

Analyzing Loss of Heterozygosity in Partek[®] Genomics Suite[™] v6.6

Introduction

Loss of heterozygosity (LOH) is the loss of function of one allele of a gene in which the other allele was already inactivated. The LOH algorithm, as implemented by Partek[®] Genomics Suite[™] (Partek GS), looks across many continuous markers (provided by array vendors) to detect regions heterozygous in 'normal' samples (e.g. 1×A, 1×B allele) that are homozygous (e.g. AA or BB) in cancer (or study) samples.

The workflow can be successfully applied for the three main applications:

- detection of copy-neutral LOH;
- confirmation of allele deletions detected by the copy number analysis;
- detection of homozygosity in germline environment.

The detection of copy-neutral LOH 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*). Detection of homozygosity in germline environment (also known as homozygosity mapping or uniparental disomy (UPD) mapping) may be used in conjunction with duo or trio workflows to detect identity by descent (IBD). To learn more on duo and trio workflows, please consult the respective documentation.

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. 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.

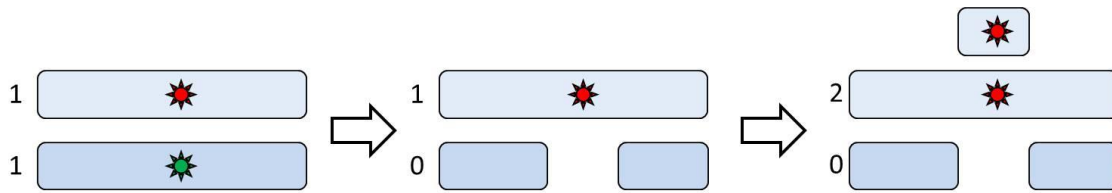


Figure 1: Possible mechanisms of LOH and their impact on copy number. Left panel: heterozygous SNP; numbers indicate the number of copies of each allele (normal = green, mutant = red). Middle panel: hemizygous deletion leading to the loss of normal allele. Right panel: duplication of the normal allele. The situation in the middle panel changes the gene copy number, while the situation in the right panel is copy number neutral.

The integration of copy number workflow with LOH 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 LOH, amplification with LOH, deletion without LOH, deletion with LOH, and copy-neutral LOH (Figure 2). The last category, copy-neutral LOH is the added value of the workflow integration. Please note that the same five categories can be obtained by the allele-specific copy number (AsCN) workflow as well (as discussed in the subsequent sections).

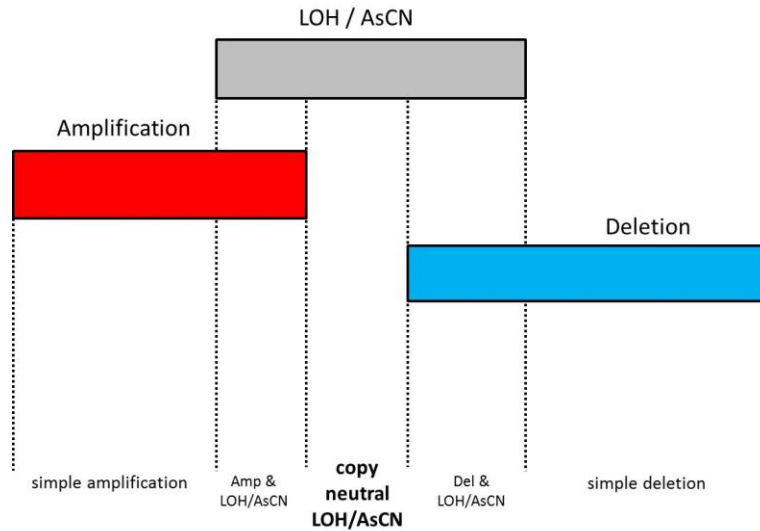


Figure 2: Integration of copy number workflow with loss of heterozygosity (LOH) or allele-specific copy number (AsCN) workflows enables the identification of copy-neutral events.

Unfortunately, LOH has the limitations on its own: 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×A and 1×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×A and 1×B, or 1:1 ratio). 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 allelic imbalance (e.g. 1×A, B allele deleted) (Figure 3). Therefore, due to its better robustness, the AsCN can be suggested as a preferred application in tumor focused applications. To learn more about AsCN, please refer to the AsCN tutorial (*Help > On-line Tutorials*).

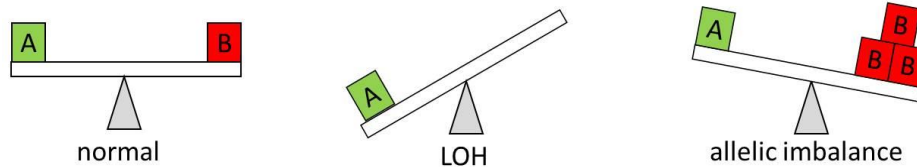


Figure 3: Loss of heterozygosity (LOH) as a special case of allelic imbalance. The situation on the left represents a normal heterozygous SNP, with one copy of each allele.

In this tutorial we will illustrate how to:

- Import a Partek GS project (.ppj) file
- Perform quality assurance (QA)/quality control (QC)
- Detect LOH regions
- Overlap LOH data with copy number data
- Annotate genes and regions of known genomic variation to our regions detected
- Filter our regions to the narrow list of regions of interest
- Visualize the results of LOH analysis

Importing a Partek Project

This example data set consists of 20 samples from an ovarian cancer study, in which paired fresh frozen tumor and peripheral blood samples were obtained from 10 women (*PLoS One* 2010;5:e9983, GSE19539). Please note that Affymetrix data have been chosen for the illustration purposes only and that the same workflow can be used to analyze data generated by other platforms.

The raw data files (.CEL for probe intensities and .CHP for genotype calls) have already been imported into Partek GS and organized as a Partek project. All the files are provided on Partek's 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.

To open a Partek project, please go to **File > Open Project...** and browse to the folder containing the tutorial data set, and select *LOH_project.ppj*. Partek will automatically open 4 spreadsheets with proper parent/children relationships between them (Figure 4).

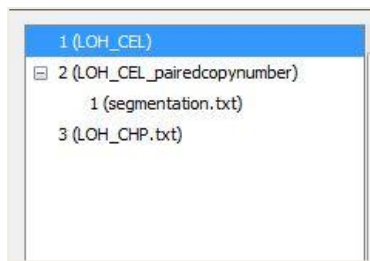


Figure 4: Viewing the Spreadsheet Tree that displays all of the spreadsheets that are currently open in Partek

A brief description of each spreadsheet and how it was obtained is outlined below.

1 (LOH_CEL)

This spreadsheet is generated from the imported SNP 6.0 CEL files showing all 20 samples on rows. Columns 1-9 contain information describing the experiment such as file names of the CEL files, Subject ID, Patient ID, etc. The rest of the columns are individual markers from the microarray with the detected intensities associated with each marker/sample.

2 (LOH_CEL_pairedcopynumber)

This spreadsheet was generated from the *LOH_CEL* spreadsheet by using the normal samples as a baseline for the paired tumor sample and converting the intensities into copy number space. For details please refer to the *Copy Number* tutorial (*Help > On-line Tutorials*).

2/1 segmentation

This spreadsheet is a child of the *LOH_CEL_pairedcopynumber* spreadsheet and was generated from the same spreadsheet by applying the Partek genomic segmentation algorithm to detect regions of amplification or deletion in the genome of the tumor samples. When applying the genomic segmentation algorithm the number of minimum genomic markers was set to **50**, and the other settings were default. Each row of the segmentation spreadsheet describes one region of amplification or deletion detected per sample. The columns describe the genomic location of the aberrant region, the sample the aberrant region was found in, and other features of the aberration such as length, # of markers in the region, etc. For details please refer to the *Copy Number* tutorial (*Help > Online Tutorials*).

3 (*LOH_CHP*)

This spreadsheet is generated from the imported SNP 6.0 .CHP files showing all 20 samples on rows. Columns 1-4 contain information describing the experiment such as file names of the .CHP files, Subject ID, etc. The rest of the columns are the SNP probes on the microarray with the genotype calls associated with each SNP/sample.

QA/QC of the LOH samples

On *LOH_CHP.txt* spreadsheet (spreadsheet **3**), please select the **LOH** workflow from the drop-down list in the upper right corner (Figure 5).

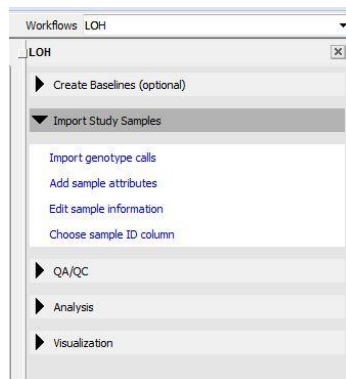


Figure 5: Selecting the Loss of Heterozygosity workflow within the Partek Genomics Suite main window.

Select the **QA/QC** section of the workflow and go to **Sample QC**. The resulting spreadsheet, *sample_QC*, shows the average no call (NC) and heterozygous (het) rate of each sample (Figure 6).

Current Selection IC_22N.birdseed.chp						
	1. GenomeWideSNP_6_Filename	2. SampleID	3. Tumor	4. PatientID	5. Sample NC Rate	6. Sample Het Rate
1.	IC_22N.birdseed.chp	IC_22N	N	22	0.0132077	0.291605
2.	IC_22T_FF.birdseed.chp	IC_22T_FF	T	22	0.0364657	0.232838
3.	IC_95N.birdseed.chp	IC_95N	N	95	0.0148996	0.297695
4.	IC_95T_FF.birdseed.chp	IC_95T_FF	T	95	0.0421307	0.262692

Figure 6: Viewing the sample_QC spreadsheet. Average no call (NC) and heterozygous (het) rates of each sample are given.

Next, go to the **Hardy-Weinberg equilibrium** step of the QA/QC section (please, make sure that the LOH_CHP spreadsheet is selected). The resulting spreadsheet, *HWE_1*, is shown in Figure 7.

Current Selection SNP_A-8343709										
	1. SNP	2. p-value (Exact)	3. Chi ²	4. p-value (Chi ²)	5. A Freq	6. B Freq	7. MAF	8. NC Freq	9. Het Freq	10. Homoz Freq
157686	SNP_A-8343709	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157687	SNP_A-8568070	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157688	SNP_A-8515425	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157689	SNP_A-8283240	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157690	SNP_A-8464522	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157691	SNP_A-8595454	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157692	SNP_A-1785667	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157693	SNP_A-8325006	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157694	SNP_A-8515584	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157695	SNP_A-2026997	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157696	SNP_A-8681526	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157697	SNP_A-8349763	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157698	SNP_A-8320714	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157699	SNP_A-8483225	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157700	SNP_A-8297783	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35
157701	SNP_A-8393126	0.109069	4.04734	0.0442411	0.684211	0.315789	0.315789	0.05	0.6	0.35
157702	SNP_A-8560181	0.109069	4.04734	0.0442411	0.315789	0.684211	0.315789	0.05	0.6	0.35

Figure 7: Result of Hardy-Weinberg equilibrium analysis of single nucleotide polymorphisms included in the .CHP files. Each row is a single SNP

Partek GS takes the data from the .CHP files to calculate the frequencies of each SNP, and displays them as one SNP per row. The description of the columns is as follows.

- 1: SNP ID (provided by the vendor)
- 2: p-value of the Fisher's exact test to test the difference between the observed and expected frequencies
- 3 & 4: χ^2 and the respective p-value to test the difference between the observed and expected frequencies
- 5 & 6: frequencies of the two alleles (A & B)
- 7: minor allele frequency (MAF); corresponds to the less frequent of the two alleles
- 8: no call (NC) frequency
- 9 & 10: frequencies of heterozygous (het) and homozygous genotypes (respectively)

Note: Column #1 shows SNP IDs as provided by the manufacturer of the array. To display the dbSNP ID, right click on the header of column 1 > **Insert Annotation**, check the **dbSNP RS ID** box, and select **OK** (not shown).

LOH analysis

Partek's LOH algorithm uses the hidden Markov model (HMM) to find regions that are most likely to be loss events based on the genotyping error and the expected heterozygous frequency at each SNP. If a paired design is used, then cancer sample and normal sample are harvested from the same individual, but the normal sample is taken from a tissue not affected by the cancer (quite often the peripheral blood). Unpaired design, on the other hand, assumes that the normal sample comes from another source. A baseline file, based on healthy individuals, is then used to determine the *expected* rate of heterozygosity. It is important to emphasize that the unpaired design does not actually detect LOH in strict sense, but rather runs of homozygosity in regions that are expected to be heterozygous. As SNPs are known to be population specific, the interpretation of the data should be performed with caution in cases when baseline files of other populations are used. Therefore, the paired design is preferred whenever possible as it is more accurate in its expected genotype frequencies and does not report regions of LOH caused by common haplotype blocks within the study population. For more details on the algorithm, please refer to the *LOH Analysis in Partek Genomics Suite* white paper (*Help > On-line Tutorials*, under White Papers tab).

Since paired design is used for the present exercise (both cancer sample and peripheral blood sample available from the same individual), please refer to Appendix A to learn how to set-up the analysis of the unpaired data (once the LOH spreadsheet has been created, the downstream steps are equal for both designs).

The LOH analysis is invoked by selecting the **Create LOH** in the *Analysis* step and specifying the experiment design (Figure 8). Please select **Paired** to proceed.

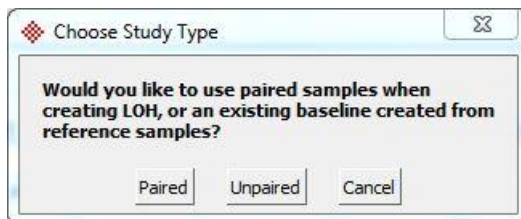


Figure 8: Choose experiment design

For paired LOH, first specify which samples are normal (*Normal column* and *Normal category* boxes) and how should the cancer samples be paired with the normal samples (*Subject column*)

Max probability specifies the maximum probability of a SNP retaining the same genotype as the previous adjacent SNP. *Genomic decay*, is the number of base pairs between neighbouring SNPs (set to "0" to inactivate and to disable the max probability effect for every SNP to SNP transition). *Genotype error* describes expected error in genotype calls of the data.

Accept the defaults and select **OK** to proceed (Figure 9).

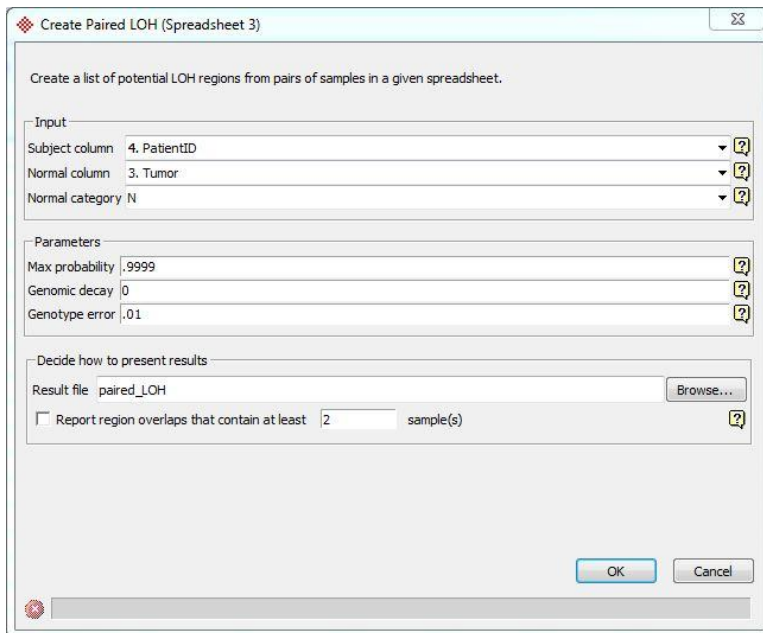


Figure 9: Configuring the Create loss of heterozygosity dialog for paired samples.

The spreadsheet *paired_loh* is created as a child of *LOH_CHP* (Figure 10). The HMM used paired normal samples as a baseline to find regions of LOH. Each row describes one region of LOH detected per study ('cancer') sample. The columns describe the following.

- 1 – 3: genomic location of the LOH region
- 4: the sample the LOH region was found in
- 5: length of the region in basepairs
- 6: heterozygous rate of the region (number of heterozygous calls divided by the total number of genotype calls in the region)
- 7: number of markers in the region

Current Selection		chr 1					
	1. chromosome	2. start	3. end	4. Sample ID	5. length(bps)	6. heterozygous rate	7. # markers
1.	chr 1	1647686	3089849	IC_22T_FF.birdseed.chp	1442164	0.0539216	204
2.	chr 1	3222627	3732639	IC_22T_FF.birdseed.chp	510013	0.064	125
3.	chr 1	4004250	4183258	IC_22T_FF.birdseed.chp	179009	0.0361446	83
4.	chr 1	4364226	4693931	IC_22T_FF.birdseed.chp	329706	0.0534351	262
5.	chr 1	4763189	4911172	IC_22T_FF.birdseed.chp	147984	0.0595238	84
6.	chr 1	4971217	6645867	IC_22T_FF.birdseed.chp	1674651	0.0488889	675
7.	chr 1	6803747	6899863	IC_22T_FF.birdseed.chp	96117	0	22
8.	chr 1	7019816	7273020	IC_22T_FF.birdseed.chp	253205	0.130435	92

Figure 10: Viewing the regions with loss of heterozygosity. Each row represents one region per study ('cancer') sample.

Finding Overlaps between LOH and Copy Number

This imported project has regions of amplification and deletion detected in the tumor samples in the *segmentation* spreadsheet, while the regions of LOH have just been identified in the *paired_loh* spreadsheet. With the *Overlap with copy number regions* tool in Partek, you can compare the *segmentation* and *paired_loh* spreadsheets to see how the aberrant regions and LOH overlap with each other. This will allow for the identification of the five regions described in the Introduction (please see Figure 2).

In order to compare the *segmentation* spreadsheet with the *paired_loh* spreadsheet, the sample ID must match between the two spreadsheets. If they do not, the *Overlap With Copy Number Regions* tool will not be able to match regions of LOH with the CN regions. When there is no common sample identifier between two spreadsheets the overlapping regions will be called incorrectly, (i.e. regions of Amplification with LOH and Deletion with LOH will not be identified and all regions of LOH will be reported as Copy-Neutral LOH).

There are different types of copy number and LOH tables that may be overlapped. Tables such as the *segmentation* and *paired_loh* report one region per sample. Using the **Find Regions in Multiple Samples** tool in Partek can generate copy number or LOH tables, often referred to as a *sig-region* table that only report regions that are common in n number of samples.

Sig-region tables exclude any regions that are not common in n number of samples. Because of this copy number regions of amplification or deletion are excluded from the *sig-region* table if they do not meet the n number threshold. If there is a region of LOH reported in a region that is rarely amplified, and the overlap is performed on *sig-region* tables, this whole region will be called copy-neutral LOH instead of amplification with LOH in those samples that were amplified. This is because the amplification is filtered out in the *sig-region* table and there appears to be no copy number variation in the region of LOH, despite the fact there are some amplifications with LOH in rare samples. The same holds true if there is a region of LOH that does not meet the n threshold. If there is a region of deletion in the same region as the excluded LOH, then it will be reported as deletion without LOH when in fact there are some samples in that region that are deletion with LOH. Using the *sig-region* tables to find overlap will decrease the number of regions reported, but at the cost of increasing the error rate of the overlap calls. The same holds true if filtering is performed on the tables before performing the *Overlap with copy number regions* step. To get the most inclusive and accurate overlap calls, use the *segmentation* and *paired_loh* results with one sample per region reported in each row.

To proceed with the exercise and to perform the copy number overlap with LOH, please follow these steps.

- From the *Analysis* section of the workflow, select **Overlap with copy number regions**

- A dialog box will appear asking to specify the spreadsheet with LOH regions and another with genomic regions. Select the **3/paired_loh** and **2/segmentation** spreadsheets so that the dialog matches the figure below. (Figure 11)
- After the spreadsheets with LOH regions and copy number regions are selected, select **OK**
-

Note: For the overlap function to work properly the entries in the sample ID columns (i.e. unique identifier of each sample) have to match between the two spreadsheets. To identify the sample ID, please use the *Choose sample ID column* option of the workflow.

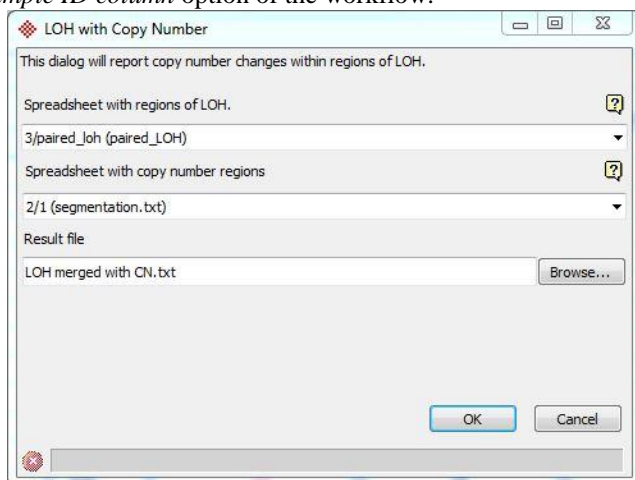


Figure 11: Setting up the loss of heterozygosity with copy number overlap dialog

Partek will create a new spreadsheet called *LOH merged with CN* (Figure 12). Each row of this spreadsheet is one genomic region of CN/LOH reported per sample. The genomic region detected in each row will be described in one of five ways (Figure 2):

Copy-Neutral LOH: A region of LOH detected where there is no overlapping amplification or deletion detected in the same sample.

Deletion with LOH: A region of LOH detected where there is an overlapping deletion detected in the same sample.

Deletion without LOH: A region of deletion is detected where there is no overlapping LOH detected in the same sample.

Amplification with LOH: A region of LOH detected where there is an overlapping amplification detected in the same sample.

Amplification without LOH: A region of amplification is detected where there is no overlapping LOH detected in the same sample.

The columns of the *LOH merged with CN.txt* contain the following information:

- 1-3: Identifies the genomic location of the region
- 4: Identifies the sample
- 5: Identifies the region by the type of overlap detected
- 6: Contains the average copy number of the region

- 7: Reports the heterozygous rate of the region. The heterozygous rate is the number of heterozygous calls (AB) divided by the total number of genotype calls. “?” indicates that no LOH was detected.

Current Selection 1							
	1. chromosome	2. start	3. end	4. Sample ID	5. Description	6. Average copy number from segmentation result	7. Heterozygous rate from LOH result
1.	1	61735	761550	IC_201T_FF	Amplification without LOH	2.3404	?
2.	1	767031	789973	IC_504T	Amplification without LOH	2.63728	?
3.	1	767031	789973	IC_22T_FF	Amplification without LOH	3.03345	?
4.	1	767031	789973	IC_399T_FF	Amplification without LOH	2.45245	?
5.	1	789973	792480	IC_504T	Amplification without LOH	2.63728	?
6.	1	789973	792480	IC_22T_FF	Amplification without LOH	3.03345	?
7.	1	789973	792480	IC_399T_FF	Amplification without LOH	2.45245	?
8.	1	789973	792480	IC_594T_FF	Deletion without LOH	1.53329	?

Figure 12: Viewing the overlap between loss of heterozygosity (LOH) and copy number (CN) regions. Each row is one genomic region of LOH/CN overlap reported per sample

To have a graphical overview of the result, invoke the **Chromosome view** from the *Visualization* stage of the workflow. Make sure that **Copy number** spreadsheet is selected for visualization in the *Track Wizard*, and click **Create**. The *Regions* track of the Chromosome view depicts the LOH/CN regions by color-coding, one chromosome at a time. Each track represents one sample (Figure 13).



Figure 13: Chromosome view depicting the overlap of loss of heterozygosity (LOH) and copy number (CN) regions. Tracks (from the top): reference track (in this example RefSeq transcripts), regions track (LOH/CN regions are color-coded, each trace represents one sample), cytoband track (chromosome 1 is shown by default), genomic label. hg18: human reference genome, revision 18

LOH with Copy Number tables can become very large because we are looking at all of the possible overlaps of copy number and loh for each sample. As shown in Figure 11, one region can be broken into multiple regions. This is why when the *LOH merged with CN.txt* spreadsheet is created there are 109 689 regions detected.

Filtering is one possibility in making this table more manageable, but perhaps the quickest way to hone in on regions of interest is to find regions in the *LOH merged with CN.txt* spreadsheet that are common over multiple samples.

To perform this function, follow these steps:

- Select **Find Regions in Multiple Samples** in the *Analysis* section of the workflow. A *Find Regions in Multiple Samples* dialog box appears (Figure 14) asking to specify the spreadsheet to perform this function on as well as the number of samples that a region must be common to before being reported in the resulting table
- Ensure that your *Find Regions in Multiple Samples* dialog box is set up as shown below (Figure 14)
- Select **OK**

Partek GS will report regions that are similar in at least 5 of the 10 samples in the cancer study from the *LOH merged with CN.txt* spreadsheet. Of course, this value should be independently determined for each experimental design.

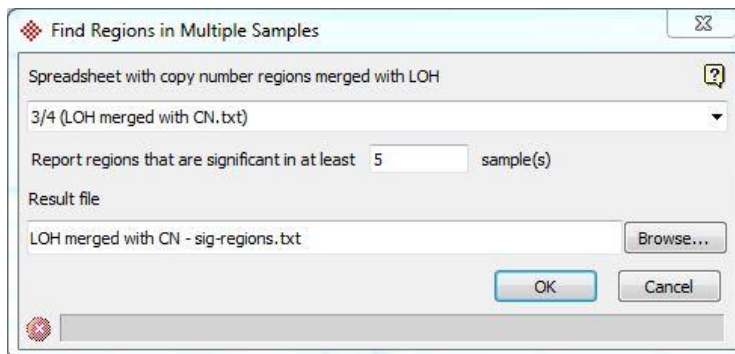


Figure 14: Configuring the *Find Regions in Multiple Samples* dialog box

A new table is (Figure 15) created as a child of the *LOH merged with CN.txt* spreadsheet called *sig-regions (CN with LOH-sig-regions.txt)* that reports genomic regions with the same CN\LOH description in at least 5 samples. By finding common regions of CN\LOH overlap amongst 50% of the samples, the number of regions reported has dropped from 109 689 rows down to 4412 rows.

Each row of the *sig-regions* spreadsheet shows one shared region (between CN and LOH) and the columns are organized as follows.

- 1-4: Genomic location of the region
- 5: Length of the region (in base pairs)
- 6: Number of samples in which the region was detected
- 7: Identifies the samples with overlapped regions
- 8: Identifies the region by the type of overlap detected
- 9: Average copy number of the regions shared across samples in column 7
- 10: Reports the heterozygous rate of the region. The heterozygous rate is the number of heterozygous calls (AB) divided by the total number of genotype calls. “?” indicates that no LOH was detected.

	1. chromosome	2. start	3. end	4. cytoband	5. length (bps)	6. # samples	7. Samples	8. Description	9. average (from 3/4)	10. Heterozygous rate from LOH result
3098.	20	62911343	62956154		44812	6	IC_151T_FF IC_201T_FF IC_258T_FF IC_399T_FF IC_504T IC_580T	Amplification without LOH	3.29009	?
3099.	MT	408	16149		15742	10	IC_151T_FF IC_201T_FF IC_22T_FF IC_258T_FF IC_315T_FF IC_399T_FF IC_504T IC_580T IC_594T_FF IC_95T_FF	Amplification without LOH	3.9144	?
3100.	17	113034	140486	17p13.3	27453	5	IC_151T_FF IC_201T_FF IC_258T_FF IC_399T_FF IC_95T_FF	Copy-Neutral LOH	1.93362	0.0893624
3101.	17	700998	705057	17p13.3	4060	5	IC_151T_FF IC_258T_FF IC_399T_FF IC_580T IC_95T_FF	Copy-Neutral LOH	1.89727	0.0879161


Figure 15: Viewing the sig-regions spreadsheet showing the overlap of loss of heterozygosity and copy number regions shared across user-defined number of samples (5 in this example). Each row represents one overlapped region.

A way to quickly see how these significant regions of CN/LOH overlap over the whole genome is to generate a *Chromosome View* plot. To invoke this visualization tool go to **Visualization > Chromosome View** in the workflow. When prompted by the *Track Wizard*, check the box **Regions found in multiple samples 4/sig-regions**, and select **Create**:

The five different CN/LOH overlap types are incorporated in the chromosome view (one chromosome at a time) to visually describe the regions reported from the *sig-regions (CN with LOH sig-regions.bed)* spreadsheet. Figure 16, for instance shows copy-neutral LOH, deletion with LOH, and deletion without LOH regions on the chromosome 17. To focus on another chromosome, please use the drop-down list in the upper right corner of the central pane. Again, as we are viewing the results of the sig regions, all of these reported regions being viewed are common in at least 50% of the samples in the data set.



Figure 16: Viewing the chromosome view displaying the overlap of loss of heterozygosity and copy number regions shared across at least 70% of the samples. Tracks (from the top): reference track (in this example RefSeq transcripts), regions track (LOH/CN regions are color-coded, y-axis shows the number of samples sharing a region), cytoband track (chromosome 17 is set in this example), genomic label. hg18: human reference genome, revision 18

From the *sig-regions* (CN with LOH-sig-regions.txt) click on the interactive filter button () and select **8. Description** from the *Column* pull-down menu to see a histogram listing the labels of the CN/LOH overlap. Moving the cursor over the labels will list the number of regions associated with label. For example, by moving the cursor green label displays that there are 280 regions of copy-neutral LOH in the *sig-regions* spreadsheet.

Annotating Genes and Genomic Variation to Copy Number and LOH

From any of the region tables that include genomic location, it is possible to annotate genes located in the regions detected using the **Find Overlapping Genes** tool from the copy number, LOH, or Tiling workflow. Follow these steps to familiarize yourself with this tool.

While viewing the *sig-regions* (CN with LOH sig-regions.bed), select **Find Overlapping Genes** from the *Analysis* portion of the LOH workflow. The *Output Overlapping Features* dialog box will open. Please, select the **RefSeq Transcripts** database for this tutorial. However, other genomic databases can be used, including custom databases by browsing to the database file by using the *Manage available annotations* button in the dialogue. Select **OK** to proceed.

The *sig-regions* (*CN with LOH-sig-regions.txt*) table now has nine new columns listing the transcripts (one transcript per row) associated with each region appended to the left (Figure 17). The new columns are briefly overviewed here.

- 1-4: Genomic location of the transcript from the RefSeq database
- 5: Transcript ID from the RefSeq database
- 6: Gene symbol (please note that some genes encode multiple transcripts)
- 7: Distance to transcription start site (positive values indicate downstream, while negative values indicate upstream)
- 8: Percent overlap with gene: length of overlap of the gene and the region, divided by the length of the gene
- 9: Percent overlap with region: length of overlap of the gene and the region, divided by the length of the region
- 10+: Correspond to the original columns of the *sig-regions* spreadsheet

Current Selection NM_015836											
	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
1.	1	118297811	118529372	-	NM_206996	SPAG17	0	0.416452	0.178253	1	118432938
2.	1	119227189	119333703	-	NM_152380	TBX15	0	1	0.253583	1	119026674
3.	1	119375362	119484819	-	NM_015836	WARS2	38109	0.651836	0.169862	1	119026674
4.	1	119375362	119484819	-	NM_015836	WARS2	5991	0.29343	1	1	119446710
5.	1	119375362	119484819	-	NM_201263	WARS2	38109	0.651836	0.169862	1	119026674
6.	1	119375362	119484819	-	NM_201263	WARS2	5991	0.29343	1	1	119446710
7.	1	119759077	119767186	+	NM_001166120	HSD3B2	0	1	0.18737	1	119758783
8.	1	119759266	119767186	+	NM_000198	HSD3B2	0	1	0.183003	1	119758783


Figure 17: Viewing the gene annotations added to the copy number with loss of heterozygosity *sig-regions* spreadsheet

Filtering Data Results in Partek

Depending on the research objective, it may be desired to focus on just a portion of the data. Partek's row, column, and interactive filters were designed for this purpose. They are powerful and flexible and can be used to easily filter-in or filter-out portions of the data based on literally any criteria in the spreadsheet.

Row filters are used to determine which rows are retained in the spreadsheet for analysis. They can easily be configured to select observations of the data based on any criteria. Column filters have the same function and use case, except on columns instead of rows.

Partek's interactive filter is part of the graphical user interface and has the same functionality as a row filter with the added benefit of displaying a bar chart of the categorical or histogram of numerical data being filtered.

Select the interactive filter button ()

When it is first opened, the interactive filter will, by default, display the class variable or the first categorical variable. The bar chart will represent a distribution of each category, and the colors of the bars are coded according to the category. The category name and the number of observations contained in a category will be displayed by using the mouseover function over each colored bar.

The first categorical variable is column 1, chromosome. The bar chart displays the distribution of regions detected per chromosome. By moving the mouse over the different bars of the bar chart, the numbers of rows/regions detected are displayed for each chromosome.

<Left-clicking> on a bar will toggle the filtering of the chosen category, right clicking on a bar will filter out all the other categories but the chosen one.

<Left-click> on the bar associated with chromosome 7 in the interactive filter. All the regions associated with chromosome 7 are now filtered out of the spreadsheet. To remove the filter, left click on the bar associated with chromosome 7 a second time.

<Right-click> on the region associated with chromosome 7 in the interactive filter to filter all the regions not associated with chromosome 7.

Notice that as the filters are applied that a yellow and black bar appears on the right side of the spreadsheet. This is a filter bar that conveniently displays that a filter is applied to the spreadsheet. The amount of black and yellow shown in the filter bar represents how much of the sheet has been filtered out of the spreadsheet (black) and how much of sheet has not been filtered out (yellow).

Now clear the filter so that all rows are returned to the spreadsheet. To clear the filter, right click on the black and yellow filter bar and select **Clear Filter**. The filter bar will disappear and all of the filtered rows will reappear in the spreadsheet.

Filtering on Region Length

In the interactive filter, click on the column drop-down menu and select column **14. Length (bps)**. If the column selected in the interactive filter is a continuous variable such as length, the *minimum value* and the *maximum value* of the specified column will be displayed to the right of the *Column* selector. These values change to reflect the status of the range tabs and can be directly typed in to set the precise *minimum* and *maximum* for the filter. The smallest region in the spreadsheet is 22 base pairs in length and the largest is 8 275 172 base pairs in length.

In the *Min* window, change the value from **11.0** to **1000** and select **Enter**. All the regions less than 1000 base pairs in length have now been filtered out. In the *Max* window, change the value from 8 275 172 to 1,000,000 and click enter. All the regions larger than 100,000 have been filtered out now, leaving only those regions between 1000 bps and 100,000 bps in our spreadsheet.

Applying Multiple Filters

It is possible to layer several filters on top of each other. With the length filter still applied, click on the column drop-down menu and select **17. Description**. There are five categories in the *Description* column, one for regions of *Amplification without LOH*, *Amplification with LOH*, *Copy-Neutral LOH*, *Deletion with LOH*, and *Deletion without LOH*. Left click on the bars for **Amplification without LOH** and **Deletion without LOH** to filter out these two categories. The spreadsheet has now been filtered down to 352 regions that are between 1000-100,000 base pairs that are described as Amplification with LOH; Copy-Neutral LOH or Deletion with LOH.

The interactive filter is a quick way to jump between columns of a spreadsheet to apply multiple filters in order to narrow the analysis results as well as to see the distribution of categories of a column. To monitor all filters that have been applied, go to **Filter > Filter Rows > Row Filter Manager**. The row filter manager lists all of the interactive filters currently applied to the spreadsheet.

As the spreadsheet has the genes associated with the regions detected included, it is possible to filter out all the regions that either contain genes or those that do not.

Filtering Regions Without Genes

Right click on the column header containing the overlap proportion **8. Percent overlap with gene** (alternatively *9. Percent overlap with region*), select **Find/Replace/Select**. A new dialogue box appears with the search/replace fields for the column. In the *Find What* field type in **0** and set the rest of the fields so that they match Figure 18 and then click **Select All**.

All of the rows that do not have a gene associated with them will now be highlighted. Right click on one of the highlighted row headers and select **Filter Exclude**. This filters out all of the rows without genes annotated to them leaving just the rows with genes annotated to the regions.

To filter out all of the rows that have genes associated with the region, instead of selecting **Filter Exclude**, choose **Filter Include**. Also, it is possible to search for a specific gene by using the **Find/Replace/Select** tool. Type in the name of the gene in the *Find What* field and click on **Select All** or **Next**.

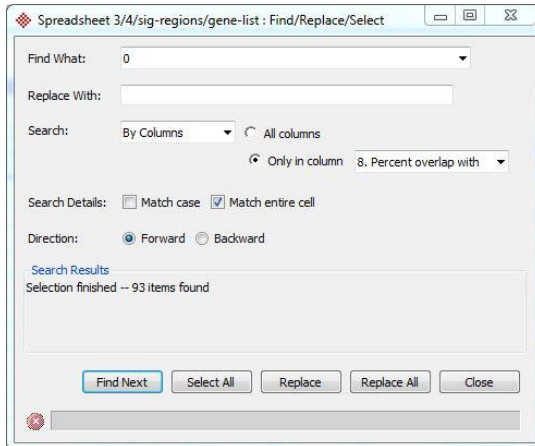


Figure 18: Configuring the Find/Replace/Select dialog box

Once a filtered list of interest is generated, a quick and convenient way to save the spreadsheet of just the regions that passed the filter is to clone the spreadsheet.

Please select the spreadsheet in the spreadsheet tree (on the left) and go to **Edit > Clone Spreadsheet**. A *Clone Spreadsheet* dialog box will appear. Leave the name of resulting copy as the default name. Under *Create as a child of spreadsheet*, select **4 (LOH merged with CN.txt)** and select **OK**. A new spreadsheet will be created as a child of **4 (LOH merged with CN.txt)** with only those regions associated with genes present.

End of Tutorial

In summary, from this tutorial, QA/QC was performed on the data, the LOH regions were detected using normal samples as baseline and the copy number regions of amplification and deletion were overlapped with the regions of LOH. To narrow down the regions reported, regions that were common in 5 out of the 10 were reported. Three more filters were applied and gene annotations were inserted to narrow the genes and segments that may play a role in controlling the phenotypes studied within this experiment. 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.

Appendix A: Setting-up the analysis of unpaired data

If unpaired data are used, the baseline file first needs to be created using normal samples. Please go to **Create baseline** option of the LOH workflow and import the genotype calls (Figure A1).

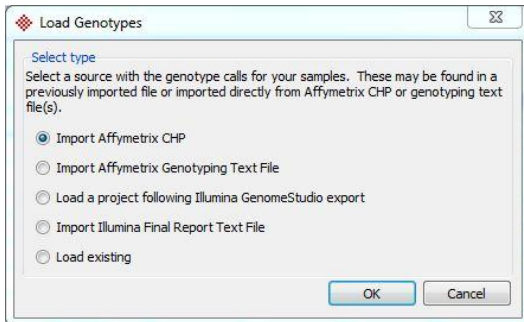


Figure A1. Importing data files with genotype calls.

Once the appropriate data files are imported, the baseline has to be created by selecting **Create LOH baseline** option of the LOH workflow (make sure that the spreadsheet with the data is selected) and specifying the baseline file name (.pgb format) (Figure A2).

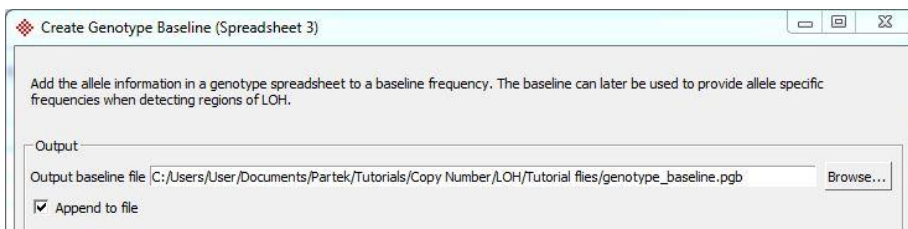


Figure A2. Creating the baseline genotype file for unpaired loss of heterozygosity analysis.

The baseline file has to be specified in the unpaired analysis dialog (Figure A3). *Default frequency* describes the expected frequency of heterozygous calls in a normal region. It is used in cases when an expected frequency of heterozygous calls cannot be found in the baseline. For the explanation of the remaining options please refer to the description of the paired analysis dialog.

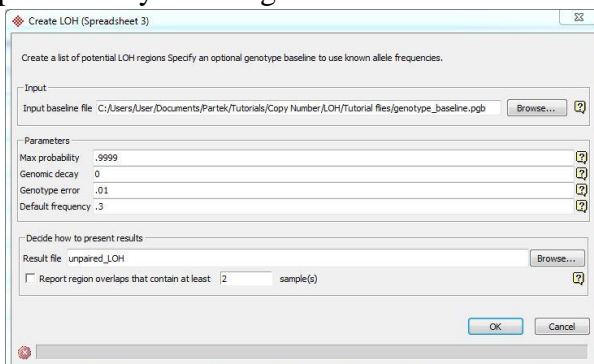


Figure A3. Configuring the Create loss of heterozygosity dialog for unpaired samples.

