Abstract
Developing and validating a bioinformatics pipeline for a clinical assay is frequently a costly process; in some cases it includes the acquisition of limited patient samples with extremely rare genotypes. To aid in the validation process, we have developed a Galaxy workflow to generate synthetic FASTQ files with known mutations provided by the user along with those sourced from dbSNP. Researchers can use these FASTQ files to mimic various mutations types including single nucleotide variants, insertions, deletions, translocations, and copy number variations. Researchers can optimize and evaluate the expected efficiency of their bioinformatics pipelines using synthetic samples simulated at various sequencing depths to mimic both germline and somatic events. In addition, changes to existing pipelines can easily be re-verified using static synthetic datasets as the gold standard reference. The workflow uses several open source tools to retrieve the genomic location of variants in HGVS notation (TransVars) and simulate reads from common sequencing platforms (ART).

Materials and Methods
We created a workflow that converts variants defined in HGVS nomenclature to their respective genomic positions using TransVar and injects the variants into a reference sequence for read simulation by ART. The first step in the workflow uses TransVar to retrieve the reference and alternate bases, genomic coordinates, and alteration type of each variant. The second step in the workflow is the retrieval of the corresponding reference sequence for the target regions using subcommand getrefseq from bedtools. The third step uses the outputs from both TransVar and bedtools to generate a synthetic reference in fasta format for the target region that incorporates the variants that were passed in. The synthetic reference generation tool transforms the annotation from TransVar into a set of variant definitions used to modify the reference. Supported variant definitions are: single nucleotide polymorphisms, deletions, insertions, translocations, inversions, and copy number gains. Optionally, users can provide a variant definition file in tabular format. The final step requires passing the synthetic reference through ART, a read simulation tool. The command line version of the tool allows users to generate and pass in a base call and quality profile. This option can be used to emulate sequencer specific biases.

A total of 1000 variants (SNPs, insertions, and deletions) for chromosome 1 were randomly sourced from dbSNP v138 for GRCh37. Variants were passed into TransVar to retrieve the genomic coordinates and altered bases. The variants were then incorporated into the reference sequence for the regions of chromosome 1 covered in Agilent's clinical research exome. To simulate a typical exome sample, the following parameters were used: paired end sequencing, mean target coverage of 100X, read length of 75 bases, insert size of 125 with a standard deviation of 35 bases. The parameters for read simulation were chosen to mimic the data derived from wet lab experiments. A sample representing targeted sequencing by a custom panel as simulated by sequencing regions on MET, ALK, EML4, ROS1, and SLC34A2.

Results

Table 1. Variants identified in synthetic samples. For the synthetic exome, TransVar was able to retrieve the genomic positions corresponding to 964 variants. For the synthetic tumor, all structural variant events were identifiable at the 1:3 dilution. The simulated translocation between ROS1 and SLC34A2 was not identified in the 1:9 dilution, and the log2 ratio for copy number between the tumor and normal was 0.34.

Conclusions

Synthetic reference generation provides an additional means to confirm the specificity of a bioinformatics pipeline in cases where cell lines or patient samples containing specific variants are difficult or impossible to procure.

Synthetic datasets provide a consistent test set that not only aid in validation but also help revalidate any updates to an existing pipeline.

Identify regions in the target panel with high homology that could possibly interfere with variant calling tools due to sequencing constraints or biological makeup.

Although, synthetic datasets offer many advantages over in vitro approaches, laboratories validating an assay should incorporate both the usage of synthetic datasets as well as patient samples and cell lines. Both the sample type and quality will play a major role in producing high sequencing quality data from the assay.

References