PILLAR BIOSCIENCES
Pillar Biosciences, Inc. 9 Strathmore Rd Natick, MA 01760
(800) 514-9307 | www.pillarbiosci.com

Detecting Contamination in Cell-free DNA Sequencing Libraries using Germline Markers

Andrew B Conley, Lavanya Rishishwar, Nachen Yang, Zonghan Liu, Brian Wright, Sean M Polvino, Yue Ke, Zhaohui Wang

INTRODUCTION

Introduction. Sample-to-sample contamination within sequencing batches can introduce false positive calls in cell-free DNA (cfDNA) assays, where variant allele fractions (VAFs) are already low (<0.5%). While not all sources of contamination are detectable post hoc, contamination among individuals can be uncovered by tracking genotype discordance across samples. We present a panel-agnostic Bayesian framework to infer contamination fractions in cfDNA libraries, leveraging germline single nucleotide polymorphisms (SNPs) from secondary analysis outputs. The method applies to any panel with adequate germline SNP coverage and operates independently of variant calling or library preparation strategy.

Methods. The algorithm works by 1) identifying SNPs across all samples in the sequencing batch, 2) genotyping each sample, 3) identifying candidate contaminant samples that share observed alleles for each homozygous SNP site in each sample, and 4) computing posterior probability that observed read counts between samples are a result of contamination versus stochastic noise. Posterior probabilities are computed across a range of fractions using the joint likelihood of observed VAF and zygosity of the implicated donor; a nonzero posterior indicates probable contamination. To evaluate performance, we simulated contaminated samples by mixing cfDNA sequencing data from a two-gene panel at known contamination fractions. We also constructed contrived contaminated libraries using cfDNA from cell lines mixed at defined ratios and evaluated CNV-positive tumor samples. Libraries were sequenced using panel-specific protocols, and secondary analysis was performed with PiVAT®.

Results. We evaluated in silico mixes, contrived mixes, and CNV-positive tumor samples for contamination. In silico mixes of healthy samples showed sensitive contamination detection down to 1% contamination when using 30 homozygous SNPs. Contrived cell line mixes (1-10% contaminant) were accordingly subsampled to 30 SNPs. Subsampled data had a median of 7 detectable contaminant SNPs, and the contaminant was detected 92% of the time. Pure cell lines showed 100% specificity. CNVs in tumor samples can yield false positive contamination calls by altering the frequency of germline SNPs. After downsampling the CNV-positive samples to 30 SNPs, our approach achieved 99% specificity.

Conclusion. Contamination is a persistent challenge in cfDNA sequencing, particularly in low-VAF contexts. We present a robust, panel-agnostic approach for contamination detection using germline markers and Bayesian inference. This scalable and generalizable framework enhances quality assurance in cfDNA sequencing and can inform the development of future post-analytical contamination surveillance methods.

MATERIALS AND METHODS

- We developed a method for identifying within-batch cross-sample contamination (Figure 1) in cfDNA panels
- The method uses germline SNPs and their VAF to detect and identify potential contamination
- Samples in a batch are genotyped and sites with concerning VAF (Off-SNPs) are identified
- The extent of potential contamination is inferred from SNP VAFs and genotypes

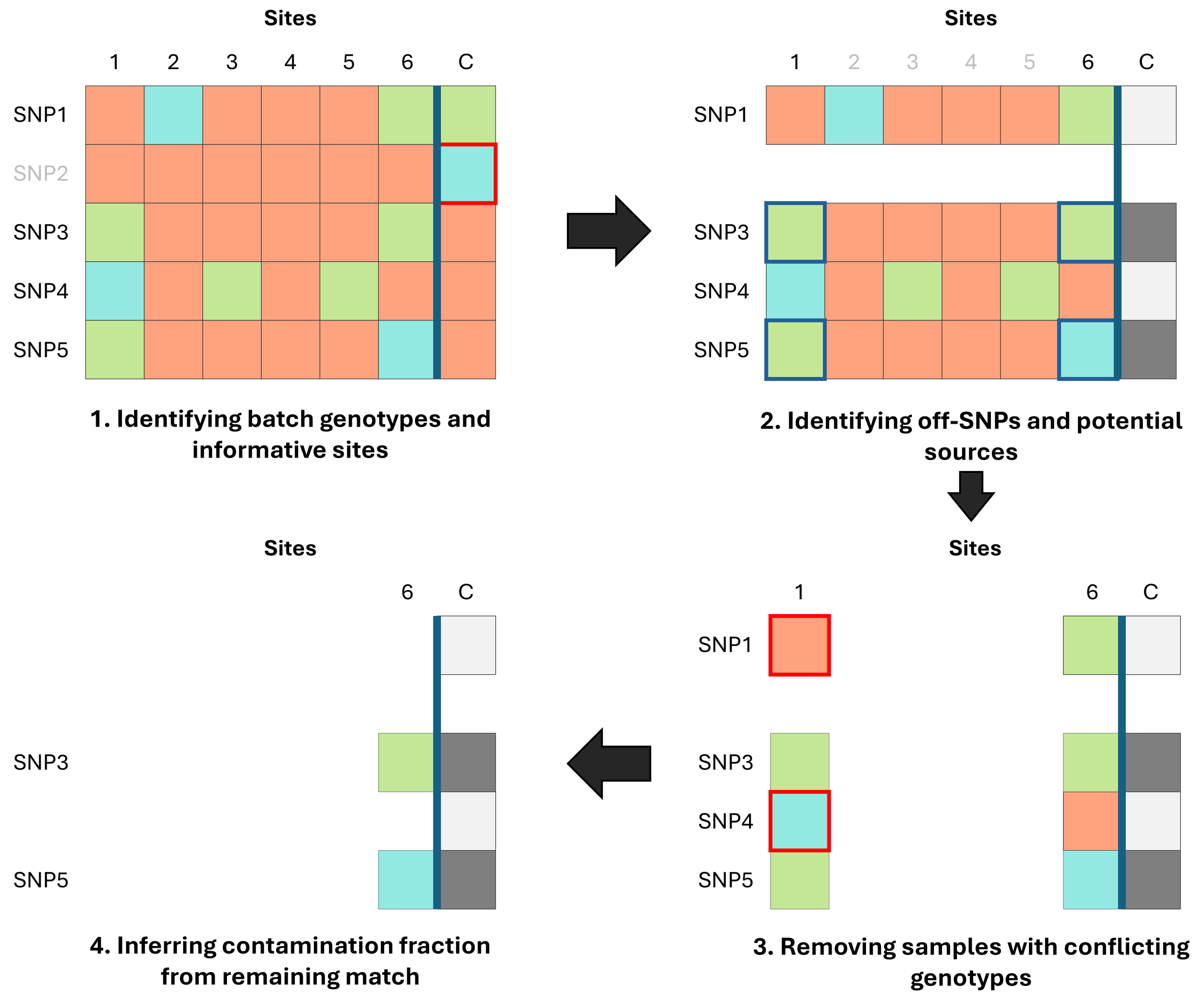


Figure 1. Withing-batch contamination detection workflow. Germline SNPs are used to identify informative sites and to include or exclude potentially contaminating within-batch samples

RESULTS

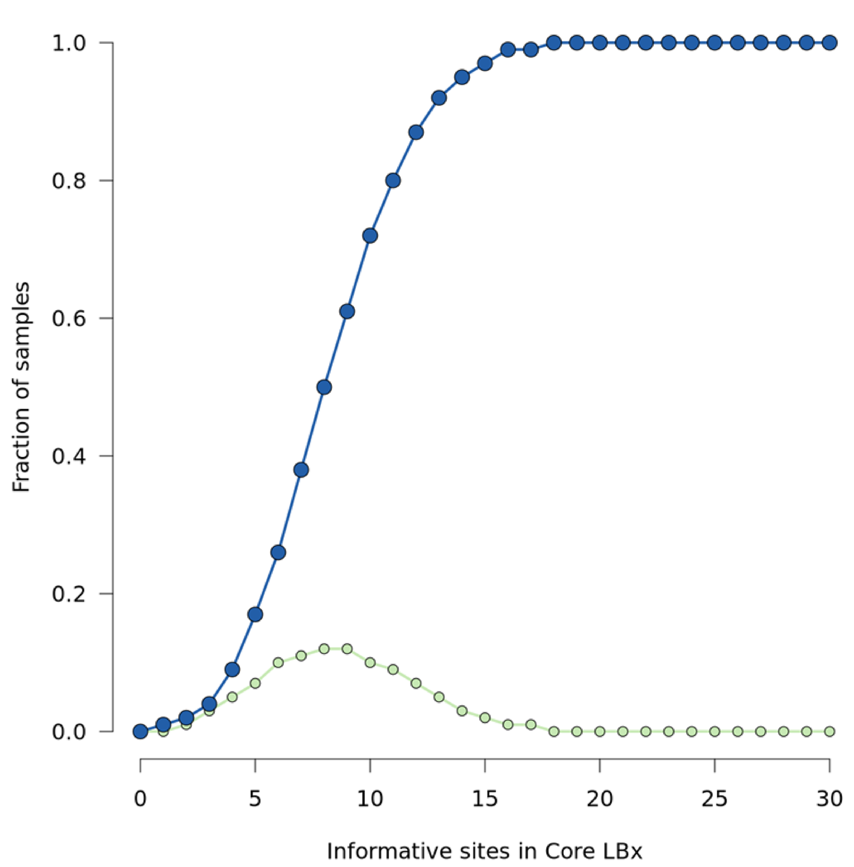
- We first tested our approach in a small series of contrived sample-to-sample mixes (Table 1) in a 150 SNP amplicon panel. Our approach was able to identify contamination in each sample
- To evaluate the broader applicability, we downsampled the number of SNPs to 30 and carried out simulated mixes (Figure 2, Table 2)
- Downsampling the contrived samples to 30 SNPs reduced sensitivity to contamination detection, but maintained specificity (Table 3)
- CNVs can cause shifts in SNP VAFs similar to real contamination. In downsampled cell lines and samples with CNVs, we were able to distinguish between a CNV and contamination ~99% of the time (Table 4)

Table 1. Contamination calls and metrics of contrived contamination mixes in a 150 SNP panel.

Background	Contamination fraction	Total SNPs	# Off	Off SNPs?	Has match?	Called?
Sample 1	0	100	2	FALSE	TRUE	FALSE
Sample 1	1	100	30	TRUE	TRUE	TRUE
Sample 1	2.5	100	32	TRUE	TRUE	TRUE
Sample 1	5	100	35	TRUE	TRUE	TRUE
Sample 1	10	100	34	TRUE	TRUE	TRUE
Sampel 2	0	112	0	FALSE	TRUE	FALSE
Sampel 2	2.5	112	46	TRUE	TRUE	TRUE
Sampel 2	5	112	47	TRUE	TRUE	TRUE
Sampel 2	10	112	46	TRUE	TRUE	TRUE

Contamination detection over different fractions

Figure 2 & Table 2. Left - Number of informative SNPs between individuals using 30 sites from 1KGP. Right – Contamination detection from *in silico* mixes.



Contamination fraction	Called contaminated	Correct contaminant
0.0	0	0
0.5	796	144
1.0	917	599
2.5	917	903
5	917	903
10	931	917
15	917	903
20	917	903
25	944	931
30	972	958
35	983	983
40	1000	1000
45	1000	1000
50	1000	1000

RESULTS

Impact of Downsampling on detection accuracy

Table 3. Downsampling and contamination detection in contrived contaminated mixes. Detection is less sensitive but maintains specificity.

Background	Contamination fraction	Median off SNPs	Off SNPs?	Match?	Contamination call
Sample 1	0	0	0	0.968	0
Sample 1	1	6	0.86	0.916	0.784
Sample 1	2.5	6	0.896	0.976	0.872
Sample 1	5	6	0.916	1	0.916
Sample 1	10	6	0.916	1	0.916
Sample 2	0	0	0	1	0
Sample 2	2.5	8	0.992	1	0.992
Sample 2	5	8	0.984	1	0.984
Sample 2	10	8	0.988	1	0.988

Table 4. Downsampling and contamination detection in cell lines and samples with and without CNVs. CNV-positive samples show off SNPs but low FP rate.

Sample	CNV?	Median off SNPs	Off SNPs?	Match?	Contamination call
497	No	3	0.232	0.228	0
508-497-A1	No	3	0.26	0.304	0
BT474	No	1	0	0.052	0
DLD-1	No	0	0	0.184	0
H1975	No	1	0	0	0
H596	No	1	0	0.016	0
JAK2-Pos-01	No	0	0	0.128	0
K562	No	1	0	0.004	0
MDA-MB-231	No	1	0	0.32	0
MDA-MB-453	No	1	0	0.004	0
S11-3313-B1-DS	No	2	0.112	0.444	0.004
C32	Yes	13	1	0.04	0.008
SYD17-10	Yes	9	1	0.1	0.020
SYD17-16	Yes	8	0.964	0.104	0.012

CONCLUSIONS

- We developed a specific and sensitive method for calling contamination in small, cfDNA amplicon panels using germline SNPs
- Our method is robust to shifts in germline SNP VAF caused by CNVs
- Our method functions well with a small number of germline variants (30) though benefits from additional SNPs
- Including germline SNP targets in an amplicon panel allows for robust contamination detection