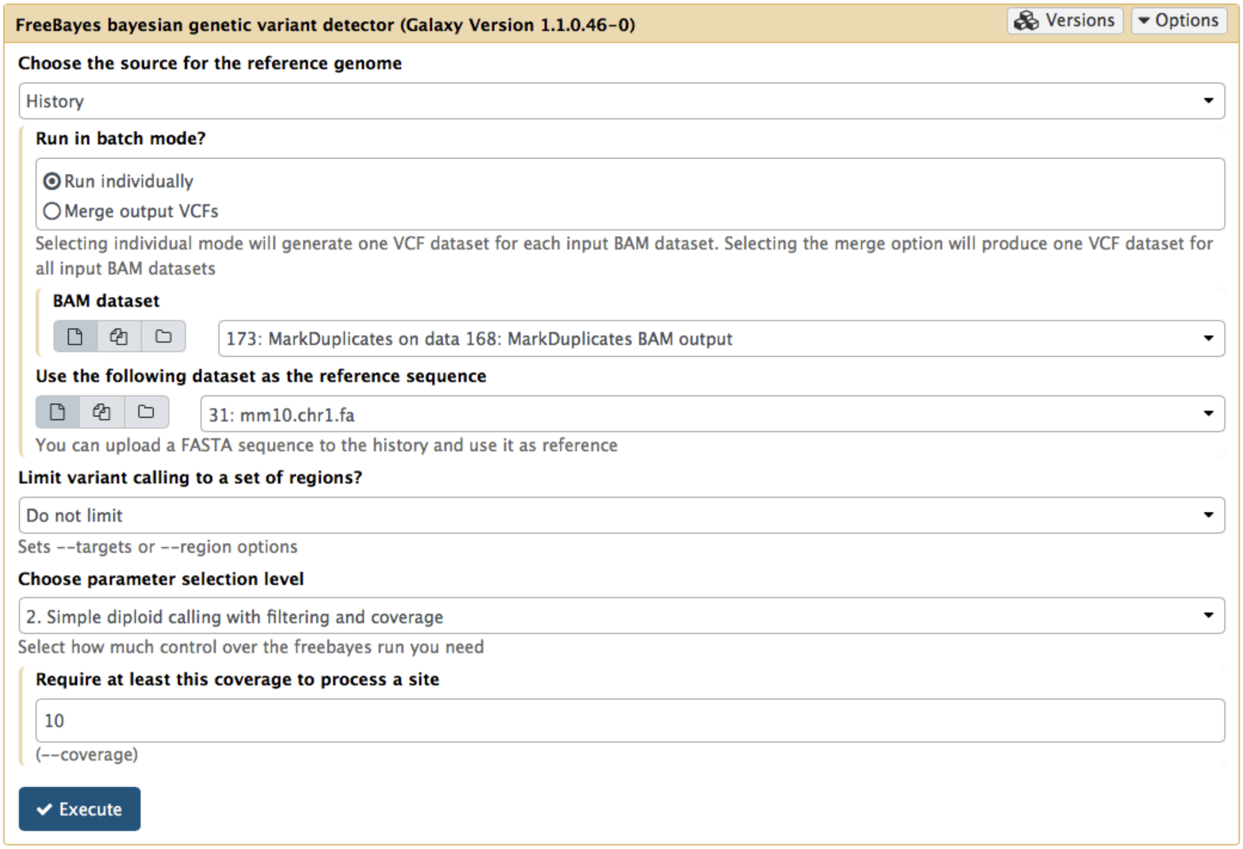
Select validation stringency

Silent

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.

9. Generating a Variant Call File

After aligning to a reference genome and removing duplicate reads, a VCF file is produced using FreeBayes. As we know that the Mouse variant maps to chromosome #1 we ask for sequence variants only on that chromosome. Sometimes you may want to ask FreeBayes to search variants in target regions by specifying a '.bed' file.



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

Choose the source for the reference genome

History

Run in batch mode?

● Run individually
○ Merge output VCFs

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

173: MarkDuplicates on data 168: MarkDuplicates BAM output

Use the following dataset as the reference sequence

31: mm10.chr1.fa

You can upload a FASTA sequence to the history and use it as reference

Limit variant calling to a set of regions?

Do not limit

Sets --targets or --region options

Choose parameter selection level

2. Simple diploid calling with filtering and coverage

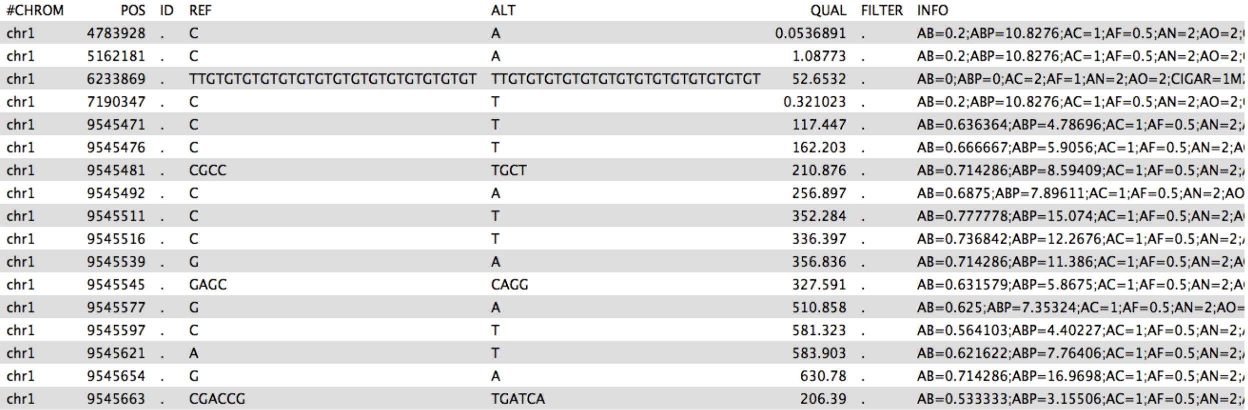
Select how much control over the freebayes run you need

Require at least this coverage to process a site

10

(--coverage)

A VCF file should include all genetic changes at a scale smaller than the imputed read sizes (~50-150 bp in length). The VCF file generated by FreeBayes is shown below. The VCF QUAL score is a Phred Quality score scaled to the probability that a base is incorrectly called.

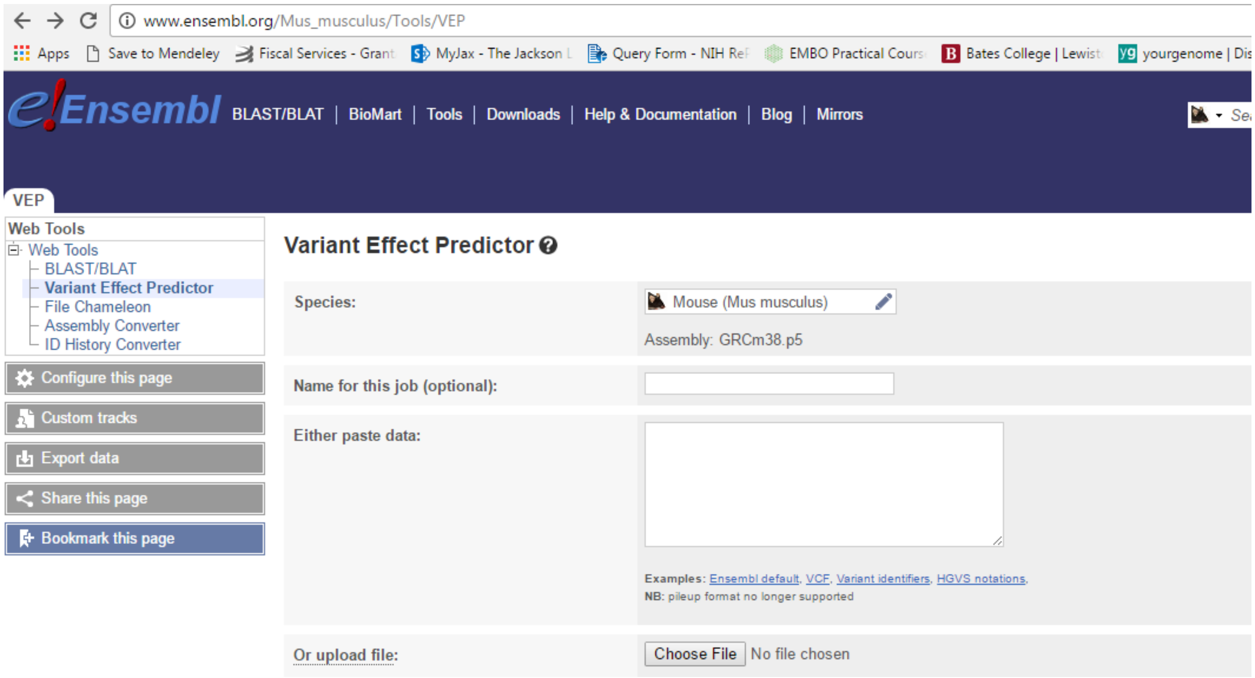


#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr1	4783928	.	C	A	0.0536891	.	AB=0.2;ABP=10.8276;AC=1;AF=0.5;AN=2;AO=2;
chr1	5162181	.	C	A	1.08773	.	AB=0.2;ABP=10.8276;AC=1;AF=0.5;AN=2;AO=2;
chr1	6233869	.	TTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	TTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	52.6532	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1M
chr1	7190347	.	C	T	0.321023	.	AB=0.2;ABP=10.8276;AC=1;AF=0.5;AN=2;AO=2;
chr1	9545471	.	C	T	117.447	.	AB=0.636364;ABP=4.78696;AC=1;AF=0.5;AN=2;
chr1	9545476	.	C	T	162.203	.	AB=0.666667;ABP=5.9056;AC=1;AF=0.5;AN=2;A
chr1	9545481	.	CGCC	TGCT	210.876	.	AB=0.714286;ABP=8.59409;AC=1;AF=0.5;AN=2;
chr1	9545492	.	C	A	256.897	.	AB=0.6875;ABP=7.89611;AC=1;AF=0.5;AN=2;AO
chr1	9545511	.	C	T	352.284	.	AB=0.777778;ABP=15.074;AC=1;AF=0.5;AN=2;A
chr1	9545516	.	C	T	336.397	.	AB=0.736842;ABP=12.2676;AC=1;AF=0.5;AN=2;
chr1	9545539	.	G	A	356.836	.	AB=0.714286;ABP=11.386;AC=1;AF=0.5;AN=2;A
chr1	9545545	.	GAGC	CAGG	327.591	.	AB=0.631579;ABP=5.8675;AC=1;AF=0.5;AN=2;A
chr1	9545577	.	G	A	510.858	.	AB=0.625;ABP=7.35324;AC=1;AF=0.5;AN=2;AO=
chr1	9545597	.	C	T	581.323	.	AB=0.564103;ABP=4.40227;AC=1;AF=0.5;AN=2;
chr1	9545621	.	A	T	583.903	.	AB=0.621622;ABP=7.76406;AC=1;AF=0.5;AN=2;
chr1	9545654	.	G	A	630.78	.	AB=0.714286;ABP=16.9698;AC=1;AF=0.5;AN=2;
chr1	9545663	.	CGACCG	TGATCA	206.39	.	AB=0.533333;ABP=3.15506;AC=1;AF=0.5;AN=2;

Not all the variation calls in a VCF file are correct or worth further exploration. Galaxy's FreeBayes tool, which produces a VCF file from a BAM file, only declares variations that are corroborated by at least 2 reads or 20% of reads, a relatively low threshold that allows for extraneous variation calls. So, it is important to view a VCF file in some sort of visualization software, such as IGV, which aligns the reads and variation calls against a reference genome, making it easier to see which variation calls are strong and which are weak. Galaxy also has a tool (slice VCF) that limits VCF data to a specific part of the genome as specified by a bed file. Variations are then used to identify sequence variants which may be deleterious.

10. Variant Identification using (Ensembl Variant Effect predictor, VEP)
http://www.ensembl.org/Mus_musculus/Tools/VEP

The first step is to download and save the VCF file produced by FreeBayes. In case your analysis goes awry a vcf file is available in the shared data library, named ‘FreeBayes on data 14 and data 13’. Upload the file through the ‘Choose File’ option on Ensembl VEP. **Important: for this exercise use Mouse mm10 as your reference. This will match the alignment to reference mm10; this is important.**



Variant Effect Predictor

Species: Mouse (Mus musculus)

Assembly: GRCm38.p5

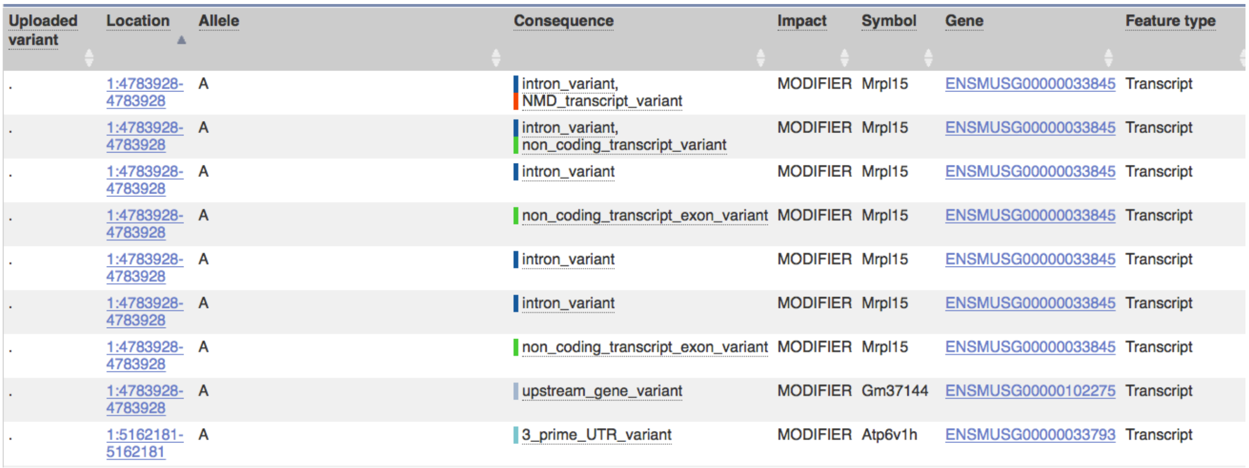
Name for this job (optional):

Either paste data:

Examples: Ensembl default, VCF, Variant identifiers, HGVS notations.
NB: pileup format no longer supported

Or upload file: Choose File No file chosen

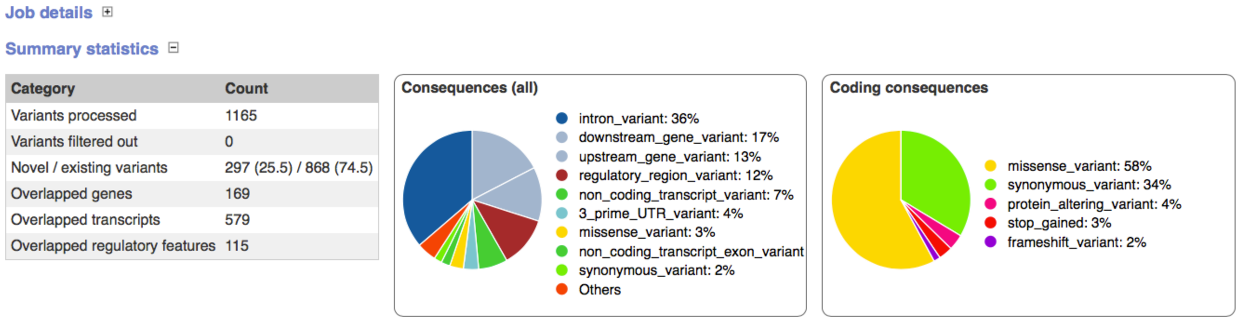
The VEP tool will look at the mouse genome in Ensembl and then compile a list of variants that are may be causative of a phenotype, in this case within the genetic region entered, mouse chr. 1. The annotated VEP file is below.



Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type
.	1:4783928-4783928	A	intron_variant, NMD_transcript_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	intron_variant, non_coding_transcript_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	intron_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	non_coding_transcript_exon_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	intron_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	intron_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	non_coding_transcript_exon_variant	MODIFIER	Mrpl15	ENSMUSG00000033845	Transcript
.	1:4783928-4783928	A	upstream_gene_variant	MODIFIER	Gm37144	ENSMUSG00000102275	Transcript
.	1:5162181-5162181	A	3_prime_UTR_variant	MODIFIER	Atp6v1h	ENSMUSG00000033793	Transcript

Info: Consequence describes the change in gene function. IMPACT type is the proposed extent of the amino shift.

This may take 3-5 minutes to calculate in VEP. Output in this module will look like this:



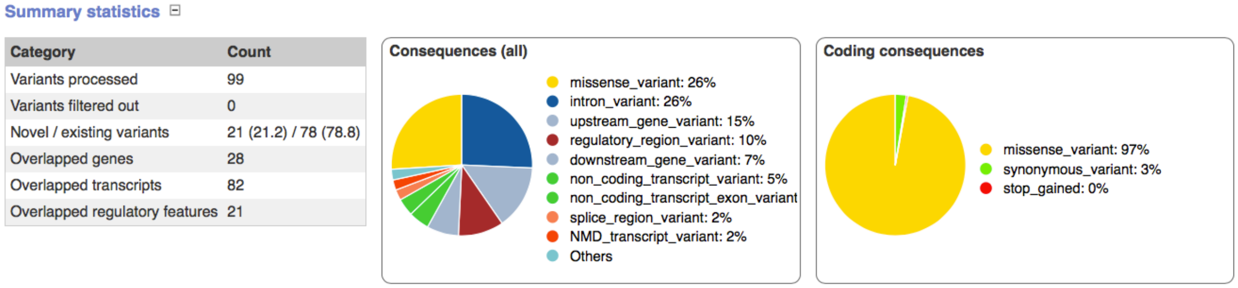
Job details ⊞

Summary statistics ⊟

Category	Count
Variants processed	1165
Variants filtered out	0
Novel / existing variants	297 (25.5) / 868 (74.5)
Overlapped genes	169
Overlapped transcripts	579
Overlapped regulatory features	115

Consequences (all)
● intron_variant: 36%
● downstream_gene_variant: 17%
● upstream_gene_variant: 13%
● regulatory_region_variant: 12%
● non_coding_transcript_variant: 7%
● 3_prime_UTR_variant: 4%
● missense_variant: 3%
● non_coding_transcript_exon_variant
● synonymous_variant: 2%
● Others
Coding consequences
● missense_variant: 58%
● synonymous_variant: 34%
● protein_altering_variant: 4%
● stop_gained: 3%
● frameshift_variant: 2%

The results of VEP can be sorted consequence or gene symbol by toggling between options in the Ensembl table. You will need to use the **Filters** function in the center to pull 'consequence' is 'missense_variant', :



Summary statistics ⊟

Category	Count
Variants processed	99
Variants filtered out	0
Novel / existing variants	21 (21.2) / 78 (78.8)
Overlapped genes	28
Overlapped transcripts	82
Overlapped regulatory features	21

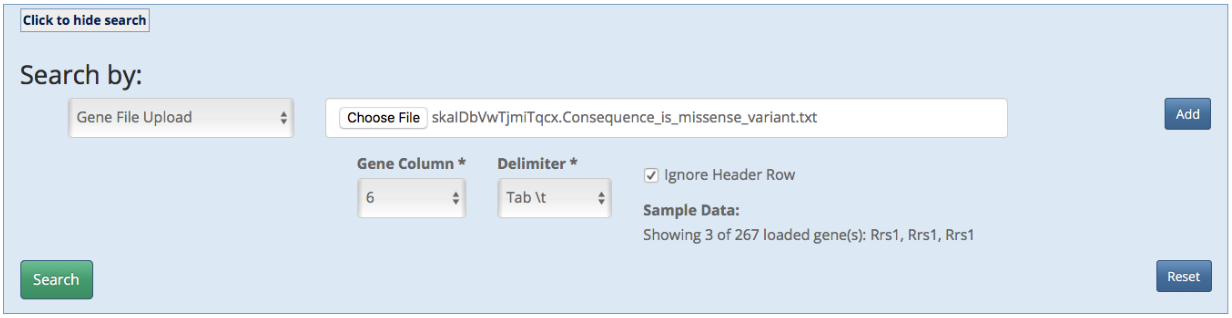
Consequences (all)
● missense_variant: 26%
● intron_variant: 26%
● upstream_gene_variant: 15%
● regulatory_region_variant: 10%
● downstream_gene_variant: 7%
● non_coding_transcript_variant: 5%
● non_coding_transcript_exon_variant
● splice_region_variant: 2%
● NMD_transcript_variant: 2%
● Others
Coding consequences
● missense_variant: 97%
● synonymous_variant: 3%
● stop_gained: 0%

Once you have missense variants you should download the text file (TXT) in the download section here just to the left of the filter section.

You will use the gene symbols/names to see if any variants found are connected to the mouse phenotype we see. In the download the gene symbol/names are most likely in column 6 or 7. In this case they are in column 6.

11. Go to the JAX Mouse Genome Informatics Human-Mouse Genome Connection tool: http://www.informatics.jax.org/humanDisease.shtml

You will need to use “Gene File Upload” setting and identify what column the gene names fall in. Should be column 6.



Search by:

Gene File Upload

Choose File skalDbVwTjmiTqcx.Consequence_is_missense_variant.txt

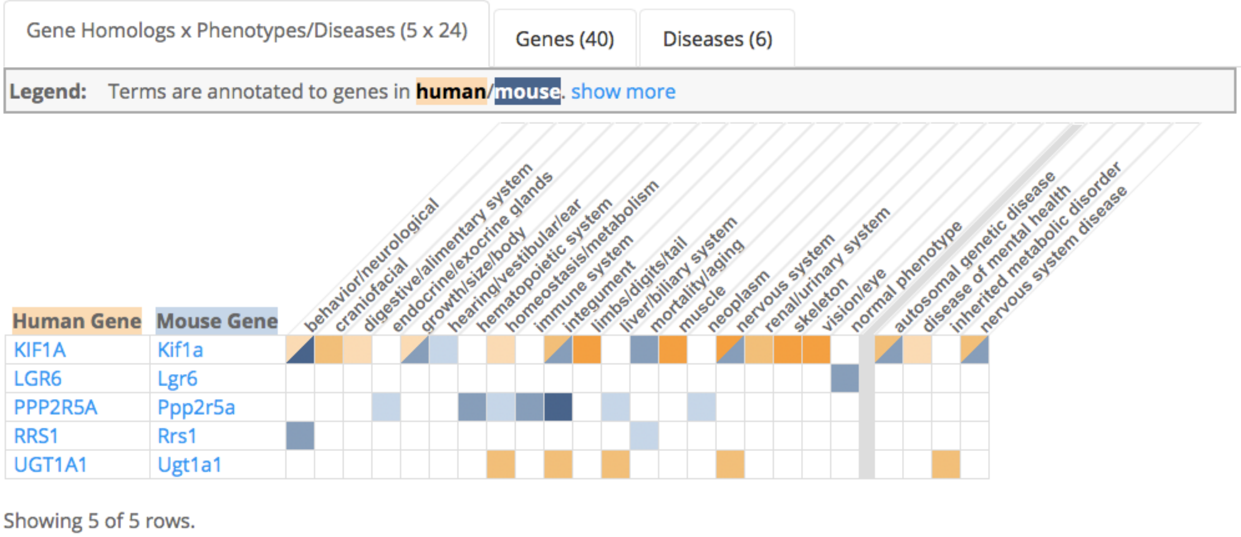
Gene Column * 6

Delimiter * Tab \t

☑ Ignore Header Row

Sample Data:
Showing 3 of 267 loaded gene(s): Rrs1, Rrs1, Rrs1

The Human-Mouse Disease Gene query will give you a list of human and mouse genes linked to disease from your variant call/VCF file. These are the exome variants with High consequence from Mouse Chromosome #1.



Legend: Terms are annotated to genes in human/mouse. show more

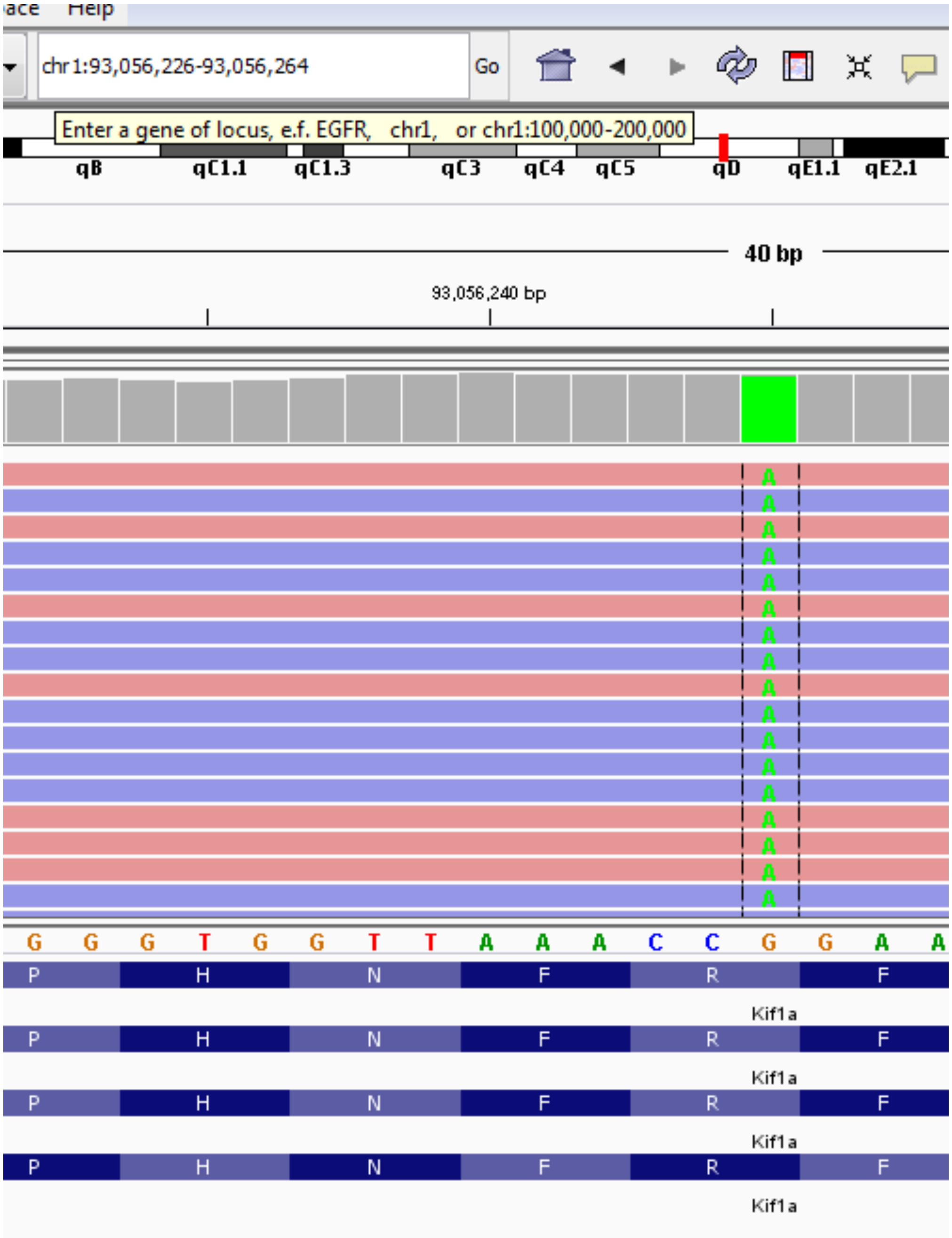
Human Gene	Mouse Gene	behavior/neurological	craniofacial	digestive/alimentary system	endocrine/exocrine glands	growth/size/body	hearing/vestibular/ear	hematopoietic system	homeostasis/metabolism	immune system	integument	limbs/digits/tail	liver/biliary system	mortality/aging	muscle	neoplasm	nervous system	renal/urinary system	skeleton	vision/eye	normal phenotype	autosomal genetic disease	disease of mental health	inherited metabolic disorder	nervous system disease
KIF1A	Kif1a	◩	🟧	🟧		◩	🟦		🟧		◩	🟧		🟦	🟧		◩	🟧	🟧	🟧		◩	🟧		◩
LGR6	Lgr6																				🟦				
PPP2R5A	Ppp2r5a				🟦			🟦	🟦	🟦	🟦		🟦			🟦									
RRS1	Rrs1	🟦												🟦											
UGT1A1	Ugt1a1								🟧		🟧		🟧				🟧							🟧	

Showing 5 of 5 rows.

Students should scan anatomical systems to look for limb phenotypes by gene. By scanning down one will note that Human gene KIF1A and mouse gene Kif1A are implicated in a wide range of phenotypes; in the yellow to blue boxes, the darker the color in the square the more evidence that exists for a gene to disease connection.

Alternatively one can apply a filter on the phenotypes/disease and see what genes might be connected to weak back legs or ‘leg dragger’, in the mouse. Filtering on limbs and neurological/behavioral will limit the number of genes.

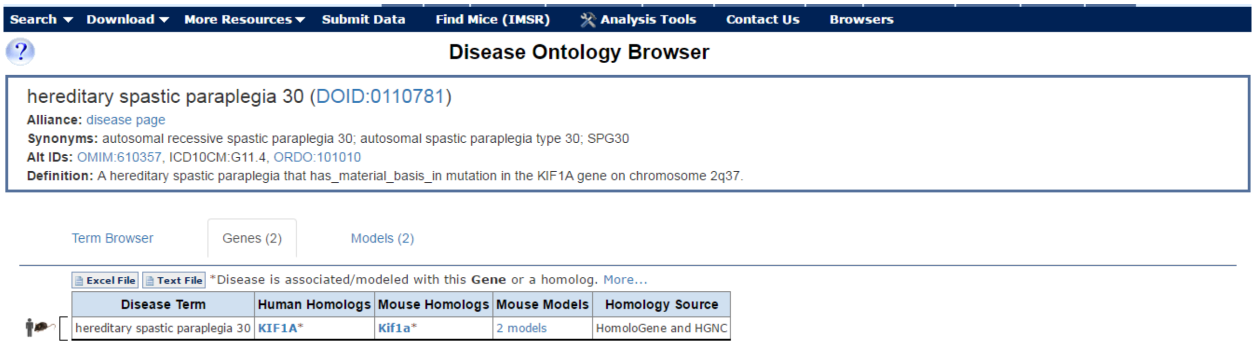
By filteringon phenotype disease two genes Dst and Kif1a will appear to be involved in limbs and neurological conditions. Kif1a appears to be the gene altered in the mouse we started with.



chr1:93,056,226-93,056,264

Enter a gene of locus, e.f. EGFR, chr1, or chr1:100,000-200,000

12. Students can also use OMIM to investigate genes involved in the apparent phenotype: ‘spastic paraplegia 30’. In doing so they will find OMIM record 610357



Disease Ontology Browser

hereditary spastic paraplegia 30 (DOID:0110781)

Alliance: disease page

Synonyms: autosomal recessive spastic paraplegia 30; autosomal spastic paraplegia type 30; SPG30

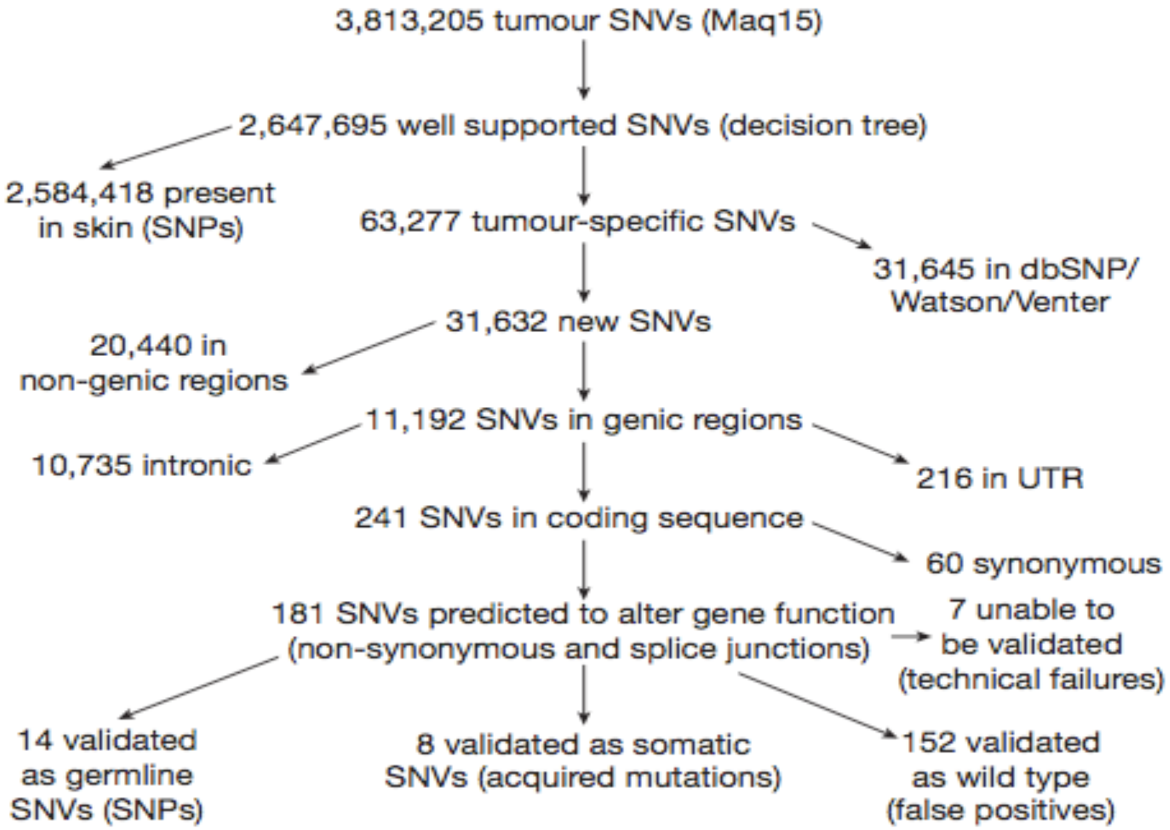
Alt IDs: OMIM:610357, ICD10CM:G11.4, ORDO:101010

Definition: A hereditary spastic paraplegia that has_material_basis_in mutation in the KIF1A gene on chromosome 2q37.

*Disease is associated/modeled with this **Gene** or a homolog. More...

Disease Term	Human Homologs	Mouse Homologs	Mouse Models	Homology Source
hereditary spastic paraplegia 30	**KIF1A***	**Kif1a***	2 models	HomoloGene and HGNC

Additional Context:
Finding a needle in a haystack:


3,813,205 tumour SNVs (Maq15)
2,647,695 well supported SNVs (decision tree)
2,584,418 present in skin (SNPs)
63,277 tumour-specific SNVs
31,645 in dbSNP/ Watson/Venter
31,632 new SNVs
20,440 in non-genic regions
11,192 SNVs in genic regions
10,735 intronic
216 in UTR
241 SNVs in coding sequence
60 synonymous
181 SNVs predicted to alter gene function (non-synonymous and splice junctions)
7 unable to be validated (technical failures)
14 validated as germline SNVs (SNPs)
8 validated as somatic SNVs (acquired mutations)
152 validated as wild type (false positives)