## Presentation Document (The GINNS)

* **Slide #1: Title**
* **Slide #2: Index**
* **Slide #3: Systemic Lupus Erythematosus (SLE)** - Icíar
  + SLE is an autoimmune chronic condition that causes inflammation in connective tissues
  + Signs & symptoms vary by individual and can involve many organ systems, but a characteristic feature is this “butterfly rash” across the cheeks and nose.
  + SLE is considered a multifactorial disorder: evidence of genetic susceptibility, but also associated to drug exposures and viral infections (not a single specific trigger)
  + Female-biased like all autoimmune disorders, important to take sex into account as we will see in our analyses
  + Multiple genetic factors have been found to increase the risk of developing SLE & DNA methylation is known to mediate processes relevant to SLE
  + Important to understand these in order to comprehend disease etiology & better treatments
* **Slide #4: The Cohort -** Sierra
  + Our project aims to find probes from these datasets that are significantly differential between normal and disease as well as within disease types (LN- and LN+)
  + We expect to be able to recreate the results of the paper
  + We expect that genes/pathways found are specific to connective tissues, while keeping in mind that inflammation is a common results of enrichment analyses and the disease is specifically involving inflammation
  + We aim to incorporate both types of data into an analysis, or examine how the individual results of both compare to each other
* **Slide #5: Gene expression data (QC & Norm) -** Naila & Sierra
  + Quantile normalization was performed (box plots ). preliminary PCA shows samples don’t group based on disease vs control.
  + After using genefilter(), diagnostic plots were run (e.g. heatmaps of sample correlations, pheatmap with clustering, and simple t-tests). These t-tests show that for SLE-LN- vs SLE-LN+ the null hypothesis that “the p-values are normally distributed” does not hold (right figure, right plot). The results of the tests which require this assumption to hold should not be taken as valid since this assumption does not hold
* **Slide #6: Methylation data (QC & Norm)** - Icíar
  + Input was a .txt file (output of GenomeStudio software based on the paper), no IDATs (raw data) available, which also limited our normalization options
  + Lack detailed knowledge about the authors’ methods in pre-processing methylation data, so we chose our own pipeline using a combination of R packages designed for this purpose - not attempting to reproduce paper here
  + Start: 55 samples, ~480k probes
  + Check sample quality based on the log median intensity in both the methylated (M) and unmethylated (U) channels. Good samples cluster together & have high median intensities - all were good
  + Check sample sex based on X & Y chromosome probes, found 4 samples to be misannotated when clustering in a phylogenetic tree plot. Removed these as we cannot trust the rest of the annotation, + given female-bias of SLE, this may introduce severe errors in downstream analyses.
  + Remove SNP probes (can bias the data), cross-hybridizing probes (bind to multiple locations in the genome), polymorphic probes, poor quality probes (detection P.val < 0.01), non-variable probes (to reduce data dimensionality), and sex probes.
  + Tried a few normalization methods & decided on an intra-array normalization type, BMIQ, which resulted in the “best” beta density plot of our final dataset after filtering of samples and probes (51 samples, ~300k probes) - see before & after beta density plots. A reminder that beta values are the estimate of methylation levels using the ratio of intensities between methylated & unmethylated alleles - values range from 0 (unmethylated) to 1 (fully methylated).
* **Slide #7: Gene expression - Reproducing the paper -** Naila

We first attempted reproducing the analysis conducted in the study.

**Part A:** we highlighted inconsistencies in reporting numbers of DE genes.This table summarizes our attempt where we quantified the lists of differentially expressed genes provided in the supplementary data. The numbers provided in the paper don’t match the supplementary!

**Part B:** Authors performed a t-test with no multiple testing error correction at p value<0.05. This is already a red flag. We performed t-test on quantile normalized data. Venn diagrams show overlaps between our results and supplementary lists.

We run GSEA ( fgsea and ermineR) on author’s supplementary lists and our t-test output.

Indeed, amongst the top significantly enriched pathways(padj<0.05) in common with the paper are shown in the bar plot( Normalized Enrichment score is averaged between every three comparisons ). This is enrichment of upregulated genes in all three comparisons. The same pathways were obtained when we performed differential expression with a t-test . No common significant enrichment in downregulated gene set was observed in part A and B. In B, only SLELN+ vs Normal down regulated genes yielded few enriched pathways, notably: Natural killer cells mediated immunity (reported in the paper) and leukocyte cytotoxicity.

**Part C:** DE using limma to compare results of moderated t-test at FDR<0.05. Less number of differentially expressed genes resulted compared to authors t-test. We selected one of the 3 volcano plots( SLE Vs Normal) . Generally, we can see top upregulated hits are interferon related genes such as IFI27, IFI44L and IF44.Whereas top common downregulated hits comprise natural killer cells genes such as KLRG1, KLRF1, KLRC3. Enrichment analysis for part C agreed with parts A and B. Pathways related to interferon type I, leukocyte, cytokines and granulocytes response were enriched in upregulated genes. Downregulated gene sets didn’t highlight interesting specific pathways but rather few general ones like “regulation of immune response” and “defense response”.

* **Slide #8: Gene expression - Multivariate Analysis -** Sierra
  + In models 2 and 3, age is a continuous variable
  + Result of contrasts in first model: no significant DE genes between SLE-LN- and SLE-LN+
  + The top right figure shows the top DE genes when sex and interactions are used in the model
  + The bottom right figure shows top DE genes when all covariates are additively in the model
  + There is some overlap between these two sets of top genes
  + Overall, the amount of differential expression (magnitude) for each each seems low, while still called as significant
* **Slide #9: Gene expression -** Naila

Decided to focus the downstream analysis on an additive model that adjusts for age and sex main effects.

· Here we are highlighting some of the top significant hits ( FDR<0.05 and lg2FC>1) for genes that were upregulated/downregulated in all 3 comparisons. These top candidate genes that surfaced in all of our analysis. In terms of up regulation, we see response to interferon related genes such as IFI44L, IFI44, IFI27. We also see RNASE2 which is an Eosinophil-derived neurotoxin. ERCP is also an Eosinophil related gene. In downregulation, KLRG1 (Natural killer cells related gene) shows the strongest stepwise decreasing pattern.

· The venn diagrams highlight the total number of significant up/down genes (FC threshold of 1.2 and FDR<0.05). Using this model, we ended up with 186 (compared to 209 in univariate part C limma model) commonly up regulated genes in diseased patients in all comparisons. We also obtained 371 (375 in univariate partC) common down regulated genes in diseased patients. In terms of gene set enrichment, similar results were obtained as in part C. Interestingly, Leukocyte migration was enriched in SLELN+ upregulated gene set. Response to viral infections /symbiont responses also highlighted.

* **Slide #10: Methylation #1** - Nikita
  + Decided on a statistical p-value threshold: different thresholds gave a different number of significant probes:
    - Benjamini-Hochberg at 0.05: 59930
    - Benjamini-Hochberg at 0.01: 29126
    - Bonferroni (which is the most conservative): 1319

→ Decided on BH (FDR) of 0.05 in line with the authors

* + Paper gave an additional cutoff of Fold Change > 1.2. Which is kind of different than usual, where for methylation expression, delta beta values are used. But we were trying to reproduce their analyses.
  + Fig 1 (on the left) shows the **Differentially Methylated Probes (DMPs)** - both hyper and hypomethylated between cases (SLE-LN+ and SLE-LN- considered together as cases) and between controls

→ ~39k probes of ~485k

* + Fig 2 (on the right) shows the DMPs between the two case types -- SLE-LN+ and LN- .

→ ~ 322k probes which was quite surprising

* **Slide #11: Methylation #2** - Nikita
  + Now, of the hyper and hypo DMPs, after converting the logFC values back to FC and selecting for FC >= 1.2, wanted to see how many match and overlap

→ ST3 and ST4 from the paper showed how many probes overlapped between cases and controls separately for hyper and hypo methylation

* + Downloading their data and then joining our generated data by probe name

→ 26 hypermethylated probes matched

→ 16,489 hypomethylated probes matched

* + Further, see how many of the matched probes actually matched for the FC.

→ 0 for hypermethylation

→ 5 for hypomethylation

* + Both figures show the difference in the FC between their and our analysis for the top 10 probes selected by highest FC from the original authors’ results

→ Quite discordant

* **Slide #12: Methylation #3** - Nikita
  + For the 5 hypomethylated probes that matched by FC, wanted to see their genomic location, and if they were involved in any enrichment pathways
  + Annotation info downloaded via the FDb.InfiniumMethylation.hg19 package for hg19
  + Fig 1 (left) shows the corresponding gene name for the 5 probes and where they’re located in the genome (quite a cool plot)
  + The 5 probes were analysed for any Gene Ontology terms they might be associated with, and even though a list of ~23k GO terms was returned, none were statistically significant
  + Fig 2 (right) shows the DMPs for SLE Cases (both) and Controls in only females → since SLE presents at a higher incidence in females. ~330k probes of the ~485k total promes were differentially methylated.

→ ~198k hypermethylated

→ ~131k hypomethylated

* **Slide #13: X-Chromosome** - Icíar
  + As mentioned earlier, autoimmune diseases (which include SLE) are heavily female-biased at a 9:1 female-male prevalence ratio.
  + “All biological sex differences are the result of the inequality in effects of the sex chromosomes, which are the only factors that differ in XX vs. XY zygotes” - Arthur Arnold
  + Analysis limited to female case vs. female control - not fair to compare X chr DNAm between males & females due to genomic imbalance (X inactivation escape genes)
  + Workflow: 44 female samples, extracted 11232 probes that map to the X chr > QC where we removed probes > Dasen normalization, which we chose as opposed to others based on the beta value density plot before & after normalization. Dasen does background adjustment followed by between-array normalization performed separately by probe design. Result was 44 samples, 9835 probes.
  + 61% of variance is explained by the first 10 PCs. In the PC1/PC2 plot, seems like control samples are clustering together and same with disease samples (roughly), but there are some disease samples clustering with disease and vice versa.
* **Slide #14: X-Chromosome** - Icíar
  + We used a single model for this analysis, DNAm ~ Condition + Age, which we fit using the limma package.
  + With an FDR < 0.05 and a delta Beta > 5%, 120 probes were found to be significantly differentially methylated. We also tried using a higher delta Beta (>10%), but there were no hits at this threshold.
  + To calculate the delta Beta, we subtracted the average beta values of controls from those of cases.
  + Out of the 120 hits, 114 were hypomethylated, and 6 were hypermethylated.
  + We ran a gene ontology analysis using gometh() from the missMethyl R package, but no GO pathways were enriched for within our top hits at an FDR < 0.05. This function takes into account the number of CpG sites per gene on the 450K array, which is important because the CpG sites per gene on the array differ, which can introduce a bias in gene ontology analyses of methylation data.

* **Slide #15: Integration -** Sierra
  + Used glmnet package to perform a logistic regression where the response is disease status (case or control) for each sample, and the data matrix is signals where the rows correspond to sample and the columns are the probes
  + It is not done with the intention of finding the best model in order to make predictions, as we have already seen discrepancies in our results and the author’s. Instead, this analysis aims to include both data types at once and explore which variables (probes) are relevant under that framework.
  + Orange arrows mean they were found in both linear models 1 and 3 (sex and status with interaction, and additive effects, respectively); Purple arrow means it was only found in the additive model
  + In the results when using both data types, the probe found using only methylation data was also found here. However, only a subset of the probes found using only expression data were found as relevant when using both data, and a set of different probes were found as relevant
  + Going forward, we would have compared the location of the relevant methylation probes to that of the relevant genes found using the expression data to see if they are close together or if the methylation probes could affect these genes

* **Slide #16: Integration -** Naila

Here we are focusing on IFN genes. Some Interferon response related genes were upregulated and the hypothesis is that they could play a role in SLE pathogenesis. RTPCR in the paper confirmed this observation. We expected to see more hypo methylated probes corresponding to IFN genes.

**A**: distribution of logFC’s of all DMRs corresponding to IFN genes in SLE vs normal. Most of them are hyper methylated.

**B**: logFC’s of top IFNs ("IFI27", "IFI44L","IFI44"). Most of the probes are hyper methylated.

**C:** focuses on IFI27 which was the top hit in all the comparisons. Most of the probes corresponding to this gene seem to be hyper methylated as well.

**D-I:** Same rationale for remaining two comparisons.

Overall, most of those probes are hypermethylated rather than hypomethylated as reported by the authors. IFI27 is the strongest upregulated gene with the largest fold change in our analysis. The authors don’t discuss this gene and it does not display a decrease in methylation. There seems to be only one probe/CpG hypomethylated. We investigated the top hits and found similar variation.

* **Slide #17: Conclusions** *(text in slides)* - Naila