Detecting HTS Barcode Contamination

Mallory A. Clark, Sara H. Stankiewicz, Vincent Barronette, & Darrell O. Ricke

Abstract

DNA barcoding enables multiple samples to be characterized in parallel with high throughput sequencing (HTS) experiments for cost efficiencies. Cross-contamination of DNA barcode reagents can result in the detection of HTS sequences for barcodes that were not originally added to a particular sample. Cross-contamination of data between multiplexed samples can also occur. Avoidance and detection of contaminated barcodes is relevant for DNA forensic samples analysis, accurate cancer diagnosis, clinical research applications, metagenomic analysis, etc. We present recommendations for the avoidance of contamination and a tool, TallyBarcodes, to aid in the detection of DNA barcode contamination.

Key Words

Forensic Science, High-Throughput Sequencing, DNA Forensics, DNA Barcode, Next Generation Sequencing, Multiplexing

Introduction

DNA barcodes are short distinct segments of DNA that are used to identify one genetic sample from another. DNA barcodes enable scientists to run multiple samples in parallel, known as multiplexing[1]. Multiplexing reduces the costs and time of sequencing. However, DNA barcode reagents are susceptible to cross-contamination[2]. This contamination is easily seen when large counts for barcodes that were not used in a particular experiment are detected in the data. Finding non-random amounts of incorrect barcodes is evidence of cross-contamination in DNA barcode reagents[3]. Cross-contamination can also arise from laboratory experimental mistakes[4] [Bandelt & Salas 2009], during primer synthesis, multiple errors in HTS sequencing, or sample carryover from previous sequencing runs on the same instrument[5].

Sample misidentification was first observed by Kircher *et al.* [6] who recommended double indexing on the Illumina platform to mitigate the issue. Quail *et al.* [7] proposed SASI-Seq with sample assurance Spike-Ins to enable sample assurance. Wright & Vetsigian [8] reported read misassignments averaging at 0.24% on a Illumina HiSeq 2500; they propose quality filtering to alleviate sample cross-talk. Similar index misassignment rates between 0.06 and 0.21% are reported by[5, 9]. Sinha *et al.* [10] found that 5-10% of sequencing reads (signals) in a multiplexed pool of samples are mislabeled due to index switching on the Illumina HiSeq 4000 platform. Echoing Kircher *et al.* [6], MacConaill *et al.* [11] propose dual-indexed sequence adapters to eliminate index cross-talk. Esling *et al.* [2] also found that cross-contamination can occur between DNA barcodes and various primers and reagents. Integrated DNA Technologies (IDT)[12] reports 0.1% to 0.5% cross-contamination of HPLC-purified DNA oligos and 0.01% to

0.05% for Ultramer DNA oligos. To mitigate cross contamination risks in a laboratory, both Esling *et al.* [2] and Sinha *et al.* [10] recommend to construct bar code arrangements that do not overlap and to use uniformed lab handling and cleaning procedures[2, 10]. Some other recommended precautions are to label PCR replicates with one-time use tagged primers, prohibit single-tagging and double-tagging sample arrangements, and to not use long primer constructs for mixture samples [2].

Materials & Methods

The software tool, TallyBarcodes, was developed to facilitate the detection of DNA barcode cross-contamination within experiments. TallyBarcodes enables the detection of contamination by quantifying the barcodes detected on the reads in the experimental data. TallyBarcodes is written in the Scala programming language.

Ion Torrent HTS Sequencing

Thermo Fisher PGM, Proton, and S5 libraries were prepared and sequenced as described in Ricke *et al.* [13].

Results

Two possible different types of sample misidentification from barcode cross-contamination are laboratory and n-1 series contamination. Both types of contamination have been detected in HTS experiment results. First, experimental results are consistent with laboratory contamination of barcode IX-08 with IX-16 during a specific experiment at a level of 0.0105 with standard deviation of 0.0016; the tube for barcode IX-16 is adjacent to IX-08 in a box of 16 barcodes with two rows of 8 tubes. Second, the n-1 series contamination patterns are characterized by trace contamination for a previous barcode reagent (n-1, n-2, etc.) in a sequentially numbered series for barcode reagents (n). Example n-1 series contamination patterns for different barcodes average 1.0×10^{-4} for n-3 and n-2 to 1.5×10^{-4} for n-1 and a maximum of 3.0×10^{-4} to 5.0×10^{-4} for n-1. Background random barcodes counts are less than 10 for HTS experiments averaging 100 million sequences.

Discussion

While the source(s) of cross-contamination of DNA barcodes can remain unknown, instances of accidental cross-contamination of DNA barcode reagents has been observed in a research laboratory setting. The n-1 series contamination pattern that was observed is inconsistent with misread bases within the barcode sequence identified, assuming $<1\%$ base calling error rate[8]. Possible contributors of cross-contamination are splatter between tubes, incomplete purification (e.g., reuse of HPLC purification columns), etc.

Detection of individuals in complex DNA forensic mixtures, cancer cells, metagenomic community characterization, etc. are pushing the detection levels lower with advances in HTS

technologies. Ricke *et al.* [13] can identify DNA contributors down to 1 in 400 concentration level. The n-1 series contamination levels of $3e-4$ places a limit of 1 in 3,333, which is below current minor contributor detection level. The laboratory contamination of barcode IX-08 with IX-16 at 0.0105 is detectable at 1 in 95, which will prevent future combinations of these two barcodes in experiments until barcode IX-08 is replaced. A hypothetical combination of a normal individual sample with barcode IX-08 and a cancer sample or DNA mixture on barcode IX-16 could lead to incorrect conclusions from the results.

While certified laboratories have standard operating procedures (SOPs) to minimize cross-contamination between reagents, we present additional recommendations to reduce and detect contamination.

Recommendations:

- Consecutive experiments should use different barcodes
- Barcodes used together in one experiment should be combined with different barcodes in the next experiment to aid in detection of possible laboratory contamination
- Selected barcodes should be spaced to mitigate n-1 series contamination
- Double labeling of samples with two barcodes
- Clinical and DNA forensic samples should use special grade barcodes that have not been purified on equipment (e.g., HPLC columns) used to purify other barcodes
- Patient samples should not be multiplexed in the same experiments as cancer/tumor samples
- Forensic reference samples should be run in different experiments rather than mixture or crime scene samples
- Periodic screening of barcodes for possible contamination
- Replication of samples with different barcodes if contamination is suspected

The software tool, TallyBarcodes, is being provided as open source (<https://github.com/doricke/TallyBarcodes>) to facilitate characterization of HTS data.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Air Force.

References

- [1] A. M. Smith *et al.*, "Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples," *Nucleic Acids Research*, vol. 38, no. 13, pp. e142-e142, 2010.

- [2] P. Esling, F. Lejzerowicz, and J. Pawlowski, "Accurate multiplexing and filtering for high-throughput amplicon-sequencing," *Nucleic Acids Research*, vol. 43, no. 5, pp. 2513-2524, 2015.
- [3] K. H. Wong, Y. Jin, and Z. Moqtaderi, "Multiplex Illumina Sequencing Using DNA Barcoding," *Current Protocols in Molecular Biology*, vol. 101, no. 1, pp. 7.11.1-7.11.11, 2013/01/01 2013.
- [4] A. Alonso *et al.*, "Mitochondrial DNA haplotyping revealed the presence of mixed up benign and neoplastic tissue sections from two individuals on the same prostatic biopsy slide," *Journal of Clinical Pathology*, vol. 58, no. 1, pp. 83-86, 08/03/accepted 2005.
- [5] R. D'Amore *et al.*, "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling," *BMC Genomics*, vol. 17, no. 1, p. 55, 2016/01/14 2016.
- [6] M. Kircher, S. Sawyer, and M. Meyer, "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform," *Nucleic Acids Research*, vol. 40, no. 1, pp. e3-e3, 2012.
- [7] M. A. Quail *et al.*, "SASl-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing," *BMC Genomics*, vol. 15, no. 1, p. 110, 2014/02/07 2014.
- [8] E. S. Wright and K. H. Vetsigian, "Inhibitory interactions promote frequent bistability among competing bacteria," *Nature Communications*, Article vol. 7, p. 11274, 04/21/online 2016.
- [9] M. C. Nelson, H. G. Morrison, J. Benjamino, S. L. Grim, and J. Graf, "Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys," *PLOS ONE*, vol. 9, no. 4, p. e94249, 2014.
- [10] R. Sinha *et al.*, "Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing," *bioRxiv*, 10.1101/125724 2017.
- [11] L. E. MacConaill *et al.*, "Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing," *BMC Genomics*, vol. 19, p. 30, 2018.
- [12] I. D. Technologies. *Tips for minimizing sample misassignment during multiplexed NGS*. Available: <https://www.idtdna.com/pages/education/decoded/article/tips-for-minimizing-sample-misassignment-during-multiplexed-ngs>
- [13] D. O. Ricke *et al.*, "The Plateau Method for Forensic DNA SNP Mixture Deconvolution," *bioRxiv*, 10.1101/225805 2017.