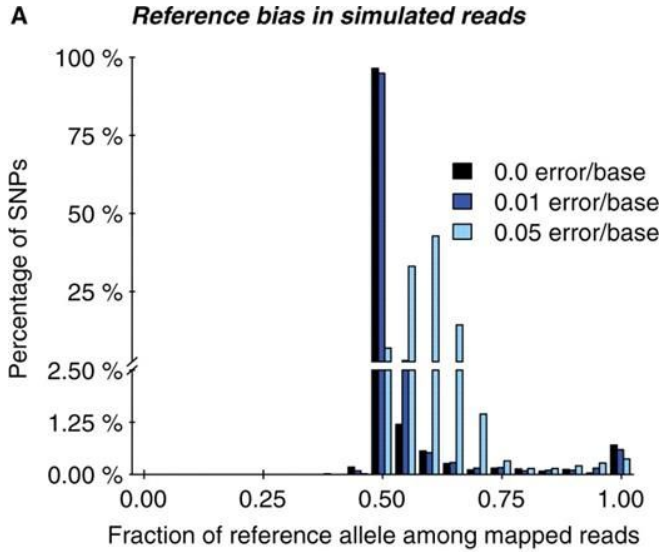
A new strategy to reduce allelic bias in RNA-Seq readmapping

Ravi Vijaya Satya, Nela Zavaljevski, Jaques Reifman
Presented by Sam Kovaka

Reference Bias

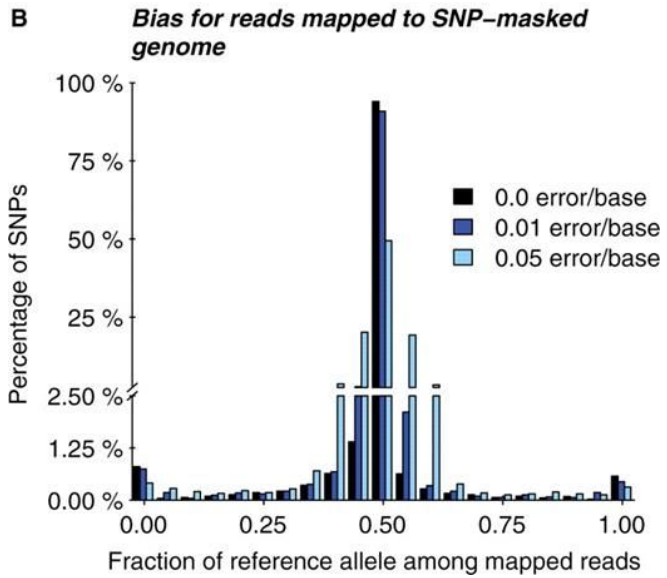
- Reads matching reference genome are more likely to align than those with variants
- Makes measuring allele-specific expression (ASE) difficult
- Multiple ways to compensate
  - Create personalized genome
  - Filter problematic reads
  - Mask variants
  - Enhanced reference genome



Degner et al. 2009

Masking Variants

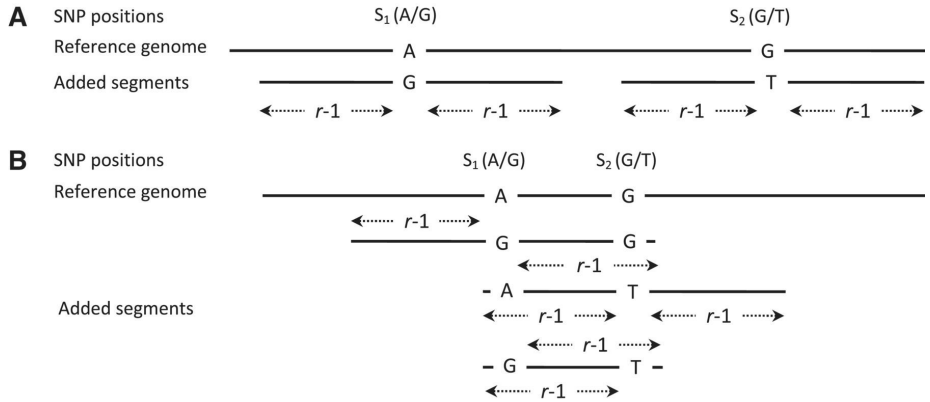
- “Each known single-nucleotide polymorphism (SNP) locus is masked with a third base that is neither the reference allele nor non-reference allele”
  - Aligners at the time couldn’t map to regions with ambiguity (Degner et al. 2009)
- Eliminates systematic bias for reference alleles
- Widespread bias still exists at individual loci



Degner et al. 2009

Enhanced Reference Genome

- Add short SNP containing segments to the reference
- Need to have one segment per possible combination of SNPs in length-*r* region
- Avoids ambiguity by not including same length-*r* sequence twice
- Not represented as a graph, just additional sequences



Data

- Created masked and enhanced references from HapMap Yoruba SNPs
  - 3.7 million SNPs
  - Used at most first 5 SNPs in any length *r*-window
- Simulated reads at every position overlapping each SNP
  - Equal number of reference and non-reference alleles
  - Lengths of 37bp, 70bp, and 100bp
  - Added error with Bernoulli probability of 0.01, 0.02, or no error
- Used two real RNA-seq datasets from Degner et al.
  - 15.6 and 16.8 million 35bp reads from Yoruba lymphoblastoid cell lines

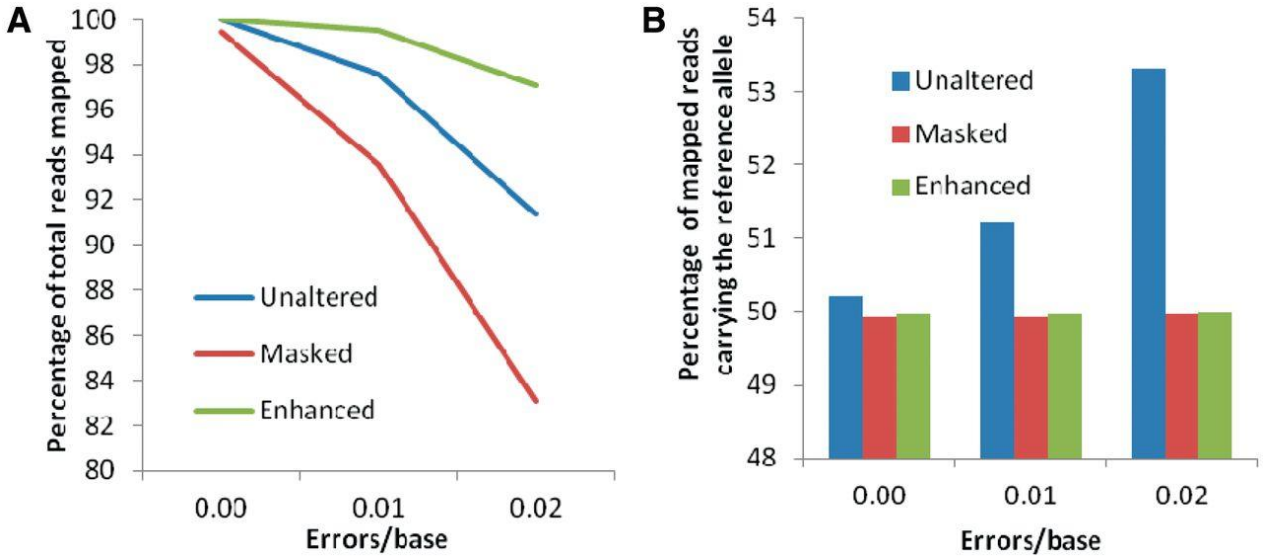
Enhanced Reference vs. hg38

- 275 Mbp worth of segments added to enhanced reference
  - 8.9% of original genome size
- hg38 contains 117 Mbp of “alt” and “random” scaffolds
  - Similar concept to enhanced reference
  - Far fewer scaffolds than enhanced segments, but much longer
  - Length-*r* segments not unique in hg38 - problematic

Simulated Reads - Overview

Masked reference has fewer reads aligned

Still, globally seems to have little reference bias

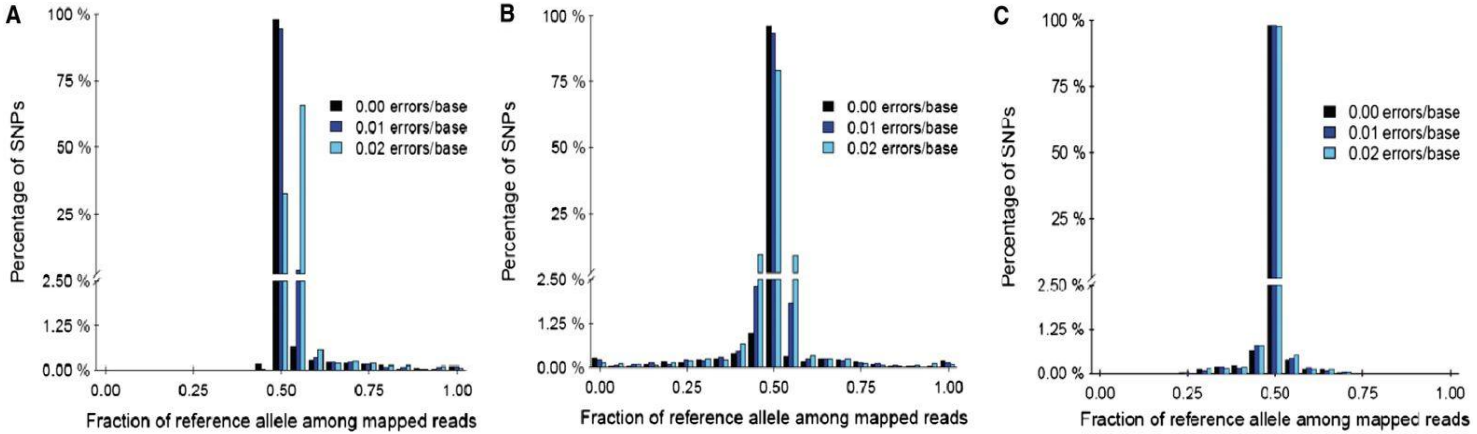


Aligned to each reference using MAQ

Simulated Reads - Individual Loci

Like before, masked reference has many loci with bias - more than unaltered

Enhanced reference has many fewer biased loci



Unaltered Reference **Masked Reference** **Enhanced Reference**

Real Data Summary

Individual	GM19238					GM19239				
Method	**Ref**	**Non-ref**	**Other**	**Total**	**Ref %**	**Ref**	**Non-ref**	**Other**	**Total**	**Ref %**
Unaltered	43 647	35 672	644	79 963	55.0	38 727	32 299	655	71 681	54.5
Masked	37 167	35 447	692	73 306	51.2	33 575	32 218	716	66 509	51.0
Enhanced	43 840	40 211	654	84 705	52.2	38 847	35 490	679	75 016	52.3

Summary of number of reads mapped with each allele using each method

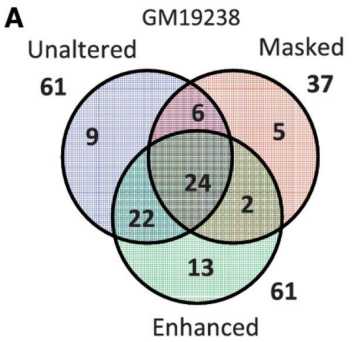
Highly Expressed Loci

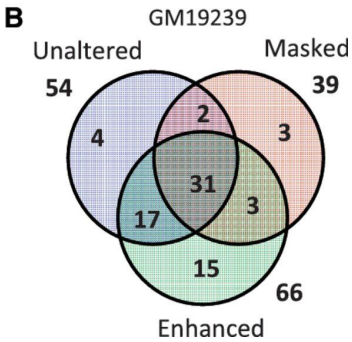
Individual	GM19238						GM19239					
Method	**Loci**	**Ref**	**Non-ref**	**Other**	**Total**	**Ref %**	**Loci**	**Ref**	**Non-ref**	**Other**	**Total**	**Ref %**
Unaltered	685	29 084	22 230	425	51 739	56.7	772	23 882	18 358	463	42 703	56.5
Masked	645	23 268	21 868	408	45 544	51.6	727	19 390	18 038	459	37 887	51.8
Enhanced	715	29 445	26 150	432	56 027	53.0	802	24 171	20 902	483	45 556	53.6

Number of reads mapped to loci with at least 20 reads mapped

Genes with ASE

Individual	GM19238			GM19239		
Method	**Ref**	**Non-ref**	**Total**	**Ref**	**Non-ref**	**Total**
Unaltered	55	6	61	43	12	54
Masked	24	13	37	21	18	39
Enhanced	40	21	61	38	28	66





Used binomial tests to see if ref or non-ref alleles were significantly expressed > 50%

Interesting Loci

			Unaltered			Masked			Enhanced						
SNP ID	**Chromosome**	**Position (hg 18)**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Gene**	**Unaltered Significant**	**Masked Significant**	**Enhanced Significant**
rs11619791	chr13	24568984	5	3	0	2	1	0	4	140	0	PABPC3	No	No	Yes
rs2814966	chr6	34820210	121	0	0	6	0	0	130	0	0	LOC646785	Yes	No	Yes
rs1042448	chr6	33162320	57	2	0	50	2	0	57	61	0	RPL32P1	Yes	Yes	No
rs1803621	chr12	6517370	2694	1690	18	2269	1695	21	2705	2541	18	LOC100133042	Yes	Yes	No

Conclusions

- Unaltered reference genomes have reference bias
- Masking SNPs with a third base does not fix bias, and can make it worse in many cases
- The enhanced reference genome reduces bias drastically