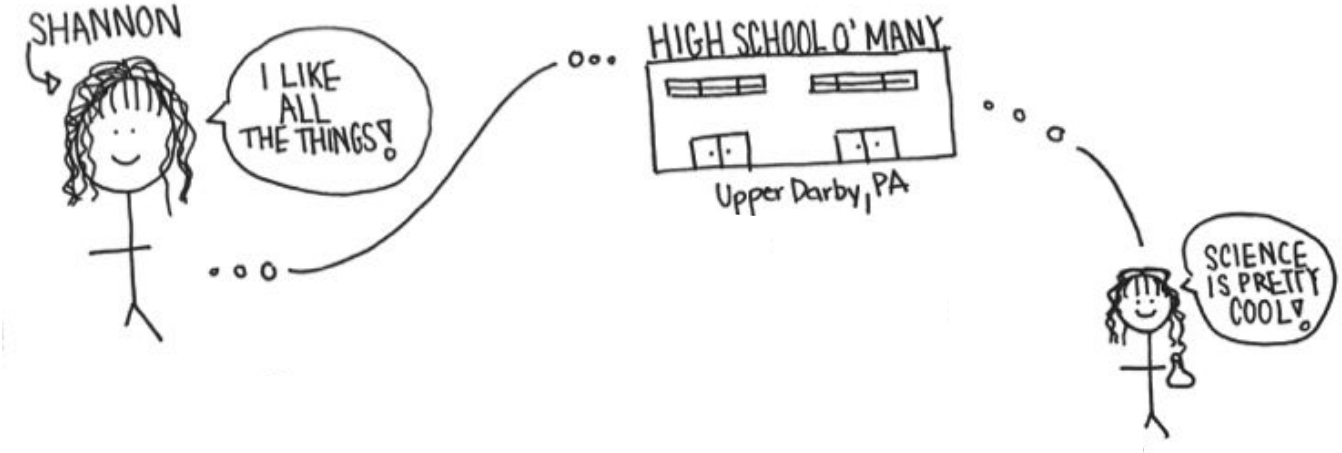
Improving the value of public data with *recount2* and phenotype prediction

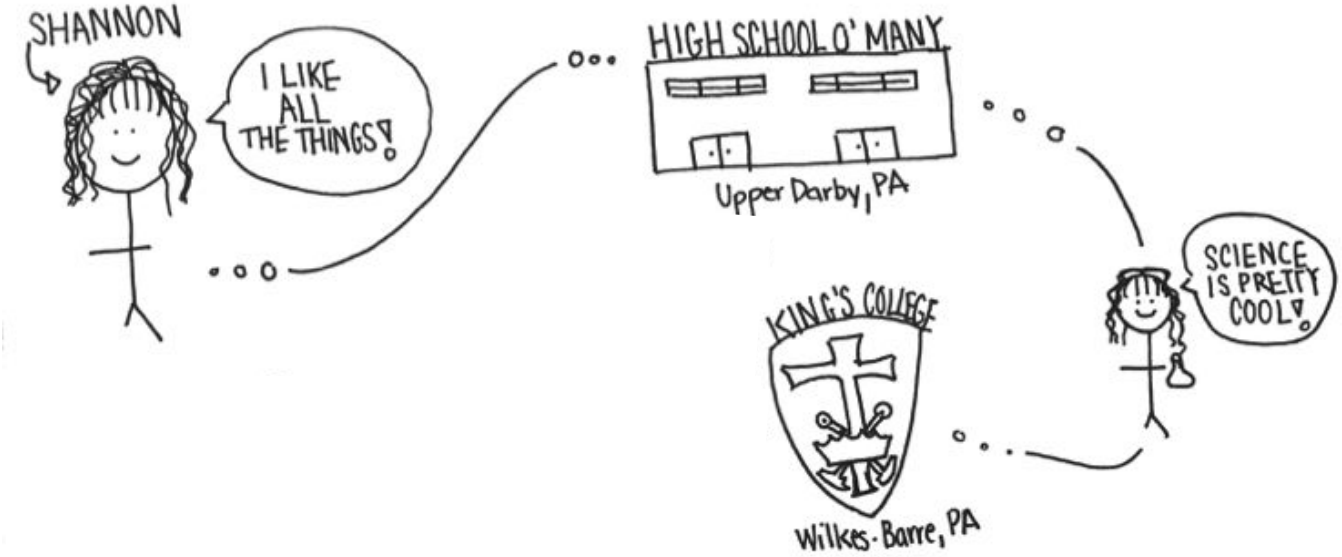
Shannon E. Ellis, PhD
Assistant Teaching Professor
Cognitive Science

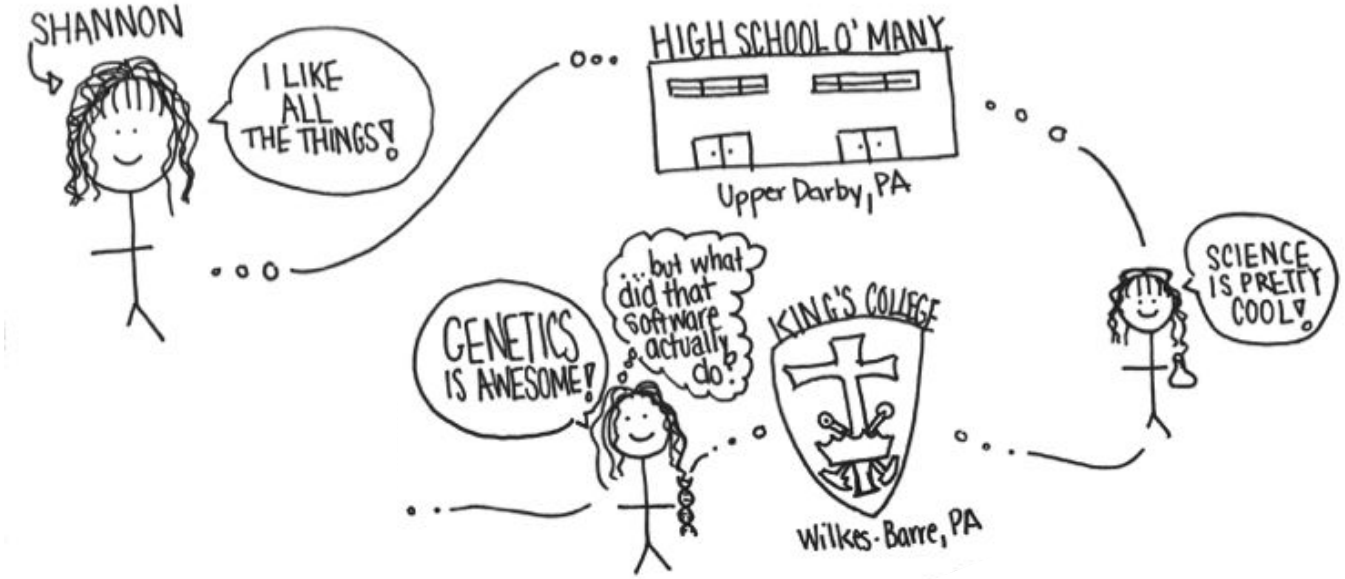
2020 02 10
shanellis.com/talks

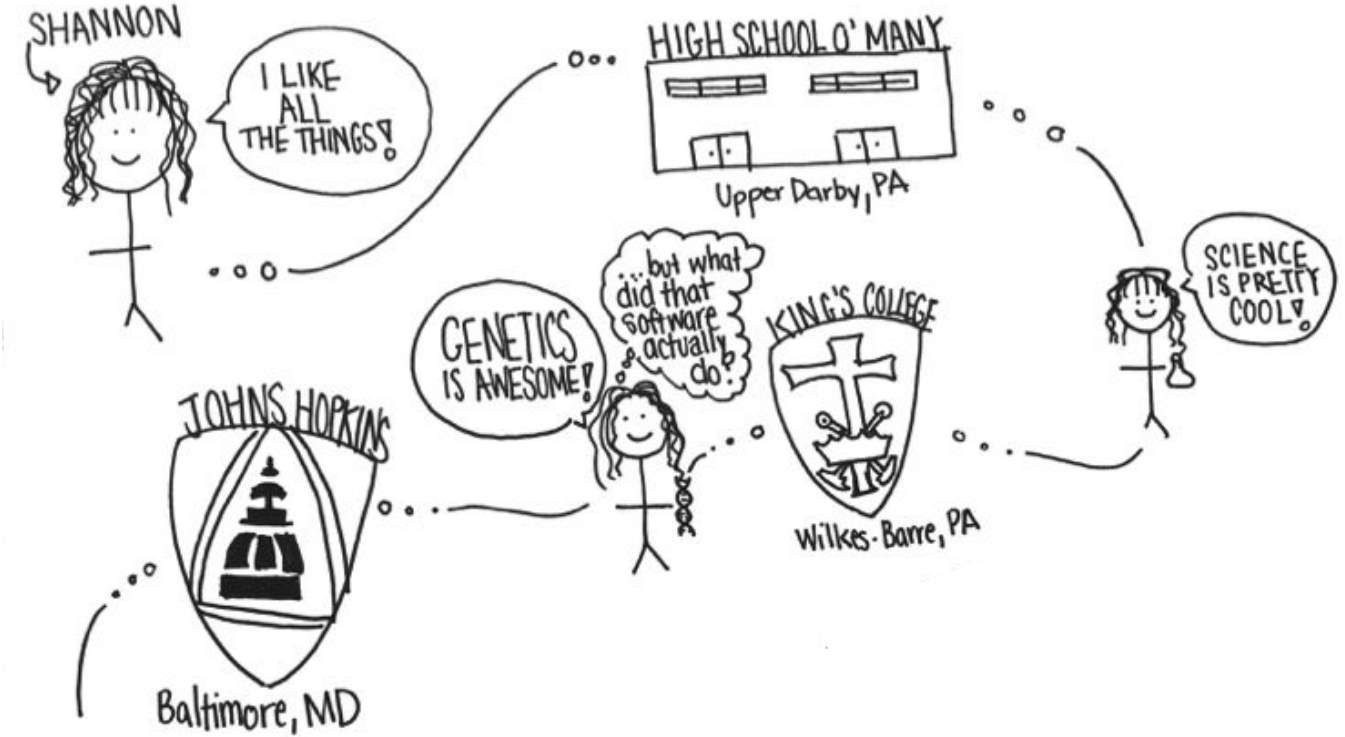

SHANNON
I LIKE ALL THE THINGS!

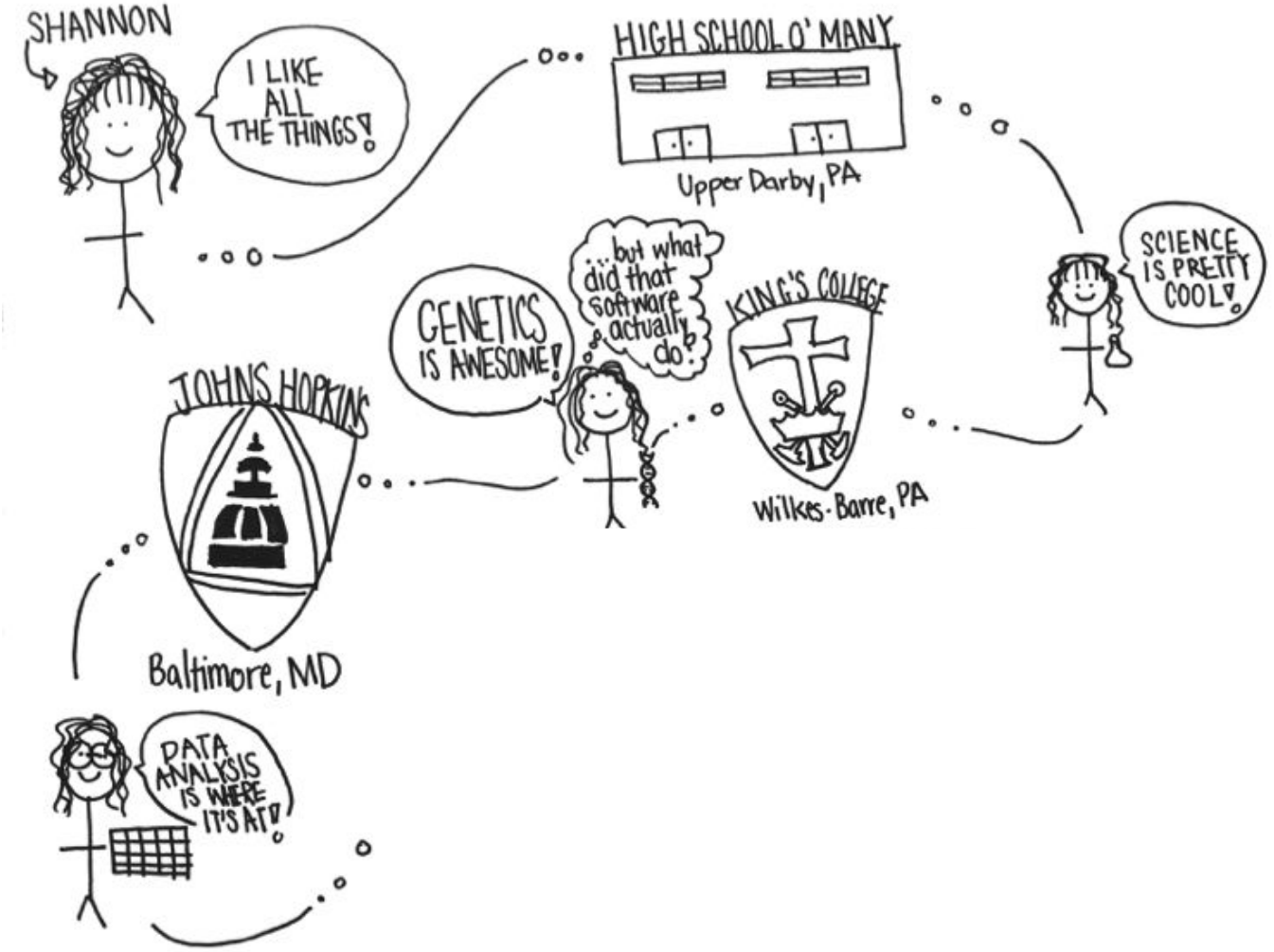

TODAY
DATA ANALYSIS
HUMAN GENETICS
TEACHING
OPEN SCIENCE

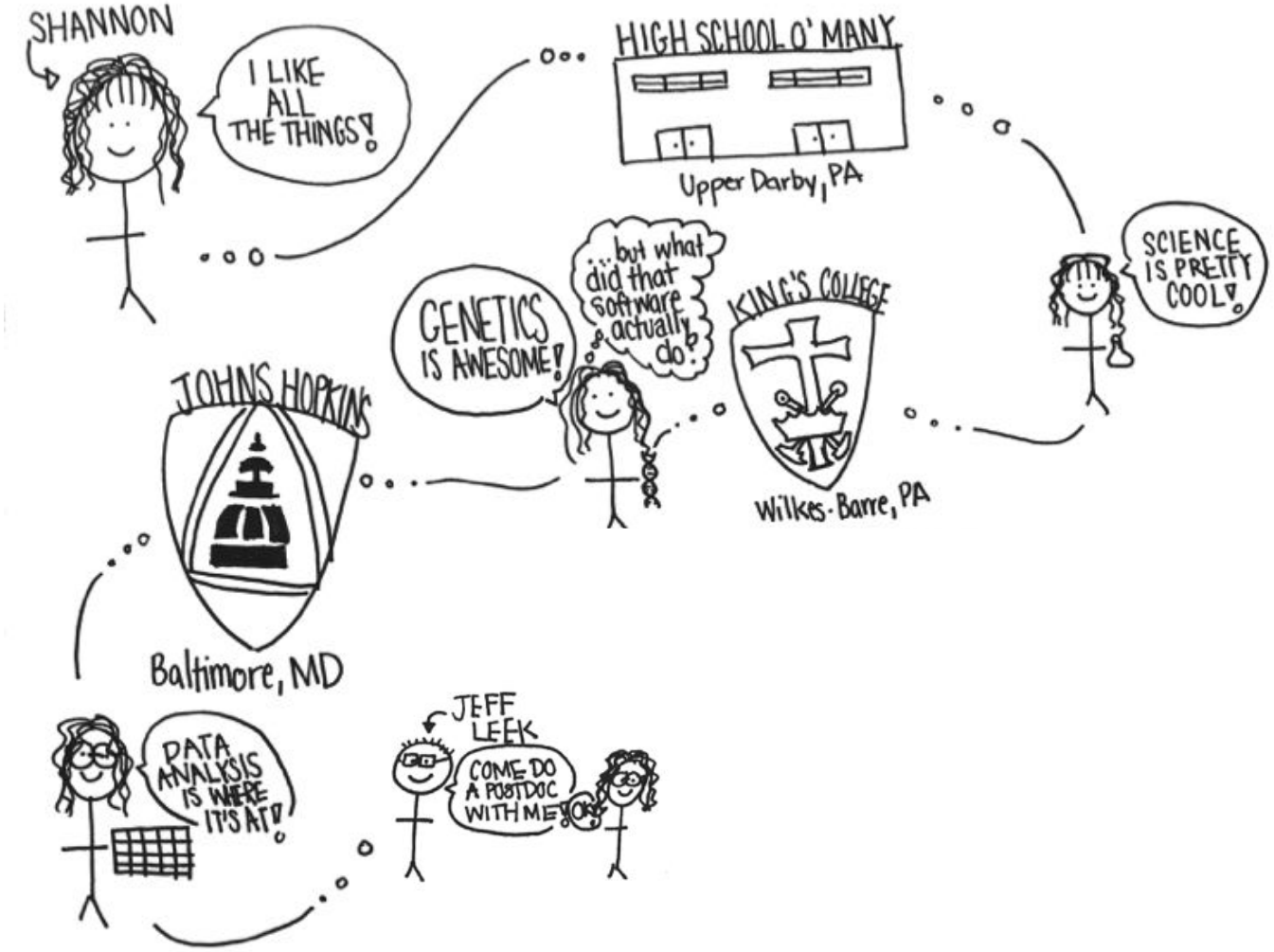

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!

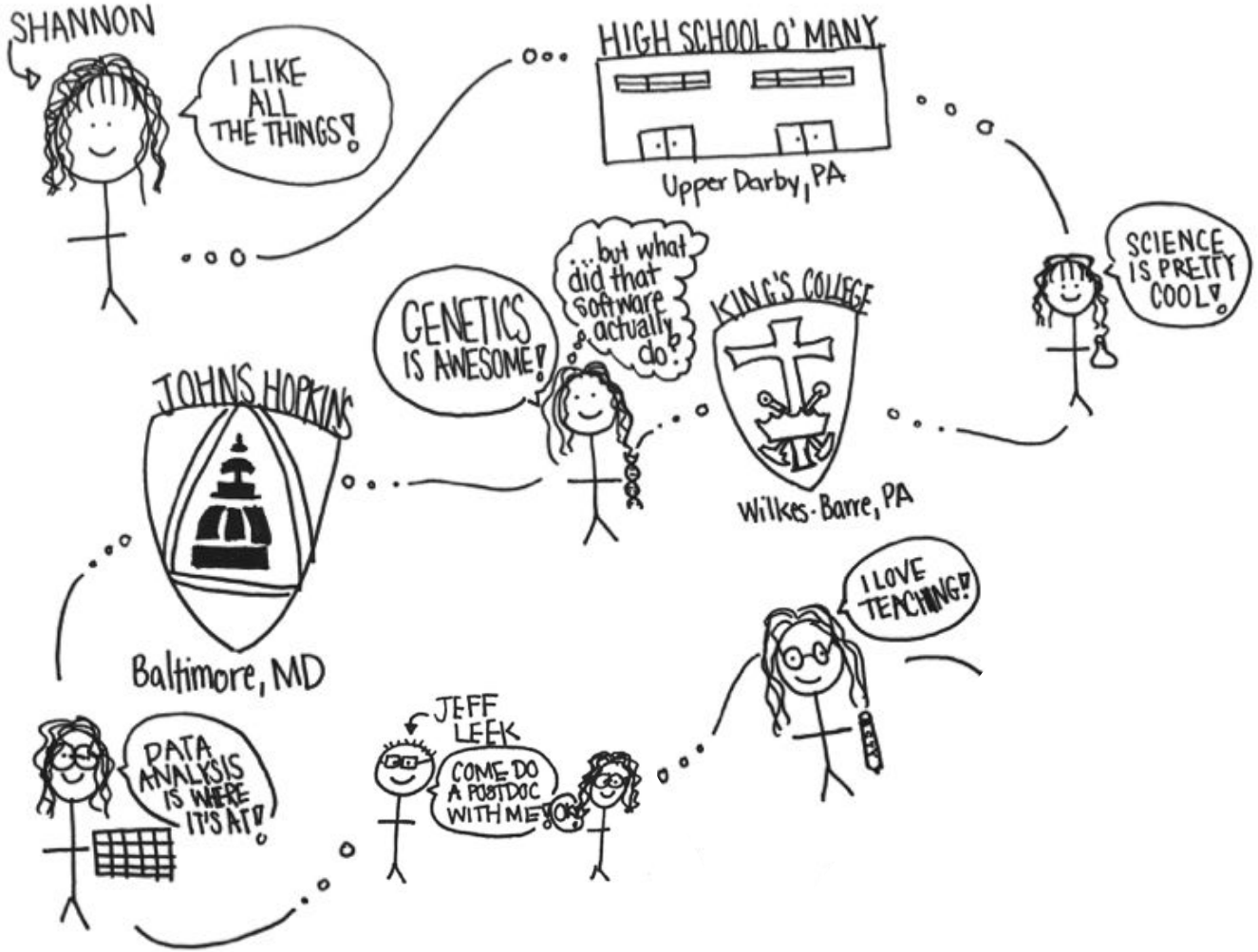

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
Wilkes-Barre, PA

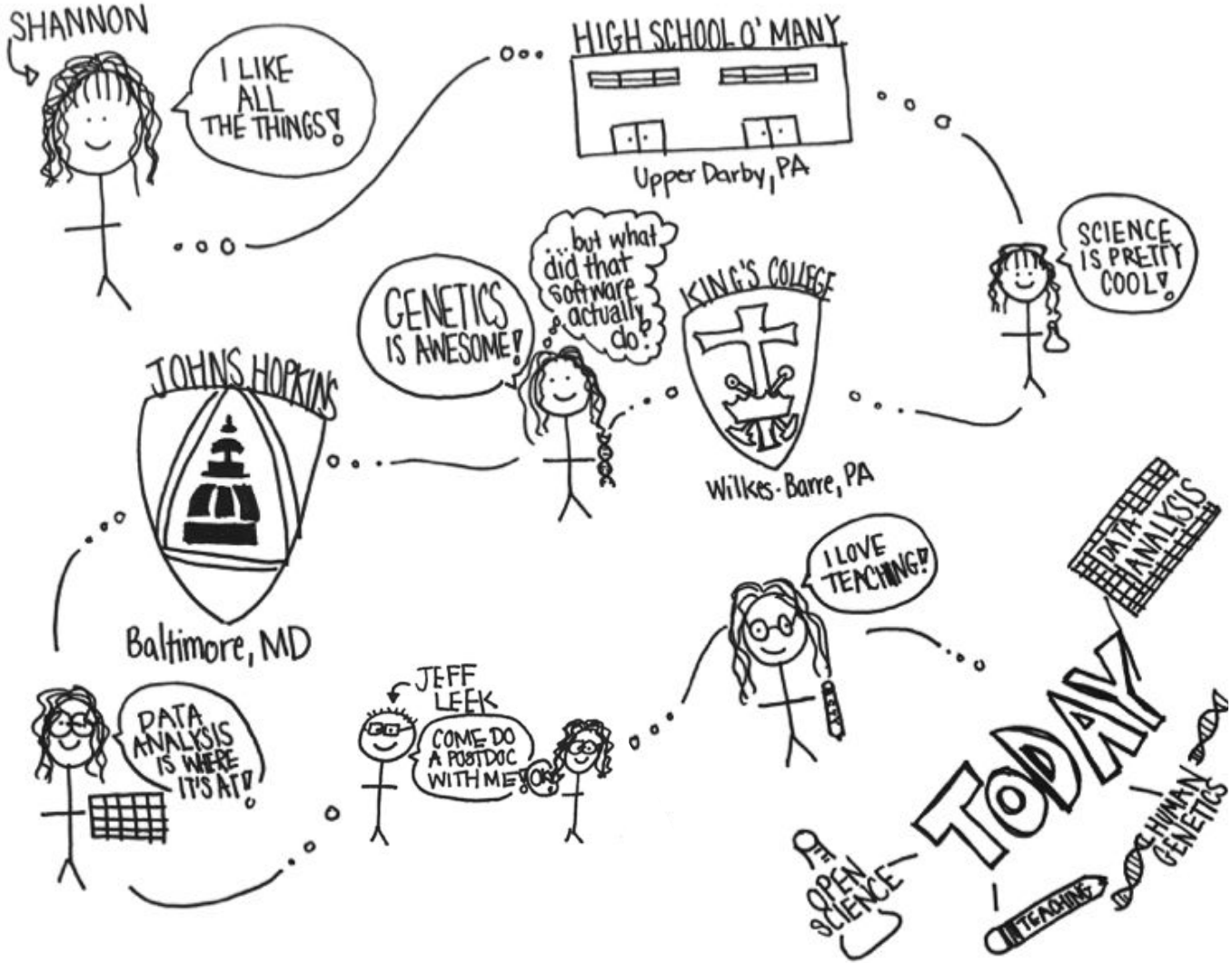

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
GENETICS IS AWESOME!
...but what did that software actually do?
Wilkes-Barre, PA


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

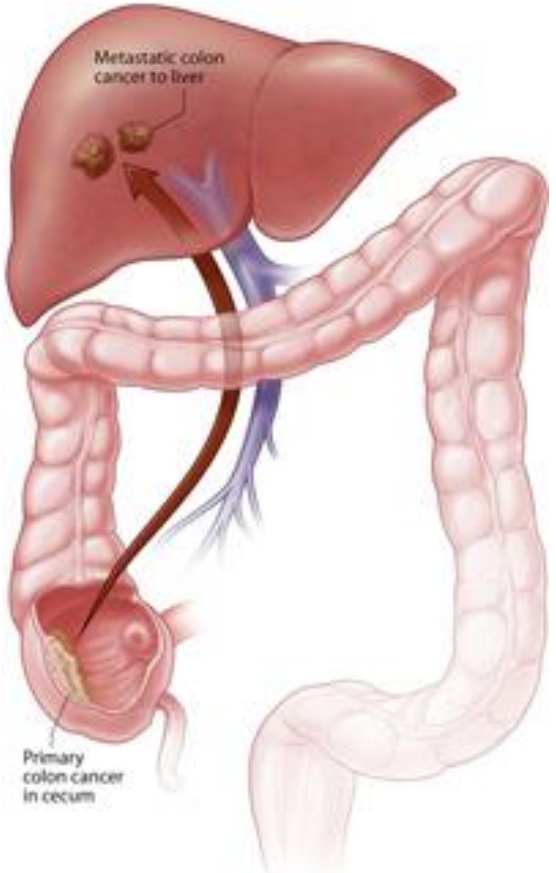

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
KING'S COLLEGE
...but what did that software actually do?
GENETICS IS AWESOME!
JOHNS HOPKINS
Wilkes-Barre, PA
Baltimore, MD
DATA ANALYSIS IS WHERE IT'S AT!


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!

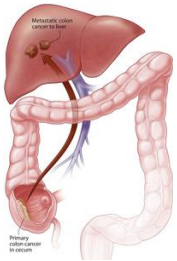

SHANNON
I LIKE ALL THE THINGS!
HIGH SCHOOL O' MANY
Upper Darby, PA
SCIENCE IS PRETTY COOL!
...but what did that software actually do?
GENETICS IS AWESOME!
KING'S COLLEGE
Wilkes-Barre, PA
JOHNS HOPKINS
Baltimore, MD
I LOVE TEACHING!
DATA ANALYSIS IS WHERE IT'S AT!
JEFF LEEK
COME DO A POSTDOC WITH ME!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


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

What makes primary cancer different than metastatic cancer?

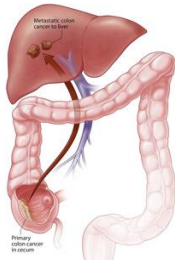


www.hopkinsmedicine.org



What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

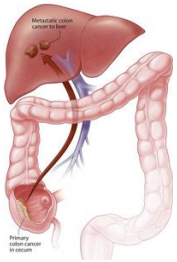


What makes primary cancer different than metastatic cancer?

Collect patient samples and information

~6 months

Find a researcher with access to patient samples



What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

~6 months

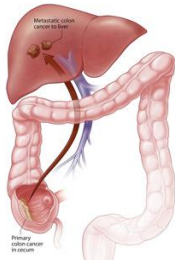
Collect patient samples and information

Extract DNA/RNA from samples

1-2 wks

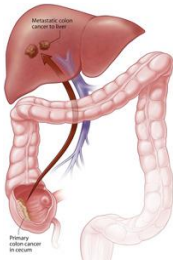
Sequence samples

2-4 wks



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

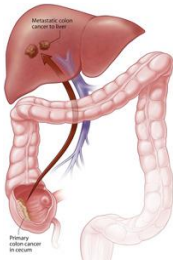


What makes primary cancer different than metastatic cancer?

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.



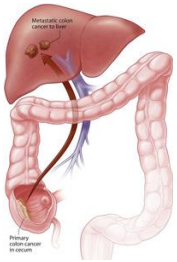
What makes primary cancer different than metastatic cancer?

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 easily accessible and well-annotated.



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

Analyze data and answer biological question

1 month - 1+ years

Total: 1+ years

Project Goal: Take publicly available RNA-Seq data and make it available and easy-to-use

- ☐ Comprehensive
- ☐ Easy to Get
- ☐ Useful for future study

Genetics101

The Central Dogma of Genetics

DNA

```
ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT
TGACTGGATCTAGTCAGCTAGCTAGCATATGCTAATGTTTTAGTAGCCGTA
```

slide adapted from jeff leek

The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

slide adapted from jeff leek

The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

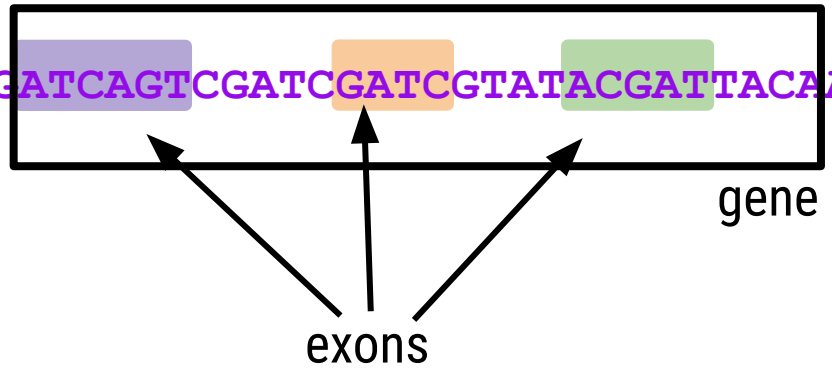
gene

slide adapted from jeff leek

The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT



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

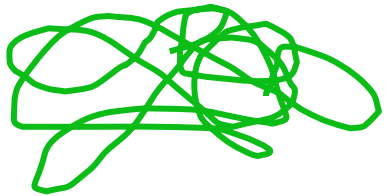
↓ **transcription**

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 				

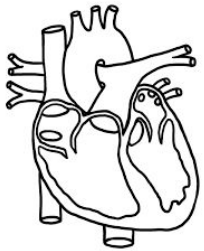
Two copies of **DNA** -> many **transcripts** -> many **proteins**

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

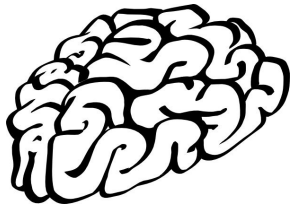
Two copies of **DNA** -> many **transcripts** -> many **proteins**

	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



!=

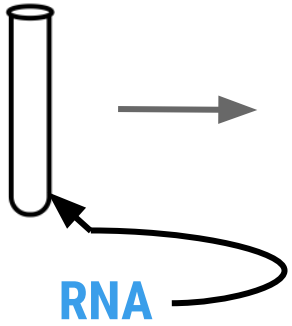


Measuring RNA levels

slide adapted from jeff leek

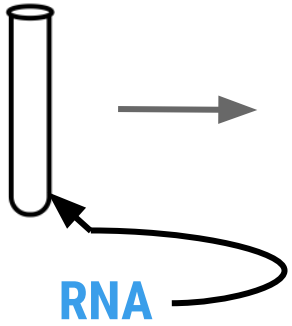
Next Generation Sequencing (NGS) Has Completely Revolutionized How We Study Genetics

Next Generation Sequencing (NGS) in one slide

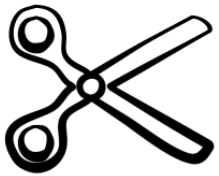


Step 1: Extract RNA to get sample of interest

Next Generation Sequencing (NGS) in one slide

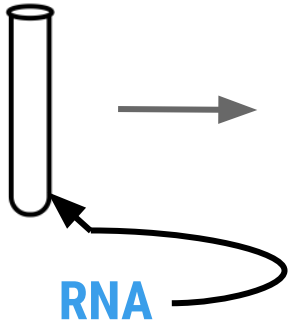


Step 1: Extract RNA to get sample of interest

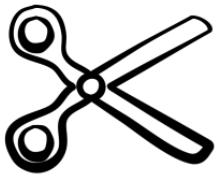


Step 2: Chop up RNA into smaller pieces

Next Generation Sequencing (NGS) in one slide



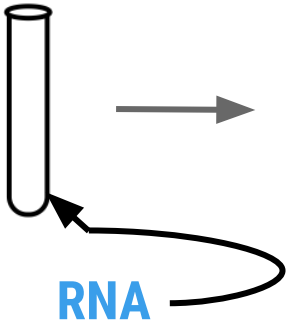
Step 1: Extract RNA to get sample of interest



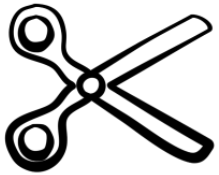
Step 2: Chop up RNA into smaller pieces

Step 3: Sequence the sample

Next Generation Sequencing (NGS) in one slide

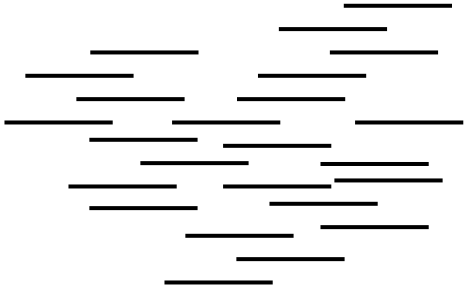


Step 1: Extract RNA to get sample of interest



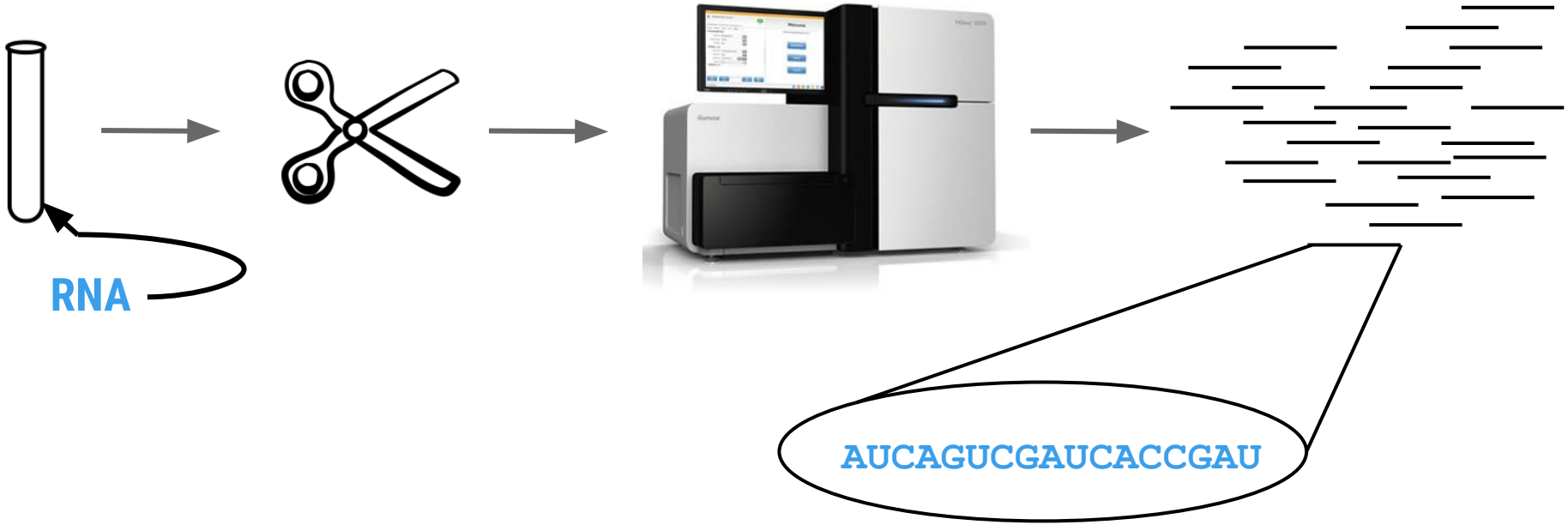
Step 2: Chop up RNA into smaller pieces

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 the RNA in that read

```
@22:16362385-16362561W:ENST00000440999:2:177:-40:244:S/2
CCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCATCCTGGTGGAACTCTC
CTGTGAGGACAGCCA
+
GGFF<BB=>GBGIIIIIIIIIIIIIIEGEHGHHIIIIIIIIHFHBB2/:=??EGGGEGFHH
IHHEDBD?@@DDHHD
@22:16362385-16362561W:ENST00000440999:3:177:-56:294:S/2
GCGTGAGCCACAGGGCCCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCA
TCCTGGTGGAACTCT
+
@=ABBBBIIIIIIIIHHGGGGIIDBDIIIIIIGIIIIHIIIIHFDD@BBDBGGFIDEE8DC
C/29>BGFCGHHHGF
@22:16362385-16362561W:ENST00000440999:4:177:137:254:S/1
TCACCATCCTGGTGGAACTCTCCTGTGAGGACAGCCAAGGCCTGAACTACCTGCaGTGGGG
AGCACCTCAGGGTTT
+
DDGBBCGGGIGGGBDDDHIIGGDGD77=BDIIIIIIIFHHH[illegible]>DE(
HHIFDDHH8@BEDDI
@22:16362385-16362561W:ENST00000440999:5:177:[illegible]:S/2
AGGGTTTGCCCAGGCAACCAGCCAGCCCTGGTCCAAGGCATCCTGGAGCGAGTTGTGGATG
GCAAAAAGACNCGCC
+
HIGHIHFHEGE4111:.;8@?@HDIIIIIIIEGGIHHIIGA?=:FIIIDD8.02506A8=
```

40M+

slide adapted from jeff leek

Sequence Identifier

Sequence

Quality scores

```
@22:16362385-16362561W:ENST00000440999:2:177:-40:244:S/2
CCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCATCCTGGTGGAACTCTC
CTGTGAGGACAGCCA
+
GGFF<BB=>GBGIIIIIIIIIIIIIIEGEHGHHIIIIIIIIHFHBB2/:=??EGGGEGFHH
IHHEDBD?@@DDHHD
@22:16362385-16362561W:ENST00000440999:3:177:-56:294:S/2
GCGTGAGCCACAGGGCCCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCA
TCCTGGTGGAACTCT
+
@=ABBBBIIIIIIIIHHGGGGIIDBDIIIIIIGIIIIHIIIIHFDD@BBDBGGFIDEE8DC
C/29>BGFCGHHHGF
@22:16362385-16362561W:ENST00000440999:4:177:137:254:S/1
TCACCATCCTGGTGGAACTCTCCTGTGAGGACAGCCAAGGCCTGAACTACCTGCaGTGGGG
AGCACCTCAGGGTTT
+
DDGBBCGGGIGGGBDDDHIIGGDGD77=BDIIIIIIIFHHH[illegible]>DE(
HHIFDDHH8@BEDDI
@22:16362385-16362561W:ENST00000440999:5:1[illegible]:S/2
AGGGTTTGCCCAGGCAACCAGCCAGCCCTGGTCCAAGGCATCCTGGAGCGAGTTGTGGATG
GCAAAAAGACNCGCC
+
HIGHIHFHEGE4111:.;8@?@HDIIIIIIIEGGIHHIIGA?=:FIIIDD8.02506A8=
```

40M+

slide adapted from jeff leek

We’ve got 40M+ reads.
What does that all mean?

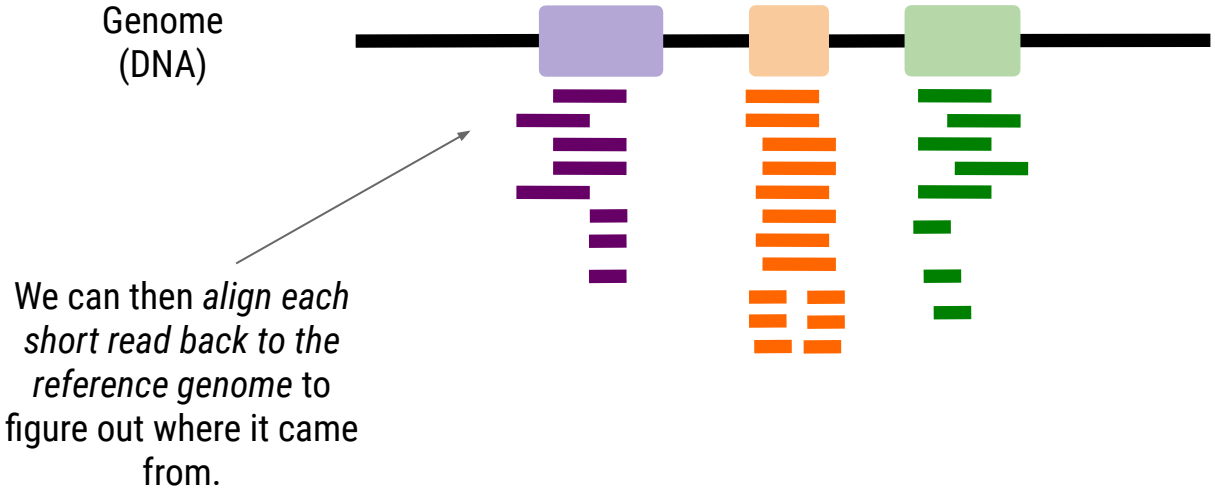
We first need to align these reads back to the genome

Genome (DNA)

The Human Genome Project (HGP) determined the *reference sequence* of the human genome in 2001.

slide adapted from jeff leek

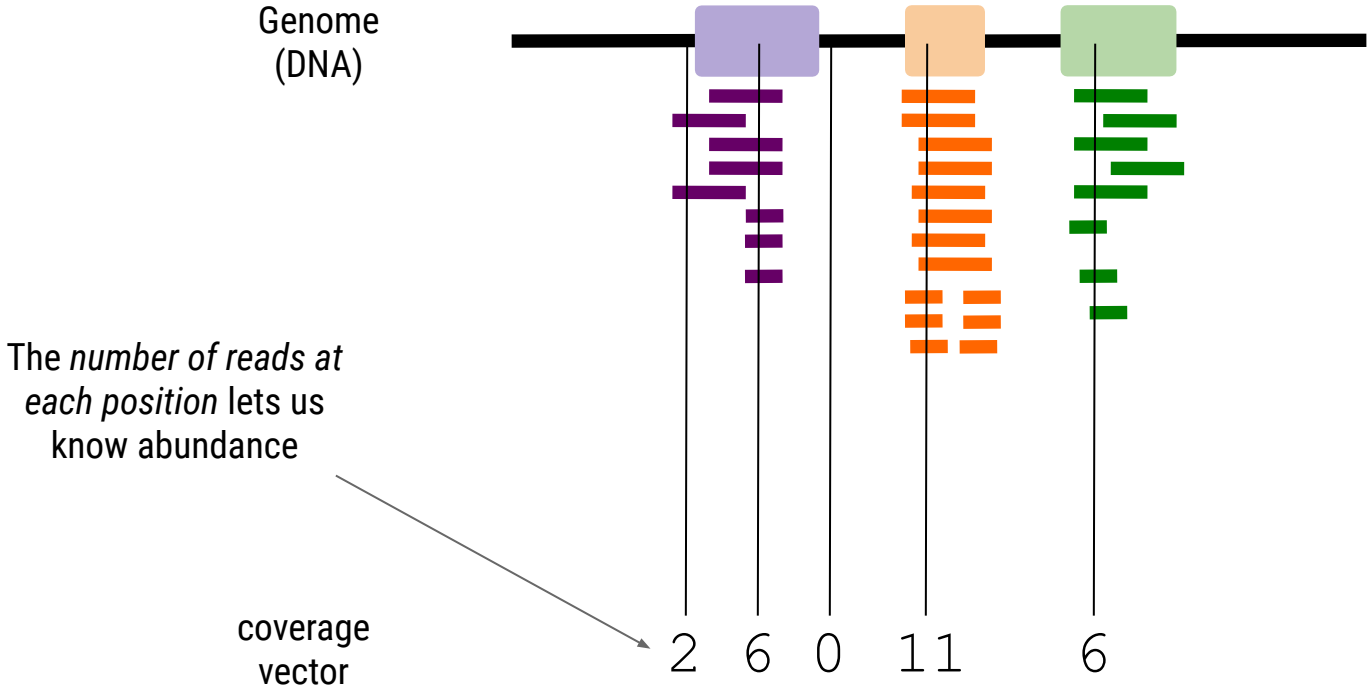
We first need to align these reads back to the genome



Nellore *et al.* (2016)
Bioinformatics
http://rail.bio/

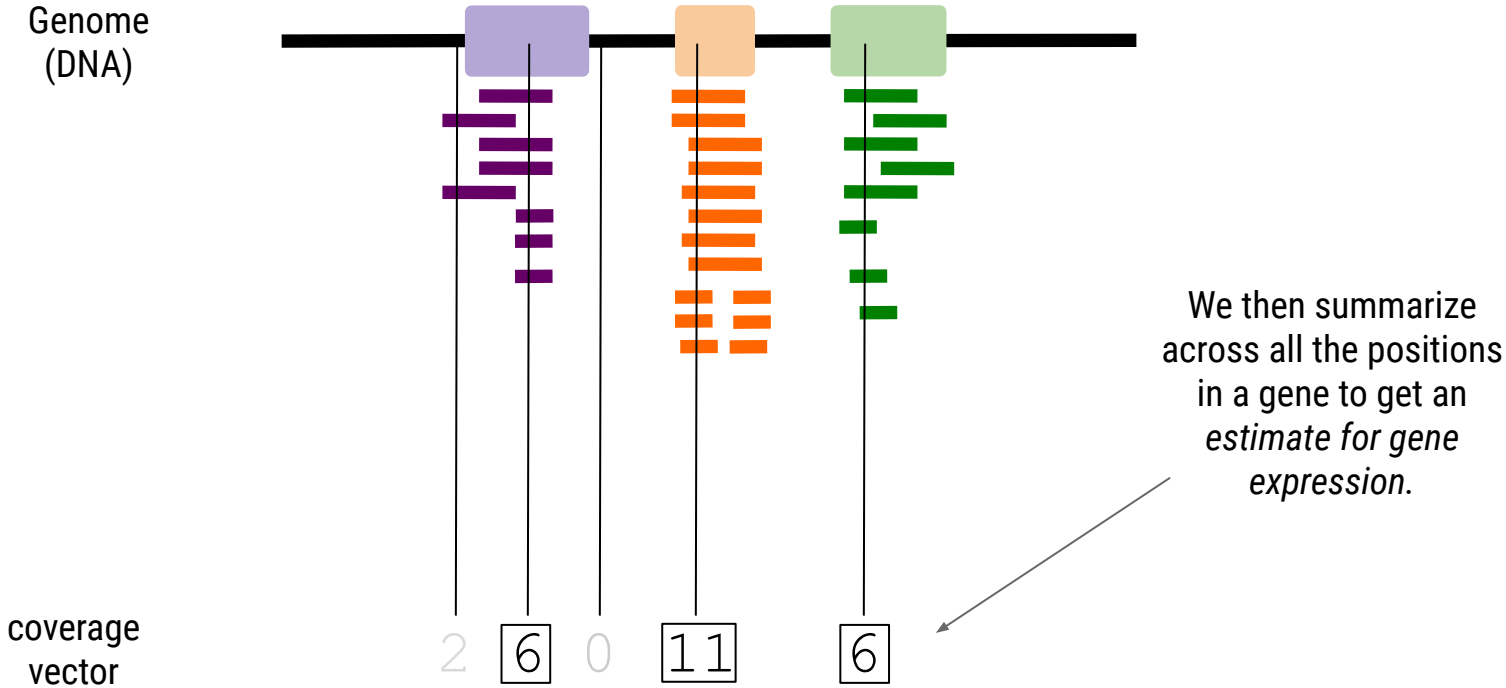
slide adapted from jeff leek

We first need to align these reads back to the genome



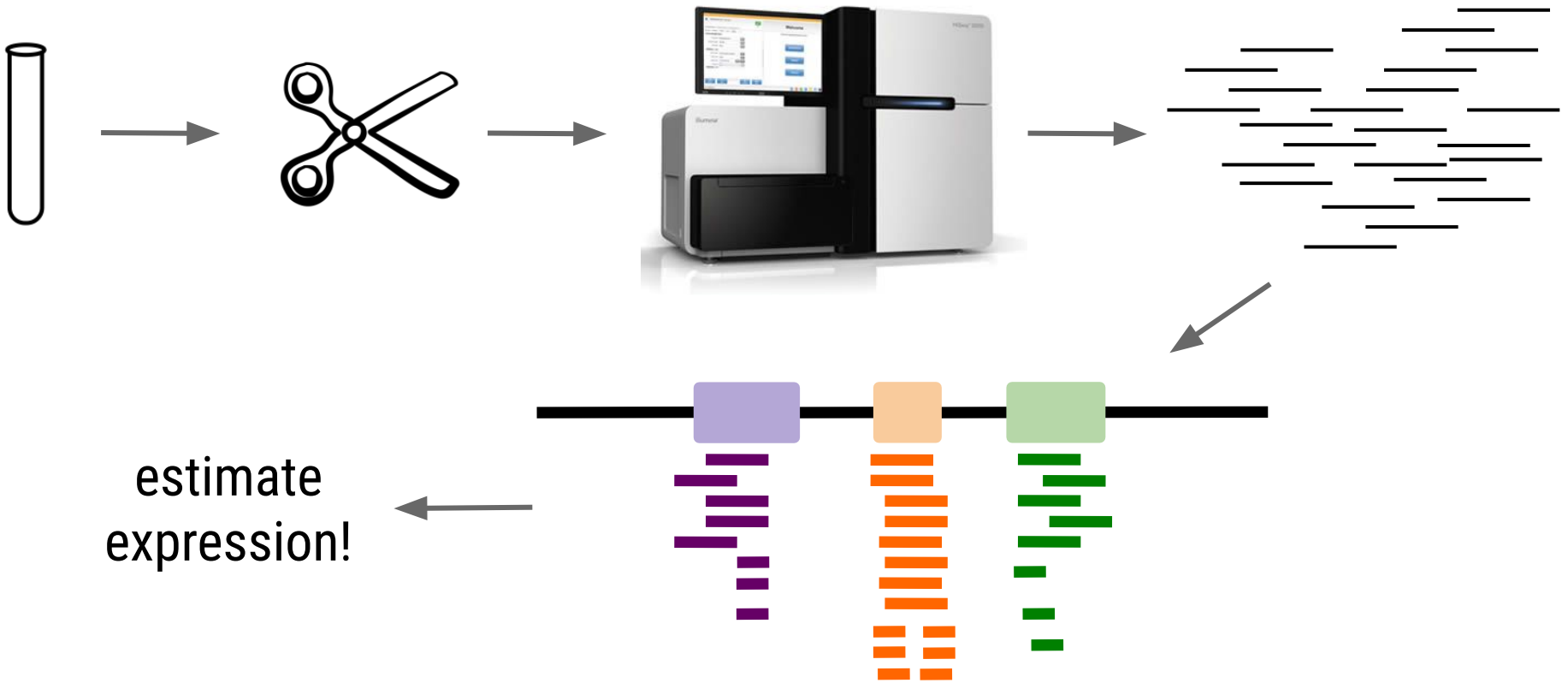
slide adapted from jeff leek

We first need to align these reads back to the genome

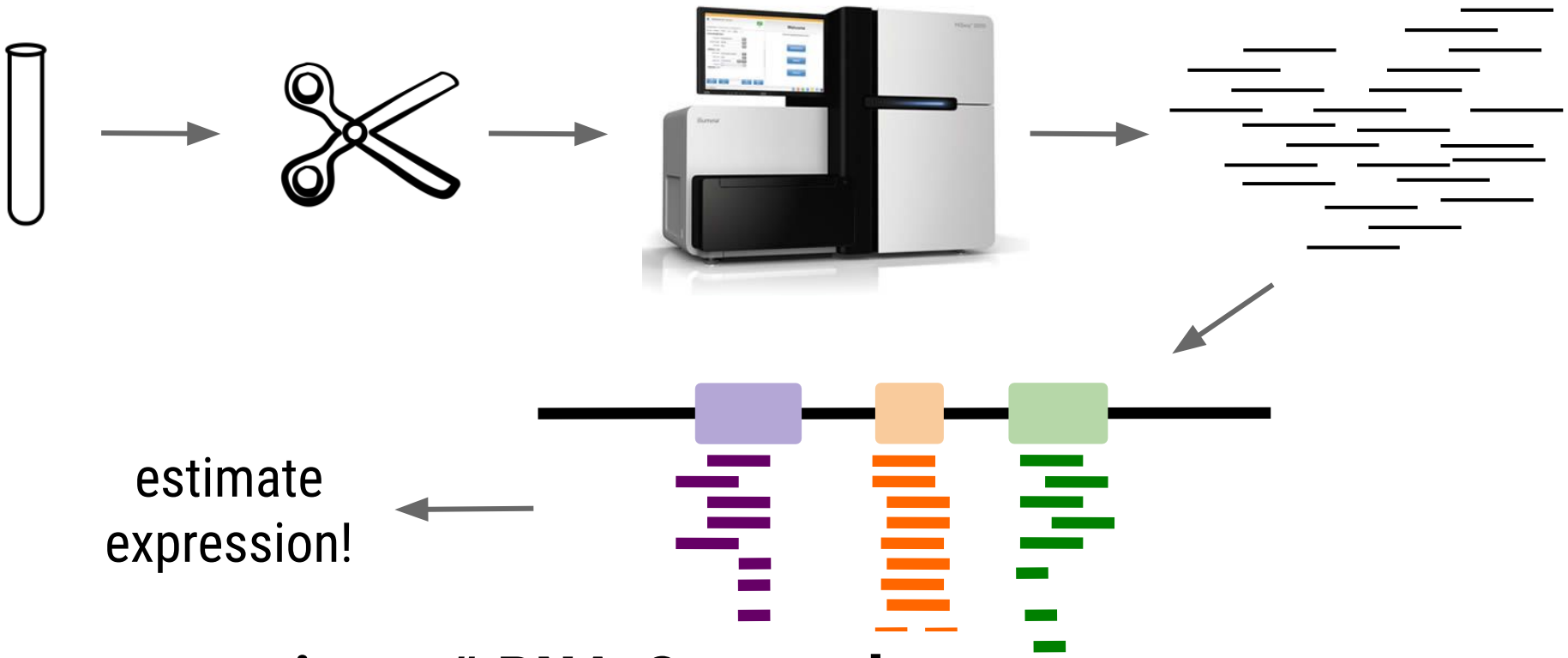


slide adapted from jeff leek

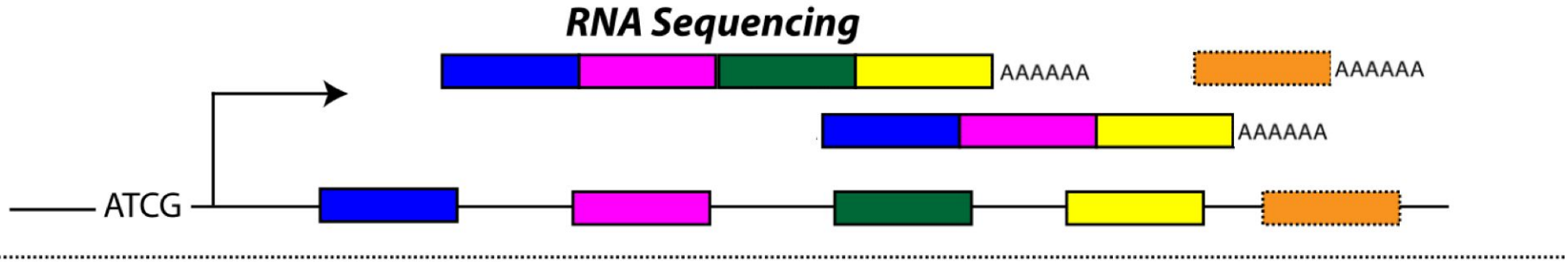
RNA-Seq = estimate expression across entire genome



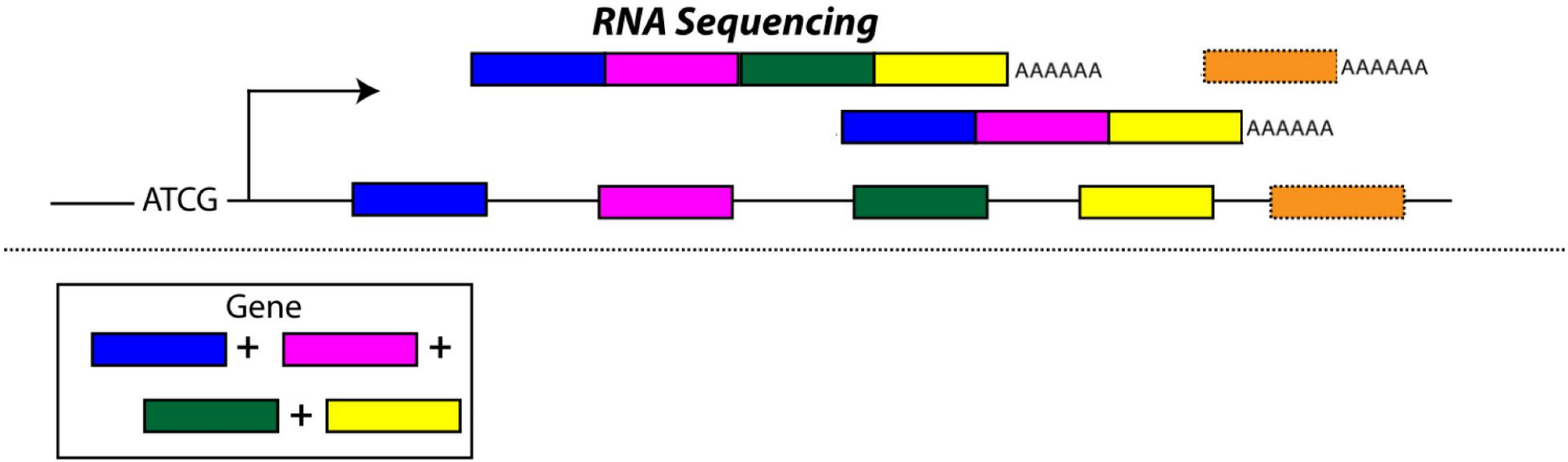
RNA-Seq = estimate expression across entire genome



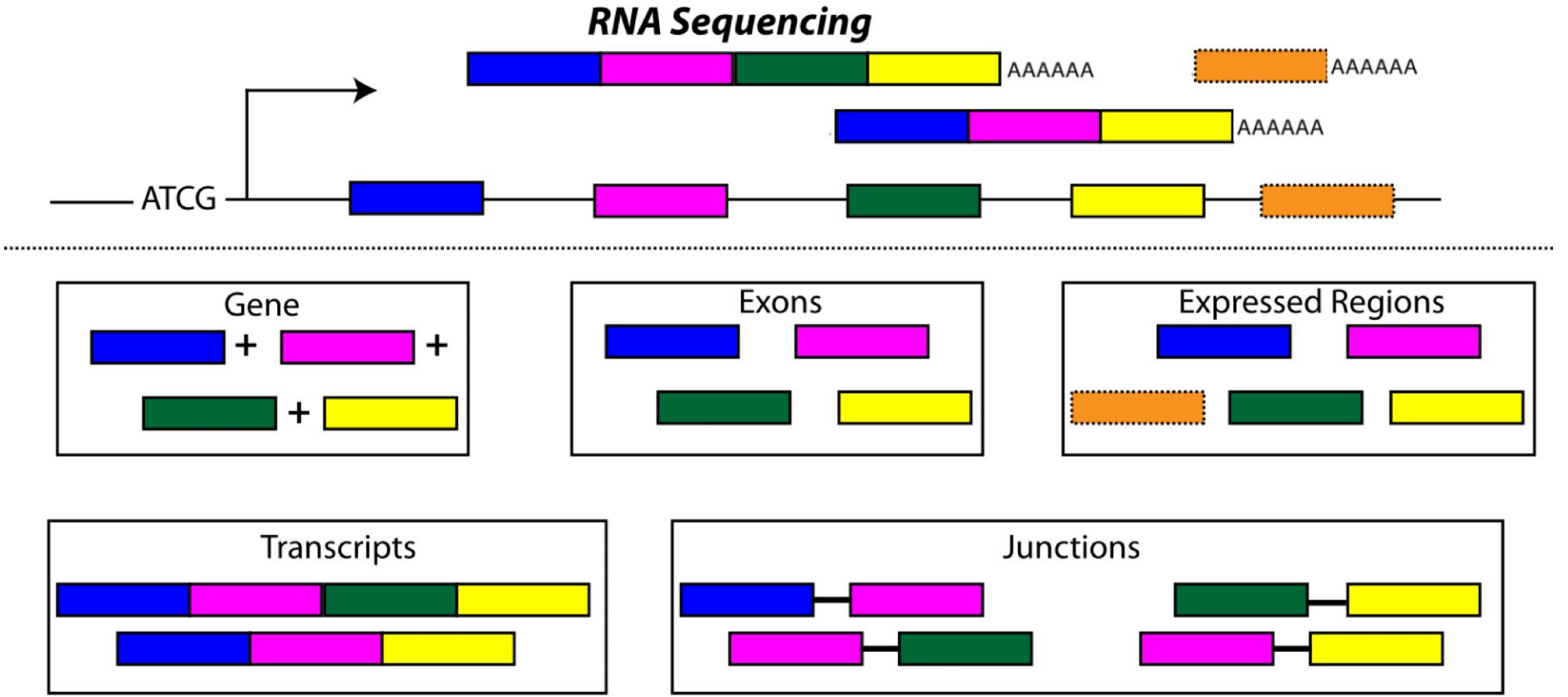
expression ≅ # RNA-Seq reads



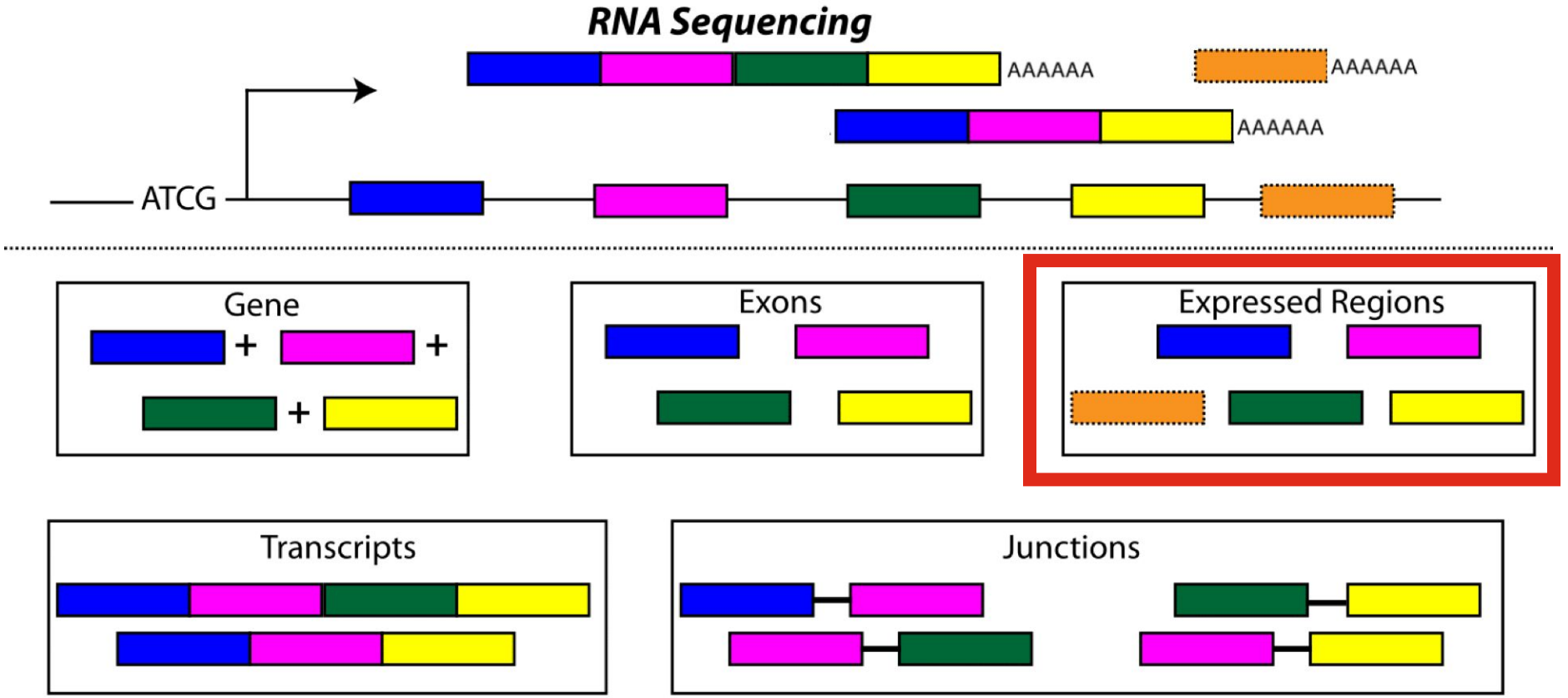
slide adapted from andrew jaffe



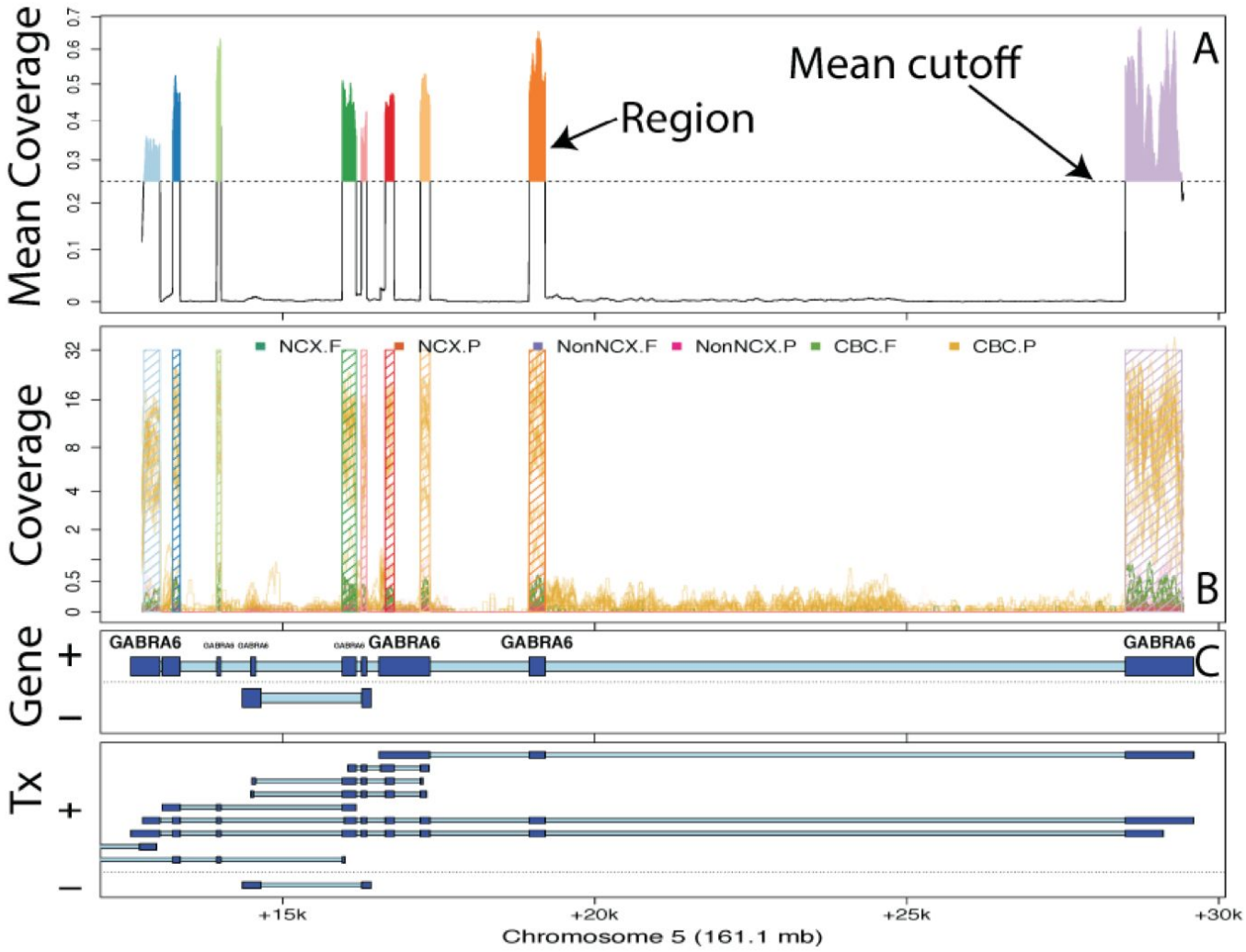
slide adapted from andrew jaffe



slide adapted from andrew jaffe



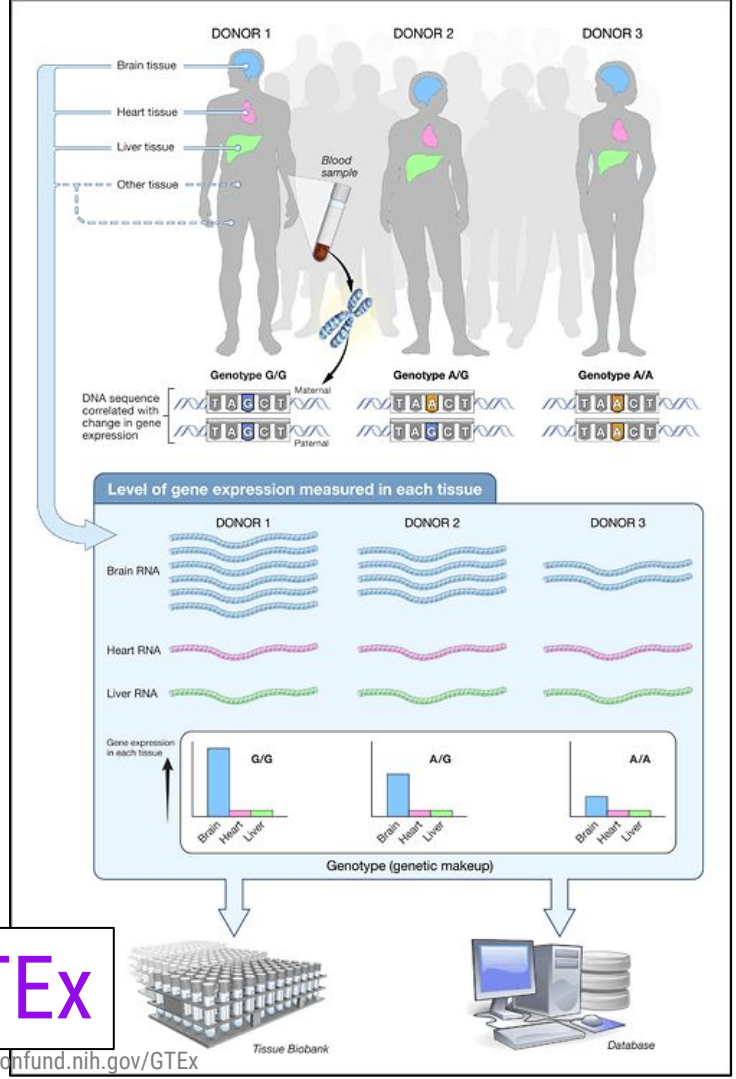
slide adapted from andrew jaffe



slide from jeff leek

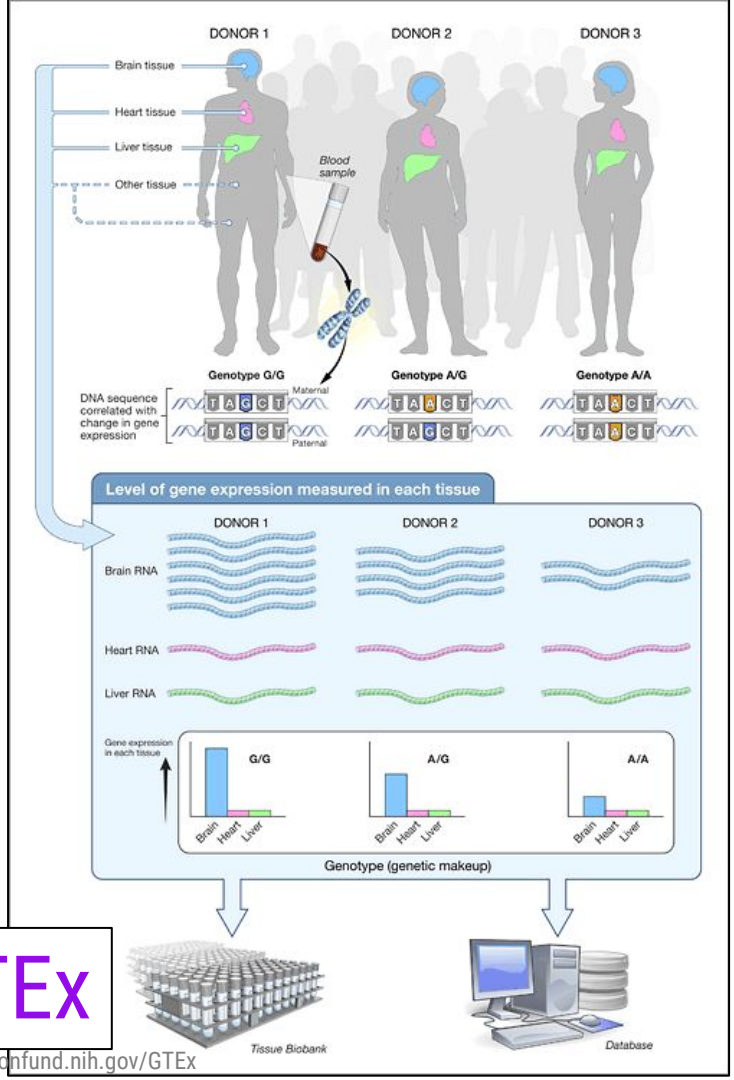
Project Goal: Take publicly available RNA-Seq data and make it available and easy-to-use

- ☐ Comprehensive
- ☐ Easy to Get
- ☐ Useful for future study



GTEx

https://commonfund.nih.gov/GTEx



GTEx

https://commonfund.nih.gov/GTEx

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES ...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

...using **7** DIFFERENT DATA TYPES

TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM

20 COLLABORATING INSTITUTIONS across the United States and Canada

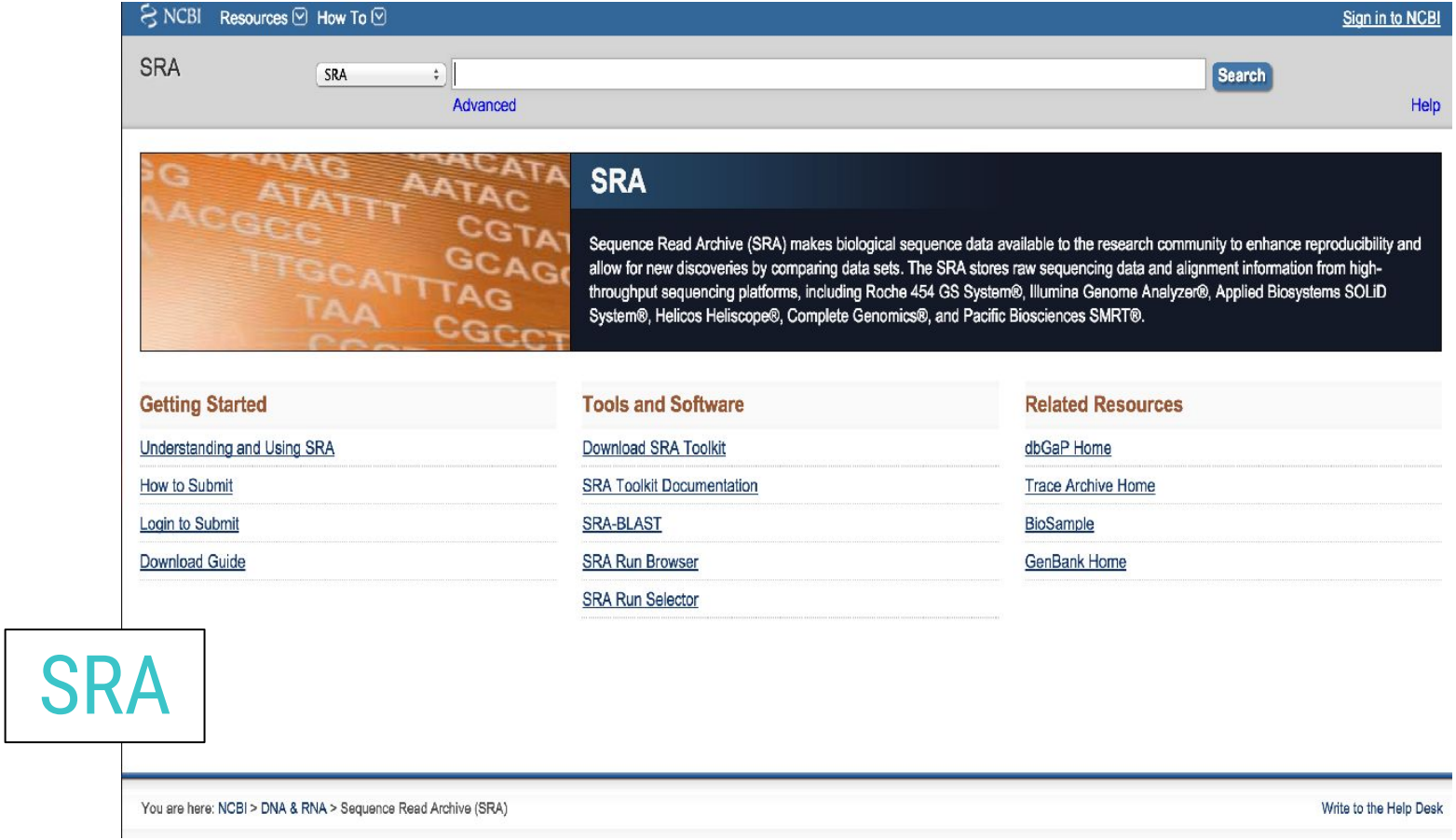
WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

...lysis of stomach cancer revealed that it is not a single disease, but a disease composed ...types, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

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

SRA

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

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

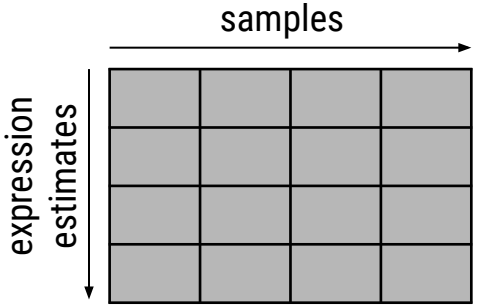
We'll take these ~70,000 samples, align each back to the reference genome, and then, for each sample, we'll estimate expression across the genome.

Project Goal: Take publicly available RNA-Seq data and make it available and easy-to-use

- [x] Comprehensive
- [ ] Easy to Get
- [ ] Useful for future study

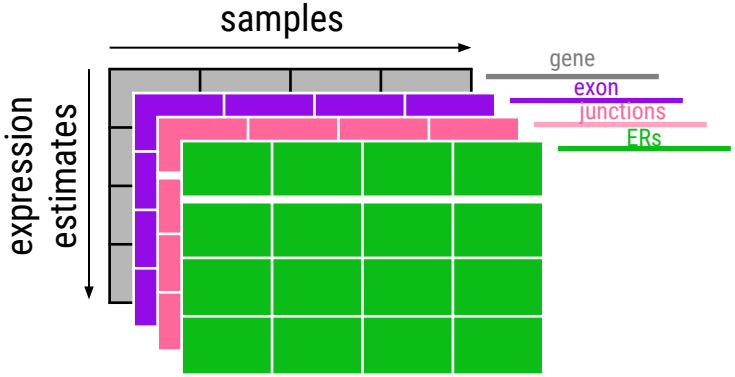
recount2

expression data for ~70,000 human samples



recount2

expression data for ~70,000 human samples



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

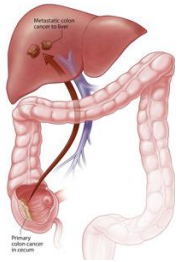
Collado-Torres L, **Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. recount: A large-scale resource of analysis-ready RNA-seq expression data. *bioRxiv* **068478**.

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



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/R... from sampl...

1-2 wks

Sequence samples

2-4 wks

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

Process sequencing data

3-6 months

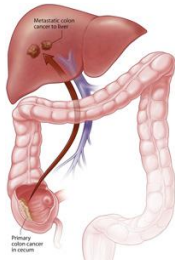
Data cleaning

1-3 months

Determine expression differences between primary cancer and metastatic cancer

1 month - 1+ years

Total: 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

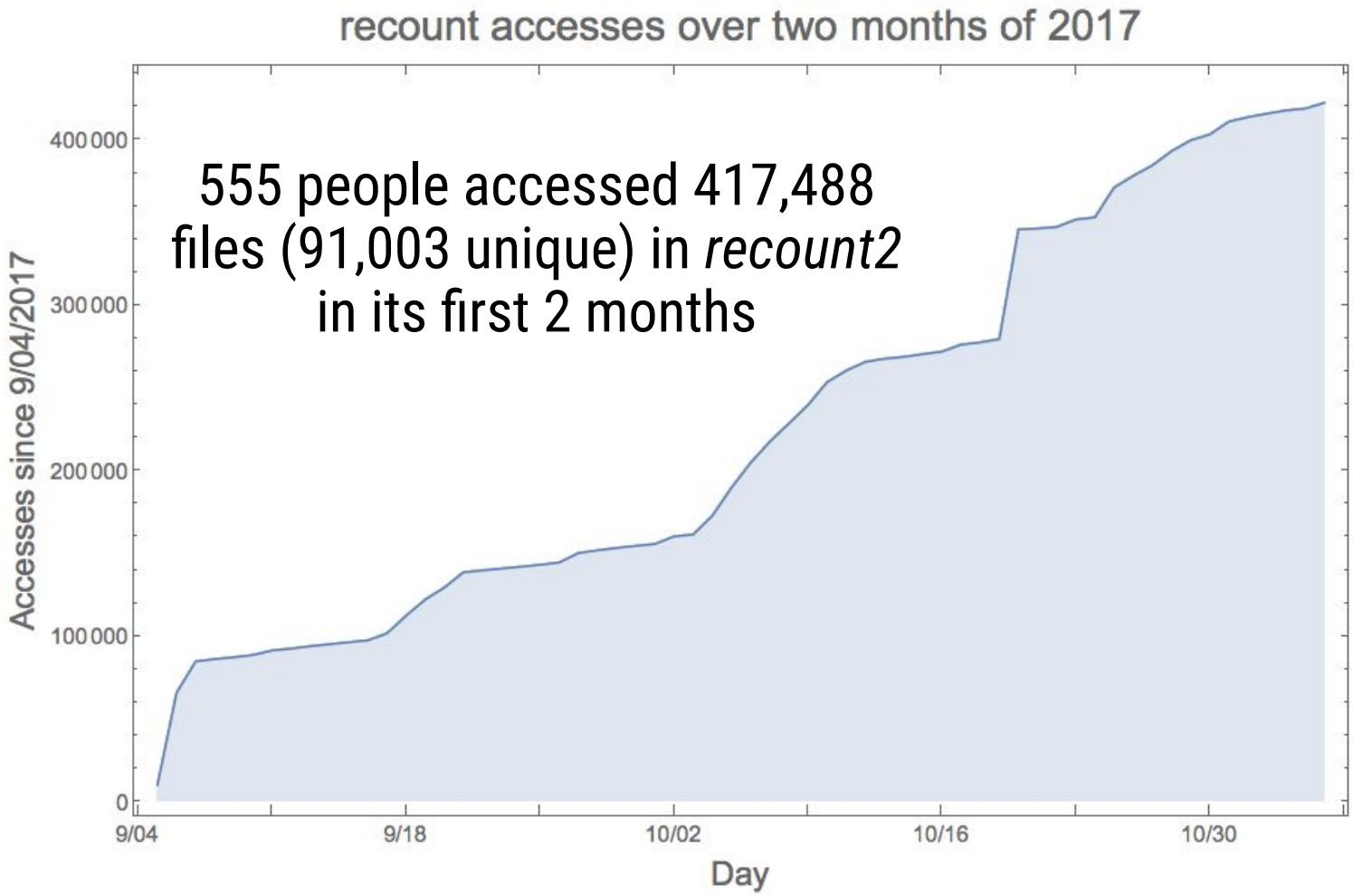
3~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

Find a researcher with access to patient samples

Extract DNA/RNA from samples

Data cleaning

Total: months

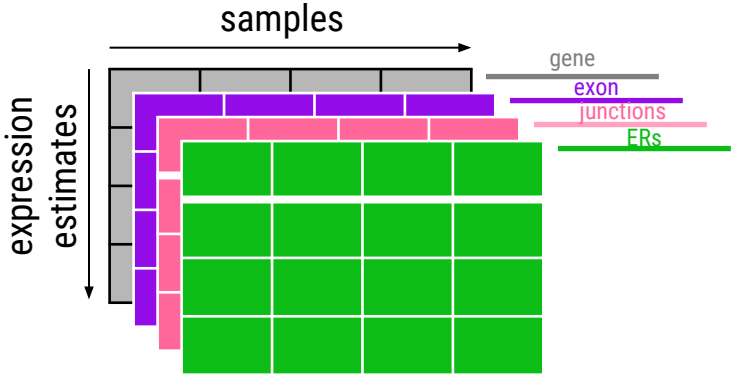

recount accesses over two months of 2017
555 people accessed 417,488 files (91,003 unique) in *recount2* in its first 2 months
Accesses since 9/04/2017
400 000
300 000
200 000
100 000
0
9/04
9/18
10/02
10/16
10/30
Day

Project Goal: Take publicly available RNA-Seq data and make it available and easy-to-use

- [x] Comprehensive
- [x] Easy to Get
- [ ] Useful for future study

recount2

expression data for ~70,000 human samples

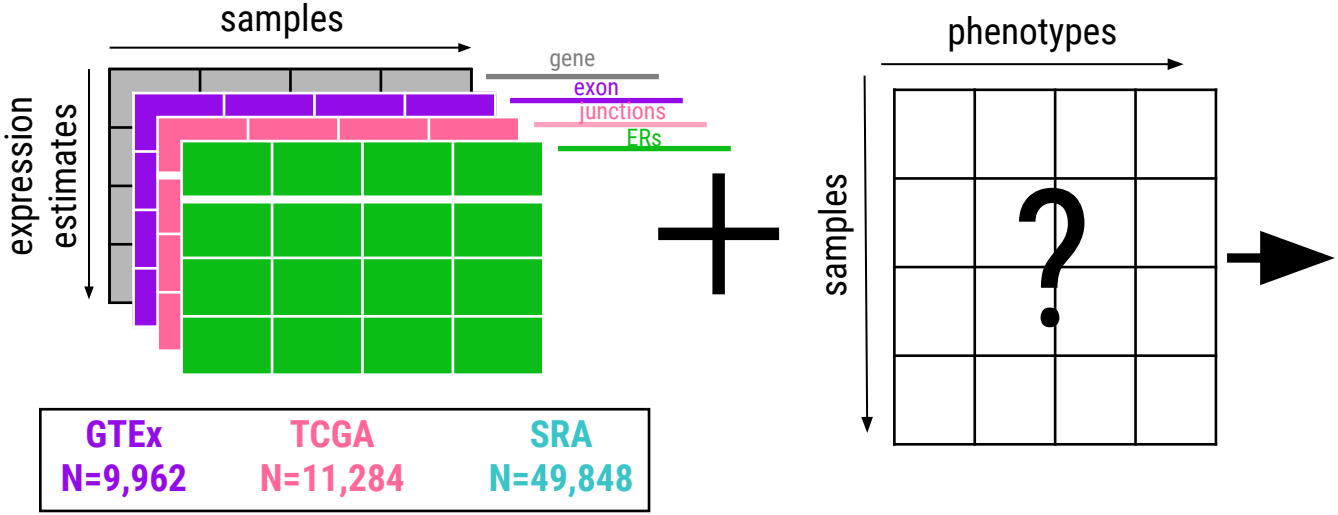


GTEx	TCGA	SRA
N=9,962	N=11,284	N=49,848

Answer meaningful questions about human biology and expression

recount2

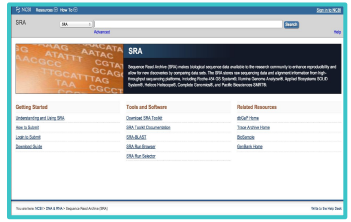
expression data for ~70,000 human samples



Answer meaningful questions about human biology and expression

SRA phenotype information is far from complete

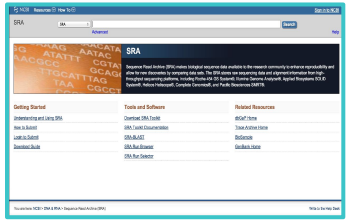
	Sex	Tissue	Race	Age
6620	female	liver	NA	NA
6621	female	liver	NA	NA
6622	female	liver	NA	NA
6623	female	liver	NA	NA
6624	female	liver	NA	NA
6625	male	liver	NA	NA
6626	male	liver	NA	NA
6627	male	liver	NA	NA
6628	male	liver	NA	NA
6629	male	liver	NA	NA
6630	male	liver	NA	NA
6631	NA	blood	NA	NA
6632	NA	blood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



SRA

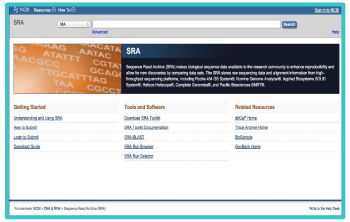
SRA phenotype information is far from complete

	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

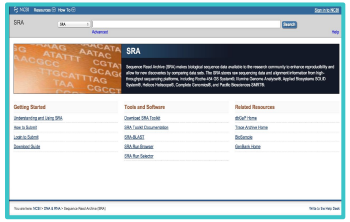
Even when information *is* provided, it's not always clear…



SRA

Category	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…

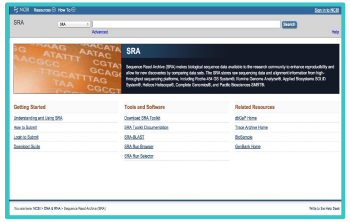


SRA

Category	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…



SRA

Category	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

in-silico Phenotyping

slide adapted from jeff leek

Goal:

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

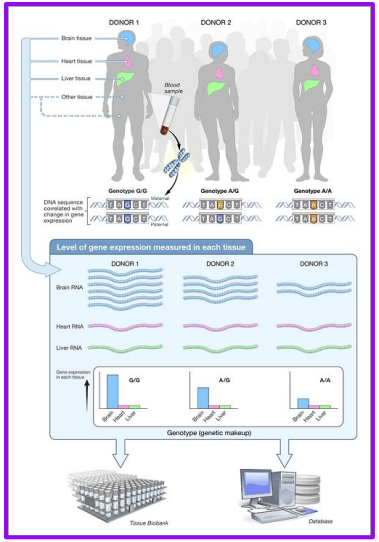
recount2

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

GTEx

Genotype Tissue Expression Project

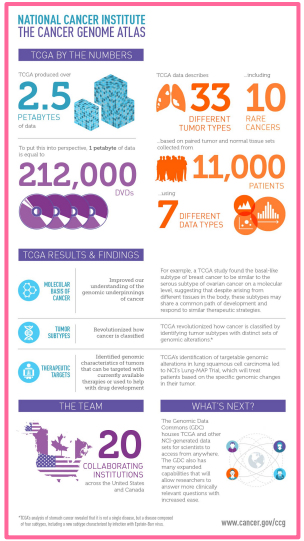
N=9,662



TCGA

The Cancer Genome Atlas

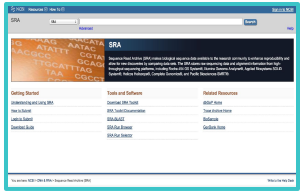
N=11,284



SRA

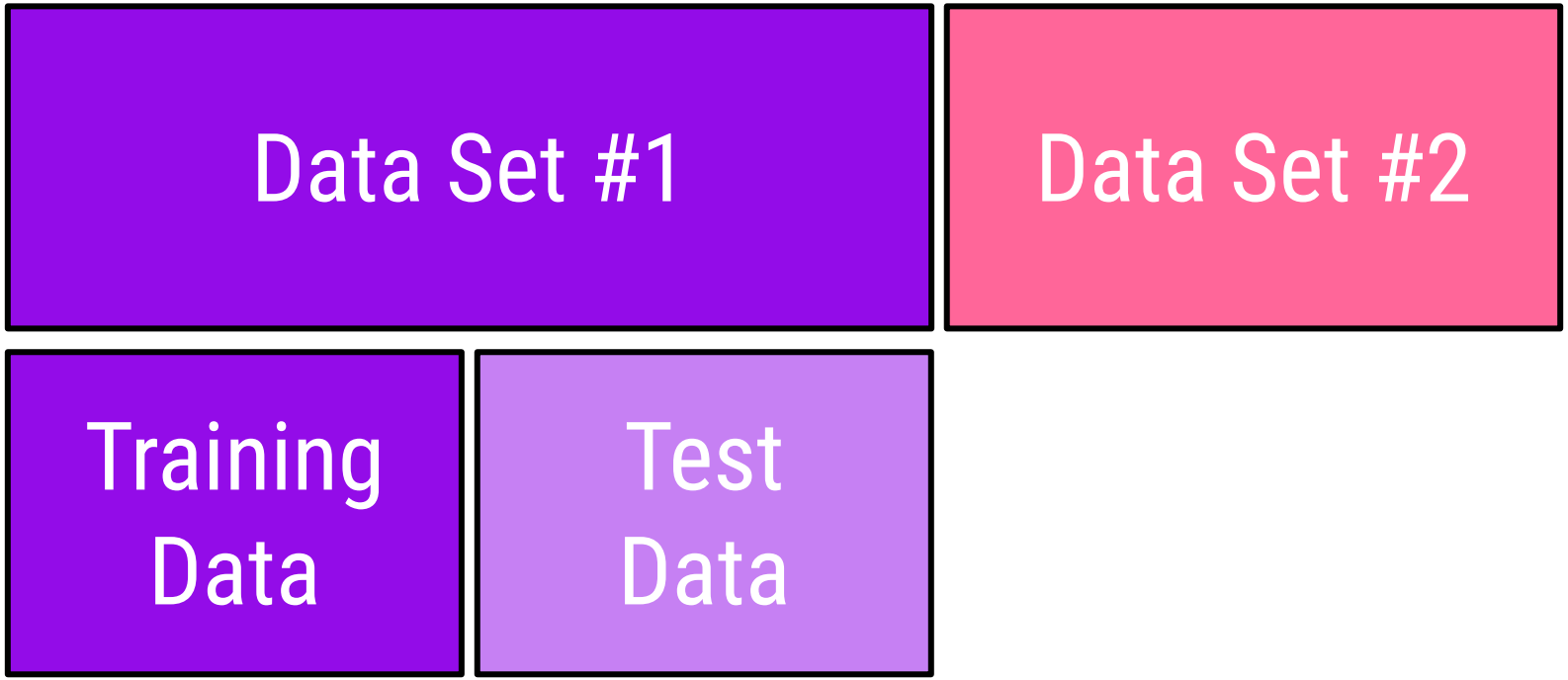
Sequence Read Archive

N=49,848

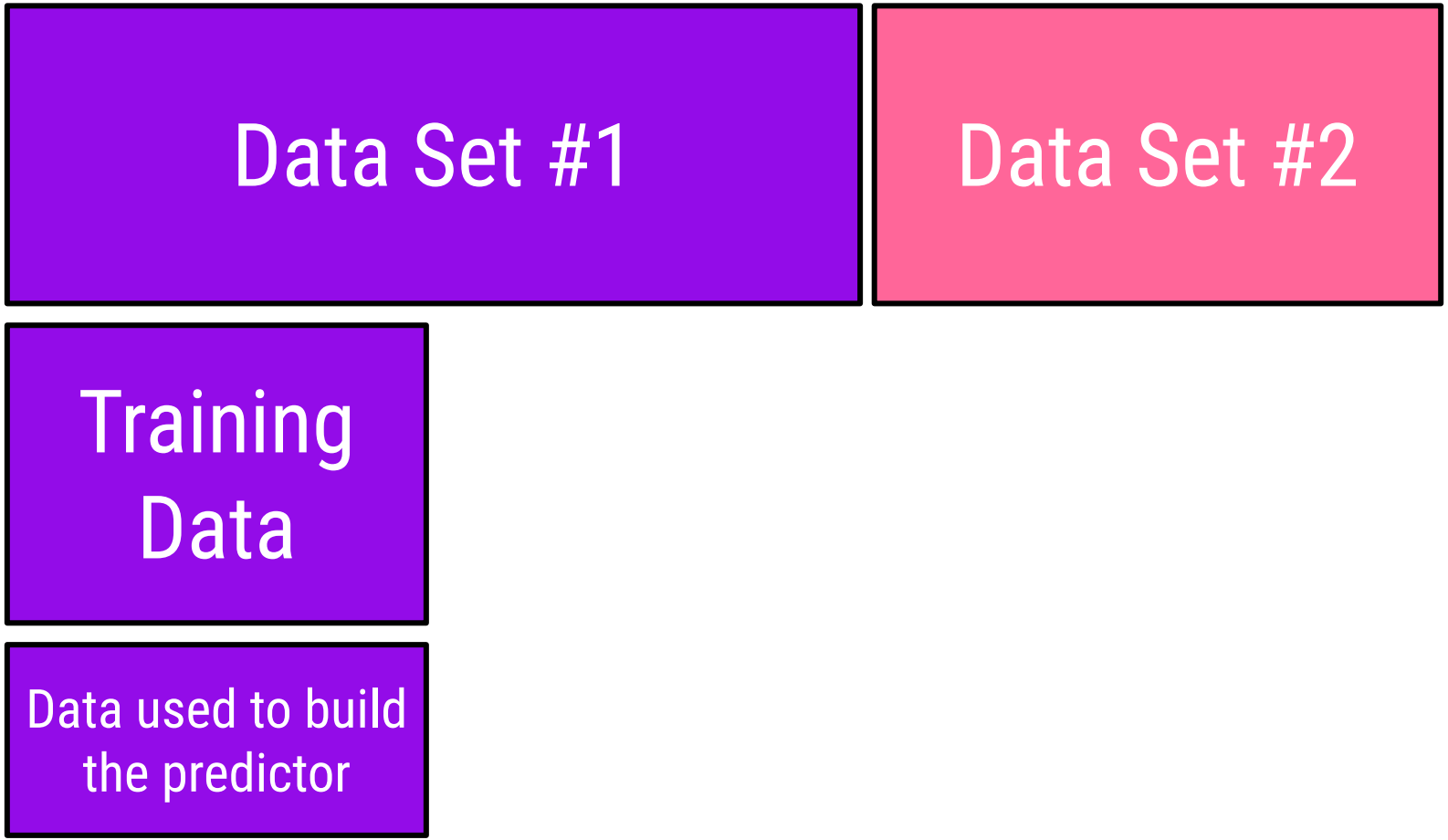


Machine Learning: Making predictions from data

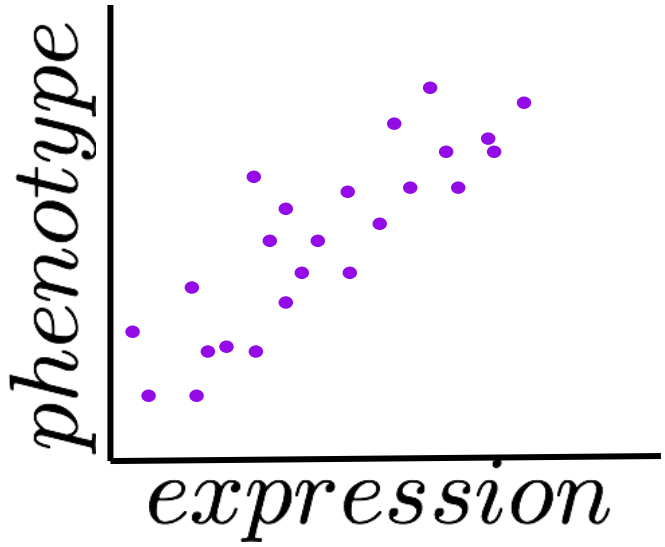
Machine Learning: Making predictions from data

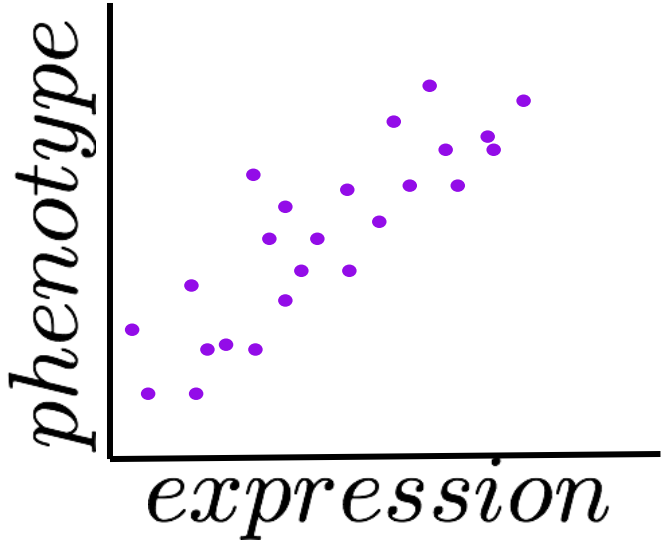


Machine Learning: Making predictions from data

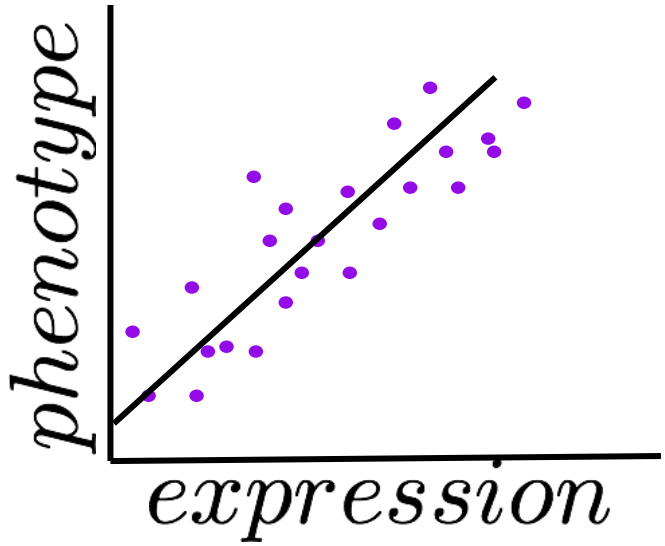


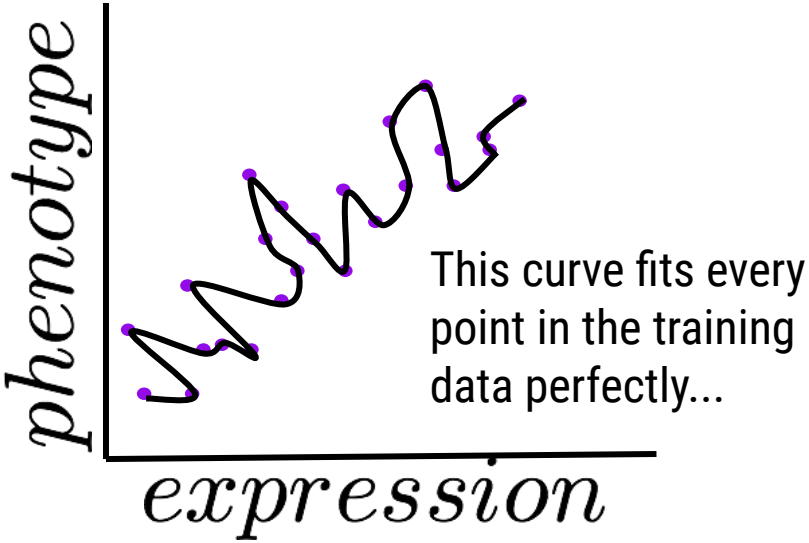
We’re interested in predicting phenotype from expression data...



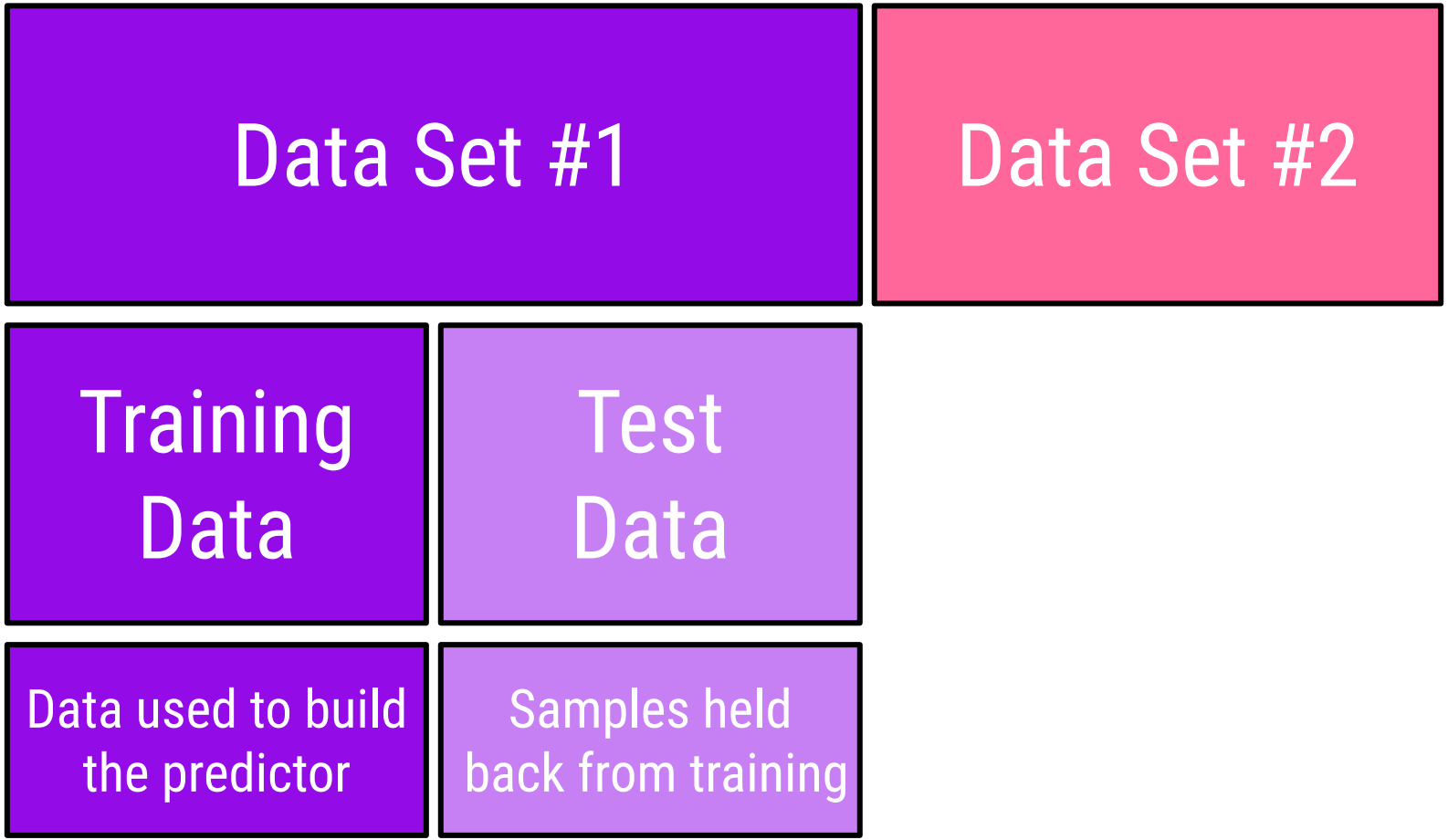


There are a number of different curves you could fit through these data

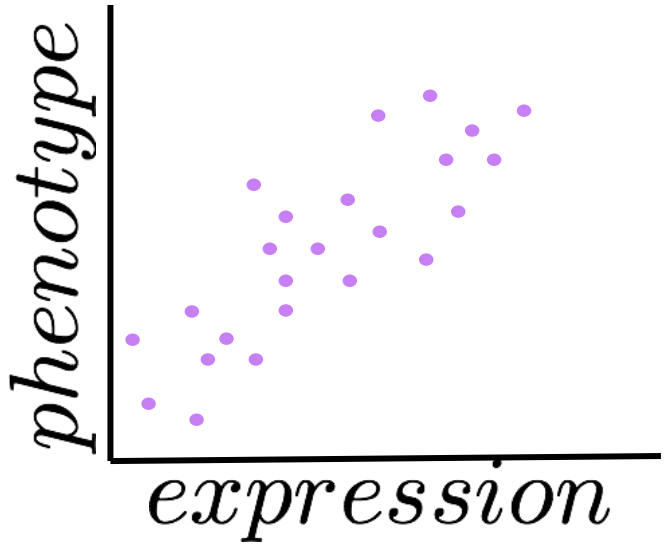


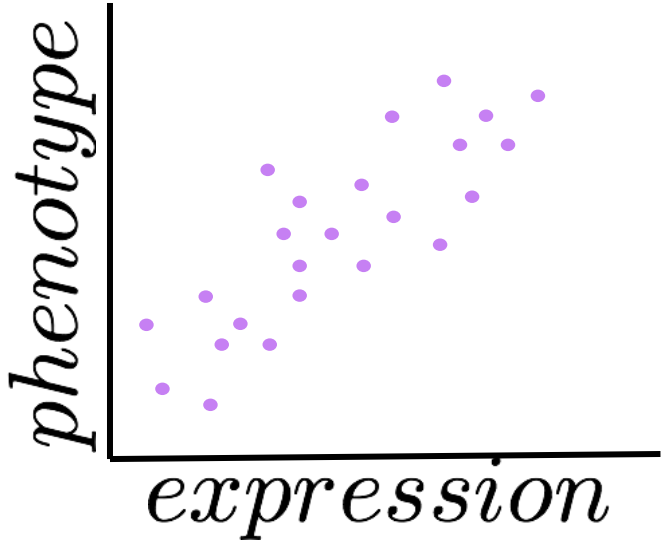


Machine Learning: Making predictions from data

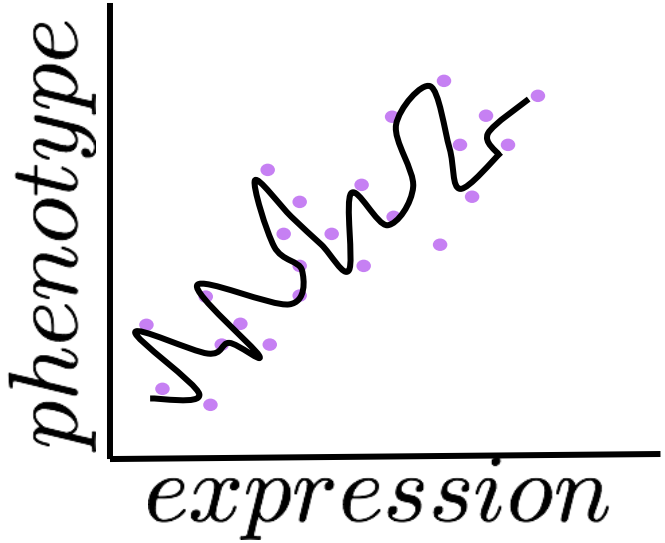


What if we tried to predict phenotype in the test data?

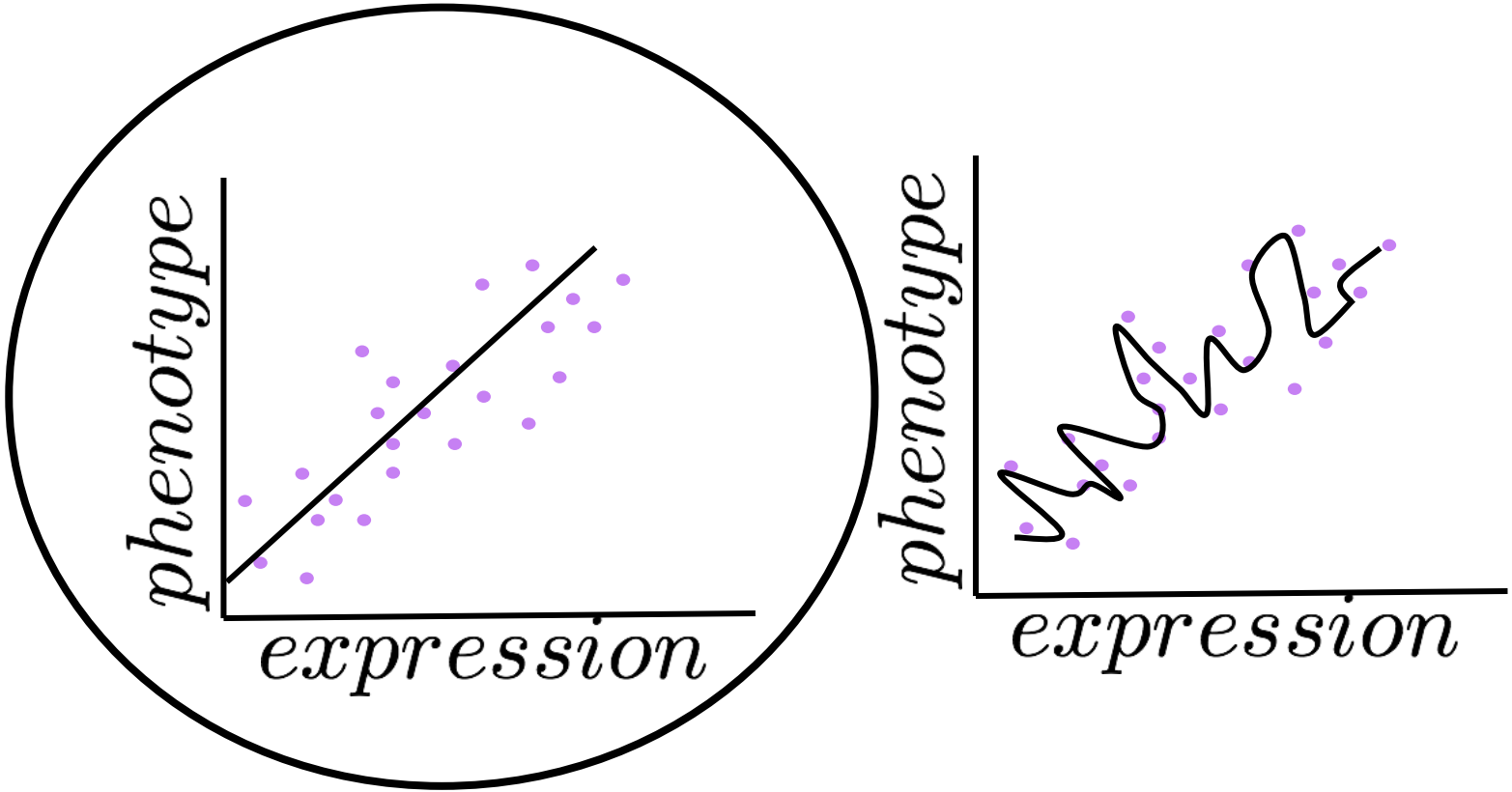




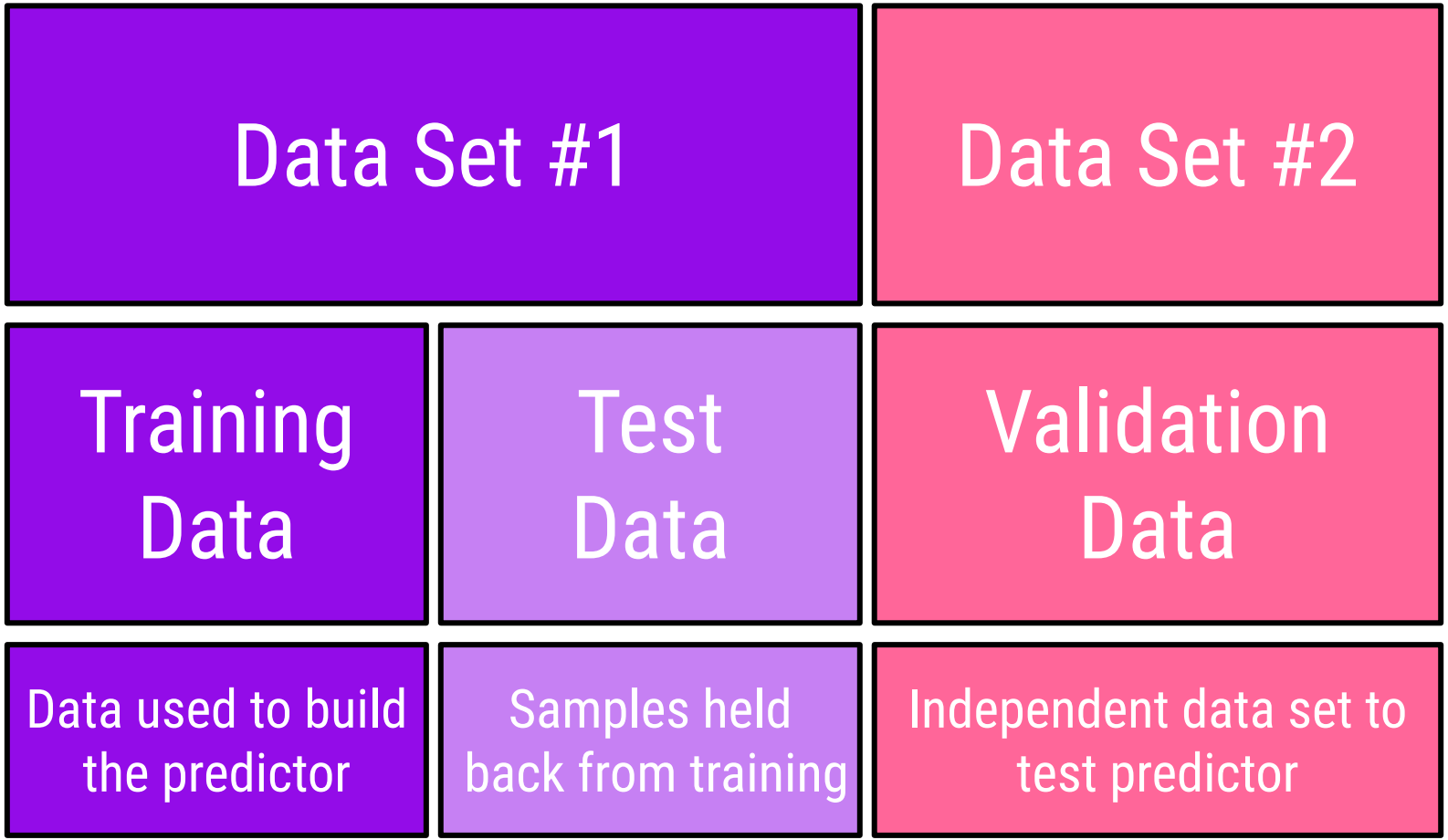
The curve no longer fits every point perfectly.



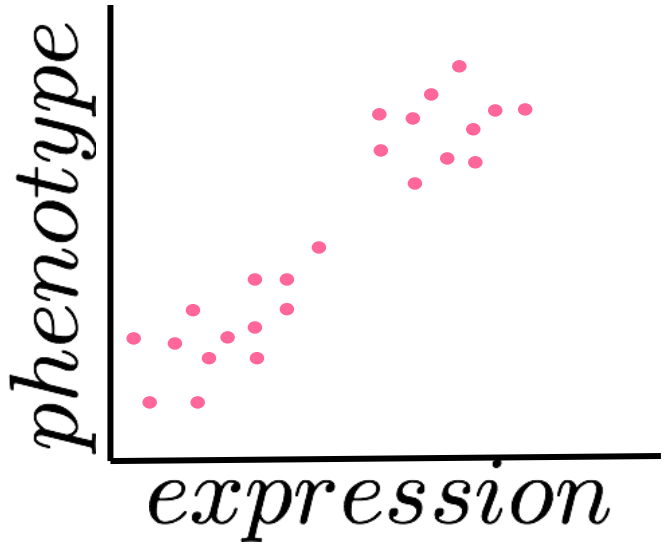
The curve no longer fits every point perfectly.



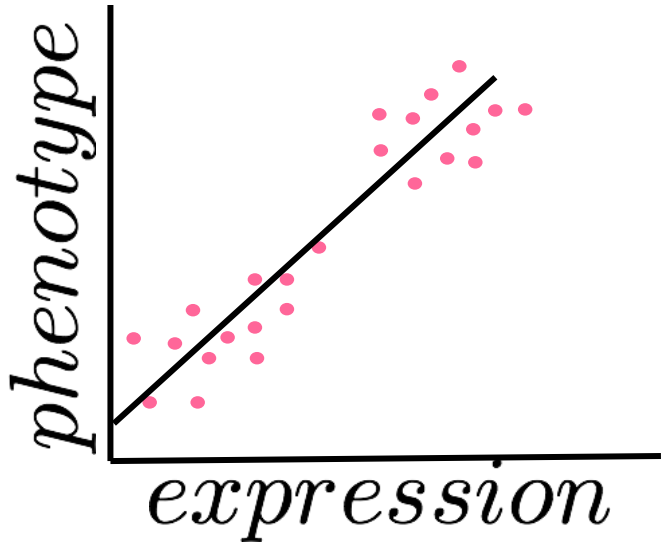
Machine Learning: Making predictions from data



We can now test prediction accuracy in an independent set of samples.



The line generated from the training data accurately predicts phenotype in the validation data



Prediction can be done using **linear regression**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

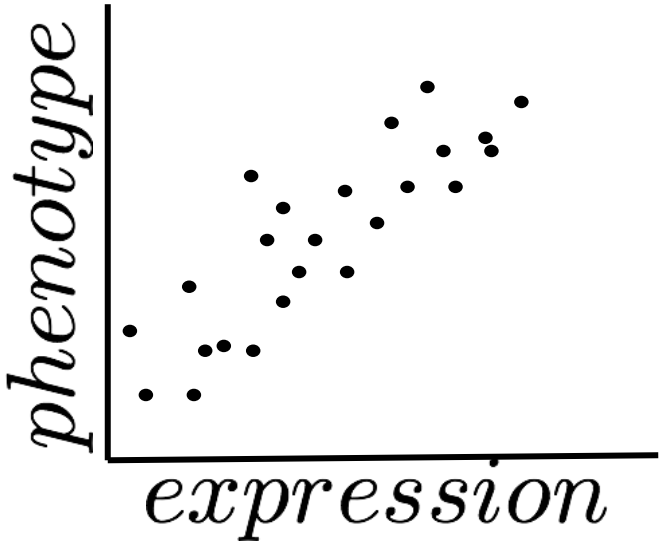
Prediction can be done using **linear regression**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$phenotype = \beta_0 + \beta_1 (expression) + \epsilon$$

Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$

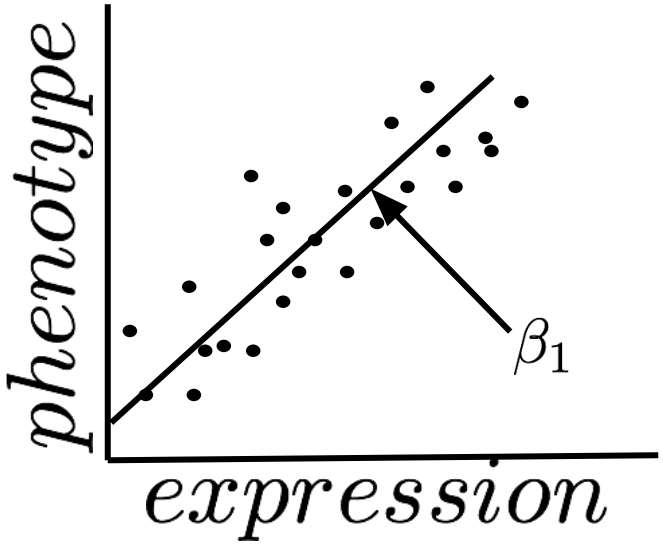


Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$

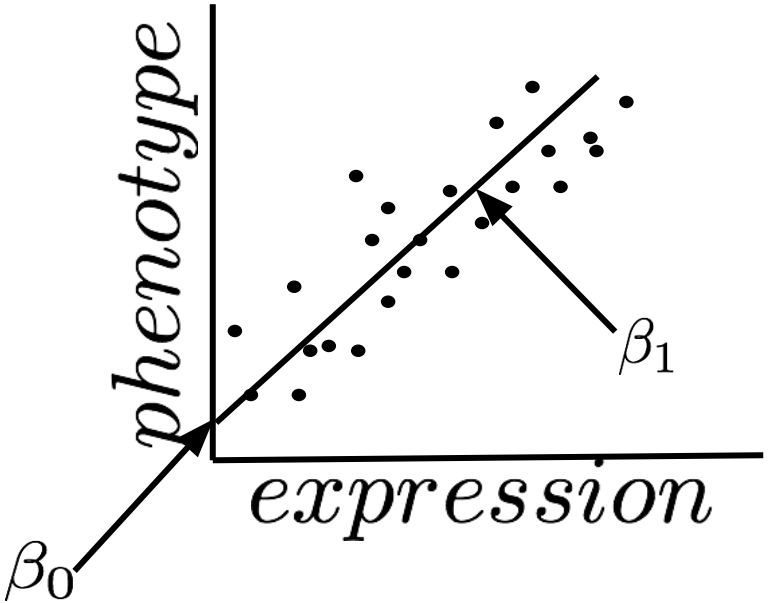
Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



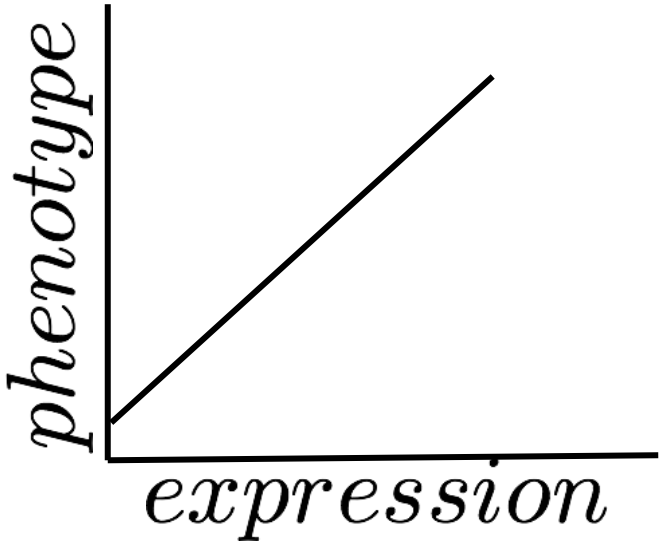
Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



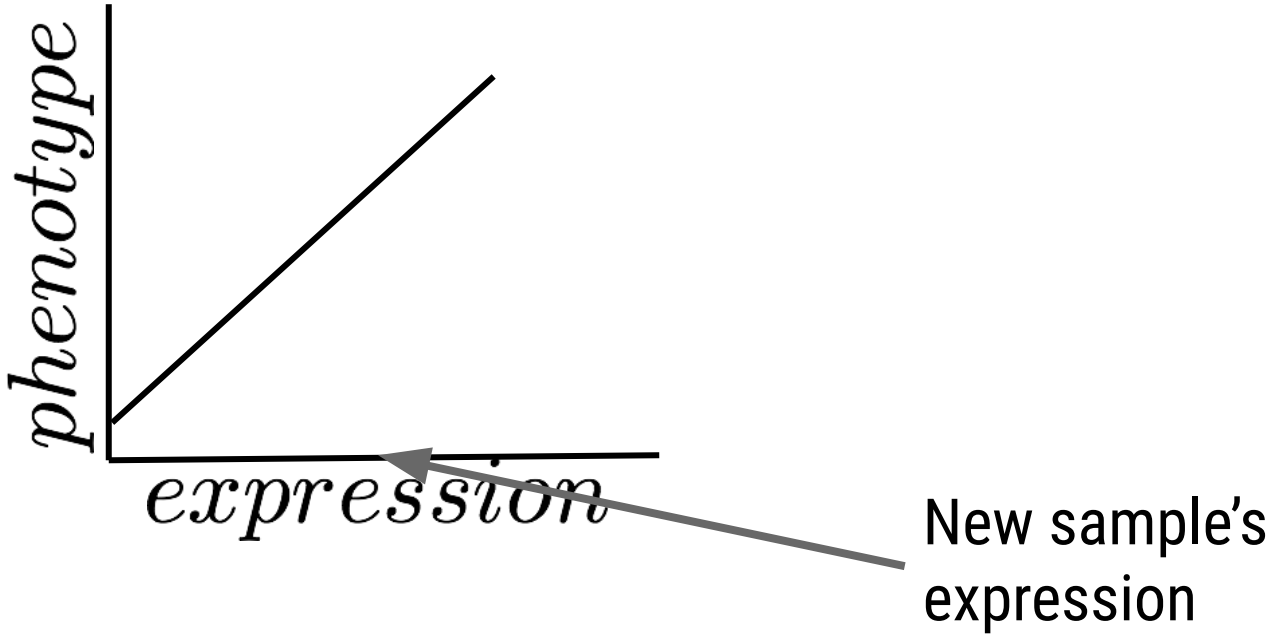
Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



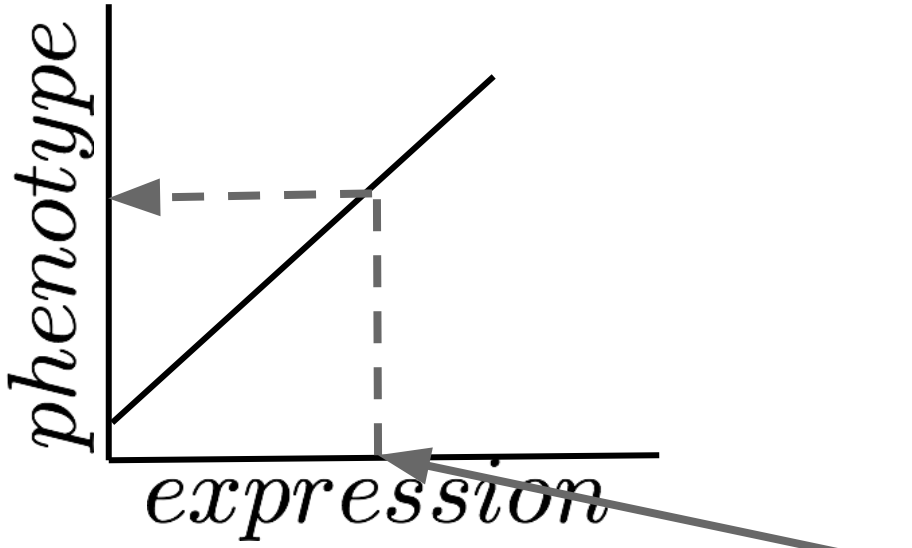
Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



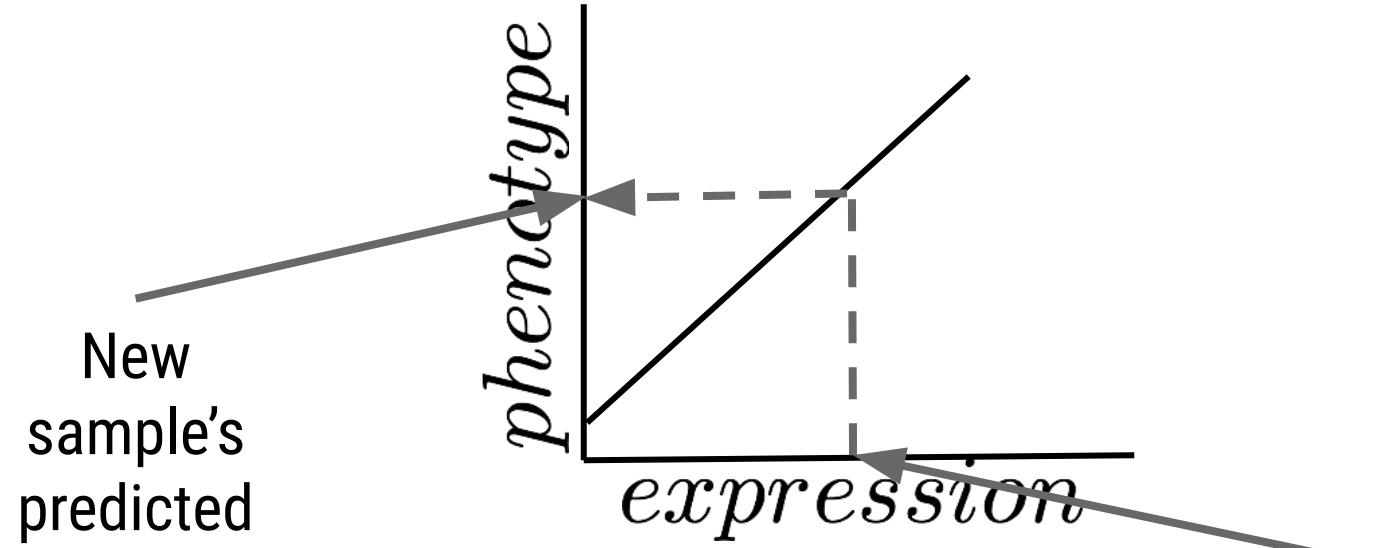
Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



Prediction is done using **linear regression**

$$phenotype = \beta_0 + \beta_1(expression) + \epsilon$$



Goal:

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

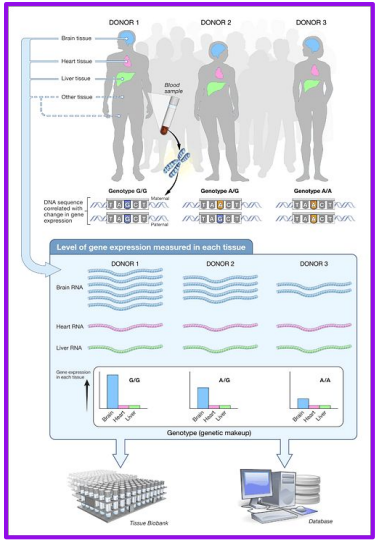
recount2

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

GTEx

Genotype Tissue Expression Project

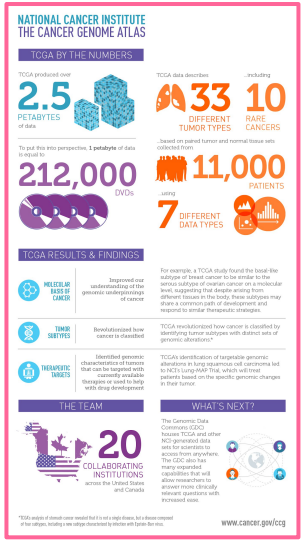
N=9,662



TCGA

The Cancer Genome Atlas

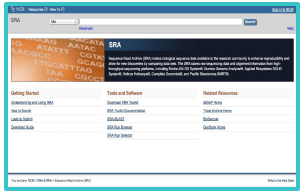
N=11,284



SRA

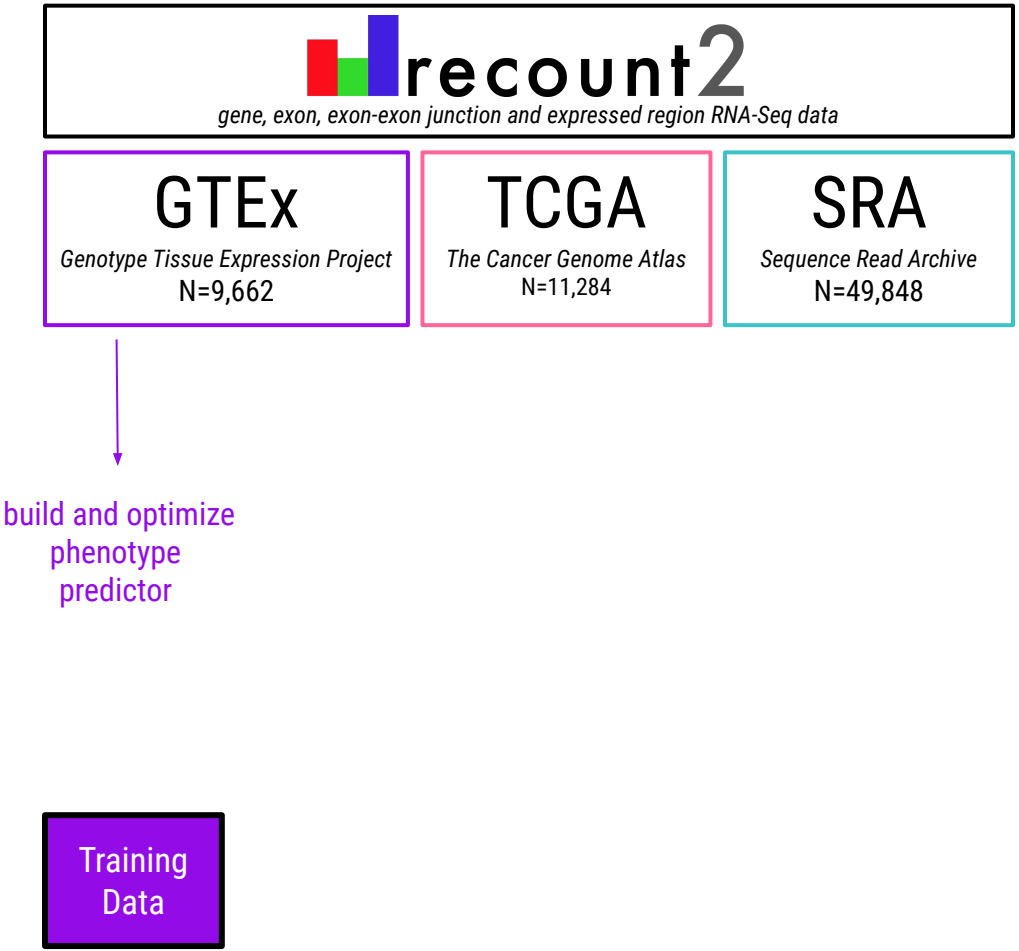
Sequence Read Archive

N=49,848



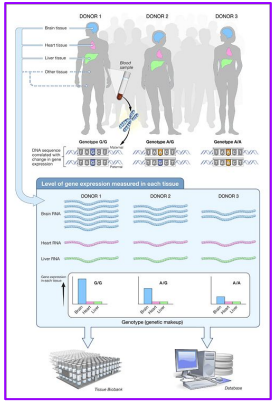
Goal :

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



Missingness limited in **GTEx phenotype data**

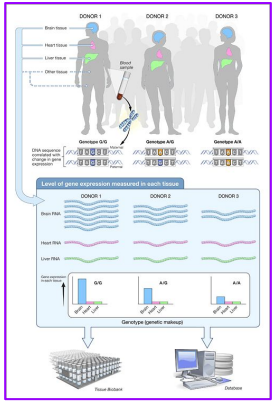
	Sex	Tissue	Race	Age
1	male	Lung	White	59
2	male	Brain	White	27
3	female	Heart	Black or African American	23
4	male	Brain	White	51
5	male	Skin	White	27
6	male	Lung	White	68
7	female	Brain	White	61
8	female	Adipose Tissue	White	42
9	male	Brain	White	40
10	female	Uterus	White	33
11	female	Nerve	White	60
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

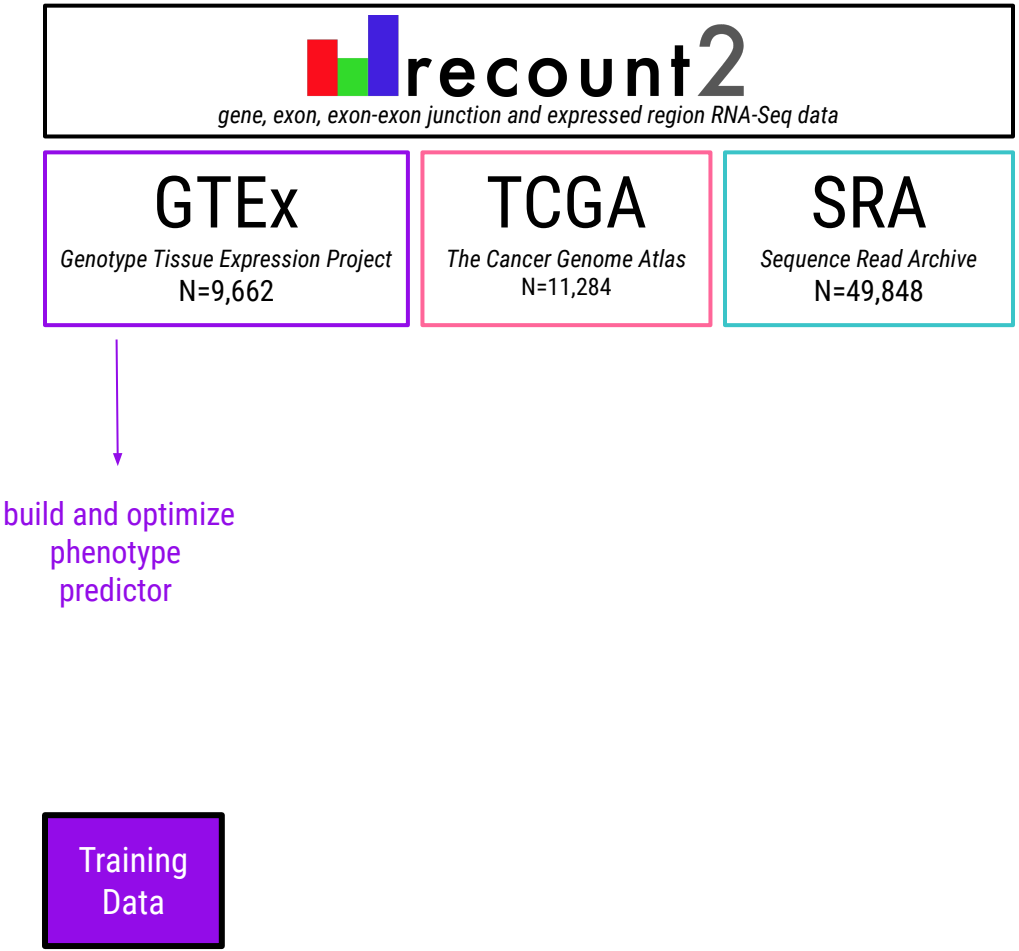


GTEx

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

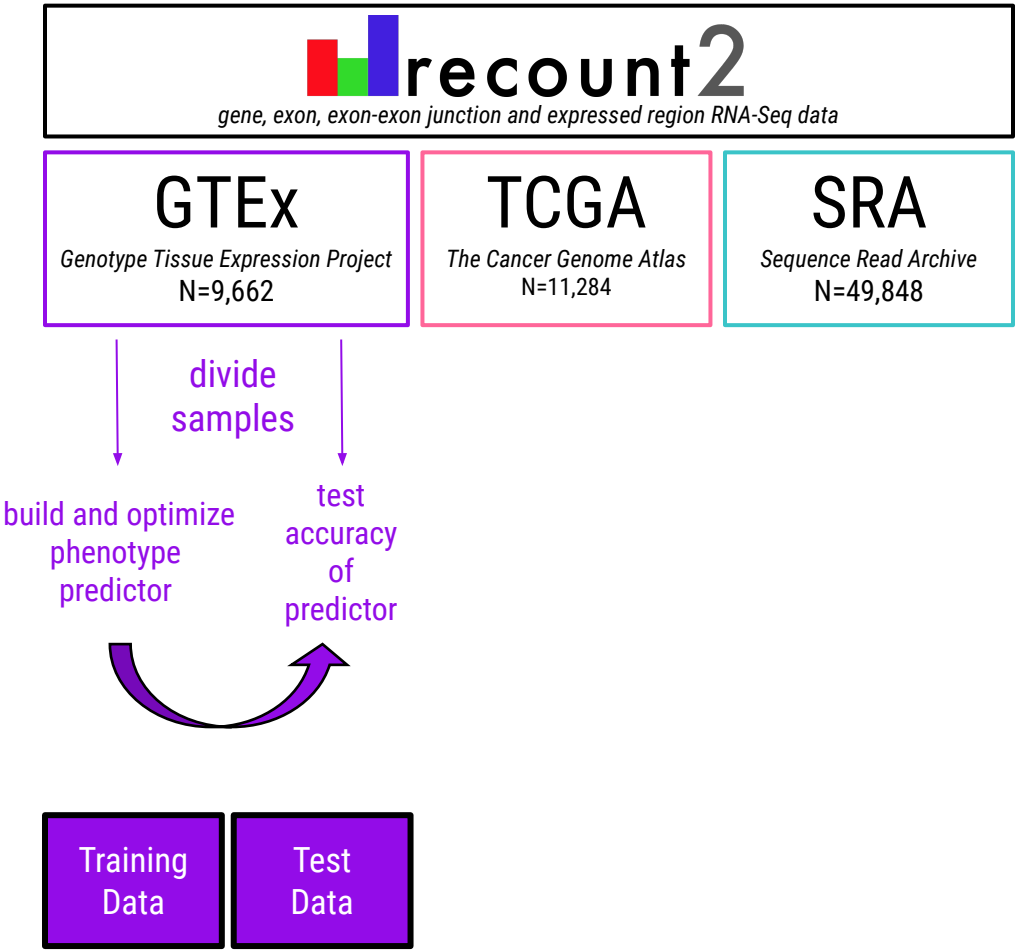
Goal :

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



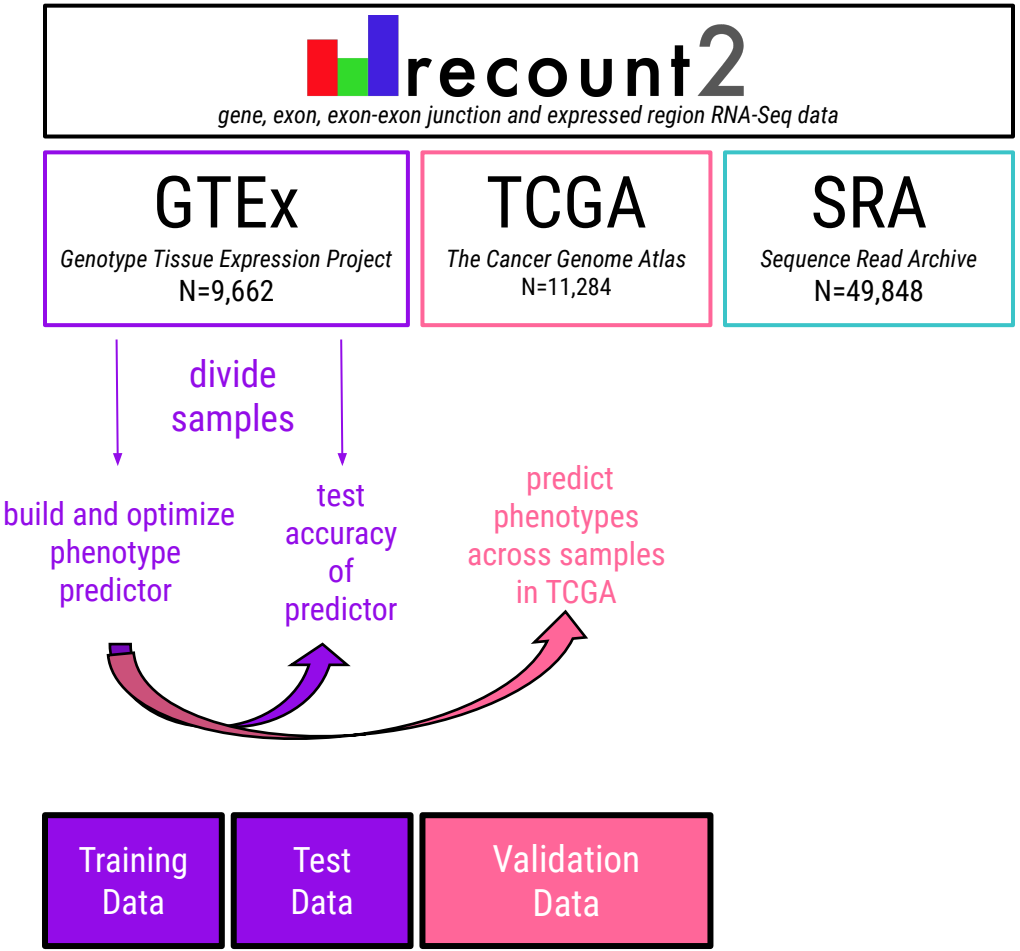
Goal :

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



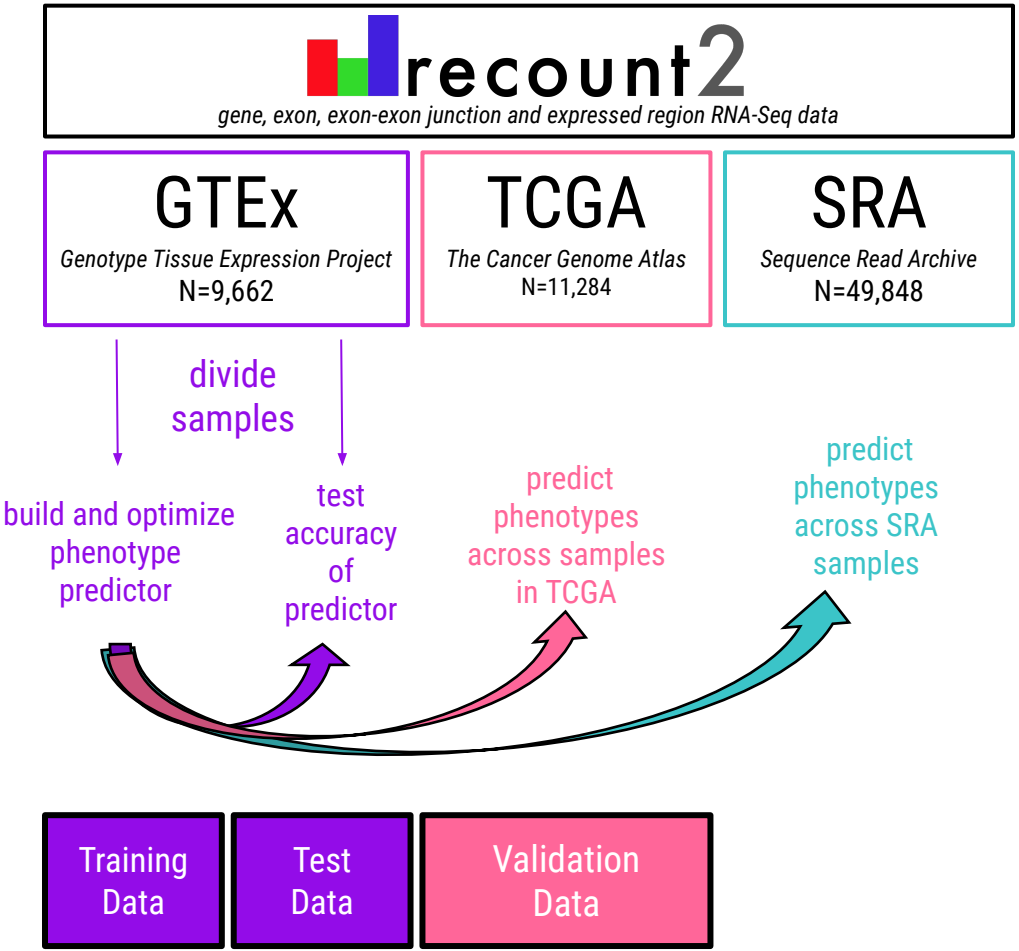
Goal :

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

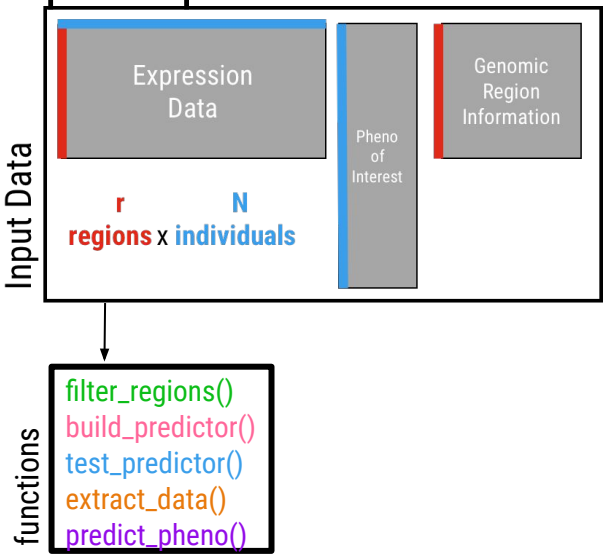


Goal :

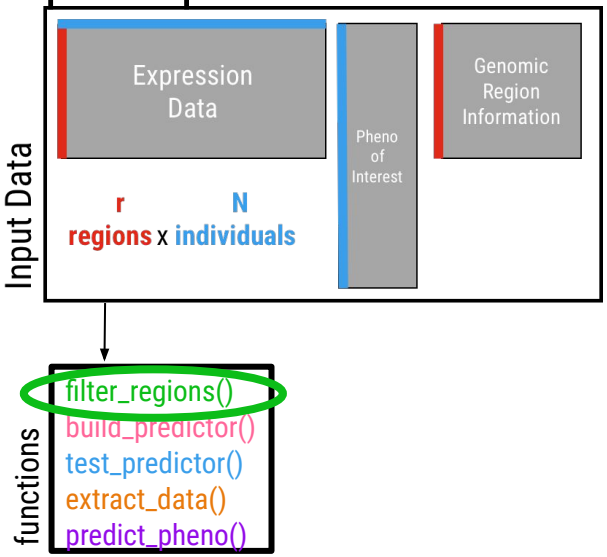
to accurately predict critical phenotype information for all samples in *recount2*



phenopredict



phenopredict



filter_regions()

Identify regions with differential expression for each level

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

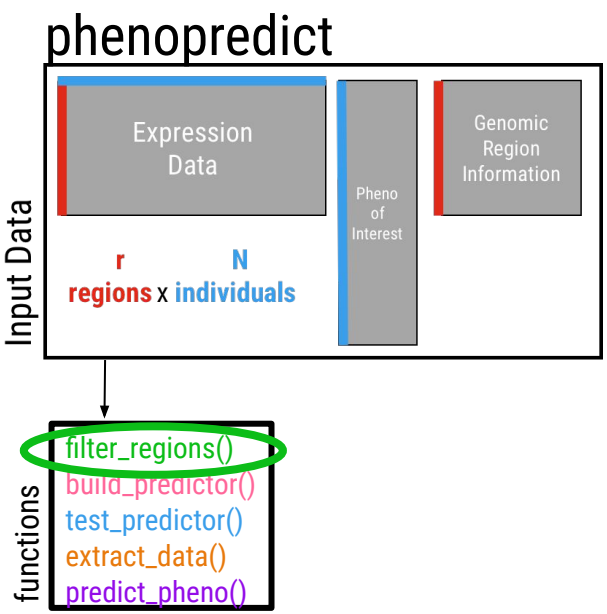
functions

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

filter_regions()

Identify regions with differential expression for each level

Set of discriminatory regions



filter_regions()

Identify regions with differential expression for each level

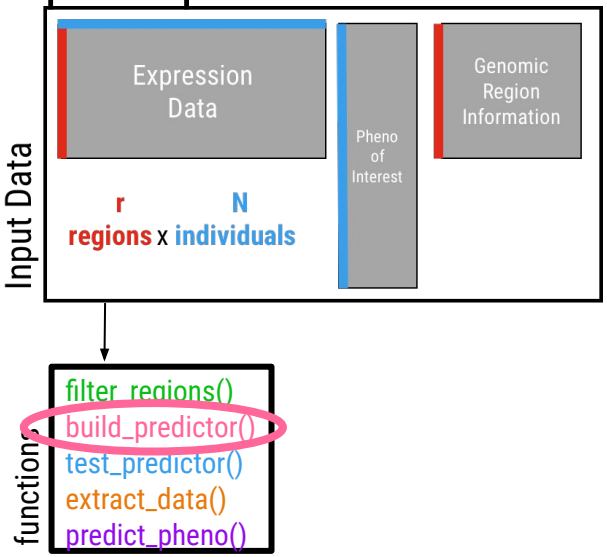
Set of discriminatory regions

$$E[\mathbf{E}_r \mid \mathbf{P}_l] = \alpha_{0r} + \mathbf{P}_l \alpha_r$$

$\mathbf{E}_r$ = expression at region r

$$\mathbf{P}_l = \begin{cases} 1, & \text{if level of phenotype} \\ 0, & \text{otherwise} \end{cases}$$

phenopredict



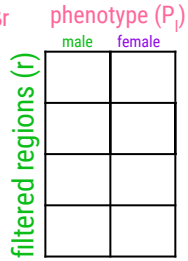
filter_regions()

Identify regions with differential expression for each level

build_predictor()

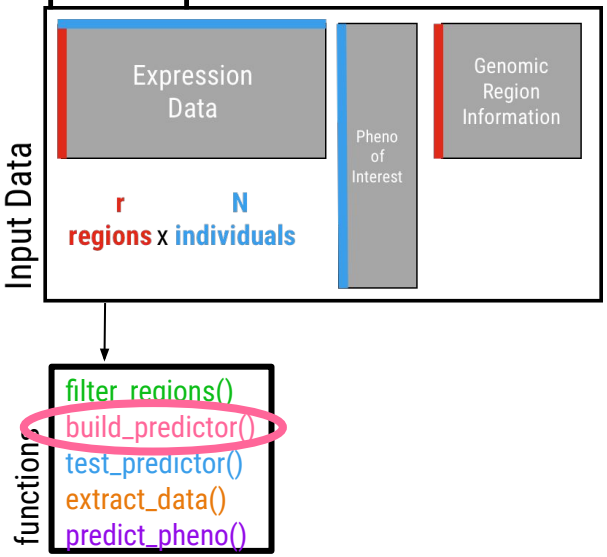
Extract coefficient estimates across regions

expression ~ phenotype



Set of discriminatory regions

phenopredict



filter_regions()

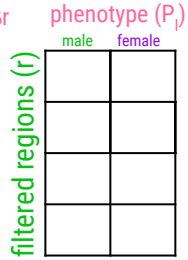
Identify regions with differential expression for each level

Set of discriminatory regions

build_predictor()

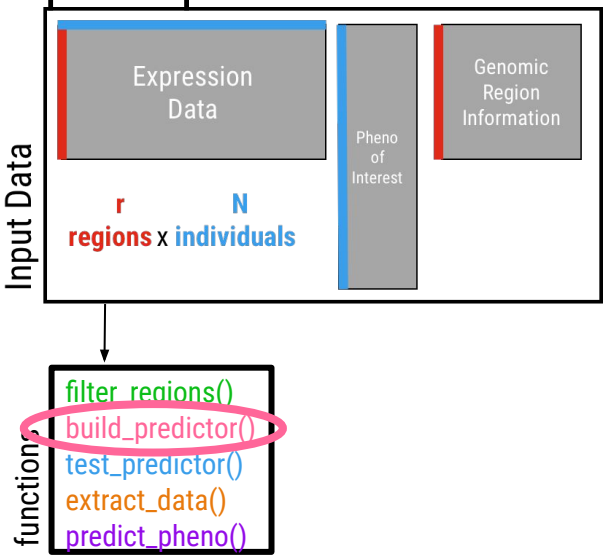
Extract coefficient estimates across regions

expression ~ phenotype



Relationship between expression and phenotype

phenopredict



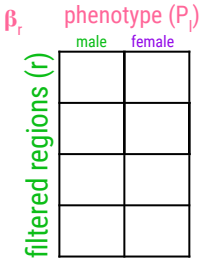
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



Set of discriminatory regions

Relationship between expression and phenotype

$$E[\{\mathbf{E}_r\}_{r\in\mathcal{S}} \mid \mathbf{P}] = \beta_r \mathbf{P}$$

$$\beta_r > 0 \forall \beta_r$$

$\mathbf{P} = l \times N$ matrix

l = number of levels in the phenotype

N = number of samples

$\mathbf{E}$ = expression for every region (r) in the selected set of regions $(\mathcal{S})$

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

filter_regions()

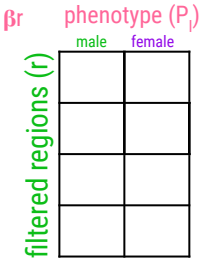
Identify regions with differential expression for each level

Set of discriminatory regions

build_predictor()

Extract coefficient estimates across regions

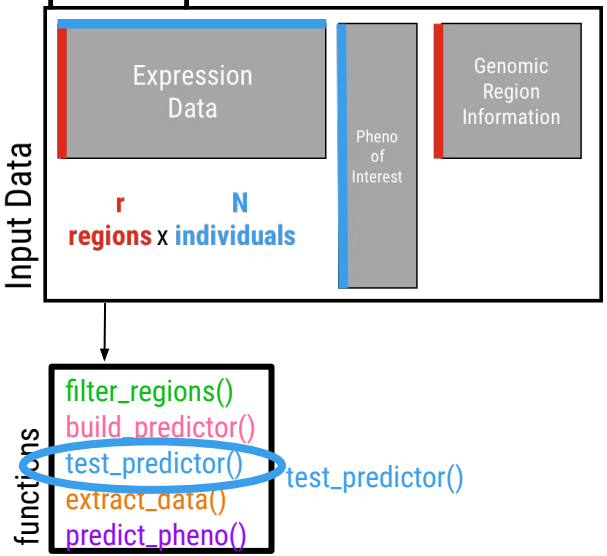
expression ~ phenotype



Relationship between expression and phenotype

Likelihood of phenotype for each individual

phenopredict



$\hat{\beta}_r$ = estimates of mean expression for each level (l) in phenotype ($\mathbf{P}$)

filter_regions()

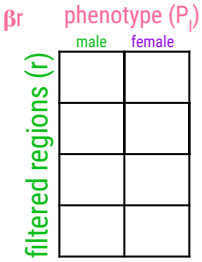
Identify regions with differential expression for each level

Set of discriminatory regions

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype

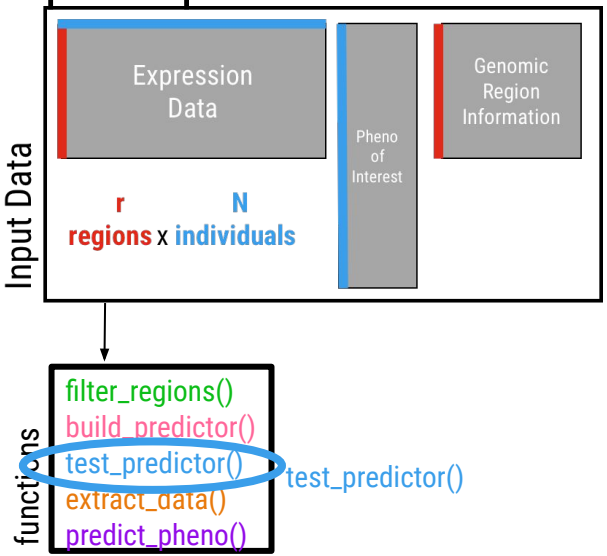


Relationship between expression and phenotype

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

Likelihood of phenotype for each individual

phenopredict



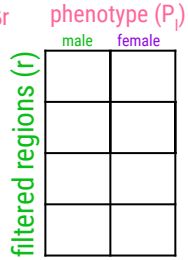
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



Set of discriminatory regions

male	0.99	0.02	0.04	0.98
female	0.01	0.98	0.96	0.02

Relationship between expression and phenotype

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

$\hat{\beta}_r =$ estimates of mean expression for each level (l) in phenotype $(\mathbf{P})$

$\hat{\gamma} = l \times N$ matrix containing the likelihood of belonging to each level in the phenotype

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions

filter_regions()

Identify regions with differential expression for each level

Set of discriminatory regions

male	0.99	0.02	0.04	0.98
female	0.01	0.98	0.96	0.02

$max_l(\hat{\gamma})$

Likelihood of phenotype for each individual

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype

β_r phenotype (P_l)

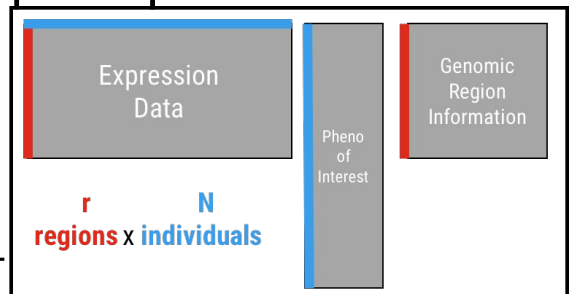
male female

filtered regions (r)

Relationship between expression and phenotype

phenopredict

Input Data



functions

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

filter_regions()

Identify regions with differential expression for each level

Set of discriminatory regions

male	0.99	0.02	0.04	0.98
female	0.01	0.98	0.96	0.02

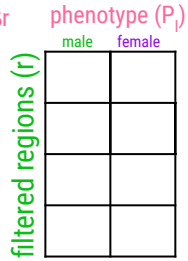
$max_l(\hat{\gamma})$

Likelihood of phenotype for each individual

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype

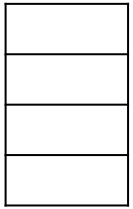


Relationship between expression and phenotype

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions



Assign phenotype to most likely category

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions

male

Assign phenotype to most likely category

filter_regions()

Identify regions with differential expression for each level

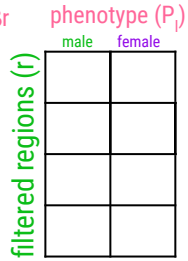
Set of discriminatory regions

male	**0.99**	0.02	0.04	0.98
female	0.01	0.98	0.96	0.02

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



Relationship between expression and phenotype

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions

male
female
female
male

Assign phenotype to most likely category

filter_regions()

Identify regions with differential expression for each level

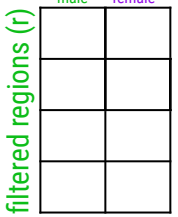
Set of discriminatory regions

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype

βr phenotype (P_i)



Relationship between expression and phenotype

male	**0.99**	0.02	0.04	0.98
female	0.01	**0.98**	**0.96**	0.02

phenopredict

Input Data

Expression Data

r regions x N individuals

Pheno of Interest

Genomic Region Information

functions

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

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions	reported
male	male
female	female
female	female
male	male

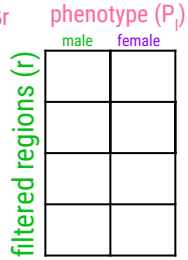
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

test_predictor()

Predict phenotype and assess accuracy in training set data

predictions	reported
male	male
female	female
female	female
male	male

Prediction accuracy: 100%

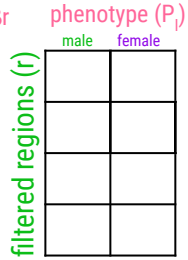
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N individuals

functions

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

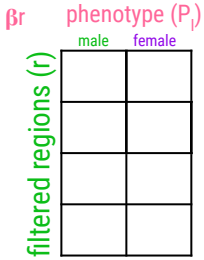
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



test_predictor()

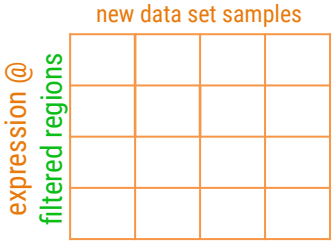
Predict phenotype and assess accuracy in training set data

predictions	reported
male	male
female	female
female	female
male	male

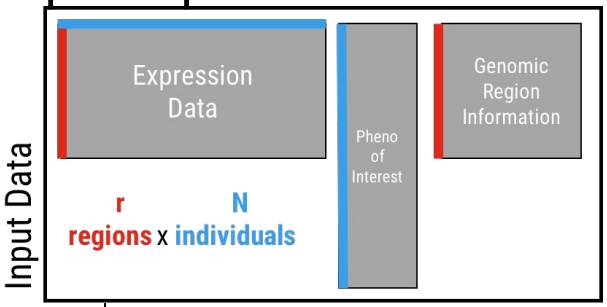
Prediction accuracy: 100%

extract_data()

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



phenopredict



functions
- filter_regions()
- build_predictor()
- test_predictor()
- extract_data()
- predict_pheno()

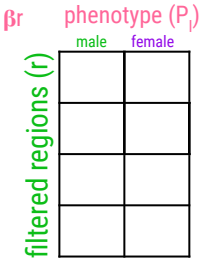
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



test_predictor()

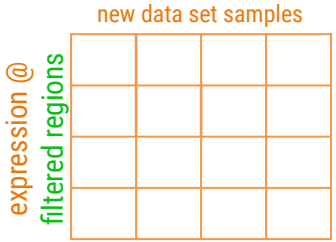
Predict phenotype and assess accuracy in training set data

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



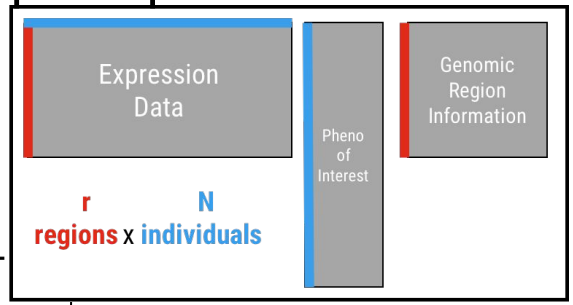
predict_pheno()

Predict phenotypes across samples in this new data set

apply coefficient estimates to the extracted data

phenopredict

Input Data



functions

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

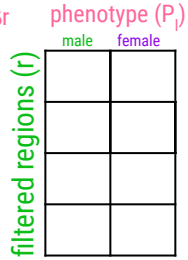
filter_regions()

Identify regions with differential expression for each level

build_predictor()

Extract coefficient estimates across regions

expression ~ phenotype



test_predictor()

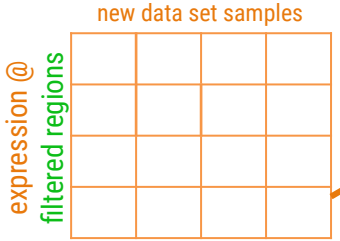
Predict phenotype and assess accuracy in training set data

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



predict_pheno()

Predict phenotypes across samples in this new data set

apply coefficient estimates to the extracted data

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N 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

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

phenopredict

Input Data

Expression Data

Pheno of Interest

Genomic Region Information

r regions x N 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)

filtered regions (r)	male	female

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				

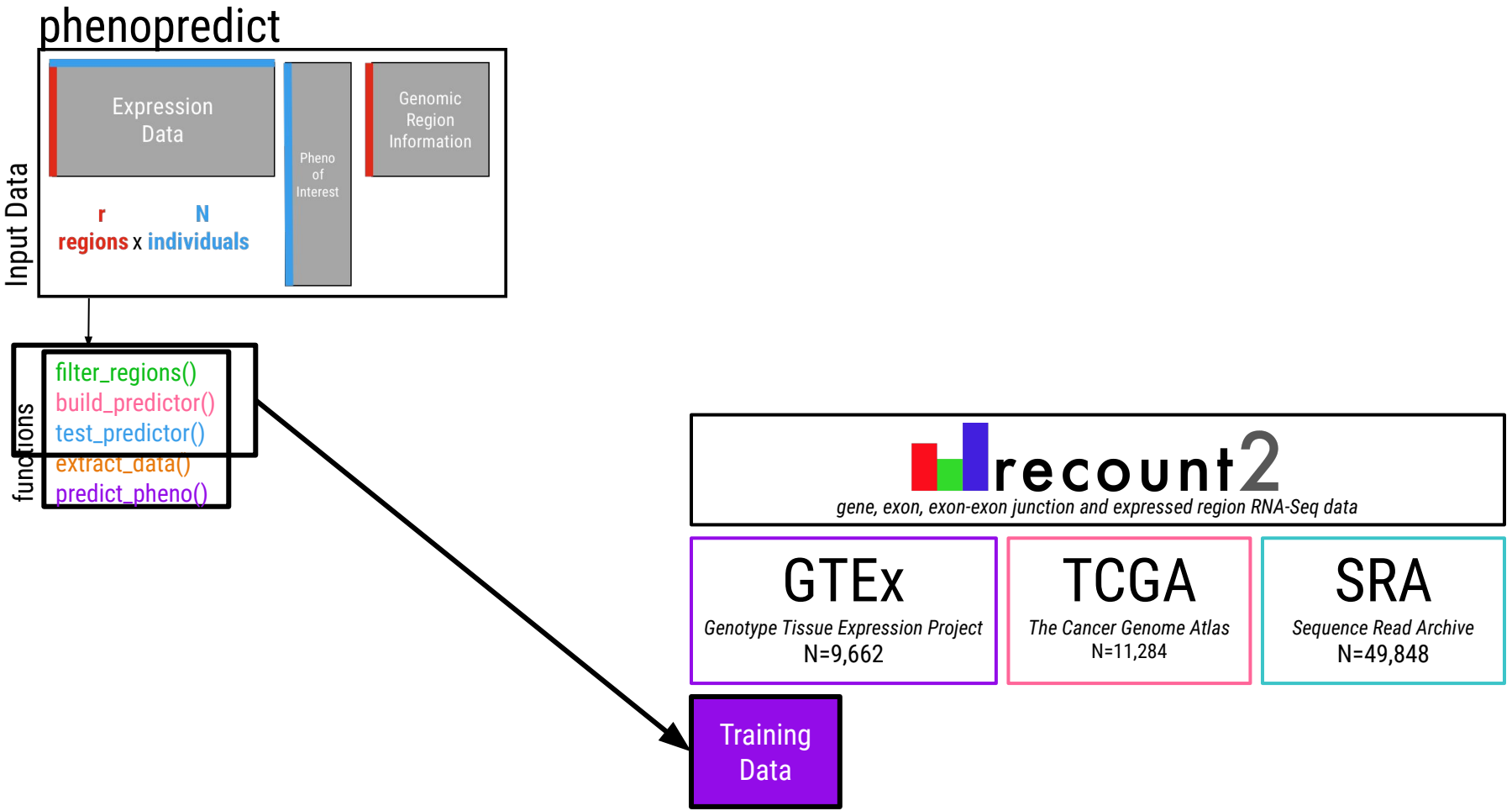
predict_pheno()

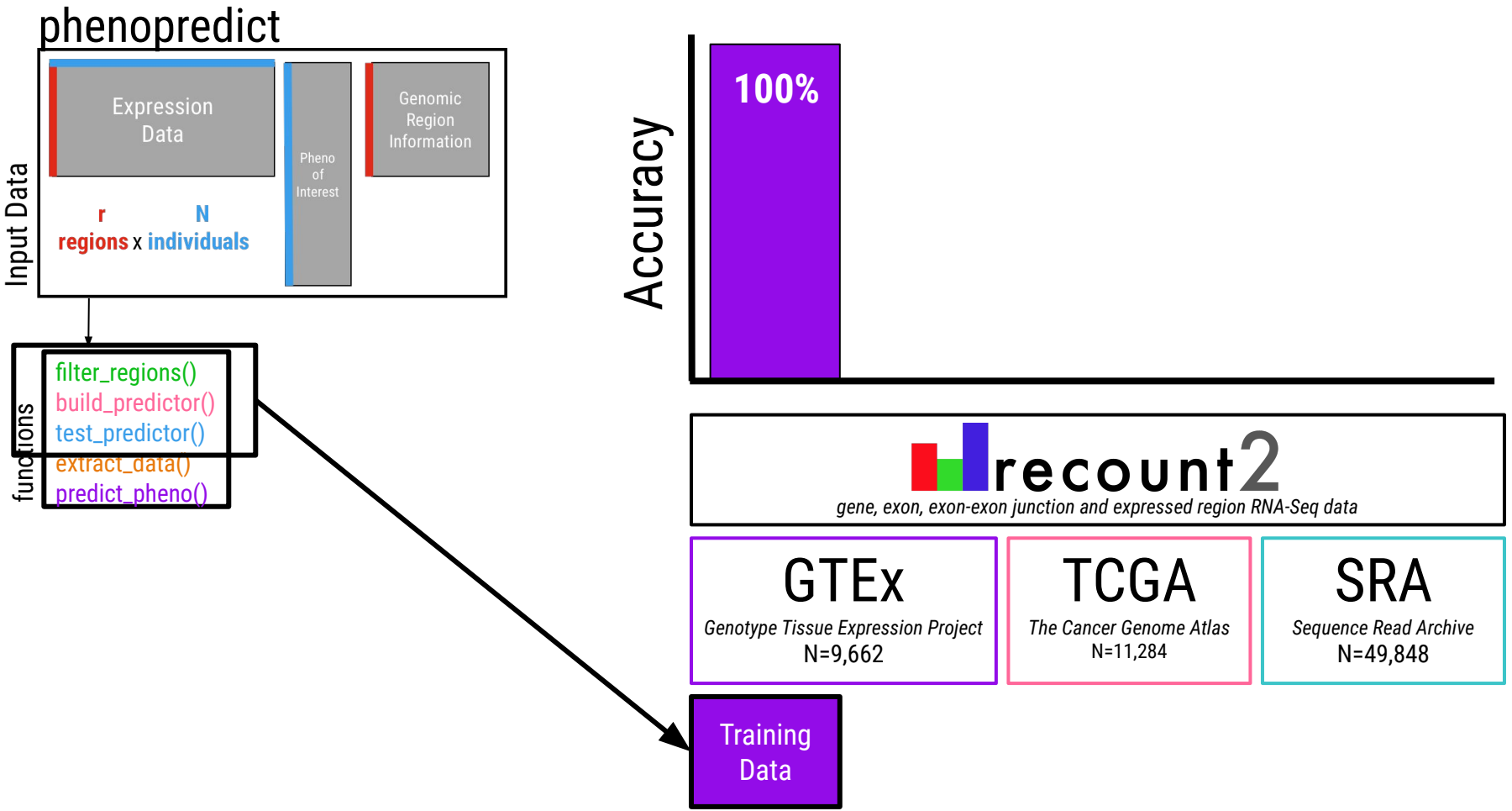
Predict phenotypes across samples in this new data set

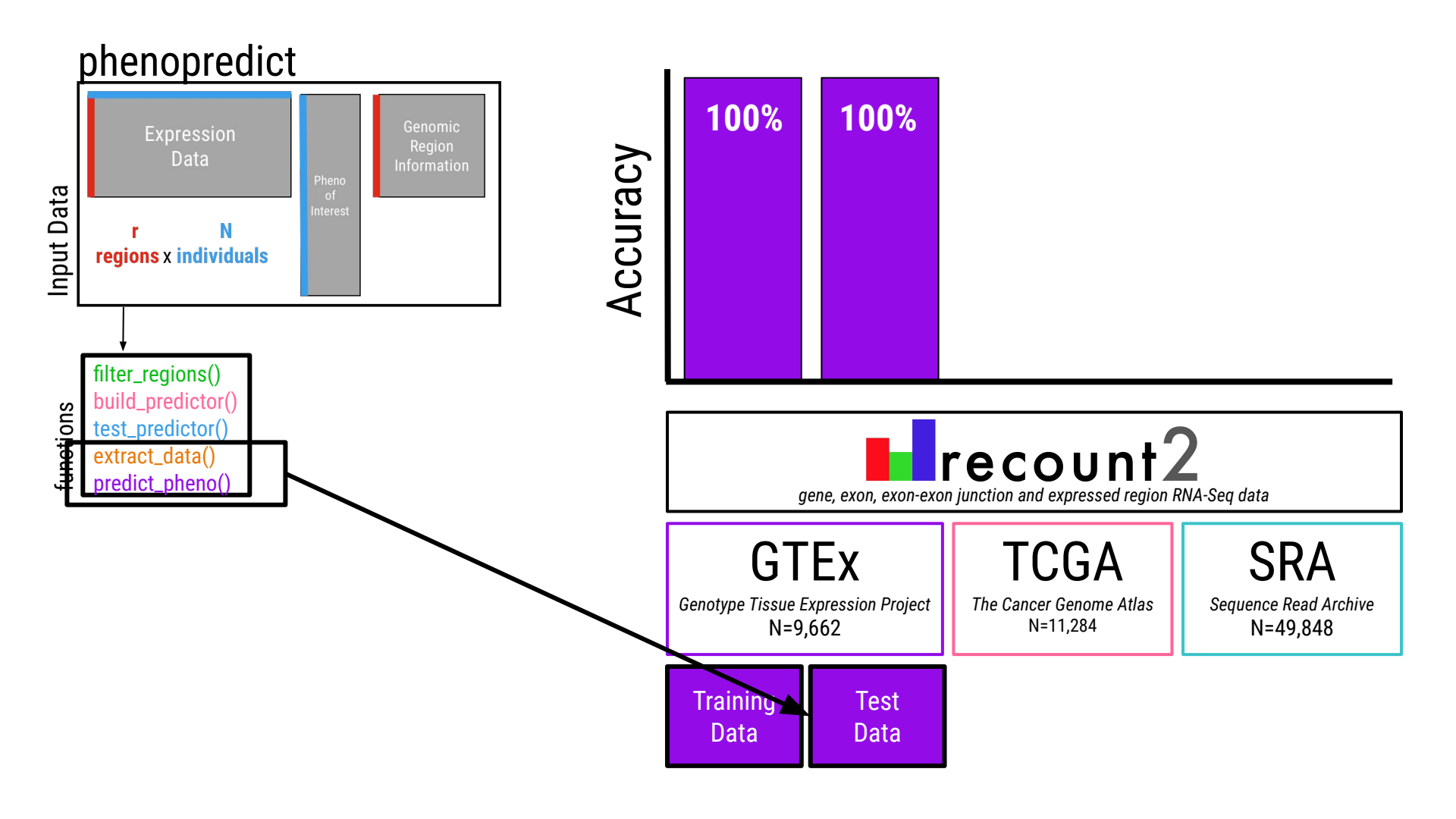
apply coefficient estimates to the extracted data

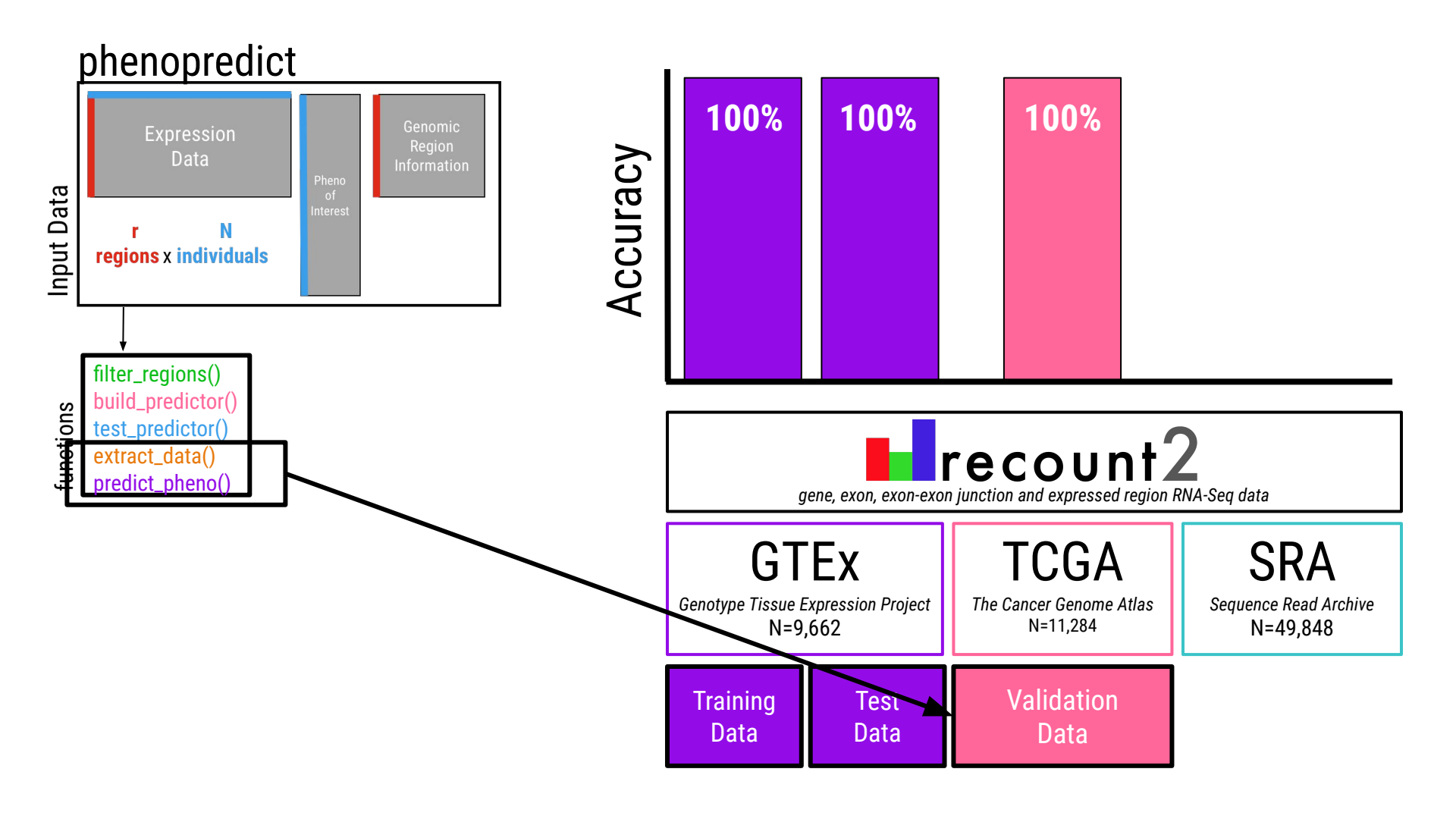
predictions in new data set
male
male
male
female

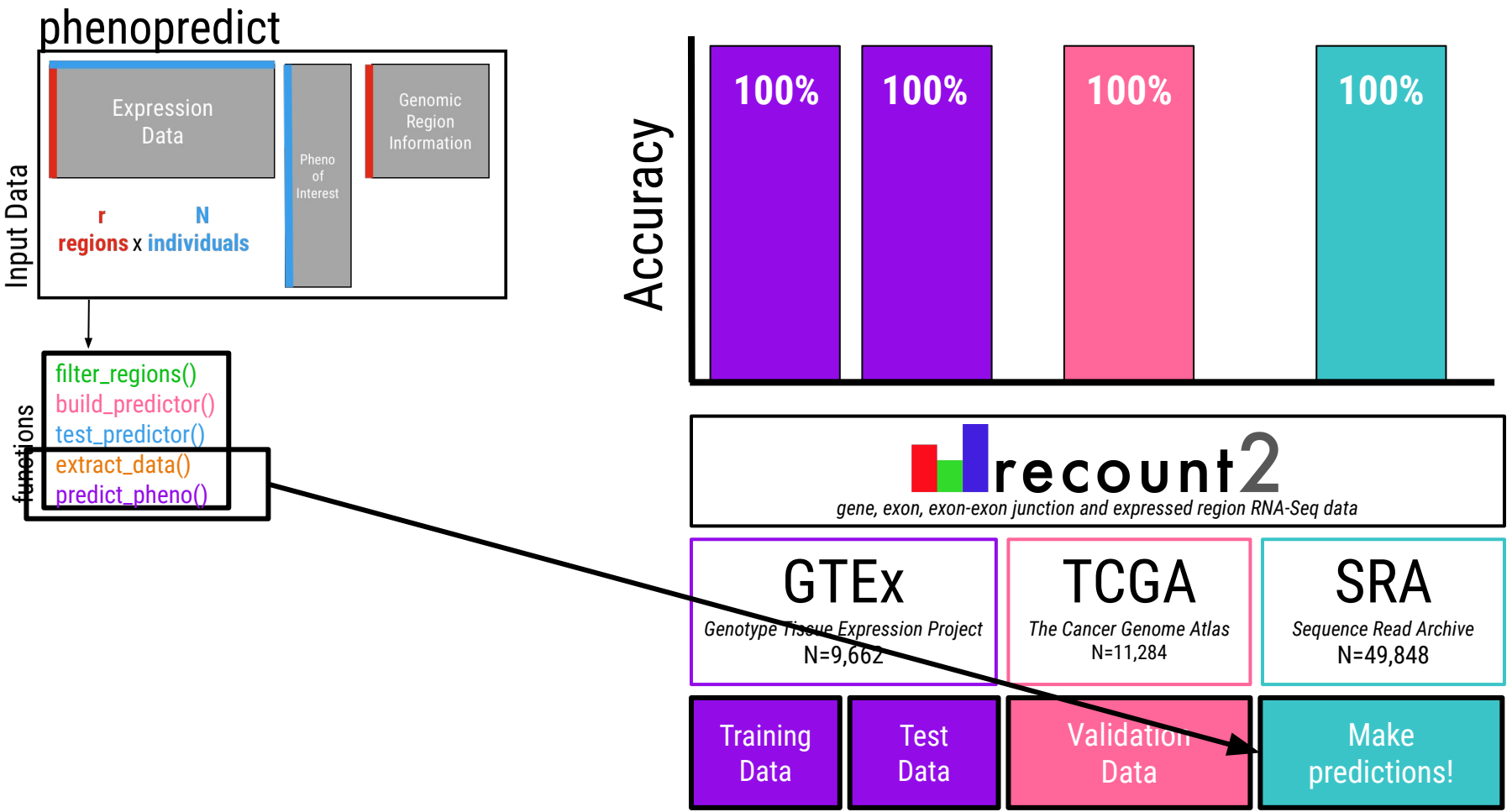
$$E[\{\mathbf{E}_r^*\}_{r\in\mathcal{S}} \mid \hat{\beta}_r] = \hat{\beta}_r\gamma^*$$

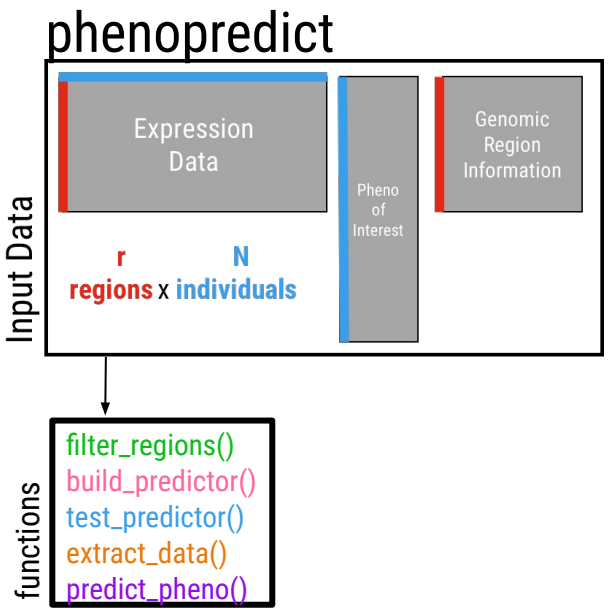

phenopredict
Input Data
Expression Data
Pheno of Interest
Genomic Region Information
r
N
regions x individuals
functions
filter_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848
Training Data

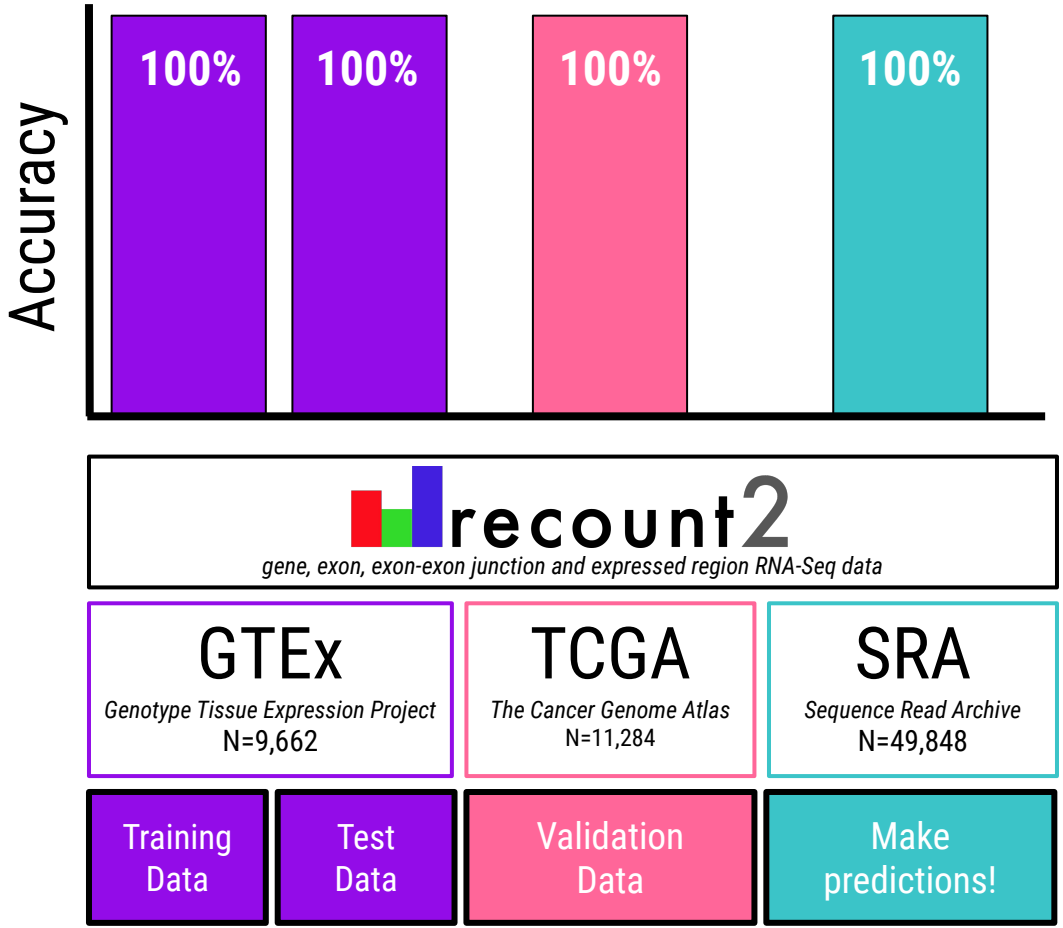

phenopredict
Input Data
Expression Data
Pheno of Interest
Genomic Region Information
r
N
regions x individuals
functions
filter_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()
Training Data
Accuracy
100%
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848


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()
Accuracy
100%
100%
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848
Training Data
Test Data


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()
Accuracy
100%
100%
100%
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848
Training Data
Test Data
Validation Data

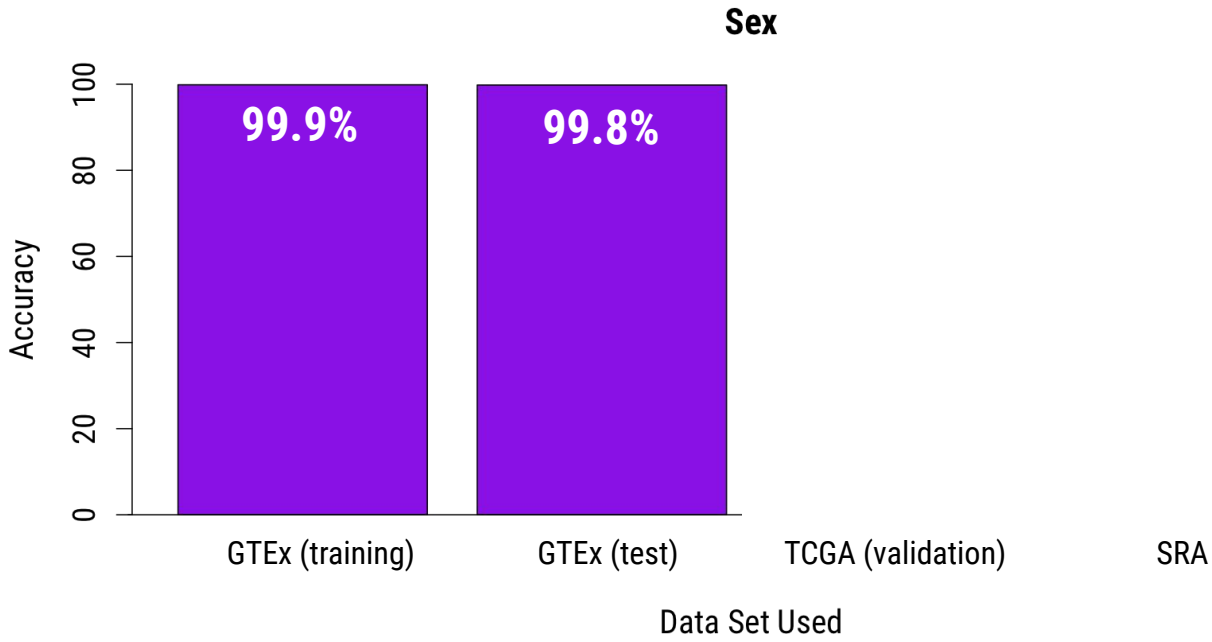

phenopredict
Input Data
Expression Data
Pheno of Interest
Genomic Region Information
r
N
regions x individuals
functions
filter_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()
Accuracy
100%
100%
100%
100%
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848
Training Data
Test Data
Validation Data
Make predictions!


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()


Accuracy
100%
100%
100%
100%
recount2
gene, exon, exon-exon junction and expressed region RNA-Seq data
GTEx
Genotype Tissue Expression Project
N=9,662
TCGA
The Cancer Genome Atlas
N=11,284
SRA
Sequence Read Archive
N=49,848
Training Data
Test Data
Validation Data
Make predictions!

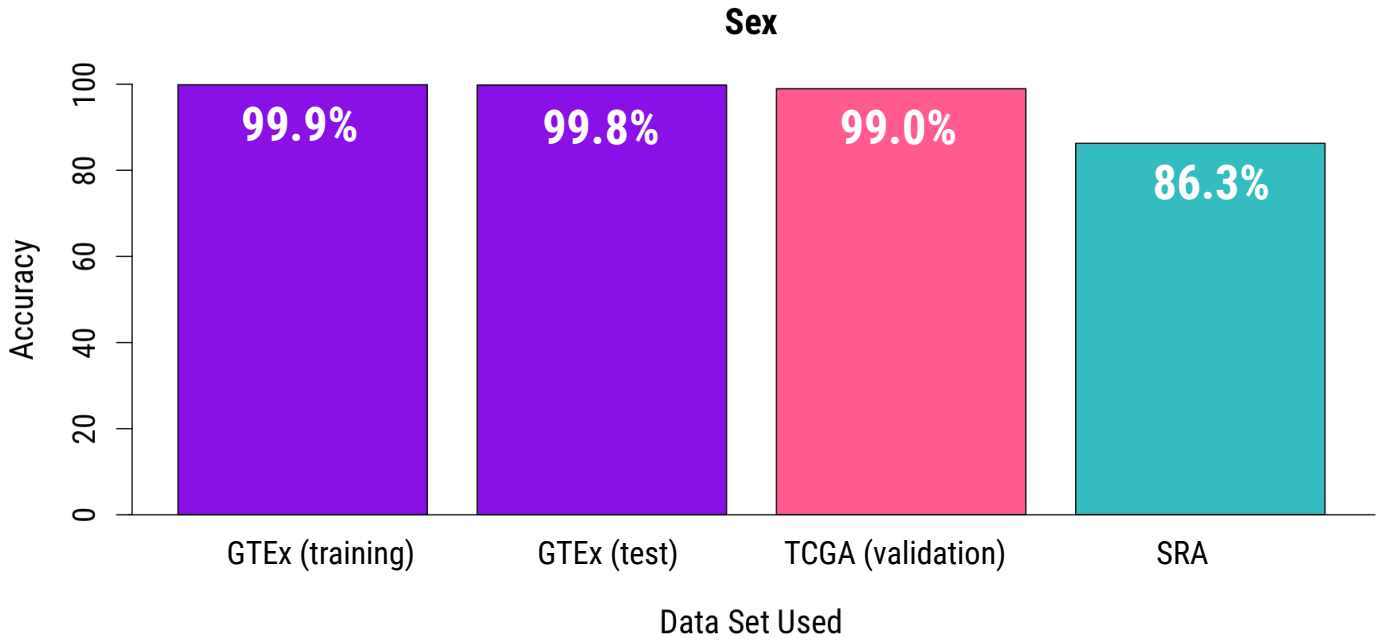
Let’s get predicting...

Sex prediction is accurate across data sets



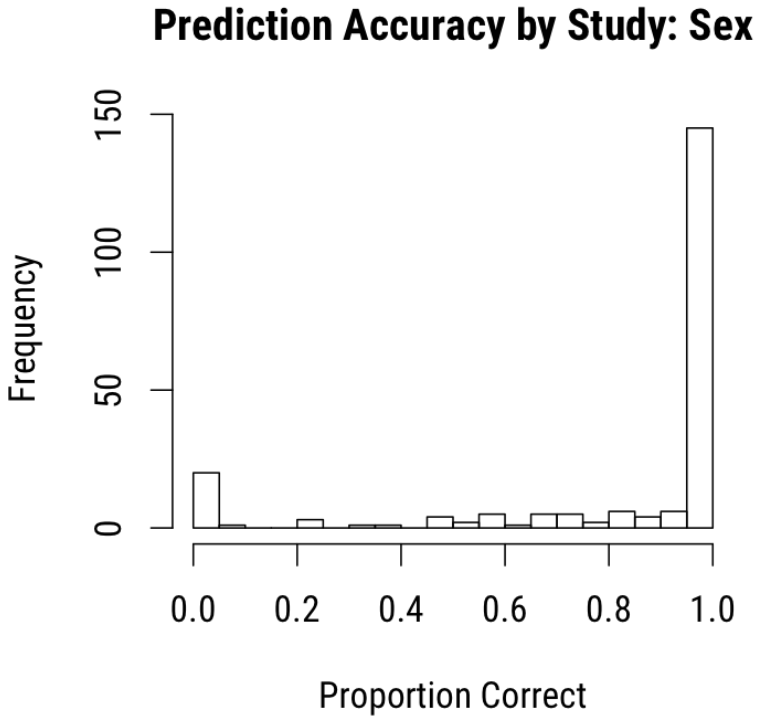
Number of Regions	**40**	**40**
Number of Samples (N)	**4,769**	**4,769**

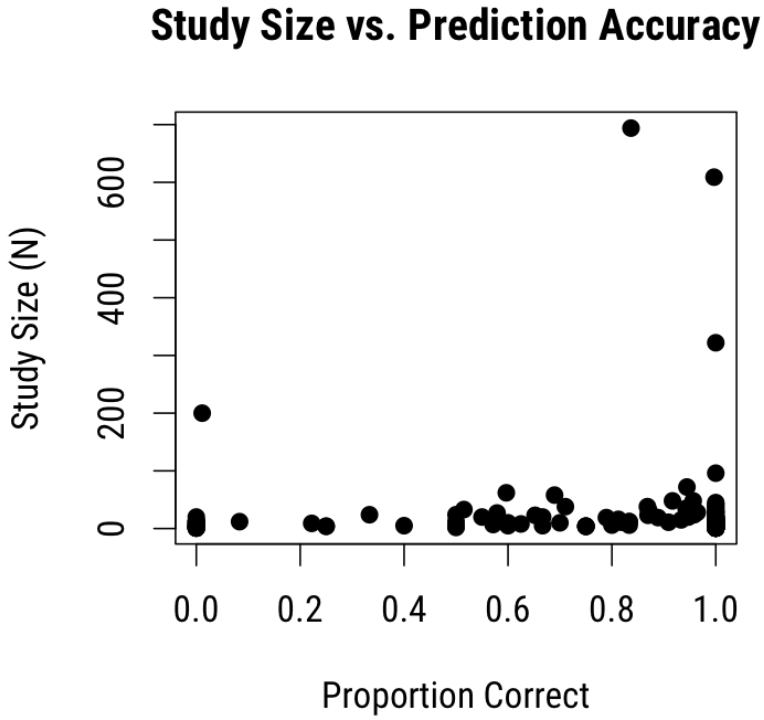
Sex prediction is accurate across data sets



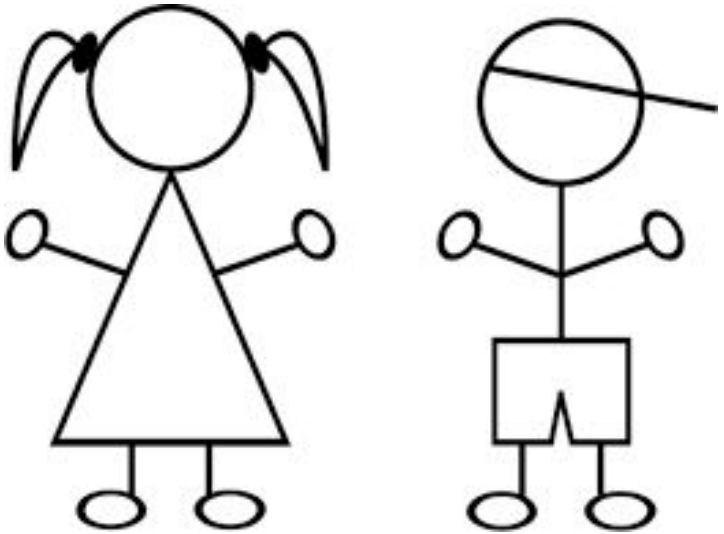
Number of Regions	40	40	40	40
Number of Samples (N)	4,769	4,769	11,245	3,640

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



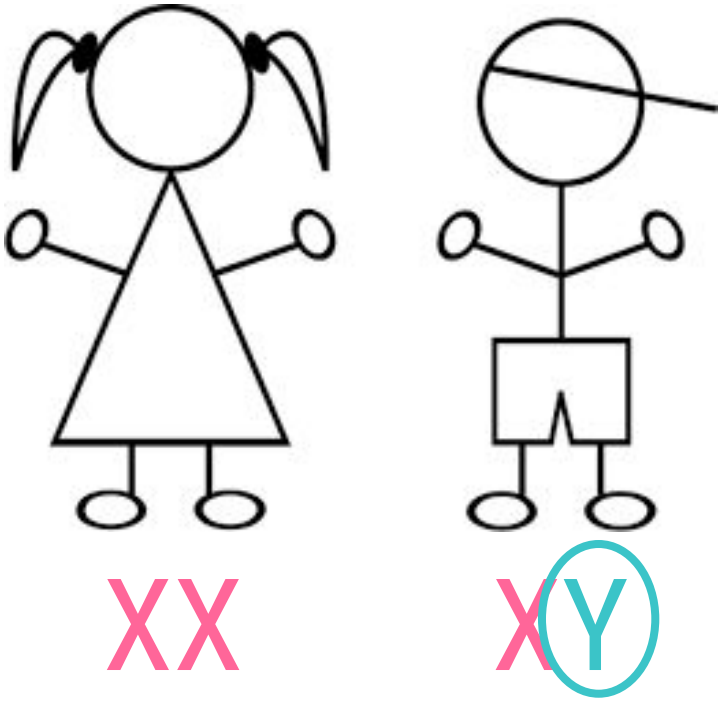


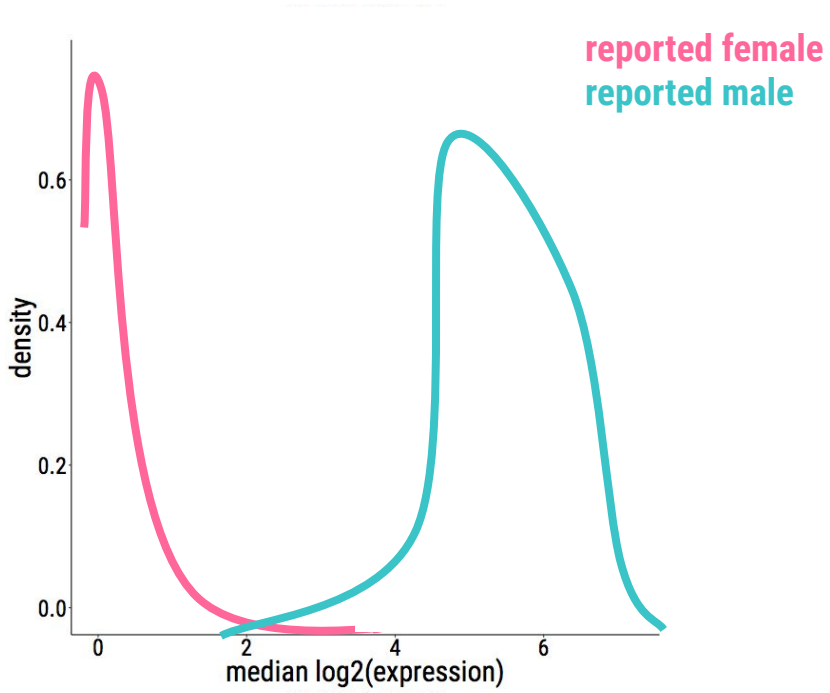
To assess misreporting of sex in the SRA, we can use Y-chromosome expression



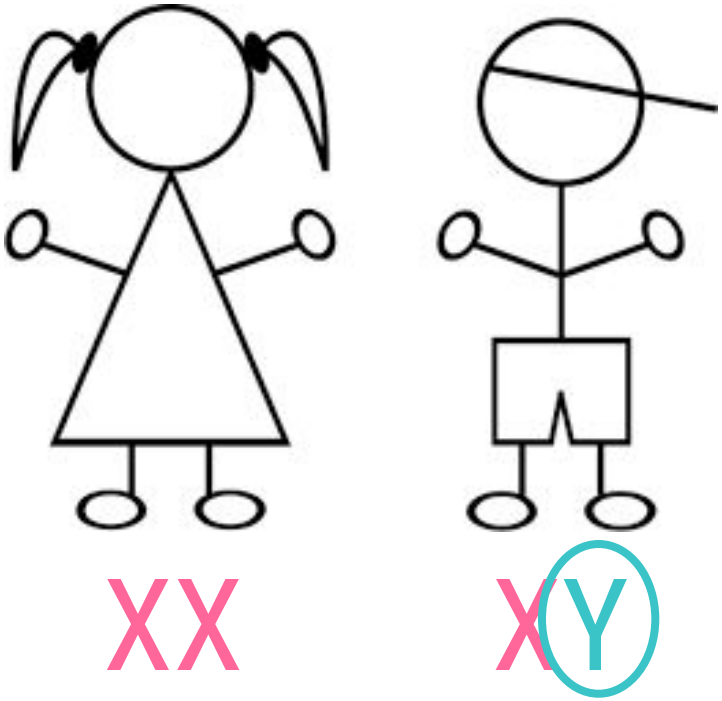
XX XY

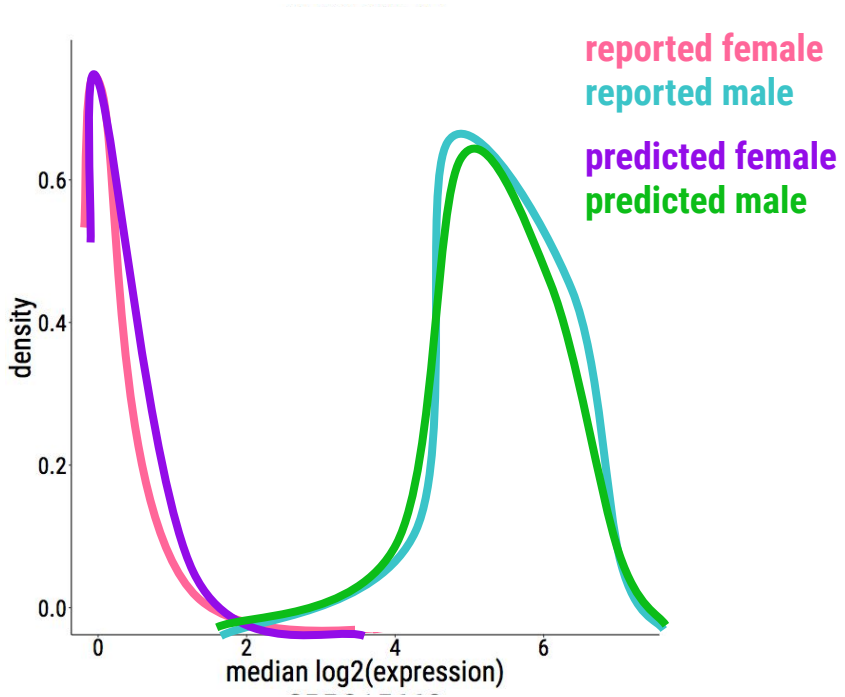
To assess misreporting of sex in the SRA, we can use Y-chromosome expression



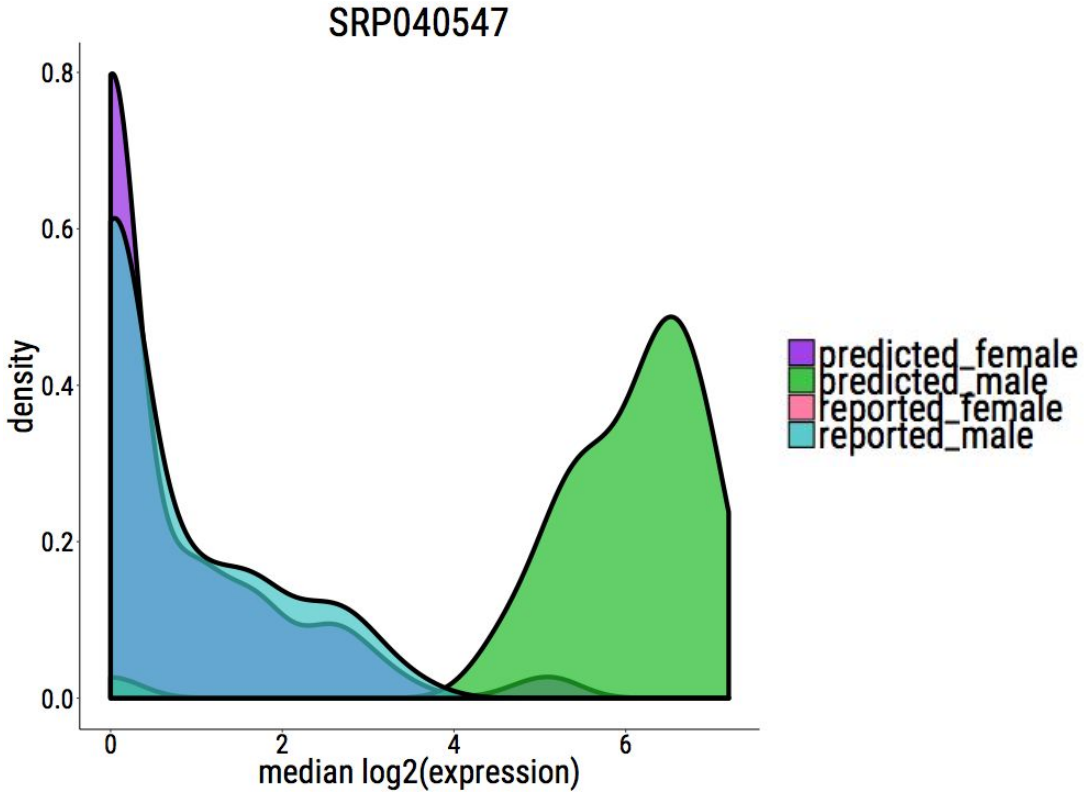


To assess misreporting of sex in the SRA, we can use Y-chromosome expression

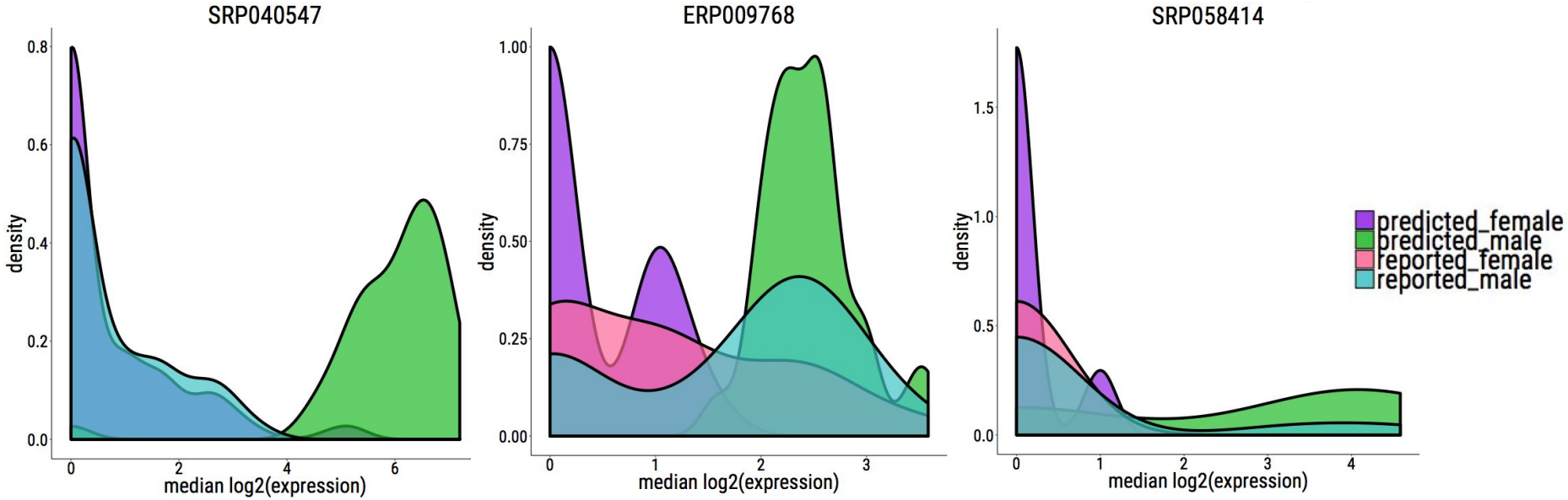


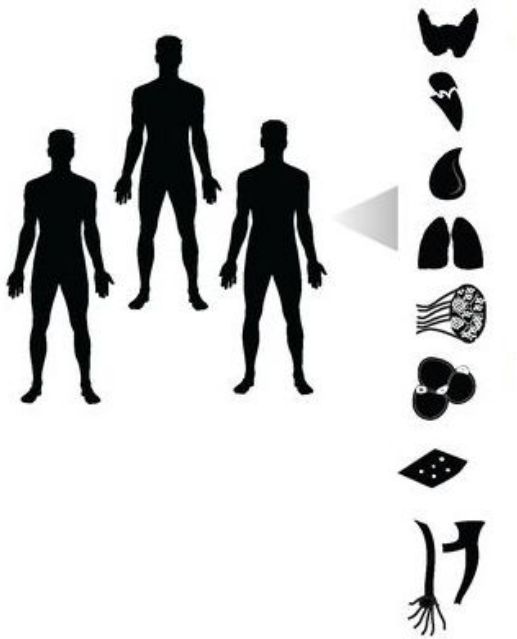


Expression from the Y chromosome suggests misreporting of sex in the SRA



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

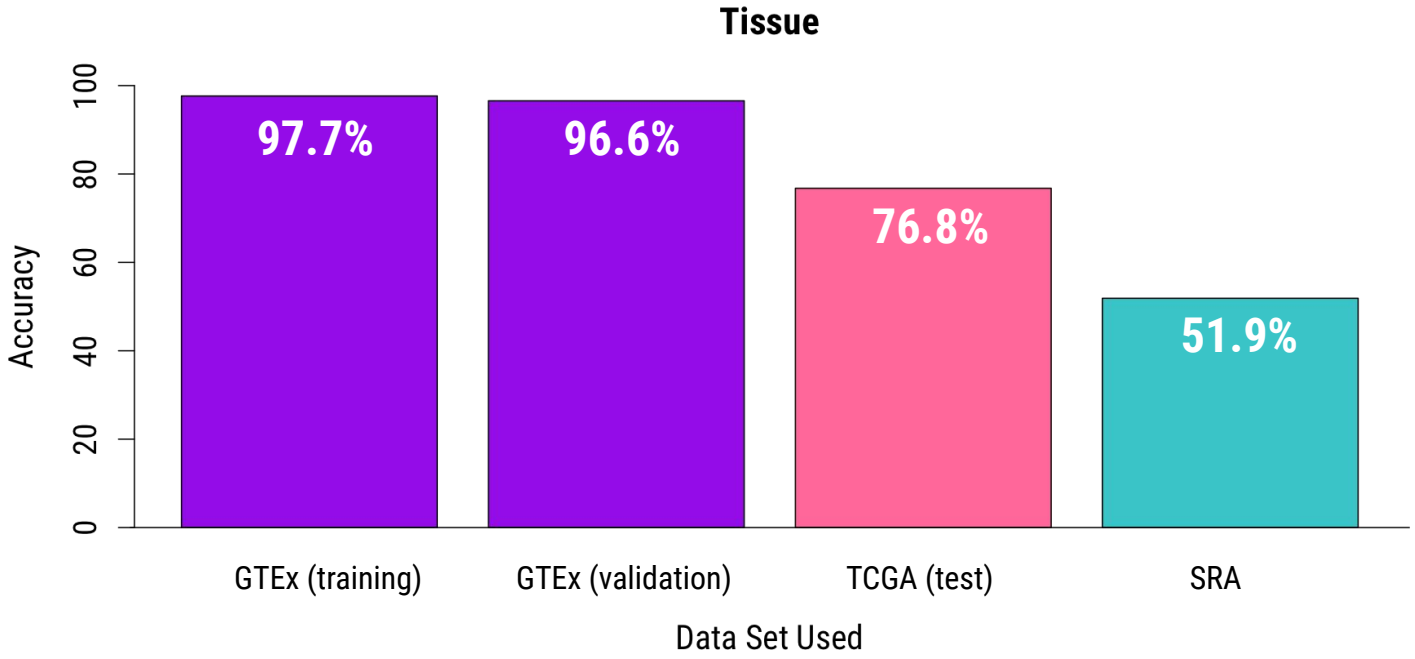




http://www.rna-seqblog.com/

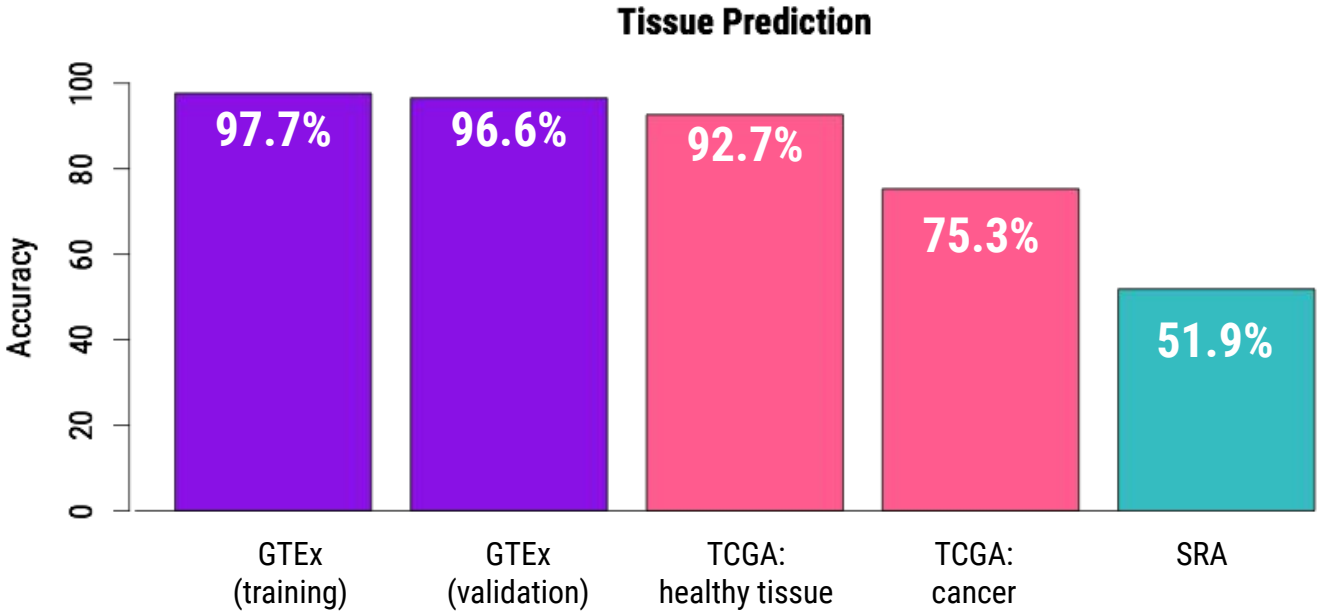
Can we use expression data to predict tissue?

Tissue prediction is accurate across data sets



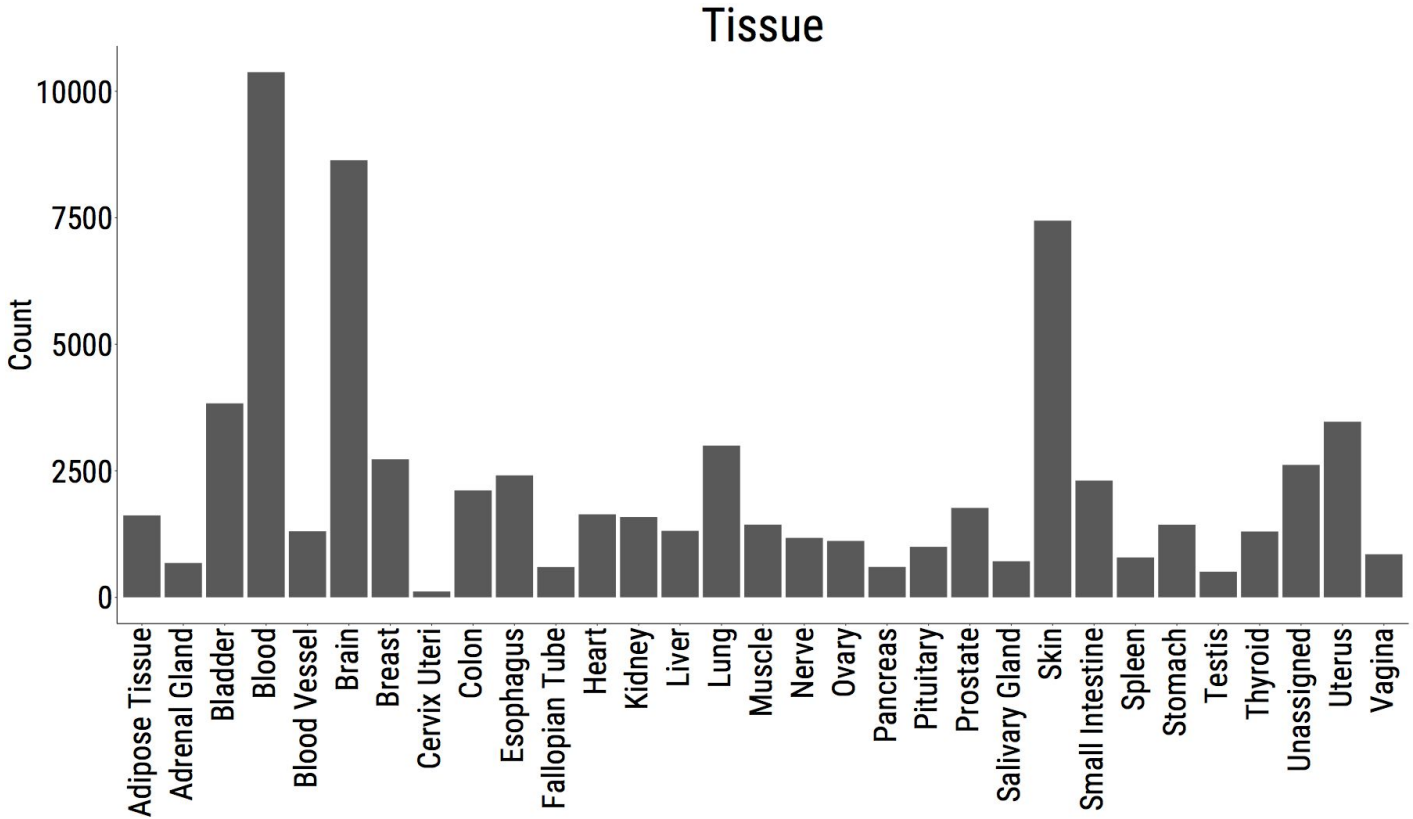
Number of Regions	2,281	2,281	2,281	2,281
Number of Samples (N)	4,769	4,769	7,317	8,951

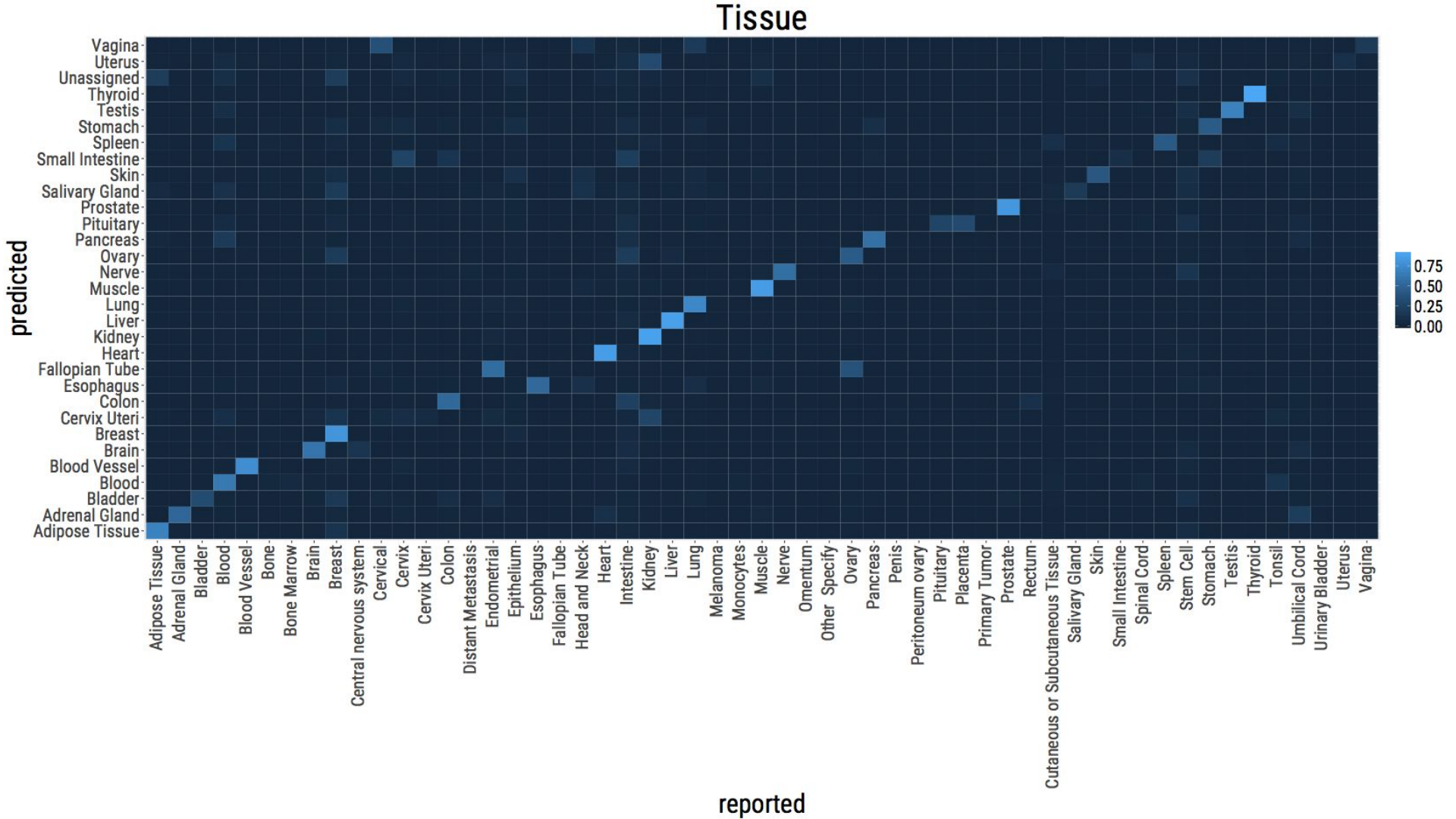
Prediction is more accurate in healthy tissue

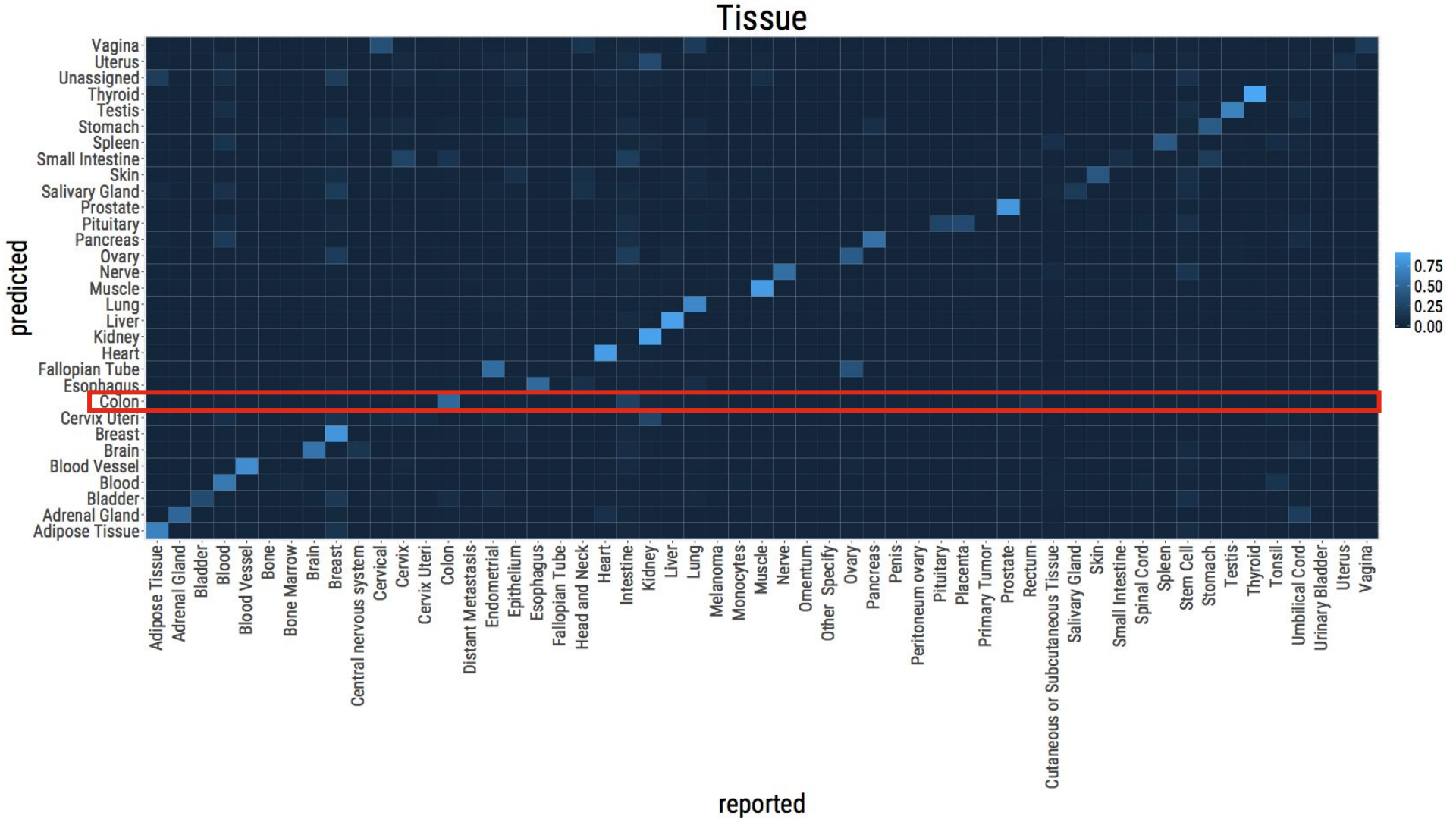


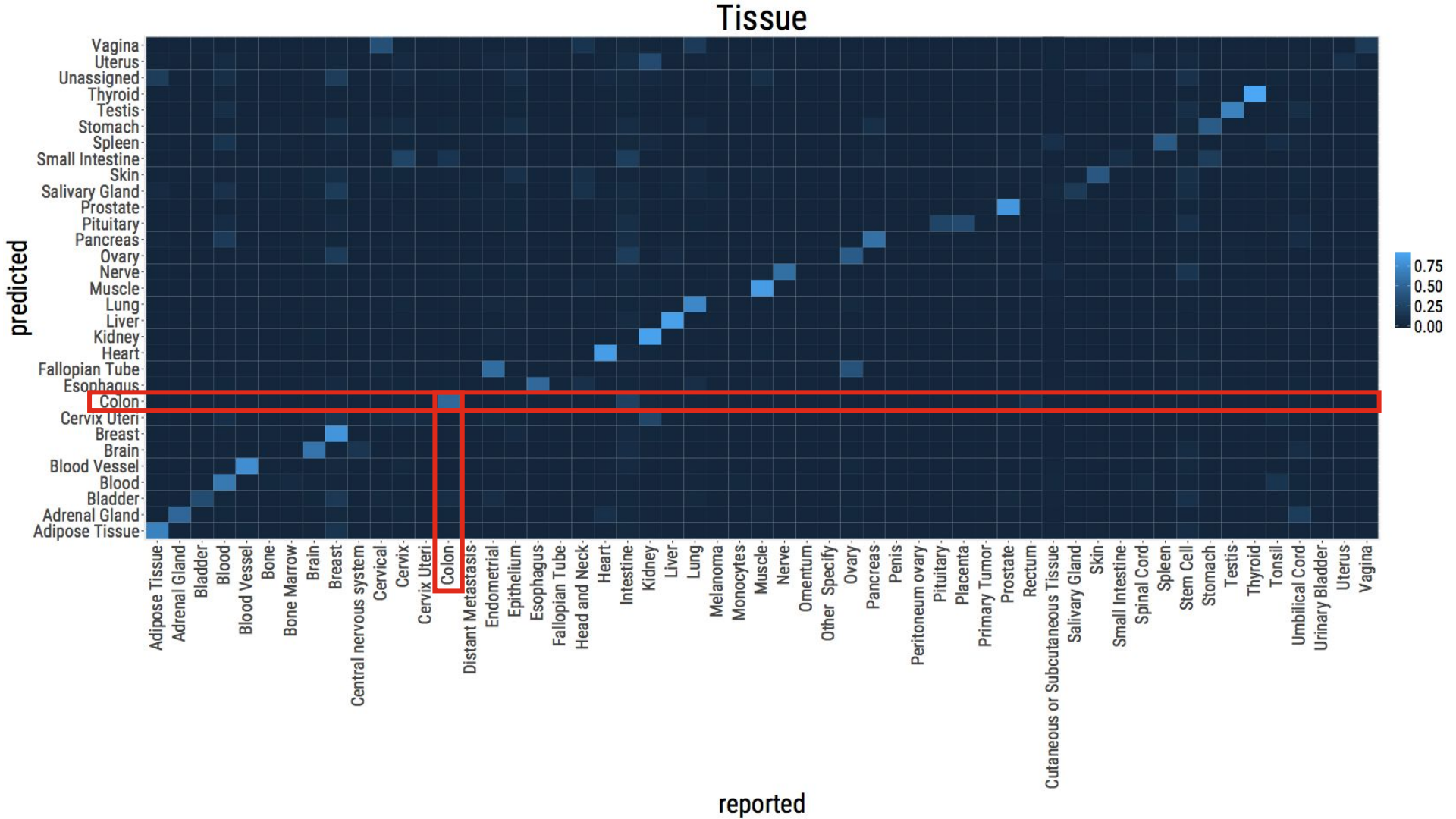
Number of Regions	2,281	2,281	2,281	2,281	2,281
Number of Samples (N)	4,769	4,769	613	6,704	8,951

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

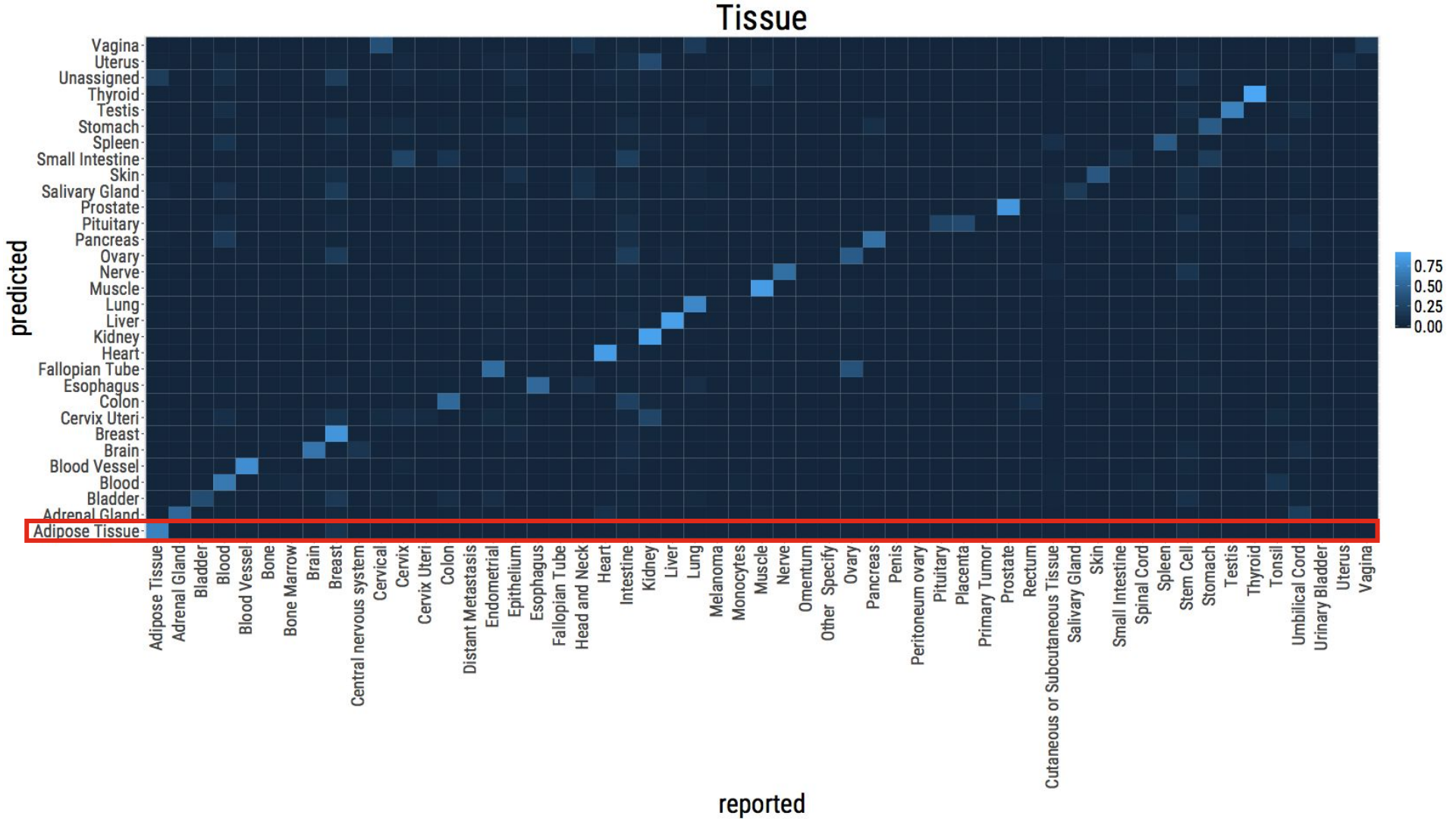


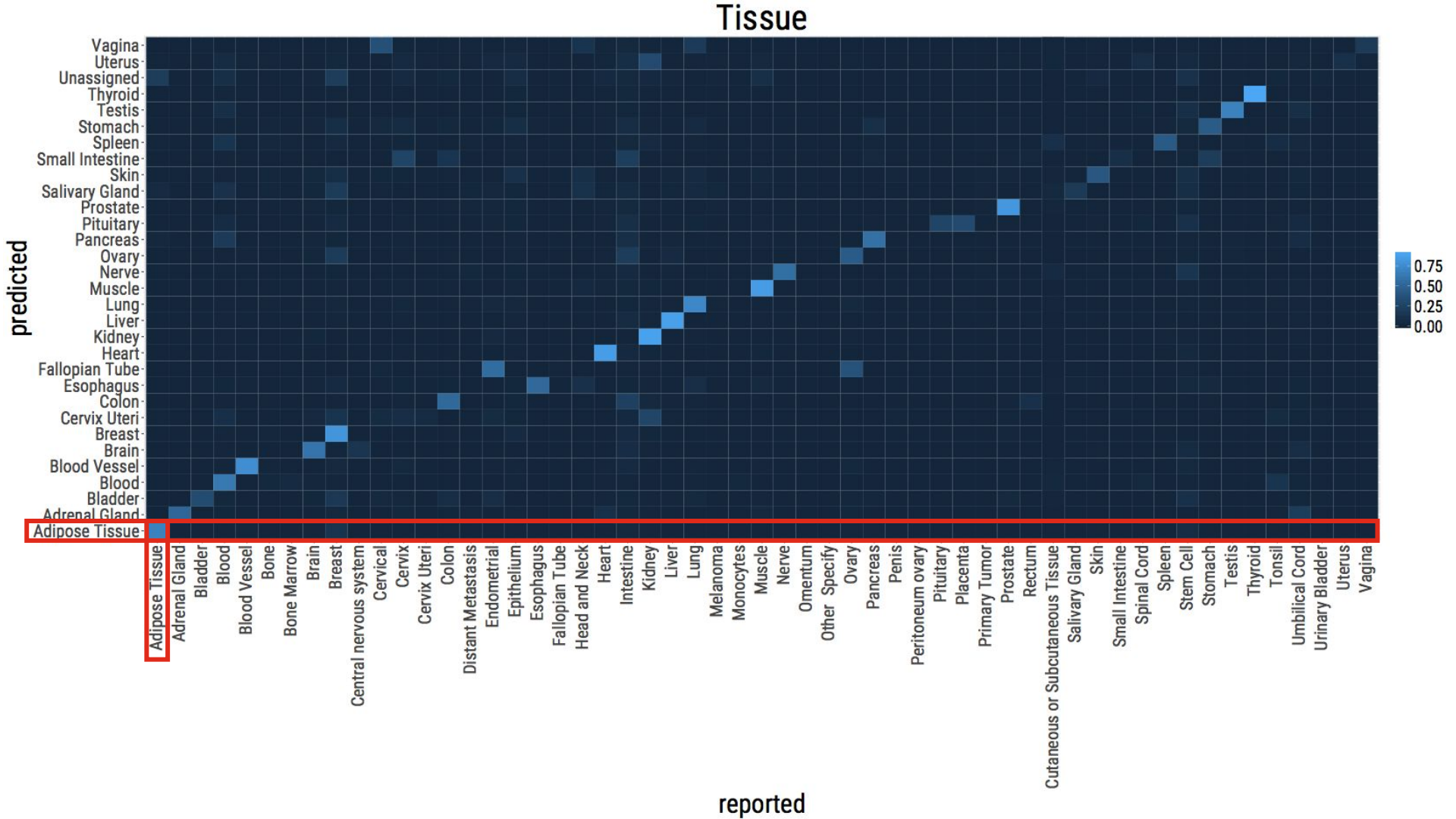

Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
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

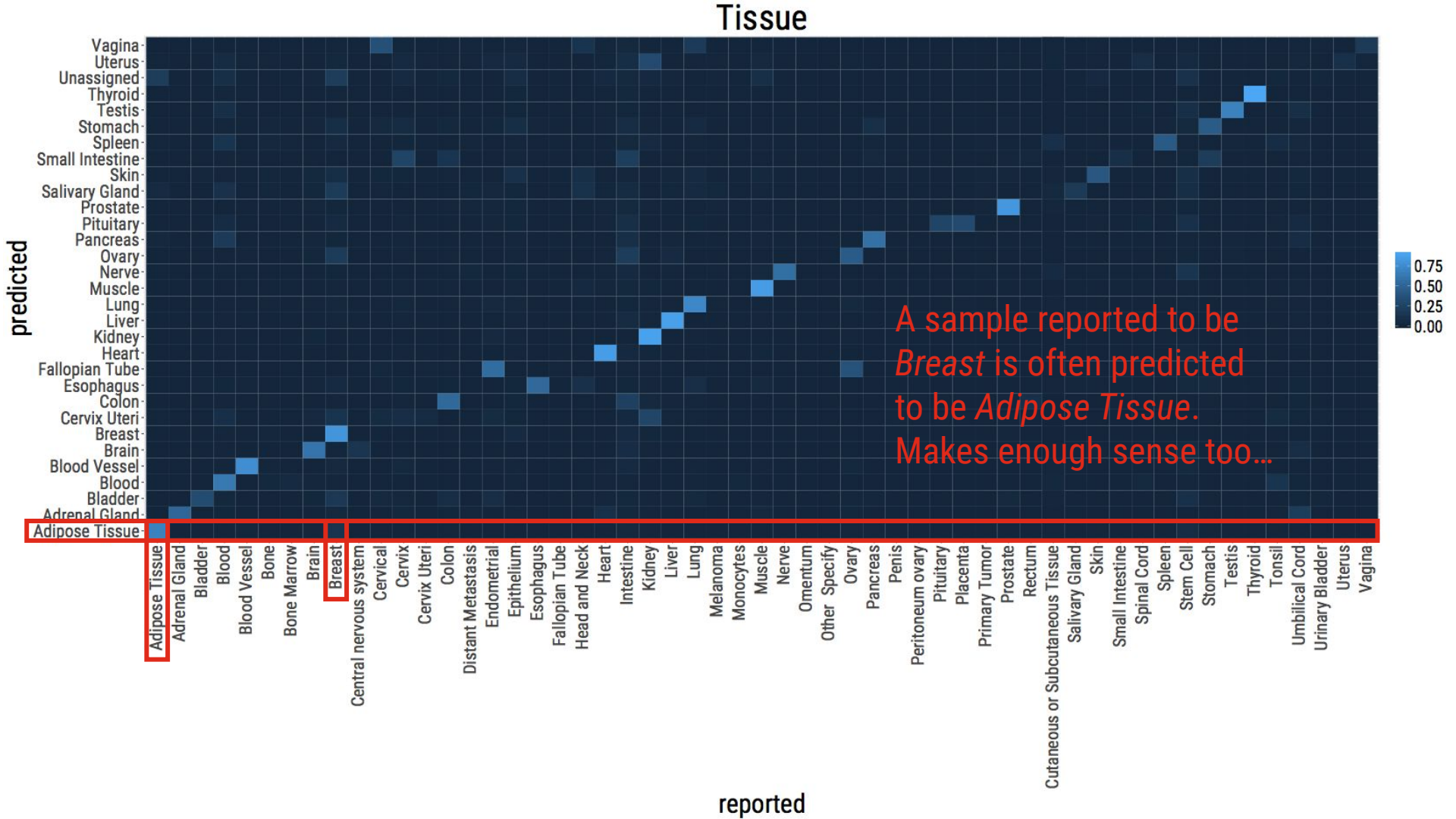

Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
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

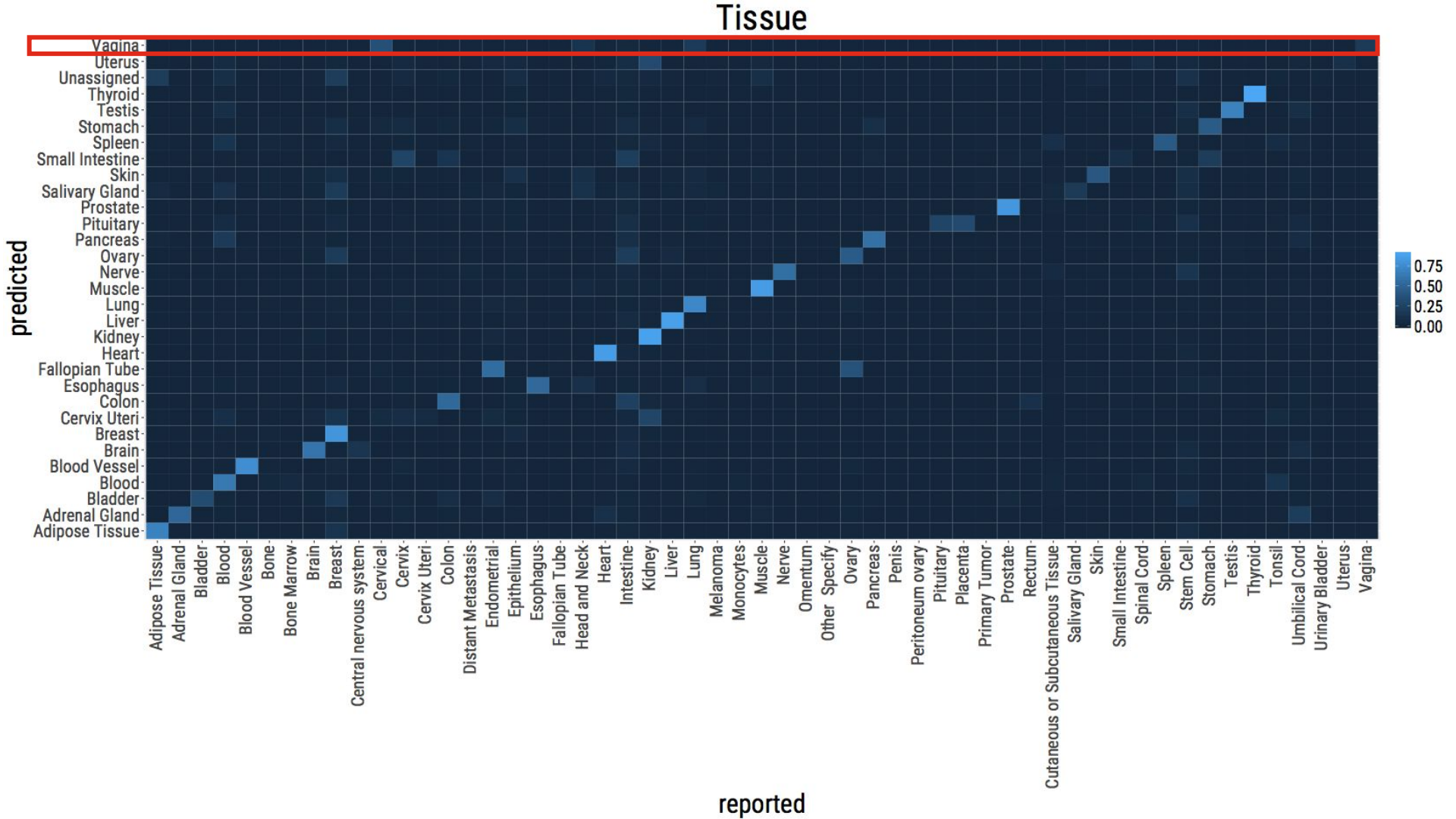

Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
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

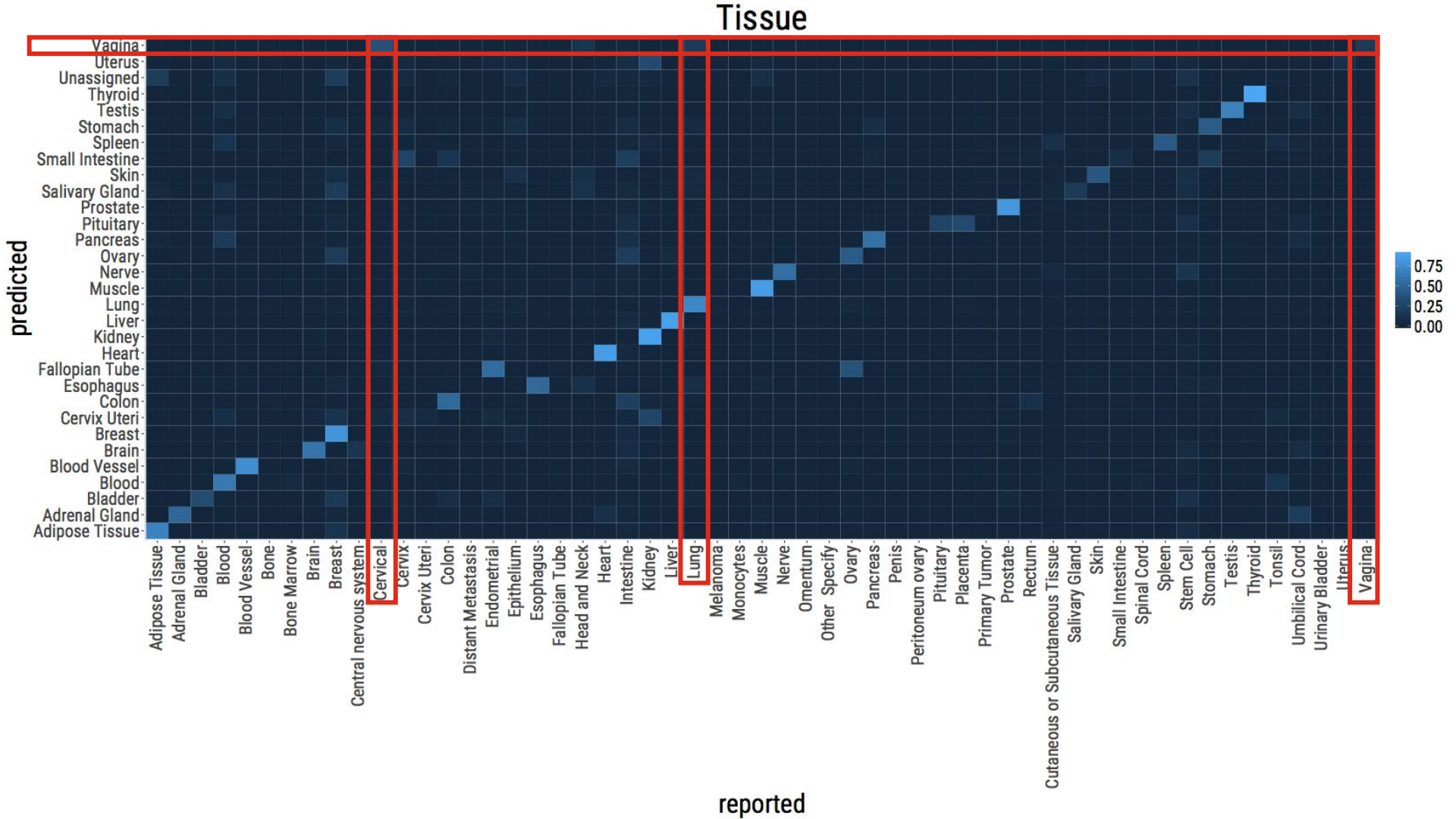

Tissue
A sample reported to be *Intestine* is predicted to be *Colon*. That makes good sense.
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
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

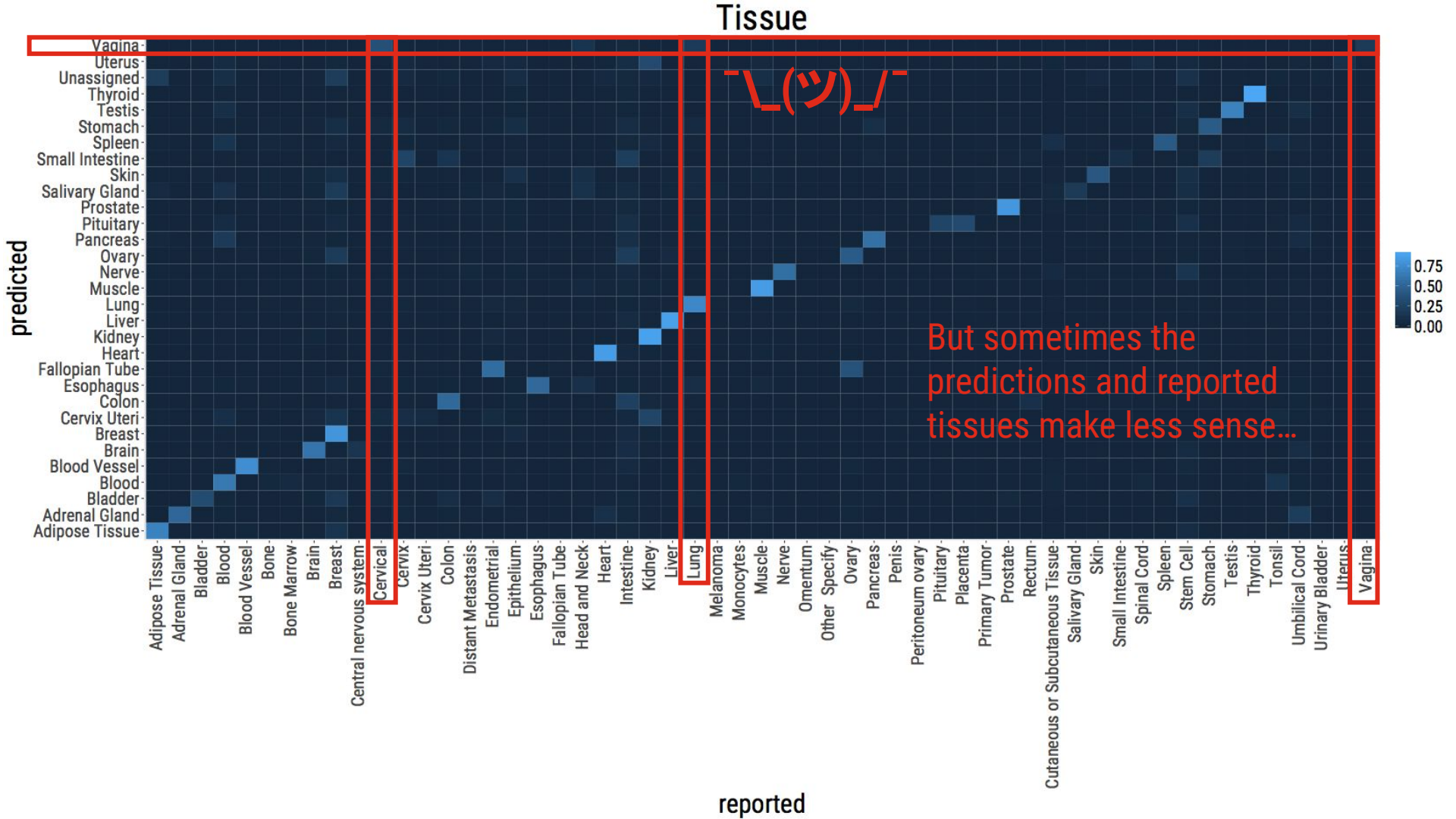

Tissue
predicted
reported
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
Colon
Cervix Uteri
Breast
Brain
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
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
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
A sample reported to be
Breast is often predicted
to be *Adipose Tissue*.
Makes enough sense too…


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

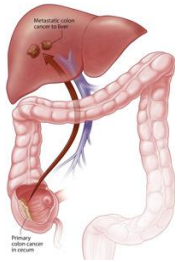

Tissue
predicted
reported
¯_(ツ)_/¯
But sometimes the predictions and reported tissues make less sense…
0.75
0.50
0.25
0.00
Vagina
Uterus
Unassigned
Thyroid
Testis
Stomach
Spleen
Small Intestine
Skin
Salivary Gland
Prostate
Pituitary
Pancreas
Ovary
Nerve
Muscle
Lung
Liver
Kidney
Heart
Fallopian Tube
Esophagus
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

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

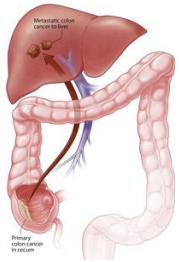
3~6 months	1-2 wks	2-4 wks	3-6 months	1-3 months	1 month - 1+ years

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**

Determine expression differences between primary cancer and metastatic cancer

Collect patient samples and information

Sequence samples

Process sequencing data

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

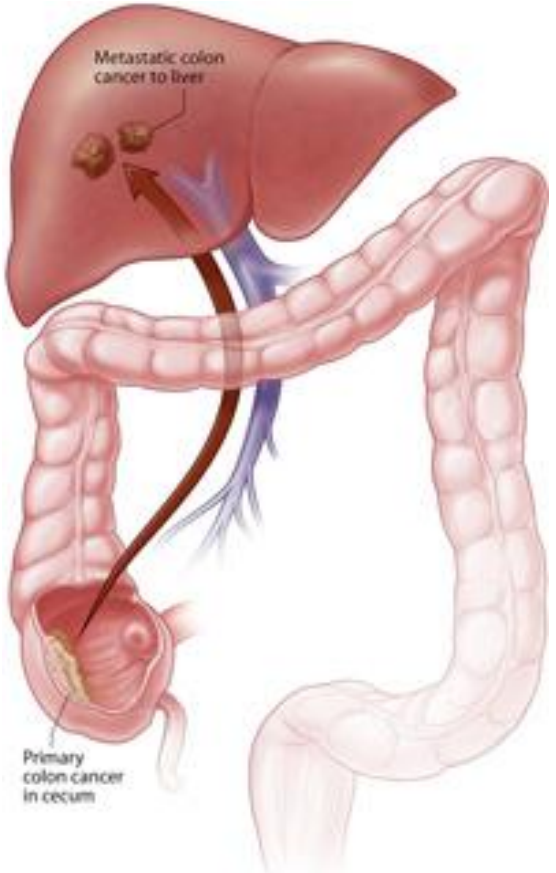
Total: months

Project Goal: Take publicly available RNA-Seq data and make it available and easy-to-use

- [x] Comprehensive
- [x] Easy to Get
- [x] Useful for future study

Ok. Ok. What about actually *using* all of these predictions…?

What makes primary cancer different than metastatic cancer?



www.hopkinsmedicine.org

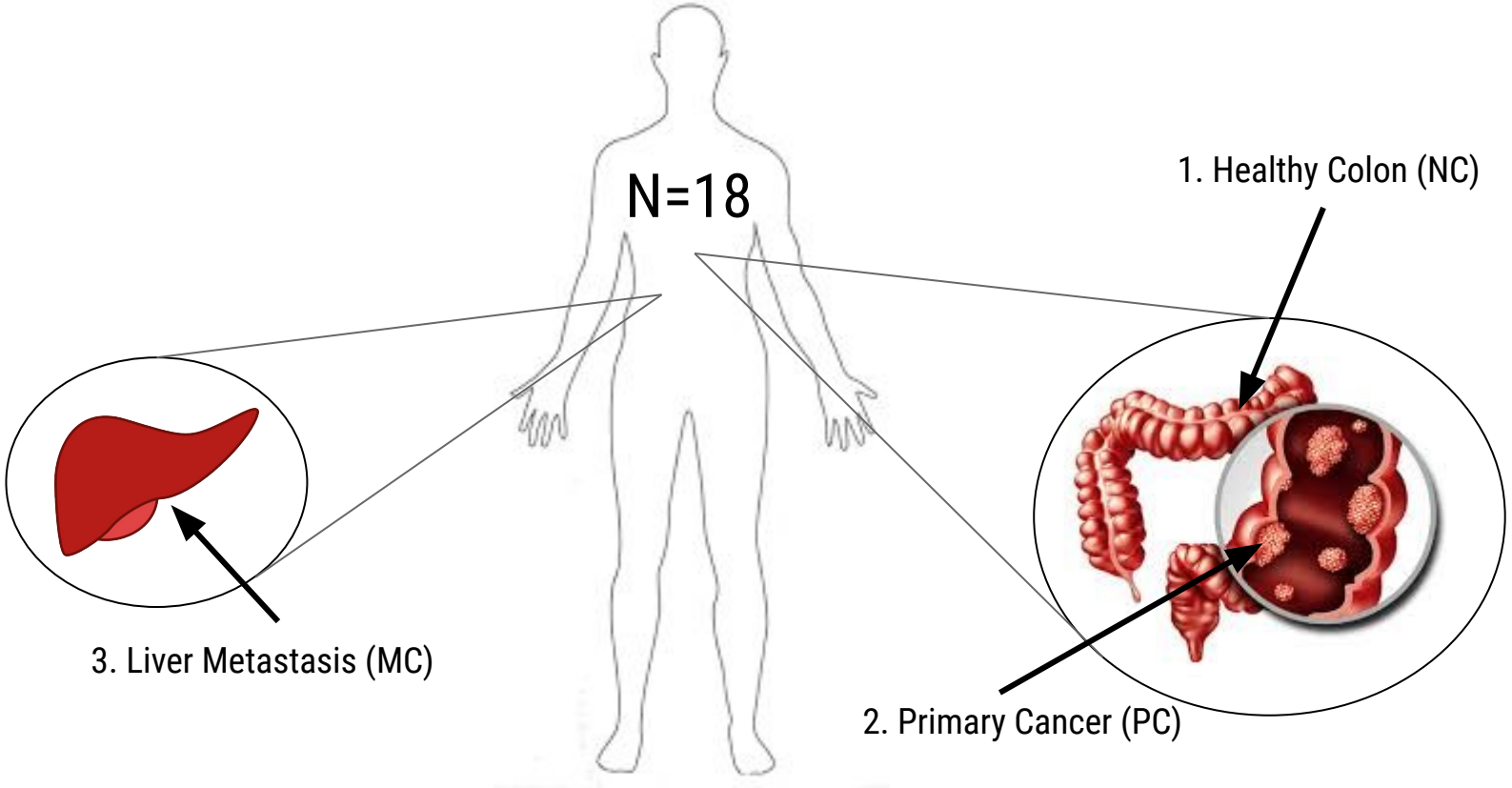
A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients

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,*]

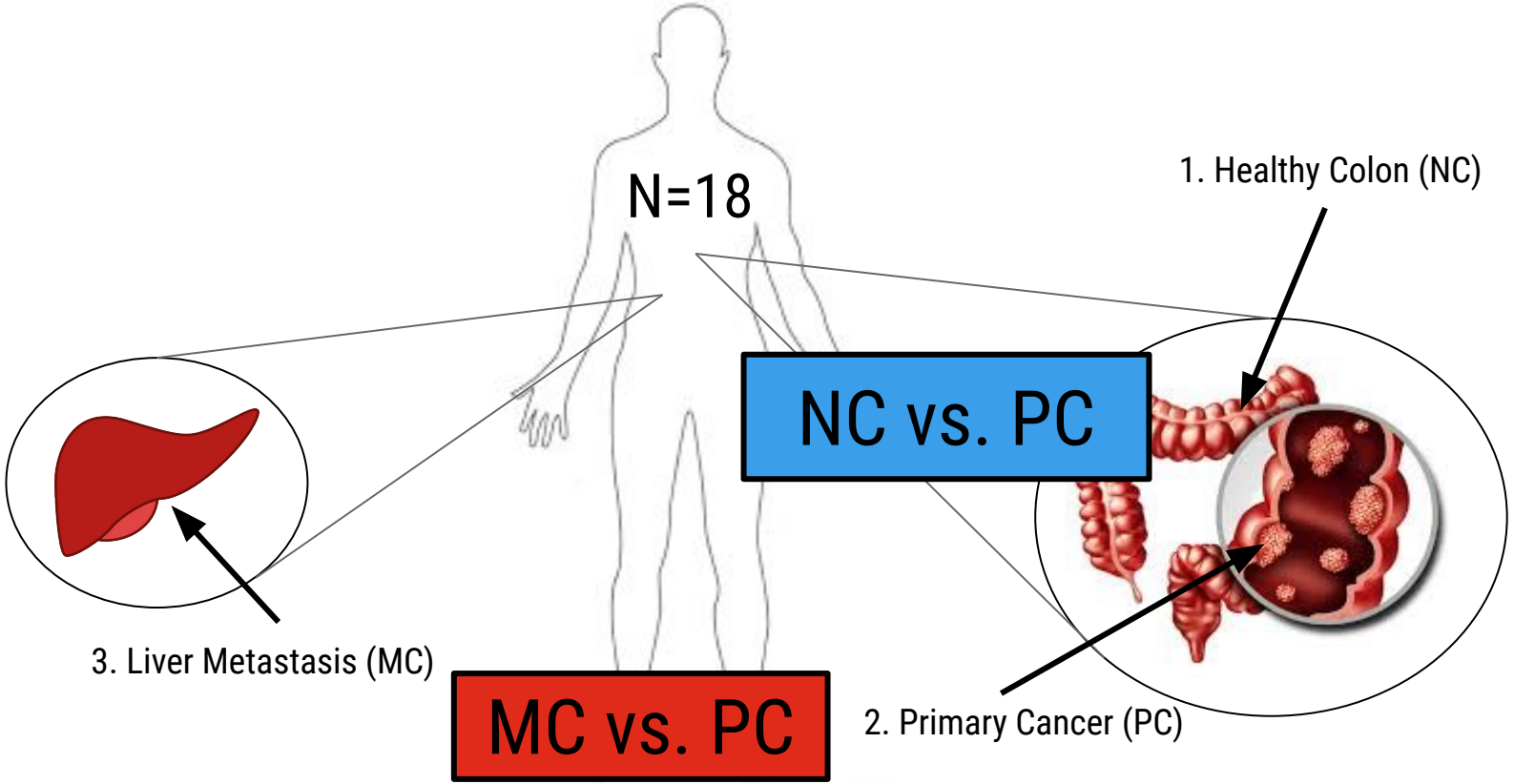
[a]Medical Genomics Research Centre, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea
[b]Department of Surgery, University of Ulsan College of Medicine, Seoul, Korea
[c]Department of Cancer Research, Institute of Innovative Cancer Research and Asan Institute for Life Sciences, 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.



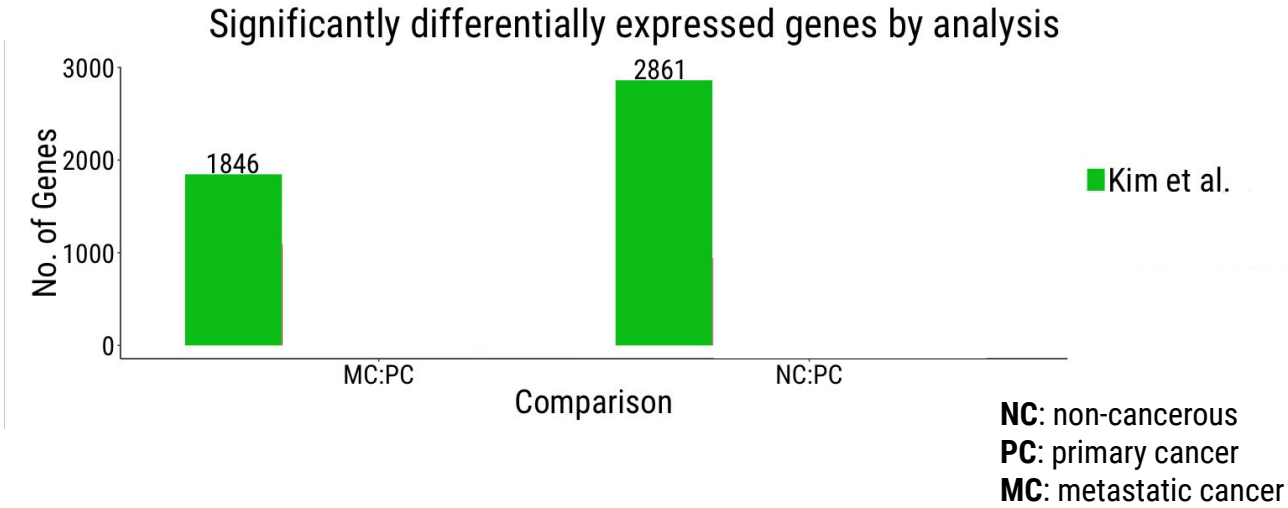
Kim et al. analysis looked to identify genes that contribute to metastasis in colon cancer.



Predictions can be used to:

(1) Identify studies of interest

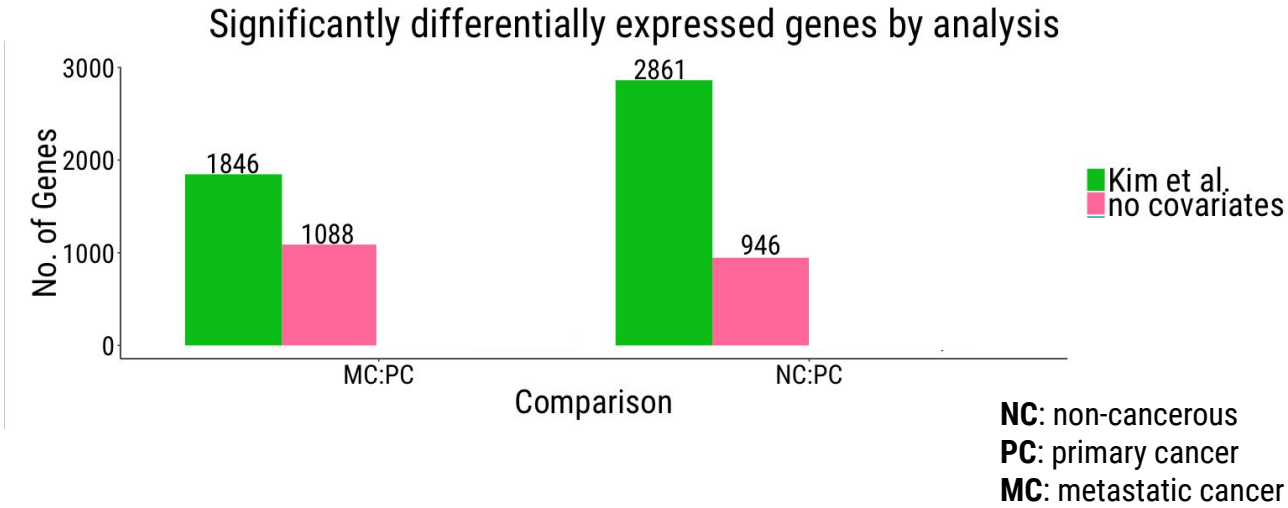
(2) appropriately analyze data



Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data

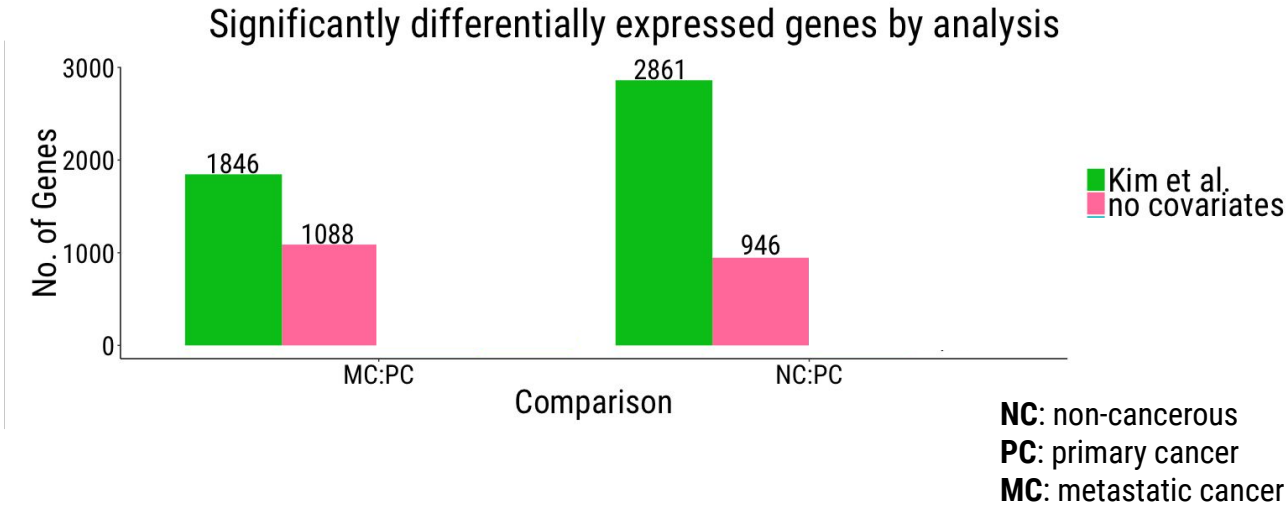


NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data



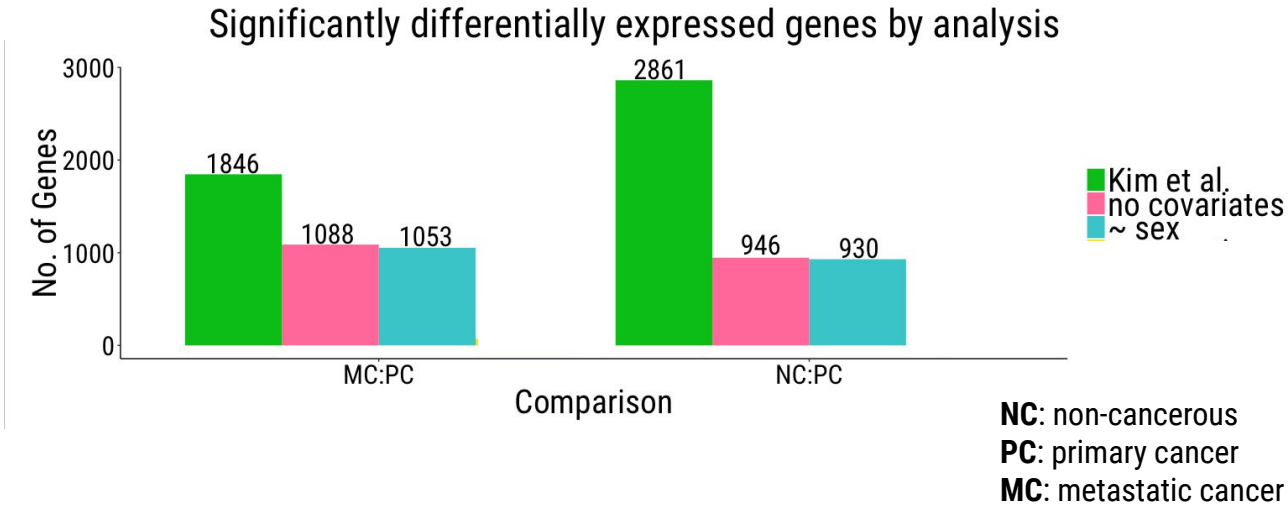
NC: non-cancerous
PC: primary cancer
MC: metastatic cancer

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

Predictions can be used to:

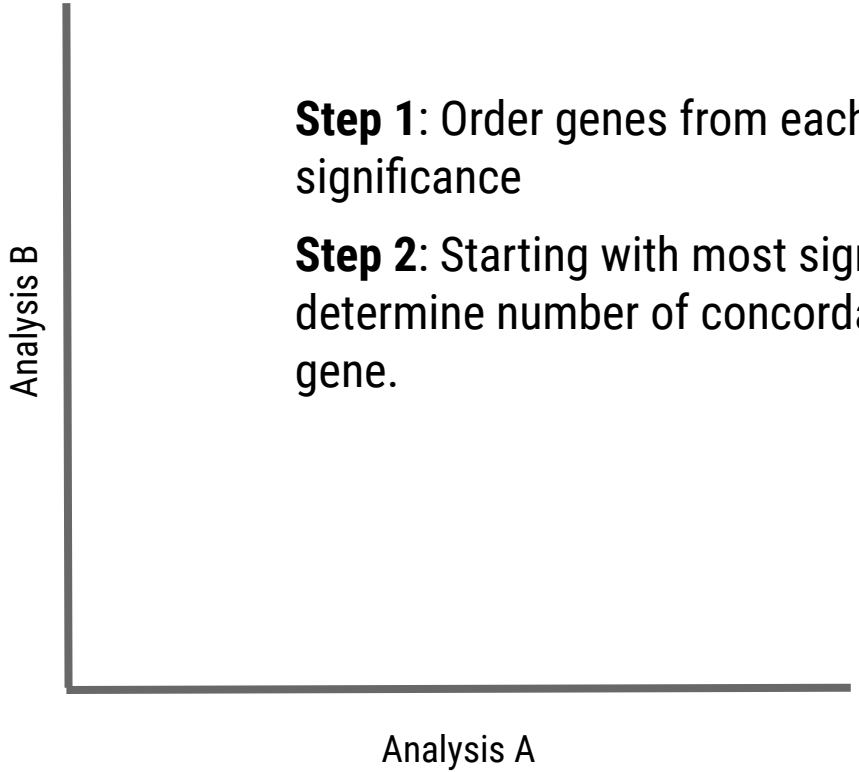
(1) Identify studies of interest

(2) appropriately analyze data



Concordance at the Top (CAT) Plots

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

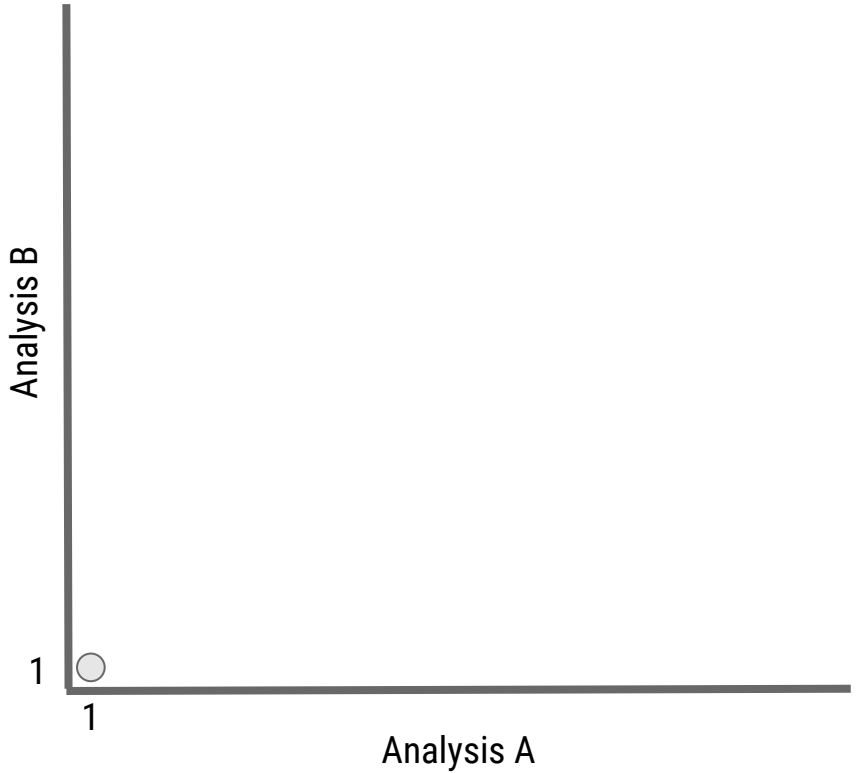


Step 1: Order genes from each analysis by significance

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

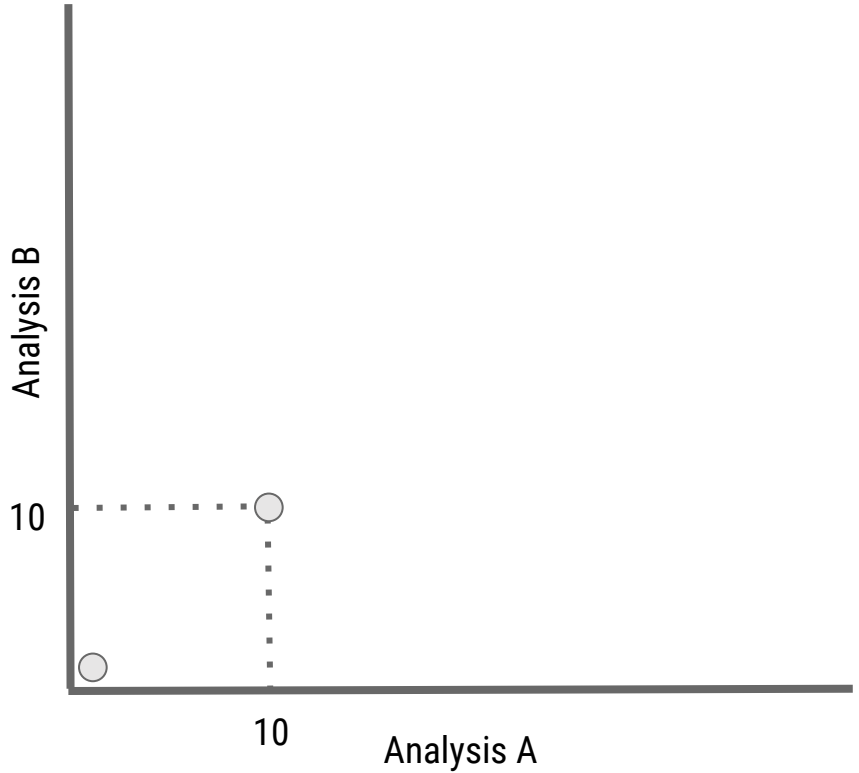
Concordance at the Top (CAT) Plots

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



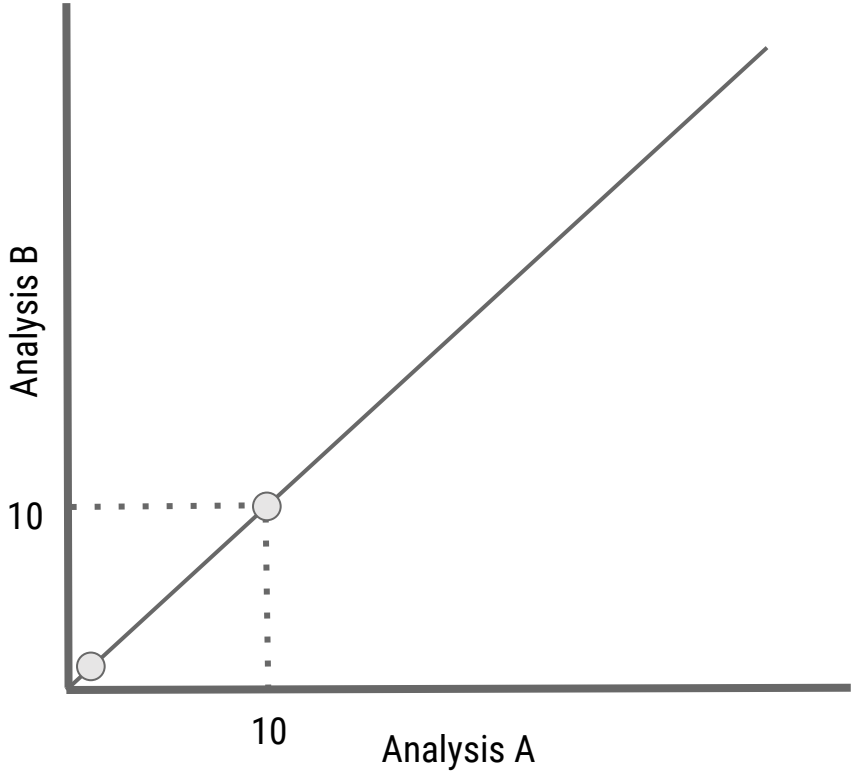
Concordance at the Top (CAT) Plots

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



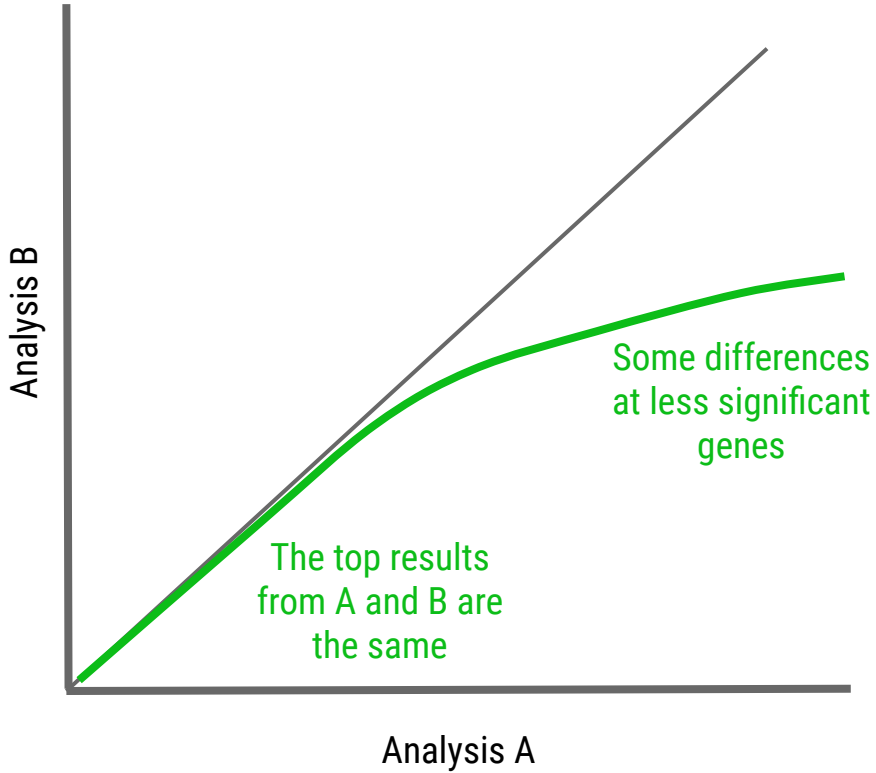
Concordance at the Top (CAT) Plots

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



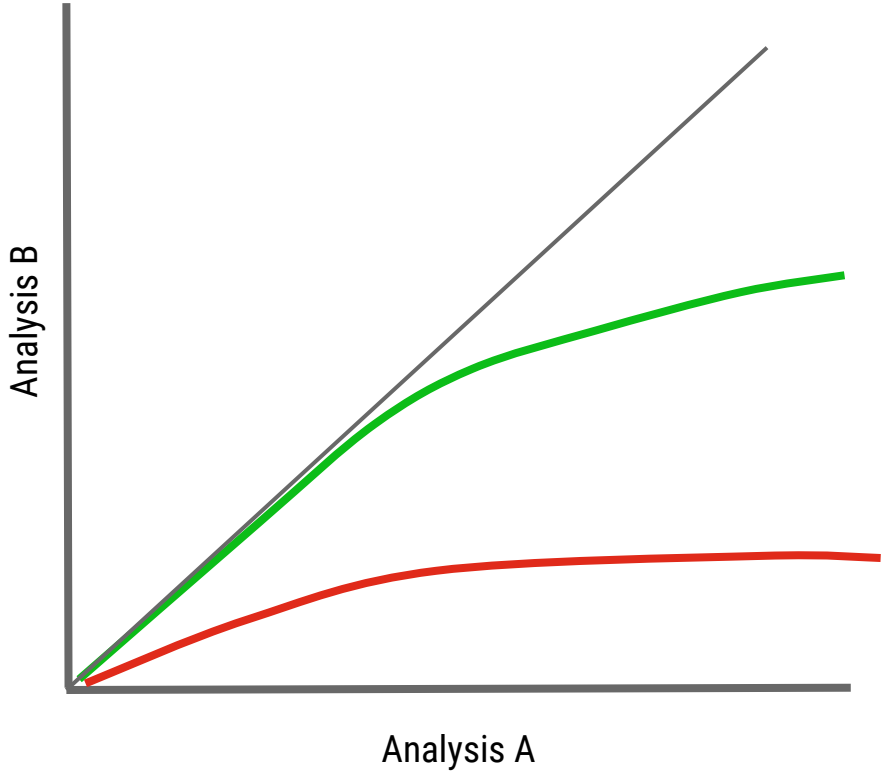
Concordance at the Top (CAT) Plots

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 ?

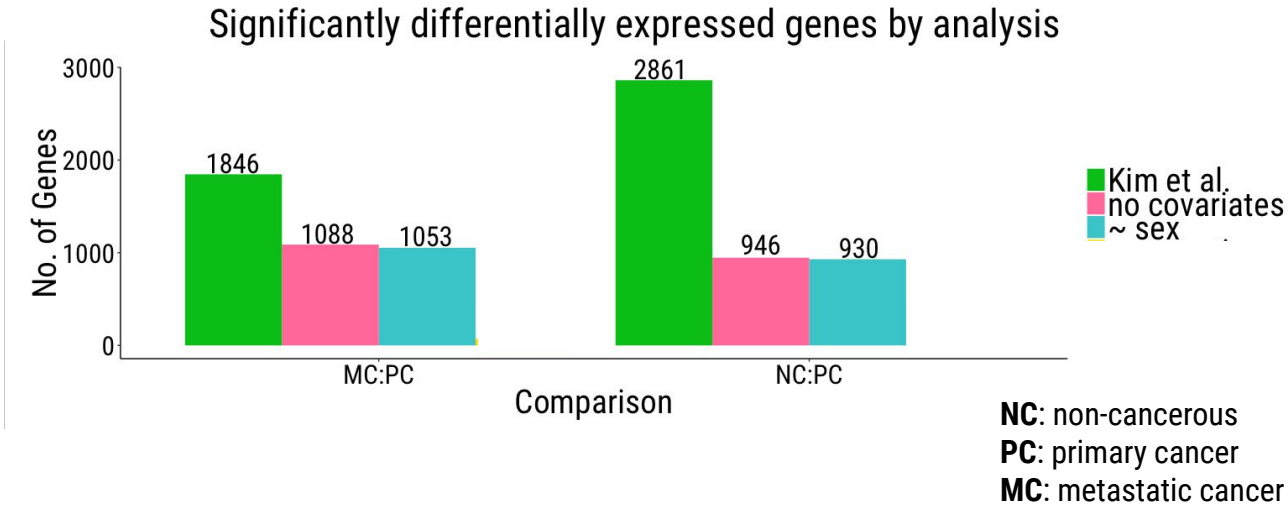


The results of the red condition are less similar between Analysis A and B than the green condition

Predictions can be used to:

(1) Identify studies of interest

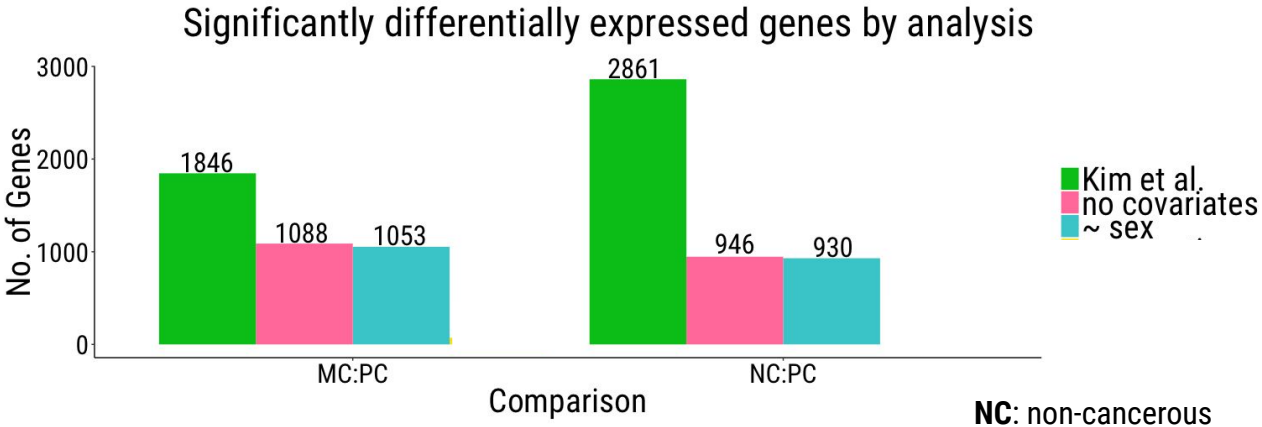
(2) appropriately analyze data



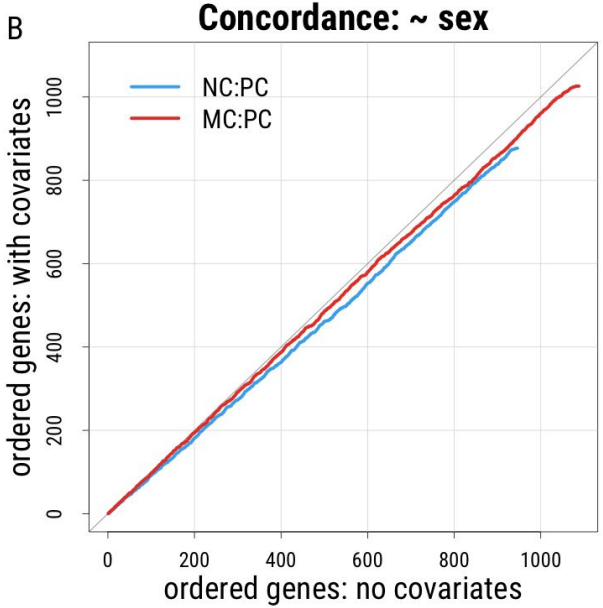
Predictions can be used to:

(1) Identify studies of interest

(2) appropriately analyze data



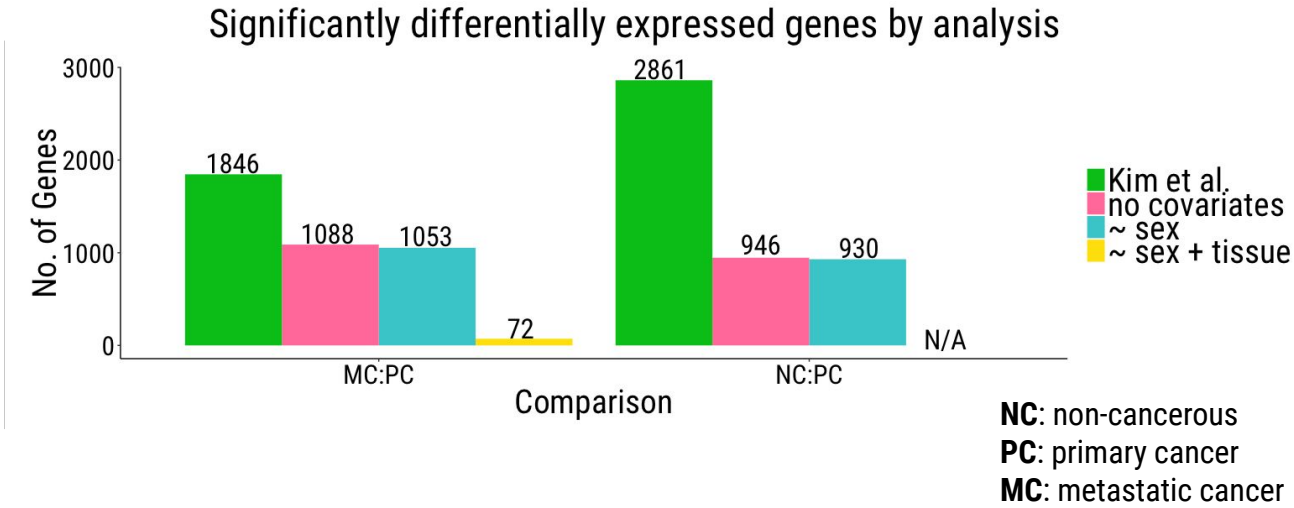
NC: non-cancerous
PC: primary cancer
MC: metastatic cancer



Predictions can be used to:

(1) Identify studies of interest

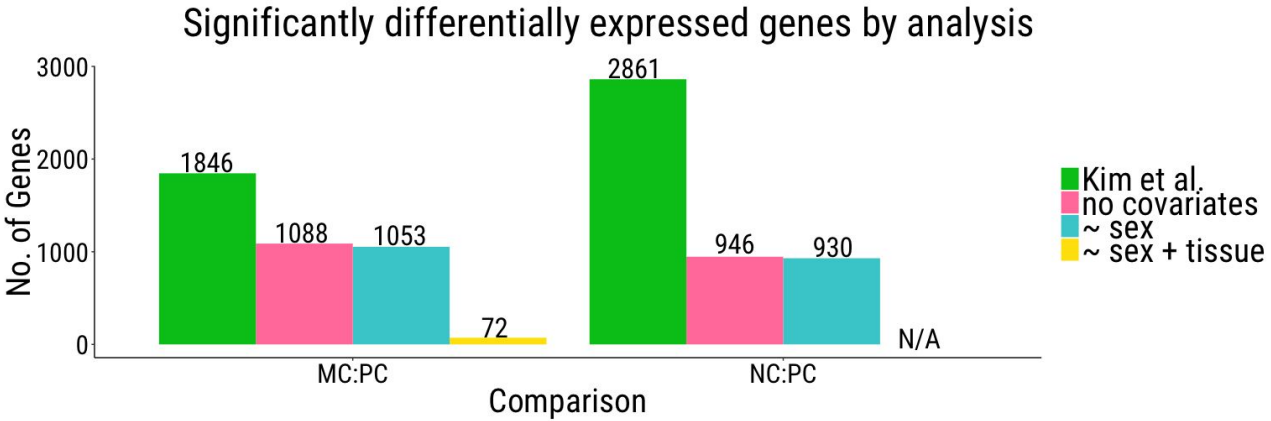
(2) appropriately analyze data

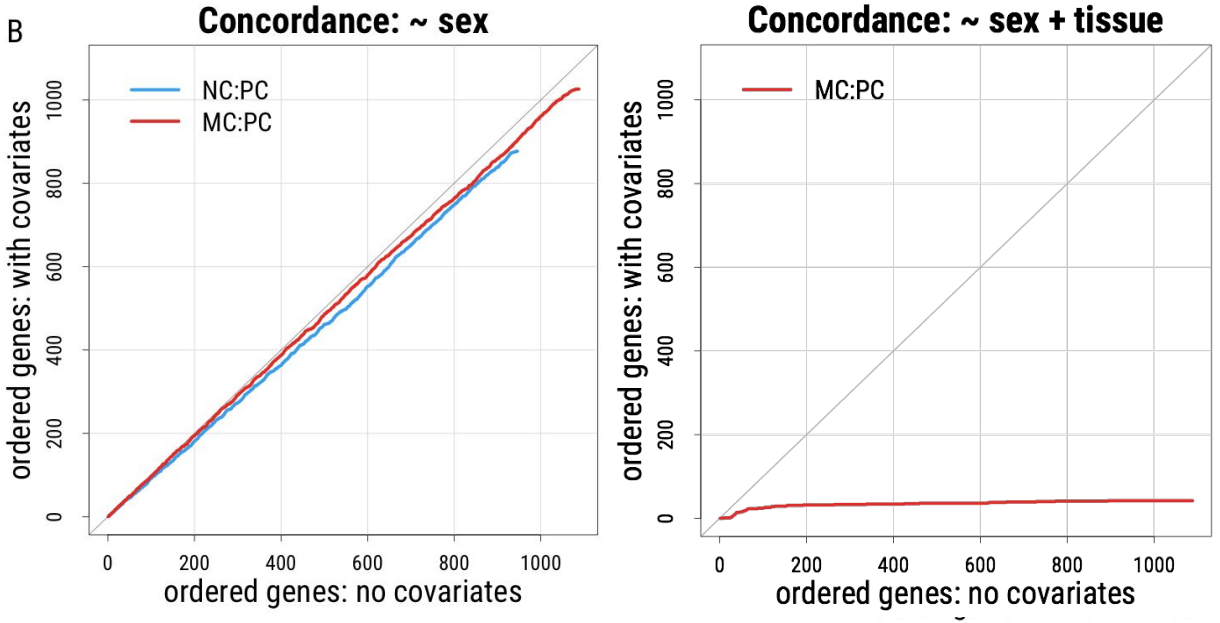


Predictions can be used to:

(1) Identify studies of interest

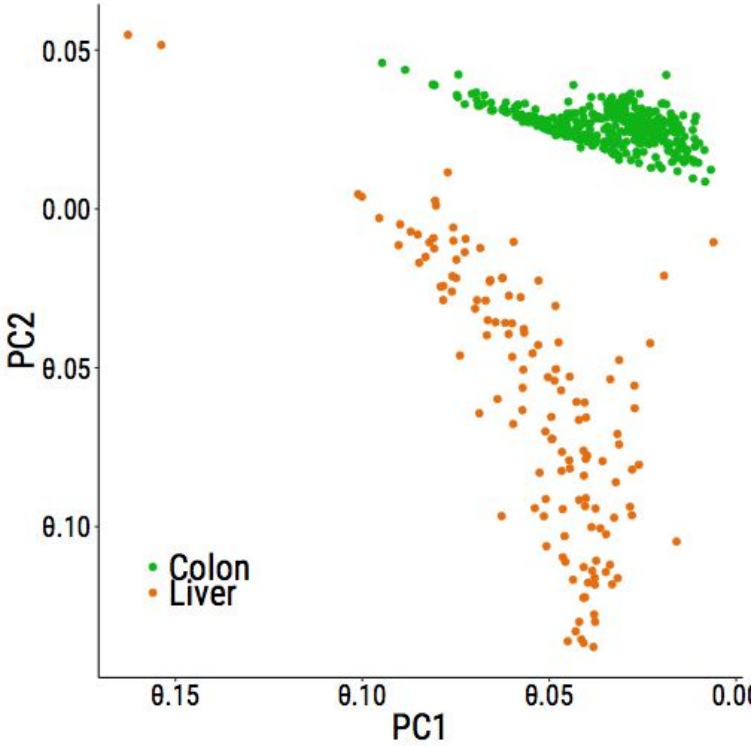
(2) appropriately analyze data



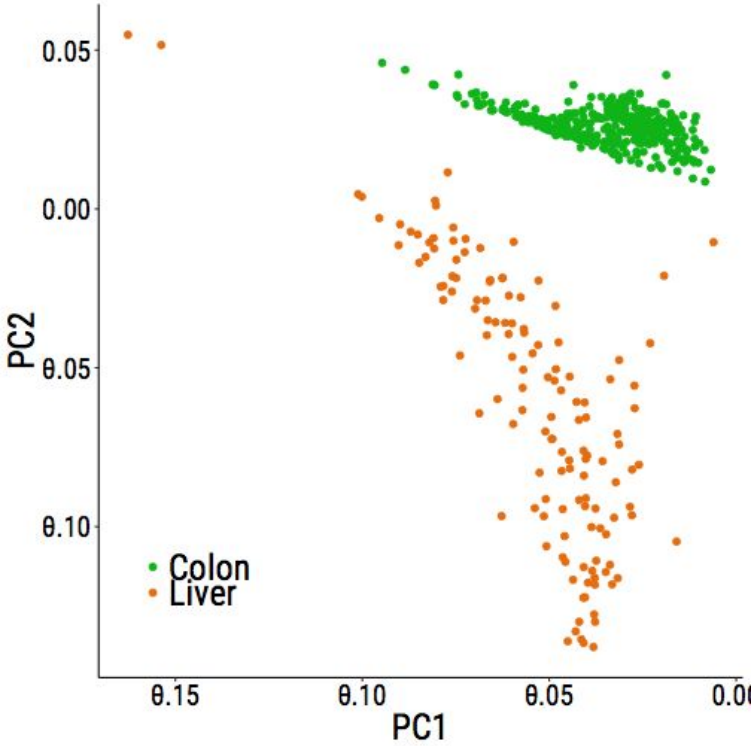


Loss of concordance suggests that differential expression is detecting tissue differences, not cancer-related changes.

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

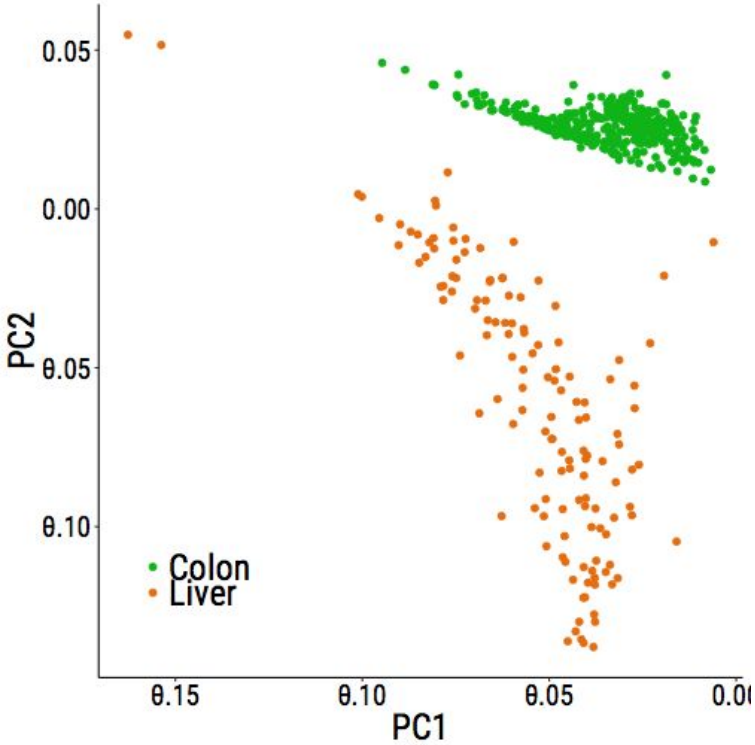


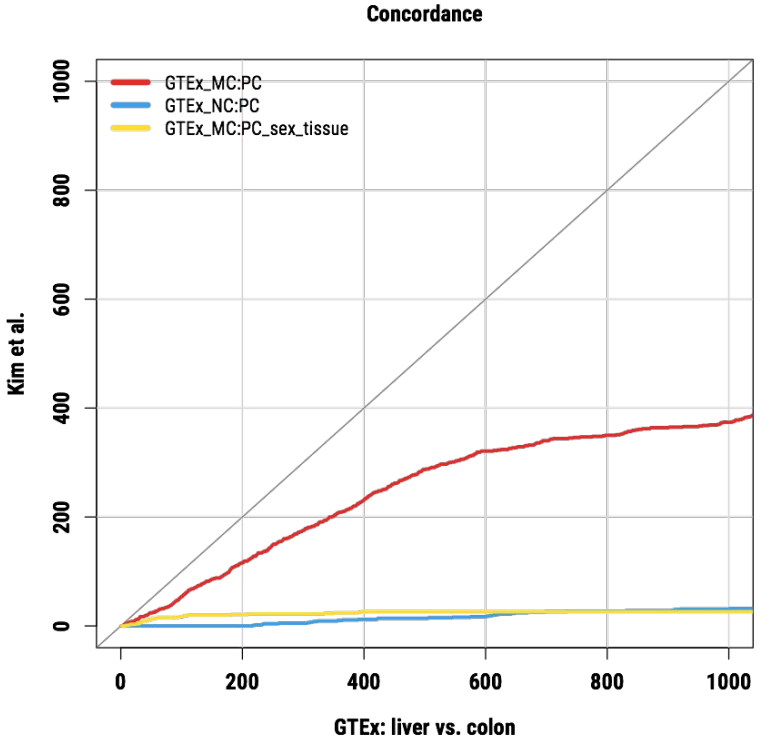
So…what if we compared the MC:PC results with differential expression between colon and liver?

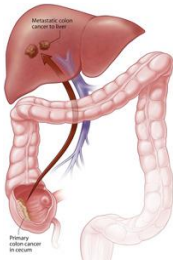


Hypothesis: MC:PC results should be most similar to GTEx colon vs. liver

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	Extract DNA/RNA from samples	Sequence samples	Process sequencing data	Data cleaning	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

Total: days-months

Finally, what if YOU want to use *recount2*...?

predictions (v0.0.06)

sample_id	dataset	reported_sex	predicted_sex	accuracy_sex	…	reported_tissue	predicted_tissue	accuracy_tissue
SRR660824	gtex	male	male	0.999	…	lung	lung	0.977
SRR2166176	gtex	male	male	0.999	…	brain	brain	0.977
SRR606939	gtex	female	female	0.999	…	heart	heart	0.966
SRR2167642	gtex	male	male	0.999	…	brain	brain	0.966
SRR2165473	gtex	male	male	0.999	…	skin	skin	0.966

Expression data and predictions available in `recount` R package

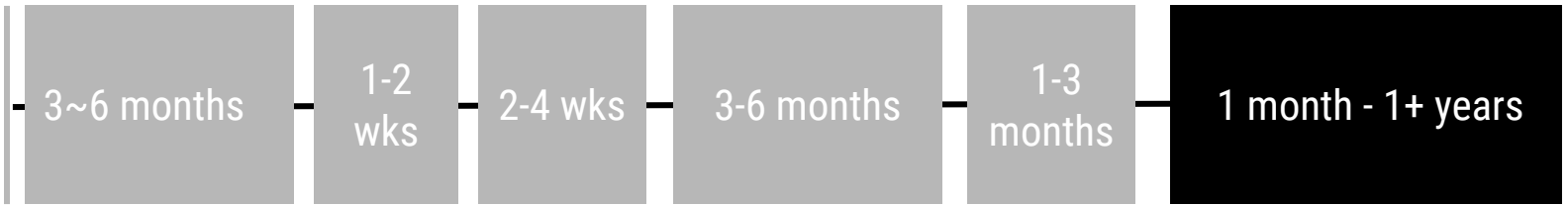
```
>library('recount')

> download_study('ERP001942', type='rse-gene')
> load(file.path('ERP001942', 'rse_gene.Rdata'))
> rse <- scale_counts(rse_gene)

> rse_with_pred <- add_predictions(rse)
```

There are *a lot* of questions an undergraduate could answer with *recount2*...

1. Which genes are expressed in which tissues?
2. Which genes contribute to X-Inactivation?
3. Which genes play a role in cancer? In autism? In Alzheimer's?
4. How different is expression between individuals?



Improving the value of public RNA-seq expression data by phenotype prediction

Shannon E Ellis, Leonardo Collado-Torres, Andrew Jaffe, Jeffrey T Leek

Nucleic Acids Research, Volume 46, Issue 9, 18 May 2018, Page e54,
https://doi.org/10.1093/nar/gky102
Published: 05 March 2018 **Article history** ▼

https://academic.oup.com/nar/article/46/9/e54/4920847

…and all of this is published in Nucleic Acids Research

If you want to…	
Align RNA-Seq data	Rail-RNA Scalable RNA-seq alignment *http://rail.bio*
Learn about human expression	recount2 *https://jhubiostatistics.shinyapps.io/recount/*
Predict phenotype information	*phenopredict* *https://github.com/leekgroup/phenopredict*

The Leek group

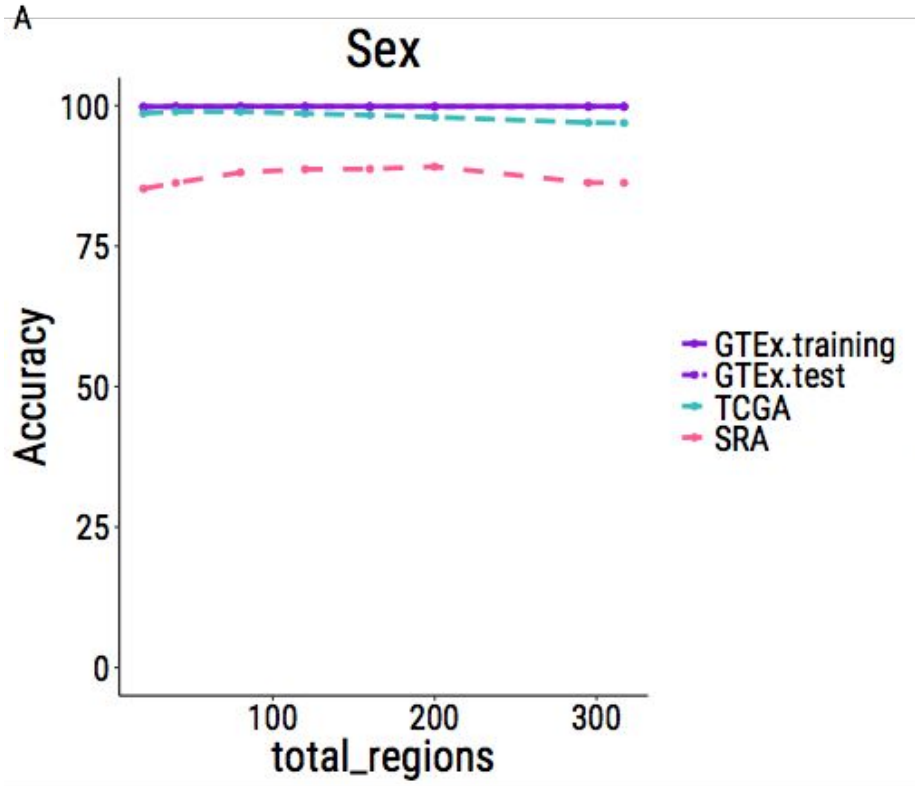
- Huan Chen
- Jack Fu
- Aboozar Hadavand
- Leslie Myint
- Kayode Sosina
- Sara Wang
- **Jeff Leek**

Collaborators

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

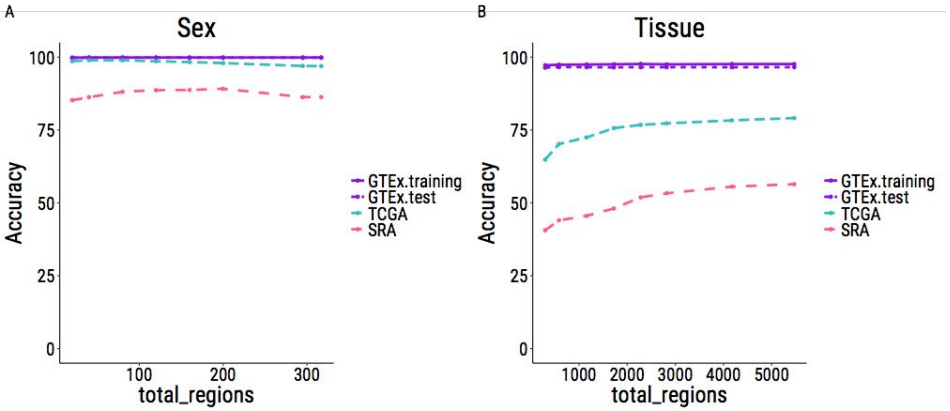
If you want to…	
Align RNA-Seq data	Rail-RNA Scalable RNA-seq alignment *http://rail.bio*
Learn about human expression	recount2 *https://jhubiostatistics.shinyapps.io/recount/*
Predict phenotype information	*phenopredict* *https://github.com/leekgroup/phenopredict*

Number of regions used for prediction was optimized in the training set.



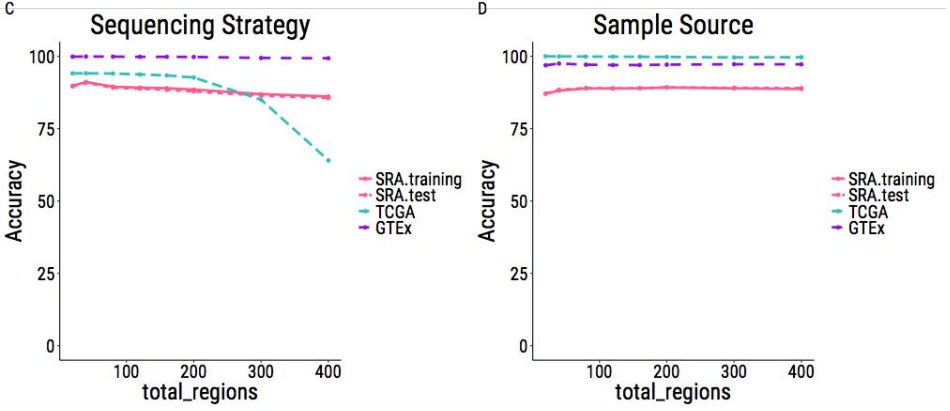
numRegions	10	20	40	60	80	100	150	200
total_regions	20	40	80	120	160	200	295	317
GTEx.training	99.83	99.87	99.87	99.87	99.87	99.87	99.87	99.87
GTEx.test	99.77	99.81	99.83	99.83	99.83	99.81	99.81	99.81
TCGA	98.62	98.95	98.94	98.59	98.32	97.96	96.99	96.93
SRA	85.22	86.26	88.1	88.68	88.74	89.15	86.29	86.26

Phenotypes are largely insensitive to number of regions used to build predictor



numRegions	10	20	40	60	80	100	150	200
total_regions	20	40	80	120	160	200	295	317
GTEx.training	99.83	99.87	99.87	99.87	99.87	99.87	99.87	99.87
GTEx.test	99.77	99.81	99.83	99.83	99.83	99.81	99.81	99.81
TCGA	98.62	98.95	98.94	98.59	98.32	97.96	96.99	96.93
SRA	85.22	86.26	88.1	88.68	88.74	89.15	86.29	86.26

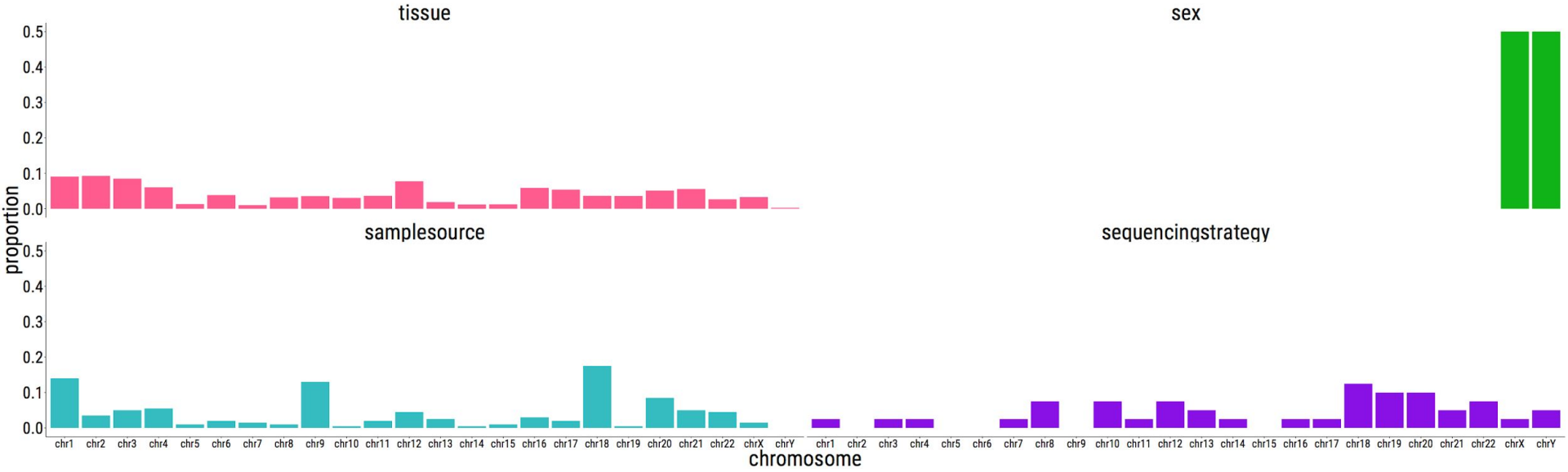
numRegions	10	20	40	60	80	100	150	200
total_regions	297	583	1156	1724	2281	2820	4174	5470
GTEx.training	97.21	97.4	97.46	97.55	97.65	97.53	97.63	97.63
GTEx.test	96.41	96.6	96.56	96.5	96.56	96.56	96.54	96.56
TCGA	64.87	70.2	72.47	75.64	76.77	77.28	78.3	79.11
SRA	40.6	44.02	45.54	48.07	51.88	53.29	55.56	56.41



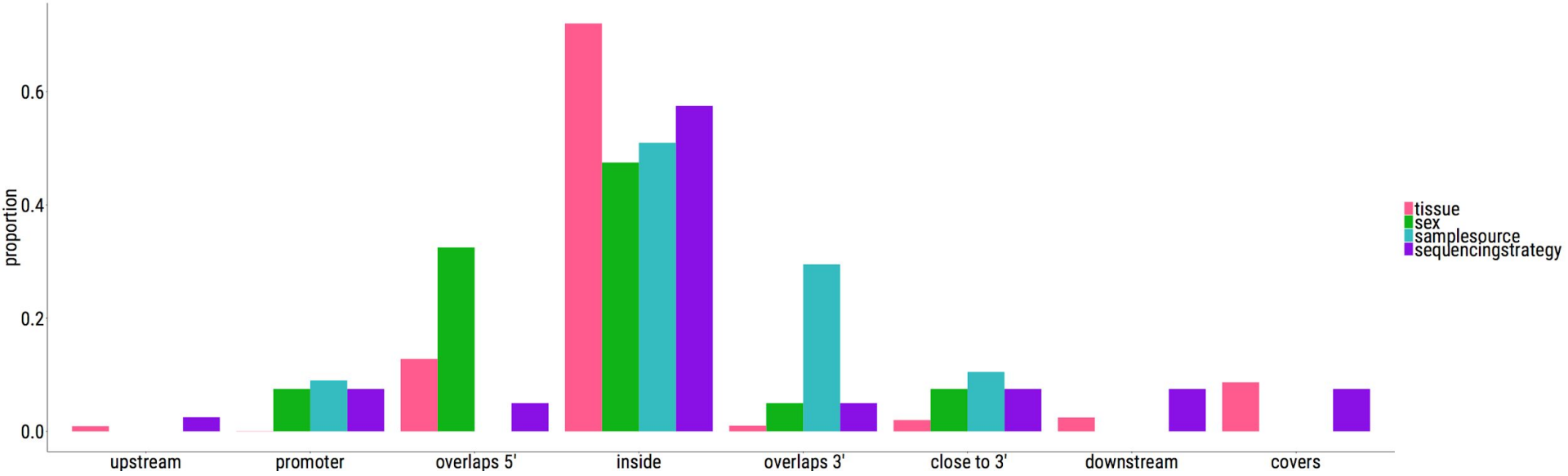
numRegions	10	20	40	60	80	100	150	200
total_regions	20	40	80	120	160	200	300	400
SRA.training	89.83	91.14	89.5	89.17	88.96	88.5	86.94	86.16
SRA.test	89.63	90.86	89.15	88.82	88.43	87.86	86.41	85.57
TCGA	94.11	94.13	94.05	93.76	93.41	92.68	85.04	64.05
GTEx	99.88	99.94	99.9	99.83	99.83	99.79	99.44	99.34

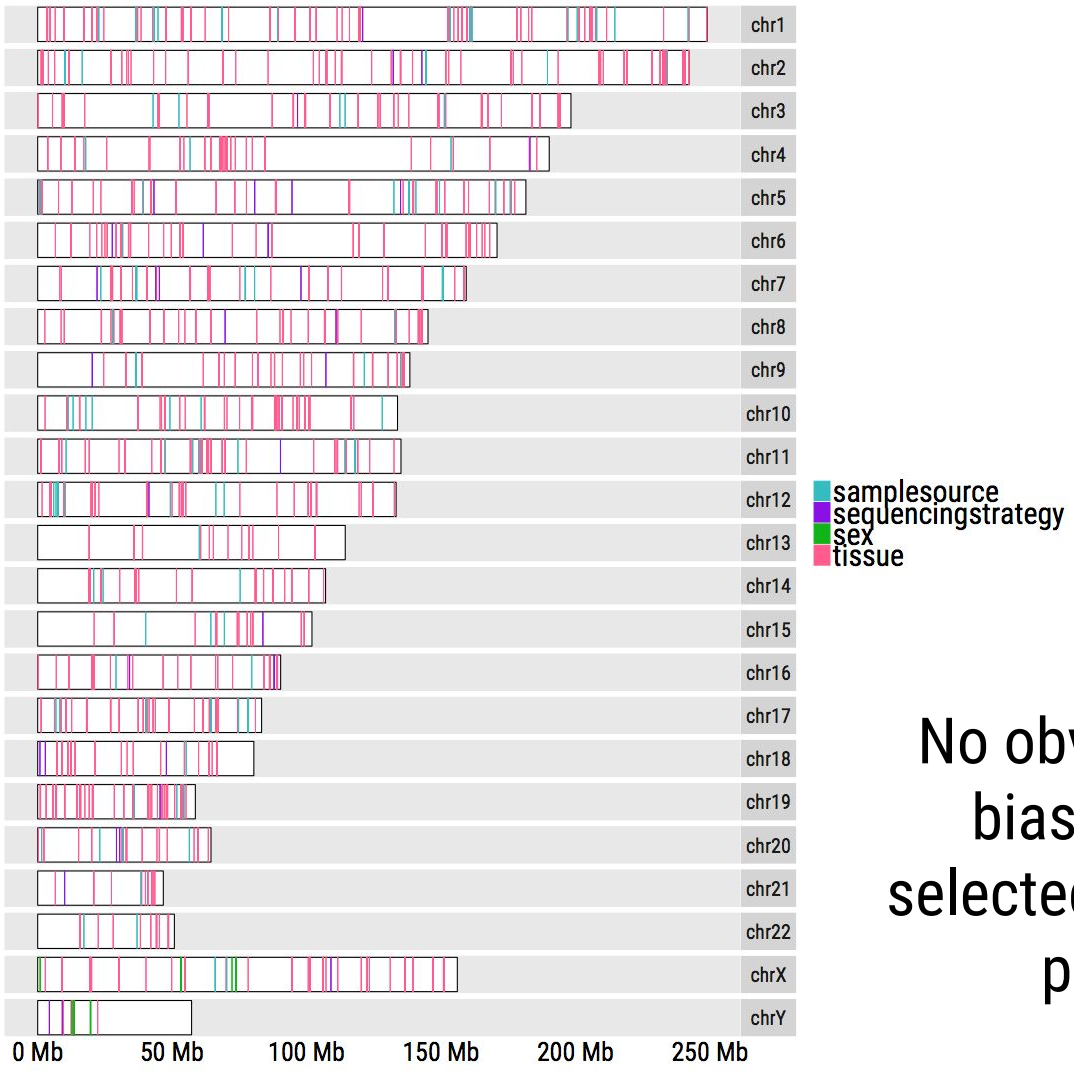
numRegions	10	20	40	60	80	100	150	200
total_regions	20	40	80	120	160	200	300	400
SRA.training	87.05	88.12	88.87	88.89	88.92	89.2	88.9	88.71
SRA.test	87.11	88.45	89.01	88.88	88.93	89.23	89.08	88.98
TCGA	99.97	99.95	99.88	99.84	99.81	99.77	99.59	99.67
GTEx	96.87	97.47	97.09	96.95	96.96	97.1	97.23	97.2

Aside from sex prediction, regions were selected across all chromosomes

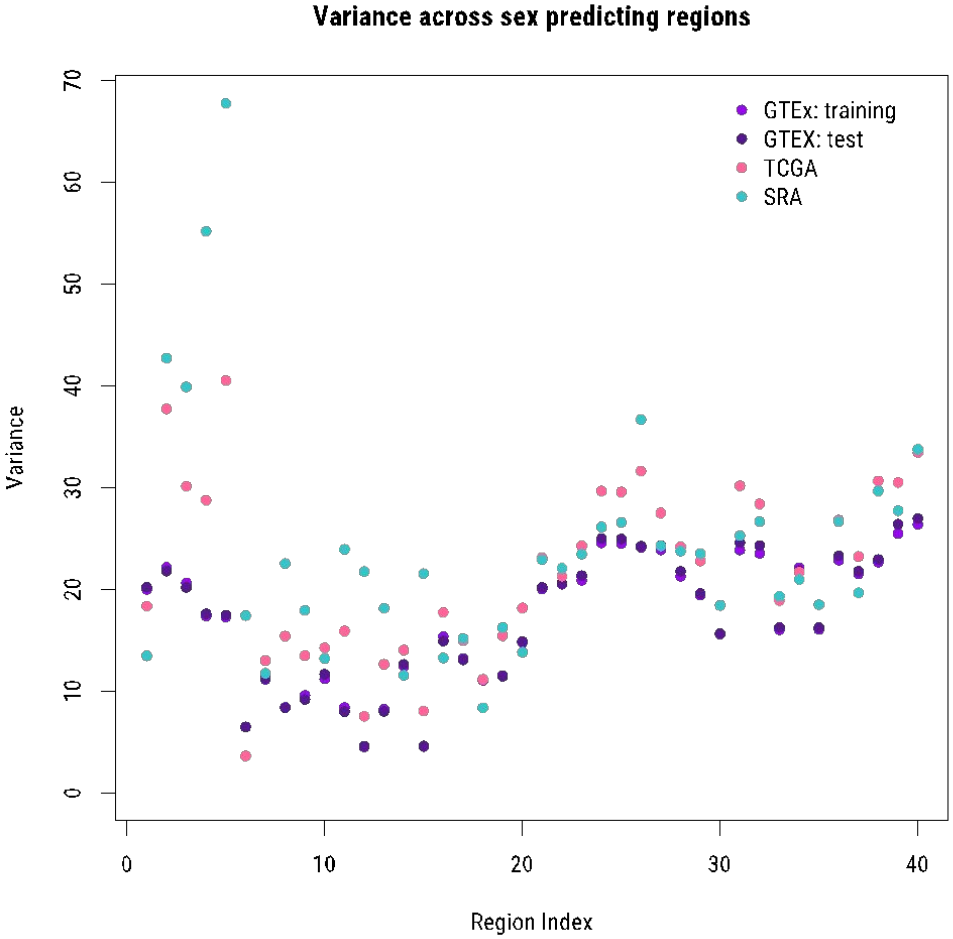


Most regions used for prediction fell within the annotated gene; however sequencing strategy shows some bias for regions outside the annotated transcriptome.

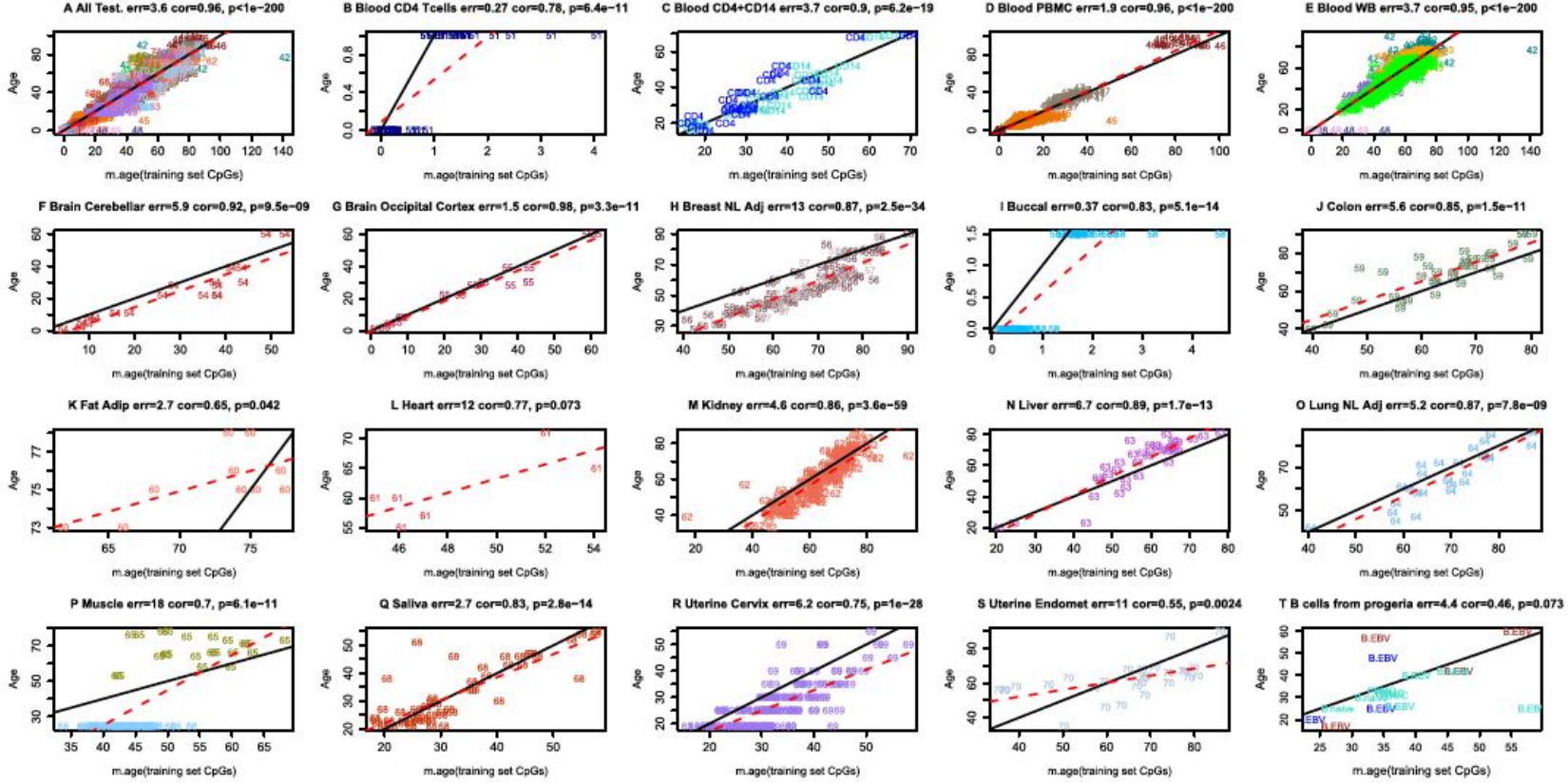




No obvious location bias for regions selected for any of the predictors

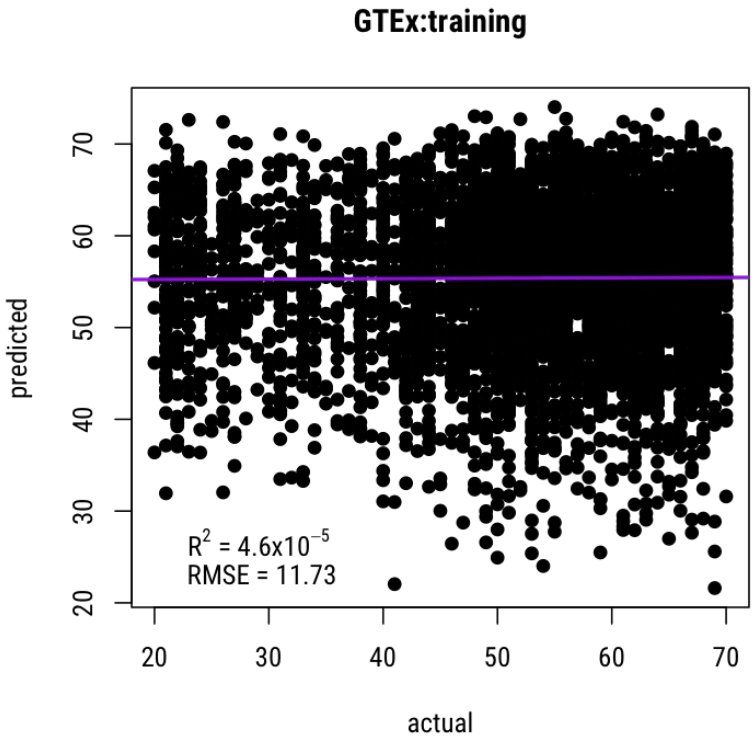

Variance across sex predicting regions
GTEx: training
GTEX: test
TCGA
SRA
Variance
Region Index
0
10
20
30
40
50
60
70
0
10
20
30
40

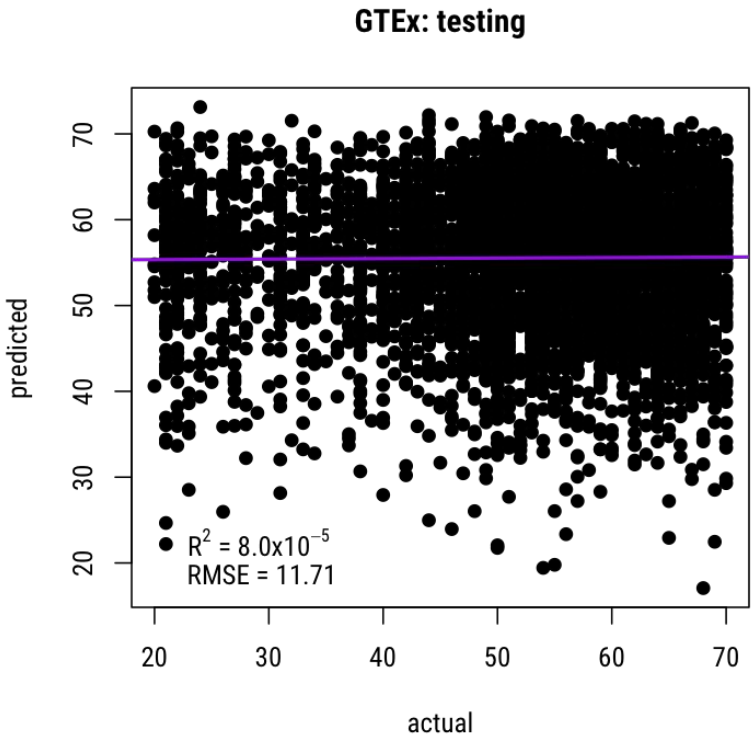
Horvath demonstrates that 353 CpGs can accurately predict age



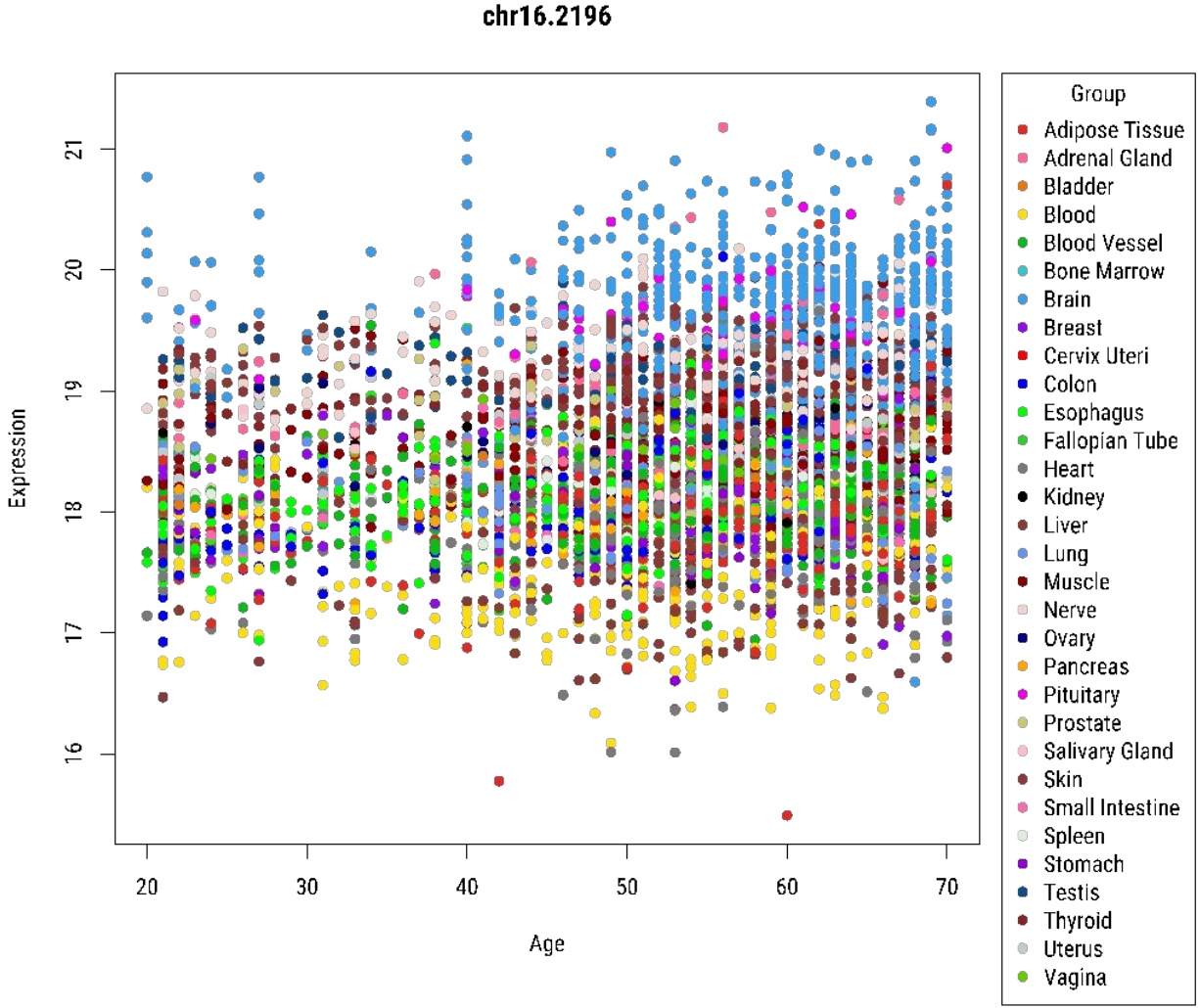
Horvath. *Genome Biology*

How well can we predict age in GTEx?

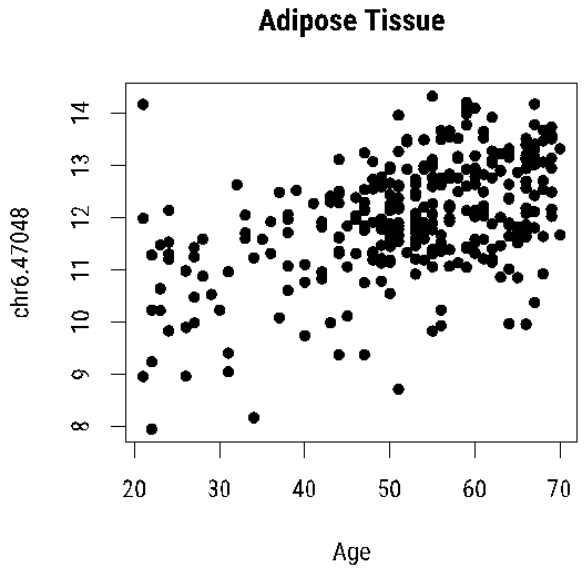


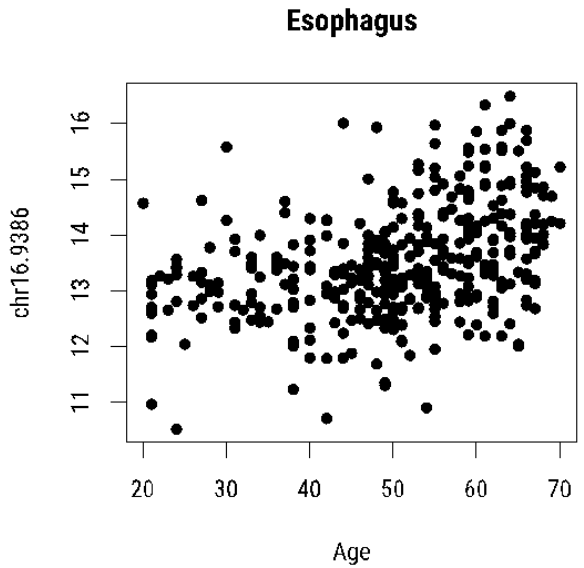


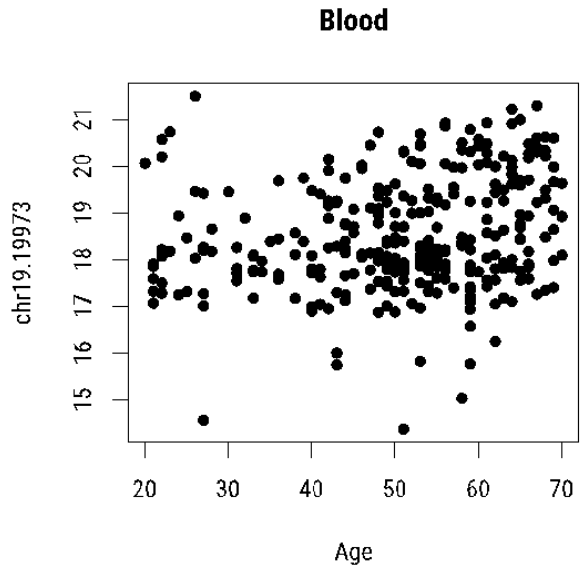
Tissue poses a problem for prediction…

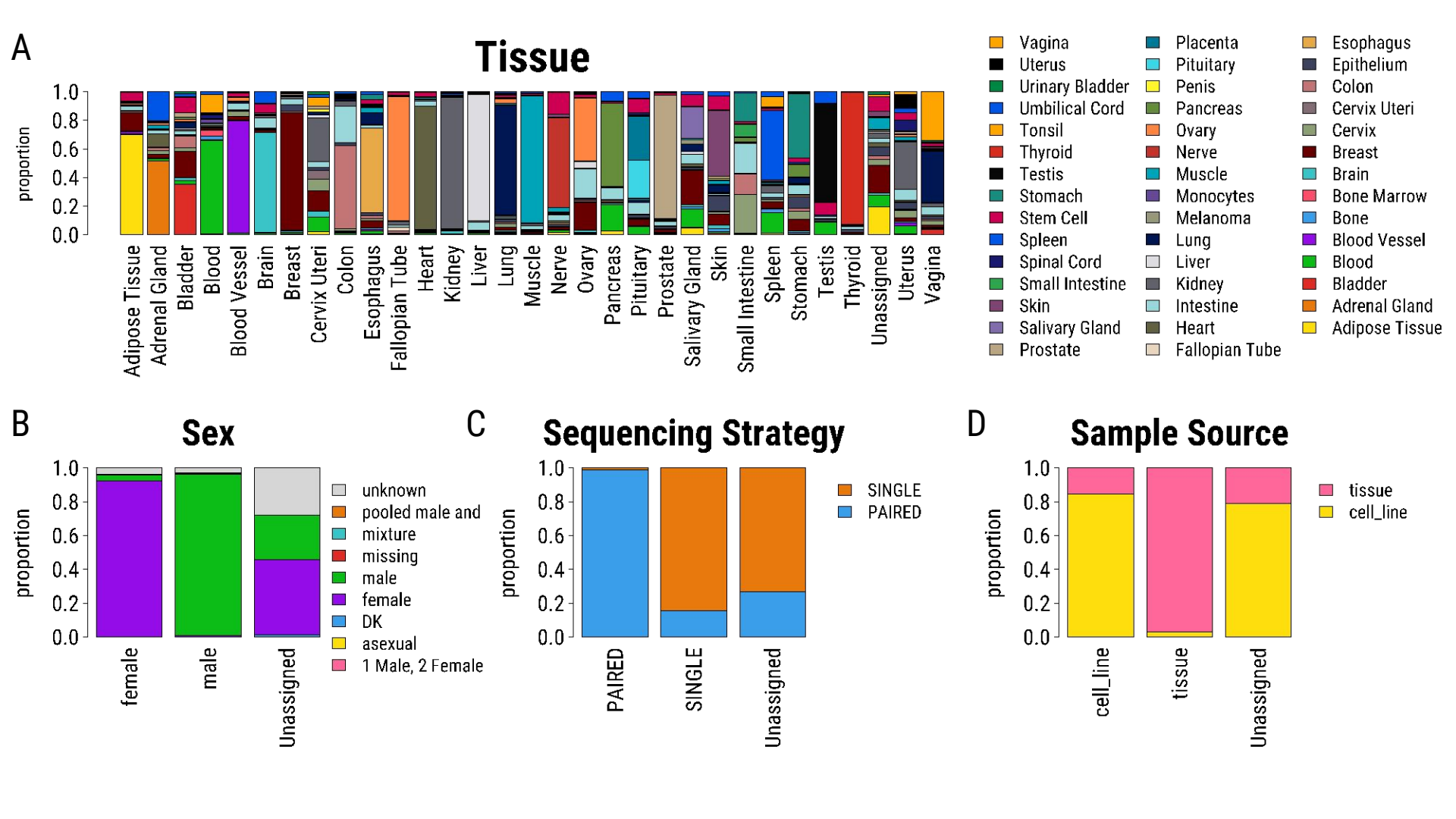


Even within tissue, signal is pretty weak…

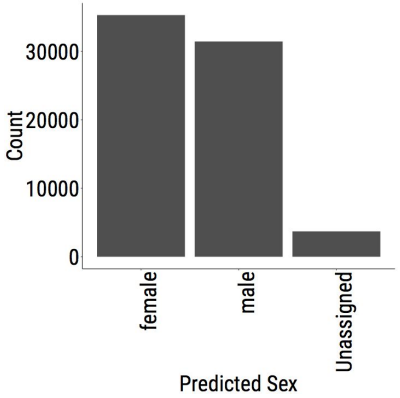


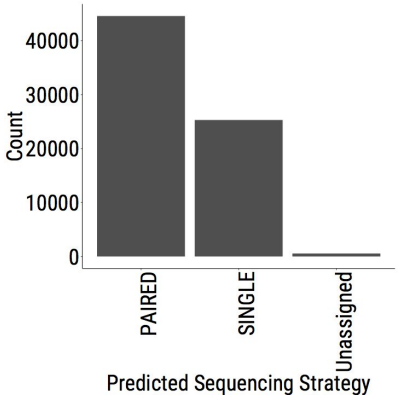


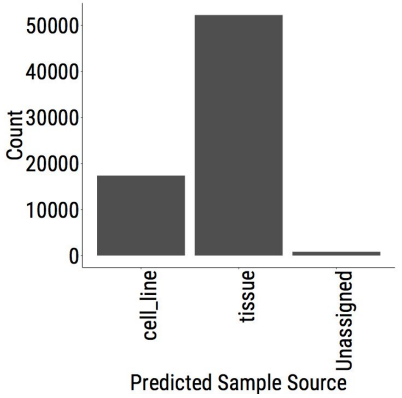


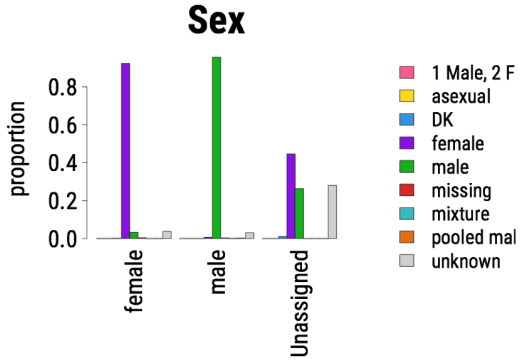

A
Tissue
proportion
1.0
0.8
0.6
0.4
0.2
0.0
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Brain
Breast
Cervix Uteri
Colon
Esophagus
Fallopian Tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid
Unassigned
Uterus
Vagina
Vagina
Uterus
Urinary Bladder
Umbilical Cord
Tonsil
Thyroid
Testis
Stomach
Stem Cell
Spleen
Spinal Cord
Small Intestine
Skin
Salivary Gland
Prostate
Placenta
Pituitary
Penis
Pancreas
Ovary
Nerve
Muscle
Monocytes
Melanoma
Lung
Liver
Kidney
Intestine
Heart
Fallopian Tube
Esophagus
Epithelium
Colon
Cervix Uteri
Cervix
Breast
Brain
Bone Marrow
Bone
Blood Vessel
Blood
Bladder
Adrenal Gland
Adipose Tissue
B
Sex
proportion
1.0
0.8
0.6
0.4
0.2
0.0
female
male
Unassigned
unknown
pooled male and
mixture
missing
male
female
DK
asexual
1 Male, 2 Female
C
Sequencing Strategy
proportion
1.0
0.8
0.6
0.4
0.2
0.0
PAIRED
SINGLE
Unassigned
SINGLE
PAIRED
D
Sample Source
proportion
1.0
0.8
0.6
0.4
0.2
0.0
cell_line
tissue
Unassigned
tissue
cell_line

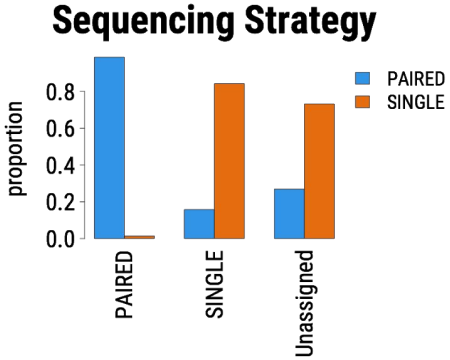
When we look at the other three predictions…

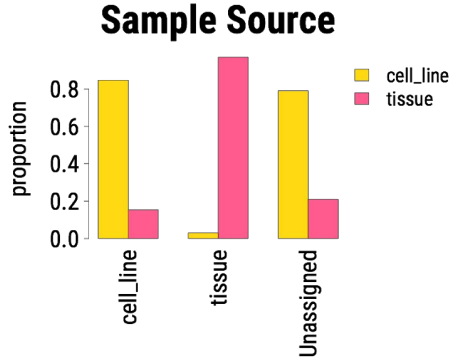












When we look at the other three predictions…