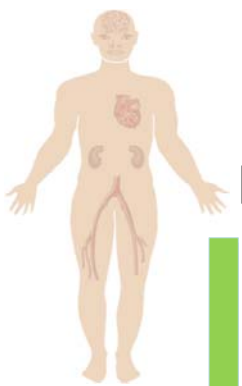
GTEx – Sample Molecular Quality and Data Production

Kristin Ardlie

23 February, 2012

BROAD INSTITUTE

GTEx Sample Workflow



LDACC

BSS

Bloods & Fresh Skin

1. Whole Blood → DNA
2. Whole Blood → Cell Culture (EBV)
3. Fresh Skin → Fibroblast Culture
4. PAXgene Tube → RNA

Genotyping

Analysis

Donor

Tissue Samples

CBR

PAXgene Preserved Tissues → RNA

Brain

Brain Bank

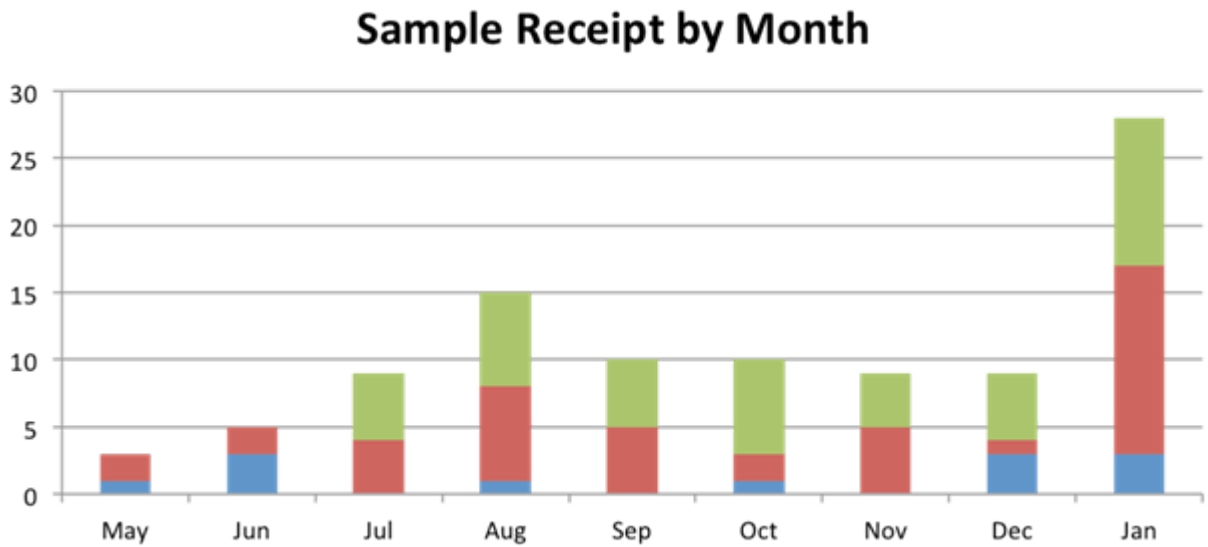
Snap Frozen Tissues → RNA

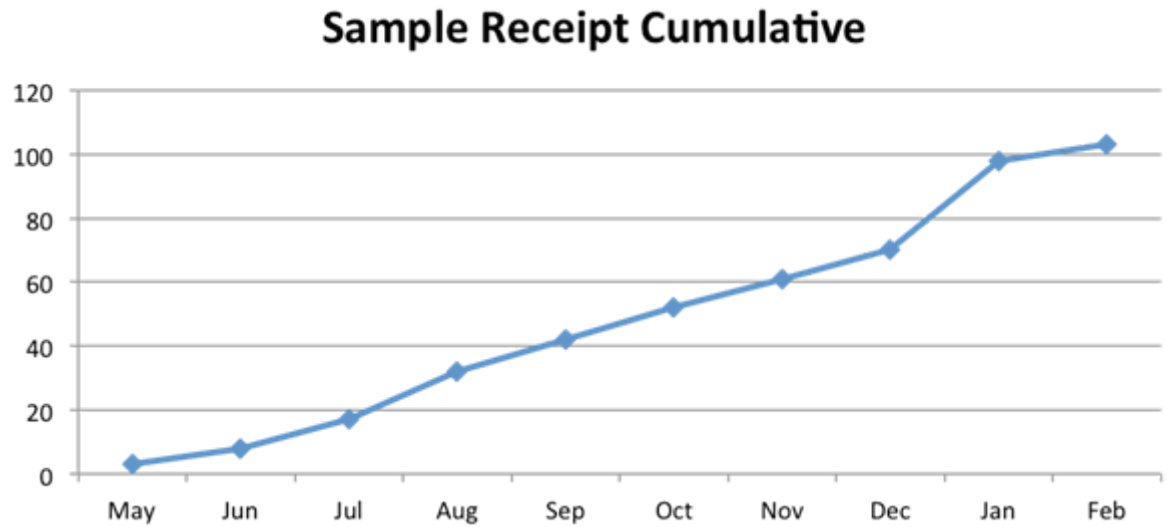
Expression arrays and RNA Seq

Analysis

dbGAP
R01 Analysis Centers

Sample Receipt by LDACC

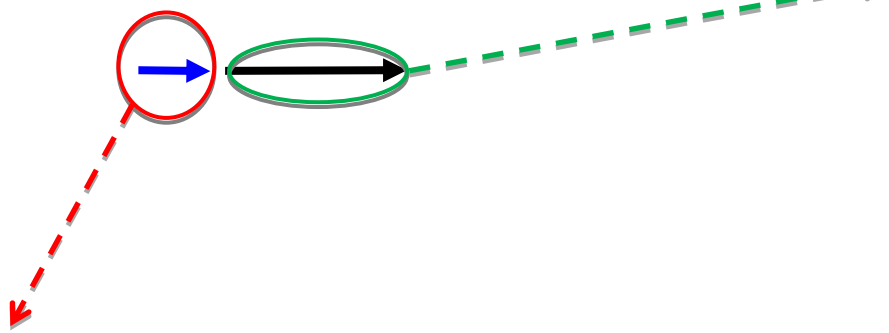




N=105 donors

GTEx Project Current Status

Base Pilot, Period-1 (12 months)	Base Pilot, Period-2 (18 months)



May onwards

GTEx Beta & Accrual Phase

105 Donors (x 19-30 tissues each)

* PaxGene Preservative only

Bloods → Rcvd - 105 Donors
Processed DNA - 104
QC'd – 100
→ Genotyped – 64

Tissues → Rcvd – 104 Donors PAX (+29 Brains)
Processed RNA – 91 (+25)
QC'd – 80 (+20)
→ Expression Arrays - 4 batches of 93-95
→ RNA SEQ –290 samples

Jan to April

GTEx Pilot Experiment

3 Donors (x 20-25 tissues each)

* Test of 3 tissue preservation methods
  - Snap-freezing
  - RNAlater
  - PaxGene preservative

Target Tissues

Paxgene Preserved Tissues

- Adipose
- Adrenal Gland
- Artery – Aorta
- Artery – Coronary
- Artery – Tibial
- Bladder
- Breast - Mammary Tissue
- Cervix – Ecto & Endocervix
- Colon
- Esophagus – Mucosa & Muscularis
- Fallopian Tube
- Heart
- Kidney – Cortex & Medulla
- Liver
- Lung
- Muscle – Skeletal
- Nerve – Tibial
- Ovary
- Pancreas
- Pituitary
- Prostate
- Skin
- Spleen
- Stomach
- Testis
- Thyroid
- Uterus
- Vagina
- Brain – Cerebellum
- Brain – Cortex

Fresh Frozen Brains

- Cerebellar Hemisphere
- Frontal Cortex (BA9)
- Hippocampus
- Substantia nigra
- Anterior cingulate cortex (BA24)
- Amygdala
- Caudate (basal ganglia)
- Nucleus accumbens (basal ganglia)
- Putamen (basal ganglia)
- Hypothalamus
- Spinal cord (cervical c-1)

Currently collecting 9-30 tissues per donor

Target Tissues – High Priority

Paxgene Preserved Tissues

- **Adipose**
- Adrenal Gland
- Artery – Aorta
- Artery – Coronary
- **Artery – Tibial**
- Bladder
- Breast - Mammary Tissue
- Cervix – Ecto & Endocervix
- Colon
- Esophagus – Mucosa & Muscularis
- Fallopian Tube
- **Heart**
- Kidney – Cortex & Medulla
- Liver
- **Lung**
- **Muscle – Skeletal**
- **Nerve – Tibial**
- Ovary
- Pancreas
- Pituitary
- Prostate
- **Skin**
- Spleen
- Stomach
- Testis
- **Thyroid**
- Uterus
- Vagina
- **Brain – Cerebellum**
- **Brain – Cortex**

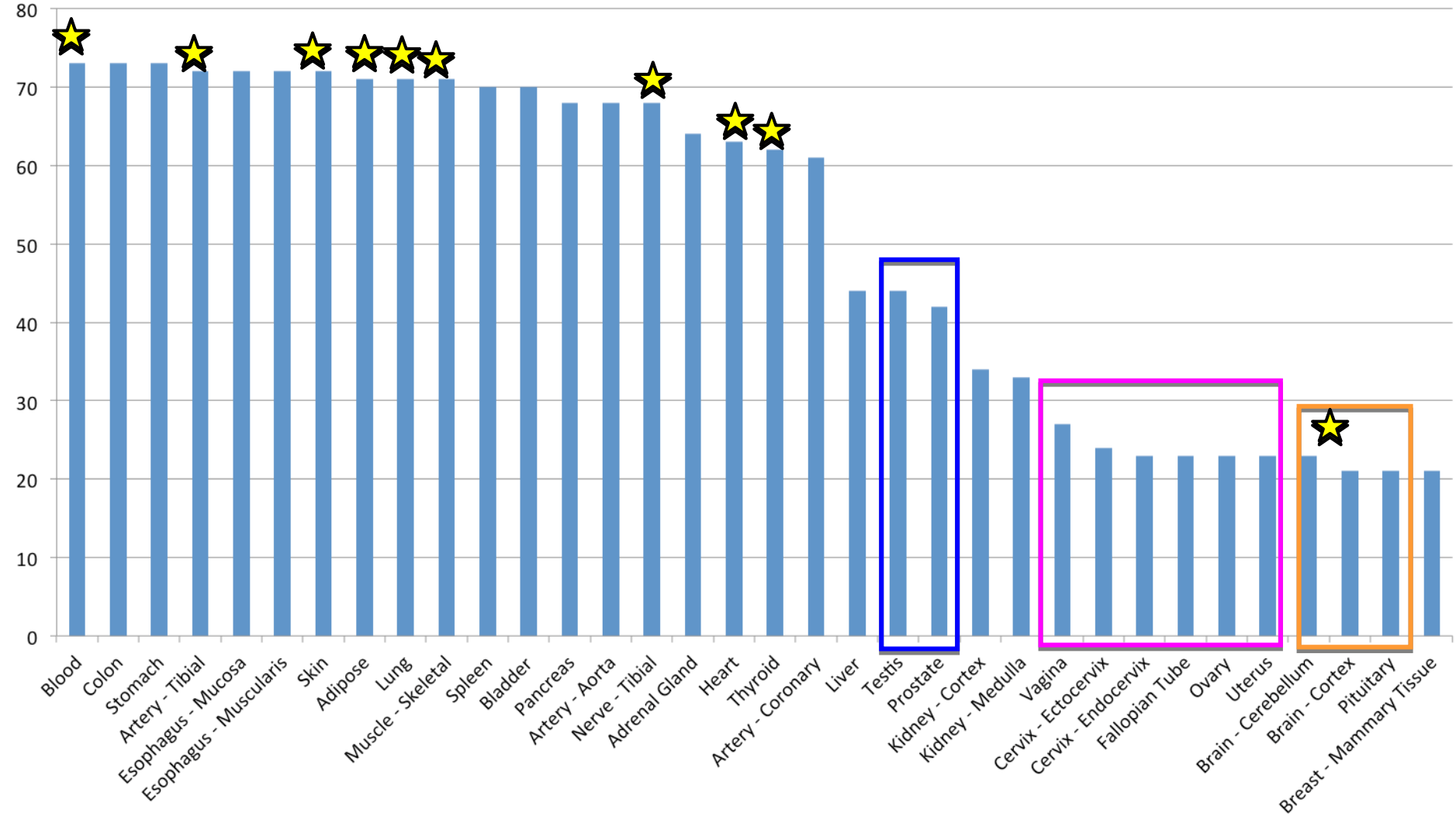
- **Fibroblast Cell Line**
- **Blood**
- **Lymphoblastoid Cell Line**

Fresh Frozen Brains

- **Cerebellar Hemisphere**
- **Frontal Cortex (BA9)**
- **Hippocampus**
- **Substantia nigra**
- **Anterior cingulate cortex (BA24)**
- **Amygdala**
- **Caudate (basal ganglia)**
- **Nucleus accumbens (basal ganglia)**
- **Putamen (basal ganglia)**
- **Hypothalamus**
- **Spinal cord (cervical c-1)**

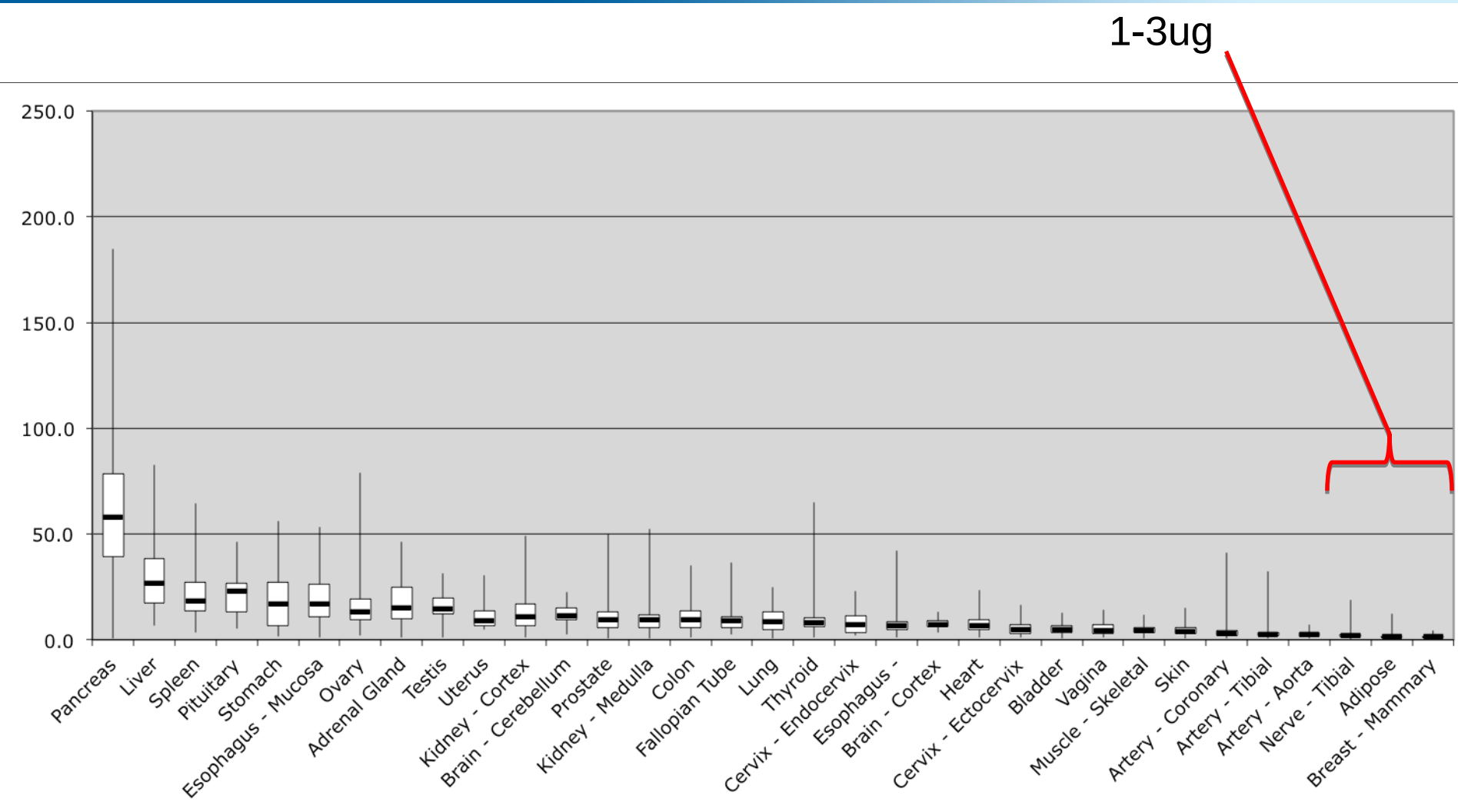
21 Tissues + 3 = 24 Priority

Tissue Distribution



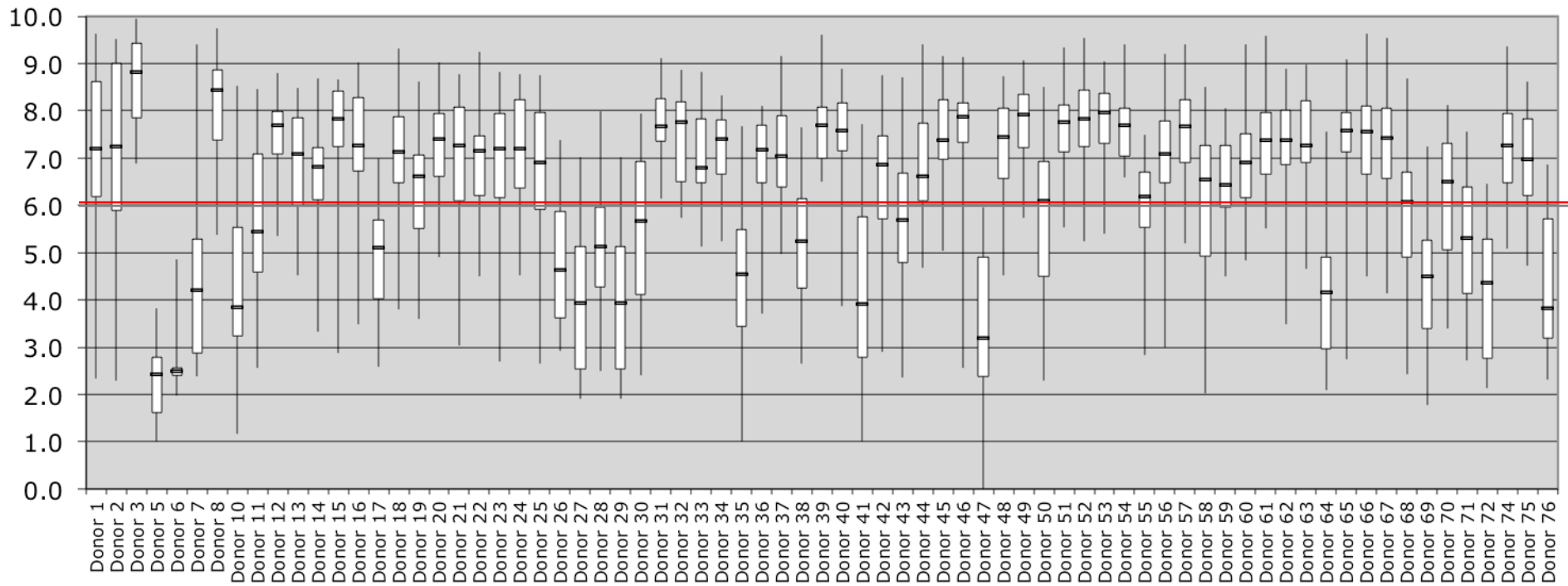
N = 75 Donors

RNA Yields



N = 1667 Paxgene Tissues

RNA Quality – Paxgene Tissues



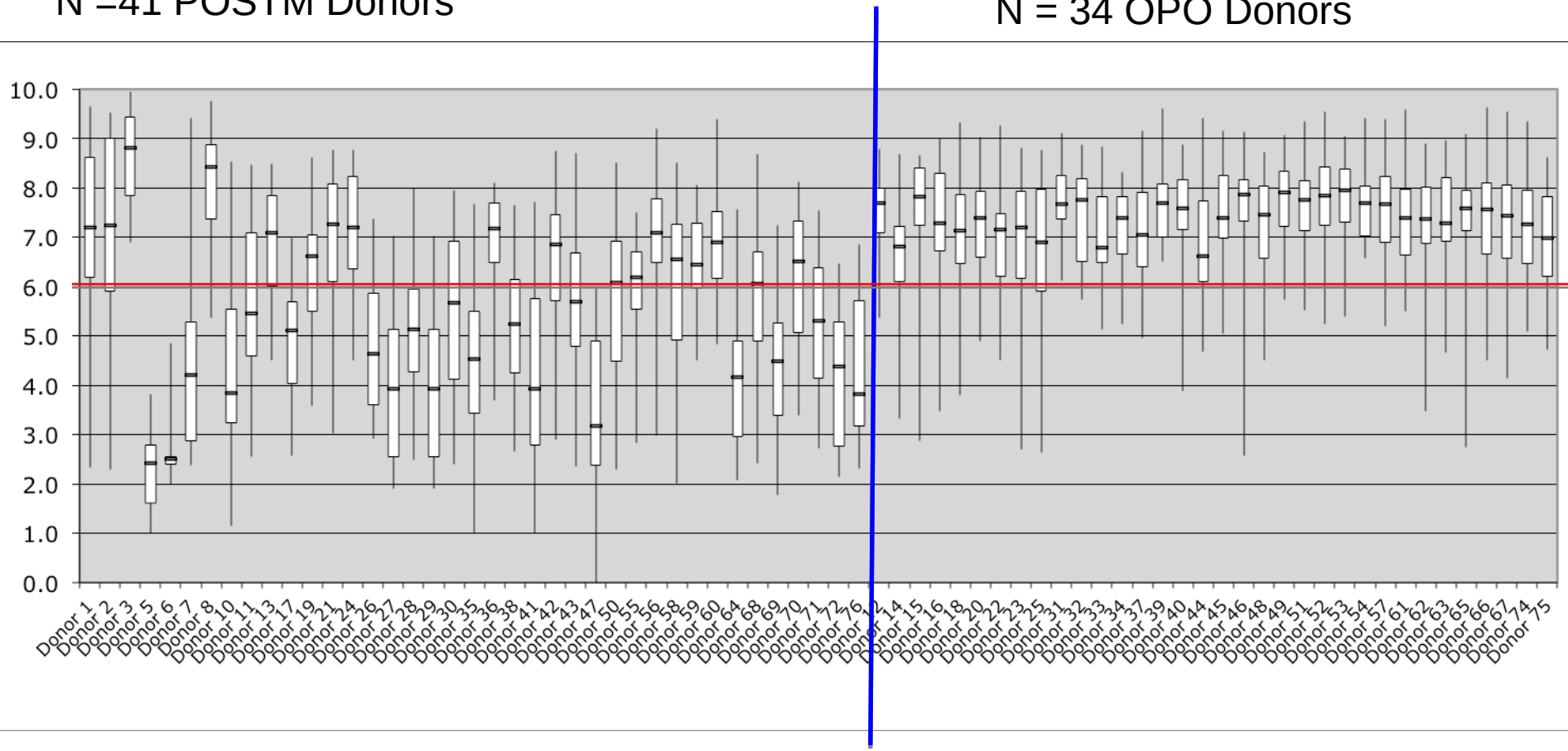
N=75 donors

RNA Quality varies by Donor

RNA Quality – Paxgene Tissues

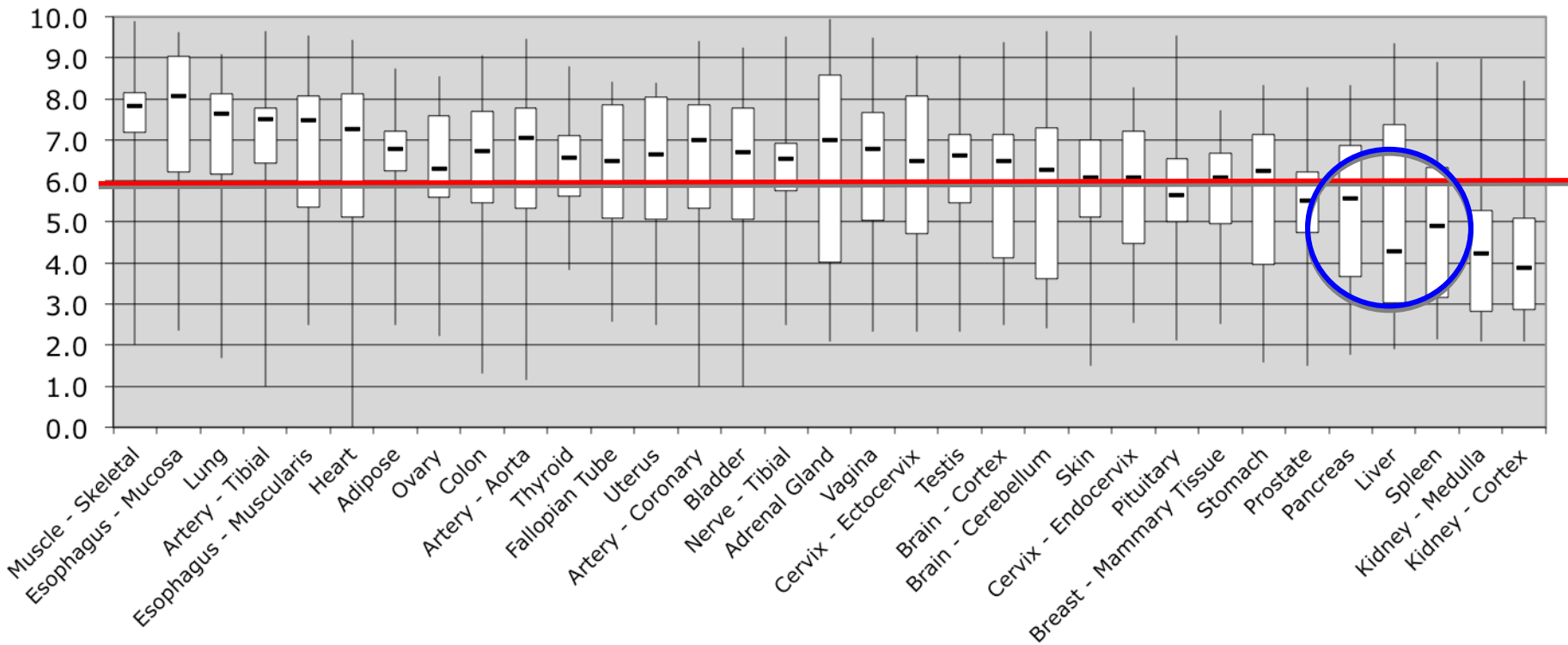
N =41 POSTM Donors

N = 34 OPO Donors



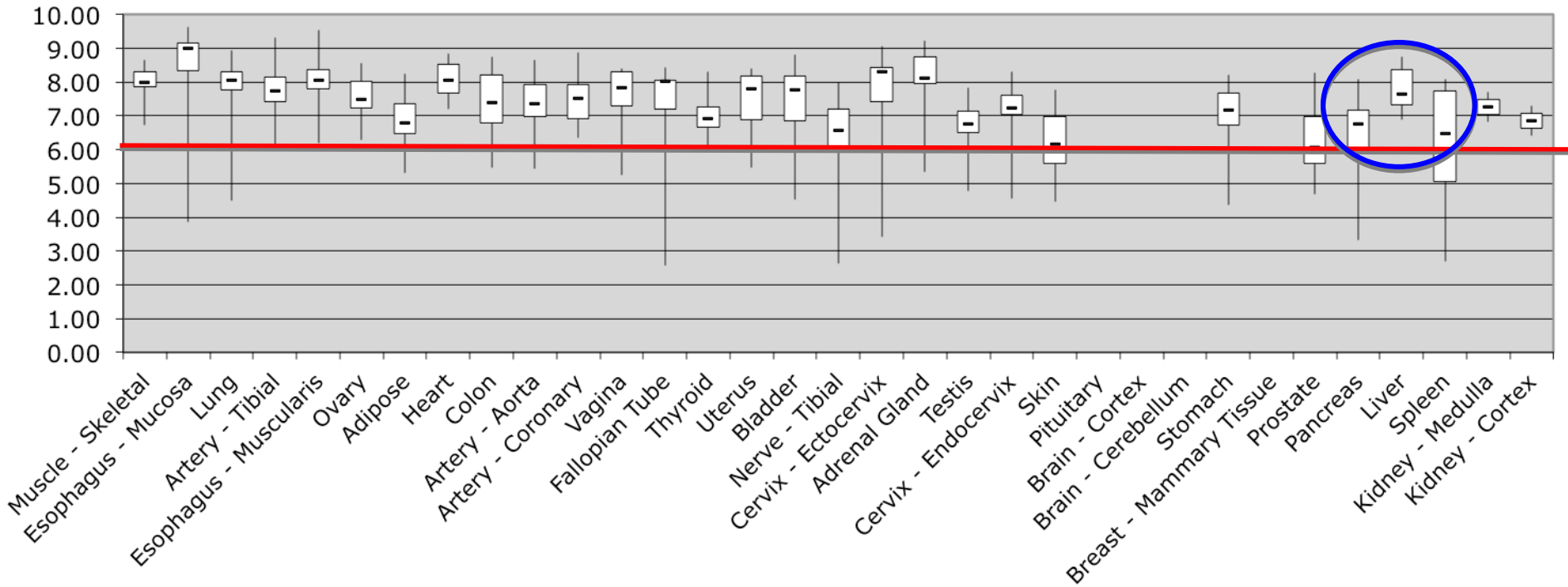
Donor RNA variation influenced by collection type

RNA Quality by Tissue Type



N=1667 tissues

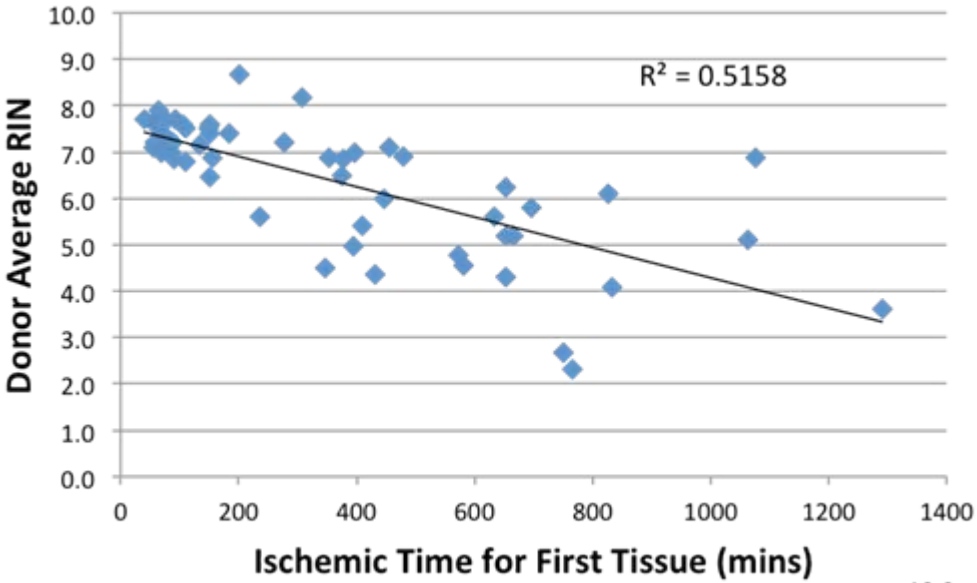
RNA Quality by Tissue Type



OPO donors only

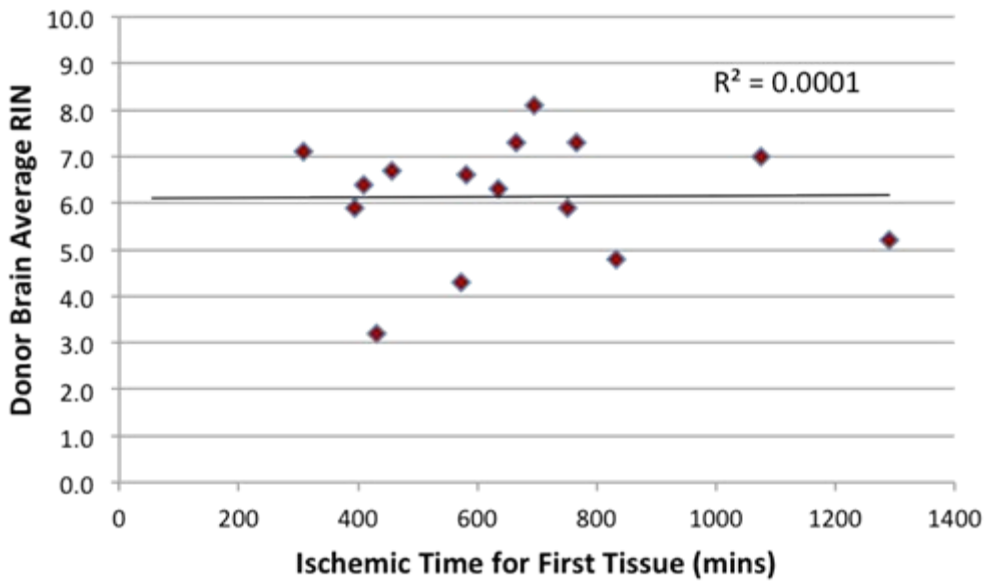
Tissue-specific RNA quality also donor-driven

RNA Quality and PMI



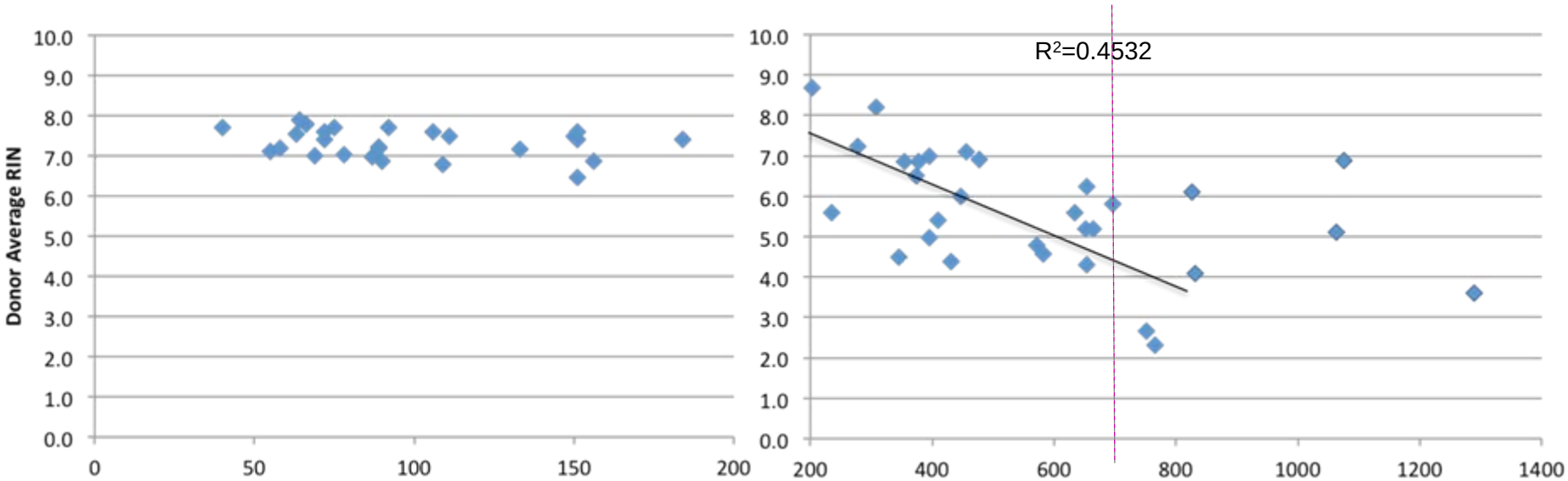
N=56 donors

N=15 donors



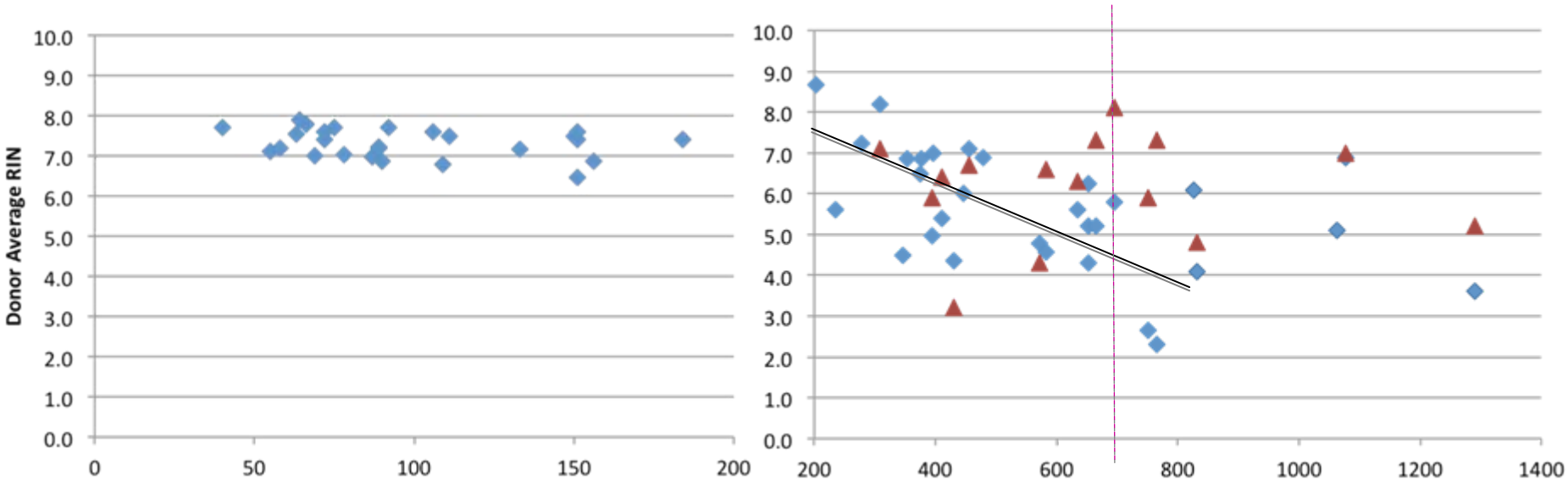
RIN decreases with increasing PMI for PAXgene Tissues, but not for Fresh Frozen Brains

RNA Quality declines with PMI



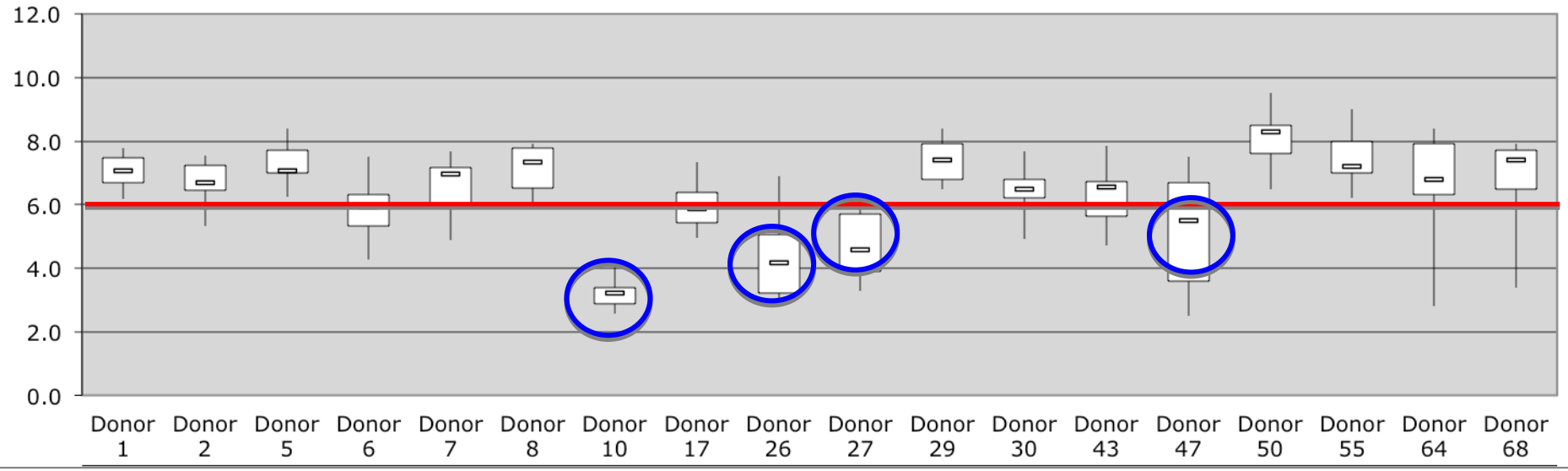
RINs are largely stable out to about 3 hours, then decline from 3- ~12 hours. More variable thereafter.

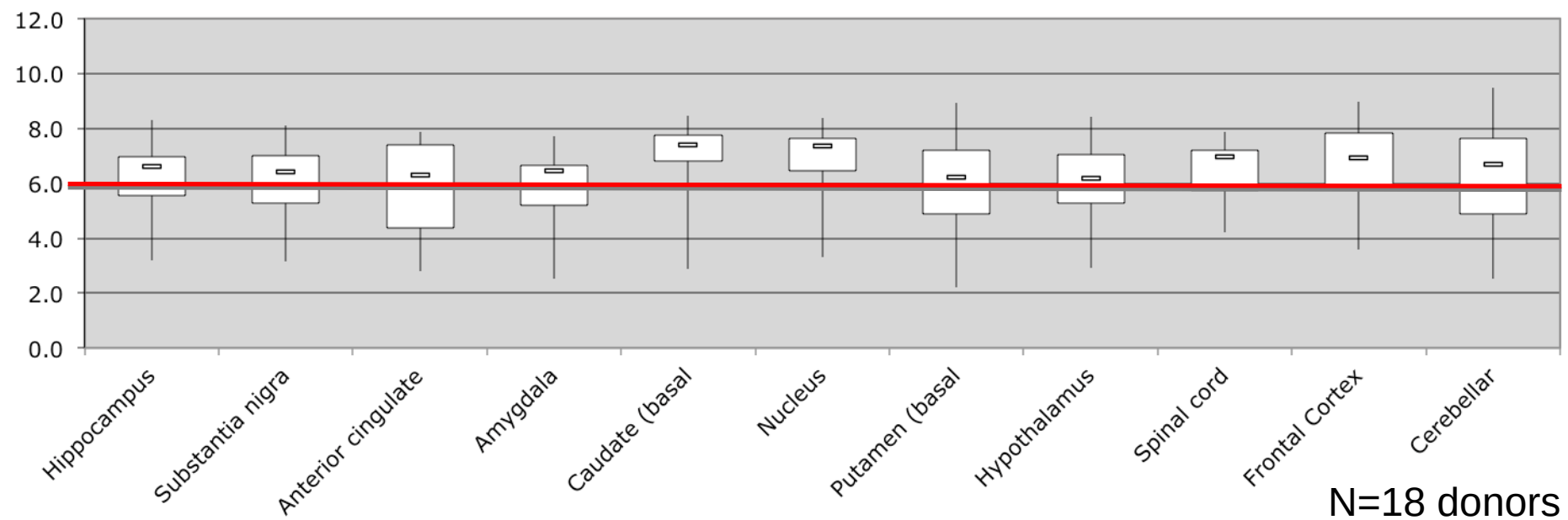
RNA Quality declines with PMI



Brain RINs less correlated with PMI, and not always the same as mean PAXgene RIN from same donor

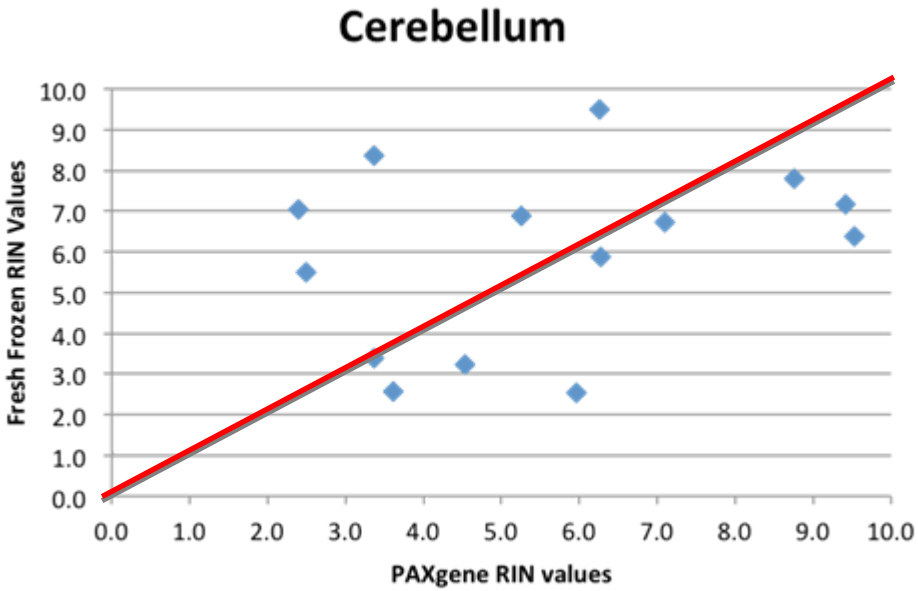
Fresh Frozen Brains



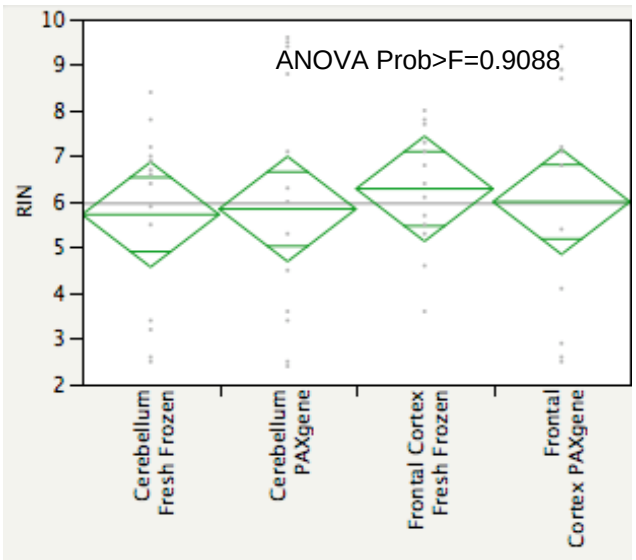


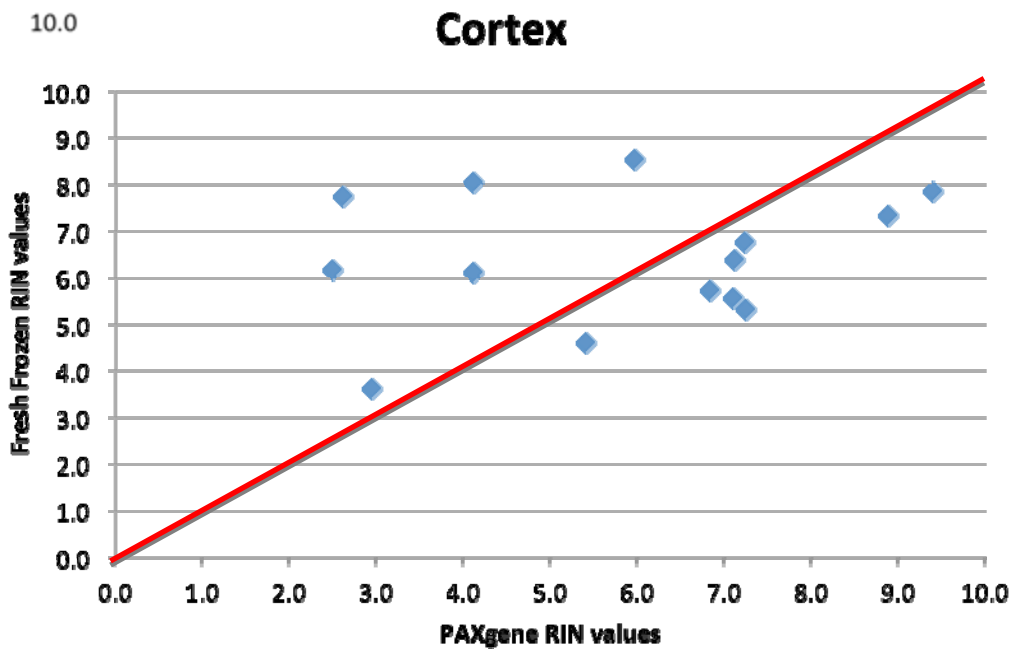
N=18 donors

PAXgene vs FF RNA Quality - Brains

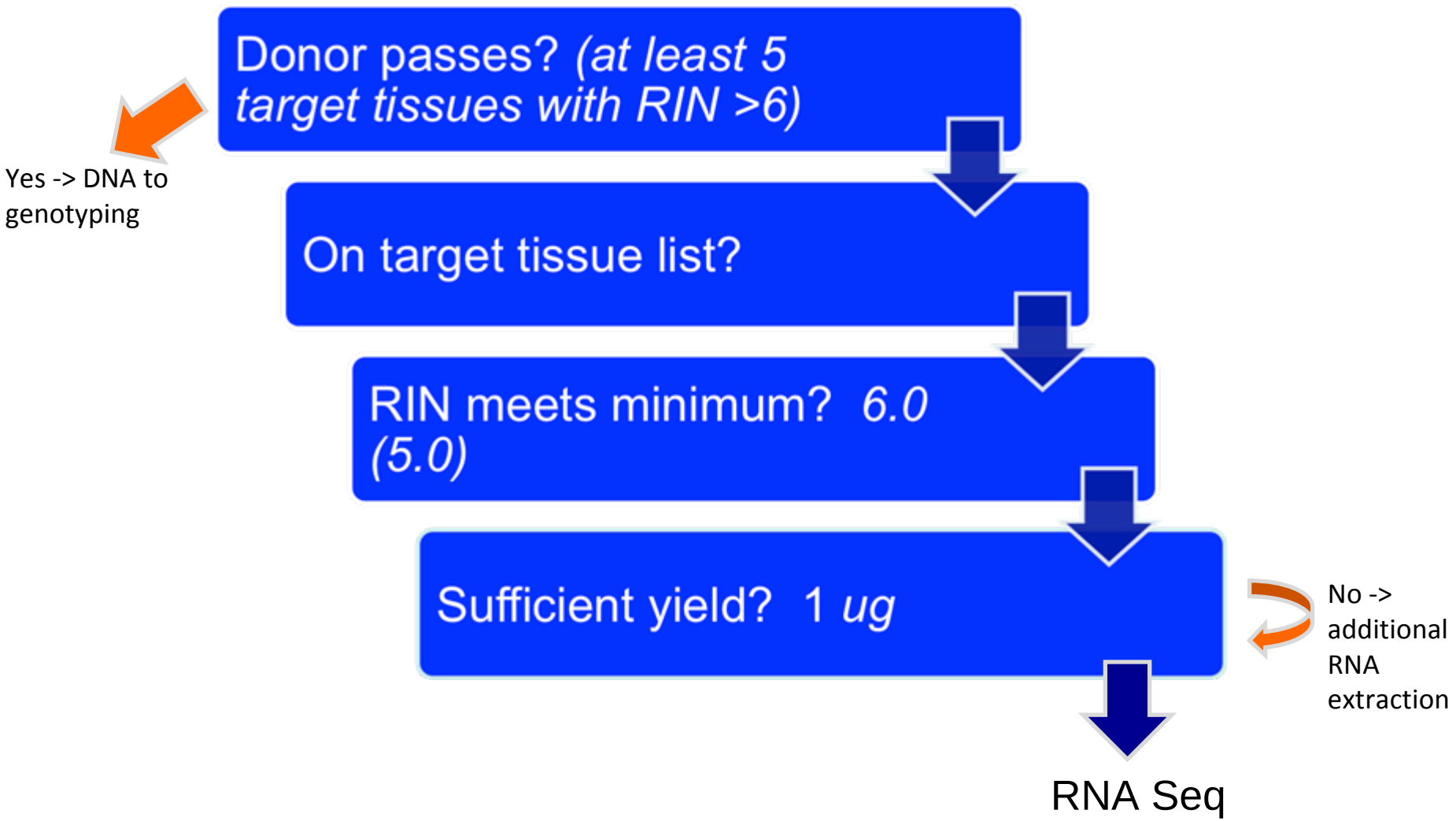


N=15 donors

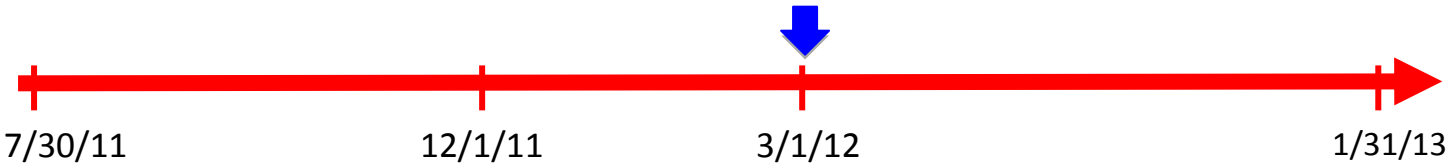




Sample Qualification

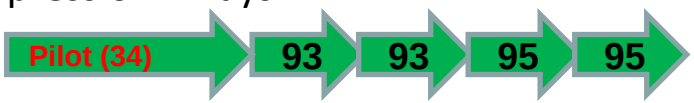


Molecular Analysis Current Status

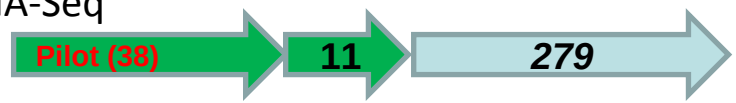


Genotyping Arrays

Expression Arrays



RNA-Seq



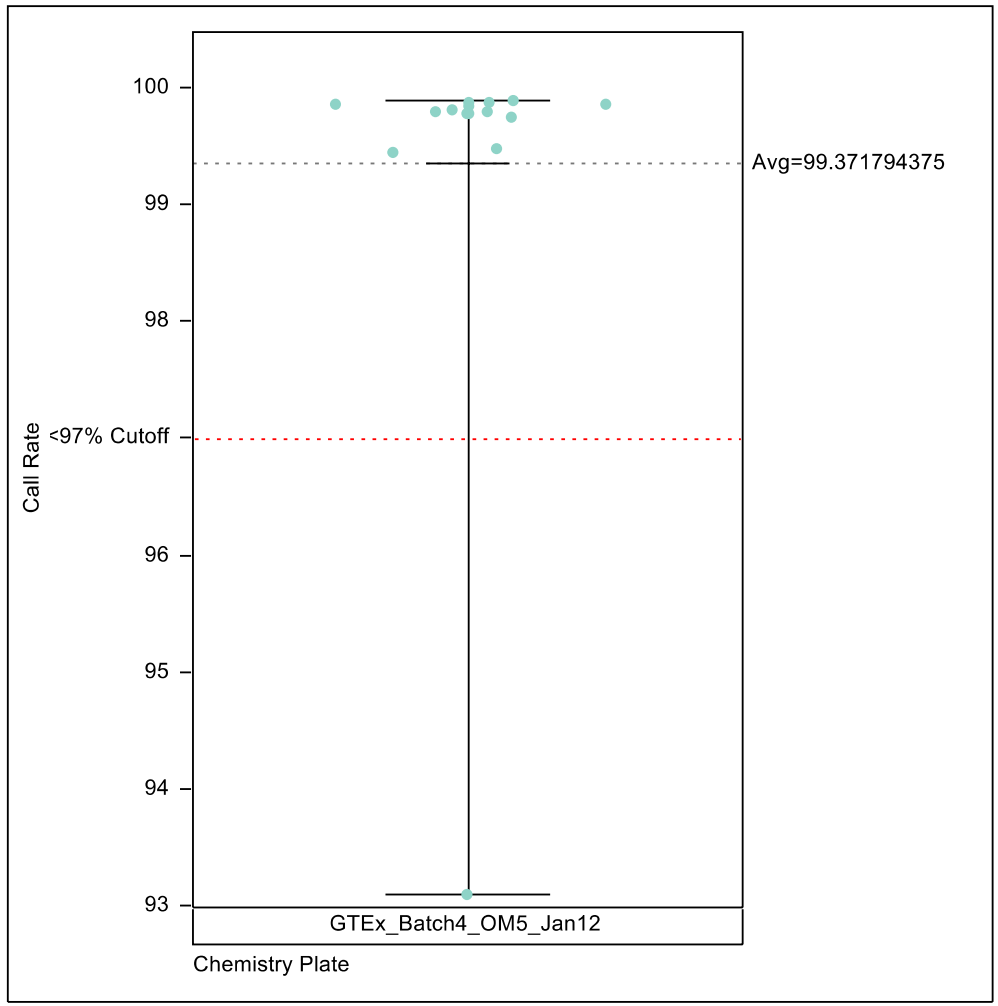
Completed by 3/1/12 Data freeze	Pilot Goal
64	300
376 (34)	1000
290 (38)	1500

Illumina 5M Genotyping

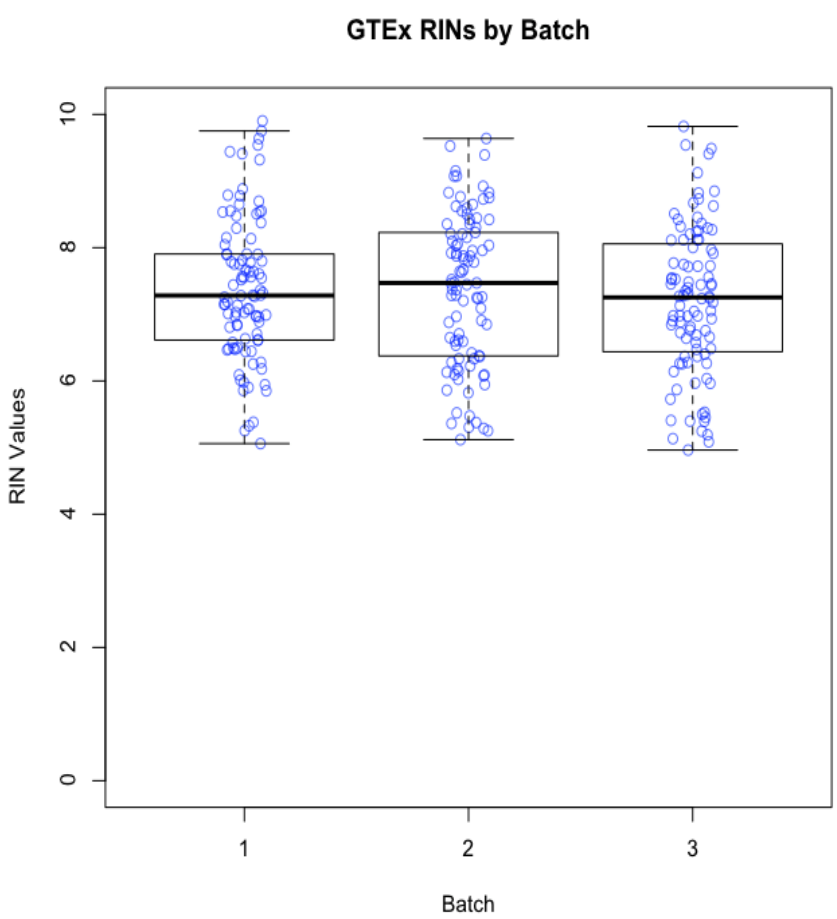
- Allele Frequency Distribution of ~5 million SNPs in GTEx 5M Batch 1
  - 100% pass rate
  - 99.35% average call rate
  - ~3.1 million polymorphic sites

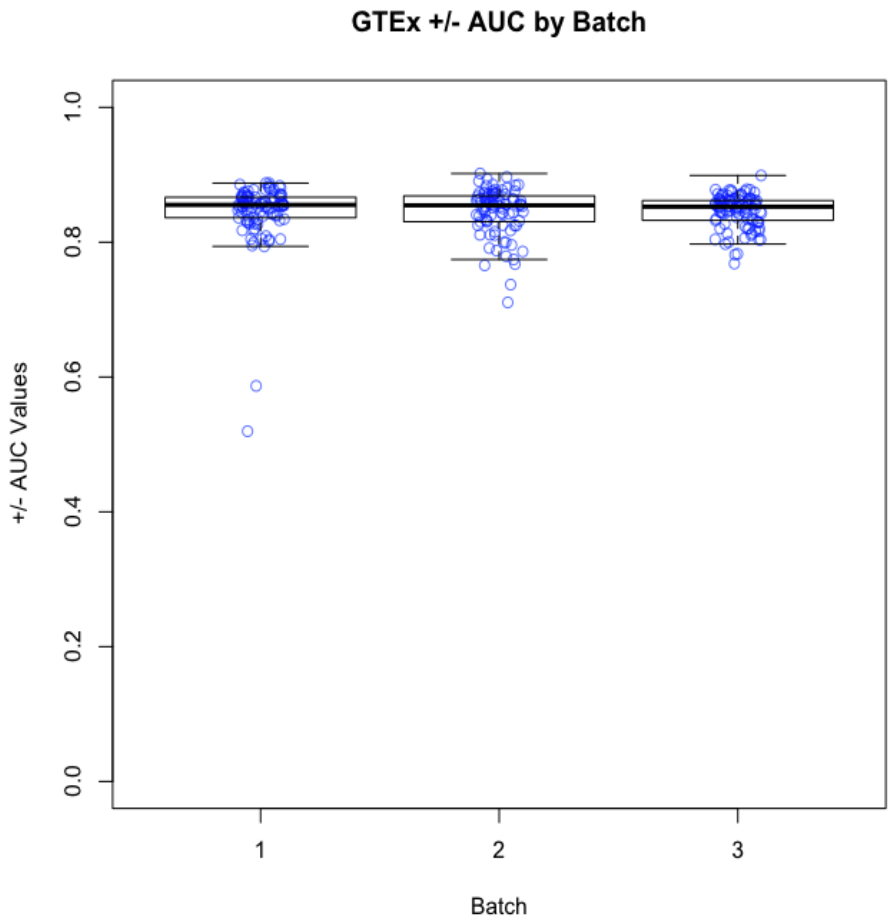
MAF	SNP Count
0	1156894
0.03-0.1	1615589
0.1-0.2	530866
0.2-0.3	362103
0.3-0.4	337539
0.4-0.5	298341

Batch 4



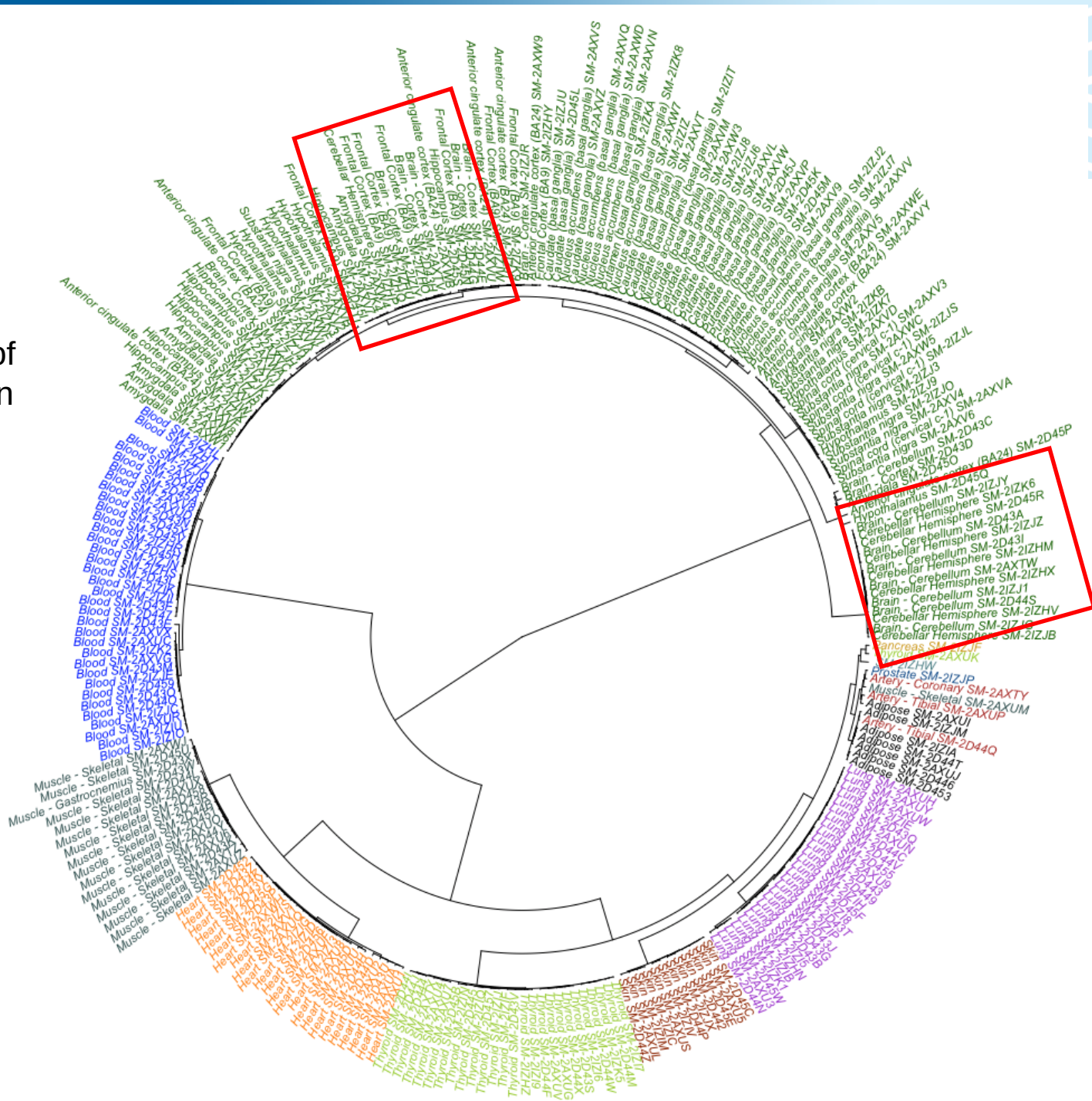
Expression Array QC by Batch





Hierachical Clustering

Distance measure:
Spearman Correlation of all Affymetrix expression levels



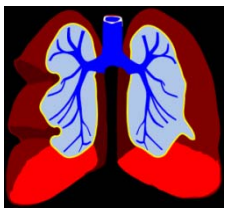
RNA Sequencing

- Tech dev efforts have focused on identifying a cDNA library construction method that provides high quality transcript profiles and:
  - Uses < 1ug total RNA input
  - Works with degraded RNA
  - Can be readily scaled to high-throughput
  - Works with preserved tissues (e.g. Paxgene Tissue)

Poly-A method (Tru-Seq, dUTP) vs non-poly-A DSN

GTEx Pilot RNA Experiments

GOAL → Select tissues for which the 3 methods produced similar, and at least mid-range, RIN scores for all samples, and for which there was sufficient RNA yield for both Affymetrix GeneArrays and RNASeq experiments.



Fresh Frozen, Liquid N2

→ RNA

RNAlater

→ RNA

→ RNA seq w/ DSN

PAXgene Tissue

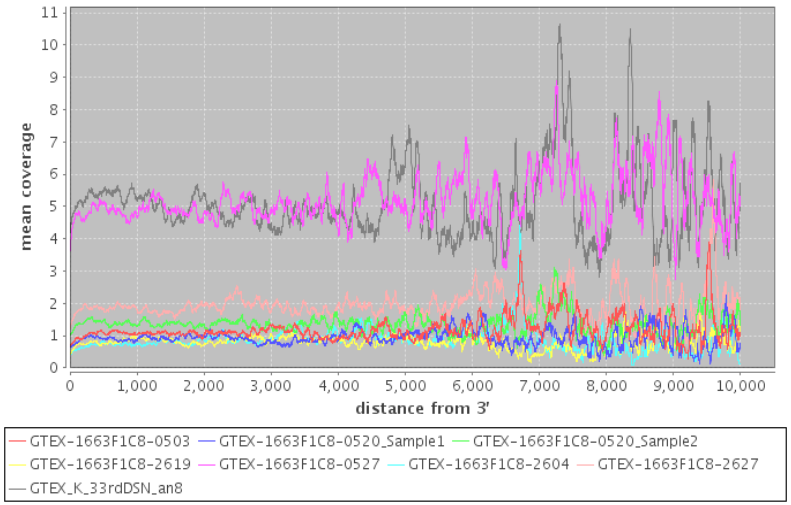
→ RNA

RNA Stable

RNA stable

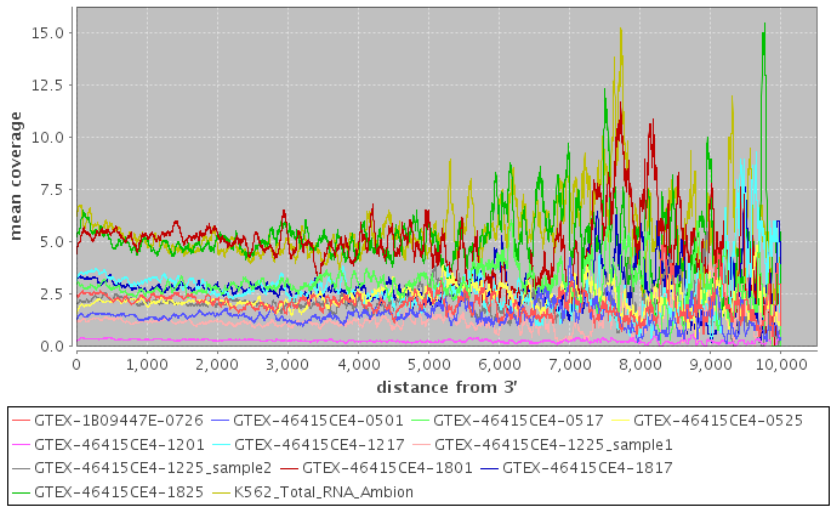
GTEx Pilot Tissue experiment results

*Based on 1000 middle expressed transcripts



Donor 2

Mean dup. rate: 73%
Mean CV: 1.23
Mean Gap %: 41%

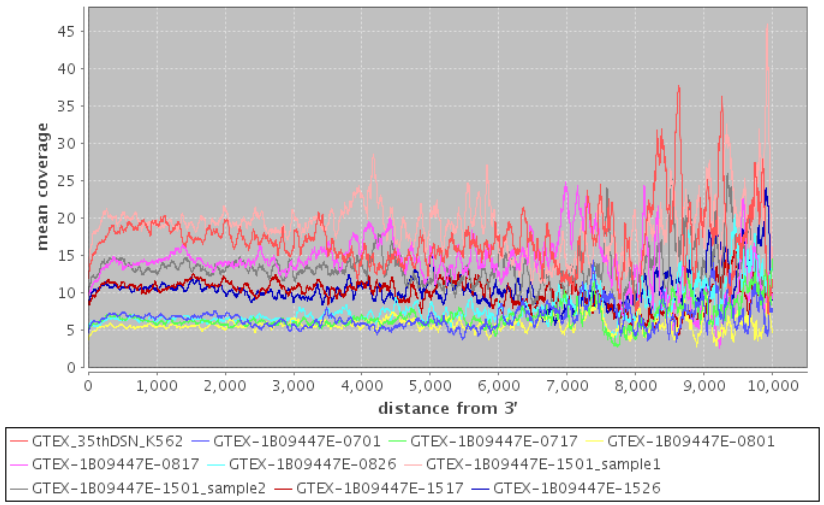


Donor 4

Mean dup. rate: 64%
Mean CV: 1.00
Mean Gap %: 19%

Donor 3

Mean dup. rate: 54%
Mean CV: 0.93
Mean Gap %: 11%



No evident difference between preservation methods.

RNA Seq First Production Batch

Donor	Tissue	RIN
GTEX-N7MS	Blood	8.2
GTEX-N7MS	Brain	8.7
GTEX-N7MS	Brain	9.6
GTEX-N7MT	Lung	9.1
GTEX-N7MT	Brain	9.5
GTEX-NPJ8	Blood	5.5
GTEX-NPJ8	Lung	8.8
GTEX-NPJ8	Brain	8.8
GTEX-NPJ8	Brain	9.4
GTEX-O5YV	Lung	6.1
GTEX-O5YV	Muscle	7.1

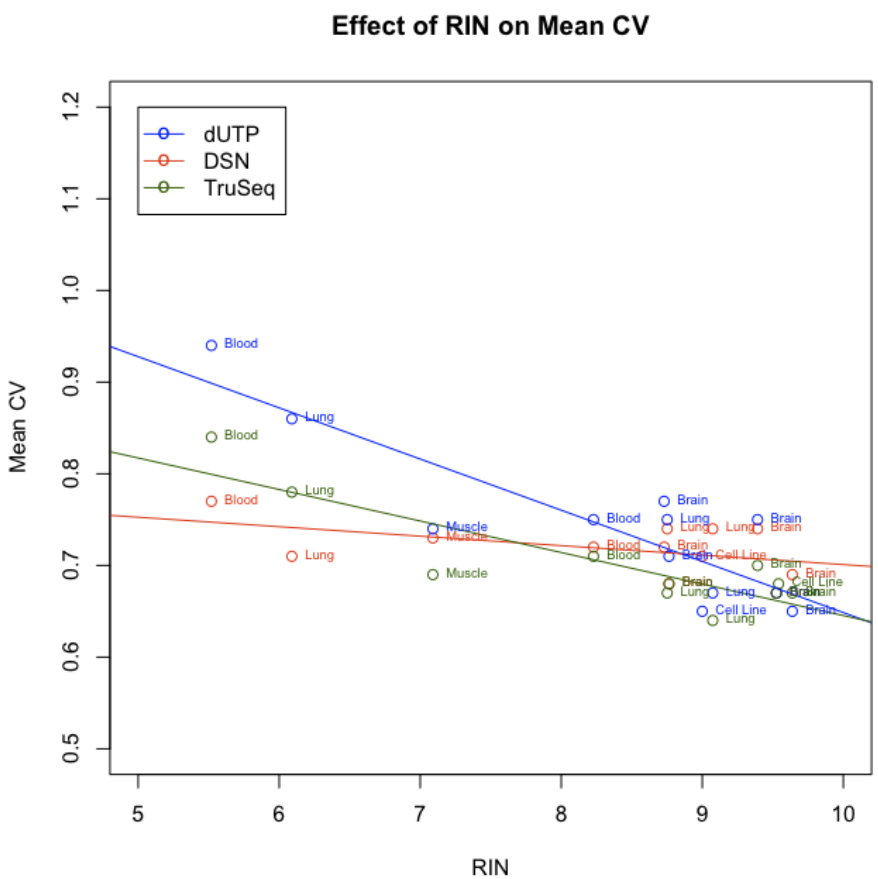
4 ug polyA+ LC
(2 methods)

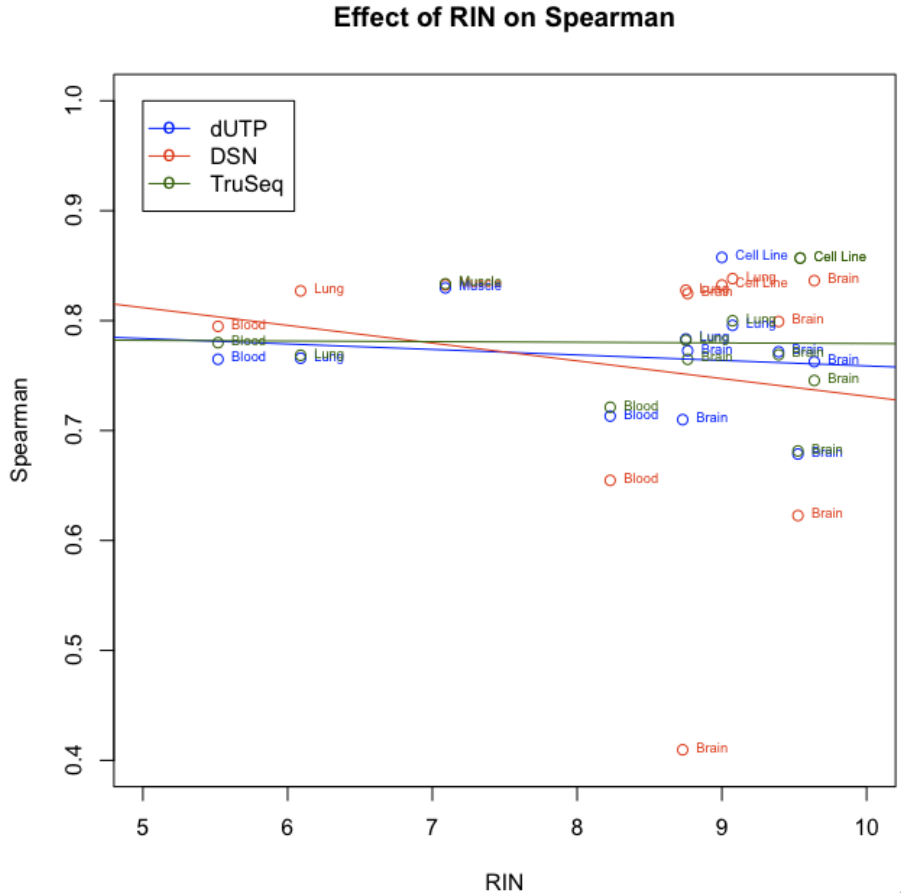
1 ug DSN LC

1/6th lane, 101 PE

FIREHOSE
Genotype-Tissue Expression

Coverage metrics





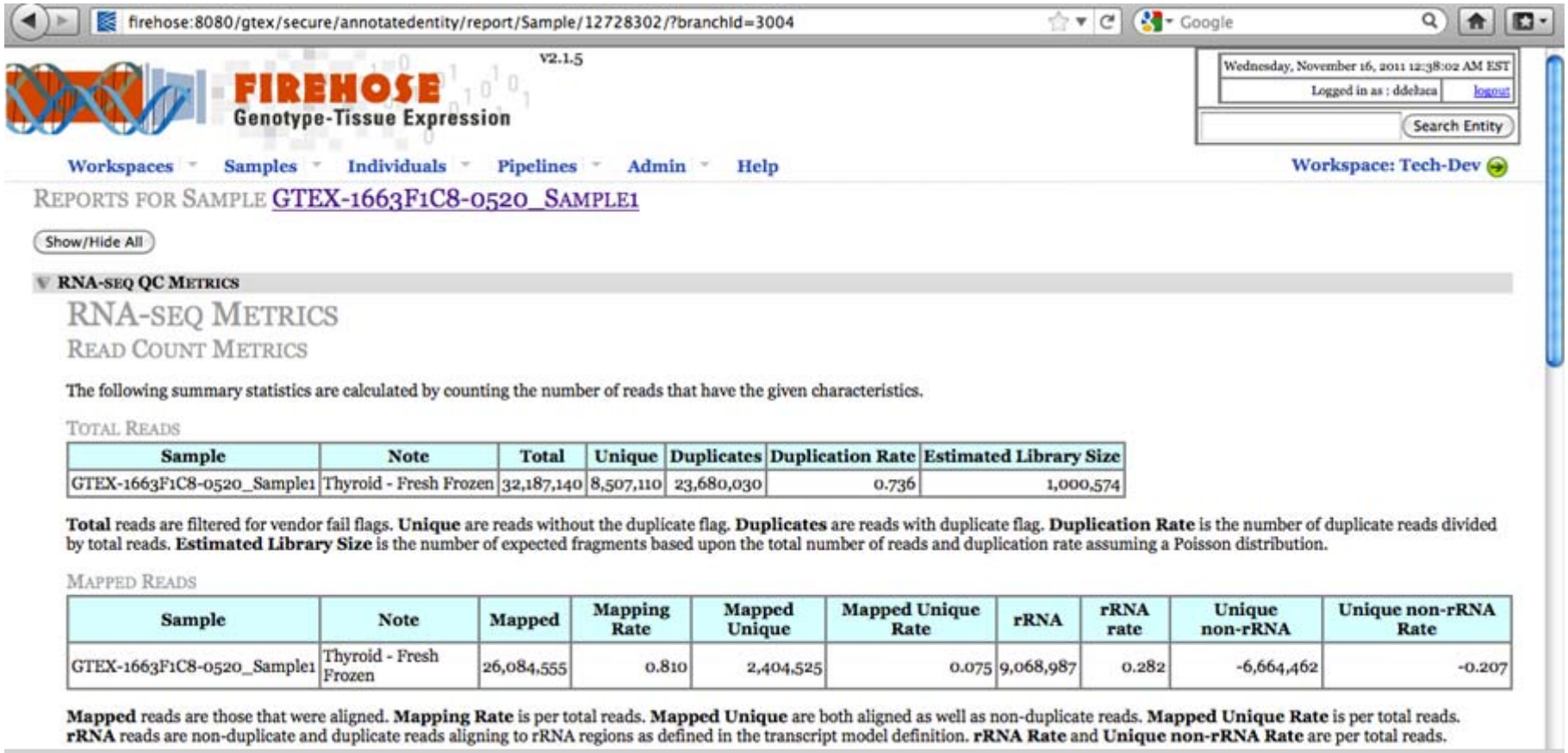
For middle expressed 1000 transcripts, downsampled to 20m reads.

LC Protocol Summary

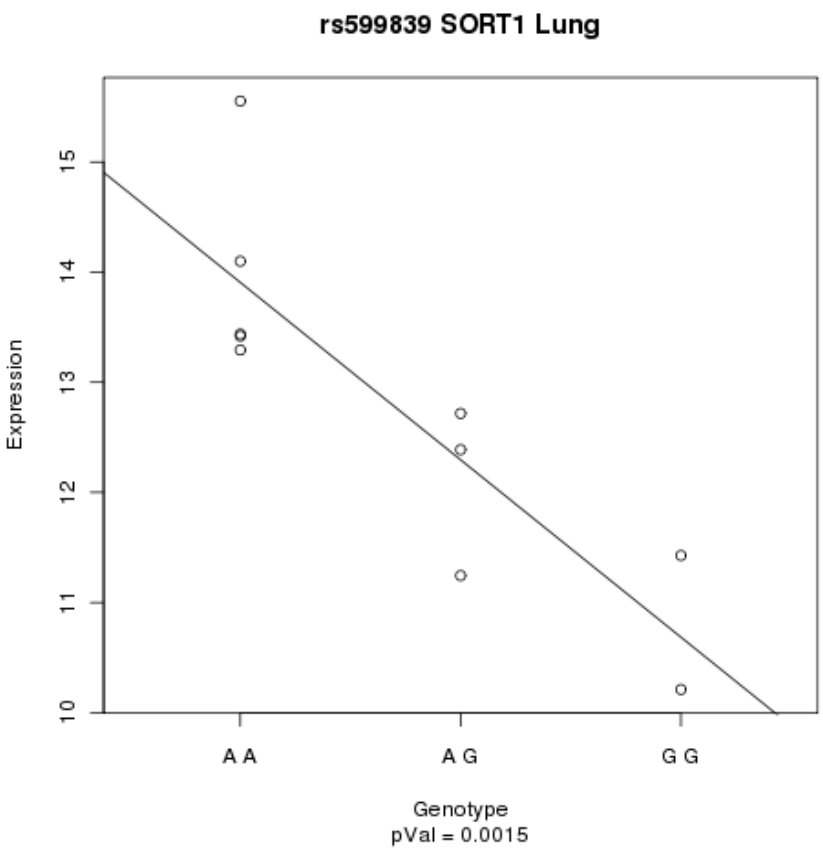
- For RINs >=7, polyA+ and DSN perform similarly
  - polyA does seem to be adversely affected by RINs <6
- BUT, DSN has significant disadvantages compared to polyA+ in terms of protocol robustness, GC bias, and cost
- polyA+ protocol currently requires 1-5ug RNA, but ongoing testing aims to reduce that to 0.1-1ug

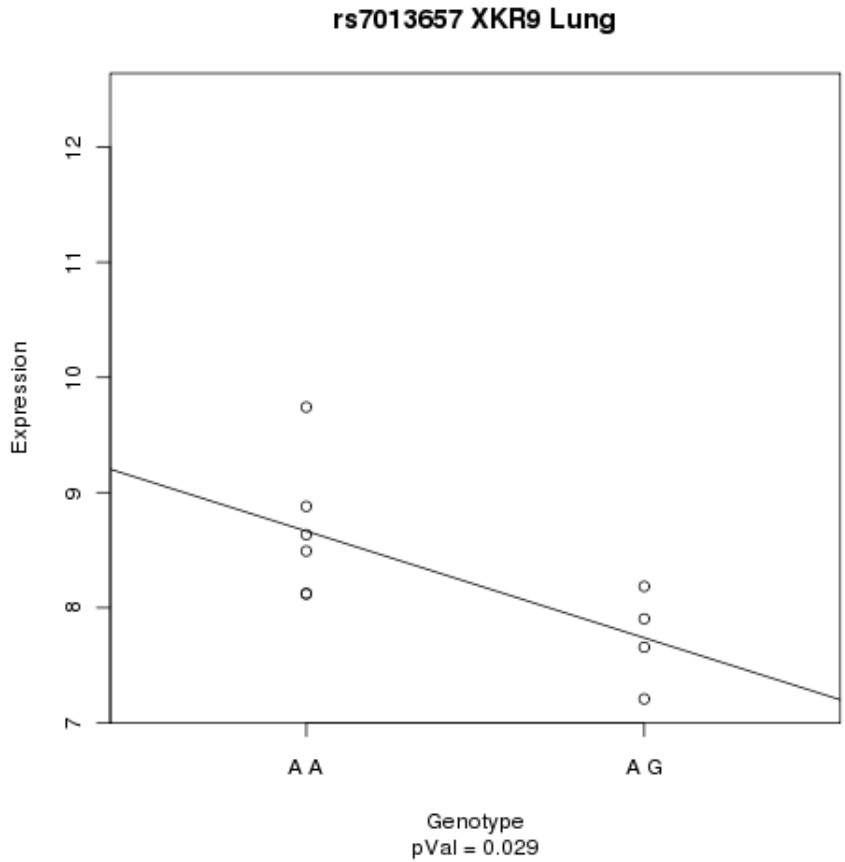
CURRENT PRODUCTION – polyA+ protocol, with 1ug RNA input and RIN>=6

RNA Seq Analysis Pipeline

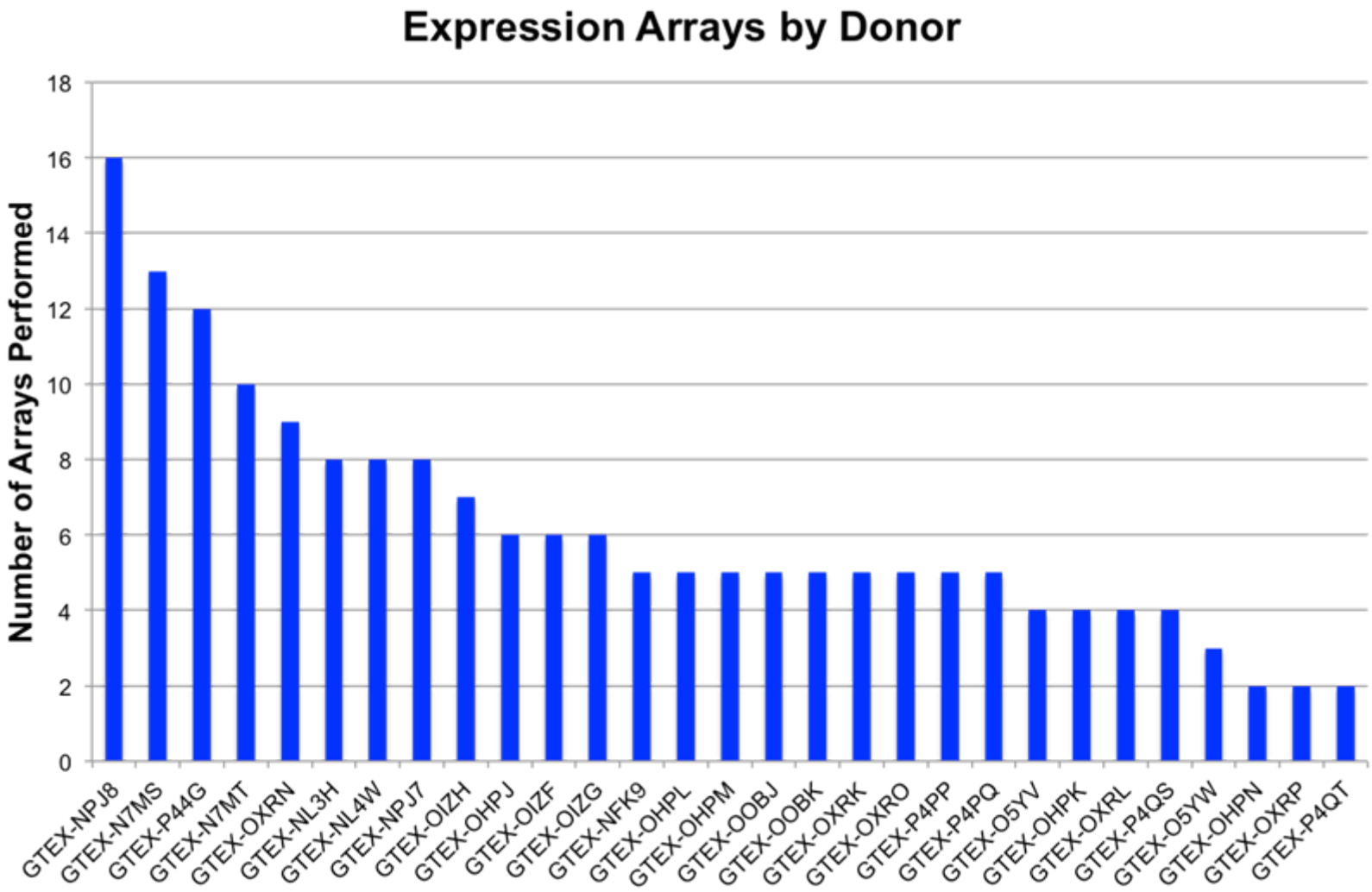


Known Liver eQTLs found in Lung





Current Power Limitations



Broad LDACC Acknowledgements

BSP
Julie Ann
Clint Chalk
Cara Fischer
Kim Kanki
Tsheko Makuwa
Paula Morais
BSP Staff
Kristin Ardlie

INFORMATICS
Michael Dinsmore
George Grant
Alex Thomson
Taylor Young
Tim Fennell
GAP, BSP, GSP teams

GAP
Yves Boie
Andrew Crenshaw
Supriya Gupta
Melissa Parkin
Prapti Pokharel
GAP Staff
Wendy Winckler (co-PI)

PROJECT MANAGEMENT
Molly Donovan
Ellen Gelfand
Kristin Thompson

ANALYSIS
David Deluca
Andrey Sivachenko
Gad Getz

GSP
Sheila Fisher
Ryan Johnson
Paula Maness
Jane Wilkinson
GSP staff
Jen Baldwin

GSAP
Xian Adiconis
Jim Bochicchio
Joshua Levin
Carsten Russ
Chad Nusbaum

Harvard Skin Disease Research Center
Jim Rheinwald
Patty Barron