Title:

Sensitive detection of DNA contamination in tumor samples via microhaplotypes

Short Title:

Detection of DNA contamination in tumor samples via microhaplotypes

Authors:

Brett Whitty and John F. Thompson[1*]

Research and Development, PGDx, 2809 Boston Street, Baltimore, MD 21224

[1]Current address: Homology Medicines Inc., 1 Patriots Park, Bedford, MA 01730

[*]Corresponding Author: Thompson.john.f@gmail.com

Abstract

Background: Low levels of sample contamination can have disastrous effects on the accurate identification of somatic variation in tumor samples. Detection of sample contamination in DNA is generally based on observation of low frequency variants that suggest more than a single source of DNA is present. This strategy works with standard DNA samples but is especially problematic in solid tumor FFPE samples because there can be huge variations in allele frequency (AF) due to massive copy number changes arising from large gains and losses across the genome. The tremendously variable allele

frequencies make detection of contamination challenging. A method not based on individual AF is needed for accurate determination of whether a sample is contaminated and to what degree.

Methods: We used microhaplotypes to determine whether sample contamination is present. Microhaplotypes are sets of variants on the same sequencing read that can be unambiguously phased. Instead of measuring AF, the number and frequency of microhaplotypes is determined. Contamination detection becomes based on fundamental genomic properties, linkage disequilibrium (LD) and the diploid nature of human DNA, rather than variant frequencies. We optimized microhaplotype content based on 164 single nucleotide variant sets located in genes already sequenced within a cancer panel. Thus, contamination detection uses existing sequence data and does not require sequencing of any extraneous regions. The content is chosen based on LD data from the 1000 Genomes Project to be ancestry agnostic, providing the same sensitivity for contamination detection with samples from individuals of African, East Asian, and European ancestry.

Results: Detection of contamination at 1% and below is possible using this design. The methods described here can also be extended to other DNA mixtures such as forensic and non-invasive prenatal testing samples where DNA mixes of 1% or less can be similarly detected.

Conclusions: The microhaplotype method allows sensitive detection of DNA contamination in FFPE tumor samples. These methods provide a foundation for examining DNA mixtures in a variety of contexts. With the appropriate panels and high sequencing depth, low levels of secondary DNA can be detected and this can be valuable in a variety of applications.

Background

In many genetics applications, contamination is not a significant issue because small amounts of extraneous DNA do not affect the experimental outcome. For example, variant frequency in germ-line

45 DNA is approximately 0%, 50% or 100% non-reference so it requires a substantial fraction of DNA from
46 other sources to affect variant calls. Contamination of up to 5% has been shown to have virtually no
47 impact on Mendelian variant calling (1). However, even with low-level DNA contamination, there can be
48 a significant impact on some applications like detecting low frequency somatic variants in tumor DNA
49 samples or deconvoluting complex mixtures in forensic samples. Clinical diagnostic samples are
50 presumed to be pure but there is always a concern that unexpected contamination could affect results
51 and escape standard detection methods. Knowledge of whether a sample is contaminated with another
52 human sample and, if so, the extent of that low-level contamination, is critical for assessing whether
53 variant calls can be trusted.

54 Variant detection is often carried out by measuring the variable lengths of short tandem repeats (STRs)
55 or by genotyping/sequencing single-nucleotide variants (SNVs). When unusual frequencies are observed,
56 the degree of contamination is estimated by quantitating the frequency of the extra variants (2-4).
57 However, accurate estimation becomes difficult at low variant frequency because the true signal can be
58 obscured by the competing noise arising from technical issues like STR stutter or genotyping/sequencing
59 errors. This is especially problematic with analog measurement systems that are often used with STR,
60 genotyping, and classical sequencing methods. In contrast, next-generation sequencing (NGS) methods
61 produce a digital signal with both STRs and SNVs, allowing greater sensitivity at low contamination
62 levels.

63 There are a number of other factors that can further complicate the analysis of DNA contamination.
64 With clinical samples, the contaminating DNA can be of different quality than the primary DNA sample.
65 This can add to the noise in the system because the DNAs may respond differently depending on the
66 assay. When tumor samples are being studied, there is the added complication of varying copy numbers
67 as large regions of the tumor genome may be highly amplified or deleted and samples typically contain
68 variable amounts of normal and tumor tissue. When large fractions of the genome have altered copy

69 number, the allelic frequencies (AFs) of SNVs can vary substantially and change from the typical 50-50
70 AF to much more extreme values. These altered allele ratios can overlap the range of AFs for what is
71 typically seen with contamination and thus a sample that is 100% pure can mistakenly appear to be
72 significantly contaminated. Thus, pure samples can be incorrectly reported as contaminated, resulting in
73 a false positive (FP) signal with perfectly suitable samples being discarded. In order to avoid these
74 problems, a method that is independent of genotype AF must be used for determining contamination
75 status with tumor DNA. Analysis based on microhaplotype (MH) counting can be used to achieve this
76 goal.

77 For individuals with a normal set of chromosomes, each autosomal region of any size has two sets of
78 alleles. Those alleles may be the same (homozygous) or different (heterozygous); but, under normal
79 circumstances with a few specific exceptions, there should never be more than two haplotypes in a
80 genomic region in any individual. More than two haplotype versions in a given region indicates
81 contamination. Each haplotype appears with a characteristic frequency in different populations and
82 there may be several MHs at the same genomic positions in any population. When DNA from two
83 individuals is mixed, accidentally or intentionally, there is a possibility for 1-4 MHs in each region.

84 Microhaplotypes can be defined as multiple variants on the same sequence read such that the phasing
85 of the variants is known and the haplotype can be unambiguously determined. MHs have been used for
86 forensic purposes in order to detect contamination and identify individuals (5-8), to detect transplant
87 DNA in a host background (9), and to determine ancestry (10-13). Use of MHs for routine detection and
88 estimation of contamination in clinical samples has not been described but is a powerful application of
89 these methods.

90 *Methods*

91 **Sequencing Samples**

92 De-identified DNA and tumor samples were obtained from several commercial and academic providers.
93 These included BioIVT, Conversant Bio, Cureline, Duke, Folio Bio, Fundacio Institute, Indivumed,
94 iSpecimen, NCI, and Proteogenex. Samples were received either as FFPE material or extracted DNA.
95 Extraction of DNA from FFPE specimens was carried out using standard techniques. DNA was prepared
96 for sequencing using hybridization capture with panels directed primarily at exons in >500 genes
97 implicated in cancer. Depending on the gene, some intronic and promoter DNA was also targeted. DNA
98 was sequenced using an Illumina NextSeq500. The PGDx elio™ tissue complete assay was used for
99 sample preparation and analysis (14, 15).

100 **Variant calling**

101 Low frequency variants were called using a machine learning approach (16) with replicate reads
102 compressed to a single sequence. Even with this approach, technology limitations can induce false
103 positives at frequencies of 1% or less. Because the false positive variant calls could artificially inflate the
104 number of $3^{rd}/4^{th}$ MHs, MH calculations included 3rd MHs only if they surpassed a minimum fraction of
105 variant calls (typically >0.2%). Each read covering the positions of all component SNVs for the MH set
106 was tagged and counted.

107 ***In silico* read mixing**

108 Artificially contaminated samples were made by *in silico* mixing at multiple artificial contamination
109 levels. Contamination levels of x% were generated by combining randomly selected 100-x% reads from
110 the sample with x% reads from the contaminant and running the resultant set of reads through the
111 standard analysis pipeline. It was found that the same results were achieved if only the reads mapping
112 to the MH regions being examined were included in the mixing. These analyses ran on the pipeline much
113 faster. Thus, later experiments used only the truncated data set to speed analysis and allow more
114 conditions to be tested.

115

116 *Results*

117 **Choice of initial variant sets**

118 There are many MHs that have been characterized in the literature, and their usefulness for
119 distinguishing individuals and ancestries is well documented (7. 11, 12). However, most of the well-
120 characterized MHs are not in regions chosen for commercial panel sequencing. In addition, the criteria
121 for selecting MHs for discriminating individuals and ancestry are not the same as for detecting
122 contamination where the goal is not distinguishing individuals but maximizing the number of MHs with
123 >2 versions. Since high depth of sequencing is required for sensitive contamination detection, regions
124 already targeted for analysis were examined so that additional, non-productive sequencing in known
125 MH regions would not be necessary. For gene panel sequencing, this means that literature MH data is of
126 limited value. Each novel MH set we chose thus requires testing for behavior across samples and
127 ancestries. We have used a >500 gene cancer panel directed against FFPE tumor DNA. The panel
128 included regions outside of the target exons and these regions were also inspected for potentially
129 informative SNV sets.

130 The criteria for inclusion in the final set of variants included maximizing the frequency of 3rd MHs while
131 minimizing the likelihood of sequencing errors in those variants. This was achieved by avoiding regions
132 with higher inherent sequencing error rates such as insertions, deletions, and variants within long
133 homopolymeric regions and other repetitive elements. The regions targeted for panel sequencing were
134 examined in the gnomAD database (gnomad.broadinstitute.org/; 17) for multiple SNVs within 180 bp of
135 each other and with 5-95% AF. Variants not in a segmental duplication or a low confidence region were
136 examined for LD in the 1000 Genomes Project samples (ldlink.nci.nih.gov/?tab=ldhap; 18). SNV sets in
137 complete or near complete LD were discarded. Pairs and triplets with at least three haplotypes and the

138 3rd and greater haplotypes having a total frequency of >1% were selected for evaluation in real samples.
139 Variant sets with median coverage >100x for the relevant variants on the read were used to generate a
140 phased haplotype. Variant sets prone to higher frequency 3rd/4th microhaplotypes in purportedly pure
141 samples were eliminated because they would be likely to produce high noise relative to signal. When
142 nearby SNV sets provided duplicate information because of high LD between SNV sets, those with the
143 highest coverage and fewest 3rd/4th microhaplotypes in pure samples were included in the panel and the
144 alternatives omitted to minimize over-counting particular genomic regions. The 106 SNV sets that
145 passed these initial evaluations are shown in Supplemental Table 1. In addition to the genomic
146 coordinates and dbSNP identifiers, the frequency of MHs is shown for the five major ancestry groupings
147 defined by 1000 Genomes as well as for the African American sub-grouping. In addition, the median
148 coverage for the relevant SNVs across >700 samples, the number of times the SNV set generated more
149 than 2 MHs in "pure" samples, and the DNA length containing the SNVs are also listed.

150 **Assessing and estimating contamination levels**

151 In order to determine the criteria with which to detect contamination, *in silico* mixing of reads from
152 individual pure samples was carried out so that the precise level of contamination would be known and
153 the sensitivity of the MH assay could be determined. The initial samples used for establishing
154 contamination methods need to be as close to 100% pure as possible for maximizing sensitivity. After
155 potentially artifactual MH sets were eliminated as described above, samples were examined using the
156 initial MH sets to ensure that they were as pure as possible by including only those samples with the
157 fewest 3rd/4th MH sites.

158 The gnomAD database was used as a starting point for selecting SNVs and, at the time it was initially
159 queried, had a higher proportion of individuals of European ancestry (~60%) than individuals with other
160 geographic ancestries. Because of this, variation in European ancestry individuals had a larger impact on

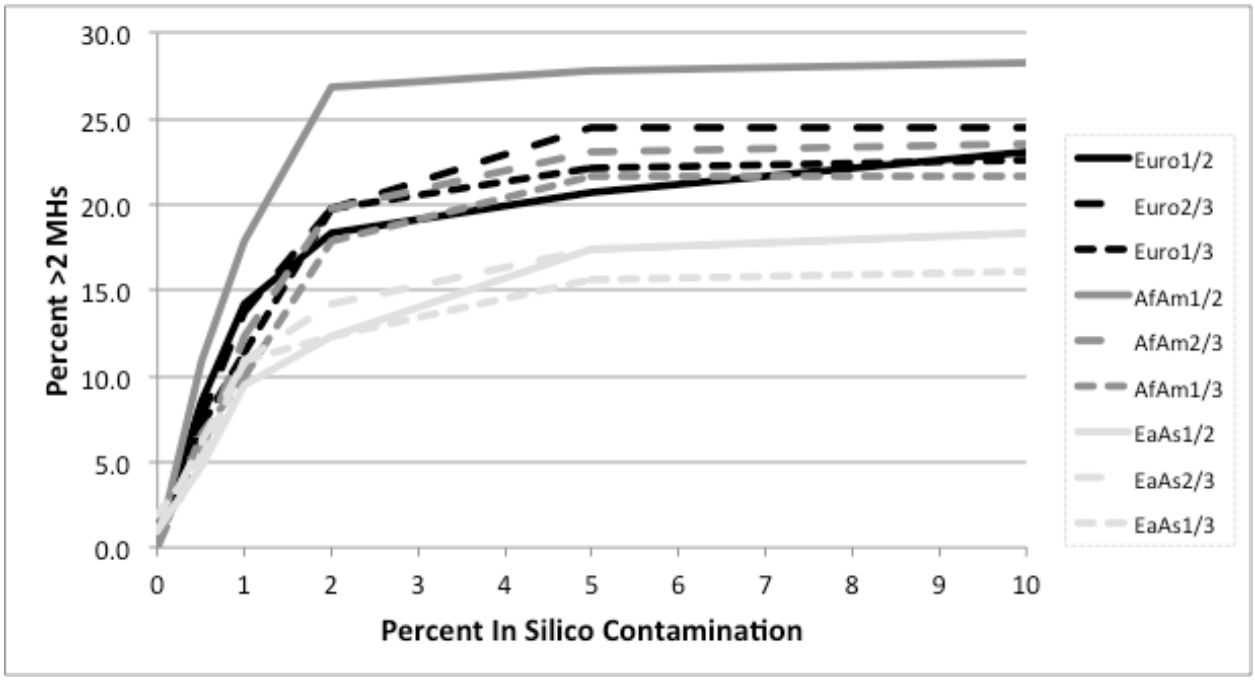
161 the initial choice of SNVs than variation in other ancestries. Thus, there is a possibility that samples from
162 different ancestral backgrounds could respond differently when examined independently using SNV sets
163 selected from gnomAD and not filtered based on ancestry. The initial testing of MH sets for establishing
164 contamination criteria was carried out using three individuals of African American ancestry, three
165 individuals of East Asian ancestry, and three individuals of European ancestry. These ancestries were
166 self-declared by the individuals and confirmed using Principle Component Analysis (PCA) of the MH sets
167 based on 1000 Genome MH frequencies. The self-declared and PCA-determined ancestries of these
168 samples were in agreement. However, with other samples, the self-declared Asian ancestries were
169 sometimes South Asian rather than East Asian. Only individuals with PCA-confirmed East Asian ancestry
170 were used in this and subsequent analyses. In addition to diverse ancestral histories, samples were
171 selected for mixing experiments based on having two or fewer SNV sets with >2 MHs.

172 Precise levels of read mixing can be generated *in silico* but the functional level of contamination was
173 found to often differ significantly from the known mixing proportions. This issue was particularly obvious
174 when samples with divergent DNA qualities were mixed. The longer, higher quality DNA appeared to
175 have a higher concentration than what had been added to the mix. This arises because longer reads map
176 better and are more likely to contain all the variants being examined and thus this DNA appears to be
177 relatively more frequent than the shorter, degraded DNA. This effect is seen with all types of DNA but is
178 especially noticeable with FFPE DNA where DNA quality is highly dependent on the methods used to fix
179 and store the sample. This is consistent with previous results that showed DNA with reduced extraction
180 yields tend to be lower quality and more susceptible to contamination (19). To minimize the impact of
181 this effect, all DNA mixes were examined in a pairwise fashion where one sample was the primary
182 sample and the second was the contaminant. These roles were then switched and the results averaged.
183 This allowed an over-contamination with one sample set to be partially mitigated by the paired under-
184 contamination of the opposite sample set.

185 As shown in Fig 1, the maximum number of MHs with 3rd haplotypes ranged from 15-30% with this set of
186 MHs. Most mixes had 20-25% 3rd haplotypes with significant overlap of individuals of European and
187 African American ancestry while those with East Asian ancestry had fewer 3rd MHs, in the range of 15-
188 20%. These results mirror the 1000 Genomes MH frequencies where the European and African American
189 frequencies for 3rd MHs are expected to be about 50% higher than East Asian frequencies. When mixes
190 were made with DNA from individuals of different ancestry, the fraction of MHs with 3rd
191 microhaplotypes increased with some combinations having >30% MHs with >2 haplotypes.

192 **Figure 1 Percent of SNV sets with >2 MHs for individual samples with 106 panel**

193 NGS reads from samples from different individuals of European (Euro), African American (AfAm), and
194 East Asian (EaAs) ancestry were mixed at frequencies of 0-10%. The number of MH sets with >2
195 haplotypes were counted and plotted for each individual pair.



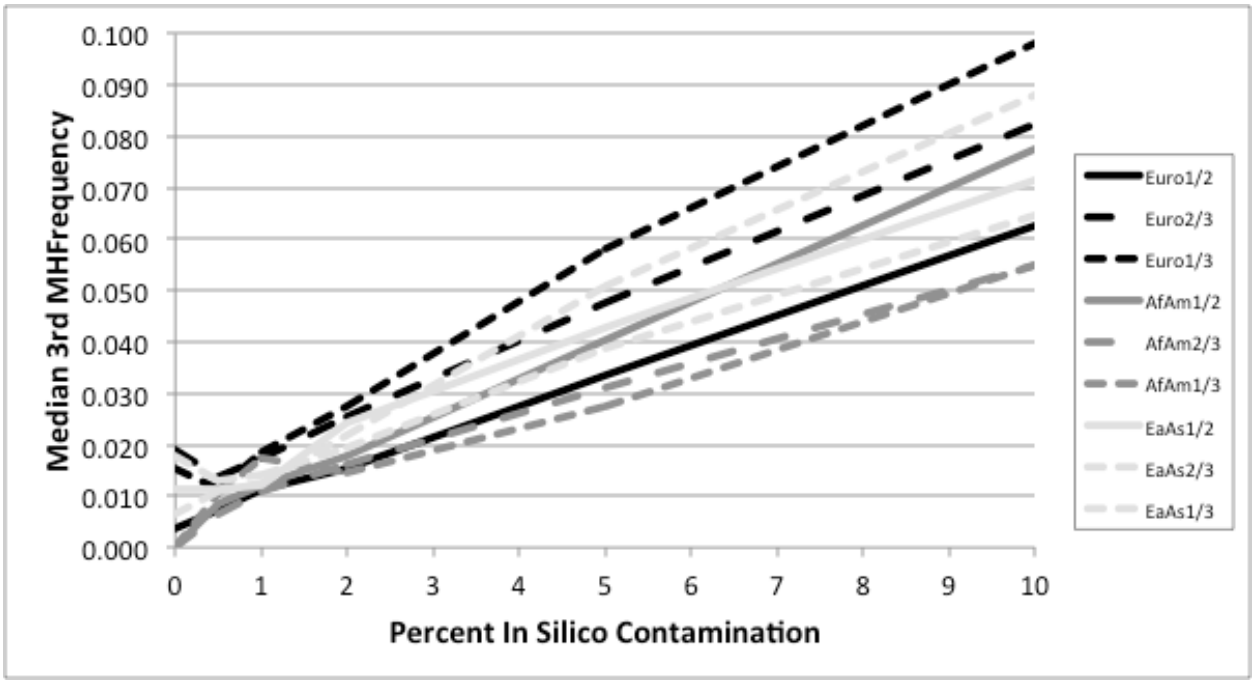
196

197 While the number of 3rd MHs is useful for assessing whether a sample is contaminated, it is less useful
198 for estimating the level of contamination as this value reaches a maximum at around 2% contamination

199 with these coverage levels. The 3rd MH frequency is more useful in this regard so we have used its
200 median frequency to estimate contamination level. As shown in Fig 2, there is a direct correlation
201 between the level of contamination and the 3rd MH median frequency. While this comparison does
202 provide an estimate of contamination levels, this panel and set of analysis parameters yields a nearly
203 two-fold difference in estimates of contamination for different sample pairs.

204 **Figure 2 Median Frequency for 3rd MHs in individual samples**

205 NGS reads from samples from different individuals of European (Euro), African American (AfAm), and
206 East Asian (EaAs) ancestry were mixed at frequencies of 0-10%. For all SNV sets with a 3rd MH, the
207 median frequency was calculated for each pair.



208

209 To solve the problem of reduced sensitivity with East Asian samples as well as the high variability of
210 contamination estimation for different individuals, new MH sets were identified and evaluated for
211 inclusion in a larger panel. Both old and new MHs were examined for optimal properties. 11 MHs from
212 the 106 panel with >5 SNV sets with >2 MHs were eliminated to reduce noise. Another 7 SNV sets with

213 median coverage <250x were eliminated so that more precise 3rd MH frequencies could be generated.
214 No new SNV sets with those properties were added to the revised panel.

215 In addition to basic quality metrics, the contamination assay should also be ancestry agnostic so that
216 contamination can be detected regardless of the DNA source. The population frequencies of 3rd and 4th
217 MHs were compared based on 1000 Genome data. SNV sets without such data were eliminated from
218 consideration on the revised panel. While it would have been desirable to use all of the population
219 groupings from 1000 Genomes in this analysis, there were not enough available SNV sets to allow
220 balancing of all five major ancestry groupings so only data from African, East Asian, and European
221 ancestry were applied to achieve balanced variation.

222 28 of the 106 MHs in the original panel had no 3rd MHs among East Asians. 23 of these were eliminated.
223 Other MHs that adversely affected ancestry balance were also eliminated. The average frequency of
224 3rd/4th MHs in the original panel ranges from a low of 8.2% for East Asians to a high of 12.7% in
225 Europeans. In the revised 164 MH panel, the average frequency of 3rd MHs was 13.9% for all three
226 ancestries. These results are shown in Table 1 for the complete 106 and 164 panels as well as the SNV
227 subsets that were shared or unique to each panel. The number of SNV sets with no 3rd MH frequency is
228 also shown for each ancestry.

229 **Table 1** Panel characteristics

230 Properties of the 106 and 164 SNV set panels are compared. In addition, the subsets of SNVs in only one
231 or both panels were also compared. Comparisons include the median coverage, the mean value of SNV
232 sets with >2 MHs, the total frequency of 3rd and 4th MHs among individuals of African (Afri), East Asian
233 (EaAs), European (Euro). Admixed American (AdAm), and African Americans (AfAm) as defined by the
234 1000 Genomes project. The number of SNV sets with no 3rd MH is also shown for each group.

					Mean Fraction of Population with 3rd and 4th MHs						Number of MHs with no examples of 3rd/4th MHs					
Panel	MHs with LD data	Med Cov	Mean Pure MH>2	Med Len	Afri	EaAs	Euro	AdAm	SoAs	AfAm	Afri	EaAs	Euro	AdAm	SoAs	AfAm
106	102	716	2.83	71	0.110	0.081	0.130	0.125	0.129	0.127	9	28	7	3	5	1
164	164	768	1.78	59	0.139	0.139	0.139	0.156	0.153	0.155	9	10	12	1	6	2
Both 106/164	50	776	1.66	46	0.122	0.120	0.127	0.142	0.143	0.140	4	5	3	1	2	1
Only 106	52	695	3.88	92	0.098	0.044	0.132	0.108	0.115	0.114	5	23	4	2	3	0
Only 164	114	768	1.83	61	0.146	0.147	0.144	0.161	0.157	0.161	5	5	9	0	4	1

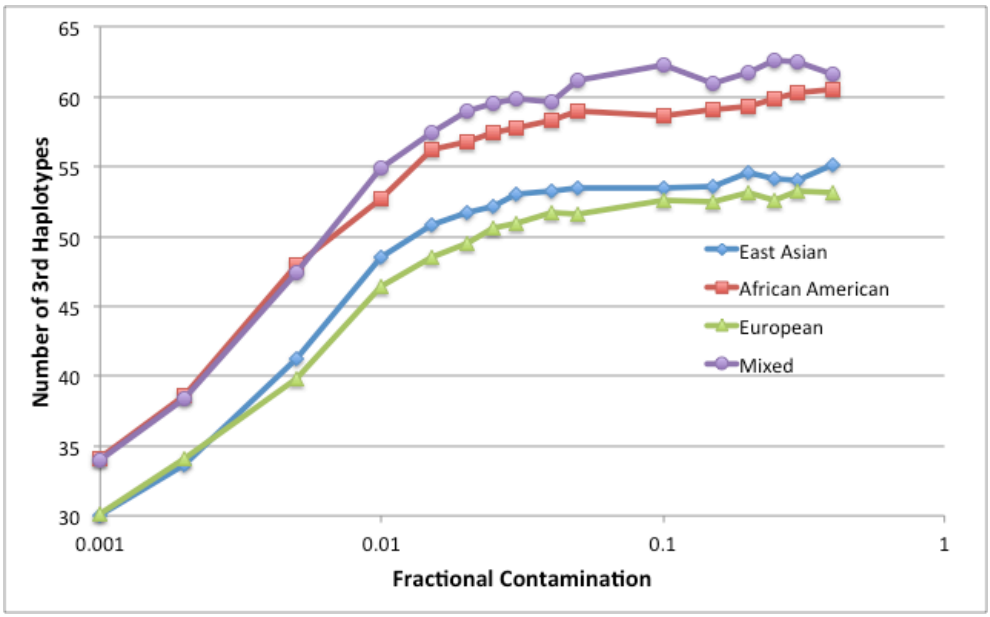
To assess the impact of the revised SNV sets, *in silico* contamination pairs were examined as before but with added constraints to minimize noise. To be considered contaminated, we used a threshold of a minimum of 25 SNV sets with >2 MHs with frequency >0.2%. 3rd MHs with frequency <0.2% are usually technical artifacts due to sequencing and thus artificially lower the median frequency if not eliminated. 15 sets of DNA from each ancestry were mixed as sample/contaminant pairs with averaged results shown in Fig 3. As expected, the East Asian fraction of >2 MHs was nearly identical to the European results. The African American samples that we examined had a higher frequency of >2 MHs than the other groupings but nearly identical to the samples mixed across ancestries. This is predicted by 1000 Genomes data. Their subgroupings of African ancestry included seven different sub-populations, only one of which was US-based. When the African American subgrouping is compared to the other African populations, it has a higher 3rd/4th MH frequency (Table 1) and thus the higher frequency in our population reflects the fact our African ancestry samples are primarily from African American individuals.

Figure 3 Percent of SNV sets with >2 MHs for individual samples with 164 panel

250 The number of 3rd MHs for pairs of individuals of East Asian, African American, and European ancestry is

251 shown for mixes of 0.1 – 40%. Results from pairs of individuals with different ancestry is included as

252 "Mixed", all on a log scale.



253

254 With the ability to confidently detect whether a sample is contaminated via the number of SNV sets with

255 >2 MHs, the next goal was to better estimate the level of contamination. The *in silico* mixing studies

256 provide an empirical calibration, but it is also useful to understand the theoretical basis for those

257 findings. As shown in Fig 4, the expected frequency of the contaminating 3rd MH depends on the nature

258 of the starting and incoming genotypes. Some combinations will generate no 3rd MH, others will

259 generate a 3rd MH with a frequency half the level of incoming contaminant, and others will generate a

260 genotype that is the same frequency as the contaminant. Since most variant combinations will be at the

261 50% level, examination of the median value for the 3rd MH should yield a value that is half of the true

262 contamination level.

263 **Figure 4 Expected frequencies for 3rd MHs based on genotype**

Primary Sample		Contaminant (X)		Rate of 3rd MH Accumulation
MH1/MH1	+	MH1/MH3	→	no 3rd MH
		MH2/MH3	→	0.5 X
		MH3/MH3	→	no 3rd MH
		MH4/MH3	→	0.5 X
MH1/MH2	+	MH1/MH3	→	0.5 X
		MH2/MH3	→	0.5 X
		MH3/MH3	→	1X
		MH4/MH3	→	0.5X

The primary limitations on the sensitivity of contamination detection are the practical sequencing depths for the SNV sets in question and the inherent error rates for sequencing. The higher the sensitivity desired, the greater sequencing depth required. The usefulness of added depth is limited by sequencing error rates that are typically on the order of 0.1% for Illumina. It is possible to employ various error correction methods to add to sensitivity, but the need for that is dependent on the sensitivity required for the application and its usefulness limited by the available sequencing depth. Based on these limitations, we have aimed for detection of contamination at least as low as 1% in these samples.

For determining the 3rd allele frequency, only those MHs with >0.2% frequency are used to avoid over-inclusion of artifactual MHs that arise from sequencing errors. Additionally, if there are only a few 3rd MHs, the median frequency could be heavily skewed by having so few values. To avoid this, only samples with >25 of the potential 164 3rd MHs are considered contaminated. As seen in Fig 3, the average value for the number of 3rd MHs among all ancestries with 1% contamination is >45 with this panel so this threshold should not eliminate any truly contaminated samples. All 3rd MHs with frequency >0.2% are then used for determining the median frequency.

280 As shown in Fig 5, contamination level using this panel is independent of ancestry. All three ancestries
281 and mixed samples behave nearly identically. A more detailed listing of results is listed in Table 2 where
282 the individual ancestries and mixed samples are shown with various thresholds for detection. 15
283 sample/contaminant pairs were tested for each ancestry with the same number of each mixed ancestry
284 pairs. The observed frequencies of the median 3rd MH (e.g. 0.005) are multiplied by 2 based on the
285 discussion above to indicate the actual contamination levels within the sample. When all of the samples
286 are aggregated, 40% of samples with 0.5% contamination were detected. If the threshold is set at 1%, all
287 contaminated samples were detected. The detection threshold can be set wherever required based on
288 the detection needs and the likely impact of contamination on final results.

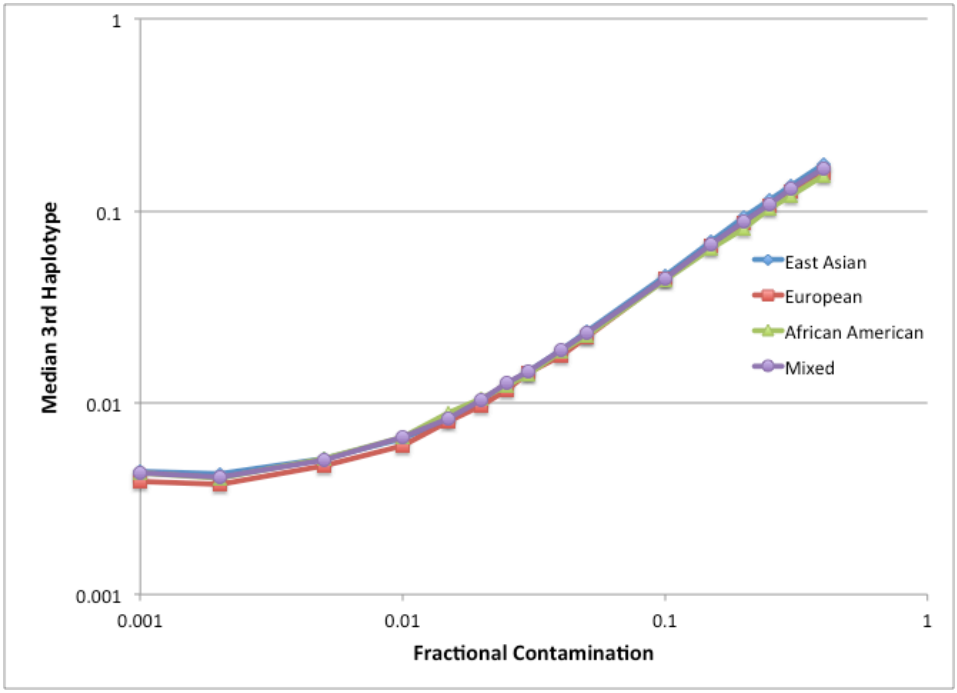
289 **Table 2 Detection efficiency versus contamination level**

		Observed 3rd MH Frequency (x2)								
East Asian		1	1.5	2	2.5	3	4	5	7	9
In silico Mixing Levels	0.5	8	0	0	0	0	0	0	0	0
	1	15	2	0	0	0	0	0	0	0
	1.5	15	12	0	0	0	0	0	0	0
	2	15	14	10	0	0	0	0	0	0
	2.5	15	15	15	8	0	0	0	0	0
	3	15	15	15	15	6	0	0	0	0
	4	15	15	15	15	15	3	0	0	0
	5	15	15	15	15	15	15	1	0	0
	10	15	15	15	15	15	15	15	15	9
African American		1	1.5	2	2.5	3	4	5	7	9
In silico Mixing Levels	0.5	3	0	0	0	0	0	0	0	0
	1	15	0	0	0	0	0	0	0	0
	1.5	15	10	0	0	0	0	0	0	0
	2	15	14	5	0	0	0	0	0	0
	2.5	15	15	15	4	0	0	0	0	0
	3	15	15	15	14	5	0	0	0	0
	4	15	15	15	15	13	1	0	0	0
	5	15	15	15	15	15	12	2	0	0
	10	15	15	15	15	15	15	15	14	7
European		1	1.5	2	2.5	3	4	5	7	9

	0.5	8	0	0	0	0	0	0	0	0
	1	15	4	0	0	0	0	0	0	0
	1.5	15	13	4	0	0	0	0	0	0
In silico	2	15	15	12	0	0	0	0	0	0
Mixing	2.5	15	15	15	8	0	0	0	0	0
Levels	3	15	15	15	13	4	0	0	0	0
	4	15	15	15	14	14	3	0	0	0
	5	15	15	15	15	15	12	1	0	0
	10	15	15	15	15	15	15	15	13	7
Mixed		1	1.5	2	2.5	3	4	5	7	9
	0.5	5	0	0	0	0	0	0	0	0
	1	15	3	0	0	0	0	0	0	0
	1.5	15	14	0	0	0	0	0	0	0
In silico	2	15	15	11	0	0	0	0	0	0
Mixing	2.5	15	15	15	7	1	0	0	0	0
Levels	3	15	15	15	15	6	0	0	0	0
	4	15	15	15	15	15	2	0	0	0
	5	15	15	15	15	15	14	0	0	0
	10	15	15	15	15	15	15	15	14	9
All (%)		1	1.5	2	2.5	3	4	5	7	9
	0.5	40	0	0	0	0	0	0	0	0
	1	100	15	0	0	0	0	0	0	0
	1.5	100	82	7	0	0	0	0	0	0
In silico	2	100	97	63	0	0	0	0	0	0
Mixing	2.5	100	100	100	45	2	0	0	0	0
Levels	3	100	100	100	95	35	0	0	0	0
	4	100	100	100	98	95	15	0	0	0
	5	100	100	100	100	100	88	7	0	0
	10	100	100	100	100	100	100	100	93	53

Figure 5 Contamination Detection as a function of ancestry

The median frequency of 3rd MHs for pairs of individuals of East Asian, African American, and European ancestry is shown. Results from pairs of individuals with different ancestry is included as “Mixed”, all on a log scale for mixing levels of 0.1 – 40%.



295

296 In addition to detecting contamination, it is also possible to use this method to identify sources of
297 contamination. A partial set of genotypes for the contaminating DNA can be generated from the MH
298 data and then matched with samples that have been tested previously or contemporaneously to
299 determine if they are the source of contamination. In a contaminated sample, MHs with 1 or 4 different
300 haplotypes can be assigned their correct genotypes directly. MHs with 2 or 3 different haplotypes can
301 generally be assigned a partial or complete genotype based on frequency. The frequency of the $2^{nd}/3^{rd}$
302 haplotypes will provide at least one and sometimes two genotypes corresponding to the contaminant.
303 Comparison of the known and partial genotypes with other samples allows them to be ruled in or out as
304 potential sources of contamination. This procedure has been applied to multiple contaminated samples
305 to successfully determine the source in real clinical laboratory situations. Often, the contamination
306 source can be shown not to arise from any samples known to have been handled in the laboratory so
307 the contamination must have originated prior to sample receipt.

308

309 **Discussion**

310 The level of DNA contamination that is an issue varies across applications. In some cases, contamination
311 levels that impact results are obvious. With other applications, signals arising from contamination can be
312 easily confused with real signals. With diagnosis of somatic variants in cancer, this is especially a
313 problem because variant identification is attempted at and below the very limits of variant detection.
314 Being able to separate the tumor signal from the contamination noise is critical for the proper diagnosis
315 of tumor etiology. Because tumor samples often have highly unusual copy number patterns, use of
316 simple genotype frequencies can cause errors in contaminant detection. Multiple samples that we
317 assessed as contaminated based on standard methods using raw genotype frequencies were found to
318 be pure by MH analysis. In contrast to the FFPE tumor samples, circulating free DNA (cfDNA) purified
319 from the plasma of cancer patients is not as prone to the same issues because most of the cfDNA
320 purified from plasma is from normal cells so typical Mendelian ratios are observed. MH analysis in such
321 samples is still valuable for sensitive contamination detection, but such samples are not as likely to
322 generate false positives as DNA from FFPE tumor samples where copy number variation is much more
323 significant.

324 For most purposes, the approximation that the percent contamination is twice the median frequency for
325 the 3rd MH is sufficient. This relationship arises from the likelihood that a 3rd MH is most likely
326 heterozygous so is present on only one of the two incoming alleles. The uncertainty around this
327 probability introduces the potential for multiple sources of error and those should be considered if a
328 more exact measurement is necessary. At low coverage, two confounding factors arise. First, low
329 coverage leads to high AF variability due to the stochastic nature of read accumulation in poorly covered
330 SNV sets. Only SNV sets with 3rd MHs are included in the median calculation rather than all SNV sets so,
331 at low coverage, AF inflation occurs among the variants that are counted. The observed AF is higher than
332 reality because there are not enough reads to generate the proper AF value. By requiring a large number

333 of 3rd MHs prior to calling a sample contaminated, this effect is mitigated. Superimposed on this effect is
334 the appearance of homozygous 3rd MHs before heterozygous 3rd MHs as coverage increases. If coverage
335 is deep, this is not an issue but, at low coverage, homozygous variants will appear first because they are
336 being double counted. Further complicating the relationship is the presence of low frequency technical
337 artifacts and the copy number issues that can occur with cancer samples. Many cancer samples have
338 large copy number variations so different SNV sets will appear at different fractional rates depending on
339 whether the 3rd MH is under- or over-represented in copy number changes. The more MH sets used in
340 the calculations, the more likely it is that these effects can be averaged out.

341 The MH approach is perfectly suited for use when somatic variants are sought because such studies use
342 very high coverage in order to detect rare variants. The sensitivity of contamination detection is
343 dependent on the coverage of the regions used in the analysis. Thus, other panel-based and exome NGS
344 approaches are also immediately extendable to similar studies. A typical whole genome approach with
345 less than 50x coverage would not be amenable to sensitive detection using individual MH sets because
346 the low coverage would prevent the observation of infrequent 3rd MHs. To overcome this, combinations
347 of MH sets could be used. The simplest approach would be to combine MH sets that are in high LD and
348 combine the appropriate SNV calls in order to achieve the needed depth of coverage across multiple
349 high LD regions. Other methods for pooling MH sets could also be used. With the entire genome as a
350 source, MH sets that behave as needed and are subject to the lowest error rates could be employed.

351 The same techniques used for these cancer samples can also be used with forensic samples to detect
352 major and minor contributors to a sample. Unlike tumor samples, copy number should not be an issue,
353 which makes interpretation easier. Like the tumor samples we have examined, the major and minor
354 contributors to a sample may have different DNA qualities that can affect the accuracy of MH
355 frequencies. Some MHs used in these panels have more than 4 potential haplotypes so could indicate
356 more than one contaminant. Panels designed with a larger number of >4 potential MHs would be

357 appropriate when forensic samples are being examined in order to clarify the number of contributors to
358 a sample. As with contamination assignment described above, MH assignment can be accomplished at
359 high enough efficiency to identify possible sources of the contamination as well as ancestry.

360 Another useful application of the MH methods is for more sensitive detection of fetal DNA in Non-
361 invasive prenatal testing (NIPT). Panels can be designed to focus on genomic regions with the most
362 likely potential copy number variation (e. g. chr 5, chr13, chr18, or chr21) and the frequency of 3rd MHs
363 in targeted regions compared to the rest of the genome. In all regions, fetal DNA will share one
364 genotype with the mother (main contributor of DNA) and have one from the father. In genomic regions
365 where the mother is heterozygous and the father's contribution to the fetal DNA is a different allele,
366 there will be a low frequency 3rd MH. If the frequency of that 3rd MH varies across genomic regions, it is
367 indicative of a copy number variation. Because the fetal "contamination" signal will always be just a
368 single copy of non-matching genotype for a single birth, the frequency distributions will be less variable
369 than with a typical contaminant. For fraternal twin births, MH frequencies should be the sum of the
370 individual fetal contributors. For SNV sets where both parents are heterozygous with four different MHs,
371 it would be possible to see 4 MHs rather than just 3. With the potential for high copy depth and lower
372 detection levels, NIPT could be done earlier in a pregnancy using the MH approach rather than standard
373 methods.

374 The methods described here provide a foundation for examining DNA mixtures in a variety of contexts.
375 With the appropriate choice of SNV panels and high sequencing depth, very low levels of secondary DNA
376 can be detected and this information can be valuable in a variety of applications.

377

378 *Abbreviations*

379 **AdAm:** Admixed American

380 **AF:** Allele Frequency

381 **AfAm**: African American

382 **Afri:** African

383 **cfDNA:** circulating free DNA

384 **EaAs:** East Asian

385 **Euro:** European

386 **FFPE:** Formalin-fixed Paraffin-embedded

387 **FP:** False Positive

388 **LD:** Linkage Disequilibrium

389 **MH:** Microhaplotype

390 **NGS:** Next-generation Sequencing

391 **NIPT:** Non-invasive Prenatal Testing

392 **PCA:** Principle Component Analysis

393 **SNV:** Single Nucleotide Variant

394 **SoAs:** South Asian

395 **STR:** Short Tandem Repeat

396

397 **References**

398 1. Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM. Correcting for Sample Contamination in

399 Genotype Calling of DNA Sequence Data. Am J Hum Genet. 2015 Aug 6;97(2):284-90.

400 2. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-

401 contamination of human samples in next-generation sequencing data. Bioinformatics. 2011 Sep

402 15;27(18):2601-2.

403 3. Sakarya, 2017, https://grail.com/wp-

404 content/uploads/2018/05/HiTSeq_2017_Sakarya_Conta_POS_Final.pdf

405 4. Fiévet A, Bernard V, Tenreiro H, Dehainault C, Girard E, Deshaies V et al. ART-DeCo: easy tool for
406 detection and characterization of cross-contamination of DNA samples in diagnostic next-generation
407 sequencing analysis. Eur J Hum Genet. 2019 May;27(5):792-800.

408 5. Kidd KK, Pakstis AJ, Speed WC, Lagace R, Chang J, Wootton S, et al. Microhaplotype loci are a powerful
409 new type of forensic marker. FSI Supplement Series. 2013. 4 (1):e123-e124.

410 6. Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, et al. Current sequencing technology
411 makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet.
412 2014 Sep;12:215-24.

413 7. Kidd KK, Speed WC. Criteria for selecting microhaplotypes: mixture detection and deconvolution.
414 Investig Genet. 2015 Jan 28;6:1.

415 8. Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. Forensic Sci Int Genet. 2019 Jan;
416 38:54-69.

417 9. Debeljak M, Freed DN, Welch JA, Haley L, Beierl K, Iglehart BS, et al. Haplotype counting by next-
418 generation sequencing for ultrasensitive human DNA detection. J Mol Diagn. 2014 Sep;16(5):495-503.

419 10. Bulbul O, Pakstis AJ, Soundararajan U, Gurkan C, Brissenden JE, Roscoe JM, et al. Ancestry inference
420 of 96 population samples using microhaplotypes. Int J Legal Med. 2018 May;132(3):703-711.

421 11. Kidd KK, Speed WC, Pakstis AJ, Podini DS, Lagacé R, Chang J, et al. Evaluating 130 microhaplotypes
422 across a global set of 83 populations. Forensic Sci Int Genet. 2017 Jul;29:29-37.

423 12. Oldoni F, Yoon L, Wootton SC, Lagacé R, Kidd KK, Podini D. Population genetic data of 74
424 microhaplotypes in four major U.S. population groups. Forensic Sci Int Genet. 2020; 49:102398.

425 13. Phillips C, McNevin D, Kidd KK, Lagacé R, Wootton S, de la Puente M, et al. MAPlex - A massively
426 parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. Forensic Sci Int Genet. 2019
427 Sep;42:213-226.

428 14. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers
429 using circulating tumor DNA. Sci Transl Med. 2017 9:2415.

430 15. Georgiadis A, Durham JN, Keefer LA, Bartlett BR, Zielonka M, Murphy D, et al. Noninvasive Detection
431 of Microsatellite Instability and High Tumor Mutation Burden in Cancer Patients Treated with PD-1
432 Blockade. Clin Cancer Res. 2019 25:7024-7034.

433 16. Wood DE, White JR, Georgiadis A, Van Emburgh B, Parpart-Li S, Mitchell J, et al. A machine learning
434 approach for somatic mutation discovery. Sci Transl Med. 2018 10:7939. PMID: 3018562

435 17. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint
436 spectrum quantified from variation in 141,456 humans. Nature. 2020 581: 434-443.

437 18. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific
438 haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015 Jul
439 2

440 19. Sehn JK, Spencer DH, Pfeifer JD, Bredemeyer AJ, Cottrell CE, Abel HJ, et al. Occult Specimen
441 Contamination in Routine Clinical Next-Generation Sequencing Testing. Am J Clin Pathol. 2015
442 Oct;144(4):667-74.

443