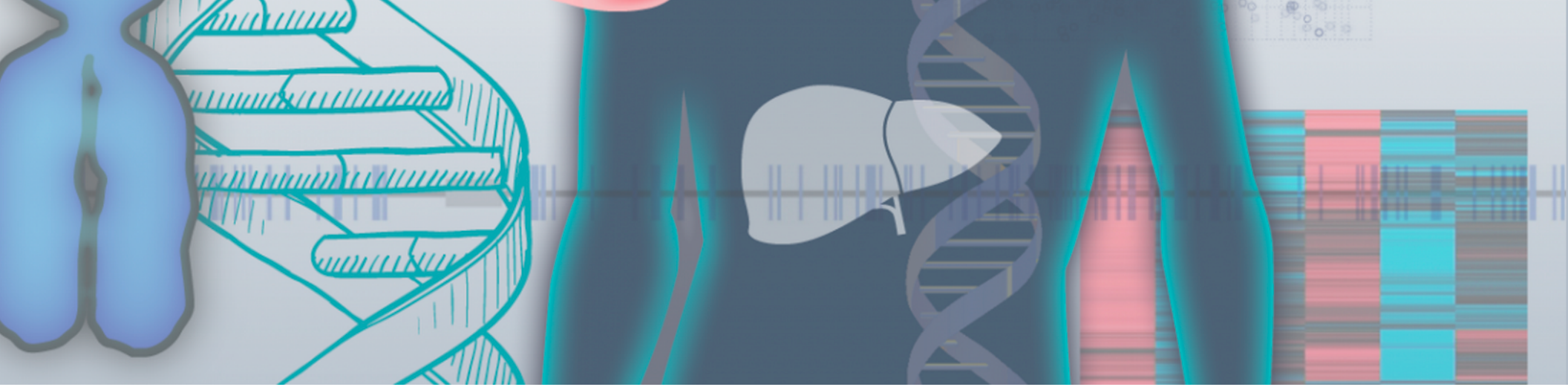


TOPMed multi-omics: transcriptomics in the MESA cohort

François Aguet

MESA Steering Committee Meeting :: 03/29/2018

BROAD INSTITUTE

NHLBI TOPMed

MESA

UNIVERSITY OF WASHINGTON 1861

The TOPMed MESA RNA-seq pilot

Exam 1 (baseline)
July 2000 - July 2002
1163 participants

- PBMCs

Exam 5
2010 - 2012
966 participants

- PBMCs
- Sorted cells (subset)
  - T cells
  - Monocytes

The TOPMed RNA-seq pilot:

Establish robust, multi-center transcriptomics pipeline and assess feasibility in the MESA cohort, with the goal to reproducibly detect changes in gene expression that are correlated with other phenotypes.

The TOPMed MESA RNA-seq pilot

Cohort	Broad	NWGC	Total
PBMC (Exam 1)	580	583	1163
PBMC (Exam 5)	498	468	966
T cell (Exam 5)	201	204	405
Monocyte (Exam 5)	202	199	401
PBMC (Control)	15	15	30
Whole Blood (Exam 5)	4	4	8
Total	1500	1473	**2973**

- **Pre-pilot:** 55 samples at each sequencing center
  - MESA samples: 20 Exam 5 PBMC, 16 Exam 1 PBMC, 4 Exam 1 Whole Blood
  - 15 control samples, replicating Exam 1 conditions
- **MESA cohort:** 1349 participants
  - 1195 with genotype in Freeze 5 VCF

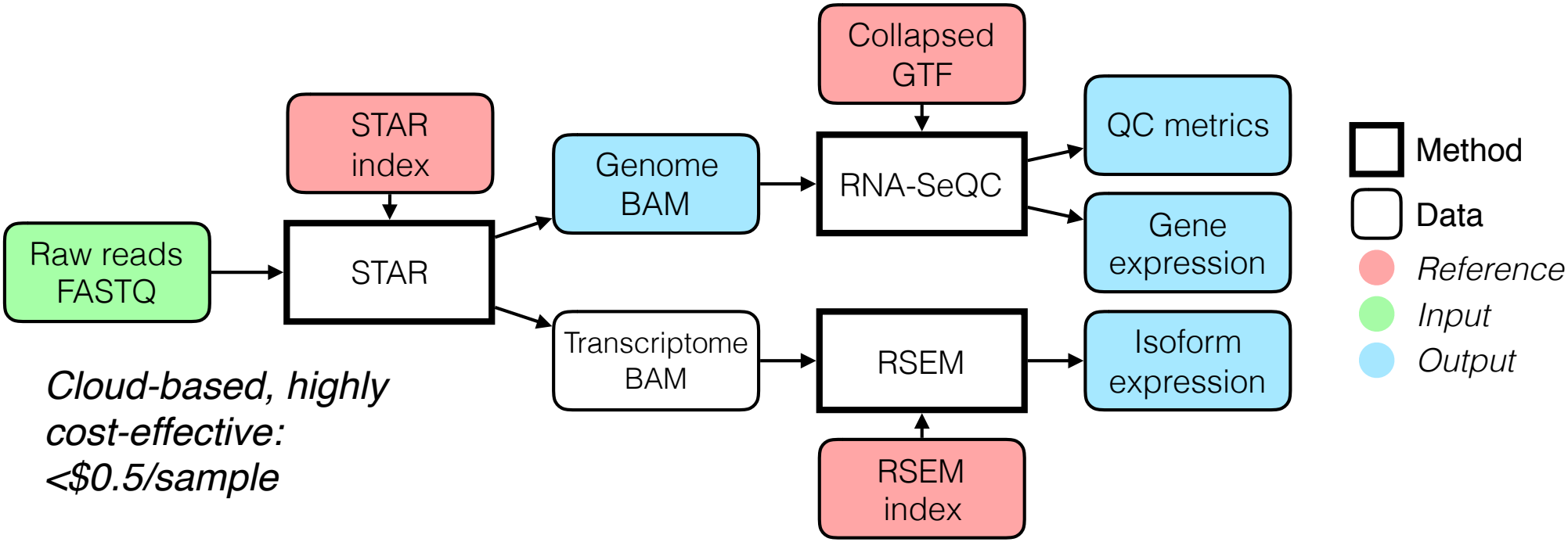
Harmonized pipelines to minimize batch effects

- **RNA-seq is highly sensitive to technical variation**
  - Experimental design: samples from the same participant sequenced at same center; within- and cross-center replicates were included
- **Sequencing protocols matched to the extent possible**
  - All samples were sequenced on HiSeq 4000 to ≥ 40M reads
- **Identical data processing pipelines**
  - Based on methods and guidelines initially developed for the GTEx Consortium
  - Implemented at both sequencing centers

’t Hoen et al., *Nat. Biotech.*, 2013

Harmonized processing pipeline

Developed and based on benchmarks for the GTEx Consortium



Cloud-based, highly cost-effective: <$0.5/sample

Pipeline components

Alignment	STAR v2.5.3a
Gene expression	RNA-SeQC v1.1.9
Transcript expression	RSEM v1.3.0
QC metrics	RNA-SeQC v1.1.9

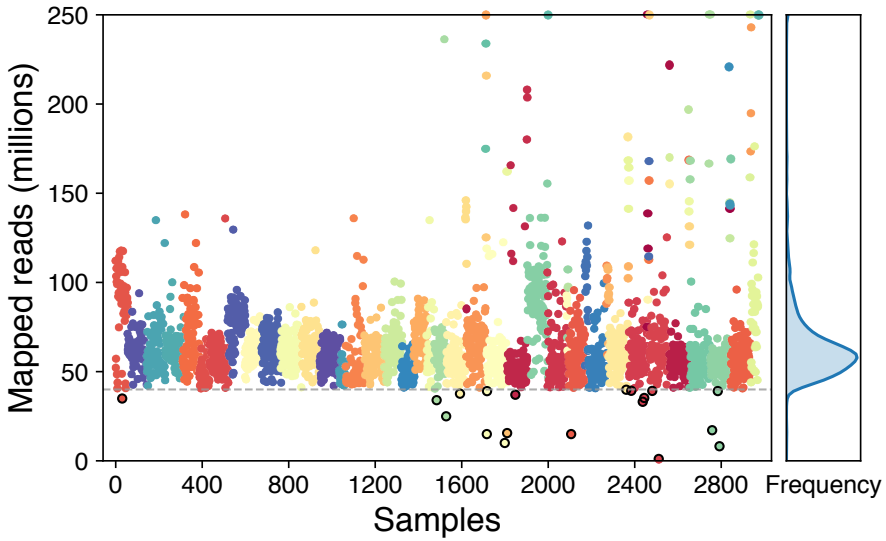
References

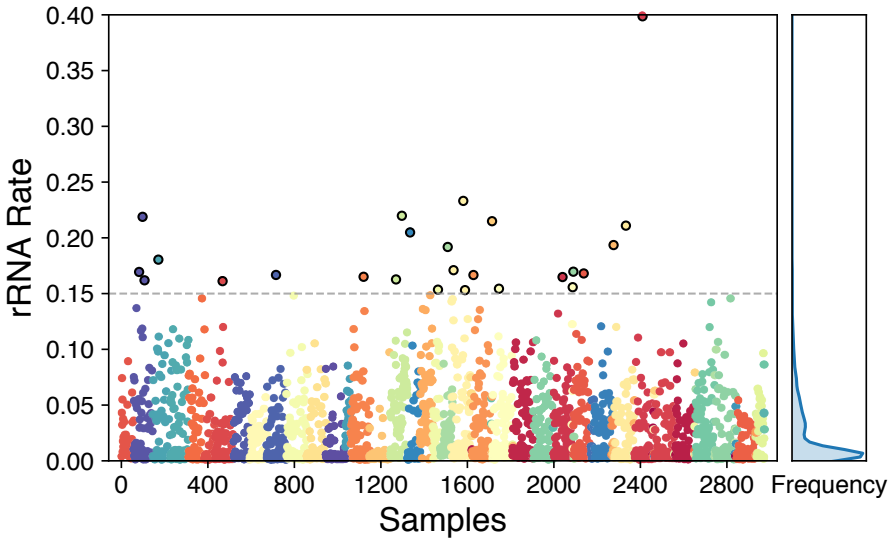
- GRCh38
- GENCODE v26

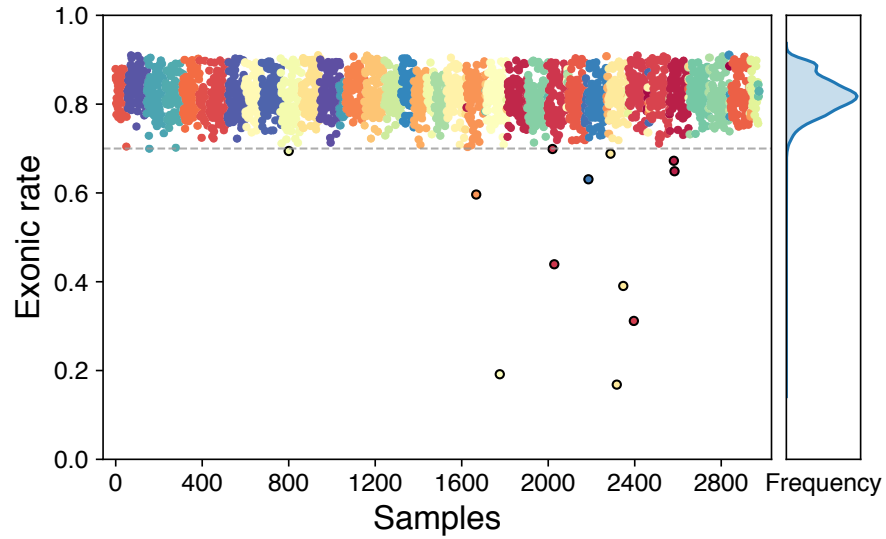
www.firecloud.org : namespace broadinstitute_gtex
https://github.com/broadinstitute/gtex-pipeline
- see TOPMed_RNAseq_pipeline.md

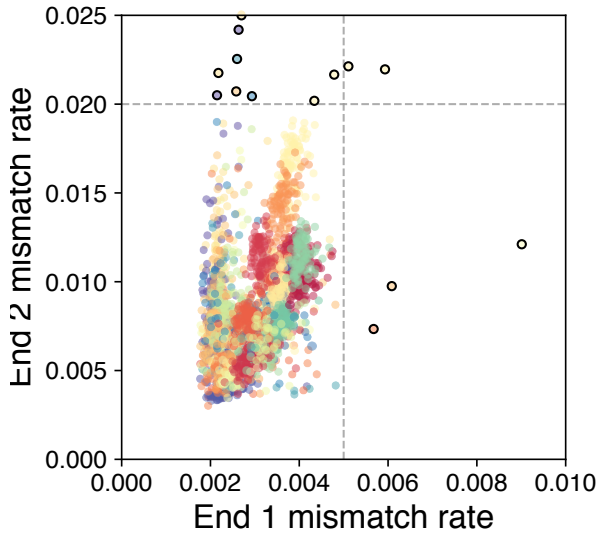
STAR: Dobin et al., *Bioinformatics*, 2013
RSEM: Li et al., *Bioinformatics*, 2010
RNA-SeQC: DeLuca et al., *Bioinformatics*, 2012

Post-sequencing quality control of RNA-seq samples



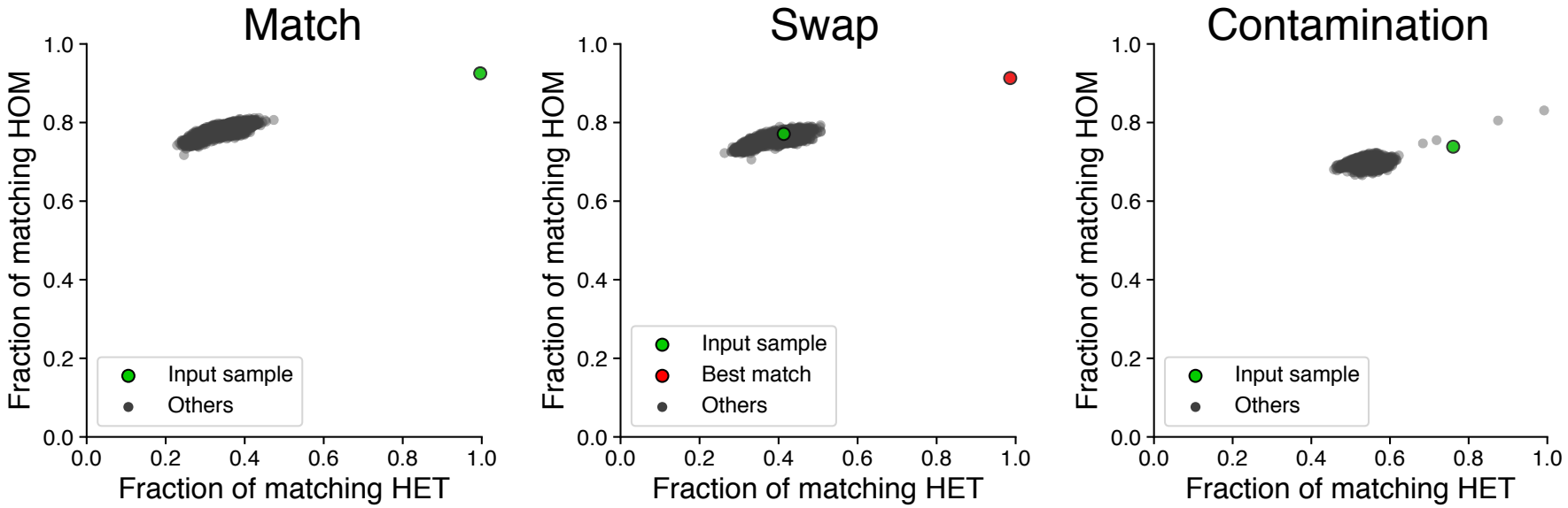






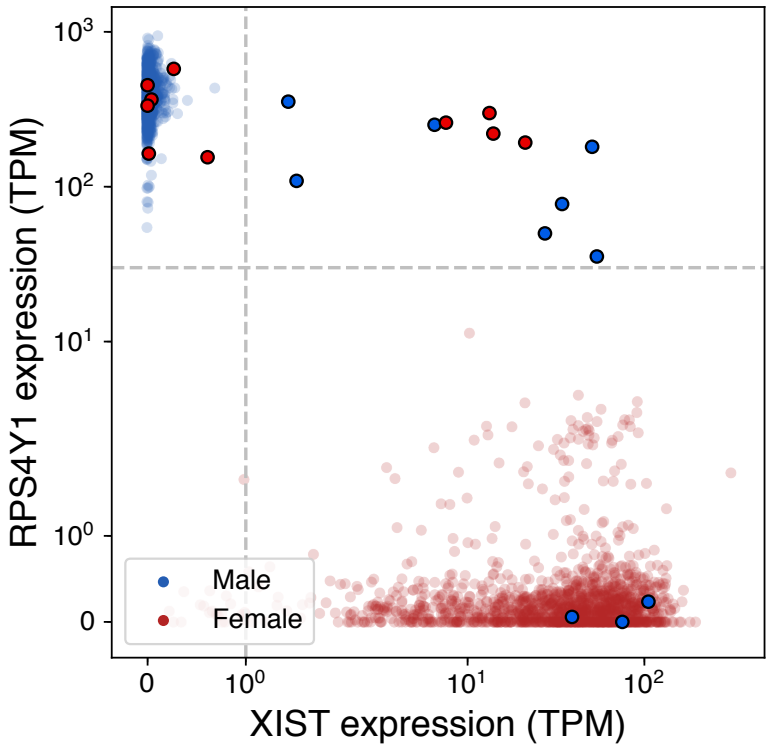
- Highly similar across sequencing centers
- Coverage goal of 40M reads exceeded for most samples

Post-sequencing quality control of RNA-seq samples



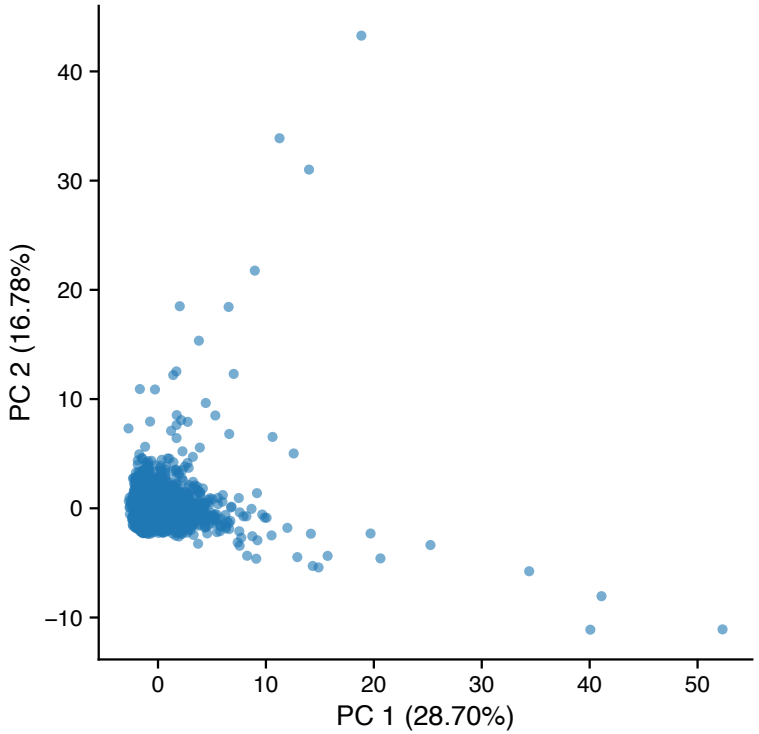
Sample swap detection

- **Fingerprinting:** concordance of heterozygous and homozygous sites between RNA-seq samples and VCF (43 mismatches)
- Expression-based sex check (20 mismatches)

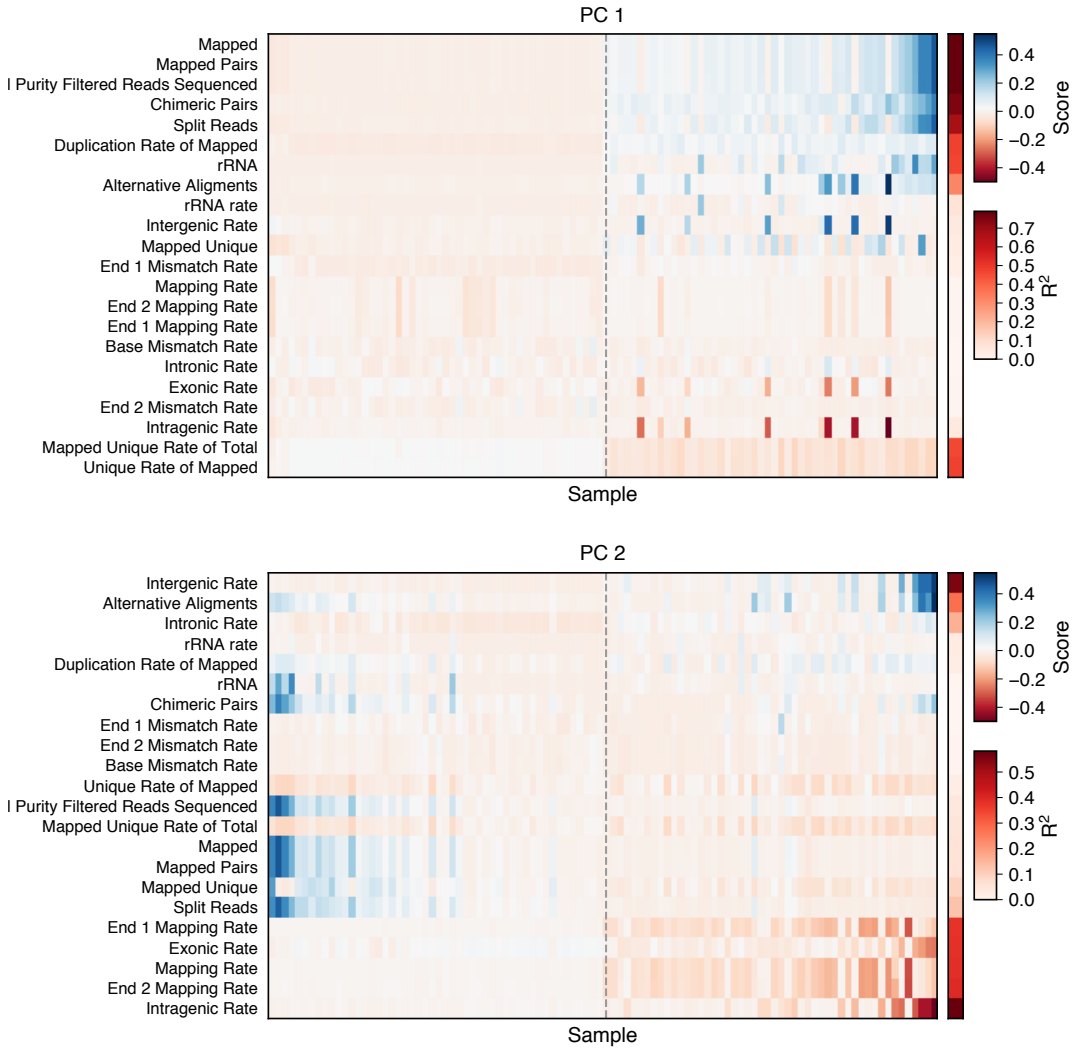


Identification of low-quality samples based on sequencing metrics

Normalized metrics

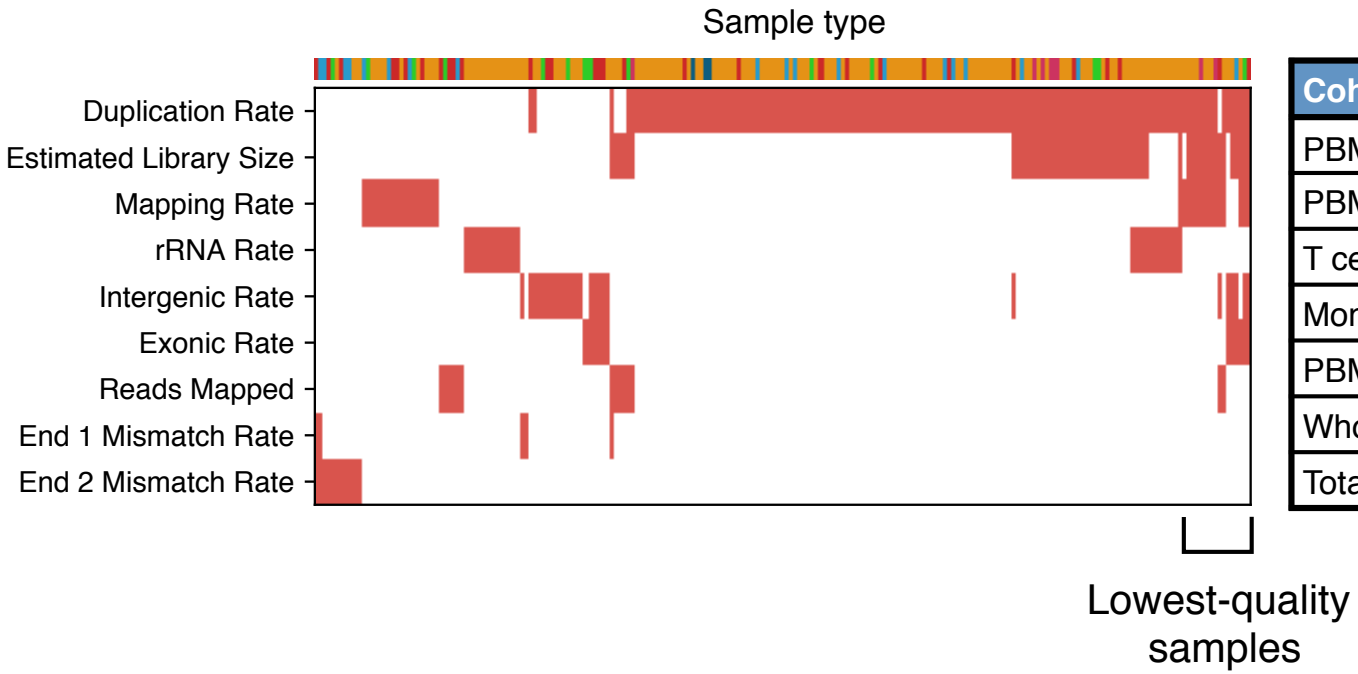


Association of PCs with metrics



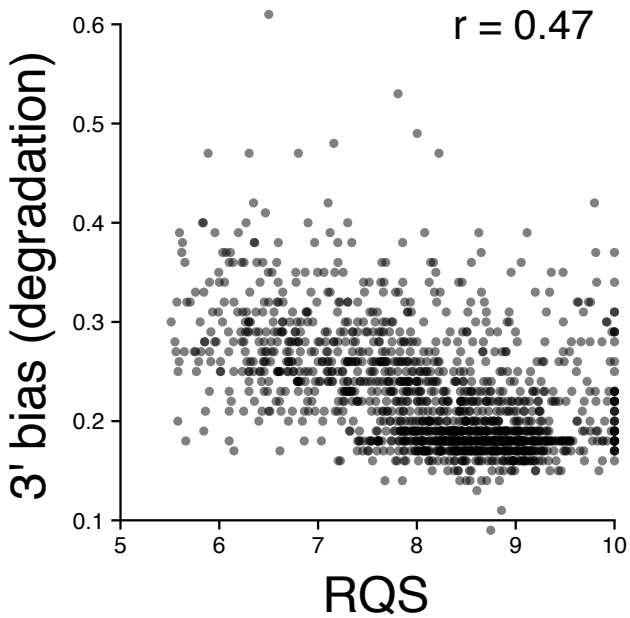
- Lowest-quality samples are outliers for multiple QC metrics

Identification of low-quality samples based on sequencing metrics



Cohort	Samples	Failed	%
PBMC (Exam 1)	1163	145	12.5
PBMC (Exam 5)	966	31	3.2
T cell (Exam 5)	405	14	3.5
Monocyte (Exam 5)	401	19	4.7
PBMC (Control)	30	3	10.0
Whole Blood (Exam 5)	8	7	87.5
Total	**2973**	**219**	**7.4**

- Small set of poor quality samples failed multiple metrics
- RNA quality: lab metrics weakly correlated with sequencing metrics
- Stringent thresholds applied to define high-quality analysis set
- Recommendation to exclude 219 samples



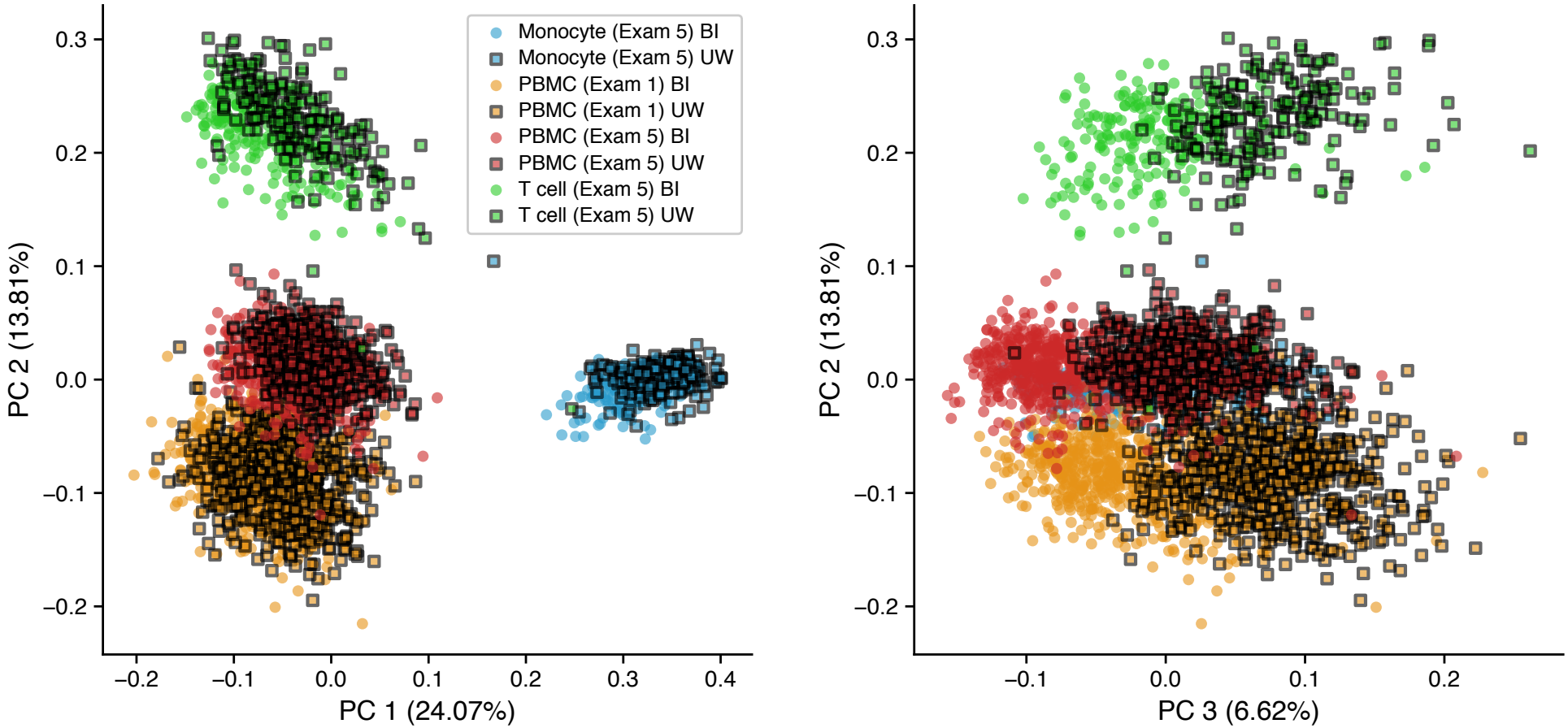
Analysis freeze (preliminary)

Cohort	PBMC Exam 1	PBMC Exam 5	T cell Exam 5	Monocyte Exam 5	Total
All samples	**972**	**916**	**385**	**375**	**2648**
With genotype (freeze 5)	863	811	343	336	2353

PBMC Exam 1	PBMC Exam 5	T cell Exam 5	Monocyte Exam 5	Total
●	●	●	●	**198**
	●	●	●	**331**
●	●			**608**

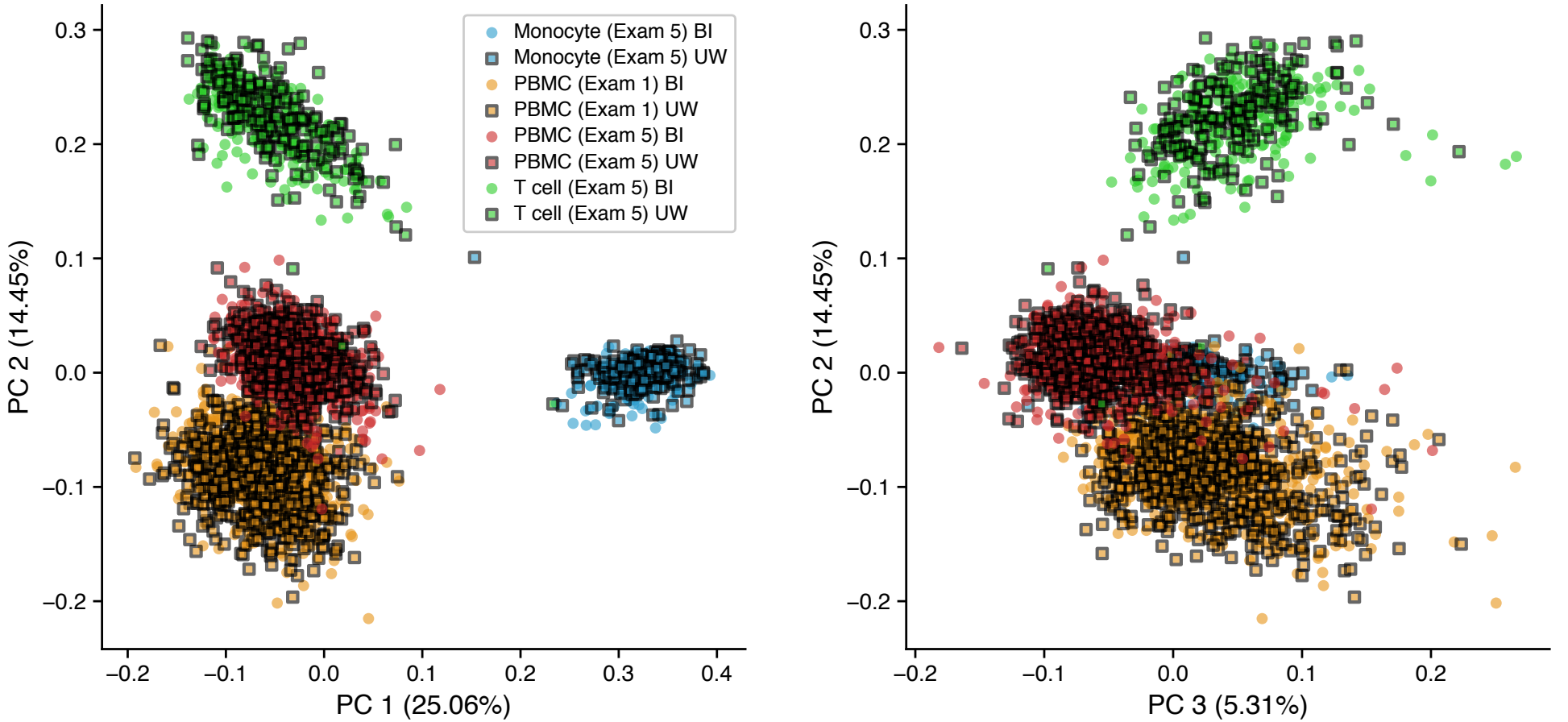
- Combination of QC steps:
  - Technical (QC metrics)
  - Expression outliers (sex check)
  - Unresolved swaps
- Freeze 5: 1149 participants with genotype data

Expression variation across sample types



- Sequencing center batch effect observed, but weak as a result of harmonization
- **Confounders: RNA isolation method, blood draw tube, and exam are linked:**
  - Trizol / Citrate CPT / Exam 1
  - All prep / Heparin CPT / Exam 5

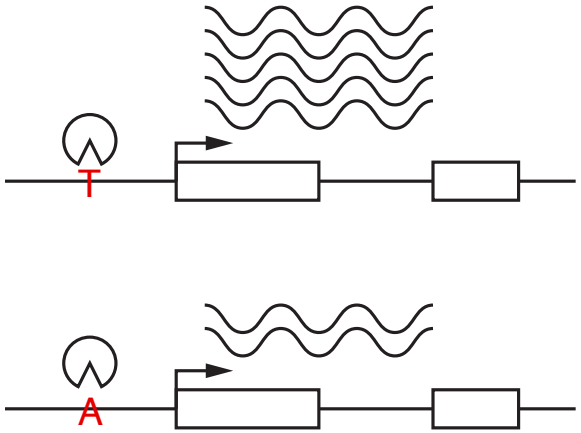
Expression variation across sample types

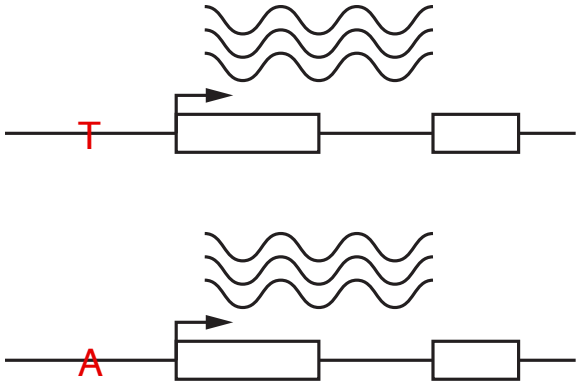


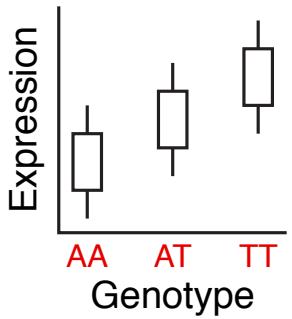
Corrected for sequencing center

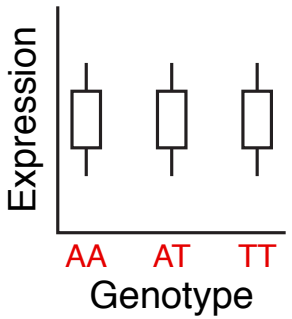
- Batch effects between Exam 1 and 5 hinder direct comparison of gene expression
- Comparisons can be made through relative measurements in each cohort
  - *Do regulatory effects on gene expression replicate between Exam 1 and Exam 5?*

Expression quantitative trait loci



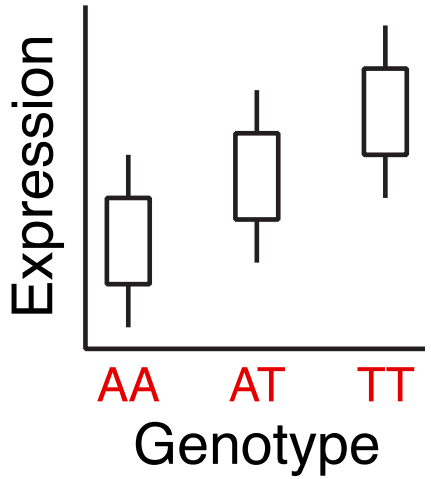






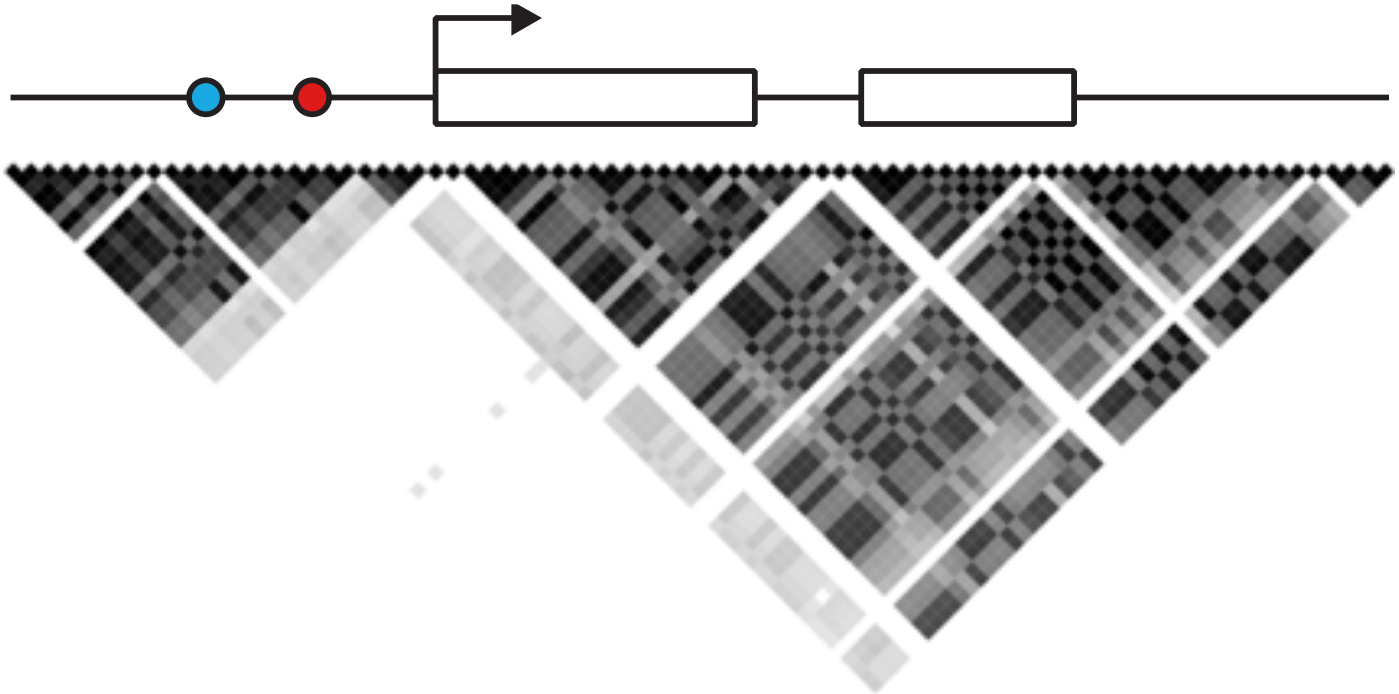
Regulatory variation is measured as expression quantitative trait loci (eQTLs)

Definition of *cis*-eQTLs



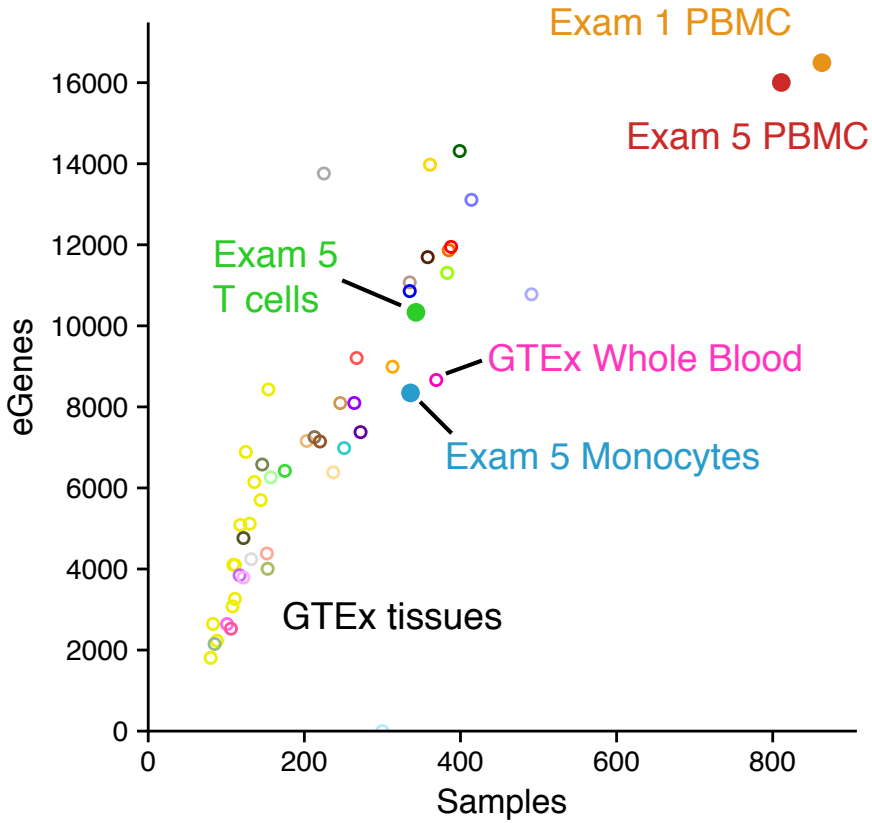
- ***cis*-eQTL**: genome-wide significant association between ≥ 1 eVariant and eGene, with associations tested within ±1Mb *cis*-window around TSS. Does not imply evidence of allelic effects at each locus.
- **eGene**: gene with at least one significant eQTL (at 5% FDR).
- **eVariant**: variant with a significant association to ≥1 eGene.
- **Effect allele**: ALT allele (not necessarily the minor allele).

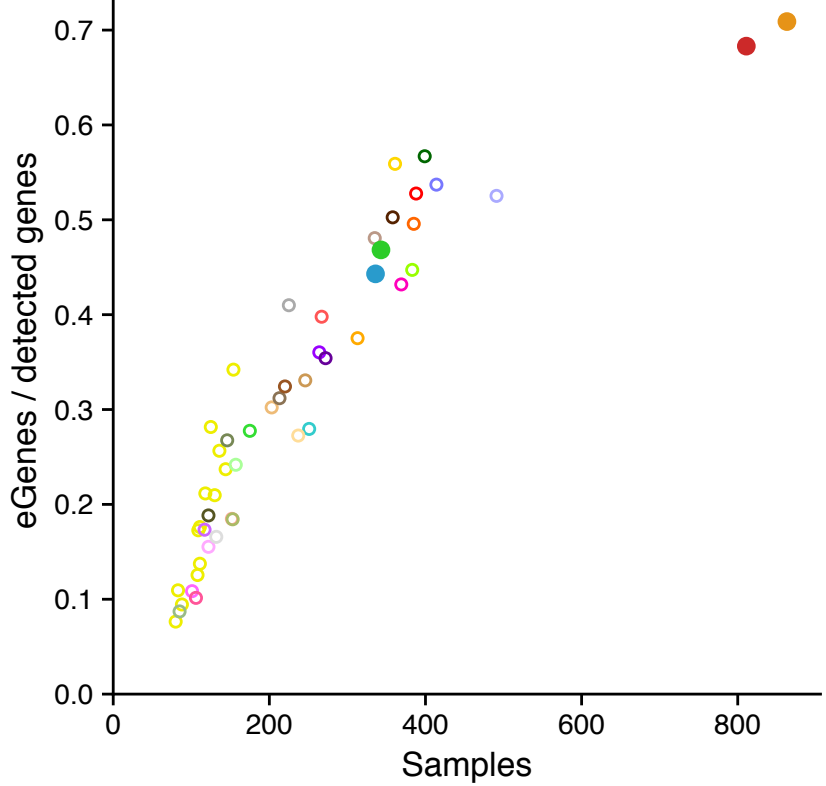
eQTL mapping and eGene discovery



- Variants in *cis*-window (±1Mb from TSS) correlated due to linkage disequilibrium (LD)
- LD must be incorporated in multiple hypothesis testing correction for establishing genome-wide significance
  - Empirical p-values from permutation of genotypes
- Methods developed for GTEx (GTEx Consortium, 2017)

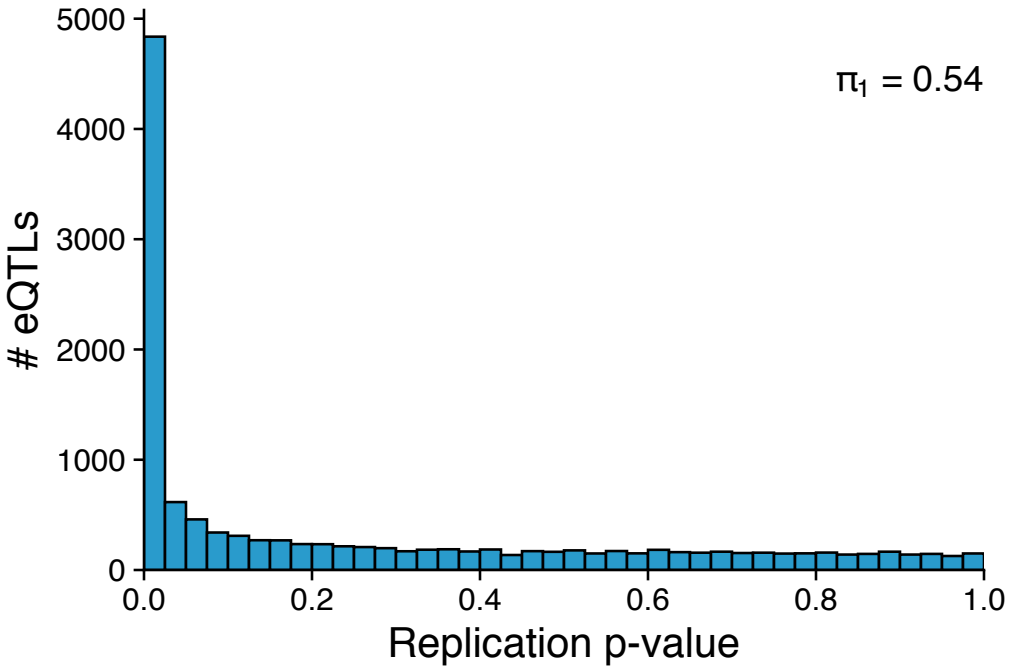
Expression quantitative trait loci detected in MESA





- *cis*-eQTL: variants within ±1Mb of TSS
- eGene: gene with ≥ 1 significant eQTL at ≤ 0.05 FDR
- GTEx data: V7

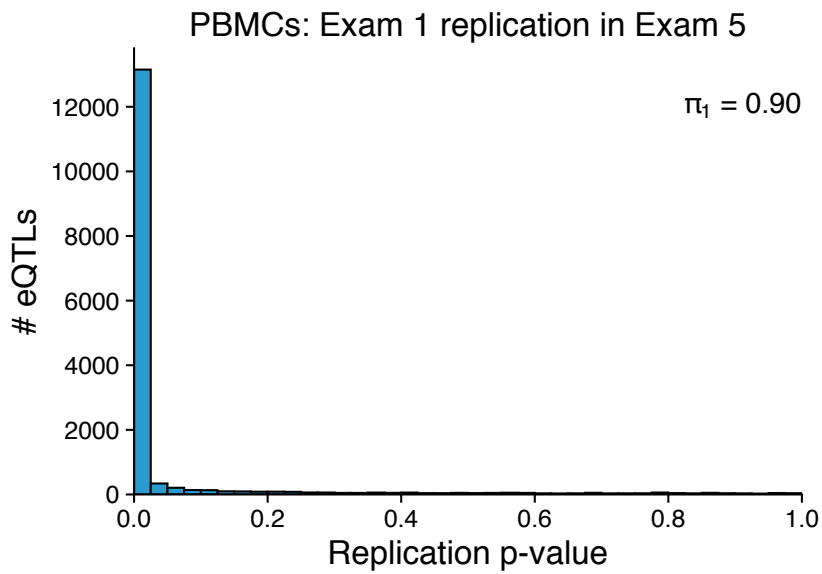
Replication of eQTLs between cohorts

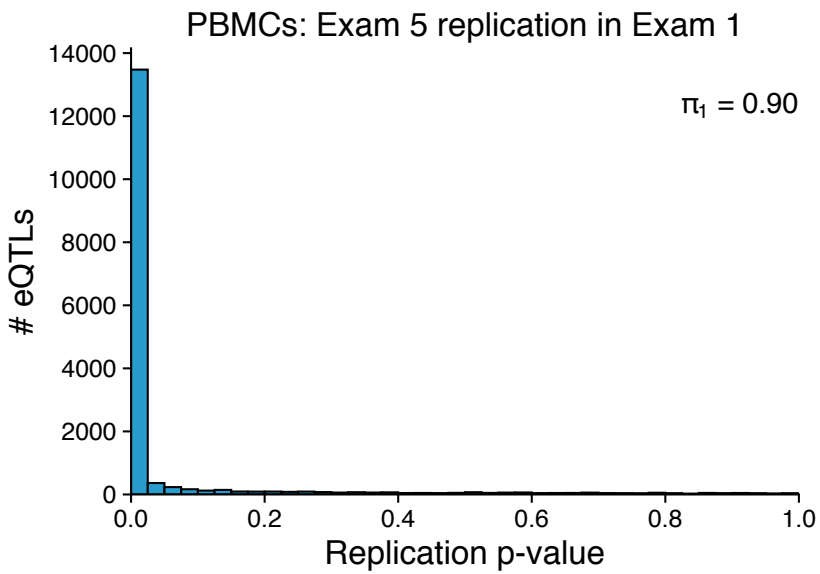


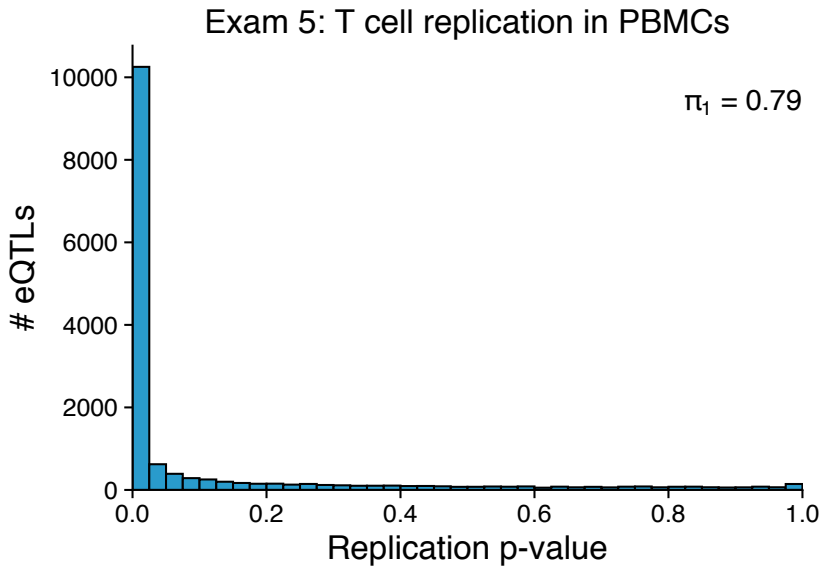
- For all genome-wide significant eQTLs in the discovery cohort, association p-values are calculated in the replication cohort
- An enrichment for small p-values indicates replication
- The proportion of replicating eQTLs is measured using the $\pi 1$ statistic, which measures the proportion of true positives

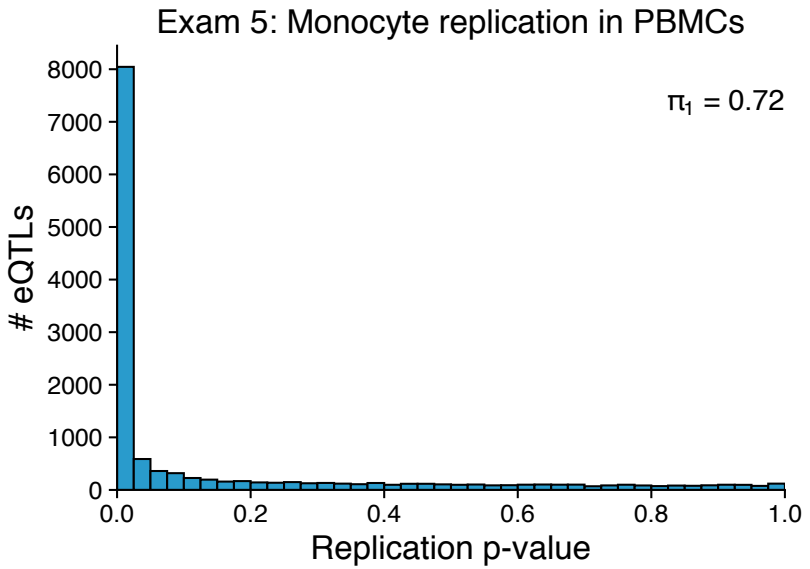
Storey & Tibshirani, *PNAS*, 2003

eQTL replication across exams and sample types



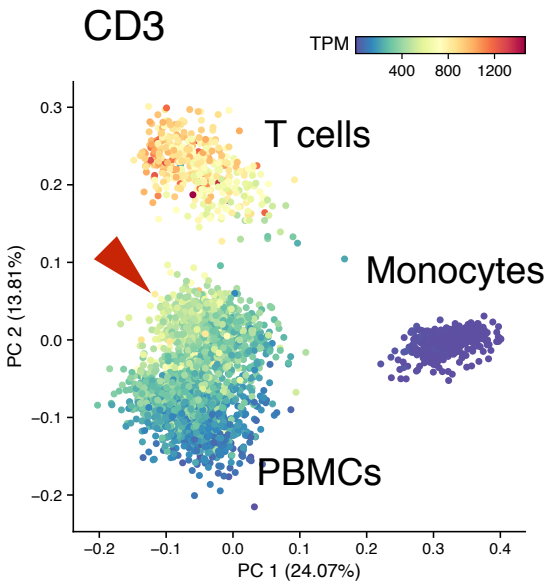


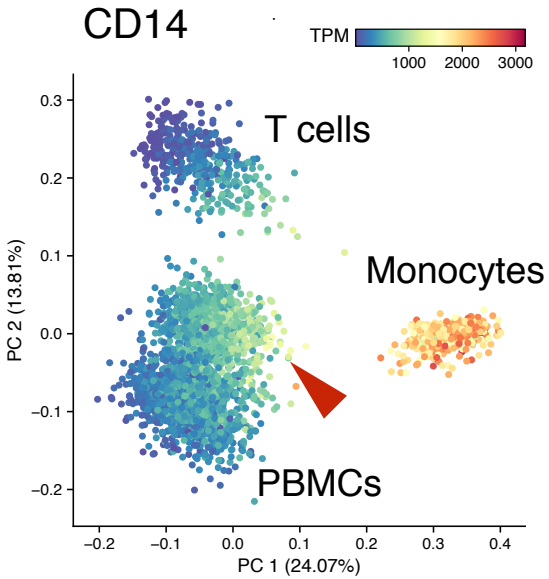


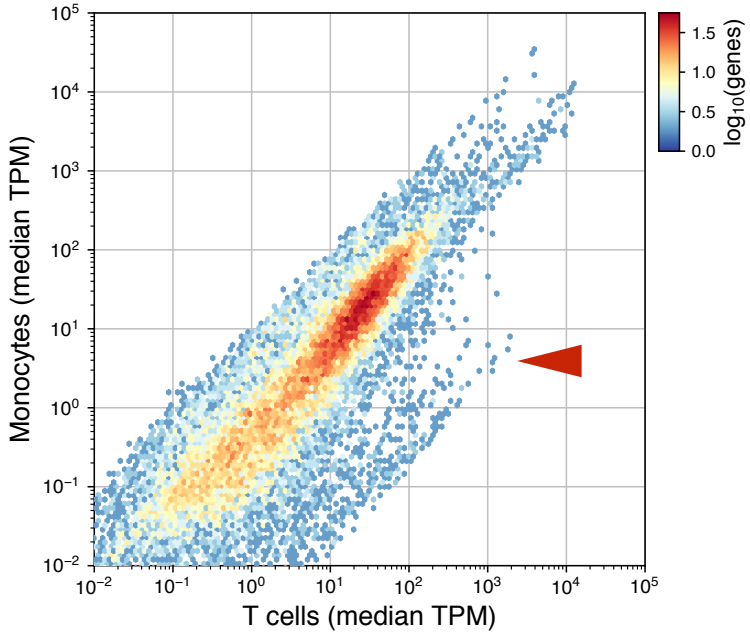


π_1: proportion of true positives

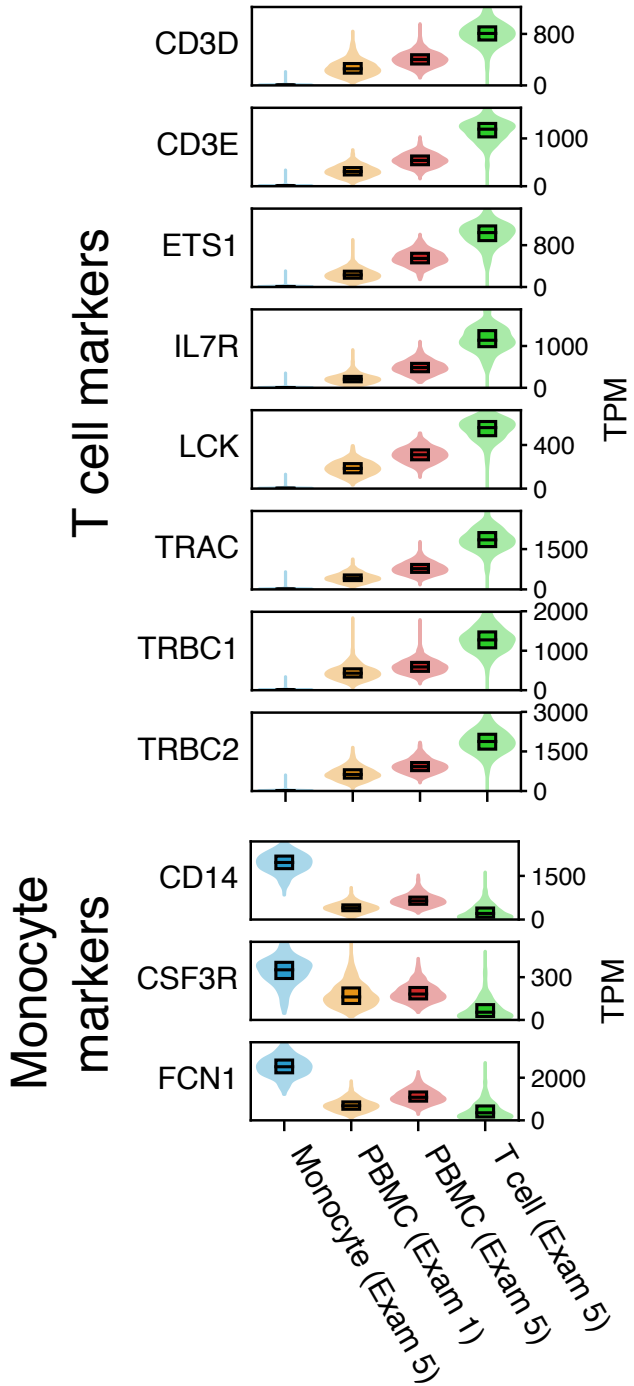
Towards identification of cell-type specific effects







Multiple highly specific (unique) markers for T cells; no unique markers for monocytes



Summary and outlook

- MESA RNA-seq data is of overall excellent quality, across exams and sample types
- Additional data being generated:
  - Splicing QTLs
  - Allele-specific expression
- Planned analyses:
  - Integration with methylation data (methylQTLs), other omics
  - Deconvolution approaches to detect cell type-specific effects in PBMCs and compare these between exams and omics data types
  - Identification of changes as a function of time and participant age (QTLs and expression)

Acknowledgements

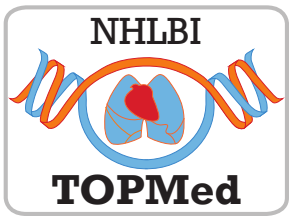
Stacey Gabriel
Namrata Gupta
Kristin Ardlie
Katie Larsson
Aaron Graubert

NWGC

Debbie Nickerson
Josh Smith
Stephanie Krauter
Chris Frazar
Daniel McGoldrick
Colleen Davis

MESA DCC (UW):
Craig Johnson, Kayleen Williams
MESA Central Lab (UV):
Peter Durda, Elaine Cornell, Russ Tracy
MESA Genetics (WF):
Yongmei Liu, Tracey Young
MESA Genetics (UVA):
Steve Rich
MESA Genetics (LABioMed/Harbor):
Kent Taylor, Jerry Rotter

George Papanicolaou
Lorraine Silsbee
Pankaj Qsaba
Jennifer Swift
Rebecca Beer



TOPMed DCC:
Cathy Laurie
Deepti Jain
Quenna Wong
Stephanie Gogarten