Copy Number Analysis in Partek® Genomics Suite™ 6.6

Introduction

At a simple level, copy number analysis can be thought of as a gene dosage question: where are the regions of the genome that have altered abundances (increased or decreased)? Furthermore, what genetic elements are in those regions, and how might a change in the abundance of the genes in those chromosomal regions affect phenotypes?

To answer those questions, the copy number analysis workflow in Partek® Genomics Suite™ (PGS) uses a variety of commercially available SNP-genotyping arrays with closely spaced genomic markers (Affymetrix® and Illumina®) as well as comparative genomic hybridization (CGH) arrays (Agilent®, NimbleGen®, or custom spotted arrays) to detect amplified or deleted regions within and shared across samples. The workflow (Figure 1) begins with either intensity data or copy-number data (depending on the array platform), allows for organization of numerous samples into projects, provides multiple options for data analysis, and enables easy integration with other workflows.

This tutorial will illustrate how to:
- Import the data in Partek® Genomics Suite™
- Perform exploratory data analysis
- Estimate copy number for each marker
- Detect and analyze regions of copy number variation
- Find regions shared across samples
- Create a list of regions that meet certain criteria
- Find genes that overlap the regions of interest
- Learn about other annotations for regions
- Visualize the data at any of the steps

Please be aware of an inherent limitation of the analysis of copy number regions: the inability to detect copy-neutral events, i.e., copy-neutral loss of heterozygosity (LOH) or copy-neutral allelic imbalance. One way to deal with these issues is to supplement the copy number analysis with SNP genotyping data, currently available with Affymetrix® or Illumina® arrays (Figure 1). For more information on integration of copy number with LOH or Allele-Specific Copy Number (AsCN), please consult Appendix C.

Note: The following tutorial was prepared using PGS version 6.6. As PGS is a rapidly evolving software application, future versions of PGS may show different screenshots than what are displayed within this tutorial. To ensure that you are using the most current version of Partek® Genomics Suite™, please use **Help > Check for Updates** from within PGS. There may be slight differences between the screenshots shown in this tutorial and what you observe due to different operating systems and different versions of PGS.
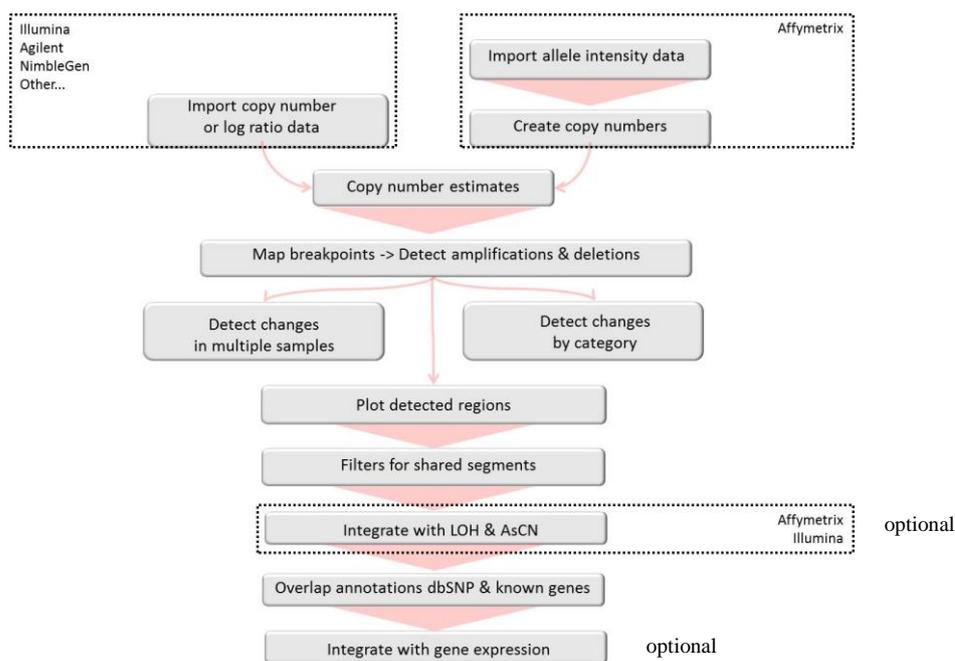
Figure 1: Overview of the copy number analysis in PGS. The analysis of Affymetrix[®] CEL files starts with the import of allele intensities, and the copy number calculation is performed by PGS. For other vendors (including Affymetrix CHP files, Agilent[®], NimbleGen[®], and Illumina[®]), PGS imports copy number or log ratio data as provided by the individual vendor software. Integration with loss of heterozygosity (LOH) or allele specific copy number (AsCN) workflows requires SNP genotype data. Integration with the gene expression workflow requires gene expression data

Importing the data

When starting a new experiment, the first step involves importing samples.
- Open the **Copy Number** workflow within PGS by selecting it from the *Workflows* drop-down list in the upper right corner and select **Import Samples** (Figure 2)
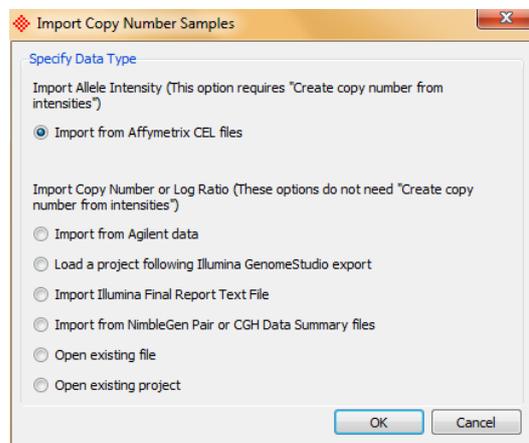
Figure 2: Viewing the Import Copy Number Samples dialog

For Affymetrix® arrays, PGS can import .CEL files containing allele intensities, and copy number estimates will be calculated from those intensities during the subsequent steps. If you are using Agilent®, Illumina®, NimbleGen®, or Affymetrix .CHP files, PGS can import files containing calculated copy numbers or log ratios. To learn more about particular import procedures, please consult the respective vendor-specific tutorials available under *Help > On-line tutorials* or contact our technical support team at support@partek.com.

A common artifact affecting all types of arrays is "GC waves" which have been shown to cause false positive copy number variations. The exact cause of GC waves is not fully understood, but they have been associated with low quality sample DNA and/or DNA degradation. If your Affymetrix® data are affected by GC waves, PGS enables you to implement a GC-wave correction procedure when importing Affymetrix CEL files, following the protocol by Diskin SJ *et al.* (Nucl Acids Res, 2008). For details, please consult Appendix A of this tutorial.

- As this tutorial will not read raw data files but will start from the data already imported into PGS, please close the import dialog (**Cancel)** and proceed with the next section to learn how to calculate copy number estimates for Affymetrix® .CEL data.

The analysis of copy number or log ratios provided by Agilent®, Illumina®, NimbleGen®, or Affymetrix® .CHP platforms is described in the section *Detection of Regions with Copy Number Variations*.

Tutorial data set

This example data set consists of 20 paired samples from an ovarian cancer study in which a fresh-frozen tumor sample and peripheral blood were obtained from each of 10 female patients (Ramakrishna et al., PLoS One, 2010, GSE19539). All 20 samples were analyzed using the Affymetrix® Genome Wide Human SNP Array 6.0. For a discussion on formalin-fixed paraffin-embedded (FFPE) tissue samples, please refer to Appendix B of this tutorial.
- Select **Help > On-line Tutorials** from the PGS main menu and download the data for the *Overlapping Copy Number with LOH* project (in .zip format)
- Unzip the files into a directory of your choosing
- Select **File > Open,** browse to the folder containing unzipped the tutorial data, select the file **IC_Intensities_SNP6.fmt,** and select **Open**. The datasheet will now open in the PGS main window (Figure 3)

The spreadsheet was generated from the import of SNP6 CEL files and shows all 20 samples on rows. Columns 1-6 contain sample information describing the experiment such as file names, Subject ID, Gender, etc. The rest of the columns are individual markers from the microarray with the $\log_2$ normalized intensities associated with each marker (marker labels are the column headers). Importing the IC_Intensities_SNP6.fmt is the equivalent of importing the 20 samples files and adding sample attributes.

Figure 3: Viewing the spreadsheet with the intensity data for the copy number tutorial. Each row represents one sample

Setting the SampleID

If you plan to do any kind of genomic integration with other studies (e.g., gene expression), then the results of both studies must be tagged with the same sample identifiers. The sample identifier must be specified before analysis results are created for copy number experiments.

- Under the *Import* section of the *Copy Number* workflow, select **Choose sample ID column**. Select **4. SubjectID** for *Sample ID Column*

Exploratory analysis

As an introduction to the copy number analysis, you might want to explore the intensity data for insight into possible groupings within the dataset or to detect any outliers or sample swaps. One method is the principal component analysis (PCA) scatterplot.

- In the **QA/QC** section of the workflow, select **Principal component analysis (PCA).** The rotated PCA plot is shown in Figure 4

Each dot on the plot corresponds to a single sample (i.e., one row of the spreadsheet) and can be thought of as a 'summary' of all normalized marker intensities in the sample. PGS automatically uses the first categorical column, in this example column *3 (Tumor)*, to color the dots. Depressing the mouse wheel while moving the mouse or using the **Rotate mode** will rotate the plot to find the optimal angle of view. As seen in the figure, the peripheral blood samples (i.e., "normals") cluster together whereas the cancer tissue samples are more dispersed and show considerable variability with respect to their total allele intensity profiles. That finding corresponds very well to the underlying biology and the genomic variability of cancer cells. To learn more on how to customize the PCA plot, please consult Partek® On-line Documentation (*Help > User's Manual*).

Note: the PCA plot can be invoked on allele intensities as shown here, but also on the copy number spreadsheet (generated in the next step) as well. Therefore, PCA is also available to users analyzing data from other vendors.
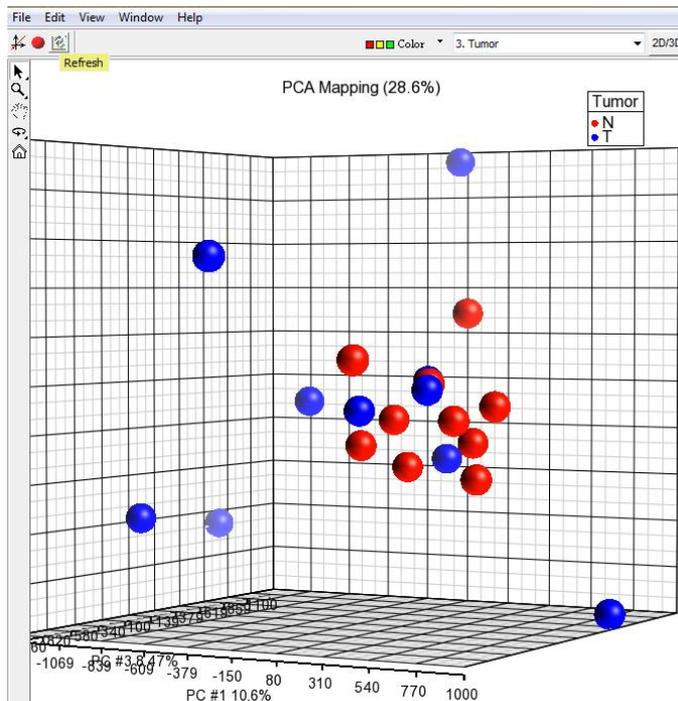
Figure 4: Principal component analysis scatterplot showing the total allele intensities of normal (red dots) and cancer (blue dots) samples. Each dot represents a single sample. Note the tumor samples (blue) are more dispersed than the normal samples (red)

Estimating copy number from marker intensities

The first step in the analysis of Affymetrix® intensity data is to estimate the number of copies of each marker (allele). The *Create copy number (from allele intensities only)* function can be accessed through the *Import* section of the *Copy Number* workflow and will estimate the copy number of each marker by comparing it to a reference. Depending on the design of the experiment, the user first has to choose between paired and unpaired samples (Figure 5). A paired design assumes that each sample has its own reference sample; PGS will generate a copy number spreadsheet. This is the typical and recommended situation for tumor/normal sample pairs for finding somatic cancer mutations. Unpaired samples, on the other hand, use a common reference (single or group of samples), and PGS will create an allele ratio spreadsheet (intended primarily for visualization purposes) in addition to the copy number spreadsheet. Unpaired analysis is typically used for studying inherited effects.
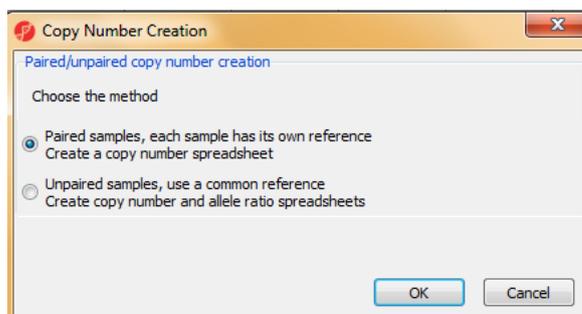
Figure 5: Viewing the Copy Number Creation dialog

In the current example, to estimate the copy number from allele intensities, PGS will compare a cancer sample ("tumor") to the peripheral blood sample ("normal") from the same individual, using the *SubjectID* column to identify the sample pairs. In other words, the allele intensities of each normal sample will serve as the baseline for the calculation (the algorithm assumes that normal samples have 2 copies of DNA at each locus). If the intensity of a probe is 2 times brighter than the baseline, it has twice as much DNA at the location on the genome which the probe targets. Conversely, half the normal intensity will point to a single copy of DNA at given locus.

- In the *Import* section of the *Copy Number* workflow, select **Create copy number (from allele intensities only)**
- To proceed with the analysis, select **Paired samples** in the *Copy Number Creation* dialog (Figure 5) and select **OK**
- The *Create Copy Number from Pairs* dialog will appear (Figure 6). The baseline samples are defined by *Column* **3. Tumor** and the *Baseline category* should be **N** (normal)
- *Match the sample pairs* uses Column **4. SubjectID** as the parameter that pairs the samples from each patient. Set these options accordingly and select **OK**
- The *IC_Intensities_SNP6_pairedcopynumber* spreadsheet is created (Figure 7)
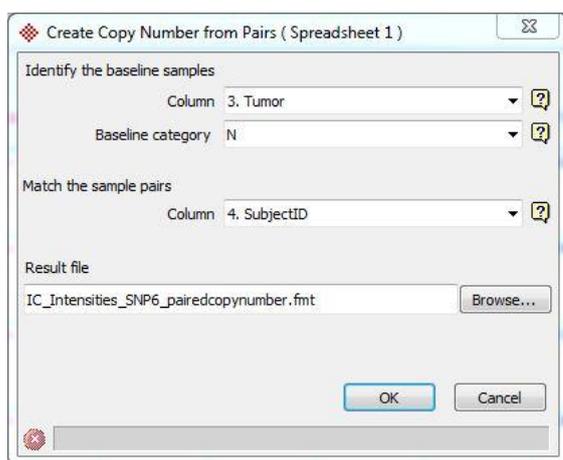
Figure 6: Viewing the Create Copy Number from Pairs dialog

Alternatively, if paired samples are not available, the *Unpaired samples* option should be selected in the *Copy Number Creation* dialog shown in Figure 5.

Both paired and unpaired copy number procedures produce a copy number spreadsheet (in this tutorial, *IC_Intensities_SNP6_pairedcopynumber)* (Figure 7). Each row represents one of the tumor samples. Columns 7+ include copy numbers of each marker (columns 1-6 are identical to those in the *IC_Intensities_SNP6* spreadsheet).

Current Selection | IC_151T_FF.CEL.pimg

	1. Original	2. File	3. Tumor	4. SubjectID	5. Gender	6. Scan Date	7. AFFX-5Q-123	8. AFFX-5Q-456	9. AFFX-5Q-789	10. AFFX-5C
1.	IC_151T_FF.CE L.pimg	IC_151T_FF.CE	T	151	Female	07/27/07	2.42484	1.9396	2.86727	2.18389
2.	IC_201T_FF.CE	IC_201T_FF.CE	T	201	Female	07/27/07	1.2475	1.72845	1.25524	1.45485
3.	IC_22T_FF.CEL	IC_22T_FF.CEL	T	22	Female	07/27/07	1.38206	1.91484	1.32699	1.55652
4.	IC_258T_FF.CE	IC_258T_FF.CE	T	258	Female	07/27/07	2.45219	2.26834	2.17861	2.08497
5.	IC_315T_FF.CE	IC_315T_FF.CE	T	315	Female	07/27/07	1.85366	1.84564	2.01329	1.88579
6.	IC_399T_FF.CE	IC_399T_FF.CE	T	399	Female	07/27/07	2.41627	1.57144	2.42373	1.98325
7.	IC_504T.CEL.pi	IC_504T.CEL	T	504	Female	07/27/07	2.77112	5.42197	2.04939	2.1622
8.	IC_580T.CEL.pi	IC_580T.CEL	T	580	Female	07/28/07	1.96063	1.80866	2.01355	1.94614
9.	IC_594T_FF.CE	IC_594T_FF.CE	T	594	Female	07/27/07	1.95618	2.32267	2.17715	1.97389
10.	IC_95T_FF.CEL	IC_95T_FF.CEL	T	95	Female	07/27/07	1.13677	1.48696	1.0251	1.40598

Figure 7: Viewing the paired copy number spreadsheet. Each row represents one cancer sample; the marker columns show the copy number (ratio of tumor to normal) for each pair of samples

Creating a reference baseline of normal samples

All copy number experiments give relative copy numbers (sample compared to a reference). Basically, PGS needs to be given a normal reference or baseline for the copy number calculation. It could be an existing reference file provided by Partek® (*Option 1* in Figure 8) (see the Affymetrix® and Illumina® baseline files available at *Help > On-line Tutorials > Baseline Files)* or a reference file previously created from a set of normal samples (*Option 2*) or a subset of samples within the current project (*Option 3*). There are two different ways you may create a reference baseline (.cnmodel) file: create a reference from a collection of normal samples not included in this experiment or use samples from the current experiment as the reference baseline. The resulting dialog is shown in Figure 8.

To generate a pooled reference file from a collection of normal samples, first import only the normal samples and then use **Tools > Create Copy Number Baseline** with the intensity spreadsheet selected. Once you have created a .cnmodel file this way, use the top radio button (*Option 2* in Figure 8).

The other way to generate a baseline for copy number calculations is to use some or all of the samples in the existing experiment (*Option 3* in Figure 8). Using this option, copy number estimates will also be generated for each normal sample (compared to the pool of normal samples).

For more information about using unpaired samples in copy number calculations, please consult Understanding Copy Number Creation (*Help > On-line Tutorials > White Papers*).
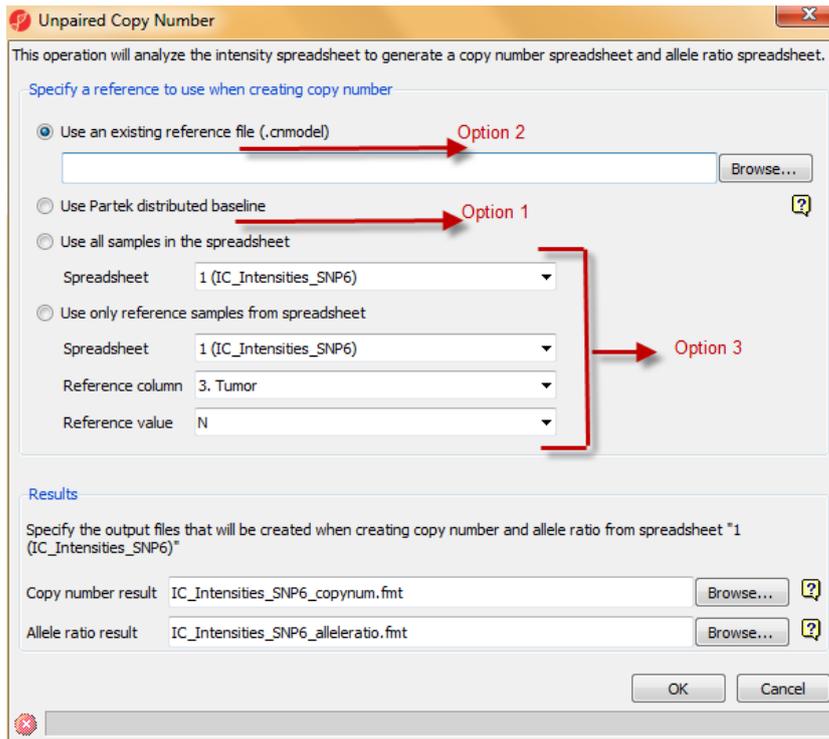
Figure 8: Viewing the Unpaired Copy Number dialog. Three kinds of baseline references are possible: using a baseline file distributed by Partek® (Option 1), using a previously created (cnmodel) reference file (Option 2), or using some or all of the samples in the current experiment as the reference (Option 3)

Visualization of copy number of allele intensity data

The result of copy number estimation per marker can be visualized by invoking the chromosome view from the *Copy Number* workflow. At this stage of the workflow, you can visualize only the raw copy number estimates of each marker (Figure 9).

- Select **Plot chromosome view** from either **QA/QC** or **Visualization** (Figure 9). If prompted to *Select an annotation source,* select **Ensembl Transcripts**
- The *Track Wizard* may prompt for which spreadsheet should be used for visualization*:* either the intensity or copy number spreadsheet may be chosen. At this step, select the **Copy Number 2** spreadsheet and **Create**

Each of the ten cancer samples is represented by a single row in the *Heatmap* track (below *Ensembl Transcripts*), with the color pattern ranging from blue (no copies) to red (four or more copies of a marker). The name of the selected sample can be found to the left of the selected row in the *Heatmap* track. The *Profile* track is located below the heat map track and shows the copy number value (y-axis) of each marker (grey dots) of the sample highlighted in the heat map track. In the other words, grey dots are a visualization of copy numbers in the cells of the copy number spreadsheet. While the grey dots represent the raw copy numbers, the bold (heavy) dots represent the smoothed copy number (average of 30 adjacent markers by default). More about the chromosome view and different ways to customize it can be found in the *Chromosome Viewer* User Guide (*Help > On-line Tutorials > User Guides*).

- To examine a different sample, click on another row of the Heatmap track. Toggle through different samples and notice the different pattern in copy numbers across samples

- To zoom in, use the magnifying glass icon [🔍] at the top of the window
- To view another chromosome, use the down arrow next to the chromosome location box at the top of the window
- When you have finished visualizing the raw copy numbers, select the red X at the top right of the window
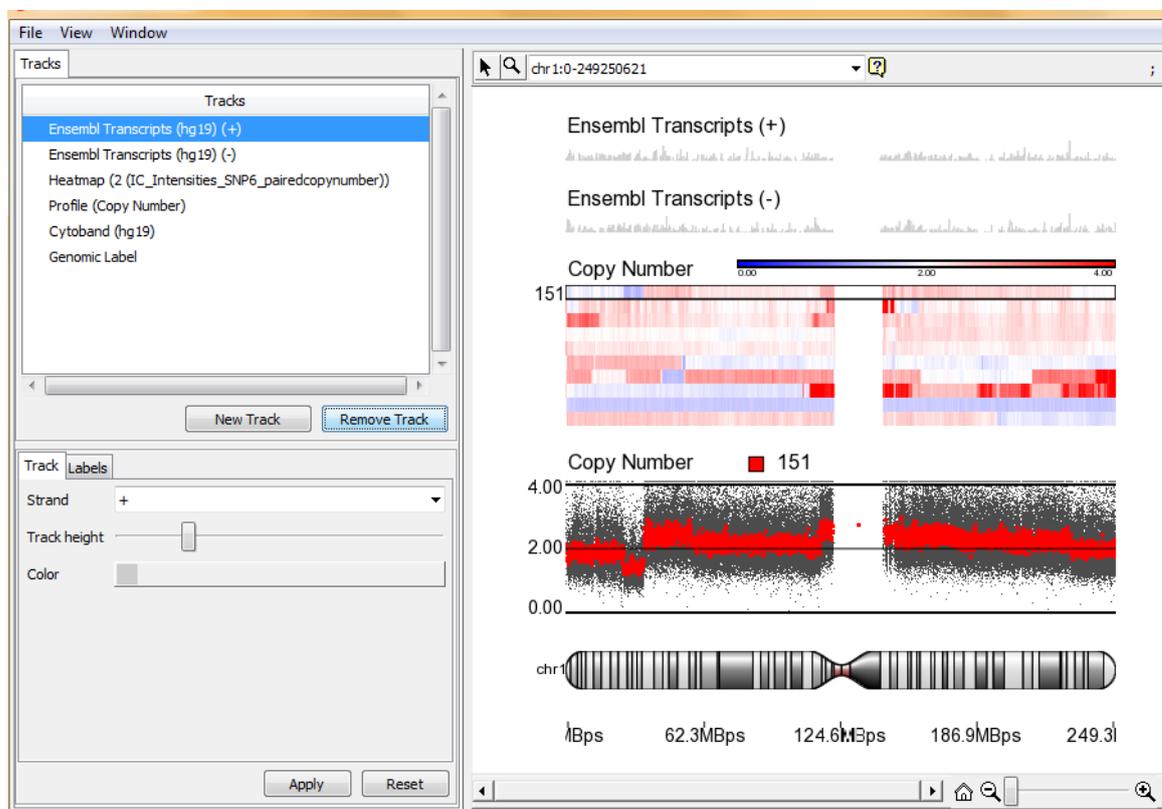
Figure 9: Chromosome view of copy number spreadsheet. The default tracks from the top are Genomic features from a selected annotation source (Ensembl Transcripts shown here), Heatmap track, Profile track, and Cytoband track. Chromosome 1 is shown by default. The Genomic label (genomic coordinates) is shown below the Cytoband track

Detection of regions with copy number variation

Starting with copy number estimates for each marker (either taken directly from the vendor's input file or calculated previously), the goal is to derive a list of regions where adjacent markers share the same copy number. PGS offers two algorithms for region detection: *Genomic segmentation* and *Hidden Markov Model* (HMM). Basically both algorithms examine trends across multiple adjacent markers. The genomic segmentation algorithm identifies breakpoints in the data, i.e., changes in copy number between two neighboring regions. The HMM algorithm looks for discrete changes of whole number copy number states (e.g., 0, 1, 2 … with no limit on the upper limit) and will find regions with those numbers of copies. Therefore, the HMM model performs better in cases of homogenous samples when the copy numbers can be anticipated (such as clinical syndromes with underlying chromosome or germ-line gene aberrations). Genomic segmentation is preferred for heterogeneous samples with unpredictable copy numbers (such as cancer because tissue biopsies often contain "contaminating" healthy

tissue, and cancer cells are quite heterogeneous with respect to multiple chromosome aberrations).

The number of copies of each marker created in the previous step will now be used to detect the genomic regions with copy number variation, i.e., to identify amplifications and deletions across the genome.

- Open the **Copy Number Analysis** section of the *Copy Number* workflow
- Select **Detect amplifications and deletions**. The dialog in Figure 10 will appear
- Select **Genomic Segmentation** and then select **OK**
- Specify the *Minimum genomic markers* as **50** and leave the remaining options at their default settings shown in Figure 11. To start the segmentation, select **OK**
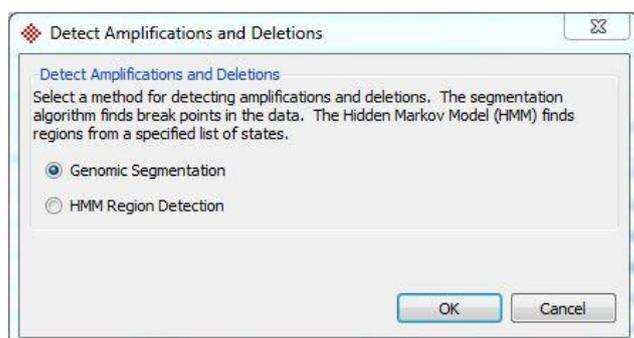
Figure 10: Viewing the Detect Amplifications and Deletions dialog

Genome segmentation itself is divided into two steps. In the first step, each region is compared to an adjacent region in order to tell whether both have the same average copy number and if a breakpoint can be inserted by using a two-sided t-test (the t-test actually compares the average intensities of regions and if the corresponding cut-off p-value is below the *P-value threshold*). The genomic size of a region is defined by the number of genomic markers included in the region (*Minimum genomic markers*), while the magnitude of significant difference between two regions is controlled by *Signal to noise* (simplified, it could be thought of as the difference in copy numbers between the regions). If the t-test is significant, it can be concluded that the region differs significantly from its nearest neighbors with respect to copy number. However, a second step is needed to identify the exact nature of the difference, i.e., whether the difference is an amplification or deletion. In this stage, two one-sided t-tests are used to compare the mean copy number in the region with the expected (normal) diploid copy number. For a detailed explanation of the genomic segmentation procedure, please consult the *Understanding Genomic Segmentation* white paper (*Help > On-line Tutorials > White Papers*). A reader interested in in-depth optimization of the segmentation procedure is encouraged to refer to the tutorial *Optimizing Copy Number Segmentation in Partek* (*Help > On-line Tutorials > Copy Number*).
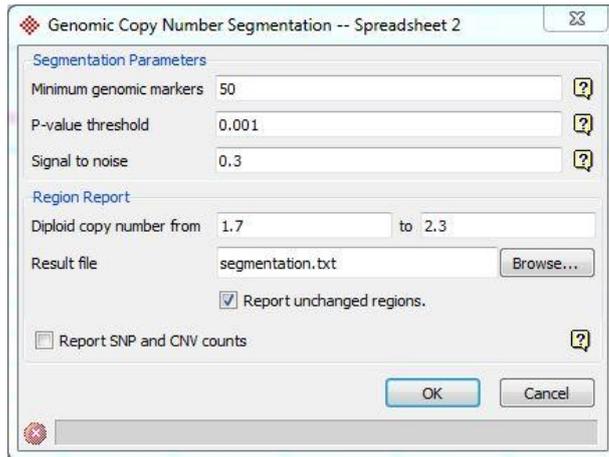
Figure 11: Setting the key parameters of the genomic segmentation procedure

The resulting spreadsheet (*segmentation*) shows one row per genomic region per sample (Figure 12). The columns provide the following information:

1 – 4:	Genomic location of the region
5:	Sample ID
6:	Description of the copy number change (amplification, deletion, etc.)
7:	The length of the region (in base pairs)
8:	The number of markers in the region
9:	Marker density in the region (region length divided by the number of markers)
10:	Geometric mean of the copy number of all the markers in the region
11:	Minimum p-value of the one-sided t-tests of the difference of the copy number in column 10 vs. the diploid range

Current Selection chr1

	1. chromosome	2. start	3. end	4. cytoband	5. Sample ID	6. Copy Number	7. length(bps)	8. # markers	9. marker density	10. mean
1.	chr1	61735	761550	1p36.33	151	Unchanged	699815	50	13996.3	2.25919
2.	chr1	761550	12192261	1p36.33 -	151	Unchanged	11430711	6216	1838.92	1.8545
3.	chr1	12192261	12315320	1p36.22	151	Deletion	123059	84	1464.99	1.50716
4.	chr1	12315320	25588234	1p36.22 -	151	Unchanged	13272914	8062	1646.35	1.87173
5.	chr1	25588234	25699232	1p36.11	151	Deletion	110998	50	2219.96	1.27463
6.	chr1	25699232	26328074	1p36.11	151	Unchanged	628842	288	2183.48	1.81903
7.	chr1	26328074	33393918	1p36.11 -	151	Deletion	7065844	3535	1998.82	1.41851
8.	chr1	33393918	33606055	1p35.1	151	Deletion	212137	107	1982.59	1.1932
9.	chr1	33606055	35546698	1p35.1 - 1p34.3	151	Deletion	1940643	1339	1449.32	1.49033
10.	chr1	35546698	57198030	1p34.3 - 1p32.2	151	Amplification	21651332	12623	1715.23	2.39176
11.	chr1	57198030	72760190	1p32.2 - 1p31.1	151	Unchanged	15562160	11233	1385.4	2.14522
12.	chr1	72760190	72818834	1p31.1	151	Amplification	58644	57	1028.84	2.6741
13.	chr1	72818834	115322035	1p31.1 - 1p13.2	151	Unchanged	42503201	27612	1539.3	2.10198
14.	chr1	115322035	145107483	1p13.2 - 1q21.1	151	Amplification	29785448	3883	7670.73	2.55175
15.	chr1	145107483	229146879	1q21.1 -	151	Unchanged	84039396	54664	1537.38	2.27473
16.	chr1	229146879	249224389	1q42.13 - 1q44	151	Unchanged	20077510	14732	1362.85	1.94588

Figure 12: Viewing the segmentation.txt spreadsheet. Each row of the spreadsheet represents one genomic region per sample

You may use the *Tools > Merge Adjacent Regions* to combine similar regions (like the first two unchanged regions shown above).

Visualizing regions of interest

The regions identified by the segmentation algorithms may be visualized in the chromosome viewer either at a single region or entire chromosome level.

- To visualize a region of interest, you may right-click on a row header and then choose **Browse to Location**
- Alternatively, to get an overview of the results, select **Visualization > Plot chromosome view,** and the *Track Wizard* will appear
- Please select the **Genomic Segmentation 2/segmentation (segmentation.txt)** when asked to *Choose the spreadsheets that you would like to add to the plot.* This adds the *Regions* track, which is the primary visualization of the segmentation results, to the plot
- For a more comprehensive visualization, also add the *Copy Number 2(IC_Intensities_SNP6_pairedcopynumber)* track in the *Track Wizard* and **Create** (Figure 13)

The *Regions track* (entitled "Genomic segmentation") depicts the segmentation results: each line represents a single sample (sample names can be found on the left side of the selected line), while the amplified, deleted, and unchanged regions are shown as red, blue, and white, respectively. The *Heatmap* and *Profile track* have already been discussed previously. However, please note that the *Profile track* (depicting the sample selected in either *Heatmap* or *Regions* track) now also includes the average aberrations for each region.
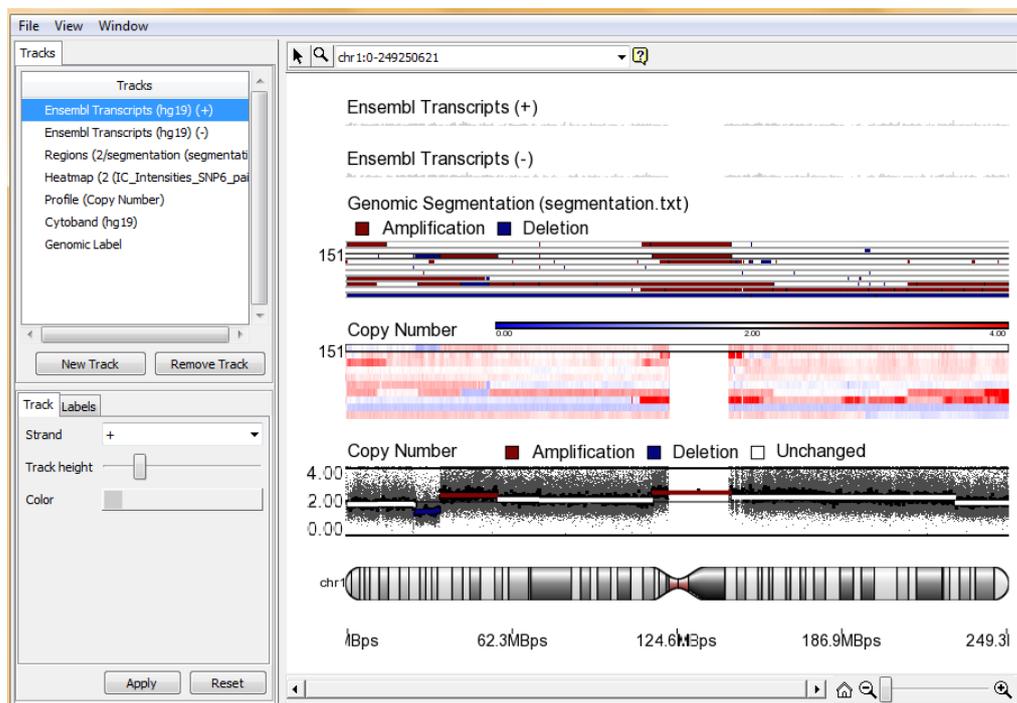
Figure 13: Chromosome view of segmentation spreadsheet. Tracks may be reordered by dragging track names in the Tracks *navigator on the left*

Analysis of shared regions with copy number variation

Once the regions with amplifications and deletions have been detected, the next step is to compare the regions across the samples and detect those which are shared by multiple samples (Figure 14). For instance, cancer samples, such as the ones used in this tutorial,

are characterized by genomic instability and multiple mutations. Therefore, it is most interesting to pick up only those copy number aberrations that appear in multiple or even all of the samples as those might be good candidates for diagnostic or prognostic markers or might involve key genes responsible for cancer pathogenesis. The *Analyze detected segments* function will identify genomic segments across all samples. Please note that since aberrations may be broken into smaller regions, the common regions may contain fewer than the minimum number of markers specified in the segmentation parameters.
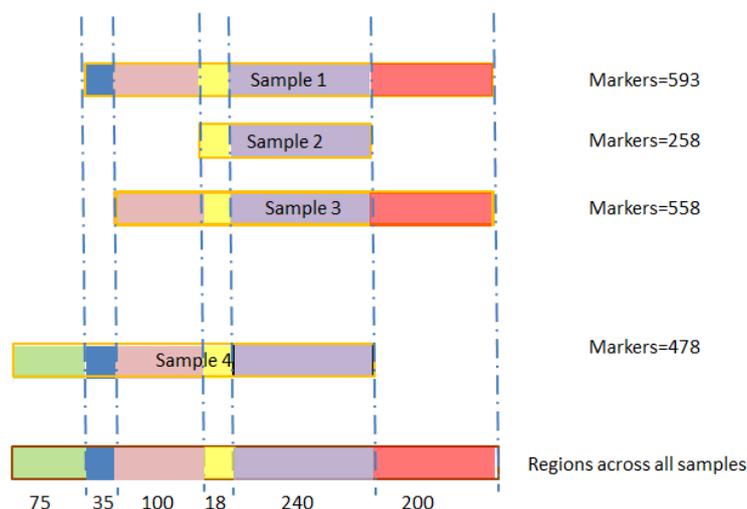
Figure 14: The Analyze Detected Segments *function detects and reports the regions across all samples. The original detected region for each sample is outlined in gold with the number of markers contained in the region shown to the right. Regions overlapping multiple segments are drawn in color with the number of markers shown below each region*

- Select **Analyze detected segments** function from the *Copy Number Analysis* section of the *Copy Number* workflow
- As shown in Figure 15, the *Analyze Segments* dialog can test for associations between copy number variations and sample categories by using the χ^2 test. For example, is there a difference in aberrations between cancer grades? Since in this paired analysis, all pairs (ovarian cancer versus their normal) share the same phenotype, associations are not tested with this paired tutorial data. Hence, leave all the boxes unchecked and select **OK** to proceed
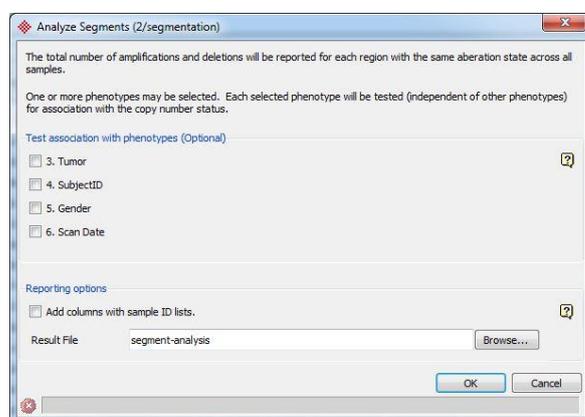
Figure 15: Viewing the Analyze Segments dialog

The *summary (segment-analysis)* spreadsheet (Figure 16) shows one region per row. The columns provide the following information:

- 1 – 4: Genomic locations of the region
- 5: Total number of samples
- 6 – 7: Number of samples with amplifications and the average amplified copy number, respectively
- 8 – 9: Number of samples with deletions and the average deleted copy number, respectively
- 10: Total number of samples with copy number aberrations
- 11 – 12: Number of samples with no change in copy number and the average copy number in those samples, respectively
- 13: Number of markers in the region
- 14: Length of region (in base pairs)
- 15+: Two columns per sample: the average copy number in each sample as well as the copy number change status of the same sample (e.g., amplified, deleted, unchanged, depending on the copy number and the threshold for unchanged defined in the segmentation dialog)

A "?" indicates that a region with particular characteristic does not exist or cannot be computed (e.g., if a region is not amplified in none of the samples, the average amplified copy number will be shown as "?"). This list may be filtered to contain only the regions that meet user-specified criteria (next section).

Current Selection 4

	1. Chromosome	2. Start	3. Stop	4. cytoband	5. Total Number of Samples	6. Total Amplifications	7. Amplification Average Copy Number	8. Total Deletions	9. Deletion Average Copy Number	10. Total Aberrations
1.	4	47231926	48070430	4p12	10	3	2.4993	7	1.55235	10
2.	4	48070430	48414701	4p12 - 4p11	10	2	2.45194	8	1.49239	10
3.	4	48414701	48814423	4p11	10	3	2.42121	7	1.55235	10
4.	4	49658417	52689480	4p11 - 4q11	10	3	2.38827	7	1.38959	10
5.	4	52689480	52710591	4q11 - 4q12	10	2	2.40201	8	1.42554	10
6.	4	52874472	52887805	4q12	10	3	2.41097	7	1.43174	10
7.	4	52887805	53426984	4q12	10	4	2.38421	6	1.46571	10
8.	4	53426984	55077283	4q12	10	4	2.38421	6	1.5426	10
9.	4	55141568	55262530	4q12	10	4	2.3787	6	1.5426	10
10.	4	103740604	104289463	4q24	10	0	?	10	1.45503	10
11.	4	182912966	183060840	4q34.3	10	1	2.4397	9	1.41905	10
12.	6	107088052	107306530	6q21	10	7	4.32094	3	1.58767	10
13.	8	39233131	39241139	8p11.22	10	6	2.83275	4	1.09004	10
14.	8	39386079	39387632	8p11.22	10	6	2.83678	4	1.09004	10
15.	8	122965748	124297671	8q24.13	10	7	2.90256	3	1.45791	10
16.	8	124297671	127416908	8q24.13 -	10	7	3.00294	3	1.45791	10
17.	8	127416908	127630013	8q24.21	10	7	2.88703	3	1.45791	10

Figure 16: Viewing the Summary (segment-analysis) spreadsheet. Each row of the spreadsheet represents one genomic region shared across multiple samples

Visualization of shared regions

To visualize the regions shared across the samples, PGS offers two plots: histogram and copy number classification (Figure 17). Both are intended to give an overview of the common aberrations in the group of samples over the entire genome and to decide on the parameters for the next step, *List Creation*.

- To obtain the histogram plot, select **Plot Histogram** from the *Plot Detected Regions* dialog. Ensure the proper spreadsheet is selected and select **OK**
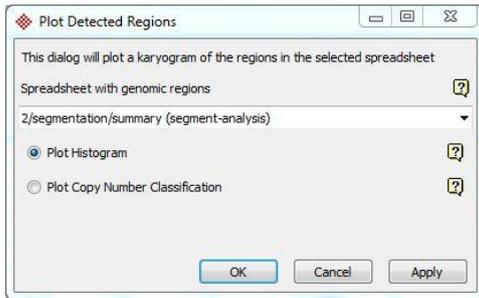
Figure 17: Viewing the Plot Detected Regions dialog

The *Karyogram View* (Figure 18) shows an overview of the shared regions across the genome with amplified regions coded in red and deleted regions coded in blue. The histogram heights reflect the number of samples that share that kind of aberration at a particular location. For example, the long arms of chromosomes 3 and 7 seem to be amplified in the majority of tutorial samples, and most samples share the deletion in the long arm of chromosome 4.

- Use the mouse-over function to get the information on cytobands as well as on the exact number of shared regions at each position, and the number of samples sharing that type of aberration
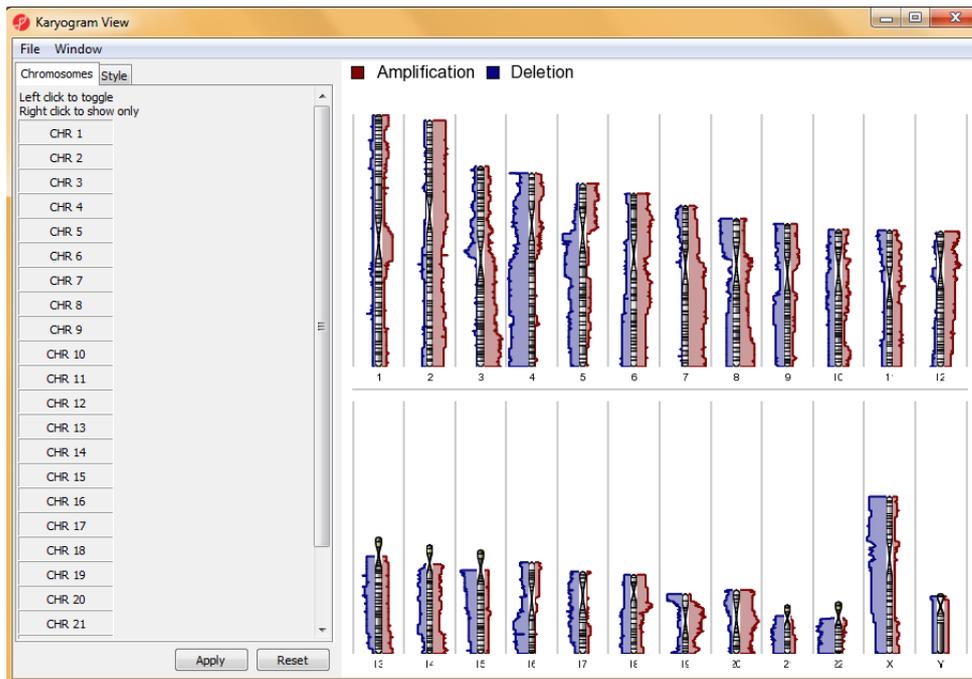
Figure 18: Histogram plot providing an overview of amplified (red) and deleted (blue) regions shared across the samples on either the entire genome (as shown) or on selected chromosomes. Both the menu and the display may be used to control which chromosomes are displayed; left click in the menu to toggle on/off; right click in menu or graph to show only that chromosome

For a sample-centric visualization, invoke the copy number classification plot.

- Close the histogram plot by selecting the red X in the upper right corner of the plot
- Select **Plot detected regions** again. Choose the **Plot copy number classification** option and select **OK**

Copy Number Analysis in Partek® Genomics Suite™ 6.6   15

The *copy number classification* view provides an overview of all the samples and the copy number regions on each chromosome (Figure 19). Each sample is drawn as a separate column next to the chromosome. Amplified regions are depicted in red, deleted regions in blue, and the regions with no copy number change are depicted in white. Sample names are given across the top. As shown in the insert in Figure 19, displaying fewer chromosomes will show the detail for each sample more clearly.
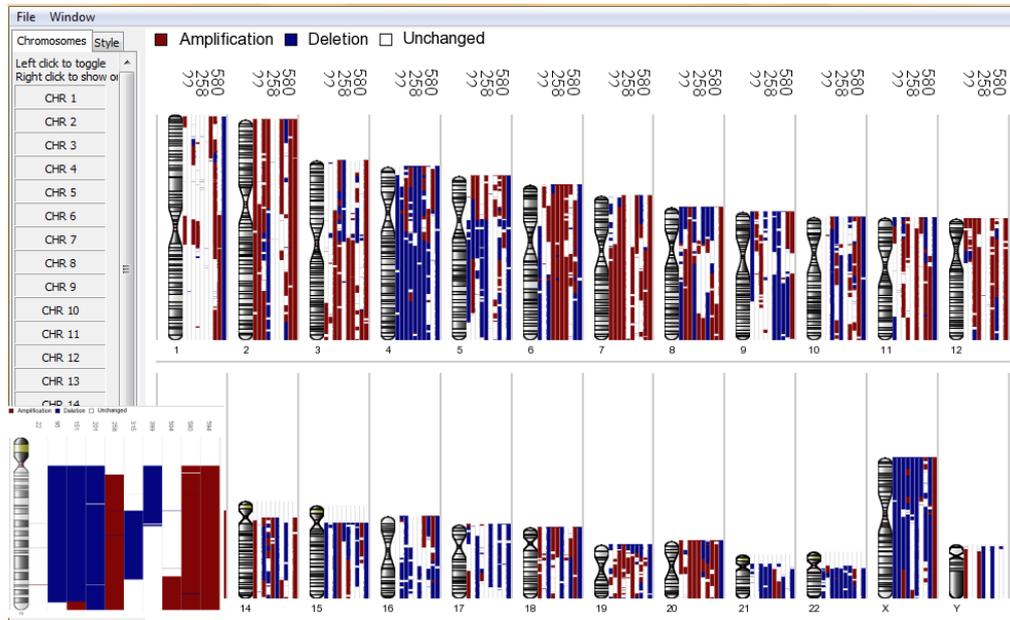
Figure 19: Copy number classification plot providing an overview of amplified (red), deleted (blue), and the regions with no copy number change (white) in each sample. The small inset on the bottom left shows the increased detail if just one chromosome is visualized

Creation of a list of regions

Creating lists of shared regions can be described as the core step of the copy number analysis in a project with multiple samples/patients. For instance, one can select all the deleted regions on chromosome 4, find all the regions spanning more than 50000 bp, or pick-up all the regions containing a certain number of markers. In this exercise, all the samples have same underlying phenotype, and two lists will be created: deleted and amplified regions across the genome shared by 8 or more samples. Those lists can then later be used for integration with loss of heterozygosity and allele-specific copy number tutorials (refer to Appendix C for more information).

- To invoke the *List Creator,* please go to **Create region list** in the *Copy Number Analysis* section of the *Copy Number* workflow
- Start defining a new criterion by selecting **Specify New Criteria**. To filter out all the amplified regions across the genome shared by at least eight samples, make the following changes in the *Configure Criteria* dialog (Figure 20)
- Set the *Name* as **Amplified**
- Set the *Spreadsheet* to **2/segmentation/summary (segment-analysis)**
- Set the *Column* drop list to: **6. Total amplifications**
- Set the box *Include values* **greater than or equal to 8**. Uncheck the box *Include values less than or equal to.* Select **OK**
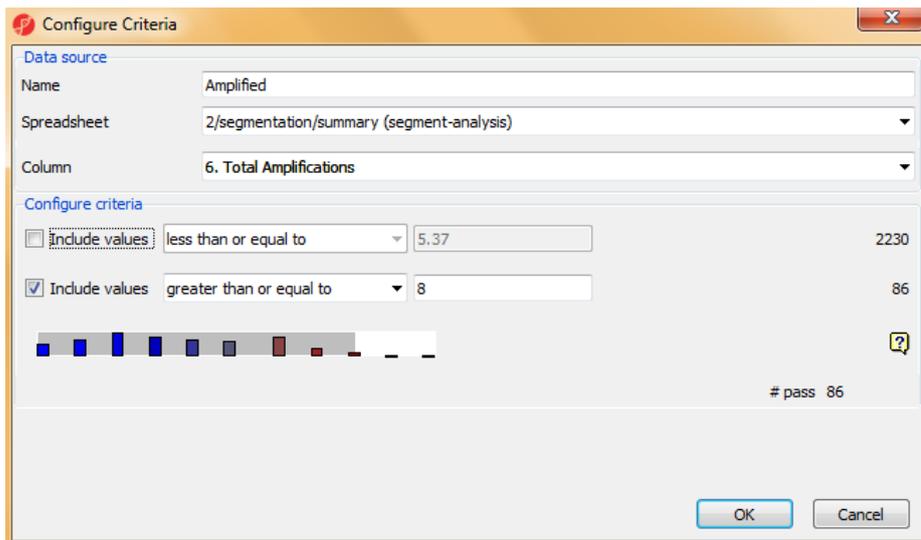
Figure 20: Viewing the Configure Criteria dialog

- Select **Save** to save the list of 86 amplified regions (confirm the name of the list in the following window), **OK**, and **Close** to exit the *List Creator* dialog

However, please note that although the list does contain the regions amplified in eight or more samples, some samples may also contain deletions in the same regions (Column 8 in Figure 21). For the downstream analyses, we may want to filter out those regions, i.e., to have the final list of regions that are only amplified in 8 or more samples. There are various options to perform the filtering in PGS; for this tutorial, we will use the interactive filter.

Figure 21: Viewing the initial list of regions amplified in 8 or more samples

- To invoke the interactive filter, please click on the interactive filer icon in the main window (📊) with the *Amplified* spreadsheet selected
- Set the *Column* drop down list to **8. Total Deletions**
- Type **0** in the *Max* box and press **enter** (Figure 22)

Figure 22: Configuring the interactive filter tool

A filtered list of 60 regions will appear. Please note that the deletions are no longer present in the list. The yellow and black bar on the right-hand side of the spreadsheet (not shown) indicates that the spreadsheet has been filtered; the height of the bar depicts the proportion of the filtered entries with respect to the number of entries in the original spreadsheet.

- To save the filtered list, right-click on the title of the *Amplified* spreadsheet in the list of spreadsheets (left pane of the main window) and select **Clone**…
- Set the *Name of resulting copy* to **Amplifiedonly**, and set the drop-list of *Create as a child of spreadsheet* to **2/segmentation (segmentation.txt)** (Figure 23). Select **OK**
- Note that the cloned spreadsheet has (*ptmp*#)* after its name in the spreadsheet navigator. This indicates that the spreadsheet has changed but has not been saved
- Save the new spreadsheet, *amplifiedonly*, by selecting the Save icon in the main window (💾). For *File name:* type in **Amplifiedonly** and **Save.** Notice that both the *ptmp* and * no longer appear
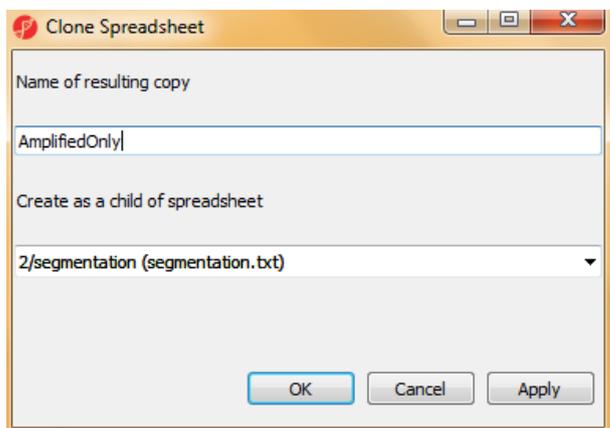
Figure 23: Configuring the Clone Spreadsheet dialog

- Examine the *Amplifiedonly* spreadsheet and notice that it contains 60 rows. Select the *Amplified* spreadsheet and notice that it also contains 60 rows when it contained 86 rows when it was created. The reduced number of rows is due to filter that was just applied. The yellow and black bar next to the scroll bar on the right of the spreadsheet indicates that a filter has been applied.
- Right-click on the yellow/black scroll bar and **Clear Filter.** The number of rows should return to 86

For this exercise, please perform the same sequence of steps to create a list of deleted regions shared across eight or more samples and remove the amplifications (make sure that you have the *summary (segment-analysis)* spreadsheet selected when you invoke the *List Creator*). You should have a list (e.g. *deletedonly*) of 94 regions.

Now you will create a single list containing all of the amplifications-only regions that occurred in 8 or more samples and all of the deletions-only regions that occurred in 8 or more samples.

- To merge both lists (amplified and deleted regions), please select the *amplifiedonly* spreadsheet in the spreadsheet pane on the left

- Use **File > Merge Spreadsheets…** and select the **Append Rows** tab
- Select the source and destination spreadsheets as shown in Figure 24
- Select **OK**. This will append the rows of the *deletedonly* spreadsheet into the *amplifiedonly* spreadsheet. The *amplifiedonly* spreadsheet now contains 154 rows (60 amplifications and 94 deletions)
- To save the new, joined list as a separate spreadsheet, right-click on the *amplifiedonly* spreadsheet in the left pane and select **Clone…** For *Name of resulting copy,* select an appropriate file name such as **amplified_or_deleted**. For *Create as a child of spreadsheet,* select **2/segmentation (segmentation.txt).** Select **OK**
- The *Merge Spreadsheets* command appended the deleted regions onto the amplified list. Now the *amplifiedonly* list contains both the regions amplified in 8 or more samples and regions deleted in 8 or more samples. To return the *amplifiedonly* spreadsheet to its original state where it only contains the regions amplified in 8 or more samples, right-click on the *amplifiedonly* spreadsheet in the spreadsheet view on the left and select **Revert to Last Saved State**. The *amplifiedonly* spreadsheet will now contain only the 60 regions

The joined list (*amplified_or_deleted)* will be the starting point for the following steps of the tutorial and may be used for LOH and AsCN workflows for integration with copy number data (please see the respective tutorials, as well as appendix C of this tutorial).
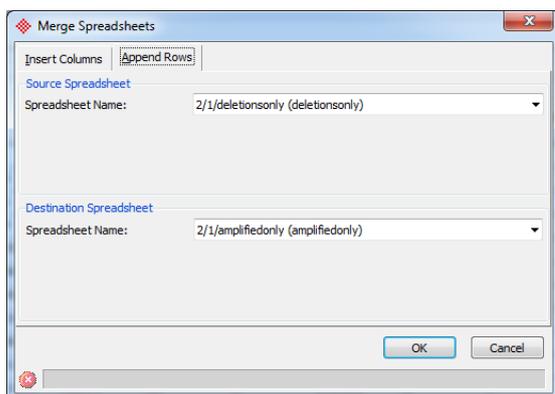
Figure 24: Configuring the Merge Spreadsheets dialog

Find overlapping genes

While the list of regions containing aberrations is interesting, the underlying biology is best investigated by identifying the genes contained in the aberrations. One way to annotate interesting regions in PGS is to generate a list of genes which overlap the regions. The *Find overlapping genes* function allows two options (Figure 25); the main difference being the focus of the output. For a region-centered view, which may be more appropriate for cytogenetic studies, you may add a new column with the genes contained in the regions. The gene-centered view, suitable for genomics integration, is available by creating a new spreadsheet with genes that overlap with the regions.
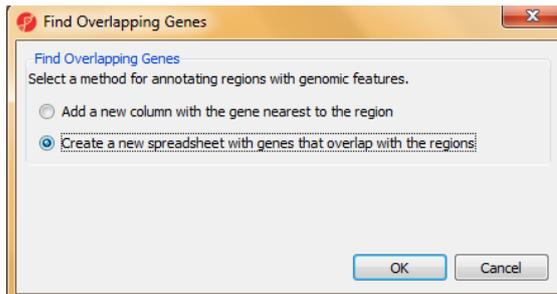
Figure 25: Options in Find Overlapping Genes dialog

- Make sure the *amplified_or_deleted* spreadsheet is selected in the spreadsheet list
- Select **Find overlapping genes** in the *Copy Number Analysis* section of the workflow
- As this tutorial will explore genomic integration, select **Create a new spreadsheet with genes that overlap with the regions** and select **OK**
- The database that associates genes (annotations) with cytogenomic locations must be specified in the next dialog (Figure 26). PGS offers a number of possibilities such as RefSeq and Ensembl or custom annotations (for the latter option, please use *Manage available annotations*). If a database file is outdated or not present, the user will be prompted to download the updated version before the analysis. Select **RefSeq Transcripts** and **OK** to proceed
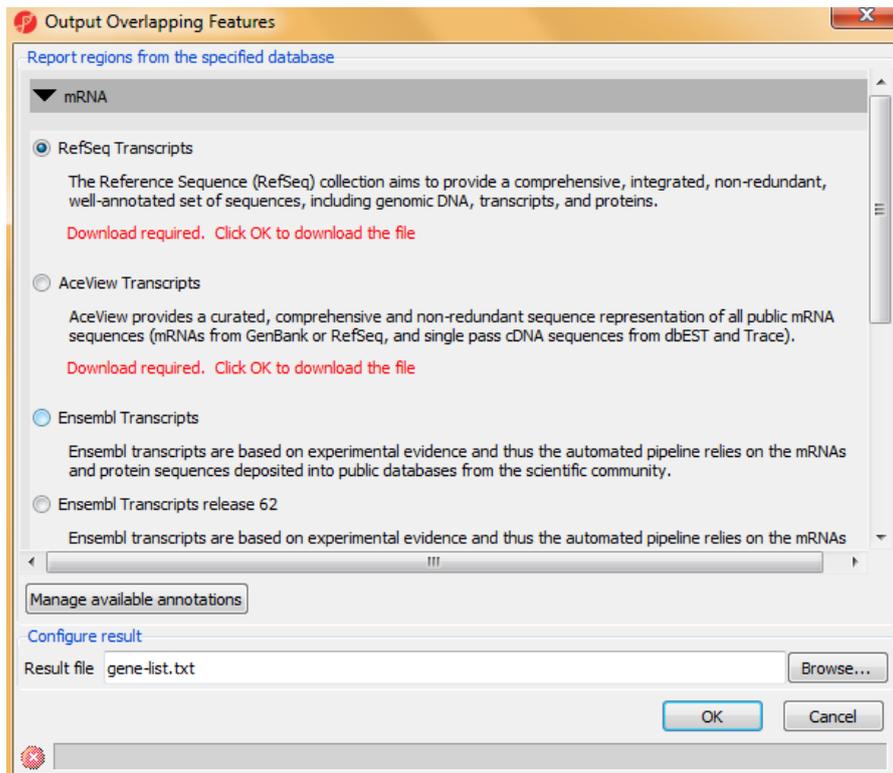
Figure 26: Viewing the Output Overlapping Features dialog. Database files not present on the computer display Download required *in red*

The resulting spreadsheet, *gene-list*, is shown on Figure 27. Each row corresponds to a transcript (RefSeq in this tutorial), and the columns are as follows:

1 – 3: Genomic coordinates of the transcript

4: Coding strand
5: Transcript ID
6: Gene symbol
7: Minimum distance of the region to the transcription start site (positive values indicate downstream, while negative values indicate upstream)
8: Percent overlap with gene: length of gene to region overlap divided by the length of the gene
9: Percent overlap with region: length of overlap divided by the length of region
10+: Correspond to the columns 1+ in the *segment-analysis* spreadsheet

	1. transcript chromosome	2. transcript start	3. transcript stop	4. strand	5. Transcript ID	6. Gene Symbol	7. Distance to TSS	8. Percent overlap with	9. Percent overlap with	10. Chromosome	11. Start	12. St
1.	22	28315364	28398668	+	NR_026963	TTC28-AS1	70641	0.15201	0.0227762	22	28386005	28
2.	22	28374002	29075854	-	NM_001145418	TTC28	133875	0.792153	1	22	28386005	28
3.	22	35462130	35483381	+	NM_001008494	ISX	14984	0.294904	0.00233832	22	35477114	38
4.	22	35653445	35691801	+	NM_001003681	HMGXB4	0	1	0.0143113	22	35477114	38
5.	22	35653445	35691801	+	NR_027780	HMGXB4	0	1	0.0143113	22	35477114	38
6.	22	35695268	35743988	+	NM_001135729	TOM1	0	1	0.0181783	22	35477114	38
7.	22	35695797	35743988	+	NM_001135730	TOM1	0	1	0.0179809	22	35477114	38
8.	22	35695797	35743988	+	NM_001135732	TOM1	0	1	0.0179809	22	35477114	38
9.	22	35695797	35743988	+	NM_005488	TOM1	0	1	0.0179809	22	35477114	38
10.	22	35695797	35743988	+	NR_024194	TOM1	0	1	0.0179809	22	35477114	38
11.	22	35695797	35743988	+	NR_024195	TOM1	0	1	0.0179809	22	35477114	38
12.	22	35731633	35731752	+	NR_037471	MIR3909	0	1	4.44009e-005	22	35477114	38
13.	22	35777060	35790208	+	NM_002133	HMOX1	0	1	0.00490574	22	35477114	38
14.	22	35796116	35820496	+	NM_006739	MCM5	0	1	0.00909659	22	35477114	38
15.	22	35937352	35950046	+	NM_014310	RASD2	0	1	0.00473634	22	35477114	38
16.	22	36002811	36013385	-	NM_005368	MB	0	1	0.00394534	22	35477114	38

Figure 27: Viewing the gene-list spreadsheet, a result of overlapping genes with regions of copy number changes. Each row of the table represents one RefSeq transcript

As previously mentioned, this type of regions-to-genes overlap is gene-centric and enables genomic integration. For instance, GO enrichment can be directly invoked on the *gene-list* spreadsheet to detect the functional groups affected by copy number changes (to learn more, please see the *GO Enrichment* Tutorial, available at *Help > On-line Tutorials > User Guides* tab

The *gene-list* spreadsheet can also be used to find possible fusion genes. To do that, you might use the interactive filter on column 8 (*Percent overlap with gene*) to get the list of genes with overlap of up to, for example, 50% (not shown). If only a part of a gene is amplified or deleted, it is possible that a translocation event took place that split the gene in two parts.

Further annotation options

PGS offers even more options for annotating the list of shared regions. For instance, the *Find overlapping genes* tool enables annotating the regions with annotations from the *Database of Genomic Variants*. As this database is an overview of known structural alterations of the human genome spanning more than one thousand base pairs, it can be used as a starting point for detection of previously described copy number alterations or detection of novel copy number alterations in these samples.

Overlap with known SNPs (under *Copy Number Analysis* stage of the workflow) invokes the dialog quite similar to the one shown in Figure 26, but with an additional option to annotate the regions with information from the dbSNP database. Two additional columns will be added to the right of the initial spreadsheet: the list of SNPs described in each region (rs numbers) and the total number of SNPs in the region. If the list of SNPs is very long, you may need to right-click on the row header and use the *Create list of dbSNP* command.

Another functionality based on annotation is *Test for known abnormalities* (also under *Copy Number Analysis* in the workflow), which enables the comparison of the regions listed in the *segmentation* spreadsheet with a database of genomic abnormalities characteristic for particular diseases or syndromes and quick identification of possible matches. The user can choose between the Partek®-distributed database (60 syndromes) or use a custom one (requires the following information organized by column: the name of the abnormality, chromosome number, starts location, and stop location). The input to *Test for known abnormalities* should be a list of aberrations for each sample; do not include unchanged regions in the input or every syndrome will be shown as positive. The power of this feature lies in the information contained in the database which can be easily created or customized.

End of Tutorial

This is the end of the Copy Number in Partek® Genomic Suite™ v6.6 tutorial. The tutorial described all the steps needed for copy number analysis: data import, detection of genomic regions with copy number changes, analysis of regions with copy number changes, and comparison of those regions with genomic databases. If you need additional assistance with this data set, please call our technical support staff at +1-314-878-2329 x350 or email at *support@partek.com*.

Last revision: October 2014

Appendix A: GC wave correction (Affymetrix® CEL files)

To normalize for GC content, use the custom import settings during the import procedure (**Customize…**) (Figure A1) and under the *Algorithm* tab of the *Advanced Import* dialog, check the **Adjust for GC content** box (Figure A2).
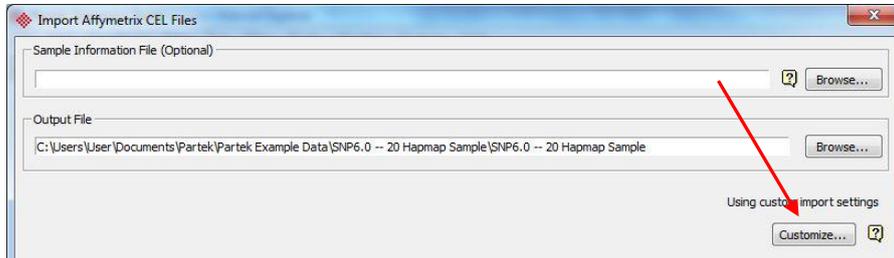
Figure A1: Accessing the custom import settings during the import of Affymetrix® CEL files
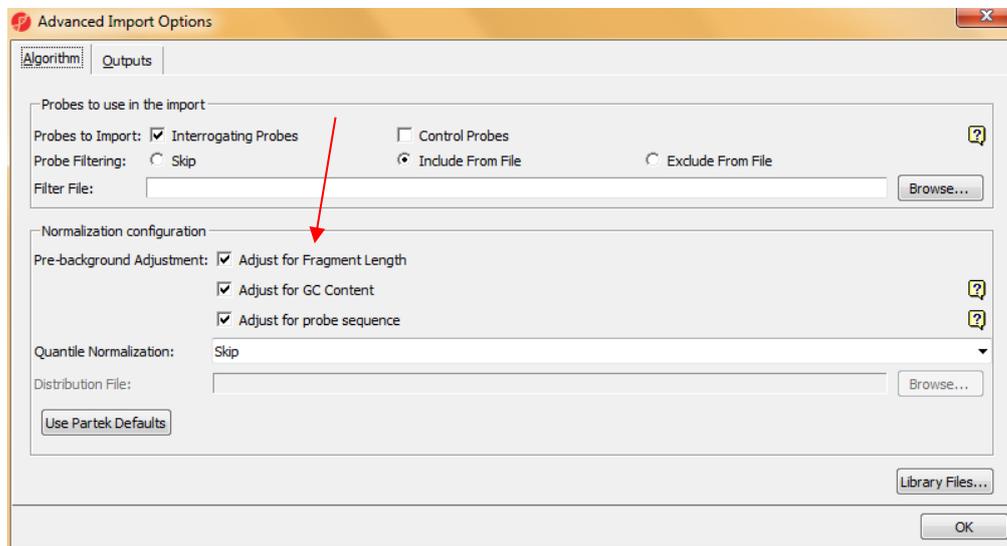
Figure A2: Adjusting for GC content during the import of CEL files

Appendix B: Filtering the FFPE samples by fragment length

Many samples used in medical research have been formalin-fixed and paraffin-embedded (FFPE). Unfortunately, storage and recovery of nucleic acids in this form can increase the rate of breaks in nucleic acids. The high density of Affymetrix® arrays allows *longer* restriction fragments which are therefore more likely to contain breaks in FFEP samples. PGS can preferentially remove probesets from long fragments from the analysis to increase the resulting signal-to-noise ratio within the sample and can result in cleaner data.
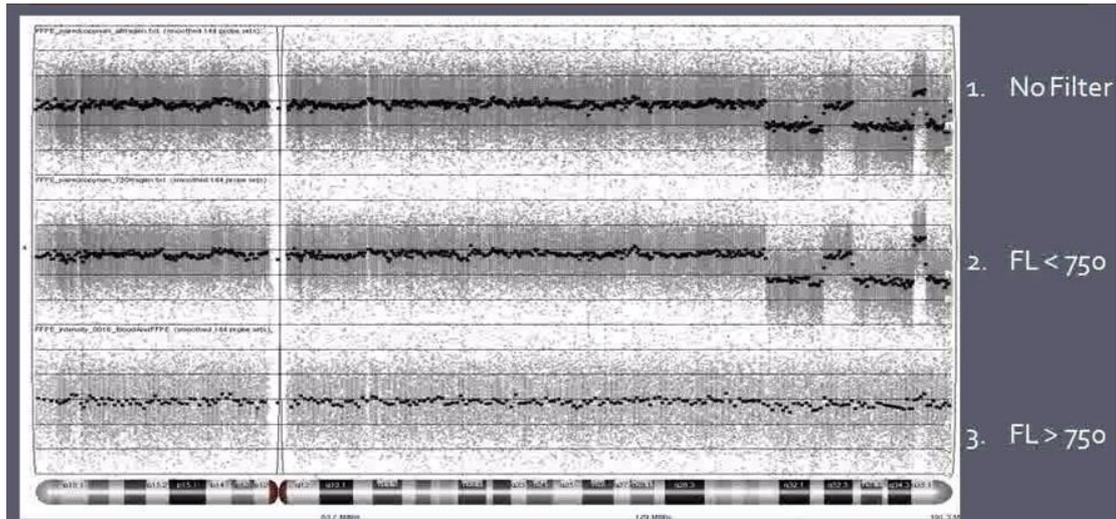
Figure B1: Probesets that hybridize to longer fragment lengths have higher signal-to-noise ratios in FFPE samples. The top track shows the inclusion of all probesets; areas of copy number deletions can be observed. The middle track shows the inclusion of probesets with fragment lengths (FL) less than 750 bp, and the same copy number deletions can be detected. The bottom track shows the data for the probesets that were filtered out; note that aberrations are smaller and harder to differentiate from the background noise. Removing such probesets decreases noise and improves statistical power

To change the distribution of restriction fragments used in the analysis, from the **Filter** menu, select **Filter Column** > **Filter by Fragment Length…** (Figure B2).
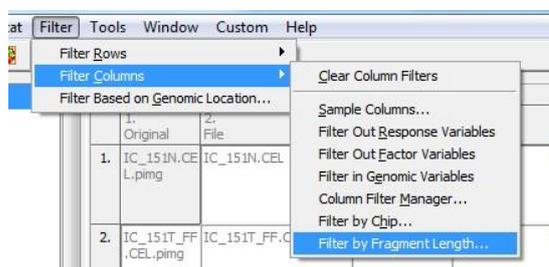
Figure B2: Accessing the Filter by Fragment Length dialog

PGS might prompt for a file containing a list of expected fragment sizes. If this file is not found, the software will download it from the Partek® servers (Figure B3) when you select **Download**.
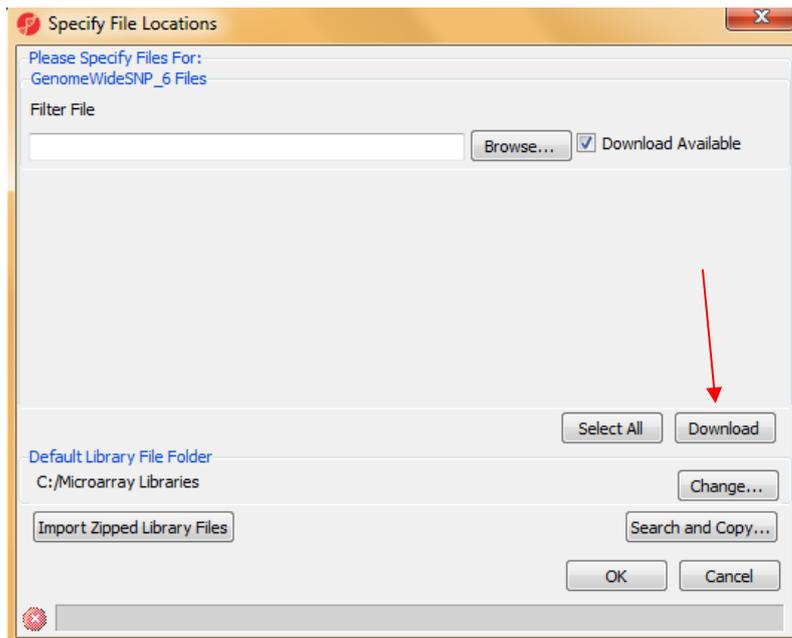
Figure B3: Download of fragment size filter from Partek® server

Once the filter file is downloaded, you'll be prompted to set the minimum and maximum fragment length (in base pairs) that you'd like to use for the analysis. A histogram of the expected overall expected distribution is displayed to aid in your selection. You can set the *Min* and *Max* by either typing in values or using the slider buttons (Figure B4). After setting the filter, select **OK** and note that a horizontal yellow bar has been introduced on the bottom of the spreadsheet indicating that the spreadsheet has been filtered on columns (not shown).
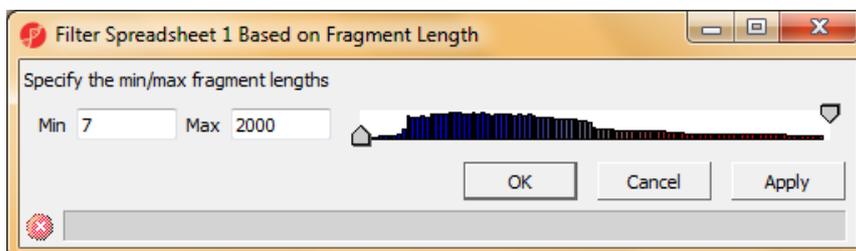
Figure B4: Setting the fragment lengths for filters

Unfortunately, there are no universal answers on how to set the limits for the fragment filter. In general, you should monitor the distribution of the fragments on agarose gels prior to hybridization on the array and use the size information from the gels to optimize the limits used for analysis as the loss of longer fragments may be different for each dataset. If no physical size data is available, removing fragments above 500 or 700 base pairs is a recommended starting point as the nucleic acids in the sample above this size may have been damaged.

Appendix C: Integration of copy number with LOH and AsCN

Although copy number analysis is a powerful tool for studying genomic aberrations, it lacks the capacity to detect genotypic changes which are copy-neutral. If we consider loss of heterozygosity (LOH), this event may be caused by a hemizygous deletion in which one allele is lost and the other allele remains present (Figure C1, middle panel). That type of LOH can be recognized not only by SNP-genotyping, but by copy-number analysis as well. However, an allele may lost initially, but the subsequent amplification of the remaining copy creates a copy-neutral LOH (Figure C1, right panel), also known as uniparental disomy. Different mechanisms have been described to create copy-neutral LOH in meiosis and mitosis; the common feature is that copy-neutral LOH can only be detected when copy number is studied in combination with SNP genotype. Please note that irrespective of the preservation of total number of copies, the biological effect is still important as recessive mutations would no longer be masked by their dominant normal counterparts.
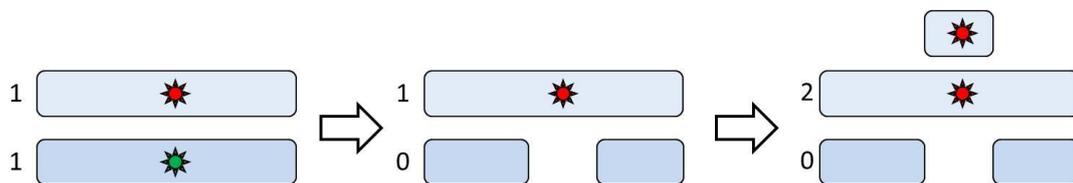
Figure C1: Possible mechanisms of LOH and their impact on copy number. Left panel: heterozygous SNP; numbers indicate the number of copies of each allele ("normal" or most common allele = green, "mutant" = red). Middle panel: hemizygous deletion leading to the loss of normal allele. Right panel: duplication of the "mutant" allele. The situation in the middle panel changes the gene copy number, while the situation in the right panel is copy-number neutral

A solution for detection of copy-neutral events is the integration of copy number workflow with LOH or the use of allelic imbalance (AI) under the Allele-Specific Copy Number (AsCN) workflow (advantages of AI over LOH are discussed below). With this approach, the copy number data are supplemented with SNP genotyping data (currently available with Affymetrix® and Illumina®) to label the genomic regions in the following fashion: amplification without LOH/AI, amplification with LOH/AI, deletion without LOH/AI, deletion with LOH/AI, copy-neutral LOH/AI (Figure C2). The last category, copy-neutral LOH/AI, is the added value of the workflow integration.
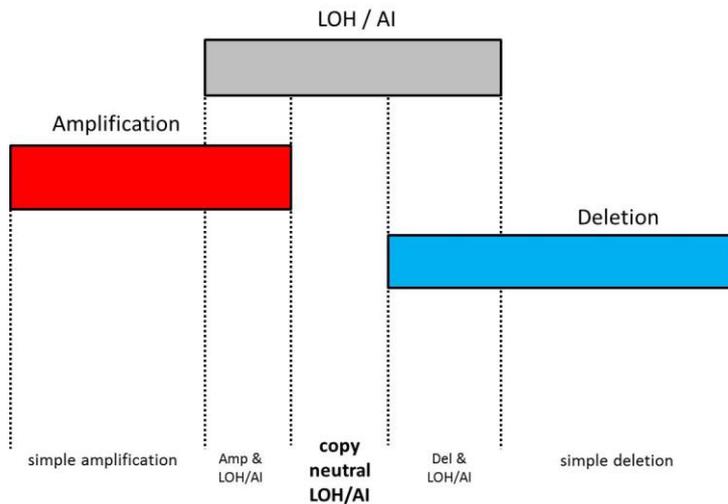
Figure C2: Integration of copy number workflow with loss of heterozygosity (LOH) or allelic imbalance (AI) under allele-specific copy number (AsCN) workflows enables the identification of copy-neutral events

A further consideration is that the correct interpretation of currently available algorithms for LOH has been proven complex and difficult because cancer cells frequently deviate from the diploid state, and tumor specimens often contain a significant proportion of normal cells. For instance, it has been shown that as the proportion of tumor cells in a sample decreases and approaches 50% or less, the capacity to detect the LOH diminishes (Yamamoto *et al*., Am J Hum Gen 2007). Moreover, the genotyping algorithms fail to call a heterozygote SNP accordingly in a situation when only one of two alleles gets amplified (e.g. 3×A and 1×B): a false positive LOH call can be the consequence.

AsCN analysis, on the other hand, is a method that enables reliable detection of allelic imbalance in tumor samples even in the presence of large proportions of normal cells. Unlike LOH, it does not require a large set of normal reference samples. For a heterozygous SNP (only those are informative), a balance is expected between the two alleles (1×A and 1×B, or 1:1 ratio). The AsCN algorithm provides an estimated number of copies of each allele and therefore enables the detection of allelic imbalance even in cases when alleles are amplified or deleted (e.g. 3×A and 1×B). Moreover, LOH can be considered a special case of AI (e.g., 1×A, B allele deleted) (Figure C3). Therefore, due to its improved robustness, AsCN may be a preferred application in tumor-focused applications.
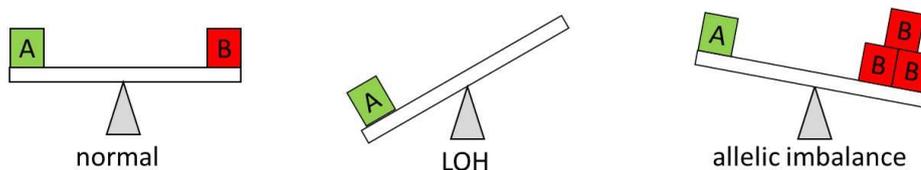
Figure C3: Loss of heterozygosity (LOH) as a special case of allelic imbalance. The situation on the left represents a normal heterozygous SNP, with one copy of each allele

References

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K. *Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms*. **Nucleic Acids Res**. 2008 Nov;36(19):e126.

Ramakrishna M, Williams LH, Boyle SE, Bearfoot JL, Sridhar A, Speed TP, Gorringe KL, Campbell IG. *Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis*. **PLoS One**. 2010 Apr 8;5(4):e9983.

Yamamoto G, Nannya Y, Kato M, Sanada M, Levine RL, Kawamata N, Hangaishi A, Kurokawa M, Chiba S, Gilliland DG, Koeffler HP, Ogawa S. *Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of Affymetrix single-nucleotide-polymorphism genotyping microarrays*. **Am J Hum Genet**. 2007 Jul;81(1):114-26.