

## Same-Species Contamination Detection with Variant Calling Information from Next Generation Sequencing

Tao Jiang <sup>1,2,3</sup>, Martin Buchkovich <sup>3#</sup> and Alison Motsinger-Reif <sup>1,2#\*</sup>

<sup>1</sup> *Department of Statistics, North Carolina State University, Raleigh, NC 27695 USA*

<sup>2</sup> *Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695 USA*

<sup>3</sup> *Translational Genomics Department, Q2 Solutions / EA Genomics, a Quintiles Quest Joint Venture, Morrisville, NC 27560 USA*

\*To whom correspondence should be addressed. Email: [alison.motsinger@gmail.com](mailto:alison.motsinger@gmail.com)

#Equal contribution

## Abstract

**Motivation:** Same-species contamination detection is an important quality control step in genetic data analysis. Compared with widely discussed cross-species contamination, same-species contamination is more challenging to detect, and there is a scarcity of methods to detect and correct for this quality control issue. Same-species contamination may be due to contamination by lab technicians or samples from other contributors. Here, we introduce a novel machine learning algorithm to detect same species contamination in next generation sequence data using support vector machines. Our approach uniquely detects such contamination using variant calling information stored in the variant call format (VCF) files (either DNA or RNA), and importantly can differentiate between same species contamination and mixtures of tumor and normal cells.

**Methods:** In the first stage of our approach, a change-point detection method is used to identify copy number variations or copy number aberrations (CNVs or CNAs) for filtering prior to testing for contamination. Next, single nucleotide polymorphism (SNP) data is used to test for same species contamination using a support vector machine model. Based on the assumption that alternative allele frequencies in next generation sequencing follow the beta-binomial distribution, the deviation parameter  $\rho$  is estimated by maximum likelihood method. All features of a radial basis function (RBF) kernel support vector machine (SVM) are generated using either publicly available or private training data. Lastly, the generated SVM is applied in the test data to detect contamination. If training data is not available, a default RBF kernel SVM model is used.

**Results:** We demonstrate the potential of our approach using simulation experiments, creating datasets with varying levels of contamination. The datasets combine, *in silico*, exome sequencing data of DNA from two lymphoblastoid cell lines (NA12878 and NA10855). We generated VCF files using variants identified in these data, and then evaluated the power and false positive rate of our approach to detect same species contamination. Our simulation experiments show that our method can detect levels of contamination as low as 5% with reasonable false positive rates. Results in real data have sensitivity above 99.99% and specificity at 90.24%, even in the presence of DNA degradation that has similar features to contaminated samples. Additionally, the approach can identify the difference between mixture of tumor-normal cells and contamination. We provide an R software implementation of our approach using the defcon()function in the vanquish: Variant Quality Investigation Helper R package on CRAN.

## 1. Introduction

High throughout next generation sequencing (NGS), has shown its advantages over traditional Sanger sequencing and microarrays in terms of accuracy, cost and speed (Metzker, 2010; van Dijk, Auger, Jaszczyzyn, & Thermes, 2014). As NGS technologies have matured, best practices for quality control and data processing procedures have also developed (Patel & Jain, 2012). One key step in a robust pipeline is the detection of sample contamination detection. Sample contamination detection is a necessary data quality control step of NGS data analysis pipeline since contamination might be introduced during sample preparation and sequencing analysis. Sample contamination affects all downstream sample analysis and may even generate misleading results, which in turn lead to false positive associations and genotype misclassification (Jun et al., 2012).

In this paper, our working definition of contamination is when any sample contains tissues from more than one contributing source. There are two major types of contamination: (1) cross-species contamination; (2) same-species or within-species contamination. Additionally, it is very common for a single cancer sample to contain both normal and tumor cells from the same patient or cell line.

Contamination can emerge in next generation sequencing samples for various reasons. Despite best practices, unclean lab devices are one of the most common causes, with the introduction of unexpected materials like *Mycoplasma* (Schmidt, Hummel, & Herrmann, 1995). This same problem exists in even more recent, large scale research projects, such as the 1000 Genomes Project (Langdon, 2014). Contamination can also arise from sample handling, sample extraction, library preparation and amplification, sample multiplexing and inaccurate barcode sequencing (Simion et al., 2018). Available methods are mainly based on the sequencing and allele

frequency information of the testing sample and can be categorized into two groups, according to the source of contaminants: cross-species contamination and same-species contamination.

Cross-species contamination has been well studied, with several methods that have emerged to detect this type of contamination (Korneliussen, Albrechtsen, & Nielsen, 2014; Laurence, Hatzis, & Brash, 2014; Schmieder & Edwards, 2011; Strong *et al.*, 2014). In fact, modern metagenomics approaches are extensions of cross species contamination detection approaches. For example, Schmieder and Edwards developed DeconSeq, a framework for identification and removal of human contamination from microbial metagenomes (Schmieder & Edwards, 2011), providing a framework for cross-species contamination detection during sequencing alignment. Later in 2014, Merchant *et al.* used microbiome analysis software to scan a sample of domestic cow, *Bos taurus*, and found small contigs from microbial contaminants (Merchant, Wood, & Salzberg, 2014). Generally, data is assembled from available Sanger reads to the known species, then the unmapped contigs within the assembly are classified by *k*-mer matching where a database containing all bacteria, archaea, and viruses from the RefSeq database. As a result, contigs aligning to other genomes are found as a sign of contamination.

By contrast, same-species or within-species contamination is relatively more challenging, with fewer methods to implement valid and robust approaches. Arguably the most commonly implemented approach, and the earliest developed is ContEst (Cibulskis *et al.*, 2011), a module within the Genome Analysis ToolKit (GATK) software (McKenna *et al.*, 2010). ContEst uses a Bayesian method to calculate the posterior probability of each contamination level and find the maximum *a posteriori* probability (MAP) estimate of the contamination level at homozygous loci. Assuming a uniform prior distribution,  $\text{Unif}(0,1)$ , on the contamination level, the posterior distribution of contamination level is proportional to the joint distribution of seeing observed

alleles given the qualities of base calling and the probabilities of observing true alleles in contaminated sample. Thus, ContEst requires VCF and BAM format input and general population frequency information, i.e., base identities and quality scores from sequencing data.

Later in 2012, Jun et al. developed the VerifyBamID package for detecting same-species contamination of human DNA samples in both sequence and array-based data (Jun et al., 2012). VerifyBamID implements both likelihood-based and regression-based approaches that assume no more than one contaminant contained in a testing DNA sample. The likelihood of a contamination level is maximized by using a grid search over each contamination level for a maximizer. While VerifyBamID has demonstrated good sensitivity in real normal data experiments, copy number alterations (CNAs) in tumor samples will shift allele frequencies away from those outside of CNA regions, which leads to misinterpreting copy number-driven shift as contamination (Bergmann, Chen, Arora, Vacic, & Zody, 2016a).

The statistical model introduced in VerifyBamID, was further developed by Bergmann *et al.* in their Conpair method for detecting another source of same-species contamination - the mixture of tumor and normal cells from the same patient sample (Bergmann, Chen, Arora, Vacic, & Zody, 2016b). Conpair focuses on homozygous loci within tumor-normal paired samples. Given that homozygous markers are invariant to copy number changes; pre-selected highly informative genomic homozygous markers are provided with Conpair to perform contamination detection.

Other more recent methods developments have used haplotype structure for contamination detection in NGS data (Sehn et al., 2015). Closely spaced SNP pairs within sequencing region are identified from the 1000 Genomes database (The 1000 Genomes Project Consortium, 2012). Then read haplotypes for these selected SNP pairs are inferred. Human-human admixture is suggested

if more than two read haplotypes are observed at a given locus in a sample. The estimated level of contamination for each testing sample are twice the mean frequency of the minor haplotype.

While the current approaches have been successful in a broad range of applications, there are major limitations that we address with our current approach. These developments represent substantial improvements in both the practical implementation of the quality control procedures and the statistical model used. Current approaches rely on the sizeable human reference genome data, as well as at least two large, memory intensive files. They each require either tumor and normal BAM files (Conpair), or VCF file and BAM file (VerifyBamID and ContEst). By using a combination of a Beta-Binomial assumption and support vector machines within our algorithm to detect same species contamination, it works directly from information in VCF file. Even for a tumor-normal paired sample, no other information is needed to perform contamination detection. The change points of B allele frequencies (from the VCF file) are detected and then all chromosomes are separated into shorter sequences. Each sequence overlapping any copy number variation or abberation region is detected and filtered. In the enclosed study, we demonstrate our method in both real and simulated data, and show that it demonstrates excellent sensitivity and specificity for both simulated and real data. We also describe an R package implementation of the method.

## 2. Materials and Methods

### 2.1. Beta-binomial Model of Allele Frequency in Next Generation Sequencing

The proposed method is designed for human applications and assumes a diploid genome. For each locus that contains a single nucleotide variant (SNV) called from next generation sequencing (NGS) data, we define the allele frequency as the number of counts for the alternative

(non-reference genome) allele over the total number of depth. For any diploid genome, if an individual is homozygous for the alternative allele (denoted as alternative/alternative, 1/1), the expected allele frequency is 1; meanwhile, if an individual is heterozygous (denoted as reference/alternative, 0/1) at a locus, 0.5 is the expected allele frequency. Using these theoretical expectations motivates using the Binomial distribution for the number of reads at each locus,

$$\Pr(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x},$$

where  $n$  is the total number of depth at the locus;  $p$  is the theoretical allele frequency;  $x$  is the number of counts for the alternative allele.

While such a simple model is intuitively appealing, previous studies have discovered extra binomial dispersion, specifically, overdispersion of allele frequency distributions (Esteve-Codina et al., 2011; Pickrell et al., 2010; Skelly, Johansson, Madeoy, Wakefield, & Akey, 2011; Zhang et al., 2014). This overdispersion results in a higher variability than the Binomial distribution, so a distribution that models such large variance is needed. Previous studies proposed and demonstrated the Beta-Binomial distribution as an appropriate model for allele frequencies at a particular locus in a subpopulation. The Beta-binomial distribution is a discrete hierarchical model containing the Beta distribution and binomial distribution, where the probability follows the Beta distribution; and the response follows the Binomial distribution. Hence the probability mass function of the Beta-Binomial distribution is

$$\Pr(x; n, a, b) = \binom{n}{x} \frac{B(x+a, n-x+b)}{B(a, b)},$$

where  $n$  is the total number of reads at the locus;  $B(a, b)$  is the beta function theoretical allele frequency;  $x$  is the number of counts for the alternative allele. This model has been applied in a number of studies since, and the advantages of the Beta-Binomial over the Binomial distribution

in dealing with overdispersion has been repeatedly demonstrated (Chen et al., 2016; Mayba et al., 2014). This work motivates our use of the Beta-Binomial distribution.

## 2.2. Quality Control of Variant Call Format (VCF) files

The input format for our method is the well-established Variant Call Format (VCF< Danecek et al., 2011). To our knowledge, this is the first method to detect same-species contamination from the VCF format. The VCF format contains all single nucleotide variation (SNV) information that is needed in the next steps. Before a VCF file is used in our software, quality control is needed to filter noise and unnecessary information. The recommended quality control and processing steps are outlined below, and are additional quality control steps beyond the processing to produce the VCF file itself.

### Step 1. Indel Filtering

The SNVs used in our method must be substitution variants, not copy number variations like insertions and deletions. Only substitution mutations result in heterozygous and homozygous genotypes that are appropriately modeled by the Beta-Binomial distribution. Insertions and deletions are identified as any mutation segments with a length of more than one base pair, and are then filtered/dropped in this step.

### Step 2. Homozygous and Heterozygous Genotype Calling

The next step of processing is to call the genotypes for modeling. As mentioned above, there are two genotypes for alternative allele at any SNV: homozygous and heterozygous. Suggested genotypes are listed in the GT field of VCF file where 0/0 means homozygous reference; 0/1 means heterozygous; and 1/1 means homozygous alternative. In step 2 of

processing, new information is generated that summarizes the genotype in reference to the alternative allele, in which “Het” reflects 0/1 in GT, and “Hom” reflects 1/1 in GT. This results in two categories of called variants, each corresponding to its own Beta-binomial model. Homozygous reference (0/0) and other heterozygous genotypes (1/2, 2/3, and so on) are not considered to simplify computations, and are labeled as “Complex” and are not included in further calculations.

### Step 3. Low and High Depth Filtering

NGS data is not completely error free, so it is important to identify whether a sequence is a true call or a sequencing error. One of the efficient methods is to set thresholds for coverage depth (Morgan et al., 2010). A previous study suggested read depths of >50 provide acceptable sensitivity and specificity in mutation detection. A reasonable read depth threshold should be chosen according to the average read depths of a testing sample.

### Step 4. Change point Detection for CNV

In a pure sample, when there is a copy number variation (CNV) region, its features look very similar to the region with more than one contributors (same species contamination). Hence it is necessary to filter CNV region before generating features. If there is CNV information already generated for testing sample, the function `vanquish::defcon()` can filter the CNV region directly. Otherwise, a change point detection method is used to detect the CNV region. It has been reported that the variances of B-Allele Frequency (BAF, alternative allele frequency) at heterozygous loci are different among normal, duplication, deletion and loss of heterozygous (Ku et al., 2013). Therefore, change point analysis can be employed to detect change point of variance, i.e., border

of copy number region. The package “changepoint” is applied on only heterozygous positions for multiple changepoint searching of variance change (Rebecca Killick & Eckley, 2014).

### 2.3. Distribution and Likelihood based Features

The next step of our approach is to generate variables/features that will be used for model building to predict whether a sample contains same-species contamination. Briefly, there are two types of features that are generated and used in model building: distribution-based feature and likelihood-based feature.

#### (1) Distribution based features

By its definition, allele frequency is a real number between 0 and 1. To generate distribution-based features, allele frequency is binned into 4 regions as shown in **Figure 1**: Low Alternative Allele Frequency (LowRate), Heterozygous Alternative Allele Frequency (HetRate), High Alternative Allele Frequency (HighRate), and Homozygous Alternative Allele Frequency (HomRate). Here 0, 0.3, 0.7, and 0.99 are their corresponding cut-off values.

From the bins in **Table 1**, there are 8 distribution-based features generated for the following model-building steps. These features reflect the distribution of allele frequencies in the whole file, instead of at each variant calling position. Therefore, each input sample/ VCF file has one set of features to represent itself.

#### (2) Average log-likelihood, the likelihood based feature

The only likelihood based feature is the average likelihood of all loci in VCF file. The likelihood is calculated by applying the Beta-binomial distribution. In the current study, NA 10855 and other available pure samples are selected as a reference genome to calculate maximum

likelihood estimator for both parameters,  $p$  and  $\rho$ , in the Beta-binomial distribution (Sequenced at Q2 Solutions). With  $\hat{p}$  and  $\hat{\rho}$ , log likelihood of all loci are calculated so that the last feature, their average value is also generated.

#### **2.4. Support Vector Machine**

After generating features, a classification method is used to identify whether a sample is from single contributor or multiple contributors. In the current study, we apply a Support Vector Machine (SVM) model because of the complexity in pattern recognition within feature space (Cortes & Vapnik, 1995). The SVM method fits a hyperplane between single and multiple contributor regions to optimally discriminate between classifications. In our approach, we implement the SVM method using the e1071 R package (Meyer et al., 2018) (<https://CRAN.R-project.org/package=e1071>). Since a linear model is not guaranteed, the Gaussian (Radial Basis Function) kernel is used to avoid such parametric assumptions. There are two parameters in SVM analysis that need to be tuned: cost and gamma. They are tuned using the parallel searching method, where a grid search is conducted on an exponentially growing sequence of cost and gamma to look for optimized paired values. It is possible for the estimated parameter to be different, given another training data set.

### **3. Results**

#### **3.1. Changepoint analysis for approximate copy number region detection**

In the scenario where the copy number information of a sample is not provided, change point analysis was conducted to find copy number regions of a sample. The rmChangePoint() function within the vanquish package imports cpt.var() from changepoint package (Rebecca Killick & Eckley, 2014). The pruned exact linear time (PELT) method (R. Killick, Fearnhead, &

Eckley, 2012) and the Changepoints for a Range of PenaltieS (CROPS) algorithm (Haynes, Eckley, & Fearnhead, 2014) are employed to search for variance changepoints. In **Figure 2(A)**, the B-allele frequencies (only those between 0.05 and 0.95) of corresponding loci contained within the input VCF files are plotted. It is clear that there exist some CNV patterns. Red vertical lines are where the variance changes detected. The whole plot is separated into multiple parts by these change points. For each part, if the percentage of loci of the B-allele frequency is between 0.45 and 0.55 is larger than 10% and the skewness is larger than 0.5, the part will be kept in further analysis. The result after filtering is shown in **Figure 2(B)**. See the documentation of the vanquish package for more details.

### **3.2. Beta-binomial parameter estimation for reference sample(s)**

In order to calculate likelihood-based features for further analysis, maximum likelihood estimators of  $\rho$  for beta-binomial distribution of heterozygous and homozygous models are estimated. For the B-allele frequency, the theoretical value of the parameter  $p$  is 0.5 in heterozygous model and 1 in homozygous model.  $p$  is fixed at 0.5 and 0.999 to search for  $\hat{\rho}$  in corresponding model. L-BFGS-B (Byrd, Lu, Nocedal, & Zhu, 1995) is applied for maximum value searching. For instance, NA10855 was chosen as a reference sample, and five replicates were sequenced by Q2 Solutions. The maximum likelihood estimator of  $\rho$  in each sample was estimated by rho\_est() in the vanquish package (**Table 2**) and the sample averages were achieved for further analysis. The value the of estimator highly depends on the variant caller, so it is suggested to keep using the same variant caller for the reference sample, training sample, and test samples.

### **3.3. Features in the classification and regression model**

To train the classification and regression model, 238 samples were sequenced at Q2 Solutions as a training data set. 124 out of the 238 samples were pure; and the rest of the training set were contaminated. Some of the contaminated samples were mixed on purpose in wet lab; and others were simulated by two pure FASTQ format files (Cock, Fields, Goto, Heuer, & Rice, 2010). Pure and contaminated samples show different B-allele frequency patterns. As shown in **Figure 3**, only heterozygous loci detected in samples are plotted. The pure sample (**Figure 3A**) had a narrow horizontal band; while the contaminated sample (**Figure 3B**) had a relatively uniform distribution for B-allele frequency. Eight boxplots and t-tests (null hypothesis of no differences) were conducted to show the difference of pure and contaminated samples considering each feature (**Figure 4**). Among all eight features after t-tests, HomVar, HetVar and HighRate have significant p-values  $1.786^{-9}$ ,  $1.750^{-6}$ , and  $4.540^{-20}$  correspondingly. See Supplementary data for details of each feature.

### 3.4. Tune cost and gamma parameters in radial kernel SVM

Monte Carlo method (1000 times) and tune() from R package e1071 (Meyer et al., 2018) (<https://CRAN.R-project.org/package=e1071>) were used to tune parameters cost and gamma in support vector machine. All 238 samples were split into training (70%, 167 samples) and test (30%, 71 samples) sets. For the training set, we used grid search to tune parameter cost in the range of  $(2^{-4}, 2^{12})$ , and gamma in the range of  $(2^{-4}, 2^4)$ . Then sensitivity and specificity were calculated from the test set by using tuned cost and gamma. The results of Monte Carlo simulation are in **Table 3**. Median values of cost (16) and gamma (0.25), mean values of sensitivity (97.65%) and specificity (96.27%) were presented and the tuned cost and gamma were used in radial kernel SVM model for prediction.

### 3.5. Simulated data test results

Two reference samples, NA12878 and NA10855, were sequenced at Q2 Solutions. Two pairs of FASTQ format files were obtained from sequencing results, and were re-sampled and mixed according to different proportions as shown in **Table 4** by seqtk (<https://github.com/lh3/seqtk>). In this simulated test, NA12878 was treated as the sample, while NA10855 was assumed to be mixed into the NA12878 sample. The percentages of the artificial contaminated mixtures range from 0.5% to 20%, and the detection rate at different levels of contamination were calculated. The total number of reads were ~50 million for all 6 mixture samples. After running contamination analysis, contamination percentages above 5% were readily detected, while lower percentages were not detectable (**Table 4**). Based on this result, the contamination detection analysis has a detection sensitivity down to 5% in this case. If contaminant is less similar to the sample that it is mixed with, the detection sensitivity will be lower; on the other hand, if the similarity is larger, then contamination detection analysis will be more challenging.

### 3.6. Two groups of real data test results

After quantitative simulation test, the trained model was applied on a real data set (22 samples in all); and the results are shown in **Table 5**. The samples are ranked by regression values from `e1071::svm()`. Twenty out of twenty-two samples had the same prediction as prior identification. However, a pair of Human\_T-Lymphoblast samples (see bold text in **Table 5**) were considered as contaminated but predicted as pure. Therefore, we zoomed in to check the distribution of B-allele frequency in this pair of samples (**Figure 5**). The copy number variation pattern (two bands and three bands in chromosomes) was shifted below (middle area of CNV pattern shifted from 0.5 to 0.3) since the sample was a mixture of tumor normal cells from the

same person. The pattern was clear: the source of sample was uniform. Based on the distance of shift, the percentage of tumor cells and normal cells of the sample is even predictable.

A second data set (contains 53 samples in all) was also used to test the model (See **supplementary information** for more details). Within this data set, 12 samples were mixed with contaminant on purpose; 41 samples were pure. The test results show that the sensitivity is > 99.99%, and the specificity is 90.24%. All four false positive samples were formalin-fixed paraffin-embedded (FFPE) tissues, and were very likely to be degraded (**Figure 6**). When generating features from this degraded sample in **Figure 6**, its features looked very similar to features from a contaminated sample.

## 4. Discussion

In this study, we introduced a novel strategy that detected same-species or within-species contamination by using B-allele frequency from variant call information only. An R package, vanquish: Variant Quality Investigation Helper, was produced to conduct this analysis. A brief flowchart (**Figure 7**) introducing contamination detection procedure shows the necessary steps as following:

- (1) Variant call format (VCF) generated by a variant caller was read into R by vanquish::read\_vcf function. So far, the supported variant callers are GATK, VarDict and strelka2.
- (2) Copy number variation regions existing in VCF file were approximately detected and filtered by vanquish::update\_vcf function.
- (3) Features for radial kernel SVM model were extracted from each sample by vanquish::generate\_feature function.

- (4) Parameter cost and gamma for kernel SVM were tuned.
- (5) A test sample was predicted whether it was contaminated.

Additionally, two scenarios were found where the performance of this approach may be affected. First, if the sample is a mixture of tumor and normal cells from the sample individual, which is not considered contamination by our definition. The second scenario is when the test sample has very low quality so that no clear B-allele frequency pattern can be found.

Finally, we have produced a user friendly R package to enable rapid analysis for same species analysis. Our tool uniquely performs this important step of QC from VCF files, improving the performance, memory requirements, etc. The running time of CNA region removal and feature generating is summarized in **Figure 8** (Hardware: Dell R820, 512GB of RAM). Five samples without any known change point are run 10 times each to achieve their average running time, keeping their maximum numbers of runs of algorithm are uniform. Same as our expectation, larger samples need more time to do change point detection and feature generation. Besides, if a sample contains more change points, it also needs a longer running time.

## Acknowledgements and funding

This work was supported by Q2 Solutions | EA Genomics, a Quintiles Quest Joint Venture and P01 CA142538 from the National Cancer Institute. The authors also gratefully thank Dr. Chad Brown for discussions and inspiration on this contamination detection project.

## References

Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V., & Zody, M. C. (2016a). Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics*, 32, 3196–3198.

Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V., & Zody, M. C. (2016b). Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics*, 32, 3196–3198.

Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16, 1190–1208.

Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., ... Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nature Communications*, 7, 11101.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., & Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics*, 27, 2601–2602.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38, 1767–1771.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158.

Esteve-Codina, A., Kofler, R., Palmieri, N., Bussotti, G., Notredame, C., & Pérez-Enciso, M.

(2011). Exploring the gonad transcriptome of two extreme male pigs with RNA-seq.

*BMC Genomics*, 12. doi:10.1186/1471-2164-12-552

Haynes, K., Eckley, I. A., & Fearnhead, P. (2014). Efficient penalty search for multiple

changepoint problems. *ArXiv:1412.3617 [Stat]*. Retrieved from

<http://arxiv.org/abs/1412.3617>

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., ... Kang,

H. M. (2012). Detecting and Estimating Contamination of Human DNA Samples in

Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*,

91, 839–848.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal Detection of Changepoints With a

Linear Computational Cost. *Journal of the American Statistical Association*, 107, 1590–

1598.

Killick, Rebecca, & Eckley, I. A. (2014). **changepoint** : An *R* Package for Changepoint

Analysis. *Journal of Statistical Software*, 58. doi:10.18637/jss.v058.i03

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next

Generation Sequencing Data. *BMC Bioinformatics*, 15. doi:10.1186/s12859-014-0356-4

Ku, C. S., Polychronakos, C., Tan, E. K., Naidoo, N., Pawitan, Y., Roukos, D. H., ... Cooper, D.

N. (2013). A new paradigm emerges from the study of de novo mutations in the context

of neurodevelopmental disease. *Molecular Psychiatry*, 18, 141–153.

Langdon, W. B. (2014). Mycoplasma contamination in the 1000 Genomes Project. *BioData*

*Mining*, 7. doi:10.1186/1756-0381-7-3

Laurence, M., Hatzis, C., & Brash, D. E. (2014). Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLoS ONE*, 9, e97876.

Mayba, O., Gilbert, H. N., Liu, J., Haverty, P. M., Jhunjhunwala, S., Jiang, Z., ... Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biology*, 15. doi:10.1186/s13059-014-0405-3

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303.

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, 2, e675.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11, 31–46.

Morgan, J. E., Carr, I. M., Sheridan, E., Chu, C. E., Hayward, B., Camm, N., ... Taylor, G. R. (2010). Genetic diagnosis of familial breast cancer using clonal sequencing. *Human Mutation*, 31, 484–491.

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, 7, e30619.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464, 768–772.

Schmidt, T., Hummel, S., & Herrmann, B. (1995). Evidence of contamination in PCR laboratory disposables, 9.

Schmieder, R., & Edwards, R. (2011). Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE*, 6, e17288.

Sehn, J. K., Spencer, D. H., Pfeifer, J. D., Bredemeyer, A. J., Cottrell, C. E., Abel, H. J., & Duncavage, E. J. (2015). Occult Specimen Contamination in Routine Clinical Next-Generation Sequencing Testing. *American Journal of Clinical Pathology*, 144, 667–674.

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., ... Telford, M. J. (2018). A software tool ‘CroCo’ detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology*, 16. doi:10.1186/s12915-018-0486-7

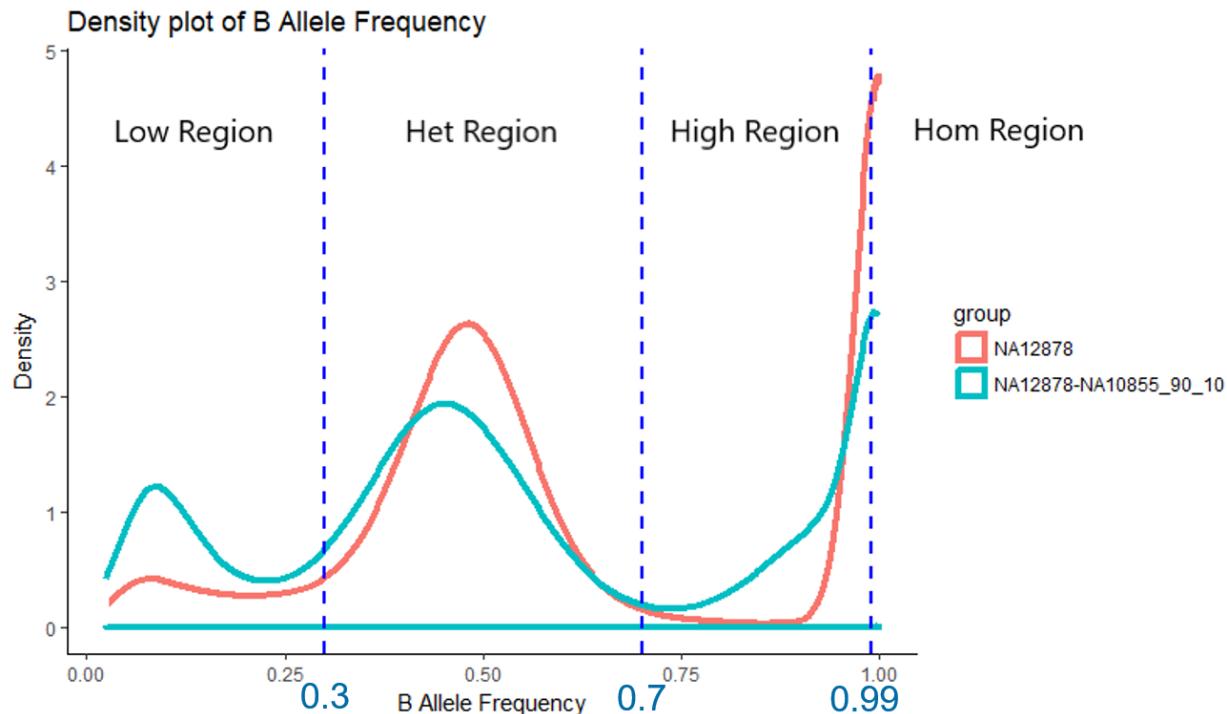
Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., & Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 21, 1728–1737.

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., ... Flemington, E. K. (2014). Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens*, 10, e1004437.

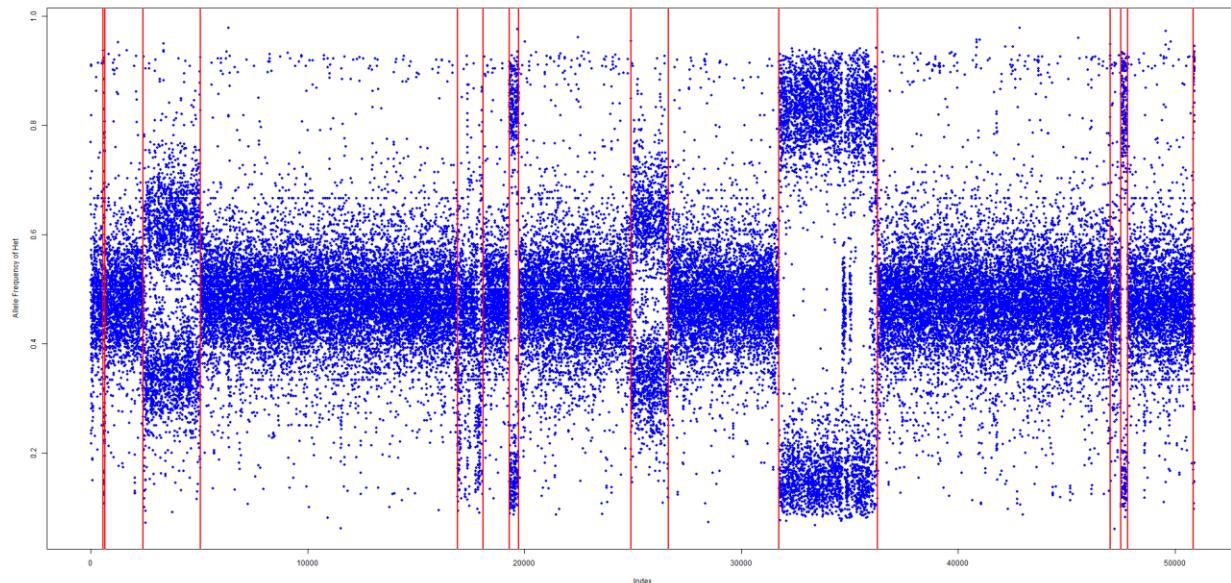
The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.

van Dijk, E. L., Auger, H., Jaszczyzyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30, 418–426.

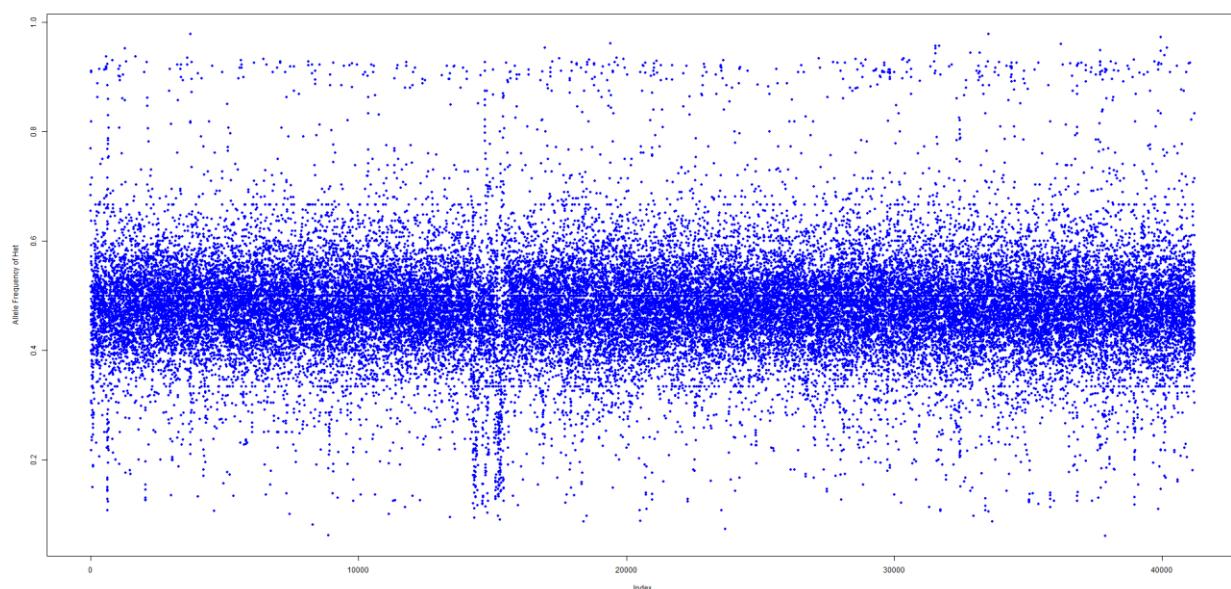
Zhang, S., Wang, F., Wang, H., Zhang, F., Xu, B., Li, X., & Wang, Y. (2014). Genome-wide identification of allele-specific effects on gene expression for single and multiple individuals. *Gene*, 533, 366–373.



**Figure 1. Four regions for B-Allele frequency: LowRate, HetRate, HighRate, and HomRate and their corresponding cut-off values.** LowRate is the region where B-Allele frequency is below 0.3; HetRate is the region where B-Allele frequency is between 0.3 and 0.7; HighRate is the region where B-Allele frequency is between 0.7 and 0.99; HomRate is the region where B-Allele frequency is above 0.99. **B Allele frequency density plots of pure (NA12878) and contaminated (NA12878-NA10855\_90\_10) samples.** The features used in SVM classifier are based on the difference between pure and contaminated curves.

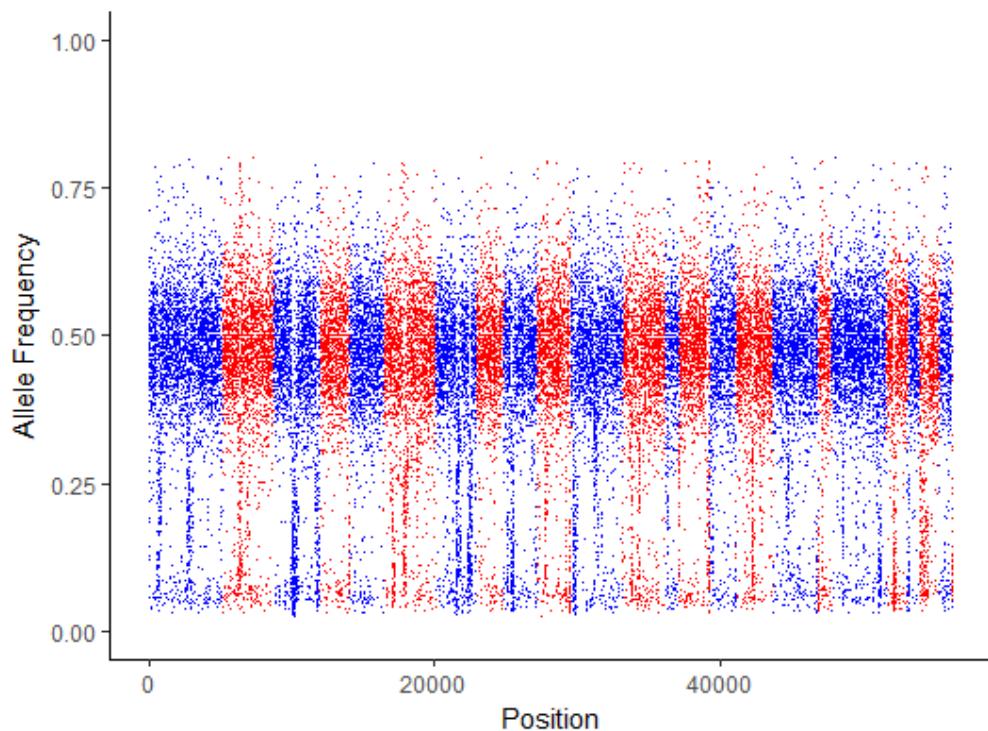


**(A) Before copy number region filtering**

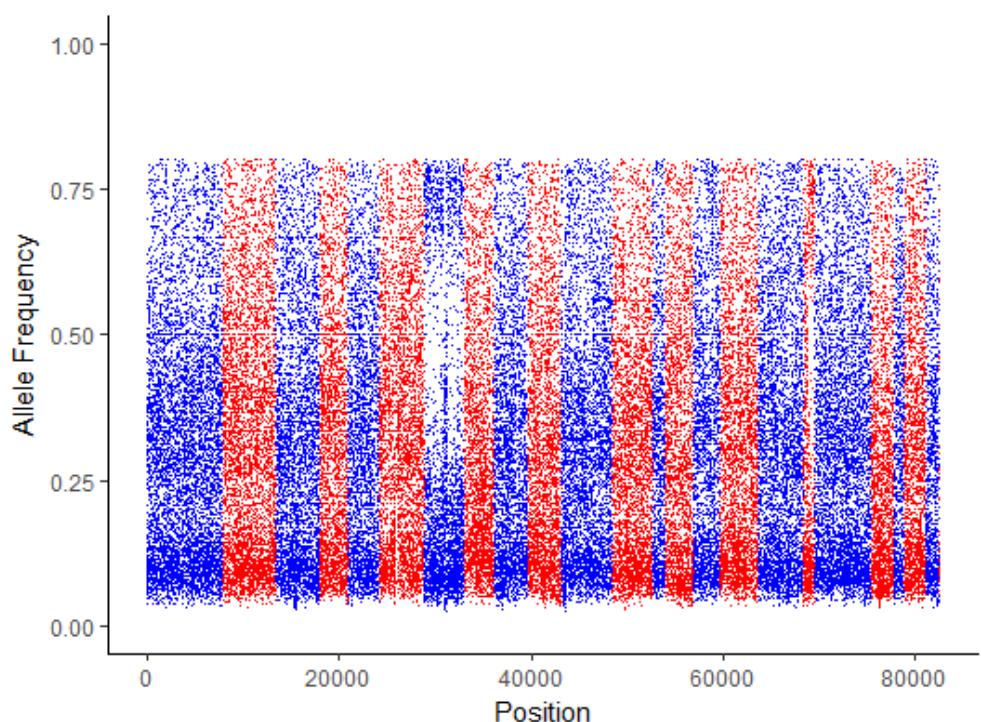


**(B) After copy number region filtering**

**Figure 2. Change point analysis for copy number region detection.** The Y-axis shows B-Allele frequency; X-axis is shows the location number of each variant from chromosome 1 to 22. (A) is before copy number region filtering. Red lines indicate where the variance change based on detection results. It is clear that there exist some copy number patterns. (B) is after copy number region filtering and in which most copy number patterns are removed.



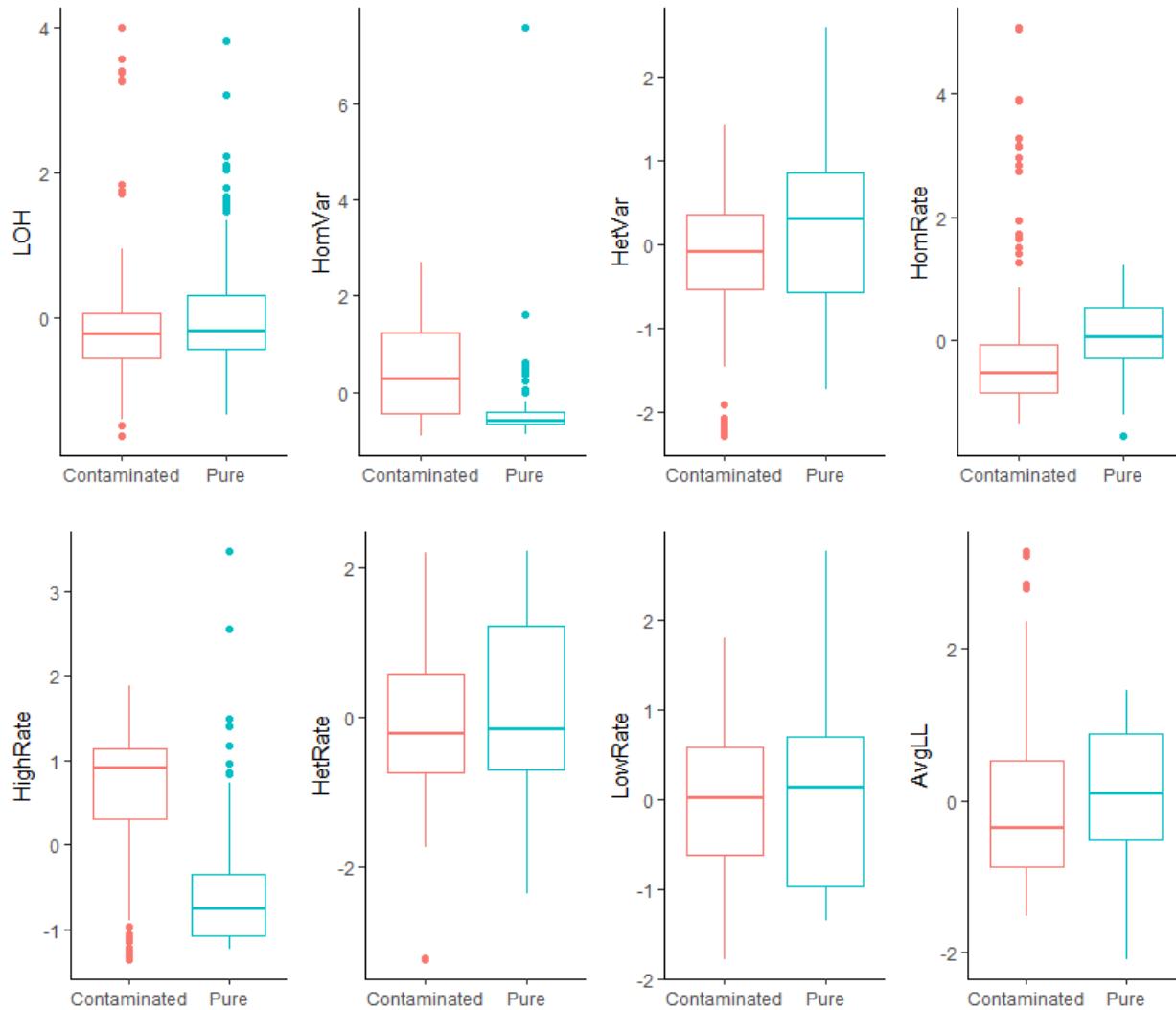
(A) Pure sample: NA24143\_gDNA\_Run2\_L1-2-3-4.SSHT42<sup>a</sup>



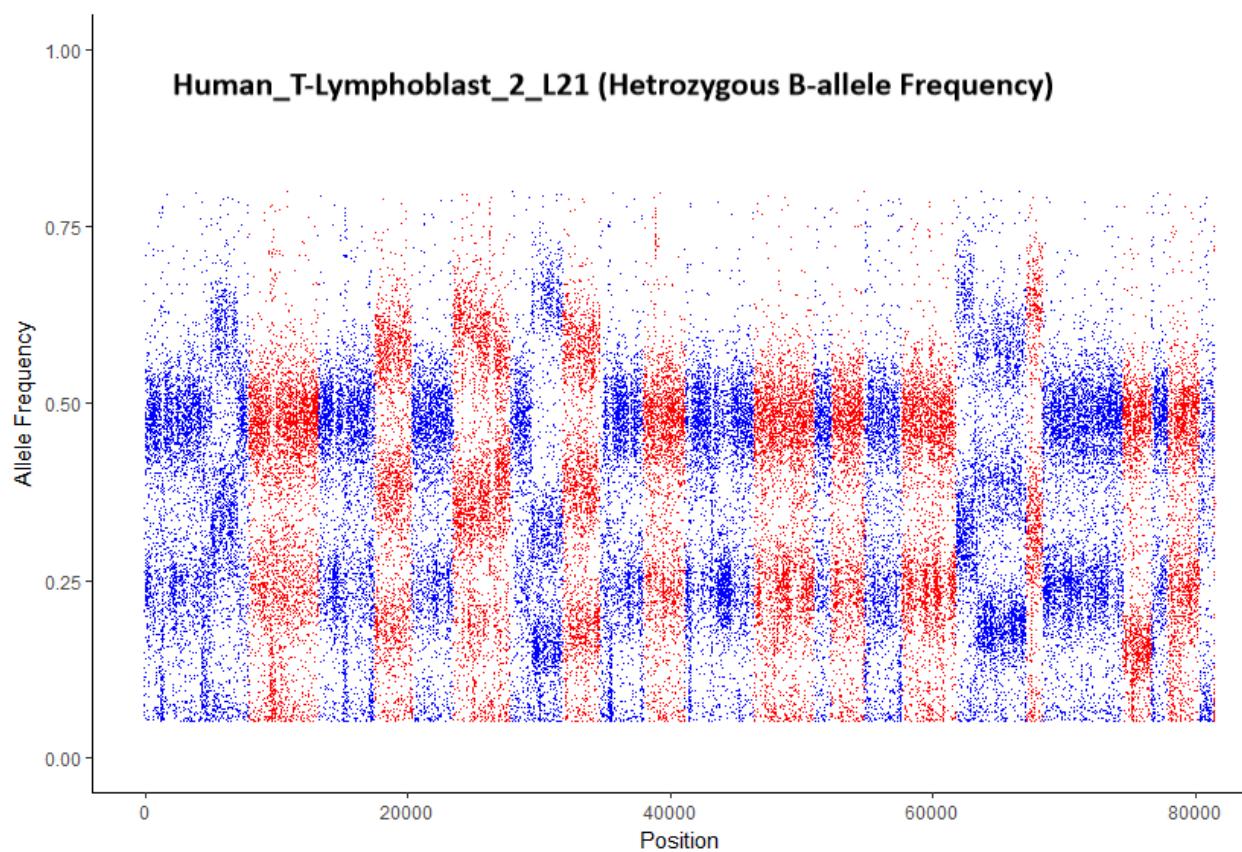
(B) Contaminated sample: Multiplex\_Reference\_Standard\_rep1\_Run1\_L1-2-3-4.SSXT96-96<sup>a</sup>

<sup>a</sup> Sequenced at Q2 Solutions | EA Genomics, a Quintiles Quest Joint Venture

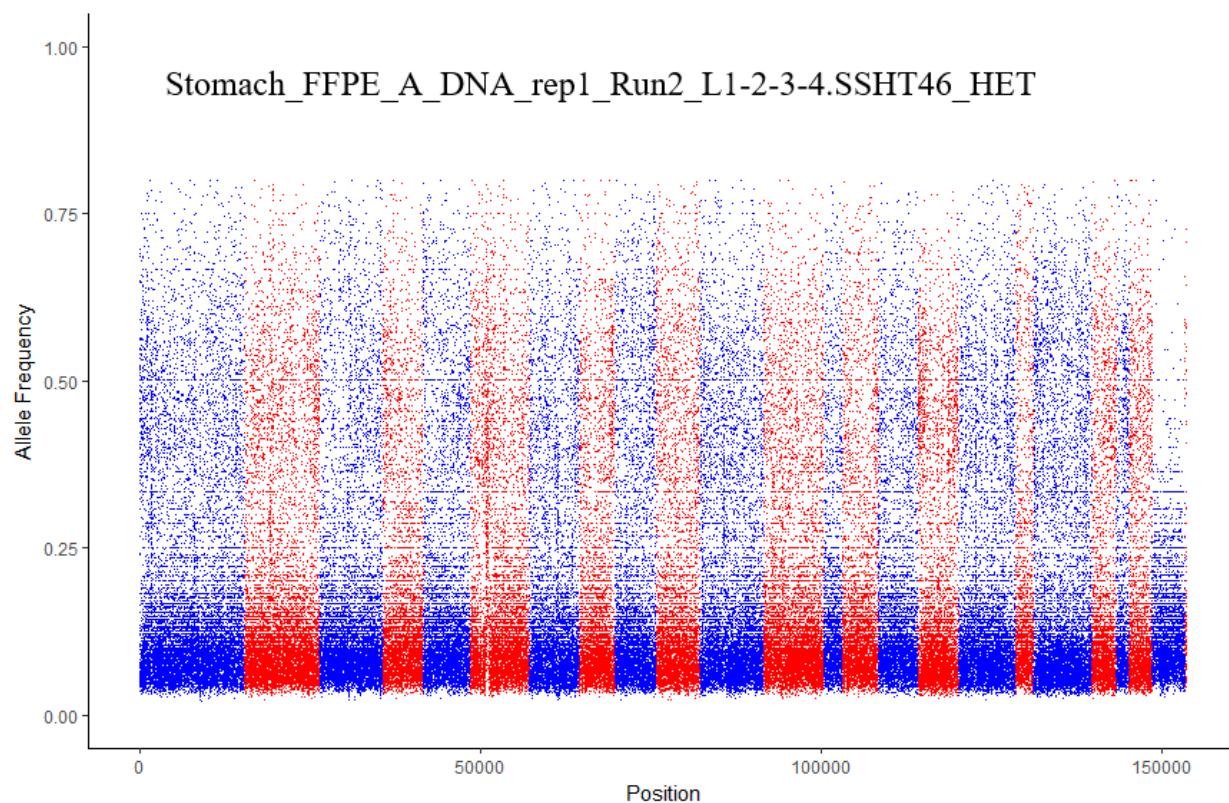
**Figure 3. Heterozygous B-allele frequency patterns of pure (A) and contaminated (B) samples.** Blue and red bands indicate different chromosomes.



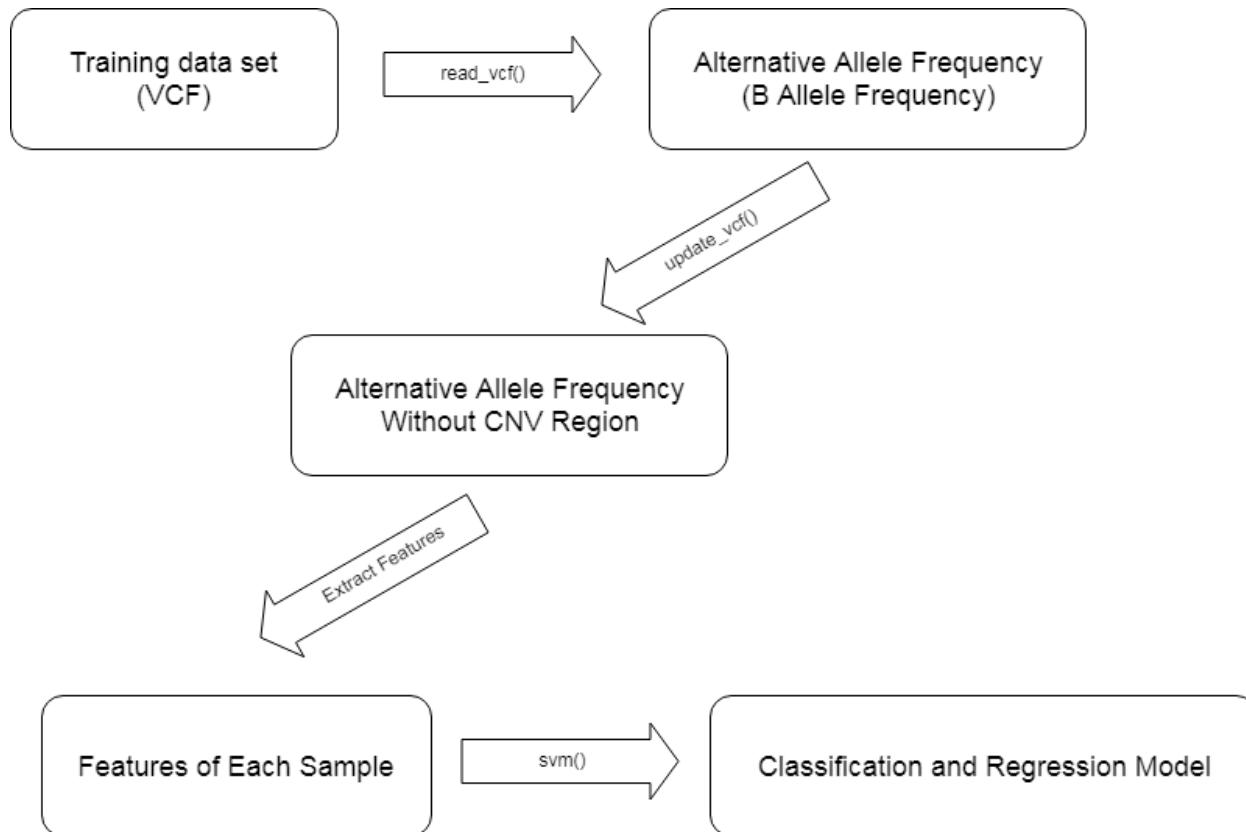
**Figure 4. Boxplots of all features for pure samples (124) vs contaminated samples (114).** Among all eight t-tests (null hypothesis of no difference hypothesis), HomVar, HetVar and HighRate have significant p-values  $1.786^{-9}$ ,  $1.750^{-6}$ , and  $4.540^{-20}$  correspondingly.



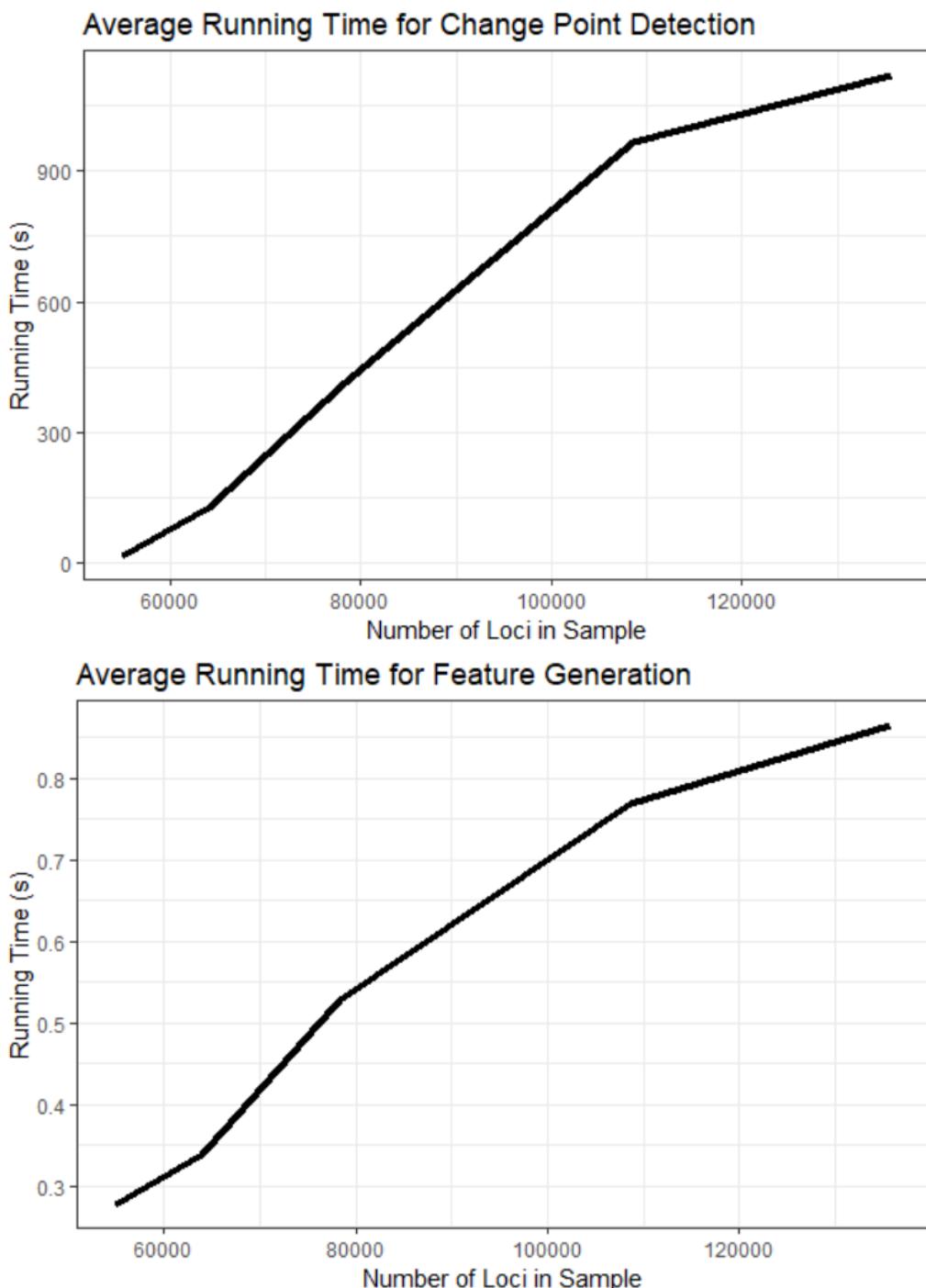
**Figure 5.** A sample that was previously considered as contaminated, but was detected as pure by `vanquish::defcon()`. The copy number variation pattern was shifted lower (middle area of CNV pattern shifted from 0.5 to 0.3) since the sample was a mixture of tumor normal cells from the same person.



**Figure 6. Formalin-fixed paraffin-embedded (FFPE) tissue with low quality.** This sample was a false positive prediction result because of the similarity between its low quality and contaminants.



**Figure 7. A brief flowchart of contamination detection procedure.**



**Figure 8. Average running time (of 10 runs) for change point detection (Top) and feature generation (Bottom).** Five samples with different numbers of loci but the same maximum number of runs of algorithm (= 2).

### ***Features for Classification and Regression Model***

Name	Description
LOH	$\frac{Het}{Hom}$ , the ratio of heterozygous and homozygous markers within a sample.
HomRate	The percentage of the loci in HomRate <sup>a</sup> region.
HighRate	The percentage of the loci in HighRate <sup>b</sup> region.
HetRate	The percentage of the loci in HetRate <sup>c</sup> region.
LowRate	The percentage of the loci in LowRate <sup>d</sup> region.
HomVar	The variance of allele frequencies in HomRate region.
HetVar	The variance of allele frequencies in HetRate region.
AvgLL	Average likelihood of all loci assuming Beta-Binomial distribution.

<sup>a</sup> HomRate: the region where B-Allele frequency is above 0.99

<sup>b</sup> HighRate: the region where B-Allele frequency is between 0.7 and 0.99

<sup>c</sup> HomRate: the region where B-Allele frequency is between 0.3 and 0.7

<sup>d</sup> HomRate: the region where B-Allele frequency is below 0.3

**Table 1. All features for classification and regression model and their descriptions.**

### **Maximum Likelihood Estimator $\hat{\rho}$ of NA10855 Samples**

	Heterozygous $\hat{\rho}$	Homozygous $\hat{\rho}$
NA10855_1	0.154	0.0269
NA10855_2	0.223	0.0253
NA10855_3	0.177	0.0210
NA10855_4	0.187	0.0310
NA10855_5	0.169	0.0274
<b>Sample Mean</b>	<b>0.182</b>	<b>0.0263</b>

**Table 2. Maximum Likelihood Estimator  $\hat{\rho}$  of NA10855 Samples.**  $\hat{\rho}$  of heterozygous and homozygous models were estimated for each sequencing replicate. The sample mean can be used for generating features from training data set.

### ***Parameter Tuning and Performance Test***

Median Cost	Median Gamma	Average Sensitivity	Average Specificity
16	0.25	97.65%	96.27%

**Table 3. Monte Carlo test results for parameter tuning and performance test.**

### ***Contamination Detection for Simulated Data***

Sample Component	Reads (NA12878)	Reads (NA10855)	Test Results
NA12878 (80%) + NA10855 (20%)	40M <sup>a</sup>	10M	Contaminated
NA12878 (90%) + NA10855 (10%)	45M	5M	Contaminated
NA12878 (95%) + NA10855 (5%)	47.5M	2.5M	Contaminated
NA12878 (97.5%) + NA10855 (2.5%)	48.75M	1.25M	Pure
NA12878 (99%) + NA10855 (1%)	49.5M	0.5M	Pure
NA12878 (99.5%) + NA10855 (0.5%)	49.75M	0.25M	Pure

<sup>a</sup>M: million

**Table 4. Contamination Detection for a series of simulated data.**

### Contamination Detection for Real Data

Sample Name	Classification	Regression	Prior Identification
Human_B-Lymphocyte_L8	1	1.9243094	1
Human_B-Lymphocyte_2_L20	1	1.9209875	1
Human_Breast_2_L16	0	1.483925	0
Human_Breast_L4	0	1.463376	0
<b>Human_T-Lymphoblast_2_L21<sup>a</sup></b>	<b>0</b>	<b>1.3622305</b>	<b>1</b>
<b>Human_T-Lymphoblast_L9</b>	<b>0</b>	<b>1.3472358</b>	<b>1</b>
Human_Brain_L3	0	1.3147938	0
Human_Brain_2_L15	0	1.303287	0
Human_Testis_L12	0	1.245767	0
Human_Cervix_2_L17	0	1.2429423	0
Human_Testis_2_L24	0	1.2424441	0
Human_Cervix_L5	0	1.203943	0
Human_Macrophage_L10	0	1.158416	0
Human_Macrophage_2_L22	0	1.1582528	0
Human_Liver_2_L18	0	1.1442246	0
Human_Liposarcoma_L7	0	1.1406007	0
Human_Liposarcoma_2_L19	0	1.132044	0
Human_Skin_2_L23	0	1.1209464	0
Human_Skin_L11	0	1.1194772	0
Human_Liver_L6	0	1.1170909	0
Human_Reference_DNA_Male_L1	0	1.0945151	0
Human_Reference_DNA_Male_2_L13	0	1.0906951	0

<sup>a</sup>This is a mixture of tumor and normal cells. See **Figure 5** for distribution of B-allele frequency of this sample.

**Table 5. Contamination Detection for a series of real data.** Twenty out of twenty-two samples had the same prediction as prior identification. However, a pair of Human\_T-Lymphoblast samples (bold text) were considered as contaminated but predicted as pure.