SeqShop Day 2:

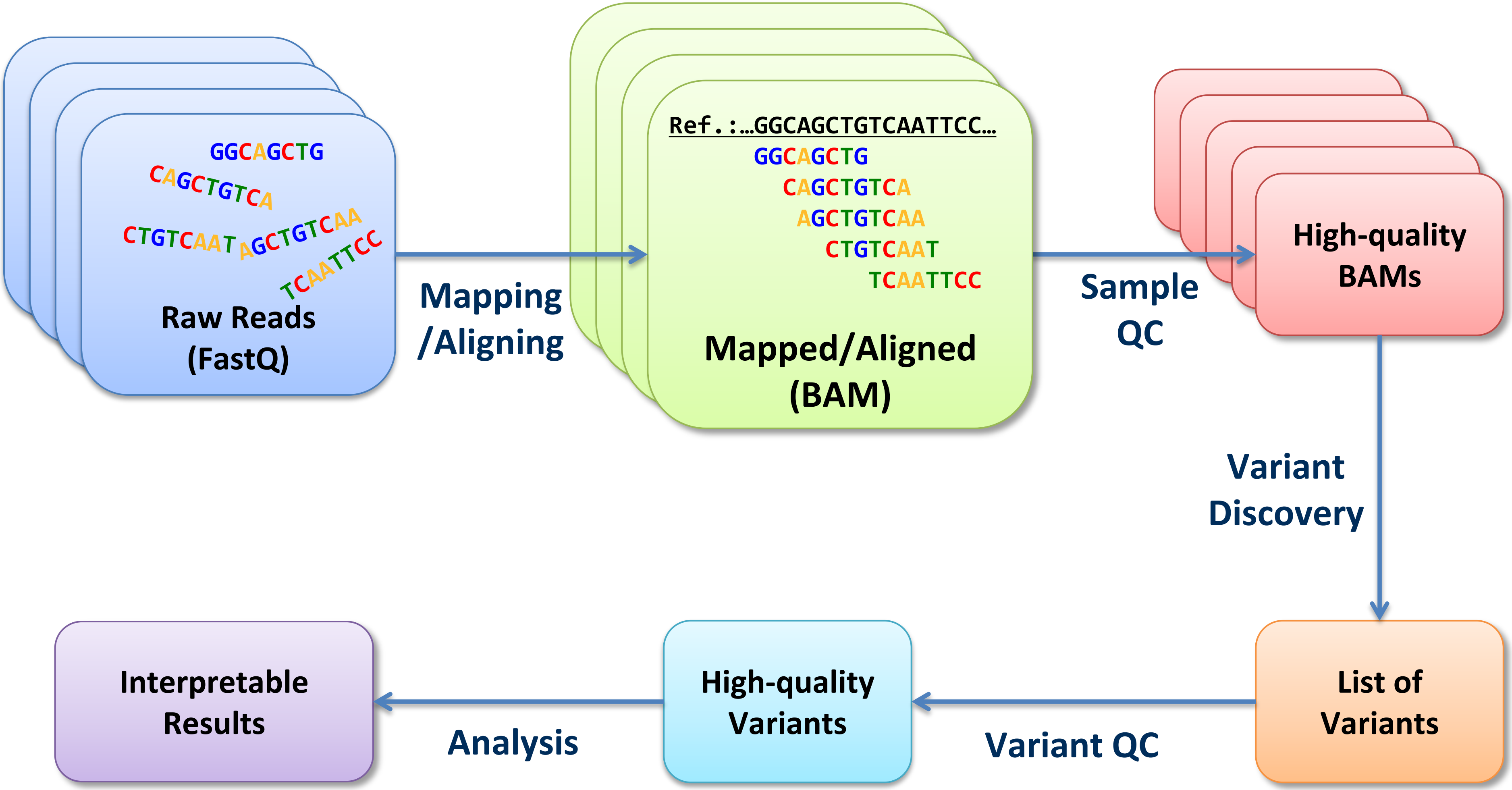
Detecting Contamination & SNP Calling

Goo Jun

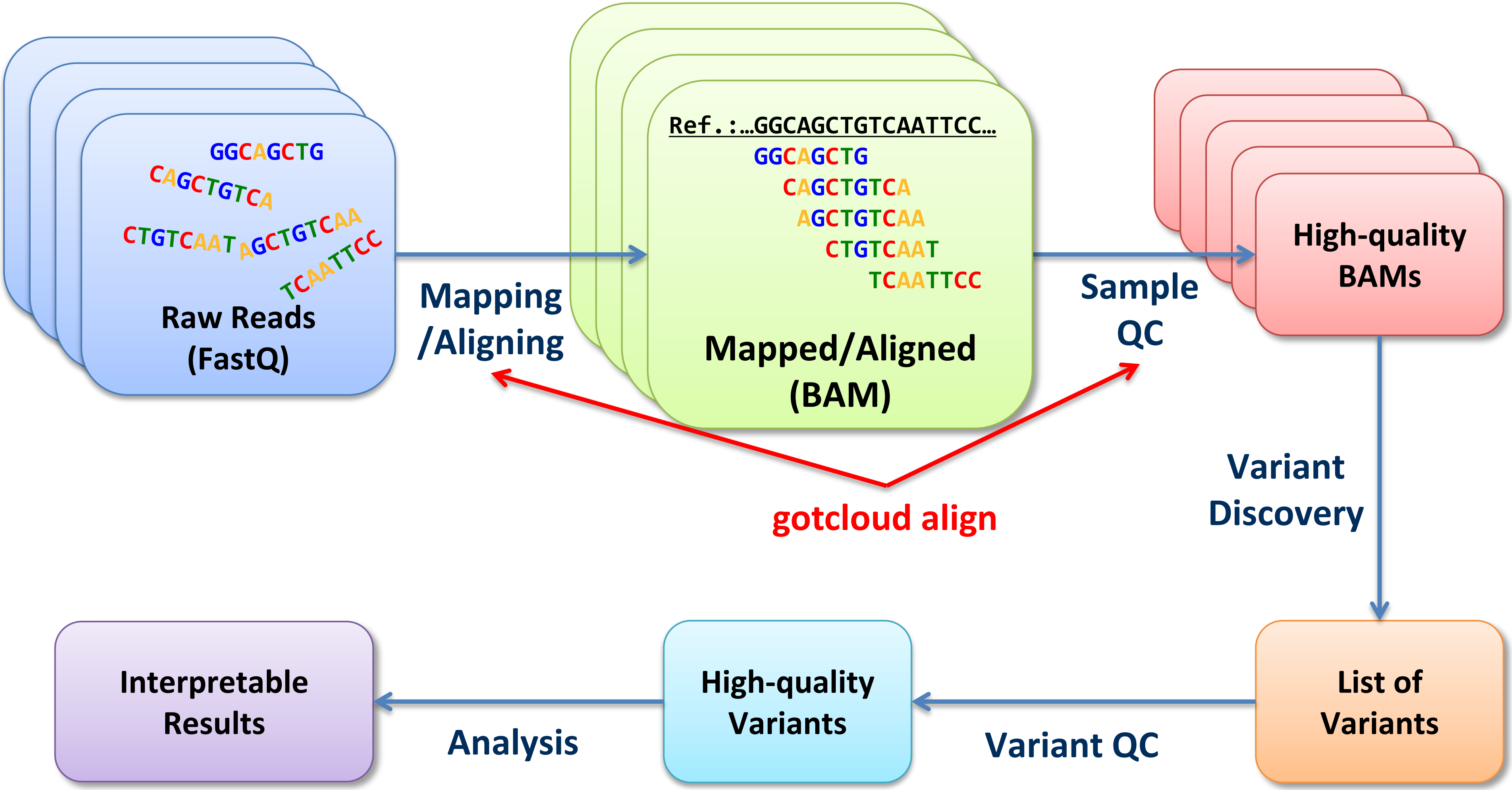
Center for Statistical Genetics & Dept. of Biostatistics

University of Michigan

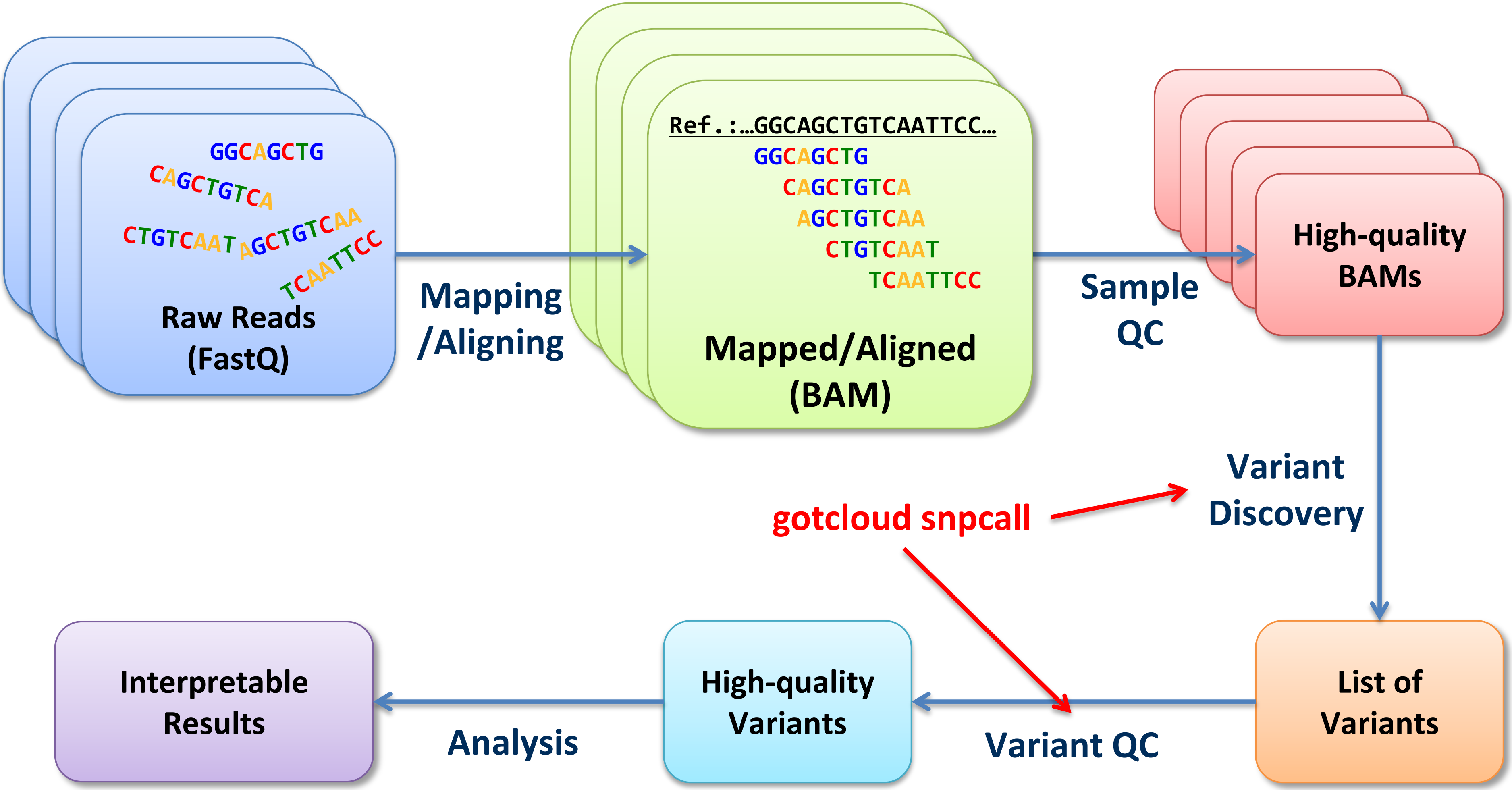
(Re)sequencing Data Analysis Flow



(Re)sequencing Data Analysis Flow



(Re)sequencing Data Analysis Flow



Part I

Estimating (and correcting) DNA sample contamination

DNA Sample Contamination

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

Contamination in Sequencing Data

- DNA contamination is common and serious
  - Timely feedback could save multi-million dollar project
  - Exact estimation and correction could save TB of data
- *In-silico* approach can solve *In-vitro* problems

Reference-Aligned Sequence Reads

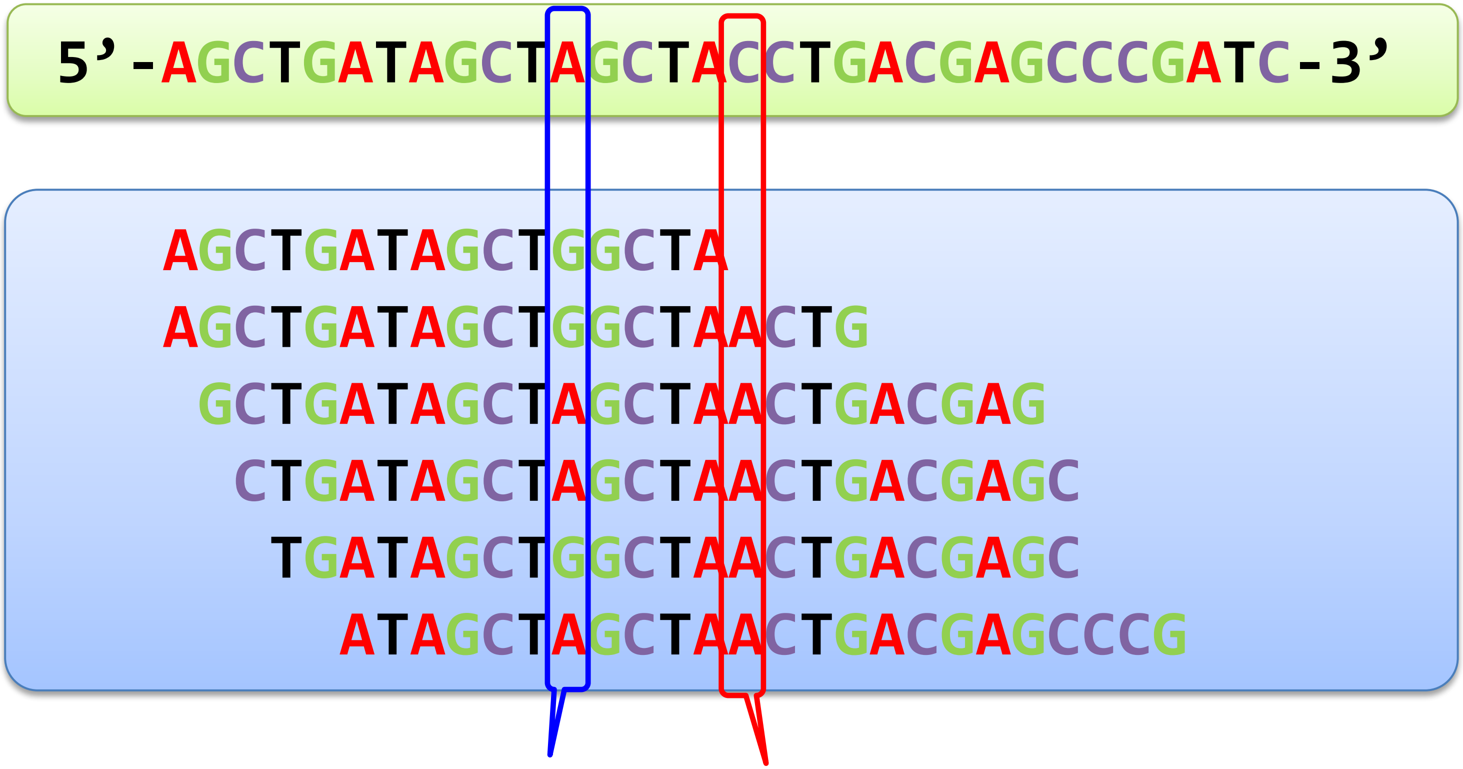
Reference

```
5’-AGCTGATAGCTAGCTACCTGACGAGCCCGATC-3’
```

Sample

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

Single-Nucleotide Polymorphism (SNP)



Reference

```
5’-AGCTGATAGCTAGCTACCTGACGAGCCCGATC-3’
```

Sample

```
AGCTGATAGCTGGCTA
AGCTGATAGCTGGCTAACTG
 GCTGATAGCTAGCTAACTGACGAG
  CTGATAGCTAGCTAACTGACGAGC
   TGATAGCTGGCTAACTGACGAGC
     ATAGCTAGCTAACTGACGAGCCCG
```

Genotype: GA
(Heterozygote)

Genotype: AA
(Homozygote)

Base Distribution in Two Samples

Reference

```
5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’
```

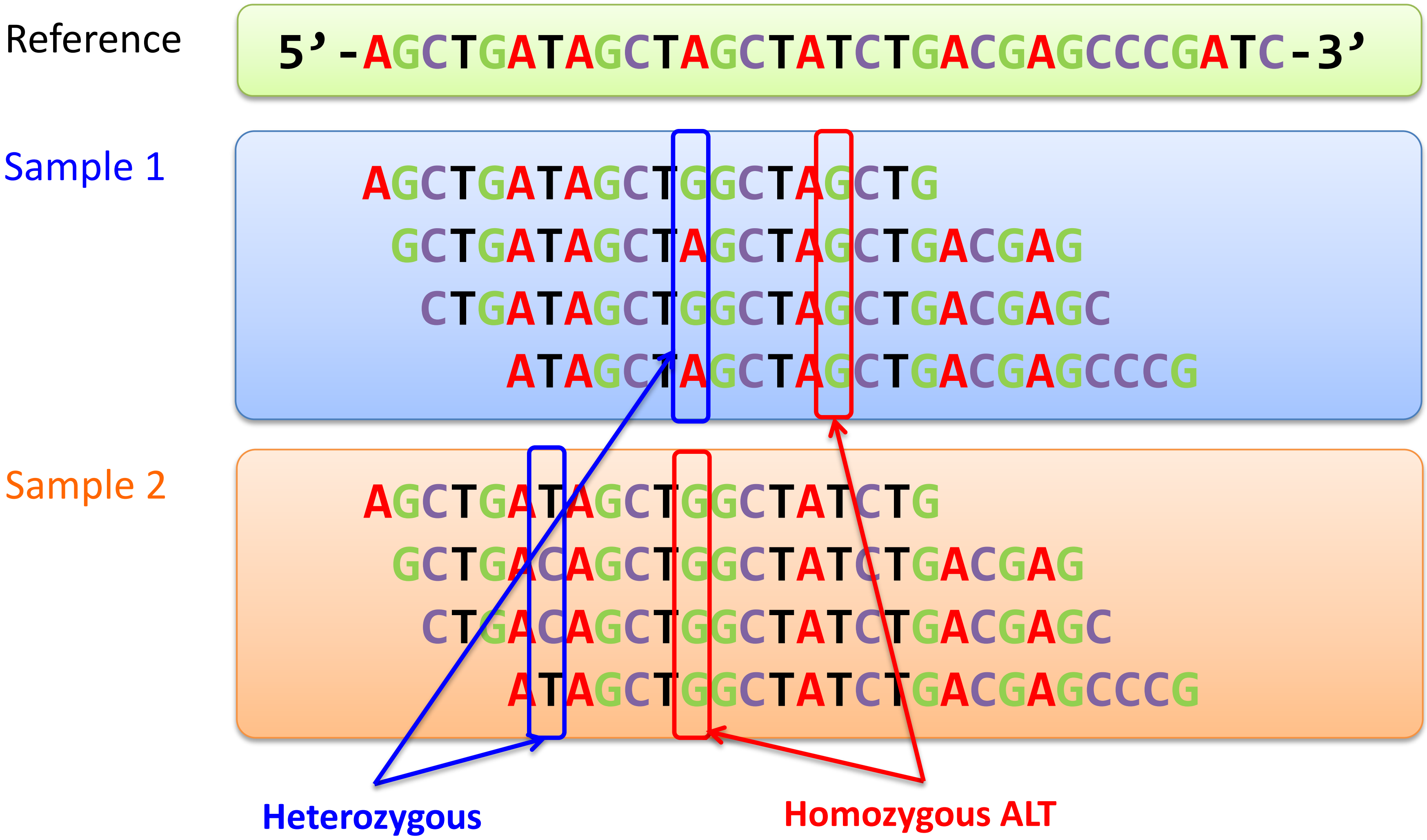
Sample 1

```
   AGCTGATAGCTGGCTAGCTG
    GCTGATAGCTAGCTAGCTGACGAG
     CTGATAGCTGGCTAGCTGACGAGC
        ATAGCTAGCTAGCTGACGAGCCCG
```

Sample 2

```
   AGCTGATAGCTGGCTATCTG
    GCTGACAGCTGGCTATCTGACGAG
     CTGACAGCTGGCTATCTGACGAGC
        ATAGCTGGCTATCTGACGAGCCCG
```

Base Distribution in Two Samples



Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1

AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG

Sample 2

AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG

Heterozygous

Homozygous ALT

Contamination: Mixture of Samples

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2

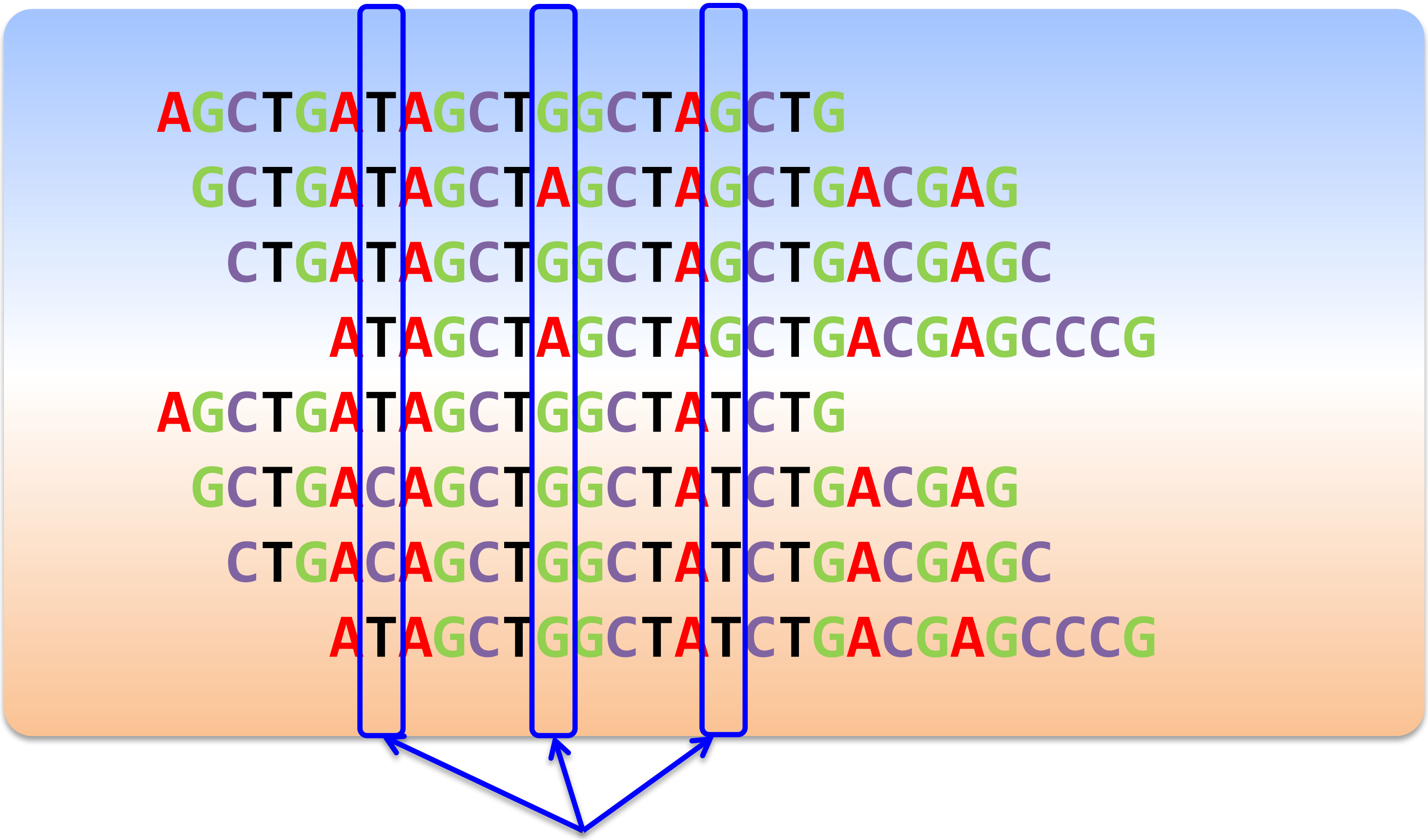
```
AGCTGATAGCTGGCTAGCTG
 GCTGATAGCTAGCTAGCTGACGAG
  CTGATAGCTGGCTAGCTGACGAGC
     ATAGCTAGCTAGCTGACGAGCCCG
AGCTGATAGCTGGCTATCTG
 GCTGACAGCTGGCTATCTGACGAG
  CTGACAGCTGGCTATCTGACGAGC
     ATAGCTGGCTATCTGACGAGCCCG
```

Contamination: Changes Base Distributions

Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1+2



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

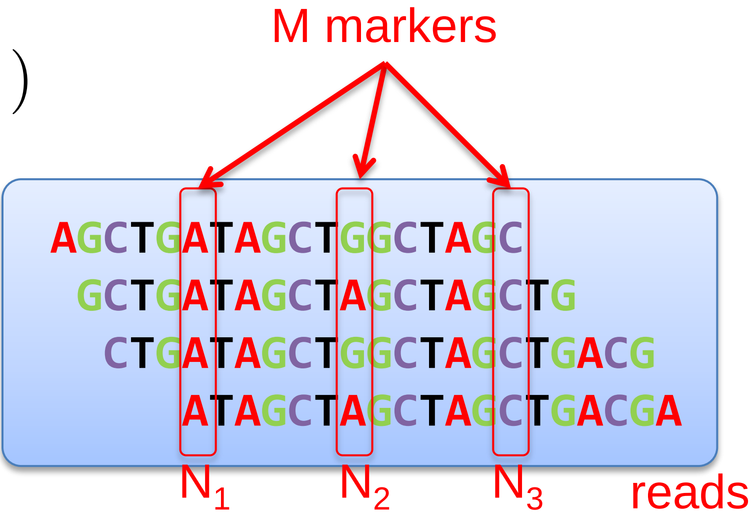
More heterozygote SNPs with biased distribution

Likelihood of Base Reads

- M markers
- N_i base reads: $\mathbf{b}_i = (b_{i1}, b_{i2}, \ldots, b_{iN_i})$

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

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



Likelihood of observed bases at *i*-th marker, given $G_i \in \{\text{AA}, \text{AB}, \text{BB}\}$

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

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

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

Likelihood from original sample

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

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

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

Likelihood from *contaminating* sample

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

Known genotypes for M sites (CHIPMIX)

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

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

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

From population allele freq. under HWE

Two-sample Mixture Model

- Likelihood with mixing proportion α

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

Contamination level: MLE of α

Simple Likelihood Model

- Probability of observing a base (b) depends on
  - Underlying (true) genotype (G)
  - Occurrence of base read error (e)
  - Example
    - P(b = A | G = AA, **no error (e=0)**) = 1
    - P(b = G | G = TT, ***error (e=1)***) = 1/3

    (In case of base read error, assume all possibilities are equally likely)
  - P(b | G) = P(b | G, e=0) P(e=0) + P(b | G, e=1) P(e=1)
  - P(e) from phred-scale base quality for j-th read in i-th site:

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

Estimation with Sequence Data Only (FREEMIX)

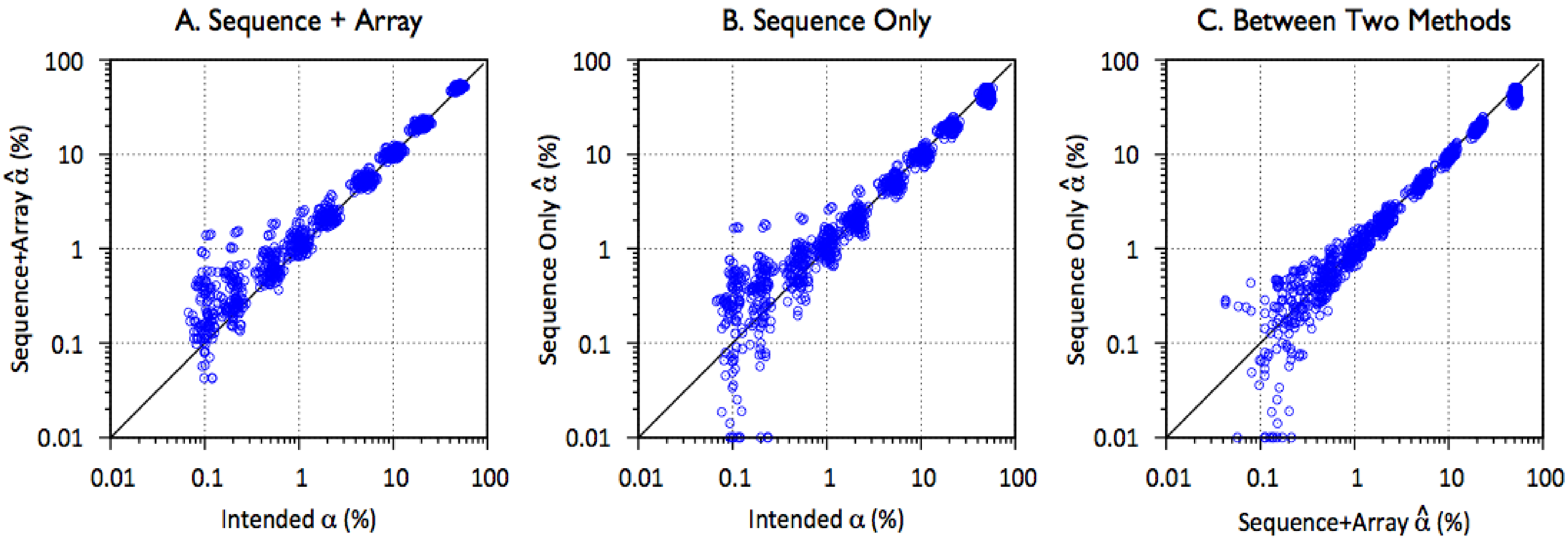
- If sequenced sample does not have external genotypes
  - Model both genotypes from population allele frequency

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

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

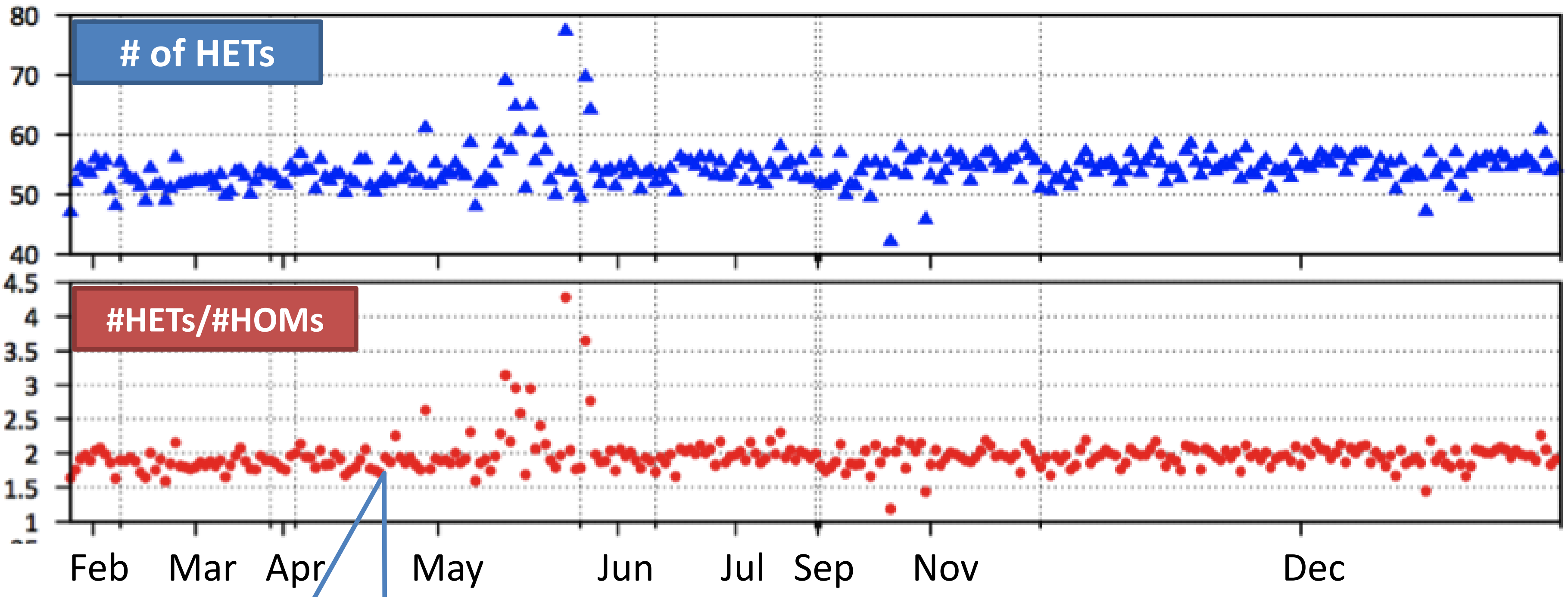
Results: Simulation

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



Results: Type-2 Diabetes Sequencing Study

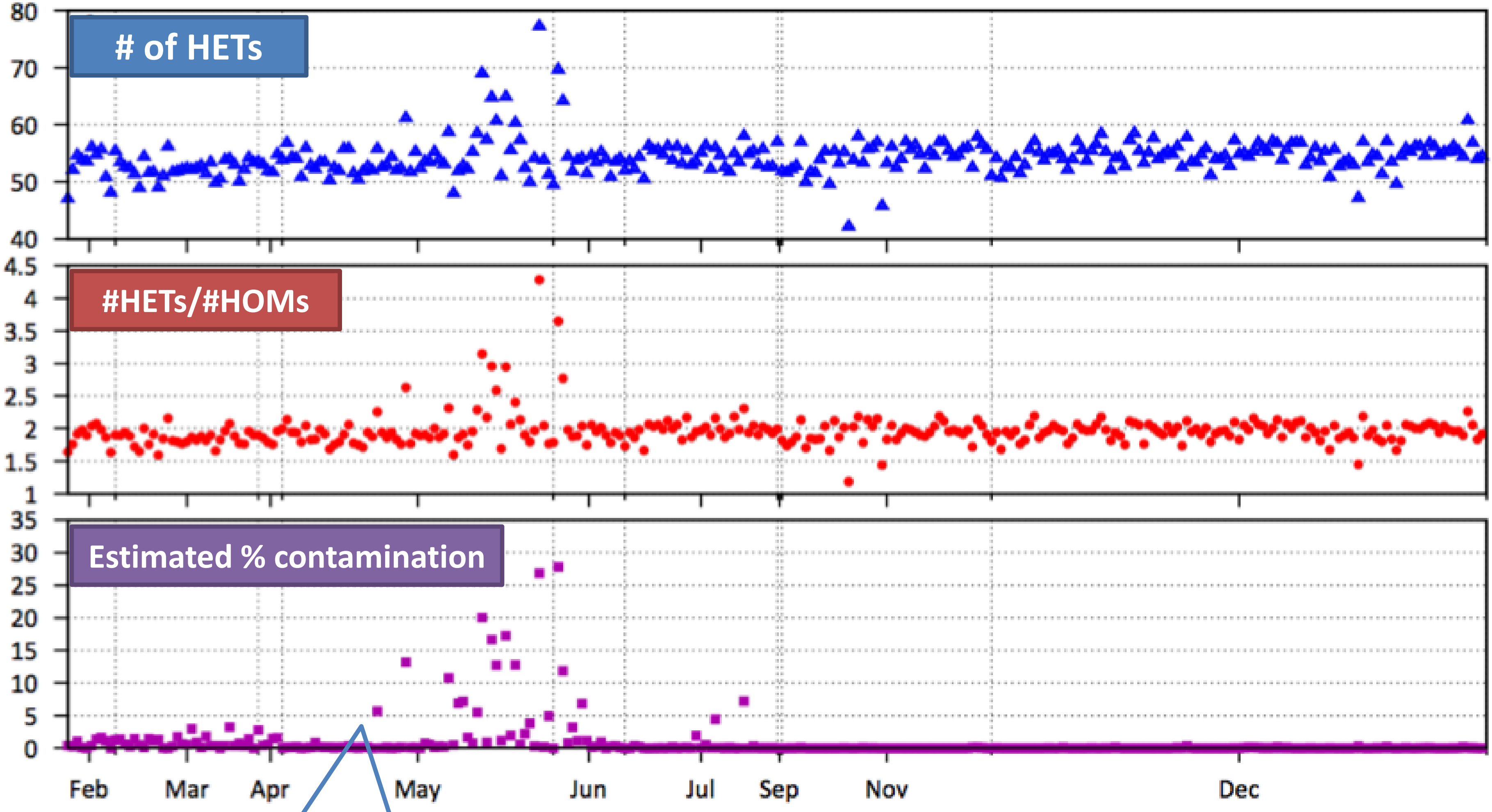
~2800 Whole genome sequences



Sequencing Date in 2010

Something changed

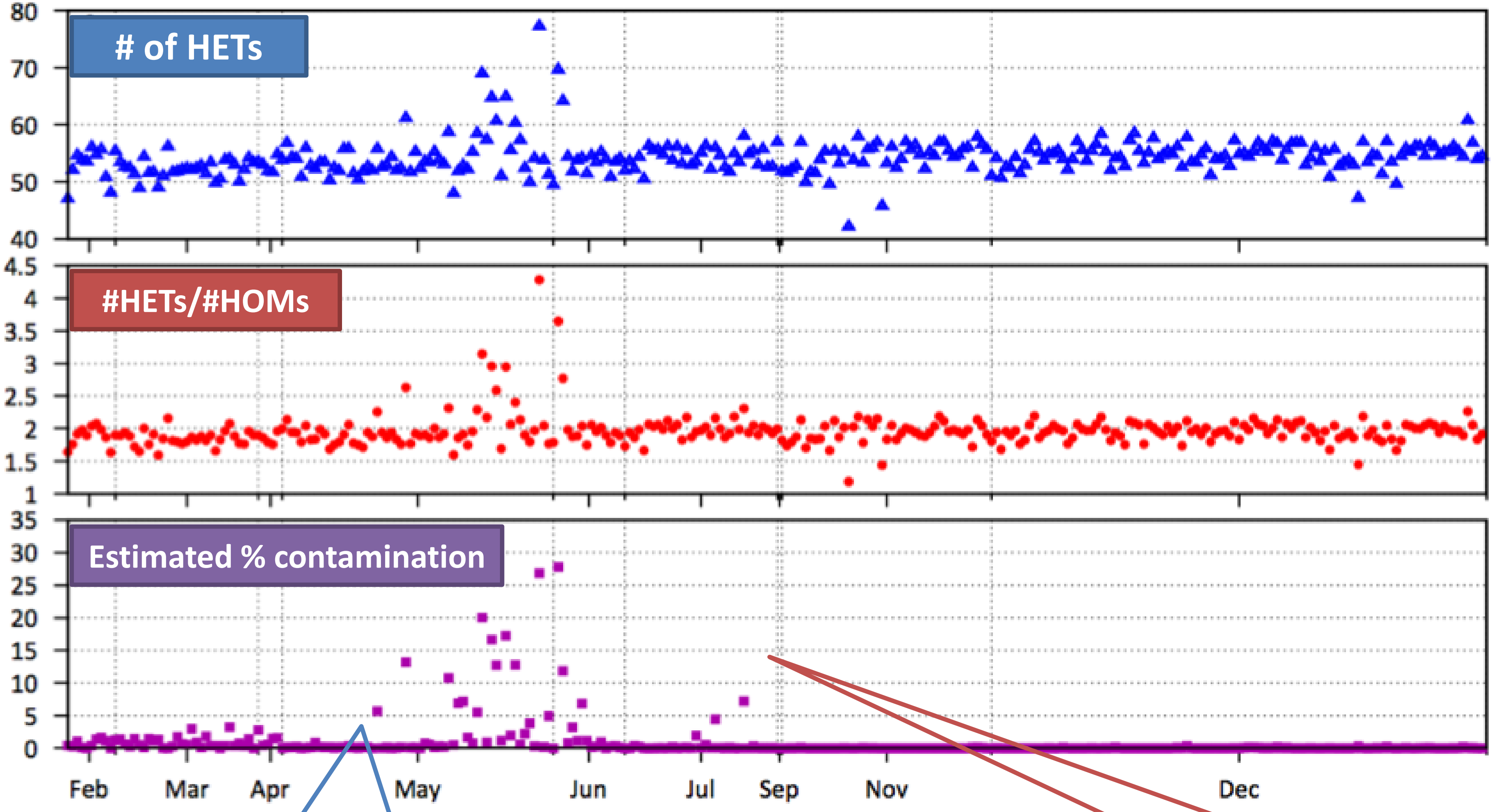
Results: Type-2 Diabetes Sequencing Study



Sequencing Date

Something changed

Results: Type-2 Diabetes Sequencing Study

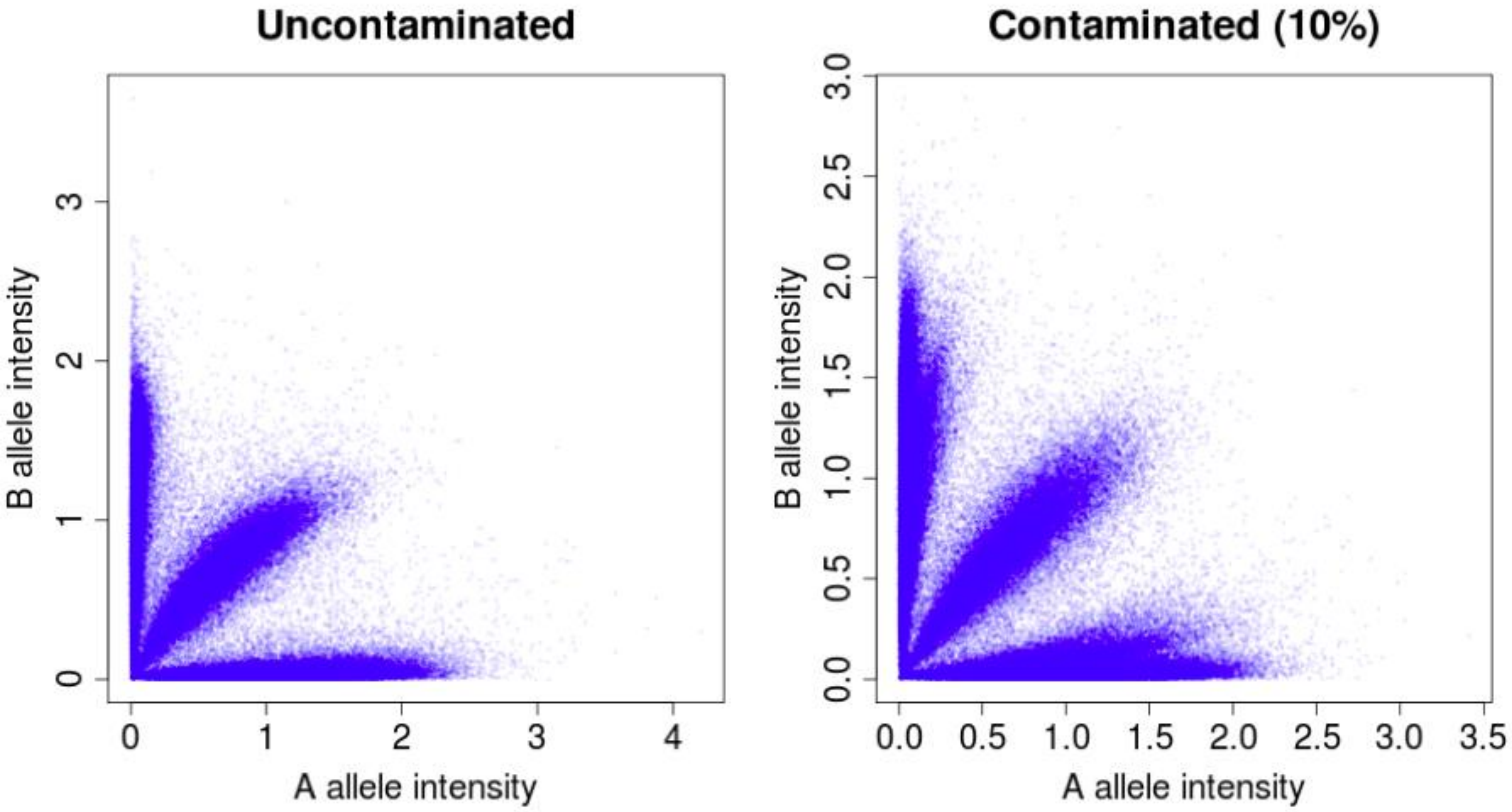


Sequencing Date

Something changed

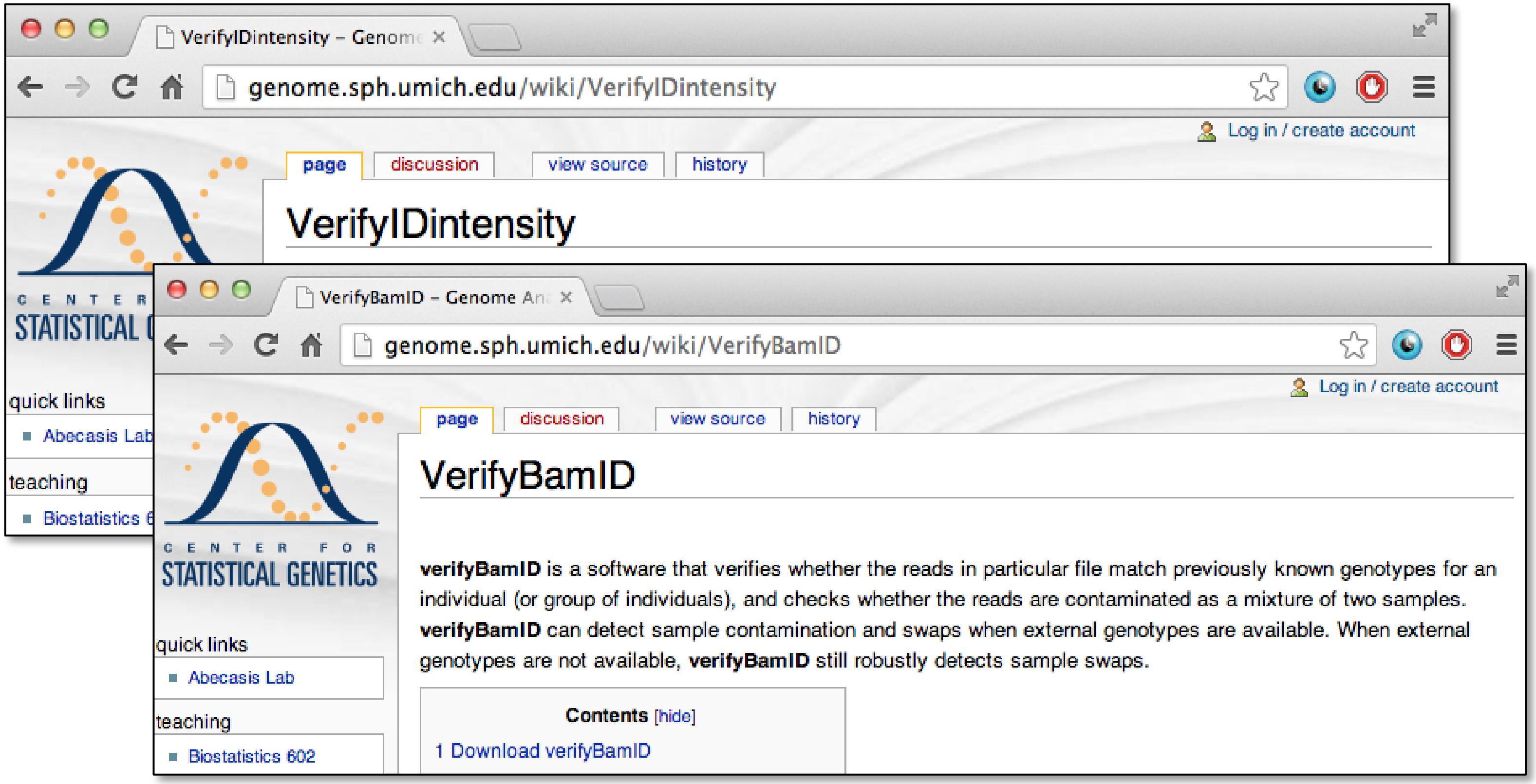
No contamination after a protocol change

Contamination in Array Intensity Data



Software for Contamination Problems

- Software tools to check contamination:
  - http://genome.sph.umich.edu/wiki/VerifyBamID
  - http://genome.sph.umich.edu/wiki/VerifyIDintensity



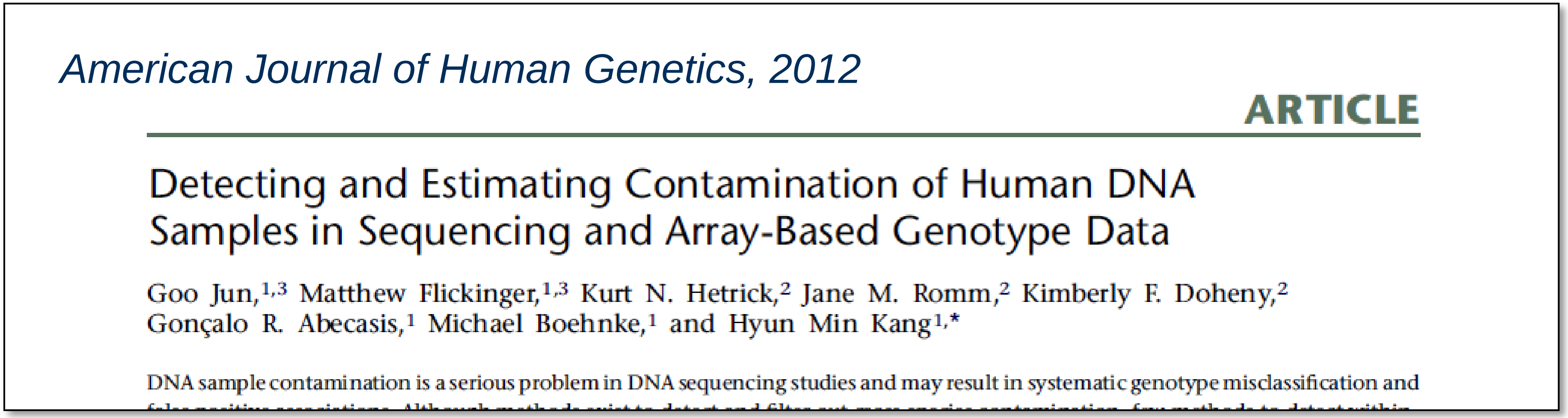
VerifyBamID

verifyBamID is a software that verifies whether the reads in particular file match previously known genotypes for an individual (or group of individuals), and checks whether the reads are contaminated as a mixture of two samples. **verifyBamID** can detect sample contamination and swaps when external genotypes are available. When external genotypes are not available, **verifyBamID** still robustly detects sample swaps.

Contents [hide]

Estimation & Correction of DNA Contamination

- Likelihood-based model accurately estimates of % of potential sample contamination.



American Journal of Human Genetics, 2012

ARTICLE

Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data

Goo Jun,[1,3] Matthew Flickinger,[1,3] Kurt N. Hetrick,[2] Jane M. Romm,[2] Kimberly F. Doheny,[2] Gonçalo R. Abecasis,[1] Michael Boehnke,[1] and Hyun Min Kang[1,*]

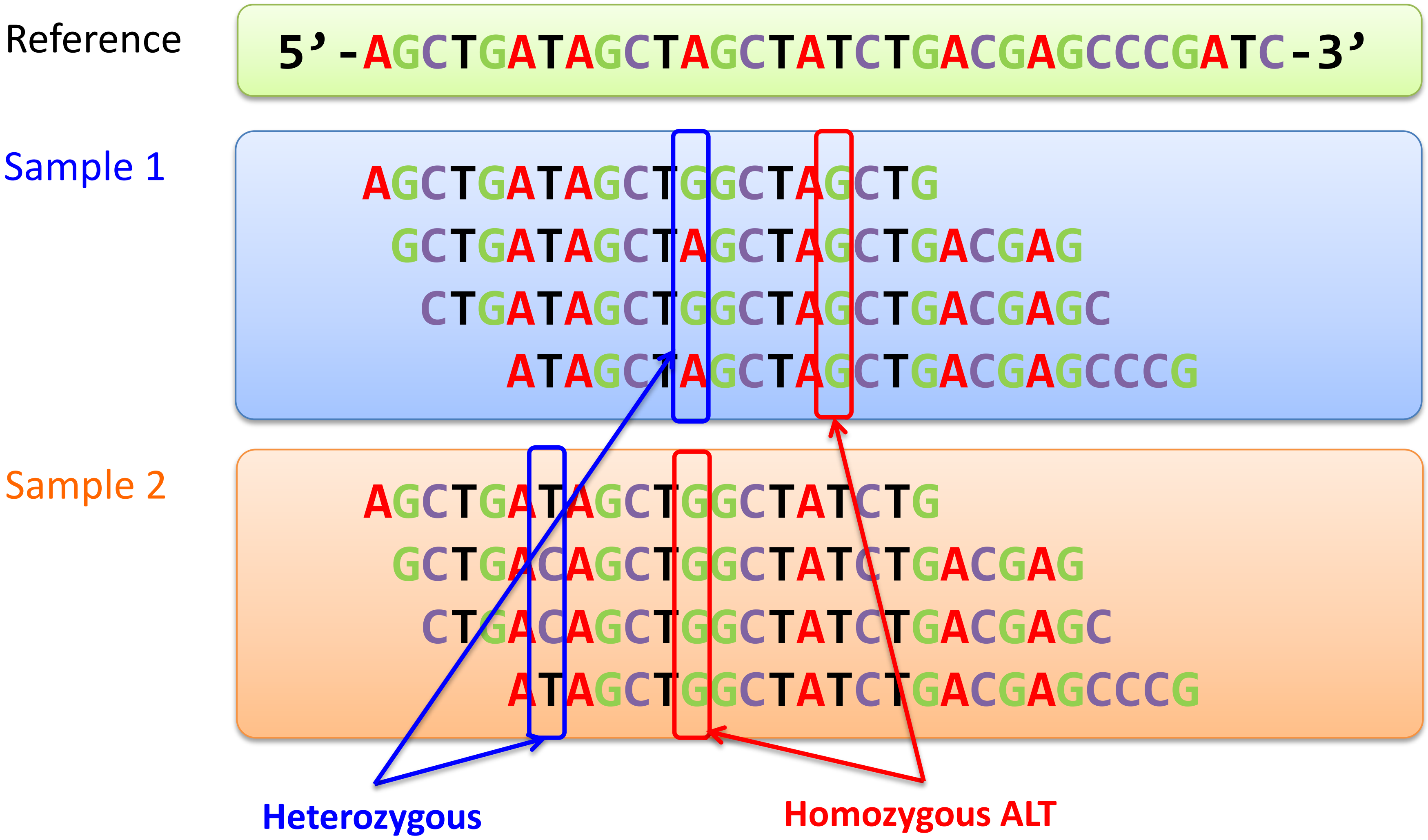
DNA sample contamination is a serious problem in DNA sequencing studies and may result in systematic genotype misclassification and false positive associations. Although methods exist to detect and filter out cross-species contamination, few methods to detect within

- The sample likelihood model can be used to correct genotype likelihoods, which greatly improves genotype accuracies.
    - Manuscript in progress (w/ M. Flickinger)

Part II

Efficient and Scalable Software Pipeline for Large-scale Sequence Data

Base Distribution in Two Samples



Reference

5’-AGCTGATAGCTAGCTATCTGACGAGCCCGATC-3’

Sample 1

AGCTGATAGCTGGCTAGCTG
GCTGATAGCTAGCTAGCTGACGAG
CTGATAGCTGGCTAGCTGACGAGC
ATAGCTAGCTAGCTGACGAGCCCG

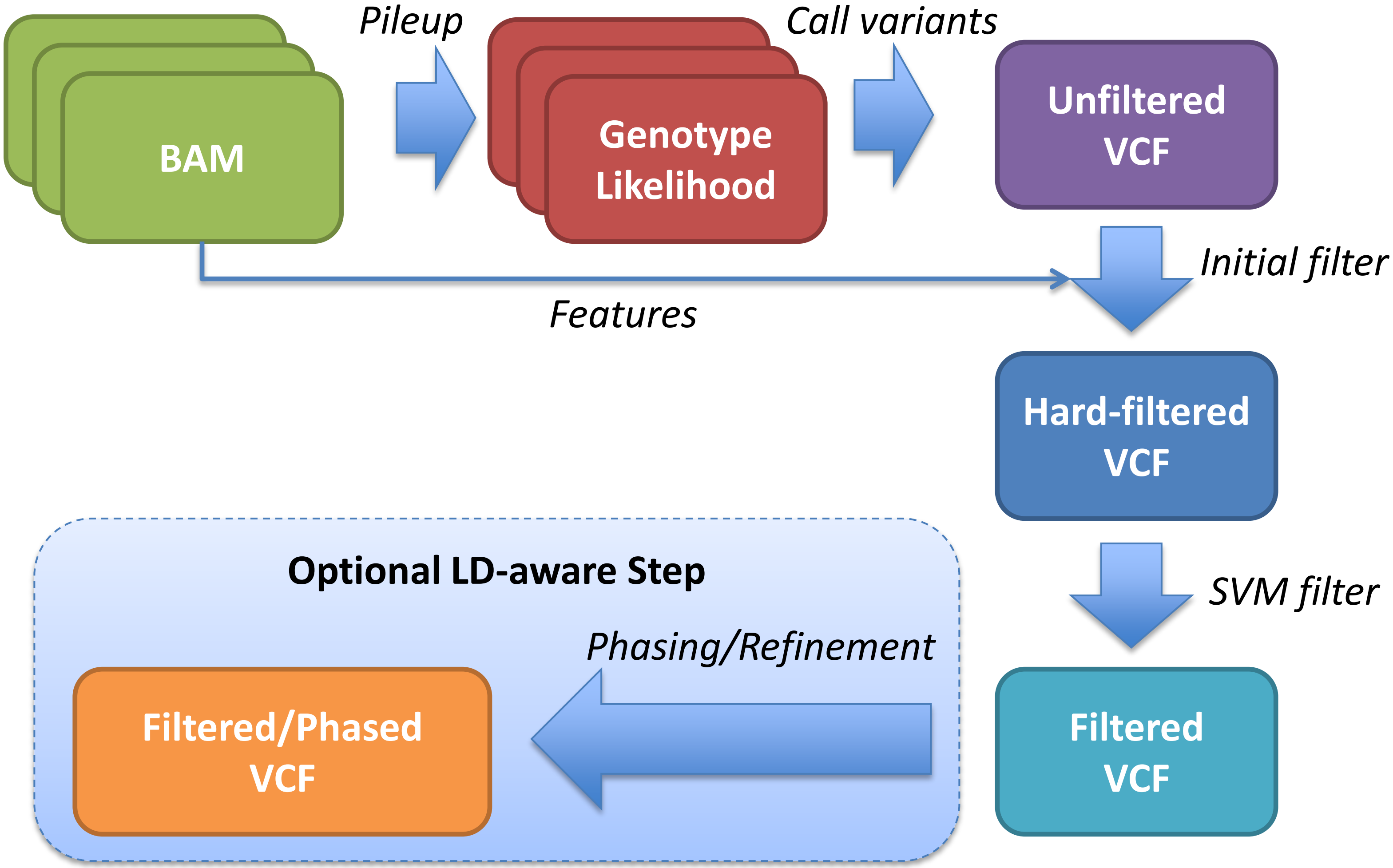
Sample 2

AGCTGATAGCTGGCTATCTG
GCTGACAGCTGGCTATCTGACGAG
CTGACAGCTGGCTATCTGACGAGC
ATAGCTGGCTATCTGACGAGCCCG

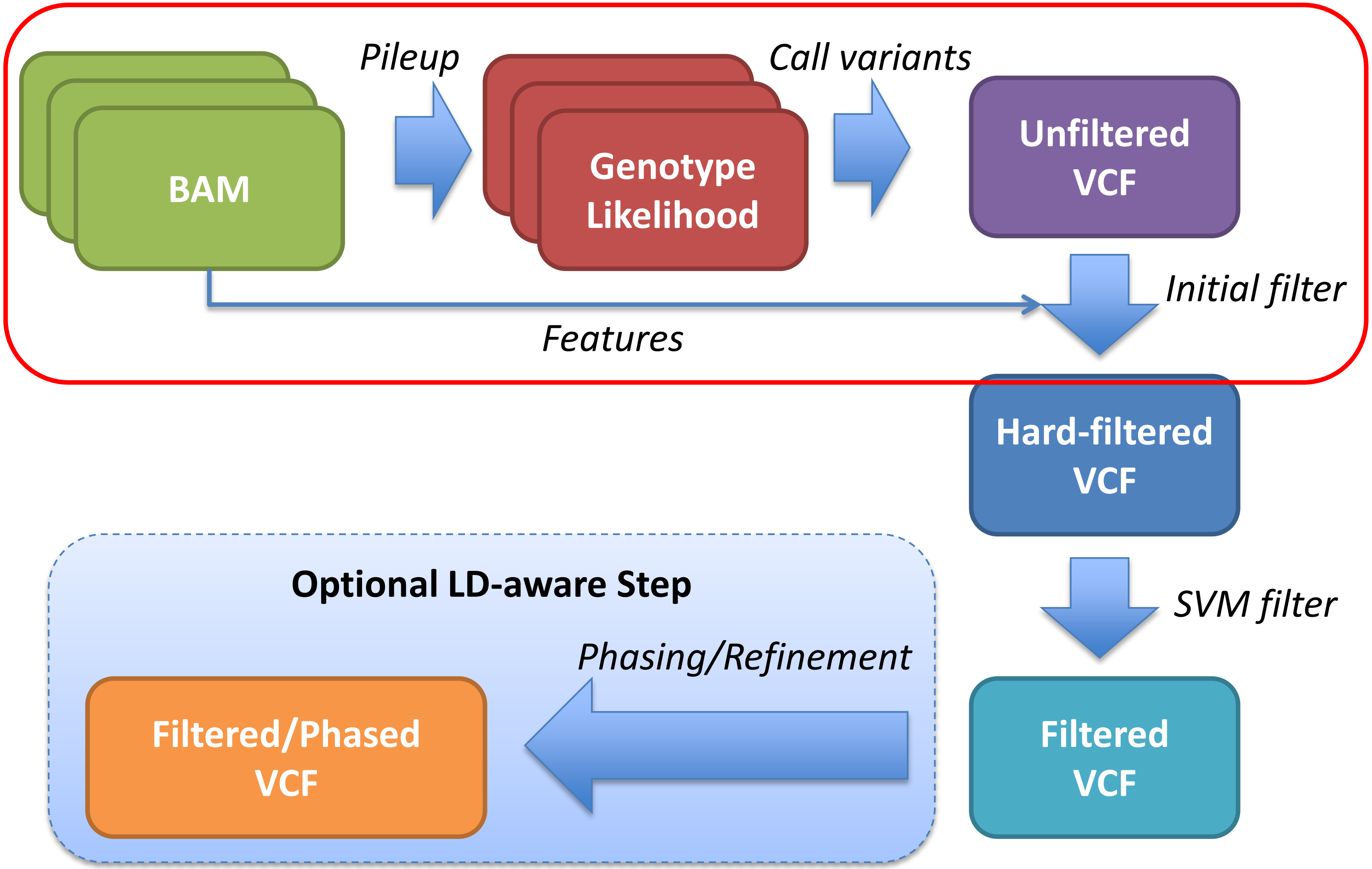
Heterozygous

Homozygous ALT

GotCloud SNP Calling Pipeline



Variant Calling From Sequence Reads



Calling Consensus Genotypes

- 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

Shotgun Sequence Data

★

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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

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

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

Possible Genotypes

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

Individual Based Prior

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

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

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

Population 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) =** 0.04	**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

Population-based Prior for a Bi-allelic SNP

- Prior probability of a site being a SNP with alleles {a,b}:

$$Pr(\mathrm{SNP}) = \theta \sum_{i=1}^{2n} \frac{1}{i}, \quad \theta = 10^{-3}$$

  - *n* : number of individuals
  - Based on neutral coalescence model

- Simple prior for each {a,b} pair

$$Pr(\mathrm{SNP}_{\{a,b\}}) = \theta \sum_{i=1}^{2n} \frac{1}{n} \times \begin{cases} 1/3 & \text{for } \mathrm{SNP}_{\{REF,ALT\}} \\ 10^{-3} & \text{all others} \end{cases}$$

Posterior Probability of Being an Bi-allelic SNP

- Posterior probability of being a SNP with reads

Prior

$$\Pr(\mathrm{SNP}_{\{a,b\}}|\mathrm{reads}) = \frac{\Pr(\mathrm{reads}|\mathrm{SNP}_{\{a,b\}})\Pr(\mathrm{SNP}_{\{a,b\}})}{\sum_{\{a,b\}}\Pr(\mathrm{reads}|\mathrm{SNP}_{\{a,b\}})\Pr(\mathrm{SNP}_{\{a,b\}})}$$

$$\Pr(\mathrm{reads}|\mathrm{SNP}_{\{a,b\}}) = \prod_{i=1}^{n}\sum_{g}\Pr(G_i = g|\mathrm{SNP}_{\{a,b\}})\Pr(\mathrm{reads}_i|G_i = g)$$

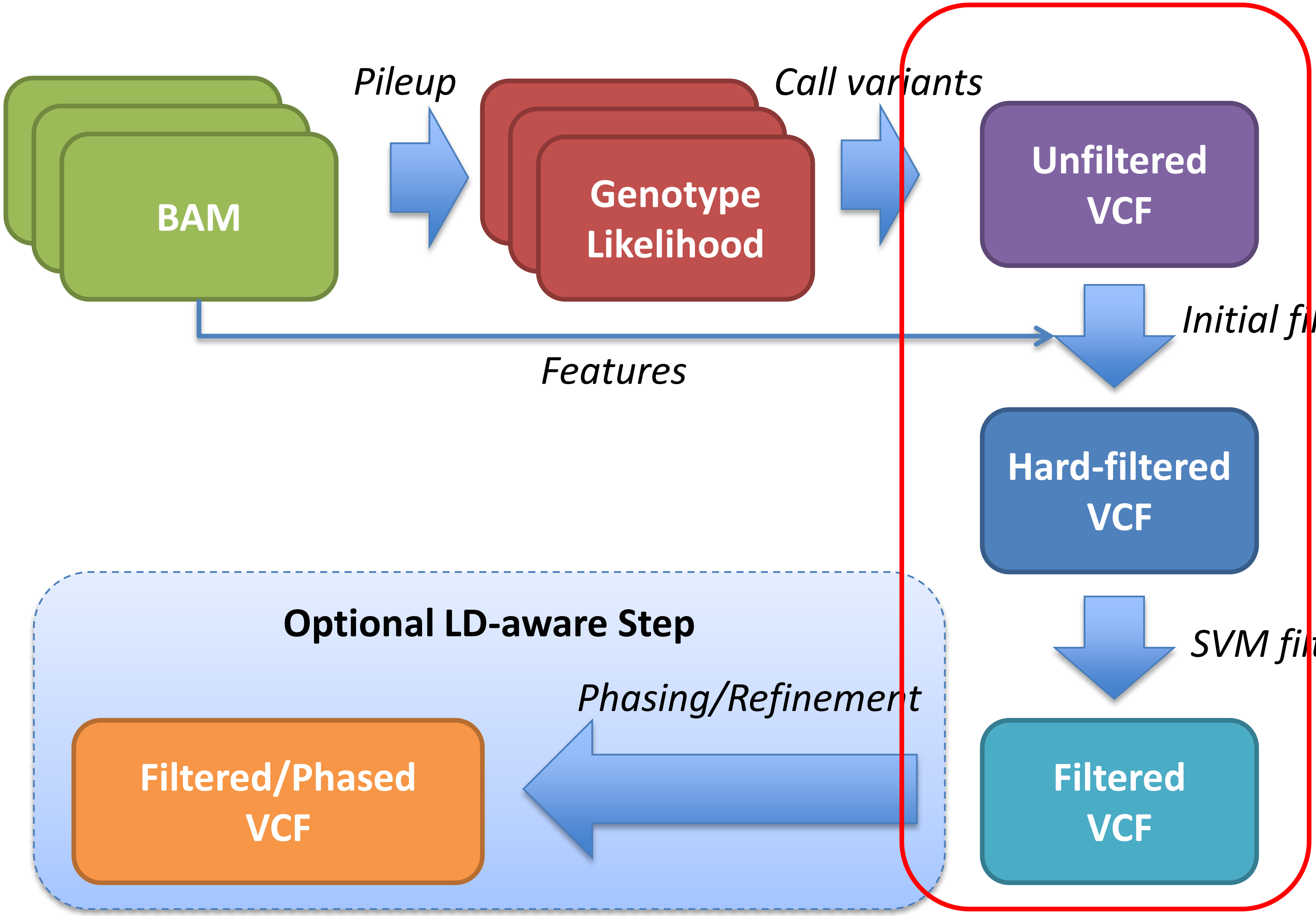
From HWE at MLE of allele freq.

Genotype Likelihood

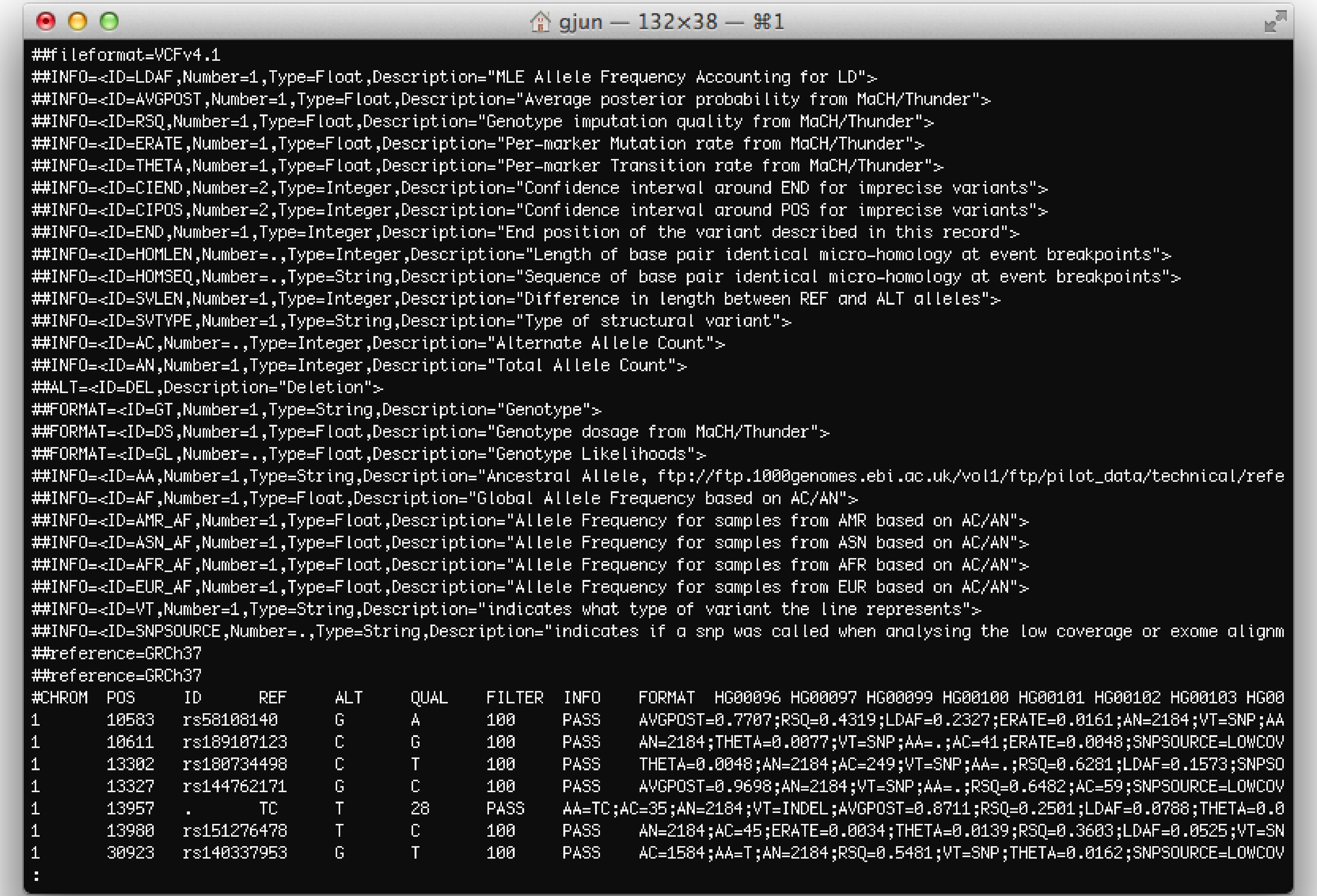
- Multi-sample statistic minimizes false discoveries!

**Other toolsets have different models for likelihood and posterior*

Variant Filtering



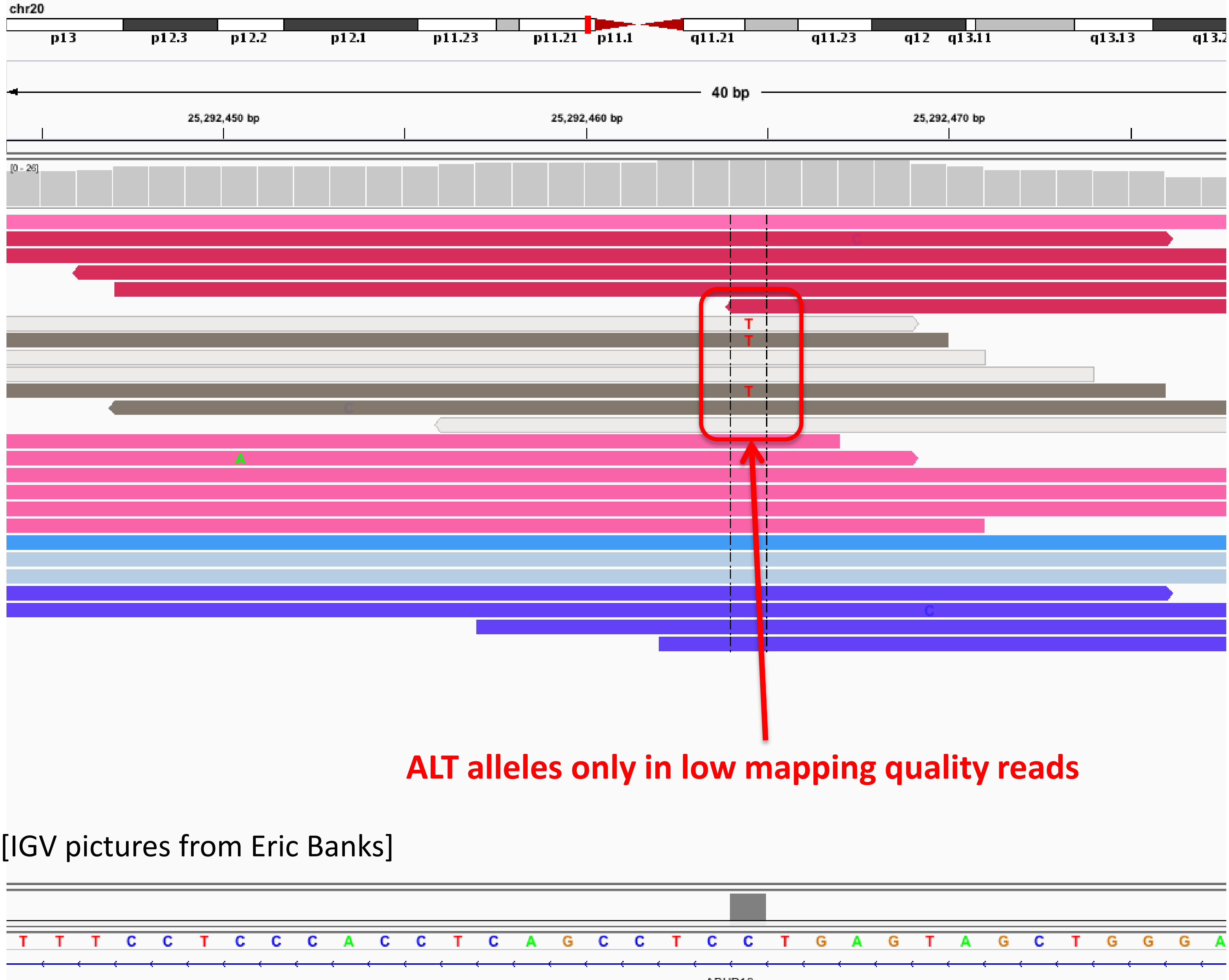
VCF (Variant Call Format)



```
##fileformat=VCFv4.1
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
##INFO=<ID=AVGPOST,Number=1,Type=Float,Description="Average posterior probability from MaCH/Thunder">
##INFO=<ID=RSQ,Number=1,Type=Float,Description="Genotype imputation quality from MaCH/Thunder">
##INFO=<ID=ERATE,Number=1,Type=Float,Description="Per-marker Mutation rate from MaCH/Thunder">
##INFO=<ID=THETA,Number=1,Type=Float,Description="Per-marker Transition rate from MaCH/Thunder">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Count">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total Allele Count">
##ALT=<ID=DEL,Description="Deletion">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/refe
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN">
##INFO=<ID=AMR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AMR based on AC/AN">
##INFO=<ID=ASN_AF,Number=1,Type=Float,Description="Allele Frequency for samples from ASN based on AC/AN">
##INFO=<ID=AFR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AFR based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from EUR based on AC/AN">
##INFO=<ID=VT,Number=1,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=SNPSOURCE,Number=.,Type=String,Description="indicates if a snp was called when analysing the low coverage or exome alignm
##reference=GRCh37
##reference=GRCh37
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00
1       10583   rs58108140      G       A       100     PASS    AVGPOST=0.7707;RSQ=0.4319;LDAF=0.2327;ERATE=0.0161;AN=2184;VT=SNP;AA
1       10611   rs189107123     C       G       100     PASS    AN=2184;THETA=0.0077;VT=SNP;AA=.;AC=41;ERATE=0.0048;SNPSOURCE=LOWCOV
1       13302   rs180734498     C       T       100     PASS    THETA=0.0048;AN=2184;AC=249;VT=SNP;AA=.;RSQ=0.6281;LDAF=0.1573;SNPSO
1       13327   rs144762171     G       C       100     PASS    AVGPOST=0.9698;AN=2184;VT=SNP;AA=.;RSQ=0.6482;AC=59;SNPSOURCE=LOWCOV
1       13957   .       TC      T       28      PASS    AA=TC;AC=35;AN=2184;VT=INDEL;AVGPOST=0.8711;RSQ=0.2501;LDAF=0.0788;THETA=0.0
1       13980   rs151276478     T       C       100     PASS    AN=2184;AC=45;ERATE=0.0034;THETA=0.0139;RSQ=0.3603;LDAF=0.0525;VT=SN
1       30923   rs140337953     G       T       100     PASS    AC=1584;AA=T;AN=2184;RSQ=0.5481;VT=SNP;THETA=0.0162;SNPSOURCE=LOWCOV
:
```

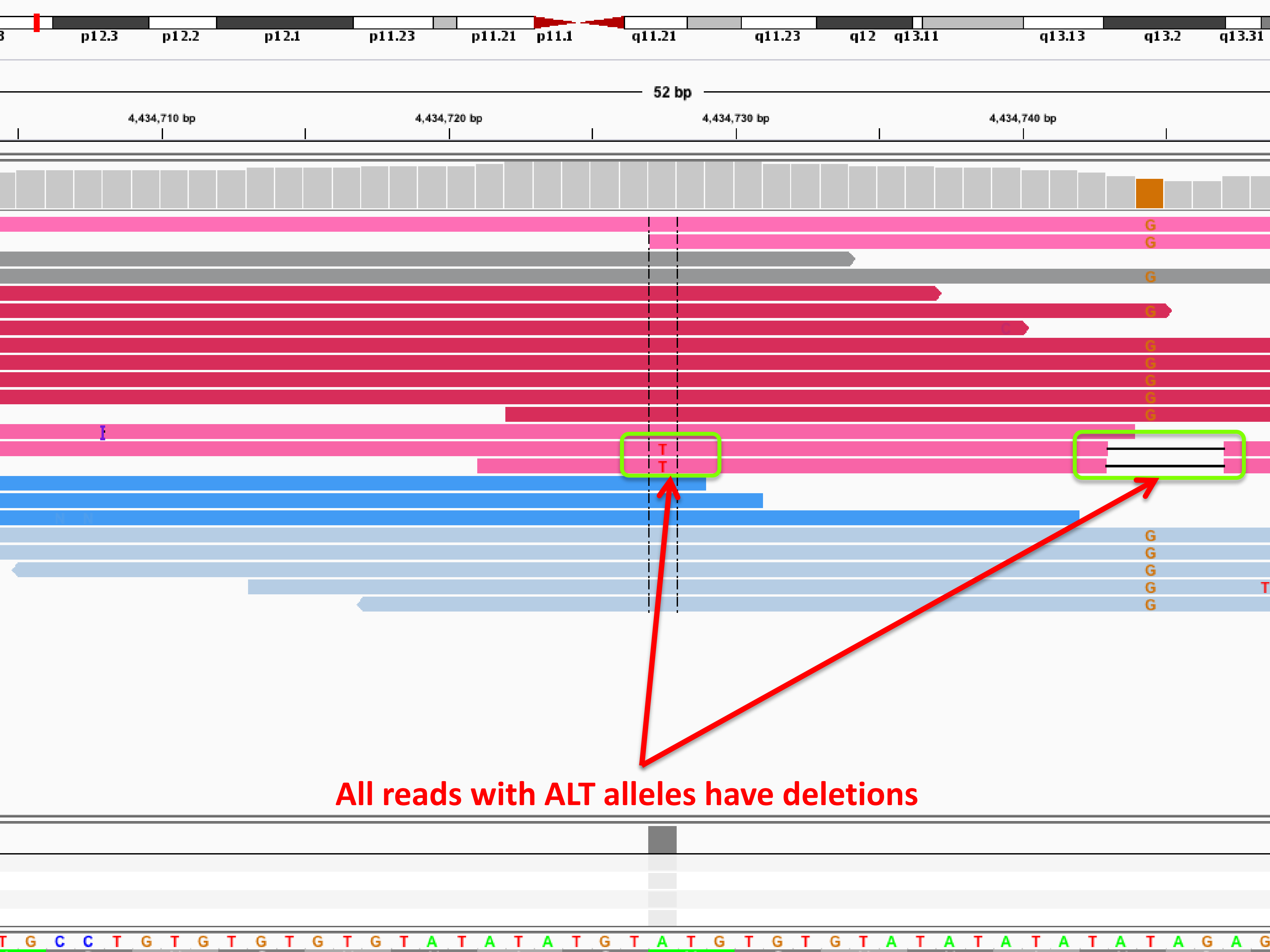
SNP Filtering

- Even with proper modeling of population-based prior, false discoveries do occur
- False discoveries affects the overall quality, not only for the problematic sites but many other sites in LD
- There are many indicators
  - Base read distribution, base quality, mapping quality, …
  - Multi-sample statistics are often more informative



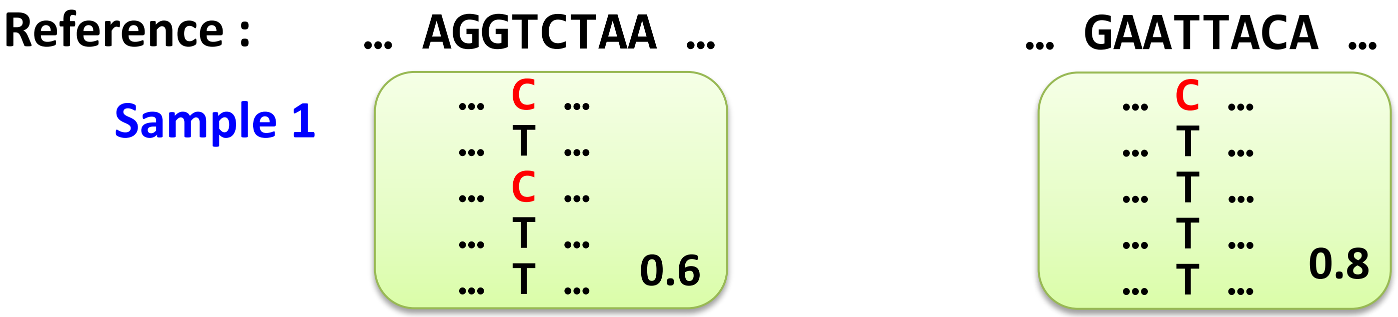
ALT alleles only in low mapping quality reads

[IGV pictures from Eric Banks]



All reads with ALT alleles have deletions

How to Tell Good from Bad: Example



Reference :

… AGGTCTAA …

Sample 1

… C …
… T …
… C …
… T …
… T …

0.6

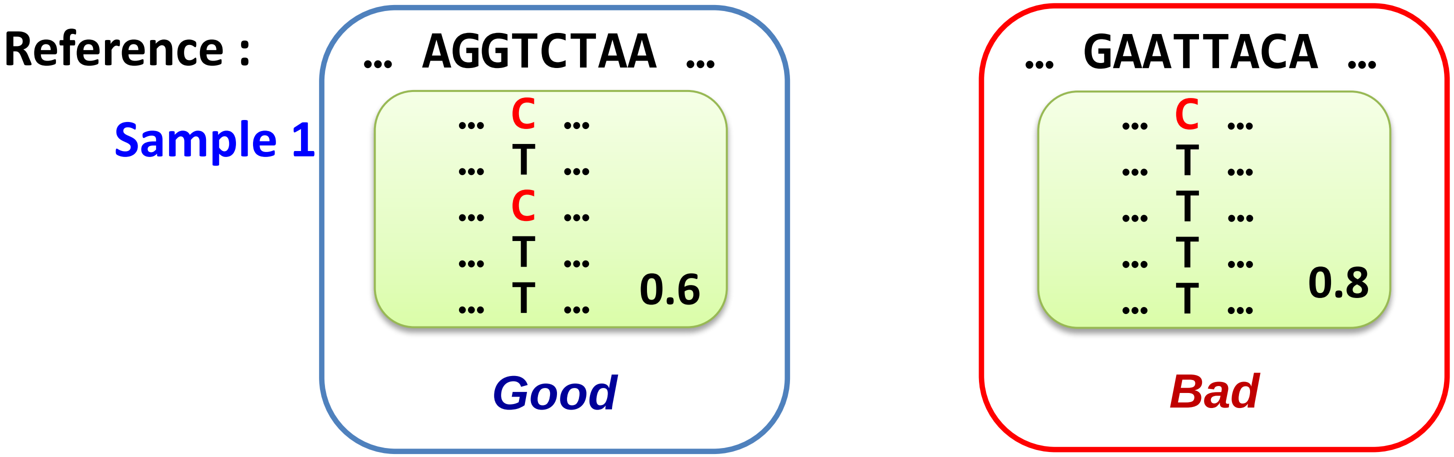
… GAATTACA …

… C …
… T …
… T …
… T …
… T …

0.8

We expect 50:50 read distribution for HET sites

How to Tell Good from Bad: Example



Reference :

Sample 1

… AGGTCTAA …

… C …
… T …
… C …
… T …
… T … 0.6

Good

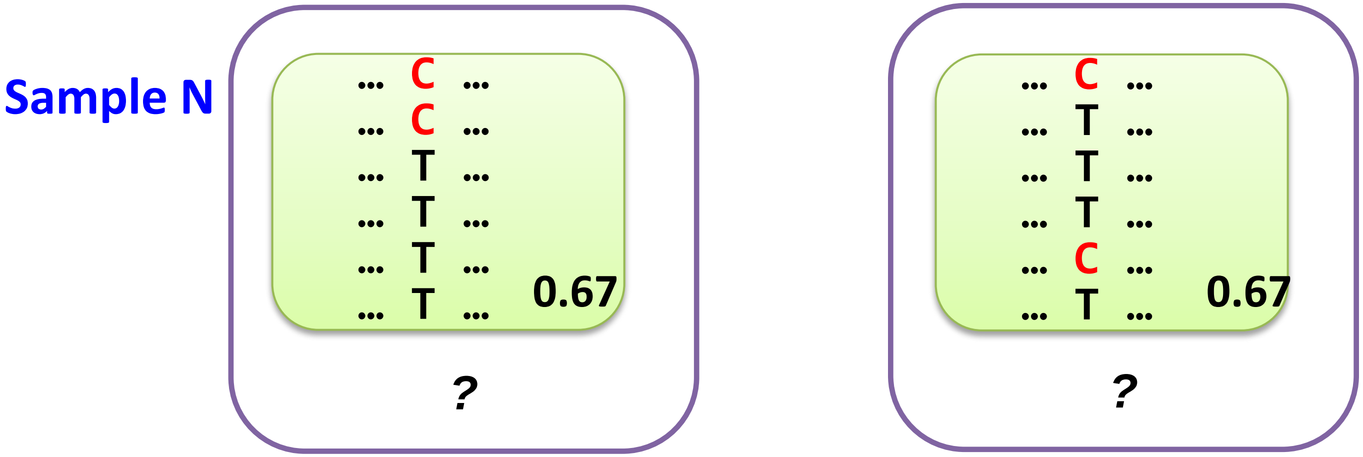
… GAATTACA …

… 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

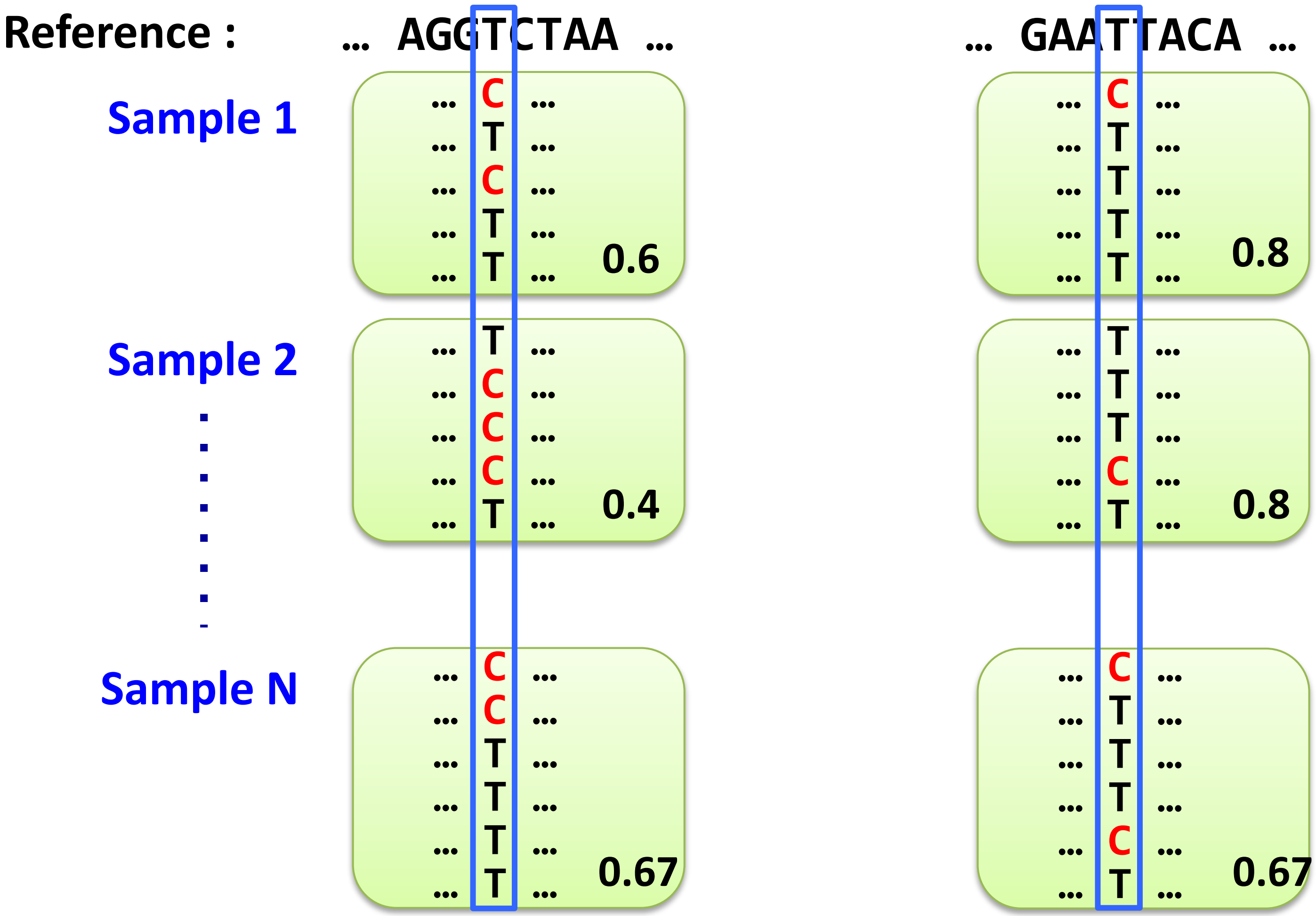
... C ...
... C ...
... T ...
... T ...
... T ...
... T ... 0.67

?

... C ...
... T ...
... T ...
... T ...
... C ...
... T ... 0.67

?

Multi-sample Filtering is Informative



Reference : … AGGTCTAA … … GAATTACA …

Sample 1

… C … … T … … C … … T … … T … 0.6

… C … … T … … T … … T … … T … 0.8

Sample 2

… T … … C … … C … … C … … T … 0.4

… T … … T … … T … … C … … T … 0.8

Sample N

… C … … C … … T … … T … … T … … T … 0.67

… C … … T … … T … … T … … C … … T … 0.67

Overall Balance: 0.56

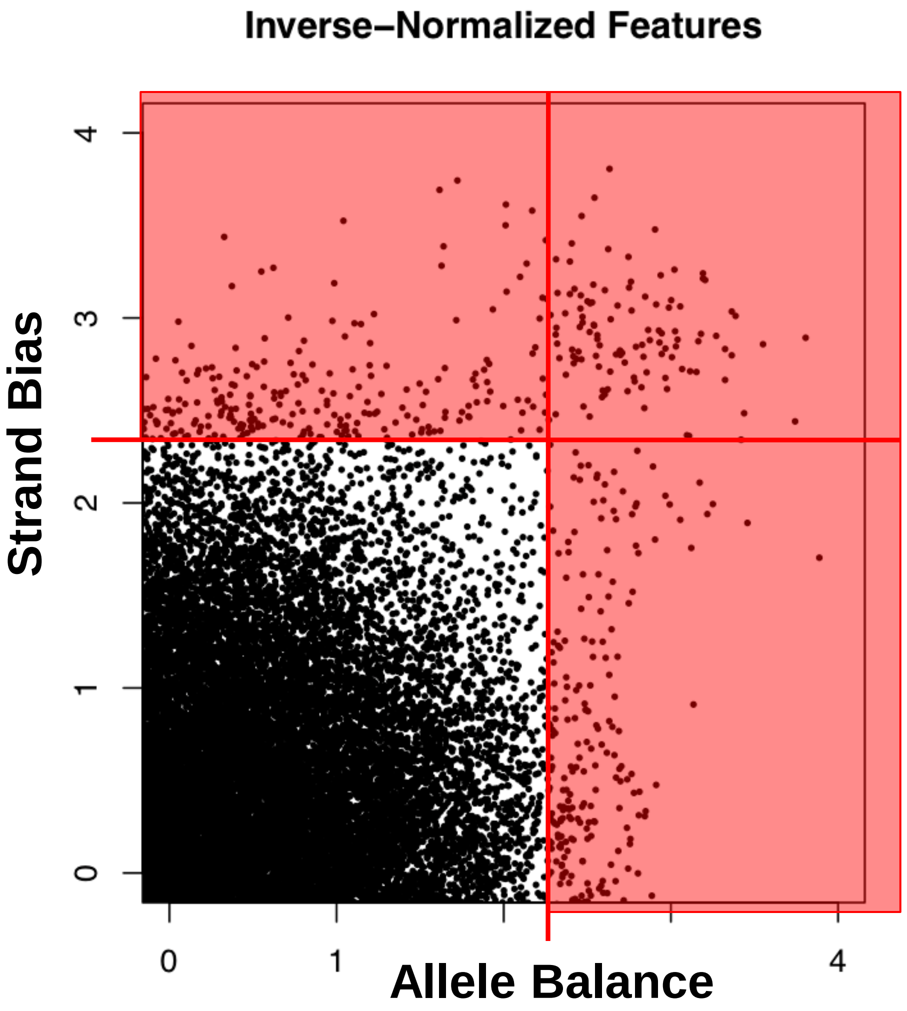
Overall Balance: 0.75

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…

Hard Filtering by Individual Thresholds

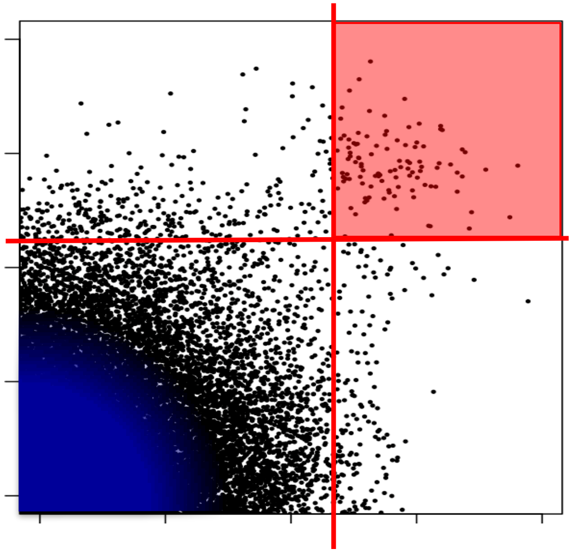
- Problems
  - False negative increases with number of filters
  - Too many knobs to turn (thresholds)



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

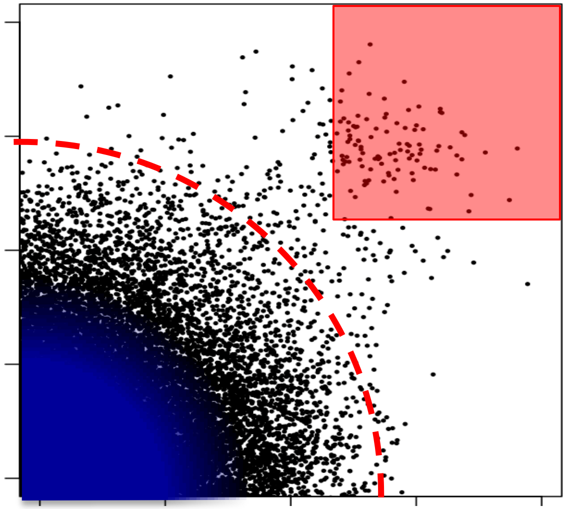
- Training
  - Positive examples
    - Known polymorphic sites
  - Negative examples
    - Filtered out by multiple hard filters
  - Input
    - All individual features collected for each site



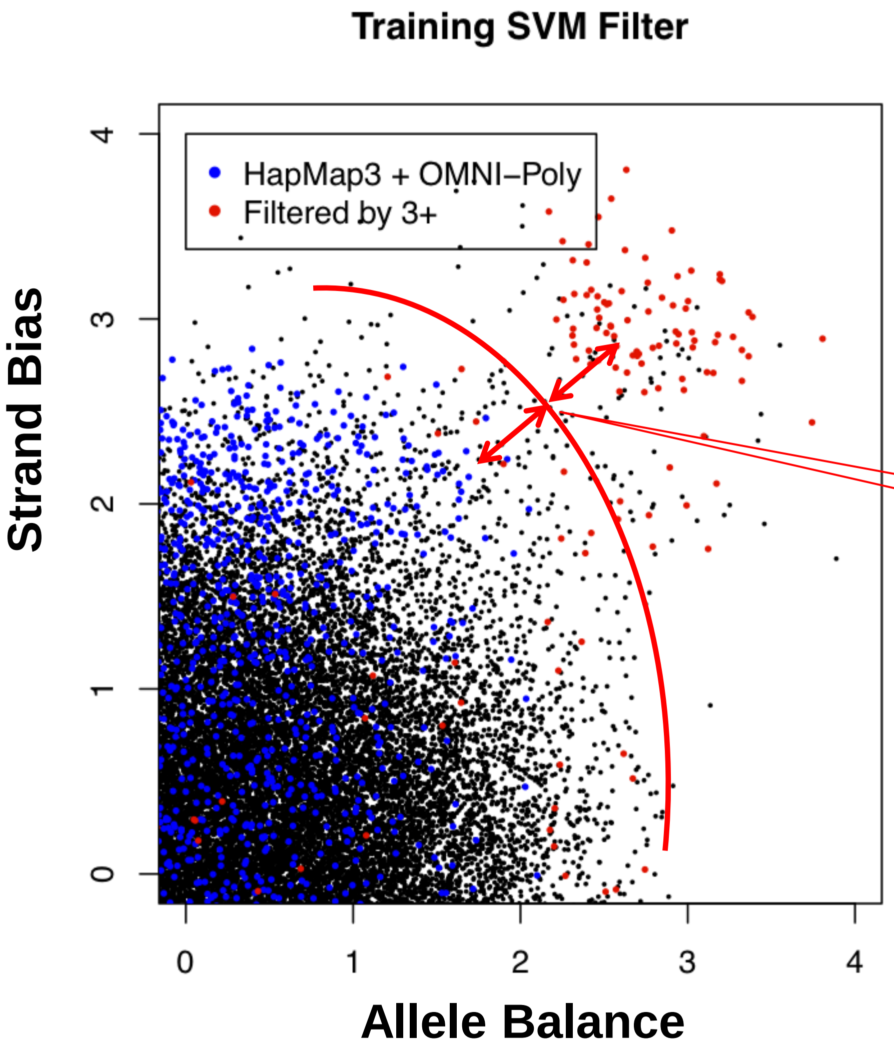
Filtering by Supervised Learning

- 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

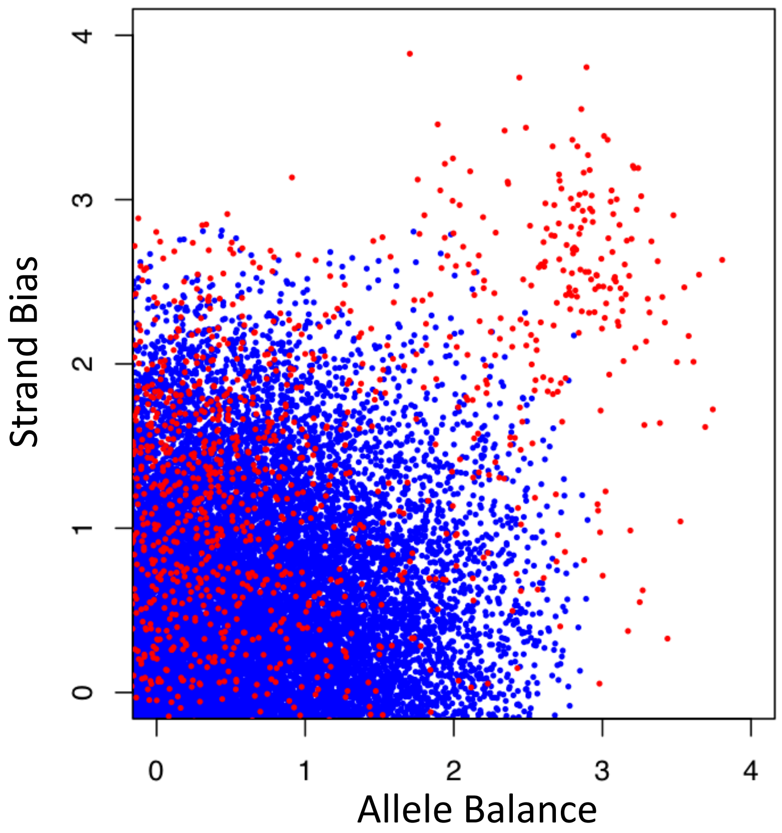


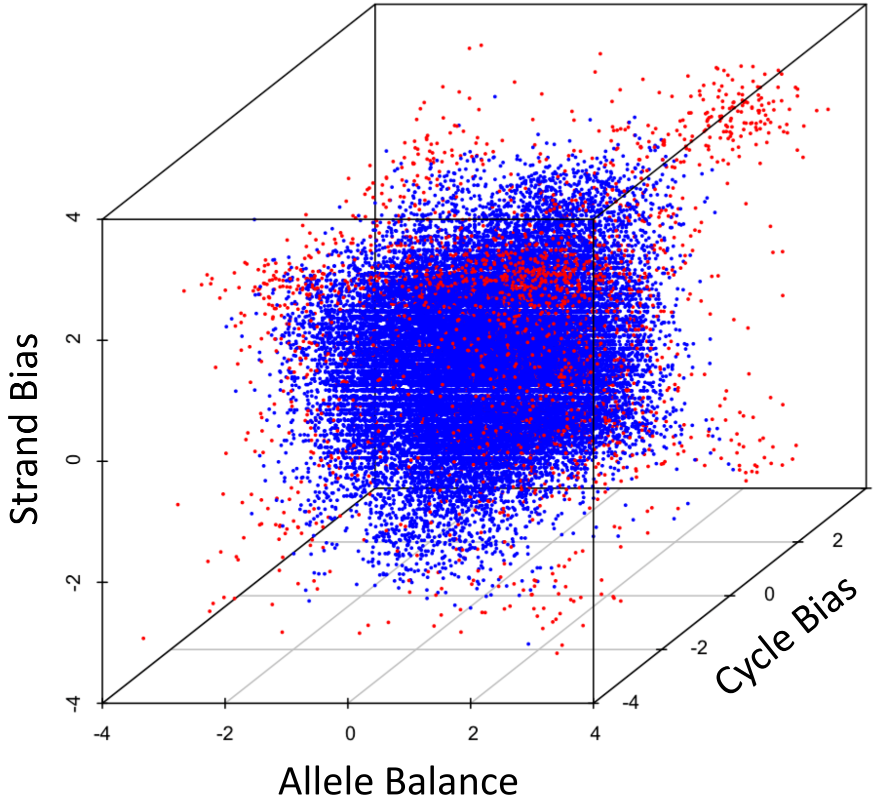
Maximize margin

- Positive example
- Negative example

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

SVM Output in Multi-dimensional Space

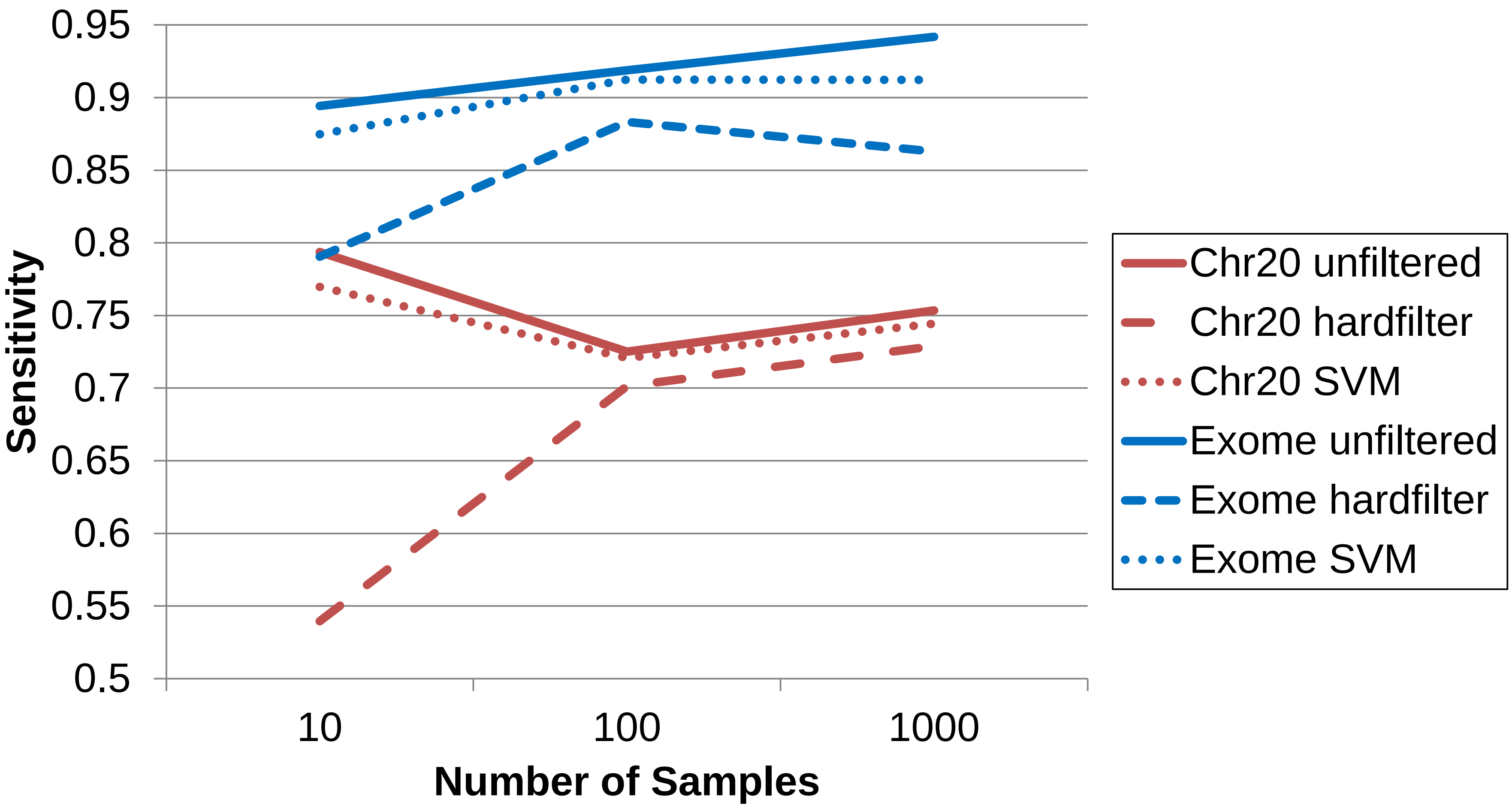




- 🟦 Filter PASS
- 🟥 Filter FAIL

Most of FAIL SNPs are outliers in higher-dimensional view

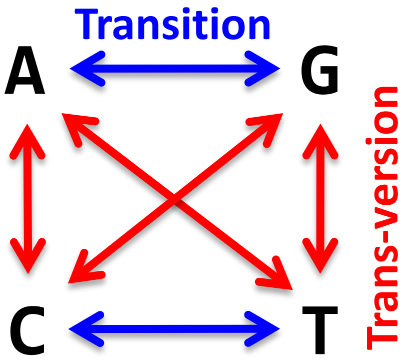
Improved Sensitivity by SVM



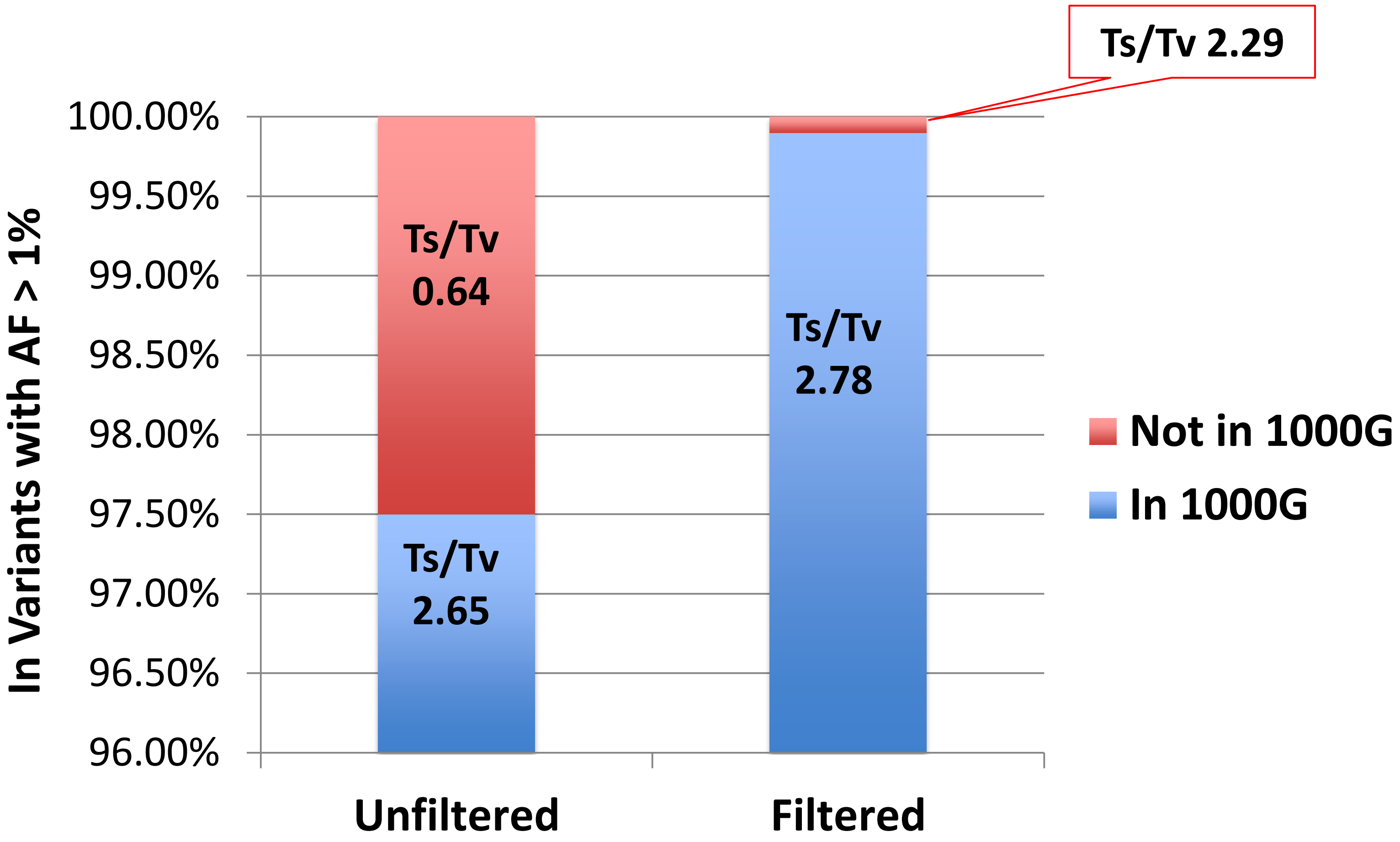
Evaluation of SNP Callsets

- Sensitivity on known SNP data
  - dbSNP, HapMap, 1000G, etc.

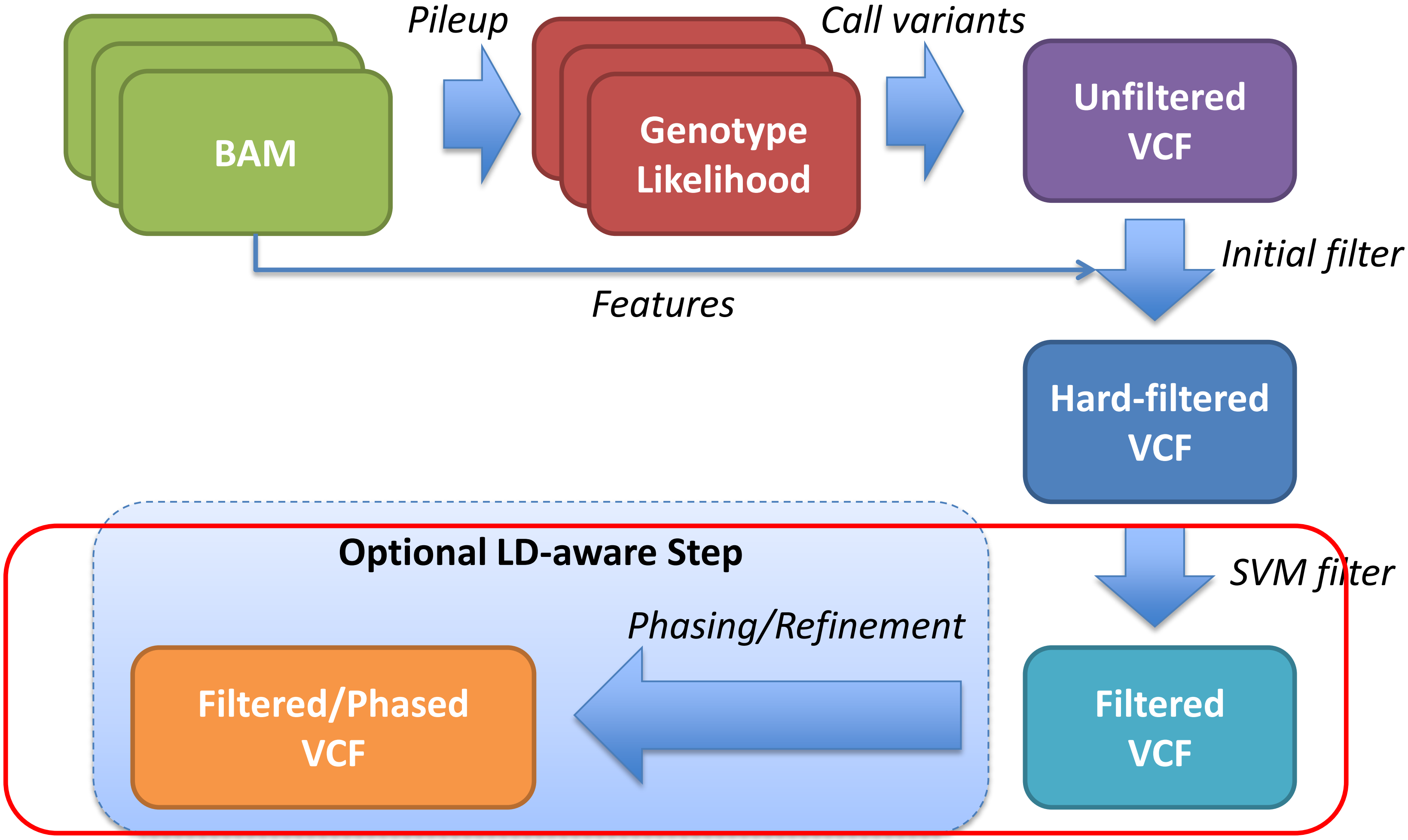
- Transition to transversion ratio (Ts/Tv)
  - Transition is easier to occur.
  - Typical Ts/Tv values
    - Whole genome: 2.2~2.4
    - Whole exome: 2.7~3.0



Results: Exome Sequencing Project (GO-ESP)



LD-aware Genotype Refinement


BAM

Pileup

Genotype Likelihood

Call variants

Unfiltered VCF

Initial filter

Features

Hard-filtered VCF

SVM filter

Filtered VCF

Optional LD-aware Step

Phasing/Refinement

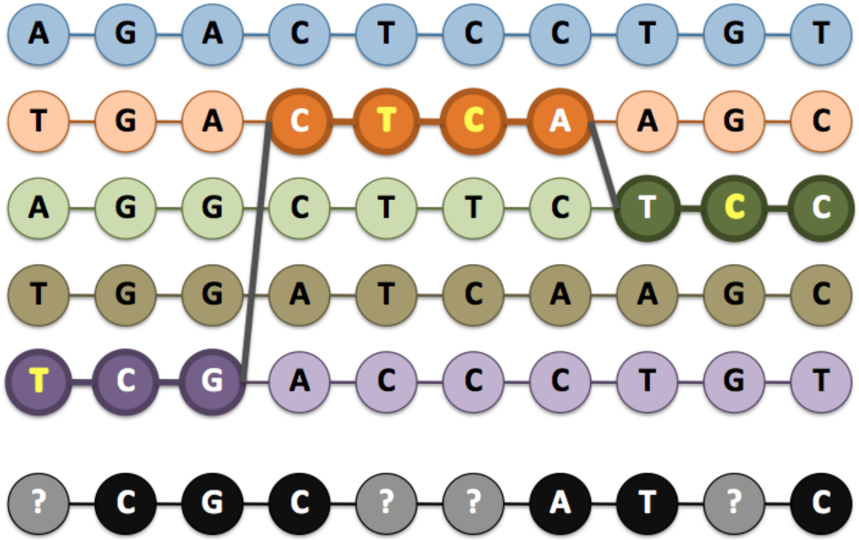
Filtered/Phased VCF

Sequence Based Genotype Calls - Haplotypes

- **Individual Based Prior**
- **Population Based Prior**
- **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

Haplotype-aware Genotype Refinement

- People share 'blocks' of genotypes
- Haplotype-phasing improves genotype accuracy by correcting unlikely genotypes and filling in missing genotypes
- gotCloud takes two-steps
  - Beagle (Step 1)
  - ThunderVCF (Step 2)



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 .
```

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

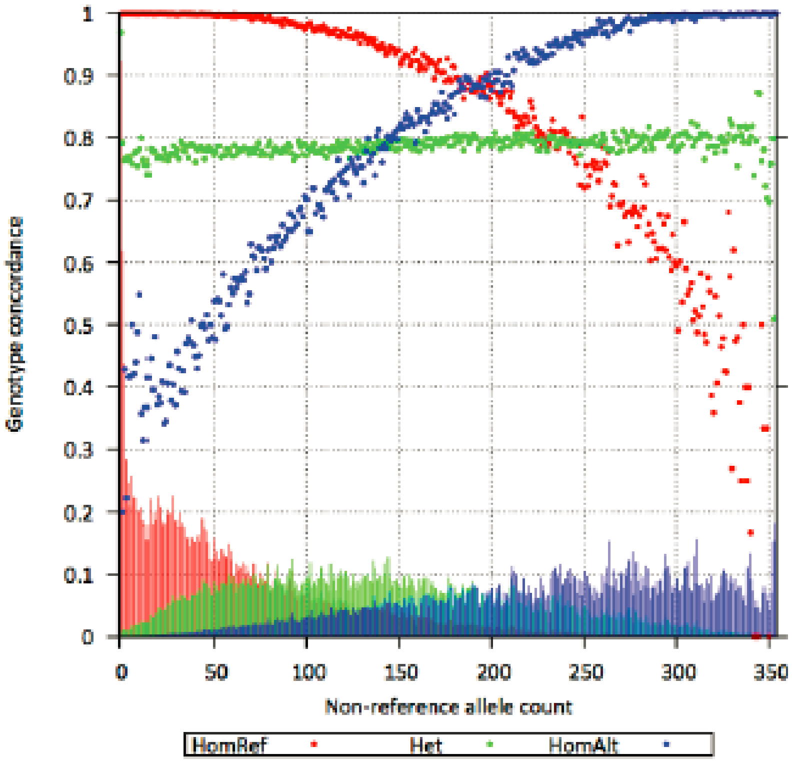
Does Haplotype Information Really Help?

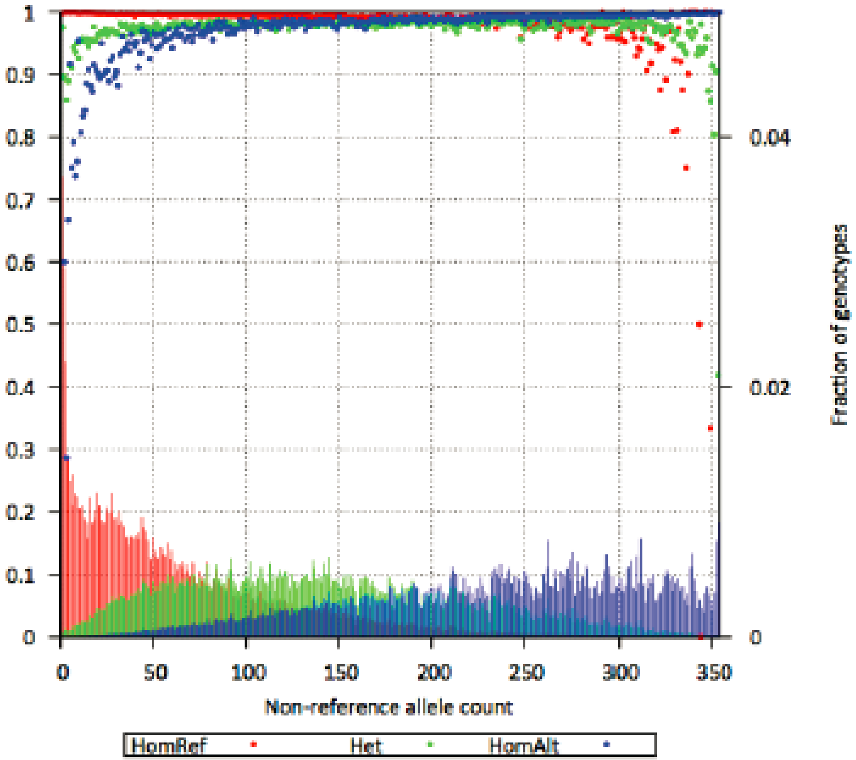
Single Site Analysis

– 21.4% HET errors

Haplotype Aware Analysis

– 2.0% HET errors



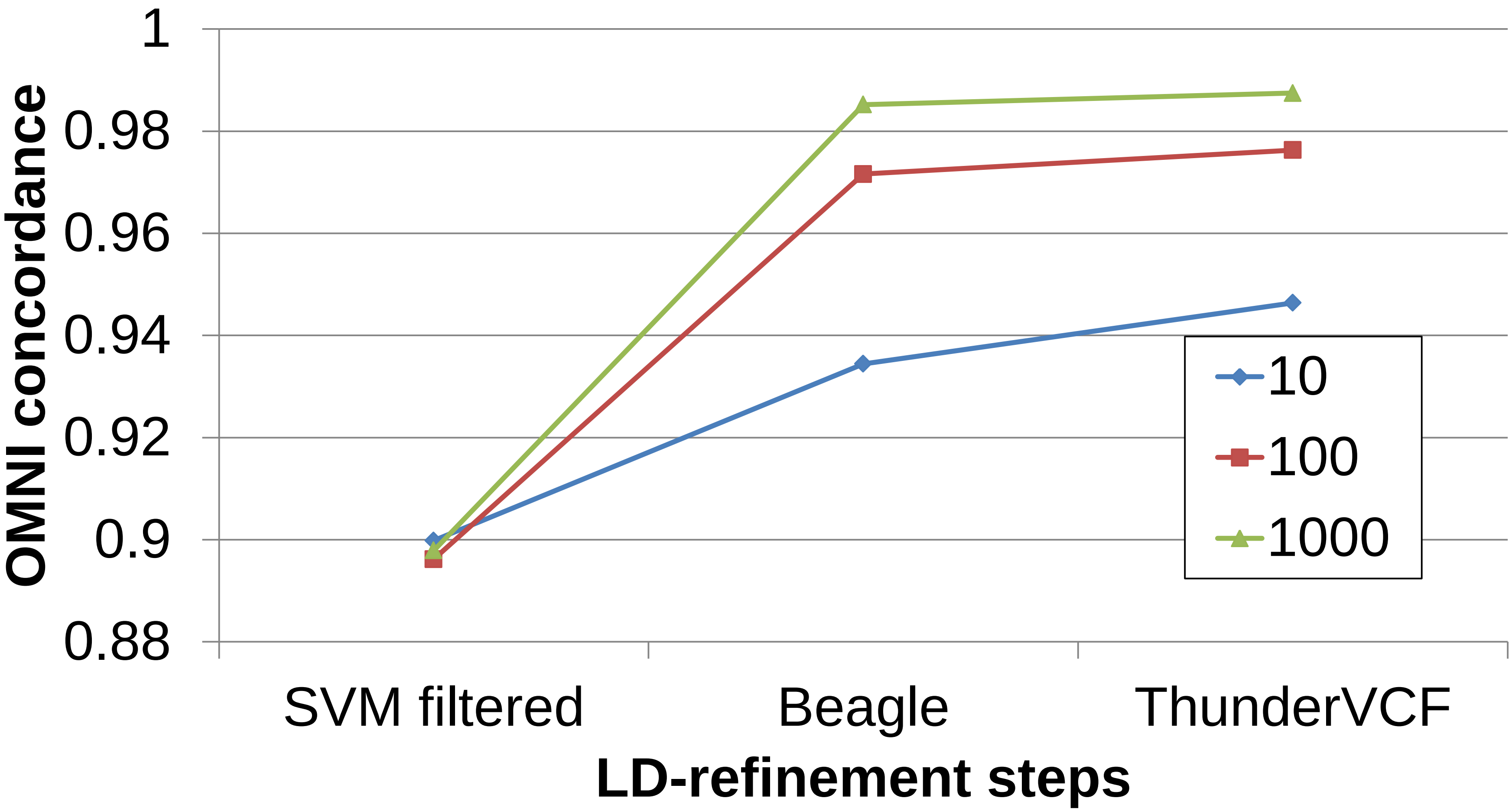


Low-pass Sequencing Improves with More Samples

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 at the University of Michigan

Low-pass Sequencing Improves with More Samples



Quality of 1000G Phase 1 Genotypes

TYPE	EVAL	N	#Variants (Overlap)	HOMREF (EVAL)	HET (EVAL)	HOMALT (EVAL)	OVER-ALL
SNP	Omni2.5	1,092	2.1M	99.87%	**99.09%**	99.35%	99.65%
SNP	CGI	34	13M	99.87%	**98.63%**	98.75%	99.60%
INDEL	CGI	34	820k	98.69%	**95.64%**	96.35%	98.01%
SV	Conrad	248	1.1k	99.92%	**99.01%**	99.47%	99.82%

- Genotype likelihood adjusting for individual BAM's bias statistic reduces ~10% of non-ref genotype discordance

- MaCH/Thunder refinement starting with beagle haplotypes provide an additional ~15% reduction.

Low-pass Sequencing with Many Samples

- For a given budget, should we sequence deeper or sequence more?

- 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 co mplex disease association studies.

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 1,000 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

Sequencing Study Design - Considerations

- Sequencing Depth
  - Improved throughput enables more samples with moderate (~10x) coverage at reasonable costs
- Whole genome vs Whole Exome vs Targeted Genes
- Sequence + Array
  - Which samples to be sequenced?

Suggested Resources

- Michigan Mapping/Variant calling pipeline on the cloud
  - http://genome.sph.umich.edu/wiki/GotCloud
- 1000 Genomes Project http://http://www.1000genomes.org/
  - Includes sequence data, variant genotypes, and many more
- VCF and other file formats: https://github.com/samtools/hts-specs