Working with a whole bunch of genetic data

...and teaching data science

Shannon E. Ellis
Assistant Teaching Professor
COGS1 (Spring 2019)

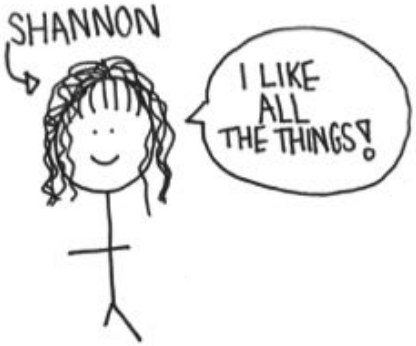
A quick tour of a geneticist turned data scientist

Background

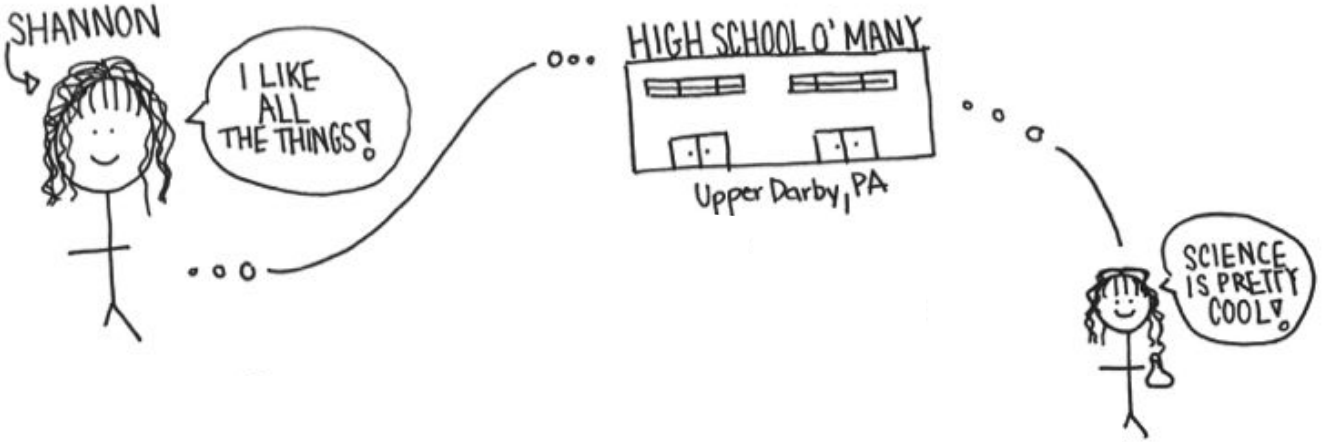
Projects

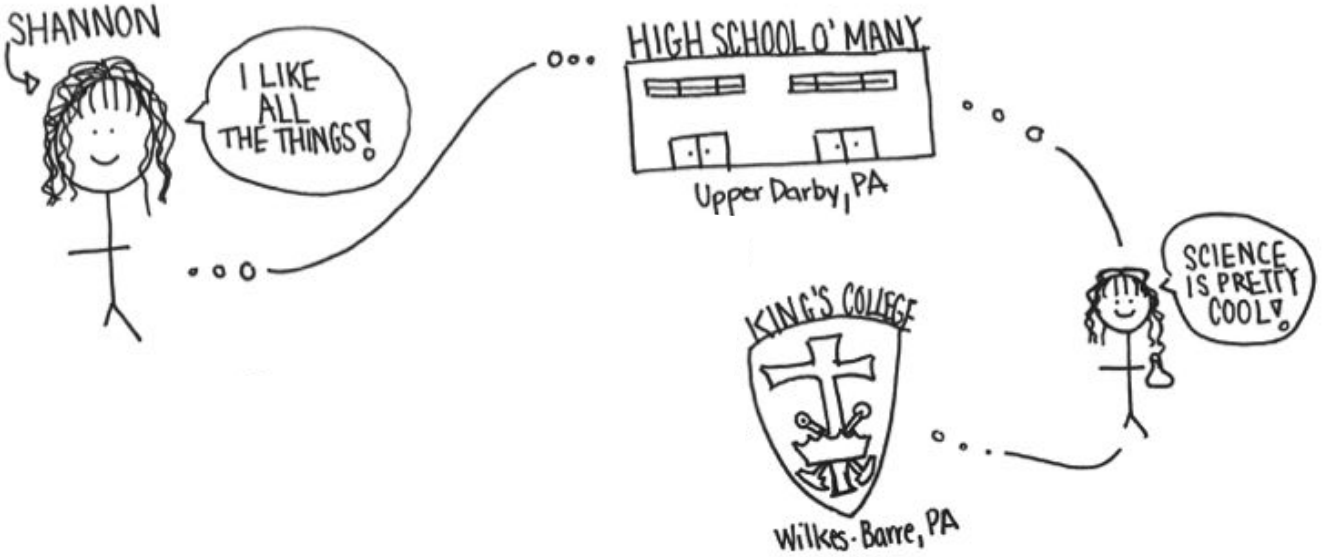
1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

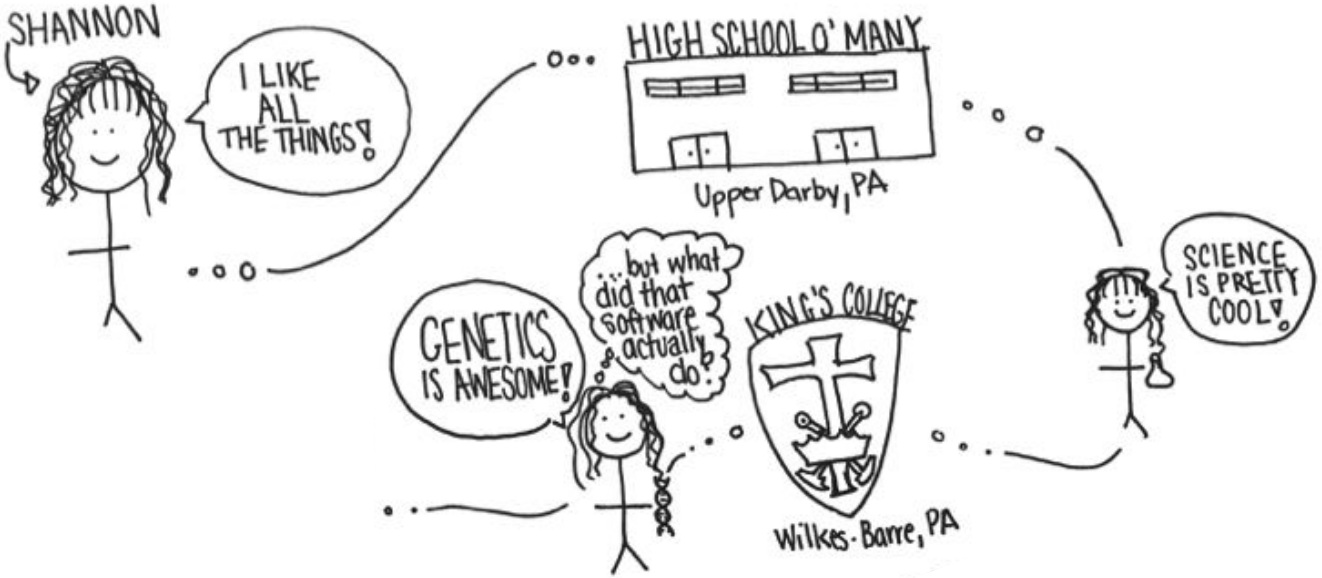
What I do here at UCSD

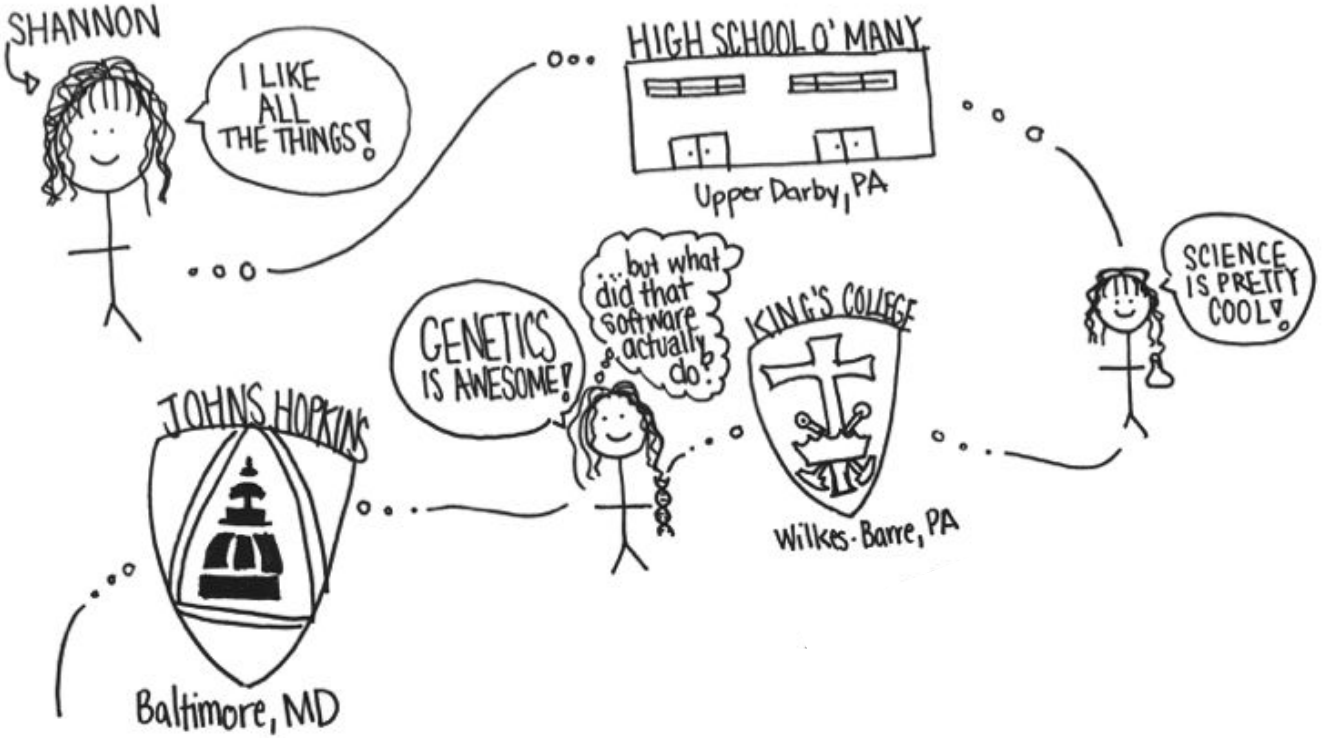

SHANNON
I LIKE ALL THE THINGS!

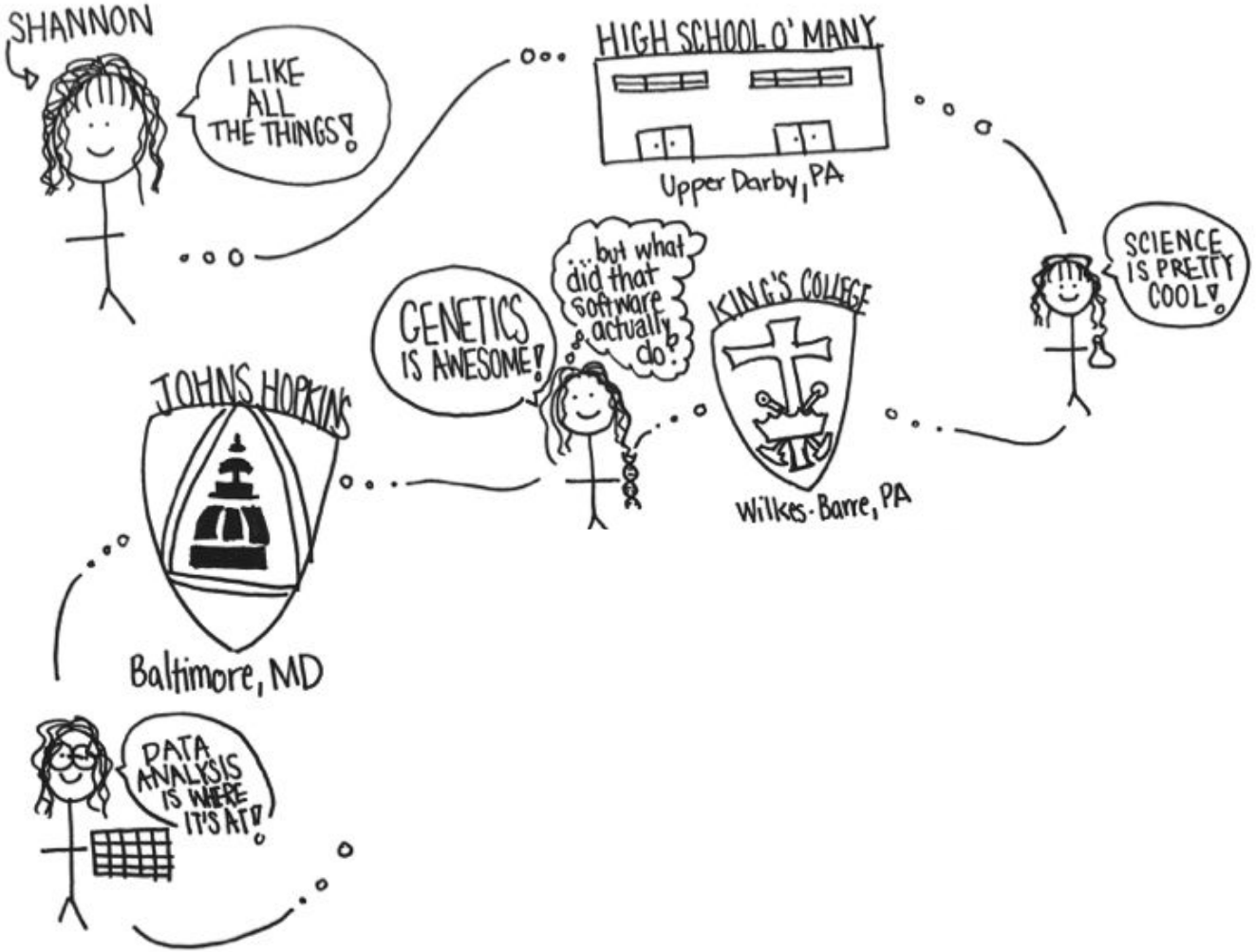

OPEN SCIENCE
TODAY
INFECTIONS
HUMAN GENETICS
DATA ANALYSIS

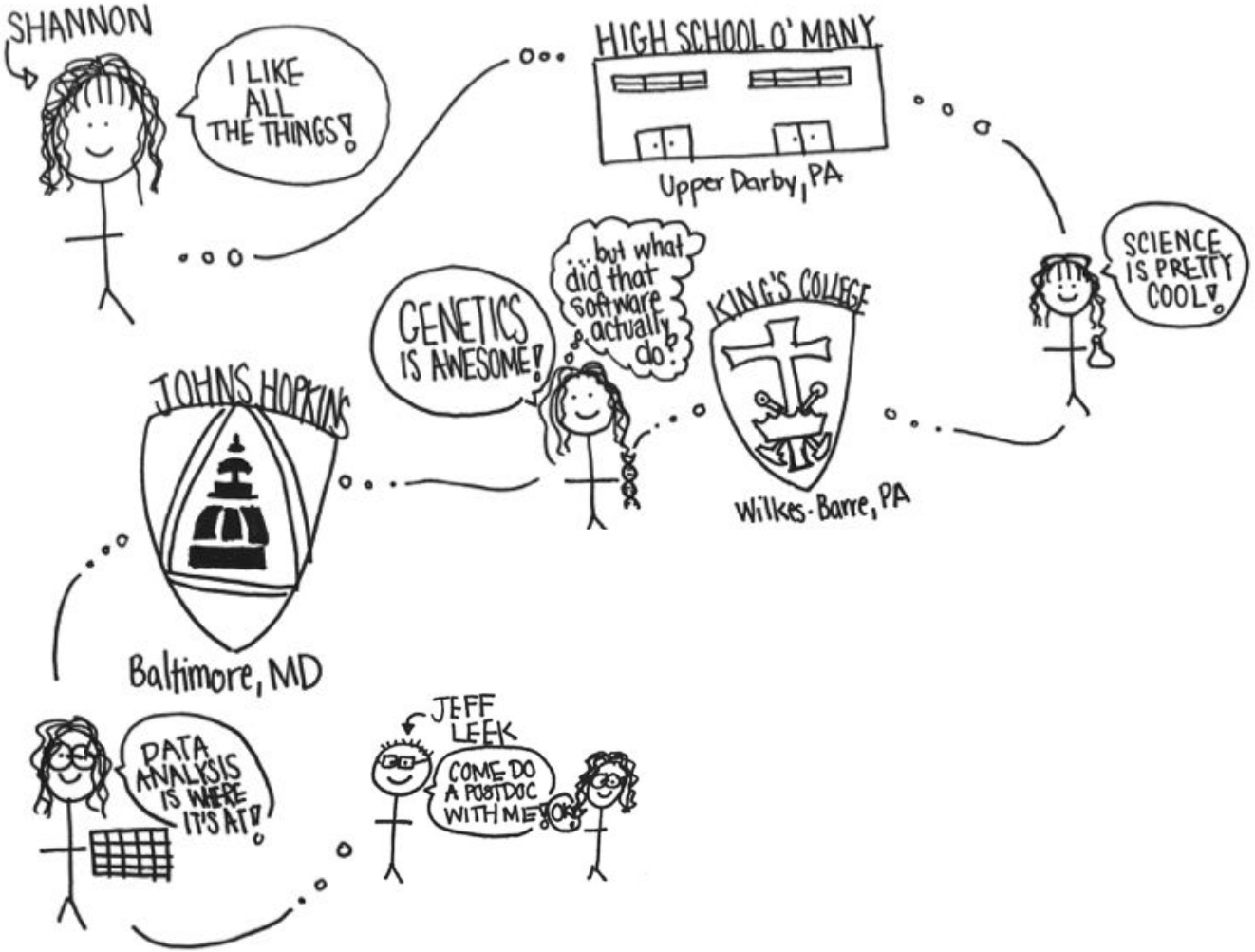

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!

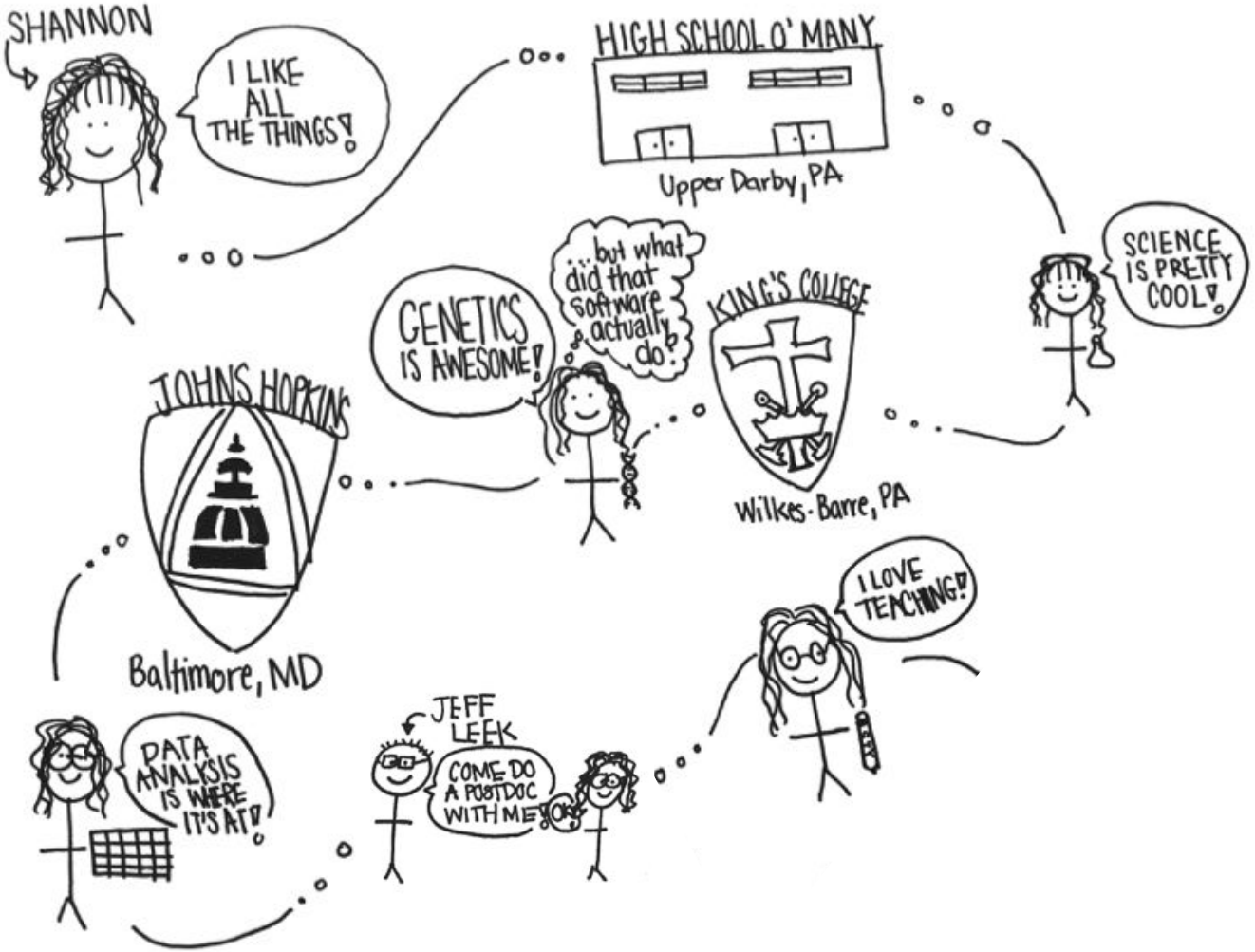

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
KING'S COLLEGE
Wilkes-Barre, PA
SCIENCE IS PRETTY COOL!

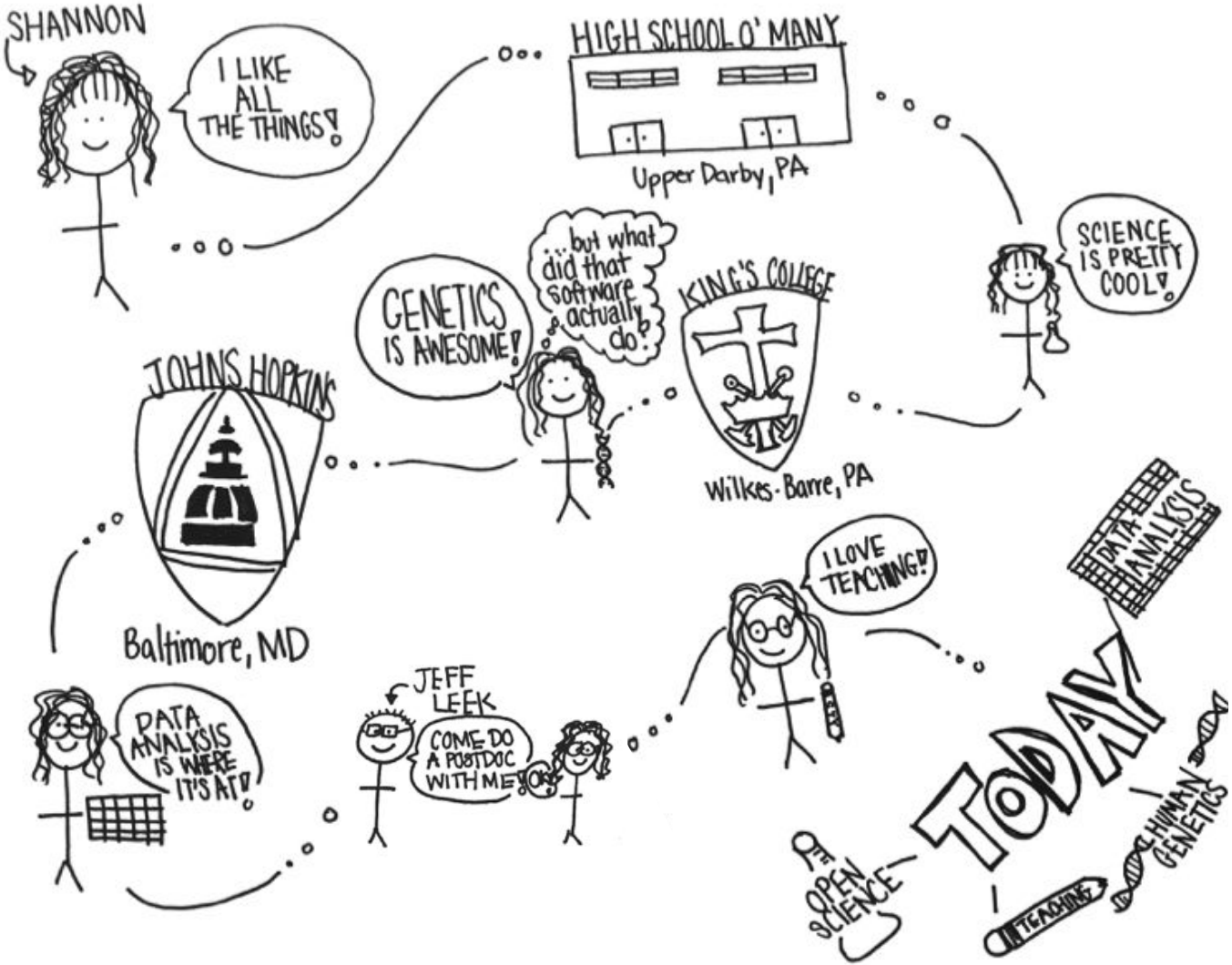

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
GENETICS IS AWESOME!
...but what did that software actually do?


SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Baltimore, MD


SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Baltimore, MD
DATA ANALYSIS IS WHERE IT'S AT!


SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Baltimore, MD
DATA ANALYSIS IS WHERE IT'S AT!
JEFF LEEK
COME DO A POSTDOC WITH ME!
OK!


SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Baltimore, MD
I LOVE TEACHING!
JEFF LEEK
COME DO A POSTDOC WITH ME!
OK!
DATA ANALYSIS IS WHERE IT'S AT!


SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Baltimore, MD
I LOVE TEACHING!
DATA ANALYSIS
DATA ANALYSIS IS WHERE IT'S AT!
JEFF LEEK
COME DO A POSTDOC WITH ME!
OK!
TODAY
OPEN SCIENCE
TEACHING
HUMAN GENETICS

A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD

The quickest history of human genetics

1900 - Rediscovery of Mendel’s work

1950s - DNA is the genetic material & Structure of DNA

2001 - Human Genome Project

- Thought this would unlock understanding of all disease!
- Not exactly how things worked out - way more complex than initially thought

Mid-2000s - GWAS set out to sort all this out

- 2007: 240 papers published
- 2017: 3800+

2010s - a move toward additional approaches

Multi-omic Data Provide a More Complete Understanding of the Autistic Brain

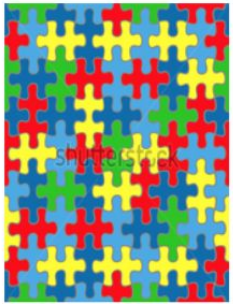
That's the title of my thesis dissertation. I'm not great at coming up with titles...ever.

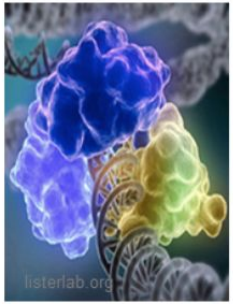
General Outline

I. Autism Background

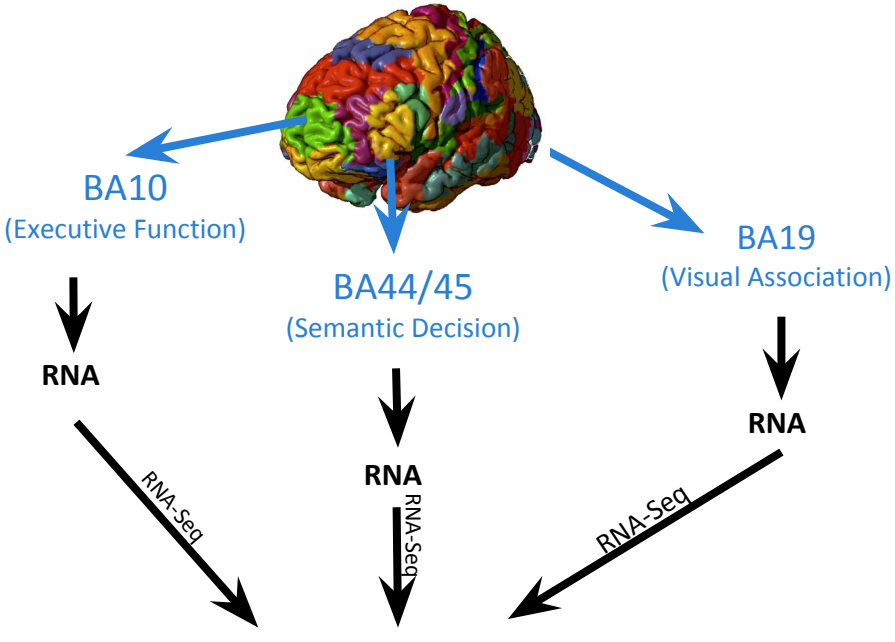
II. Transcriptome Analyses
   A. Gene expression differences in the autistic brain
   B. Cross-disorder transcriptomic overlap

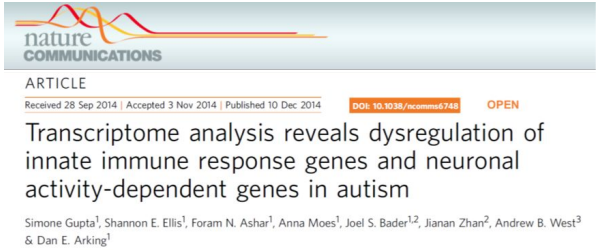
III. Epigenome of the Autistic Brain
   A. CpG methylation
   B. nonCpG methylation





RNA-Sequencing in Autism Brains





nature COMMUNICATIONS

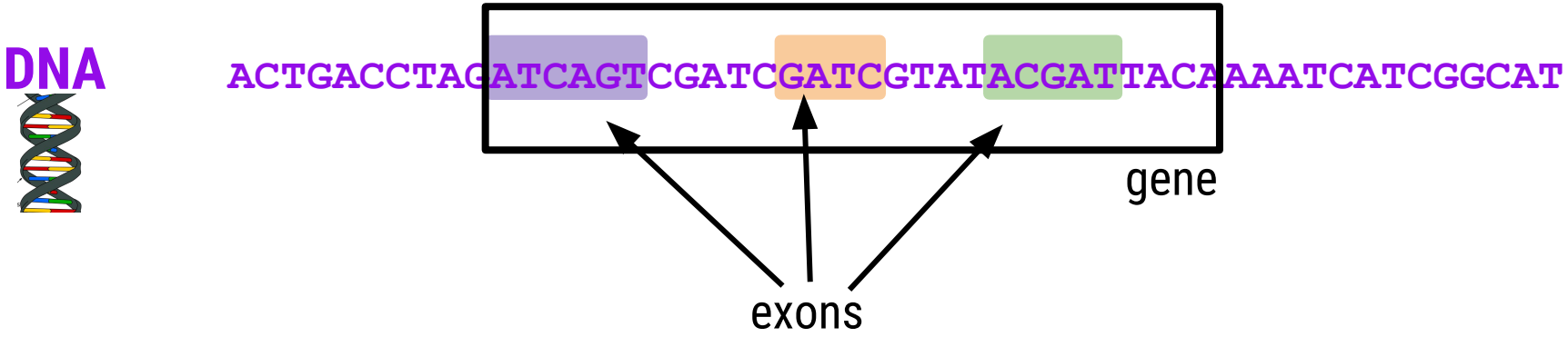
ARTICLE

Received 28 Sep 2014 | Accepted 3 Nov 2014 | Published 10 Dec 2014 DOI: 10.1038/ncomms6748 OPEN

Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism

Simone Gupta[1], Shannon E. Ellis[1], Foram N. Ashar[1], Anna Moes[1], Joel S. Bader[1,2], Jianan Zhan[2], Andrew B. West[3] & Dan E. Arking[1]

The Central Dogma of Genetics



slide adapted from jeff leek

The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

↓ **transcription**

RNA

AUCAGUCGAUCACCGAU

slide adapted from jeff leek

The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

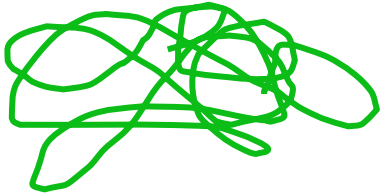
↓ **transcription**

RNA

AUCAGUCGAUCACCGAU

↓ **translation**

proteins



slide adapted from jeff leek

Two copies of **DNA** -> many **transcripts** -> many **proteins**

	role in the cell	# copies/cell	functional unit	# unique functional units
DNA 	*blueprint*	*2*	*gene*	*20,000*
RNA 				
proteins 				

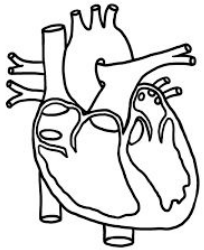
Two copies of **DNA** -> many **transcripts** -> many **proteins**

	role in the cell	***# copies/cell***	***functional unit***	***# unique functional units***
DNA 	*blueprint*	*2*	*gene*	*20,000*
RNA 				
proteins 	*carry out cellular functions*	*varies* $\sim 10^{10}$	*proteins* *(metabolites, hormones, etc.)*	*~100,000*

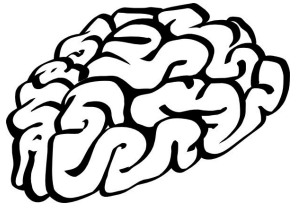
Two copies of **DNA** -> many **transcripts** -> many **proteins**

	role in the cell	***# copies/cell***	***functional unit***	***# unique functional units***
DNA 	*blueprint*	*2*	*gene*	*20,000*
RNA 	*messenger*	*varies ~360,000*	*transcript*	*~100,000*
proteins 	*carry out cellular functions*	*varies ~10^{10}*	*proteins (metabolites, hormones, etc.)*	*~100,000*

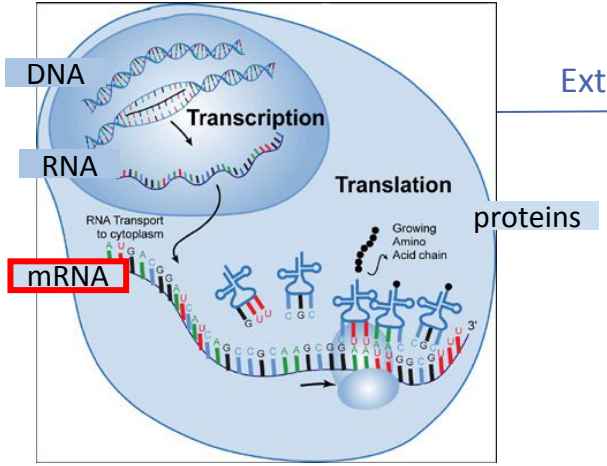
Variability at the level of RNA allows for a heart cell to function differently than a brain cell



!=

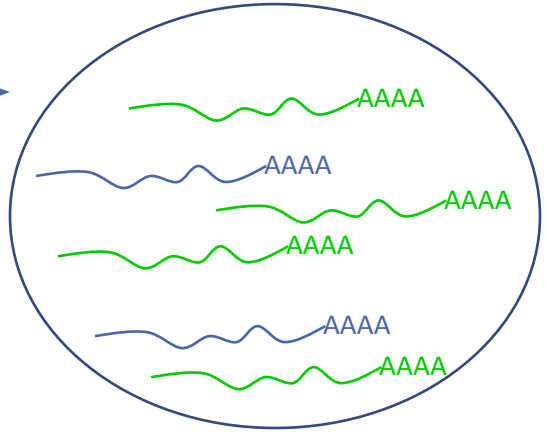


RNA-Sequencing



Extract mRNA

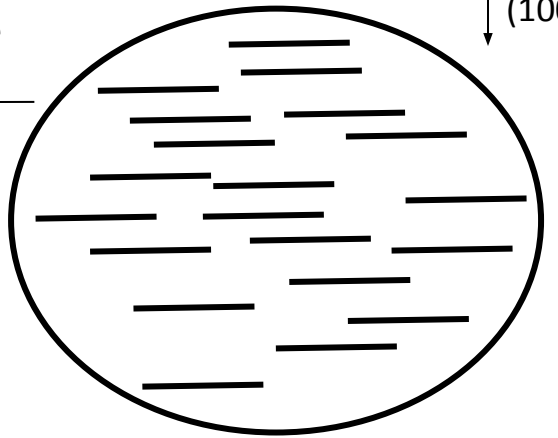
Pool of polyA+-selected mRNA



Prepare a sequencing library

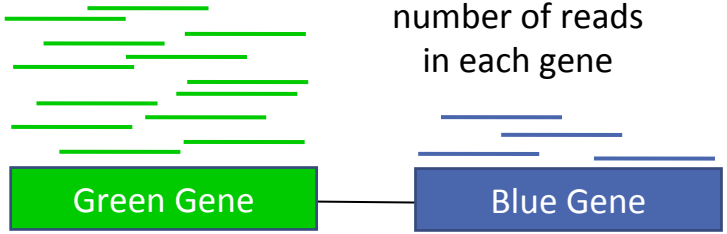
Illumina Sequencing (100bp SE)

Sequencing Output



Align those reads to the reference genome

Approximate gene expression by counting the number of reads in each gene



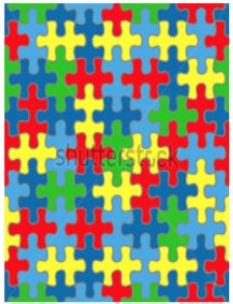
Green Gene > Blue Gene

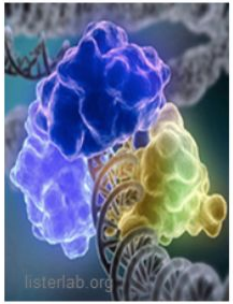
Results in a single slide

I. Autism Background

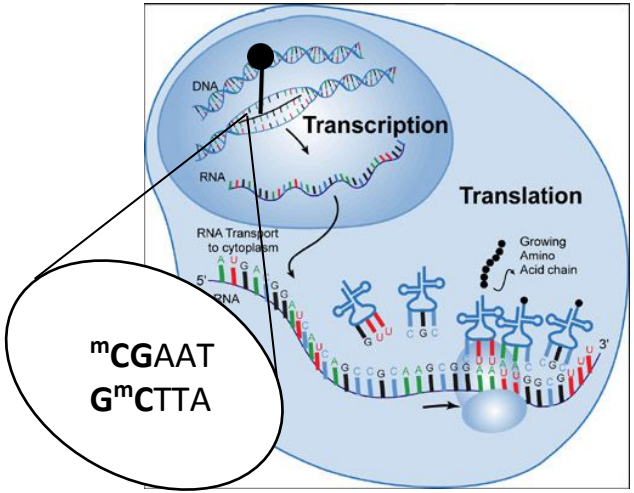
II. Transcriptome Analyses
   A. Microglia playing a role in the autistic brain
   B. RNA levels show similar patterns across conditions

III. Epigenome of the Autistic Brain
   A. CpG methylation
   B. nonCpG methylation



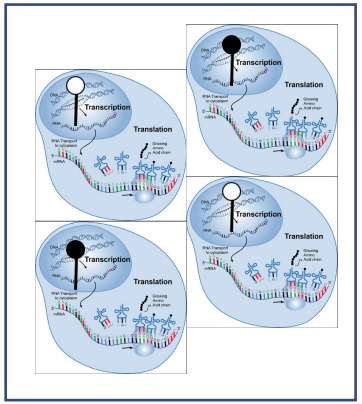


DNA methylation is most often studied at CpG dinucleotides

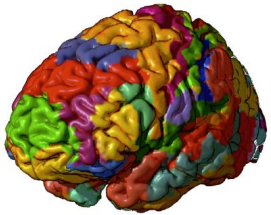


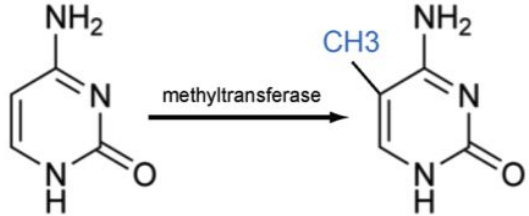
http://www.tokresource.org/

tissue



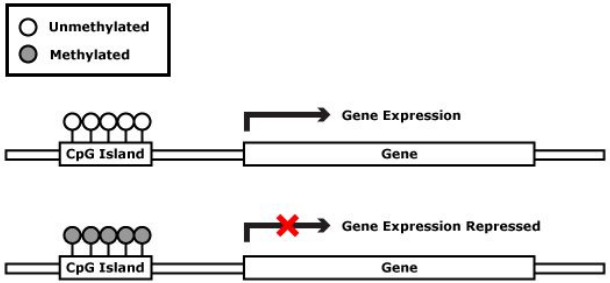
50% methylated





Cytosine (C)

5mC

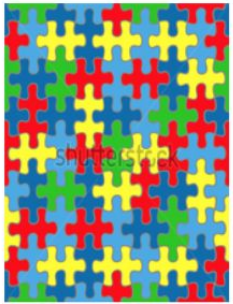


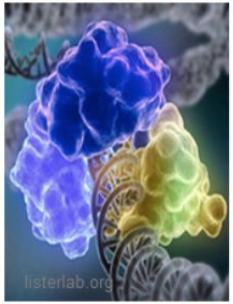
Results in a single slide

I. Autism Background

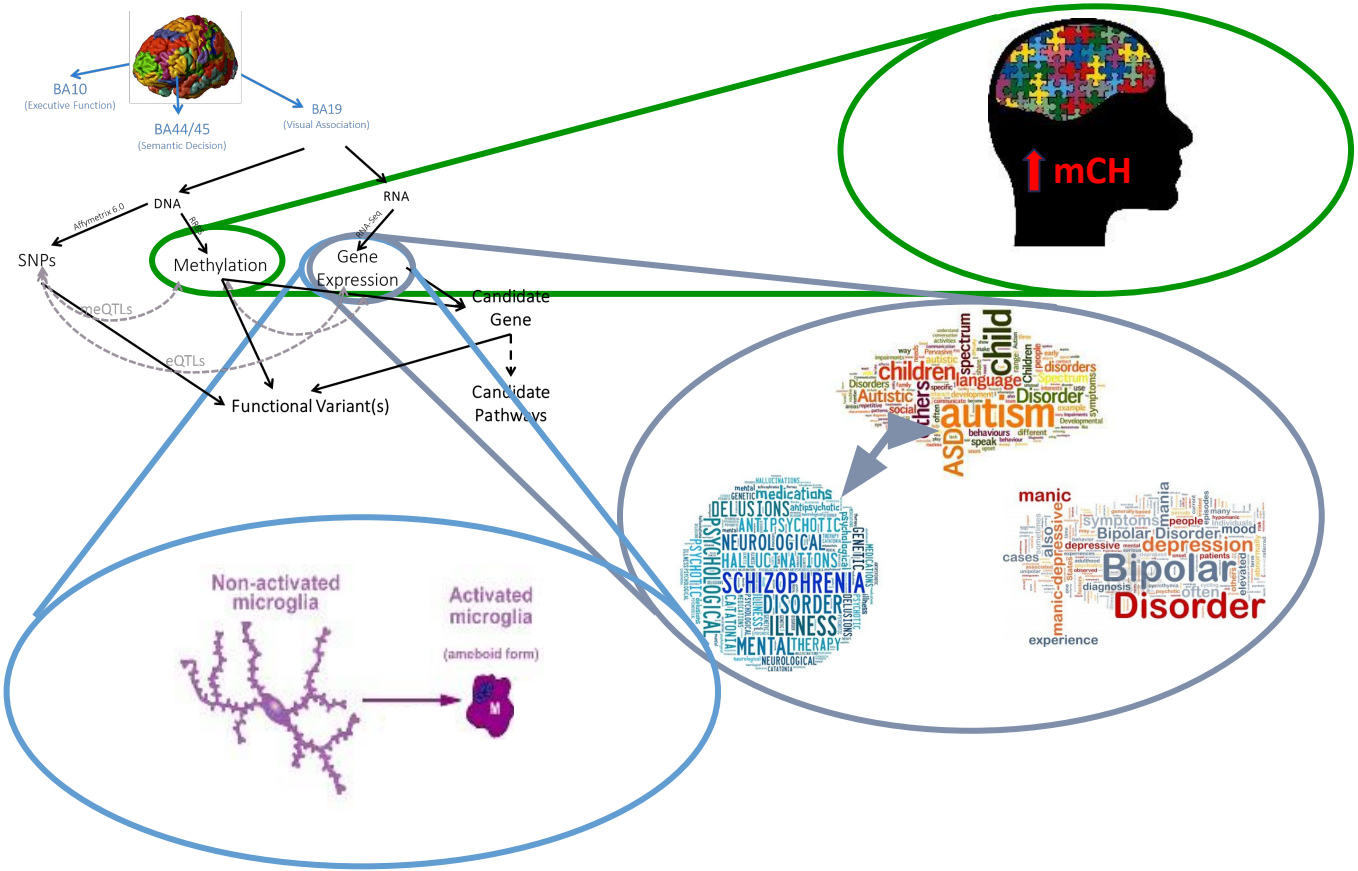
II. Transcriptome Analyses
   A. Microglia playing a role in the autistic brain
   B. RNA levels show similar patterns across conditions

III. Epigenome of the Autistic Brain
   A. CpG methylation does not differ
   B. Increased global nonCpG methylation





Conclusions: Toward a More Complete Understanding of the Autistic Brain



Scientific Acknowledgments (PhD work)

The Arking Lab


Dan E. Arking, PhD


Foram N. Ashar

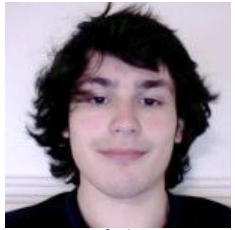

Nathan Bilhmayer


Pallav Bhatnagar, PhD


Christina Castellani, PhD


Rebecca Eggebeen


Simone Gupta, PhD


Naftali Horowitz


Ryan Longchamps


Anna Moes, MS


Rebecca Panitch


Elizabeth Vincent

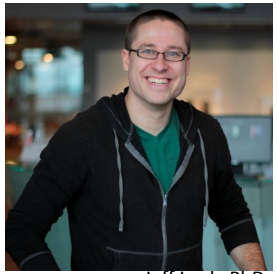
Collaborators

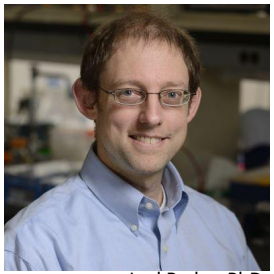

Shan Andrews


Andrew West, PhD

Thesis Committee


Dani Fallin, PhD


Jeff Leek, PhD


Joel Bader, PhD


Juan Troncoso, MD

Human Genetics Program


David Valle, MD


Kirby Smith, PhD


Sandy Muscelli

A quick tour of a geneticist turned data scientist

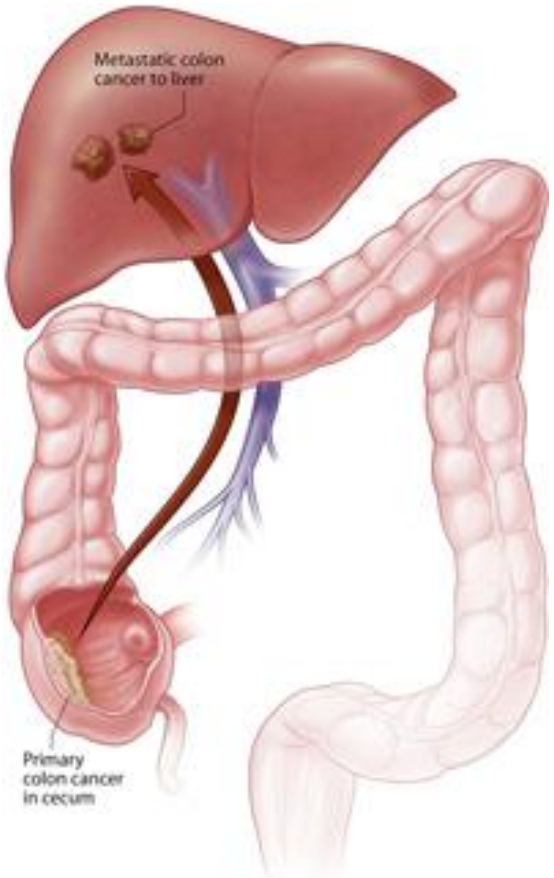
Background

Projects

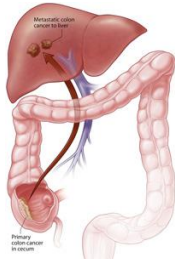
1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD

What makes primary cancer different than metastatic cancer?

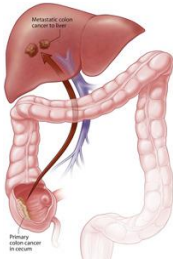


www.hopkinsmedicine.org



What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

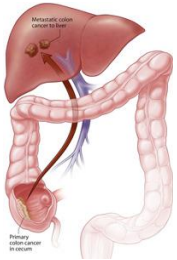


What makes primary cancer different than metastatic cancer?

Collect patient samples and information

~6 months

Find a researcher with access to patient samples



What makes primary cancer different than metastatic cancer?

Collect patient samples and information

Sequence samples

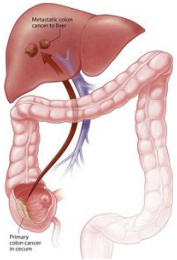
~6 months

1-2 wks

2-4 wks

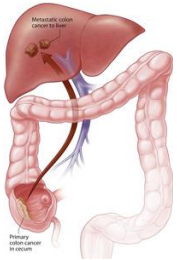
Find a researcher with access to patient samples

Extract DNA/RNA from samples



What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples	Collect patient samples and information	Extract DNA/RNA from samples	Sequence samples	Process sequencing data	Data cleaning	Analyze data and answer biological question
	~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

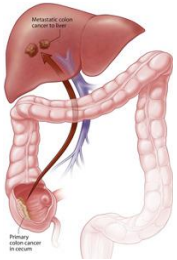


What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples	Collect patient samples and information	Extract DNA/RNA from samples	Sequence samples	Process sequencing data	Data cleaning	Analyze data and answer biological question
	~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

Total: 2+ years

Biologists have recently gotten pretty good at making their data available to the public.



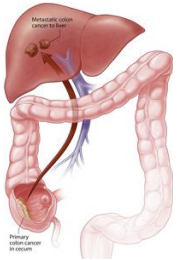
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples	Collect patient samples and information	Extract DNA/RNA from samples	Sequence samples	Process sequencing data	Data cleaning	Analyze data and answer biological question
	3~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

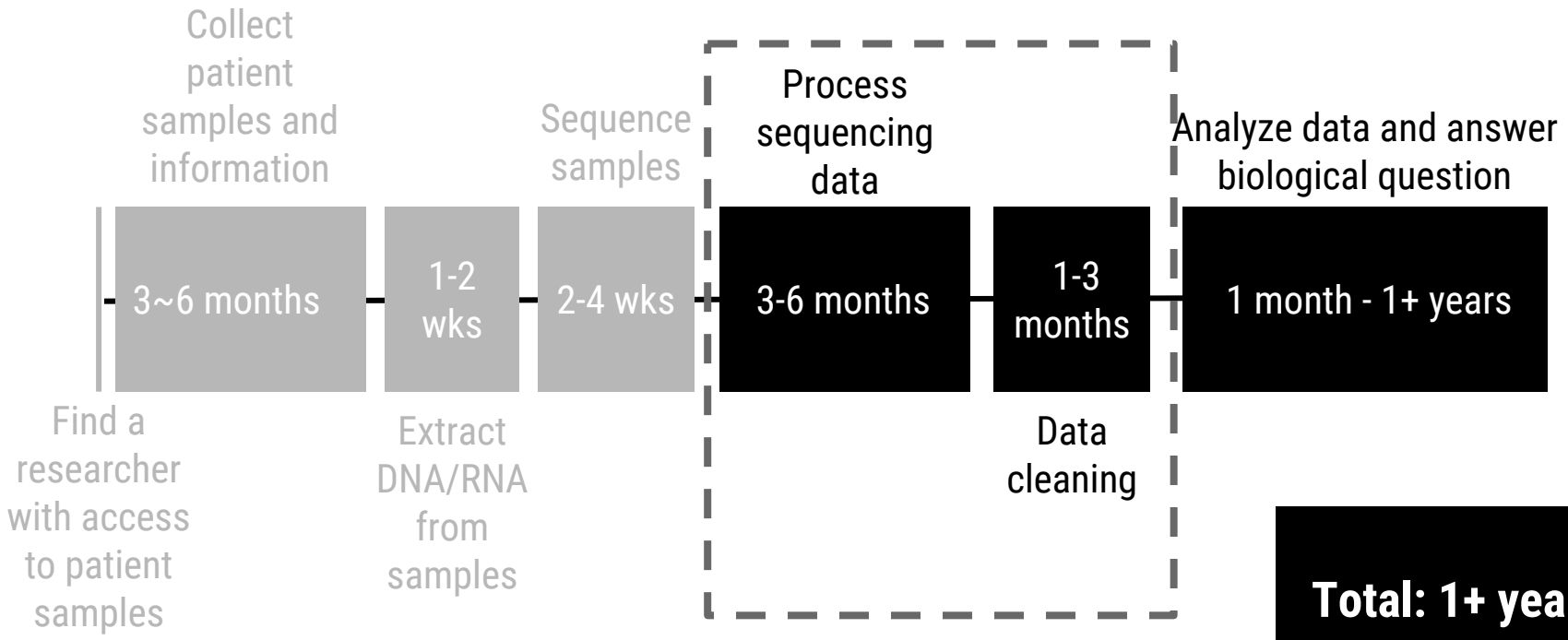
Total: 1+ years

Biologists have recently gotten pretty good at making their data available to the public.

...but they're *not great* at making these data <u>easily accessible</u> and <u>well-annotated</u>.



What makes primary cancer different than metastatic cancer?

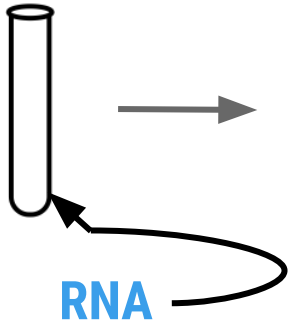


Total: 1+ years

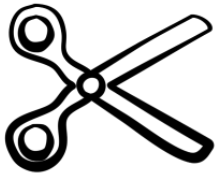
Measuring Transcription

slide adapted from jeff leek

Next Generation Sequencing (NGS) in one slide

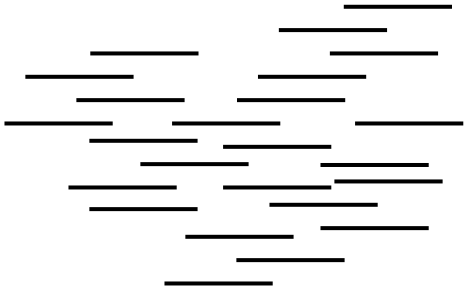


Step 1: Extract RNA to get sample of interest



Step 2: Chop up RNA into smaller pieces

Step 3: Sequence the sample

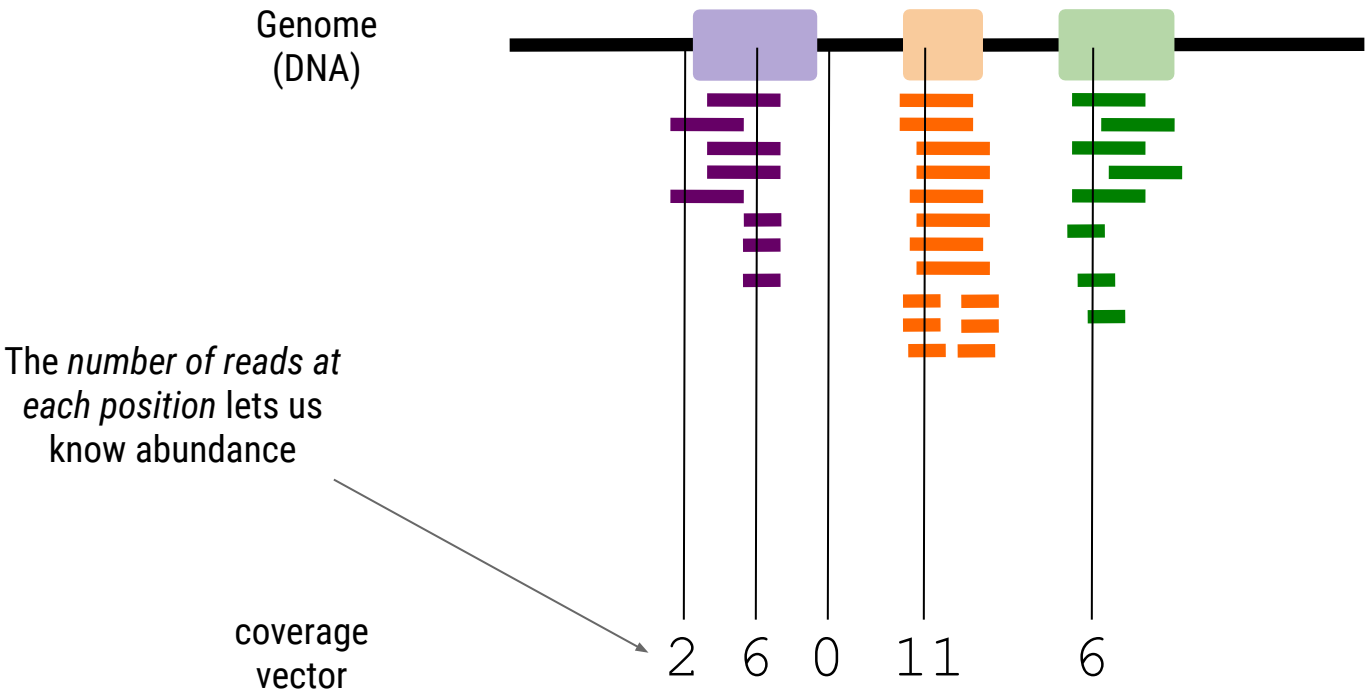


Step 3: Obtain short read data from the sequencer

Next Generation Sequencing (NGS) in one slide

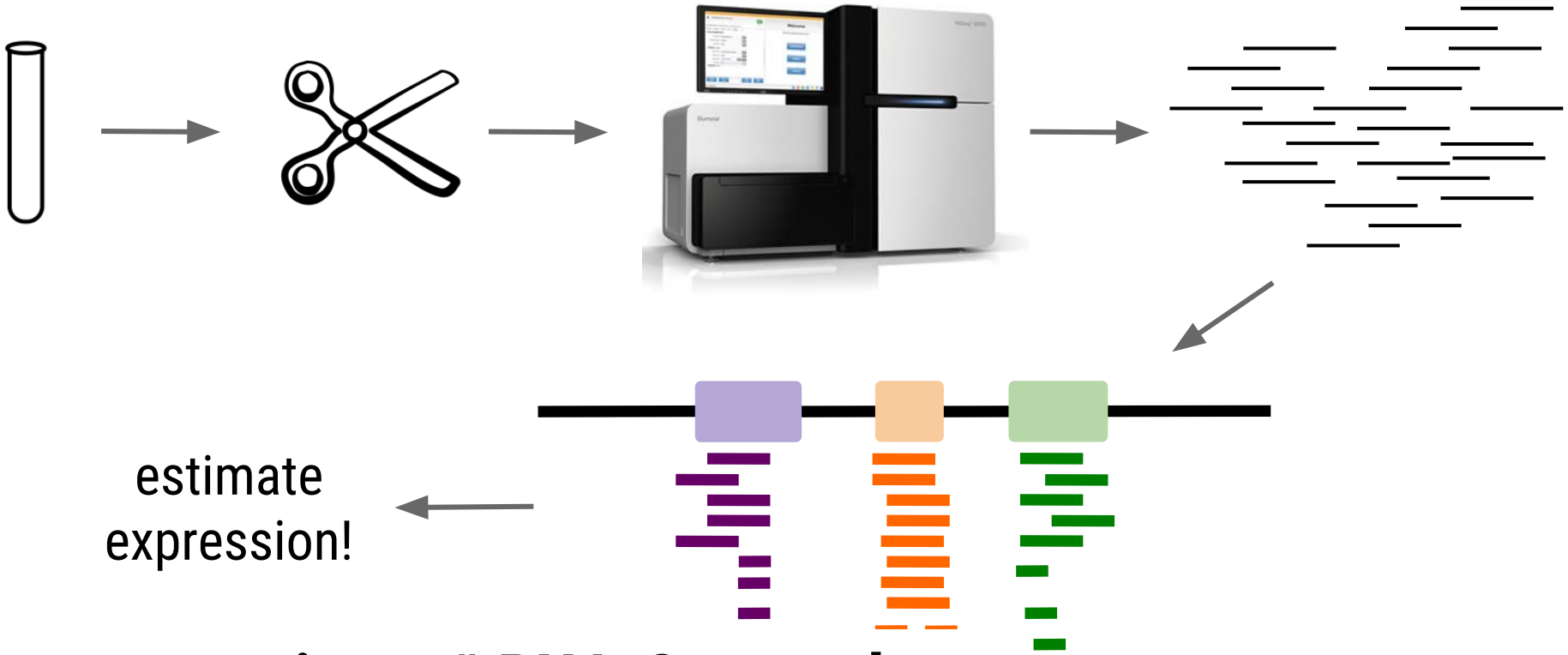
A short read tells you the sequence of that read

We first need to align these reads back to the genome



slide adapted from jeff leek

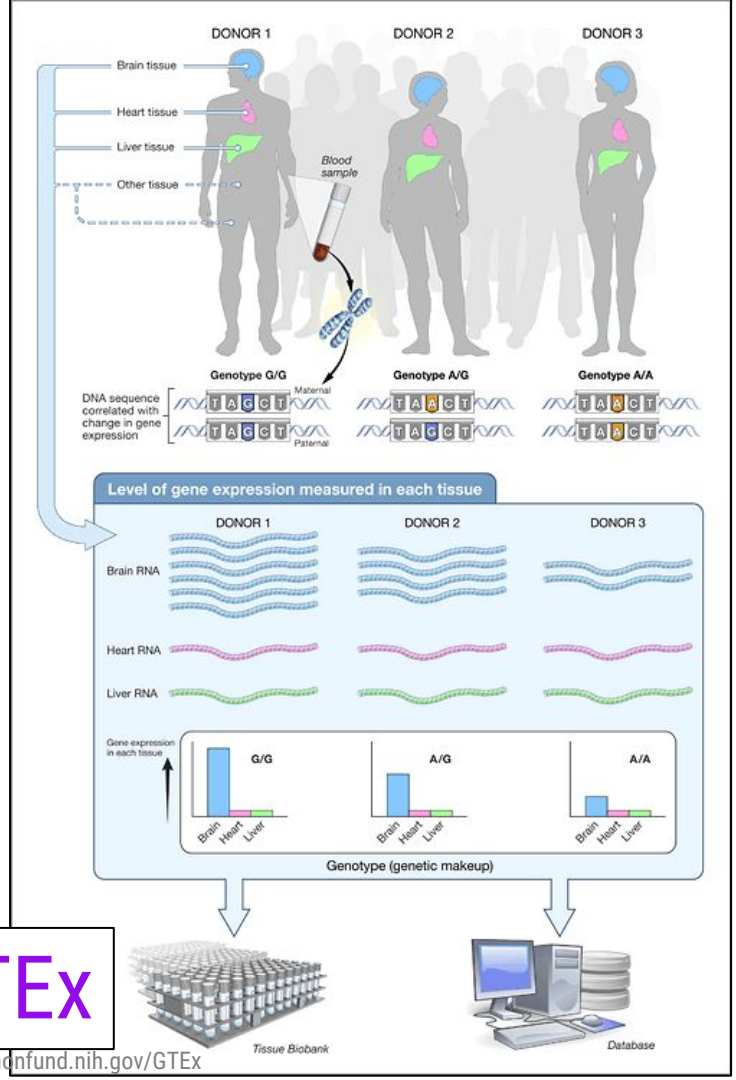
RNA-Seq = estimate expression across entire genome



expression ≅ # RNA-Seq reads

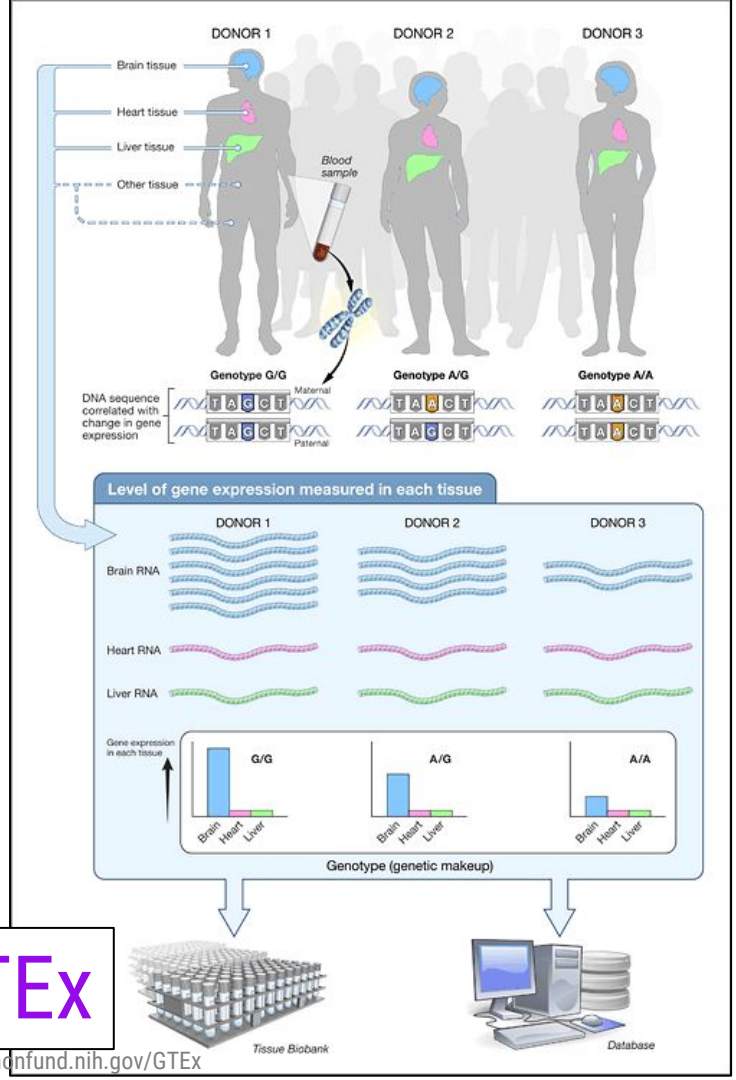
Scaling Up

slide adapted from jeff leek



GTEx

https://commonfund.nih.gov/GTEx



GTEx

https://commonfund.nih.gov/GTEx

TCGA

NCBI Resources How To Sign in to NCBI

SRA

SRA Search

Advanced Help

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Getting Started

- Understanding and Using SRA
- How to Submit
- Login to Submit
- Download Guide

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- dbGaP Home
- Trace Archive Home
- BioSample
- GenBank Home

SRA

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA) Write to the Help Desk

Project	No. of Sample
GTEx Genotype-Tissue Expression Project	9,962
TCGA The Cancer Genome Atlas	11,284
SRA Sequence Read Archive	49,848

https://jhubiostatistics.shinyapps.io/recount/

recount2: analysis-ready RNA-seq gene and exon counts datasets

Datasets | Popular datasets | GTEx | TCGA | Documentation | Download data with R | Accessing recount2 via SciServer | Contribute your data

A multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets

recount2 is an online resource consisting of RNA-seq gene and exon counts as well as coverage bigWig files for 2041 different studies. It is the second generation of the ReCount project. The raw sequencing data were processed with Rail-RNA as described at bioRxiv 038224 which created the coverage bigWig files. For ease of statistical analysis, for each study we created count tables at the gene and exon levels and extracted phenotype data, which we provide in their raw formats as well as in RangedSummarizedExperiment R objects (described in the SummarizedExperiment Bioconductor package). We also computed the mean coverage per study and provide it in a bigWig file, which can be used with the derfinder Bioconductor package to perform annotation-agnostic differential expression analysis at the expressed regions-level as described at bioRxiv 015370. The count tables, RangedSummarizeExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and file information tables are ready to use and freely available here. We also created the recount Bioconductor package which allows you to search and download the data for a specific study . By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, we make finding and analyzing RNA-seq data considerably more straightforward.

Related publications

Collado-Torres L, **Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. recount: A large-scale resource of analysis-ready RNA-seq expression data. *bioRxiv* **068478**.

The Datasets

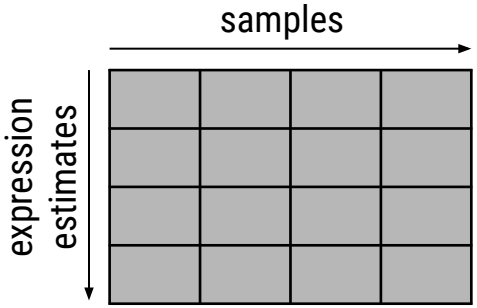
Show 10 entries

Search:

accession	number of samples	species	abstract	gene	exon	junctions	phenotype	files info
All	All	All	All			All	All	
SRP025982	1720	human	We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the United States Food and Drug Administration. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, we assess RNA sequencing (RNA-seq) performance for sequence discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics. At all sequencing depths, we discover unannotated exon-exon junctions, with >80% validated by qPCR. We find that	RSE counts	RSE counts	RSE jx_bed jx_cov counts	link	link

recount2

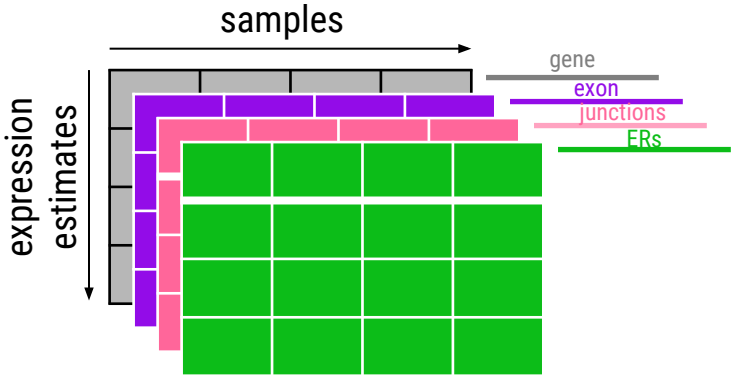
expression data for ~70,000 human samples



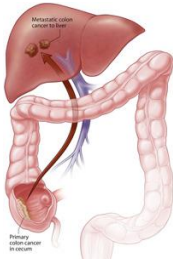
GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284

recount2

expression data for ~70,000 human samples



GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284



What makes primary cancer different than metastatic cancer?

Collect patient samples and information

Sequence samples

Process sequencing data

Determine expression differences between primary cancer and metastatic cancer

3~6 months — 1-2 wks — 2-4 wks — 3-6 months — 1-3 months — 1 month - 1+ years

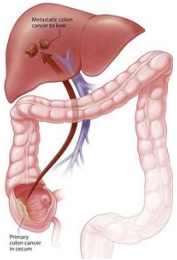
Find a researcher with access to patient samples

Extract DNA/R... from sampl...

Find publicly available **RNA-sequencing data** from primary cancer and metastasis

Data cleaning

Total: 1+ years



What makes primary cancer different than metastatic cancer?

Get already processed and summarized RNA-Seq data from *recount2*

Collect patient samples and information

Sequence samples

Process sequencing data

Determine expression differences between primary cancer and metastatic cancer

3~6 months | 1-2 wks | 2-4 wks | 3-6 months | 1-3 months | 1 month - 1+ years

Find a researcher with access to patient samples

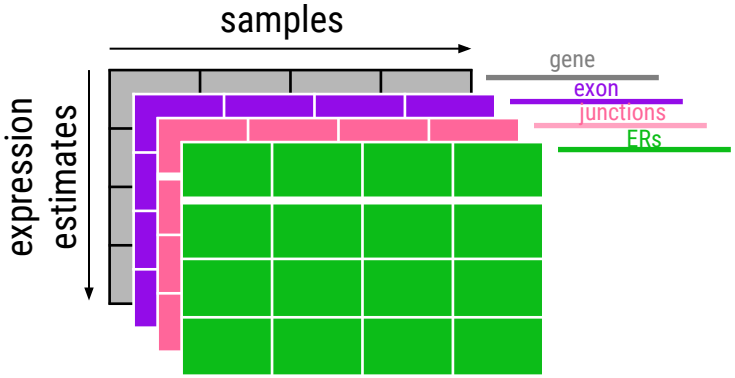
Extract DNA/RNA from samples

Data cleaning

Total: months

recount2

expression data for ~70,000 human samples

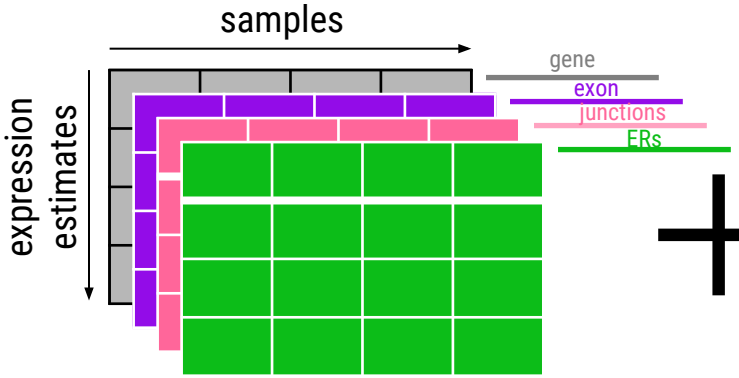


GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284

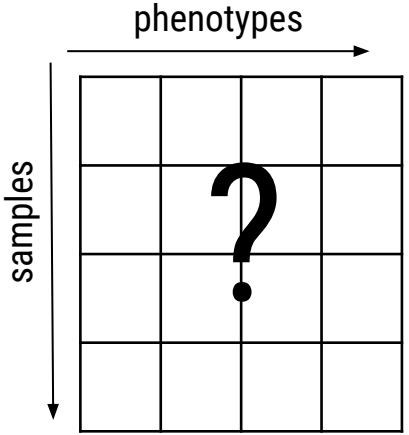
→ Answer meaningful questions about human biology and expression

recount2

expression data for ~70,000 human samples



GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284



Answer meaningful questions about human biology and expression

in-silico Phenotyping

slide adapted from jeff leek

SRA phenotype information is far from complete

	Sex	Tissue	Race	Age
6620	female	liver	NA	NA
6621	female	liver	NA	NA
6622	female	liver	NA	NA
6623	female	liver	NA	NA
6624	female	liver	NA	NA
6625	male	liver	NA	NA
6626	male	liver	NA	NA
6627	male	liver	NA	NA
6628	male	liver	NA	NA
6629	male	liver	NA	NA
6630	male	liver	NA	NA
6631	NA	blood	NA	NA
6632	NA	blood	NA	NA
6633	NA	blood	NA	NA
6634	NA	blood	NA	NA
6635	NA	blood	NA	NA
6636	NA	blood	NA	NA

SRA phenotype information is far from complete

	Sex	Tissue	Race	Age
6620	female	liver	NA	NA
6621	female	liver	NA	NA
6622	female	liver	NA	NA
6623	female	liver	NA	NA
6624	female	liver	NA	NA
6625	male	liver	NA	NA
6626	male	liver	NA	NA
6627	male	liver	NA	NA
6628	male	liver	NA	NA
6629	male	liver	NA	NA
6630	male	liver	NA	NA
6631	NA	blood	NA	NA
6632	NA	blood	NA	NA
6633	NA	blood	NA	NA
6634	NA	blood	NA	NA
6635	NA	blood	NA	NA
6636	NA	blood	NA	NA

Even when information *is* provided, it's not always clear…

Sex across the SRA:

Level	Frequency
F	95
female	2036
Female	51
M	77
male	1240
Male	141
Total	**3640**

Even when information *is* provided, it's not always clear…

Sex across the SRA:

Level	Frequency
F	95
female	2036
Female	51
M	77
male	1240
Male	141
Total	**3640**

"1 Male, 2 Female", "2 Male, 1 Female", "3 Female", "DK", "male and female" "Male (note: ….)", "missing", "mixed", "mixture", "N/A", "Not available", "not applicable", "not collected", "not determined", "pooled male and female", "U", "unknown", "Unknown"

Even when information *is* provided, it's not always clear…

Sex across the SRA:

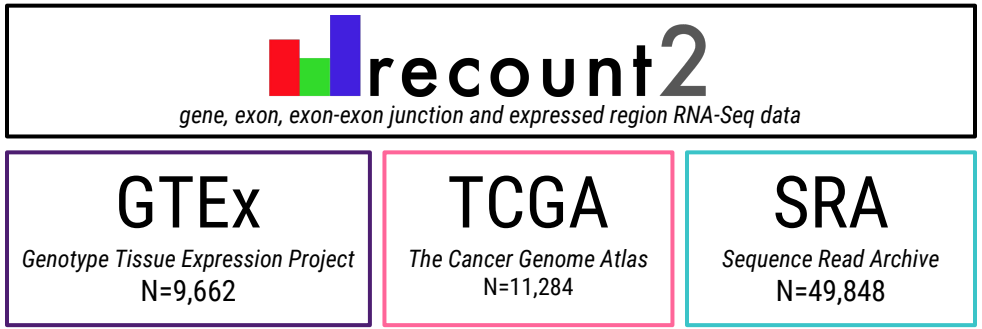
Level	Frequency
F	95
female	2036
Female	51
M	77
male	1240
Male	141
Total	**3640**

"1 Male, 2 Female", "2 Male, 1 Female", "3 Female", "DK", "male and female" "Male (note: ….)", "missing", "mixed", "mixture", "N/A", "Not available", "not applicable", "not collected", "not determined", "pooled male and female", "U", "unknown", "Unknown"

# of NAs	# w/sex assigned
44,957	4,700

Goal :

to accurately predict critical phenotype information for all samples in *recount2*



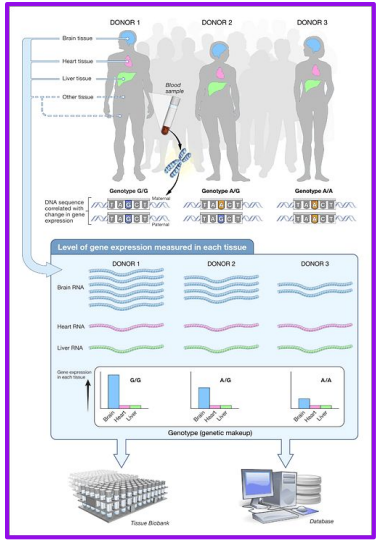
Goal :

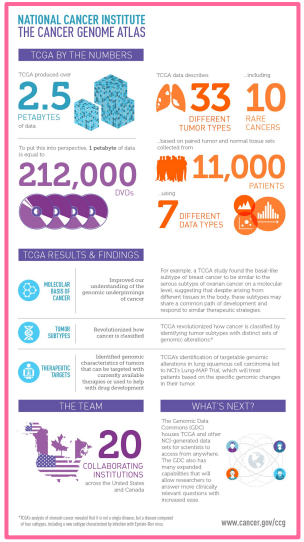
to accurately predict critical phenotype information for all samples in *recount2*

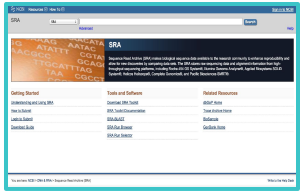
recount2

gene, exon, exon-exon junction and expressed region RNA-Seq data

GTEx	TCGA	SRA
Genotype Tissue Expression Project	*The Cancer Genome Atlas*	*Sequence Read Archive*
N=9,662	N=11,284	N=49,848

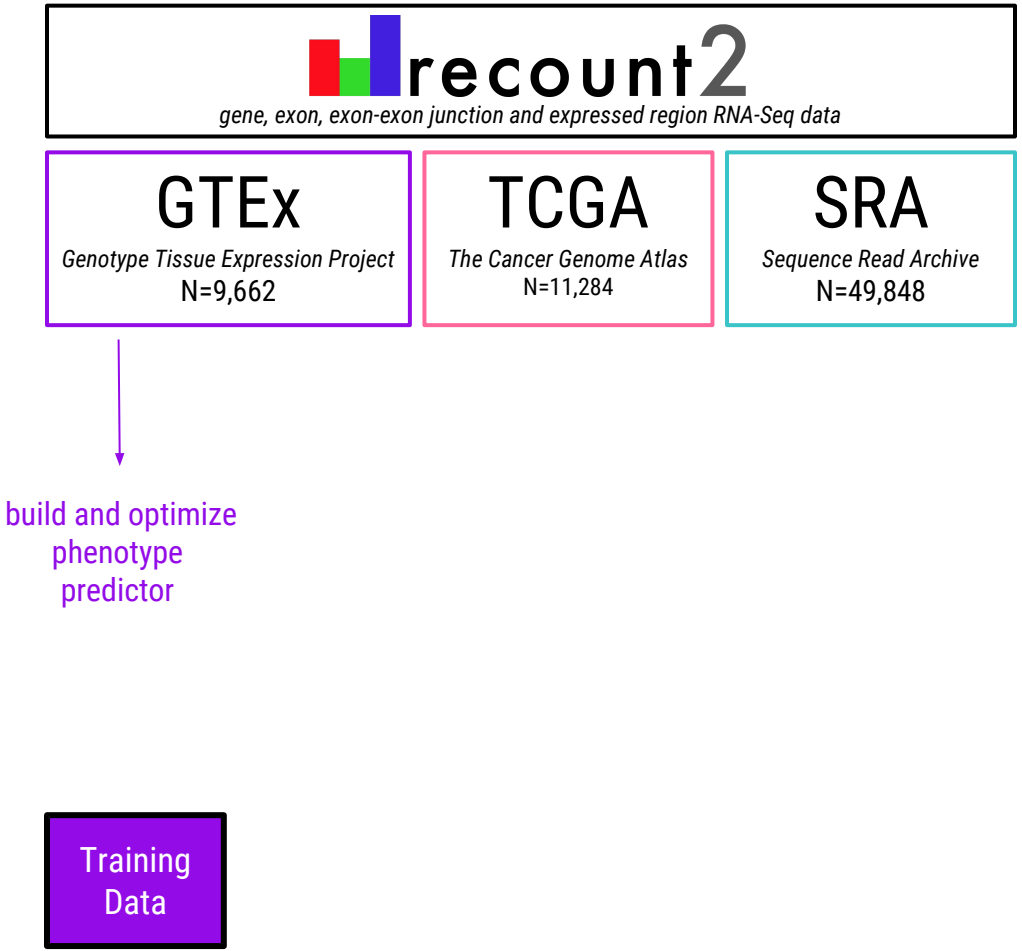






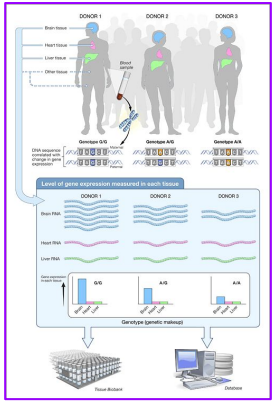
Goal :

to accurately predict critical phenotype information for all samples in *recount2*



Missingness limited in GTEx phenotype data

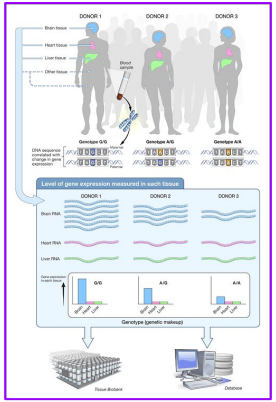
	Sex	Tissue	Race	Age
1	male	Lung	White	59
2	male	Brain	White	27
3	female	Heart	Black or African American	23
4	male	Brain	White	51
5	male	Skin	White	27
6	male	Lung	White	68
7	female	Brain	White	61
8	female	Adipose Tissue	White	42
9	male	Brain	White	40
10	female	Uterus	White	33
11	female	Nerve	White	60
12	male	Muscle	White	54
13	female	Ovary	White	31
14	male	Blood	White	53
15	female	Brain	White	56
16	male	Muscle	White	44



GTEx

Missingness limited in GTEx phenotype data

	Sex	Tissue	Race	Age
1	male	Lung	White	59
2	male	Brain	White	27
3	female	Heart	Black or African American	23
4	male	Brain	White	51
5	male	Skin	White	27
6	male	Lung	White	68
7	female	Brain	White	61
8	female	Adipose Tissue	White	42
9	male	Brain	White	40
10	female	Uterus	White	33
11	female	Nerve	White	60
12	male	Muscle	White	54
13	female	Ovary	White	31
14	male	Blood	White	53
15	female	Brain	White	56
16	male	Muscle	White	44



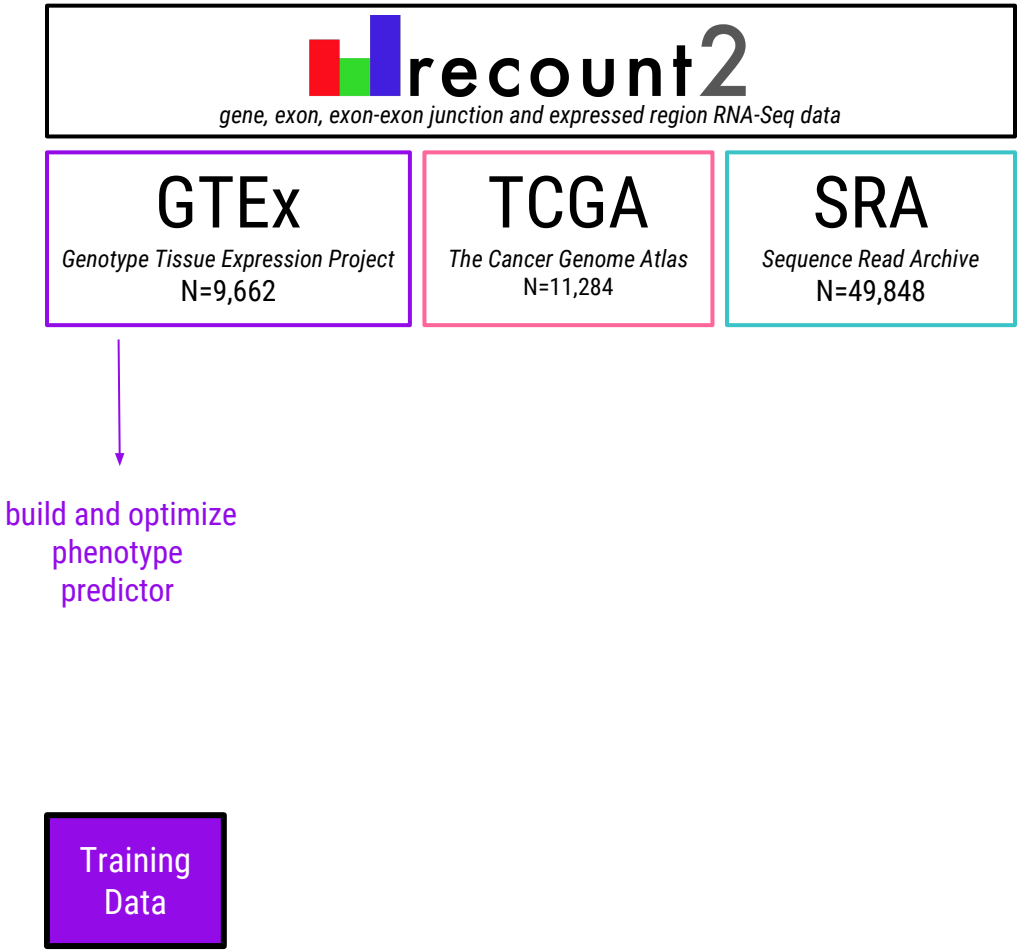
GTEx

Sex across GTEx:

level	Frequency
female	3,626
male	6,036
NA	0

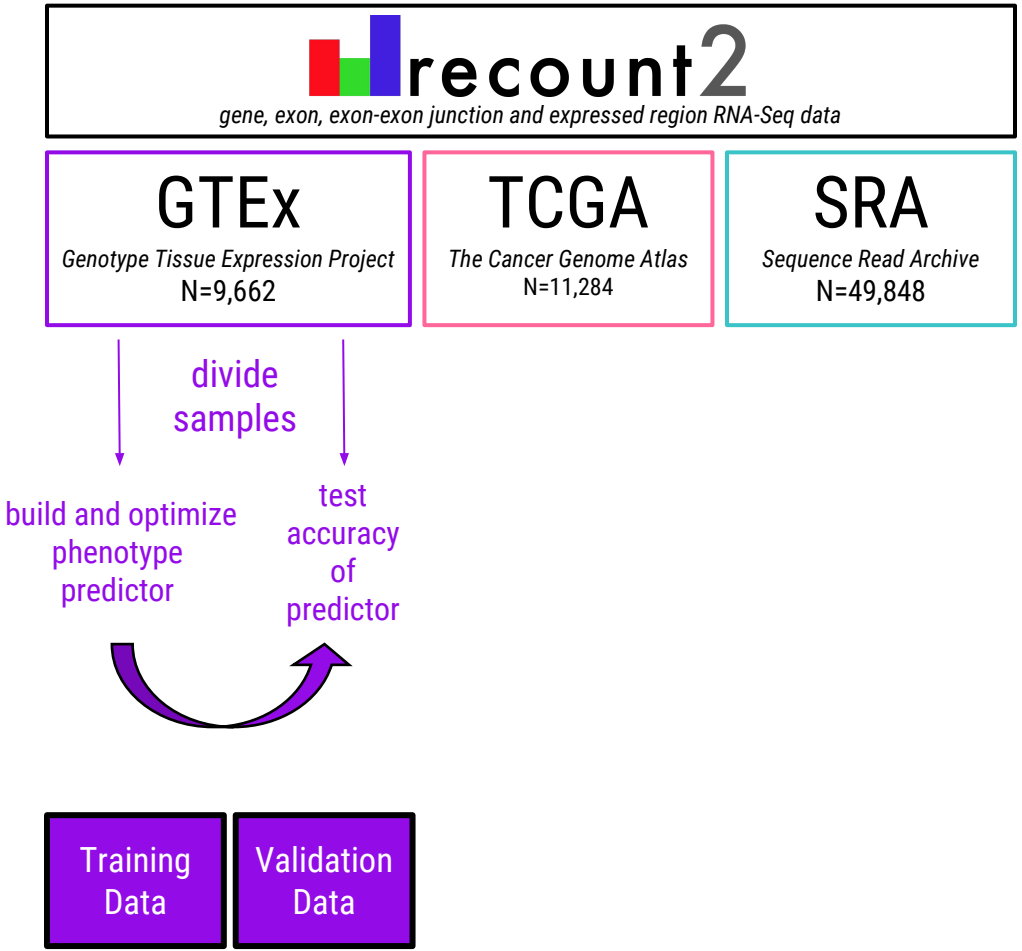
Goal :

to accurately predict critical phenotype information for all samples in *recount2*



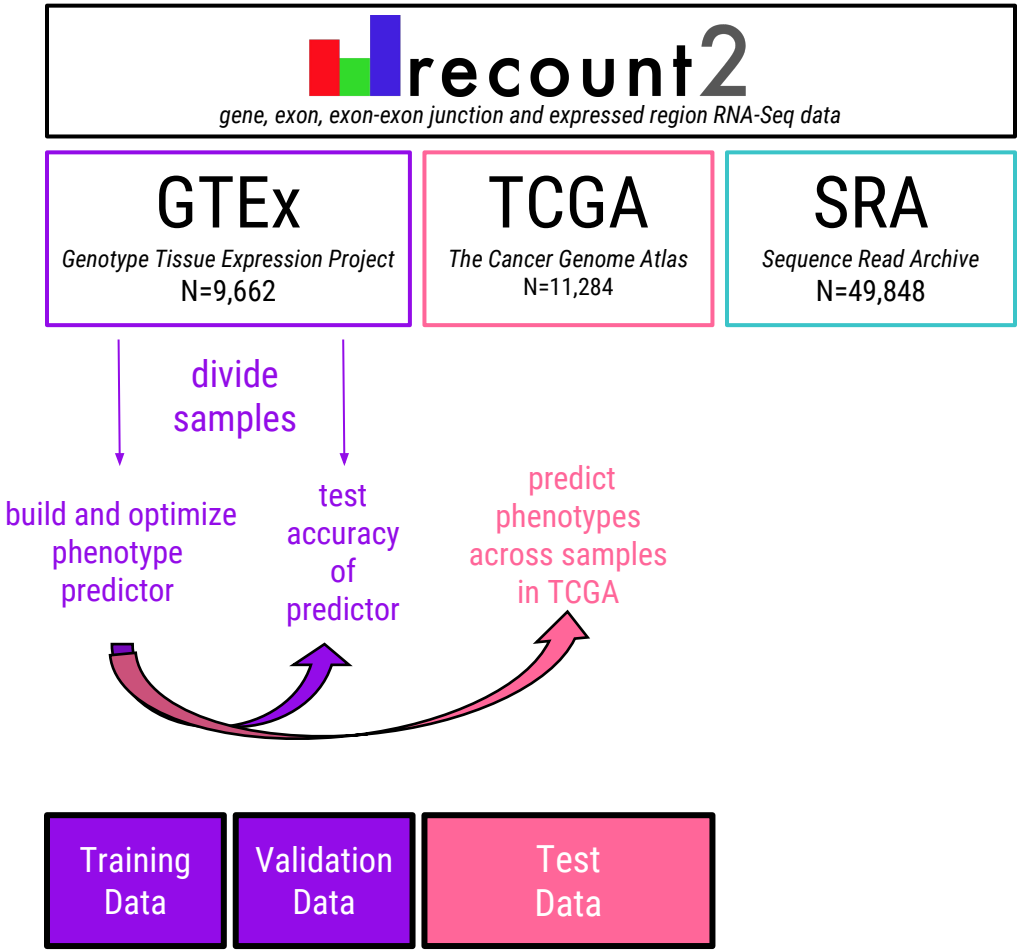
Goal :

to accurately predict critical phenotype information for all samples in *recount2*



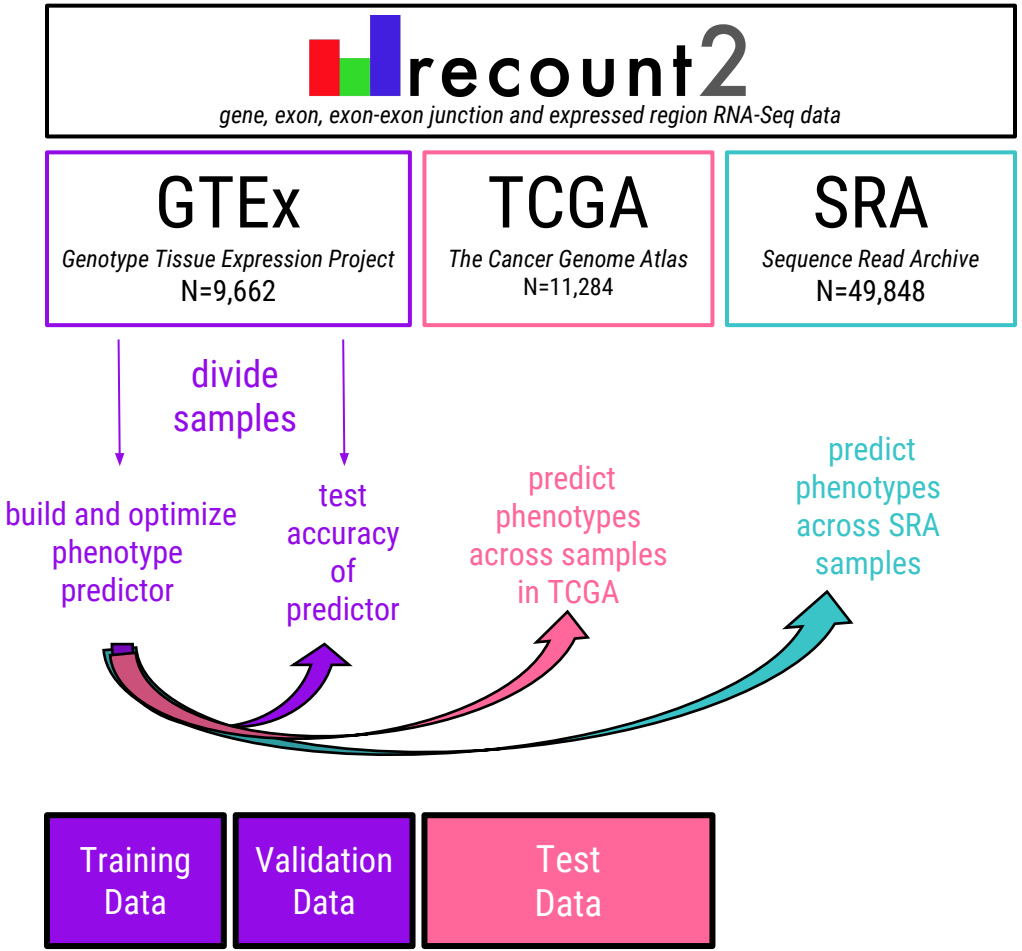
Goal :

to accurately predict critical phenotype information for all samples in *recount2*



Goal :

to accurately predict critical phenotype information for all samples in *recount2*



phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r N

regions x individuals

functions

filter_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()

filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype

βr phenotype (P_i)

male female

filtered regions (r)

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions	reported
male	male
female	female
female	female
male	male

Prediction accuracy: 100%

extract_data()

Extract expression information at regions identified by filter_regions() in a new data set

new data set samples

expression @ filtered regions

predict_pheno()

Predict phenotypes across samples in this new data set

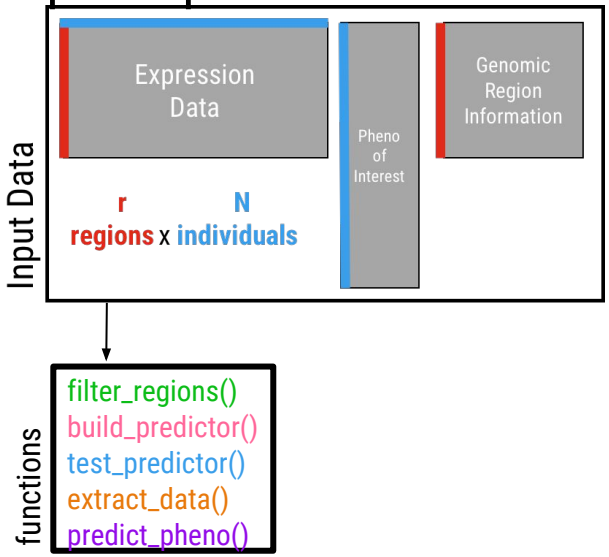
apply coefficient estimates to the extracted data

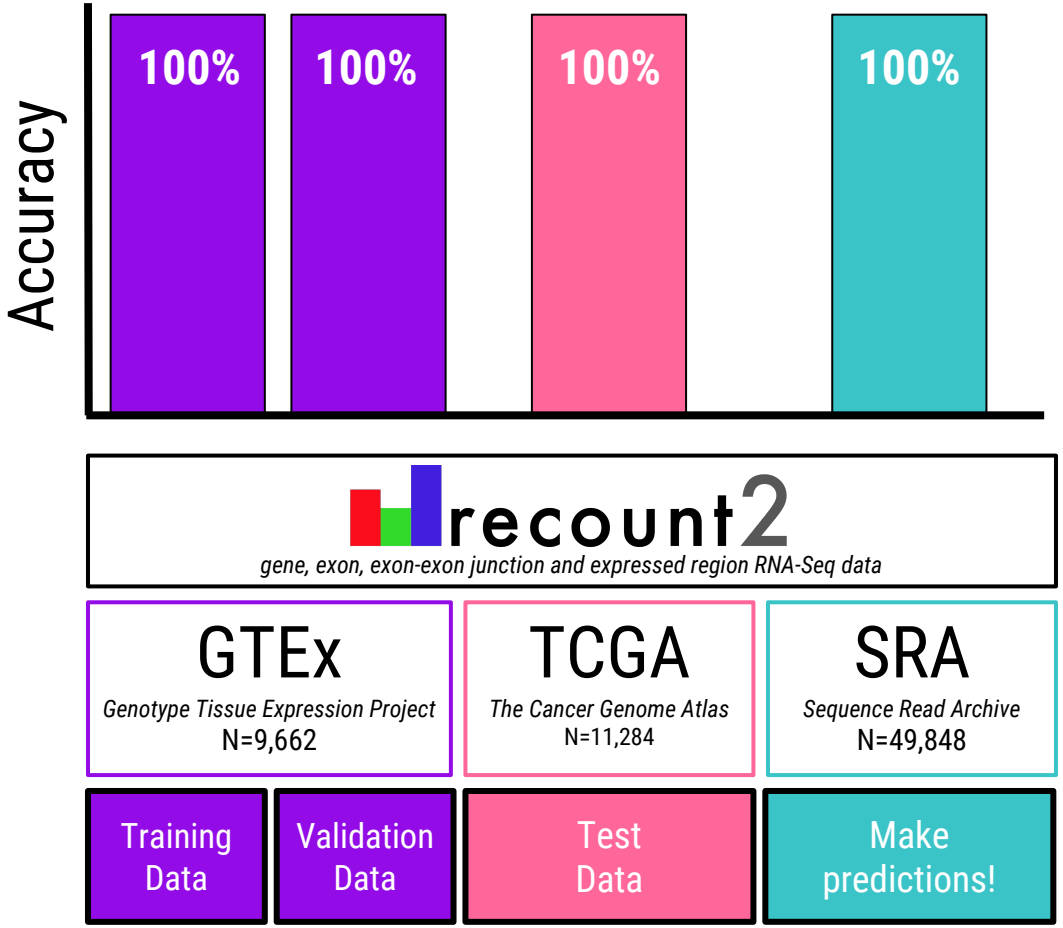
predictions in new data set

male
male
male
female

$$E[\{\mathbf{E}_r^*\}_{r\in\mathcal{S}} \mid \hat{\beta}_r] = \hat{\beta}_r\gamma^*$$

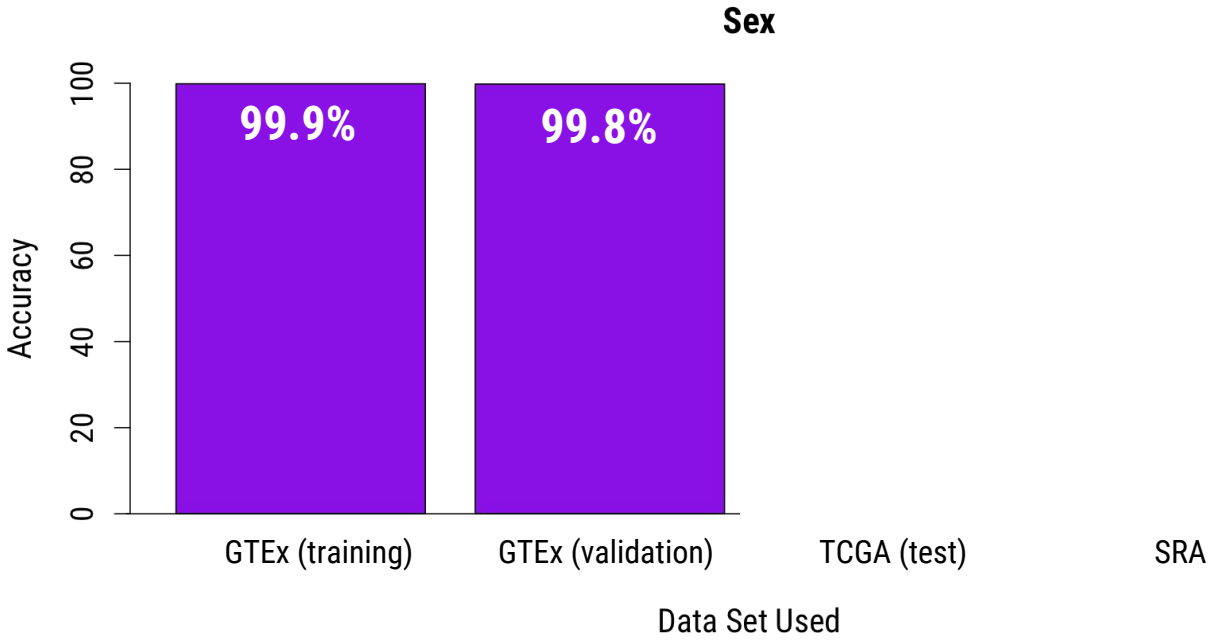
phenopredict





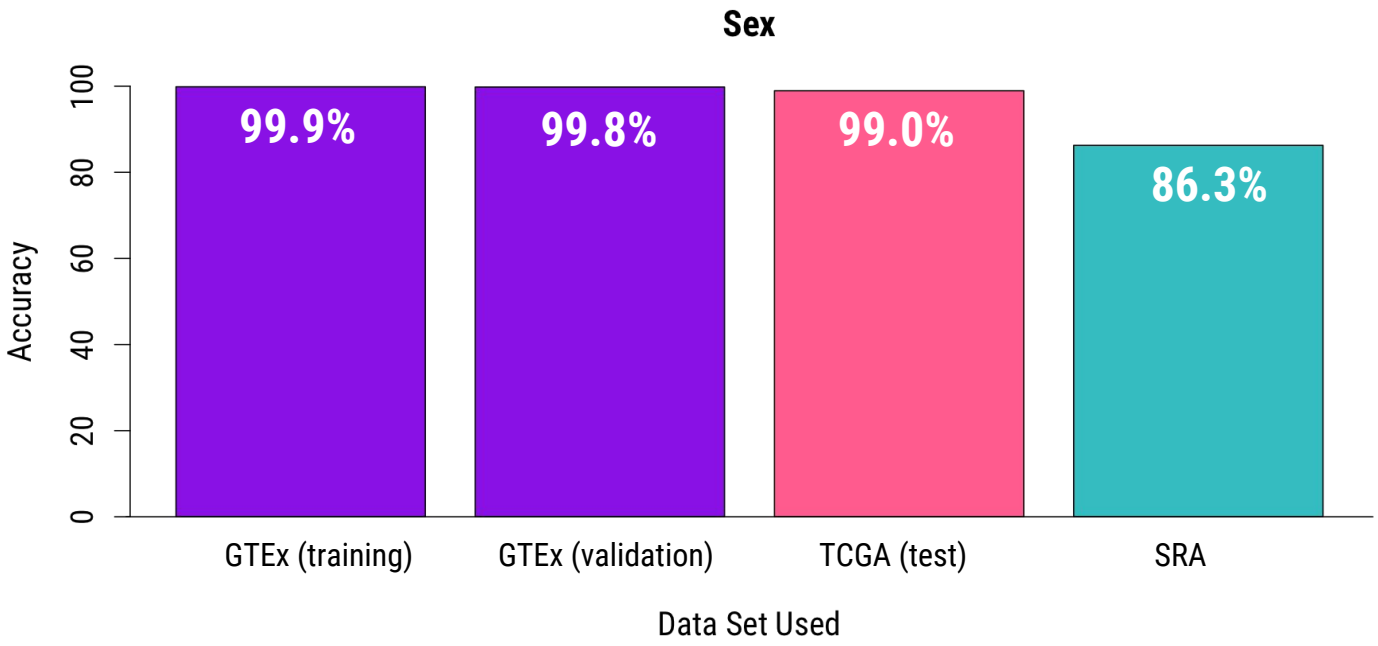
Let’s get predicting...

Sex prediction is accurate across data sets



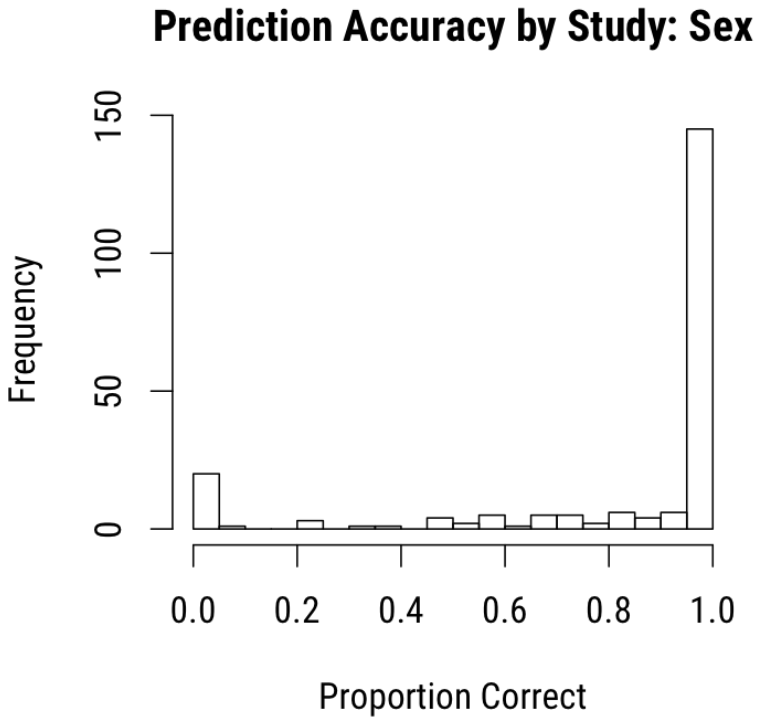
Number of Regions	**40**	**40**
Number of Samples (N)	**4,769**	**4,769**

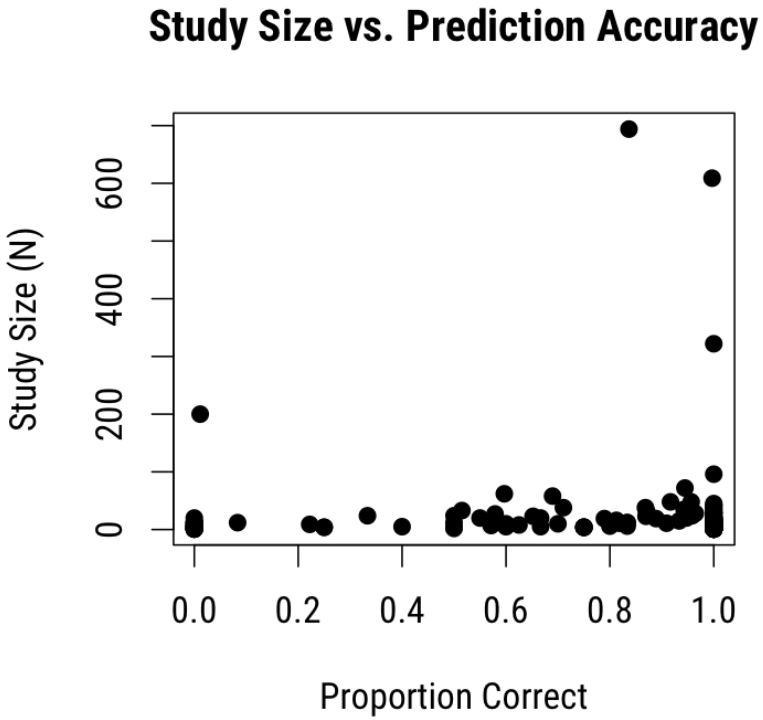
Sex prediction is accurate across data sets



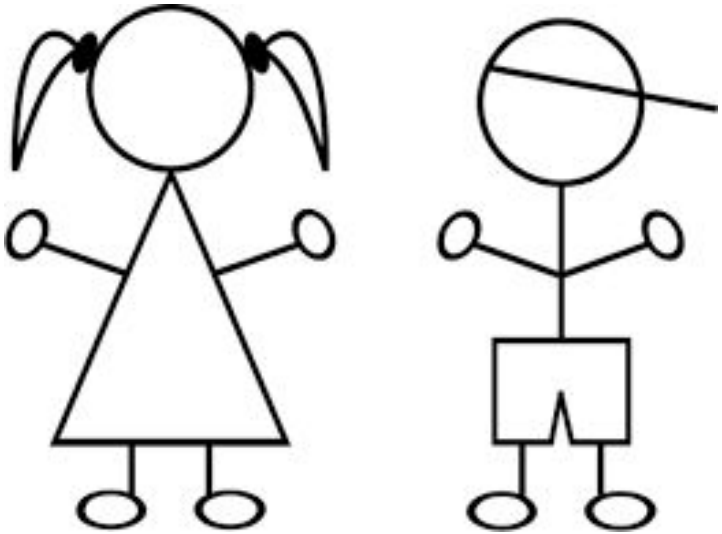
Number of Regions	**40**	**40**	**40**	**40**
Number of Samples (N)	**4,769**	**4,769**	**11,245**	**3,640**

Are a few studies driving decrease in accuracy across the SRA samples?



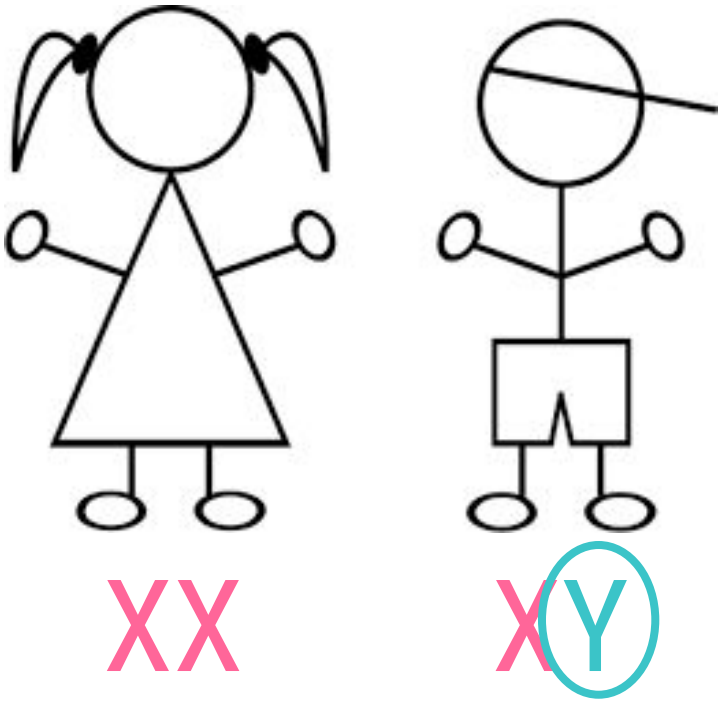


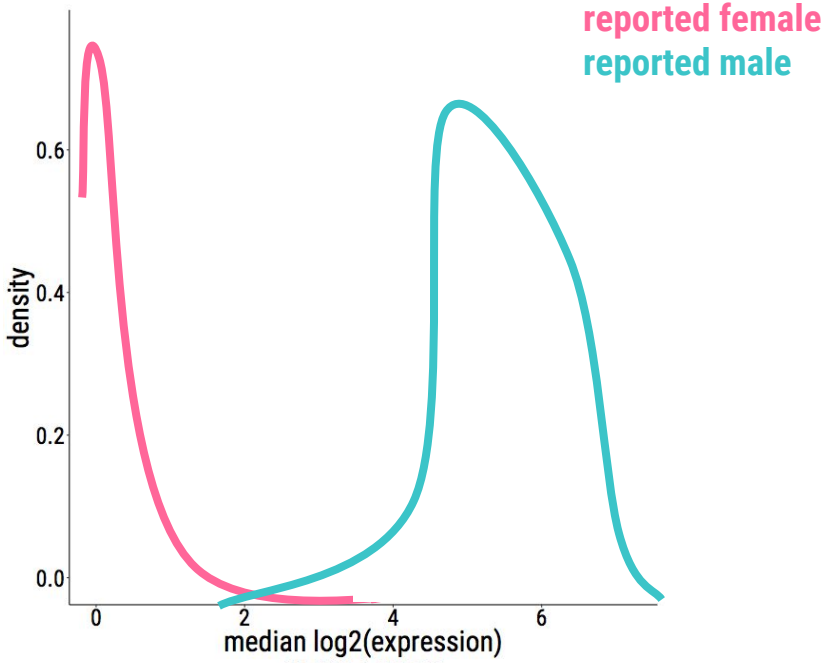
To assess misreporting of sex in the SRA, we can use Y-chromosome expression



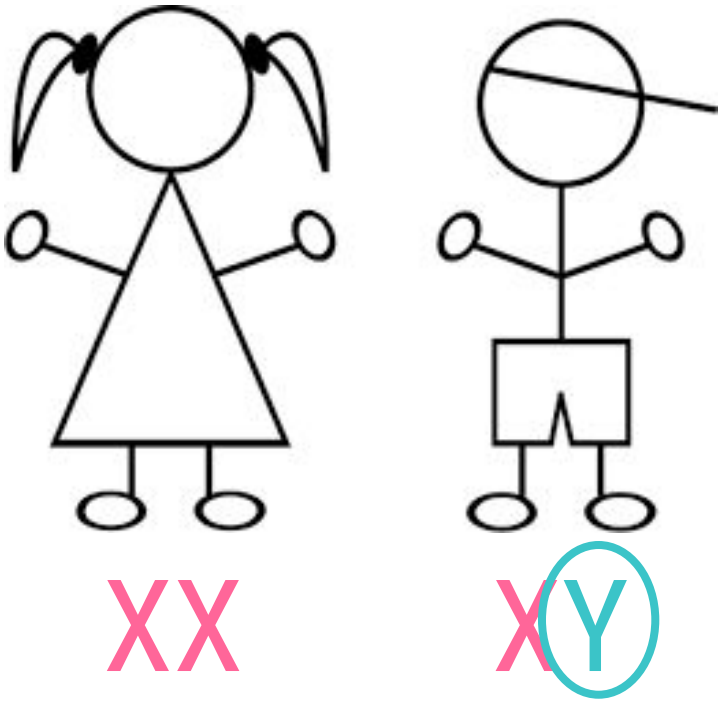
XX XY

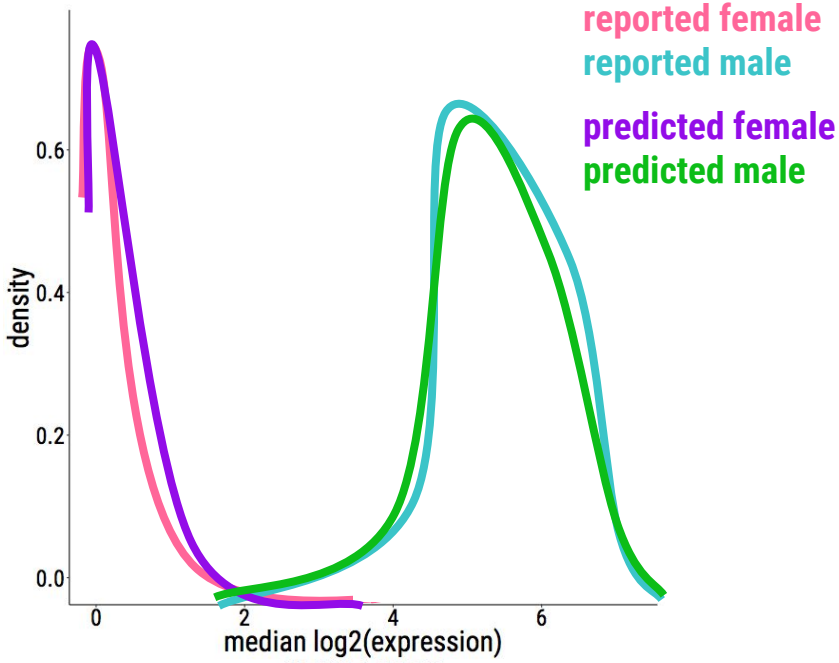
To assess misreporting of sex in the SRA, we can use Y-chromosome expression



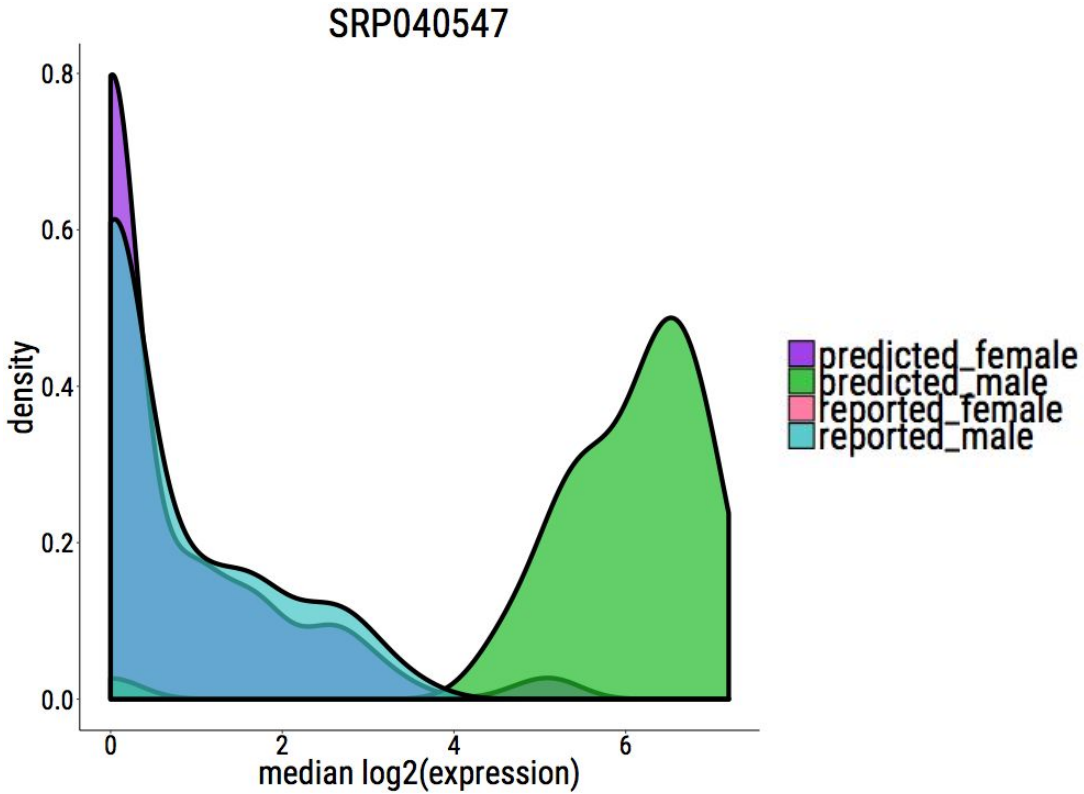


To assess misreporting of sex in the SRA, we can use Y-chromosome expression

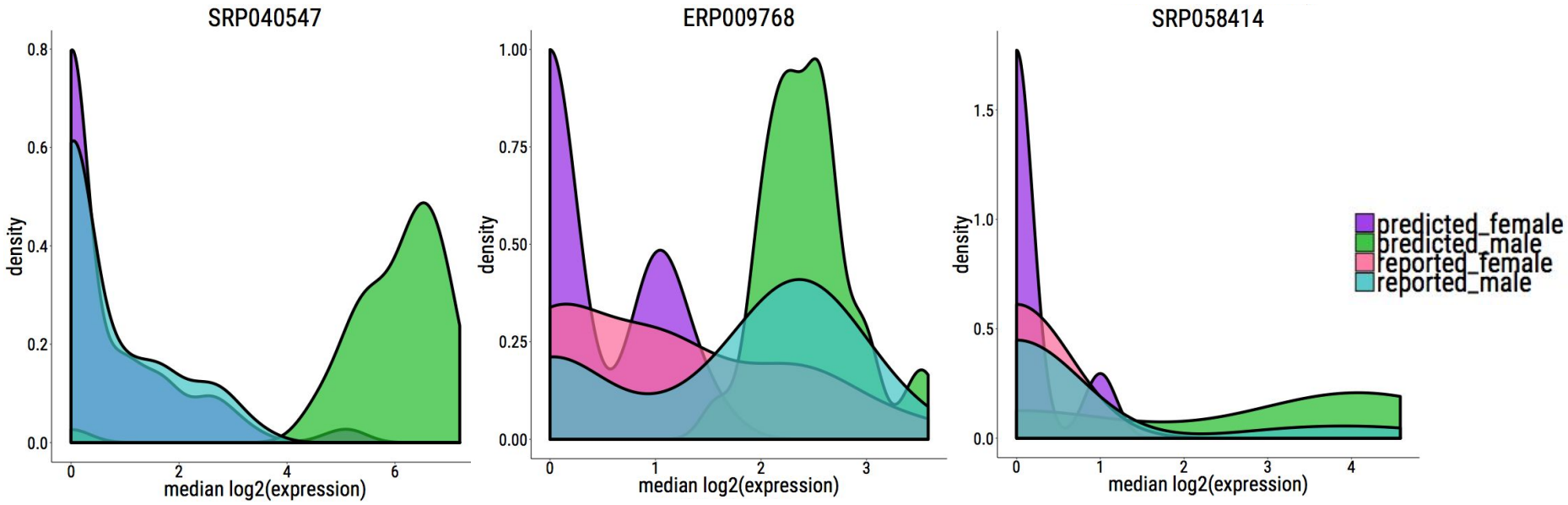




Expression from the Y chromosome suggests misreporting of sex in the SRA



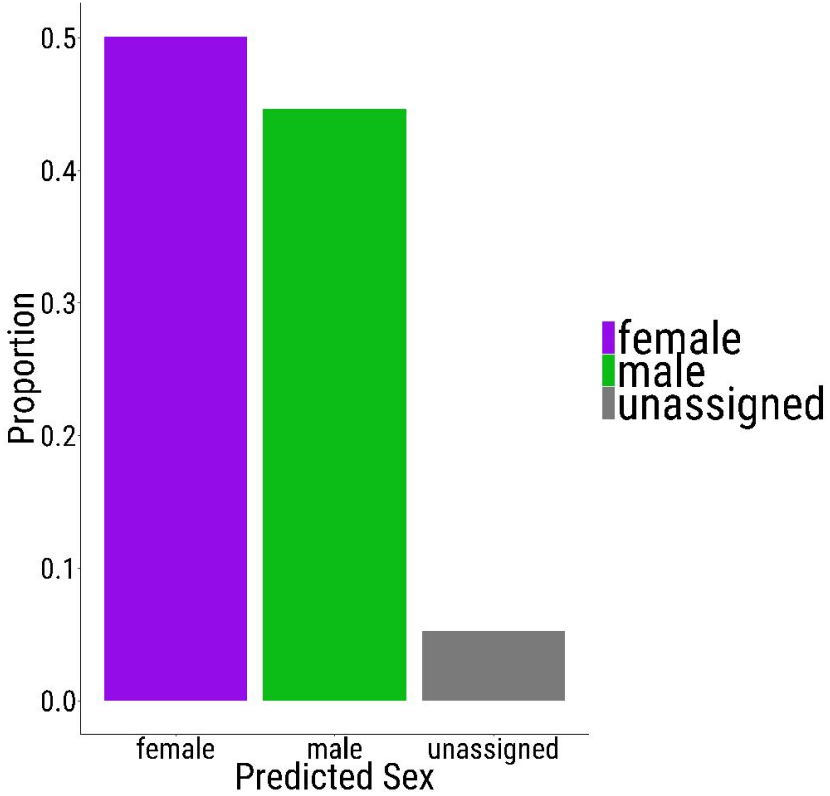
Expression from the Y chromosome suggests misreporting of sex in the SRA



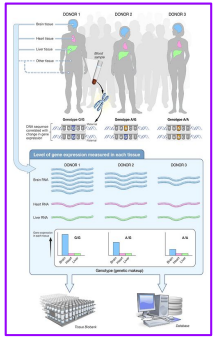
There’s a well-documented history of male sex-bias in biomedical research…

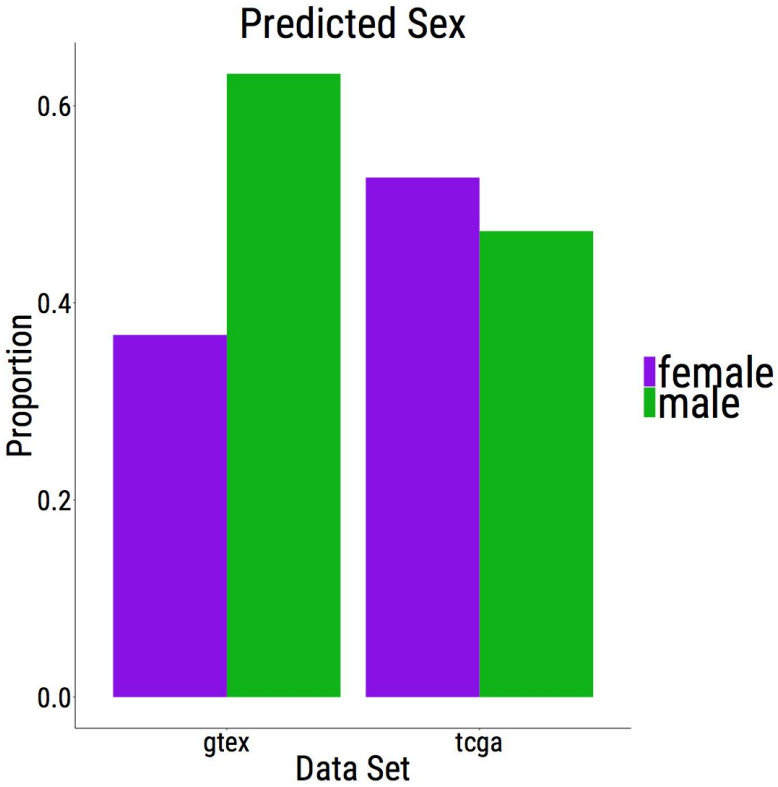
…so let’s take a closer look within *recount2*

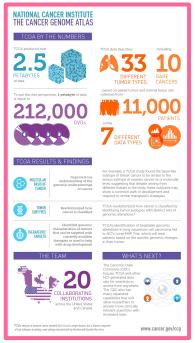
Across the ~70,000 samples in *recount2*, there are more samples predicted to be female than male.



GTEx is male-biased ; TCGA is female-biased



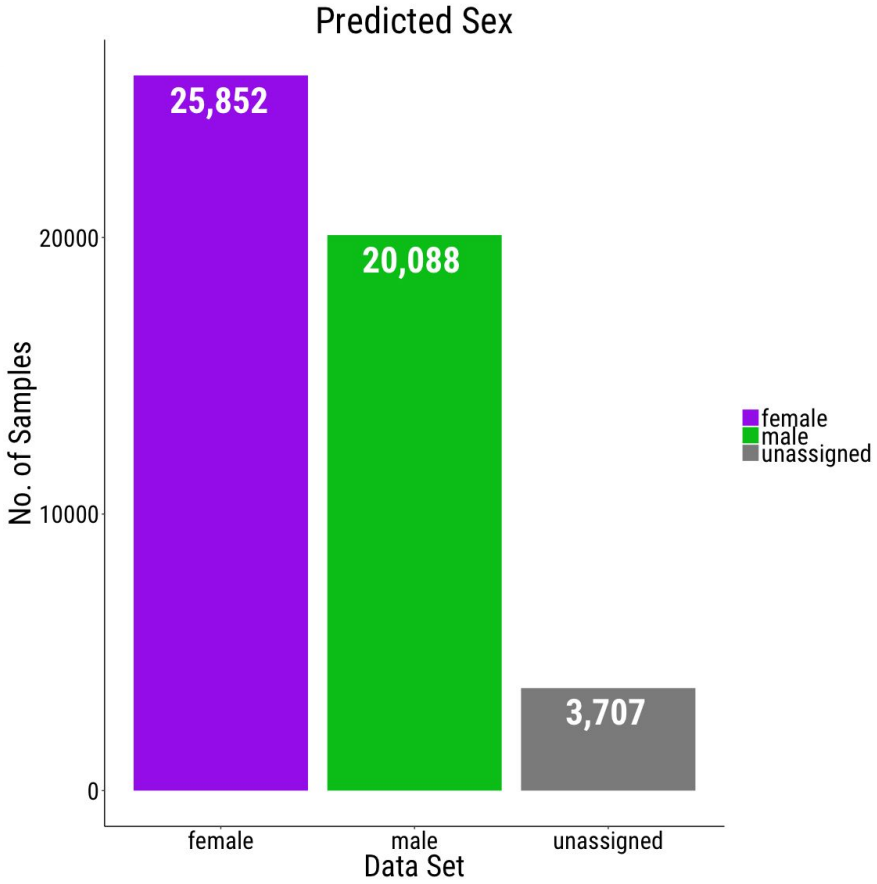




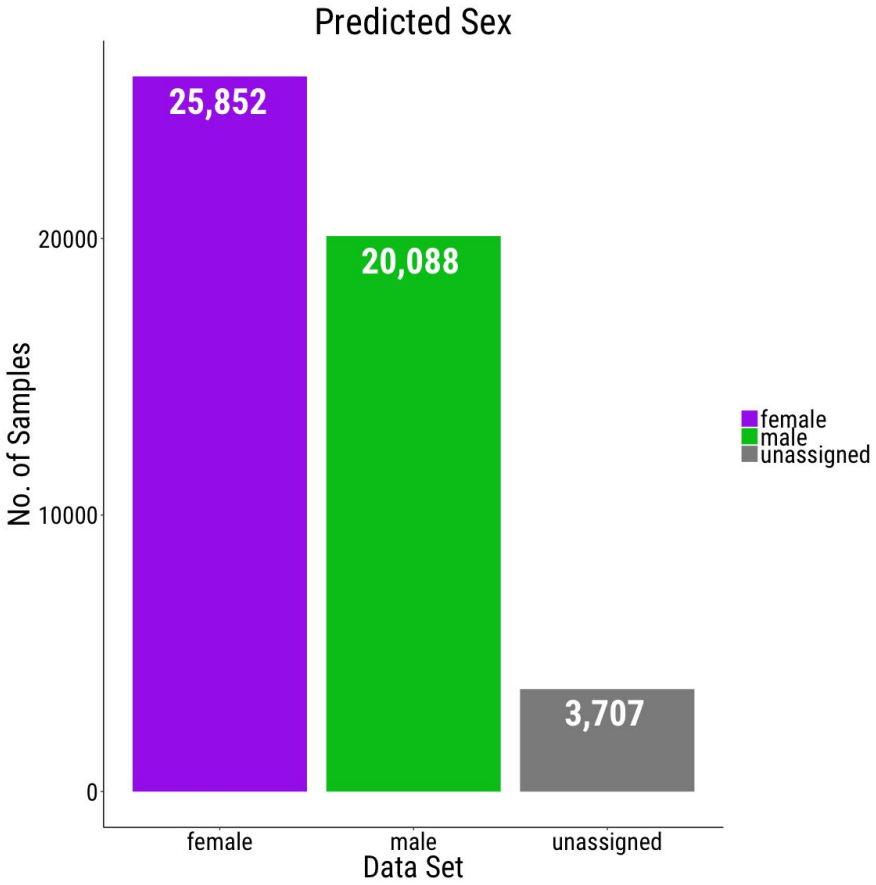
…but that has been previously reported by the consortia

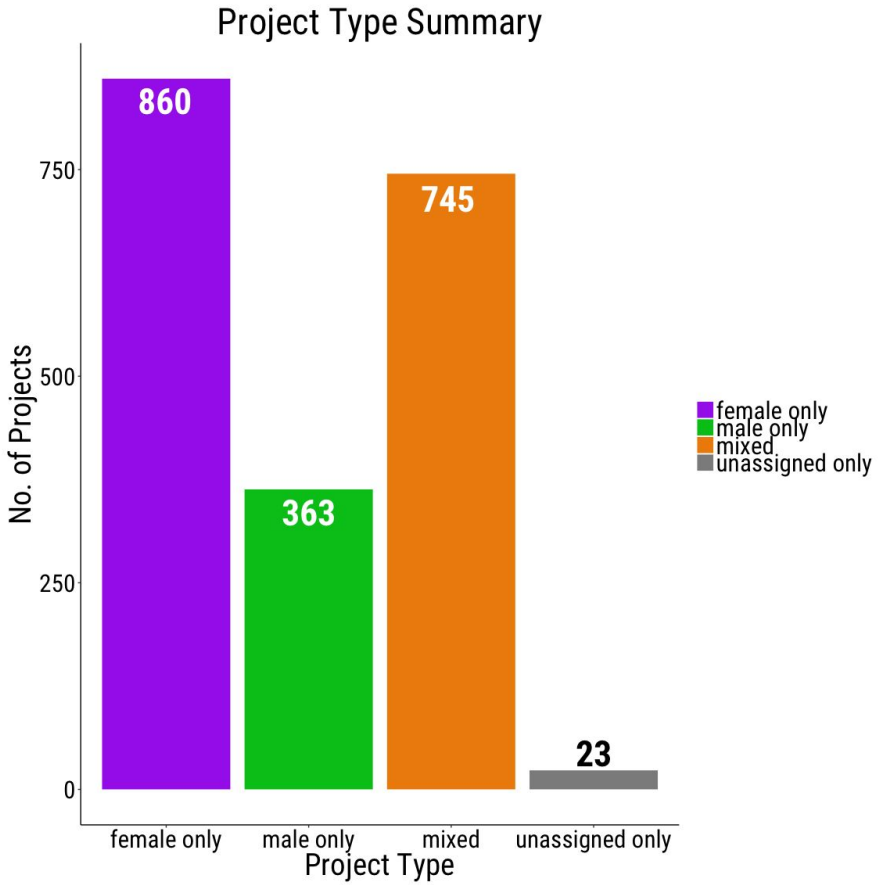
Is there a sex bias across the SRA?

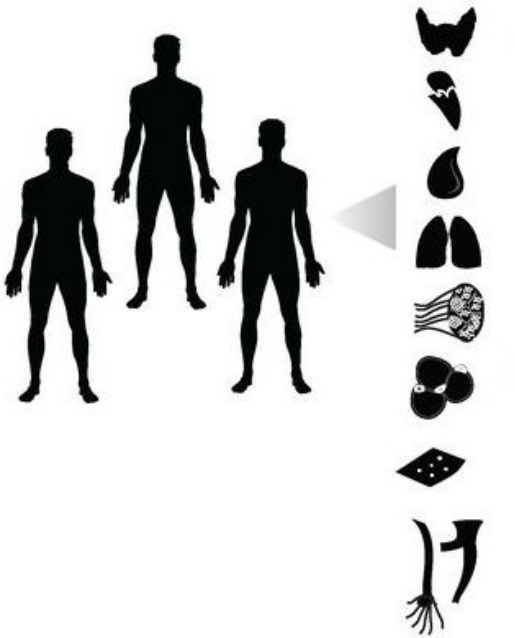
No evidence for sex bias in samples across the SRA



There are many female-only, male-only, and male-female projects available in *recount*



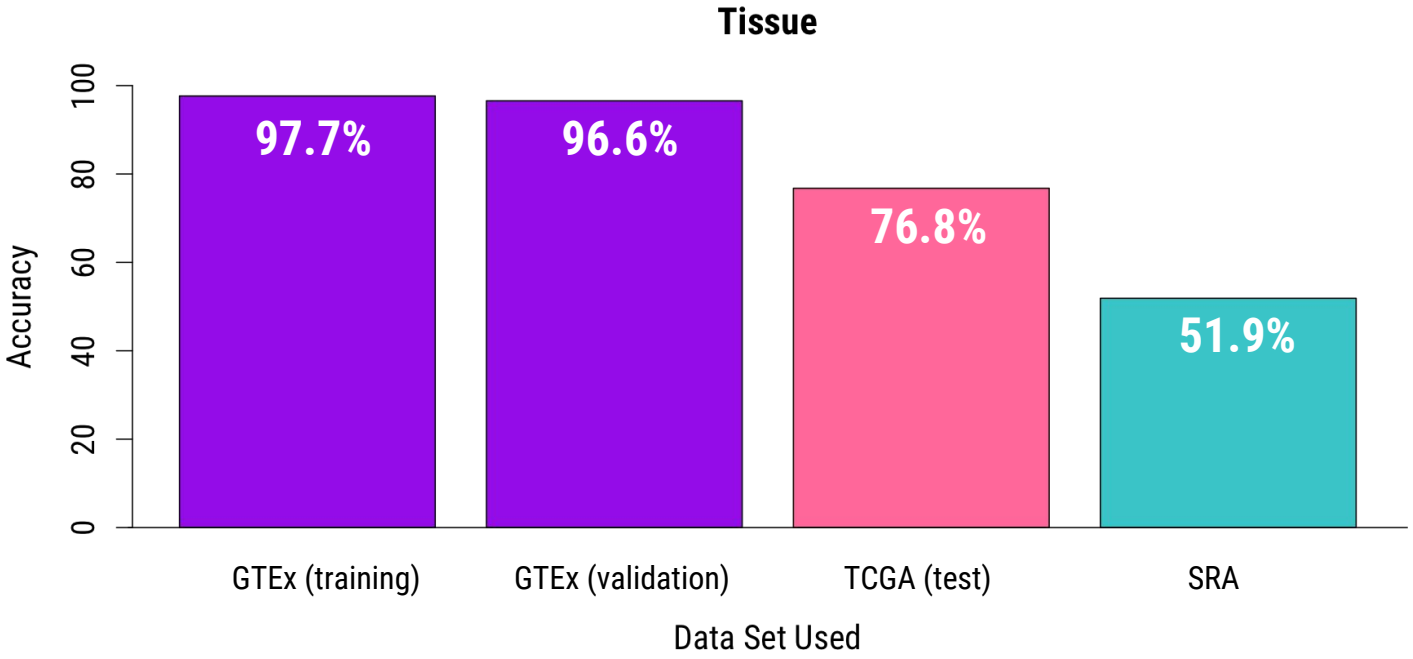




Can we use expression data to predict tissue?

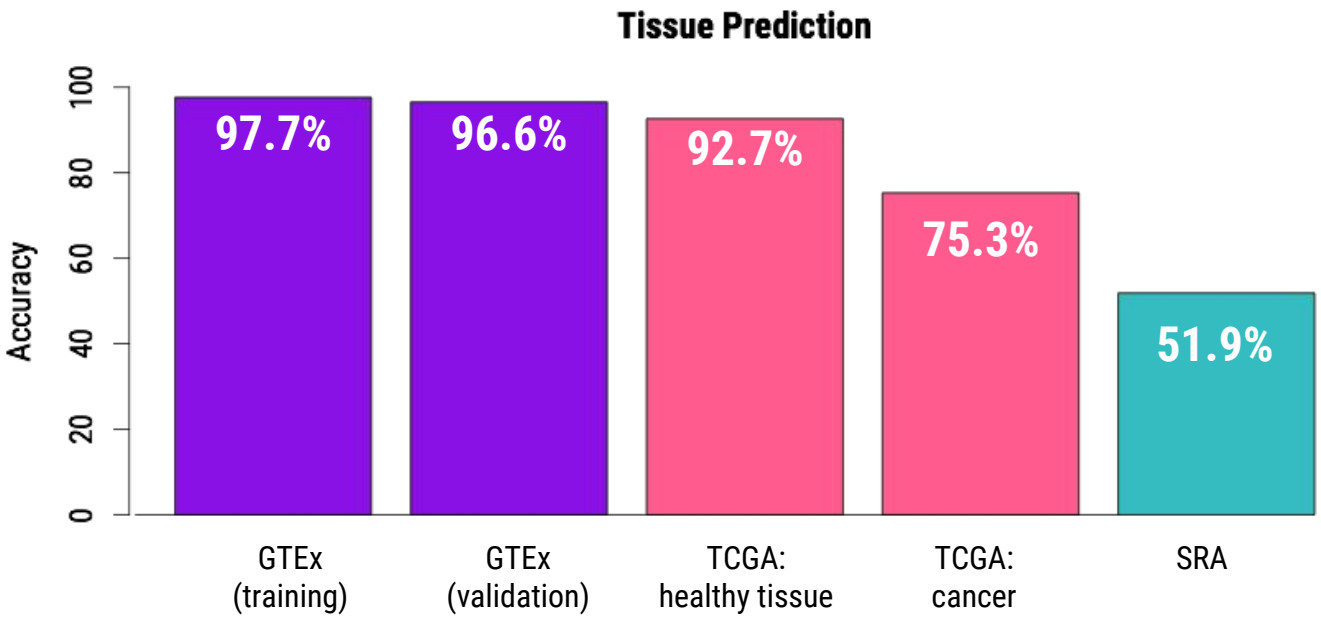
http://www.rna-seqblog.com/

Tissue prediction is accurate across data sets



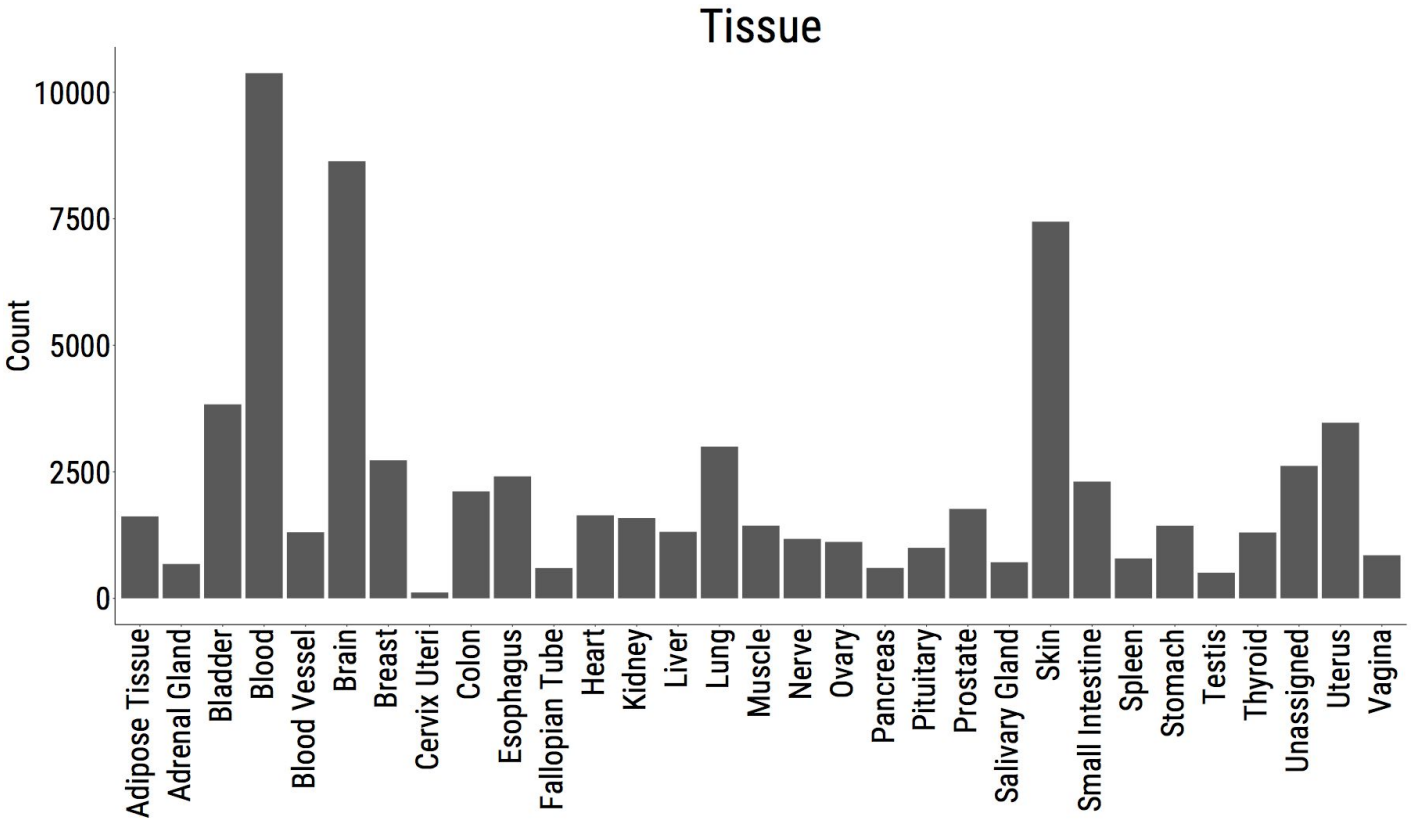
Number of Regions	2,281	2,281	2,281	2,281
Number of Samples (N)	4,769	4,769	7,317	8,951

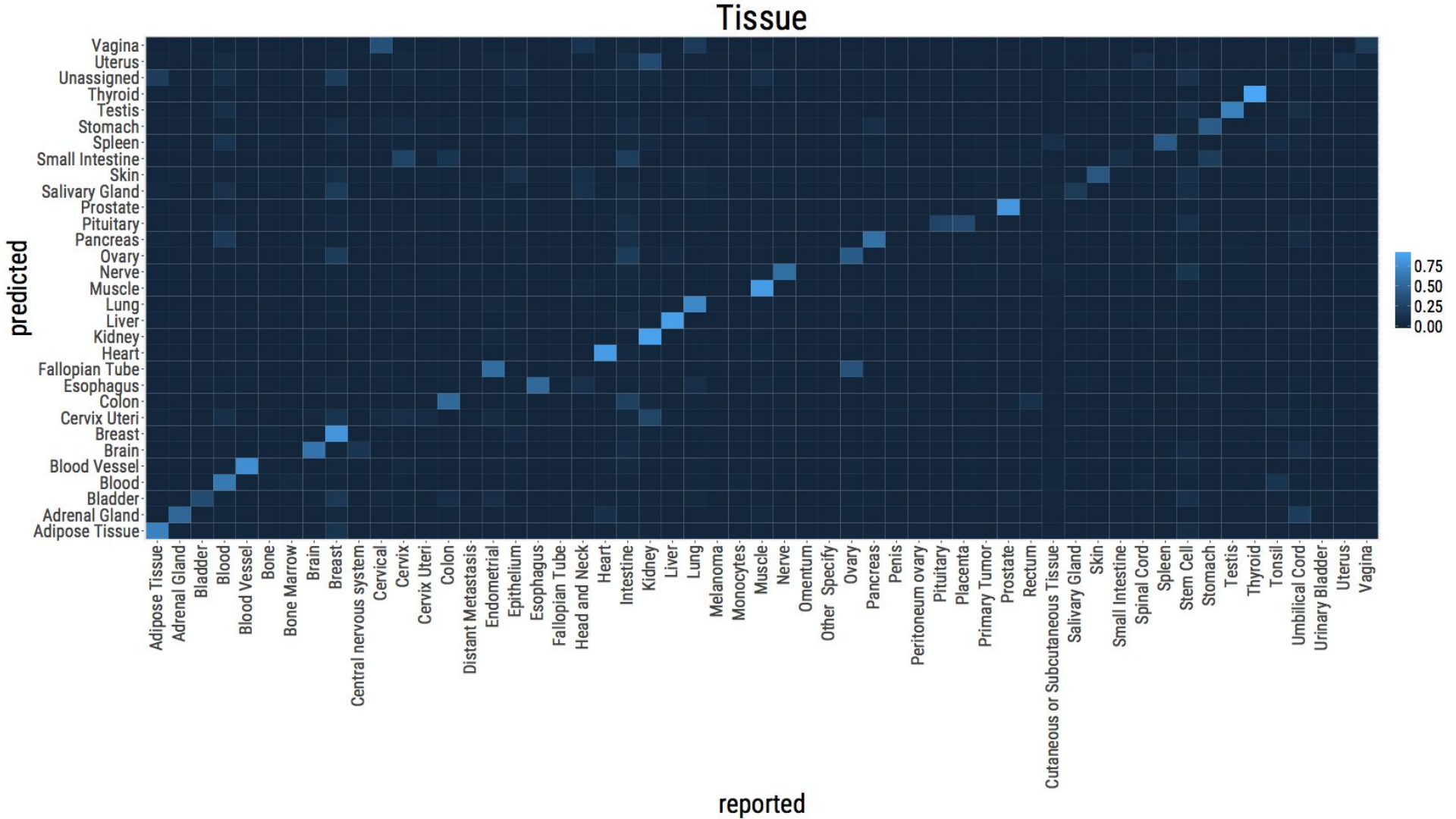
Prediction is more accurate in healthy tissue

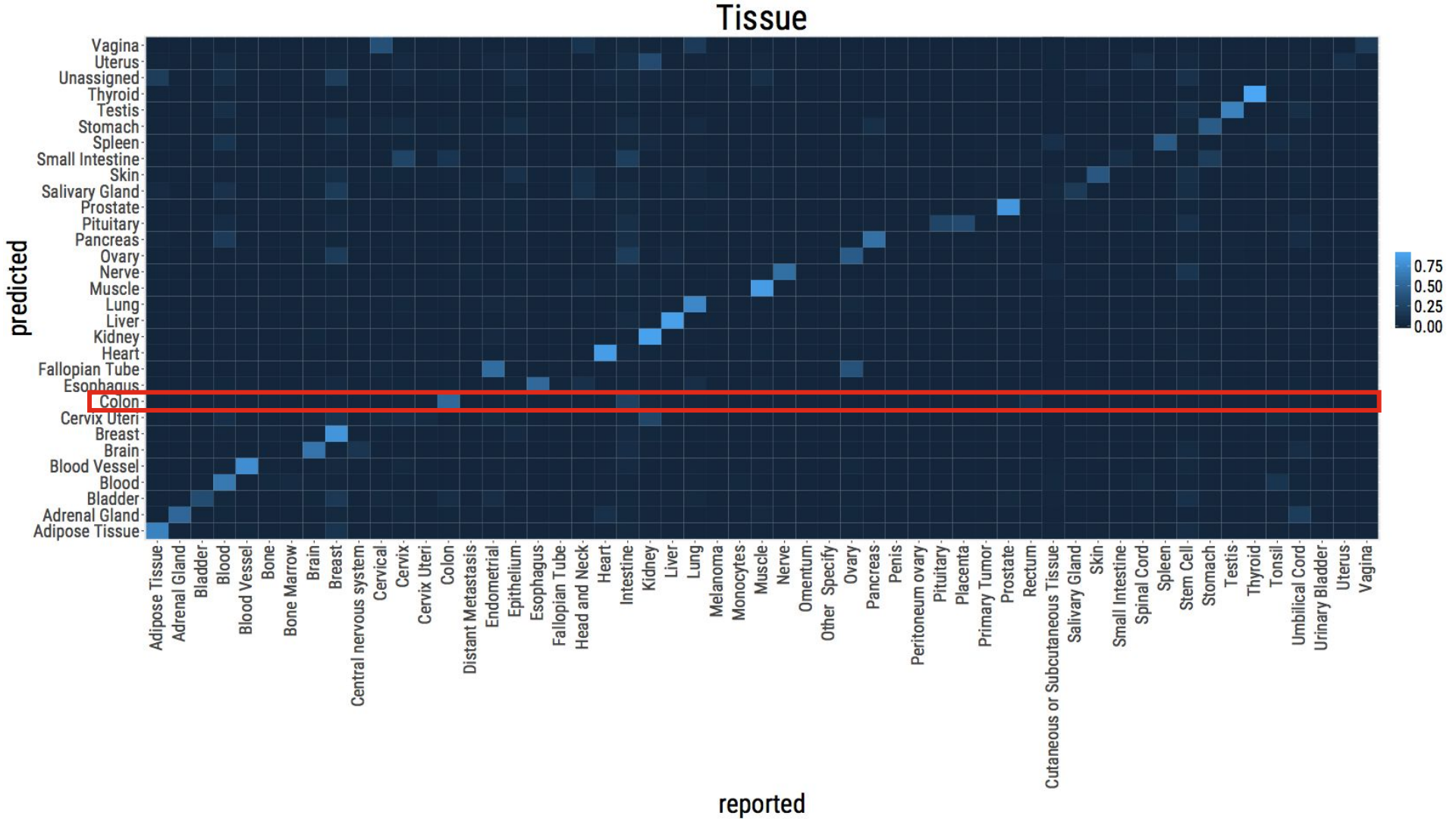


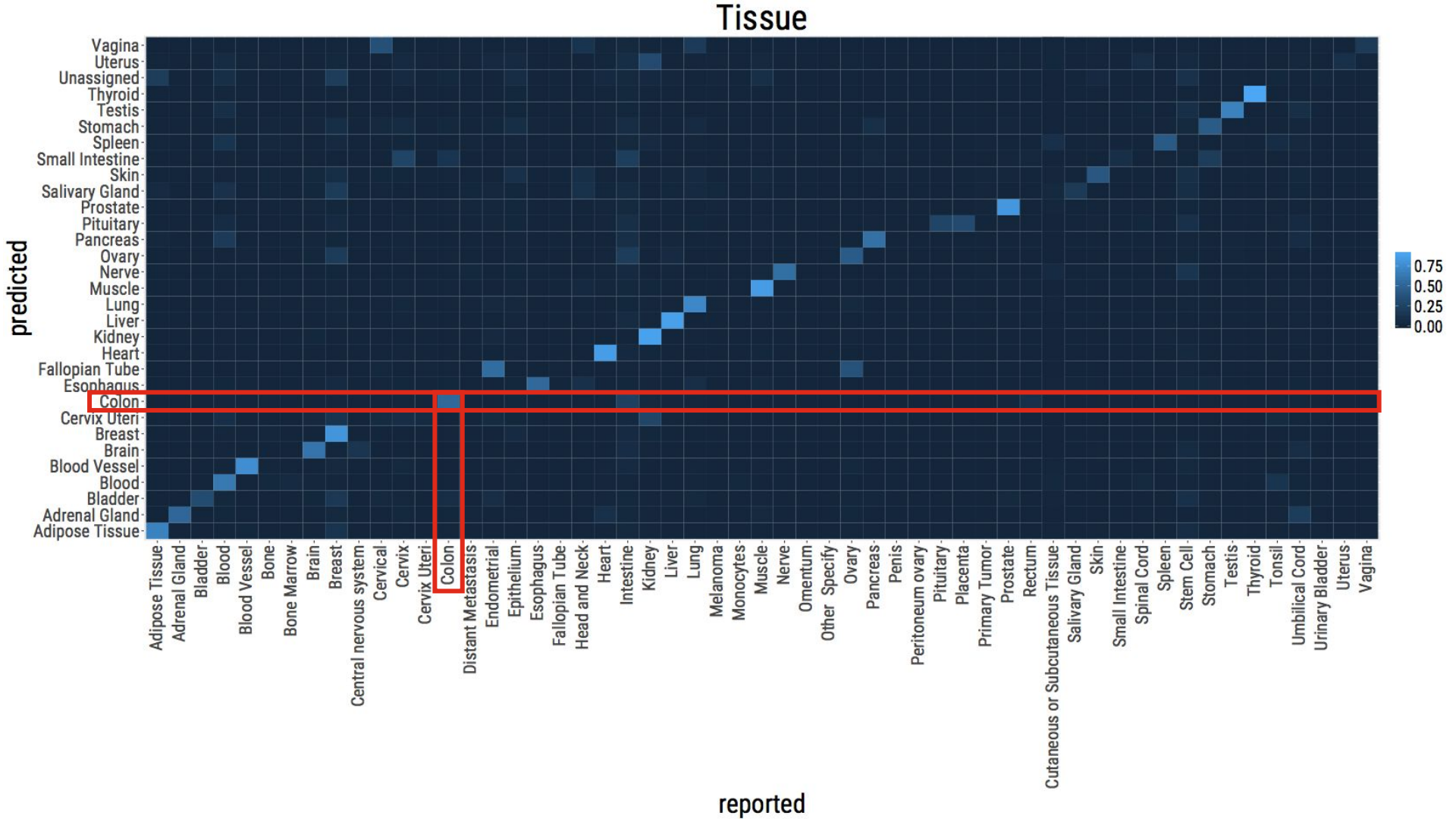
Number of Regions	2,281	2,281	2,281	2,281	2,281
Number of Samples (N)	4,769	4,769	613	6,704	8,951

Across the samples in *recount*, brain, blood, and skin are the three most frequently predicted tissues types

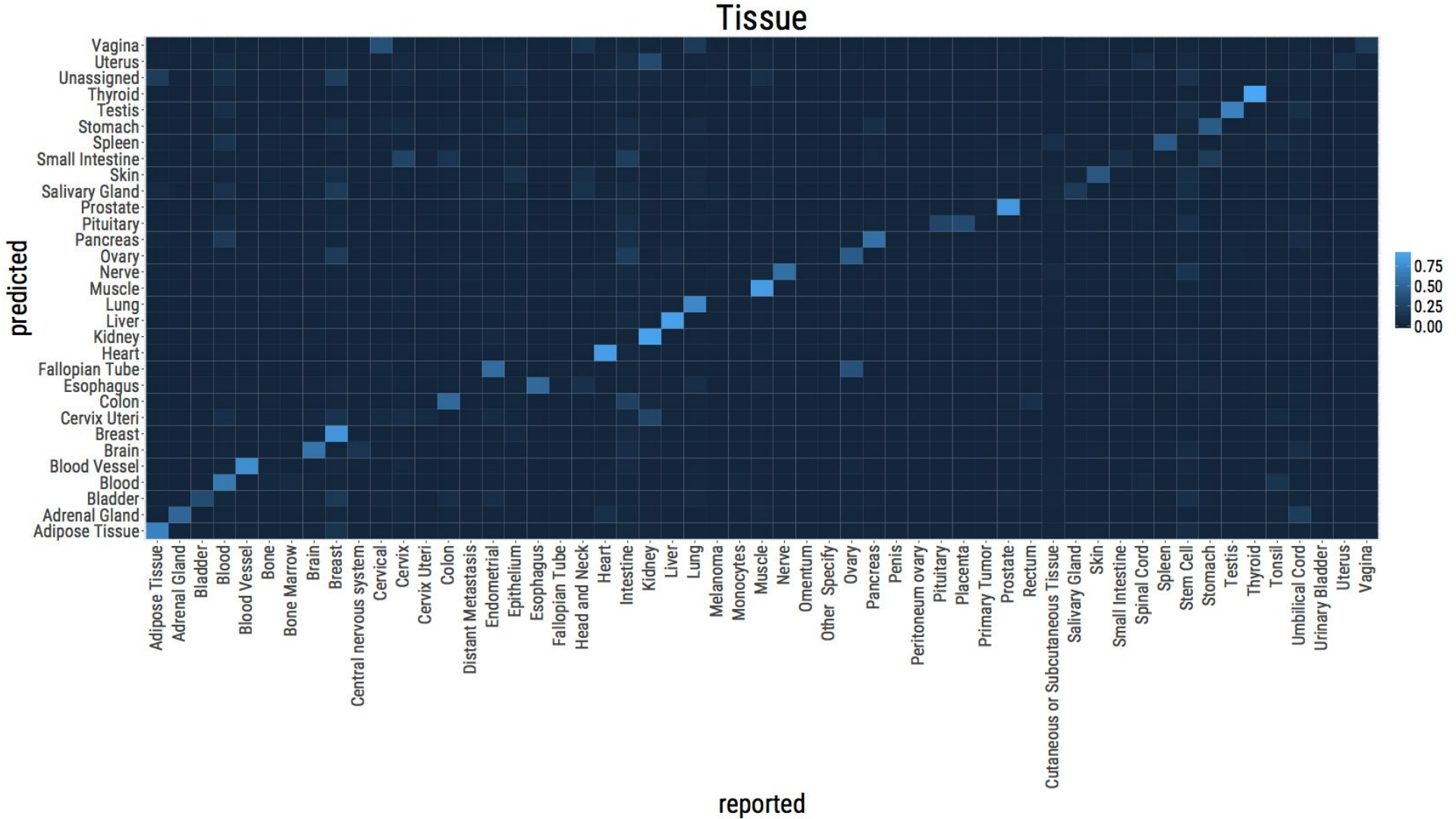



Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Bone
Bone Marrow
Brain
Breast
Central nervous system
Cervical
Cervix
Cervix Uteri
Colon
Distant Metastasis
Endometrial
Epithelium
Esophagus
Fallopian Tube
Head and Neck
Heart
Intestine
Kidney
Liver
Lung
Melanoma
Monocytes
Muscle
Nerve
Omentum
Other Specify
Ovary
Pancreas
Penis
Peritoneum ovary
Pituitary
Placenta
Primary Tumor
Prostate
Rectum
Cutaneous or Subcutaneous Tissue
Salivary Gland
Skin
Small Intestine
Spinal Cord
Spleen
Stem Cell
Stomach
Testis
Thyroid
Tonsil
Umbilical Cord
Urinary Bladder
Uterus
Vagina
0.75
0.50
0.25
0.00


Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Bone
Bone Marrow
Brain
Breast
Central nervous system
Cervical
Cervix
Cervix Uteri
Colon
Distant Metastasis
Endometrial
Epithelium
Esophagus
Fallopian Tube
Head and Neck
Heart
Intestine
Kidney
Liver
Lung
Melanoma
Monocytes
Muscle
Nerve
Omentum
Other Specify
Ovary
Pancreas
Penis
Peritoneum ovary
Pituitary
Placenta
Primary Tumor
Prostate
Rectum
Cutaneous or Subcutaneous Tissue
Salivary Gland
Skin
Small Intestine
Spinal Cord
Spleen
Stem Cell
Stomach
Testis
Thyroid
Tonsil
Umbilical Cord
Urinary Bladder
Uterus
Vagina
0.75
0.50
0.25
0.00


Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Bone
Bone Marrow
Brain
Breast
Central nervous system
Cervical
Cervix
Cervix Uteri
Colon
Distant Metastasis
Endometrial
Epithelium
Esophagus
Fallopian Tube
Head and Neck
Heart
Intestine
Kidney
Liver
Lung
Melanoma
Monocytes
Muscle
Nerve
Omentum
Other Specify
Ovary
Pancreas
Penis
Peritoneum ovary
Pituitary
Placenta
Primary Tumor
Prostate
Rectum
Cutaneous or Subcutaneous Tissue
Salivary Gland
Skin
Small Intestine
Spinal Cord
Spleen
Stem Cell
Stomach
Testis
Thyroid
Tonsil
Umbilical Cord
Urinary Bladder
Uterus
Vagina
0.75
0.50
0.25
0.00


Tissue
A sample reported to be *Intestine* is predicted to be *Colon*. That makes good sense.
predicted
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
reported
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Bone
Bone Marrow
Brain
Breast
Central nervous system
Cervical
Cervix
Cervix Uteri
Colon
Distant Metastasis
Endometrial
Epithelium
Esophagus
Fallopian Tube
Head and Neck
Heart
Intestine
Kidney
Liver
Lung
Melanoma
Monocytes
Muscle
Nerve
Omentum
Other Specify
Ovary
Pancreas
Penis
Peritoneum ovary
Pituitary
Placenta
Primary Tumor
Prostate
Rectum
Cutaneous or Subcutaneous Tissue
Salivary Gland
Skin
Small Intestine
Spinal Cord
Spleen
Stem Cell
Stomach
Testis
Thyroid
Tonsil
Umbilical Cord
Urinary Bladder
Uterus
Vagina
0.75
0.50
0.25
0.00

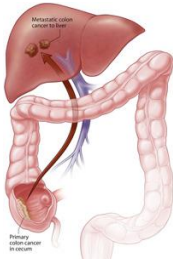

Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Bone
Bone Marrow
Brain
Breast
Central nervous system
Cervical
Cervix
Cervix Uteri
Colon
Distant Metastasis
Endometrial
Epithelium
Esophagus
Fallopian Tube
Head and Neck
Heart
Intestine
Kidney
Liver
Lung
Melanoma
Monocytes
Muscle
Nerve
Omentum
Other Specify
Ovary
Pancreas
Penis
Peritoneum ovary
Pituitary
Placenta
Primary Tumor
Prostate
Rectum
Cutaneous or Subcutaneous Tissue
Salivary Gland
Skin
Small Intestine
Spinal Cord
Spleen
Stem Cell
Stomach
Testis
Thyroid
Tonsil
Umbilical Cord
Urinary Bladder
Uterus
Vagina
0.75
0.50
0.25
0.00

Tissue prediction is largely accurate across *recount2*

Tissue can be accurately predicted from expression data.

Discordant predictions are often made to biologically similar tissues.

Sometimes, predictions are inaccurate.



What makes primary cancer different than metastatic cancer?

Get already processed and summarized RNA-Seq data from *recount2*

Collect patient samples and information

Sequence samples

Process sequencing data

Determine expression differences between primary cancer and metastatic cancer

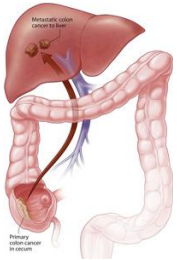
3~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

Find a researcher with access to patient samples

Extract DNA/RNA from samples

Data cleaning

Total: months



What makes primary cancer different than metastatic cancer?

Get already processed and summarized RNA-Seq data from *recount2* **with sample information**

Collect patient samples and information

Sequence samples

Process sequencing data

Determine expression differences between primary cancer and metastatic cancer

3~6 months | 1-2 wks | 2-4 wks | 3-6 months | 1-3 months | 1 month - 1+ years

Find a researcher with access to patient samples

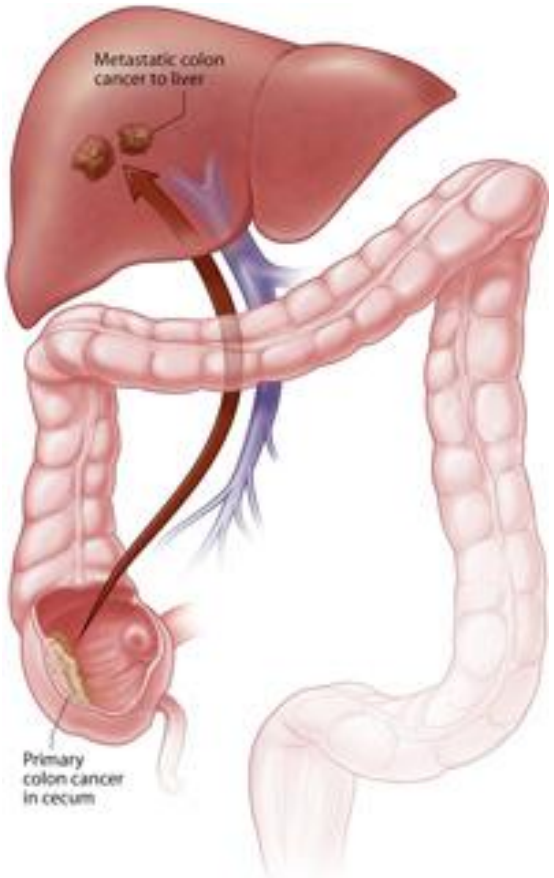
Extract DNA/RNA from samples

Data cleaning

Total: months

Ok. Ok. What about actually *using* these data and predictions…?

What makes primary cancer different than metastatic cancer?



www.hopkinsmedicine.org

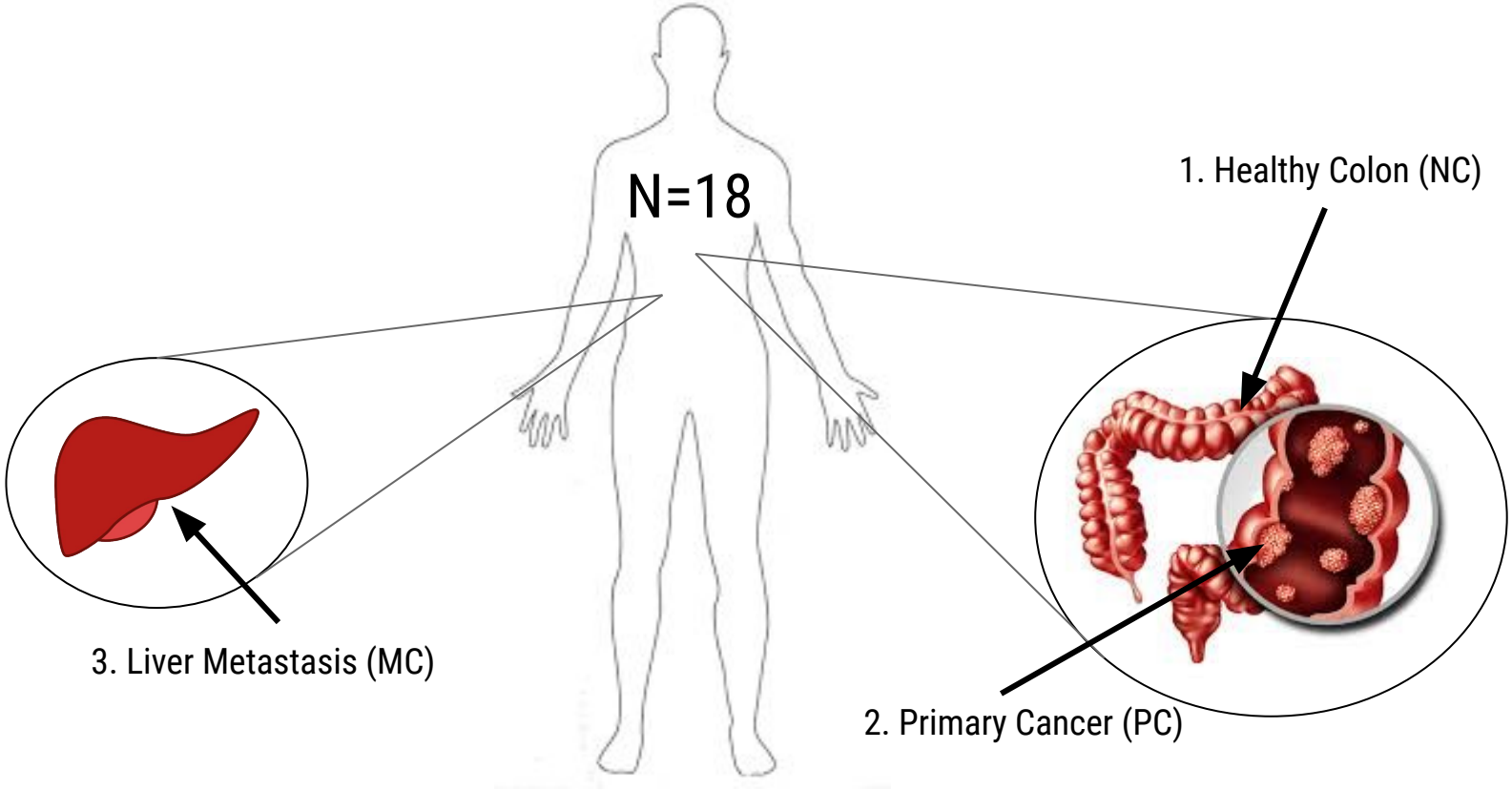
A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients

Seon-Kyu Kim[a,1], Seon-Young Kim[a,1], Jeong-Hwan Kim[a], Seon Ae Roh[b,c], Dong-Hyung Cho[c,d], Yong Sung Kim[a,c,**], Jin Cheon Kim[b,c,*]

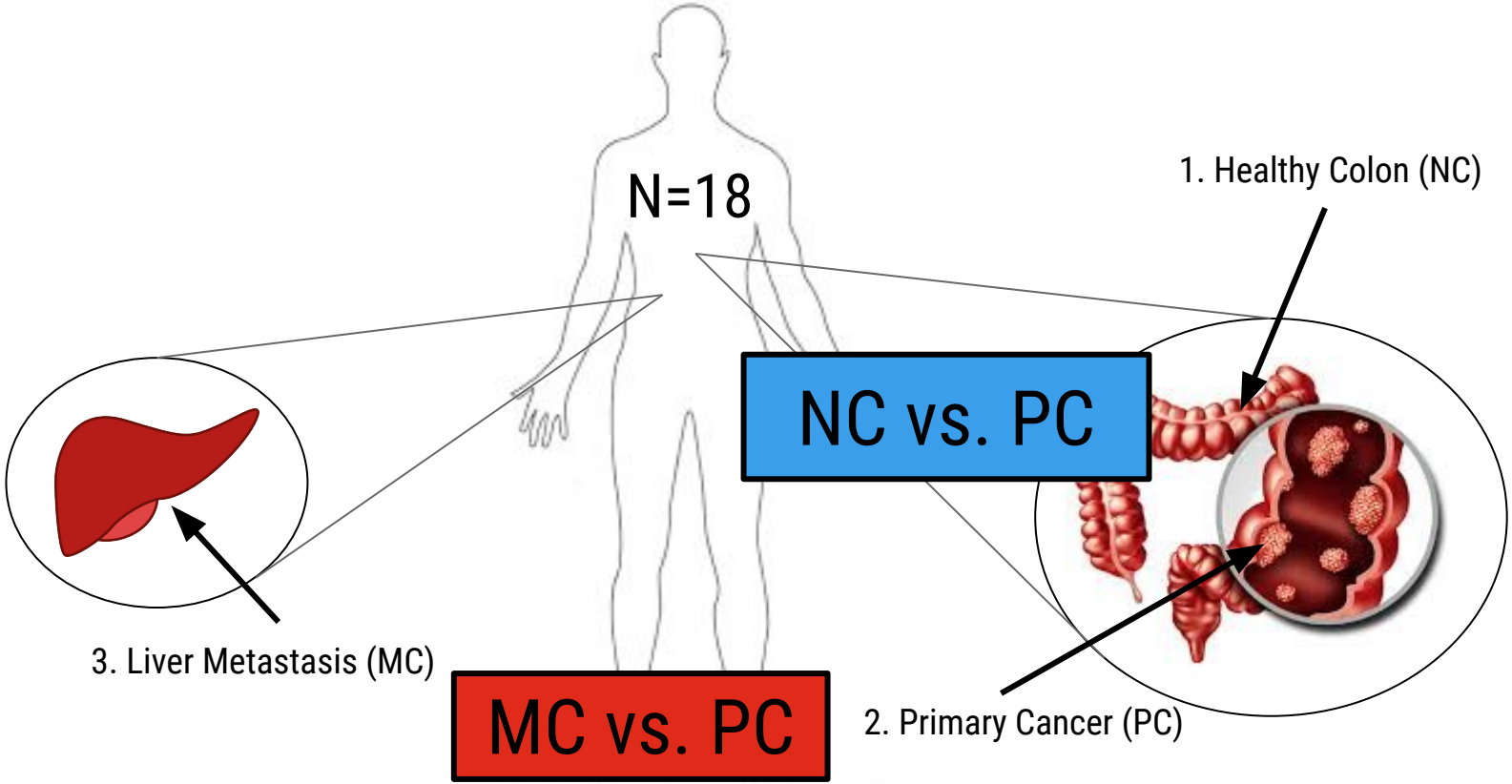
[a]Medical Genomics Research Centre, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea
[b]Department of Surgery, University of Ulsan College of Medicine, Seoul, Korea
[c]Department of Cancer Research, Institute of Innovative Cancer Research and Asan Institute for Life Sciences, Asan Medical Centre, Seoul, Korea
[d]Graduate School of East-West Medical Science, Kyung Hee University, Gyeonggi-do, Korea

Molecular Oncology, July 2014

Kim et al. analysis looked to identify genes that contribute to metastasis in colon cancer.



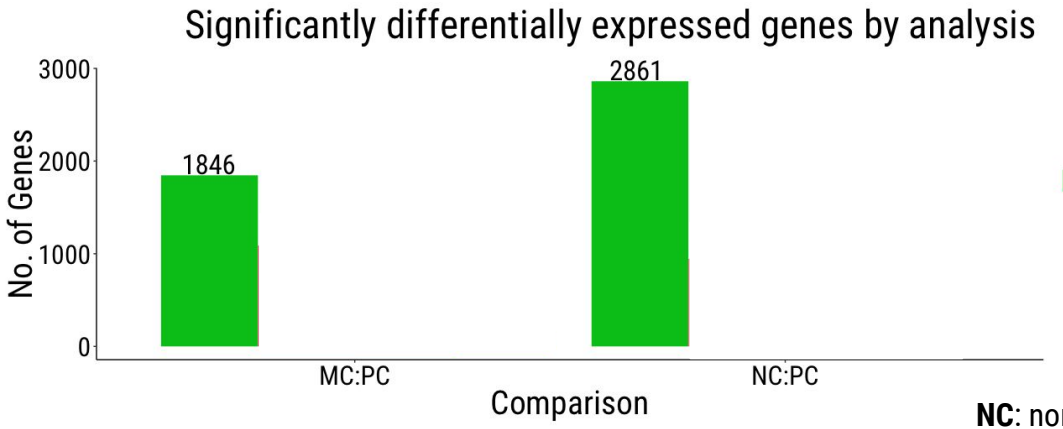
Kim et al. analysis looked to identify genes that contribute to metastasis in colon cancer.



Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data

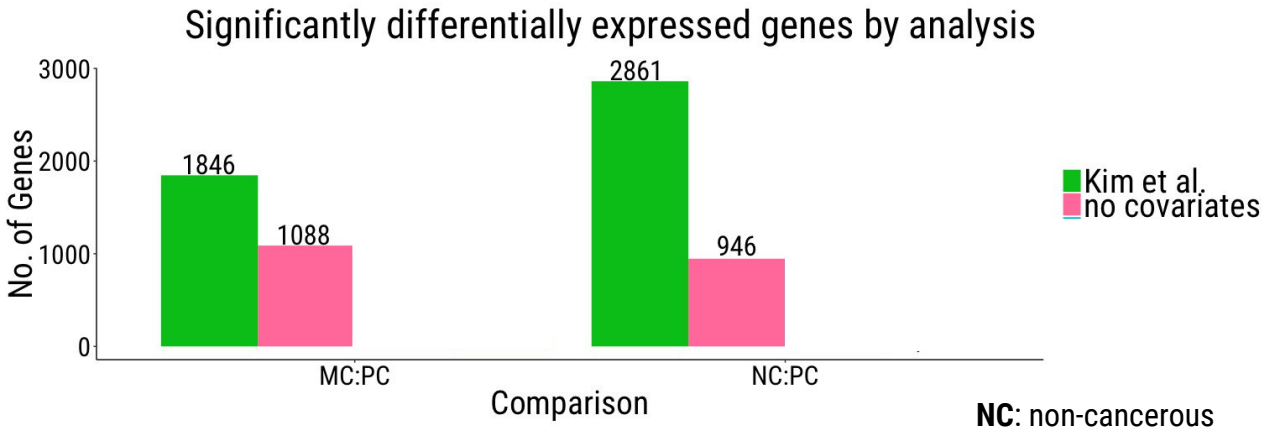


NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data

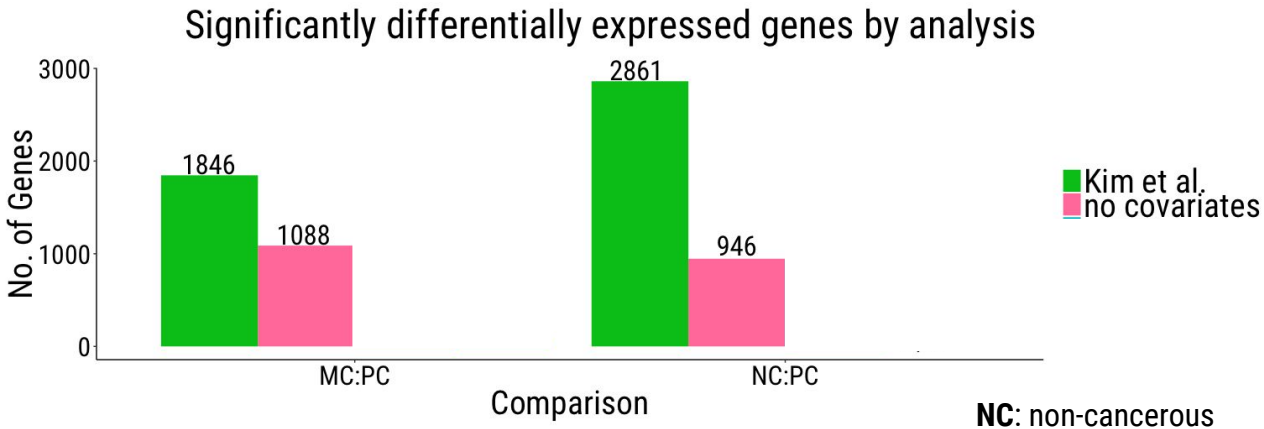


NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data



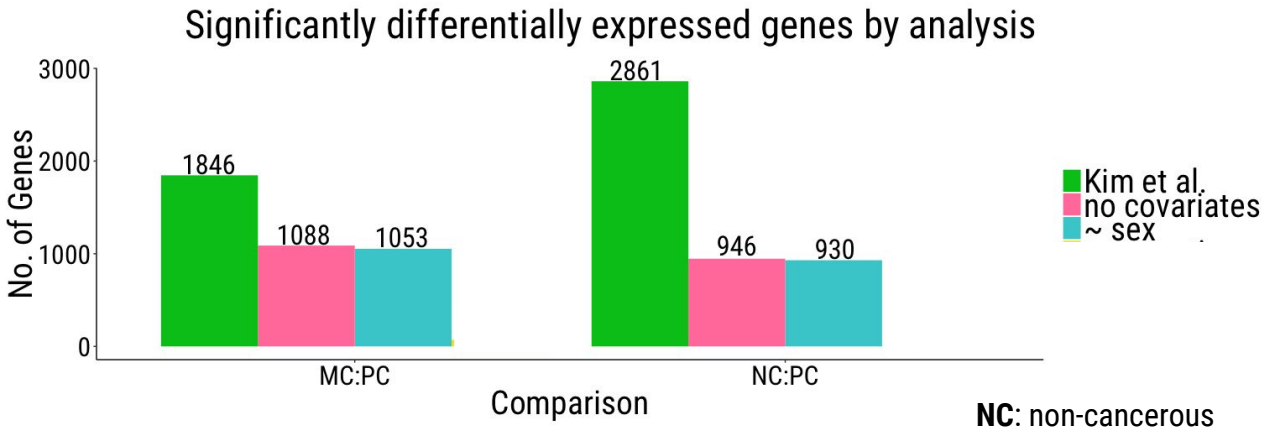
NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

Are the same genes found when sex is included in the analysis?

Predictions can be used to:

(1) Identify studies of interest

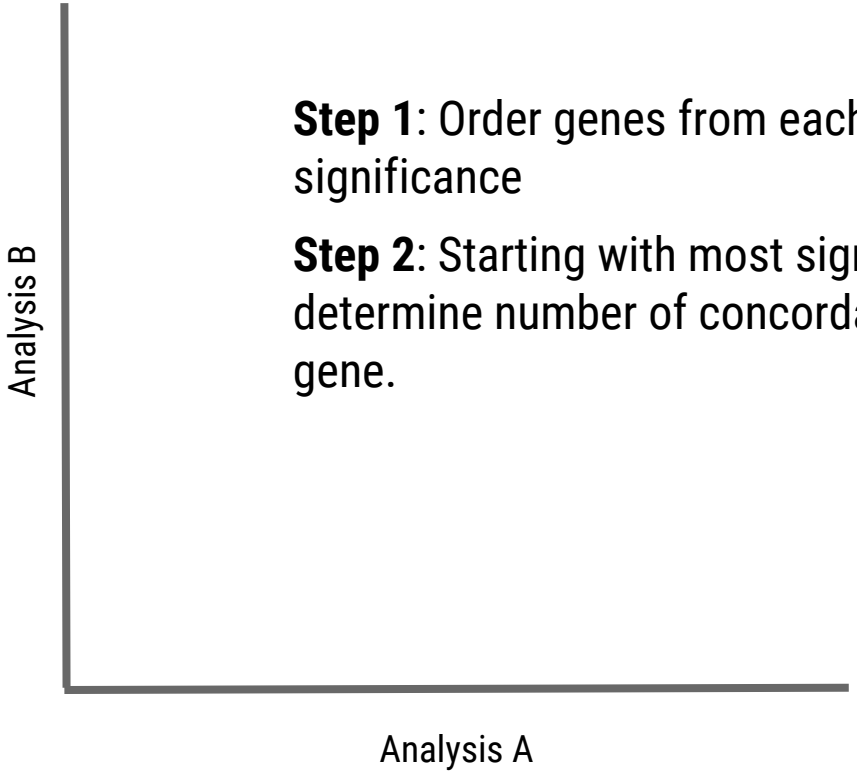
(2) appropriately analyze data



NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?

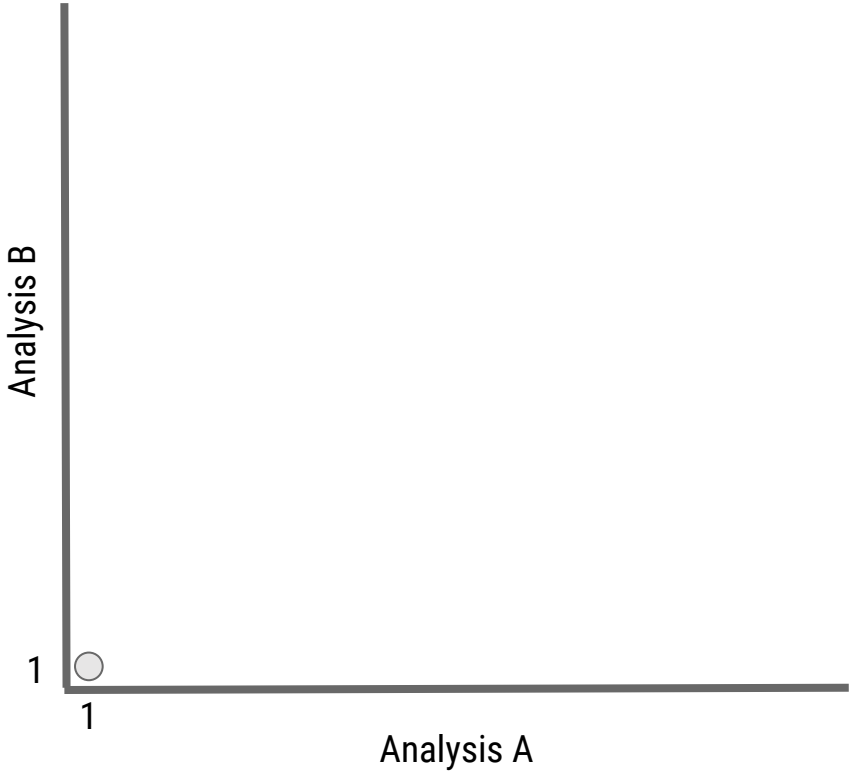


Step 1: Order genes from each analysis by significance

Step 2: Starting with most significant gene, determine number of concordant genes at each gene.

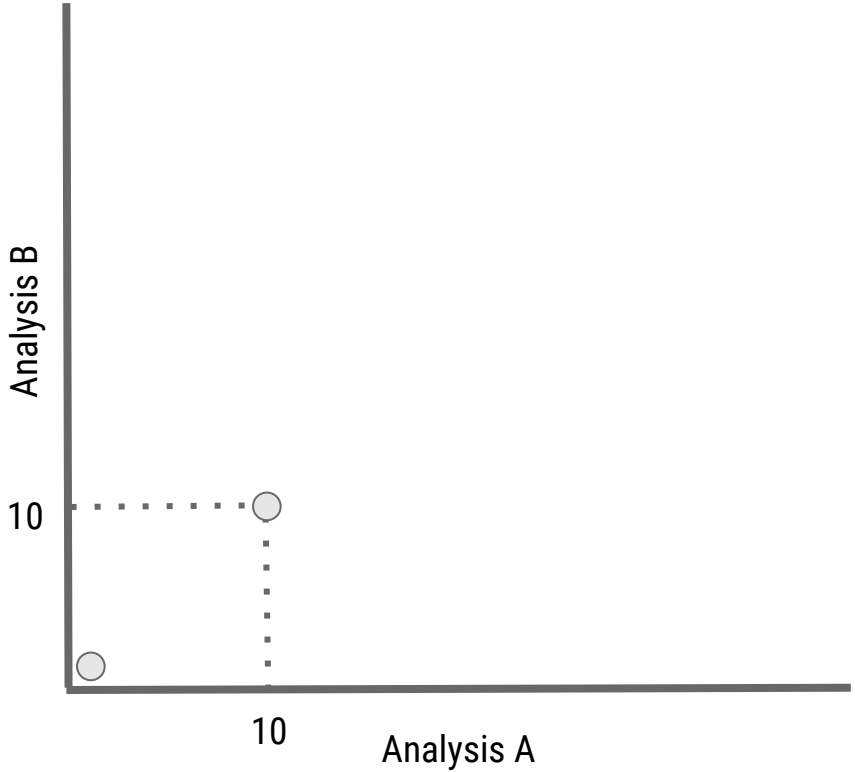
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?



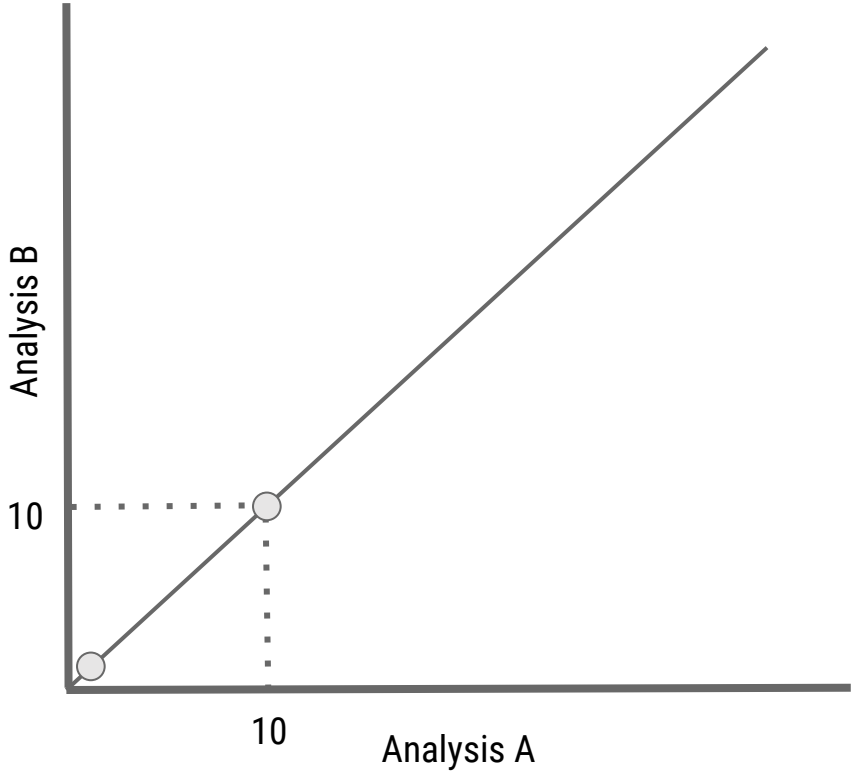
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?



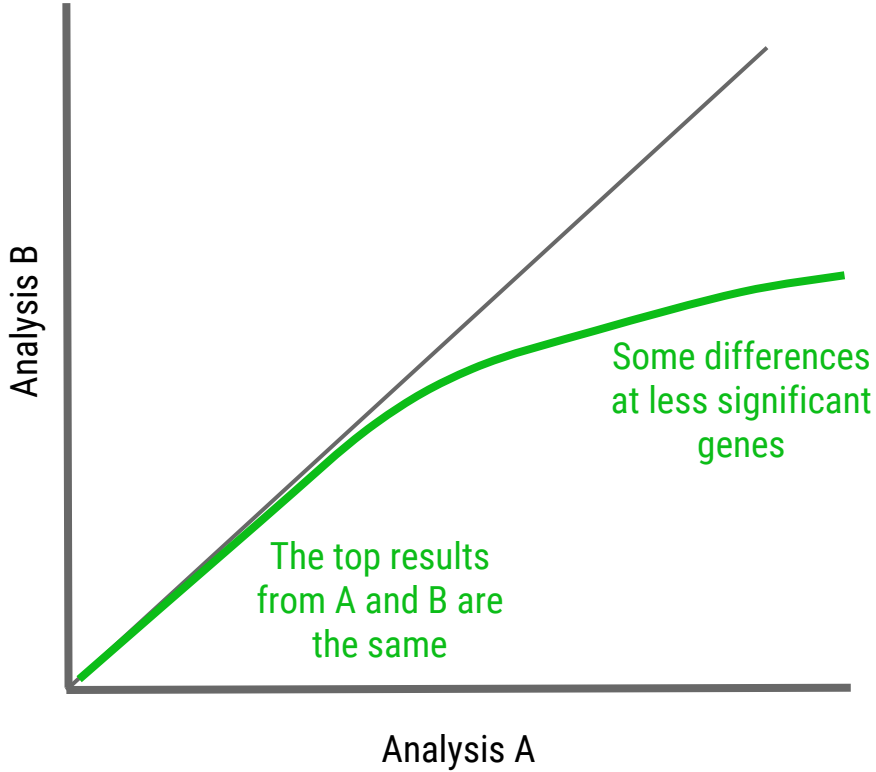
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?



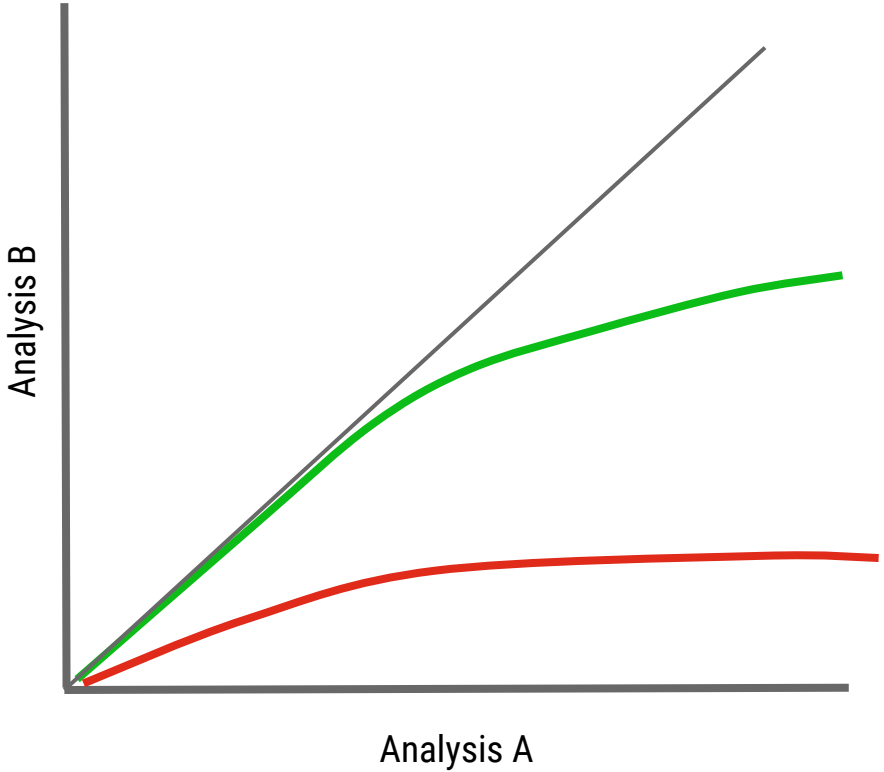
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?



Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B ?

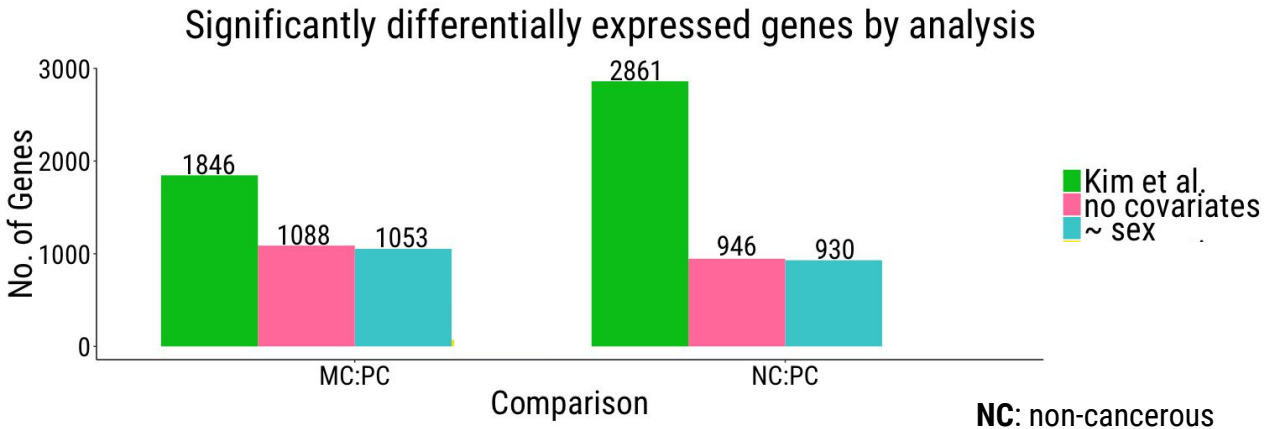


The results of the red condition are less similar between Analysis A and B than the green condition

Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data

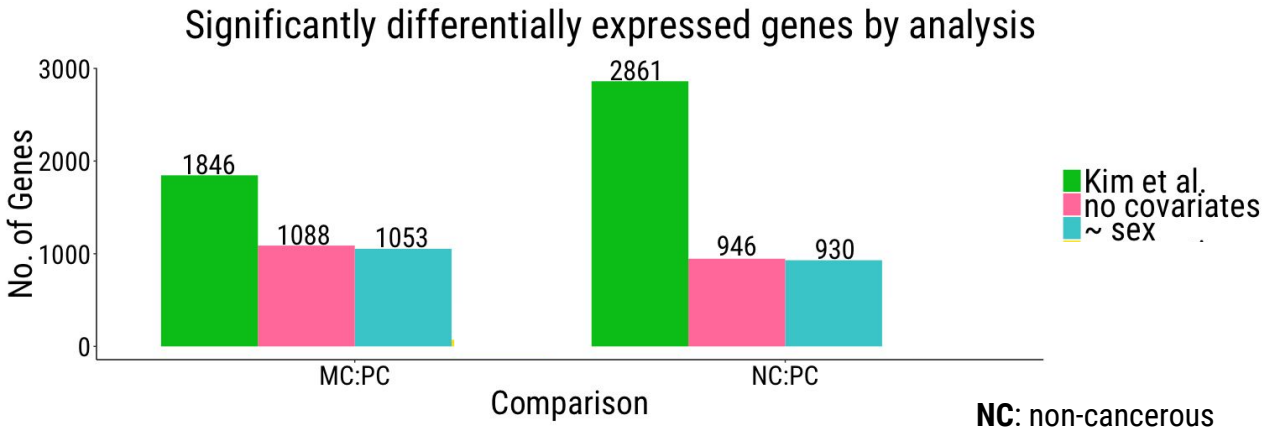


NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

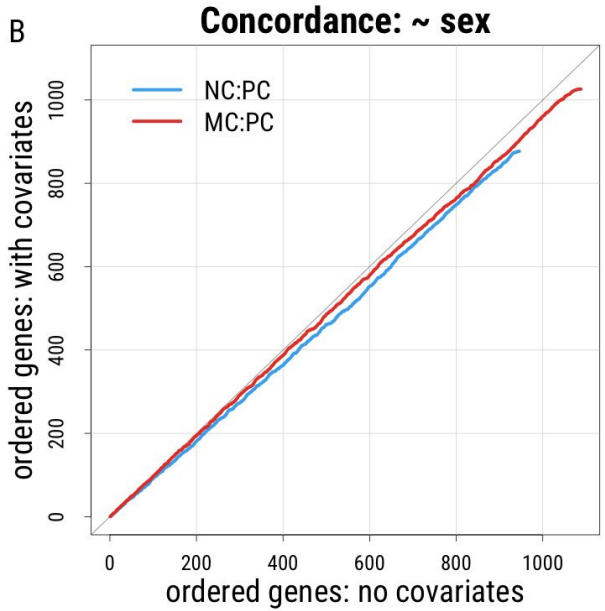
Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data



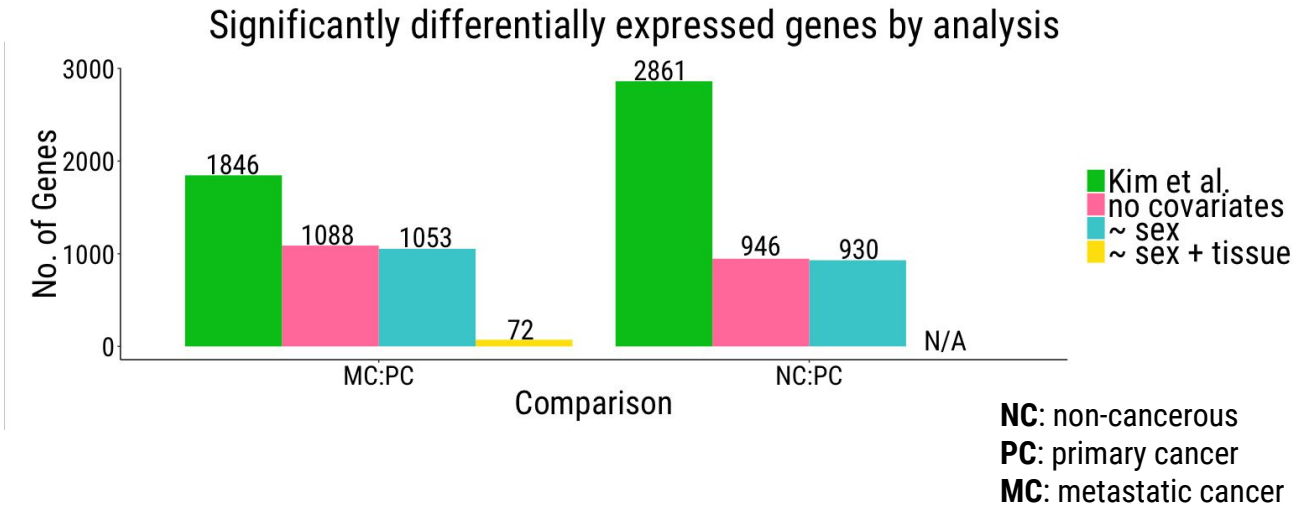
NC: non-cancerous
PC: primary cancer
MC: metastatic cancer



Predictions can be used to:

(1) Identify studies of interest

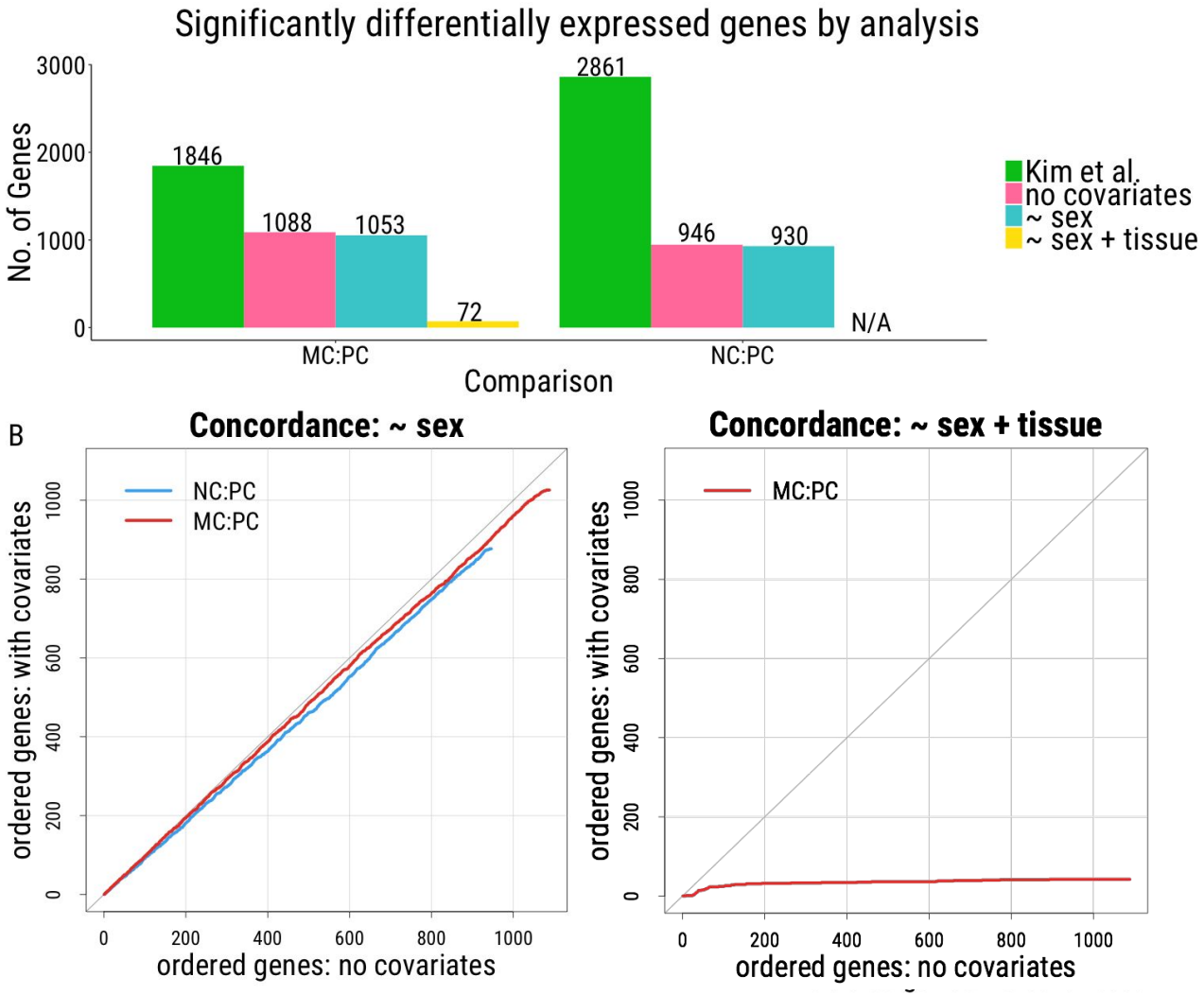
(2) appropriately analyze data



Predictions can be used to:

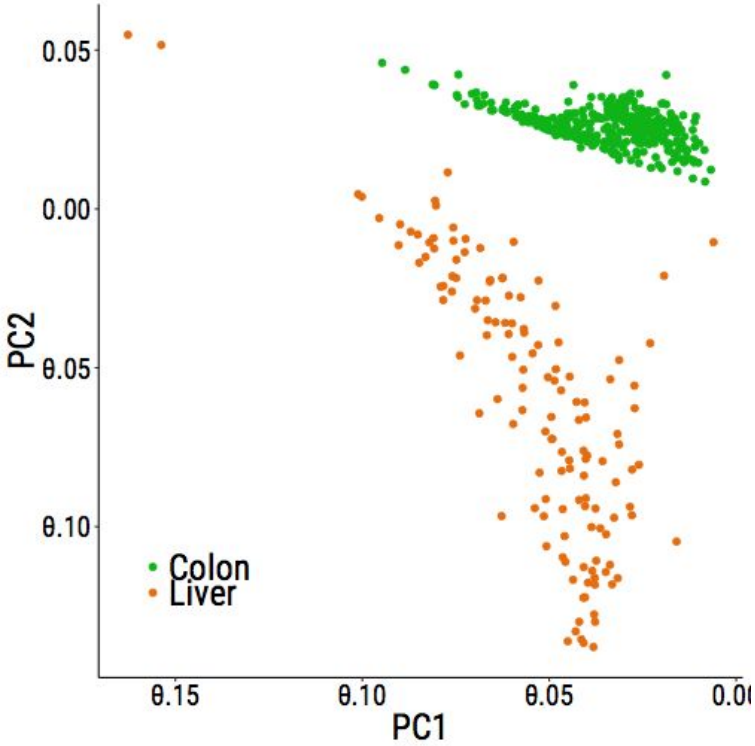
(1) Identify studies of interest

(2) appropriately analyze data

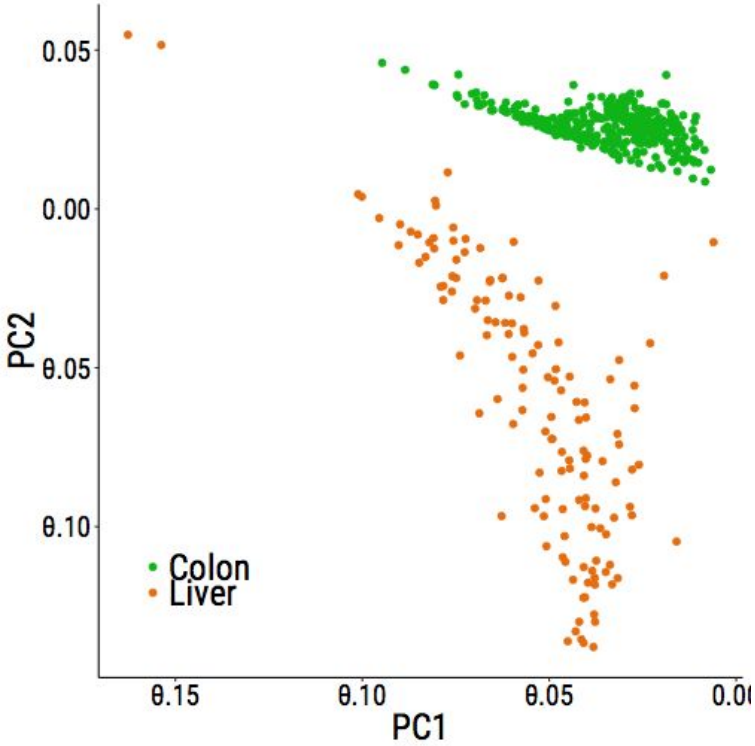


Loss of concordance suggests that differential expression is detecting tissue differences, not cancer-related changes.

We have expression data from both healthy liver and colon samples (GTEx)...

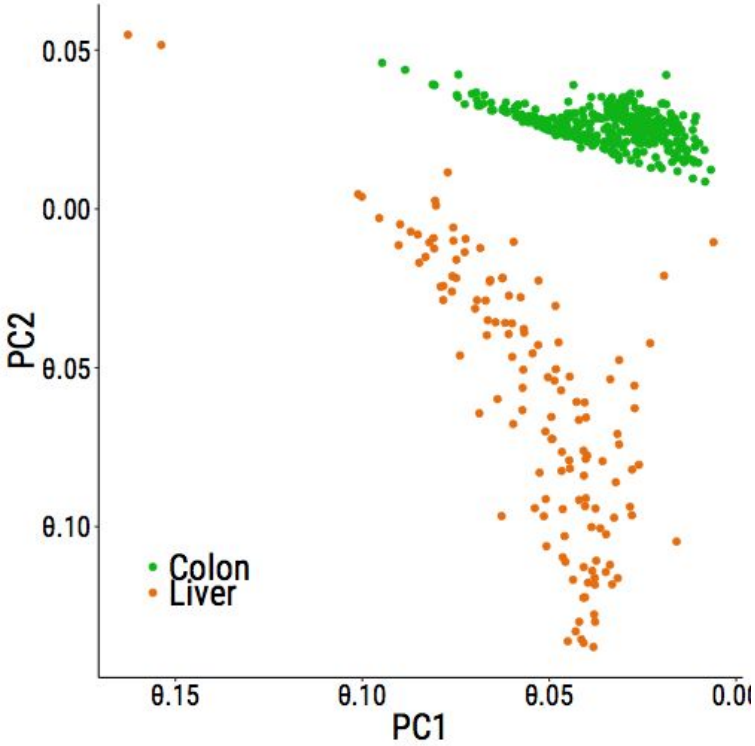


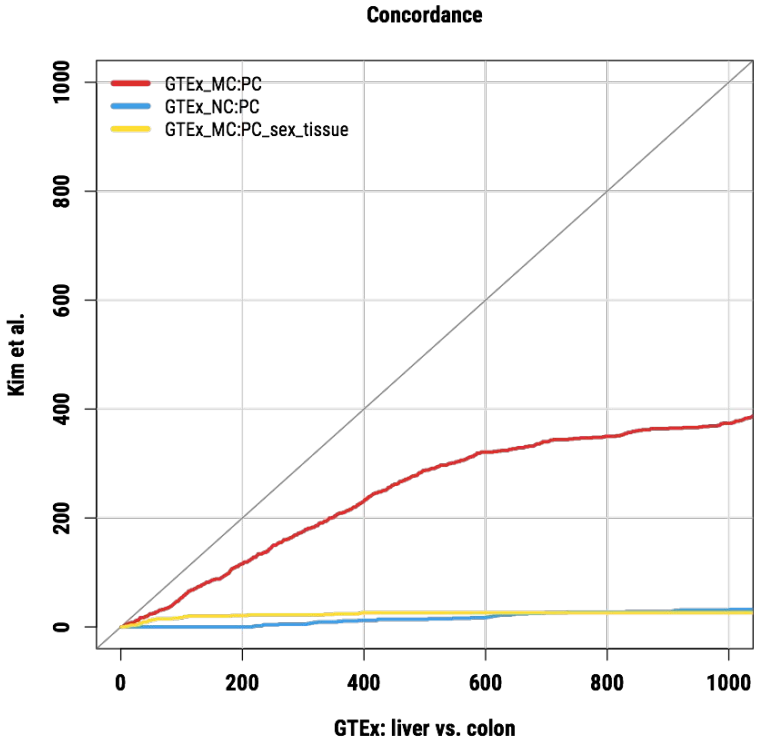
So…what if we compared the MC:PC results with differential expression between colon and liver?

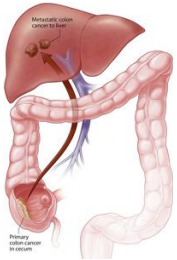


Hypothesis: MC:PC results should be most similar to GTEx colon vs. liver

Comparison of results with GTEx colon vs. liver suggests differential expression results detecting tissue differences







What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

Collect patient samples and information

3~6 months

Extract DNA/RNA from samples

1-2 wks

Sequence samples

2-4 wks

Process sequencing data

3-6 months

Data cleaning

1-3 months

Determine expression differences between primary cancer and metastatic cancer

1 month - 1+ years

Total: days-months

The Leek group

- Jack Fu
- Aboozar Hadavand
- Leslie Myint
- Kayode Sosina
- Sara Wang
- **Jeff Leek**

Collaborators

- **Andrew Jaffe**
- Kasper Hansen
- Margaret Taub
- Leah Jager
- **Sean Kross**
- **Ben Langmead**
- **Abhi Nellore**
- Kai Kammers
- **Leo Collado-Torres**
- Ashkaun Razmara

A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD

Chromebook Data Science (CBDS)

Find a **partner organization**

Collaboratively **develop course content**

Develop **new technology** as needed

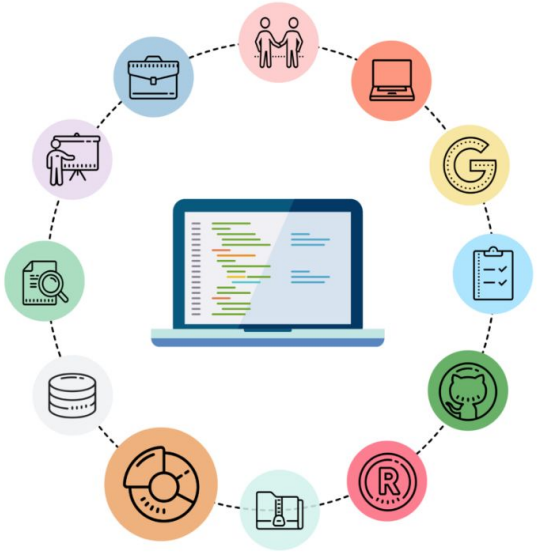
Design in-person **tutoring program**

Launch program, **teach the stuff** & get learners **jobs**

CBDS

(Chromebook Data Science)

Library Author Community Support



Chromebook Data Science

Course 7: Data Visualization

The instructor has published 100% of this course.

Data Visualization

Jeffrey Leek

$0.00 / MINIMUM $25.00 / SUGGESTED

YOU PAY

$25.00

YOU PAY (US$)

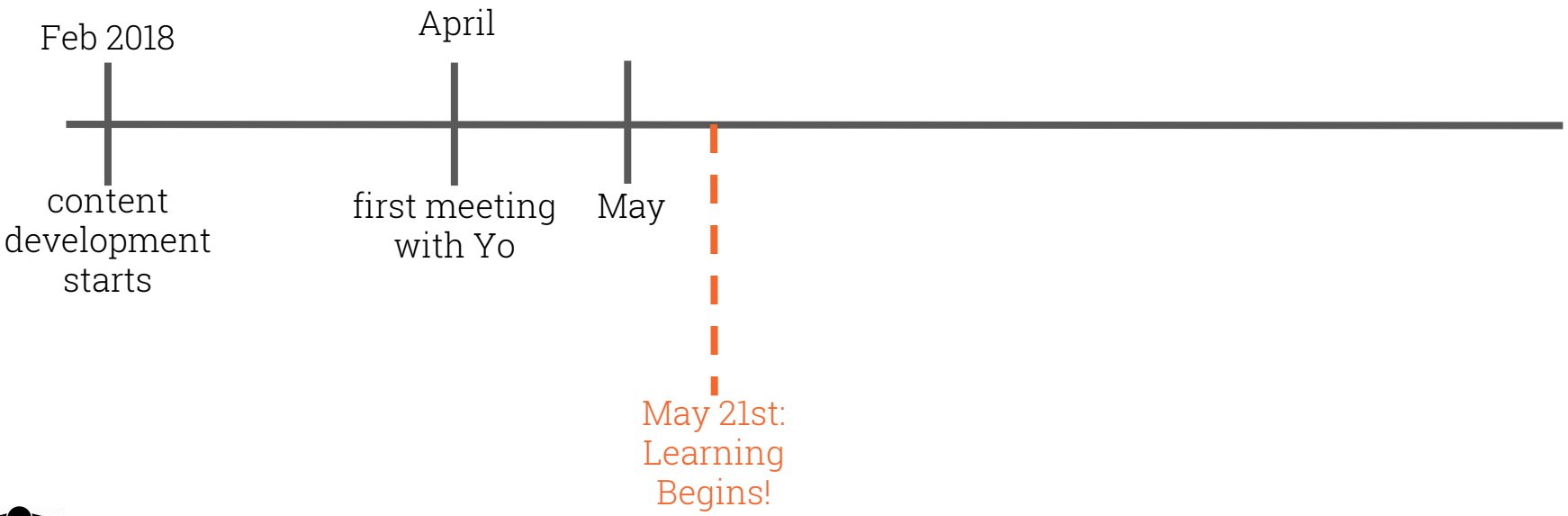
$25.00

EU customers: Price excludes VAT. VAT is added during checkout.

Add Course To Cart

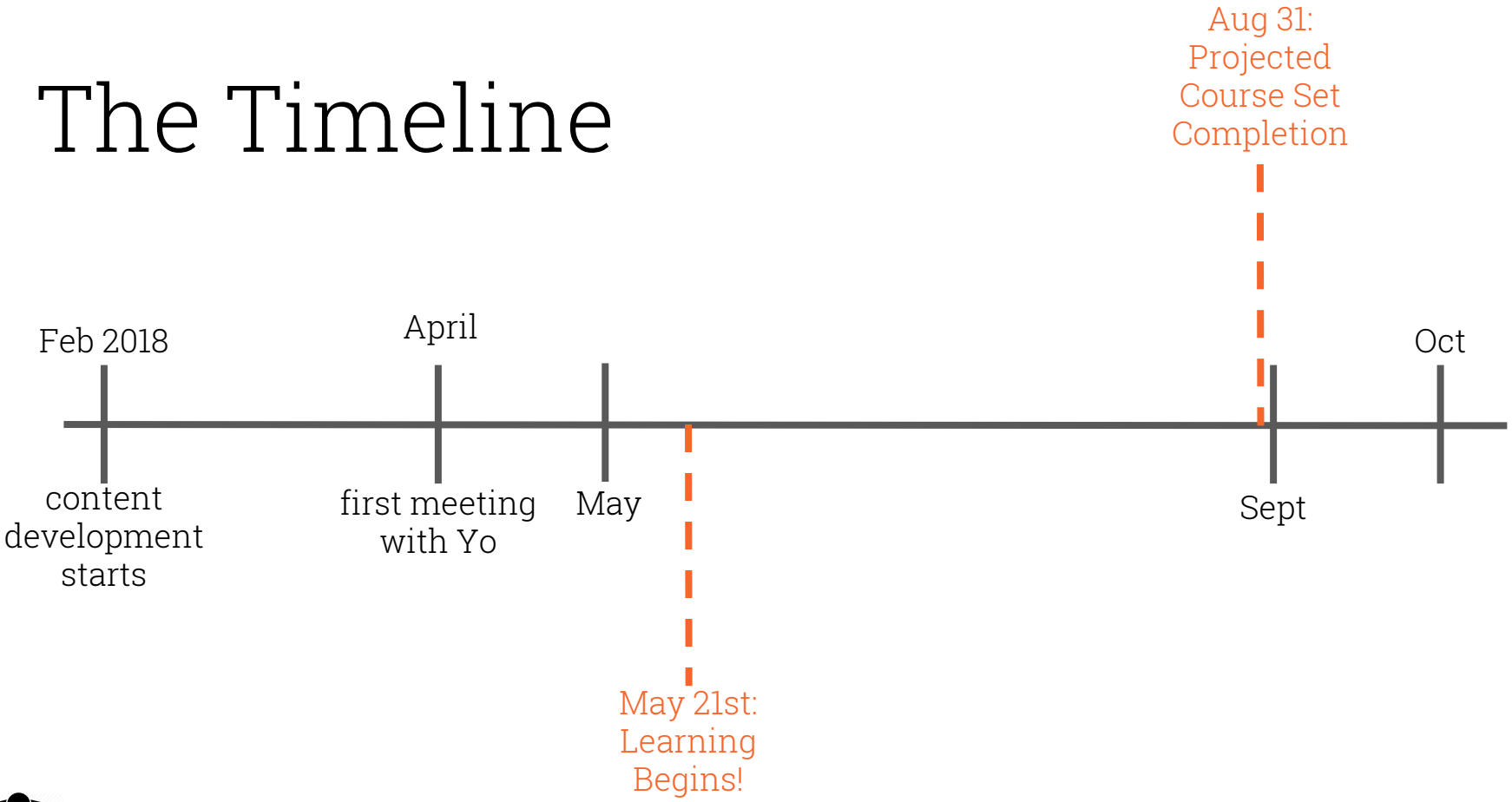
Course Info Course Materials Instructors Community

The Timeline



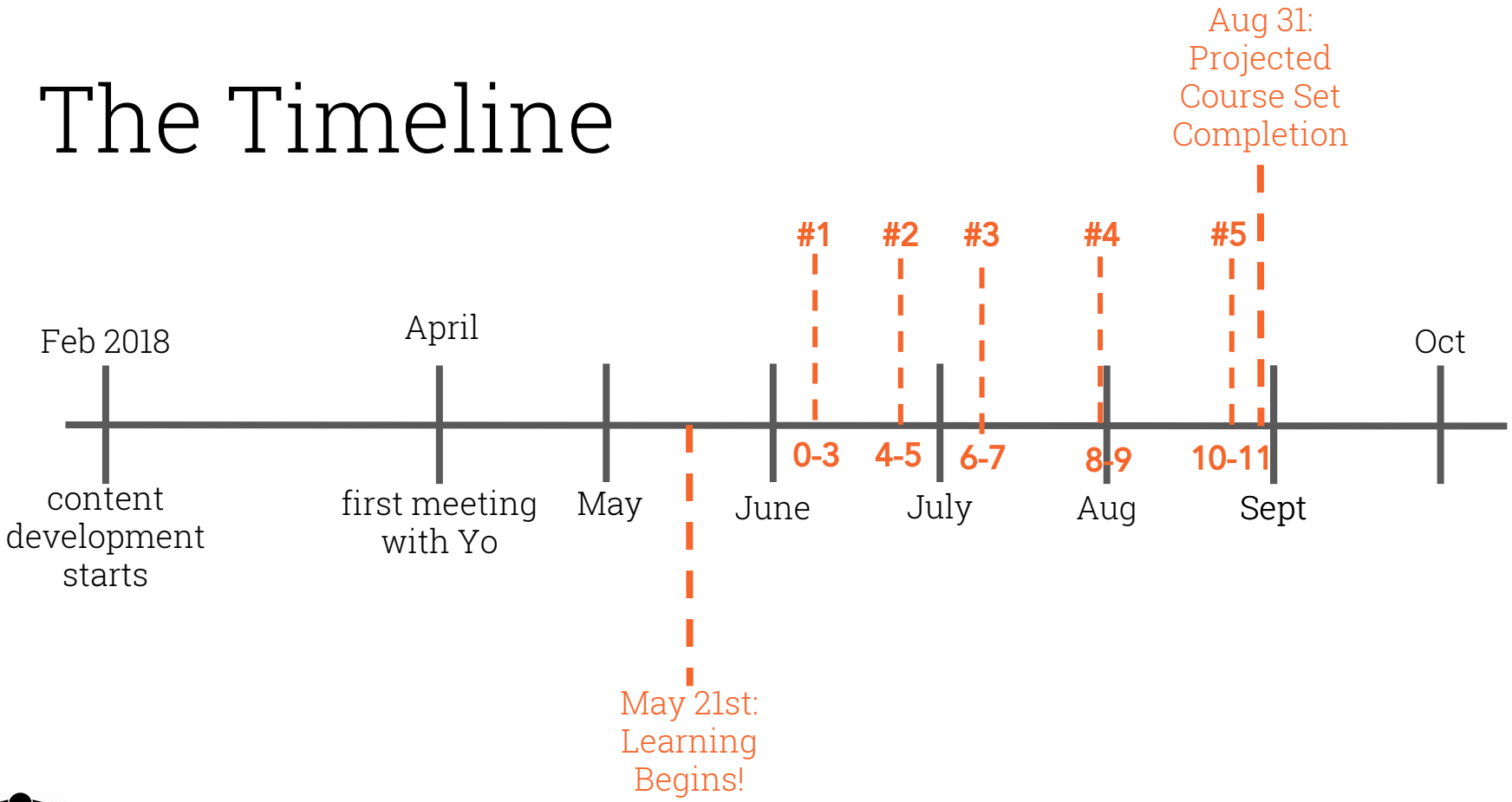
Launch program, **teach the stuff** & get learners **jobs**

The Timeline



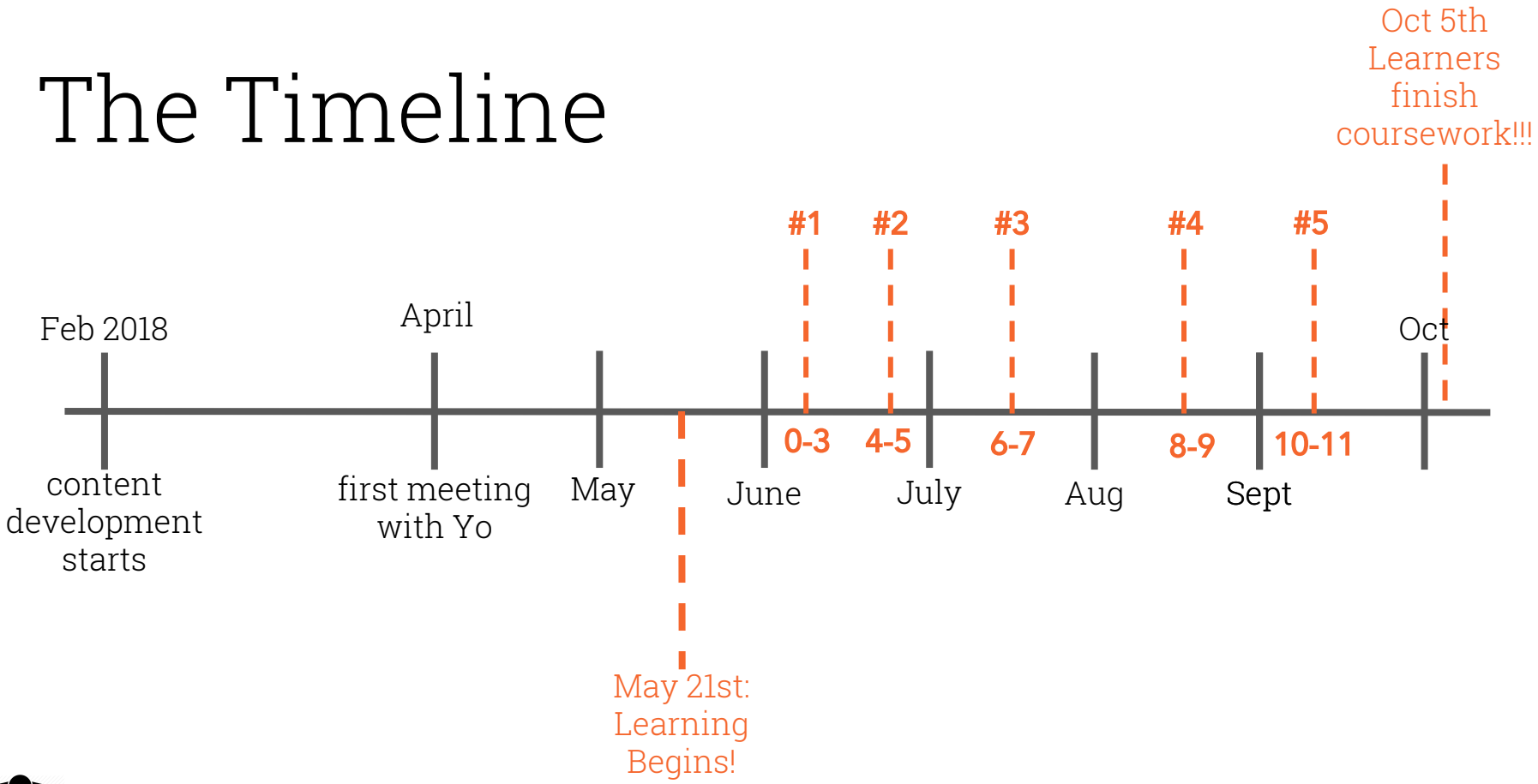
Launch program, **teach the stuff** & get learners **jobs**

The Timeline



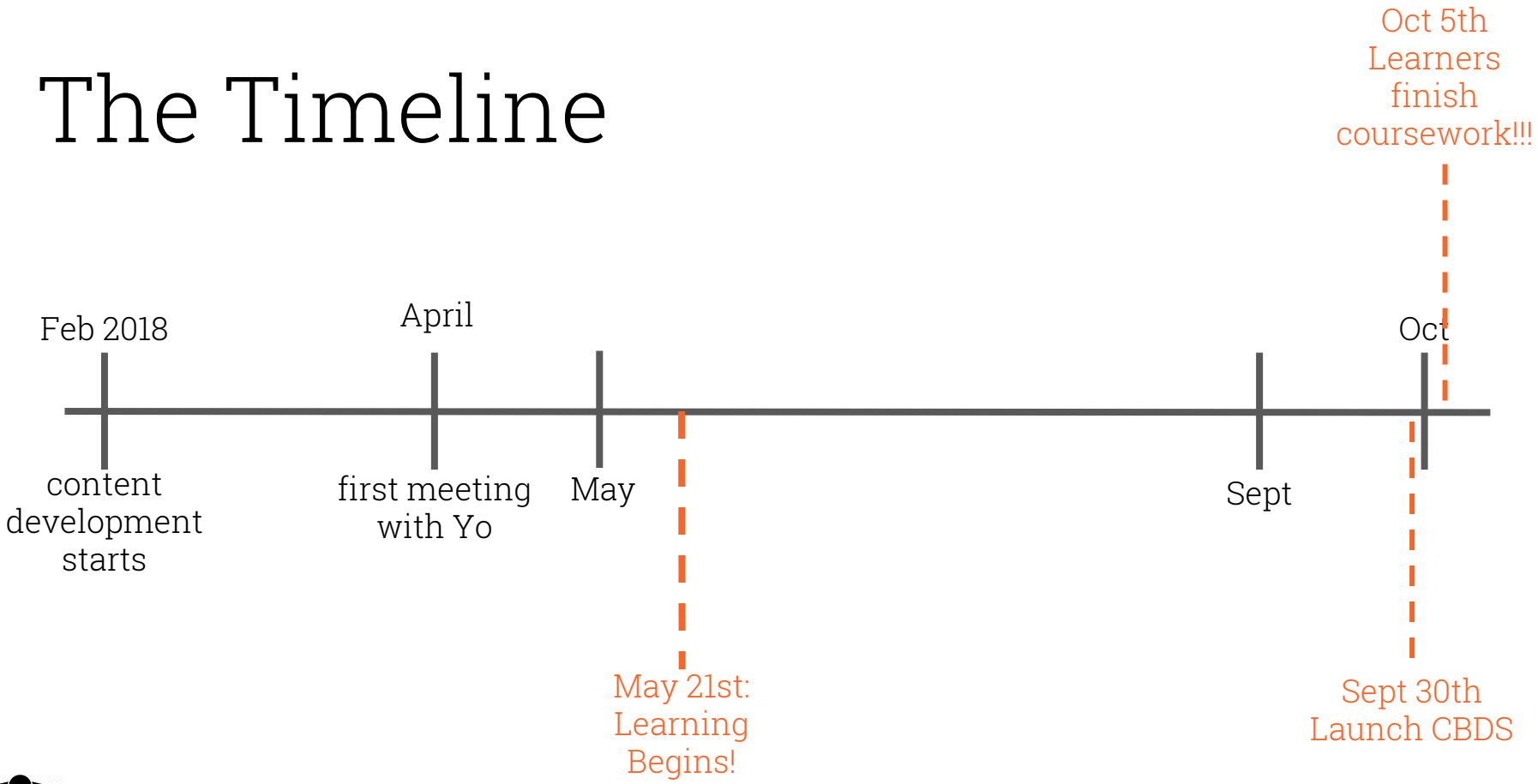
Launch program, **teach the stuff** & get learners **jobs**

The Timeline



Launch program, **teach the stuff** & get learners **jobs**

The Timeline



Launch program, **teach the stuff** & get learners **jobs**

john "as b/4" fink
@adr

Follow

rstudio and Chromebook Data Science are so fun I am working on them on a Saturday night so thanks a buncha buncha @kierisi for the recommendation.

9:40 PM - 6 Apr 2019

2 Retweets **10** Likes

2 2 10

https://twitter.com/adr/status/1114704504295972866

Chromebook Data Science Team

	Leek	Johnson	Ellis	Hadavand	Muschelli	Kross	Collado-Torres	Jager	McClymont	Myint
Content Development	●	○	●	●	○	○	●	●	●	●
Administration & Tutoring	●	●	●	●	○	○	○	○	○	○
Technology	○	○	●	○	●	●	○	○	○	○

Collaboratively **develop course content**

A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD

COGS9 : Intro to Data Science

Brad has previously written about COGS9 in his super interesting and thoughtful blog post Data Science at UC San Diego, but very briefly here, COGS9 is an intro level course designed to get undergraduate students interested in data science, familiar with what data science is, and excited to learn more. It is neither math nor computationally-heavy, but is rather taught through concepts and examples. When it first ran there were 24 students. Now, each time it is offered, there are hundreds. This quarter, when numbers settled out, I had 326 students in the course.

http://www.shanellis.com/blog/cogs9-introduction-to-data-science/

COGS 18 - Introduction To Python

- Home
- Materials
- Coding Labs
- Assignments

Welcome to COGS18: Introduction to Python!

COGS 18 · Shannon Ellis · Spring 2019 · UCSD

Overview

Introduction to Python (COGS18) is a course offered by the Department of Cognitive Science of UC San Diego, taught by Shannon Ellis. It is a hands-on programming course, focused on teaching students in Cognitive Science and related disciplines an introduction on how to productively use Python.

Current Iteration

Introduction to Python is currently running for Spring Quarter 2019, for which you can check out the current syllabus and schedule. Course lectures are recorded and are publicly available as screencasts from here.

https://cogs18.github.io/intro/

Hands-On Data Science

COGS108 - Data Science in Practice

Course materials for Hands-On Data Science.

UC San Diego COGS108@gmail.com

Repositories 23 People 19 Teams 4 Projects 0 Settings

Pinned repositories

Customize pinned repositories

Overview

Overview and map of the organization, which services COGS108: Hands-On Data Science, from UCSD.

★ 41 17

Lectures-Sp19

Slides and Notebooks used in Lecture for Sp19 COGS108

★ 5 22

Section_Workbooks

Workbooks for practice during discussion section

Jupyter Notebook 25

Tutorials

Tutorial notebooks for hands-on data science, following along with the course topics.

Jupyter Notebook ★ 38 117

Projects

Final Project materials and description.

Jupyter Notebook ★ 3 95

Readings

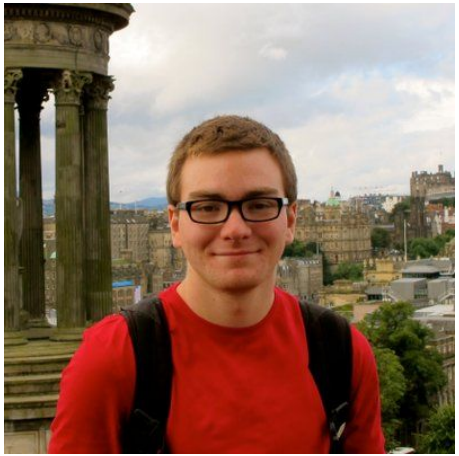
A curated list of suggested reading materials.

★ 6 2

https://github.com/COGS108

UCSD

Brad Voytek



Tom Donoghue

Instructional staffs, students,
& COGS faculty/staff