Whole Genome Sequencing
Low Pass Sequencing

Gonçalo Abecasis

Previous Lecture

- Introduction to Whole Genome Sequencing
  - What will we learn from whole genome sequencing?
- Challenges with Read Mapping
- Interpreting Mismatches: Variant or Error
  - Single individual analyses require deep sequencing
  - Multi-individual analyses can use shallower data
- Information contained in paired reads

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 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

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

From Sequence to Genotype: Individual Based Prior

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)= 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: Population 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

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

- **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

The Challenge

- Whole genome sequence data will greatly increase our understanding of complex traits
- Although a handful of genomes have been sequenced, this remains a relatively expensive enterprise
- Dissecting complex traits will require whole genome sequencing of 1,000s of individuals
- **How to sequence 1,000s of individuals cost-effectively?**

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

Current Genome Scale Approaches

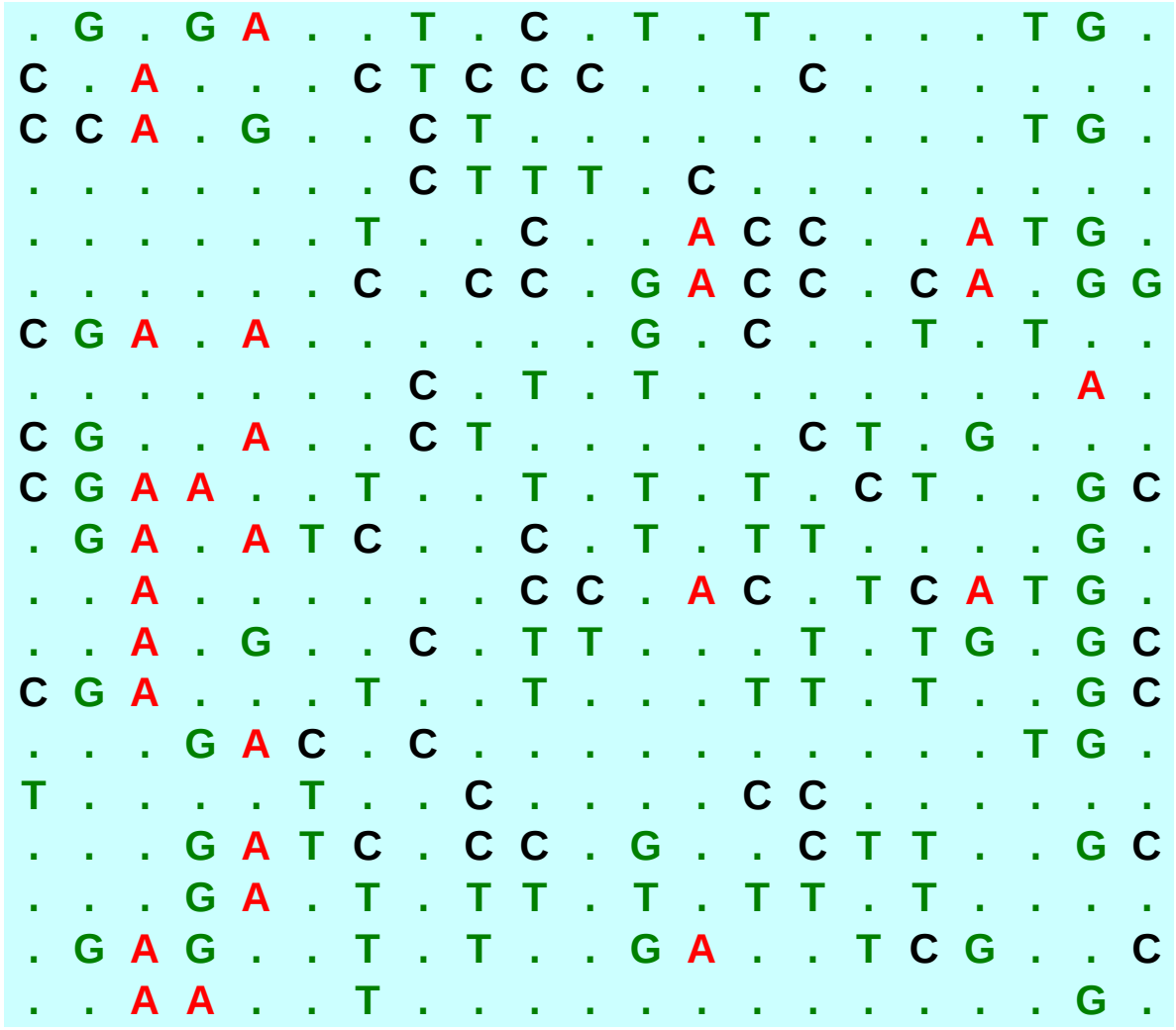
- Deep whole genome sequencing
  - Can only be applied to limited numbers of samples
  - Most complete ascertainment of variation
- Exome
  - Can
  - SNPs and in
- Low coverage whole genome sequencing
  - Can be applied to moderate numbers of samples
  - Very complete ascertainment of shared variation
  - Less complete ascertainment of rare variants

Our Focus For Today

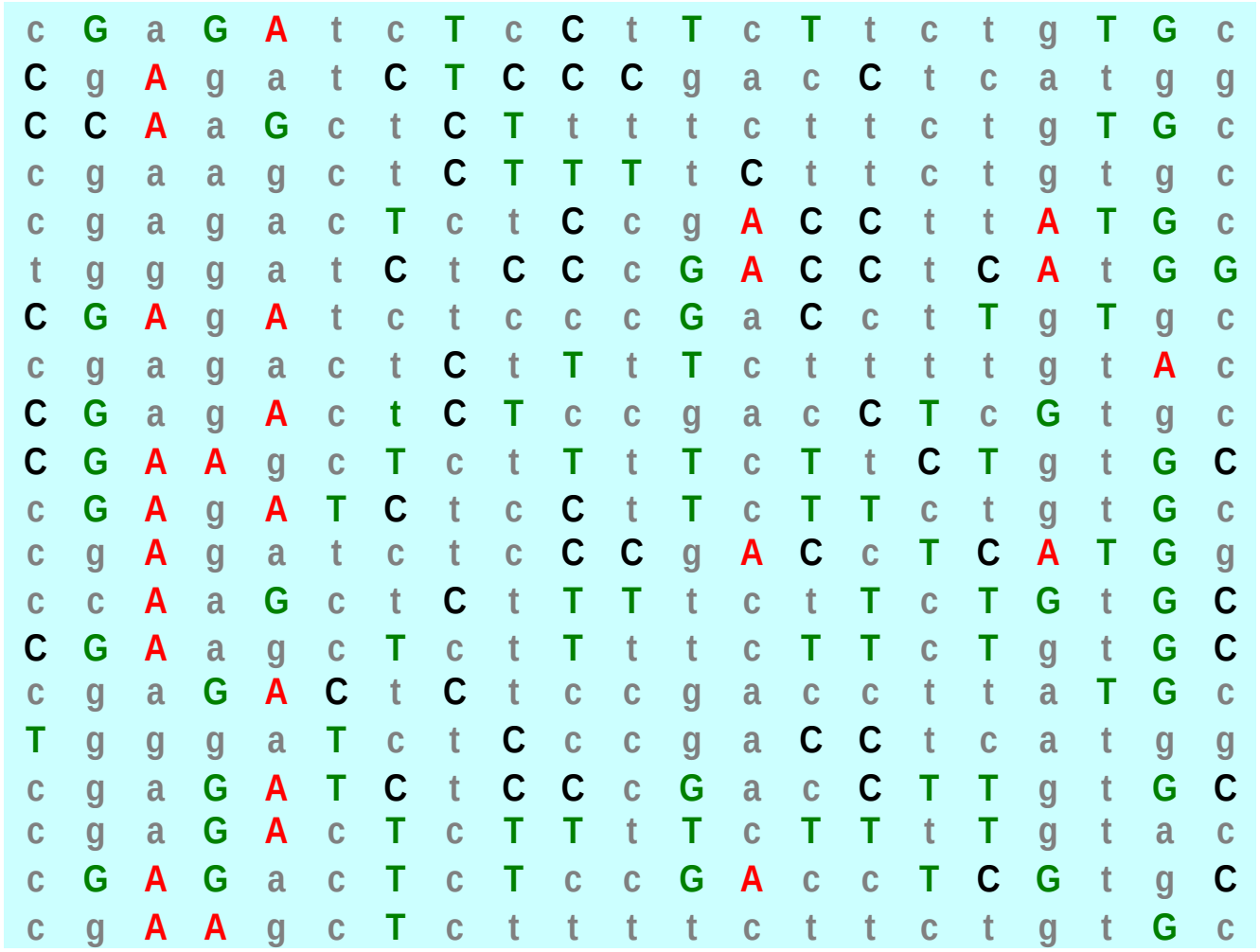
Recipe For Imputation With Shotgun Sequence Data

- 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



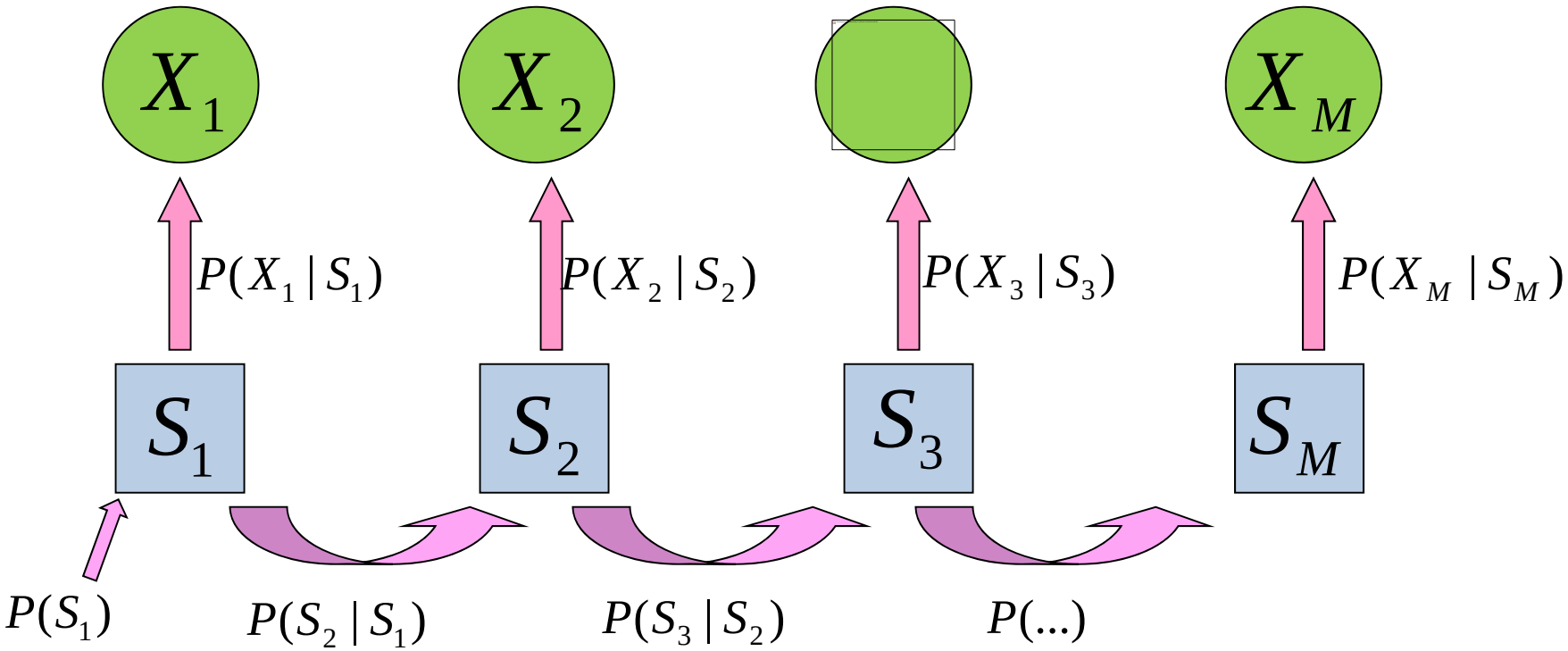
Silly Cartoon View of Shot Gun Data



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 imputatoin…

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
- $P(S_i \mid S_{i-1})$ depends on estimated population recombination rate
- $P(X_i \mid S_i)$ 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

Project Goals

- >95% of accessible genetic variants
  with a frequency of >1%
  in each of multiple continental regions

- Extend discovery effort to lower frequency variants in coding regions of the genome

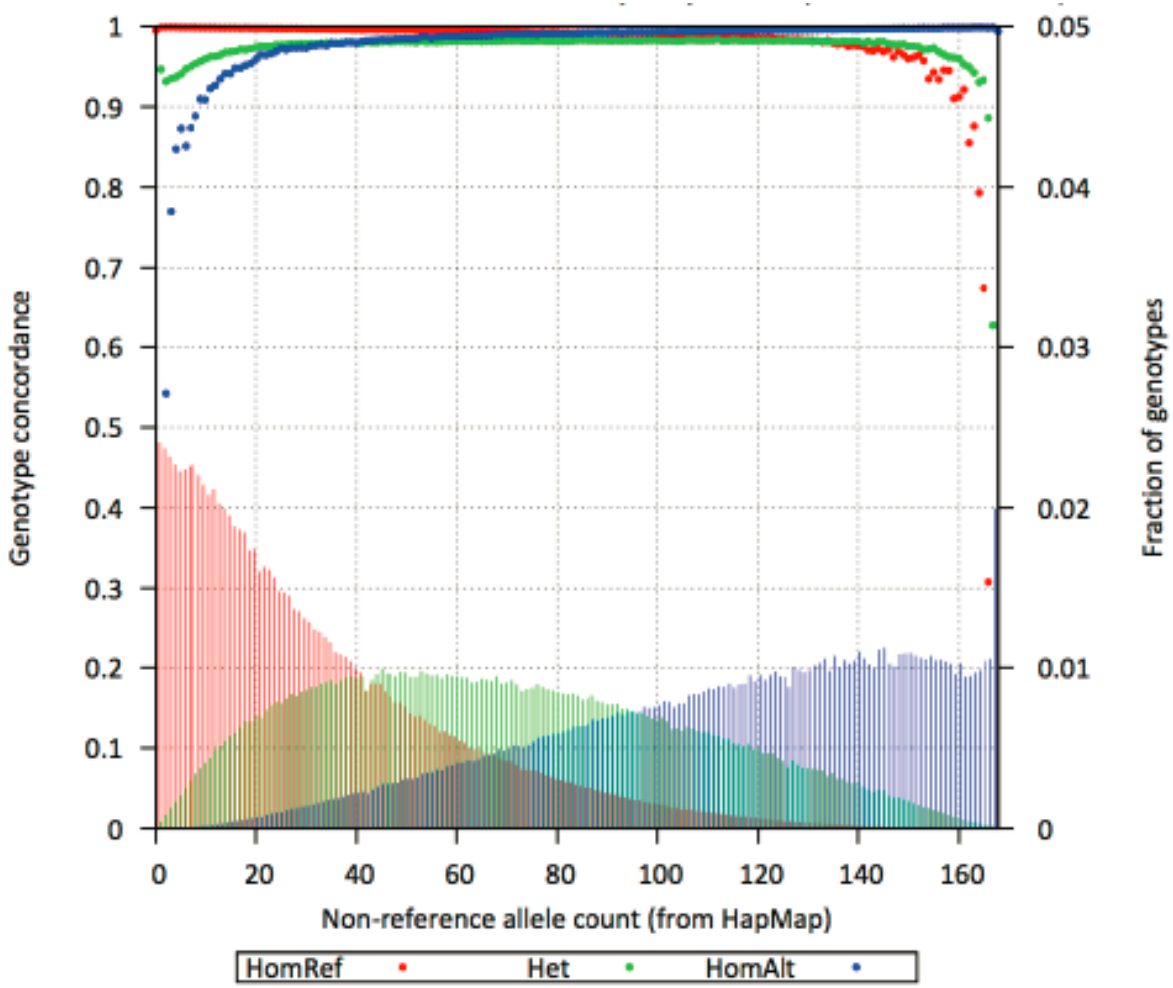
- Define haplotype structure in the genome

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



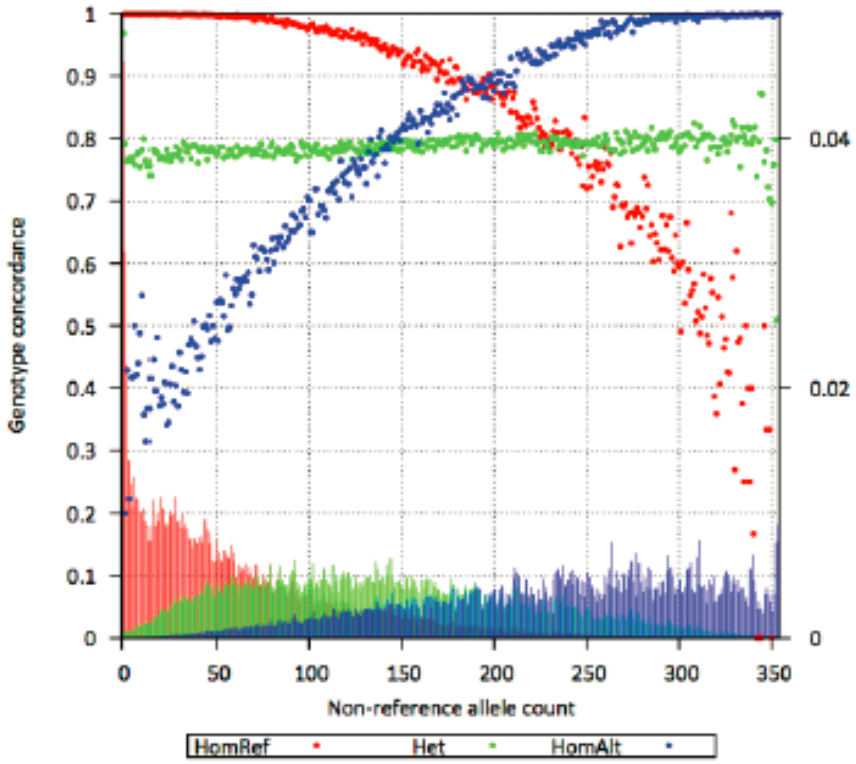
Genotype accuracy for rare genotypes is lowest, but definition of rare changes as more samples are sequenced.

Hyun Min Kang

Does Haplotype Information Really Help?

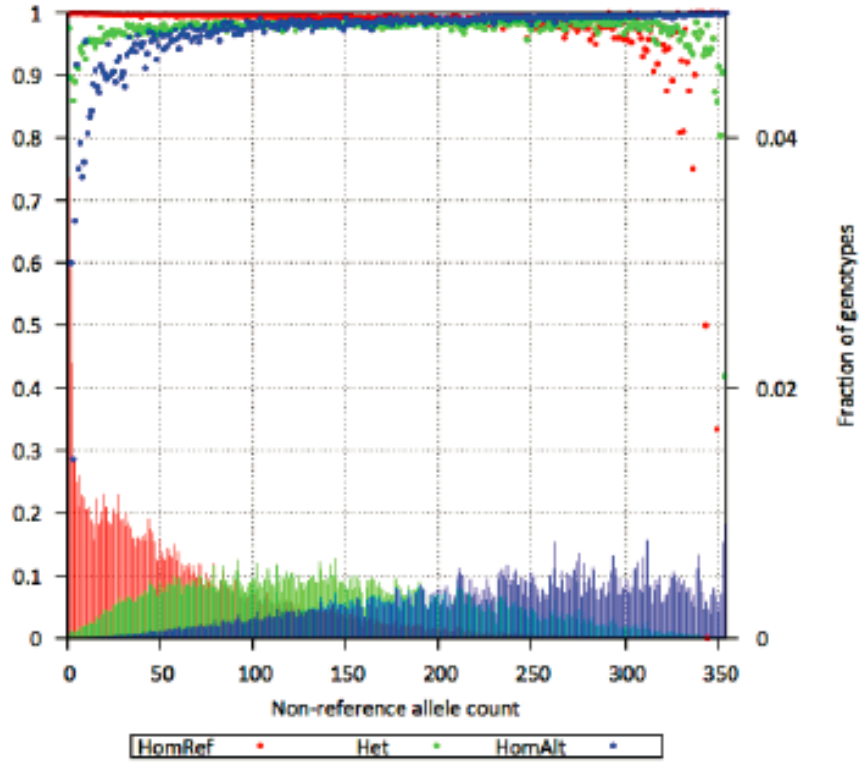
Single Site Analysis

– 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

Implications for Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 67 individuals at 30x

	Sequencing of 67 individuals at 30x depth			
Minor Allele Frequency	**0.5 – 1.0%**	**1.0 – 2.0%**	**2.0 – 5.0%**	**>5%**
Proportion of Detected Sites	59.3%	90.1%	96.9%	100.0%
Genotyping Accuracy	100.0%	100.0%	100.0%	100.0%
…. Heterozygous Sites Only	100.0%	100.0%	100.0%	100.0%
Correlation with Truth (r^2)	99.8%	99.9%	99.9%	100.0%
Effective Sample Size ($n \cdot r^2$)	67	67	67	67

Implications for Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 1000 individuals at 2x

	Sequencing of 1000 individuals at 2x depth			
Minor Allele Frequency	**0.5 – 1.0%**	**1.0 – 2.0%**	**2.0 – 5.0%**	**>5%**
Proportion of Detected Sites	79.6%	98.8%	100.0%	100.0%
Genotyping Accuracy	99.6%	99.5%	99.5%	99.8%
…. Heterozygous Sites Only	78.8%	89.5%	95.9%	99.8%
Correlation with Truth (r^2)	56.7%	76.1%	88.2%	97.8%
Effective Sample Size ($n \cdot r^2$)	567	761	882	978

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

Summary So Far

- Analysis of Low Pass Sequence Data
  - Single sample analyses produce poor quality variants.
  - Single site analyses produce poor quality genotypes.
  - Multi-sample, multi-site analyses can work quite well.

- Intuition for why low pass analyses are attractive for complex disease association studies.

Design A Whole Genome Low Pass Sequencing Study

Gonçalo Abecasis

David Schlessinger

Francesco Cucca

SardiNIA Whole Genome Sequencing

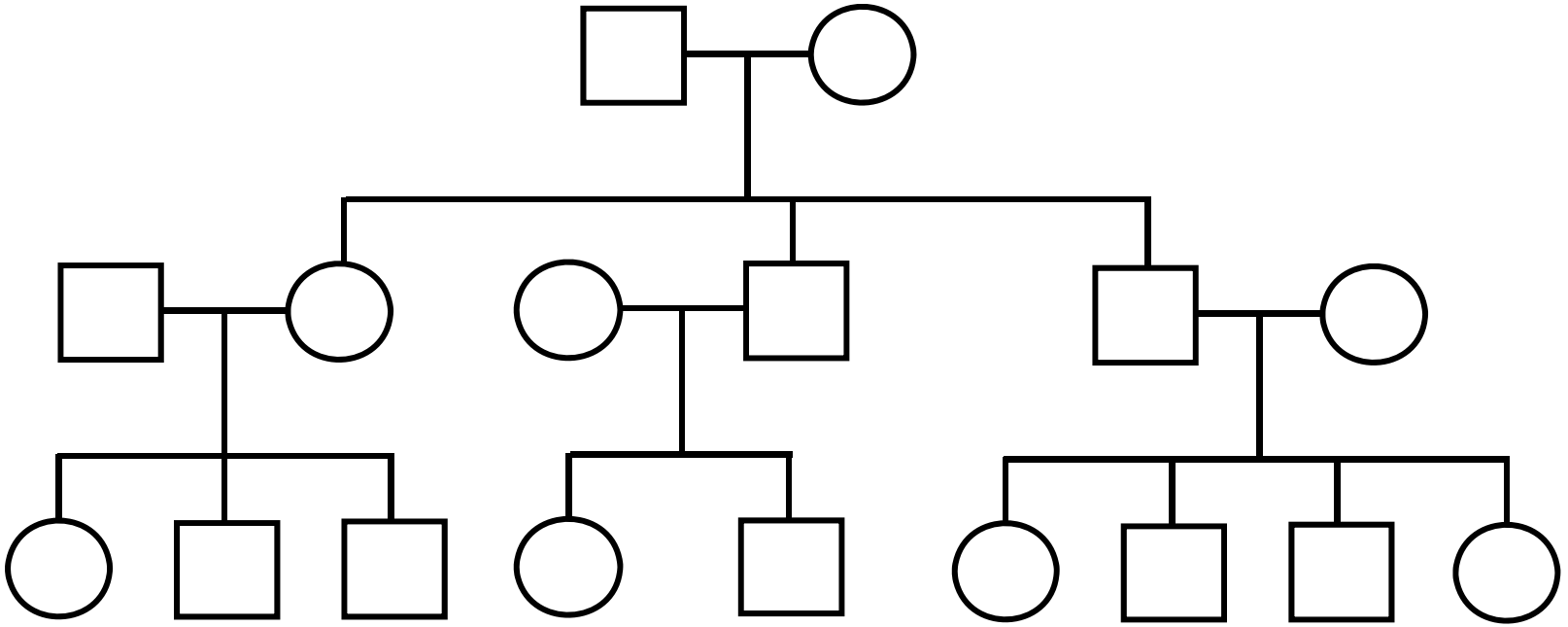
- 6,148 Sardinians from 4 towns in the Lanusei Valley, Sardinia
  - Recruited among population of ~9,841 individuals
  - Sample includes >34,000 relative pairs

- Measured ~100 aging related quantitative traits

- Original plan:
  - Set out to sequence >1,000 individuals at 2x to obtain genomes
  - Genotype all individuals, impute sequences into relatives

Who To Sequence?

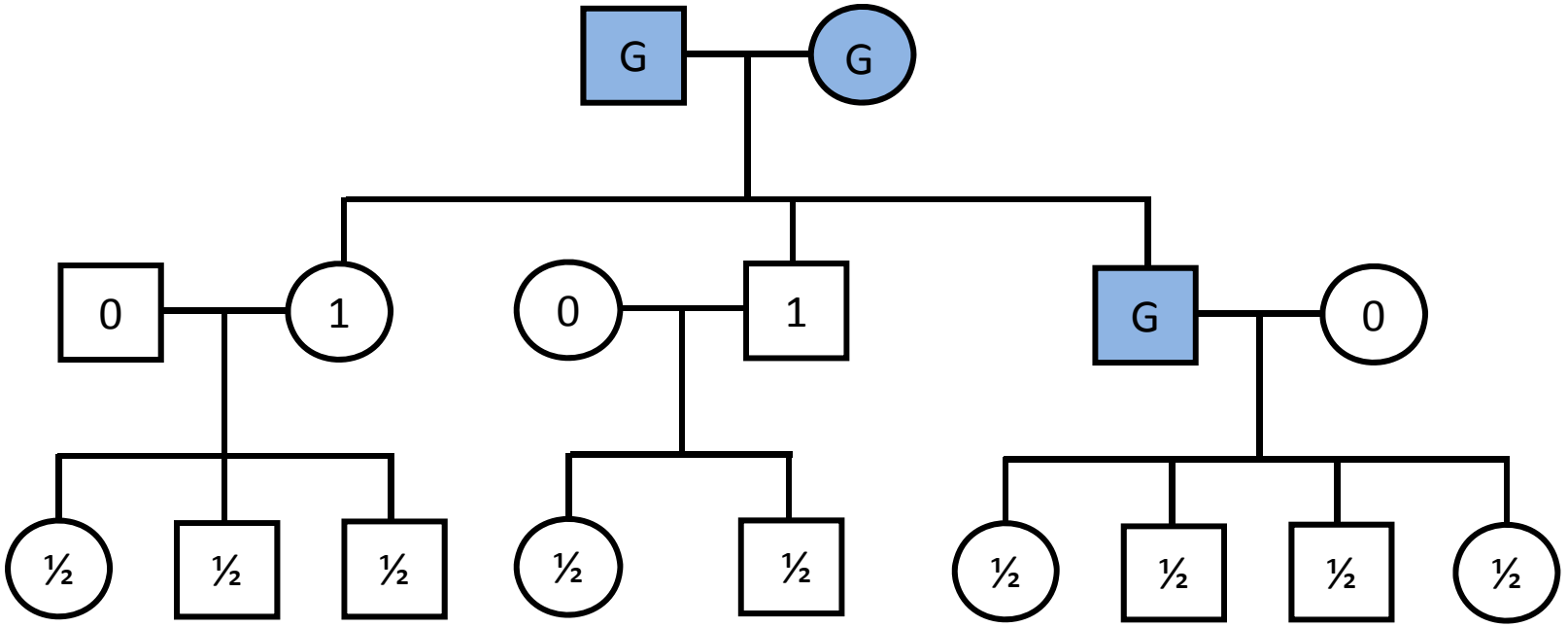
Assuming All Individuals Have Been Genotyped



0 Genomes Sequenced, 0 Genomes Analyzed

Who To Sequence?

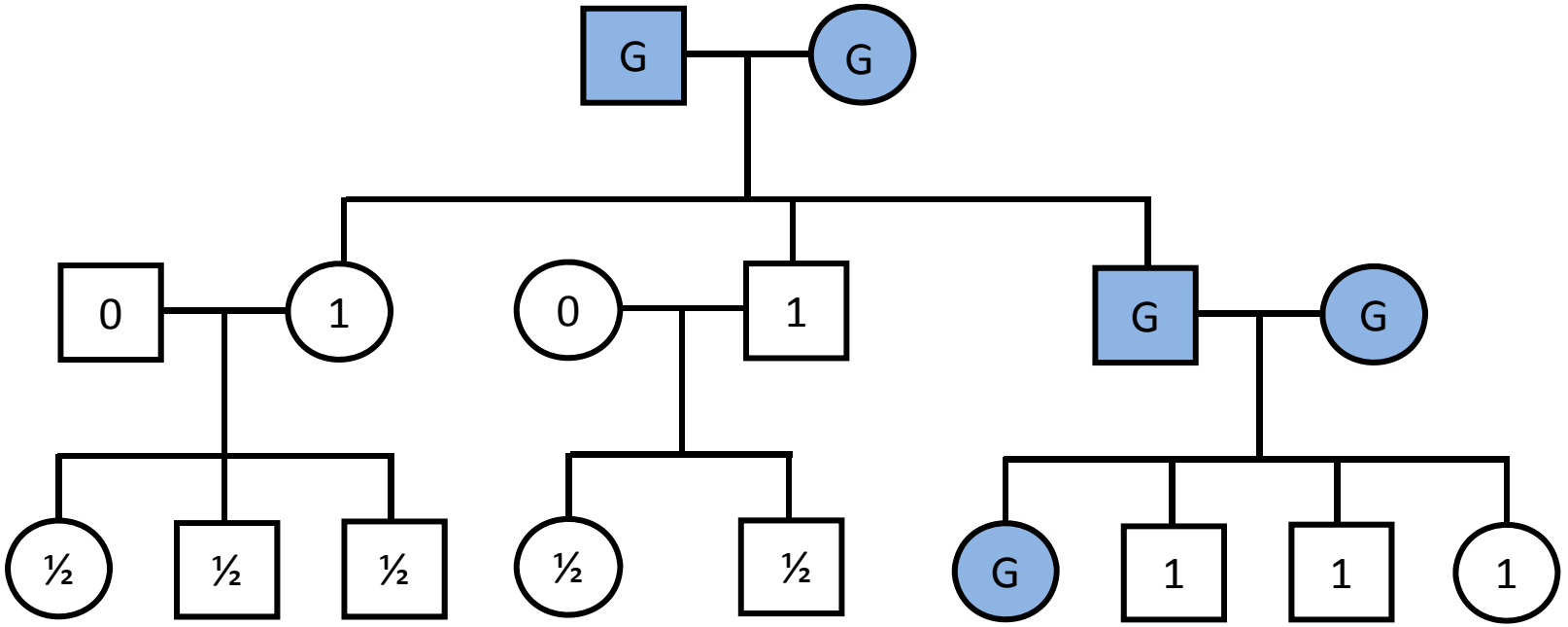
Assuming All Individuals Have Been Genotyped



3 Genomes Sequenced, 9.5 Genomes Analyzed

Who To Sequence?

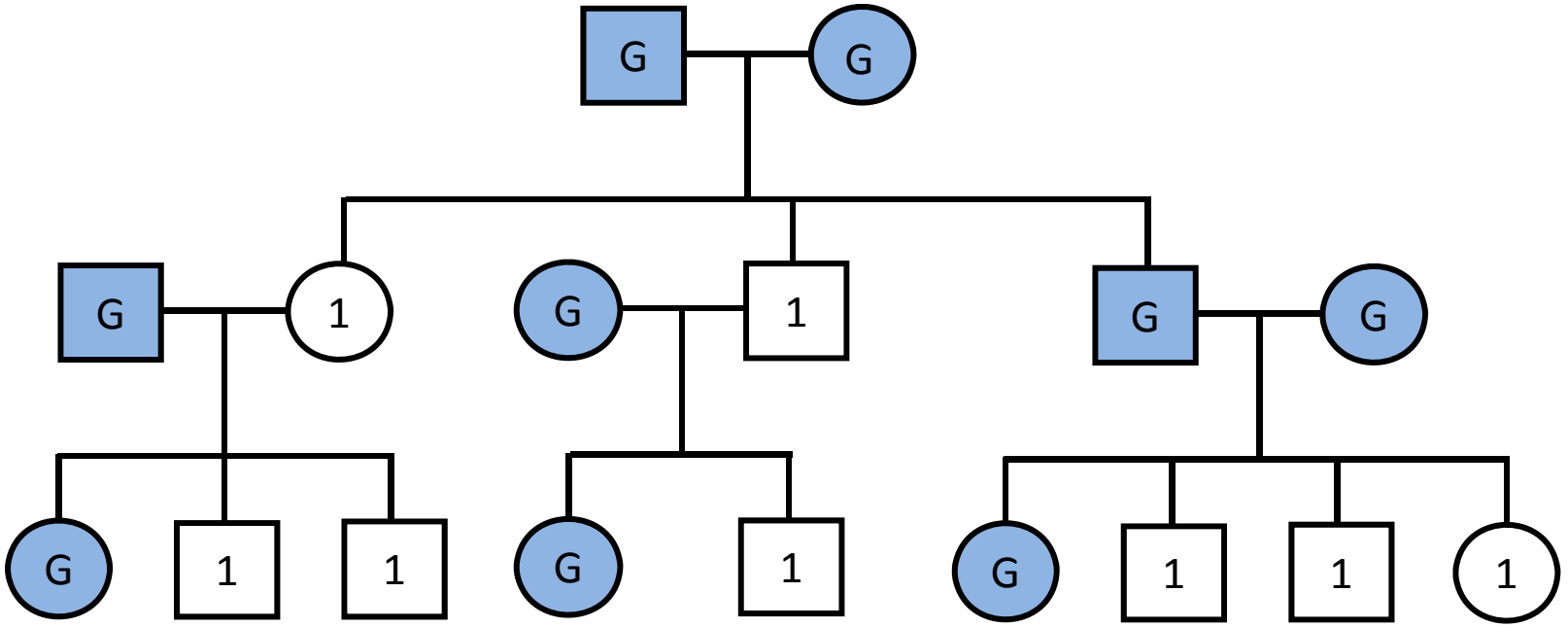
Assuming All Individuals Have Been Genotyped



5 Genomes Sequenced, 12.5 Genomes Analyzed

Who To Sequence?

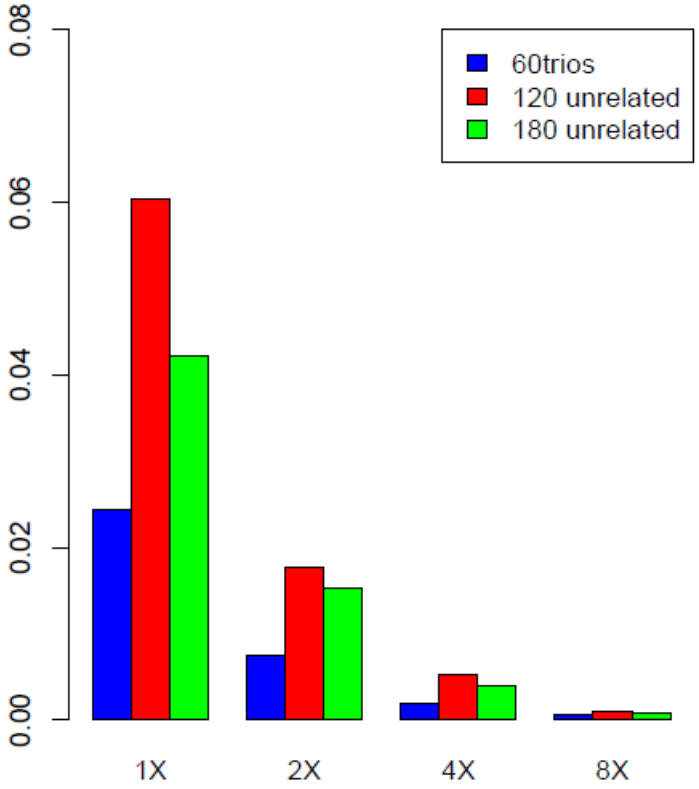
Assuming All Individuals Have Been Genotyped



9 Genomes Sequenced, 17 Genomes Analyzed

Anything to Gain from Sequencing Trios?

Improved Accuracy at Heterozygous Sites



- Sequencing trios improves genotype call accuracy
  - At low coverage …
  - Smaller gain w/deep coverage
- Leads to similar numbers of detected variants
  - At low coverage …
  - No gain w/deep coverage
- Improved haplotype accuracy

Wei Chen and Bingshan Li

Assembling Sequences In Sardinia

Sardinian team led by Francesco Cucca, Serena Sanna, Chris Jones

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: ~35 Gb/lane

- Discovered and genotyped >17M genetic variants so far.

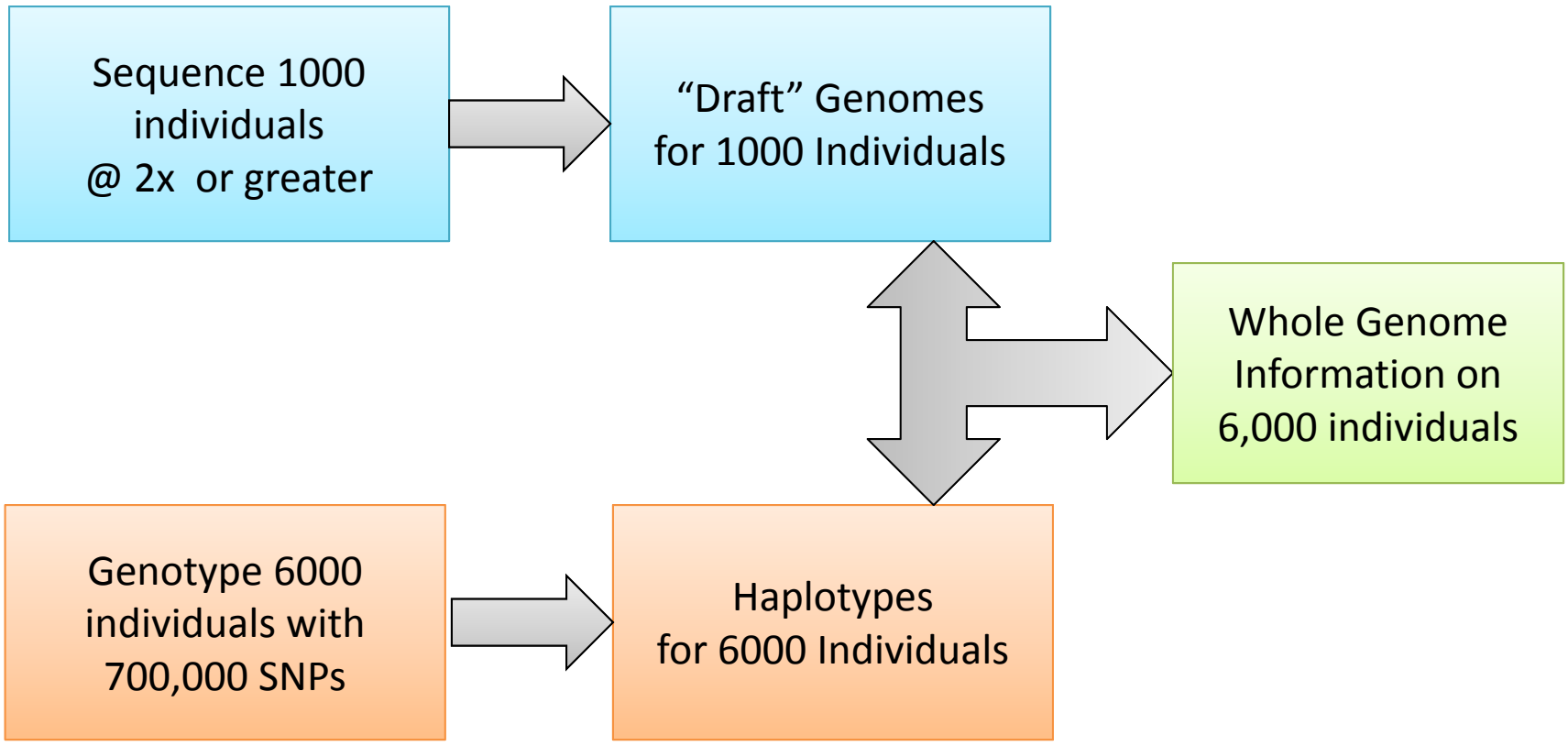
Fabio Busonero, Hyun Min Kang, Bingshan Li

Accuracy Of Variant Calls

	Genotype Class		
Sample Set	**Homozygous Reference**	**Heterozygotes**	**Homozygous Non-Reference**
Analysis Ignoring Relatedness			
66 Samples	2.1	8.7	3.2
226 Samples	1.0	5.5	1.9
508 Samples	0.2	1.3	0.4
Trio-Aware Analysis			
66 Samples	1.0	5.4	1.5
226 Samples	0.6	3.6	1.1
508 Samples	0.2	0.6	0.4

Carlo Sidore, Hyun Min Kang, Serena Sanna

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		Imputation Accuracy (r^2) IN SARDINIA		
Panel	Chr	MAF 1-3% (SNP)	MAF 3-5% (SNP)	MAF >5% (SNP)
1000G (563)	20	0.75	0.88	0.94
Sardinia (508)	20	0.90	0.95	0.97

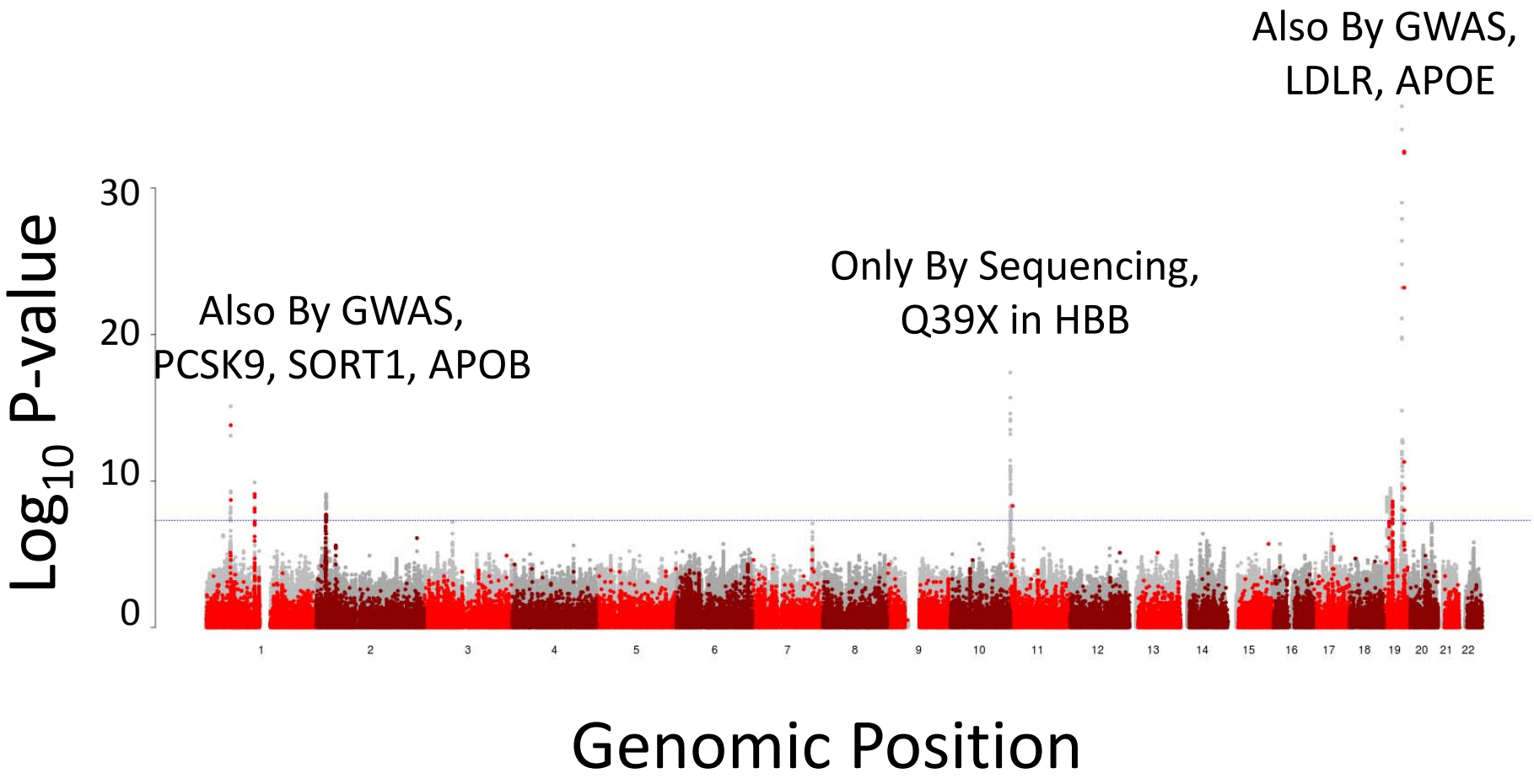
Data: Sardinia data set; chr20; Imputation-panel: Affy1M; Evaluation-panel: Metabochip

Sardinian Haplotypes Are Not Great for Imputation Outside Sardinia

Reference		Imputation Accuracy (r^2) OUTSIDE SARDINIA		
Panel	**Chr**	**MAF 1-3%**	**MAF 3-5%**	**MAF >5%**
1000G Nov (563)	20	0.83	0.85	0.94
Sardinia (508)	20	0.77	0.83	0.92

Data: GAIN data set; chr20; Imputation-panel: Affy1M; Evaluation-panel: Perlegen Custom Array

What Do We See Genomewide?
LDL Cholesterol



LDL Genetics In Lanusei, 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.

Parting Thoughts …

- Sequencing enables new genetic discoveries
- Achieving sufficient sample sizes is a challenge
  - Take advantage of efficient study designs
  - Take advantage of interesting sample sets
- Many challenges remain in analyzing data
  - At least as tough as generating it!

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)