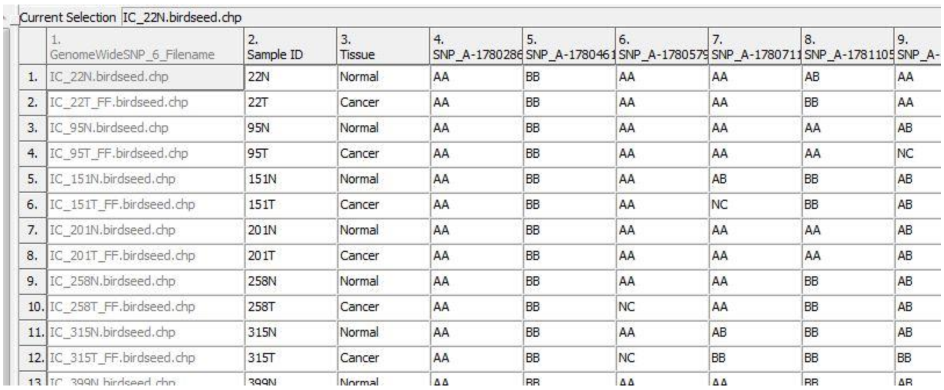
Using the Association Workflow in Partek® Genomics Suite™

This user guide will illustrate the use of the Association workflow in Partek® Genomics Suite™ (PGS) and discuss the basic functions available within the workflow. The workflow requires files with genotype calls, such as Affymetrix .CHP or genotyping text files, a project file exported from Illumina Genome Studio or Illumina final report text file. For the purpose of this example a set of 20 samples (10 normal and 10 cancer) was chosen. The samples were genotyped on Affymetrix SNP 6.0 arrays, and imported in CHP format (Figure 1).



Current Selection IC_22N.birdseed.chp

	1. GenomeWideSNP_6_Filename	2. Sample ID	3. Tissue	4. SNP_A-1780286	5. SNP_A-1780461	6. SNP_A-1780579	7. SNP_A-1780711	8. SNP_A-1781105	9. SNP_A-
1.	IC_22N.birdseed.chp	22N	Normal	AA	BB	AA	AA	AB	AA
2.	IC_22T_FF.birdseed.chp	22T	Cancer	AA	BB	AA	AA	BB	AA
3.	IC_95N.birdseed.chp	95N	Normal	AA	BB	AA	AA	AA	AB
4.	IC_95T_FF.birdseed.chp	95T	Cancer	AA	BB	AA	AA	AA	NC
5.	IC_151N.birdseed.chp	151N	Normal	AA	BB	AA	AB	BB	AB
6.	IC_151T_FF.birdseed.chp	151T	Cancer	AA	BB	AA	NC	BB	AB
7.	IC_201N.birdseed.chp	201N	Normal	AA	BB	AA	AA	AA	AB
8.	IC_201T_FF.birdseed.chp	201T	Cancer	AA	BB	AA	AA	AA	AB
9.	IC_258N.birdseed.chp	258N	Normal	AA	BB	AA	AA	BB	AB
10.	IC_258T_FF.birdseed.chp	258T	Cancer	AA	BB	NC	AA	BB	AB
11.	IC_315N.birdseed.chp	315N	Normal	AA	BB	AA	AB	BB	AB
12.	IC_315T_FF.birdseed.chp	315T	Cancer	AA	BB	NC	BB	BB	BB
13.	IC_399N.birdseed.chp	399N	Normal	AA	BB	AA	AA	BB	AB

Figure 1: Genotype calls loaded into the Association workflow.

Quality Assessment/Quality Control (QA/QC)

Following the sample import please follow the workflow to perform the QA/QC steps. Based on the results of the QA/QC, you might decide to omit some samples from the downstream steps.

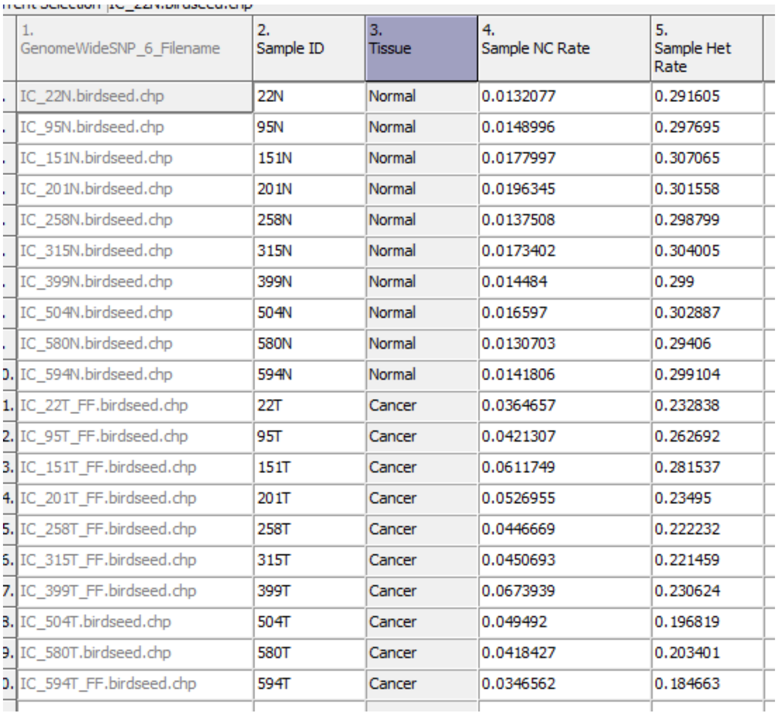
Sample QC

The Sample QC option will invoke the *sample_QC* spreadsheet (Figure 2), which shows one sample per row. The rate of missing genotype calls for each sample is given in the *Sample NC Rate* column. The rate is determined by dividing the number of no-calls (NC) by the total number of genotypes in the sample and an unusually high number in this column indicates that the overall genotyping quality in the sample is poor. As a rough guide, one can tolerate a NC rate of up to 5%.

The *Sample Heterozygosity Rate* can also be used as a quality indicator. As a rule of a thumb, you might want to reconsider the samples with the heterozygosity rate which falls out of the interval mean ± 3 × standard deviations of all the samples. To calculate the mean and standard deviation for the heterozygosity rate of the samples please use **Stat > Descriptive > Column Statistics…** and select mean and standard deviation from the list of available Candidate Measures (not shown).

In this example the average sample heterozygosity was 0.26335, with the standard deviation of 0.04237. Therefore, one might decide to remove the samples with the heterozygosity rate greater than 0.39035 or less than 0.13635.

By applying both criteria, i.e. sample no-call rate and sample heterozygosity rate to the current example, three samples were removed from the study, thus resulting in 10 normal and 7 tumor samples.



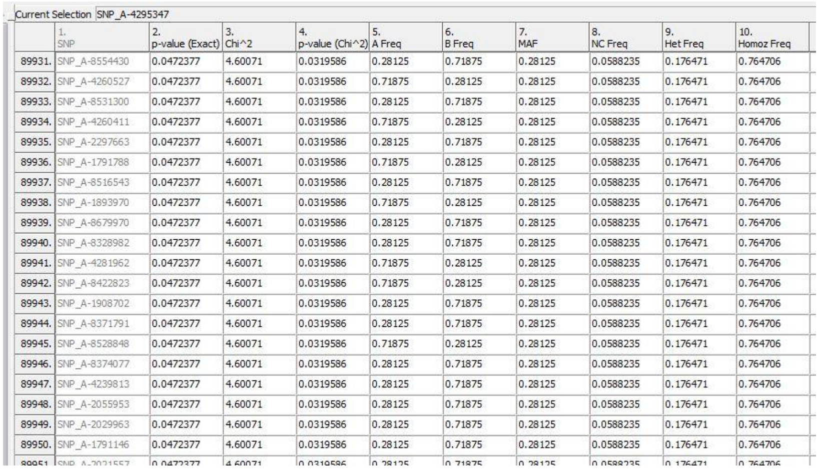
	1. GenomeWideSNP_6_Filename	2. Sample ID	3. Tissue	4. Sample NC Rate	5. Sample Het Rate
.	IC_22N.birdseed.chp	22N	Normal	0.0132077	0.291605
.	IC_95N.birdseed.chp	95N	Normal	0.0148996	0.297695
.	IC_151N.birdseed.chp	151N	Normal	0.0177997	0.307065
.	IC_201N.birdseed.chp	201N	Normal	0.0196345	0.301558
.	IC_258N.birdseed.chp	258N	Normal	0.0137508	0.298799
.	IC_315N.birdseed.chp	315N	Normal	0.0173402	0.304005
.	IC_399N.birdseed.chp	399N	Normal	0.014484	0.299
.	IC_504N.birdseed.chp	504N	Normal	0.016597	0.302887
.	IC_580N.birdseed.chp	580N	Normal	0.0130703	0.29406
0.	IC_594N.birdseed.chp	594N	Normal	0.0141806	0.299104
1.	IC_22T_FF.birdseed.chp	22T	Cancer	0.0364657	0.232838
2.	IC_95T_FF.birdseed.chp	95T	Cancer	0.0421307	0.262692
3.	IC_151T_FF.birdseed.chp	151T	Cancer	0.0611749	0.281537
4.	IC_201T_FF.birdseed.chp	201T	Cancer	0.0526955	0.23495
5.	IC_258T_FF.birdseed.chp	258T	Cancer	0.0446669	0.222232
6.	IC_315T_FF.birdseed.chp	315T	Cancer	0.0450693	0.221459
7.	IC_399T_FF.birdseed.chp	399T	Cancer	0.0673939	0.230624
8.	IC_504T.birdseed.chp	504T	Cancer	0.049492	0.196819
9.	IC_580T.birdseed.chp	580T	Cancer	0.0418427	0.203401
0.	IC_594T_FF.birdseed.chp	594T	Cancer	0.0346562	0.184663

Figure 2: Sample QC spreadsheet showing no-call rate and heterozygosity rate of each sample.

Hardy-Weinberg Equilibrium

After removing the samples which did not pass the sample QC criteria you can proceed to the next QC step, Hardy-Weinberg equilibrium, which is essentially QA/QC on SNP level. The resulting spreadsheet (*HWE*) features one SNP per row (Figure 3).

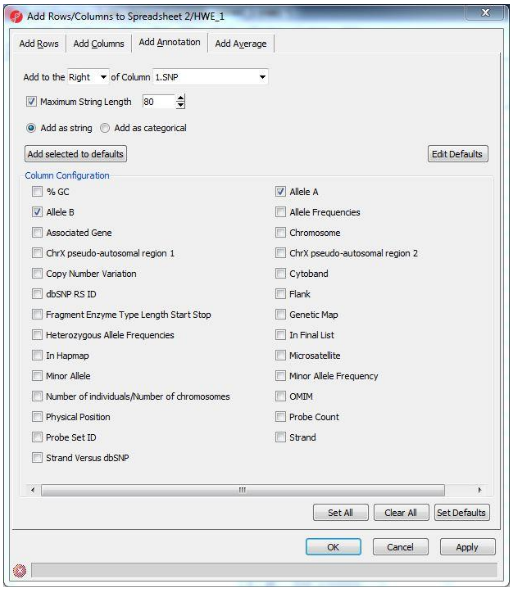
The difference between the observed and expected frequencies of each allele at each locus (or SNP) are tested by Fisher's exact test (*p-value(Exact)*) and χ^2 test (*Chi^2* and *p-value(Chi^2)*). Frequencies of both alleles are provided in the columns *A Freq* and *B Freq*, while the minor allele frequency (*MAF*) corresponds to the one with lower frequency. The remaining three columns contain the no-call frequency (*NC Freq*), heterozygous frequency (*Het Freq*), and homozygous frequency (*Homoz Freq*) at the given locus. Depending on the annotation provided by the array vendor, it may be possible to annotate the SNPs with exact base calls at each locus. Please right-click on a column header and select **Insert Annotation**. In the *Add Rows/Columns to Spreadsheet* dialog, please tick mark the **Allele A** and **Allele B** boxes (Figure 4). Two new columns will be added to the *HWE* spreadsheet and will contain the genotype of each allele (not shown).



Current Selection SNP_A-4295347

	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
89931.	SNP_A-8554430	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89932.	SNP_A-4260527	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89933.	SNP_A-8531300	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89934.	SNP_A-4260411	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89935.	SNP_A-2297663	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89936.	SNP_A-1791788	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89937.	SNP_A-8516543	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89938.	SNP_A-1893970	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89939.	SNP_A-8679970	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89940.	SNP_A-8328982	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89941.	SNP_A-4281962	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89942.	SNP_A-8422823	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89943.	SNP_A-1908702	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89944.	SNP_A-8371791	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89945.	SNP_A-8528848	0.0472377	4.60071	0.0319586	0.71875	0.28125	0.28125	0.0588235	0.176471	0.764706
89946.	SNP_A-8374077	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89947.	SNP_A-4239813	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89948.	SNP_A-2055953	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89949.	SNP_A-2029963	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89950.	SNP_A-1791146	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706
89951.	SNP_A-2021557	0.0472377	4.60071	0.0319586	0.28125	0.71875	0.28125	0.0588235	0.176471	0.764706

Figure 3: The Hardy-Weinberg equilibrium spreadsheet. The spreadsheet features one SNP per row and is intended to be used for QA/QC on SNP level.



Add to the Right ▾ of Column 1.SNP ▾

☑ Maximum String Length 80

● Add as string ○ Add as categorical

Column Configuration

□ % GC	☑ Allele A
☑ Allele B	□ Allele Frequencies
□ Associated Gene	□ Chromosome
□ ChrX pseudo-autosomal region 1	□ ChrX pseudo-autosomal region 2
□ Copy Number Variation	□ Cytoband
□ dbSNP RS ID	□ Flank
□ Fragment Enzyme Type Length Start Stop	□ Genetic Map
□ Heterozygous Allele Frequencies	□ In Final List
□ In Hapmap	□ Microsatellite
□ Minor Allele	□ Minor Allele Frequency
□ Number of individuals/Number of chromosomes	□ OMIM
□ Physical Position	□ Probe Count
□ Probe Set ID	□ Strand
□ Strand Versus dbSNP	

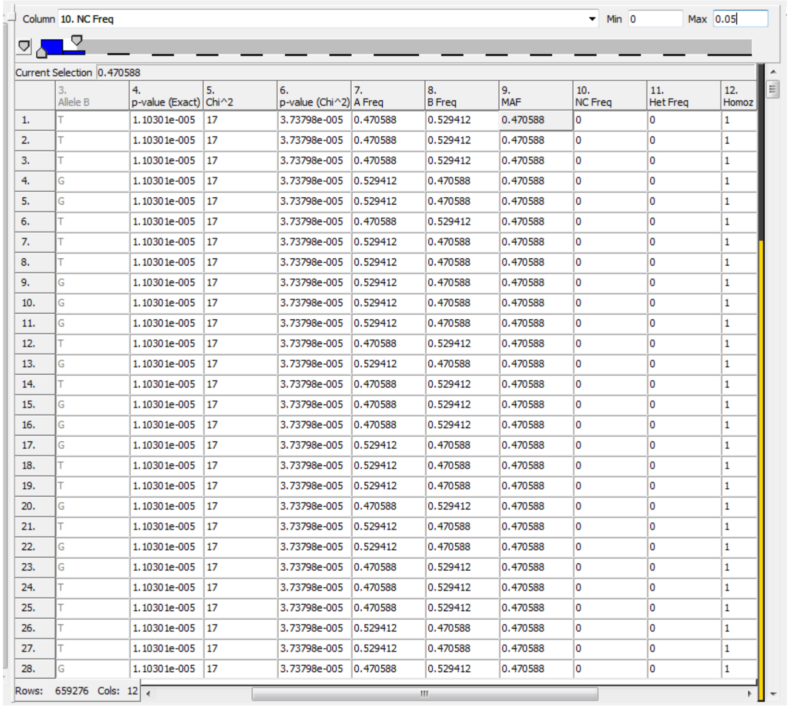
Figure 4: Adding additional annotation to a spreadsheet

At this stage, the following two filters may be considered:

1. A SNP no-call rate should be less than 5%.
2. Minor allele frequency of a SNP should be greater than 5%.

After that, you might want to additionally remove the SNPs that are not in Hardy-Weinberg equilibrium. For that purpose, a multiple testing correction should be applied to the exact p-value: the cut-off p-value after the correction equals 0.05 / (number of SNPs left after 1st and 2nd filter) (in the other words, Bonferroni's correction).

The filtering can be performed by the interactive filter (the icon 📊), to first filter in the SNPs with the NC frequency less than 0.05 (Figure 5), and then to filter in the SNPs with MAF greater than 0.05 (Figure 6). Please note that the effects of the interactive filter are additive.



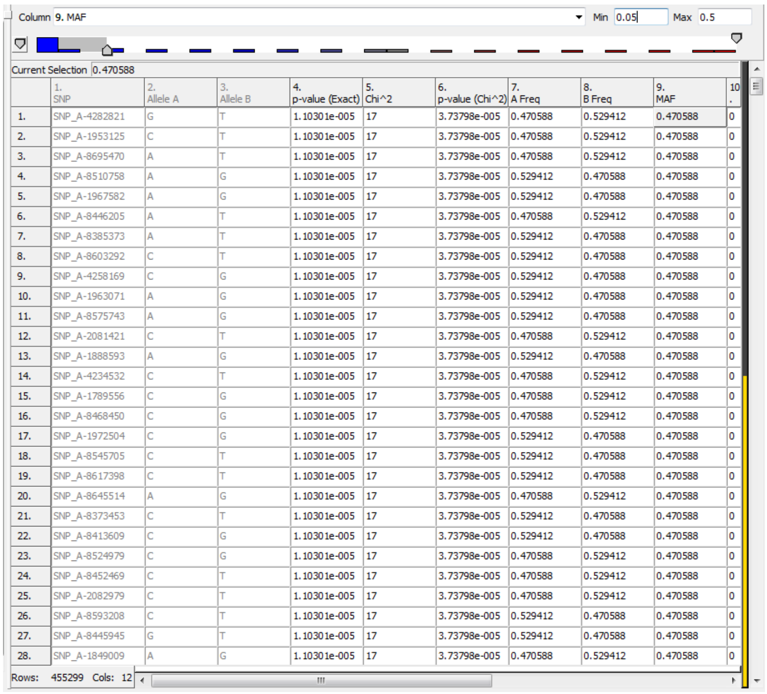
Column 10. NC Freq Min 0 Max 0.05

Current Selection 0.470588

	3. Allele B	4. p-value (Exact)	5. Chi^2	6. p-value (Chi^2)	7. A Freq	8. B Freq	9. MAF	10. NC Freq	11. Het Freq	12. Homoz
1.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
2.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
3.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
4.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
5.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
6.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
7.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
8.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
9.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
10.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
11.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
12.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
13.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
14.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
15.	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
16.	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
17.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
18.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
19.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
20.	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
21.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
22.	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
23.	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
24.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
25.	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1
26.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
27.	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0	0	1
28.	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0	0	1

Rows: 659276 Cols: 12

Figure 5: Using the interactive filter to filter in the SNPs with no call frequency less than 5% (max 0.05). After the filtering 659 276 SNPs remain in the spreadsheet.



Column 9. MAF Min 0.05 Max 0.5

Current Selection 0.470588

	1. SNP	2. Allele A	3. Allele B	4. p-value (Exact)	5. Chi^2	6. p-value (Chi^2)	7. A Freq	8. B Freq	9. MAF	10 .
1.	SNP_A-4282821	G	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
2.	SNP_A-1953125	C	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
3.	SNP_A-8695470	A	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
4.	SNP_A-8510758	A	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
5.	SNP_A-1967582	A	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
6.	SNP_A-8446205	A	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
7.	SNP_A-8385373	A	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
8.	SNP_A-8603292	C	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
9.	SNP_A-4258169	C	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
10.	SNP_A-1963071	A	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
11.	SNP_A-8575743	A	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
12.	SNP_A-2081421	C	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
13.	SNP_A-1888593	A	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
14.	SNP_A-4234532	C	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
15.	SNP_A-1789556	C	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
16.	SNP_A-8468450	C	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
17.	SNP_A-1972504	C	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
18.	SNP_A-8545705	C	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
19.	SNP_A-8617398	C	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
20.	SNP_A-8645514	A	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
21.	SNP_A-8373453	C	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
22.	SNP_A-8413609	C	G	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
23.	SNP_A-8524979	C	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
24.	SNP_A-8452469	C	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
25.	SNP_A-2082979	C	T	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0
26.	SNP_A-8593208	C	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
27.	SNP_A-8445945	G	T	1.10301e-005	17	3.73798e-005	0.529412	0.470588	0.470588	0
28.	SNP_A-1849009	A	G	1.10301e-005	17	3.73798e-005	0.470588	0.529412	0.470588	0

Rows: 455299 Cols: 12

Figure 6: Using the interactive filter to filter in the SNPs with minor allele frequency greater than 5% (min 0.05). After the filtering 455 299 SNPs remain in the spreadsheet.

Two applied filters reduced the number of available SNPs to 455 299. Using the Bonferroni's correction, the threshold p-value is: 0.05 / 455 299 = 1.098e-7. As the minimum p-value of the remaining SNPs in this example is 1.103e-5 (as can be seen in Figure 6), no further filtering needs to be performed.

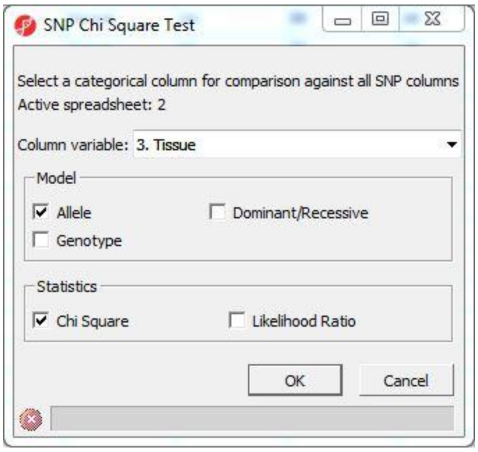
However, in order to proceed to the next step of the workflow, the changes (i.e. filtering in of SNPs which met the chosen QA/QC criteria) have to be applied to the parent spreadsheet; the same SNPs need to be filtered in. To do that please select the parent spreadsheet (in this example this is the one with 17 samples on rows) and then choose **Filter > Filter Columns > Filter Columns Based on a List…** In the *Filter Columns on Spreadsheet* dialog, please set the *Filter based on spreadsheet* to the spreadsheet containing the final 455 299 SNPs and set the *Key column* to SNP.

Association Analysis

The association analysis in this example will start with the *Run chi-square test* option of the workflow. If you would like to know more about the *Generate sample IBS* functionality, please refer to the Trio user guide (available by **Help > On-line Tutorials**).

Chi-square Test

The dialog of the chi-square test is shown in Figure 7. The column variable enables you to set the phenotype (categorical variable) which will be tested for association with the SNPs. In the other words, the allele/genotype frequencies as specified by the model (please see the discussion below) will be compared between the categories. By setting the *Column variable* to *Tissue*, in this example, one will test the association of the SNPs with cancer.



Select a categorical column for comparison against all SNP columns
Active spreadsheet: 2

Column variable: 3. Tissue

☑ Allele ☐ Dominant/Recessive
☐ Genotype

☑ Chi Square ☐ Likelihood Ratio

Figure 7: Setting the options of the SNP Chi Square Test dialog, to test the association of the SNPs with the variable selected in the Column variable box.

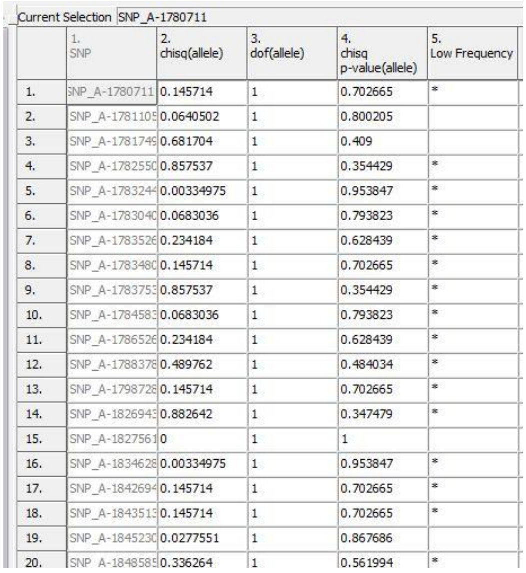
The model section allows for specification of the statistical model.

1. Allele: frequencies of alleles (A vs. B) are compared across the categories of the selected variable (i.e. phenotype).
2. Genotype: frequencies of three possible genotypes (AA, AB and BB) are compared across the categories of the selected variable.
3. Dominant/Recessive: two combinations of genotypes are compared across the categories of the selected variable.

   Dominant: AA + AB versus BB (A is the causal variant)
   Recessive: AA versus AB + BB (A is the causal variant)

Significant p-value indicates that the allele/genotype frequencies are different between the categories of the selected variable, i.e. that an association exists between the genotype and the phenotype.

In the present example χ^2 statistic was used to assess the difference in allele frequencies (allele model) between the normal and cancer samples. The resulting spreadsheet (*ChiSquare*) shows one SNP per row (Figure 8). PGS provides the value of χ^2 statistic (*chisq*), degrees of freedom (*dof*), and the associated p-value (*chisq p-value*) for each SNP. In addition, please note the *Low Frequency* column, which flags the SNPs with possibly unreliable p-values. The flags are based on contingency tables, which are used to calculate the χ^2 statistic. A 2 × 2 contingency table is flagged if any cell in the table has an expected frequency of less than 5. Any other dimension table is flagged if the mean expected cell frequency is less than 5 or any single expected frequency is less than one.

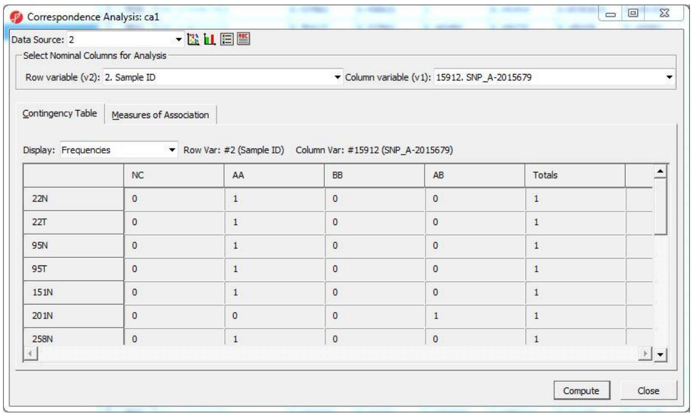


Current Selection SNP_A-1780711

	1. SNP	2. chisq(allele)	3. dof(allele)	4. chisq p-value(allele)	5. Low Frequency
1.	SNP_A-1780711	0.145714	1	0.702665	*
2.	SNP_A-1781105	0.0640502	1	0.800205	
3.	SNP_A-1781749	0.681704	1	0.409	
4.	SNP_A-1782550	0.857537	1	0.354429	*
5.	SNP_A-1783244	0.00334975	1	0.953847	*
6.	SNP_A-1783040	0.0683036	1	0.793823	*
7.	SNP_A-1783526	0.234184	1	0.628439	*
8.	SNP_A-1783480	0.145714	1	0.702665	*
9.	SNP_A-1783753	0.857537	1	0.354429	*
10.	SNP_A-1784583	0.0683036	1	0.793823	*
11.	SNP_A-1786526	0.234184	1	0.628439	*
12.	SNP_A-1788378	0.489762	1	0.484034	*
13.	SNP_A-1798728	0.145714	1	0.702665	*
14.	SNP_A-1826943	0.882642	1	0.347479	*
15.	SNP_A-1827561	0	1	1	
16.	SNP_A-1834628	0.00334975	1	0.953847	*
17.	SNP_A-1842694	0.145714	1	0.702665	*
18.	SNP_A-1843513	0.145714	1	0.702665	*
19.	SNP_A-1845230	0.0277551	1	0.867686	
20.	SNP_A-1848585	0.336264	1	0.561994	*

Figure 8: Chi square table provides the results of test of allele/genotype frequencies between the categories of the chosen grouping variable (i.e. between the phenotypes) for each SNP.

A convenient way to take a look at genotypes of all the samples is to right-click on a row header and select **Correspondence Analysis (Orig Data)**. The resulting window (Figure 9) will contain a table with samples on rows and genotypes on columns. In addition, the correspondence analysis is available from all the spreadsheets within this workflow that contain SNPs on rows.



Data Source: 2

Select Nominal Columns for Analysis

Row variable (v2): 2. Sample ID

Column variable (v1): 15912. SNP_A-2015679

Display: Frequencies

Row Var: #2 (Sample ID) Column Var: #15912 (SNP_A-2015679)

	NC	AA	BB	AB	Totals
22N	0	1	0	0	1
22T	0	1	0	0	1
95N	0	1	0	0	1
95T	0	1	0	0	1
151N	0	1	0	0	1
201N	0	0	0	1	1
258N	0	1	0	0	1

Figure 9: Correspondence analysis window showing genotypes at the chosen locus (found in the Column variable box) across all the samples.

Moreover, it is easy to visualize the genotypes. Please right-click on the row header of a SNP and select **Dot Plot (Orig Data)**. An example of a dot plot is seen in *Figure 10*. PGS will automatically use the name of the associated gene (as provided by the array vendor) as the plot title (in this example CA10).

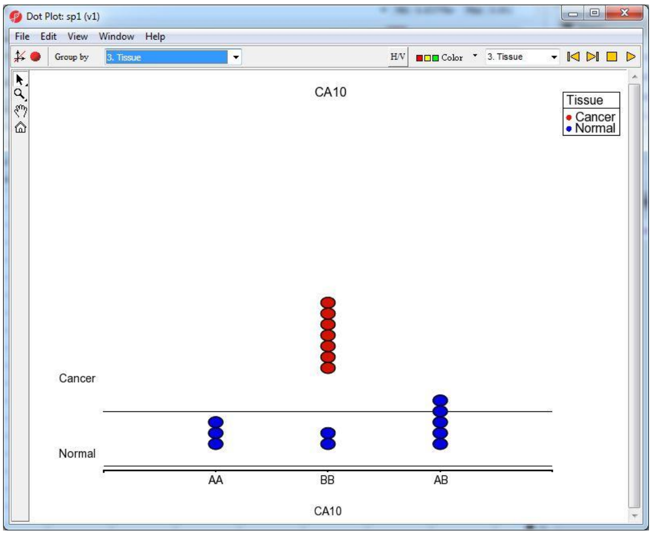
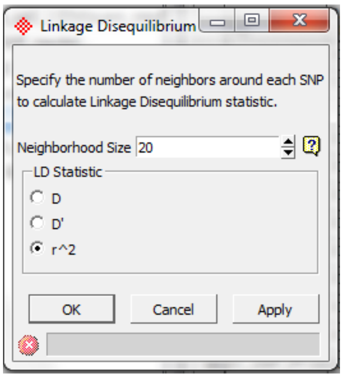


Figure 10: Dot plot of the original data showing genotypes across samples. PGS will automatically use the name of the associated gene (if provided by the array vendor) as the plot title. The samples are grouped by the tissue column and colored by the same column to highlight the difference between the cancer samples and the normal samples.

Linkage Disequilibrium

For the linkage disequilibrium analysis one first has to set the neighborhood size of each SNP (Figure 11). In the other words, each SNP will be compared to that number of neighboring SNPs.

Furthermore, PGS lets you to choose your linkage disequilibrium statistic: D, D’ or r^2 (for discussion on the implementation please consult the User’s Manual). The D statistic is the deviation of the expected frequency of alleles from the observed frequency. If the D statistic becomes the value of 0, that indicates that two SNPs are in equilibrium. On the other hand, D <> 0 indicates the existence of linkage disequilibrium. D’ is the normalized value of D, calculated by dividing D by the theoretical maximum for the observed allele frequencies, and is also used to assess the linkage disequilibrium between two alleles. If two SNPs were inherited together (meaning that the disequilibrium exists), the expected value of D’ is 1. r^2 is an associative measurement. It will determine how well one can predict the genotype of one SNP, given the genotype of another SNP. The $r^2 = 1$ means that two SNPs are tightly associated (i.e. in linkage disequilibrium). Knowing the genotype of the first, you will be able to tell the genotype of the second.



Linkage Disequilibrium

Specify the number of neighbors around each SNP to calculate Linkage Disequilibrium statistic.

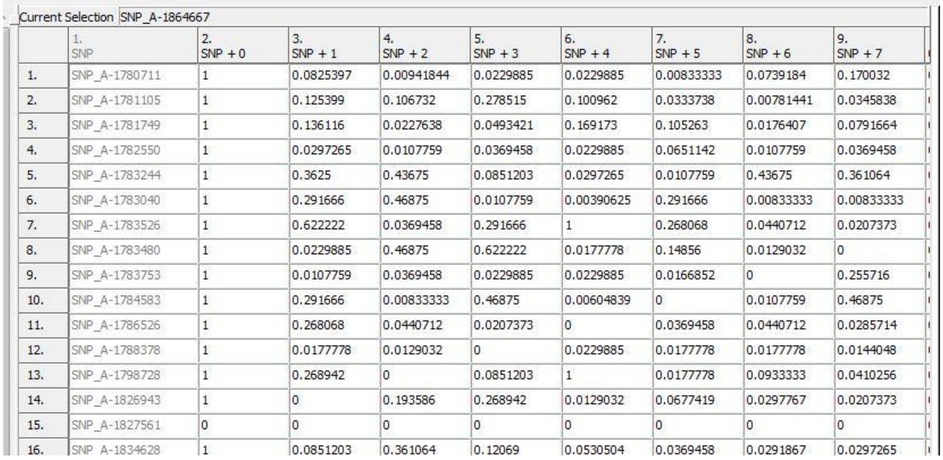
Neighborhood Size 20

LD Statistic

○ D
○ D'
● r^2

Figure 11: Setting the options to run the linkage disequilibrium test.

For the purpose of this guide, linkage disequilibrium was performed on the data set, using the neighborhood size of 20 (PGS default) and r^2 as the disequilibrium statistic. The resulting spreadsheet (*LD_r^2*) is shown in Figure 12.



Current Selection SNP_A-1864667

	1. SNP	2. SNP + 0	3. SNP + 1	4. SNP + 2	5. SNP + 3	6. SNP + 4	7. SNP + 5	8. SNP + 6	9. SNP + 7
1.	SNP_A-1780711	1	0.0825397	0.00941844	0.0229885	0.0229885	0.00833333	0.0739184	0.170032
2.	SNP_A-1781105	1	0.125399	0.106732	0.278515	0.100962	0.0333738	0.00781441	0.0345838
3.	SNP_A-1781749	1	0.136116	0.0227638	0.0493421	0.169173	0.105263	0.0176407	0.0791664
4.	SNP_A-1782550	1	0.0297265	0.0107759	0.0369458	0.0229885	0.0651142	0.0107759	0.0369458
5.	SNP_A-1783244	1	0.3625	0.43675	0.0851203	0.0297265	0.0107759	0.43675	0.361064
6.	SNP_A-1783040	1	0.291666	0.46875	0.0107759	0.00390625	0.291666	0.00833333	0.00833333
7.	SNP_A-1783526	1	0.622222	0.0369458	0.291666	1	0.268068	0.0440712	0.0207373
8.	SNP_A-1783480	1	0.0229885	0.46875	0.622222	0.0177778	0.14856	0.0129032	0
9.	SNP_A-1783753	1	0.0107759	0.0369458	0.0229885	0.0229885	0.0166852	0	0.255716
10.	SNP_A-1784583	1	0.291666	0.00833333	0.46875	0.00604839	0	0.0107759	0.46875
11.	SNP_A-1786526	1	0.268068	0.0440712	0.0207373	0	0.0369458	0.0440712	0.0285714
12.	SNP_A-1788378	1	0.0177778	0.0129032	0	0.0229885	0.0177778	0.0177778	0.0144048
13.	SNP_A-1798728	1	0.268942	0	0.0851203	1	0.0177778	0.0933333	0.0410256
14.	SNP_A-1826943	1	0	0.193586	0.268942	0.0129032	0.0677419	0.0297767	0.0207373
15.	SNP_A-1827561	0	0	0	0	0	0	0	0
16.	SNP_A-1834628	1	0.0851203	0.361064	0.12069	0.0530504	0.0369458	0.0291867	0.0297265

Figure 12: The linkage disequilibrium spreadsheet. Each row represents one SNP while the columns show neighboring SNPs (relative to the SNP in the row header). The cells contain the values of the chosen linkage disequilibrium statistic (r^2).

The *LD_r^2* spreadsheet features one SNP per row, while the cells contain the values of the chosen linkage disequilibrium statistic (r^2 in this example). The columns represent neighboring SNPs, relative to the SNP in a given row: the SNP in the row header is labeled SNP + 0, the next one SNP + 1, followed by SNP + 2 and so on. The total number of columns depends on the size of the neighborhood, as selected in the linkage analysis set-up dialog.

If the genotype of a given locus (SNP) is the same across all the samples, then the the r^2 values for that SNP will be 0 (an example can be seen in Figure 12).

Results of the linkage disequilibrium analysis can be visualized by the LD plot. To invoke the plot, please select a SNP, right-click on the row header and select **LD Plot** from the contextual menu. An example of the plot is given in Figure 13. The SNPs in the neighborhood are on rows and columns, with the chosen SNP in the middle. The r^2 values are visualized in a form of a heat map, and the map than can be used to quickly identify blocks of disequilibrium.

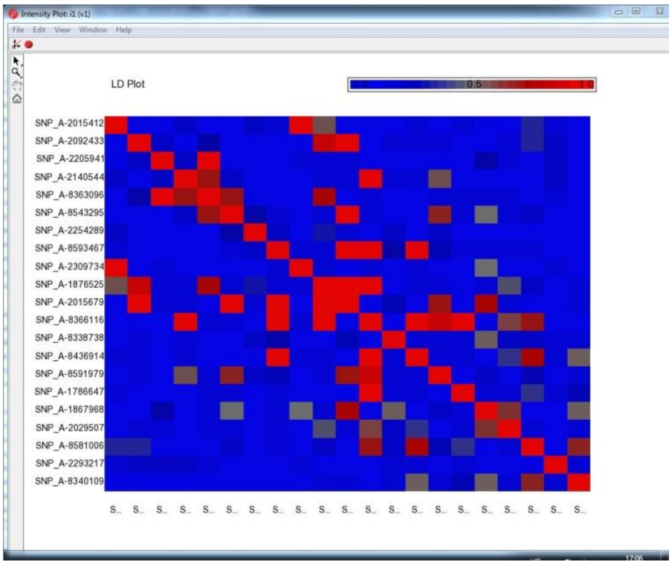


Figure 13: Linkage disequilibrium plot. The chosen SNP is in the middle (in this example SNP_A-2015679), while the map shows r2 values.

End of User Guide

This is the end of the user guide. If you need additional assistance, please call our technical support staff at +1-314-878-2329 or email support@partek.com.

Note

Last revision: January 2012