Analysis of Next Generation Sequence Data

02/13/2019

Qi Yan
Department of Pediatrics
Children's Hospital of Pittsburgh of UPMC
University of Pittsburgh

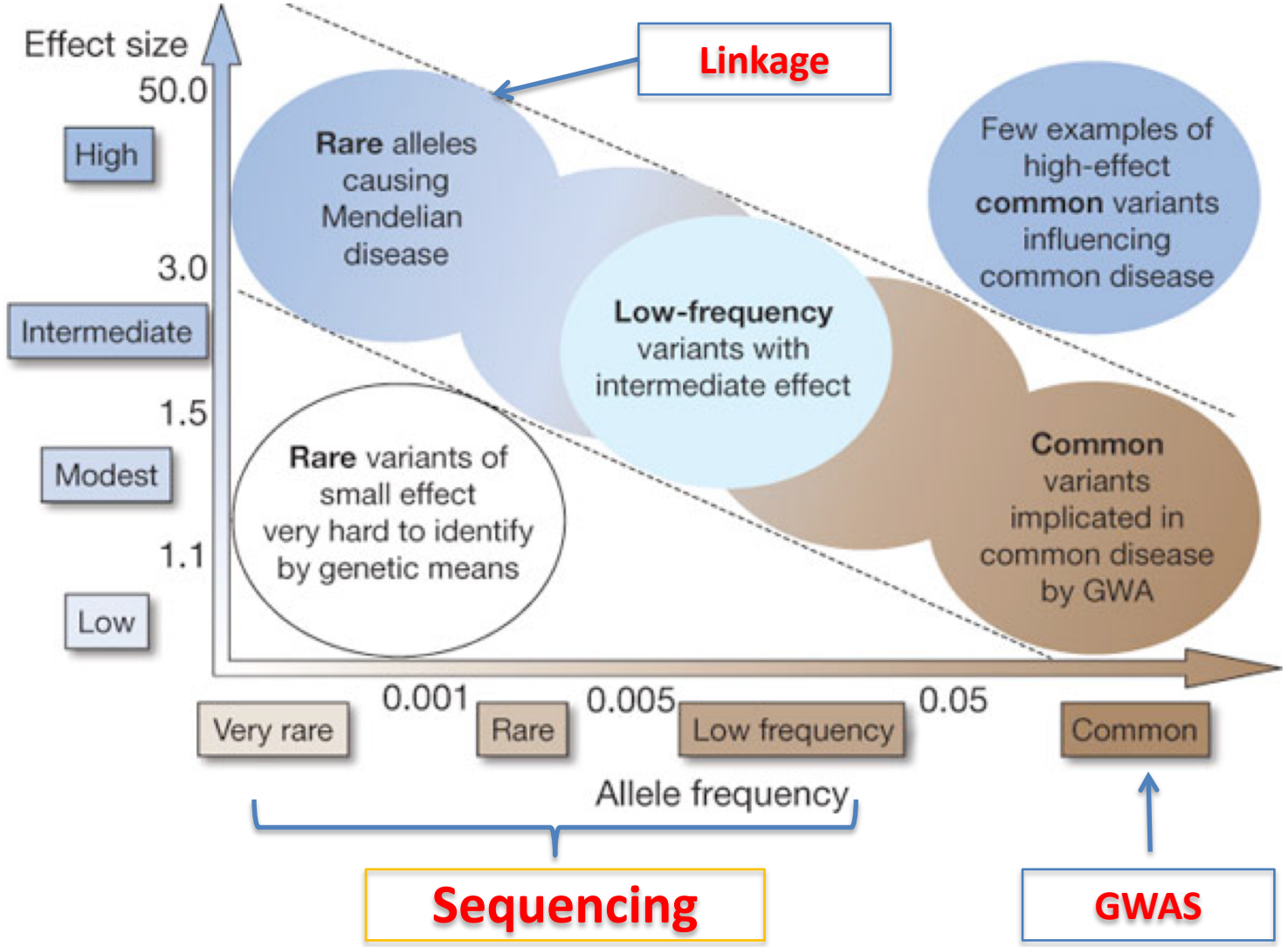
Last Lecture

- Genome-wide association study has identified thousands of disease-associated loci
- Large consortium performs meta-analysis to further increase the sample size (power) to detect additional loci
- GWAS is limited by the chip design and rare variants are rarely explored

Outline

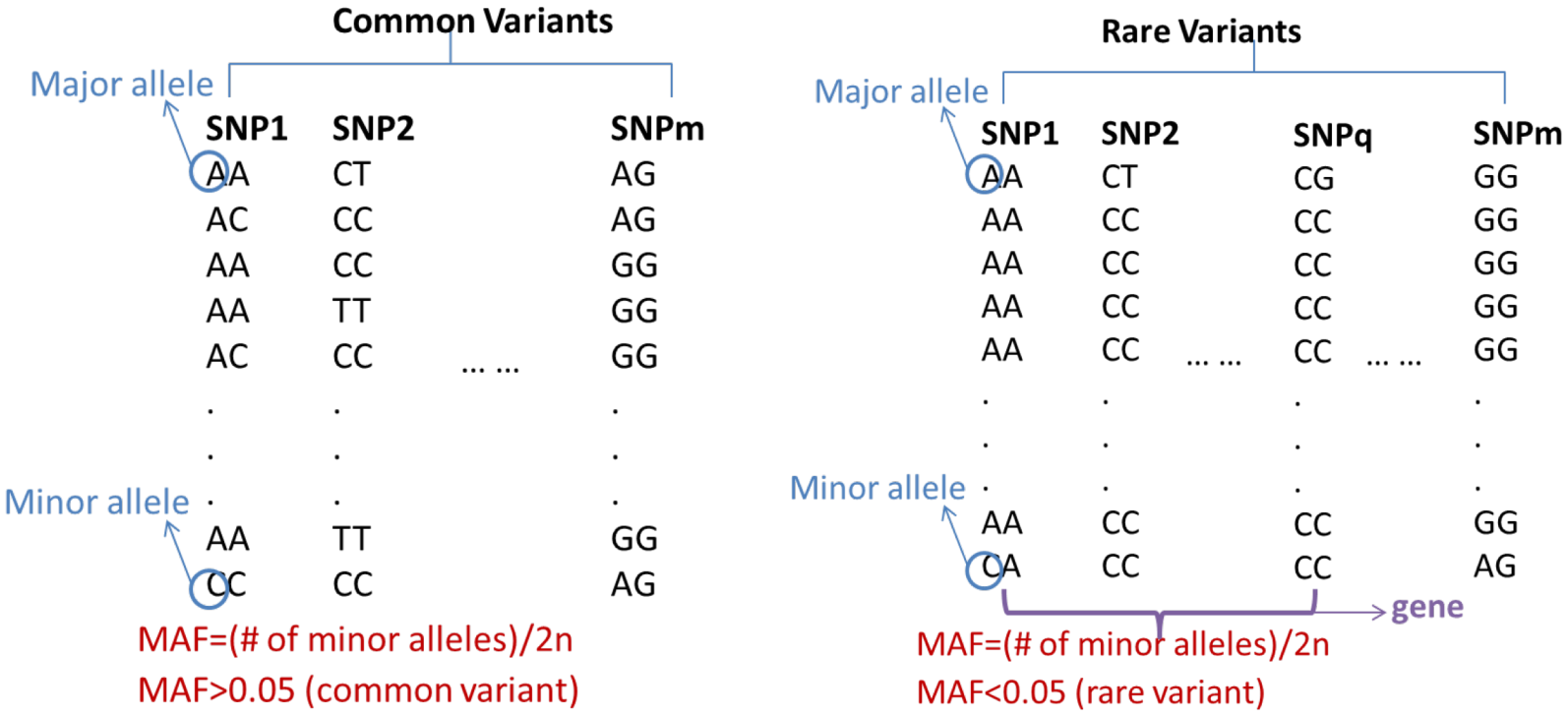
- Background of DNA sequencing
- From sequence data to genotype
- Rare variant tests

Genetic Spectrum of Complex Diseases



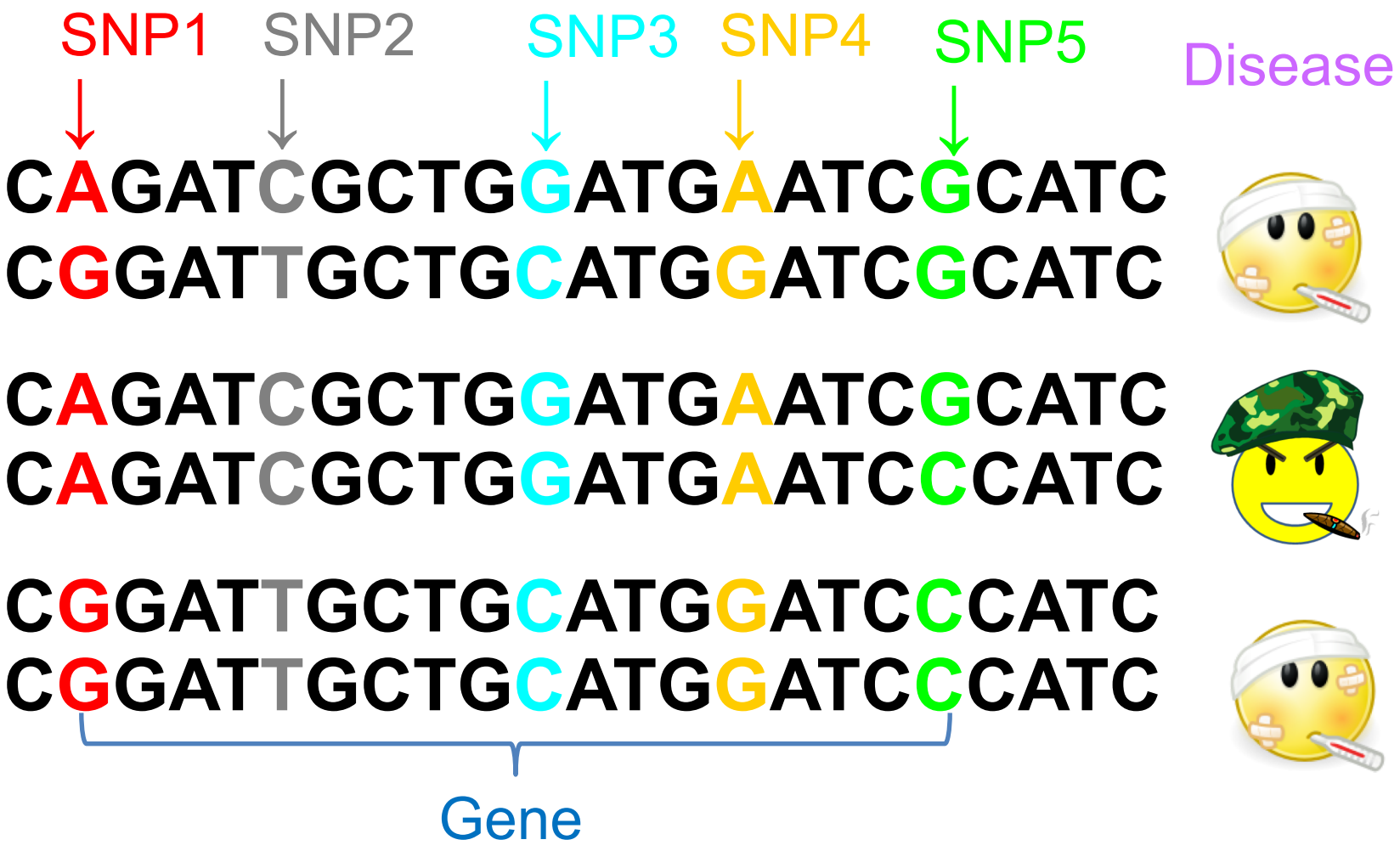
Common VS Rare

- Genotypes:
  - Common variants (e.g. MAF≥0.05): single marker test;
  - Rare variants (e.g. MAF<0.05): test at gene level

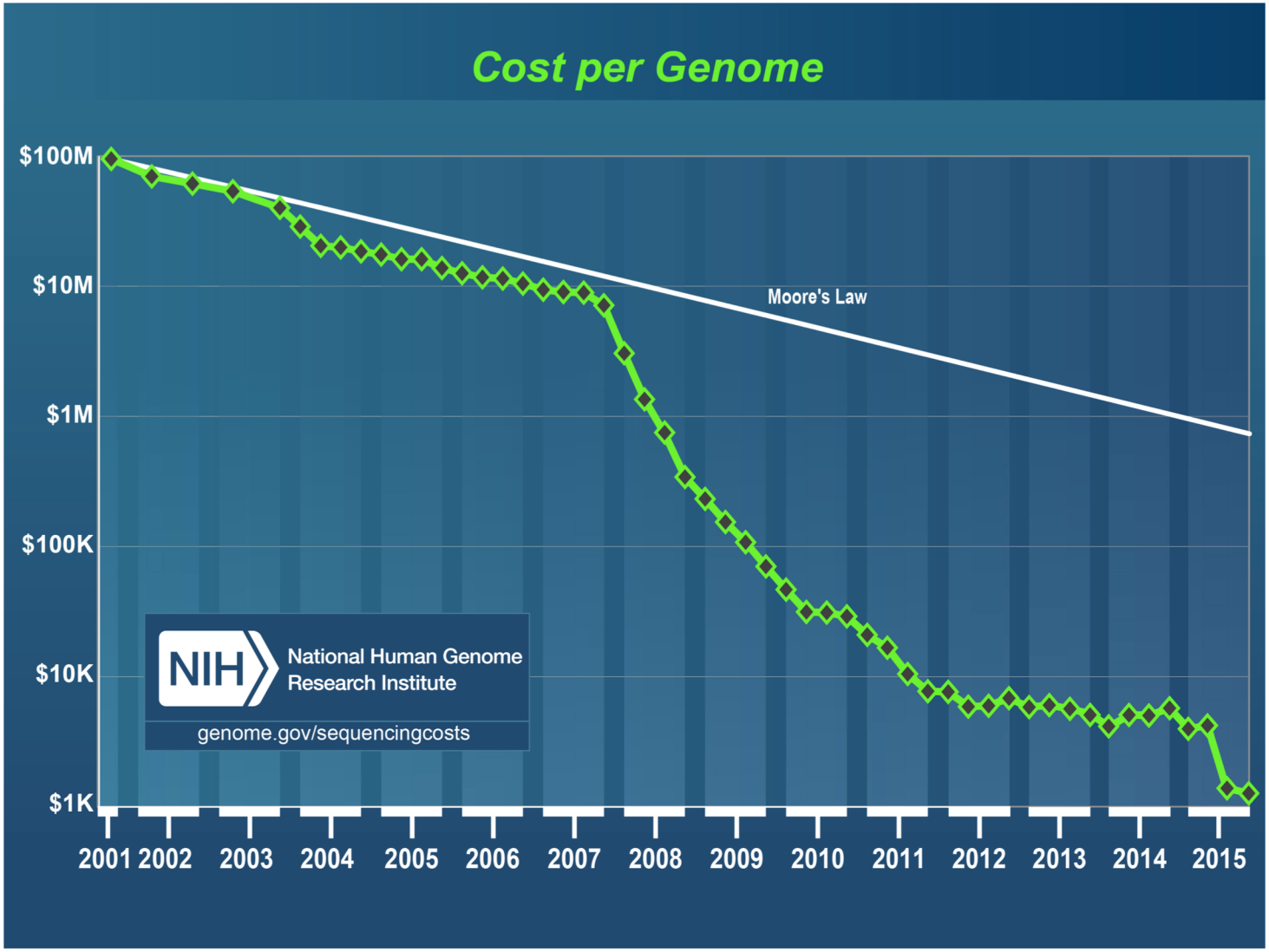


- **Only subset of functional elements include common variants**
- **Rare variants are more numerous and thus will point to additional associated loci**

Association Study in Case Control Samples



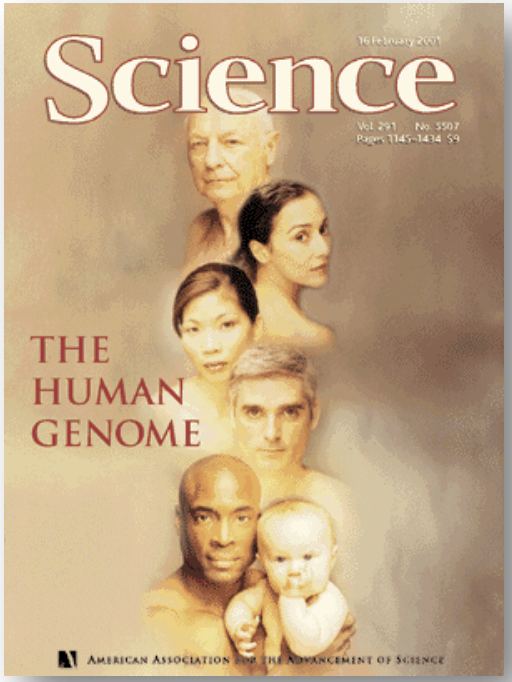
Sequencing Cost



http://www.genome.gov/

A Road to Discover Human Genome

- A lot of efforts have been made to discover and understand human genome.
- In the past 20 years, several projects involving multiple countries have been finished or on progress.

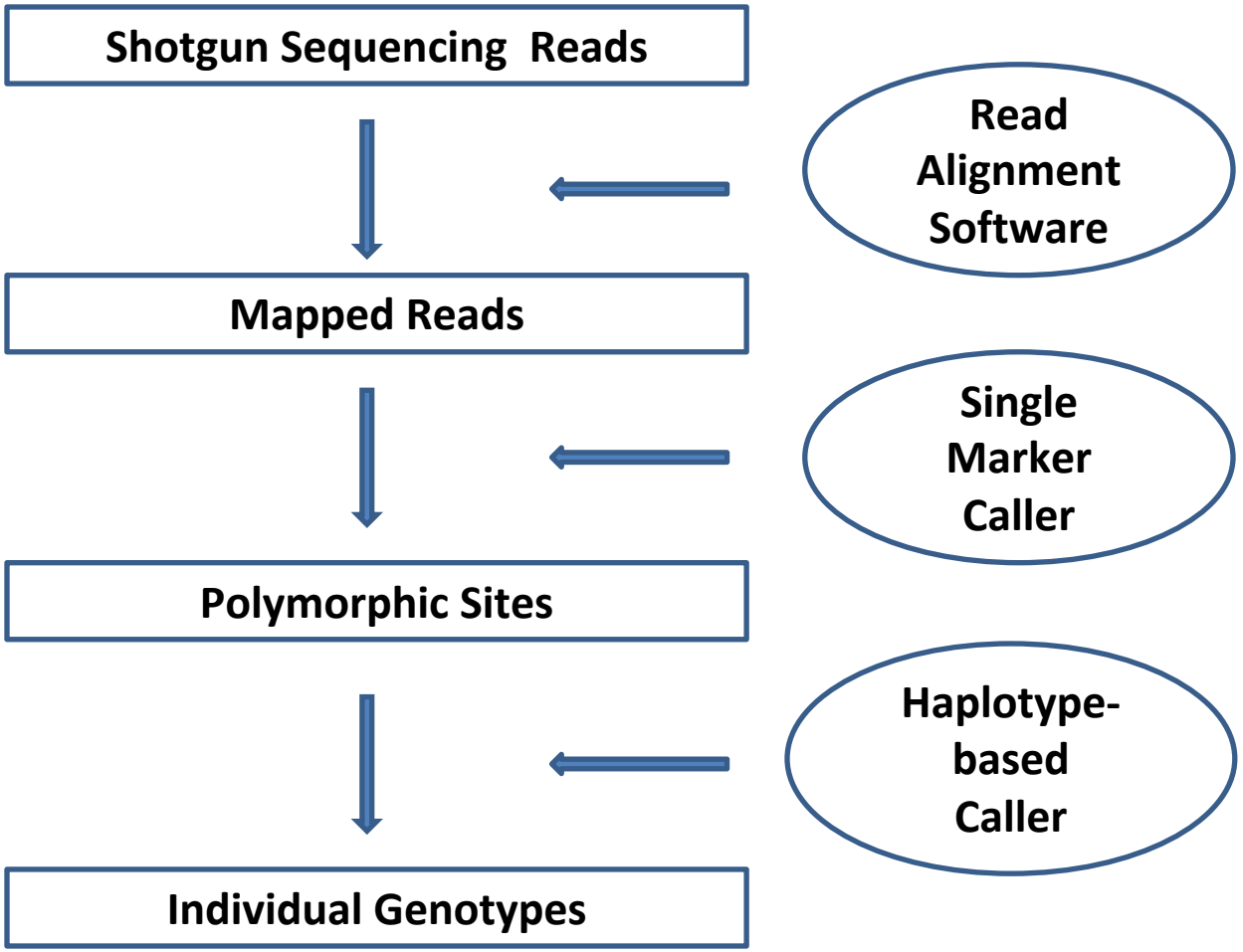


- The Human Genome Project identifies 20,000-25,000 genes in human DNA, determine the sequences of the 3 billion chemical base pairs that make up human DNA and boost the sequencing technology in industry.
- The HapMap Project determines the common patterns of DNA sequence variation in the human genome from populations with ancestry from Africa, Asia and Europe.
- The One Thousand Genome Project sequenced the genomes of more than 1,000 individuals from more than 10 different ethnic groups using next-generation sequencing technology and provide a much deeper catalog of human genetic variation.

Next Generation Sequencing

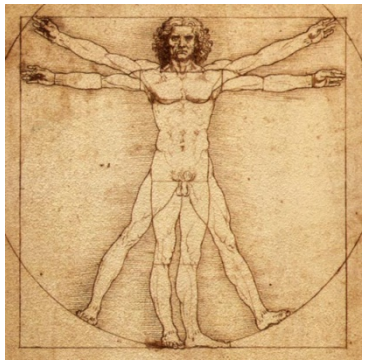
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
  - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, and others…
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy 0.5 – 1.0% error rates per base may be typical
- High-throughput but hard to assemble

A Typical Pipeline



Short read alignment

Sequencer



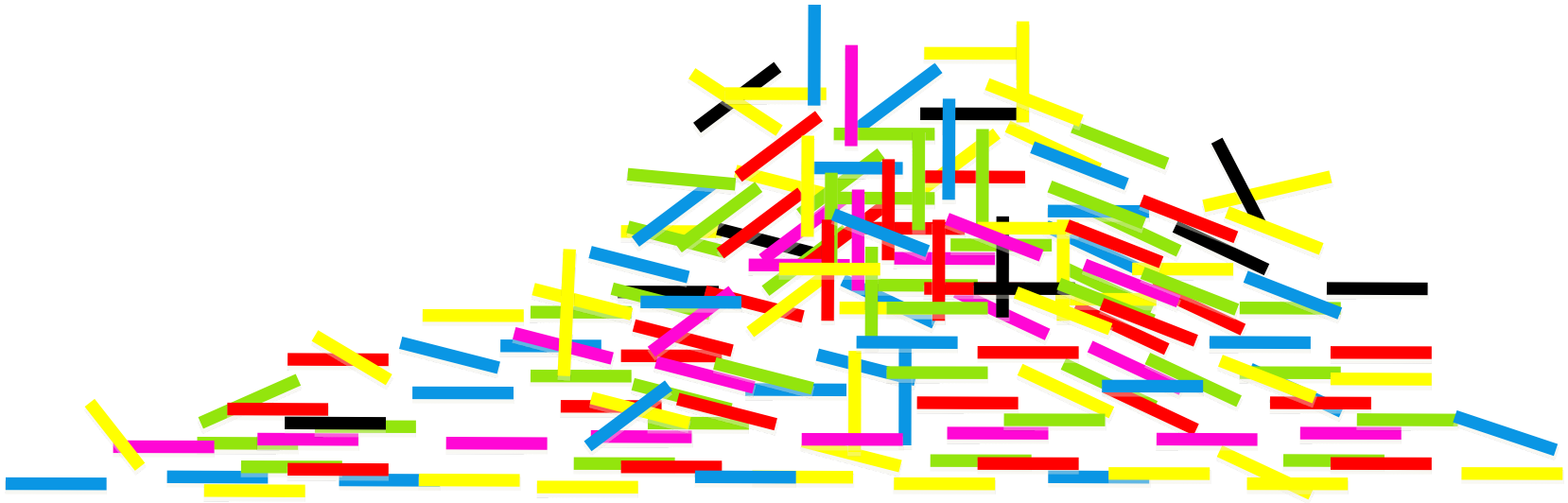
Human source

Reads from sequencing machines are short: 30-400 bp

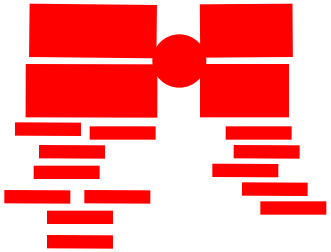
Short read alignment

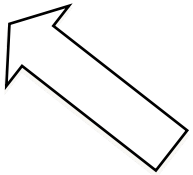
Sequencing machine

And you get
MILLIONS of them

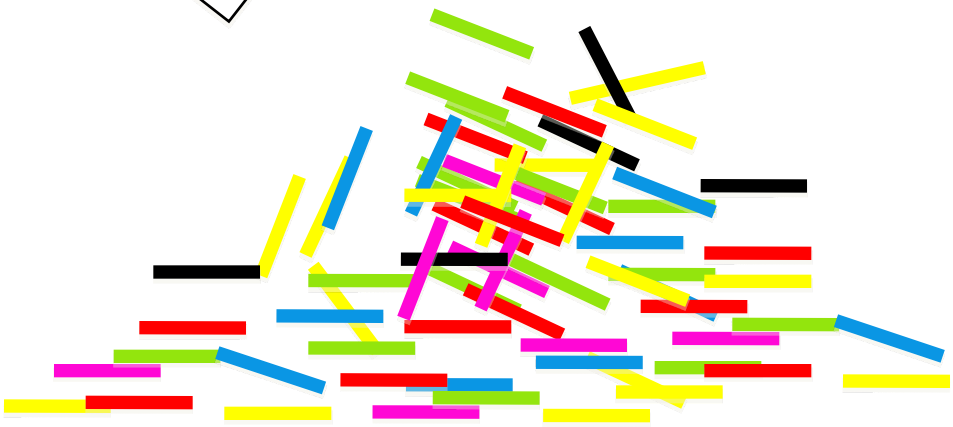


Short read alignment





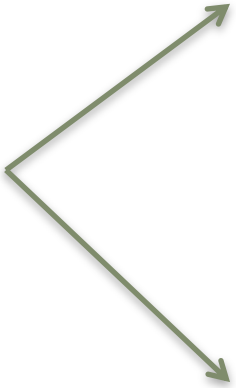
Need to map them back to human reference



Alignment

Reference sequence:

actgtagattag**ccgagtagctag**ctagtcgat



Find best match for each read in a reference sequence

ccgagaagctag

Challenges:

- Hashing is time and memory consuming for millions of reads and billion-base long reference
- Errors in reads
- Each read may be mapped to multiple positions
- Individual polymorphisms

Existing Alignment by Category

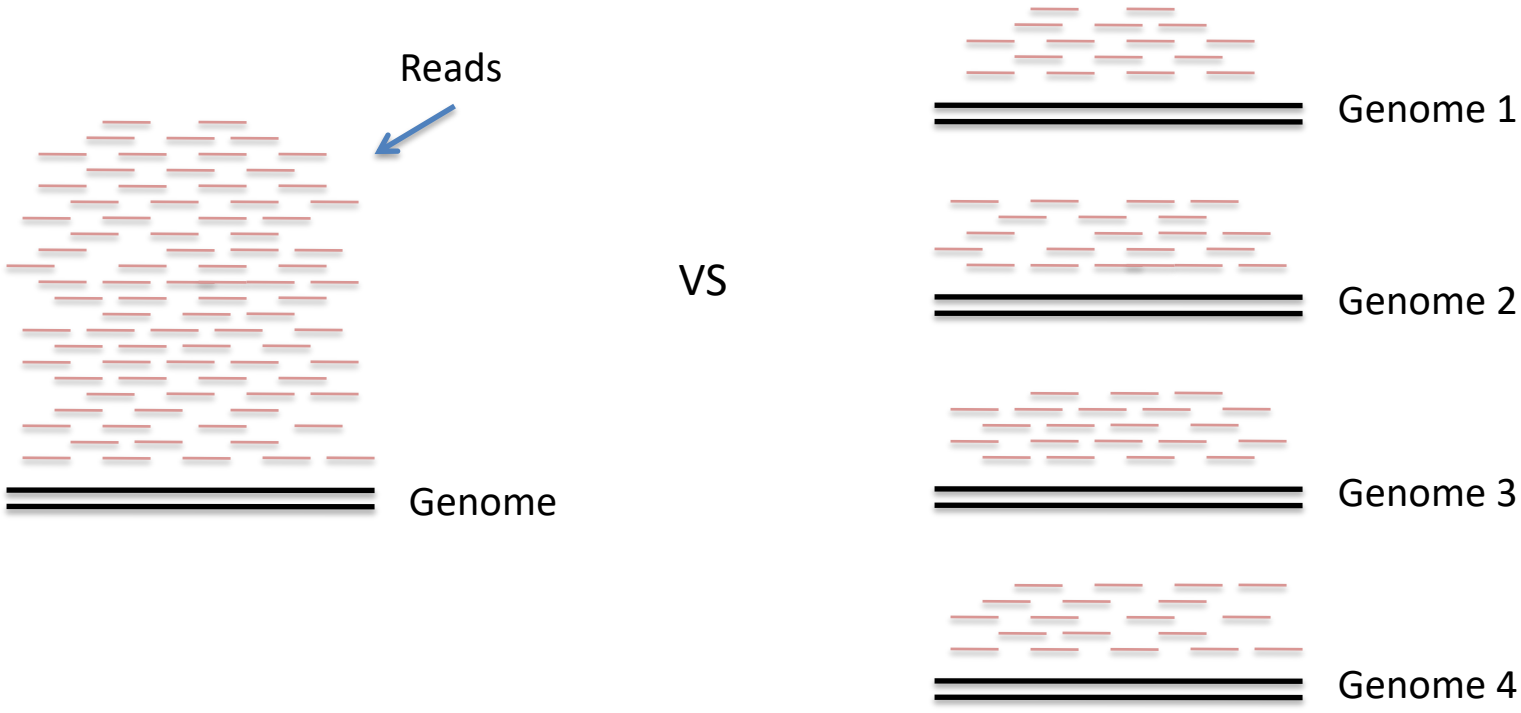
- Hashing reference genome
  - SOAP1, MOSAIK, PASS, BFAST, …
- Hashing short reads
  - Eland, MAQ, SHRiMP, …
- Merge-sorting reference together with reads
  - Slider
- Based on Burrows-Wheeler Transform
  - BWA, SOAP2, Bowtie, …

Li and Durbin (2009), *Bioinformatics* 25 (14): 1754-60

After Alignment

- Each read is mapped to reference genome with tolerated number of mismatches
  - Mismatches allow us to discover the individual variation
- Each site of reference genome is covered by multiple un-evenly distributed reads
  - Some sites might not be covered

Coverage (High vs Low)



- Which one has more power to detect variations?

Different Approaches

- Deep whole genome sequencing
  - Expensive, only can be applied to limited samples currently
  - Most complete ascertainment of all variations
- Low coverage whole genome sequencing
  - Modest cost, typically 100-1000 samples
  - Complete ascertainment of common variations
  - Less compete ascertainment of rare variants
- Exome capture and targeted region sequencing
  - Modest cost, high coverage
  - Most interesting part of the genome

Genotype Calling

- Types of study
  - Population-based (unrelated samples)
  - Family-based (multiple families)

- Types of method
  - Single sample, single site
  - Multiple samples, single site
  - Multiple samples, multiple sites

Genotype Calling

- One of the most important steps in next generation sequencing downstream analysis is genotype inference

- The essential question in genotype calling is:

$$P(G|R)$$

where R denotes base call at all loci, G denotes hypothetical true genotype

Genotype Calling from Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

2A and 3C

Observed Data

A/C or A/A or C/C

Predicted Genotype

A Simple Model

At one site, n_A reads carry A, n_B reads carry B

$$N_A \sim \begin{cases} Binomial(n_A + n_B, 1-\delta) & G = A/A \\ Binomial(n_A + n_B, 0.5) & G = A/B \\ Binomial(n_A + n_B, \delta) & G = B/B \end{cases}$$

# of reads carrying A

total # of reads

Inference with no reads

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 1.0

P(reads|A/C)= 1.0

P(reads|C/C)= 1.0

Possible Genotypes

Inference with short read data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= P(C observed, read maps |A/A)

P(reads|A/C)= P(C observed, read maps |A/C)

P(reads|C/C)= P(C observed, read maps |C/C)

Possible Genotypes

Inference assuming error of 1%

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01

P(reads|A/C)= 0.50

P(reads|C/C)= 0.99

Possible Genotypes

As data accumulate …

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01*0.01**=**0.0001

P(reads|A/C)= 0.5*0.5**=**0.25

P(reads|C/C)= 0.99*0.99**=**0.98

Possible Genotypes

As data accumulate …

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01*0.01*0.01**=**0.000001

P(reads|A/C)= 0.5*0.5*0.5**=**0.125

P(reads|C/C)= 0.99*0.99*0.99**=**0.97

Possible Genotypes

As data accumulate …

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01*0.01*0.01*0.99**=**0.00000099

P(reads|A/C)= 0.5*0.5*0.5*0.5**=**0.0625

P(reads|C/C)= 0.99*0.99*0.99*0.01**=**0.0097

Possible Genotypes

In the "end"

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

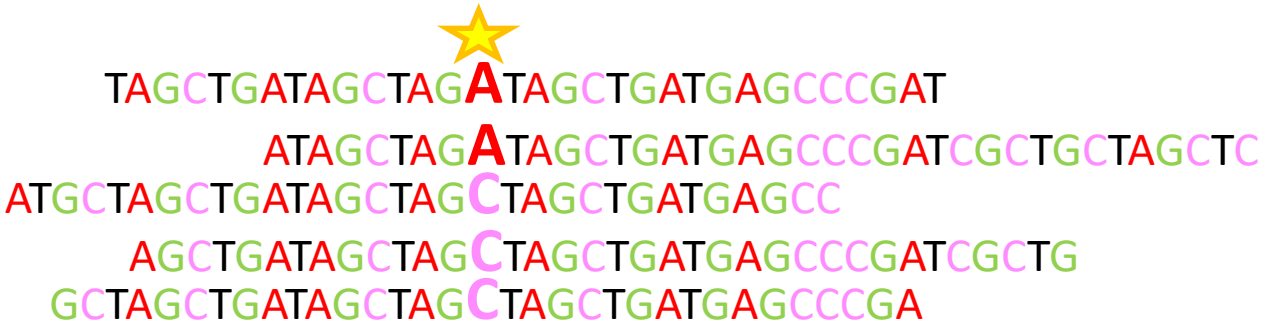
Reference Genome

P(reads|A/A)= 0.00000098

P(reads|A/C)= 0.03125

P(reads|C/C)= 0.000097

Not the "end" yet



Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

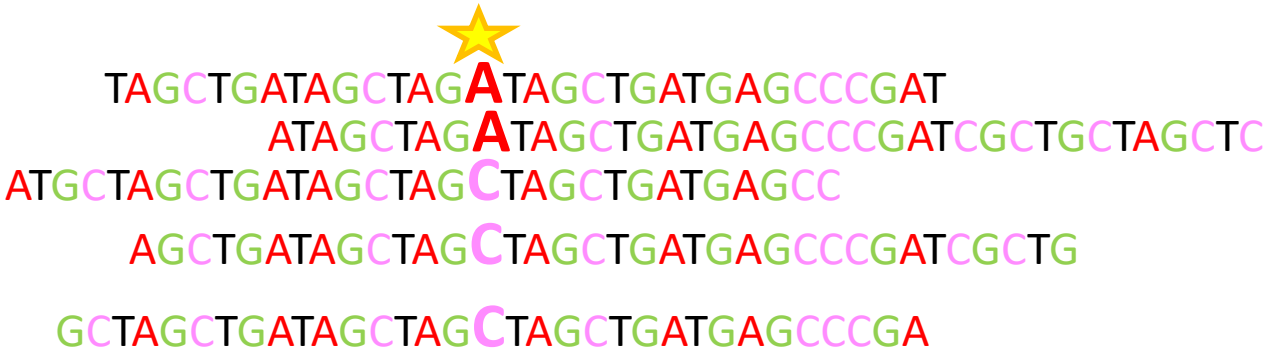
Reference Genome

P(reads|A/A) = 0.00000098
P(reads|A/C) = 0.03125
P(reads|C/C) = 0.000097

Making a genotype call requires
combining sequence data with prior information

$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Not the "end" yet



Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

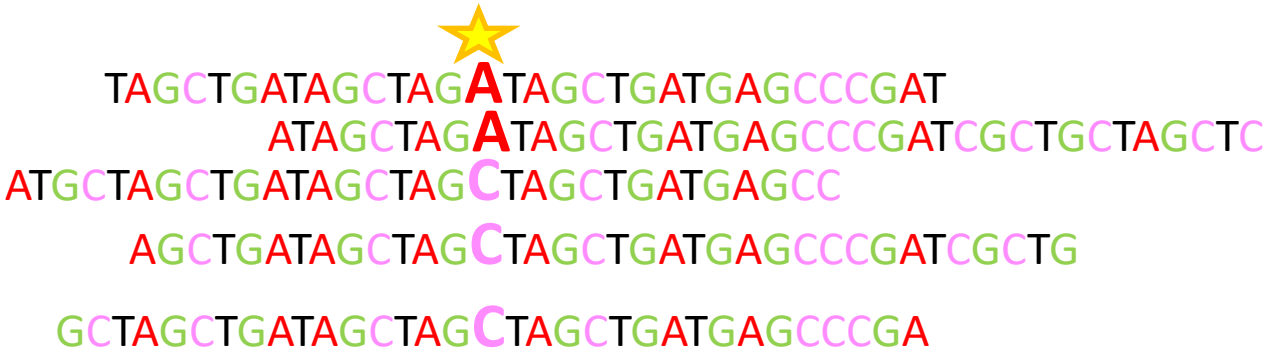
P(reads|A/A)= 0.00000098
P(reads|A/C)= 0.03125
P(reads|C/C)= 0.000097

Prior(A/A) = 0.00034
Prior(A/C) = 0.00066
Prior(C/C) = 0.99900

P(A/A|reads) < 0.01
P(A/C|reads) = 0.175
P(C/C|reads) = 0.825

Base Prior: every site has 1/1000 probability of varying

Population Based Prior



Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

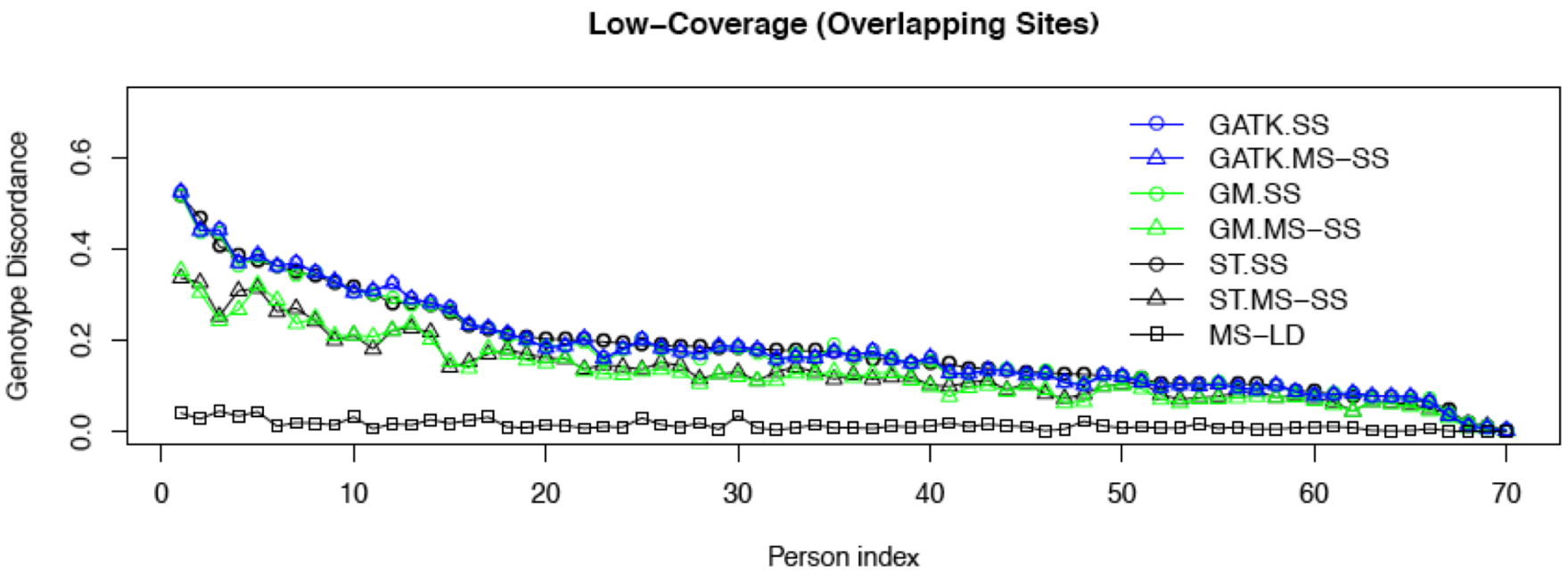
P(reads|A/A)= 0.00000098	Prior(A/A) = 0.04	P(A/A|reads) < .001
P(reads|A/C)= 0.03125	Prior(A/C) = 0.32	P(A/C|reads) = 0.999
P(reads|C/C)= 0.000097	Prior(C/C) = 0.64	P(C/C|reads) = <.001

Population Based Prior: Use frequency information from examining others at the same site. E.g. P(A) = 0.2

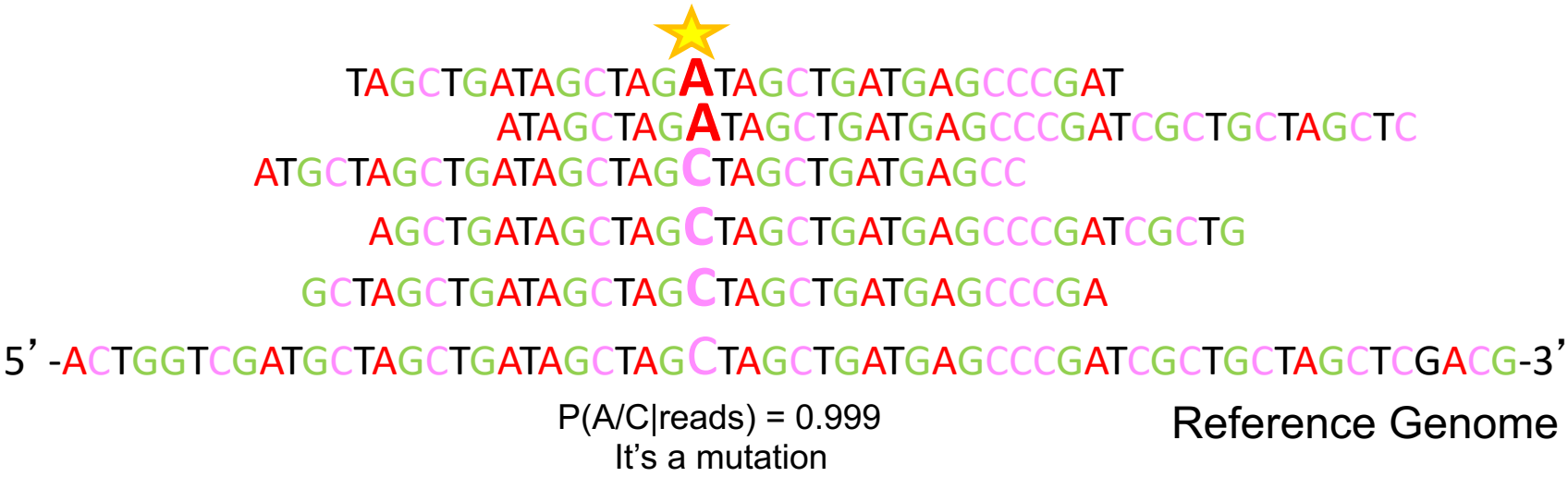
Prior Information

- Individual based prior
  - Equal probability of showing polymorphism
  - 1/1000 bases different from reference
  - Error Free and Poisson distribution
  - Single sample, single site
- Population based prior
  - Estimate frequency from many individuals
  - Multiple sample, single site
- Haplotype/Imputation based prior
  - Jointly model flanking SNPs, use haplotype information
  - Important for low coverage sequence data
  - Multiple samples, multiple sites

Comparisons of Different Genotype Calling Methods



Somatic SNPs

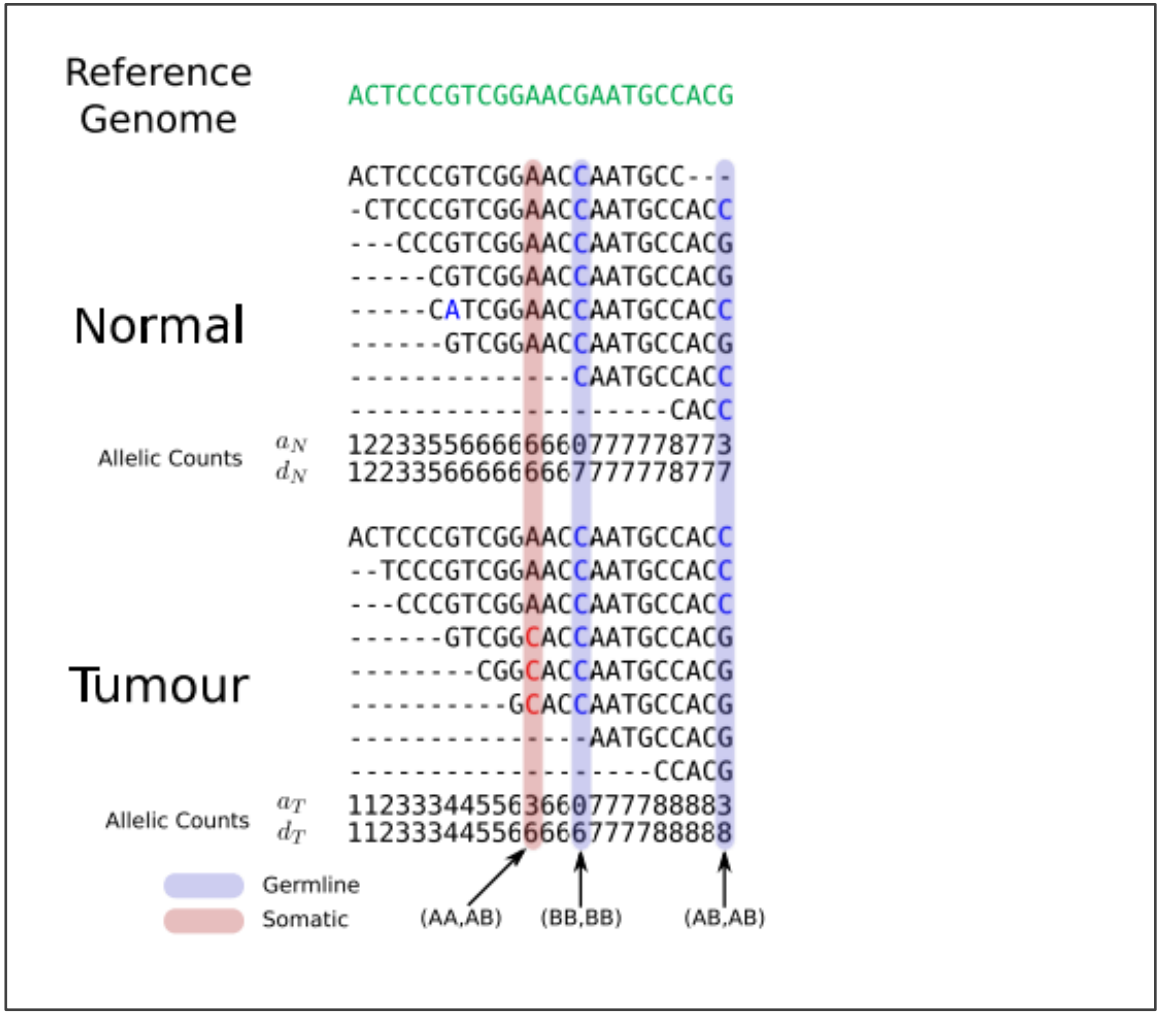


- *Germline mutations* (all normal cells) – occur in gametes and can be passed onto offspring (every cell in the entire organism will be affected)

- *Somatic mutations* (tumor cells) – occur in a single body cell and cannot be inherited (only tissues derived from mutated cell are affected)

- In cancer studies, we want to know if they are germline or somatic mutations in order to better understand cancer biology, diagnose cancer and improve cancer therapies

Somatic SNP calling

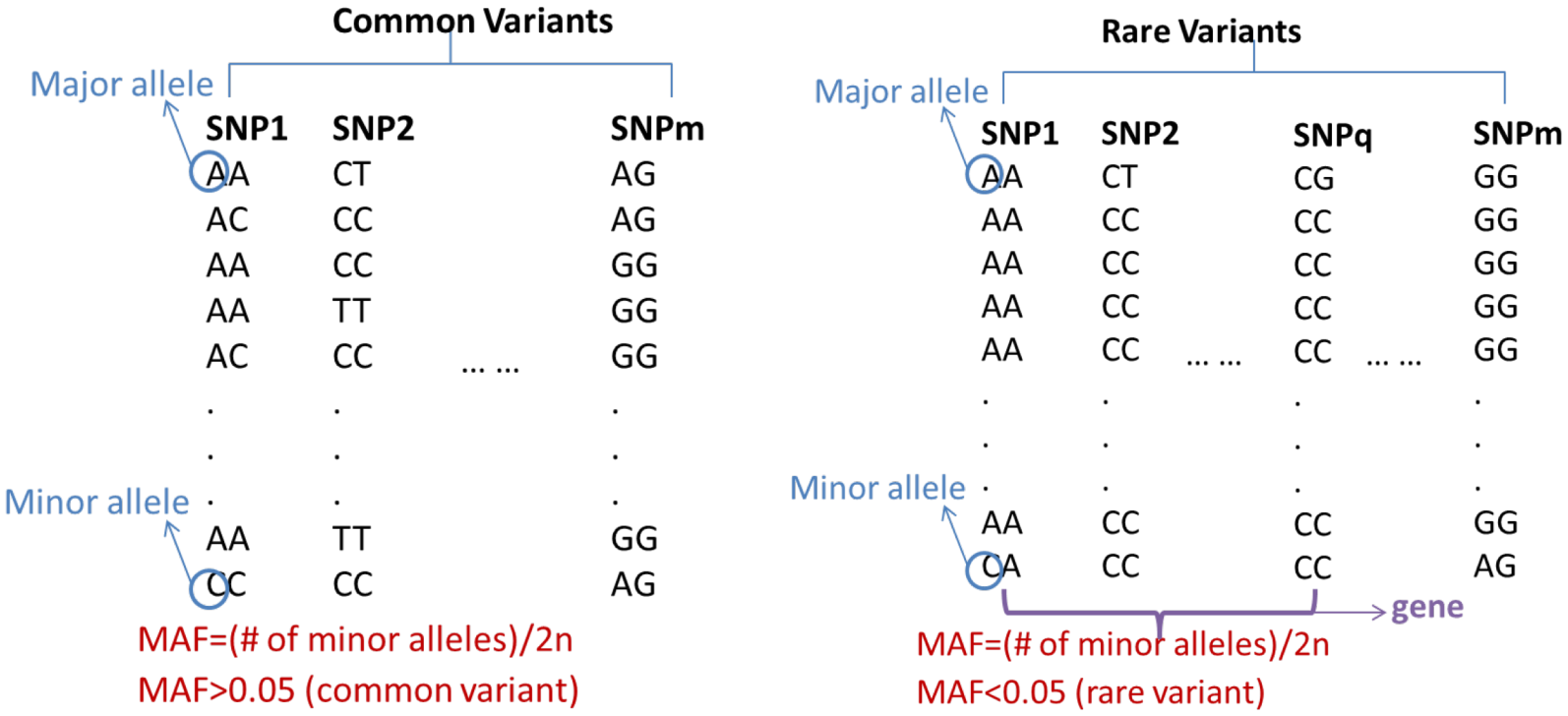


Rare Variant Tests

- 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

Common VS Rare

- Genotypes:
  - Common variants (e.g. MAF≥0.05): single marker test;
  - Rare variants (e.g. MAF<0.05): test at gene level



- **Only subset of functional elements include common variants**
- **Rare variants are more numerous and thus will point to additional loci**

Single Marker Test for Rare Variant

- Rare variants are hard to detect
- Rare variants have low frequency that makes single marker test less powerful
- Rare causal SNPs are hard to identify even with large effect size

Single Marker 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 (MAF) = 0.1, 0.01, 0.001
- Required sample size = 486, 3545, 34322,

Alternatives to Single Marker Test Collapsing Method (Burden Test)

- Group rare variants in the same gene/region
- Score each individual
  - Presence or absence of rare copy $X_j = \begin{cases} 1 & \text{rare variants present} \\ 0 & \text{otherwise} \end{cases}$
  - Weight each variant
- Use individual score as a new “genotype”
- Test in a regression framework

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

Burden Test

SNP1 SNP2 SNP3 SNP4 SNP5 Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

New “Genotype” = SNP1 + SNP2 + … + SNP5

New “Genotype” = W1*SNP1 + W2*SNP2 + … + W5*SNP5

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 Variants

	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 reduces power
- Still better than single marker test

Challenges

- Assume all causal rare variants have the same effect direction
- It is hard to separate causal and null SNPs
  - Including all rare variants will dilute the true signals
- Assume the effect size of each rare variant the same

Sequence Kernel Association Test (SKAT)

Let there be n subjects with q genetic variants. The $n \times 1$ vector of the quantitative trait $\boldsymbol{y}$ follows a linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $\mathbf{G}$ is an $n \times q$ genotype matrix for the q rare genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the q genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error.

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_E^2\mathbf{I}\right)$$

where $\mathbf{W}$ is a predefined $q \times q$ diagonal weight matrix for each variant

Thus, the null hypothesis H_0: $\boldsymbol{\gamma} = 0$ is equivalent to H_0: $\tau = 0$, which can be tested with a variance component score test.

Wu, M. C.*, Lee, S.*, Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011) Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). American Journal of Human Genetics , 89, 82-93.

Sequence Kernel Association Test (SKAT)

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\varepsilon}$$

$$\text{Var}(\mathbf{y}) = \tau\mathbf{GWG}' + \sigma_E^2\mathbf{I}$$

SKAT test statistic: $Q = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\underbrace{\mathbf{GWG}'}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$

- Called "**kernel**".
- Linear combination used here. Could be more flexible form.

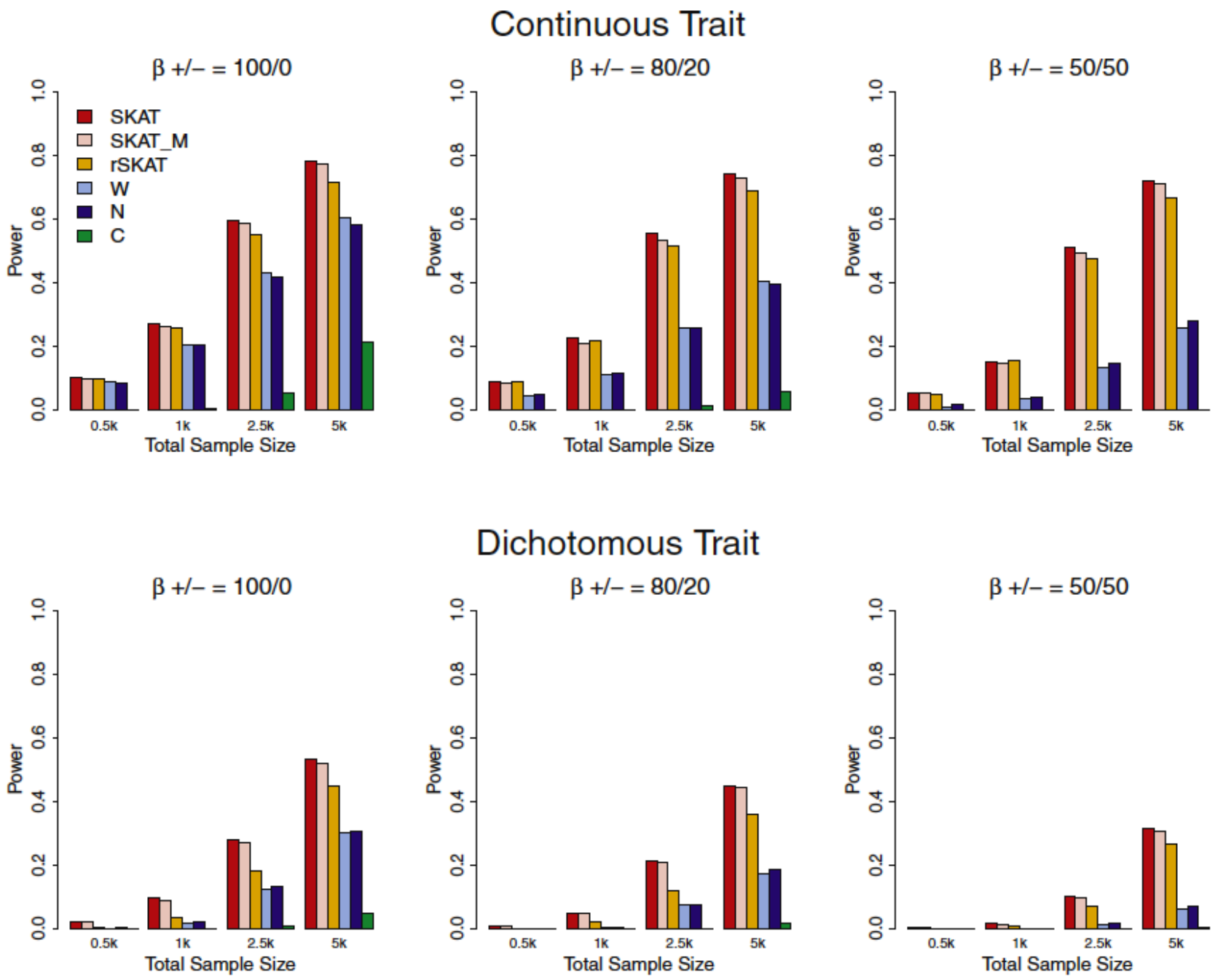
where the parameters are estimated under H_0 (i.e., H_0: $\tau = 0$)

Thus, under H_0: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

$$\widehat{\boldsymbol{\Sigma}} = \hat{\sigma}_E^2\mathbf{I}$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

Q follows a mixture of Chi-square distribution: $Q \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2$


Continuous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
SKAT
SKAT_M
rSKAT
W
N
C
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.5k
1k
2.5k
5k
Total Sample Size
Dichotomous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
Power
Total Sample Size

Discussion

- Analysis of rare variants is (was?) an active research area
- Weight for each SNP is the key
- What to do if the samples are related
- Most tests reply on permutation
  - Computationally intensive

Reference

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

- Nielsen R, Paul JS et al. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*

- Li Y, Chen W et al. (2012) Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data. *Statistics in Biosciences.*

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

- Chen W, Li B et al. (2013) Genotype calling and haplotyping in parent-offspring trios. *Genome Research.*

Reference

- http://genome.sph.umich.edu/wiki/Rare_variant_tests
- 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
- Wu M, Lee S, et al. (2011) *Am J Hum Genet*