Bioinformatics, 37(15), 2021, 2215–2217
doi: 10.1093/bioinformatics/btaa936
Advance Access Publication Date: 2 November 2020
Applications Note

Genetics and population analysis

read_haps: using read haplotypes to detect same species contamination in DNA sequences

Hannes P. Eggertsson[1] **and Bjarni V. Halldorsson**[1,2,*]

[1]deCODE Genetics, Reykjavík 102, Iceland and [2]Department of Engineering, School of Technology, Reykjavík University, Reykjavík 102, Iceland

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on February 10, 2020; revised on September 10, 2020; editorial decision on October 20, 2020; accepted on October 22, 2020

Abstract

Motivation: Data analysis is requisite on reliable data. In genetics this includes verifying that the sample is not contaminated with another, a problem ubiquitous in biology.

Results: In human, and other diploid species, DNA contamination from the same species can be found by the presence of three haplotypes between polymorphic SNPs. read_haps is a tool that detects sample contamination from short read whole genome sequencing data.

Availabilityand implementation: github.com/DecodeGenetics/read_haps.

Contact: bjarni.halldorsson@decode.is

1 Introduction

Short read sequencing (SRS) is commonly used to study human genetic variation. Reads that overlap polymorphic positions in the genome can be used to determine polymorphisms, most frequently SNPs and indels. Reads or read pairs that overlap two SNPs can be used to assign alleles to haplotypes (Halldórsson *et al.*, 2002; Lippert, 2002). In human and other diploid species, it should be possible to assign the alleles of two SNPs carried to exactly two haplotypes.

A number of reasons exist why more than two haplotypes may be observed in sequence data; Sequencing error is ubiquitous at the base pair level, reads are commonly incorrectly mapped to the reference genome, parts of the genome are variable in copy number and finally the sample may be contaminated with DNA from another sample.

Contamination is commonly detected with the tool verifyBamID Jun et al. (2012), which estimates contamination using a genotype likelihood model. Contamination may be incorrectly estimated, particularly for admixed samples. We developed read_haps to detect contamination without relying on population genetic assumptions. Vanquish is based on principles similar to verifyBamID Jiang *et al.* (2019) and ConFindr Low et al. (2019) is a tool similar to read_haps trained on bacterial data.

2 Methodology

read_haps takes as input a sam/bam/cram file of reads aligned to a reference genome, a VCF file of variants called and a set of reliably genotyped markers. read_haps starts by finding all heterozygous markers that are reliably genotyped in the VCF file, by default conditioning the genotypes on those that have a PHRED-scaled minimum likelihood of 40 (presumed error rate at most 10^{-4}) for a genotype alternate from the most likely genotype.

Only biallelic SNPs are considered and we refer to the marker alleles as 0 and 1. A two SNP haplotype can be written as xy where x is the allele at the first marker and y is the allele at the second marker. Two heterozygous markers can have phase 00-11 (parity) or 01-10 (non-parity), i.e. for parity phasing the individual has two haplotypes one with allele 0 on both markers and the other with allele 1 on both markers. Observing both parity and non-parity suggests the presence of three haplotypes between the two markers, but can also be explained by genotyping and sequencing error.

To limit the impact of genotyping error we examined the sequencing data of a 15.220 individuals (Jónsson et al., 2017), including 1548 trios and determined a set of SNPs that showed low levels of inheritance errors and could generally be reliably genotyped across samples. A detailed description of this filtering is found in the Supplementary Note of Halldorsson *et al.* (2016) under the subheading ‘Description and filtering of sequencing data’.

To limit mapping and sequencing errors, only reads with high mapQ and bases with high bp quality are considered. Three haplotypes should generally not be observed in a sample from a single diploid individual, but may occur in regions of structural variation. Three haplotypes are commonly observed in a sample that contains the mixture of the DNA of two individuals. Pairs of heterozygous markers overlapped by at least two read pairs in parity and at least two reads pairs in non-parity are considered evidence for three haplotypes at the marker pair. Samples with multiple such pairs are considered contaminated, by default at most 0.2% of marker pairs can have two reads showing both parity and non-parity.

read_haps is implemented in C++ using Seqan (Döring *et al.*, 2008) and relies on htslib (Li *et al.*, 2009).

Table 1. Definition of output from read_haps

Column	Type	Explanation
SNP_PAIRS	int	Number of read phasable SNP pairs
ERROR_PAIRS	int	SNP pairs with at least one read showing parity and non-parity
DOUBLE_ERROR_PAIR_COUNT	int	SNP pairs with at least two reads showing parity and non-parity
DOUBLE_ERROR_FRACTION	float	DOUBLE_ERROR_PAIR_COUNT/SNP_PAIRS
REL_ERROR_FRACTION	float	ERROR_PAIRS/SNP_PAIRS
NONSENSE_FRACTION	float	Average fraction of reads representing base pairs different from the two alleles
PASS_FAIL	bool	PASS/FAIL
REASON		COVERAGE or CONTAMINATION

Table 2. Effect of contaminating a sample on read_haps and verifyBamID output

Contamination (%)	read_haps				verifyBamID			
	Mean	SD	Median	Failed (%)	Mean	SD	Median	Failed (%)
0	0.00003	0.00004	0.00002	0	0.00080	0.00027	0.00077	0
1	0.00030	0.00014	0.00026	0	0.01041	0.00250	0.01077	0
2	0.00074	0.00035	0.00068	0	0.01797	0.00570	0.01971	45
3	0.00156	0.00070	0.00149	27	0.02786	0.00843	0.02985	79
4	0.00251	0.00094	0.00239	73	0.03883	0.00972	0.03919	92
5	0.00334	0.00137	0.00336	74	0.04483	0.01321	0.04793	97
10	0.00980	0.00321	0.01005	100	0.09293	0.02592	0.09537	100
20	0.02111	0.00753	0.02262	100	0.18036	0.04387	0.18374	100

Note: Contamination is simulated at 1,2,3,4,5,10 and 20% in 100 simulations. Results for error rate of 0% are computed by subsampling to 30X each of 7 files 100 times. Mean, sd and median values of DOUBLE_ERROR_FRACTION and FREEMIX are reported, respectively for read_haps and verifyBamID. Failure rate is computed using a threshold of 0.002 and 0.02 for read_haps and verifyBamID, respectively.

3 Usage

In addition to previously defined inputs, read_haps also requires a fasta file containing the reference genome (defaults to genome.fa). A set of reliable markers is given along with the software. All tests are performed and tuned to Illumina WGS data, assuming −30× sequencing coverage human genome data mapped to GRCh38 using BWA mem (Li, 2013). Single sample genotyping results have been tested using GATK (McKenna et al., 2010) and GraphTyper (Eggertsson *et al.*, 2017). The program has been tested with the default options for selecting which reads and variants to consider, but a number of options can also be set by the user.

A definition of the program output is given in Table 1. A sample can be given the flag 'FAIL' when either very few SNP pairs can be phased, generally due to low coverage sequencing, or a sample contamination is suspected.

The level of contamination at each marker pair can be output using the option '–pairs'. This will point to marker pairs where there exists evidence of both parity and non-parity phasing.

4 Results

We ran read_haps, using the default DOUBLE_ERROR_THRESHOLD of 0.002 (0.2%), on seven samples from the Genome in a Bottle consortium (GiaB). These samples have been sequenced to a high coverage and were subsampled to 30× to represent a more typical dataset.

We simulated (Table 2) sample contamination by creating a bam file that is a mix of bam file sequences from two individuals. The simulations were done 100 times, each time randomly selecting two files from the set of 7, with the contaminating sample constituting 1, 2, 3, 4, 5, 10 and 20% of the total sample. The results for read_haps and verifyBamID are presented in Table 2. For comparison purposes, the results for the 7 samples, without contamination, are presented as contamination of 0% in Table 2, where we subsample each file 100 times to 30× coverage.

Both programs fail all samples with 10% contamination or higher in our simulations. Both DOUBLE_ERROR_FRACTION and FREEMIX were elevated at all levels in our simulations (Table 2).

Acknowledgement

The authors thank collaborators at deCODE genetics.

Financial Support: none declared.

Conflict of Interest: Authors are employees of deCODE genetics.

References

Döring,A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Eggertsson,H.P. *et al.* (2017) Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.*, **49**, 1654–1660.

Halldórsson,B.V. *et al.* (2002) A survey of computational methods for determining haplotypes. In: Istrail, S., Waterman, M., Clark, A. (eds.) *RECOMB Workshop on Computational Methods for SNPs and Haplotype Inference*. Springer, Berlin, Heidelberg, pp. 26–47.

Halldorsson,B.V. *et al.* (2016) The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.*, **48**, 1377–1384.

Jiang,T. *et al.* (2019) Same-species contamination detection with variant calling information from next generation sequencing. *bioRxiv*, 531558.

Jónsson,H. *et al.* (2017) Whole genome characterization of sequence diversity of 15,220 icelanders. *Sci. Data*, **4**, 170115.

Jun,G. *et al.* (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.*, **91**, 839–848.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv:1303.3997*.

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lippert,R. (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinf.*, **3**, 23–31.

Low,A.J. *et al.* (2019) Confindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ*, 7, e6995.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.