Detection of Intraspecies Contamination as a Quality Control Tool in The OncoDEEP® Kit Comprehensive Genomic Panel

Raphaël R. Léonard[1], Jessie Hong[1], Sébastien Sauvage[1], Karim Ashour Garrido[1], Jean-François Vanbellinghen[1], Isabella Giovannoni[2], Sabina Barresi[2] , Gabriele Capone[3]and Maxime Lienard[1].

[1] OncoDNA SA, Rue Louis Breguet 1, Gosselies, Belgium; [2] Pathology Unit, IRCCS Bambino Gesù Children's Hospital, Rome, Italy; [3] Devyser, Italy.

Introduction : Next-Generation sequencing (NGS) emerged as a revolutionary technology in genomics research aiding precision medicine for almost two decades, but it has its challenges. One of which is the intraspecies contamination which could affect assay's (1) sensitivity due to the presence of contaminating DNA decreasing the observed allele fraction of variants in the actual specimen (2) accuracy due to the presence of pathogenic variants in contaminating DNA leading to false-positive result.

OncoDEEP® Kit is a pan-cancer NGS assay developed with oncology expertise and supported by BioIT solutions. It is based on a comprehensive panel of 638 genes allowing the detection of single nucleotide variant, insertions/deletions, loss of heterozygosity, copy number variation. Additionally, genomic signatures, such as 1p/19q codeletion, microsatellite instability, tumor mutational burden and homologous recombination deficiency can also be assessed.

To improve the kit quality control, a check for intraspecies contamination was needed. We selected the methodology presented by Li et al. 2021 in "Contamination Assessment for Cancer Next-Generation Sequencing" due to its ease of implementation, speed and feasibility to scale up for Whole Genome/Exome Sequencing. It is based on an α/β ratio where α is the number of dbSNP variants with 100% of variant allele fraction different from the genomic reference and β is the number of dbSNP variants different from the genomic reference.

The OncoDEEP kit allowed the full characterization, from the DNA extraction to the final report, of the samples in less than 5 working days (**Figure 1**). After extraction, libraries were constructed (**3h**), enriched (**hands on time: 4h**) based on Twist Biosciences Technology and sequenced (**20h**) on an Illumina NextSeq 500 or NextSeq 2000, depending on the center. Finally, FastQ files were uploaded and analyzed through OncoDNA dedicated BioIT pipeline.

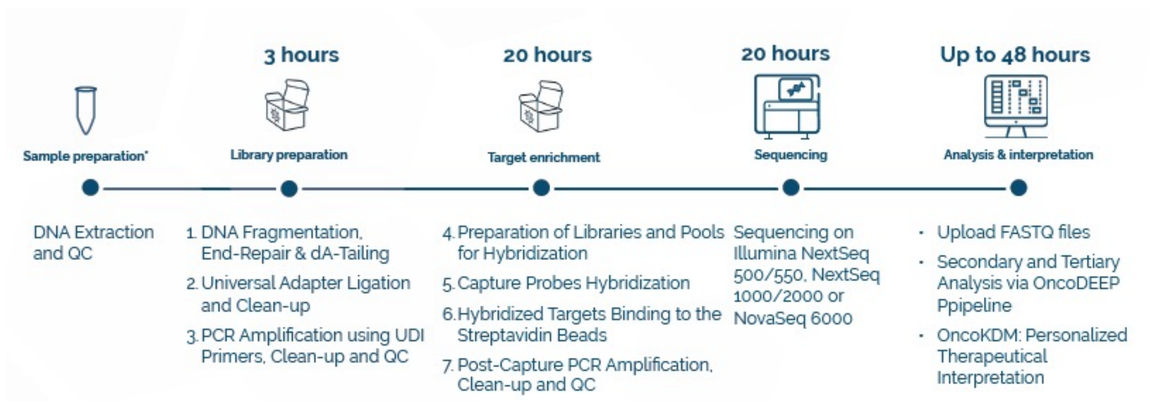

Figure 1: Full OncoDEEP workflow from the wetlab part (DNA extraction to sequencing) to FASTQ files upload and secondary/tertiary analysis.

Material and Methods: Following the method described by Li *et al.* 2021, we artificially created contaminated samples by concatenating two samples' FASTQs, one fully and one in known proportions of reads for several contaminated fractions (1/20, 1/10, 1/8, ¼) and for four pairs of samples.

We retrieved the UCSC dbSNP v155 covered by our kit panel to create two reference files, one for GRCh38/hg38 and the other for GRCh37/hg19. First, we downloaded the dbSNP v155 GRCh38/hg38 for the whole genome. Then, we filtered it using our OncoDEEP Kit BED file with all targets extended by 25 base pairs (BP) on both sides, taking indiscriminately all Single Nucleotide Polymorphisms (SNPs), regardless of the Variant Frequency (VF). Due to the differences in terms of the number of SNPs available in dbSNP v155 for GRCh37/hg19, we used CrossMap (Zhao *et al.* 2013) to lift over the GRCh38/hg38 reference file to GRCh37/hg19 using the Ensembl chain file.

Using the artificially contaminated samples, we computed the α/β ratios for the four contamination fractions for each pair of samples, then created a plot of the mean ratios on the contamination fraction (Fig.2). An exponential curve was fitted on the graph and the resulting formula (R^2=0.9865) was used in our detection module.

Equation 1: Contamination fraction = 46.47 exp(-75.14 α/β)

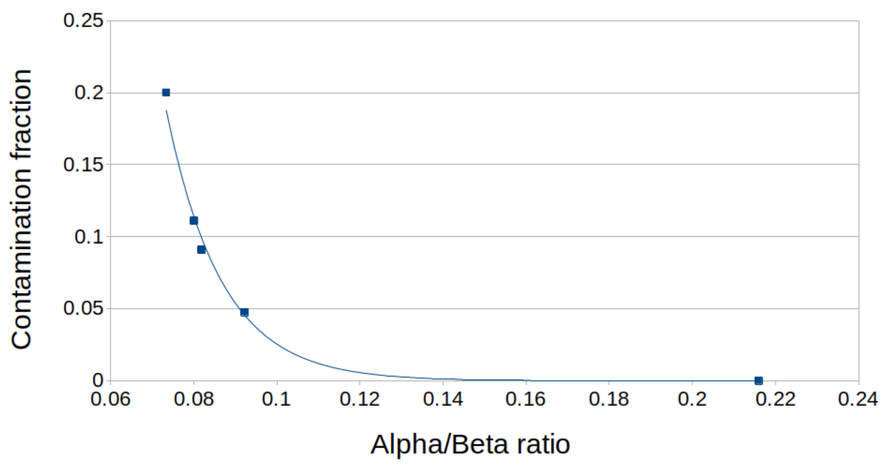

Figure 2 Contamination fraction curve: mean α/β ratio of four pairs of samples against the four contamination fractions and the mean ratio of the uncontaminated samples. An exponential curve is fitted to the scatter plot.

The next step was applying the exponential formula obtained by fitting the data in Fig. 2 to our artificially contaminated samples to assess the efficiency and limitations of the Li *et al.* 2021 method (Table 1). Finally, the sensitivity of the method was assessed on 48 clinical samples.

Table 1: Results obtained using equation 1 compared to the expected results for each artificially contaminated sample.

ID	Contamination	Result
S1	0	3.52E-07
S1_S2_20	1/20	0.015983611137258
S1_S2_10	1/10	0.045885662203302
S1_S2_8	1/8	0.055375718087316
S1_S2_4	1/4	0.081851929678063
S2	0	8.54E-06
S2_S3_20	1/20	0.052175798848453
S2_S3_10	1/10	0.072236838756698
S2_S3_8	1/8	0.072041054177688
S2_S3_4	1/4	0.083656775347825
S3	0	7.80E-07
S3_S4_20	1/20	0.040871245004403
S3_S4_10	1/10	0.110491109560217
S3_S4_8	1/8	0.136096502528166
S3_S4_4	1/4	0.299717260928125
S4	0	0.000128566639236
S4_S5_20	1/20	0.124977372543735
S4_S5_10	1/10	0.263875380954789
S4_S5_8	1/8	0.307158696768063
S4_S5_4	1/4	0.60784913528297

Results and discussion: For 48 samples, we used the formula fitted to our kit gene panel (equation 1). Four contaminated samples were used from the technical validation of the kit, six other clinical samples were diluted with a reference sample to assess the sensitivity of the module.

Table 2: Contamination fraction for the 48 clinical samples used to determine the sensitivity of the detection tool.

ID	Result	contaminated	ID	Result	contaminated
C1	2.21E-08	N	C25	1.10232286	Y
C2	2.64E-03	N	C26	7.29E-06	N
C3	5.89E-06	N	C27	7.95314694	Y
C4	3.64E-08	N	C28	3.47011522	Y
C5	7.47E-07	N	C29	0.74599986	Y
C6	2.10E-06	N	C30	3.47813702	Y
C7	4.21E-07	N	C31	0.44277133	Y
C8	1.85E-05	N	C32	0.0002189	N
C9	5.71E-08	N	C33	0.00055034	N
C10	2.89E-06	N	C34	0.00171114	N
C11	3.02E-04	N	C35	0.00011793	N
C12	3.27E-08	N	C36	5.58E-06	N
C13	7.46E-06	N	C37	3.47E-06	N
C14	7.21E-08	N	C38	7.50E-05	N
C15	2.39E-07	N	C39	0.00204377	N
C16	8.49E-06	N	C40	7.68E-05	N
C17	6.43E-06	N	C41	0.00059225	N
C18	2.80E-05	N	C42	2.45E-05	N
C19	1.74E-07	N	C43	0.00033967	N
C20	4.60E-05	N	C44	0.0001799	N
C21	0.1402149	Y	C45	2.40E-05	N
C22	0.1642037	Y	C46	2.70E-05	N
C23	0.0507646	Y	C47	7.69E-05	N
C24	1.8258629	Y	C48	1.87E-08	N

Contamination detection module steps:
1) Open BAM file
2) Find SNPs VF and count α and β
3) Use equation 1
4) Print Results

Based on the results of Table 1 and 2, we realized that, while capable of detecting contaminations, the method we used was too imprecise to determine an accurate contamination fraction. This is due to the exponential nature of the formula, which makes the estimation extremely sensitive and even more rapidly progressing. The contamination fraction is supposed to vary between 0 and 1 but in several cases, we had results largely over 1 (100% contamination), our worst sample in our clinical list being C27 (Table 2) with a contamination fraction of 7.95. Therefore, instead of using the method as a quantitative estimation of the contamination fraction, we decided to use it as a qualitative method and, based on the results showed in Table 1, we chose a threshold of 4.5% which allows to cover most cases of in-silico contaminated samples and all contaminated clinical samples.

Conclusion: The tool using the Li et al. 2021 method works properly and gives an overview of the contamination level of samples. Nevertheless, this method is only a qualitative tool due to the exponential nature of the formula. The threshold set at 4.5% provided a sensitivity of 100% on the clinical sample's cohort and is compatible with our variants minimum reporting frequency of 5% for the OncoDEEP kit. The next step would be to verify if it could be used with our liquid biopsy analysis pipeline using Unique Molecular Identifiers which allows a lower minimum reporting frequency.