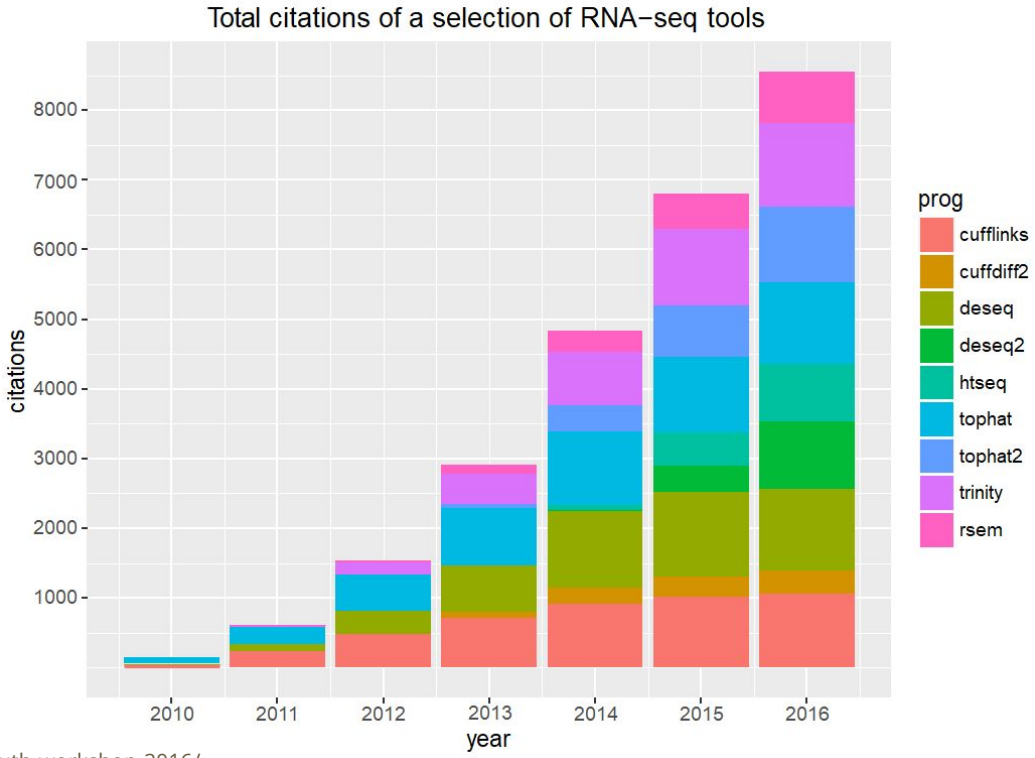
Analyze of multimodality of gene expression profiles using publicly available databases

Amir Vajdi

Outline

- Kallisto Method
- Running Kallisto
- TCGA and GTEx database
- Multimodality Analysis

The growth of RNA-seq



https://pachterlab.github.io/kallisto-sleuth-workshop-2016/

kallisto-sleuth project

k-mers **a**lone **l**ose **l**ots of **i**nformation; **s**trong **t**ogether **o**nly

kallisto

quantification

Sleuth

analysis

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525.

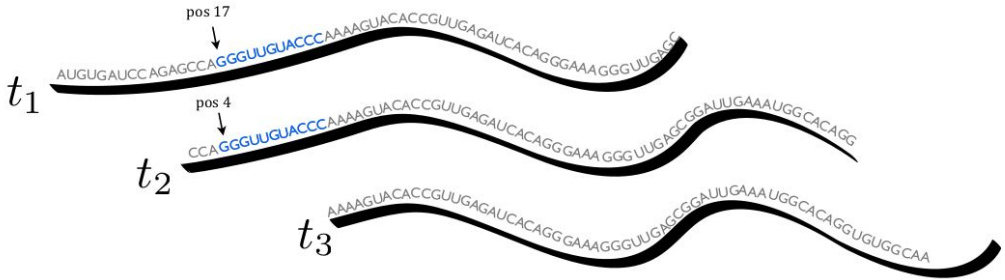
Alignment based analysis

read alignment information

read 1	GGGTTGTACCC
read 2	ATGTGATCC
read 3	CCGTTG
read 4	GAAAGGGTTG
read 5	CACAGGTGTGG

https://pachterlab.github.io/kallisto-sleuth-workshop-2016/

Alignment based analysis

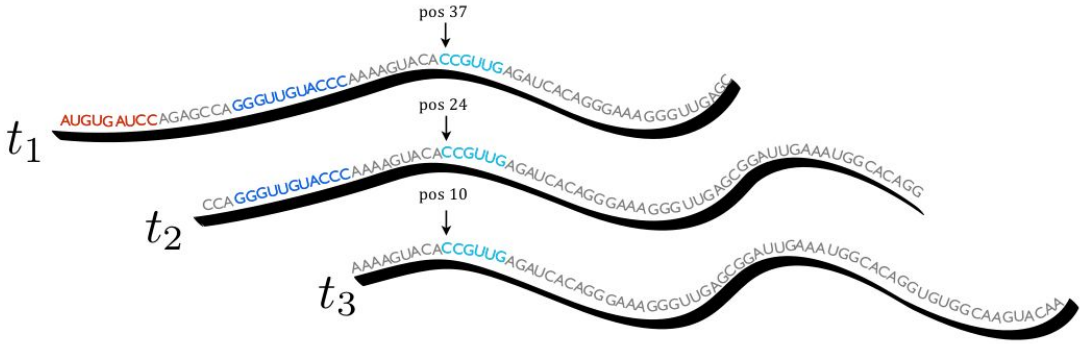


read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	
read 3	CCGTTG	
read 4	GAAAGGGTTG	
read 5	CACAGGTGTGG	

https://pachterlab.github.io/kallisto-sleuth-workshop-2016/

Alignment based analysis

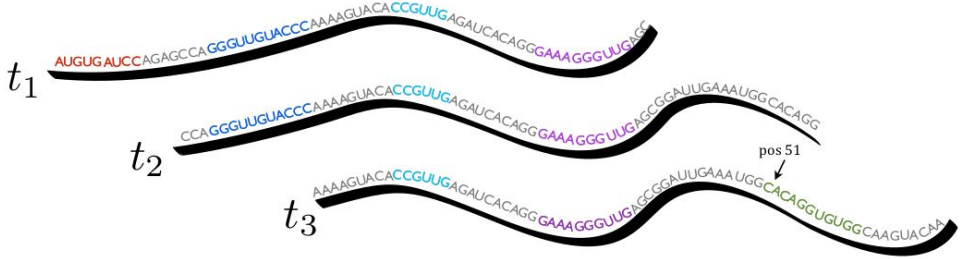


read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	
read 5	CACAGGTGTGG	

https://pachterlab.github.io/kallisto-sleuth-workshop-2016/

Alignment based analysis



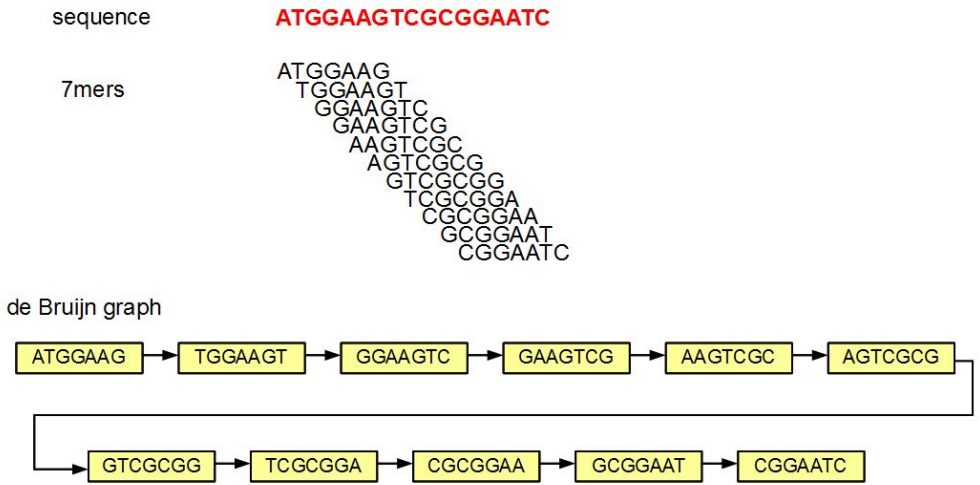
read alignment information

read 1	GGGTTGTACCC	t_1 @position 17, t_2 @position 4
read 2	ATGTGATCC	t_1 @position 1
read 3	CCGTTG	t_1 @position 37, t_2 @position 24, t_3 @position 10
read 4	GAAAGGGTTG	t_1 @position 53, t_2 @position 40, t_3 @position 26
read 5	CACAGGTGTGG	t_3 @position 51

- Even ultra-fast alignment is still pretty slow
- Alignments contain information that we don’t usually care about

De Bruijn Graph

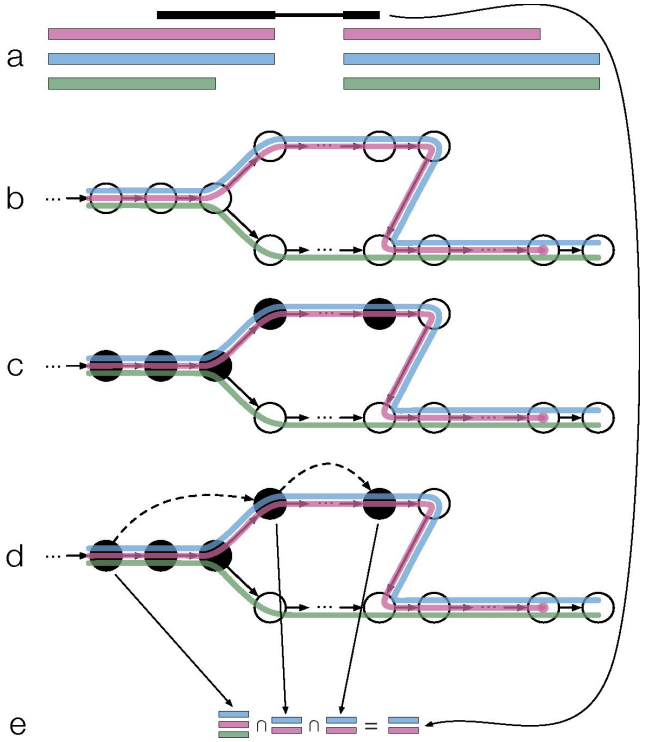
an ***n*-dimensional De Bruijn graph** of ***m*** symbols is a directed graph representing overlaps between sequences of symbols. It has m^n vertices, consisting of all possible length-***n*** sequences of the given symbols.



Kallisto Terms

- ***A pseudo-alignment of a read to a set of transcripts**, **T**, is a subset, **S** ⊆ **T** without specific coordinates mapping each base in the read to specific positions in each of the transcripts in **S**.*

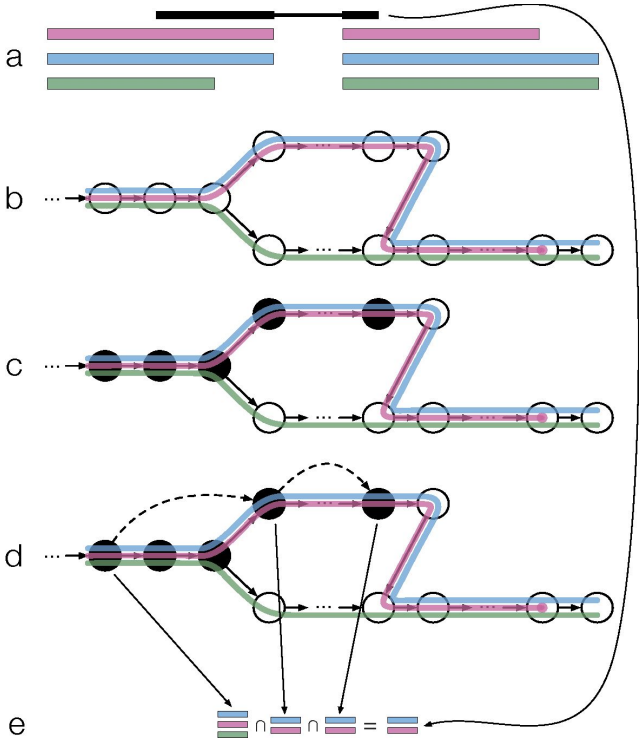
- ***A path covering of the graph**: a set of paths whose union covers all edges of the graph, where the paths correspond to transcripts.*



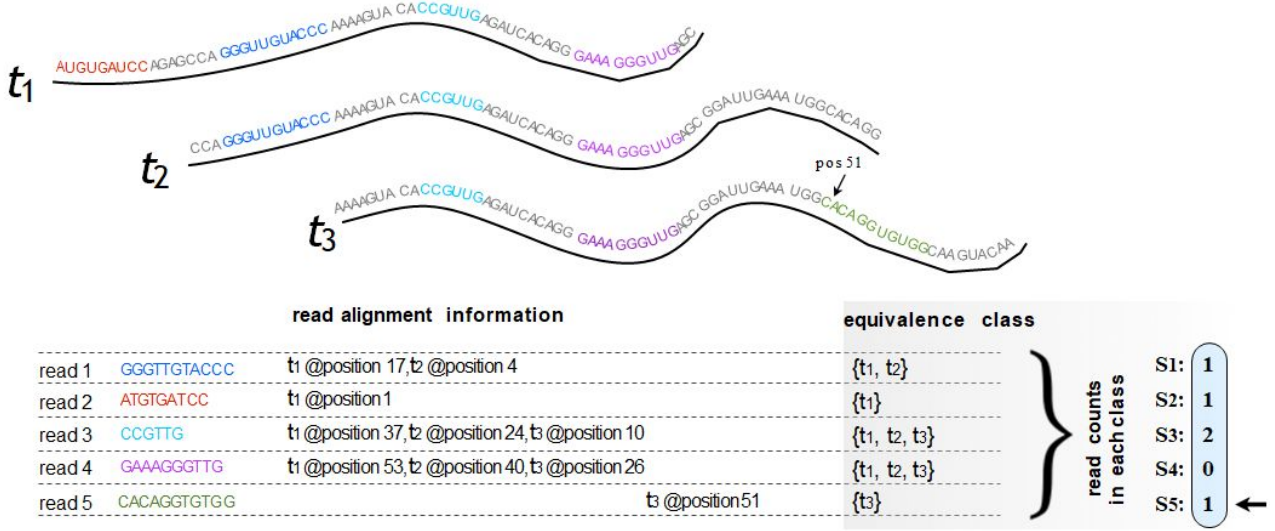
Kallisto Terms

- This path covering of a T-DBG induces multi-sets on the vertices, called ***k*-compatibility classes**
- Each read is **compatible** with one or more Transcripts
- **An equivalence class** is a group of reads that are compatible with the same set of transcripts

  A key point is that the *k*-compatibility class of an error-free read coincides with the minimal equivalence class consisting of transcripts containing the read for large *k*.



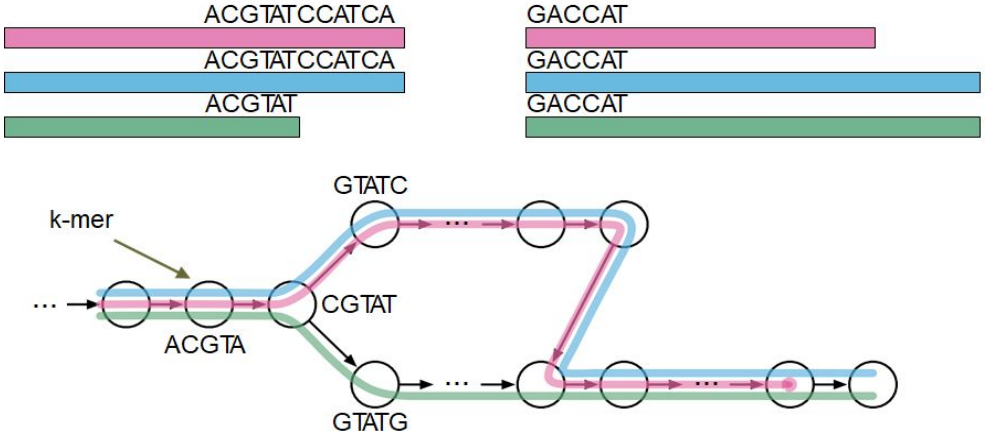
Kallisto Idea



Identifying the transcripts from which the read could have originated

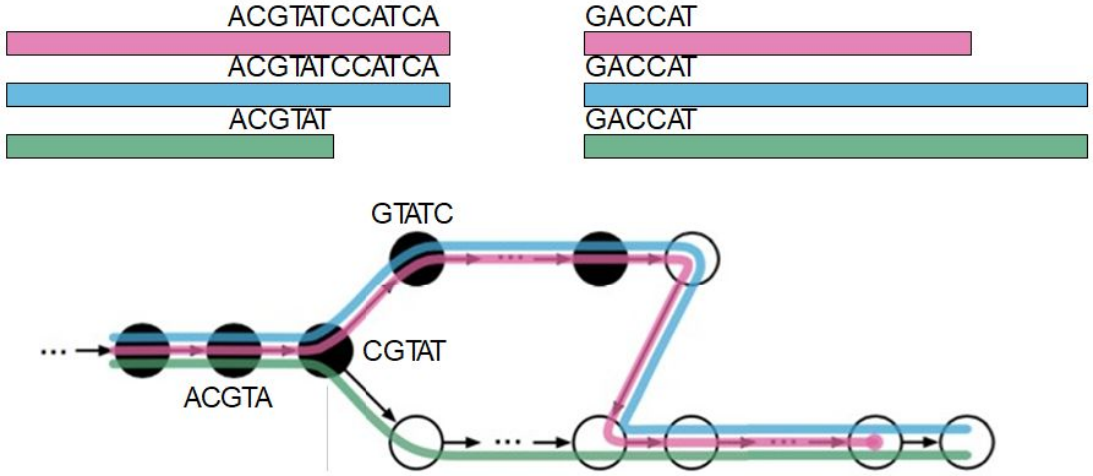
idea: let's compute that directly rather than a basepair-level alignment that has more information than we need

How kallisto computes pseudoaligments



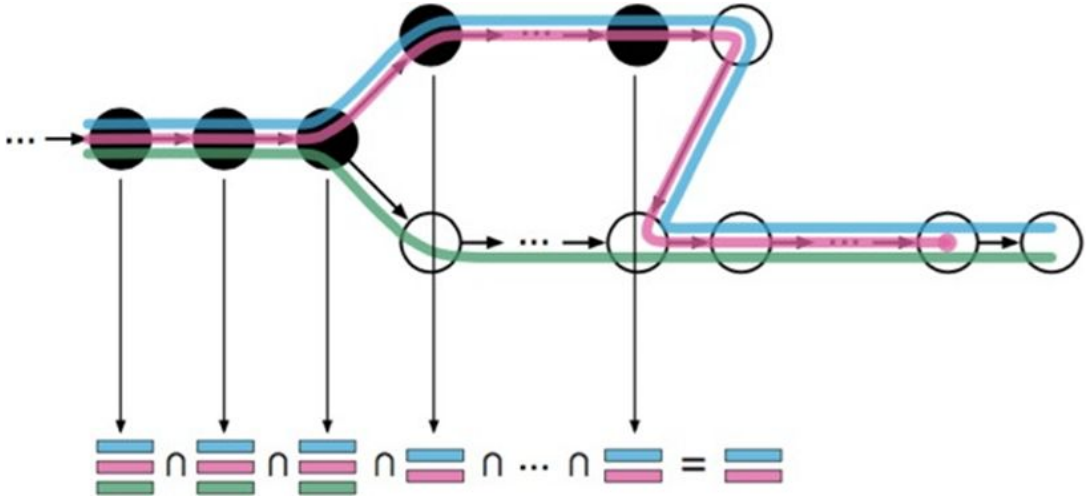
- Given our reference transcriptome,
  we first construct its *target de Bruijn Graph (T-DBG)*
- Only has to be done *once* per transcriptome (and is fast)

How kallisto computes pseudoaligments



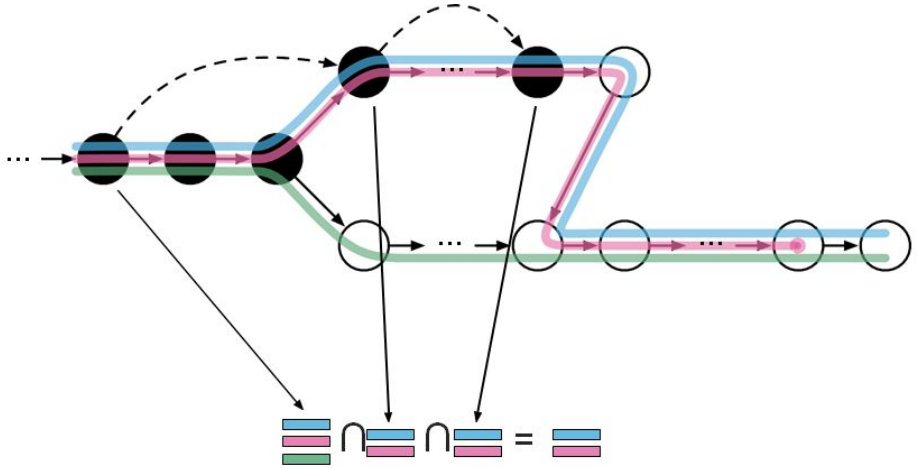
- Given a read, finding its *k*-mers in the T-DBG gives you information about where the read could have come from
- This can be done *very* fast
- But individual *k*-mers might be more ambiguous than the read as a whole

How kallisto computes pseudoaligments



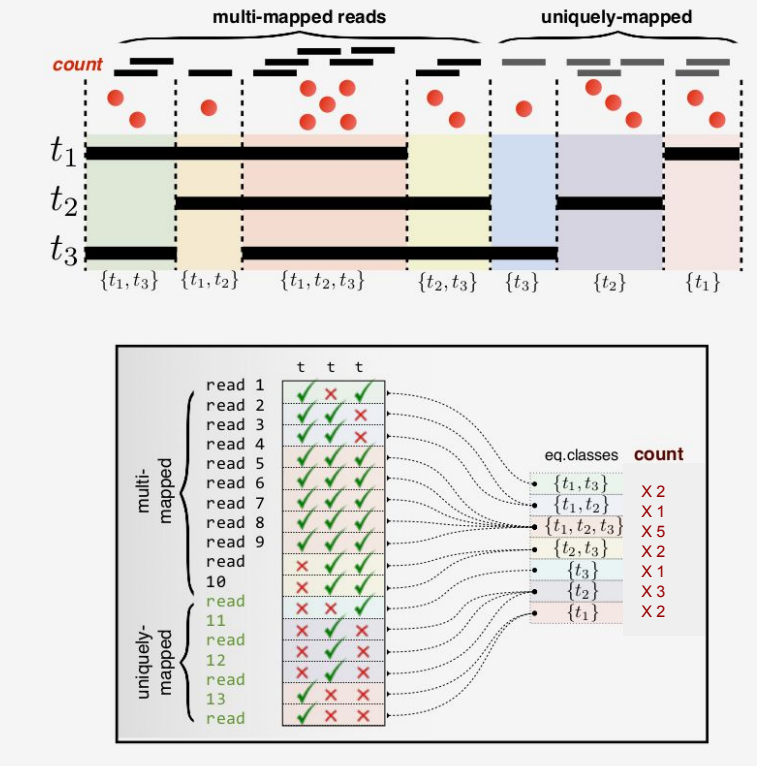
- Combining information across the k-mers can recover lost information
- For each k-mer we have the set of transcripts it could have come from. Intersecting them gives the set of transcripts that *all* k-mers could have come from
- It's possible for their combination to have information equivalent to the entire read, even if no single k-mer does by itself

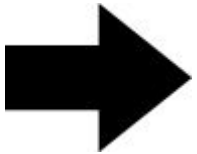
How kallisto computes pseudoaligments

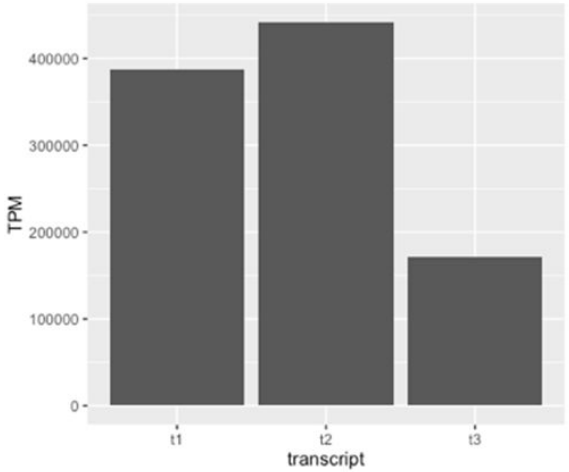


- Knowing the T-DBG, we can predict ahead of time which k-mers will be potentially interesting
- By only processing those k-mers, kallisto runs ~8 times faster

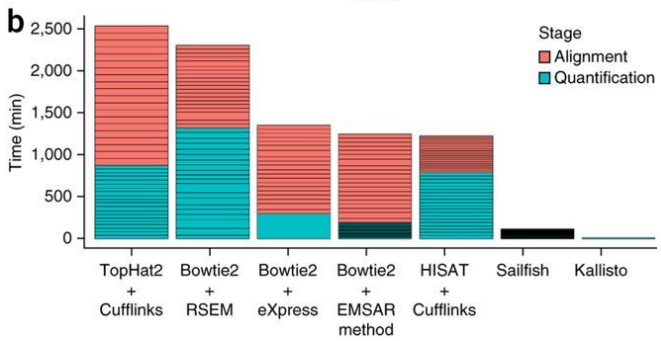
Transcript compatibility counts





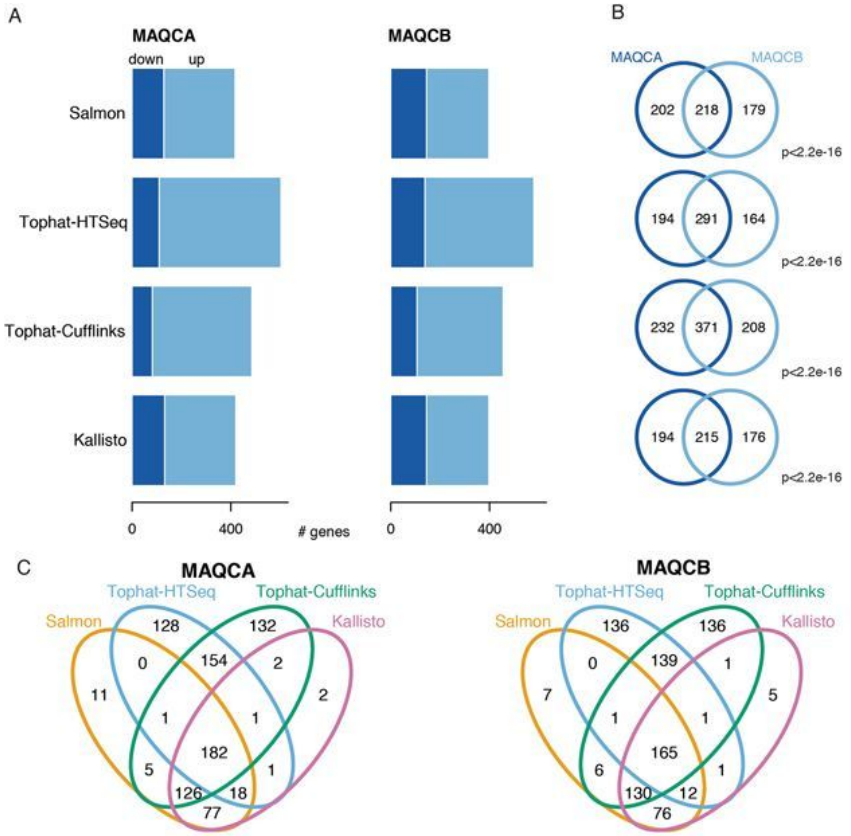


Benchmarking



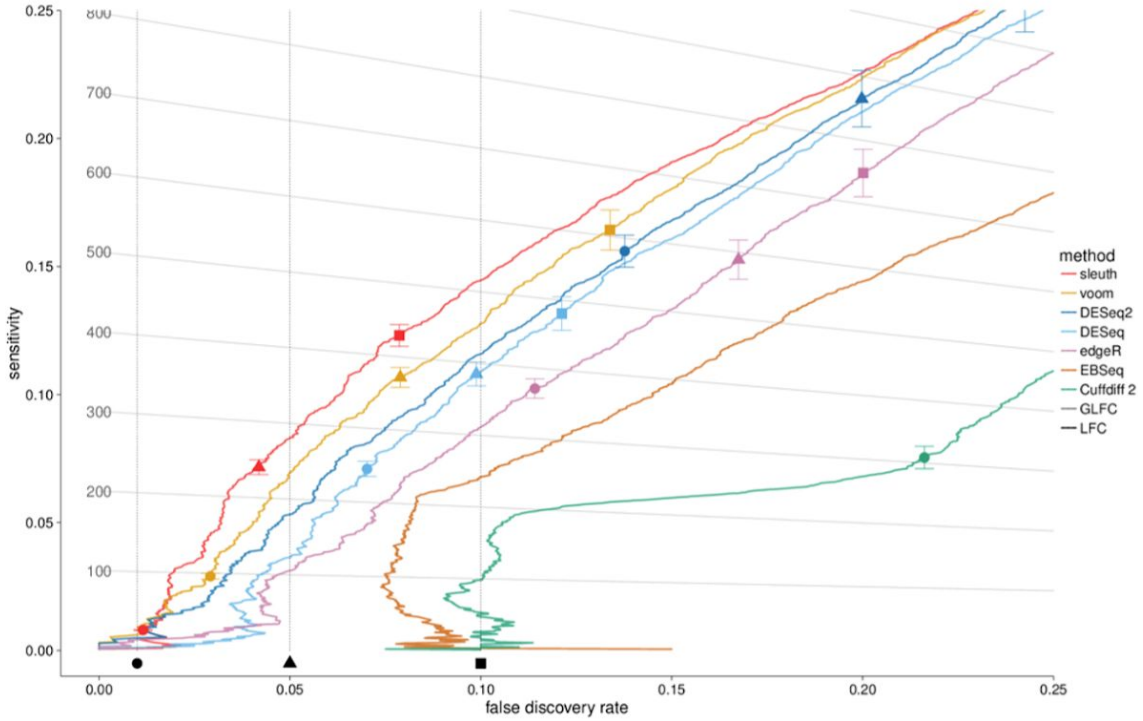
Computation Cost:

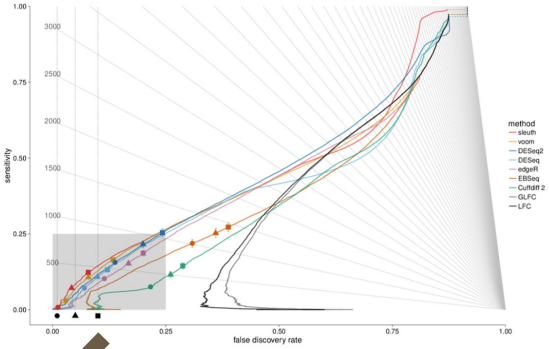
TCGA pipeline	$39/Sample
Kallisto	$0.09/Sample



Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data JO - Scientific Reports : 1559 (2017)

Benchmarking





Running Kallisto

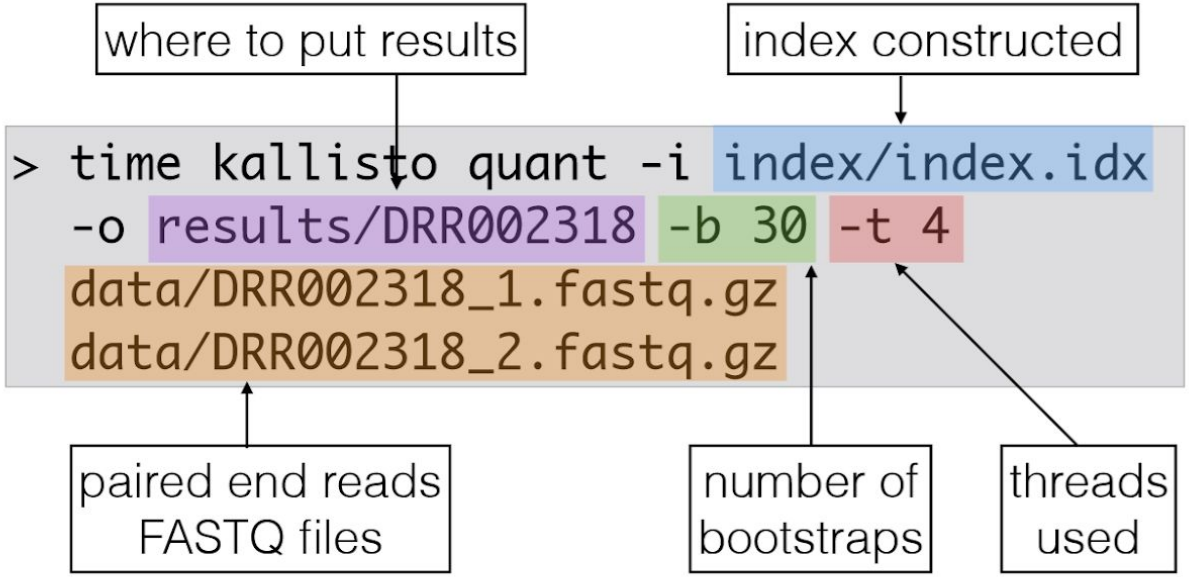
```
> time kallisto index -i index.idx
  Latimeria_chalumnae.LatCha1.cdna.all.fa.gz
```

- Requirements for running kallisto
  - Reference transcriptome
  - Paired end reads, FASTQ files

- Creates an index of the transcriptome reference that kallisto will use for quantification
  - k-mer tradeoffs
    - high k - more specific
    - low k - robust to sequencing errors

❏ for 75bp reads, use default, k=31
❏ for 50bp reads, use default except if known issues
❏ shorter reads, lower k=25 or k=21.

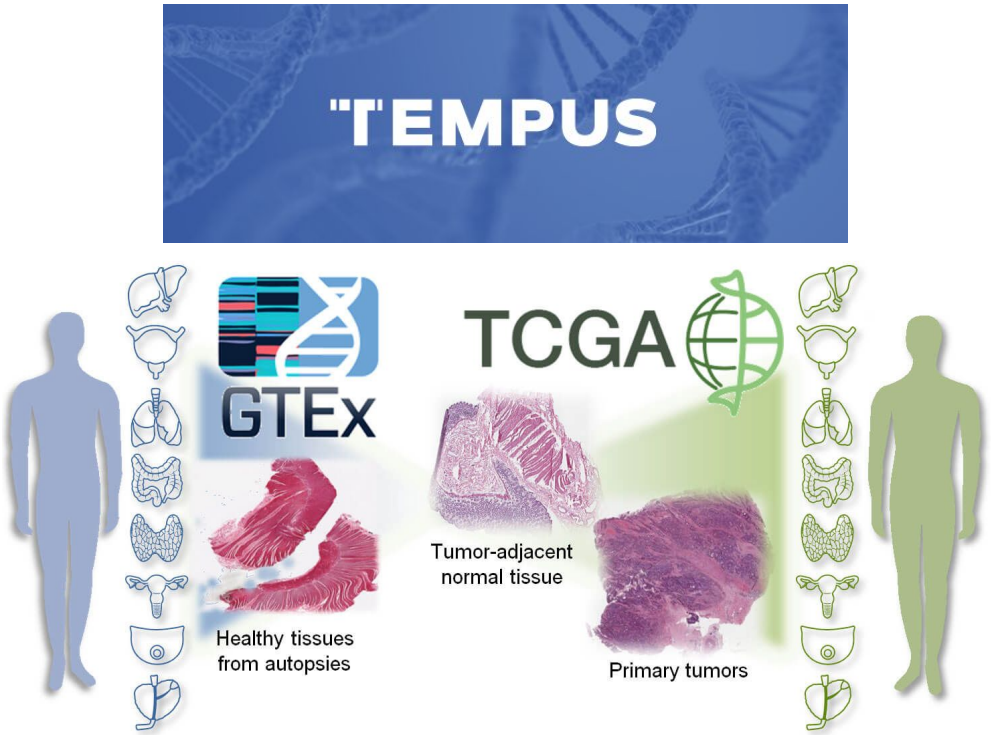
Quantification



Result

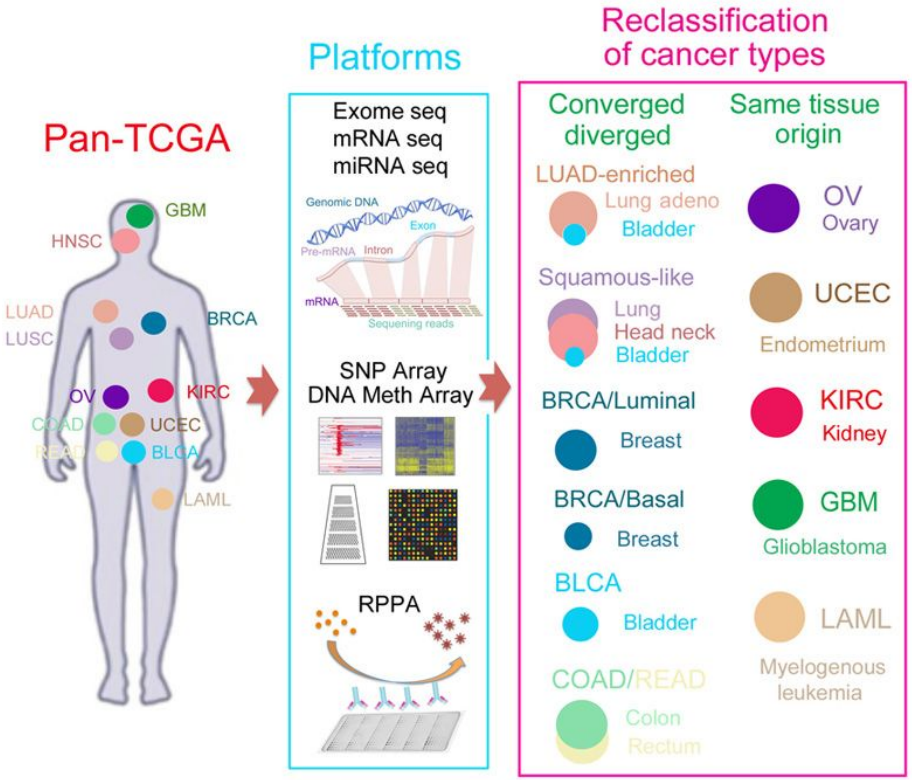
- The output is stored in the specified directory
  - **abundance.h5** - HDF5 compressed file, not human readable contains the quantifications, all bootstraps and other information
  - **abundance.tsv** - Tab Separated file abundances and counts for each transcript
  - **run_info.json** - JSON formatted file information about the run

Analyzing Multimodality of gene expression data



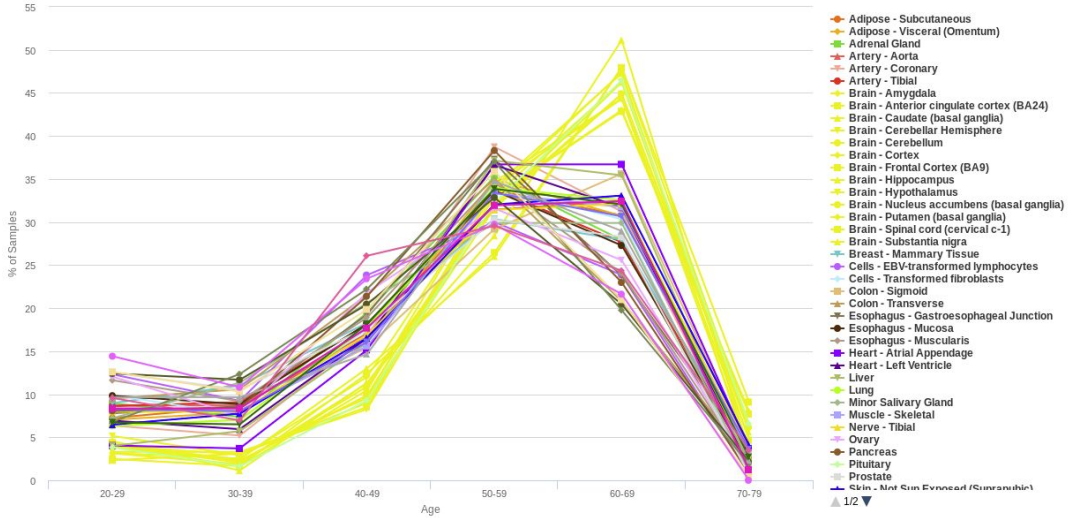
TCGA Database

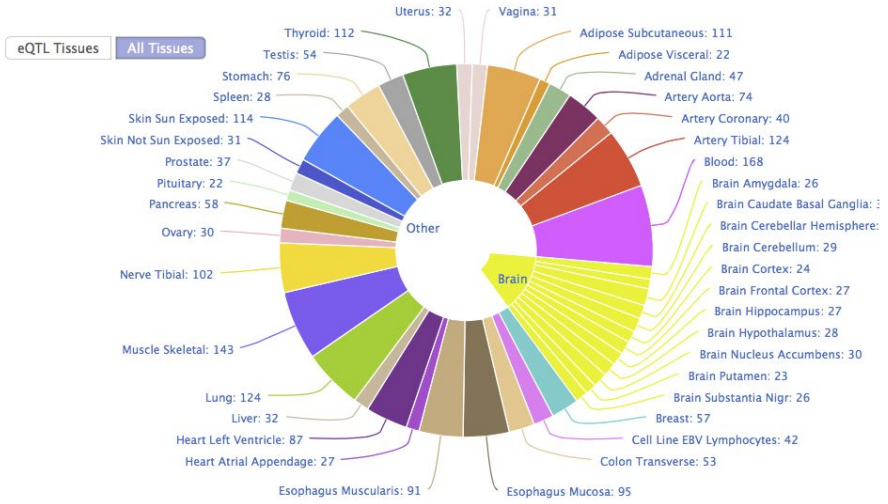
Omics Characterizations
Mutations
Copy Number
DNA Methylation
Gene Expression
MicroRNA
RPPA
Clinical Data
Imaging

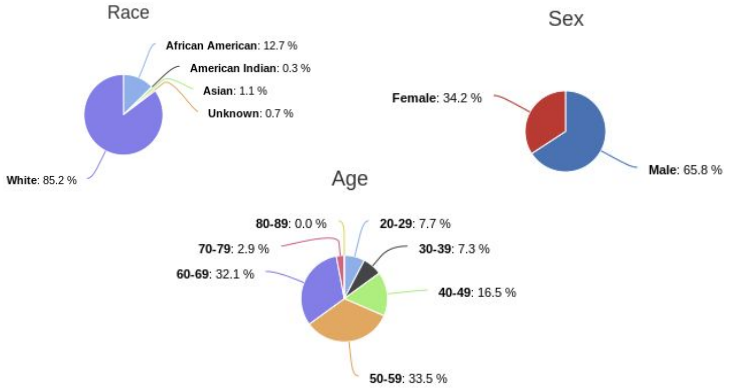


GTEx Database

11,688 Samples





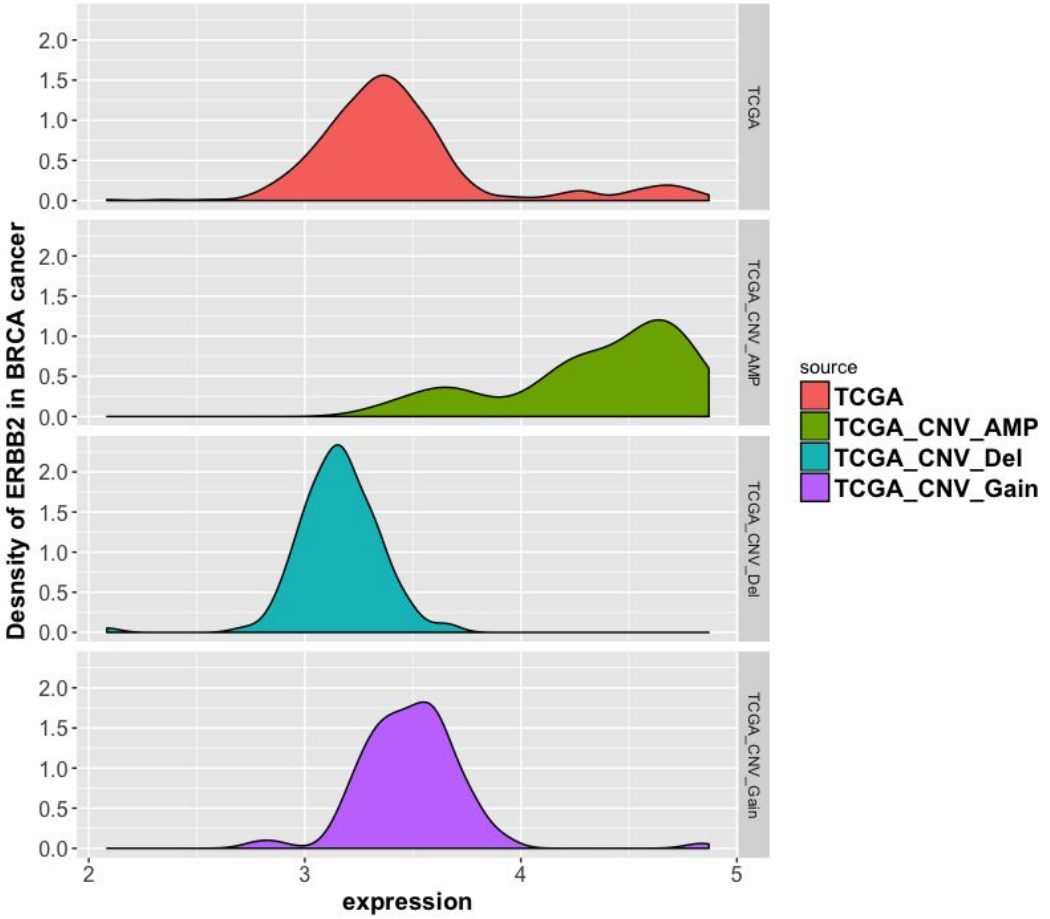


Motivation

- Understanding molecular and biological features that influence the expression of a gene
- Distinguish genes with multimodal distributions
- Explore over/under expressions calls of a subset of genes for some cancer types

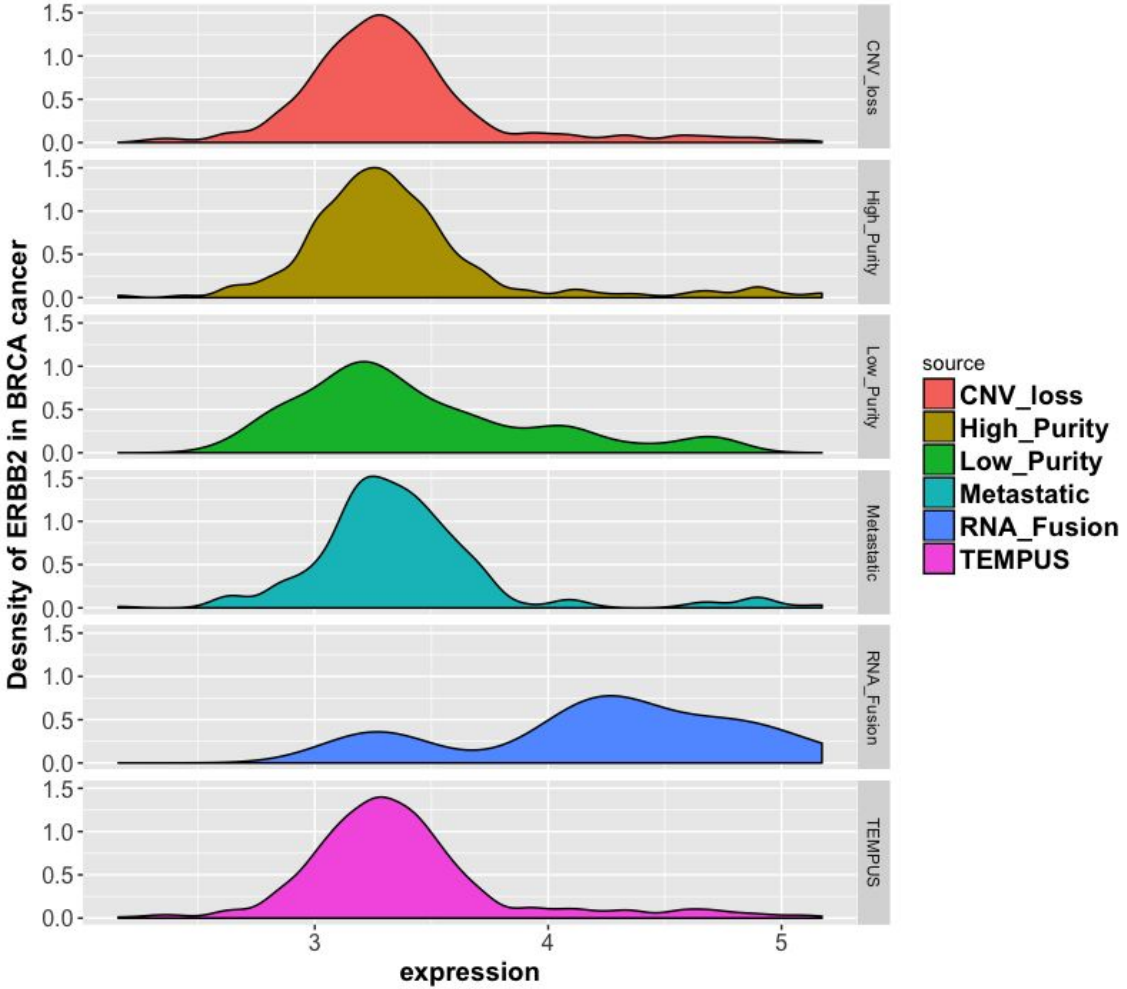
CNV

The expression of *ERBB2* gene in TCGA breast cancer has a direct relation with copy number of variation (CNV)



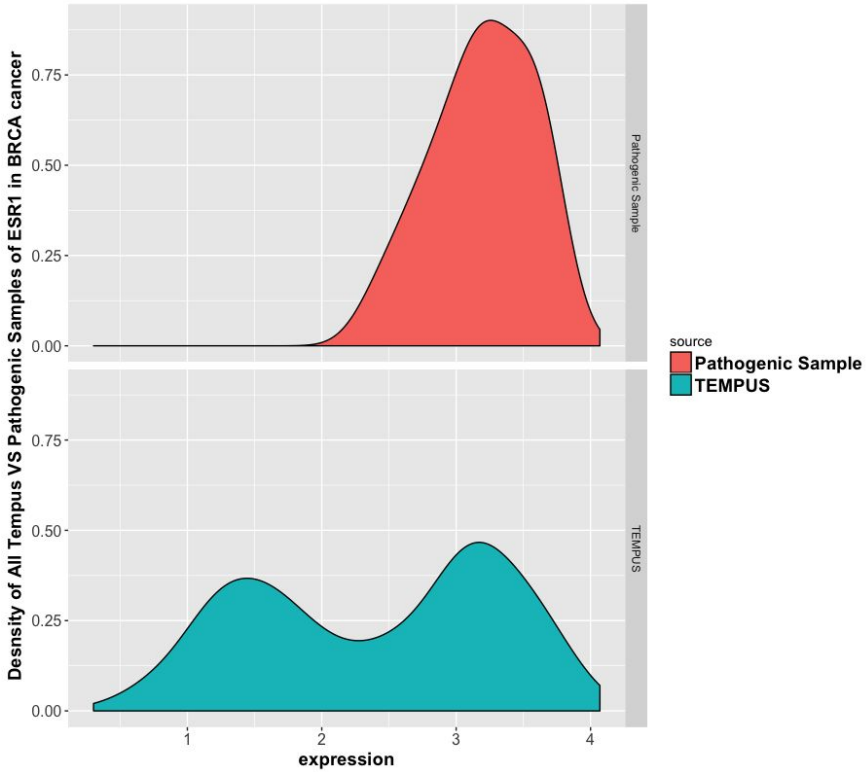
Fusion

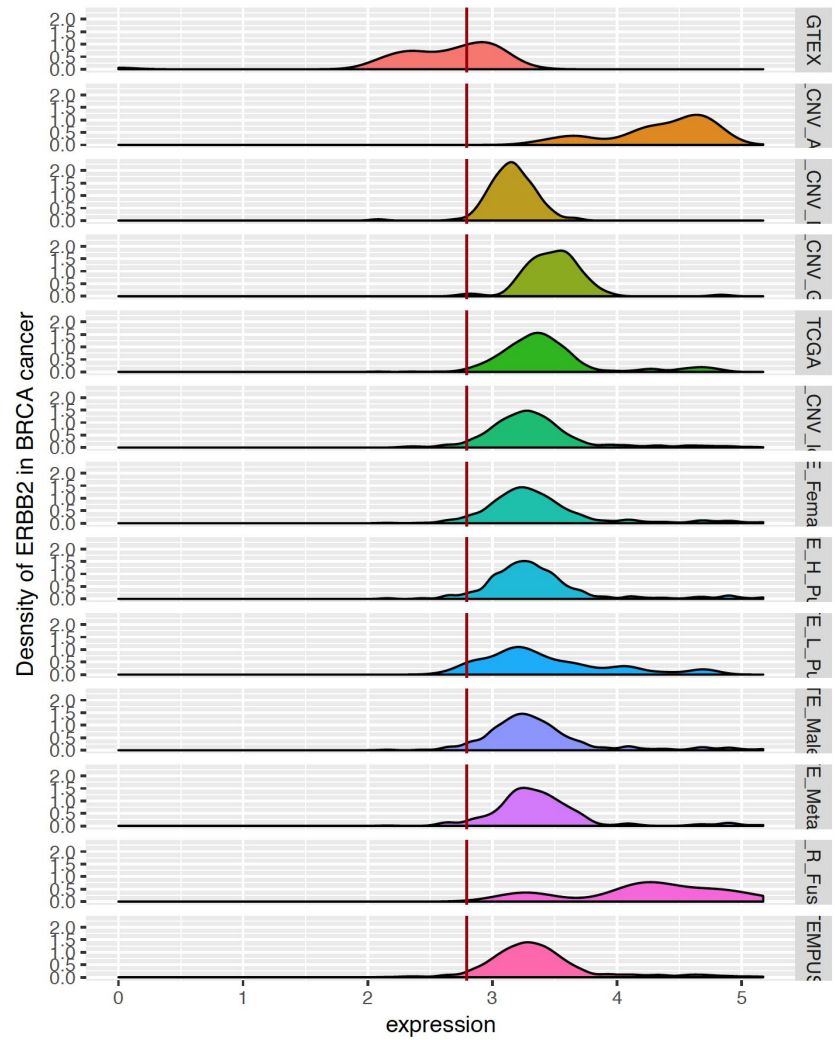
Tempus samples with RNA fusions involving *ERBB2* have higher expression of the gene in breast cancer

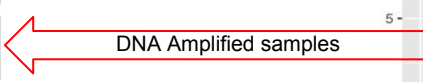


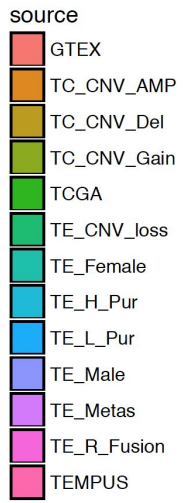
Pathogenic Variation

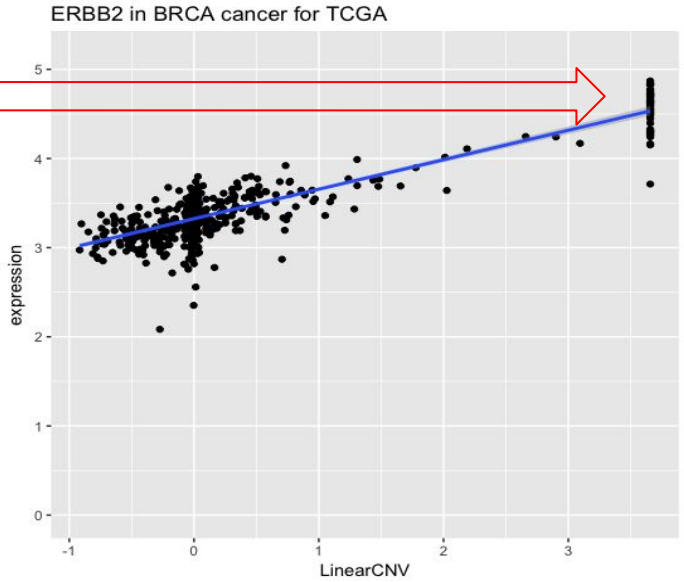
Tempus samples with Pathogenic variants in breast cancer have different expression patterns in *ESR1* gene

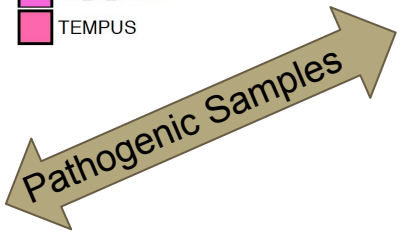


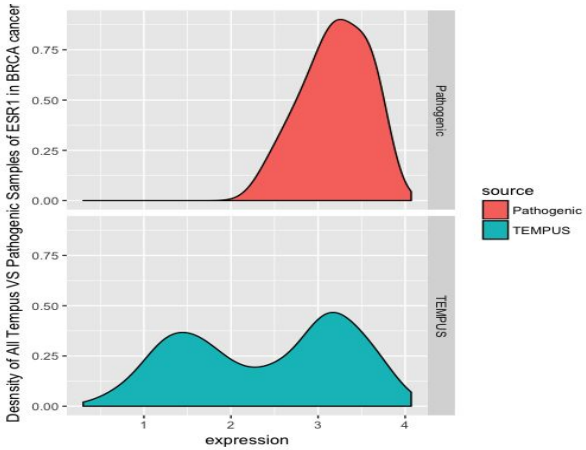

Desnsity of ERBB2 in BRCA cancer
GTEX
TC_CNV_AMP
TC_CNV_Del
TC_CNV_Gain
TCGA
TE_CNV_loss
TE_Female
TE_H_Pur
TE_L_Pur
TE_Male
TE_Metas
TE_R_Fusion
TEMPUS
expression


DNA Amplified samples


source
GTEX
TC_CNV_AMP
TC_CNV_Del
TC_CNV_Gain
TCGA
TE_CNV_loss
TE_Female
TE_H_Pur
TE_L_Pur
TE_Male
TE_Metas
TE_R_Fusion
TEMPUS


ERBB2 in BRCA cancer for TCGA
expression
LinearCNV


Pathogenic Samples


Desnsity of All Tempus VS Pathogenic Samples of ESR1 in BRCA cancer
Pathogenic
TEMPUS
expression
source
Pathogenic
TEMPUS

Challenges

Complexity of gene expression data

Samples Size

Noise Samples

Uncertainty of features

Non-linear relationship between gene expression and meta-data

Need a metric to distinguish between gene of interest

Data

13 Cancer Genes (Preliminary Analysis)

- *EGFR, MLH1, MGMT ,AR ,ERBB2 ,ESR1 ,PGR ,MAP3K1 ,MAP2K2 ,AKT1 ,AKT2 ,AKT3 ,MYC*

32 Cancer Types

25,469 Expression Samples (TCGA, GTEx, Tempus)

Meta data

- Biological: Race, Gender, Age, Vital Status
- Molecular: CNV, Tumor Purity, Metastasis, Tumor Stage, Methylation Status

Workflow

Gene Expression

Preprocessing & Building Expression Matrix

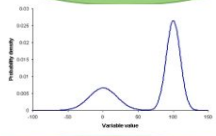
Cancer Cohorts (TCGA VS Tempus)

All Cohorts together (Cancer VS Normal)

Bimodality Test (LP-method, Hybrid method using GMM and kernel density)

if modes>1

Find the cutoff



Individual Cohort (TCGA, TEMPUS, GTEx)

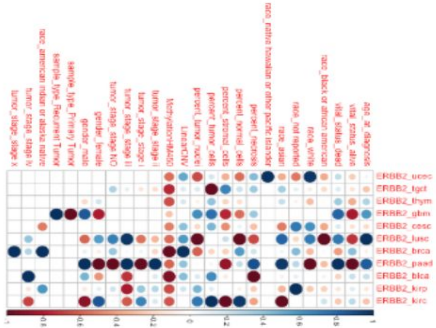
Bimodality Test

if modes>1

Preprocessing & Building Meta data Matrix

Ridge Regression (Response= gene expression)

$R^2 > 0.4$



LP-Method for mode detection

Using both Gaussian Mixture Model (GMM) and Kernel Density methods to simultaneously fit the modes

Returns actual mode(s) and 95% interval

Extremely fast

Sensitive to outliers

Mukhopadhyay, Subhadeep. "Large-scale mode identification and data-driven sciences." *Electronic Journal of Statistics* 11.1 (2017): 215-240.

Ridge Regression (On TCGA)

Response (Y):

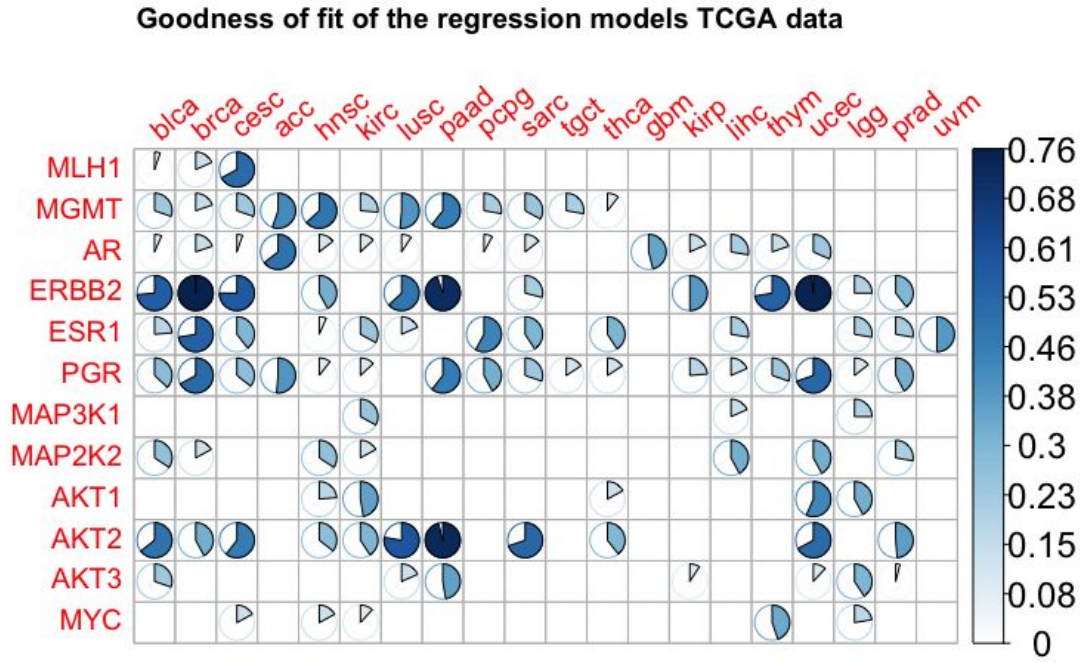
Gene expression vector

Predictors (X):

32 features extract from meta-data

Filter those models that has

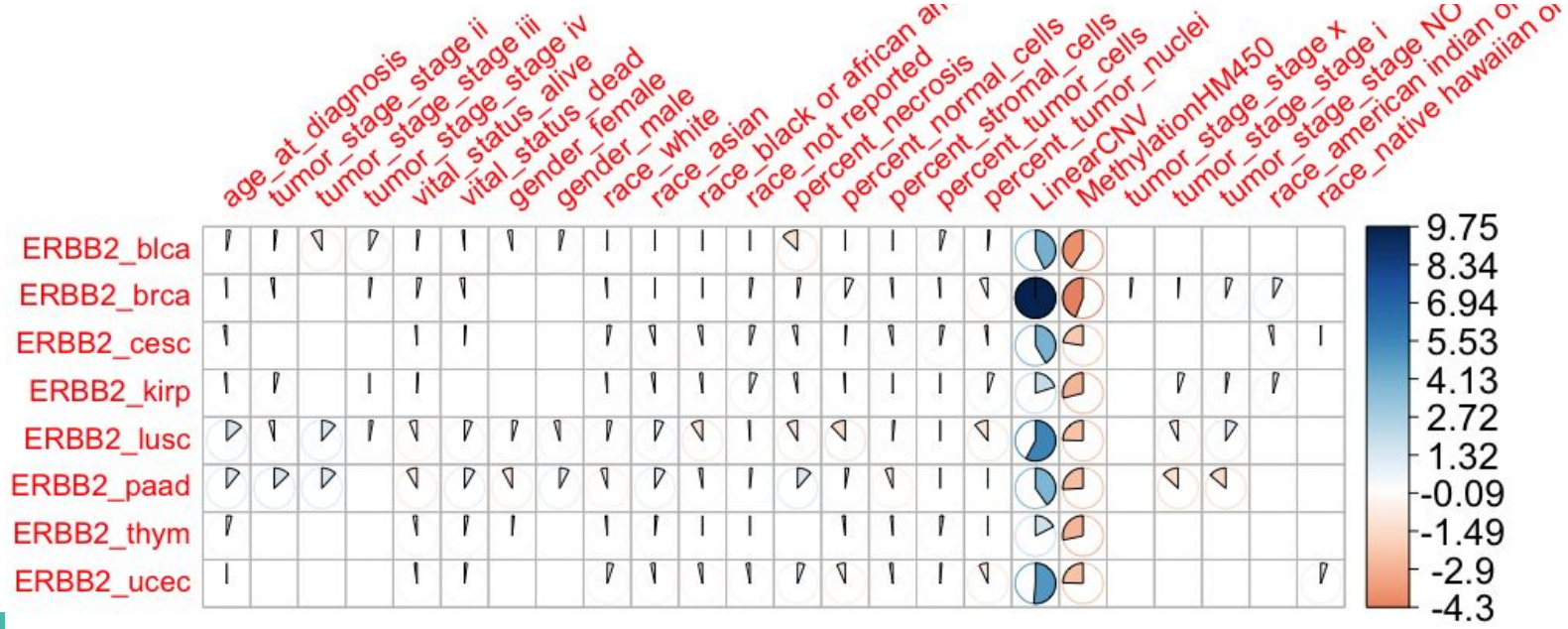
$$R^2 > 0.4$$



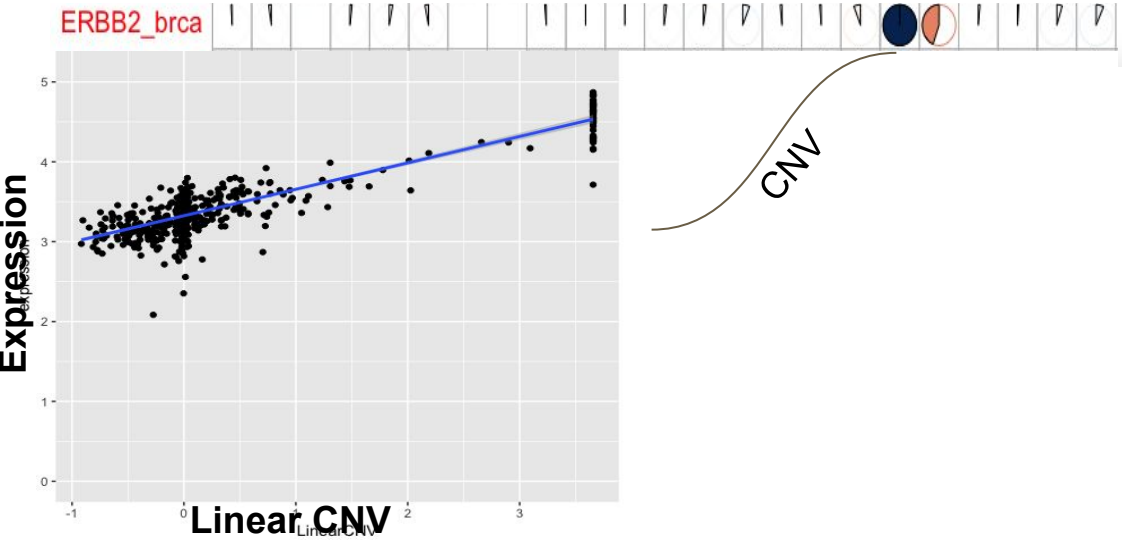
Interpret Result of The Model For *ERBB2*

Correlation plot for cancer types with $R^2 > 0.4$

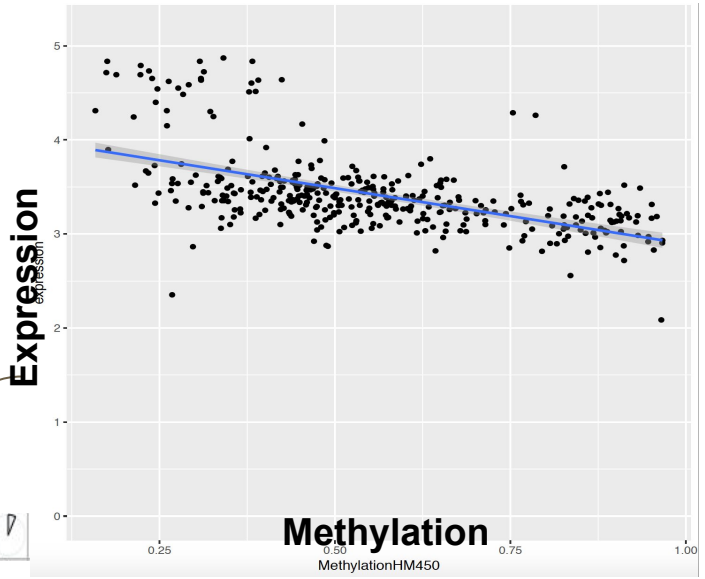
Non-zero coefficient of the regression model, represents the correlation between that features and gene expression.



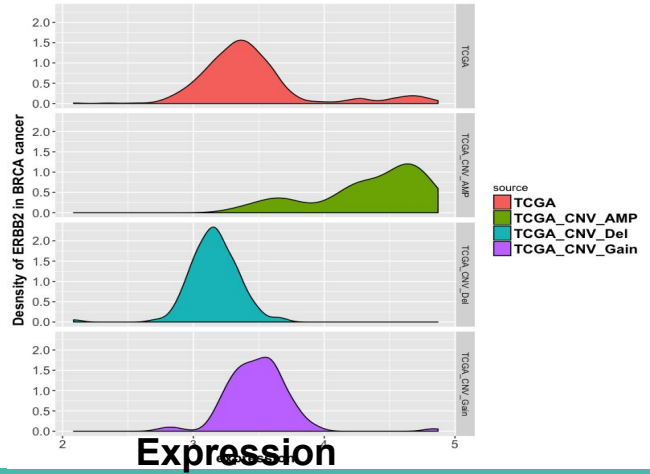
ERBB2 in Breast Cancer



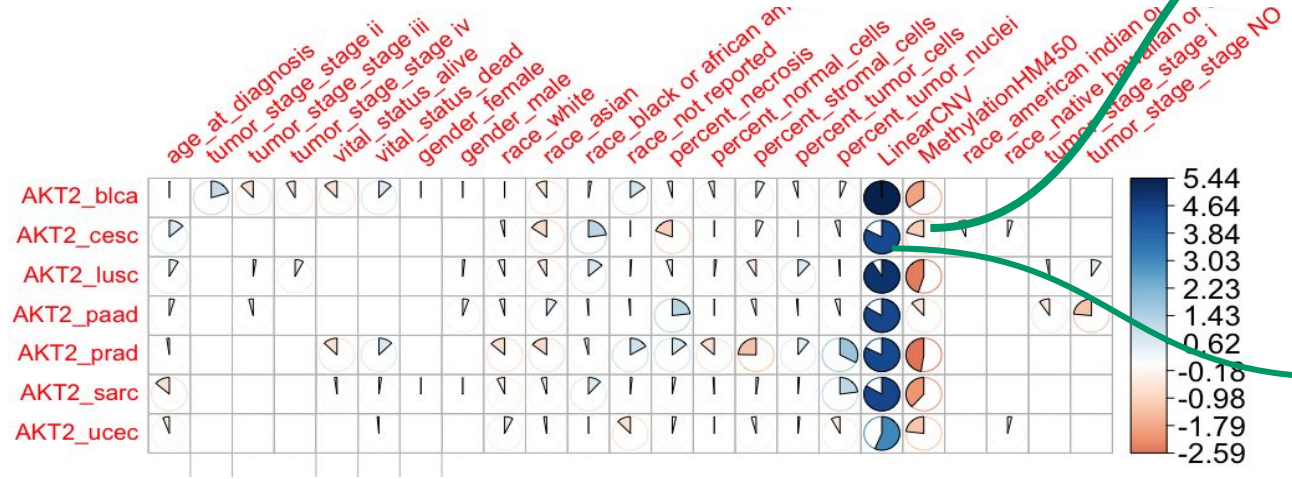
Methylation

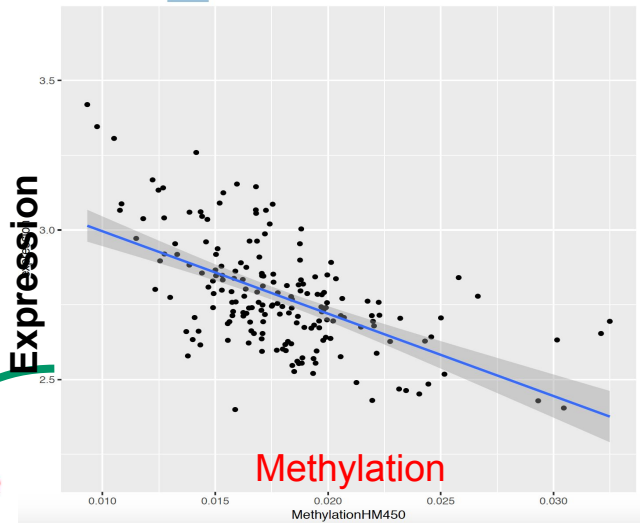


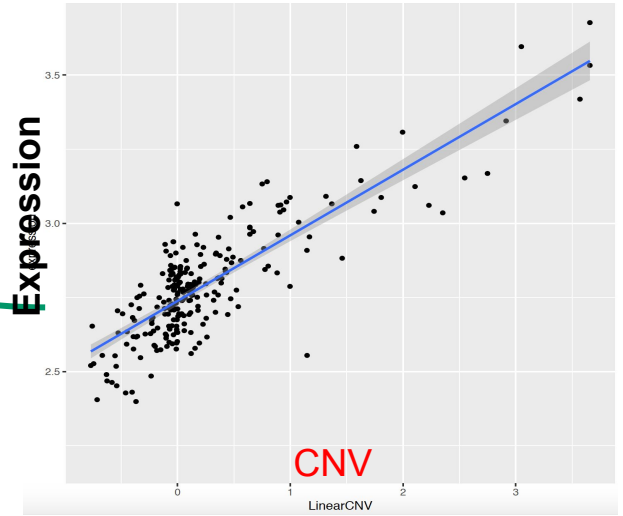
CNV



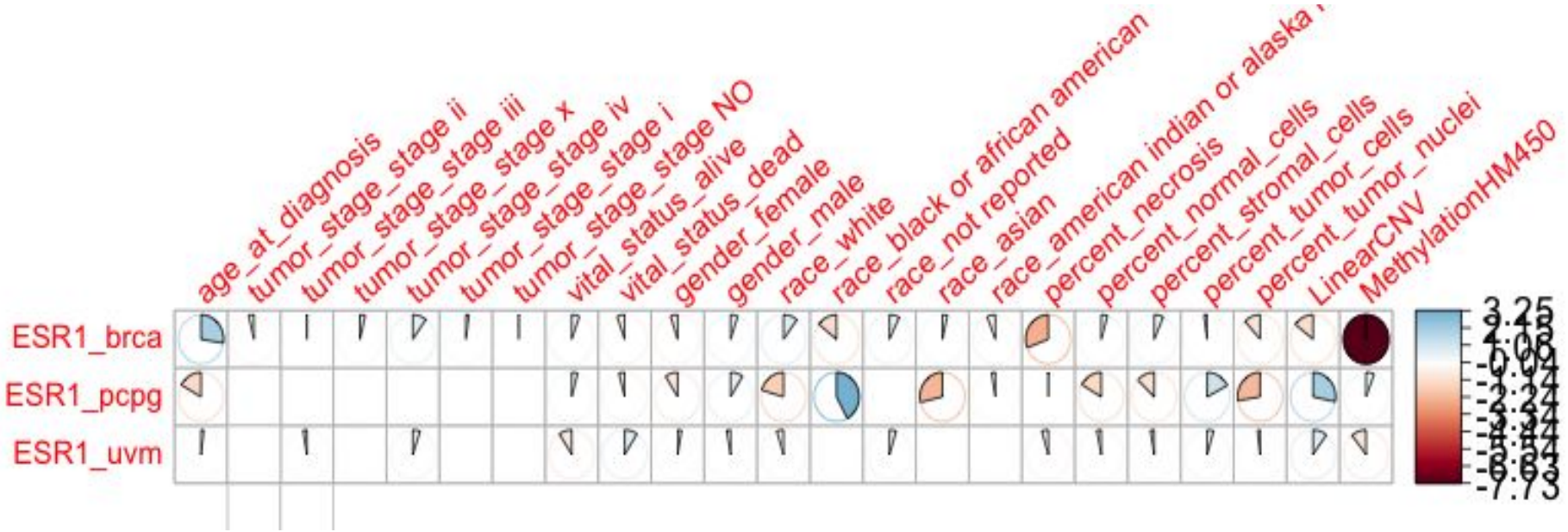
Result of *AKT2* Model







Result of *ESR1* Model



Conclusion

- Expression data is complex
- An expression of a gene from a population can have a high or low variation which can be caused by different meta-data
- This pipeline performs better for those genes that have a high variation
- Multivariable models may help detect biological and molecular effects that result in bimodal gene expression distributions
- These models can help improve over/under expression calls

Improving Result

How can we add pathway data into a model?

How about clinical data?

Performing more statistical analysis for each gene across different cancer types and try to interpret changes of the correlation and cluster them into meaningful biological groups

Is there any relationship between coefficient of **oncogenes**, **tumor suppressor, housekeeping, Chromatin Accessibility** , etc genes?