Gretel Documentation

Release 0.0.1a

Sam Nicholls

Mar 11, 2017

Contents

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

CHAPTER 1

Gretel

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

What is it?

Gretel is a Python package providing a command line tool for the recovery of haplotypes from metagenomic data sets. **Gretel** parses an alignment of reads into a **Hansel** matrix and uses the evidence of SNP pairs observed to appear on the same reads to probabilistically reconstruct the most likely haplotypes.

Gretel uses an L'th order Markov chain model to reconstruct likely sequences of variants that constitute haplotypes in the real metagenome. Our approach involves graph-like traversal of the data within the **Hansel** matrix. Edges are probabilitically weighted based on the evidence on the reads, as well as the haplotype as it has been reconstructed so far.

What can I use it for?

Gretel is designed to recover haplotypes from your data set, without the need for setting (or optimisation) of any parameters. **Gretel** does not require a priori knowledge of your input data (such as its contents, or the true number of haplotypes) and makes no assumptions regarding the distributions of alleles at variant sites and uses the available evidence from the aligned reads without altering or discarding the observed varations.

Why should I use it?

Gretel is the first tool capable of recovering haplotypes from metagenomes. Whilst tools exist for analogous haplotyping problems, such as the assembly of viral quasispecies, typically these tools rely on overlap approaches that create too many unranked haplotypes. **Gretel** is capable of ranking the haplotypes it outputs by their likelihood.

Gretel requires no parameters and our approach is robust to sequencing error and misalignment noise.

Requirements

```
$ pip install numpy hanselx pysam PyVCF
```

Install

```
$ pip install gretel
```

Usage

You will require a sorted BAM containing your reads, aligned to some pseudo-reference. You can use any sequence as your reference, such as a consensus assembly of the metagenomic reads, or a known strain reference (such as HIV-1). You must bgzip and tabix your VCF.

```
$ gretel <bam> <vcf.gz> <contig> -s <1-start> -e <1-end> --master <master.fa> -o
↪<outdir>
```

Citation

Paper pending...

License

Hansel and Gretel are distributed under the MIT license, see LICENSE.

CHAPTER 2

Protocol

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

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.

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>
```

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.

Gretel Outputs

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.

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 3

History

0.0.2-wip

- 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.

0.0.1

- Import repository from *claw*.

CHAPTER 4

Indices and tables

- genindex
- modindex
- search