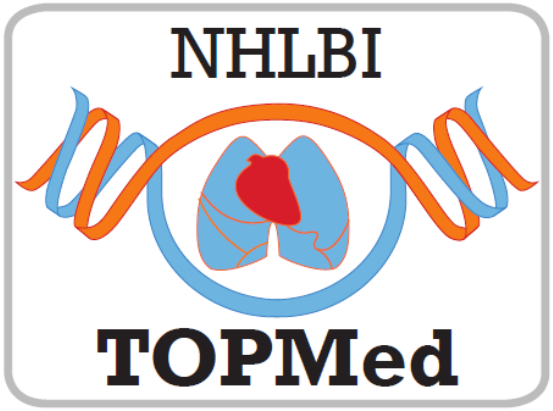


Variant Annotation for TOPMed

Deepti Jain

TOPMed Analysis Workshop

August 8, 2017

Overview of variant annotation session

- Part I – Introduction to variant annotations for TOPMed
- Part II – Defining and filtering aggregation units using variant annotations
- Part III – Hands on exercise for generating variants list used for aggregation tests
- Part IV – Hands on exercise for conducting association tests using aggregate units

Part I
Introduction to variant annotations for TOPMed

What is variant annotation?

- Annotation: a note of explanation or comment

- is the variant
  - Within a gene or intergenic
  - Within which transcript
  - Conserved
  - Deleterious
  - …

How are annotations used?

1. Rare variant association tests
   - To define and filter aggregation units
   - Use as weights

2. Fine map novel and previously known significantly associated loci to identify likely causal variants

What is the annotation source?

- Lots of resources!

Annotation can be generated by any lab or consortium

- NCBI
- Ensemble
- UCSC
- ENCODE
- Roadmap Epigenomics Consortium
- FANTOM5
- …

WGSA

J Med Genet. 2016 February ; 53(2): 111–112. doi:10.1136/jmedgenet-2015-103423.

WGSA: an annotation pipeline for human genome sequencing studies

Xiaoming Liu[1,2], **Simon White**[3], **Bo Peng**[4], **Andrew D. Johnson**[5,6], **Jennifer A. Brody**[7], **Alexander H. Li**[1], **Zhuoyi Huang**[3], **Andrew Carroll**[8], **Peng Wei**[1,9], **Richard Gibbs**[3], **Robert J. Klein**[10], and **Eric Boerwinkle**[1,2,3]

https://sites.google.com/site/jpopgen/wgsa/

Annotation for TOPMed variants

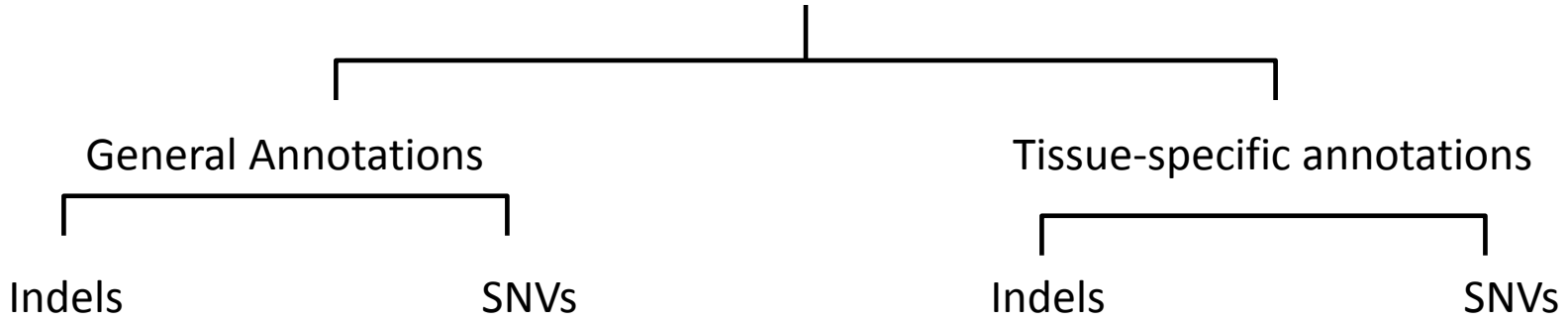
- Generated by Xiaoming Liu
- Variants
  - Includes 236,191,939 variants combined over TOPMed freeze 2, 3 and 4 (SNVs and Indels)
  - Includes variants flagged as failed by IRC pipeline
- Annotations
  - Used WGSA v065
  - 1,502 annotations
- Typically an annotation set is generated for every IRC freeze
- All WGSA annotation releases are made available in the exchange area

1,502 annotations for TOPMed

- Gene based location and consequence
  - Softwares : SnpEff, ANNOVAR, VEP
  - Gene models: Ensembl ,RefSeq ,UCSC
- Transcript-specific annotation
- Loss-of-function annotations (eg: LOFTEE)
- Deleteriousness predictions(CADD, MetaSVM, ssSNV etc)
- Allele frequencies (1000G, UK10K, EXAC etc)
- Regulatory annotations (ENCODE, Roadmap, FANTOM5)
- Conservation scores
- Mappability scores
- rsIDs
- Many more ….

It is guaranteed that you will not use all of the annotations for an analysis.
We recommend starting with a subset of frequently used annotations

TOPMed WGSA annotation directory structure



- 23 .gz file (one per chromosome)
- 1 .txt Data dictionary
- 1 .tsv file with first 1000 lines of chr1

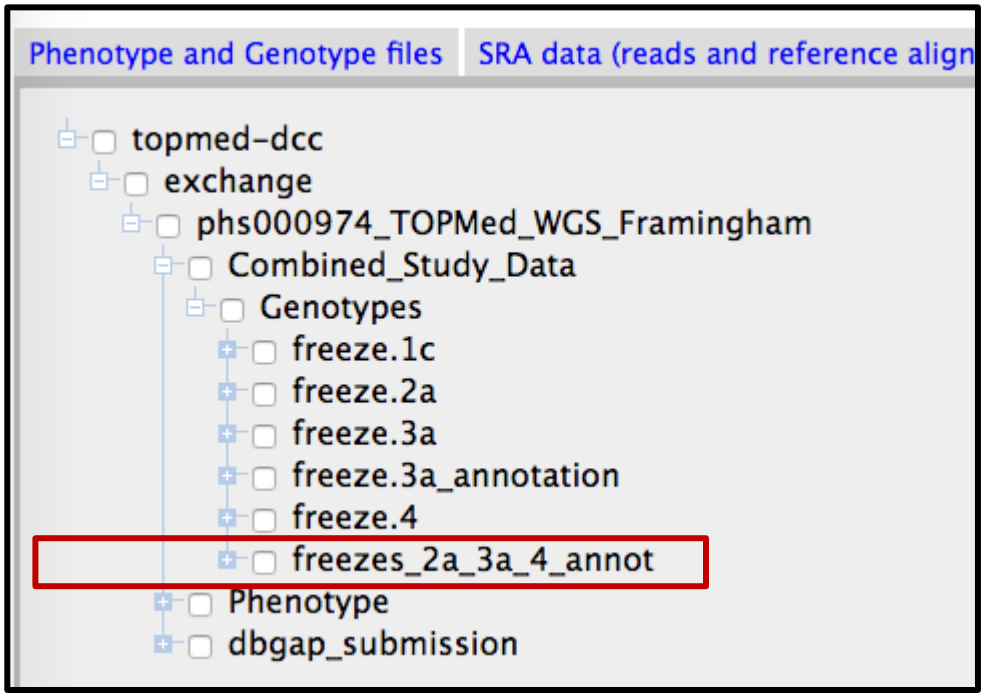
$\times \quad 4 \quad = 100$

The size of the annotation directory is **102G**
Chr1 general annotation file for SNVs is **5GB**

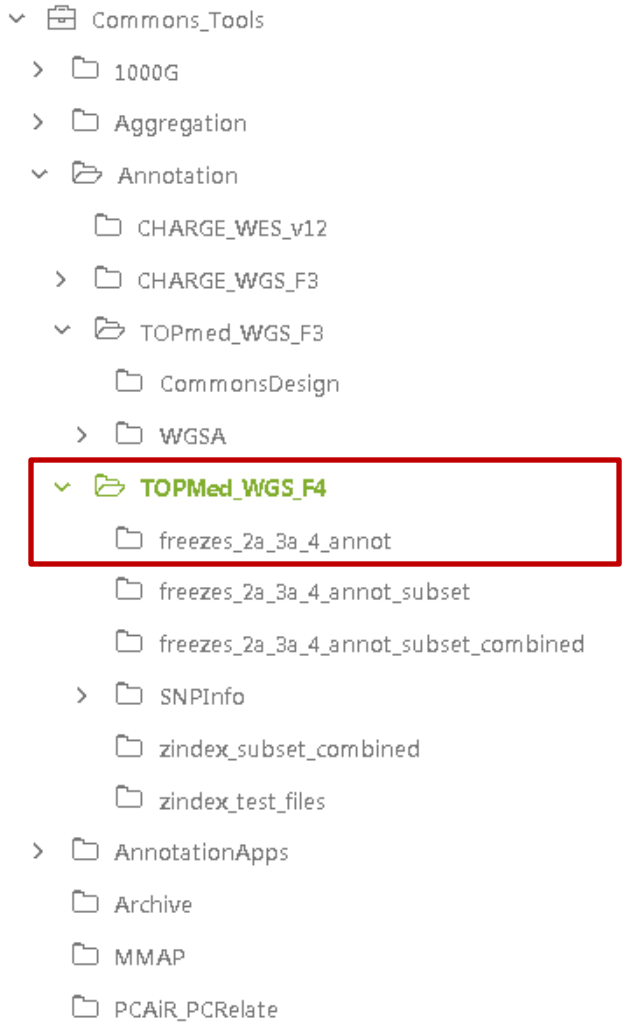
*Review data dictionary "*List of resources v0.65.pdf*" before using annotations*

Where to find WGSA annotations for TOPMed?

Exchange Area



Analysis commons



Gene-based annotations are at transcript level

chr:10273 T>C

VEP_ensembl_Transcript_ID

ENST00000456328|ENST00000488147|ENST00000438504|ENST00000515242|ENST00000541675|ENST00000423562|ENST00000450305|ENST00000538476|ENST00000518655

VEP_ensembl_Consequence

upstream_gene_variant|downstream_gene_variant|downstream_gene_variant|upstream_gene_variant|downstream_gene_variant|downstream_gene_variant|upstream_gene_variant|downstream_gene_variant|*splice_region_variant*

VEP_ensembl_Gene_Name

DDX11L1|WASH7P|WASH7P|DDX11L1|WASH7P|WASH7P|DDX11L1|WASH7P|DDX11L1

VEP_ensembl_Gene_ID

ENSG00000223972|ENSG00000227232|ENSG00000227232|ENSG00000223972|ENSG00000227232|ENSG00000227232|ENSG00000223972|ENSG00000227232|ENSG00000223972

Ensembl_Regulatory_Build_Overviews

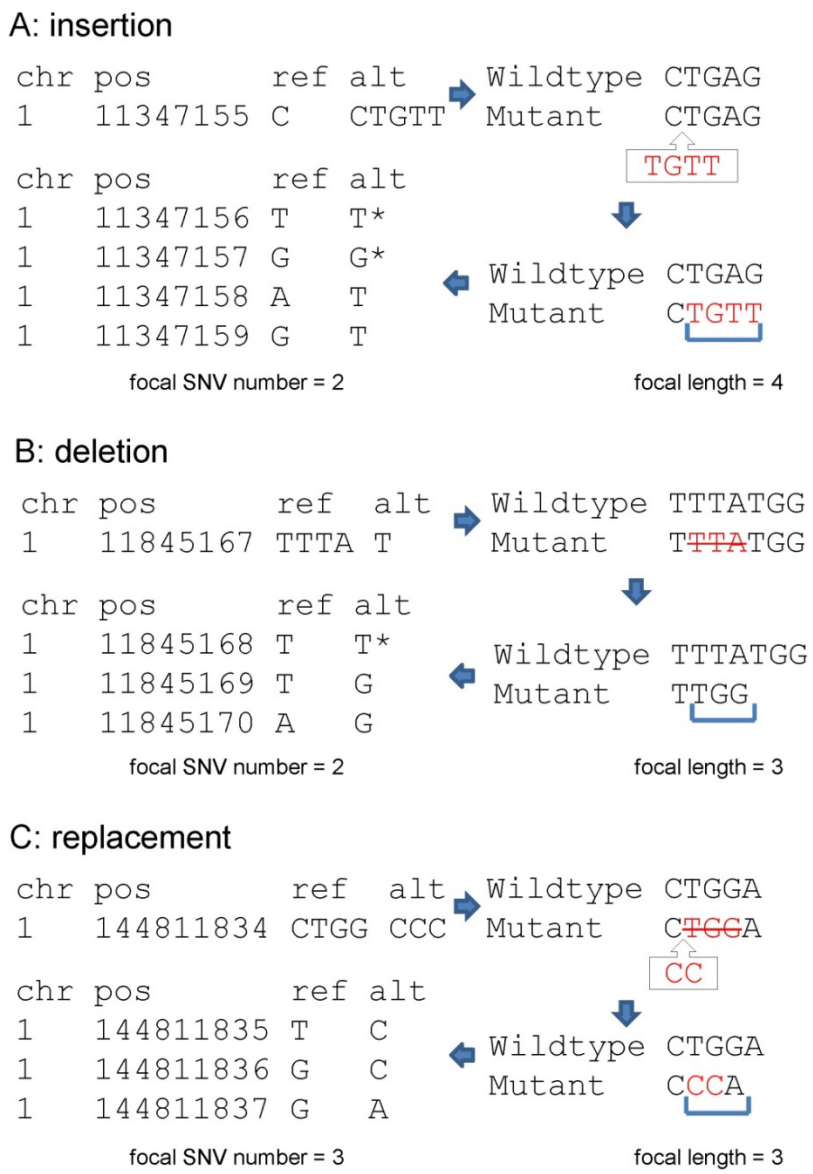
ctcf

VEP_ensembl_LoF

.|.|.|.|.|.|.|.|HC

NOTE: Missingness is denoted as "." in all annotation fields

For some annotations indels are translated into SNVs



GERP++ RS score
0.325{1}0.097{1}0.392{1}2.010{1}

Genomic Evolutionary Rate Profiling (GERP) identifies constrained elements in multiple alignments by quantifying substitution deficits. the larger the score, the more conserved the site

Part II
Defining and filtering aggregation units using variant annotations

How are annotations used?

1. Rare variant association tests
   - **To define and filter aggregation units**
   - Use as weights [1]

2. Fine map novel and previously known significantly associated loci to identify likely causal variants

[1]Morrison AC, Huang Z, Yu B, et al. Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am J Hum Genet. 2017;100(2):205-215.

Steps involved in generating aggregate variant list for association testing

STEP1: Define aggregation units

- which genomic regions will be included in each unit

STEP2: Decide on filtering criteria

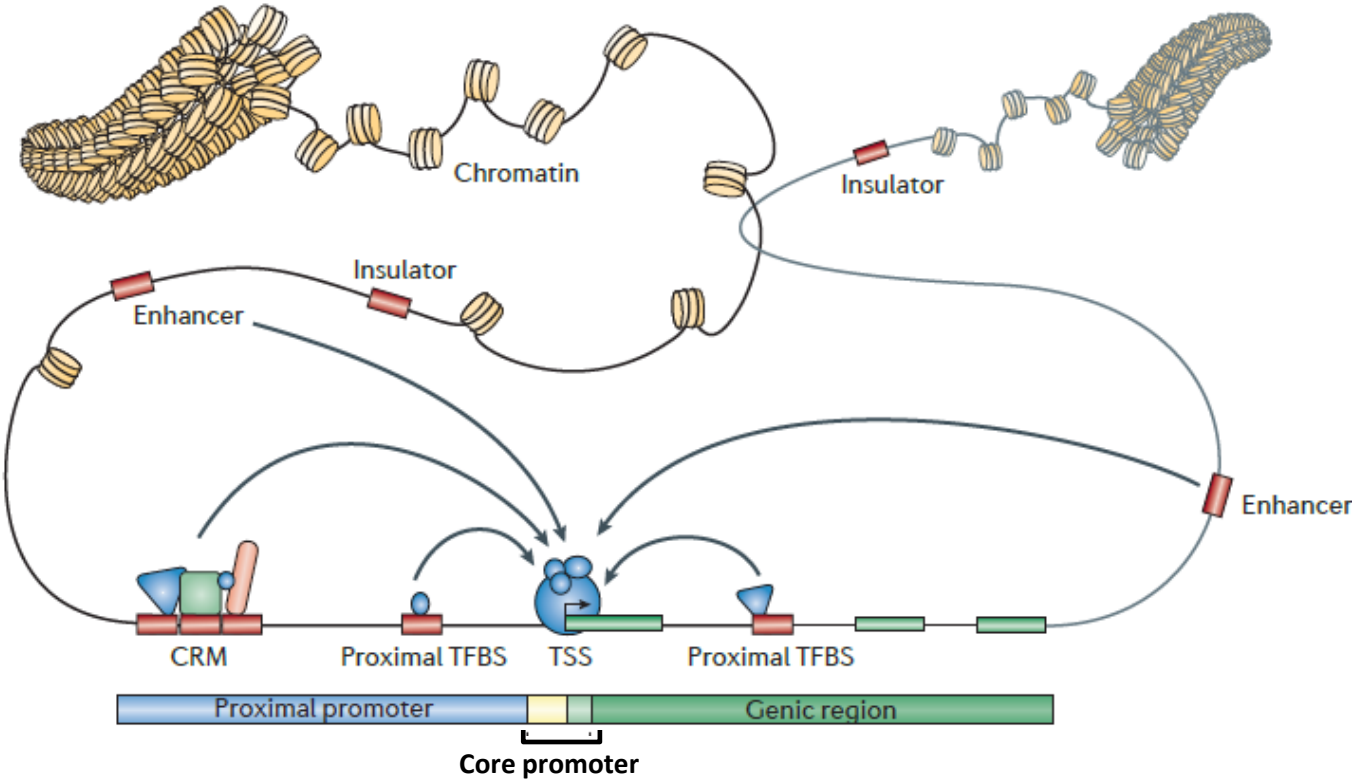
- which variants will be filtered within each unit

Goal is to create list of variants in each aggregation unit which can be used in multiple variants association tests (example Burden and SKAT tests)

PART II
STEP1: Define aggregation units

Gene is one of the fundamental units of biology and gene-based aggregation units are frequently used in rare variant association testing

What is a gene?



TFBS: transcription factor binding site,

CRM: cis-regulatory module

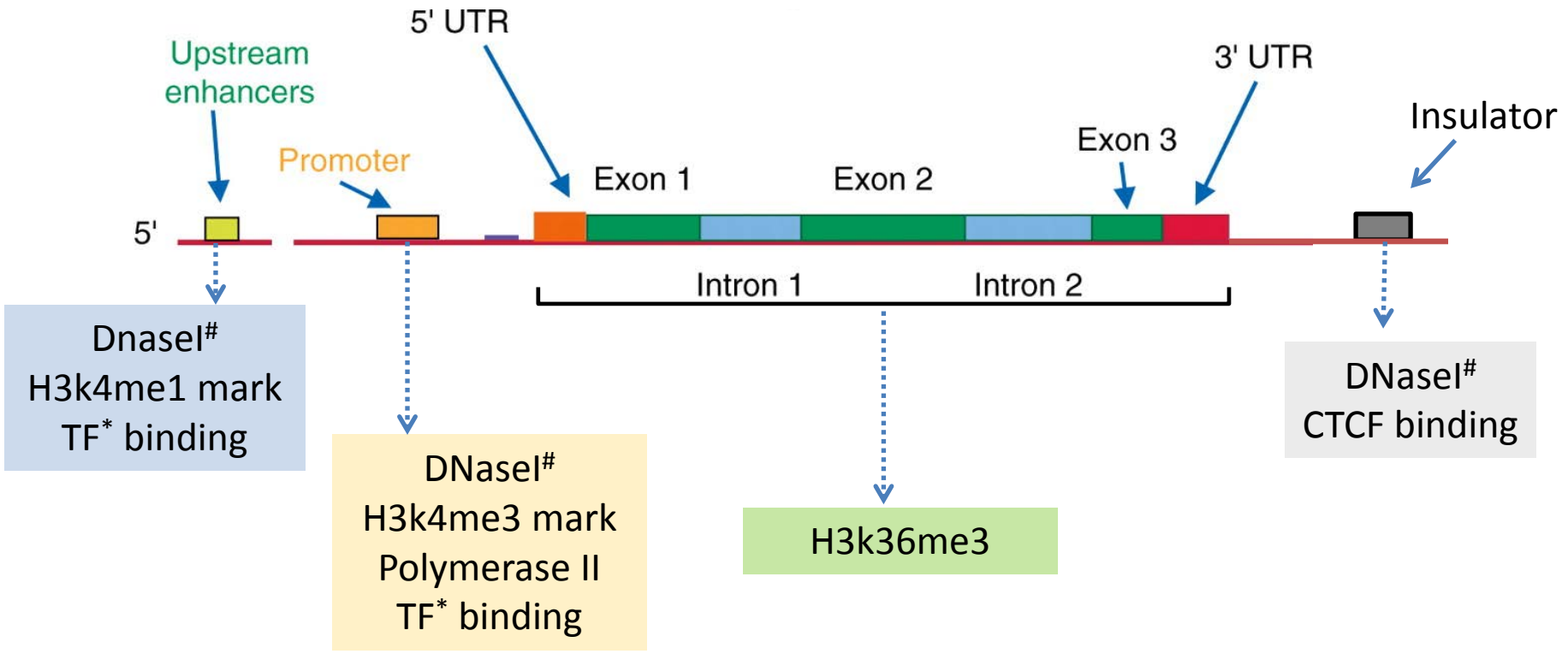
UTR: untranslated region

Functional gene unit = transcript + promoter + enhancers

Transcript = Exon + Intron + 3'UTR + 5'UTR

Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012;13(4):233-45.

Biochemical signatures typically associated with non-coding functional elements



Enhancer : Interacts with promoter can be involved in repression or induction of a gene
Promoter : Genomic element where the transcription machinery assembles
UTR : Untranslated region
EXON : Coding part of a transcript (mRNA)
INTRON : Non-Coding part of a transcript (mRNA)
Insulator : Barriers that protect genes from influence of outside enhancers or inactivating chromatin structures

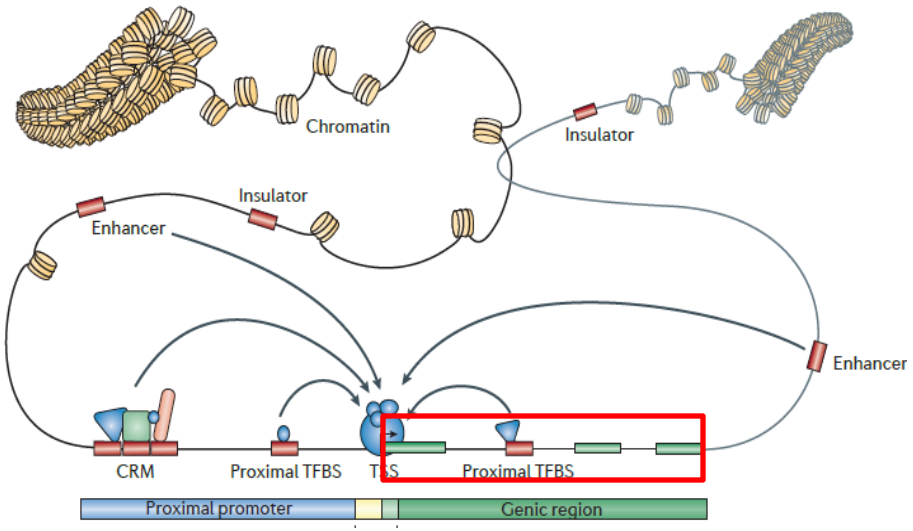
NOTE: These biochemical marks are tissue-specific . Additionally, these may also show temporal and treatment specific variations within a cell type

* TF : transcription factor, # DNaseI Hypersensitivity, which is an indicator of chromatin accessibility

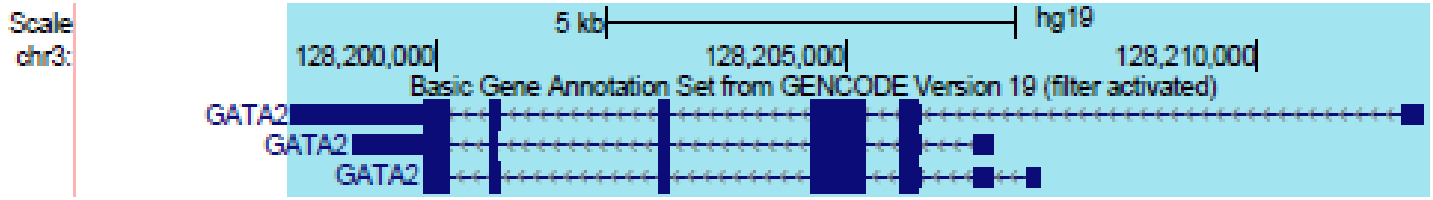
Modified from R. Searle and P. M. Hopkins, 2009

Functional gene unit = transcript + promoter + enhancers

Transcript and gene models

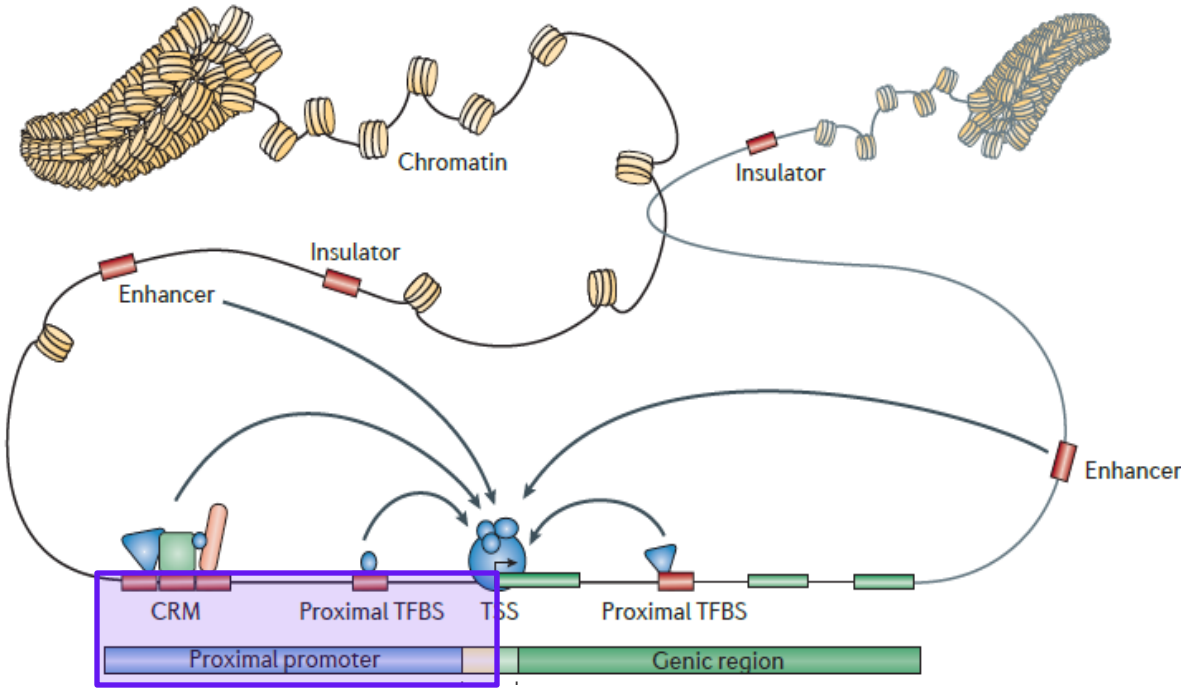


- Gene models: GENCODE, RefSeq, UCSC



- Gene
  - Genomic region spanning all the transcripts of a gene
  - A single transcript (example expressed transcripts in RNA-Seq data of relevant tissue)

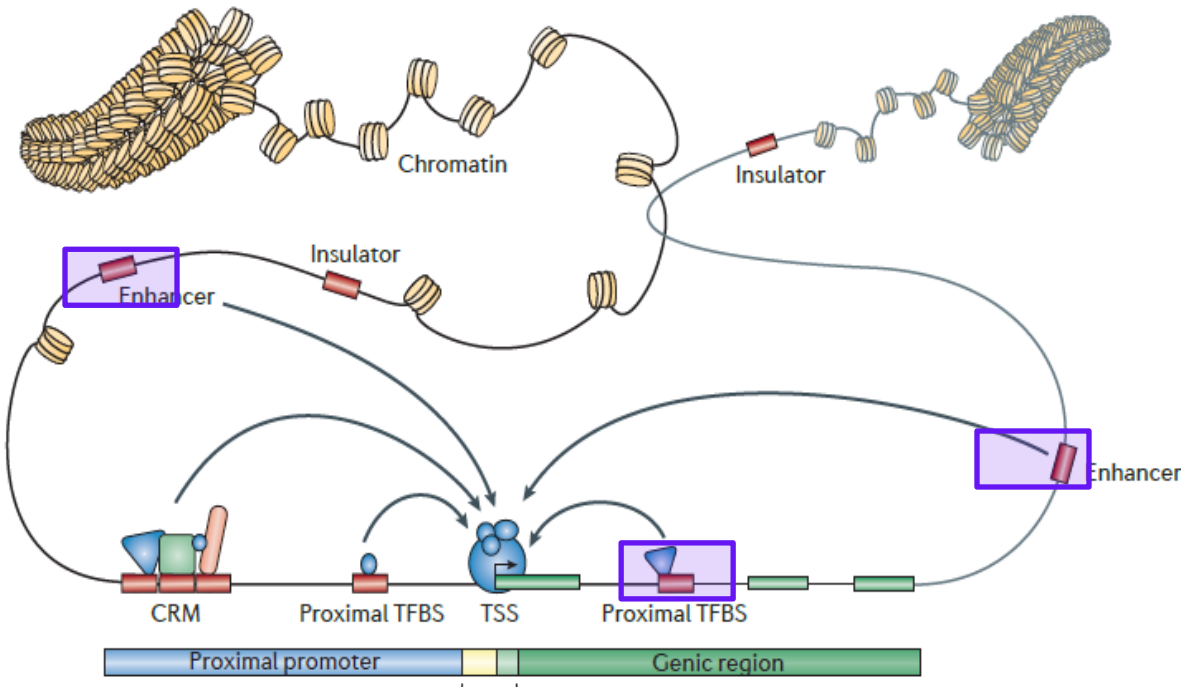
Promoters



- Some distance upstream from TSS (typically 5Kb)
- 5Kb upstream overlapping with H3K4me3 and or H3K27ac mark
- 5Kb upstream overlaps with DNaseI hypersensitive regions
- 5Kb upstream that overlaps with CAGE peaks[1]

[1]Morrison AC, Huang Z, Yu B, et al. Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am J Hum Genet. 2017;100(2):205-215.

Enhancers



- Flanking regions overlapping with H4K4me1 and or H3K27ac
- Flanking regions overlapping with DNaseI hypersensitive regions
- Enhancer-gene link predictions[1,2]
- Chromosome conformation capture (3C,4C,Hi-C etc.)

[1]*Thurman RE et.al Nature. 2012 Sep 6; 489(7414):75-82.*
[2]*Forrest AR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.*

Example gene-based aggregation units

- Gene + flanking regions
- Gene + enhancer + promoter
- UTR's+ enhancer + promoter
- Promoter of a gene
- First intron of a transcript

Aggregation units that are continuous genomic ranges

chromosome	start	end	gene_id
chrX	99883667	99894988	ENSG00000000003.10
chrX	99839799	99854882	ENSG00000000005.5
chr20	49551404	49575092	ENSG00000000419.8
chr1	169818772	169863408	ENSG00000000457.9
chr1	169631245	169823221	ENSG00000000460.12
chr1	27938575	27961788	ENSG00000000938.8

Aggregation units that are non-continuous genomic ranges

chromosome	start	end	gene_id	type
chr1	27938575	27961788	ENSG00000000938.8	gene
chr1	27945001	27945200	ENSG00000000938.8	7Enh
chr1	27945201	27945400	ENSG00000000938.8	7Enh
chr1	27945401	27945600	ENSG00000000938.8	7Enh
chr1	27945601	27945800	ENSG00000000938.8	7Enh

Other units of aggregation
(not covered in this workshop)

- Genes in a pathway[1]
- Protein-protein interaction domains[2]
- Topological associated domains[3]
- Specific histone modification marks[3]
- DNaseI Hypersensitive sites/nucleosome depleted regions
- …

[1]Richardson TG, Timpson NJ, Campbell C, Gaunt TR. A pathway-centric approach to rare variant association analysis. Eur J Hum Genet. 2016;25(1):123-129.
[2]Richardson TG, Shihab HA, Rivas MA, et al. A Protein Domain and Family Based Approach to Rare Variant Association Analysis. PLoS ONE. 2016;11(4):e0153803.
[3]Lumley, T., Brody, J., Peloso, G., & Rice, K. (2016). Sequence kernel association tests for large sets of markers: tail probabilities for large quadratic forms. doi:10.1101/085639

PART II
STEP2: Decide on filtering criteria

Scenario 1: simple filtering

- Genic unit

  Transcript range + 20 kb flanking region upstream and downstream

- Filters:

  CADD phred score>=10 and MAF<=1%

Combined Annotation Dependent Depletion (CADD) is a tool for scoring the deleteriousness of variants. A scaled C-score of greater or equal 10 indicates that these are predicted to be the 10% most deleterious substitutions that you can do to the human genome

Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37.

Scenario 2: Using multi-tissue regulatory regions

- Genic unit
  Gene + 20 kb flanking region upstream and downstream

- Filters:
  A. Flanking region
     – Overlaps with “Ensembl_Regulatory_Build_Overviews”

  A. Gene region
     – Overlaps with “Ensembl_Regulatory_Build_Overviews” OR
     – Overlaps with LOF variants

- Ensembl_Regulatory_Build_Overviews
  - genome segment prediction based on 17 cell types from ENCODE and Roadmap.
  - ctcf - CTCF binding sites,
  - distal - Predicted enhancers
  - open - Unannotated open chromatin regions
  - proximal - Predicted promoter flanking regions
  - tfbs - Unannotated transcription factor binding sites
  - tss - Predicted promoters

- ENCODE_Dnase_cells: number of cell lines supporting a DNase I hypersensitive site

Scenario 3: Using tissue specific regulatory regions -I

- All epigenomics marks of erythroleukemia cell line K562 (EID: E123) in WGSA
  - E123-DNase.macs2.narrowPeak
  - E123-H2A.Z.narrowPeak
  - E123-H3K27ac.narrowPeak
  - E123-H3K27me3.narrowPeak
  - E123-H3K36me3.narrowPeak
  - E123-H3K4me1.narrowPeak
  - E123-H3K4me2.narrowPeak
  - E123-H3K4me3.narrowPeak
  - E123-H3K79me2.narrowPeak
  - E123-H3K9ac.narrowPeak
  - E123-H3K9me1.narrowPeak
  - E123-H3K9me3.narrowPeak
  - E123-H4K20me1.narrowPeak

Scenario 3: Using tissue specific regulatory regions -II

- Genic unit

  Gene + 20 kb flanking region upstream and downstream

- Filters:

  A. Flanking region

  – Overlaps with either H3K4me3 or H3K4me1 enriched regions) & DNaseI hypersensitivity sites in k562 cells

  B. Gene region

  – overlap (either H3K4me3 or H3K4me1 enriched regions) & DNaseI hypersensitivity sites in k562 cells OR

  – Overlaps with LOF variants

DCC will provide frequently used aggregation units on the EA

- GENCODE Gene + flanking regions
  - 0kb,5kb ,20kb,100kb,200kb
  - Filtered to keep variants overlapping “Ensembl_Regulatory_Build_Overviews” and LOF variants
- Promoter
  - 5Kb upstream of GENCODE genic unit
  - Filtered to keep “tss” and “proximal” overlapping variants from “Ensembl_Regulatory_Build_Overviews”
- First intron of GENCODE transcripts
- DCC will try to accommodate requests for assistance in defining units and lists of variants within them
- Requests for creating aggregation units and list of variants within them will be accepted as time permits

Summary

- A large set of annotations are available for TOPMed through WGSA
- Annotations can be used to create and filter aggregation units
- Frequently used aggregation units will be made available by DCC on the Exchange Area

Part III
Hands-on exercise for generating variants list used for aggregation tests

Key libraries and functions

Parse the WGSA annotation file	
library (wgsaparsr)	Package for working with WGSA output files
get_fields	list the annotation fields available in a WGSA output file
parse_to_file	Converts list-fields into multiple rows
parse_indel_to_fille	Same as “parse_to_file” but for indel annotations
Create aggregation units file using GENCODE genes	
library(genetable)	Package for working with .gtf gene model files
import_gencode	import the gtf file to a tidy data frame
filter_gencode	filter gtf file on different features and tags
define_boundaries	define the boundaries of the feature of interest
Aggregate variants by genic units and create input file for association testing	
- R code for the workshop - DCC uses a MySQL server for creating and filtering variant list in aggregation units using WGSA annotations	

Aggregation unit input file

- Aggregation unit is a gene and 20 kb flanking region upstream and downstream of it
- Only subset of 1000K variants used for the workshop were used
- No annotation based filtering was performed on the variants
- Indels are not included

Header of aggregation unit input file

group_id	chromosome	position	ref	alt
ENSG00000188157.9	1	970546	C	G
ENSG00000242590.1	1	970546	C	G
ENSG00000188157.9	1	985900	C	T
ENSG00000217801.5	1	985900	C	T
ENSG00000242590.1	1	985900	C	T
ENSG00000273443.1	1	985900	C	T

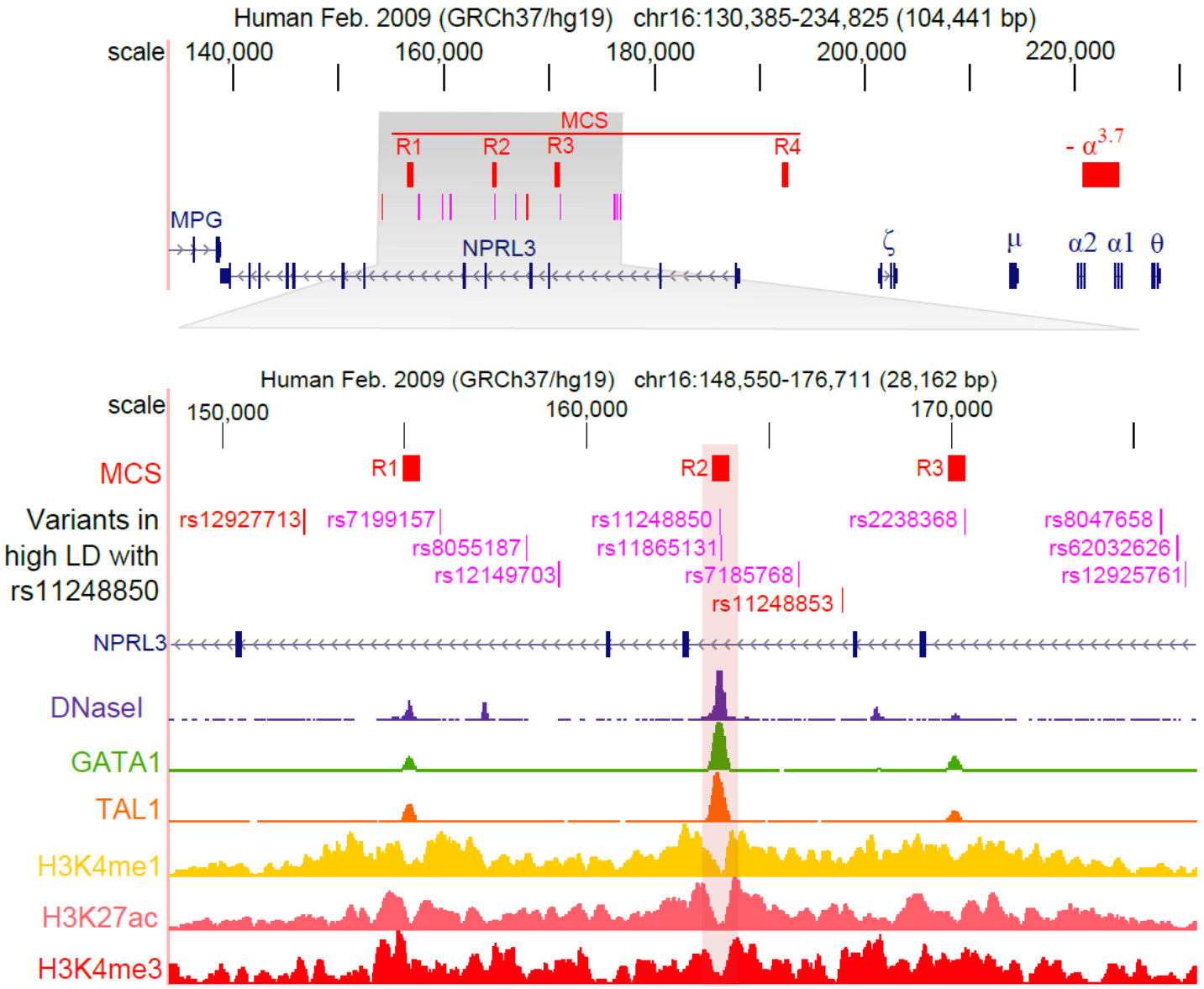
Aggregation unit identifier to which the variant belongs

Variant information

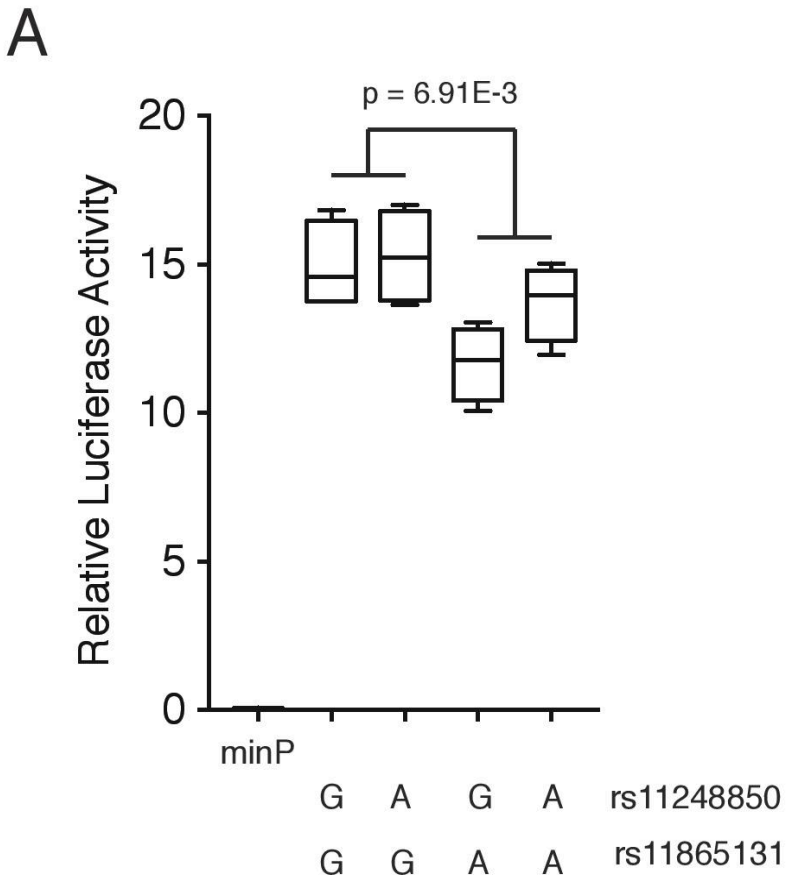
NOTE: A given variant can belong to multiple aggregation unit

EXTRA SLIDES

Predicting putative causal variants

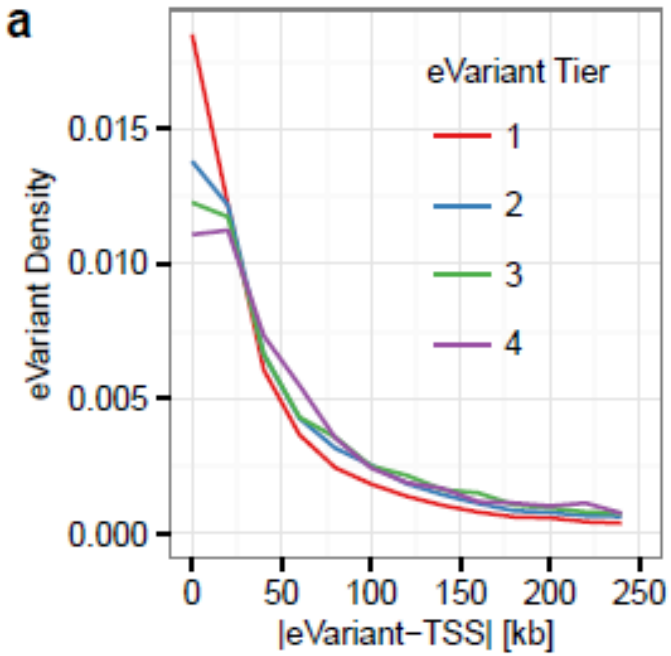


Functional assays confirm allele-specific activity of the predicted causal variants



Courtesy Vijay Shankaran

Choosing the length of flanking regions for gene-based aggregation units



Aguet, F., Brown, A., Castel, S., Davis, J., Mohammadi, P., & Segre, A. et al. (2016). Local genetic effects on gene expression across 44 human tissues. doi:10.1101/074450