CallHap: A Pipeline for Analysis of Pooled Whole-genome Haplotypes

Last edited: 04/06/2017 By: Brendan F Kohrn

**Licensing information**

With the exception of the Genome Analysis Toolkit, all programs are freely available under either the Gnu Public License or the MIT License. The Genome Analysis Toolkit is free for non-commercial use; other use should contact the Broad Institute at softwarelicensing@broadinstitute.org. Python and bash scripts for the CallHap pipeline are available at https://github.com/cruzan-lab/CallHap.

**Introduction**

CallHap is a pipeline designed for the analysis of pooled haplotype data. It depends on the presence of two types of sequencing libraries; either single sample libraries (SSLs) or pooled libraries (Pool). Ideally, a Pool should contain equimolar genetic material from 20 individuals, and one of those individuals should be prepared separately as a SSL. This pipeline picks up following sequencing on an Illumina HiSeq or similar high-throughput sequencer.

**Requirements**

* A LINUX/UNIX/MacOS system with the following programs installed:
  + Cutadapt (http://cutadapt.readthedocs.io/en/stable/index.html)
  + Sickle (http://bioinformatics.ucdavis.edu/research-computing/software/)
  + BWA (http://bio-bwa.sourceforge.net/)
  + Samtools (http://samtools.sourceforge.net/)
  + PicardTools (https://broadinstitute.github.io/picard/)
  + GATK (https://software.broadinstitute.org/gatk/)
  + Freebayes (<https://github.com/ekg/freebayes>)
  + Python 2.7x (<https://www.python.org/>) with NumPy (<http://www.numpy.org/>)
  + Java Development Kit

**Quick Start:**

**Setup:**

program-config.sh:

Edit program-config.sh so that each of the variables is set to the absolute path of the program in question (basically, whatever you would type to start the program).

Reference Preparation:

Obtain a reference genome (in FASTA format) for your species of interest (or closely related other species), and prepare it for use by using the following commands:

$ bwa index {reference}.fasta

$ samtools faidx {reference}.fasta

$ java -jar /path/to/picardtools/picard.jar \

CreateSequenceDictionary \

R={reference}.fasta \

O={reference}.dict

**Preprocessing:**

Note that there are two basic processing pipelines provided; one with automated trimming (CallHap\_Preproc\_0.01.23.sh) and one without automated trimming (CallHap\_Preproc\_NoTrimming\_0.01.23.sh). It is strongly suggested that at least a few (2-5) samples per flow cell be run manually (one step at a time), at least through trimming for quality control and to see if those samples need any additional trimming beyond the basic trimming steps (adapter and quality trimming). If you are doing trimming separately, be sure to use the locations of the trimmed files in the preconfig instead of the locations of the raw files.

Create a preconfig file in Excel with the following columns:

* Read1File
* Read2File
* RGLB
* RGSM
* RGPU
* Mode
* Reference

Each row should represent one sequencing library (SSL or Pooled). Read1File and Read2File should give the absolute path to the locations of the raw data for the Read 1 and Read 2 files (in the case of single end data, give the file location under Read1File, and put a period (.) for Read2File). RGLB should be some identifier for the library (e.g. library number). RGSM should be a sample name, preferably indicating the species of the library, the location the sample came from, and whether the sample is a SSL or Pool (Example: SpenamLocS#SSL, SpenamLocS#Pool). RGPU should indicate the barcoding used for this library during library prep (Example: ATTACTCG-TATAGCCT). Mode should be one of se (single-end) or pe (paired-end). Finally, Reference should indicate the reference genome you would like this library aligned to. If all samples are of the same species, the reference genomes for all libraries should be the same.

Save the preconfig file as a .csv. Convert it to a config file using:

$ python /path/to/CallHap/CallHap\_ConfigCreator.py \

--input preconfig.csv \

--adapt1 {SequencingAdapter} \

--adapt2 {SequencingAdapter} \

--sequencer {Sequencer used to produce data} \

--minBaseQual 30 \

--minReadQual 30 \

--runID {Identifier for this run}

This will output a .sh file with the run ID as the name (for example, of you put --runID runID, the file would be called runID.sh)

Then use the following command to run the rest of the pre-processing (replacing the script name if you did trimming separately):

$ bash /path/to/CallHap/CallHap\_Preproc\_0.01.23.sh \

program-config.sh {runID}.sh

**SNP Calling:**

Set up an input list of files using:

$ ls -1 /path/to/files/\*SSL\*.rg.ra.bam > {RunID}.txt

$ ls -1 /path/to/files/\*Pool\*.rg.ra.bam >> {RunID}.txt

Call FreeBayes using:

$ /path/to/freebayes/freebayes -L {RunID}.txt \

-p 1 -f /path/to/reference/{reference}.fasta \

-v {RunID}\_SNPs.vcf --use-best-n-alleles 2 \

--min-repeat-entropy 1 --no-partial-observations \  
--min-alternate-fraction 0.05

--min-alternate-fraction should be set to 1/poolsize or lower. This step may take a while, and while running, may look like it isn’t doing anything.

**SNP Filtering:**

SNP filtering is accomplished by use of a custom python script, which can be run with the command:

$ python /path/to/CallHap/CallHap\_VCF\_Filt.py \

-i {RunID}\_SNPs.vcf -o {RunID}\_d600q20\_Haps.vcf \

-O {RunID}\_d600q20\_Pools.vcf -n <number of SSLs> \

-N <number of Pools> -d 600 -q 20 -p 20

You may need to trim off one or more columns from the VCF file if one sample was not called at a majority of positions; if a single sample is not called at a particular position, the variant at that position will be discarded. To determine if a column needs to be removed, look at your VCF file in Excel, and see if there are any columns that are periods (“.”) for the majority of rows. Removing the column can also be done in Excel, but you need to be careful because Excel likes to add quotes when it saves files with commas in the cells, as do most spreadsheet editors I’ve found.

**Haplotype Calling:**

Before running this step, check how many cores are available on the system you’re using with htop. Make sure you don’t overload the system you’re working on; don’t set -t to higher than the number of available cores, and don’t take up all the cores on the machine.

Haplotype calling can be run using:

$ python /path/to/CallHap/CallHap\_HapCallr.py \

--inputHaps {RunID}\_d250q20\_Haps.vcf \

--inputFreqs {RunID}\_d250q20\_Pools.vcf \

-o {RunID} -p 20 -t 5 -l 2 --numRandom 100 –numTopRSS 3

This program generates four to six output file per solution output (within the minimum number of RSS values):

A NEXUS file (RunID\_solNum\_haps.nex) for network phylogeny creation; PopART (<http://popart.otago.ac.nz/index.shtml>) works fairly well. I’ve been using the TCS algorithm.

A VCF file (RunID\_solNum\_PredFreqs.vcf) containing the estimated SNP frequencies based on the estimated haplotype frequencies, and the per-SNP average residuals in the INFO field

A CSV file (RunID\_solNum\_freqs.csv) containing the per-pool haplotype frequencies and RSS values for each pool.

A CSV file (RunID\_solNum\_Regression.csv) containing paired observed and predicted SNP frequencies from the Least Squared algorithm.

(Optional): A Structure formatted file (RunID\_solNum\_iterNum.str) containing the expanded haplotype frequencies

(Optional): A Genpop file containing the haplotype frequencies for use in Adigenet.

In addition, outputs are generated describing the original haplotypes network (RunID\_Initial.nex), the unique haplotypes network (RunID\_Unique.nex), raw topologies observed from each random order (RunID\_RAW.csv), the frequency of each unique topology generated (RunID\_topologies.csv), the frequency of occurrence for each haplotype found in any random order (RunID\_summary.csv).

In terms of population-genetics analysis, haplotypes should be treated as independent alleles at a single locus.

**Detailed Instructions:**

**Adapter/Quality trimming:**

Adapter and quality trimming should be performed before any other step in the pipeline. This ensures better read alignment and higher quality of the final data. The automated pipeline uses cutadapt for adapter trimming and sickle for quality trimming; however, you can use other trimming programs if so desired.

Cutadapt is available at <http://cutadapt.readthedocs.io/en/stable/index.html> under the MIT License and can be run using:

$ /path/to/cutadapt -a {inAdapter1} -A {inAdapter2} \

-o {output\_read\_1}\_at.fastq.gz \

-p {output\_read\_2}\_at.fastq.gz \

{input\_read\_1}.fastq.gz {input\_read\_2}.fastq.gz

for paired-end reads or

$ /path/to/cutadapt -a {inAdapter1} \

-o {output\_read\_1}\_at.fastq.gz {input\_read\_1}.fastq.gz

for single-end reads.

If you aren’t certain what adapter sequence you have, running FastQC (freely available at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> under GPLv3) may help determine what adapters are present. Otherwise, consult your library preparation protocol.

While cutadapt can also do quality trimming (using the -q option), or remove a fixed number of bases (using the -u option), the default pipeline uses a second program, (sickle) for quality trimming. Sickle is available at <http://bioinformatics.ucdavis.edu/research-computing/software/> under the MIT License and can be run with

$ /path/to/sickle pe -f {output\_read\_1}\_at.fastq.gz \

-r {output\_read\_2}\_at.fastq.gz -o {output\_read\_1}ut\_at\_qt.fastq.gz -p {output\_read\_2}\_at\_qt.fastq.gz -t sanger -s {SampleName}\_extras.fastq.gz -q {minBaseQuality} -g

for paired-end reads or

$ /path/to/sickle se -f {output\_read\_1}\_at.fastq.gz \

-o {output\_read\_1}\_at\_qt.fastq.gz -t sanger \

-q {minBaseQuality} -g

for single-end reads.

For more details on these programs, consult their respective manuals.

Following trimming, it is recommended that at least 2-5 samples per flow cell be quality-checked using FastQC. For this pipeline, check that there are almost no remaining adapters of any type in the AdapterContent page of the report and that you are satisfied with the quality scores in the Per base sequence quality section and the base percentages in the Per base sequence content section.

Note that FastQC will generate output files in the same directory as the input files.

**Read alignment:**

The automated pipeline uses BWA-mem to align reads with default options. BWA can be obtained from <http://bio-bwa.sourceforge.net/> under GPLv3, and run using

$ /path/to/bwa mem -M {reference}.fasta \

{output\_read\_1}\_at\_qt.fastq.gz \

{output\_read\_2}\_at\_qt.fastq.gz > \

{SampleName}.sam

For paired-end reads and

$ /path/to/bwa mem -M {reference}.fasta \

{output\_read\_1}\_at\_qt.fastq.gz > {SampleName}.sam

After alignment, the file is converted to a bam file using

$ /path/to/samtools view -Sbu -F 4 {SampleName}.sam | \

/path/to/samtools sort – {SampleName}.sort

and indexed using:

$ /path/to/samtools index {SampleName}.sort.bam

At this time, any unaligned reads are also removed. Samtools can be obtained from <http://www.htslib.org/>.

**Readgroup Creation:**

PicardTools is used to add readgroups to the files. These are a requirement for local realignment with GATK, and for SNP calling with FreeBayes. For later analysis, it is useful if each sample have a different sample name (RGSM) and readgroup ID (RGID), since Freebayes (our SNP caller) uses the readgroup ID to differentiate samples. I used the library number as the readgroup ID.

PicardTools is available at <https://broadinstitute.github.io/picard/>, and can be run using

$ java -jar /path/to/picard AddOrReplaceReadGroups \

INPUT={SampleName}.sort.bam \

OUTPUT={SampleName}.sort.rg.bam \

RGID={ReadGroupID} \

RGLB={ReadGroupLibrary} \

RBPL={ReadGroupSequencingPlatform} \

RGPU={ReadGroupRunBarcode} \

RGSM={ReadGroupSampleName} \

CREATE\_INDEX=true

RGLB, RBPL, RGPU, and RGSM are required for this tool to run, and RGID needs to be different for each library.

**Local Realignment:**

Local realignment is carried out using the Genome Analysis Toolkit (GATK, available at <https://software.broadinstitute.org/gatk/>). The first step in this process is to locate targets for local realignment using:

$ java -jar /path/to/GATK -T RealignerTargetCreator \

-R {reference}.fasta \

-I {SampleName}.sort.rg.bam \

-o {SampleName}.sort.rg.intervals

Following this, local realignment can be run using:

$ java -jar /path/to/GATK -T IndelRealigner \

-R {reference}.fasta \

-I {SampleName}.sort.rg.bam \

-targetIntervals {SampleName}.sort.rg.intervals \

-o {SampleName}.sort.rg.ra.bam \

-dt NONE \

--maxReadsForRealignment 200000

**SNP Calling:**

Set up an input list of files using:

$ ls -1 /path/to/files/\*SSL\*.rg.ra.bam > {RunID}.txt

$ ls -1 /path/to/files/\*Pool\*.rg.ra.bam >> {RunID}.txt

Or whatever identifier you used to differentiate PLs and SSLs. The important thing is that this file list all SSLs, followed by all PLs.

Call FreeBayes using:

$ /path/to/freebayes/freebayes -L {RunID}.txt \

-p 1 -f /path/to/reference/{reference}.fasta \

-v {RunID}\_SNPs.vcf --use-best-n-alleles 2 \

--min-repeat-entropy 1 --no-partial-observations \  
--min-alternate-fraction 0.05

--min-alternate-fraction should be set to 1/poolsize or lower. This step may take a while, and while running, may look like it isn’t doing anything.

This step may take a while, and while running, may look like it isn’t doing anything.

FreeBayes can be found at <https://github.com/ekg/freebayes>.

**SNP Filtering:**

Before running SNP filtering, it may be necessary to trim off one or more columns from the VCF file if one sample was not called at a majority of positions; if a single sample is not called at a particular position, the variant at that position will be discarded, so a single sample uncalled (or at low depth) at a majority of positions can result in no data making it through the filtering step. To determine if a column needs to be removed, look at your VCF file in Excel, and see if there are any columns that are periods (“.”) for the majority of rows. Removing the column can also be done in Excel, but you need to be careful because Excel likes to add quotes when it saves files with commas in the cells, as do most spreadsheet editors I’ve found.

If desired, sample depth can be assessed using the GATK DepthOfCoverage tool (see <https://software.broadinstitute.org/gatk/documentation/tooldocs/org_broadinstitute_gatk_tools_walkers_coverage_DepthOfCoverage.php> for instructions). This tool takes a similar amount of time to SNP calling.

SNP filtering is accomplished by use of a custom python script, which can be run with the command:

$ python /path/to/CallHap/CallHap\_VCF\_Filt.py \

-i {RunID}\_SNPs.vcf -o {RunID}\_d600q20\_Haps.vcf \

-O {RunID}\_d600q20\_Pools.vcf -n {number of SSLs} \

-N {number of Pools} -d 600 -q 20 -p 20

-i is the input VCF file from FreeBayes

-o is the output haplotypes file, containing haplotypes found in the SSLs

-O is the output Pool SNP frequencies file, containing frequency of the more common allele in each pool

-n is the number of SSLs in the input file

-N is the number of Pools in the input file

-d, --minDepth: This option sets the minimum read depth that must be present at a position in ALL libraries in order for that position to be considered as a variant. It should be set based on the number of individuals in a PL. For haploid sequence, a depth of 15 per individual in the pool is recommended (Sims et al., 2014), so that for a pool of 20 individuals, a depth of 300 is required at a site to be able to call variants.

-q, --minQual: Controls the minimum PHRED-scaled variant quality needed to use a variant. Mostly useful for filtering out super-low quality variants, but can be set higher as necessary. -p is the number of individuals in each pool.

--minCallPrev: Controls the maximum allowable error in SSLs for a variant to be called. It can range from 1 (all reads in each SSL need to have the same identity) to 0.5 (Up to half the reads in a SSL can have a different identity). At a setting of 1, some real SNPs could be removed based on unavoidable errors in the SSLs, while at a setting of 0.5, confidence in the identity call for SSLs, and thus in the identity of haplotypes, will be significantly decreased. I set this parameter at a default of 0.9, to allow for some sequencing error in the SSLs while still maintaining a high accuracy of SSL identity calls.

--minSnpPrev: Coupled with poolSize, this option controls how much of a PL must be the alternate identity for a SNP to be at that position when there is no variation in the SSLs. The value is a positive floating-point decimal, which gets multiplied by 1/poolSize to yield the proportion of reads that must be of a different identity in a PL to yield a variant. At a value of zero, all positions would be called as variants if there was any variation in a PL. I set this at a default value of 0.75 in order to allow for some error in low-frequency haplotypes, while removing the majority of low-frequency sequencing errors from consideration.

--indelDist: How far away from indels a variant should be for use. IndelDist takes an integer value greater than 0; at a value of 0, distance from an indel will not be considered as a filter. I set this at a relatively conservative value of 100 (the length of my raw sequencing reads) as being the maximum distance at which the presence of an indel could have any effect on variant discovery.

You may want to try running this program with different sets of parameters to determine what your optimum parameters will be for a particular run.

**Haplotype Calling:**

Before running this step, check how many cores are available on the system you’re using with htop. Make sure you don’t overload the system you’re working on; don’t set -t to higher than the number of available cores, and don’t take up all the cores on the machine.

Haplotype calling can be run using:

$ python /path/to/CallHap/CallHap\_HapCallr.py \

--inputHaps {RunID}\_d600q20\_Haps.vcf \

--inputFreqs {RunID}\_d600q20\_Pools.vcf \

-o {RunID} -p 20 -t 5 -l 2 --numRandom 100 –numTopRSS 3

Where

--inputHaps is the haplotypes file from SNP filtering

--inputFreqs is the Pools file from SNP filtering

-o is a unique output prefix for this run of haplotype caller

-p is the number of individuals in each pool

-t is the number of threads to use during processing

-l is the number of times to iterate across the SNPs within each order

-r denotes how high a residual should be able to exist after adding a SNP, and is used to defer processing of a SNP where the residual doesn’t reduce enough to another iteration

--dropFinal is a flag which pairs with -r to remove SNPs with a high residual entirely at the end if they don’t reduce the residual enough. May not work with current random ordering algorithm; don’t use for now.

--genpop is a flag that instructs CallHap to generate genpop output

--structure is a flag that instructs CallHap to generate structure formated output

--numRandom: Controls how many psudo-random orderings of SNPs to use, and should be a value greater than zero. I set this value at 100 as a compromise between run time and increased chance of finding the correct solution; in practice, this value should be set based on the number of starting haplotypes relative to the number of SNPs present. If the number of starting haplotypes is close to the number of SNPs, this value can be low; the maximum number of haplotypes in the network is one more than the number of SNPs. However, if the number of SNPs is greater than the number of haplotypes, more attempts may be needed to help resolve the best network topology.

--numTopRSS: This option just influences how many RSS values down are processed for the final output solutions, and should be an integer greater than zero. I set it at a value of 3 so I could examine the higher RSS value solutions.

This program generates four to six output file per solution output (within the minimum number of RSS values):

A NEXUS file (RunID\_solNum\_haps.nex) for network phylogeny creation; PopART (<http://popart.otago.ac.nz/index.shtml>) works fairly well. I’ve been using the TCS algorithm.

A VCF file (RunID\_solNum\_PredFreqs.vcf) containing the estimated SNP frequencies based on the estimated haplotype frequencies, and the per-SNP average residuals in the INFO field

A CSV file (RunID\_solNum\_freqs.csv) containing the per-pool haplotype frequencies and RSS values for each pool.

A CSV file (RunID\_solNum\_Regression.csv) containing paired observed and predicted SNP frequencies from the Least Squared algorithm.

(Optional): A Structure formatted file (RunID\_solNum\_iterNum.str) containing the expanded haplotype frequencies

(Optional): A Genpop file containing the haplotype frequencies for use in Adigenet.

In addition, outputs are generated describing the original haplotypes network (RunID\_Initial.nex), the unique haplotypes network (RunID\_Unique.nex), raw topologies observed from each random order (RunID\_RAW.csv), the frequency of each unique topology generated (RunID\_topologies.csv), the frequency of occurrence for each haplotype found in any random order (RunID\_summary.csv).

In terms of population-genetics analysis, haplotypes should be treated as independent alleles at a single locus.

**Common Errors:**

|  |  |
| --- | --- |
| Problem | Solution |
| Quick-start pipeline produces empty files | Check that input files defined in the preconfig exist |
| Multiple best-RSS solutions | If one occurs more frequently than the other, use that one.  If both occur equally, check to see if the network phylogenies for each solution look the same, and if the generated haplotype frequencies look the same. If the generated haplotype frequencies are identical, it doesn’t matter which haplotype is actually present.  If generated haplotype frequencies differ, create non-biologically relevant pools containing the same DNA samples, but shuffled in new ways (perhaps by using individual 1 from each population as one pool, individual 2 from each population as a second, and so on). |
|  |  |