Resources and computational analyses
TCGA, PCA, GDAN, GDC, regulons

Gordon Robertson
Canada's Michael Smith Genome Sciences Centre
British Columbia Cancer Agency
Vancouver BC Canada

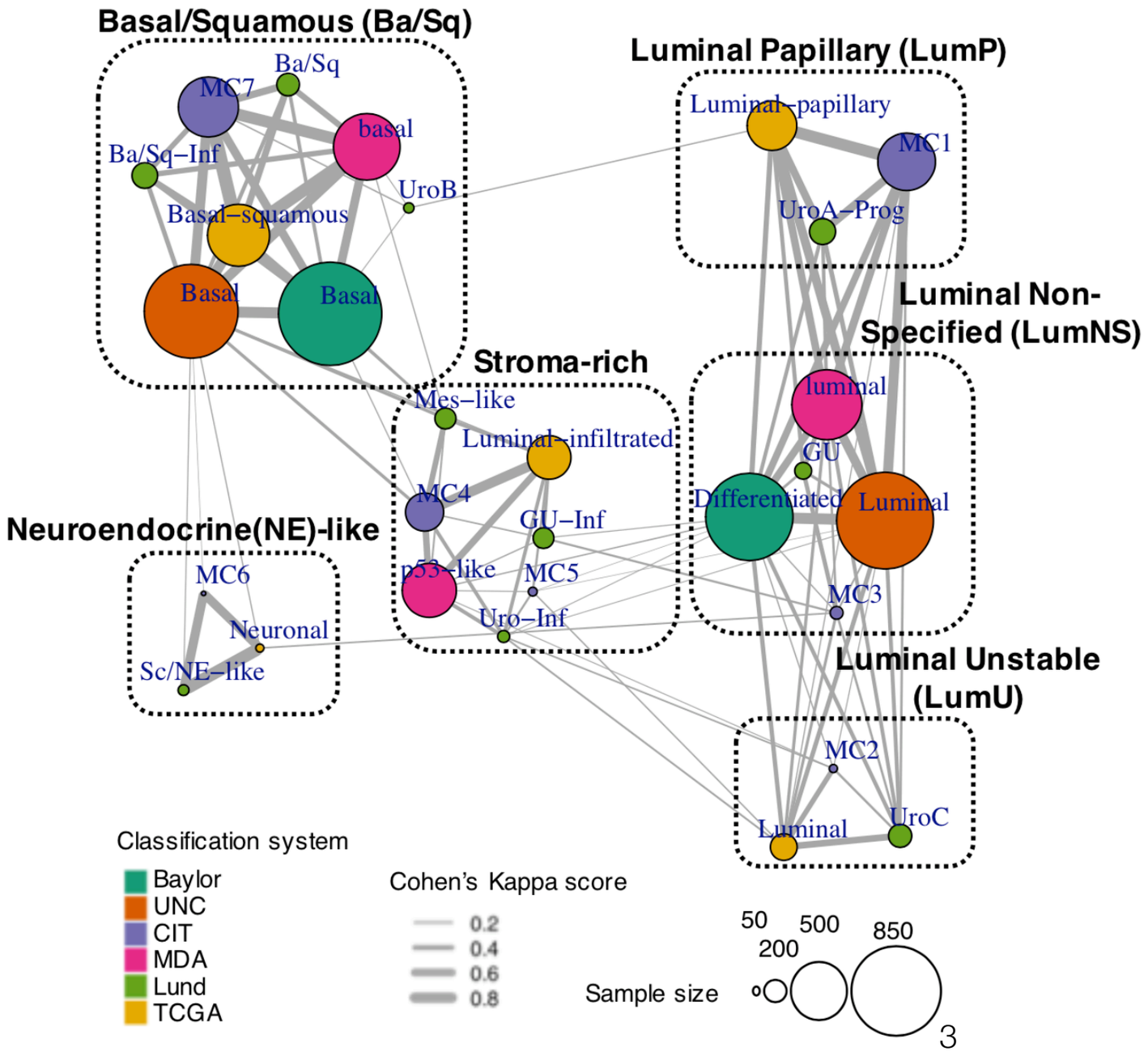
Saturday, 10 August 2019, 08h00 - 10h00

BCAN Think Tank 2019, Washington DC

- Consensus MIBC subtypes
- TCGA
  - miRNA-seq data generating process
- TCGA BLCA
  - LncRNA-based MIBC subtypes
- PanCancer Atlas resources
  - Publications, clinical / batch-corrected expression data
- The GDAN and the GDC, interim projects
  - GDC-QC: hg19 vs. hg38 data for all TCGA platforms
    - miRNA-seq data
- Regulon analysis
  - Two method publications, with case studies

The consensus molecular classification of muscle-invasive bladder cancer

Aurélie Kamoun, Aurélien de Reyniès, Yves Allory, Gottfrid Sjödahl, A. Gordon Robertson, Roland Seiler, Katherine A. Hoadley, Hikmat Al-Ahmadie, Woonyoung Choi, Clarice S. Groeneveld, Mauro A. A. Castro, Jacqueline Fontugne, Pontus Eriksson, Qianxing Mo, Alexandre Zlotta, Arndt Hartmann, Colin P. Dinney, Joaquim Bellmunt, Thomas Powles, Núria Malats, Keith S. Chan, William Y. Kim19, David J. McConkey, Peter C. Black, Lars Dyrskjøt, Mattias Höglund, Seth P. Lerner, Francisco X. Real, François Radvanyi, The Bladder Cancer Molecular Taxonomy Group



Input systems		Consensus subtypes
Baylor		LumP
UNC		LumU
CIT	→	LumNS
MDA		Ba/Sq
Lund		Stroma-rich
TCGA		NE-like

- bioRxiv preprint first posted online Dec. 7, 2018
- Manuscript now in review
- R packages

TCGA: NIH-funded molecular profiling consortium

- >11k samples, 33 cancers, ~10 years
- Clinical data
- DNA/methylation, RNA, protein, by sequencing and arrays

Influence medical practice

Enable a global research community

Each project: an integrative ‘marker’ paper

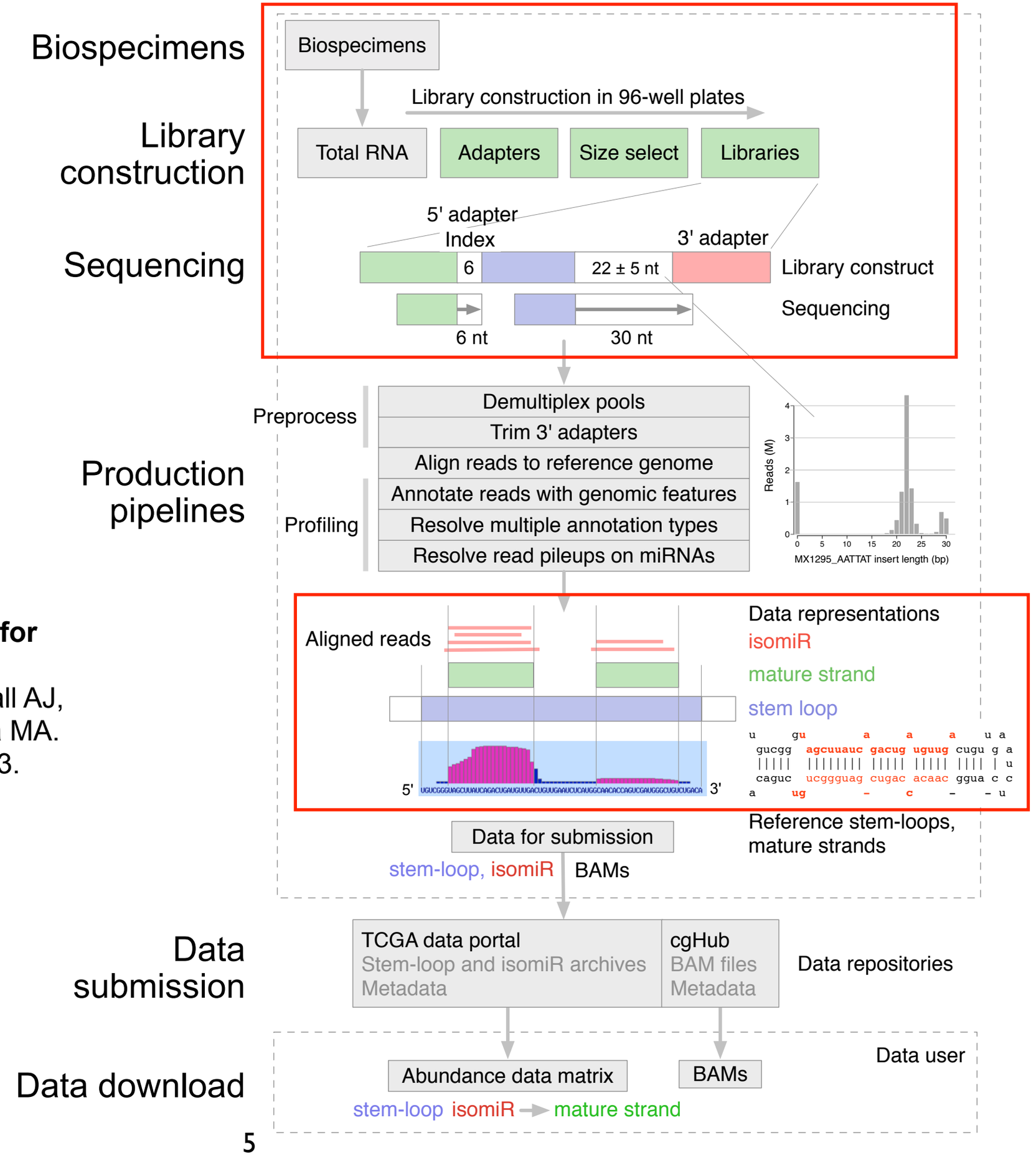
- 5+ genomic platforms in ~3000 words
- No functional validation
- AWGs, 35-person telecons
- Publication context
- Listening to medical

http://crosstalk.cell.com/blog/a-resource-10-years-in-the-making

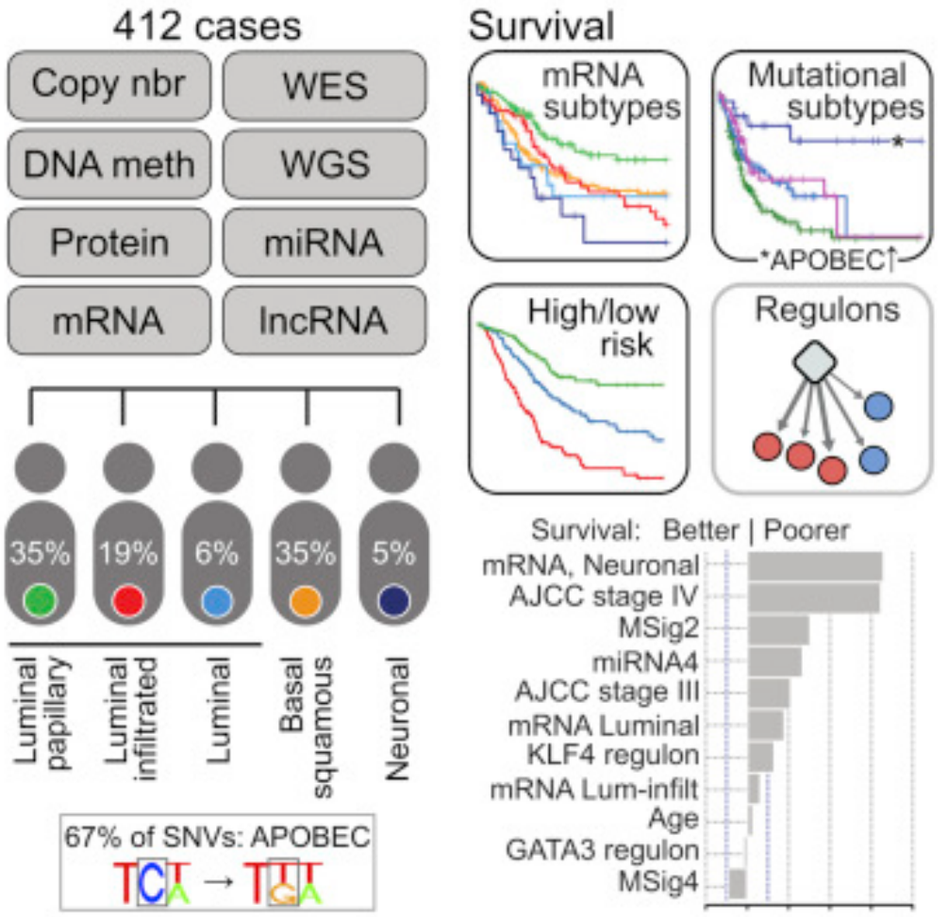
An NIH contract

~11k miRNA-seq data sets

Large-scale profiling of microRNAs for The Cancer Genome Atlas.
Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, Ma Y, Jones S, Marra MA.
Nucleic Acids Res. 2016 Jan 8;44(1):e3.
PMID: 26271990



TCGA BLCA: Data types and analyses



Cell. October 2017

- Clinical data
- Mutations and short/long indels, mutational signatures, APOBEC: WES
- Copy number: Affy SNP6 arrays
- DNA methylation: Illumina 450k arrays
- Messenger RNA-seq, poly(A)-selected
- microRNA-seq
- RPPA: ~200 total and phosphorylated proteins

- Subtypes: mSig, mRNA, lncRNA, miRNA; COCA
- Immune infiltration
- EMT scores

- Microbes: screening, genomic integration
- lncRNAs from mRNA-seq reads, Ensembl v82
- Regulon analysis, 23 BCa-associated regulators
- Multivariate survival analysis
  ~200 features, ~100 univariate, 15 multivariate

LncRNA expression is more specific

Ewan A Gibb

Genome Med. 2015
PMID: 25821520

RNA-seq reads
STAR, Cufflinks
Ensembl v82 (Sept '15)

BLCA
CHOL
MESO
PAAD
Pan-Kidney
UVM

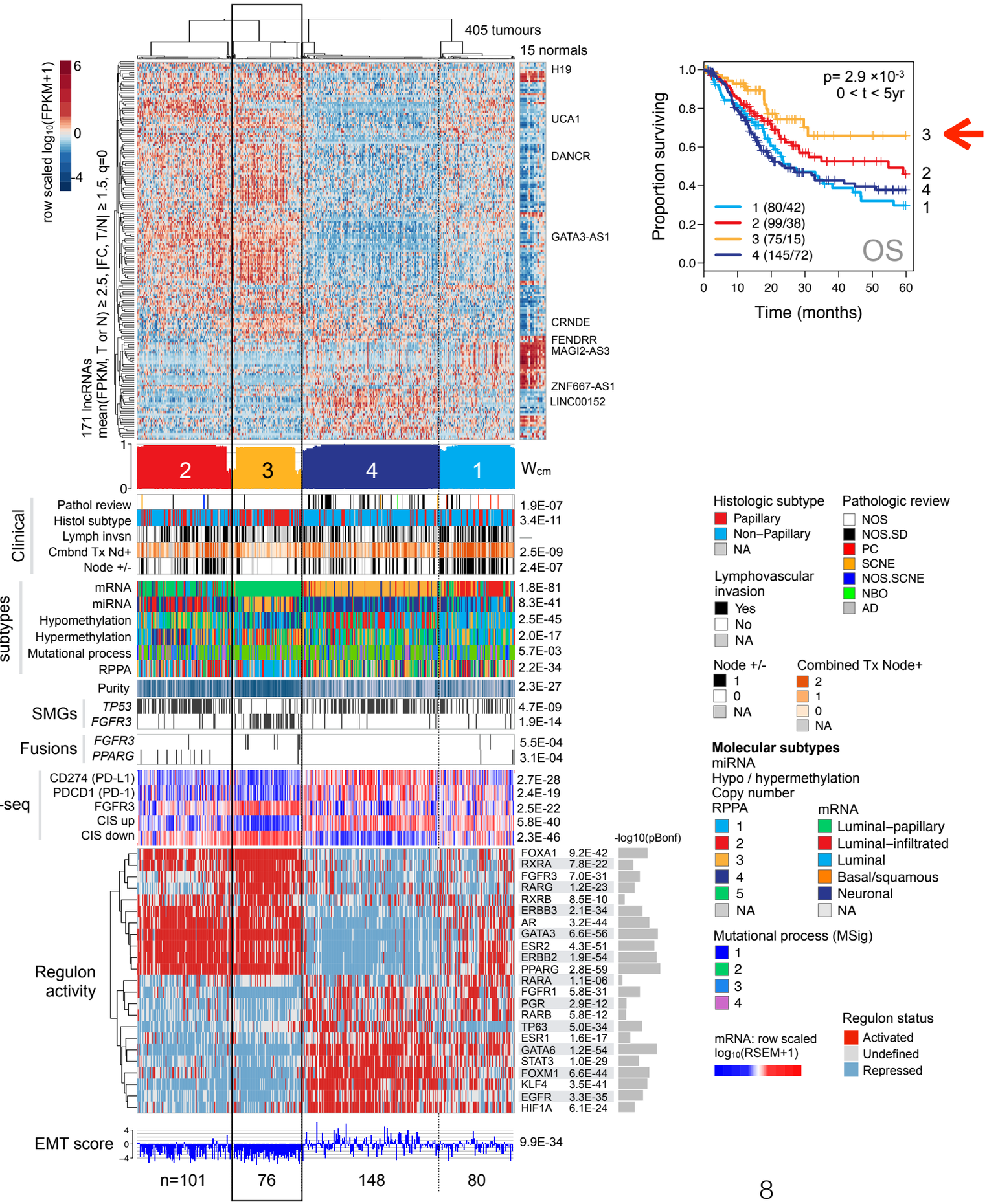
Expression Specificity of Disease-Associated lncRNAs: Toward Personalized Medicine.

Nguyen Q, Carninci P.
Curr Top Microbiol Immunol. 2016;394:237-58.
PMID: 26318140

Abstract Long noncoding RNAs (lncRNAs) perform diverse regulatory functions in transcription, translation, chromatin modification, and cellular organization. Misregulation of lncRNAs is found linked to various human diseases. Compared to protein-coding RNAs, lncRNAs are more specific to organs, tissues, cell types, developmental stages, and disease conditions, making them promising candidates as diagnostic and prognostic biomarkers and as gene therapy targets. The functional annotation of mammalian genome (FANTOM) consortium utilizes cap analysis of gene expression (CAGE) method to quantify genome-wide activities of promoters and enhancers of coding and noncoding RNAs across a large collection of human

Misregulation of lncRNAs is found linked to various human diseases. Compared to protein-coding RNAs, lncRNAs are more specific to organs, tissues, cell types, developmental stages, and disease conditions, making them promising candidates as diagnostic and prognostic biomarkers and as gene therapy targets. The functional

human diseases. In this chapter, we discuss lncRNA expression specificity, review diverse functions of disease-associated lncRNAs, and present perspectives on their potential therapeutic applications for personalized medicine. The future development of lncRNA applications relies on technologies to identify and validate their functions, structures, and mechanisms. Comprehensive understanding of genome-wide interaction networks of lncRNAs with proteins, chromatins, and other RNAs in regulating cellular processes will allow personalized medicine to use lncRNAs as highly specific biomarkers in diagnosis, prognosis, and therapeutic targets.



lncRNA consensus subtypes

Ewan A Gibb

RNA-seq reads
STAR, Cufflinks
Ensembl v82 (Sept ’15)
lncRNA FPKMs

Harmonized GDC hg38 mRNA-seq data use GENCODE v22 annotations, so include all biotypes

Subtype with favourable outcomes confirmed in an independent cohort (EAG)

The PanCancer Atlas

TCGA
33 cancer projects

PanCancer Atlas
Issues across cancers

CCG GDAN
GDC

https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html

PanCancer Atlas publications

https://cancergenome.nih.gov/publications
https://gdc.cancer.gov/node/977

- Cell-of-Origin Patterns Dominate the **Molecular Classification** of 10,000 Tumors from 33 Types of Cancer.
- An Integrated TCGA Pan-Cancer **Clinical Data** Resource to Drive High-Quality Survival Outcome Analytics.
- The **Immune** Landscape of Cancer.
- Oncogenic **Signaling Pathways** in The Cancer Genome Atlas.
- Comprehensive Characterization of Cancer **Driver Genes and Mutations**; Pathogenic **Germline Variants** in 10,389 Adult Cancers.
- Genomic and Molecular Landscape of **DNA Damage Repair Deficiency** Across The Cancer Genome Atlas.
- Machine Learning Identifies **Stemness Features** Associated with Oncogenic Dedifferentiation.
- Genomic and Functional Approaches to Understanding **Cancer Aneuploidy**.
- Comprehensive Analysis of **Alternative Splicing** Across Tumors from 8,705 Patients; Somatic **Mutational** Landscape of **Splicing Factor** Genes and Their Functional Consequences across 33 Cancer Types; Systematic Analysis of **Splice-Site-Creating Mutations** in Cancer.
- Molecular Characterization and Clinical Relevance of **Metabolic Expression Subtypes** in Human Cancers.
- **Driver Fusions** and Their Implications in the Development and Treatment of Human Cancers.
- Integrated Genomic Analysis of the **Ubiquitin Pathway** Across Cancer Types.
- Genomic, Pathway Network, and Immunologic Features Distinguishing **Squamous Carcinomas**.
- Machine Learning Detects Pan-cancer **Ras Pathway Activation** in The Cancer Genome Atlas.
- Pan-cancer Alterations of the **MYC Oncogene** and Its Proximal Network across the Cancer Genome Atlas.
- …

PanCancer Atlas data

https://gdc.cancer.gov/node/977

- **Sample Annotations**
  - Analyte level annotations - merged_sample_quality_annotations.tsv
- **Mutation Files**
  - Controlled mutation annotation file - mc3.v0.2.8.CONTROLLED.maf.gz
  - Public mutation annotation file - mc3.v0.2.8.PUBLIC.maf.gz
  - ABSOLUTE-annotated MAF file - TCGA_consolidated.abs_mafs_truncated.fixed.txt.gz
  - Molecular Signatures - tcga_pancancer_082115.vep.filter_whitelisted.context.maf.signatures.txt
  - Mutation Load - mutation-load-updated.txt
- **DNA copy number Files**
  - SNP6 whitelisted copy number segments file - broad.mit.edu_PANCAN_Genome_Wide_SNP_6_whitelisted.seg
  - GISTIC2.0 all_thresholded.by_genes file - all_thresholded.by_genes_whitelisted.tsv
  - GISTIC2.0 all_data_by_genes file - all_data_by_genes_whitelisted.tsv
  - ISAR-corrected SNP6 whitelisted copy number segments file - ISAR_corrected.PANCAN_Genome_Wide_SNP_6_whitelisted.seg
  - gzipped ISAR-corrected GISTIC2.0 all_thresholded.by_genes file -

- **RNA and Protein Files**
  - RNA batch corrected matrix - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv
  - miRNA batch corrected matrix - pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv
  - miRNA sample information - PanCanAtlas_miRNA_sample_information_list.txt
  - RPPA batch corrected matrix - TCGA-RPPA-pancan-clean.txt

usc.edu_PANCAN_HumanMethylation450.betaValue_whitelisted.tsv
- Leukocyte score - TCGA_all_leuk_estimate.masked.20170107.tsv
- **RNA and Protein Files**
  - RNA batch corrected matrix - EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv
  - miRNA batch corrected matrix - pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv
  - miRNA sample information - PanCanAtlas_miRNA_sample_information_list.txt
  - RPPA batch corrected matrix - TCGA-RPPA-pancan-clean.txt
- **Other Files**
  - PARADIGM Pathway Inference Matrix - merge_merged_reals.tar.gz
  - DNA methylation Stemness signatures (lists of probes and genes) - DNAmethylation and RNAexpression Stemness Signatures.xlsx
  - DNA methylation and RNA stemness scores - SupplementalTable_S1.xlsx
  - iCluster input features - pancan33.iCluster.features.csv

The CCG GDAN and the GDC

TCGA
33 cancer projects

→

PanCancer Atlas
Issues across cancers

→

CCG GDAN
GDC

→

Interim projects GDC-QC
…

NCI's Genomic Data Commons (GDC)

https://gdc.cancer.gov/

GRCh37 and GRCh38

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

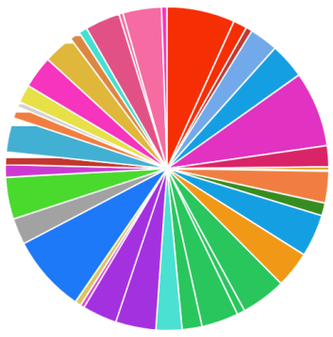
CCG Web Site | Contact Us | Launch Data Portal | GDC Apps

Search this website

About the GDC | About the Data | Analyze Data | Access Data | Submit Data | For Developers | Support | News

The Next Generation Cancer Knowledge Network

Case Distribution by Disease Type



Data Availability Summary

Programs	2
Projects	39
Disease Types	38
Cases	14,551

Launch Data Portal

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET).

More about the GDC

High Quality Data Sharing Enables Precision Medicine

The GDC obtains validated datasets from NCI programs in which the strategies for tissue collection couples quantity with high quality.

The GDC encourages data sharing in support of precision medicine. Tools are provided to guide data submissions by researchers and institutions.

Learn more about submitting data

Analyze Data

The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

More about Analyzing Data

Access Data

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. The GDC also provides a **GDC Data Transfer Tool** and a **GDC API** for programmatic access.

More about Accessing Data

Submit Data

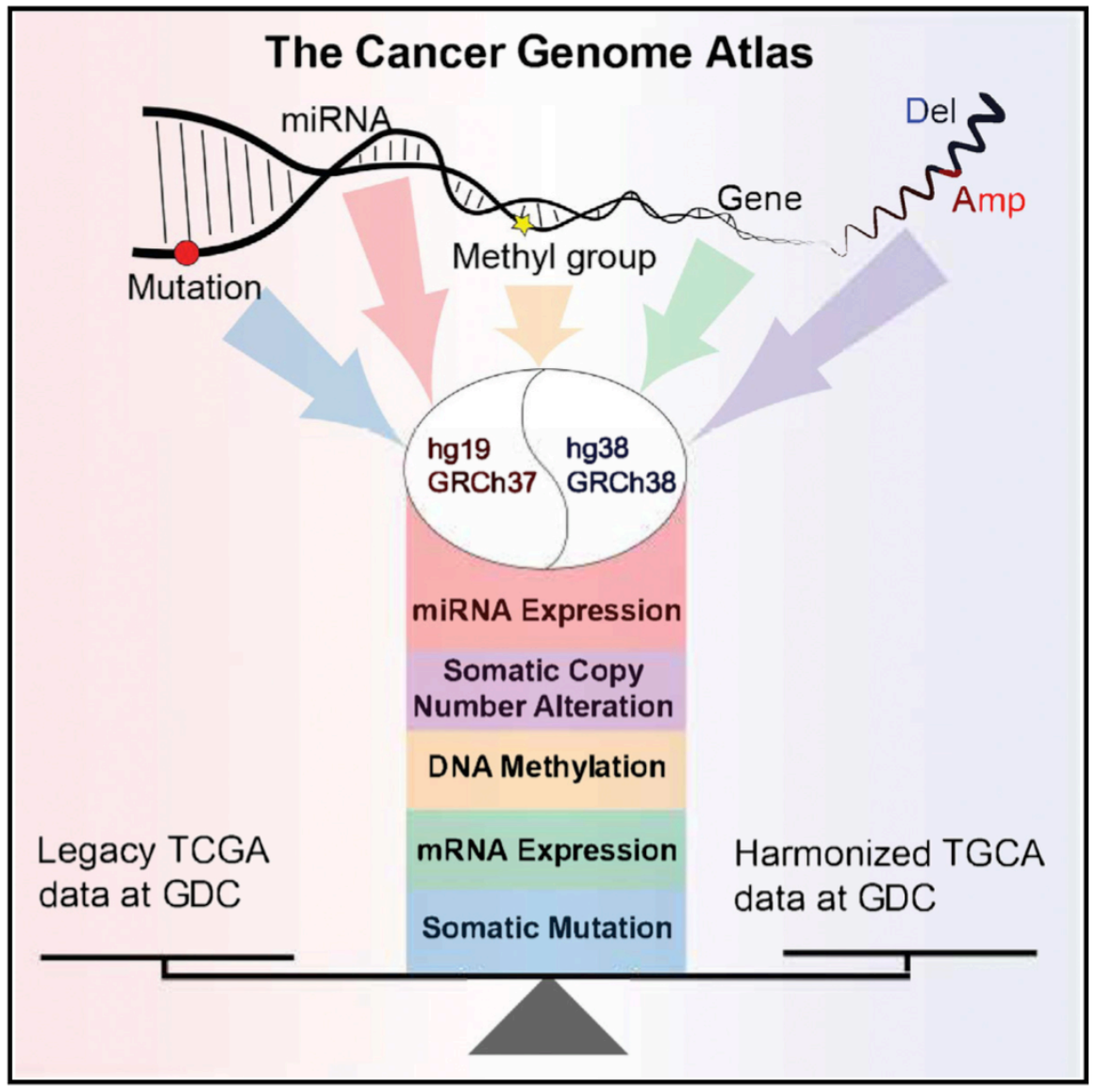
The GDC provides tools to guide data submission including the **GDC Data Submission Portal**, a web-based tool for submitting clinical, biospecimen and small volumes of molecular data as well as the **GDC Data Transfer Tool**, a client-based tool for submitting large, high volume molecular data. A secure **GDC API** is also available for batch data submissions.

More about Submitting Data

Cell Systems

Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data

Graphical Abstract



Authors

Galen F. Gao, Joel S. Parker, Sheila M. Reynolds, ..., The Genomic Data Analysis Network, Han Liang, Michael S. Noble

Correspondence

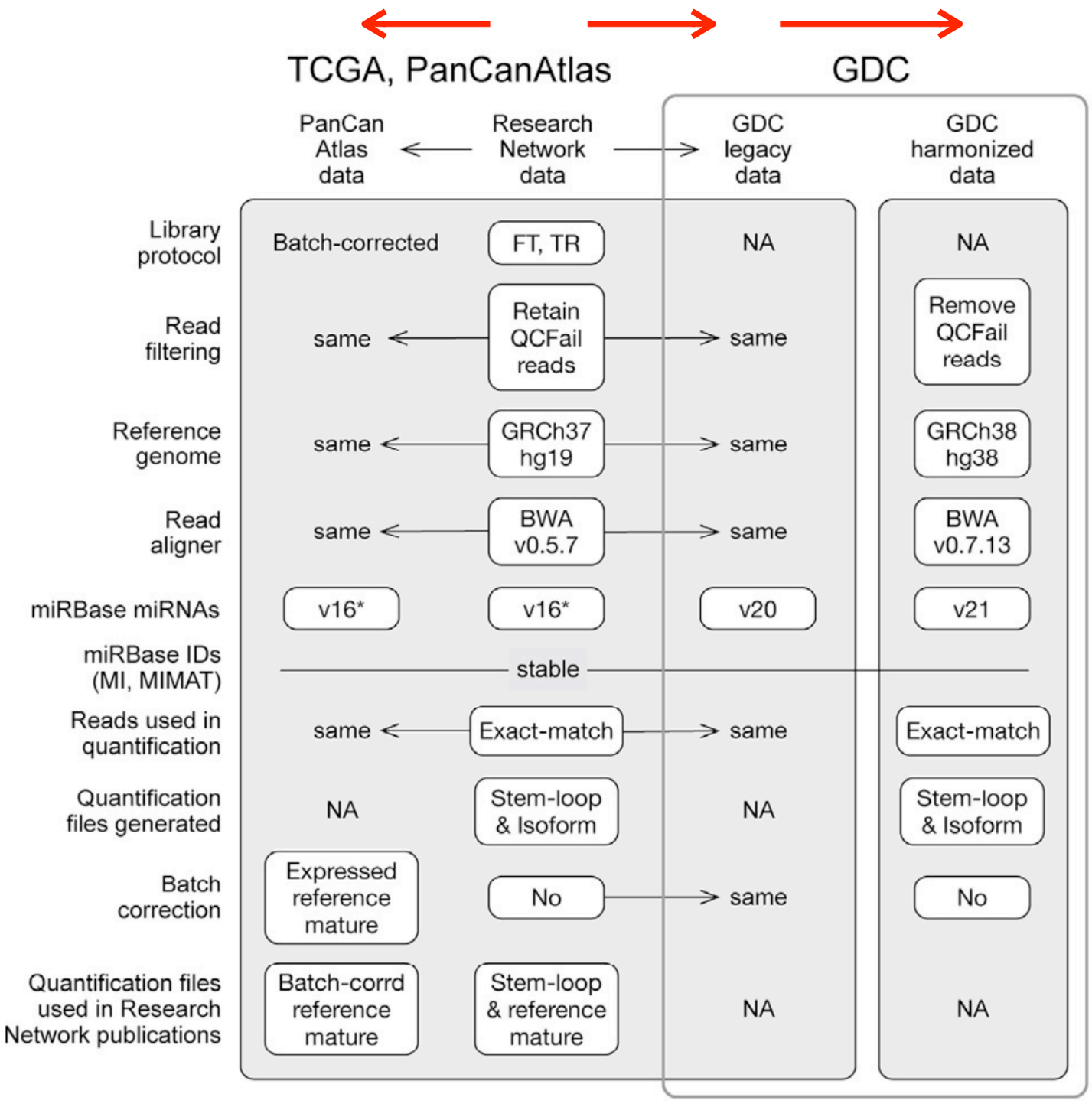
hliang1@mdanderson.org (H.L.), mnoble@cogenimmune.com (M.S.N.)

In Brief

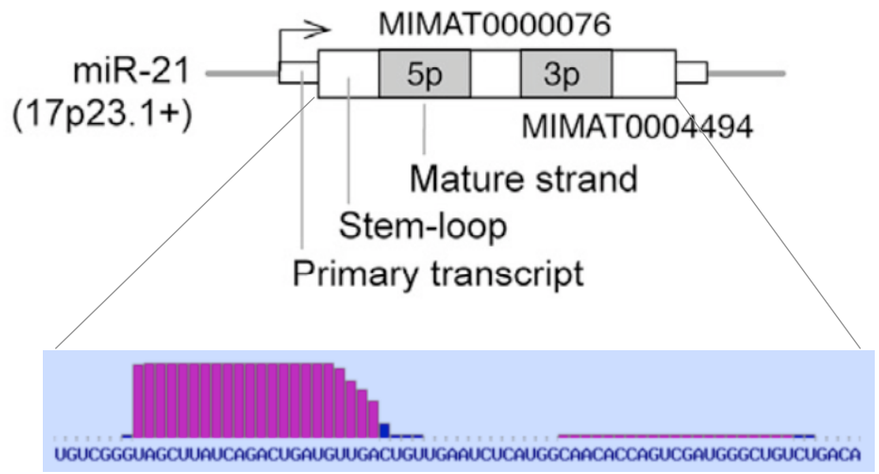
Gao et al. performed a systematic analysis of the effects of synchronizing the large-scale, widely used, multi-omic dataset of The Cancer Genome Atlas to the current human reference genome. For each of the five molecular data platforms assessed, they demonstrated a very high concordance between the 'legacy' GRCh37 (hg19) TCGA data and its GRCh38 (hg38) version as 'harmonized' by the Genomic Data Commons.

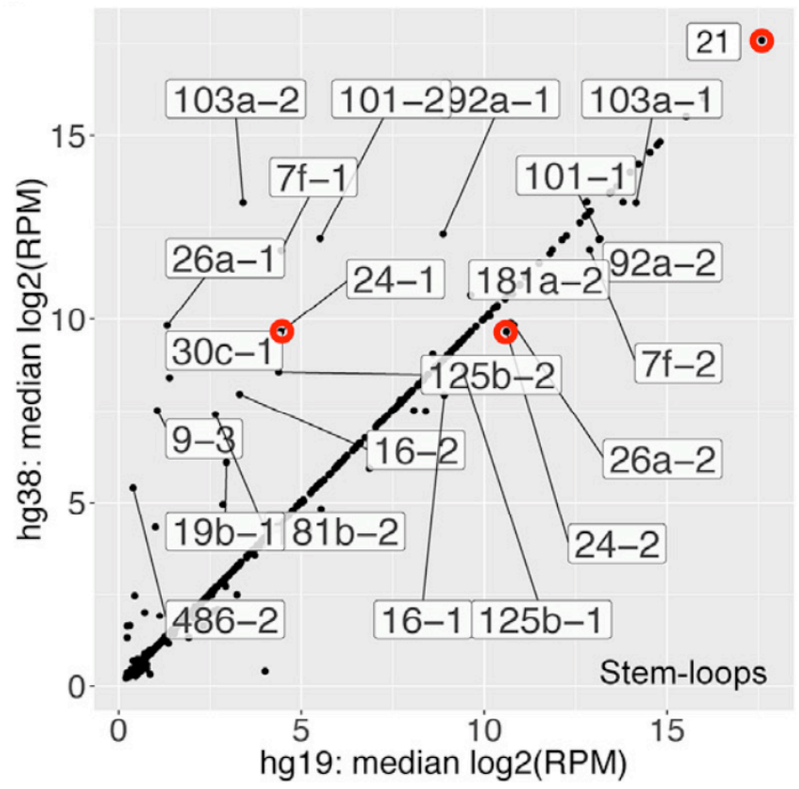
Gao et al., 2019, Cell Systems 9, 24–34
July 24, 2019
https://doi.org/10.1016/j.cels.2019.06.006

Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data



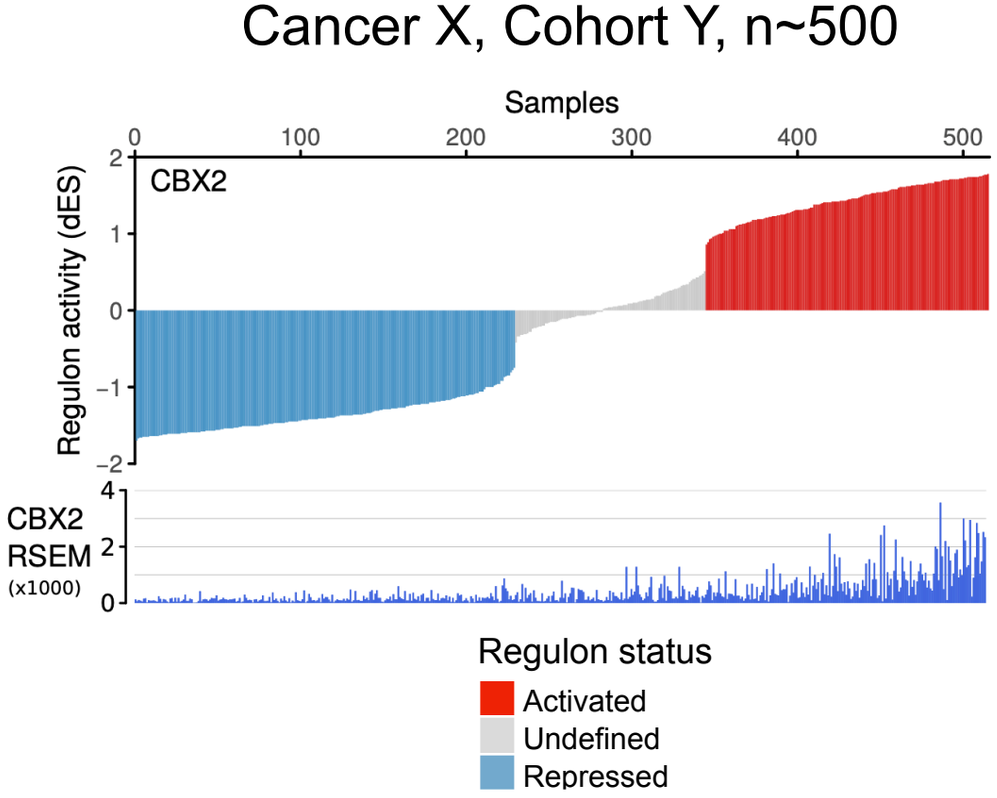
Cell Systems 2019, 9:24-34





Regulon analysis

Mauro A.A. Castro
Clarice S. Groeneveld
Vinicius S. Chagas
Bioinformatics and Systems Biology Laboratory
Federal University of Paraná Polytechnic Center, Curitiba, Brazil



A regulon activity profile (RAP) across a cohort is a nonlinear transformation of gene expression data. By reporting on genes that respond to a regulator, RAPs offer a functional readout that informs on biological state.

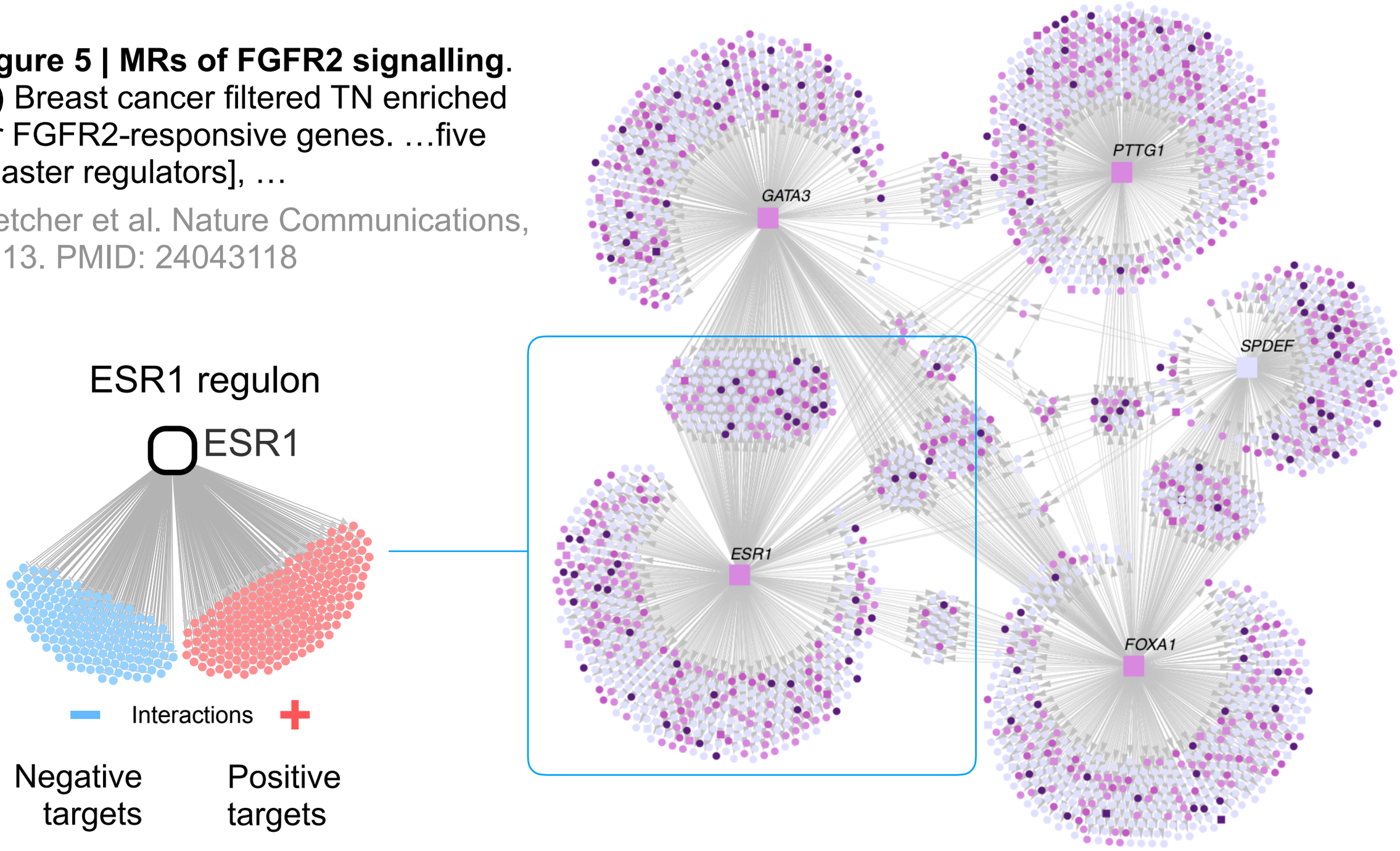
Regulon analysis

Mauro Castro
Clarice Groeneveld
Vinicius Chagas

Transcriptional network → regulons → a regulator's target genes

Figure 5 | MRs of FGFR2 signalling. **(c)** Breast cancer filtered TN enriched for FGFR2-responsive genes. …five [master regulators], …

Fletcher et al. Nature Communications, 2013. PMID: 24043118



Regulon activity can sort a cohort, regulon status can stratify

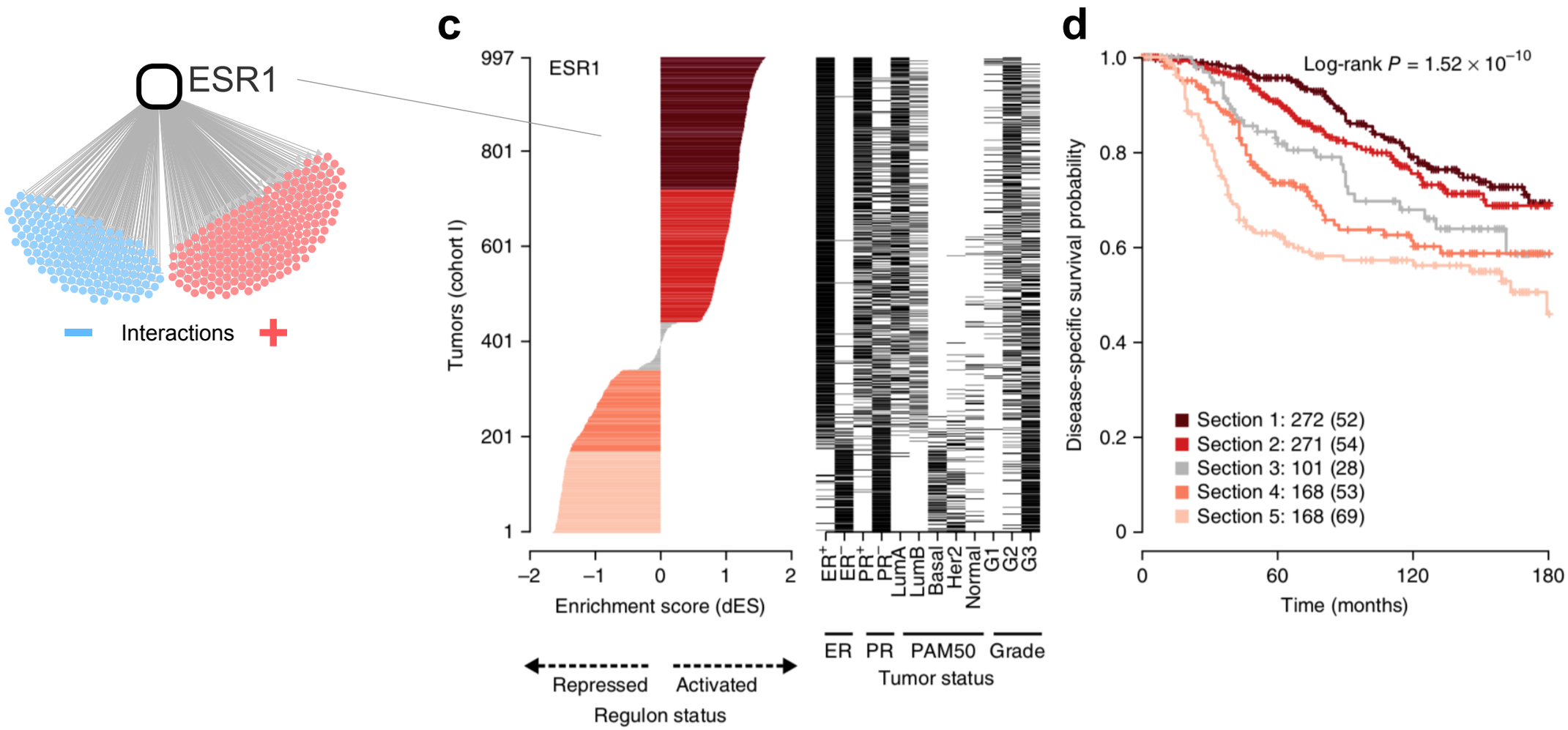
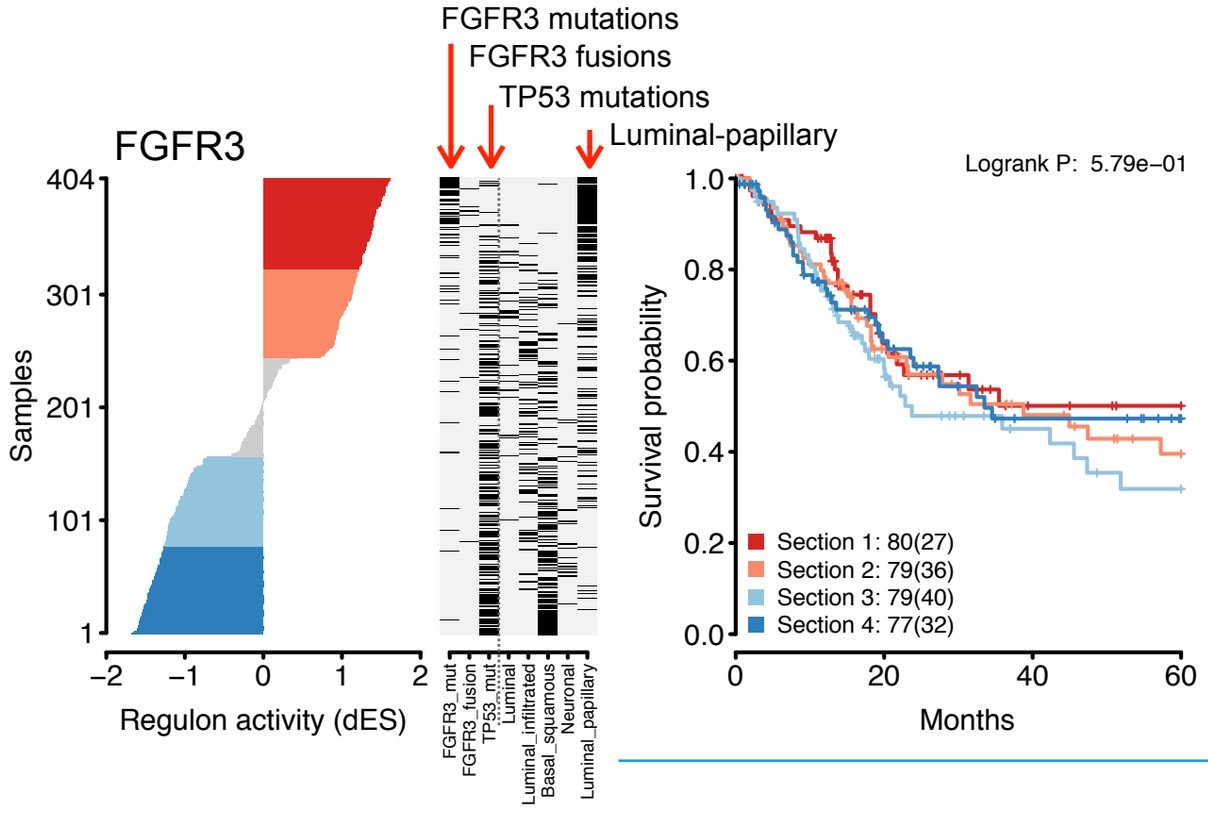


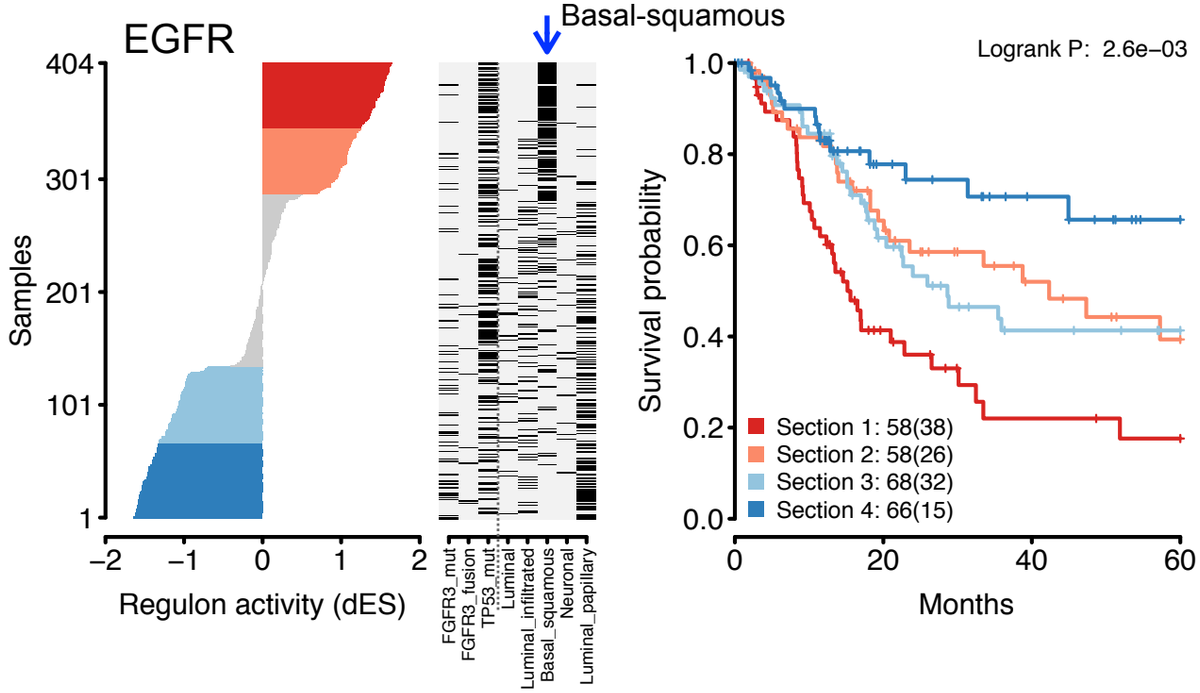
Figure 7. The ESR1 regulon as readout of cell state. …(**c**) Differential enrichment scores calculated for all tumors in METABRIC cohort 1. …ER status, PAM50 subclass and tumor grade.… (**d**) Kaplan-Meier … disease-specific survival … tumor subgroups highlighted in **c**. …

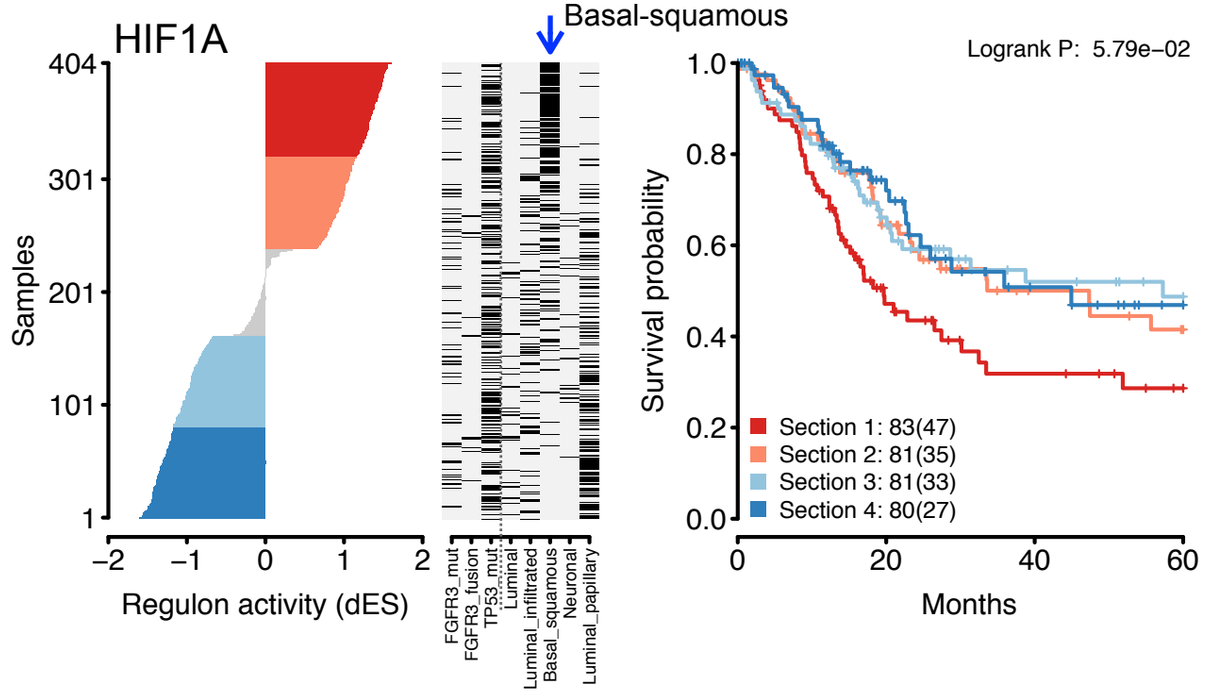
Castro MAA *et al*. *Regulators of genetic risk of breast cancer identified by integrative network analysis*. Nature Genetics, 2016. 48(1):12-21.

TCGA MIBC

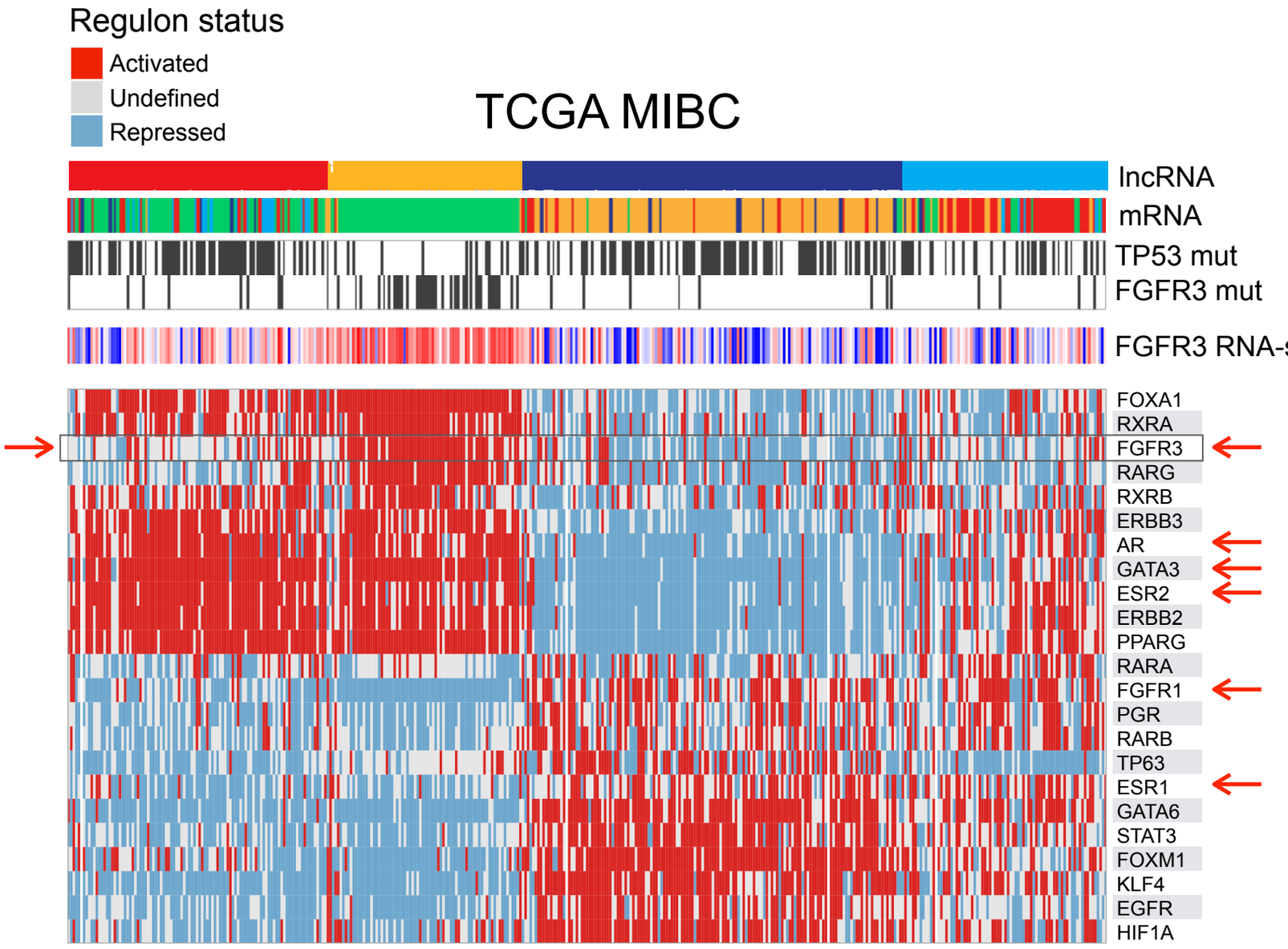


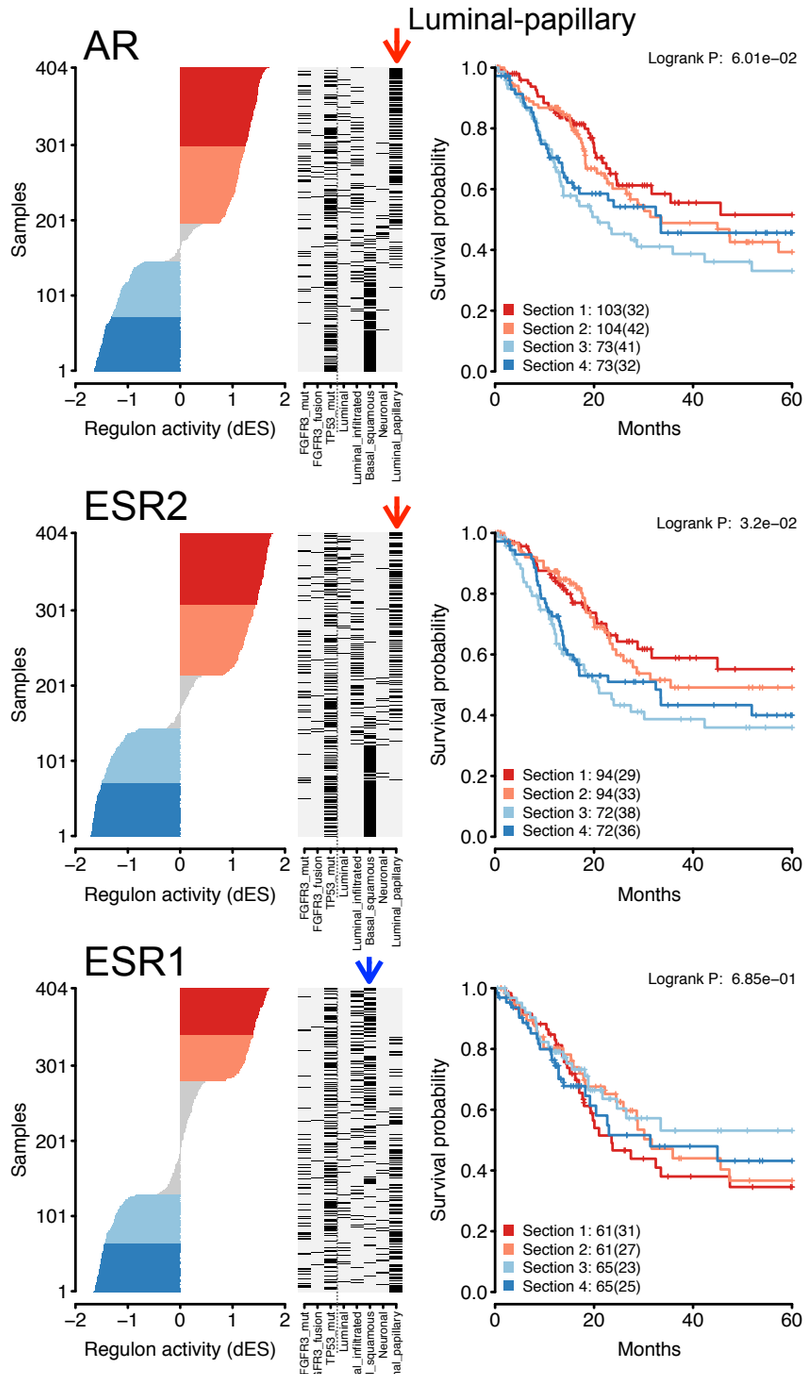
FGFR3_mut
FGFR3_fusion
TP53_mut
Luminal
Luminal_infiltrated
Basal_squamous
Neuronal
Luminal_papillary

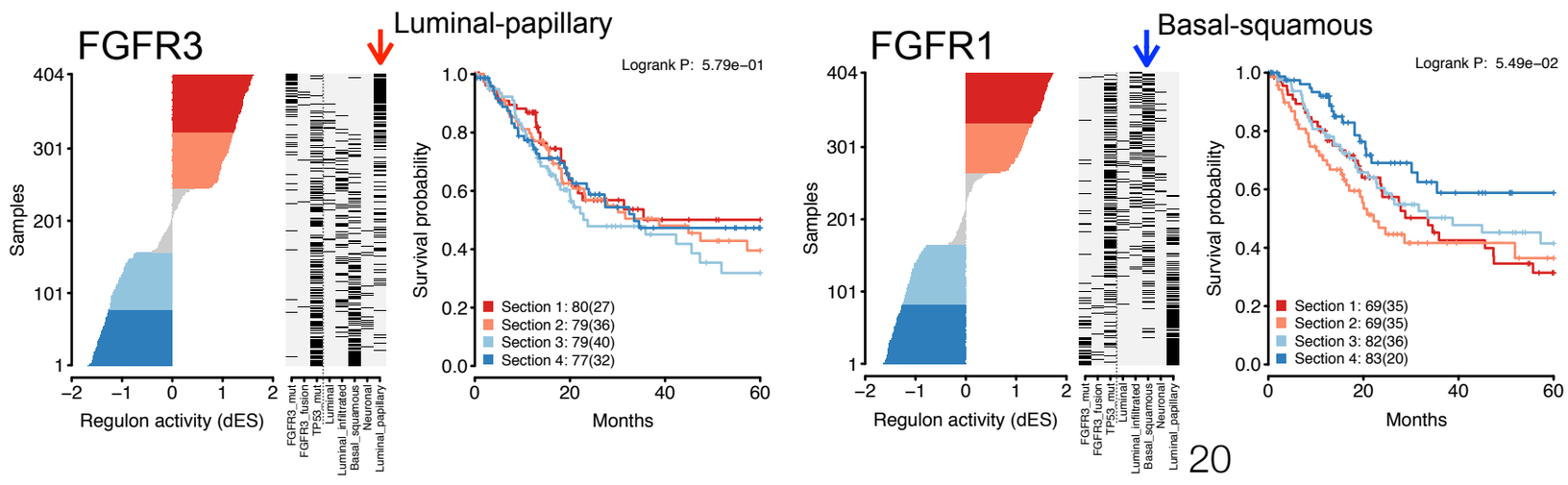




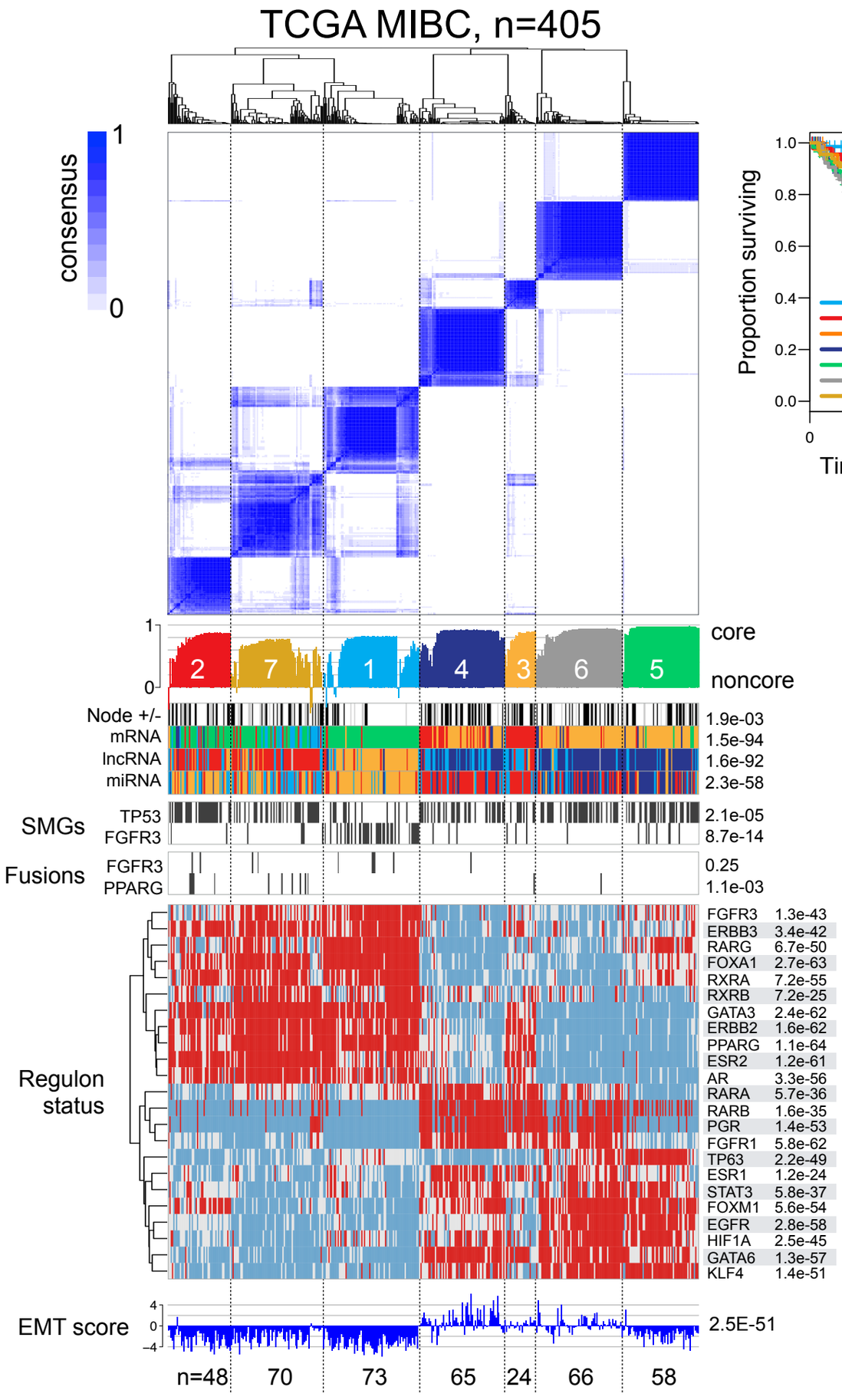
Regulon activity/status and subtypes

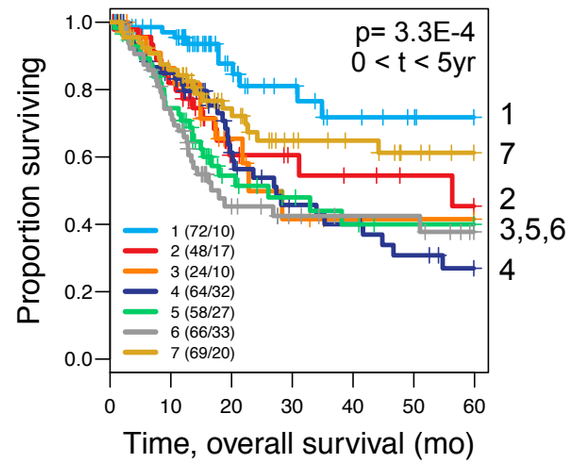


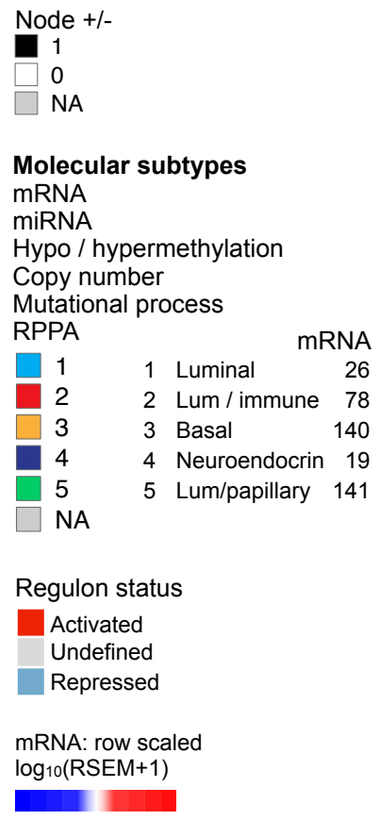




Finer-grained consensus subtypes from regulon activity profiles







Unsupervised consensus clustering of RAPs for 23 BLCA-associated regulators

AR, EGFR, ERBB2, ERBB3, ESR1, ESR2, FGFR1, FGFR3, FOXA1, FOXM1, GATA3, GATA6, HIF1A, KLF4, PGR, PPARG, RARA, RARB, RARG, RXRA, RXRB, STAT3, TP63

Methods publications, with case studies

RTNsurvival: an R/Bioconductor package for regulatory network survival analysis

Clarice S. Groeneveld, Vinicius S. Chagas, Steven J. M. Jones, A. Gordon Robertson, Bruce A. J. Ponder, Kerstin B. Meyer and Mauro A. A. Castro

Bioinformatics. 2019 Mar 28. [Epub ahead of print]

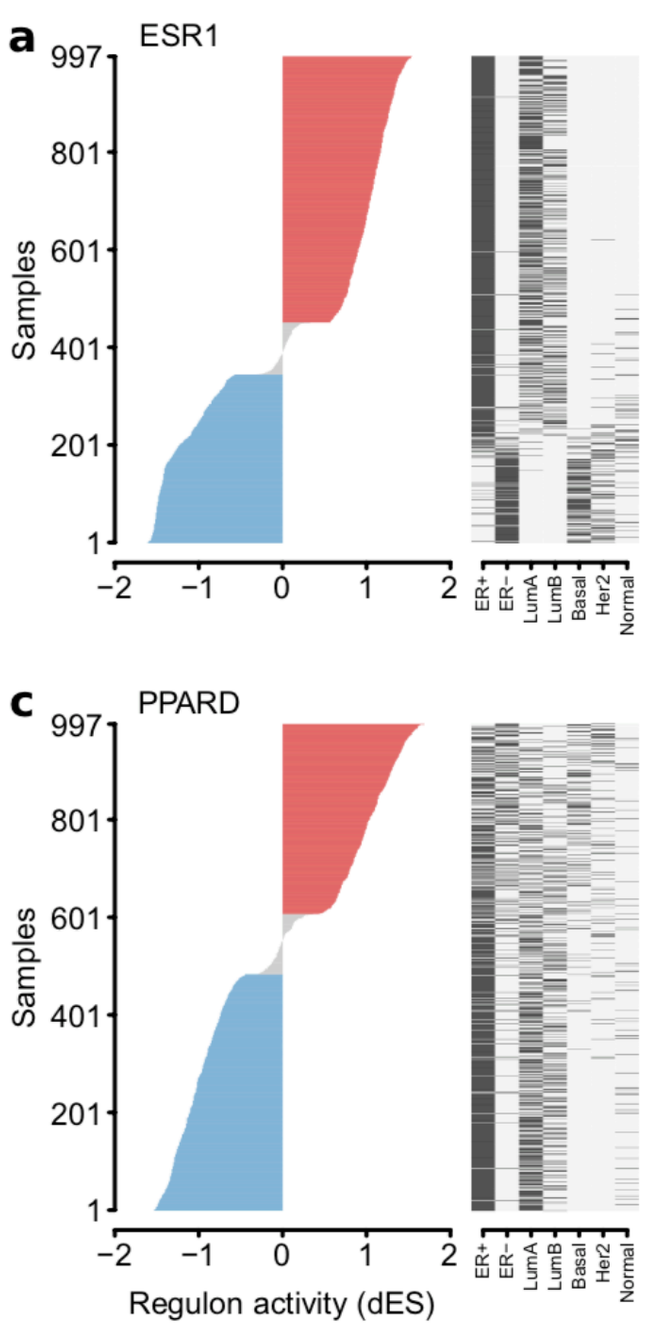
RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons

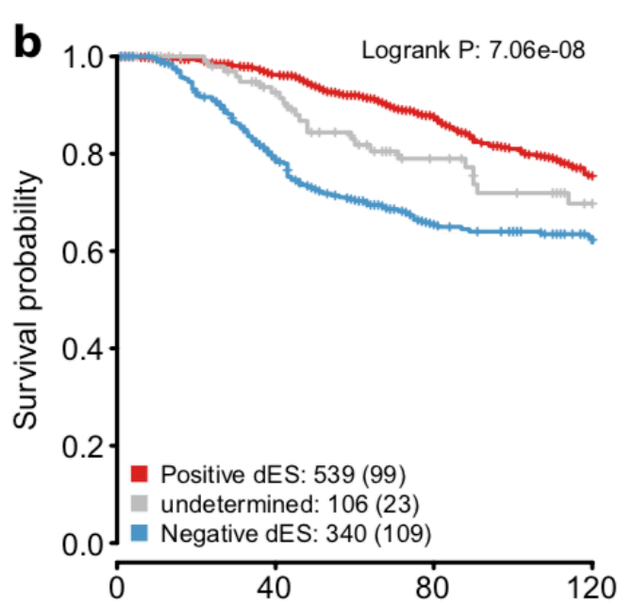
Vinicius S. Chagas†, Clarice S. Groeneveld†, Kelin G. Oliveira, Sheyla Trefflich, Rodrigo C. de Almeida, Bruce A. J. Ponder, Kerstin B. Meyer, Steven J. M. Jones, A. Gordon Robertson and Mauro A. A. Castro

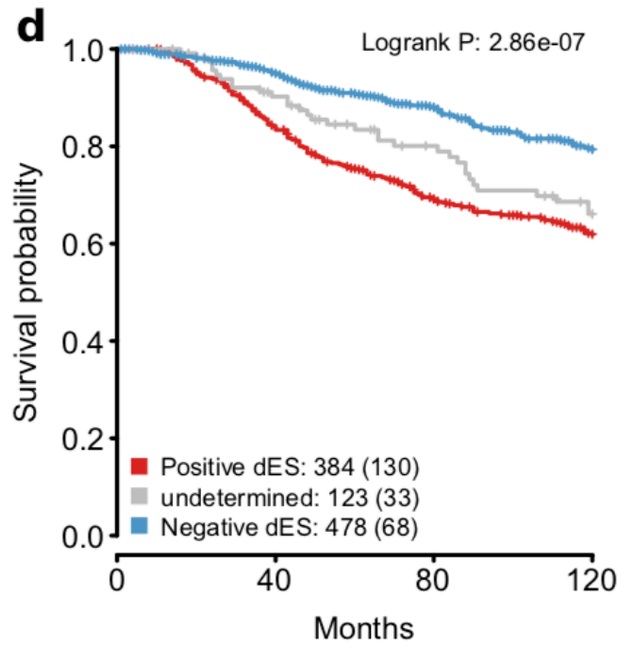
Bioinformatics, btz534, Published: 28 June 2019

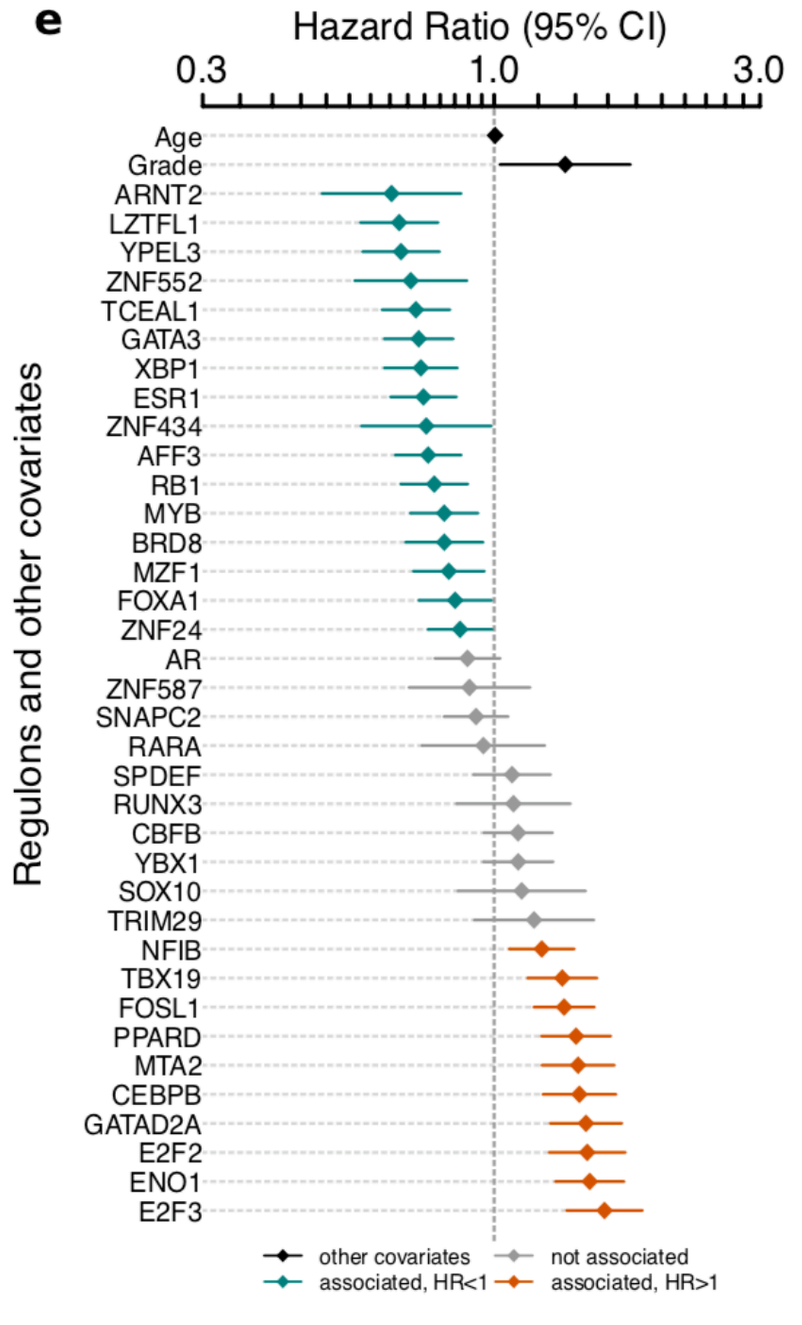
RTNsurvival: an R/Bioconductor package for regulatory network survival analysis

METABRIC cohort 1





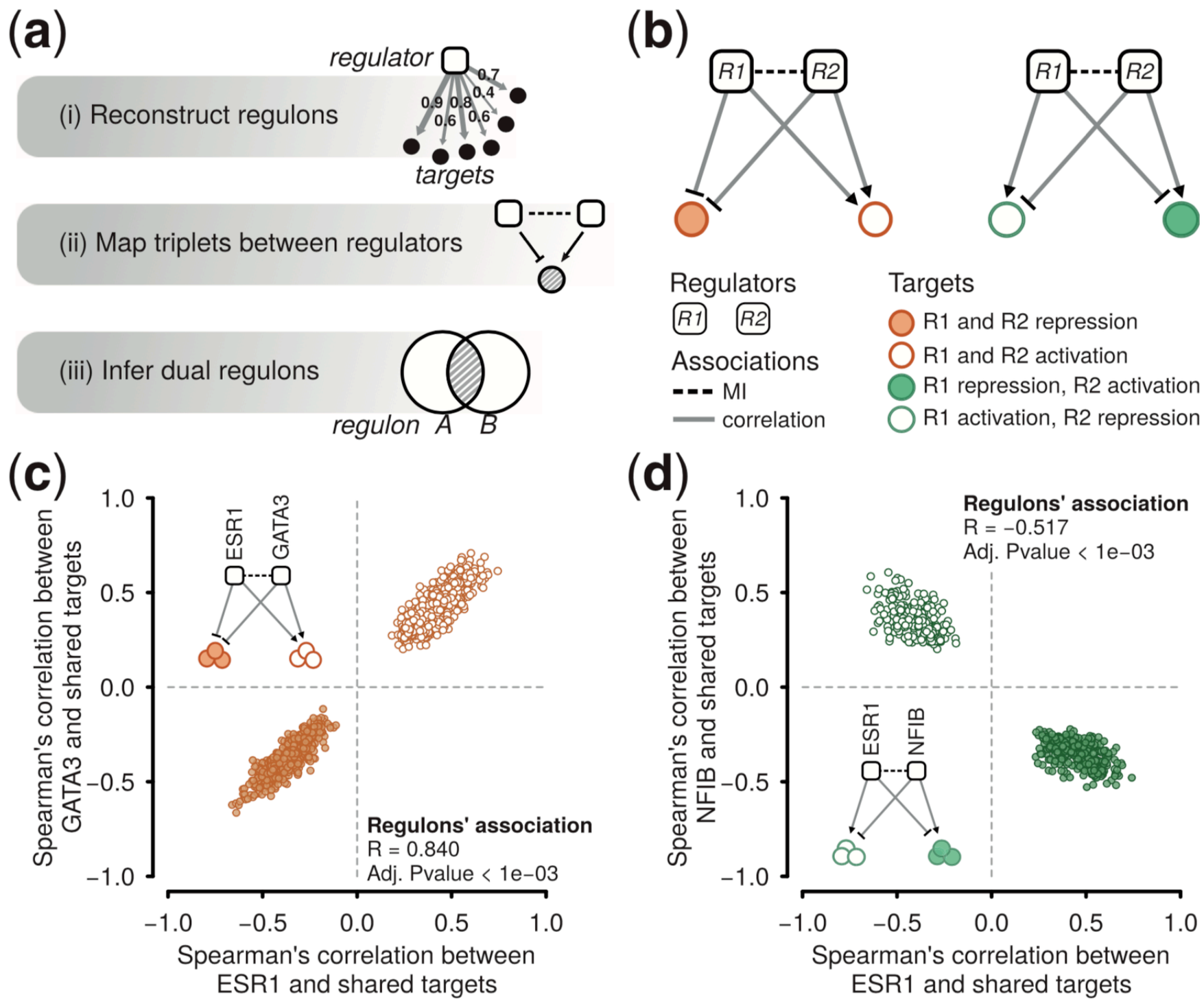




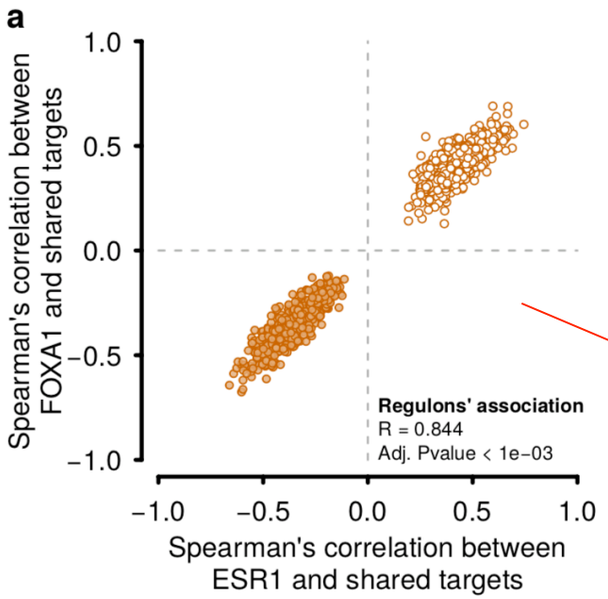
Supplementary Information

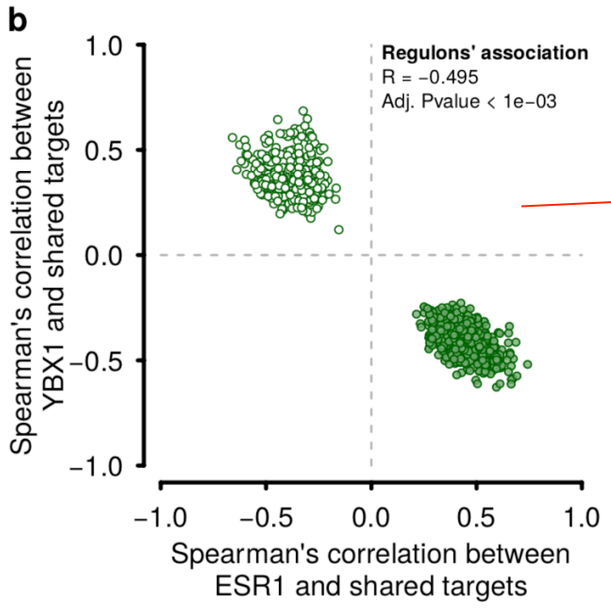
RTNsurvival case studies: regulon activity as a predictor variable in univariate and multivariate survival analyses.

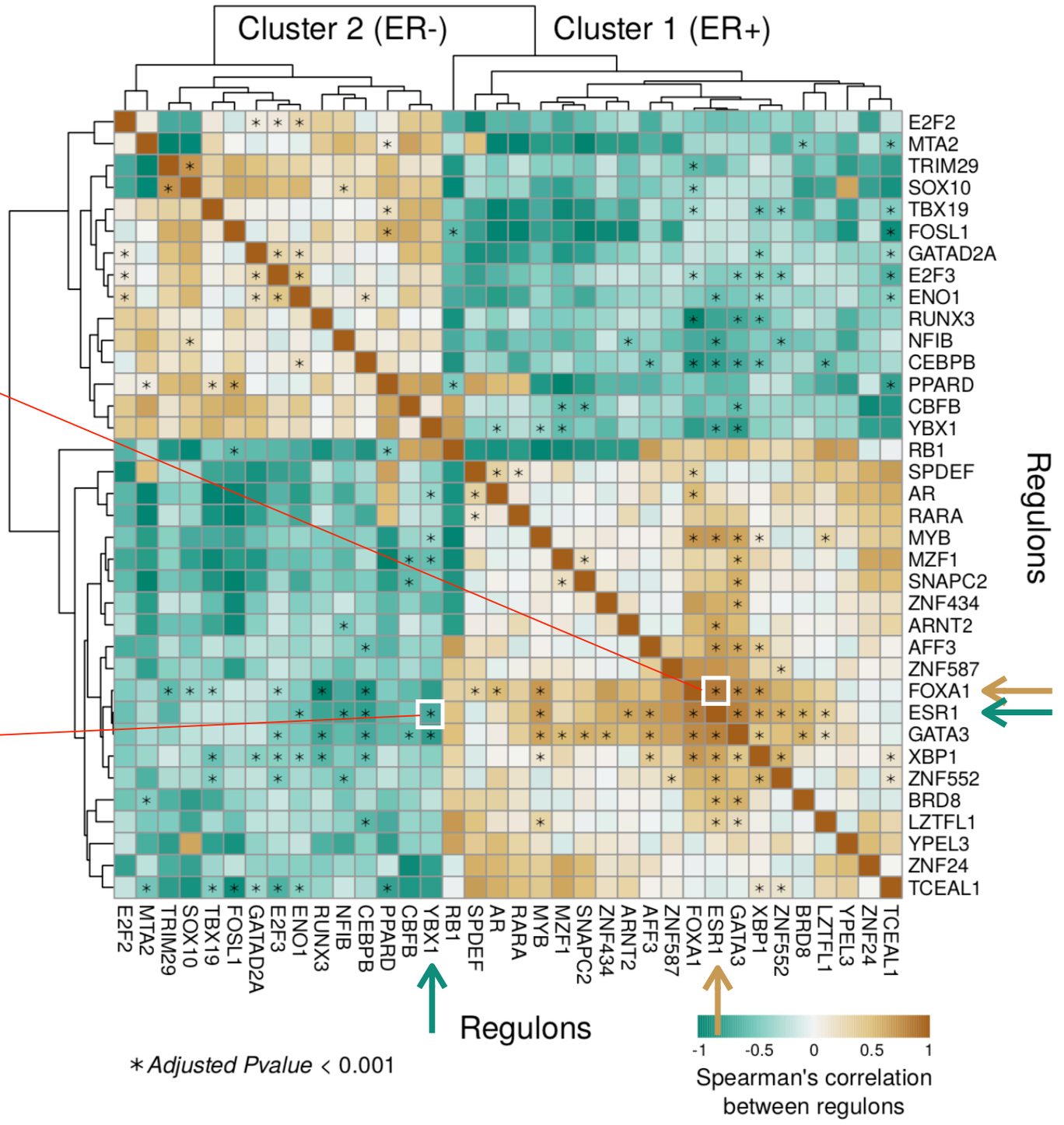
RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons



METABRIC, n=997







Supplementary Figure 2: Heatmap showing the correlation matrix between regulons for 36 [risk-associated] transcription factors. Each [cell] in the heatmap summarizes the relationship between a regulon's shared targets as shown in the scatter plots of Supplementary Figure 1. Significant associations (P < 0.001, BH adjusted) are indicated with asterisks. "Cluster 1" and "Cluster 2", as named in Castro et al. (2016), represent regulons associated with ER+ and ER- tumours, respectively.

Supplementary Information

RTNduals case studies: exploring dual regulons in breast cancer regulatory networks.

Contents

- Consensus MIBC subtypes
- TCGA
  - miRNA-seq data generating process
- TCGA BLCA
  - LncRNA-based MIBC subtypes
- PanCancer Atlas resources
  - Publications, clinical / batch-corrected expression data
- The GDAN and the GDC, Interim projects
  - GDC-QC: hg19 vs. hg38 data for all TCGA platforms
    - miRNA-seq data
- Regulon analysis
  - Two method publications, with case studies

Special thanks to

Mauro A.A. Castro
Vinicius S. Chagas
Clarice S. Groeneveld
Bioinformatics and Systems Biology Laboratory, Federal University of Paraná Polytechnic Center, Curitiba, Brazil

Benilton de Sa Carvalho
Biostatistics and Computational Biology Laboratory, Department of Statistics, University of Campinas, São Paulo, Brazil

Ewan A. Gibb
Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada
Now with Decipher Biosciences, Vancouver, Canada

Karen L. Mungall, Sara Sadeghi
Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada

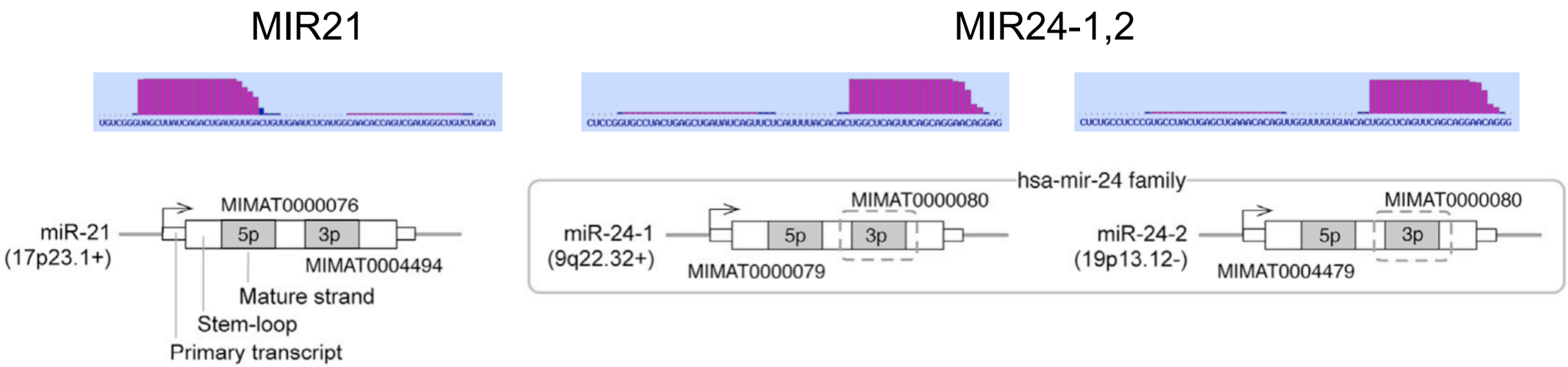
Tara M. Lichtenberg
Biospecimen Core Resource, The Research Institute at Nationwide Children's Hospital, Columbus, OH

Hikmat Al-Ahmadie
Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY

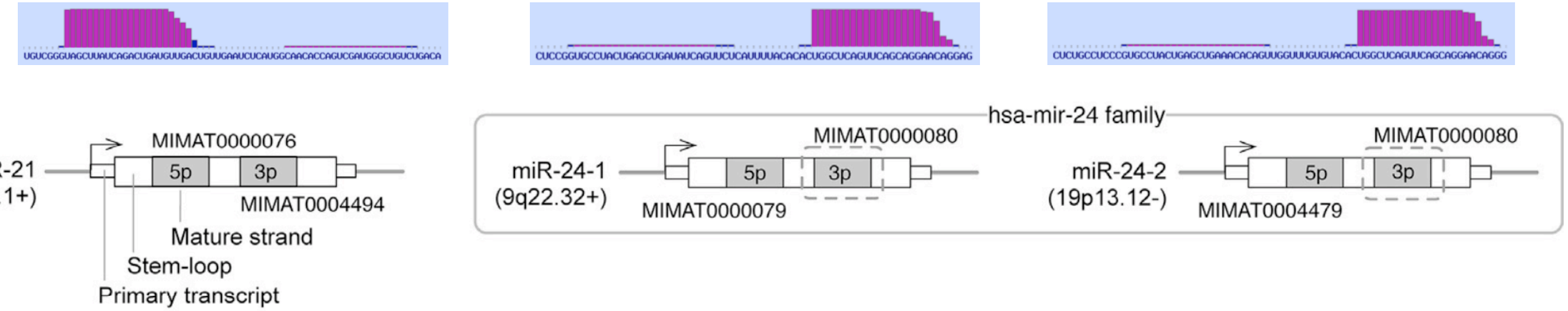
John N. Weinstein, David J. Kwiatkowski and Seth P. Lerner

Thank you

When a mature strand can be expressed from more than one genomic location, the TCGA miR-seq data do *not* indicate which location expressed a read

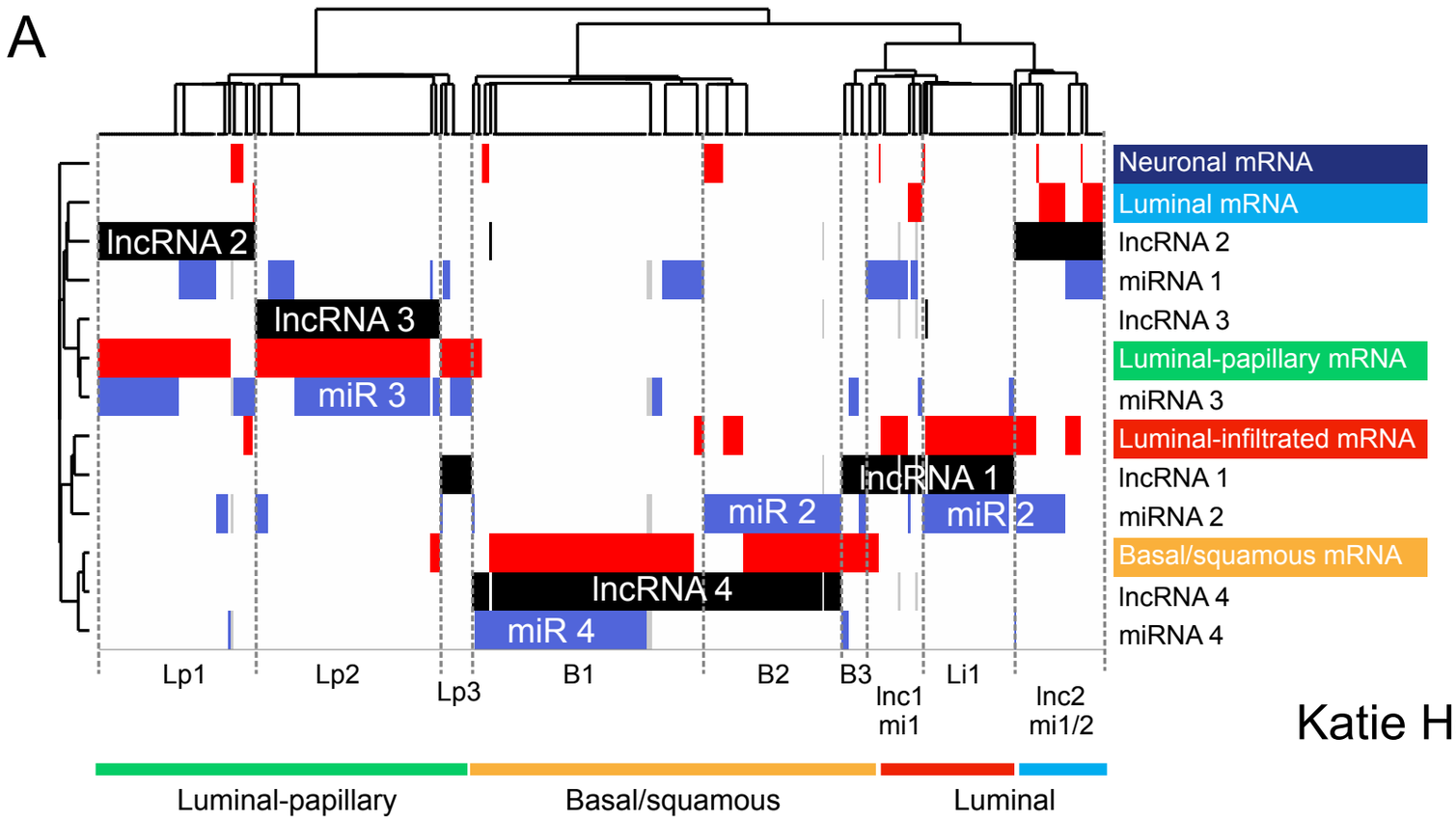


Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data



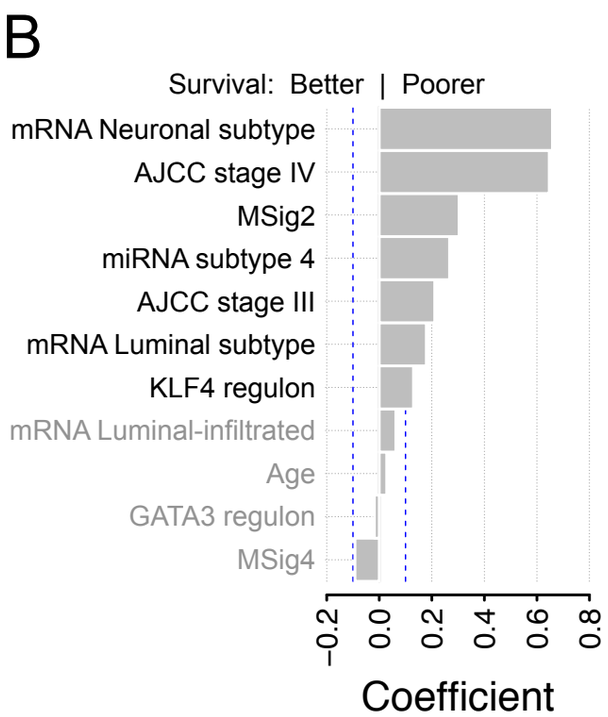
Analyses involving: only mature strand RPMs, vs. locations (stem-loop RPMs)

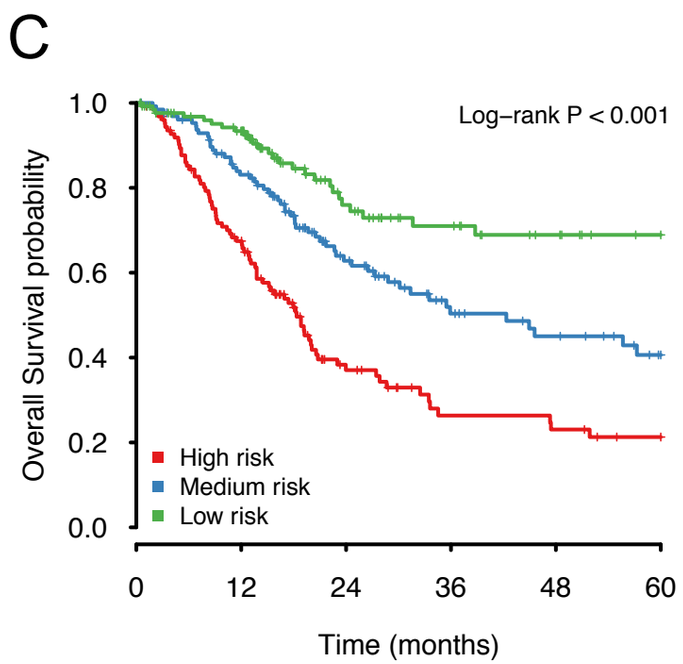
COCA subtypes



Katie Hoadley, UNC

Multivariate OS





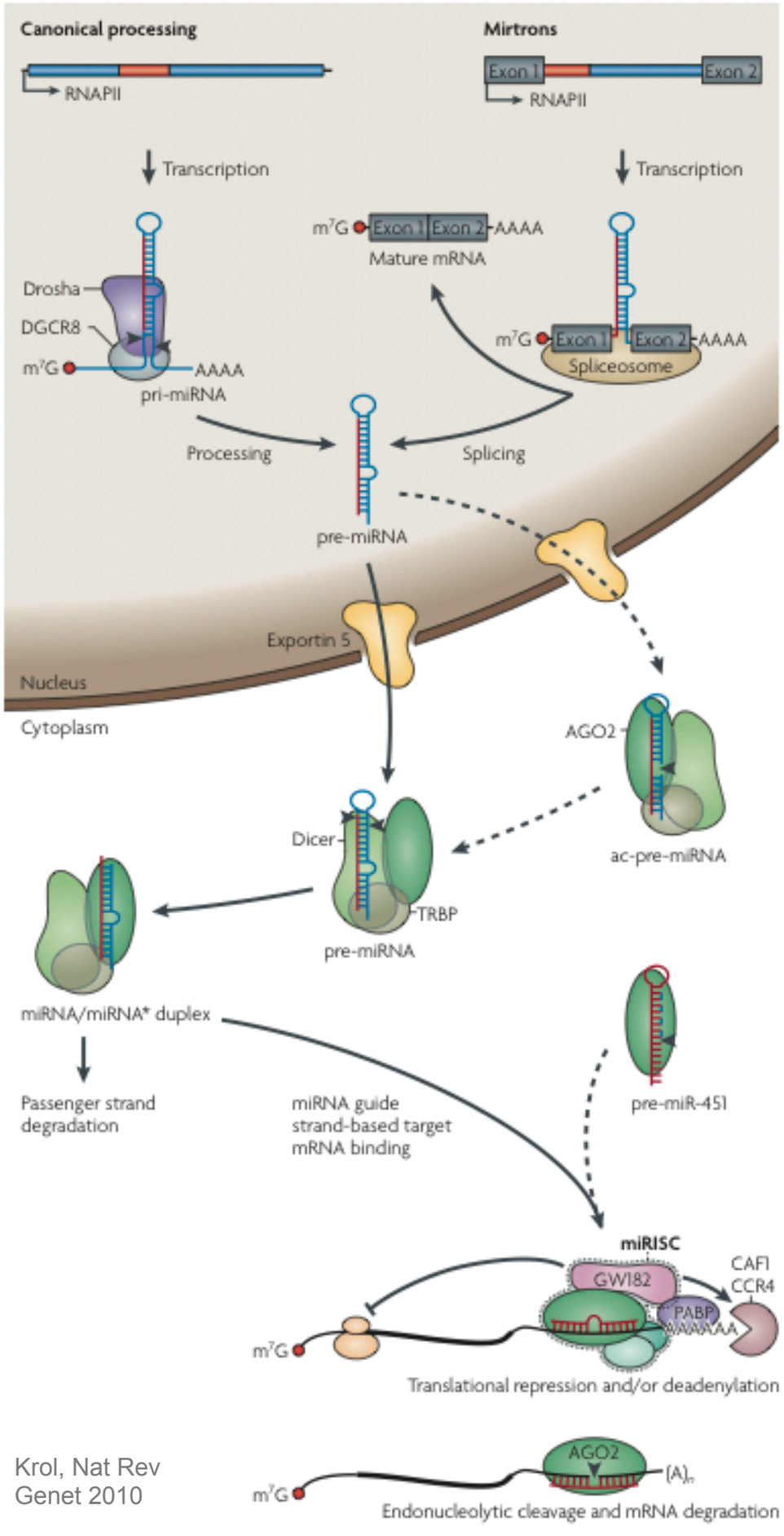
Using the best-OS subtype as the reference

Mauro AA Castro
Benilton de Sa Carvalho

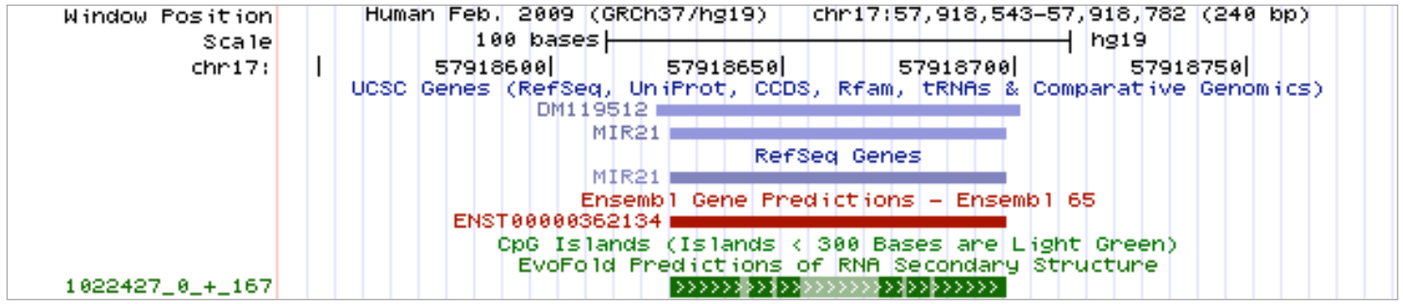
Getting miRNA-seq data from the GDC: isomiRs


```
library(TCGAbiolinks)

#-- 2. Harmonized hg38 isoform data

query.mirna.isoform <- GDCquery(project = "TCGA-BLCA",
                        data.category = "Transcriptome Profiling",
                        data.type = "Isoform Expression Quantification",
                        workflow.type = "BCGSC miRNA Profiling",
                        experimental.strategy = "miRNA-Seq",
                        sample.type = c("Primary solid Tumor")
                        )

GDCdownload(query.mirna.isoform,
            method = "api",
            directory = "GDCdata_hg38_isoforms",
            files.per.chunk = 50)

#-- Prepare: save an .RData file into working folder, then delete all intermediate files
hg38_isoform_data <- GDCprepare(query.mirna.isoform,
                                directory = "GDCdata_hg38_isoforms",
                                save = T,
                                save.filename = "hg38_mirna_isoforms.RData",
                                remove.files.prepared = TRUE
                                )

#-- load the .RData from the working directory
hg38_isoform_datafile <- get(load("hg38_mirna_isoforms.RData"))

dim(hg38_isoform_datafile)
# 2025288       7

# For each sample, the file has a read_count, RPM, and crossmapped column
head(hg38_isoform_datafile)
#   miRNA_ID     isoform_coords                read_count  RPM    `cross-mapped` miRNA_region        barcode
#   <chr>        <chr>                         <int>       <dbl>  <chr>          <chr>               <chr>
# 1 hsa-let-7a-1 hg38:chr9:94175961-94175982:+          1   0.35  N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
# 2 hsa-let-7a-1 hg38:chr9:94175961-94175983:+          4   1.41  N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
# 3 hsa-let-7a-1 hg38:chr9:94175961-94175984:+         13   4.60  N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
# 4 hsa-let-7a-1 hg38:chr9:94175961-94175985:+          1   0.35  N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
# 5 hsa-let-7a-1 hg38:chr9:94175962-94175981:+         33   11.7  N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
# 6 hsa-let-7a-1 hg38:chr9:94175962-94175982:+       1662  588.   N    mature,MIMAT0000062 TCGA-E7-A7PW-01A-11R-A358-
```

miRNA biogenesis, resources



Krol, Nat Rev Genet 2010



Stem-loop sequence MI0000102

Accession	MI0000102
ID	hsa-mir-100
Symbol	HGNC:MIR100
Description	Homo sapiens miR-100 stem-loop
Stem-loop	-uu c a cg c a - uau ccug g caca acc uagau cga cuugug g u |||| | |||| ||| ||||| ||| |||||| | ggau c gugu ugg aucua guu gaacac c a ugu u a au u c g cug

Mature sequence MIMAT0000098

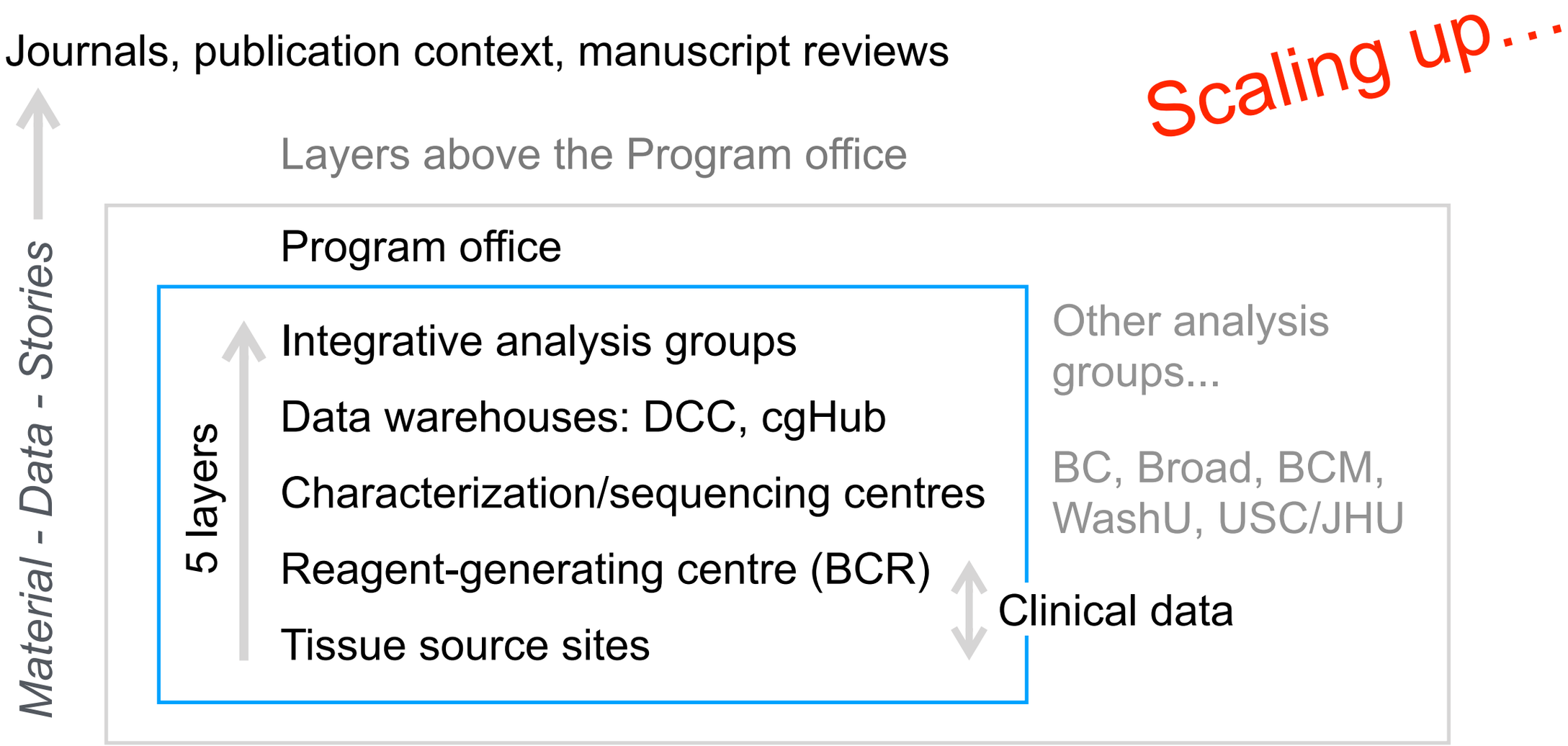
Accession	MIMAT0000098
ID	hsa-miR-100
Sequence	13 - aacccguagauccgaacuugug - 34

Minor miR* sequence MIMAT0004512

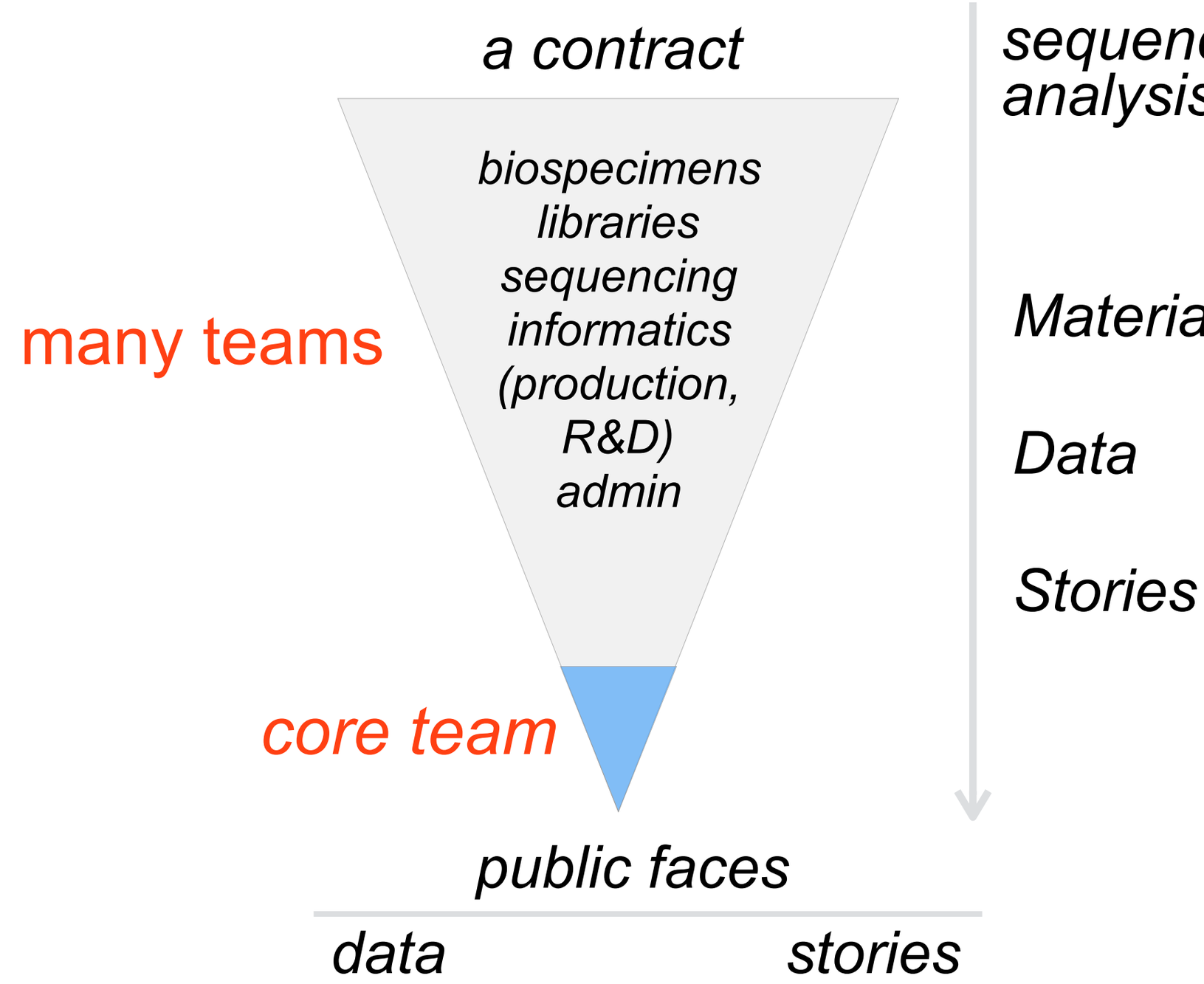
Accession	MIMAT0004512
ID	hsa-miR-100*
Sequence	48 - caagcuuguaucuauagguaug - 69 Get sequence

miRBase 16

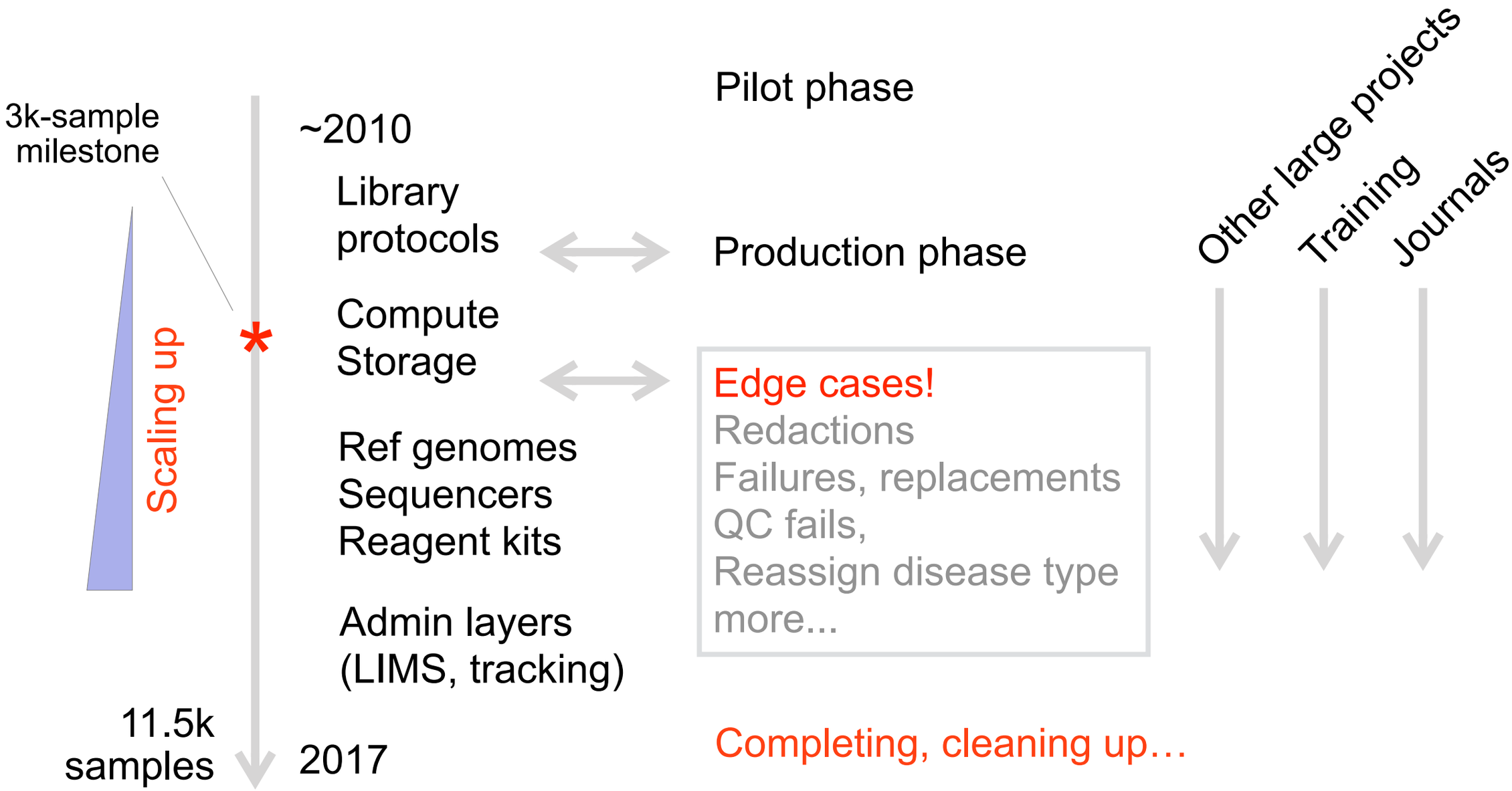
Consortium structure, flow, and scale



The BC lab: ~300 people



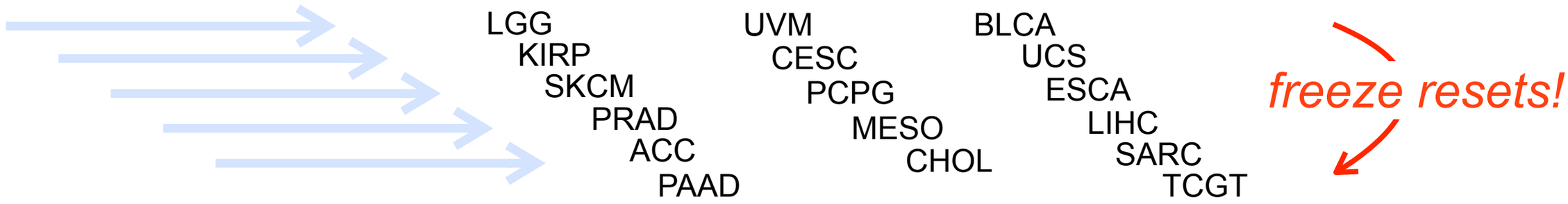
Persistent challenges: generating data



Persistent challenges: analysis

1. miRNA-seq: clustering, differential abundance, targeting
2. RNA-seq: coding and noncoding RNAs
3. Microbes: read screening and *de novo* assembly
4. Copy number and miRNAs
5. DNA methylation and miRNAs, TSSs
6. Mutation calls
7. Tumour purity
8. Clinical data, path review, platforms, and data freezes
9. Context: medical and genomic research, publications

Overlapping projects



The BCCA Genome Sciences Centre in TCGA

Core team
Reanne Bowlby
Denise Brooks
Andy Mungall
Gordon Robertson
Payal Sipahimalani

Microbes/assembly
Karen Mungall
Sara Sadeghi
Lynette Lim
Richard Mar
Victoria Trinh
Caleb Choo

Strelka: long indels
Katy Kasaian

lncRNAs
Ewan Gibb

Team leads
Adrian Ally
Miruna Balasundaram
Yaron Butterfield
Eric Chuah
Amanda Clarke
Qixia Deng
Noreen Dhalla
Ranabir Guin
Carrie Hirst
Darlene Lee
Haiyan I. Li
Michael Mayo
Angela Tam
Nina Thiessen
Tina Wong
Natasja Wye
Kelsey Zhu

Marco Marra
Steven Jones
Yussanne Ma bioIT group
Robin Coope engineering
Richard Moore sequencing
Robert Holt sequencing
Lance Bailey systems
Jacqueline Schein biospecimens

An NIH contract: ~2k RNA-seq datasets

OV, LAML, STAD, ESCA

Karen Mungall
Readman Chiu
Caleb Choo
Richard Mar

Swanson *et al. BMC Genomics* 2013, **14**:550
http://www.biomedcentral.com/1471-2164/14/550

BMC Genomics

METHODOLOGY ARTICLE **Open Access**

Barnacle: detecting and characterizing tandem duplications and fusions in transcriptome assemblies

Lucas Swanson[1,2], Gordon Robertson[1], Karen L Mungall[1], Yaron S Butterfield[1], Readman Chiu[1], Richard D Corbett[1], T Roderick Docking[1], Donna Hogge[3], Shaun D Jackman[1], Richard A Moore[1], Andrew J Mungall[1], Ka Ming Nip[1], Jeremy DK Parker[1], Jenny Qing Qian[1], Anthony Raymond[1], Sandy Sung[1], Angela Tam[1], Nina Thiessen[1], Richard Varhol[1], Sherry Wang[1], Deniz Yorukoglu[1,2,5], YongJun Zhao[1], Pamela A Hoodless[3,4], S Cenk Sahinalp[2], Aly Karsan[1] and Inanc Birol[1,2,4*]

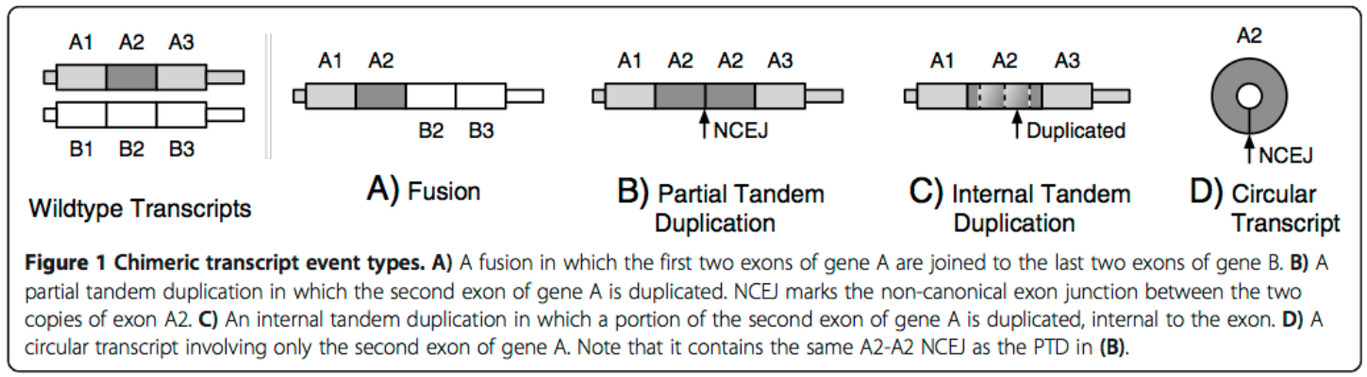


Figure 1 Chimeric transcript event types. A) A fusion in which the first two exons of gene A are joined to the last two exons of gene B. **B)** A partial tandem duplication in which the second exon of gene A is duplicated. NCEJ marks the non-canonical exon junction between the two copies of exon A2. **C)** An internal tandem duplication in which a portion of the second exon of gene A is duplicated, internal to the exon. **D)** A circular transcript involving only the second exon of gene A. Note that it contains the same A2-A2 NCEJ as the PTD in **(B)**.

De novo assembly and analysis of RNA-seq data

Gordon Robertson[1], Jacqueline Schein[1], Readman Chiu[1], Richard Corbett[1], Matthew Field[1], Shaun D Jackman[1], Karen Mungall[1], Sam Lee[2], Hisanaga Mark Okada[1], Jenny Q Qian[1], Malachi Griffith[1], Anthony Raymond[1], Nina Thiessen[1], Timothee Cezard[1,4], Yaron S Butterfield[1], Richard Newsome[1], Simon K Chan[1], Rong She[1], Richard Varhol[1], Baljit Kamoh[1], Anna-Liisa Prabhu[1], Angela Tam[1], YongJun Zhao[1], Richard A Moore[1], Martin Hirst[1], Marco A Marra[1,3], Steven J M Jones[1,3], Pamela A Hoodless[2,3] & Inanc Birol[1]

We describe Trans-ABySS, a *de novo* short-read transcriptome assembly and analysis pipeline that addresses variation in local read densities by assembling read substrings with varying stringencies and then merging the resulting contigs before analysis. Analyzing 7.4 gigabases of 50-base-pair paired-end Illumina reads from an adult mouse liver poly(A) RNA library, we identified known, new and alternative structures in expressed transcripts, and achieved high sensitivity and specificity relative to reference-based assembly methods.

NATURE METHODS | VOL.7 NO.11 | NOVEMBER 2010 | 909

An NIH contract: ~2k RNA-seq datasets

OV, LAML, STAD, ESCA

Karen Mungall
Readman Chiu
Caleb Choo
Richard Mar

Swanson *et al. BMC Genomics* 2013, **14**:550
http://www.biomedcentral.com/1471-2164/14/550

BMC

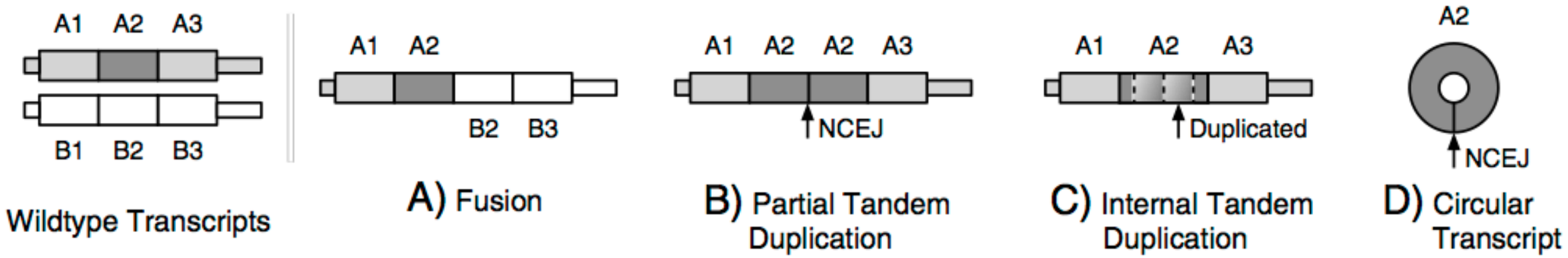


Figure 1 Chimeric transcript event types. A) A fusion in which the first two exons of gene A are joined to the last two exons of gene B. **B)** A partial tandem duplication in which the second exon of gene A is duplicated. NCEJ marks the non-canonical exon junction between the two copies of exon A2. **C)** An internal tandem duplication in which a portion of the second exon of gene A is duplicated, internal to the exon. **D)** A circular transcript involving only the second exon of gene A. Note that it contains the same A2-A2 NCEJ as the PTD in **(B)**.

assembly and analysis pipeline that addresses variation in local read densities by assembling read substrings with varying stringencies and then merging the resulting contigs before analysis. Analyzing 7.4 gigabases of 50-base-pair paired-end Illumina reads from an adult mouse liver poly(A) RNA library, we identified known, new and alternative structures in expressed transcripts, and achieved high sensitivity and specificity relative to reference-based assembly methods.

NATURE METHODS | VOL.7 NO.11 | NOVEMBER 2010 | 909

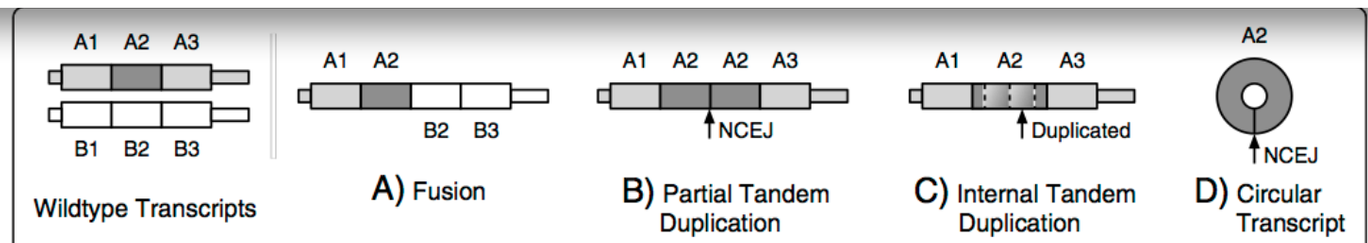


Figure 1 Chimeric transcript event types. A) A fusion in which the first two exons of gene A are joined to the last two exons of gene B. **B)** A partial tandem duplication in which the second exon of gene A is duplicated. NCEJ marks the non-canonical exon junction between the two copies of exon A2. **C)** An internal tandem duplication in which a portion of the second exon of gene A is duplicated, internal to the exon. **D)** A circular transcript involving only the second exon of gene A. Note that it contains the same A2-A2 NCEJ as the PTD in **(B)**.

Microbes: screening and genomic integration

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3402–3404
doi:10.1093/bioinformatics/btu558

Sequence analysis

Advance Access publication August 20, 2014

BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters

Justin Chu*, Sara Sadeghi, Anthony Raymond, Shaun D. Jackman, Ka Ming Nip, Richard Mar, Hamid Mohamadi, Yaron S. Butterfield, A. Gordon Robertson and Inanç Birol*

Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada

Associate Editor: Alfonso Valencia

De novo assembly and analysis of RNA-seq data

Gordon Robertson[1], Jacqueline Schein[1], Readman Chiu[1], Richard Corbett[1], Matthew Field[1], Shaun D Jackman[1], Karen Mungall[1], Sam Lee[2], Hisanaga Mark Okada[1], Jenny Q Qian[1], Malachi Griffith[1], Anthony Raymond[1], Nina Thiessen[1], Timothee Cezard[1,4], Yaron S Butterfield[1], Richard Newsome[1], Simon K Chan[1], Rong She[1], Richard Varhol[1], Baljit Kamoh[1], Anna-Liisa Prabhu[1], Angela Tam[1], YongJun Zhao[1], Richard A Moore[1], Martin Hirst[1], Marco A Marra[1,3], Steven J M Jones[1,3], Pamela A Hoodless[2,3] & Inanc Birol[1]

We describe Trans-ABySS, a *de novo* short-read transcriptome assembly and analysis pipeline that addresses variation in local read densities by assembling read substrings with varying stringencies and then merging the resulting contigs before analysis. Analyzing 7.4 gigabases of 50-base-pair paired-end Illumina reads from an adult mouse liver poly(A) RNA library, we identified known, new and alternative structures in expressed transcripts, and achieved high sensitivity and specificity relative to reference-based assembly methods.

NATURE METHODS | VOL.7 NO.11 | NOVEMBER 2010 | 909

Sara Sadeghi
Karen Mungall

BLCA (RNA & DNA)
CESC (RNA)
CHOL
ESCA
LGG
LIHC
MESO
Pan-GI
Pan-SCC
SARC
TGCT
THYM
THCA
UCS
UVM (RNA)

1. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer

28 iCluster subtypes		Basal-sqam	Luminal	Luminal-inf	Luminal pap	Neuronal
C1	STAD (EBV-CIMP)	0	0	0	0	0
C2	BRCA (HER2 amp)	1	0	3	4	2
C3	Mesenchymal (Immune)	14	0	1	0	2
C4	Pan-GI (CRC)	0	0	0	0	0
C5	CNS/Endocrine	0	0	0	0	1
C6	OV	0	0	0	0	0
C7	Mixed (chr9 del)	10	1	2	36	1
C8	UCEC	0	0	0	3	2
C9	ACC/KICH	0	0	0	0	0
C10	Pan-Squamous	25	9	20	33	2
C11	LCC (IDH mut)	0	0	0	0	0
C12	THCA	0	0	0	2	0
C13	Mixed (chr8 del)	2	8	5	14	1
C14	LUAD	0	0	0	1	0
C15	SKCM/UVM	0	0	0	0	0
C16	PRAD	0	0	0	0	0
C17	BRCA (chr8q amp)	1	1	2	3	3
C18	Pan-GI (MSI)	0	0	1	0	0
C19	BRCA (Luminal)	0	0	0	0	0
C20	Mixed (Stromal/Immune)	17	3	36	5	2
C21	DLBC	0	0	0	0	0
C22	TGCT	0	0	0	2	1
C23	GBM/LHH (IDH1 wt)	0	0	0	0	0
C24	LAML	0	0	0	0	0
C25	Pan-SCC (chr11 amp)	6	3	3	9	1
C26	LIHC	0	0	0	0	0
C27	Pan-SCC (HPV)	63	0	3	29	0
C28	Pan-Kidney	0	0	0	0	0

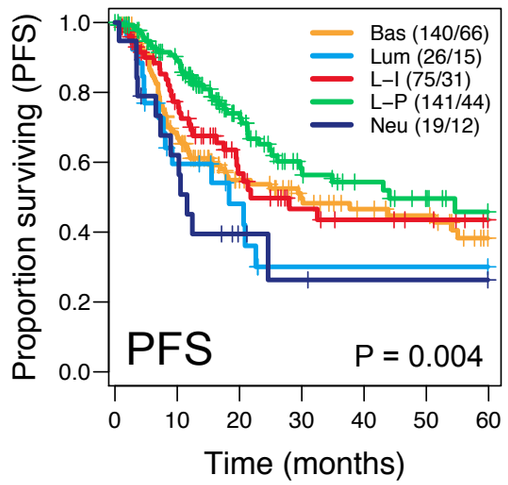
TCGA mRNA subtypes (n = 399)

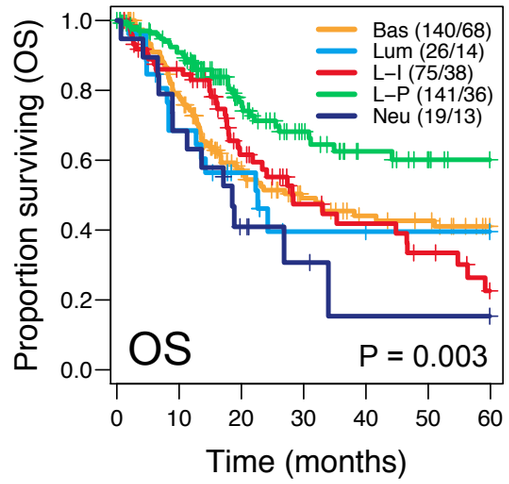
2. The Immune Landscape of Cancer

6 Immune subtypes		Basal-squamous	Luminal	Luminal-infiltrated	Luminal papillary	Neuronal	
C1	Wound healing	26	19	44	69	14	172
C2	IFN-γ dominant	105	4	24	28	3	164
C3	Inflammatory	0	1	8	12	0	21
C4	Lymphocyte depleted	5	1	0	29	0	35
C5	Immunologically quiet	0	0	0	0	0	0
C6	TGF-β dominant	1	0	0	2	0	3
		137	25	76	140	17	395

TCGA mRNA subtypes

3. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics





TCGA mRNA subtypes

Note: Table 3 gives an Assessment and Recommended Use of the Endpoints of OS, PFI, DFI, and DSS

NCI's Center for Cancer Genomics Genome Data Analysis Network

NIH NATIONAL CANCER INSTITUTE

20 Oct 2016

1-800-4-CANCER | Live Chat | Publications | Dictionary

ABOUT CANCER | CANCER TYPES | RESEARCH | GRANTS & TRAINING | NEWS & EVENTS | ABOUT NCI | search

Home > About NCI > NCI Organization > CCG > Insights & Innovations Blog

CCG Welcomes a New Genomic Data Analysis Network

Subscribe

October 20, 2016, by Jean Claude Zenklusen, Ph.D.

As NCI's Center for Cancer Genomics (CCG) shifts its focus from The Cancer Genome Atlas (TCGA) project to new research, our strategy is to maintain the efficient workflow that made TCGA a success while adding key functionalities and expertise. The new members of our Genomic Data Analysis Network (GDAN), four of whom are first-time NIH grant recipients, each bring unique knowledge to the network, creating an exciting blend of scientific capabilities. This team, expanded from seven to thirteen centers, will deliver actionable insights from CCG's genomic and clinical data to the entire cancer research community.

Jean Claude Zenklusen, Ph.D.

There are three types of centers in the GDAN, each designed to contribute to a different facet of genomic analysis: Processing, Specialized, and Visualization.

Featured Posts

Clouds Democratize CCG Data
March 7, 2017, by Izumi Hinkson, Ph.D.

Researcher Studies Own Cancer
December 6, 2016, by Amy E. Blum, M.A.

TCGA Data Inform LOXO-101 Drug Development
September 6, 2016, by Amy E. Blum, M.A.

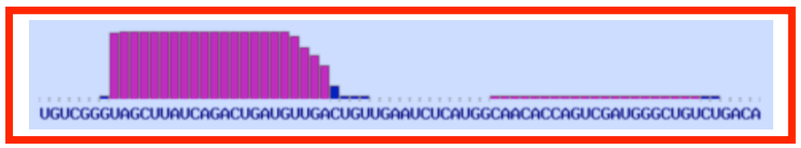
Archive

2017 (6)
2016 (14)
2015 (10)

Specialized Center: Steven Jones (BCCA)/Theo Knijnenberg(ISB): miRNA data analysis

https://www.cancer.gov/about-nci/organization/ccg/blog/2016/new-genomic-data-analysis-network

Getting miRNA-seq data from the GDC: stem-loops




```
library(TCGAbiolinks) # v2.10.4

getGDCInfo()$data_release
# "Data Release 15.0 - February 20, 2019"

#-- 1. Harmonized stem-loop data for the TCGA-BLCA cohort
query.mirna <- GDCquery(project = "TCGA-BLCA",
                        data.category = "Transcriptome Profiling",
                        data.type = "miRNA Expression Quantification",
                        workflow.type = "BCGSC miRNA Profiling",
                        sample.type = c("Primary solid Tumor"),
                        experimental.strategy = "miRNA-Seq"
)
GDCdownload(query.mirna,
            method = "api",
            directory = "GDCdata_hg38_stemloops",
            files.per.chunk = 50 )
# Downloading data for project TCGA-BLCA
# GDCdownload will download 417 files. A total of 20.98 MB
```

```
#-- Save an .RData file into the working folder, then delete all intermediate files
hg38_stemloop_data <- GDCprepare(query.mirna,
                                 save = T,
                                 save.filename = "hg38_mirna_stemloops.RData",
                                 directory = "GDCdata_hg38_stemloops",
                                 remove.files.prepared = TRUE
                                 )
#-- Load the .RData file
hg38_stemloop_RPMs <- get(load("hg38_mirna_stemloops.RData"))

dim(hg38_stemloop_RPMs)
# 1881 1252

hg38_stemloop_RPMs[1:5,1:4]
#   miRNA_ID     TCGA-FD-A6TE-01A-12R-A33A-13 TCGA-C4-A0F7-01A-11R-A085-13 TCGA-GU-AATO-01A-11R-A39B-13
# 1 hsa-let-7a-1                    9736.1845                     4526.388                    6204.6717
# 2 hsa-let-7a-2                    9634.7070                     4398.605                    6190.3701
# 3 hsa-let-7a-3                    9827.2697                     4500.592                    6289.6559
# 4   hsa-let-7b                   11958.4499                    11669.622                    7707.9844
# 5   hsa-let-7c                     205.8587                     1211.836                     908.9734
```