Using personal genomes for ENCODE data

Jieming Chen

Gerstein Lab, Yale University

AWG Call

Mar 11 2016

Reference Genome (RG)

Hg19/GRCh37/Hg38/GRCh38

Reference Genome (RG)

Haploid
Reference
genome

Hg19/GRCh37/Hg38/GRCh38

Transitioning from Reference Genome (RG) to Personal Genome (PG)

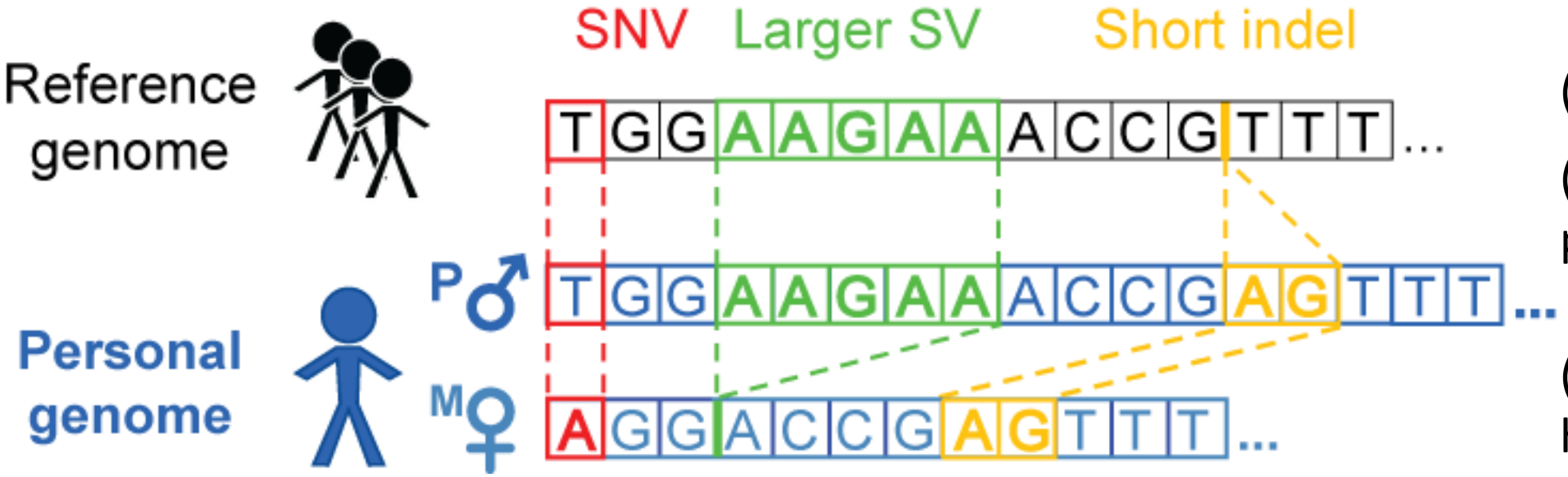
Why the personal genome (PG) should be the platform for functional genomics

- **<u>Advantages of PGs</u>**

1. **Diploid**
   --Ability to incorporate private variants of any size
   --exhibit phase information
2. **Scale easily with more samples and improving sequencing technologies: longer reads and more accurate phase information**
   --We have developed a tool for PG construction
3. **Very useful in functional genomic assay analyses**
   a) read alignment
   b) RNA-seq quantification
   c) allele-specific analyses

Constructing a personal genome

Personal genome construction using *vcf2diploid* tool allows variants of different sizes



(fasta; reference)

+

(vcf, variants phased or unphased)

↓

(fasta; for each haplotype)

Rozowsky *et al. Mol Syst Biol* (2011)

Some construction considerations

1. **Choice of call set(s)**

-- e.g. different versions of 1000GP call sets

2. **Choice of variants**

-- e.g. SVs or indels or SNVs only

3. **Choice of reference**

-- choose the reference genome in which the call set is derived from

4. **Assessment of call set quality**

-- e.g. analysis of Mendelian inconsistency in family data

Assessment of call set quality: Mendelian inconsistency (e.g. GATK HC PCR-free CEU trio)

NA12891 Father	NA12892 Mother	NA12878			total	%Err
		RR	RA	AA		
RR	RA	518631	505499	1215	1025345	0.12
RR	AA	1659	194589	1806	198054	1.75
RA	RR	507750	506699	1110	1015559	0.11
RA	RA	194409	397233	195245	786887	---
RA	AA	742	194722	206720	402184	0.18
AA	RR	1485	193636	1551	196672	1.54
AA	RA	653	198416	202366	401435	0.16
AA	AA	113	1316	816825	818254	0.17

*Autosomes only

NA12878 family of PGs we already have

	Source	Refgen	Depth	Variants
1	1000 Genomes Project (1000GP) pilot	hg18	60x	SNVs, indels, deletions (including 33 from fosmid sequencing)
2	GATK Best Practices v3 (UnifiedGenotype)	hg19	64x	SNVs, indels
3	GATK Best Practices v4 (HaplotypeCaller, PCR-free)	hg19	64x	SNVs, indels
4	1000GP Phase 3 SNVs-only	hg19	7.4x	SNVs
5	1000GP Phase 3 SNVs-indels	hg19	7.4x	SNVs, indels
6	1000GP Phase 3 SNVs-indels-SVs	hg19	7.4x	SNVs, indels, SVs

Other possible choices of NA12878 call sets for ENCODE:

1. **Genome In A Bottle (GIAB)**
   --HiSeq2500 300x, 150x150bp read, hg19
   --44x PacBio SV calls (from Mt Sinai School of Medicine)
   (https://sites.stanford.edu/abms/content/giab-reference-materials-and-data, updated Sep 2015)

2. **Complete Genomics**
   --80x, SNVs, indels and SVs, GRCh37
   (http://www.completegenomics.com/public-data/69-Genomes/)

3. **Illumina Platinum Genomes**
   --HiSeq2000, PCR-free, 50x and 200x, SNVs and indels, available for both hg19 and hg38
   (http://www.illumina.com/platinumgenomes/)

4. **1000 Genomes Project SV group**
   --SV calls using longer reads: PacBio, Moleculo

Useful list of NA12878 public datasets (google sheet on GIAB site):
https://docs.google.com/spreadsheets/d/1iL45zPit9-kVmk-9sDJEGMhxsUWf52n1nw_duTdNwcE/edit#gid=0

Highly scalable:
Use of personal genomes with ENCODE data

- Constructed 382 personal genomes from 1000GP Phase 1 data
  --SNVs and indels
  --match with their corresponding RNA-seq and/or ChIP-seq sets (from 8 different studies)
- Most of our ChIP-seq sets are from ENCODE
  --14 GM cell lines with available 1000GP Phase 1 DNA data

**note that there are 35 GM cell lines in ENCODE, with ChIP-seq data (including CEU and YRI trios)

Utility of PGs:
Read alignment

Alignment gets better as variant sets get more complete: NA12878 Pol2 ChIP-seq (ENCODE)

	Ref genome	Pgenome: SNVs only	Pgenome: SNVs + indels only	Pgenome: SNVs + indels + SVs
Reads processed	208,051,087			
# reads uniquely aligned	171,944,588 (82.65%)	172,591,380 (82.96%)	172,738,321 (83.03%)	172,743,175 (83.03%)
# reads that multimap	17,826,675 (8.57%)	17,795,258 (8.55%)	17,782,167 (8.55%)	17,779,800 (8.55%)

Almost 1M increase in reads

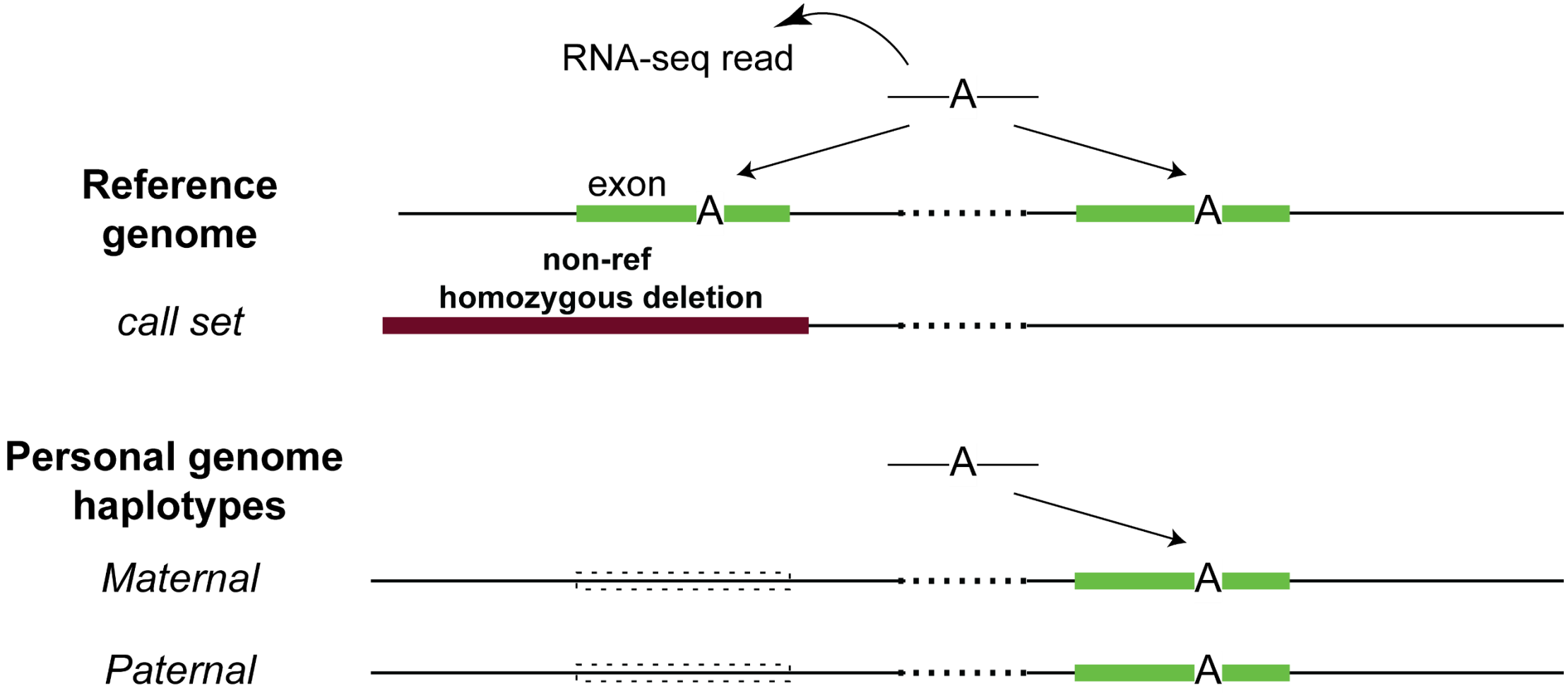
Alignment gets better as variant sets get more complete: NA12878 RNA-seq (Kilpinen *et al.* 2013)

	Ref genome	Pgenome: snvs only	Pgenome: snvs + indels only	Pgenome: snvs + indels + SVs
Reads processed	37,558,398			
# reads uniquely aligned	25,303,498 (67.37%)	25,486,837 (67.86%)	25,538,449 (68.00%)	25,568,042 (68.08%)
# reads that multimap	4,041,495 (10.76%)	4,010,417 (10.68%)	4,012,297 (10.68%)	3,972,990 (10.58%)

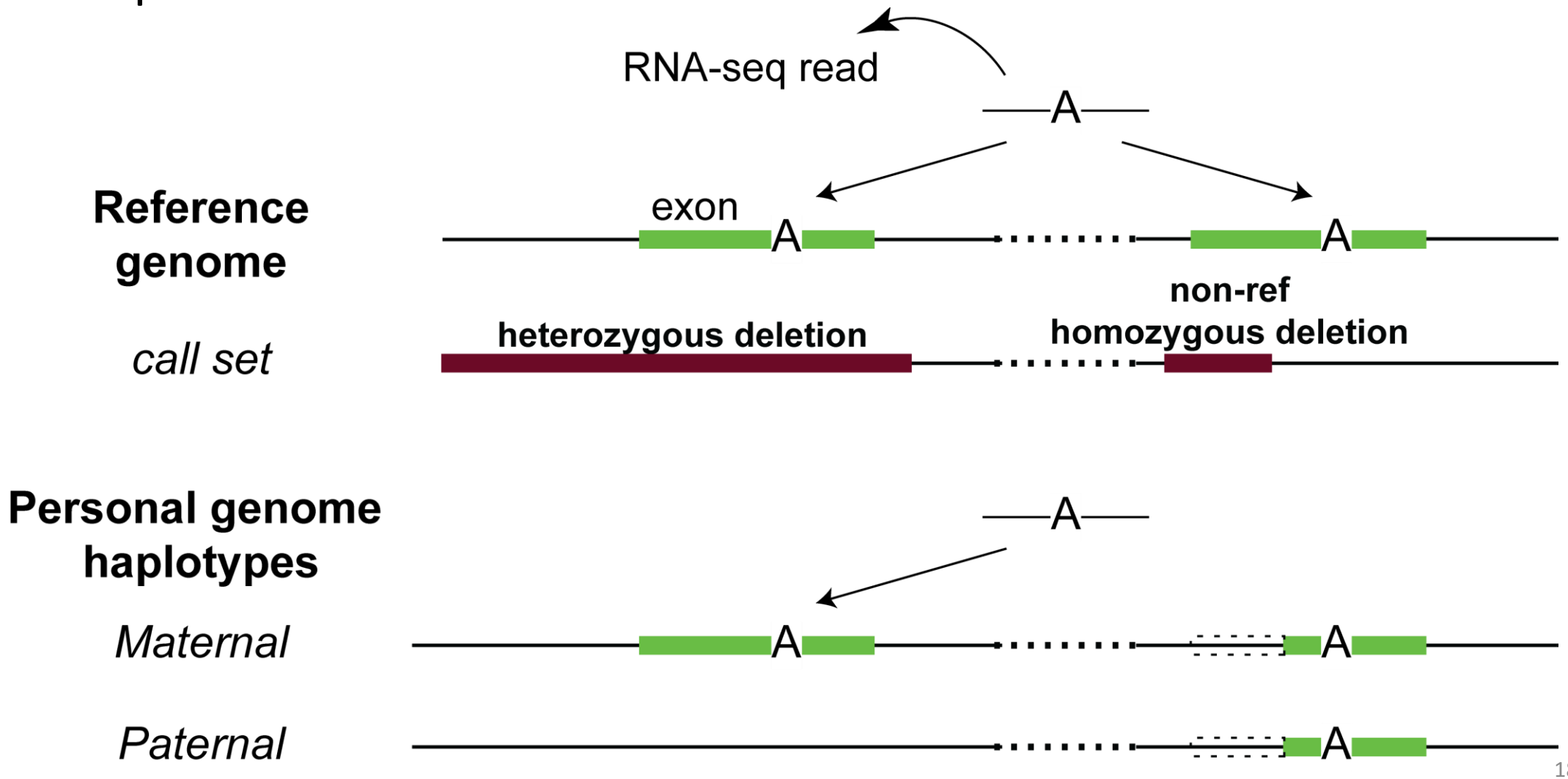
Over 260K increase in reads

Utility of PGs:
RNA-seq quantification with SVs incorporated

Including SVs in diploid personal genomes: a simple example



Including SVs in diploid personal genomes : more complex



SVs-in-PG analyses in Sudmant *et al.*, *Nature* (2015)

- Constructed 2 personal genomes of NA12878 based on GRCh37 reference genome

1. 1000GP P3 **SNVs and indels** integrated call set (low coverage)
2. 1000GP P3 **SNVs, indels and SVs** with breakpoint information

Utility of SVs: Exons with direct SV overlap

Comparing between vanilla GRCh37 ref and Pgenome-SVs

- 18 exons (5 genes) with a direct SV overlap
- 6/18 exons were affected in terms of expression (±10 reads)
- 4/6 showed substantial changes in expression using RNA-seq data from Kilpinen *et al.* (±10 reads, 2x change)

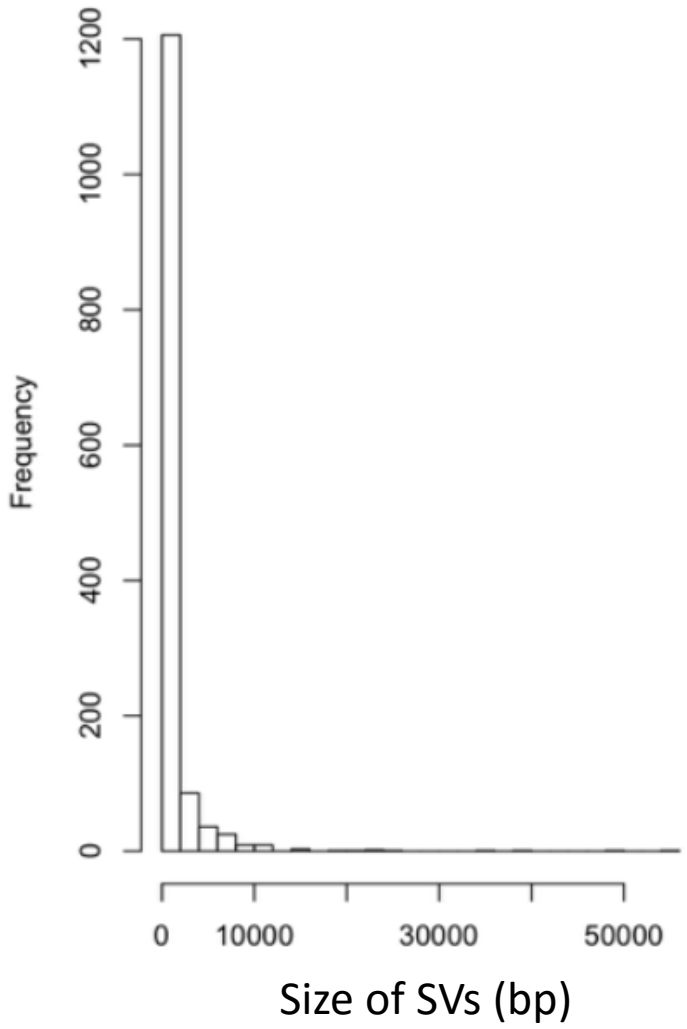
Oliver Stegle, Sudmant *et al., Nature* (2015)

SVs-in-PG analyses in Sudmant *et al.*, *Nature* (2015)

- Constructed 2 personal genomes of NA12878 based on GRCh37 reference genome

1. 1000GP P3 **SNVs and indels** integrated call set (low coverage)
2. 1000GP P3 **SNVs, indels and SVs** with breakpoint information

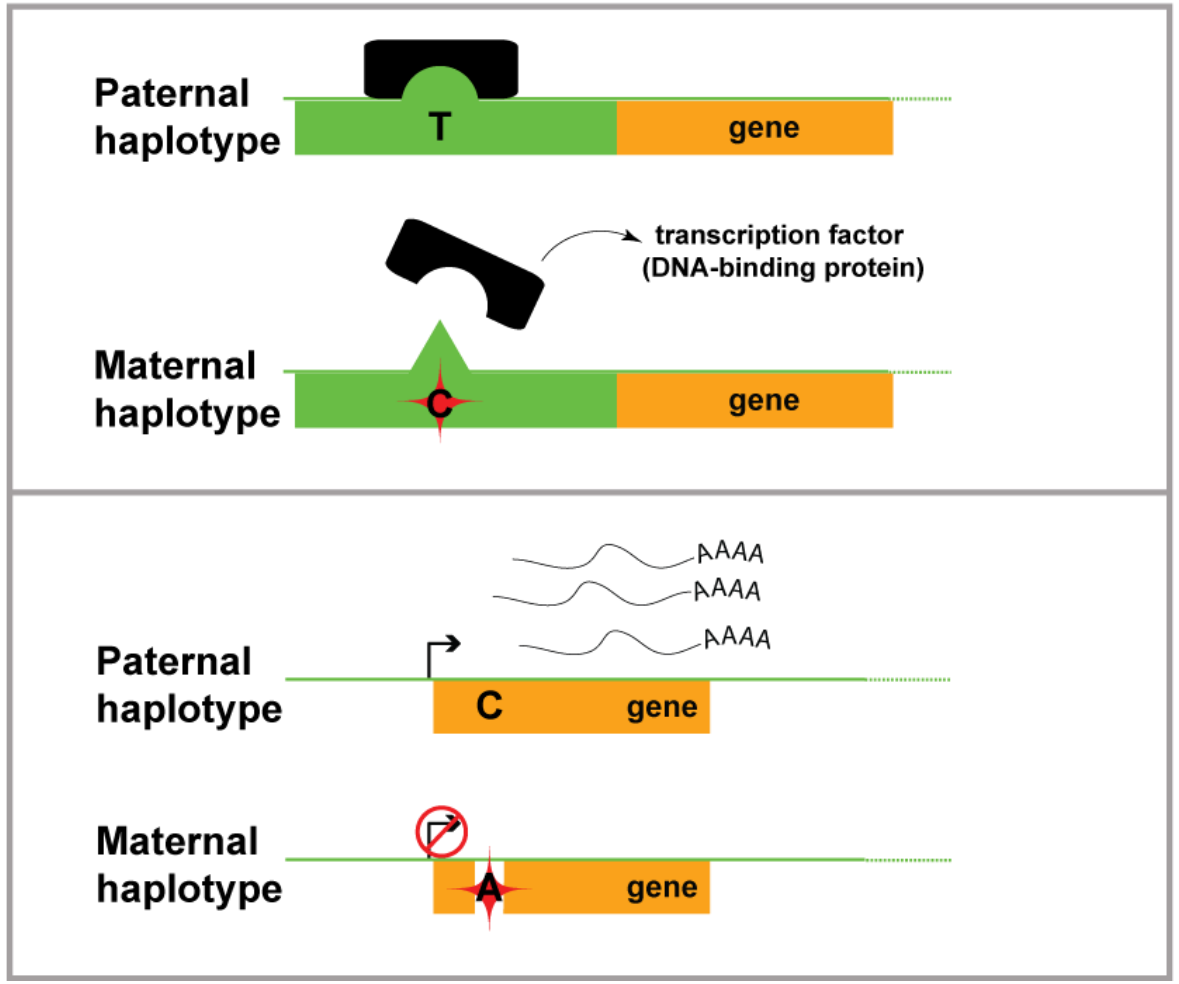
**1,383 /68,000 SVs
--with precise breakpoint information
--most are still sub-1kb



Oliver Stegle

Utility of PGs:
Alleviate biases in detection of allele-specific variants

Allele-specific (AS) behavior



Compute allelic ratio (i.e. read difference at the 2 alleles)

e.g. a SNV @ chr 7 position 4345325

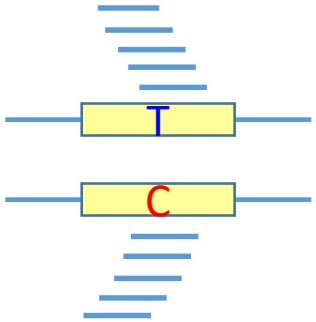
RNA-/ChIP-Seq Reads

```
ACTTTGATAGCGTCAACG
 CTTTGATAGCGTCAACGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAACGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
```

5 x T (ref)

5 x C

Allelic ratio = 0.5
(i.e. 'null' expectation)



e.g. a SNV @ chr 5 position 12455

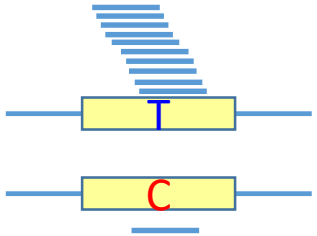
RNA-/ChIP-Seq Reads

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAATGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
```

9 x T (ref)

1 x C

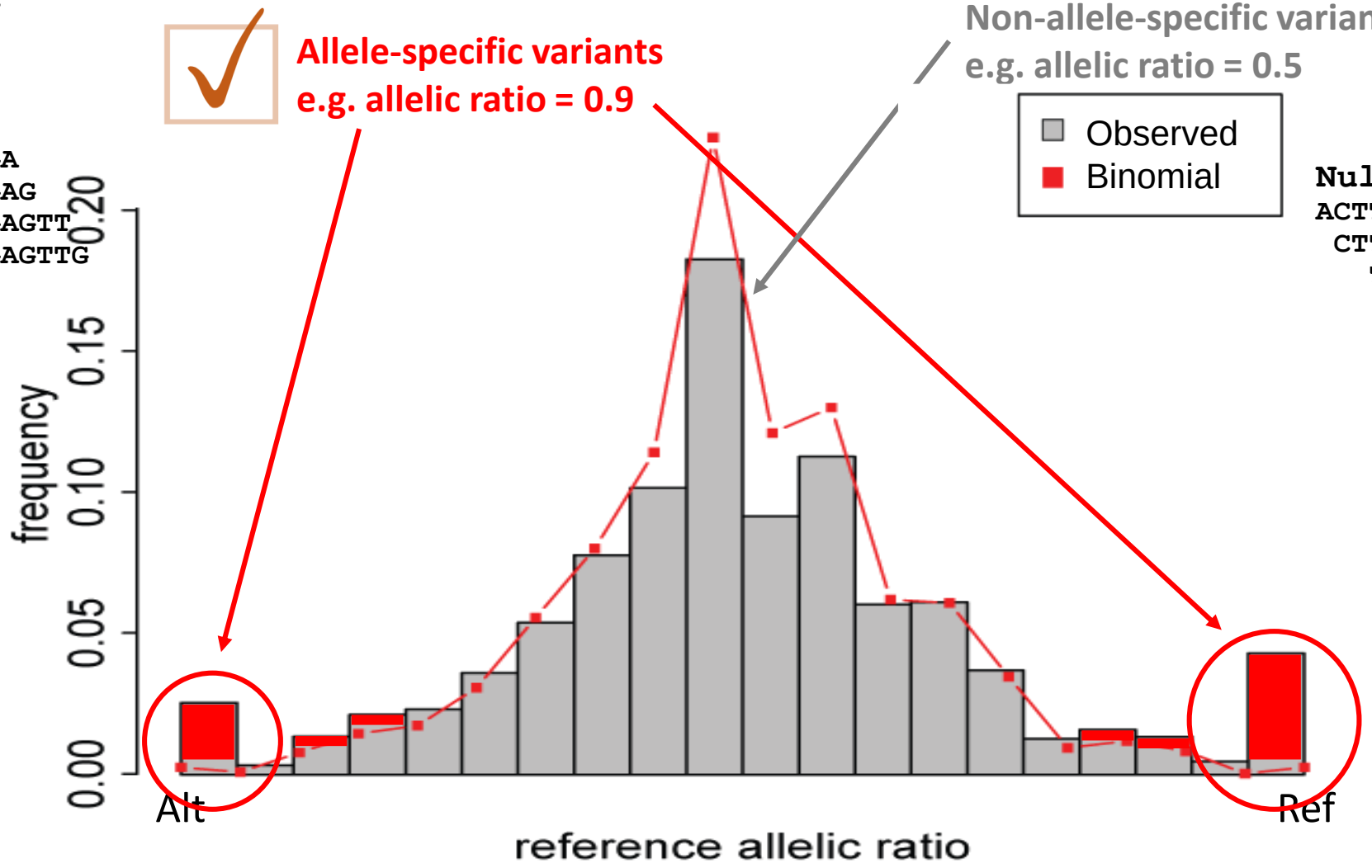
Allelic ratio = 0.9



Binomial model on observed distribution

ASE/ASB Example:

```
...GTCAATGCAC
...GTCAATGCACG
...GTCAATGCACGTC
...GTCAATGCACGTCG
...GTCAACGCACGTCGGGA
 GTCAATGCACGTCGAGAG
   CAATGCACGTCGGGAGTT
    AATGCACGTCGGGAGTTG
```



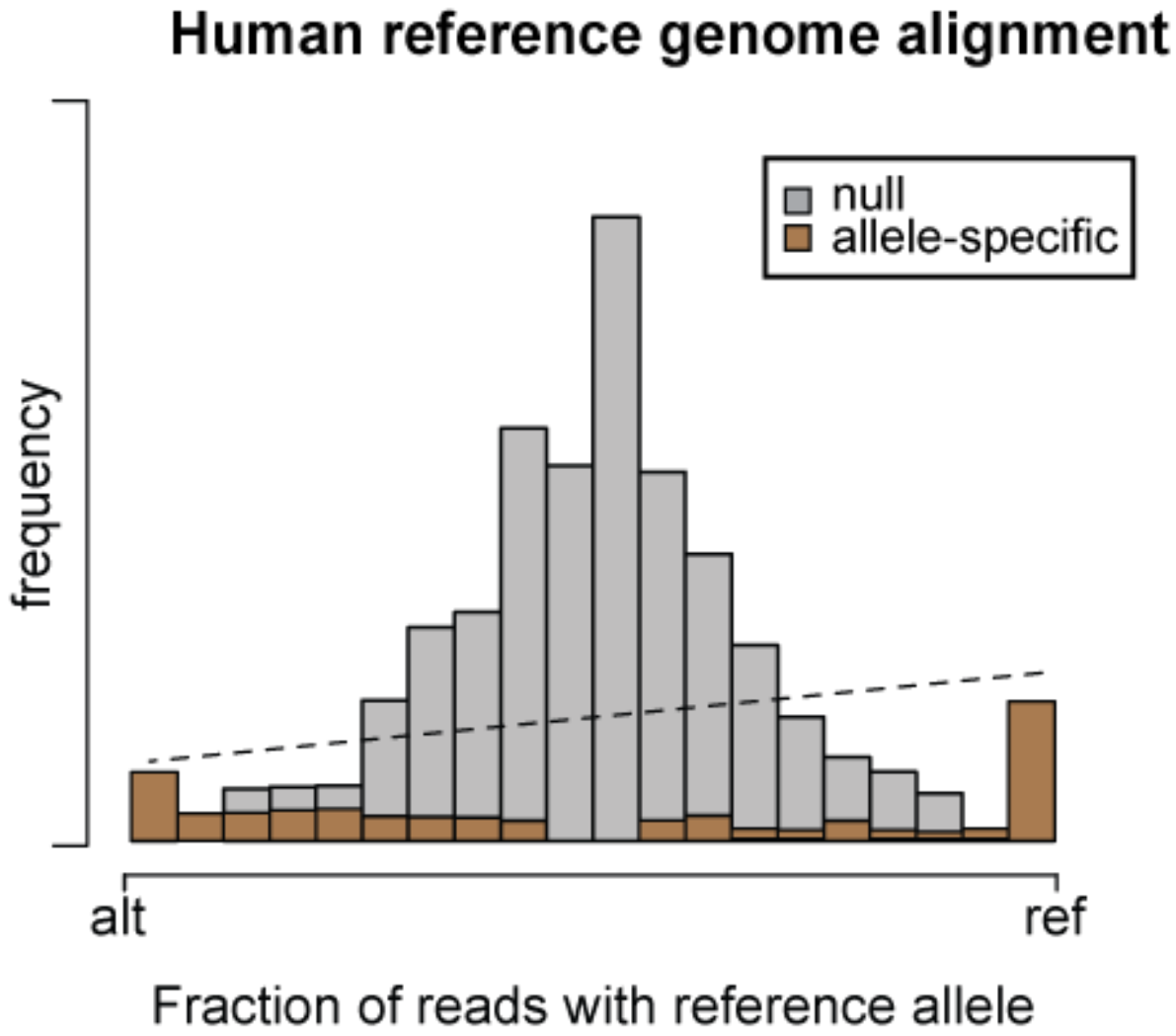
Null Example:

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
      ATAGCGTCAATGCACGT...
       TAGCGTCAACGCACGT...
          CGTCAACGCACGT...
             CAATGCACGT...
              AATGCACGT...
```

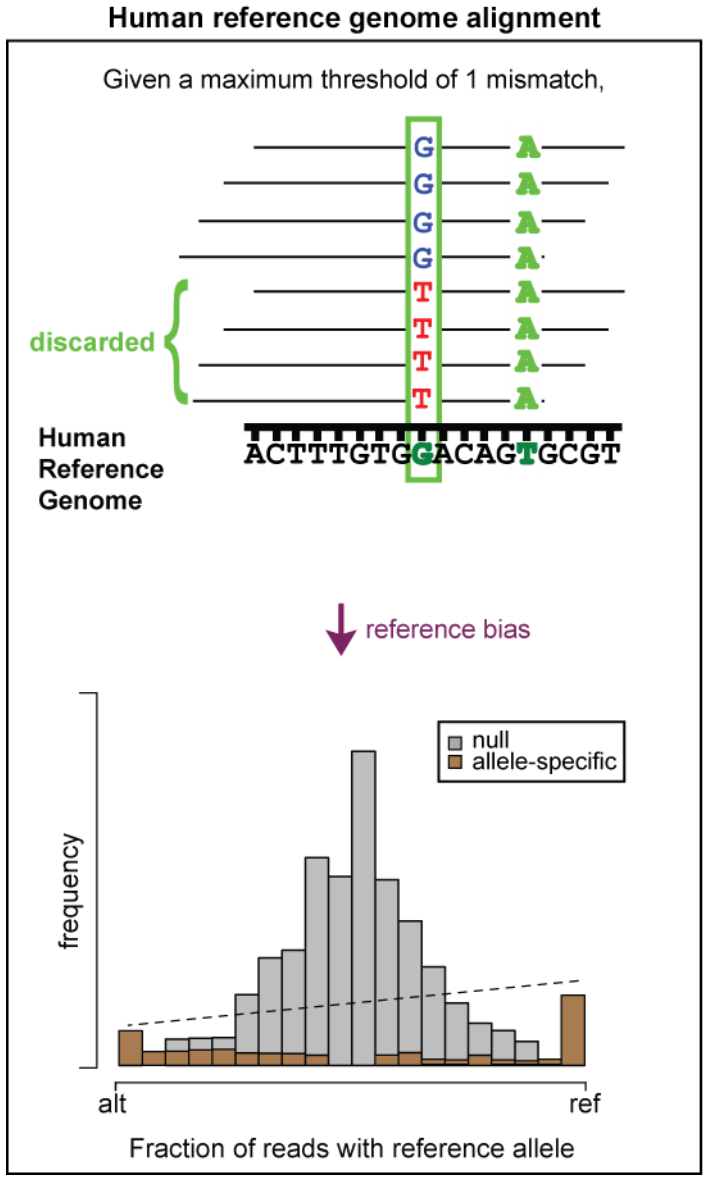
Some issues in AS analyses

1. Reference bias
2. Ambiguous mapping bias
3. Over-dispersion

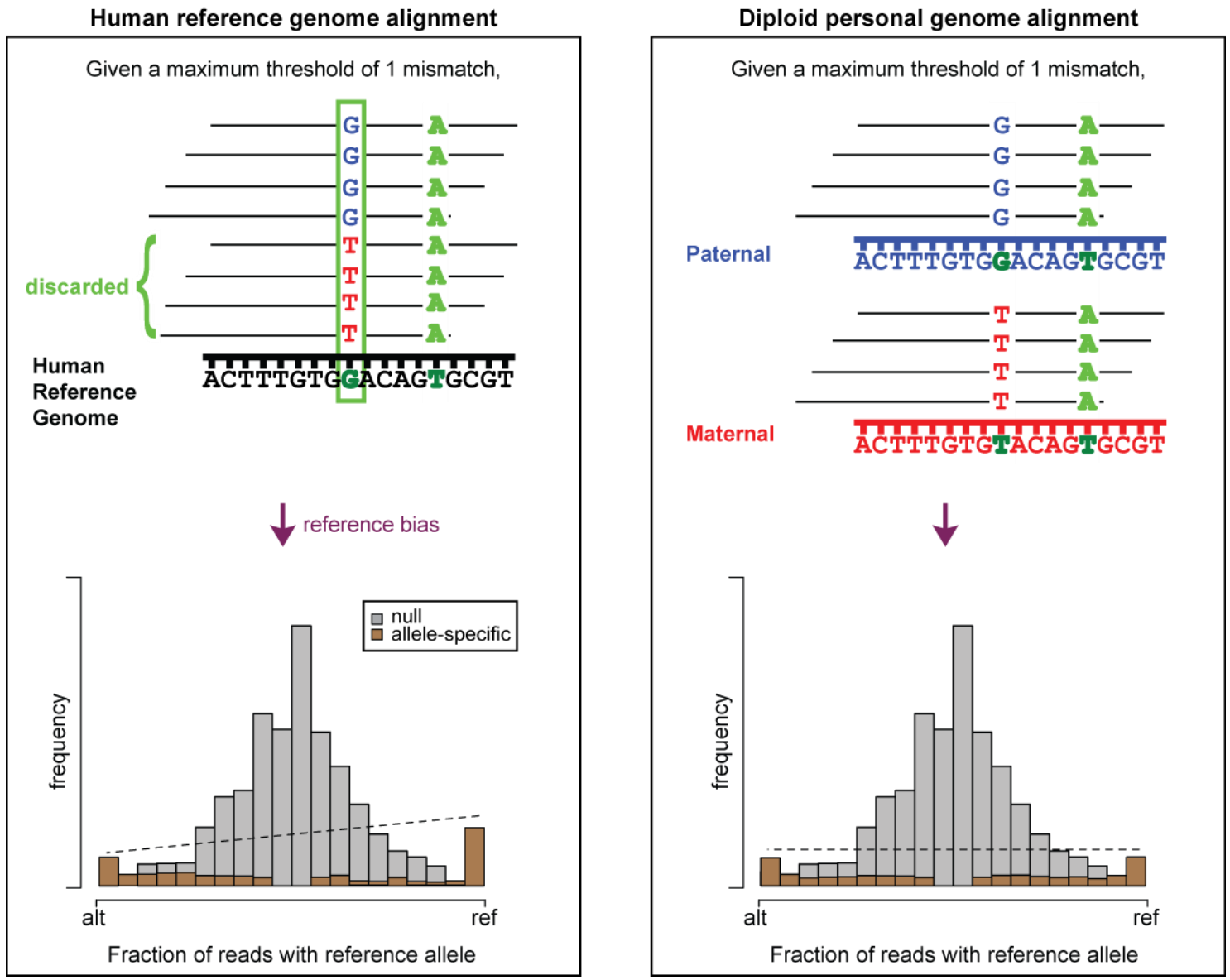
Reference bias
(naïve alignment to the human reference genome)



PG alleviates reference bias in alignment



PG alleviates reference bias in alignment

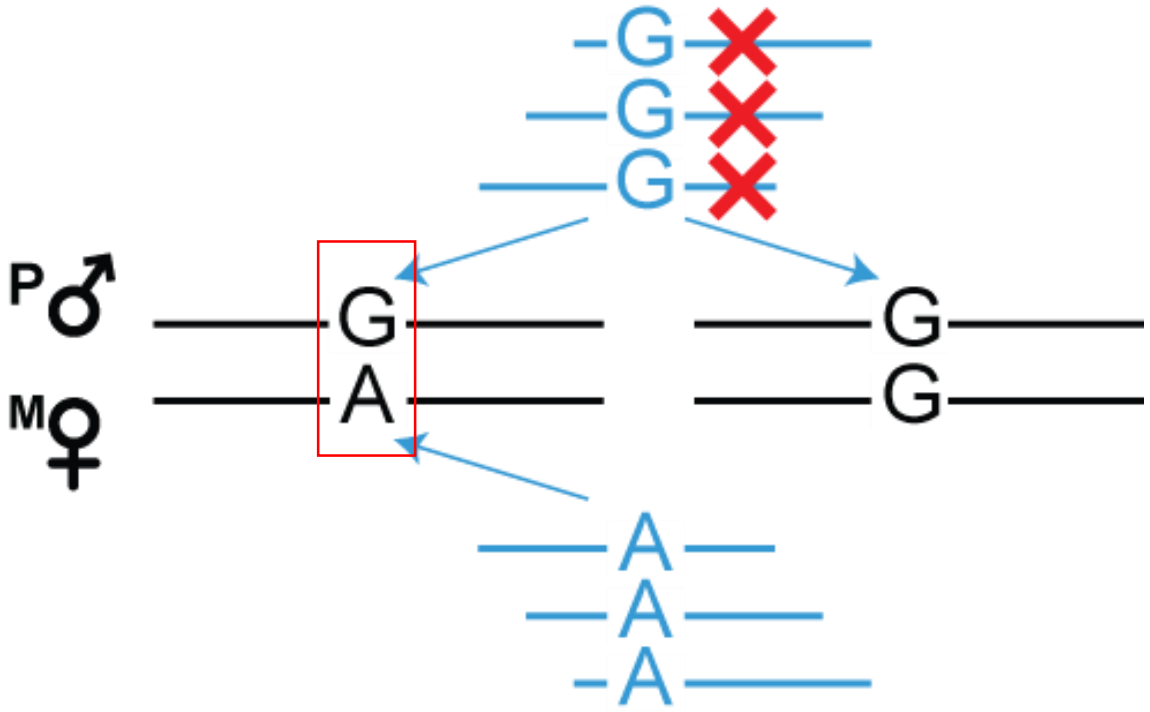


Some issues in AS analyses

1. Reference bias
2. Ambiguous mapping bias
3. Over-dispersion

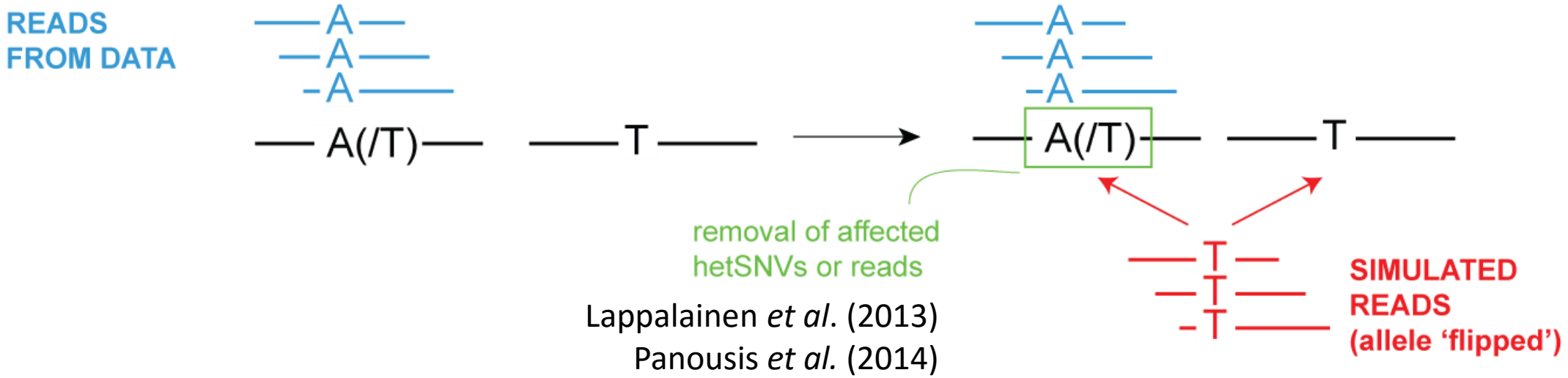
Ambiguous mapping bias due to sequence similarity

- For AS analyses, discard reads that multi-map



Account for ambiguous mapping bias

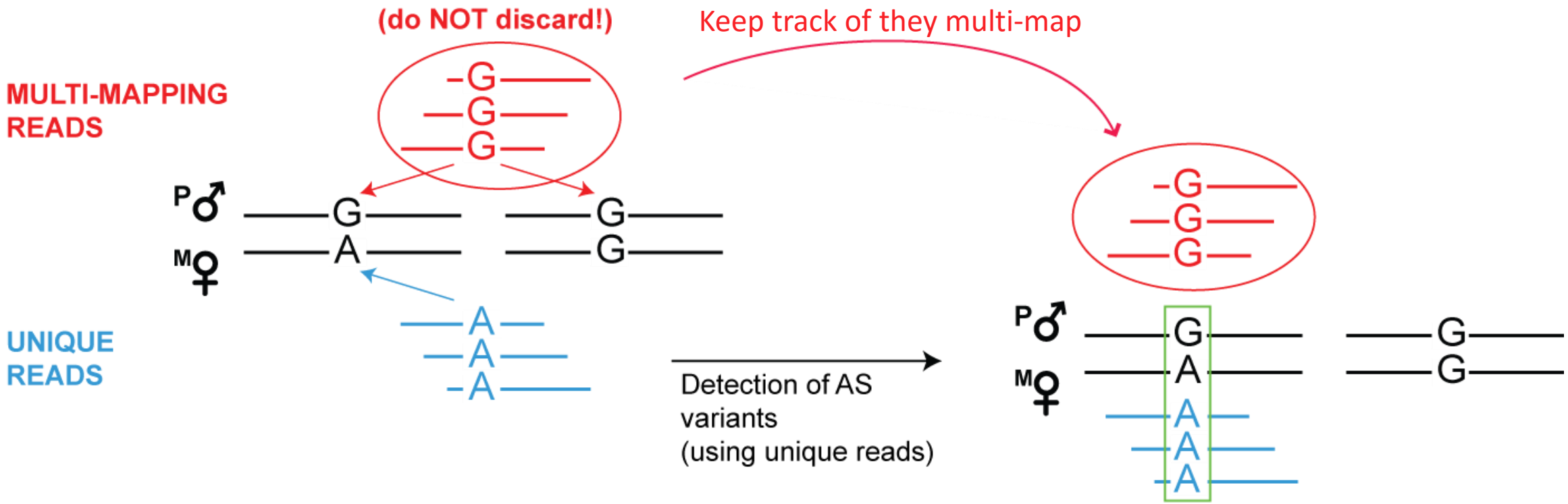
- Using the reference genome, new simulated reads are created where alleles of the original reads are flipped (at het SNV positions)



Lappalainen *et al*. (2013)
Panousis *et al.* (2014)
Van de Geijn *et al.* (2015)

PG facilitates the resolution of ambiguous mapping bias

- Using the personal genome, we do not need to simulate reads.
- We can directly test affected sites using multi-mapping read pile



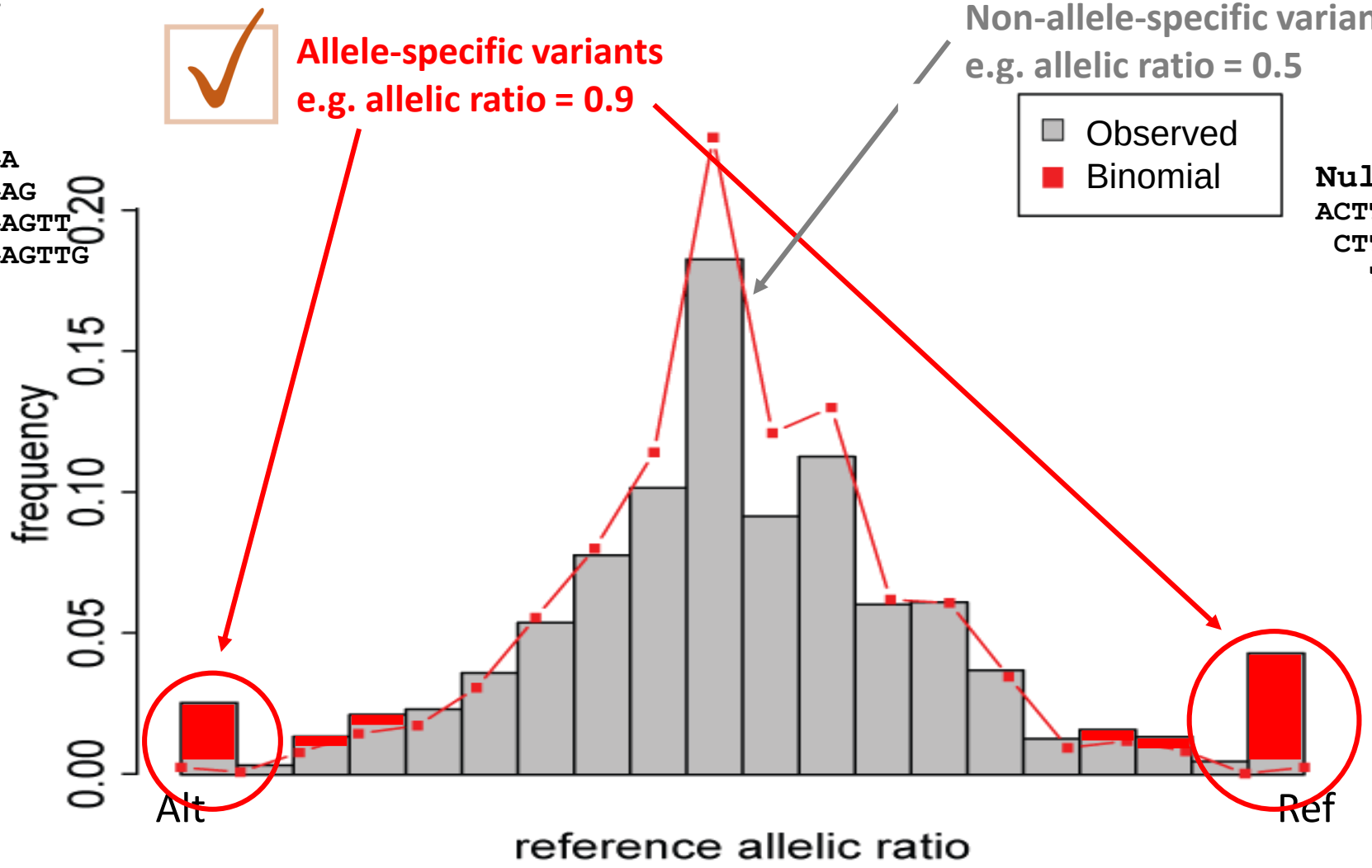
Some issues in AS analyses

1. Reference bias
2. Ambiguous mapping bias
3. Over-dispersion

Binomial model on observed distribution

ASE/ASB Example:

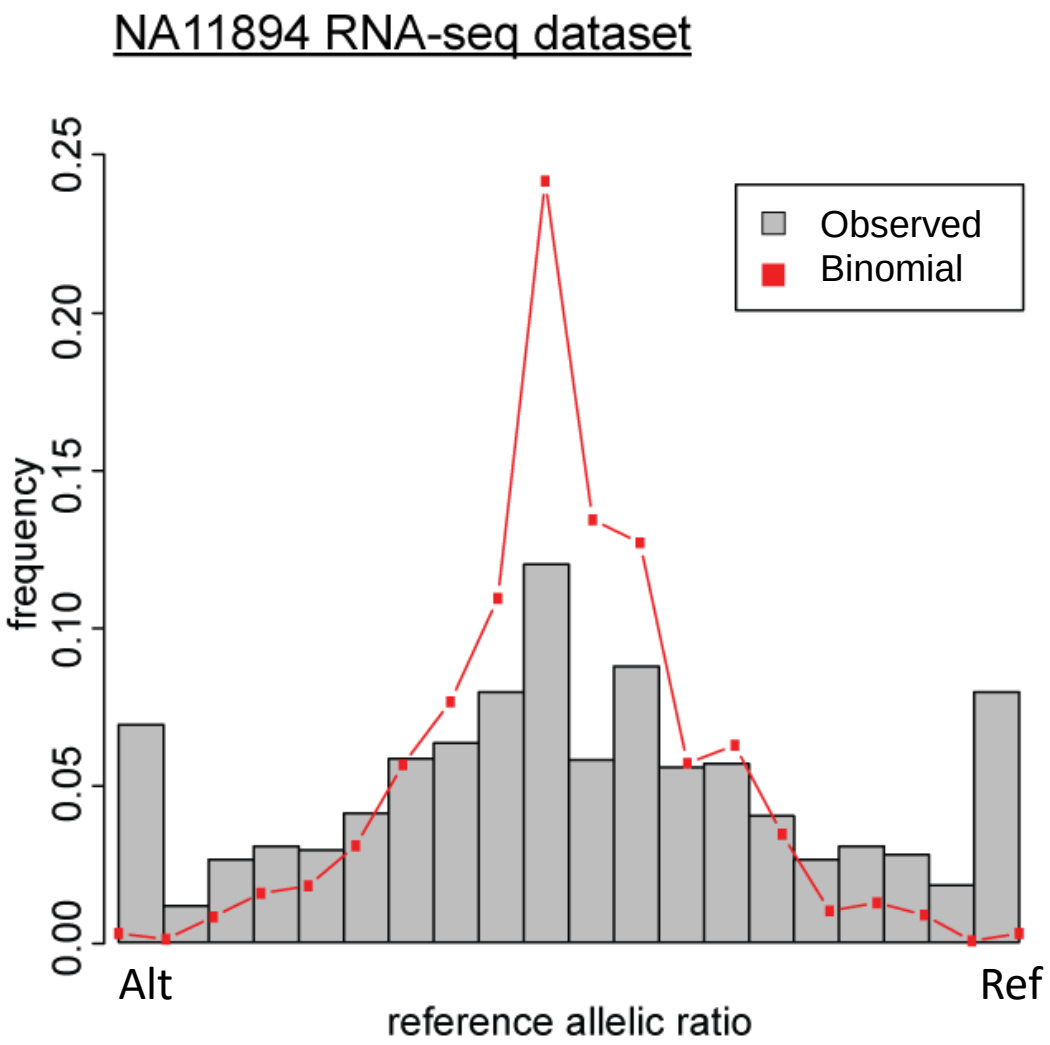
```
...GTCAATGCAC
...GTCAATGCACG
...GTCAATGCACGTC
...GTCAATGCACGTCG
...GTCAACGCACGTCGGGA
 GTCAATGCACGTCGAGAG
   CAATGCACGTCGGGAGTT
    AATGCACGTCGGGAGTTG
```



Null Example:

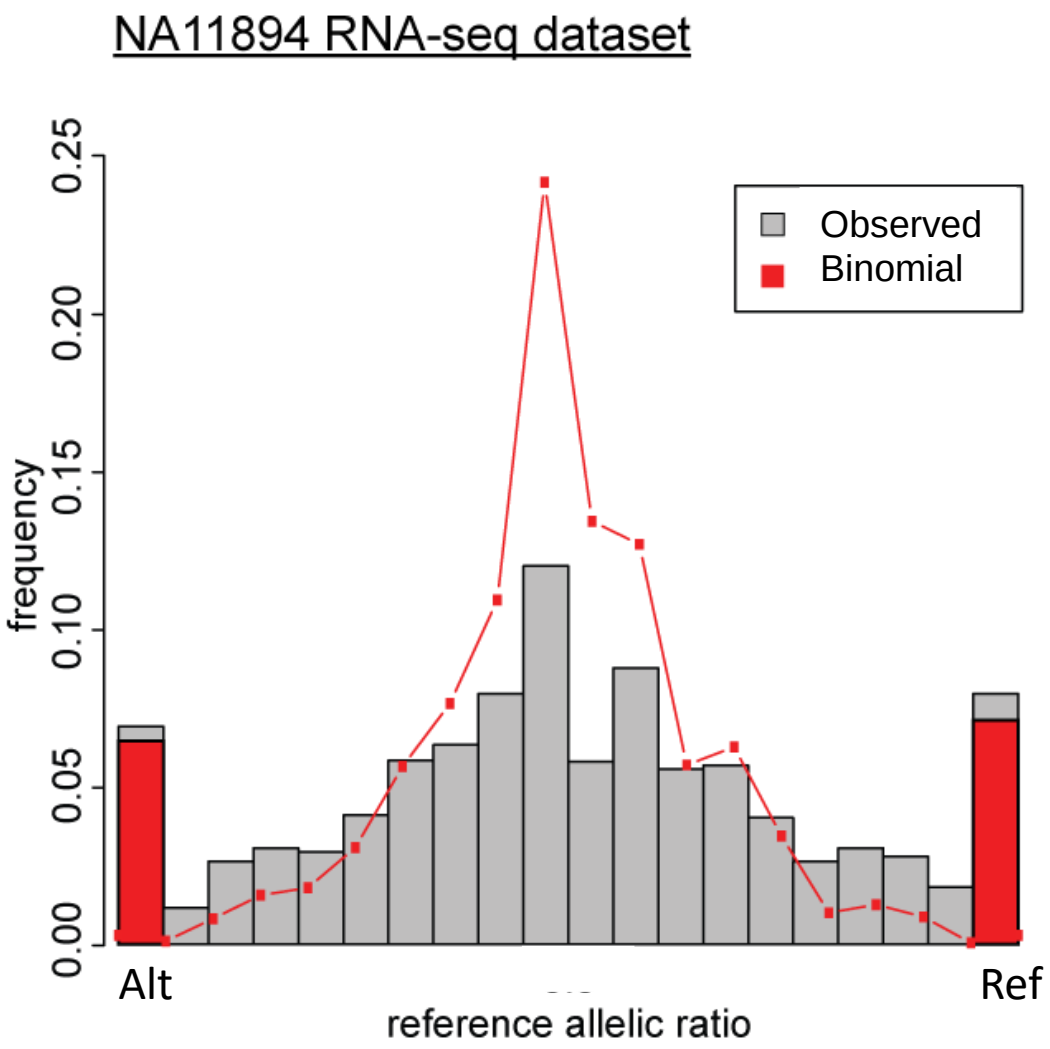
```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
      ATAGCGTCAATGCACGT...
       TAGCGTCAACGCACGT...
          CGTCAACGCACGT...
             CAATGCACGT...
              AATGCACGT...
```

Over-dispersion
– broader distribution than expected



- Binomial distribution insufficient to explain over-dispersed observed distribution

Over-dispersion
– broader distribution than expected

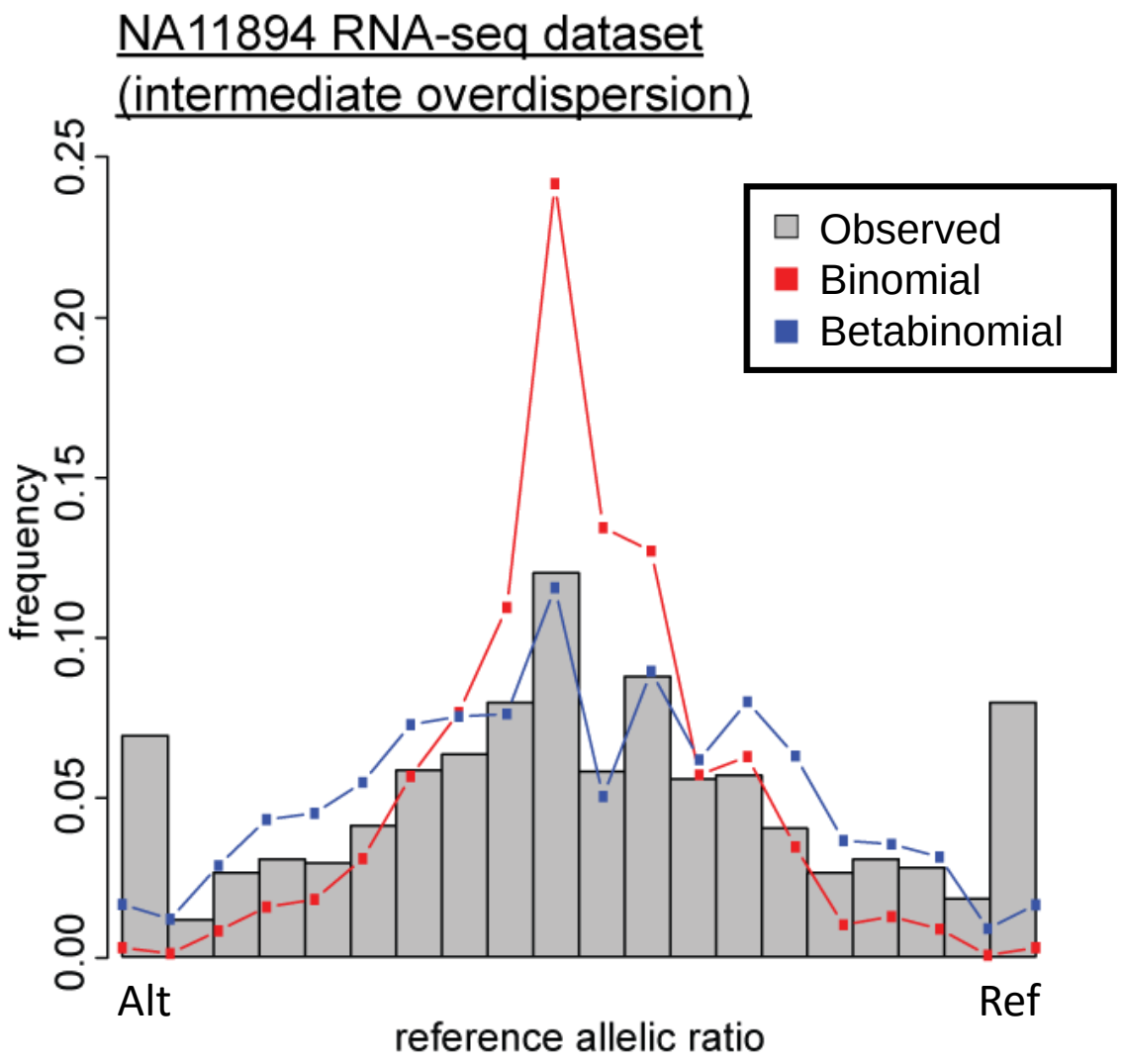


- Binomial distribution insufficient to explain overdispersed empirical distribution
- Binomial test over-calls allele-specific variants

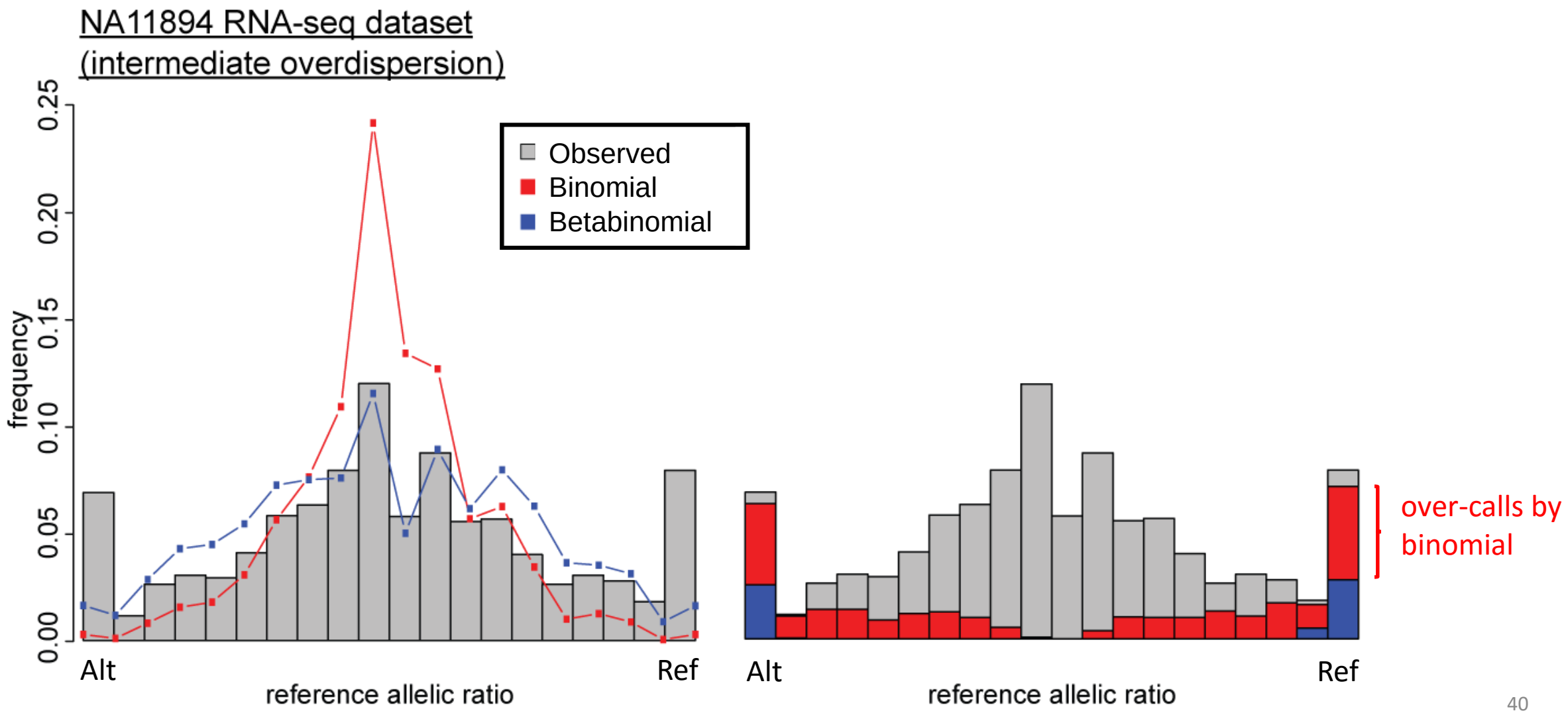
Beta-binomial distribution – to account for over-dispersion

Binomial	Beta-binomial
$\binom{n}{k} p^k (1-p)^{n-k}$	$\binom{n}{k} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha,\beta)}$
• 2 parameters -- X ~ (n,k)	• 4 parameters -- X ~ (n,k) -- k ~ Beta(α,ß)
n = total number of reads k = allelic ratio	n = total number of reads k = allelic ratio **accounting for the overdispersion** of allelic ratio distribution

Accounting for overdispersion: Binomial VS Beta-binomial

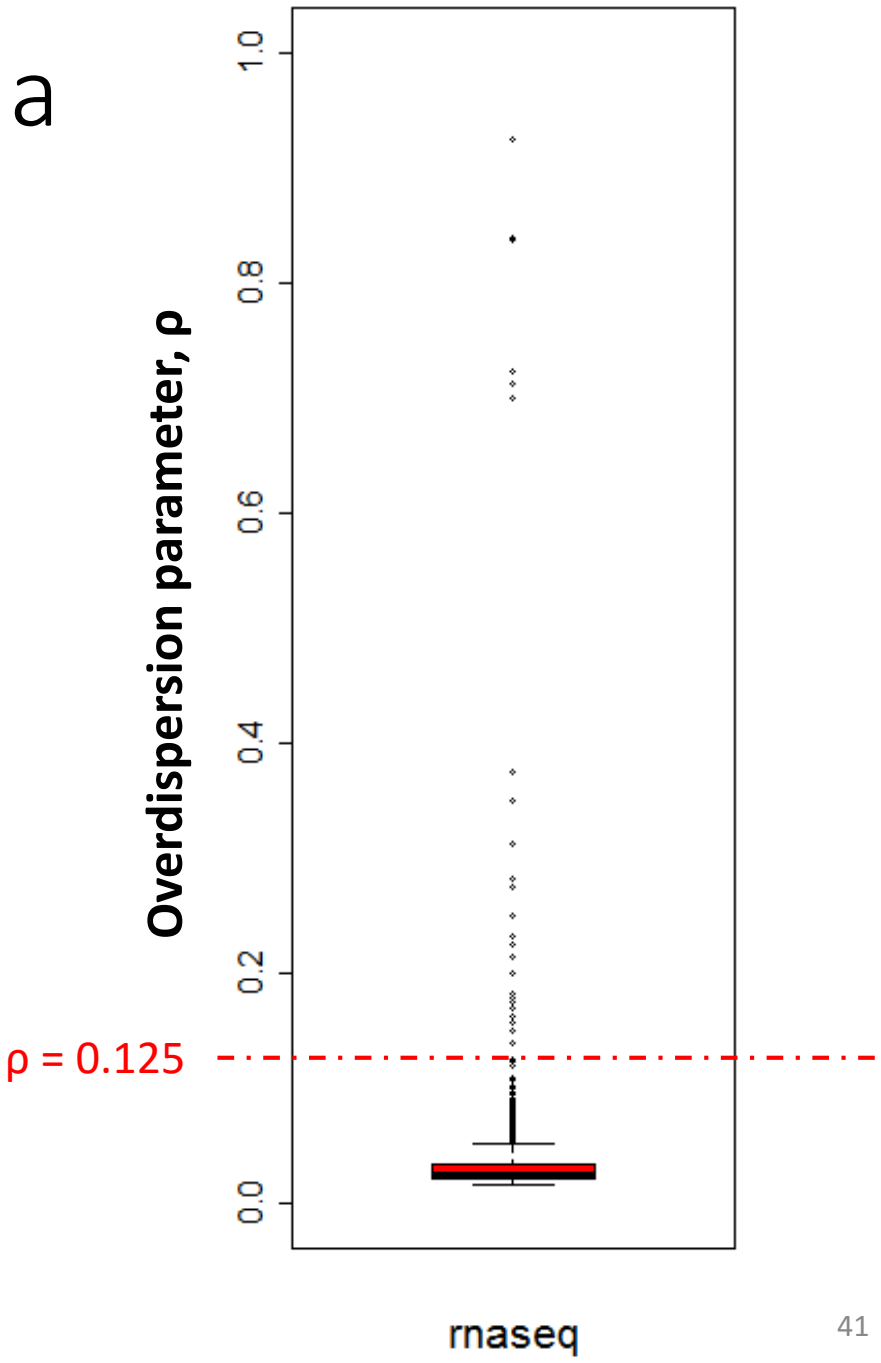


Accounting for overdispersion: Binomial VS Beta-binomial

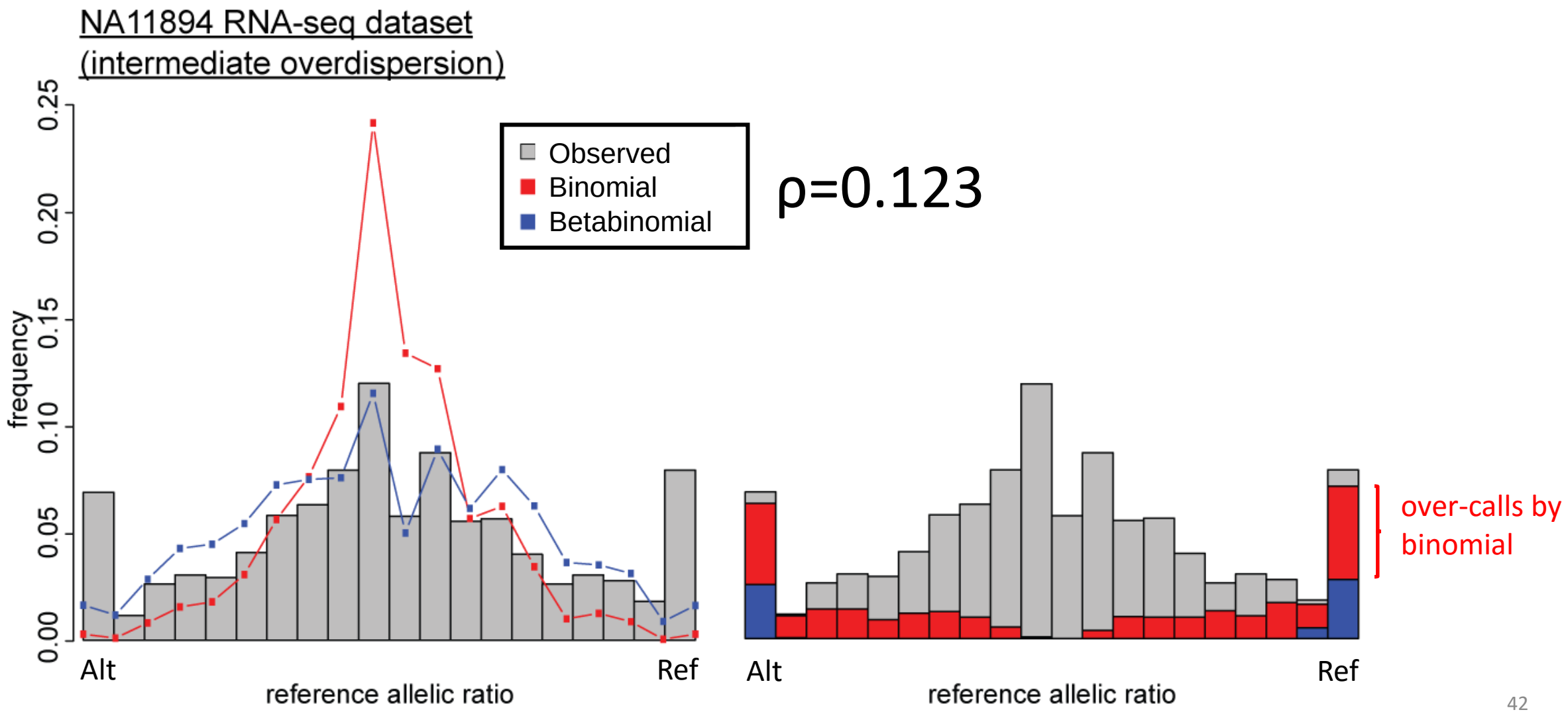


Overdispersion parameter (ρ) as a useful QC metric for data harmonization

- Curated 987 RNA-seq datasets from eight different studies (including ENCODE)
  -- removed 32 datasets with ρ ≥ 0.125 (1 sd from ρ_{mean})



Accounting for overdispersion: Binomial VS Beta-binomial



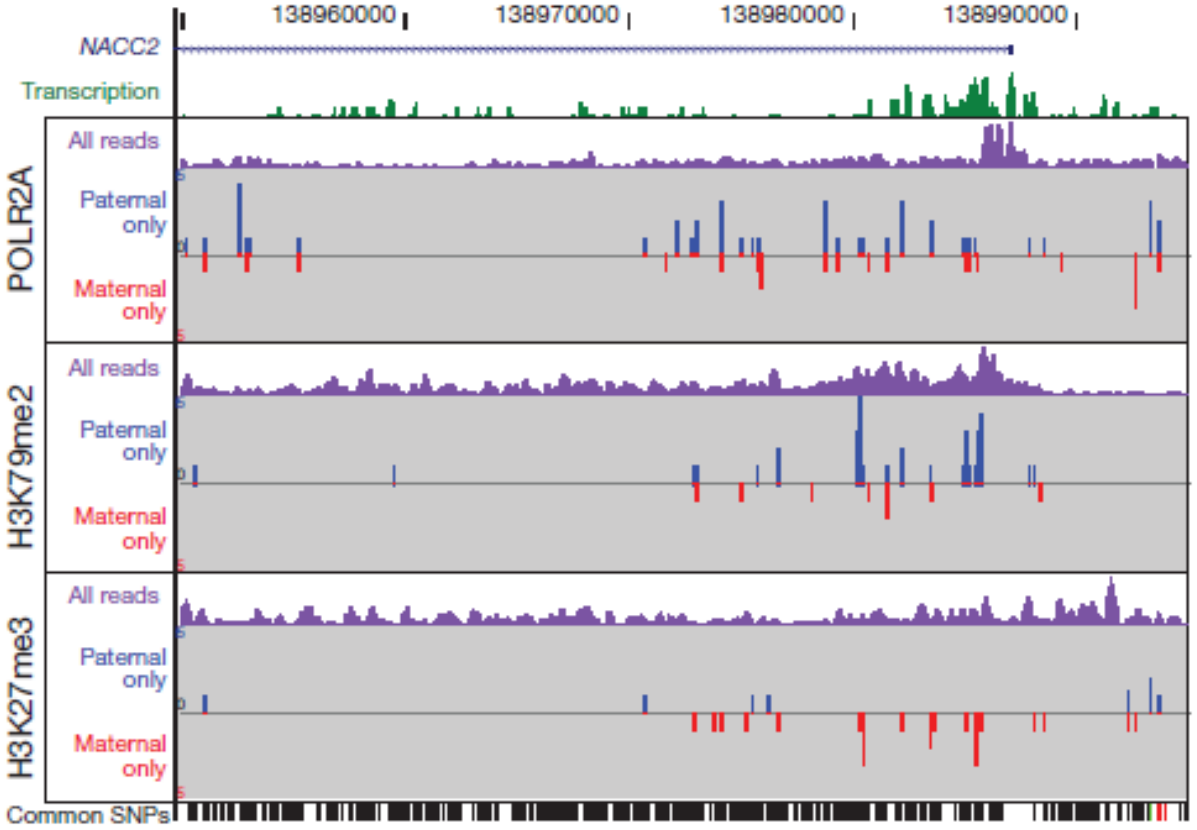
Summary

1. PG is a **more realistic and intuitive** representation of the human genome
2. PG can **incorporate variants of any size** (e.g. SNVs, indels and SVs)
   --improves alignment of reads from functional genomic assays
   --improves accuracy of quantification
3. PG is able to **include phase information and diploid nature**
   --alleviates reference bias
4. PG construction is **highly scalable**
   --Scale easily with more samples and improving sequencing technologies, e.g. longer reads and more accurate phase information
   --more than 300 PGs have been built
   --rapid construction of a PG with *vcf2diploid* tool (~1h for NA12878 full set of variants: SNVs, indels and SVs)
5. PG is **useful in processing and analyses of functional genomic assays**
   --e.g. read alignment, RNA-seq quantification and allele-specific analyses

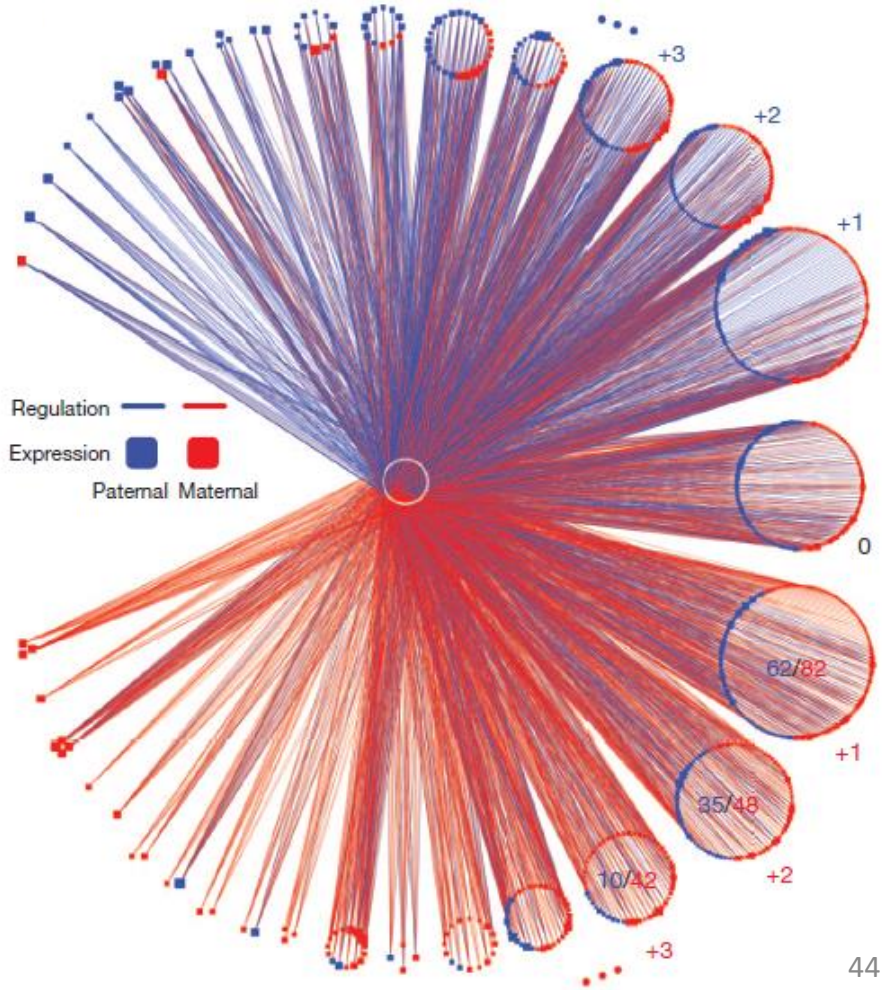
Previous use of a personal genome for a single individual (NA12878) in ENCODE

The Encode Consortium, *Nature* (2012)

Djebali *et al.*, Nature (2012)



Gerstein *et. al.*, *Nature* (2012)



Future prospects of PGs in ENCODE

1. **Diploid genomes for all GM cell lines and H1**
   --14/35 GM cell lines in ENCODE with available 1000GP Phase 1 DNA data

2. **Dealing with non-diploid genomes, e.g. cancer cell lines (polyploid)**
   --make use of available HeLa genome (with phase and ploidy info)

3. **Entex**
   --PGs and functional assays for multiple tissues

Acknowledgement

Yale University

- Joel Rozowsky
- Timur Galeev
- Arif Harmanci
- Rob Kitchen
- Mark Gerstein

EMBL-EBI

- Oliver Stegle

Mayo Clinic

- Alexej Abyzov