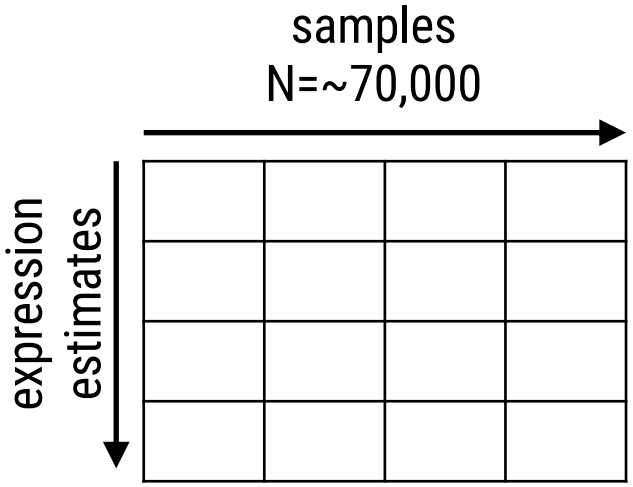
In silico phenotyping to improve the usefulness of public data

Shannon E. Ellis

Leek Group

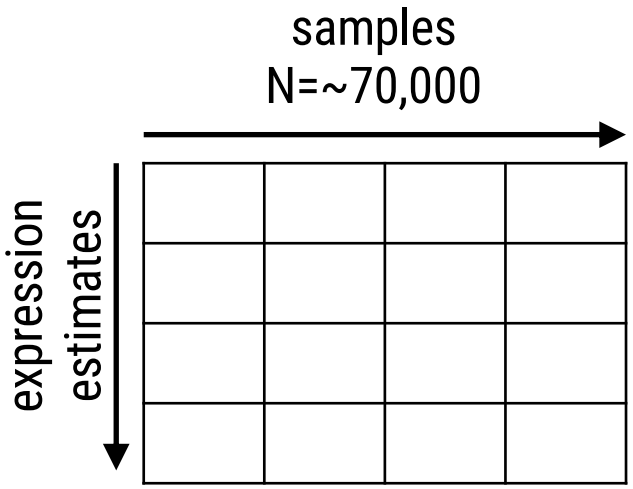
@Shannon_E_Ellis

recount2



1. Have ‘novel’ isoforms ever been seen previously?
2. What regions of the human genome are transcribed in humans?

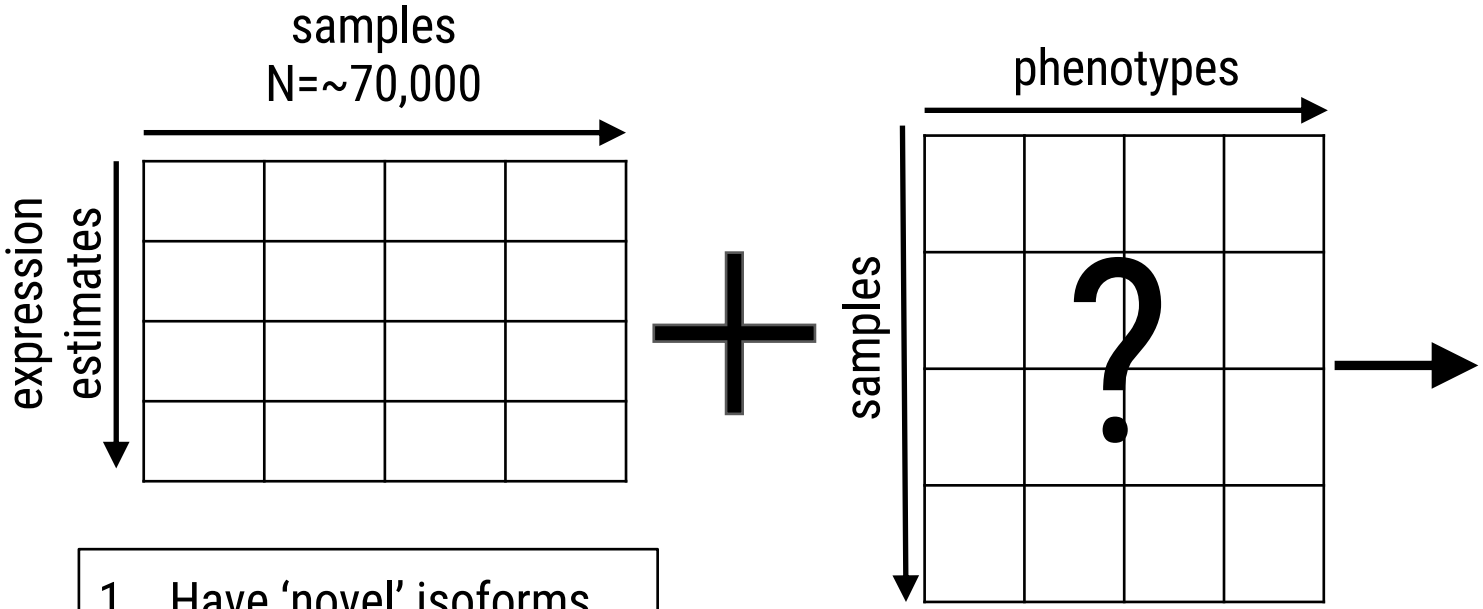
recount2



1. Have ‘novel’ isoforms ever been seen previously?
2. What regions of the human genome are transcribed in humans?

1. Have ‘novel’ isoforms ever been seen previously? **In what tissue? At what levels?**
2. What regions of the human genome are transcribed in humans **and in what tissues**?
3. Do the same genes escape X Inactivation across all tissues?
4. What expression changes occur as we age?
5. …….

recount2



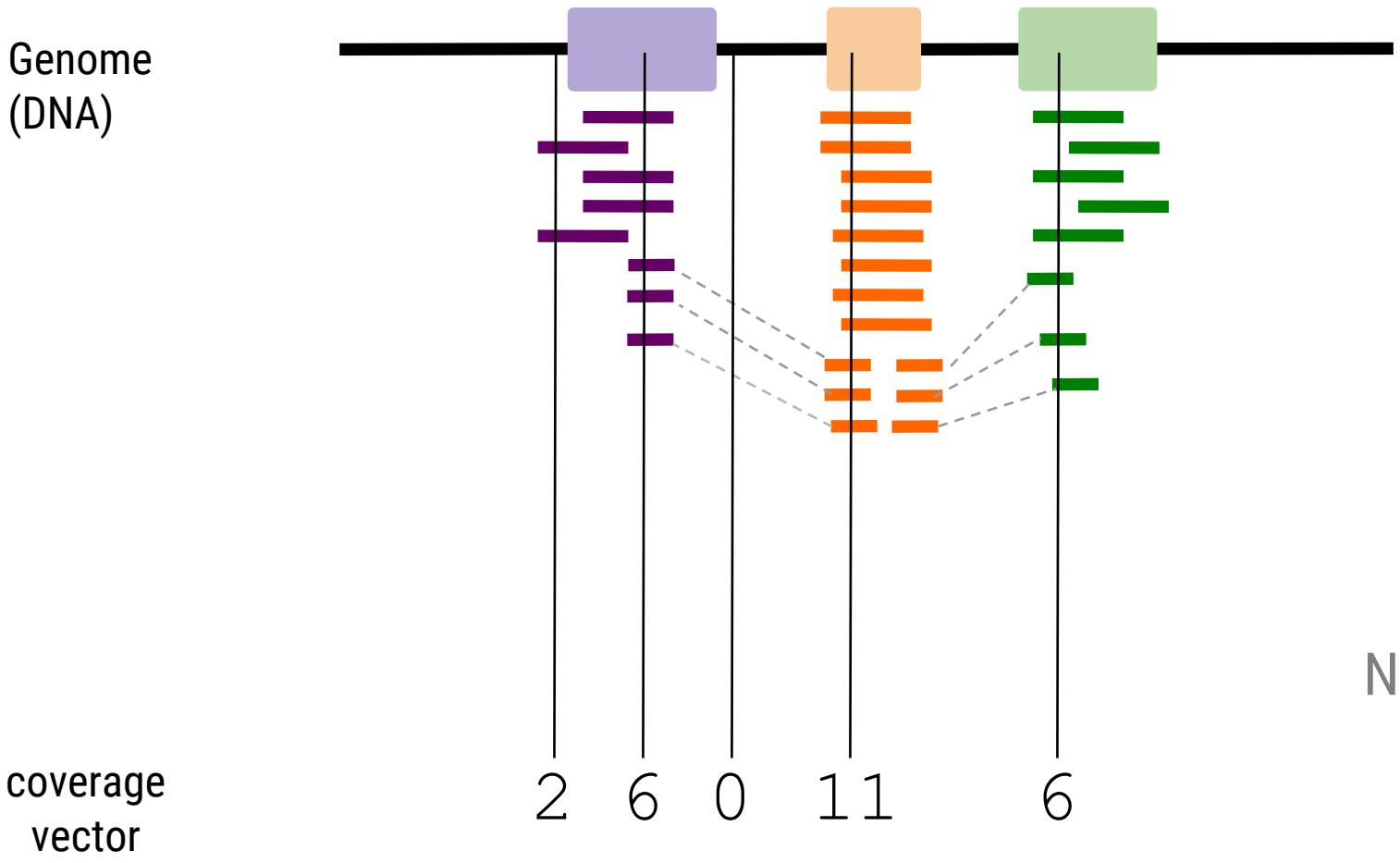
1. Have ‘novel’ isoforms ever been seen previously?
2. What regions of the human genome are transcribed in humans?

1. Have ‘novel’ isoforms ever been seen previously? **In what tissue? At what levels?**
2. What regions of the human genome are transcribed in humans **and in what tissues**?
3. Do the same genes escape X Inactivation across all tissues?
4. What expression changes occur as we age?
5. …….

Measuring Transcription

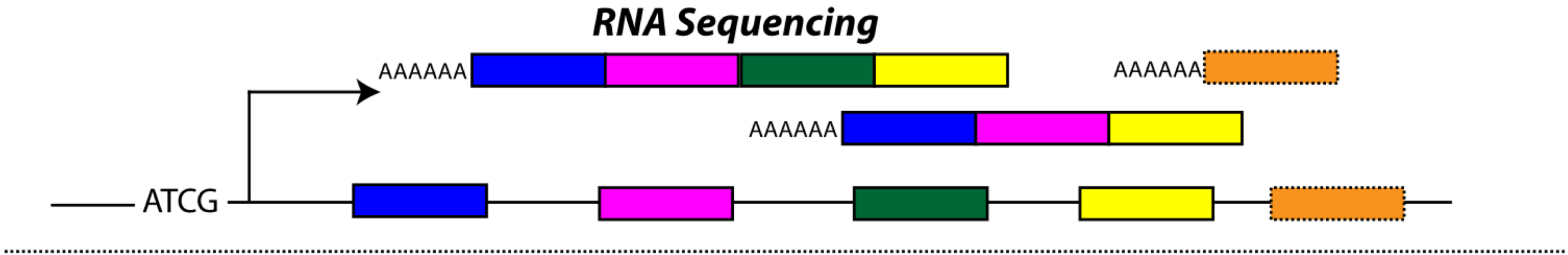
slide adapted from jeff leek

RNA-Sequencing: Alignment using Rail-RNA

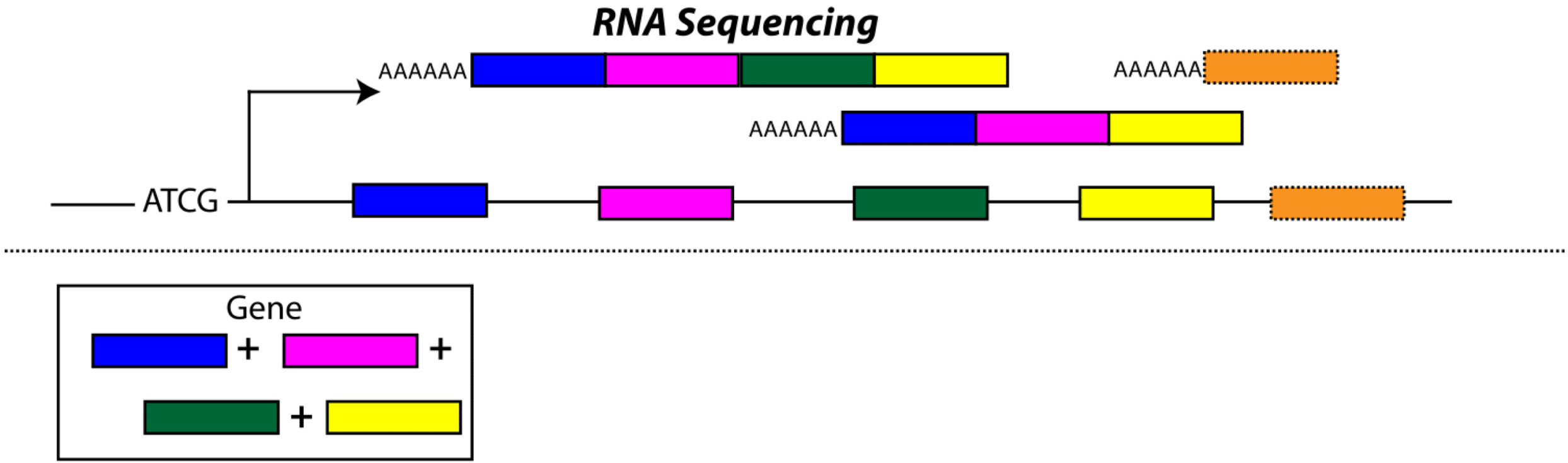


Nellore *et al.* (2016)
Bioinformatics
http://rail.bio/

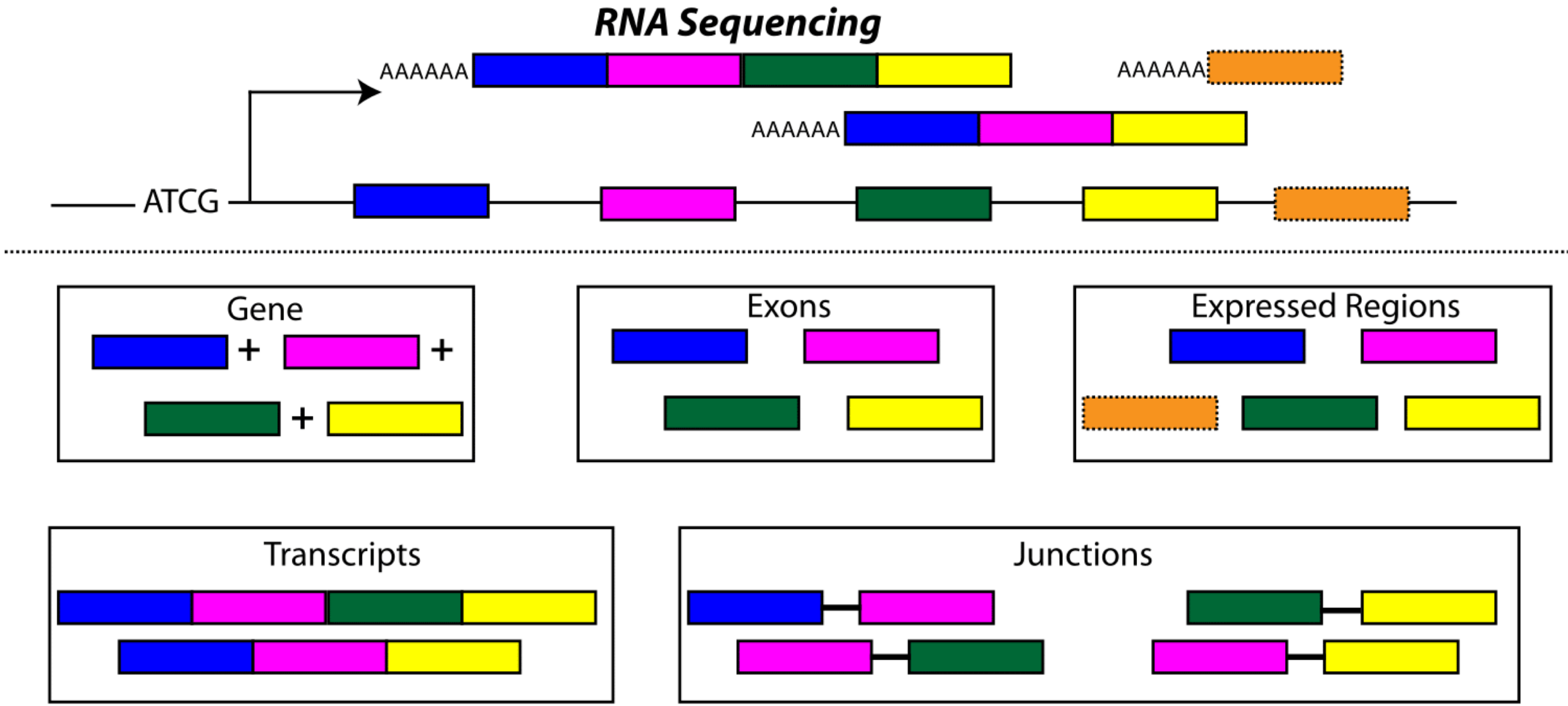
slide adapted from jeff leek



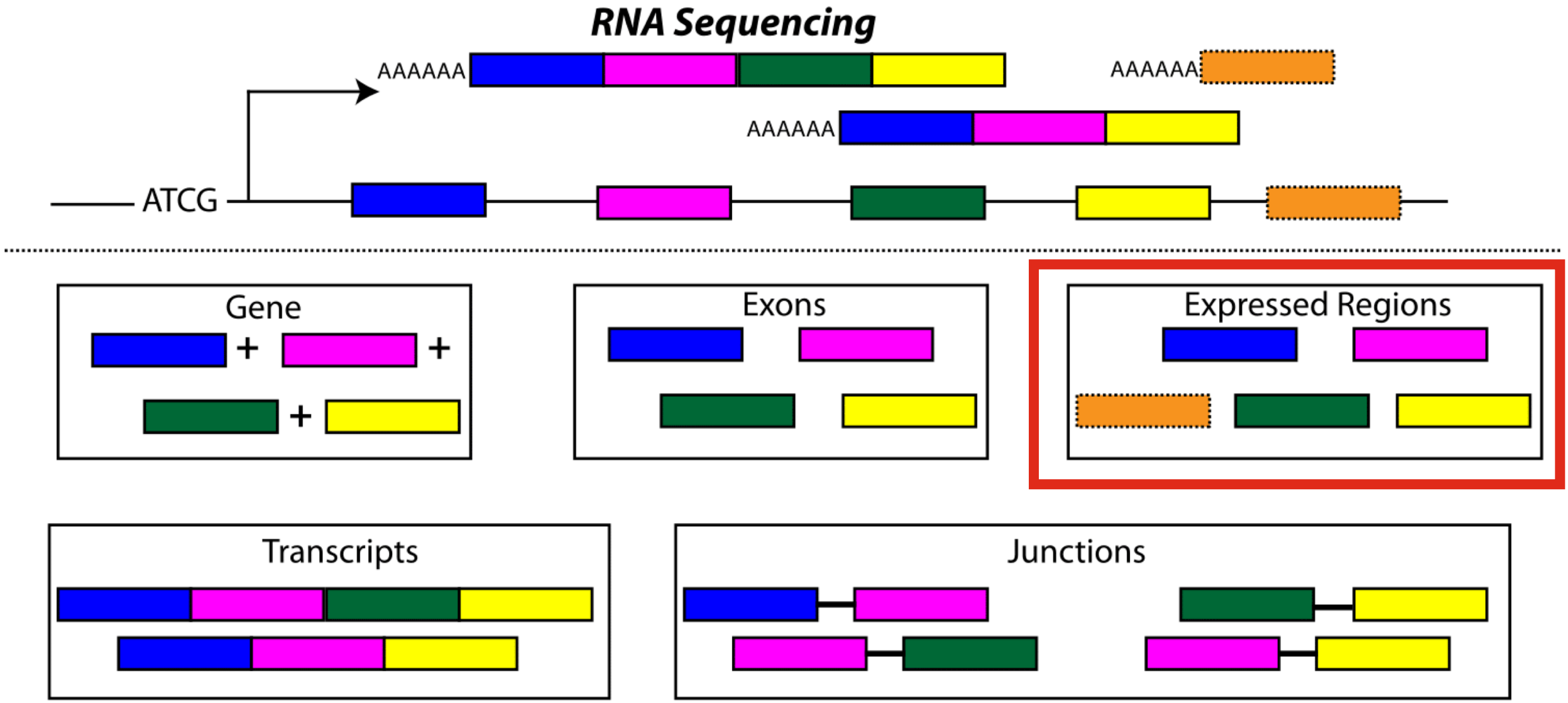
slide adapted from andrew jaffe



slide adapted from andrew jaffe



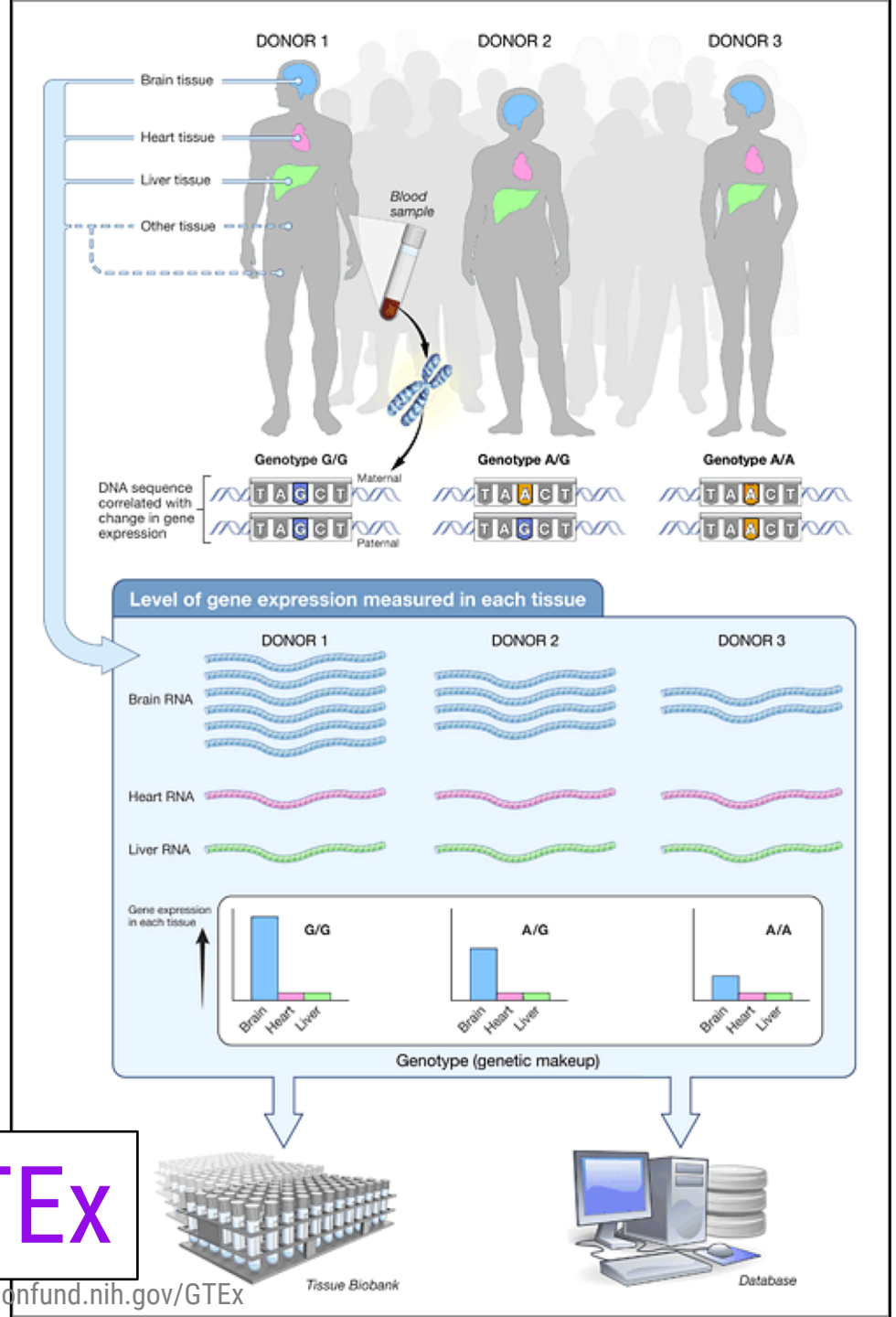
slide adapted from andrew jaffe



slide adapted from andrew jaffe

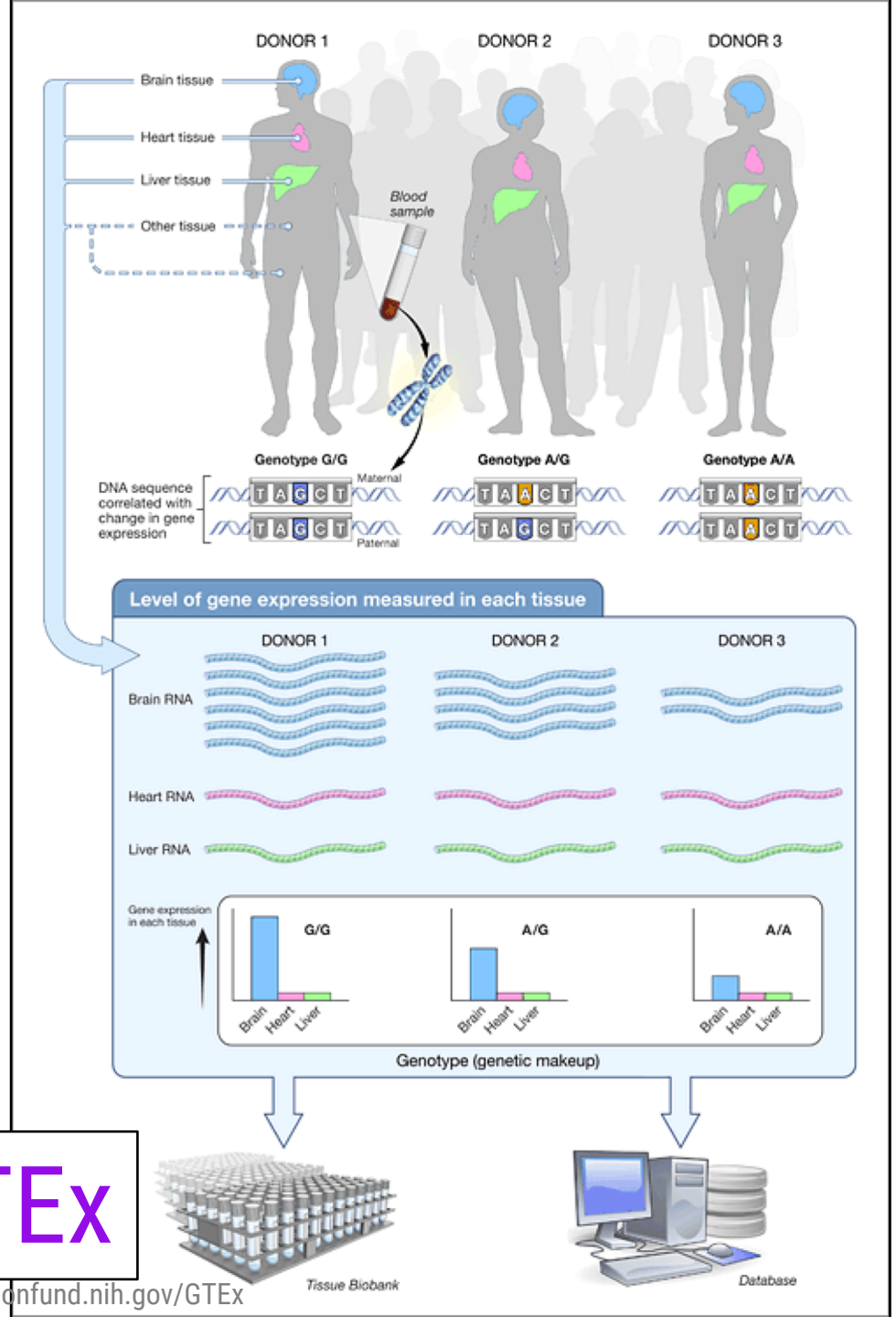
Scaling Up

slide adapted from jeff leek



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

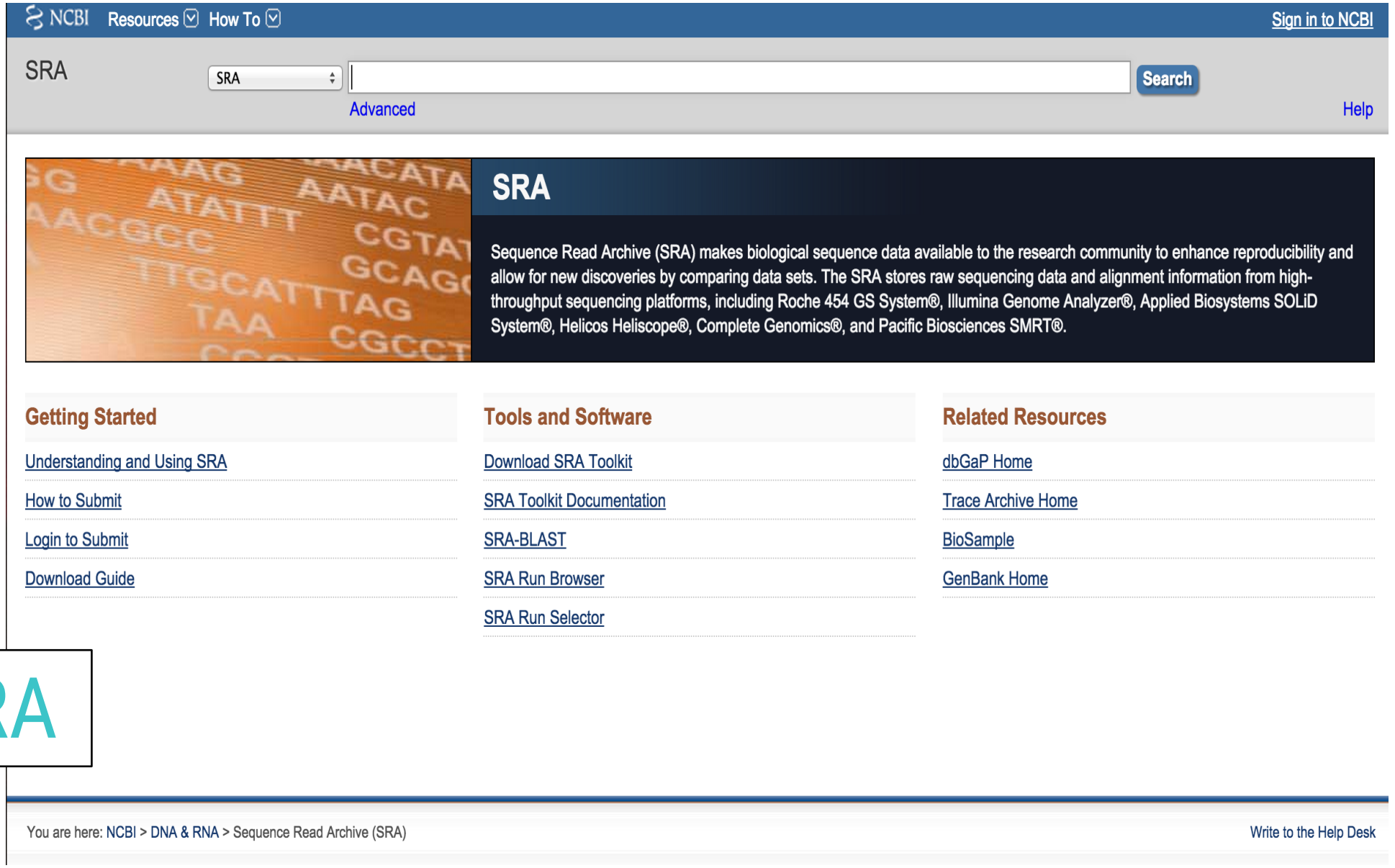
WHAT'S NEXT?

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.

*...alysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, 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

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

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

Related publications

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

The Datasets

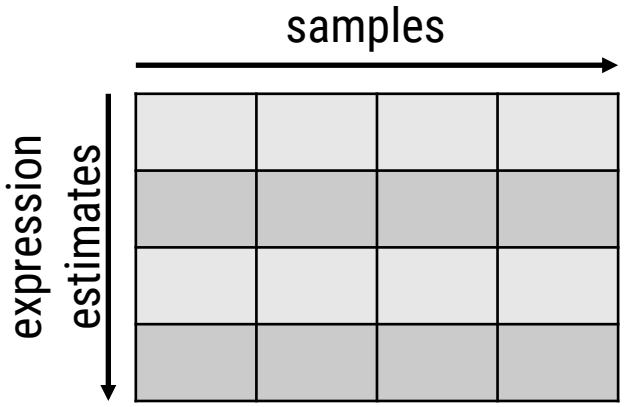
Show 10 entries

Search:

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

recount2

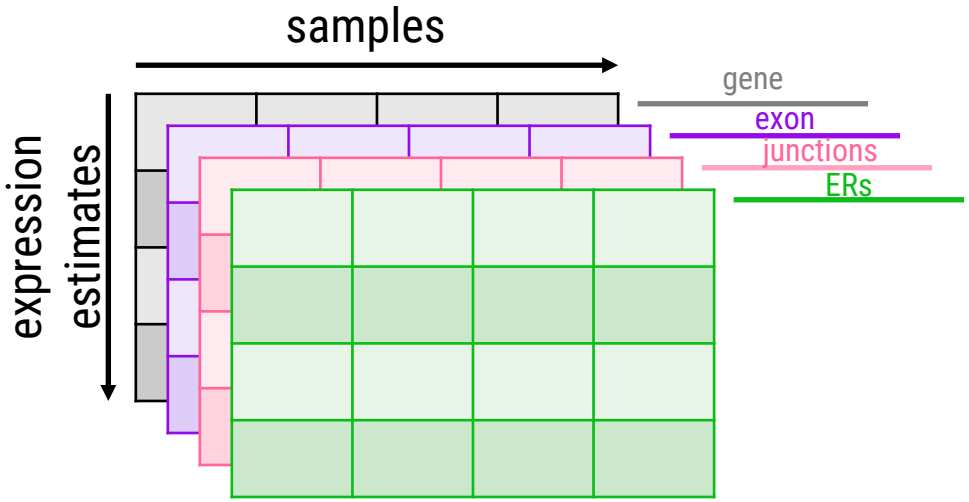
expression data for ~70,000 human samples



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

recount2

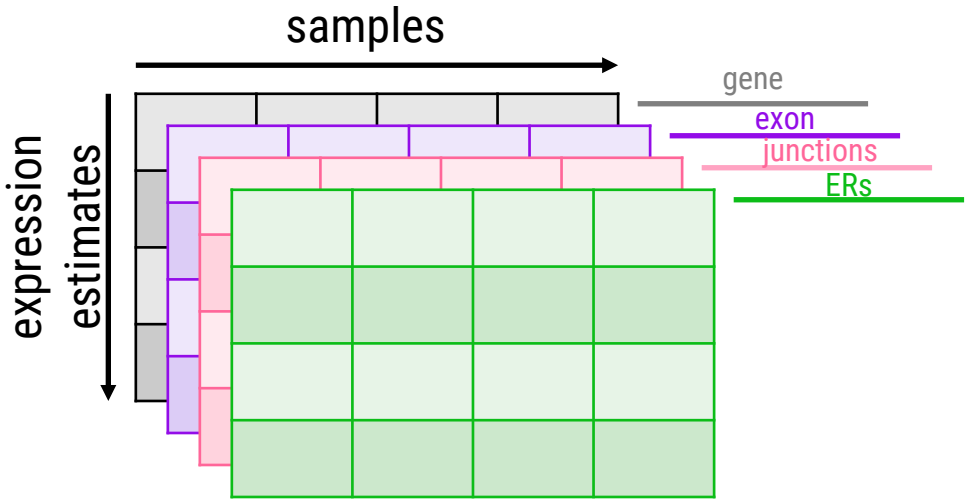
expression data for ~70,000 human samples



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

recount2

expression data for ~70,000 human samples



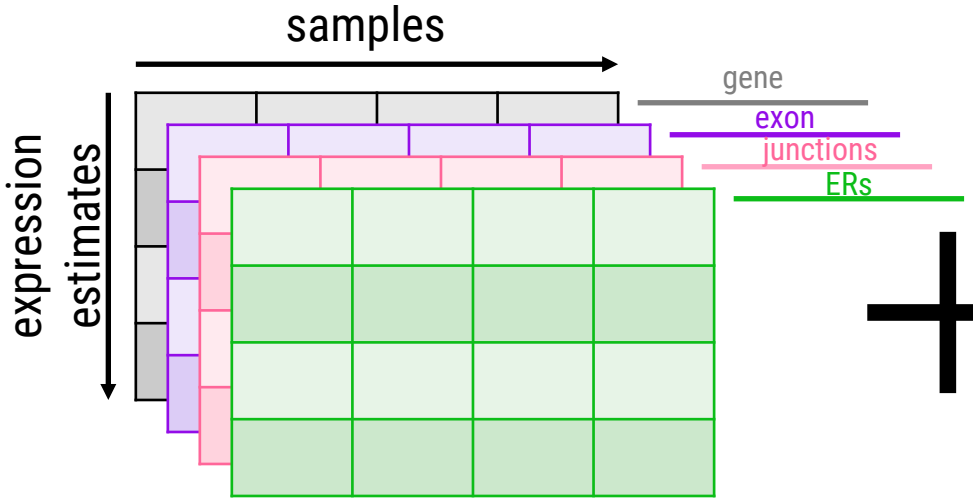
GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284

→

Answer meaningful questions about human biology and expression

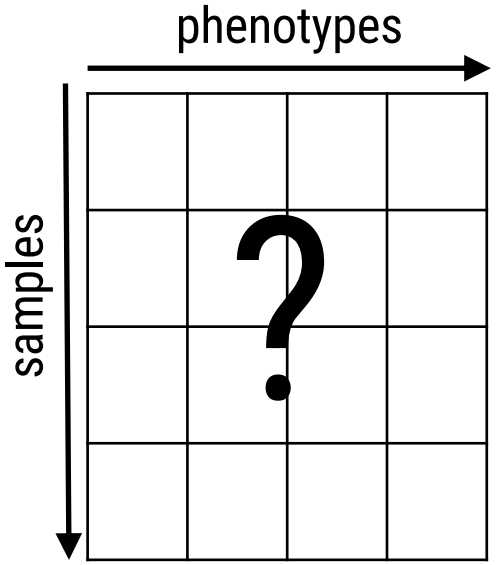
recount2

expression data for ~70,000 human samples



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

+

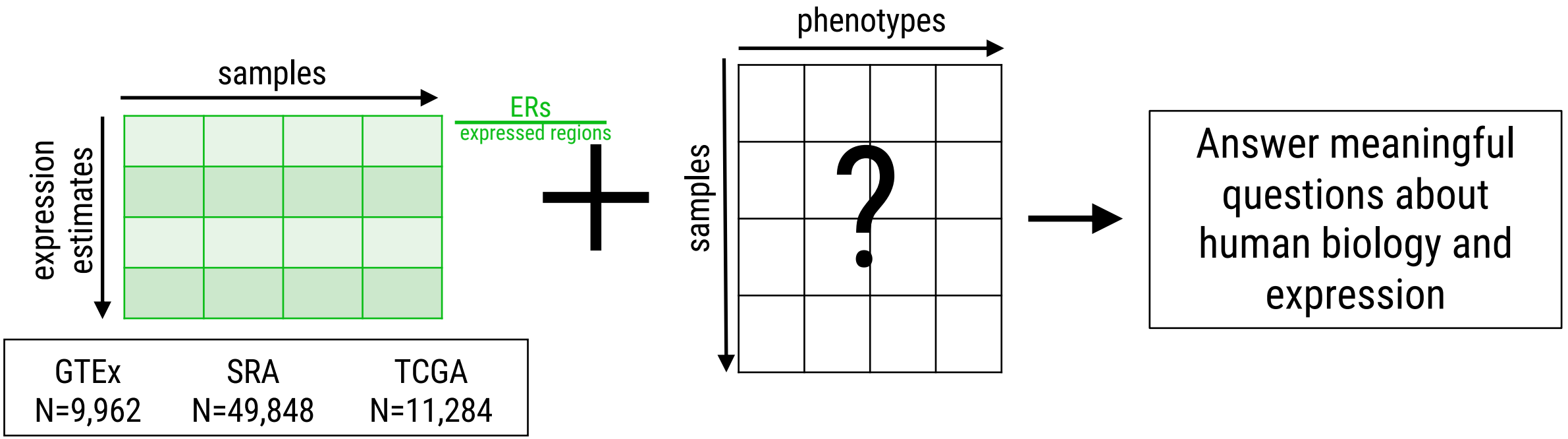


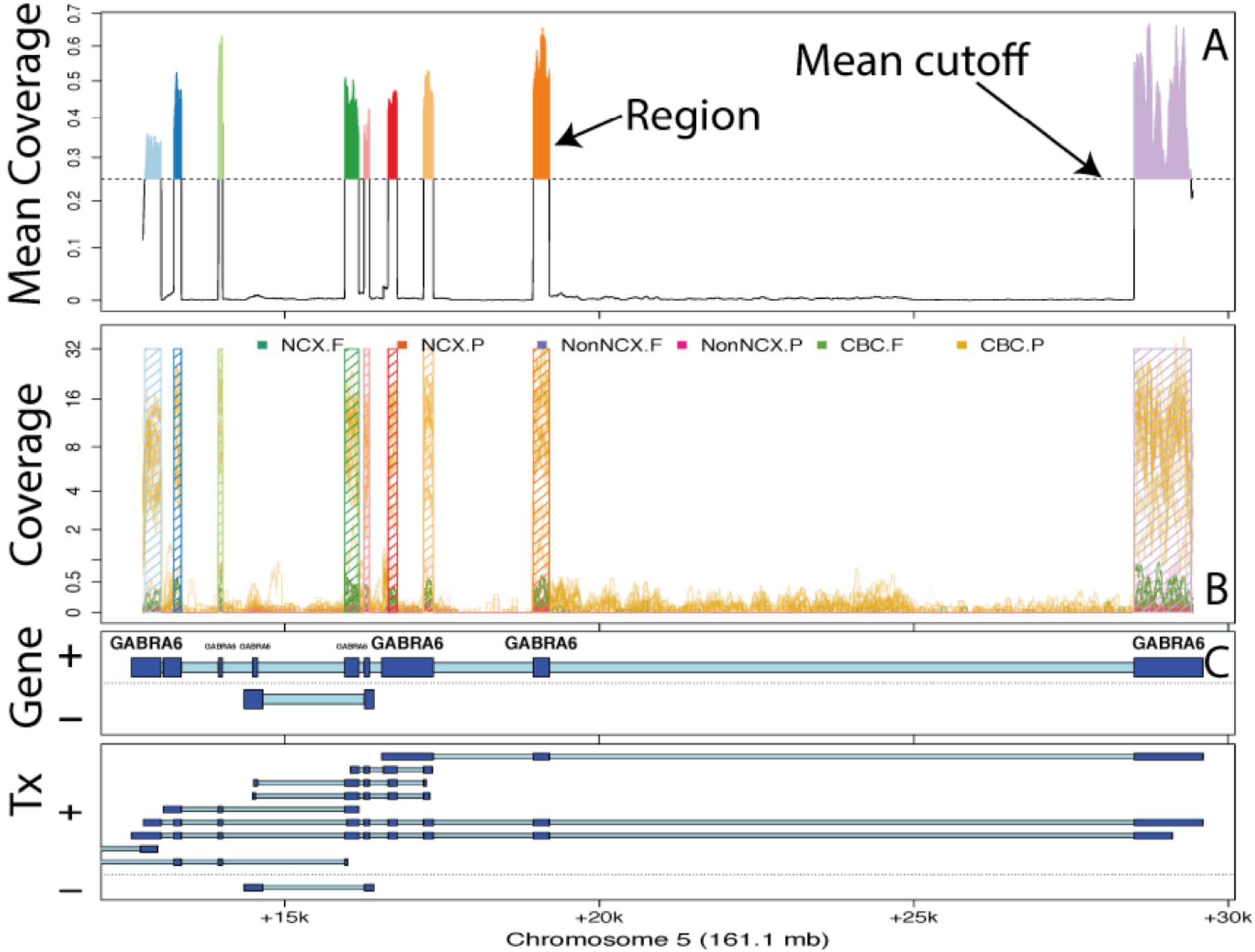
→

Answer meaningful questions about human biology and expression

recount2

expression data for ~70,000 human samples





slide from jeff leek

in-silico Phenotyping

slide adapted from jeff leek

SRA phenotype information is far from complete

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

SRA phenotype information is far from complete

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

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

sra_meta$Sex

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_meta$Sex

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_meta$Sex

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

Missingness limited in **GTEx data**

	SubjectID	Sex	Tissue	Race	Age
1	678145	male	Lung	White	59
2	706551	male	Brain	White	27
3	590954	female	Heart	Black or African American	23
4	706553	male	Brain	White	51
5	706551	male	Skin	White	27
6	590947	male	Lung	White	68
7	590933	female	Brain	White	61
8	706560	female	Adipose Tissue	White	42
9	678142	male	Brain	White	40
10	590945	female	Uterus	White	33
11	706562	female	Nerve	White	60
12	721000	male	Muscle	White	54
13	984968	female	Ovary	White	31
14	720990	male	Blood	White	53
15	985156	female	Brain	White	56
16	985125	male	Muscle	White	44

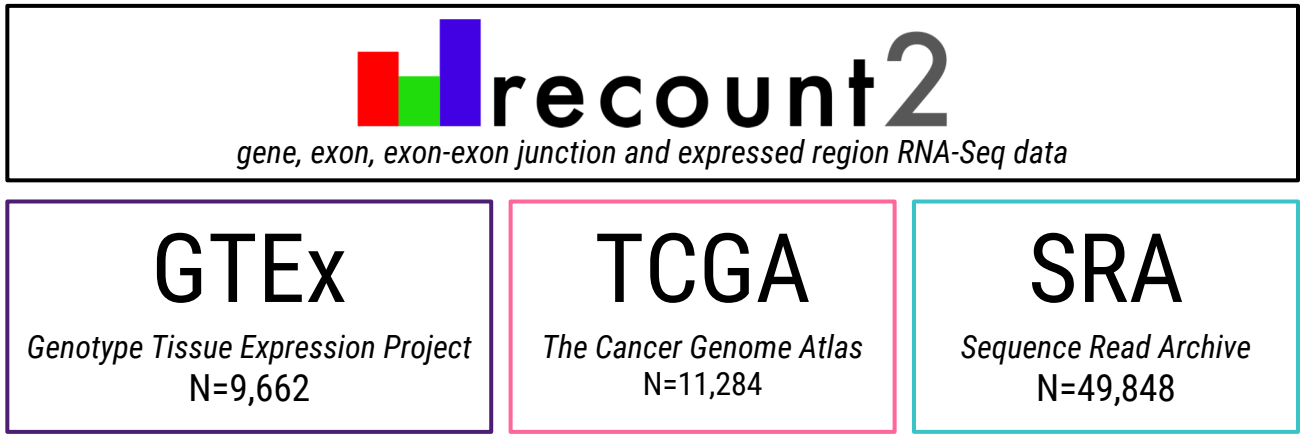
Missingness limited in **GTEx data**

	SubjectID	Sex	Tissue	Race	Age
1	678145	male	Lung	White	59
2	706551	male	Brain	White	27
3	590954	female	Heart	Black or African American	23
4	706553	male	Brain	White	51
5	706551	male	Skin	White	27
6	590947	male	Lung	White	68
7	590933	female	Brain	White	61
8	706560	female	Adipose Tissue	White	42
9	678142	male	Brain	White	40
10	590945	female	Uterus	White	33
11	706562	female	Nerve	White	60
12	721000	male	Muscle	White	54
13	984968	female	Ovary	White	31
14	720990	male	Blood	White	53
15	985156	female	Brain	White	56
16	985125	male	Muscle	White	44

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

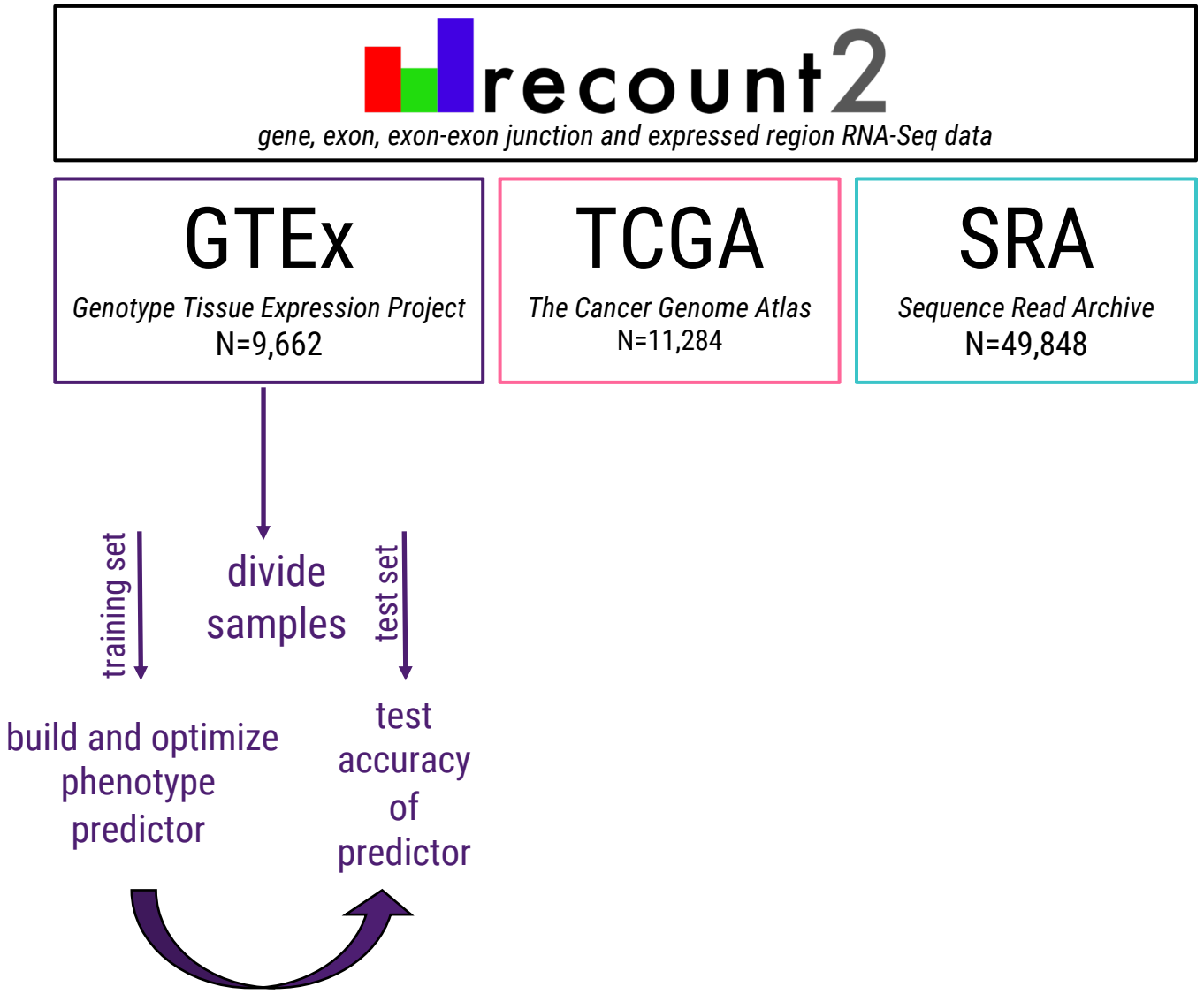
Goal :

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



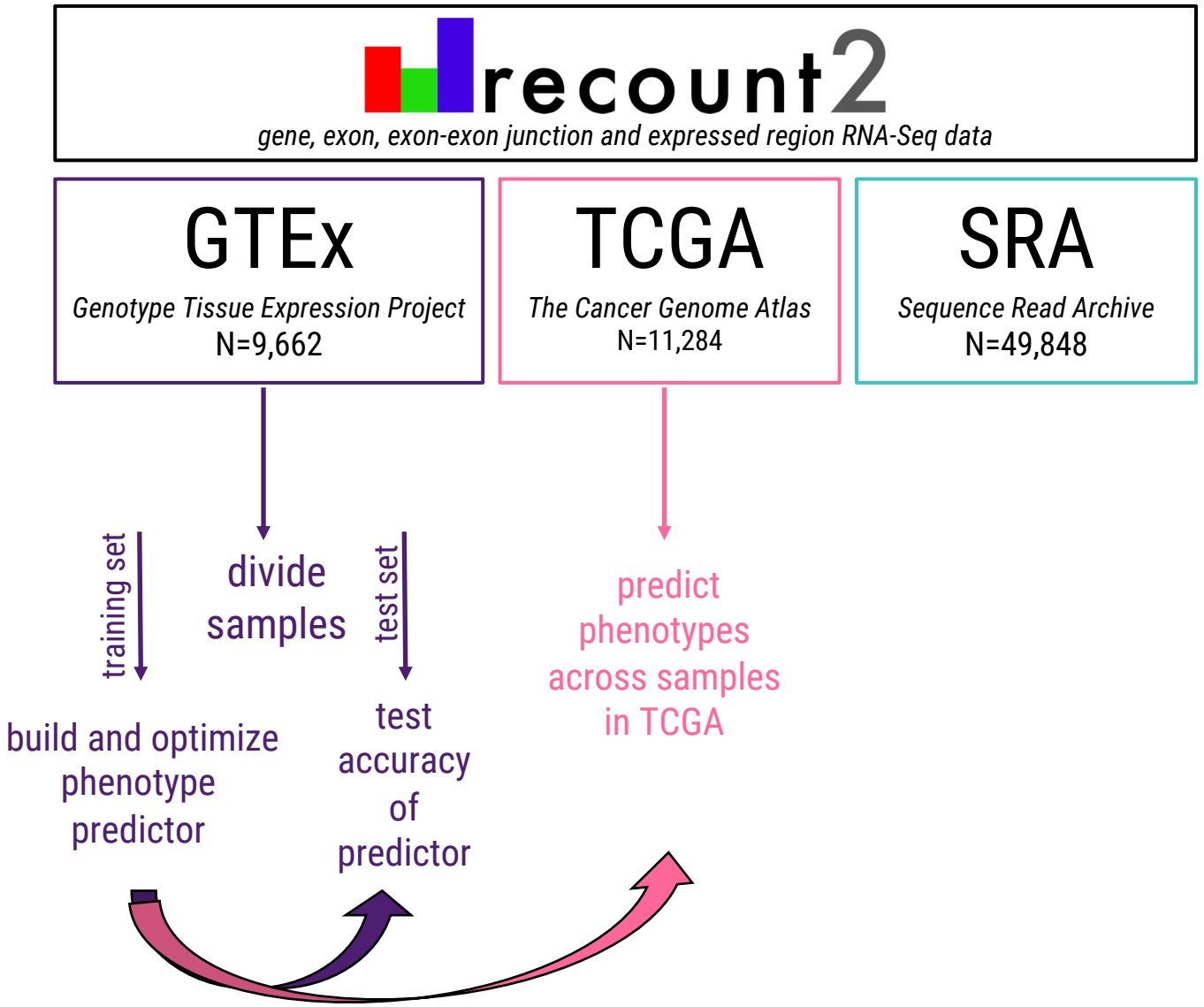
Goal :

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



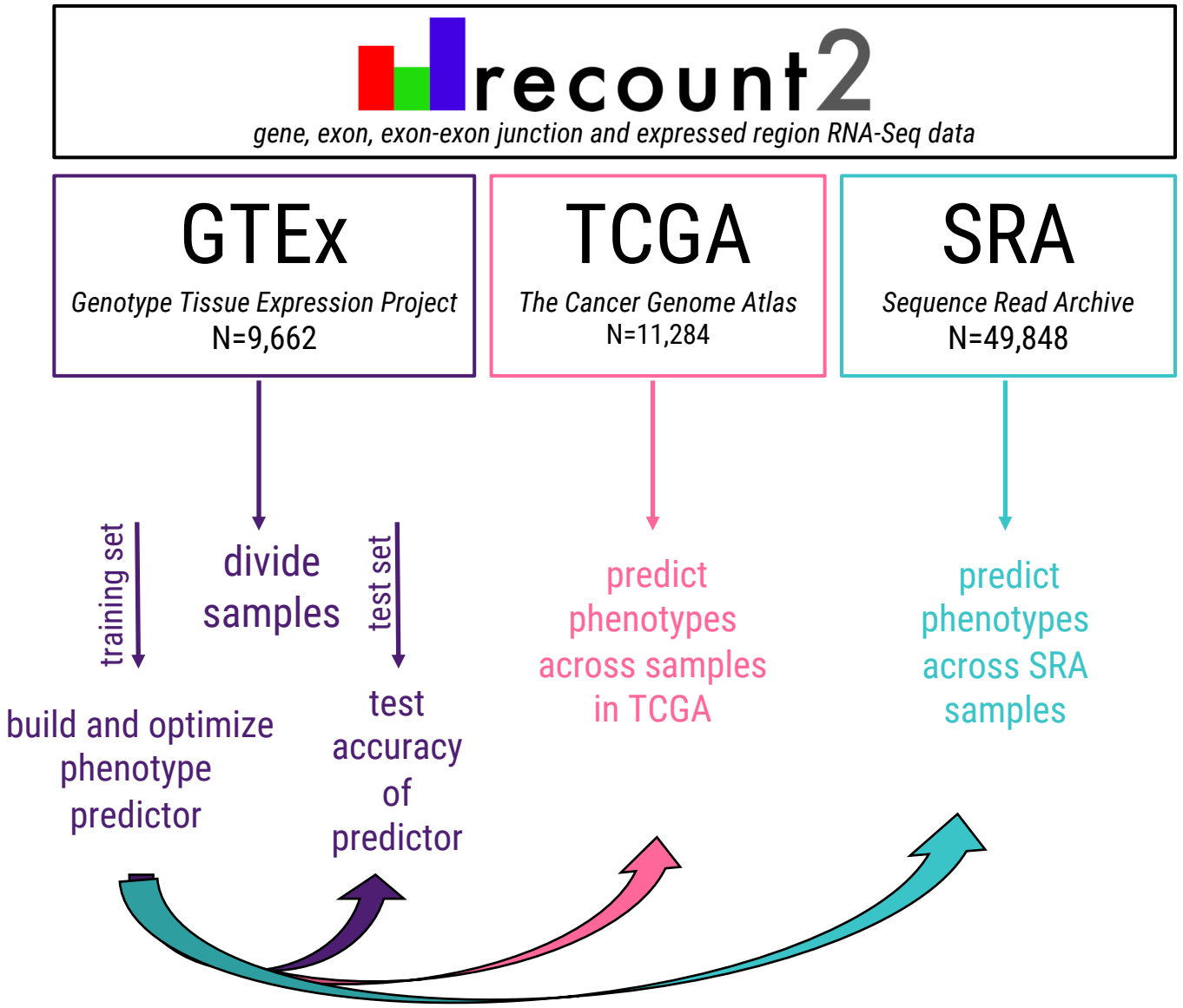
Goal :

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

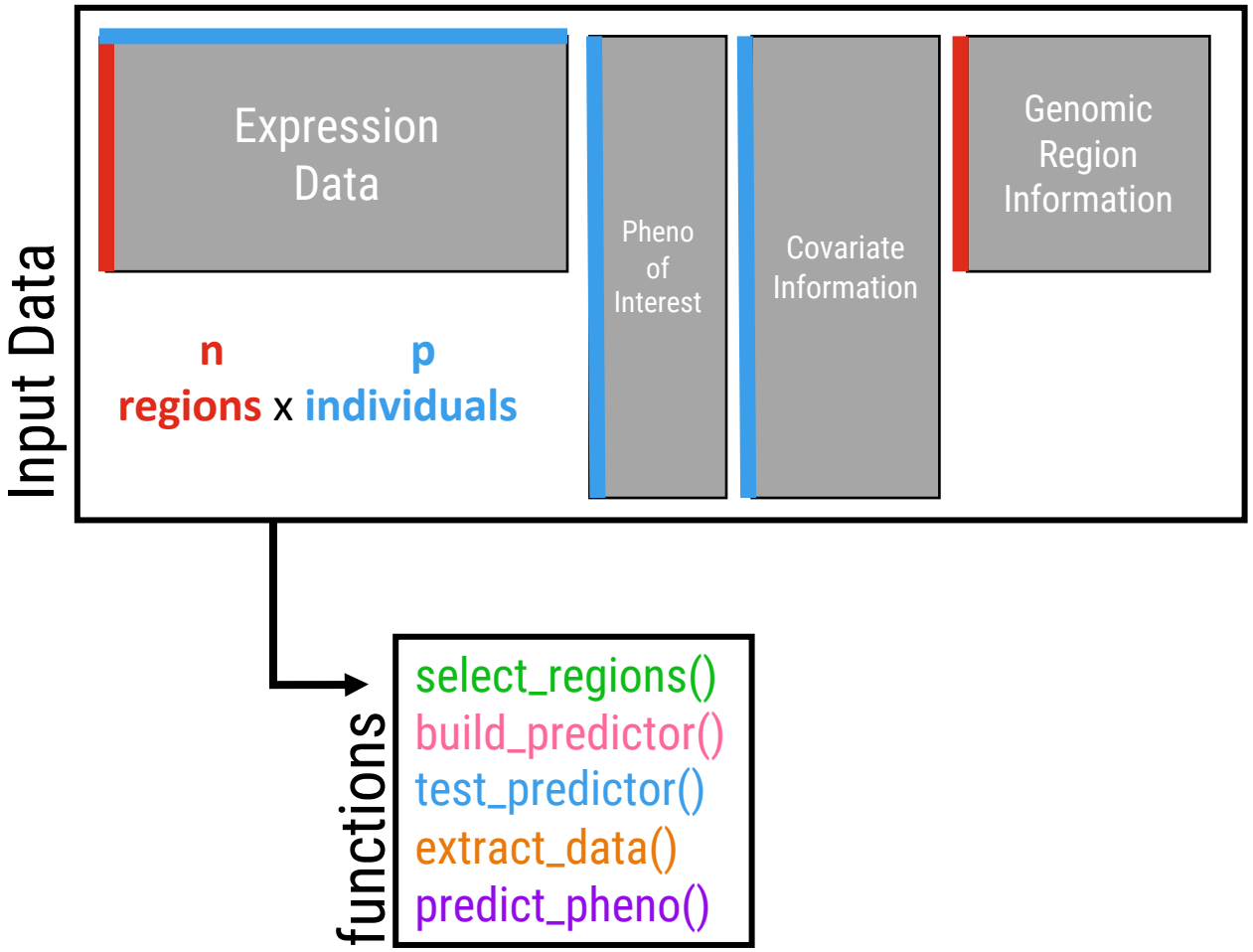


Goal :

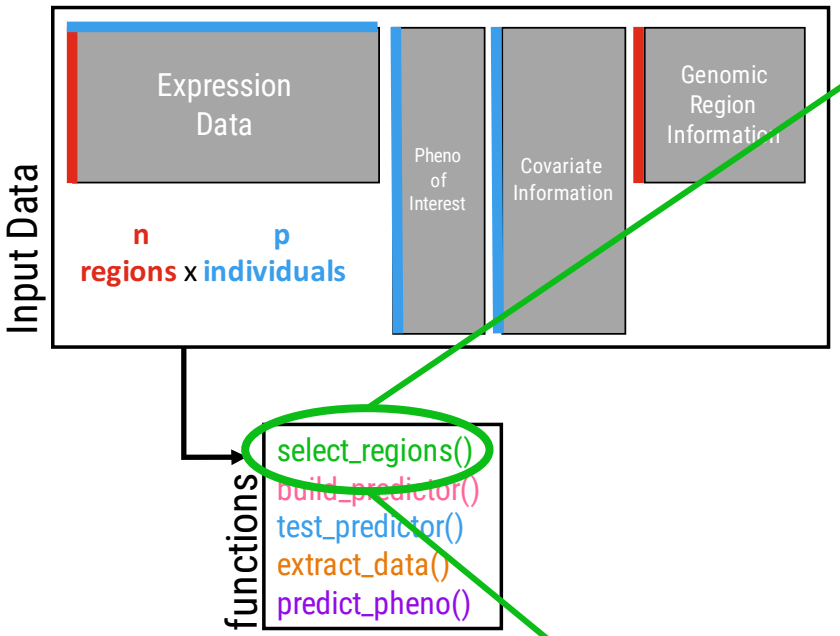
to accurately predict critical phenotype information for all samples in *recount*

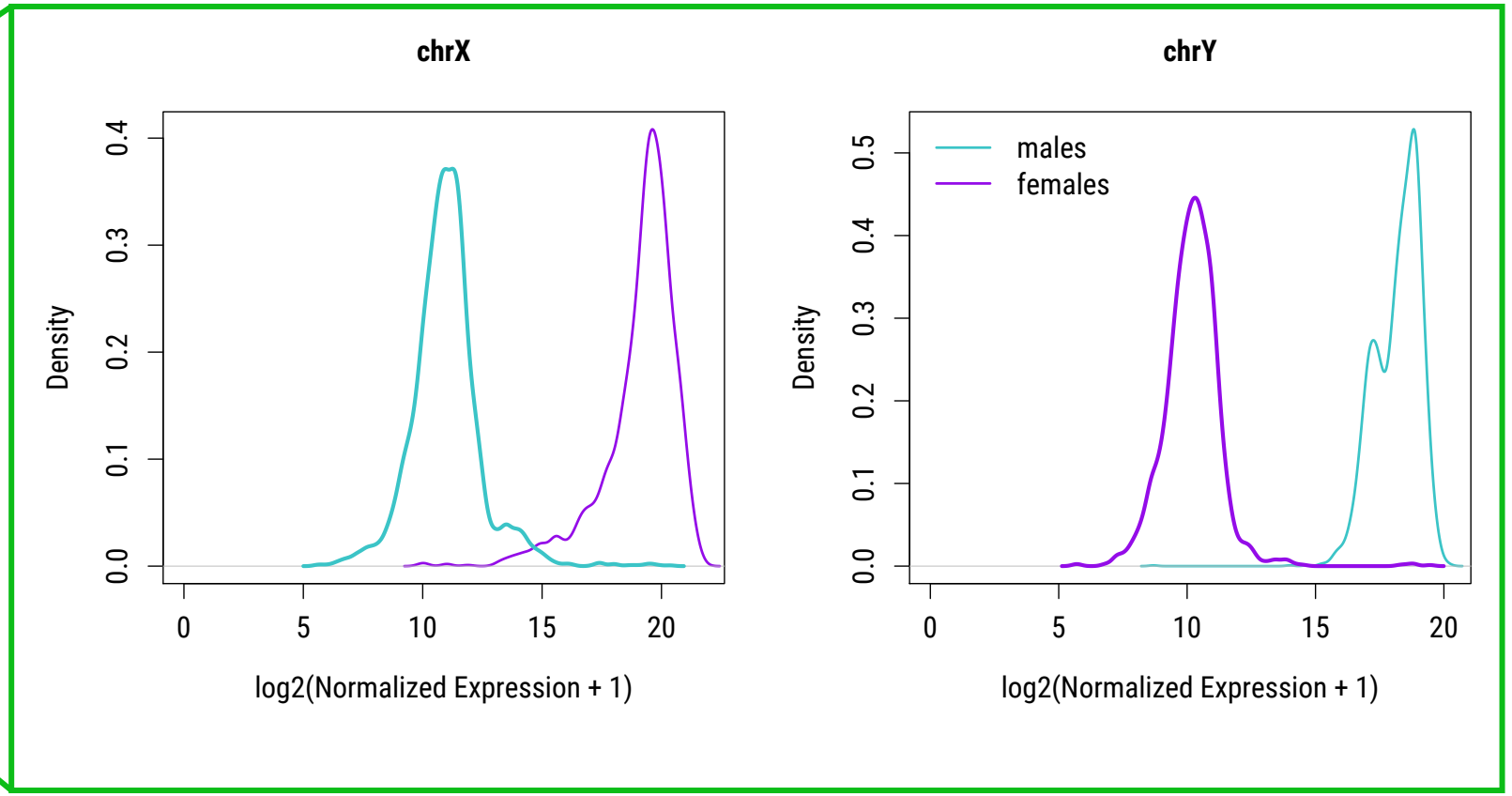


phenopredict



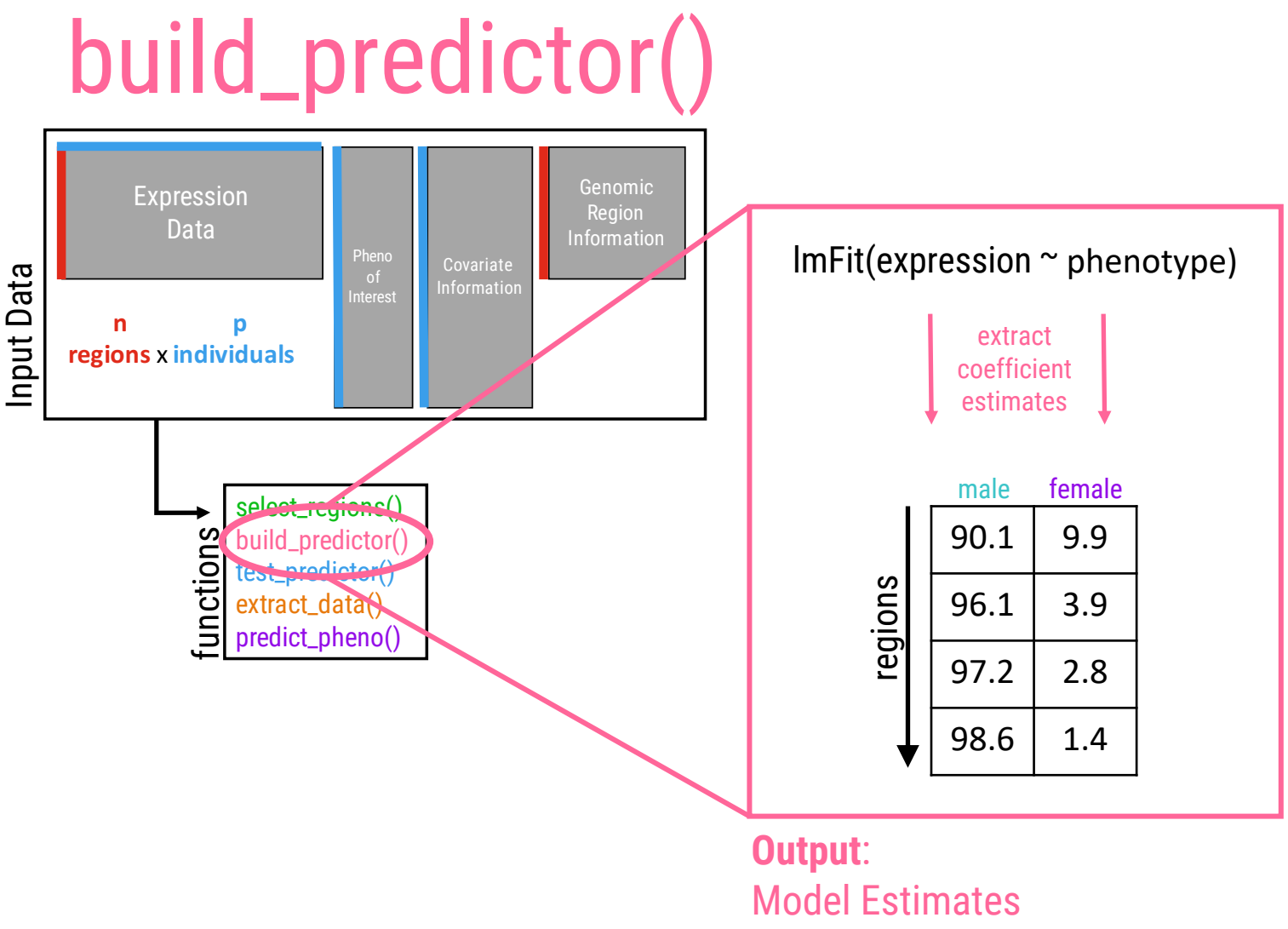
select_regions()





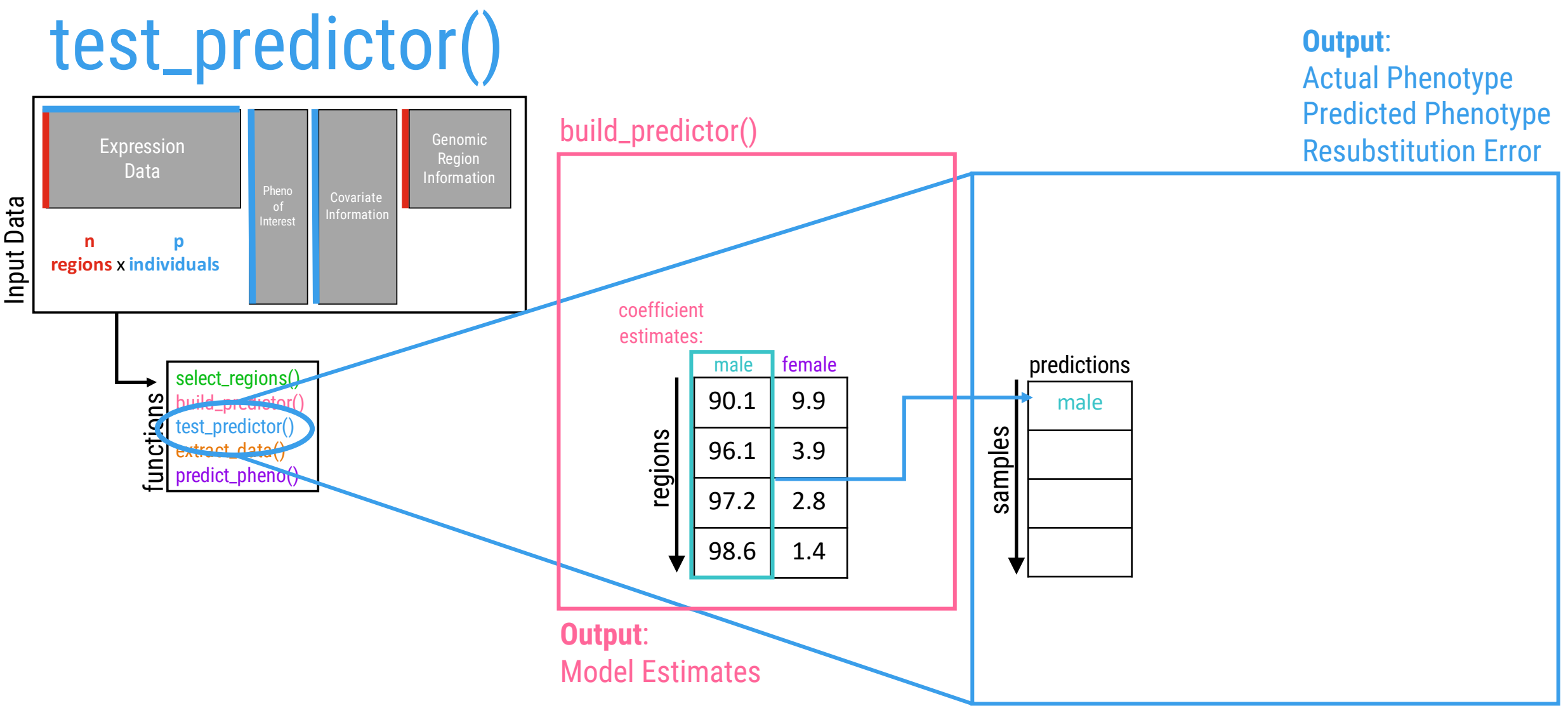
Output:
Coverage matrix (data.frame)
Region information (GRanges)

build_predictor()

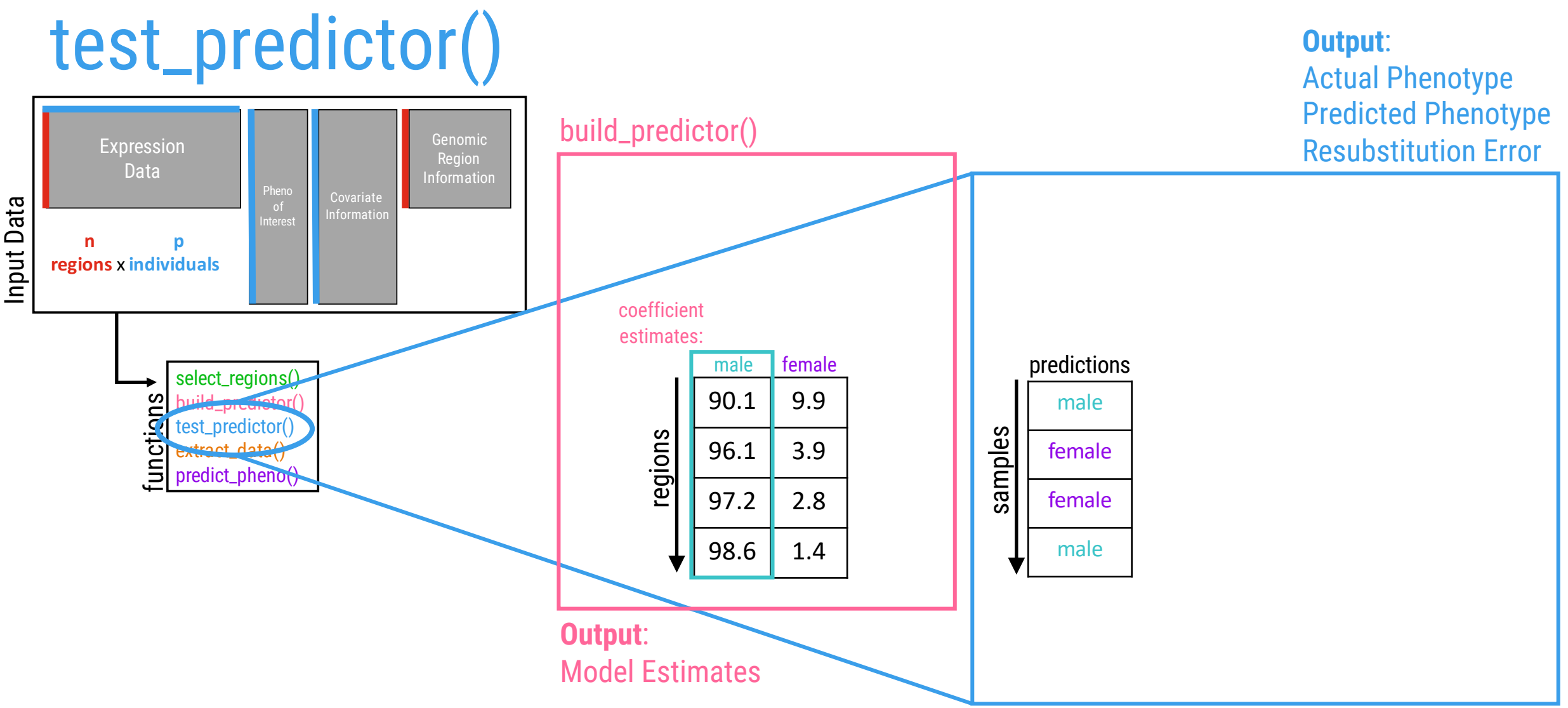


Output:
Model Estimates

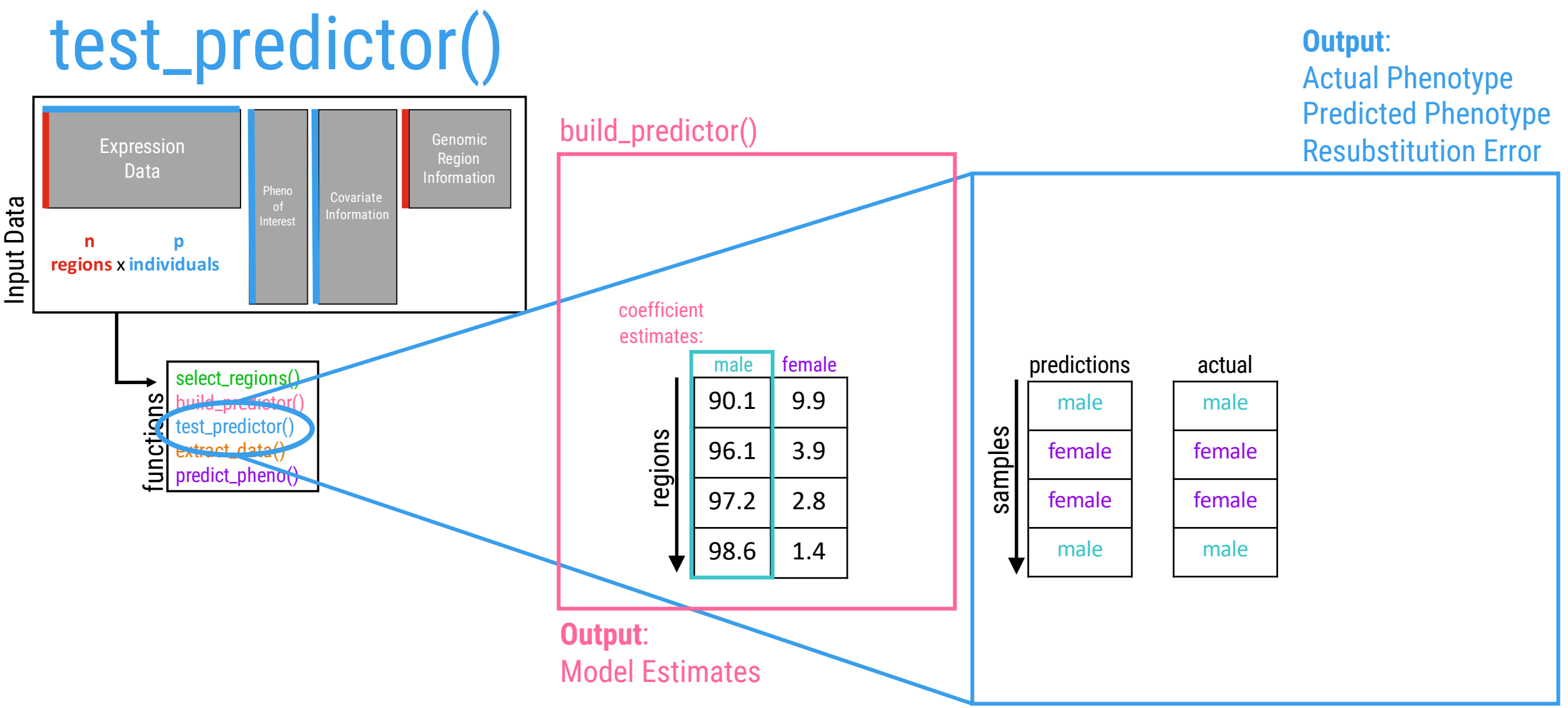
test_predictor()



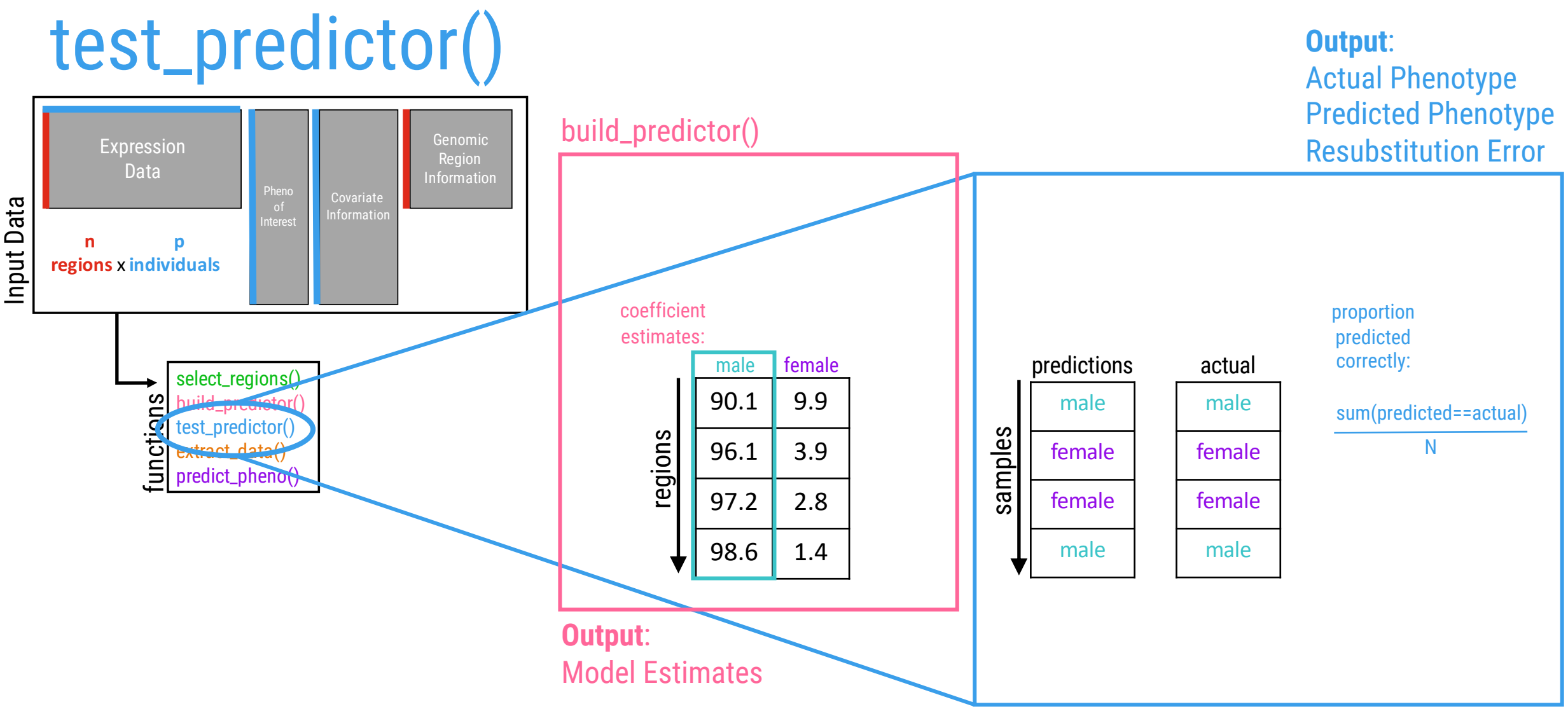
test_predictor()



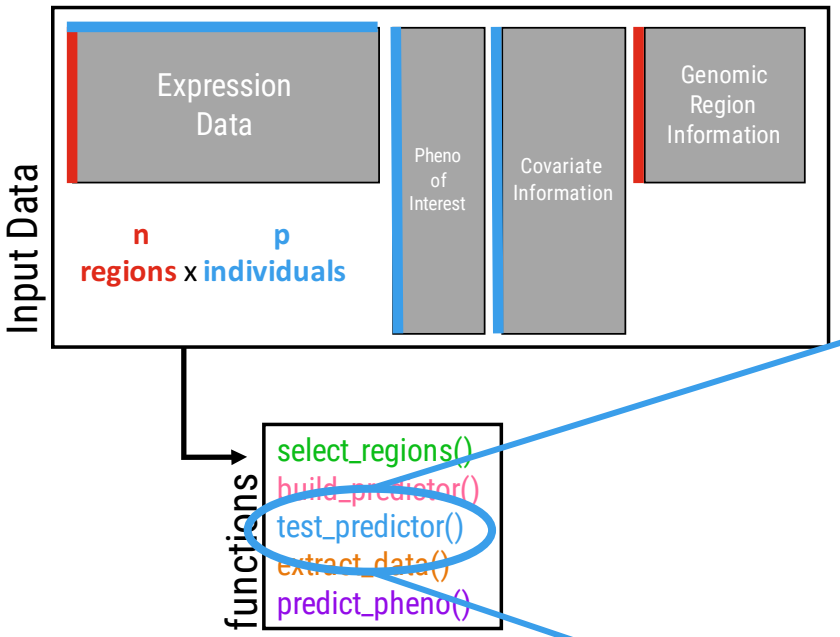
test_predictor()



test_predictor()



test_predictor()



build_predictor()

coefficient estimates:

regions	male	female
	90.1	9.9
	96.1	3.9
	97.2	2.8
	98.6	1.4

Output:
Model Estimates

Output:
Actual Phenotype
Predicted Phenotype
Resubstitution Error

samples	predictions	actual
	male	male
	female	female
	female	female
	male	male

proportion predicted correctly:

$$\frac{\text{sum(predicted==actual)}}{N}$$

$$\frac{\text{(predicted==actual) = 4}}{N=4}$$

100%

extract_data()

Input Data

Expression Data

n regions x p individuals

Pheno of Interest

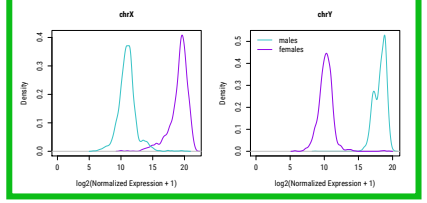
Covariate Information

Genomic Region Information

functions

select_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()

select_regions()



Extract expression estimates for regions used to build the predictor

test set samples

expression estimates

Output:
Coverage matrix (data.frame)
Region information (GRanges)

extract_data()

Input Data

Expression Data

n regions x p individuals

Pheno of Interest

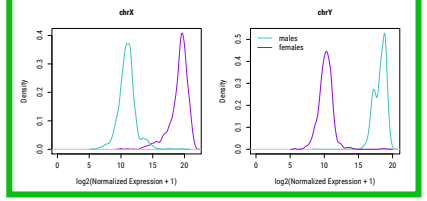
Covariate Information

Genomic Region Information

functions

select_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()

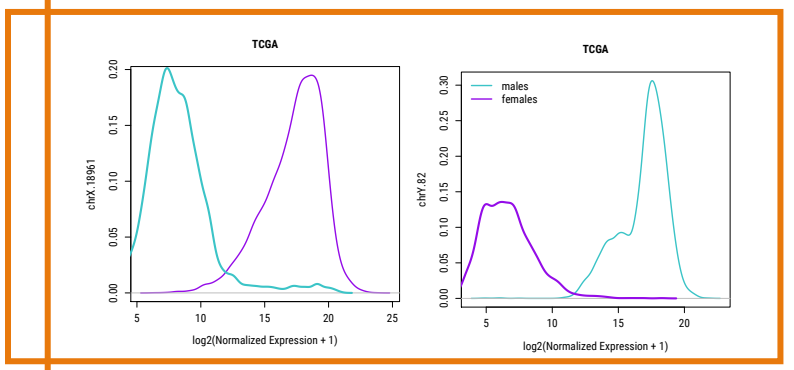
select_regions()



Extract expression estimates for regions used to build the predictor

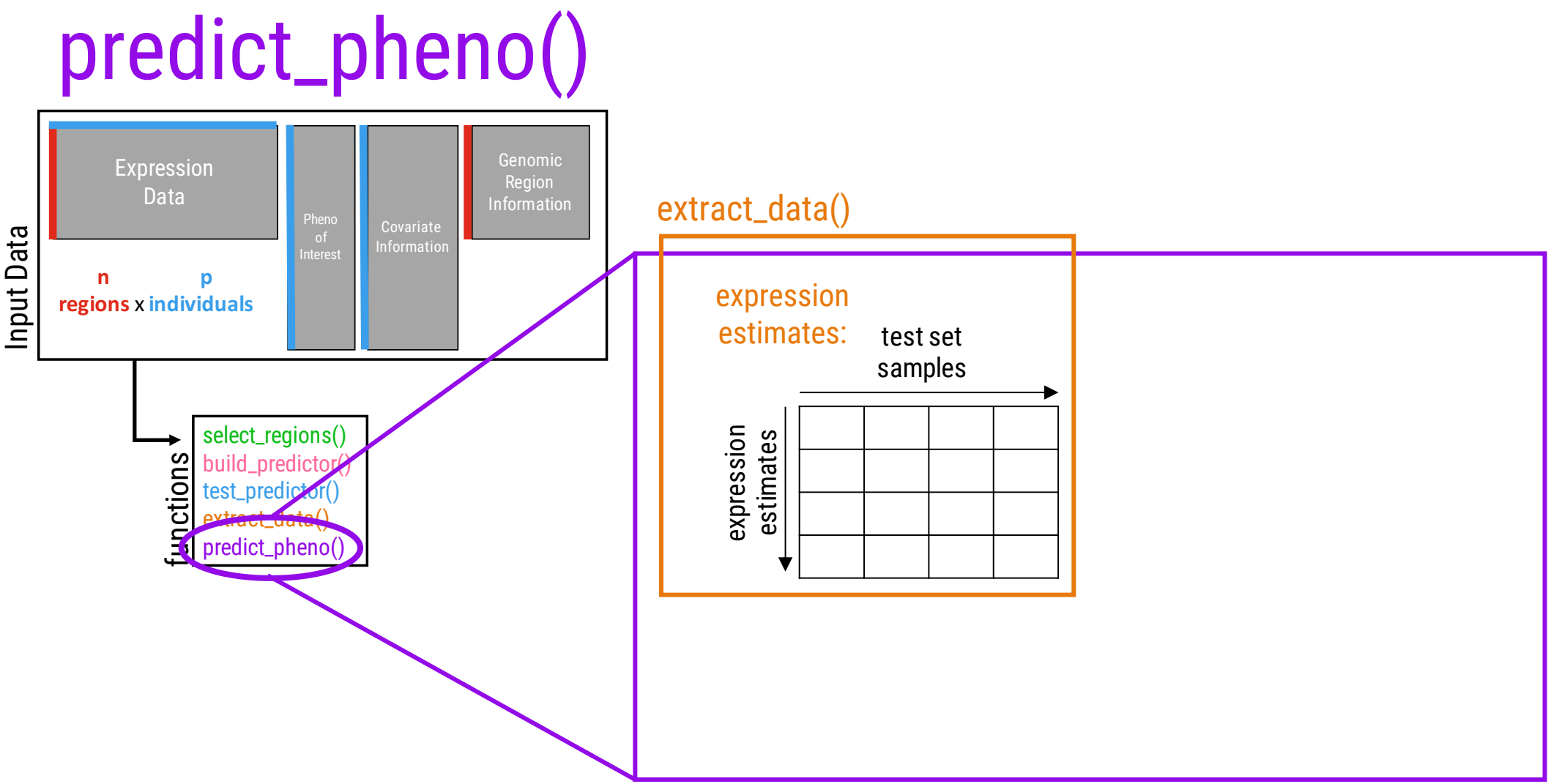
test set samples

expression estimates

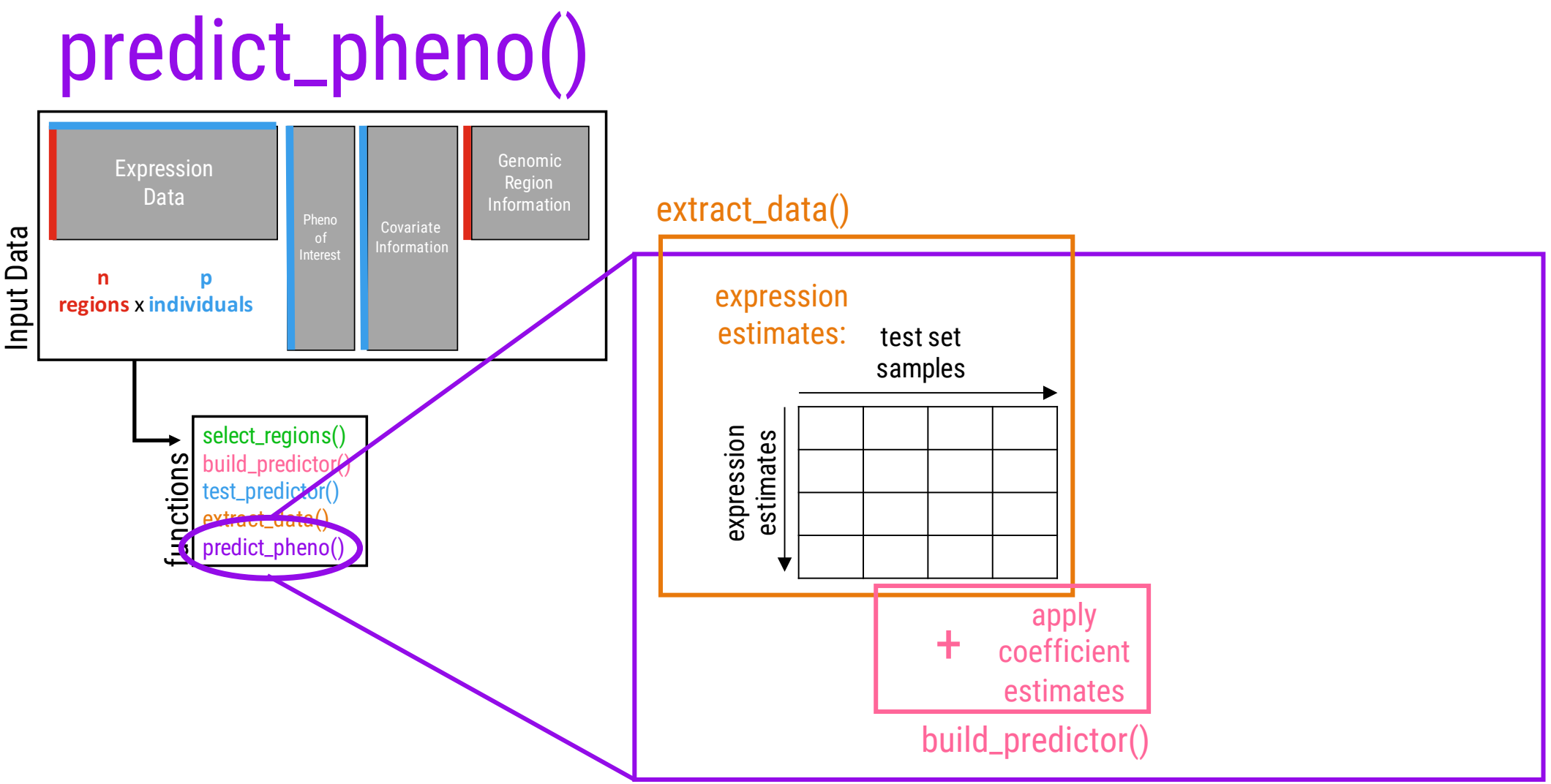


Output:
Coverage matrix (data.frame)
Region information (GRanges)

predict_pheno()

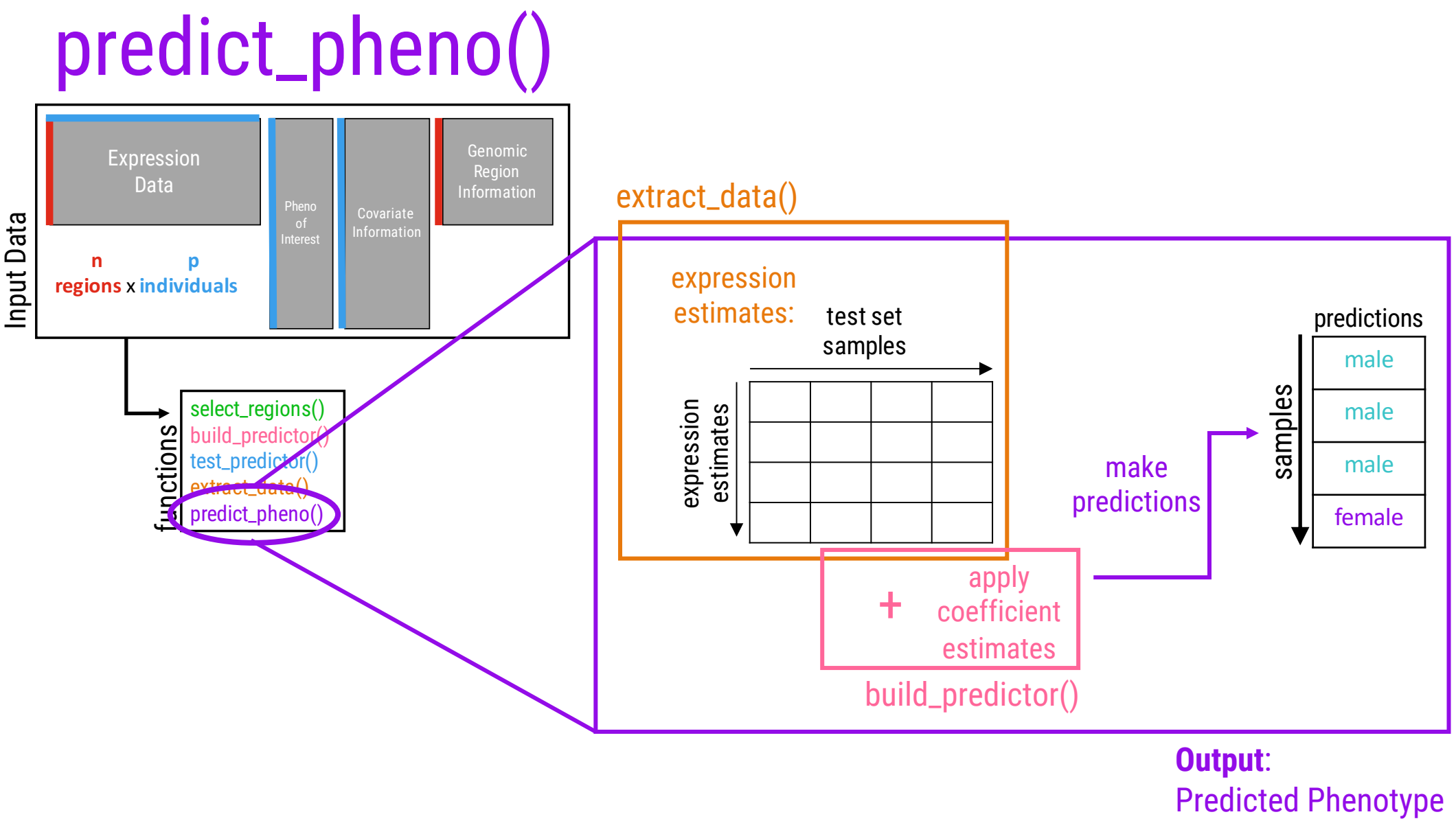


predict_pheno()



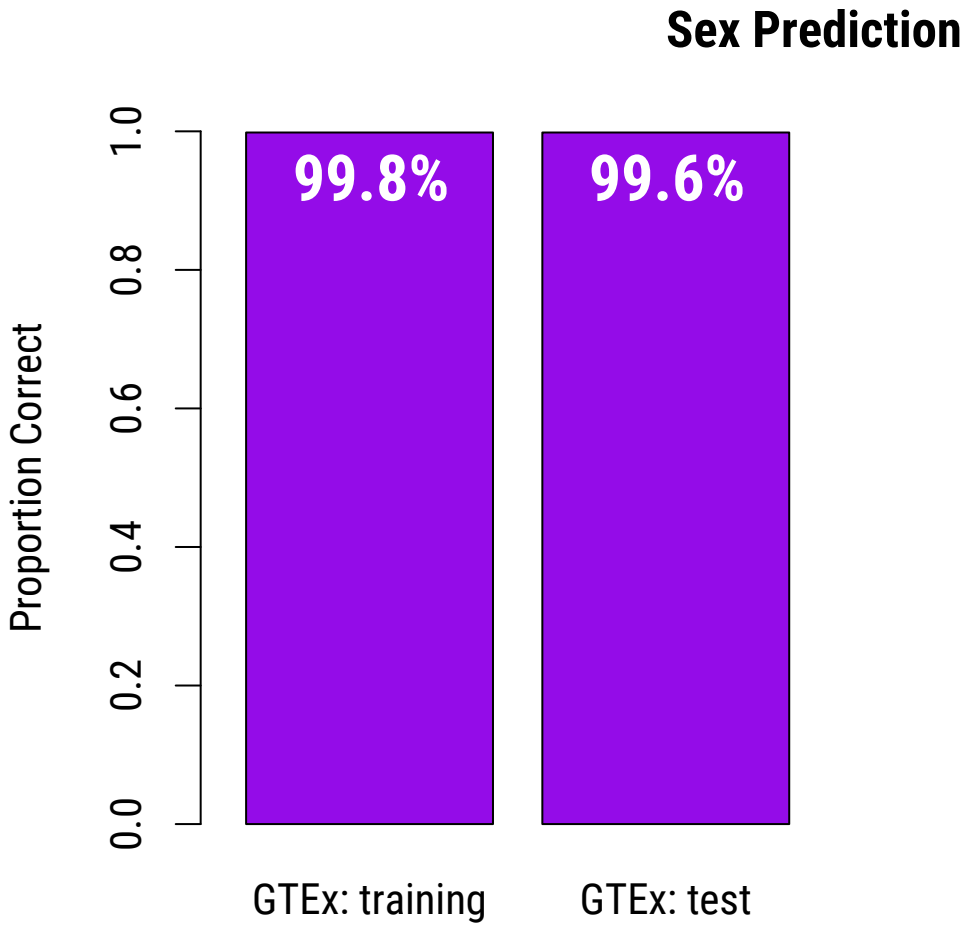
Output:
Predicted Phenotype

predict_pheno()



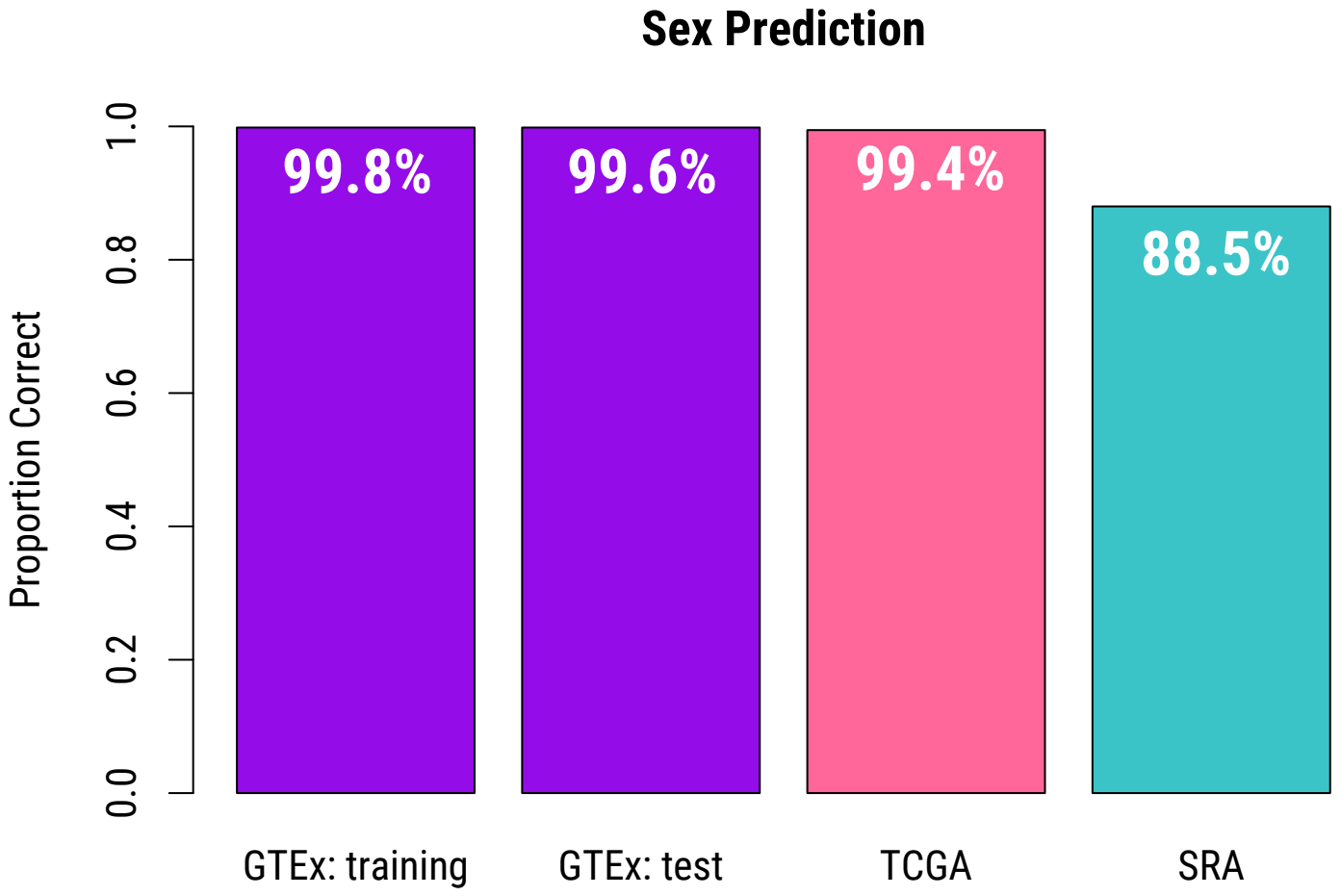
Output:
Predicted Phenotype

Sex prediction is accurate across data sets

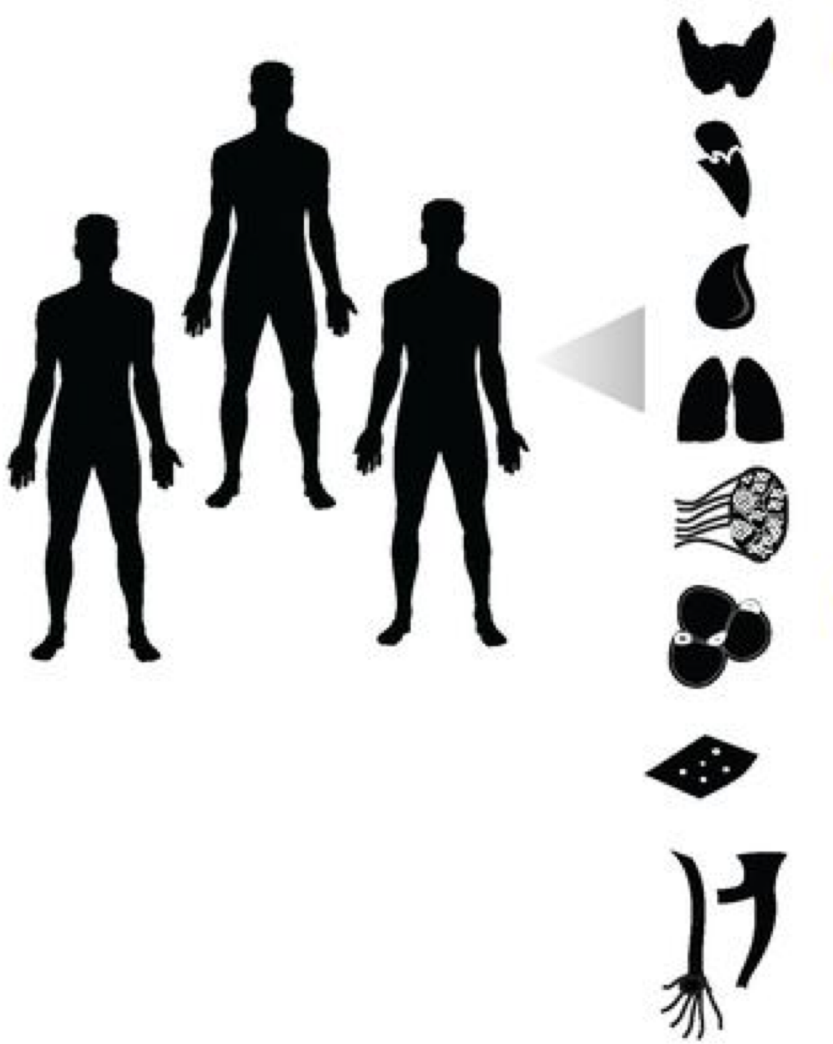


Number of Regions	**20**	**20**
Number of Samples (N)	**4,769**	**4,769**

Sex prediction is accurate across data sets



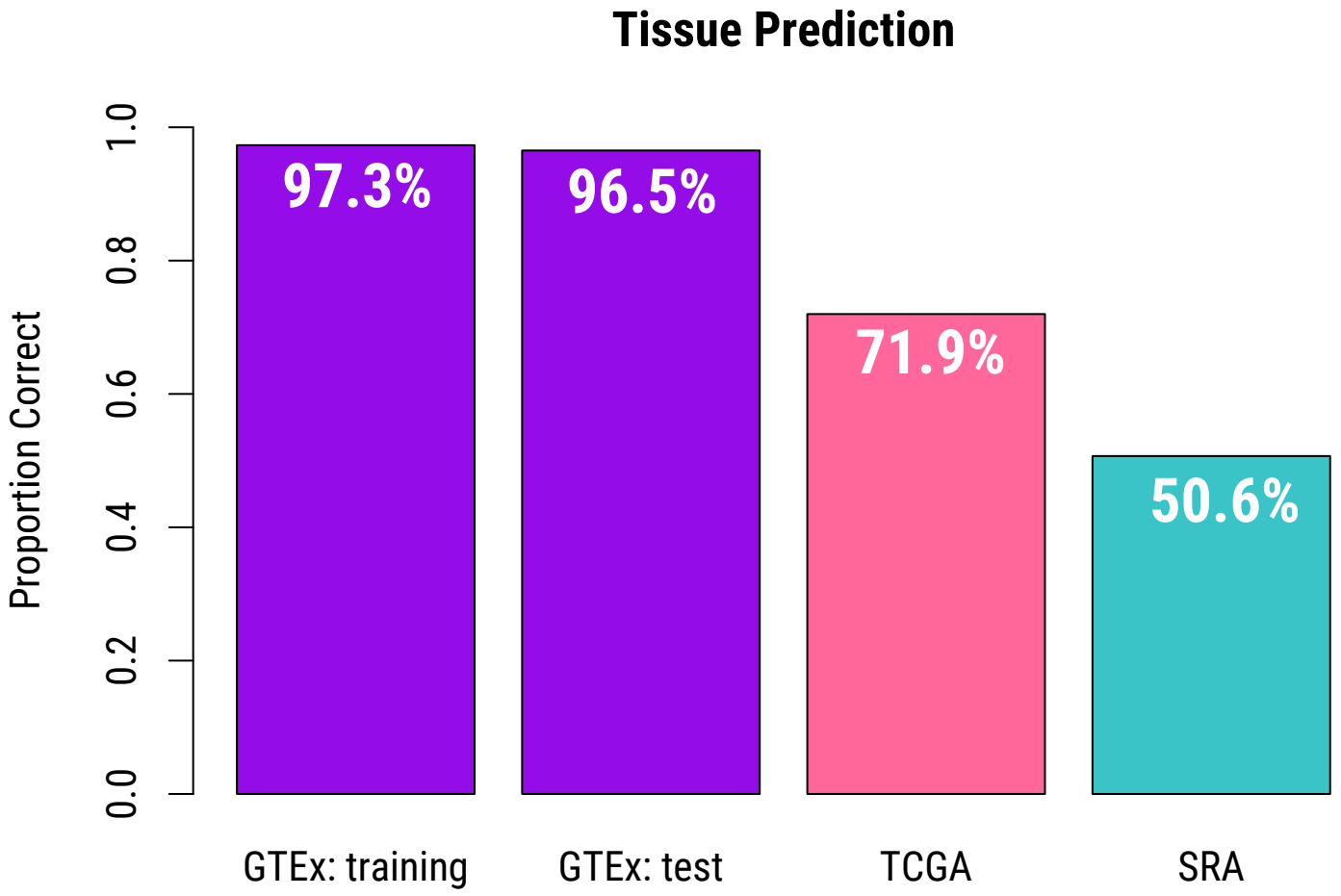
Number of Regions	20	20	20	20
Number of Samples (N)	4,769	4,769	11,245	3,640



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

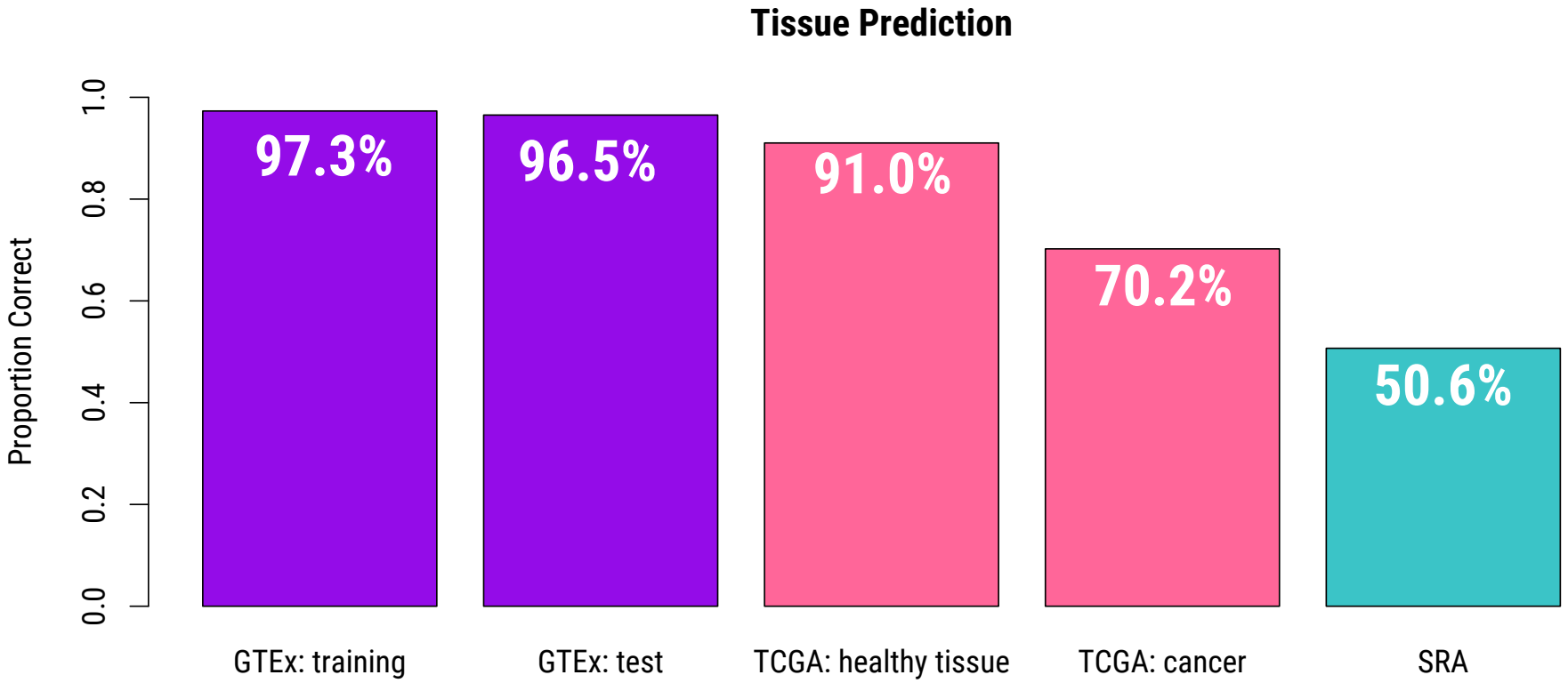
Can we use expression data to predict tissue?

Tissue prediction is accurate across data sets



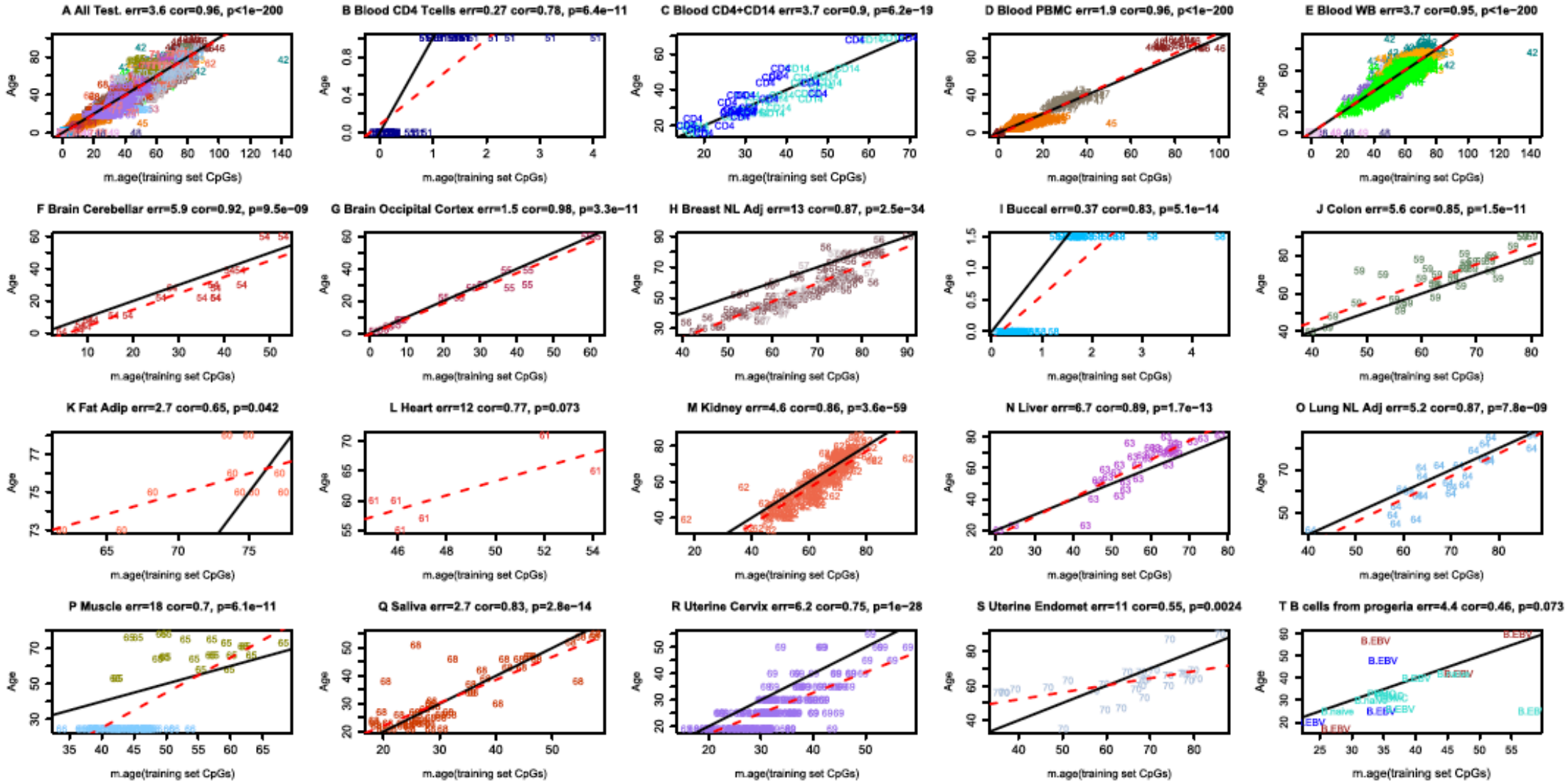
Number of Regions	**589**	**589**	**589**	**589**
Number of Samples (N)	**4,769**	**4,769**	**7,193**	**8,951**

Prediction is more accurate in healthy tissue



Number of Regions	589	589	589	589	589
Number of Samples (N)	4,769	4,769	613	6,579	8,951

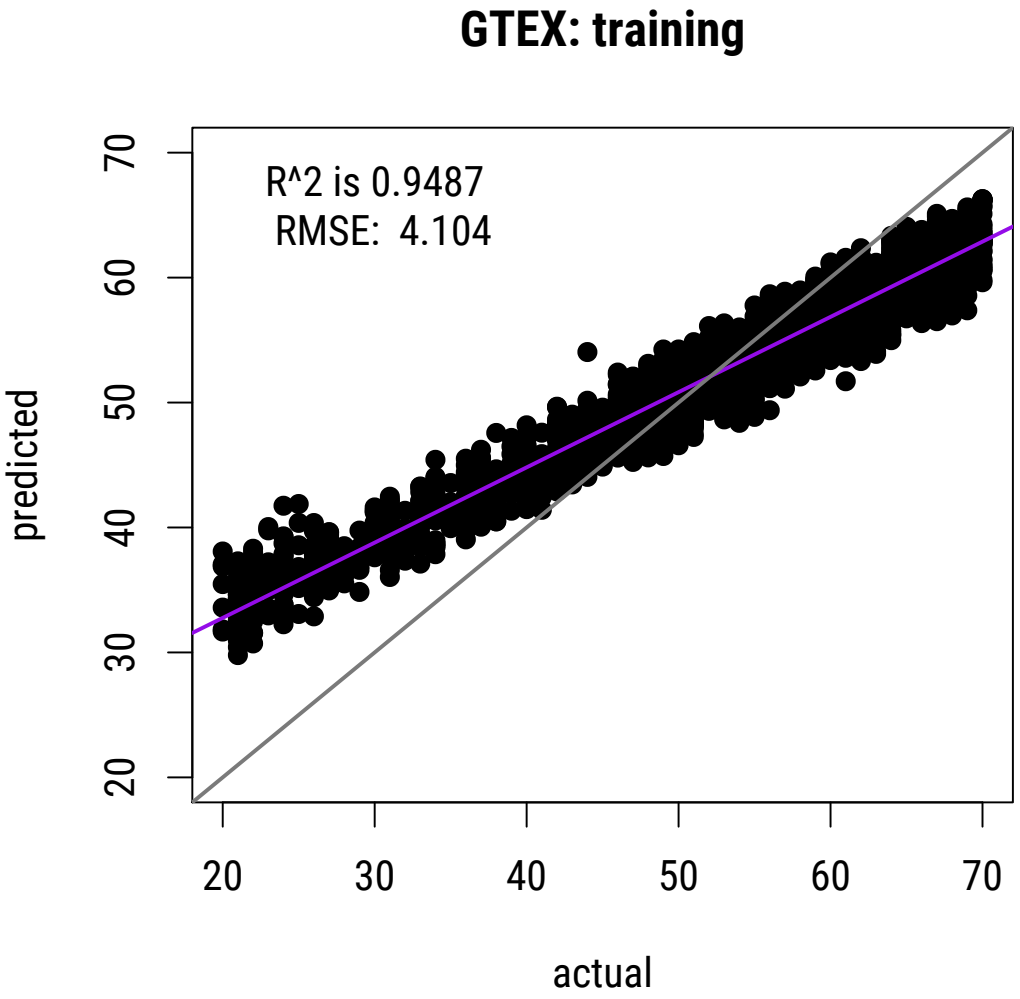
Horvath demonstrates that 353 CpGs can accurately predict age



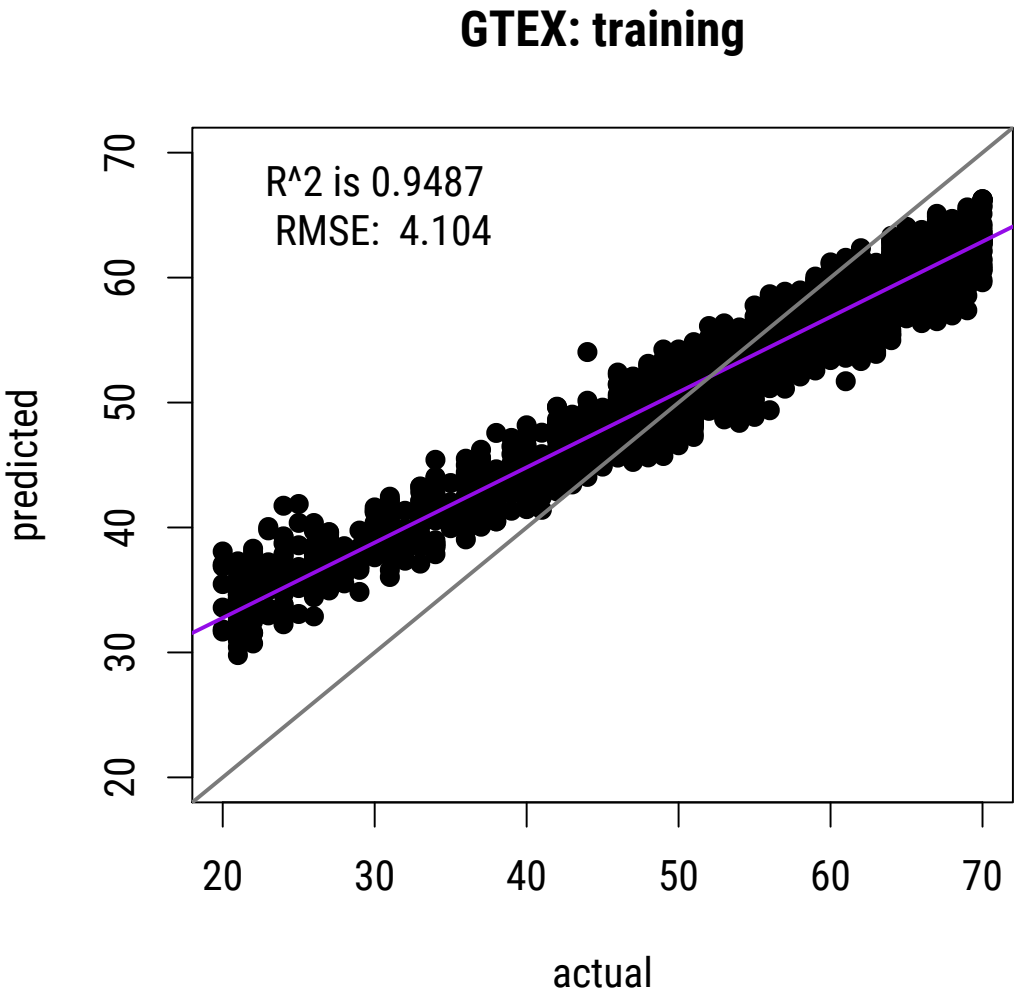
Horvath. *Genome Biology* (2013)

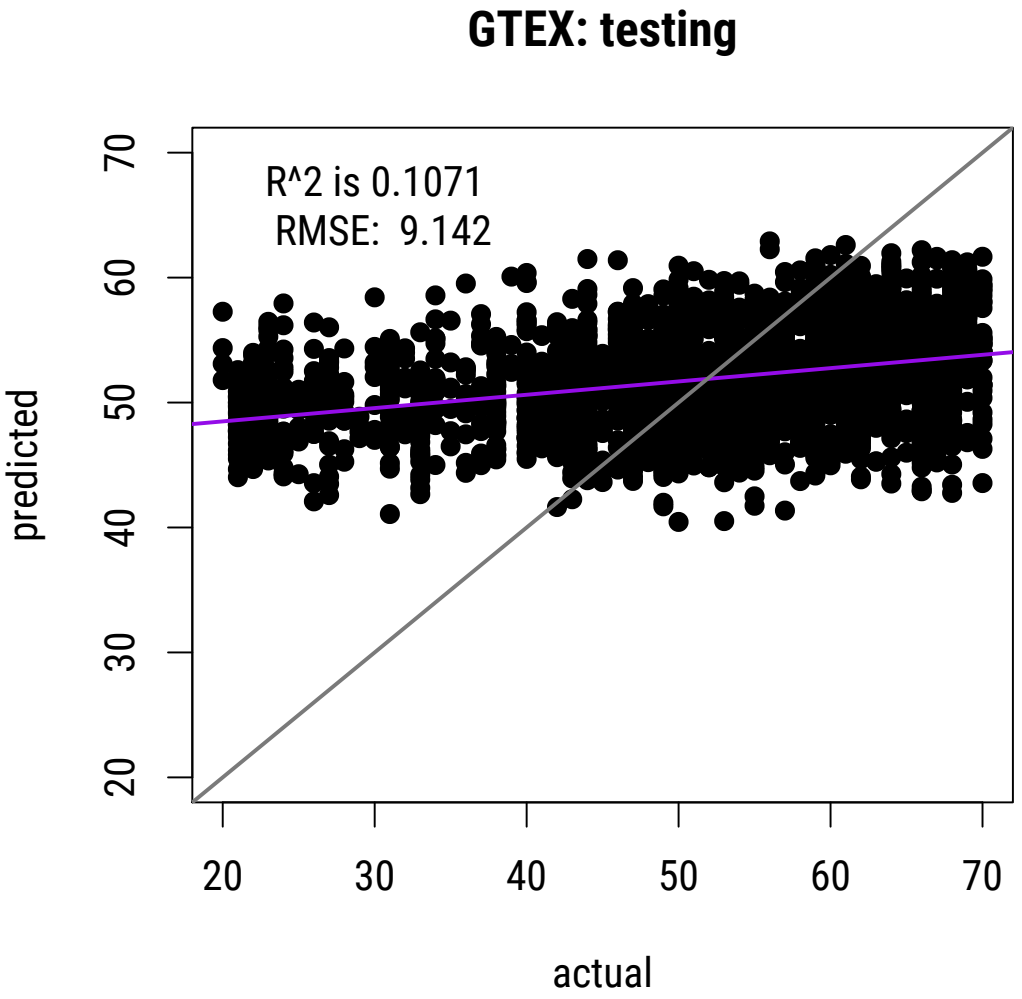
Can we predict age from gene expression data?

How well can we predict age in GTEx?

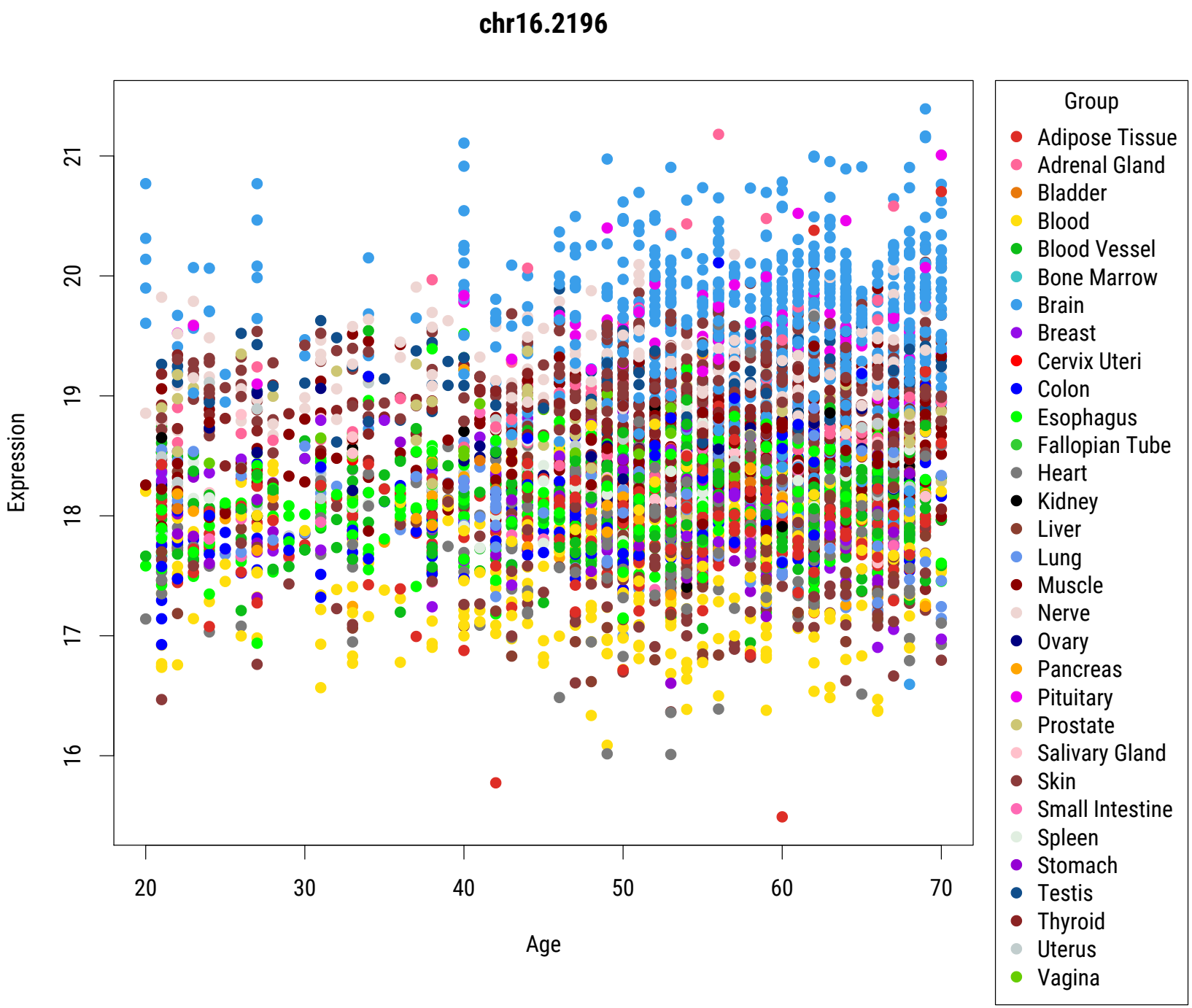


How well can we predict age in GTEx?

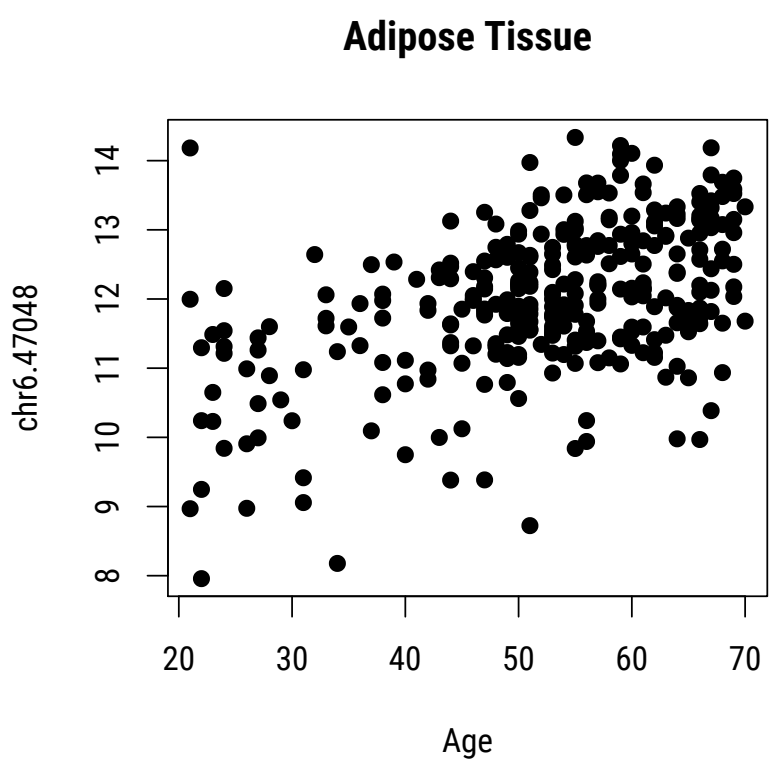


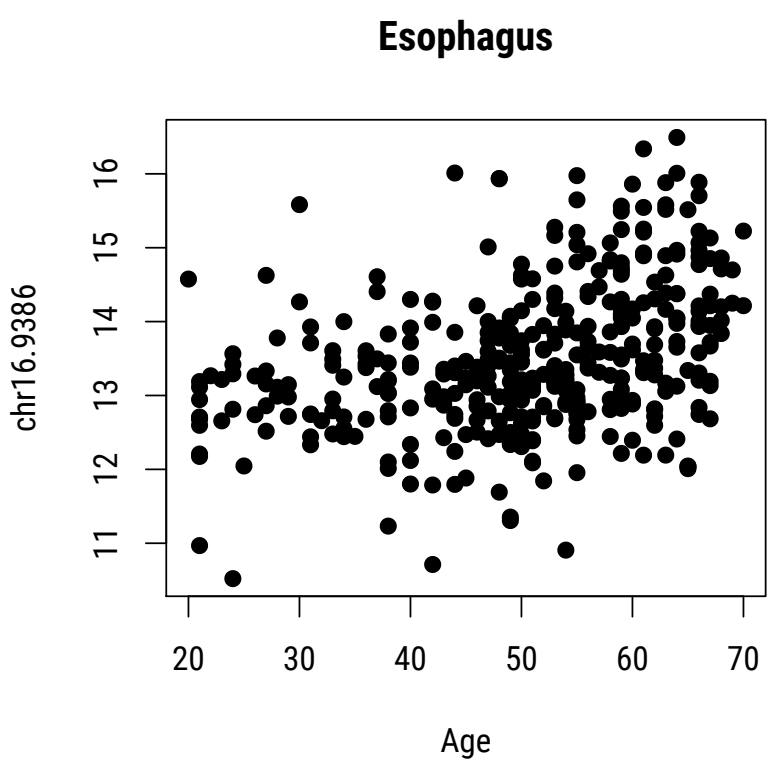


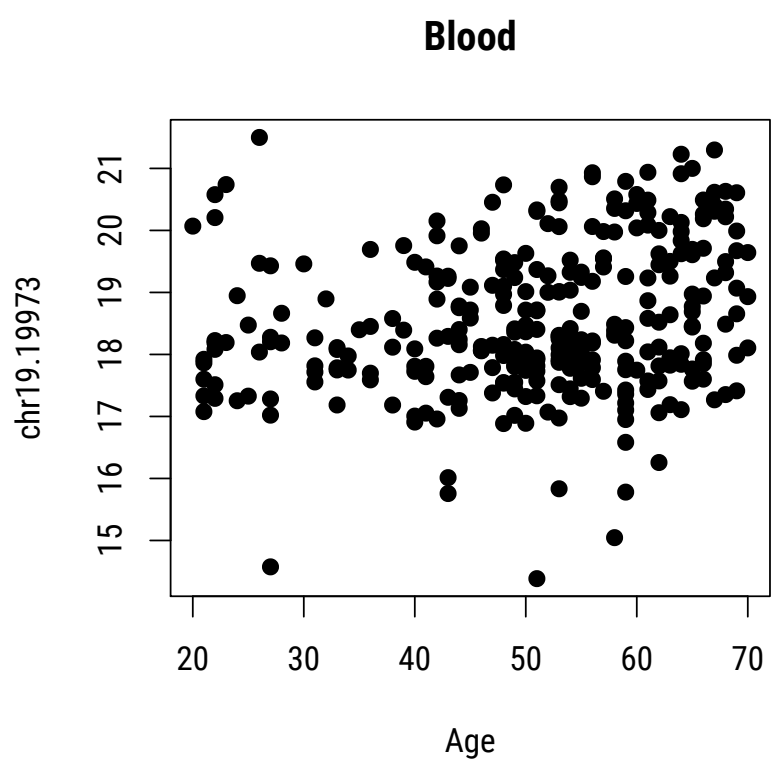
Tissue poses a problem for prediction…



Even within tissue, signal is pretty weak…

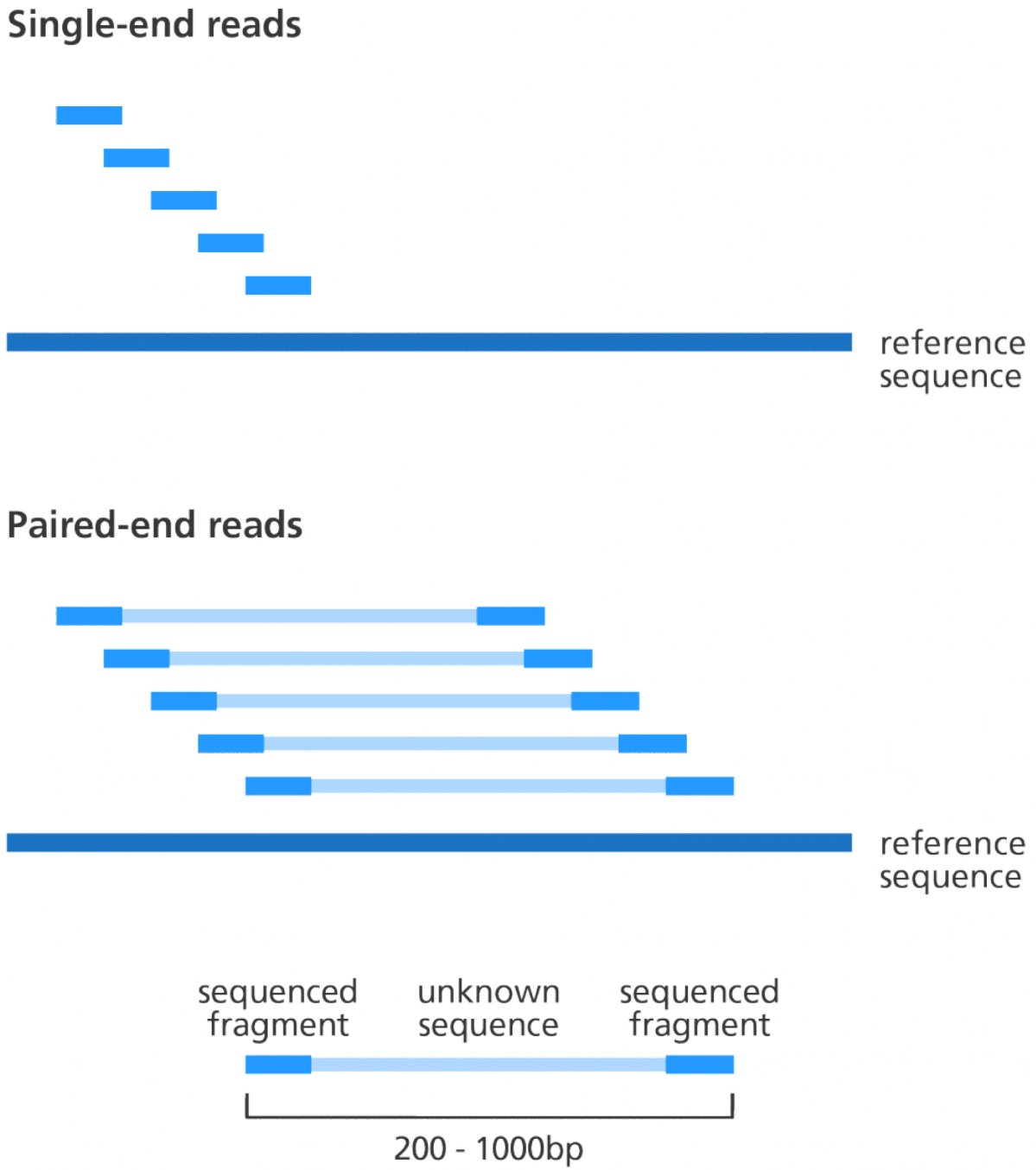




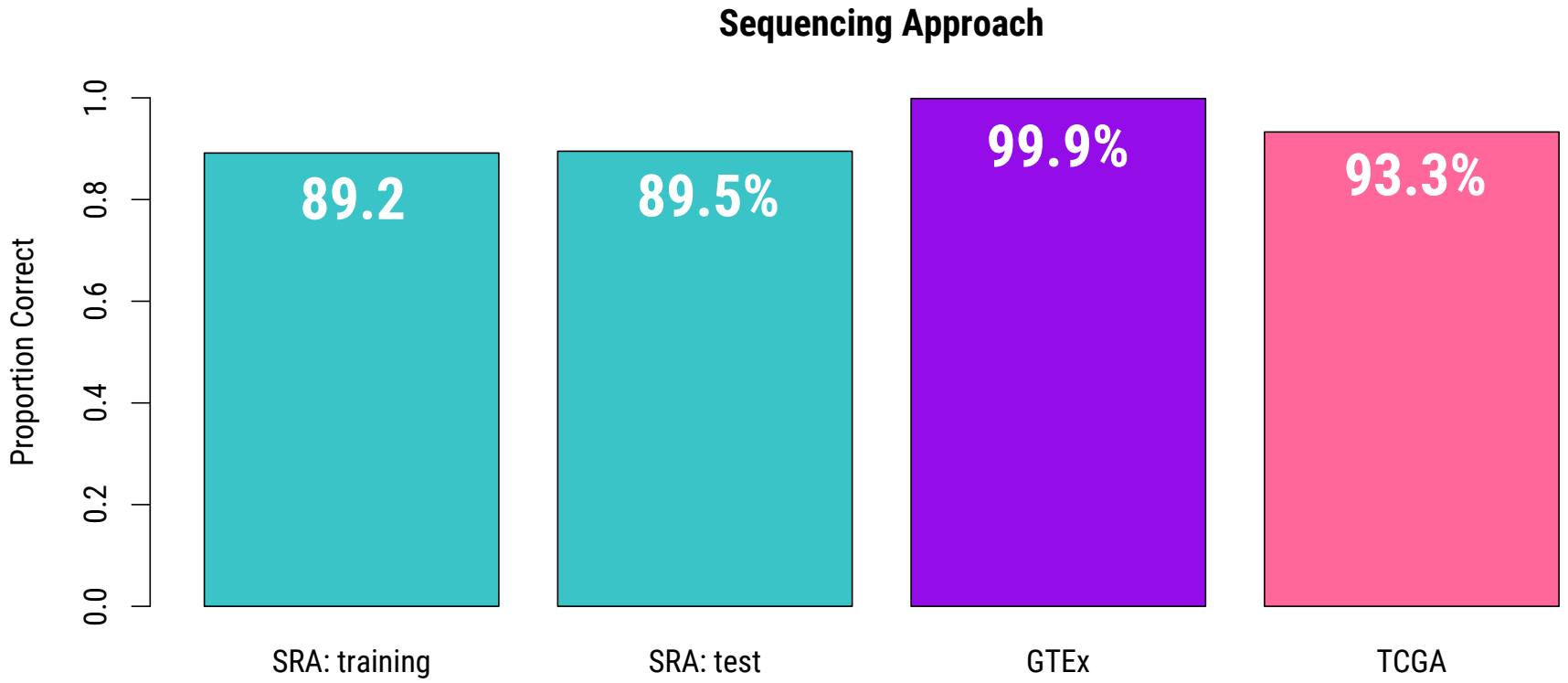


What about predicting a technical aspect of sequencing?

Can we predict which sequencing approach was employed from expression data?



Sequencing approach prediction is accurate across data sets



Number of Regions	80	80	80	80
Number of Samples (N)	24,829	24,828	9,538	9,929

predictions (v0.0.01)

sample_id	study	pred_sex	accuracy_sex	pred_tissue	accuracy_tissue	pred_PE_SE	accuracy_PE_SE
SRR660824	gtex	male	0.999	Lung	0.961	PAIRED	0.999
SRR2166176	gtex	male	0.998	Brain	0.951	PAIRED	0.999
SRR606939	gtex	female	0.999	Heart	0.961	PAIRED	0.999
SRR2167642	gtex	male	0.999	Brain	0.961	PAIRED	0.999
SRR2165473	gtex	male	0.999	Skin	0.961	PAIRED	0.999

If you want to…	
Align RNA-Seq data	Rail-RNA Scalable RNA-seq alignment
Learn about human expression	recount2
Predict phenotype information	*phenopredict* *https://github.com/ShanEllis/phenopredict*

The Leek group

- Jack Fu
- Sean Kross
- Leslie Myint
- Divya Narayanan
- Claire Ruberman
- **Jeff Leek**

Collaborators

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

http://rail.bio
$ ./install-rail-rna-V

https://jhubiostatistics.shinyapps.io/recount/
> biocLite("recount")