Analysis of Short Read Sequences

Gonçalo Abecasis

University of Michigan School of Public Health

Genomewide Association Studies

- Survey 500,000 SNPs in a large sample

- An effective way to skim the genome and …
- … find common variants associated with a trait of interest

- Rapid increase in number of known complex disease loci
  - For example, ~50 genes now identified for type 2 diabetes.

- Techniques for genetic analysis are changing rapidly
  - What are some of the potential benefits and challenges for replacing genotyping with sequencing in complex trait studies?

Questions that Might Be Answered With Complete Sequence Data…

- What is the contribution of each identified locus to a trait?
  - Likely that multiple variants, common and rare, will contribute

- What is the mechanism? What happens when we knockout a gene?
  - Most often, the causal variant will not have been examined directly
  - Rare coding variants will provide important insights into mechanisms

- What is the contribution of structural variation to disease?
  - These are hard to interrogate using current genotyping arrays.

- Are there additional susceptibility loci to be found?
  - Only subset of functional elements include common variants …
  - Rare variants are more numerous and thus will point to additional loci

Shotgun Sequence Reads

ACTGGTCGATGCTAGCTGATAGCTAGCTA

GCTGATGAGCCCGATCGCTGCTAGCTCG

AGCTGATAGCTAGCTAGCTGATGAGCCCGA

GAGCCCGATCGCTGCTAGCTCGACG

- Typical short read might be <25-100 bp long and not very informative on its own
- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences
- Sequencing errors are much more common than true variation

Base Qualities

Short Read Sequence

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Short Read Base Qualities

30.30.28.28.29.27.30.29.28.25.24.26.27.24.24.23.20.21.22.10.25.25.20.20.18.17.16.15.14.14.13.12.10

- Each base is typically associated with a quality value
- Measured on a “Phred” scale, which was introduced by Phil Green for his Phred sequence analysis tool

$$BQ = -\log_{10}(\epsilon)\,, where\ \epsilon\ is\ the\ probability\ of\ an\ error$$

Read Alignment

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Short Read (30-100 bp)

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome (3,000,000,000 bp)

- The first step in analysis of human short read data is to align each read to genome, typically using a hash table based indexing procedure

- This process now takes no more than a few hours per million reads …

- Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies

Read Alignment – Food for Thought

- Typically, all the words present in the genome are indexed to facilitate read mapping …
  - What are the benefits of using short words?
  - What are the benefits of using long words?

- How matches do you expect, on average, for a 10-base word?
  - Do you expect large deviations from this average?

Mapping Quality

- Measures the confidence in an alignment, which depends on:
  - Size and repeat structure of the genome
  - Sequence content and quality of the read
  - Number of alternate alignments with few mismatches

- The mapping quality is usually also measured on a “Phred” scale

- Idea introduced by Li, Ruan and Durbin (2008) *Genome Research* **18:**1851-1858

Mapping Quality Definition

- Given a particular alignment **A**, we can calculate

$$
\begin{aligned}
P(\mathbf{S}|\mathbf{A},\mathbf{Q}) &= \\
&= \prod_{i}^{n} P(\mathbf{S}_i|\mathbf{A},\mathbf{Q}) \\
&= \prod_{i}^{n} \left\{\frac{1}{3}10^{-\mathbf{Q}_i/\mathbf{10}}\right\}^{I(S_i\ mismatch|\mathbf{A})} \left\{1-10^{-\mathbf{Q}_i/\mathbf{10}}\right\}^{I(S_i\ match|\mathbf{A})}
\end{aligned}
$$

- Then, the mapping quality is:

$$
MQ(\mathbf{S}|\mathbf{A}_{best},\mathbf{Q}) = \frac{P(\mathbf{S}|\mathbf{A}_{best},\mathbf{Q})}{\sum_i P(\mathbf{S}|\mathbf{A}_i,\mathbf{Q})}
$$

- In practice, summing over all possible alignments is too costly and this quantity is approximated (for example, by summing over the most likely alignments).

Refinements to Mapping Quality

- In their simplest form, mapping qualities apply to the entire read
- However, in gapped alignments, uncertainty in alignment can differ for different portions of the read
  - For example, it has been noted that many wrong variant calls are supported by bases near the edges of a read
- Per base alignment qualities were introduced to summarize local uncertainty in the alignment

Per Base Alignment Qualities

Short Read

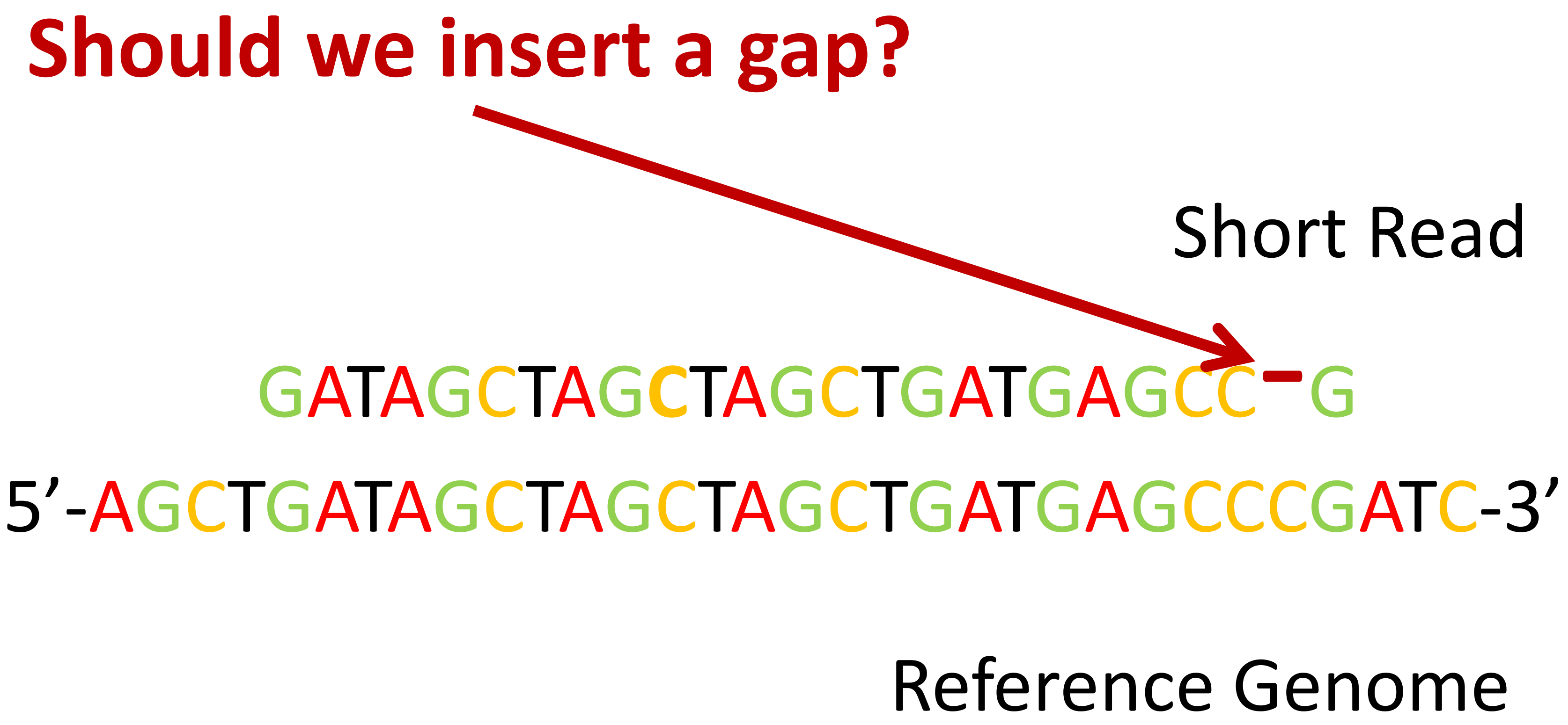
GATAGCTAG**C**TAGCTGATGA GCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities



Should we insert a gap?

Short Read

GATAGCTAGCTAGCTGATGAGCC-G

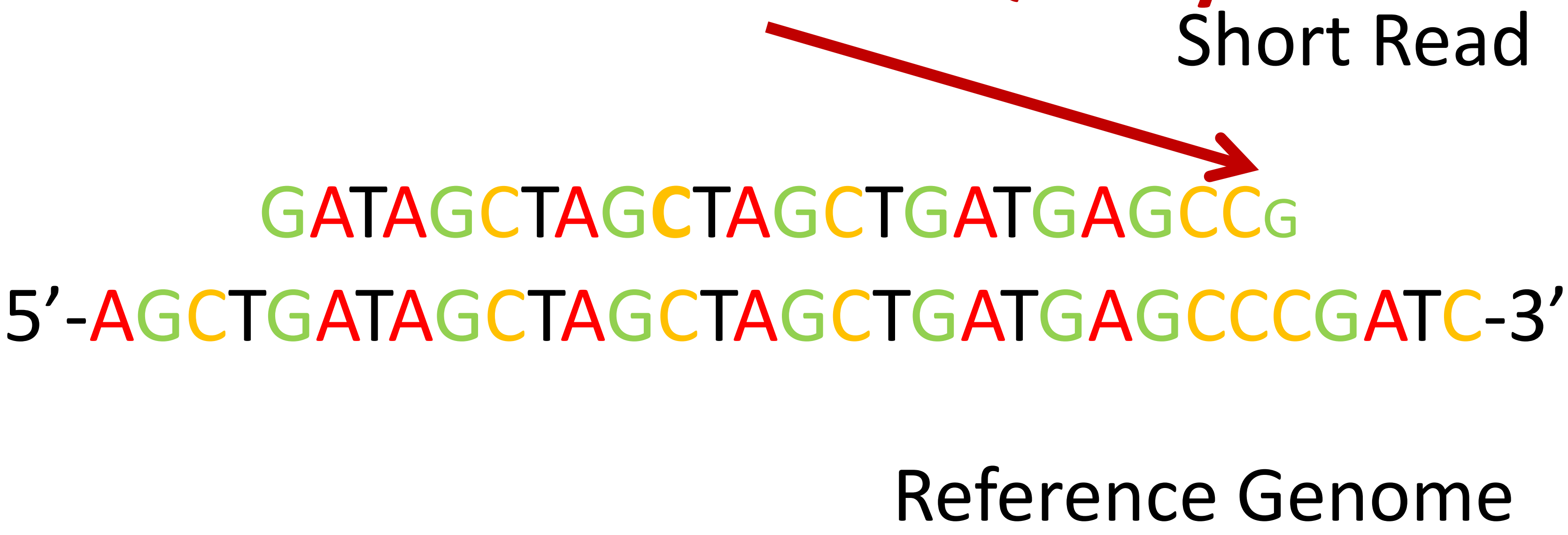
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality



Short Read

GATAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Shotgun Sequence Data

★

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 1.0

P(reads|A/C, read mapped)= 1.0

P(reads|C/C, read mapped)= 1.0

Possible Genotypes

Shotgun Sequence Data

★

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

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

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

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

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 0.01

P(reads|A/C, read mapped)= 0.50

P(reads|C/C, read mapped)= 0.99

Possible Genotypes

Shotgun Sequence Data

★

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 0.0001

P(reads|A/C , read mapped)= 0.25

P(reads|C/C , read mapped)= 0.98

Possible Genotypes

Shotgun Sequence Data

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A , read mapped)= 0.000001

P(reads|A/C , read mapped)= 0.125

P(reads|C/C , read mapped)= 0.97

Possible Genotypes

Shotgun Sequence Data

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A , read mapped)= 0.00000099

P(reads|A/C , read mapped)= 0.0625

P(reads|C/C , read mapped)= 0.0097

Possible Genotypes

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A , read mapped)= 0.00000098

P(reads|A/C , read mapped)= 0.03125

P(reads|C/C , read mapped)= 0.000097

Possible Genotypes

Shotgun Sequence Data

★

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A, read mapped)= 0.00000098

P(reads|A/C, read mapped)= 0.03125

P(reads|C/C, read mapped)= 0.000097

Combine these likelihoods with a prior information to assign a genotype.

Ingredients That Go Into Prior

- Most sites don’t vary
  - P(non-reference base) ~ 0.001

- When a site does vary, it is usually heterozygous
  - P(non-reference heterozygote) ~ 0.001 * 2/3
  - P(non-reference homozygote) ~ 0.001 * 1/3

- Mutation model
  - Transitions account for most variants (C↔T or A↔G)
  - Transversions account for minority of variants

From Sequence to Genotype: Individual Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

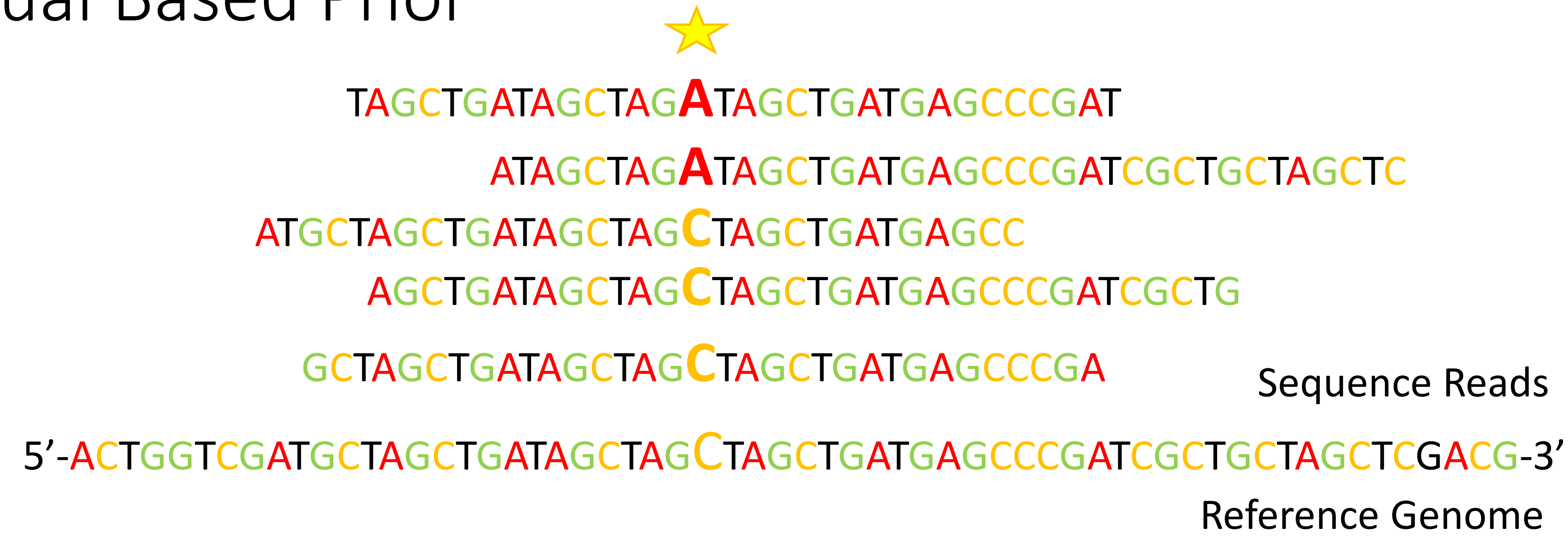
5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A)** = 0.00034	Posterior(A/A) = <.001
P(reads|A/C)= 0.03125	**Prior(A/C)** = 0.00066	Posterior(A/C) = 0.175
P(reads|C/C)= 0.000097	**Prior(C/C)** = 0.99900	Posterior(C/C) = 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

From Sequence to Genotype: Individual Based Prior



TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.00034	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.00066	**Posterior(A/C) =** 0.175
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.99900	**Posterior(C/C) =** 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

Shotgun Sequence Data

Haplotype Based Prior

★

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

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

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 20% of similar chromosomes carry allele A.

Shotgun Sequence Data

Haplotype Based Prior

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

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

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 20% of similar chromosomes carry allele A.

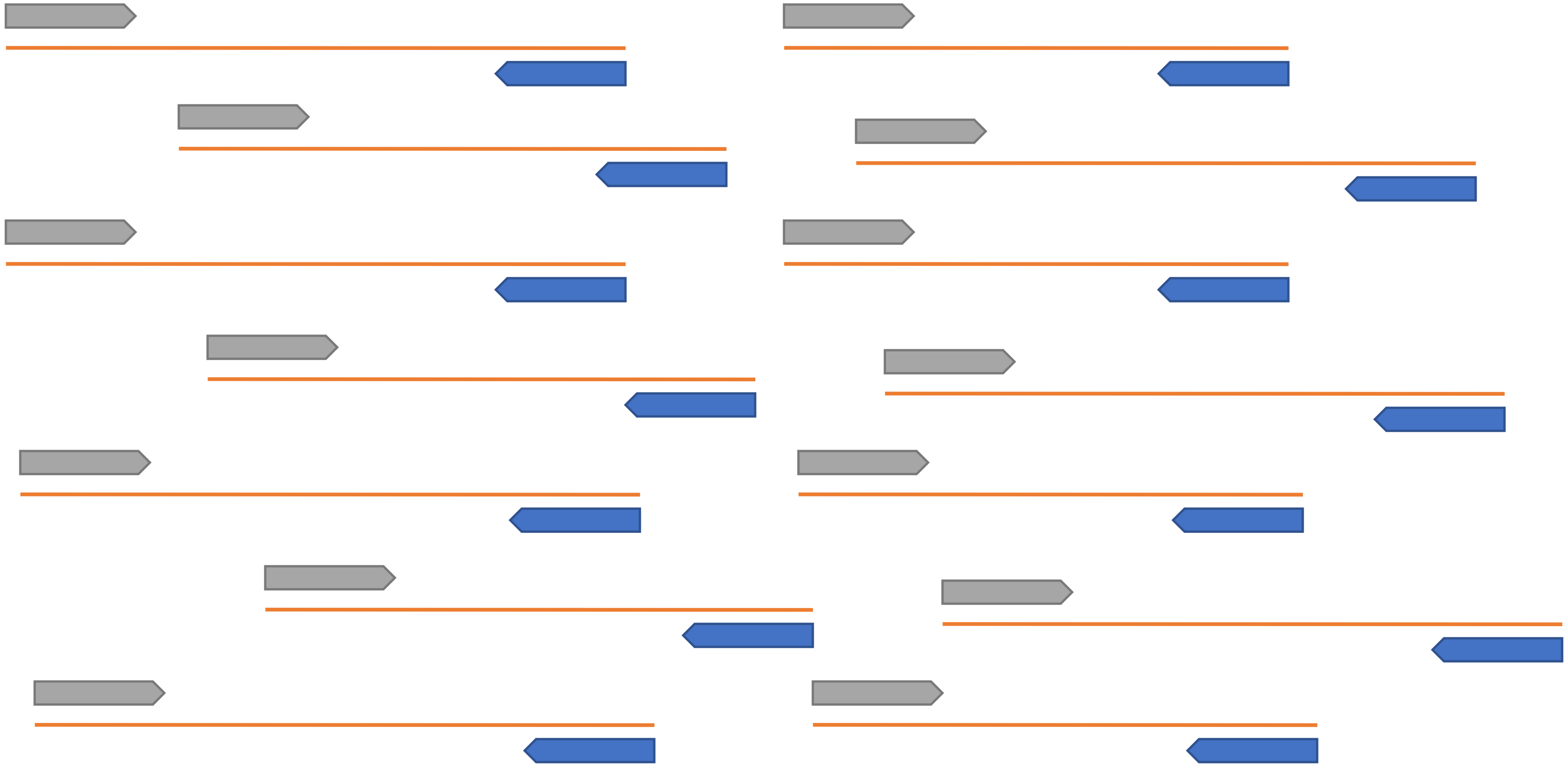
Sequence Based Genotype Calls

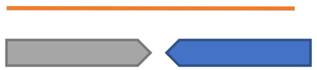
- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced

Paired End Sequencing



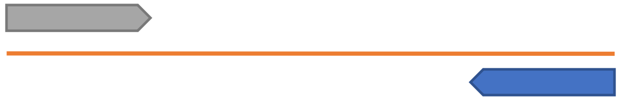


Population of DNA fragments of known size (mean + stdev)

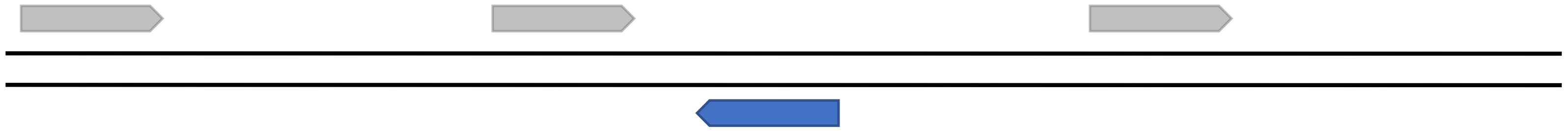
Paired end sequences

Paired End Sequencing

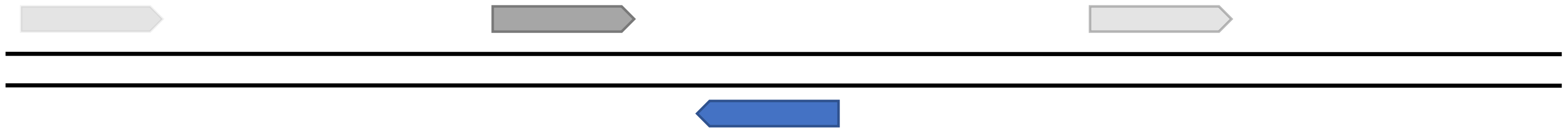
Paired Reads



Initial alignment to the reference genome



Paired end resolution



Detecting Structural Variation

- Read depth
  - Regions where depth is different from expected
    - Expectation defined by comparing to rest of genome …
    - … or, even better, by comparing to other individuals

- Split reads
  - If reads are longer, it may be possible to find reads that span the structural variation

- Discrepant pairs
  - If we find pairs of reads that appear to map significantly closer or further apart than expected, could indicate an insertion or deletion

  - For this approach, “physical coverage” which is the sum of read length and insert size is key

- De Novo Assembly

How Much Variation is There?

- An average genome includes:
    - 3.6M SNPs
    - 350K indels
    - 700 large deletions

- Numbers are probably underestimates …
- … some variants are hard to call with short reads

- 1000 Genomes Project (2012) *Nature* **491**:56-65

How Much Variation is There?

SNPs Per Individual in Gene Regions

Primarily European Ancestry

European Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	10127	6174	3953	38.2	5.10
MISSENSE	8541	5184	3357	72.2	2.16
NONSENSE	86	57	29	2.1	1.70

Primarily African Ancestry

African Ancestry	# SNP	# HET	# ALT	# Singletons	Ts/Tv
SILENT	12028	8038	3990	53.2	5.19
MISSENSE	9870	6502	3367	94.2	2.16
NONSENSE	92	57	35	2.4	1.57

NHLBI Exome Sequencing Project

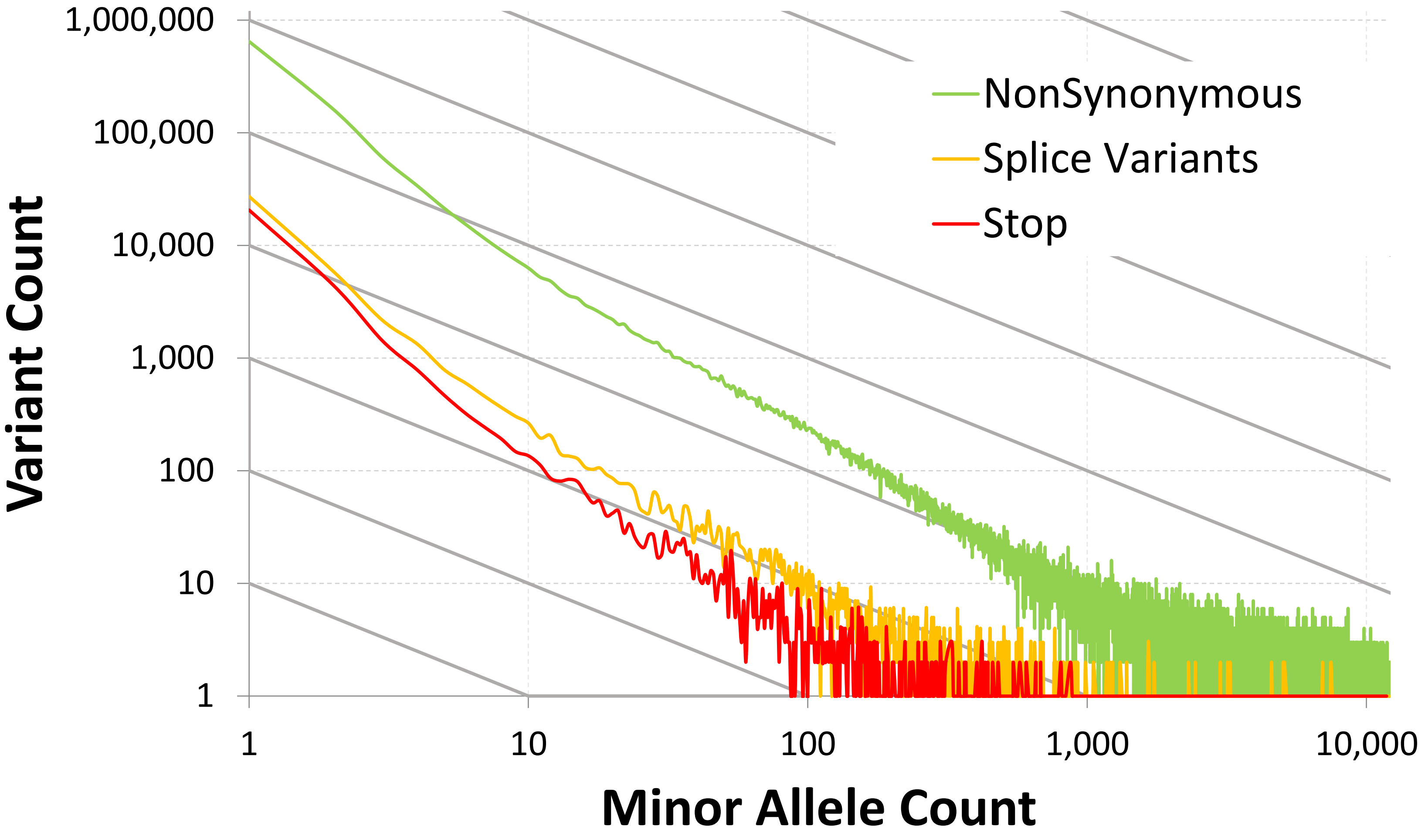
Lots of Rare Functional Variants to Discover

SET	# SNPs	Singletons	Doubletons	Tripletons	>3 Occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46,740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Non-Syn / Syn Ratio		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

There is a very large reservoir of extremely rare, likely functional, coding variants.
(Results above correspond to approximately 5,000 individuals)

NHLBI Exome Sequencing Project

Allele Frequency Spectrum (After Sequencing 12,000+ Individuals)



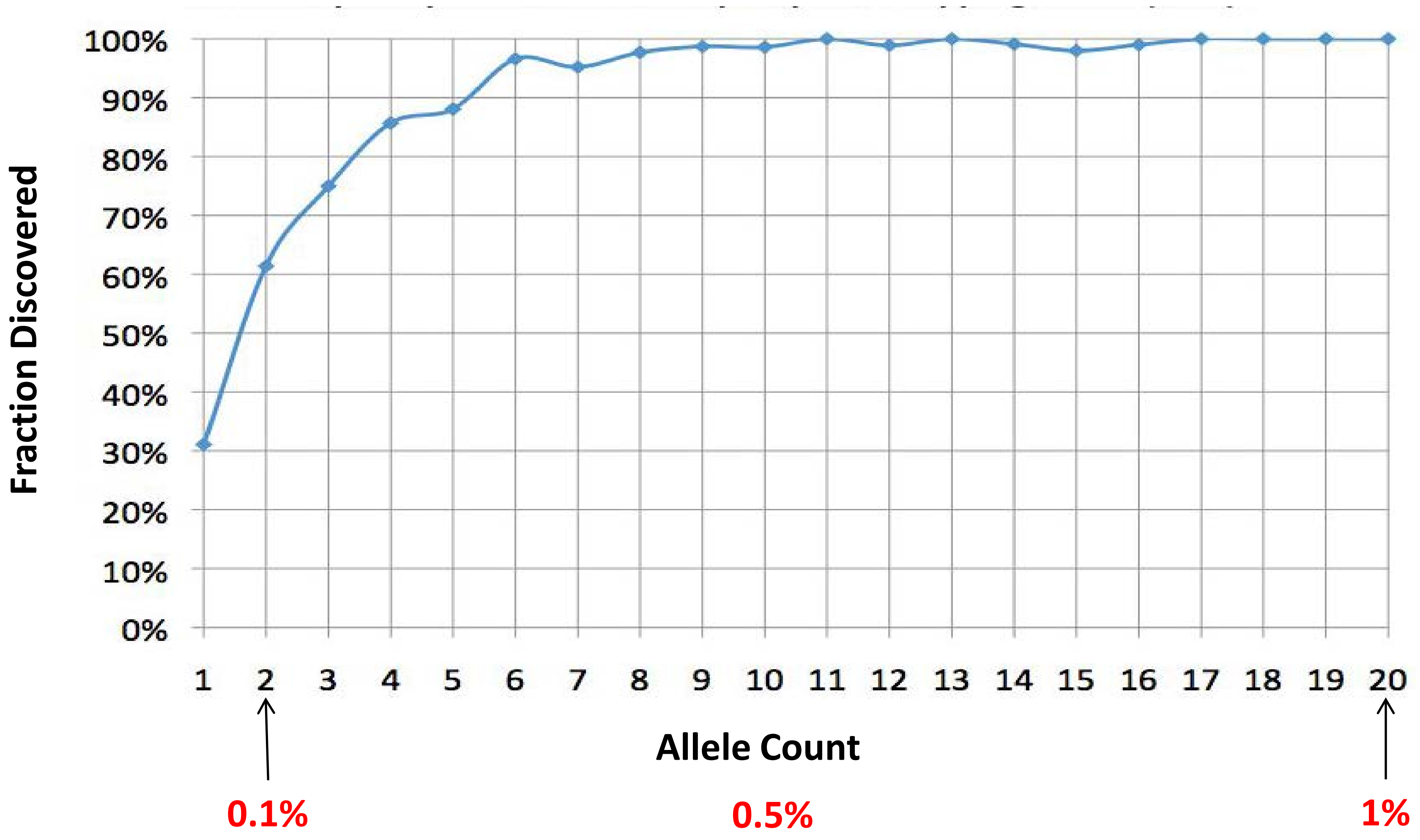
http://genome.sph.umich.edu/wiki/Exome_Chip_Design

Genome Scale Approaches For Studying Rare Variation

- 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 pass whole genome sequencing
  - Can be applied to moderate numbers of samples
  - SNPs and indels present in multiple individuals

Empirical Variant Discovery Power

1000 Genomes Project, 4x Low Pass Sequencing

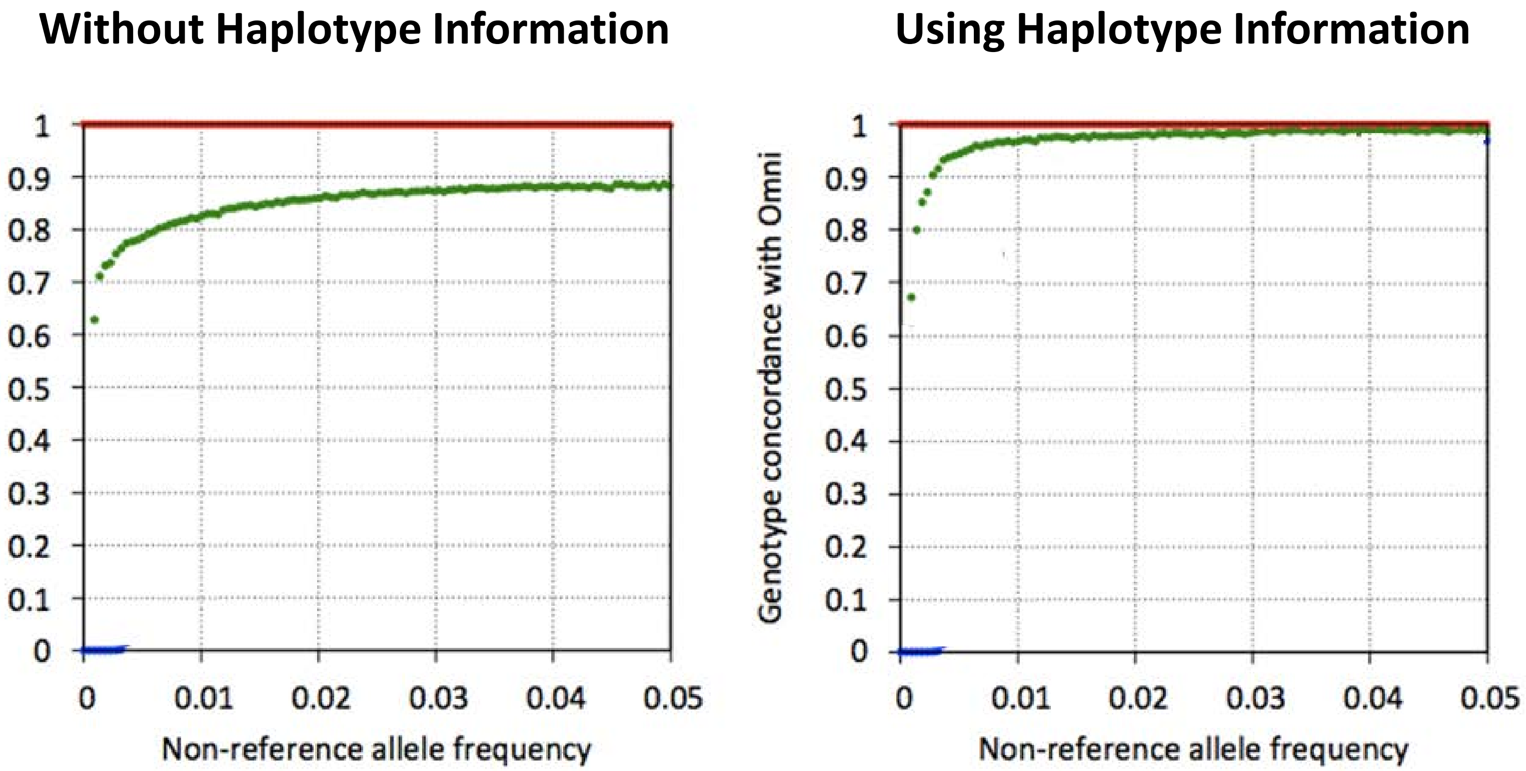


Fraction of variants discovered in low pass sequencing, estimated by comparison with External data.

Hyun Min Kang

Empirical Evaluation of Haplotype Callers

1000 Genomes Project, 4x Low Pass Sequencing



Homozygote Sites, Heterozygote Sites

What Was the Optimal Model for Analysis of 1000 Genomes Pilot Data?

1000 Genomes Call Set (CEU)	Homozygous Reference Error	Heterozygote Error	Homozygous Non-Reference Error
Broad	0.66	4.29	3.80
Michigan	0.68	3.26	3.06
Sanger	1.27	3.43	2.60
Majority Consensus	0.45	2.05	2.21

- Pilot analyzed with different haplotype sharing models
  - Sanger (QCALL), Michigan (MaCH/Thunder), Broad (BEAGLE)
  - Consensus of the three callers clearly bested single callers

Given Fixed Capacity,
Should We Sequence Deep or Shallow?

	.5 – 1%	**1 – 2%**	**2-5%**
400 Deep Genomes (30x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	100%	100%	100%
Effective N	400	400	400
3000 Shallow Genomes (4x)			
Discovery Rate	100%	100%	100%
Het. Accuracy	90.4%	97.3%	98.8%
Effective N	2406	2758	2873

Li et al, *Genome Research,* 2011

Design A Whole Genome Sequencing Study in Sardinia

Gonçalo Abecasis

David Schlessinger

Francesco Cucca

SardiNIA Whole Genome Sequencing

- 6,148 Sardinians from 4 towns in the Lanusei Valley, Sardinia
  - Recruited among population of ~9,841 individuals
  - Sample includes >34,000 relative pairs

- Measured ~100 aging related quantitative traits

- Original plan:
  - Sequence >1,000 individuals at 2x to obtain draft sequences
  - Genotype all individuals, impute sequences into relatives

How Is Sequencing Progressing?

- NHGRI estimates of sequencing capacity and cost …
  - Since 2006, for fixed cost …
  - … ~4x increase in sequencing output per year

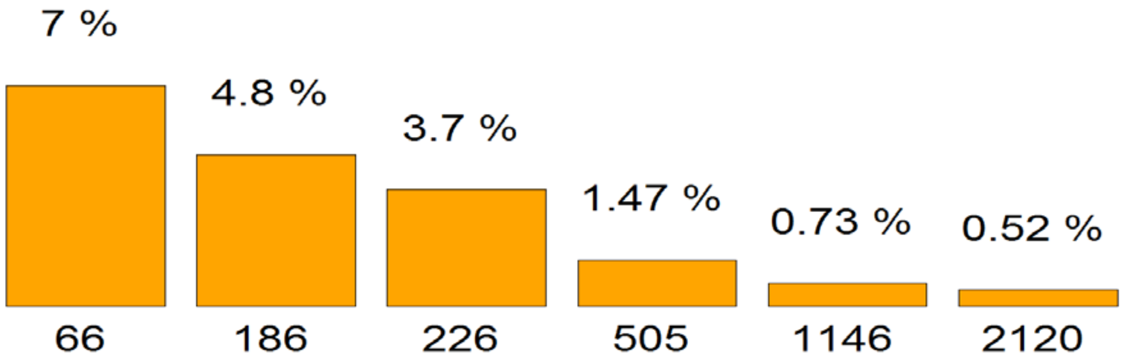
- In our own hands…
  - Mapped high quality bases
  - March 2010: ~5.0 Gb/lane
  - May 2010: ~7.5 Gb/lane
  - September 2010: ~8.6 Gb/lane
  - January 2011: ~16 Gb/lane
  - Summer 2011: ~45 Gb/lane

- Other small improvements
  - No PCR libraries increase genome coverage, reduce duplicate rates

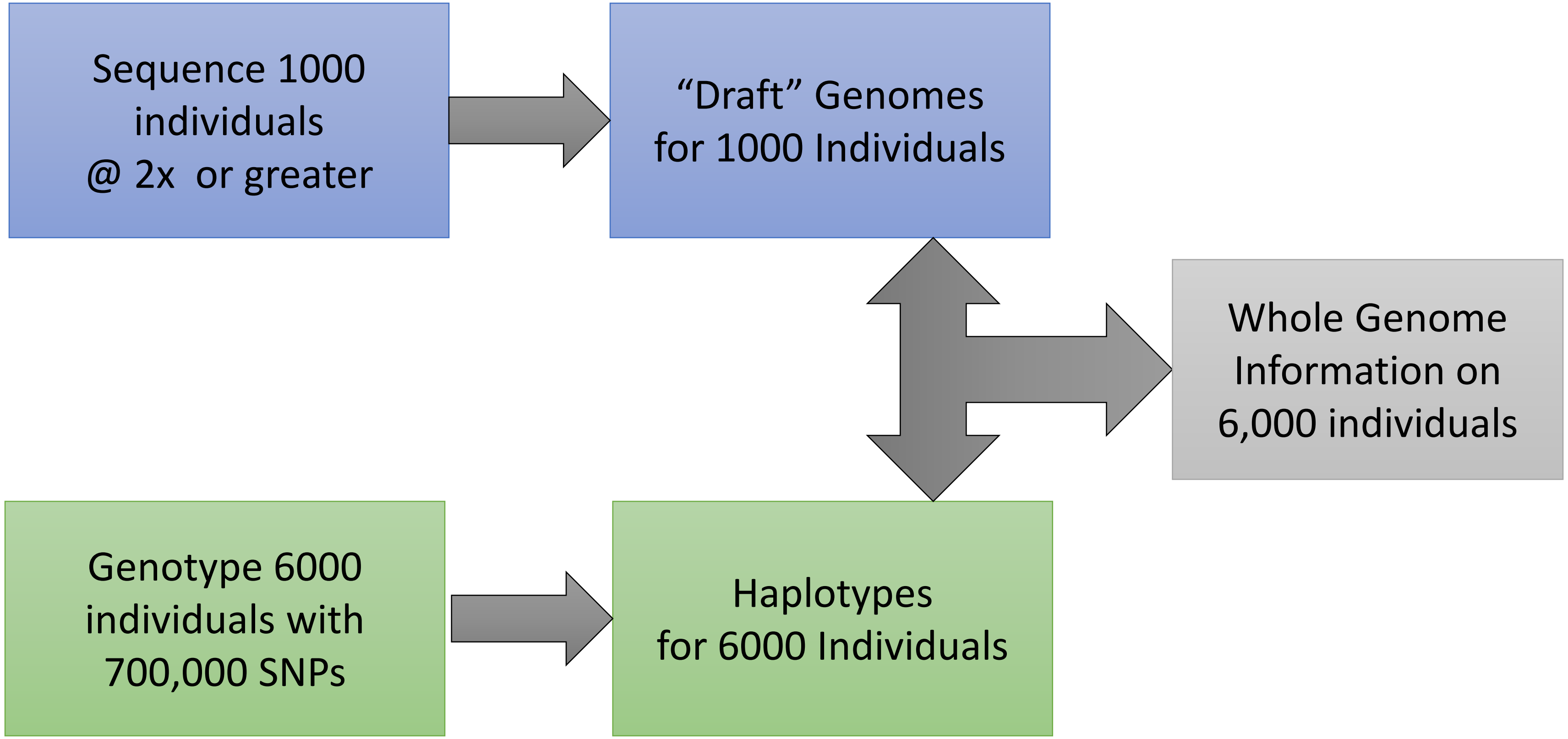
Fabio Busonero, Andrea Maschio

As more samples are sequenced, Accuracy increases

Heterozygous Mismatch Rate (in %)



Design



Sequence 1000 individuals @ 2x or greater

“Draft” Genomes for 1000 Individuals

Whole Genome Information on 6,000 individuals

Genotype 6000 individuals with 700,000 SNPs

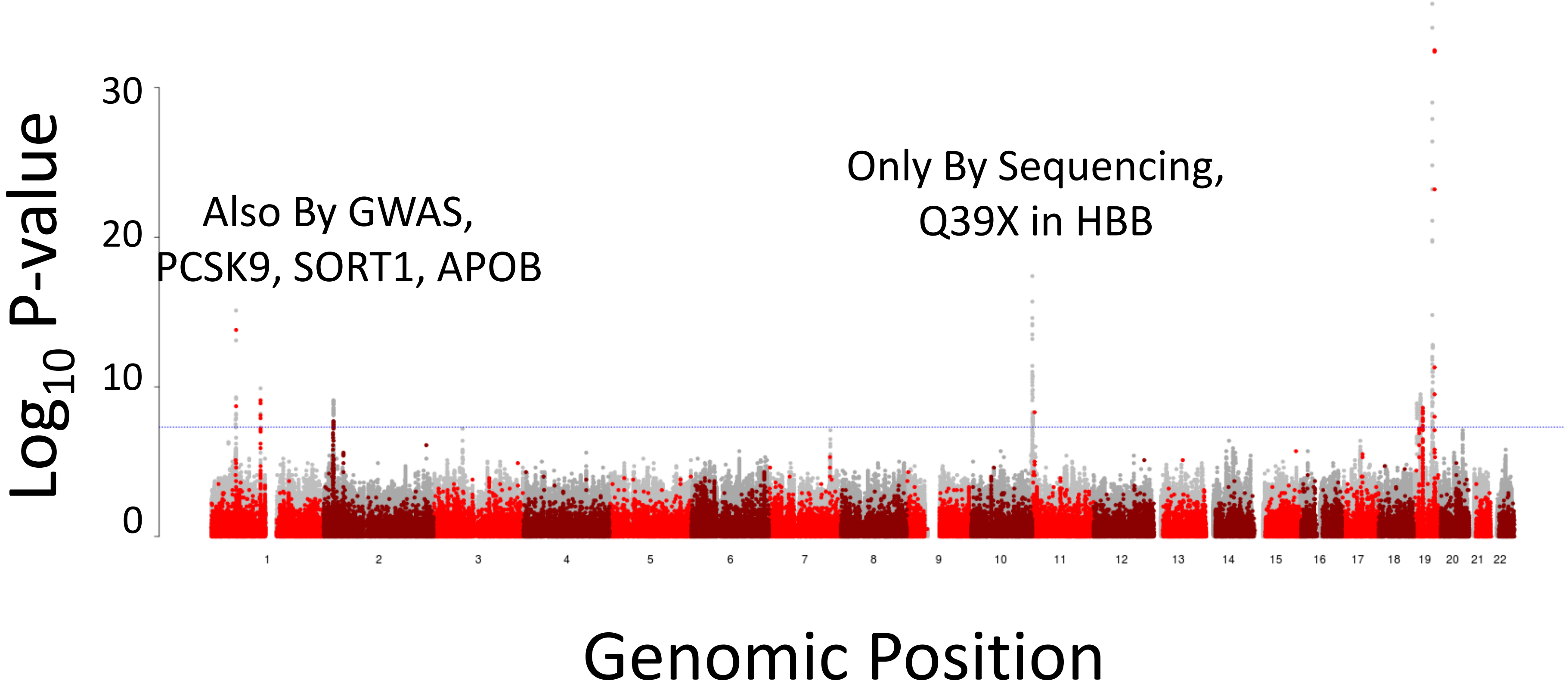
Haplotypes for 6000 Individuals

Sardinian Haplotypes Are Great For Imputation In Sardinia

Reference Panel		SNP Imputation Accuracy (r^2) IN SARDINIA		
Population	**Size**	**MAF 1-3%**	**MAF 3-5%**	**MAF >5%**
1000G (Worldwide)	563	0.75	0.88	0.94
Sardinia	508	0.90	0.95	0.97
Sardinia	831	0.92	0.97	0.98
Sardinia	1488	0.95	0.98	0.99

Data: Sardinia data set; chr20; Imputation-panel: Affy1M; Evaluation-panel: Metabochip

What Do We See Genomewide?
LDL Cholesterol



Also By GWAS,
PCSK9, SORT1, APOB

Only By Sequencing,
Q39X in HBB

Also By GWAS,
LDLR, APOE

LDL Genetics In Lanusei Valley, Sardinia, Current Sequenced Based View

Locus	Variants	MAF	Effect Size (SD)	H^2
HBB	**Q39X**	.04	0.90	8.0%??
APOE	R176C, C130R	.04, .07	0.56, 0.26	3.3%
PCSK9	R46L, rs2479415	.04, .41	0.38, 0.08	1.2%
LDLR	rs73015013, **V578R**	.14, .005	0.16, 0.62	1.2%
SORT1	rs583104	.18	0.15	0.6%
APOB	rs547235	.19	0.19	0.5%

- Most of these variants are important across Europe, extensively studied.
- **Q39X** variant in HBB is especially enriched in Sardinia.
- **V578R** in LDLR is a Sardinia specific variant, particularly common in Lanusei.

Tools for Sequence Analysis

Useful Pointers

MAQ and BWA

- Two popular read mappers developed by Heng Li and Richard Durbin at Sanger
- MAQ uses short sequences to build an index; it is relatively slow but very accurate
- BWA uses a special technique to index much longer sequences; it is much faster and nearly as accurate
- http://maq.sourceforge.net/index.shtml

SAM/BAM format and SAMTOOLS

- Generic format for storing aligned reads
    - Sequence, base quality, indels, mate information

- SAM is a plain text format, easy to generate
- BAM is an indexed binary format, compact and fast

- Very active mailing lists available

- Li et al, *Bioinformatics,* **25:**2078–2079
- http://samtools.sourceforge.net
- http://samtools.sourceforge.net/SAM1.pdf

Picard & GATK

- Set of *java* tools for manipulating SAM/BAM
  - Developed at the Broad

- Particularly useful for:
  - Removing duplicate reads
  - Recalibrating base quality scores
  - Removing variant calls due to artifacts

- http://picard.sourceforge.net
- http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit

VerifyBamID

- Identify contamined samples
  - Contamination is surprisingly common in short read data
  - Contamination, if ignored, will result in greatly degraded genotypes

- Contamination can be estimated by comparing sequence data to known genotypes or using only sequence data

- http://genome.sph.umich.edu/wiki/VerifyBamId

UMAKE / GotCloud

- Pipelines for processing sequence data

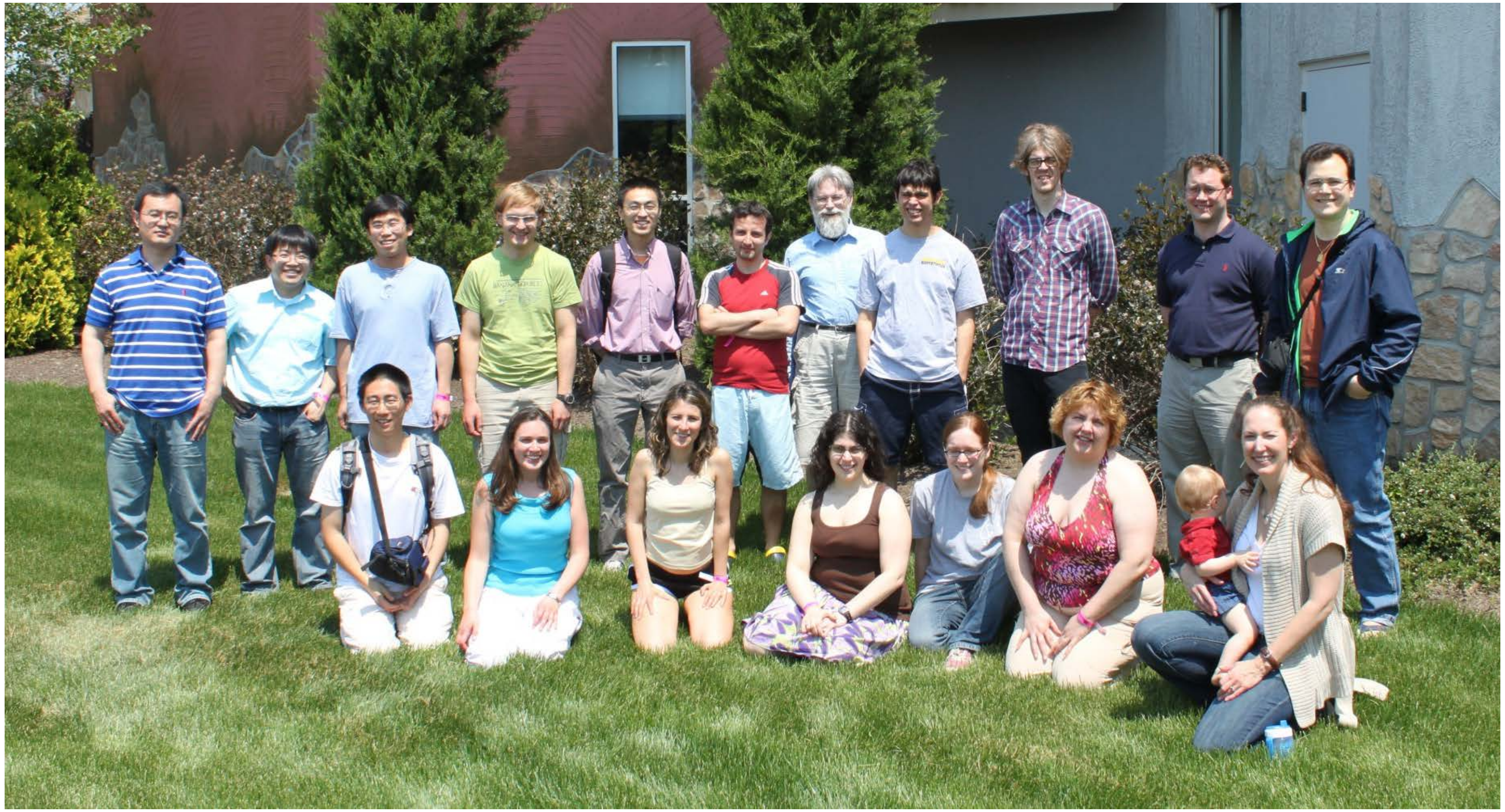
- Glue together a variety of steps and tools
  - Mapping, scrubbing of alignments, variant calling and filtering, genotyping

- http://genome.sph.umich.edu/wiki/GotCloud
- http://genome.sph.umich.edu/wiki/UMAKE

LASER:
Locating Ancestry from Sequence Reads

- Tool for estimating ancestry of a sequenced sample
- Uses reference set of genotyped samples to establish PCA coordinates
- Can handle targeted, exome or whole genome sequence data
- Available from:
  http://genome.sph.umich.edu/wiki/LASER

Acknowledgements



Thank you to the National Institutes of Health (NHGRI, NEI, NHLBI)
for supporting our work.