Performance evaluation of three DNA sample tracking tools in a whole exome sequencing workflow

Evaluation of the Human Sample ID Kit in clinical practice, in collaboration with Ghent University Hospital

Introduction

Whole-exome (WES) and whole-genome sequencing (WGS) have become routine practice in clinical genetic laboratories. However, the complex workflows, custody transfers and large datasets impose challenges on data integrity that range from the initial sample collection to the downstream data analysis. It is estimated that up to 3% of all samples may be compromised by provenance errors, raising serious concerns about the integrity and reliability of massively parallel sequencing (MPS) data[1]. Both in the clinic and the research laboratory, identity mix-ups can have detrimental consequences. A wrong diagnosis resulting in an incorrect or delayed treatment can cause severe harm to the patient, while erroneous data in a research context can impair discovery of new causal variants by yielding misleading variant candidates[2,3]. As sample mix-up errors are difficult to detect or to prevent, implementation of appropriate measures are critical for the unambiguous re-identification of samples throughout all stages of the MPS workflow[4,5]. An independent post hoc verification that the sequence results have been correctly assigned to each patient is therefore highly desirable. By using single nucleotide polymorphisms (SNPs), a unique fingerprint can be determined for each individual sample, ensuring sample mislabeling and handling errors are no longer part of the workflow[1,2].

Over the last years, several SNP based sample identification panels specifically designed for MPS have been commercialized, including our own Human Sample ID Kit (www.pxlence.com). In this study, an external comparison of the performance of our Human Sample ID Kit with two other commercially available SNP sample tracking methods is provided.

Methods

Patient samples

In total 46 different genomic DNA (gDNA) samples were used in this study, isolated from either blood (40 samples), FFPE tissue (3 samples) or fresh frozen tissue (1 sample). For one donor, three 3 biological gDNA replicates were included.

SNP sample tracking

The Human Sample ID Kit (#PXL-SID-001, pxlence) was compared with two other commercially available SNP sample tracking kits, designated kit A and kit B throughout the text. All three kits were used as recommended by the manufacturer using a DNA input of 20 ng per reaction. An experienced laboratory technician from the Center for Medical Genetics, Ghent University Hospital, Belgium, performed all the lab work.

Sequencing

The DNA sample tracking libraries of all three kits were pooled and spiked in a WES workflow and simultaneously sequenced on the Illumina NovaSeq system with a paired-end read length of 2 × 150 bp.

Data Analysis

For assessing on-target specificity and coverage uniformity, reads were first aligned to the human reference genome by means of the Burrows-Wheeler aligner (BWA v0.7.17)[6]. Mosdepth (v0.2.3) and total sample read-depth were used to calculate per-nucleotide normalized coverage to determine coverage uniformity of the various SNPs per patient[7]. To assess specificity, only regions having a non-normalized minimum per-nucleotide coverage of 25,

and overlapping with a SNP included in the corresponding kit, were considered to be on-target. For analyzing genotype similarities between WES and sample tracking data, individual libraries were downsampled to 100,000 reads. Genotype matches through logarithm of the odds (LOD) scores were used for comparison of genetic fingerprints between samples using the CrosscheckFingerprints tool from the Picard software package (v2.1.1)[8]. In this analysis, a near zero LOD score indicates an inconclusive comparison, while a sample match or mismatch are given a positive or negative LOD score, respectively.

Results

Coverage uniformity

Sequencing coverage uniformity is a measure of the amplification efficiency of each individual SNP within the multiplex PCR reactions performed as part of each of the corresponding method's workflow. Perfect equimolar assay coverage means minimal sequencing capacity is required to attain a minimal coverage per assay (for example, in this setting each assay would be covered exactly 30x times), resulting in an optimal cost-efficiency. Deviation of such a perfect situation results in increased sequencing capacity – and sequencing cost – required to achieve similar results (e.g. if a method includes a sub-performing assay, additional sequencing capacity would be needed to bring that assay up to 30x coverage). Coverage uniformity is typically reported as the percentage of assays having a coverage above 0.2 times the median coverage in a specific sample. However, since this measure ignores highly-efficient assays with excessive coverage – resulting in decreased coverage uniformity – the percentage of assay falling within the range of 2-fold around the median sample coverage (calculated across all samples) will also be reported here. Gender markers were omitted from this analysis. Results indicate that kit A scores best on coverage uniformity, with 89.55% of the datapoints within 2-fold of the median (Figure 1). It is followed by our Human Sample ID Kit and kit B, with 84.14% and 81.75% of the datapoints within a 2-fold range around the median, respectively (Figure 1). These findings are confirmed when calculating the per-sample standard deviation (SD) of the normalized coverage: 0.0029 for kit A,

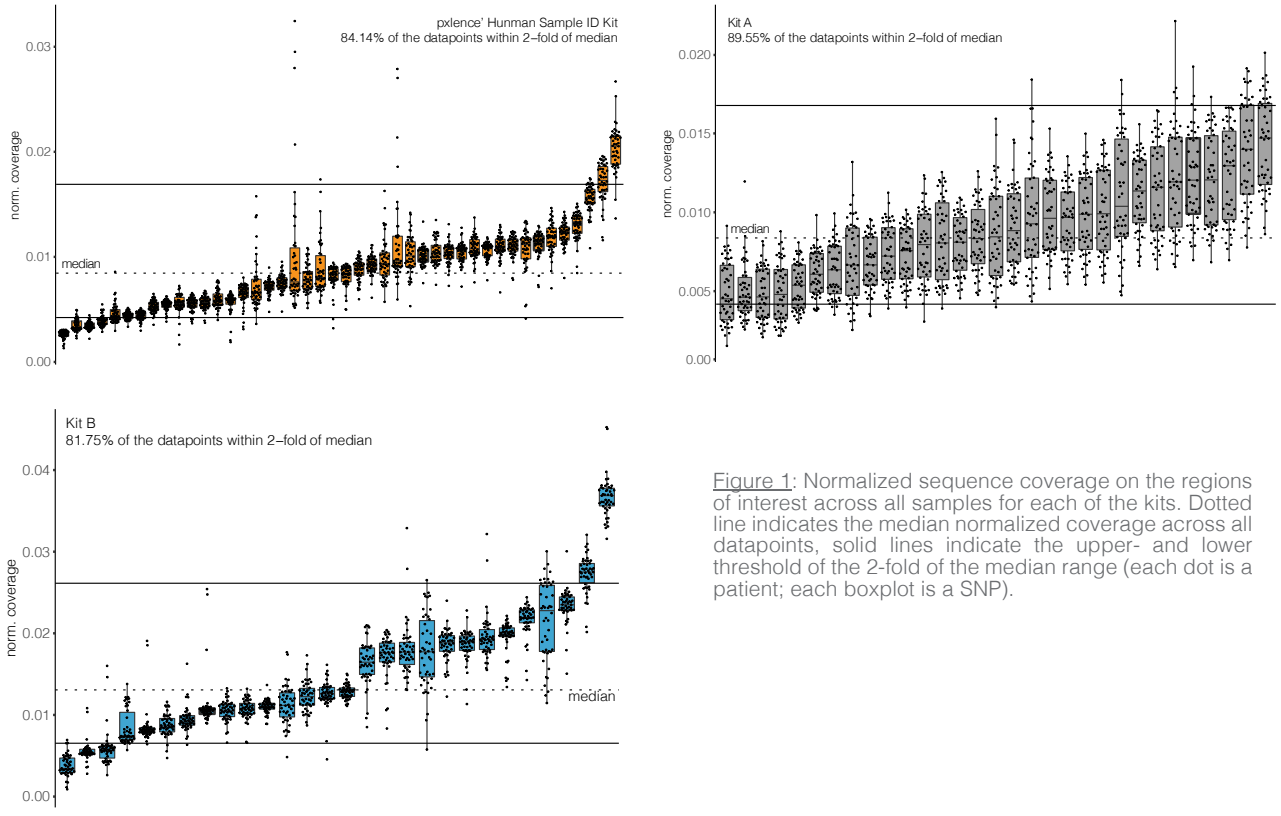

Figure 1: Normalized sequence coverage on the regions of interest across all samples for each of the kits. Dotted line indicates the median normalized coverage across all datapoints, solid lines indicate the upper- and lower threshold of the 2-fold of the median range (each dot is a patient; each boxplot is a SNP).

0.0040 for the Human Sample ID Kit, and 0.0076 for kit B (Figure 2A). Remarkably, kit A shows a much larger cross-sample intra-assay variability compared to the other two methods tested.

On-target specificity

Anlogous to coverage uniformity, on-target specificity has an impact on the sequencing capacity required per sample, and thus per-sample sequencing cost. In an ideal scenario, all reads produced by a method will generate useful data for the targets of interest. A lower on-target specificity – and thus higher level of non-specificity – will lead to lower coverage for the SNPs thus a need for higher per-sample sequencing capacity to achieve the same minimal per-assay coverage. This analysis shows that kit B performs best in the context of on-target specificity, with a median of 4.43% of the reads aligning off-target (Figure 2B). Our Human Sample ID Kit scores also scores very well with approximately 9.90% off-target reads. In this context, kit A performs worst with a median of 22.82% and a significant portion of the samples showing off-target percentages above 30 (n = 11) and even up to 40% (n = 4). A much smaller range in per-sample off-target percentages could be observed for both kit B and the Human Sample ID Kit. This means that, in comparison to kit B and our Human Sample ID Kit, kit A will need up to 30-35% more per-sample sequencing capacity for a significant portion of the samples, resulting in an overall higher per-sample sequencing cost.

Genotyping and sample discrimination

LOD scores were used for comparing the genotypes obtained by the three SNP sample tracking panels and the data from the WES (Figure 3A). Only for the Human sample ID Kit, unambiguous sample discrimination and identification was obtained for all samples. For kit A, one genotyping sample was not identified correctly, showing as a mismatch with the correct WES data, being inconclusive with another unrelated sample. With kit B, only 32 of the 46 samples were matched with its corresponding WES data and - more importantly - only 20 samples showed a correct match while not having an inconclusive result with another sample. For none of the three kits a match was found with an unrelated sample. The good performance of the Human Sample ID Kit – and to a lesser extend kit A – is further substantiated by looking at the individual LOD scores of the matching samples and mismatch samples (Figure 3B).

For a good discriminatory performance, LOD scores should be as decisive as possible, meaning LOD scores of matching samples should be as high as possible above zero and LOD scores of mismatch samples should be as low as possible below zero. As expected from the genotyping, kit B shows the lowest discriminatory performance, with average LOD scores of ±6 eing just above the inconclusi › threshold. In compa ison, LOD scores of the Huma Sample ID Kit and kit A showed a much better discrimination with

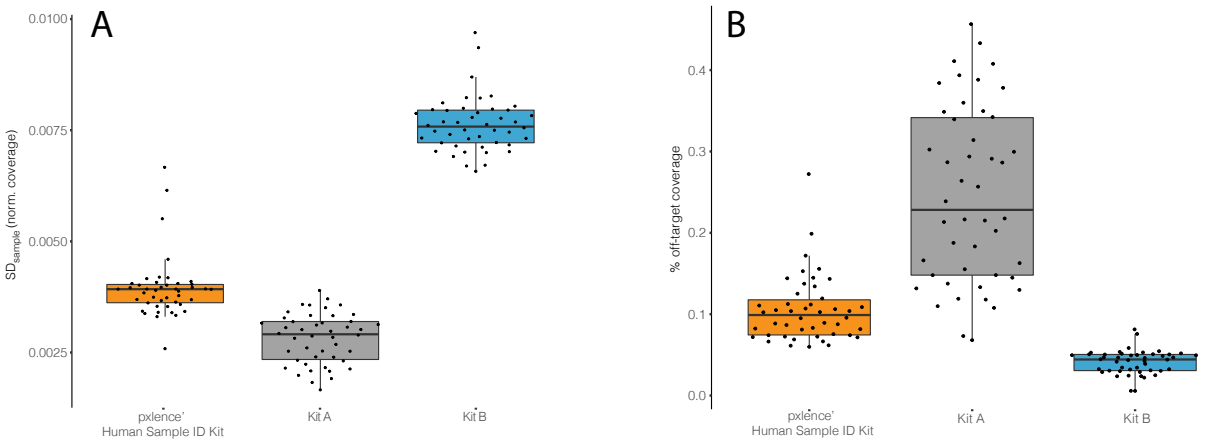

Figure 2: (A) Standard deviation of the normalized coverage per sample across all regions of interest for each of the kits, (B) percentage of off-target coverage per sample for each of the kits (lower is better; each dot is a patient).

LOD scores of the matching samples of respectively ±19 and ±13, and LOD scores of mismatch samples of ±-48 and ±-30, respectively. The excellent performance of the Human Sample ID Kit in discriminating samples can be partly explained by the larger number of SNP assays in this kit. A larger number of assays has the additional benefit that – in case of sub-optimal sequencing with lower coverage values – sufficient high-quality markers remain for robust sample identification. In contrast, kits with lower SNP numbers can suffer from lack in discrimination power when combined with sub-optimal sequencing results (due to insufficient high-quality marker remaining).

Conclusion

In a clinical setting, the three tested SNP sample tracking methods displayed significant differences in their sample identification and genotyping performance. Overall, kit B showed to be unreliable with many samples that showed undecisive correlation although on-target specificity was observed t be highest in this kit. The Kit A performed best on coverage uniformity but showed poor on-target rates. From the three kits, the Human Sample ID Kit excelled in sample identification and discrimination. The high sample discrimination performance allows for a high-confident and robust genotyping, assuring correct sample identification and avoids samples to be reanalyzed. Combined with an above-average on-target specificity and coverage uniformity, our Human Sample ID Kit shows the overall best per-sample cost-efficiency.

References

1. Sehn, J. K. et al. Occult specimen contamination in routine clinical next-generation sequencing testing. Am. J. Clin. Pathol. 144, 667–674 (2015).

2. Pengelly, R. J. et al. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med. 5, 89 (2013).

3. Pedersen, B. S. & Quinlan, A. R. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am. J. Hum. Genet. 100, 406–413 (2017).

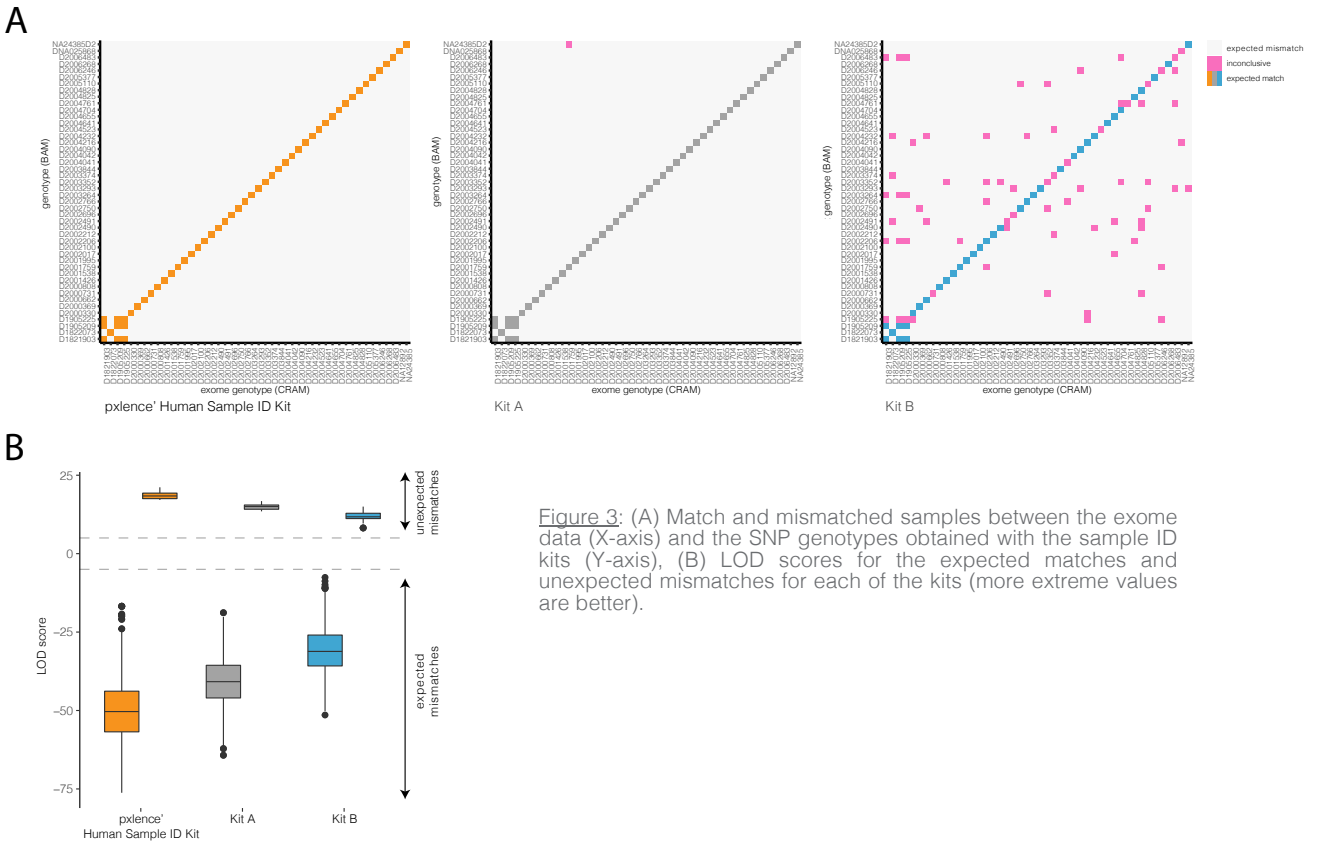

Figure 3: (A) Match and mismatched samples between the exome data (X-axis) and the SNP genotypes obtained with the sample ID kits (Y-axis), (B) LOD scores for the expected matches and unexpected mismatches for each of the kits (more extreme values are better).

4. Matthijs, G. et al. Guidelines for diagnostic next-generation sequencing. Eur. J. Hum. Genet. 24, 2–5 (2016).

5. Moore, R. A. et al. Sample Tracking Using Unique Sequence Controls. J. Mol. Diagnostics 22, 141–146 (2020).

6. Li H. and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics (2010).

7. Pedersen B. and Quinlan A. Mosdepth: quick coverage calculation for genomes ane exomes. Bioinformatics (2018).

8. “Picard Toolkit.” 2019. Broad Institute, GitHub Repository. Broad Institute