The identity predicament of human genetics. How to avoid patient sample mix-ups?

Vincent Dunon, Jo Vandesompele, Steve Lefever, Frauke Coppieters, pxlence bvba, Belgium

Reading time

7 minutes

Summary

- DNA sample mix-ups occur up to 3% in sequencing workflows
- These errors have substantial diagnostic and economic consequences
- Several elegant and affordable solutions exist to detect such sample mix-ups
- The Human Sample ID Kit enables straightforward resequencing of a SNP panel for unambiguous samples identification

Introduction

Throughout the years, massively parallel sequencing (MPS) has become an indispensable technology in molecular diagnostics. The combination of high-throughput data generation at rapidly decreasing cost allowed the implementation of whole-exome sequencing (WES) and whole-genome sequencing (WGS) in clinical genetic laboratories. MPS starts to play a crucial role in realizing personalized medicine and precision oncology.

Performing WES/WGS studies in-house involves a high start-up cost, complex workflows, and handling of large datasets. This imposes severe challenges on data integrity that range from the initial sample collection to the downstream data analysis. It is estimated that between 0.3% and 3% of all samples are compromised by provenance errors, raising major concerns about the integrity and reliability of NGS data[1]. In addition, human genetic research is suffering from the widespread use of misidentified and contaminated cell lines, occurring in up to 36% of cases, jeopardizing research outcomes[2,3].

Sources of errors and consequences

WES and WGS involve complex sample preparations with various manipulations that are often carried out by multiple individuals, including sample collection, sample storage, DNA extraction, barcoding, target enrichment, sequencing, and data analysis. Due to the high need of expertise and resources, some laboratories prefer to outsource the sample preparation, sequencing and/or bioinformatic analysis. Therefore, the multiple sample handlings and transfer of sample custody make these samples especially susceptible to provenance errors such as sample mix-up, cross-contamination and mislabeling[4,5,6]. As these errors are difficult to detect or to eliminate, implementation of appropriate measures are critical for the unambiguous re-identification of samples throughout all stages of the MPS workflow[7,8].

In both clinical diagnostics and research environments, identity mix-ups can have detrimental consequences. A wrong diagnosis leading to an incorrect or delayed treatment can cause severe harm to the patient. In research, erroneous data will impair discovery of new causal variants by yielding misleading variant candidates[5,9]. The same applies for contaminations leading to a loss of diagnostic and discovery power due to false-positive and false-negative variant calls. In biomedical research, cell lines are indispensable as in vitro models and the quality of the generated data greatly depends on the correct identification. Therefore, to increase reliability of published data, requirements on the authentication and purity of cell lines are imposed by various funding agencies and publishers[2]. Consequently, provenance errors and cell line misidentification also pose severe economic costs on healthcare systems and research funds[3,10]. Therefore, additional efforts to guarantee sample identification and cell line authentication are highly recommended.

Keeping track

Over the recent years, different guidelines were established, emphasizing the importance of implementing a sample tracking or authentication system. Risks are significantly reduced by carrying out good practice in sample handling, thorough documentation and implementation of (semi-)automated processes[8,11]. However, to assure continuous sample tracking throughout the entire MPS workflow, an additional independent confirmation of sample identity is highly desirable. Such a sample identification tool should therefore allow post hoc verification that the sequence results have been correctly assigned to each patient. By using genetic labels that are inherently linked to the sample from the initial sampling up to the data analysis and reporting, sample mislabeling and handling errors are removed from the workflow[1,5]. This also applies to cell line authentication, where genetic labels can be used for assessing the identity of the cell line and research results[2,3]. Several options are available to keep track of DNA samples during MPS workflows (Figure 1).

Short tandem repeats

Short tandem repeats (STRs), also known as microsatellites, have been widely adopted in forensic profiling as genetic markers. STRs consist of tandem repeated DNA units of 1 to 6 nucleotides that are abundantly present in the human genome. They are multiallelic and highly polymorphic. The American National Standards Institute (ANSI) and the American Type Culture Collection (ATCC) provided authentication standards for human cell lines based on STR profiling[2].

While suitable for occasional cell line authentication, STRs as genetic markers for patient sample tracking poses several limitations. The standard practice for

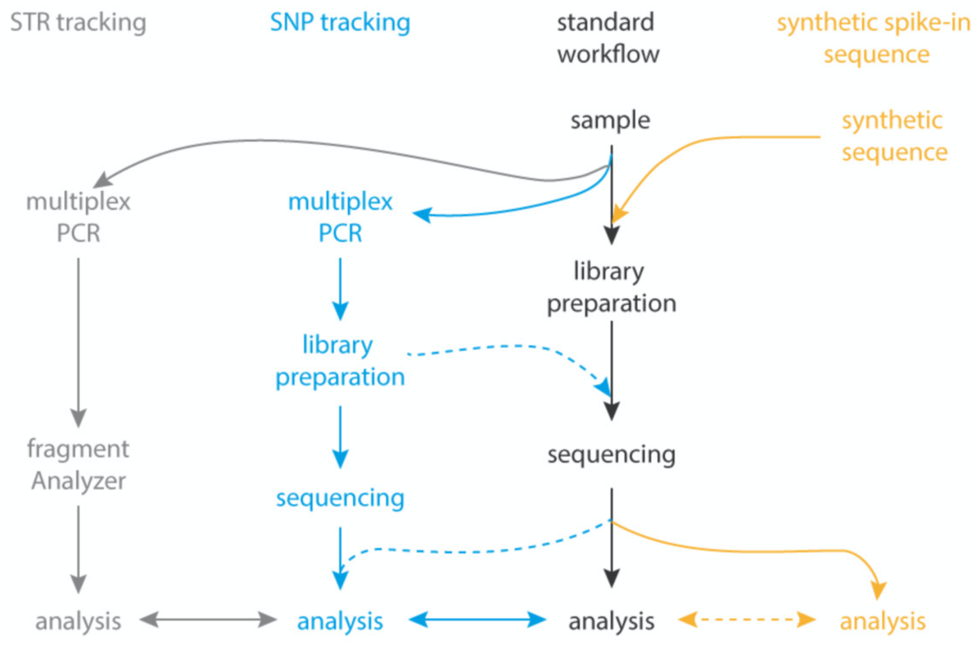

Figure 1: Different methods to achieve sample tracking or authentication and how they can be integrated into a general MPS workflow.

STR profiling requires capillary electrophoresis, limiting the throughput. Furthermore, profiling of STR with MPS is challenging due to the high variability in amplicons sizes reaching up to 350 bp, the repetitive nature and the high mutation rate. Most importantly, STRs are mainly found in non-exonic regions, making STR profiling unsuitable as sample tracking method for WES studies or other MPS approaches focusing on coding regions. Finally, MPS data analysis pipelines do not routinely screen for STRs due to a lack of adequate bioinformatic tools for high-quality STR genotyping[1,3,12,13].

Single nucleotide polymorphisms

Genotypic profiling using single nucleotide polymorphisms (SNPs) is another method to identify samples. SNPs are single nucleotide variations that occur widespread in the general population. They are the most common type of genetic variations and are found genome-wide[14,15]. As they are typically biallelic, they are less informative than STRs and therefore more SNPs are needed to achieve a similar level of discrimination[16]. Nonetheless, SNPs are becoming increasingly popular as genotyping tool for the identification of human samples as they offer several advantages compared to STRs. SNPs are stable genetic markers with a low mutation rate, providing a solution for correct authentication. Analysis can be performed on fragmented DNA samples (such as DNA from liquid biopsies or FFPE tissue) as only short amplicon sizes (< 100 bp) are needed, and due to their biallelic nature, genotyping can be more easily automated[14,17].

All of these qualities make SNPs suitable genetic markers for a cost-efficient genotyping method that is higly reliable, reproducible and transferable among laboratories[18]. Crucial is that all SNPs have a sufficiently high minor allele frequency and that the SNPs are represented in the MPS enrichment approach. This can be done by designing the SNP panel to be compatible with the specific MPS approach or by adding additional capture probes targeting the SNPs for identification. Different SNP panels have been established for sample identificatoin or forensic applications, including panels specifically designed for WES[5]. Our own panel consists of 44 SNPs and 6 gender markers and displays superior coverage uniformity and discrimination potential.

Synthetic spike-in DNA

As a third alternative, unique synthetic genetic sequences are added directly to the biological sample (ideally) at the time of receipt or during downstream steps in the workflow. As the genetic barcode remains inherently linked to the sample, it simultaneously undergoes the same handlings, allowing to confirm identity throughout the whole process. The spiked-in DNA sequences will be sequenced along with the normal MPS workflow and allow to verify sample identity and absence of contamination. The synthetic sequences should not impair with the MPS workflow of the biological sample, as can be done using artificial sequences sharing low similarity with any known DNA sequence in the human genome[8] or using the principle of mirrored human sequences[19]. For target enrichment applications, this also implies that additional probes will need to be added to the hybridization panel to pick-up the synthetic spike-in DNA.

Synthetic DNA spike-ins do offer the benefit of allowing to differentiate between samples with very similar or identical genotypes, e.g. twin studies or longitudinal studies on the same individuals[6,8,19]. Importantly, this method cannot be applied for cell line authentication nor for forensic studies. Moreover, many diagnostic laboratories may be reluctant to 'contaminate' patient samples by intentionally adding synthetic sequences.

Conclusion

With the rise of MPS, most sample identification methods are currently using an optimized SNP panel. Below, we made an overview of the applicability of the different identification markers for forensics, sample tracking and cell line authentication. The main advantages of SNPs over STRs are that they are widely available throughout the genome, including the exonic regions, and can be tested using small amplicons with fixed lengths. STRs remains the gold standard for sample identification in forensic applications and for cell line authentication. The principal reason is that from a historical perspective the forensic and cell line authentication databases are compiled using STR profiles. As an adequate alternative, synthetic sequences can be spiked into the biological sample, offering the advantage that sample is inherently linked to the identification marker and therefore are simultaneously processed and analyzed. However, this method is not applicable for forensic or cell line authentication purposes. Undoubtedly, SNP based sample identification will continue to increase popularity due to their ease of use and versatility.

	STR	SNP	spike-in DNA
forensics	x*	x	
WES/WGS		x*	x
cell line authentication	x*	x	

* standard practice

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. Almeida, J. L., Cole, K. D. & Plant, A. L. Standards for Cell Line Authentication and Beyond. PLoS Biol. 14, 1–9 (2016).

3. Freedman, L. P. et al. Reproducibility: Changing the policies and culture of cell line authentication. Nat. Methods 12, 493–497 (2015).

4. Lohr, M. et al. Identification of sample annotation errors in gene expression datasets. Arch. Toxicol. 89, 2265–2272 (2015).

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

6. Tourlousse, Di. M., Ohashi, A. & Sekiguchi, Y. Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls. Sci. Rep. 8, 1–9 (2018).

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

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

9. 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).

10. Lippi, G., Mattiuzzi, C., Bovo, C. & Favaloro, E. J. Managing the patient identification crisis in healthcare and laboratory medicine. Clin. Biochem. 50, 562–567 (2017).

11. Jennings, L. J. et al. Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. J. Mol. Diagnostics 19, 341–365 (2017).

12. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. LobSTR: A short tandem repeat profiler for personal genomes. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 7262 LNBI, 62–63 (2012).

13. Halman, A. & Oshlack, A. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Research 9, 1–29 (2020).

14. Tillmar, A., Grandell, I. & Montelius, K. DNA identification of compromised samples with massive parallel sequencing. Forensic Sci. Res. 4, 331–336 (2019).

15. Yousefi, S. et al. A SNP panel for identification of DNA and RNA specimens. BMC Genomics 19, 1–12 (2018).

16. Budowle, B. & Van Daal, A. Forensically relevant SNP classes. Biotechniques 44, 603– 610 (2008).

17. Kidd, K. K. et al. Developing a SNP panel for forensic identification of individuals. Forensic Sci. Int. 164, 20–32 (2006).

18. Roques, S., Chancerel, E., Boury, C., Pierre, M. & Acolas, M. L. From microsatellites to single nucleotide polymorphisms for the genetic monitoring of a critically endangered sturgeon. Ecol. Evol. 9, 7017–7029 (2019).

19. Blackburn, J. et al. Use of synthetic DNA spike-in controls (sequins) for human genome sequencing. Nature Protocols 14 (2019).