Bioinformatics approaches to the discovery of novel therapeutics in immunological disorders

Pramod Akula Bala

Bioinformatics Projects

- **Microarray** analysis of macrophages from HMDP, *mouse* adipocytes and leukocytes and GEO database
- **Bulk RNA Seq** analysis of human PBMCs from WIHS cohort, human and mouse neutrophils and T cells
- **Single cell RNA Seq** analysis of mouse aorta and lymph nodes
- **GWAS and eQTL Analysis** of the SNP data and RNA seq data from WIHS cohort

HIV-CVD: WIHS Cohort Study

RNASeq and GWAS/ eQTL Analysis

Bioinformatics Projects: Bulk RNASeq and GWAS/eQTL

- **Bulk RNASeq and GWAS/eQTL** analysis of human PBMC RNA Seq and genomic data from WIHS cohort

- The Women's Interagency HIV Study (WIHS) is an ongoing longitudinal cohort study of over 4,000 women with or at risk of HIV infection that was initiated in 1994 at six (now expanded to ten) U.S. sites

- Measurements of carotid artery focal plaque (cIMT and lesion number) , a marker of subclinical atherosclerosis, were obtained

- 92 Patients divided into 4 subgroups of 23 each (HIV+CVD+, HIV-CVD+, HIV-CVD-, HIV-CVD+) for RNASeq analysis

- RNA Seq analysis was followed by GWAS and eQTL analysis on 1794 and 92 patients, respectively

Challenges in RNA-Seq

- RNA Degradation – Sample collection,handling, treatment, processing and preservation

- Contamination with external agents: bacteria, viruses that could degrade RNA by RNAses

- RNA Quality – proportion of other RNAs compared with mRNA; mRNA comprises only 1-3% of total RNA; Ribosomal RNA, on the other hand, makes up >80% of total RNA samples, with the majority of that comprised by the 28S and 18S rRNA species

- Sequencing depth to get the optimum counts of the reads (25 to 50 M reads/sample)

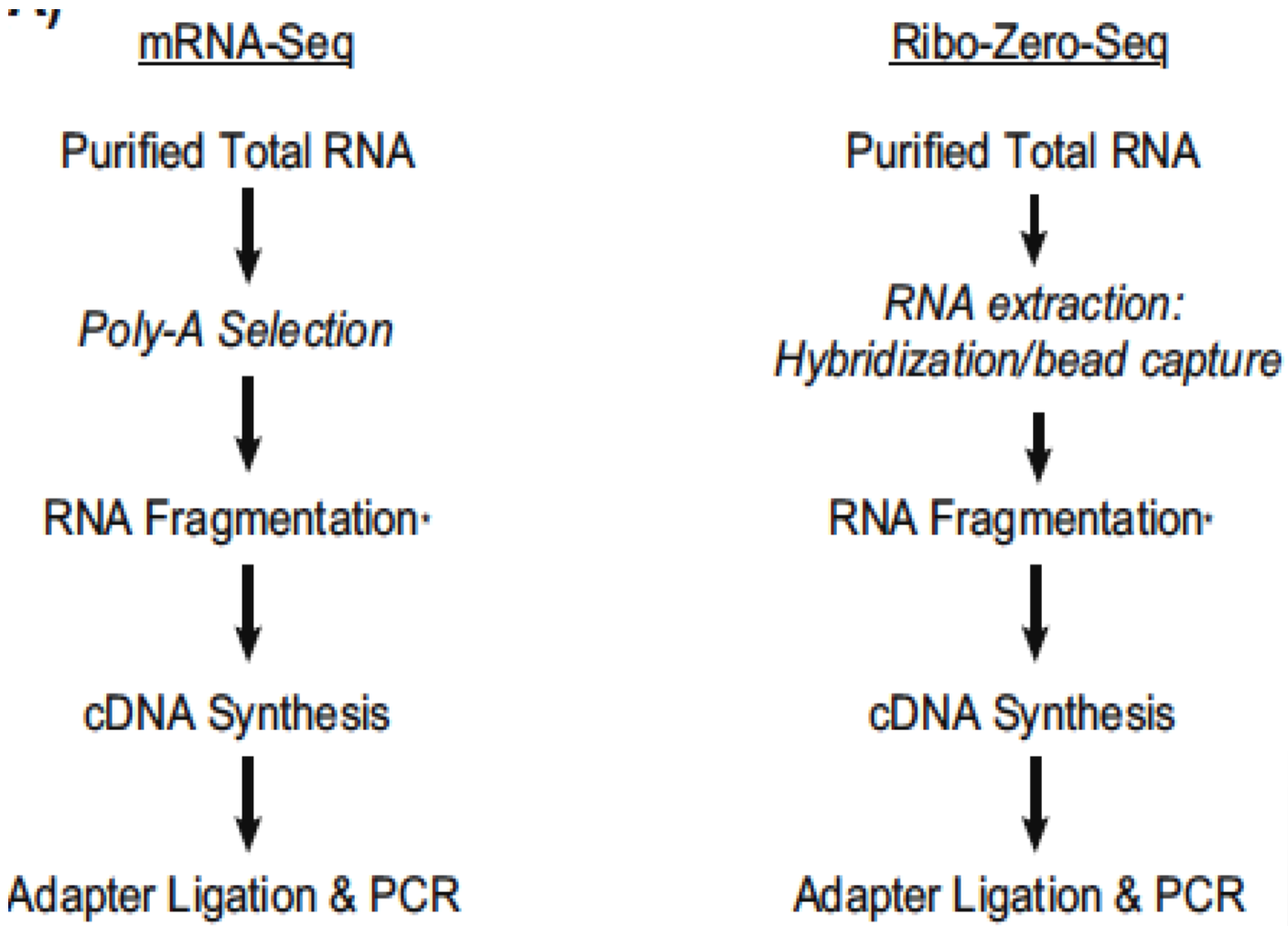
Quality Check: Tapestation RIN

Extraction #	WIHS_ID#	FACS Cell #	Cell # for Extraction	Vol for Extraction	Elution volume (ul)	Extraction date	Theoretical RIN	rRNA Ratio 28/18	Conc (pg/ul)	Tota Extracted ng/ul
22	10100038	215,163	153,688	500	50	5. Oct 15	6.7	1.2	2800	114.80
39	10100381	483,696	241,848	500	50	5. Oct 15	9.6	1.9	2070	84.87
1	10100610	243,177	151,986	500	50	5. Oct 15	5.2	0.9	3330	136.53
4	10100634	105,794	75,567	500	50	5. Oct 15	8.6	1.9	2310	94.71
20	10100684	212,171	212,171	500	50	5. Oct 15	7.4	1.5	2970	121.77
6	10100696	236,521	147,826	500	50	5. Oct 15	6.5	1.2	2750	112.75
11	10100711	338,311	211,444	500	50	5. Oct 15	7.7	1.5	4330	177.53
5	10101016	169,517	121,084	500	50	5. Oct 15	7.6	1.6	3420	140.22
18	10101066	152,599	152,599	500	50	5. Oct 15	7.2	1.4	3480	142.68
24	10101131	312,600	173,667	500	50	5. Oct 15	6.6	1.5	2630	107.83
42	10101179	250,000	125,000	500	50	5. Oct 15	8.6	1.4	3720	152.52
83	10101434	108,445	54,223	300	50	5. Oct 15	8.8	1.6	1070	43.87
76	10101446	846,541	423,271	500	50	5. Oct 15	8.7	2	4610	189.01
44	10101509	347,544	173,772	500	50	5. Oct 15	8.8	2	3910	160.31
2	10101561	129,618	92,584	500	50	5. Oct 15	9.2	2.4	2190	89.79
32	10101888	375,792	234,870	500	50	5. Oct 15	4	1.3	4420	181.22
9	10102068	244,783	152,989	500	50	5. Oct 15	6.3	1	3400	139.40
46	10102412	218,130	109,065	500	50	5. Oct 15	9.6	1.8	1990	81.59
14	10102436	253,963	158,727	500	50	5. Oct 15	9.4	1.6	3200	131.20
58	10102537	193,367	96,684	500	50	5. Oct 15	8.3	1.5	1590	65.19
48	10102640	88,896	88,896	500	50	5. Oct 15	8.7	1.2	820	33.62
60	10102739	284,815	142,408	500	50	5. Oct 15	8.8	1.4	2670	109.47
56	10103111	229,079	114,540	500	50	5. Oct 15	8.8	2	3080	126.28
52	10103147	547,190	273,595	500	50	5. Oct 15	8.1	1.8	3480	142.68
40	10103236	201,234	100,617	500	50	5. Oct 15	7.7	1.5	2200	90.20
7	10205030	475,890	264,383	500	50	5. Oct 15	4.7	1.1	3080	126.28
36	10205270	250,000	125,000	500	50	5. Oct 15	9.4	2	1840	75.44
68	10205321	146,203	73,102	500	50	5. Oct 15	8.6	1.4	1780	72.98
19	10205559	150,552	150,552	500	50	5. Oct 15	7.9	1.7	2100	86.10
3	10205749	162,154	115,824	500	50	5. Oct 15	7.1	1.6	3220	132.02
54	10205775	654,021	327,011	500	50	5. Oct 15	8.5	0.9	5640	231.24
84	10205814	328,295	164,148	500	50	5. Oct 15	8.3	1.9	4790	196.39
8	10205989	411,703	228,724	500	50	5. Oct 15	4.2	0.9	3220	132.02
92	10206474	250,000	125,000	500	50	5. Oct 15	8.7	2.2	4270	175.07
55	10310069	627,475	313,738	500	50	5. Oct 15	8.8	1.4	3880	159.08
16	10310538	135,609	96,864	500	50	5. Oct 15	7.8	1.8	2180	89.38
13	10310603	139,068	99,334	500	50	5. Oct 15	7.7	1.5	2930	120.13
31	10310627	274,514	171,571	500	50	5. Oct 15	5.1	1	2940	120.54
66	10311403	500,000	250,000	500	50	5. Oct 15	8.2	1.2	3530	144.73
15	10311465	169,277	120,912	500	50	5. Oct 15	8.1	1.9	2690	110.29
64	10311554	232,530	116,265	500	50	5. Oct 15	8.5	1.3	1700	69.70
17	12120001	695,251	347,626	500	50	5. Oct 15	4.6	1	1970	80.77
74	12120050	1,055,650	527,825	500	50	5. Oct 15	8.9	2	5170	211.97
21	12120186	320,503	228,931	500	50	5. Oct 15	5.3	1	2770	113.57
82	12120249	475,196	237,598	500	50	5. Oct 15	7.4	1.4	6940	284.54
25	12120301	427,487	213,744	500	50	5. Oct 15	4	1.2	5610	230.01

Qualiy check Tapestation RIN

Extraction #	WIHS_ID#	FACS Cell #	Cell # for Extraction	Vol for Extraction	Elution volume (ul)	Extraction date	Theoretical RIN	rRNA Ratio 28/18	Conc (pg/ul)	Tota Extracted ng/ ul
43	12120313	307,575	153,788	500	50	5. Oct 15	9.5	1.6	1820	74.62
23	12120598	479,650	239,825	500	50	5. Oct 15	5	N/A	2510	102.91
30	12120613	107,700	76,929	500	50	5. Oct 15	8.6	1.9	1470	60.27
62	12120962	206,284	103,142	500	50	5. Oct 15	7.9	1	2340	95.94
80	12121449	366,364	183,182	500	50	5. Oct 15	8.5	1.7	4020	164.82
27	20100028	1,941,540	882,518	500	50	5. Oct 15	4.7	0.8	4360	178.76
71	20100167	397,223	198,612	500	50	5. Oct 15	8.9	2.3	3950	161.95
72	20100321	393,022	196,511	500	50	5. Oct 15	8.5	2.3	4670	191.47
47	20100446	249,142	124,571	500	50	5. Oct 15	8.8	1.1	2780	113.98
63	20100559	611,242	305,621	500	50	5. Oct 15	8.4	1.2	1270	52.07
35	20100852	236,266	118,133	500	50	5. Oct 15	9.6	1.8	2180	89.38
67	20101107	593,591	296,796	500	50	5. Oct 15	8	1.4	4870	199.67
79	20101486	250,000	125,000	500	50	5. Oct 15	8.6	1.2	1850	75.85
91	20101854	586,894	293,447	500	50	5. Oct 15	8	1.9	15100	619.10
28	20102539	306,941	191,838	500	50	5. Oct 15	7.6	1.1	2880	118.08
70	20204042	187,613	93,807	500	50	5. Oct 15	8.5	2	3000	123.00
90	20204078	296,493	148,247	500	50	5. Oct 15	7.2	1.7	11100	455.10
29	20204244	1,204,794	602,397	500	50	5. Oct 15	2.5	N/A	3590	147.19
33	20204307	306,679	170,377	500	50	5. Oct 15	5.8	1.3	3430	140.63
78	20204523	324,357	162,179	500	50	5. Oct 15	8.2	N/A	3070	125.87
50	20204561	774,097	387,049	500	50	5. Oct 15	8.1	1.2	3920	160.72
38	20205323	393,486	196,743	500	50	5. Oct 15	4.8	N/A	1600	65.60
26	22107264	660,713	330,357	500	50	5. Oct 15	3.7	N/A	3120	127.92
59	22208028	350,630	175,315	500	50	5. Oct 15	8.8	1.7	2900	118.90
75	22208092	250,000	125,000	500	50	5. Oct 15	7.8	1.3	2610	107.01
10	30306406	119,604	85,431	500	50	5. Oct 15	7.8	1.9	4960	203.36
37	30409048	250,000	125,000	500	50	5. Oct 15	9.5	2.2	2890	118.49
41	40718324	276,175	138,088	500	50	5. Oct 15	9.2	2.2	6020	246.82
45	40718362	500,000	250,000	500	50	5. Oct 15	9.6	2.1	2060	84.46
49	40718449	577,523	288,762	500	50	5. Oct 15	8.6	1.7	7570	310.37
53	40718475	435,986	217,993	500	50	5. Oct 15	9	2.3	2920	119.72
34	41130266	422,534	211,267	500	50	5. Oct 15	4.7	1.5	3350	137.35
51	42204010	182,946	91,473	500	50	5. Oct 15	8.2	1.3	1920	78.72
57	42719290	433,377	216,689	500	50	5. Oct 15	7.9	2.3	6650	272.65
61	50110277	47,455	47,455	500	50	5. Oct 15	8.2	1.3	1970	80.77
65	50220395	108,046	54,023	300	50	5. Oct 15	6.4	N/A	877	35.96
69	50220410	146,738	73,369	500	50	5. Oct 15	8.5	1.1	1690	69.29
12	50330716	675,934	337,967	500	50	5. Oct 15	7.4	1.8	6070	248.87
86	50331263	121,834	121,834	500	50	5. Oct 15	8.5	1.3	1720	70.52
73	60100329	500,000	250,000	500	50	5. Oct 15	8.8	2	6670	273.47
87	60100331	574,910	287,455	500	50	5. Oct 15	9	2.4	5000	205.00
88	60100420	237,638	118,819	500	50	5. Oct 15	9.1	2	2850	116.85
77	60100874	500,000	250,000	500	50	5. Oct 15	8.6	1.5	2660	109.06
81	60101042	111,414	55,707	300	50	5. Oct 15	9.2	N/A	1690	69.29
85	60101131	578,020	289,010	500	50	5. Oct 15	8.6	2	5340	218.94
89	60101484	352,555	176,278	500	50	5. Oct 15	8.5	1.8	6350	260.35

RNA seq-method options



Brief Bioinformatics Pipeline

1
- RNA-Seq reads obtained in Fastq files
- FastQC done to check the read qulaity (12 parameters)

2
- Only mRNA reads retained, rRNA,tRNA, adaptors, low complexity reads removed
- RNA-seq reads mapped to reference genome using Tophat

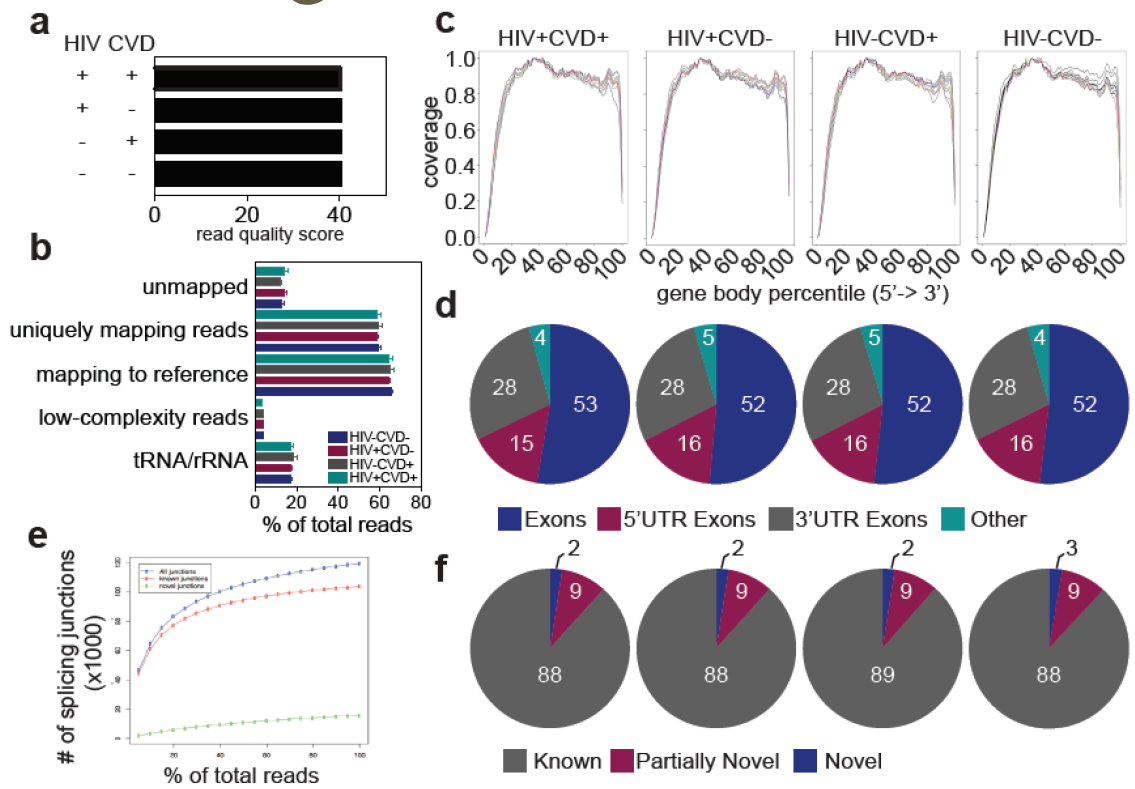
3
- Mapped reads output: BAM files, counts and RPKM values
- QC for mapping using RSEQC(10 parameters)

4
- Differential analysis using DESeq2 R/ Bioconductor packages
- Expression levels for house keeping and apoptotic genes analyzed

5
- Differential genes sorted at FDR of 10% and linked to genecards

6
- Differential gene sets analyzed for Gene set enrichment analysis and GO terms

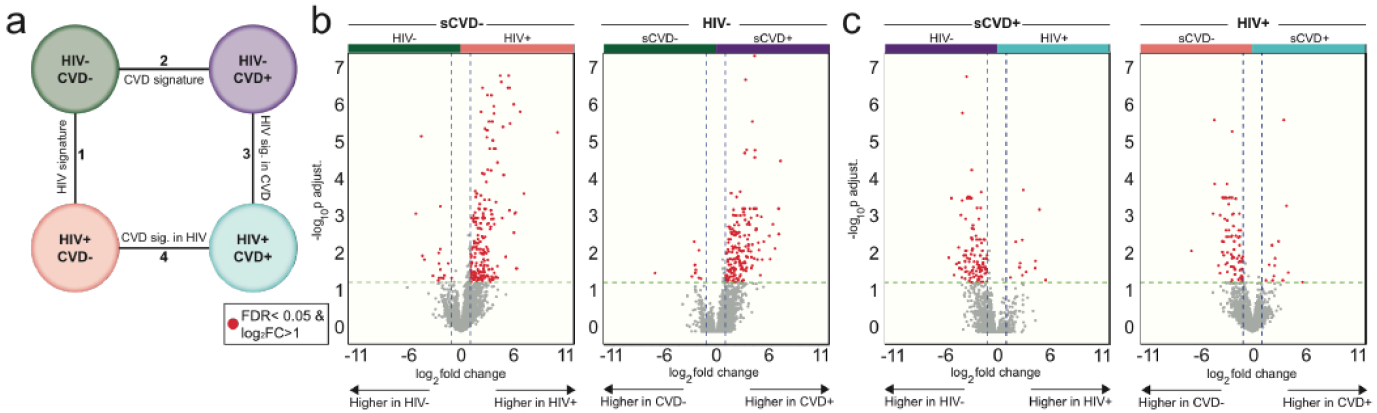
RSeQC




```
@folder = `cat ../number.list`;
        @job = `cat control.job`;
        $i = 0;
        while ($i <= $#folder){
                        chomp $folder[$i];
                        open(out,">cluster.$folder[$i].job");
                        print out @job;
                         print out "cp /path/accepted_hits_filtered_dust_4_unique.bam $folder[$i]_accepted_hits_filtered_dust_4_unique.bam\n";
                        print out "samtools index $folder[$i]_accepted_hits_filtered_dust_4_unique.bam\n";
                        close out;
                        `qsub cluster.$folder[$i].job`;
                        $i++;
        }
python2.7 ../scripts/geneBody_coverage.py  -i /path/accepted_hits_filtered_dust_4_unique.bam -r ../refseqexpthg19 -o ../genebody/Result.$s-$e-genebody
python ./scripts/read_NVC.py -i /path/accepted_hits_filtered_dust_4_unique.bam -o ./read_NVC/#_NVC
python ./scripts/read_quality.py -i /path/accepted_hits_filtered_dust_4_unique.bam -o ./read_quality/#_quality
```

```
#!/bin/bash -ex
#PBS -l nodes=1:ppn=1
#PBS -l mem=20GB
#PBS -l walltime=50:00:00
#PBS -m ae
#PBS -j eo
#PBS _V
source ~/.bashrc
hostname
TMPDIR=/scratch
cd $PBS_O_WORKDIR
```

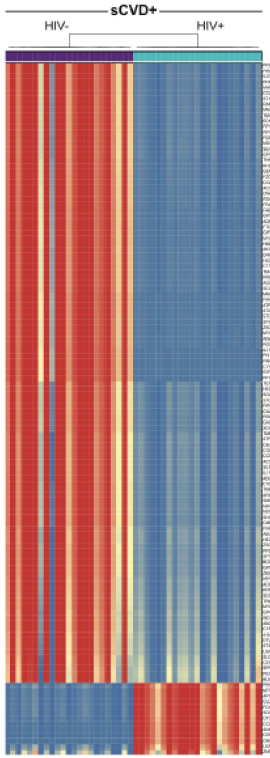
DE and Volcano plots

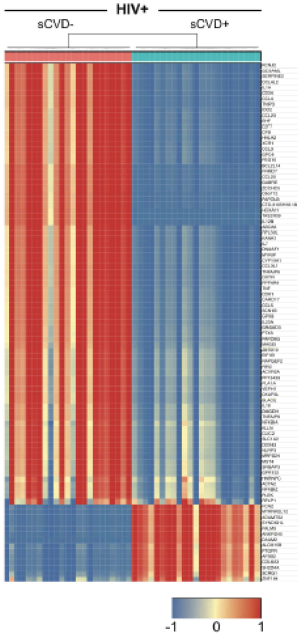


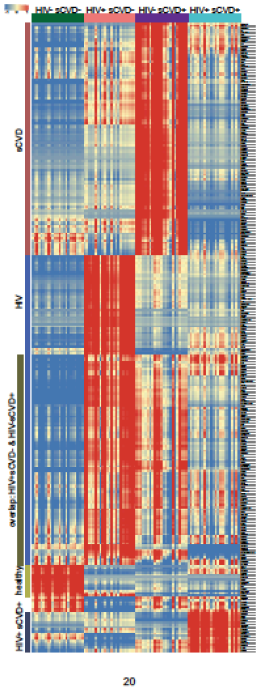

```
fc <-read.table("HIV_samples_for_DE_order_PC_nodup.txt", row.names = 1, header = TRUE)
x <- as.factor(rep(c("HIVMCVDM", "HIVMCVDP", "HIVPCVDM", "HIVPCVDP"), each = 23))
fc<-as.matrix(fc)
design <- model.matrix(~x)
y <- DGEList(counts = fc, group = x)
y <- calcNormFactors(y, method = "upperquartile")
y <- estimateGLMCommonDisp(y, design, verbose = TRUE)
y <- estimateGLMTagwiseDisp(y, design)
fit <- glmFit(y, design)
lrt <- glmLRT(fit, coef = 2)
topUQFC <- topTags(lrt, n = Inf)$table
write.table(topUQFC, file = "topUQFC_HIVMCVDP_HIVMCVDM_uq.xls", sep="\t",)
my.contrasts <- makeContrasts(xHIVPCVDMvsxHIVMCVDP=xHIVPCVDM-xHIVMCVDP, xHIVPCVDPvsxHIVMCVDP=xHIVPCVDP-xHIVMCVDP, xHIVPCVDPvsxHIVPCVDM=xHIVPCVDP-xHIVPCVDM, levels=design)
lrt <- glmLRT(fit, contrast=my.contrasts[,"xHIVPCVDMvsxHIVMCVDP"])
top_coeff5 <- topTags(lrt, n = Inf)$table
write.table(top_coeff5, file = "HIVPCVDM_vs_HIVMCVDP_uq.xls", sep="\t",)
```

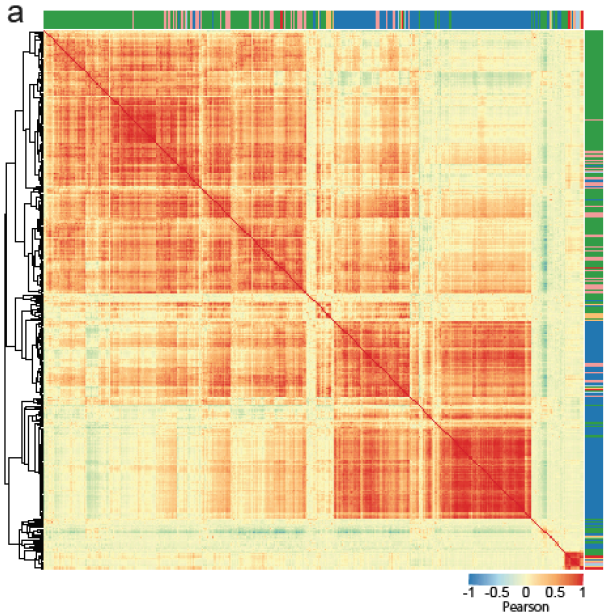
```
results = read.table(file = "HIVPCVDM_vs_HIVMCVDM_2_volplot.txt", header=TRUE)
results$FDR0.05LogFC1 = as.factor(abs(results$log2FoldChange) > 1 & results$padj < 0.05)
results$log10padj = -log10(results$padj)
write.table(results, file="results3.xls", sep="\t",)
results3 = read.table(file = "results3.txt", header=TRUE)
p=ggplot(results3, aes_string(x="log2FoldChange", y="log10padj", colour="FDR0.05LogFC1"))
p1=p+geom_point( size=1.3, show.legend = T)
+coord_cartesian(ylim=c(0, 4))
+scale_y_continuous(breaks=seq(0, 12, 0.5))
+xlim(-2,2)+xlab("log2 fold change") + ylab("-log10 padj")
+scale_color_manual(values=c("red", "blue"))
+theme(panel.background = element_rect(fill = "white", colour = "grey50"),legend.position = "right")
+geom_text_repel(data=filter(results3, padj_mod<0.05), aes(label=Gene),fontface = 'bold',segment.color = 'blue',size=1.8,colour = "black")+labs(title = "HIVPCVDM_vs_HIVMCVDM_5_1_Volcanoplot")
+theme_classic(base_size = 16)
```

Heatmap and correlation plots





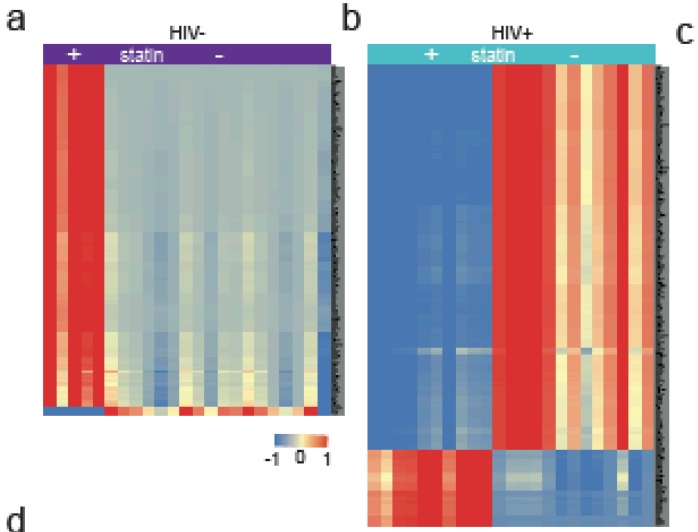


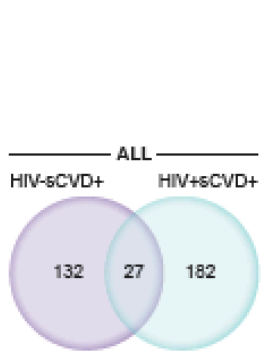


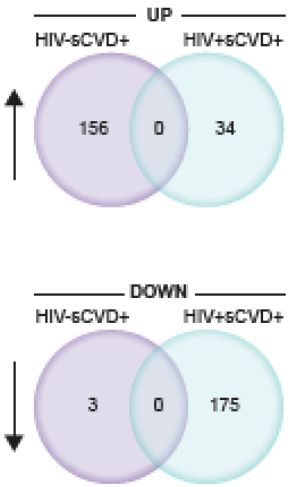

```
PC1_A=read.table("HIV_samples_for_DE_order_PC_nodup_sort_statin
_B_DE_GS.txt", header=T, row.names = 1,)
annotation1 <- data.frame(Type = factor
(rep(c("HIVMCVDM","HIVMCVDP","HIVPCVDM","HIVPCVDP"),c("24",
"24", "24","24"))))
rownames(annotation1) <- colnames(PC1_A)
z <- (-20:20)/10
ppi <- 600
png("HIV_GS_DE_heatmap.png", width=9*ppi, height=9*ppi, res=ppi)
pheatmap(PC1_A, color = colorRampPalette(c("blue", "white", "red"))
(41), cluster_rows = F, cluster_cols = F, show_rownames = T,
show_colnames = T, scale = "row", fontsize = 4, fontsize_row = 1,
fontsize_col = 5, annotation = cbind(annotation,annotation1),
border_color = "NA", breaks = z, cellwidth = 5, cellheight =
1.0,treeheight_row = 0)
dev.off()
```

```
PC1_A_cor=read.table("correlation_395_redo_original_KL.txt", header=T, row.names = 1,)
M<-cor(t(PC1_A_cor),method = "pearson",)
head(round(M,2))
corrp_original_redo_venn<-corrplot(M, method = "color", type="full", order="original", tl.cex = 0.12, tl.col="black",
cl.cex = 1.2, cl.pos = "b")
write.table(corrp_original_redo_venn, file="corrp_original_redo_venn.txt", sep="\t",)
annotation_KL<- data.frame(Cohort = factor (rep(c("Group1", "Group2","Group3", "Group4",
"Group5","Group6","Group7"),c("185", "144", "11","44","5","5","1"))))
cor=read.table("corrp_original_redo_venn.txt", header=T, row.names = 1,) cormat <- as.matrix((cor))
rownames(annotation_KL) <- colnames(cormat)
rownames(annotation_1_KL) <- rownames(cormat)
ann_colors = list(Cohort = c(Group1 = "#bf8bff", Group2 = "#0000ff", Group3 = "#111111",Group4 = "#008744", Group5
= "#00ffff", Group6 = "#e30074",Group7 = "#8d5524" ))
pheatmap(cormat, colors = c("#E46726", "white", "#6D9EC1"), cluster_rows = F, cluster_cols = F, show_rownames = T,
show_colnames = T, scale = "none",  fontsize_row = 1.1, fontsize_col = 1.1, border_color = "NA", cellwidth = 1.0,
cellheight = 1.0,fontface = "bold", annotation_row = annotation_1_KL, annotation_col = annotation_KL,
annotation_colors = ann_colors)
```

Pathway Analysis





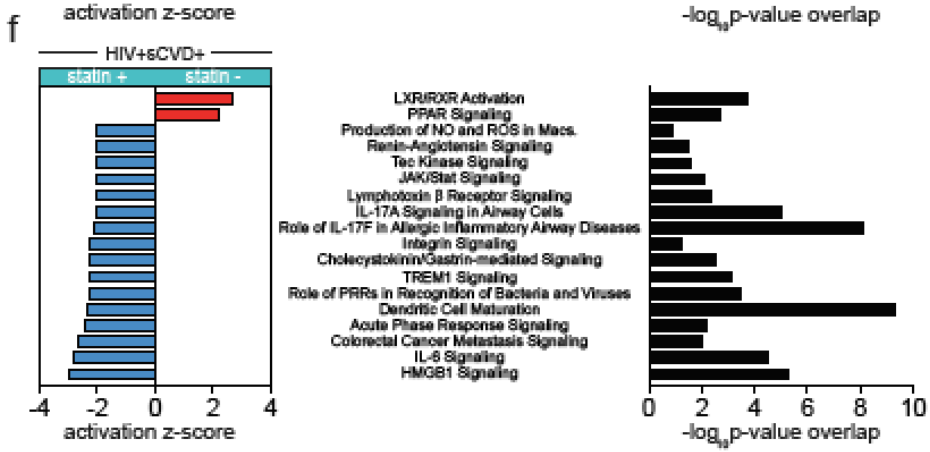


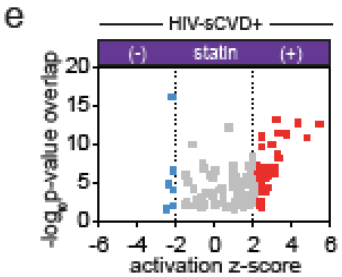
Top Genes - UP

HIV-sCVD+	HIV+sCVD+
IL12B	POU5F1B
MAS1	CFI
IL2RA	RNASE9
CCL22	MEDAG
HBD	MTUS1
HBB	C17orf100
HBM	OR52K1
CHI3L1	KRT15

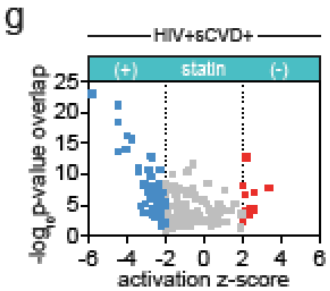
Top Genes - DOWN

HIV-sCVD+	HIV+sCVD+
RP11-166B2	CSF3
RFPL1	CCL24
C5orf63	IL36G
	KIAA1644
	TFPI2
	WNT5A
	C1QTNF1
	RNF17





statin -		statin +	
inhibited		activated	
CK	TF	CK	TF
IL10	NFE2L2	TNF	RELA
IL1RN	BCL6	IL1β	NFKB1
	HDAC1	IFNγ	IRF7
	ZFP36	CD40LG	IRF5
		IL1α	NFATC2



statin +		statin -	
inhibited		activated	
CK	TF	CK	TF
TNF	RELA	IL10	BCL6
IL1α	HMGB1	IL37	MEOX2
IL1β	FOXO1		GFI1
IFNγ	NFKB1		ZFP36
CSF2	IFI16		PRDM1

GWAS and eQTL

Human Omni 2.5M-quad (n=1794) and 91 have both RNA-seq and genotype data.

- QC
- Sex Check
- SNP Call rate
- Person Call rate
- Hardy-Weinberg Statistics
- Relationship Check
- Heterozygosity Check
- PCA Analysis
- Imputation

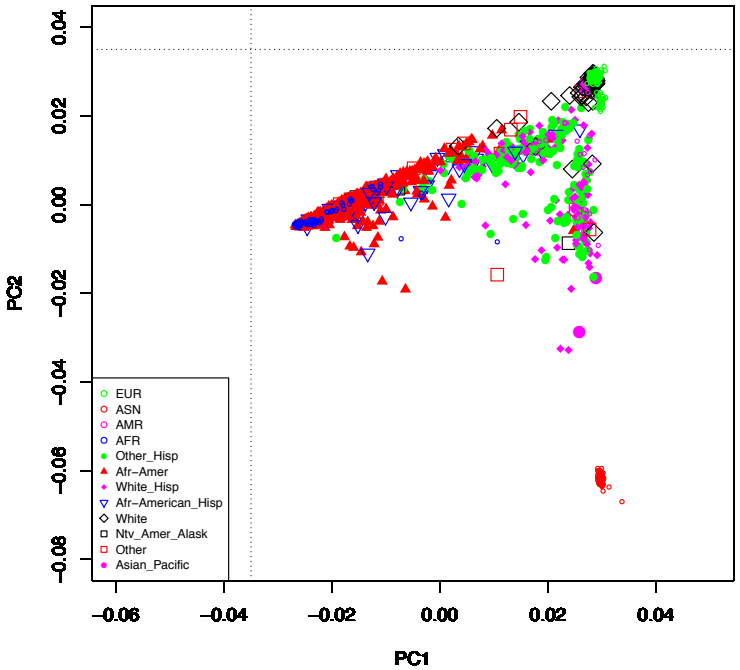
- plink1.9 --bfile WIHS-IMT-final-obfuscated --geno 0.2 --make-bed --out HapMap_3_r3_2
- plink1.9 --bfile HapMap_3_r3_2 --mind 0.2 --make-bed --out HapMap_3_r3_3
- plink1.9 --bfile HapMap_3_r3_3 --geno 0.02 --make-bed --out HapMap_3_r3_4
- plink1.9 --bfile HapMap_3_r3_4 --mind 0.02 --make-bed --out HapMap_3_r3_5
- plink1.9 --bfile HapMap_3_r3_5 --check-sex
- grep "PROBLEM" plink.sexcheck| awk '{print$1,$2}'> sex_discrepancy.txt
- plink1.9 --bfile HapMap_3_r3_5 --remove sex_discrepancy.txt --make-bed --out HapMap_3_r3_6
- plink1.9 --bfile HapMap_3_r3_6 --freq --out MAF_check
- plink1.9 --bfile HapMap_3_r3_6 --maf 0.01 --make-bed --out HapMap_3_r3_8
- plink1.9 --bfile HapMap_3_r3_8 --hwe 1e-6 --make-bed --out HapMap_3_r3_9
- plink1.9 --bfile HapMap_3_r3_9 --remove het_fail_ind.txt --make-bed --out HapMap_3_r3_10
- plink1.9 --bfile HapMap_3_r3_10 --filter-founders --make-bed --out HapMap_3_r3_11
- #perl make_par_file.pl MDS_merge2_LDpruned 0 > MDS_merge2_LDpruned.par
- #smartpca -p MDS_merge2_LDpruned.par
- R --vanilla < plot.pca.r --args MDS_merge2_Ldpruned

- #plink1.9 --freq --bfile HapMap_3_r3_11 --out HapMap_3_r3_11_freq

- #perl HRC-1000G-check-bim-NoReadKey.pl -b HapMap_3_r3_11.bim -f HapMap_3_r3_11_freq.frq --1000g
- #sh Run-plink.sh

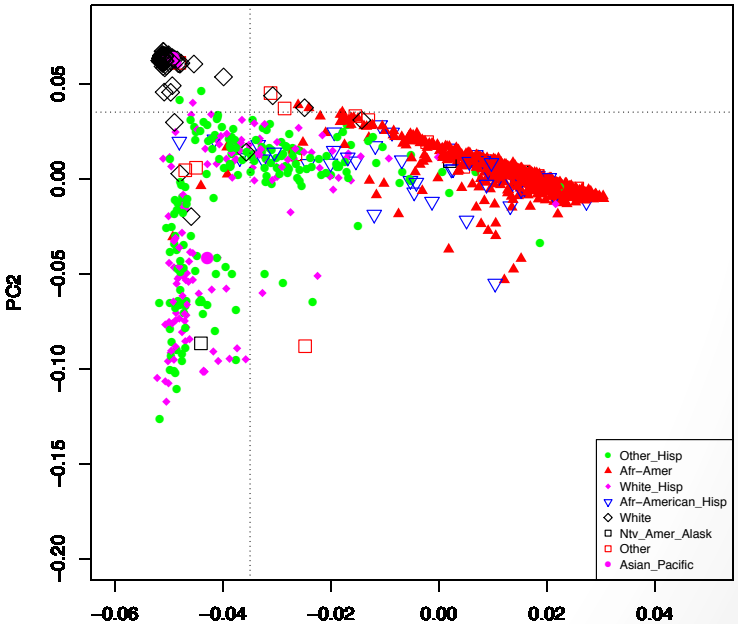
- #export PERL5LIB=/Users/pramod/Desktop/WIHS_GWAS/plink_final_obfuscated-10252017_2/vcftools/src/perl
- #for i in {1..23}; do plink2 --bfile HapMap_3_r3_11-updated --chr ${i} --recode vcf --out HapMap_3_r3_11-updated_for_imput_chr${i}; done
- #for i in {1..23}; do sed 's/^##fileformat=VCFv4.3/##fileformat=VCFv4.2/' HapMap_3_r3_11-updated_for_imput_chr${i}.vcf > HapMap_3_r3_11-updated_for_imput_chr${i}_v2.vcf; done
- #for i in {1..23}; do vcf-sort HapMap_3_r3_11-updated_for_imput_chr${i}_v2.vcf | bgzip -c > HapMap_3_r3_11-updated_for_imput_chr${i}_v2.vcf.gz; done

GWAS and eQTL

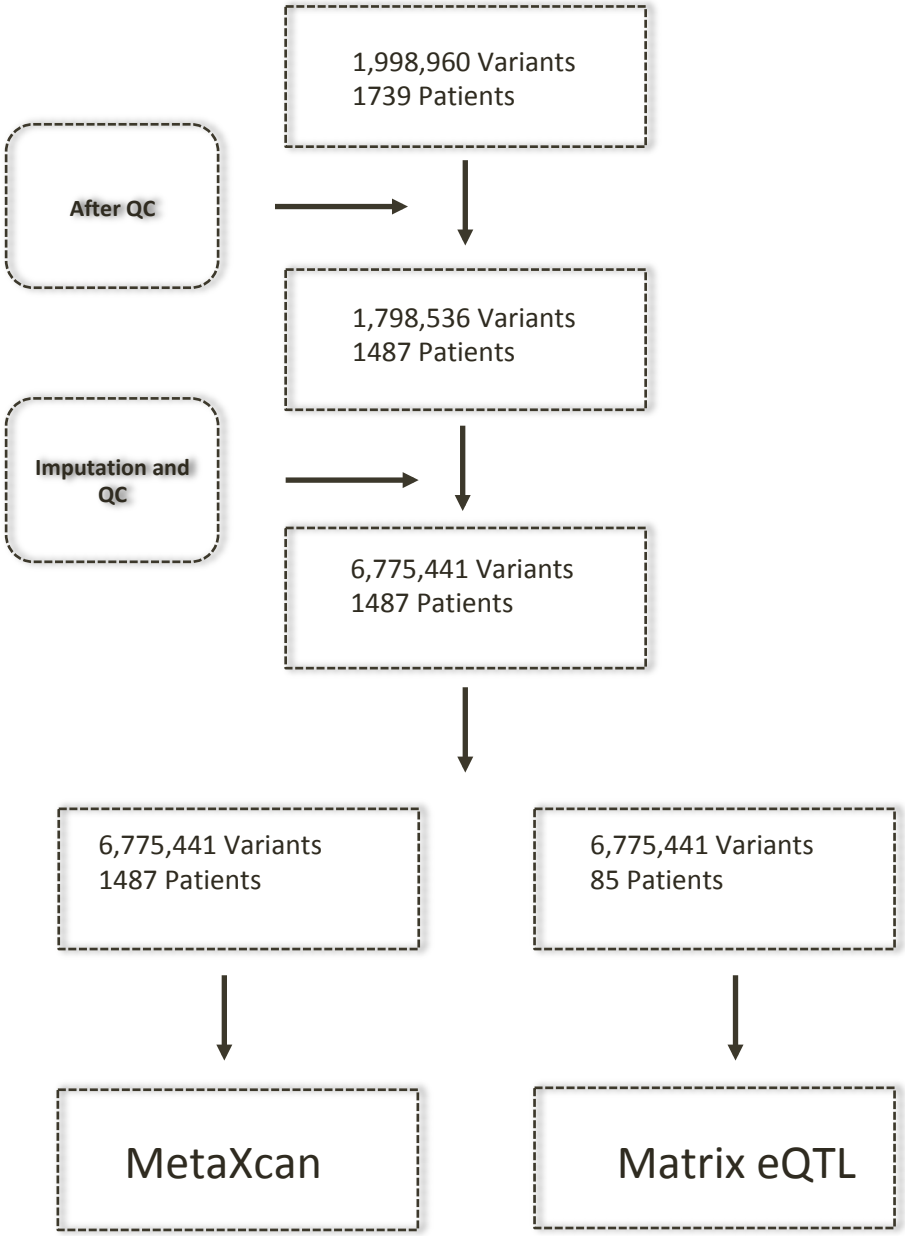


- PCA of 1487 patients (Post Imputation and 100G)
- Included Afr-Amer,white Hisp,Afr-Amer-Hisp, Other-Hisp,White, Nat-Amer-Alask, Other, and Asian-Pacific
- 1000G: ASN, EUR, AMR and AFR

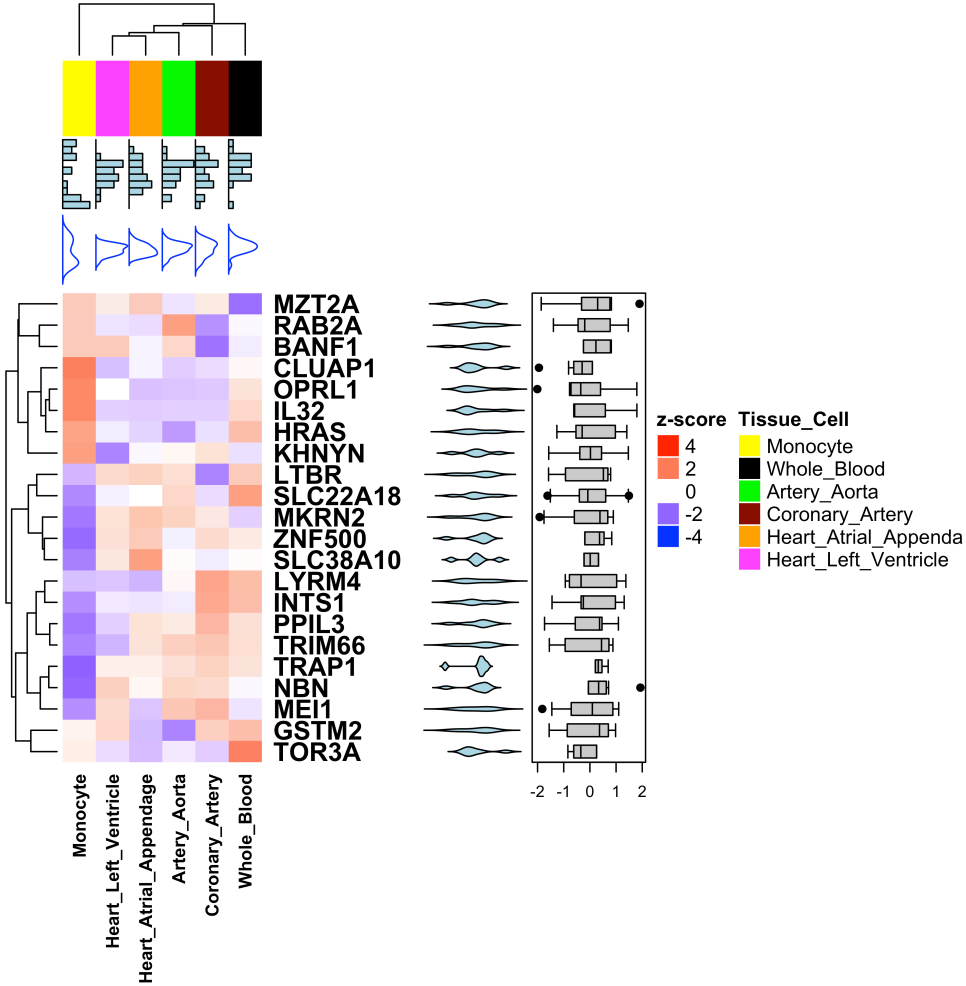
```
data <- read.table(file="MDS_merge2_LDpruned.evec_85.txt",header=TRUE)
pdf("PCA_WIHS_Postimpute_85_1000G.pdf",width=7,height=7)
for (i in 1:nrow(data))
{
    if (data[i,22]=="EUR") {plot(data[i,2],data[i,3],type="p",xlim=c(-0.06,0.05),ylim=c(-0.08,0.04),xlab="PC1",ylab="PC2",pch=1,cex=0.5,col="green")}
 par(new=T)
    if (data[i,22]=="ASN") {plot(data[i,2],data[i,3],type="p",xlim=c(-0.06,0.05),ylim=c(-0.08,0.04),xlab="PC1",ylab="PC2",pch=1,cex=0.5,col="red")}
 par(new=T)
}

 legend("bottomleft",pch=c(1,1,1,1,16,17,18,6,5,0,0,16),c("EUR","ASN","AMR","AFR","Other_Hisp","Afr_Amer","White_Hisp","Afr_American_Hisp","White","Ntv_Amer_Alask","Other","Asian_Pacific"),col=c("green","red",470,"blue","green","red",470,"blue","black","black","red","magenta"),bty="o",cex=0.7)
dev.off()
```



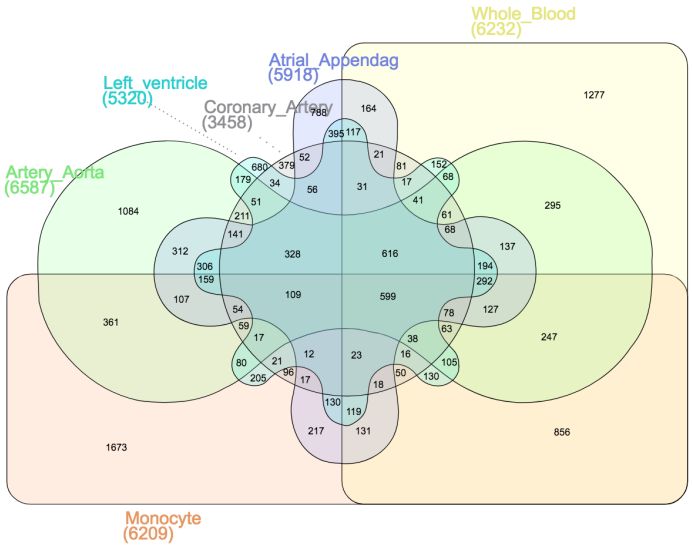
GWAS and eQTL



GWAS and eQTL

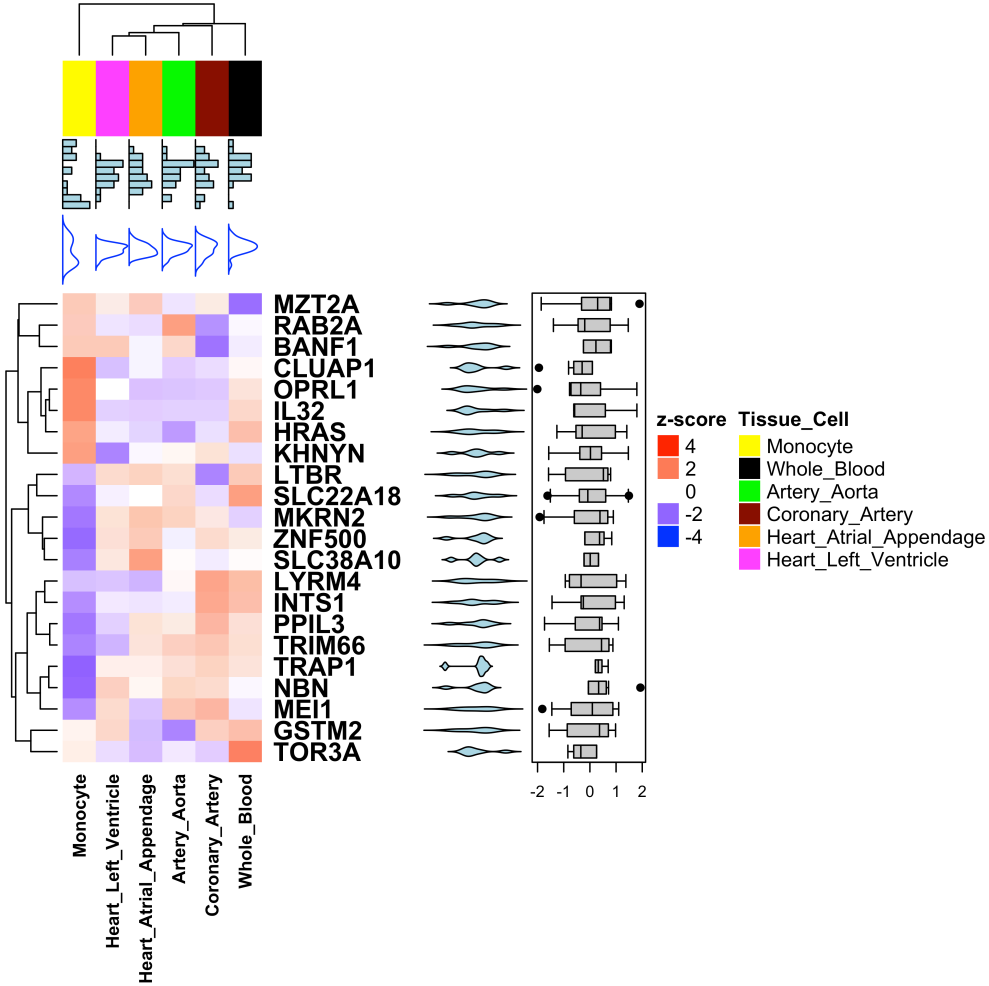


MetaXcan: Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.
Phenotype : IMT and Lesion number
Genotypes : imputed SNPs from 1487 and 85 samples
Reference DB: MESA population (monocyte data) and GTex tissues



- **MetaXcan-predicted association of predicted gene transcript levels with IMT across biologically relevant tissues in GTEx and MESA monocytes.**
- Data are shown for all genes at which altered transcription was significantly associated with IMT in at least one biologically relevant tissue and monocytes, after accounting for multiple testing
- Data are *z*-scores of transcript level association with higher IMT, clustered by tissue and monocytes
- Direction of *z*-score indicates whether higher or lower gene expression is associated with IMT

GWAS and eQTL



MetaXcan: Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.
Phenotype : IMT and Lesion number
Genotypes : imputed SNPs from 1487 and 85 samples
Reference DB: MESA population (monocyte data) and GTex tissues

```
dfa<-read.table(file = "Metaxcan_heatmap_zscore.txt", header = T, row.names = 1)
dfam<-as.matrix(dfa)
mat_scaled = t(apply(dfam, 1, scale))
ppi=600
rownames(df)=colnames(dfam)
head(df)
ann_colors1 = list(Tissue_Cell = c(Monocyte = "Yellow", Whole_Blood = "Black", Artery_Aorta = "green",Coronary_Artery =
"darkred", Heart_Atrial_Appendage="orange",  Heart_Left_Ventricle="magenta"))
ha1 = HeatmapAnnotation(df, col=ann_colors1, show_legend = TRUE)
Heatmap(mat_scaled, name = "Zscore",cluster_columns=F, top_annotation = ha1, col = colorRamp2(c(-2, 0, 2), c("blue", "w
"green")))
dev.off()
.violin = anno_density(mat_scaled, type = "violin",
                gp = gpar(fill = "lightblue"), which = "row")
.boxplot = anno_boxplot(mat_scaled, which = "row")
ha_mix_right = HeatmapAnnotation(violin = .violin, bxplt = .boxplot,
                which = "row", width = unit(4, "cm"), height = unit(4, "cm"))

png("complex_heatmap_metazcore.png", width=7*ppi, height=7*ppi, res=ppi)
Heatmap(mat_scaled, name = "Zscore",row_names_gp = gpar(fontsize = 13, fontface = "bold"),
      cluster_columns=T, top_annotation = ha1, col = colorRamp2(c(-3.0, 0, 3.0), c("blue", "white", "red")))+ha_mix_right
dev.off()
.hist = anno_histogram(mat_scaled, gp = gpar(fill = "lightblue"))
.density = anno_density(mat_scaled, type = "line", gp = gpar(col = "blue"))
ha_mix_top = HeatmapAnnotation(hist = .hist, density = .density, df, col=ann_colors1)
png("complex_heatmap_metazcoreA.png", width=7*ppi, height=7*ppi, res=ppi)
Heatmap(mat_scaled, name = "z-score",row_names_gp = gpar(fontsize = 13, fontface = "bold"),
      cluster_columns=T, show_column_names = T,column_names_gp = gpar(fontsize = 9,fontface = "bold"),top_annotatio
ha_mix_top,top_annotation_height = unit(4, "cm"), col = colorRamp2(c(-3.0, 0, 3.0), c("blue", "white", "red")))+ha_mix_rig
dev.off()
```

- for x in $(cat metadatabase)
- do
- python2.7 MetaXcan.py --model_db_path data/${x}_imputed_10_peer_3_pcs_v2.db --covariance data/${x}_nested_cv_all_covariances_10_peer_3pcs.txt.gz --gwas_folder data/GWAS/ --gwas_file_pattern ".*gz" --snp_column SNP --effect_allele_column A1 --non_effect_allele_column A2 --beta_column BETA --pvalue_column P --output_file results/${x}_COVAR.csv
- done

MetaXcan

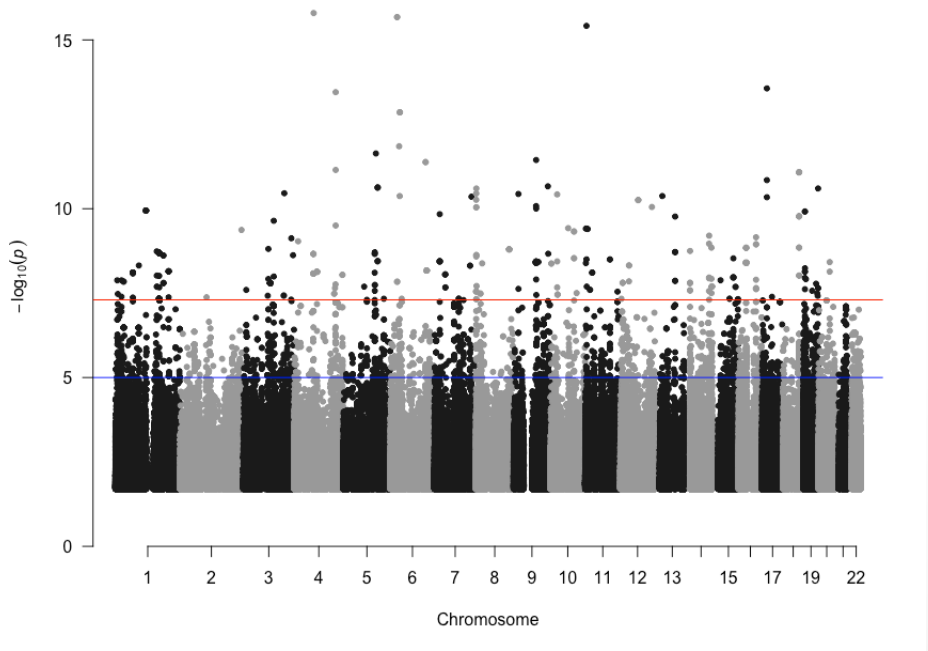
- gene: a gene's id: as listed in the Tissue Transcriptome model. Ensemble Id for some, while some others (mainly DGN Whole Blood) provide Genquant's gene name
- gene_name: gene name as listed by the Transcriptome Model, generally extracted from Genquant
- zscore: S-PrediXcan's association result for the gene
- effect_size: S-PrediXcan's association effect size for the gene
- pvalue: P-value of the aforementioned statistic.
- pred_perf_r2: R2 of tissue model's correlation to gene's measured transcriptome (prediction performance)
- pred_perf_pval: pval of tissue model's correlation to gene's measured transcriptome (prediction performance)
- pred_perf_qval: qval of tissue model's correlation to gene's measured transcriptome (prediction performance)
- n_snps_used: number of snps from GWAS that got used in S-PrediXcan analysis
- n_snps_in_cov: number of snps in the covariance matrix
- n_snps_in_model: number of snps in the model
- var_g: variance of the gene expression, calculated as $W' * G * W$ (where W is the vector of SNP weights in a gene's model, W' is its transpose, and G is the covariance matrix)

MetaXcan

- MetaXcan shares most of the benefits of PrediXcan: a) it directly tests the regulatory mechanism through which genetic variants affect phenotype; b) it provides gene-level results which are better functionally characterized than genetic variants, easier to validate within model systems, and carry a smaller multiple testing burden; c) the direction of the effects are known, facilitating identification of therapeutic targets; d) reverse causality is largely avoided since predicted expression levels are based on germline variation, which are not affected by onset of disease; e) it can be systematically applied to existing GWAS studies; f) tissue-specific analysis can be performed using all the models that are available.

GWAS and eQTL

- Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
- RPKMs > 0.1 in 20 percent of samples
- Counts > 6 in 20 percent of the samples
- Cis_eQTL results at FDR < 0.05
- Gene-SNP pairs : 1668
- eGenes: 362
- eSNPs: 1499
- Manhattan Plot

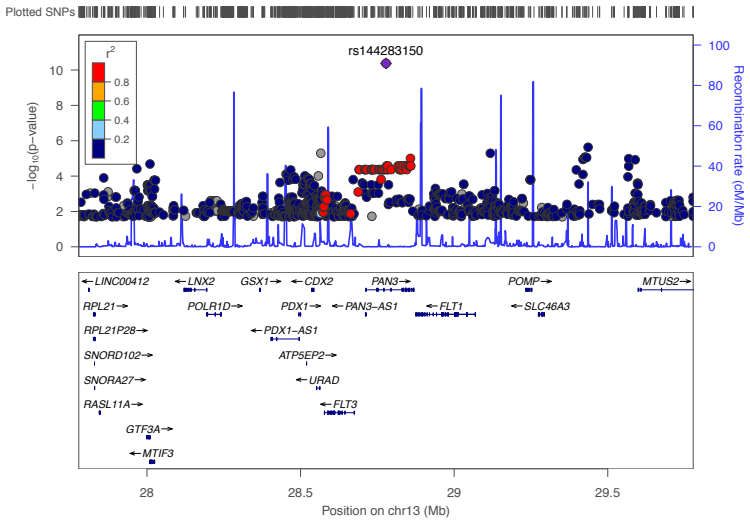


GWAS and eQTL

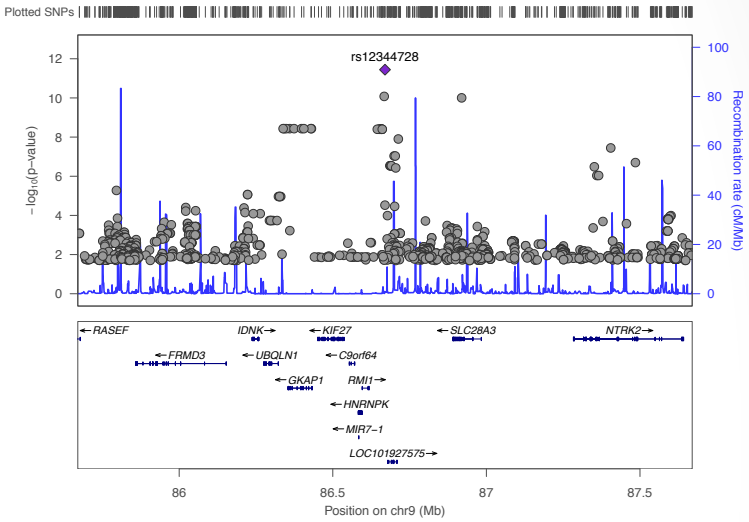
```
base.dir = ("/mnt/BioAdHoc/overflow/Groups/KLey/pramod/WIHS_QC_whole/imputation_Whole_Eagle_autosomes_copy/85_Matrixeqtl_RPKM/")
base.dir
SNP_file_name = paste(base.dir, "snps_expression.txt", sep="")
SNP_file_name
expression_file_name = paste(base.dir, "EXtracted_RPKMS_85_Fqtl_10.txt", sep="")
expression_file_name
snps_location_file_name = paste(base.dir, "snps_loc_head.txt", sep="")
snps_location_file_name
gene_location_file_name = paste(base.dir, "extracted_gene_loc_orderA.txt", sep="");
gene_location_file_name

output_file_name_cis = tempfile();
output_file_name_tra = tempfile();
#covariates_file_name = paste(base.dir, "/data/Covariates.txt", sep="")
#output_file_name = tempfile()
expr = fread("EXtracted_RPKMS_85_Fqtl_10.txt",sep="\t",header=T,)
expr[ 1, ]
snps = fread("snps_expression.txt",sep="\t",header=T,)
snps[1,]
#cvrt = read.table(covariates_file_name,sep="\t",header=T,row.names=1)
#cvrt[ 1, ]
snpspos = read.table("snps_loc_head.txt", header = TRUE, stringsAsFactors = FALSE);
genepos = read.table("extracted_gene_loc_orderA.txt", header = TRUE, stringsAsFactors = FALSE);

pvOutputThreshold_cis = 2e-2;
pvOutputThreshold_tra = 1e-2;

errorCovariance = numeric()
cisDist = 1e6;

useModel = modelLINEAR
snps = SlicedData$new()
snps$fileDelimiter = "\t"     # the TAB character
snps$fileOmitCharacters = "NA" # denote missing values;
snps$fileSkipRows = 1         # one row of column labels
snps$fileSkipColumns = 1      # one column of row labels
snps$fileSliceSize = 2000     # read file in pieces of 2,000 rows
snps$LoadFile( SNP_file_name )
gene = SlicedData$new()
gene$fileDelimiter = "\t"     # the TAB character
gene$fileOmitCharacters = "NA" # denote missing values;
gene$fileSkipRows = 1         # one row of column labels
gene$fileSkipColumns = 1      # one column of row labels
gene$fileSliceSize = 2000     # read file in pieces of 2,000 rows
gene$LoadFile(expression_file_name)

me = Matrix_eQTL_main(
 snps = snps,
 gene = gene,
 output_file_name     = output_file_name_tra,
 pvOutputThreshold     = pvOutputThreshold_tra,
 useModel = useModel,
 errorCovariance = errorCovariance,
 verbose = TRUE,
 output_file_name.cis = output_file_name_cis,
 pvOutputThreshold.cis = pvOutputThreshold_cis,
 snpspos = snpspos,
 genepos = genepos,
 cisDist = cisDist,
 pvalue.hist = "qqplot",
 min.pv.by.genesnp = FALSE,
 noFDRsaveMemory = FALSE);
unlink(output_file_name_tra);
unlink(output_file_name_cis);
```

GWAS and eQTL

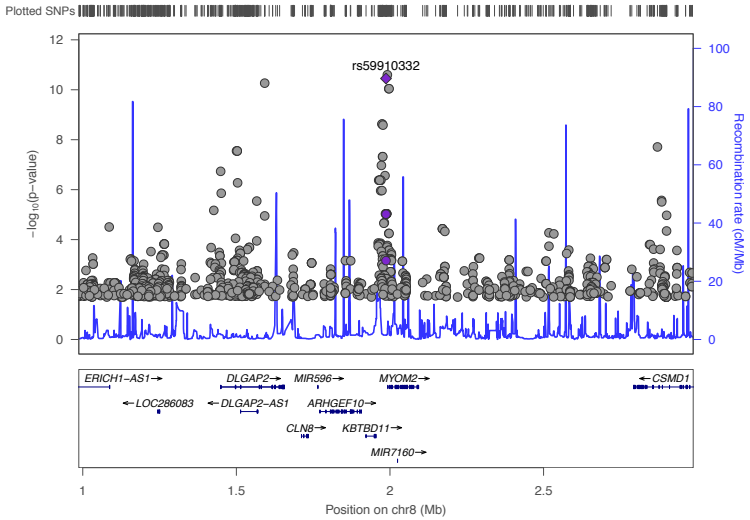
FLT3



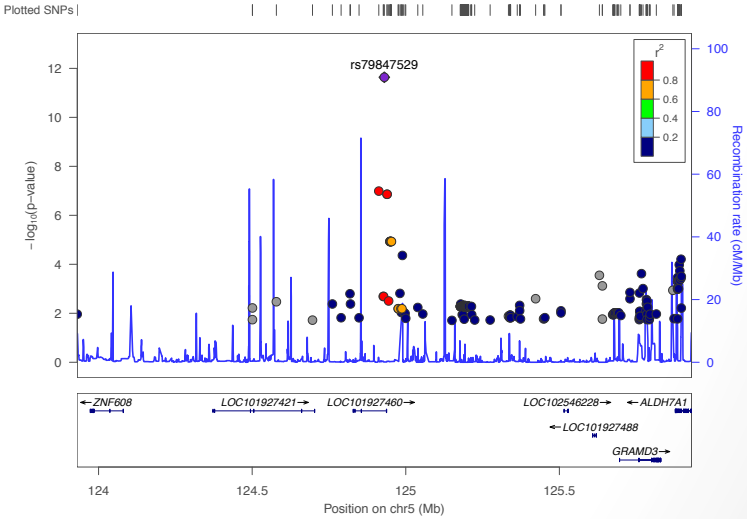
SLC28A3



ARHGEF10



GRAMD3



GWAS and eQTL

- Flt3 Signaling-Dependent Dendritic Cells Protect against Atherosclerosis. Immunity (2011)
- FLT3 activation improves post-myocardial infarction remodeling involving a cytoprotective effect on cardiomyocytes. J.Am.Coll.Cardiol. (2014)
- SLC28A3: Rare coding variants associated with blood pressure variation in 15 914 individuals of African ancestry. J.Hypertens (2017)
- SLC28A3 rs7853758 as a new biomarker of tacrolimus elimination and new-onset hypertension in Chinese liver transplantation patients. Biomarkers in medicine (2017)
- GRAMD3: Genetic Determinants of "Cognitive Impairment, No Dementia”. JAD. (2012)
- Functional SNP of ARHGEF10 confers risk of atherothrombotic stroke. Hum.Mol.Gen (2010)
- *ARHGEF10* knockout inhibits platelet aggregation and protects mice from thrombus formation. Journal of thrombosis and haemostatsis. (2017)

GWAS and eQTL

- **Protein QTL based on MFI and ELISA results**
- **Association tests with clinical parameters**

- gcta64 --mlma-loco --bfile obfuscated_org --pheno 1482_creatinine.phen --covar 1482_covar.txt --qcovar HapMap_3_r3_11.ldpruned.forPCA.evec_1482_gcta_qcovar.txt --reml-maxit 2000 --out obfuscated_org_9_1482_cov_IMT

- #cp obfuscated_org_9_1482_cov_IMT.loco.mlma obfuscated_org_9_1482_cov_IMT.loco.mlma.txt
- #awk -F "\t" '{ if(($9 != "-nan") && ($2)) { print $2} }' obfuscated_org_9_1482_cov_IMT.loco.mlma.txt > extract_No_NAN_snps.txt

- #plink1.9 --bfile obfuscated_org --extract extract_No_NAN_snps.txt --make-bed --out obfuscated_org_extract_SNPs_no_nan
- #gcta64 --mlma-loco --bfile obfuscated_org_extract_SNPs_no_nan --pheno 1482_IMT.phen --covar 1482_covar.txt --qcovar HapMap_3_r3_11.ldpruned.forPCA.evec_1482_gcta_qcovar.txt --out obfuscated_org_extract_SNPs_no_nan

- #cp obfuscated_org_9_1482_cov_IMT.loco.mlma obfuscated_org_9_1482_cov_IMT.loco.mlma.txt
- #awk '{print $9}' obfuscated_org_9_1482_cov_IMT.loco.mlma.txt > pointwise-p_obfuscated.txt
- #awk 'NR > 1 { print }' pointwise-p_obfuscated.txt > pointwise-pa_obfuscated.txt
- #sort pointwise-pa_obfuscated_snps_no_nan_extract.txt > sort.txt

- #plink1.9 --bfile obfuscated_org --recodeA --out obfuscated_org
- #awk '{ $1=""; $2=""; $3=""; $4="";$5=""; $6=""; print}' obfuscated_org.raw > obfuscated_org_col6
- #cat obfuscated_org_col6 | java -jar ./transpose.jar > obfuscated_org_col6_9_input
- #cut -d" " -f2- obfuscated_org_col6_9_input > obfuscated_org_col6_9a_input
- #./slide.1.0/slide_1prep -QG obfuscated_org_col6_9a_input 1482 100 obfuscated_org_col6_9a_input_prep1482
- #plink1.9 --bfile obfuscated_org --recode 12 --out obfuscated_org
- #awk '{print $2, $3, $4}' obfuscated_org.map > obfuscated_org_234.map
- #perl maxstat_W.pl

- #cat obfuscated_org_col6_9a_input_prep1482_maxstat* > obfuscated_org_col6_9a_input_prep1482_maxstat_2000000
- #./slide.1.0/slide_3sort obfuscated_org_col6_9a_input_prep1482_maxstat_2000000_sort obfuscated_org_col6_9a_input_prep1482_maxstat_2000000
- ./slide.1.0/slide_4correct -p obfuscated_org_col6_9a_input_prep1482_maxstat_2000000_sort pointwise-pa_obfuscated.txt obfuscated_org_correct_pa obfuscated_org_234.map
- #sort pointwise-pa_obfuscated_snps_no_nan_extract.txt > sort.txt
- ####./slide.1.0/slide_4correct -ip obfuscated_org_9_extract_SNPs_no_nan_col6_9a_input_prep91_maxstat_2000000_sort 0.000100157
- sort -g -k 3 obfuscated_org_correct_pa > correct_pa_sort_all.txt
- awk -F "\t" '{print $3}' obfuscated_org_correct_pa > correct_pa.txt
- sort correct_pa.txt > correct_pa.sort.txt

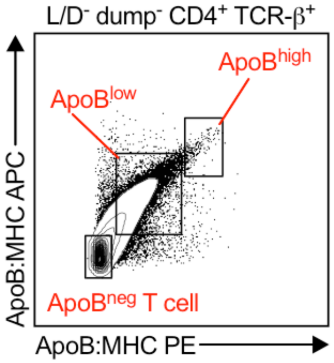
```
#Result based on 20000 sampling
#SNP_id Pointwise-P     Corrected-P     Approx.STDEV(% of P)    Effective#ofTests       Note
rs2166923       2.32072000e-10  4.50000000e-04  1.49966246e-04(33.32583%)       1939053
rs7844006       2.64789000e-10  7.00000000e-04  1.87017379e-04(26.71677%)       2643614
rs60759353      2.82568000e-10  7.50000000e-04  1.93576535e-04(25.81020%)       2654228
```

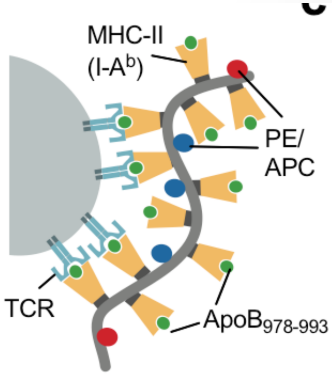
Single Cell and Bulk RNASeq

- **Single cell and bulk RNA Seq** analysis of mouse aorta and lymph nodes
- Objective: *To demonstrate the ability of ApoB Tregs to control the autoimmunity in the natural course of atherosclerosis at transcriptomic level using bulk and single cell RNA Seq*

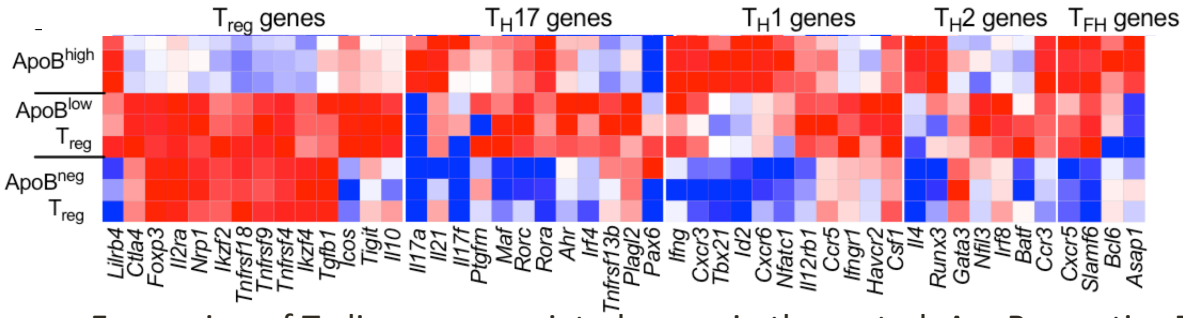
- Dextramer of recombinant mouse MHC-II fused to $ApoB_{978-993}$ that specifically bound to the TCR was used to detect ApoB-reactive T-helper cells ($ApoB^{+}$) from lymph nodes in flow cytometry in atherosclerosis-prone **8-week-old** *Apoe*$^{-/-}$ and *Ldlr*$^{-/-}$ mice

- $ApoB^{low}$ and $ApoB^{high}$ T cells distinguished by ApoB:MHC-staining in flow cytometry

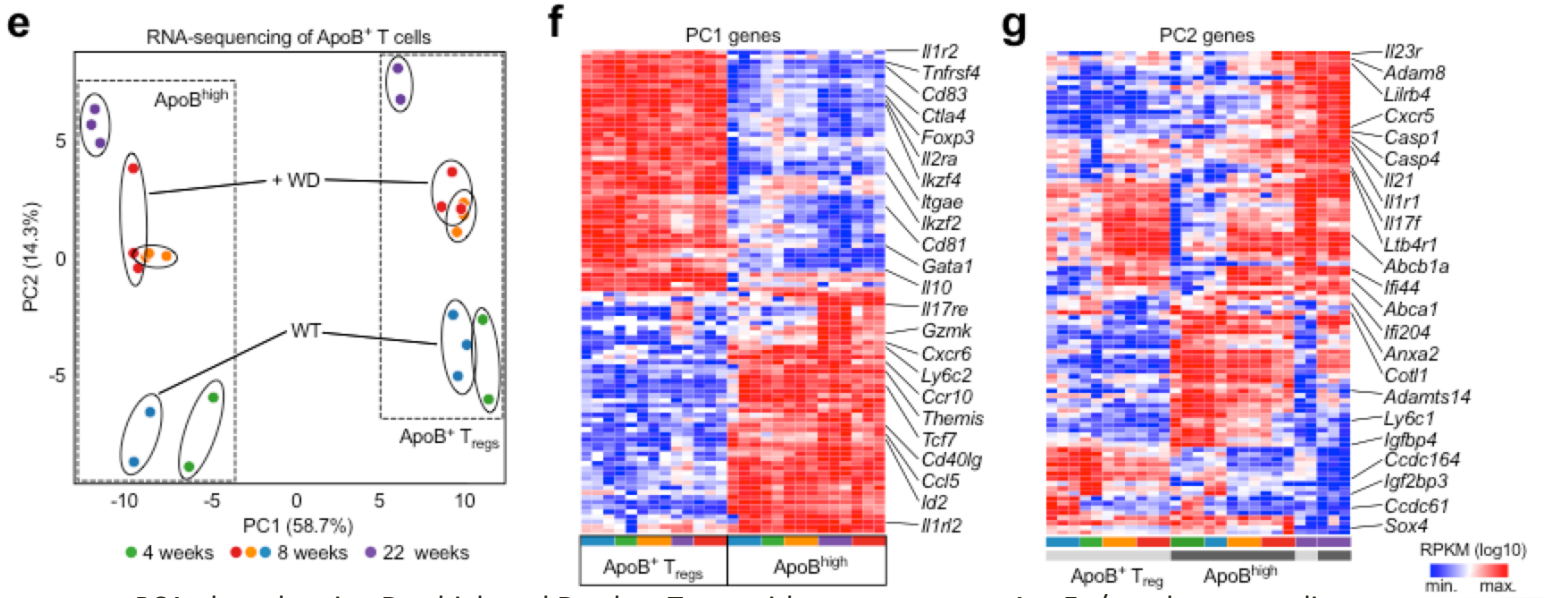




Single Cell and Bulk RNASeq

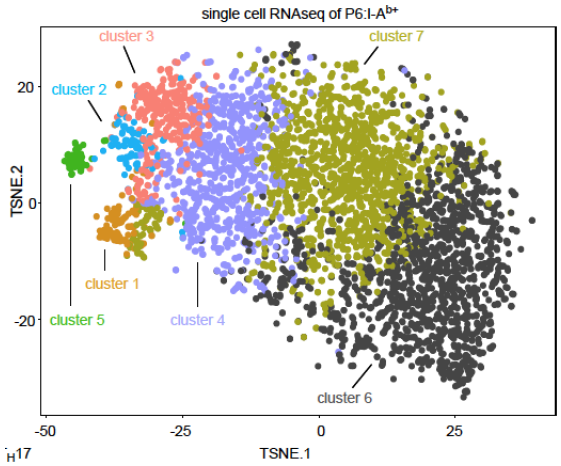


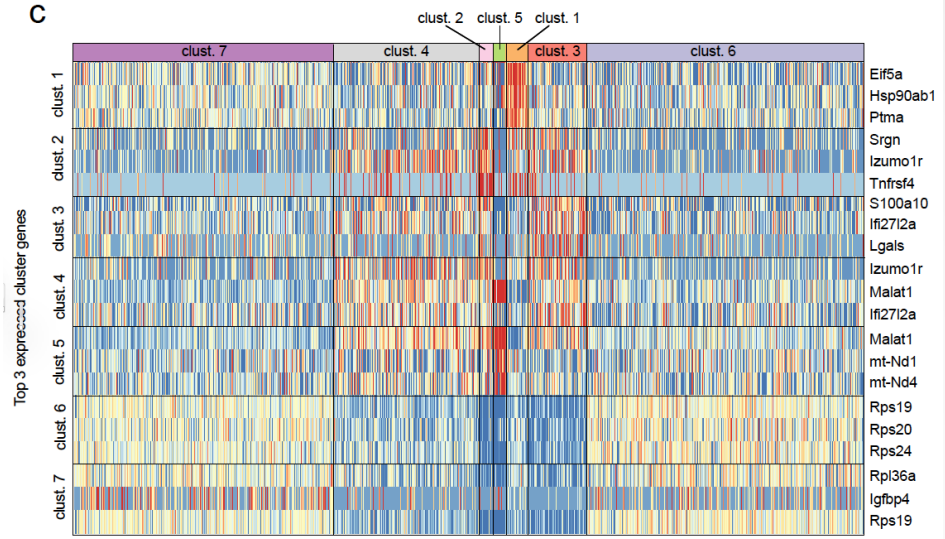
- Expression of T_H-lineage associated genes in the sorted ApoB reactive T cells after *Bulk RNA Sequencing*



- PCA plots showing Dex high and Dex low Tregs with respect to age, ApoE -/- and western diet
- ApoB^{low} T_{regs} and ApoB^{high} cells were distinguishable by core T_{reg} genes, such as *Foxp3*, *Il2ra*, *Ctla4*, and *Itgae*
- Up regulation of *Il23r*, *Il21*, and *Il17f* in 22-week-old mice and WD-fed 8-week-old mice in both ApoB$^+$ T cell subsets, indicative of a T_H17 phenotype that developed during lesion formation

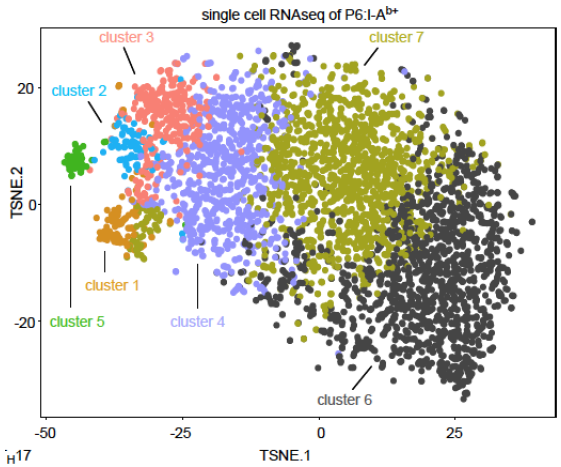
Single Cell and Bulk RNASeq

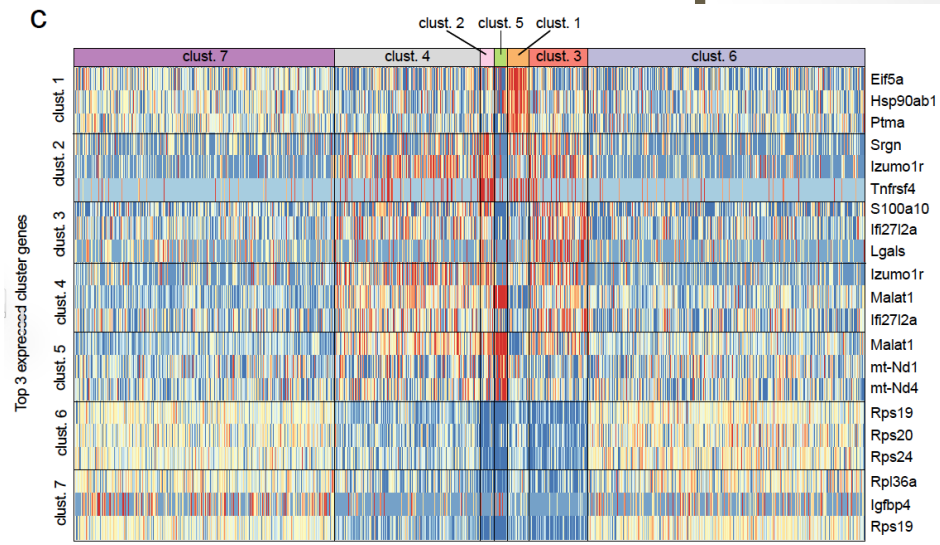


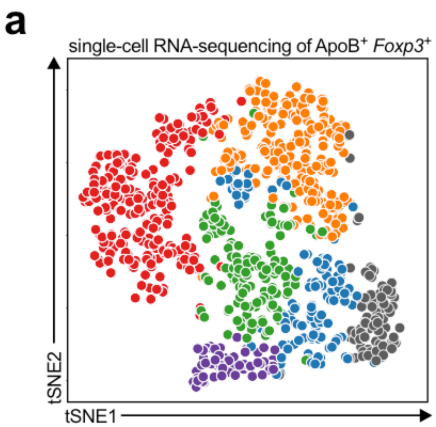


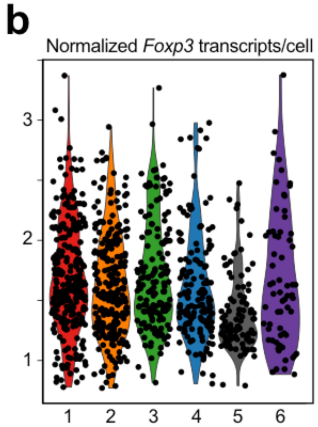
- ./cellranger-1.3.1/cellranger count --id=9_13 --fastqs=./A9_S5 --transcriptome=./refdata-cellranger-mm10-1.2.0/ --sample=A9 −nopreflight
- Outs directory: analysis and gene_bc_matrices directories, cloupe file, web_summary.html
- Analysis directory: clustering, tsne, pca, diffexp
- From gene_bc_matrices for Cell ranger R kit and Seurat package
- ./cellranger-1.3.1/cellranger reanalyze --id=KL_HFDA_13_WT_40_kmeans_40B --matrix=/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/KL_HFDA_13_WT_40/outs/filtered_gene_bc_matrices_h5.h5 --params=./HFD_kmeans.csv --agg=./HFD.csv
- head HFD.csv
- library_id,molecule_h5
- KL_HFDA_13_WT_40,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/KL_HFDA_13_WT_40/outs/molecule_info.h5
- head HFD_kmeans.csv
- max_clusters,40
- ./cellranger-1.3.1/cellranger aggr --id=7X_sec --csv=./7X.csv --normalize=mapped −nosecondary
- library_id,molecule_h5
- DW-A1,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-A1/outs/molecule_info.h5
- DW-B1,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-B1/outs/molecule_info.h5
- DW-C1,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-C1/outs/molecule_info.h5
- DW-A3,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-A3/outs/molecule_info.h5
- DW-B3redo,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-B3redo/outs/molecule_info.h5
- DW-C3,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-C3/outs/molecule_info.h5
- DW-2redo,/mnt/overflow/BioAdHoc/Groups/KLey/pramod/v1.3redo/DW-2redo/outs/molecule_info.h5

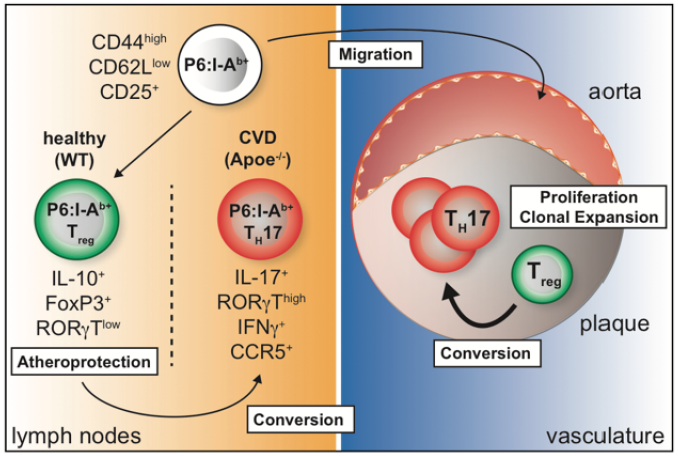
Single Cell and Bulk RNASeq











- P6:I-Ab+ CD4+ T cells flow sorted and single cell sorted and sequenced by 10x Genomics drop-sequencing workflow
- 4485 single cell libraries were successfully generated.
- Gene expression was mapped on the mm10 transcriptome using the CellRanger pipeline.
- Gene expression clustering after dimensionality reduction on 10-k means clustering in t-SNE
- Heatmap generated using cell ranger R kit
- FoxP3 cells sorted based on normalized expression and clustered by SEURAT

Bulk RNASeq

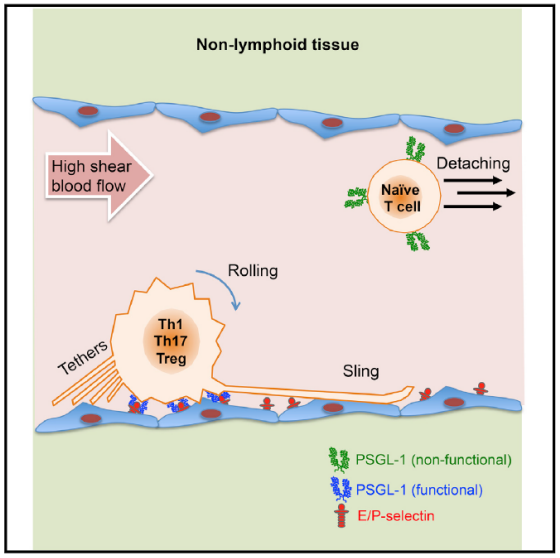
- Objective: *To demonstrate homing of T cells to peripheral tissues/ inflammation sites involves formation of* ***tethers and slings*** *enabled by the transcriptomic changes during differentiation*
- **Bulk RNA Seq** analysis of mouse and human neutrophils and T cells

- *Tethers and slings, discovered in neutrophils, facilitate cell rolling at high shear stress. (Nature. 2012)*

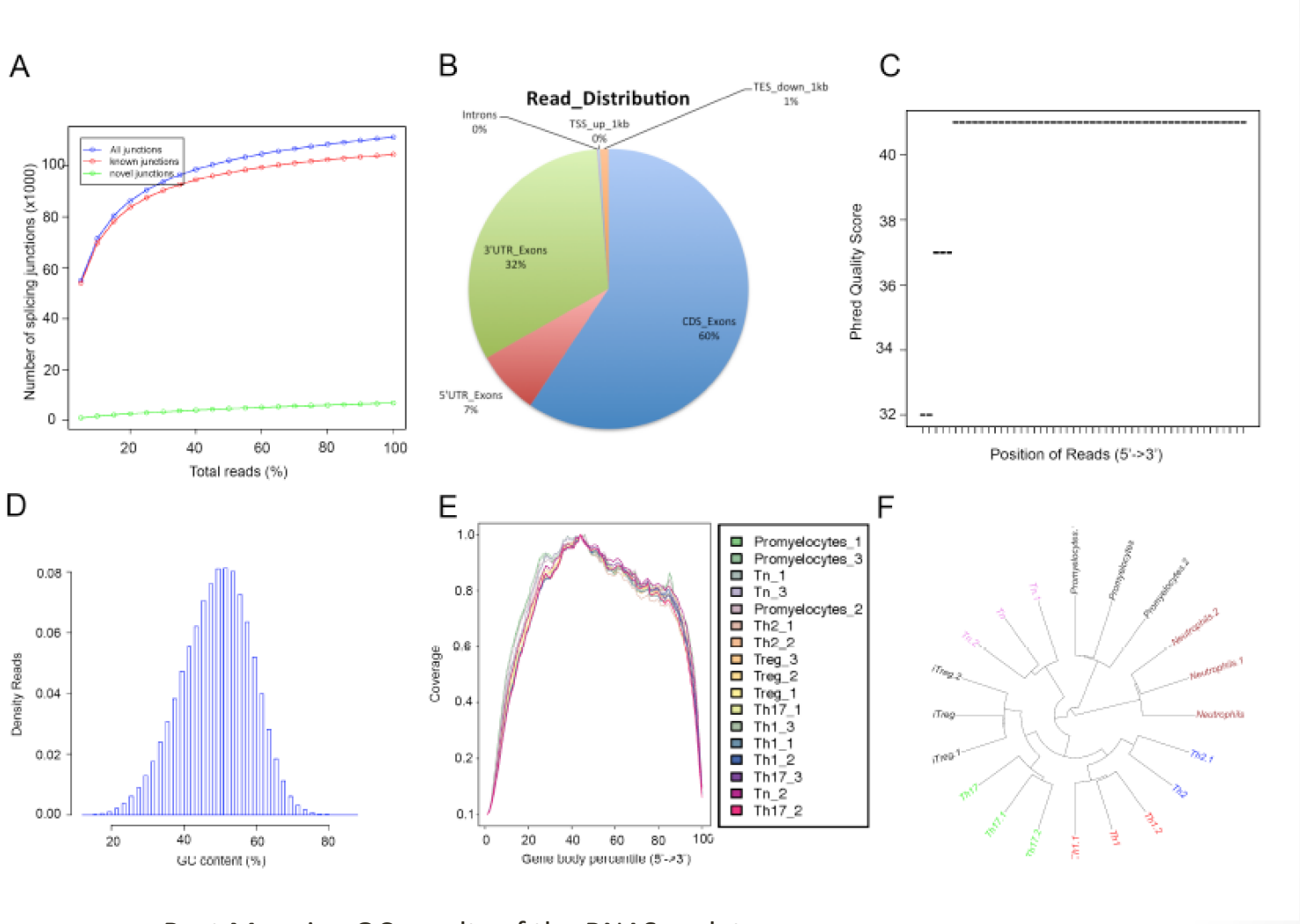
- The isolation of mRNA was done using the polyA method.

- DESeq2 (R package) was used to perform the differential expression.
  DE genes below adjusted p value of <0.01.

- Disease and biological functions of DE genes revealed by pathway analysis (Qiagen IPA)



Bulk RNASeq

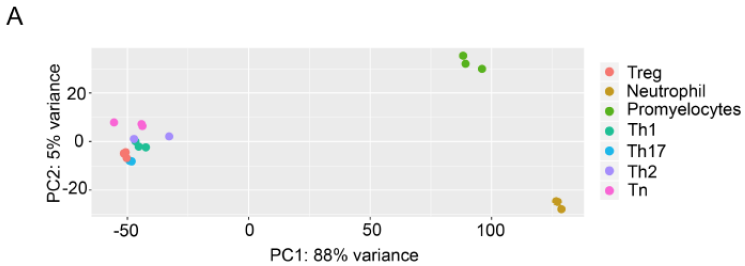


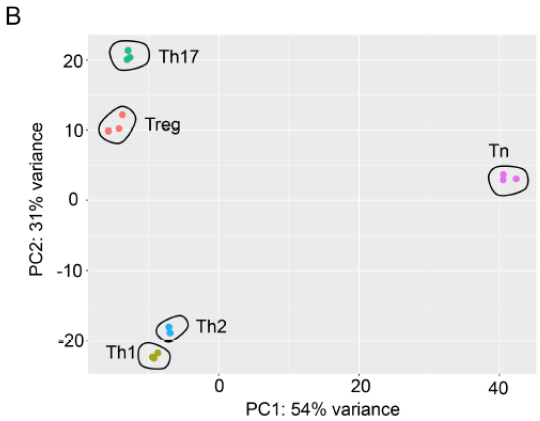
- Post Mapping QC results of the RNASeq data
- Mapping was done by TopHat2 to mm10 UCSC genome
- RSeQC was used for evaluating mapping quality
- Uniform gene body coverage with no RNA degradation

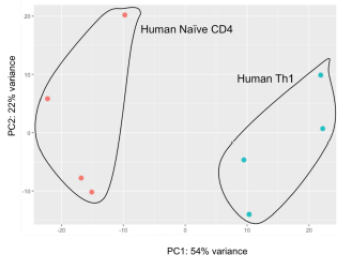
CD4 T cell subsets

Subset	Transcription Factor	Cytokines
Th1	T-bet	IFN-γ
Th2	GATA-3	IL-4, 5, 13
Th17	ROR-γt	IL-17A, F
TFH	Bcl6	IL-21
Treg	FoxP3	TGF-β, IL-10
Tr1	unknown	IL-10

Bulk RNASeq





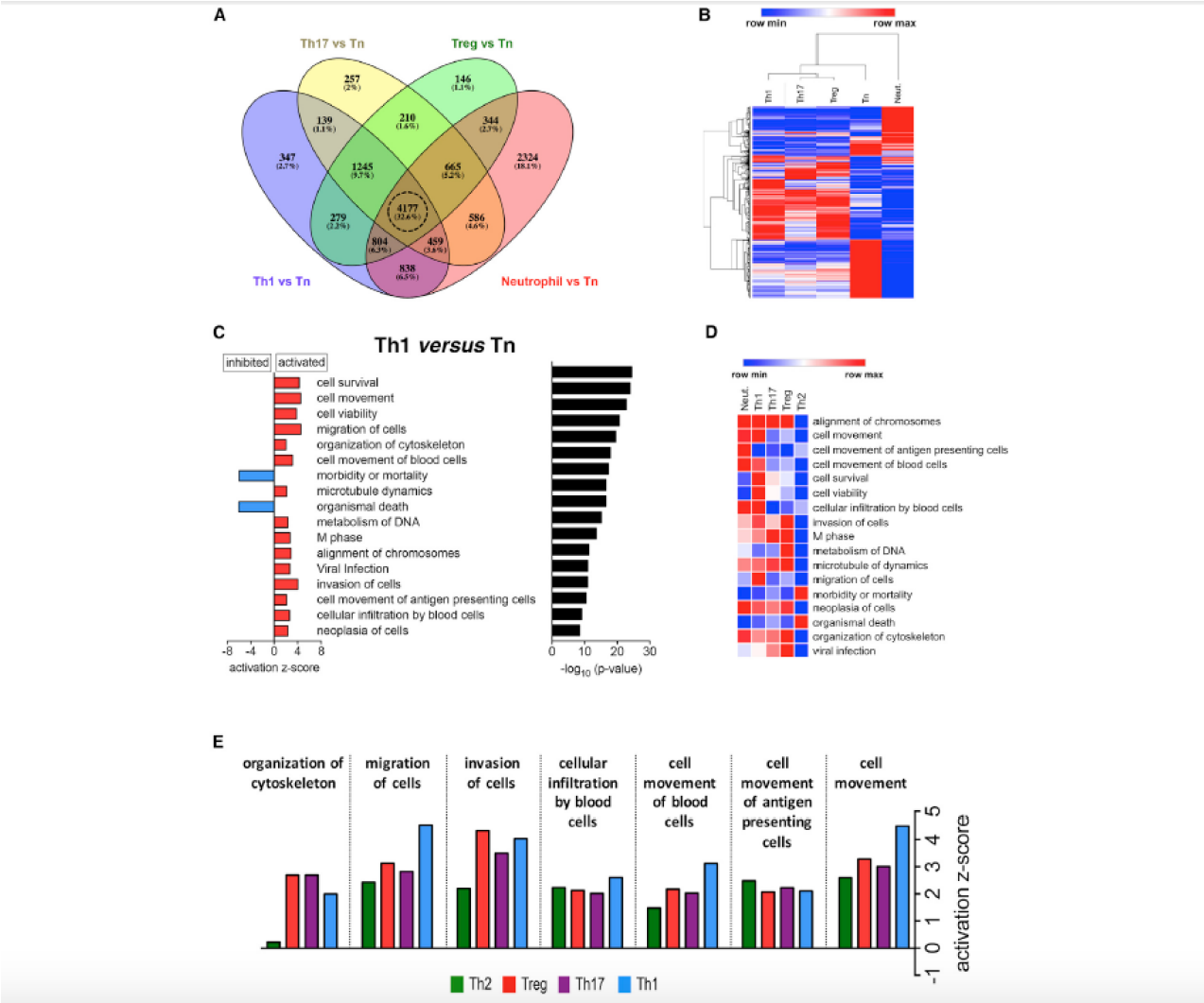


- PCA plots demonstrate the grouping of naïve T cells and differentiated T cells in mouse and human samples

Bulk RNASeq

```
Counts_M8<-read.table("MA-counts_KY.txt", header = T, row.names = 1)
head(Counts_M8)
condition<-rep(c("Th1","Th17","Treg","NaiveT"), c("3","3","3","3"))
condition
colnames<-colnames(Counts_M8)
samplenames <-colnames(Counts_M8)
coldata<-data.frame(condition,colnames,samplenames)
coldata
dds_M8<-DESeqDataSetFromMatrix(countData =Counts_M8, colData = coldata, design = ~condition )
nrow(dds_M8)
keep<-rowSums(counts(dds_M8) >7)>=1
dds_M8<-dds_M8[keep,]
nrow(dds_M8)
dds_M8_counts<-assay(dds_M8)
write.table(dds_M8_counts,"dds_M8_counts_KY.xls",sep="\t",)
rld_M8 <- rlog(dds_M8)
rld_M8_LF<-assay(rld_M8)
write.table(rld_M8_LF, "rld_M8_LF_7_1_KY.xls",sep="\t",)
plotPCA(rld_M8)
de_dds_M8 <- DESeq(dds_M8)
(res_M8 <- results(de_dds_M8))
baseMean_S8 = counts(de_dds_M8, normalized = TRUE)
write.table(baseMean_S8, "baseMean_Mgrant_7_1_KY.xls", sep="\t",)
res_M8_Th1_Th17<-results(de_dds_M8, contrast=c("condition", "Th1", "Th17"))
res_M8_Th1_Treg<-results(de_dds_M8, contrast=c("condition", "Th1", "Treg"))
res_M8_Th1_NaiveT<-results(de_dds_M8, contrast=c("condition", "Th1", "NaiveT"))
res_M8_Th17_Treg<-results(de_dds_M8, contrast=c("condition", "Th17", "Treg"))
res_M8_Th17_NaiveT<-results(de_dds_M8, contrast=c("condition", "Th17", "NaiveT"))
res_M8_Treg_NaiveT<-results(de_dds_M8, contrast=c("condition", "Treg", "NaiveT"))
write.table(res_M8_Th1_Th17, "res_M8_Th1_Th17_KY.xls", sep="\t",)
write.table(res_M8_Th1_Treg, "res_M8_Th1_Treg_KY.xls", sep="\t",)
write.table(res_M8_Th1_NaiveT, "res_M8_Th1_NaiveT_KY.xls", sep="\t",)
write.table(res_M8_Th17_Treg, "res_M8_Th17_Treg_KY.xls", sep="\t",)
write.table(res_M8_Th17_NaiveT, "res_M8_Th17_NaiveT_KY.xls", sep="\t",)
write.table(res_M8_Treg_NaiveT, "res_M8_Treg_NaiveT_KY.xls", sep="\t",)
```

Bulk RNASeq



- RNA-Seq Reveals Activation of Cell Migration-Related Genes in Sling-Forming Cells
- DESeq2 DE Analysis between neutrophils, naïve T cells, Th1, Th17 and Tregs
- IPA on the DE genes to find the cellular processes involved in the cell functionality
- Cell specific gene signatures

Bioinformatics Projects

- Objective: *To translate **macrophage gene signature** for LPS responsiveness based on* Interleukin-12β and arginase expression ratio (Polarization Factor Ratio) from HMDP (83 inbred strains) for diagnosis and survival prediction of macrophage related human immunological diseases in clinical settings

- **Microarray** analysis of peritoneal macrophages from hybrid mouse diversity panel (HMDP) consisting of 83 inbred strains

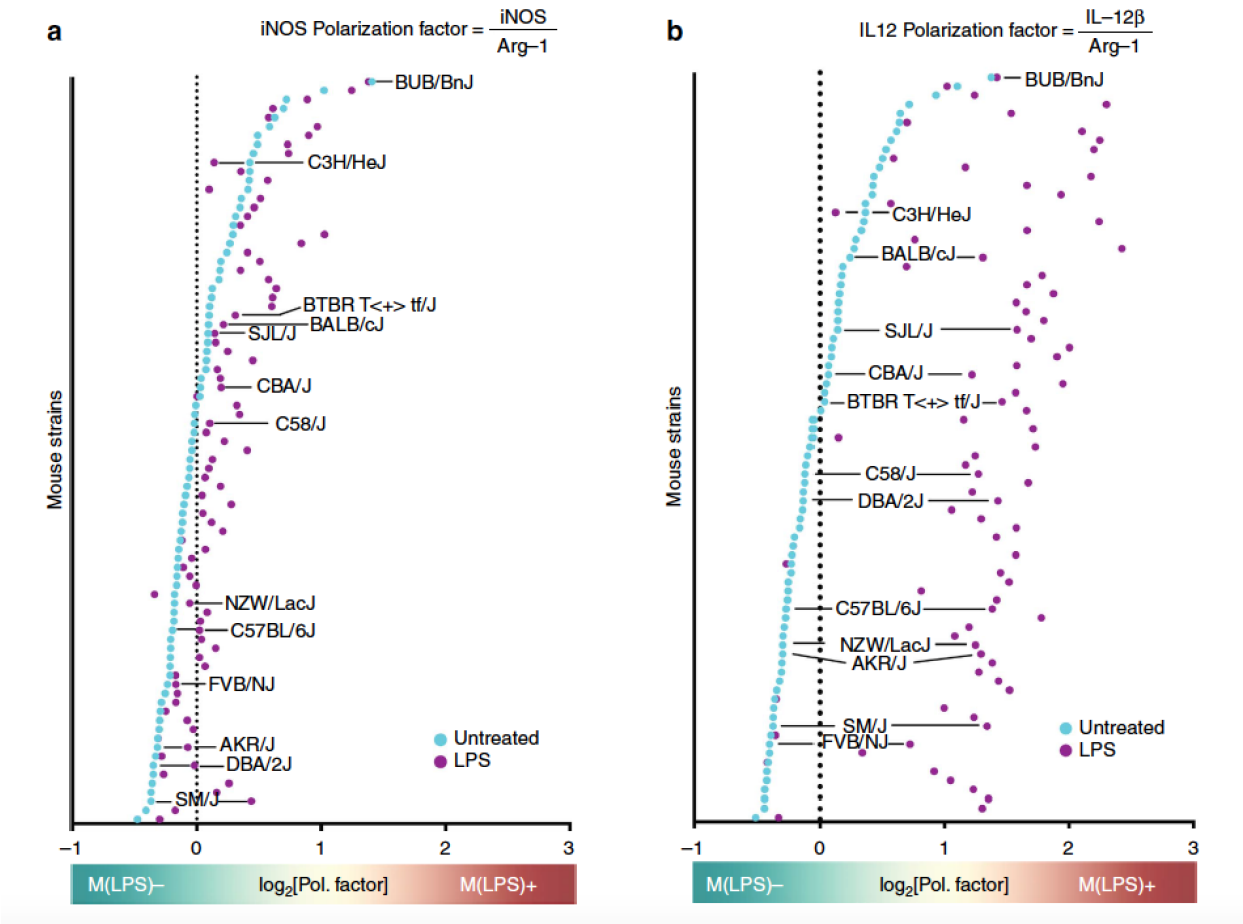
- HMDP acts as **surrogate** for human natural immune variation

- Each transcript of the Affymetrix chip (was correlated to the PFR) in 83 mouse strains.

- Positive and negative correlations with $P < 0.0001$ (Pearson) and FDR < 5% were selected as M(LPS)+ and M(LPS)- gene signature lists, respectively

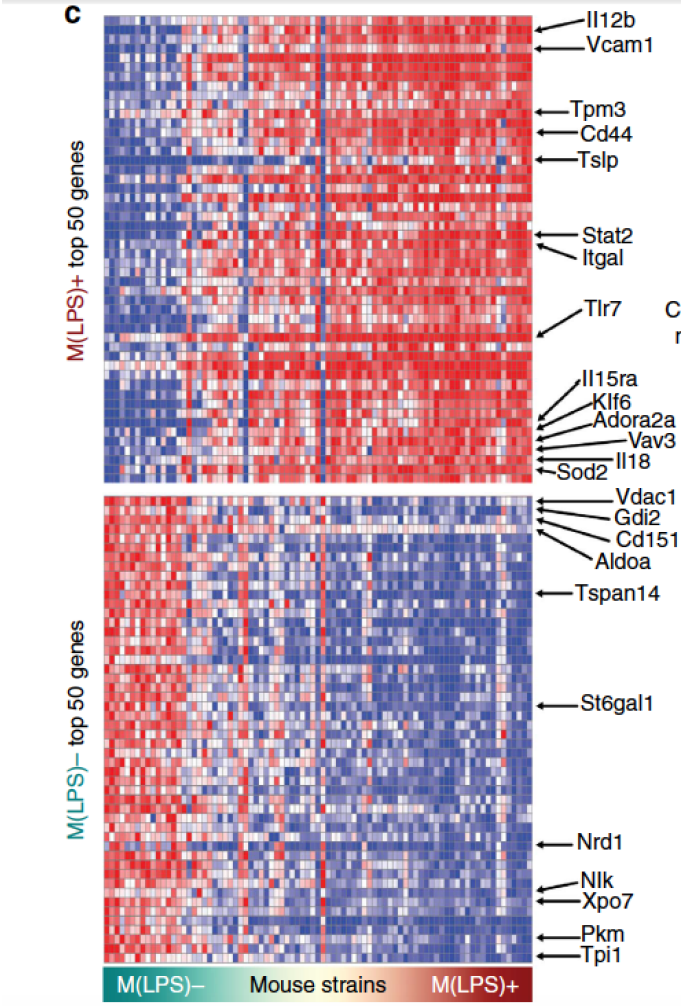
- Details on the HMDP data is accessible online (http://systems.genetics.ucla.edu/data)

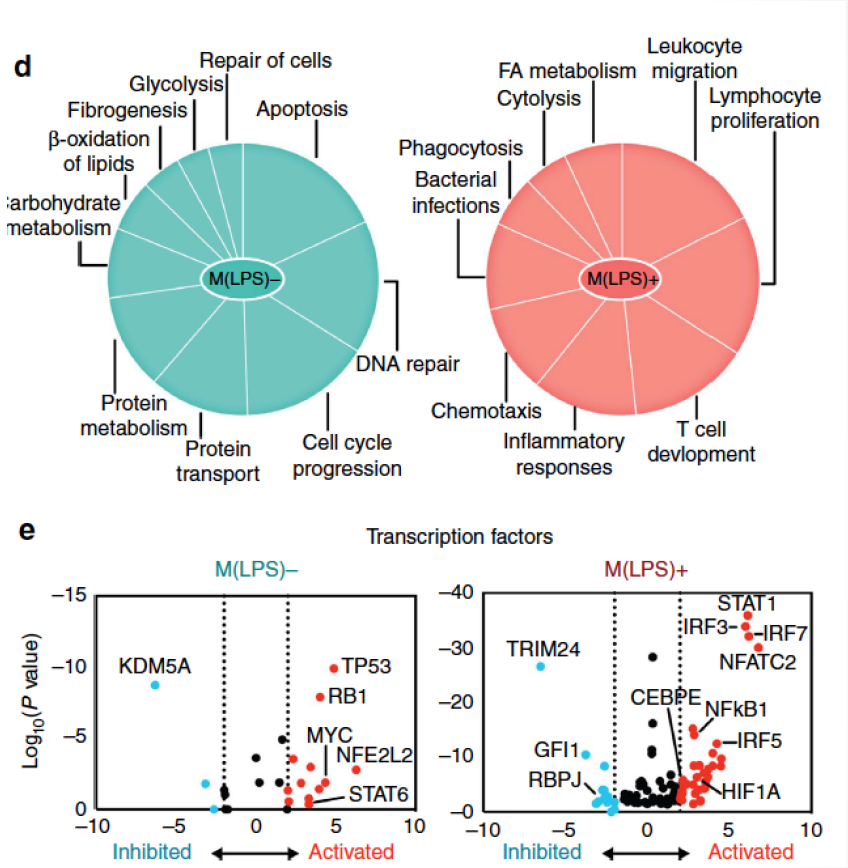
Microarray



- Inter-strain differences of LPS-induced macrophage activation across 83 mouse strains of the HMDP
- HMDP without (baseline, cyan) and with LPS (purple) treatment

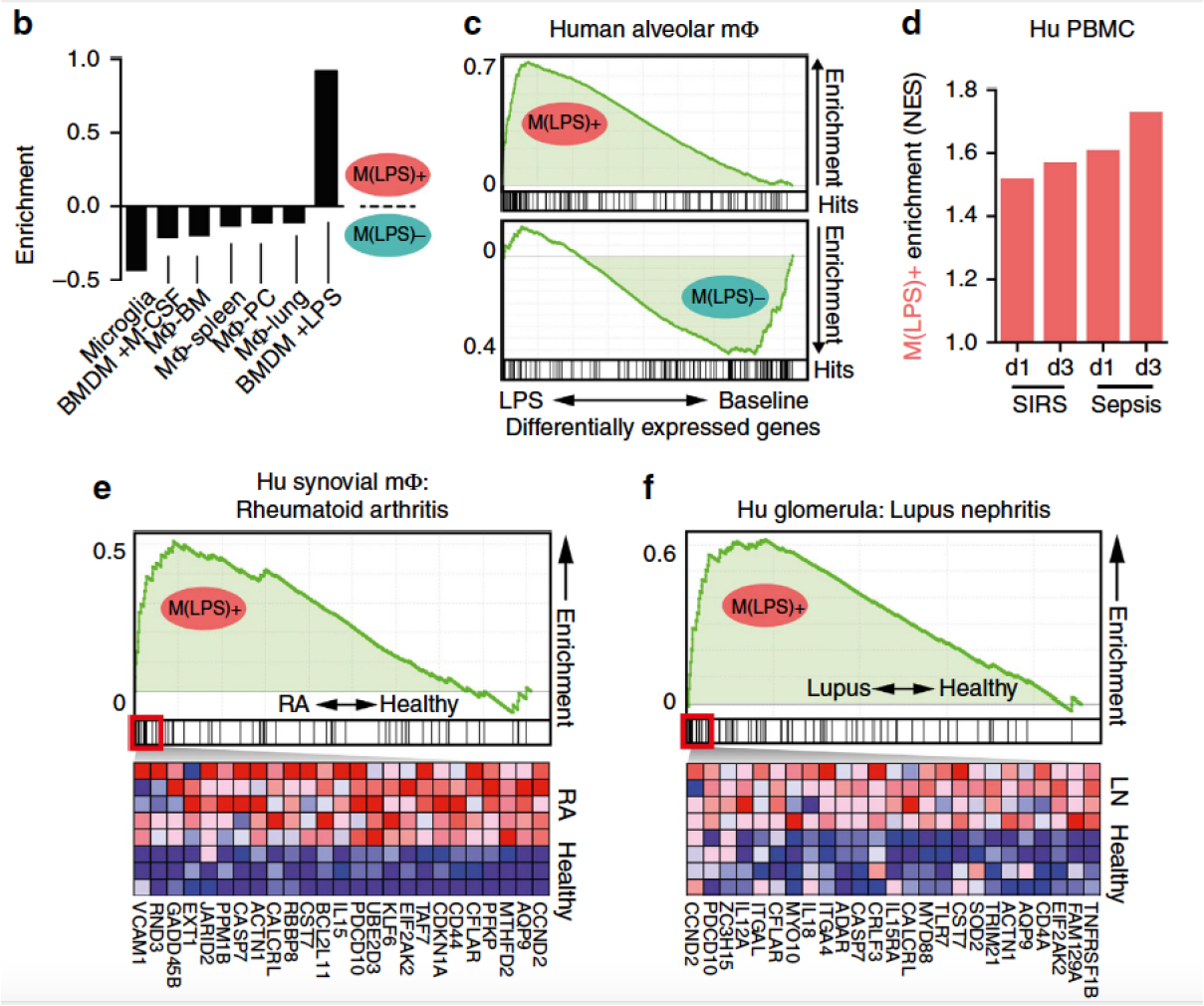
Microarray





- Core gene signatures and functional programs of LPS-responder and non-responder macrophages.
- Heatmaps for top 50 M(LPS)+ (top:c) and M(LPS)-(bottom:d) genes.
- Key biological processes enriched and TF's in M(LPS)+ and M(LPS)-gene signatures as determined by IPA

Microarray



- Macrophage activation phenotypes in mice and humans under homoeostatic and disease conditions
- Genesets: top 200 genes from the macrophage signatures
- Data from GSE15907, GSE40885, GSE13904, GSE10500 and GSE32591

Konrad, B., Ehinger, E., Gupta, P., Pramod, A.B., eta al., Nature Communications (2017)