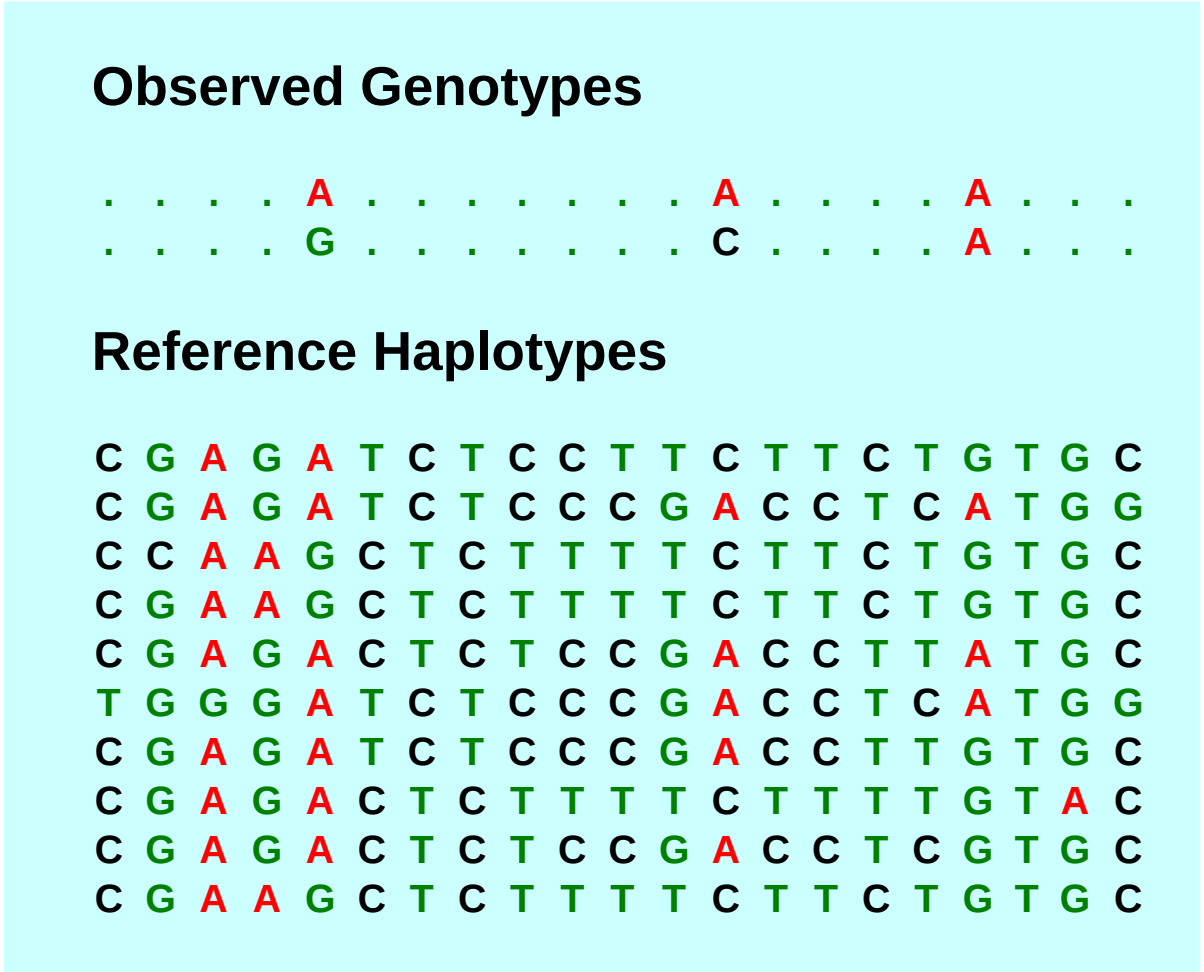
Genotype Imputation

Class Discussion for January 19, 2016

Intuition

- Patterns of genetic variation in one individual …
- … guide our interpretation of the genomes of other individuals
- Imputation uses previously seen combinations of genetic variants to interpret new genomes.

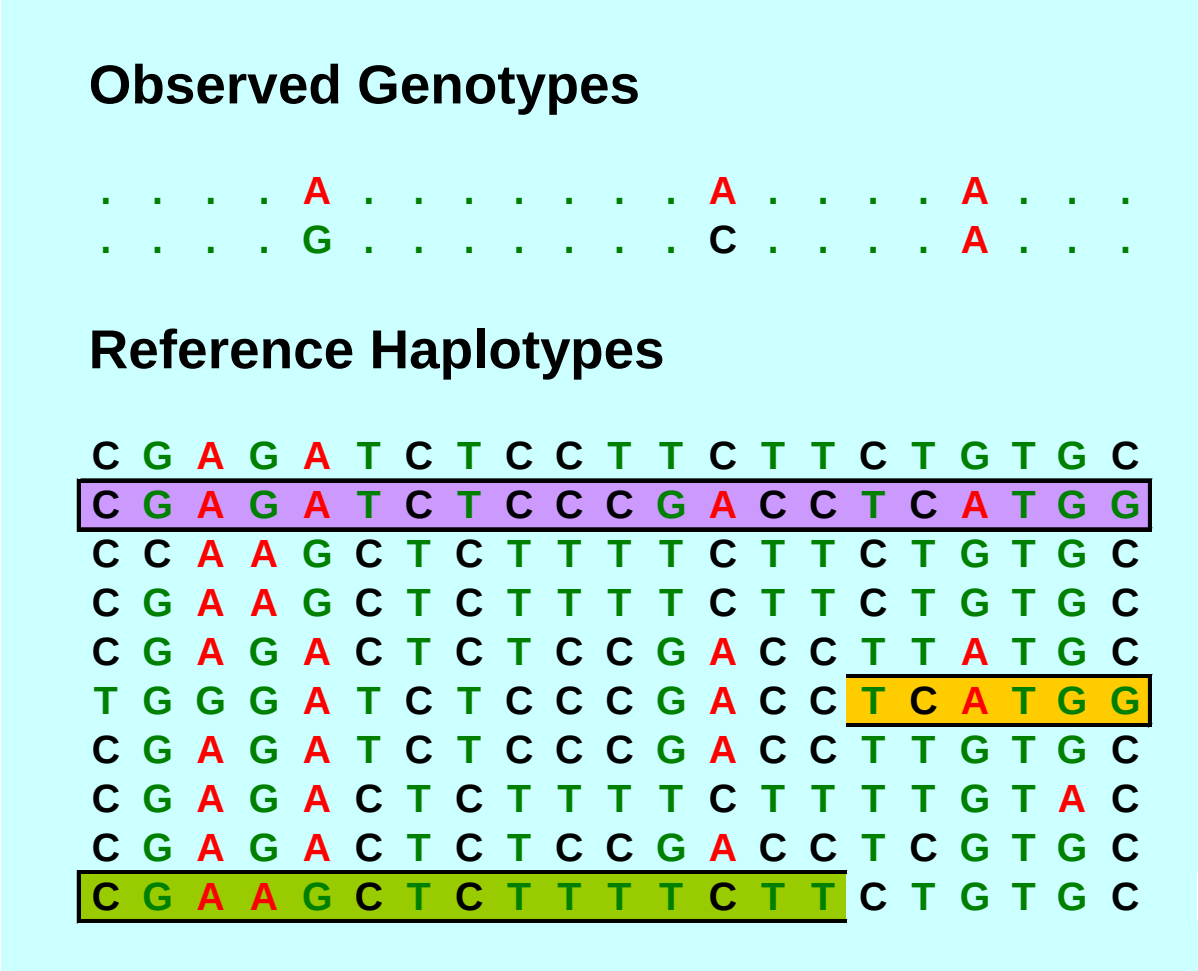
Observed Genotypes



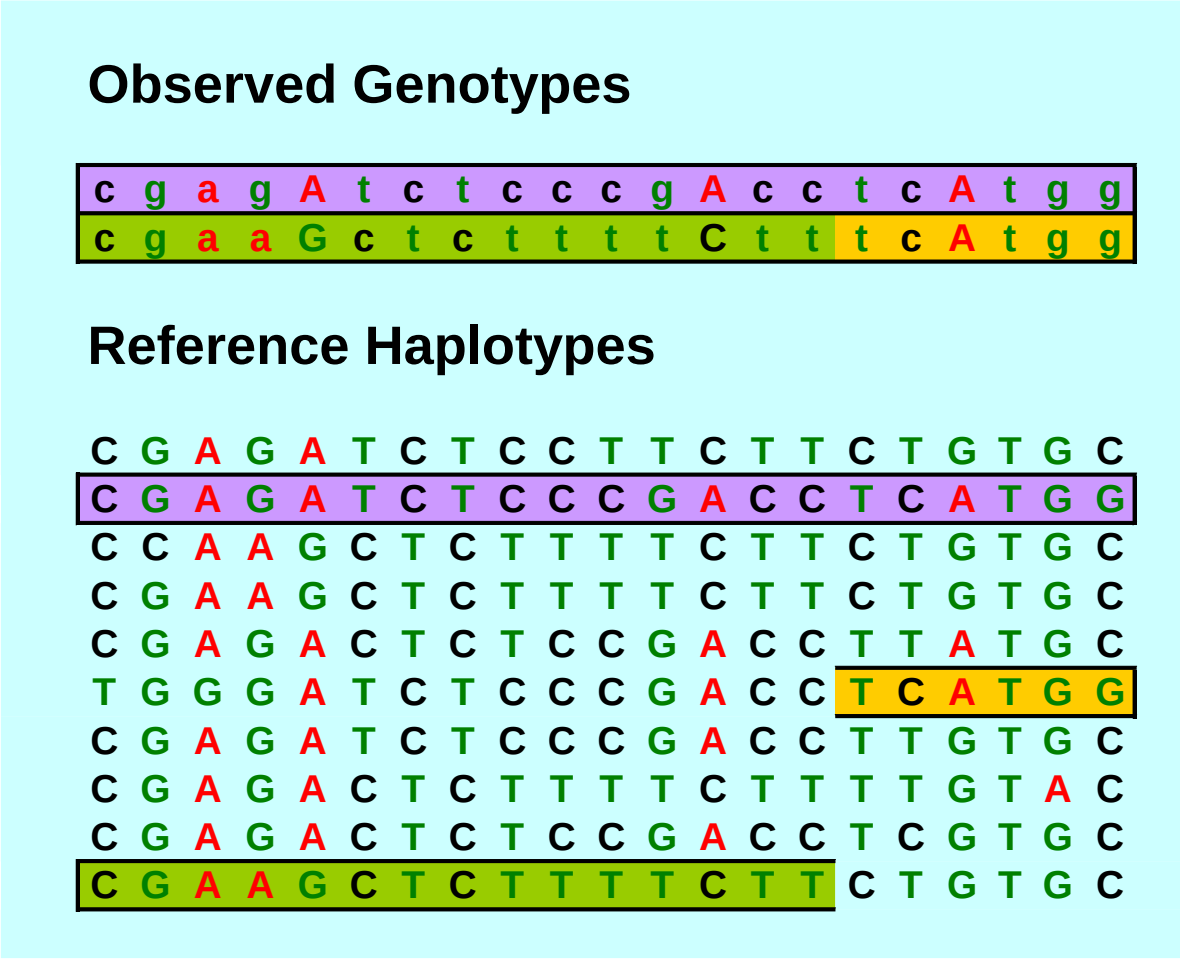
Study Sample
Inexpensive measurements
at 100,000s of markers

Reference Sample
Detailed measurements
of 1,000,000s of markers

Identify Match Among Reference



Fill-in Missing Genotypes



Howie et al (2012) *Nat Genet* 44:955

- Questions…
- Reviewing table 2, can you summarize the factors that influence imputation quality and their relative contributions?
- What struck you most about the paper?

Table 2 Accuracy of different imputation methods and 1000 Genomes reference panels applied to various GWAS data sets

GWAS data set	Imputation method[a]	Reference panel[b]	Imputation accuracy (mean R^2)[c]		
			MAF 1–3%	MAF 3–5%	MAF >5%
GAIN psoriasis	MaCH or **minimac**	60 CEU individuals	0.67	0.76	0.91
(European American; $N = 2,759$)			**0.69**	**0.77**	**0.91**
		283 EUR individuals	**0.73**	**0.78**	**0.92**
		381 EUR individuals	**0.83**	**0.85**	**0.94**
WTCCC2	IMPUTE2	60 CEU individuals	0.66	0.78	0.88
(UK; $N = 2,490$)	(sampling or **pre-phasing**)		**0.65**	**0.77**	**0.87**
		283 EUR individuals	0.77	0.82	0.89
			0.75	**0.81**	**0.88**
		381 EUR individuals	0.84	0.88	0.92
			0.82	**0.86**	**0.91**
WHI	MaCH or **minimac**	60 CEU and 59 YRI individuals	0.51	0.73	0.83
(African-American; $N = 8,421$)			**0.49**	**0.70**	**0.80**
		283 EUR and 172 AFR individuals	**0.55**	**0.72**	**0.81**
		381 EUR and 174 AFR individuals	**0.61**	**0.75**	**0.83**
1000 Genomes EUR	IMPUTE2	380 EUR individuals	0.82	0.86	0.92
(European ancestry; $N = 381$)	(sampling or **pre-phasing**)	(WTCCC2 SNPs)	**0.81**	**0.85**	**0.91**
		380 EUR individuals (sequence SNPs)	0.66	0.79	0.91
			0.64	**0.78**	**0.90**

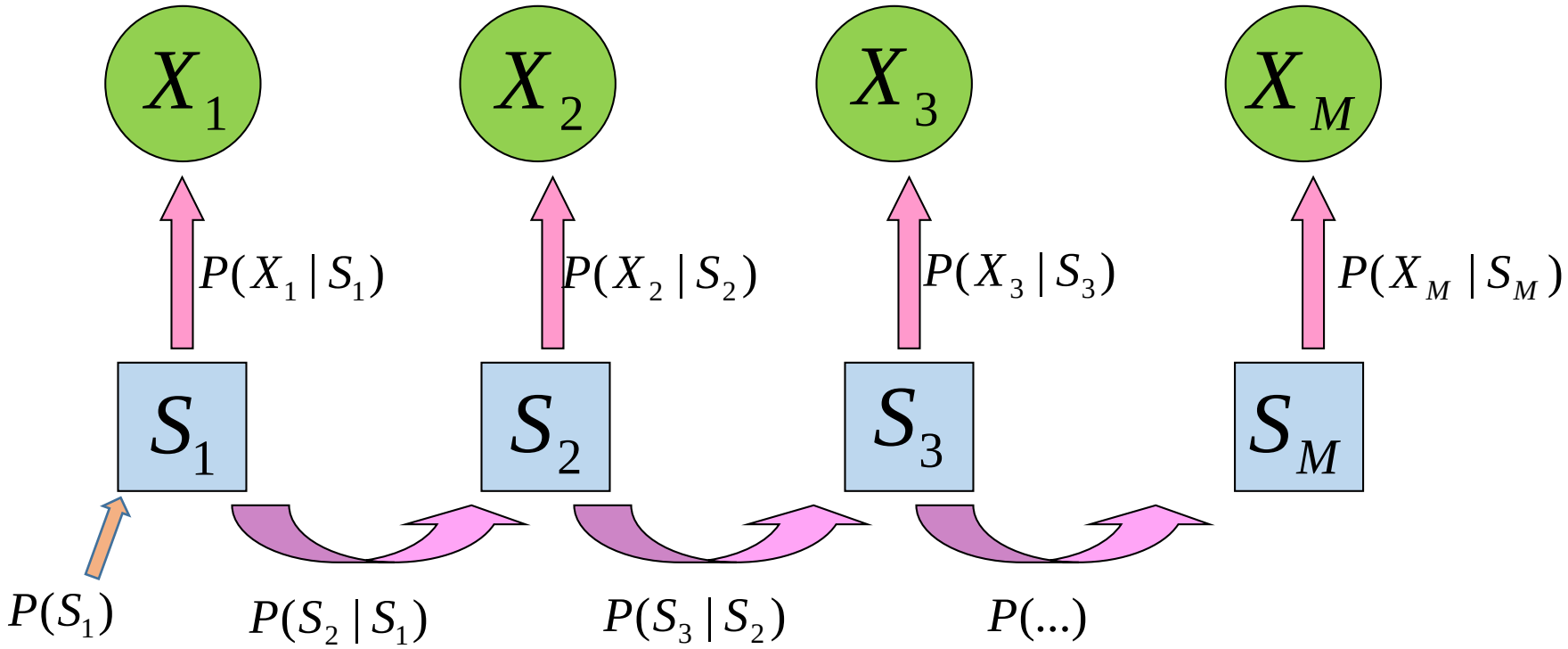
YRI, Yoruba from Ibadan, Nigeria; AFR, African population; CEU, EUR, European populations; from 1000 Genomes.

[a]We imputed each GWAS data set with an existing imputation method and its pre-phasing counterpart. [b]Reference panels used to impute each GWAS data set included the 1000 Genomes low-coverage Pilot (June 2010), the 1000 Genomes interim release (August 2010) and the 1000 Genomes interim Phase 1 release (November 2010). [c]Each cell shows the mean R^2 between true genotypes and imputed dosages for the specified MAF window and reference panel. For a given GWAS data set, all accuracy values within a MAF window were calculated on the same set of SNPs; the corresponding SNP counts are shown in **Supplementary Figure 1**. Accuracy values from pre-phasing are shown in bold (some analyses were performed only with pre-phasing).

Implementation

- Markov model is used to model each haplotype, conditional on all others
- At each position, we assume that the haplotype being modeled copies a template haplotype
- Each individual has two haplotypes, and therefore copies two template haplotypes

Markov Model



The final ingredient connects template states along the chromosome …

Possible States

- A state *S* selects pair of template haplotypes
  - Consider S_i as vector with two elements ($S_{i,1}$, $S_{i,2}$)

- With *H* possible haplotypes, H^2 possible states
  - $H(H+1)/2$ of these are distinct

- A recombination rate parameter describes probability of switches between states
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a, S_{i+1,2} = b))$ $(1-\theta)^2$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b))$ $(\theta(1-\theta))/H$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b^*))$ $(\theta/H)^2$

Emission Probabilities

- Each value of *S* implies expected pair of alleles
- Emission probabilities will be higher when observed genotype matches expected alleles
- Emission probabilities will be lower when alleles mismatch
- Let *T(S)* be a function that provides expected allele pairs for each state *S*

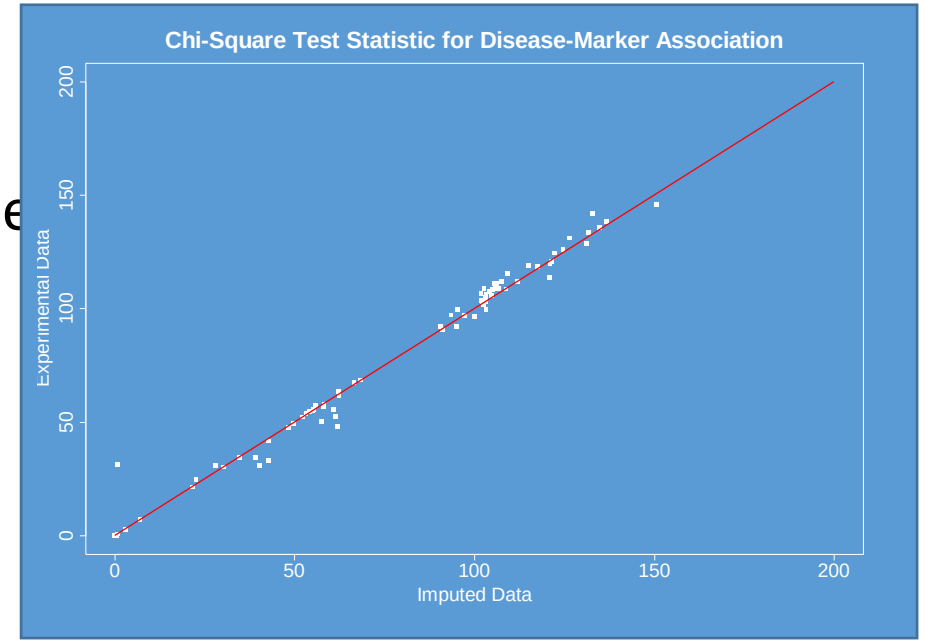
Emission Probabilities

$$P(G_j|S_j) = \begin{cases} (1-\varepsilon_j)^2+\varepsilon_j^2, & T(S_j)=G_j \text{ and } G_j \text{ is heterozygote,} \\ 2(1-\varepsilon_j)\varepsilon_j, & T(S_j)\neq G_j \text{ and } G_j \text{ is heterozygote,} \\ (1-\varepsilon_j)^2, & T(S_j)=G_j \text{ and } G_j \text{ is homozygote,} \\ (1-\varepsilon_j)\varepsilon, & T(S_j) \text{ is heterozygote and } G_j \text{ homozygote,} \\ \varepsilon_j^2, & T(S_j) \text{ and } G_j \text{ are opposite homozygotes.} \end{cases}$$

Does This Really Work? Preliminary Results

- Used 11 tag SNPs to predict 84 SNPs in CFH
- Predicted genotypes differ from original ~1.8% of the time
- Reasonably similar results possible using various haplotyping methods

Comparison of Test Statistics, Truth vs. Imputed



Does this really, really work?

- 90 GAIN psoriasis study samples were re-genotyped for 906,600 SNPs using the Affymetrix 6.0 chip.

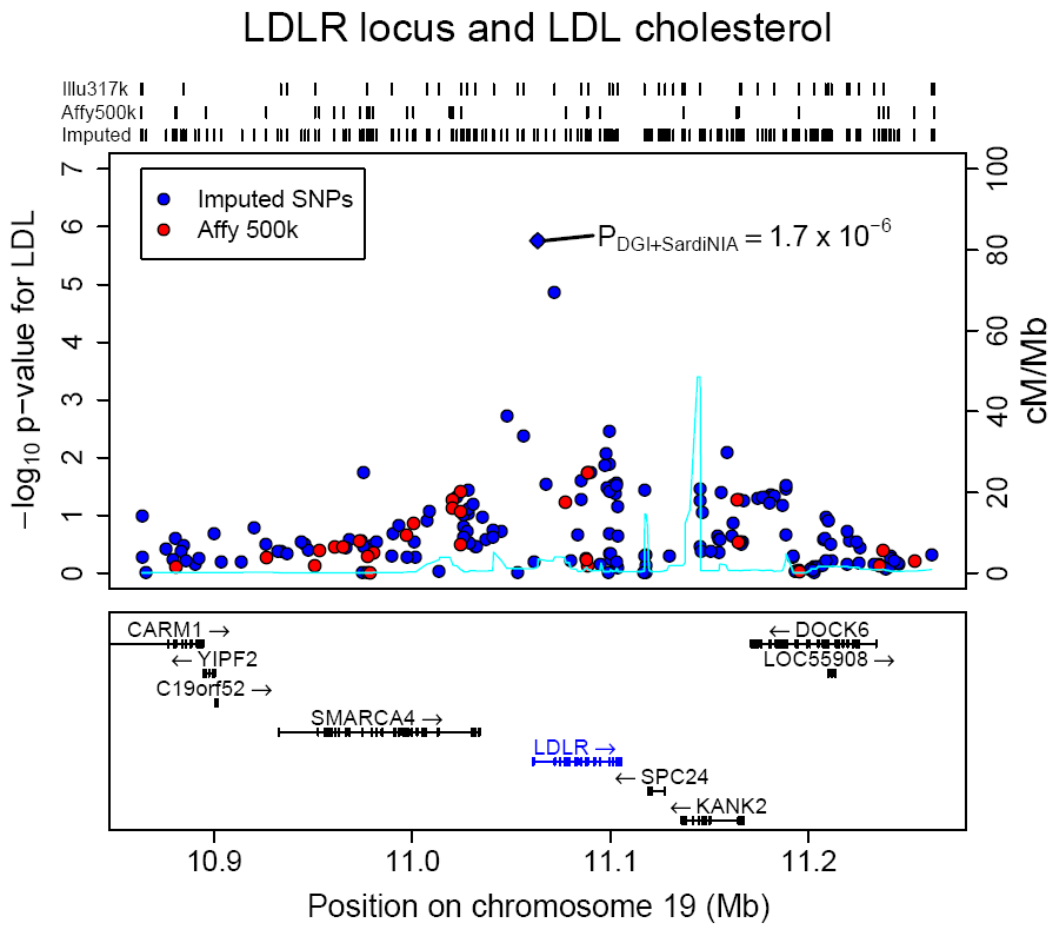
- Comparison of 15,844,334 genotypes for 218,039 SNPs that overlap between the Perlegen and Affymetrix chips resulted in discrepancy rate of 0.25% per genotype (0.12% per allele).

- Comparison of 57,747,244 imputed and experimentally derived genotypes for 661,881 non-Perlegen SNPs present in the Affymetrix 6.0 array resulted in a discrepancy rate of 1.80% per genotype (0.91% per allele).

- Overall, the average r^2 between imputed genotypes and their experimental counterparts was 0.93. This statistic exceeded 0.80 for >90% of SNPs.

Nair et al, *Nature Genetics*, 2009

LDLR and LDL example



Willer et al, *Nature Genetics*, 2008

Li et al, Annual Review of Genomics and Human Genetics, 2009

Impact of HapMap Imputation on Power

	Power	
Disease SNP MAF	**tagSNPs**	**Imputation**
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotyped SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

Combined Lipid Scans

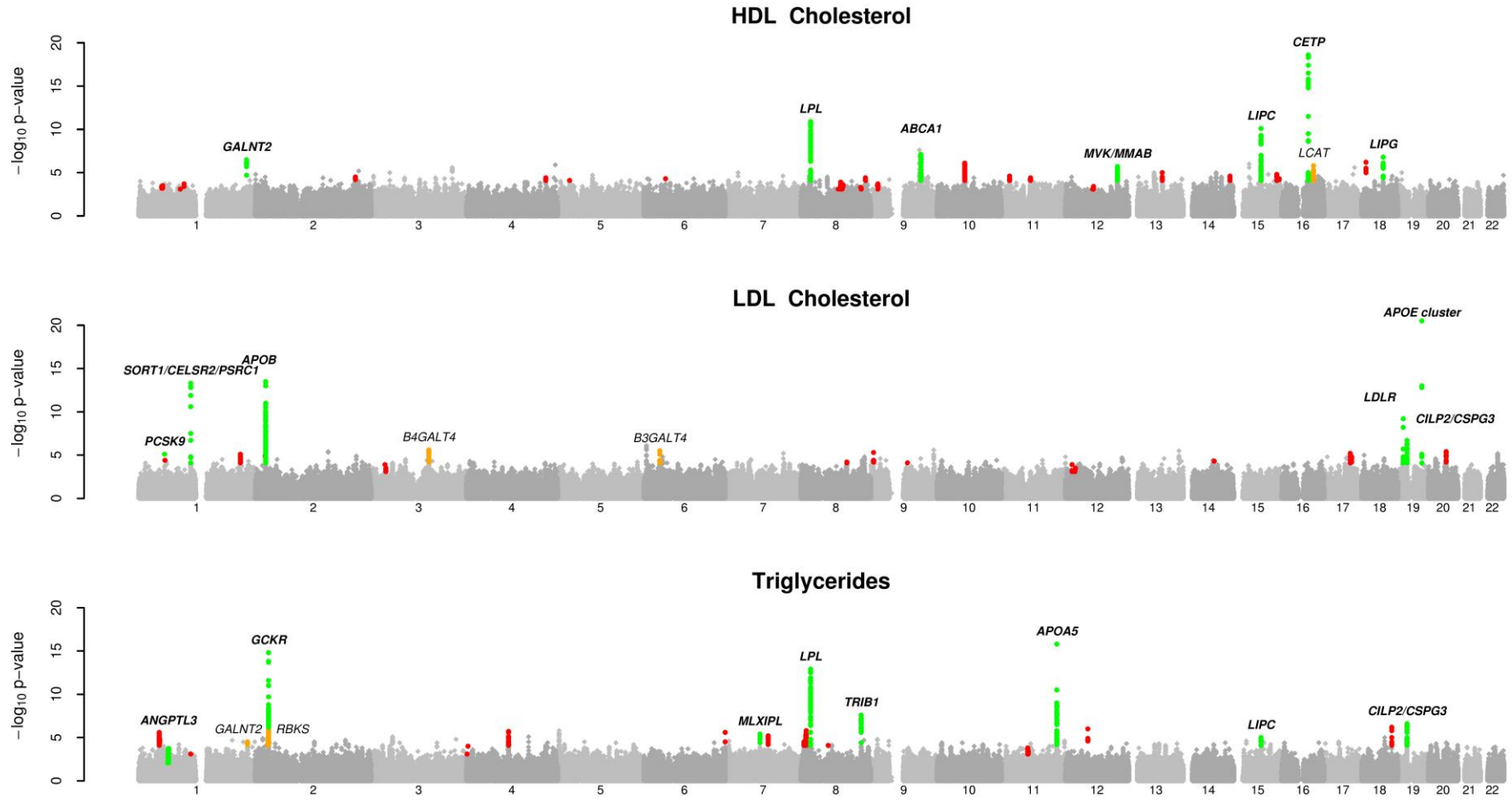
- SardiNIA (Schlessinger, Uda, et al.)
  - ~4,300 individuals, cohort study
- FUSION (Mohlke, Boehnke, Collins, et al.)
  - ~2,500 individuals, case-control study of type 2 diabetes
- DGI (Kathiresan, Altshuler, Orho-Mellander, et al.)
  - ~3,000 individuals, case-control study of type 2 diabetes

- Individually, 1-3 hits/scan, mostly known loci

- Analysis:
  - Impute genotypes so that all scans are analyzed at the same “SNPs”
  - Carry out meta-analysis of results across scans

Willer et al, *Nature Genetics*, 2008

Combined Lipid Scan Results
18 clear loci!



Willer et al, *Nature Genetics*, 2008

Summary

- Genotype imputation can be used to accurately estimate missing genotypes
- Genotype imputation is usually implemented through using a Hidden Markov Model
- Benefits of genotype imputation
  - Increases power of genetic association studies
  - Facilitates analyses that combine data across studies
  - Facilitates interpretation of results

Code Tidbits

Try to Sketch Transition() function for HMM…

Conditioning Probabilities on Observed Data

```
void MarkovModel::Condition(float * vector, char ** haplotypes, int position,
                            char observed, double e, double freq)
   {
   if (observed == 0) return;

   double pmatch = (1. - e) + e * freq;
   double prandom = e * freq;

   for (int i = 0; i < states; i++)
      if (haplotypes[i][position] == observed)
         vector[i] *= pmatch;
      else
         vector[i] *= prandom;
   }
```

Any idea why we use e * freq for error model?

Applying Transition Probabilities …

```
void MarkovModel::Transpose(float * from, float * to, double r)
   {
   if (r == 0)
      for (int i = 0; i < states; i++)
         to[i] = from[i];
   else
      {
      double sum = 0.0;

      for (int i = 0; i < states; i++)
         sum += from[i];

      sum *= r / states;

      double complement = 1. - r;

      // avoid underflows
      if (sum < 1e-10)
         {
         sum *= 1e15;
         complement *= 1e15;
         }

      for (int i = 0; i < states; i++)
         to[i] = from[i] * complement + sum;
      }
   }
```

- What are the inputs?
  - float * from, float * to, double r
- Why calculate the sum?
  - What is the alternative?
  - Why multiply it by 1/r?
- Why is there a section guarding against underflow?

Scanning Along the Chromosome …

```
void MarkovModel::WalkLeft(char * observed, char ** haplotypes, float ** freqs)
   {
   // Initialize likelihoods at first position
   for (int i = 0; i < states; i++)
      matrix[0][i] = 1.;

   // Scan along chromosome
   for (int i = 0; i < markers - 1; i++)
      {
      if (observed[i])
         Condition(matrix[i], haplotypes, i, observed[i], E[i], freqs[observed[i]][i]);
      Transpose(matrix[i], matrix[i+1], R[i]);
      }

   if (observed[markers - 1])
      Condition(matrix[markers - 1], haplotypes, markers - 1, observed[markers - 1],
                E[markers - 1], freqs[observed[markers - 1]][markers - 1]);
   }
```

Connection Between Imputation and Low-Pass Sequencing

Shotgun Sequence Data

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

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

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-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

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

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

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

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

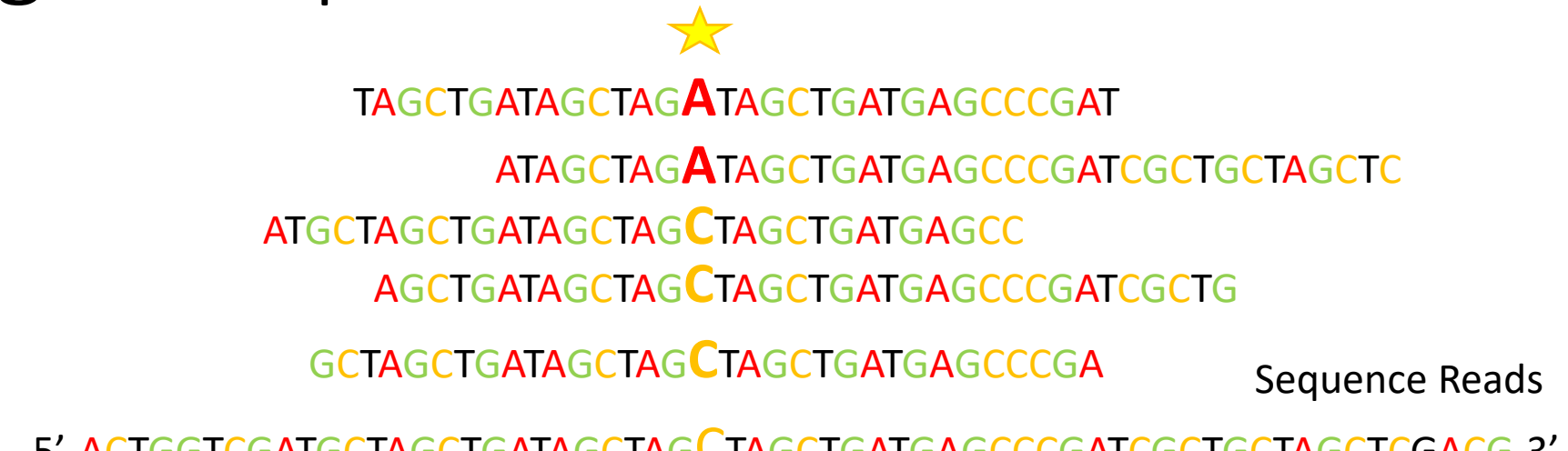
P(reads|A/A , read mapped)= 0.00000098

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

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

Possible Genotypes

Shotgun Sequence Data



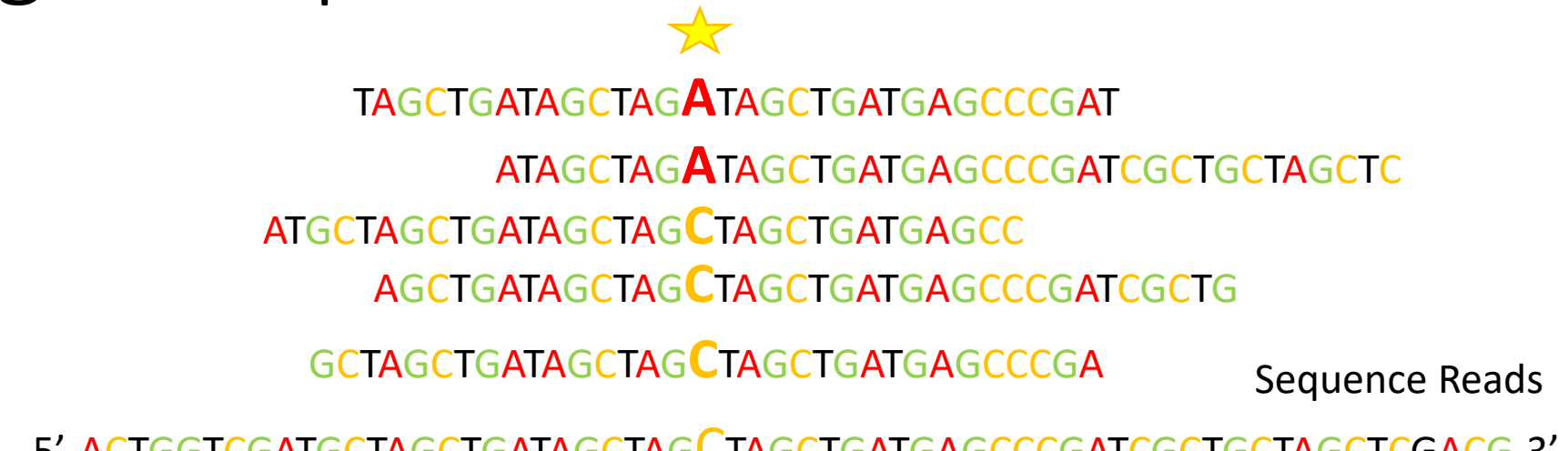
P(reads|A/A, read mapped)= 0.00000098

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

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

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Shotgun Sequence Data



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

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Ingredients That Go Into Prior

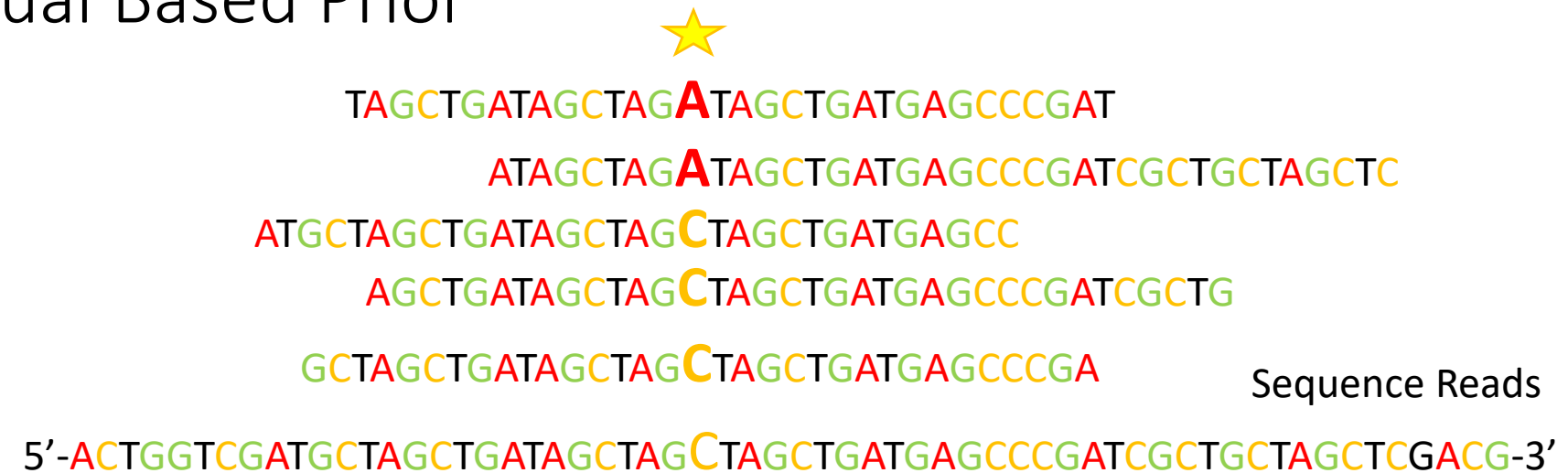
- Most sites don’t vary
  - P(non-reference base) ~ 0.001
- When a site does vary, it is usually heterozygous
  - P(non-reference heterozygote) ~ 0.001 * 2/3
  - P(non-reference homozygote) ~ 0.001 * 1/3
- Mutation model
  - Transitions account for most variants (C↔T or A↔G)
  - Transversions account for minority of variants

From Sequence to Genotype: Individual Based Prior

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



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

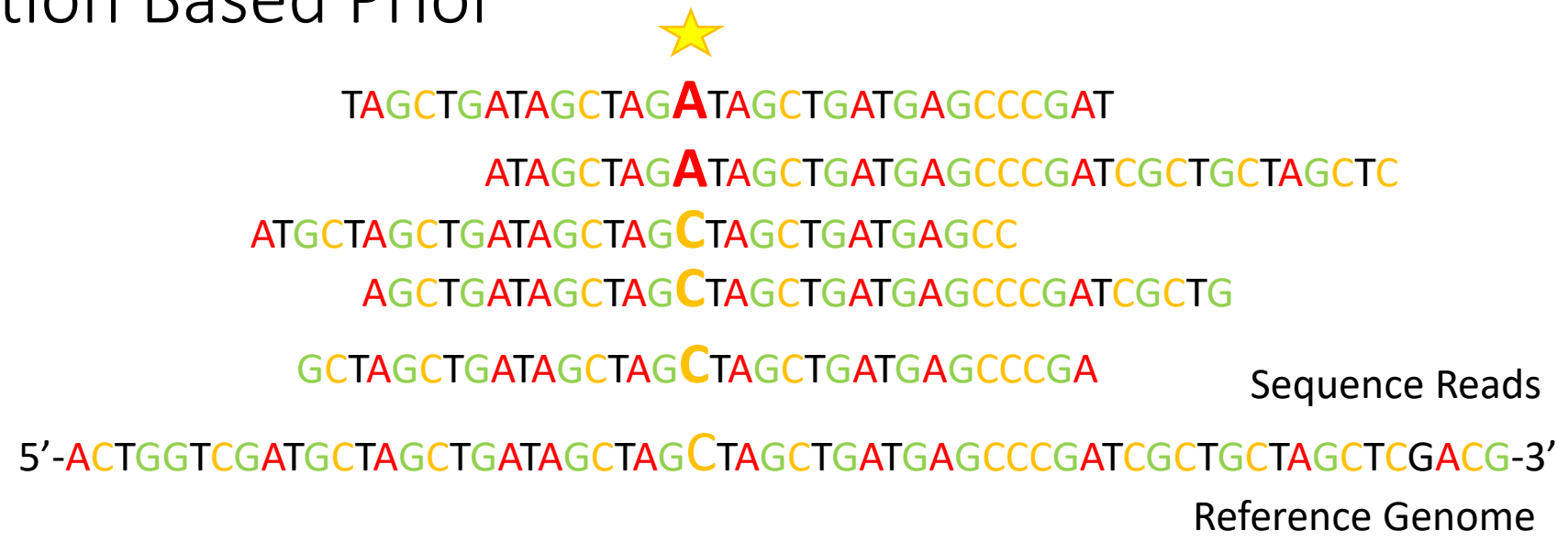
P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.99900 **Posterior(C/C) =** 0.825

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

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

From Sequence to Genotype: Population Based Prior



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

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

From Sequence To Genotype: Population Based Prior

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

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

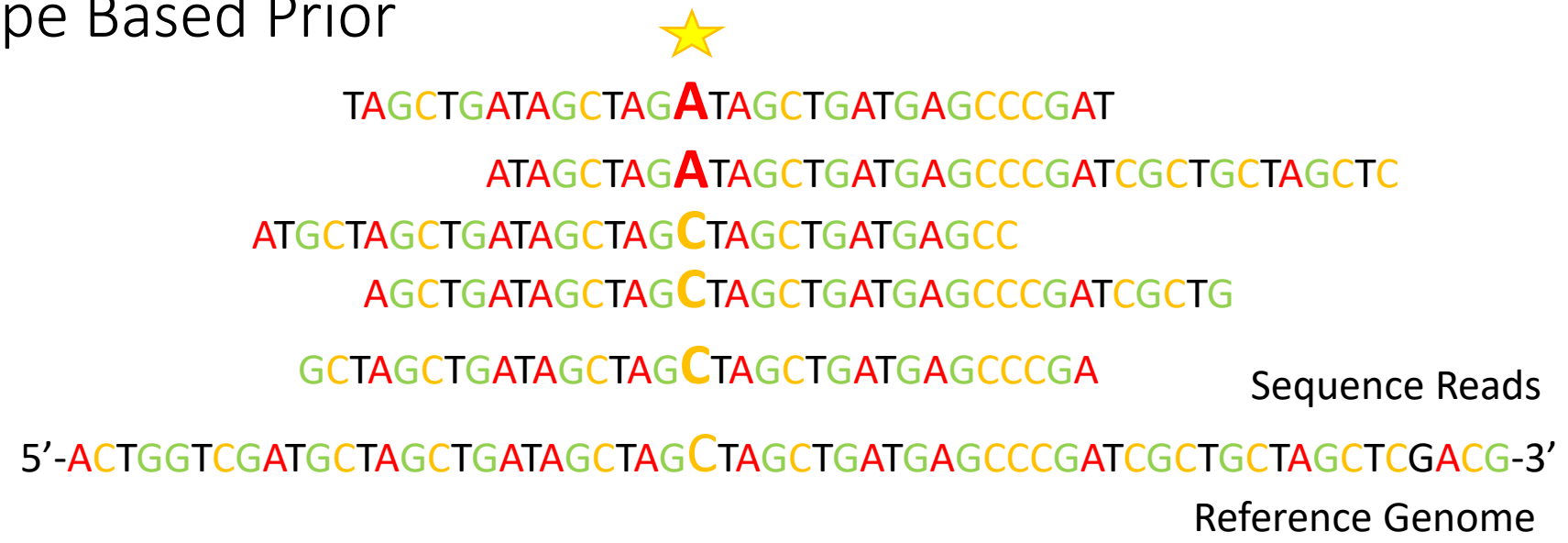
Sequence Based Genotype Calls

- **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

Shotgun Sequence Data

Haplotype Based Prior



P(reads|A/A)= 0.00000098	**Prior(A/A) =** 0.81	**Posterior(A/A) =** <.001
P(reads|A/C)= 0.03125	**Prior(A/C) =** 0.18	**Posterior(A/C) =** 0.999
P(reads|C/C)= 0.000097	**Prior(C/C) =** 0.01	**Posterior(C/C) =** <.001

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.
In the example above, we estimated that 90% of similar chromosomes carry allele A.

Shotgun Sequence Data

Haplotype Based Prior

P(reads|A/A)= 0.00000098 **Prior(A/A) =** 0.81 **Posterior(A/A) =** <.001

P(reads|A/C)= 0.03125 **Prior(A/C) =** 0.18 **Posterior(A/C) =** 0.999

P(reads|C/C)= 0.000097 **Prior(C/C) =** 0.01 **Posterior(C/C) =** <.001

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.
In the example above, we estimated that 90% 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

Current Genome Scale Approaches

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

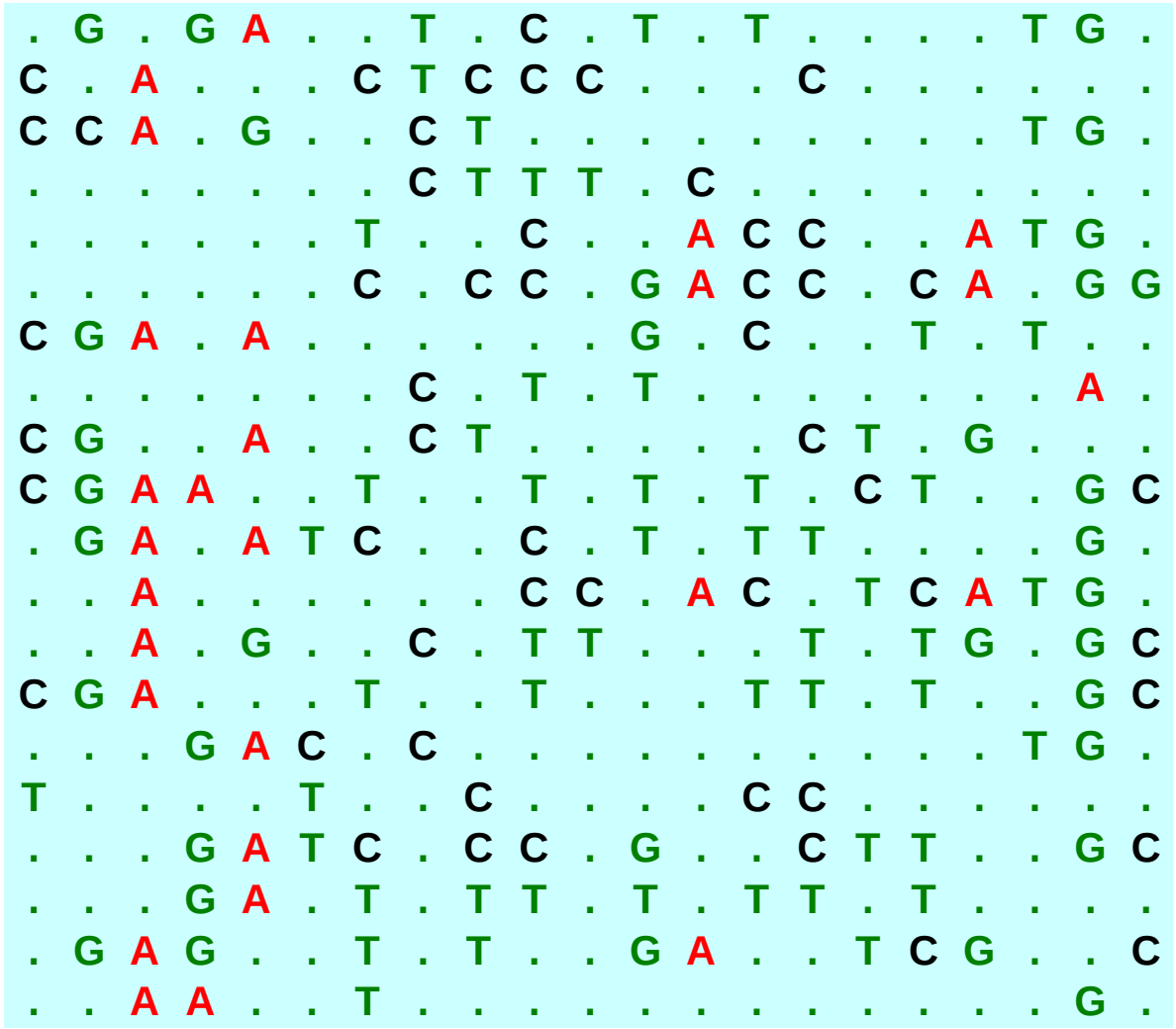
- Exome capture and targeted sequencing
  - Can be applied to moderate numbers of samples
  - SNPs and indels in the most interesting 1% of the genome

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

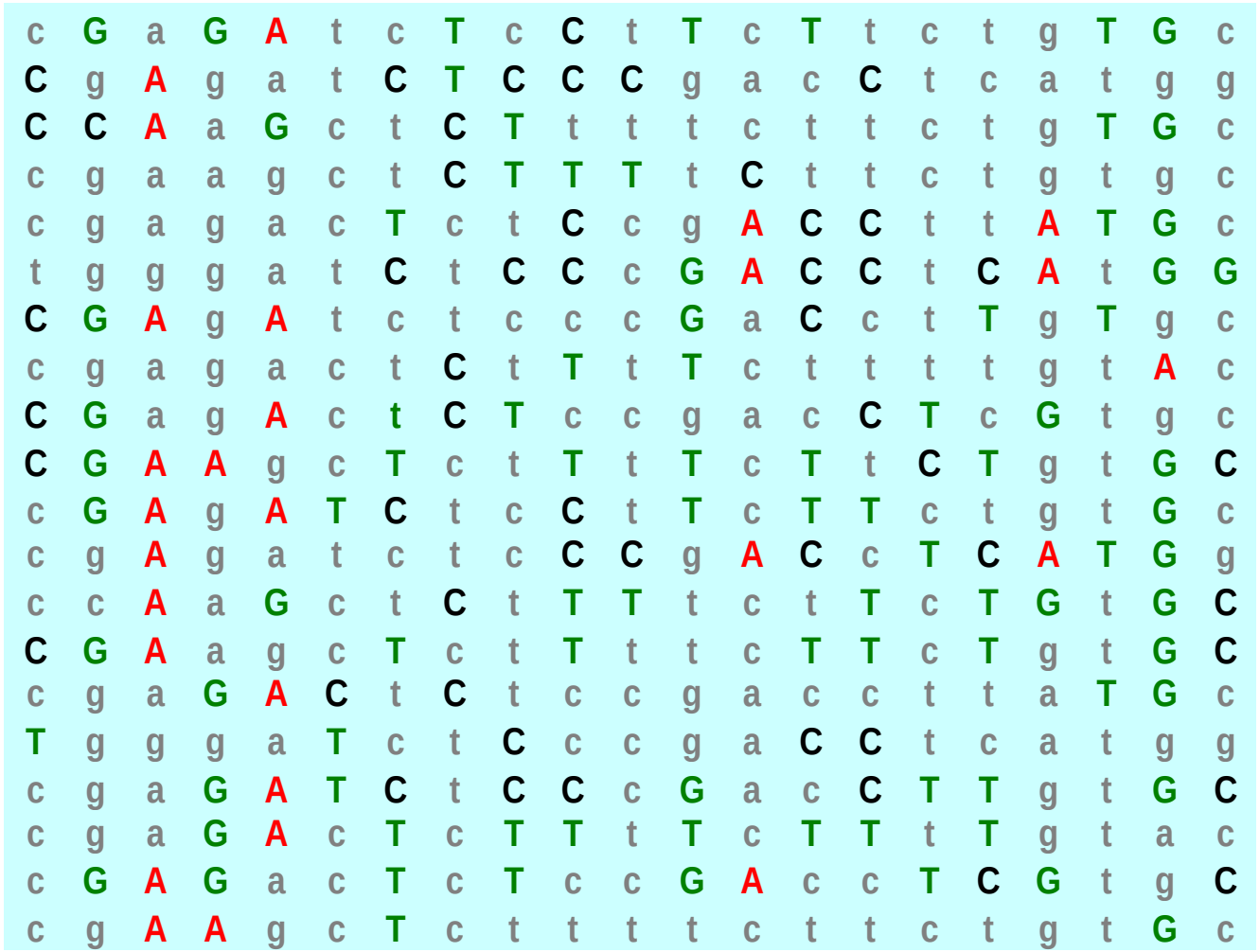
Recipe For Imputation With Shotgun Sequence Data

- Start with some plausible configuration for each individual
- Use Markov model to update one individual conditional on all others
- Repeat previous step many times
- Generate a consensus set of genotypes and haplotypes for each individual

Silly Cartoon View of Shot Gun Data



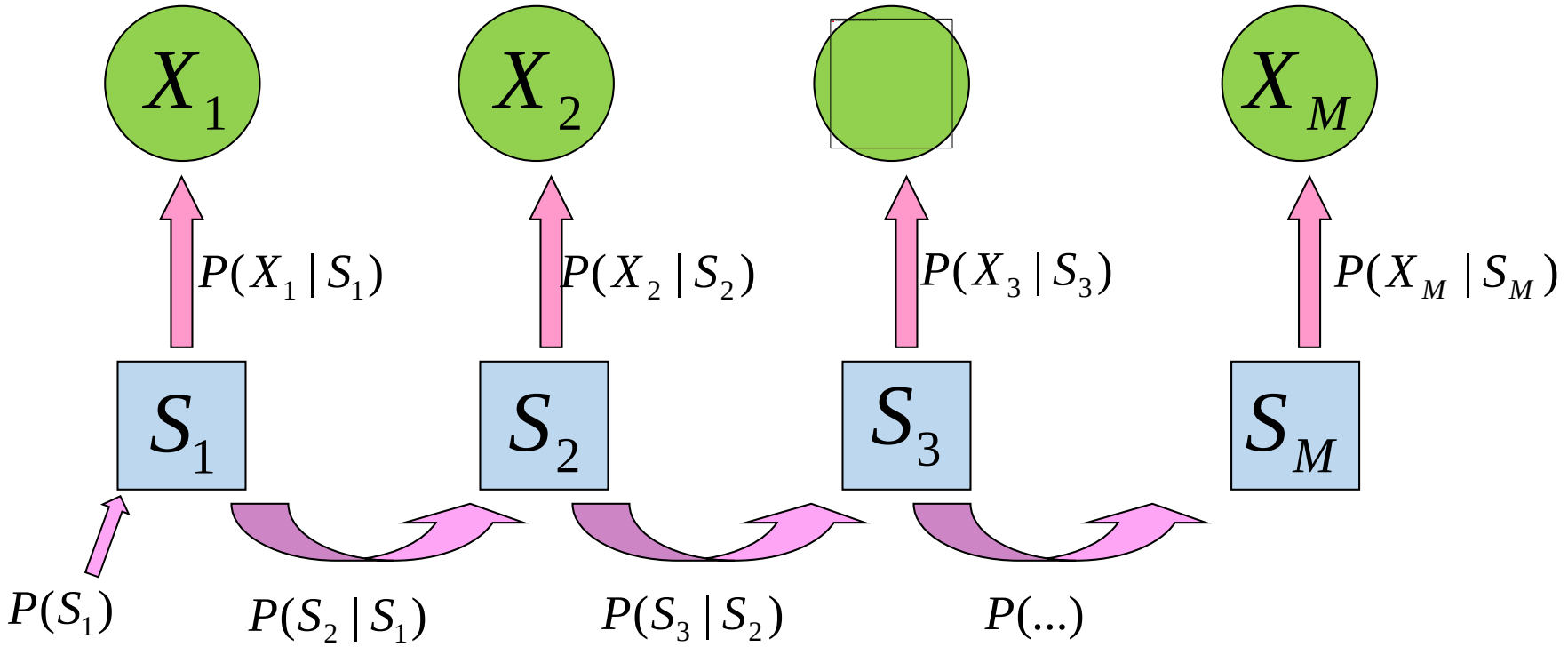
Silly Cartoon View of Shot Gun Data



How Do We Update One Pair Of Haplotypes?

- Markov model similar to that for genotype imputation
- To carry out an update, select one individual
  - Let X_i be observed bases overlapping position *i* for individual
- Assume (temporarily) that current haplotype estimates for all other individuals are correct
- Model haplotypes for individual being updated as mosaic of the other available haplotypes
  - $S_i = (S_{i1}, S_{i2})$ denotes the pair of haplotypes being copied

Markov Model



Model is very similar to the one we previously used for imputation…

Likelihood

$$L = \sum_{S_1} \sum_{S_2} \cdots \sum_{S_M} P(S_1) \prod_{i=2}^{M} P(S_i \mid S_{i-1}) \prod_{i=1}^{M} P(X_i \mid S_i)$$

- $P(S_1) = 1 / H^2$ where H is the number of template haplotypes
- $P(S_i|S_{i-1})$ depends on estimated population recombination rate
- $P(X_i|S_i)$ are the genotype likelihoods

Simulation Results: Common Sites

- Detection and genotyping of Sites with MAF >5% (2116 simulated sites/Mb)

  - **Detected Polymorphic Sites: 2x coverage**
  - 100 people 2102 sites/Mb detected
  - 200 people 2115 sites/Mb detected
  - 400 people 2116 sites/Mb detected

  - **Error Rates at Detected Sites: 2x coverage**
  - 100 people 98.5% accurate, 90.6% at hets
  - 200 people 99.6% accurate, 99.4% at hets
  - 400 people 99.8% accurate, 99.7% at hets

Simulation Results: Rarer Sites

- Detection and genotyping of Sites with MAF 1-2% (425 simulated sites/Mb)

  - **Detected Polymorphic Sites: 2x coverage**
  - 100 people 139 sites/Mb detected
  - 200 people 213 sites/Mb detected
  - 400 people 343 sites/Mb detected

  - **Error Rates at Detected Sites: 2x coverage**
  - 100 people 98.6% accurate, 92.9% at hets
  - 200 people 99.4% accurate, 95.0% at hets
  - 400 people 99.6% accurate, 95.9% at hets

That’s The Theory … Show Me The Data!

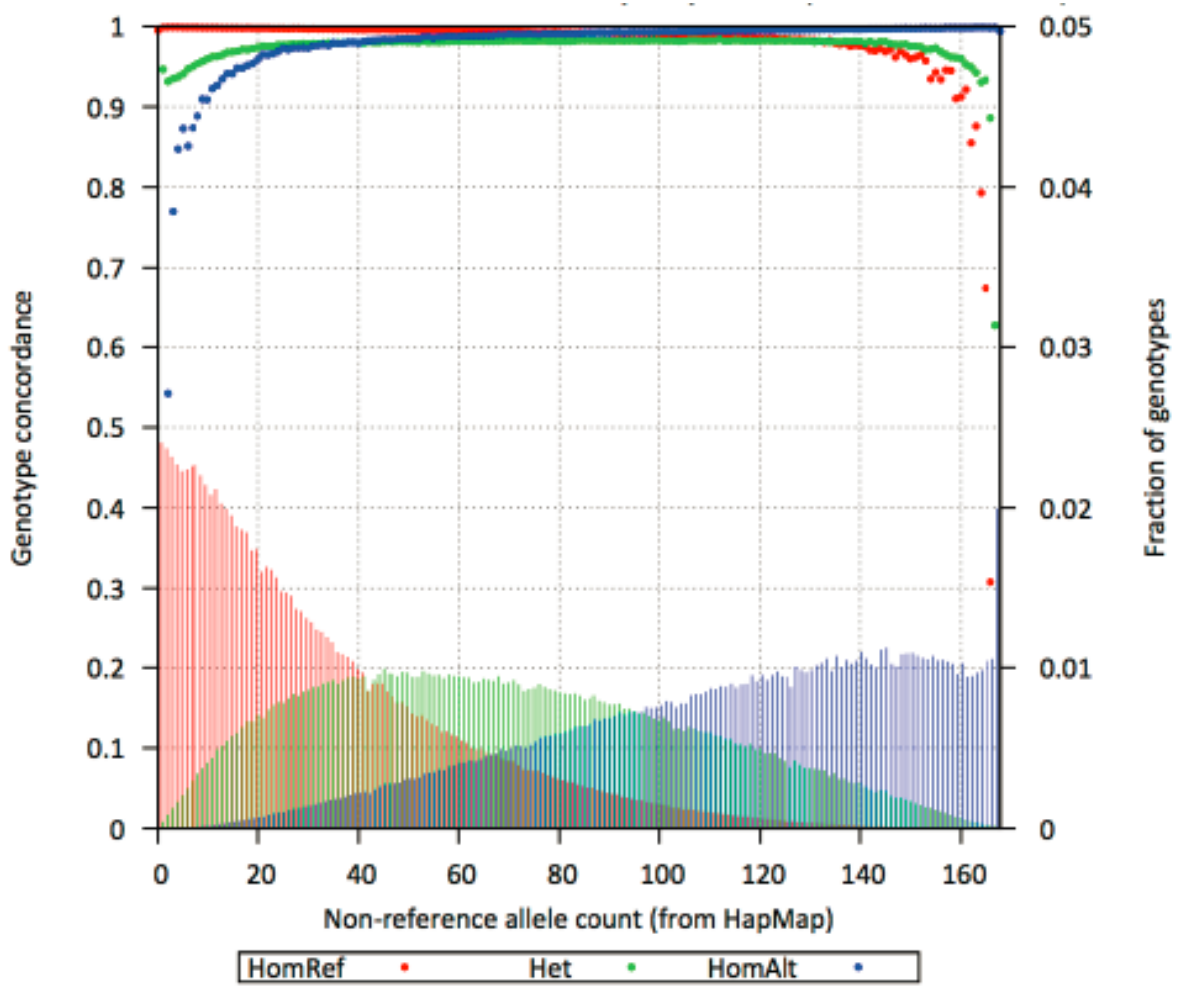
Results from 1000 Genomes Project

1000 Genomes Pilot Completed

- 2 deeply sequenced trios
- 179 whole genomes sequenced at low coverage
- 8,820 exons deeply sequenced in 697 individuals

- 15M SNPs, 1M indels, 20,000 structural variants

Accuracy of Low Pass Genotypes



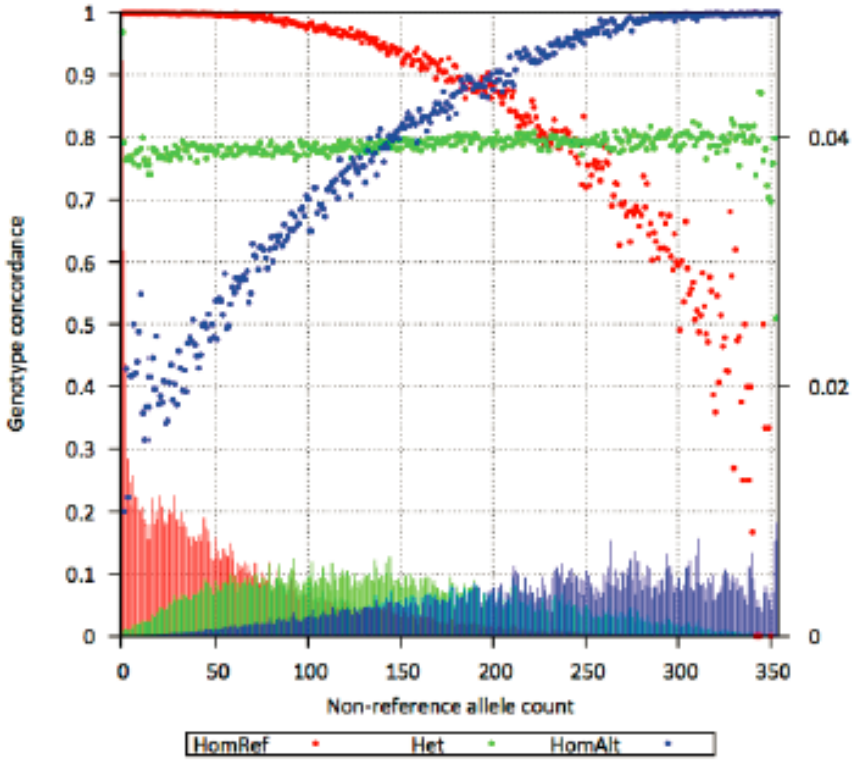
Genotype accuracy for rare genotypes is lowest, but definition of rare changes as more samples are sequenced.

Hyun Min Kang

Does Haplotype Information Really Help?

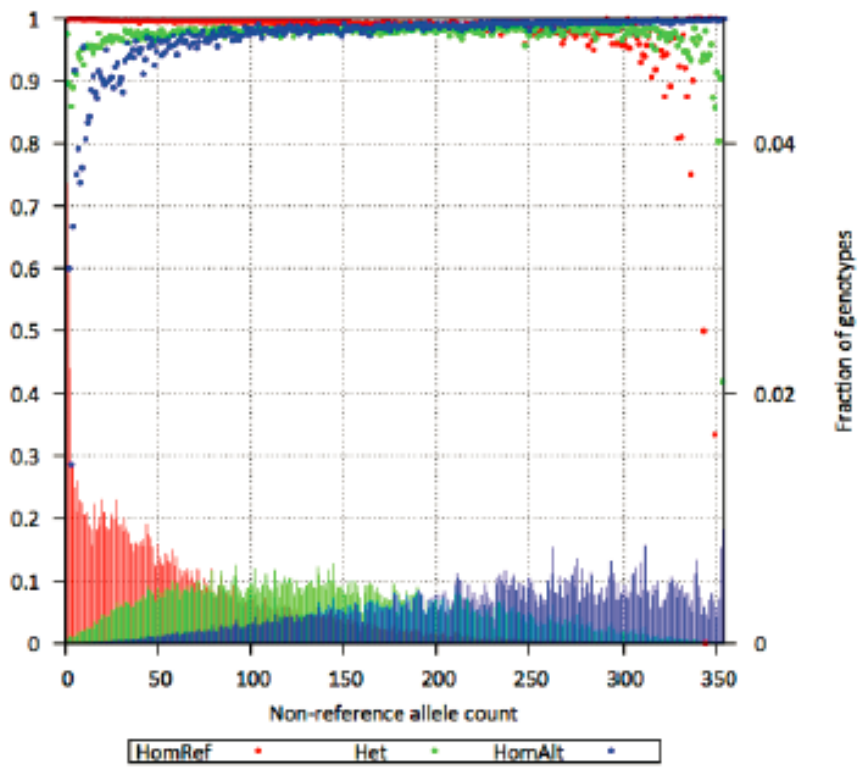
Single Site Analysis

– 21.4% HET errors



Haplotype Aware Analysis

– 2.0% HET errors



As More Samples Are Sequenced, Low Pass Genotypes Improve

Analysis	#SNPs	dbSNP%	Missing HapMap %	Ts/Tv	Accuracy at Hets*
March 2010 Michigan/EUR 60	9,158,226	63.5	7.0	1.91	96.74
August 2010 Michigan/EUR 186	10,537,718	52.5	5.6	2.04	97.56
October 2010 Michigan/EUR 280	13,276,643	50.1	1.8	2.20	97.91**

Accuracy of Low Pass Genotypes Generated by 1000 Genomes Project, When Analyzed Here At the University of Michigan

Some Important Notes

- The Markov model we described is one of several possible models for analysis of low pass data
- Alternative models, based on E-M algorithms or local clustering of individuals into small groups exist
- Currently, the best possible genotypes produced by running multiple methods and generating a consensus across analysis their results.

What Was Optimal Model for Analyzing 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

Recommended Reading

- 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.

- Le SQ and Durbin R (2010) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research* (in press)