Quality Control in Sequence Data

Goo Jun, Gonçalo Abecasis, Mary Kate Wing

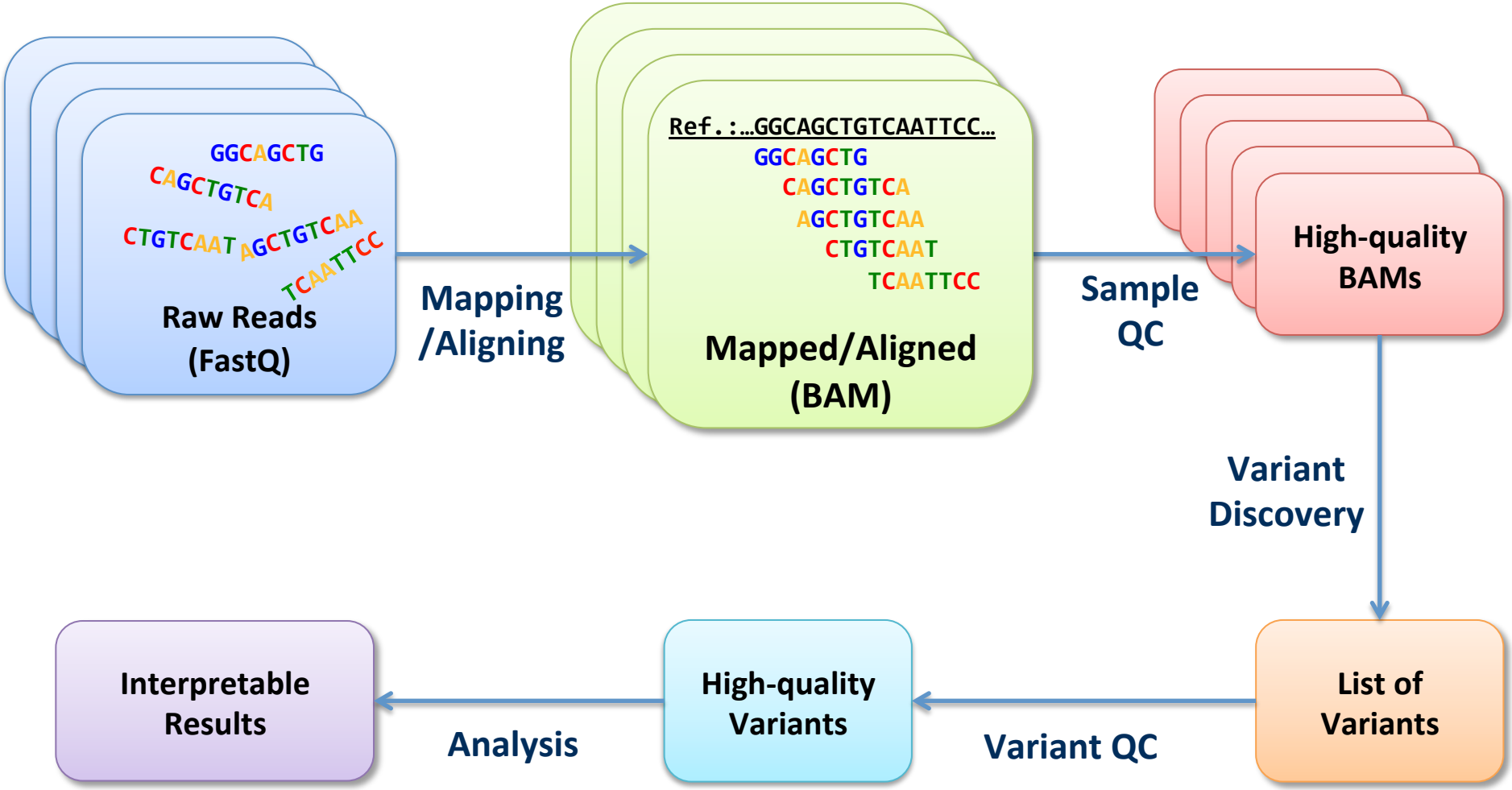
Center for Statistical Genetics & Dept. of Biostatistics

University of Michigan

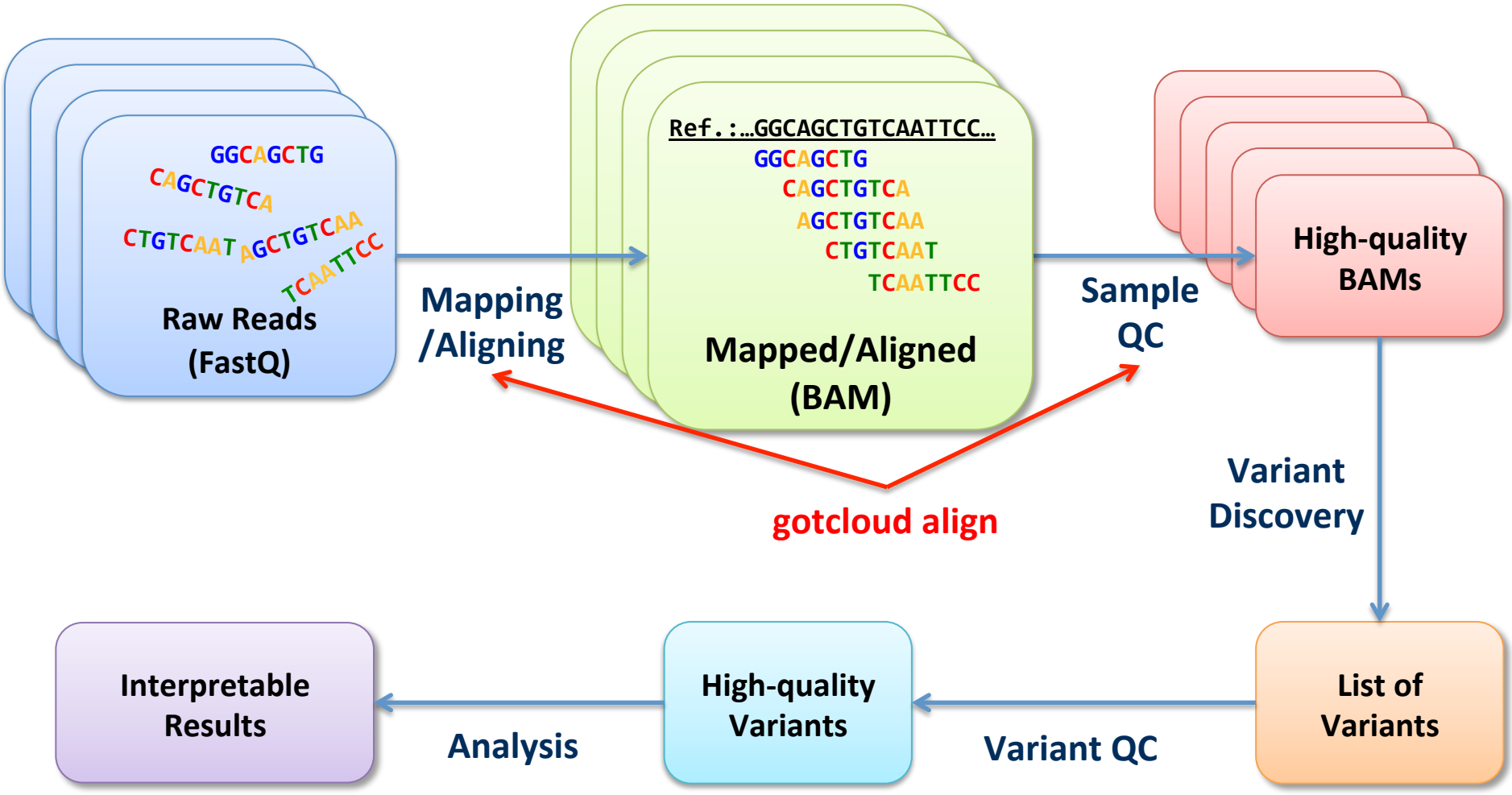
MICHIGAN

CENTER FOR STATISTICAL GENETICS

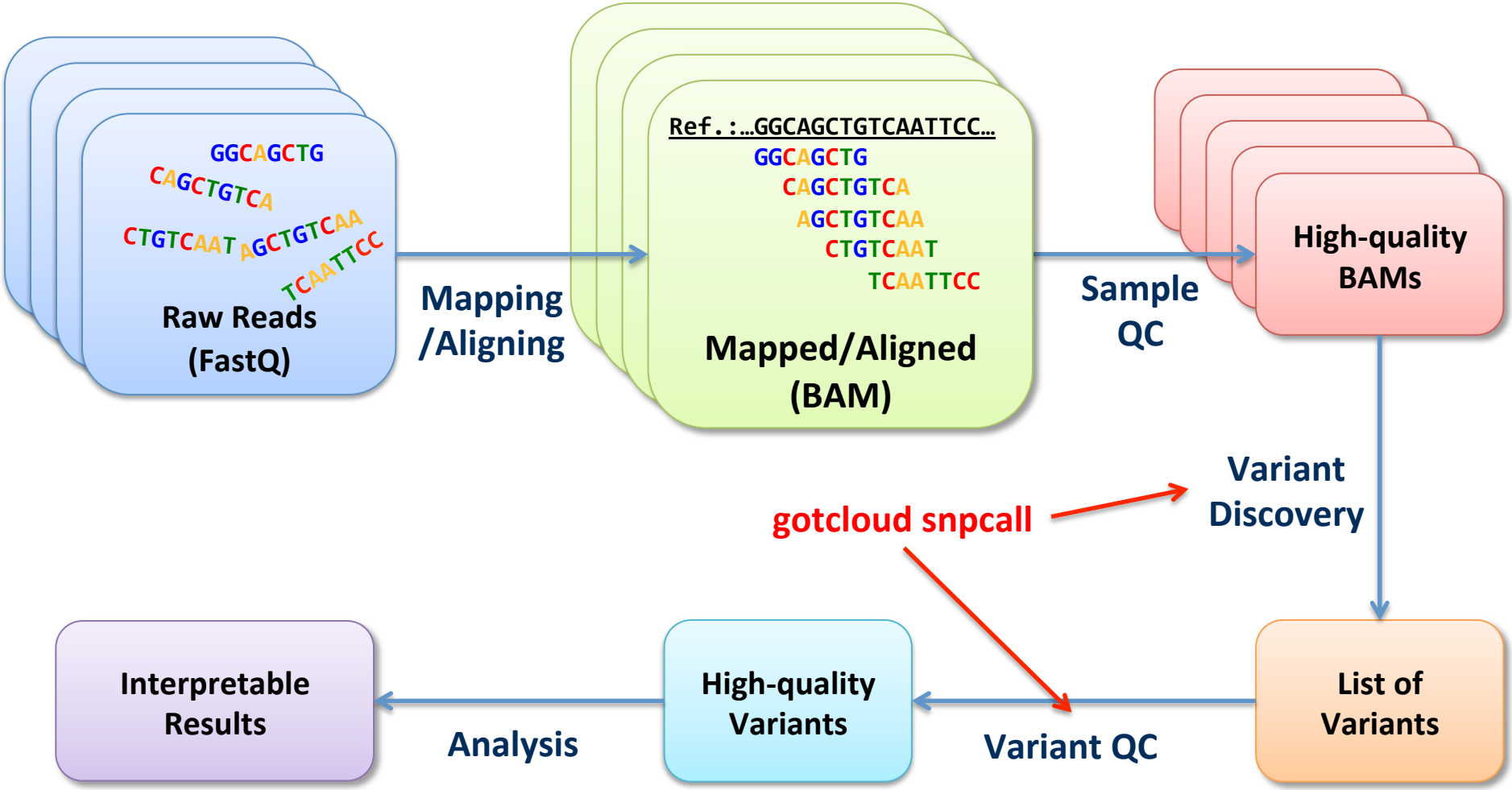
(Re)sequencing Data Analysis Flow



(Re)sequencing Data Analysis Flow



(Re)sequencing Data Analysis Flow



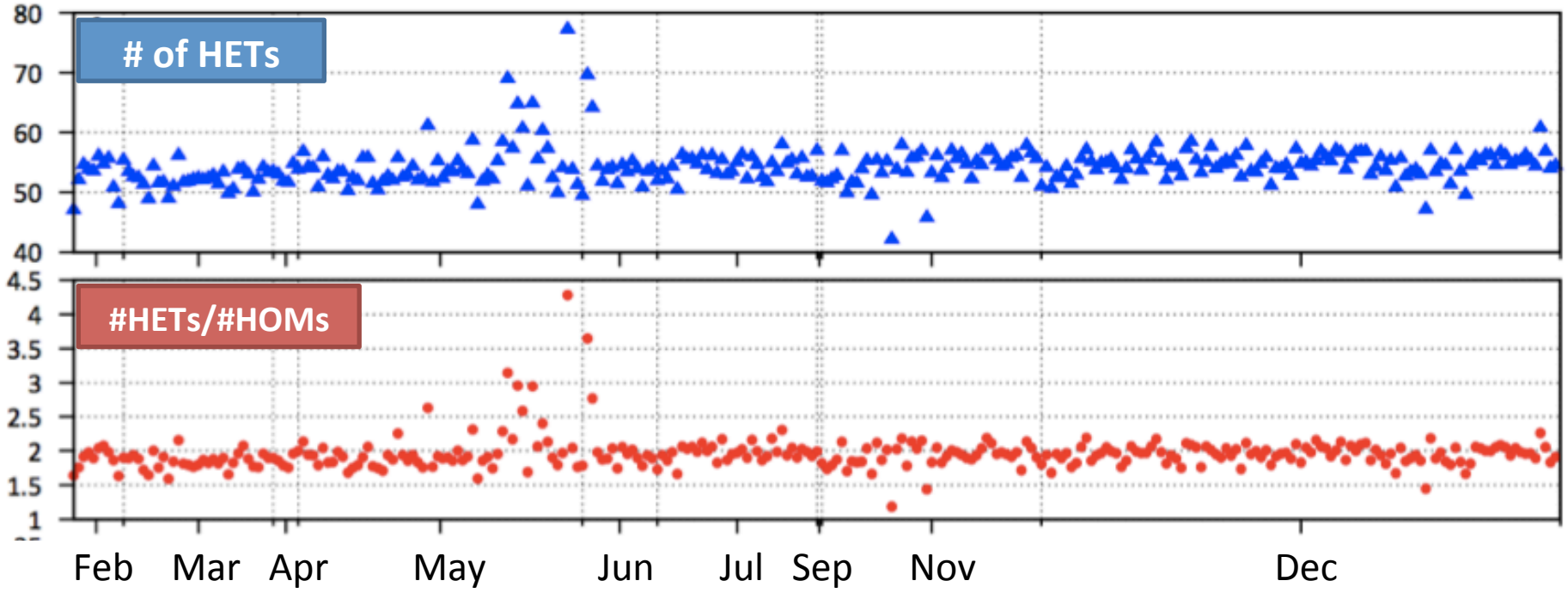
Part 1.

Detecting and estimating sample contamination

DNA Sample Contamination

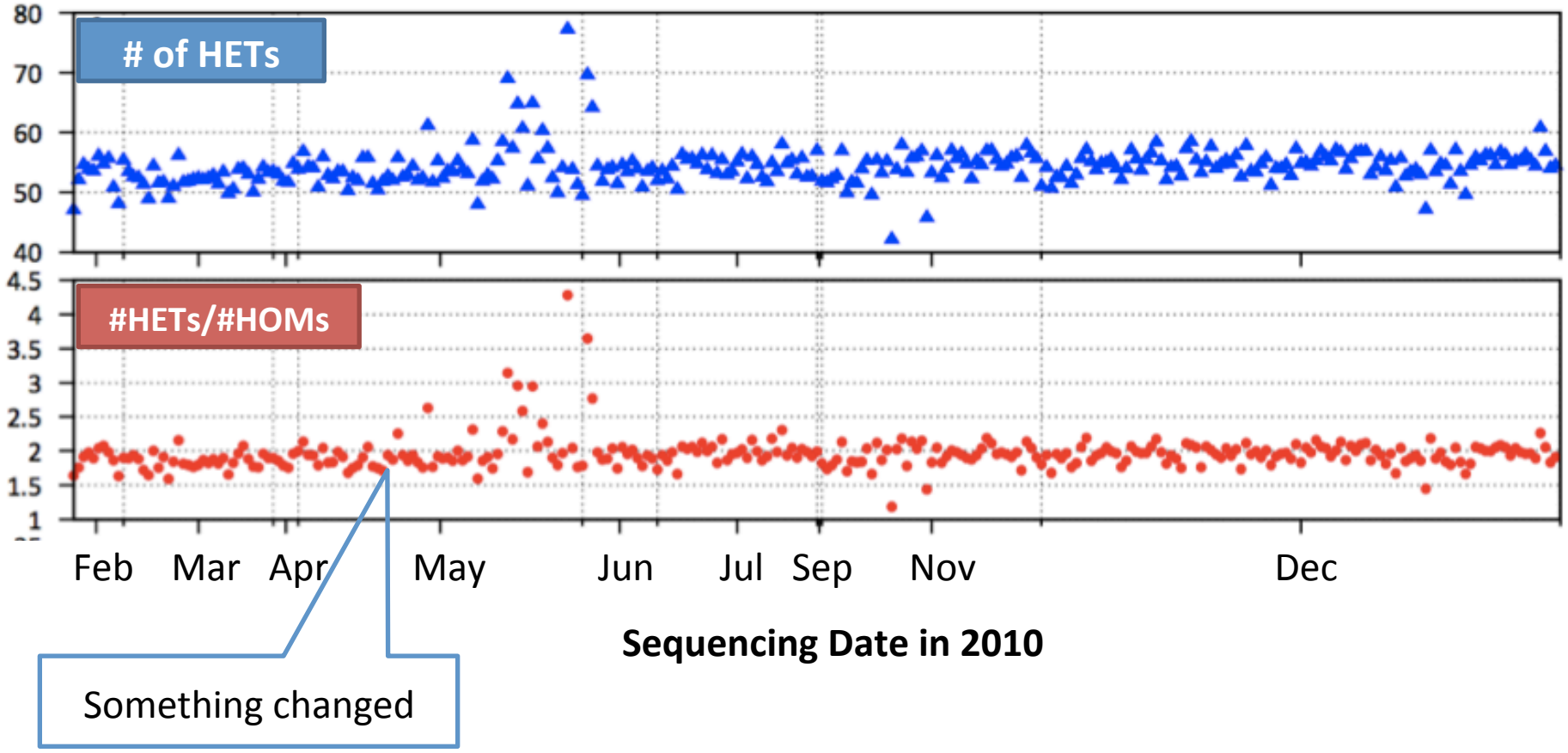
- Sample contamination is a *common* problem.
- Sample contamination degrades genotype accuracy and power of analysis.
- Timely feedback about contamination is important.

Contamination in Sequence Data



Sequencing Date in 2010

Contamination in Sequence Data



Reference-Aligned Sequence Reads

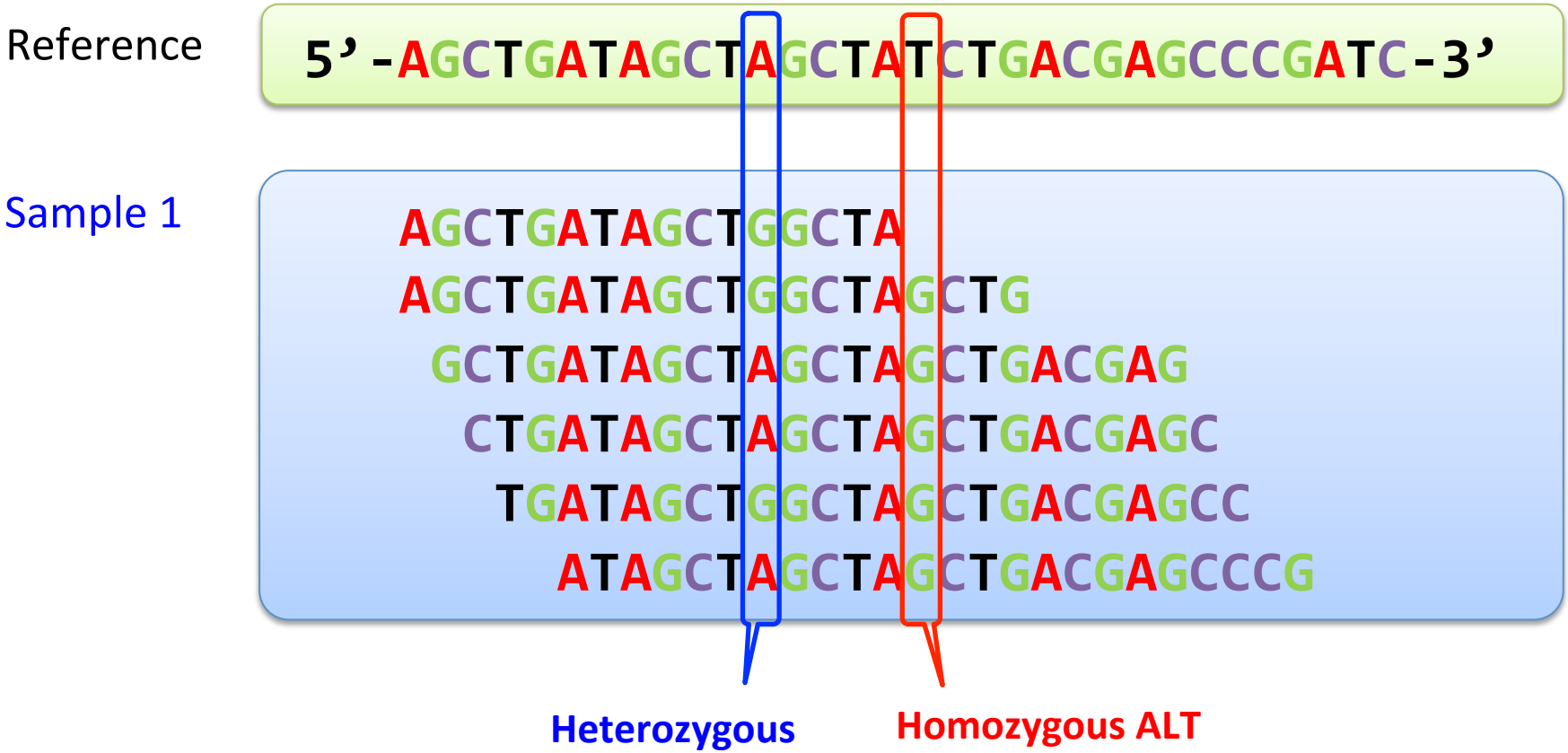
Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1

AGCTGATAGCTGGCTA
AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTAGCTAGCTGACGAGC
TGATAGCTGGCTAGCTGACGAGCC
ATAGCTAGCTAGCTGACGAGCCCG

Identifying SNPs



Base Distribution in Two Samples

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

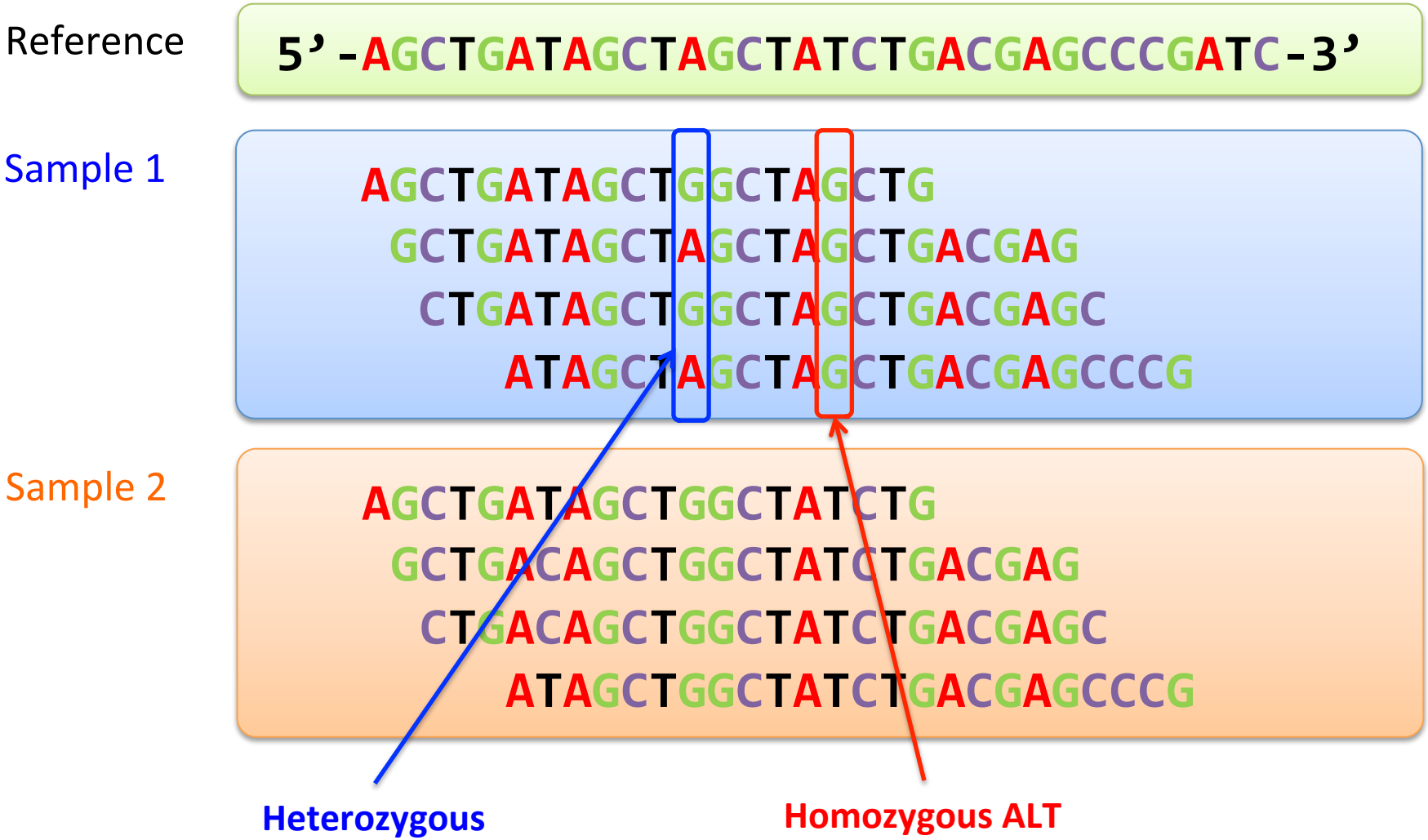
Sample 1

AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG

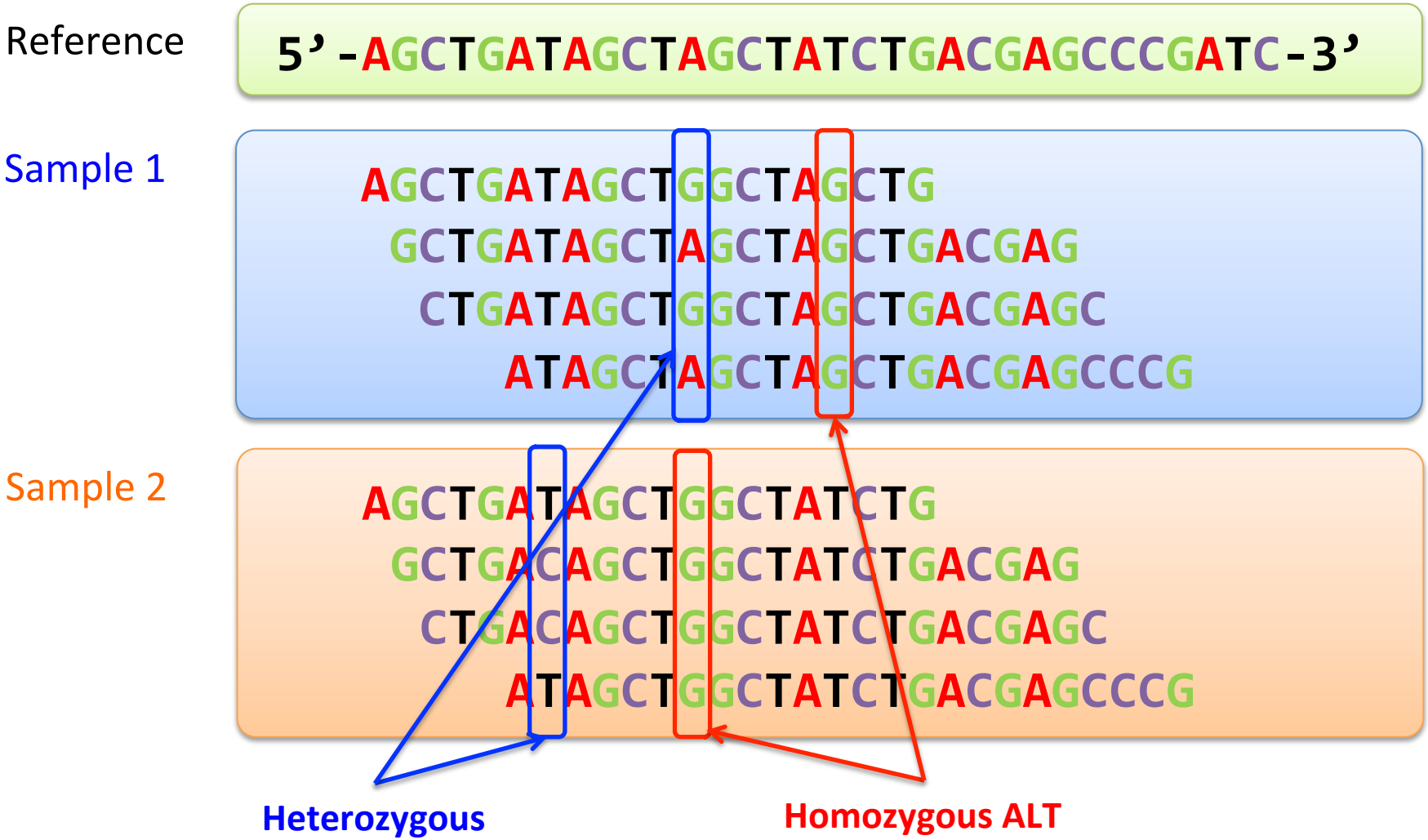
Sample 2

AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG

Base Distribution in Two Samples



Base Distribution in Two Samples



Contamination: Mixture of Samples

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2

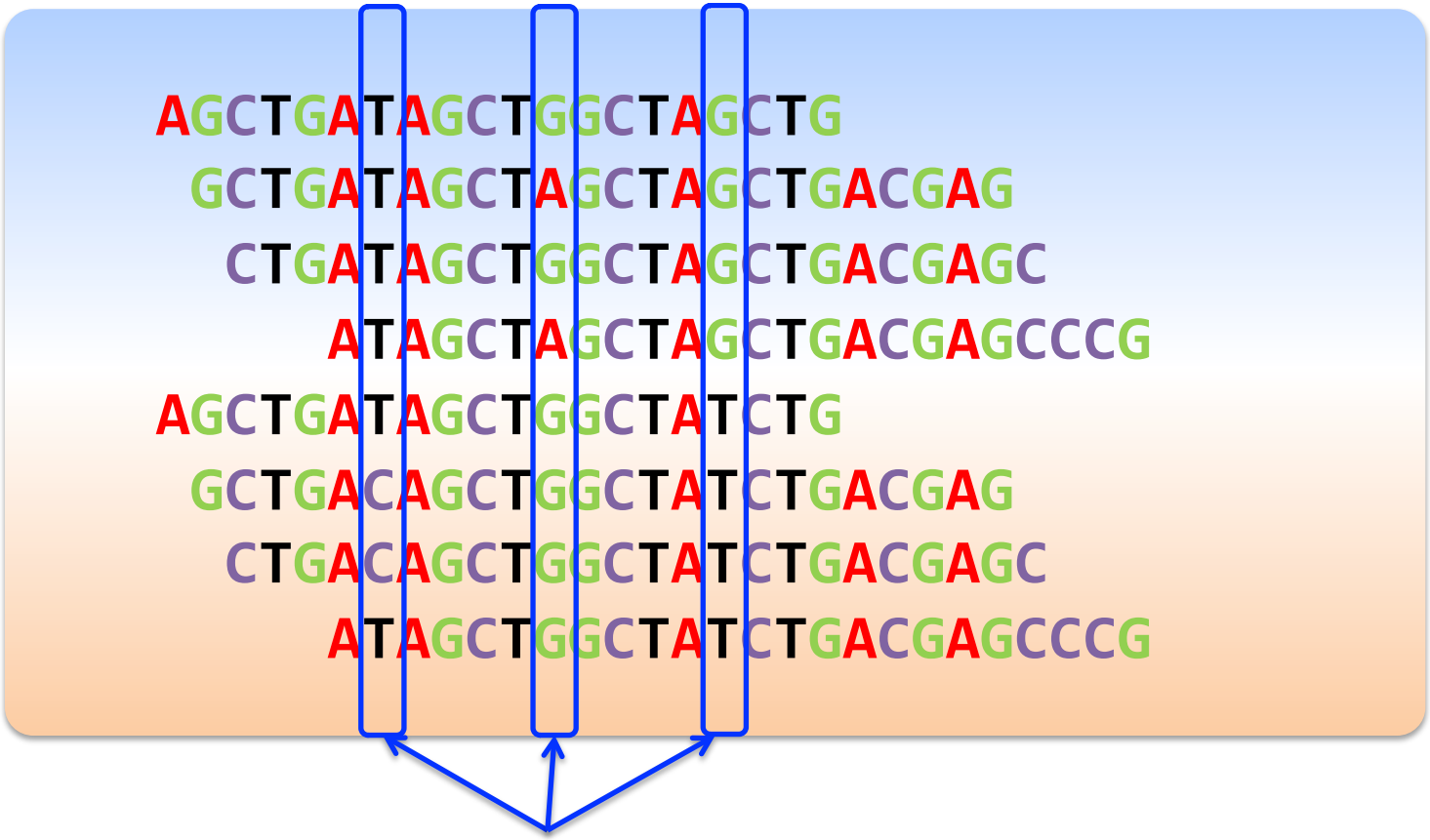
AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG
AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG

Contamination: Excessive Heterozygous SNPs

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2



Sample mixture leads to more heterozygosity

Quantifying Sample Contamination

- (# HETs) / (# HOMs)
  - No clear way to quantify the amount of contamination
  - Varies by population (admixed samples)
  - Requires variant calling
- Better solution
  - Estimate amount of contamination directly from base reads.

Likelihood of Sequence Data

- For M SNPs with known genotypes

$$L = \prod_{i}^{M} P(\mathbf{b}_i | G_i)$$

$$= \prod_{i}^{M} \prod_{j=1}^{N_i} P(b_{ij} | G_i)$$

  - Bases at i-th SNP site with N_i reads: $\mathbf{b}_i = (b_{i1}, b_{i2}, ..., b_{iN_i})$
  - Genotype of i-th marker: $G_i \in \{\mathrm{AA}, \mathrm{AB}, \mathrm{BB}\}$

Two-sample Mixture Model

- Likelihood with mixing proportion α

$$L(\alpha) = \prod_{i}^{M} \prod_{j=1}^{N_i} P(b_{ij}|G_i;\alpha)$$

$$= \prod_{i}^{M} \sum_{g_i \in \{AA,AB,BB\}} \prod_{j=1}^{N_i} P(b_{ij}|G_i,g_i;\alpha)P(g_i)$$

$$= \prod_{i}^{M} \sum_{g_i} \prod_{j=1}^{N_i} \{(1-\alpha)P(b_{ij}|G_i) + \alpha P(b_{ij}|g_i)\} P(g_i)$$

  - Estimated contamination: MLE of α

Two-sample Mixture Model

- Likelihood with mixing proportion α

$$L(\alpha) = \prod_{i}^{M} \prod_{j=1}^{N_i} P(b_{ij}|G_i;\alpha)$$

$$= \prod_{i}^{M} \sum_{g_i \in \{AA,AB,BB\}} \prod_{j=1}^{N_i} P(b_{ij}|G_i, g_i;\alpha) P(g_i)$$

$$= \prod_{i}^{M} \sum_{g_i} \prod_{j=1}^{N_i} \{(1-\alpha)P(b_{ij}|G_i) + \alpha P(b_{ij}|g_i)\} P(g_i)$$

From population allele frequency under HWE

Two-sample Mixture Model

- Likelihood with mixing proportion α

$$L(\alpha) = \prod_{i}^{M} \prod_{j=1}^{N_i} P(b_{ij}|G_i;\alpha)$$

$$= \prod_{i}^{M} \sum_{g_i \in \{AA,AB,BB\}} \prod_{j=1}^{N_i} P(b_{ij}|G_i, g_i;\alpha) P(g_i)$$

$$= \prod_{i}^{M} \sum_{g_i} \prod_{j=1}^{N_i} \{(1-\alpha) P(b_{ij}|G_i) + \alpha P(b_{ij}|g_i)\} P(g_i)$$

Probability of each base read given genotype

Probability of Each Base Read

- Base read observation
  - Reference (A), Alternative (B), or Other (O)
- Probability of observing a base depends on
  - Underlying (true) genotype
  - Occurrence of base read error
- Example
  - P(read = A | genotype = AA, **no error**) = 1
  - P(read = O | genotype = BB, ***error***) = 2/3

(In case of base read error, assume all possibilities are equally likely)

Likelihood Table

- b_{ij} : j-th read at i-th marker ($j = 1 \,..\, N_i$)
- G_i : true genotype of i-th marker
- e_{ij} : occurrence of error (1: error, 0: no error)

G_i	e_{ij}	$P(b_{ij} \mid G_i, e_{ij})$		
		b_{ij} =A	b_{ij} =B	b_{ij} =O
AA	0	1	0	0
	1	0	1/3	2/3
AB	0	1/2	1/2	0
	1	1/6	1/6	2/3
BB	0	0	1	0
	1	1/3	0	2/3

Computing Full Likelihood

- Likelihood over M markers :

$$L(\alpha) = \prod_{i}^{M} \sum_{g_i} \prod_{j=1}^{N_i} P(b_{ij}|G_i, g_i; \alpha) P(g_i)$$

$$= \prod_{i}^{M} \sum_{g_i} \prod_{j=1}^{N_i} \sum_{e_{ij} \in \{0,1\}} P(b_{ij}|G_i, g_i, e_{ij}; \alpha) P(e_{ij}) P(g_i)$$

- $P(e_{ij})$: probability of base read error, from Phred score Q_{ij}

$$P(e_{ij} = 1) = 10^{-\frac{Q_{ij}}{10}}$$

Estimation with Sequence Data Only

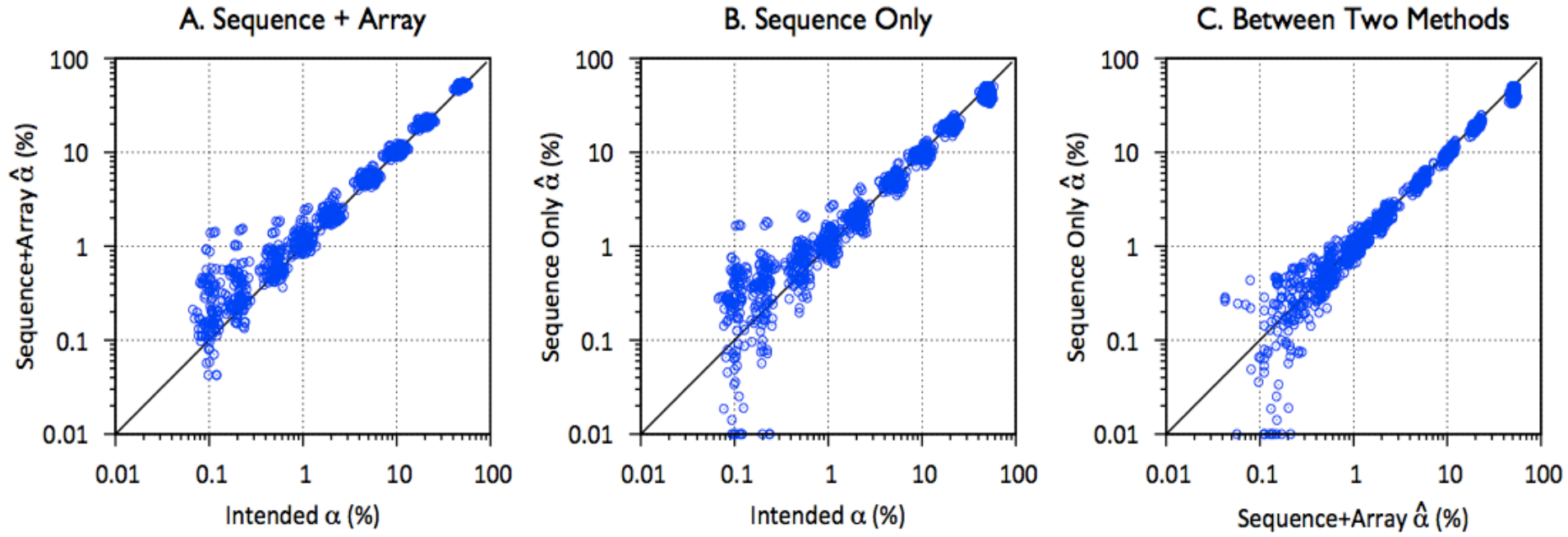
- If sequenced sample does not have external genotypes
  - Model both genotypes from population allele frequency

- Latent variables
  - G_i : true genotype of the contaminated sample
  - g_i : true genotype of the contaminating sample

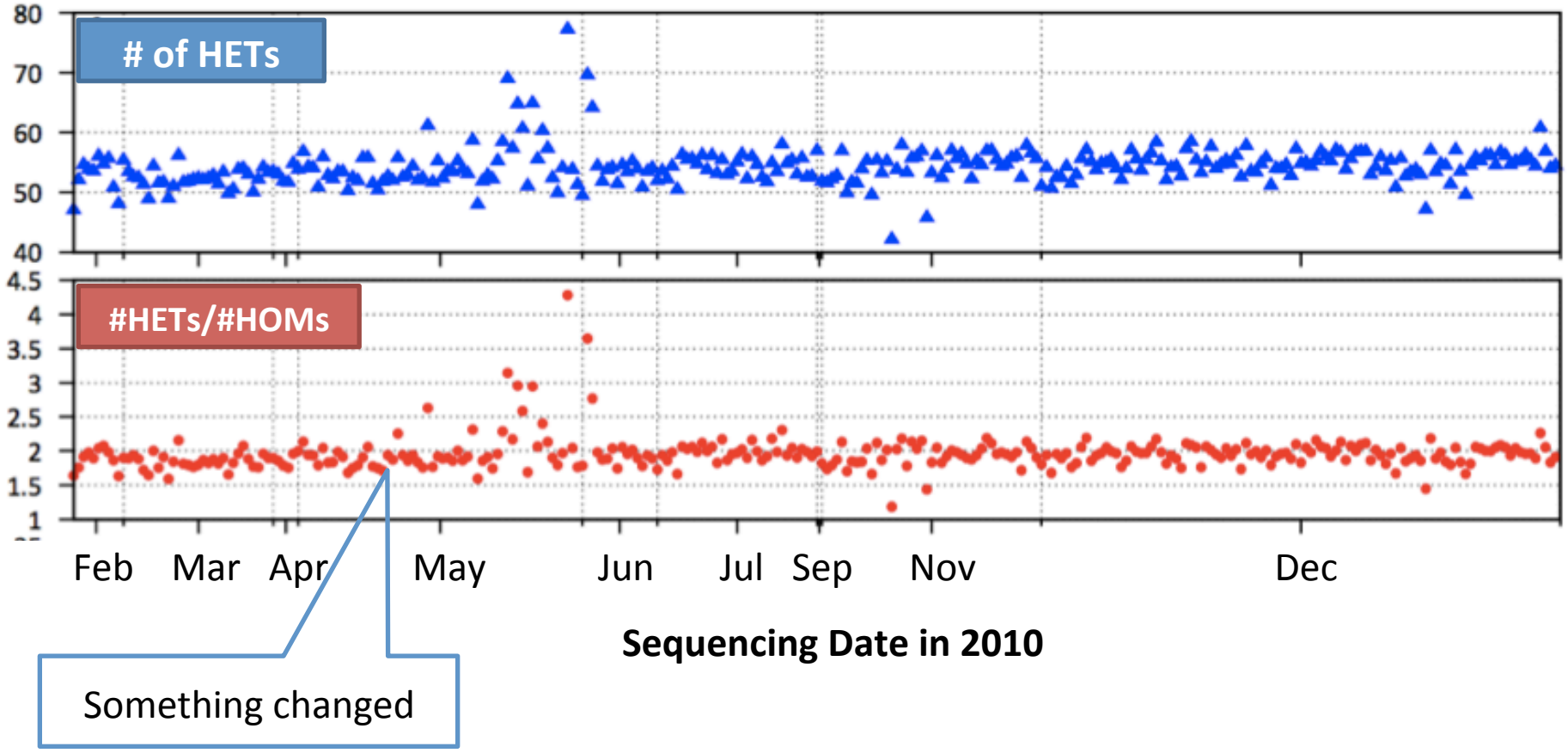
$$L(\alpha) = \prod_{i}^{M} \sum_{G_i} \sum_{g_i} \prod_{j}^{N_i} P(b_{ij}|G_i, g_i; \alpha) P(G_i) P(g_i)$$

Results: Simulation

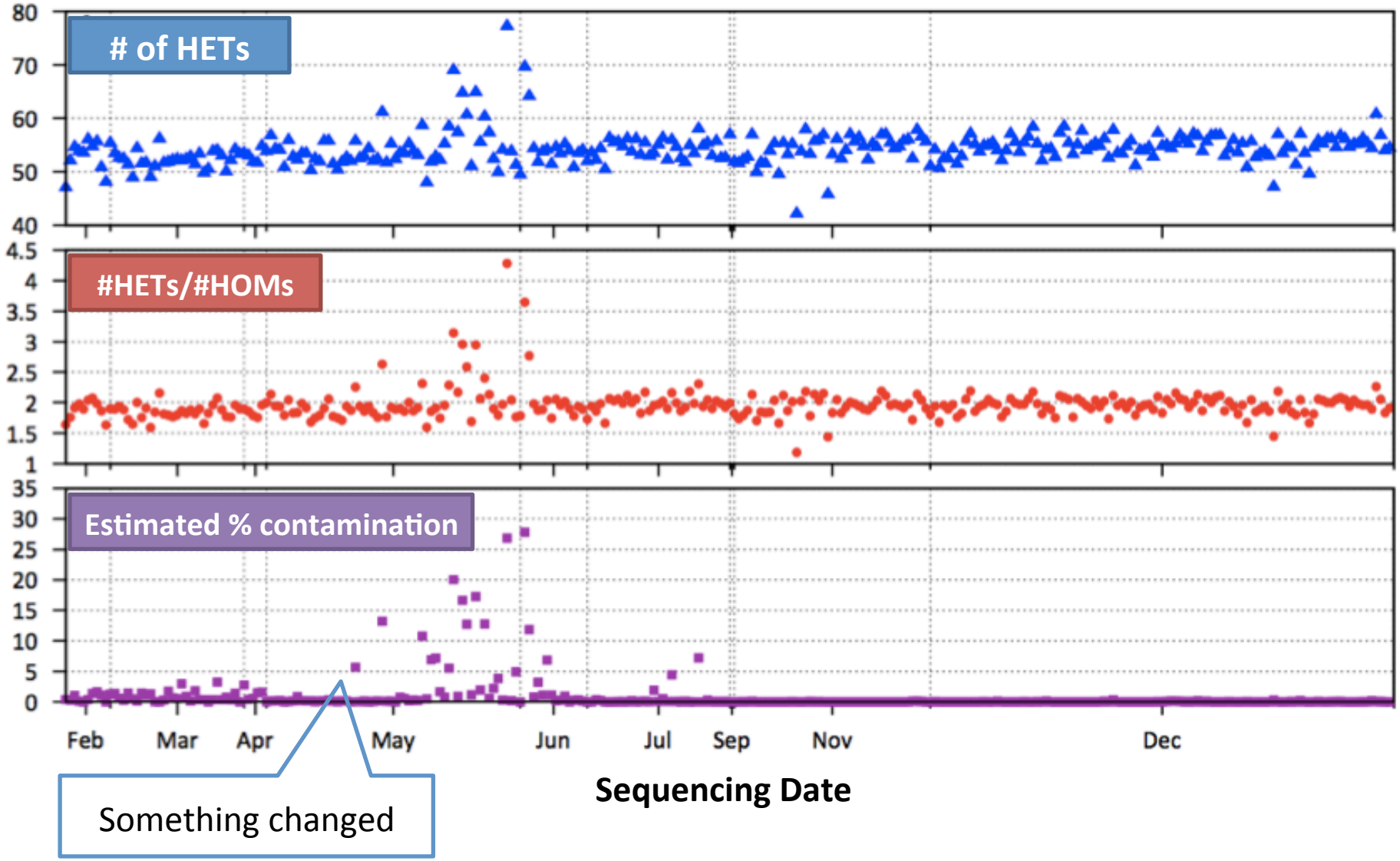
- Simulated contamination
  - Randomly mixed reads from two samples in 1000 Genomes
  - Simulated from 0.1% to 50% contamination



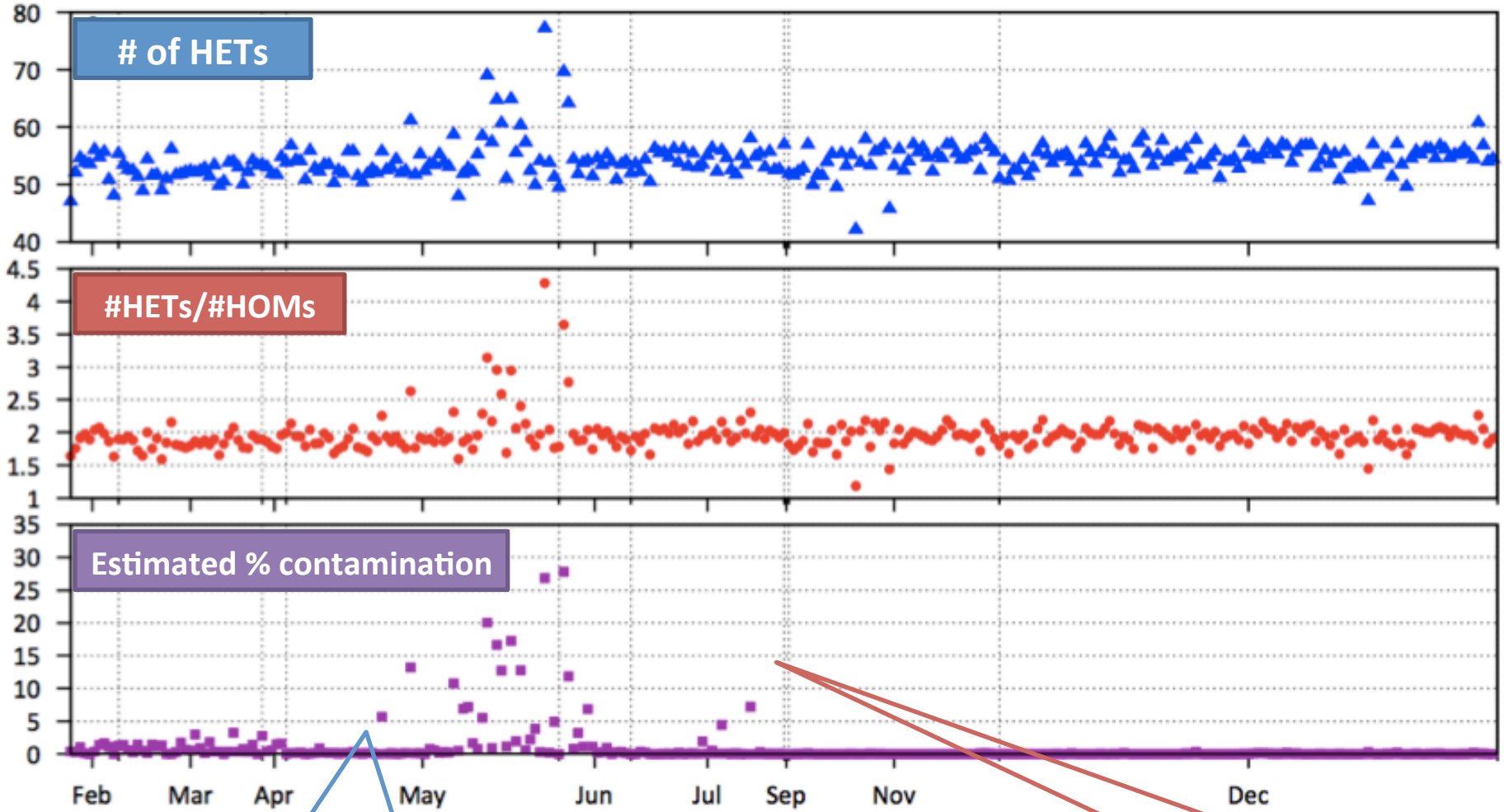
Contamination in Sequence Data



Results: Real Data



Results: Real Data



gotcloud Sample QC Steps after Alignment

- Remove duplicate reads
- Base quality recalibration
  - Adjust reported base quality with empirically measured quality
- BAM quality report (QPLOT)
- Check sample contamination (verifyBamID)

```
cd ~goo/gotcloudTutorial/QCFiles
ls *.selfSM
tail *.selfSM
```

gotcloud Example

```
> tail ~goo/gotcloudTutorial/QCFiles/*.selfSM
#SEQ_ID         RG      CHIP_ID         #SNPS   #READS  AVG_DP  FREEMIX
HG00096         ALL     NA              36298   392     0.01    0.00000

FREELK1         FREELK0         FREE_RH         FREE_RA
86.22           86.22           NA              NA

CHIPMIX         CHIPLK1         CHIPLK0         CHIP_RH         CHIP_RA
NA              NA              NA              NA              NA

DPREF   RDPHET  RDPALT
NA      NA      NA
```

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

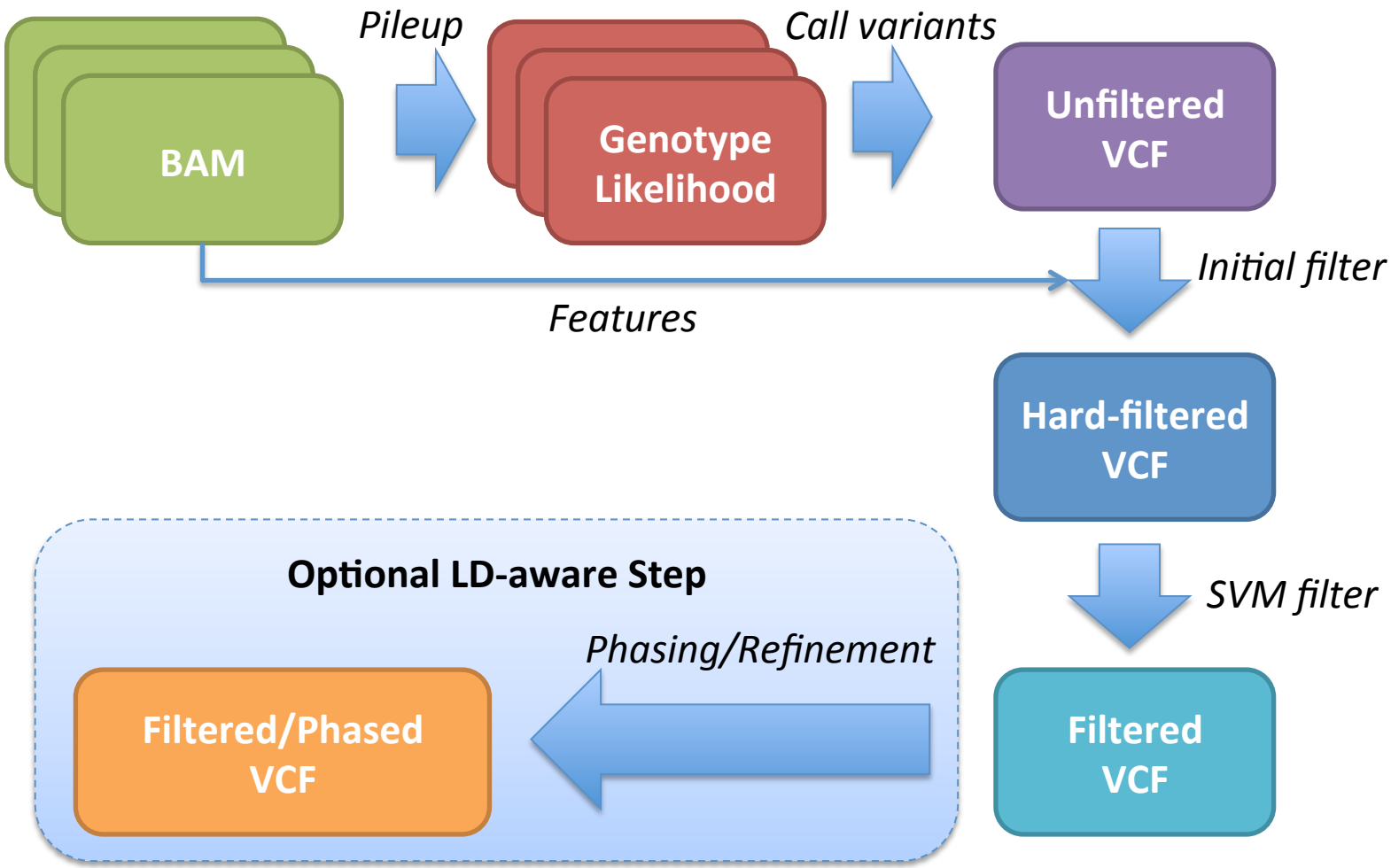
Summary: Part 1

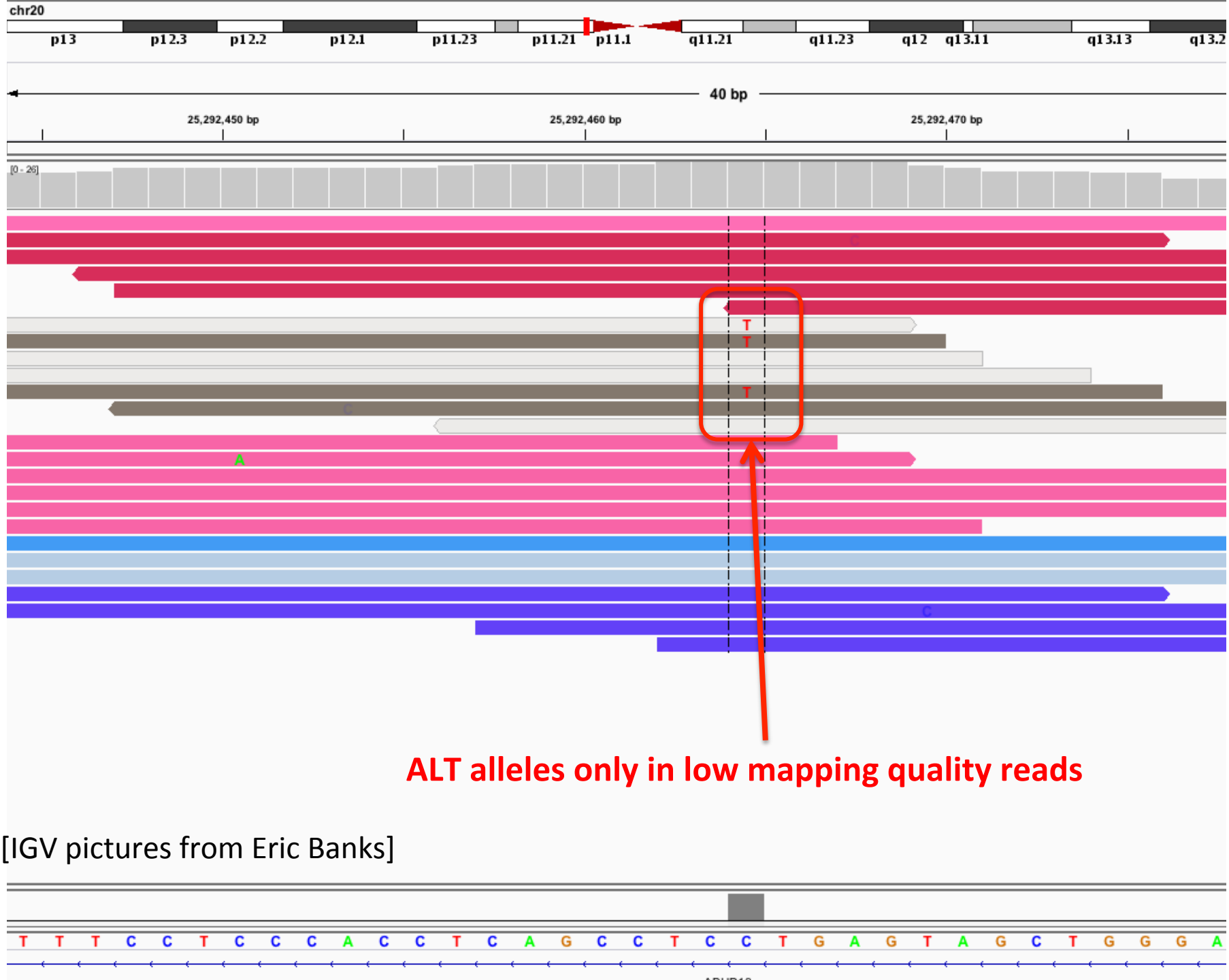
- Likelihood-based model accurately estimates of % of potential sample contamination.
- The sample likelihood model can be used to correct genotype likelihoods, which greatly improves genotype accuracies. (skipped)
- Stand-alone tools to check contamination:
  - http://genome.sph.umich.edu/wiki/VerifyBamID
  - http://genome.sph.umich.edu/wiki/VerifyIDintensity

Part 2.

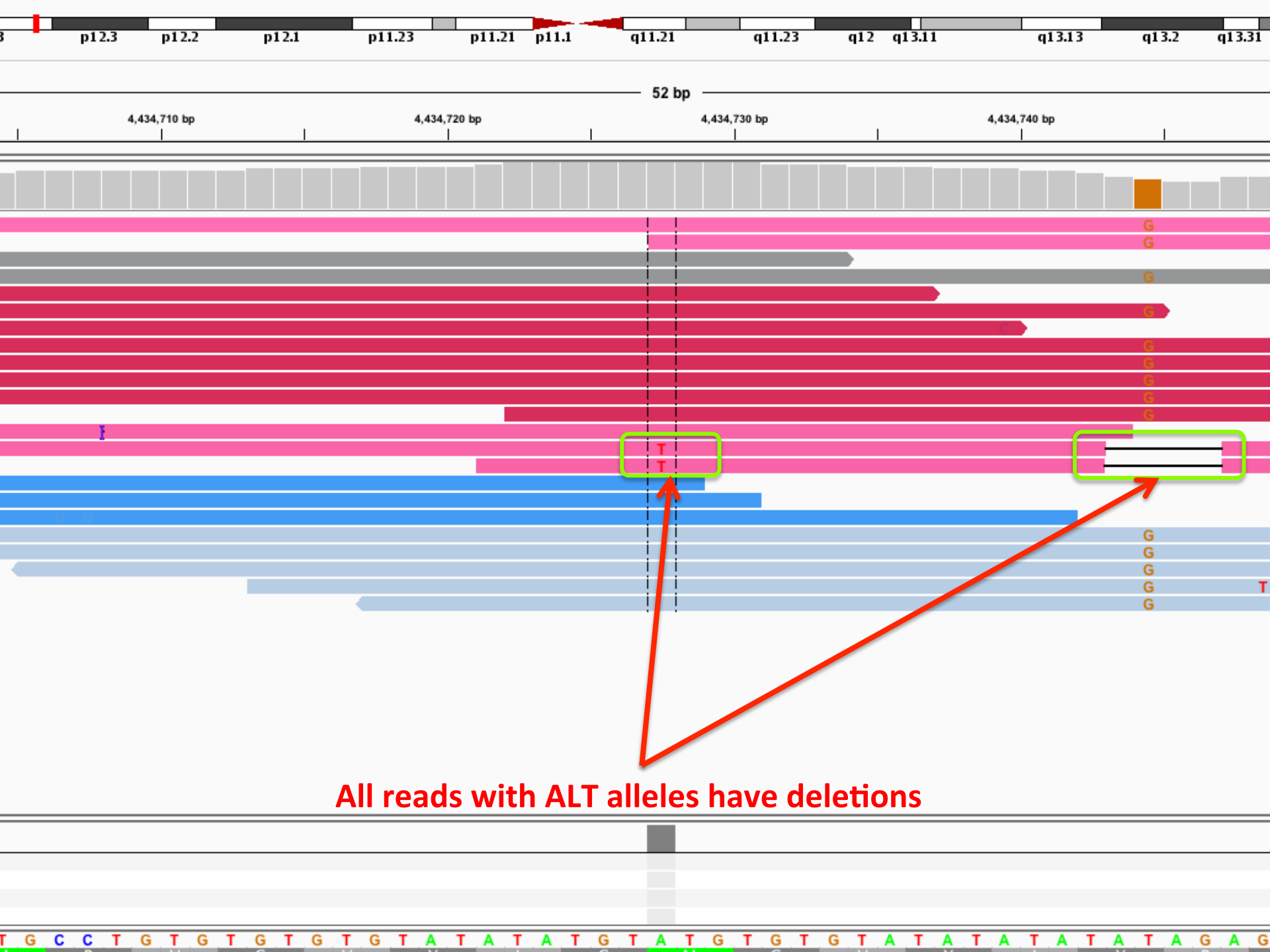
Variant Calling and Filtering

Michigan SNP Calling Pipeline

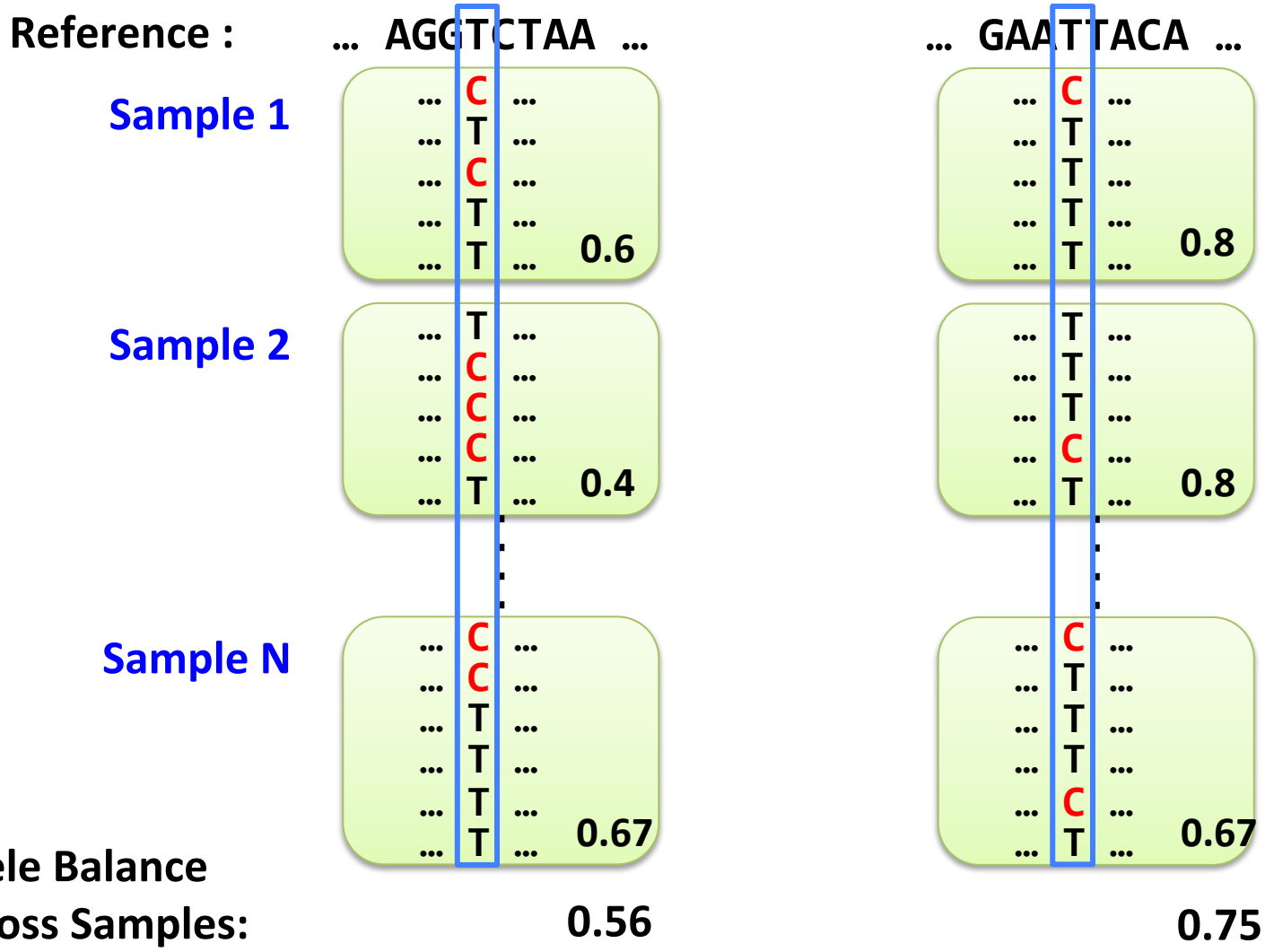




[IGV pictures from Eric Banks]


p12.3
p12.2
p12.1
p11.23
p11.21
p11.1
q11.21
q11.23
q12
q13.11
q13.13
q13.2
q13.31
52 bp
4,434,710 bp
4,434,720 bp
4,434,730 bp
4,434,740 bp
All reads with ALT alleles have deletions

Multi-sample Filtering is Effective



Filtering Criteria Examples

Feature	Description
Depth	Overall depth across samples
QUAL	Overall genotype confidence
Call Rate	Proportion of genotyped samples
Allele Balance	(# REF)/(# ALT) in HET sites
Strand Bias	Correlation of ALT allele with +/- strand
Cycle Bias	Correlation of ALT allele with read cycle
Etc.	And many more…

gotcloud Example: SNP Filtering Results

- Let's check out SNP calling results from gotcloud

```
cd ~goo/gotcloudTutorial
cd vcfs/chr20
```

- Filtering summary file

```
less chr20.filtered.sites.vcf.summary
```

- Let's see what sites are filtered out

```
grep AB70 chr20.filtered.sites.vcf|less -S
```

Dig into VCF

- Check which sample has ALT alleles in those SNPs

```
tabix -h chr20.filtered.vcf.gz 20:42925764-42925764
tabix -h chr20.filtered.vcf.gz 20:42925764-42925764|tail -2
tabix -h chr20.filtered.vcf.gz 20:42925764-42925764|tail -2|cut –f10
```

- (Don’t need to do this yourself)

```
for i in {10..69};do tabix -h chr20.filtered.vcf.gz 20:42925764-42925764 |tail -2|cut -f$i;done
```

```
HG00103 HG00111 HG00112 HG00114 HG00125 HG00127 HG00131 HG00137
HG00142 HG00239 HG00246
```

Check BAM Files at Problematic Sites

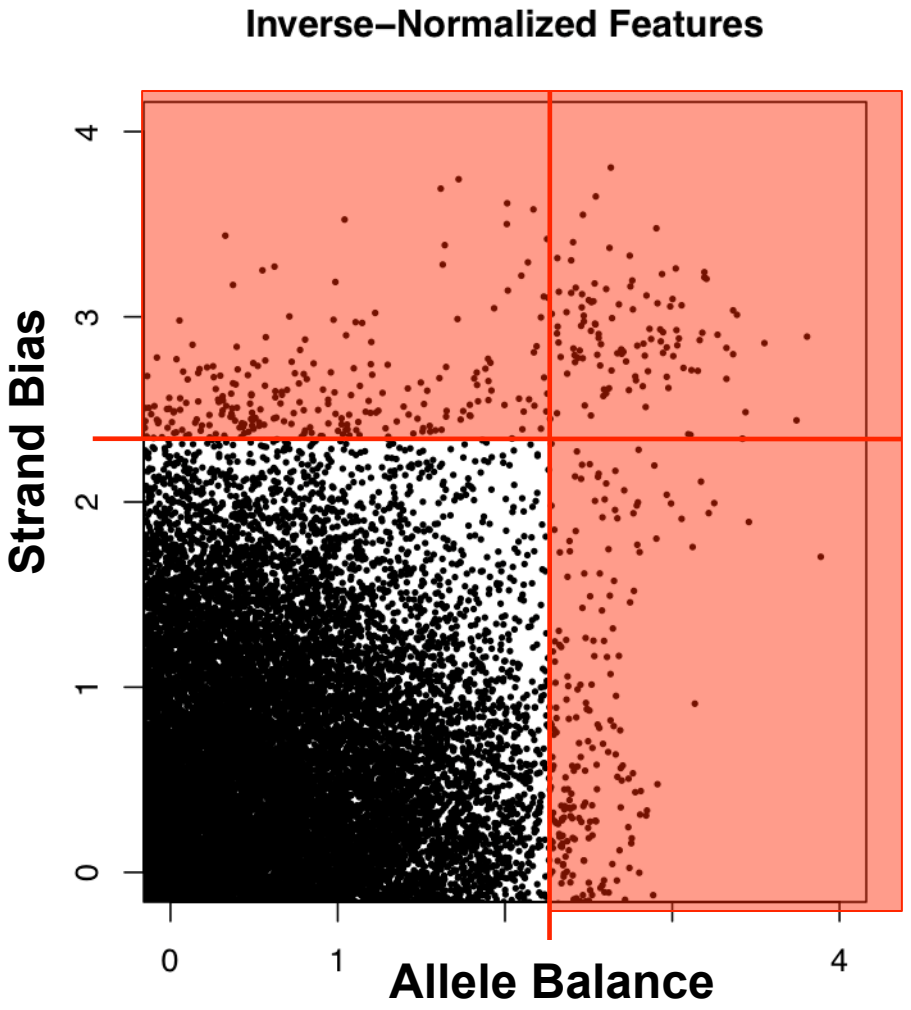
- Use samtools to see original BAM files

```
cd /opt/gotcloudExample/
cd bams

samtools tview HG00103.bam ../chr20Ref/humak_g1k_v37_chr20.ref
(type g) 20:42925764
samtools tview HG00111.bam ../chr20Ref/humak_g1k_v37_chr20.ref
(type g) 20:42925764
```

Hard Filtering by Individual Thresholds

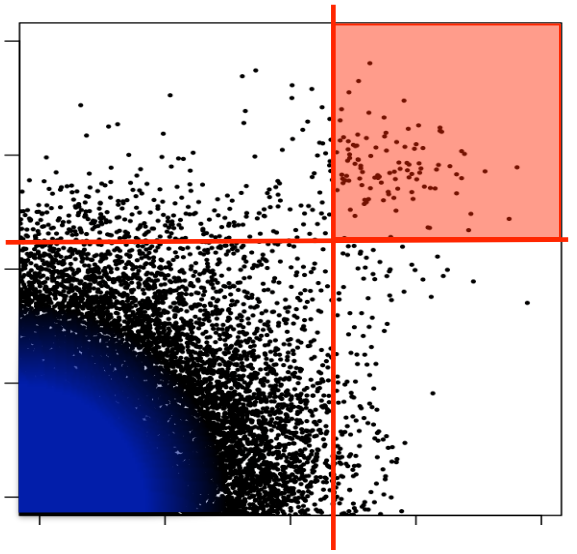
- Problems
  - False negative increases with number of filters
  - Too many knobs to turn (thresholds)



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

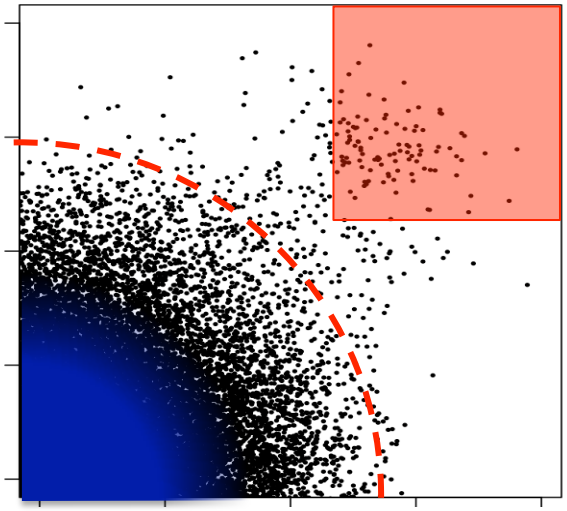
- Training
  - Positive examples
    - Known polymorphic sites
  - Negative examples
    - Filtered out by multiple hard filters
  - Input
    - All individual features collected for each site



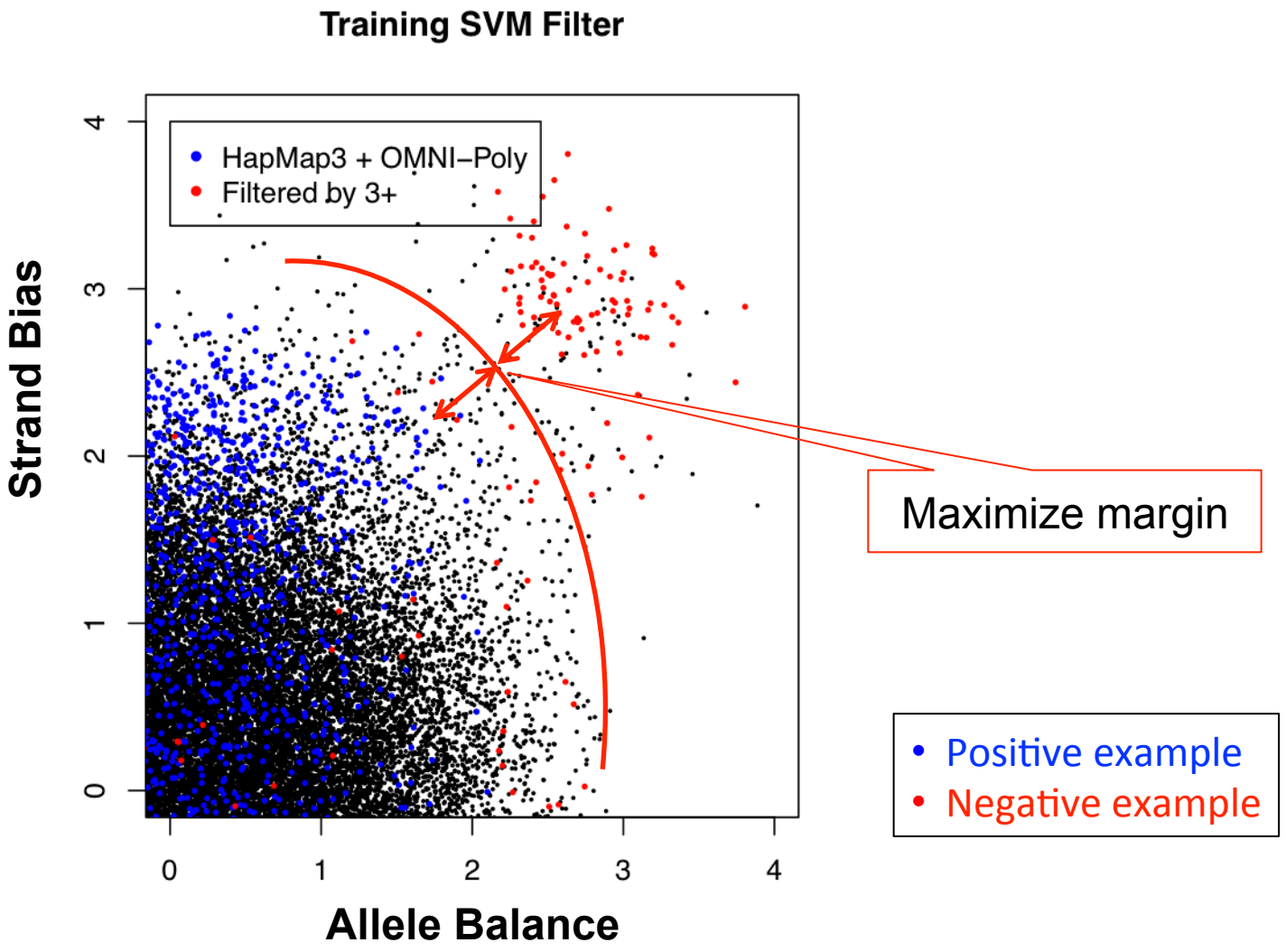
Filtering by Supervised Learning

- 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

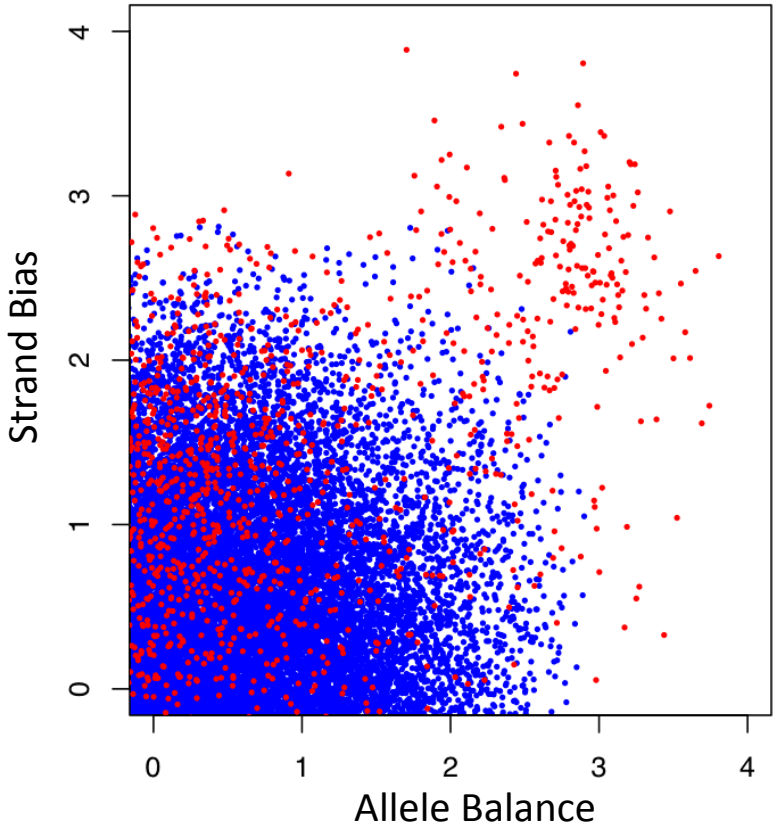


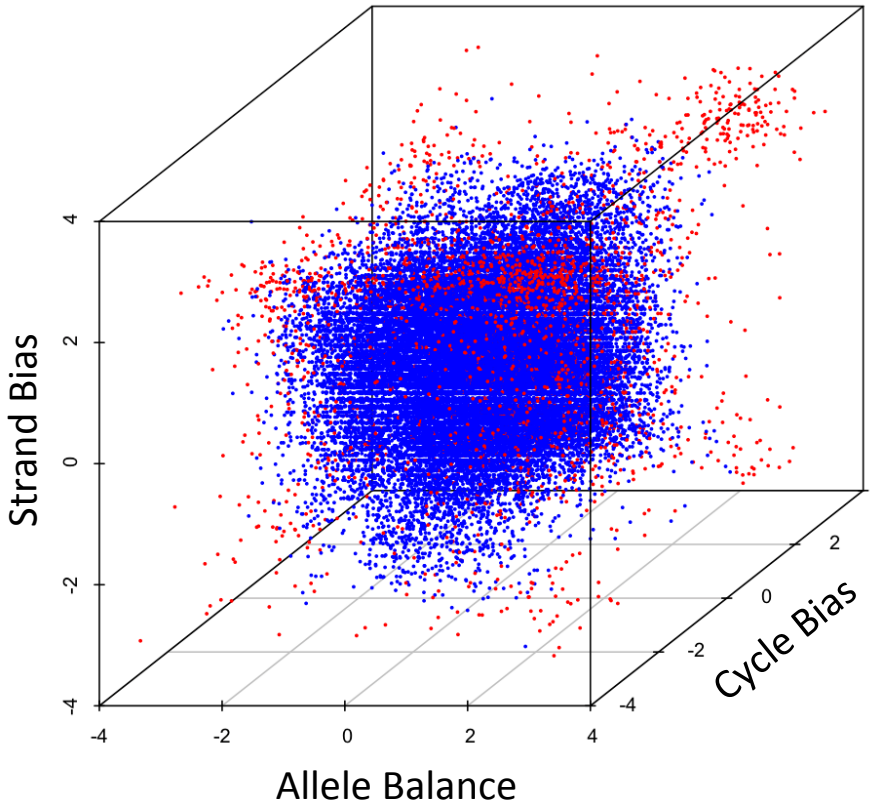
Training SVM with Examples



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

SVM Output in Multi-dimensional Space



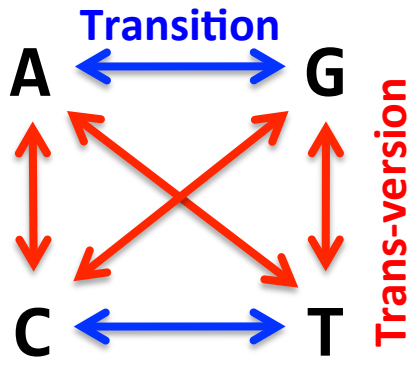


- Filter PASS
- Filter FAIL

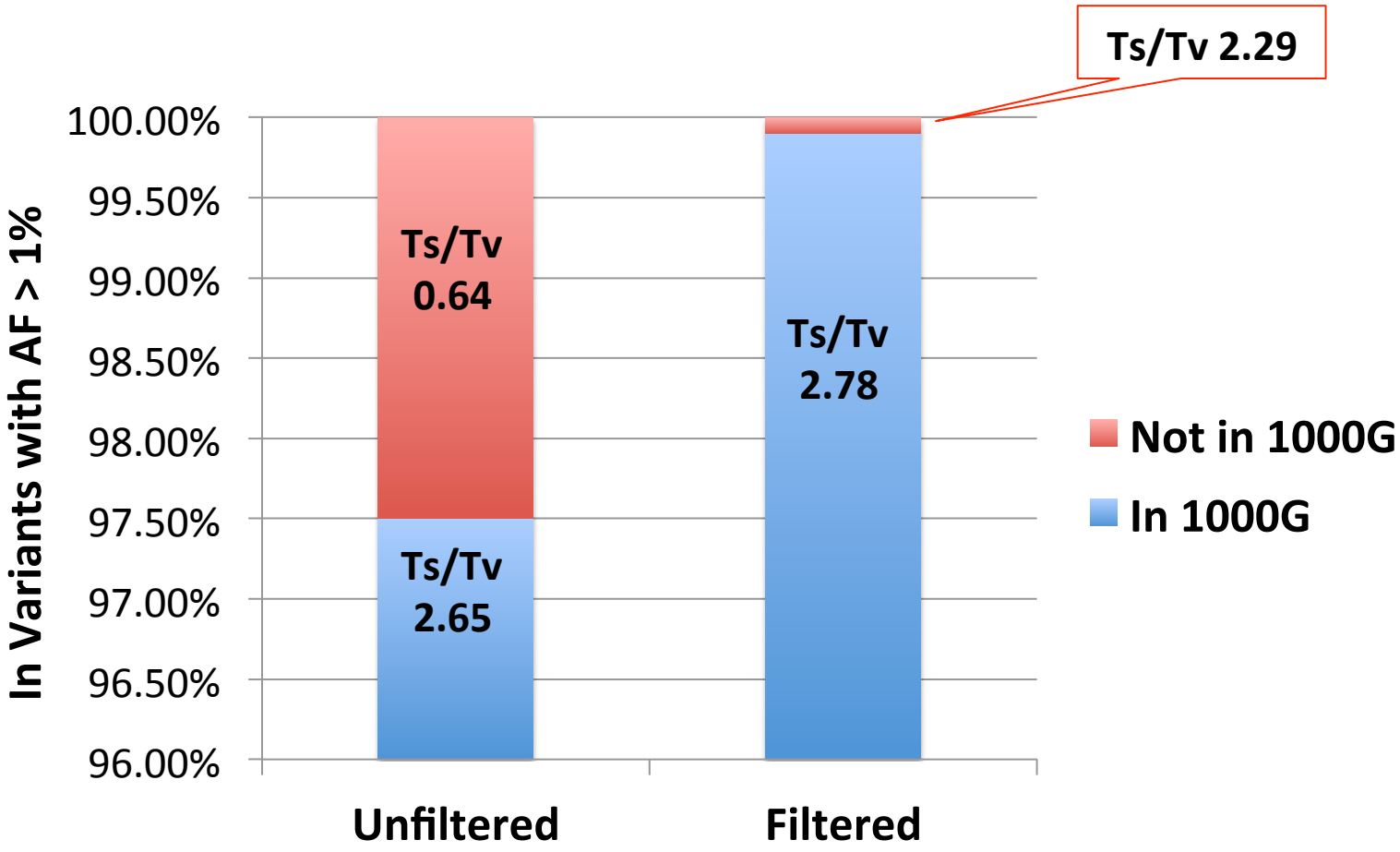
Most of FAIL SNPs are outliers in higher-dimensional view

Evaluation of SNP Callsets

- Sensitivity on known SNP data
  - dbSNP, HapMap, 1000G, etc.
- Transition to trans-version ratio
  - Transition is easier to occur.
  - Typical Ts/Tv values
    - Whole genome: 2.2~2.4
    - Whole exome: 2.7~3.0



Results: Exome Sequencing Project (GO-ESP)



Summary: Part 2

- Multi-sample strategy delivers high-quality variants
- SVM-based filtering provides effective and flexible methods for variant filtering
  - Applicable to GATK-generated variants and features
  - Trained SVM models are re-usable
  - Applicable to sequencing data from multiple platforms
- Michigan Mapping/Variant calling pipeline on the cloud
  - http://genome.sph.umich.edu/wiki/GotCloud

Backup

Genotype Calling with Contamination Parameter

- SNP calling with genotype likelihood model

$$\arg\max_{G_i} L = \arg\max_{G_i} \prod_{j=1}^{N_i} P(b_{ij}|G_i)$$

Genotype Calling with Contamination Parameter

- SNP calling with genotype likelihood model

$$\arg\max_{G_i} L = \arg\max_{G_i} \prod_{j=1}^{N_i} P(b_{ij}|G_i)$$

- Genotype likelihood model with contamination

$$\arg\max_{G_i} L(\alpha) = \arg\max_{G_i} \sum_{g_i} \prod_{j=1}^{N_i} \{(1-\alpha)P(b_{ij}|G_i) + \alpha P(b_{ij}|g_i)\} P(g_i)$$

Genotype Calling with Contamination Parameter

- SNP calling with genotype likelihood model

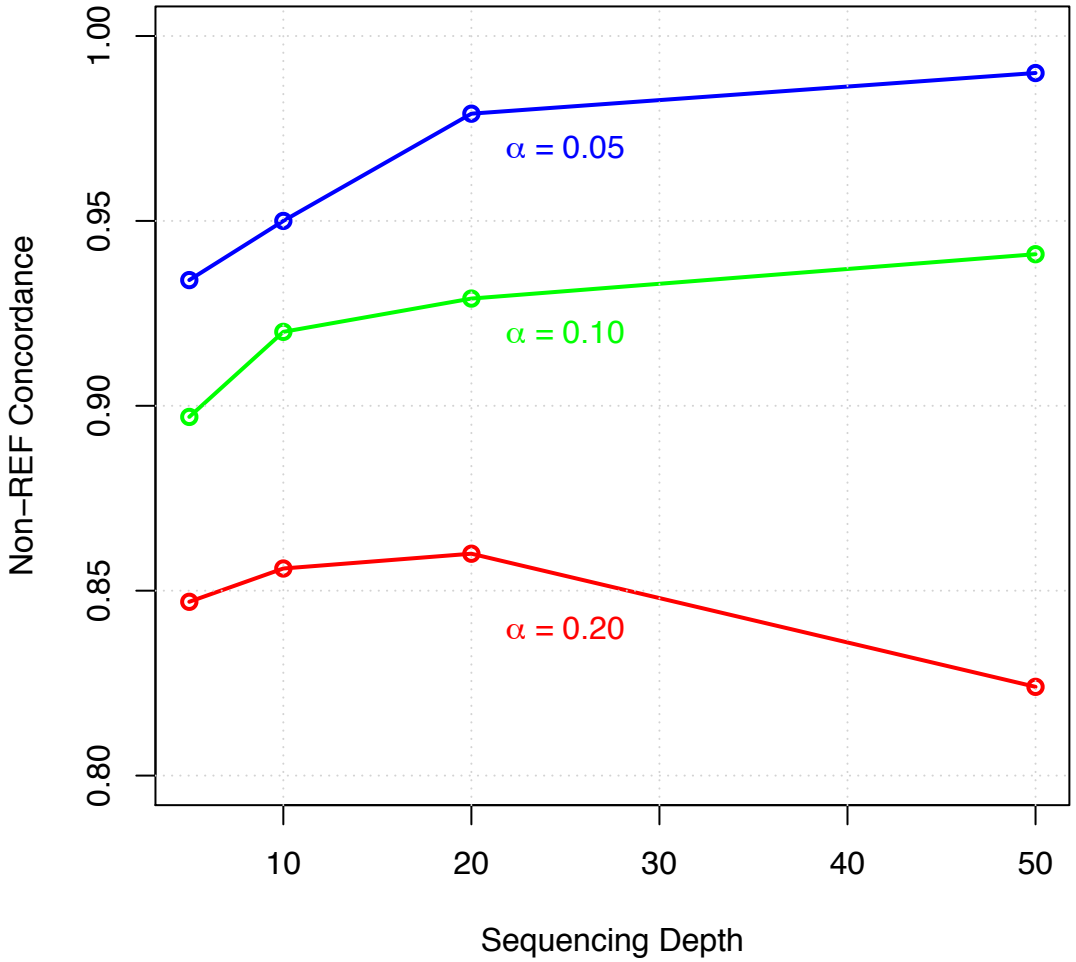
$$\arg\max_{G_i} L = \arg\max_{G_i} \prod_{j=1}^{N_i} P(b_{ij}|G_i)$$

- Genotype likelihood model with contamination

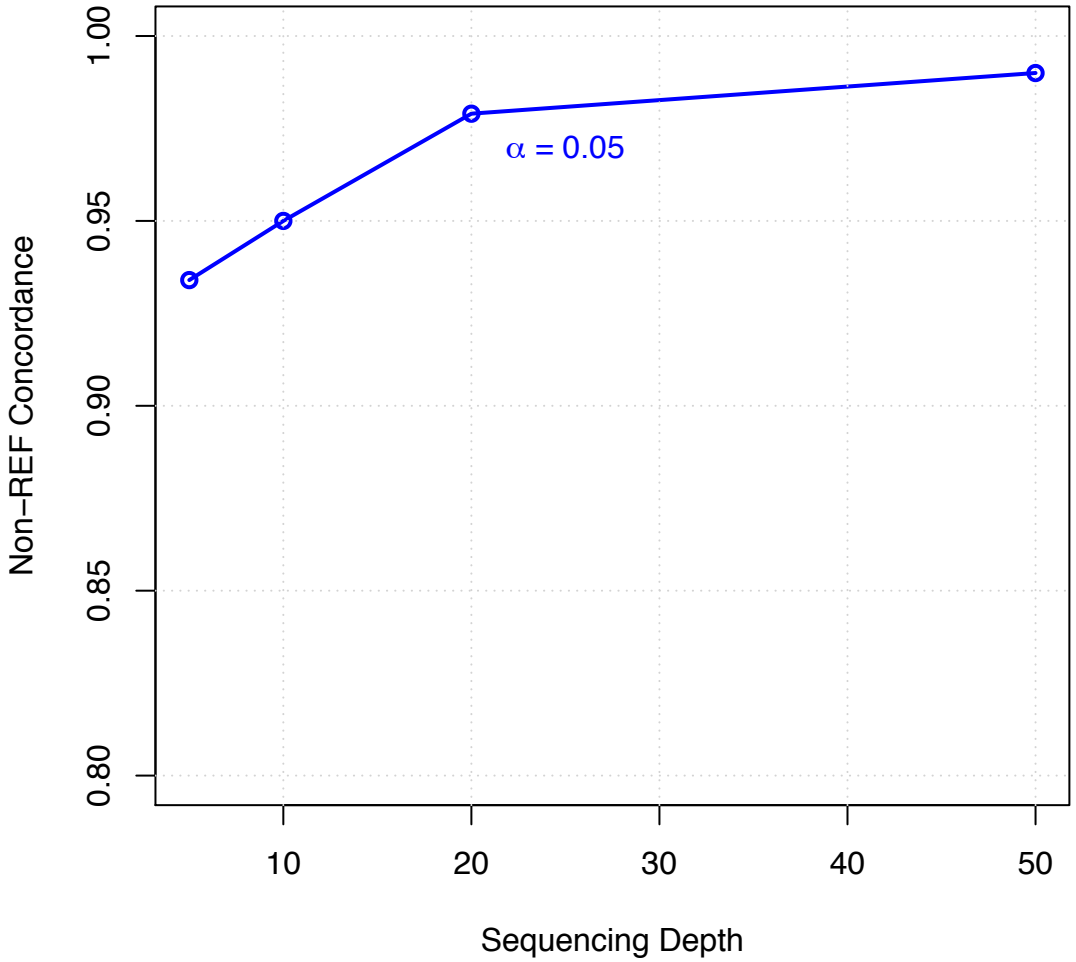
$$\arg\max_{G_i} L(\alpha) = \arg\max_{G_i} \sum_{g_i} \prod_{j=1}^{N_i} \{(1-\alpha)P(b_{ij}|G_i) + \alpha P(b_{ij}|g_i)\} P(g_i)$$

Estimated Contamination

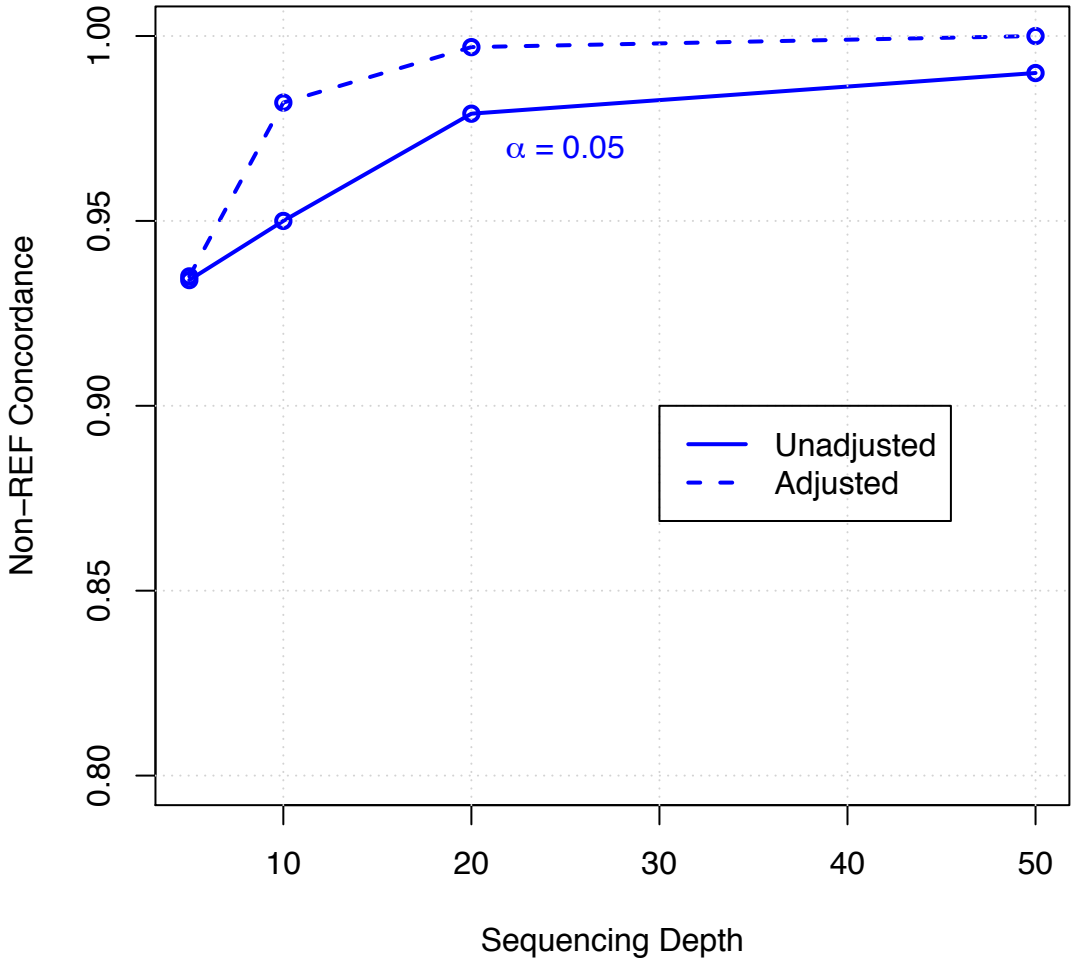
Simulation: Correcting Contamination



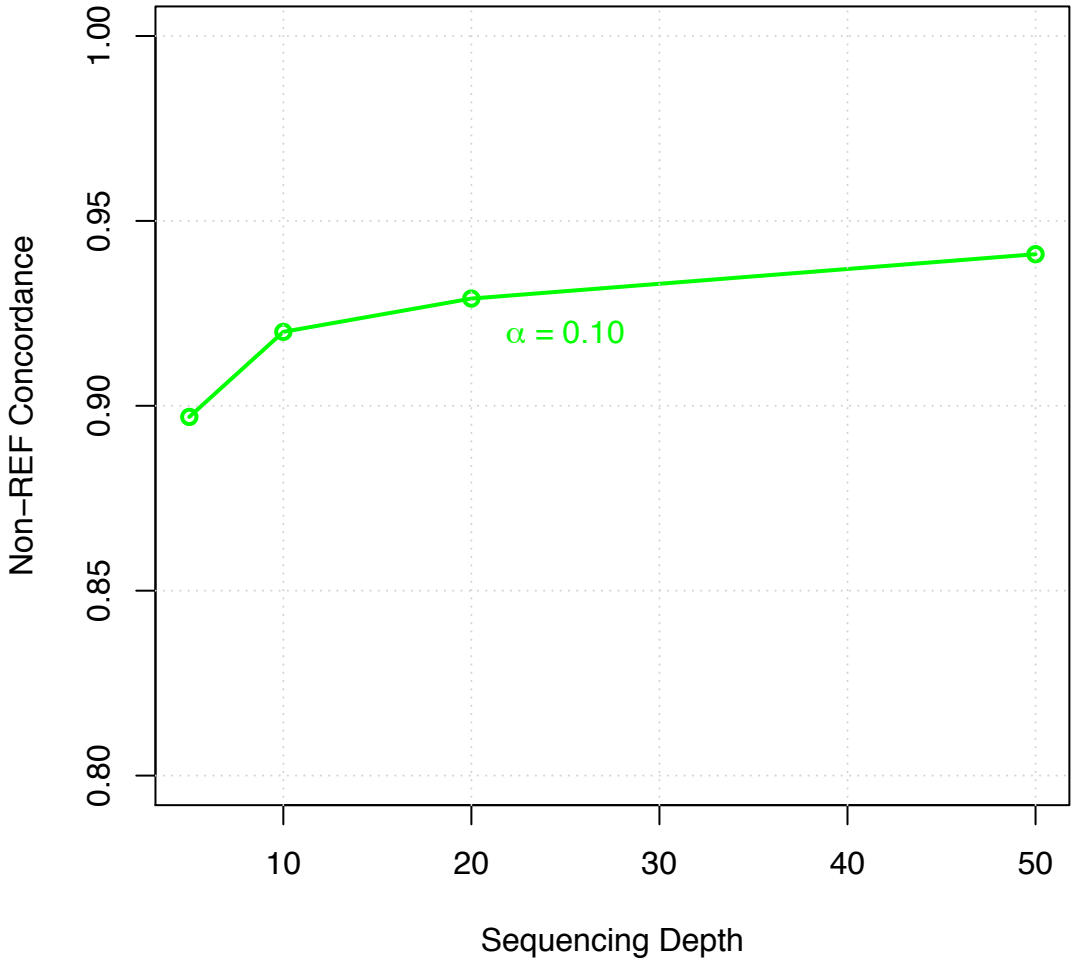
Simulation: Correcting Contamination



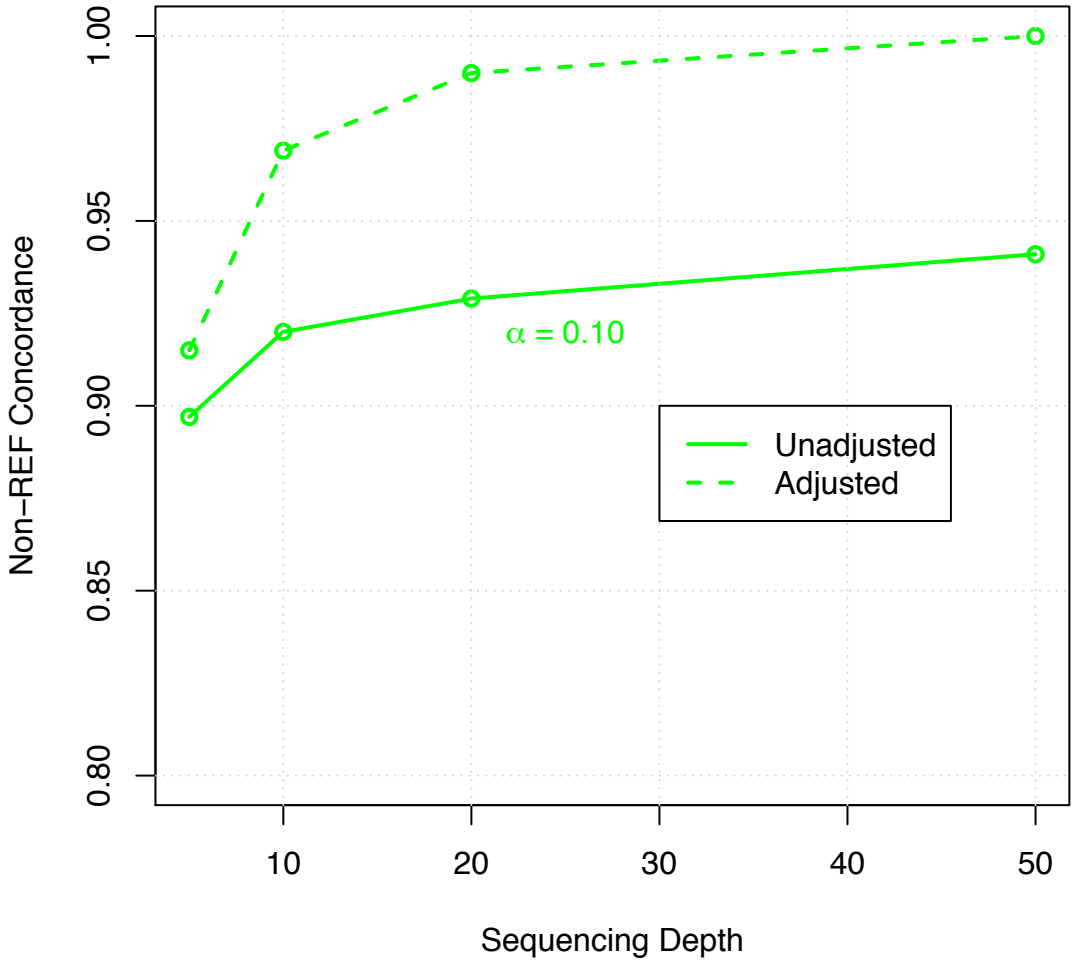
Simulation: Correcting Contamination



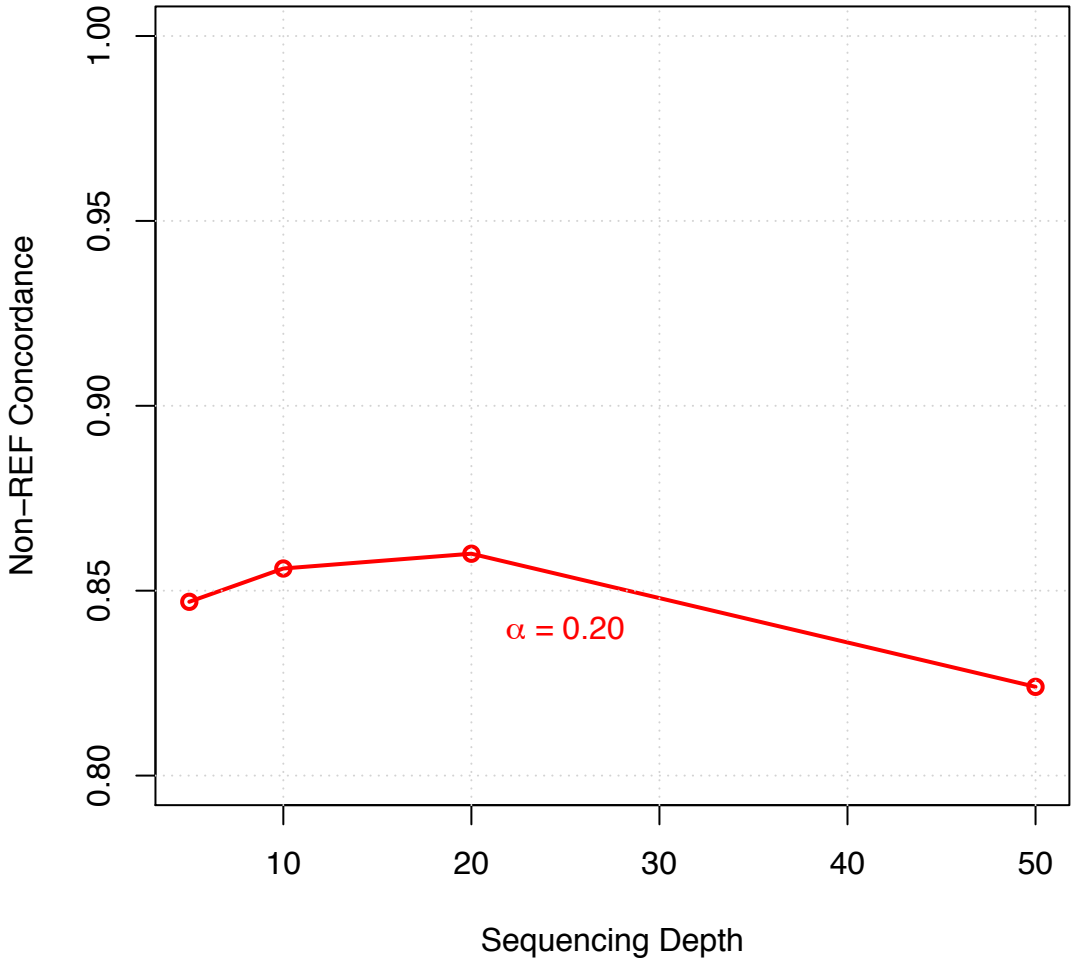
Simulation: Correcting Contamination



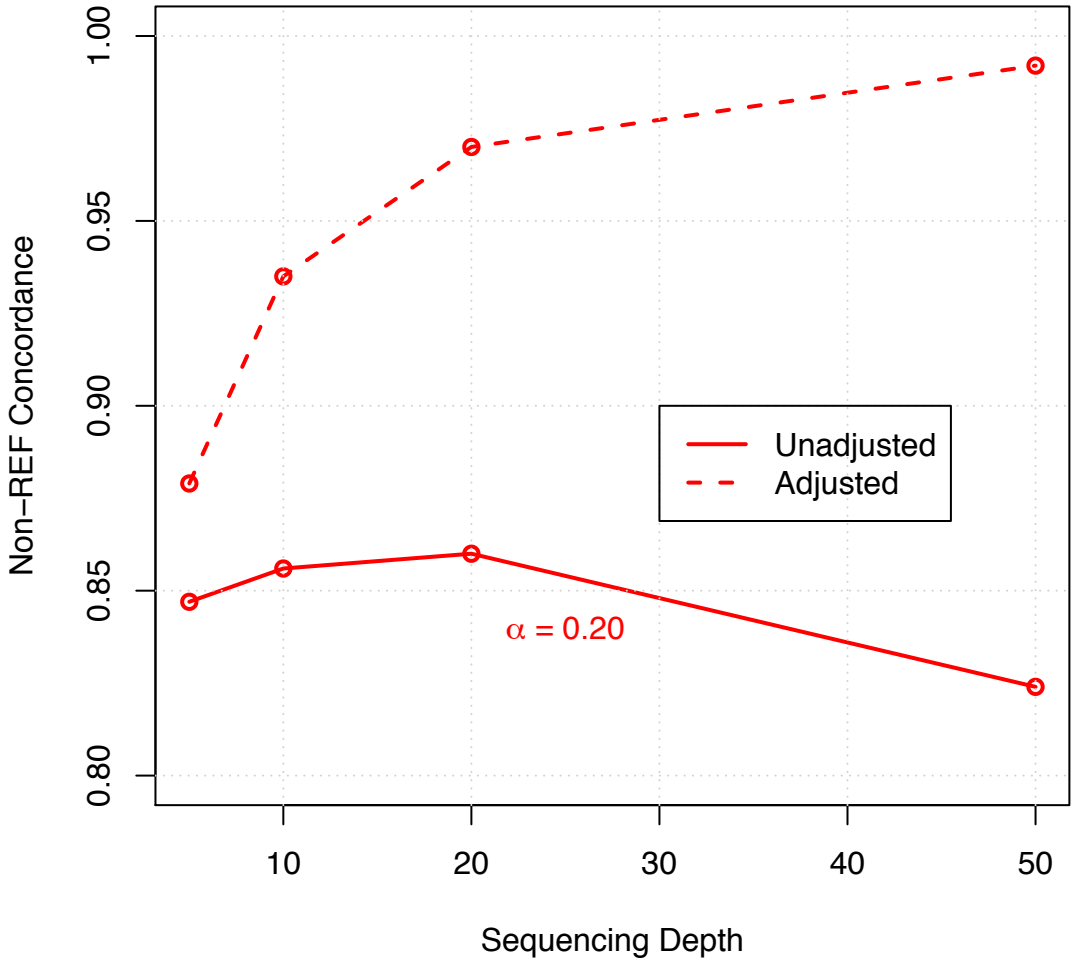
Simulation: Correcting Contamination



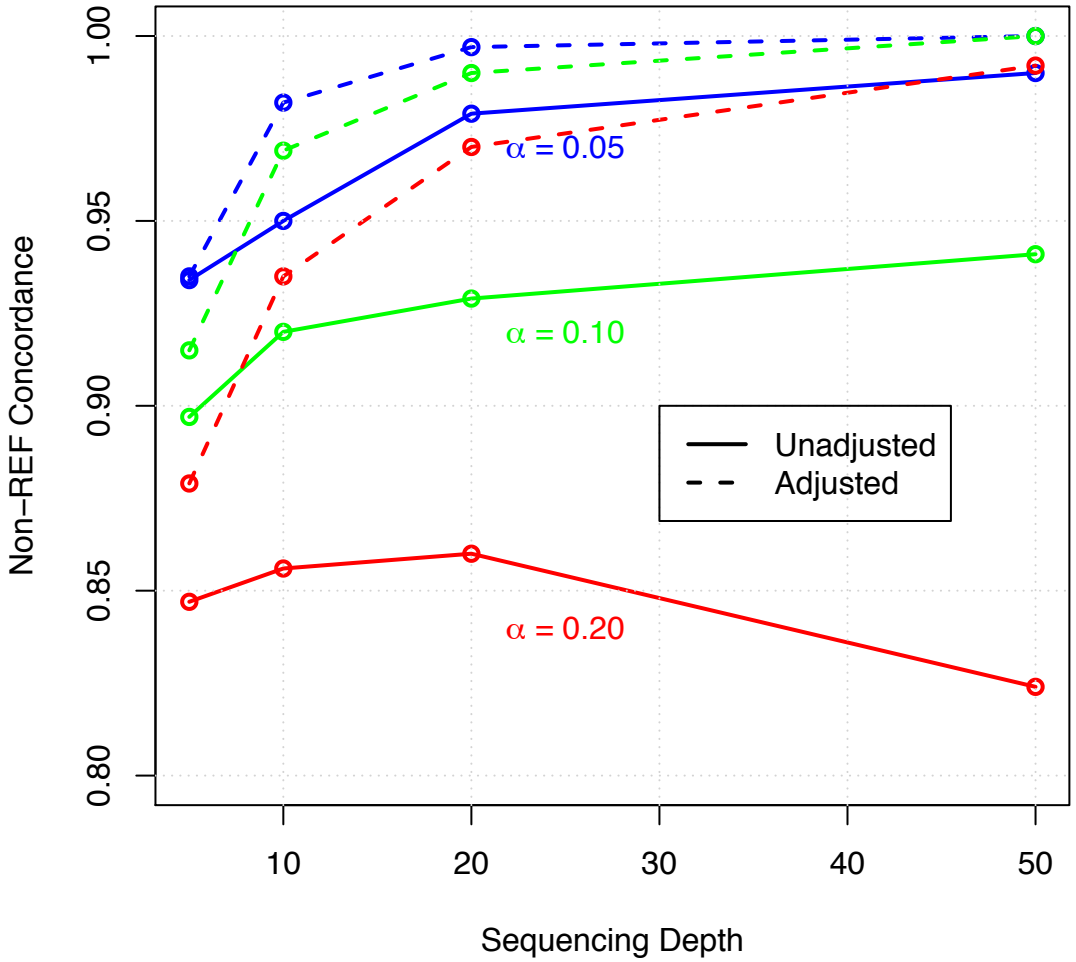
Simulation: Correcting Contamination



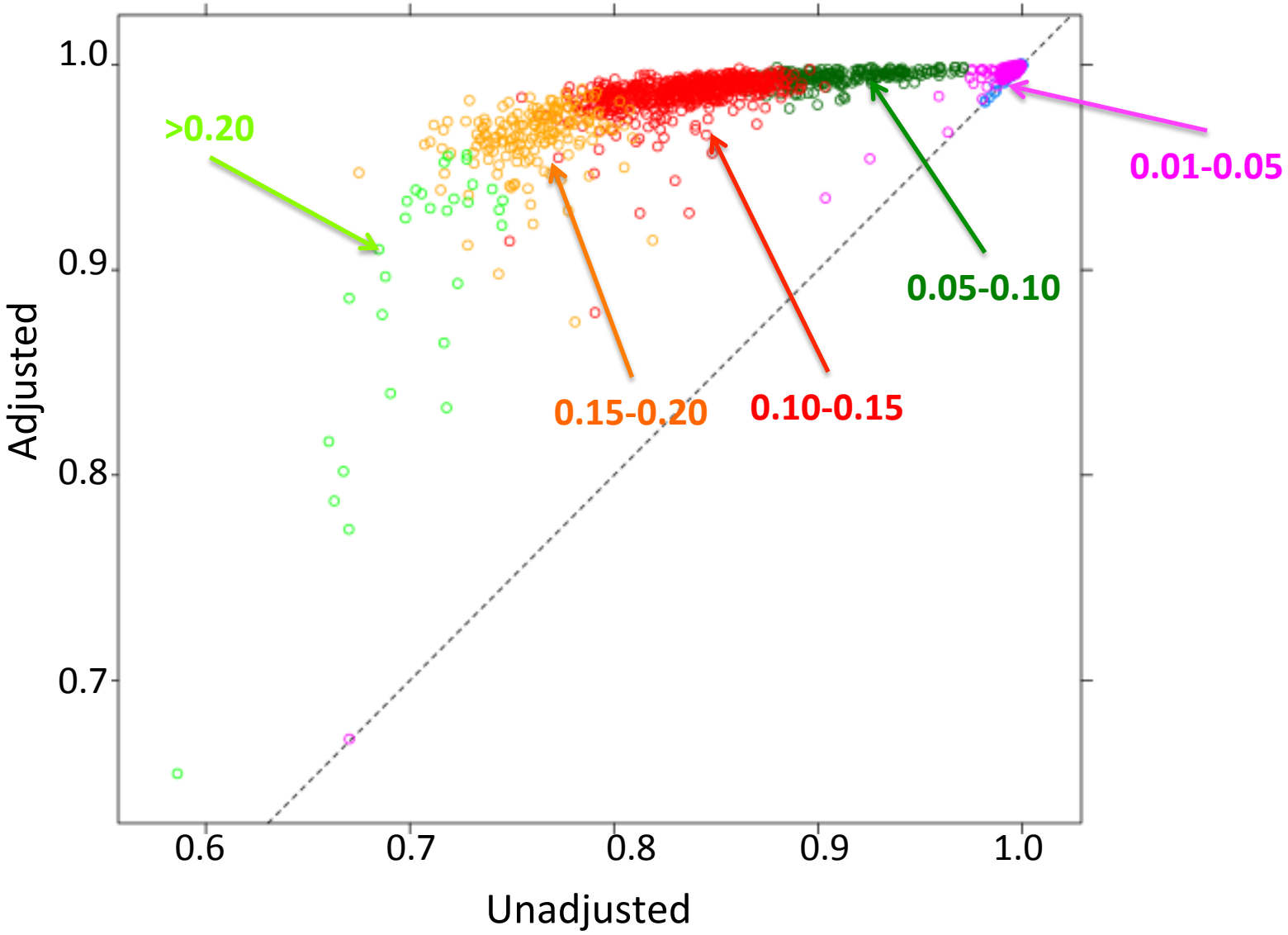
Simulation: Correcting Contamination



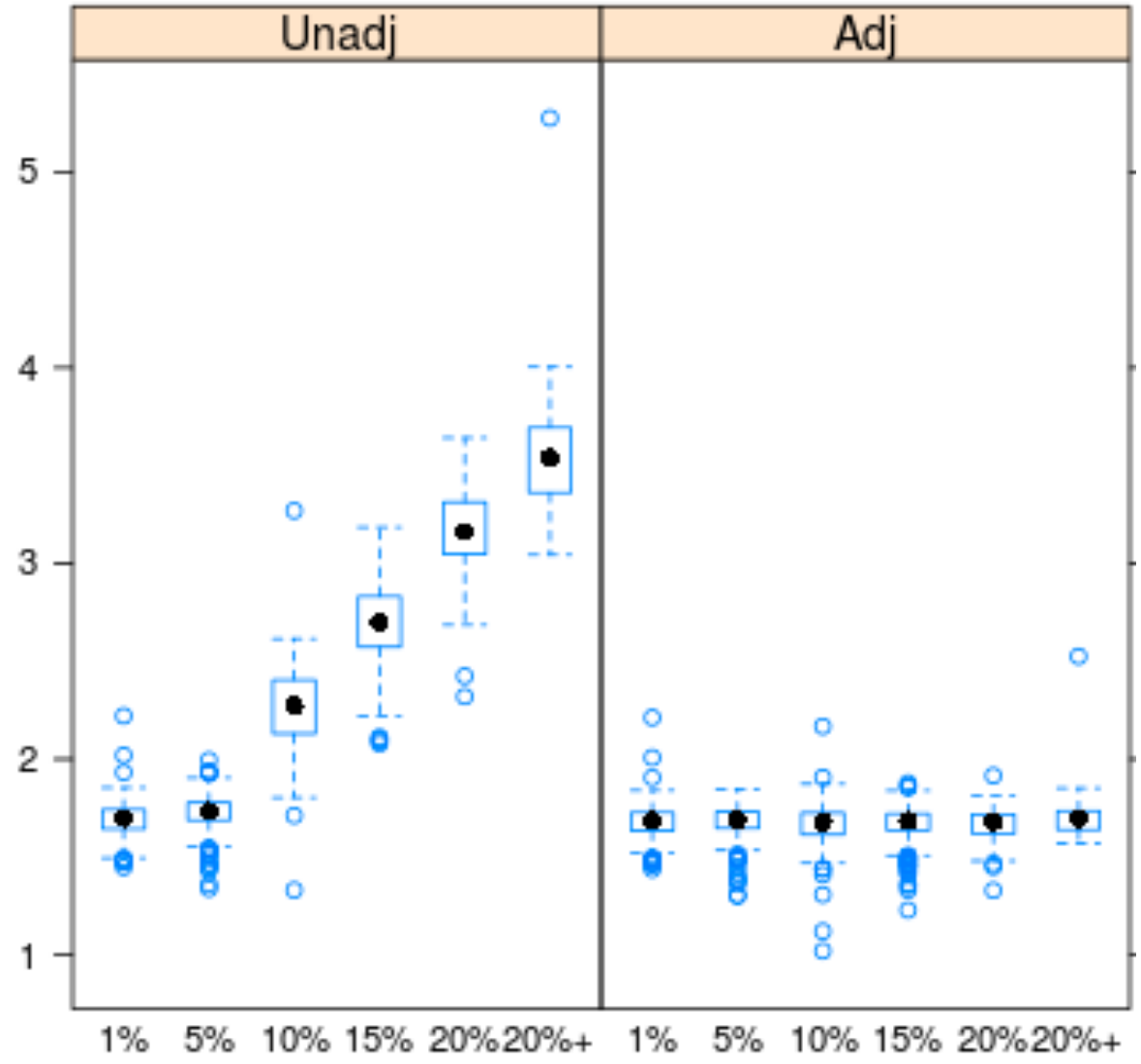
Simulation: Correcting Contamination



Results: Non-REF Genotype Concordance

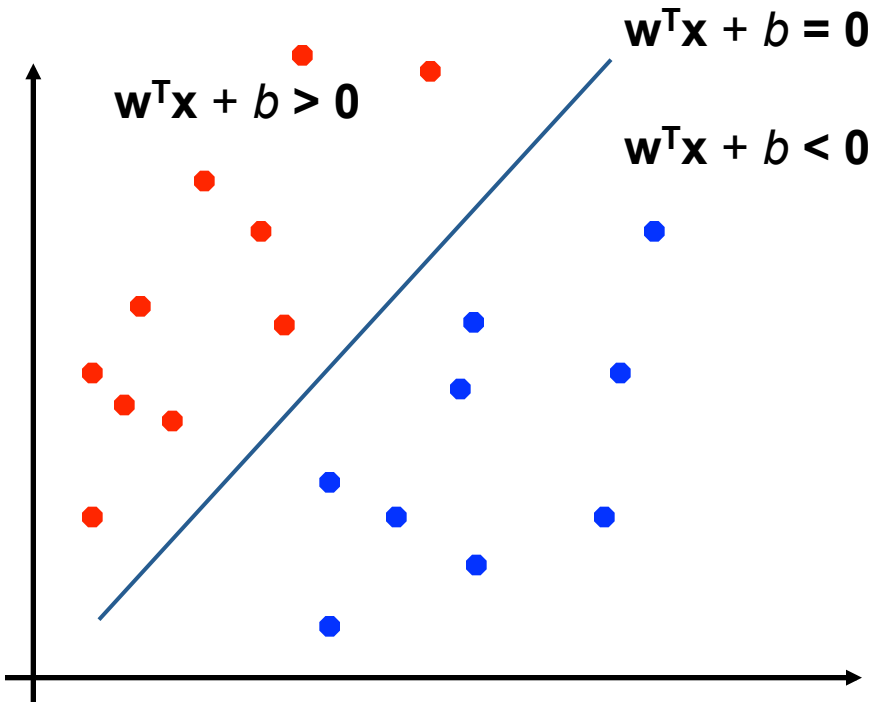


Results: HET/HOM Change



SVM Basics: Linear Separators

- Binary classification can be viewed as the task of separating classes in feature space:

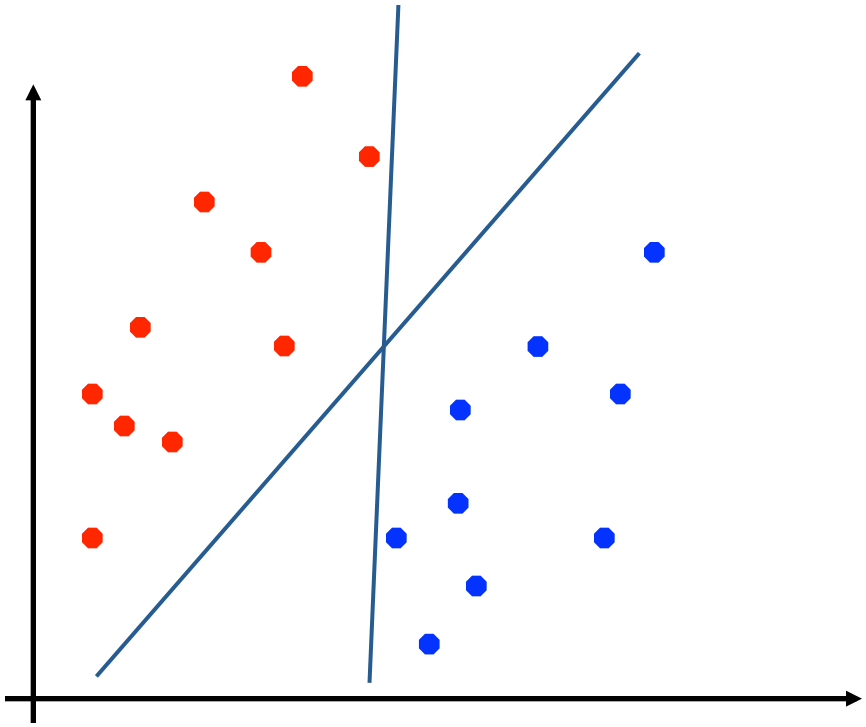


$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$$

[Figures: R. Mooney]

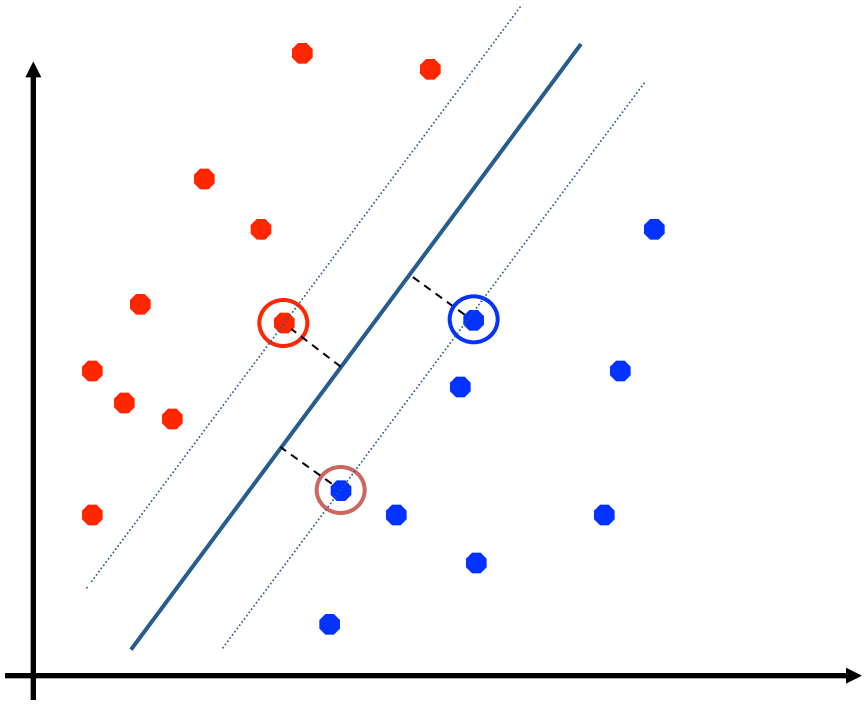
SVM Basics: Optimal Linear Separator

- Which of the linear separators is optimal?



SVM Basics: Maximum Margin

- Maximizing the margin is good according to intuition and PAC theory.
- Implies that only support vectors matter; other training examples are ignorable.



Non-linear SVMs: Kernel Trick

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable: