Copy Number Variation Analysis Pipeline

Introduction

The copy number variation (CNV) pipeline uses either NGS or Affymetrix SNP 6.0 (SNP6) array data to identify genomic regions that are repeated and infer the copy number of these repeats. Three sets of pipelines have been used for CNV inferences.

- ASCAT
- ABSOLUTE
- DNAcopy

The first set of CNV pipelines are built upon the ASCAT [1] algorithm for both WGS and SNP6 data. ASCAT is able to generate Allele-specific Copy Number Segment data with integer copy number values, and the derived integer Gene-Level Copy Number. 1.) The WGS copy number analysis pipeline, ascatNGS, is described in detail here. 2.) The SNP6 copy number analysis pipeline, ASCAT2, is adopted from the example ASCAT analysis. 3.) The SNP6 copy number analysis pipeline, ASCAT3, is an updated version of ASCAT2. The ASCAT3 analysis in TCGA was done by the Vanloo lab, and the GDC released a reformatted version of these calls. Both ASCAT2 and ASCAT3 generates data similar to ascatNGS.

The second CNV pipeline, ABSOLUTE, also uses Affymetrix SNP 6.0 (SNP6) array data as input. The hg19 version of the segments were published as one of the TCGA PanCancer analysis papers and the data is available in the GDC publication page. These calls have been manually curated and thus are considered of good quality. The GDC performed segment liftover and generated gene-level copy numbers. Note that the intermediate output of GRCh38 segments contain liftover artifacts and were not released in the GDC. Users can also obtain corresponding purity and ploidy measurements from the GDC publication page mentioned above.

The third set of CNV pipelines are built onto the existing TCGA level 2 SNP6 data generated by Birdsuite and uses the DNAcopy R-package to perform a circular binary segmentation (CBS) analysis [2]. CBS translates noisy intensity measurements into chromosomal regions of equal copy number. The final output files are segmented into genomic regions with the estimated copy number for each region. The GDC further transforms these copy number values into segment mean values, which are equal to log2(copy-number/ 2). Diploid regions will have a segment mean of zero, amplified regions will have positive values, and deletions will have negative values.

ASCAT Pipelines

Data Processing Steps

Copy Number Segmentation

The Somatic Copy Number Workflow uses a tumor-normal pair of either SNP6 raw CEL data, or WGS data as input. The ASCAT algorithm derives allele-specific copy number segments while estimating and adjusting for tumor purity and ploidy [1]. Because there are two parental strands, the resulting Copy Number Segment or Allele-Specific Copy Number Segment files contain 3 different copy number integer values: Major_Copy_Number refers to the larger strand copy

number, Minor_Copy_Number refers to the smaller strand copy number, Copy_Number is the sum of Major_Copy_Number and Minor_Copy_Number, and thus equals to the total copy number at the locus.

I/O	Entity	Format
Input	Submitted Genotype_Array	CEL
Output	Copy Number Segment or Allele-Specific Copy Number Segment	TXT

I/O	Entity	Format
Input	Aligned Reads	BAM
Output	Copy Number Segment or Allele-Specific Copy Number Segment	TXT

Gene-Level Copy Number

Gene-level Copy Number is generated by inheriting the Copy_Number value of the residing segment in the Copy Number Segment file generated from ASCAT2, ASCAT3, or ascatNGS workflows.

In some occasions, one gene may overlap with more than one segment. In this case, min_copy_number is the minimum value of all segments it overlaps, max_copy_number is the maximum value of all segments it overlaps, and copy_number is calculated as the weighted (on length of overlapped regions) median of copy number values from all overlapped segments. When there is a tie (very rare), the smaller number is used. If a gene overlaps with only one segment, copy_number = min_copy_number = max_copy_number. If a gene overlaps with no segments, the gene gets empty value "" in copy_number, min_copy_number and max_copy_number.

I/O	Entity	Format
Input	Copy Number Segment or Allele-Specific Copy Number Segment	TXT
Output	Copy Number Estimate	TXT

File Access and Availability

Type	Description	Format
Copy Number Segment	A table that associates contiguous chromosomal segments with genomic coordinates, and integer copy numbers.	TXT
Allele-Specific Copy Number Segment	A table that associates contiguous chromosomal segments with genomic coordinates, and integer copy numbers.	TXT

Type	Description	Format
Copy Number Estimate	A Gene-level Copy Number file that displays integer copy number on a gene level. Generated from Copy Number Segment or Allele-Specific Copy Number Segment files.	TXT

ABSOLUTE Copy Number

Data Processing Steps

The source data were generated by external groups. Please check the corresponding publication for details.

File Access and Availability

File Access and Availability is similar to that from the ASCAT pipelines, except that only gene-level copy numbers are available, but not segmentation calls.

DNACopy Pipeline

Data Processing Steps

The GRCh38 SNP6 probe-set was produced by mapping probe sequences to the GRCh38 reference genome and can be downloaded at the GDC Reference File Website.

Copy Number Segmentation

The Copy Number Liftover Workflow uses TCGA level 2 tangent.copynumber files. These files were generated by first normalizing array intensity values, estimating raw copy number, and performing tangent normalization, which subtracts variation that is found in a set of normal samples.

The Copy Number Liftover Workflow performs CBS analysis using the DNACopy R-package to process tangent normalized data into Copy Number Segment files, which associate contiguous chromosome regions with log2 ratio segment means in a tab-delimited format. The number of probes with intensity values associated with each chromosome region is also reported (probes with no intensity values are not included in this count). During copy number segmentation probe sets from Pseudo-Autosomal Regions (PARs) were removed from males and Y chromosome segments were removed from females.

Masked copy number segments are generated using the same method except that a filtering step is performed that removes the Y chromosome and probe sets that were previously indicated to be associated with frequent germline copy-number variation.

I/O	Entity	Format
Input	Submitted Tangent Copy Number	TXT

I/O	Entity	Format
Output	Copy Number Segment or Masked Copy Number Segment	TXT

[1] Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W. et al. "Allele-specific copy number analysis of tumors." Proceedings of the National Academy of Sciences, 107.39 (2010): 16910-16915.

[2] Olshen, Adam B., E. S. Venkatraman, Robert Lucito, and Michael Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data." Biostatistics 5, no. 4 (2004): 557-572.