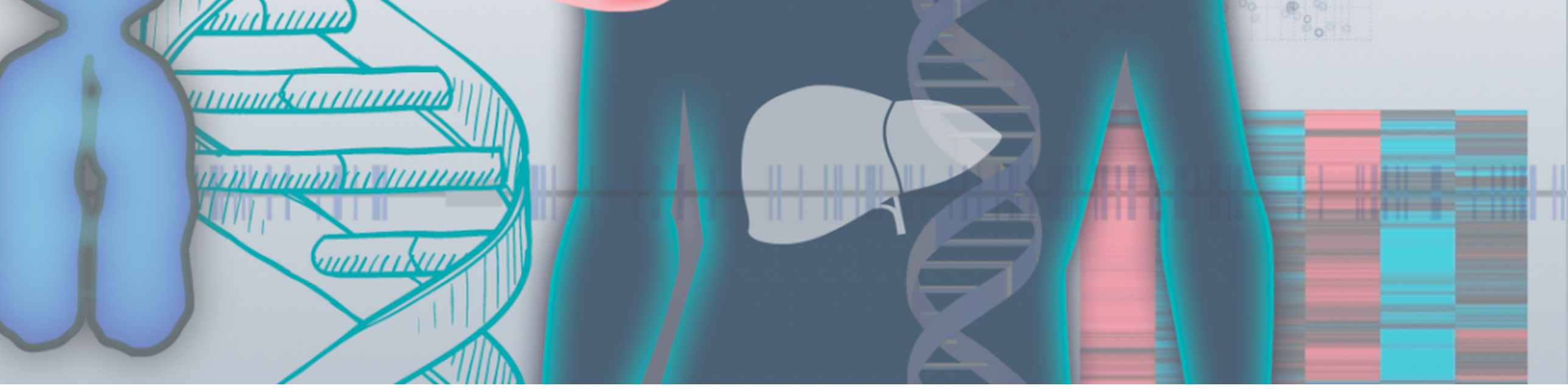


The v7 data release

GTEx Community Meeting :: 07/11/2016

François Aguet, Broad Institute

BROAD INSTITUTE

GTEx

Outline

- Summary of changes from v6
- Summary of raw and derived data produced for v7
- Summary of benchmarking results for alignment & isoform quantification
- New pipelines:
  - WGS/WES sample and variant QC
  - RNA-seq alignment, quantification, and QC
  - eQTL discovery
- Planned changes and additions for v8 release

Summary of changes from v6

- **Genotyping:** microarrays => WGS/WES
- **RNA-seq alignment:** TopHat 1.4 => STAR 2.4.2a
- **Gene expression:** new collapsed gene model
- **Isoform quantification:** FluxCapacitor => RSEM
- **eQTL discovery:** MatrixEQTL => FastQTL

Core derived data

Expression

- Read counts for genes, transcripts, exons, junctions
- Normalized expression for genes, transcripts (TPM)
- Coverage tracks (bigWig)

eQTL

- Gene-level summary: best variant, q-value, etc.
- Significant variant-gene pairs
- All variant-gene pairs
- Expression matrices (BED format); normalized + TPM
- Covariates

All derived data will be available on the GTEx Portal (http://gtexportal.org/)

Additional derived data

- Splicing QTL
  - Altrans [Ongen & Dermitzakis, 2015]
  - sQTLseekeR [Monlong et al., 2014]

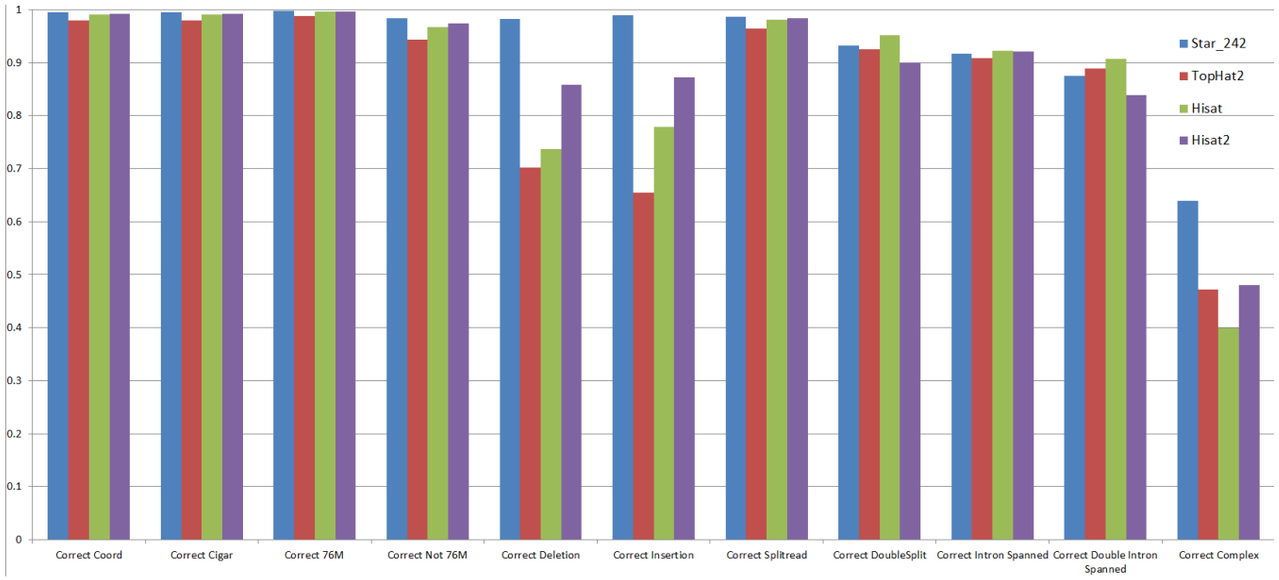
- Allele-specific expression
  [Castel et al., 2015; van de Geijn et al., 2015]

- Multi-tissue eQTL
  - Metasoft [Han & Eskin, 2012]

All derived data will be available on the GTEx Portal (http://gtexportal.org/)

Benchmarking

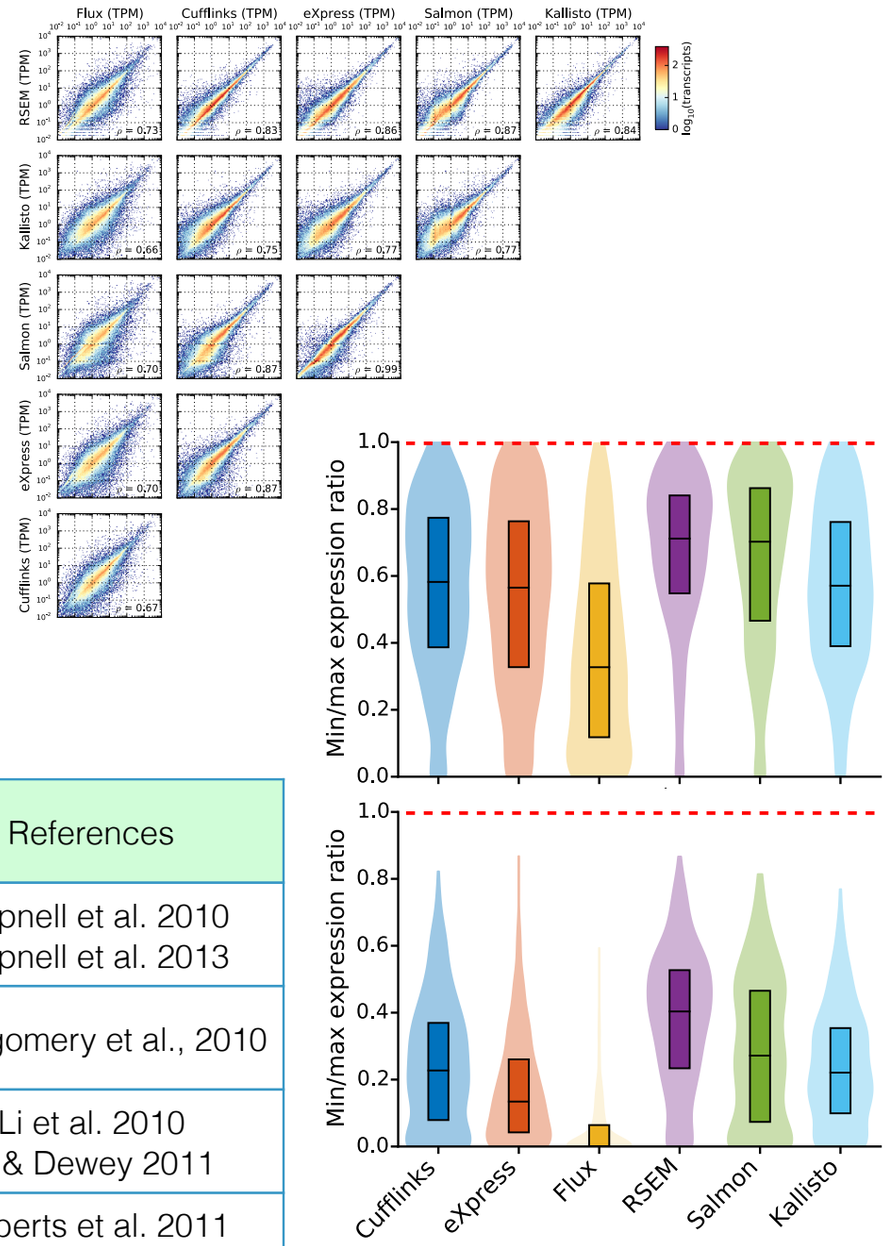
Spliced transcript alignment



Tim Sullivan, Broad Institute

	References
TopHat	Trapnell et al. 2009
TopHat2	Kim et al. 2013
STAR	Dobin, 2013
HISAT2	Kim et al., 2015

Transcript isoform expression estimation



	Input alignment	References
Cufflinks	Genome	Trapnell et al. 2010 Trapnell et al. 2013
FluxCapacitor	Genome	Montgomery et al., 2010
RSEM	Transcriptome	Li et al. 2010 Li & Dewey 2011
eXpress	Transcriptome	Roberts et al. 2011 Roberts & Pachter 2012
Sailfish/ Salmon	Transcriptome / Raw reads	Patro et al. 2014 Patro et al., 2016
Kallisto	Raw reads	Bray et al. 2015

Gene-level expression quantification

- Quantification based on collapsed annotation (GENCODE v19)
  - Exclude exons from transcripts annotated as *retained_intron* or *read_through*
- GTEx RNA-seq protocol is unstranded
  - Exclude exon domains shared by overlapping genes

- Effect on eQTL discovery:
  ~10-15% more eGenes discovered vs. gene-level quantification from RSEM

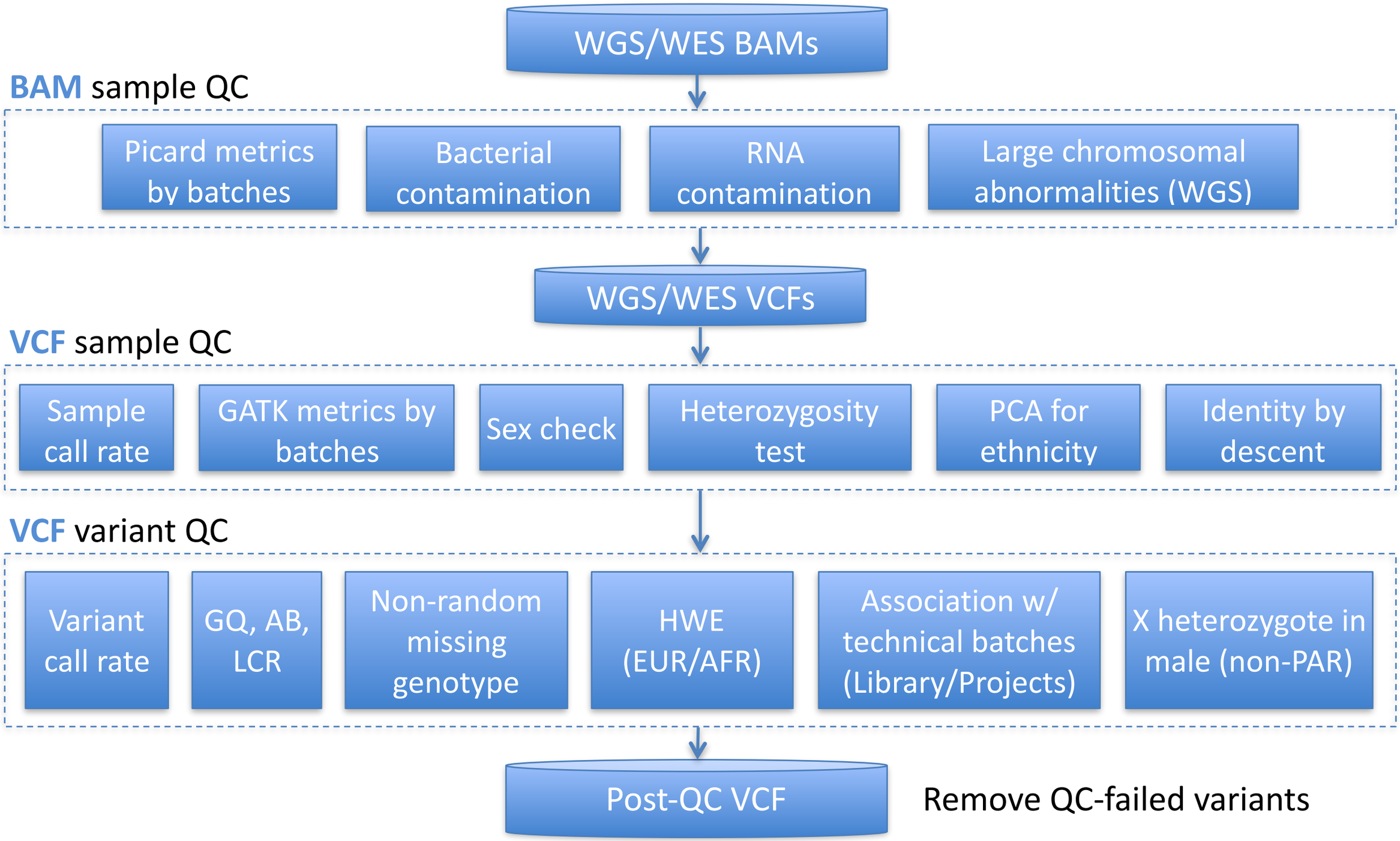
v6p release

- **Update of derived data only** (hosted on GTEx Portal)
  - Gene expression: read counts + RPKM GCT files.
  - eQTL: FastQTL instead of MatrixEQTL, otherwise identical. Includes chr. X eQTL.

Genotyping data and pipeline

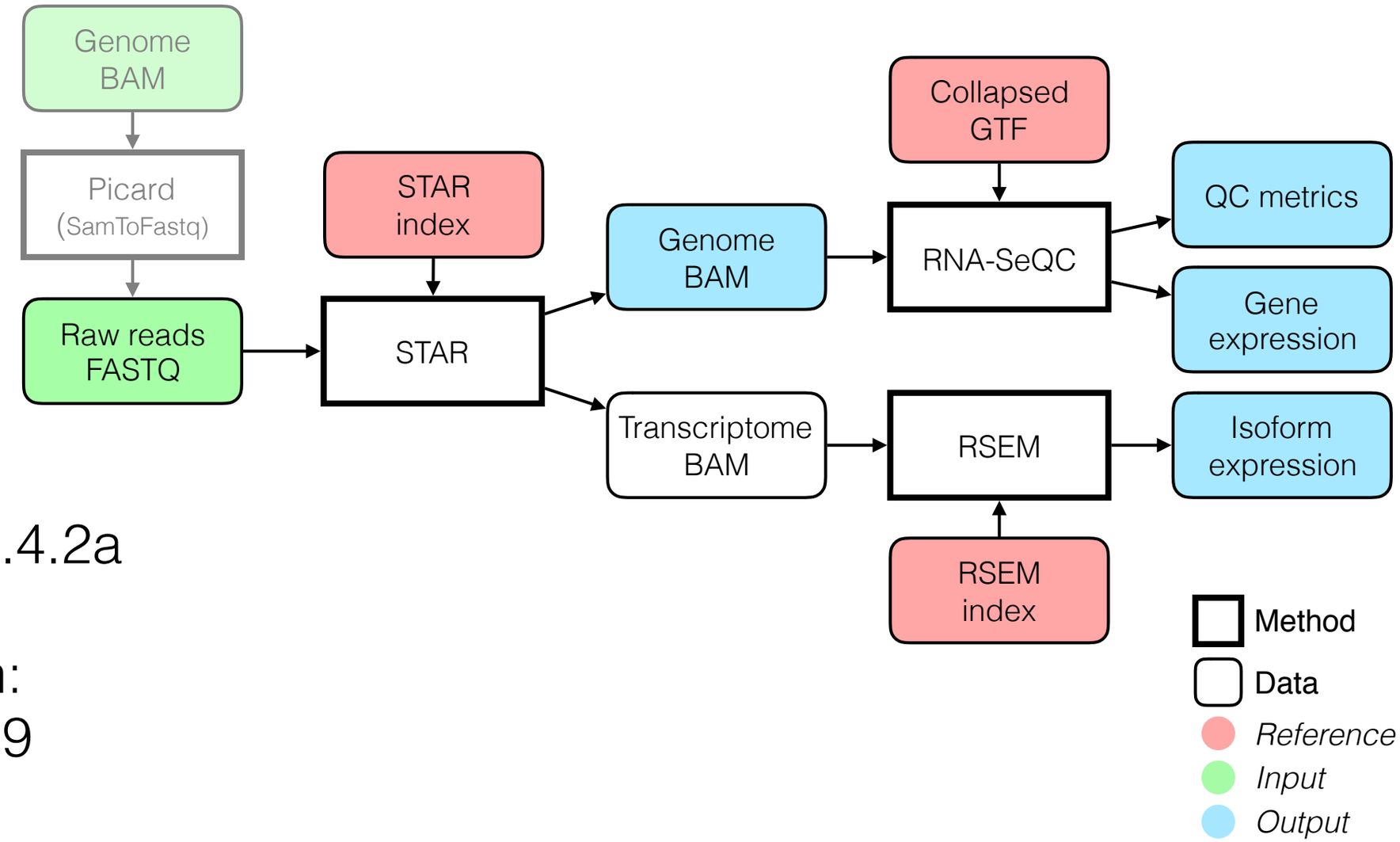
	WES	WGS
# donors	603	652
# donors (post sample QC for eQTL analysis)	603	635
Sequencing coverage	100x	30x
Alignment	BWA	BWA-MEM
Joint variant calling	HaplotypeCaller v3.4 (GATK)	HaplotypeCaller v3.4 (GATK)
Variant QC	-	GATK, Hail, Custom code
Functional and LoF annotations	Ensembl's Variant Effect Predictor + LOFTEE	Ensembl's Variant Effect Predictor + LOFTEE
Phasing of SNPs and indels	Local (in sequence read)	Local and long range with SHAPEIT
Structural variant calling	-	GenomeSTRiP, LUMPY (merged call set)

Overview of WGS/WES QC pipeline



See poster #20 (Li et al.)

RNA-seq alignment and quantification

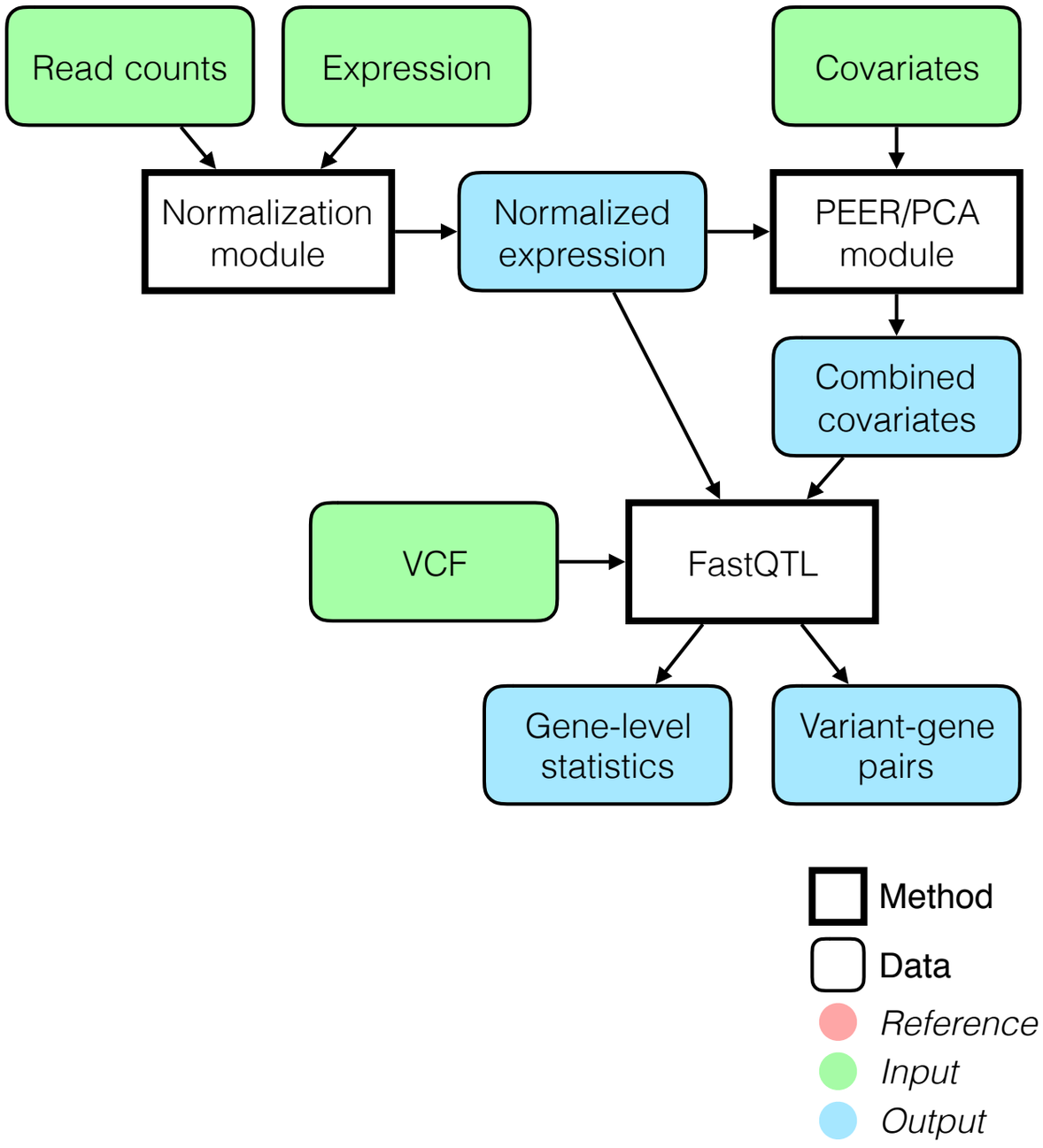


- Aligner: STAR v2.4.2a
- Gene expression:
  RNA-SeQC v1.1.9
- Transcript expression:
  RSEM v1.2.22
- QC metrics:
  RNA-SeQC v1.1.9

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

eQTL discovery

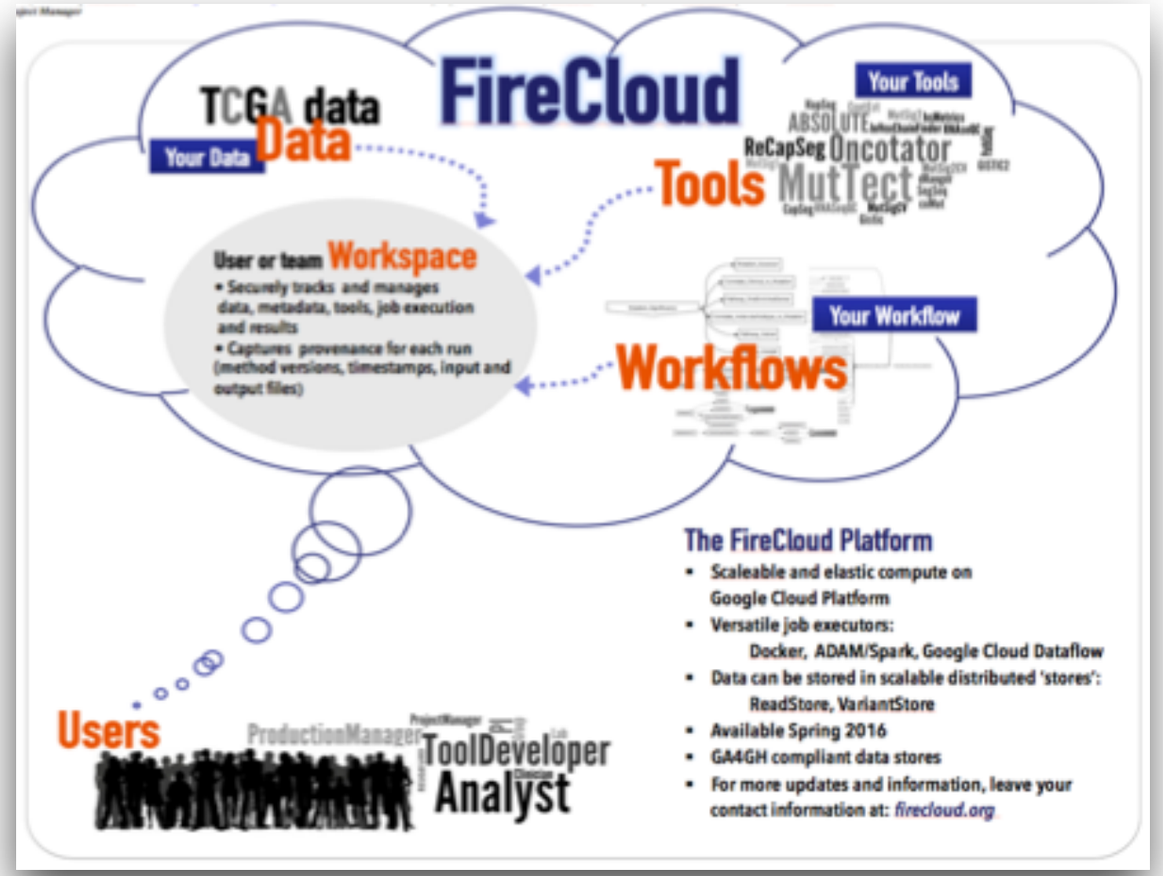
- QTL mapper: FastQTL
- Covariate correction:
  - PEER factors
  - Explicit covariates:
    Genotype PCs, gender
- *cis* window: ±1Mb
- MAF ≥ 0.01 and ≥10 samples containing minor allele

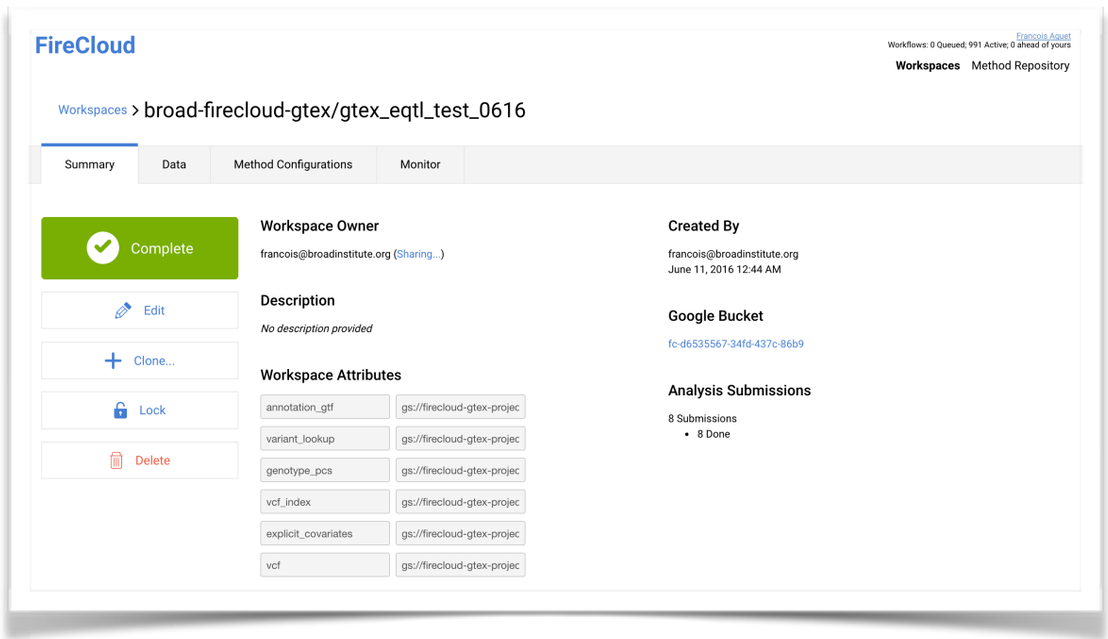


FastQTL: Ongen et al., *Bioinformatics*, 2016
PEER: Stegle et al., *PLoS Comp Bio*, 2010

Public release of pipelines on FireCloud

- Cloud-based genomics analysis platform developed at the Broad Institute: http://firecloud.org
  - Part of the NCI Cloud Pilot initiative; currently hosts TCGA data.
- Several GTEx pipelines already implemented (RNA-seq and eQTL); public release is imminent.
  - Also available as Docker images.





Outlook: planned changes/additions for v8 release

- Realignment/quantification to hg38/GRCh38 (+ latest GENCODE release) using FireCloud
- Re-evaluation of isoform quantification methods
- Small RNA-seq pipeline (alignment, QC, quantification)
- FireCloud will facilitate collaborating on pipelines (Docker-based).
  Let us know if you're interested in contributing!

Acknowledgments

BROAD INSTITUTE

LDACC

K. Ardlie, G. Getz, A. Segrè, T. Sullivan, X. Li

E. Gelfand, C. Trowbridge

D. MacArthur, M. Kellis, J. Hirschhorn

Genomics Platform

GTEx Portal

J. Nedzel, K. Huang, K. Hadley,

S. Meier, M. Noble

GTEx

The GTEx Project Consortium

Benchmarking Subgroup

eQTL Subgroup

Transcriptome Subgroup

Gender Subgroup

NIH

The Common Fund

Donors and their families