SeqShop Day 2

Variant Calling & Filtering : SNPs

Hyun Min Kang
Department of Biostatistics
University of Michigan

In this Lecture…

- Calling single nucleotide polymorphisms (SNPs)
- Improving quality of SNP calls by filtering
- Designing large-scale sequencing studies
- Detecting and estimating DNA contamination

Shotgun Sequence Reads

ACTGGTCGATGCTAGCTGATAGCTAGCTA

GCTGATGAGCCCGATCGCTGCTAGCTCG

AGCTGATAGCTAGCTAGCTGATGAGCCCGA

GAGCCCGATCGCTGCTAGCTCGACG

- Typical short read might be 70-150 bp long and not very informative on its own

- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

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

Read Alignment

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

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

Calling Consensus Genotype - Details

- Each aligned read provides a small amount of evidence about the underlying genotype
  - Read may be consistent with a particular genotype …
  - Read may be less consistent with other genotypes …
  - A single read is never definitive

- This evidence is cumulated gradually, until we reach a point where the genotype can be called confidently

- Let's outline a simple approach …

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-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

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-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

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

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 incorporating information from other individuals and flanking sites to assign a genotype.

Shotgun Sequence Data

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$\Pr(G|R) = \frac{\Pr(R|G)\pi(G)}{\sum_g \Pr(R|g)\pi(g)}$$

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

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) =** 0.00034	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

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) =** 0.00034	**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

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

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

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

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

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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

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) =** 0.81	**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.

Summary : 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

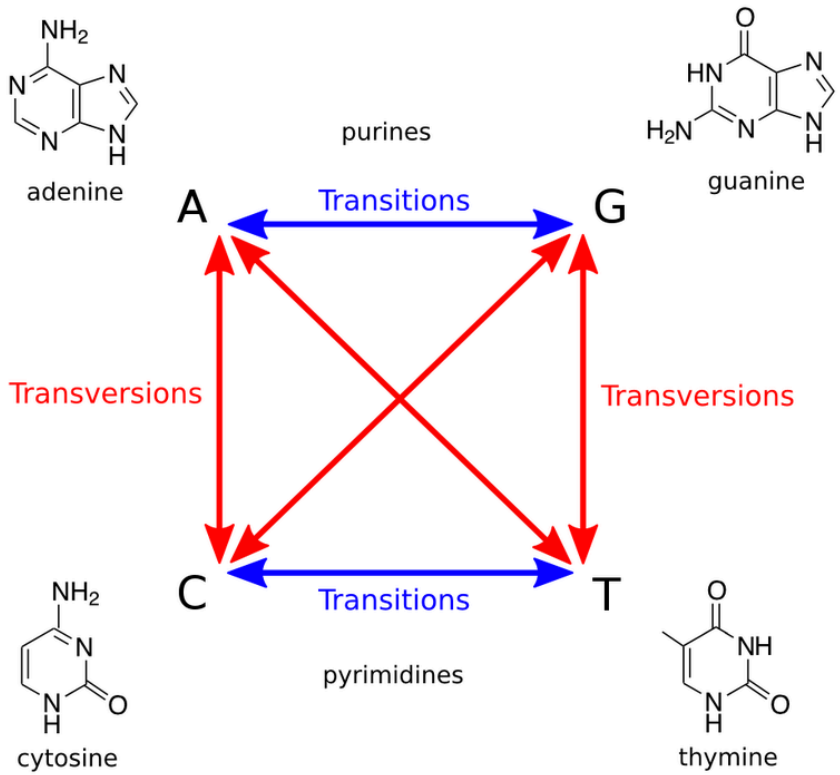
False Positive / Negative SNP calls

- In an ideal world (no errors in alignment or base quality)
  - Posterior probability quantifies statistical evidence
  - Determine posterior probability cutoff to calibrate between false positive vs. false negative

- In the real world..
  - Alignment errors create false positive SNPs
    - Segmental duplication mapped to the same position
    - Alignment can be ambiguous at the end of the read
  - Low depth limits the power to detect rare SNPs
    - If the SNP is not shared across individuals, insufficient sequence depth limits the SNP detection power

How can we Know the Quality of a SNP call set?

- Overlap with existing SNPs (dbSNP, HapMap)
- Ts/Tv ratio is an effective surrogate for SNP quality



Genome-wide
In a true SNP, expected Ts/Tv ~ 2

In a (random) false SNP
expected Ts/Tv ~ 0.5

In exomes (coding sequences)
True SNP has expected Ts/Tv ~ 3

Nonsynonymous SNPs : Ts/Tv ~ 2
Synonymouys SNPs : Ts/Tv > 5
(Calculated from codon frequency)

Summarizing Quality of SNP Calls

1000 Genomes Phase 3 call sets using GotCloud

Call set	# SNPs (# WGS)	in dbSNP (129)	Known Ts/Tv	Novel Ts/Tv	HapMap3 Sensitivity
Initial Call Set	1,787,005 (79M)	11.5%	2.29	2.15	98.8%
Filtered (PASS) Call Set	1,541,273 (68M)	12.4%	2.35	2.37	98.5%
Filtered-out (FAIL) Call Set	245,732 (11M)	5.7%	1.68	1.29	0.3%

Per Base Alignment Qualities

Short Read

GATAGCTAGCTAGCTGATGA GCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Per Base Alignment Qualities

Should we insert a gap?

Short Read

GATAGCTAGCTAGCTGATGAGCC-G

5'-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3'

Reference Genome

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality

Short Read

GATAGCTAGCTAGCTGATGAGCCG

5'-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3'

Reference Genome

How to Tell Good from Bad: Example

Reference : … AGGTCTAA … … GAATTACA …

Sample 1

…	C	…	
…	T	…	
…	C	…	
…	T	…	
…	T	…	0.6

…	C	…	
…	T	…	
…	T	…	
…	T	…	
…	T	…	0.8

We expect 50:50 read distribution for HET sites
(Allelic imbalance is indicator of alignment artifacts)

How to Tell Good from Bad: Example

Reference : … AGGTCTAA …

Sample 1

…	C	…
…	T	…
…	C	…
…	T	…
…	T	…

0.6

Good

… GAATTACA …

…	C	…
…	T	…
…	T	…
…	T	…
…	T	…

0.8

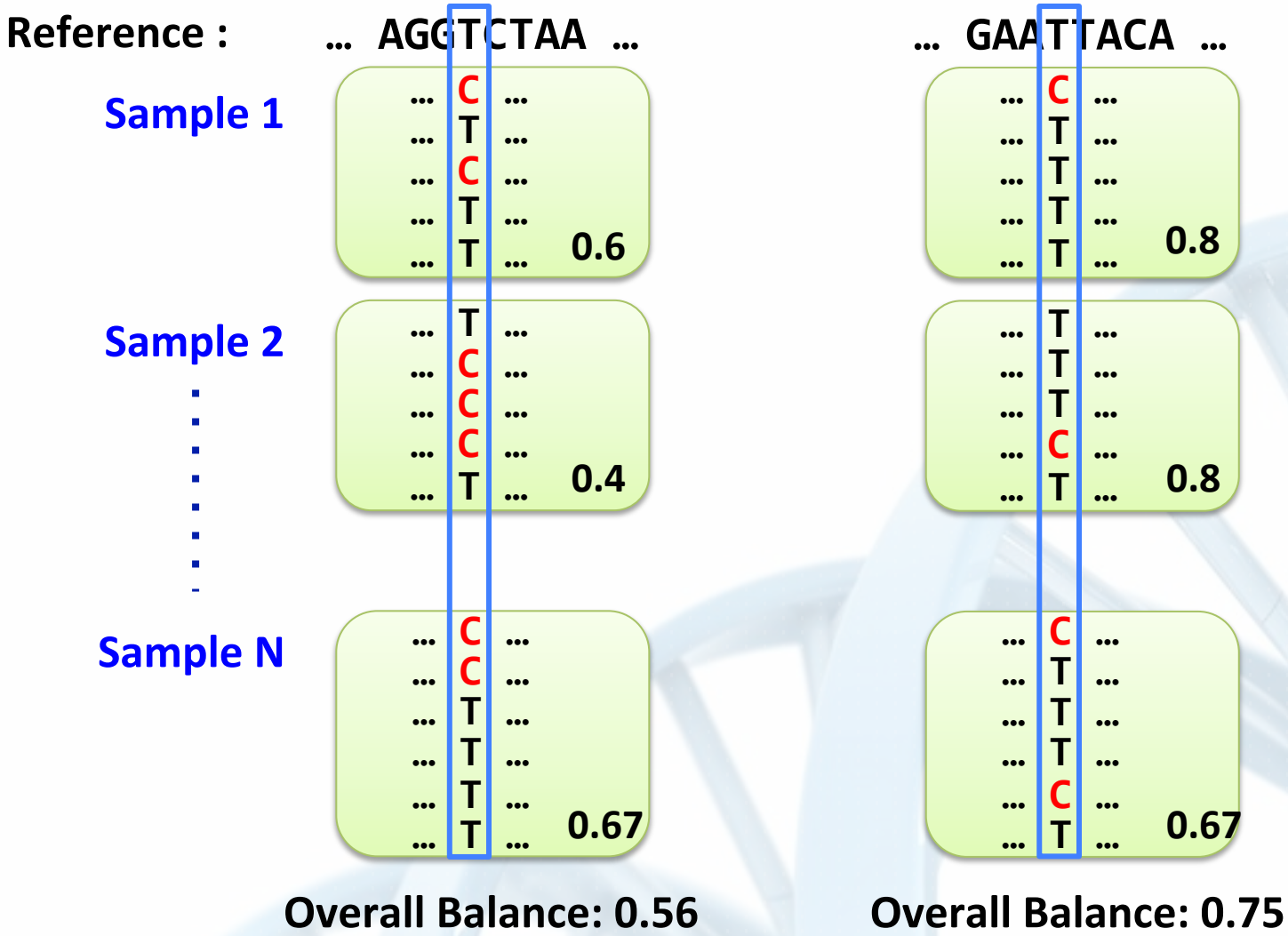
Bad

How to Tell Good from Bad: Example

Hard to tell whether it's random deviation or not on a single sample

Sample N

Multi-sample Filtering is Informative

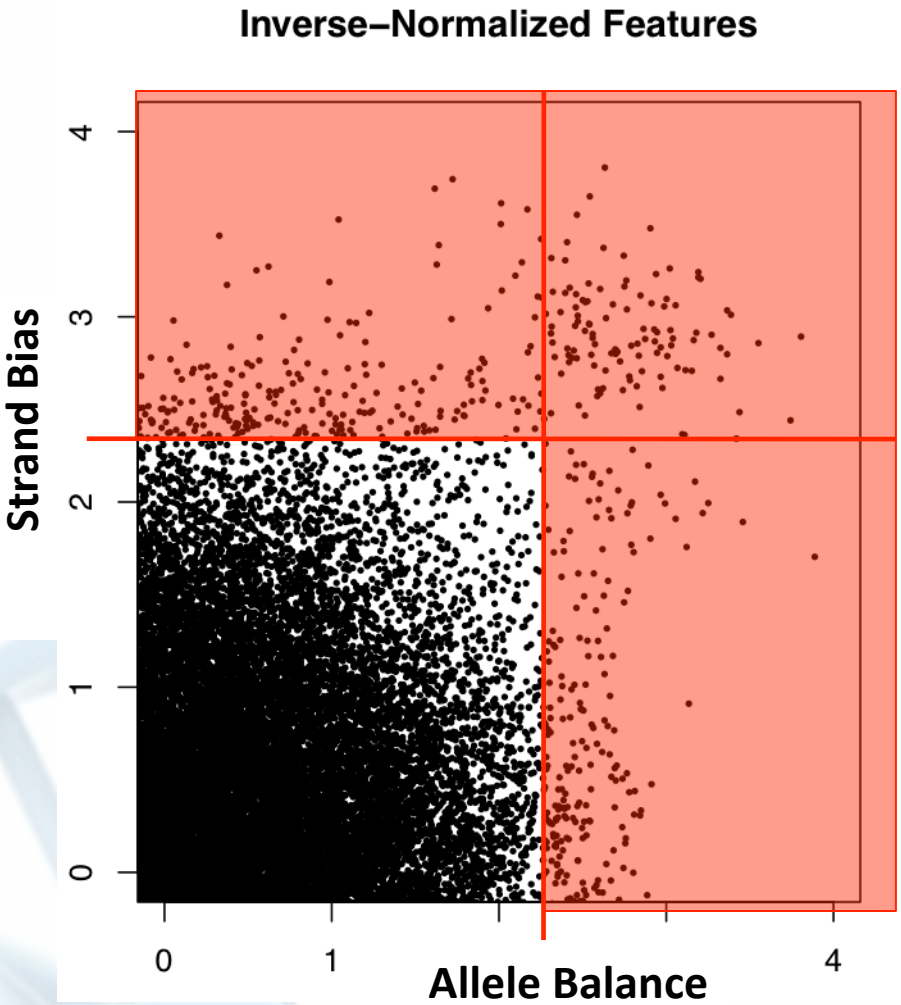


Filtering Criteria Examples

Feature	Description
Depth	Overall depth across samples
QUAL	Overall genotype confidence
Allele Balance	(# REF)/(# ALT) in HET sites
Strand Bias	Correlation of ALT allele with +/- strand
Cycle Bias	Correlation of ALT allele with read cycle
Mapping Quality	Average quality of aligned sequences
HWE	Hardy-Weinberg Equilibrium Test Statistics

Hard Filtering by Individual Thresholds

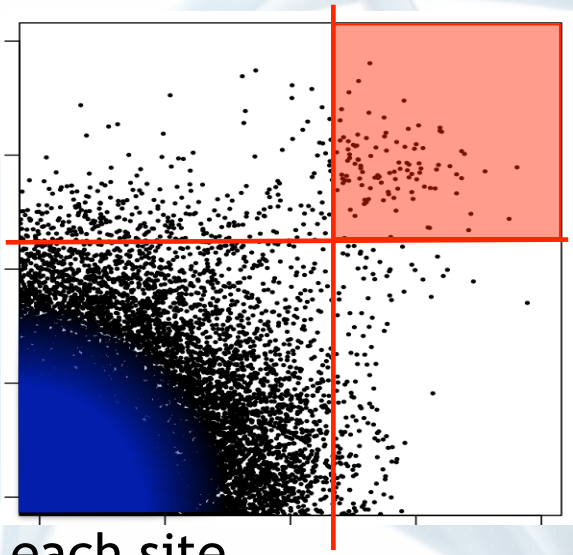
- Problems
  - False negative increases with number of filters
  - Too many knobs to turn (thresholds)



Filtering by Supervised Learning

- Use features to train a support vector machine (SVM)
  - Can be trained using suspected positive/negative examples
  - Provides single score from all features combined

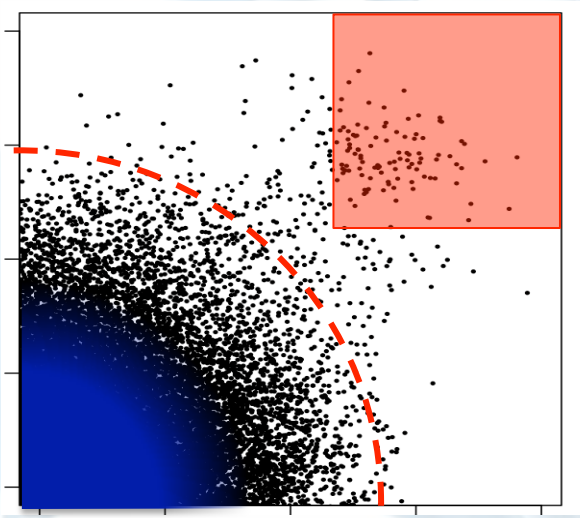
- Training
  - Positive examples
    - Known polymorphic sites
  - Negative examples
    - Filtered out by multiple hard filters
  - Input
    - All individual features collected for each site



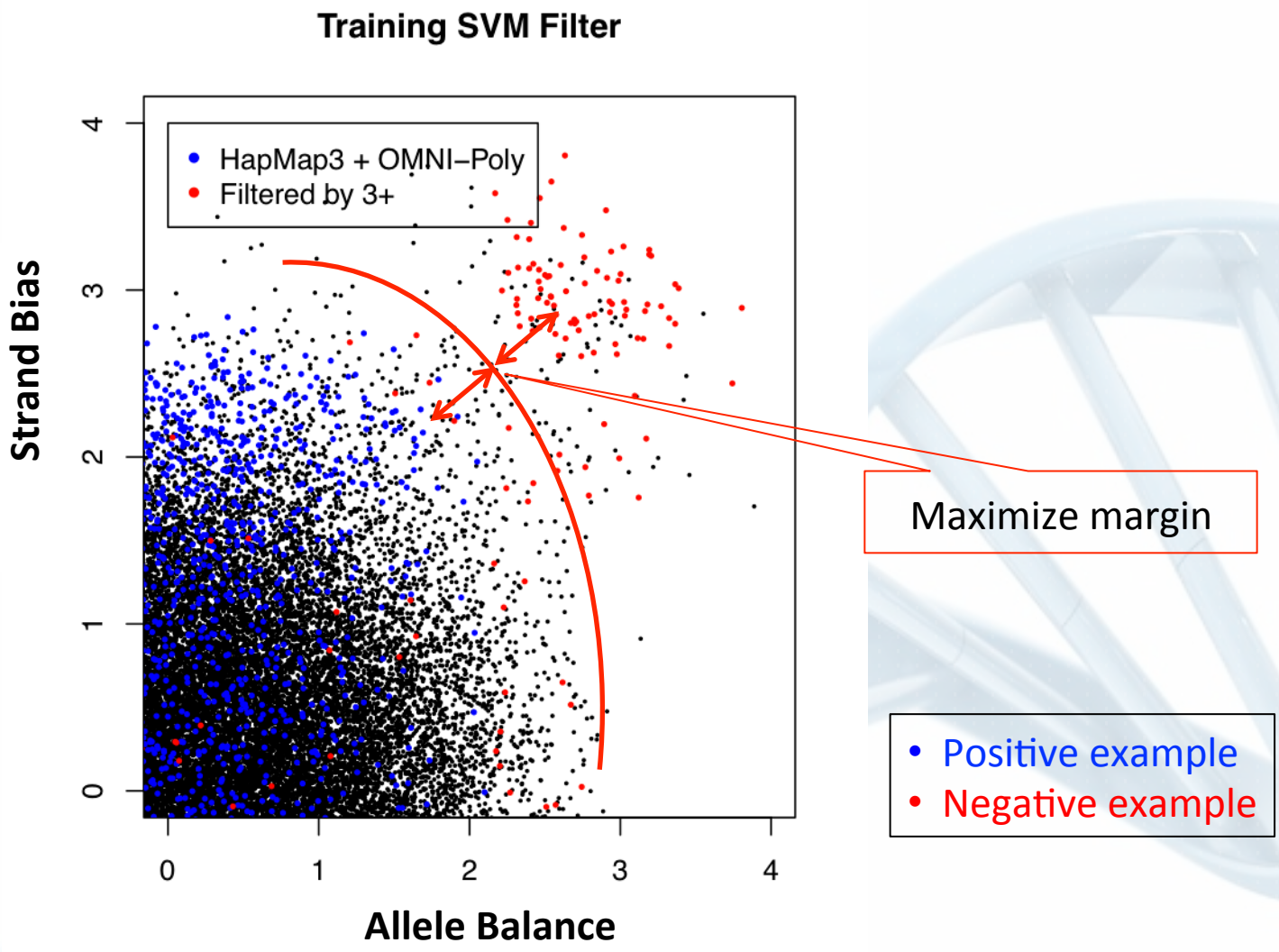
Filtering by Supervised Learning

- Use features to train a support vector machine (SVM)
  - Can be trained using suspected positive/negative examples
  - Provides single score from all features combined

- Training
  - Positive examples
    - Known polymorphic sites
  - Negative examples
    - Filtered out by multiple hard filters
  - Input
    - All individual features collected for each site

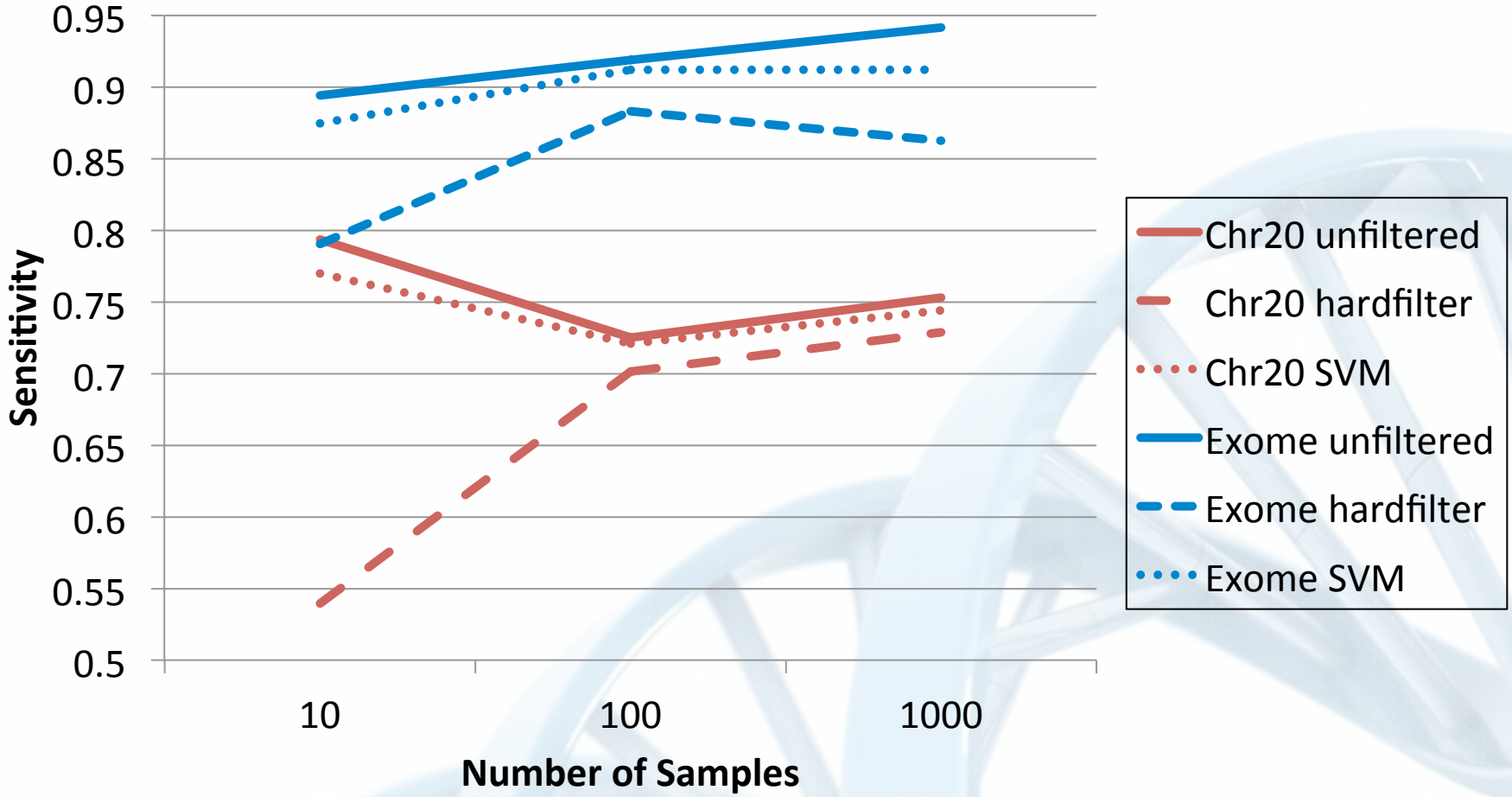


Training SVM with Examples

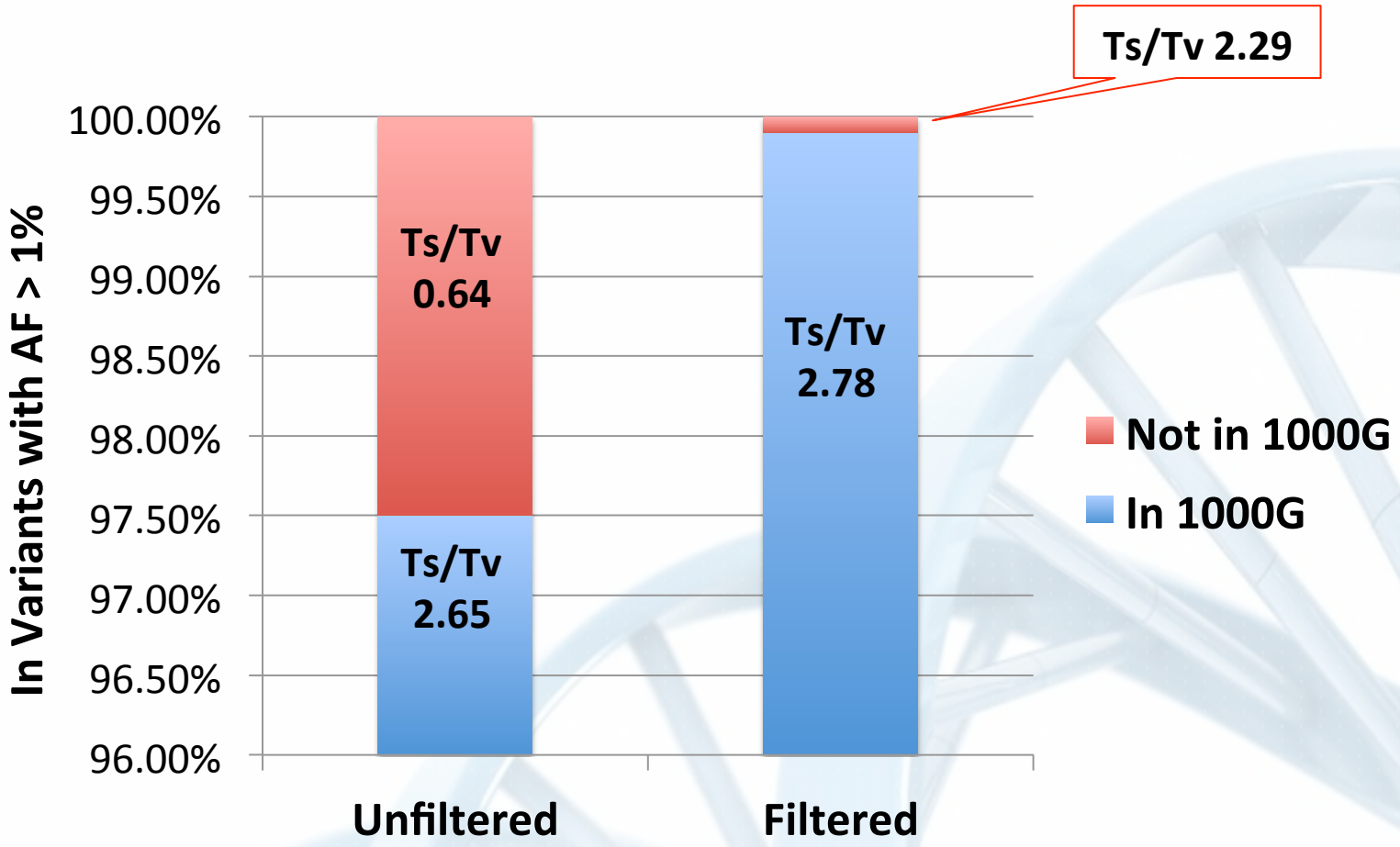


>20 dimensional feature set was used for final filtering under nonlinear kernel space

Improved Sensitivity by SVM



Results in Exome Sequencing Project



Summary : Variant Filtering

- Initial SNP calls typically contain false positive variants
- Ts/Tv, overlap with external resources are good surrogates of SNP quality
- Various features from sequence data can be used to extract high quality variants
- Combining information across multiple samples using SVM further improves the performance of variant filtering

How Much Variation In a Genome?

- 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 In an Exome?

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

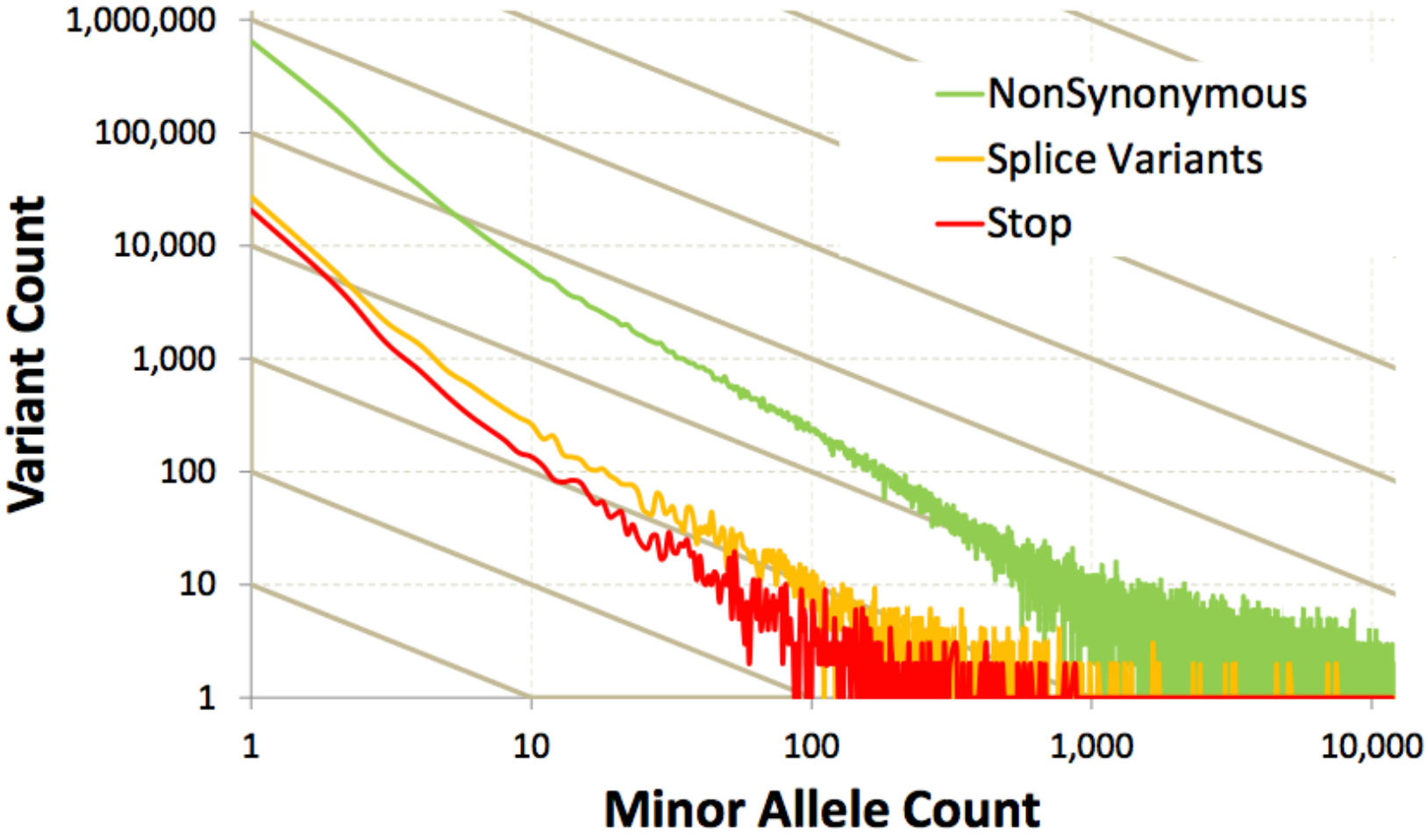
Lots of Rare Functional Variants to Discover

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

Low-coverage or Deep Sequencing?

- 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?**

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

c	G	a	G	A	t	c	T	c	C	t	T	c	T	t	c	t	g	T	G	c
C	g	A	g	a	t	C	T	C	C	C	g	a	c	C	t	c	a	t	g	g
C	C	A	a	G	c	t	C	T	t	t	t	c	t	t	c	t	g	T	G	c
c	g	a	a	g	c	t	C	T	T	T	t	C	t	t	c	t	g	t	g	c
c	g	a	g	a	c	T	c	t	C	c	g	A	C	C	t	t	A	T	G	c
t	g	g	g	a	t	C	t	C	C	c	G	A	C	C	t	C	A	t	G	G
C	G	A	g	A	t	c	t	c	c	c	G	a	C	c	t	T	g	T	g	c
c	g	a	g	a	c	t	C	t	T	t	T	c	t	t	t	t	g	t	A	c
C	G	a	g	A	c	t	C	T	c	c	g	a	c	C	T	c	G	t	g	c
C	G	A	A	g	c	T	c	t	T	t	T	c	T	t	C	T	g	t	G	C
c	G	A	g	A	T	C	t	c	C	t	T	c	T	T	c	t	g	t	G	c
c	g	A	g	a	t	c	t	c	C	C	g	A	C	c	T	C	A	T	G	g
c	c	A	a	G	c	t	C	t	T	T	t	c	t	T	c	T	G	t	G	C
C	G	A	a	g	c	T	c	t	T	t	t	c	T	T	c	T	g	t	G	C
c	g	a	G	A	C	t	C	t	c	c	g	a	c	c	t	t	a	T	G	c
T	g	g	g	a	T	c	t	C	c	c	g	a	C	C	t	c	a	t	g	g
c	g	a	G	A	T	C	t	C	C	c	G	a	c	C	T	T	g	t	G	C
c	g	a	G	A	c	T	c	T	T	t	T	c	T	T	t	T	g	t	a	c
c	G	A	G	a	c	T	c	T	c	c	G	A	c	c	T	C	G	t	g	C
c	g	A	A	g	c	T	c	t	t	t	t	c	t	t	c	t	g	t	G	c

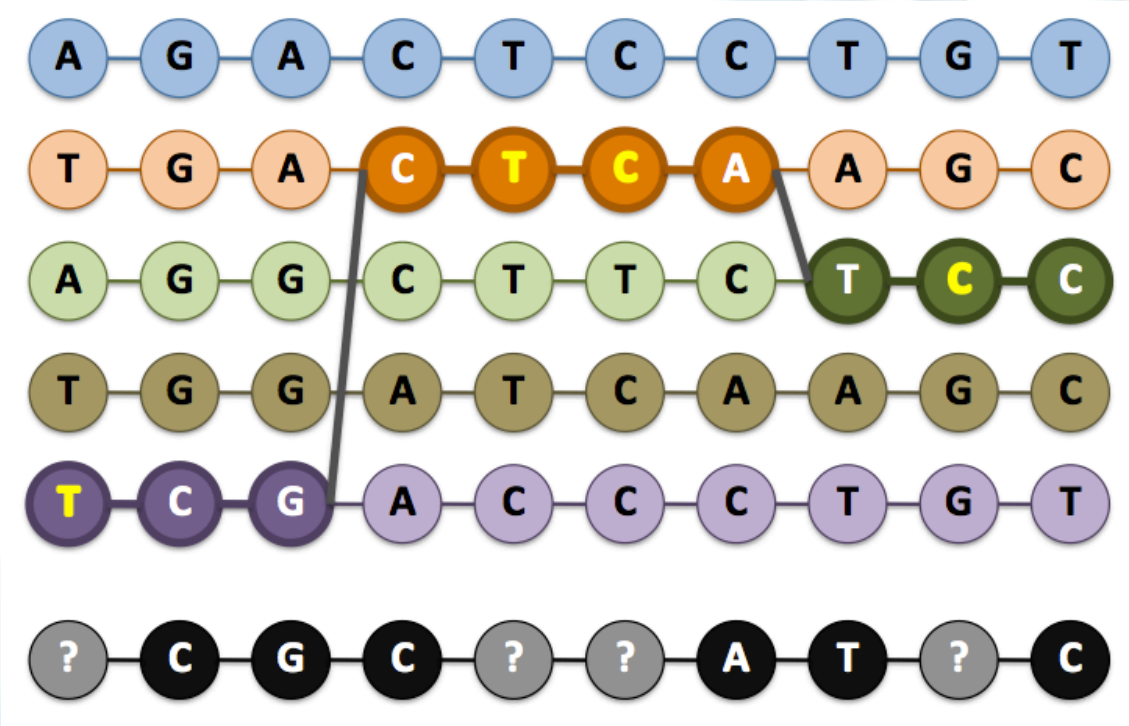
Silly Cartoon View of Shot Gun Data

```
. G . G A . . T . C . T . T . . . . T G .
C . A . . . C T C C C . . . C . . . . . .
C C A . G . . C T . . . . . . . . . T G .
. . . . . . . C T T T . C . . . . . . . .
. . . . . . T . . C . . A C C . . A T G .
. . . . . . C . C C . G A C C . C A . G G
C G A . A . . . . . . G . C . . T . T . .
. . . . . . . C . T . T . . . . . . A .
C G . . A . . C T . . . . . C T . G . . .
C G A A . . T . . T . T . T . C T . . G C
. G A . A T C . . C . T . T T . . . . G .
. . A . . . . . . C C . A C . T C A T G .
. . A . G . . C . T T . . . T . T G . G C
C G A . . . T . . T . . . T T . T . . G C
. . . G A C . C . . . . . . . . . . T G .
T . . . . T . . C . . . . C C . . . . . .
. . . G A T C . C C . G . . C T T . . G C
. . . G A . T . T T . T . T T . T . . . .
. G A G . . T . T . . G A . . T C G . . C
. . A A . . T . . . . . . . . . . . G .
```

Haplotype-aware Genotype Refinement

- People share ‘blocks’ of genotypes

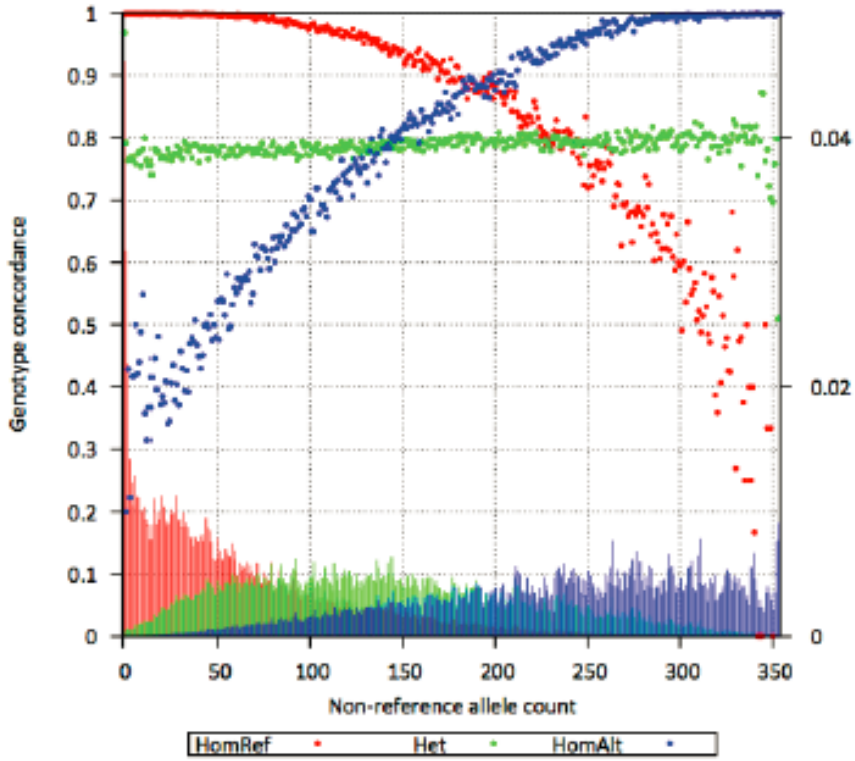
- Haplotype-phasing improves genotype accuracy by correcting unlikely genotypes and filling in missing genotypes



Haplotype-Aware Analysis Improves Genotype Accuracy

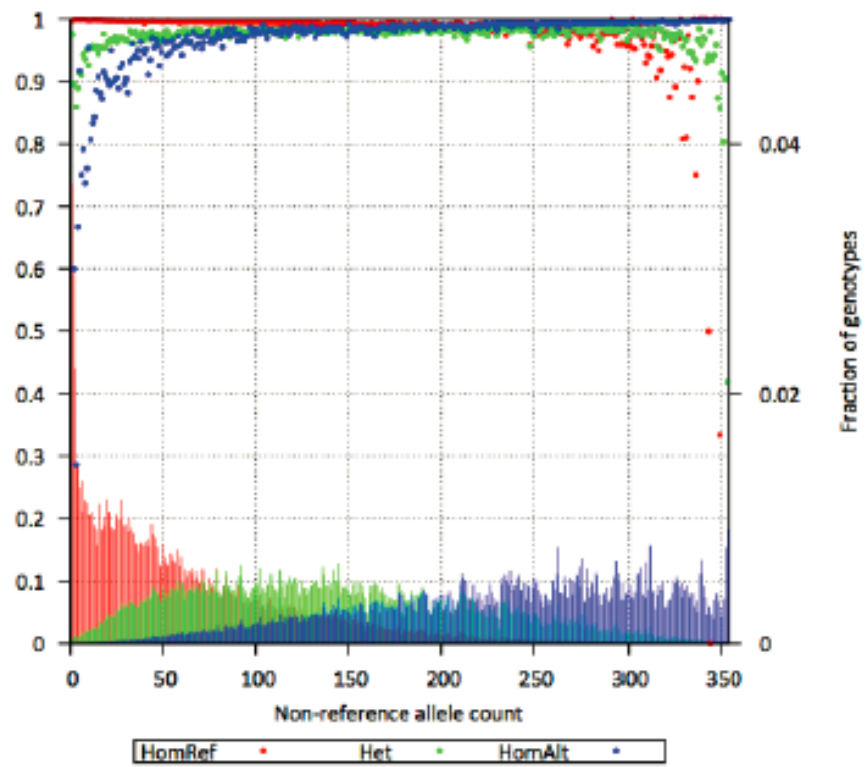
Single Site Analysis

– 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

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

Genotype Accuracy Further Improves with Sample Sizes

Concordance between 1000G Phase 3 and Deep Sequence Data (CGI)

Gold	R/R	R/A	A/A	All
SNPs	0.9998	0.9930	0.9983	0.9994
Indels	0.9990	0.9889	0.9923	0.9982

Summary : Design of Sequencing Studies

- 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.
  - Haplotype-aware analysis works even better

- Low pass analyses are attractive for complex disease association studies.

- To study rarer variants better, what is the optimal coverage?

DNA Sample Contamination

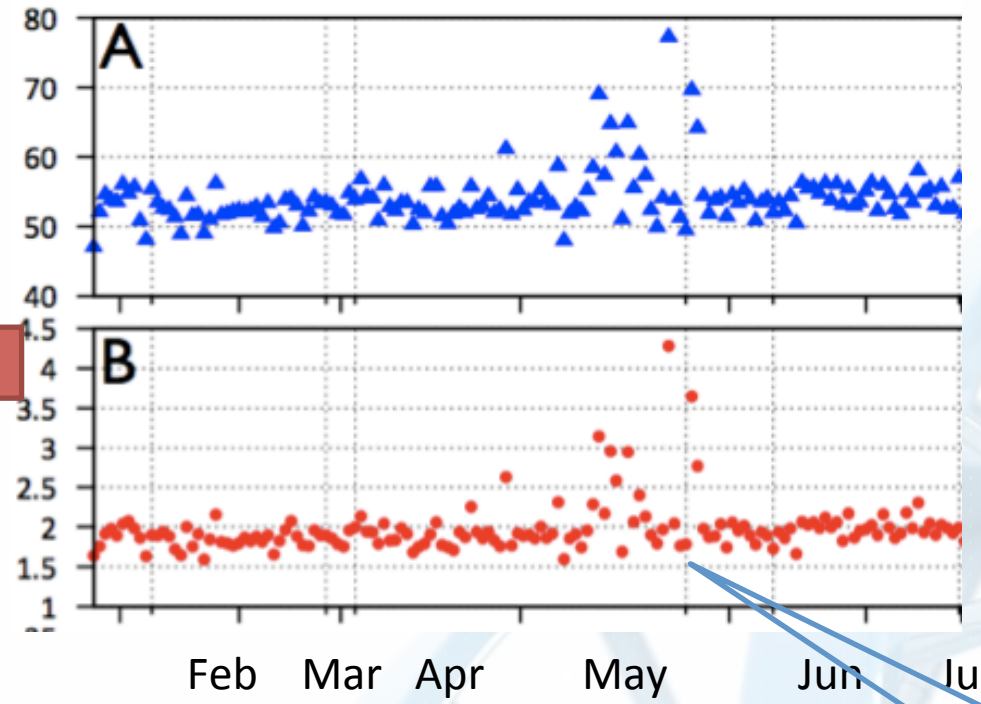
*Picture from D. Figarelli, *National Forensic Science Tech. Center*

Unexpected Results in SNP calls

Per-sample genotype count

# of HETs

#HETs/#HOMs



Sequencing Date in 2010

Ref	Dad	Mom
A	A	A
A	A	B
A	B	A
A	B	B

Expected Het:Hom is 2:1 (or less)

Something changed?

Reference-Aligned Sequence Reads

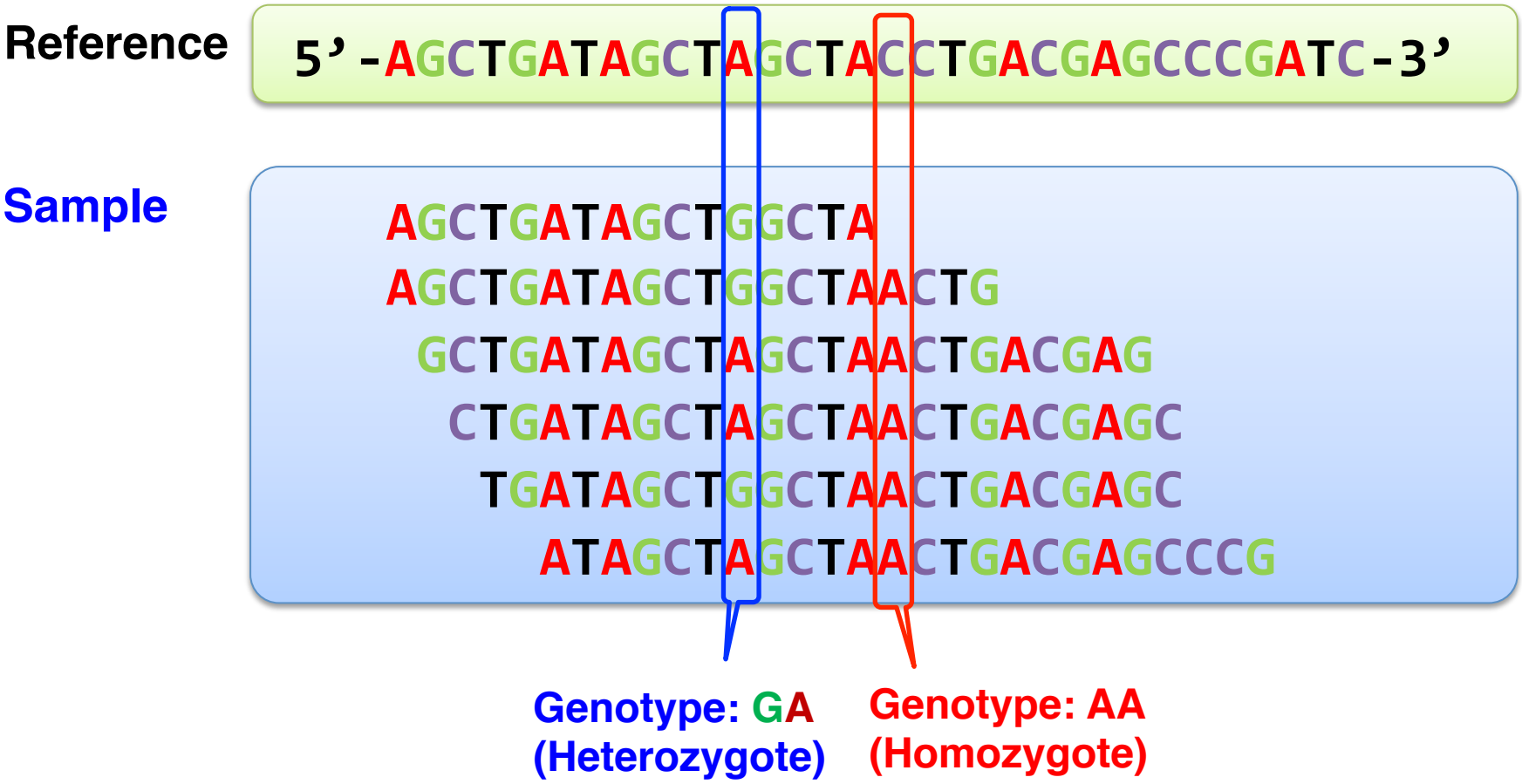
Reference

5'-AGCTGATAGCTAGCTACCTGACGAGCCCGATC-3'

Sample

```
AGCTGATAGCTGGCTA
AGCTGATAGCTGGCTAACTG
 GCTGATAGCTAGCTAACTGACGAG
  CTGATAGCTAGCTAACTGACGAGC
   TGATAGCTGGCTAACTGACGAGCC
     ATAGCTAGCTAACTGACGAGCCCG
```

Single-Nucleotide Polymorphism (SNP)



Base Distribution in Two Samples

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

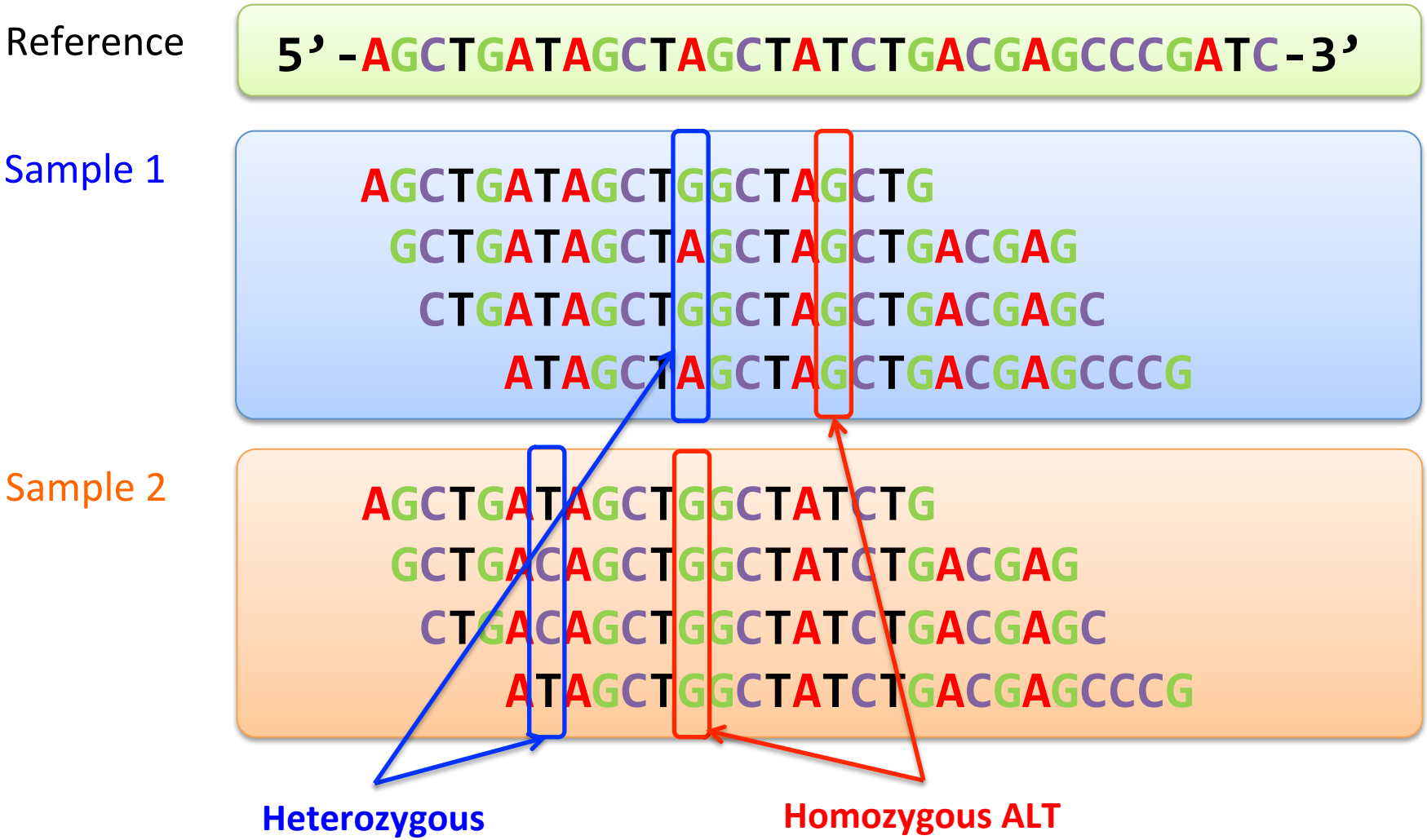
Sample 1

AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG

Sample 2

AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG

Base Distribution in Two Samples



Contamination: Mixture of Samples

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2

```
AGCTGATAGCTGGCTAGCTG
 GCTGATAGCTAGCTAGCTGACGAG
  CTGATAGCTGGCTAGCTGACGAGC
      ATAGCTAGCTAGCTGACGAGCCCG
AGCTGATAGCTGGCTATCTG
 GCTGACAGCTGGCTATCTGACGAG
  CTGACAGCTGGCTATCTGACGAGC
      ATAGCTGGCTATCTGACGAGCCCG
```

Contamination: Changes Base Distributions

Reference

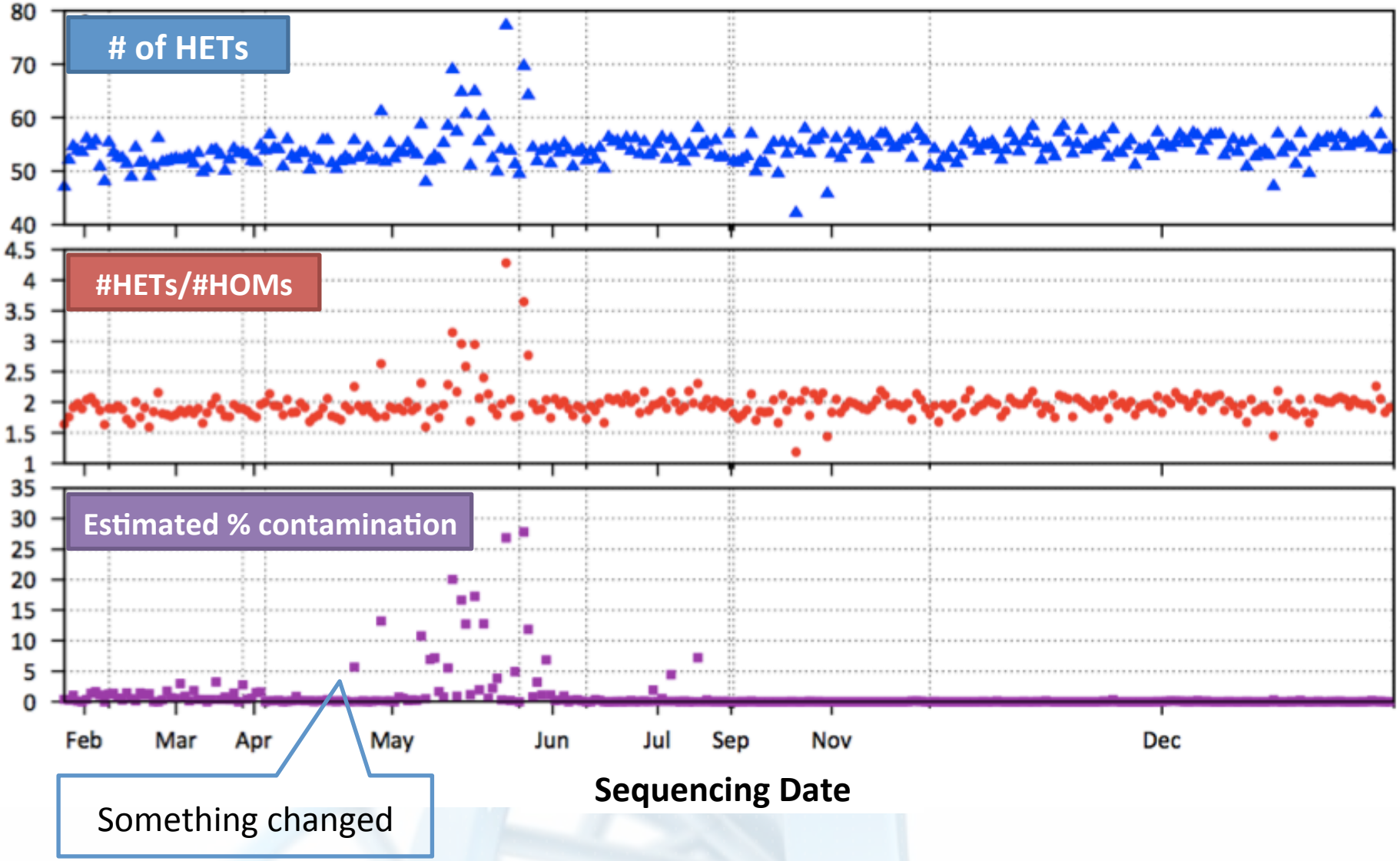
5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2

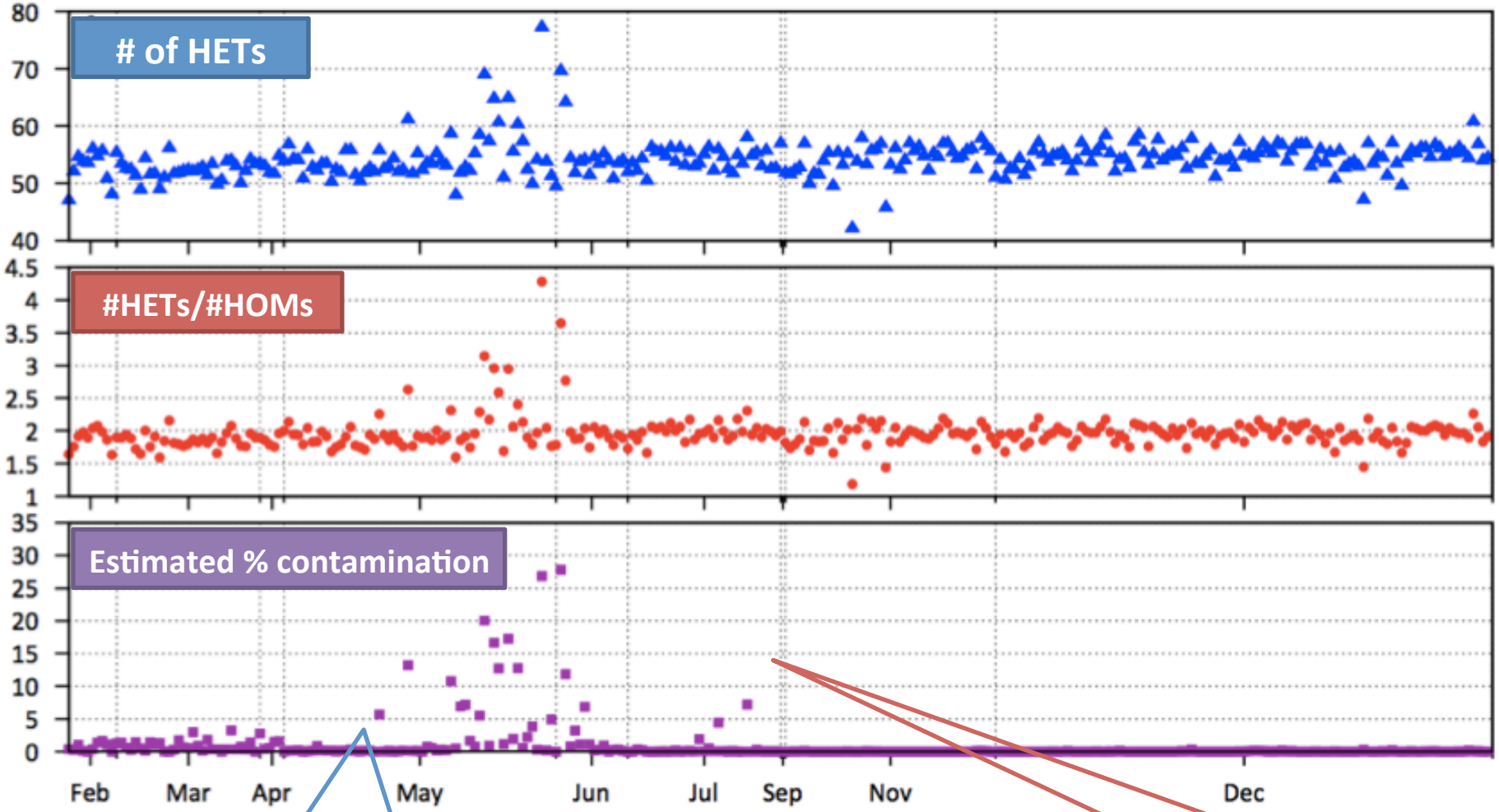
```
AGCTGATAGCTGGCTAGCTG
 GCTGATAGCTAGCTAGCTGACGAG
  CTGATAGCTGGCTAGCTGACGAGC
      ATAGCTAGCTAGCTGACGAGCCCG
AGCTGATAGCTGGCTATCTG
 GCTGACAGCTGGCTATCTGACGAG
  CTGACAGCTGGCTATCTGACGAGC
      ATAGCTGGCTATCTGACGAGCCCG
```

More heterozygote SNPs with biased distribution

Using Mixture Model to Detect and Estimate Contamination



Using Mixture Model to Detect and Estimate Contamination

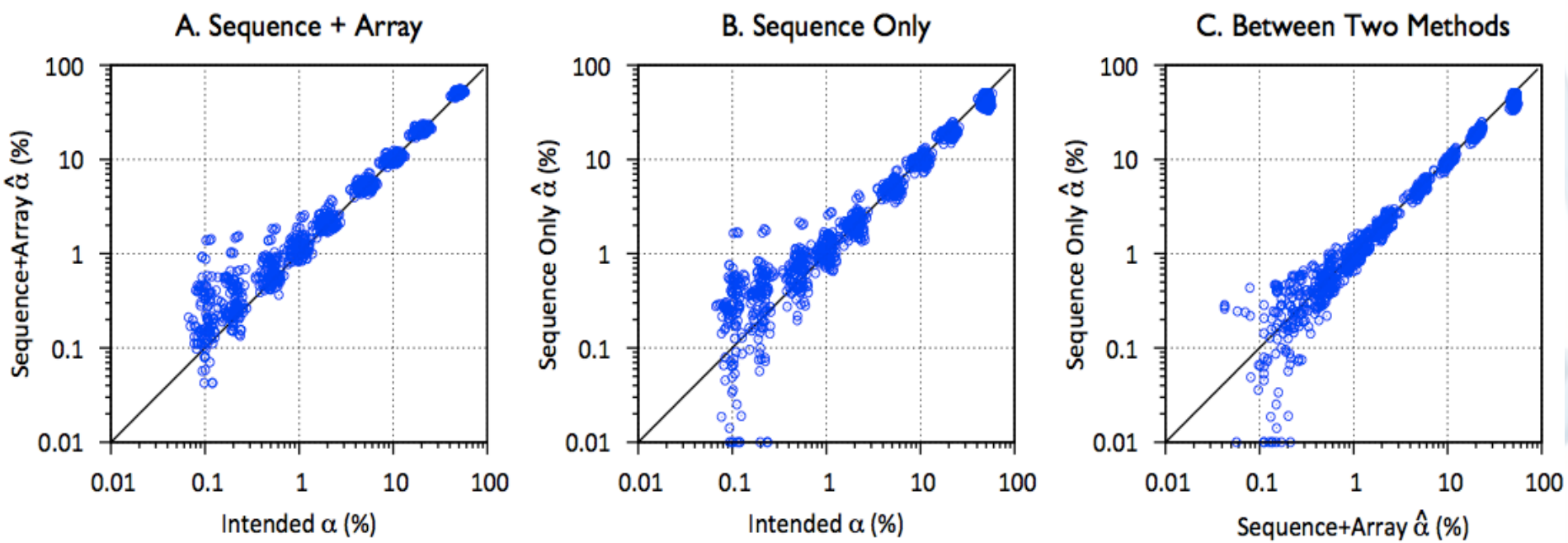


Something changed

No contamination after a protocol change

DNA Contamination can be Precisely Estimated

- Simulated contamination from real sequence data
  - Can accurately detect as low as 1% contamination
  - Works with or without known genotype data



Summary : DNA contamination

- DNA contamination is a practically important problem in the analysis of sequence data
- Detecting and estimating contamination from sequence data is possible via mixture models
- Is it also possible to correct for the contamination effect if exists (not shown today)