Supporting Online Material for

A reannotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data

Nuno L. Barbosa-Morais[‡], Mark J. Dunning, Shamith A. Samarajiwa, Jeremy F. J. Darot, Matthew E. Ritchie, Andy G. Lynch and Simon Tavaré

[‡]Corresponding author
Mailing address: CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
Fax: +44 1223 404208 Phone: +44 1224 404297
E-mail: Nuno.Barbosa-Morais@cancer.org.uk

Contents:

Supporting Methods

BLAST parameters

An e-value of 10^{-6} was chosen as a threshold for BLAST because it was found to be loose enough to select alignments representative of all putative targets susceptible of providing considerably sensitive signal. As illustrated in **Figure S19**, for $e = 10^{-6}$ the probe-target similarity is always under 90% and well below the 96% associated with the two mismatches allowed. An even less stringent threshold would increase the computing time with no additional benefit.

The DUST filter was turned off because low complexity sequences are also putative targets for microarray probes and therefore have to be considered in the assessment of probe specificity.

Filtering strategies for the analysis of the MAQC-V1 dataset

Detection: the function *detectionCall* in *lumi* (1) was used to determine how many of the 30 arrays each gene was detected on by using the detection p-values provided by Illumina and a fixed p-value threshold, with values below this threshold deemed to be detected; genes detected on at least one array passed the filter.

Expression: the *kOverA* function from the *genefilter* package (2) was applied to the non-normalised expression values to determine which genes had expression level higher than *A* on at least *k* of the 30 arrays. The value of *k* was set to 10 for varying values of *A*.

IQR: the inter-quantile range of each probe was calculated from $\log_2$ normalized data and genes greater than a given cut-off passed the filter. A range of different cut-offs was used for each of the methods, and the number of probes in each of the annotation categories retained by the filter was recorded.

More details can be found in (3).

References

1. Du, P., Kibbe, W.A. and Lin, S.M. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547-1548.
2. Gentleman, R., Carey, V., Huber, W. and Hahne, F. Bioconductor, genefilter: methods for filtering genes from microarray experiments.
3. Dunning, M.J. (2008) Genome-wide analyses using bead-based microarrays. PhD Thesis, University of Cambridge. http://www.dspace.cam.ac.uk/handle/1810/218542

Legends for Supporting Figures

Figure S1. Snapshots of ReMOAT. **A** – Input screen for conversion of identifiers; a list of Illumina probe IDs is used to search for selected alternative IDs. **B** – HTML output of the ID Converter for previous search, exhibiting gene information, several alternative IDs (Ensembl, Entrez, HGNC, Unigene, Lumi), and links to the corresponding databases, as well as to iHOP and Wikigenes. **C** – HTML output of the Probe Re-annotation tool for the same Illumina probe IDs.

Figure S2. A – Relative proportion of quality scores for probes on each of the three human WG platforms. **B** – Cumulative barplots of number of probes assigned each quality score, separated by primary source of derivation (RefSeq and UniGene (UG)), for each of the three human WG platforms.

Figure S3. A – UCSC genome browser graphic for the human *BCKDHB* gene on chromosome 6, which contains a SNP (rs7740958 – T/C) at position 7. The Illumina probe targeting this sequence (GI_34101271-I) contains a C at this location. There is also a probe (GI_34101266-A) targeting a constitutive splice junction of *BCKDHB* and matching no SNPs. **B** – Boxplots of the log-intensity of expression from probe GI_34101271-I in the Japanese HapMap population according to the rs7740958 genotype. **C** – Boxplots of the log-intensity of expression from probe GI_34101266-A in the Japanese HapMap population according to the rs7740958 genotype. **D** – Scatter-plot of log-intensity of expression from probes GI_34101271-I and GI_34101266-A in the Japanese HapMap population, with samples coloured according to the rs7740958 genotype. **E** – Boxplots of the log2 expression ratios in the Japanese HapMap population according to the rs7740958 genotype.

Figure S4. UCSC genome browser graphic for human probes ILMN_1692545 and ILMN_1670800, which differ in only one nucleotide corresponding to SNP rs13082444. Obsolete target transcripts XM_933970.1 and XM_944901.1 are also represented.

Figure S5. UCSC genome browser graphics for 5 different regions matched by a set of six human probes genomically clustered.

Figure S6. UCSC genome browser graphics for the human *GAGE* cluster on chromosome X. Represented is also a set of more than a dozen probes, each targeting transcripts from multiple genes from that cluster.

Figure S7. Histograms of distribution of the number of probes per gene for each of the human WG platforms, according to both the original Illumina annotation and our re-annotation.

Figure S8. UCSC genome browser graphics for the human *CAST* gene, including custom tracks representing the logged expression levels (red for brain samples and green for

reference samples) and associated log ratios in the MAQC dataset for all the probes targeting the *CAST* locus for each of the three human Illumina WG platforms.

Figure S9. Boxplots of expression ranks (out of all the DASL 'Perfect' probes) of probes targeting splice junctions for each of the 6 DASL samples.

Figure S10. UCSC genome browser graphic for the human *CPNE1*/*RBM12* locus, including the targeting Illumina probes.

Figure S11. UCSC genome browser graphic for the human *PCDHG* locus, including the targeting Illumina probes.

Figure S12. UCSC genome browser graphic for the murine *CDKN2A* locus, including the targeting Illumina probes.

Figure S13. UCSC genome browser graphic for the human intron/exon junction of *BC067758* covered by probe ILMN_1690644.

Figure S14. Box plots of distribution of average expression ranks of probes, across the GEO arrays, according to: **A** - the proportion of transcripts associated with a UCSC gene targeted by the probe; **B** - the proportion of transcripts associated with an Ensembl gene targeted by the probe; **C** - the number of GenBank transcripts associated with an UniGene gene targeted by the probe.

Figure S15. A – UCSC genome browser graphic for the human *BRCA1* locus, including the targeting Illumina probes. **B –** Histogram of distribution of log-ratios of expression between probes ILMN_2311089 and ILMN_1738027 for the 10 samples of the Miranda et al dataset (GEO series GSE13733).

Figure S16. UCSC genome browser graphic for the human *HYDIN* locus, including the targeting Illumina probes.

Figure S17. UCSC genome browser graphic for the human *FAM90A1* locus, including the targeting Illumina probes.

Figure S18. Venn diagrams of conservation of probe composition between the different WG-6 platform versions for Human (**A**) and Mouse (**B**).

Figure S19. Scatter plots comparing BLAST e-values (y-axis) and similarity (x-axis) between Human WG-6 V3 probes and putative transcriptomic (**A**) and genomic (**B**) targets. The points circled in blue correspond to alignments including gaps. For e = 10^{-6} the probe-target similarity is always under 90% (red lines) and well below the 96% (green lines) associated with the two mismatches allowed.