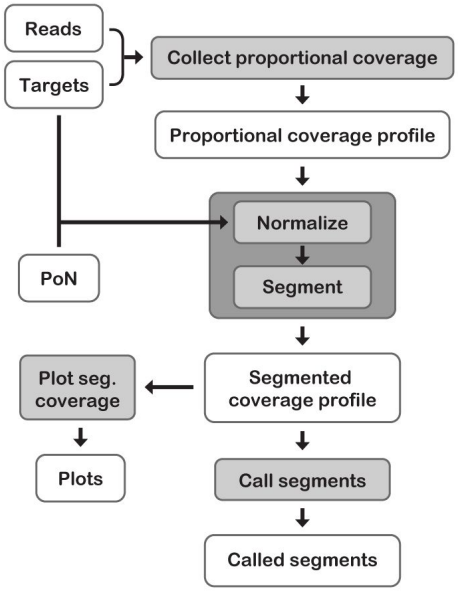
Call somatic copy number variants using GATK CNV

About the workflow



This workflow detects *somatic* copy number variation using a panel of normals (PoN). The workflow is optimized for Illumina short-read whole exome sequencing (WES) data. Our team is working on optimizing application to whole genome sequencing (WGS) data and also a separate CNV workflow for germline calling.

The underlying algorithms and current workflow options, e.g. syntax, may change during development. The presented basic approach and general concepts will still be germaine. Please check the GATK website (https://software.broadinstitute.org/gatk/) for updates. For detailed examples of workflows, see (http://gatkforums.broadinstitute.org/gatk/discussion/6791). For the mathematics behind the workflow, see (https://github.com/broadinstitute/gatk-protected/blob/master/docs/CNVs/CNV-methods.pdf).

The workflow uses the following tools from the *gatk-protected-1.0.0.0-alpha1.2.3* pre-release (https://github.com/broadinstitute/gatk-protected/releases/tag/1.0.0.0-alpha1.2.3). Note other tools in this release may be unsuitable for analyses.

GATK CNV tools		page
1. **CalculateTargetCoverage**	This is pre-computed for you in this tutorial.	**2**
2. **CombineReadCounts and CreatePanelOfNormals**	Creates the Panel of Normals (PoN) using data from (1).	**3**
3. **NormalizeSomaticReadCounts**	Uses data from (1) and (2).	**4**
4. **PerformSegmentation**	Uses data from (3).	**4**
5. **PlotSegmentedCopyRatio**	Optional visualization step uses data from (3) and (4).	**5**
6. **CallSegments**	Uses data from (3) and (4).	**6**

About the example data

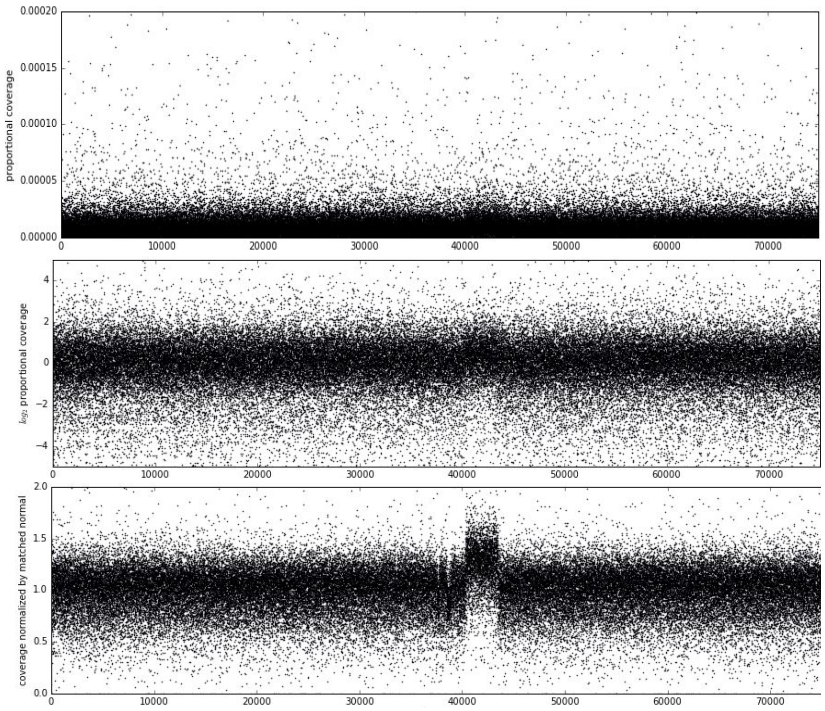
We have whole exome capture sequence data for chromosomes 1–7 of matched normal and tumor samples. Because the data is from real cancer patients, we have anonymized them at multiple levels. The anonymization process preserves the noise inherent in real samples. The data is representative of Illumina sequencing technology from five years ago (2011).

1. Collect proportional coverage using target intervals and read data.

This first step is done for you ahead of time and the command is here for your reference.
Process each BAM, whether normal or tumor. The tool collects coverage per read group at each target and divides these counts by the total number of reads per sample.

```
bYnY ÈbYj _YlcýÇbYj #Yd[mdYl]4Yj_]l#gn]jY_] ê
      È) ĒafhmlëZYeë^ad]ē ê
      È4 ĒafhmlëlYj_]lëlknē ê
      ÈljYfk^gje 0#/6 ê
      È_jgmh"q 3!-0,% ê
      ÈlYj_]l)f^g &5,, ê
      Ùc]]h\mhk ê
      È/ Ēgmlhmlëh[gnë^ad]ē
```

The target file (**È4**) is a *padded* intervals list of the baited regions. You can add padding to a target list using the PadTargets tool. For us, padding each exome target 250bp on either side increases sensitivity. The **ÈlYj_]l)f^g &5,,** option keeps the original target names from the target list. The **Ùc]]h\mhk** option asks the tool to include alignments flagged as duplicate.



The top plot shows the raw proportional coverage for our tumor sample for chromosomes 1–7. Each dot represents a target. The y-axis plots proportional coverage and the x-axis targets.

The middle plot shows the data after a median-normalization and log2-transformation.

The bottom plot shows the tumor data after normalization against its matched-normal.

→ For each of these progressions, how certain are you that there are copy-number events? How many copy-number variants are you certain of? What is contributing to your uncertainty?

---

2. Create the CNV panel of normals (PoN) with two commands.

The normals should represent the same sequencing technology, e.g. sample preparation and capture target kit, as that of the tumor samples under scrutiny. The PoN is meant to encapsulate sequencing noise and may also capture common germline variants. Like any control, you should think carefully about what sample set would make an effective panel. At the least, the PoN should consist of ten normal samples that were ideally subject to the same batch effects as that of the tumor sample, e.g. from the same sequencing center. Our current recommendation is 40 or more normal samples. Depending on the coverage depth of samples, adjust the number.

→ *What is better, tissue-matched normals or blood normals of tumor samples?*
→ *What makes a better background control, a matched normal sample or a panel of normals?*

The first step combines the proportional read counts from the multiple normal samples into a single file. The **`Èafhml,akl`** parameter takes a file listing the relative file paths, one sample per line, of the proportional coverage data of the normals.

```
bYnY ÈbYj _YlcýÇbYj #geZaf]2]Y\#gmflk ê
       Èafhml,akl fgjeYdkÇlpl ê
       È/ kYf\Zgp×[geZaf]\ÈfgjeYdkÇlkn
```

The second step creates a single CNV PoN file. The PoN stores information such as the median proportional coverage per target across the panel and projections of systematic noise calculated with PCA (principal component analysis). Our tutorial's PoN is built with 39 normal blood samples from cancer patients from the same cohort (not suffering from blood cancers).

```
bYnY ÈbYj _YlcýÇbYj #j]Yl]0Yf]d/^.gjeYdk ê
       È) kYf\Zgp×[geZaf]\ÈfgjeYdkÇlkn ê
       È/ kYf\Zgp×fgjeYdkÇhgf ê
       Èfg1# ê
       ÈÈ\akYZd]3hYjc ê
       ÈÈeafaeme4Yj_]l&Y[lgj0]j[]flad]4`j]k`gd\ þ
```

This results in two files, the CNV PoN and a **`lYj_]lëo]a_`lkÇlpl`** file that typical workflows can ignore. Because we have a small number of normals, we include the **`Èfg1#`** option and change the **`ÈÈeafaeme4Yj_]l&Y[lgj0]j[]flad]4`j]k`gd\`** to 5%.

→ *Based on what you know about PCA (http://setosa.io/ev/principal-component-analysis/), what do you think are the effects of using more normal samples? A panel with some profiles that are outliers?*

---

3. Normalize a raw coverage profile using the PoN.

We normalize the tumor coverage against the PoN's target medians and against the principal components of the PoN.

```
bYnY ÈbYj _YlcýÇbYj .gjeYdar]3geYla[2]Y\#gmflk ê
       È) [gn×lmegjÇlkn ê
       È0/. kYf\Zgp×fgjeYdkÇhgf ê
       È04. kYf\Zgp×lmegjÇhlfÇlkn ê
       È4. kYf\Zgp×lmegjÇlfÇlkn
```

This produces the *pre-tangent-normalized* file (**È04.Þ** and the *tangent-normalized* file (**È4.**), respectively. Resulting data is log2-transformed.

Denoising with a PoN is critical for calling copy-number variants from WES coverage profiles. It can also improve calls from WGS profiles that are typically more evenly distributed and subject to less noise. Furthermore, denoising with a PoN can greatly impact results for (i) samples that have more noise, e.g. those with lower coverage, lower purity or higher activity, (ii) samples lacking a matched normal and (iii) detection of smaller events that span only a few targets.

---

4. Segment the normalized coverage profile.

Segmentation groups contiguous targets with the same copy ratio.

```
bYnY ÈbYj _YlcýÇbYj 0]j^gje3]_e]flYlagf ê
       È4. kYf\Zgp×lmegjÇlfÇlkn ê
       È/ kYf\Zgp×lmegjÇk]_ ê
       È,/'
```

For our tumor sample, we reduce the ~73K individual targets to 14 segments. The **È,/'** parameter tells the tool that the input coverages are log2-transformed.

This command will error if you have not installed R and certain R components. If you need, take a few minutes to install R from <https://www.r-project.org/>. Then install the components with the following command.

```
2k[jahl afklYddë2ëhY[cY_]kÇ2
```

We include **afklYddë2ëhY[cY_]kÇ2** in the tutorial data bundle. Alternatively, download it from (https://github.com/broadinstitute/gatk-protected/blob/master/scripts/install_R_packages.R).

View the resulting file with **[Yl kYf\Zgp×lmegjÇk]_**.

Sample	Chromosome	Start	End	Num_Probes	Segment_Mean
tumor	1	1	249238921	17585	0.99647119
tumor	2	1	243193925	13159	0.995918783
tumor	3	1	21484649	1139	0.999307093
tumor	3	21502423	22485593	54	0.798851916
tumor	3	22503367	92855531	3740	0.997231251
tumor	3	92873305	96090725	174	0.816146473
tumor	3	96108499	109371107	705	0.994470169
tumor	3	109388881	115519763	326	0.816995479
tumor	3	115537537	141901787	1390	0.994263396
tumor	3	141919561	198008273	3016	1.146312186
tumor	4	1	191131706	6818	0.998268633
tumor	5	1	180899061	7849	0.999514915
tumor	6	1	171097489	8860	0.995090745
tumor	7	1	159125275	8166	0.996402123

→ *Which chromosomes have events?*

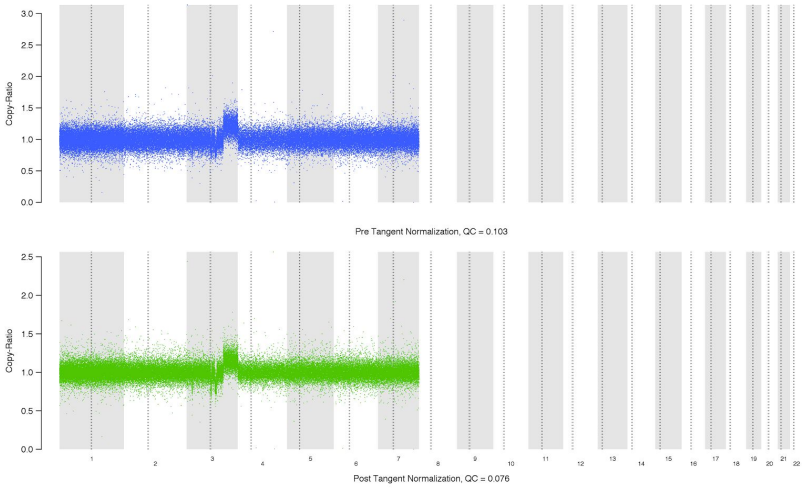
---

5. (Optional) Plot segmented coverage.

This command requires XQuartz <https://www.xquartz.org/> installation. If you do not have this dependency, then view the results in the precomputed_results *folder instead.* Currently plotting only supports human assembly b37 autosomes. Going forward, this tool will accommodate other references and the workflow will support calling on sex chromosomes.

```
bYnY ÈbYj _YlcýÇbYj 0dgl3]_e]fl]\#ghq2Ylag ê
      È4. kYf\Zgp×lmegjÇlfÇlkn ê
      È04. kYf\Zgp×lmegjÇhlfÇlkn ê
      È3 kYf\Zgp×lmegjÇk]_ ê
      È/ kYf\Zgp ê
      Èhj] lmegj ê
      È,/'
```

The **È/** defines the output directory, and the **Èhj]** defines the basename of the files. Again, the **È,/'** parameter tells the tool that the inputs are log2-transformed. The output folder contains seven files--three PNG images and four text files.



- **"]^gj]ë!^l]jÇhf_** (shown above) plots copy-ratios pre (top) and post (bottom) tangent-normalization across the chromosomes.

The plot automatically adjusts the y-axis to show all available data points. Dotted lines represent centromeres.

- **`"]^gj]ë!^l]jë#2ë,aeëýÇhf_`** shows the same but fixes the y-axis range from 0 to 4 for comparability across samples.
- **`&mdd']fge]Çhf_`** colors differential copy-ratio segments in alternating blue and orange. The horizontal line plots the segment mean. Again the y-axis ranges from 0 to 4.

→ *Open each of these images. How many copy-number variants do you see?*

Each of the four text files contain a single quality control (QC) value. This value is the median of absolute differences (MAD) in copy-ratios of adjacent targets. Its calculation is robust to actual copy-number variants and should decrease post tangent-normalization.

- **`hj]1[Çlpl`** gives the QC value before tangent-normalization.
- **`hgkl1[Çlpl`** gives the post-tangent-normalization QC value.
- **`\1[Çlpl`** gives the difference between pre and post QC values.
- **`k[Yd]\ë\1[Çlpl`** gives the fraction difference (preQc - postQc)/(preQc).

---

6. Call segmented copy number variants.

This final step makes one of three calls for each segment--neutral (0), deletion (-) or amplification (+). These deleted or amplified segments could represent somatic events.

```
bYnY ÈbYj _YlcýÇbYj #Ydd3]_e]flk ê
       È4. kYf\Zgp×lmegjÇlfÇlkn ê
       È3 kYf\Zgp×lmegjÇk]_ ê
       È/ kYf\Zgp×lmegjÇ[Ydd]\
```

View the results with **`[Yl kYf\Zgp×lmegjÇ[Ydd]\`**.

Sample	Chromosome	Start	End	Num_Probes	Segment_Mean	Segment_Call
tumor	1	1	249238921	17585	0.99647119	0
tumor	2	1	243193925	13159	0.995918783	0
tumor	3	1	21484649	1139	0.999307093	0
tumor	3	21502423	22485593	54	0.798851916	-
tumor	3	22503367	92855531	3740	0.997231251	0
tumor	3	92873305	96090725	174	0.816146473	-
tumor	3	96108499	109371107	705	0.994470169	0
tumor	3	109388881	115519763	326	0.816995479	-
tumor	3	115537537	141901787	1390	0.994263396	0
tumor	3	141919561	198008273	3016	1.146312186	+
tumor	4	1	191131706	6818	0.998268633	0
tumor	5	1	180899061	7849	0.999514915	0
tumor	6	1	171097489	8860	0.995090745	0
tumor	7	1	159125275	8166	0.996402123	0

→ *Besides the last column, how is this result different from that of step 4?*
