Analysis of Next Generation Sequence Data

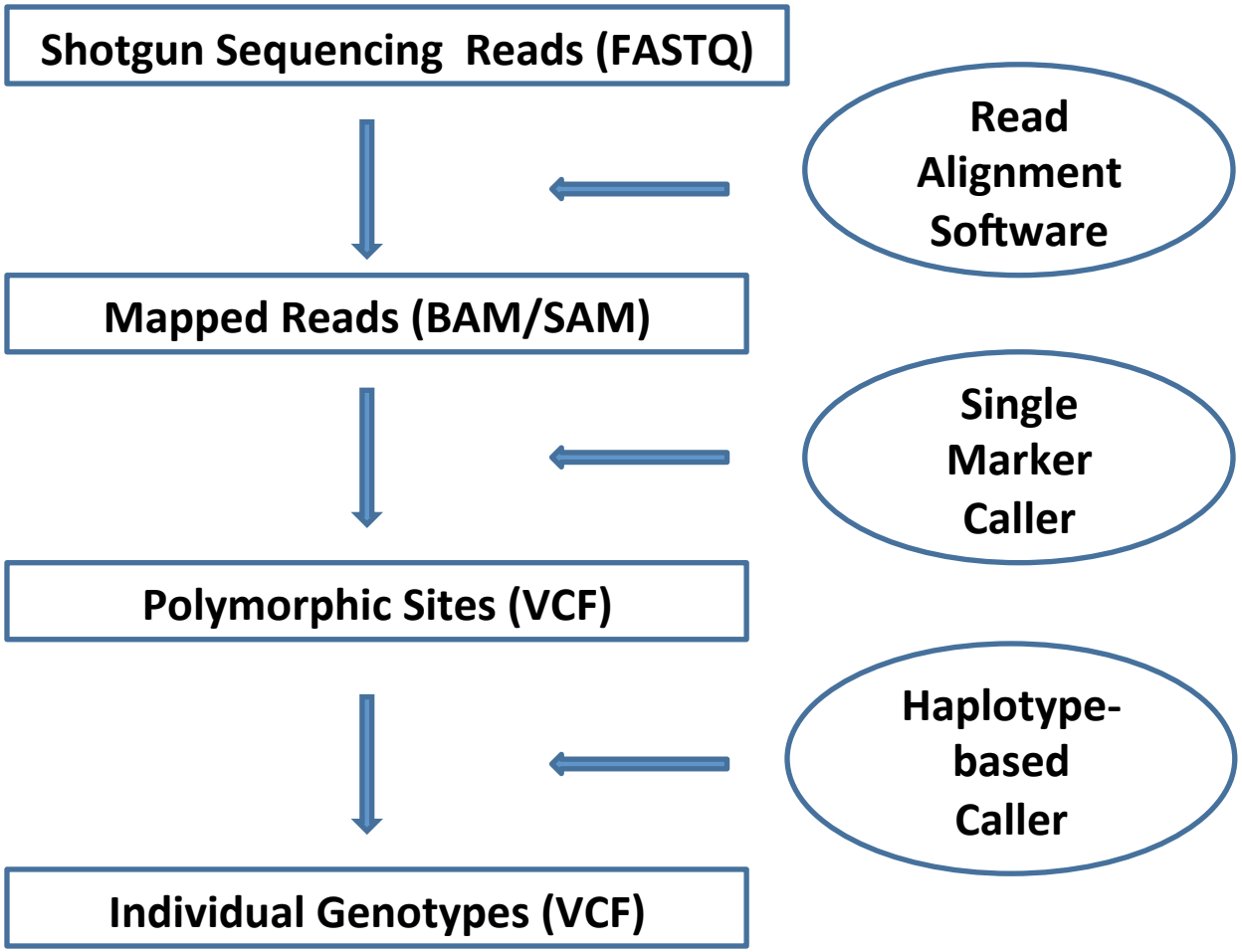
BIOST 2055

03/23/2012

Last Lecture

- From sequence data to genotypes
- Alignment
  - Hash table
  - Burrows-Wheeler transforms
- SNP calling and genotype inference
  - Single site, single sample
  - Single site, multiple sample
  - Multiple site, multiple sample

Workflow


Shotgun Sequencing Reads (FASTQ)
Read Alignment Software
Mapped Reads (BAM/SAM)
Single Marker Caller
Polymorphic Sites (VCF)
Haplotype-based Caller
Individual Genotypes (VCF)

An Example to Walk Though

- http://genome.sph.umich.edu/wiki/TrioCaller
- http://genome.sph.umich.edu/wiki/Tutorial:_Low_Pass_Sequence_Analysis

- File formats
  - FASTQ/FASTA
  - BAM/SAM
  - VCF

- Software
  - BWA: mapping reads
  - Samtools/bcftools: detect SNPs and call genotypes
  - Thunder/Triocaller: refine genotypes using flanking haplotypes

FASTA

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCC
CAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
```

FASTQ

```
@ERR009169.17725968 IL18_2954:8:100:1790:1881/2
CTAAAAATACAAAAAAAAAAAAAGAAAAAAAATGCTGAGCATCGTGGCGGATGGCTGTAAACCCAGCTACTCGGGA
+
@BBBBBBBBABCBCBBBBBCBB@=@BBBBBB@<@7?=?;15)9=/@@6AB6*6%(%5&2=%*,'2))-2.12?A:(
@ERR009169.17725969 IL18_2954:8:100:1790:1768/2
ACTACCTATGAAGTGGGAACATTTTAAAGGCAAGAAATCAGAGCTCAGAAGTCAAGTAACTTACTCAAGATCACAC
+
BBBBABCBBCBBBBBBCBBBBBBBBBBBBB@BBBBBBBABBAC@B>BCB@B>>AABAAB@BAB@B>B@B@@=BABA
@ERR009169.17725970 IL18_2954:8:100:1790:480/2
ACAAAATACAGCCAATTCTTGCTATTTGCAGTAGTGAGGTTCTAGAAAGTCACCGTGAACGCTGAGCTGCCACTCC
+
??@@?=@@???>????@???@?@????<;>?=??9?>>?8=????>>??=>><=?7?:?=?====>=4==6=<===
@ERR009169.17725972 IL18_2954:8:100:1790:1563/2
CGGTAACTGCTATGTGTAAAGGCTTAGGGCACTTTACACCTGTCAGACTGACAAATCAGACAGTGGAATCATGCAA
+
=>>??@?@@??@??>@????@@>??????>?<????>?>???>???>=???7=>??2>?===>7?>==?4>=?=6&
@ERR009169.17725973 IL18_2954:8:100:1790:1246/2
TTCCTTTGAGTAAGATATGGGATGTTATTAATTGATTAATCTCCCTCCCTATTCTTAAAAATGATTTAAGGAGGGT
+
BA@ABB8%4A=A?A<@?BAA?=A@;A=BA:3@A??AA=;?=>?>6==6?=9;1'@=2+282=>3?8997=44=778
```

Base Quality Score

Phred Quality Score

$Q = -10 \times \log_{10}(e)$ where e is the per-base sequencing error.

$Q = \text{ASCII} - 33$

BAM/SAM

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1         TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGCCAT


@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

http://genome.sph.umich.edu/wiki/SAM

CIGAR

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

VCF file

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2    GT:GQ:DP:HQ 0|0:48:1:51,51 20
17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 20
1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 20
1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Reference

- Examples:
- http://genome.sph.umich.edu/wiki/Tutorial:_Low_Pass_Sequence_Analysis
- http://genome.sph.umich.edu/wiki/TrioCaller

- The 1000 Genomes Project (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-73

- Li Y et al (2011) Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research* **21**:940-951.

- Li H et al (2009) The Sequence Alignment/Map format and SAMtools

Rare Variant Tests

- Genotype calling is the first step of the journey
- Identify SNPs/genes associated with phenotype
- Sequencing provides more comprehensive way to study the genome
  - Discover more rare variants

Association Study in Case Control Samples

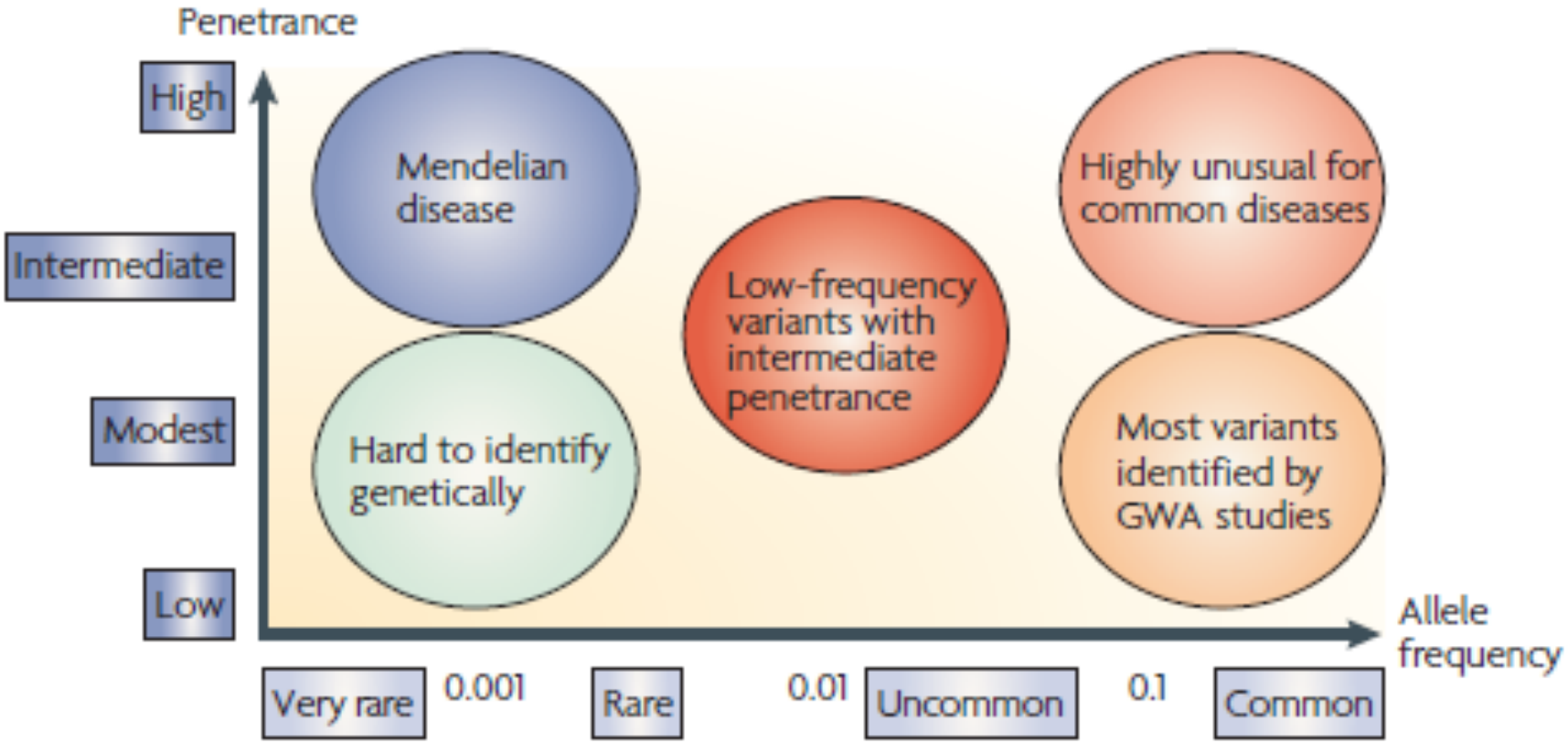
SNP1 SNP2 SNP3 SNP4 SNP5 Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

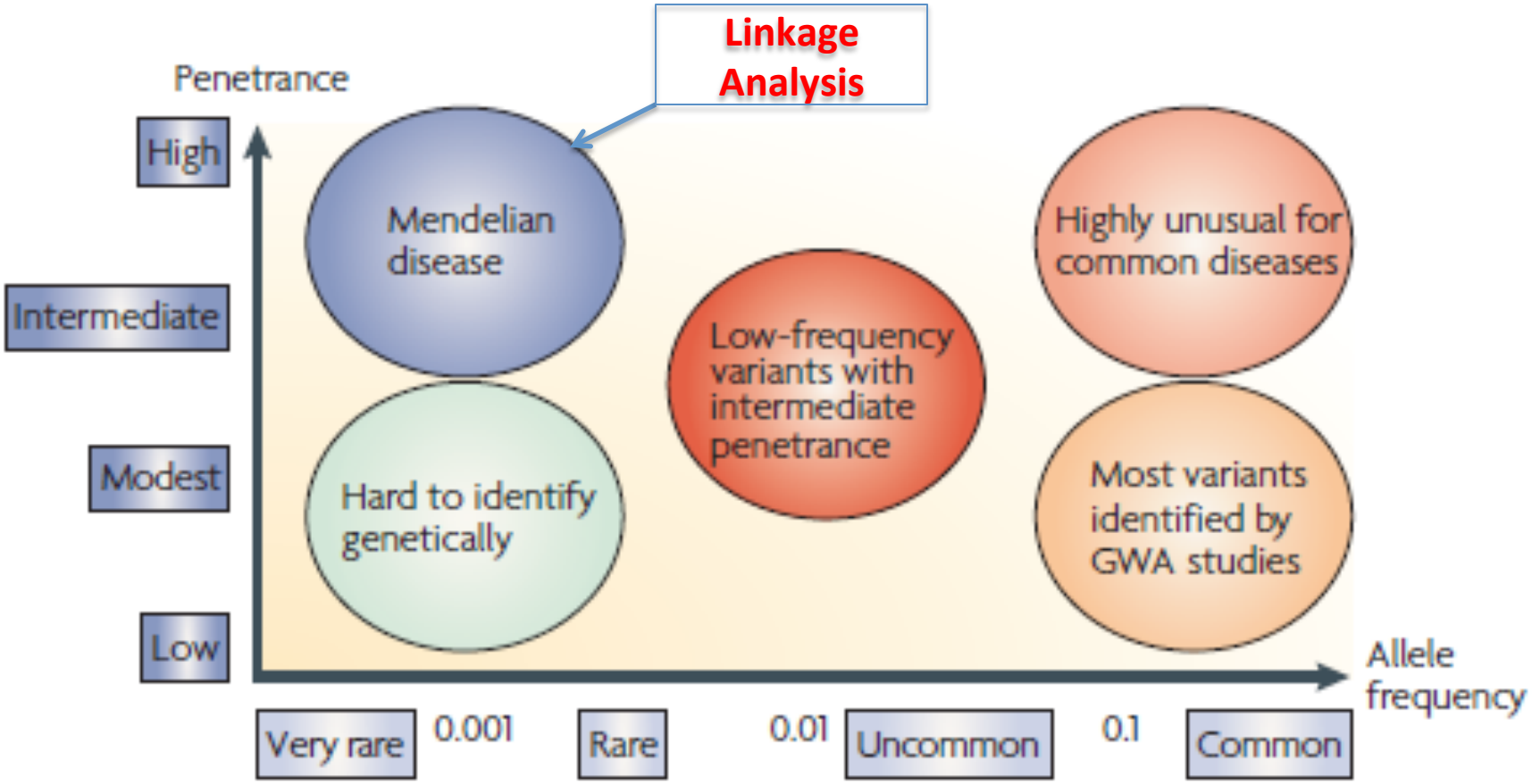
CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

Genetic Spectrum of Complex Diseases



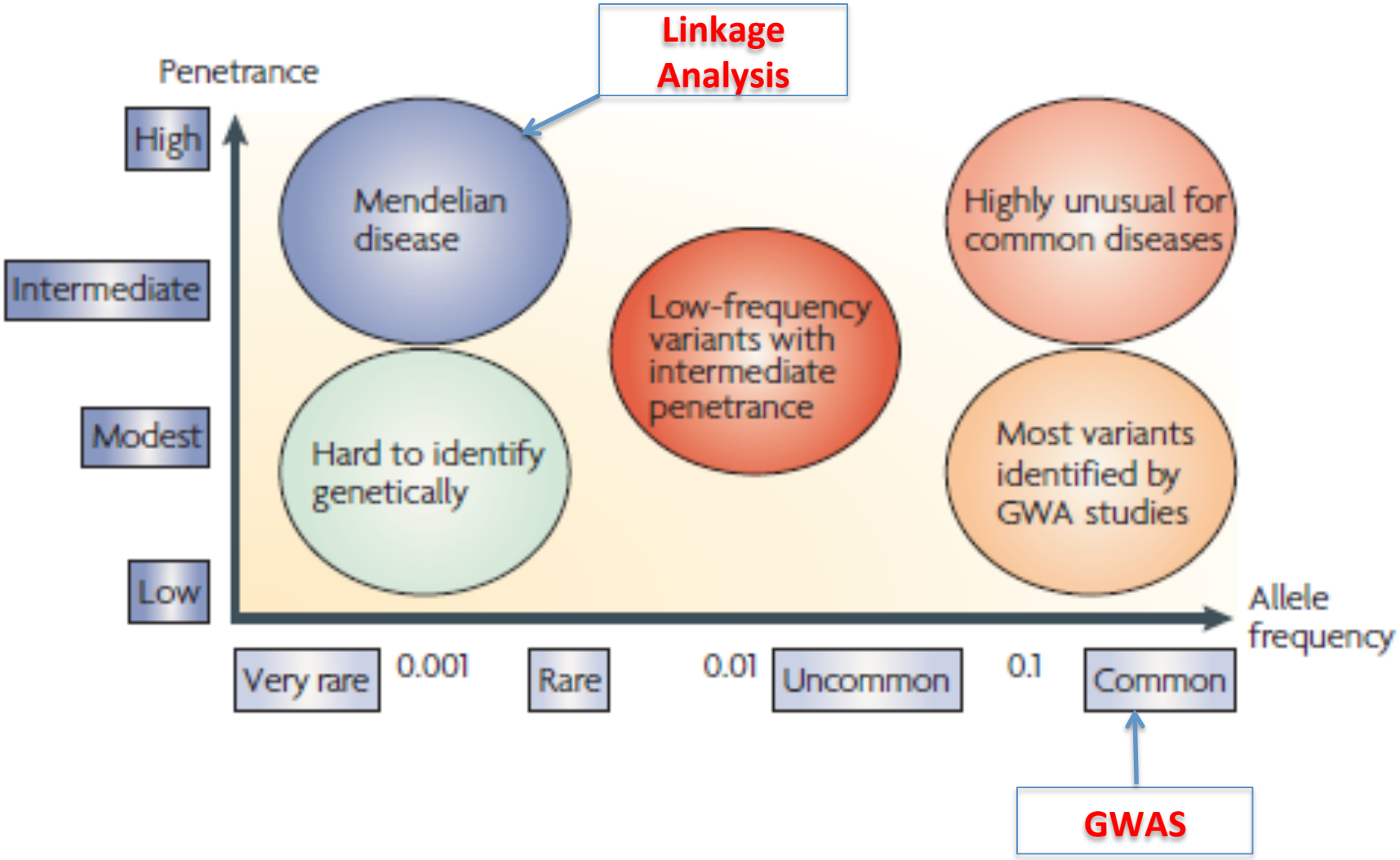
McCarthy MI et al. Nat Rev Genet. 2008

Genetic Spectrum of Complex Diseases



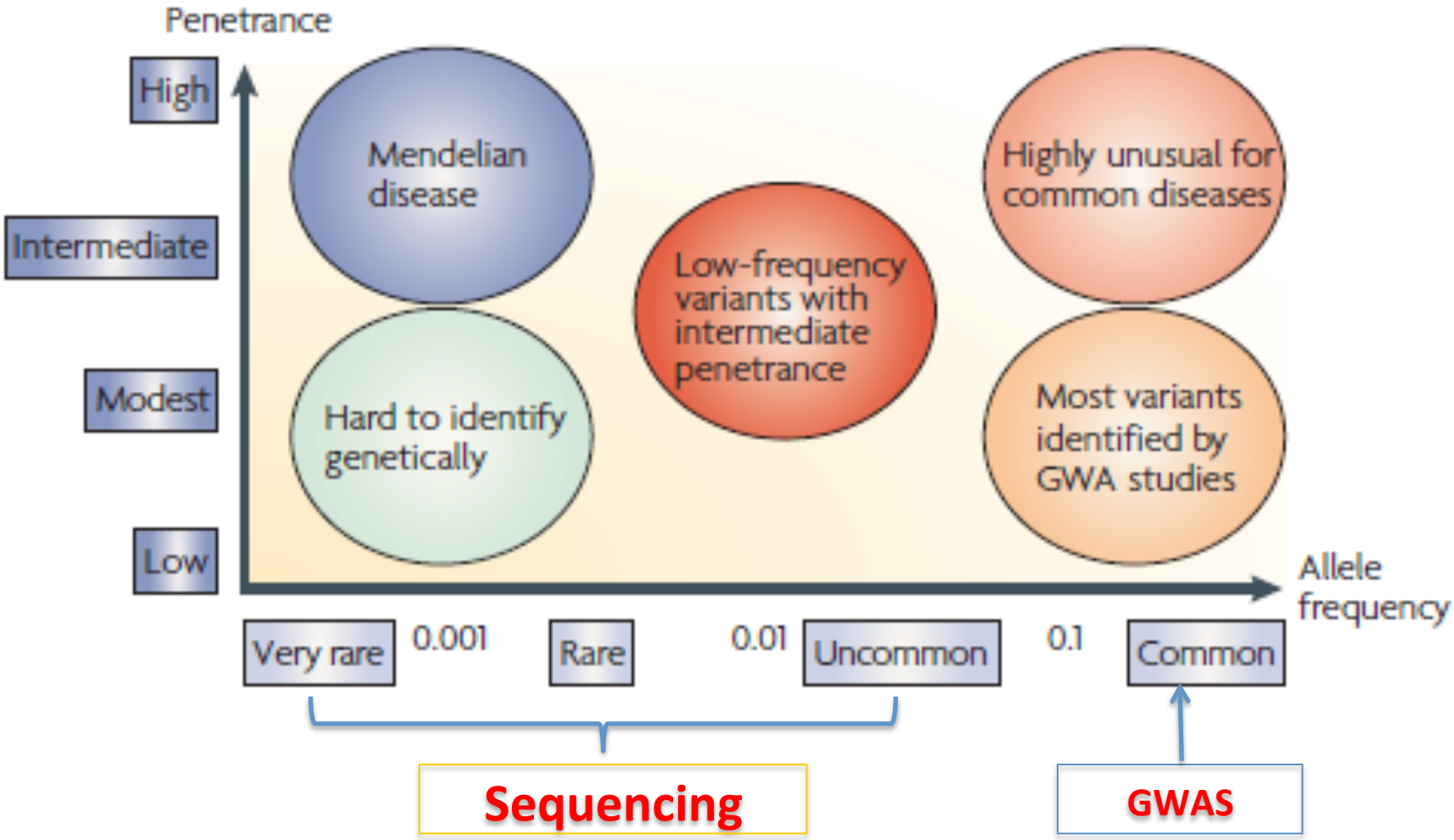
McCarthy MI et al. Nat Rev Genet. 2008

Genetic Spectrum of Complex Diseases



McCarthy MI et al. Nat Rev Genet. 2008

Genetic Spectrum of Complex Diseases



McCarthy MI et al. Nat Rev Genet. 2008

Why Study Rare Variants

- Identify additional susceptibility loci
- Find missing heritability
- Lead to functional analysis

Several Approaches to Study Rare Variants

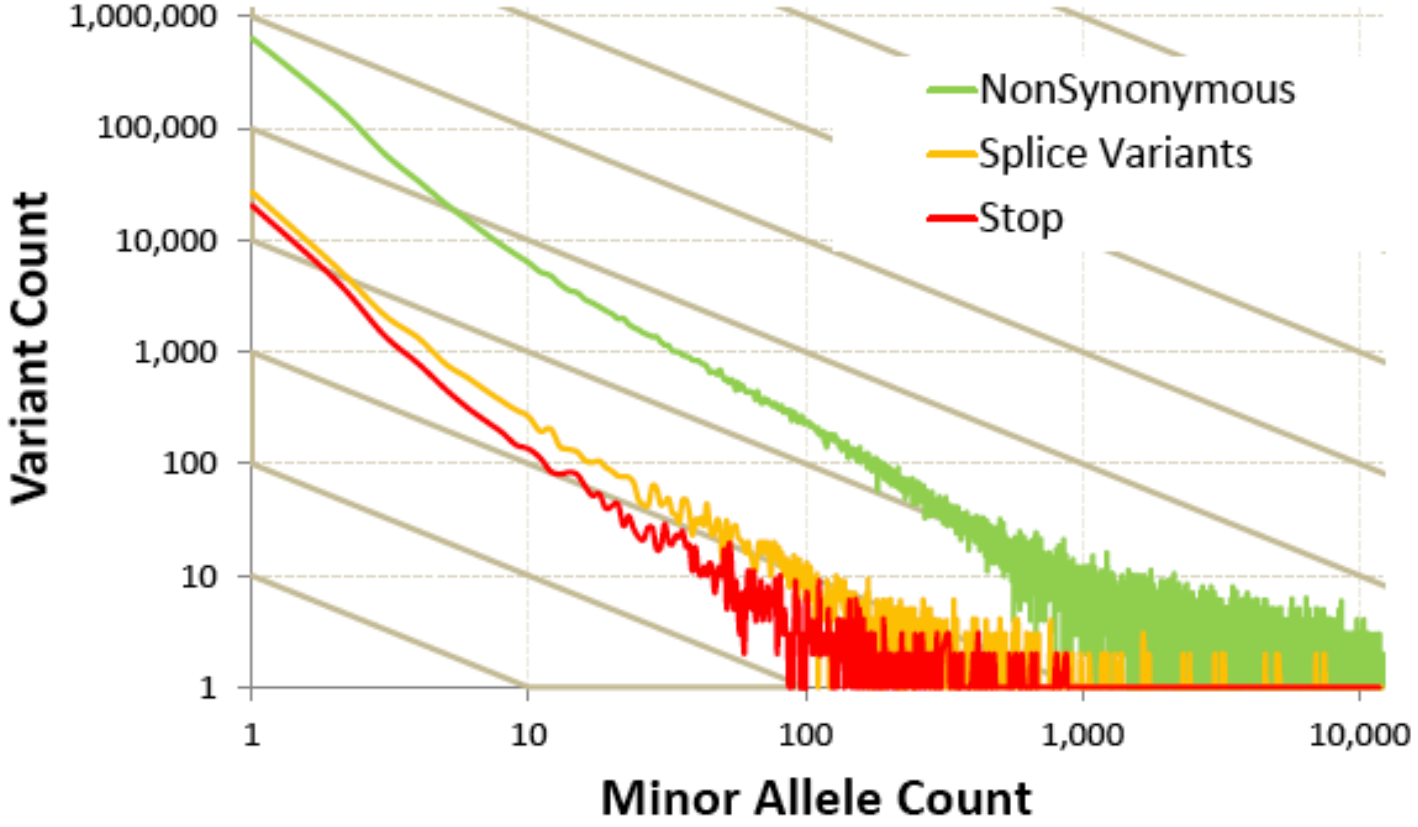
- **Deep whole genome sequencing**
  - Can only be applied to limited numbers of samples
  - Most complete ascertainment of variation

- **Exome capture and targeted sequencing**
  - Can be applied to moderate numbers of samples
  - SNPs and indels in the most interesting 1% of the genome

- **Low coverage whole genome sequencing**
  - Can be applied to moderate numbers of samples
  - Very complete ascertainment of shared variation

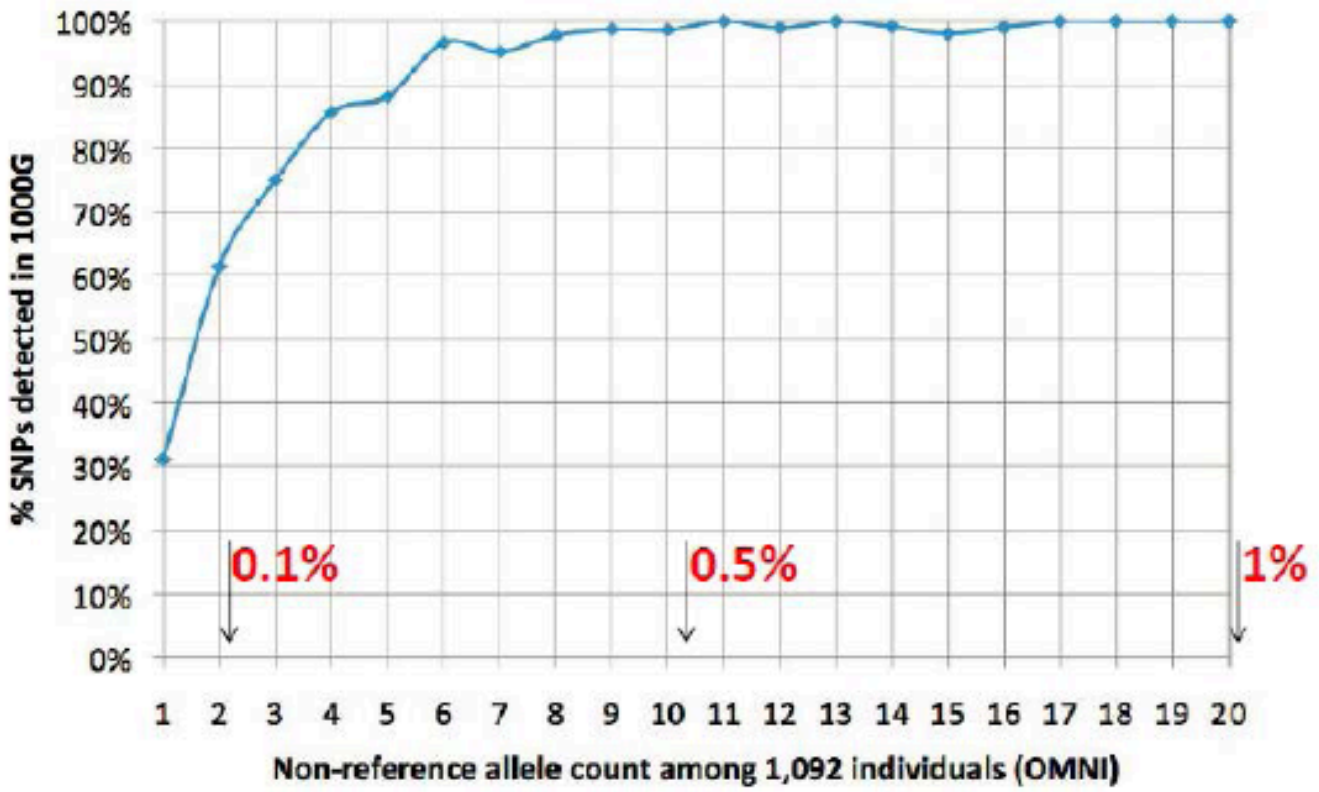
- **New Genotyping Arrays and/or Genotype Imputation**
  - Examine low frequency coding variants in 100,000s of samples
  - Current catalogs include 97-98% of sites detectable by sequencing an individual

Allele Frequency Spectrum (Sequenced 12,000+ Individuals)



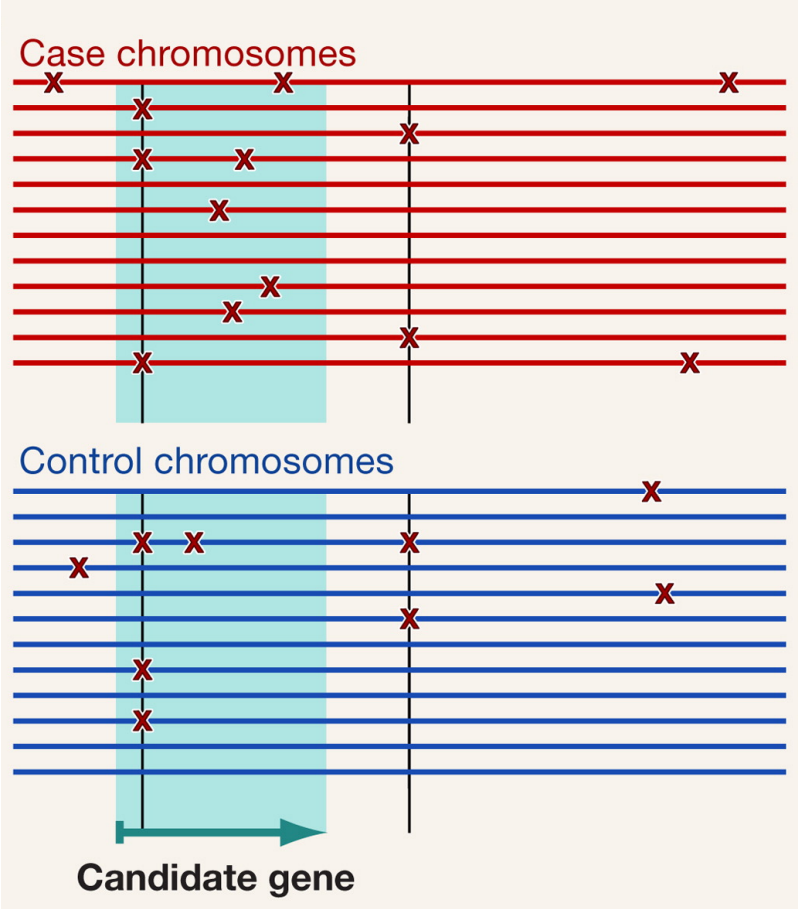
http://genome.sph.umich.edu/wiki/Exome_Chip_Design

SNP Detection in Low Pass Sequencing



In 1000 Genomes Project Phase I (1094 samples @ 4x), Hyun Min Kang

Rare Variants



Single SNP Test for Rare Variant

- Disease prevalence ~10%
- Type I error $5x10^{-6}$
- To achieve 80% power
- Equal number of cases and controls

- Minor Allele Frequency = 0.1, 0.01, 0.001
- Required sample size = 486, 3545, 34322,

Single SNP Test for Rare Variant

- Rare variants are hard to detect
- Power/sample size depends on both frequency and effect size
- Rare causal SNPs are hard to identify even with large effect size

Alternatives to Single Variant Test Collapsing Method

- Group rare variants in the same gene/region
- Score each individual
  - Presence or absence of rare copy
  - Weight each variant
- Use individual score as a new “genotype”

Challenges

- Disease is caused by multiple rare variants in an additive manner
- It is hard to separate causal and null SNPs
  - Including all rare variants will dilute the true signals
- The effect size of each rare variant varies

Power of Burden Test

	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

- Power tabulated in collections of simulated data
- Combining variants can greatly increase power
- Currently, appropriately combining variants is expected to be key feature of rare variant studies.

Impact of Null Alleles

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants + 5 null variants	.04	.70
10 disease associated variants + 10 null variants	.03	.55
10 disease associated variants + 20 null variants	.03	.33

- Including non-disease variants reduces power
- Power loss is manageable, combined test remains preferable to single marker tests

Impact of Missing Disease Alleles

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants, 2 missed	.05	.72
10 disease associated variants , 4 missed	.05	.52
10 disease associated variants , 6 missed	.04	.28
10 disease associated variants, 8 missed	.03	.08

- Missing disease alleles loses power
- Still better than single variant test

Refining Rare Variant Test

- Counting the number of rare variants per individual
- Weighting rare variants according to frequency
- Weighting rare variants according to function
- Including imputed variants in the analysis

- Each of these methods may improve power, but few practical examples provide guidance

Weighted Sum Statistic

- Assumption: effect size is inversely proportional to minor allele frequency

- Weight $\hat{w}_i = \sqrt{n_i \cdot q_i(1-q_i)}, \qquad q_i = \frac{m_i^U + 1}{2n_i^U + 2},$

- Individual genetic score $\gamma_j = \sum_{i=1}^{L} \frac{I_{ij}}{\hat{w}_i},$

CMAT: Combined Minor Allele Test

Consider gene with k variants in sample of N cases and N controls.

For polymorphism i define:

- w_i, a weight based on functional annotation, minor allele frequency, imputation accuracy
- g_{ij}, the expected posterior minor allele count in individual j.
- Set $m_A = \sum_{i=1}^{k} w_i \sum_{j=case} g_{ij} \quad M_A = \sum_{i=1}^{k} w_i \sum_{j=case} (2 - g_{ij})$

The test statistic is then $\Sigma_{CMAT} = \dfrac{m_A M_U - m_U M_A}{N(m_A + m_U)(M_A + M_U)}$

Significance of the test statistic evaluated by permutation of affection status.

Zawistowski et al (2010)

Discussion

- An active research area
- What to do if there are causal alleles in the opposite directions
- What to do if the samples are related
- Most tests reply on permutation
  - Computationally intensive

Reference

- Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. *Cell*. 2011 Sep 30;147(1):57-69.

- Li and Leal (2008) *Am J Hum Genet* **83:**311-321

- Madsen BE, Browning SR (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet 5(2)

- Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S (2010) *Am J Hum Genet* **87**:604-617