**Quality Control of NGS Data to Identify Sample Contamination, Identity and Ethnicity**

**Abstract**

Quality control (QC) of next-generation sequencing (NGS) data requires accurate identification of samples with cross-individual contamination. Here, we demonstrate a method to QC NGS samples by calculating the pair-wise Identity-by-descent (IBD) probability from exome genotyped SNP data, and accurately estimating the IBD probability for the contaminated, related and distantly related samples. The implications of this method is in quality control of NGS genotype data, including those sequenced by targeted sequencing panels by estimating the level of cross-individual contamination, and inferring the relationships and ethnicity of the samples.

1. **Introduction**

Quality control (QC) of next generation sequencing (NGS) data is one of the vital first steps warranted in avoiding any misclassification of sequence genotypes leading to false positive association of variants in downstream analyses. QC of NGS data includes detection of sample misclassification due to sample swaps, plate swaps or cross-individual contaminations during sample preparation and sequencing process. DNA contamination is one of the most common sources of error in genotype data analysis. This is attributable to the fact that the samples are processed in a batch within the same sequencing plate involving multiple steps of sample handling, reagent mixing, adapter barcoding, and subsequently leading to potential cases of contamination. Samples in adjacent wells of the same plate and samples in the same batch during library preparation are therefore more likely to be contaminated. DNA sample contamination could be of three major types: occurring at the cross-individual, cross-species and individual levels [1]. It is easier to detect cross species contamination and can be filtered out through series of steps during sequence reads alignment [2], whereas within species contamination is difficult to detect and can greatly reduce the genotype data quality [3]. Since these genotype data are used in several downstream analyses such as identification of disease causing variants in genes, driver mutations in cancer related genes including methods (such as burden/SKAT tests [4]) that could be used to detect rare variants of large effect in diseases like cancer. The objective of this study is to develop a quality control method for NGS data that could detect samples with cross-individual contamination, identify relationship, and infer ethnicity of all individuals in the cohort. Several methods have been developed in the past that claim to detect the source of sample contamination using SNP based genotypes [1, 5], but they are still not very efficient method of detection as discussed below.

In this method, we have used the concept of IBD (identity by descent) which is calculated based on the proportion of allele types descended from the common ancestors that are shared by the individuals in a base population [6]. Alleles from two different individuals are considered to be IBD if they have descended from the same parental allele in a population, whereas alleles that look the same but have not descended from the common ancestor are simply known to be identical by state (IBS) [6, 7]. The conventional method of calculating the IBD using known pedigree has become obsolete due to the fact that SNP data are now commonly used to estimate the IBD when there is no information on the reference pedigree or basal population [6]. Implementation of this method to identify relatedness between the individuals is shown to predict varying degree of relatedness between every pair of individuals in the cohort as shown in the result below. The IBD thus estimated can differentiate duplicates or monozygotic twins (IBD=1), first-degree relatives, (IBD=0.5), first cousins (IBD=0.25) and second cousins (IBD=0.125), thereby inferring actual relationship between the samples [8]. Whereas, cryptic IBD values (any values between expected IBD [6]) are obtained for the pair of individuals with cross contamination problems (Figure 1). In contrast to the concept of population structure where coalescent ancestry is characteristically common to all individuals within the population, cryptic relatedness is limited to one or a fewer pairs of individuals [9]. It is argued that the cryptic relatedness, like population stratification (confounding biases due to inherent differences in allelic frequencies among subpopulations from different ethnicities), can have serious effects leading to spurious associations in case-control studies [10]. In a cohort of samples of ethnic admixture, the proportion of alleles shared among ethnic minorities could be different to those shared among individuals that are ethnically predominant in the cohort. IBD values are inherently susceptible to the confounding biases due to population stratification [11]. Therefore, in any cohort of ethnic admixture, the IBD values for individuals from ethnic minorities are likely to be inflated/deflated making them purportedly more related to one another that can be confused with cryptic relatedness arising from sample contamination.

There are also several other methods previously published that estimate sample contamination (eg. Contest [1]) or sample swaps (eg. HYSYS [14]) or both sample swaps and sample contamination (eg. Conpair [15]). ContEst method is based on the Bayesian probability that calculates the *a posterior probability* to estimate the level of sample contamination [1] and the HYSYS method is based on the concordance of homozygous germ line variants (known to be least affected by disease or by loss of heterozygosity) between the samples to determine the metrics of relatedness [14], whereas the Conpair method utilizes a pair of tumour matched normal to perform concordance verification between the tumour matched normal samples, estimating further for cross-individual contamination in whole-genome and whole-exome sequencing data. However, the utility of these methods is limited to QC of NGS samples where only paired patient-matched tumour and normal samples are available, and therefore these tools are not useful when either unmatched tumour or unmatched germline samples need to be analysed independently [1, 14].

In addition to these methods, methods implementing algorithms estimating IBD values can also be used to determine the cross-sample contamination. There are several tools [1, 5, 12, 13] available for the calculation of IBD between the individuals, however very few tools have addressed the problem of sample contamination. IBD values have been traditionally derived from two different approaches: maximum likelihood (MLEs) [16, 17] and the method of moments (MMEs) [5, 18]. The MLEs are found to be more biased when analysing small size loci, and yet are computationally intensive [18], whereas MMEs are comparatively less efficient in deriving accurate biological inferences [17]. Tools like Plink[6] [5], estimates the IBD values based on the method-of-moments approach calculated for two individuals taken from the same homogenous, randomly mating population [5]. Thus reported IBD probabilities between many individuals that are distantly unrelated are also reported as zero IBD.

In this method, we implement the raw unified additive relationship (UAR) discussed previously [6] using SNPs (maf >= 1% and maf <= 10%) from the genotype panel to estimate the relatedness coefficients that are highly correlated with pedigree relationships compared to Plink estimates. This method is expected to detect sample contamination, relationship and ethnicity of samples with high level of precision, hence can be an effective tool to quality control NGS data, including those sequenced by targeted sequencing panel. Although, the raw unified additive relationship (UAR) model [6] has been used previously in vcftools (**The variant call format and VCFtools)**, the kinship function of this tool has not taken various things into consideration such as ethnicity of individuals and filtering of good quality genotype that are going to affect the overall quality control process. Particularly, when having large numbers of samples, from multiple ethnicities, pedigrees, the relatedness of 3rd degree or distantly related individuals starts to show.

1. **Methods**
   1. **Cohort description**

In this study, SNPs from 258 multi-ethnic samples were genotyped using GATK [20] (see methods) creating multi-sample VCF files. These samples also include individuals with known relationships (first-degree, second-degree and mislabelled samples), unrelated individuals, along with the individuals with plate contamination. This cohort also includes seven samples from simulated serial dilution experiment (discussed below) used as positive controls for sample contamination.

* 1. **Sequence Genotyping**

Sequenced reads were aligned to the build of human genome (UCSC hg19 version) using Novoalign sequence alignment tool [V2.07.09, ([http://www.novocraft.com](http://www.novocraft.com/))] and merged as BAM (a compressed binary version of a SAM file which is used to store aligned sequences) files. These sequence alignment files were converted/handled using tools like SAMtools [19] and Picard tools (version 2.5.0, available at: <https://broadinstitute.github.io/picard>) for further analysis. SNPs were called using Genome Analysis Toolkit (GATK version 3.5) which uses statistical evidence to call a variant having enough depth of coverage for each variant site [20]. The VCF (Variant Call Format) files with SNP calls generated from GATK were used for downstream analysis. GATK incorporated quality filter were used to filter good quality SNPSs with the following values, for combined recalibrated quality score (QUAL) > 50, Quality by sequencing depth at the SNP position or QUAL/DP ratio (QD) > 0.5, Fisher strand (FS) < 60, maximal length of the homopolymer run (HRUN) < 5, Filter = PASS, strand bias at the SNP position (SB) < 0.1. Only the good quality SNPs derived from whole exome sequencing were retained that were further analysed and filtered using custom scripts employing R and Bioconductor packages.

* 1. **Simulation of sample contamination**

Two of the samples from European origin (A-02-0010 and A-02-0063) were selected for dilution experiment using serial dilution method of reads for each input BAM file. The total numbers of reads in unmapped regions in each of the BAM file were insignificant, and therefore all reads (both mapped and unmapped) were considered for serial dilution. Picard tool’s ‘Downsample’ option was used for down sampling of AOGC-02-0010 to 1%, 5%, 10%, 20%, 30%, 40% and 50% BAM reads. Reads from each of these BAM files were added to A-02-0063 BAM file to simulate sample contamination thereby creating contaminated BAM files with 1%, 5%, 10%, 20%, 30%, 40% and 50% contamination levels by proportionally down-sampling the reads in one sample also taking into account the total read counts of another sample. For example, 5% contamination of sample “A-02-0010” (in sample “A-02-0063”) can be achieved by calculating the proportion of reads to down-sample from sample A-02-0010 [i.e., (0.05 \* total reads in A-02-0063)/ total reads in A-02-0010)]. Since for both samples, the total reads were not significantly different from mapped reads, the down-sampled reads from A-02-0010 were simply added to sample A-02-0063 using ‘MergeSamFiles’ option from Picard tools to create contaminated samples. ‘AddOrReplaceReadGroups’ option from Picard tools was used to change the SM tags of newly created BAM files. Replacing a new SM tag in newly created BAM files ensures that all the read groups were assigned to a single new read group.

* 1. **Population Stratification**

The sample pairs from ethnic minorities are found to have inflation in IBD also for unrelated pairs. This is often expected considering the effects of population stratification due to ethnic differences in samples within the cohort. Various different methods were attempted to correct for inflated IBD as a result of confounding biases due to ethnicity (population stratification). Principal component (PCA) analysis was performed using the common SNPs from samples and HumanHap650Y Genotyping tagSNPs derived from international HapMap Project [21] to infer the ethnicity of each sample in the cohort (**Figure 1**). Ethnicity of each sample was predicted by spiking the common SNPs from HumanHap650Y Genotyping tagSNPs with the SNPs from the samples in this cohort and plotting the, which was then used to perform PCA analysis. The shortest multidimensional Euclidean distances were calculated for all sample points in our cohort with the ethnically defined HapMap samples using first four principal components. The sample points lying in the shortest multidimensional Euclidean distance from the HapMap samples were then assigned with their respective ethnicity.

Additionally, several methods were also tried to implement to solve the problem of inflation in IBD among individual pairs of ethnic minorities due to population stratification. Multidimensional Euclidean distance was calculated between the centroid of the PCA and the mean eigenvector for each pair of samples as a measure of variance due to ethnicity. Then generalized linear regression models (GLM) was applied to a response variable of Euclidean distance against paired IBD in an attempt to correct for IBD inflation. However, none of these methods could solve the problem of population stratification inflating the IBD values. We have also tried to correct the IBD inflation using another method known as snpload implemented in a parallel PCA and data processing tool known as Shellfish.py. Snpload is a method by which a list of SNPs that are used in generating clusters for a PCA is garnered. We had removed top 500 of these SNPs that contribute most of the variance in a PCA assuming that the removal of these SNPs from the analysis would solve the problem. However, this method also failed to solve the problem of inflation in IBD. In our attempt to use various different approaches to subdue confounding biases affecting the IBD values, we also tried a method inferring the ethnicity of each sample in the cohort as described above and analysed all samples that belonged to one specific ethnicity along with the HapMap samples from the same ethnicity. Since our analyses have shown that the minimum of 10 samples are required to calculate more accurate IBD values between the sample pairs, inclusion of HapMap samples helped to address this problem. This ethnic specific analysis with enough number of samples was expected to correct the inflation, but it also failed to achieve the expected results. When none of these methods solved the problem of population stratification inflating the IBD values, we finally, performed PCA analysis (**Figure 1**) to infer the ‘closest’ ethnicity of each sample using multidimensional Euclidean distance in the PCA based on the ethnicity of samples from HapMap genotype data. IBD thresholds for different sample types were determined: sample pairs with inflated IBD due to ethnicity (confounding), contaminated samples, mislabelled samples, relatives (first-degree, second-degree and identical pairs), cancer-germline extensions and unrelated sample pairs. Inflation in IBD was distinct among sample pairs from ethnic minorities. In this way, IBD thresholds were defined for each sample group including sample pairs from ethnic minorities for which IBDs were found inflated due to confounding biases.

* 1. **Genotype filtering**

Any indels were excluded from the analysis and only the good qualities of SNPs in VCF file were filter passed for IBD calculation based on GATK filtering as discussed above. Similarly, allele depth thresholds were applied for different genotypes as following: the minimum depth of coverage of 20 folds was set as a requirement to apply genotype filtering with the additional filtering of minimum of 10 folds for heterozygous calls and minimum of 5 folds for homozygous calls. In addition to GATK recalibration, several hard filtering was also applied to SNPs to include only good quality SNPs suitable to calculate accurate IBD values. Hard filtering was also applied for each variant site based on the minor allele frequency (maf) to exclude SNPs that are too rare and too common in the cohort. Variant site filtering was applied for minor allele frequency (denoted by ‘p’) in the cohort to be p>= 1% and p <= 10%. IBD methods are likely to be affected by the change in allele frequencies contributed by copy number variations (CNVs) in cancer samples hence contributing to SNPs that are too rare or too common in the cohort and this method we tend to avoid this flaw by avoiding any SNP outside these filtering range used here. Genotype quality (GQ) of 20 (representing phred score of 99% accuracy for SNP calls) was also applied to hard filtering ensuring only good quality genotypes were used to calculate the IBD.

* 1. **Calculation of IBD and development of algorithm**

Pairwise genetic relationships between the individuals were calculated from the genotype markers as discussed previously (equation i and ii) [6]. Using the raw UAR, where the relationship between a pair of individuals denoted as *i* and *j* using the weighted average across the SNPs is given by;

*Aij = 1/m ∑ k [1]/2pk (1 – pk)* ---- equation (i)

And for the relationship of an individual with itself by;

*Aii = 1 + 1/m ∑ k {(x2ik – (1 + 2pk) xik + 2p2k)} / 2pk (1 – pk)* --equation (ii)

Where, the parameter X is the indicator for the SNP *k* in *i* and *j* individuals, *pk* denotes the allele frequency of SNP *k*, where m is the number of SNPs in the given individual.

Custom R-script was written to develop the algorithm to calculate the IBD estimates of paired-samples in the cohort. In order to determine the minimum number of SNPs required to calculate the IBD from the NGS genotyping data, a cohort of 266 samples from different ethnic origins, related, unrelated and contaminated samples were analysed. The filtered SNPs were iteratively passed in chunks to calculate the pairwise IBD and the total SNPs used in each chunk were recorded. The IBD values for paired individuals were plotted against the number of SNPs used in the analysis to determine the minimum number of SNPs required to calculate the coherent IBD between the pair of individuals. Sample pairs within relatedness spectrum of IBD > 0.06 were analysed for the detection of sample contamination.

1. **Results and Discussion**
   1. **Samples with inflated IBD**

IBD values are inherently susceptible to the problem of inflation or deflation due to systematic difference in allele frequency among the individual of different ethnicities [11]. The pairwise IBD was calculated for all possible combinations of sample pairs. The problem of inflation in IBD was addressed using principal component analysis (PCA) (**Figure 1**) and by inferring the ethnicity of each sample (see methods).

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\Rplot01.tiff

Figure 1. Principal component analysis (PCA) was performed to separate the HapMap samples from different ethnic origins that are used to infer the ethnicity of the samples with unknown ethnicity in the cohort. These samples are dispersed at various spatial distances in a PCA. -9 indicates the cases of unknown ethnicities from our cohort and all other samples are of known ethnicities derived from HapMap study.

The unrelated sample pairs with confounding IBDs (purportedly related) were identified to be ethnic minorities (also PCA outliers) (**Figure 2**). On the other hand, the sample pairs from European origin did not show any inflation and these unrelated sample pairs were found to have IBD closed to zero as expected for unrelated individuals (**Figure 2**). Similarly, related sample pairs were also identified based on the calculated IBDs for these samples that are associated with both the first-degree and second-degree relatives.

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\threshold.tiff

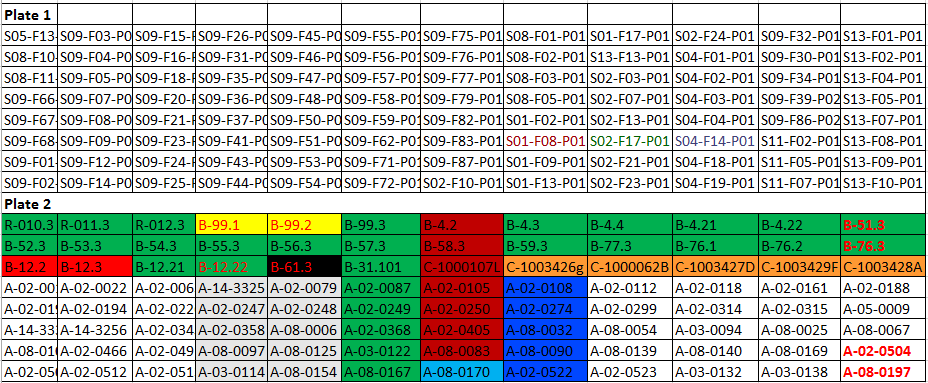
**Figure 2. IBD threshold determined for different sample pairs in the cohort. IBD values range from 0 to 1 for unrelated to identical paired individuals.**

* 1. **Contaminated samples**

The IBD calculated for the contaminated samples showed cryptic relatedness between the paired individuals. The contaminated samples have cryptic IBD and the contaminating samples (contaminants) have IBD close to 1 for self-pairs (i.e. normal IBD). We were also able to detect samples contaminated by multiple samples (**Table 1**). Here, contaminated samples were found to have cryptic IBD for self-pairs, and cryptic to seemingly related IBDs for cross-sample pairs. All sample pairs that were not contaminated and did not belong to ethnic minorities were found to have IBD close to zero (see rules to detect sample contamination discussed below) and their self-pairs to have IBD close to 1. Thus, we were able to separate both the contaminating and the contaminated samples from the pairs (**Table 1**). Sample pairs with cryptic IBD values were those samples sequenced in the same sequencing plate and were also located in the wells adjacent to one another (**Figure 3**). This strongly suggests that the samples from the same plate could have been contaminated during sample handling and manipulation process in the lab. Cryptic IBD was also detected in sample pairs that were simulated for contamination. The cryptic relatedness detected for these simulated sample pairs (self-pairs) are found correlated (r = -0.96) with the level of contamination as they proportionally exhibit higher degree of cryptic relatedness with the increasing level of contamination (**Figure 4**).

**Table 1. Identification of contaminated samples using IBD values for the paired individuals where IBD (cross-samples) is the IBD for contaminated sample paired with the respective contaminants and IBD (self-pairs) is the IBD for contaminated sample paired with itself). Samples highlighted in same colour were sequenced in the same sequencing plate, whereas sample pairs in the same rows were at the adjacent cells of the same sequencing plates.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Contaminated** | **Contaminant** | **IBD**  **(cross-samples)** | **IBD**  **(self-pairs)** |
| S09-F83-P01 | S09-F87-P01 | 0.266 | 0.805 |
| S01-F08-P01 | S01-F09-P01 | 0.327 | 0.786 |
| S04-F14-P01 | S04-F18-P01 | 0.332 | 0.773 |
| S11-F02-P01 | S11-F05-P01 | 0.297 | 0.77 |
| S13-F08-P01 | S13-F09-P01 | 0.269 | 0.77 |
| A-08-0083 | A-03-0122 | 0.373 | 0.763 |
| A-08-0090 | 0.567 |
| A-02-0105 | A-02-0087 | 0.406 | 0.718 |
| A-02-0108 | 0.554 |
| A-02-0250 | A-02-0274 | 0.459 | 0.716 |
| S02-F17-P01 | S02-F21-P01 | 0.361 | 0.708 |
| A-02-0405 | A-02-0368 | 0.405 | 0.682 |
| A-08-0032 | 0.642 |
| A-08-0170 | A-08-0167 | 0.286 | 0.648 |
| A-02-0522 | 0.753 |
| A-08-0083 | A-03-0122 | 0.373 | 0.763 |
| A-08-0090 | 0.567 |
| A-02-0105 | A-02-0108 | 0.554 | 0.718 |
| A-02-0087 | 0.406 |
| A-02-0250 | A-02-0274 | 0.459 | 0.716 |
| A-02-0249 | 0.557 |
| A-02-0405 | A-02-0368 | 0.405 | 0.682 |
| A-08-0032 | 0.642 |
| S02-F17-P01 | S02-F21-P01 | 0.361 | 0.708 |
| S04-F14-P01 | S04-F18-P01 | 0.332 | 0.773 |
| S01-F08-P01 | S01-F09-P01 | 0.327 | 0.786 |
| S11-F02-P01 | S11-F05-P01 | 0.297 | 0.77 |
| S13-F08-P01 | S13-F09-P01 | 0.269 | 0.77 |
| S09-F83-P01 | S09-F87-P01 | 0.266 | 0.805 |



**Figure 3. Samples that were found to be contaminated from these two different sequencing plates belong to the sample plate and were located in the adjacent wells. Thus identified contaminated samples (having cryptic IBD) located in adjacent wells also indicates that they were accurately identified since they are more likely be contaminated by one another.**

To detect whether the IBD value is correlated with the level of contamination, we have analysed seven samples with different levels of contamination (**Figure 4**). Using seven samples with 1%, 5%, 10%, 20%, 30%, 40% and 50% contamination levels, we have determined that the level of sample contamination is negatively correlated (r = -0.96) with the IBD estimates for the samples paired with itself (**Figure 4**).

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\serial_dilution.tiff

**Figure 4. Comparison of estimated coefficient of IBD for seven samples paired with itself and the actual percentage of contamination in these samples**.

* 1. **Unrelated samples**

Majority of the sample pairs were found to be unrelated (**Figure 2**). These sample pairs include samples with inflated IBD (in the case of confounding unrelated samples; see **Table 2**) and samples with IBD close to zero. Since this method uses the current generation as the base population, the IBD values for some of the sample pairs are calculated to be negative (**Figure 2**) inferring the correlation of homologous alleles from different gametes rather than the probability of IBD among those paired individuals [6]. This is because, a correlation has the properties of being bounded at –1 and +1, whereas as probability is (typically) between 0-1. If we use the current population as the base (as used to be done), then pairs of individuals that are less related than average will have a negative relationship coefficient. This causes a problem as a lot of methods (which initiated from using a true ancestral population as a base) used probabilities of relationship coefficients. However, the correlation of homologous alleles between gametes (i.e. individuals) creates a standardised relationship coefficient that is a correlation.

* 1. **Related samples**

Relationships of known sample pairs used in this study were accurately identified based on the IBD values. These samples include first-degree relatives and a pair of second-degree relatives (**Figure 2**). Since these sample pairs belong to ethnic minorities, their IBDs are much inflated and do not have IBD values as expected for related individuals (**Table 2**). Any related sample pairs not from ethnic minorities are expected to have IBD that are normal (not inflated). However, these samples do not have cryptic IBDs for self-pairs as we see in contaminated self-pairs (IBD between 1.05 -0.685). To determine whether the sample pairs from ethnic minorities are contaminated, it is also strongly suggested to perform a separate analysis of ethnic minorities only. This should be aided with plate mapping to determine the possibility for barcode contamination of samples within the plate. Additionally, we were also able to identify a sample that was mislabelled in this cohort (**Figure 2**). This sample pair had IBD close to 1.0 indicating that one of the samples was mislabelled (or duplicated).

* 1. **Cancer-germline extension**

Another group of samples analysed in this study was cancer-germline samples. These samples include both germline and cancer samples from the same patients. These sample pairs are therefore identical, but also are relatively “contaminated” by the cancer sub-clones giving them different IBD threshold as compared to the true identical sample pairs (**Figure 2**).

**Table 2. Description of ethnicities and IBD range for different sample types analysed in this study.**

|  |  |  |
| --- | --- | --- |
| Type | Range | Ethnicity |
| Identical | 0.974-1.33 | All ethnicities |
| Mislabelled (identical) | 0.984- | African |
| Cancer extension | 0.928-1.05 | Mostly (or all) Europeans |
| Contaminated | 0.266-0.805 | All Europeans |
| Serial Dilution (contaminated) | 0.0616-0.986 | All Europeans |
| First- degree | 0.456 -0.885 | Ethnic minorities |
| Second- degree | 0.227- | European/Russian |
| Confounding (unrelated) | 0.06-0.395 | All Ethnic minorities |
| Unrelated | < 0.06 | All ethnicities |

* 1. **Minimum SNPs for coherent IBD**

In order to determine the minimum number of SNPs required for IBD, SNP counts were recorded against the observed IBD for every double number of SNPs used for the analysis. This method indicates that as few as 3000 SNPs are required for the calculation of coherent IBD values (**Figure 5**). IBD values for different sample types are distinctly linear after using as few as 3,000 SNPs for the analysis, indicating that this method can also be used with genotyped data sequenced by some of the available target sequencing panels. This method is particularly useful to perform QC on NGS samples sequenced by targeted sequencing panels such as TruSight sequencing panels covering 4800 SNPs, and possibly by TruSight Myeloid and TruSight tumours covering 1765 SNPs (not tested).

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\threshold_snpcount.tiff

Figure 5. The number of SNPs (nearly 3,000 SNPs) required to infer the IBD relatedness between the paired individuals. IBD value of 1.0 indicates the sample pairs compared with monozygotic twins or itself. Sample pairs within IBD close to 0.0 indicate unrelated sample pairs and the individual with cryptic IBD lie adjacent to the dotted line of 0.5 IBD.

* 1. **Comparison with other methods**

We have compared this method with available QC methods such as PLINK [5]. The standard IBD methods within PLINK tool is known to estimate many of the relationships to be zero as it is not capable of defining distantly related pairs in absence of defined base [6]. This is why most of the IBD calculated for distantly related sample pairs are not well resolved by PLINK method and do not have strong correlation (r = 0.76) with the IBD method used herein (**Figure 6**) as compared to (r = 0.94) those sample pairs defined as closed relatives (**Figure 7**).

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\PLink_vs_IBD_all_pairs.tiff

**Figure 6. Comparison of IBD calculated using PLINK method for all the sample pairs with IBD calculated with our method.**

C:\Users\Akanshya\Dropbox\PhD documents\qc paper\Plink_vs_IBD.tiff

**Figure 7. Comparison of IBD calculated using PLINK method for the related sample pairs (including self-pairs) with IBD calculated with our method**.

Another problem with the calculation of IBD using Plink method is that many false positives were identified in the case of cryptically related sample pairs. This clearly demonstrates that the Plink method could not account for correct IBD for the samples that are “unrelated”, in addition to identifying too many false positives for the cryptically related sample pairs. Additionally, Plink also doesn’t have a clear basis of defining IBD threshold for unrelated ethnic outliers, contaminated samples, related and unrelated samples.

* 1. **Rules to Identify contaminated samples**

This method can be used to identify related samples, infer their ethnicities and detect cross-sample contamination. In detecting sample contamination using this method, there are several points need to be considered:

1. All sample pairs with cryptic IBD for both self-pairs and cross-sample pairs need to be shortlisted from the output file.
2. The shortlisted sample pairs with cryptic IBD that belong to ethnic minorities need to be flagged since any inflation/deflation in these sample pairs could be just because of confounding biases due to population stratification (e.g. confounding ethnic minorities in Table 2).
3. Samples with cryptic IBD for both self-pairs along with cross-sample pairs are most likely to be contaminated.
4. Two sample pairs that have cryptic IBD and belong to two different ethnic origins are most likely to be contaminated.
5. This method can also identify both the contaminated and contaminating samples. In a cross-sample pair with cryptic IBD, if any one of the two samples has cryptic IBD for self-pair, that sample can be identified as contaminated. The other sample with normal IBD for self-pair is therefore identified as contaminating sample. If both samples have cryptic IBD for self-pairs, then both of these samples are likely to contaminate each other.
6. Sample pairs with cryptic IBDs that are located in adjacent well of the same sequencing plate are also more likely to be contaminated regardless of their ethnic origin.
7. This method can also detect the level of contamination based on the cryptic IBD values. Samples with higher level of contamination proportionally exhibit higher degree of cryptic relatedness as shown in the result above (**Figure 4**).
8. Cross sample pairs with cryptic IBD, but self-pairs with normal IBD (i.e. IBD~1) for both samples are least likely to be contaminated. Inflation/deflation in IBD for these cross sample pairs could be the result of population stratification given these samples belong to ethnic minorities.

* **Advantages of this method**
  + Can be used for quality control of NGS genotype data, including those sequenced by targeted sequencing panels on target genes capturing fewer SNPs. As few as 3000 SNPs are sufficient enough to QC the NGS data with this method.
  + This method can estimate the level of cross-individual contamination, while also inferring the relationships and ethnicity of the samples.
  + Can efficiently detect contaminated samples having contamination as low as 1 percent.
  + This method does not require normal-matched tumour samples. Individual cancer samples, unmatched normal samples and samples from different ethnic origins can be analysed together for quality control.
  + This method is computationally intensive and still very fast and would take less than 2.5 hours to complete the analysis of more than 200 samples in 40 core (1.0 X40 GB RAM) machine.
  + More accurate measure of IBD than Plink IBD
  + Unlike with ContEst, this method can be used to QC any number of samples independently without having a need for patient-matched tumour and normal.
* **Disadvantages of using this method**
  + Requires more elaborate process to complete the analysis, such as identification of ethnicity using principal component analysis.
  + This method can be difficult to use when discriminating cryptically related sample pairs from ethnic minorities that also have confounding effect on the IBD values.

In conclusion, genetic studies such as in population association studies, various statistical inferences are used that require analysis of independent samples. This method can be used as a multi-platform analysis that can quality control NGS samples thereby avoiding possibility for spurious association in presence of any nominally related and contaminated samples. Although this method is computationally intensive and requires more elaborate process to QC the samples, it can be more accurate in identifying any contaminated samples compared to any other methods as shown in this study by using as few as 3000 SNPs.

**References**

1. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G: **ContEst: estimating cross-contamination of human samples in next-generation sequencing data**. *Bioinformatics (Oxford, England)* 2011, **27**(18):2601-2602.

2. Schmieder R, Edwards R: **Fast identification and removal of sequence contamination from genomic and metagenomic datasets**. *PloS one* 2011, **6**(3):e17288.

3. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data**. *Am J Hum Genet* 2012, **91**(5):839-848.

4. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare-variant association testing for sequencing data with the sequence kernel association test**. *Am J Hum Genet* 2011, **89**(1):82-93.

5. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ *et al*: **PLINK: A tool set for whole-genome association and population-based linkage analyses**. *American Journal of Human Genetics* 2007, **81**(3):559-575.

6. Powell JE, Visscher PM, Goddard ME: **Reconciling the analysis of IBD and IBS in complex trait studies**. *Nat Rev Genet* 2010, **11**(11):800-805.

7. Weir BS, Anderson AD, Hepler AB: **Genetic relatedness analysis: modern data and new challenges**. *Nat Rev Genet* 2006, **7**(10):771-780.

8. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: **Data quality control in genetic case-control association studies**. *Nat Protoc* 2010, **5**(9):1564-1573.

9. Astle W, Balding DJ: **Population Structure and Cryptic Relatedness in Genetic Association Studies**. *Statistical Science* 2009, **24**(4):451-471.

10. Devlin B, Roeder K: **Genomic control for association studies**. *Biometrics* 1999, **55**(4):997-1004.

11. Stevens EL, Heckenberg G, Roberson ED, Baugher JD, Downey TJ, Pevsner J: **Inference of relationships in population data using identity-by-descent and identity-by-state**. *PLoS Genet* 2011, **7**(9):e1002287.

12. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: **Whole population, genome-wide mapping of hidden relatedness**. *Genome Research* 2009, **19**(2):318-326.

13. Browning BL, Browning SR: **A Fast, Powerful Method for Detecting Identity by Descent**. *American Journal of Human Genetics* 2011, **88**(2):173-182.

14. Schroder J, Corbin V, Papenfuss AT: **HYSYS: have you swapped your samples?** *Bioinformatics (Oxford, England)* 2017, **33**(4):596-598.

15. Bergmann EA, Chen BJ, Arora K, Vacic V, Zody MC: **Conpair: concordance and contamination estimator for matched tumor-normal pairs**. *Bioinformatics (Oxford, England)* 2016, **32**(20):3196-3198.

16. Thompson EA: **The estimation of pairwise relationships**. *Ann Hum Genet* 1975, **39**(2):173-188.

17. Milligan BG: **Maximum-likelihood estimation of relatedness**. *Genetics* 2003, **163**(3):1153-1167.

18. Lynch M, Ritland K: **Estimation of pairwise relatedness with molecular markers**. *Genetics* 1999, **152**(4):1753-1766.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

21. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site**. *Genome Res* 2005, **15**(11):1592-1593.