STAR+WASP reduces reference bias in the allele-specific mapping of RNA-seq reads

Rebecca Asiimwe[1], Dobin Alexander[1,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724, USA

*Corresponding author
†Current address:
Dr. Alexander Dobin
Cancer Center
Cold Spring Harbor Laboratory, Cold Spring Harbor, United States
dobin@cshl.edu

Abstract

Summary
Allele-specific expression is an important genetic phenomenon that impacts an individual's phenotype and is relevant in various biological and medical contexts. Next-generation RNA sequencing technologies provide an unprecedented opportunity to measure ASE genome-wide across all heterozygous alleles expressed in a given sample. One of the major obstacles to the accurate calculation of ASE from RNA-seq data is the reference mapping bias, i.e., the preferential misalignment of the reads to the reference allele. Here, we present STAR+WASP, our re-implementation of WASP, a highly accurate algorithm for reducing the reference bias (Van De Geijn *et al.* 2015). We show that STAR+WASP is an order of magnitude faster than WASP while significantly reducing reference bias and providing ASE estimations similar to the original WASP algorithm.

Availability and Implementation
STAR+WASP is implemented within STAR as an integrated C++ module. STAR+WASP is open-source software, freely accessible at: http://code.google.com/p/rna-star/.

Contact: dobin@cshl.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Introduction

Allele-specific expression (ASE) is a genetic phenomenon where the two alleles of a particular gene are not expressed equally in an individual's cells. The importance of allele-specific expression lies in its potential to influence an individual's phenotype and its relevance in various biological and medical contexts, such as genetic diversity and evolution, gene regulation, disease susceptibility, pharmacogenetics, etc. Next-generation RNA sequencing technologies provide an unprecedented opportunity to measure ASE genome-wide across all heterozygous alleles expressed in a given sample. Calculating allele-specific expression in RNA-seq data involves mapping reads to the reference genome, identifying reads that overlap heterozygous variants, assigning the reads to one of the alleles, and calculating allelic count imbalance (Castel *et al.* 2020). This approach was successfully utilized for multiple ASE-based analyses, such as identifying cis-regulated gene expression variations, nonsense-mediated mRNA decay, rare regulatory variation, and imprinting (Castel *et al.* 2015; Cosentino, Brink, and Siegel 2021).

One of the major obstacles to the accurate calculation of ASE from RNA-seq data is the mapping bias. Conventionally, quantification of ASE from RNA-seq data begins with the alignment of sequence reads to a linear reference genome, where each genomic position constitutes only the reference allele. This often leads to reference bias, i.e. preferential alignment to the reference allele and incorrect alignment of reads containing non-reference alleles (Degner *et al.* 2009; Stevenson, Coolon, and Wittkopp 2013). Although methods such as the use of consensus genomes (Kaminow *et al.* 2022), personalized genomes (Rozowsky *et al.* 2011; Groza *et al.* 2020), variant/SNP-aware aligners (Hach *et al.* 2014), graph-based aligners (Paten *et al.* 2017; Garg *et al.* 2018; Kim *et al.* 2019; Rakocevic *et al.* 2019; Rautiainen and Marschall 2020), multiple population reference genomes (Chen *et al.* 2021), and de novo assembly (Zerbino and Birney 2008), have been developed and applied to reduce reference bias, this challenge remains a substantial impediment in ASE studies.

To mitigate this problem, van de Geijn et al. (Van De Geijn *et al.* 2015) developed WASP (Workflow for Allele-Specific Processing), an allele-specific analysis method that eliminates reads with potential reference bias. WASP was shown to provide a highly accurate estimate of the ASE; however, it significantly increased computational costs. In this note, we present STAR+WASP, our re-implementation of the WASP algorithm built directly into the RNA-seq aligner STAR. We show that STAR+WASP is an order of magnitude faster than WASP while performing very similarly to the original WASP algorithm in estimating ASE and reducing the reference bias.

Algorithm

The WASP algorithm (Van De Geijn *et al.* 2015) is based on a simple but highly effective idea **(Figure 1a)**. First, all RNA-seq reads are mapped to the reference genome, and uniquely mapped reads overlapping heterozygous single-nucleotide variants are identified. The variant alleles are flipped for each of the variants in these reads: if a read sequence contains a reference (REF) base, it is switched to the alternative (ALT) base, and if a read contains an ALT base, it is switched to the REF base. Then, the modified sequence is remapped to the reference genome. If the modified read fails to map or maps to a different locus or multiple loci, the read is flagged as reference-biased and is excluded from the ASE calculations. The WASP method significantly reduced the reference bias in simulations and real data (Van De Geijn *et al.* 2015). However, these

improvements have come at the cost of substantially increased processing time and computational pipeline complexity. The main bottleneck of the WASP's original implementation is its multistep nature, which requires writing and reading BAM files twice. To mitigate this issue, we re-implemented the WASP algorithm inside our RNA-seq aligner STAR. We follow the WASP idea of remapping the reads with a flipped genotype; however, this is done on the fly immediately after mapping the original read sequence. This avoids the input/output bottlenecks and significantly increases the processing speed. STAR+WASP attaches the "vW" tag to the alignments in the BAM output file, which indicate whether alignments pass the WASP filtering, or the reason why they are rejected (**Supplementary Table 1**).

Benchmarks

Publicly available human RNA sequencing (RNA-seq) datasets were downloaded from the 1000 genomes (The 1000 Genomes Project Consortium *et al.* 2015) and ENCODE (The ENCODE (ENCyclopedia of DNA Elements) Project 2004) projects **(Supplementary Figure 1)**. The 16 samples included in this study represent various long RNA populations: polyA+, polyA-, and total RNA. The reads were aligned to the human reference genome (GRCh38 with Gencode v39 annotations) using STAR. The reads overlapping personal variants for each sample were filtered with WASP and STAR+WASP algorithms (see Supp. Methods).

First, we compared the reads that were deemed reference-biased and filtered out by WASP and STAR+WASP **(Supplementary Table 2, Supplementary Figures 2,3)**. Since the STAR+WASP algorithm was designed to resemble the WASP algorithm closely, STAR+WASP and WASP filtering results are very similar. The insignificant differences can be attributed to slight variations in filtering rules (e.g., the maximum allowed number of variants per read).

Next, we evaluated STAR+WASP performance in alleviating allelic bias. Overall, all samples **(Figure 1b)** had small but sizable proportions (1.4-6.5%) of reads that overlapped variants (these reads are labeled with a "vW" tag in the STAR+WASP BAM output). As expected, samples with longer read lengths (>90 million reads) had a higher percentage of reads overlapping variants. **Figure 1c** shows the proportions of reads that were deemed reference-biased and thus were rejected by WASP filtering for various reasons (labeled with different "vW" tag values ≠1). The main reason for rejection was the inability to remap reads after flipping the allelic variant, followed by those reads that were re-mapped to a different locus after variant flipping **(Figure 1c)**. Samples with shorter read lengths have higher rejection rates because shorter sequences are typically harder to align with high fidelity.

We further assessed the extent of reference bias introduced by aligning reads to a conventional linear reference (GRCh38) and evaluated the performance of STAR+WASP in reducing this bias. Since ground truth is unknown in real RNA-seq data, we use a proxy metric REF-ALT, defined as the difference between proportions of reads aligning to the reference and alternative alleles, averaged over all variants. Note that such differences represent allelic imbalances for each individual variant, but after averaging over all variants, the REF-ALT is expected to be close to zero. Without STAR+WASP filtering, the REF-ALT varies between 4% and 10% (median = 6.36%) for different samples **(Figure 1d, Supplementary Figure 4)**. Reads rejected by STAR+WASP filtering have large REF-ALT (44-80%) (median = 64.19%), confirming that the

filtering algorithm rejects reads with strong reference bias. Alignments that passed WASP filtering have substantially smaller REF-ALT (0-5%, media = 2.03%), demonstrating that the STAR+WASP algorithm can significantly reduce the reference bias.

To demonstrate STAR+WASP computational efficiency, we benchmarked STAR, WASP, and STAR+WASP mapping speed and random access memory (RAM) usage by computing the overall runtime and RAM usage for each mapping job run with a different number of threads **(Figure 1e, Supplementary Figure 5)** on a dedicated server. STAR+WASP alignments were considerably faster (6.5 to 10.5 times) than WASP. While STAR+WASP and WASP both use STAR for the read alignment to the genome, the on-the-fly implementation of the WASP algorithm in STAR+WASP allows for much faster re-mapping and filtering of the reads. The peak memory usage of STAR+WASP was only insignificantly higher (<0.2%) than that of STAR or WASP algorithms **(Supplementary Figure 6)**. We also ran these benchmarks for all samples and tools on a shared computer cluster, which resembles more closely real-life applications, and obtained very similar results **(Supplementary Figures 7 and 8)**.

Summary

Reference mapping bias is one of the major sources of errors in the estimation of genome-wide ASE from high-throughput RNA-seq data. We developed STAR+WASP, a re-implementation of the WASP reference bias reduction algorithm. We demonstrated that STAR+WASP produces results very similar to the original WASP tool and significantly reduces the reference bias. Since STAR+WASP is implemented within the STAR aligner codebase, it performs the filtering of reference-biased reads on the fly, allowing to decomplexify the computational pipeline, eliminate input/output bottlenecks, and increase the processing speed by an order of magnitude.

Acknowledgments

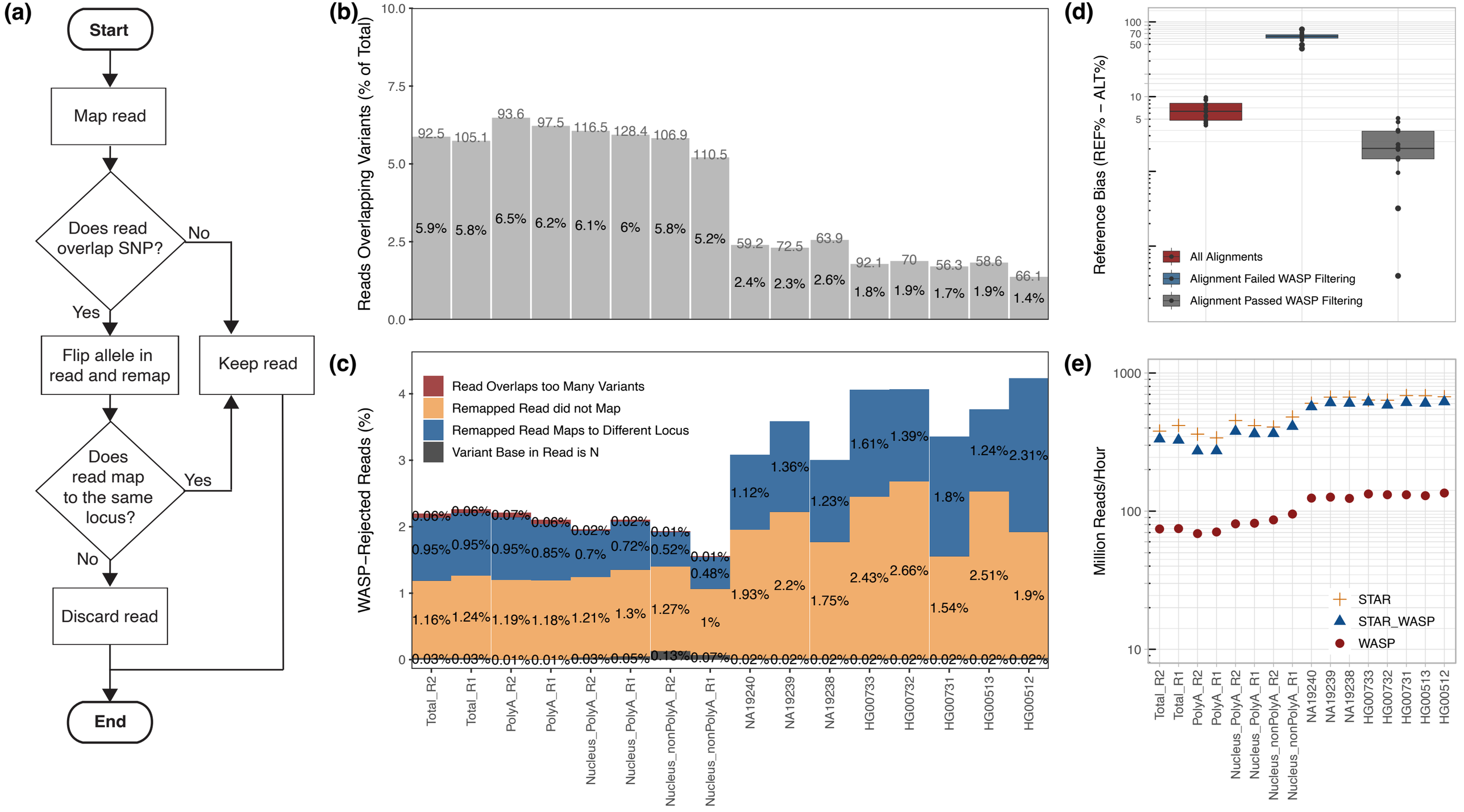
We are grateful to Bryce van de Geijn, Graham McVicker, and Jonathan Pritchard for their fruitful discussions about the WASP algorithm.

Figure 1 caption

(a) WASP's filtering workflow which was re-implemented in the STAR+WASP tool to reduce reference bias. Adopted from Bryce van de Geijn et al. (Van De Geijn *et al.* 2015). **(b)** Percentage of reads overlapping variants in each sample (read length is shown above each bar per sample). **(c)** Percentages of reads rejected by STAR+WASP filtering (out of reads overlapping variants) for each rejection cause. **(d)** Reference bias (REF-ALT) for unfiltered and filtered reads. **(e)** Mapping speed for STAR (no filtering), original WASP (with STAR as the mapper), and STAR+WASP algorithm.

References

1. Castel SE, Aguet F, Mohammadi P *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol* 2020;**21**:234.
2. Castel SE, Levy-Moonshine A, Mohammadi P *et al.* Tools and best practices for data processing in allelic expression analysis. *Genome Biol* 2015;**16**:195.
3. Chen N-C, Solomon B, Mun T *et al.* Reference flow: reducing reference bias using multiple population genomes. *Genome Biol* 2021;**22**:8.
4. Cosentino RO, Brink BG, Siegel TN. Allele-specific assembly of a eukaryotic genome corrects apparent frameshifts and reveals a lack of nonsense-mediated mRNA decay. *NAR Genomics Bioinforma* 2021;**3**:lqab082.
5. Degner JF, Marioni JC, Pai AA *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 2009;**25**:3207–12.
6. Garg S, Rautiainen M, Novak AM *et al.* A graph-based approach to diploid genome assembly. *Bioinformatics* 2018;**34**:i105–14.
7. Groza C, Kwan T, Soranzo N *et al.* Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol* 2020;**21**:124.
8. Hach F, Sarrafi I, Hormozdiari F *et al.* mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* 2014;**42**:W494–500.
9. Kaminow B, Ballouz S, Gillis J *et al.* Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. *Genome Res* 2022;**32**:738–49.
10. Kim D, Paggi JM, Park C *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15.
11. Paten B, Novak AM, Eizenga JM *et al.* Genome graphs and the evolution of genome inference. *Genome Res* 2017;**27**:665–76.
12. Rakocevic G, Semenyuk V, Lee W-P *et al.* Fast and accurate genomic analyses using genome graphs. *Nat Genet* 2019;**51**:354–62.
13. Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol* 2020;**21**:253.
14. Rozowsky J, Abyzov A, Wang J *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011;**7**:522.
15. Stevenson KR, Coolon JD, Wittkopp PJ. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* 2013;**14**:536.
16. The 1000 Genomes Project Consortium, Corresponding authors, Auton A *et al.* A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
17. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* 2004;**306**:636–40.
18. Van De Geijn B, McVicker G, Gilad Y *et al.* WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* 2015;**12**:1061–3.
19. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.


(a)
Start
Map read
Does read overlap SNP?
No
Yes
Flip allele in read and remap
Keep read
Does read map to the same locus?
Yes
No
Discard read
End
(b)
Reads Overlapping Variants (% of Total)
10.0
7.5
5.0
2.5
0.0
92.5
105.1
93.6
97.5
116.5
128.4
106.9
110.5
59.2
72.5
63.9
92.1
70
56.3
58.6
66.1
5.9%
5.8%
6.5%
6.2%
6.1%
6%
5.8%
5.2%
2.4%
2.3%
2.6%
1.8%
1.9%
1.7%
1.9%
1.4%
(c)
WASP−Rejected Reads (%)
Read Overlaps too Many Variants
Remapped Read did not Map
Remapped Read Maps to Different Locus
Variant Base in Read is N
Total_R2
Total_R1
PolyA_R2
PolyA_R1
Nucleus_PolyA_R2
Nucleus_PolyA_R1
Nucleus_nonPolyA_R2
Nucleus_nonPolyA_R1
NA19240
NA19239
NA19238
HG00733
HG00732
HG00731
HG00513
HG00512
(d)
Reference Bias (REF% − ALT%)
All Alignments
Alignment Failed WASP Filtering
Alignment Passed WASP Filtering
(e)
Million Reads/Hour
STAR
STAR_WASP
WASP