Informatic Updates for Whole-Genome Microarrays

Illumina has released new product files (.bpm, .egt, and .xml) for all currently available human whole-genome genotyping microarrays. These new files provide a number of improvements requested by researchers, including:

- **Marker positions in human genome build 37/hg19**
- **Genomic strand information (+/-)**
- **Removal of indeterminate loci**

What is a new genome build?

While the human genome reference sequence has been available to researchers for many years, groups such as the Genome Reference Consortium are constantly striving to improve the sequence accuracy. Periodically (every few years), they release a new reference genome that incorporates the latest data, fills sequence gaps, corrects any known errors, and maps any newly discovered markers or features across the genome.

What is the most recent genome build?

The most recent human genome build is commonly referred to as hg19 or build 37, released in February 2009. The new genome build can be visualized using publically available tools like the UCSC Genome Browser (http://genome.ucsc.edu/).

What is the difference between build 36 and build 37?

Build 37 offers a large number of small improvements to the accuracy of the reference sequence. For complete release notes, please review the information provided by the Genome Reference Consortium, which lists the changes made to each chromosome in detail. The release notes are available from the following web address: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml.

How does using build 37 information help my research?

Build 37 represents the most up-to-date information from the human genetics community regarding the location of features (i.e. genes, exons, SNPs, repeat regions etc.) across the human genome. It is important that research findings be placed in the correct context across the genome for data interpretation.

What is the benefit of strand information?

SNPs can be assayed against either strand of DNA. Knowing this information allows researchers to convert genotypes to a consistent strand, which enables data sets to be merged across different microarrays for imputation analysis and facilitates comparison with publicly available data such as that from the HapMap and 1000 Genomes Projects.

Will I need a new version of GenomeStudio® software to read these new manifest files?

GenomeStudio v1.6 (version 2010.1) or higher is needed to read the new manifest files that include +/-strand information.

What do you mean by “indeterminate loci”?

As researchers begin to incorporate more rare variants into their studies, having very high genotype accuracy is increasingly important. To address this, Illumina scientists are incorporating additional tests into the quality control (QC) pipeline for whole-genome genotyping arrays. These tests include more diverse DNA samples during the cluster training process, increased standards for reproducibility of genotypes (within and across products), as well as concordance with high-quality sequence data. Moving forward, these QC tests will be a standard part of the Gentrain process (cluster training process) for new products. These stringent standards have been applied to all currently available whole-genome genotyping arrays. Markers that do not pass the increased stringency of these new tests are referred to as “indeterminate loci.”

How should I use these files if I am just starting a new project?

Researchers should process samples according to standard protocol using the new manifest, cluster, and product descriptor files.

How should I use these files if I am In the middle of a project?

There are two options for this situation. Illumina recommends that all data be reloaded into GenomeStudio to re-call genotypes and SNP metrics using the new files. Alternatively, the project can be continued and completed using the original files. The build 37 map positions can then be updated and sub-standard loci removed at a later time using a second-stage analysis program. Name and position conversion files (i.e. build 36 to build 37) along with lists of indeterminate loci are available for all currently available arrays upon request from Illumina Technical Support (TechSupport@illumina.com).

How should I use these files if I am finished with my project?

A researcher interested in updating their completed data to build 37 can use the information in the new manifest files to update chromosome and base pair positions in secondary analysis files (i.e., map files). In addition, a list of indeterminate loci for each currently available whole-genome microarray is available upon request from Illumina Technical Support (TechSupport@illumina.com). Researchers should feel free to use their own professional judgment in removing these loci from analysis entirely or flagging them for additional review to avoid spurious associations.

How should I use these files if I am a LIMS customer?

Illumina strongly recommends converting to the newest version of these files as soon as possible in any stage of a project. As a reminder, in laboratory instrument management systems (LIMS), the product files are bound to the product that is accessioned into Project Manager. Also, Autocall will use the current file in the LIMS database for all future *.gtc file generation for that BeadChip product. Therefore, researchers should only update Project Manager with the new product files when they are confident that all projects containing the specific BeadChip product can be transitioned to the new versions, and that data previously produced with the older versions should be re-queued for Autocall.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, HiScan, Eco, and TruSeq are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 370-2010-028 Current as of 18 Oct 2010

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA