# ImmGen microarray gene expression data: Data Generation and Quality Control pipeline.

*Jeff Ericson*, Scott Davis*, Jon Lesh#, Melissa Howard#, Diane Mathis and Christophe Benoist*

Division of Immunology, Dept. of Microbiology and Immunobiology
Harvard Medical School, 77 Ave Louis Pasteur, Boston, MA, 02445
#Expression Analysis, 4324 S. Alston Avenue, Durham, NC. 27713

*These authors contributed equally

Table of Content

I. Data Generation
II. Data Preprocessing
III. Quality Control on ImmGen Microarray data
IV. Interpreting microarray expression values in ImmGen microarray data

Much of the gene expression profiles generated so far by the ImmGen project were obtained using hybridization to microarray supports. Preliminary tests and results supporting the choice of the Affymetrix 1.0 ST platform have been reported[1]. Here, we will document the data generation pipeline adopted for the project, the quality control (QC) steps and the criteria for data exclusion, as well as some of the basic analysis guidelines used to analyze expression across ImmGen data. The criteria evolved with experience and with the profiling of substantial numbers of cell-types, and those described are those adopted as of Feb 2012. Data and QC metrics for all ImmGen datasets as of this date are listed in Table 1.

## I. Data Generation

Per ImmGen SOP, the final cytometric sorts (typically from 10,000 to 30,000 cells, although 10% of attempted samples were below 10,000 cells) were performed directly in Trizol (no more than $5.10^4$ cells in 500 ul Trizol), frozen after 2 minutes, and sent to the ImmGen core team in Boston. RNA was prepared from the Trizol lysate by Chloroform extraction and isopropanol precipitation in the presence of Glycoblue carrier. The pellet was washed with ethanol, air-dried, and finally taken up in 12.5 µl dH2O. Early pilots showed that RNA quantitation by micro-spectrofluorimetry in preparations from the low cell numbers typical in ImmGen was rather unreliable, and not a valuable predictor of labeling efficiency and microarray data quality. Thus, each RNA sample were used in its entirety, without prior quantitation, decisions on whether to hybridize to the microarray made on the basis of the amount of ssDNA probe obtained.

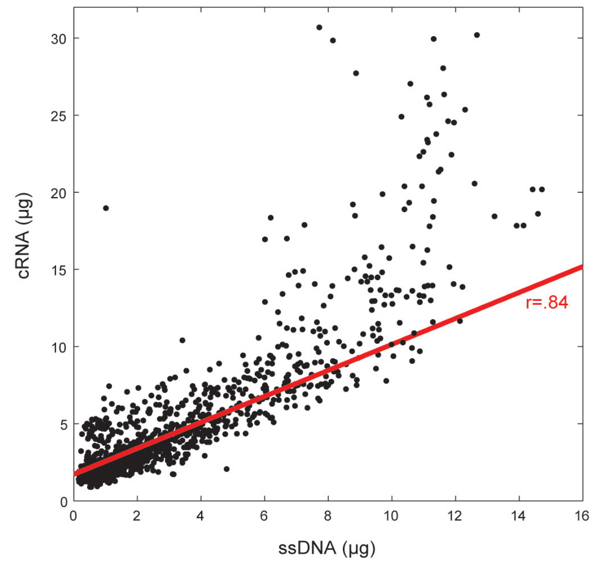

**Fig. 1:** Relationship between the yield of intermediate cRNA and labeled ssDNA probe, for all ImmGen samples.

Probe amplification and labeling and hybridizations were performed, for essentially all samples, at Expression Analysis, Durham, NC. The starting total RNA was converted and amplified into antisense cRNA, and then was converted into ssDNA, which was fragmented and labeled with Biotin before being hybridized to Affymetrix Mouse Gene ST 1.0 microarrays.

RNA was first converted to sense-strand cDNA using the Ambion WT Expression Kit, in a reverse-transcription reaction with primers designed using a proprietary oligonucleotide-matching algorithm to avoid rRNA binding, thereby providing comprehensive coverage of the transcriptome while significantly reducing coverage of rRNA. This method also avoids the 3’ bias inherent in methods that prime exclusively with oligo-dT-based primers. During processing, the concentrations of the intermediate cRNA and sense-strand cDNA samples were evaluated using a NanoDrop micro-spectrophotometer. Up to 10ug of cRNA was used for sense-strand cDNA synthesis. Subsequently, 0.7 to 2.75ug of the resulting sense-strand cDNA was fragmented and labeled using uracil-DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE1). APE1 recognizes and fragments the cDNA at dUTP residues, which were incorporated during the 2nd-cycle. Finally, ssDNA was labeled by terminal deoxynucleotidyl tranferase (TdT) using the Affymetrix DNA Labeling Reagent.

There was strong correlation between yields of cRNA and ssDNA (Fig. 1), discounting cRNA levels above 10 μg, since only 10 μg were used for ssDNA synthesis.

The yield of both the cRNA and ssDNA intermediates proved to be reliable metrics, and good predictors of data quality. Since ssDNA was the material that was actually hybridized to the microarray, ssDNA yield was taken as the benchmark to decide whether or not to hybridize each preparation. Fig.2 shows the relationship between the ssDNA yield and the final data quality, assessed here by the signal’s dynamic range (DR); as will be discussed below, samples with a dynamic range >60 were considered to be of very good quality, those with DR between 40 and 60 acceptable. After the first batches, samples yielding less than 0.7 μg ssDNA were not used for microarray hybridization, unless stemming from rare cell-types (and only a small minority, <50%, eventually passed the later QC steps). As might be predicted, these low yields corresponded to rare populations which were the most difficult to sort.

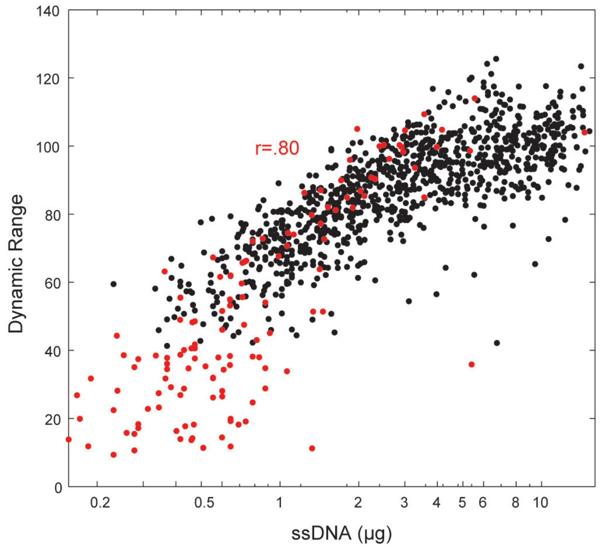

**Fig. 2:** Relationship between the yield of labeled ssDNA probe and the dynamic range of the resulting ST1.0 data, across all ImmGen samples. Red dots are those that eventually failed QC, some because of poor dynamic range, others for contamination or other issues.

All ImmGen data posted on the web server and deposited in the GEO database were generated in 37 independent batches. To monitor batch effects, each batch included a pair of common RNA samples, from whole CD4+ and CD19+ splenocytes. These consistently passed the ssDNA threshold.

Hybridization cocktail was prepared using the Hybridization, Wash and Stain kit (Affymetrix), applied to Mouse Gene ST 1.0 arrays, and incubated at 45°C for 16 hours. Following hybridization, arrays were washed and stained with fluorescent streptavidin using standard Affymetrix procedures before scanning on the Affymetrix GeneChip Scanner and data extraction using the Affymetrix Expression Console. A primary array QC metric at this step was Positive versus Negative AUC (area under the curve), which is akin to Signal to Noise – it relates to the ability to distinguish true signal from noise. A Pos. vs. Neg. AUC value of $\geq$ 0.8 passed array QC metrics at Expression Analysis.

## II. Data Preprocessing

Microarrays were scanned in the Affymetrix GeneChip Scanner 3000 7G instrument, and the image data processed to generate primary .cel files, which were then used for normalization and for QC analysis. Quality control statistics were generated for each sample, and the samples that passed (see section III below) were then included in a new normalization run together with all previous ImmGen samples that had previously passed quality control.

### I. Preprocessing and Normalization

Each new batch of .cel files returned from Expression Analysis to the Boston core team was normalized, using Affymetrix Power Tools' "apt-probeset-summarize" executable with the rma-sketch standard method. Robust Multichip Average (RMA, ref2) was the algorithm used for feature-level normalization, the *de facto* standard for Affymetrix microarrays.

Once every three months, all ImmGen samples (including samples that had previously failed QC) were re-normalized together and analyzed with quality control metrics for verification. These regular releases were passed on to the ImmGen group for biological curation and analysis, and were also the source of the data loaded onto ImmGen's public data browsers. We analyzed the possible drift in ImmGen data over time, resulting from the inclusion of additional datasets. A simulation experiment was performed in which the effect of adding additional batches of 30 samples was tested, reading out the reported expression value (normalized to the value obtained with 750 samples). As shown in Fig. 3, the effect of additional batches of samples was very small, in comparison to the range of expression values found across ImmGen data, particularly after the first few hundred samples.

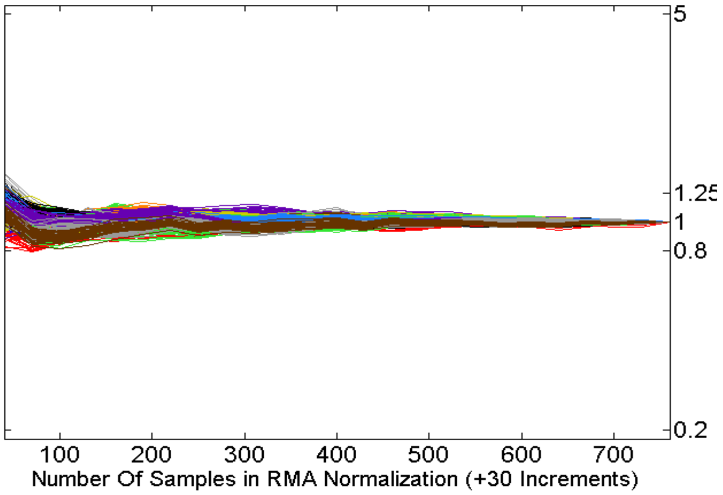

**Fig. 3: Drift in expression values from inclusion of additional datasets**. To evaluate the impact of including additional batches of data into the ImmGen datagroup, the post-normalization expression values for a set of 20 representative genes were recorded (in 40 populations) after iterative inclusion of additional batches to 30 datasets picked at random. The lines represent the observed values (standardized to the value observed when including all 750 datasets), and are color-coded by gene (each line representing a different population). The Y-axis on the display is scaled to the 0.2-5 range of differential expression commonly observed across ImmGen populations. The drift in post-normalization values is minimal, and reduced even further beyond 350 samples; note also that the slight drift in post-normalization intensities coordinately affect all populations, such that the drift in fold change between populations for the same gene would be even smaller.

### 2. Removal of erroneous or misleading probes

The Affymetrix Mouse Gene ST1.0 Array has over 750,000 distinct 25-mer probes ("features"). Each feature targets a specific exon of the target genes, with good coverage throughout, allowing analysis of exon-level transcription. For normal use, the features that correspond to a gene are consolidated into one "probesets" (where each probeset corresponds to a gene or locus, for 35,518 probesets on the MouseGene ST1.0 array). Each probeset summarizes values from 24-40 features, with considerably less noisy data. Most genes are represented by a single probeset on the ST1.0 array.

A number of features or probesets present in the primary data were removed prior to normalization. These probes represent glitches in the processing, are uninformative or give suspect of erroneous data, and were thus removed from the data releases.

*Un-annotated or Duplicate-Read probes*: The CSV annotation file for the ST1.0 arrays was downloaded from the Affymetrix website. Probesets whose Gene Symbol assignment was listed as "---" were removed, as these correspond to array controls or to intergenic sequences on unknown significance, which often exhibit very high inter-replicate noise. We also realized that the Affymetrix processing algorithm generates results for "fantom" probesets that are merely multiple replicates of the same probeset (e.g. for Snord115 probesets 10564013 and 10564017), with expression values identical to the 9th decimal. Including multiple copies of these probesets would distort pre-processing or later cluster analyses. Thus, only one example of these probesets was retained.

*Remove features with spuriously high values:* While the majority of features yielded signals that correlated with the probeset as a whole, a minority of features yielded very high signals, which were uniformly high and inconsistent with the expression of the probeset or with other features mapping to the same exon. While they might possibly correspond to unknown RNA species, these aberrant features were considered irrelevant to the gene as a whole and most likely due to spurious cross-hybridization. They were removed since they might erroneously suggest expression in cell-types in which a given gene was otherwise silent (Fig. 4). A total of 2372 features (1 per gene at most) were removed from consideration prior to RMA normalization, if the following criteria were met: feature with high expression (>10.0 after log2

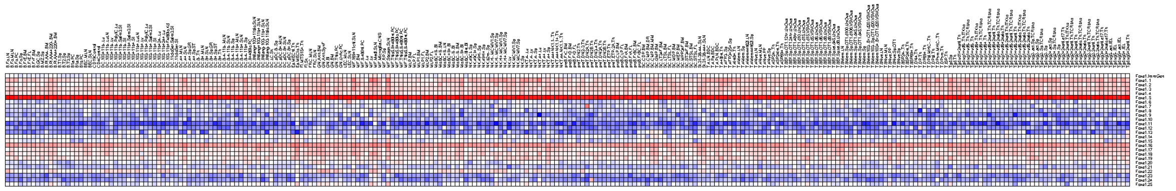

**Fig. 4: Example of cross-hybridizing feature**. Each row of this heatmap representation depicts the expression of a single feature of Foxe1, for all ImmGen populations (in columns). The 6th feature from the top show uniformly high expression, without the variation seen in all other features of the gene, and correspondingly did not correlate well with the entire gene's expression.

transformation), and with low max/min range across all ImmGen data (<=3-fold change after log2 transformation), and with low correlation to its probeset in comparison to other features of the same probeset (Pearson coefficient < .4 and in bottom 0.2 quantile of features within its probeset).

*Remove probes that are discordant with RNA-seq results: The* same CD4 and CD19 controls included in each batch of microarray processing were also analyzed by deep RNA sequencing (RNA-seq; ~2x10$^8$ Illumina paired end reads; details reported elsewhere, Doran et al, in preparation). Comparison of the RNAseq results with the microarray data (Fig. 5) showed generally good agreement with the microarray results, as many of the genes scoring as expressed by RNA-seq (>1 FPKM - fragments per kilobase of exon per million fragments mapped) were also positive by microarray (> standard ImmGen threshold of 7, after log2 transformation). A fraction (18%, 1599/9031) of the transcripts detected by RNA-seq scored below threshold on the microarray (log2 expression <7) or were simply absent from the microarray (7%, 666/9031). More a problem were transcripts that scored robustly on the microarray but were essentially not detected in either of the RNA-seq runs (red dots on Fig. 5). These were considered to result from cross-hybridization on the microarray, a hypothesis supported by the nature of the transcripts, many of which corresponded to very homologous multigene families (e.g. *Hist1*). These transcripts were removed from the released datagroups.

Overall, uninformative or questionable probesets and features have been removed as described from the data presented on ImmGen data browsers and smartphone supports, and from custom data supplied on request through the ImmGen site. On the other hand, all probesets and features are retained in the raw .cel files available from the NCBI GEO database (accession # GSE15907).

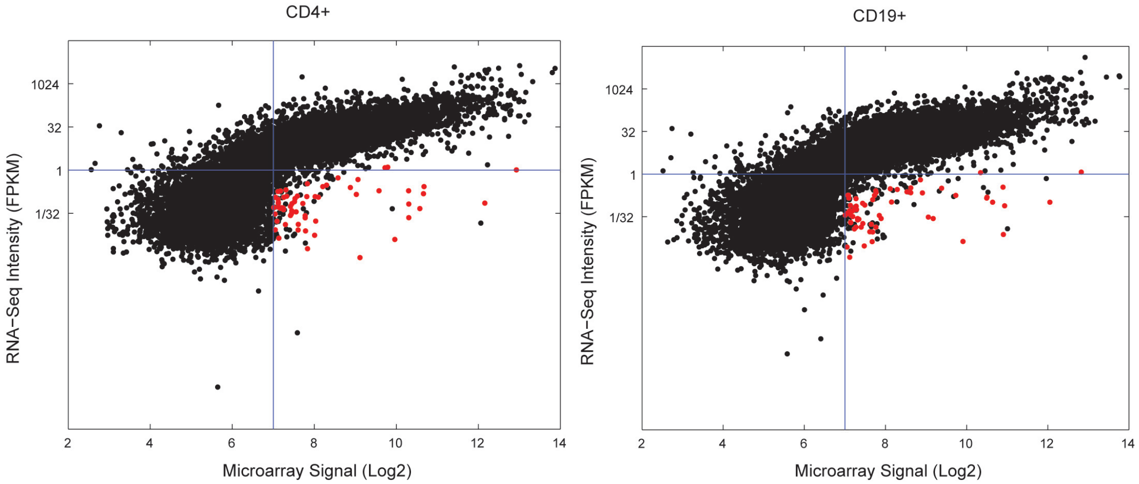

**Fig. 5: Comparison of RNA-seq and microarray-based profiles.** RNAs from the same bulk CD4+ T or CD19+ B splenocytes were profiled on ST1.0 microarrays as usual, or by deep RNA-seqencing on the Illunina platform (Doran G., in preparation). Sequence reads were mapped to individual genes represented on the microarray. Blue lines denote the expression thresholds with either technique; Red dots are genes that gave suspiciously high signals on the microarray and were removed from the public datagroups.

## III. Quality Control on ImmGen Microarray data

Several QC steps were implemented during the pre-processing of ImmGen microarray data. These were used to remove datasets that were of low intrinsic quality, showed indications of likely to resolve discrepancies. Note that these QC standards evolved over the course of the project, and continue to be refined, such that the public data may slightly evolve over time.

### I. Dynamic Range

After experience, the dynamic range (DR, or the ratio between the highest and lowest signal values in a single dataset), became the primary metric of quality for individual expression profiles. To avoid confounding by single outliers, the DR was calculated for each dataset by dividing the 95th by the 5th percentile of expression values. In practice, the dynamic range was calculated after pre-processing, but this value was highly correlated with dynamic range in the raw data (Fig. 6).

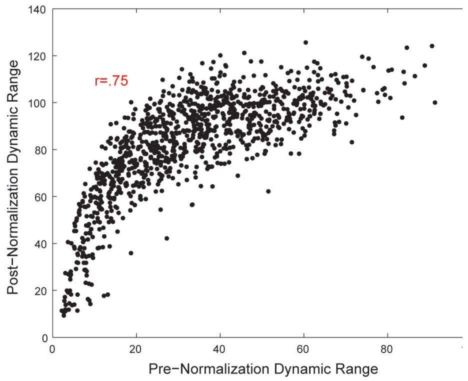

**Fig. 6:** Dynamic range of all ImmGen samples, before or after data normalization

Low dynamic range denoted low signal intensity on the chip, and generally corresponded to samples from lower cell numbers, and thus limited amount of ssDNA probe (Fig. 2). The corrections introduced during the initial steps of data processing and of normalization resulted, for poor data, in an amplification of the signal intensities of un-expressed and negative control probes, as illustrated in Fig. 7 for two samples of different dynamic ranges. Indeed, the position of the main peak (calculated as the "primary mode" of a Gaussian decomposition) was another measure of data quality, highly correlated to the DR. The expression/expression plots of Fig. 8 show that with two well-amplified samples,

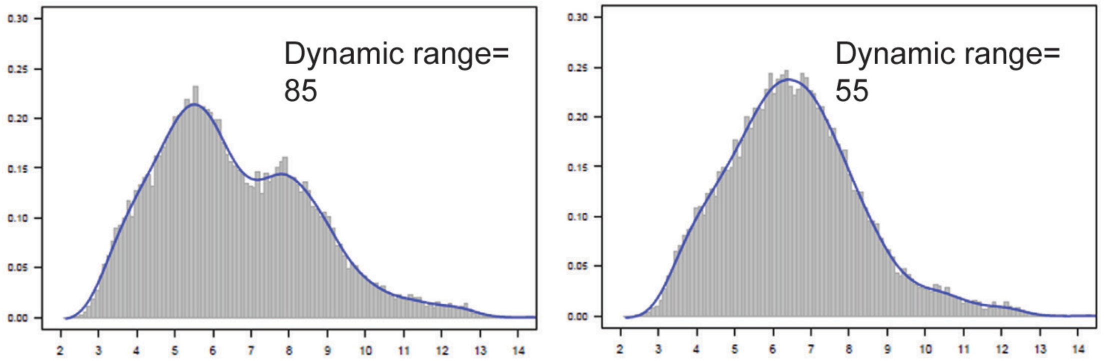

**Fig. 7:** Post-normalization expression histograms of two representative datasets, one with good dynamic range (left, DR=85) or borderline dynamic range (right, DR=55). X-axis: expression value (log2-transformed). Note how negative transcripts are clearly demarcated on the left profile, but creep up towards the positive range in the lower-quality dataset at right.

there was symmetry along the x=y axis, but that comparing samples with very different DR values led to a distortion in the comparison (for illustration only, an unacceptably bad sample is shown here).

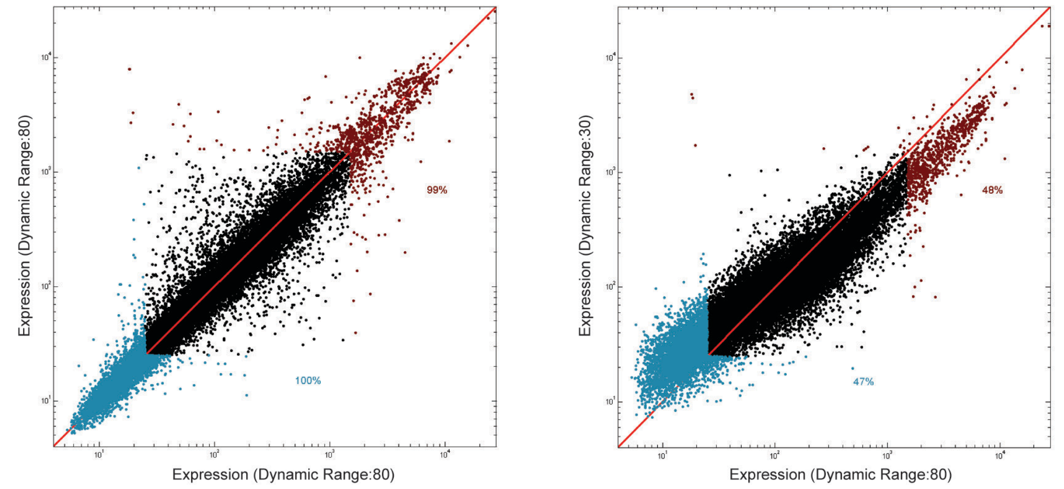

**Fig. 8:** Expression/expression plots comparing samples with good (80) or bad (30) DR scores.

In practice, all samples with dynamic range below 40 were systematically dropped. Samples with DR scores between 40 and 60 were closely inspected for other metrics (in particular, examining the match with other replicates for the same cell-type), and were retained if they originated from difficult-to-sort populations and/or did not generate high intra-population variation (see section III.3). The overall distribution of DR metric for all ImmGen samples is shown in Fig. 9; the median DR score for retained samples is of 86.8, and only 7.5% of retained datasets had a DR<60.

We assessed whether and how the inclusion of datasets with somewhat different dynamic range generated false-positives in the identification of variably expressed genes across the

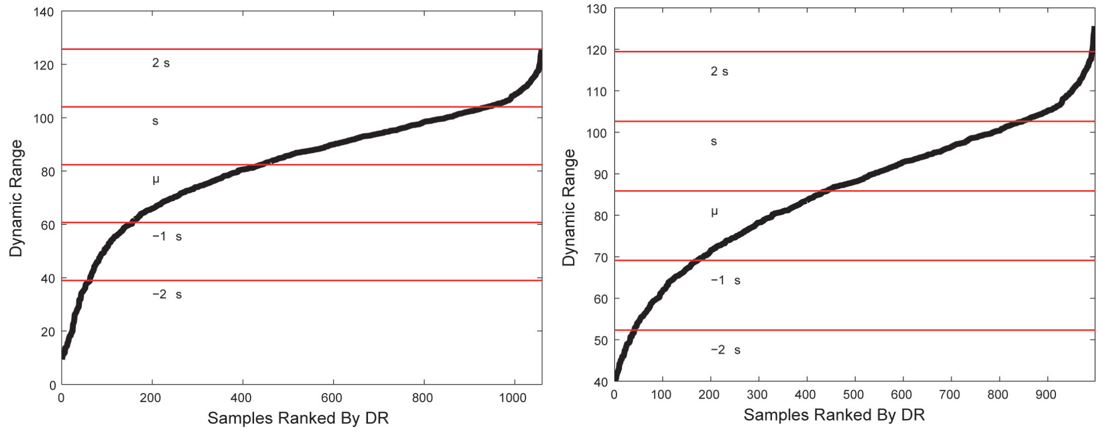

**Fig. 9:** Distribution of Dynamic Range scores across all ImmGen samples. The plot at left depicts all acquired samples, prior to QC, the plot at right depicts the DR scores of retained samples (mean DR score of 86.8). Some data with a DR score below 60 were retained in the ImmGen datagroup, but are flagged on the web display, in particular because of their higher values for non-expressed genes.

ImmGen datagroup, an element of key importance in a compendium of this nature. In Fig. 10, we plotted the number of genes varying by more than 3-fold in a pairwise comparison; these were measured in pairs of replicate datasets from a given cell-type, where one dataset had a dynamic range between 70 and 90 and the other varied across the entire range of DR values. As expected, very poor quality datasets (DR<40) gave rise to very high numbers of false-positives (500 to 10,000), but datasets with DR between 60 and 120 yielded a roughly even proportion of false-positives (0.05, 0.95 quantiles 1 and 264 genes, median 14 genes). This range was well distinguished from similar inter-population comparisons (0.05, 0.95 quantiles 337 and 3021, median 1439).

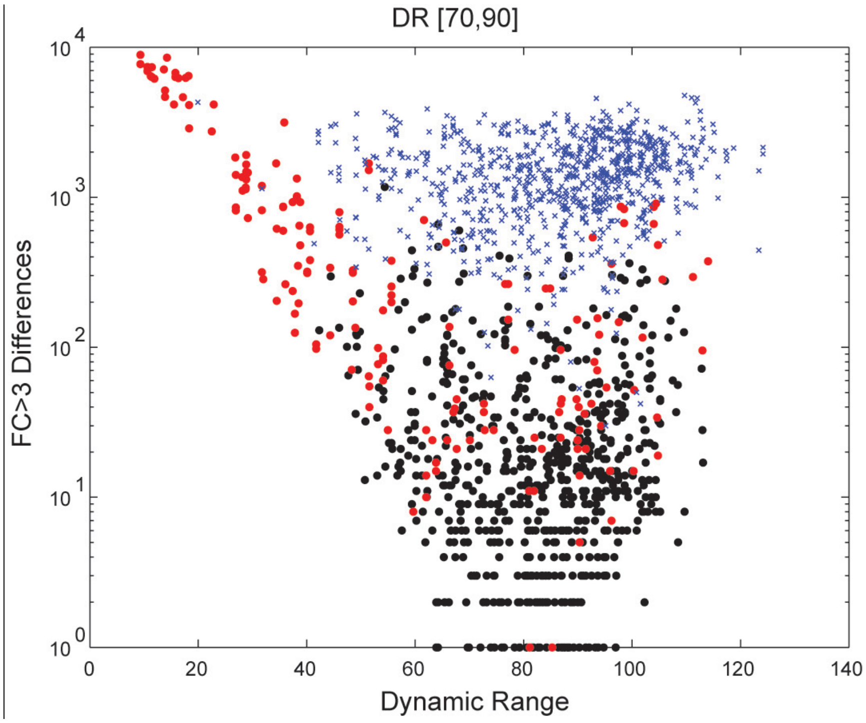

**Fig. 10:** Influence of the difference in dynamic range in a pairwise comparison between two individual ImmGen datasets and the false-positive rate. Pairwise comparisons where one population had a DR of 70-90, and the other any DR value along the range, counting the number of transcripts with a FoldChange >3. Black and red dots are measures of: false-positives, from comparisons of replicates within individual populations (red dots are for those failing QC metrics). Blue dots are a reference from similar pairwise comparisons between populations.

## 2. ImmGen-Wide Correlation

Another important flag for data QC was a determination of the correlation of new datasets to the ImmGen-wide data. Since ImmGen encompasses a wide variety of different immune system cells, there was no expectation of uniform results. But a low level of correlation with any other dataset was usually a flag for suspicious data (with the exception of very different cell types like stromal cells).

A correlation matrix was drawn for each dataset using the "ImmGenQC" GenePattern plug-in, facilitating the identification of outlier datasets (as a "blue streak" in Fig. 11). Samples that did not have an ImmGen-wide CC above .97 with any other sample were flagged as suspect samples, except for those cell-types that naturally showed little correlation to other ImmGen populations, yet strong intra-group correlation (e.g. neutrophils or stromal cells). Maximum correlation coefficients for all ImmGen populations are shown in Table 1.

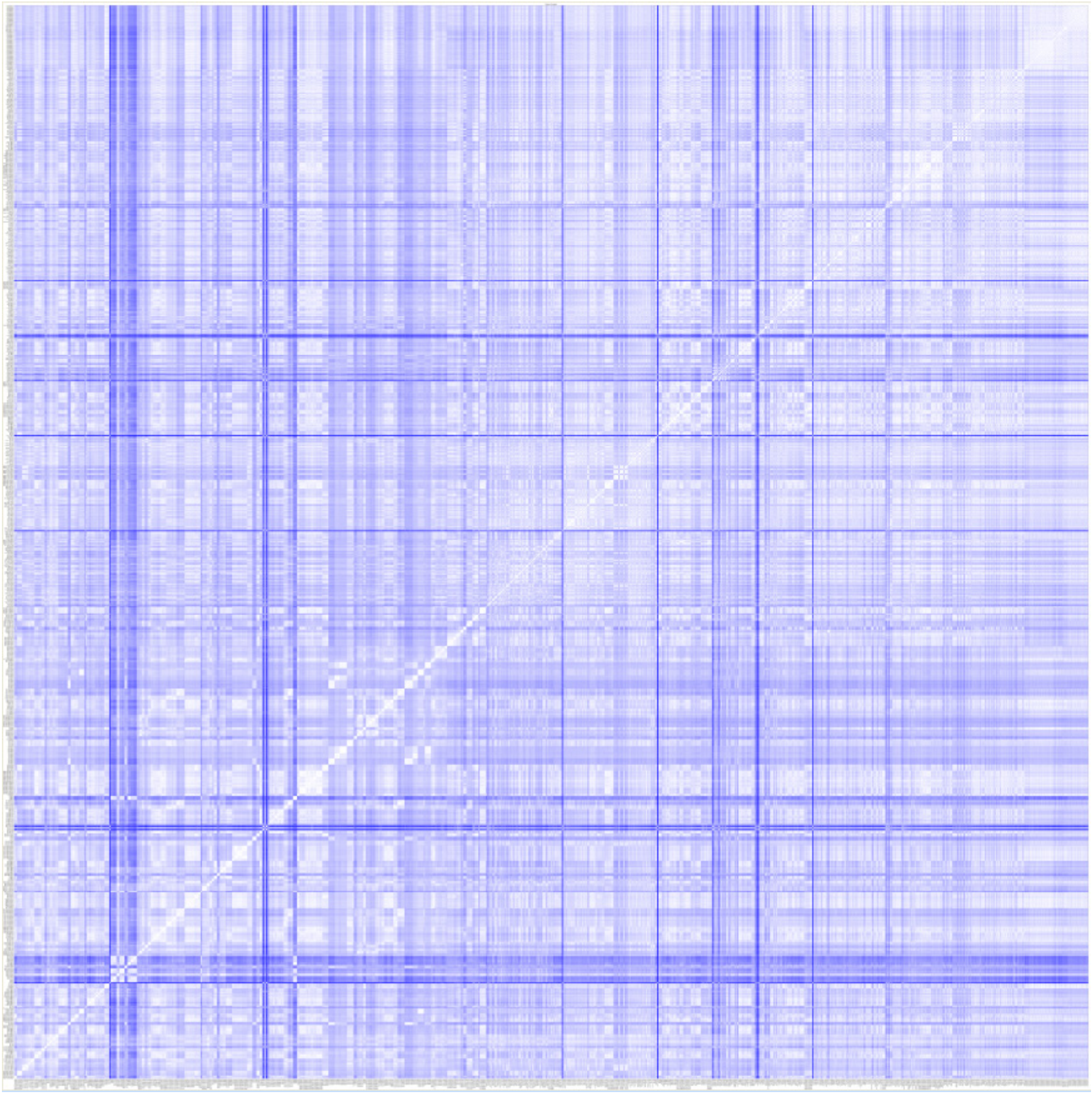

**Fig. 11:** Correlation matrix across all hematopoietic cell-types among ImmGen samples (bad data included). The isolated blue streak denote poor quality samples, although the grouped streak (~samples 50-65) represent neutrophils.

## 3. Population Coefficient of Variation

The ImmGen standard was to profile three biological replicates per cell-type examined. With these triplicates, a coefficient of variation (CV; computed as the Standard Deviation divided by the Mean of triplicate data) was determined for each gene on the microarray and for each cett-type. As previously discussed[1], the distribution of these gene-wise CVs was a good indication of the aggregate data quality for a given population. Any population with a median CV above 0.20 (such as the example shown in Fig. 12, left panel) was examined to determine why there was so much variation between biological replicates. The usual reason was that one of the replicates either had a poor or borderline DR score. This replicate was then removed, based on consideration of other data metrics (on the aggregate 14 datasets were removed for poor concordance and intrinsically borderline quality). Post-QC, the median intra-population CV ranged from .056 to .185 (median .102) across retained ImmGen data.

In some cases, more than three replicates were profiled for one cell type, all of the replicates passed the DR and ImmGen-wide CC metric for individual datasets, and all replicates were classified as quality samples. However, in some cases, a few replicates were on the lower end of the ImmGen QC assessment, and by dropping them and re-classifying them as poor quality samples, the population as a whole had a lower median CV value (dropping below 0.20); in such cases, the three best replicates were kept.

In some cases, triplicates were profiled and initially passed QC standards, but were later re-classified as poor quality samples, with evolving standards. Such populations were not necessarily re-profiled, and thus the population does not have three quality replicates. In most cases there were at least two replicates; in 5 cases, there was only one quality replicate (which should be treated with caution).

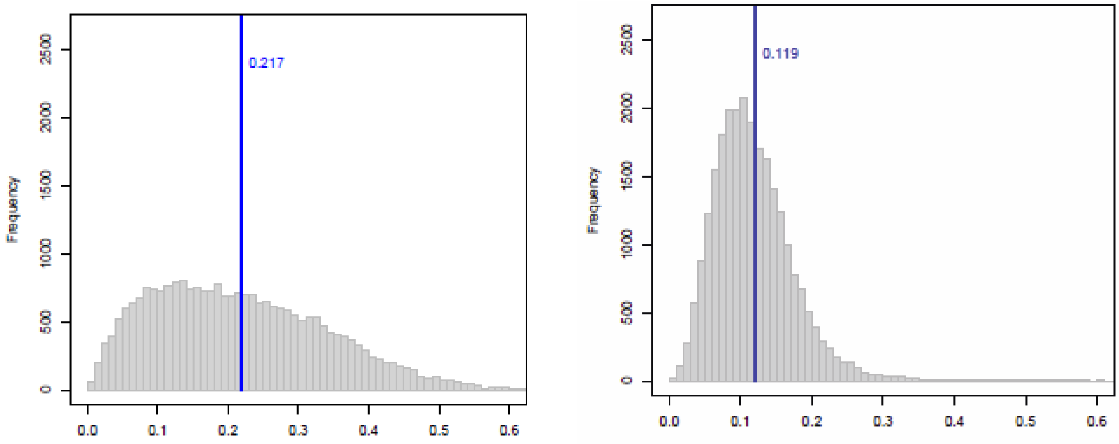

**Fig. 12 : Distribution of gene-wise inter-replicate Coefficients of Variation for individual cell-types.** The example shown at right has a good CV distribution, with a median between 0.1 and 0.15, while the example at left shows a far greater proportion of noisy data, with median CV at 0.21.

## 4. Detection of contaminated populations

The quality control metrics described above deal with quality of the expression datasets. In addition, the datasets were searched for signs of contamination by analyzing transcripts expressed at high levels specifically in particular populations. This search was straightforward for some cell-types for which highly specific transcripts could be defined, such as immunoglobulins for B cells or hemoglobins for erythrocytes, but not as clearcut for others (in particular for contamination with myeloid cells. The transcripts included:

For T lymphocytes, *Lat , Fyb, Thy1, Tcf7, Cd3g*; for B lymphocytes, *Igh-6, Ms4a, Igj, Cd79b, Pax5, Igk*; for myeloid cells, *Tlr13, Anxa3, IL13ra1, Alox5ap*. For red blood cells, *Hba-a1, Hba-a2*.

In addition, *Hspa8* was used to denote possible stress in the sorted cell preparation.

Some level of contamination with erythrocyte RNA is unavoidable for some cell preparations (understandably, those from blood or bone marrow), even with stringent exclusion at the sort stage. The extremely high levels of hemoglobin mRNA in red cells (30% of total mRNA) ensures that traces contamination will lead to visible signals for Hb genes. In practice, datasets showing traces of Hb signals were not removed from the public data groups, but expression of the following genes should be treated with caution: Hba, Hba, Hbb, Gm5226, Alas2, Gypa, Epb4.2, Slc4a1.

ImmGen data were generated from male mice to allow coverage of *ChrY*-encoded genes, with the exception of samples from fetal cells for which both sexes were pooled. Expression of *Xist* was used to identify gender errors (the cell preparation was repeated in such cases).

## 5. Batch effects.

Batch effects are an important source of confounders in gene expression profiling. ImmGen data were acquired over a 3-year period in 47 different batches (as of June 2011). Batch analysis with tools commonly used to extract batch-specific variation (e.g. PCA) was difficult to apply in this instance, as different batches were usually composed of different cell-types over time. Analysis of the constant samples from CD4+ and CD19+ controls included in most batches batch served as an indicator of serious defects in processing of individual batches. As illustrated in Fig. 13, no such effects were observed.

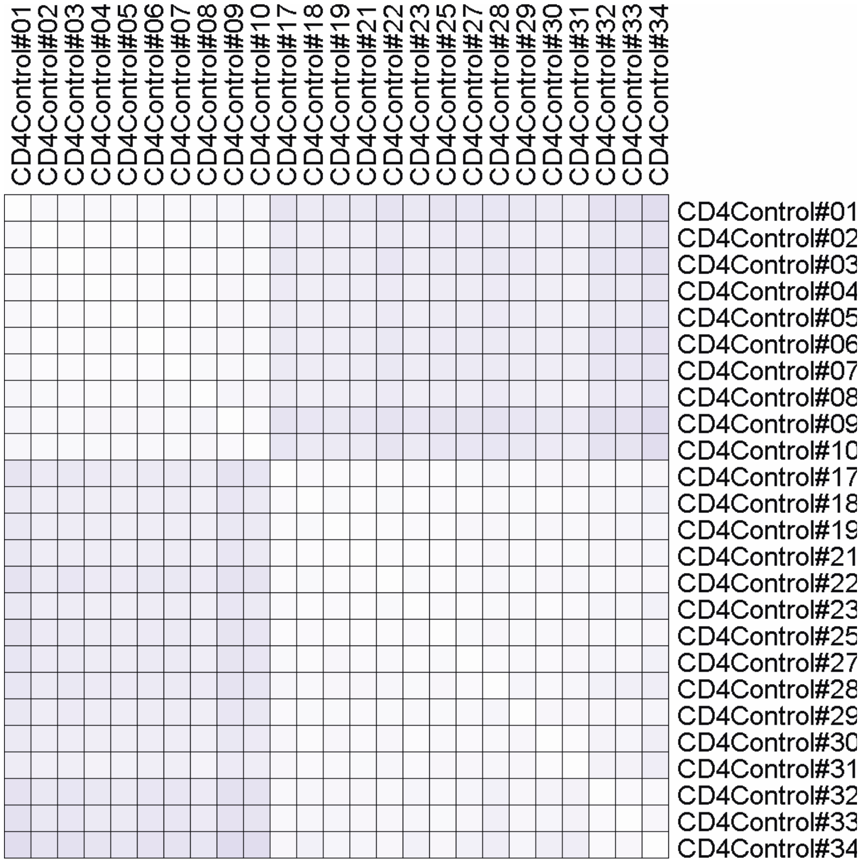

**Fig. 13:** Correlation matrix across the constant CD4 control samples included with each batch as technical replicates. Three RNA preparations were used over the course of the program (batches numbered here 1-10, 17-31, 32-34). Although there are subtle differences between the RNA samples themselves, no individual batch particularly stands out. Correlation coefficients between the same RNA samples in different batches ranged from .978 to .998, which compares favorably to inter-population comparisons where correlation coefficient range down to 0.65.

## IV. Interpreting microarray expression values in ImmGen microarray data

An important outcome of expression profiling is to determine whether a gene is actually expressed, or not, in a given cell-type. The question is not as trivial as it sounds, because it is difficult to determine a clear boundary between true expression vs background signals on the microarray; similarly, for RNA-seq, whether rare reads in deep RNA-seq data represent true activity or simple transcriptional background noise. For genes with signals in the borderline or background range, the answer is probabilistic.

For the ImmGen microarray data, the post-normalization expression values range from approximately 10 to 20,000 (presented in a linear scale, considered more intuitively significant to experimental biologists). Within this range, thresholds for likelihood of true gene expression were determined from the distribution of values across the microarrays. Because the ST1.0 arrays do not include reliable negative controls, an empirical Gaussian mixture model was used to evaluate probabilistic thresholds of expression. The expression histograms can be decomposed into two main Gaussian distributions (and further into a more finely-tuned set of four Gaussian distributions) with the gmdistribution.fit function in MATLAB (Fig. 13). It was assumed that the low distribution (blue line) corresponds to background noise, and the high distributions (red and yellow) correspond to true signal. From these, conservative probabilities of expression were calculated on the basis of the low distribution (>120 is a 95% or greater probability of true expression), and conversely the probability that a gene is silent was deduced from the high distribution (<47 is a 95% or greater probability of a silent gene). In the 50 to 120 range, there are intermediate probabilities that a gene is truly expressed (50% probability at the intersection point, between 80 and 90 for most samples.

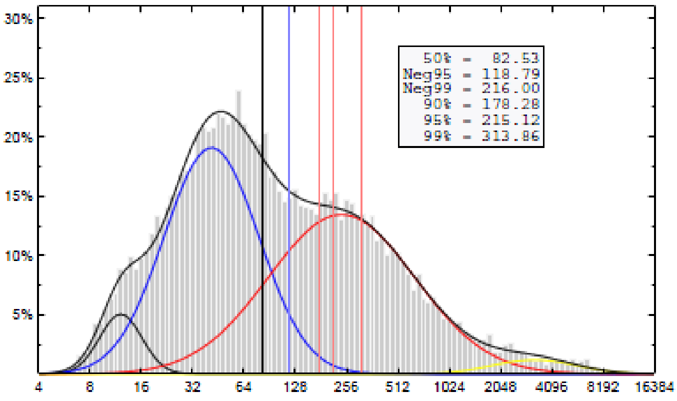

**Fig. 13: Thresholds for positive expression in microarray profiles.** The distribution of expression values was decomposed into the sum of Gaussian distributions. The probabilities of true expression was computed from the negative (blue) distribution.

These empirical determinations were supported by an analysis of the same RNA samples run on four different microarray platforms, some of which included true negative controls[1]. Similarly, detection thresholds from deep RNA-seq corresponded reasonably well to these determinations (the threshold value of 1 FPKM corresponded to ~90 on the microarray – see Fig. 5 above – showing that the 120 value for actual expression is conservative).

While these values apply to the majority of genes, there may be departure for individual genes. For instance, one can observe consistent, and biologically plausible, signals with a clear population-specific pattern in the 20-40 range.

In addition, these thresholds need to be adapted for some datasets of borderline quality, for which the primary mode of the low distribution were artificially raised by the normalization process (see II.1 above). Thus, the ImmGen data QC tables list the threshold for 95% probability of expression, when they depart from the value of 120 that is applied to most datasets.

## REFERENCES

1. Painter, M.W. *et al*. Transcriptomes of the B and T lineages compared by multiplatform microarray profiling. *J Immunol*. **186**, 3047-3057 (2011).

2. Irizarry, R.A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).