Gretel Documentation

Release 0.0.1a

Sam Nicholls

Dec 18, 2020

Contents

An algorithm for recovering haplotypes from metagenomes. Sister to Hansel.

CHAPTER 1

Protocol

Gretel provides a command line tool for the recovery of haplotypes. We recommend the following protocol.

1.1 Read Alignment

Gretel requires your reads to be aligned to a common reference. This is to ensure that reads share a co-ordinate system, on which we can call for variants and recover haplotypes. The reference itself is of little consequence, though dropped reads will lead to evidence to be unavailable to Gretel.

Construction of a *de novo* consensus assembly for a metagenome is left as an exercise for the reader. Our lab has traditionally been using *velvet*, but recommendations have led me to find *Ray*.

We used *bowtie2* during our experiments. We increased its sensitivity with the following parameters to increase alignment rates:

```
bowtie2 --local -D 20 -R 3 -L 3 -N 1 -p 8 --gbar 1 --mp 3
```

See the blog post **'bowtie2: Relaxed Parameters for Generous Alignments to Metagenomes <https://samnicholls.net/2016/12/24/bowtie2-metagenomes/'_** for more information.

Sort and index the alignment.

1.2 Variant Calling

Gretel is robust to sequencing error and misalignment noise, thus the calling of variants need not be carefully conducted. Typically we have used *samtools*, but for our own Gretel pipeline, we have aggressively called all heterogenous sites in an alignment as a SNP using the *snpper* tool in our gretel-test repository.

For somewhat questionable reasoning, we currently require a compressed and indexed VCF:

```
bgzip <my.vcf>
tabix <my.vcf.gz>
```

1.3 Invocation of Gretel

As described in the README, Gretel is invoked as follows:

```
gretel <my.sort.bam> <my.vcf.gz> <contig> [-s 1startpos] [-e 1endpos] [--master␣
↪master.fa] [-o output_dir]
```

You must provide your sorted BAM, compressed VCF, and the name of the contig on which to recover haplotypes. Use *-s* and *-e* to specify the positions on the aligned reads between which to recover haplotypes from your metagenome.

By default, Gretel will output a FASTA containing the recovered SNPs, in order, for each haplotype. Providing an optional "master" FASTA sequence will permit Gretel to "fill in" the non-SNP positions (*i.e.* the positions between *-s* and *-e* that do not appear in the VCF) with the nucleotide from the pseudo-reference.

1.4 Gretel Outputs

1.4.1 out.fasta

A **FASTA** containing each of the recovered sequences, in the order they were found. Each sequence is named <*iteration*>__-<*log10 likelihood*>. Sequences are not wrapped.

1.4.2 gretel.crumbs

Additionally, Gretel outputs a whimsically named *crumbs* file, containing some potentially interesting metadata, as well as a record of each recovered haplotype. The first row is a comment containing the following (in order):

- The number of SNPs across the region of interest
- Unused (currently)
- Unused (currently)
- The suggested value of L for the L'th order Markov chain used to reconstruct haplotypes
- The chosen value of L for the L'th order Markov chain
- The average likelihood of the returned haplotypes given the state of the Hansel matrix at the time the haplotypes were each recovered
- The average likelihood of the returned haplotypes given the state of the Hansel matrix at the time the reads were parsed
- The average number of observations removed from the Hansel matrix by the reweighting mechanism

The rest of the file contains tab-delimited metadata for each recovered haplotype:

- The iteration number, starting from 0
- The *weighted* likelihood of the haplotype, given the Hansel matrix at the time the haplotype was recovered
- The *unweighted* likelihood of the haplotype, given the Hansel matrix at the time the reads were parsed

In practice, we rank with the **weighted** likelihoods to discern the haplotypes most likely to exist in the metagenome. One may attempt to use the *unweighted* likelihoods as a means to compare the abundance, or read support, **between the returned haplotypes** (*i.e.* not necessarily the metagenome as a whole).

CHAPTER 2

History

2.1 0.0.94

- Added *–pepper* option to for permissive pileups by overriding the pysam pileup stepper to *all* instead of *samtools*.

2.2 0.0.93

- Move *process_vcf* to *util* module. I may drop use of *pyvcf* in future as I don't like the API.
- Dropped pointless *append_path* stub.
- Fixed an edge case where reads beginning with a SNP that aligned to the start of a parallel parsing window are counted twice.
- Added a small test package to help detect future regressions.
- Added *–version* argument to print program version number.
- Removed *–lorder* argument as users should not need to select the chain order.

2.3 0.0.92

- Adds *–dumpmatrix* and *–dumpsnps* debugging options.
- Clean up Hansel matrix initialisation.
- Add *gretel-snpper* command for generating naive VCF.
- Fix a regression where the *L* parameter of the matrix is incorrectly left unset.

2.4 0.0.90

Resolves a bug whereby SNPs are incorrectly parsed from the BAM if either:

- its quality score is below 13
- the read is overlapped by its primary mate

Well covered data sets need not be overly affected by the additional noise that may have been introduced, but the problem is more noticeable with low coverage and you may wish to reapply Gretel to affected data. Sorry.

2.5 0.0.81

- Add warning and advice when an entry in Hansel is missing evidence.
- Make the 'Unable to select' warning sound much less bad because it is normal.

2.6 0.0.8

- Docs
- Deprecate *gretel-crumbs* command

2.7 0.0.7

- Further improvements to parallel read processing
- Add - symbol to enable support for deletions

2.8 0.0.6b

- Fix setting of *L* parameter

2.9 0.0.6

- MULTIPROCESSING
- Re-write read handling, again

2.10 0.0.5

- *-s* and *-e* introduced to allow specification of positions between which to recover haplotypes
- Attempt some basic indel handling
- Fix a bug where the master sequence was altered by the output of each reported haplotype

2.11 0.0.4

- Add experimental *–sentinels* option
- Improve docs

2.12 0.0.3

- Hansel is now seperate from Gretel
- [Hansel] *get_marginal_at* is *now get_counts_at*
- [Hansel] *selext_next_edge_at* deprecated
- Gene recovery and likelihood plots are now on seperate panels
- Re-write methods to add observations to matrix to be less awful to read
- Drop *–hit* and *–gene* options to verification
- Replace verification script to *gretel-crumbs* command

2.13 0.0.2

- Improve documentation.
- Provide *util* subpackage for filling *Hansel* structure with BAM observations.
- Explicitly provide possible symbols to *Hansel*.
- Improve plotting
- Remove *process_hits* and *process_refs* as these are no longer needed.
- Rename *establish_path* to *generate_path*
- Rename *add_ignore_support3* to *reweight_hansel_from_graph* so we have some sort of indication of what it does.
- Altered Sphinx configuation.

2.14 0.0.1

- Import repository from *claw*.

CHAPTER 3

Indices and tables

- genindex
- modindex
- search