National Human Genome Research Institute

Analyses and Annotation for Family-based Genome-wide Data

Anthony Musolf, PhD
Research Fellow
Computational and Statistical Genomics Branch
National Human Genome Research Institute/National Institutes of Health

The Forefront of Genomics®

Overview

1. Background

NIH
NHGRI

Overview

1. Background
2. Gene-based TDT

NIH
NHGRI

Overview

1. Background
2. Gene-based TDT
3. Extended family linkage analysis

NIH
NHGRI

Overview

1. Background
2. Gene-based TDT
3. Extended family linkage analysis
4. Gene-based association with related individuals

NIH
NHGRI

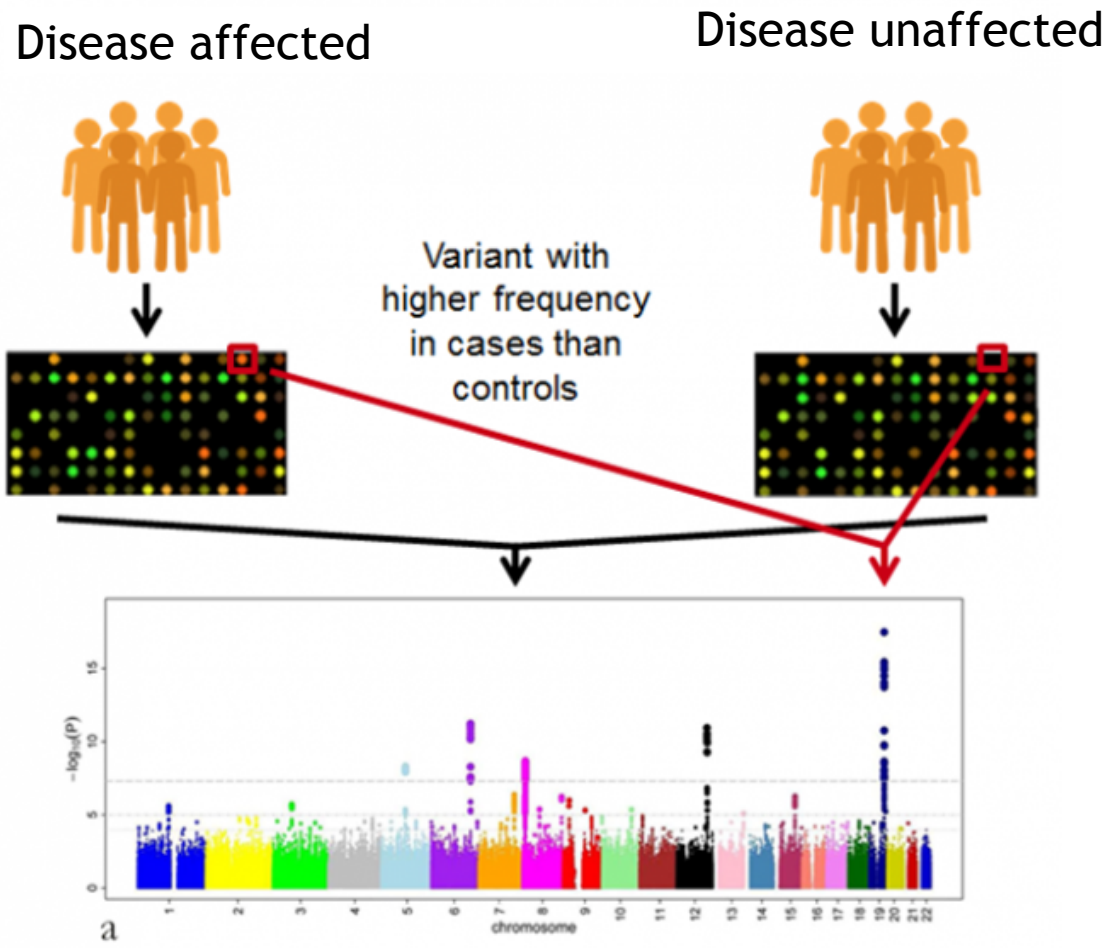
Overview

1. Background
2. Gene-based TDT
3. Extended family linkage analysis
4. Gene-based association with related individuals
5. Annotation of results

NIH
NHGRI

Background – Genetic Analysis

- Analysis of genotypic data assumes you have genotype and phenotypic data on set of subjects
  - Genotype = microarray, exome chip, whole exome sequence, whole genome sequence
  - Phenotype = Binary (affected/ unaffected), quantitative



NIH
NHGRI

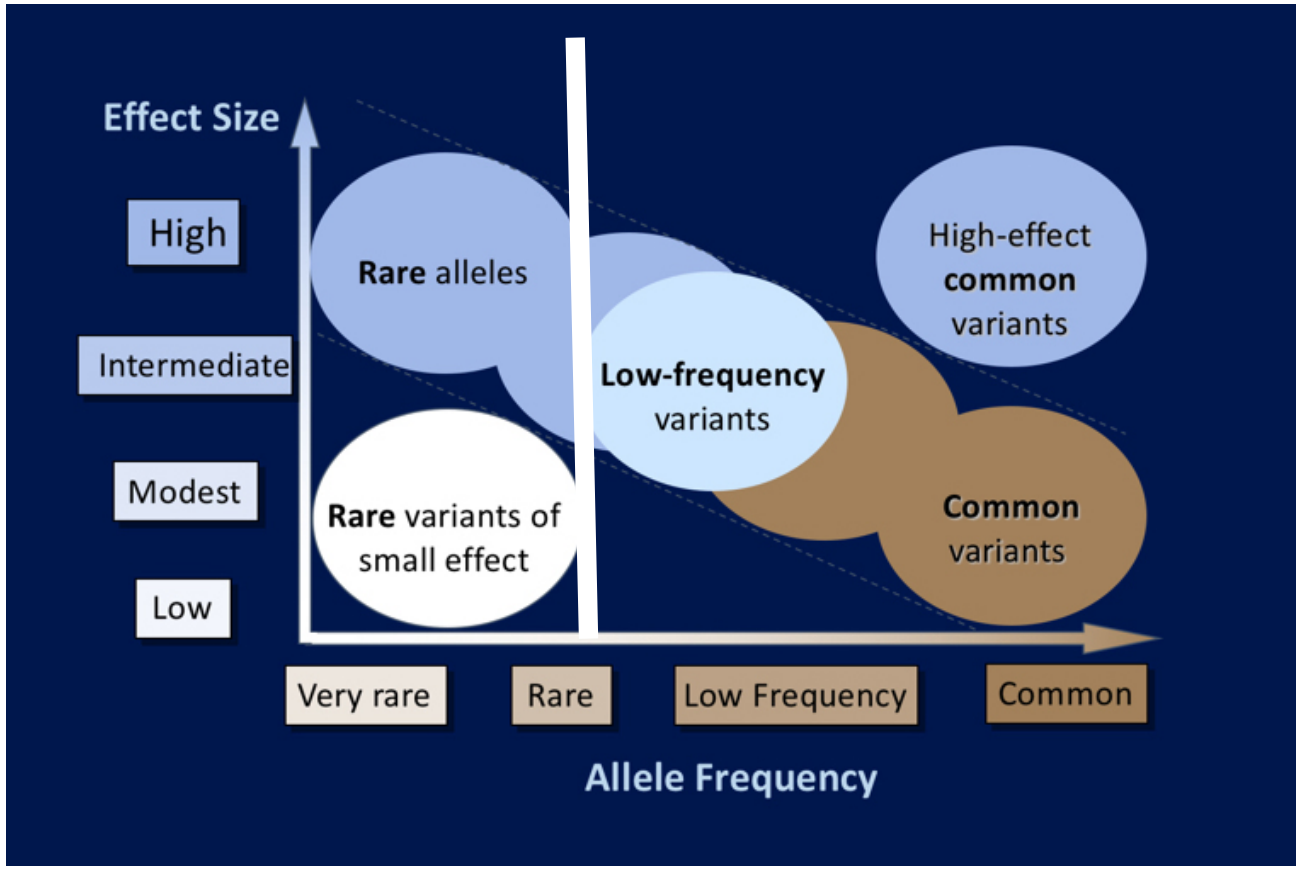
Background – Family Studies

- Most genetic studies are performed with unrelated individuals
  - Population-based study
- Some genetic studies are performed with related individuals
  - Family-based study
- Different kinds of family data
- Determine which kind of analysis will be performed

NIH
NHGRI

Why Family-based Studies?

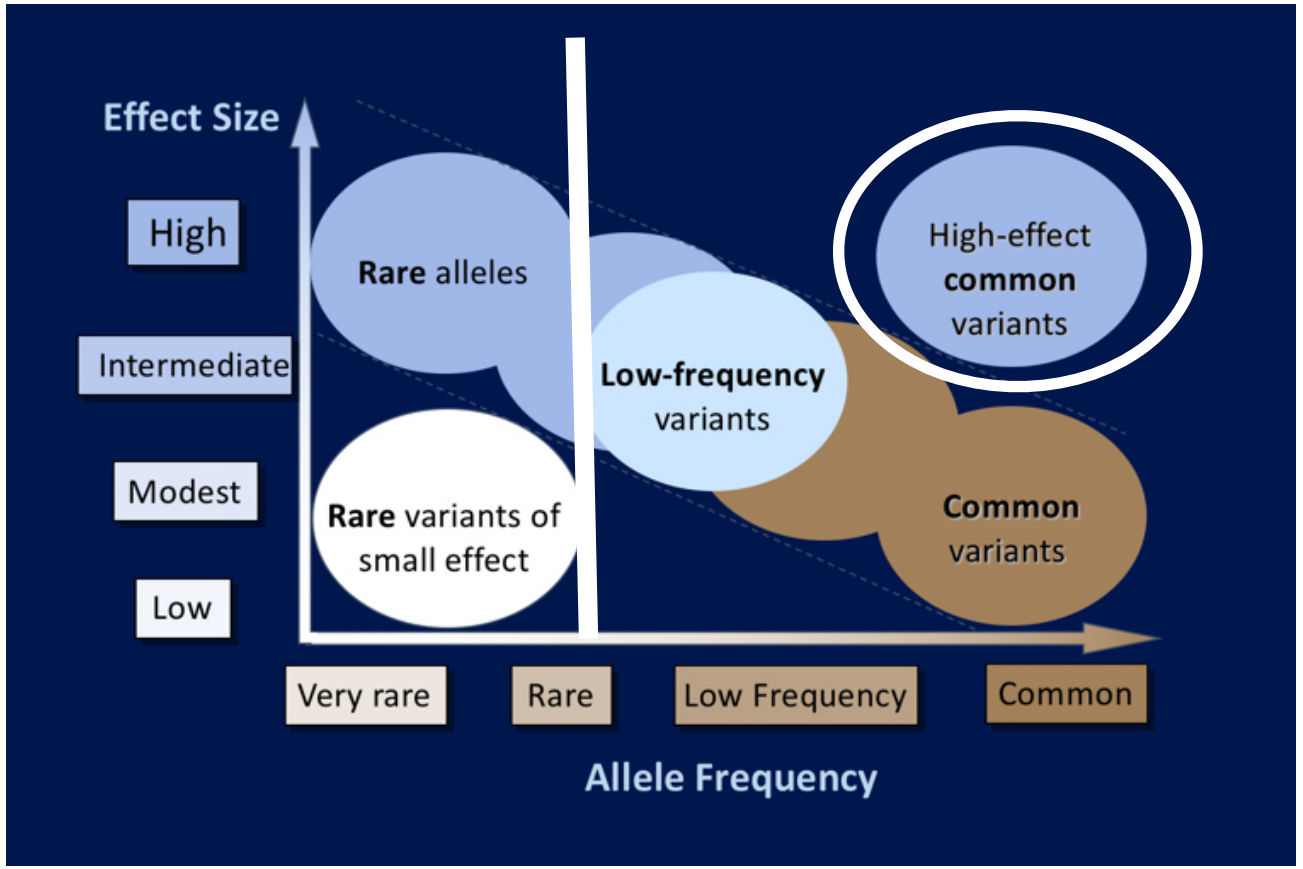
- Advantages over population-based studies
  - Aggregated families are often enriched for rare variants that are potentially highly penetrant



NIH
NHGRI

Why Family-based Studies?

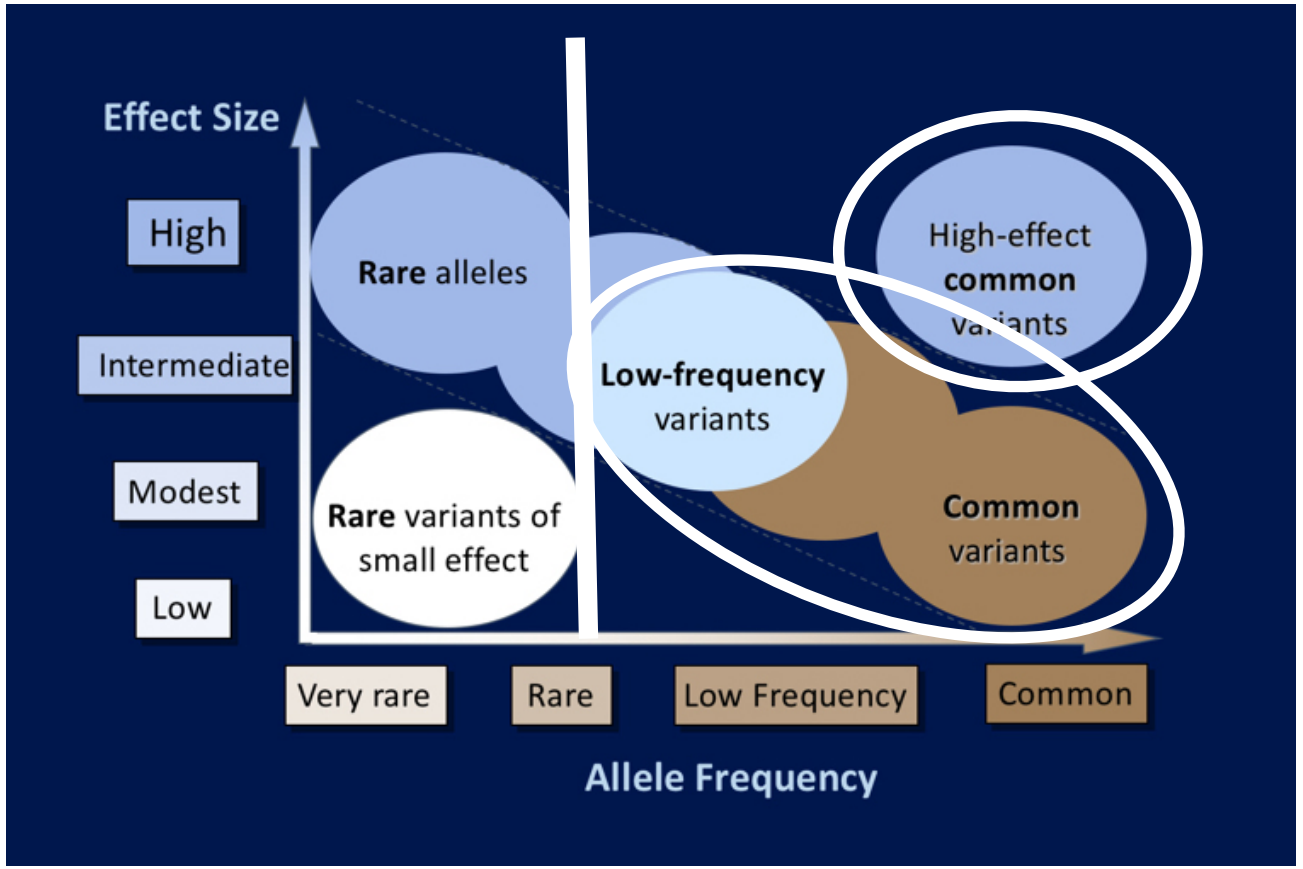
- Advantages over population-based studies
  - Aggregated families are often enriched for rare variants that are potentially highly penetrant



NIH
NHGRI

Why Family-based Studies?

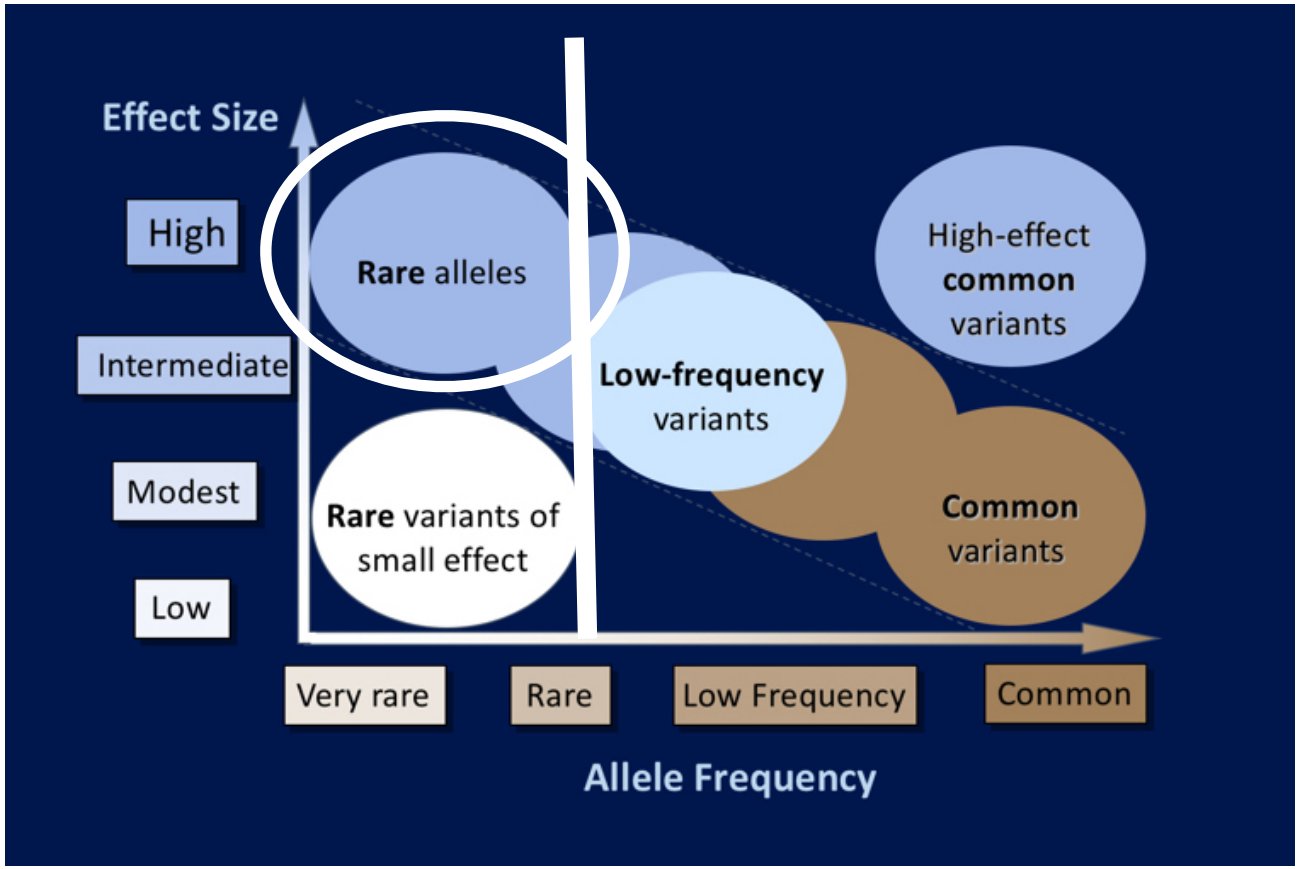
- Advantages over population-based studies
  - Aggregated families are often enriched for rare variants that are potentially highly penetrant



NIH
NHGRI

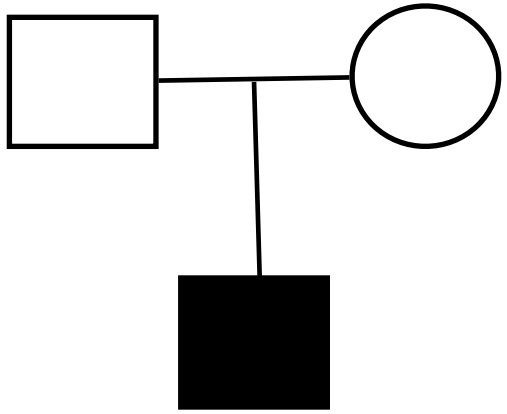
Why Family-based Studies?

- Advantages over population-based studies
  - Aggregated families are often enriched for rare variants that are potentially highly penetrant



NIH
NHGRI

Types of Family Data

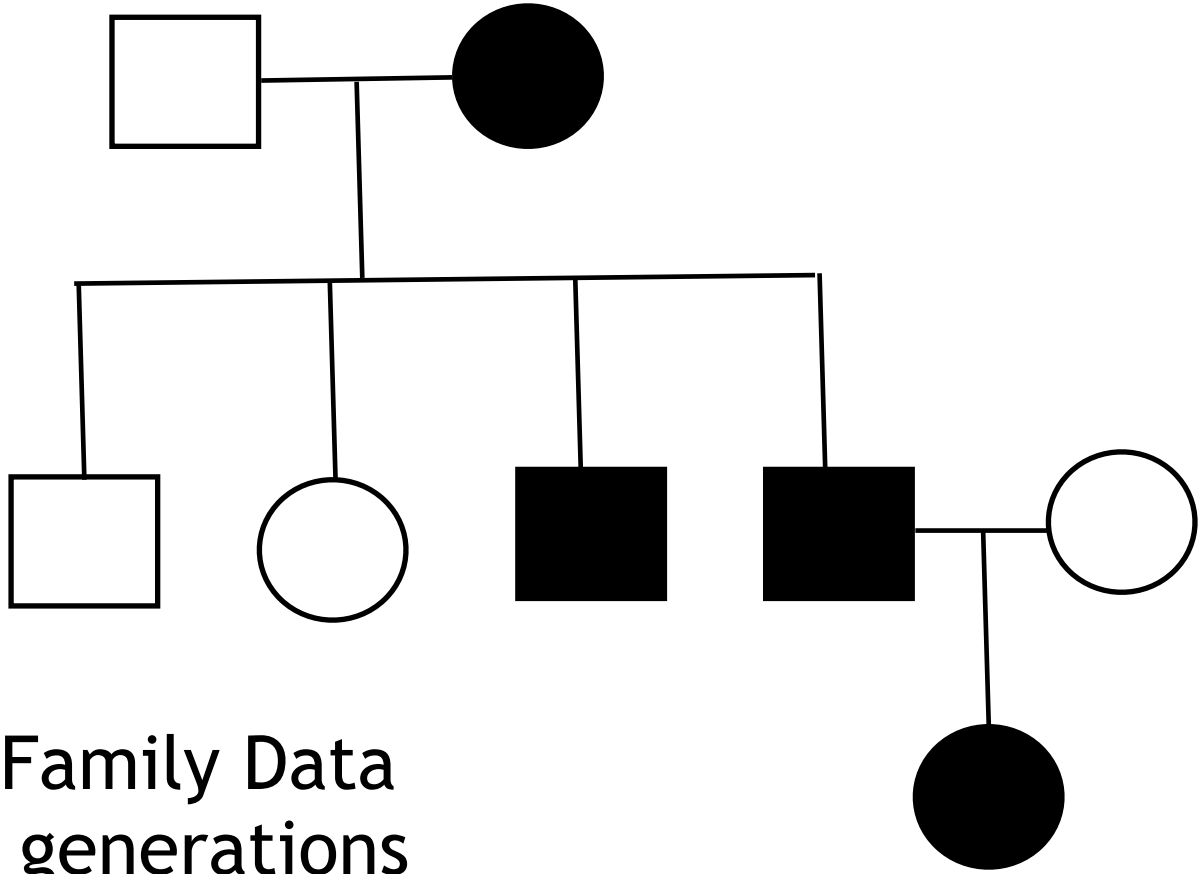


Trio Data

- Two genotyped parents, one genotyped affected child
- TDT, Gene-based TDT

NIH NHGRI

Types of Family Data

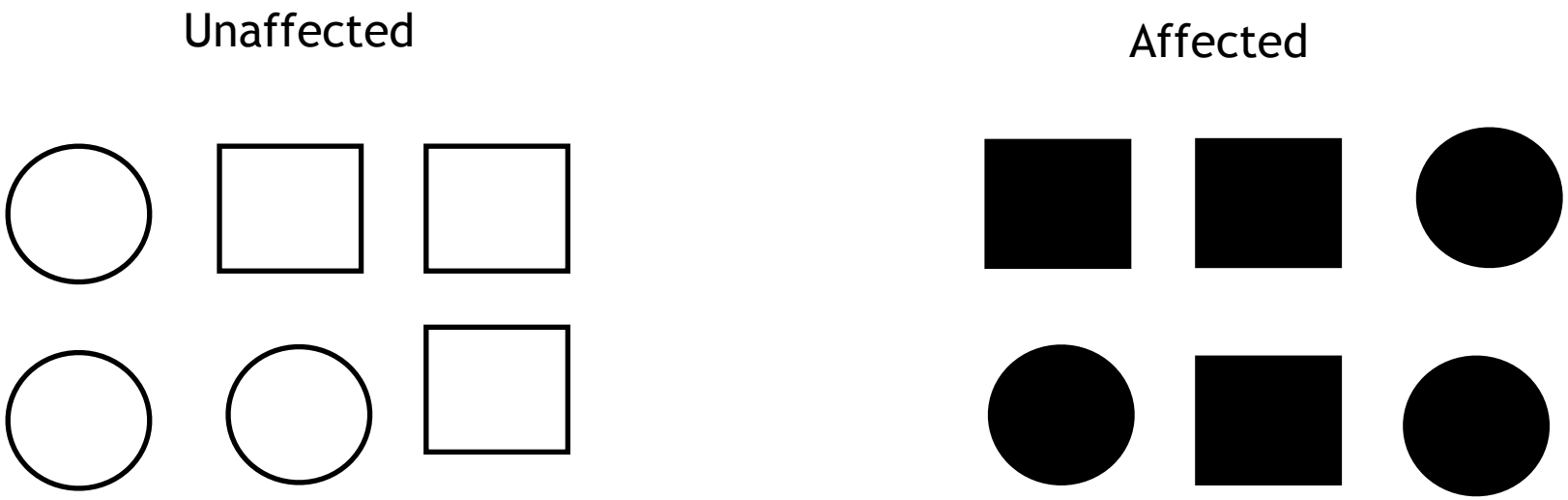


Extended Family Data

- Multiple generations
- Linkage analysis

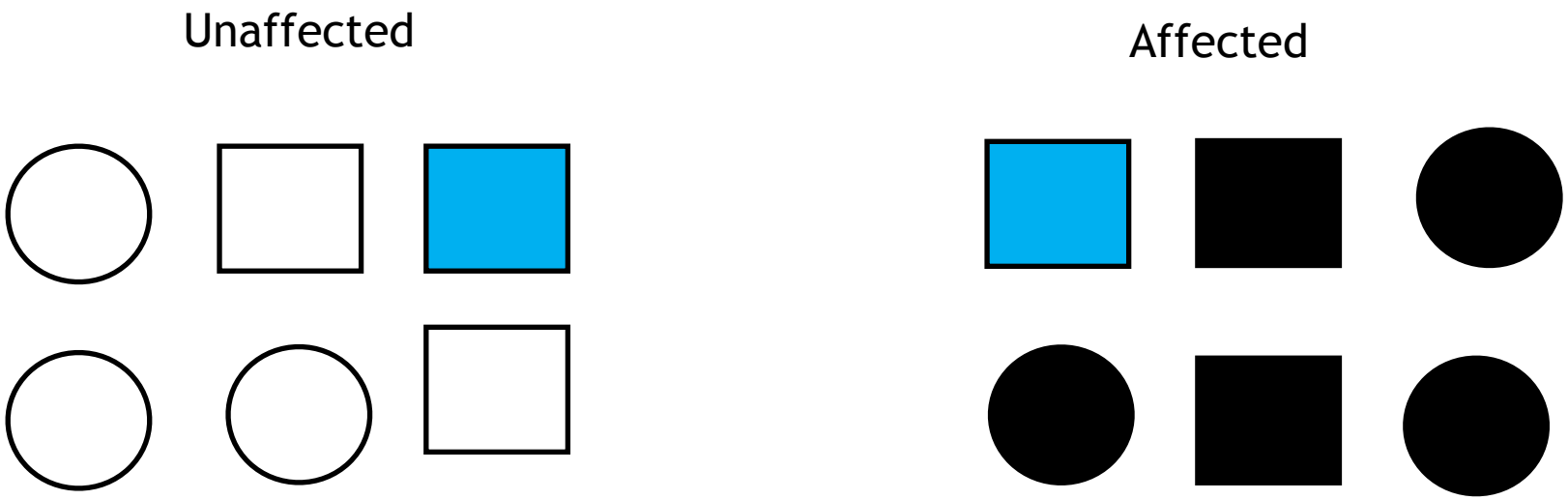
NIH NHGRI

Types of Family Data



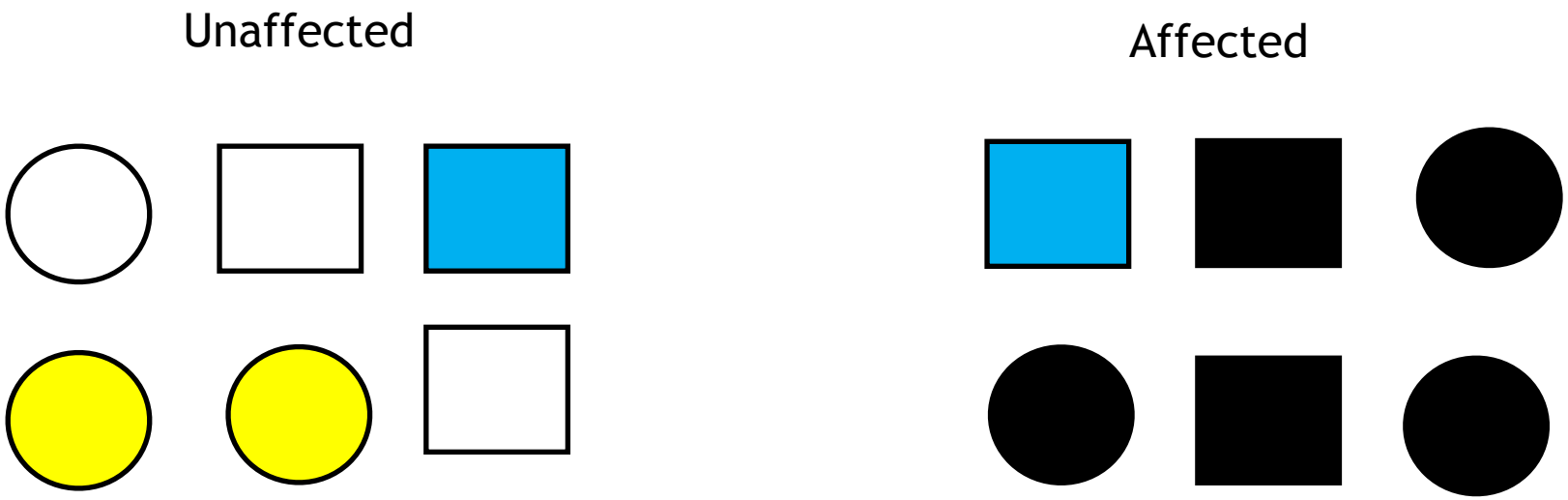
NIH
NHGRI

Types of Family Data



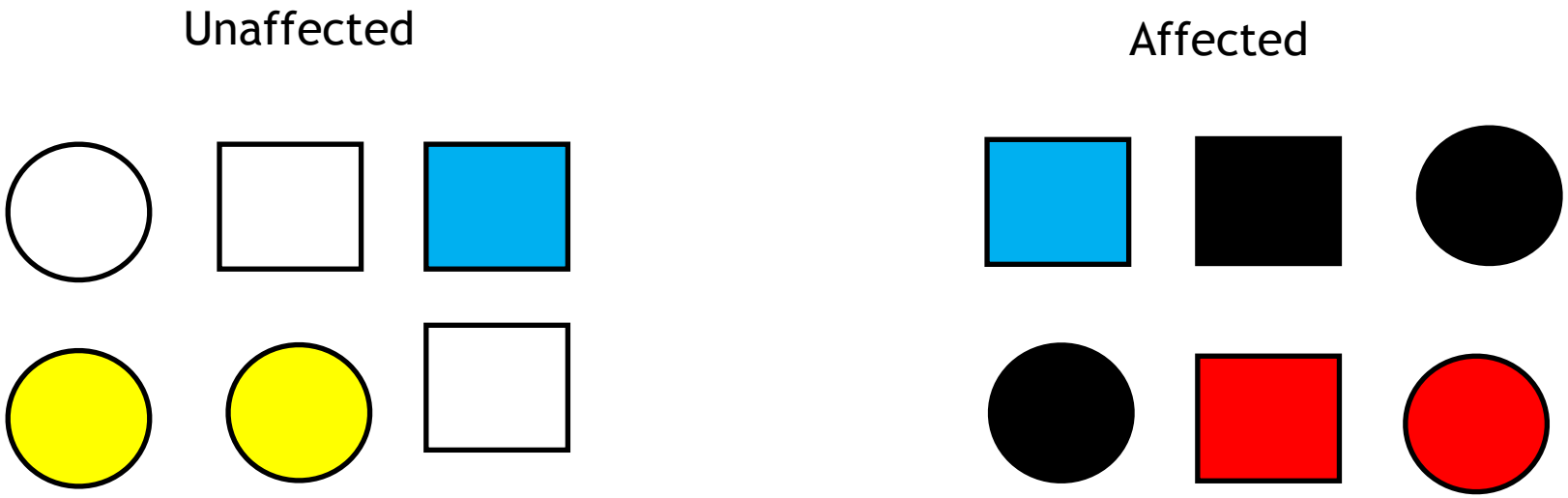
NIH
NHGRI

Types of Family Data



NIH
NHGRI

Types of Family Data



Population-based study with related individuals

- Mix of related unrelated individuals
- Could remove related individuals
- Can use kinship matrix to keep both in

NIH
NHGRI

Analysis of Trio Data

- Transmission Disequilibrium Test (TDT)
- Use trio data to find which alleles are transmitted to child from parent at a single varaint
- Finds linkage in the presence of genetic association
- Robust to population stratification

NIH
NHGRI

Gene-based TDT

- Gene-based TDT is similar to standard TDT
- Standard TDT looks for transmission at single marker level
- Gene-based does this at gene level
- Particularly useful for rare variants

NIH NHGRI

How do we perform a gene-based TDT?

- Many gene-based tests; not many adapted to TDT
  - rvTDT uses gene-based tests adapted for TDT
  - https://github.com/statgenetics/rv-tdt
- Uses PLINK files
- WARNING – rvTDT is not user friendly!

NIH
NHGRI

Running rvTDT

- Step 1 – rvTDT requires phased data
  - Phase data to determine haplotypes
  - Allows us to know which allele is on with chromosome
  - Identify recombination
  - We can phase PLINK data using a program called SHAPEIT
    - https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

```
shapeit --input-bed PLINK.bed PLINK.bim PLINK.fam --seed 123456789  --output-max PLINK.haps
PLINK.phased.sample
```

- Input bed = binary PLINK files

NIH
NHGRI

SHAPEIT Output

- SHAPEIT will output two files
- .sample file which contains subject information
  - Family/Individual IDs
  - Parental IDs
  - Sex and phenotype
- .haps file with phased genotype information

20:2228212-SNV 20:2228212-SNV 2228212 T C 1 0 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
1 1 1 1 1 1

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

20:2228212-SNV 20:2228212-SNV 2228212 T C 1 0 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 0 1 1 1 1 1 1

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

20:2228212-SNV **20:2228212-SNV 2228212 T C** 1 0 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 0 1 1 1 1 1 1

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

    20:2228212-SNV **20:2228212-SNV 2228212 T C** 1 0 1 1 1 1 1 1
    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1
    1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
    1 1 1 0 1 1 1 1 1 1
  - Remove three columns (cut or awk)

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

  20:2228212-SNV 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1

  - Remove three columns (cut or awk)

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

  20:2228212-SNV 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
  - Remove three columns (cut or awk)
- Map file (extension .map)

  DPY19L1P1   rs541054536   0.002577

NIH
NHGRI

rvTDT Input Files

- Requires three files
  - Genotype file (extension .tped)

  20:2228212-SNV 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1

  - Remove three columns (cut or awk)
- Map file (extension .map)

  DPY19L1P1 rs541054536 0.002577
- Phenotype file (extension .phen)

182-1 182 0 0 1 0

NIH
NHGRI

Running rvTDT with phased data

rvTDT PROJECT-NAME –G INPUT.tped –P INPUT.phen –M INPUT.map –u MAF

- Output will provide two folders
  - PROJECT-NAME_pval
  - PROJECT-NAME_rvTDT
- Pval contains a list of each gene and its pvalues
- Each gene has its own file

NIH
NHGRI

P-value Output file

- Output file will contain gene name and p-values for different tests

#gene	CMC-Analytical	BRV-Haplo	CMC-Haplo	VT-BRV-Haplo	VT-CMC-Haplo	WSS-Haplo
BMP7	0.439633	0.642715	0.439122	0.788423	0.40519	0.62475

- rvTDT runs six different tests
- Each test varies on how the gene markers are created

NIH
NHGRI

Pvalue Output file

- Output file will contain gene name and pvalues for different tests

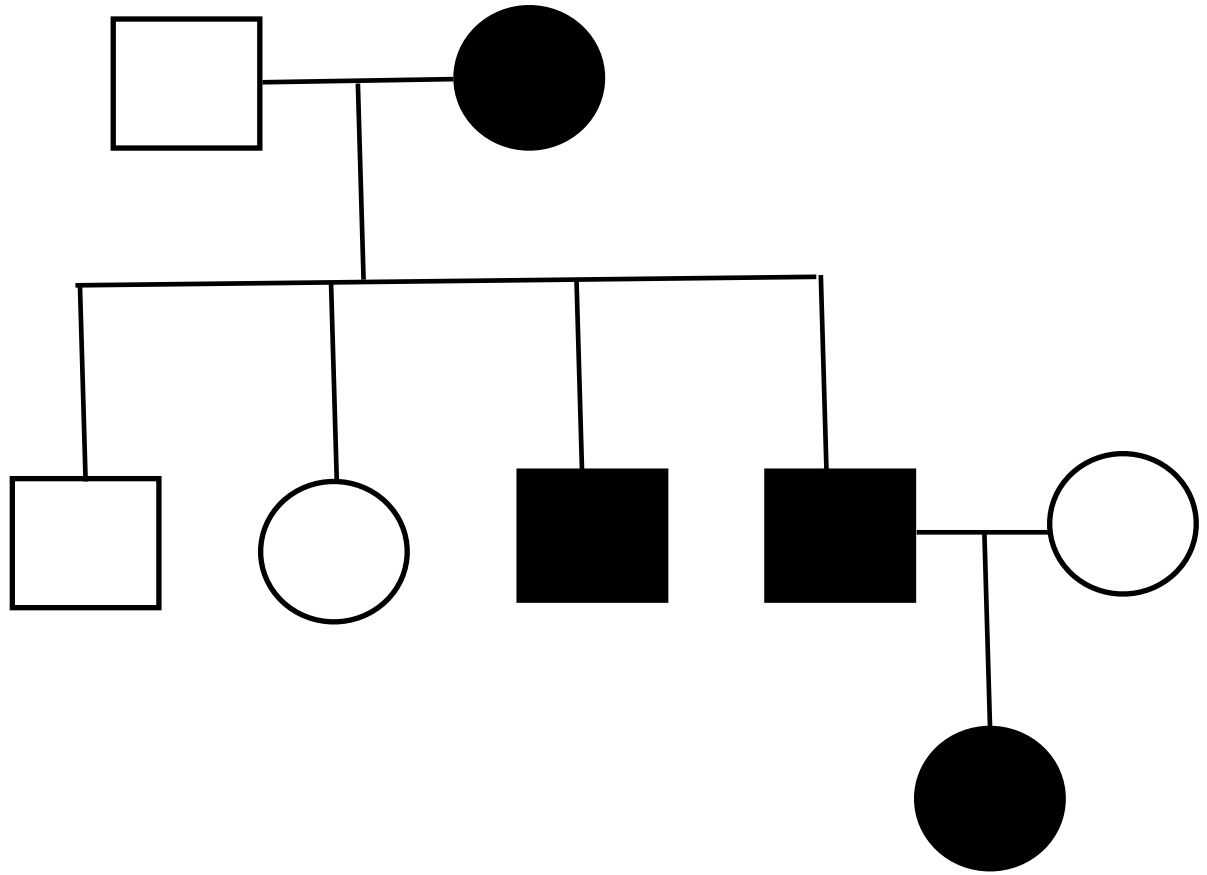
#gene	CMC-Analytical	BRV-Haplo	**CMC-Haplo**	VT-BRV-Haplo	VT-CMC-Haplo	WSS-Haplo
BMP7	0.439633	0.642715	**0.439122**	0.788423	0.40519	0.62475

- rvTDT runs six different tests
- Each test varies on how the gene markers are created
- Personally, I prefer CMC-Haplo

NIH
NHGRI

Analysis of Extended Family Data

- Genetic linkage analysis
- Slightly different than association



NIH
NHGRI

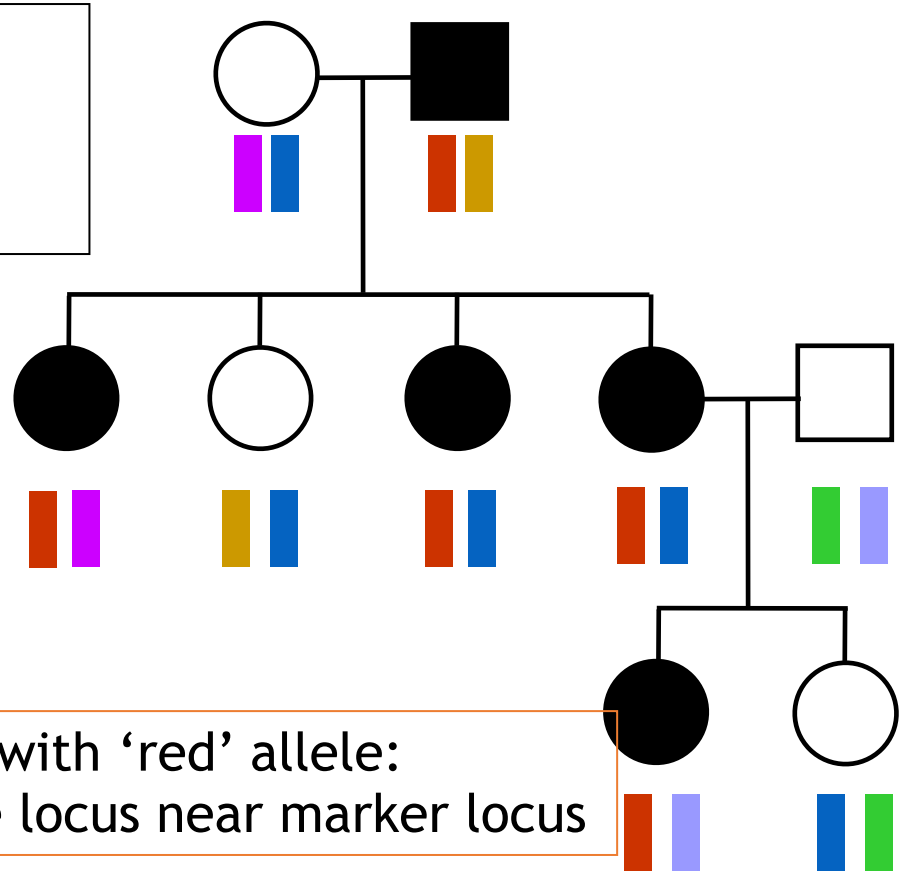
What is Genetic Linkage Analysis

- Type of analysis to locate a gene by seeing which genetic marker segregates with the disease in families
- Tendency of adjacent alleles on the same chromosome to be transmitted together
- Violations of Mendel's Law of Independent Assortment
- Measured in LOD scores for each family
- Cumulative across families
- Add parameter for heterogeneity (HLOD score) across families

NIH
NHGRI

An example of genetic linkage

Is there a genetic marker that co-segregates with the disease?



Disease tracks with ‘red’ allele:
Implies disease locus near marker locus

NIH
NHGRI

Types of Linkage Analysis

- Two main types of linkage – parametric and nonparametric
- Parametric linkage assumes initial parameters such as disease allele frequency, penetrance, and inheritance model (dominant/recessive)
  - Can be two-point (analysis between single variant and phenotype) or multipoint (multiple variants and phenotype)
- Nonparametric linkage has no parameters, calculates linkage based purely on relationship differences between family members (identity by descent)

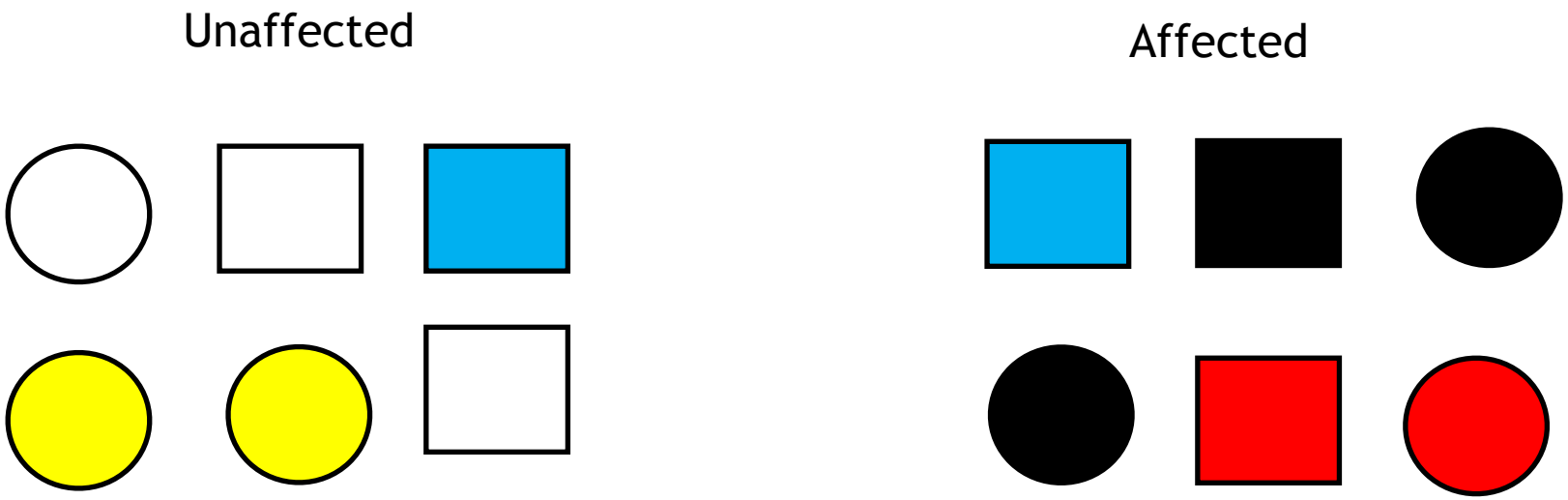
Software for Linkage Analysis

- Linkage analysis programs will require similar data to PLINK TDT
- File with genotype information
- File with subject information including family information
- Also will require
  - Model file – Parametric linkage
  - Frequency file containing MAFs of all variants

- Merlin
  - http://csg.sph.umich.edu/abecasis/merlin/tour/input_files.html
- Morgan
  - https://sites.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml
- GENEHUNTER
  - https://bio.tools/GENEHUNTER

NIH NHGRI

Analysis of Population-based Data with Related Subjects



- How can we perform an association analysis with related individuals?

NIH
NHGRI

EMMAX

- EMMAX is an association analysis program that can handle related data
- Does so by use of a kinship matrix
  - Determines how related each person is to every other person
  - Similar to IBD
- EMMAX requires a VCF file to run, along with a phenotype file

EPACTS

- EMMAX can be run through the EPACTS software
  - https://genome.sph.umich.edu/wiki/EPACTS
- EPACTS can make kinship matrix via the command
  - epacts make-kin --vcf INPUT.vcf --ped INPUT.ped --out KIN.kinf
- Once kinship matrix is made, can run association analysis with:
  - epacts single --vcf INPUT.vcf --ped INPUT.ped --kinf KIN.kinf --pheno PHEN --out OUTPUT
- Works with binary/quantitative phenotypes
- Also can be run as a gene-based test

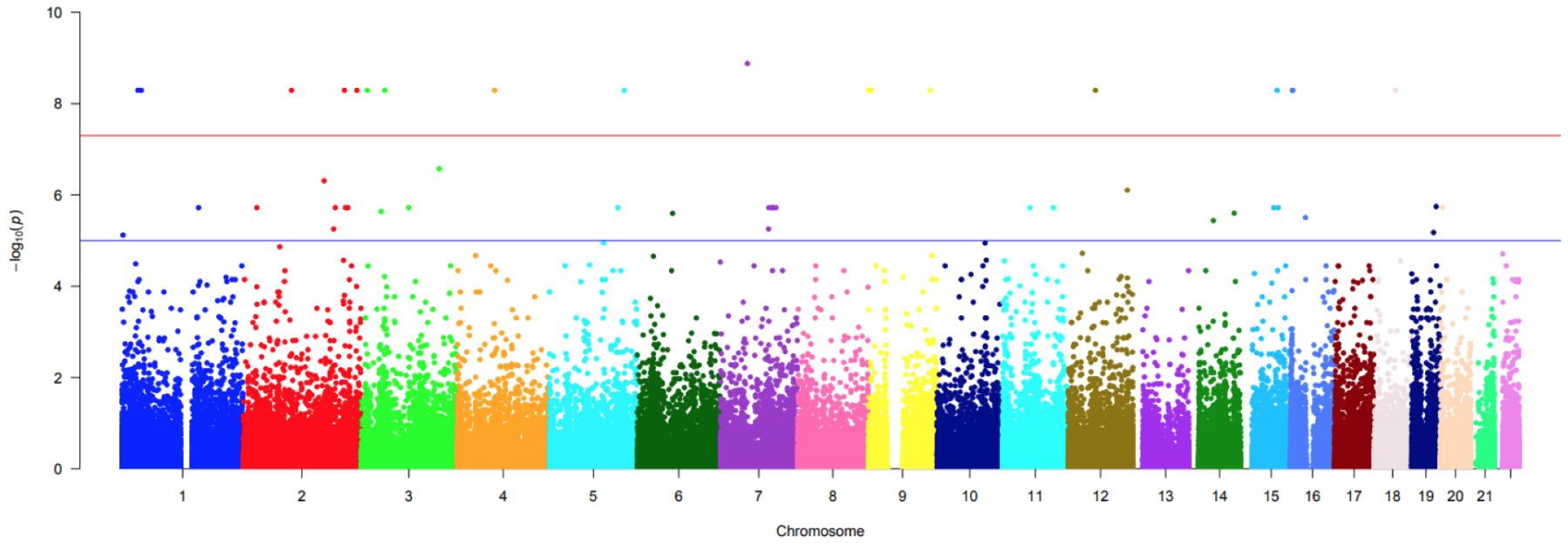
EPACTS Output

- EPACTS output will contain a variety of metrics including p-value

#CHROM	BEG	END	MARKER_ID	NS	AC	CALLRATE	GENOCNT	MAF	STAT	PVALUE	BETA	SEBETA	R2
10	1.06E+08	1.06E+08	10:105765430_A/G_10:105765430	1740	3475	1	0/5/1735	0.00144	4.705	2.74E-06	5.429	1.154	0.01259
10	1.14E+08	1.14E+08	10:113935379_A/G_10:113935379	1740	2111	1	274/821/645	0.39339	4.6773	3.13E-06	0.4298	0.09188	0.01245

NIH
NHGRI

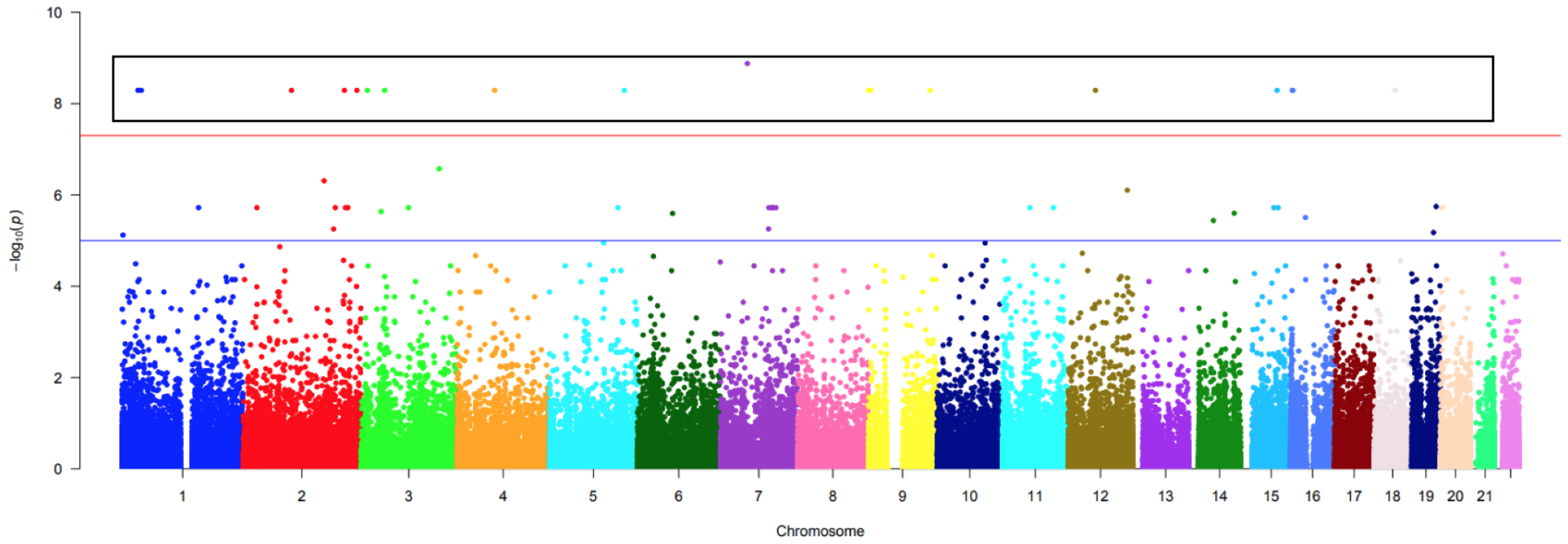
What do I do once my analysis is completed?



- After running analysis, how do I interpret my results?

NIH
NHGRI

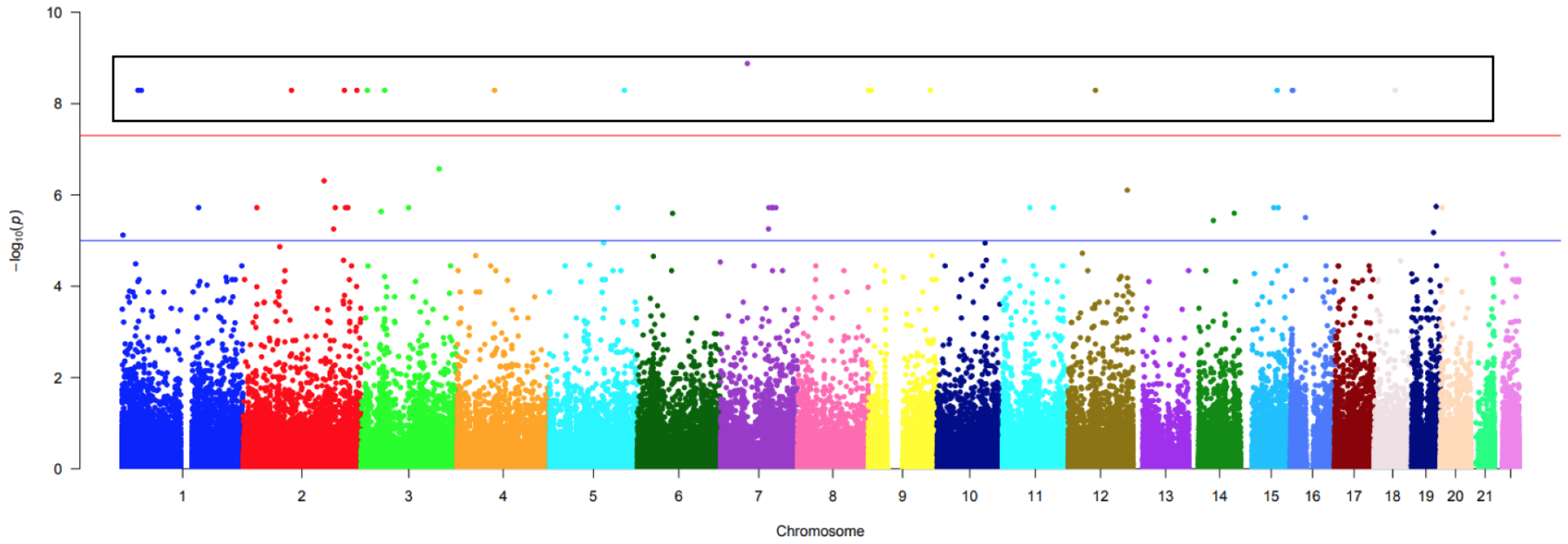
What do I do once my analysis is completed?



- After running analysis, how do I interpret my results?

NIH
NHGRI

What do I do once my analysis is completed?



- After running analysis, how do I interpret my results?
- **Annotation!**

NIH
NHGRI

Annotation

- Annotation allows us to differentiate aspects of variants to determine which variants are the best candidates moving forward
  - Gene location
  - Exonic/intronic/intergenic
  - Minor allele frequency
  - Protein pathogenicity prediction
- Many different annotation programs exist
- wANNOVAR
  - https://wannovar.wglab.org/

NIH
NHGRI

wANNOVAR

- Web-based database
- Collates data from different sources to give extensive annotation on variants
- Takes simple input of:
  - Chromosome
  - Basepair position start
  - Basepair position stop
  - Major allele
  - Minor allele

NIH
NHGRI

wANNOVAR Output

- wANNOVAR will email you a link with your output file
- Will be a very large excel file with *a lot* of information
  - Gene/exonic functions
  - Amino acid change
  - Allele frequency information from 1000Genomes, gnomAD
  - dbSNP info
  - Protein prediction info from SIFT, PolyPhen2, CADD and others

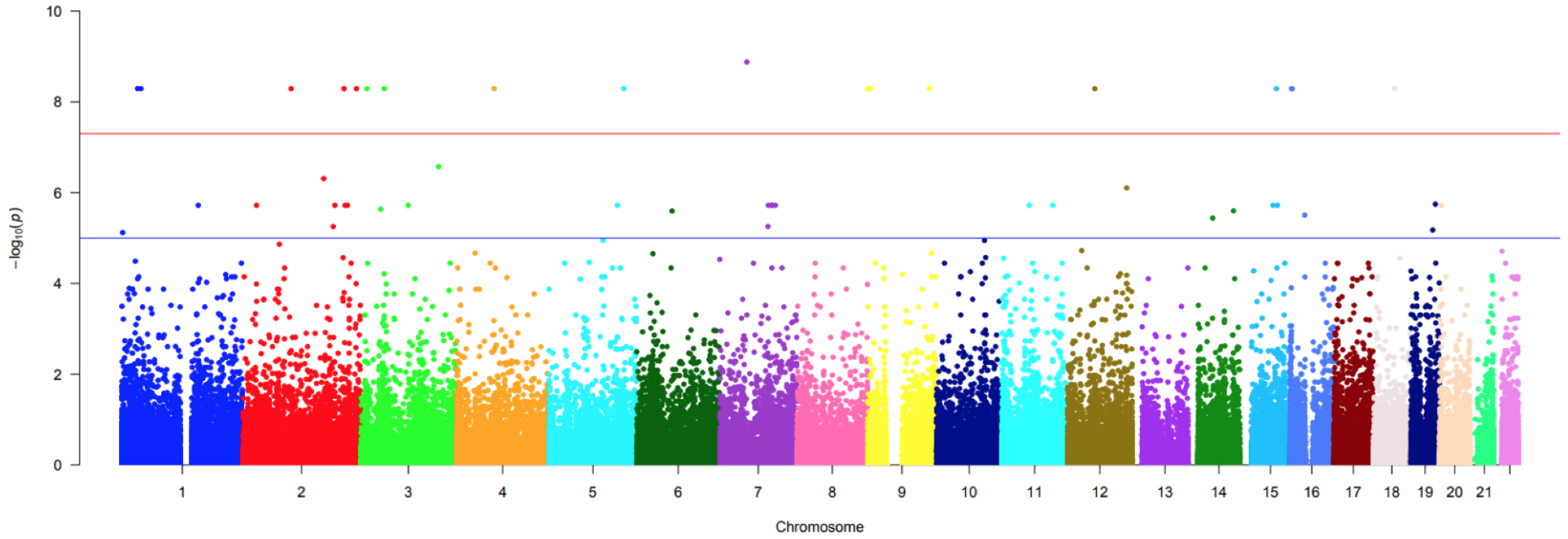
NIH
NHGRI

Annotation on Noncoding Data

- Databases like wANNOVAR give much information for coding variants including protein prediction
- However, do not offer much for noncoding variants, except to say they're noncoding!

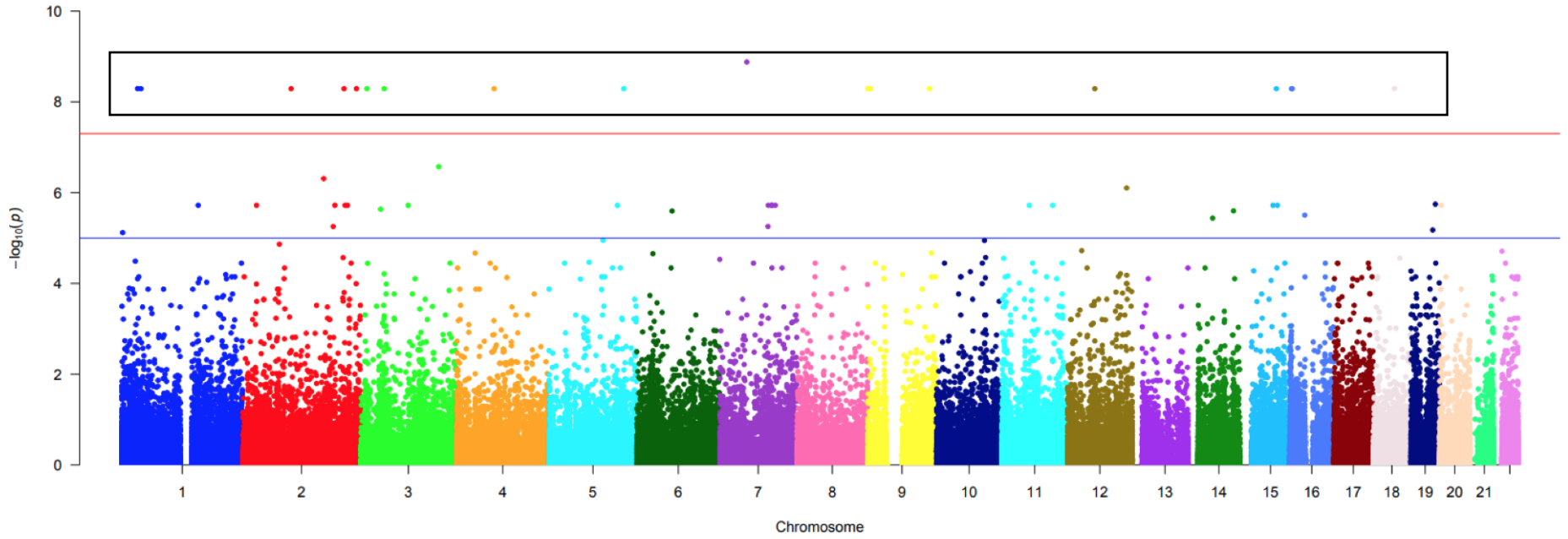
NIH
NHGRI

Prioritization between noncoding variant



NIH
NHGRI

Prioritization between noncoding variant



NIH
NHGRI

Noncoding Annotation Databases

- Not a lot of databases for noncoding variants

- RegulomeDB
  - https://regulomedb.org/regulome-search/
  - Database for transcription factor binding sites (TFBS)
  - Search by BP position or rsID
  - Returns a score (%) of how likely a variant is to be in a TFBS

NIH
NHGRI

Noncoding Annotation Databases

- SPANR
  - http://tools.genes.toronto.edu/
  - Database for determining whether noncoding variant is a splice site
  - Input file requires – chromosome, position, ID, major allele, minor allele
  - Output will tell you whether variant is in known splice site

NIH
NHGRI

Conclusions

- What did we learn?
  - Analysis techniques for:
    - Gene-based TDT
    - Genetic Linkage Analysis for Extended Families
    - Association Analysis with a mix of related and unrelated individuals
  - Annotation techniques with:
    - wANNOVAR
      - Provides large amounts of annotation information
      - Very useful for coding variants
    - RegulomeDB
      - Transcription factor binding sites
    - SPANR
      - Splice sites

NIH
NHGRI

Comments/Questions

- Feel free to email me any questions at:
  - musolfam@mail.nih.gov

NIH
NHGRI