SOFTWARE TOOL ARTICLE

REVISED *SNPsplit*: Allele-specific splitting of alignments between genomes with known SNP genotypes [version 2; referees: 3 approved]

Felix Krueger, Simon R. Andrews

Bioinformatics Group, The Babraham Institute, Cambridge, UK

v2 **First published:** 23 Jun 2016, **5**:1479 (doi: 10.12688/f1000research.9037.1)
Latest published: 27 Jul 2016, **5**:1479 (doi: 10.12688/f1000research.9037.2)

Abstract

Sequencing reads overlapping polymorphic sites in diploid mammalian genomes may be assigned to one allele or the other. This holds the potential to detect gene expression, chromatin modifications, DNA methylation or nuclear interactions in an allele-specific fashion. SNPsplit is an allele-specific alignment sorter designed to read files in SAM/BAM format and determine the allelic origin of reads or read-pairs that cover known single nucleotide polymorphic (SNP) positions. For this to work libraries must have been aligned to a genome in which all known SNP positions were masked with the ambiguity base 'N' and aligned using a suitable mapping program such as Bowtie2, TopHat, STAR, HISAT2, HiCUP or Bismark. SNPsplit also provides an automated solution to generate N-masked reference genomes for hybrid mouse strains based on the variant call information provided by the Mouse Genomes Project. The unique ability of SNPsplit to work with various different kinds of sequencing data including RNA-Seq, ChIP-Seq, Bisulfite-Seq or Hi-C opens new avenues for the integrative exploration of allele-specific data.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
REVISED version 2 published 27 Jul 2016	✓		✓
version 1 published 23 Jun 2016	✓ report	✓ report	✓ report

1 **Andrew Keniry**, Walter and Eliza Hall Institute of Medical Research Australia

2 **Nicolas Servant**, Institut Curie France

3 **Prasoon Agarwal**, Karolinska Institutet Sweden

Discuss this article

Comments (0)

Corresponding author: Felix Krueger (felix.krueger@babraham.ac.uk)

How to cite this article: Krueger F and Andrews SR. ***SNPsplit*: Allele-specific splitting of alignments between genomes with known SNP genotypes [version 2; referees: 3 approved]** *F1000Research* 2016, **5**:1479 (doi: 10.12688/f1000research.9037.2)

Grant information: Research was supported by the Babraham Institute and the UK Biotechnology and Biological Sciences Research Council (BBSRC).

Competing interests: No competing interests were disclosed.

First published: 23 Jun 2016, **5**:1479 (doi: 10.12688/f1000research.9037.1)

REVISED **Amendments from Version 1**

This new version primarily adds a new Figure 1 to illustrate the process of generating N-masked genomes for single- or dual-hybrid genomes. We have also added a new section to the SNPsplit User Guide describing in more detail the process of filtering and processing high confidence SNPs so that the process can be adapted to other genomes more easily. We have also included a new paragraph in the manuscript acknowledging this improvement.

See referee reports

Introduction

Most functional NGS studies performed today still ignore the fact that many model organisms are diploid, and work on the averaged signal from the two alleles. However, a complete understanding of the biology of diploid organisms requires that the two alleles be measured separately. Allele-specific analysis of next-generation sequencing reads is becoming an important tool to identify events such as allele-specific expression of genes (ASE), allele-specific binding of transcription factors or histones (ASB) or allele-specific methylation (ASM). These techniques allow a more detailed investigation of the effects of genetic or epigenetic variation on genome regulation or studying parent of origin effects such as genomic imprinting or allelic imbalance.

There are two main use cases for the investigation of allele-specific events: If both parental genotypes are clean and known in advance, e.g. for defined crosses of inbred mouse strains, parent of origin specific effects can be studied by comparing the two parental genotypes. Alternatively, allele-specific analyses require the more complex procedure of whole genome haplotype reconstruction (e.g. as described in 1). For the purposes of this manuscript we will use the terms 'Allele 1' or 'Allele 2' to refer to the maternal or paternal genotype, respectively, or to a reference and alternative strain or genome if the distinction between maternal/paternal is not meaningful.

The detection of allele-specific events relies on the ability to distinguish the two alleles of a diploid organism, which can be accomplished by looking at reads covering heterozygous single nucleotide polymorphisms (SNPs), small insertions or deletions (InDels) or greater structural variations. While the allele-specific analysis of InDels has been found to be challenging[2], the use of SNPs to discriminate alleles is the most widely used approach because it allows for the maintenance of a common set of reference genome coordinates.

Several approaches have been taken to perform allele-specific alignments. The simplest is to align all reads to a single reference genome, but this introduces a bias as reads from the allele which is more similar to the reference are able to map more efficiently[3]. Another approach involves the generation of two personalised genomes by incorporating known SNP positions (and possibly InDels) followed by an alignment to both genomes and finally a post-processing step to compute the union of the separate alignments (used in different flavours in 4–6). This approach is slower as it requires two separate mapping steps, and can still result in allelic bias because reads from one allele might not map uniquely or to an incorrect location in one of the genomes[3]. A more recent improvement[7] aims to reduce mapping biases by first aligning reads to the reference genome, then realigning reads that overlap SNP positions in all possible allele combinations and keeping only reads that align to the same position regardless of their genotypes - this reduces bias, but is computationally complex. Finally, the issue of bias can be tackled by masking polymorphic sites with the ambiguity nucleobase 'N' (henceforth called 'N-masking'), performing a single alignment to the N-masked genome and then assigning reads based on the sequence found underneath the masked positions. The rationale for N-masking in allele-specific alignments is that the mapping bias towards the reference allele is eliminated and both alleles of the same read get placed in the same position in the genome equally well. N-masking the genome is a one-off exercise and this approach has the advantage of requiring only a single alignment to a reference which noticeably reduces the computational load. Despite the fact that N-masking effectively avoids allelic biases it may occasionally result in a minor loss of sensitivity when the density of N covered by a read is getting too high.

A requirement for N-masking is that SNP positions are known, e.g. via a public resource like the Mouse Genomes Project which provides high quality variant calls for a large number of mouse strains[8] (hosted at http://www.sanger.ac.uk/science/data/mouse-genomes-project). If the genotype is not known, SNP positions may be called from the data itself, or from genome re-sequencing performed in parallel. The quality of the genotype calls is crucial for allele assignment, so the genotype data needs to be collected carefully and quality control and filtering is required to avoid biases and false positive hits[9]. Further downstream analysis of allele-specific data is highly dependent on the experiment type and is beyond the scope of this manuscript.

To our knowledge there are currently no user-friendly solutions available for the allele-specific splitting of sequencing reads aligned to N-masked genomes. We sought to address this by creating SNPsplit, an easy-to-use tool for assigning allele-specifc reads. In its generic mode SNPsplit is not tied to any particular aligner and operates across several different experiment types including RNA-Seq, genomic DNA-alignments, DNA methylation (Bisulfite-Seq) and 3-D genome organisation (Hi-C). While a similar allele-specific functionality has been integrated into specialised applications, e.g. HiC-Pro[10], the unique capability to work with several different data types renders SNPsplit an ideal choice for correlation studies using allele-specific sequencing reads.

Methods

Implementation

SNPsplit is written in Perl and consists of three separate scripts that can be run individually on the command line. It takes alignment files in BAM/SAM format as input and further requires an annotation file containing the positions of all SNPs in the genome. SNPsplit determines for each aligned read whether it overlaps with a known SNP position and adds a tag to the alignment that indicates whether the read can be assigned to a specific allele

or is unassignable. The reads are then sorted into different sub-files depending on the library type, i.e. single-end or paired-end, and the nature of the sample, e.g. RNA-Seq, BS-Seq or Hi-C.

Generating N-masked genomes

As long as a SAM/BAM file that was aligned to an N-masked genome is provided as input SNPsplit should perform well regardless of how the N-masking itself was accomplished. Since there is an ever growing number of genomes and different SNP annotation files and file formats it would be too much to ask to provide a generally applicable way of constructing N-masked genomes that fits all cases.

We do however provide an automated solution to generate N-masked versions of the genome for all strains in the Mouse Genomes Project (http://www.sanger.ac.uk/science/data/mouse-genomes-project). The genome preparation step supports the generation of single hybrid strains where one allele is the same as the mouse reference sequence (which is based on strain C57BL/6J, hereafter called Black 6) and one alternative allele, e.g. SPRET/EiJ. It also supports the generation of dual hybrid strains where both alleles are different from the Black 6 reference, e.g. CAST/EiJ and 129S1/SvImJ. At the time of writing the Mouse Genomes Project encompassed variation information for 36 different mouse strains; the SNP annotation data for all strains relative to Black 6 reference sequence may be found in the variant call format (VCF) file 'mgp.v5.merged.snps_all.dbSNP142.vcf.gz' (VCF v4.2; last modified 13 May 2015; download available at: ftp://ftp-mouse.sanger.ac.uk/current_snps/). The SNPsplit genome preparation first reads the SNP annotations for the strain in question from the VCF file and then constructs the N-masked genomes based on the Black 6 reference sequence using only high confidence homozygous positions. The process is slightly different for single or dual hybrid strains (Figure 1).

Single-Hybrid Strains. This generates a new genome sequence, with SNPs either N-masked or included as full sequence, where Allele 1 (or Genome 1) is the Black 6 reference and Allele 2 (or Genome 2) is the alternative strain.

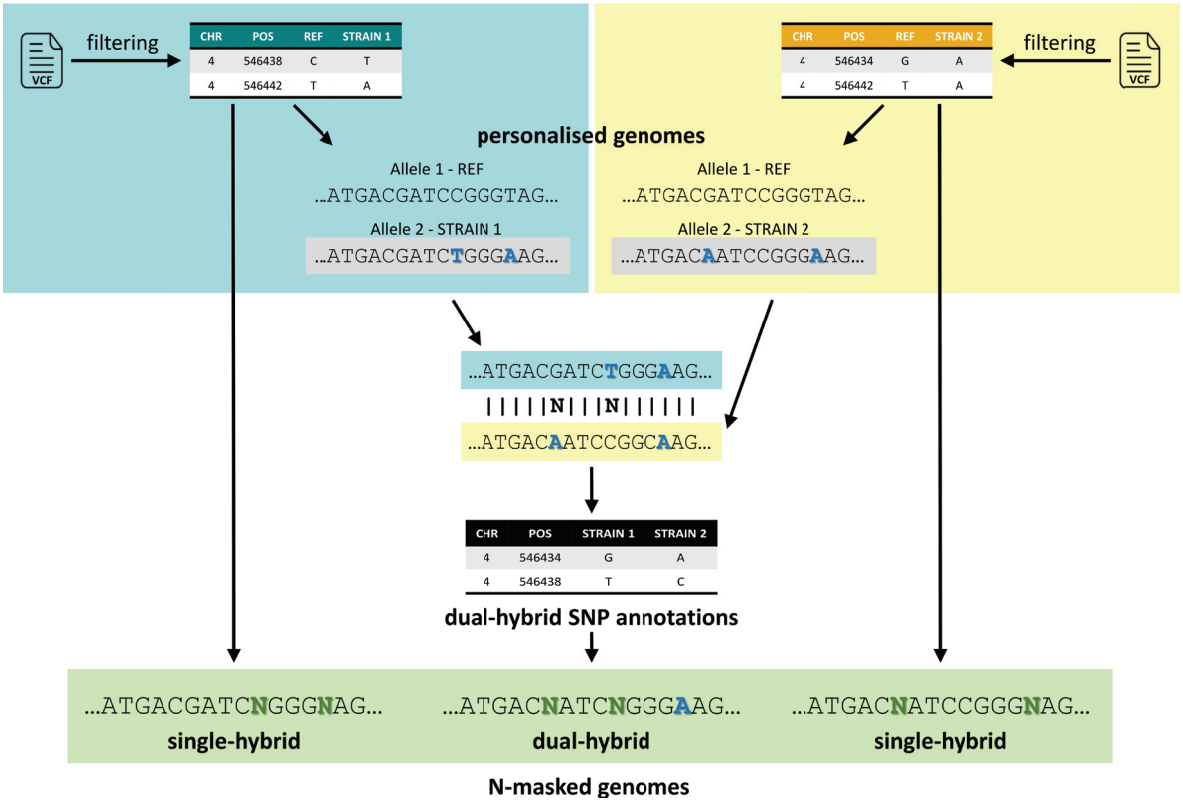

Figure 1. Generating N-masked or personalised genomes. Single-hybrid strains require SNP information/filtering from only one strain (Strain 1, blue box). SNP positions against the referene genome (REF) can either be masked by Ns (N-masked genome) or incorporated unmasked as full sequence (personalized genome). Dual hybrid strains require the SNP information/filtering also from a second strain (Strain 2, yellow box). The SNP information of Strain 1 and Strain 2 are then compared to create dual-hybrid SNP annotations (note that some positions where both strains had the same variation relative to the reference genome are no longer regarded as SNP and are now missing in the new annotations, e.g. A at position 546442). The dual-hybrid annotations are then used to N-mask SNP positions using the Strain 1 genome as new reference. The N-masked or personalised genomes for Strain 2 are technically not required to generate dual hybrid genomes but may be written out for convenience reasons.

1) The VCF file is read and filtered for high-confidence SNPs for the strain specified

2) The Black 6 reference genome is read into memory, and the filtered high-confidence SNP positions are incorporated either as N-masking (default) or full sequence (optional)

Dual-Hybrid Strains. This generates a new genome sequence where neither allele is the Black 6 reference. SNPs can be either N-masked or included as full sequence, where Allele 1 (or Genome 1) is the strain specified as strain 1 and Allele 2 (or Genome 2) is the strain specified as strain 2.

1) The VCF file is read and filtered for high-confidence SNPs in strain 1

2) The Black 6 reference genome is read into memory, and the filtered high-confidence SNP positions are incorporated as full sequence and N-masking (optional)

3) The VCF file is read and filtered for high-confidence SNPs in strain 2

4) The filtered high-confidence SNP positions of strain 2 are incorporated as full sequence and N-masking (optional)

5) The SNP information of strain 1 and strain 2 relative to the Black 6 reference genome build are compared and a new Ref/SNP annotation is constructed whereby the new Ref/SNP information will be strain 1/strain 2

6) The full genome sequence of strain 1 is read into memory, and the high-confidence SNP positions between strain 1 and strain 2 are incorporated as full sequence and N-masking (optional)

The N-masked sequences (or sequences containing the full sequence SNPs) are written out in FASTA format and ready to be indexed with the alignment software of your choice. Alignments to N-masked genomes are not very different to regular mapping except that they require the aligner to support ambiguity DNA bases such as N. Software confirmed to be working for this approach include (but are not limited to) Bowtie2[11], BWA[12], HISAT2[13], STAR[14] or any tool wrapping one of these aligners.

Even though the automated genome preparation is optimised to work with the VCF file from the Mouse Genomes Project the process can be easily adapted to work with any other genome as long as the genotypes are known and well defined. The SNPsplit manual provides more detailed information about the SNP filtering from VCF files and which entries are required to make it work also with other genomes.

Running SNPsplit on aligned files

SNPsplit operates in two stages which are run sequentially: I) read tagging and II) read sorting. Both steps generate detailed reports for record keeping.

Stage I: Tagging SNPsplit analyses reads for overlaps with known SNP positions for which it requires the mismatch position field (MD:Z:) in the SAM entry, and writes out a tagged BAM file in the same order as the original file. This process requires a list of all known SNP positions between the two different genomes (supplied as a SNP file) and works on a read-by-read basis.

Read tagging generally works as a multi-step process:

1. Determine the position(s) in the read that overlap genomic N(s)
2. Adjusting position for insertions/deletions
3. Determine equivalent genomic position
4. Determine if the SNP is present in the list of SNP positions, and if yes whether the position in the read was the Allele 1 or Allele 2 base

Depending on the collected SNP information the tagging module then determines whether a read can be assigned to a certain allele and appends an additional optional field 'XX:Z:tag' to the SAM entry of each read. The tag can be one of the following:

- UA - Unassigned
- G1 - Genome 1-specific (Allele 1, the reference)
- G2 - Genome 2-specific (Allele 2, the alternative strain)
- CF - Conflicting

Reads are considered unassignable (UA) if they do not overlap any known SNP position. Reads harbouring at least one SNP specific for both genomes at the same time are classified as conflicting (CF).

The determination of overlaps is geared to handle the CIGAR operations M (match to the reference), D (deletion in the read), I (insertion in the read) and N (skipped regions, used for splice mapping). Other CIGAR operations (see the SAM format specification for further details[15]) are currently not supported. This means that SNPsplit requires reads to be a full match from end-to-end and thus soft-clipping (CIGAR operation: S), which may introduce artefactual alignments to poorly annotated regions in the genome[16] is not supported (see also section Use Cases for RNA-Seq below on how to avoid soft-clipping issues).

Stage II: Sorting The tagged BAM file is read in again and sorted into allele-specific files according to their XX:Z: tag. For paired-end or Hi-C experiments the combination of tags for both Read 1 and Read 2 are considered (see below for examples). Conflicting reads, or also disagreeing read-pairs for paired-end samples, are not printed out by default. The sorting process may also be run stand-alone on tagged BAM files to try out different sorting options (e.g. separating out paired-end and singleton alignments or enabling reporting of conflicting alignments).

Operation

SNPsplit runs on any Linux-based operating system with Perl installed (tested using CentOS v6.2 and Perl v5.10.1). In addition, a functional version of SAMtools[15] (v0.1.18 or later) is required

for handling of SAM/BAM files. Memory requirements depend directly on the genome size and the total number of heterozygous SNPs to be stored, but as a guideline 5–10 GB RAM should be sufficient to process data for most mouse strains.

Use cases

Standard genomic DNA alignments

SNPsplit is able to handle any kind of standard genomic alignment file irrespective of the method employed to generate the library as long as the CIGAR operation requirements are met (see Stage I: Tagging above). A non-exhaustive list of supported applications includes genome re-sequencing, histone or protein ChIP-Seq (chromatin immunoprecipitation sequencing) or ATAC-Seq (Assay for Transposase-Accessible Chromatin by sequencing).

A use case of ChIP-Seq for the transcription factor ZFP57 is shown in Figure 2 (data re-analysed from 17). Alignments to an N-masked reference genome were performed for reciprocal crosses between Black 6 and Cast/EiJ mice using Bowtie 2, followed by SNPsplit sorting. This process was able to identify allele-specific binding of ZFP57 to several different imprinting control regions in a parental origin-specific manner, exemplified for the SNRPN locus in Figure 2.

RNA-Seq

In addition to standard linear alignments with or without small InDels, SNPsplit also handles spliced read alignments containing large gaps (CIGAR operation: N), such as reads spanning exon boundaries in RNA-Seq experiments. Spliced read aligners that have successfully been used for allele-specific alignments in conjunction with SNPsplit include Tophat[18], STAR[14] (Spliced Transcripts Alignment to a Reference) and HISAT2[13]. To work smoothly together with SNPsplit, HISAT2 and STAR require the user to disable soft-clipping which is performed by default (CIGAR operation: S), and STAR also needs to be instructed to print out the mismatch position (MD:Z:) field. More detailed instructions may be found in the SNPsplit User Guide.

Hi-C

As a variant of the chromatin conformation capture assay Hi-C is a proximity-ligation based assay which allows the investigation of the three-dimensional structure of the genome by massively parallel sequencing[19]. This is accomplished by measuring the frequency at which different parts of the genome sequence come into close physical contact. While standard Hi-C cannot discriminate whether an interacting fragment originated from the same or the other allele, allele-specific interaction maps can separate

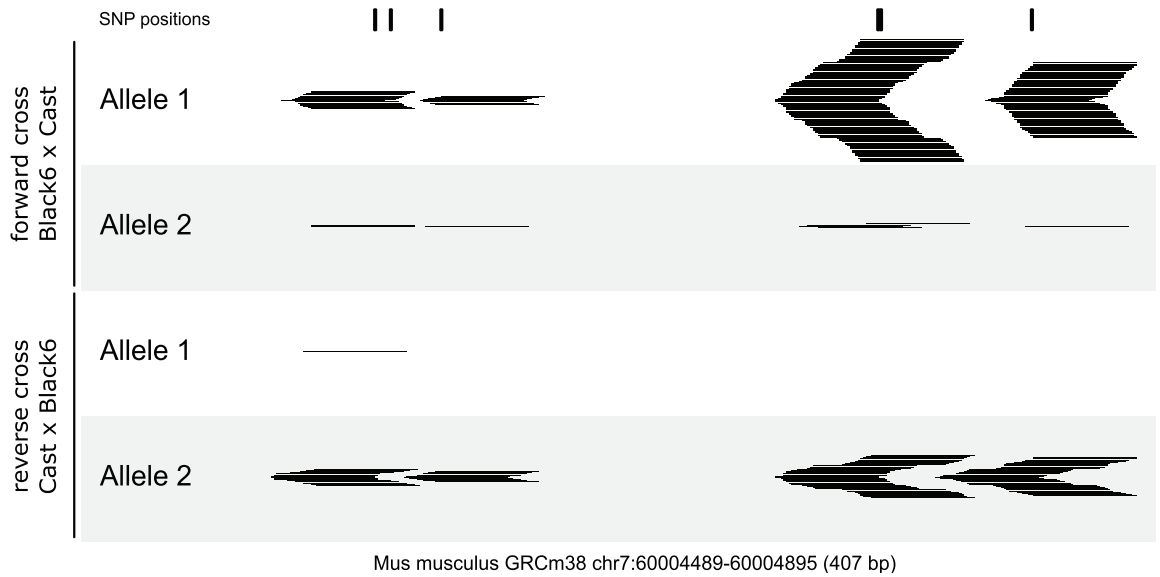

Figure 2. ChIP-Seq for the transcription factor ZFP57 identifies parental-origin allele-specific binding at the differentially methylated region (DMR) of the SNRPN locus. The binding of ZFB57 is methylation dependent and can be found exclusively on the maternal allele (genetic background of mother in forward cross: Black 6; mother in reverse cross: Cast). SNP positions were N-masked and used for allele-specific splitting of sequencing reads (shown as horizontal lines in black). Allele 1: Black 6 reference. Allele 2: Cast/EiJ strain (Cast). The area shown depicts the DMR only in part. Data taken from 17 (GEO accession: GSE55382).

cis-allele from trans-allele interactions, thereby greatly improving the analysis of chromatin dynamics and gene regulation[20,21].

The Hi-C mode of SNPsplit assumes that the input data is in the Hi-C format produced by the HiCUP pipeline[22], i.e. the input BAM files are by definition paired-end and Read 1 and Read 2 follow each other. It discriminates several additional read combinations to distinguish between cis- and trans-allele interactions:

- G1-G1
- G2-G2
- G1-UA
- G2-UA
- G1-G2
- UA-UA

For mixed allele groups such as G1-G2 there is no need to create the reverse group (G2-G1) since Hi-C interactions have no directionality. Again, read pairs containing at least one conflicting read (tag: CF) are not printed out by default, but this may be optionally enabled.

Bisulfite-Seq

Bisulfite sequencing is a method to interrogate DNA methylation patterns using the chemical properties of sodium bisulfite to convert cytosines to uracil but leaving methylated cytosines largely unaffected.

The bisulfite mode of SNPsplit assumes that the input data has been processed with the bisulfite alignment tool Bismark[23]. SNPsplit runs a quick check at the start of a run to see if the file provided appears to be a Bismark file, and sets the appropriate flags for bisulfite and/or paired mode automatically. Paired-end mode requires Read 1 and Read 2 of a pair to follow each other in consecutive lines so the BAM file will be sorted by read name if necessary.

Utilisation of SNP positions and allele assignment of bisulfite treated reads In contrast to the standard mode, C>T SNPs may not always be used for allele-specific sorting in a bisulfite setting since they could either be a genuine SNP or rather reflect the methylation state. Since the majority of known SNPs actually involves C to T transitions (due to spontaneous deamination of methylated CpG dinucleotides), the ability to assign aligned bisulfite treated reads is thus somewhat reduced compared to regular DNA-based alignments. The number of SNP positions that have been skipped because of this bisulfite ambiguity is documented in the report file.

Positions requiring special treatment include all of the following Allele 1/Allele 2 combinations: C/T or T/C for forward strand alignments and G/A or A/G for reverse strand alignments. These positions may however be used to assign opposing strand alignments since they do not involve C to T transitions directly. For that reason, the bisulfite call processing also extracts the bisulfite strand information from the alignments in addition to the basecall at the position involved. For any SNPs involving C positions that are not C to T SNPs both methylation states, i.e. C and T, are allowed to match the C position.

For SNPs which were masked by Ns in the genome no methylation call will have been performed during the alignment step, i.e. they will receive a '.' (dot) in the methylation call string. This means that SNP positions themselves may be used for allele-sorting but do not participate in calling methylation. While this reduces slightly the number of total methylation calls it effectively eliminates the problem of assigning potentially incorrect methylation states to these positions.

To demonstrate the effectiveness of sorting bisulfite treated reads we reprocessed publicly available bisulfite sequencing data from reciprocal mouse crosses reported by Xie and colleagues[6] (GEO accession: GSE33722). First we generated a dual hybrid genome for 129X1/SvJ (129) (as near-enough relative we used the SNP annotations for strain 129S1/SvImJ) and Cast/EiJ (Cast) mice, and then aligned the data to the N-masked genome using Bismark (v0.16.1, default parameters, read trimming was performed using Trim Galore[24] v0.4.1, default parameters). The data was then processed with SNPsplit and all datasets for the F1 forward cross 129 (mother) x Cast (father), and F1 reverse cross Cast (mother) x 129 (father) were merged and analysed using SeqMonk (v.0.33.0[25]) (Figure 3).

Whilst the majority of the genome shows very similar methylation levels on both alleles of the hybrid mice, this approach also allows the detection of allele-specific methylation events. This can be readily spotted at imprinted loci where one parental allele is fully methylated while the other remains completely unmethylated. The Gnas/Nespas locus in the mouse genome shows both a paternally methylated region (more upstream) and maternally methylated region (more downstream) where the allele-specific methylation pattern is maintained in a parent-of-origin dependent manner (Figure 3). This demonstrates that the combination of bisulfite mapping and read sorting by SNPsplit is an effective tool to identify allele-specific methylation in diploid genomes.

Summary

Analysing next-generation sequencing data in an allele-specific fashion holds the potential to uncover regulatory events or mechanisms that would otherwise be obscured in bulk data. SNPsplit is designed to enable researchers to quickly and easily perform allele-specific analysis of their sequencing data as long as the SNP genotypes of the organism in question are known. For hybrid mouse strains covered by the Mouse Genomes Project, SNPsplit offers an easy solution from generating N-masked genomes to allele-specific sorting of reads without requiring the user to possess excessive computational skills. SNPsplit is not tied to any specific application and indeed it has been used already to answer questions for a variety of different data types such as ChIP-Seq, RNA-Seq, Bisulfite-Seq and Hi-C. This gives SNPsplit the unique capability of bringing together allele-specific data including gene-expression, DNA methylation, genomic accessibility or architecture which holds great potential for studying genome regulation.

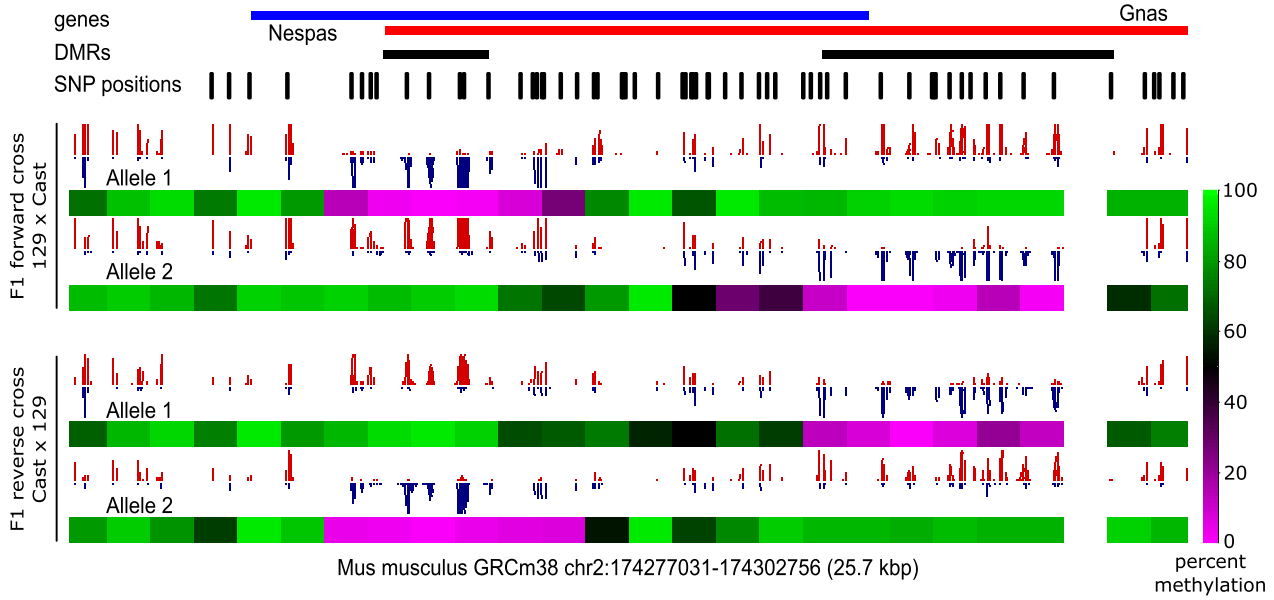

Figure 3. Allele-specific methylation at differentially methylated regions (DMRs) is maintained in a parent-of-origin specific way at the Gnas/Nespas locus. The upstream DMR is methylated exclusively on the paternal allele, while the more downstream DMR is methylated exclusively on the maternal allele in both forward (129 x Cast) and reverse crossed (Cast x 129) hybrid mice. SNP positions were N-masked and used for allele-specific sorting with SNPsplit. Allele 1: 129X1/SvJ reference (129). Allele 2: Cast/EiJ strain (Cast). Red or blue dots in the graph represent calls for methylated or unmethylated cytosines, respectively (CpG context only). The percentage methylation was determined for 2000 bp windows for the region shown using the Bisulfite Methylation Pipeline in Seqmonk[25] (default options). Data taken from 6, GEO accession: GSE33722.

Software availability

1. Software available from: http://www.bioinformatics.babraham.ac.uk/projects/SNPsplit/

2. Latest source code: https://github.com/FelixKrueger/SNPsplit

3. Archived source code as at time of publication: https://zenodo.org/record/55477#.V18PoDb93ww[26]

4. Software license: GNU GPL v3 or later

Author contributions

FK designed and wrote SNPsplit and the manuscript, SRA was involved in study design and wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Research was supported by the Babraham Institute and the UK Biotechnology and Biological Sciences Research Council (BBSRC).

References

1. Selvaraj S, Dixon JR, Bansal V, *et al.*: **Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing.** *Nat Biotechnol.* 2013; **31**(12): 1111–1118.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Rivas MA, Pirinen M, Conrad DF, *et al.*: **Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome.** *Science.* 2015; **348**(6235): 666–669.
PubMed Abstract | Publisher Full Text | Free Full Text

3. 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**(24): 3207–3212.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Crowley JJ, Zhabotynsky V, Sun W, *et al.*: **Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance.** *Nat Genet.* 2015; **47**(4): 353–360.
PubMed Abstract | Publisher Full Text | Free Full Text

5. 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**(1): 522.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Xie W, Barr CL, Kim A, *et al.*: **Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome.** *Cell.* 2012; **148**(4): 816–831.
PubMed Abstract | Publisher Full Text | Free Full Text

7. 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**(11): 1061–1063.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Keane TM, Goodstadt L, Danecek P, *et al.*: **Mouse genomic variation and its**

effect on phenotypes and gene regulation. *Nature.* 2011; **477**(7364): 289–294.
PubMed Abstract | Publisher Full Text | Free Full Text

9. 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.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Servant N, Varoquaux N, Lajoie BR, *et al.*: **HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biol.* 2015; **16**: 259.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–595.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods.* 2015; **12**(4): 357–360.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

16. **QC Fail.** 2016.
Reference Source

17. Strogantsev R, Krueger F, Yamazawa K, *et al.*: **Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression.** *Genome Biol.* 2015; **16**(1): 112.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Kim D, Pertea G, Trapnell C, *et al.*: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol.* 2013; **14**(4): R36.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Lieberman-Aiden E, van Berkum NL, Williams L, *et al.*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science.* 2009; **326**(5950): 289–293.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Dixon JR, Jung I, Selvaraj S, *et al.*: **Chromatin architecture reorganization during stem cell differentiation.** *Nature.* 2015; **518**(7539): 331–336.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Rao SS, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text

22. Wingett S, Ewels P, Furlan-Magaril M, *et al.*: **HiCUP: pipeline for mapping and processing Hi-C data [version 1; referees: 2 approved, 1 approved with reservations].** *F1000Res.* 2015; **4**: 1310.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics.* 2011; **27**(11): 1571–1572.
PubMed Abstract | Publisher Full Text | Free Full Text

24. **Trim Galore.**
Reference Source

25. **SeqMonk.**
Reference Source

26. Krueger F: **SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes.** *Zenodo.* 2014.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 2

Referee Report 01 August 2016

doi:10.5256/f1000research.9980.r15257

Andrew Keniry
Walter and Eliza Hall Institute of Medical Research, Parkville, Vic, Australia

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 27 July 2016

doi:10.5256/f1000research.9980.r15258

Prasoon Agarwal
Division of Clinical Immunology, Department of Laboratory Medicine, Karolinska Institutet, Huddinge, Sweden

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 08 July 2016

doi:10.5256/f1000research.9725.r14556

Prasoon Agarwal
Division of Clinical Immunology, Department of Laboratory Medicine, Karolinska Institutet, Huddinge, Sweden

This article by Krueger F *et al*. describes in detail a new software SNPsplit which is capable of sorting reads or read–pairs in the allele specific modus that covers know single nucleotide polymorphic (SNP), or single nucleotide variation (SNV) locations. For sorting the reads the inputs to the software are a 'N' masked genome, created using known SNP or SNV locations, to which the reads are aligned using any available aligner and a VCF file containing the positions of all SNPs. The authors claim that this is the only user-friendly solution available for the allele-specific splitting of sequencing reads aligned to N-masked genomes. Overall the manuscript is very well written. However, I have some minor queries regarding the software performance:

1. Masking known SNP positions in the genome sequence eliminated the reference bias but, in case of heterozygous SNPs there could be a chance of having significant bias toward higher mapping rates of the allele in the reference sequence, is there any provision in the software to remove this noise and bias from mapped reads? This kind of bias can lead to false signal of allelic imbalance.

2. It is stated that the software can construct the N-masked genomes, so is it restricted to mouse alone or can be extended in case of humans or other species?

3. Can SNPsplit be used to split reads for indels and deletions?

I have personally used the software for ATAC-seq data from patients and it works perfectly fine. I have not used the genome builder module of the software. The data looks perfect in the UCSC genome browser for the 'Allele 1' or 'Allele 2' and the unassigned. Overall I felt it is a very user friendly software available.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 18 Jul 2016

Felix Krueger, The Babraham Institute, UK

We would like to thank the reviewer for their kind and approving comments about SNPsplit. Some specific comments may be found below.

Masking known SNP positions in the genome sequence eliminated the reference bias but, in case of heterozygous SNPs there could be a chance of having significant bias toward higher mapping rates of the allele in the reference sequence, is there any provision in the software to remove this noise and bias from mapped reads? This kind of bias can lead to false signal of allelic imbalance.

SNPsplit is really designed for mapping against genomes with clean parental genotypes; this is also why the genome preparation step is very strict and only uses high quality homozygous parental SNPs (see the new section on this in the SNPsplit manual). If the SNPs are heterozygous already in one or even both of the parental strains we propose that these positions could be masked by Ns for the mapping step, but not be used at all for the allele-specific read assignment as such. For the allele-specific splitting, i.e. the file supplied to SNPsplit with --snp_file, only homozygous high quality SNP positions should be used.

It is stated that the software can construct the N-masked genomes, so is it restricted to mouse alone or can be extended in case of humans or other species?

The SNPsplit approach does in theory work on any N-masked genome irrespective of the species. The preparation is currently optimised to work the VCF file provided by the Mouse Genomes Project but as long as the VCF file conforms to the same standards it should also work for other genomes. If the VCF file looks different, e.g. you got the file from a collaborator and it came without header lines or format description one would have to adapt the relevant section(s) in the preparation script. We have now added a section to the SNPsplit User Guide that explains in more detail the basis of SNP filtering and which parameters are needed to run properly. We hope that this will help generating N-masked genomes when the SNP data looks different than the file provided by Mouse Genomes Project.

Can SNPsplit be used to split reads for indels and deletions?

SNPsplit supports the processing of reads that contain indels, but indels themselves are not used to sort reads to different genomes at the current time. The reason for this is mainly that indel annotations are often more tricky to come by and alignments over indels (e.g. at the ends of reads) are notoriously more difficult. There are currently no plans to extend the functionality to include indels, but we might look into this in the future if the demand arises.

Competing Interests: No competing interests were disclosed.

Referee Report 01 July 2016

doi:10.5256/f1000research.9725.r14557

Nicolas Servant
Institut Curie, Paris, France

This manuscript by Krueger and collaborators describes SNPsplit, an alignment sorter for allele specific analysis. SNPsplit is designed to work on N-masked alignment and provides additional utilities to generate the appropriate reference.

The main interest of SNPsplit is its ability to operate across different experiment types and alignment software. It therefore allows to analyse and to integrate in an allele specific and unbiased manner heterogeneous dataset.

The manuscript is well written, and is divided in two main parts. The first part presents the software and its implementation and the second part, presents user cases.

Other than my few comments below, I am happy with the manuscript and want to note that I have successfully downloaded and used SNPsplit in the context of several projects.

1. The way the N-masked genome is generated based on SNPs information is not that easy to understand for non expert users. I would recommend a figure to help in understanding this point.

   In the context of Dual-Hybrid strain, the interest of first generating the strain 1 (S1) and the strain 2 (S2) genomes is not clear to me.

If I'm correct, the final masked genome will be generated from the S1 fasta reference only (unmasked). If we use a simple example with a SNP which is Bl6=A, S1=T, S2=T, using this strategy will avoid a mismatch at this position, compared to the strategy of masking only heterozygous SNPs between S1/S2 on the reference (Bl6) genome. But what is the interest of generating the S2 reference genome (step 4) ? Using S1 or S2 should be enough?

2. How paired-end sequencing is managed in practice ? The authors present an application with Hi-C data but is it the same for any NGS application?

3. The authors present SNPsplit in the context of the Mouse Genome Project. I'm wondering how difficult it would be to transpose it to any genome and organism as long as the genotype is known. A few words about that in the manuscript would be interesting.

4. In the Introduction, the authors mentioned the WASP software[7]. To avoid any misunderstanding with the parental genome strategy, I would suggest something like; “A more recent improvement aims to reduce mapping biases by first aligning reads on the reference genome, then realigning reads that overlap SNP positions in all possible allele combinations and keeping only reads that align to the same position regardless their genotypes - this reduces bias, but is computationally complex”

5. At the end of the introduction, the authors mentioned that “SNPsplit is not tied to any particular aligner” which is not exactly true as the authors explain later that “SNPsplit can be used as long as the CIGAR operation requirements are met”

6. Typo:

   “The determination of overlaps is geared” in **Stage I: Tagging**
   Reference 26 and 27 doesn't seem to be used in the text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 18 Jul 2016

Felix Krueger, The Babraham Institute, UK

Many thanks for the constructive review and the thoughtful suggestions on how to improve the manuscript. Please find our point-by-point responses below.

The way the N-masked genome is generated based on SNPs information is not that easy to understand for non expert users. I would recommend a figure to help in understanding this point.

We have now added a new Figure 1 that aims to explain the process of generating single or dual hybrid N-masked genomes in more detail. We feel that this new figure makes the whole process substantially easier to understand, many thanks for suggestion!

In the context of Dual-Hybrid strain, the interest of first generating the strain 1 (S1) and

the strain 2 (S2) genomes is not clear to me.
If I'm correct, the final masked genome will be generated from the S1 fasta reference only (unmasked). If we use a simple example with a SNP which is Bl6=A, S1=T, S2=T, using this strategy will avoid a mismatch at this position, compared to the strategy of masking only heterozygous SNPs between S1/S2 on the reference (Bl6) genome. But what is the interest of generating the S2 reference genome (step 4) ? Using S1 or S2 should be enough?

Technically it would be sufficient to use the Strain 1 unmasked genome as new reference and then use the SNP annotations of both strains to compute a new Strain1/Strain2 SNP annotation file (see also the new Figure 1). Since writing out a new genome only takes a few seconds and it might be of potential use for other projects we do also write out versions of Strain 2, it can always be removed later if desired.

How paired-end sequencing is managed in practice? The authors present an application with Hi-C data but is it the same for any NGS application?

We do mention in the manuscript that the paired-end mode uses both ends for the genome-specific assignments even though it is not as complicated as the Hi-C mode. The SNPsplit User Guide already contains more detailed information about paired-end file handling and also provided paired-end sorting reports to illustrate this, so we would kindly refer the user to the manual rather than adding an extra section here.

The authors present SNPsplit in the context of the Mouse Genome Project. I'm wondering how difficult it would be to transpose it to any genome and organism as long as the genotype is known. A few words about that in the manuscript would be interesting.

Indeed, as long as the genotypes are well defined and the list of SNP positions is known the genome preparation should also work well with any other genome (see also the reply to Prasoon Agarwal's comment). To enable this process we have added a new section to the SNPsplit User Guide that explains the SNP filtering and processing in more detail, and we have added a short paragraph about this to the manuscript.

In the Introduction, the authors mentioned the WASP software[7]. To avoid any misunderstanding with the parental genome strategy, I would suggest something like; "A more recent improvement aims to reduce mapping biases by first aligning reads <u>on the reference genome</u>, then realigning reads that overlap SNP positions in all possible allele combinations and keeping only reads that align to the same position <u>regardless their genotypes</u> - this reduces bias, but is computationally complex"

We have changed this sentence to make it clearer.

At the end of the introduction, the authors mentioned that "SNPsplit is not tied to any particular aligner" which is not exactly true as the authors explain later that "SNPsplit can be used as long as the CIGAR operation requirements are met"

At least in its generic mode SNPsplit is not tied to any specific aligner, but this does not necessarily mean that every single option of any aligner will be supported. To this end, SNPsplit was initially designed to work with Bowtie2 and Tophat, but we have later also seen other software such as

STAR or HISAT2 work fine as well. Aligner-specific comments can be found in the SNPsplit User guide.

Reference 26 and 27 doesn't seem to be used in the text

Thanks for spotting this, we have now removed these references from the bibliography.

Competing Interests: No competing interests were disclosed.

Referee Report 01 July 2016

doi:10.5256/f1000research.9725.r14558

Andrew Keniry
Walter and Eliza Hall Institute of Medical Research, Parkville, Vic, Australia

Krueger and Andrews report SNPsplit, a tool for sorting mapped next generation sequencing reads into separate files depending on the allelic origin of the read. The process requires genetic heterogeneity between the alleles such that the reads can be assigned to a particular allele based on known SNP positions. Such a tool allows for allele specific analysis and becomes useful for studies on phenomena such as genomic imprinting, allelic imbalance and X-chromosome inactivation. SNPsplit improves on current methods for analysis of allele specific sequencing reads by providing built in tools for the analysis of bisulfite and HiC data, which are otherwise more complicated to analyse. A tool for creating an N-masked genome is also provided, which overcomes mapping bias towards the reference genome.

To date I have used SNPsplit to process data from RNA-seq, bisulfite-seq, transcription factor ChIP-seq and histone ChIP-seq. I have not tested the built in N-masked genome creator. SNPsplit has proven to be easy to use and very stable in my run environment. Typically, a sample is processed in approximately 3 hours depending on read depth. As far as I can tell, by assessing known imprinted genes and the silent female X chromosome, SNPsplit does an excellent job of assigning reads to the correct genome, with known phenomena appearing as expected. The built in option for analysis of bisulfite data works very well on reads mapped with the bismark program.

I've found SNPsplit to work very well for all the data types I have used it for: RNA-seq, bisulfite-seq, transcription factor ChIP-seq and histone ChIP-seq. It should be noted however that due to the requirement for a SNP to be present in the read, assignable reads from narrow transcription factor ChIP-seq peaks can be sparse and some peaks may not assignable at all. This is an unavoidable limitation, and the authors show successful assignment of transcription factor ChIP-seq reads in Figure 2, however this is perhaps something researchers should consider. Reads deriving from broad histone ChIP-seq peaks suffer no such limitation.

I have no issues with the performance of SNPsplit, and neither the accuracy of how the tool is presented in the paper. Perhaps the authors could include some details that will aid researchers in experimental design? For example, for a typical SNP density what percentage of reads would be assigned to each genome, unassigned and conflicting? This would help researchers in estimating required read depths for their particular question. Could the authors also explain how vcf files are filtered for high confidence SNPs when preparing the N-masked genome? A discussion of post processing techniques for the removal of

incorrectly annotated SNPs would also be beneficial. Perhaps the benefit of longer read length for assigning reads to a particular genome could also be mentioned?

In summary, SNPsplit will prove to be a very useful tool for the analysis of epigenomic data from next generation sequencing experiments.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 18 Jul 2016

Felix Krueger, The Babraham Institute, UK

We would like to thank the reviewer for their kind and positive feedback. Please find our specific replies below.

Could the authors also explain how vcf files are filtered for high confidence SNPs when preparing the N-masked genome?

We have expanded the SNPsplit User Guide considerably to now include a detailed section on how the SNP filtering is accomplished and which of the parameters are required for the process to work with other VCF files. We hope that this will make the process easier to understand and allow adapting the procedure for other genomes as well.

Perhaps the authors could include some details that will aid researchers in experimental design? For example, for a typical SNP density what percentage of reads would be assigned to each genome, unassigned and conflicting? This would help researchers in estimating required read depths for their particular question. A discussion of post processing techniques for the removal of incorrectly annotated SNPs would also be beneficial. Perhaps the benefit of longer read length for assigning reads to a particular genome could also be mentioned?

It would be undoubtedly helpful to provide some more guidance about experimental design, however we are not sure that this should be added in the context of software manuscript. As a general guideline the percentage of reads that can be assigned allele specifically increases proportionally with the number of heterozygous SNPs present between two strains, and increasing the read length also increases the chances to hit a SNP. Furthermore paired-end data can be assigned with a much rate than single-end reads. The SNPsplit User Guide contains a few example reports to help users to get an idea about typical values. We also feel that post-processing techniques probably warrant more discussion than just being briefly mentioned here, a good paper to read to get started is also cited in the manuscript (Castel et al.).

Competing Interests: No competing interests were disclosed.