

# Allele Specific Copy Number Analysis in Partek® Genomics Suite™ v6.6

---

## Introduction

---

Allele-specific copy number (AsCN) is a method for detection of allele dosage and allele imbalance (AI) (e.g. 3×A, 1×B allele). It provides an insight in losses and gains of alleles and, in addition, is not burdened by the limitations inherent to the loss of heterozygosity (LOH) workflow.

The major applications of the AsCN workflow include:

- detection of copy-neutral genomic events (e.g. 2×A, 0×B);
- confirmation of allele deletions detected by the copy number analysis.

Please note that the detection of copy-neutral genomic events and the confirmation of allele deletions require integration with the copy number workflow and the discussion proceeds below (for more information on the copy number analysis itself please refer to the respective tutorial available under *Help > On-line Tutorials*).

In the context of genetics, the term LOH refers to the loss of function of one allele of a gene in which the other allele was already inactivated. An advantage of the LOH analysis is that it provides a solution to a problem associated with the copy number approach: the inability to detect genotypic changes which are copy-neutral. The LOH may be caused by a hemizygous deletion in which the normal allele is lost and the mutated allele remains present (Figure 1, middle panel). That type of LOH can be recognized not only by SNP-genotyping, but by copy-number analysis as well. However, an allele can get lost initially, but the subsequent amplification of the remaining copy creates a copy-neutral LOH (Figure 1, right panel), first described as UPD (uniparental disomy). Different mechanisms have been described to create copy-neutral LOH in meiosis and mitosis, and 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, irrespectively of the preservation of total number of copies, the biological effect is still important as the recessive mutations are no longer masked by their dominant normal counterparts

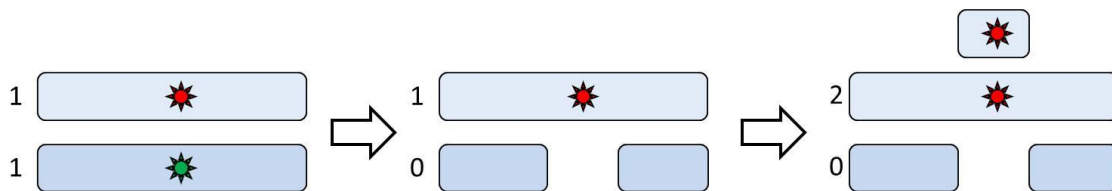


Figure 1: Possible mechanisms of loss of heterozygosity and their impact on copy number. Left panel: heterozygous SNP; numbers indicate the number of copies of each allele (normal = green, mutant = red). Middle panel: hemizygous deletion leading to the loss of normal allele. Right panel: duplication 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.

Unfortunately, LOH analysis has some important limitations: the correct interpretation of currently available algorithms for LOH has been proven complex and difficult, because cancer cells frequently deviate from diploid state and tumor specimens often contain 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 G *et al.* Am J Hum Gen 2007). Moreover, genotyping algorithms fail to call a heterozygote SNP accordingly in a situation when only one of two alleles gets amplified (e.g.  $3 \times A$  and  $1 \times B$ ): a false positive LOH result can be the consequence.

AsCN analysis, on the other hand, is a method that enables a reliable detection of allele imbalance in tumor samples even in the presence of large proportions of tumor 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 ( $1 \times A$  and  $1 \times B$ , or 1:1 ratio). AsCN algorithm provides an estimated number of copies of each allele and therefore enables the detection of AI even in cases when alleles are amplified or deleted (e.g.  $3 \times A$  and  $1 \times B$ ). Furthermore, please note that LOH can be considered a special case of AI (e.g.  $1 \times A$ , B allele deleted) (Figure 2). Therefore, due to its better robustness, the AsCN can be suggested as a preferred application in tumor focused applications.

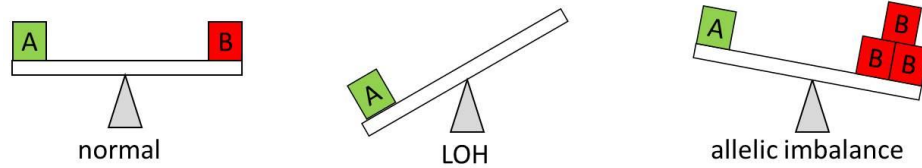


Figure 2: Allelic imbalance as a general case of loss of heterozygosity (LOH). The situation on the left represents a normal heterozygous SNP, with one copy of each allele.

The integration of copy number workflow with AsCN workflow relies on the supplementation of the copy number data with the SNP genotyping data (currently available by Affymetrix and Illumina) to label the genomic regions in the following fashion: amplification without AI, amplification with AI, deletion without AI, deletion with AI, and copy-neutral AI (Figure 3). The last category, copy-neutral AI is the added value of the workflow integration. Please note that the same five categories can be obtained by the LOH workflow as well (as discussed in the respective tutorial, available under *Help > On-line Tutorials*).

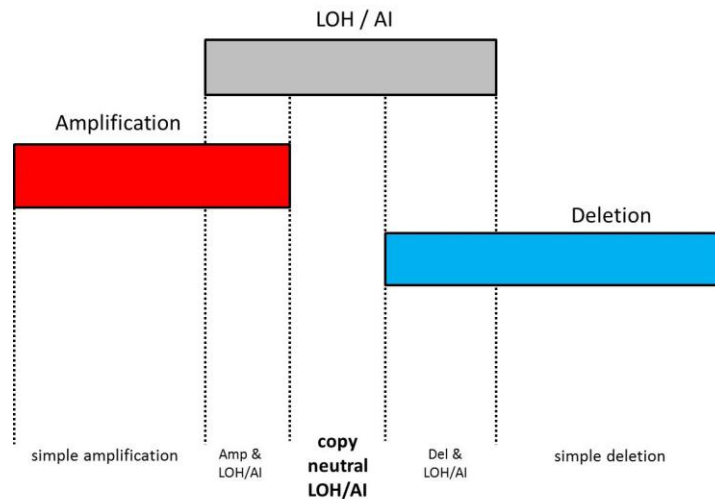


Figure 3: Integration of copy number workflow with loss of heterozygosity (LOH) or allelic-specific copy number (AsCN) workflows enables the identification of copy-neutral LOH and copy-neutral allelic imbalance (AI).

This tutorial will give you an overview of Allele Specific Copy Number workflow (Figure 4) in Partek® Genomics Suite™ (Partek GS) and will illustrate how to:

- Create allele specific copy number estimates
- Identify regions of allelic imbalance
- Visualize the results

Overlap the results with CN data

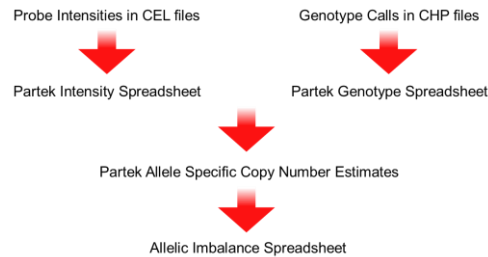


Figure 4: Key steps of the allele specific copy number analysis in Partek Genomics Suite

## Importing the Data

This example data set consists of 20 samples from an ovarian cancer study, in which fresh frozen tumor and peripheral blood samples were obtained from 10 subjects (paired design) (*PLoS One* 2010;5:e9983, GSE19539).

The raw data files (.CEL for probe intensities and .CHP for genotype calls) have already been imported into Partek GS and saved in .fmt format. They are provided on the Partek tutorials page (under Copy Number tab), found by selecting **Help > On-line Tutorials** in the Partek GS main menu. To proceed with the exercise, please download the .zip file to your computer, and unzip it.

Select **Allele Specific Copy Number** from the *Workflows* drop-down menu on the right side of the main window (Figure 5).

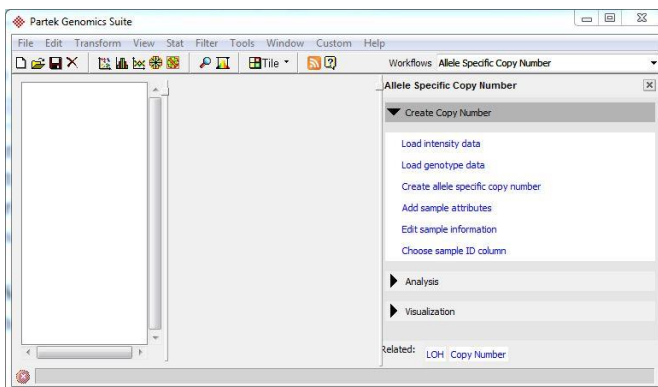


Figure 5: Selecting the Allele Specific Copy Number workflow within the Partek Genomics Suite main window

Select **Load Intensity Data**. You will be asked to import .CEL files or load an existing file as shown in Figure 6. Load the data from the files provided (*AsCN\_CEL.fmt*).

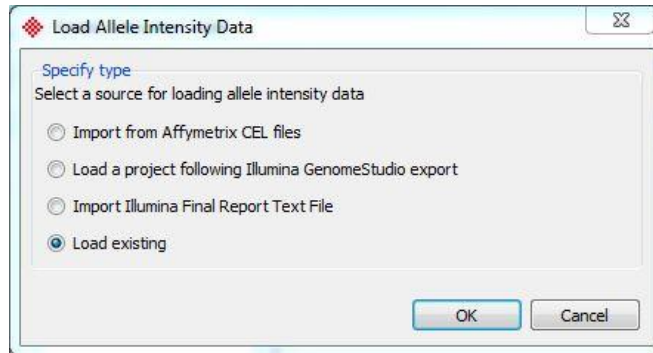


Figure 6: Loading allele intensity data.

Next, select **Load Genotype Data** from the workflow. Select **Load Existing** in the *Load Genotypes* and import the provided file (*AsCN\_CHP.fmt*). Two spreadsheets will be loaded in the main window as shown in Figure 7

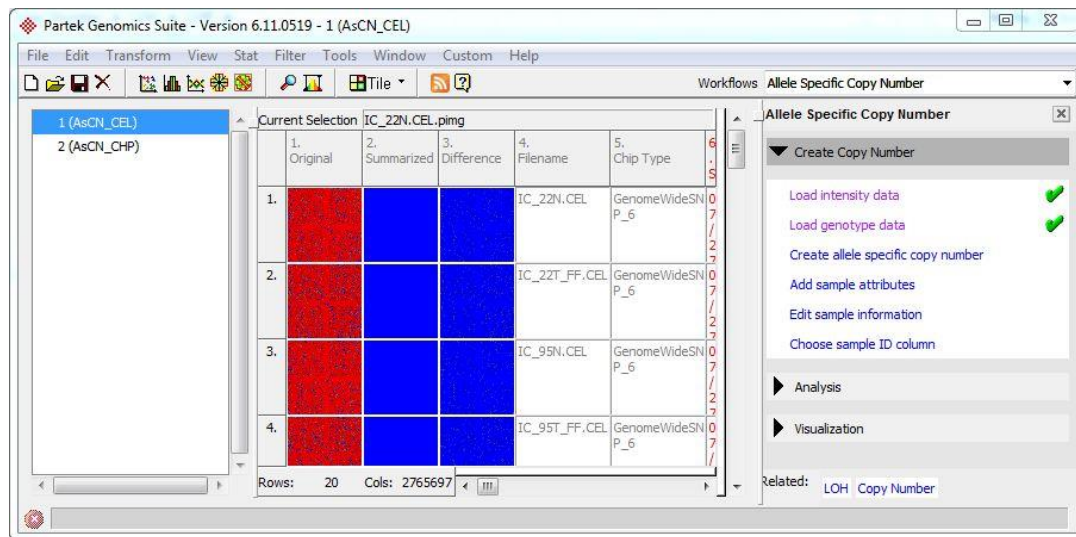


Figure 7: Viewing the Partek Genomics Suite main window after loading allele intensity and genotype spreadsheets.

## Estimating Allele Specific Copy Number

The algorithm will estimate AsCN using genotype calls of SNVs (.CHP file) and signal intensity of each allele (.CEL file). That will enable to assess allele gains or allele losses and to detect allelic imbalance with respect to reference (normal) samples. The algorithm, however, will not take into account all SNVs, as some will

be uninformative and treated as missing values. The exact procedure depends on the experimental design:

- paired analysis: reference sample is normal tissue taken from the same individual as the study (tumor) sample
- unpaired analysis: reference sample is taken from a different, healthy individual

For paired analysis Partek GS requires genotype calls for the reference sample and allele intensities for both reference and study (tumor) samples. SNVs that are heterozygous in reference samples are considered informative.

Unpaired analysis, on the other hand, requires genotype calls and allele intensities for both reference and study samples. Informative SNVs are defined as being heterozygous in study samples. As long stretches of homozygosity (i.e. loss of heterozygosity) will therefore be uninformative, paired design should be used whenever possible.

For more details on the algorithm implemented in Partek GS, please refer to the *Allele Specific Copy Number* white paper (Help > On-line Tutorials, under White Papers).

To proceed with the tutorial, please select **Create Allele Specific Copy Number** from the workflow; the dialog shown in Figure 8 will appear.

Although Partek GS will automatically pick up the intensity and genotype spreadsheet, make sure that the entries in the dialog match the spreadsheets you would like to analyze.

Select the column containing file names in each respective spreadsheet for the **Unique sample column**. This column should be unique for every sample and have the same value for both spreadsheets, as it will be used to identify the samples. Set the **Column that identifies the normal samples column** as 5. *Tumor* from the drop down menu. Choose *N* for the **Normal category**. These two fields will specify which samples are used as references.

Select *SubjectID* from the **Column that Matches the Pairs column** drop down menu. This will match reference sample to study sample taken from the same individual.

Select **OK**.

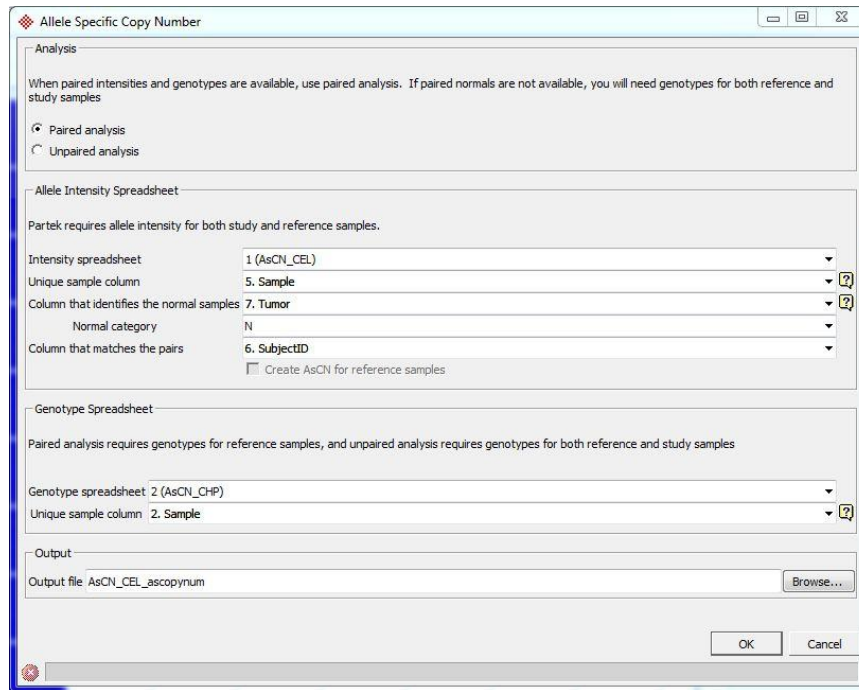


Figure 8: Configuring the Allele Specific Copy Number dialog

When complete, a new spreadsheet (*AsCN\_CEL\_ascopynum*) will be created containing AsCN for the heterozygous calls in the study (tumor) samples. Each row of the table is one of the alleles in each sample. Columns 11+ in Figure 9 represent one allele each, while the copy number values are given in cells. Max and min are determined by the values, with max to min ratio in diploid regions being around 1. The “?” indicates that the probeset is not informative.

5. Sample	6. SubjectID	7. Tumor	8. Chip Type	9. Scan Date	10. Allele	11. AFFX-SNP_10000979	12. AFFX-SNP_10009702	13. AFFX-SNP_10015773	14. AFFX-SNP_10021569
IC_22T_FF	22	T	GenomeWideSN	07/27/07	max	2.81319	0.722309	?	1.31005
IC_22T_FF	22	T	GenomeWideSN	07/27/07	min	0.442941	0.376798	?	0.549047
IC_95T_FF	95	T	GenomeWideSN	07/27/07	max	1.05918	?	?	?
IC_95T_FF	95	T	GenomeWideSN	07/27/07	min	0.480281	?	?	?
IC_151T_FF	151	T	GenomeWideSN	07/27/07	max	1.26118	1.26435	?	1.34542
IC_151T_FF	151	T	GenomeWideSN	07/27/07	min	0.289516	0.5782	?	0.599225
IC_201T_FF	201	T	GenomeWideSN	07/27/07	max	1.66667	?	?	?
IC_201T_FF	201	T	GenomeWideSN	07/27/07	min	0.403129	?	?	?
IC_258T_FF	258	T	GenomeWideSN	07/27/07	max	?	?	?	0.856292
IC_258T_FF	258	T	GenomeWideSN	07/27/07	min	?	?	?	0.733333
IC_315T_FF	315	T	GenomeWideSN	07/27/07	max	?	?	0.87704	?
IC_315T_FF	315	T	GenomeWideSN	07/27/07	min	?	?	0.676508	?

Figure 9: Viewing the allele specific copy number estimates. Each row represents one allele per sample.


## Detecting allelic imbalances

The spreadsheet with AsCN estimates is the starting point for the detection of allelic imbalances. To perform that step, please go to **Detect allelic imbalances** under the *Analysis* section of the workflow.

The algorithm will first calculate allelic imbalance scores for each SNP, defined as  $(\max - \min) / (\max + \min)$ . The score ranges 0 – 1 with the values close to one being indicative of allelic imbalance (for a comprehensive explanation please refer to the *Allele Specific Copy Number* white paper). Subsequently, genomic segmentation will be performed to find regions with similar imbalance scores and the mode of the score of each region (one per row) will be reported in the resulting spreadsheet (*imbalance*) (Figure 10).

1. Chromosome	2. Start	3. Stop	4. Length	5. Sample ID	6. Number of Markers	7. Proportion
1	564621	20887873	20323252	IC_594T_FF	1813	0.604491
1	20887873	22222356	1334483	IC_594T_FF	35	0.409049
1	22222356	27975756	5753400	IC_594T_FF	441	0.601613
1	27975756	29068117	1092361	IC_594T_FF	20	0.27999
1	29068117	40476647	11408530	IC_594T_FF	912	0.598836
1	40476647	41247203	770556	IC_594T_FF	10	0.193159
1	41247203	52073191	10825988	IC_594T_FF	855	0.611266
1	52073191	52993559	920368	IC_594T_FF	10	0.223802
1	52993559	53878129	884570	IC_594T_FF	105	0.625142
1	53878129	54036702	158573	IC_594T_FF	10	0.293625
1	54036702	59903372	5866670	IC_594T_FF	767	0.603331
1	59903372	60430710	527338	IC_594T_FF	12	0.19915
1	60430710	64541363	4110653	IC_594T_FF	475	0.601951
1	64541363	64892260	350897	IC_594T_FF	10	0.172144
1	64892260	72540269	7648009	IC_594T_FF	829	0.590336
1	72540269	73350894	810625	IC_594T_FF	10	0.138736
1	73350894	73946116	595222	IC_594T_FF	64	0.676889
1	73946116	74991961	1045845	IC_594T_FF	31	0.425126

Figure 10: Viewing the allelic imbalance spreadsheet. Each row is a region containing similar allelic imbalance scores across single nucleotide variations.

To narrow down the list of regions we will filter out the regions with lowest scores. Please select the **Interactive Filter** icon () from the main window toolbar. In the drop-down list set *Column* to **7. Proportion**. To filter out the lowest proportion scores set the *Min* value to **0.15**. The filter will adjust as shown in Figure 11 .



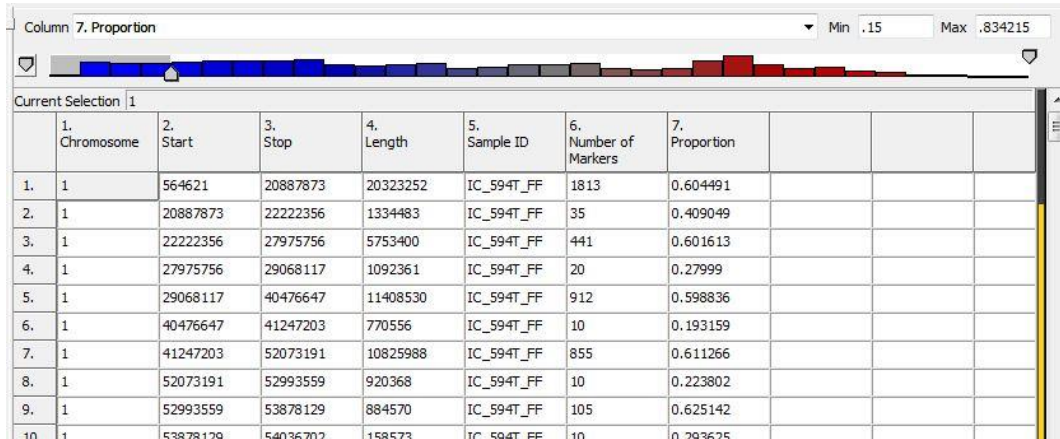


Figure 11: Setting the interactive filter to exclude the regions with proportion scores lower than 0.15. The golden column at the right indicates that the spreadsheet has been filtered.

To invoke the **Chromosome View** on the spreadsheet and get a graphic representation of the regions, please select it under the *Visualization* step of the workflow. When prompted by the *Track Wizard*, choose **Allele Specific Copy Number 3 (AsCN\_CEL\_ascopynum)** and **3/1 (imbalance.txt)** for addition to the plot (Figure 12).

Select **Create**.

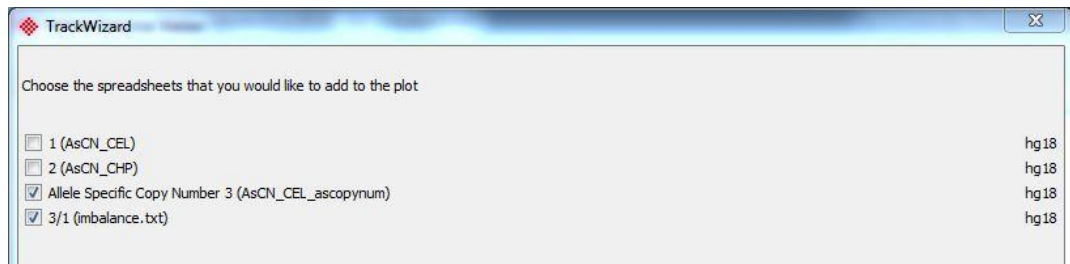


Figure 12: Configuring the Track Wizard

The *Regions track* of the Genome Viewer provides an overview of detected regions across the study sample, one sample per trace. Color codes represent the imbalance proportion. AsCN of alleles in the selected sample is given in the *Profile track* (Figure 13).



Figure 13:Chromosome view invoked on allelic imbalance spreadsheet. The tracks (from the top): Annotation track with genomic features from a selected annotation source (in this example RefSeq database), Regions track displaying the regions of similar allelic imbalance (color represents imbalance proportion), Profile track showing the selected sample (y-axis is the copy number value, dots represent alleles), Cytoband track (chromosome 1 is shown by default), and Genomic label.

## Overlapping allele specific copy number with copy number regions

The AsCN data can be overlapped with CN data using **Overlap with copy number** function of the Workflow. That allows to identify the following regions: copy-neutral allelic imbalance, amplification with allelic imbalance, amplification without allelic imbalance, deletion with allelic imbalance, and deletion without allelic imbalance (Figure 3).

The integration step requires a spreadsheet with copy number regions, the result of genomic segmentation (*Detect amplifications and deletions*) step of the *Copy Number* workflow. It is important that both spreadsheets have column with the unique identifier of each sample. For the purpose of this exercise we provided the *segmentation.txt* spreadsheet which will be used for integration. If you want to learn more about Copy Number workflow, please refer to the respective tutorial (Help > On-line tutorials).

Import the *segmentation.txt* spreadsheet in Partek GS (**File > Open...**) and select **Overlap with copy number regions** under *Analysis* section of the AsCN workflow. In the resulting dialog specify the spreadsheet with allelic imbalances (*imbalance.txt*) and the spreadsheet with copy number regions (*segmentation.txt*) (Figure 14). Select **OK**.

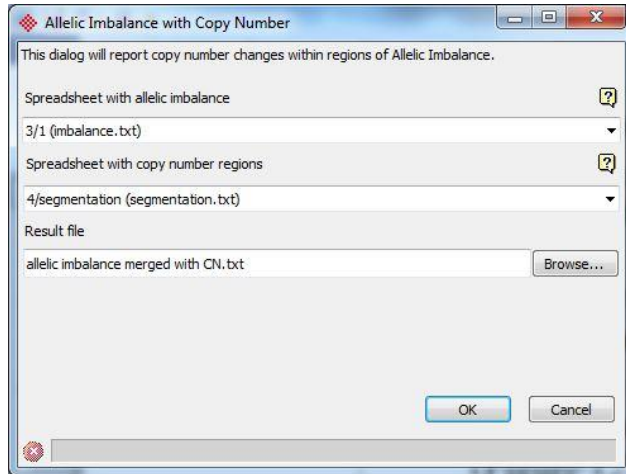


Figure 14: Configuring the integration of allelic imbalance with copy number regions

Each row of the new spreadsheet, *allelic imbalance merged with CN.txt*, is an overlapped region per sample (Figure 15).

Columns 1-3 are genomic coordinates of the region, sample is identified by column 4, and the description of the region is given in column 5. Average copy number from the segmentation (expressed as mean of the samples with the same CN result) can be found in column 6, with “?” indicating no CN change. Column 7 provides the average copy number of markers in the genomic region, while the number of CN markers is given in the column 8. Number of samples with CN changes or allelic imbalance can be found in columns 9-11.

Current Selection 2.06509											
	1. chromosome	2. start	3. end	4. Sample ID	5. Description	6. Average copy number from segmentation result	7. Average copy number of markers in the genomic region	8. # copy number markers in the genomic region	9. # Amplification Samples	10. # Deletion Samples	11. # Allelic Imbalance Samples
4601.	2	13178648	14522253	IC_315T_FF	Copy-Neutral Allelic Imbalance	?	2.06509	900	6	1	6
4602.	2	14522253	14643966	IC_580T	Amplification with Allelic Imbalance	2.98917	2.38702	74	6	1	6
4603.	2	14522253	14643966	IC_201T_FF	Amplification with Allelic Imbalance	2.67144	2.8863	74	6	1	6
4604.	2	14522253	14643966	IC_22T_FF	Amplification with Allelic Imbalance	2.73143	2.6653	74	6	1	6
4605.	2	14522253	14643966	IC_399T_FF	Amplification with Allelic Imbalance	3.09584	5.01636	74	6	1	6
4606.	2	14522253	14643966	IC_594T_FF	Amplification without Allelic Imbalance	2.75867	2.67857	74	6	1	6
4607.	2	14522253	14643966	IC_151T_FF	Amplification without Allelic Imbalance	2.63549	2.48746	74	6	1	6
4608.	2	14522253	14643966	IC_504T	Deletion with Allelic Imbalance	1.21544	1.23261	74	6	1	6
4609.	2	14522253	14643966	IC_315T_FF	Copy-Neutral Allelic Imbalance	?	2.00353	74	6	1	6
4610.	2	14643966	14687786	IC_580T	Amplification	2.79334	2.25837	38	6	1	6
4611.	2	14643966	14687786	IC_201T_FF	Amplification	2.86084	2.89537	38	6	1	6

Figure 15 Viewing allelic imbalance regions merged with copy number change regions. Each row is one region per sample.

Overlapped regions shared by multiple samples can then be filtered by using **Find regions in multiple samples** option of the AsCN workflow (under *Analysis*). For this tutorial we shall find the regions that are significant in at least **5** samples (Figure 16). Configure the dialog as shown in Figure 16 and select **OK**.

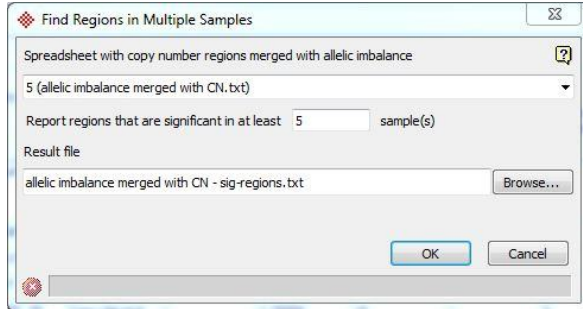


Figure 16: Configuring the Find Regions in Multiple Samples dialog.

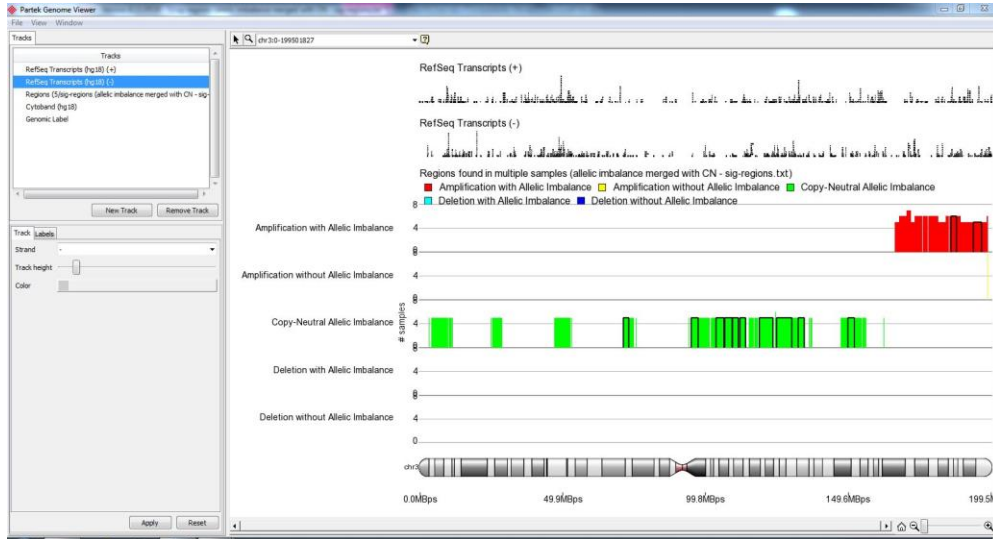
The *sig-regions.txt* spreadsheet (Figure 17) shows one significant region per row, and the columns are organized as follows:

- 1 – 5: genomic coordinates of the significant region
- 6: number of samples sharing the region
- 7: the list of samples sharing the region
- 8: description of the significant region
- 9: the average value in the region detected by genomic segmentation across all the significant samples (“?” indicates copy neutral region)

Current Selection   2									
	1. chromosome	2. start	3. end	4. cytoband	5. length (bps)	6. # samples	7. Samples	8. Description	9. average (from 5)
1.	2	27756526	28085489	2p23.2	328964	5	IC_201T_FF IC_22T_FF IC_399T_FF IC_504T IC_580T	Amplification with Allelic Imbalance	2.79686
2.	2	45762177	45865519	2p21	103343	5	IC_151T_FF IC_201T_FF IC_22T_FF IC_399T_FF IC_580T	Amplification with Allelic Imbalance	2.6612
3.	2	45865519	45975550	2p21	110032	5	IC_151T_FF IC_201T_FF IC_22T_FF IC_399T_FF IC_580T	Amplification with Allelic Imbalance	2.6612
4.	2	64704901	65140078	2p14	435178	5	IC_201T_FF IC_22T_FF IC_399T_FF IC_580T IC_594T_FF	Amplification with Allelic Imbalance	2.81309
5.	2	65399085	65497154	2p14	98070	5	IC_151T_FF IC_201T_FF IC_22T_FF IC_399T_FF IC_580T	Amplification with Allelic Imbalance	2.82558
6.	2	65497154	65588160	2p14	91007	5	IC_151T_FF	Amplification	2.86407
7.	2	65588160	65627144	2p14	38985	5	IC_151T_FF	Amplification	2.86407
8.	2	65627144	65724880	2p14	97737	5	IC_151T_FF	Amplification	2.86407

Figure 17: Significant regions found in multiple samples.

The significant regions can be visualized by invoking the **Chromosome view** on the *sig-regions.txt* spreadsheet. In the *Track Wizard* accept the default to include the **Regions found in multiple samples** in the view and select **Create**. The height of the stacks in the *Regions* track shows the count of significant regions, while the colors of the stacks describe the nature of the region (Figure 18).



*Figure 18:Chromosome view invoked on significant regions spreadsheet. The tracks (from the top): Annotation track with genomic features from a selected annotation source (in this example RefSeq database), Profile track displaying significant regions shared across samples (y-axis represents number of samples sharing the region, color describes the region), Cytoband track (here set to chromosome 3), and Genomic label.*

## End of Tutorial

---

This is the end of the allele specific copy number analysis tutorial. If you need additional assistance with this data set, you can call our technical support staff at +1-314-878-2329 or email our technical support staff at [support@partek.com](mailto:support@partek.com).

Last revision: July 2011