Exorcise.

Exome-guided reannotation of nucleotide sequences (exorcise) is available at:

<https://github.com/SimonLammmm/exorcise/>

This document describes the basic usage and syntax of exorcise.

Contents:

* Introduction
* Syntax
* Modes
* Outputs

Introduction

Sequences are often prescribed with annotations based on a certain genome assembly. It is not always appropriate to accept these annotations for all use cases of those sequences. For instance, a library of sgRNA sequences for CRISPR-Cas9 may be designed based on GRCh37 and sgRNAs annotated as "targeting" a certain gene or "non-targeting". Sequence search of those sgRNAs in GRCh38 might reveal discrepancies because of updates to the assembly; further, cell lines which do not perfectly reflect GRCh37 will show differential reactivity to the CRISPR guides as compared their annotations. Exorcise reannotates sequences based on their presence or absence in a user-supplied genome and exome. If the supplied genome and exome reflect the subject under consideration, then the user can be confident as to the presence or absence of sequences in the subject and the validity of the annotations.

Exorcise aligns sequences to the genome and implements set mathematics operations between alignment targets and exon coordinates to transfer annotations to the perfect alignments. It is written entirely in R and relies on BLAT for perfect sequence alignment.

Exorcise calculates sequence-level reannotations whereby each sequence is considered in turn. If provided a grouping variable (for example, original annotations), exorcise also calculates group-level reannotations where a consensus sequence-level reannotation is determined for each group and applied to all sequences in the group. We refer to this behaviour as harmonisation.

Exorcise is developed and maintained by Dr Simon Lam and is available at <https://github.com/SimonLammmm/exorcise/>.

Syntax

Exorcise takes mandatory and optional arguments.

|  |  |  |  |
| --- | --- | --- | --- |
| Long flag (short flag) | Mandatory? | Value | Description |
| --infile (-i) | Yes | File | File containing sequences to be exorcised. Arbitrary columns will be returned in the output file. |
| --outdir (-o) | Yes | Directory | Working directory to place intermediate files and the output file exorcise.tsv. |
| --seq (-g) | Yes | Number | Column number in the infile that corresponds to sequences to be exorcised. 1-based integer, must not be greater than the number of columns in the infile. |
| --pam (-z) | No | Nucleotide | Sequence to be appended to the 5' end of each sequence for BLAT purposes only. Must be in the single-letter nucleotide alphabet (ACGTN). |
| --library (-l) | Yes, if -v -w not specified | File | Exorcised library to be used for the re-annotation. If this is not specified, then genome and exome must be specified. Must be an exorcise library. Enables post-hoc mode (see Modes section). |
| --genome (-v) | Yes, if -l not specified | File | Genome in 2bit format. Enables ad-hoc mode (see Modes section). Ignored if library is specified. |
| --exome (-w) | Yes, if -l not specified | File | Exome from UCSC Table Browser. Enables ad-hoc mode (see Modes section). Ignored if library is specified. |
| --harm (-n) | No | Number | Column number in the infile containing groups. 1-based integer, must not be greater than the number of columns in the infile. Enables harmonisation (see Modes). |
| --priorities (-y) | Yes, if -n specified and not 0 | File | Feature priorities list in NCBI Datasets format. Ignored if harm is not specified. Enables harmonisation (see Modes). |
| --control (-c) | No | String | Comma-separated list of strings to treat as control sequences. R regex allowed. Ignored if harm is not specified. Enables control reannotation (see Modes). |
| --control\_type (-d) | No | String | Comma-separated list of strings of the same length as control indicating control type. Ignored if control is not specified. Ignored if harm is not specified. |
| --help (-h) | No | Flag | Show help and then exit. |

Modes

Exorcise can use previous exorcise outputs (exorcise libraries) to reannotate sequences without recalculation of the entire sequence vector each time. This is called post-hoc mode and is extremely fast. Post-hoc mode takes precedence over ad-hoc mode (see below).

When an exorcise library is not specified, then it is calculated from the genome and exome. This is called ad-hoc mode. Ad-hoc mode can take a long time (from 20 minutes to over a week, depending on inputs) to run BLAT. The output can be used as an exorcise library the next time that sequences need to be reannotated in the same genome/exome context.

Both post-hoc and ad-hoc modes support harmonisation on the harm column if specified along with a feature priorities file. The harmonisation algorithm can take a long time if there are many groups in the harm column.

The following table shows exorcise behaviours when different sets of inputs are received. Combined behaviour is allowed by specifying the necessary combinations of inputs. (Note that the inputs -i, -o, and -g are mandatory and so are omitted from the table.)

|  |  |
| --- | --- |
| Inputs | Behaviours |
| -v, -w | Ad-hoc mode: aligns guides to the genome and uses set theory to find exonic hits. |
| -l | Post-hoc mode: uses an existing exorcise library to reannotate input sequences. |
| -n, -y | Harmonisation: also performs group-level reannotation. |
| -n, -y, -c (-d) | Control reannotation: also reannotates controls to the optional vector control\_type. |
| -l, omit -n, -y | Use post-hoc mode and inherit existing harmonisations from the library, if any. |
| -l, -n 0 | Use post-hoc mode and explicitly specify not to inherit existing harmonisations. |

Outputs

Exorcise outputs a single exorcise.tsv file as well as intermediate files for diagnostic purposes. If you are happy with the results, then the intermediate files can be removed.

Exorcise uses checkpointing. It checks for the presence of an intermediate file and uses it rather than regenerating it if it exists from a previous run.

|  |  |
| --- | --- |
| File | Description |
| exorcise.1-seq.fa | Sequences plus PAM to be sent to BLAT. |
| exorcise.2-genome.2bit\_BLAT.psl | BLAT genome alignment results. |
| exorcise.3-genome.2bit\_genomicRanges.tsv | Coordinates of BLAT alignments and calculated cut sites. |
| exorcise.3-genome.2bit\_genomicSeqSpecs.tsv | Coordinates of BLAT alignments to be sent for sequence validation. |
| exorcise.3-genome.2bit\_genomicSeqs.fa | Validated sequences. |
| exorcise.4-exome.gz\_exonHits.tsv | BLAT alignments with annotations inherited from the exome. |
| exorcise.5-exome.gz\_exonDist.tsv | BLAT alignments and exome annotations with distance to the nearest exon. |
| exorcise.tsv | Main exorcise output file. |

The main exorcise output file is exorcise.tsv. Exorcise outputs are in columns starting with exo\_. Arbitrary columns included in the input file are appended after the exorcise columns.

|  |  |
| --- | --- |
| Column | Description |
| exo\_id | Unique identifier for this exorcise reannotation. |
| exo\_seq | Sequence. |
| exo\_symbol | Sequence-level annotation from exome. |
| exo\_harm | Group-level annotation (harmonisation) from exome. |
| exo\_orig | Original annotation. |
| exo\_target | Genome coordinates of alignment between sequence plus PAM and the genome. |
| exo\_cut | Genome coordinates of the cut site. |
| ... | Arbitrary columns included in the input file. |