Automated immune repertoire sequence correction using ErrorX

Endeavor Bio

Overview

Our flagship software, ErrorX, is an automated pipeline for correction of sequencing errors in B-cell and T-cell repertoire sequencing data. During the sequencing process, errors are introduced both by PCR amplification during library preparation, and by base miscalling on the sequencing instrument. These errors are indistinguishable from naturally occurring variants that arise from somatic mutation. This causes endless frustration for researchers who can't distinguish between true, biologically interesting mutations, and artifacts of sequencing error. ErrorX solves this problem by introducing a fully automated method to identify sequencing errors in datasets before they can be mistaken for biological mutants.

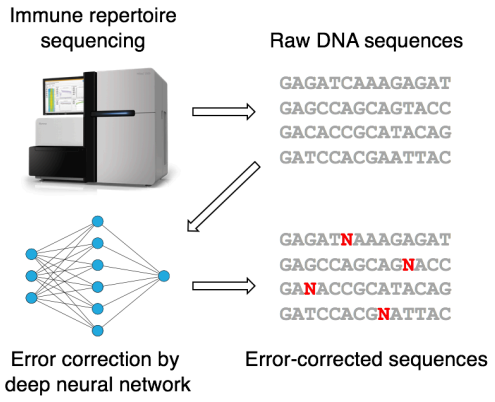

Figure 2. The ErrorX pipeline uses deep neural networks to correct immune repertoire sequencing data.

Our technology

ErrorX uses deep neural networks trained on real sequencing datasets to identify errors introduced by sequencing or PCR (Figure 1). Properties such as nucleotide motifs, quality scores, and inferred germline sequences are extracted from the dataset and fed into the network. ErrorX then uses deep learning to identify patterns that lead to a higher chance of sequencing error. New datasets can then be scanned for these patterns, pinpointing positions that have a high chance of being errors, so that these are not mistaken for true biological variants.

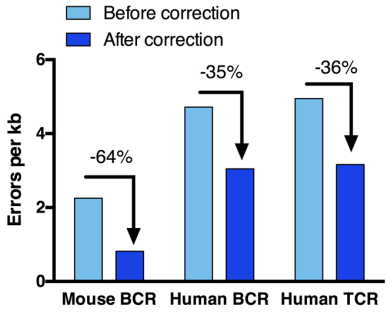

Figure 1. Error rate before and after ErrorX correction on three independent datasets.

Performance

In benchmark studies, ErrorX was applied to B-cell and T-cell repertoire sequencing data from public datasets. **We found that ErrorX error correction was able to reduce error rates by up to 63%** (Figure 2). Error identification was highly specific, with a false positive rate of < 0.001% in all cases.

ErrorX performance was similar when errors were present in the framework region, CDR loops, or non-templated junction region. In addition, sequences were corrected with great efficiency, needing only 15 seconds per kb with germline assignment and 3 seconds per kb when the germline has already been assigned beforehand. This allows ErrorX to be integrated seamlessly into an existing analysis pipeline.

The ErrorX advantage

- Trained on real data, not synthetic repertoires
- Training is done beforehand, with rapid application to new datasets
- Can be applied to previously collected datasets with no change in library preparation
- Runs on unlabeled data – no need for unique molecular identifiers (UMI)
- No loss of rare clones
- Can be easily integrated into an existing pipeline

Contact us at contact@endeavorbio.com to learn more about how ErrorX can improve your data!