

# Combining and Analyzing Data from Different BeadChips

A merged project in BeadStudio software can be used for integrated analysis of data generated from different versions of Infinium® BeadChips.

## INTRODUCTION

Illumina's family of Infinium HD BeadChips enables higher sample throughput per BeadChip and lower DNA input per sample, while maintaining the high quality of data generated by the robust Infinium Assay. As a result, researchers can significantly reduce the time it takes to go from data acquisition to analysis and publication. The BeadStudio GT module (v3.2 or higher) lets customers seamlessly upgrade their research projects to the Infinium HD product line by allowing cross-product data analysis.

Cross-product integrated analysis and streamlined data export to third-party downstream analysis packages are possible from within a single BeadStudio project. This is important for researchers in the middle of a project started with Infinium II products, who want to migrate to Infinium HD products. It is also important for labs that want to ensure they are using a stable assay platform

amenable to lengthy longitudinal or replication studies. This Technical Note describes the types of data that can be combined, how to set up a merged project, and how to analyze and interpret the combined data.

## COMBINING DATA FROM TWO PRODUCTS

Creating a merged data set from two projects in BeadStudio begins by creating a sample sheet that links the samples to be loaded into the project with the manifest file (\*.bpm) for each product each sample was run on. Genotype calls and CNV data (i.e., B-allele frequency and Log R ratios) can be analyzed in a merged project. However, only the data for the SNPs and probes in common between the two manifests listed are displayed in the merged project. Data for SNPs and probes that are not common are not loaded into the project and will not be used in subsequent analyses.

FIGURE 1: SAMPLE SHEET OF MERGED DATA SET

	A	B	C	D	E	F	G	H	I
1	[Header]								
2	Investigator Name	Joe Investigator							
3	Project Name	610 & 550							
4	Experiment Name	610 & 550							
5	Date	3/19/2008							
6									
7	[Manifests]								
8	Manifest	Name	Ver						
9	A	Human610-QuadV1_B		1					
10	A	HumanHap550-2v3_B		3					
11	[Data]								
12	Sample_ID	SentrixBarcode_A	SentrixPosition_A	Version	Sample_Name	Replicate	Parent1	Parent2	Gender
13	NA06991_600	4155472198	R02C01	1	NA06991_600				Female
14	NA07029_600	4155472247	R02C01	1	NA07029_600				Male
15	NA07348_600	4155472226	R02C02	1	NA07348_600				Female
16	NA10830_600	4155472078	R01C01	1	NA10830_600				Male
17	NA10851_600	4155472247	R01C01	1	NA10851_600				Male
18	NA06991_550	1983150061	A	3	NA06991_550				Female
19	NA07029_550	1983150065	B	3	NA07029_550				Male
20	NA07348_550	1983150058	A	3	NA07348_550				Female
21	NA10830_550	1983150060	A	3	NA10830_550				Male
22	NA10851_550	1983150065	A	3	NA10851_550				Male

To create a merged project from two data sets, both manifests must be identified in the [Manifests] section of the Sample Sheet. Individual samples in the [Data] section should reference the applicable version number.

FIGURE 2: SNP TABLE OF MERGED PROJECT

Name	Chr	Position	ChITest100	Het Excess	AA Freq	AB Freq	BB Freq	Call Freq	Minor Freq	10% GC	50% GC	HW Equil	# Calls	# no calls	1		3	
															GenTrain Score	Cluster Sep	GenTrain Score	Cluster Sep
rs12700934	7	29145...	0.0000	0.5000	0.3333	0.6667	0.0000	1.0000	0.3333	0.4453	0.8846	0.0633	12	0	0.8502	0.5970	0.7890	0.3234
rs12700978	7	29783...	0.0000	-0.6571	0.3333	0.1667	0.5000	1.0000	0.4167	0.8062	0.8062	0.0228	12	0	0.7939	1.0000	0.8261	1.0000
rs12700997	7	29904...	0.0000	0.5000	0.3333	0.6667	0.0000	1.0000	0.3333	0.8609	0.8609	0.0833	12	0	0.8316	1.0000	0.8441	1.0000
rs12701006	7	30113...	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000	0.0000	0.8330	0.8499	1.0000	12	0	0.8668	0.6365	0.8543	0.7198
rs12701011	X	14678...	0.0000	-0.6571	0.3333	0.1667	0.5000	1.0000	0.4167	0.8482	0.9133	0.0228	12	0	0.8877	0.7326	0.8223	0.6746
rs12701014	7	36544...	0.0009	0.3333	0.0000	0.5000	0.5000	1.0000	0.2500	0.9203	0.9496	0.2482	12	0	0.9163	1.0000	0.8829	1.0000
rs12701020	7	30661...	0.3633	0.0909	0.0000	0.1667	0.8333	1.0000	0.0833	0.8078	0.8078	0.7528	12	0	0.7950	1.0000	0.8407	0.6721
rs12701035	7	30798...	0.3652	0.0906	0.8333	0.1667	0.0000	1.0000	0.0833	0.9514	0.9558	0.7537	12	0	0.9245	1.0000	0.9187	1.0000
rs12701041	7	30812...	0.0455	0.2000	0.0000	0.3333	0.6667	1.0000	0.1667	0.8696	0.8696	0.4884	12	0	0.8383	1.0000	0.9278	1.0000
rs12701044	7	36974...	0.0455	0.2000	0.6667	0.3333	0.0000	1.0000	0.1667	0.9371	0.9592	0.4885	12	0	0.9292	1.0000	0.9011	1.0000
rs12701064	7	31212...	0.0455	0.2000	0.6667	0.3333	0.0000	1.0000	0.1667	0.9405	0.9416	0.4885	12	0	0.9064	1.0000	0.9275	1.0000
rs12701070	7	31256...	0.0009	0.3333	0.5000	0.5000	0.0000	1.0000	0.2500	0.8547	0.8547	0.2482	12	0	0.8270	1.0000	0.8387	1.0000
rs12701102	7	31522...	0.0000	0.7143	0.0000	0.8333	0.1667	1.0000	0.4167	0.8752	0.9635	0.0133	12	0	0.9355	1.0000	0.8428	0.6826

The SNP Table of a merged project displays separate parent columns for each of the data sets indicated by the Version number (1 or 3 in the blue parent columns of this example).

For example, HumanHap550-Duo v3.0 data can be merged with Human610-Quad v1.0 data. However, to take advantage of the additional ~60,000 CNV-specific markers from the Human610-Quad v1.0 product (or any other HD product), analysis must be performed without merging data between these two products.

CREATING A MERGED PROJECT SAMPLE SHEET

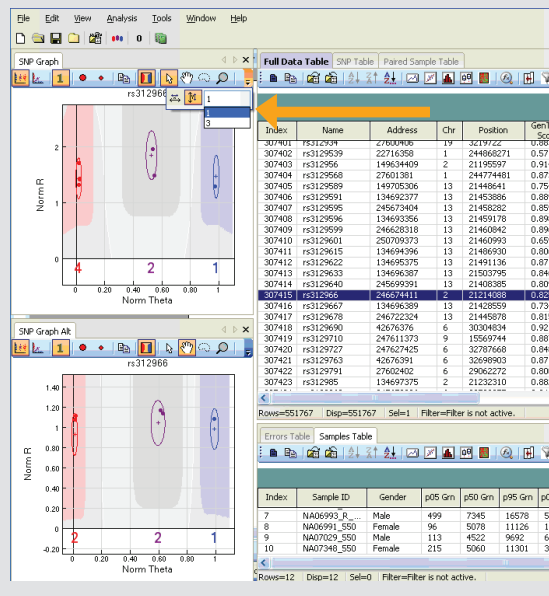
Figure 1 shows an example of a sample sheet that is correctly formatted to load a merged data set. Under the [Manifests] section, both manifests are identified by name and a new version (Ver) column has been added. If there are differences in the annotation data between manifests, BeadStudio uses the annotation data from the first manifest listed in the sample sheet. Researchers should list the most recent manifest first, so the data loaded into BeadStudio will be based on the most recent annotation of the genome.

In the [Data] section of the sample sheet, the standard formatting for each sample is used and Sample\_ID, SentrinxBarcode, and SentrinxPosition are listed. Values entered into the new Version column link the SentrinxBarcode with the correct manifest as specified in the [Manifest] section. This value is also populated in a new Version column in the BeadStudio Samples Table.

NEW COLUMNS IN THE SNP TABLE

A merged project created in BeadStudio using a sample sheet similar to the one shown in Figure 1 permits researchers to analyze the two data sets in the same project. However, there are some differences in the SNP

FIGURE 3: DISPLAYING SNP GRAPHS OF MERGED DATA SETS



Two SNP graphs (SNP Graph and SNP Graph Alt) can be displayed at one time in BeadStudio. Clicking on the arrow at the far right of the SNP Graph toolbar (marked with orange arrow) lets you select which data set to display. The numbers in the dropdown menu correspond to the Lot/Version number of the product.

In this example, seven samples processed with the Human610-Quad v1.0 product are displayed by selecting "1" in the SNP Graph dropdown menu (top plot) and five samples processed with the HumanHap550-Duo v3.0 product are displayed by selecting "3" in the SNP Graph Alt dropdown menu (bottom plot).

**TABLE 1: CONCORDANCE SUMMARY OF CNVS DETECTED USING HUMANHAP550 AND HUMAN610-QUAD BEADCHIPS**

	550	610	550 $\cap$ 610	IN 610, NOT IN 550	IN 550, NOT IN 610	% REGION COVERAGE OF 610 REGIONS BY 550	% BASE COVERAGE OF 610 REGIONS BY 550
<b>AVERAGE</b>	4	2.8	2.3	0.5	1.8	84.9	83.0
<b>MEDIAN</b>	4	3	2	0	1	100	93
<b>25th %ile</b>	3	2	1	0	1	75	78
<b>75th %ile</b>	5	4	3	1	2	100	100

Table for a merged project compared to a standard single-product project.

After the merged project is loaded in BeadStudio, new columns are populated in the SNP Table (Figure 2). The SNP Table shows new parent columns for each manifest and version or lot-specific subcolumns.

A dropdown menu on the SNP Graph header lets you select which product data to display (Figure 3). Each product is referred to in this menu by the lot number defined in the sample sheet (Figure 1). For this procedure, Lot number and Version number are identical and this value is defined in the Ver column of the [Manifest] section of the sample sheet (Figure 1).

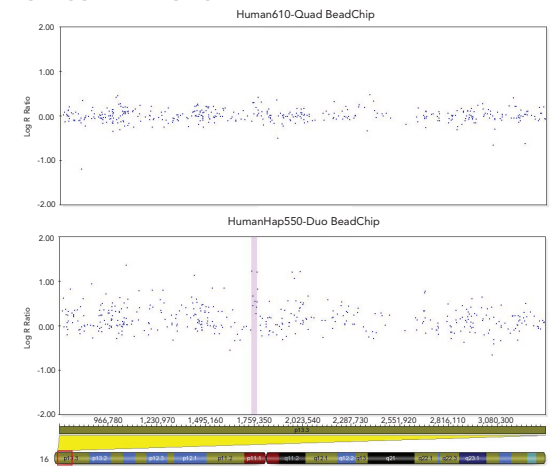
**CLUSTER GENERATION**

A separate cluster file (\*.egt) must be loaded for each version or lot listed in the sample sheet to generate the highest quality data for both genotyping and downstream CNV analysis.

Noise in the data impacts the log R calculations used by CNV identification algorithms. This noise can be reduced by using the proper cluster file and by following strict experimental procedures. This begins with precise quantitation of DNA input amounts using PicoGreen to ensure that experimental input amounts match that used to create the Illumina-supplied standard cluster file (750 ng for Infinium II BeadChips, 400 ng for two-sample Infinium HD BeadChips, and 200 ng for four-sample Infinium HD BeadChips). During processing, oven temperatures and staining temperatures should also be routinely checked to ensure precise protocol conditions.

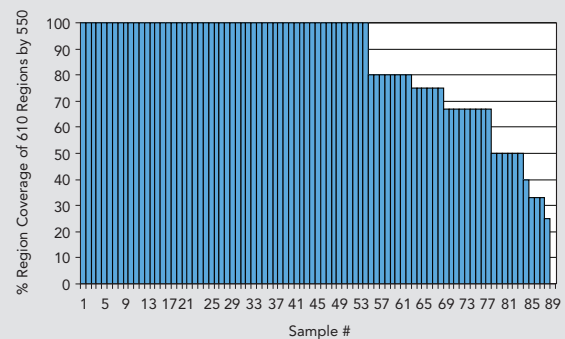
During the initial creation of a BeadStudio project, only one cluster file can be specified with the Project Creation wizard. Additional cluster files must be loaded after project creation using the Import Cluster Positions function. After selecting **File | Import Cluster Positions**, you are prompted to associate a cluster file (\*.egt file) with a Version/Lot number.

**FIGURE 4: EXAMPLE OF DISCORDANT DETECTION OF A LOW-CONFIDENCE CNV**




In a region of chromosome 16, a CNV is detected by cnvPartition using the HumanHap550-Duo data (bottom), but not from the Human610-Quad data (top). This difference could be the result of false positive detection due to greater noise in the HumanHap550-Duo data. The CNV was given a relatively low confidence value by cnvPartition.

**FIGURE 5: HISTOGRAM OF OVERLAP BETWEEN CNVS DETECTED ON TWO DIFFERENT BEADCHIPS**



Most samples have 100% concordance of CNV regions found using both BeadChips. Some regions are not found by both BeadChips, possibly due to having lower CNV confidence values or biologically derived differences.

In this example, a popup window appears, asking “Is this the cluster file for Lot 1?” Clicking Yes allows you to navigate to the cluster file for Lot 1. Clicking No prompts a second popup window asking, “Is this the cluster file for Lot 3?” Clicking Yes at this point allows you to navigate to the appropriate cluster file for Lot 3.

Once all the cluster files have been imported, update the sample statistics by clicking the calculate button in the Samples Table . Separate SNP graphs (SNP Graph and SNP Graph Alt) are available for each cluster file version (Figure 3). To specify the active cluster file version used to generate the sample data, use the dropdown menu located at the far right of the SNP Graph tool bar. The bottom of each SNP Graph shows genotype counts specific to that cluster file.

#### CNV ANALYSIS OF MERGED DATA

Once a merged project is created, it can be used to identify regions of CNV across all samples in the merged data set. The method for performing CNV analysis using *cnvPartition* is the same for merged data sets as it is for single data sets. The starting point is the merged project, rather than a single-product project.

#### EXPERIMENTAL VALIDATION OF CNV ANALYSIS

As with all Illumina products, internal scientists have rigorously tested the robustness of combining data sets for use in CNV analysis. CNV analysis using the Illumina *cnvPartition* plug-in algorithm was performed on the data obtained for 89 HapMap samples that were processed using both the HumanHap550-Duo v3.0 and the Human610-Quad v1.0 BeadChip products. A stringent confidence threshold was set (75) for the purpose of this analysis in order to reduce the detection of false positives, which might prevent an accurate comparison between the products.

In the majority of cases (54 out of 89 samples), the boundaries of all detected CNV regions showed 100% concordance between the two data sets (Figure 4 and Table 1). Discrepancies in CNV detection between the remaining pairs of data sets resulted from several contributing factors. Many of the uniquely identified CNV regions using data from either the HumanHap550-Duo v3.0 or the Human610-Quad v1.0 scored at relatively lower confidence (CNV confidence < 100), and therefore, may represent false positive detection (Figure 4). In addition, the HapMap samples are from immortalized cell lines of phenotypically normal individuals with very small (if any) regions

of CNV that can be difficult to detect. The data analyzed from the two BeadChip products were collected at different times, and it is possible that instability in the cell line caused the chromosomal composition of the cell line to change over time. Thus, differences in detection between the two data sets may be biologically derived through an artifact of cell culture. In general, the data from the new Infinium HD products showed a lower standard deviation of the Log R ratio than the Infinium II data, which will enable researchers to refine regions of CNV in their cross-platform research projects.

#### SUMMARY

The Illumina BeadStudio Genotyping Module allows researchers to merge data sets from different products. A merged project supports the analysis of both Infinium II and Infinium HD genotyping and CNV data in a single project. The data generated from the two platforms are highly stable, reflecting the intrinsic robustness of the Infinium Assay and allowing researchers to combine projects across BeadChip versions and pursue downstream analyses with confidence.

#### ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina’s DNA Analysis and Software products.

Illumina, Inc.  
**Customer Solutions**  
 9885 Towne Centre Drive  
 San Diego, CA 92121-1975  
 1.800.809.4566 (toll free)  
 1.858.202.4566 (outside North America)  
 techsupport@illumina.com  
 www.illumina.com

#### FOR RESEARCH USE ONLY