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

Gertjan Wils[1,], Céline Helsmoortel[3,**], Pieter-Jan Volders[2], Inge Vereecke[3], Mauro Milazzo[3], Jo Vandesompele[1,2], Frauke Coppieters[1,2], Kim De Leeneer[2,3,Δ], and Steve Lefever [1,*,Δ]**

[1] pxlence bvba, Dendermonde, Belgium
[2] Department of Biomolecular Medicine, Ghent University, Ghent, Belgium
[3] Center for Medical Genetics Ghent, Ghent University Hospital, Ghent, Belgium
[*] Correspondence: steve.lefever@pxlence.com
[**] Shared first authors
[Δ] Shared last authors

Abstract

Next-generation sequencing applications are becoming the building blocks for clinical diagnostics. These experiments require numerous wet- and drylab steps, each one increasing the probability of a sample swap and/ or contamination. Therefore, an identity confirmation at the end of the process is required to ensure the right data is used for each patient. We tested three commercially available, SNP bases sample tracking kits in a diagnostic workflow to evaluate their performance. The coverage uniformity, on-target specificity, sample identification and genotyping performance were determined to measure the reliability and estimate the cost-effectiveness of each kit. Our findings showed that the kit from Swift didn't perform up to standards as only 20 out of the 46 samples were correctly genotyped. The kit provided by Nimagen identified all but one sample and the kit from pxlence unambiguously identified all samples, making it the most reliable and robust kit of this evaluation. The kit from Nimagen showed poor on-target rates, resulting in deeper sequencing needs and higher sequencing costs compared to the other two kits.

Key point
When choosing a sample tracking tool, reliability, robustness and cost-effectiveness have to be considered. Our evaluation showed that the kit from pxlence is the best chose when looking at these parameters in a real-life setting.

Introduction

Whole-exome (WES) and whole-genome sequencing (WGS) have become routine practice in clinical genetic laboratories [1]. 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 [2–4]. 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 [5,6]. 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 [7,8]. An independent post hoc verification that the sequence results have been correctly assigned to each patient is therefore highly desirable.

Already in 2013, the American College of Medical Genetics and Genomics (ACMG) advised to track sample identity throughout the MPS process as part of adequate quality control [9]. The need for sample tracking was also included in more recent guidelines for (diagnostic) MPS issued by for instance the European Society of Human Genetics [7] and the Canadian College of Medical Geneticists (CCMG) [10]. Different methods exist for DNA sample tracking such as spike-in synthetic DNA standards [8,11,12] or single nucleotide polymorphisms (SNP) panels. By genotyping SNPs through an independent analysis, a unique fingerprint can be determined for each individual sample without interfering with the original DNA, ensuring sample mislabeling and handling errors are no longer part of the workflow [2,13–15]. Over the last years, several SNP-based sample identification panels specifically designed for MPS have been commercialized. In this study, a comparison of the performance of three commercially available SNP sample tracking methods is provided.

Materials and Methods

Patient samples

In total 46 different genomic DNA (gDNA) samples were used in this study, isolated from either blood (40 samples, MagCore Genomic DNA Large Volume Whole Blood Kit, MagCore Automated Nucleic Acid Extractor), Formalin-Fixed Paraffin-Embedded (FFPE) tissue (3 samples) and fresh frozen tissue (1 sample, QiaAmp Blood mini kit, QIAcube, QIAgen). Two samples are reference samples (NA24385 and NA12892) from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. For one FFPE donor, three 3 biological gDNA replicates were included.

SNP sample tracking library preparation

The following commercially available SNP sample tracking kits were evaluated: Human Sample ID Kit PXL-SID-001 V1.0 (pxlence, kit A), Human Identification and Sample tracking kit RC-HEST V2.2 (Nimagen, kit B) and Accel-Amplicon Sample_ID Panel CP-UZ6128 V3.0 (Swift Biosciences custom panel, kit C). An overview of SNPs and gender markers (GM) in each kit are given in Table A1, Table A2 and Table A3. All three kits were used as recommended by the manufacturer. A 20 ng/µl dilution was made for each sample. Following the manual, different amount of input was used (kit A 20 ng, kit B 80 ng and kit C 20 ng). Quality control of the resulting library preparations was performed using concentration measurement (Fluoroskan, ThermoFisher, Invitrogen Quant-iT dsDNA Assay Kit, high sensitivity) and automated electrophoresis (TapeStation, Agilent, Agilent High Sensitivity D1000 ScreenTape).

Sequencing

Per kit, library preps were pooled equivolumetrically, followed by bead purification (AMPure XP, Beckman Coulter) and concentration measurement of the final pools using qPCR (Kapa Library Quantification Kit, Roche) or Fluoroskan (ThermoFisher, Invitrogen Quant-iT dsDNA Assay Kit, high sensitivity). Subsequently, the three pools were spiked in a diagnostic WES workflow containing 186 exomes (SureSelectXT Low Input Target Enrichment System, Human All Exon V7 probes, Bravo Automated Liquid Handling Platform, Agilent), and 2 whole genome preparations (NEXTFLEX Rapid XP DNA-seq kit, PerkinElmer). The following ratios were applied for pooling: 0.23% for kits A and C, 0.47% for kit B, 87.23% for the 186 exomes, and 11.74% for the 2 genomes. 1.5 nM of the final pool, including 1% PhiX, was sequenced on an Illumina NovaSeq 6000 system (S4 Reagent Kit, 300 cycles, paired-end sequencing).

Data Analysis

For assessing on-target specificity and coverage uniformity, reads were first aligned to the human reference genome (GRCh38) by means of the Burrows-Wheeler aligner (BWA v0.7.17) [16]. 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 [5]. To assess specificity, only regions having a non-normalized minimum per-nucleotide coverage of 25x 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) [17]. 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.

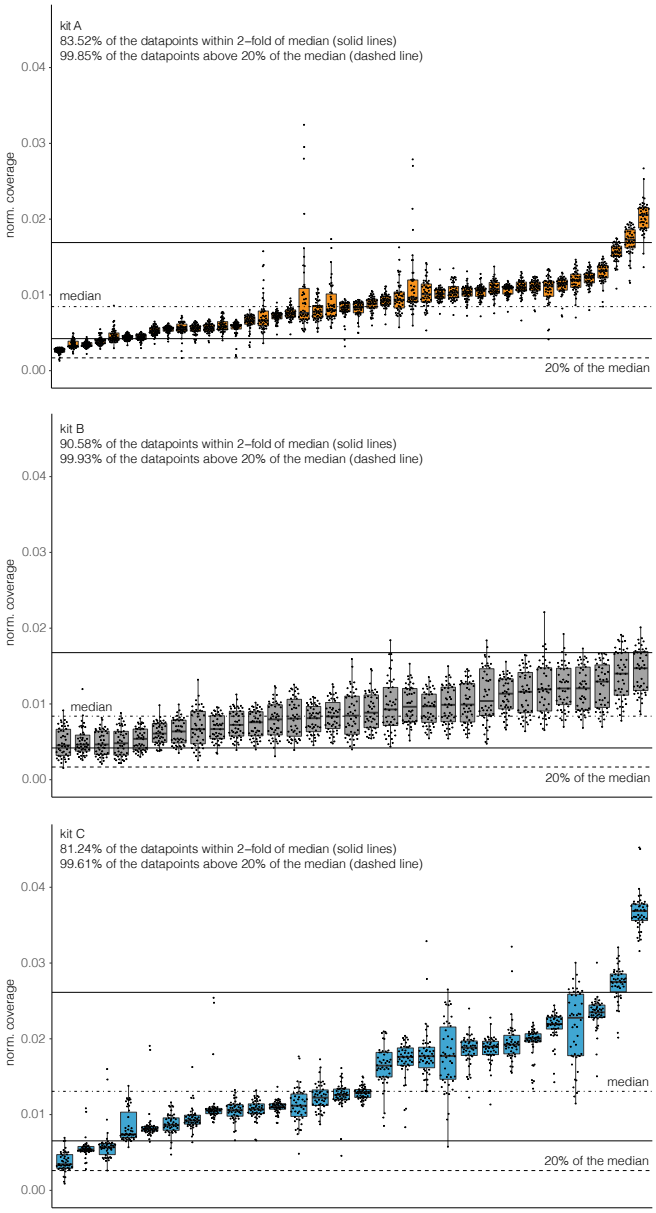

Figure 1. Normalized sequence coverage on the regions of interest across all samples for each of the kits. Dotted-dashed line indicates the median normalized coverage across all datapoints, solid lines indicate the upper- and lower threshold of the 2-fold of the median range, and dotted line indicates the 20% of the median coverage (each dot is a patient; each boxplot is a SNP).

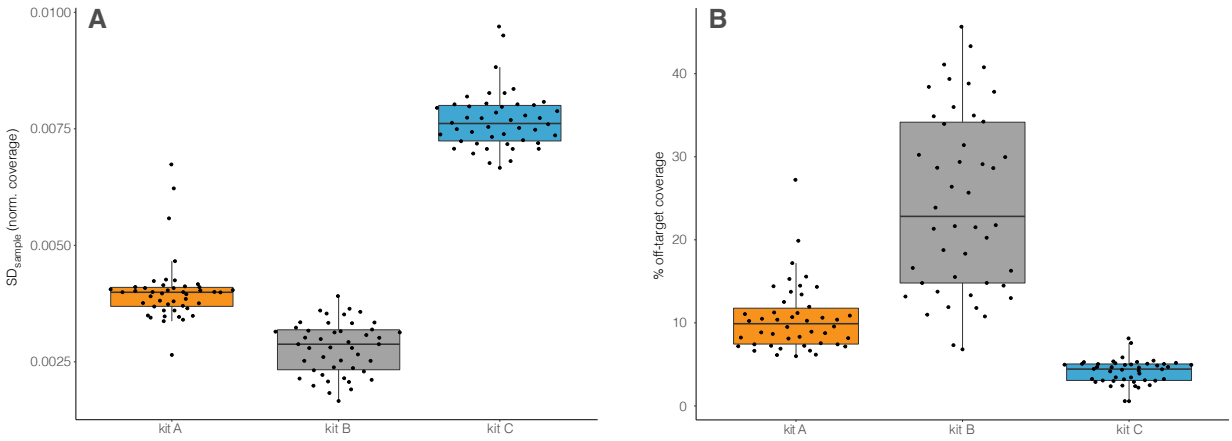

Figure 2. (A) Standard deviation of the normalized coverage per sample across all regions of interest for each of the kits, (B) percentage of off-target coverage per sample for each of the kits (lower is better; each dot is a patient).

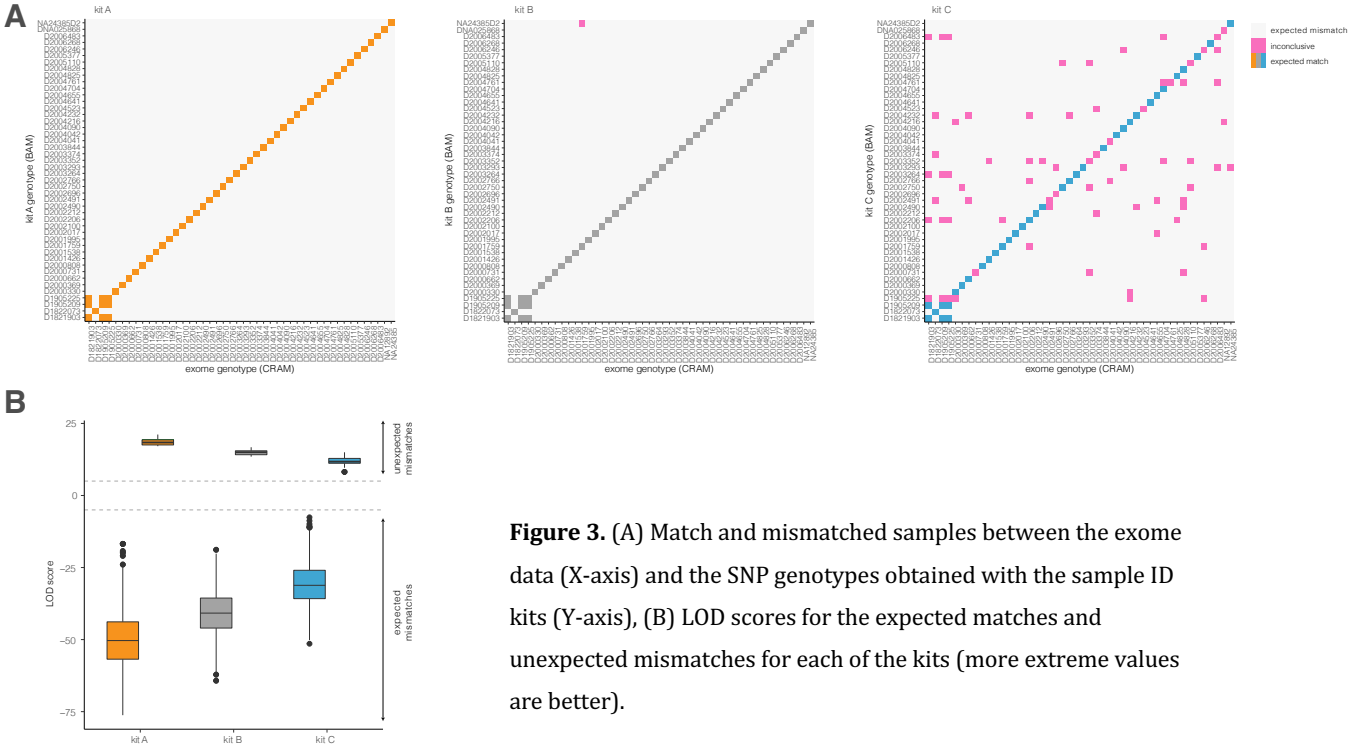

Figure 3. (A) Match and mismatched samples between the exome data (X-axis) and the SNP genotypes obtained with the sample ID kits (Y-axis), (B) LOD scores for the expected matches and unexpected mismatches for each of the kits (more extreme values are better).

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 kit A, unambiguous sample discrimination and identification was obtained for all samples. For kit B, one genotyping sample was not identified correctly, showing as a mismatch with the correct WES data, and being inconclusive with another unrelated sample. With kit C, 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, indicating that no sample mix-ups occurred. The excellent performance of kit A – and to a lesser extent kit B – 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 C shows the lowest discriminatory performance, with average LOD scores of ±6 being just above the inconclusive threshold. In comparison, LOD scores of kit A and kit B showed a much better discrimination with 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 kit A 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.

Conclusions

In a clinical setting, the three tested SNP sample tracking methods displayed significant differences in their sample identification and genotyping performance (Table 1). Overall, kit C showed to be unreliable with many samples that showed undecisive correlations, although on-target specificity was observed to be highest in this kit. Kit B performed best on coverage uniformity but showed poor on-target rates resulting in higher sequencing costs. From the three kits, kit A 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 the need for reanalyzing samples. Combined with an above-average on-target specificity and coverage uniformity, kit A shows the overall best per-sample cost-efficiency.

Table 1 Overview of performance parameters evaluated in this study (best score in bold)

Parameters	Human Sample ID kit (pxlence, kit A)	Human Identification and Sample tracking kit (Nimagen, kit B)	Accel-Amplicon Sample ID Panel (Swift Biosciences, kit C)
Number of SNPs	**44**	31	28
Number of gender markers	**6**	5	1
Coverage > 20% of mean	99.85%	**99.93%**	99.61%
Coverage within 2-fold of median	84.14%	**89.55%**	81.75%
Standard deviation of coverage	0.0040	**0.0029**	0.0076
Median % off-target	9.90%	22.82%	**4.43%**
Median LOD match (more is better)	**19**	13	6
Median LOD mismatch (less is better)	**-50.04**	-40.81	-31.17
Conclusive genotypes	**46/46**	45/46	20/46

Statements and declarations

Author Contributions

Conceptualization, Jo Vandesompele, Steve Lefever and Frauke Coppieters; methodology, Inge Vereecke and Mauro Milazzo; software, Pieter-Jan Volders and Steve Lefever; validation, Inge Vereecke and Mauro Milazzo; formal analysis, Pieter-Jan Volders and Steve Lefever; investigation, Mauro Milazzo and Inge Vereecke; writing—original draft preparation, Gertjan Wils; writing—review and editing, Frauke Coppieters, Jo Vandesompele, Mauro Milazzo, Steve Lefever and Kim De Leeneer; visualization, Steve Lefever; supervision, Jo Vandesompele; project administration, Céline Helsmoortel; All authors have read and agreed to the published version of the manuscript.

Funding

This research was in part funded by VLAIO, grant number HBC.2020.2643.

Informed Consent

Statement: Patient consent was waived due to the experiments not adversely affecting the rights and welfare of the subjects.

Conflicts of Interest

Gertjan Wils is employee of pxlence. Jo Vandesompele, Steve Lefever and Frauke Coppieters are co-founders of pxlence. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Shickh S, Mighton C, Uleryk E, Pechlivanoglou P, Bombard Y. The clinical utility of exome and genome sequencing across clinical indications: a systematic review. Vol. 140, Human Genetics. 2021. p. 1403–16.

2. Hu H, Liu X, Jin W, Ropers HH, Wienker TF. Evaluating information content of SNPs for sample-tagging in re-sequencing projects. Sci Rep. 2015 May 15;5:10247.

3. 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 [Internet]. 2015;144(4):667–74. Available from: https://academic.oup.com/ajcp/article/144/4/667/1767338

4. Kim J, Park W-Y, D Kim NK, Jin Jang S, Chun S-M, Sung C-O, et al. Good Laboratory Standards for Clinical Next-Generation Sequencing Cancer Panel Tests. J Pathol Transl Med [Internet]. 2017;51:191–204. Available from: https://doi.org/10.4132/jptm.2017.03.14

5. Pedersen BS, Quinlan AR. Mosdepth: Quick coverage calculation for genomes and exomes. Bioinformatics [Internet]. 2018;34(5):867–8. Available from: https://bioconda.github.io/

6. Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med [Internet]. 2013 Sep 27 [cited 2021 Dec 14];5(9):1–7. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/gm492

7. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Vol. 24, European Journal of Human Genetics. 2016. p. 2–5.

8. Moore RA, Zeng T, Docking TR, Bosdet I, Butterfield YS, Munro S, et al. Sample Tracking Using Unique Sequence Controls. J Mol Diagnostics. 2020;22(2):141–6.

9. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013 Sep;15(9):733–47.

10. Hume S, Nelson TN, Speevak M, McCready E, Agatep R, Feilotter H, et al. CCMG practice guideline: Laboratory guidelines for next-generation sequencing. J Med Genet. 2019;56(12):792–800.

11. Blackburn J, Wong T, Madala BS, Barker C, Hardwick SA, Reis ALM, et al. Use of synthetic DNA spike-in controls (sequins) for human genome sequencing. Nat Protoc. 2019;14(7):2119–51.

12. Deveson IW, Chen WY, Wong T, Hardwick SA, Andersen SB, Nielsen LK, et al. Representing genetic variation with synthetic DNA standards. Nat Methods. 2016;13(9):784–91.

13. Du Y, Martin JS, McGee J, Yang Y, Liu EY, Sun Y, et al. A SNP panel and online tool for checking genotype concordance through comparing QR codes. PLoS One [Internet]. 2017;12(9). Available from: https://doi.org/10.1371/journal.pone.0182438

14. Helsmoortel C, Kooy RF, Vandeweyer G. Multiplexed High Resolution Melting Assay for Versatile Sample Tracking in a Diagnostic and Research Setting. J Mol Diagnostics. 2016 Jan;18(1):32–8.

15. Westphal M, Frankhouser D, Sonzone C, Shields PG, Yan P, Bundschuh R. SMaSH: Sample matching using SNPs in humans. BMC Genomics. 2019 Dec 30;20(Suppl 12):1001.

16. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2010;26(5):589–95. Available from: https://academic.oup.com/bioinformatics/article/26/5/589/211735

17. Tools P. Broad Institute, GitHub Repository. 2019.