Canvas Software Design Document

Bioinformatics

Exported on Mar 27, 2017

Table of Contents

1 Introduction

Canvas is an algorithm for calling copy number variants from either (a) a mostly diploid germline sample, or (b) a germline sample together with a tumor sample from the same individual. The vast majority of normal germline samples will be diploid, that is, having two copies. However, tumor samples may be much more extensively rearranged. Canvas seeks to identify regions of the sample's genome that are present in zero, one, or more than two times in the genome. Briefly, this is achieved by scanning the genome for regions that have an unexpected number of short read alignments. Regions with fewer than the expected number of alignments are classified as losses. Regions having more than the expected number of alignments are classified as gains.

Canvas was originally developed to work on low-depth sequencing data from cytogenetic, low-depth single-cell and whole genome sequencing experiments. Since then it has been extended to handle tumor/normal samples and targeted protocols such as whole exome tumor/normal data. Canvas can also be run on large custom targeted enrichment panels (e.g. a 100,000-probe panel such as exome data); however, using Canvas on targeted amplicon data (TruSeq Amplicon) is not appropriate.

The Canvas source code is distributed as an open-source software, and can be accessed at: https://github.com/Illumina/canvas

Canvas is described in the publication **Canvas: versatile and scalable detection of copy number variants** in the journal Bioinformatics.

Publication link:
http://bioinformatics.oxfordjournals.org/content/early/2016/04/19/bioinformatics.btw163

Access to full paper:
http://bioinformatics.oxfordjournals.org/cgi/reprint/btw163?ijkey=muXEP3dplL8aWqJ&keytype=ref

BioRxiv link: http://biorxiv.org/content/early/2016/01/13/036194

Canvas supports a number of different workflows depending on the input sequencing data. The available modes are:

- **Germline-WGS:** CNV calling of a diploid germline sample from whole genome sequencing data
- **Somatic-Enrichment:** CNV calling of a somatic sample from targeted sequencing data
- **Somatic-WGS:** CNV calling of a somatic sample from whole genome sequencing data
- **Tumor-normal-enrichment:** CNV calling of a tumor/normal pair from targeted sequencing dat

2 Usage

Each of these workflows can be launched from the Canvas.exe command-line tool. An example of the usage information for the Somatic-WGS workflow is shown below.

```
[~]$/illumina/sync/software/unofficial/Canvas/latest/Canvas Somatic-WGS -h
Canvas 1.13.0 Copyright © Illumina 2016

Somatic-WGS - CNV calling of a somatic sample from whole genome sequencing
data

Usage: Canvas.exe Somatic-WGS [OPTIONS]+

Options:
  -b, --bam=VALUE            tumor sample .bam file (required)
      --somatic-vcf=VALUE    .vcf file of somatic small variants found in
the
                               tumor sample. Used as a fallback for
estimating
                               the reported tumor sample purity. This
option
                               has no effect on called CNVs
      --b-allele-vcf=VALUE   vcf containing SNV b-allele sites (only sites
                               with PASS in the filter column will be
used)
                               (required)
      --exclude-non-het-b-allele-sites
                             exclude SNV b-allele sites from the vcf that
do
                               not have sufficient read evidence for a
                               heterozygous genotype in the sample. This
option
                               should be used when the b-allele vcf is not
                               matched to the sample (e.g. from dbSNP)
      --ploidy-bed=VALUE     .bed file containing regions of known ploidy
in
                               the sample. Copy number calls matching the
known
                               ploidy in these regions will be considered
non-
                               variant
  -o, --output=VALUE         output directory (required)
  -r, --reference=VALUE      Canvas-ready reference fasta file (required)
  -g, --genome-folder=VALUE  folder that contains both genome.fa and
                               GenomeSize.xml (required)
  -f, --filter-bed=VALUE     .bed file of regions to skip (required)
  -n, --sample-name=VALUE    sample name (required)
      --custom-parameters=VALUE
                             use custom parameter for command-line tool.
                               VALUE must contain the tool name followed
by a
                               comma and then the custom parameters.
Option can
                               be specified multiple times.
  -c=VALUE                   continue analysis starting at the specified
                               checkpoint. VALUE can be the checkpoint
name or
                               number
  -s=VALUE                   stop analysis after the specified checkpoint
is
                               complete. VALUE can be the checkpoint name
or
                               number
  -h, --help                 show this message and exit
```

```
  -v, --version             print version and exit
```

Below is an example script to launch the workflow using GRCh37 as the reference genome. It must be run on a dedicated compute node (e.g. through SGE using qsub: qsub -l excl=true example_script.sh).

```
#!/usr/bin/env bash
/illumina/sync/software/unofficial/Canvas/1.13.0/Canvas Somatic-WGS --bam
'/bioinfoSD/eroller/COLO829C/COLO829C_S1.bam' --sample-name 'COLO-829C-
3lanes' --reference
'/illumina/sync/software/unofficial/Isas/Genomes/Homo_sapiens/Ensembl/GRCh
37/Annotation/Canvas/kmer.fa' --genome-folder
'/illumina/sync/software/unofficial/Isas/Genomes/Homo_sapiens/Ensembl/GRCh
37/Sequence/WholeGenomeFasta' --filter-bed
'/illumina/sync/software/unofficial/Isas/Genomes/Homo_sapiens/Ensembl/GRCh
37/Annotation/Canvas/filter13.bed' --output
'/bioinfoSD/eroller/CanvasTest' --b-allele-vcf
'/bioinfoSD/eroller/COLO829BL/COLO829BL_S1.vcf.gz' --somatic-vcf
'/bioinfoSD/eroller/COLO_smoke/COLO-829BL-2lanes_COLO-829C-
3lanes_G1_P1.somatic.vcf.gz' --ploidy-bed
'/bioinfoSD/eroller/COLO_smoke/DetectCNVTemp/COLO-829C-
3lanes/ploidy.bed.gz'
```

3 Canvas Overview

Outline The Canvas workflow comprises five distinct modules designed to (1) process aligned read data and calculate coverage bins, (2) perform outlier removal and normalization of coverage estimates, (3) identify segments of uniform copy number, (4) calculate minor allele frequencies (MAF) and (5) assign copy number / allelic states and infer genome-wide parameters. Separate workflows exist for somatic, germline and exome sequencing data. The latter workflow can be run either with or without a matched normal control sample and requires a manifest file with locations of targeted regions. For brevity, here we highlight core functionalities of individual modules,

Coverage binning Binning is performed by fixing a total number of unique k-mer alignments within each bin, leading to variable window size that depends on repeat content and alignment complexity of genomic regions.

Normalization Options for both bin-level and fragment-based GC-content normalization exist, along with a number of outlier removal strategies.

Partitioning and MAF calculation Canvas provides implementation of two coverage partitioning algorithms: unbalanced Haar wavelet transform (used by default) and circular binary segmentation (Fryzlewicz, 2007; Venkatraman et al., 2007). MAFs are derived at heterozygous SNV positions determined either by germline SNV variant calls in the reference (normal) sample or by direct assessment using population SNV panels (such as the 1000 Genomes Project or dbSNP database).

Somatic model and copy number (CN) assignment for tumour-normal whole-genome and exome workflows Accurate somatic copy number assignment requires knowledge of genome ploidy, sample contamination and location of heterogeneous variants. To infer these parameters Canvas fits a deviation model based on the principle that for a given combination of purity (normal contamination level) and diploid coverage level, MAF and coverage values for each CN state can be easily derived. Iterating over different purity and coverage combinations Canvas calculates a model deviation *D* by comparing observed and expected MAF and coverage values. To better discriminate models with similar deviation values, a separate penalty term *P* is introduced to measure the complexity of each model. The penalty is related to the total number of implied somatic events needed to produce a given tumor genome. Variables that best predict genome complexity are inferred using logistic regression on a training set of sequenced tumor genomes with known karyotypes. Finally, a separate EM clustering module is used to group *MAF, Coverage* tuples into clusters and assess if they could represent heterogeneous variants. This is done by computing distance between observed cluster centroids and expected *MAF, Coverage* values under the null (no heterogeneity) model. The resulting cluster deviation *C* is combined with other two metrics to provide a total deviation *T = D (model deviation) + P (penalty term) + C*. Ploidy, contamination and heterogeneity values from the model with the smallest total deviation are then used to assign CN and genotype to each segment.

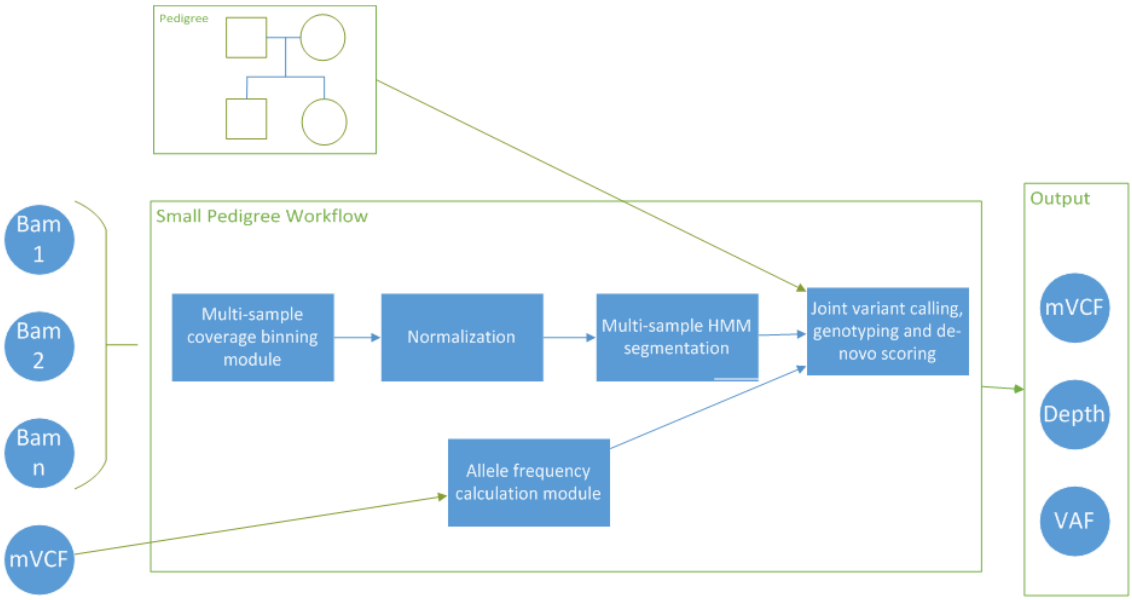
CN assignment for germline variants Canvas can also detect germline CNVs from single-sample whole genome sequencing experiments. Since large scale changes in genome ploidy or sample contamination are rarely, if at all, present in "normal" samples, Canvas assigns copy numbers based on the premise that the majority of genomic regions are diploid and that there is no contamination (as seen with somatic genomes).

Below is a detailed description of individual modules.

4 Canvas Multi-Sample

Canvas Small Pedigree Workflow (SPW) is designed to provide ability to call de-novo variants in parents-offspring trios when the pedigree information is available. In cases of unrelated samples, Canvas SPW will perform a joint multi-sample CNV calling to ensure that variants called are normalized and have the same breakpoints across samples. Similarly to other Canvas workflows, SPW is composed of a number of modules:

1) Input data processing and binning (CanvasBin, CanvasSNW)
2) Normalization and outlier removal (CanvasClean)
3) Multi-sample segmentation (CanvasPartition)
4) Variant calling, genotyping and de novo calling (CanvasPedigreeCaller)



5 Modules

Canvas workflows are composed of a number of distinct steps. Using the single executable described above will launch the appropriate steps using default parameters. The steps used in Canvas are:

(1) Counting alignments in genomic bins (implemented as CanvasBin.exe)

(2) Removal of systematic biases and outliers from the counts (implemented as CanvasClean.exe)

(3) Partitioning the counts into homogenous regions (implemented as CanvasPartition.exe)

(4) Quantifying SNV allele frequency for each region, including heterozygous-only SNVs identified in the normal sample for tumor/normal workflow (implemented as CanvasSNV.exe)

(5) Assigning a copy number to each homogenous region (implemented as CanvasSomaticCaller.exe or CanvasDiploidCaller.exe)

Below is a description of individual steps.

5.1 CanvasBin

The CanvasBin procedure creates genomic windows, or bins, across the genome and counts the number of observed alignments that fall into each bin. The alignments are provided in the form of a BAM file. Canvas has been validated using BAM files created using the Isaac and Bowtie aligners; results on .bam files from another aligner (e.g. bwa) would also likely be appropriate. The specific bowtie command appropriate for Canvas is of the form

```
zcat R1.fastq.gz | cut -b 1-35 | bowtie --quiet -S -q -v0 -m1 hg19 - |
samtools view -Sbo FromBowtie.bam -
```

In this example, hg19 is a bowtie index generated for the hg19 genome built using bowtie's bowtie-build command. Note that only the first 35 bases of read 1 are aligned and stored in the BAM file. It follows that CNV-only applications such as low-depth cytogenetics require the sequencing of just 35 bases from one end of each sequencing template.

Isaac BAM files are created through calls made within the whole genome sequencing (WGS) workflow. Canvas applies special treatment to BAM files when measuring coverage. Specifically, Canvas will only keep alignments from proper pairs that align to the forward strand. Additionally, the alignment's CIGAR string must begin with at least thirty-five matches. It was found that restricting Isaac alignments in this way yielded very similar results to using Bowtie alignments of the same input data. To invoke this behavior at the command-line, the "-p" flag must be used when calling CanvasBin.exe.

CanvasBin stores observed alignments in RAM as a collection of byte arrays, one array for each chromosome. Each array has the same length as the the size of a corresponding chromosome. As the BAM file is read in, CanvasBin records the position of each alignment's left-most base within the chromosome-appropriate byte array. After all alignments in the BAM file have been read, the byte arrays have a nonzero integer wherever an alignment was observed and a zero everywhere else. The -m (coverage mode) argument to CanvasBin controls the details of how these counts are accumulated; the default mode "TruncatedDynamicRange" counts up to 10 hits at each position.

In addition to a BAM file, CanvasBin accepts FASTA **kmer.fa** file as input (provided in the Annotation/Canvas subfolder for a reference genome) that contains genomic sequences that were used for alignment. The first base of each 35-mer within this FASTA file is marked as unique or non-unique with uppercase and lowercase letters. If a 35-mer is unique then its first nucleotide is capitalized, otherwise, it is not capitalized. Note: In previous workflow versions (2.6.6 and earlier), the kmer file was named **genome.fa**; the new name is clearer and includes correct handling of PAR-masked regions.

```
For example, in the sequence
```

```
acgtttaATgacgatGaacgatcagctaagaatacgacaatatcagacaa
, the three 35-mers marked as unique would be
ATGACGATGAACGATCAGCTAAGAATACGACAATA
TGACGATGAACGATCAGCTAAGAATACGACAATAT
GAACGATCAGCTAAGAATACGACAATATCAGACAA
```

The genomic locations of unique 35-mers are stored by CanvasBin in a collection of bit arrays analogous to those used to store alignment positions. Unique positions and non-unique positions are marked with “1”s and “0”s, respectively. The tool **FlagUniqueKmers**, from the Canvas solution, can be used to generate this uniqueness-flagged annotation for a reference FASTA file.

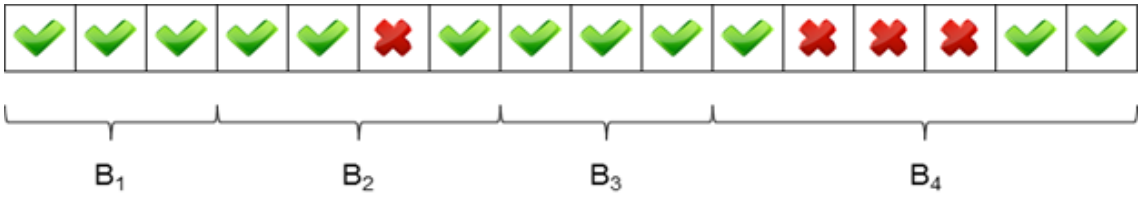
If a Bowtie BAM is used, then the alignment byte arrays will never have a nonzero value in a position where the corresponding position in the unique 35-mer byte array is a “0”. This is because bowtie, if used in the appropriate way for Canvas (see example earlier in this section), will only return exact 35-mer matches that are unique in the genome. Isaac is a more sensitive aligner and can place reads in locations that are not prefixed by a unique 35-mer. Therefore, CanvasBin next iterates through the byte array for hit counts, and sets this to “0” everywhere that the unique 35-mer byte array is set to “0”. This guarantees that we only use Isaac alignments that start at unique 35-mer positions in the genome.

5.1.1 Excluded Regions

Optionally, a subset of unique 35-mers can be masked from analysis. A BED file can be provided at the command line that specifies regions of the genome to ignore from analysis. After the alignments and unique 35-mers are loaded, the regions in this BED file are read in. For each entry in the file, the elements of both the alignment and unique 35-mer byte arrays that are contained within the entry’s range are set to “0”. This effectively excludes the region from further analysis and is therefore a means to ignore problematic regions of the genome. A default masking file is provided with Isis/Canvas (now named filter13.bed - was named filter.bed in previous releases) which includes such regions. It was generated by sequencing a pool of 100 normal individuals and calling CNVs with Canvas. Since the “sample” is really a pool of many samples, any detected region is presumably an artifact of the reference genome or the secondary analysis pipeline - e.g. telomeres and centromeres (and some pericentromeric regions) are excluded.

5.1.2 Bin Size

When CanvasBin is invoked, the user supplies a desired average number of alignments per bin for diploid regions. By default this value is set to 100 alignments per bin for WGS data. (This default was chosen to provide roughly the same signal-to-noise ratio as an Agilent CGH array.) For enrichment data, the default is set to 300. This value must then be converted into a “bin size.” In typical usage, “bin size” would refer to the size of some fixed genomic window. An example of this might be a bin size of 10kb. Here, we use the term “bin size“ to refer to the number of unique genomic 35-mers per bin. Because some regions of the human genome are more repetitive than others, bins’ physical sizes (in genomic coordinates) will not be identical. Below is a toy example to illustrate this concept. Each box is a position along the genome. Each checkmark represents a unique 35-mer while each X represents a non-unique 35-mer. The bin size here is three (three check marks per bin). Note that the physical size of each bin is not constant. B_1 and B_3 have a physical size of three but B_2 and B_4 have physical sizes of four and six, respectively.



To compute bin size, the ratio of observed alignments to unique 35-mers is calculated for each autosome. The desired number of alignments per bin is then divided by the median of these ratios to yield bin size. For whole genome sequencing, bin sizes will typically be in the range of 800-1000 unique 35-mers. Correspondingly, the majority of physical window sizes will be in the 1-1.2kb range. The advantage of this approach relative to using fixed genomic intervals for bins is that we now expect to have the same number of reads mapping to each bin, regardless of the bins' "mappability" or "uniqueness". Optionally, CanvasBin takes a Nextera manifest file through command line argument -t <path to manifest> (–manifest=<path to manifest>). In this case, only the targeted regions specified in the manifest are used to compute the bin size such that bins overlapping the targeted regions (the on-target bins) have the desired median number of alignments.

5.1.3 Bin Size for multiple samples

Data processing and coverage binning is done in the same manner as in single sample whole-genome re-sequencing workflow. The only difference involves calculation of bin sizes. As in the case of single sample workflow, to compute bin size, the ratio of observed alignments to unique 35-mers is calculated for each autosome. The desired number of alignments per bin is then divided by the median of these ratios to yield bin size. The change in the multi-sample workflow is that the median of "observed alignments / unique 35-mers" ratio is taken across all samples for each chromosome. This enables aggregation of coverage biases and ensurers that bins in all samples have the same genomic coordinates.

5.1.4 Bins

Once bin size is computed, bins are defined as consecutive genomic windows such that each bin contains the same number, bin size, of unique 35-mers. The number of observed alignments present within each bin's boundaries is then counted from the alignment byte arrays. The GC content of each bin is also calculated.

CanvasBin supports user-defined bins specified by -n <path to a BED file>. This option is particularly useful for enrichment data. Probe regions and exons are commonly used for binning. Since user-defined bins are not guaranteed to contain the same number of unique 35-mers per bin, downstream coverage normalization is needed such that each bin follows the same distribution.

5.1.5 Fragment Binning

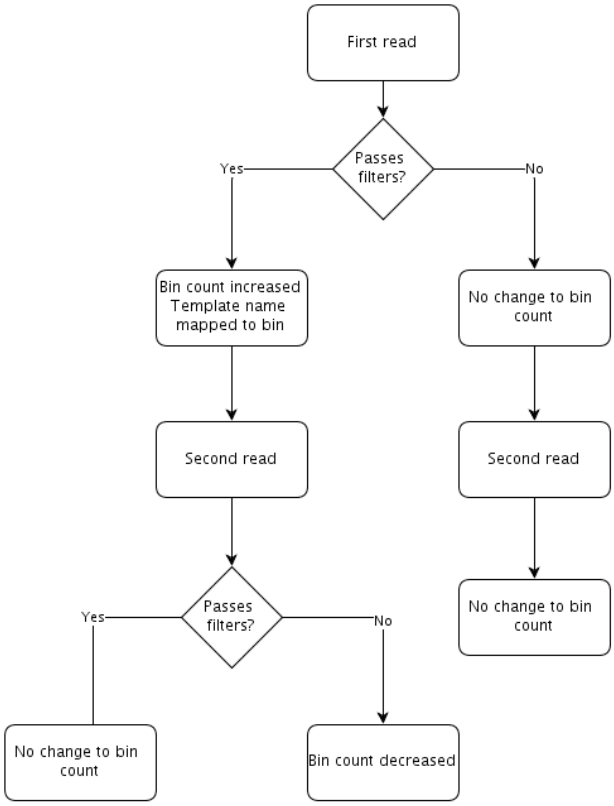
CanvasBin supports fragment binning with user-defined bins. The fragment binning mode is turned on by command line argument -m Fragment. The algorithm works by assigning sequenced fragments to bins. The aim is to associate each bin with a number of fragments and report these numbers as the counts for the bins. The start and end positions of fragments are calculated from the locations of paired-end reads in the input BAM file. A fragment will only contribute towards the count for a particular bin if both reads in the pair pass a set of filters. Each fragment is assigned to a bin according to which bin the fragment overlaps the most. If a fragment overlaps more than one bin by the same amount, it will be assigned to the left-most (smallest start position) bin.

Each read is subject to a number of quality filters before it can form part of a valid fragment. These filters are split into three categories:

1. Are the read and its mate properly paired?
2. Is the read a unique?
3. Is the read a high-quality read with a good alignment?

The algorithm operates in such a way that only fragments where both reads in the pair pass the above filters contribute to a bin's count and it requires one pass through the BAM file. To do this, the first and second reads are considered separately. The fragment start and end can be identified from each individual alignment record. If the first read in a pair passes the filters, the count for the corresponding bin will be incremented and a unique name associated with this pair

of reads will be mapped to the bin. When the second read in the pair is encountered in the pair, if it doesn't pass the filters and the first read has been used to increment a count, the bin's count will be decreased by one. The following flowchart outlines the changes to the bin counts as the first and second reads are encountered in the BAM file.



The following table shows the changes to the bin count that will be made as the first and second read in a pair are encountered. The result of this algorithm is an integer count associated with each bin.

First Read	Bin Count Change	Second Read	Bin Count Change	Overall Change
Passes filters	+1	Passes filters	+0	+1
Passes filters	+1	Fails filters	-1	+0
Fails filters	+0	Passes filters	+0	+0
Fails filters	+0	Fails filters	+0	+0

5.1.6 Output Format

The output of CanvasBin is a **G1_P1_0.binned** gzip bed file containing chromosome, genomic start, genomic stop, observed counts and GC content for each bin. For illustration, the first five lines of one such file is shown below:

```
chr1    754148  755200  90      46
chr1    755200  756030  93      54
chr1    756030  756783  90      54
chr1    756783  757489  89      53
chr1    757489  758242  80      55
```

5.2 CanvasNormalize

CanvasNormalize performs coverage normalization for enrichment data. This module is usually used after CanvasBin and before CanvasClean. A reference sample is either given as an input file, computed based on one or more input sample, or computed using a PCA model as an input file. The reference sample must contain the same bins in the same order as the input sample. Coverage normalization is done by dividing the sample bin count by the reference bin count. The ratio is then multiplied by (20 * "the normal copy number of the bin") to convert it to "coverage". That is, for a bin with normal copy number of 2, the ratio will be multiplied by 40. This makes sure the diploid coverage is around 40 so the CanvasSomaticCaller works properly. The ploidy information is given as an input ploidy BED file.

CanvasNormalize can be run in 3 different modes, WeightedAverage, BestLR2 and PCA. The modes differ in how a reference sample is obtained, how the bins are filtered out and how the ratios are computed. We provide the details for each mode in the following sections.

5.2.1 WeightedAverage

To compute the reference sample, one or more normal sample is given as input. If only one normal sample is given, it is used as the reference sample. Otherwise, a weighted average of the normal samples is computed. The weights are inversely proportional to the median bin count and they add up to one. If a manifest is given, the median is computed over only the on-target bins, which are the bins overlapping the targeted regions specified in the manifest.

To calculate the sample to reference ratios, we loop through each bin. A bin is skipped if the reference bin count is less than 1. This helps prevent introducing too much noise into segmentation and CNV calling. Library-size normalization is also done by dividing the bin counts by the median bin count for the input and reference samples. The is done only when both input and reference median bin counts are greater than 0. When a manifest is given, the median is computed using only the on-target bins.

5.2.2 BestLR2

As in the WeightedAverage mode, if only one normal sample is given, it is used as the reference sample. Otherwise, one of the normal samples is chosen as the reference sample for the input sample. For each normal sample, the mean squared log-transformed input-to-normal ratio is calculated across all bins. The reference sample is then selected to be the normal sample with the least mean squared log ratio.

Calculation of the sample to reference ratios is done in the same way as in the WeightedAverage mode.

5.2.3 PCA

In the PCA mode, an input PCA model is given in place of normal samples. The PCA model file is in a BED-like format containing columns chromosome, genomic start, genomic stop, mean, axis1, axis 2, ..., axis *p,* where each row corresponds to a bin. The input sample and the PCA model must contain the same bins in the same order.

To create the reference sample, the input bin count vector, **y**, is first centered by the mean vector, **μ**, given in the PCA model, projected onto the space spanned by the *p* axes specified in the model, and translated back using the mean vector. That is, $\mathbf{r} = (\mathbf{y}-\boldsymbol{\mu})^T\mathbf{x}_1\,\mathbf{x}_1 + ... + (\mathbf{y}-\boldsymbol{\mu})^T\mathbf{x}_{p+1}\,\mathbf{x}_p + \boldsymbol{\mu}$, where $\mathbf{x}_i$'s are the axes in the model and **r** is the bin count vector for the reference sample. We then set each reference bin count to 1 if it is below 1. The reference bin counts are further normalized such that the median sample to reference ratio is 1, where the median is calculated with respect to only the on-target bins if a manifest is given.

To calculate the sample to reference ratios, an sample bin count is set to 1 if it is less than 1. A bin is skipped if the reference bin count is below the user-specified minimum or above the user-specified maximum.

5.2.4 Output Formats

CanvasNormalize outputs the normalized bin counts in the same BED format as the output from CanvasBin. It outputs another copy number data file in CSV containing columns "Fragment Count", "Reference Count", Chromosome, Start, End and "Unsmoothed Log Ratio". The "Fragment Count" column contains the input sample bin counts, the "Reference Count" column contains the reference bin counts, and the "Unsmoothed Log Ratio" column contains log-transformed ratios of "Fragment Count" to "Reference Count".

```
Fragment Count,Reference Count,Chromosome,Start,End,Unsmoothed Log Ratio
709,209.99,chrM,3386,3466,3.376351
448,139.64,chrM,3586,3666,3.20825
600,176.41,chrM,3786,3866,3.401168
496,146.24,chrM,3986,4066,3.391685
```

5.3 CanvasClean

CanvasClean is comprised of four procedures for removing outliers and systematic biases from the count data computed in CanvasBin. These are

(1) Single point outlier removal

(2) Physical size outlier removal

(3) GC content correction

(4) Normalization of formalin-fixed, paraffin-embedded (FFPE) samples (somatic workflows only)

CanvasClean takes as input the BED file created by CanvasBin.

5.3.1 Single Point Outlier Removal

The aim of this step is to remove individual bins that represent extreme outliers. These are bins whose counts are very different from the counts present in the bins upstream and downstream from themselves. We define two values, *a* and *b*, to be very different if their difference is greater than we would expect by chance, assuming *a* and *b* come from the same underlying distribution. Specifically, we make use of the Chi-squared distribution:

$\mu = 0.5a + 0.5b$

$\chi^2 = ((a - \mu)^2 + (b - \mu)^2)\ \mu^{-1}$

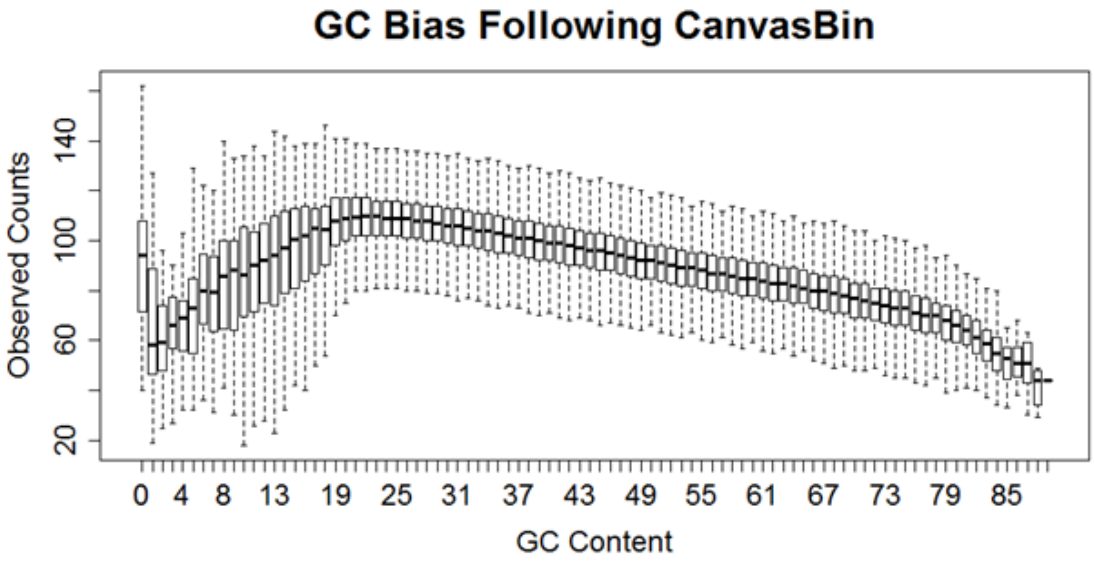
A value of χ^2 greater than 6.635, which is the 99th percentile of the Chi-squared distribution with one degree of freedom, is considered “very different.” If a bin’s count is very different from both its immediate upstream and downstream neighbors’ counts, then the bin is deemed an outlier and removed.

5.3.2 Physical Size Outlier Removal

Recall that bins will likely not have the same physical (genomic) size. The average for whole genome sequencing runs might be approximately 1kb but some may be several megabases in size if they cover repetitive regions of the genome. Example regions might include centromeres and telomeres. The counts in these regions tend to be unreliable so we remove bins with extreme physical size. Specifically, we calculate the 98th percentile of observed physical sizes and remove bins whose sizes are larger than this threshold.

5.3.3 GC Content Correction

The main driver of variability in bins' counts is their GC contents. CanvasClean provides a mechanism for correcting any GC content bias that may be present in the counts calculated in CanvasBin. An example of the magnitude of the bias is presented in the following figure.



Two GC normalization modes are available, MedianByGC and LOESS.

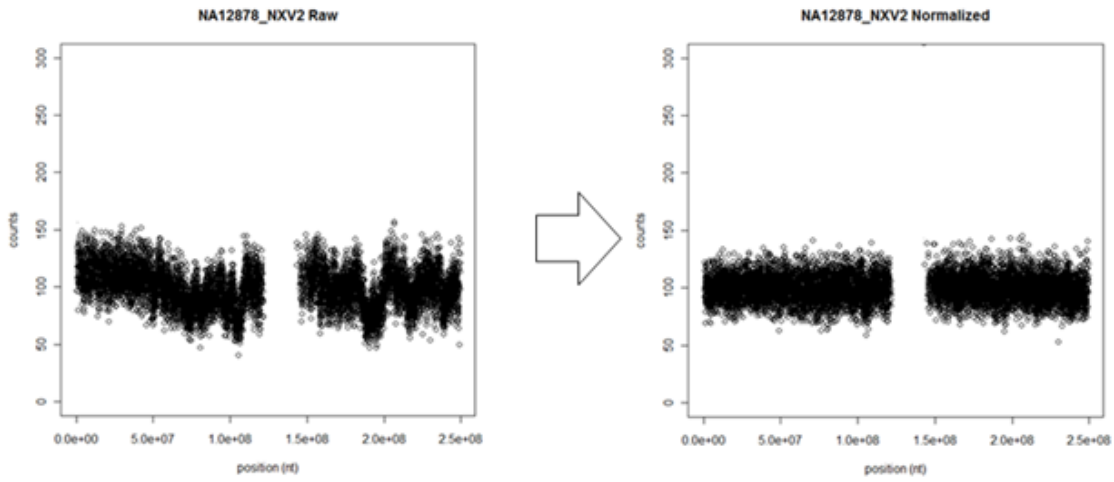
5.3.3.1 MedianByGC

The correction is simple. Bins are first aggregated by their GC content, which has been rounded to the nearest integer. Second, each bin's count is divided by the median count of bins having the same GC content. Finally, this value is multiplied by the global median. The effect is to flatten the midpoints of the bars in the box-and-whisker plot displayed above.

Some values for GC content will have very few bins so the estimate of its median is not very robust. To guard against this, we discard any bin where the number of bins sharing the same GC content is fewer than a threshold computed by max(minNumberOfBinsPerGCForWeightedMeidan, number of bins/100). If a GC value has fewer than 100 bins, the weighted median is computed using bins of neighboring GC values. Bins of the target GC value get full weight, bins of the two neighboring GC values get half weight, two-away bins get 1/4 weight, etc. Neighboring GC values are included until we have at least 100 bins to estimate the weighted median bin count. Parameter minNumberOfBinsPerGCForWeightedMeidan can be set using command line argument -w <number of bins>.

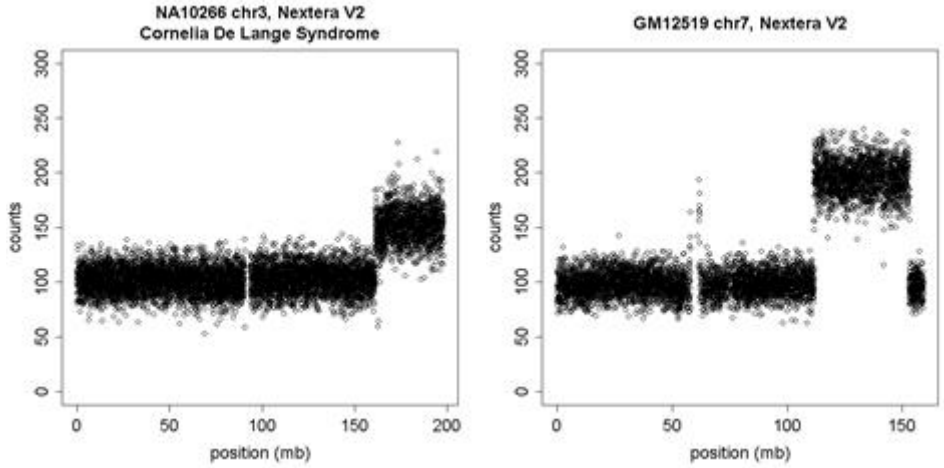
Optionally, CanvasClean takes a Nextera manifest file through command line argument -t <path to manifest> (–manifest=<path to manifest>). In this case, only on-target bins are used to estimate the count distribution at each GC content level. Similarly, we discard any any bin where the number of on-target bins sharing the same GC content is fewer than a threshold computed by max(minNumberOfBinsPerGCForWeightedMeidan, number of on-target bins/100). The weighted median is similarly computed when there are less than 100 bins for a GC value.

For some sample preparation schemes, GC content correction has a dramatic effect. The figure below illustrates this for a low depth sequencing experiment using the Nextera library preparation method. The figure on the left shows bins counts as a function of chromosome position before normalization. The figure on the right shows the same, following GC content correction.



For whole genome sequencing experiments, we typically see median absolute deviations (MADs) of 10.3, which is very close to the expected value of 10 predicted by the Poisson model for an average count of 100, indicating that little bias remains following the simple procedures implemented in CanvasClean.

Importantly, the normalization signal does not dampen signal from CNVs. Below are two figures, one for a chromosome known to harbor a single copy gain and one for a double copy gain.



5.3.3.2 LOESS

This GC normalization mode can be turned on by -m LOESS or --mode LOESS.

In the LEOSS mode, GC normalization works the same as in the MedianByGC mode except that the fitted bin count for each GC percentage is obtained from fitting a LOESS model. The bin counts are first log-transformed for variance stabilization, normalized by GC content, and transformed back after normalization.

The degree of the polynomials used in LOESS is 1. The bandwidth is searched in [0.3, 0.75] to find the value that minimize an objective function**without using chromosome Y**. Given a bandwidth, GC percentages and bin counts, the fitted bin counts are obtained by LOESS to calculate the normalized bin counts:

"normalized bin count" = "bin count" - "fitted bin count" + "median bin count".

Another LOESS is performed on the normalized bin counts and the standard deviation of the fitted values is calculated. This is because, after two passes of LOESS, the standard deviation is expected to be small when there is no CNV.

Once the best bandwidth is found, a LOESS model is fitted using **all the chromosomes**. The normalized bin counts are calculated and exp-transformed.

5.3.4 Normalization of FFPE samples

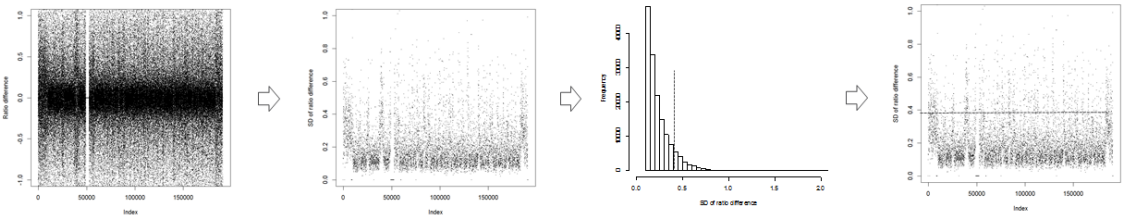
FFPE procedure is a common way of preserving tumor tissues following a cancer treatment surgery and is therefore a frequent source of material for tumor sequencing. However, FFPE fixation introduces noise and biases during sample preparation (fragmentation, need of high temperatures for DNA extraction) that lead to difficult-to-interpret coverage depth anomalies. While some of it is traced to the dropout of AT-rich sequencing fragment (in part corrected by GC-normalization), it doesn't fully explain the biases suggesting that other poorly understood factors, like chromatin confirmation and 2/3D DNA interaction might be involved. Canvas uses a two-step strategy to normalize FFPE samples:

(1) FFPE de-noising algorithm to remove localized biases

(2) Fragment-based normalization

5.3.4.1 FFPE de-noising algorithm

Canvas uses a separate FFPE de-noising algorithm that utilizes FFPE-specific properties of noise and variance. In particular, the main feature of FFPE-induced noise is an increase in coverage variance in regions of high bias, which could not be corrected by normalization routings relying on mean and related summary statistics (median, mode, etc.). Specifically, the method is composed of the following steps (also show on the pictogram):

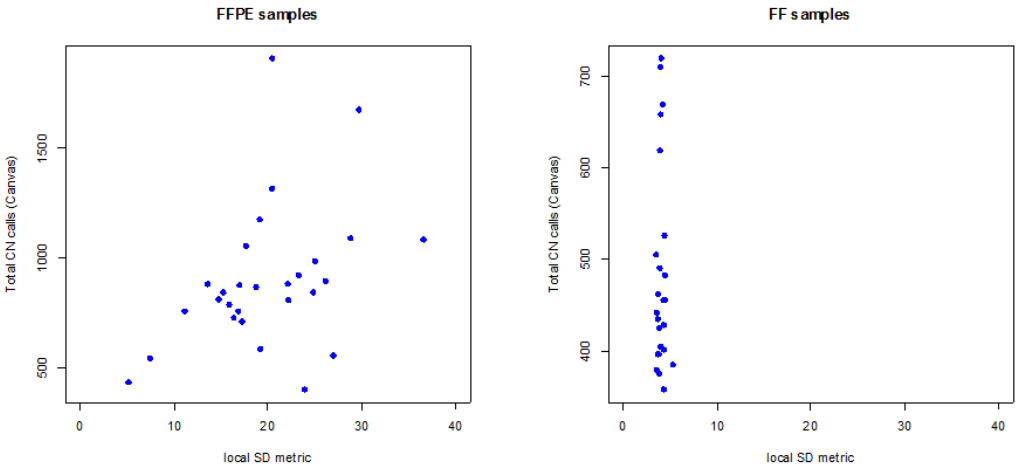


1) Estimate time-lag difference of consecutive coverage bins. Time-lagged difference reveals CN-independent spikes of variance in FFPE-bias regions. The advantageous properties of the transformation include:

1. The mean for all regions (including large scale CN changes) is zero;
2. The variance is doubled (http://en.wikipedia.org/wiki/Skellam_distribution) - making regions of noise easier to detect
3. Correct breakpoint regions can be distinguished by sharper transition between CN changes than the regions of choppiness.

2) Such spikes can be used to implement a simple local FFPE bias/variance de-noising algorithm. Canvas runs a sliding window of coverage bins (20 bins by default) and calculates standard deviation in each window. The average (SD) over all windows and chromosomes is then estimated for each sample; for simplicity such a metric is called local standard deviation (localSD). Bins with localSD threshold are removed at a CanvasClean stage. Since this is a lossy transformation, the threshold is set such that only noisy FFPE samples are de-noised, while normal and less noisy FFPE samples are not affected (step 3 below for details).

3) Comparing localSDs of coverage data from a reference set of FFPE and FF samples derived from the same individuals clearly separates FFPE and FF groups by the metric (**localSD threshold set to 10** from examples below). This property is used to set a localSD threshold specifically to insure that only noisy FFPE samples undergo de-nosing.



5.3.4.2 Fragment-based normalization

While bin-level normalization corrects most of the noise, it still does not capture variability that occurs at the level of individual fragment sizes. The later is particularly influenced by PCR amplification step occurring during FFPE library preparation workflow. To improve it, Canvas normalizes bin counts through a weighted GC-content correction factor, derived by averaging contribution of individual fragments towards the total bin counts. A fragment average is weighted by ratio of observed versus expected GC-content counts for a given fragment sequence.

CanvasClean writes to disk a file of the same format as was written by CanvasBin.

5.4 Multi-sample binning, normaliztion and MAF calculation

Canvas SPW uses the same coverage binning (CanvasBin) and normalization (CanvasClean) steps as other Canvas workflows (i.e. single sample germline and somatic). Briefly, binning is performed by fixing a total number of unique k-mer alignments within each bin, leading to variable window size that depends on repeat content and alignment complexity of genomic regions. Normalization strategy includes the bin-level GC-content normalization, the single point outlier removal and the physical size outlier removal.

The only two difference specific for Canvas SPW are:

1) Coverage binning is caluclated jointly accross a pedigree by estimating median multi-sample depth, which ensures that bins in different samples are represented by the same genomic corrdinates.

2) Bin outlier removal is trigged if at least one sample within a pedigree passes outlier removal thresholds.

Allele frequency calculation is also done by invoking the same components (CanvasSNV) as other Canvas workflows: SNV sites are derived from a supplied VCF file that could either contain SNV / small variant calls for a pedigree or a list common SNVs. The only difference is that multi-sample pedigree SNV file is required and vcf sample names should match the ones present in bam files. Temorary files from these modules are written into separate folders, one for each sample.

5.5 CanvasPartition

CanvasPartition takes as input normalized CanvasClean BED files and segments all bins into set of distinct segments having equal mean coverage values. CanvasPartition provides

implementation of a number of segmentation algorithms: currently these include unbalanced Haar wavelet segmentation (the default) and circular binary segmentation (CBS).

5.5.1 Single sample segmentation

5.5.1.1 Unbalanced Haar wavelet

Haar wavelet segmentation is an alternative partitioning method introduced in (Fryzlewicz P, 2007). It is based on a traditional discrete wavelet transform (DWT), but in contrast to DWT allows for an arbitrary size segments (rather then size factor of two). This is achieved by using adaptation of matching pursuit algorithm (see paper for details) that allows breakpoint basis selection in O(Nlog(N)) time. The resulting DWT coefficients along with original Haar basis vectors can be used to recontact a de-noised signal after setting all DWT coefficients below user specified threshold to zero. This smoothed piecewise constant signal can be used to characterize breakpoints in CanvasClean output by looking at a difference between consecutive data points. The -f argument allows adjusting the MAD factor (default: 2) to control the number of output segments.

5.5.1.2 Circular binary segmentation

CBS is an algorithm for identifying regions of the genome such that their average counts are statistically different than their neighboring regions' average counts. The implementation is a port of the circular binary segmentation (CBS) algorithm from its original C/Fortran to C#. The algorithm is fully described in (Olshen AB, 2004) and (Venkatraman ES, 2007). Briefly, each chromosome is considered as a segment. The algorithm looks into each chromosome/segment and identifies the pair of bins for which the counts in the bins between them are maximally different than the rest of the bins' counts. The statistical significance of the maximal difference is assessed via permutation testing. If the difference is found to be statistically significant, then the procedure is applied recursively to the two or three segments created by partitioning the current segment by the identified pair of points. Input to the algorithm is the output generated by the CanvasClean algorithm. The computational complexity of the algorithm $O(N^2)$ so the problem is divided into sub-chromosome problems followed by merging, in practice. Heuristics are used to speed up the permutation testing. The reader is directed to the referenced articles for a detailed discussion.

5.5.2 Multi-sample segmentation

5.5.2.1 HMM

Canvas SPW uses a Hidden Markov Model (HMM) to produce segmentation of the normalized coverage data.

Hidden States By default for each sample HMM uses 5 discrete copy number states to capture coverage distribution within bins representing homozygous copy deletion, heterozygous deletion, normal diploid state, a gain and multiple-copy duplication respectively. To jointly model pedigree samples, each hidden Markov state is associated with a set of all possible copy numbers. With complete enumeration this will scale exponentially in a number of samples. For example, with three samples such a strategy will produce a total of 125 possible copy number combinations as hidden Markov states. To control statistical model complexity and computational scalability (in both memory and runtime), hidden state specification is corrected in the following two ways. First, when assigning copy numbers to samples within a pedigree it is assumed that each CNV variant has only one ALT allele at each site. This allows hidden states to be grouped according to the copy number of the ALT allele. Second, achieving such a grouping, a maximization step will select only one hidden state (with the maximal likelihood) and will only report group's copy number in the output. This approach effectively reduces the number of hidden states from exponential to linear in the number of CNs (five by default).

Emission Distribution To capture over-dispersion of count data in bins, particularly for heterozygous and homozygous deletions, multivariate negative binomial distribution is used to model emission probabilities. Number of dimensions of such a distribution is set to be equal to the number of CN states and the distribution of bin counts is set to depend on a state's copy number. An alternative parametrization of negative binomial distribution via mean (μ) and

variance (σ) was used to model distribution of counts in a bin as $Prob(x\ counts\ in\ a\ bin) = \left(1 + \frac{\mu}{x}\right)^{-k} \frac{(k+x-1)!}{x!(k-1)!}\left(\frac{\mu}{\mu+x}\right)^{x}$, where *k* is the clumping parameter set to $k = \frac{\mu^2}{\sigma^2-\mu^2}$. Initial values of coverage means and variances of emission distribution are estimated from the observed mean / variance.

Transition Distribution Transition distribution is represented by a simple square matrix with dimensions equal to the number of CN states.

HMM Partitioning Following HMM parameter initialization, Baum–Welch expectation maximization (EM) algorithm is used to optimize emission and transition parameters. Once EM parameter calculation has completed, identifying the optimal segmentation is done using the Viterbi partitioning. It should be noted that while hidden states are set to correlate with copy numbers (i.e. state zero correspond to CN=0), they do not represent the final CN assignments as the actual CNV identification is carried out at the variant calling stage.

5.5.3 Common variants and output format

There is also an option to pass a BED file (**via --custom-parameters=CanvasPartition,-c=.bed** or via **--custom-parameters=CanvasPartition,--commoncnvs**=.bed) with common variant regions that will be represented as separate segments during Canvas partition stage. The input bed file should only contain non-overlapping intervals. This option will trigger the following behavior:

- Segmentation breakpoints spanning regions of common variants will be skipped
- A common variant region with either start or stop position more than distanceThreshold (10kb by default) away will be skipped
- A common variant region can be merged with adjacent segment if they share similar mean coverage

A processed annotation BED file (commoncnvs_*.bed) containing a list of common variants is available for the standard *Homo sapiens* reference genomes. These variants were derived from the analysis of population-wide common variants from over 2,000 germline samples. Canvas common CNV regions are based on https://git.illumina.com/Bioinformatics/cascadia-data/blob/master/copy-number/final_events.bed and were derived by applying the following filtering steps:

- overlapping variants were merged
- variants less than 5kb were filtered out

CanvasPartition outputs a gzipped-BED-like file (**G1_P1.partitioned**) containing chromosome, genomic start, genomic stop, bin coverage and an integer label that identifies which partitioned region the bin belongs to. For illustration, the first five lines of one such file are shown below:

```
chr1    754148  755200  90      1
chr1    755200  756030  93      1
chr1    756030  756783  90      1
chr1    756783  757489  89      1
chr1    757489  758242  80      1
```

5.6 CanvasSNV

CanvasSNV step is used to estimate SNV allele frequencies (aka minor allele frequencies, MAF). It takes as input (1) the variant calls (.vcf file) for the normal/reference sample or path to

a dbSNP / reference vcf file, and (2) the aligned reads (.bam file) for the tumor/reference sample.

For the tumor/normal workflow, germline SNV calls from the normal sample are used to identify all heterozygous SNVs (GT = 0/1 or 1/0) that pass filters (PASS in the FILTER column). We also require that the GQX tag (if present) is >= 30. For germline calling, CanvasSNV can use either the germline SNV calls, or has the option to review all dbSNP sites - see Resequencing Workflow Software Design Document)

For each SNV site, CanvasSNV computes how many observations were made for the REF and ALT alleles. This calculation considers all reads which are not flagged as PCR duplicates, secondary or supplementary alignments. Next counts (for all reads whose alignment to the reference covers a site of interest) of how many reads have the reference base, and how many reads have the alternative base are made.

Given A and B reads for the two alleles being considered (where A+B>0), we define the minor allele frequency (MAF) as min(A, B) / (A + B). Note that MAF will be approximately 0.5 for diploid regions. (In practice, the MAF at diploid positions will be a bit less - the higher the coverage, the closer to 0.5). For regions with a given **ploidy** (copy number and major allele count), we can predict the minor allele frequency of SNVs in the region:

- Ploidy A: MAF 0
- Ploidy AB (normal): MAF 0.5
- Ploidy AA (copy-neutral LOH): MAF 0.5
- Ploidy AAB: MAF 0.33333
- Ploidy AAA: MAF 0
- Ploidy AABB: MAF 0.5
- Ploidy AAAB: MAF 0.25
- Ploidy AAAA: MAF 0
- (etc.)

The output of CanvasSNV is a temporary file **VFResultsG1_P1.txt.gz**. It is a gzipped tab-delimited file containing one header line (beginning with #), an where each record contains the following fields: chromosome name, position, reference allele, alternate allele, # of reads for the reference allele, # of reads for the alternate allele. Here are few example lines from such as file:

```
#Chromosome Position Ref Alt CountRef CountAlt
chr1 12783 G A 53 135
chr1 13116 T G 24 70
chr1 13118 A G 25 69
chr1 14907 A G 73 102
chr1 14930 A G 48 98
chr1 15190 G A 160 13
chr1 16298 C T 18 54
```

The output of CanvasSNV also includes a temporary file **VFResultsG1_P1.baf**. It is a comma-delimited file containing one header line (beginning with #), an where each record contains the following fields: chromosome name, position, b-allele frequency (BAF). Here are few example lines from such as file:

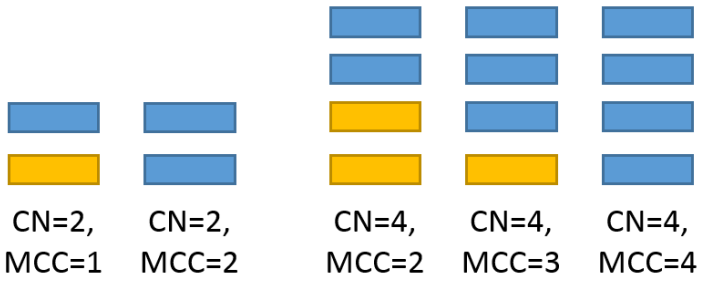
```
Chromosome,Position,BAF
chr22,10559508,0.192307692307692
chr22,10672164,0.5
chr22,10673144,0.947368421052632
chr22,10684424,0.772727272727273
chr22,10684460,0.444444444444444
```

5.7 CanvasDiploidCaller

The final step of the Canvas workflow for calling germline CNV variants is to assign discrete copy number states to each region identified by CanvasPartition and to determine variant's major copy number (MCC, copy number of major allele) where possible. CanvasDiploidCaller takes coverage and MAF values for each CanvasPartition segments as input. It uses Gaussian Mixture model with a separate component for each copy number and MCC states: e.g. for CN=2 model has two states representing diploid case (MCC=1, CN=2) and LOH (MCC=2, CN=2). Means and covariance matrices are estimated from the data for the diploid model and then adjusted via Expectation Maximization (EM) algorithm. For example, if the mean, μ, and standard deviation, σ, are estimated to be 100 and 15 in the diploid model, then the corresponding estimates in the haploid model would be initialized to μ/2 and σ/2. Following EM fitting, [CN,MCC] model tuple with the highest posterior probability is used to assign copy number to each segment. In cases where a segment has no spanning SNPs and hence empty MAF values, LOH states are skipped and not considered.

The diagram below illustrates CN and MCC values for several examples. For simplicity, assume all these examples are on an autosome. From left to right:

- Wild-type: Two copies, one maternal and one paternal
- Copy-neutral LOH: Two copies of the same chromosome
- Copy number 4 without LOH (2 copies of each chromosome)
- Copy number 4 with partial LOH (3 + 1)
- Copy number 4 with full LOH (four copies of one chromosome)



Following assignment of copy number states, neighboring regions that received the same copy number call are merged into a single region. This can occur if CanvasPartition over-segments a region.

Then Phred-like *q*-scores are assigned to each region. A model has been trained on a test corpus of ~10 curated germline samples to compute the probability that each segment has the correct direction - i.e. true copy number is 2 and it was assigned copy number 2, true copy number is <2 and it was assigned copy number <2, or true copy number is >2 and it was assigned copy number >2. A logistic regression model was trained to compute this probability. The three features used are:

- *LogBinCount* $\log_{10}$(1+number of bins). Segments with more bins can be assigned a copy number more confidently
- *ModelDistance*: Distance to the model. (Lower distances reflect a better fit)
- *DistanceRatio:* Ratio between between the distance to the best model point, and the distance to the next runner-up. The larger this value, the greater the confidence that the assigned ploidy is indeed correct.

The probability is then scaled to generate a Phred-scaled likelihood that the call is correct. This estimate is likely to be conservative as it is derived from array CGH data which is likely not 100% accurate. Importantly, *q*-scores are a function of number of bins, not genomic size, so

they are applicable to experiments of any sequencing depth, including low-depth cytogenetics screening.

The coordinates of non-diploid regions and their *q*-scores are output to a VCF file. Two filters are applied to PASS variants. First, a variant must be have a *q*-score of Q10 or greater. Second, a variant must be of size 10kb or greater.

5.8 CanvasPedigreeCaller

Small pedigree variant calling and genotyping module performs multi-sample CNV calling and the identification of de novo variants when pedigree information is available. As in the case of HMM partitioning, multivariate negative binomial distribution is used to model coverage data. However, in addition to bin depth, a depth of each allele at SNV sites is used along with the segmentation results. For the variant calling and de-novo scoring within a pedigree the following algorithm is used to estimate copy number of each segment.

1) Estimate median bin coverage for each segment. Determine whereas to use either median bin coverage or median allele coverage at SNV positions when fitting negative binomial model. To use median allele coverage, the segment should contain at least 4 SNPs and have1kb the spacing between adjacent variants. The latter constraint is set to prevent incomplete segmentation from interfering with alleles counts. For example, assume a small 2kb HET deletion that contains two hemizygous SNPs with a 100bp of flanking false positive segments on both sides of the true breakpoints. These diploid segments in turn each contain a variant. Allele counts from these diploid variants might lead to incorrect median allele coverage and a false negative diploid call for the segment. The choice of parameters was guided by the performance on training and test data.

2) Create a pedigree-consistent set of copy number assignments given a maximal copy number threshold CN_{max} (set to 6 by default):

- When using median allele coverage, derive all possible combinations of genotypes with the constraint that the total copy number is equal to or less than MAXCN for each parent. Next, derive a set of inherited genotypes and possible de-novo calls with de novo mutation rate R_{denovo} (set by default to 0.00001):
- For segments where allele depth information is not available, for each CN assignment in parents, a set of all CN states in offerings that are compatible with CNs in parents are enumerated.

In each case a pedigree probabilistic model M_p is fitted to the depth data to estimate likelihood (*L*) of copy number (*CN*) assignments within a pedigree given coverage (*D,* either total or allele) as

$$L(CN_M, CN_F, CN_C / D) \sim P(D_M/CN_M)\, P(D_F/CN_F)\, P(D_C/CN_C) \times P(CN_C/CN_M, CN_F)$$

, where the last term represents Mendelian transmission probabilities or the estimated de novo rate of CNVs R_{denovo}. Copy numbers from the model with the highest likelihood M_{pMAX} are used to make final CNV calls.

3) Estimate quality scores (Q-score and DQ-score) for variant calls in each sample. Q-scores are obtained by calculating marginal probabilities $P(D_j/CN_{i \neq ALT}) \in M_{pMAX}, j =$ {parents, offspring} and i = {0,1,2,3,4,5} of copy number variant not been ALT. DQ-scores represent probability of CN=2 in parents for trios (adding CN=2 in a sibling for quads) given an ALT call in a proband.

4) With no pedigree information, samples are considered IID (independently and identically distributed). Variant calls are made independently for each sample reducing the terms of the pedigree model to $L(CN_M, CN_F, CN_C / D) \sim P(D_M/CN_M)\, P(D_F/CN_F)\, P(D_C/CN_C)$.

The primary output from Canvas SPW are multi-sample and single-sample .vcf file compliant with the version 4.1. The copy number (CN) and major chromosome count (MCC) are indicated in the per-sample data along with variant q-score (QS). De novo calls also have a DQ format field that represents a de novo Phred-scaled quality score and dq20 filter flag for variants with

novo quality score above 20. An example lines representing de novo variant calls with PASS filter and dq20 flag:

```
12      29851548        Canvas:LOSS:12:29851549-29853545        N
<CNV>   19.35   PASS
SVTYPE=CNV;dq20;END=29853545;CNVLEN=1997;CIPOS=-466,458;CIEND=-540,458
RC:BC:CN:MCC:QS:DQ     86:2:1:0:23.33:12.12
```

For multi-sample .vcf file the following rules are applied to assign PASS filter flag:

- PASS variant if proband is PASS (with pedigree information)
- If at least once sample has PASS (with no pedigree information)
- Both parents and a proband are PASS (for de novo variants)

5.9 CanvasSomaticCaller

CanvasSomaticCaller is used in tumor-normal and somatic workflows to identify regions with an abnormal copy number (e.g. for autosomes the copy number != 2 or b-allele frequency != ~50%). It does this by accounting for any changes in ploidy, normal contamination (purity) and polyclonality. The correction logic for these confounding factors is implemented in somatic model, key principles of which are:

Principle 1: For a given set of ploidy and purity values one can derive **expected** values of coverage and MAF **for each copy number state.**

Definition: Model Deviation = (expected – observed) coverage and MAF values for a given set of ploidy and purity values.

Principle 2: For two models with similar deviations, the model that implies a smaller number of CN changes (i.e. closer to diploid) should be preferred.

Principle 3: Given training samples with known ploidy/purity values, optimization techniques can be used to infer the best feature weights to use for fitting a particular somatic model.

Principle 4: Given most likely purity and ploidy values, for a given segment, departure from **observed coverage and MAF values** can be indicative of segment's polyclonality

As mentioned, CanvasSomaticCaller attempts to estimate overall variant heterogeneity if the tumor is polyclonal. While finding ploidy/purity prediction is linear in the number of states, including heterogeneity makes model complexity exponential. CanvasSomaticCaller uses a number of approximations (described below) to detect polyclonality.

CanvasSomaticCaller is composed of a number of steps.

5.9.1 Input and pre-processing

1. Load coverage segments identified by CanvasPartition
2. Load variant frequency information captured by CanvasSNV; retain all sites with ref+alt coverage >= 10. CanvasSNV output contains only variant frequencies of heterozygous SNVs if the VCF of a matched normal sample was used. However, a dbSNP VCF can be used in place of a normal VCF and hence not all positions in CanvasSNV are heterozygous. In this case, the command line argument -d (--dbsnpvcf) should be used to prune MAFs of homozygous positions for each segment.
    a. We filter the zero MAFs for a segment if more than 10% of MAFs are non-zero or at least one of the positions has a MAF >= 0.2. This is because homozygous positions are expected to have MAFs of zero.

   b. When filtering is done, we further take only MAFs > 0.1 if there are any MAFs > 0.1 in the segment. This is because some homozygous positions have non-zero MAFs that are very close to zero.

   c. If the number of MAFs before filtering is greater than the given MAF count threshold (default: 50) but less than this threshold after filtering, we lower the threshold to be the number of MAFs after filtering.

3. For each segment identified by CanvasPartition, note the median number of counts per bin ("coverage") and the median minor allele frequency (MAF).

   a. The number of MAFs of each segment must meet the MAF count threshold (default: 50) to be used in model fitting. However, the MAF count threshold may be relaxed so that we have at least 20 segments for model fitting.

   b. Each segment is assigned a weight, which is used in computing the deviation described below. The weight is the length of the segment if the total number of segments is greater than 100. Otherwise, the weight is the number of bins in the segment. In case we have fewer than 10 MAFs in the segment, the weight is multiplied by the ratio of the number of MAFs over 10. This is because a smaller number of MAFs will provide a less reliable estimation of the true median MAF for the segment. This weighting technique makes model fitting more robust to noise.

5.9.2 Somatic model fitting

1. Model the median diploid coverage **D** and the overall purity **p**: Given a candidate tuple **(D, p)**, compute the expected (MAF, Coverage) points for pre-specified set of ploidy and purity values. Loop over a collection of all segments of reasonable size (>50kb) and informative numbers of SNVs (>=50). The **model deviation** for (D, p) is the sum, across all segments, of ||(SegmentMAF, SegmentCoverage) - (ExpectedMAF, ExpectedCoverage)||. The final **total deviation** is derided by combining **model deviation** and **penalty term** that includes additional coefficient-weighted factors as follows:
   +*W1**DeviationScore: Model deviation
   +*W2**DiploidDistance: Scaled number of CN events needed to transform a given genome into diploid
   +*W3**CN2: Percentage of CN = 2
   +*W4**PercentNormal: Percentage of normal CNs
   To derive the formula coefficients *W1, W2, W3* and *W4*, a model training workflow used ~140 annotated tumour samples was used as described in https://git.illumina.com/Bioinformatics/CNAT/tree/master/CanvasSomaticTrainingData.

2. The model with the lowest score is then selected as a **preliminary best model of overall coverage and purity.**

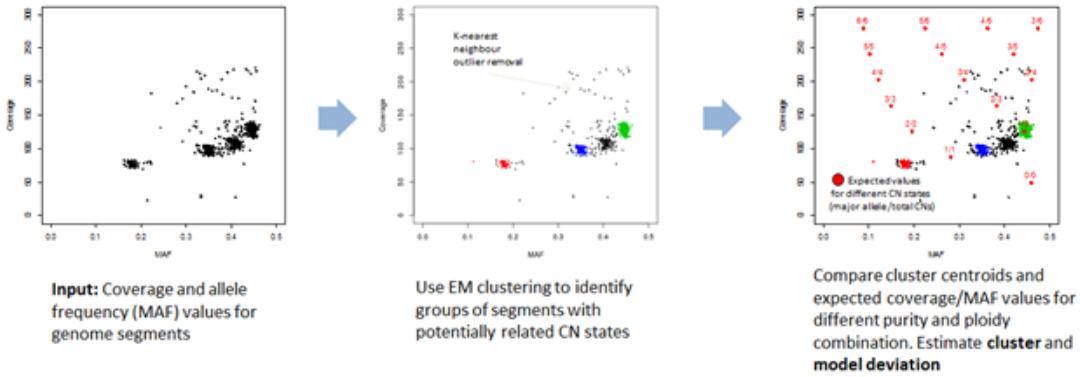
3. Preliminary calling: For each segment, identify the ploidy which best fits the MAF and coverage of that segment, according to our model of overall coverage of purity.

4. A **merge** step is carried out. Adjacent segments with same copy number are combined into a single segment, and segments with size under a threshold (currently 50000 bases) are merged into the best (highest q-score) adjacent segment. Segments are considered adjacent if they are on the same chromosome, have no other segments between them, and have no forbidden intervals (from the filtering .bed file) between them)

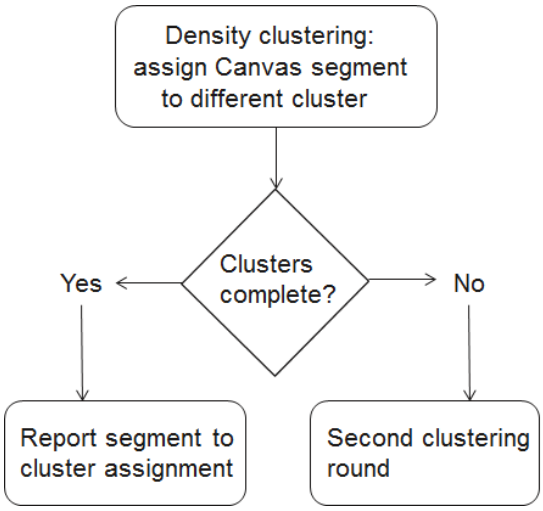
5.9.3 Detection of polyclonal variants

1. Somatic caller also attempts to account for **tumor polyclonality** by implementing Principle 4 described above. The assumption used by CanvasSomaticCaller is that observed coverage and MAF values from polyclonal regions will tend to group into a limited set of clusters. Each segment tuple (SegmentMAF, SegmentCoverage) is then assigned to these distinct clusters using non-parametric density clustering (Rodriguez A., 2014) by default (--custom-parameters=CanvasSomaticCaller ,CanvasSomaticClusteringMode=Density) or

Gaussian Mixture Expectation-Maximization clustering (--custom-parameters=CanvasSomaticCaller ,CanvasSomaticClusteringMode=GaussianMixture). **Cluster deviation** for cluster (D, p) is defined as the sum of distances between segments belonging to the cluster and the closest expected (MAF, Coverage) tuple for a given ploidy and purity model as calculated above. In general, clusters located farthest from expected (MAF, Coverage) centroid are more likely to be heterogeneous**. The adjusted total deviation is than defined as the average of total (model) deviation and** cluster deviation. Pictorial example of this approach is shown in figure 1.



Sometimes due to highly non-spherical nature of clusters (i.e. when several heterogeneous states exist between two copy number states), density clustering under-segments highly heterogeneous clusters. In such a situation a second round of density clustering is carried out on these under-partitioned clusters leading to a clustering logic as shown below.



2. Best somatic purity and ploidy model found in step 1 is used to identify polyclonal variants and correct CN assignment if necessary (from CN=2 to CN=1 or CN=3). The correction is done by finding the variants that satisfy following requirements:
    a. most likely CN=2
    b. second most likely CN=1 or 3
    c. best model purity > 0.2
    d. variant segment is assigned to heterogeneous clusters
    e. ratio of segment distance to the best and runner-up model is above threshold W5
3. Assign Phred-like *q*-scores to each region. A model has been trained on a test corpus of 33 curated tumor/normal samples to compute the probability that each segment has the correct direction - i.e. true copy number is 2 and it was assigned copy number 2, true copy number is <2 and it was assigned copy number <2, or true copy number is >2 and it was assigned

copy number >2. A logistic regression model was trained to compute this probability. The three features used are:

a. *LogBinCount* $\log_{10}$(1+number of bins). Segments with more bins can be assigned a copy number more confidently

b. *ModelDistance*: Distance to the model. (Lower distances reflect a better fit)

c. *DistanceRatio:* Ratio between between the distance to the best model point, and the distance to the next runner-up. The larger this value, the greater the confidence that the assigned ploidy is indeed correct.

The probability is then scaled to generate a Phred-scaled likelihood that the call is correct. Note that this q-scoring model uses the same features as the model for germline calling (CanvasDiploidCaller), but re-trained on tumor/normal data. The overall ROC curve area on the test data was 0.829. (Note that truth data is not available, so some of the "false positives" may reflect correct copy number calls which have not been annotated correctly in the truth data. Ten of the highest-scoring "bad" calls on the training data were assessed to ensure they are plausible copy number calls)

5.9.4 Estimate purity from somatic SNVs

- If somatic SNV calls are available (typically generated by Strelka as part of the tumor/normal workflow), then the tumor purity is also estimated from these SNV calls. We take all SNV calls which pass filters and which have a variant allele frequency <0.5; the estimated tumor purity is twice the mean variant allele frequency of these somatic SNVs. This estimate is written to the Canvas log file along with CNV based purity estimate (line "Purity estimates: X from CNVs, Y from SNVs"). SNVs-estimated purity values are used in the *EstimatedTumorPurity* VCF flag if the following conditions are met:
    - Canvas did not assign many CNVs (93% or more of the genome is at copy number 2)
    - CNV-estimated purity is less than 0.5 (50%)

  In tests across several data-sets (HCC1187 admixture data, Gel35, Gel9), the CNV-derived and SNV-derived tumor purity estimates agreed within 3%.

5.9.5 Detecting samples with choppiness coverage profile

Samples with significantly degraded DNA can exhibit a characteristic choppiness profile (see figure below) even in regions of diploid coverage after GC-content normalization. This can lead to a large number of false positive calls. As no significant correlation was found between choppiness and characteristics of genomic regions (GC content, repeats), these patterns so far can not be corrected by normalization procedures. Canvas CheckNonDiploidMAFs module within CanvasSomaticCaller attempts to identify these regions by using an observation that choppy coverage still leads to a normal distribution of B-allele ratios. Hence by testing whereas CN=1 calls have a distribution of B-allele ratios around 0.5 (a feature of CN=2 regions) it would be possible to identify these coverage anomalies. Canvas uses a one-way t-test with equal variances to estimate t-statistics (for the null hypothesis test that B-allele ratios for CN=1 equal 0.5). If the null hypothesis can not be rejected, Canvas sets ploidy and purity model to 100% purity and ploidy=2. This is done to ensure that for low purity samples choppiness doesn't interfere with true CN calls.

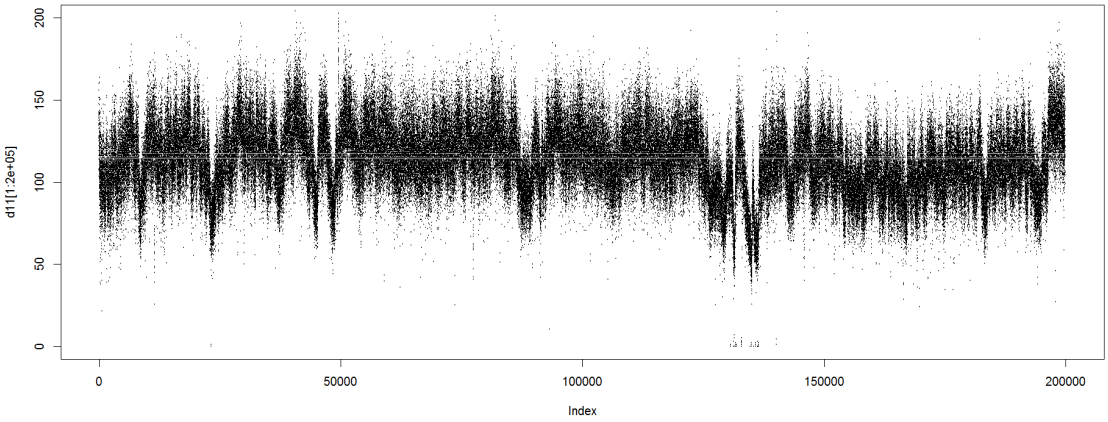


Figure. Example of choppiness at the beginning of small arm for chromosome 1, coverage represents output of CanvasClean

The main output of CanvasSomaticCaller is a variant call file (.vcf 4.1 format) listing all intervals which were assigned a copy number. See CanvasSomaticCaller outputs, below.

Copy numbers are reported as a REF, LOSS, or GAIN depending on the copy number count and expected count. The expected count is generally 2 (copy number 2 is REF, copy number 0 or 1 is LOSS, copy number 3 or greater is GAIN). The exception is that if a ploidy .bed file is passed in to Canvas, it can update the expected number of copies for particular chromosomes. This feature is used in the analysis workflow, based on the normal allosome (sex chromosome) copy count reported in the SNP/Indel .vcf header line ##ExpectedSexChromosomeKaryotype. For instance, for chrX and chrY on an XY sample, copy number calls of 2 on X or Y would be called as a GAIN, with the exception of the PAR region on chrX which still has expected copy number 2.

5.10 Parameter training

To facilitate data-driven parameter optimization, all parameters related to model fitting of CanvasSomaticCaller are stored in **SomaticCallerParameters.json** configuration file (under CanvasSomaticCaller subproject). optimizeSomaticCanvasModel.py training workflow takes this configuration file and a set of training samples as an input and attempts to optimize parameters given the data. The training data andscript reside under https://git.illumina.com/Bioinformatics/CNAT repository. The training procedure is fully described in https://git.illumina.com/Bioinformatics/CNAT/blob/master/README.md

5.11 Module operations

As exemplified in the previous sections, Canvas comprises different executable components wrapped into easy-to-use workflows for different sample types and enrichment / whole genome cases. While they can be invoked through a single command line to generate a set of variant call without any further interventions, users can also run individual components to generate intermediate files or to rerun a particular module with different parameter settings.

Below is an example command for CanvasBin

CanvasBin

```
mono CanvasBin.exe
>>>Command-line arguments:
CanvasBin [version number]
Please specify the Canvas k-uniqueness reference file.
```

```
Usage: CanvasBin.exe [OPTIONS]+
Bin alignments into variable-sized genomic intervals.
Options:
  -b, --bam=VALUE            bam file containing unique alignments
  -r, --reference=VALUE      Canvas-ready reference fasta file
  -c, --chr=VALUE            for bam input, only work on this
chromosome.
                               Output intermediate binary data. Must
follow-up
                               with a single CanvasBin call passing
all the
                               intermediate binary data files (see -i
option)
  -i, --infile=VALUE         intermediate binary data file from
individual
                               chromosome. Pass this option multiple
times,
                               once for each chromosome
  -f, --filter=VALUE         bed file containing regions to ignore
  -d, --bindepth=VALUE       median counts desired in each bin
  -z, --binsize=VALUE        bin size; optional
  -o, --outfile=VALUE        text file to output containing computed
bins, or
                               if -c option was specified the
intermediate
                               binary data file to output
  -y, --binsizeonly          calcualte bin size and exit
  -h, --help                 show this message and exit
  -p, --paired-end           input .bam is a paired-end alignment
(e.g. from
                               Isaac)
  -m, --mode=VALUE           coverage measurement mode
  -t, --manifest=VALUE       Nextera manifest file
  -n, --bins=VALUE           bed file containing predefined bins
```

Default parameters at the individual modules step can be passed on to the main executable by invoking --custom-parameters=[module_name],-m=[parameter_name]. For example, to enable fragment-based GC content normalization beneficial for processing FFPE samples, CanvasBin parameter “mode” can be altered through command line argument of the main Canvas executable as follows: --custom-parameters=CanvasBin,-m=TruncatedDynamicRange.

In addition to running individual Canvas modules, Canvas binary allows running a subset of modules by utilizing checkpoints via -c and –s arguments:

```
  -c=VALUE                   continue analysis starting at the
specified checkpoint. VALUE can be the checkpoint name or number
```

```
  -s=VALUE                    stop analysis after the specified
checkpoint is complete. VALUE can be the checkpoint name or number
```

Canvas modules have the following checkpoint numbers: 1 for CanvasBin, 2 for CanvasClean, 3 for CanvasPartition, 4 for CanvasSNV and 5 for variant calling (Somatic or Germline). Omitting –s will execute Canvas until the final step in the workflow has completed. For example, specifying -c 3 in the whole-genome somatic workflow will run CanvasPartition, CanvasSNV and CanvasSomaticCaller. This could be advantageous if users want to rerun Canvas with different partitioning algorithm without the need to repeat time consuming I/O or coverage binning operations.

Ease of module reuse is further expanded by the fact that Canvas intermediate files are saved as bed files, facilitating access by third-party software such as bedtools.

6 Somatic Outputs

6.1 Vcf file

The primary output from CanvasSomaticCaller is a .vcf file, generally named SampleName_S1.CNV.vcf. (Later in the tumor/normal workflow, this output gets merged together with output of the SV caller, Manta, into a combined vcf file SampleName_S1.SV.vcf).

The .vcf file header includes the estimated tumor purity (1.0 for pure tumor), the estimated overall ploidy (mean copy number across all bases; 2.0 for normal data), tumor heterogeneity proportion and the estimated chromosome count (total number of chromosomes weighted by their copy number). Both OverallPloidy and EstimatedChromosomeCount are (as of 1.7.0) based on all PF calls (records with PASS in the FILTER column in the .vcf file). Example:

```
##EstimatedTumorPurity=0.28
##OverallPloidy=2.45
##EstimatedChromosomeCount=61.91
##HeterogeneityProportion=0.02
```

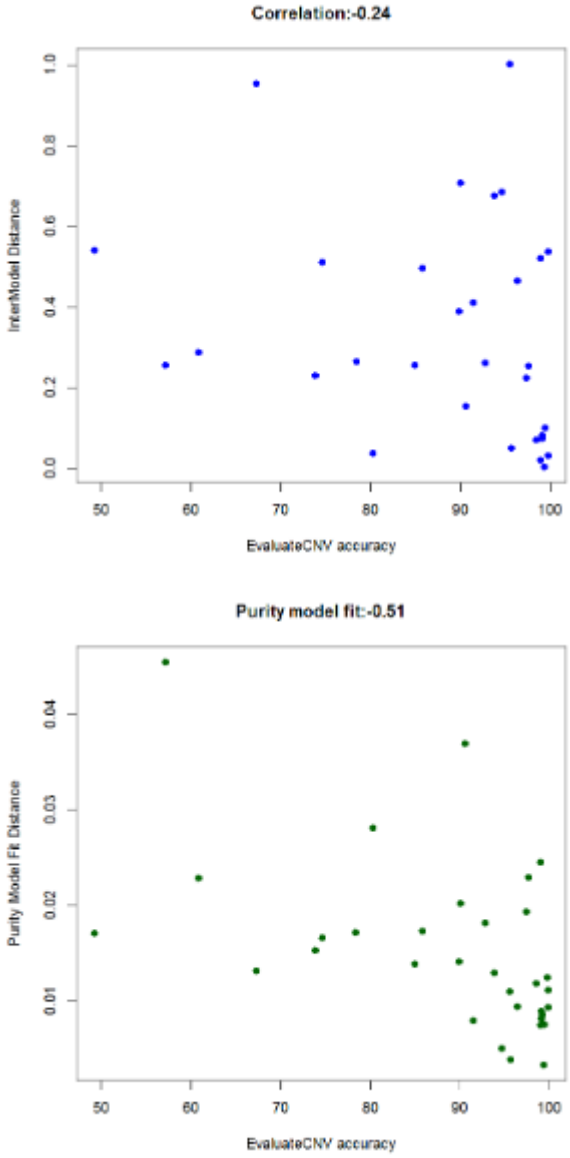
The header also includes an indication of the goodness of fit of the coverage/purity model to the data; lower values are better and should indicate more reliable calling overall:

```
##PurityModelFit=0.0259
##InterModelDistance=0.6853
```

The header line also includes the estimated diploid coverage which can be used to covert the hit count within a bin to a copy number:

```
##DiploidCoverage=60.94
```

PurityModelFit is defined as a total deviation D of the best model. InterModelDistance shows genome edit distance between the best model and other top models (defined by MaximumRelatedModels). The premise is that if the top models provide widely different genome baseline (leading to high interModelDistance), the overall modeling approach might be more unstable. Distribution of these parameters on the https://git.illumina.com/Bioinformatics/CNAT/tree/master/CanvasSomaticTestData is shown below.



PurityModelFit is more predicative of overall model quality than InterModelDistance. By selecting 70% genomewide accuracy as a threshold for high vs, low quality predictions, at inter model distance above 0.8 false discovery rate (FDR) is 50% and below 0.8 - 10%. While few observation with very high inter-model distance exist, **InterModelDistance threshold of 0.8** seem to be discriminative of poor model fit. PurityModelFit provides even better resolution - at the same 70% genomewide accuracy cut-off, and **PurityModelFit threshold of 0.022** leads to FDR of 7%. These metric value cut-offs are therefore recommended for use when assessing quality of the Canvas somatic model fit.

`Header also contains a`

`##LocalSDmetric`

`that assesses genome-wide coverage spikiness associated with FFPE` sample prep biases. Details of the metric derivation are available at [FFPEde-noisingalgorithm](#) .

The copy number (CN) and major chromosome count (MCC) are indicated in the per-sample data. Below are three sample lines: The first is for a region of LOH (copy number 2, major chromosome count 2); the second is for a copy number variant (copy number 4, major chromosome count 2); the third is for a reference call (copy number 2, major chromosome count 1).

```
1 753815 Canvas:GAIN:1:753816:813096 N <CNV> 10 PASS
SVTYPE=LOH;END=813096 RC:BC:CN:MCC 87:18:2:2

3 26664138 Canvas:GAIN:3:26664139:32101159 N <CNV> 61 PASS
SVTYPE=CNV;END=32101159 RC:BC:CN:MCC 118:6520:4:2

5 1000000 Canvas:REF:3:1000000:2000000 N . 61 PASS END=2000000
RC:BC:CN:MCC 118:6520:2:1
```

One minor note on formatting: According to the VCF spec, if any of the ALT alleles is a symbolic allele (an angle-bracketed ID String “<ID>”) then the padding base is required and POS denotes the
coordinate of the base preceding the polymorphism. So, for non-reference calls (i.e. records where ALT is not equal to .), the POS column stores the base **before** the variant.

The output .vcf file includes (in Canvas 1.14 and newer) the CIPOS and CIEND info tags, providing a confidence interval for the POS end END coordinates using relative positions. For example, if POS=12345 and CIPOS=-10,20 then the true start is somewhere between 12345-10 and 12345+20. Confidence intervals are set using the midpoints of bins at the start or end of a segment, using the working assumption that the true breakpoint falls in the half of the bin closest to a segment boundary.

- If the **last** bin in a segment has endpoints (start, end), then the first value for CIEND is (start-end)/2. The second value is either symmetric or (if the next segment is immediately adjacent) set to half the size of the next bin.
- Similarly, the second value for CIPOS is set using the endpoints of the **first** bin in a segment. The first value for CIPOS is either symmetric or (if the next segment is immediately adjacent) is set using half the size of neighboring bin from the previous segment.

6.2 Additional outputs

CNV-CoverageAndVariantFrequency.txt - Summarized coverage and MAF data across the genome, suitable for plotting. The columns for each row of this text file are:

1. chromosome
2. start position
3. end position
4. copy number
5. major chromosome count
6. median hits for bins in this region: each region is composed of multiple bins where the size of each bin is set so that it contains the same number of uniquely mapped positions in the reference genome. The MedianHits is the median number of reads that mapped to bins in this region so it will correlate with the copy number of the region. The bin sizes are also scaled so that the median number of hits per bin is a fixed value (default 100 for WGS). This means that a sample sequenced at different overall depth of coverage (e.g. 30x vs 90x) should have similar values for median hits within a given region.
7. normalized coverage: floating point copy number. A value of 2.0 represents copy number 2. Conversion from median hits above:

	normalized coverage = median hits * 2 / diploid coverage

   A rough approximation of sequencing depth is also possible using the mean coverage of the tumor sample:

	sequencing depth = normalized coverage * mean coverage / Overall Ploidy
8. median minor allele frequency

9. reference ploidy
10. extra columns providing a histogram of variant allele frequency (b allele frequency) for SNVs in this region. The collection of SNVs is typically taken from the collection of heterozygous SNVs in the normal .vcf file.

This table does not report coverage for regions where very few (or no) bins are available - the centromeres (and other intervals included in the filter .bed file for the reference genome) are excluded entirely from CNV calling, and so will not have any coverage or b allele frequency reported (these columns are left blank).

CNVModeling.txt - (MAF, Coverage) data points for segments and for the model. Tab-delimited file with two tables. The first gives the variant frequency and coverage coordinates (MedianMAF, MedianHitCount) for the model. The remaining table has one record for each segment used in model fitting. The values per segment are: MedianMAF, MedianHitCount, Deviation, chromosome, start, end, length, truth set copy number.

CNVConfusionMatrix.txt - available if (and only if) a truth set was specified up-front, for developer testing. Provides a confusion matrix with actual (truth set) copy number calls in the rows, and called copy numbers in the columns; the matrix entries are the number of bases with the corresponding (TruthSet, CanvasCall) copy numbers.

7 Regression tests

See the README from the CanvasTest git repository

8 References

Roller E., et al. (2016). Canvas: versatile and scalable detection of copy number variants. *Bioinformatics,* doi: 10.1093/bioinformatics/btw163.

Olshen AB V. E. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5** , 557-572.

Fryzlewicz P. (2007). Unbalanced Haar Technique for Nonparametric Function Estimation. *Journal of the American Statistical Association,* **102**, 1318-1327.

Rodriguez A. et. all (2014). Clustering by fast search and find of density peaks. *Science*, **6191**, 1492-1496.

Rousseeuw P. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*. **20.** 53-65.

Skellam J. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations, *Journal of the Royal Statistical Society*, **109**, 296.

Venkatraman ES, O. A. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **31** , 657-663.

normalized coverage = median hits * 2 / diploid coverage