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 submitted)

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.

Usage

HeteroGenesis is implemented in two parts:

The first, **HeteroGenesis_vargen.py**, takes a FASTA genome sequence, an optional VCF file containing known germline SNV and InDel locations from dbSNP, and a JSON file containing a set of parameters, and outputs a JSON file with a lists of variants for each clone (including the germline genome) in the simulated tumour, as well as files containing the order that mutations occurred.

The second part, **HeteroGenesis_varincorn.py** is then run, once for each clone, and incorporates the list of variants for the clone into the reference genome. It outputs: i) the FASTA genome sequence (one file for each copy of a chromosome), ii) a VCF file of SNP/SNV and InDel positions and frequencies, and iii) a file containing CNV/CNA positions and copy numbers.

HeteroGenesis_vargen.py

-v/--version : Version

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

Parameter	Description	Default value
prefix	String added to output file names.	""
reference	FASTA file containing the sequence of a reference or other input genome.	Required
genome	A ‘genome’ file containing chromosome names in the first column and chromosome lengths in the second. Alternatively, a FATSA index file (fasta.fai) may be provided, which also includes these columns.	Required
dbsnp	VCF file from dbSNP containing known germline SNPs and InDels.	none
directory	Directory to output all files to.	"./"
structure	Structure of clones in the tumour, in the format: “clone1_name, clone1_distance_from_parent, clone1_parent_name, clone2_name, 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-	"clone1,0.2,germline,clone2,0.8,clone1"

	>clone3->clone1. Distances from parent clones can be any fraction or number as they are used relative to each other.	
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.	0.000002
cnvrepsomatic	Number of somatic replication CNVs.	250
cnvdelsomatic	Number of somatic deletion CNVs.	250
dbsnpsnvproportion	Proportion of germline SNVs taken from dbSNP	0.9
dbsnpindelproportion	Proportion of germline InDels taken from dbSNP	0.5
chromosomes	List of chromosomes to include in the model - these must match with chromosome names in the reference and genome files. The user may wish to leave out chrY, as all chromosomes start with 2 copies. Alternatively, the user can set this parameter to “all” which will include ‘chr1’,	”all”

	‘chr2’…’chrX’ etc. (no chrY), if the reference and genome files also contain these names.	

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.py

-v/--version : Version

-j/--json : JSON file containing parameters. The same JSON file as used for HeteroGenesis_vargen.py can be given but only the following parameters are used. These should contain the same values as given for HeteroGenesis_vargen.py:

Parameter	Description	Default value
prefix	String added to output file names.	""
reference	FASTA file containing the sequence of a reference or other input genome.	Required
genome	A ‘genome’ file containing chromosome names in the first column and chromosome lengths in the second. Alternatively, a FATSA index file (fasta.fai) may be provided, which also includes these columns.	Required
directory	Directory containing JSON output from HeteroGenesis_vargen.py and where output files will be written to.	"./"
chromosomes	List of chromosomes included in the model.	”all”

FreqCalc

-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.

Technical Notes

1. Variants are limited by the following rules in order to reduce complexity of the program:
    - On the same copy of a chromosome whithin a clone:
        - CNVs and deletion InDels cannot partially overlap on the same copy of a chromosome in a clone. However, fully overlapping on the same chromosome, or partially overlapping on different copies of a chromosome, can occur.
        - No variants can occur within a deleted region, even if there are additional copies of the region that haven't been deleted. A CNV deletion may occur over a previous variant but that variant will be ignored when writing output files if all copies have been deleted.
        - InDel deletions can not be placed over a region if any copy of that region on the same chromosome (i.e. from a CNV) contains an SNV or another InDel.
2. Insertion sequences are taken from copying random locations in the genome.
3. CNV lengths are >50 bases and InDel lengths are ≤50 bases.

4. The copy number output does not take into account deletions from InDels (i.e. deletions ≤50 bases).

5. Chromosomes are selected for variant placement at random while taking into account length (of the original reference chromosome), with the exception of aneuploid events, for which all chromosomes are selected with equal probabilities. After an aneuploid replication event has occurred, additional copies of the original chromosome, containing the same set of existing variants, are then available for further variants to be added to.