



# ART-DeCo: easy tool for detection and characterization of cross-contamination of DNA samples in diagnostic next-generation sequencing analysis

Alice Fiévet<sup>1,2</sup> · Virginie Bernard<sup>3</sup> · Henrique Tenreiro<sup>1</sup> · Catherine Dehainault<sup>1</sup> · Elodie Girard<sup>4,5,6,7</sup> · Vivien Deshaies<sup>4,5,6,7</sup> · Philippe Hupe<sup>4,5,6,7</sup> · Olivier Delattre<sup>1,2</sup> · Marc-Henri Stern<sup>1,2</sup> · Dominique Stoppa-Lyonnet<sup>1,2,8</sup> · Lisa Golmard<sup>1</sup> · Claude Houdayer<sup>1,2,8,9</sup>

Received: 28 June 2018 / Revised: 20 October 2018 / Accepted: 4 December 2018 / Published online: 25 January 2019  
© European Society of Human Genetics 2019

## Abstract

Next-generation sequencing (NGS) is routinely used for constitutional genetic analysis. However, cross-contamination between samples constitutes a major risk that could impact the results of the analysis. We have developed ART-DeCo, a tool using the allelic ratio (AR) of the Single Nucleotide Polymorphisms sequenced with regions of interest. When a sample is contaminated by DNA with a different genotype, unexpected ARs are obtained, which are in turn used for detection of contamination with a screening test, followed by identification and quantification of the contaminant. Following optimization, ART-DeCo was applied to 2222 constitutional DNA samples. The screening test was positive for 191 samples. In 33 cases (contamination percentages: 1.3% to 29.2%), the contaminant was identified and was mostly located in adjacent wells. Three other positive cases were due to barcoding errors or mixture of two DNA samples. Interestingly, the last contaminated sample corresponded to a bone marrow transplant recipient. Lastly, no contaminant was identified in 154 weakly positive (<4%) samples that were considered to be irrelevant to constitutional genetic analysis. ART-DeCo lends itself to mandatory quality control procedures, also highlighting the delicate steps of library preparation, resulting in practice improvement. Importantly, ART-DeCo can be implemented in any NGS workflow, from gene panel to genome-wide analyses. <https://sourceforge.net/projects/ngs-art-deco/>.

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41431-018-0317-x>) contains supplementary material, which is available to authorized users.

✉ Alice Fiévet  
Alicefievet@hotmail.com

- <sup>1</sup> Department of Genetics, Institut Curie, 75005 Paris, France
- <sup>2</sup> Institut Curie, PSL Research University, INSERM U830, Institut de recherche, F-75005 Paris, France
- <sup>3</sup> Institut Curie Genomics of Excellence (ICGex) platform, Institut Curie Research Center, Paris, France
- <sup>4</sup> Institut Curie, 26 rue d'Ulm, F-75005 Paris, France
- <sup>5</sup> PSL Research University, F-75005 Paris, France
- <sup>6</sup> Inserm, U900, F-75005 Paris, France
- <sup>7</sup> Mines Paris Tech, F-77305 cedex Fontainebleau, France
- <sup>8</sup> University Paris Descartes, Sorbonne Paris Cité, France
- <sup>9</sup> Department of Genetics, F76000 and Normandy Univ, UNIROUEN, Inserm U1245, Normandy Centre for Genomic and Personalized Medicine, Rouen University Hospital, Rouen, France

## Introduction

Next-generation sequencing (NGS) is now routinely used in many diagnostic genetic laboratories, as it allows simultaneous sequencing of multiple genes for multiple patients, which is a more cost-effective and more rapid approach than Sanger sequencing [1].

The main steps of the NGS technique are: enrichment of DNA regions of interest via DNA hybridization capture or PCR amplification, identification of each patient's DNA with a barcode, pooling of patient DNA, sequencing by NGS sequencer and bioinformatics analysis of the raw data.

During the NGS preparation step, patient DNA samples are processed in parallel until the barcoding step. This step can be performed at different times, according to the type of NGS preparation technique, but always before pooling of samples to create the library. After sequencing of the library, each amplicon sequenced is assigned to a patient, according to its barcode.

Sequenced DNA is compared to the reference genome and the patient's variants are listed in order to identify relevant variants, corresponding to the variant calling step. The majority of the variants listed are usually single nucleotide polymorphisms (SNPs). In constitutional genetic analysis, the variants detected are characterized by their allelic ratio (AR): for a given variation, the AR is defined as the (number of reads supporting the variant)/(number of reads at this position). A variant is expected to have an AR around 0.5 when the individual is heterozygous for this variant. It is noteworthy that, in the case of mosaicism, a variant is present in only a portion of the individual's cells and the expected AR is then  $<0.5$ . If the individual is homozygous for the reference allele or for the alternative allele, the expected AR is 0 or 1, respectively.

Sample contamination is a major risk in NGS diagnosis, and needs to be controlled, as series of samples are processed in parallel. Sample contamination can lead to failure of identification of a variant affecting function in the patient, which can be masked by the large quantity of contaminant. Another major risk is to wrongly conclude on the presence of a variant, which actually corresponds to the contaminant. This risk is especially relevant for diseases associated with de novo variant, in which mosaicism can occur.

In order to address this important issue, we have therefore developed a tool (ART-DeCo: Allelic Ratio-based Tool for Detection of Contamination) designed to detect contamination in constitutional NGS analysis. The strategy of this tool is based on the detection of SNPs presenting ARs not usually expected in constitutional analyses. ART-DeCo can be easily implemented in any NGS workflow to control for sample contamination.

## Materials and methods

### Library preparation and sequencing

Library preparation was performed manually with SureSelect QXT kit on a home-made 384 kb gene panel (Agilent Technologies). The first step of this preparation consisted of dilution of the patients' gDNA in a  $4 \times 8$  plate (4 columns of standard 96-wells plate). gDNA was then fragmented and adaptor-tagged. The library was purified using Agencourt AMPure XP beads (Beckman Coulter), amplified and re-purified. Samples were hybridized to the capture probes, and then captured using streptavidin-coated beads (Dynabeads MyOne Streptavidin T1, Life Technologies). Libraries were amplified to add barcodes and purified using beads. Libraries were then pooled and sequenced with a NextSeq 500<sup>®</sup> sequencer (Illumina<sup>®</sup>).

### Nomenclature

For each SNP, a wild-type homozygous sample is indicated as Ref/Ref, whereas a homozygous sample for the alternative allele is indicated as Alt/Alt. Heterozygous samples are indicated as Ref/Alt.

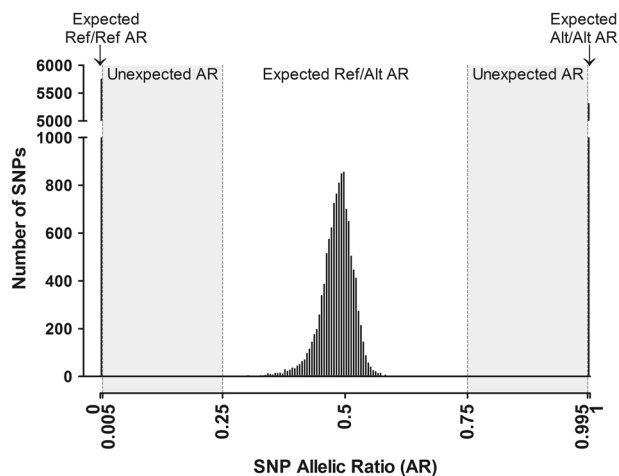
### Rationale

This strategy is based on the detection of SNPs presenting unexpected AR for constitutional analyses, i.e. distortions from 0 (homozygous wild-type Ref/Ref), 1 (homozygous alternate Alt/Alt) and 0.5 (heterozygous Ref/Alt) (Fig. 1). For each SNP, the AR of heterozygous samples should be 0.5, but it actually fluctuates around this value, e.g. due to differences in mapping quality between reads with or without mismatches. The heterozygous range was defined as [0.25–0.75]. Note that this [0.25–0.75] range could be restrained and adapted to the heterozygous distribution if this distribution is known. However, it is not needful. Indeed, homozygous SNPs with 50% maximum contamination level result in AR values fluctuating around 0.25 and around 0.75, whereas lower contamination levels result in AR values  $<0.25$  or  $>0.75$  (Supplementary Table 1, Supplementary Fig. 1). Consequently, the [0.25–0.75] range allows accurate discrimination between poorly called heterozygous SNPs and sample contamination.

For each SNP, the AR of homozygous samples is theoretically 0 or 1, but in practice is slightly different because of the background noise generated by polymerase, sequencing and alignment errors, index hopping or incomplete trimming of the adaptors (see below AR extraction) [2–4]. Background noise is generally low in Illumina<sup>®</sup> sequencing and expected values are usually observed. As the minimum depth for the SNPs under study was set at  $200\times$  (see below SNP selection), the background noise was set at 0.5%, in order to tolerate at least one read as default background. The expected AR intervals of homozygous SNPs used were therefore [0–0.005 [and] 0.995–1] for Ref/Ref and Alt/Alt genotypes, respectively. Consequently, unexpected ratios were situated in the [0.005–0.25 [and] 0.75–0.995] ranges (Fig. 1).

### AR extraction

After trimming adaptors by Cutadapt using default parameters [5], reads were aligned via Bowtie2 allowing up to one mismatch in the 22 bp-long seed and reporting only unique alignments [6]. Reads with mapping quality less than 20 were filtered out. Variant calling software was not used, as we wanted to report any frequencies within a focused list of SNPs. The Depth Of Coverage function from the Genome Analysis Toolkit (GATK) was used [7],



**Fig. 1** Trimodal distribution of allelic ratios (ARs) of SNPs. AR values for 14 SNPs in 1650 samples were used. For homozygous SNPs (Ref/Ref or Alt/Alt), the observed ARs are [0;0.005 [or] 0.995;1] respectively. For heterozygous SNPs (Ref/Alt), the observed ARs are [0.25;0.75]. SNPs with AR [0.005;0.25 [or] 0.75;0.995] are unexpected in uncontaminated DNA samples

together with additional statistical analysis detailed below to report ARs of SNPs. To ensure analysis of high-quality data, only base quality  $\geq 20$  were considered for determination of the depth of coverage of the selected polymorphisms.

### SNP selection

For each sample, the AR distribution of an SNP selection was computed by an algorithm in order to detect, identify and quantify the contaminant (Supplementary Fig. 2).

SNP selection only retained informative polymorphisms with a typical trimodal AR distribution. Among the 628 SNPs, with an European population frequency in the range 0.1–99.9% from the 1000 Genomes database and present in our 60-gene panel, those with recurrent high background noise (e.g. close to homopolymer stretches) were excluded (see “optimization step” section below). Similarly, polymorphisms within paralogous genes were excluded to avoid misinterpretation of AR spoiled by expected misalignment. A total of 547 polymorphisms were then able to be analyzed. Only SNPs with at least 200 $\times$  coverage were taken into account to allow detection of low contamination. Homozygous SNPs for the same allelic version throughout the samples were non-informative and could not be used for analysis.

### Detection of contamination: “worst-case scenario” screening test

The first step of identification of contamination consisted of a screening test for each sample of the run, based on

estimation of the “worst-case scenario” (WCS) percentage of contamination. This screening test is independent of background noise and identifies samples possibly contaminated above a certain cutoff, defined as 1% in the present study.

Following the optimization step (see below), the WCS calculation was defined as:

$WCS = \max(r \times 2; (1 - a) \times 2)$ ; with  $r$  = median of the highest 2% of ARs of Ref/Ref SNPs and  $a$  = median of the lowest 2% of ARs of Alt/Alt SNPs.

The main advantage of the WCS test is to rapidly rule out any contamination when it is negative. However, it has a low specificity and a positive WCS test must be confirmed by identification of the contaminant, as the worst scenario is never certain.

### Identification of the contaminant

Contamination was suspected when the WCS percentage of contamination was  $\geq 1\%$ . The second step consisted of identification of the contaminant in order to confirm the contamination. This identification was based on the SNPs of the contaminated sample (i.e. its genotype) compared to the genotypes of the other samples of the run.

Only homozygous SNPs of the contaminated sample were used, as heterozygous SNPs exhibited excessive variability of AR values to allow reliable identification of small variations corresponding to low-level contamination. Only SNPs with AR  $< 0.25$  or  $> 0.75$  were used, corresponding to homozygous SNPs (Ref/Ref or Alt/Alt), including contaminated ( $< 0.25$  or  $> 0.75$ ) or non-contaminated SNPs ( $< 0.005$  or  $> 0.995$ , i.e. background noise).

To identify a putative contaminant, the percentage of SNPs compatible with contamination of one sample (A) by another sample (B) was calculated according to the number of homozygous SNPs satisfying the compatibility conditions listed in Table 1. In other words, for homozygous SNPs with expected AR values, the contaminant had to have the same genotype, while for SNPs with unexpected AR values, the contaminant had to have a different genotype. The suspected contaminant was therefore identified by its genotype. For each sample, the other samples from the same run were tested, scored and ranked as putative contaminants and the sample with the highest percentage was considered to be a putative contaminant.

To determine whether the putative contaminant actually contaminated the sample under study, two conditions were then required. Firstly, the percentage of SNPs compatible with contamination of sample A by sample B had to be higher than the percentage of SNPs of sample A, compatible with absence of contamination; otherwise sample B could simply be genetically similar to sample A. The percentage

**Table 1** Genotype criteria for a contaminant (column 2) according to the genotype of the contaminated sample (column 1)

AR/genotype of a SNP of sample (A)	AR/genotype of the same SNP of the tested contaminant (B)	SNP compatible with contamination of sample A by sample B
0–0.005 (Ref/Ref)	0–0.25 (Ref/Ref)	Yes
	0.25–1 (Ref/Alt or Alt/Alt)	No
0.005–0.25 (Ref/Ref)	0.25–1 (Alt/Ref or Alt/Alt)	Yes
	0–0.25 (Ref/Ref)	No
0.75–0.995 (Alt/Alt)	0–0.75 (Ref/Alt or Ref/Ref)	Yes
	0.75–1 (Alt/Alt)	No
0.995–1 (Alt/Alt)	0.75–1 (Alt/Alt)	Yes
	0–0.75 (Ref/Alt or Ref/Ref)	No

Genotype compatibility (yes/no) is indicated in column 3

**Table 2** Calculation of the level of contamination by the contaminant

Observed AR of an SNP of sample A (AR(A))	AR of the same SNP of contaminant B	Calculation of the percentage of contaminant DNA (B) in sample (A) for each SNP
0–0.25	0.25–0.75	$(0 + \text{AR}(A)) \times 2$
0–0.25	0.75–1	$(0 + \text{AR}(A))$
0.75–1	0.25–0.75	$(1 - \text{AR}(A)) \times 2$
0.75–1	0–0.25	$(1 - \text{AR}(A))$

After identifying the contaminant, for each homozygous SNP, the difference between the theoretical AR (0 for Ref/Ref, 1 for Alt/Alt) and the observed AR (column 1) is used to calculate the contamination percentage (column 3) according to the genotype of the contaminant (column 2)

of SNPs compatible with absence of contamination was the percentage of SNPs with a normal AR  $<0.005$  or  $>0.995$  among the total number of homozygous SNP (i.e. with AR  $<0.25$  or AR  $>0.75$ ) (Supplementary Table 2). Secondly, the putative contaminant had to be significantly more compatible with the contaminated sample than the other samples of the run (Fisher's exact test with Bonferroni correction for multi-testing, limit of significance 0.05).

### Quantification of contamination

As the WCS contamination is only a rough, overestimated value, a refined percentage is calculated following identification of the contaminant and according to its genotype. The contamination percentage of a sample by its contaminant is expressed as the median of the values obtained for calculation of contamination rate for each SNP (Table 2).

All samples were collected for diagnostic and genetic counselling purposes. Appropriate individual written consent for genetic analysis was obtained from all the participating patients or their legal guardians.

### Availability

This tool is available at: <https://sourceforge.net/projects/ngs-art-deco/>

## Results

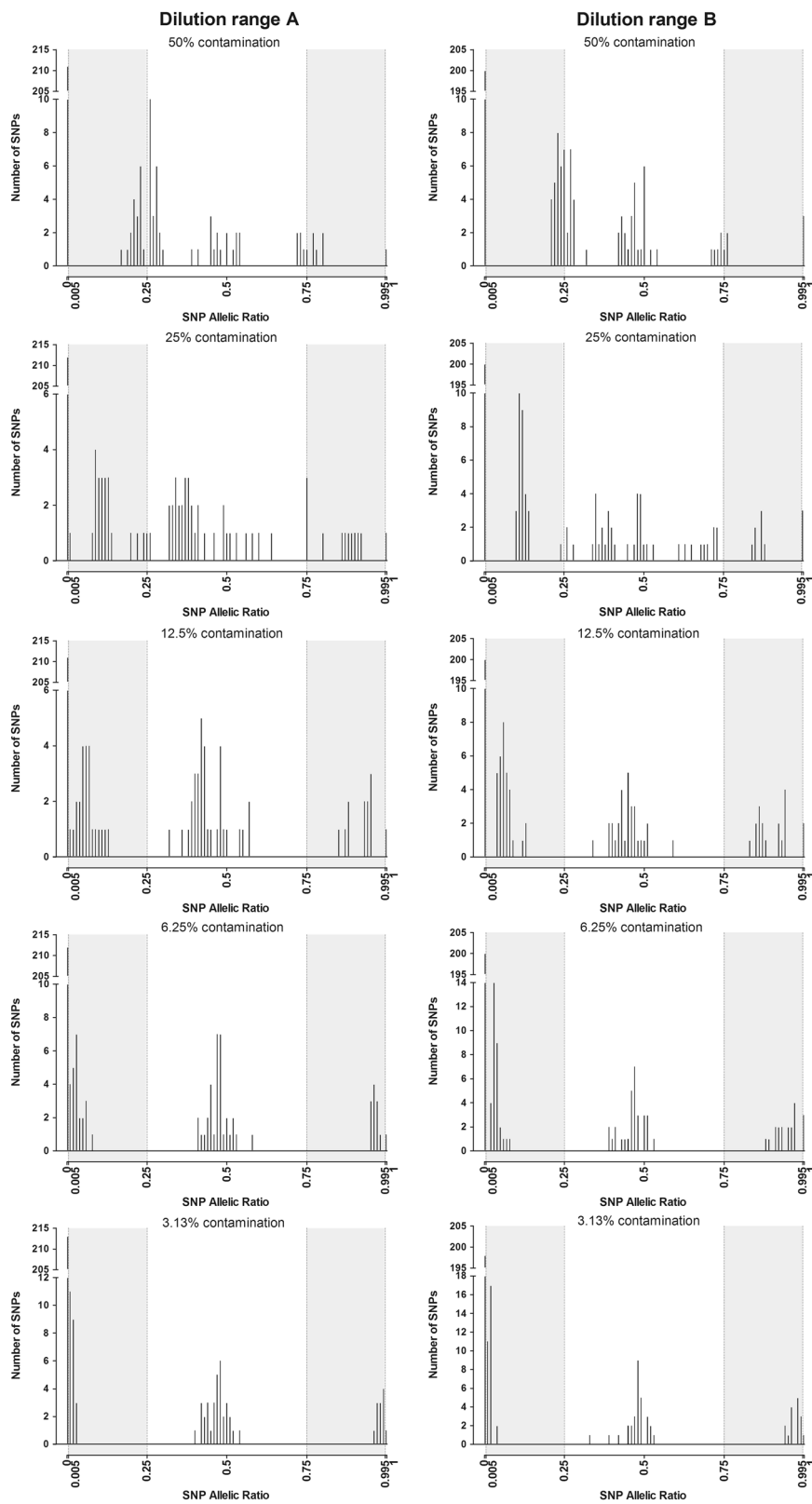
### Optimization step: dilution ranges

In order to test our strategy and optimize test parameters, two dilution ranges were prepared (dilution A and dilution B) of two DNA samples (A and B) with two other DNA samples (Contaminant\_1 and Contaminant\_2, respectively). These dilutions created 10 samples with contamination levels of 50%, 25%, 12.5%, 6.3% and 3.1%, respectively (Fig. 2). These contaminated samples were processed with the contaminant samples and 20 other samples to create a 32-well test sample plate (Supplementary Figure 3). This set was sequenced with another 32-well diagnostic sample plate processed separately to mimic routine diagnosis.

The results were used to define the SNPs to be included for contamination analysis, i.e. satisfying the quality criteria defined above and the WCS calculation.

The 10 contaminated samples were detected by the WCS contamination screening test with values higher than 1% (4.92–49.37%), whereas WCS contamination of the other 22 samples of the test plate was  $<1\%$  (0.35–0.94%). The contaminants were always correctly identified and accurately quantified, as the expected contamination percentages were obtained (Table 3).

**Fig. 2** Distribution of allelic ratios (AR) values of SNPs in contaminated DNA samples. AR values of the SNPs of the 10 contaminated DNA samples from the two dilution ranges according to the level of contamination (50%, 25%, 12.5%, 6.3%, 3.1%)



**Table 3** Detection, identification and quantification of contamination on the optimization set

Dilutions	Expected contamination percentage	WCS contamination percentage	Refined contamination percentage	Contaminant	<i>p</i> -Value
A1	50%	46.46%	45.49%	Contaminant_1	3.12E-12
A2	25%	34.29%	23.21%	Contaminant_1	4.17E-18
A3	12.50%	20.87%	11.55%	Contaminant_1	2.00E-18
A4	6.25%	11.65%	5.77%	Contaminant_1	1.32E-18
A5	3.13%	5.03%	2.98%	Contaminant_1	3.57E-20
B1	50%	49.37%	48.43%	Contaminant_2	1.05E-05
B2	25%	27.42%	24.60%	Contaminant_2	9.30E-20
B3	12.50%	23.68%	12.59%	Contaminant_2	1.56E-22
B4	6.25%	10.95%	7.30%	Contaminant_2	3.57E-21
B5	3.13%	4.92%	3.36%	Contaminant_2	1.83E-20

WCS worst-case scenario

## Routine analysis

This strategy was used for 2227 consecutive constitutional DNA samples sequenced in 36 runs in the context of routine diagnostic genetic testing. Five samples were excluded due to lack of coverage.

Among the 2222 diagnostic constitutional DNA samples tested, 2031 (91.4%) had a WCS less than 1%, excluding contamination.

Thirty-three of the 191 samples with a positive screening test comprised a contaminant identified in the run. These contaminations had WCS contamination estimates of 1.8–42.8% and real quantification based on their contaminants of 1.3% to 29.2% (Table 4). The site of the contaminant of these 33 contaminated samples was located in an adjacent lateral well for 29 cases (87.9%), which was significantly higher than expected (1.7 cases expected (5.2%)) if the identified contaminant was randomly assigned among the other samples of the run (Fisher's exact test;  $p < 10^{-11}$ ) (Table 4, Fig. 2). The other four contaminated samples involved two DNA of the same run, but not in adjacent wells.

Among the 158 other samples with a positive screening test, 154 samples had a low WCS contamination below 4%, and four had a very high WCS contamination (38.1%, 43.6%, 49.6% and 23.2%). Contamination of the 154 samples with low WCS could not be confirmed; most probably because this low level prevented unambiguous contaminant identification, but high background noise could remain a plausible option (see Discussion). In contrast, background noise could not explain the high WCS contamination observed for the last four samples (Table 4). For two of these four samples, two different samples were identified by the same barcode: two samples were identified by barcode 65 in run No. 18 and 2 samples were identified by barcode 75 in run No. 19 (WCS of 38.1% and 43.6%, respectively) (see barcode correspondence in Supplementary Figure 3).

The third sample (WCS of 49.6%) was comprised of a mixture of two DNA samples that were supposed to be distributed to two successive wells but were actually distributed into the same well. Lastly, the fourth sample with high WCS (23.2%) corresponded to DNA extracted from the saliva of an allogeneic bone marrow transplant recipient. Contamination therefore reflected the mixture of lymphocyte DNA from the donor and DNA from the cells of the patient's mouth.

Overall, 36 (1.6%) of the 2222 diagnostic constitutional DNA samples analyzed in this routine diagnostic setting were contaminated during the presequencing steps.

## Discussion

Identification of contamination in NGS analysis is important to avoid erroneous diagnostic results, especially when mosaicism is suspected. In this study, we present an easy method to detect contamination in routine NGS constitutional genetic analysis. The screening test with quantification of the WCS percentage of contamination identified possibly contaminated samples above a defined cutoff. The contaminant was then identified to confirm and precisely quantify contamination. Interestingly, this method can be used for any constitutional NGS workflow and can be customized according to the user's needs.

## SNP selection

SNP selection is of utmost importance for successful implementation of this method. In order to avoid false-positives, poor quality SNPs must be excluded, as they frequently give unexpected ARs. It is the user's responsibility to define a poor quality SNP for the panel, bearing in mind the consequences in terms of specificity.

**Table 4** Summary of contaminated diagnostic samples

Run number	Barcode	WCS contamination percentage	Contamination percentage by the contaminant	Contaminant	p-Value	Comments
4	BC_26	23.2%				Allograft
5	BC_15	5.5%	4.6%	BC_7	8.85E-18	Contamination by an adjacent lateral well
5	BC_5	6.0%	3.6%	BC_6	4.31E-07	Contamination by an adjacent lateral well
9	BC_24	3.9%	2.9%	BC_64	2.43E-17	Contamination by an adjacent lateral well
9	BC_23	4.1%	2.8%	BC_31	2.62E-12	Contamination by an adjacent lateral well
9	BC_4	4.7%	2.3%	BC_5	1.63E-16	Contamination by an adjacent lateral well
9	BC_7	3.8%	2.3%	BC_6	1.30E-09	Contamination by an adjacent lateral well
9	BC_28	3.5%	1.6%	BC_29	0.004	Contamination by an adjacent lateral well
11	BC_90	6.3%	4.3%	BC_89	4.54E-17	Contamination by an adjacent lateral well
11	BC_60	4.0%	2.6%	BC_52	1.84E-17	Contamination by an adjacent lateral well
11	BC_75	49.6%				DNA from 2 patients in the same well
13	BC_70	9.40%	5.70%	BC_71	4.53E-06	Contamination by an adjacent lateral well
13	BC_71	5.00%	3.00%	BC_72	5.95E-08	Contamination by an adjacent lateral well
16	BC_67	13.50%	7.90%	BC_66	4.45E-17	Contamination by an adjacent lateral well
18	BC_65	38.10%				DNA from 2 patients with the same barcode
19	BC_75	43.6%				DNA from 2 patients with the same barcode
21	BC_47	7.6%	5.1%	BC_46	9.26E-15	Contamination by an adjacent lateral well
23	BC_42	3.3%	2.2%	BC_41	5.46E-06	Contamination by an adjacent lateral well
25	BC_78	39.3%	17.4%	BC_77	1.03E-15	Contamination by an adjacent lateral well
25	BC_89	18.0%	9.6%	BC_81	1.74E-13	Contamination by an adjacent lateral well
25	BC_86	5.5%	3.0%	BC_85	1.32E-13	Contamination by an adjacent lateral well
27	BC_11	13.1%	10.0%	BC_3	7.60E-17	Contamination by an adjacent lateral well
27	BC_9	15.0%	7.7%	BC_1	1.34E-12	Contamination by an adjacent lateral well
27	BC_61	3.9%	1.8%	BC_60	3.88E-07	Contamination by an adjacent lateral well
27	BC_2	1.8%	1.3%	BC_3	0.005	Contamination by an adjacent lateral well
29	BC_87	42.8%	25.4%	BC_79	8.45E-16	Contamination by an adjacent lateral well
29	BC_95	9.2%	6.6%	BC_79	1.57E-08	Contamination by non-adjacent well
29	BC_80	3.0%	2.3%	BC_79	0.007	Contamination by an adjacent lateral well
29	BC_76	2.6%	1.3%	BC_77	0.010	Contamination by an adjacent lateral well
33	BC_16	36.7%	29.2%	BC_8	2.29E-07	Contamination by an adjacent lateral well
34	BC_27	20.0%	18.4%	BC_18	3.59E-18	Contamination by an adjacent diagonally well
34	BC_47	8.7%	4.2%	BC_48	7.82E-14	Contamination by an adjacent lateral well
35	BC_91	7.3%	5.6%	BC_68	2.09E-14	Contamination by non-adjacent well
35	BC_71	8.6%	3.5%	BC_72	1.24E-11	Contamination by an adjacent lateral well
35	BC_92	4.7%	3.2%	BC_69	2.07E-16	Contamination by non-adjacent well
36	BC_60	21.8%	10.7%	BC_52	3.18E-15	Contamination by an adjacent lateral well
36	BC_49	6.4%	5.0%	BC_50	1.62E-14	Contamination by an adjacent lateral well

The number of SNPs required to ensure satisfactory sensitivity depends on their allele frequency (AF) in the population. A SNP with an AF of 50% would be the most informative for the detection of contamination. We recommend including at least 30 SNPs, in linkage equilibrium, with an AF between 30% and 70%, ensuring a 99% probability of having at least 5 informative SNPs. Rare SNPs and SNPs in linkage disequilibrium included in the panel design should obviously be taken into account to consolidate predicted contaminations. For whole exome or genome sequencing, with very high number of SNPs, analysis can be restrained on most covered positions by adjusting minimal depth of coverage.

### WCS screening test

Quantification of WCS contamination was performed on all samples of a run. The WCS test is designed to provide the user with a value higher than the actual contamination value (hence the name “worst case scenario”). The actual contamination value is calculated after identification of the contaminant (see “quantification and localization” in the Discussion section below). A *p*-value could be calculated to highlight a so called “significantly high” WCS”. However, in the event of a highly contaminated run, high WCSs would not be significantly different from one another with a *A* value close to 1, leading the user to miss the contamination.

WCS calculation has two main advantages: firstly, it constitutes a rapid screening test with a customizable contamination cutoff; secondly, this screening test remains effective even in the absence of contaminant in the plate. These two aspects will be discussed successively.

A 1% cutoff was used in our experiments to demonstrate the performance of the method. However, in clinical practice, a 10% cutoff could be more compatible with the sensitivity of Sanger sequencing, as contamination less than 10% would not be detected by Sanger sequencing [8]. In addition, index hopping in pooled libraries has been observed up to 6% [4] across various methods and Illumina sequencers. Then, those “index-contaminations” might lead to contamination predictions with or without contaminant at low rate. In any case, the user can select any critical cutoff depending on the objectives of the study.

This WCS calculation enables the contamination detection even when the contaminant sample is not present in the run to confirm it. For example, the WCS screening test allowed the detection of the mixture of two different DNAs in the same well before barcoding and barcoding of two different DNAs with the same barcode. A particularly interesting example was a sample from a female patient with a high WCS percentage of contamination (23.22%) with no contaminant identified in the run and no experimental explanation. Surprisingly, an X-linked gene

included in the panel showed that this DNA sample more closely resembled a male sample than a female sample. This sample corresponded to that of a woman who underwent allogeneic bone marrow transplant for acute lymphocytic leukaemia 12 years previously. Our method suggested that the donor was likely a man and that the tested DNA sample, extracted from saliva, was composed of a maximum of 23.22% of patient DNA and a minimum of 76.78% of donor DNA. This result was not surprising, as saliva is known to contain lymphocytes [9]. This finding highlights the importance of providing laboratories with relevant clinical information to ensure reliable interpretation of the results.

In 154 samples, a WCS between 1% and 4% failed to identify any contaminant, which could be explained by high background noise and/or too low contaminant level, that could result from index hopping [4], or absence of contaminant in the plate. An AR background noise cutoff of 0.005 was used, so the theoretical lower limit of detection of contamination was 1%. However, in practice, because of the normal distribution of the heterozygous SNP AR, a low level of contamination is associated with a great number of SNPs with AR in the background noise, preventing confirmation of low levels of contamination.

Background noise determines the contamination detection cutoff, which is why detection depends on sequencing protocols used and must be adapted to the user’s specific needs.

### Localization and quantification of contaminants

After localizing the contaminant, a refined contamination percentage was calculated, taking into account the genotype of the variants of the contaminant and the contaminated sample.

Thirty-three of the 2222 diagnostic constitutional DNA samples tested were contaminated by another sample on the plate and 6 (0.3%) of them presented clinically relevant contamination  $\geq 10\%$  and 27 (1.2%) presented contamination  $< 10\%$ , deemed to be negligible for constitutional genetic analysis.

As expected, most of the contaminants were located in the adjacent lateral well (87.8%), which is highly suggestive of projection of droplets during library preparation prior to the barcoding step, as many library protocols, including SureSelectQXT, comprise washes that require up-and-down pipetting in the wells of the plates, which sometimes generates droplets that fall onto the plate or into an adjacent well. An understanding of this most common mechanism of contamination is of utmost importance to ensure increased vigilance and optimized practices by the user. Automation of library preparation might reduce contamination but in any case optimization of best practices can be monitored by measuring contamination rates over time, which should theoretically decrease.



Other methods of detection of contamination have already been described in the literature. However, most of these methods were developed for tumor analyses and can require supplemental SNP array data, e.g. the ContEst tool [10]. Alternatively, the Conpair tool [11] does not need SNP array data, but is based on tumor-constitutional pair analyses. Interestingly, Sehn et al. described a haplotype-based tool for tumor analyses [12], which should theoretically also be suitable for constitutional analyses, but with several design constraints, as loci with SNPs in low-linkage disequilibrium are needed to ensure reliable contamination detection, as this tool was developed for the frequent rearrangements found in tumors. Our method is simpler with no such constraints, as rearrangements are not frequently found in constitutional analysis. Lastly, the method described by Jun et al. and Flickinger et al. [13, 14], also described contamination detection with sequence reads, but based on larger amounts of data (at least 1000 SNPs) provided by massive sequencing such as genome-wide analysis or whole-exome sequencing. Even if such analysis are starting to be more routinely performed, genes panels such as hereditary cancer panels are still widely used for routine diagnostic.

As the proposed method is based on a standard gene panel commonly used in routine constitutional genetic testing, it constitutes a powerful and easy-to-use quality tool with educational benefits, as it also highlights the weaknesses of the process, which is why we believe it should be implemented in diagnostic pipelines as part of the accreditation process. Importantly, it can be used in any NGS workflow, from gene panel to genome-wide analyses.

**Acknowledgements** This work was supported by grants from the ANR-10-EQPX-03 from the Agence Nationale de la Recherche (Investissements d’Avenir).

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok, et al. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18:308.
2. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L, et al. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*. 2011;12:451.
3. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA*. 2011;108:9530–5.
4. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*. 2018;19:332.
5. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*. 2011;17:10–2.
6. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
7. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:1–33.
8. Davidson CJ, Zeringer E, Champion KJ, Gauthier MP, Wang F, Boonyaratanakornkit J, et al. Improving the limit of detection for Sanger sequencing: a comparison of methodologies for KRAS variant detection. *Biotechniques*. 2012;53:182–8.
9. Taniguchi S, Maekawa N, Yashiro N, Hamada T. Detection of human T-cell lymphotropic virus type-1 proviral DNA in the saliva of an adult T-cell leukaemia/lymphoma patient using the polymerase chain reaction. *Br J Dermatol*. 1993;129:637–41.
10. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G, et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics*. 2011; 27:2601–2.
11. Bergmann EA, Chen BJ, Arora K, Vacic V, Zody MC. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics*. 2016;32:3196–8.
12. Sehn JK, Spencer DH, Pfeifer JD, Bredemeyer AJ, Cottrell CE, Abel HJ, et al. Occult specimen contamination in routine clinical next-generation sequencing testing. *Am J Clin Pathol*. 2015; 144:667–74.
13. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*. 2012;91:839–48.
14. Flickinger M, Jun G, Abecasis GR, Boehnke M, Kang HM. Correcting for sample contamination in genotype calling of DNA sequence data. *Am J Hum Genet*. 2015;97:284–90.