

FINAL PORTFOLIO

Bioinformatics: Tools for Genome Analysis
Spring 2021

*Ryan Yancey
15 May 2021*

Statement of Purpose

This document serves as an aggregate of materials learned from Professor Sinjung Yun's and Professor Sajgun Yun's course on the tools and software available for bioinformatic analysis of genomes and genome variation at the Johns Hopkins University Department of Biotechnology.

The content ranges from open reading frame prediction in prokaryotic organisms, to nucleotide variant calling and ChIP-seq peak calling with real sequencing data examples for proof of concept. Tools are described to achieve various objectives and links to these sources are provided in-line *via* hyperlink wherever possible.

The structure of each section in the following table of contents will begin with a brief introduction of the topic covered, followed by an overview of the objectives accomplished for each unit. Then, a summary of the material covered will be presented prior to a proof-of-concept application of a selection of those materials covered. If provided, supplementary material for each unit will be referenced by in-line superscript citation and full references will be provided in the *References* section of this document.

Source code and resulting files in this text are available at:

<https://github.com/ryancey1/Final-Portfolio>

Table of Contents

<u>Section</u>	<u>Page</u>
I. Gene Prediction of Prokaryotic Organisms	4
<i>Introduction</i>	4
<i>Objectives</i>	4
<i>Summary</i>	4
<i>Supplemental Reading</i>	5
<i>Application: Using command line Glimmer3 to predict CDS in Spiroplasma helicoides strain TABS-2</i>	5
II. Gene Prediction of Eukaryotic Organisms	6
<i>Introduction</i>	6
<i>Objectives</i>	6
<i>Summary</i>	6
<i>Supplemental Reading</i>	9
<i>Application: Aligning a cDNA sequence to a contig using Splign</i>	9
<i>Application: Creating and uploading BED files to GDV</i>	10
III. Database Browsers and Data Retrieval	11
<i>Introduction</i>	11
<i>Objectives</i>	11
<i>Summary</i>	11
<i>Supplemental Reading</i>	13
<i>Application: Rscript to demonstrate biomaRt use</i>	13
IV. Genome Browsers and Data Retrieval	14
<i>Introduction</i>	14
<i>Objectives</i>	14
<i>Summary</i>	14
<i>Supplemental Reading</i>	16
<i>Application: Using the UCSC Table Browser to retrieve RefSeq coordinates</i>	16
V. Genome Variation: Analysis Platforms	18
<i>Introduction</i>	18
<i>Objectives</i>	18
<i>Summary</i>	18
<i>Supplemental Reading</i>	21
<i>Application: Using NCBI Variation Viewer to explore KCNJ11 variants</i>	21
VI. Genomic Data Files: Description and Manipulation	23
<i>Introduction</i>	23
<i>Objectives</i>	23

	<i>Summary</i>	23
	<i>Supplemental Reading</i>	24
	<i>Application: Converting SAM to BAM and uploading to IGV</i>	25
VII.	DNA Elements and the ENCODE Project	26
	<i>Introduction</i>	26
	<i>Objectives</i>	26
	<i>Summary</i>	26
	<i>Supplemental Reading</i>	28
	<i>Application: Exploring ENCODE data at the SOX11 promoter to predict expression patterns</i>	29
VIII.	Non-Coding RNA and Ultra-Conserved Regions	30
	<i>Introduction</i>	30
	<i>Objectives</i>	30
	<i>Summary</i>	30
	<i>Supplemental Reading</i>	32
	<i>Application: Using multiple tools to analyze an ultraconserved element</i>	33
IX.	Next Generation Sequencing	35
	<i>Introduction</i>	35
	<i>Objectives</i>	35
	<i>Summary</i>	35
	<i>Supplemental Reading</i>	38
	<i>Application: Identifying SNPs in NGS data from the 1000 Genomes Project using Galaxy</i>	39
X.	Chromatin Immunoprecipitation Sequencing Analysis (ChIP-seq)	41
	<i>Introduction</i>	41
	<i>Objectives</i>	41
	<i>Summary</i>	41
	<i>Supplemental Reading</i>	42
	<i>Application: Using Galaxy to call peaks with MACS2 on mouse ChIP-seq data</i>	43
XI.	RNA Sequencing Analysis (RNA-seq)	45
	<i>Introduction</i>	45
	<i>Objectives</i>	45
	<i>Summary</i>	45
	<i>Supplemental Reading</i>	46
	<i>Application: Using Galaxy to assess differential expression of RNA-seq data from two mouse cell lines</i>	47
XII.	References	51

I. Gene Prediction of Prokaryotic Organisms

Introduction

Bacterial genes are structured in a polycistronic way. That is, one mRNA transcript can be translated into multiple proteins. Since there are no introns in prokaryotic genomes, it is relatively accurate to annotate them with open reading frame (ORF) finding algorithms. The longest ORF is usually the operon for a polycistronic region. This section discusses the annotation process in detail and applies the concepts.

Objectives

- Run Glimmer on the command line.
- Use the ORF Finder to find bacterial open reading frames.
- Run BPPROM, Neural Network Promoter Prediction.
- Perform basic sequence analysis using seqinr.

Summary

Prokaryotic gene structure

- Bacterial mRNA is polycistronic - one transcript codes many genes
- Operons are made up of one promoter and many coding regions
- Usually only a few nucleotides between coding regions

Early annotations

- ORF finding originally used to annotate prokaryotic gDNA
- Thought process: longest ORF is usually CDS because no introns
- However, where does ORF begin? Many candidates for start codon within some ORFs
- RNA prediction is difficult. Predictors can distinguish coding DNA from noncoding DNA, but transcription start sites (TSS) or transcription termination sites (TTS) are harder

Transcript data

- RNA-seq is a worthy complement to gene sequencing
- Helps to complete gene annotation
- Good at finding mRNA locations, but not CDS
- Gene predictors + transcript data are commonly used in modern annotation pipelines

ORF Finder

- Available [here](#)
- Shows all ORFs
- usually non-overlapping and longest are real genes

FGENESB

- Available [here](#)

- Automatic bacterial genome annotation
- Based on Markov chain models
- Can determine CDS, promoters, operons, etc

GLIMMER

- UNIX-based, installed on JHU BFX server
- Finds long ORFs (since they're usually coding regions)
- Can utilize training set to annotate new sequences

Other gene-finding programs

- [GeneMark](#)
- [EasyGene 1.2b](#)

Supplemental Reading

- Automated prokaryotic genome annotation pipeline¹
- Structural features as basis for predicting transcript category²
- Gene cluster prediction using deep learning³

Application: Using command line Glimmer3 to predict CDS in *Spiroplasma helicoides* strain TABS-2

Command line arguments:

```
$ ls
sheli.fasta  sheliprt.fasta

$ long-orfs -n -t 1.15 sheli.fasta sheli.longorfs

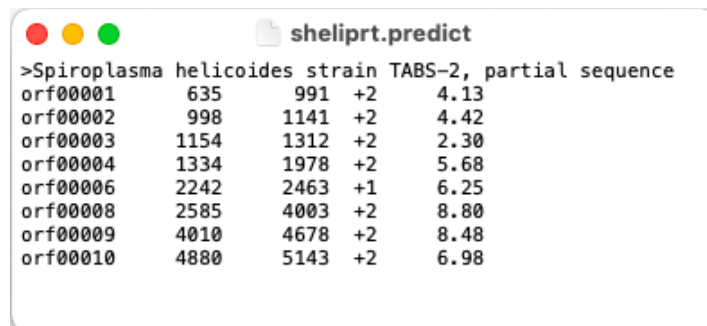
$ extract -t sheli.fasta sheli.longorfs > sheli.train

$ build-icm -r sheli.icm < sheli.train

$ glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt

$ extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
```

Resulting .predict file containing CDS predictions:



```
sheliprt.predict
>Spiroplasma helicoides strain TABS-2, partial sequence
orf00001      635      991  +2      4.13
orf00002      998     1141  +2      4.42
orf00003     1154     1312  +2      2.30
orf00004     1334     1978  +2      5.68
orf00006     2242     2463  +1      6.25
orf00008     2585     4003  +2      8.80
orf00009     4010     4678  +2      8.48
orf00010     4880     5143  +2      6.98
```

II. Gene Prediction of Eukaryotic Organisms

Introduction

While open reading frame finding works quite well for prokaryotic gene prediction, eukaryotic gene prediction is slightly more complex. Due to the presence of introns, ORF finding will inaccurately predict the length of genes and potentially predict truncated genes due to intronic stop codons. This section covers different methods for eukaryotic, intron-aware gene prediction models. Expression-based prediction is the most accurate, but its pitfall is in its reliance on multiple steps where error may be introduced. Machine learning has become more important in gene prediction and annotation of eukaryotic genomes recently^{4,5}.

Objectives

- Run *ab initio* gene prediction.
- Explain when *ab initio* methods are used instead of expression methods.
- Use comparative genomics to find exons.
- Explain how comparative genomics is used to find exons.
- Align mRNA to genomic DNA using Spidey, Splign, BLAT.
- Annotate mRNA locations after alignment of mRNA to genome.
- Annotate CDS locations after mRNA/genome alignment.
- Describe the BED and WIG file formats.
- Create BED files using a text editor.
- Load BED and WIG files at NCBI.

Summary

Expression-based gene prediction

- cDNA/protein aligned to genomic DNA
- Precisely defines CDS and alternative transcripts
- CONS: 60% to 80% of DNA is rarely expressed
 - Difficult to get cDNA/map

Single genome de novo gene prediction

- Exon locations can be predicted
- Works where expression-based fails
- Early programs were error prone ([GENSCAN](#))
- Newer algorithms: [HMMGene](#), [FGENESH](#), [Augustus](#)

Content sensors

- Algorithms dependent on features within the CDS (nt composition, etc)
- Wobble base in bacteria is much more variant in non-coding DNA, so it can be used to predict CDS
- Hexamer frequency is most reliable
- Codon bias + dicodon combinations

- HMMs use training sets to determine hexamer patterns and find similar patterns among unknown sequences

Signal sensors

- Promoters, splice sites, stop codons, poly-A tails
- Weight matrix PSSM identifies conserved positions in splice sites/promoters
- GT/AG to start/stop introns, by example

Output of Eukaryotic Gene Prediction

- We look for exons
- However UTRs are typically ignored, so predictions typically start with Start codon and end with the stop codon on first and last exons, respectively

Comparative genomics

- Uses slightly dissimilar genomes to identify conserved regions which are most likely coding DNA⁶

Expression-based Prediction

Splign

- Available [here](#)
- Aligns cDNA to genomic DNA
- Identifies splice-junction locations, frameshifts, alternative gene models where possible
- Supports cross-species alignment

Genomic BLAST

- Available [here](#)
- EST alignment to genomic DNA
- Shows there might be a gene located there

BLAT

- Available [here](#)
- Analogous to Splign
- Advantage over Splign is that it can be used to align anywhere in the genome, not just a specified chunk of DNA

Entrez Genomes database

- Complete genome database
- Can examine visually

NCBI Genome Data Viewer (GDV)

- Available [here](#)
- Genome browser supporting RefSeq genomes
- GEO and other NCBI resources utilize

Variation Viewer

- Available [here](#)
- Main focus is human genome
- Similar browser to NCBI Gene Database browser

BED Format (short)

- “Browser Extensible Data”
- 3-column table
- Row represents a single entity
- BED3
 1. Chromosome name (chrom)
 2. Start nucleotide (start)
 3. End nucleotide (end)
- Tab separated
- Can be 0-based or 1-based (important to know which yours is!)

BED Format (long)

- BED6 and BED12 (6-column and 12-column)
- BED6
 1. Chrom
 2. Start
 3. End
 4. Name
 5. Score
 6. Strand
- BED12
 7. thickStart
 8. thickEnd
 9. itemRgb
 10. blockCount
 11. blockSizes
 12. blockStarts
- Resource available [here](#)

WIG Format

- “Wiggle” format
- Useful for storing continuous data
- Resource available [here](#)
- Needs a header to be displayed in genome browser

BigWig Format

- Compressed WIG file
- Load much faster in genome browsers because they omit irrelevant info

Supplemental Reading

- The strawberry genome was re-annotated using transcript level annotation, in order to update the previous annotation which used predictors and EST data⁷

Application: Aligning a cDNA sequence to a contig using Splign

Running *Splign* against a PQBP1 transcript NM_144495.3 and human genome sequencing contig AC233300:

Input:

cDNA: NM_144495.3

Genomic: From: 1 To: max

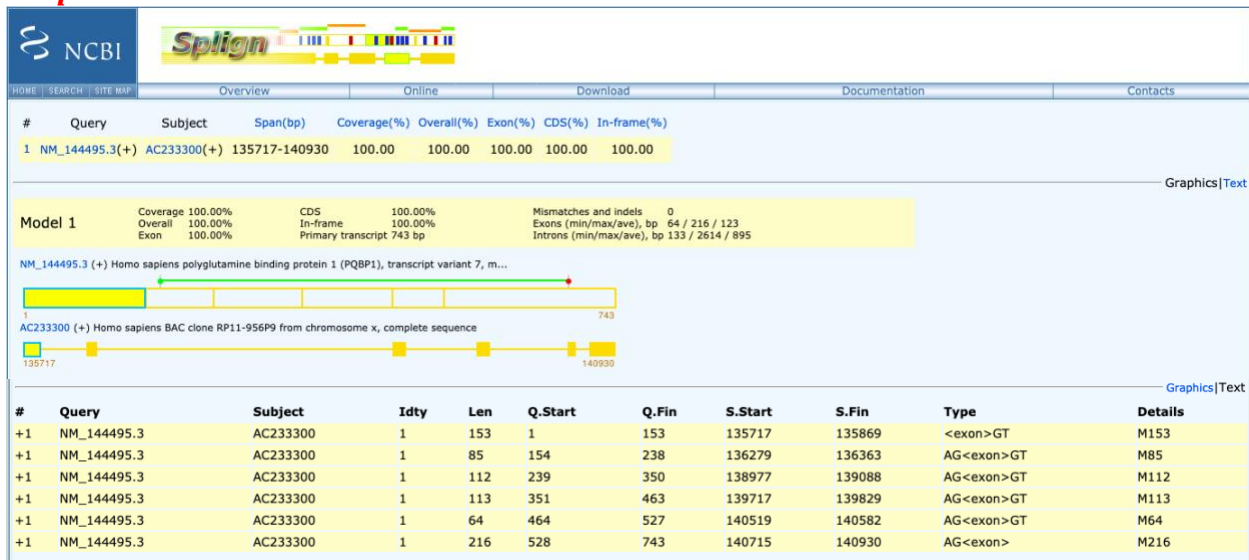
AC233300

☐ Lower quality query sequence (e.g. EST)
☐ Reverse and complement the query
☐ More partial alignments
☐ Use discontinuous megablast (e.g. for cross-species)

No file chosen

Whole genome: Not selected

Output:



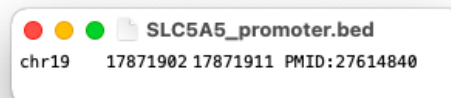
There are a total of 6 exons that map from the NM_144495.3 cDNA to the human genome contig. The green bar shows the predicted CDS, and the box below that represents the mRNA transcript (note the absence of introns). The box below that represents the alignment to the whole contig – gaps in between the boxes represent introns. The table below is a textual representation of the results.

Application: Creating and uploading BED files to GDV

A paper describes a promoter region of the human NIS gene⁸. The official symbol of the NIS gene is SLC5A5. Find the coordinates of the transcription start site for this gene. The promoter is said to be 38 bases upstream from that TSS. Given that location, create a single-line BED file with a zero-based start. The single entry in this file should represent a 9-nucleotide region centered on the nucleotide at position -38 from the TSS. The final region should be nine nucleotides long. Load the BED file onto the Variation Viewer at NCBI and narrow the view to the promoter and the first exon of this gene. Discuss your results in the Discussion area.

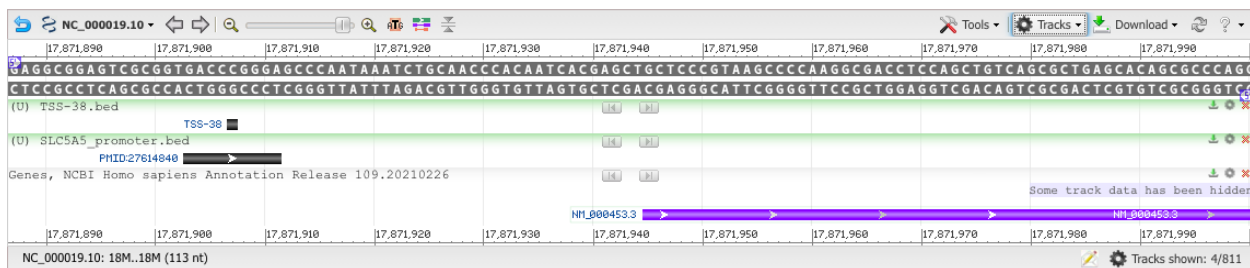
SLC5A5 is a gene present on human chromosome 19, spanning from chr19:17,871,394..17,895,175 on the chromosome⁹. However, the full region of mRNA does not encompass this whole region, rather it runs from chr19:17,871,945..17,895,174.

Thus, the transcription start site (TSS) is at nucleotide chr19:17,871,945 and the -38 position from there is chr19:17,871,907. Centering on that nucleotide, ± 4 nt from there yields the range: chr19:17871903-17871911. Compensating the start position to be zero-based, the resulting BED file would then take the following format:



SLC5A5_promoter.bed
chr19 17871902 17871911 PMID:27614840

Uploaded to the GDV with the -38nt marked:



III. Database Browsers and Data Retrieval

Introduction

Databases are common in bioinformatics, and for eukaryotic genomic data, the Ensembl database and browser are commonly used. Though it was originally designed for chordates, non-chordate *eukaryotes* have been added to the database.

The goal statement from their website: *“The goal of Ensembl [is to] automatically annotate the genome, integrate this annotation with other available biological data and make all this publicly available via the web.”*

Objectives

- Download data sets from BioMart.
- Retrieve data from biomaRt using R.
- Design searches using R to retrieve data.

Summary

Ensembl

- Provides genome browser for non-prokaryotic organisms
- Data are stored in MySQL relational databases
- Custom view/bookmarks can be saved for sharing results/publication

Homepage

- Searches can come from the [homepage](#)
 - Search form: *species* + gene
- Searches can also come from species-specific index ([Here](#), for example)
 - Search form: gene

Chromosome Views

- Chromosome Summary: bar-charts demonstrate a good visual for coding/noncoding genes and pseudogenes
- Region in Detail: narrower view of chromosome

Gene Identifiers

- ENSG = “ENSEMBL Genes” (only ENSG for humans)
- ENS**MUS**G = “ENSEMBL MUS musculus Genes” (mouse genes)
- ENS**DAR**G = “ENSEMBL DANIO rerio Genes” (Zebrafish genes)
- Species have longer identifiers that identify the organism

Location Tab

- Specific for chromosomal location view
- Other views include synteny maps and links to other genome browsers

Gene Tab

- Info on a specific gene
- All known transcripts/proteins for this gene are in this tab
- Gene based displays include variation tables, comparative genomics tables, and ontological specification

Transcript Tab

- Similar to gene tab, but specific to a single transcript of a gene
- 3D protein models can be accessed from this tab (if they're available)

Data Retrieval

BioMart

- Interface for downloading data sets from Ensembl
- Available [here](#)
- To start, one will choose the dataset to use and the organism to search

Filters

- From there, filters are used to narrow down the results
- Many categories exist, and options are even deeper within those categories
 - Region: Restrict to a specific area on a chromosome
 - Gene: Can specify what type of genes (disease associated, those only with specific IDs, etc)

Attributes

- Filters are used to narrow down to those genes of interest
- Attributes are then used to format the output (what will the table look like)

Preview

- Self explanatory
- Can be used to download either ALL results, or UNIQUE results (many genes will have multiple transcripts)

Sequences

- Once genes are filtered, click "Sequences" on attributes page and FASTA formatted sequences can be downloaded
- The header of the FASTA sequence can be customized for the information needed

Bioconductor and biomaRt data mining

Bioconductor project

- Open source initiative to mine data in a way which makes it easy to manipulate the data
- Tools available for analyzing/visualizing high-throughput data

biomaRt

- Bioconductor's R-based version of BioMart
- Video walkthrough available [here](#)

Supplemental Reading

- In 2016, Yates *et al* released an update to Ensembl and subsequently published a paper describing the new features¹⁰
- Bioconductor biomaRt [user guide](#)¹¹

Application: Rscript to demonstrate biomaRt use

I published an RMarkdown document which illustrates the process of constructing a query to biomaRt in an Rscript and verifies the results by comparing to BioMart exported results. The RMarkdown is published here: https://rpubs.com/ryancey3/unit3-4_HW_Q4

And the regular markdown file is also pushed to my GitHub [here](#)

IV. *Genome Browsers and Data Retrieval*

Introduction

The UCSC Genome Browser is a public tool hosted by the University of California, Santa Cruz. Currently, 97 genomes are available to browse and search. Tracks can be overlaid over gene annotation which display experimental data. Some of these tracks include, but are not limited to assembly data, comparative genomics, expression data, SNP and CNV data, as well as histone modifications and potential DNA regulatory elements. Other tools available through UCSC include the Table Browser and BLAT. Another tool available for genome browsing is the Integrative Genomics Viewer, hosted by the Broad Institute. Here, we include a brief summary of tracks and tools available before demonstrating some of their use.

Objectives

- Retrieve data from the UCSC Table Browser using Galaxy.
- Compute expressions to determine interval lengths.
- Create histograms on mined data.
- Operate on intervals using Join and Subtract features.
- Design and solve problems using Galaxy.
- Retrieve data from server at IGV.
- Load genomes from IGV server.
- Add data tracks to hg38 on IGV server.
- Load BED and WIG files to IGV.
- Design problems on IGV.

Summary

UCSC Genes

- UCSC genes compiles data from RefSeq, GenBank, and UniProt to predict genes
- Considered accurate
- Conservatively imports data from other databases

Conservation Track

- Gene conservation track available for ~100 species
- Allows user to quickly view levels of gene conservation
- Add/subtract species by clicking tab to the left of the track

PhastCons

- Some species have genome-wide alignments
 - Easy way to find conserved regions
- Aligned with blastz
- phastCons computes the conservation scores
- Two state phylogenetic HMM
 - One for conserved regions

- Another for non-conserved regions
- Genome broken up and max likelihood analyzes data
 - 1 Mb fragments

Misc tracks

- SNPs
- Gene Expression data
- ENCODE project data
- many others

Track Display

- Each track has 5 display modes:
 1. Hide – no display
 2. Dense – collapsed to a single line
 3. Full – annotations separated to individual lines
 4. Squish – same as full, but unlabeled and at 50% height
 5. Pack – same as squish, but with labels

Human Assemblies

- Default for human genome is GRCh38/hg38
- Good gene/mRNA tracks
- Not all tracks have been mapped from hg19 to hg38

SNPs

- Can view where an SNP lies in protein's 3D structure if the structure is known
- Click on links to "Mappings to PDB protein structures"

Table Browser

- Way to download UCSC data
- [Table Browser User Guide](#)

Galaxy

- Open-source framework
- Integrates command line tools/database information to perform scientific workflows
- Can be accessed through usegalaxy.org, a local Galaxy install, or cloud-based Galaxy

Data Conversion

- Galaxy can convert data from one format to another
 - These formats then can be accepted by other operations Galaxy provides

usegalaxy.org

- Public Galaxy website (available [here](#))
- Online tutorials ([screencasts](#)) exist for many basic functions.
 - [FASTQ file prep](#)
 - [Galaxy 101](#)

Integrative Genomics Viewer (IGV)

- Available as a local download
- Internet connection is required to load public datasets
- User data can also be loaded

Initial Projects

- IGV developed to view TCGA and 1000Genomes projects
- First release was in August 2008

Viewing Data

- Data are either continuous or annotations
- Data can be loaded by file, server, or URL specific

Supplemental Reading

- UCSC Genome Browser 2021 update¹²

Application: Using the UCSC Table Browser to retrieve RefSeq coordinates

1. **Navigate to the [UCSC Table Browser](#)**
2. **Construct a query with the following parameters:**
 - genome:** *Human*
 - assembly:** *Dec. 2013 (GRCh38/hg38)*
 - group:** *Genes and Gene Prediction Tracks*
 - track:** *NCBI RefSeq*
 - region:** *position chr7:27090000-27139300*
 - output format:** *BED - browser extensible data*

The screenshot shows the UCSC Table Browser interface with the following settings:

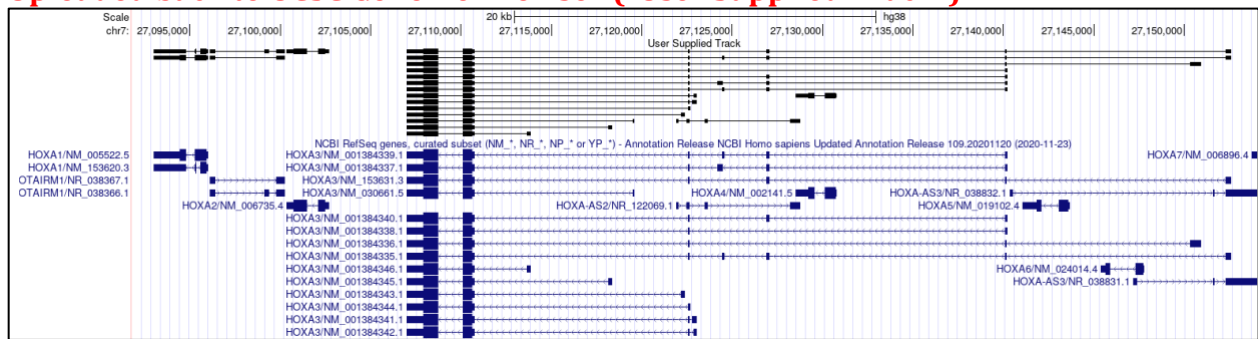
- clade:** Mammal
- genome:** Human
- assembly:** Dec. 2013 (GRCh38/hg38)
- group:** Genes and Gene Predictions
- track:** NCBI RefSeq
- table:** RefSeq All (ncbiRefSeq)
- region:** position chr7:27,090,000-27,139,300
- identifiers (names/accessions):** (empty)
- filter:** (empty)
- subtrack merge:** (empty)
- intersection:** (empty)
- correlation:** (empty)
- output format:** BED - browser extensible data
- Send output to:** ☐ Galaxy ☐ GREAT
- output file:** (empty)
- file type returned:** ☒ plain text ☐ gzip compressed

Buttons at the bottom include "get output" and "summary/statistics". A note at the bottom states: "To reset all user cart settings (including custom tracks), [click here](#)."

Click the Get Output button. Then Get BED. The Table Browser will display the records for the RefSeq accessions NM_005522, NM_153620, NR_038367, NR_038366.

chr7	27092992	27096000	NM_005522.5	0	-	27094439	27095912	0	2	1803,740,	0,2268,
chr7	27092992	27096000	NM_153620.3	0	-	27095295	27095912	0	3	1803,95,442,	0,2268,2566,
chr7	27096093	27100258	NR_038367.1	0	+	27100258	27100258	0	2	295,480,	0,3685,
chr7	27096093	27100258	NR_038366.1	0	+	27100258	27100258	0	3	295,269,480,	0,3004,3685,
chr7	27100353	27102683	NM_006735.4	0	-	27100725	27102500	0	2	1112,574,	0,1756,
chr7	27107009	27113842	NM_001384346.1	0	-	27107914	27110640	0	3	1711,646,218,	0,3105,6615,
chr7	27107009	27118357	NM_001384345.1	0	-	27107914	27110640	0	3	1711,646,211,	0,3105,11137,
chr7	27107009	27119595	NM_030661.5	0	-	27107914	27110640	0	3	1711,646,80,	0,3105,12506,
chr7	27107009	27122372	NM_001384343.1	0	-	27107914	27110640	0	3	1711,646,224,	0,3105,15139,
chr7	27107009	27122708	NM_001384344.1	0	-	27107914	27110640	0	3	1711,646,150,	0,3105,15549,
chr7	27107009	27123027	NM_001384341.1	0	-	27107914	27110640	0	4	1711,646,84,255,	0,3105,15549,15763,
chr7	27107009	27123027	NM_001384342.1	0	-	27107914	27110640	0	4	1711,646,84,158,	0,3105,15549,15860,
chr7	27107009	27140239	NM_001384339.1	0	-	27107914	27110640	0	6	1711,646,84,116,185,157,	0,3105,15549,17424,19876,33073,
chr7	27107009	27140239	NM_001384337.1	0	-	27107914	27110640	0	6	1711,646,84,288,185,157,	0,3105,15549,17177,19876,33073,
chr7	27107009	27140239	NM_001384340.1	0	-	27107914	27110640	0	5	1711,646,84,185,157,	0,3105,15549,19876,33073,
chr7	27107009	27140239	NM_001384338.1	0	-	27107914	27110640	0	4	1711,646,84,157,	0,3105,15549,33073,
chr7	27107009	27150950	NM_001384336.1	0	-	27107914	27110640	0	5	1711,646,84,104,643,	0,3105,15549,33073,43298,
chr7	27107009	27152583	NM_001384335.1	0	-	27107914	27110640	0	7	1711,646,84,116,185,104,296,	0,3105,15549,17424,19876,33073,45278,
chr7	27107009	27152583	NM_153631.3	0	-	27107914	27110640	0	6	1711,646,84,185,104,296,	0,3105,15549,19876,33073,45278,
chr7	27121918	27128760	NR_122069.1	0	+	27128760	27128760	0	4	108,200,179,561,	0,548,1541,6281,
chr7	27128524	27130757	NM_0012141.5	0	-	27129224	27130733	0	2	1047,640,	0,1593,

Uploaded back to UCSC Genome Browser ("User Supplied Track")



What is the difference between an NM and NR RefSeq transcript?

NM RefSeq transcripts represent coding mRNA, whereas NR RefSeq transcripts represent non-coding mRNA¹³

V. *Genome Variation: Analysis Platforms*

Introduction

Variations in the genome include single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) among others. Some are benign, while others may result in phenotypes. In the case of cystic fibrosis, polymorphisms in the CFTR gene result in nearly 70% of cases¹⁴. Copy number variations were first described in 2004 under the name Copy number polymorphisms (CNPs), and they refer to kilobase scale genomic insertions or deletions in the human genome relative to a reference¹⁵.

Objectives

- Find and categorize all SNPs associated with a gene.
- Retrieve and view SNPs from BioMart and Galaxy with IGV
- Navigate the NCBI 1000 Genomes Browser.
- Search ClinVar for clinical data on SNPs.
- Access Flagged SNP tracks in Galaxy.
- Access SIFT, PolyPhen data in Ensembl.
- Define copy number variation (CNV).
- Explain the ramifications and origins of CNV.
- Find CNV tracks on DGV and at UCSC.
- Analyze CNVs using the NCBI Variation Viewer.
- Add and analyze DECIPHER tracks at Ensembl.
- Use IGV to analyze CNV data from Broad Institute Cancer Genomes.

Summary

Single Nucleotide Polymorphisms (SNPs)

- Polymorphisms are mutations present in at least 1% of the population
- Single nucleotide polymorphisms differ by just one nucleotide from reference

Functional Consequences of SNPs

- SNPs in coding sequences that alter amino acid composition are the cause of some genetic disorders
- CDS SNPs which don't change amino acid sequence may alter splicing
- Promoter SNPs may affect gene expression
- Ones with no known impact are often used as chromosomal markers

Medical Implications

- SNPs can cause disease for a variety of reasons
 - Premature stop codons
 - Increased medical risks
 - Differential response to drugs

Categories of SNPs

- Synonymous SNPs – amino acid sequence stays same
- Non-synonymous SNPs – amino acid sequence changes
- Noncoding SNPs – non-coding regions

dbSNP

- Available [here](#)
- Large database of SNPs from NCBI
- Also includes **M**NPs and **s**hort indels
- Galaxy can be used to join genomic intervals and datasets to find overlapping genomic regions

SNPEffect and SNPedia

- [SNPEffect](#) focus is on protein alteration due to SNPs
- [SNPedia](#) collects publications about SNPs and clinical effects

1000 Genomes Exon Pilot Project

- NGS helped discover many new SNPs¹⁶
- Also helped understand how geography correlates with allele frequency

NCBI 1000 Genome Browser

- Available [here](#)
- Browser of 1000 Genomes Project data
- ClinVar/dbSNP tracks can be added

ClinVar

- Available [here](#)
- Acts to supplement dbSNP to help show associations with some SNPs and phenotypes
- Can be searched with gene symbol or by disease
- Results include variation ID, phenotype/condition, allele frequency, clinical significance, review status
- Filters can help narrow down searches to very specific parameters

ClinVar Evidence

- Clicking on record brings one to a page with links for supporting evidence about variation

ClinVar at UCSC Genome/Table Browser

- Available under the “Phenotype and Literature” group
- ClinVar Variants – displays genomic positions of ClinVar records
- ClinGen CNVs – clinical CNV microarray data

Ensembl

- Variant table shows ClinVar data
- Also shows Proven(Sift) and PolyPhen data

Copy Number Variation

CNV Intro

- CNVs may cause the vast differences in individuals of a population
- Individuals who vary in the number of copies of a gene might have different traits

Comprehensive Genomic Analysis

- Large scale studies showed approx 800 CNVs in at least 3% of the population¹⁷
- CNVs may be sensory related – population differences in taste and smell might be the result of CNVs

Mechanism

- Microhomology-Mediated Break Induced Replication (MMBIR)¹⁸
- Based on DNA repair model
- Only considered a possible mechanism – helps to explain hotspots in genome

More studies

- Variations may be associated with the origin of some neurological disorders (autism, mental retardation, etc)¹⁹
- CNV haplotype strategies were developed by Su *et al.* (polyHap)²⁰

Review Paper

- Tang and Amon review helps put into context how CNVs may determine phenotypes²¹

CNV Tools/Databases

Database of Genomic Variants (DGV)

- Available [here](#)
- Hosted by Center for Applied Genomics
- Contains maps to hg19 and hg38 now

Others

- DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources) – built upon patient data (available [here](#))
- NCBI dbVar database sources from DGV and other databases (available [here](#))

Bioconductor Packages

- [cn.mops](#)
- [CNVtools](#)

IGV, Cancer, and CNVs

- TCGA group runs database called Firehose
 - Different cancer types and CNV data are available
 - CNV summary track is useful to show patterns in CNV data associated with cancers

CNV within Variation Viewer

- dbVar ClinVar Large Variations track shows list of CNVs

GnomAD (Genome Aggregation Database)

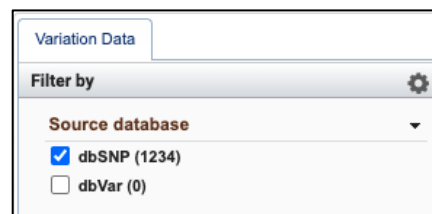
- Available [here](#)
- Browser can be used to search different identifiers
- Peaks visualize protein coding SNPs
- CNV data below that (brown = gain, red = loss)

Supplemental Reading

- Copy Number Variation review²²
- ExAC project – exome sequencing of over 60,000 humans²³
- TOPMed²⁴
- gnomAD extends from the ExAC project²⁵

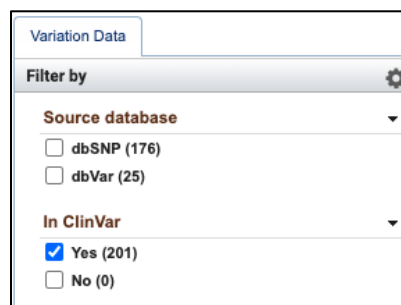
Application: Using NCBI Variation Viewer to explore KCNJ11 variants

- 1. Find the human KCNJ11 gene in the NCBI Variation Viewer. How many variants are in dbSNP? There are 1234 variants in dbSNP. Searching for KCNJ11 on Variation viewer, I filtered the results on the side bar:**



The screenshot shows the 'Variation Data' sidebar in the NCBI Variation Viewer. Under the 'Filter by' section, the 'Source database' dropdown is set to 'dbSNP (1234)'. The 'dbVar (0)' option is also visible but not selected.

- 2. How many are in ClinVar? There are 201 variants in ClinVar. Same process:**



The screenshot shows the 'Variation Data' sidebar in the NCBI Variation Viewer. Under the 'Filter by' section, the 'Source database' dropdown is set to 'dbSNP (176)'. The 'dbVar (25)' option is also visible but not selected. Below this, the 'In ClinVar' dropdown is set to 'Yes (201)'. The 'No (0)' option is also visible but not selected.

3. Filter for copy number variation. How many? There are 64 results.

Variant type

- ☐ single nucleotide variant (0)
- ☒ copy number variation (64)
- ☐ deletion (0)
- ☐ insertion (0)
- ☐ microsatellite (0)

4. Find the pathogenic CNVs. Click each pathogenic accession (nsv) to get to dbVar. Loss or gain? They are all copy number **gain** changes.

▼	nsv4351782	202,758 - 31,726,224	copy number variation	NAV2 and 716 more
Variant calls associated with nsv4351782				
Allele information				
Call ID	Change	Condition	Most severe clinical significance	
nssv15161291	copy number gain	See cases		

VI. Genomic Data Files: Description and Manipulation

Introduction

Genomic technologies output genomic data files, and further manipulation of these data files are a common practice in bioinformatics pipelines. BED files, SAM files, and BAM files are introduced in this section. Their manipulation with freely available tool suites (BEDtools and SAMtools) is also described and implemented at the end.

Objectives

- Use BEDtools in Galaxy and on the command line.
- Describe a SAM and BAM file.
- Use SAMtools in Galaxy/command line to convert SAM files to BAM files.
- Use SAMtools in Galaxy/command line to sort and index BAM files.
- Load BAM files with indexes onto IGV.

Summary

BEDtools on command line

- Suite of tools to perform genomic data manipulations
- Can be installed for command line interface usage on Mac OS X and Linux

BEDtools Functions

- Type “bedtools —help” on the command line interface version to get a list of commands and a short description of their function.

BEDtools Intersect

- Returns overlaps between two feature files
- -wa Write the original entry in A for each overlap.
- -v Only report those entries in A that have _no overlaps_ with B.

BEDtools Subtract

- Removes portions of intervals overlapping another feature

SAM Format

- SAM = “Sequence Alignment/Map” file
- Info about alignment to reference sequence
- Stores info about NGS experiment alignment to reference genome
- Alignment is performed by BWA or HISAT or BOWTIE or BOWTIE2, etc.

CIGAR String

- Variations between reference sequence and aligned read are encoded by a CIGAR string:
 - M – match, I – insertion, D – deletion
- Other CIGAR encodings include skipped bases, etc.

BAM Format

- BAM is equivalent to SAM format, but it is in **binary**
- Compression brings file size down to ~bytes = ~bases in read
- BAM files are usually sorted and then indexed prior to visualization

SAMtools on command line

- SAMtools is available for local download
- Can be used to call SNPs in genomic alignments

SAMtools in Galaxy

- SAM files are large, so space is an issue when using SAMtools in Galaxy
- Current allocation for Galaxy is ~250 GB

SAM-to-BAM

- SAM can be converted to BAM using SAMtools

COMMAND LINE EXAMPLE

samtools view -bS filename.sam > filename.bam

Sorting the BAM File

- As mentioned, BAM files are most useful when sorted

COMMAND LINE EXAMPLE

samtools sort filename.bam filename.sorted

- Output is going to be called *filename.sorted.bam*

Indexing the sorted BAM file

COMMAND LINE EXAMPLE

samtools index filename.sorted.bam

- Output will be called *filename.sorted.bam.bai*

Supplemental Reading

- An overview of BEDtools²⁶
- An overview of the SAM/BAM format and SAMtools²⁷

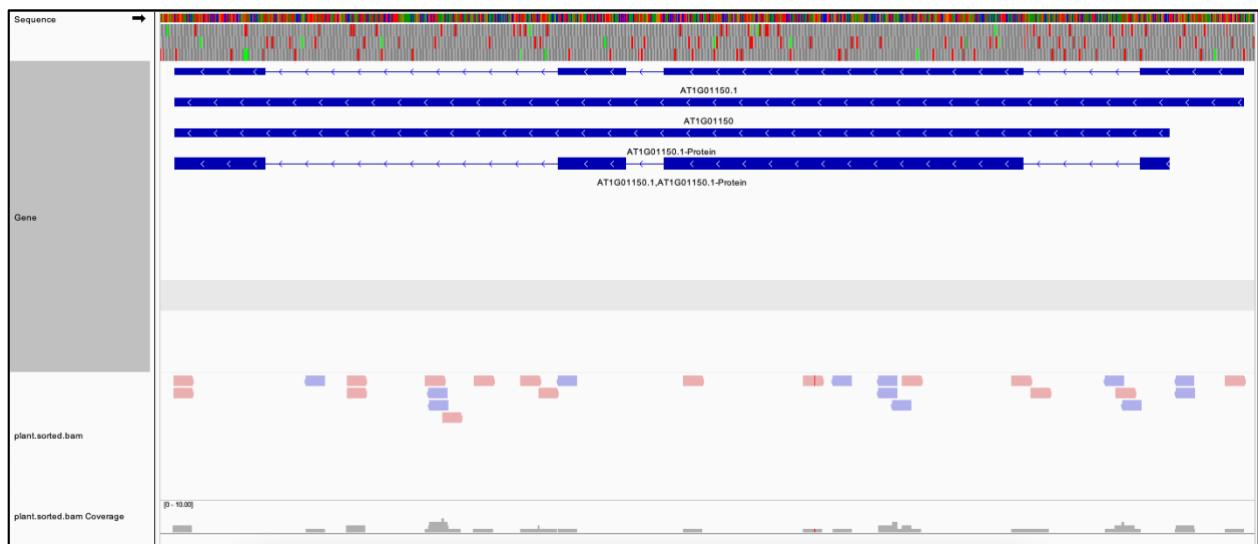
Application: Converting SAM to BAM and uploading to IGV

SAM files are generally quite large. Attached is a truncated file called plant.sam. It is from *Arabidopsis thaliana* and it was aligned to a portion of the TAIR 10 build.

- 1. Use the command line to convert plant.sam to BAM format (as in part 3 of reading), sort the BAM file, and then create an index for the BAM file. List your commands.**

```
ryanyancey@Ryans-Air Final-Portfolio % ls -lh plant*
-rw-r--r--@ 1 ryanyancey staff 19M May 13 22:28 plant.sam
ryanyancey@Ryans-Air Final-Portfolio % samtools view -bS plant.sam > plant.bam
ryanyancey@Ryans-Air Final-Portfolio % ls -lh plant*
-rw-r--r-- 1 ryanyancey staff 3.4M May 13 22:37 plant.bam
-rw-r--r--@ 1 ryanyancey staff 19M May 13 22:28 plant.sam
ryanyancey@Ryans-Air Final-Portfolio % samtools sort plant.bam -o plant.sorted.bam
ryanyancey@Ryans-Air Final-Portfolio % ls -lh plant*
-rw-r--r-- 1 ryanyancey staff 3.4M May 13 22:37 plant.bam
-rw-r--r--@ 1 ryanyancey staff 19M May 13 22:28 plant.sam
-rw-r--r-- 1 ryanyancey staff 3.4M May 13 22:37 plant.sorted.bam
ryanyancey@Ryans-Air Final-Portfolio % samtools index plant.sorted.bam
ryanyancey@Ryans-Air Final-Portfolio % ls -lh plant*
-rw-r--r-- 1 ryanyancey staff 3.4M May 13 22:37 plant.bam
-rw-r--r--@ 1 ryanyancey staff 19M May 13 22:28 plant.sam
-rw-r--r-- 1 ryanyancey staff 3.4M May 13 22:37 plant.sorted.bam
-rw-r--r-- 1 ryanyancey staff 62K May 13 22:37 plant.sorted.bam.bai
ryanyancey@Ryans-Air Final-Portfolio %
```

- 2. Load your BAM files onto IGV using *A. thaliana* (TAIR 10) as your genome. View the AT1G01150.1 locus.**



VII. DNA Elements and the ENCODE Project

Introduction

The human genome is not just a sequence of nucleotides with genes composed of exons and introns. There are non-coding elements within the DNA that serve purposes encompassing regulation and signaling. The ENCyclopedia Of DNA Elements Project began in 2004 and released its first round of findings in 2007. It has since gone on to release multiple versions and become integrated in various genome browsers like the UCSC Genome Browser and IGV.

Objectives

- Access and analyze ENCODE data at UCSC.
- Use Galaxy to download data on a particular chromosomal region.
- Display ENCODE data in IGV.

Summary

Human Genome

CpG Islands

- C followed by G on the *same* strand
- Only 1% frequency (should be ~4.4%)
- CpG islands are *concentrated* CpG clusters
 - Associated with promoters and TSS

Transposons

- Approximately half of human genome
- Consists of LINEs, SINEs, RV-like elements, transposon fossils

Gene Density

- Average GC content is ~41%, but regions of higher GC content exist
- Regions with 50%+ GC are likely gene-rich

Paralogue Clusters

- Closely oriented clusters of duplicated genes
- Tend to result from recent events
- Example: keratin genes (two clusters — type 1 and type 2)

ENCODE Project

ENCyclopedia Of DNA Elements

- Functional elements in human genome
- Selection of 44 elements chosen to represent ~1% of human genome
 - Protein coding genes
 - Non-protein coding genes
 - Transcriptional regulatory elements
 - Sequences that mediate chromosome structure

- Functional sequences yet to be determined
- Pilot data can also be viewed on the UCSC genome browser

Production Phase

- Genome-wide mapping of DNA Elements
- These are available on the UCSC Genome Browser for **hg38**
 - Transcription
 - H3K4me1
 - H3K4me3
 - H3K27Ac
 - DNase HS Clusters
 - DNase HS signals
 - DNase HS in 95 cell types
- These are available for **hg19**:
 - BU ORChID (DNA cleavage sites)
 - Gencode genes (ENCODE coding genes collection)
 - Affy RNA Loc (Subcellular loc of RNA)
 - ENCODE RNA-seq (expression data)
 - ENC Chromatin (chromatin interaction)
 - Stanf Nucleosome (nucleosome positioning)

Mouse ENCODE Project

- Started for comparative genomics with the human ENCODE data
- Est 2009
- Useful for having a reference genome

ENCODE 2012

- Major finding: non-coding **variants** reside in functional regions and many SNPs are enriched there as well

Eukaryotic Features

- Difference between eukaryotic and prokaryotic genomes
 - Larger genome
 - Linear chromosomes
 - Haploid
 - Splicing
 - Centromeres/telomeres

Repeats

- Interspersed Repeats
 - LTR-containing Interspersed Elements (LINEs)
 - Short Interspersed Nuclear Element (SINEs)
- Psuedogenes: nonfunctional genes derived from old genes
- Simple Repeats: repetitions of low-complexity regions
- Segmental Duplications: large duplications
- Tandem repeats: near telomeres, centromeres, ribosomal DNA

Aneuploidy

- Abnormalities in chromosome number
 - Typically a result of nondisjunction in meiosis

Inversions

- Large region of a chromosome invert

Chromosome Number and Synteny

- Roughly similar amount of DNA between species
 - Organization (chromosome size and number) appears proportional
- Fusion/splitting of genes may occur during speciation

Sequence Tagged Sites

- Short regions of unique DNA sequences
- Useful for PCR primer design

dbSTS and ePCR

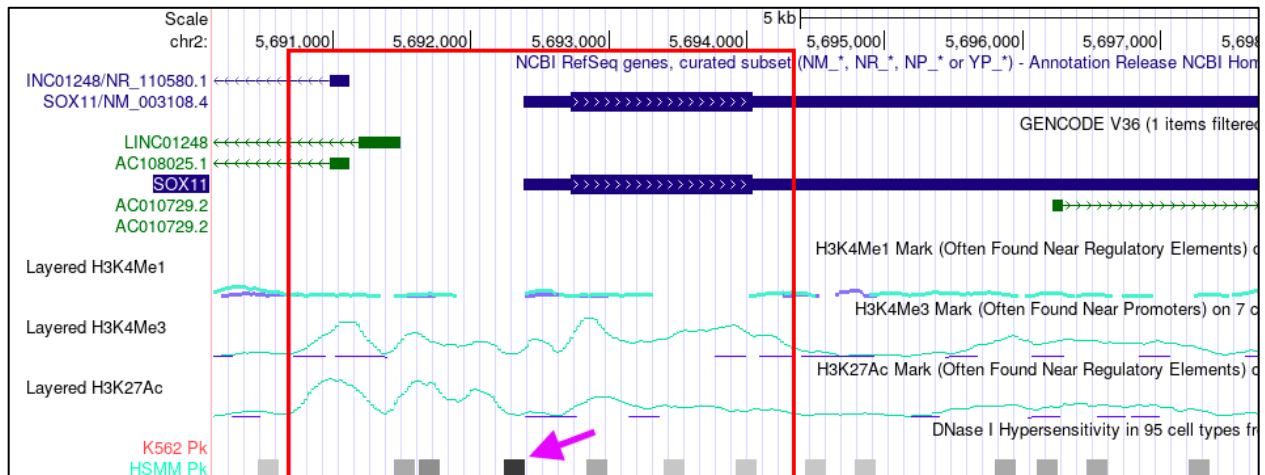
- dbSTS is discontinued but available via FTP
- Electronic PCR useful for finding STS sites for PCR primer design

Supplemental Reading

- A critique of ENCODE²⁸
- A 2020 update to ENCODE²⁹

Application: Exploring ENCODE data at the SOX11 promoter to predict expression patterns

For the human SOX11 gene (Zoom out by 1.5X), use the UCSC Genome Browser (hg38) to examine histone modifications. Eliminate all cell types except HSMM and K562.



- 1. Which cell type is more active and is that activity near the promoter? The HSMM cell line is more active. There appears to be greater enrichment of both H3K27ac marks and H3K4me3 marks near the promoter and very little evidence of the same marks in K562 cells used in this dataset (red boxed area)**
- 2. Look at DNase I hypersensitivity. Does there appear to me a modification near the promoter region of this gene?**
The promoter region of the SOX11 gene in the HSMM cell line appears to have a strong peak for DNase I Hypersensitivity. This is indicative of a nucleosome-free region and supports the evidence that H3K27ac and H3K4me3 marks are present at that cell type's promoter.

VIII. *Non-Coding RNA and Ultra-Conserved Regions*

Introduction

Once thought to be relatively unimportant, the role of non-coding RNAs in gene expression and regulation among other processes has slowly come to light. Functional non-coding RNA annotation and databases are covered in this section. A discussion of ultraconserved regions is also visited.

Objectives

- Analyze SNPs that affect the binding of miRNAs to targets.
- Explore the functional RNA database and miRBase.
- Search UCNEBase.
- Display a CNE in a genome browser.

Summary

Noncoding RNA (ncRNA)

- RNA which stays that way
 - No protein
- mRNA is **not** one
- Examples: tRNA, rRNA, lncRNA, snoRNA, snRNA, miRNA, siRNA, gRNA, etc

Functional RNA (fRNA)

- Broader term than ncRNA, but used synonymously
 - Includes secondary structures in mRNA (technically not ncRNA)
- “Coding fRNA” are fRNAs within exons
- Notable examples: riboswitches, SECIS element
- Warden et al estimates ~5% yeast genome fRNA containing

PPfold

- Secondary structure predictor for RNA
- Phylogenetic stochastic context-free grammars (phylo-SCFGs) used to predict “information entropy” of RNA 2ry structure probability distribution

Predicting miRNA Targets

- miRanda algorithm
- Finds miRNAs and miRNA targets
 - Parses 3'UTRs and uses comparative genomics alongside
- Other miRNA finding algorithms:
 - [PicTar](#)
 - [TargetScan](#)
 - [miRNEST 2.0](#)
 - [PolymiRTS](#)
 - [DIANA-TarBase v.8](#)
 - [miRDB](#)

miRBase

- Available [here](#)
- Database/repository for miRNA known sequences
- Detailed records
- Annotations
- Deep sequencing read integration

Families of ncRNAs

- Rfam database
 - European Bioinformatics Institute
- Pfam, but for ncRNA families
- Multiple sequence alignment of proteins allows for structure/function prediction
 - Same idea for RNA is applied

Ultraconserved Elements, Conserved Noncoding Elements

Describing UCEs

- 200+ bp matches with 100% identity in multiple species
 - With lower cutoff, more findings
 - Not attributed to chance due to the length of identity
- Not many believed to be coding DNA
- Human, mouse, and rat DNA
 - 40% can be aligned
 - Only 5% selectively conserved
 - 3.5% non-coding, 1.5% coding

Finding UCEs

- Sort of arbitrary which species to select for probing
- An approach exists where genomes don't need to be aligned first
- Comparative analysis

SNPs and UCEs

- SNPs are exceedingly uncommon in UCEs (ultra-conserved)
- Little room for variation
- Those mutations which are not lethal can still occur though

Evolutionary Stats

- UCEs are not typically found outside of chordates
- UCEs more constrained than exons
 - It's possible to find exon homologs in other phyla
 - Rare for UCEs
- Are UCEs unique/developed in chordates?
- If not, did they function differently before the divergence of vertebrates/invertebrates?

Intronic Enhancer Example

- UCE function is relatively unknown
- SOX6 contains a UCE (uc322) which increases gene expression in melanocytes

T-UCR

- A new class of fRNAs includes transcribed ultra conserved regions (T-UCRs)
- The original set of UCEs found in 2004 mostly encode for ncRNA
 - Most of these are tumorigenic
- T-UCR deregulation may lead to tumorigenesis in neuroblastoma

lncRNA

- Once thought to be uncommon, now they're believed to be more prevalent
- May play roles in transcriptional regulation, epigenetic, brain function, etc
 - [Review discussing function](#)
 - Role may depend on secondary/tertiary structure

NONCODE

- Includes transcripts from lncrnadb
- 500k+ seqs (16 species)
- Browse and search capabilities

LCNSs

- Long conserved noncoding sequences
 - 500+ bp with more than 95% identity with other species
 - 611 currently between human and mouse
- Study looking at vertebrates: [Sakuraba et al 2008](#)
 - Often clustered in low gene density regions
 - Mutations occur with around the same rate as other genomic regions
 - They are just eliminated by natural selection in these regions

CNE Databases

- Lists many databases:
 1. [ANCORA](#)
 2. [cneViewer](#)
 3. [UCNEBase](#)
 4. [CEGA](#)

Supplemental Reading

- [Harmston et al 2013](#) takes a good look at the focus on CNEs³⁰
- Non-coding RNA and its role in disease³¹
- NGS as a means to predict miRNA candidates

Application: Using multiple tools to analyze an ultraconserved element

Get FASTA sequence of ultraconserved element. The URL used in wget command is [here](#). Python script for extract_uc.py is available [here](#).

```
ryanyancey@Ryans-Air Final-Portfolio % wget https://users.soe.ucsc.edu/~jill/ultra.watson.fa
--2021-05-13 19:19:59-- https://users.soe.ucsc.edu/~jill/ultra.watson.fa
Resolving users.soe.ucsc.edu (users.soe.ucsc.edu)... 128.114.47.74
Connecting to users.soe.ucsc.edu (users.soe.ucsc.edu)[128.114.47.74]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 131192 (128K)
Saving to: 'ultra.watson.fa'

ultra.watson.fa      100%[=====] 128.12K  483KB/s   in 0.3s

2021-05-13 19:19:59 (483 KB/s) - 'ultra.watson.fa' saved [131192/131192]

ryanyancey@Ryans-Air Final-Portfolio % ls -l
total 472
-rw-r--r-- 1 ryanyancey staff 101271 May 13 19:19 Ryancey_FinalPortfolio.docx
-rw-r--r-- 1 ryanyancey staff 299 May 13 19:12 extract_uc.py
-rw-r--r-- 1 ryanyancey staff 131192 May 9 2004 ultra.watson.fa
ryanyancey@Ryans-Air Final-Portfolio % python3 extract_uc.py ultra.watson.fa 75
> uc.75+
AAATTGAAAAATCCCATCTCACAATTAATGTTCCAAAACACAATAAATGCTCTTCTTTACGTAAAATTTGCCCAAATGATCAACGTCATGTTCTTTTTTACTAAAATATATCTATATATTGAAGAAC
TATAATACTGTACACTACAGTATGAAATAAAATAGGAAATATAAAATGAGCCACATAAAATGTTATTTGACCTAAAATTAATGAATGCAAAAAAAA
ryanyancey@Ryans-Air Final-Portfolio % python3 extract_uc.py ultra.watson.fa 75 > uc.75.fa
ryanyancey@Ryans-Air Final-Portfolio % cat uc.75.fa
> uc.75+
AAATTGAAAAATCCCATCTCACAATTAATGTTCCAAAACACAATAAATGCTCTTCTTTACGTAAAATTTGCCCAAATGATCAACGTCATGTTCTTTTTTACTAAAATATATCTATATATTGAAGAAC
TATAATACTGTACACTACAGTATGAAATAAAATAGGAAATATAAAATGAGCCACATAAAATGTTATTTGACCTAAAATTAATGAATGCAAAAAAAA
ryanyancey@Ryans-Air Final-Portfolio %
```

Navigate to [Blat](#) and upload uc.75.fa to run against hg38:

Human BLAT Search

BLAT Search Genome

Genome: ☐ Search all ☐ Human

Assembly:

Query type:

Sort output:

Output type:

☐ All Results (no minimum matches)

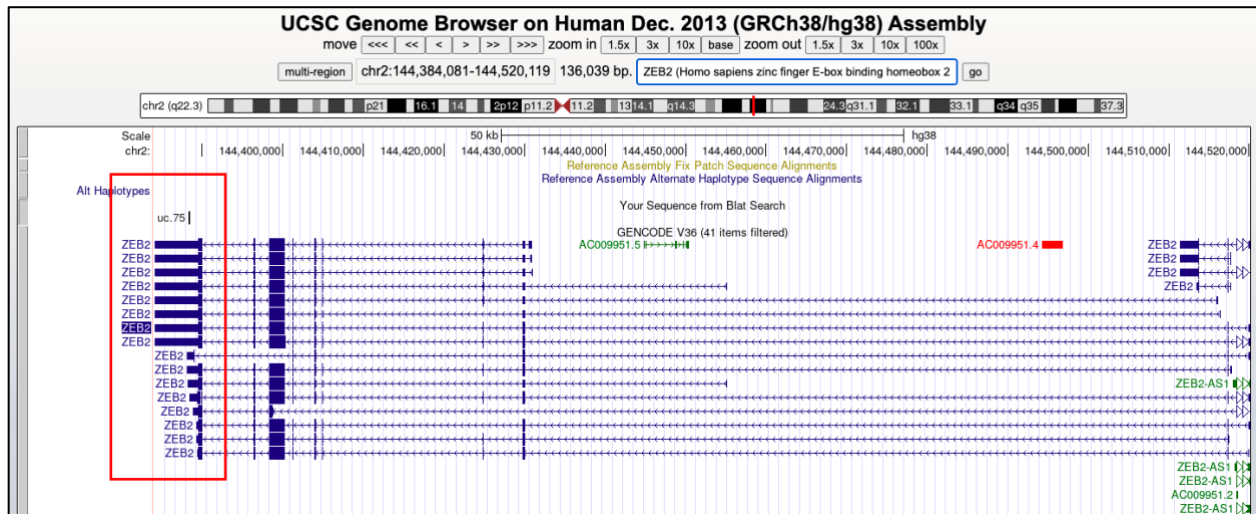
Submit I'm feeling lucky Clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

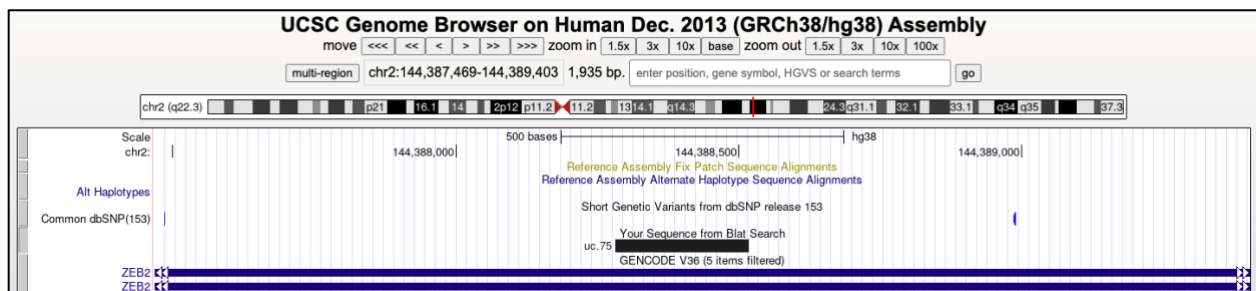
File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: uc.75.fa

Click “Browser” next to the first result. Observe on genome browser and zoom to full gene view.



This ultraconserved element appears in the 5' UTR of ZEB2. Using the dbSNP153 "Common SNPs" track, it's clear that no SNPs appear in this region, as expected.



IX. Next Generation Sequencing

Introduction

Next generation sequencing (NGS) is defined as a high-throughput and accurate means for obtaining large-scale genome sequences at a fraction of the time it would take to sequence via Sanger Sequencing. These massively parallel sequencing technologies revolutionized genomics and are a staple of scientific research. NGS technologies branch into ChIP-seq and RNA-seq. Higher-resolution NGS technologies (scRNA-seq) have become the forefront of genomic research in recent years.

Objectives

- Use Galaxy to perform NGS analyses.
- Assess NGS data quality.
- Trim NGS data based on quality scores.
- Create a VCF file, from FASTQ files through the assembly process.

Summary

Sequencing technologies

- NGS implies a first generation
 - This was Sanger Sequencing (chain termination sequencing)
 - Low-throughput, intensive
- **Next Generation (NGS)**
 - **Roche/454 Life Sciences (No longer in use)**
 - **Illumina**
 - **ABI SOLiD**
 - **Ion Torrent**
 - **Pacific Biosciences (PacBio)**
- Third Generation
 - Oxford Nanopore

Phred quality scores

- Quality scores originally used to resolve overlap discrepancies
- They remain highly important
- Quality scores imply a confidence in the specific nucleotide

FASTQ format

- FASTQ uses single characters to represent nucleotides and quality scores
- Phred can go into double digits – thus ASCII codes need to be used to represent them
- FASTQ similar to FASTA, only differences:
 - FASTQ starts with @ instead of >
 - FASTQ contains quality score lines (beings with +)
- FASTQ lines are made up of 4 lines for each row of sequence:
 - Header line (@)

- DNA seq line
 - Header line (+)
 - Quality score line
- .fq or .fastq file extension

Sequence Read Archive (SRA)

- NCBI, EMBL, DDBJ international project
 - Stores unprocessed, raw read files
- Because of how quickly sequencing is becoming accessible, storage concerns threaten the future of the SRA (though they now contain cloud instances)
- SRA toolkit freely available for accessing raw read data
 - Available [here](#)

Metagenomics

- Study of environmental genetic material
- High-throughput way to identify DNA/RNA
- Specifies which organisms are present in a sample
 1. Real-Time Metagenomics ([website](#))
 2. Rapid Annotation using Subsystem Technology (RAST) ([website](#))
 3. Pipeline Analysis for Next Generation Amplicons (PANGEA) ([download](#))
 4. Metagenomic analysis tools in Galaxy ([tutorial](#))
 5. Metagenome Analyzer (MEGAN) - ([download](#))

Genome Sequence Assembly

The challenge

- How does one take seq data and converge to a full sequence of DNA?

De novo assembly vs Reference-based assembly

- *De novo* assembly is used to align fragments from scratch³².
 - Used when a genome is sequenced for the first time
- Reference-based assembly uses a previously agreed upon sequence as a guide
 - Allows for quicker genome assembly
 - Closely related species genomes can be used with some success

Single-end vs Paired-end

- NGS involves shotgun sequencing of small fragments
 - Single-end has sequences from only one end
 - Paired-end has sequences from both ends
- Paired end provides another layer of info: sequence length
- Need to know which method is used in your seq experiments

Organism

- Assemblers can be better for certain genomes (smaller ones) and some are better for larger ones.

- SPAdes, Velvet (smaller)
- ABySS (larger)

Sequencing Platform

- Platforms may have different error/read lengths
 - Longer read length makes it easier to assemble large genomes
 - But they typically come with a cost of precision
 - Higher error rate

De novo assemblers: Velvet

- Velvet utilizes a de Bruijn graph approach
 - Good for smaller prokaryotic genomes
- Available on BFX server

Reference-based assembly tools

- Mappers use reference genome to align small fragments
- Bowtie/Bowtie2
- BWA
- HISAT/HISAT2

Variant Calling Background and Workflow

Mismatches

- Bases will sometimes not align in sequences
- There are a few reasons: Sequencing error, Heterozygous SNP, Homozygous SNP, PCR amplification errors
- Sequencing errors can be easily identified with deeper sequencing libraries
- Heterozygous SNPs will appear on ~50% of reads
- Homozygous SNPs will appear on nearly 100% of reads (different from reference)
- Aberrations in these patterns inform whether SNPs are real or not

Variant calling tools

1. FreeBayes
2. SAMtools (mpileup)
3. GATK Unified Genotyper
4. GATK Haplotype Caller
5. Platypus

Variant calling

- BAM file input, VCF file output
 - VCF (variant call format)
- Variants are lines in VCF file
 - Chromosomal position
 - Expected allele
 - Sequenced allele

- many other statistics

Suggested workflow

1. FASTQ Groomer (get files into Sanger/Illumina 1.9 format if not already)
2. FASTQ Trimmer / Trimmomatic
3. Bowtie2/BWA/HISAT
4. SAM-to-BAM
5. Filter SAM or BAM
6. Sort BAM
7. FreeBayes (variant calling step) – PRODUCES VCF file
8. SnpEff or ANNOVAR to predict variant-function interaction

Supplemental Reading

- *De novo* genome assembly³²
- An assessment of variant callers³³

Application: Identifying SNPs in NGS data from the 1000 Genomes Project using Galaxy

Forward reads:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_1.filt.fastq.gz

Reverse reads:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_2.filt.fastq.gz

PIPELINE

- **Determine quality encoding with FASTQC, trim low-quality bases with Trimmomatic, and compare.**

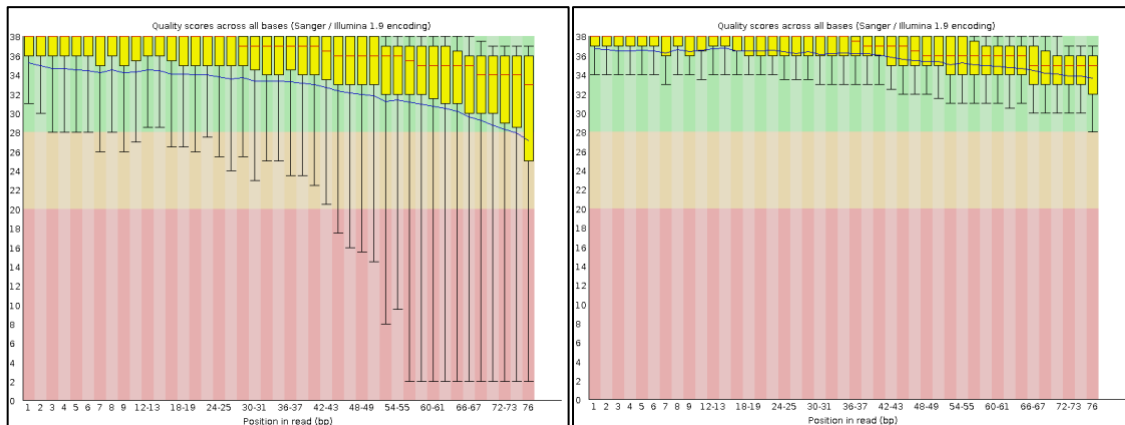


Figure 1 - FASTQC quality scores for R1 before (left) and after (right) Trimmomatic trimming.

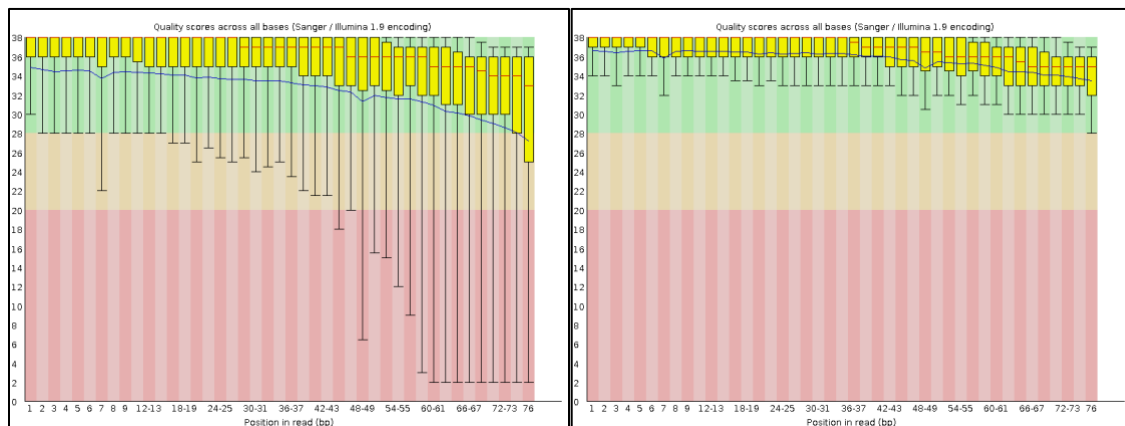


Figure 2 - FASTQC quality scores for R2 before (left) and after (right) Trimmomatic trimming.

- **Align reads to reference genome using HISAT2**
- **Identify variants with FreeBayes**
- **Filter and annotate variants with VCFfilter (AF=0.5, DP>10)**
- **Use UCSC Main to upload RefSeq genes in BED format, then VCFannotate to intersect the BED annotations with the VCF file.**

- Use command line tools to find the number of total results versus the number of intersecting results.

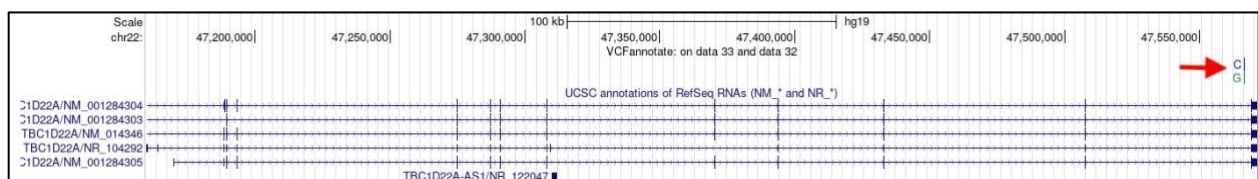
There are 30 variants found with allele frequencies at 0.5 (AF = 0.5) and read depths greater than 10 (DP > 10). Of these, 11 were annotated with a RefSeq gene (found by searching for “BED-features” in text editor – made sure to omit the result which corresponded to header text).

```
$ grep -v '#' annotated_snps.vcf | wc -l
30
```

```
$ grep -i 'bed-features=' annotated_snps.vcf | wc -l
11
```

Choose any SNP in the filtered, annotated VCF file that overlaps a gene. View that position in any genome browser. What is the nucleotide change and the gene that is affected? In which part of the gene is the SNP located? What effect might the SNP have on the gene function, if any?

```
$ grep -oiE 'bed-features=.*?;' annotated_snps.vcf
BED-features=NR_136571.1;
BED-features=NR_136571.1;
BED-features=NR_136571.1;
BED-features=NR_136571.1;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001291030.2:NM_001013647.2;
BED-features=NМ_001284303.1:NM_001284304.1:NM_014346.5:NR_104292.1:NM_001284305.1;
```



Chromosome: 22

Position: 47566394 (hg19 in UCSC)

SNP: C>G (g.C47566394G)

Gene affected: TBC1D22A

This SNP is located in the final intron of the TBC1D22A gene. Although this gene is not located in the coding region, its proximity to the end of the intron may result in mRNA splicing failure – though this is unlikely. Galaxy Workflow was extracted and available at this link [here](#). The downloaded ‘.ga’ file is available on my GitHub [here](#).

X. *Chromatin Immunoprecipitation Sequencing Analysis (ChIP-seq)*

Introduction

Chromatin Immunoprecipitation followed by sequencing by NGS (ChIP-seq) is a genomic technology which allows researchers to assess the epigenetic features associated with histone modifications and chromatin binding proteins. Advances in these technologies have improved the general understanding of the epigenome and its role in disease.

Objectives

- Analyze ChIP-seq data.
- Describe peak callers.
- Run MACS2 to create bedgraph files.
- Display WIG files and bedgraph files in IGV.

Summary

ChIP-seq as a Technology

Chromatin ImmunoPrecipitation (ChIP)

- IP: method to isolate protein of interest with antibodies to the protein
- chromatin IP refers to antibodies targeting histones
 - antibodies can be as specific as to target the modification, as well
- the procedure is used to isolate cross-linked chromatin-DNA complexes
- These can then be subject to ChIP-qPCR for specific sites, or ChIP-seq for all sites

ChIP-seq

- Chromatin immunoprecipitation followed by next-gen sequencing of bound DNA
- ChIP is performed as normal
- DNA is sequenced and aligned
- Regions with high “pileup” are believed to be associated with the assayed chromatin mark
- Negative controls are used to subtract noise from background binding to make final results clearer

ChIP-seq Data Analysis

File formats

- ChIP-seq raw data: FASTQ format
- Many deposited to SRA or GEO
 - fastq-dump can be used to get FASTQ files
- FASTQ from SRA can be uploaded by EBI SRA tool
 - Get Data > Upload File > Paste/Fetch data > paste ftp link

Alignment

- After grooming/trimming reads, aligned to reference genome
- Bowtie2/BWA can be used among others
- Once aligned, then proceed to peak calling

Peak Calling

- Important part of experiment
- IDs peaks where pileups occur
- Associates them with the DNA the chromatin peaks align to
- Input: BAM file
- Output: BED or WIG file
- MACS2 and Sicer are common programs
 - Though there are [many others](#)

ChIP-seq controls and determining noise from signal

- What is a good type of control?
- Good one can determine noise in experiment from low, meaningful signal
- Replicate experiments can be helpful
- Control experiments (DNA input, IgG antibody, or untagged strains) are great for improving detection

Differential Peak Calling

- Finding out what binding patterns change due to conditions or different tissues/cell types
- Differential Peak Calling tools are available
- Some require that replicates are present to work

Supplemental Reading

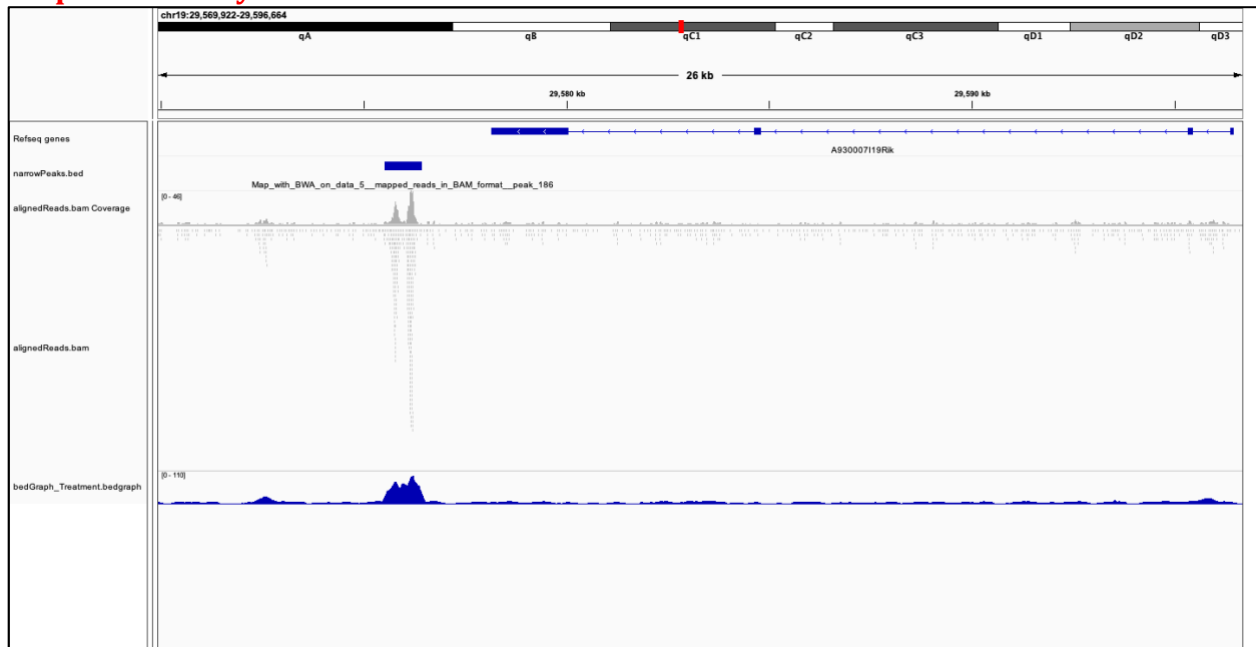
- Histone variants play important roles in neuronal development³⁴
- Introduction of the MACS ChIP-seq peak calling algorithm³⁵
- An analysis of over 100 human epigenomes³⁶

Application: Using Galaxy to call peaks with MACS2 on mouse ChIP-seq data

File: Mouse_ChIP-Seq_Example_Experimental_Data_chr19_mm9.fastq.gz (available on GitHub [here](#))

1. Run FASTQC to determine the quality score encoding
2. Run FASTQ Groomer to convert the file to Sanger/Illumina 1.9 phred encoding ONLY IF NEEDED
3. Run Trimmomatic and set the minimum phred score in a 4 nt sliding window to 25
4. Re-run FASTQC to check the quality scores and encoding scheme
5. Run Map with BWA with default settings (make sure to select for single-end reads), aligning against the mouse genome version mm9
6. Run MACS2 callpeak on the BAM file, setting the Effective genome size to the mouse genome. Use default settings for the rest of the parameters and leave the control field blank.

- a. Load the MACS2 Bedgraph Treatment file and narrow Peaks BED file from step 6 and aligned BAM file from step 5 to IGV or UCSC. Find a gene locus that has ChIP peaks nearby.



- b. MACS2 produces a Bedgraph file, not a WIG file. How do those two file types differ? Can Bedgraph files be converted to WIG format, and vice versa?

Both file formats are used to graphically display continuous data on browser tracks. Bedgraph files contain information about the original data, whereas Wiggle (WIG) formatted files compress the data into bins and store chromosomal start coordinates, along with values. Bedgraph files can be converted to WIG files, but the inverse is not possible.

- c. MACS2 has the option of generating 'broad peaks'. What type of ChIP-seq data should be analyzed for 'broad peaks' instead of 'narrow peaks'? Why?

Broad peaks are generally used for determining histone modification locations in a genome. This is because the DNA wraps around histones and thus the peaks encompass a larger (broader) region. Narrow peaks are useful for identifying locations of transcription factor or chromatin modifying factor binding sites in the genome.

- d. MACS2 has the option to remove duplicate reads before peak-calling. What are duplicate reads and why would one choose to remove them?

Duplicate reads are reads which are literally identical to each other (i.e. they would stack up perfectly in a genome browser). They are typically fragments of PCR amplification for library preparation in NGS workflows. One chooses to remove these because true, biological binding sites are more likely to be represented in ChIP-seq data by *overlapping* reads, not *duplicate* reads. For a great discussion, see [this Biostars post](#).

XI. RNA Sequencing Analysis (RNA-seq)

Introduction

RNA sequencing (RNA-seq) is a technique of NGS which sequences cDNA on a massively parallel scale. As mentioned at the beginning of this portfolio, expression-level annotation is the most accurate means of annotating eukaryotic genomes. In fact, RNA-seq has uncovered non-coding RNA transcription start sites, alternative splice products, among many other non-coding elements that are part of an organisms transcriptome.

Objectives

- Be able to describe how RNA-seq works.
- Be able to list the goals of an RNA-seq experiment.
- Discuss emerging issues involving RNA-seq.
- Run the TopHat/Cufflinks package in Galaxy and be able to explain each step.
- Interpret results from Cuffdiff.
- Visualize aligned RNA-seq data as BAM files.

Summary

RNA-seq as a Technology

RNA-seq Goals

- RNA-seq: massively parallel cDNA sequencing
- Can define TSS, reveal new ncRNAs/lncRNAs, annotate unknown splice products/sites, quantify transcript levels
- Can quantify *absolute* gene expression

Reference-based RNA-seq analysis

- Study transcriptomes
- Alignment to genome allows one to find out where original RNA was transcribed
- Many tools available to process and analyze raw RNA-seq data

Alignment issues

- RNA spans exon-exon boundaries – alignment algorithms need to be aware
- Two paired end reads may *appear* very distant in genome, but they are not because the introns between them are large
- Software specialized for transcript assembly are required for transcript assembly (TopHat)

Cufflinks, Cuffcompare, Cuffdiff

- TopHat aligns sequenced reads
- Cufflinks assembles them and puts out complete transcripts
 - Input: accepted_hits.bam, GTF annotation file (gene names to locations)
 - Output: New GTF file with gene locations/names
 - Output: 2 tables – gene expression, transcript expression levels
- Cuffcompare compares 2+ GTF files output from Cufflinks

- Differential expression from two cell types, developmental stages, etc
- Cuffdiff performs statistical analysis of DE between the two

TopHat and Cufflinks in Galaxy

- Both available through Galaxy (though TopHat depreciated)

Other analysis tools

- MATS – statistical analysis method (hypothesis testing of alt-splicing events)
- DiffSplice
- DESeq2 – popular R/Bioconductor tool
 - Differential expression testing from RNA-seq data
- edgeR – another popular R/Bioconductor tool
- All these tools rely on reference genome alignment before abundance estimation

StringTie

- Alternative to Cufflinks
- Assembly/expression estimates in one step

De novo (non-reference-based) RNA-seq analysis

- Transcript assembly/quantification can be performed in the absence of a good reference genome
- Overlapping segments used to replicate full RNA molecule
- Trinity is a popular *de novo* assembler
- Others include Salmon, Sailfish, RSEM, kallisto

RNA-seq analysis output

- Abundances of known/novel transcripts
- New transcripts continue to be discovered
- More in-depth analysis can be performed with RNA-seq over conventional microarrays

Supplemental Reading

- Review on RNA-seq best practices³⁷
- An in-depth description of TopHat/Cufflinks³⁸
- An in-depth look at Stringtie for *de novo* transcript assembly³⁹
- DESeq2 analysis of RNA-seq data⁴⁰

Application: Using Galaxy to assess differential expression of RNA-seq data from two mouse cell lines

Follow this Galaxy [RNA-seq tutorial](#) on Galaxy Main (usegalaxy.org), including the Vizualization section. It will take some time to run various steps, so don't wait until the last minute!

When you upload files, use the following links to avoid errors.

file type: fastqsanger, genome:mm10

https://zenodo.org/record/583140/files/G1E_rep1_forward_read_%28SRR549355_1%29

https://zenodo.org/record/583140/files/G1E_rep1_reverse_read_%28SRR549355_2%29

https://zenodo.org/record/583140/files/G1E_rep2_forward_read_%28SRR549356_1%29

https://zenodo.org/record/583140/files/G1E_rep2_reverse_read_%28SRR549356_2%29

https://zenodo.org/record/583140/files/Megakaryocyte_rep1_forward_read_%28SRR549357_1%29

https://zenodo.org/record/583140/files/Megakaryocyte_rep1_reverse_read_%28SRR549357_2%29

https://zenodo.org/record/583140/files/Megakaryocyte_rep2_forward_read_%28SRR549358_1%29

https://zenodo.org/record/583140/files/Megakaryocyte_rep2_reverse_read_%28SRR549358_2%29

file type: gtf, genome:mm10

https://zenodo.org/record/583140/files/RefSeq_reference_GTF_%28DSv2%29

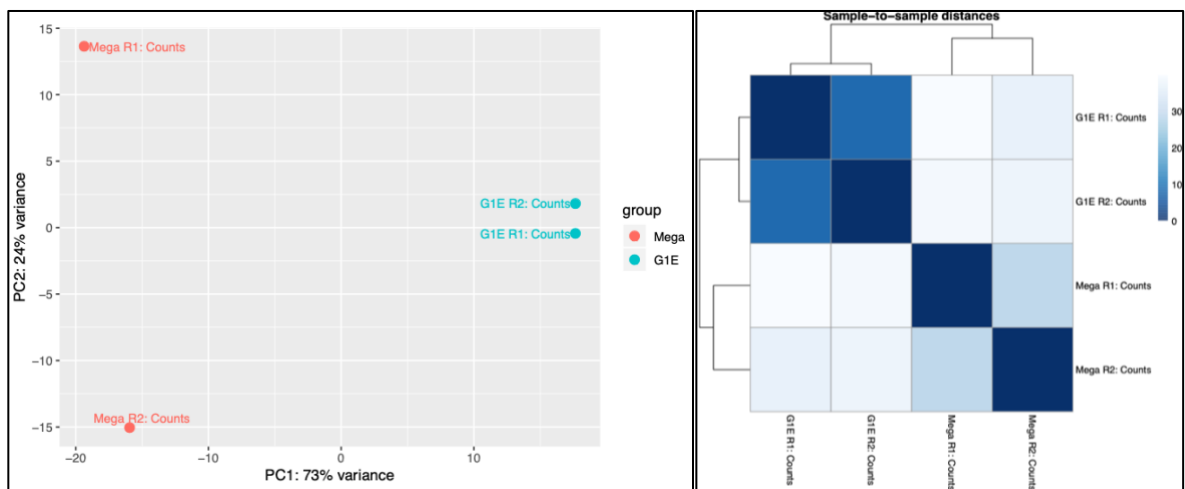
-
- Scale chr11: 96,195,000 | 5 kb | 96,200,000 | mm10 | 96,205,000
- mm10 RefSeq GTF
- G1E R1 StringTie
- G1E R2 StringTie
- PLUS: G1E R1 bigwig
- PLUS: G1E R2 bigwig
- MINUS: G1E R1 bigwig
- MINUS: G1E R2 bigwig
- MINUS: Mega R1 bigwig
- MINUS: Mega R2 bigwig
- PLUS: Mega R1 bigwig
- PLUS: Mega R2 bigwig

48

- b. How well do the G1E and Megakaryocyte RNA-seq replicates agree? What is your evidence? Submit and describe two figures that support your conclusion.**

The two G1E replicates are in agreement with each other nearly perfectly, whereas the Megakaryocyte replicates are a bit different. This is supported by two plots in the DESeq2 output. In the PCA plot below, the two groups (Mega & G1E) are separated along the first principal component (x-axis), which is a good thing. This indicates that the two groups are dissimilar along the greatest source of variance. Unfortunately, on the second greatest source, only the G1E group is clustered together. Megakaryocyte replicates separate pretty greatly along this component. This could be related to batch effects or sample contamination.

Next to the PCA plot, there is a heatmap displaying the “sample-to-sample distance” between all permutations of replicate couples. The darker blue a square is, the smaller the “distance” between the two samples is. The middle line (0-distance) represents 1-to-1 comparisons between the same sample. It’s clear that the G1E R1 and R2 counts are similarly distant – as indicated by the darker blue colors in the upper-left quadrant of the heatmap. In contrast, Megakaryocyte replicates are somewhat distant from each other, as indicated by the light blue hues.



- c. How many transcripts have a significant (adjusted p-value < 0.01) change in expression between these conditions? How many transcripts are up-regulated in G1E? How many transcripts are down-regulated in G1E?**

Using the Filter tool on Galaxy, filtering by “ $c7 < 0.01$ ” yields 51 transcripts. Filtering these 51 by $c3 > 0$ and $c3 < 0$ for upregulated and downregulated transcripts, respectively, there are 30 upregulated and 21 downregulated transcripts.

- d. Choose a transcript that is differentially expressed from part c and has a log2 fold change of at least 2 or -2. What is the transcript? What is the biological function of the gene corresponding to this transcript? In which cell type is this transcript more highly expressed, and by how much? Make a conjecture about how the difference in expression of this gene might explain or be a result of the cell types examined.

I examined the transcript: NM_023785, which is downregulated nearly 10-fold in the G1E state. This transcript corresponds to *M. musculus* pro-platelet basic protein (Ppbp), a cytokine secreted by activate platelets which promotes synthesis of extracellular matrix, mitogenesis, among other processes⁴¹.

According to the mouse transcriptome ENCODE data, this transcript is most highly expressed in mice spleen and mouse embryonic livers at a rate of 131.5 and 84.5 RPKM (reads per kilobase per million reads), respectively ^{41,42}. This is sensible, as multiple sources denote that hematopoietic stem cells (liver stem cells) are progenitors to megakaryocyte cells ^{43,44}.

The differential expression pattern found in this experiment makes sense, as well. G1E cells are embryonic stem cells which lack the GATA1 transcription factor. While it may be possible that GATA1 promotes the expression of Ppbp in mice, it is much more likely that the difference in expression observed is due to the differentiation state of the two cells. Megakaryocytes are highly differentiated and highly specific in their function. They are the source of platelets in the human body, so it is sensible that Ppbp is expressed in these cells. However, G1E cells are highly plastic (non-differentiated), so their role as a cell is not clearly defined. Thus, tissue-specific protein-coding genes are not likely to express in these cells.

XII. References

1. Sallet, E., Gouzy, J. & Schiex, T. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics* **30**, 2659–2661 (2014).
2. Kumar, A. & Bansal, M. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression. *DNA Res* dsw045 (2016) doi:10.1093/dnares/dsw045.
3. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research* **47**, e110–e110 (2019).
4. Dao, F.-Y., Lv, H., Wang, F. & Ding, H. Recent Advances on the Machine Learning Methods in Identifying DNA Replication Origins in Eukaryotic Genomics. *Front. Genet.* **9**, 613 (2018).
5. Mahood, E. H., Kruse, L. H. & Moghe, G. D. Machine learning: A powerful tool for gene function prediction in plants. *Appl Plant Sci* **8**, (2020).
6. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in Comparative Genomics: Genome Correspondence, Gene Identification and Regulatory Motif Discovery. *Journal of Computational Biology* **11**, 319–355 (2004).
7. Darwish, O., Shahan, R., Liu, Z., Slovin, J. P. & Alkharouf, N. W. Re-annotation of the woodland strawberry (*Fragaria vesca*) genome. *BMC Genomics* **16**, 29 (2015).
8. Wen, G., Pachner, L. I., Gessner, D. K., Eder, K. & Ringseis, R. Sterol regulatory element-binding proteins are regulators of the sodium/iodide symporter in mammary epithelial cells. *J Dairy Sci* **99**, 9211–9226 (2016).
9. PubChem. SLC5A5 - solute carrier family 5 member 5 (human). <https://pubchem.ncbi.nlm.nih.gov/gene/SLC5A5/human>.
10. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res* **44**, D710–D716 (2016).
11. The biomaRt users guide. <https://www.bioconductor.org/packages/release/bioc/vignettes/biomaRt/inst/doc/biomaRt.html>.
12. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research* **49**, D1046–D1057 (2021).
13. Pruitt, K., Murphy, T., Brown, G. & Murphy, M. *RefSeq Frequently Asked Questions (FAQ)*. *RefSeq Help [Internet]* (National Center for Biotechnology Information (US), 2020).
14. Kerem, B. S. *et al.* DNA marker haplotype association with pancreatic sufficiency in cystic fibrosis. *Am J Hum Genet* **44**, 827–834 (1989).

15. Sebat, J. Large-Scale Copy Number Polymorphism in the Human Genome. *Science* **305**, 525–528 (2004).
16. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
17. Wong, K. K. *et al.* A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *The American Journal of Human Genetics* **80**, 91–104 (2007).
18. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Reviews Genetics* **10**, 551–564 (2009).
19. Bruder, C. E. G. *et al.* Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *The American Journal of Human Genetics* **82**, 763–771 (2008).
20. Su, S.-Y. *et al.* Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics* **26**, 1437–1445 (2010).
21. Tang, Y.-C. & Amon, A. Gene Copy-Number Alterations: A Cost-Benefit Analysis. *Cell* **152**, 394–405 (2013).
22. Chen, L., Zhou, W., Zhang, L. & Zhang, F. Genome Architecture and Its Roles in Human Copy Number Variation. *Genomics Inform* **12**, 136 (2014).
23. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
24. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).
25. Genome Aggregation Database Consortium *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
26. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
28. Graur, D. *et al.* On the Immortality of Television Sets: ‘Function’ in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution* **5**, 578–590 (2013).

29. The ENCODE Project Consortium *et al.* Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
30. Harmston, N., Barešić, A. & Lenhard, B. The mystery of extreme non-coding conservation. *Phil. Trans. R. Soc. B* **368**, 20130021 (2013).
31. Lekka, E. & Hall, J. Noncoding RNAs in disease. *FEBS Letters* **592**, 2884–2900 (2018).
32. Henson, J., Tischler, G. & Ning, Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **13**, 901–915 (2012).
33. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
34. Santoro, S. W. & Dulac, C. Histone variants and cellular plasticity. *Trends in Genetics* **31**, 516–527 (2015).
35. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
36. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
37. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).
38. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
39. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
40. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
41. PPBP pro-platelet basic protein [Homo sapiens (human)] - Gene - NCBI. <https://www.ncbi.nlm.nih.gov/gene/5473>.
42. Ppbp pro-platelet basic protein [Mus musculus (house mouse)] - Gene - NCBI. <https://www.ncbi.nlm.nih.gov/gene/57349>.
43. Woolthuis, C. M. & Park, C. Y. Hematopoietic stem/progenitor cell commitment to the megakaryocyte lineage. *Blood* **127**, 1242–1248 (2016).
44. Deutsch, V. R. & Tomer, A. Megakaryocyte development and platelet production. *British Journal of Haematology* **134**, 453–466 (2006).