Look at

<http://www.bioconductor.org/help/workflows/oligo-arrays/>

Illumina arrays: 29mer address oligos. Hyb to RNA. Then hyb to series of decoding oligos – combinatorial pattern of +/- from these allows decoding of the 29mer address.

Notes from conversation with Zizhen Yao, May 3, 2011

Limma: might be able to set up a better design matrix than the multiple pairwise analyses to utilize data across arrays better. Test all three contrasts at once. Ryan’s analysis just pairwise (?). Also doesn’t necessarily take into account experimental batches.

Gene universe for GO analysis: any gene called as present in any one of the arrays, i.e. all expressed genes.

Zizhen’s favorite packages:

Limma, lumi, Seth Falcon’s GO analysis package, lumiHumanIDmapping, lumiHumanAll.db

Check the lumi vignette

Try BioC course materials on limma.

Venn diagrams use fixed cutoffs, so underestimate the overlap between the gene lists. To look at similarity between two conditions this isn’t too bad. But if we’re mostly interested in the differences between two samples, could use limma in a more sophisticated way to get those gene lists directly.

She does use |logFC| > 1 AND adj.P.val < 0.05, but she thinks the result would be pretty similar without using that adj.P.val filter, as the FC threshold is more stringent.

The variance filter removes things that don’t vary across any array being considered.

She uses lumi defaults.

Check out detectionCall function to get present/absent type calls. A gene being present in just one array is enough to keep it for further analysis.

Affy arrays also use a variance filter. (something about across the whole array?).

Affy’s NSfilter is also some kind of present/absent thing. Zizhen does NOT use this.

Perhaps better to do present/absent filtering on each condition separately not across all arrays together1

Affy: perhaps keeps one probe per entrezID.

Notes from conversation with Ryan Basom, May 3, 2011

1. Some GUI type tools that could be useful if I’m encouraging others to do their own analysis:

Ingenuity (3-seat license, use on own computer via website). GO analysis, etc. Has nice tutorials at the website – don’t hear back very much from users with problems. Register via form on FHCRC website (I have registered).

MEV – multi experiment viewer. Some statistical tools. Can export directly from HutchBase to this tool. Can do gene set enrichment analysis (Broad) here.

Partek genomic suite – a single licensed computer – go and sit at it to use it.

GenomeStudio – Illumina’s software for looking at beadarray data. I now have access to this – see his email of May 6th, afternoon.

Some helpful R packages: lumi, limma, genefilter, arrayQualityMetrics, Ringo, siggenes.

Also SAM (an alternative to limma = significance analysis of microarrays, is a separate program but has a Bioc equivalent and also an Excel plugin. **Look at SAM website**)

Illumina’s tech support can be contacted by email and can be quite responsive, but are not always correct.

2. His analysis

He makes his README pdfs just by copy-paste. Intending to learn Sweave.

The variance filter is to get rid of non-variable genes (nothing to do with replicates). It works PER COMPARISON (so I should redo looking at all 4 conditions simultaneously?)

Shorth is nicely explained in the bioconductor case studies book.

He has an R script that takes the targets file, and a file describing the comparisons of interest, and it makes the design matrix.

Limma users guide: check out chapter 8. Case study 11.7 (beadarray) might be useful too. Limma does Bayesian moderated t-tests. Can also do ANOVA.

GenomeStudio: does some filtering, gives some expressed/not calls using the mean of negative controls, background subtraction, but Ryan isn’t using this. It also takes rawdata (IDAP – proprietary format?) and does something with it. It also added the annotation that we see in the output file.

The quantile normalization strategy is something that he and Jeff decided on.

The array contains about 100 controls, allowing background estimation. Could subtract mean background, but he finds this gives some negative values which causes trouble for taking logs, so they don’t subtract background.

Can also get bead level data in text format, and Ryan has an R script that can take that and get the same results as GenomeStudio would have done. Beadarray package could be useful for this.

Usually he gives users significant gene list based on logFC of 0.585 (= 1.5 fold), but with Kyle’s data this gives a list of >1000 genes so he bumped up the threshold to logFC of 1 (= 2-fold change) Kyle also likes logFC=1.

In each pairwise limma comparison, probes are retained that pass filtering in either set of samples, so it’s a bigger number than passed in one set of triplicates along.

The table gives numbers per probe, and there are often >1 probe per gene.

These arrays are HT12 version 4. nuID identifiers were added by lumi.

# Things to check out

websites

SAM (significance analysis of microarrays)

R packages

beadarray (an alternative to lumi)