Why Whole Genome Sequencing Methods Differ

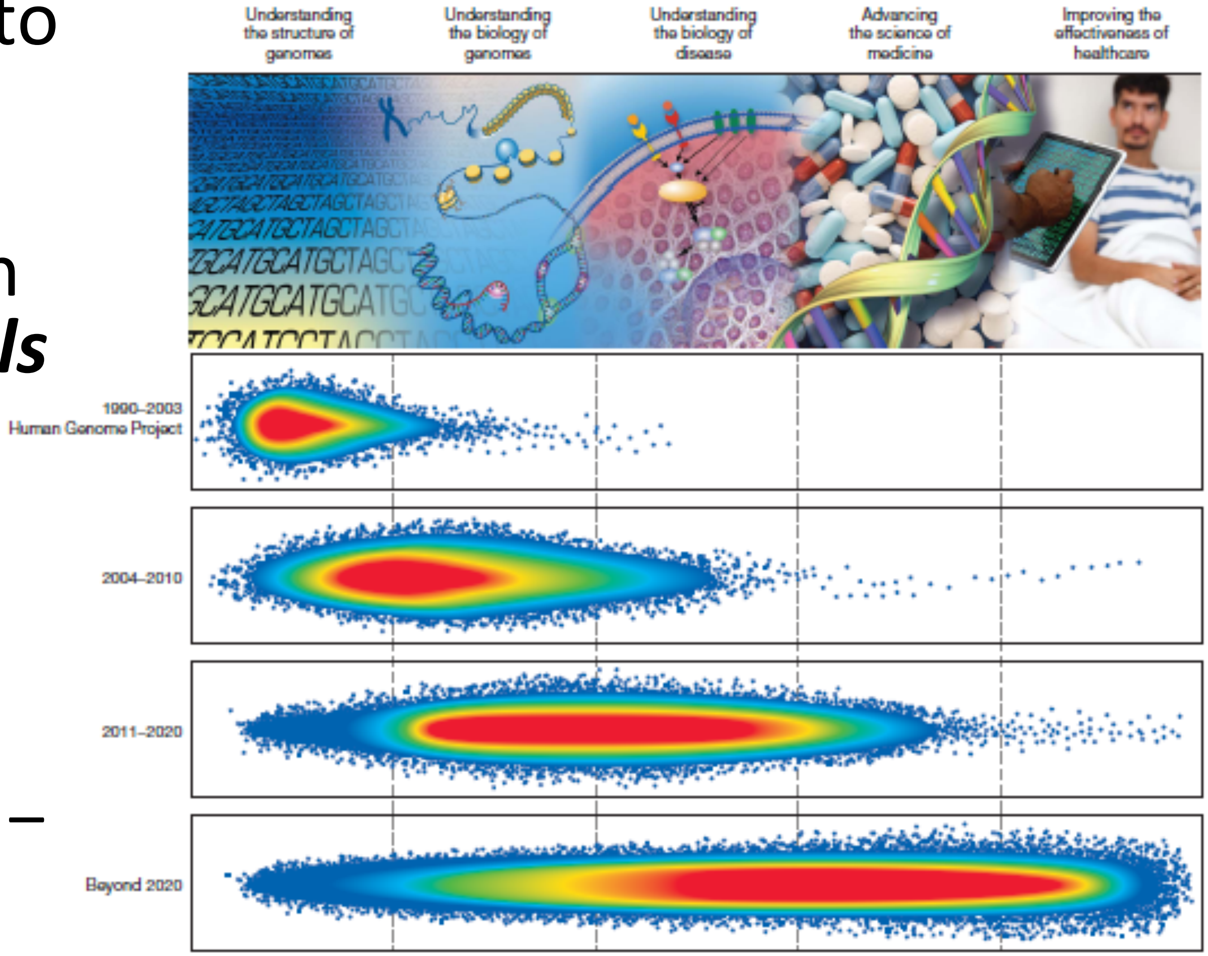
Justin Zook and Marc Salit

National Institute of Standards and Technology

June 6, 2012

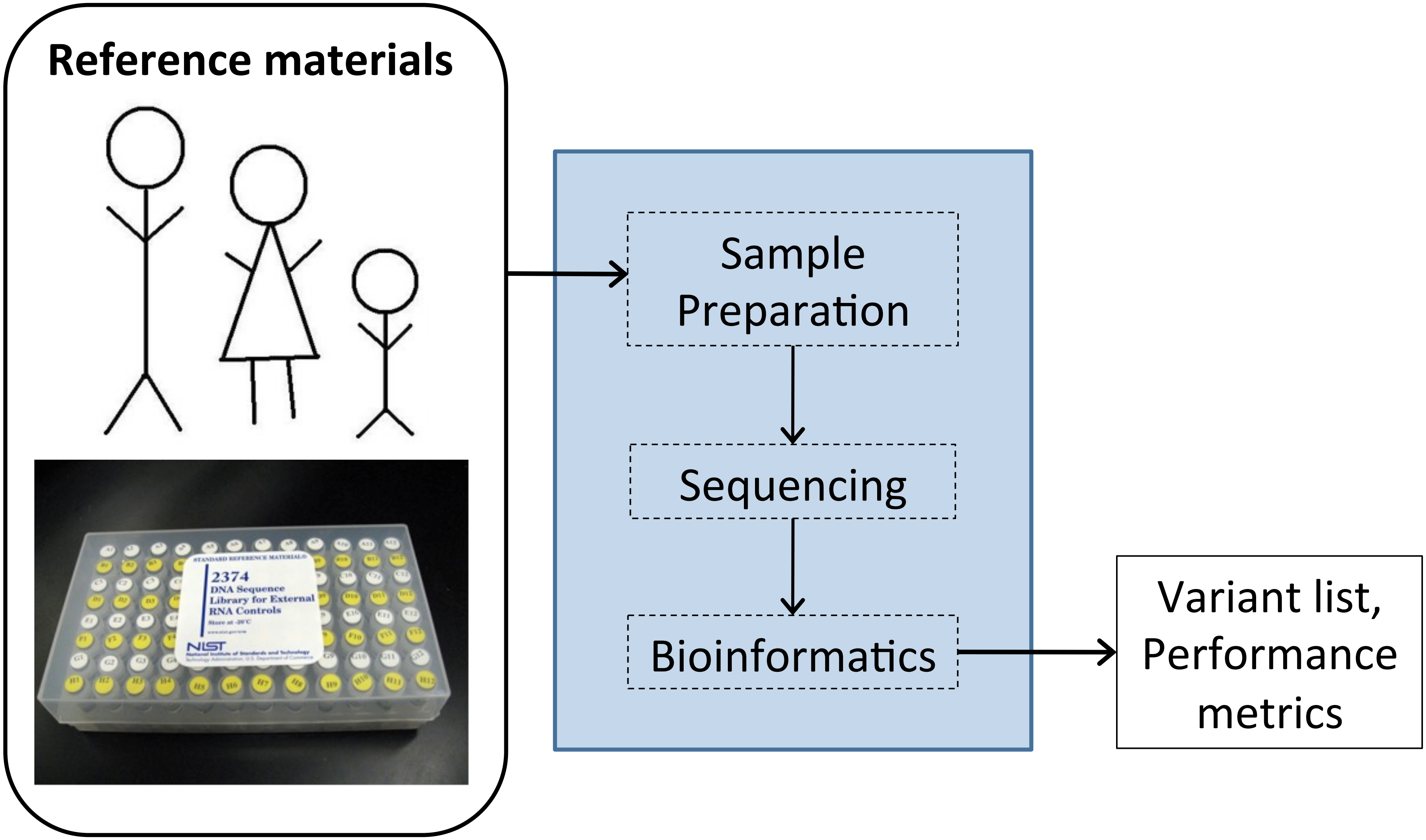
"Genome in a Bottle"

- Help enable translation of NGS to regulated clinical applications
- Select and maintain ***Reference Materials***
  - From a single, internationally-recognized source
  - Stable
  - Homogeneous
  - Well-characterized – towards "perfect" human genomes



E. Green et al. *Nature* (2011) 470: 204

Vision for NIST Genomic RMs



Reference materials

Sample Preparation

Sequencing

Bioinformatics

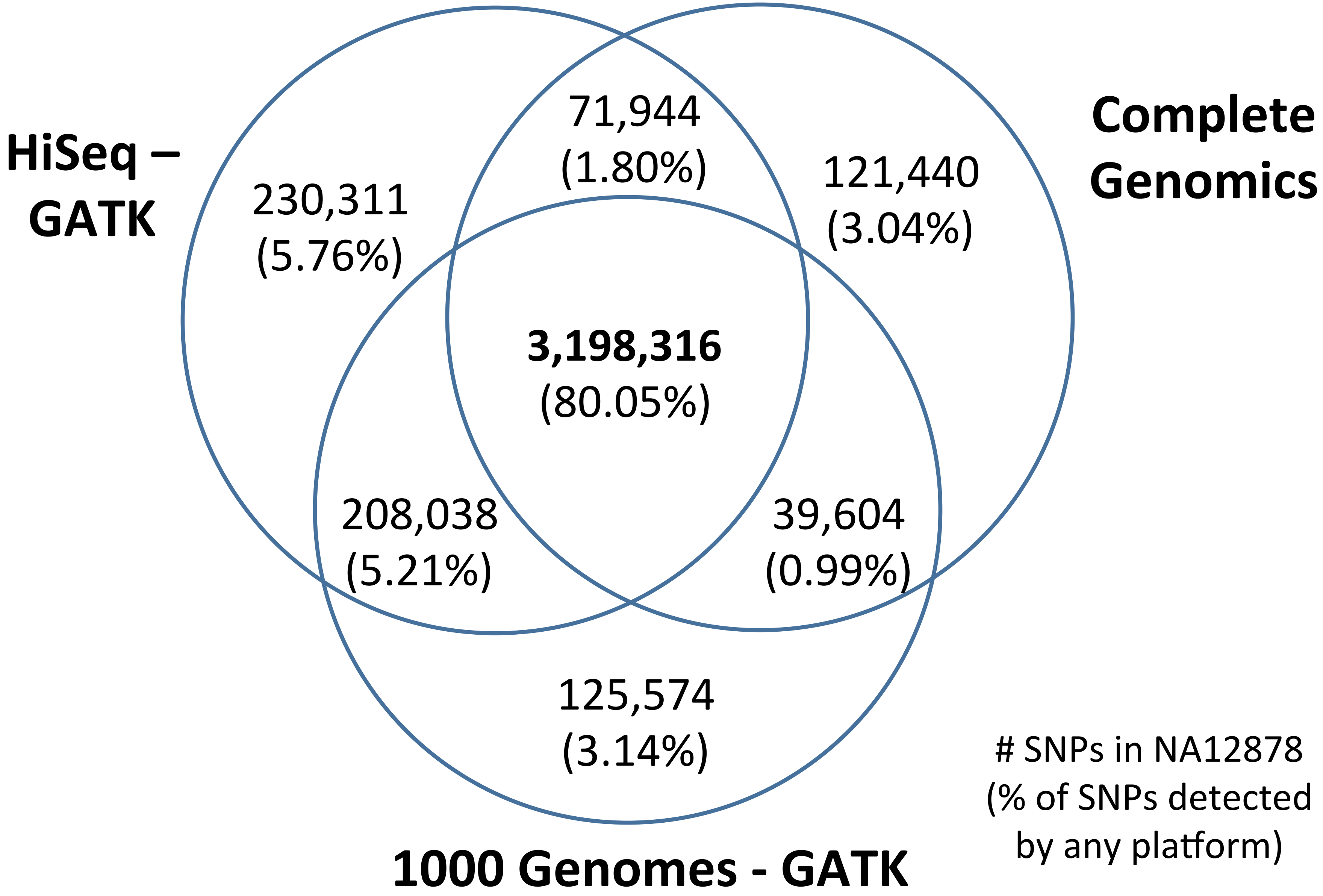
Variant list, Performance metrics

STANDARD REFERENCE MATERIAL®
2374
DNA Sequence
Library for External
RNA Controls
Store at -20°C
NIST
National Institute of Standards and Technology

NIST Genomic Reference Materials

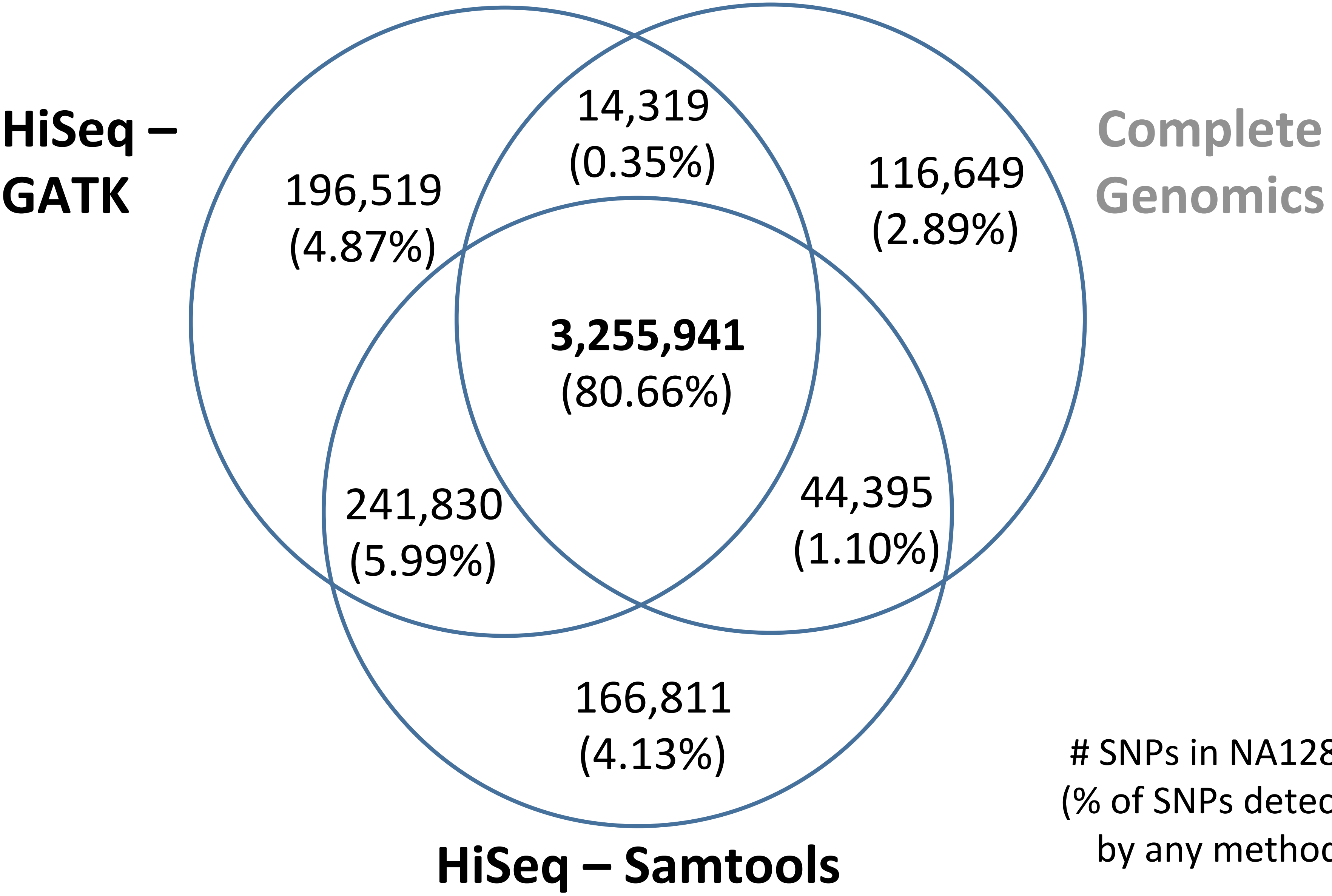
- Reference Material vs. Reference Genome or Reference Assembly
- Understanding uncertainty from bias is essential for Standard Reference Material characterization
- Comparison of SNPs in multiple datasets on a prospective Reference Material (NA12878)
- Integrating datasets to form consensus calls
- Utility of Reference Materials for understanding performance and bias

Whole genome sequencing technologies disagree about 100,000's of SNPs



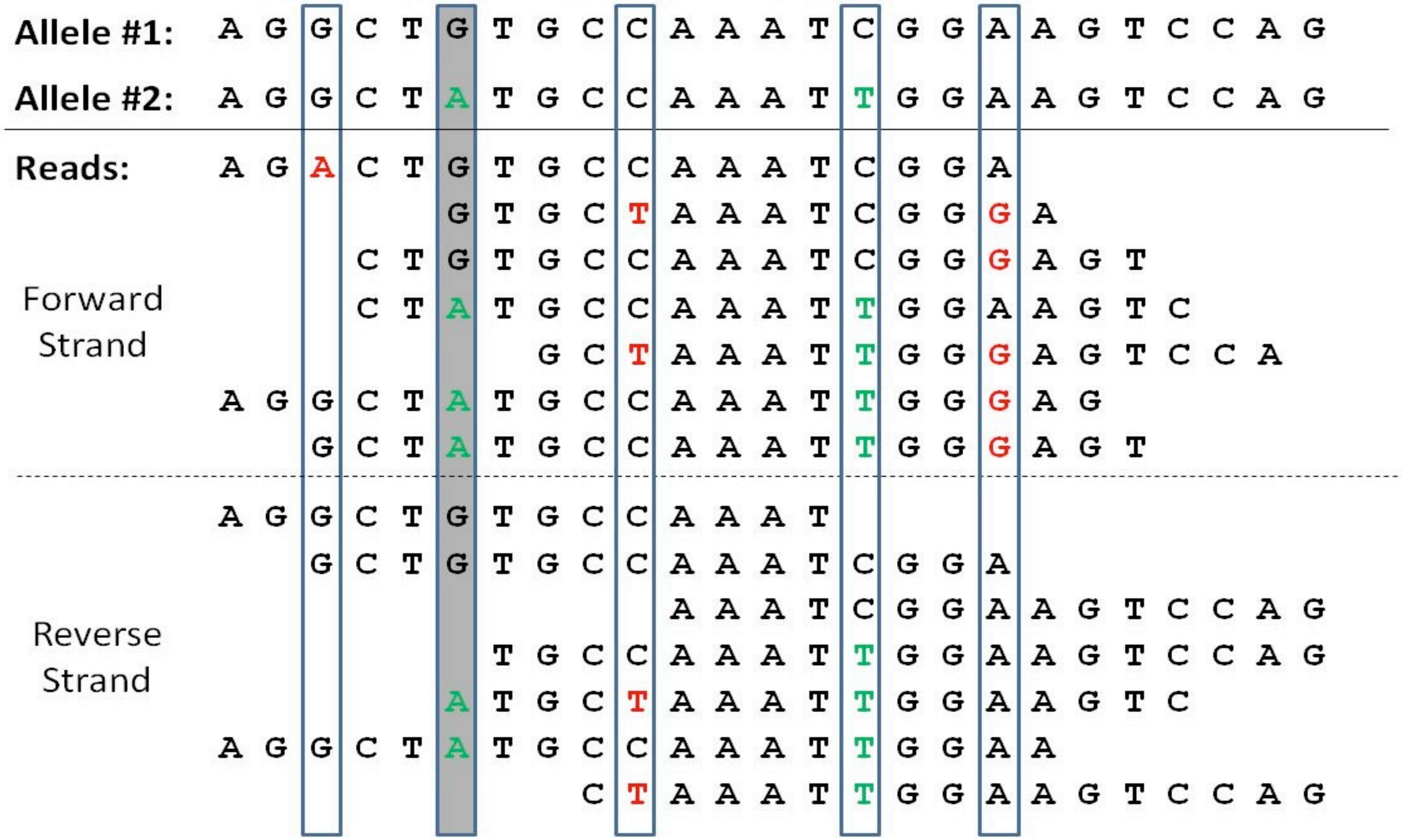
# SNPs in NA12878 (% of SNPs detected by any platform)

Different bioinformatics algorithms also disagree about 100,000's of SNPs



# SNPs in NA12878
(% of SNPs detected by any method)

Identifying characteristics of calls



```
Allele #1:  A G G C T G T G C C A A A T C G G A A G T C C A G
Allele #2:  A G G C T A T G C C A A A T T G G A A G T C C A G
-------------------------------------------------------------
Reads:      A G A C T G T G C C A A A T C G G A
                      G T G C T A A A T C G G G A
                  C T G T G C C A A A T C G G G A G T
Forward           C T A T G C C A A A T T G G A A G T C
Strand                    G C T A A A T T G G G A G T C C A
            A G G C T A T G C C A A A T T G G G A G
                G C T A T G C C A A A T T G G G A G T
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
            A G G C T G T G C C A A A T
                G C T G T G C C A A A T C G G A
                                A A A T C G G A A G T C C A G
Reverse                 T G C C A A A T T G G A A G T C C A G
Strand                A T G C T A A A T T G G A A G T C
            A G G C T A T G C C A A A T T G G A A
                            C T A A A T T G G A A G T C C A G
```

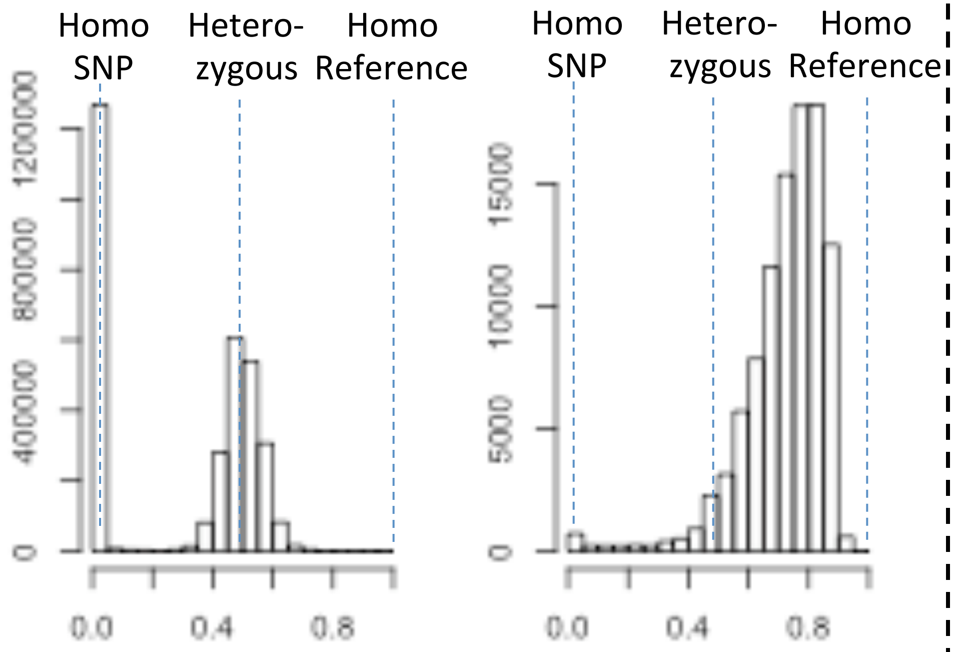
Allele Balance: 1 0.83 0.5 …

↑

Strand bias

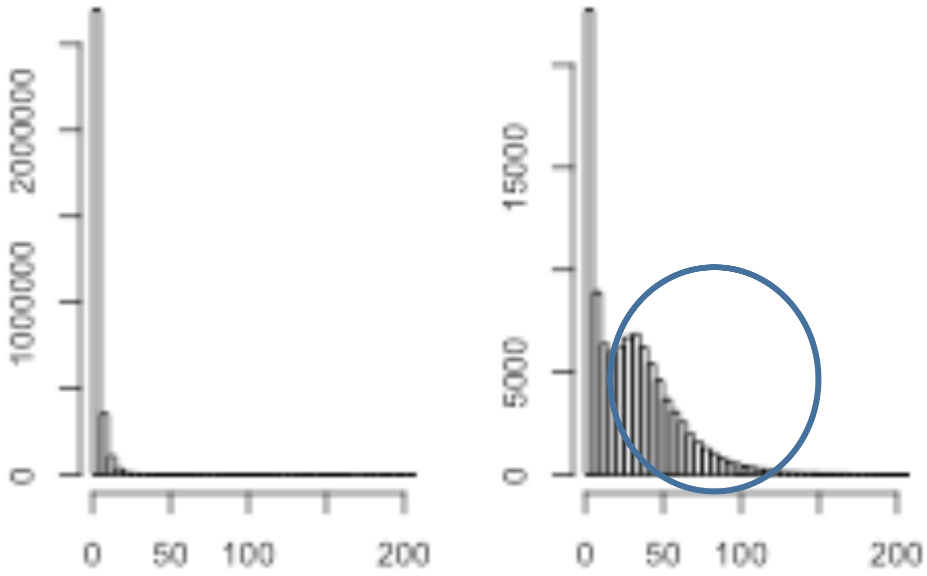
Some false positives have distinctive characteristics

True Positives **False Positives**



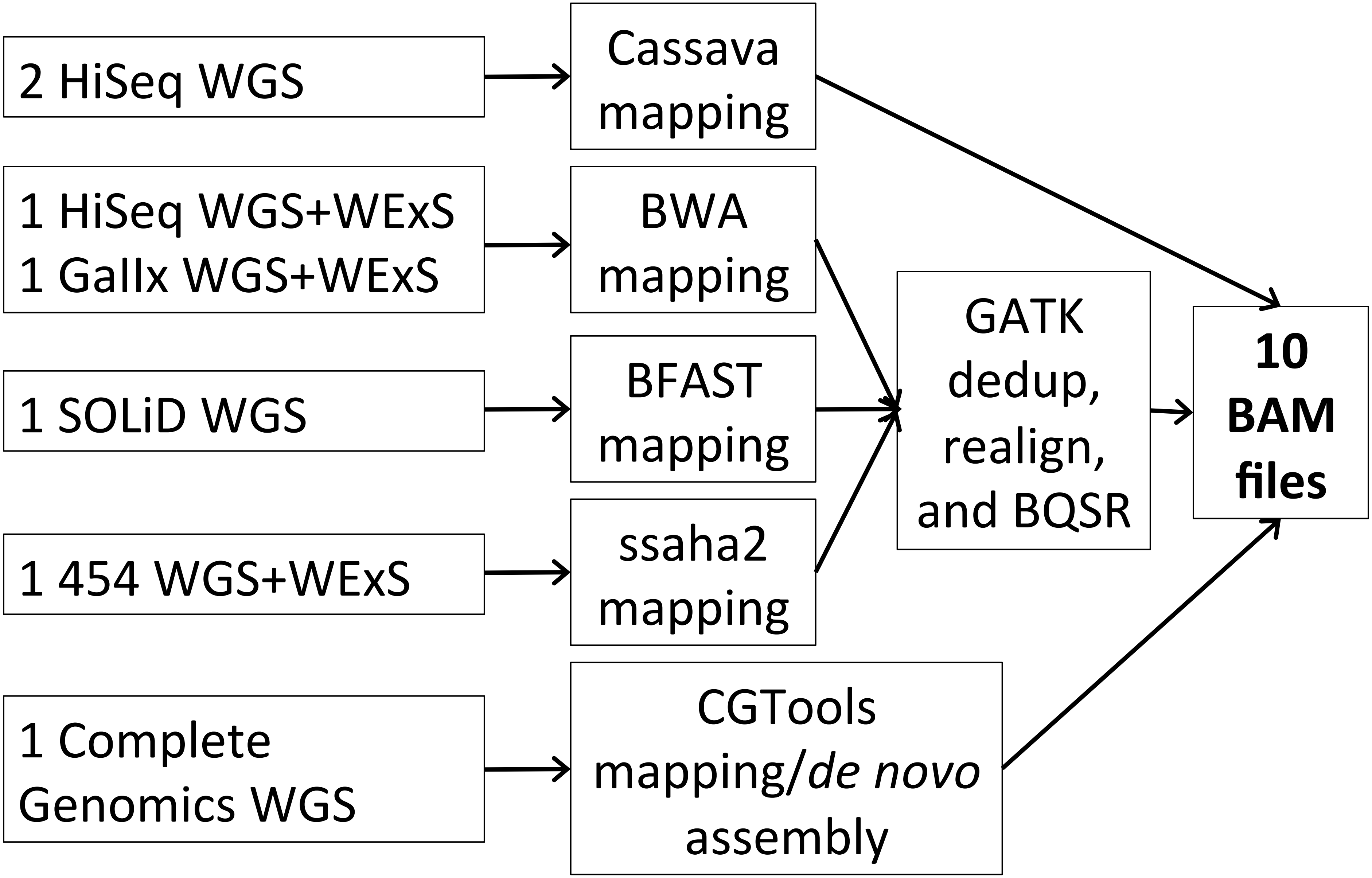
Allele Balance

True Positives **False Positives**



Strand bias test

10 datasets used for NA12878 genome



2 HiSeq WGS

1 HiSeq WGS+WExS
1 GaIIx WGS+WExS

1 SOLiD WGS

1 454 WGS+WExS

1 Complete
Genomics WGS

Cassava
mapping

BWA
mapping

BFAST
mapping

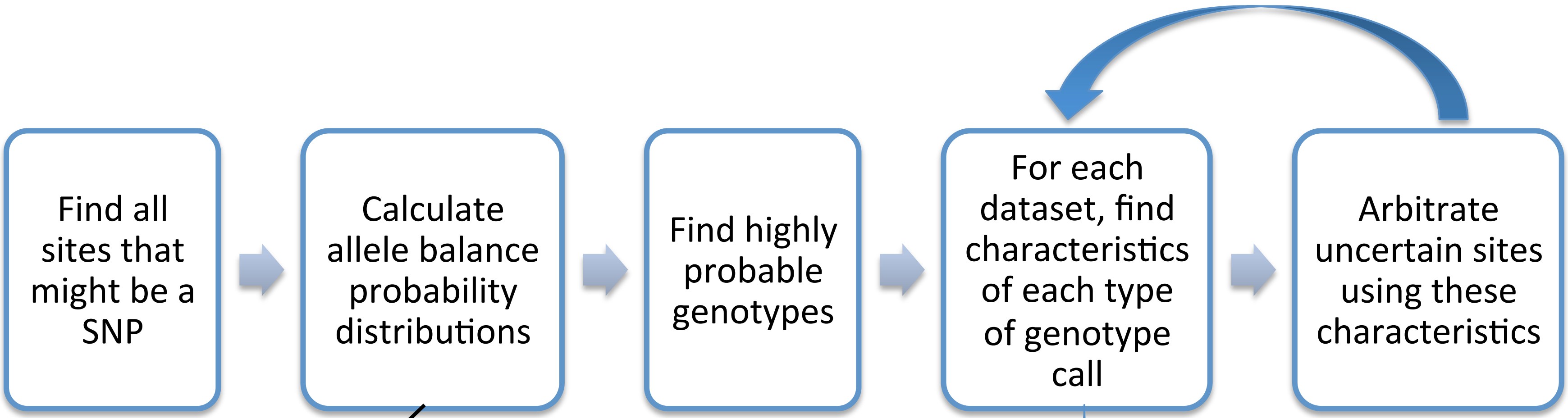
ssaha2
mapping

CGTools
mapping/*de novo*
assembly

GATK
dedup,
realign,
and BQSR

**10
BAM
files**

Using characteristics of reliable calls to arbitrate between discordant calls



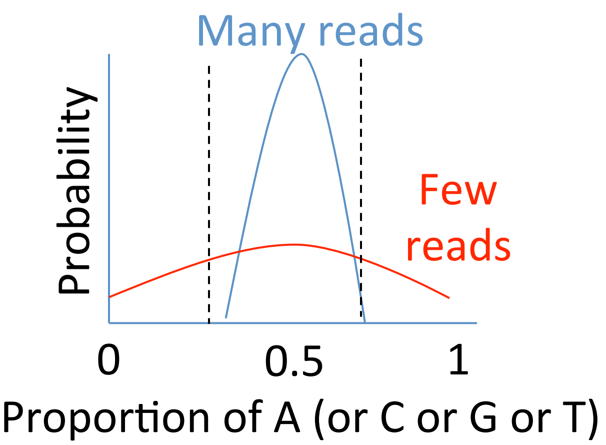
Find all sites that might be a SNP

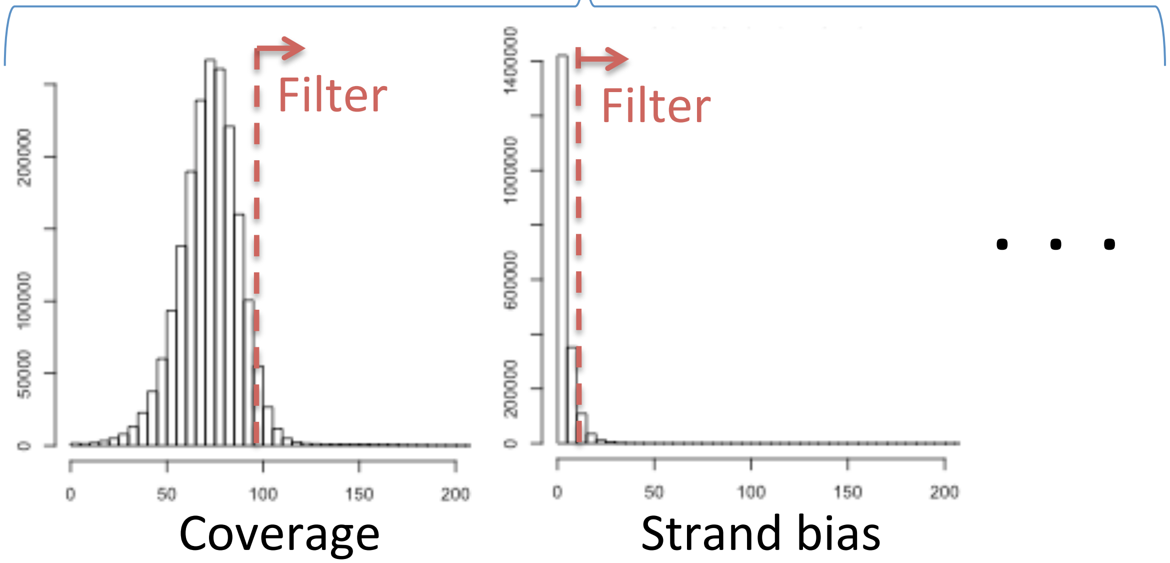
Calculate allele balance probability distributions

Find highly probable genotypes

For each dataset, find characteristics of each type of genotype call

Arbitrate uncertain sites using these characteristics





Consensus Genotype Calling

Step	Homozygous Reference	Heterozygous	Homozygous Variant	Uncertain
All possible SNP locations	-	-	-	9,367,080
Highly confident set	4,235,734	1,891,778	1,116,083	2,123,485
After 1st arbitration	293,367	107,163	40,769	1,669,184
After allele balance arbitration	806,344	0	211,471	651,369
After voting	199,983	3,705	30,579	428,659
Total	**5,535,428**	**2,002,646**	**1,398,902**	**428,659**

Performance Metrics: Algorithm Comparison

HiSeq – **Samtools**

	Integrated Consensus Genotypes			
	Homozygous Reference	**Heterozygous**	**Homozygous Variant**	**Uncertain**
Homozygous Reference/ No Call	5.44M	**FNs** *69.2k (1.81%)*	*47.2k (1.23%)*	228k (5.95%)
Heterozygous	**FPs** *90.3k (2.36%)*	1.93M (50.4%)	*2199 (0.06%)*	157k (4.10%)
Homozygous Variant	*9990 (0.26%)*	*3714 (0.10%)*	1.35M (35.2%)	42.0k (1.10%)

HiSeq – **GATK**

	Integrated Consensus Genotypes			
	Homozygous Reference	**Heterozygous**	**Homozygous Variant**	**Uncertain**
Homozygous Reference/ No Call	5.53M	**FNs** *181k (4.73%)*	*153k (3.99%)*	329k (8.58%)
Heterozygous	**FPs** *6094 (0.18%)*	1.82M (47.5%)	*317 (0.01%)*	85.9k (2.24%)
Homozygous Variant	*1934 (0.05%)*	*401 (0.01%)*	1.25M (32.5%)	13.8k (0.36%)

SNP arrays overestimate performance

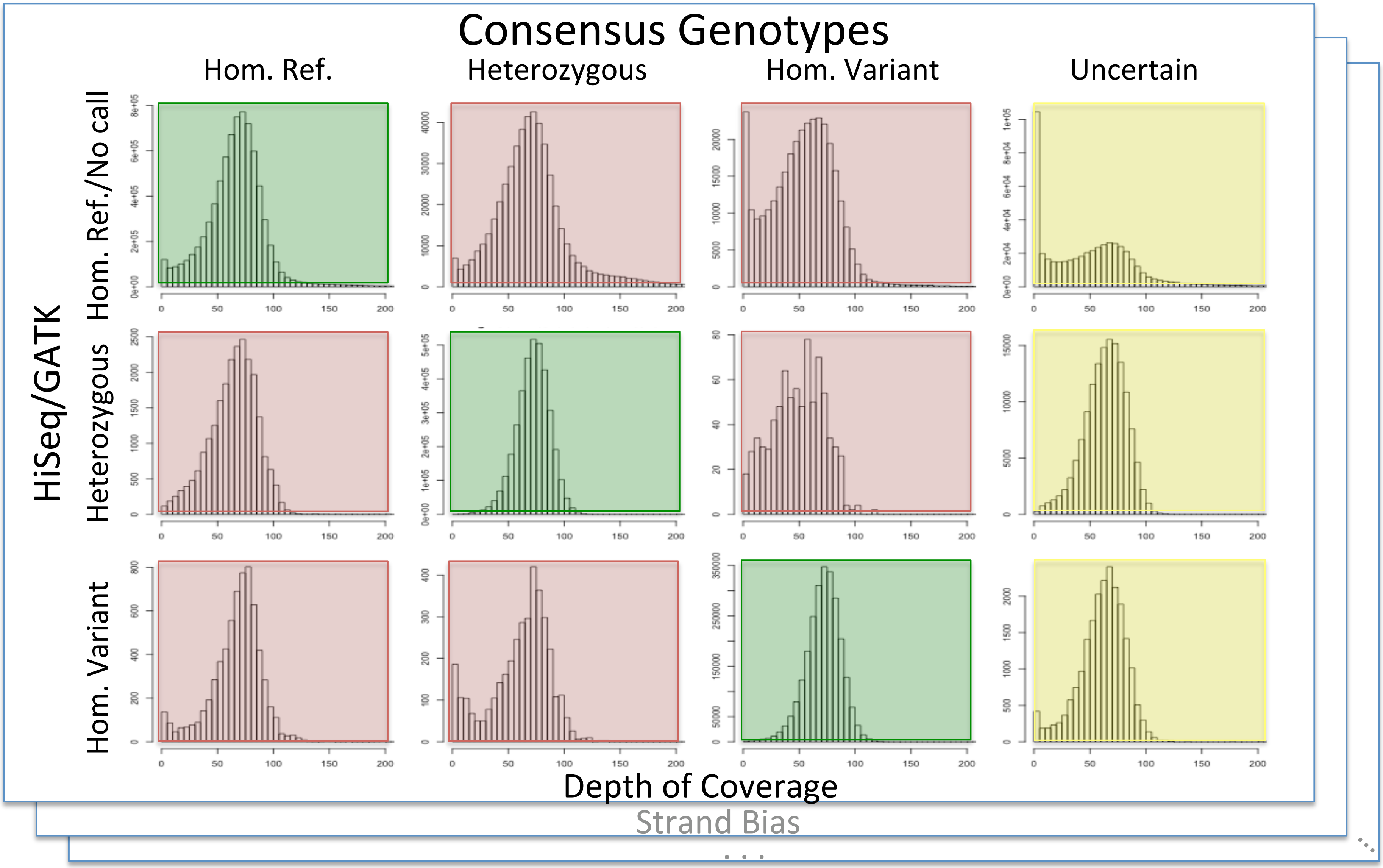
OMNI SNP Array

		Homozygous Reference	Heterozygous	Homozygous Variant	Uncertain
HiSeq – GATK	Homozygous Reference/ No Call	1.45M	**FNs** *7.24k (1.34%)*	*5.28k (0.65%)*	N/A
	Heterozygous	**FPs** *196 (0.03%)*	411k (60.7%)	*133 (0.02%)*	N/A
	Homozygous Variant	*154 (0.02%)*	*150 (0.02%)*	249k (37.0%)	N/A

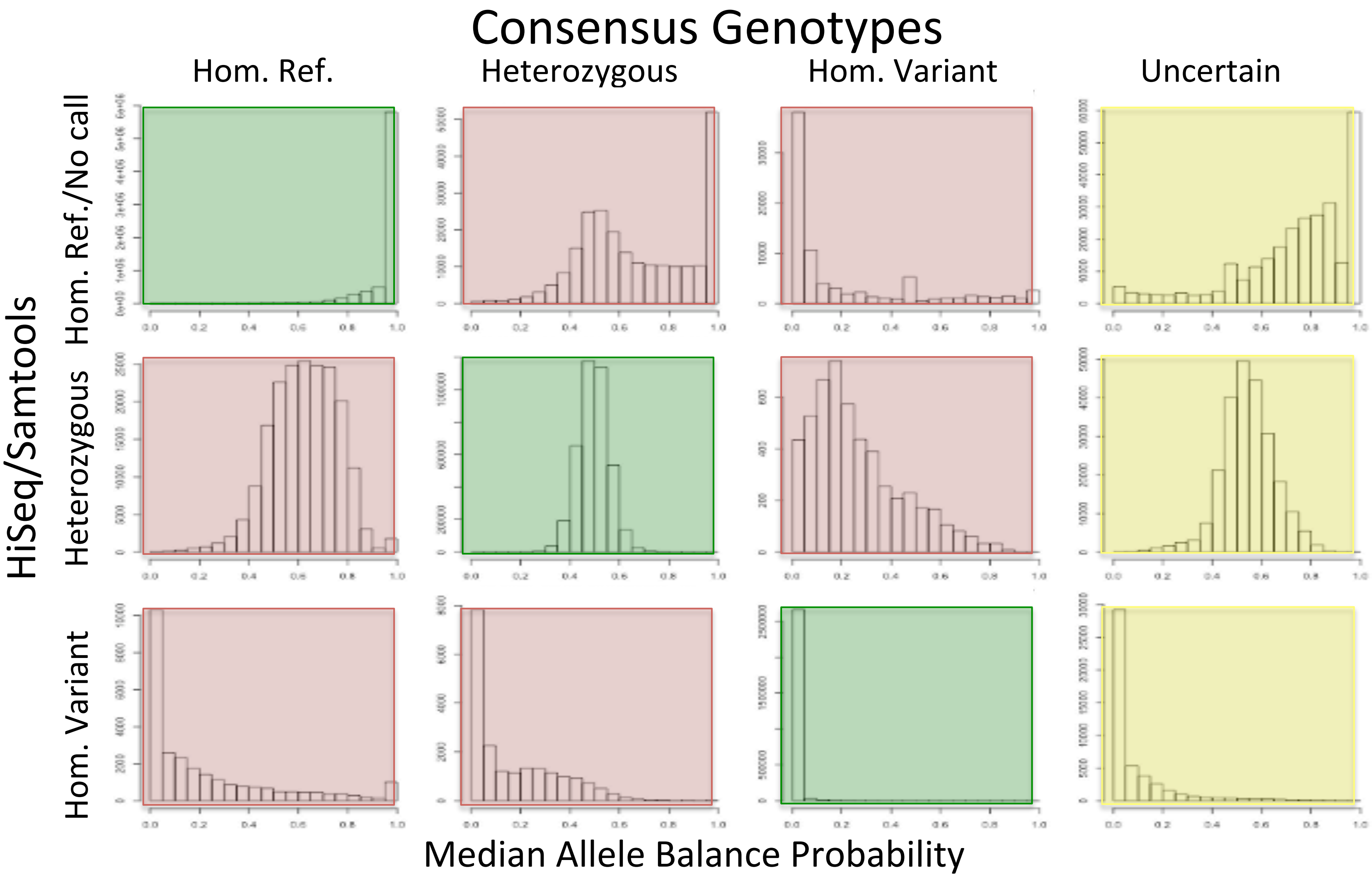
Integrated Consensus Genotypes

		Homozygous Reference	Heterozygous	Homozygous Variant	Uncertain
HiSeq – GATK	Homozygous Reference/ No Call	5.53M	**FNs** *181k (4.73%)*	*153k (3.99%)*	329k (8.58%)
	Heterozygous	**FPs** *6094 (0.18%)*	1.82M (47.5%)	*317 (0.01%)*	85.9k (2.24%)
	Homozygous Variant	*1934 (0.05%)*	*401 (0.01%)*	1.25M (32.5%)	13.8k (0.36%)

Performance Metrics: Characteristics of Mis-calls

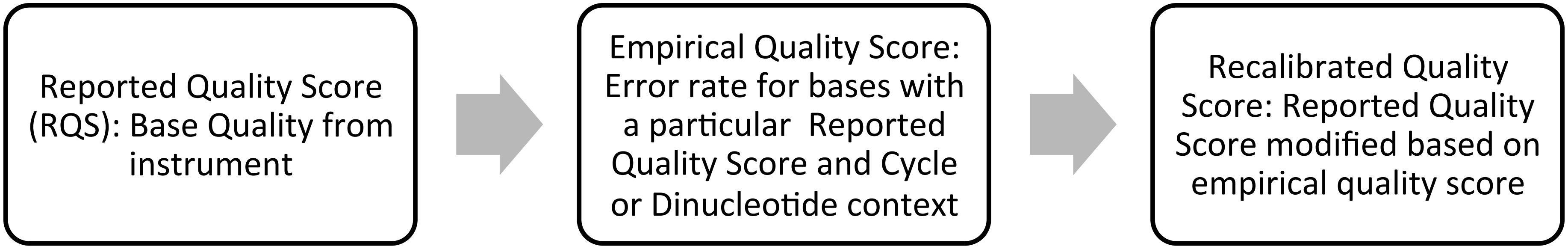


Performance Metrics: Characteristics of Mis-calls



Recalibrating base quality scores with Reference Materials

Ref	A	C	C	T	G	G	A	T	C
Read1 →	A	C	C	G	G	G	A	T	C
Cycle	1	2	3	4	5	6	7	8	9
Quality	30	30	25	10	30	20	20	30	30
Dinuc	NA	AC	CC	CG	GG	GG	GA	AT	TC
Read2 ←	T	G	G	A	C	C	T	C	G
Cycle	9	8	7	6	5	4	3	2	1
Quality	30	30	25	30	30	20	30	30	30
Dinuc	GT	GG	AG	CA	CC	TC	CT	GC	NG

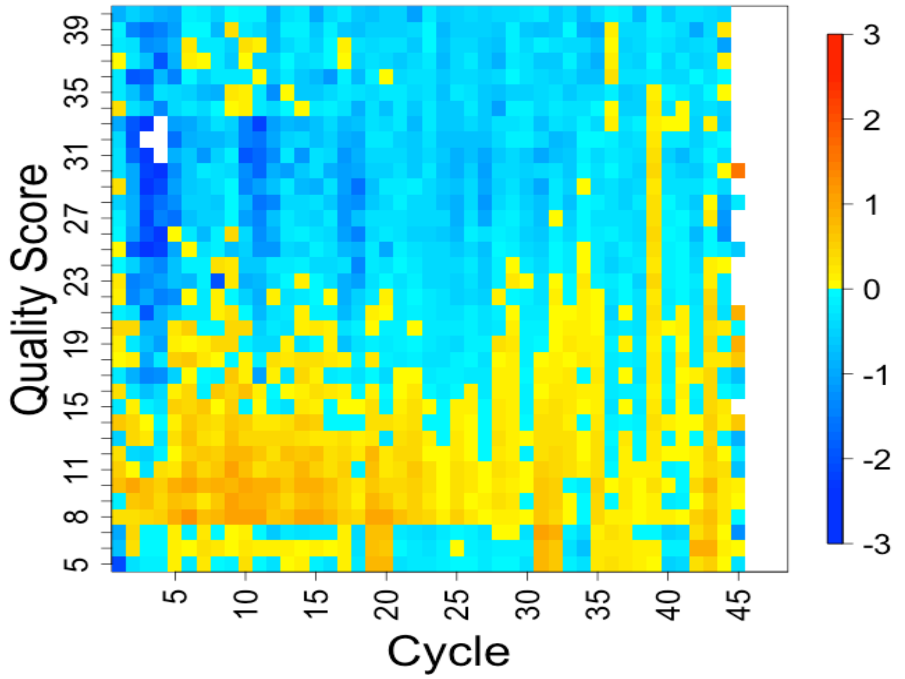


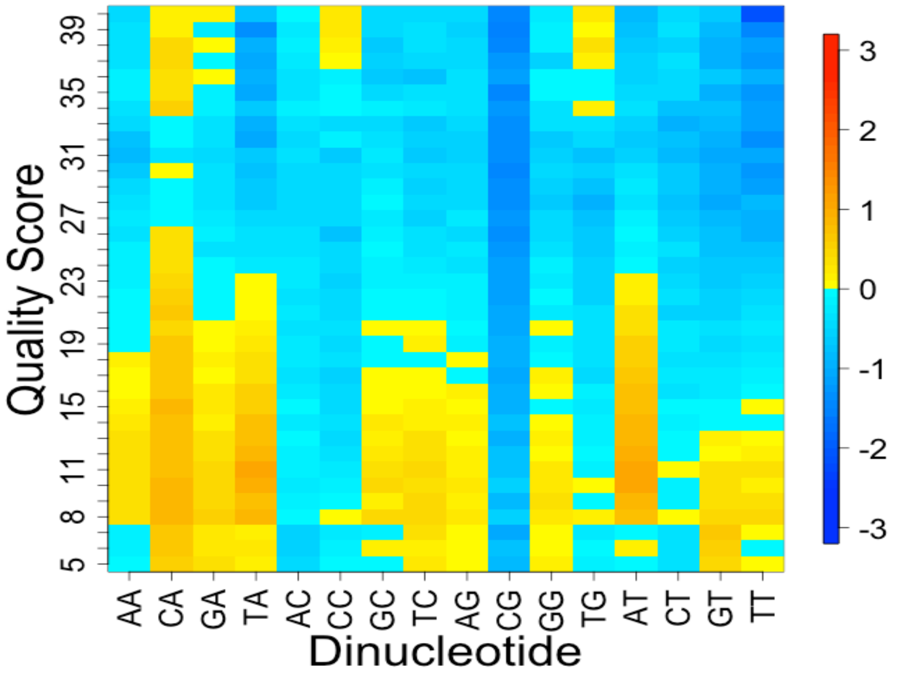
Reported Quality Score (RQS): Base Quality from instrument

Empirical Quality Score: Error rate for bases with a particular Reported Quality Score and Cycle or Dinucleotide context

Recalibrated Quality Score: Reported Quality Score modified based on empirical quality score

Genome recalibration generally lowers qualities too much (esp. at CpG sites)





Blue: Genome recalibration lowers quality scores too much
Yellow-orange: Genome recalibration raises quality scores too much

Zook et al., Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing, *PLoS One*, submitted.

Utility of Reference Materials (RMs)

- RMs to assess sequencing performance are important for many applications (research, clinical, forensic, etc.)
- Whole genome RMs
  - Characterized by multiple technologies
  - Can identify ways to improve technologies and algorithms
  - Provide constant benchmarks for rapidly changing technologies and algorithms
  - Also looking into bacterial genome RMs
- Synthetic DNA RMs
  - Can be spiked-in to any sample
  - Can test detectability of specific types of variants
  - Can be used to improve GATK Base Quality Score Recalibration
    - Zook JM, et al., Synthetic Spike-in Standards Improve Run-Specific Systematic Error Analysis for DNA and RNA Sequencing *PLoS ONE*, submitted.

“Genome in a Bottle” Consortium

- Public-private-academic consortium
- Select, characterize, and discuss applications of RMs for human whole genome sequencing
- Open meeting at Stanford University in August 2012
- Whole genome RM characterization
  - Perform sequencing with multiple platforms with replicates and family members of prospective RM(s)
  - Develop methods to integrate data from multiple sequencing platforms and bioinformatic algorithms
  - Confirm subset of variants with orthogonal technologies

Acknowledgments

- Marc Salit
- Dan Samarov
- Archon Genomics Xprize
  - Brad Chapman
- Genome in a Bottle Consortium

jzook@nist.gov

NGS has many sources of error and bias

Source of uncertainty	Solutions
Statistical sampling differences	More sequencing
PCR amplification bias	PCR-free techniques
Random sequencing errors	More sequencing, accurate quality scores, single cell sequencing
Systematic sequencing errors	**Multiple platforms, base quality score recalibration, strand bias information, orthogonal validation**
Global mapping errors (duplications)	Paired reads, longer reads, accurate mapping qualities, read coverage info, aCGH, optical mapping, fosmids, decoy reference sequence
Local alignment errors (repeats, complex variants)	*de novo* assembly, longer reads, lower sequencing error rates

Other types of variants are more difficult than SNPs

- Indels (scale – 1-10s of bases)
- Large insertions and deletions (>10s of bases)
- Copy number variants (CNVs)
- Inversions
- Complex structural rearrangements

- Many difficult SNPs are near or inside other variants