Imputing HumanHT-12 Expression BeadChip Data

GenomeStudio® Gene Expression Module offers users the option to impute data for HumanHT-12 BeadChip bead types using a well-established algorithm.

Introduction

The HumanHT-12 v4.0 BeadChip is Illumina's highest throughput, lowest cost solution for whole-genome expression profiling and expression-based quantitative trait loci (eQTL) studies. This BeadChip contains 12 arrays, each featuring more than 48,000 probes that provide genome-wide coverage of well-characterized genes, gene candidates, and splice variants. Illumina guarantees that on any given HumanHT-12 BeadChip array more than 99.98% of the bead types, or probes, will be present. On average, each bead type will be represented with 15-fold redundancy; however, up to nine bead types may be represented by only 0, 1, or 2 copies on a given array, resulting in missing data. Because many downstream analyses such as normalization, clustering, and principal component analysis need complete data sets, Illumina GenomeStudio analysis software requires researchers to impute or exclude missing HumanHT-12 BeadChip data in their gene expression projects.

Missing Data in GenomeStudio

GenomeStudio Gene Expression Module offers users the option to impute or exclude missing bead types at project creation and before reanalyzing a project. After HumanHT-12 BeadChip data have been loaded, GenomeStudio detects any missing bead types. If missing bead types are detected, a dialog box appears asking the user if they would like to impute the missing data (Figure 1). Choosing to impute missing data estimates the value for missing bead types, allowing 100% of the bead types to be analyzed in the project. In contrast, choosing not to impute missing data will result in GenomeStudio excluding the missing bead types for all microarrays in the project. When a bead type is missing on any microarray in a project, that bead type will be excluded from all the micorarrays of the project even if valid measurements were made for the other microarrays. GenomeStudio will create a new table that displays the excluded or imputed probes (Figure 2).

k-NN Data Imputation Algorithm

GenomeStudio uses the k-Nearest Neighbor (k-NN) algorithm to estimate values for missing bead types on the HumanHT-12 BeadChip. This well-established algorithm is commonly used to impute data and has been shown to be a robust method for estimating values for missing microarray data[1].

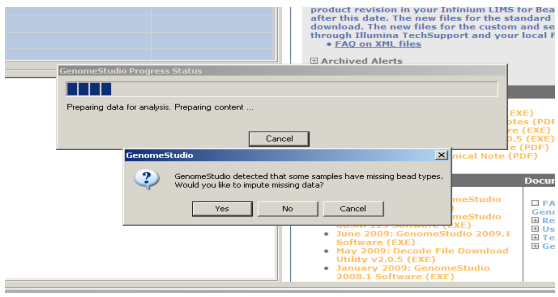
The k-NN imputation algorithm selects genes with expression values similar to the gene of interest in imputing the missing value. For each missing value, the method identifies other genes with expression most similar to the genes of interest. The weighted average of values for those respective genes is then used to estimate the missing values. In this way, the k-NN imputation algorithm provides a robust and sensitive method for missing value estimation. Refer to the following section for a more detailed description of the k-NN imputation algorithm.

Imputing Data In GenomeStudio

The following steps describe in more detail how GenomeStudio imputes missing data using the k-NN imputation algorithm.

1. **Let X_{ip} denote a matrix of HumanHT-12 BeadChip data with i samples and p bead types that contains all missing and non-missing intensity values.**
2. **Filter X_{ip} to obtain a matrix, X_{ij} containing non-missing intensity values for i samples and j remaining bead types.**
3. **Let Yim denote a matrix of m bead types across i samples that have a missing value in at least one sample. If, for example, bead type m = 1 has a missing intensity value in sample i = 1, GenomeStudio identifies the non-missing intensity values, denoted by zim, in the remaining samples where i ≠1 and m = 1.**
4. **Within each sample i ≠1, Euclidean distances are calculated from the zim intensity values to all other bead type intensity values in X_{ij}.**
5. **Within each sample, k nearest neighbors, where k = 15, are determined between zim and X_{ij}.**
6. **The distance weighted average is calculated using the 15 nearest neighbors.**
7. **The missing value for bead type m = 1 is replaced with the weighted average.**
8. **GenomeStudio repeats this for all missing bead types. On each HumanHT-12 BeadChip array, there will be no more than 0.01% of probes, or five bead types, missing.**

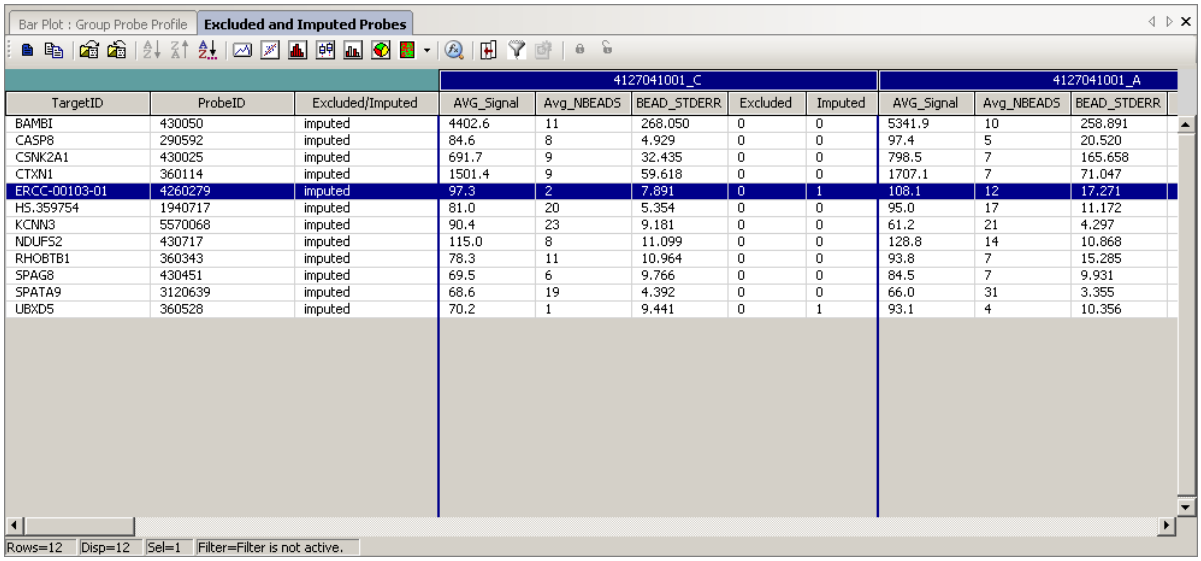
Figure 1: Impute Or Exclude Dialog Box



If GenomeStudio detects missing data, the Impute or Exclude dialog appears. If GenomeStudio does not identify missing data, the project is created and appears in the GenomeStudio main window.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 2: Excluded And Imputed Probes Table



			4127041001_C					4127041001_A		
TargetID	ProbeID	Excluded/Imputed	AVG_Signal	Avg_NBEADS	BEAD_STDERR	Excluded	Imputed	AVG_Signal	Avg_NBEADS	BEAD_STDERR
BAMBI	430050	imputed	4402.6	11	268.050	0	0	5341.9	10	258.891
CASP8	290592	imputed	84.6	8	4.929	0	0	97.4	5	20.520
CSNK2A1	430025	imputed	691.7	9	32.435	0	0	798.5	7	165.658
CTXN1	360114	imputed	1501.4	9	59.618	0	0	1707.1	7	71.047
ERCC-00103-01	4260279	imputed	97.3	2	7.891	0	1	108.1	12	17.271
HS.359754	1940717	imputed	81.0	20	5.354	0	0	95.0	17	11.172
KCNN3	5570068	imputed	90.4	23	9.181	0	0	61.2	21	4.297
NDUFS2	430717	imputed	115.0	8	11.099	0	0	128.8	14	10.868
RHOBTB1	360343	imputed	78.3	11	10.964	0	0	93.8	7	15.285
SPAG8	430451	imputed	69.5	6	9.766	0	0	84.5	7	9.931
SPATA9	3120639	imputed	68.6	19	4.392	0	0	66.0	31	3.355
UBXD5	360528	imputed	70.2	1	9.441	0	1	93.1	4	10.356

All missing data, whether imputed or excluded, appear in the Excluded and Imputed Probes table. In this table, the Excluded/Imputed column shows whether a probe was excluded or imputed. Every sample column has Excluded and Imputed subcolumns with values of 0 or 1. A value of 1 indicates that this bead type is "missing" because the number of beads present is fewer than the defined limit of three. If the user chooses to exclude a given bead type in one sample, the bead type will be excluded across all samples in the project.

Summary

In order to support higher throughput and more cost-effective expression analysis, the 12-sample HumanHT-12 v4.0 BeadChip contains fewer beads per bead type. As a result, up to nine bead types may not be represented on a given HumanHT-12 BeadChip array. To enable data to be collected on 100% of the bead types, Illumina's GenomeStudio Gene Expression Module allows users to impute missing data using a well-established algorithm for estimating microarray data.

Additional Information

For more information about Illumina gene expression solutions, please visit www.illumina.com or contact us at the address below.

References

1. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(6): 520–525.

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

© 2011 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, TruSeq, Eco, and MiSeq are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 470-2008-009 Current as of 21 January 2011

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA