Brief Overview of Index Swapping Observations

Grant Kinsler*, Kerry Geiler-Samerotte*, Dmitri Petrov
Contact: grantkinsler@gmail.com, kerry.samerotte@gmail.com, @GrantKinsler, @KSamerotte

13 February 2019

Index swapping tends to occur most frequently on Illumina machines with patterned flow cells/ use “ExAmp” technology - these include HiSeq 3000, 4000, X, and the NovaSeq (thus far). It seems to not be as big a problem on non-non-pattered flow cell machines (MiSeq, HiSeq 2500, NextSeq). We’ve been doing a lot of direct comparison between HiSeq X and NextSeq and find that NextSeq has very minimal rates of swapping - we also have collaborators who’ve tested the HiSeq 2500 and MiSeq and, similarly, find minimal rates. Unfortunately (for sake of this particular problem), it seems that Illumina may continue to use this technology in their new machines (i.e. NovaSeq), indicating that it’s something that needs to be seriously tackled/ considered for any method going forward.

There are a few different estimates of index swapping rates. For our barcode sequencing, when we used 96 separate samples (all uniquely indexed), we found that ~45% of our reads were from sample combinations that should not exist on the HiSeq X. In comparison, for the exact same library, we see ~2% of reads from swapped combinations on the NextSeq. See our Twitter thread with some more information about this: https://twitter.com/GrantKinsler/status/1040750784697131008.

Note: Our rates may be higher than more diverse samples - we have evidence (again shown in that Twitter thread), that template swapping (on the sequencing machine) may be dramatically inflating our observed rates of swapping. Because we are using sequencing many virtually identical sequences (only a small part of which is a variable barcode), this homology may be allowing for more template swapping than might be expected in a whole genome context. For instance, the original preprint (https://www.biorxiv.org/content/10.1101/125724v1) describes between 5-10% of their reads were swaps. Illumina believes the majority of this problem is due to free adapters in libraries (which we believe is NOT the primary driver of swapping in our samples). We also find that the rates that we observe somewhat vary depending on fraction of the lane taken up by our samples - our 45% rate was when we took up ~70% of the HiSeq X lane. We get something closer to a 12% rate when ~30% of the lane consists of our samples.

We haven’t explicitly tested swapping between lanes on the same run of the machine, but we’re not sure exactly how this would happen, since, presumably, the lanes are physically separated and don’t allow for molecules to transfer between them. One possible source of cross-contamination that might look like swapping between lanes would be primer contamination, but we think this is very unlikely in our case.

Again, here’s the original preprint that described the index swapping problem in the context of RNA sequencing: https://www.biorxiv.org/content/10.1101/125724v1

And here’s a section of Illumina’s website (along with a white paper) that explains what they think is the predominant reason this is happening (for non-amplicon contexts): https://www.illumina.com/science/education/minimizing-index-hopping.html