HeteroGenesis

Introduction

HeteroGenesis is used to generate genomes for multiple related clones in a heterogeneous tumour, along with a matched germline genome. For each clone and germline sample, it provides FASTA files containing the sequences for each copy of a chromosome in the genome, and files detailing the variants incorporated.

A separate tool, FreqCalc, can be used to combine the variant call format (VCF) and copy number files generated by HeteroGenesis to give overall bulk tumour outputs that reflect user defined proportions of each clone in a tumour, and its purity. This is useful, for example, when the user intends to carry out *in silico* sequencing of each clone and combine the reads to form a bulk tumour dataset.

For more information, see "Simulation of Heterogeneous Tumour Genomes with HeteroGenesis and In Silico Whole Exome Sequencing, Tanner G et al., 2018." (Manuscript pending submission)

Please also cite this when using HeteroGenesis or FreqCalc.

Requirements

Python3 is required to run both HeteroGenesis and FreqCalc. numpy is also required for running HeteroGenesis. Both tools have been tested with Python 3.5.2 and numpy 1.12.0 and 1.12.1.

Installation

In the heterogenesis directory, type: `python setup.py install`

Overview

HeteroGenesis is implemented in two parts:

The first, **heterogenesis_vargen**, takes: i) a FASTA genome sequence, ii) a .fai index file for the genome sequence, iii) an optional file containing known germline SNV and InDel locations and minor allele frequencies from dbSNP, and iv) a JSON file containing a set of parameters. It outputs a JSON file with lists of variants for the germline and each clone in the simulated tumour, as well as files containing the order that mutations occurred in each.

The second part, **heterogenesis_varincorp** is then run, once for each clone, and incorporates the list of variants for a clone into a reference genome. It outputs: i) the FASTA genome sequence (one file for each

copy of a chromosome), ii) a VCF file of SNV and InDel positions and frequencies, and iii) a file containing the copy numbers along the genome.

FreqCalc can then be run to cobine outputs from clones to generate bulk tumour outputs. It takes a file containing the proportions of each clone in a tumour, along with the outputs from heterogenesis_varincorp, and outputs equivalent files for the bulk tumour.

Inputs

heterogenesis_vargen

1. **Reference Genome:** The starting genome sequence, in FASTA format, that variants will be incorporated into.
2. **Reference Genome Index:** A .fai index file for the reference genome, created with samtools faidx. This should be saved in the same directory as the reference genome.
3. **dbSNP File:** ...
4. **Parameters File:** A JSON file containing run parameters and locations of other inputs. Any parameter that is missing from the file will be set at its default value:

   (An example parameters file is provided in the repository - 'example.json')

Parameter	Description	Default Value
prefix	String added to output file names.	""
reference	FASTA file containing the sequence of a reference or other input genome. Must have a .fai index file located in the same directory.	Required
dbsnp	VCF file from dbSNP containing known germline SNPs and InDels.	none
directory	Directory to output all files to.	"./"
	Structure of clones in the tumour, in the format: “clone1_name, clone1_distance_from_parent, clone1_parent_name, clone2_name,	

structure	clone2_distance_from_parent, clone2_parent_name…". All parent clone names must also be listed as a separate clone, ie. if clone2's parent clone is clone1, then clone1 must also be listed as a clone with a parent clone. The exception to this is when the parent clone is 'germline', and this must occur at least once as the parent clone for the root clone of the tumour. Loops in the lineage will cause the program to never end, ie. clone1->clone2->clone3->clone1. Distances from parent clones can be any fraction or number as they are used relative to each other.	"clone1,0.2,germline,clone2,0.8,clone1"
snvgermline	Rate of germline SNVs per base.	0.00014
indgermline	Rate of germline indels per base	0.000014
cnvrepgermline	Number of germline replication CNVs.	160
cnvdelgermline	Number of germline deletion CNVs.	1000
aneuploid	Number of somatic aneuploid events. i.e. replication or deletion of a chromosome. Copy number is randomly chosen from 0, 2 or 3. Germline aneuploid events are not available.	2
snvsomatic	Rate of somatic SNVs per base.	0.00001
indsomatic	Rate of somatic indels per base.	0.000001
cnvrepsomatic	Number of somatic replication CNVs.	250
cnvdelsomatic	Number of somatic deletion CNVs.	250

dbsnpsnvproportion	Proportion of germline SNVs taken from dbSNP. The default value is taken from an estimate of the proportion of SNVs found in dbSNP for coding regions, to make the genomes suitable for use with whole-exome *in silico* sequencing. The user may therefore wish to adjust this if they intend to use the genomes for other purposes.	0.9
dbsnpindelproportion	Proportion of germline InDels taken from dbSNP. Default value taken from estimates for coding regions, as above.	0.5
chromosomes	List of chromosomes to include in the model. Alternatively, "all" can be given, in which case chromosomes 1-22 will be used. This only works for genomes for which chromosomes are labelled 'chr1','chr2'... (Also note that X and Y are not included with "all")	”all”

CNV lengths and copy numbers, and indel lengths are taken from lognormal distributions, that are defined by the mean and variance of the underlying normal distribution. Values from these distributions are then scaled up by a multiplication factor for cnv lengths. Indel length distributions are the same for germline and somatic.

cnvgermlinemean	Germline CNV length lognormal mean.	-10
cnvgermlinevariance	Germline CNV length lognormal variance	3
cnvgermlinemultiply	Germline CNV length multiplication factor.	1000000
cnvsomaticmean	Somatic CNV length lognormal mean.	-1
cnvsomaticvariance	Somatic CNV length lognormal variance.	3
cnvsomaticmultiply	Somatic CNV length multiplication factor.	1000000
indmean	Indel length lognormal mean.	-2
indvariance	Indel length lognormal variance.	2
indmultiply	Indel length multiplication factor.	1
cnvcopiesmean	CNV copies lognormal mean.	1
cnvcopiesvariance	CNV copies lognormal variance.	0.5

heterogenesis_varincorp

1. **Parameters File:** The same JSON file as used for heterogenesis_vargen can be given but only the following parameters are used. These should contain the same values as given for heterogenesis_vargen:

Parameter	Description	Default Value
prefix	String added to output file names.	""
reference	FASTA file containing the sequence of a reference or other input genome. Must have a .fai index file located in the same directory.	Required
directory	Directory containing JSON output from heterogenesis_vargen and where output files will be written to.	"./"
chromosomes	List of chromosomes included in the model. Alternatively, "all" can be given, in which case chromosomes 1-22 will be used. This only works for genomes for which chromosomes are labelled 'chr1','chr2'... (Also note that X and Y are not included with "all")	”all”

FreqCalc

1. **Clones File:** File with clone proportions in format: 'clone name' \t 'fraction'.
2. **Outputs From heterogenesi_varincorp**

Outputs

heterogenesis_vargen

...

heterogenesis_varincorp

...

FreqCalc

...

Implementation

heterogenesis_vargen and **heterogenesis_varincorp** are run with the following:

```
heterogenesis\_vargen -j example.json
heterogenesis\_varincorp -j example.json
```

-v/--version : Version

-j/--json : JSON file containing parameters.

FreqCalc is run with:

```
freqcalc -c clones.txt -d {directory of HeteroGenesis outputs} -p {prefix}
```

-v/-—version : Version

-c/--clones : File with clone proportions in format: 'clone name' \t 'fraction'.

-d/--directory : Directory containing VCF and CNV files.

-p/--prefix : Prefix of VCF and CNV file names. This will be the same as what was provided for the 'prefix' parameter with HeteroGenesis.