RNA-seq Data Mining

Emerging Approaches For Tumor Analyses in Epidemiological Studies

NIH NATIONAL CANCER INSTITUTE

May 3, 2023
9:30am - 12:00pm

Session overview

- Normalization and differential expression
- Isoform discovery and alternative splicing
- Pathway analysis
- Clustering and classification
- Fusion gene detection
- Allele-specific expression
- RNA editing

Normalization and differential expression

Stephen Hartley

Staff Scientist

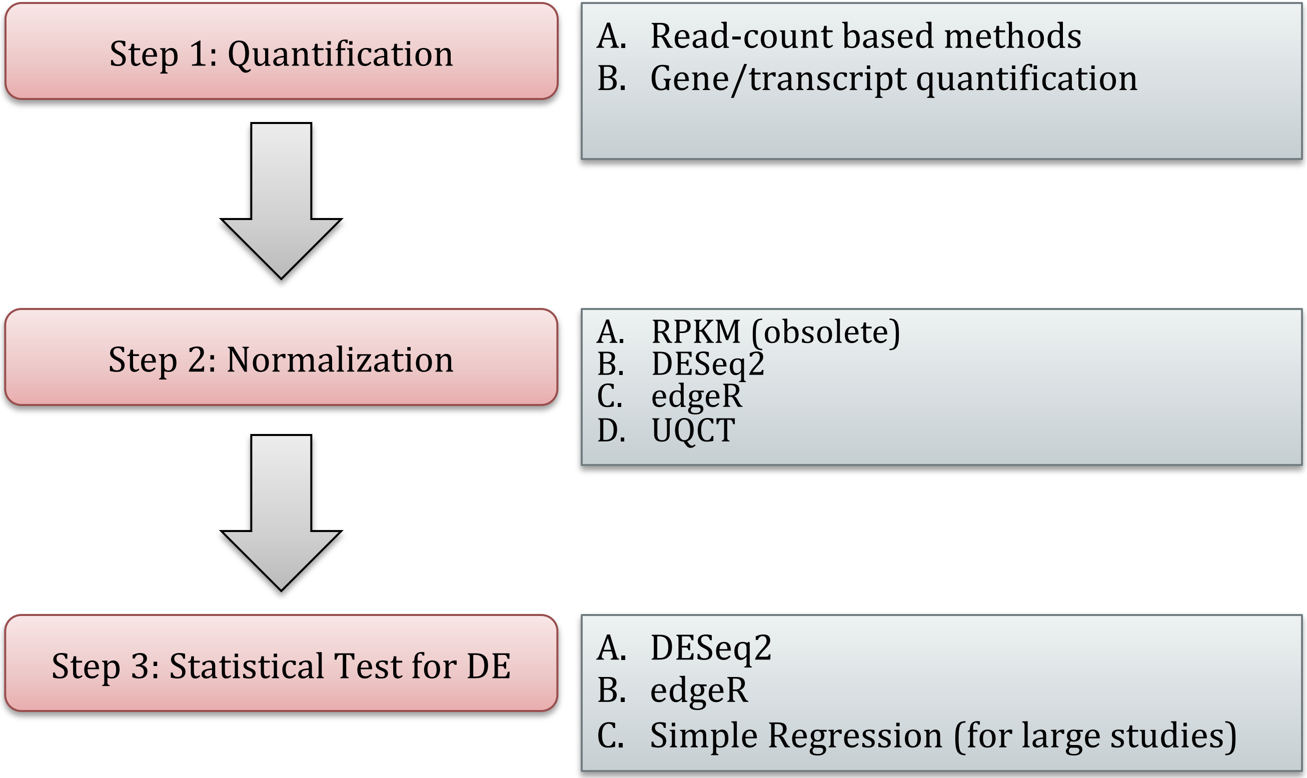
DCEG/LGS

Differential Expression Analysis

Study Question:

Are there genes that are differentially up/down regulated with respect to our variable of interest (exposure, status, etc.)

Differential Expression Analysis: Overview



Step 1: Quantification

A. Read-count based methods
B. Gene/transcript quantification

Step 2: Normalization

A. RPKM (obsolete)
B. DESeq2
C. edgeR
D. UQCT

Step 3: Statistical Test for DE

A. DESeq2
B. edgeR
C. Simple Regression (for large studies)

Step 1: Quantification

To determine if expression is different, first we must quantify expression levels:

The two main ways to quantify expression levels are:

1. Read Counts Based Methods
   - Directly count # reads covering a feature (gene, transcript, etc)
   - Must be normalized as a separate step

2. Gene/Transcript Quantification Based Methods
   - Use more complicated methods to estimate the quantity/proportions of different genes or transcripts
   - Most tools include normalization as part of the quantification step

Step 2: Normalization

Read counts must be normalized so that they are comparable to one another

- Originally, gene-level read counts would be normalized to transcript length (in kb), and to total # reads for the sample (in millions)
  - “RPKM” – Reads per Kilobase per Million
  - “RPM” – Reads per Million
  - “FPKM” – “Fragments” (aka read-pairs) per Kb per Million

NOTE: Do not do this!
Better normalization methods have been developed. Do NOT just divide by total count.

- Because RPKM became the standard, many tools use proper normalization methods but then adjust the numbers into “RPKM” or “RPM”-“like” units.
  - Sometimes estimate transcript abundances as “TPM” (transcripts per million)

Step 2: Normalization

Read counts must be ***NORMALIZED*** so that they are comparable to one another!

There are several tools that will perform this normalization properly:

- edgeR (“TMM” normalization)
- DESeq2 (“RLE” normalization)

The difference is usually minimal.

(Note: these tools also perform differential expression analysis, but you can also just use the normalization part.)

Gene/Transcript Quantification Based Methods

Use more complicated methods to estimate the quantity/proportions of different genes or transcripts

Popular options:

- CuffLinks
- Kallisto

These tools generally output normalized expression levels for each gene and/or transcript.

Step 3: Statistical Tests for Differentials

Study Question:

Are there genes that are differentially up/down regulated with respect to our variable of interest (exposure, status, etc.)

How do we test for this?

Depends on your study:

A. Traditional RNA-Seq study designs:
   - Very small (~3 replicates case/control)

B. Large scale RNA-Seq studies:
   - Much larger (10+ replicates per group)

Step 3: Statistical Tests for Differentials

A. Traditional RNA-Seq study designs:
   - Very small (~3 replicates case/control)

Require specialized methods:

- estimate dispersion with only 3 samples per group (share information between genes)
- Unusual probability distribution (negative binomial).

For count-based methods:

- edgeR and DESeq2 are the most common options

For transcript-quantification-based methods:

- CuffDiff (for use with Cufflinks)
- Sleuth (for use with Kallisto)

Step 3: Statistical Tests for Differentials

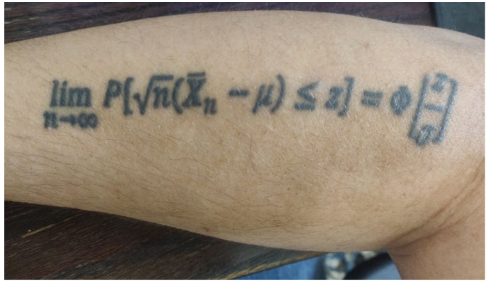
B. Large scale RNA-Seq studies:

- Much larger (10+ replicates per group)

Specialized methods *not necessary*, can just use linear regression on:

- log-scaled normalized counts (count-based methods)
- TPM estimates (transcript quantification methods)

Central Limit theorem: negative binomial distribution is irrelevant with large sample size


$\lim_{n\to\infty} P[\sqrt{n}(\bar{X}_n - \mu) \leq z] = \Phi\left[\frac{z}{\sigma}\right]$

Step 3: Statistical Tests for Differentials

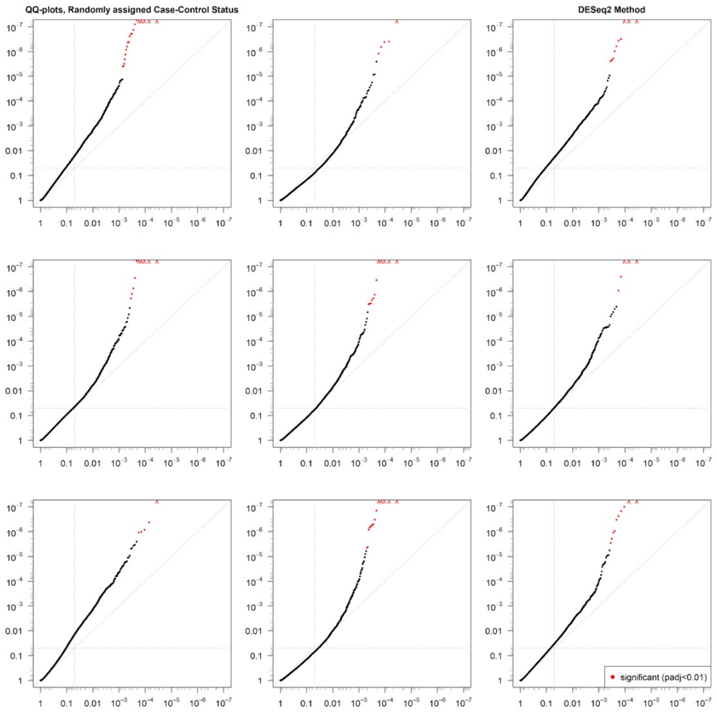
Also note: in large sample sizes, specialized methods appear to overfit / inflate p-values

- In a recent project, I took a dataset and generated 9 completely random case/control variables
- Ran DESeq2 and simple linear regression (on log-normalized counts)
- DESeq2 showed statistically significant results

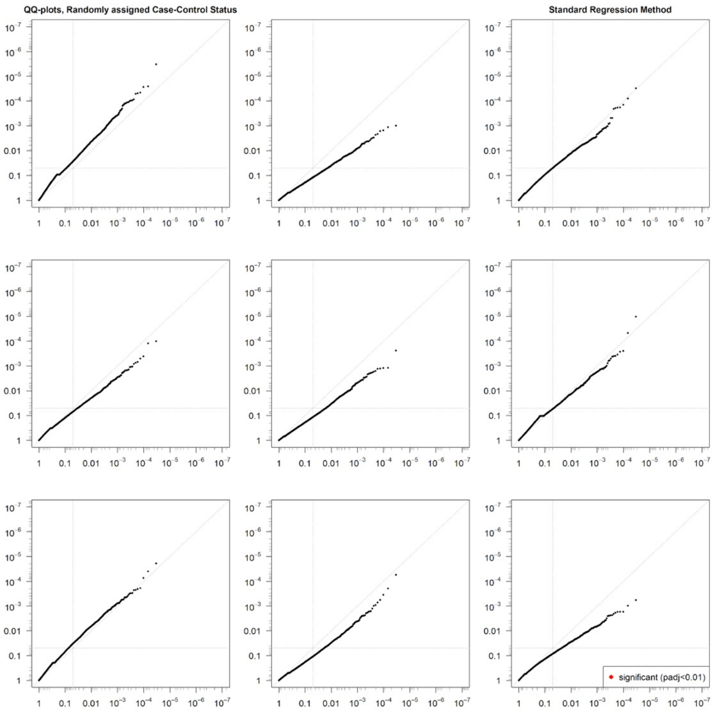
Step 3: Statistical Tests for Differentials

Also note: in large sample sizes, specialized methods appear to overfit / inflate p-values

DESeq2



Simple Regression



Step 3: Statistical Tests for Differentials

Remember to control for confounding variables (if applicable).

Most methods give the ability to add confounders, effect modifiers, batch, etc. Usually in a standard regression form such as:

$$Y \sim batch + age + sex + caseCtrlStatus$$

Step 4: Interpretation

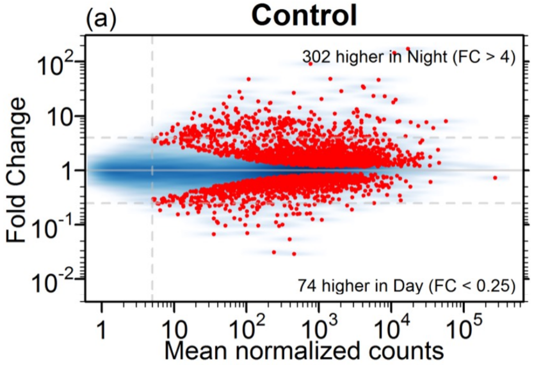
Regardless of the method, the result traditionally includes:

- Results table
  - List of genes (or transcripts) with fold change and p-values
- “MA” plot.

  x-axis: mean normalized counts (# read pairs per gene)

  y-axis: fold change

Real-life example: day/night fold change, TTC8 gene, rat pineal glands



Differential Expression Analysis
Main takeaways

- DE analysis looks for differences in expression based on some study condition/exposure
- Three steps:
  - Quantification
  - Normalization
  - Statistics
- Two primary quantification methods:
  - Count-Based
  - Transcript-Based
- Statistical Methods depend on study size
  - Small studies: Use standard DE/RNA-Seq tools
  - Large studies: Use simple regressions on log-transformed/normalized data

Differential Splicing and Alternative Transcript Usage Analysis

Stephen Hartley

Staff Scientist

DCEG/LGS

Differential Splicing
Differential Isoform usage
Differential Exon Usage
Alternative Isoform Regulation
etc…
Analysis

Study Question:

Is there any kind of differential expression regulation occurring among a gene’s transcript set, distinct from gene-level differential expression?

2 basic types of differential expression regulation:

1) Differential Gene Expression (DGE/DEG/DE) ✓
   - Entire gene is up/down regulated, depending on some experimental condition
   - Fairly straightforward: Use edgeR, DESeq2, or limma.

2) Alternative Isoform Regulation (AIR) ?
   - One or more ***specific*** transcripts/splice-variants/exons/etc are being ***independently*** up/down regulated based on some experimental condition
   - *Much trickier!*

Alternative Isoform Regulation can take many different forms:

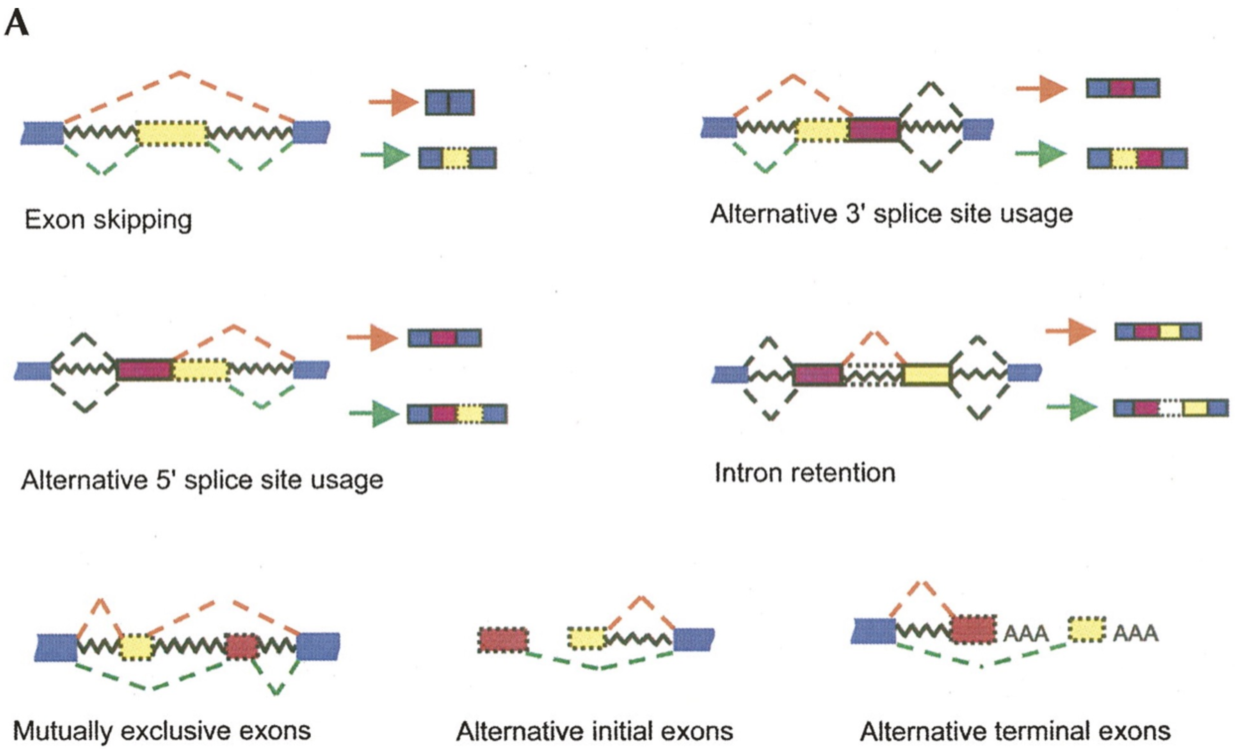
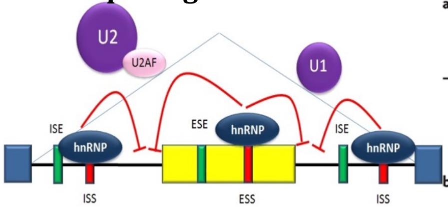

A
Exon skipping
Alternative 3' splice site usage
Alternative 5' splice site usage
Intron retention
Mutually exclusive exons
Alternative initial exons
Alternative terminal exons
AAA
AAA

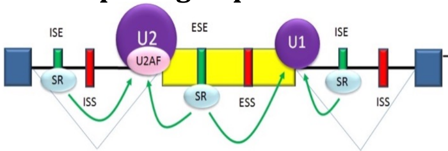
Image: Wang, Z., & Burge, C. B. (2008). *Rna*, *14*(5), 802-813.

… and can be caused by many different mechanisms

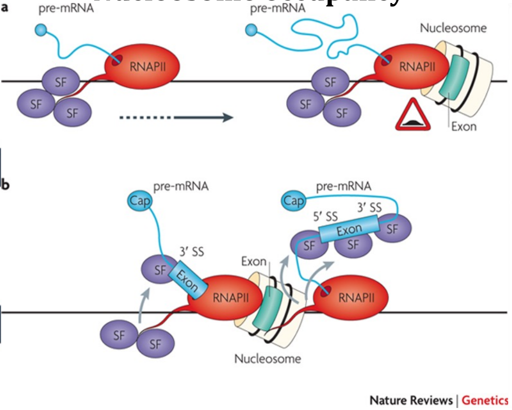
Splicing Activation[1]



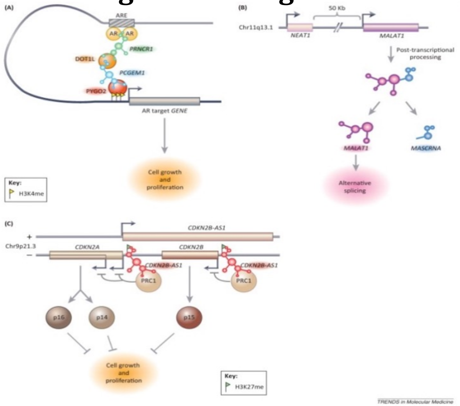
Splicing Repression[1]



Nucleosome occupancy[2]



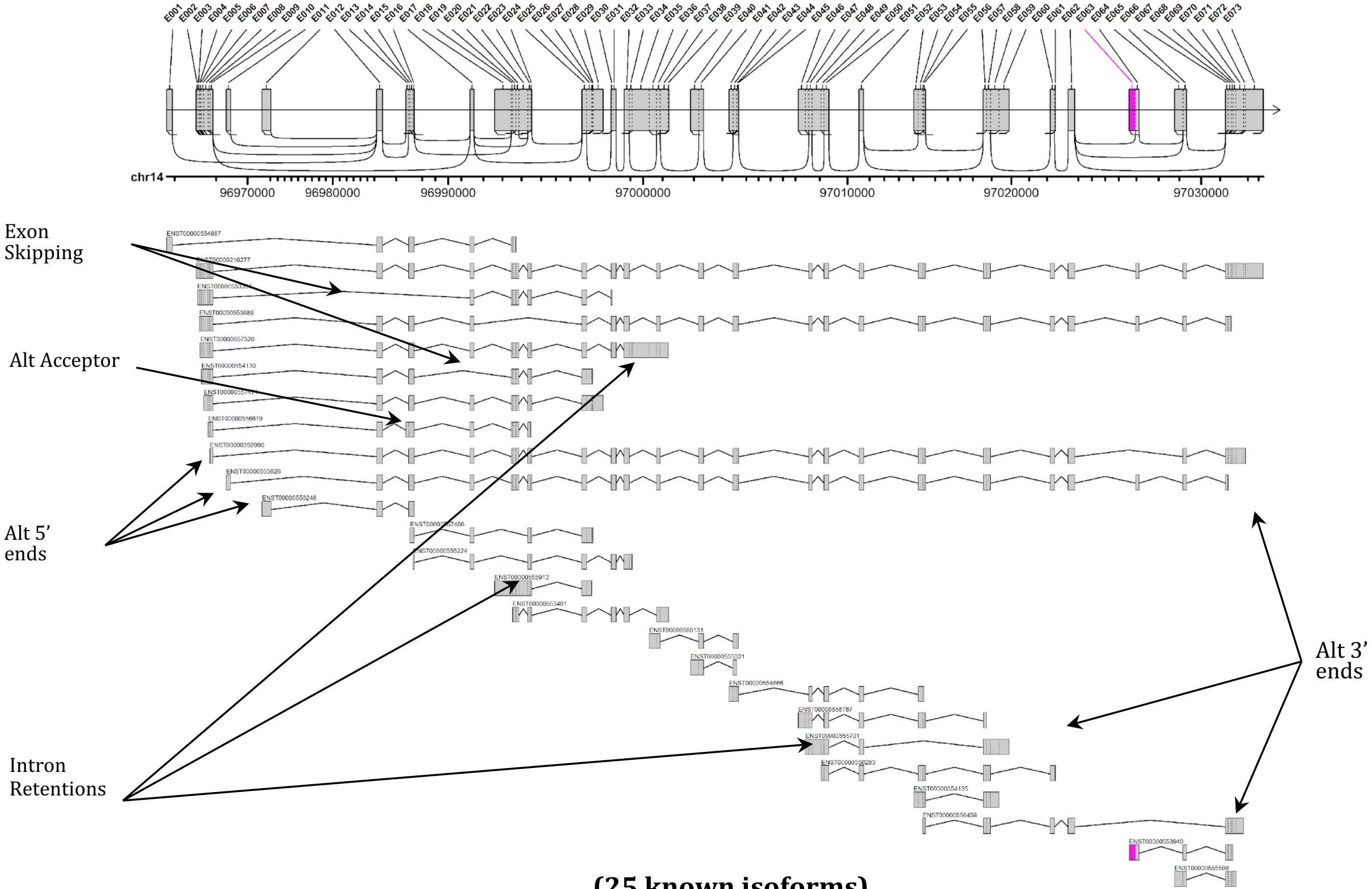
Long Non Coding RNA's[3]



[1] Wang, Zefeng, and Christopher B. Burge. *Rna* 14.5 (2008): 802-813.

[2] Keren, H., Lev-Maor, G., & Ast, G. (2010). *Nature Reviews Genetics*, *11*(5), 345-355.

The result:
Staggering complexity



Exon Skipping

Alt Acceptor

Alt 5' ends

Intron Retentions

Alt 3' ends

(25 known isoforms)

Analysis of Isoform-Level Differences is ***REALLY HARD***

Making it worse:

- Transcripts are >2kb
- Reads are <150bp
- Read coverage is not uniform
- Annotation is incomplete
  - Usually *VERY* incomplete
- Interpretation is often *very* difficult

The Interpretation Problem:

- Even if you *detect* AIR/DS/etc, *interpretation* is ***ALSO HARD***.
    - Results often complex & counterintuitive
    - Dozens of isoforms, each (may be) regulated differently
    - Annotation might be wrong/incomplete

- Bottom line: it’s not like DE analysis
    - You can’t just report fold-change & p-value and call it a day.

There are *many* tools for detecting differential splicing:

- Count-based:
    - Detects differential splicing/etc by proxy, uses counts of individual components (exons, splice sites, etc)
    - Examples: SUPPA2, rMATS, DEXSeq, edgeR, JunctionSeq (my tool)

- Isoform-based:
    - Detect AIR directly by estimating isoform abundances.
    - Most tools do not test directly for splicing changes
        - They allow you to test each isoform independently
        - May indirectly discover differential splicing if you observe some isoforms changing and not others
    - Examples: CuffDiff2, Kallisto, RSEM.

(Note: there are ***MANY*** more tools, but these at least have seen some real-life usage)

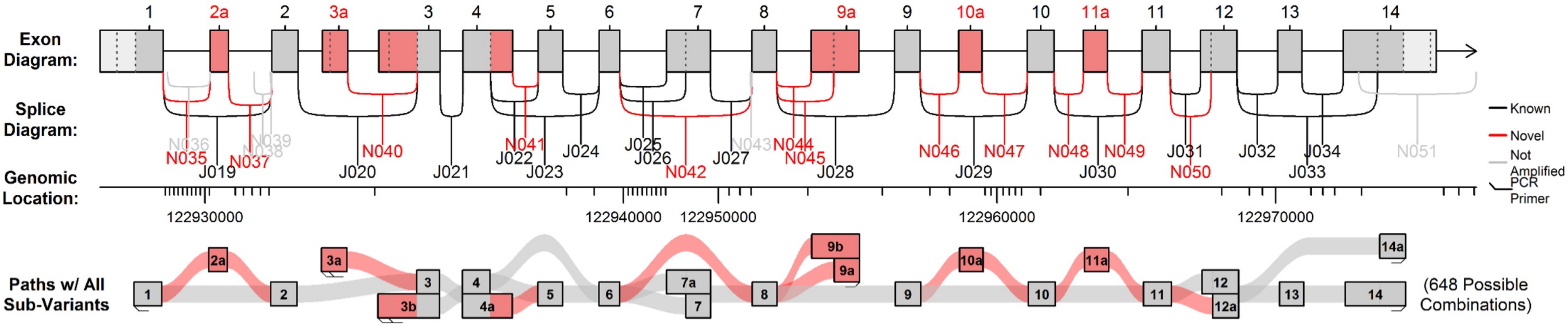
Problems with existing tools:

Problems with existing tools:

- Isoform-level tools *just don’t work very well*
  - Fundamentally hard/impossible to quantify overlapping 2kb isoforms using <150bp fragments

- Many tools have poor performance when annotation incomplete
  - Many count-based methods perform very poorly when affected transcripts are not annotated (for obvious reasons)
  - Annotation is ***<u>always</u>*** incomplete

Problem: Illumina data does not give us full-length Isoforms



648 paths!
(potential isoforms)

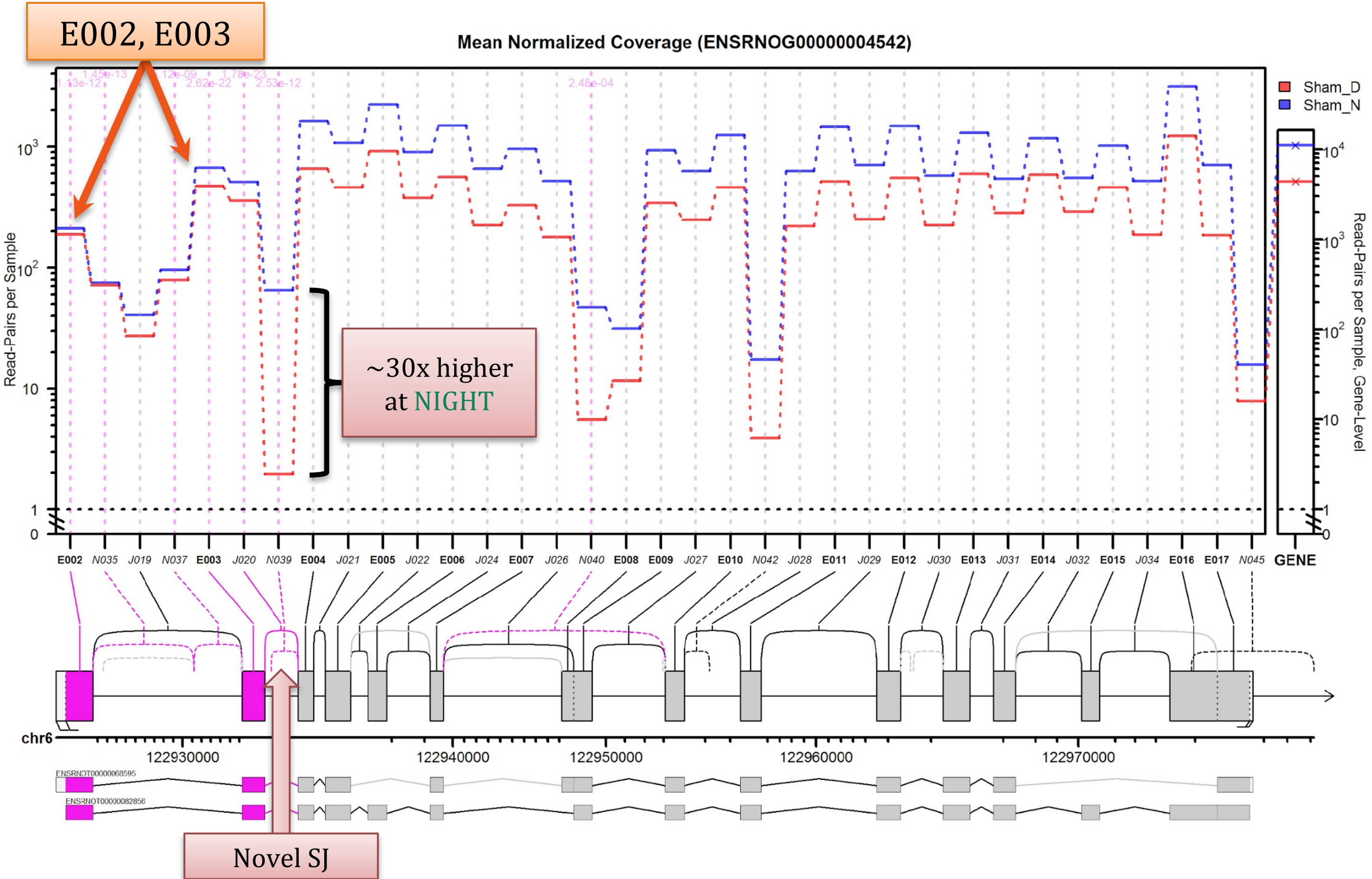
Differential Splicing/Differential Isoform usage/Differential Exon Usage/Alternative Isoform Regulation/etc… Analysis

INTERPRETATION and VALIDATION

- Results should not be trusted without substantial validation
  - With the right primers, qPCR can validate splicing differentials
  - Long-read technologies can validate the existence of splicing variants.
    - Confirm that they are coding, in-frame, full-length, etc.

- Interpretation should be in-depth:
  - Generate “wiggle” plots for UCSC browser (or IGV, etc). Examine expression levels across the gene
  - Examine gene closely: check for novel exons, novel splice junctions, genes on the opposite strand (if RNA-Seq is unstranded), overlapping genes.
    - These may cause false positives and/or misleading results

Applied example: the TTC8 Gene in rat pineal glands

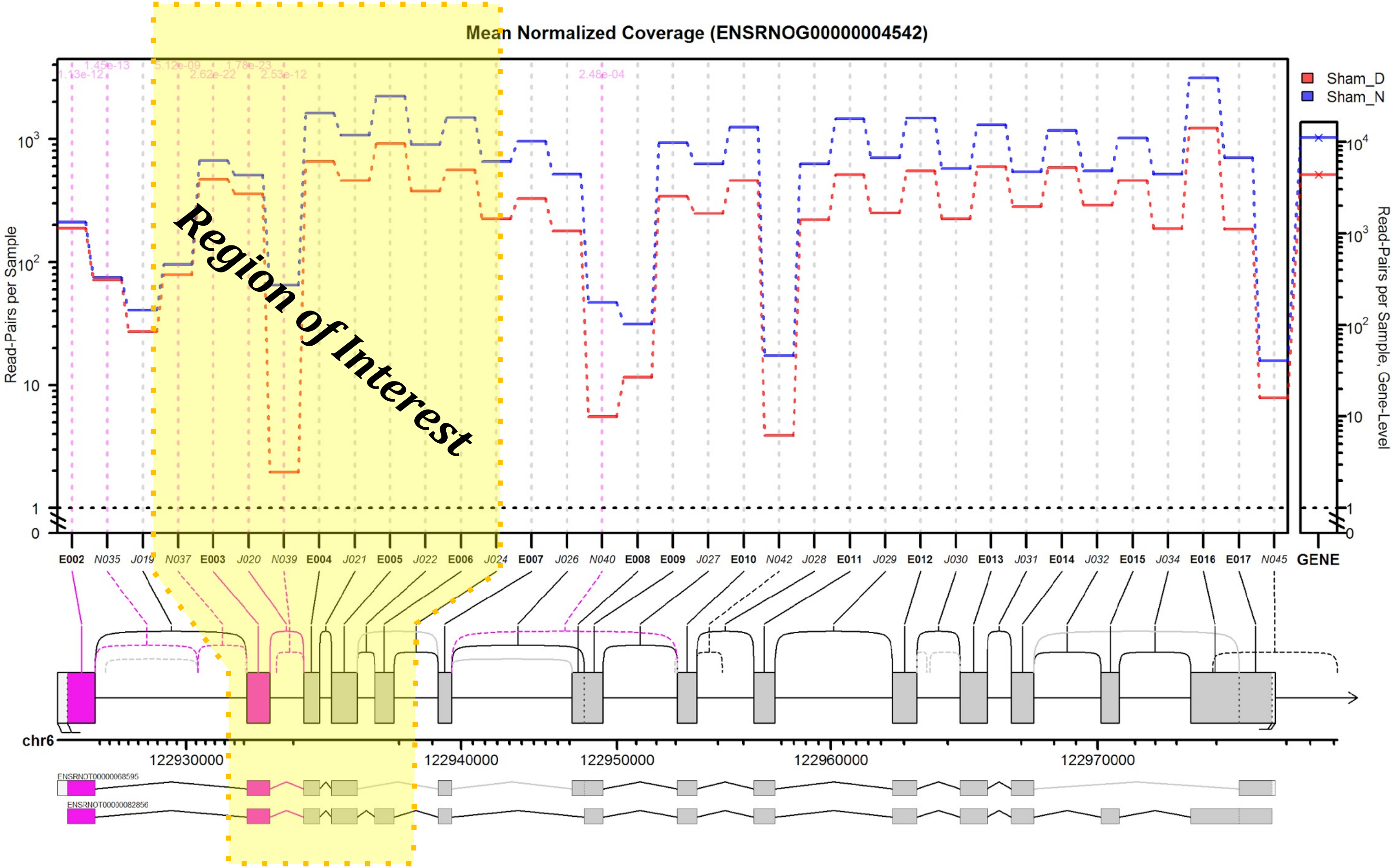


E002, E003

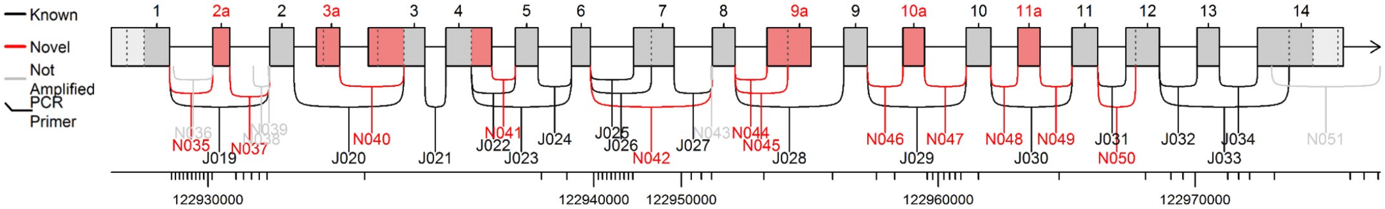
~30x higher at NIGHT

Novel SJ

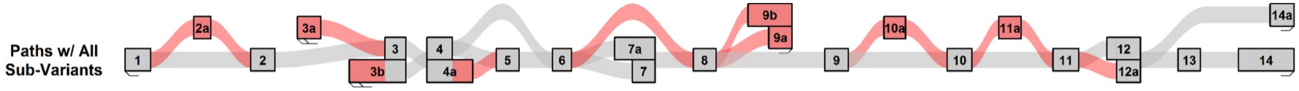
Applied example: the TTC8 Gene in rat pineal glands



Problem: Illumina data does not give us full-length Isoforms

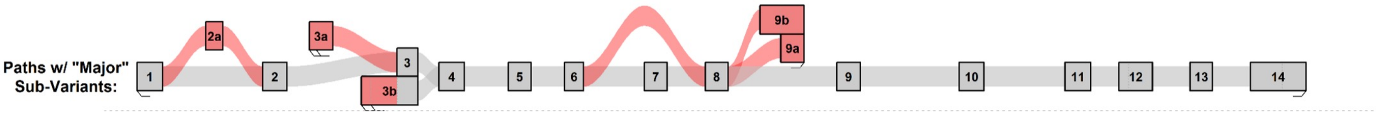


Take all possible paths through the gene:



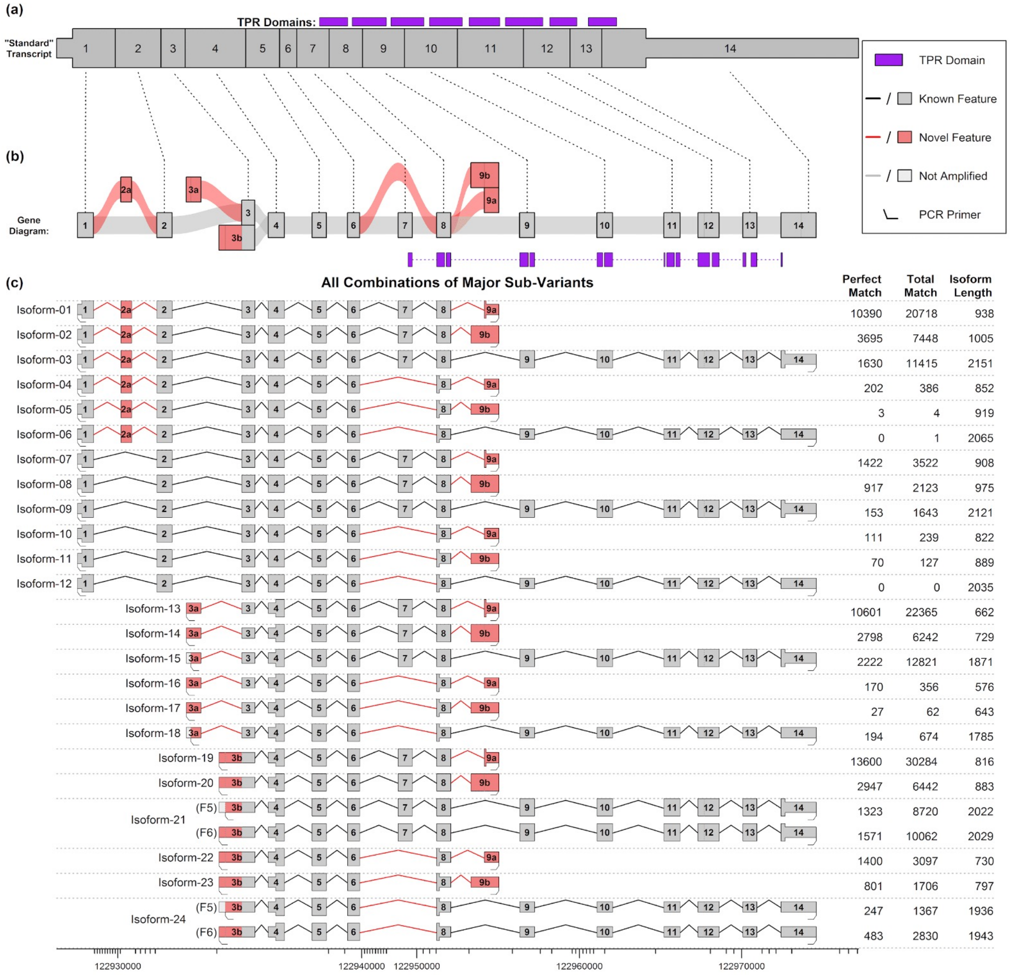
648 paths!
(potential isoforms)

Pick an (arbitrary) subset of high-coverage junctions:



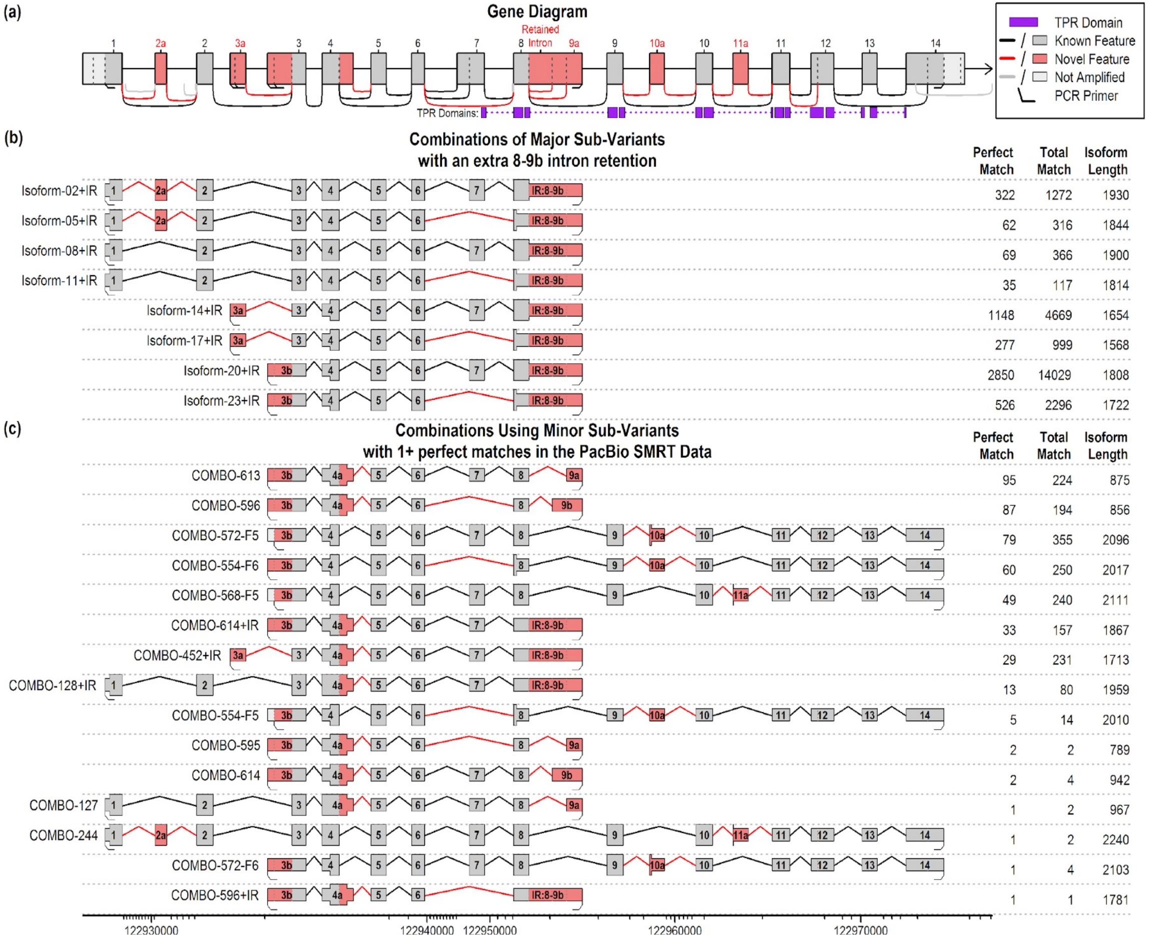
Still: 24 paths!

PacBio SMRT Sequencing Validation



		Perfect Match	Total Match	Isoform Length
Isoform-01		10390	20718	938
Isoform-02		3695	7448	1005
Isoform-03		1630	11415	2151
Isoform-04		202	386	852
Isoform-05		3	4	919
Isoform-06		0	1	2065
Isoform-07		1422	3522	908
Isoform-08		917	2123	975
Isoform-09		153	1643	2121
Isoform-10		111	239	822
Isoform-11		70	127	889
Isoform-12		0	0	2035
Isoform-13		10601	22365	662
Isoform-14		2798	6242	729
Isoform-15		2222	12821	1871
Isoform-16		170	356	576
Isoform-17		27	62	643
Isoform-18		194	674	1785
Isoform-19		13600	30284	816
Isoform-20		2947	6442	883
Isoform-21	(F5)	1323	8720	2022
	(F6)	1571	10062	2029
Isoform-22		1400	3097	730
Isoform-23		801	1706	797
Isoform-24	(F5)	247	1367	1936
	(F6)	483	2830	1943

PacBio SMRT Sequencing Validation



(b) Combinations of Major Sub-Variants with an extra 8-9b intron retention

	Perfect Match	Total Match	Isoform Length
Isoform-02+IR	322	1272	1930
Isoform-05+IR	62	316	1844
Isoform-08+IR	69	366	1900
Isoform-11+IR	35	117	1814
Isoform-14+IR	1148	4669	1654
Isoform-17+IR	277	999	1568
Isoform-20+IR	2850	14029	1808
Isoform-23+IR	526	2296	1722

(c) Combinations Using Minor Sub-Variants with 1+ perfect matches in the PacBio SMRT Data

	Perfect Match	Total Match	Isoform Length
COMBO-613	95	224	875
COMBO-596	87	194	856
COMBO-572-F5	79	355	2096
COMBO-554-F6	60	250	2017
COMBO-568-F5	49	240	2111
COMBO-614+IR	33	157	1867
COMBO-452+IR	29	231	1713
COMBO-128+IR	13	80	1959
COMBO-554-F5	5	14	2010
COMBO-595	2	2	789
COMBO-614	2	4	942
COMBO-127	1	2	967
COMBO-244	1	2	2240
COMBO-572-F6	1	4	2103
COMBO-596+IR	1	1	1781

Differential Splicing/Differential Isoform usage/Differential Exon Usage/Alternative Isoform Regulation/etc… Analysis

Main Takeaways

- Numerous tools exist.
  - Two basic types: count-based or transcript-based
- Not for the faint of heart
- Interpretation and examination of results should be extensive and in-depth
- Don’t take results at face value: validation is important
- Hot take: don’t bother unless you’re willing to spend a lot of time/resources interpreting and validating the results

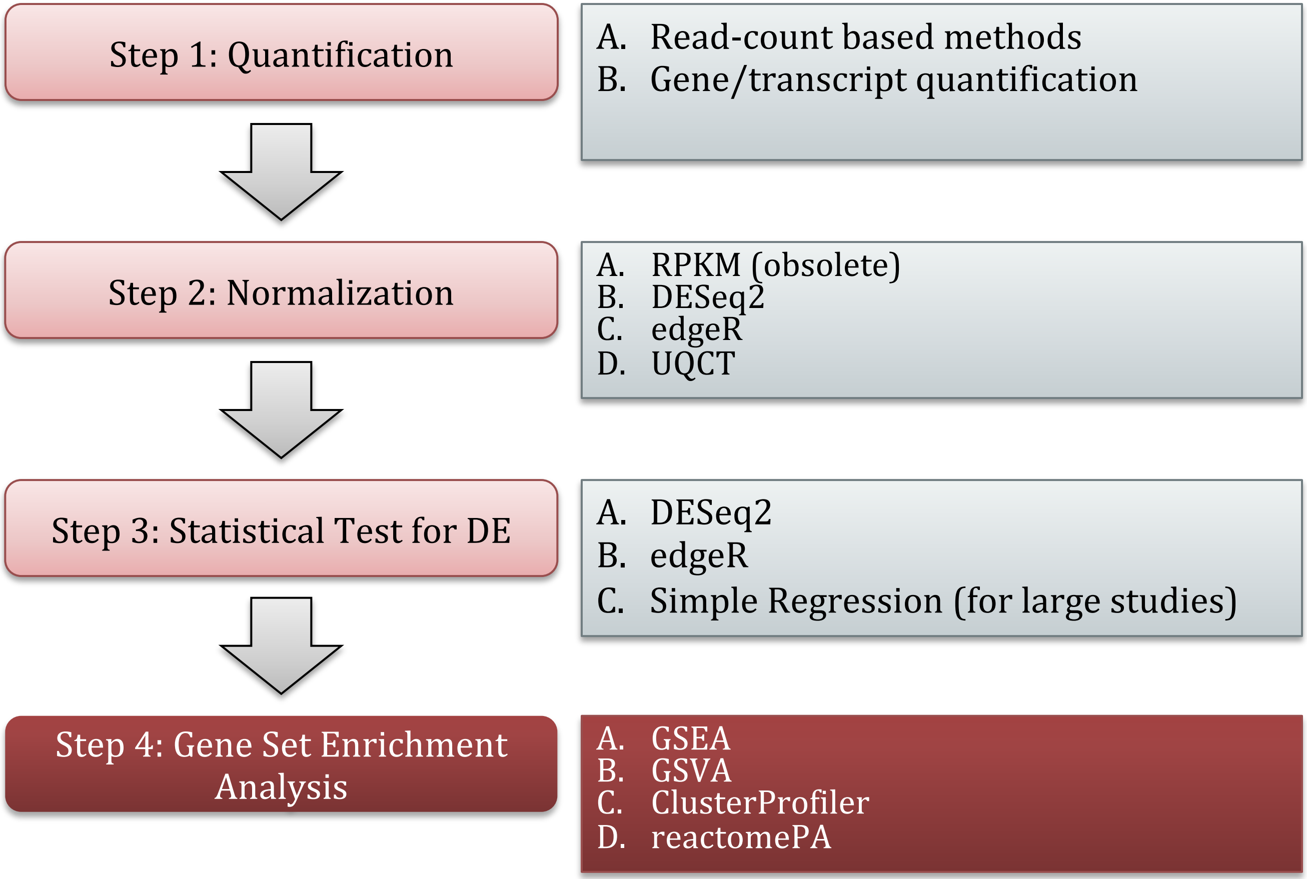
Pathway and Gene Set Enrichment Analysis

Stephen Hartley

Staff Scientist

DCEG/LGS

Gene Set Enrichment Analysis



Step 1: Quantification

A. Read-count based methods
B. Gene/transcript quantification

Step 2: Normalization

A. RPKM (obsolete)
B. DESeq2
C. edgeR
D. UQCT

Step 3: Statistical Test for DE

A. DESeq2
B. edgeR
C. Simple Regression (for large studies)

Step 4: Gene Set Enrichment Analysis

A. GSEA
B. GSVA
C. ClusterProfiler
D. reactomePA

Gene Set Enrichment Analysis

There are many different methods/tools for performing this sort of analysis, differ in details, but basic idea:

- Look at *sets* of genes
- Are the significant/substantial differentials ***concentrated*** in these sets

Study Question:

Are there pathways, gene-ontology keywords, or other gene sets that are ***disproportionately represented*** in the ***results*** of an RNA-Seq analysis?

Gene Set Enrichment Methods

What can it do?

- Give you more information about the biological processes involved
- Assist in interpretation of RNA-Seq expression results
- Provide more information

Gene Set Enrichment Methods

What is it NOT?

- Not always necessary in all differential expression analysis

- Not an easy way to get a p-value if you don’t come up with anything directly
  - It can be used for this in certain limited circumstances, but only *carefully*

What Gene Lists?

There are numerous different options for finding useful gene lists for use with GSEA.

- Molecular Signatures Database (MSigDB)
- Gene Ontology (GO)
- REACTOME pathway database

MSigDB

LOTS of gene sets, separated into several main groups:

H: Hallmark gene sets (50 sets)

- ○ Highly-curated sets that represent merge of multiple gene sets

C1: Positional gene sets (300 sets)

- ○ Gene sets corresponding to human chromosome cytogenetic bands

C2: Curated gene sets (6495 sets)

- ○ Curated from various sources, including online databases and the literature.
- ○ Many also contributed by individual experts.

C3: Regulatory target gene sets (3713 sets)

- ○ Gene sets representing potential targets of regulation by transcription factors or microRNAs.

C4: Computational Gene Sets (858 sets)

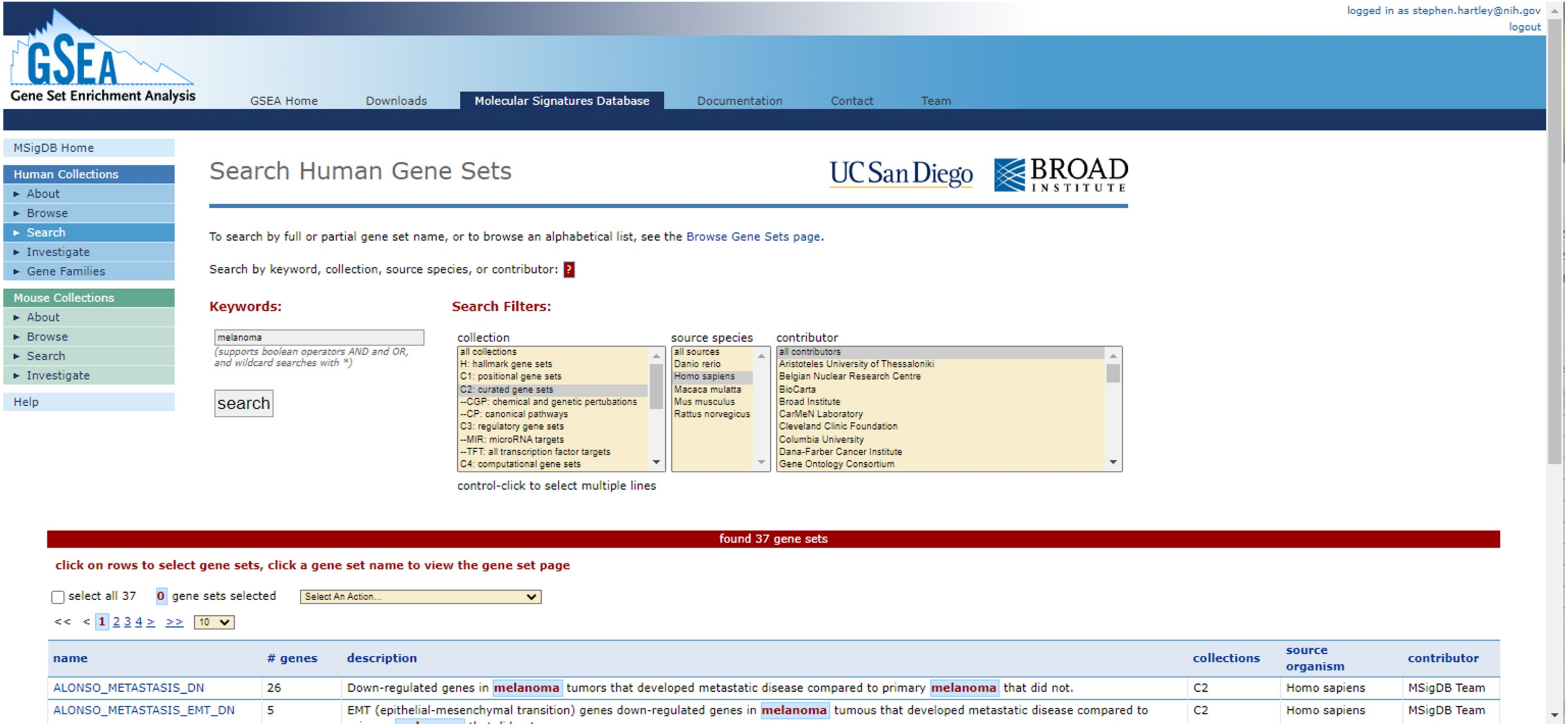
- ○ Gene sets generated *in silico* via data mining

MSigDB

Generally NOT recommended to test against entire MSigDB database.

Better to test against gene sets that are related to your study topic / known biological processes.

MSigDB



Search Human Gene Sets

To search by full or partial gene set name, or to browse an alphabetical list, see the Browse Gene Sets page.

Search by keyword, collection, source species, or contributor: ?

Keywords:

melanoma

*(supports boolean operators AND and OR, and wildcard searches with *)*

search

Search Filters:

collection
all collections
H: hallmark gene sets
C1: positional gene sets
C2: curated gene sets
--CGP: chemical and genetic pertubations
--CP: canonical pathways
C3: regulatory gene sets
--MIR: microRNA targets
--TFT: all transcription factor targets
C4: computational gene sets

source species
all sources
Danio rerio
Homo sapiens
Macaca mulatta
Mus musculus
Rattus norvegicus

contributor
all contributors
Aristoteles University of Thessaloniki
Belgian Nuclear Research Centre
BioCarta
Broad Institute
CarMeN Laboratory
Cleveland Clinic Foundation
Columbia University
Dana-Farber Cancer Institute
Gene Ontology Consortium

control-click to select multiple lines

found 37 gene sets

click on rows to select gene sets, click a gene set name to view the gene set page

☐ select all 37   0 gene sets selected   Select An Action...

<< < 1 2 3 4 > >> 10

name	# genes	description	collections	source organism	contributor
ALONSO_METASTASIS_DN	26	Down-regulated genes in **melanoma** tumors that developed metastatic disease compared to primary **melanoma** that did not.	C2	Homo sapiens	MSigDB Team
ALONSO_METASTASIS_EMT_DN	5	EMT (epithelial-mesenchymal transition) genes down-regulated genes in **melanoma** tumous that developed metastatic disease compared to	C2	Homo sapiens	MSigDB Team

Gene Ontology

The Gene Ontology Resource (GO) is a large database of genes organized into hierarchical keywords.

Examples:

- “biological process”
- “mitochondrion”
- “glucose transmembrane transport”
- “amino acid binding”

Each gene will belong to ***many*** GO terms.

Each term will often belong to ***many*** higher-level terms (ex: “metabolic process” belongs to “biological process”)

Gene Set Enrichment Methods

There are numerous different methods/tools for performing this sort of analysis:

- ClusterProfiler
  - Can use GO terms or other annotated lists

- reactomePA
  - Uses “REACTOME” database
  - Same developer as ClusterProfiler

- “GSEA”:
  - Developed by same group that created MSigDB

- GSVA: “Gene Set Variation Analysis”
  - Optionally allows for mixture of up/down regulation

Gene Set Enrichment Analysis
Main takeaways

- Takes the *results* of Differential Expression analysis as *input*
- Can help with interpretation, can give more information about biological functions involved, etc.
- Several options for gene set database
- Several options for software analysis tools

Clustering and classification

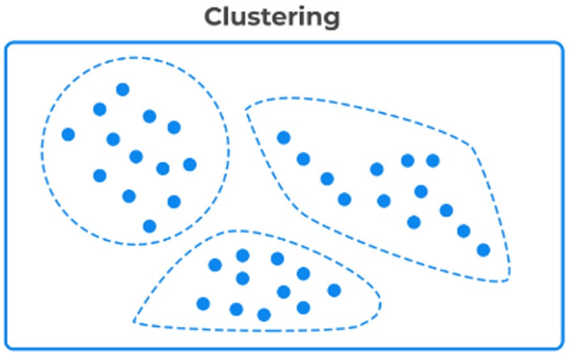
Wei Zhao

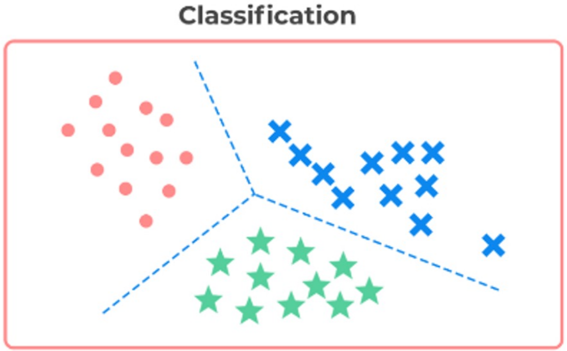
Research fellow

DCEG/ITEB

Clustering and classification – Common approaches

- Unsupervised machine learning: models to group samples with similar features together
  - Dimensionality reduction
  - Clustering
- Supervised machine learning: models to predict class of new sample
  - Feature selection
  - Classification





https://www.analyticsvidhya.com/blog/2021/11/quick-tutorial-clustering-data-science/

Clustering and classification – Unsupervised machine learning

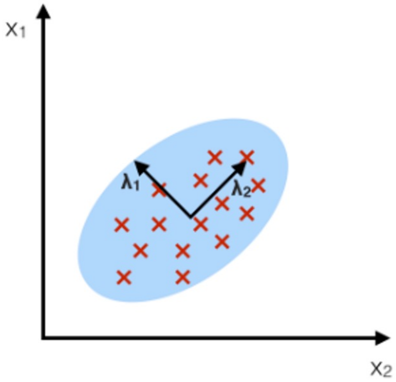
- Dimensionality reduction:
  - Principal component analysis (PCA)
  - t-distributed stochastic neighbor embedding (t-SNE)
  - Uniform Manifold Approximation (UMAP)
- Clustering
  - Hierarchical clustering
  - K-Means
  - Model-based clustering
  - Deep learning

Clustering and classification – Unsupervised machine learning

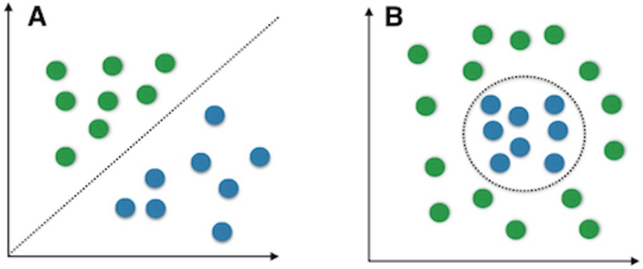
- **Dimensionality reduction** is the process of reducing the number of features when exploring the structure of high-dimensional data.
- Identifies the most relevant information
- Reduces computational time
- Does lose some information
- Used for data classification and visualization

Clustering and classification – Unsupervised machine learning

- PCA
  - Find component axes that maximize the variance of the data
  - A linear method
  - Fast and easy to apply
  - Interpretable results
  - Can be used to filter the top significant PCs
  - Limited performance when the data is not linearly separable
  - Not optimized for 2D visualization



Linear vs. nonlinear problems

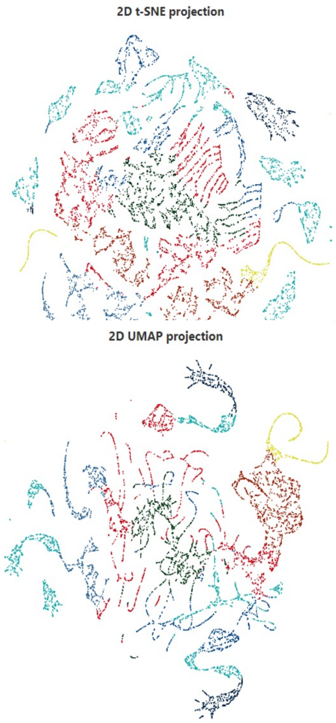


https://sebastianraschka.com/Articles/2014_python_lda.html
https://sebastianraschka.com/Articles/2014_kernel_pca.html

Clustering and classification – Unsupervised machine learning

When there are too many samples and the goal is to is to identify cell types (e.g. scRNA-seq), usually we use tSNE and UMAP.

- tSNE and UMAP are both graph-based non-linear methods. And are optimized for 2D visualization
- tSNE is a stochastic algorithm
- tSNE preserve local structure, but not global. (i.e. the distant proximity is NOT informative).
- UMAP preserves better the global structure, and can control the balance between local and global structure with parameters.
- UMAP is much faster than tSNE.
- Both requires optimization of hyper-parameters.
- Some nice posts:
  - https://distill.pub/2016/misread-tsne/
  - https://www.youtube.com/watch?v=NEaUSP4YerM
  - https://pair-code.github.io/understanding-umap/



2D t-SNE projection

2D UMAP projection

Maaten et al (2008) Journal of Machine Learning Research

Clustering and classification – Unsupervised machine learning

- Clustering
  - Hierarchical clustering (R function: hclust())
  - K-Means (R package “factoextra”,”cluster”)
  - Model-based clustering, self-organizing maps (R package: “kohonen”, “MBCluster.Seq”)
  - Deep learning
  - More methods: Oyelade et al. Bioinform Biol Insights. 2016; 10: 237–253.

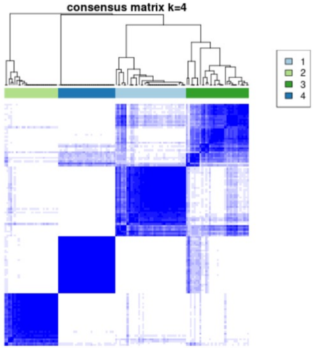
Clustering and classification – Unsupervised machine learning

- Hierarchical clustering vs. K-Means
  - Hierarchical clustering is favorable when
    - there number of the clusters needs to be explored.
    - we want to interpret the clustering structure.
  - K-Means is favorable when
    - the number of clusters is known.
    - the numbers of the features and samples are large.

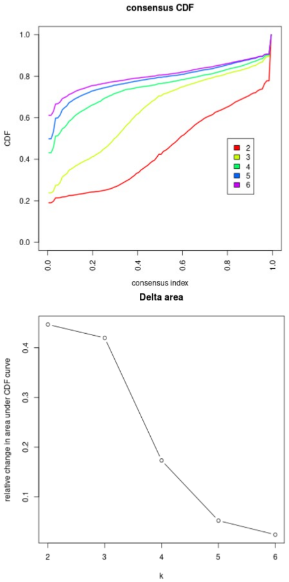
Clustering and classification – Unsupervised machine learning

- Consensus Clustering could be used to determine the number of possible clusters within gene expression data set.

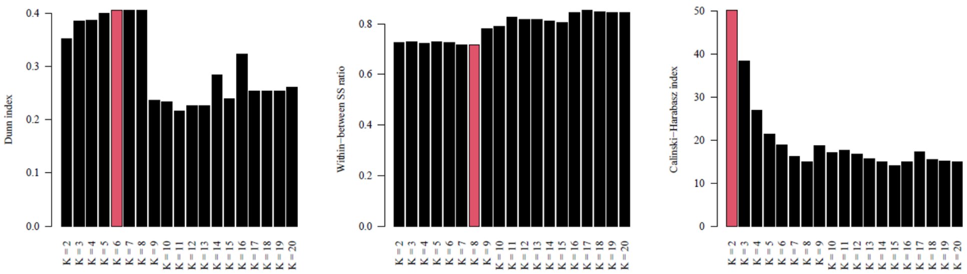
consensus



Consensus Cumulative Distribution Function (CDF) Plot



Cluster_separation_stats



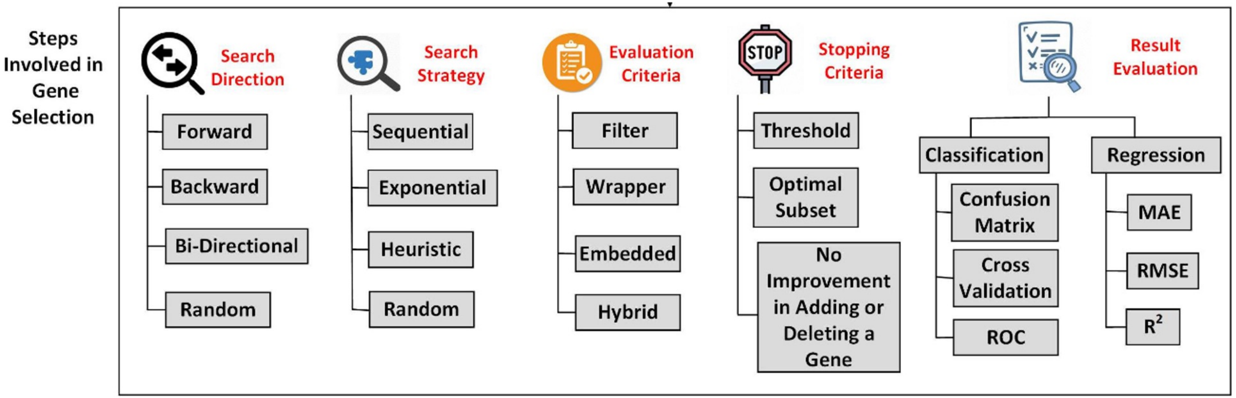
R package: ConsensusClusterPlus

Clustering and classification – Unsupervised machine learning

- How to validate the results?
  - External validation: do genes/samples clusters correspond to their known functions or biological features?
  - Internal validation: Examine the inter- and intra-cluster relationship (e.g. silhouette width.)
  - Relative validation: How does it compare to other clustering? Is there consensus in results?

Clustering and classification – Supervised machine learning

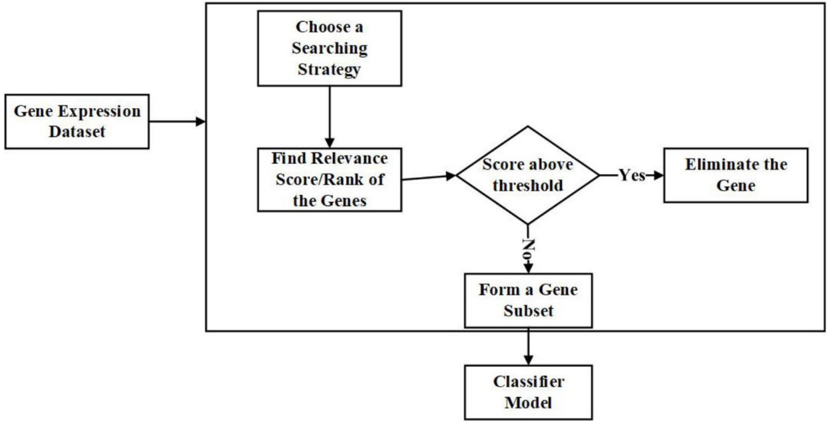
- **Feature selection** is a technique to identify the informative genes and to remove the redundant and irrelevant genes. It does NOT obtain new features. Compared to dimensionality reduction, the models are more interpretable.
- More resource:
  - https://www.frontiersin.org/articles/10.3389/fgene.2020.603808/full
  - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00441-x



Mahendran et al., Front. Genet., 10 December 2020

Clustering and classification – Supervised machine learning

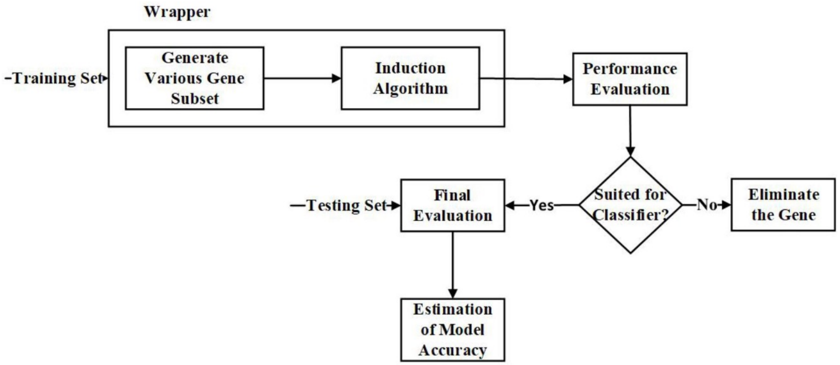
- Filter Feature selection
  - Example: Pearson correlation, Fisher score, mutual information, model-based ranking.
  - Advantage: do not depend on classifiers; fast; can scaled to large data sets.
  - Disadvantage: Some filter feature selection algorithms consider univariate features, thus potentially ignore the feature dependencies.



Mahendran et al., Front. Genet., 10 December 2020

Clustering and classification – Supervised machine learning

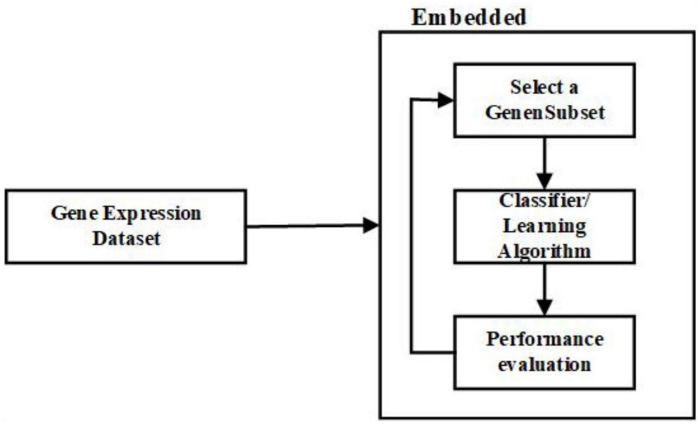
- Wrapper Feature selection
  - Example: Hill climbing, Forward selection, Backward elimination.
  - Advantage: the model tracks the feature dependencies.
  - Disadvantage: computationally intensive for data sets with high dimension.



Mahendran et al., Front. Genet., 10 December 2020

Clustering and classification – Supervised machine learning

- Embedded feature selection
  - Example: LASSO regression, L1 Regression, ID3, Random forest.
  - Advantage: improved accuracy compared to filter and wrapper methods; computationally less intensive than wrapper methods; perform feature selection and learning algorithm in parallel.



Mahendran et al., Front. Genet., 10 December 2020

Clustering and classification – Supervised machine learning

- Classification
  - Discriminant analysis
  - Random forest
  - Support Vector Machine
  - K-nearest neighbor (KNN)
  - Naïve Bayesian classifier
  - Decision tree
  - Neural network
  - Deep learning
- Resource: a review on the classifier software for gene expression data (http://dx.doi.org/10.14257/ijbsbt.2015.7.4.10)

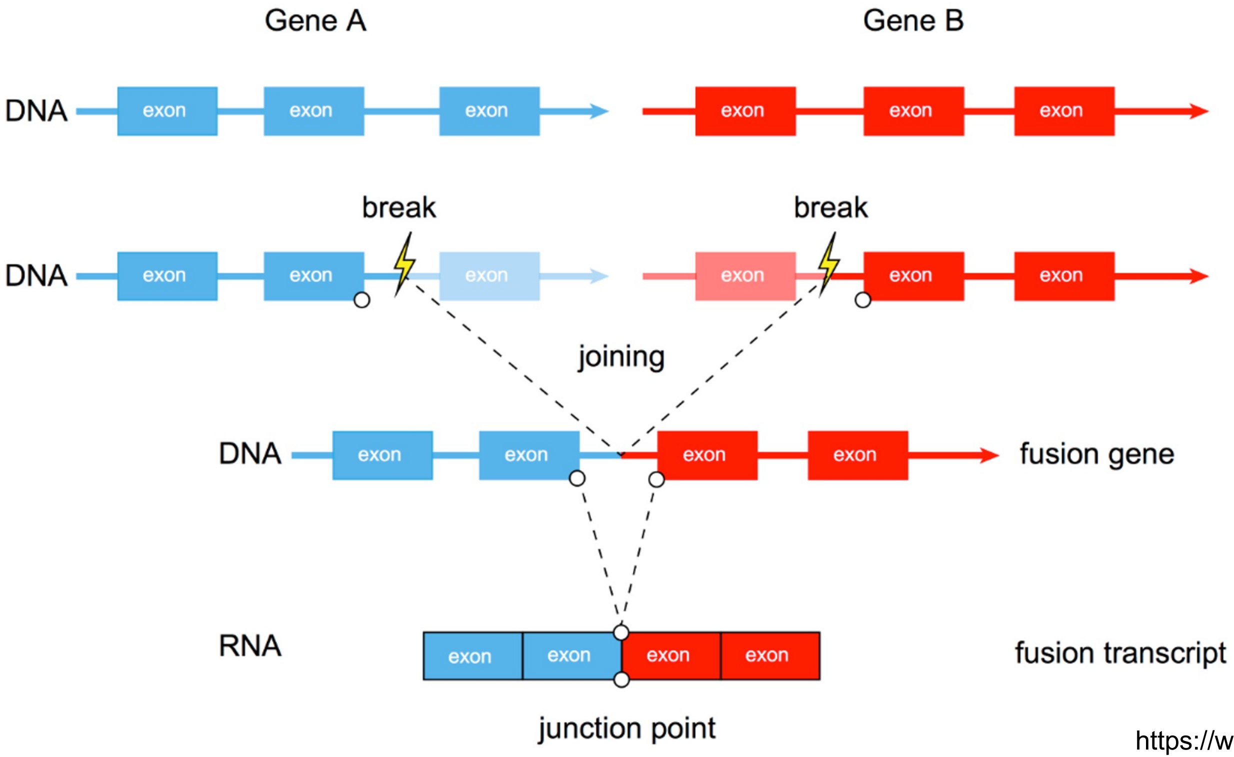
Fusion gene detection

Difei Wang

Bioinformatics Manager

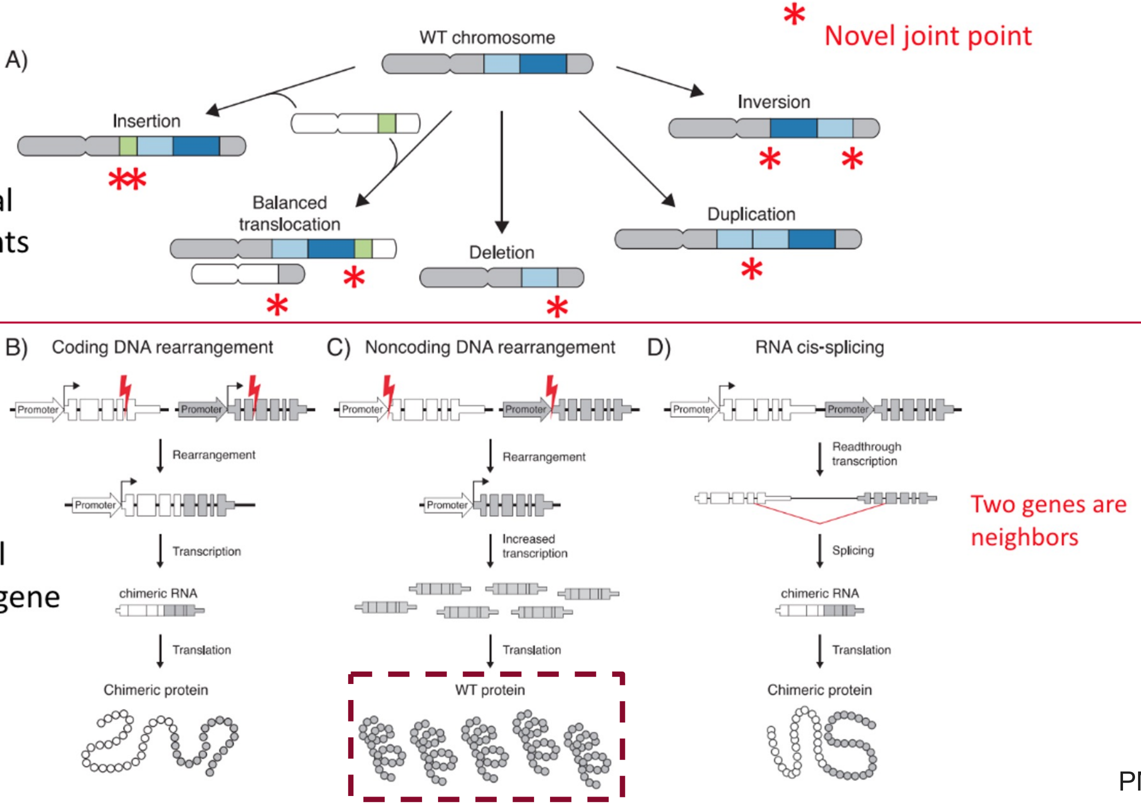
DCEG/CGR

What are fusion genes?



https://www.tumorfusions.org

What may cause fusion genes?



* Novel joint point

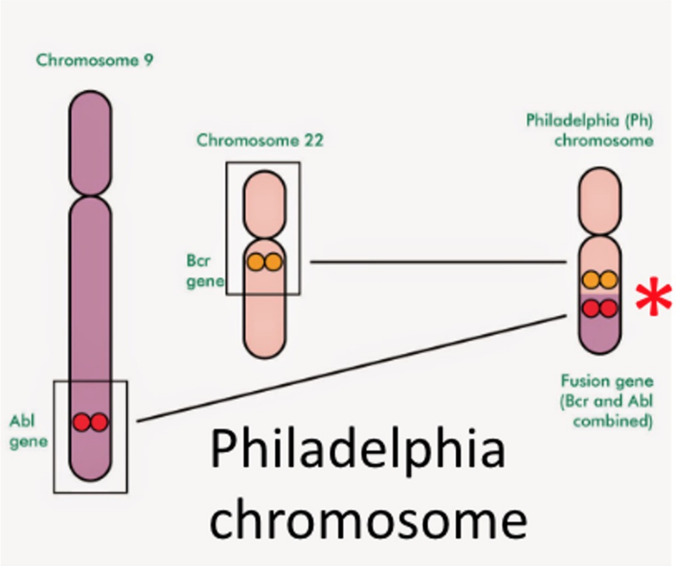
Types of chromosomal rearrangements

functional outcomes of gene fusions

Two genes are neighbors

PMID: 32307742

Why are they important?

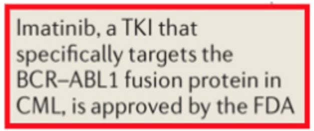


1960s
Peter C. Nowell
David A. Hungerford

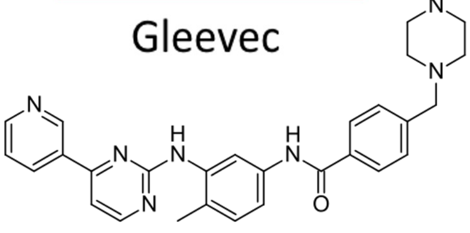
A piece of chromosome 9 and a piece of chromosome 22 break off and trade places.

~ 20 yrs later BCR-ABL fusion in CML patients

The BCR-ABL gene is formed on chromosome 22 where the piece of chromosome 9 attaches.

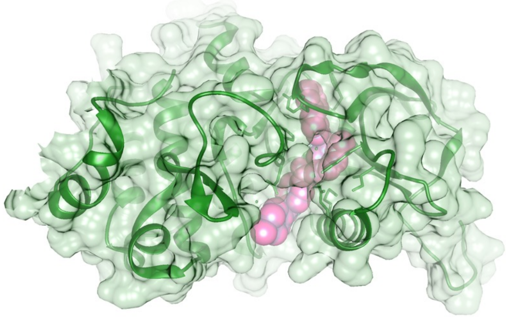


Imatinib, a TKI that specifically targets the BCR–ABL1 fusion protein in CML, is approved by the FDA



Gleevec

FDA approved in 2001



BCR-ABL kinase (green)
Gleevec (pink) 1IEP

Thanks to the next-generation sequencing and TCGA project, more novel disease-causing fusion genes were discovered. They can serve as potential targets for new therapies.

How to detect fusion gene?

- Guided approach
  - Cytogenetic data
  - Fluorescence in situ hybridization (FISH)
  - High-throughput array-based analyses

- Unbiased gene fusion detection
  - Deep-sequencing technologies/Next-generation sequencing (NGS)

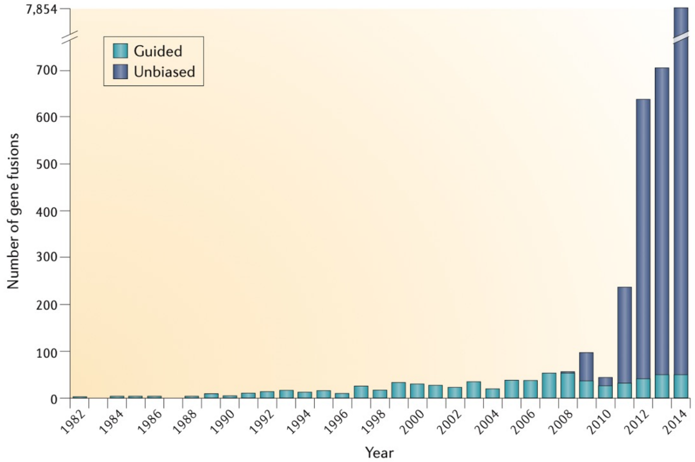
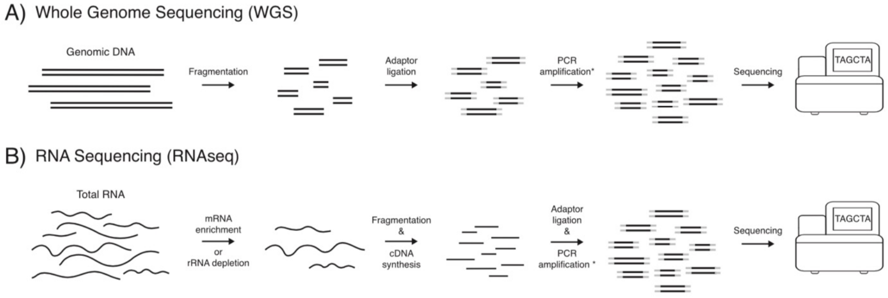


Figure 4 | **Gene fusion reports.** The number of new gene fusions reported each year from 1982 to 2014 (REF. 5) is shown. Unbiased gene fusions are those detected by deep sequencing, whereas guided gene fusions represent those identified as the result of molecular analyses directed by prior (cyto)genetic information.

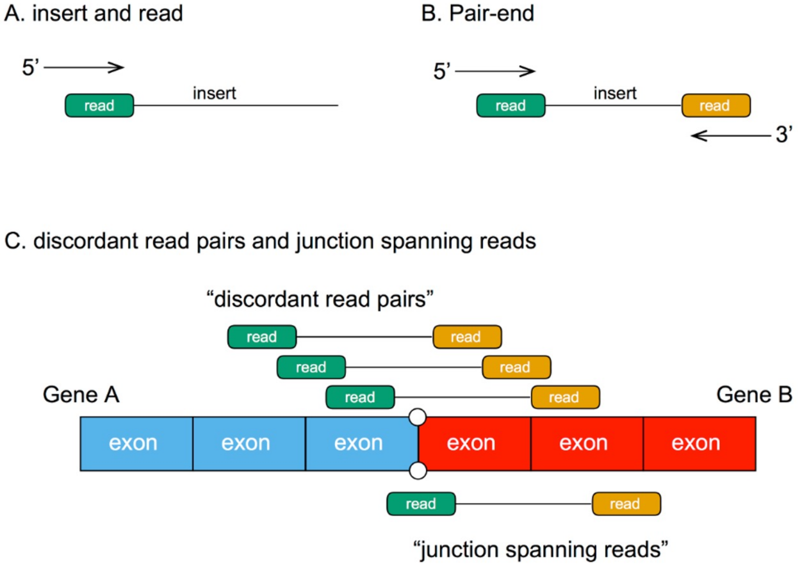
PMID: 25998716

How to detect fusion gene through sequencing?

Step1: DNA-seq and RNA-seq



Step2: in silico detection of fusion genes


A. insert and read

B. Pair-end

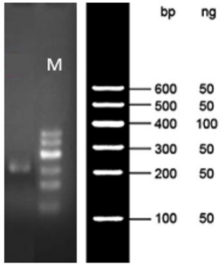
C. discordant read pairs and junction spanning reads

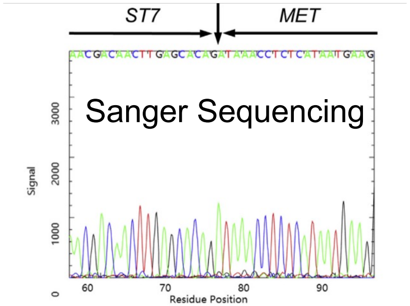
https://www.tumorfusions.org

PMID: 32307742

Step3: Validation of identified fusion genes

RT-PCR




Sanger Sequencing

Current bioinformatics tools for gene fusion detection

- Benchmark
  - Liu et al. NAR 2015
    - 24 tools, tested 15
    - Overall performance rating: SOAPfuse > **FusionCatcher** > **JAFFA**
    - Ericscript performed well on synthetic dataset
    - SOAPfuse most computational demanding.
  - Kumer et al. Sci. Rep. 2016
    - 12 tools tested
    - Performance rating: Ericscript > **FusionCatcher** > **JAFFA**
  - SOAPfuse failed to finish sometime.

Suggested tools

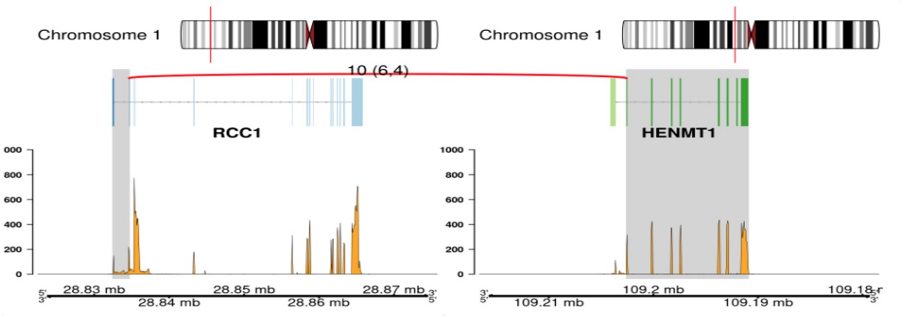
EricScript
FusionCatcher
JAFFA

STAR-Fusion (2019)
Arriba (2019)
(the winner of DREAM SMC-RNA Challenge in 2018)

PMID: 26582927
PMID: 26862001

Visualization of gene fusion

RCC1-HENMT1 Novel Fusion in Testicular Germ Cell Tumors



+ strand

- strand gene converted
to + for visualization

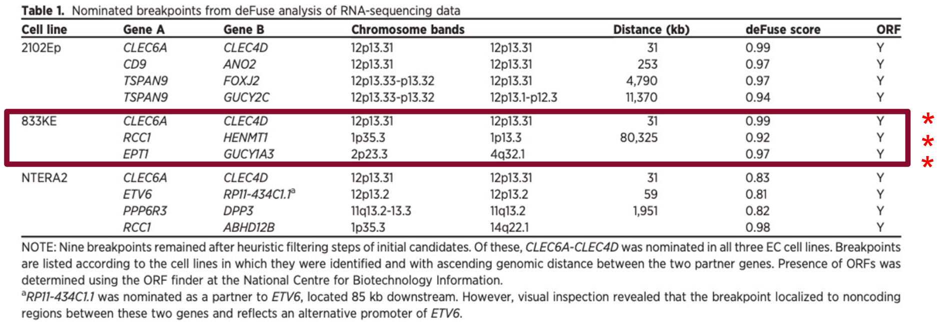
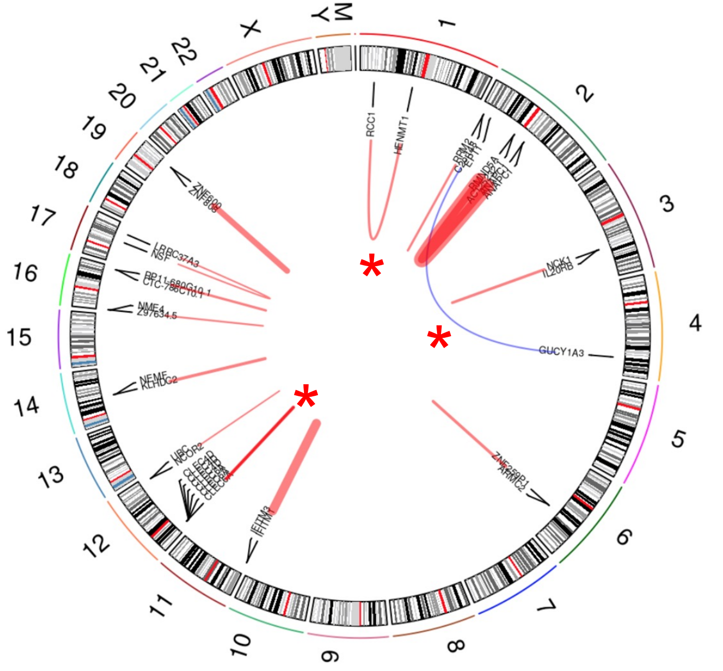


Table 1. Nominated breakpoints from deFuse analysis of RNA-sequencing data

Cell line	Gene A	Gene B	Chromosome bands		Distance (kb)	deFuse score	ORF	
2102Ep	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.99	Y	
	CD9	*ANO2*	12p13.31	12p13.31	253	0.97	Y	
	TSPAN9	*FOXJ2*	12p13.33-p13.32	12p13.31	4,790	0.97	Y	
	TSPAN9	*GUCY2C*	12p13.33-p13.32	12p13.1-p12.3	11,370	0.94	Y	
833KE	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.99	Y	*
	RCC1	*HENMT1*	1p35.3	1p13.3	80,325	0.92	Y	*
	EPT1	*GUCY1A3*	2p23.3	4q32.1		0.97	Y	*
NTERA2	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.83	Y	
	ETV6	*RP11-434C1.1*[a]	12p13.2	12p13.2	59	0.81	Y	
	PPP6R3	*DPP3*	11q13.2-13.3	11q13.2	1,951	0.82	Y	
	RCC1	*ABHD12B*	1p35.3	14q22.1		0.98	Y	

NOTE: Nine breakpoints remained after heuristic filtering steps of initial candidates. Of these, *CLEC6A-CLEC4D* was nominated in all three EC cell lines. Breakpoints are listed according to the cell lines in which they were identified and with ascending genomic distance between the two partner genes. Presence of ORFs was determined using the ORF finder at the National Centre for Biotechnology Information.

[a] *RP11-434C1.1* was nominated as a partner to *ETV6*, located 85 kb downstream. However, visual inspection revealed that the breakpoint localized to noncoding regions between these two genes and reflects an alternative promoter of *ETV6*.



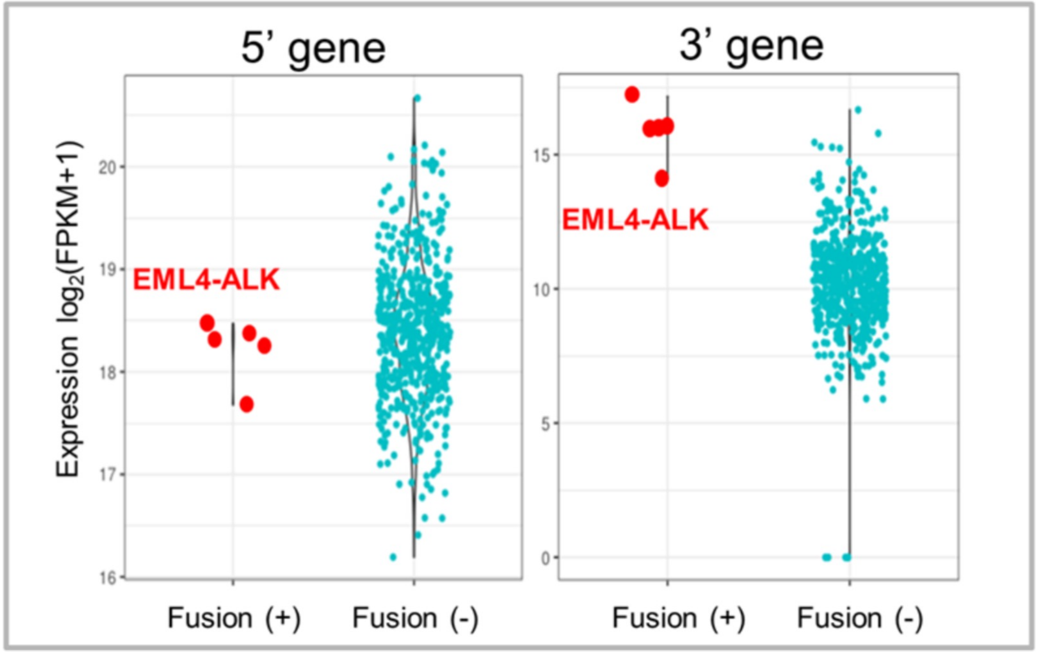
defuse 833KE, circos plot
Others either read through or intron involved.

PMID: 26659575

Further evidence of detected fusion genes

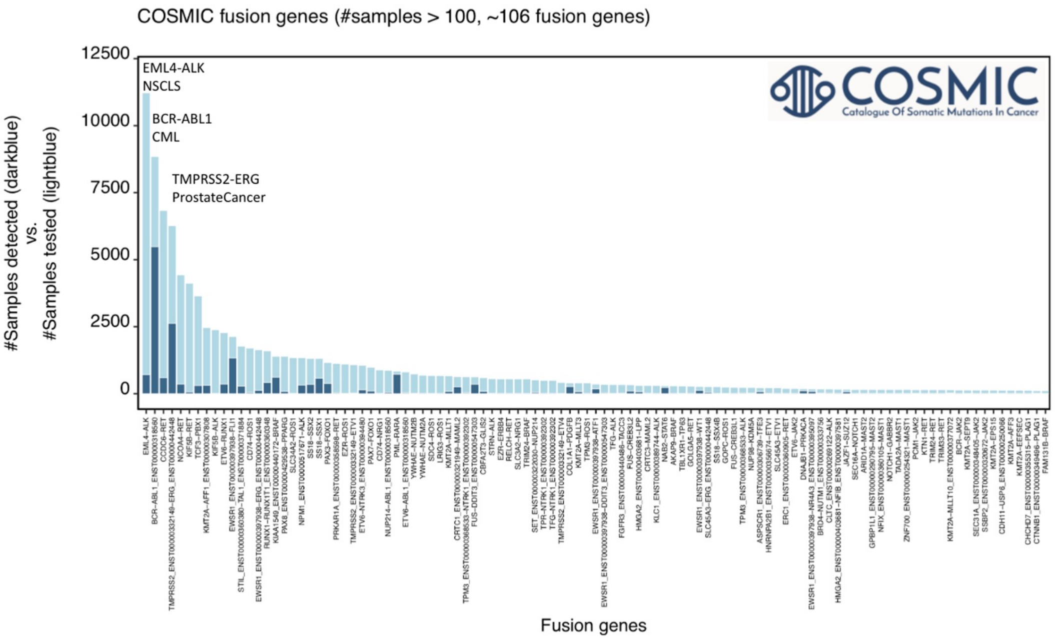
Expression changes of fusion gene

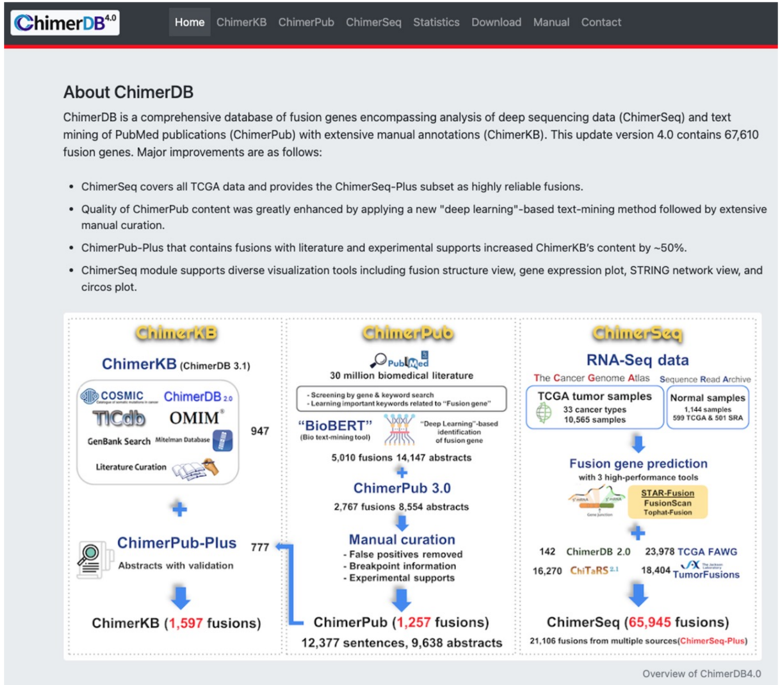
EML4-ALK fusion gene in NSCLC



ChimerDB 4.0
PMID: 31680157

Novel or known fusion genes





About ChimerDB

ChimerDB is a comprehensive database of fusion genes encompassing analysis of deep sequencing data (ChimerSeq) and text mining of PubMed publications (ChimerPub) with extensive manual annotations (ChimerKB). This update version 4.0 contains 67,610 fusion genes. Major improvements are as follows:

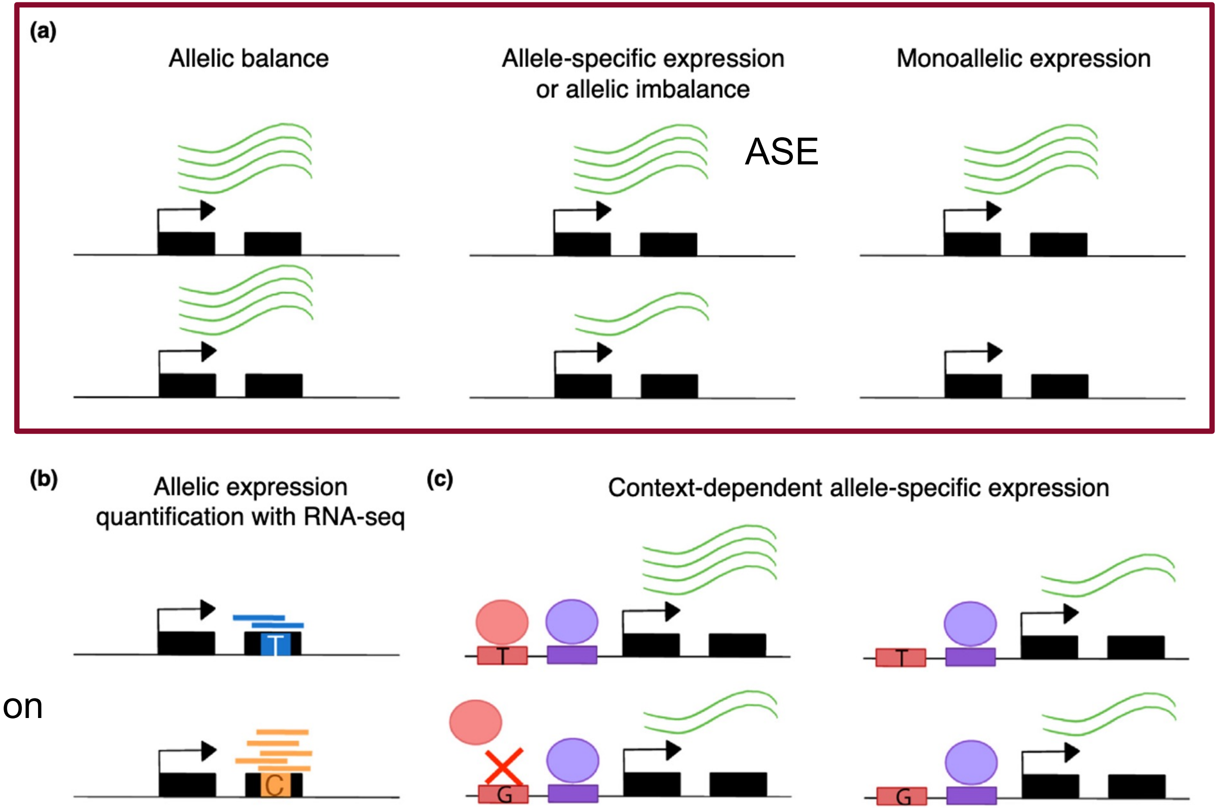
- ChimerSeq covers all TCGA data and provides the ChimerSeq-Plus subset as highly reliable fusions.
- Quality of ChimerPub content was greatly enhanced by applying a new "deep learning"-based text-mining method followed by extensive manual curation.
- ChimerPub-Plus that contains fusions with literature and experimental supports increased ChimerKB's content by ~50%.
- ChimerSeq module supports diverse visualization tools including fusion structure view, gene expression plot, STRING network view, and circos plot.

Overview of ChimerDB4.0

Allele-specific expression

Difei Wang
Bioinformatics Manager
DCEG/CGR

What is allele-specific expression (ASE)?


(a)
Allelic balance
Allele-specific expression or allelic imbalance
ASE
Monoallelic expression
(b)
Allelic expression quantification with RNA-seq
T
C
(c)
Context-dependent allele-specific expression
T
G
T
G

e.g. X-chromosome inactivation

ASE quantification

Potential mechanism

PMID: 33383480

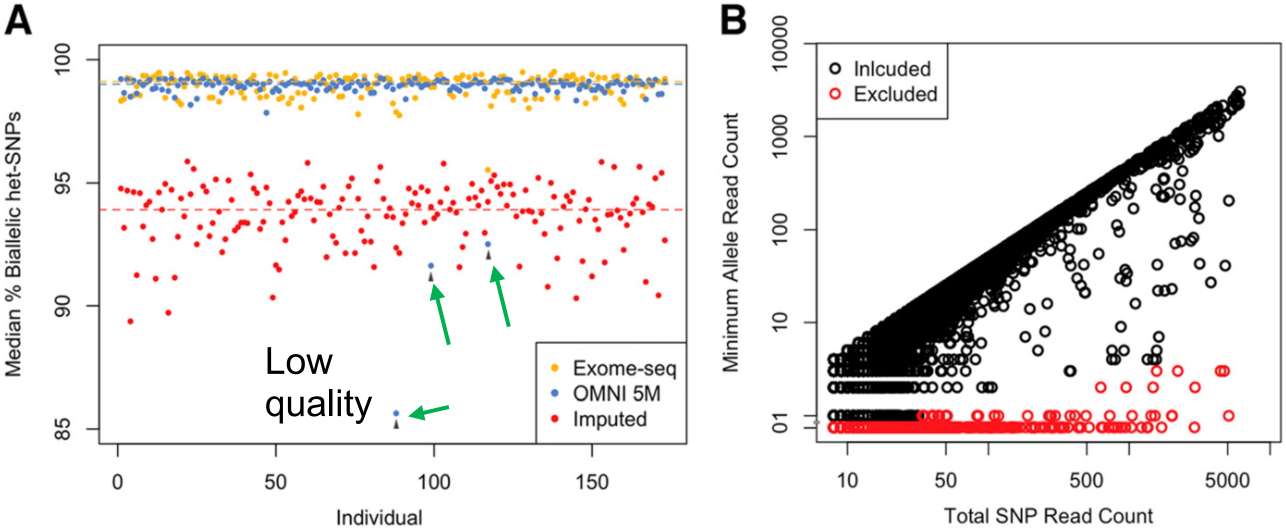
Current tools for ASE detection

- ASEReadCounter (GATK package)
  - From fastq to a table of allelic counts
- QIlelic
  - Starting with the allelic count table, estimate allelic imbalance and overdispersion
- EMASE-Zero
  - From bam -> alntools -> table
- ASEP
  - From the allelic count table
- phASER
  - From fastq
- WASP
  - From aligned BAM + SNP info to correct reads mapping, count

PMID: 34099647
PMID: 29444201
PMID: 27605262

Quality control of genotype data

- Het-SNP only
- Genotyping error can be an essential source of false signals of allelic imbalance.
- Errors are more common in imputed data



Red: not supported by RNA-seq data

PMID: 33383480

Quality control of allele counting

- All the reads counted over a site indeed originate from that genomic locus. Uniquely mapped reads only.

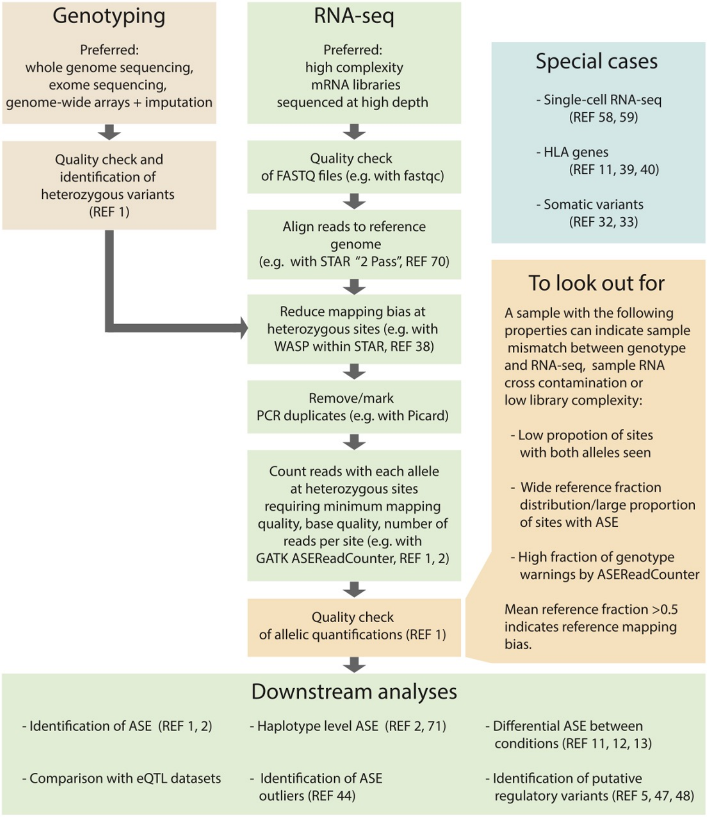
- All reads from that locus are counted.

- Allelic mapping bias. Remove ~ 20% het-SNPs fall within regions of low mappability

- Use a variant-aware aligner like GSNAP, or align to a personalized genome.

PMID: 33383480

Guidelines for allele-specific expression analysis



Genotyping

Preferred:
whole genome sequencing,
exome sequencing,
genome-wide arrays + imputation

Quality check and identification of heterozygous variants (REF 1)

RNA-seq

Preferred:
high complexity mRNA libraries sequenced at high depth

Quality check of FASTQ files (e.g. with fastqc)

Align reads to reference genome (e.g. with STAR "2 Pass", REF 70)

Reduce mapping bias at heterozygous sites (e.g. with WASP within STAR, REF 38)

Remove/mark PCR duplicates (e.g. with Picard)

Count reads with each allele at heterozygous sites requiring minimum mapping quality, base quality, number of reads per site (e.g. with GATK ASEReadCounter, REF 1, 2)

Quality check of allelic quantifications (REF 1)

Special cases

- Single-cell RNA-seq (REF 58, 59)
- HLA genes (REF 11, 39, 40)
- Somatic variants (REF 32, 33)

To look out for

A sample with the following properties can indicate sample mismatch between genotype and RNA-seq, sample RNA cross contamination or low library complexity:

- Low propotion of sites with both alleles seen
- Wide reference fraction distribution/large proportion of sites with ASE
- High fraction of genotype warnings by ASEReadCounter

Mean reference fraction >0.5 indicates reference mapping bias.

Downstream analyses

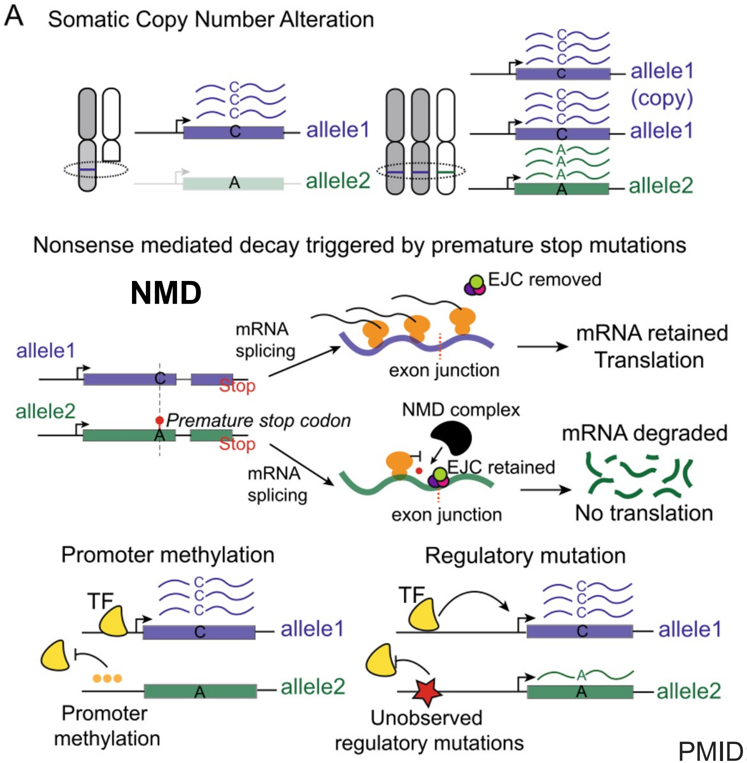
- Identification of ASE (REF 1, 2)
- Haplotype level ASE (REF 2, 71)
- Differential ASE between conditions (REF 11, 12, 13)
- Comparison with eQTL datasets
- Identification of ASE outliers (REF 44)
- Identification of putative regulatory variants (REF 5, 47, 48)

PMID: 33383480

ASE in high-risk neuroblastoma

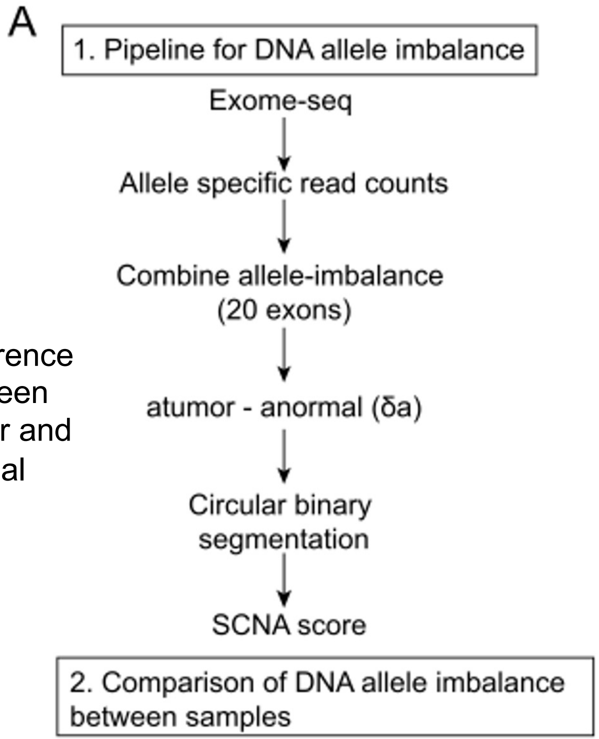
- Few recurrent somatic mutations
- Frequent somatic copy number variations (SCNAs)
- 96 high-risk neuroblastoma tumors
- Identified 1043 genes with recurrent, neuroblastoma-specific allele-specific expression (NB-ASE)

Four example mechanisms causing ASE



PMID: 35246212

DNA allele imbalance



1. Pipeline for DNA allele imbalance

Exome-seq

Allele specific read counts

Combine allele-imbalance (20 exons)

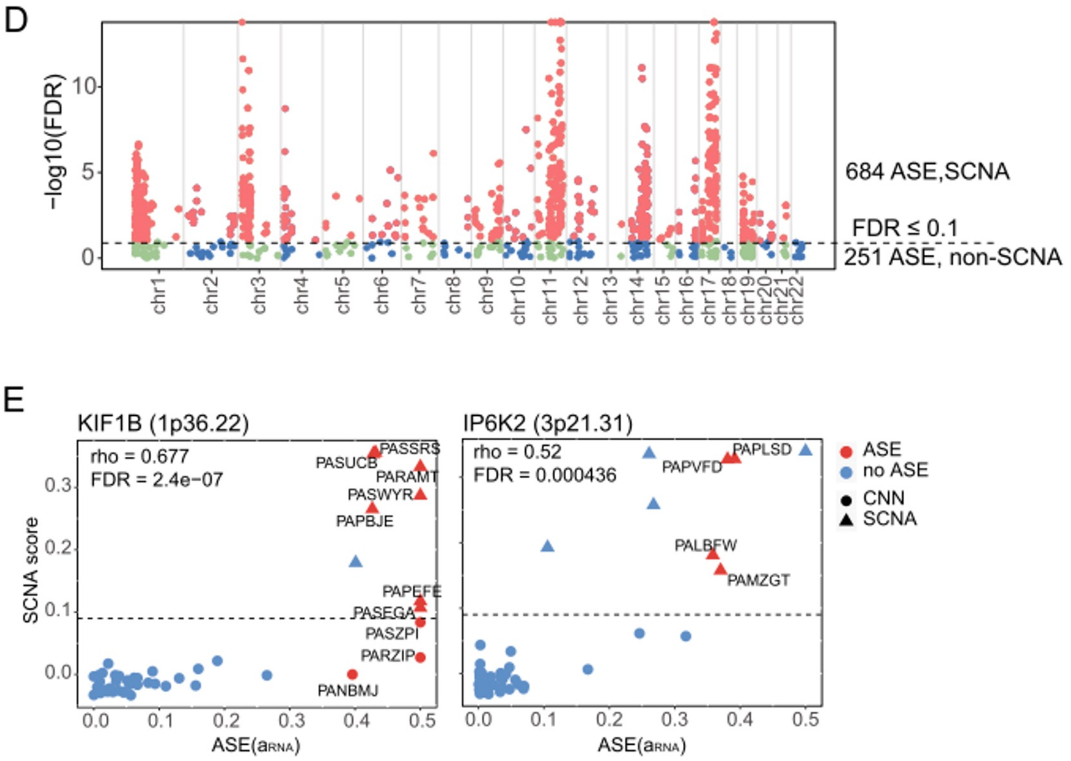
atumor - anormal (δa)

Circular binary segmentation

SCNA score

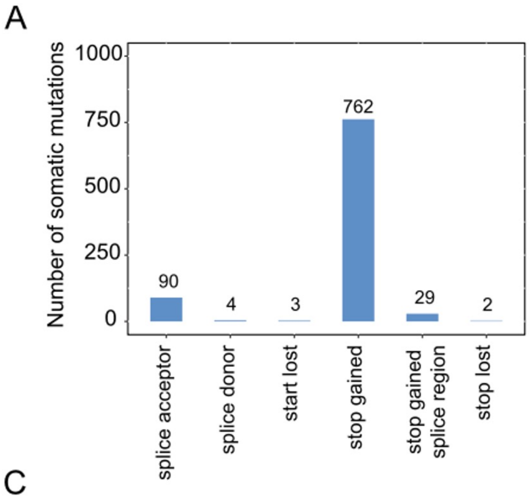
2. Comparison of DNA allele imbalance between samples

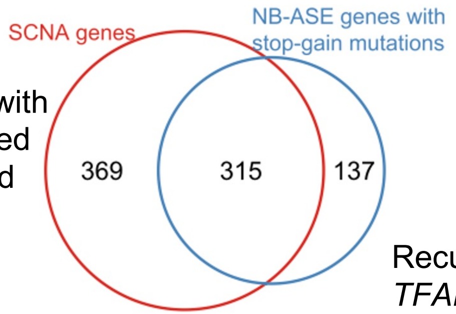
Difference between tumor and normal



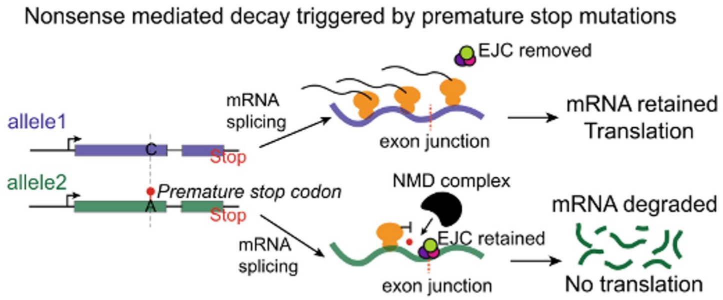
PMID: 35246212

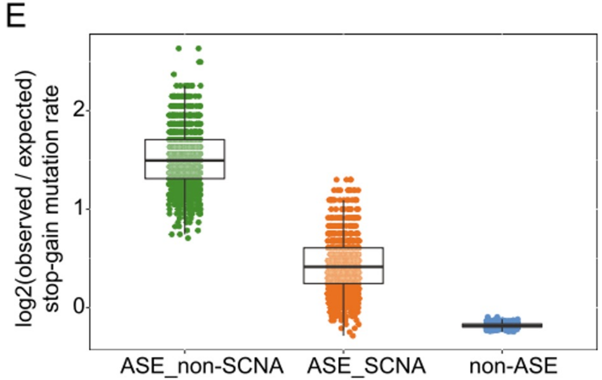
RNA allele imbalance



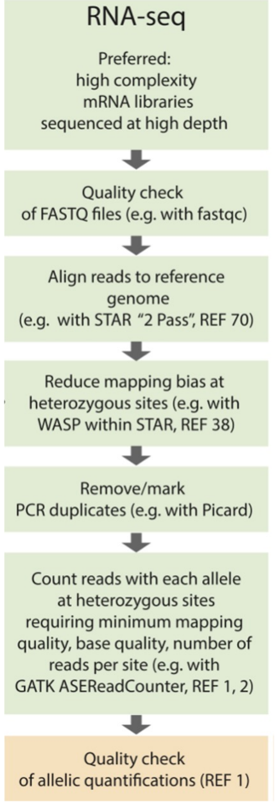


Genes with correlated ASE and SCNA





Recurrent ASE outside of common SCNA regions, including *TFAP2B* and *PTPRH*, both low expression in stage 4 disease and evidence for tumor suppressor activity.



RNA-seq

Preferred:
high complexity
mRNA libraries
sequenced at high depth

Quality check of FASTQ files (e.g. with fastqc)

Align reads to reference genome (e.g. with STAR "2 Pass", REF 70)

Reduce mapping bias at heterozygous sites (e.g. with WASP within STAR, REF 38)

Remove/mark PCR duplicates (e.g. with Picard)

Count reads with each allele at heterozygous sites requiring minimum mapping quality, base quality, number of reads per site (e.g. with GATK ASEReadCounter, REF 1, 2)

Quality check of allelic quantifications (REF 1)

PMID: 35246212

RNA Editing Detection

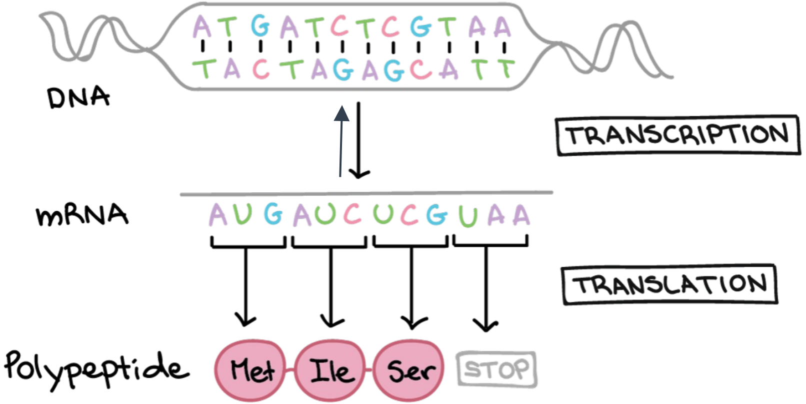
Jian Sang

Postdoc

Fellow

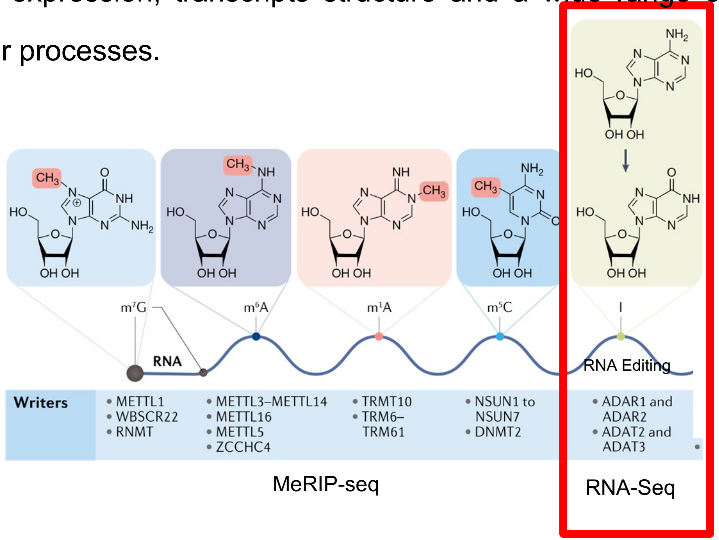
DCEG/ITEB

RNA modifications



- RNA modification is the chemical modifications that happen on RNAs after they are transcribed from DNA, before translated to protein.
- RNA modifications can change the sequence, structure, stability, and function of RNA molecules, thereby impacting genes expression, transcripts structure and a wide range of cellular processes.

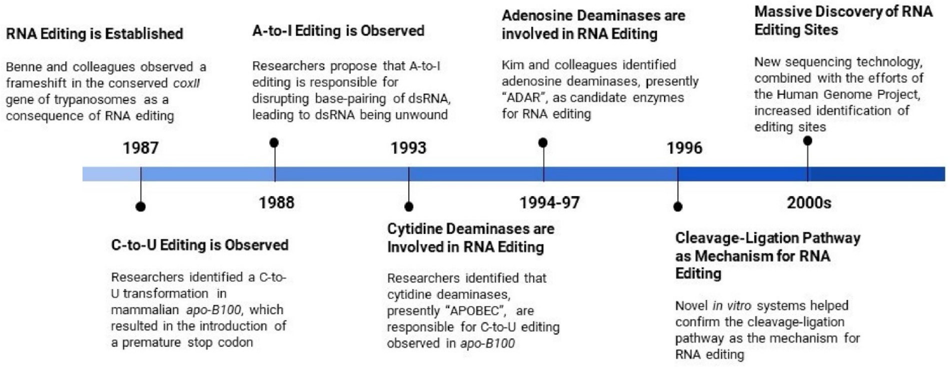
- RNA modification is one of the most important mechanisms for post-transcriptional regulation of human genome
- According to Modomics (https://genesilico.pl/modomics/), currently, more than 300 RNA modifications have been identified almost involved with all types of RNAs.



(PMID: 32300195)

Milestones in RNA Editing Discovery

- ❖ RNA editing is the RNA modification that directly changes RNA sequences without any corresponding mutation at DNA level.



RNA Editing is Established

Benne and colleagues observed a frameshift in the conserved *coxII* gene of trypanosomes as a consequence of RNA editing

1987

C-to-U Editing is Observed

Researchers identified a C-to-U transformation in mammalian *apo-B100*, which resulted in the introduction of a premature stop codon

1988

A-to-I Editing is Observed

Researchers propose that A-to-I editing is responsible for disrupting base-pairing of dsRNA, leading to dsRNA being unwound

1993

Cytidine Deaminases are Involved in RNA Editing

Researchers identified that cytidine deaminases, presently "APOBEC", are responsible for C-to-U editing observed in *apo-B100*

1994-97

Adenosine Deaminases are involved in RNA Editing

Kim and colleagues identified adenosine deaminases, presently "ADAR", as candidate enzymes for RNA editing

1996

Cleavage-Ligation Pathway as Mechanism for RNA Editing

Novel *in vitro* systems helped confirm the cleavage-ligation pathway as the mechanism for RNA editing

2000s

Massive Discovery of RNA Editing Sites

New sequencing technology, combined with the efforts of the Human Genome Project, increased identification of editing sites

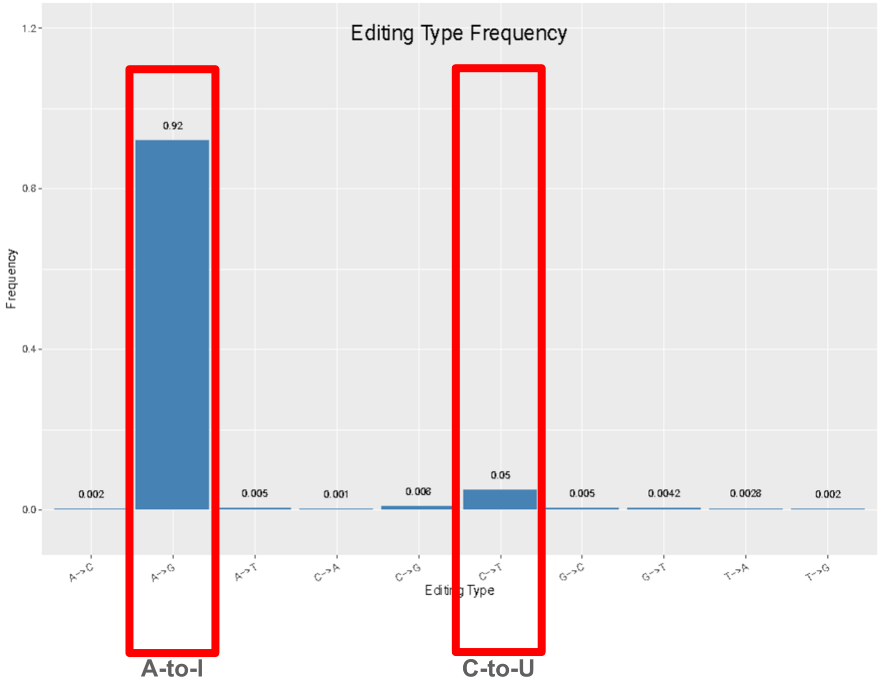
(PMID: 32650588)

RNA editing *V.S.* RNA variations

❖ RNA-editing ≠ RNA-based variations

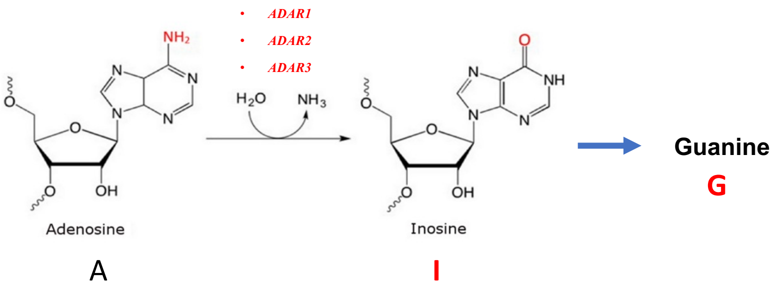
- Various kinds of substitutions can be detected at RNA Level using RNA-Seq, e.g. A>C, C>A, G>A, G>T... Most of them are originally from genomic Variation

- The A-to-I (A>G) and C-to-U (C>T) are only two existed RNA-Editing types that have been verified through molecular experiments.



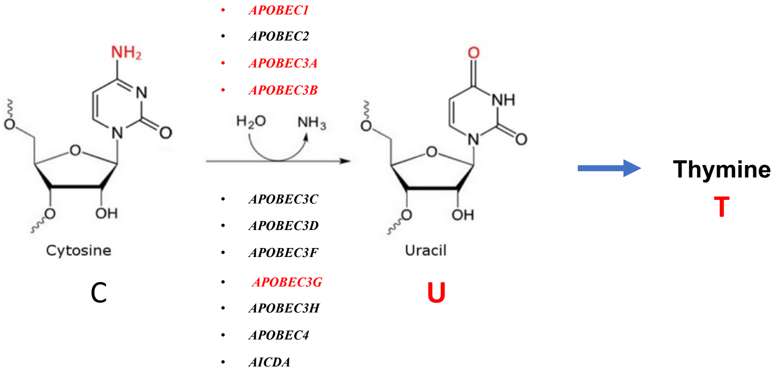
The schemes of RNA editing events

A-to-I

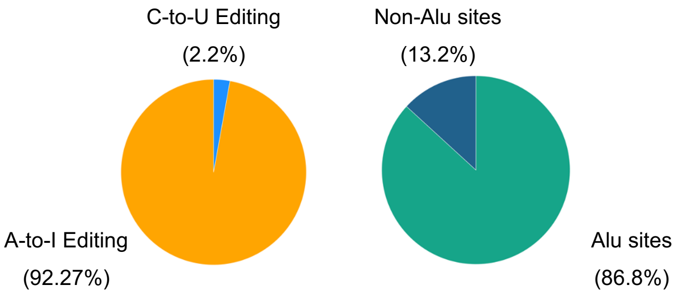


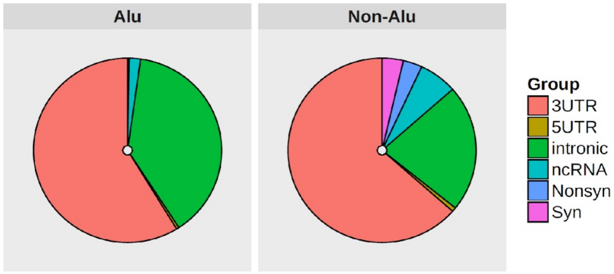
- *ADAR1*
- *ADAR2*
- *ADAR3*

C-to-U



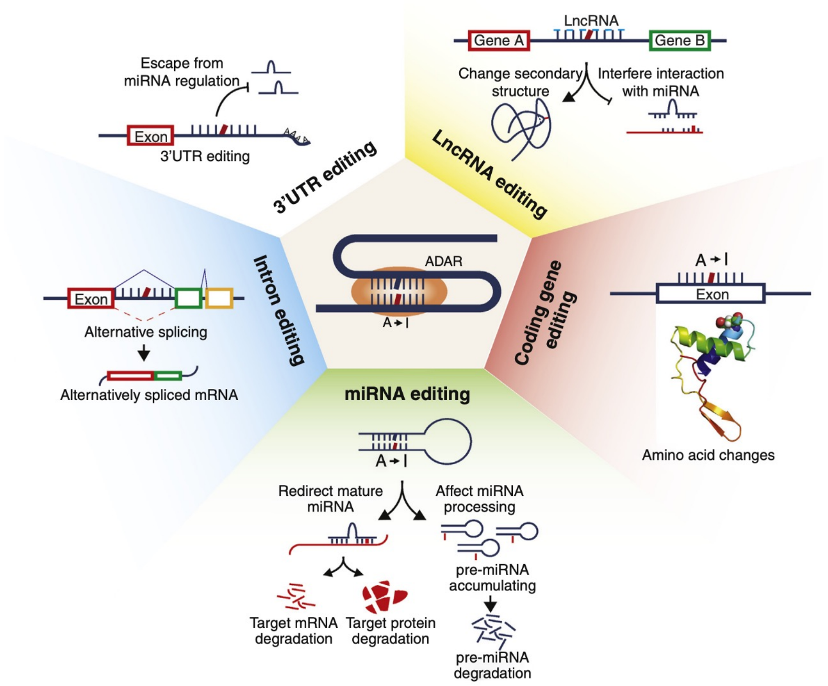
- *APOBEC1*
- *APOBEC2*
- *APOBEC3A*
- *APOBEC3B*
- *APOBEC3C*
- *APOBEC3D*
- *APOBEC3F*
- *APOBEC3G*
- *APOBEC3H*
- *APOBEC4*
- *AICDA*





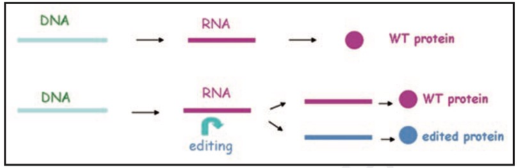
The Alu region is a family of highly repeated regions in human genomes, are responsible for regulation of tissue-specific genes

Molecular functions of RNA editing



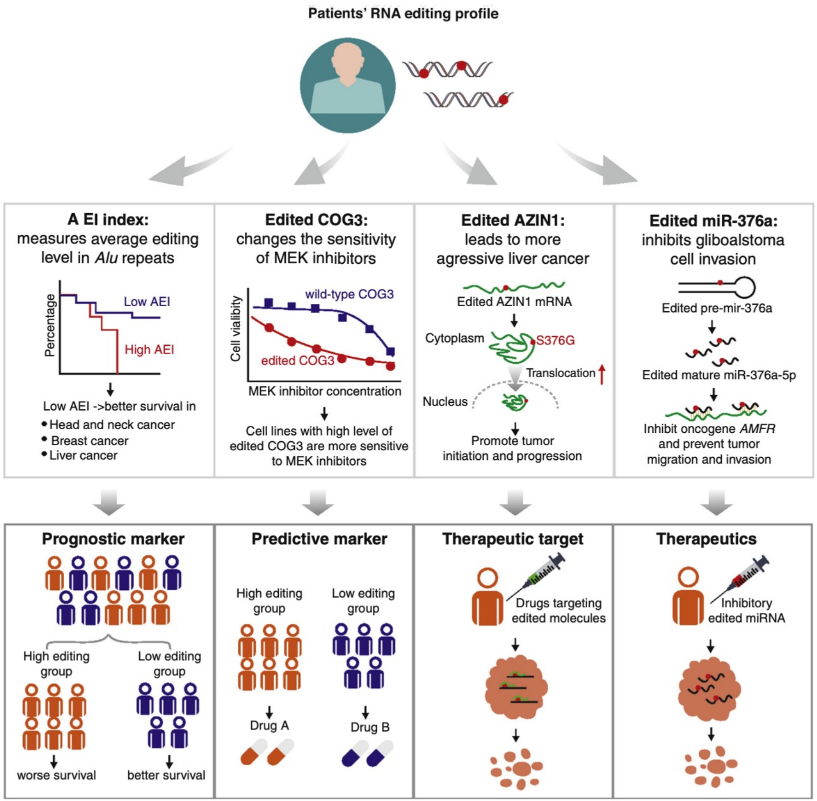
Current Opinion in Genetics & Development

- **Increase the diversity of cellular transcriptome**
- **Improve adaptation to internal and external changes**



(PMID: 29127844)

Clinical implication of RNA editing events in cancer study



Patients' RNA editing profile

A EI index: measures average editing level in *Alu* repeats

Low AEI ->better survival in

- Head and neck cancer
- Breast cancer
- Liver cancer

Edited COG3: changes the sensitivity of MEK inhibitors

Cell lines with high level of edited COG3 are more sensitive to MEK inhibitors

Edited AZIN1: leads to more agressive liver cancer

Promote tumor initiation and progression

Edited miR-376a: inhibits gliboalstoma cell invasion

Inhibit oncogene *AMFR* and prevent tumor migration and invasion

Prognostic marker

worse survival

better survival

Predictive marker

Therapeutic target

Therapeutics

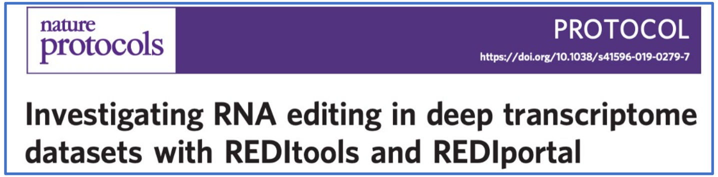
(PMID: 29127844)

Methods for detecting RNA editing events

- Detecting the differences between DNA-Seq and RNA-Seq results

  (1) DNA-Seq/RNA-Seq paired (accurate but time consuming & expensive)

  (2) RNA-Seq only, known RNA editing database is required (>20 times faster than **(1)**, most widely used strategy)

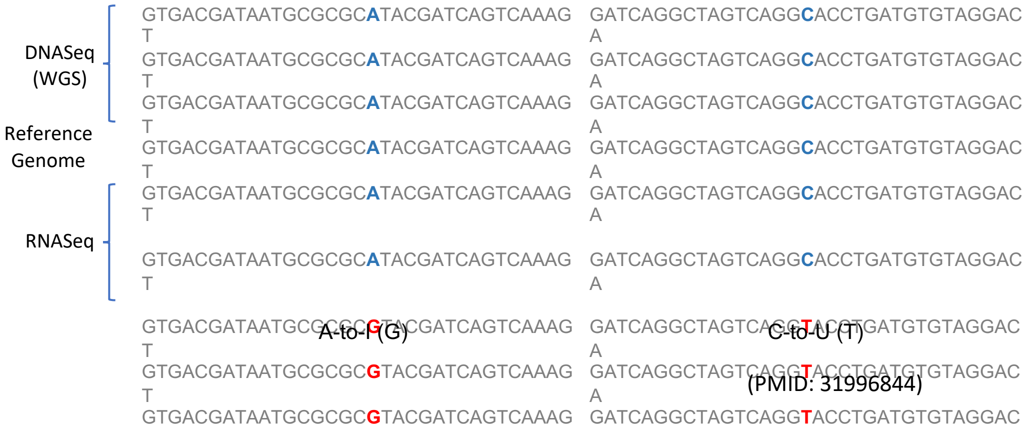


nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-019-0279-7

Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal



GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**A**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**C**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**G**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**T**ACCTGATGTGTAGGACA

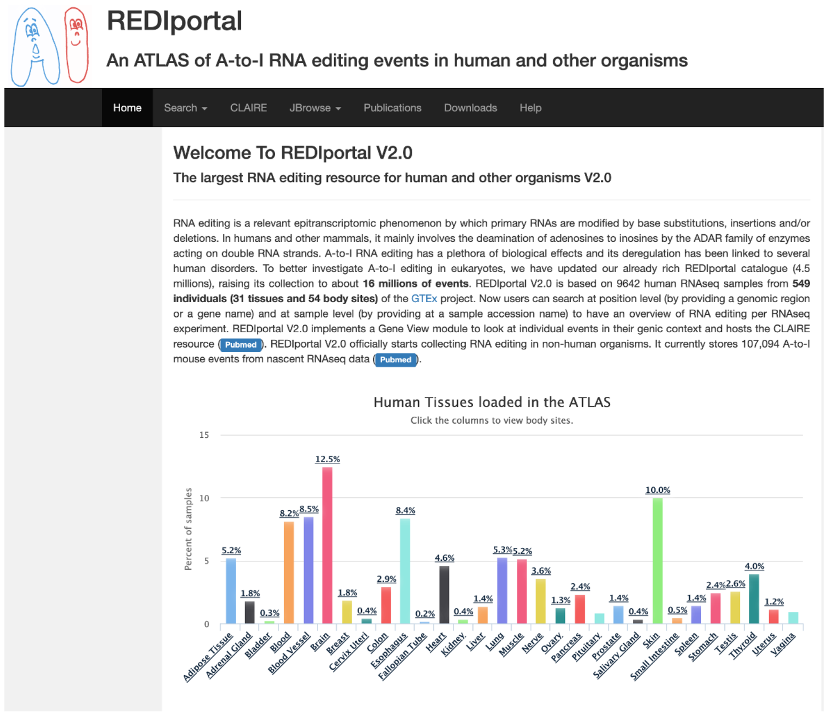
GTGACGATAATGCGCGC**G**TACGATCAGTCAAAGT GATCAGGCTAGTCAGG**T**ACCTGATGTGTAGGACA

GTGACGATAATGCGCGC**G**TACGATCAGTCAAAG GATCAGGCTAGTCAGG**T**ACCTGATGTGTAGGAC

(PMID: 31996844)

- Minimum mapping quality and base quality >20
- Minimum coverage for each site >10
- Exclude variation sites in multiple mapped and duplicated reads
- Exclude variation sites in highly repeated genomic regions
- Trim 6 bases up in every sequencing
- Maximum reads supporting variation for DNA and RNA across all samples should be at least 15
- Maximum editing level across all samples should be at lease 0.15
- Filtration of genomic mutation with dbSNP, gnomAD

Resources for RNA editing analysis



REDIportal

An ATLAS of A-to-I RNA editing events in human and other organisms

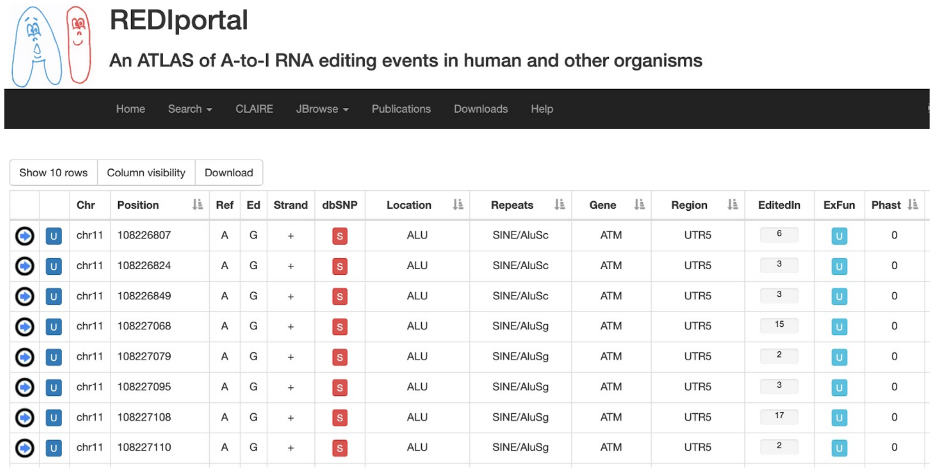
Welcome To REDIportal V2.0

The largest RNA editing resource for human and other organisms V2.0

RNA editing is a relevant epitranscriptomic phenomenon by which primary RNAs are modified by base substitutions, insertions and/or deletions. In humans and other mammals, it mainly involves the deamination of adenosines to inosines by the ADAR family of enzymes acting on double RNA strands. A-to-I RNA editing has a plethora of biological effects and its deregulation has been linked to several human disorders. To better investigate A-to-I editing in eukaryotes, we have updated our already rich REDIportal catalogue (4.5 millions), raising its collection to about **16 millions of events**. REDIportal V2.0 is based on 9642 human RNAseq samples from **549 individuals (31 tissues and 54 body sites)** of the GTEx project. Now users can search at position level (by providing a genomic region or a gene name) and at sample level (by providing at a sample accession name) to have an overview of RNA editing per RNAseq experiment. REDIportal V2.0 implements a Gene View module to look at individual events in their genic context and hosts the CLAIRE resource (Pubmed). REDIportal V2.0 officially starts collecting RNA editing in non-human organisms. It currently stores 107,094 A-to-I mouse events from nascent RNAseq data (Pubmed).

Human Tissues loaded in the ATLAS

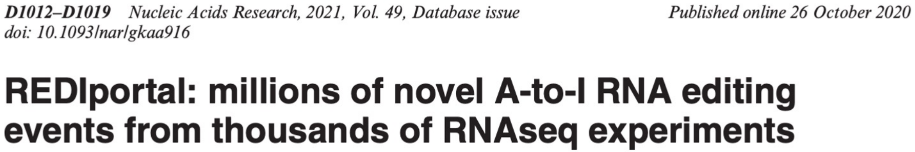
Click the columns to view body sites.



REDIportal

An ATLAS of A-to-I RNA editing events in human and other organisms

		Chr	Position	Ref	Ed	Strand	dbSNP	Location	Repeats	Gene	Region	EditedIn	ExFun	Phast
●	U	chr11	108226807	A	G	+	S	ALU	SINE/AluSc	ATM	UTR5	6	U	0
●	U	chr11	108226824	A	G	+	S	ALU	SINE/AluSc	ATM	UTR5	3	U	0
●	U	chr11	108226849	A	G	+	S	ALU	SINE/AluSc	ATM	UTR5	3	U	0
●	U	chr11	108227068	A	G	+	S	ALU	SINE/AluSg	ATM	UTR5	15	U	0
●	U	chr11	108227079	A	G	+	S	ALU	SINE/AluSg	ATM	UTR5	2	U	0
●	U	chr11	108227095	A	G	+	S	ALU	SINE/AluSg	ATM	UTR5	3	U	0
●	U	chr11	108227108	A	G	+	S	ALU	SINE/AluSg	ATM	UTR5	17	U	0
●	U	chr11	108227110	A	G	+	S	ALU	SINE/AluSg	ATM	UTR5	2	U	0



D1012–D1019 Nucleic Acids Research, 2021, Vol. 49, Database issue *Published online 26 October 2020*
doi: 10.1093/nar/gkaa916

REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments

THANKS FOR YOUR ATTENTION!
Questions?

Next: Practical session 10 (10:45am)

Lecture 12: Data visualization
May 15 (Monday) 9:30am – 12pm, Room TE406/408/410