HAPSEG (v2) BETA

(/modules/docs/HAPSEG/2)

> This module is currently in beta release. The module and/or documentation may be incomplete.

A probabilistic method to interpret bi-allelic marker data in cancer samples.

Author: Scott Carter, Matthew Meyerson, Gad Getz

Contact:

- For HAPSEG and ABSOLUTE questions:
  - Use the Biostars forums <https://www.biostars.org/t/hapseg/ (https://www.biostars.org/t/hapseg/)> and <https://www.biostars.org/t/absolute/ (https://www.biostars.org/t/absolute/)>
  - Use the CGA discussion and help forum (http://www.broadinstitute.org/cancer/cga/cga_forums), especially for help with data interpretation.
- BEAGLE questions: contact Brian Browning at browning@uw.edu (mailto:browning@uw.edu?subject=BEAGLE%20in%20context%20of%20HAPSEG)
- For GenePattern site questions, contact gp-help@broadinstitute.org (mailto:gp-help@broadinstitute.org?subject=HAPSEG)

Algorithm Version: HAPSEG 1.1.1

Summary

The HAPSEG module uses (1) Affymetrix SNP6 or SNP250KSTY array data providing allelic genotypes and copy number information, (2) the statistical phasing software BEAGLE, and optionally (3) statistical phasing information from a population haplotype reference, to estimate homologue-specific copy ratios (HSCRs). HAPSEG runs on tumor samples with or without a patient-matched normal sample and produces partially phased haplotypes without phased reference panels given the cancer samples contain allelic imbalance with which to impute phasing. Use of the optional reference haplotype is recommended and improves inference of genotypes by resolving genotypes of adjacent markers and by extension HSCR estimation with the exception of for regions of high recombination and with the assumption that normal haplotype statistics apply to cancer cells. Such phased copy ratios data, which can also be derived from next generation sequencing data, allows resolution of copy-neutral loss of heterozygosity (CN-LOH) events and are preferred in downstream ABSOLUTE analysis over copy ratios data derived from CGH (comparative genomic hybridization), FISH or cytogenetics.

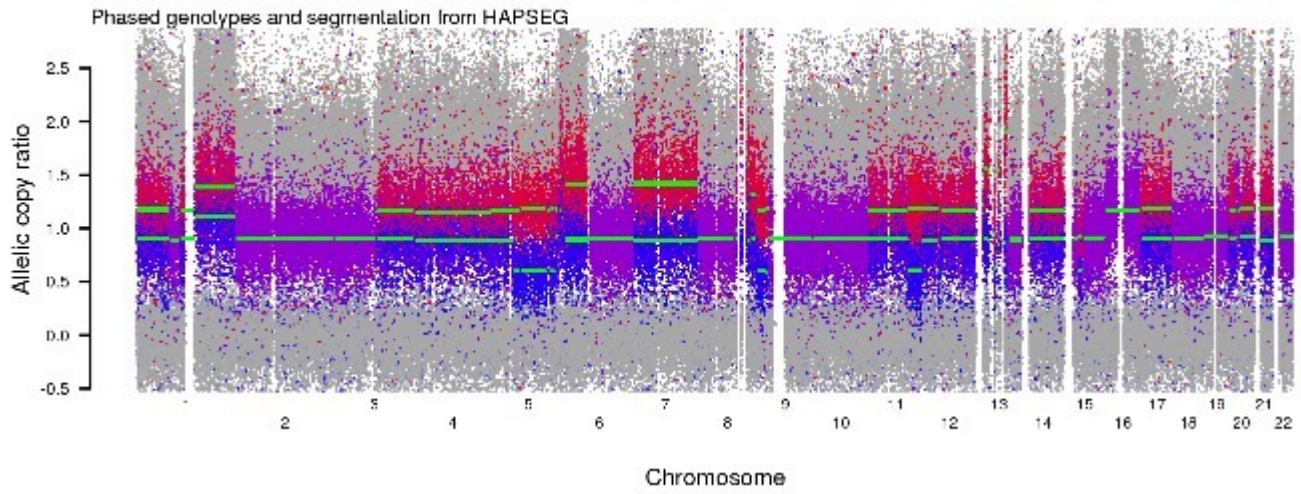
- For an overview of the analysis workflow, see *Using HAPSEG and ABSOLUTE in GenePattern (http://www.broadinstitute.org/using-hapseg-and-absolute-in-genepattern)*. The CGA group provides an example dataset for download (ftp://ftp.broadinstitute.org/pub/genepattern/example_files/HAPSEG_1.1.1/paper_example.zip) to use in the workflow.

- Additional information about HAPSEG is available from *Carter et al.* (http://precedings.nature.com/documents/6494/version/1), the Broad Institute Cancer Genome Analysis (CGA) group website (http://www.broadinstitute.org/cancer/cga/hapseg), and their *How do I run HAPSEG? (http://www.broadinstitute.org/cancer/cga/hapseg_run)* page. The latter discusses algorithm parameters as an R package.

HAPSEG produces the segmented copy ratios data in an RData format suitable for use with the ABSOLUTE module (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00309:1.2). In addition, if selected, HAPSEG provides per-chromosome plots of the fitted segments. These include calibrations of A versus B allele copy-ratios for SNPs using contours that denote error-model fit for each genotype cluster. For contours and lines, black marks homozygous and green marks heterozygous genotypes.

Background

Elucidation of the sequence of the multiple genomic events that give rise to tumorigenesis is an ongoing area of research. Genomic events include functional mutations, genomic rearrangements including translocations, gene conversion or loss of heterozygosity (LOH), and somatic copy number alterations (SCNAs) that range from regional and chromosomal amplifications and deletions to whole genome duplications (Burrell et al. (http://www.nature.com/nature/journal/v501/n7467/fig_tab/nature12625_T1.html)). SCNAs can lead to gene dosage changes impacting phenotype; SCNAs and copy neutral LOH events at heterozygous or mutant loci can lead to unequal dose contributions of one allele over the other. HAPSEG captures such allelic copy number alterations as genome-segmented copy ratios as shown in the chart. Colors indicate low (blue), high (red), and balanced (purple) allele ratios. If unprovided, HAPSEG determines segmentation, and marks distinct copy ratios with green lines.



Haplotype phasing distinguishes HAPSEG from other segmentation modules. Extend the concept of allelic variation for a given chromosome across a population. Observed rates of co-occurrence of variation on a given chromosome in a population, in *linked inheritance* or *linkage disequilibrium (http://en.wikipedia.org/wiki/Linkage_disequilibrium) (LD)*, allows derivation of a haplotype panel of statistical likelihoods of co-variation. We can use such panels to statistically impute phased haplotypes from segmented genomic data for a member of the given population.

- The HAPSEG module provides four different population haplotype panels from HapMap. Updated haplotype reference panels are publicly available for more populations from the *International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/)* and the *1000 Genomes Project (http://www.1000genomes.org/)*.

- HAPSEG uses a prior version of the current imputation software package BEAGLE from the University of Washington (http://faculty.washington.edu/browning/beagle/beagle.html). As of this writing, BEAGLE v3 documentation is still available at https://faculty.washington.edu/browning/beagle/b3.html (https://faculty.washington.edu/browning/beagle/b3.html). Alternatively, use Wayback Machine (https://archive.org/web/) for the web address to access prior version documentation.
- Wikipedia lists additional imputation software packages and also explains the preference for haplotypes from the 1000 Genomes Project as of mid-2014 (http://en.wikipedia.org/wiki/Imputation_%28genetics%29#Tools).

A precursor workflow that calculates segmentation from earlier Affymetrix chips is outlined on GenePattern's SNP Copy Number and Loss of Heterozygosity Estimation (http://genepattern.broadinstitute.org/gp/pages/protocols/SnpCN.html) page. Other GenePattern modules that produce segmentation files are CBS (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00121:2), GLAD (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00087:2) and CopyNumberInferencePipeline (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.pipeline:00023:1). The latter uses a panel of diploid CEL files in BIRDSEED and TANGENT algorithms to normalize copy number calculations and user-provided normal sample files to adjust for batch effects. Segmentation files are also used by GenePattern's GISTIC2 module (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis:00125:6.2) to identify significantly amplified or deleted genomic regions across a set of samples.

Algorithm

HAPSEG estimates homologue-specific copy ratios (HSCRs) by genomic segment. HAPSEG was optimized on error models for gene expression from Affymetrix SNP arrays in multiple cancer-derived datasets. Refer to *Carter et al. (http://precedings.nature.com/documents/6494/version/1)* for the algorithms used.

> SNP arrays provide both allele expression and allele genotype information. Although conceptually HAPSEG could utilize high throughput sequencing data, the current offered algorithm version is most suited to microarray data.
>
> > Other algorithms to similiarly process next generation sequencing data include GATK (https://www.broadinstitute.org/gatk/)'s (Genome Analysis Tool Kit, Broad Institute) ReCapSeg (http://gatkforums.broadinstitute.org/categories/recapseg-documentation). ReCapSeg is a copy-number variant detector that runs on user-defined regions--exomes or arbitrary windows--and a panel of normal samples to segment genomic regions by differing copy ratios. It is part of the Clonal Evolution Exome Suite (http://www.broadinstitute.org/cancer/cga/acsbeta) that is under development by the CGA group as of June, 2015. These tools are unavailable on GenePattern and require scripting for use.

Briefly, HAPSEG implements an error model tailored to the basic physics of Affymetrix SNP microarrays measurements, internally recomputes genomic segmentation using error-model fit to fine-tune genome segments of equal copy number, and uses LD information from phased haplotype panels to improve inference of genotypes. Copy ratio is defined as [concentration of alleles in a cancer-derived DNA sample]/[concentration of alleles in a normal diploid DNA sample] for a given genomic segment.

Alleles are defined by genotype and grouped by haplotype segments. HAPSEG calculates **haplotype phasing** using the cancer sample allelic imbalances and additionally with population reference haplotype panel information. It models four distinct genotypes in each segment in contiguous chromosomal blocks of variation using the statistical program BEAGLE that is included in HAPSEG.

Genomic segments are defined by regions of equal copy number, i.e. each segment is a section of a chromosome where all the loci have the same number of copies. HAPSEG will determine the segments or, if provided prior segmentation information, improves upon the given segmentation.

References

Carter SL, Meyerson M, Getz G. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. Available from Nature Precedings; 2011. (abstract and PDF link (http://precedings.nature.com/documents/6494/version/1))

BEAGLE:

- Current website with latest version: http://faculty.washington.edu/browning/beagle/beagle.html (http://faculty.washington.edu/browning/beagle/beagle.html).
- BEAGLE v3 documentation: https://faculty.washington.edu/browning/beagle/b3.html (https://faculty.washington.edu/browning/beagle/b3.html).
- Alternatively, use Wayback Machine (https://archive.org/web/) for the web address to access prior version documentation.

Parameters

Name	Description
plate name *	Name of the sample plate. This is used for display and reporting purposes only.
array name *	Name of the chip that was run. This is used for display and reporting purposes only.
seg file	Segmented copy number data file for this sample (e.g., from GLAD, CBS, or similar algorithms). If this file is not provided, HAPSEG will segment the data for you.
snp file *	SNP (/cancer/software/genepattern/file-formats-guide#SNP) intensity file for this sample.
out file name	The name of the output file. By default, this will be <plate.name>_<array.name>.segdat.RData
genome build *	Which corresponding phased reference genome to use, specific to BEAGLE, based on either hg19 or hg18 builds. Phased reference genome files are different for the different versions of BEAGLE as described on the BEAGLE website.

Name	Description
platform *	The microarray chip type used. The supported values are currently: • SNP_250K_STY • SNP_6.0 (default)
use pop *	HAPMAP (http://hapmap.ncbi.nlm.nih.gov/citinghapmap.html) population to use. The currently supported values are: • CEU or CEPH (default): Utah residents with ancestry from northern and western Europe • CHB or CH: Han Chinese in Beijing, China • JPT or JA: Japanese in Tokyo, Japan • YRI or YOR: Yoruba in Ibadan, Nigeria
impute gt *	If set to TRUE, the module will impute genotypes using BEAGLE (included in the HAPSEG module). The authors recommend this be TRUE.
plot segfit *	If set to TRUE, the module will plot JPG images of the segmentation fits.
merge small *	If set to TRUE, the module will merge small segments. The algorithm for merging segments can be found in the HAPSEG paper (http://precedings.nature.com/documents/6494/version/1/files/npre20116494-1.pdf).
merge close *	If set to TRUE, the module will merge close segments. The algorithm for merging segments can be found in the HAPSEG paper (http://precedings.nature.com/documents/6494/version/1/files/npre20116494-1.pdf).
min seg size *	Minimum segment size. Default: 10
normal *	If set to TRUE, the module will treat this sample as a normal sample. The default is FALSE.
out p *	Outlier probability. Default: 0.05
seg merge thresh *	The distance threshold for merging segments. Default: 1e-10
use normal *	If set to TRUE, the module will use a matched normal sample if one is provided. The default is FALSE.
drop x *	If set to TRUE, the module will remove the X chromosome from the calculation. The default is FALSE.
drop y *	If set to TRUE, the module will remove the Y chromosome from the calculation. The default is TRUE.

Name	Description
calls file	If you are using a matched normal sample, a Birdseed SNP calls file must be supplied. Birdseed (http://www.broadinstitute.org/mpg/birdsuite/birdseed.html) is a SNP genotyping algorithm, and it outputs a file containing Birdseed genotype calls of 0 (AA), 1 (AB), or 2 (BB).
mn sample	If using a matched sample (*use normal* is set to TRUE), the name of that matched normal sample.
calibrate data *	Calibration is the process by which SNP measurements are standardized to copy ratios. If *On*, the module will perform a calibration on the input data. If *Off*, no calibration will be performed. If left at the default value (*Inferred*), the calibration status will be inferred.
clusters file	If calibrate data is *On* the user must supply a Birdseed clusters file. Birdseed (http://www.broadinstitute.org/mpg/birdsuite/birdseed.html) is a SNP genotyping algorithm, and it outputs a file containing the estimates of means and variances of intensities for each SNP for AA samples, AB (heterozygous) samples, and BB samples.
prev theta file	An optional file storing the previous theta values. Theta values represent the allelic intensity ratios for SNPs on the array. Equal heterozygotes have a ratio of 0.5, while homozygous calls gives values of ~0.8 and ~0.2.

* - required

Input Files

1. <snp.file>

   A SNP (/cancer/software/genepattern/file-formats-guide#SNP) intensity file containing this sample, which can either be per-sample (default) or multi-sample. This is a tab-delimited file with two columns named A and B and the row names correspond to the chip's probeset IDs. This file can either be a text file or a saved RData file (created in the R programing language via *write.table* or the equivalent) containing that data as the object *dat*.

   In a *multi-sample SNP file*, the probeset IDs in column A will be repeated for each sample and are distinguished by having "<array name>-" prepended to each. HAPSEG will use the <array name> parameter to decide which to load on that run, taking a multi-sample file but only operating on the chosen sample.

2. <seg.file>

   A segmented copy number file (e.g., from GLAD, CBS, etc).

3. <clusters.file>

   A Birdseed clusters file, either processed by the Affymetrix SNP6 Copy Number Inference Pipeline (http://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.pipeline:00023:1) or raw from Birdseed. This is a tab-delimited file where row names are the probeset IDs. In this case there are 6 columns: AA.a, AB.a, BB.a, BB.b, AB.b and AA.b.

4. <prev.theta.file>

   A file storing theta values from previous HAPSEG runs.

Output Files

1. <plate.name>_<array.name>.segdat.RData

   The copy number data segmented by haplotype. This is suitable as an input to the ABSOLUTE GenePattern module.

2. chr*/HAPSEG_SEG*.jpg

   Per-chromosome plots of the fitted segments. There will be one subdirectory for each chromosome and one plot for each fitted segment. These are only provided if <plot.segfit> is TRUE. *Note that these files will not be created on Windows*

Example Data

A set of HAPSEG example data from the CGA group is available at:

ftp://ftp.broadinstitute.org/pub/genepattern/example_files/HAPSEG_1.1.1/paper_example.zip

This can be run through HAPSEG and the output supplied to ABSOLUTE. A README file in the ZIP archive provides the filenames and parameters you will need to run this example data through HAPSEG, ABSOLUTE, ABSOLUTE.summarize, and ABSOLUTE.review.

Requirements

Acceptance of the module license is required for its use. A copy of the license text is available at http://www.broadinstitute.org/cancer/cga/sites/default/files/images/ABSOLUTE_HAPSEG_license_2013.pdf (http://www.broadinstitute.org/cancer/cga/sites/default/files/images/ABSOLUTE_HAPSEG_license_2013.pdf).

The module runs only on GenePattern 3.4.2 or above and requires R2.15 with the following packages, each of which will automatically download and install when the module is installed:

- boot_1.3-7
- class_7.3-5
- cluster_1.14.3
- foreign_0.8-51
- KernSmooth_2.23-8
- lattice_0.20-10
- MASS_7.3-22
- Matrix_1.0-9
- mgcv_1.7-21
- nlme_3.1-105
- nnet_7.3-5
- rpart_3.1-55
- spatial_7.3-5
- Rcpp_0.9.14

- numDeriv_2012.9-1
- iterators_1.0.6
- foreach_1.4.0
- BiocGenerics_0.4.0
- DBI_0.2-5
- xtable_1.7-0
- XML_3.95-0.1
- RSQLite_0.11.2
- IRanges_1.16.2
- Biobase_2.18.0
- AnnotationDbi_1.20.1
- annotate_1.36.0
- RColorBrewer_1.0-5
- geneplotter_1.36.0
- getopt_1.17
- optparse_0.9.5
- DNAcopy_1.32.0

Please install R2.15.3 instead of R2.15.2 before installing the module. The GenePattern team has confirmed test data reproducibility for this module using R2.15.3 compared to R2.15.2 and can only provide limited support for other versions. The GenePattern team recommends R2.15.3, which fixes significant bugs in R2.15.2, and which must be installed and configured independently as discussed in *Using Different Versions of R* (http://www.broadinstitute.org/cancer/software/genepattern/administrators-guide#using-different-versions-of-r) and *Using the R Installer Plug-in (http://www.broadinstitute.org/cancer/software/genepattern/administrators-guide#using-the-r-installer-plug-in)*. These sections also provide information on patch level fixes that are necessary when additional installations of R are made and considerations for those who use R outside of GenePattern.

There is a known issue with running HAPSEG on Windows, wherein the jpg files are not output. The .segdat.RData file produced, however, is valid.

Note that HAPSEG may require several hours to run per sample. While it is not strictly required, a computational grid or dedicated multi-core server is highly recommended. The computation generally requires at least 6G of available RAM.

Platform Dependencies

Task Type:
SNP Analysis

CPU Type:
any

Operating System:
any

Language:
R2.15

Version Comments

Version	Release Date	Description
1.6	2015-10-16	Updated to make use of the R package installer.
1	2013-06-30	Initial version.