

Using the Trio Workflow in Partek® Genomics Suite™ v6.6

This user guide will illustrate the use of the Trio/Duo workflow in Partek® Genomics Suite™ (PGS) and discuss the basic functions available within the workflow. Please note that this user guide focuses on microarray data, while the NGS duo/trio analysis is discussed in the DNA-seq documentation (available at **Help > On-line Tutorials**). 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 10 samples genotyped on Affymetrix SNP 6.0 arrays were chosen (Figure 1).

Current Selection IC_22N.birdseed.chp								
1. GenomeWideSNP_6_FileName	2. SNP_A-1780286	3. SNP_A-1780461	4. SNP_A-1780579	5. SNP_A-1780711	6. SNP_A-1781105	7. SNP_A-1781550	8. SNP_A-1781774	9. SNP_A-1781774
1. IC_22N.birdseed.chp	AA	BB	AA	AA	AB	AA	AA	AA
2. IC_95N.birdseed.chp	AA	BB	AA	AA	AA	AB	AA	BB
3. IC_151N.birdseed.chp	AA	BB	AA	AB	BB	AB	AA	AB
4. IC_201N.birdseed.chp	AA	BB	AA	AA	AA	AB	AA	BB
5. IC_258N.birdseed.chp	AA	BB	AA	AA	BB	AB	AA	AB
6. IC_315N.birdseed.chp	AA	BB	AA	AB	BB	AB	AA	AB
7. IC_399N.birdseed.chp	AA	BB	AA	AA	BB	AB	AA	AB
8. IC_504N.birdseed.chp	AA	BB	AA	AA	AA	AA	AA	AB
9. IC_580N.birdseed.chp	AA	BB	AA	AA	AB	AB	AA	AB
10. IC_594N.birdseed.chp	AA	BB	AA	AA	BB	BB	AA	AA

Figure 1: Genotype calls imported into the Trio workflow.

Sample Quality Control

Following the sample import please follow the workflow to perform the quality control (QC) steps. Based on the results, you might decide to omit some samples from the downstream steps.

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 $\text{mean} \pm 3 \times \text{standard deviation}$ of all the samples. To calculate the mean and standard deviation for the heterozygosity rate of the samples, please use **Start > Descriptive > Column Statistics...** and select mean and standard deviation from the list of available Candidate Measures (not shown).

Current Selection IC_22N.birdseed.chp		
1. GenomeWideSNP_6_Filename	2. Sample NC Rate	3. Sample Het Rate
1. IC_22N.birdseed.chp	0.0132077	0.291605
2. IC_95N.birdseed.chp	0.0148996	0.297695
3. IC_151N.birdseed.chp	0.0177997	0.307065
4. IC_201N.birdseed.chp	0.0196345	0.301558
5. IC_258N.birdseed.chp	0.0137508	0.298799
6. IC_315N.birdseed.chp	0.0173402	0.304005
7. IC_399N.birdseed.chp	0.014484	0.299
8. IC_504N.birdseed.chp	0.016597	0.302887
9. IC_580N.birdseed.chp	0.0130703	0.29406
10. IC_594N.birdseed.chp	0.0141806	0.299104

Figure 2: Sample QC spreadsheet showing no-call (NC) rate and heterozygous rate for each sample

Data Analysis

The Data Analysis section of the workflow enables you to perform identity by state (IBS) analysis, duo analysis, and trio analysis. All three methods provide an insight in relationships in population data, and can be used, for instance to test paternity, detect relatives, or detect allele sharing.

Generate Sample IBS

The *Generate Sample IBS* option will create two spreadsheets: *ibs_frequency* and *ibs_matrix*. We shall first focus on the former one.

The *ibs_frequency* spreadsheet (Figure 3) shows one pairwise comparison per row. The samples being compared are listed in the first column, and then (for filtering purposes) given separately in the next two columns.

Current Selection 151N vs 399N												
1. Samples	2. Sample 1	3. Sample 2	4. IBS 0	5. IBS 1	6. IBS 2	7. IBS Mean	8. IBS Variance	9. IBS2*	10. Percent IBS2*	11. Percent Informative SNPs	12. p-value(Binomial Population)	13. p-value(Binomial Relation)
1. 22N vs 95N	22N	95N	45940	288428	552202	1.57103	0.348589	114742	0.714094	0.18124	1	0
2. 22N vs 151N	22N	151N	44938	288584	549949	1.57162	0.346601	117577	0.723484	0.183951	1	0
3. 22N vs 201N	22N	201N	45527	289066	547772	1.5692	0.348404	114613	0.715705	0.18149	1	0
4. 22N vs 258N	22N	258N	46679	287841	552944	1.57046	0.350231	115729	0.712582	0.183002	1	0
5. 22N vs 315N	22N	315N	45221	286422	552450	1.57373	0.346863	117517	0.722124	0.184073	1	0
6. 22N vs 399N	22N	399N	45124	286237	555500	1.57549	0.346063	116477	0.720769	0.182217	1	0
7. 22N vs 504N	22N	504N	44966	287103	552816	1.57392	0.346168	116925	0.722245	0.182951	1	0
8. 22N vs 580N	22N	580N	47066	287192	554058	1.57073	0.350964	114339	0.708398	0.181698	1	0
9. 22N vs 594N	22N	594N	44926	289784	552349	1.57203	0.346104	114849	0.718817	0.180118	1	0
10. 95N vs 151N	95N	151N	44470	284939	552885	1.57624	0.344993	121799	0.732542	0.188451	1	0
11. 95N vs 201N	95N	201N	44224	288056	548388	1.57248	0.34518	117223	0.726077	0.183323	1	0
12. 95N vs 258N	95N	258N	44616	286132	555231	1.57633	0.34489	118890	0.727129	0.184548	1	0

Figure 3: The *ibs_frequency* spreadsheet features one pairwise comparison of samples per row.

Columns #4 – #6 give the number of loci within particular IBS class. For a pair of individuals with genotype information, IBS can be observed at a given locus with three possible outcomes, depending on the number of alleles that the individuals have in common. If the alleles are labeled as A and B, the classification of IBS is as follows.

- IBS2: two alleles in common (AA/AA, AB/AB, or BB/BB)

- IBS1: one allele in common (AA/AB, AB/BB)
- IBS0: no common alleles (AA/BB)

The mean IBS state and its variance are found in columns #7 and #8, respectively. The calculation of the mean is:

Total = "IBS 0" + "IBS 1" + "IBS 2"

IBS Mean = $(0 * \text{"IBS 0"} + 1 * \text{"IBS 1"} + 2 * \text{"IBS 2"}) / \text{Total}$

The variance is to get the sum of square of errors for each IBS and divided by total:

$\text{sqerror_0} = (\text{"IBS 0"} - \text{"IBS Mean"})^2$


$\text{sqerror_1} = (\text{"IBS 1"} - \text{"IBS Mean"})^2$

$\text{sqerror_2} = (\text{"IBS 2"} - \text{"IBS Mean"})^2$

IBS Variance = $(\text{"IBS 0"} * \text{sqerror_0} + \text{"IBS 1"} * \text{sqerror_1} + \text{"IBS 2"} * \text{sqerror_2}) / \text{total}$

For the analysis of genetic relatedness an IBS sub-class is of special interest, namely the concordant heterozygotes (i.e. AB/AB), labeled as *IBD2** (column #9). The *percent IBD2** (column #10) is the ratio of concordant heterozygotes to the sum of concordant heterozygotes plus discordant homozygotes (that is $\text{IBS2}^*/(\text{IBS 2}^* + \text{IBS 0})$).

Column #11 (*percent informative SNPs*) gives the fraction of informative SNPs between the two individuals and is calculated using the following equation:
 $(\text{IBS0} + \text{IBS2}^*) / \text{Total}$.

To ease the within-population comparisons you can use PGS to display a plot having percent IBS2* on the x-axis and percent informative SNPs on the y-axis (the plot is commonly referred to as an IBS2* plot). To invoke the plot, please select both *percent IBS2** and *percent informative SNPs* columns, and select the **Scatter Plot** icon (). An example of the plot is shown in Figure 4. It can be used for fast identification of identical samples, siblings, and parent & child pairs. For an excellent discussion on the topic, please refer to Stevens EL et al. (2011).

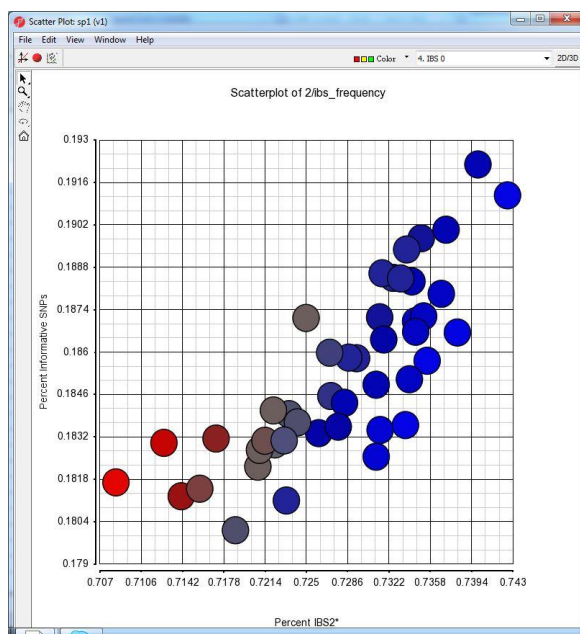


Figure 6: SNP Duo setup dialog. Paired samples need to be specified.

PGS will then generate a SNP duo spreadsheet, with each duo shown in one row (an example of a single duo is shown in Figure 7). The samples are specified in three left-most columns, while each of the remaining columns corresponds to one SNP. The codes 0, 1, 2, and ? are used as explained at the beginning of this section. The results of SNP duo analysis can be visualized, and that will be discussed in the SNP Trio section.

1. Samples	2. Sample 1	3. Sample 2	4. SNP_A-1780286	5. SNP_A-1780461	6. SNP_A-1780579	7. SNP_A-1780711	8. SNP_A-1781105	9. SNP_A-1781550	10. SNP_A-1781550
1. 22N vs 95N	22N	95N	2	2	2	2	1	1	2

Figure 7: Result of SNP duo analysis. Each column represents one SNP while the cell entries show the number of alleles shared between the samples.

SNP Trio

SNP Trio function enables the analysis of blocks of uniparental inheritance, based on SNP calls of child-mother-father trios. The following scenarios are possible:

- NI (non informative) (0): e.g. child AA, mother AA, father AA, or child AA, mother AA, father AB, etc.
- BPI (biparental inheritance) (1): e.g. child AB, mother AA, father BB.
- UPI-P (paternal uniparental disomy) (2): e.g. child BB, mother AA, father AB (both B alleles inherited from the father)
- UPI-M (maternal uniparental disomy) (3): e.g. child BB, mother AB, father AA (both B alleles inherited from the mother)
- MI-S (single allele mendelian inconsistency) (4): e.g. child AB, mother AA, father AA
- MI-D (double allele mendelian inconsistency) (5): e.g. child AA, mother BB, father BB.

Upon invocation of the *SNP Trio* function, a dialog appears which is used to specify the trios (Figure 8). It could be done by either specifying the columns with mothers and fathers (*Trios* tab) or by manually selecting samples (*Manual* tab).

Sample specification

Trios Manual

Father Column:

Mother Column:

Add Trios

Trios

Father	Mother	Child

Figure 8: SNP Duo setup dialog. Family trios need to be specified.

PGS will then generate a SNP trio spreadsheet, with each trio shown in one row (an example of a single trio is shown in Figure 9. The samples are specified in three left-most columns, while each of the remaining columns corresponds to one SNP. The codes 0 – 5 are used as explained at the beginning of this section.

1. Father	2. Mother	3. Child	4. SNP_A-1780286	5. SNP_A-1780461	6. SNP_A-1780575	7. SNP_A-1780711	8. SNP_A-1781105	9. SNP_A-1781550	10. SNP_A-1781774	11. SNP_A-1781745	12. SNP_A-1781702	13. SNP_A-1782317	14. SNP_A-1782274	15. SNP_A-1782274
22N	95N	151N	0	0	0	4	2	0	0	1	0	0	0	0

Figure 9: Result of SNP trio analysis. Each column represents one SNP while the cell entries show the analysis of uniparental inheritance.

Furthermore, the results of trio analysis can be visualized in the chromosome view (in the Visualization section of the workflow). The central part of the plot is the SNP trio track, as illustrated in Figure 10. Each dot on the plot shows a SNP, while the possible inheritance scenarios are given on the y-axis.

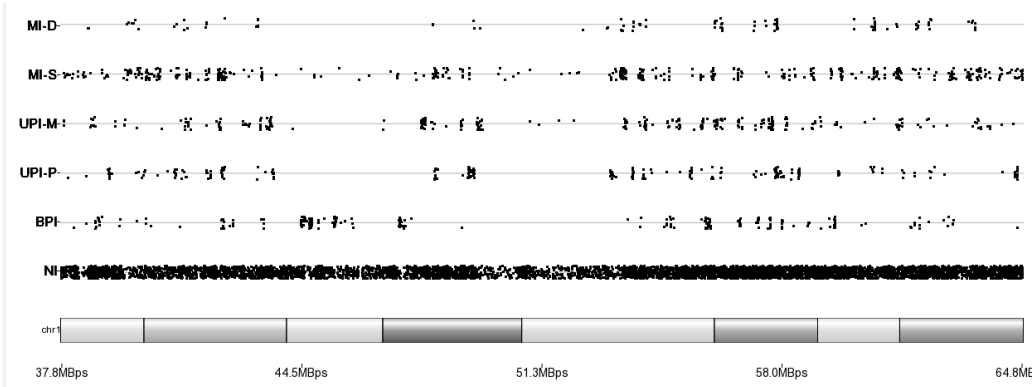


Figure 10: SNP trio track of the chromosome view. Each dot represents one SNP. Various modes of uniparental inheritance are shown on the y-axis.

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.

References

Stevens EL *et al.* Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. PLoS Genetics 2011;7: e1002287

Note

Last revision: January 2012