**Debey 2006:**

**Summary:**

Big issue with PBMC is instability and degradation during transportation. Suggest use PAXgene to stabilize, then use Globin Depletion to remove globin transcripts from samples which account for majority of PBMC, reducing variance for other genes. PBMC samples were PAXgene stabilized after collection, then comparison was done between Globin depleted samples (GRP) and non globin depleted. GRP samples had higher library coverage (termed present calls) and reduced intra group variance (technical variance). They were also able to construct 100% validation accuracy classifiers on gender with fewer features using GRP samples vs non-GRP samples. Used loocv with features being presumably unit normalized DEG expression. Their idea is that you can also use the fewer samples on GRP for disease state models. Fewer features will probably reduce over fitting. GRP as a feature selection method? No pathway analysis.

**Methods:** They used a different older globin clear method (R Nase H Digestion)

**Metrics:**

Statistical: Present Reads, Technical Variance, nFeatures for classifier (maybe see if their claim about feature selection is true, test on gender, age, race(\*))

Overarching: Reproducibility, Feasibility, Affordability

Copy their rank plot:

A graph of a number of pax

Description automatically generated

**Differences:**

Their change in present calls was less significant than ours. They had fewer percentage of non-globin clear DEGs (1%) compared to ours.

**Shin 2014**

**Summary:**

Assessed the effect of globin depletion on detection of more transcripts and what **pathways** those transcripts are from as well as that sample **quality** was higher for globin depleted samples.

**Differences:** Their RNA yield was lower following globin depletion which is the opposite of our experience. Perhaps this is because they are using an older version of globin depletion.

They also had technical controls since the globin and non-globin were sequenced on different lanes, they mixed their samples together and sequenced the mixed sample multiple times on both the globin and non-globin clear lane to (after applying the globin treatment) to assess wether processing the RNA through globin clear introduced sample variation which they found was minimal in their Q-Q plot showing little bias introduced. They also used spike in control.

**Things already confirmed:**

* Increase number of reads aligned
* Reduce proportion of globin mapped reads (heme\_prop.csv)

**Things to confirm:**

* Redo pathway analysis on GO instead of KEGG to see if we can identify the blood like pathways which were down regulated
* Proportion of heme in globin depleted vs non-globin depleted (heme prop)
* Get from Novogene the exact library prep, globin clear, and sequencing steps
* Redo unique genes with normalized counts in addition to the raw counts both with and without filtering on coding / lncRNA

**Notes**: Since we only have HS globin comparisons, we should also add some HC comparisons to see changes in what the DE genes are for HS between GC and NGC and whether the model produced by GC has greater power.

**Good Things:**

Make a copy of their experimental design graphic

A diagram of a diagram

Description automatically generated

Use their terminology GC (Globin Clear) NGC (Non Globin Clear)

**2020 Harrington:**

Summary: Compared the gene expression of whole blood samples with Ribo Zero Gold (way to get rid of rRNA) (RZG) and Ribo Zero Gold + Globin Clear (RG). Found that most of the high non-hemoglobin DEGs discovered were non-coding and had no relation to patient disease.

**Cool Points:** Had experiment design that first compared the effect of RNA quantity, then library method (RZG / RG). Found the there was little difference between high RNA and low RNA quantity.

* 3 Technical replicates with half factorial block design (RZG/RG \* High/Low RNA \* 6 Samples) / 2 = 36 – 3 low samples = 33 libraries
* Compared the RZG/RG as well as bioinformatic removal of globinclear on RZG and RG.
* Calculated **statistical power on disease state** (we can do this too if we have HC sample comps)
* Used 1 CPM threshold for considering whether a gene was read or not (low count filtering by CPM), *see how many counts that corresponds to*, do for combined

**Todo:**

* Copy plots of volcano with x and y axis showing globin / non globin and protein coding / non protein coding
* Make table comparing whether the new genes found are all low count via histogram of CPM.

**A diagram of a hill with text

Description automatically generated with medium confidence**