# Stemformatics Data Methods

## Normalization, Transformation and Annotation

Version 2.0.4

March 6, 2017

Author: Othmar Korn

www.stemformatics.org

twitter.com/stemformatics

# Table of Contents

# 1 Normalization

## *1.1 Illumina BeadChip Microarray*

### Methods

The Illumina microarray platform provides for multiple simultaneous measurements of gene expression by summarising a pool of probes (on average, each probe is represented up to 30 times) directed towards the target sequence. Illumina detection scores are calculated by ranking the z-value of a summarised probe relative to the z-value of the negative controls. The score provides a probability (P-value) that the probe is detected above the negative controls. Stemformatics uses the recommended threshold of $p < 0.01$, which provides less than 1% false positive rate.

Stemformatics uses the R/Bioconductor package *Lumi* to quantile normalize Illumina BeadChip array data. When plotted, the density of log(2) normalized expression is observed as a bi-modal distribution, segregating non-detected from detected probes. From this bi-modal distribution, an expression threshold (detection floor) is determined, which is used to assist researchers in deciding the reliability of expression of a given gene, as it is provided on each Stemformatics expression plot. Note that we do not make a hard detection call based on this artificial threshold for graphing purposes (i.e. we don't decide to include or exclude probe expressions). Some genes may pass the Illumina detection p-value filter, but may sit on or below our expression threshold.

The median of expression is provided for a given dataset and is taken to be the median of above-threshold log(2) normalized probe expression values. This allows the researcher to determine how the expression of a gene of interest compares to other genes in a given study.

### Tools

Raw expression data is normalized in-house using the *Lumi* Bioconductor package in R ( `http://www.bioconductor.org/help/bioc-views/release/bioc/html/lumi.html` ).

Processing steps can be summarised as:

- Load raw expression BeadChip data file
- Perform an “affy” background correction *
- Quantile normalize and log(2) transform
- Output background corrected, normalized, log(2) expression matrix

* The name is misleading; this type of background correction is applicable to both Illumina and Affymetrix data

## *1.2 Affymetrix GeneChip Microarray*

## Methods

The Affymetrix GeneChip platform includes a set of “mis-match” (MM) probes to complement each “perfect match” (PM) probe as a means of assessing hybridization (expression) intensity against background levels. The mis-match probe sequences serve as negative control probes and from this, chip effects can be observed and hence background correction can be applied (as in the MAS5 algorithm employed by Affymetrix). The PM probesets map to gene transcripts.

*NOTE:* We do not use the Affymetrix MAS5 normalization approach, nor make use of the detection calls provided.

We plot the density of log(2) normalized expression and we observe a bi-modal distribution, segregating non-detected and detected probes.

From this bi-modal distribution, the expression threshold (detection floor) is determined which we plot on our gene expression graphs.

The median of expression is provided for a given dataset and is taken to be the median of above-threshold log(2) normalized probe expression values.

## Tools

Two tool sets have been utilised by Stemformatics for the normalization of Affymetrix GeneChip microarrays.

### *R/Bioconductor - “simpleaffy”*

Processing steps can be summarised as:

- Load expression CEL files using `ReadAffy()` command
- Output QC plots and statistics (including RNA degradation RLE and NUSE)
- Perform GC-RMA background correction, requiring platform CDF and probe data files to incorporate probe affinity adjustments in addition to standard PM-MM (perfect match to mis-match ratio)
- Perform quantile normalization
- Output background corrected, normalized, log(2) probe expression matrix

### *Affymetrix Power Tools*

Where we cannot use “simpleaffy” (e.g. where Bioconductor probe meta file and/or CDF are unavailable, such as for uncommon platforms like Affymetrix U133AAofAV2, the “early release” version of the HT-HG-U133A platform) we use Affymetrix Power Tools “`apt-probeset-summarize`” program with the following chipstream:

```
rma-bg,quant-norm.sketch=0.usepm=true.bioc=true,pm-mm,med-polish
```

equating to:

- RMA background correction
- Quantile normalization
- PM-MM adjustment
- Median polish
- Log2

The Affymetrix Power Tools suite is released under an Open Source license and is supported by Affymetrix. Required chip design, probe grouping and annotation files are available from Affymetrix ("Library Files" packages).

Our preference is to perform GC-RMA correction using Bioconductor "simpleaffy" package wherever possible, but the Affymetrix Power Tools standard RMA variant also provides quality normalization results which are compatible with like methods implemented in Bioconductor.

## *1.3 Affymetrix Gene ST Microarray*

### Methods

The Affymetrix Gene ST platform comprises a much larger array design than the GeneChip platform, and is made up of exon-based probesets that are mapped to gene transcripts.

Unlike the GeneChip platform, the ST array design does not include MM (mis-match) probe sets which are normally used as negative controls for assessing chip effects and subsequently aiding in background correction of expression scores. Instead, antigenomic probe sets are included on the chip which can essentially be thought of as negative control probes, as for a given species (Human or Mouse), there should not be any hybridizations of gene products to these antigenomic probes.

As only PM (perfect match) probe sets are included for the target genome, when plotting the density of probeset expression we do not see a bi-modal distribution unlike GeneChip or Illumina BeadChip arrays. Instead, we expect to see an expression curve which more resembles a bell curve.

We set the expression threshold (detection floor) for these arrays to be the median of antigenomic probeset expression. Essentially, we say that if a PM probe set expresses similarly to the median of “background”, then there is low confidence in the expression status of the target gene.

The median of expression for a Gene ST array is taken to be the median of above-threshold log(2) normalized probe expression values.

### Tools

We use Affymetrix Power Tools to process Gene ST arrays.

Since GC-RMA background correction cannot be applied to Gene ST arrays, we opted for standard RMA (compatible with Bioconductor RMA implementation) and quantile normalization. This yields results almost identical to those achieved using the popular “`aroma.affymetrix`” Bioconductor package in R, but with a much faster run time and smaller memory footprint.

## *1.4 Affymetrix GeneChip Primeview*

### Methods

The Affymetrix GeneChip Primeview array differs from the previous generation of GeneChip microarrays in that only PM (perfect match) probesets are provided. This means we cannot perform a GC-RMA background correction, in this case we opt for RMA only.

Primeview arrays also include background control probesets and hence we are able to calculate the median of background to use as our detection threshold.

The median of expression is taken to be the median of above-threshold log(2) normalized probe expression values.

### Tools

***R/Bioconductor - "simpleaffy"***

Processing steps can be summarised as:

- Load expression CEL files using `ReadAffy()` command
- Output QC plots and statistics (including RNA degradation RLE and NUSE)
- Perform RMA background correction
- Perform quantile normalization
- Output background corrected, normalized, log(2) probe expression matrix

## *1.5 Agilent SurePrint G3*

### Methods

The Agilent SurePrint G3 series of microarrays are single colour oligonucleotide gene expression microarrays and are conceptually akin to Illumina or Affymetrix GeneChip arrays.

Agilent SurePrint chips include a plethora of well annotated control probesets including negative controls whose 95th percentile of expression we call the detection floor (threshold).

The median of expression is taken to be the median of above-threshold log(2) normalized probe expression values.

### Tools

We use R/limma (Ritchie et al 2015) to load and normalize Agilent sample expression files (*.txt).

Processing steps can be summarised as:

- Load raw expression sample data files (green channel only)
- Perform RMA background correction
- Perform quantile normalization
- Collapse replicate feature probe expression values to a mean value for each sample
- Output background corrected, normalized, log(2) expression matrix

## *1.6 RNAseq normalization*

### Methods

The current release of Stemformatics houses RNAseq gene expression data produced on Illumina and SOLiD platforms (Stemformatics v6.5.1, March 2017). By far the largest proportion of datasets have been sequenced on the Illumina HiSeq series of instruments in both Human and Mouse studies.

We receive barcode-trimmed FASTQ base-space or much less frequently, colour-space read libraries (e.g. from SOLiD) sequenced in either single or paired end configuration. If RNA samples have been multiplexed or pooled they are merged to a single FASTQ library per biological replicate sample before continuing.

After performing RNA sequence alignment (see “Tools” below), the matrix of annotated RNAseq read counts (i.e. raw counts) is normalized for library size (read depth) using TMM (Trimmed Mean of M-values) scaling factors (Robinson & Oshlack 2009) and further normalized for gene length, producing RPKM values (Reads Per Kilobase per Million).

RPKM counts undergo a log(2) transformation and are subject to a further median scaling.

Normalization to RPKM allows relative gene expression levels to be compared across genes in Stemformatics within a given dataset.

*NOTE*: RPKM data is not suitable for statistical analyses such as DEG (differential expression analysis) – in such cases, raw counts are usually required in order for the analysis tool to perform its own internal transformations.

### Exceptions to methods

Some RNAseq experiments in Stemformatics have been processed by third parties (e.g. as part of collaborative projects and in other special circumstances) – where normalization methods differ from the Stemformatics pipeline, these are described in the dataset summaries.

### Tools

FASTQ libraries are quality controlled using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to check for both uniformity and consistency of nucleotide distributions and Phred quality scores across read bases.

Stemformatics uses R/Subread (Liao, Smyth & Shi 2013) to align sequences and annotate RNAseq reads to the Ensembl transcriptome.

Annotated counts are TMM scaled using R/edgeR (Robinson, McCarthy & Smyth 2010) and median-scaled using R/limma (Ritchie et al 2015).

The Stemformatics RNAseq pipeline is implemented in the “s4m” (release pending) software suite.

## *1.7 Batch correction*

In both microarray and RNAseq experiments we sometimes observe batch (technical) effects.

Where the dataset's experimental design accommodates batch correction and the dataset otherwise passes our QC criteria, we run a standardised batch correction procedure.

### Standardised batch correction methods

For microarray datasets, we run **ComBat** from R **sva** package (Johnson et al 2007 Biostatistics, https://doi.org/10.1093/biostatistics/kxj037 ). The batch effect is removed from the post-normalized log2 expression data.

For RNAseq datasets we use function *`removeBatchEffect`* from R **limma** package (Ritchie et al 2015 Nucleic Acids Research, https://doi.org/10.1093/nar/gkv007 ). In this method, we fit a linear model (capturing both phenotype and batch dimensions) to the TMM normalized RPKM log2 data and the batch component is subsequently subtracted.

In both cases, QC plots before and after batch correction are used to validate the successful removal of batch.

*NOTE:* We initially trialled the ComBat method for RNAseq data but found that it removes genes from the data where there are any missing ('`NA`') values for any samples – this results in removal of a sizeable proportion of gene expression (also there are conflicting opinions in the literature as to the validity of ComBat for count data). ComBat does not exhibit this behaviour with microarray data because generally we do not observe any missing values (raw data is continuous and positive within a typical dynamic range).

However, limma's `removeBatchEffect()` is lossless i.e. preserves all gene data and gives us good results for RNAseq batch removal.

Where datasets have had batch correction applied, this information is available on the dataset summary page.

# 2 Dataset Processing

## *2.1 Basic requirements*

We choose high quality studies for inclusion in Stemformatics based on the following criteria

- Value to the stem cell community
- Study's raw expression data is available in a public repository; e.g. *Array Express* ( `http://www.ebi.ac.uk/arrayexpress/` ) or *GEO* (the Gene Expression Omnibus `http://www.ncbi.nlm.nih.gov/geo/` )
- Study includes use of biological replicates in experimental design
- Adherence to standard protocols and procedures and QC measures

## *2.2 Curation and Annotation*

Annotation is performed on a set of tab-delimited plain text files which are collectively known as a “Stemformatics Dataset” (or “s4m dataset”) and are created, annotated, validated and normalized using an in-house tool. The files are formatted for database loading, and are human or machine editable (in the same spirit as MAGE-TAB files).

Sample names and investigation metadata are hand curated by experienced biologists. Sample naming adheres to experiment's published sample names as much as possible; although we do curate sample names, cell types and descriptions as necessary for our Stemformatics audience.

We include sample metadata available from *Array Express*. These metadata adhere to the MGED ontology and although we do not currently drive Stemformatics functionality against ontological terms or relationships, we do store investigation and sample metadata in Stemformatics for user consumption, also sourcing information from published papers and material related to the study in question.

Investigation metadata are annotated using our annotation tool or can be performed directly on our tab-delimited files.

When normalization and annotation are complete, an automated validation script checks for the presence of all required files and metadata fields within files, and also performs sample expression integrity and consistency checks before allowing the curator to load the Stemformatics Dataset into the Stemformatics database.

At this point we generate *GenePattern* GCT and CLS compatible files from our normalized probe expression values.

We also compile binary versions of GCT files for use by the Stemformatics analysis pipeline for faster loading and processing.

See *Appendix 1 - Stemformatics Dataset Processing Pipeline* for dataset processing work flow.

## 2.3 Biosample replicate handling

Most studies do not include technical replicate RNA extract samples, but for those that do, we ensure these samples are grouped together (expression levels averaged) and appear in Stemformatics as a single biological sample, which may then form part of a biological replicate sample group.

These biological replicates, of which every Stemformatics study includes, are treated as individual samples except in gene expression graphing where they are aggregated by sample type - we calculate the average expression across a set of biological replicates and provide error bars and standard deviations for these expression values on our histograms (bar plots) and box plots.

We allow the user however to export all underlying values used in generating our graphs and plots.

# 3 Microarray Probeset Mapping

The majority of probe set mappings are created in-house using the same mapping approach as Ensembl.

Microarray oligonucleotide probe sequences are aligned to the Human and Mouse transcriptome (Ensembl Transcript annotated Human and Mouse genome models).

In the current release (Stemformatics v6.5.1, March 2017) we have performed probe set alignments for the majority of platforms against the GRCh37 and NCBIM37 genome assemblies (cDNA and ncRNA sequences) for Human and Mouse, respectively.

## *3.1 Sequence alignment tool and basic parameters*

We use *exonerate (* `http://www.ebi.ac.uk/~guy/exonerate/` *)* to perform ungapped alignments of probe oligo sequences to transcript-annotated genome FASTA files for assemblies described previously, for Ilumina and Affymetrix GeneChip platforms.

For both Illumina and Affymetrix we allow a 4% mismatch threshold; 2 in 50 base-pair mismatch and 1 in 25 base-pair mismatch for Illumina and Affymetrix probes respectively. Ensembl allow a 2% mismatch threshold.

We find that we are able to map more probe sets to the genome with an acceptable level of mis-hits – as we provide individual probe set intensities and multi probe set mapping information, the user is able to identify and determine the robustness of a given probe set-to-gene association.

## *3.2 Transcript clusters and gene mapping*

The Ensembl gene and transcript annotation model provides the mapping between probe set-mapped transcripts and genes.  As the sequence alignment results are summarized to probe sets at the transcript cluster level, it is simply a matter of transforming the probe set-to-transcript mappings to probe set-to-gene mappings.

We store an optimised probe set-to-gene mapping table for Stemformatics operations, bypassing the transcript layer in most cases.

Probe set to transcript mappings are however utilised in the Gene List Annotation feature of the Stemformatics Analyses workbench.

## *3.3 Control probe handling*

Control probes are removed from our mappings if present (whether created in-house or sourced from Ensembl).

We found that control probes were interfering with many gene expression displays and analyses (e.g. an abundance of cross-hybridizing Alu elements), creating “noisy” false-positive outputs.

For Affymetrix arrays, “AFFX” probes (spike in controls, background probes etc) have been removed, similarly for Illumina arrays, we remove annotated Control Sequences.

# 4 External annotations and mappings

## 4.1 Ensembl gene annotation

Stemformatics centralises around the Ensembl gene annotation. All other gene identifiers e.g. gene symbols, RefSeq IDs as provided in Gene Search or in user-uploaded gene sets are converted to Ensembl identifiers for further use.

Currently in Stemformatics release v6.5.1 (March 2017) we incorporate the genome annotations as provided by Ensembl version 67 and 69 (Mouse and Human respectively) – i.e. gene names, descriptions, genomic coordinates and external annotation mappings are as provided by Ensembl.

If a user's gene of interest is not found within Stemformatics, a common cause is use of unsupported gene ID or nomenclature. Where possible, Ensembl identifiers provide the most exact match to a gene of interest.

## 4.2 Probe mappings

Whilst the majority of our microarray probe mappings are performed in-house using Ensembl's probe mapping pipeline (see section 3. *Microarray Probeset Mapping*) we do currently source some mappings directly from Ensembl Biomart; namely for Affymetrix Gene ST and Exon ST arrays (the only current exception being the Affymetrix MoGene-2_0-st array which was mapped in-house).

## 4.3 KEGG pathways

We have incorporated KEGG (*Kyoto Encyclopedia of Genes and Genomes*, http://www.genome.jp/kegg/) pathway information for both Human and Mouse species, via the “KEGG.db” (version 2.5, 23-03-2011) Bioconductor library.

These KEGG pathways are annotated using Entrez gene identifiers. We have converted these to Ensembl identifiers via Ensembl to Entrez mappings sourced from Ensembl.

Not all Entrez identifiers have a corresponding Stemformatics Ensembl identifier. Approximately 5% of the mouse Entrez identifiers could not be converted to Ensembl and 2% of the human Entrez identifiers could not be converted to Ensembl.

# 5 Statistical Tests and Methods

This section describes in more detail, the statistical tests and methods utilised by Stemformatics for analyses and Gene Set Annotation.

## 5.1 *Fisher Exact Test (Gene Set Annotation)*

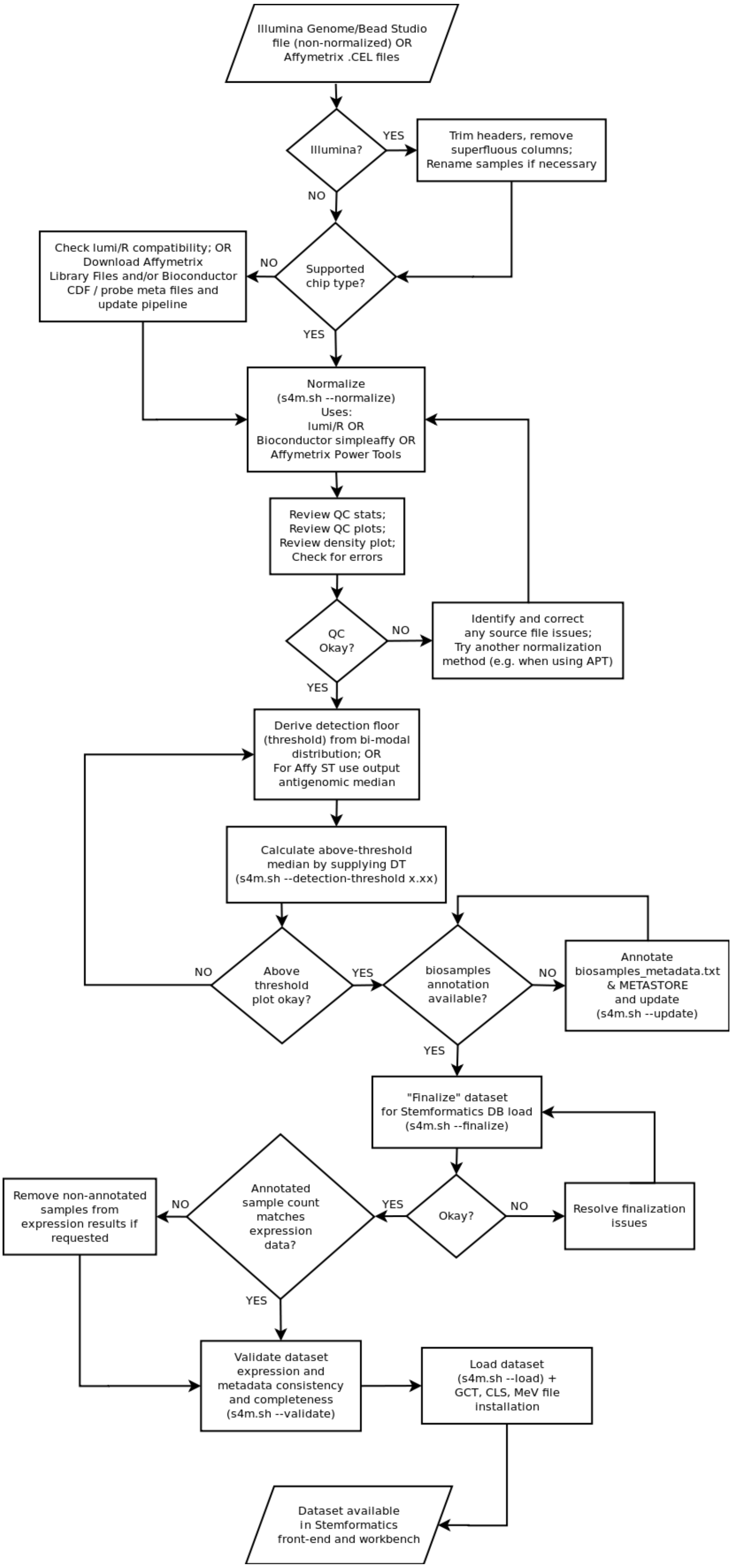
The numbers for the calculation of the 2x2 Fisher Exact Test in the Gene Set Annotation are shown below, based on an example from DAVID ( `http://david.abcc.ncifcrf.gov/` ) of human genome backround of 30,000 genes, and 40 genes involved in an example pathway. A given gene set has found 3 out of 300 belong to this pathway.

The p-value for this example is 8.85E-3. Please note that the right tail is used for display and the lower the p-value the more likely this gene set is enriched in this pathway.

| | *In the Gene Set* | *In the genome* |
|---|---|---|
| In the KEGG Pathway | # genes in the gene set that were found in the KEGG pathway | # genes in the KEGG pathway |
| Outside the KEGG Pathway | # genes in the gene set minus the # of genes found in the KEGG pathway | # genes in the human genome that are not found in the KEGG pathway |

| | *In the Gene Set* | *In the genome* |
|---|---|---|
| In the KEGG Pathway | 3 | 40 |
| Outside the KEGG Pathway | 297 | 29960 |

## **Appendix 1** – Stemformatics Microarray Processing Pipeline


Illumina Genome/Bead Studio file (non-normalized) OR Affymetrix .CEL files
Illumina?
YES
Trim headers, remove superfluous columns; Rename samples if necessary
NO
Supported chip type?
NO
Check lumi/R compatibility; OR Download Affymetrix Library Files and/or Bioconductor CDF / probe meta files and update pipeline
YES
Normalize (s4m.sh --normalize) Uses: lumi/R OR Bioconductor simpleaffy OR Affymetrix Power Tools
Review QC stats; Review QC plots; Review density plot; Check for errors
QC Okay?
NO
Identify and correct any source file issues; Try another normalization method (e.g. when using APT)
YES
Derive detection floor (threshold) from bi-modal distribution; OR For Affy ST use output antigenomic median
Calculate above-threshold median by supplying DT (s4m.sh --detection-threshold x.xx)
NO
Above threshold plot okay?
YES
biosamples annotation available?
NO
Annotate biosamples_metadata.txt & METASTORE and update (s4m.sh --update)
YES
"Finalize" dataset for Stemformatics DB load (s4m.sh --finalize)
Remove non-annotated samples from expression results if requested
NO
Annotated sample count matches expression data?
YES
Okay?
NO
Resolve finalization issues
YES
Validate dataset expression and metadata consistency and completeness (s4m.sh --validate)
Load dataset (s4m.sh --load) + GCT, CLS, MeV file installation
Dataset available in Stemformatics front-end and workbench