HeteroGenesis Workflow

heterogenesis_vargen

heterogenesis_vargen is used to generate the lists of variants for incorporation into the germline and each somatic clone's genomes.

Workflow

1. Chromosome lengths for the included chromosomes are taken from the genome index file.

2. Lists of randomly selected SNVs and InDels from the dbSNP minor allele frequency (MAF) file are generated. The two types of variants are first separated into two lists and half of the total number of each list are randomly sampled (without replacement), using the MAFs (normalised to 1) as probabilities of selecting each variant; This should be far more than enough needed for incorporation into the genomes but allows for some to be dismissed due to being in deleted regions.

3. The numbers of each type of variant to occur for the first time in each somatic clone is determined by randomly selecting a somatic clone for each variant, using the given evolutionary distances from parent clones (normalised to 1) as probabilities of selecting each clone. (Numbers of each type of germline variant are provided separately by the user.)

4. The orders of each type of variant in each clone (herein clone also refers to the germline 'clone') is determined by shuffling lists of the numbers of each variant type. E.g. if a clone needs to contain 3 SNVs, 1 replication CNV, 2 deletion CNVs, 2 InDels, and 1 aneuploid event, then the following list of variant types would be generated: "snv, snv, snv, repcnv, delcnv, delcnv, indel, indel, aneuploid" --> "indel, snv, delcnv, aneuploid, repcnv, snv, indel, snv, snv, delcnv"

5. Variants are then generated for each clone. First the variants and chromosome copies of the parent clone are copied over (with the exception of the germline 'clone', which starts with no variants and 2 copies of each chromosome - 'A' and 'B'). The list of variant is then extended by generating aditional variants for each item in the variant type list:

    - SNVs

        - A chromosome is chosen at random, with probabilities that take into account the (reference) chromosome lengths and number of copies of each chromosome. A chromosome copy and position is then also chosen at random.
        - The alternate allele is randomly chosen from the three possible alternative nucleotides.
        - If the SNV is for the germline, the SNV may be taken from the list of dbSNP SNVs, with the probability of this happening provided by the user, instead of randomly generating the SNV.

        - SNVs are discarded (and another one created) if:
            - the reference base is an 'N'.
            - on the the same copy of the chromosome, an existing SNV or InDel is loated at the same position.
            - on the the same copy of the chromosome, a deletion has occured over that base (even if there are additional undeleted copies from a previous CNV).

    - InDels

        - A chromosome is chosen at random, with probabilities that take into account the (reference) chromosome lengths and number of copies of each chromosome. A chromosome copy and position (for the base imediately preceeding the InDel) is then also chosen at random.
        - The length is chosen from a scaled lognormal distribution, where parameters for the underlying normal distribution and a scaling factor are provided by the user. Lengths are limited to ≤50 bases.
        - Insertions or deletions are chosen with equal probabilities.
        - Insertion sequences are chosen at random from elsewhere in the genome.
        - If the InDel is for the germline, the InDel may be taken from the list of dbSNP InDels, with the probability of this happening provided by the user, instead of randomly generating the InDel.
        - Indels are discarded (and another one created) if:
            - the preceding reference base is an 'N'.
            - either the inserted or deleted sequence contains more than 1/4 'N's.
            - on the the same copy of the chromosome, an existing SNV or InDel is loated at the same position.
            - on the the same copy of the chromosome, a deletion has occured over that base (even if there are additional undeleted copies from a previous CNV).
            - the InDel is a deletion and the deleted region overlaps a CNV breakpoint, or existing SNV or InDel.

    - CNVs

        - A chromosome is chosen at random, with probabilities that take into account the (reference) chromosome lengths and number of copies of each chromosome. A chromosome copy and starting position is then also chosen at random.
        - The length is chosen from a scaled lognormal distribution, where parameters for the underlying normal distribution and a scaling factor are provided by the user. Lengths are limited to >50 bases.
        - If the CNV is a replication, the copy number is chosen from a lognormal distribution (limited to >1), where parameters for the underlying normal distribution are provided by the user.
        - CNVs are discarded (and another one created) if:
            - the CNV partially overlaps another CNV on the same copy of the chromosome.
            - the CNV is in a deleted region.

- Aneuploid events
    - A chromosome and copy is chosen at random, irrespective of chromosome lengths.
    - The copy number is chosen at random from [0,2,3].
    - If a chromosome is replicated, all variants from the original are copied to the new chromosomes, which are also then available for placing new variants on.
    - Aneuploid histories are recorded in the chromosome names. E.g. If chr1A undergoes a duplication, chr1A is removed and chr1A-1 and chr1A-2 are added. If chr1A-2 then undergoes a triplication, chr1A-2 is removed and chr1A-2-1, chr1A-2-2 and chr1A-2-3 are added.

6. Lists of variants and the orders they occur in each clone are written to files.

heterogenesis_varincorp

heterogenesis_varincorp is run separately for each clone. It takes the lists of variants generated by heterogenesis_vargen and incorporates them into a reference genome sequence, as well as calculating copy numbers and variant frequencies along the genome.

Workflow

1. The lists of variants from heterogenesis_vargen is read in.
2. For each copy of each chromosome in the clone, the variants that are located on it are sequentially used to update several items:
    - **cnblocks**: This is a list of regions (or blocks) with, start and end positions, that is used to calculate copy numbers along a chromosome. The list starts with one block equal to the length of the chromosome and is updated for CNVs by splitting blocks at CNV breakpoints and either replicating all blocks included within the CNV (if the CNV is a replication) or removing them (if the CNV is a deletion). If multiple copies of a region covering the new CNV position exist in cnblocks, from a previous CNV, then the new CNV is incorporated only into the first copy. After all variants have been dealt with, the cnblocks lists for all copies of a chromosome are combined and any blocks that overlap breakpoints on any of the other copies are again split. The number of blocks that correspond to each region on the chromosome is then added up to give the copy number status for each region, which is then written to file.
    - **allblocks**: This is a list of blocks, similar to cnblocks, that acts as a blueprint for generating the final sequence, for a copy of a chromosome, from the reference sequence. Unlike cnblocks, it also includes blocks corresponding to SNVs/InDel insertions, and is also updated by InDel deletions - these were not included in cnblocks as it wouldn't be useful to have a diferent copy number status recorded everytime an InDel deletion (≤50 bases) occurs. As with the CNVs in cnblocks, if multiple copies of a region covering a variant position exist in allblocks, from a previous CNV, then the new variant is incorporated only into the first copy. Variants are used to

update allblocks as follows:

- **CNVs**: As with cnblocks, CNVs result in splitting, replication or removal of blocks.
- **InDel deletions**: These are treated as CNV deletions and result in blocks being split at the start and end of the deletion, with the resulting middle block(s) being removed.
- **SNVs and InDel insertions**: These variants become new single base blocks, with start positions the same as the end positions, and with the alternate allele sequence (including the preceeding base for InDels) recorded for the block. Blocks in the allblocks list are split immediately before and after the variant position and the resulting middle block is swapped with the variant block.
- **Aneuploid event**: This doesnt require any action as aneuploid events were taken into account when creating the variant lists for copies of chromosomes.

- **vcfcounts**: This is a list of SNVs and InDels, with the number of occurences on the chromosome recorded for each. Each SNV/InDel also has information recorded on the position of CNVs that overlap them. This allows us to know how many occurences of an SNV/InDel to replicate/remove based on whether a new CNV falls within or around a previous CNV. Variants are used to update vcfcounts as follows:

  - **SNVs/InDels**: These are added to vcfcounts with a starting copy of 1.

  - **CNVs**: For each SNV/InDel that lies within the CNV, a recursive function goes through each level of previous CNVs over the same position, determines where the new CNV fits in, and replicates/removes the number of occurences recorded within that level. As with the cnblocks and allblocks lists, CNVs affect only the first copy of the region they're within.

  Once all variants have been dealt with, the vcfcounts lists for all copies of a chromosome in the clone are combined, with shared SNV/InDels' numbers of occurences added together. The overall copy number at each SNV/InDel position is taken from the combined cnblocks list and used with the number of occurences to calculate allele frequeincies. These are then written to file.

3. The genome sequence for each copy of each chromosome is generated using the allblocks lists and reference genome. For each block in allblocks, the genome sequence is extended with either the corresponding reference sequence at the given positions or, in the case of blocks from SNVs/InDel insertions, the alternate allele sequence. The sequence is then written to file in FASTA format.