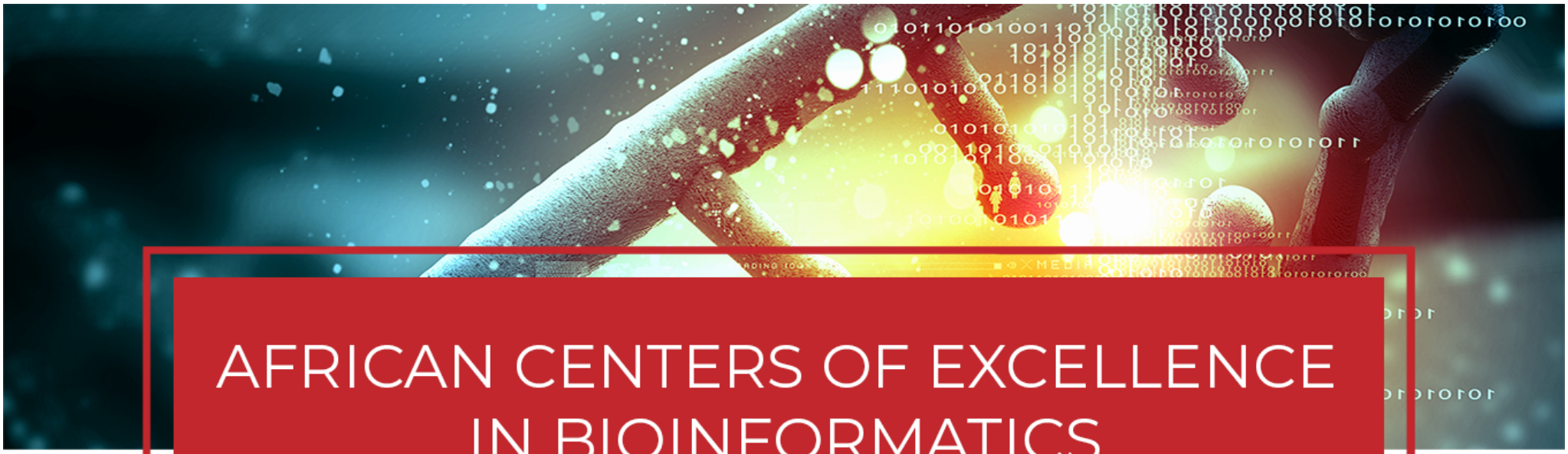


Haplotype-based Association Tests, Linkage Analysis for Binary Traits and Quantitative Traits

Today's Instructor

Andrew Oler, PhD

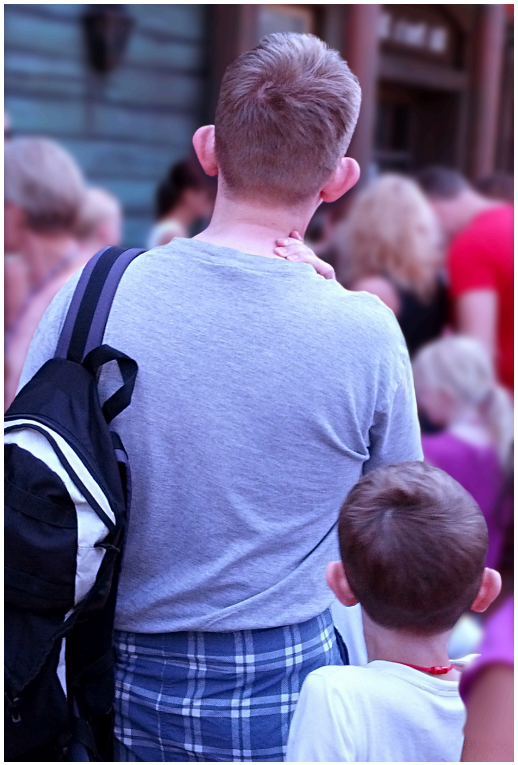
Senior Bionformatics Specialist

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: ace@icermali.org
  - Listserv: ACE-MALI-L@LIST.NIH.GOV
  - Instructors:
    - andrew.oler@nih.gov

Topics

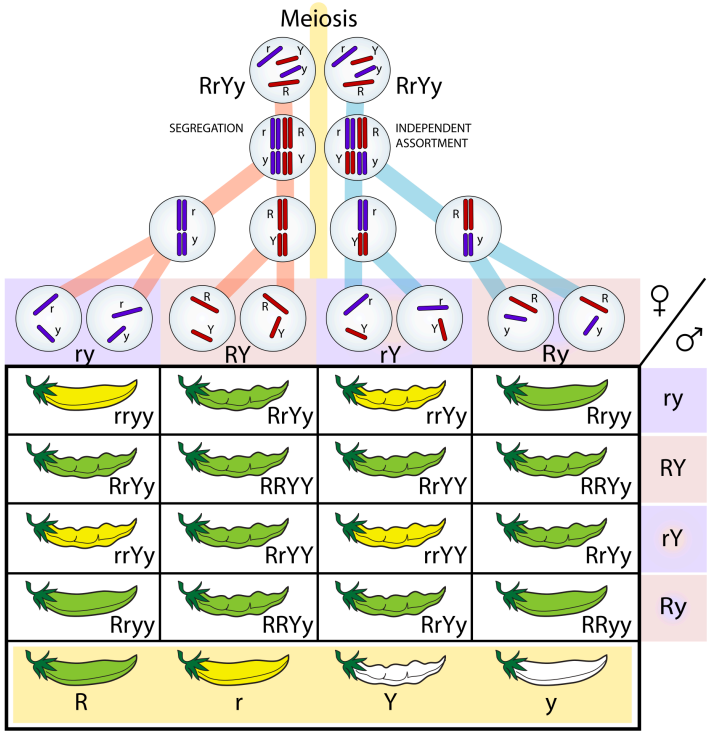
- Mendelian inheritance and Pedigree analysis
  - Inheritance models
- Mapping traits
  - Consanguinity
  - Linkage analysis
    - Parametric
    - Nonparametric
- Hands-on exercises
  - Drawing pedigrees
  - Linkage power calculation
  - Linkage analysis with paramlink, MERLIN

Inheritance of Traits



https://en.wikipedia.org/wiki/Heredity

Principles of Heredity

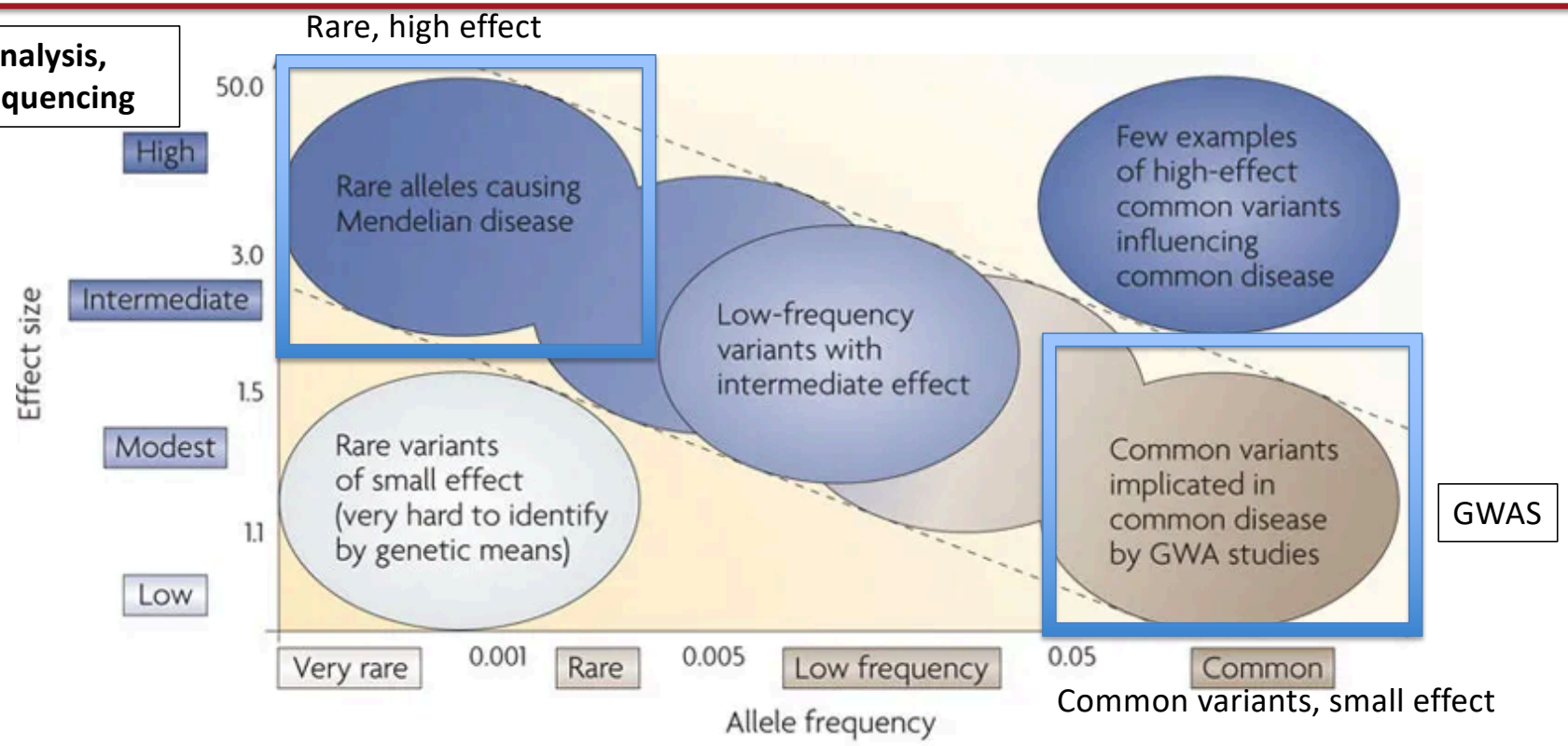


- Traits are associated with particular genetic variants
- Key aspects of heredity
  - Independent Assortment
  - Segregation of Chromosomes
  - Chromosomal crossover/recombination
  - Meiosis

https://en.wikipedia.org/wiki/Heredity

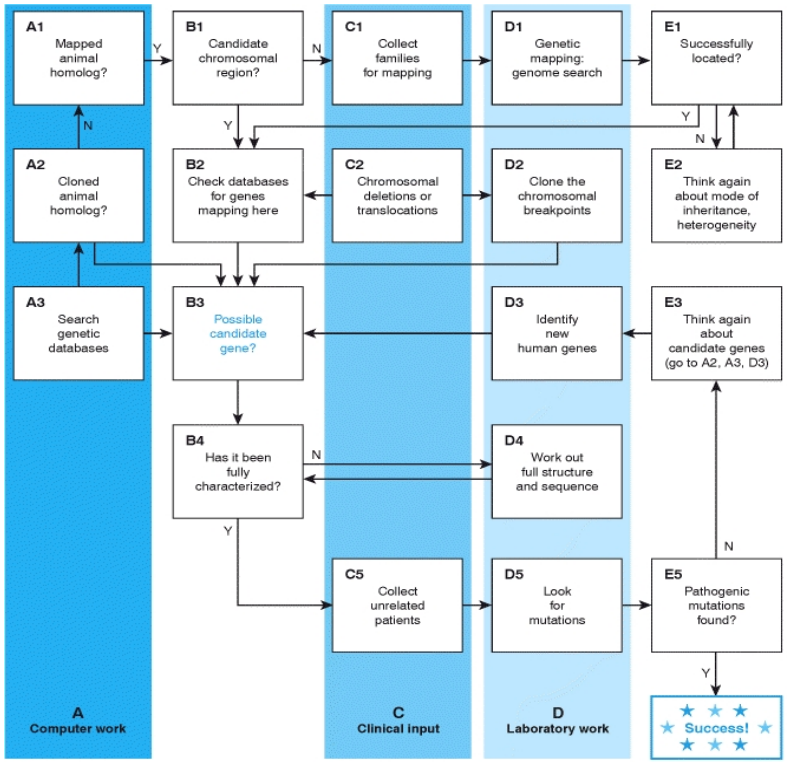
Mendelian Disorders vs. Multifactorial Traits

Linkage analysis, exome sequencing



Antonarakis, S.E., *Nature Reviews Genetics, 2010; 11*, 380-384.

“How to identify a human disease gene”



Human Molecular Genetics. 2nd edition. (1999) http://www.ncbi.nlm.nih.gov/books/NBK7561/

Strategies for identifying human disease genes

- Genome-wide Association study (GWAS)
  - thousands of unrelated individuals – common diseases, case/control
  - SNP array markers
- **Linkage study** (small number of related individuals, better for rare disease)
  - Traditional linkage analysis
    - SNP array markers to genotype individuals in affected families to maximize number of "meiosis"
    - Parametric linkage, Nonparametric linkage
  - Exome sequencing
    - Sequence genomes of related individuals, e.g., trio or multiple family members.
    - Map to reference, identify and annotate variants (coding, splicing, benign, etc.)
    - Filter for rare alleles (e.g., < 1% in population)
  - Apply genetic model to identify variants associated with the disease
    - In traditional linkage analysis, the marker SNP is *linked* to but not responsible for the disorder
    - In exome sequencing, the variant identified is likely *causal* for the disorder

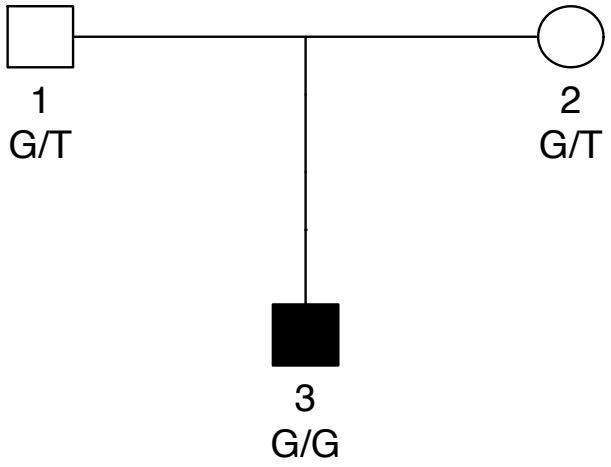
http://www.nature.com/scitable/topicpage/mendelian-genetics-patterns-of-inheritance-and-single-966

Genetic Inheritance Models

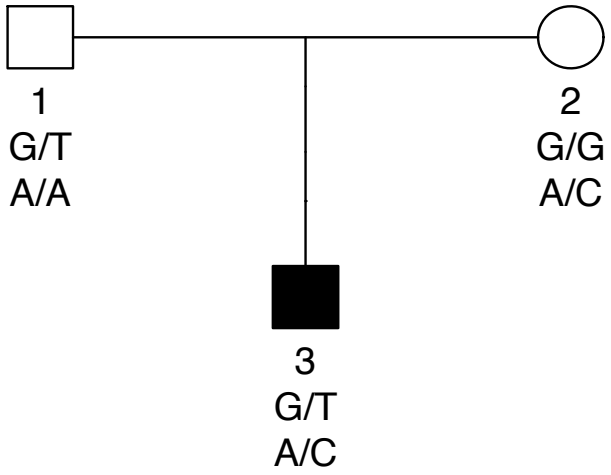
- Autosomal recessive single-gene
- Autosomal dominant single-gene
- X-linked recessive single-gene
- X-linked dominant single-gene
- Compound heterozygote
- De novo
- Other
  - Multi-gene “oligogenic” (modifier genes)
  - Dominant, variable expressivity

Autosomal Recessive

Homozygous recessive



Compound Heterozygous

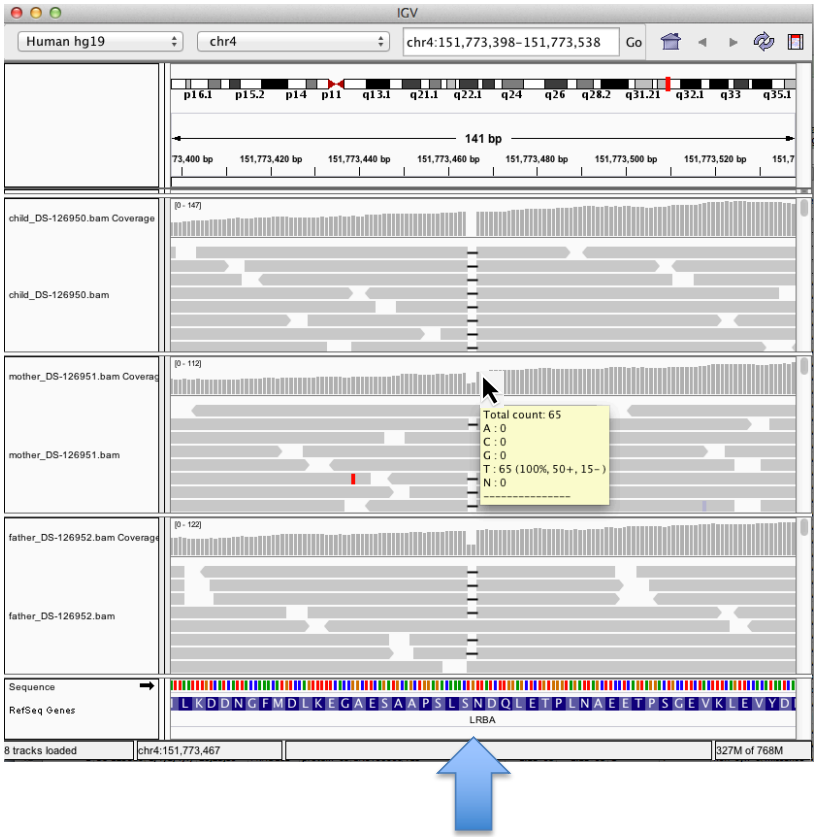


Verification of Variant and Transmission in Raw Data with IGV

Proband

Mother

Father

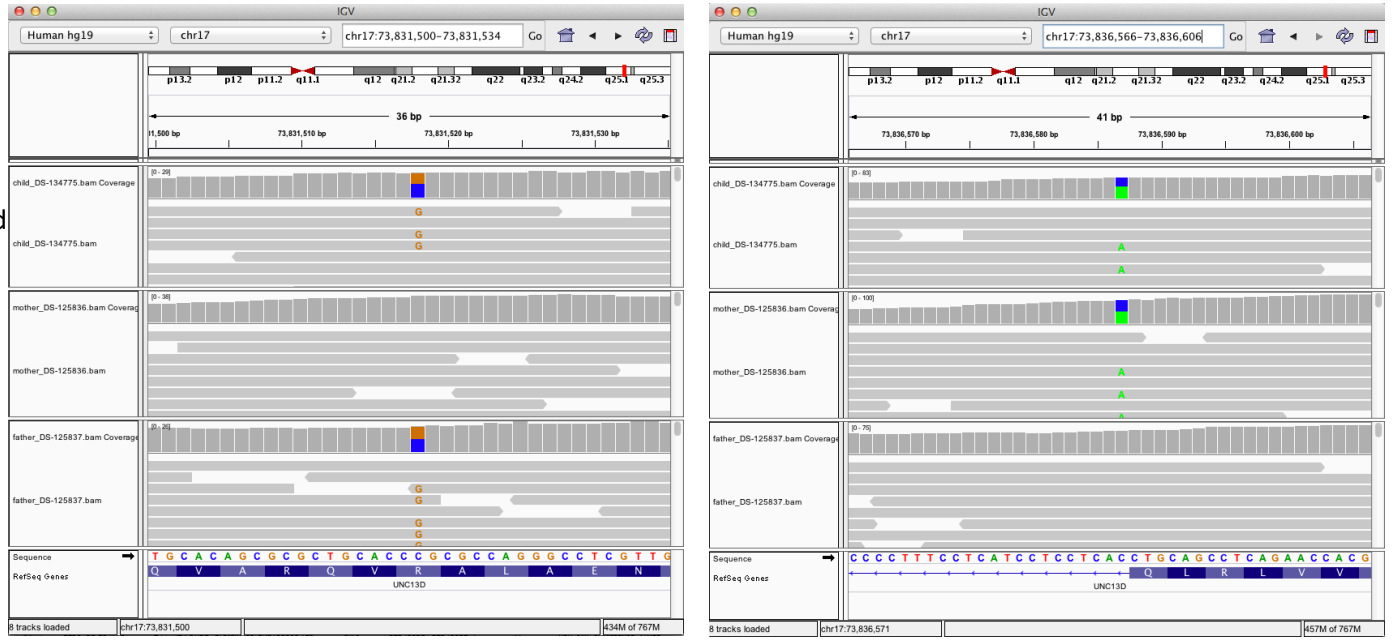


Verification of Variant and Transmission in Raw Data with IGV

Proband

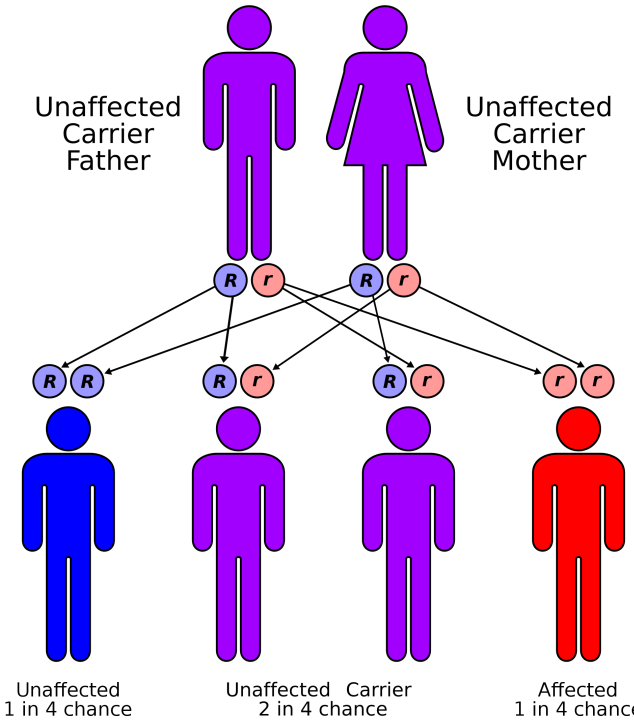
Mother

Father



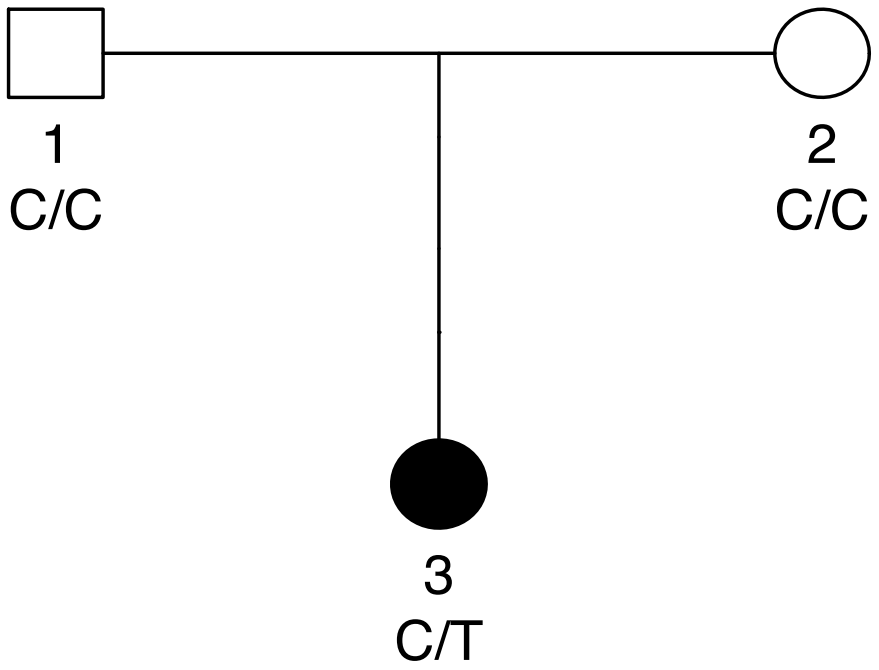
Compound Heterozygous

Sickle Cell trait is Recessive



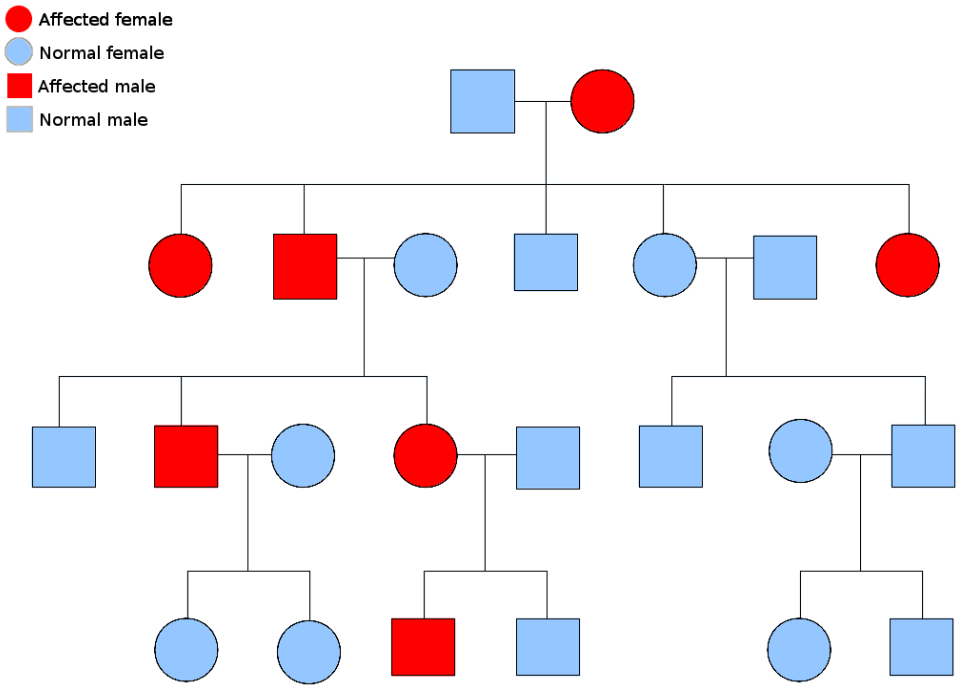
https://en.wikipedia.org/wiki/Mendelian_traits_in_humans

De novo Inheritance



e.g., STAT3, hyper-IgE syndrome

Autosomal Dominant



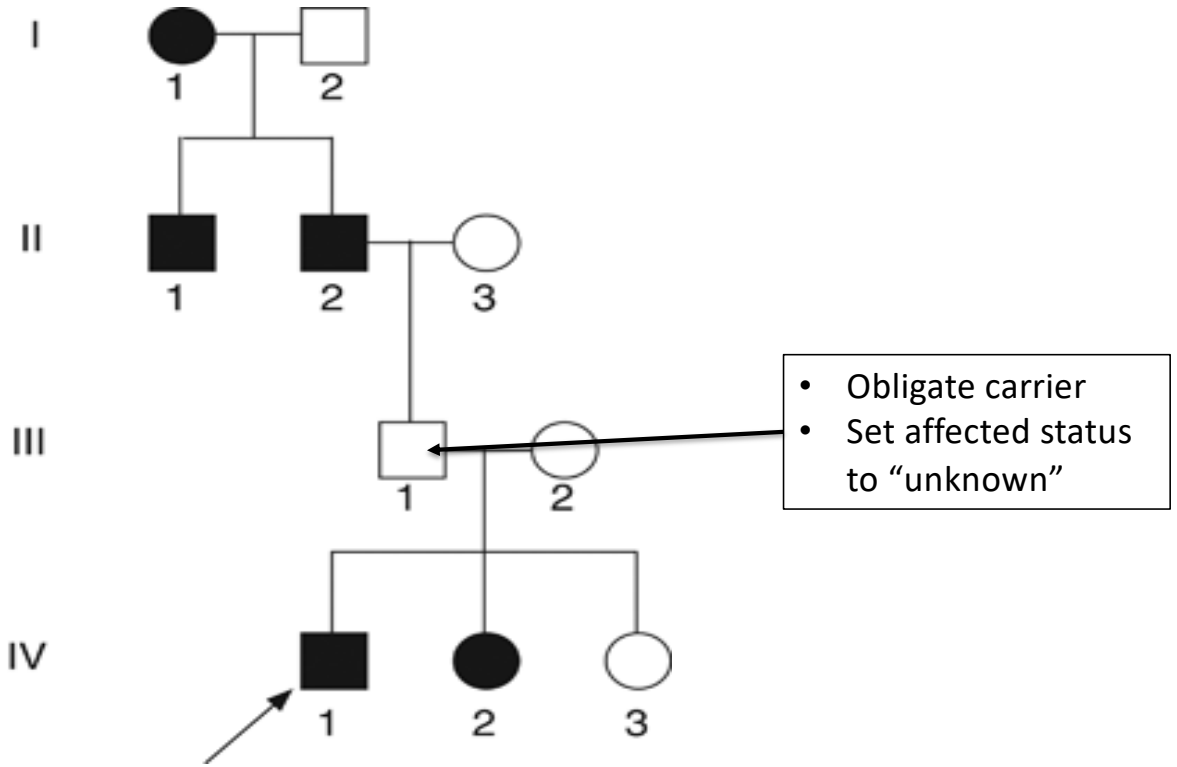
e.g., STAT3, hyper-IgE syndrome

https://en.wikipedia.org/wiki/Heredity

Autosomal Dominant Incomplete Penetrance

Penetrance:

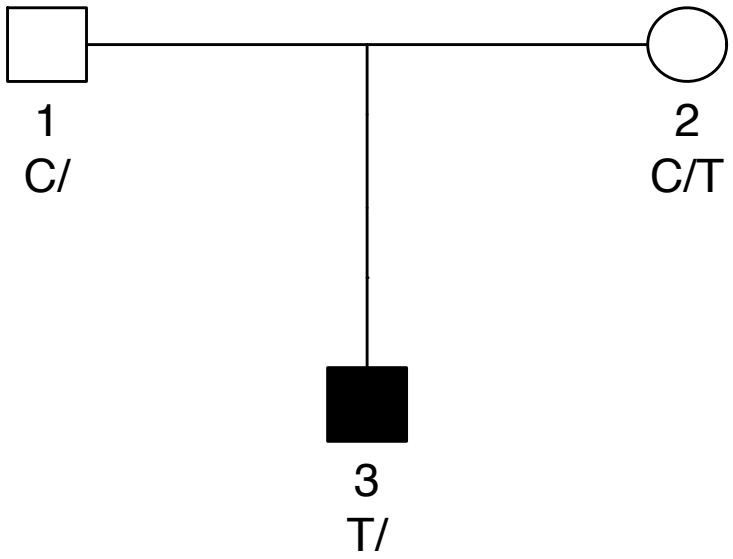
- Percentage of genotype-positive individuals with a clinical manifestation of the disorder
- Complete penetrance = 100%
- CTLA4 penetrance = ~67%
- “Incomplete penetrance” aka “Variable expressivity”



e.g., CTLA4

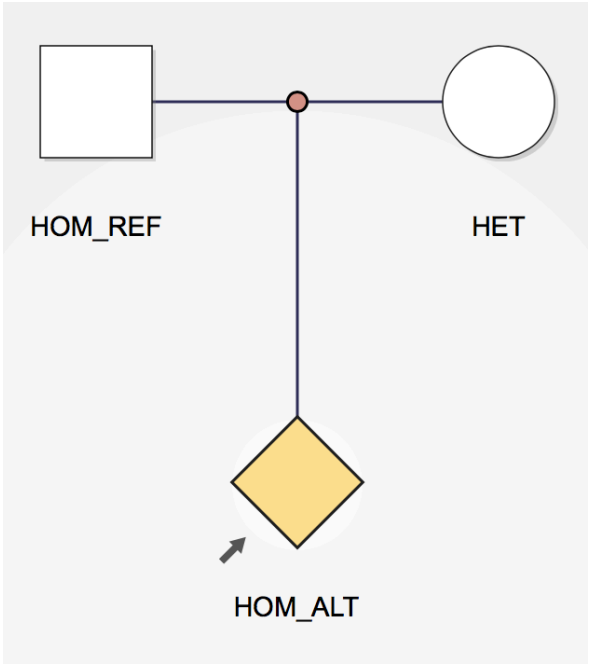
X-linked Recessive Inheritance

Actual Genotypes

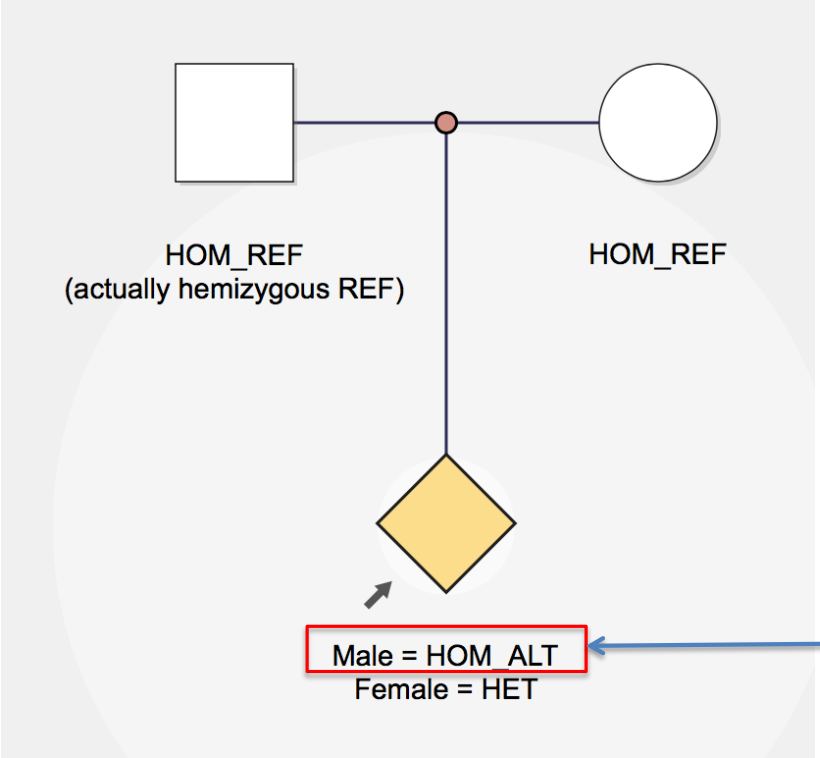


“hemizygous”

How it looks in Exome Sequencing

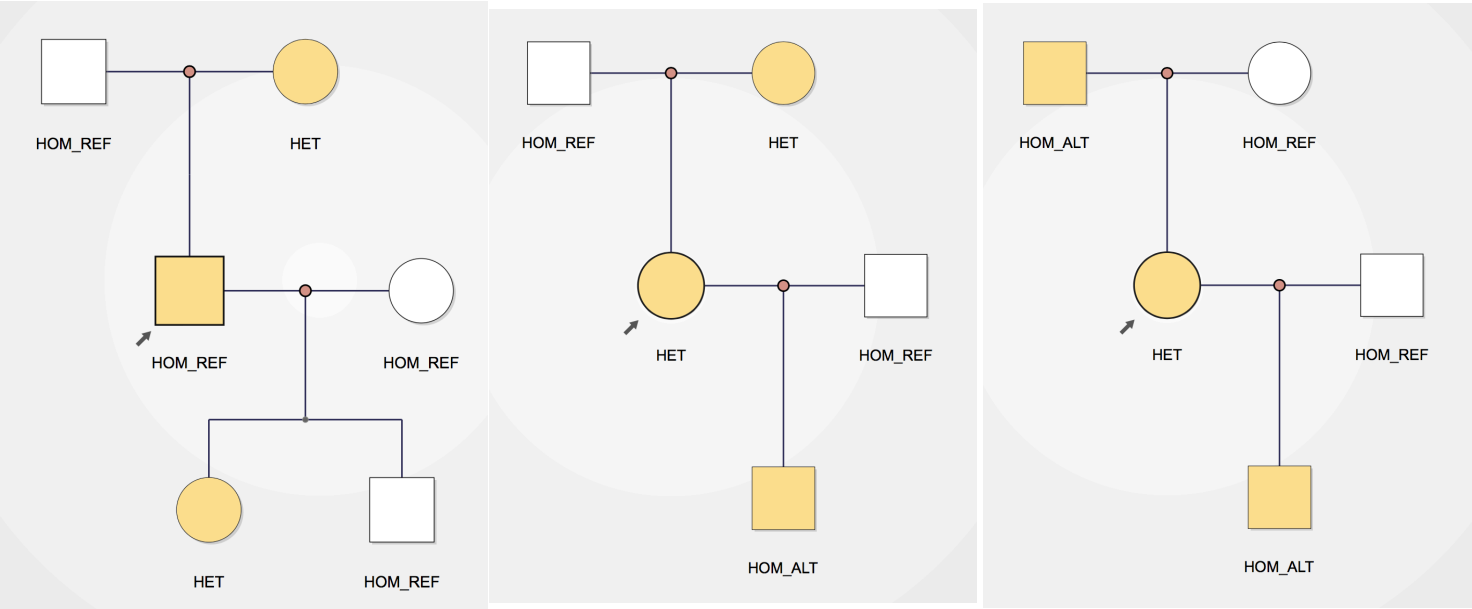


X-linked *de novo*



X-linked dominant

- Male proband
- Female proband (mat)
- Female proband (pat)



Difference from autosomal dominant:

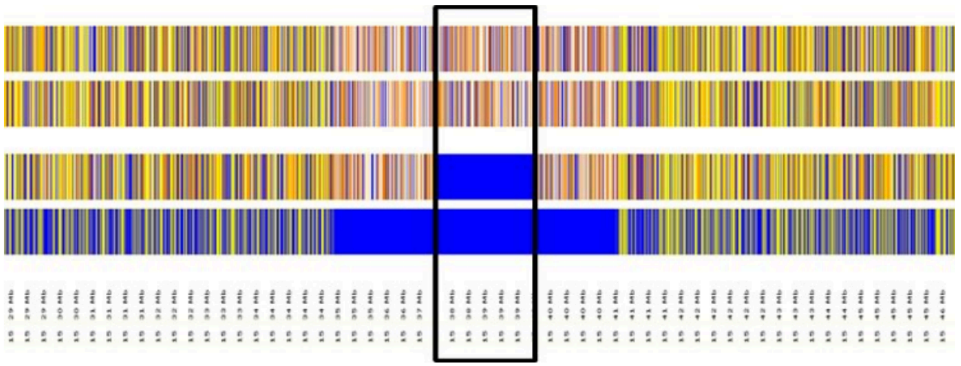
- Subset of dominant model, with specific rules about genotypes, gender, and transmission from parent to sons and daughters
- chrX

Mapping Genetic Traits

- Consanguinity
  - Homozygosity mapping
- Linkage analysis
  - Parametric vs. Nonparametric
  - Single-point vs. multi-point

Homozygosity Mapping

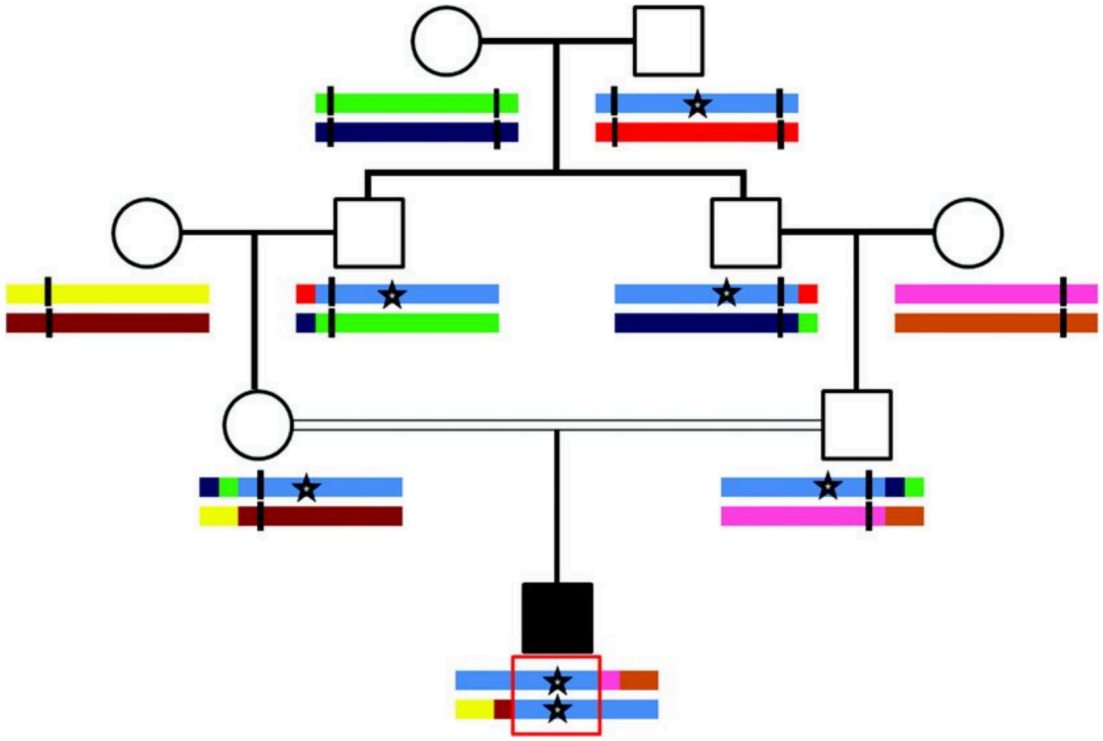
- Identifying regions that appear homozygous, or lack heterozygosity
- “Run of Homozygosity (ROH)”
- Region with “absence of heterozygosity (AOH)”
- AOH regions may signal
  - Identity by descent
    - Suggesting autosomal recessive
  - Uniparental isodisomy
  - Hemizygous copy number variant (Deletion; compare to CNV/SV calls)



HomSI tool – good GUI for mapping homozygosity

http://www.igbam.bilgem.tubitak.gov.tr/en/bioinformatics.html

Homozygosity could be autozygosity



https://www.nature.com/articles/gim2010128/figures/1

Determining degree of inbreeding/consanguinity

- Determine regions of homozygosity > 3Mb
- Add up the total length
- Divide by total autosomal length (~2881Mb) = **% identity by descent (IBD)**

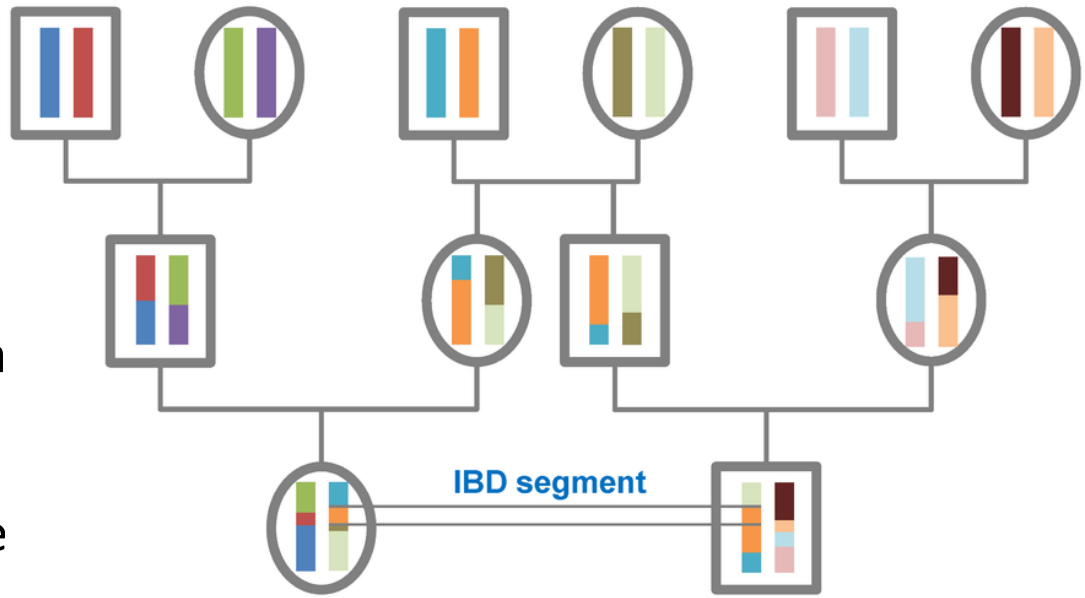
Consanguinity Degree	Theoretic Percentage	Percentage of Homozygosity (Confidence Interval)
First or closer	> 25%	> 28.7%
First	25%	21.3–28.7%
First or second		15.3–21.3%
Second	12.5%	9.7–15.3%
Second or third		8.3–9.7%
Third	6.25%	4.6–8.3%
Third or fourth		4.2–4.6%
Fourth	3.125%	2.6–4.2%
Fourth or fifth		1.6–2.6%
Fifth	1.5625%	0.5–1.6%

https://www.nature.com/articles/gim2012169

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5797403/

Identity by Descent

- "A DNA segment is identical by state (IBS) in two or more individuals if they have identical nucleotide sequences in this segment. An IBS segment is **identical by descent** (**IBD**) … in two or more individuals if they ha inherited it from a common ancestor without recombination, that is, the segment has the same ancestral origin in these individuals."



https://en.wikipedia.org/wiki/Identity_by_descent

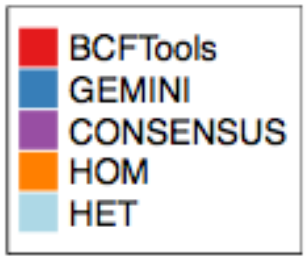
How to Determine Regions of AOH

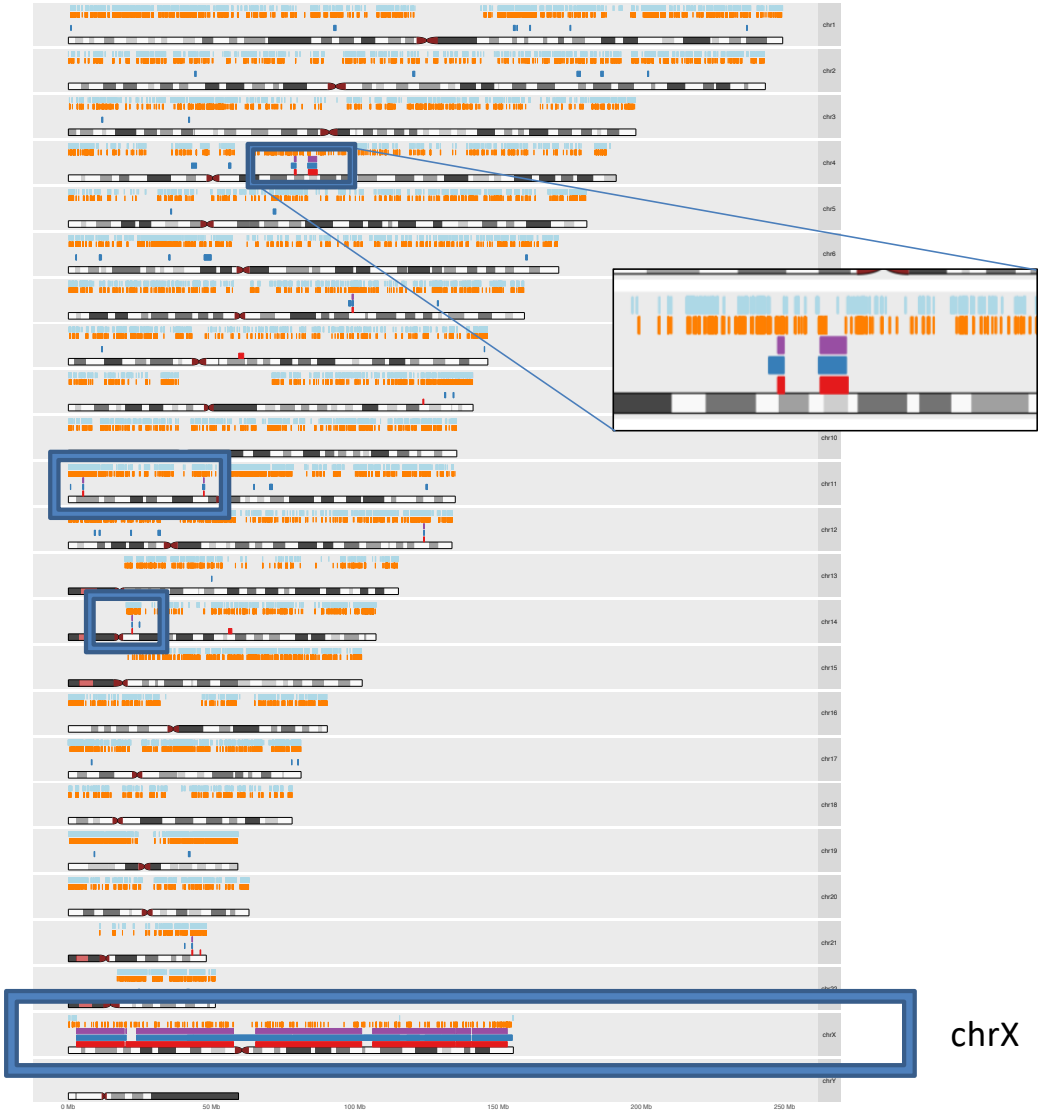
- Chromosomal microarray
- Exome -> VCF
  - PLINK
  - BCFTools
  - GEMINI
  - Others…
- Choose a few methods and make a consensus (e.g., found in at least 2 methods)
- Generally, window of a certain length with many homozygous variants and no heterozygous variants; only use high quality variants

Example Non-consanguineous Male child

- Typically, only a few small runs of homozygosity < 3Mb
- chrX is hemizygous in males so appears to be a run of homozygosity

PLINK = green (less sensitive so absent here)

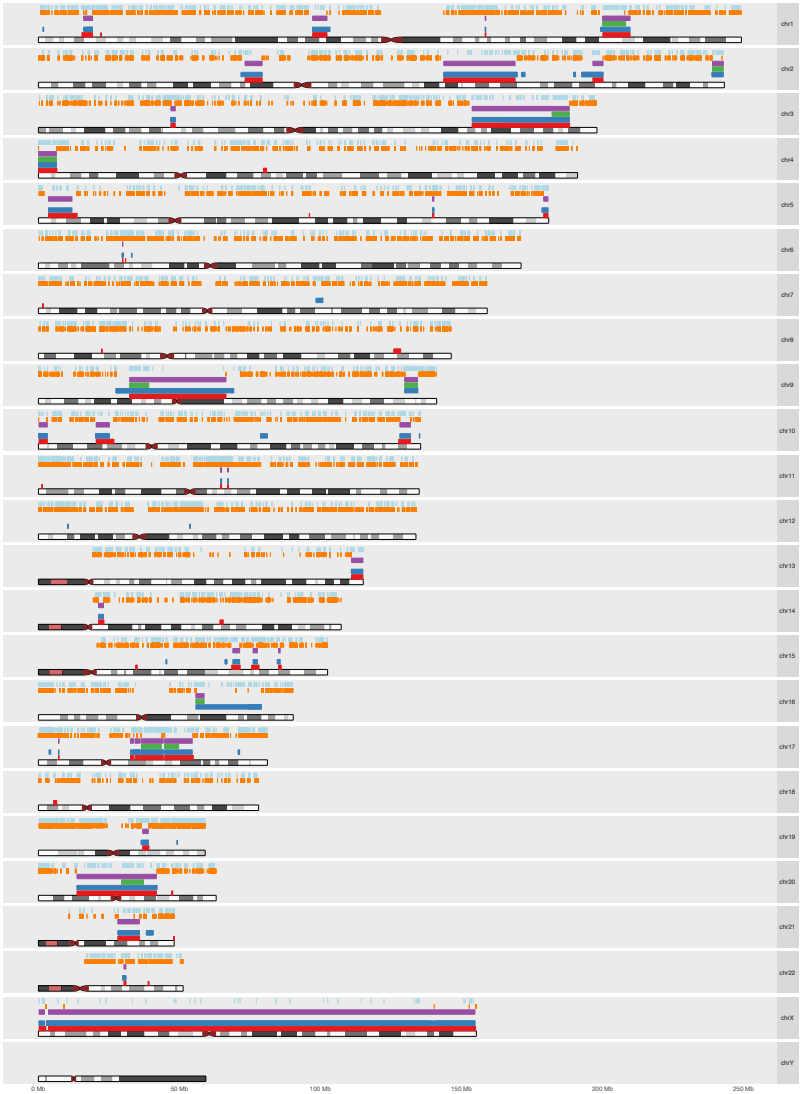




chrX

Example Consanguineous Male child

- 7.129% IBD
- 18 regions of AOH > 3Mb
- Largest is 34Mb (chr3)

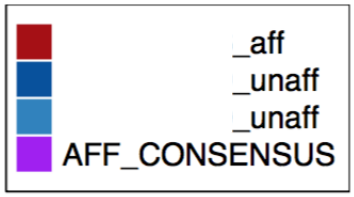


Additional unaffected siblings

5.592% IBD

1.567% IBD

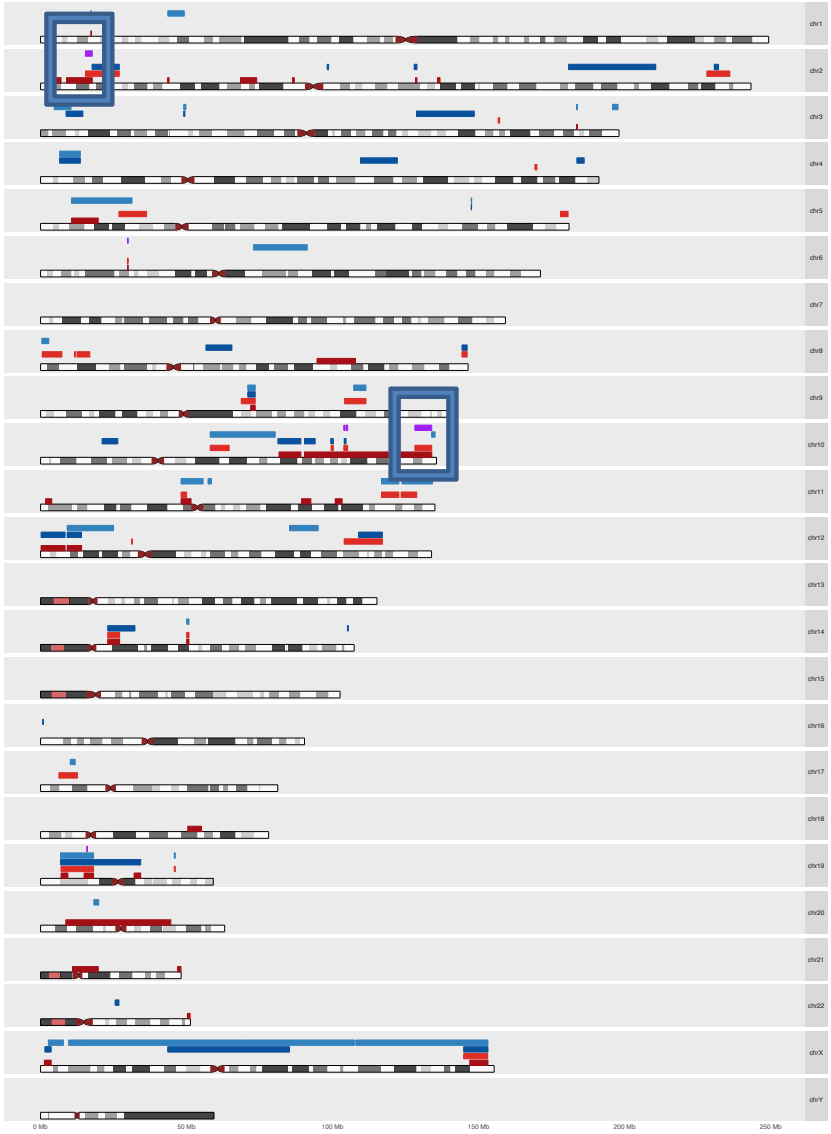
All children together narrows down region of interest



- Shades of Red for Affected
- Shades of Blue for Unaffected
- Purple for Candidate Consensus Region (i.e., Region of AOH in all affected and no unaffected)
- **Identify genes in region of interest that could be related to the proband phenotype**
- **Low resolution** -- often hundreds or thousands of genes

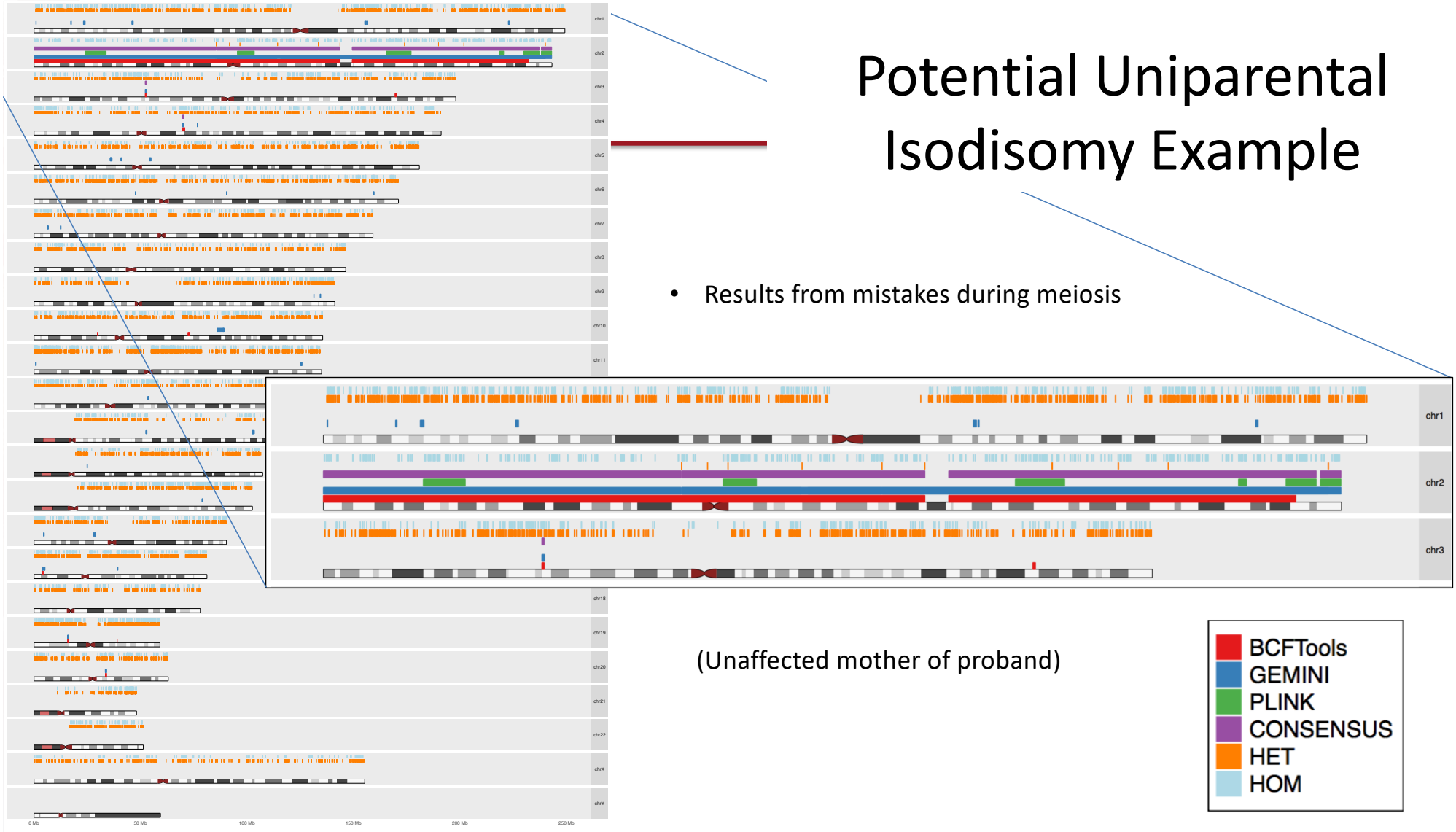
Another example of how helpful additional siblings can be

- 3.8-5.8% IBD in children
- 7-16 regions > 3Mb



Potential Uniparental Isodisomy Example

- Results from mistakes during meiosis



(Unaffected mother of proband)

Genetic Linkage Analysis

- History and theory
- Linkage analysis methods
- LOD scores
- Linkage analysis tools
- Demo/hands-on exercises

Exceptions to Mendel's Law of Independent Assortment

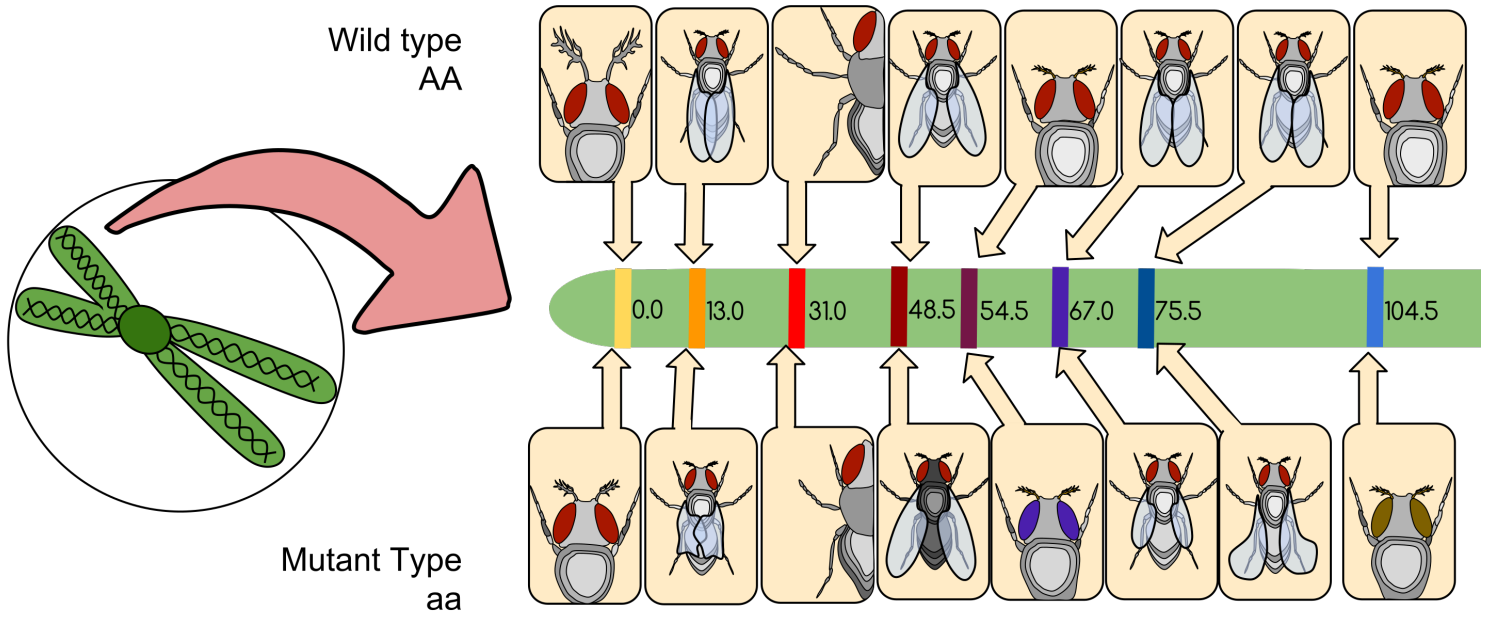
Bateson, Saunders, and Punnett experiment

Phenotype and genotype	Observed	Expected from 9:3:3:1 ratio	
Purple, long (*P_L_*)	284	216	Higher than expected
Purple, round (*P_ll*)	21	72	Lower than expected
Red, long (*ppL_*)	21	72	Lower than expected
Red, round (*ppll*)	55	24	Higher than expected

- flower colour (*P*, purple, and *p*, red) and the gene affecting the shape of pollen grains (*L*, long, and *l*, round)
- The traits were related because of their proximity on the same chromosome = linked

https://en.wikipedia.org/wiki/Genetic_linkage

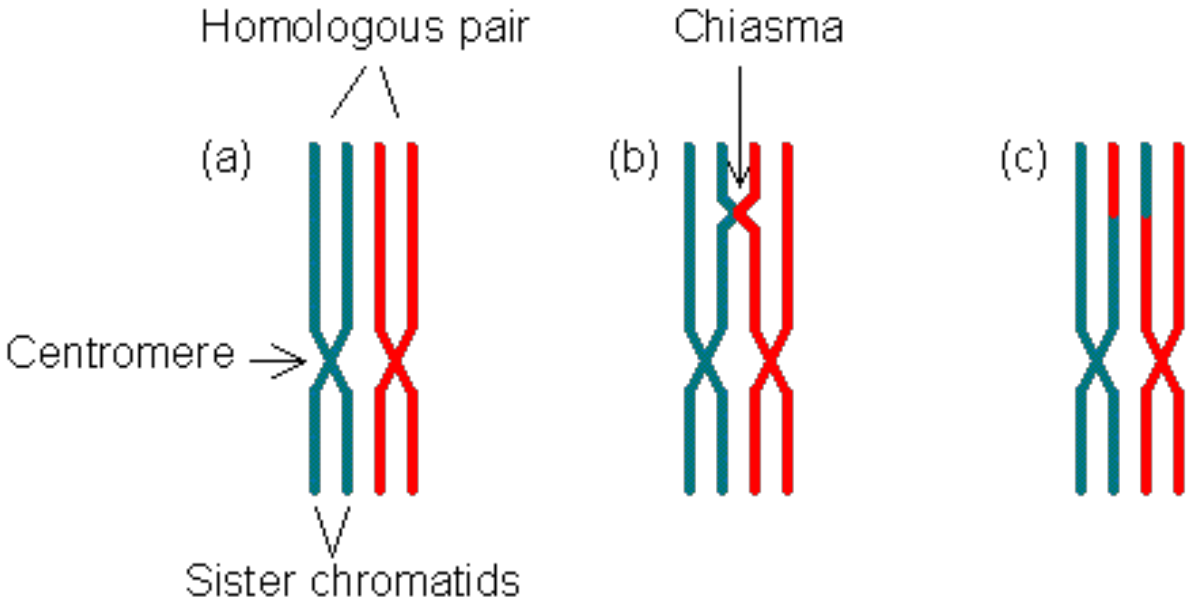
Genetic Linkage



- “Linkage” refers to genes being in proximity on chromosomes.
- The further apart two genes are, the greater chance for recombination, which promotes independent assortment
- Thomas Hunt Morgan mapped distances between fly genes => centimorgans (cM) is a measure of distance
- 1 cM ~ 1Mb

https://en.wikipedia.org/wiki/Genetic_linkage

Recombination



http://www.web-books.com/MoBio/Free/Ch8D1.htm

Linkage Analysis

- In linkage analysis, we use various markers spread throughout the genome and testing each for linkage with a disease trait
  - i.e., does any marker occur more frequently *with* the disease than expected by chance
- Binary traits => affected, unaffected
  - Any marker(s) occur more frequently with the ***affected*** state?
  - e.g., cystic fibrosis (*CFTR*), Huntington disease (*HTT*), breast cancer (*BRCA1*, *BRCA2*)
- Quantitative traits => e.g., numeric values such as height
  - Which markers are associated with ***higher*** values of the trait?

https://en.wikipedia.org/wiki/Genetic_linkage

Linkage Analysis History

- Linkage analysis became possible in the 1990s when sufficient multi-allelic markers were mapped across chromosomes (prior to human genome sequence)
- In the 2000s with the sequencing of many human genomes (1KG), we were able to generate dense single nucleotide polymorphisms for use in linkage
- Linkage analysis gives **low resolution** results – usually many genes lie in the linkage region – but it limits the number of genes required for follow-up
- Linkage analysis was largely overshadowed by GWAS in the 2010s after 1KG and HapMap projects
- GWAS is more powerful for association studies, targeted to identify common variants with small effect
- In addition, exome and genome sequencing have largely been the tool of choice recently to discover causal variants in rare disorders
- However, linkage analysis may become more popular as we realize that rare variants likely responsible for a large proportion of complex diseases
- In large families, linkage analysis can also complement traditional variant filtering approaches used in exome/genome sequencing studies

Cantor, 2013; doi:10.1016/b978-0-12-383834-6.00010-0

Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet*. 2015;16(5):275–284. doi:10.1038/nrg3908

Bailey-Wilson JE, Wilson AF. Linkage analysis in the next-generation sequencing era. *Hum Hered*. 2011;72(4):228–236. doi:10.1159/000334381

Linkage Analysis Methods

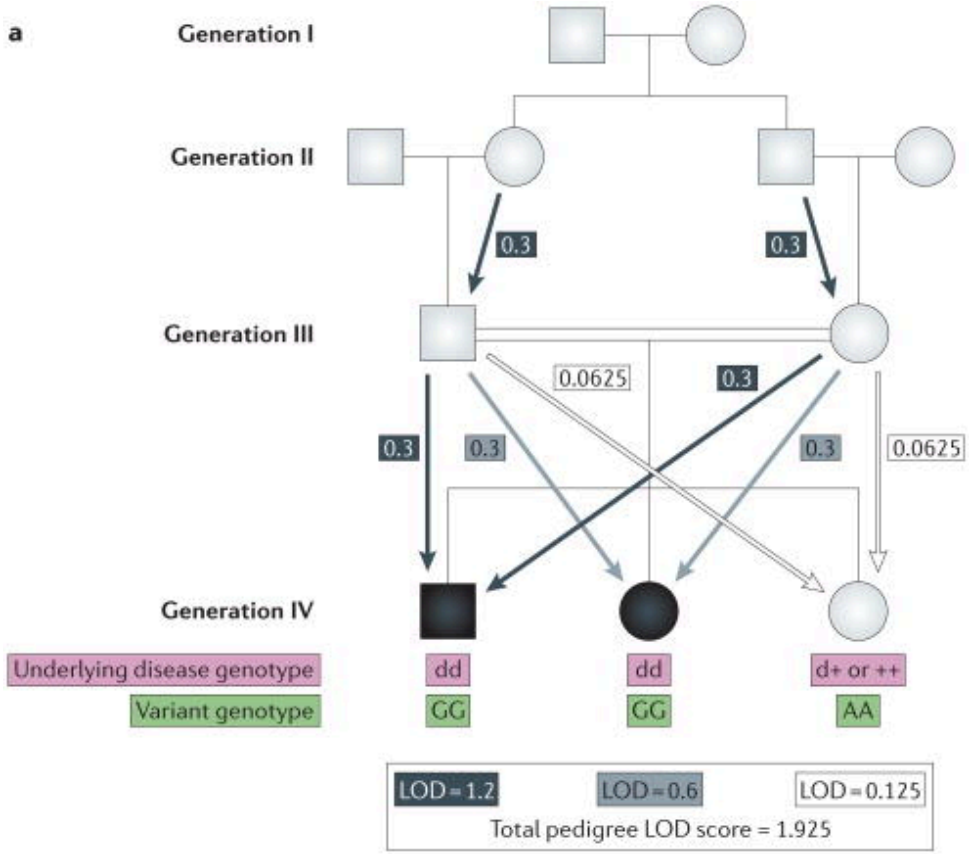
- Parametric linkage analysis assumes a specific genetic model (e.g., autosomal recessive, etc.), disease prevalence, and penetrance level
  - More powerful than nonparametric if assumed genetic model is correct
- Nonparametric (model-free) linkage analysis assumes no model of inheritance
- Two-point linkage analysis uses a single marker
- Multipoint linkage analysis (most common) uses multiple markers
- LOD score is a statistical test for linkage, comparing the likelihood of obtaining the test data if the marker and disease are linked to observing the same data by chance. log10 of the likelihood ratio
- Positive LOD scores suggest linkage, with 3 or 3.3 being the standard for statistical significance (3 ~ 1000:1 odds in favor of linkage)

Ott J, Wang J, Leal SM. *Nat Rev Genet*. 2015;16(5):275–284. doi:10.1038/nrg3908

https://en.wikipedia.org/wiki/Genetic_linkage

LOD score calculation

- General rules for Max LOD in nuclear pedigree
  - dominant: 0.3(c − 1)
    - c, genotyped children
  - aut. recessive: 0.6(a − 1) + 0.125n
    - a, affected
    - n, unaffected
  - X-linked recessive: 0.3(b − 1)
    - b, boys
- Use **linkage.power()** method in R package “paramlink” to estimate



http://folk.uio.no/magnusv/LinkageCourse/Paramlink/paramlink_power.pdf

Ott J, Wang J, Leal SM. *Nat Rev Genet*. 2015;16(5):275–284. doi:10.1038/nrg3908

Linkage Analysis tools

- MERLIN - Multipoint linkage analysis, parametric and nonparametric (Abecasis, 2002) (multiple families; binary and quantitative). http://csg.sph.umich.edu/abecasis/Merlin/index.html
- SEQLinkage – Collapsed Haplotype-based linkage, multipoint parametric linkage (Leal, 2015) (multiple families)
- RV-NPL – Rare variant nonparametric linkage analysis (Leal, 2019)
- Paramlink – R package for working with pedigrees and two-point parametric analysis (single family)
- Pre-processing for Exome/Genome sequencing -> linkage
  - vcftools, bcftools
  - plink
- Others (SimWalk2, GENEHUNTER, etc.) – see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4440411/ for more tools

Ott J, Wang J, Leal SM. *Nat Rev Genet*. 2015;16(5):275–284. doi:10.1038/nrg3908

Demo/Hands-on exercises

- Paramlink (Rstudio)
  - Drawing pedigrees
  - Estimating Max theoretical LOD scores
  - Two-point Linkage analysis for single family (dominant)
  - Multipoint linkage analysis for single family (dominant) using MERLIN

Autosomal Dominant Incomplete Penetrance

Rules:

- Convert all “unaffected” to “unknown” and re-run “gemini autosomal_dominant”
- All affected are HET
- No one is HOM_ALT
- For each affected:
  - If affected has two parents, one is HET, one is homozygous reference (HOM_REF)
  - If grandparents via HET parent, one should be HET and one HOM_REF
  - If grandparents via HOM_REF parent, both should be HOM_REF