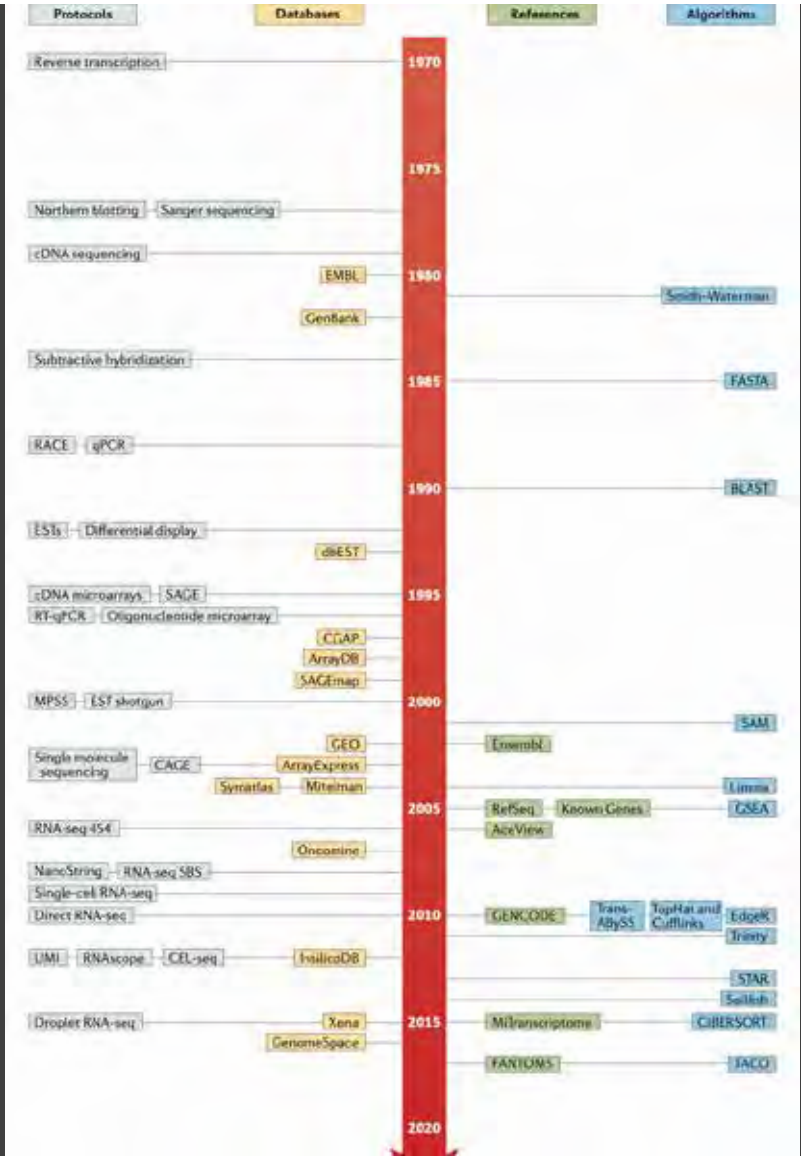
TRANSCRIPTOMICS

David W. Craig, Ph.D

USC University of Southern California

USC Institute Of Translational Genomics
Keck Medicine of USC

BACKGROUND


Protocols
Databases
References
Algorithms
Reverse transcription
1970
1975
Northern blotting
Sanger sequencing
cDNA sequencing
EMBL
1980
Smith–Waterman
GenBank
Subtractive hybridization
1985
FASTA
RACE
qPCR
1990
BLAST
ESTs
Differential display
dbEST
cDNA microarrays
SAGE
1995
RT-qPCR
Oligonucleotide microarray
CLAP
ArrayDB
SAGEmap
MPSS
EST shotgun
2000
SAM
GEO
Ensembl
Single molecule sequencing
CAGE
ArrayExpress
Limma
Syntrafas
Miteiman
2005
RefSeq
Known Genes
GSEA
RNA-seq 454
AceView
Oncomine
NanoString
RNA-seq SBS
Single-cell RNA-seq
Direct RNA-seq
2010
GENCODE
Trans-ABySS
TopHat and Cufflinks
EdgeR
Trinity
UMI
RNAscope
CEL-seq
InsilicoDB
STAR
Salkish
Dropler RNA-seq
Xena
2015
MiTranscriptome
CIBERSORT
GenomeSpace
FANTOM5
FACO
2020

Objective: Broad survey of RNA-Seq & Cancer

- Focus on breadth over depth
- Focus on methods focusing on tumor RNA-seq
  - Will not cover eQTL
- Focus on human applications and RNA
  - Epigenetics/mouse largely not covered

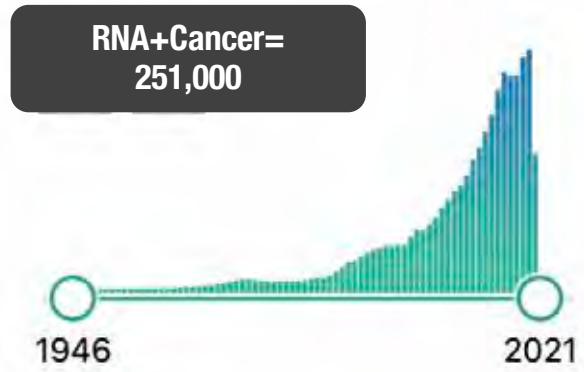
Transcriptomics:

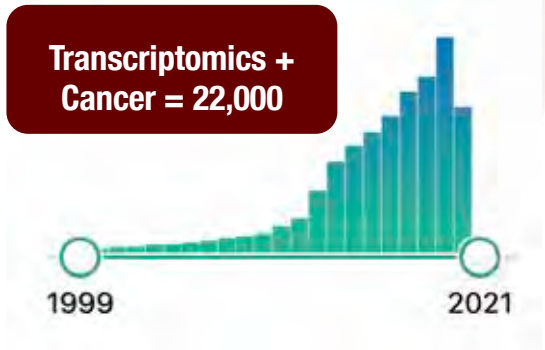
- Background & Core concepts w/ NGS

Applications

- Bulk Applications
- Emerging Methods

Basics of Analysis





Bound to disappoint those dedicated those who live transcriptomics

REVIEW PAPERS

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Translating RNA sequencing into clinical diagnostics: opportunities and challenges

Sara A. Byron[1], Kendall R. Van Keuren-Jensen[1], David M. Engelthaler[1], John D. Carpten[2] and David W. Craig[1]

Abstract | With the emergence of RNA sequencing (RNA-seq) technologies, RNA-based biomolecules hold expanded promise for their diagnostic, prognostic, and therapeutic applicability in various diseases, including cancers and infectious diseases. Detection of gene fusions and differential expression of known disease-causing transcripts by RNA-seq represent some of the most immediate opportunities. However, it is the diversity of RNA species detected through RNA-seq that holds new promise for the multi-faceted clinical applicability of RNA-based measures, including the potential of extracellular RNAs as non-invasive diagnostic indicators of disease. Ongoing efforts toward the establishment of benchmark standards, assay optimization for clinical conditions and demonstration of assay reproducibility are required to expand the clinical utility of RNA-seq.

RNA sequencing: the teenage years

Rory Stark[1], Marta Grzelak[1] and James Hadfield[2]*

Abstract | Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Cancer transcriptome profiling at the juncture of clinical translation

Marcin Cieślik[1,2] and Arul M. Chinnaiyan[1–3]

Abstract | Methodological breakthroughs over the past four decades have repeatedly revolutionized transcriptome profiling. Using RNA sequencing (RNA-seq), it has now become possible to sequence and quantify the transcriptional outputs of individual cells or thousands of samples. These transcriptomes provide a link between cellular phenotypes and their molecular underpinnings, such as mutations. In the context of cancer, this link represents an opportunity to dissect the complexity and heterogeneity of tumours and to discover new biomarkers or therapeutic strategies. Here, we review the rationale, methodology and translational impact of transcriptome profiling in cancer.

ANNUAL REVIEWS

Genomic basis for RNA alterations in cancer

https://doi.org/10.1038/s41586-020-1970-0
Received: 29 March 2018
Accepted: 11 December 2019
Published online: 5 February 2020
Open access

PCAWG Transcriptome Core Group[1,96], Claudia Calabrese[2,96], Natalie R. Davidson[3,4,5,6,7,96], Deniz Demircioğlu[8,9,96], Nuno A. Fonseca[2,96], Yao He[10,96], André Kahles[3,4,6,7,96], Kjong-Van Lehmann[3,4,6,7,96], Fenglin Liu[10,96], Yuichi Shiraishi[11,96], Cameron M. Soulette[12,96], Lara Urban[2,96], Liliana Greger[2], Siliang Li[13,14], Dongbing Liu[15,16], Marc D. Perry[15,16], Qian Xiang[15], Fan Zhang[10], Junjun Zhang[10], Peter Bailey[17], Serap Erkek[18], Katherine A. Hoadley[19], Yong Hou[13,14], Matthew R. Huska[20], Helena Kilpinen[21], Jan O. Korbel[18], Maximilian G. Marin[13], Julia Markowski[20], Tannistha Nandi[8], Qiang Pan-Hammarström[22], Chandra Sekhar Pedamallu[23,24,25], Reiner Siebert[26], Stefan G. Stark[3,4,6,7], Hong Su[13,14], Patrick Tan[8,25], Sebastian M. Waszak[18], Christina Yung[18], Shida Zhu[13,14], Philip Awadalla[18,24], Chad J. Creighton[27], Matthew Meyerson[23,24,25], B. F. Francis Ouellette[28], Kui Wu[29,30], Huanming Yang[29], PCAWG Transcriptome Working Group[1], Alvis Brazma[2,96], Angela N. Brooks[11,15,16,96,97], Jonathan Göke[8,31,96], Gunnar Rätsch[3,4,6,7,96,97], Roland F. Schwarz[2,32,33,34], Oliver Stegle[2,18,35,36], Zemin Zhang[10,96] & PCAWG Consortium[96]

Annual Review of Cancer Biology

Deciphering Human Tumor Biology by Single-Cell Expression Profiling

Itay Tirosh[1] and Mario L. Suvà[2,3]

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 7610001, Israel; email: itay.tirosh@weizmann.ac.il

[2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA

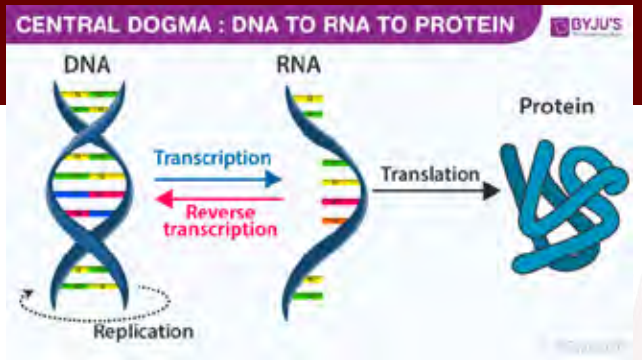
[3]Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA; email: mario_suva@mgh.harvard.edu

Next-generation computational tools for interrogating cancer immunity

Francesca Finotello, Dietmar Rieder, Hubert Hackl and Zlatko Trajanoski*

Abstract | The remarkable success of cancer therapies with immune checkpoint blockers is revolutionizing oncology and has sparked intensive basic and translational research into the mechanisms of cancer–immune cell interactions. In parallel, numerous novel cutting-edge technologies for comprehensive molecular and cellular characterization of cancer immunity have been developed, including single-cell sequencing, mass cytometry and multiplexed spatial cellular phenotyping. In order to process, analyse and visualize multidimensional data sets generated by these technologies, computational methods and software tools are required. Here, we review computational tools for interrogating cancer immunity, discuss advantages and limitations of the various methods and provide guidelines to assist in method selection.

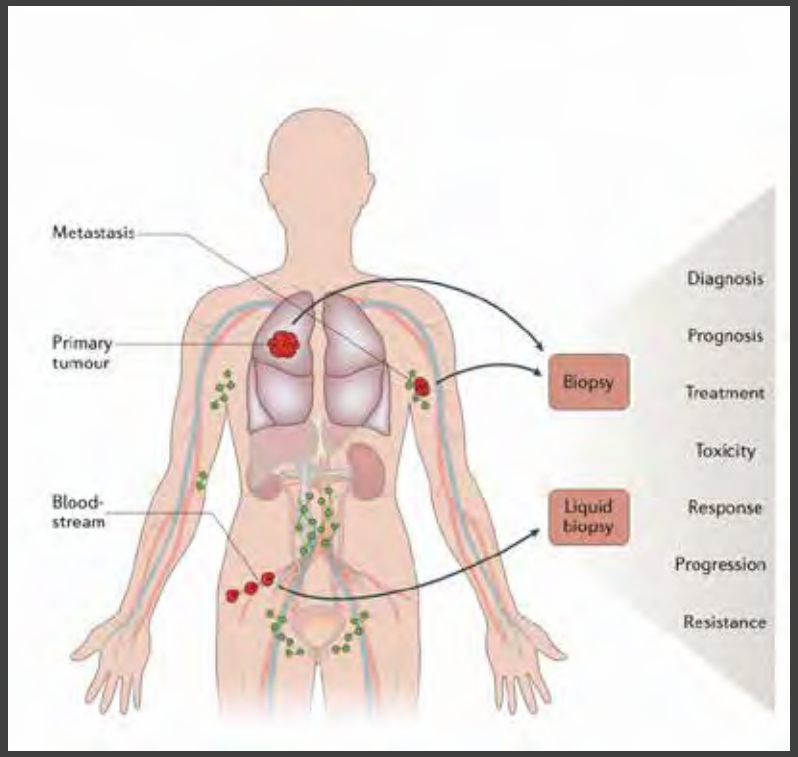
WITH IMPORTANT EXCEPTIONS…



- … you are diploid with a maternal and paternal copy
- … you have two copies of 22 chromosomes plus X and sometimes Y
- … there are four nucleotides (A, T, C, G), about 3 billion bases long (ATTATA..)
- … a copy of your genome is every cell.
- … there are 4 millions genetic variants between two people (3 billion)
- …Variants are of different types. Single nucleotide substitutions (SNVs), Insertions/Deletions (indels), Structural Variants (inversions, duplications, translocations)
- … most genetic variants are not functional. There are many variants in genes like BRCA1, having a variant does not mean you carry the BRCA1 gene.
- … changes occurring in a specific tissue or cell during our life are called somatic events
- … SNPs $\in$ SNVs. SNPs are inherited. Polymorphisms are common in population & you weren't born with cancer.
- … 1% of your genome is coded in genes, sometimes this is called your exome
- … in genes, DNA is transcribed to RNA, RNA is translated to proteins
- … genes are frequently transcribed as exons broken by introns, where the introns on spliced out of mRNA
- … a considerable number of modifications can occur to proteins (e.g. phosphorylation)
- … 99% of your genome we don't understand, but we all recognize its important.
- … two identically cloned calico cats look nothing alike because epigenetics matters

The exceptions are often the most important aspects of understand and treating diseases.

Why Transcriptomics Matters In Cancer

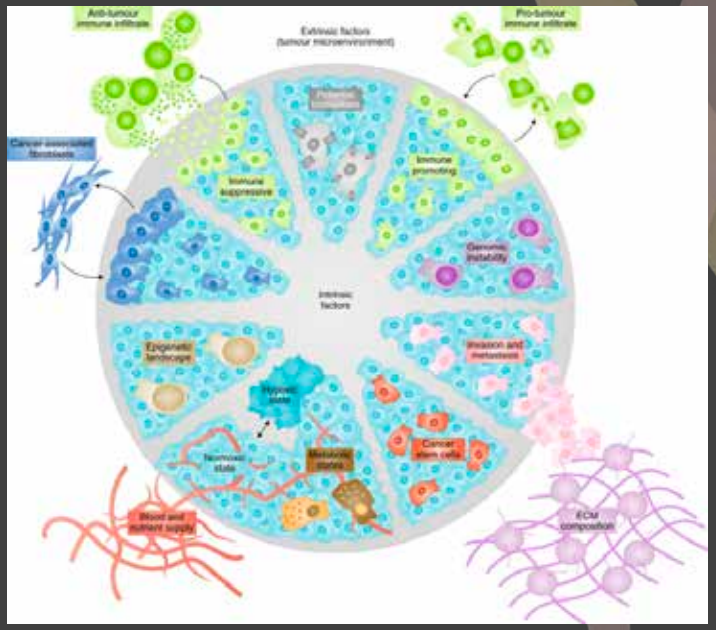


Tumour heterogeneity and metastasis at single-cell resolution

Interpretation of functional impact of DNA variation

Provide possible biomarkers for diagnosis or progression

Give insight into biological drivers, response, therapies



HUMAN GENOME PROJECT

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We

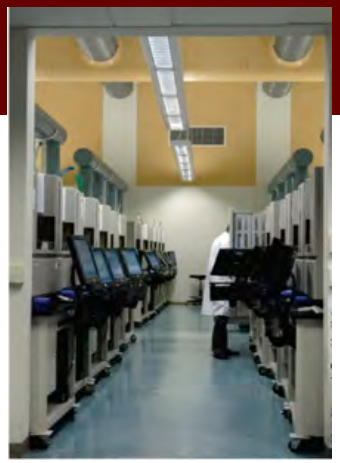
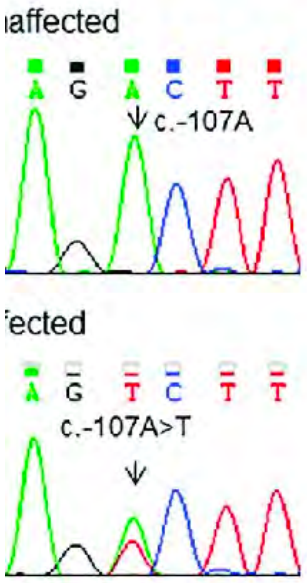


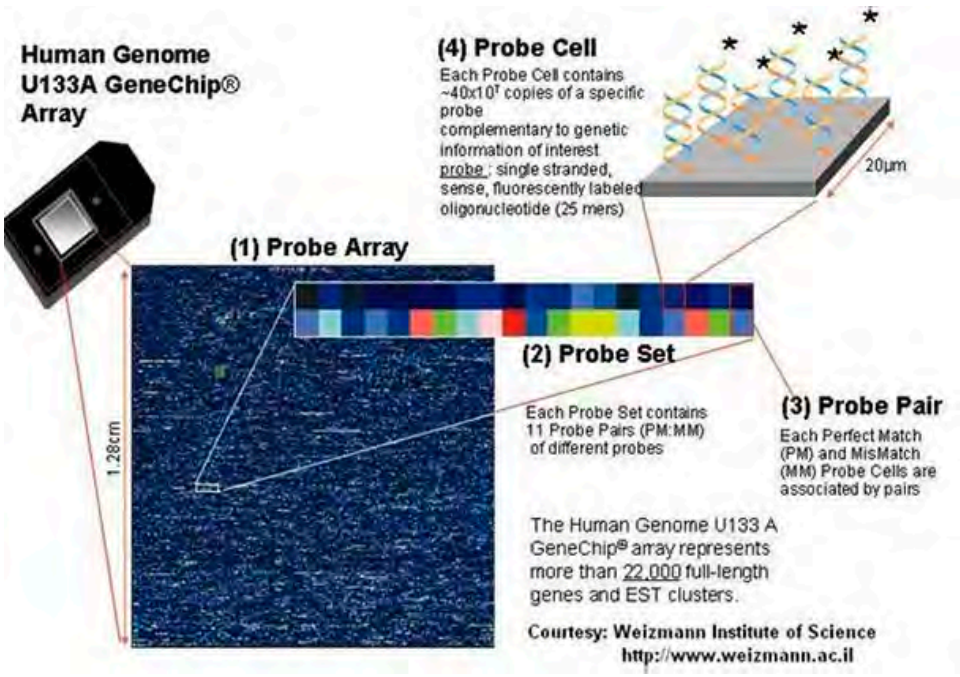
Table 8 Chromosome size estimates

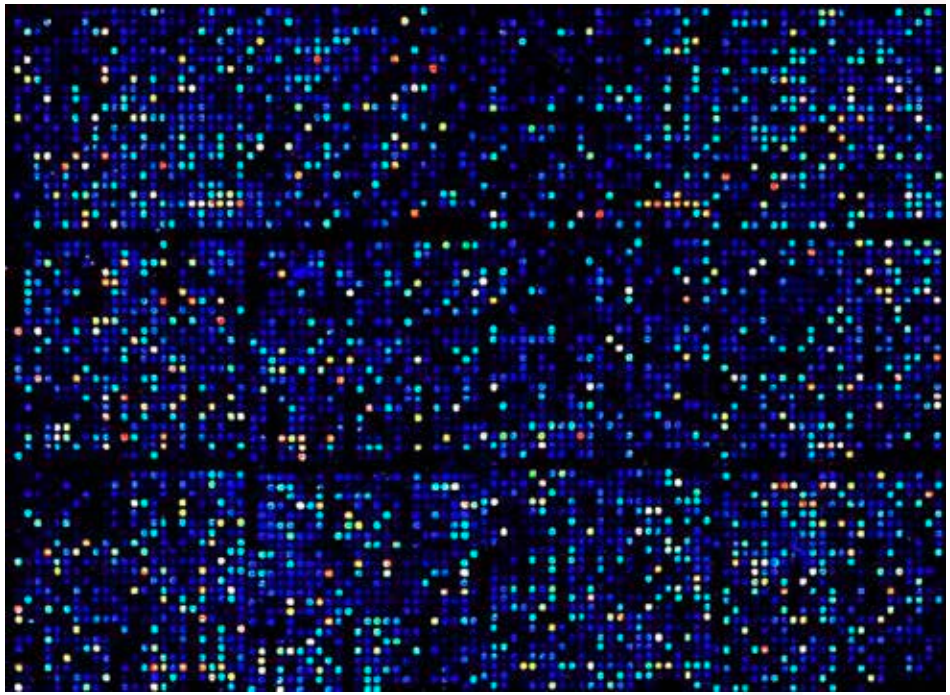
Chromosome*	Sequenced bases† (Mb)	FCC gaps‡		SCC gaps‖		Sequence gaps#		Heterochromatin and short arm adjustments**(Mb)	Total estimated chromosome size (including artefactual duplication in draft genome sequence)†† (Mb)	Previously estimated chromosome size‡‡ (Mb)
		Number	Total bases in gaps§ (Mb)	Number	Total bases in gaps¶ (Mb)	Number	Total bases in gaps# (Mb)			
All	2,692.9	897	152.0	4,076	142.7	145,514	80.6	212	3,289	3,286
1	212.2	104	17.7	347	12.1	11,803	6.5	30	279	263
2	221.6	50	8.5	296	10.4	12,880	7.1	3	251	255
3	186.2	71	12.1	336	11.8	14,689	8.1	3	221	214
4	168.1	39	6.6	343	12.0	12,768	7.1	3	197	203
5	169.7	46	7.8	337	11.8	10,304	5.7	3	198	194
6	158.1	15	2.6	275	9.6	5,225	2.9	3	176	183
7	146.2	27	4.6	195	6.8	4,338	2.4	3	163	171
8	124.3	41	7.0	249	8.7	8,692	4.8	3	148	155
9	106.9	19	3.2	122	4.3	6,083	3.4	22	140	145
10	127.1	14	2.4	163	5.7	8,947	5.0	3	143	144
11	128.8	29	4.9	193	6.8	8,279	4.6	3	148	144
12	124.5	26	4.4	168	5.9	8,226	4.6	3	142	143
13	92.9	12	2.0	115	4.0	5,065	2.8	16	118	114
14	86.9	13	2.2	40	1.4	775	0.4	16	107	109
15	73.4	18	3.1	104	3.6	5,717	3.2	17	100	106
16	73.1	55	9.4	102	3.6	4,757	2.6	15	104	98
17	72.8	41	7.0	95	3.3	4,261	2.4	3	88	92
18	72.9	22	3.7	113	4.0	4,324	2.4	3	86	85
19	55.4	49	8.3	108	3.8	2,344	1.3	3	72	67
20	60.5	7	1.2	33	1.2	469	0.3	3	66	72
21	33.8	4	0.1	0	0.0	0	0.0	11	45	50
22	33.8	10	1.0	0	0.0	0	0.0	13	48	56
X	127.7	141	24.0	182	6.4	4,282	2.4	3	163	164
Y	21.8	6	1.0	19	0.7	113	0.1	27	51	59
NA	5.1	0	0	134	0.0	577	0.3	0	0	0
UL	9.3	38	0	7	0.0	566	0.3	0	0	0



Another Form of Analog Measurements: Arrays

Quantification of complementary transcripts (25-120bp)



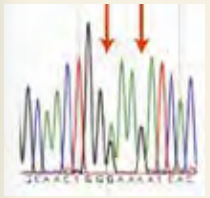


Engineering New Measurements

Traditional Sanger Sequencing: 1985+

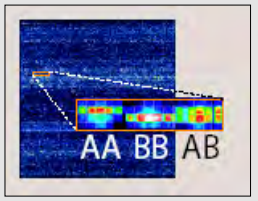
- Engineering improvements (capillaries/dyes)
- Consensus of billions of molecules

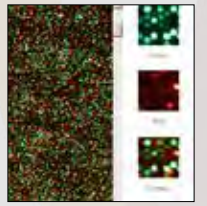

Identification of genetic variants using bar-coded multiplexed sequencing



Microarrays: 1995+

- Detection of gene variation by hybridization
- Consensus of billions of molecules





Single molecule sequencing: 2007+

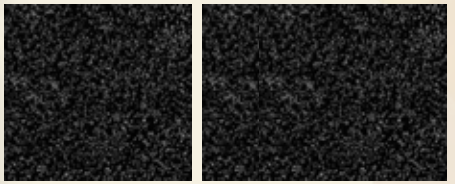
- Each re
  clonally
- Millions
  (lawn-sequencing)


Resolving Individuals Contributing Trace Amounts DNA to Highly Complex Mixtures Using High-Den SNP Genotyping Microarrays


Protecting Aggregate Genomic Data

Article Info & Metrics eLetters PDF

A paper published recently in *PLOS Genetics* (1) describes a statistical method for resolving individual genotypes within a mix of DNA samples or data sets containing aggregate single-nucleotide polymorphism data. This scientific advance may have important implications for forensics and for genome-wide association studies (GWAS). It has also changed our understanding of the risks of making aggregate genomic data publicly available. While we assess the broader scientific, ethical, and policy implications of this development, NIH has moved swiftly to remove aggregate genomic data from our publicly available Web sites. Further information about changes in NIH open-access policies for GWAS is available on the NIH's GWAS Web site (2)



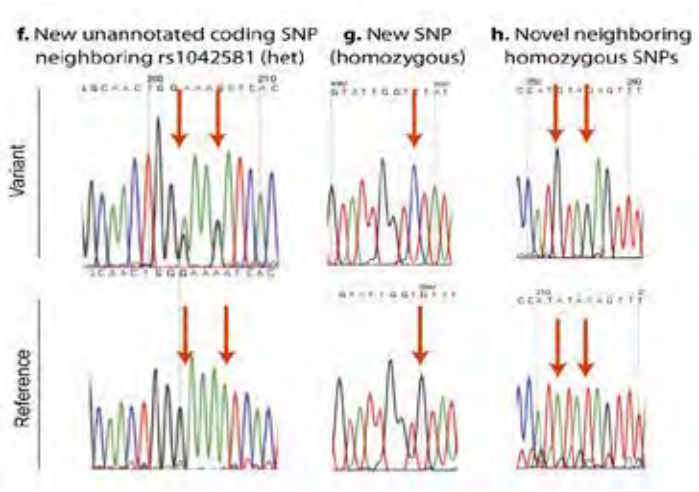
Single cell sequencing: 2012+

- Full RNA-seq on single cells in soluti
- Emerging spatial genomic

Beyond Transcript Abundance By Integration of DNA

Allele specific expression, non-sense mediated decay, PSI, intron inclusion, and more

KEY PRINCIPLES THAT YOU MUST KNOW



Pseudo-single molecule reads

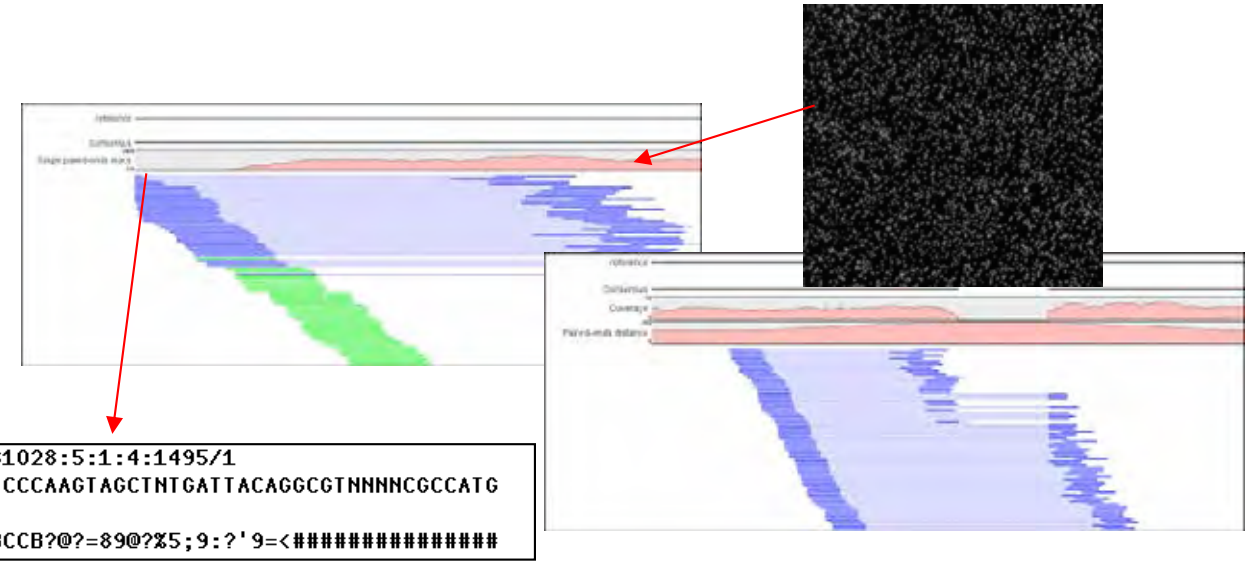
- A heterozygous SNP will give the paternal or maternal allele in a single read, not both

Paired-Reads

- First 100 bases and last 100 bases of a ~500bp DNA molecule

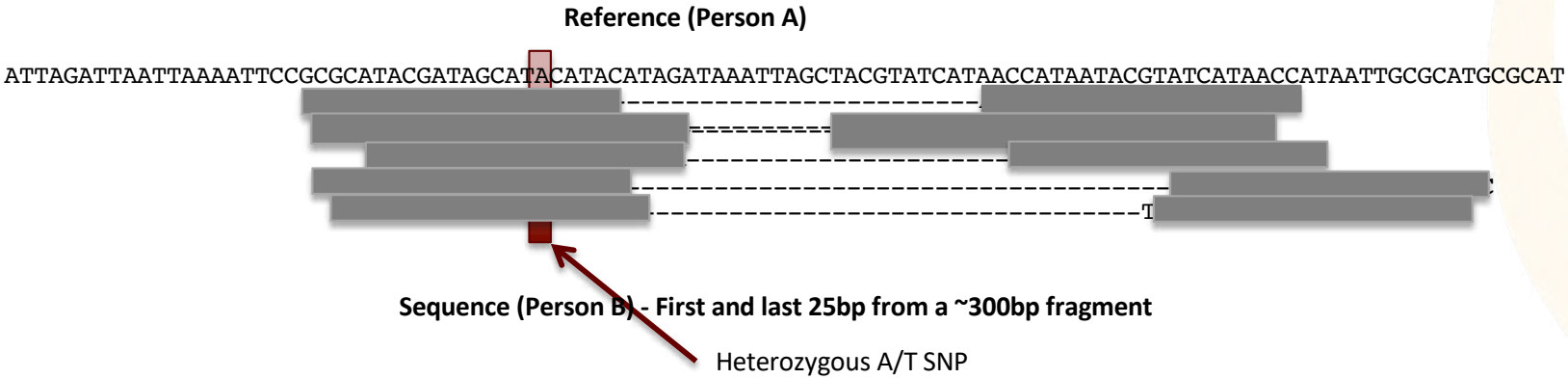
Billions of reads in a sequencing run

- Sampling matters and is how we control error



```
@unknown:81028:5:1:4:1495/1
GCTTCAGCGTCCCAAGTAGCTNTGATTACAGGCGTNNNNCGCCATG
+
BCCCCCBCCBCCB?@?=89@?%5;9:?'9=<################
```

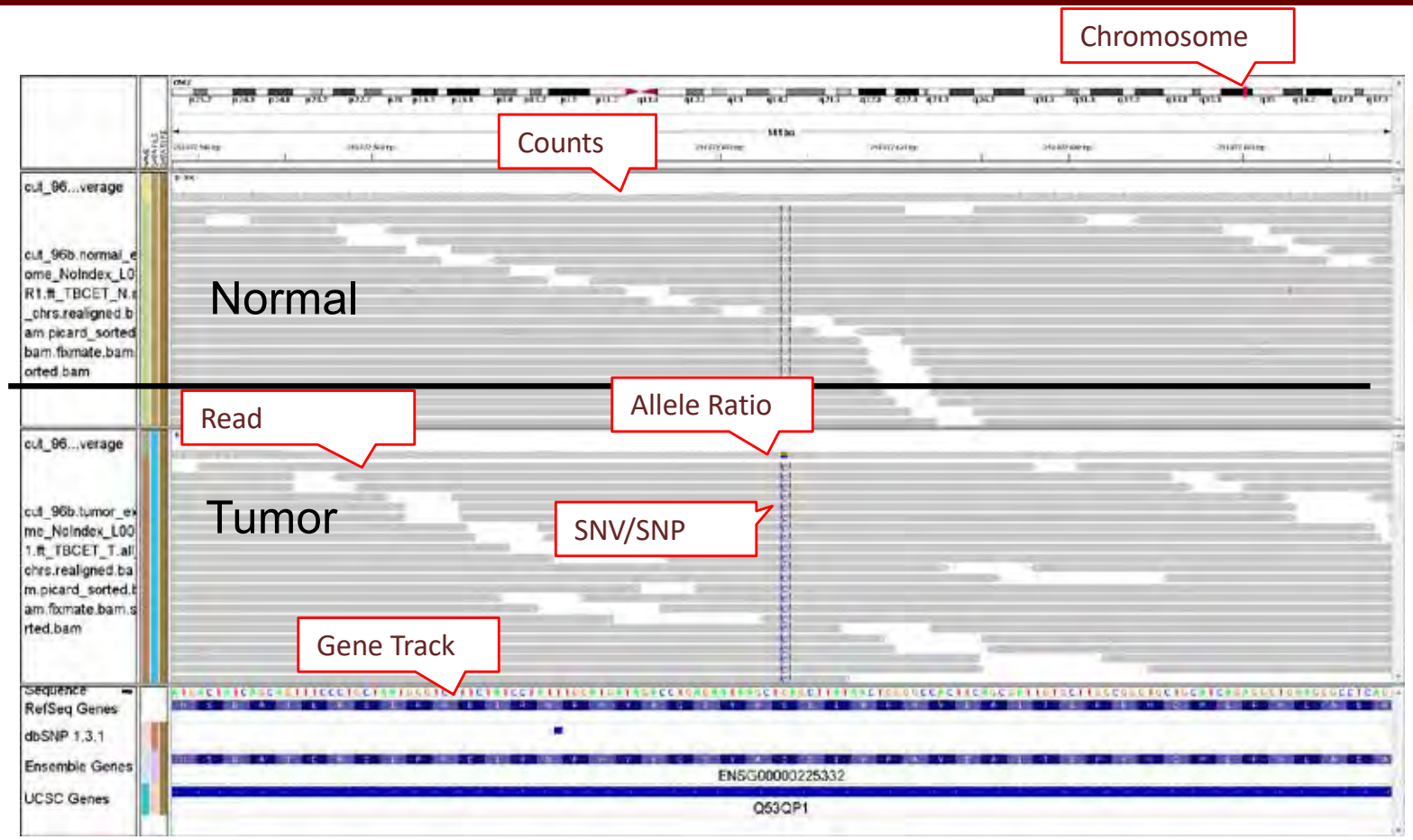
Concept of NGS Sequence Analysis



Steps to remember:

1. Alignment (produces BAM file)
2. Variant Calling (produces VCF file)
3. Interpretation (produces powerpoint)

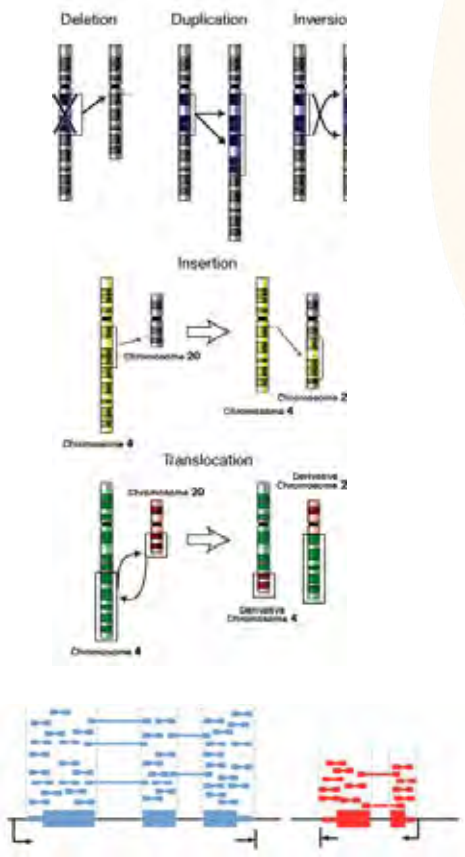
VARIANTS: EXAMPLE



Next Generation Sequencing

Quantum Measurement of Molecular Variation

- **Point Mutations** – Single Nucleotide Variation (SNVs) & Small Insertions/mutations (Indels)
- **Copy Number** – Changes in abundance – both DNA/RNA
- **Rearrangements** –Translocations and Structural variants via read mapping
- **Transcriptional Profiling** – Abundance, exon level, isoform level, and study splicing defects



VISUALIZING TUMOR / REFERENCE IN IGV

Quantification/Abundance/Differential Expression

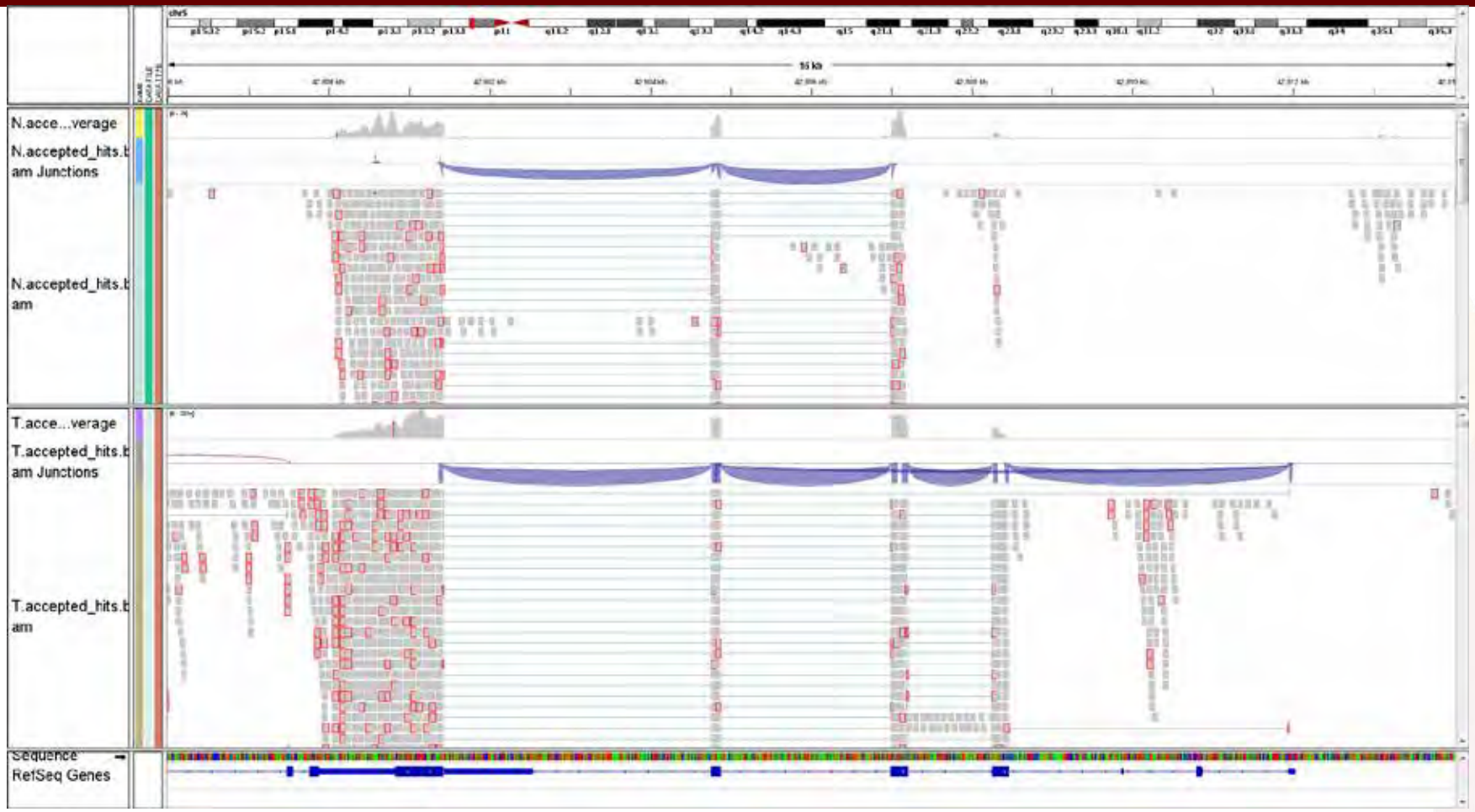
Abundance:

- FPKM: Fragments Per Exon Kilobase of Sequence Per Million Reads
  - Some genes are longer then other genes and they get counted more
- TPM: For every 1,000,000 **RNA** molecules in the **RNA-seq** sample, x came from this gene/transcript
- Transcripts? Why the use of genes….

Differential Expression Should Sot be done using Abundance

- Let's say you have erythrocytes higher in 1 sample, adding lots of globin
  - HBB: TPM in sample A is 600K
  - HBB: TPM in sample B is 300K.
    - Because TPM is fractional, all TPMs are lower in Sample B.
    - You need to normalize before differential expression!
    - You need the count level data to addresss these types of issues.

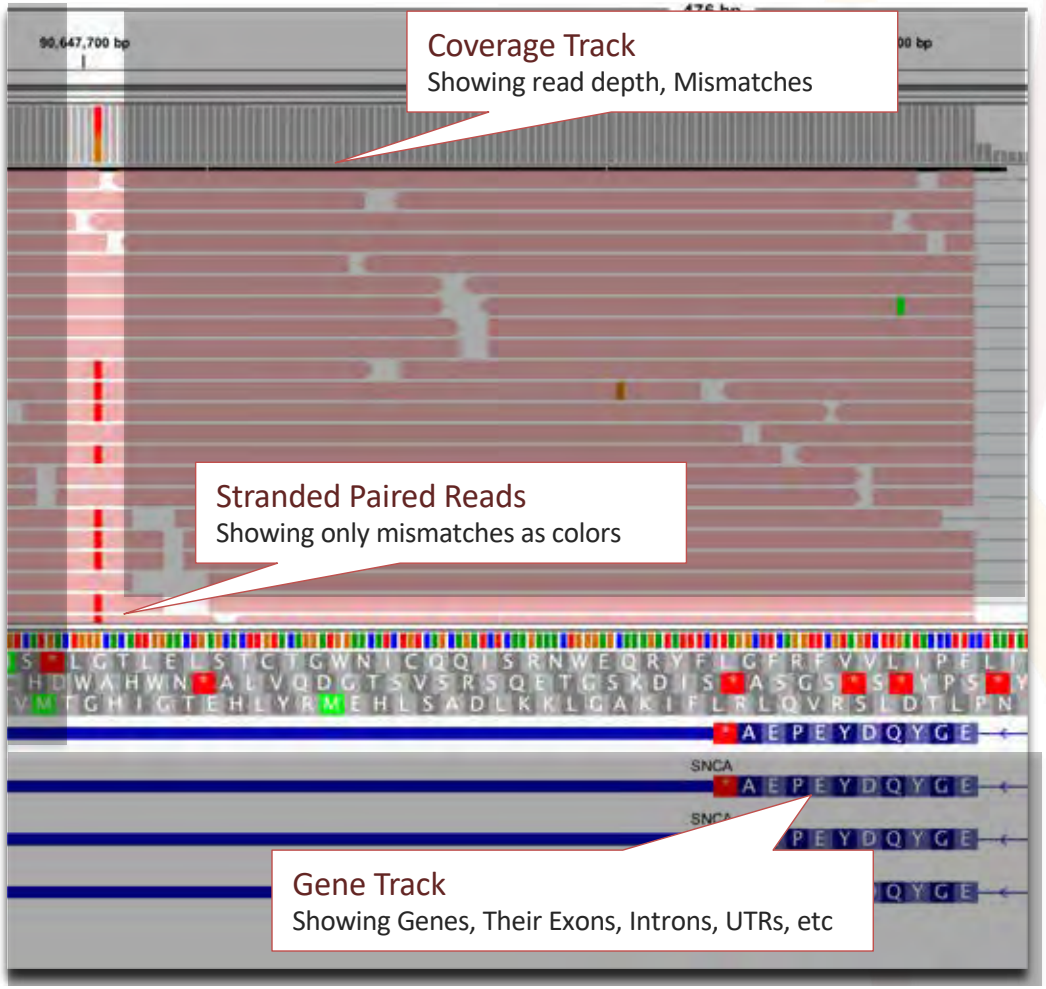
RNA


chr5
15 kb
N.acce...verage
N.accepted_hits.b
am Junctions
N.accepted_hits.b
am
T.acce...verage
T.accepted_hits.b
am Junctions
T.accepted_hits.b
am
Sequence
RefSeq Genes

TRANSCRIPTOME: BEYOND QUANTIFICATION

– Raw Data View:

- Pre-spliced, Spliced, Strandedness
- Allele counts, etc.



TRANSCRIPTOME: BEYOND QUANTIFICATION

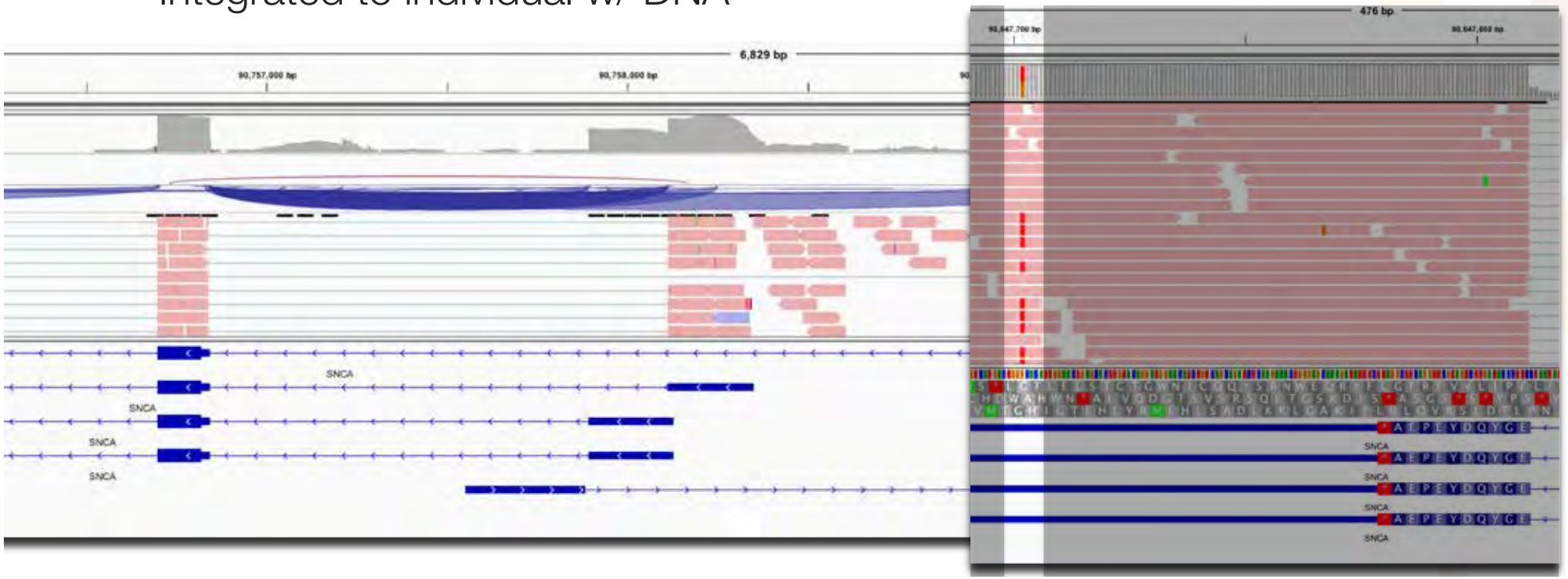
- Transcript Quantification (ZOOM OUT)
  - Pre-spliced, Spliced, Strandedness

Transcriptome: Beyond Quantification

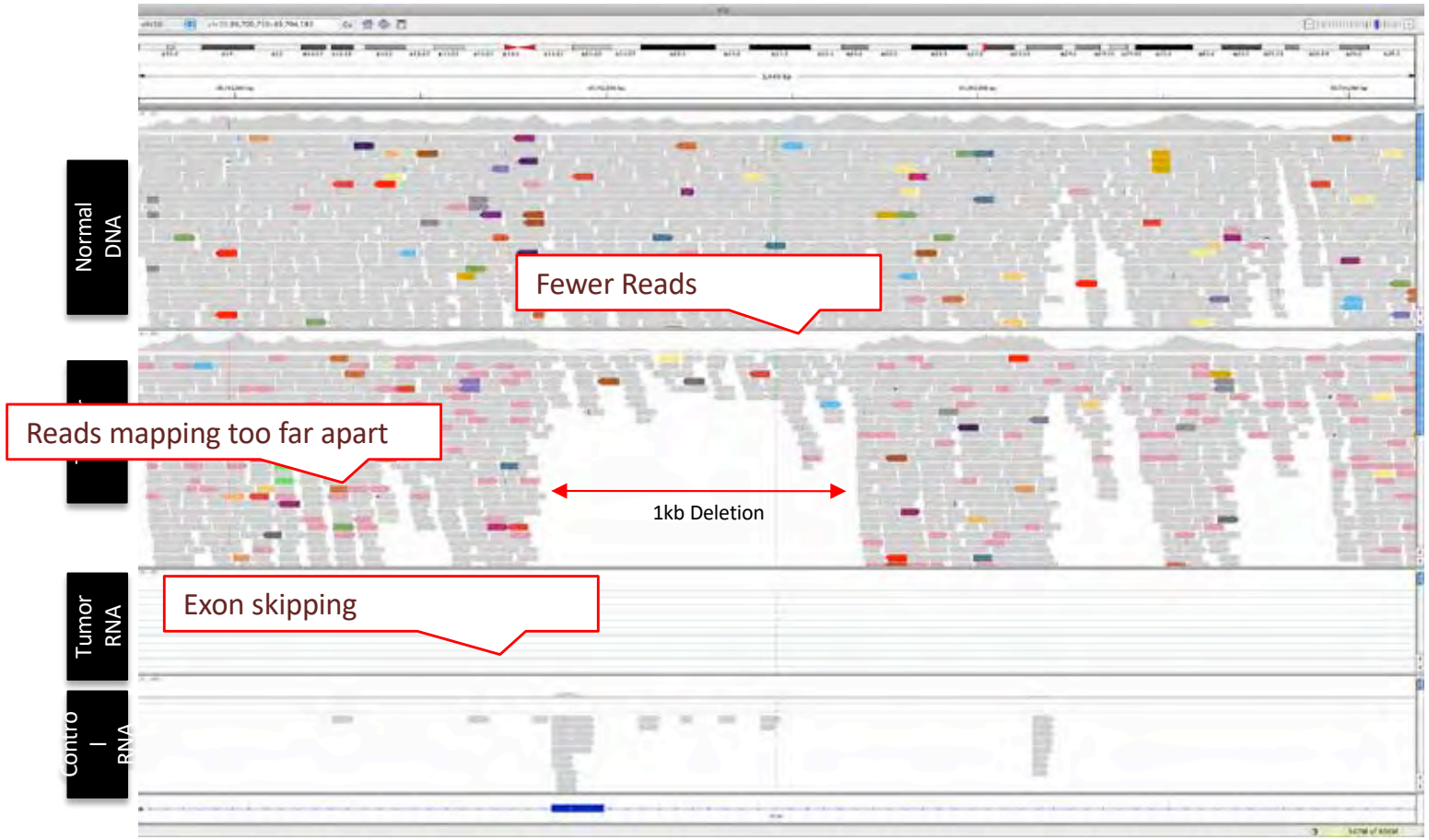
- Junctions/Isoforms
  - Alternative start-sites
  - Integrated to individual w/ DNA

SNPs

- Allele specific expression
- Non-sense mediated decay, eQTLs

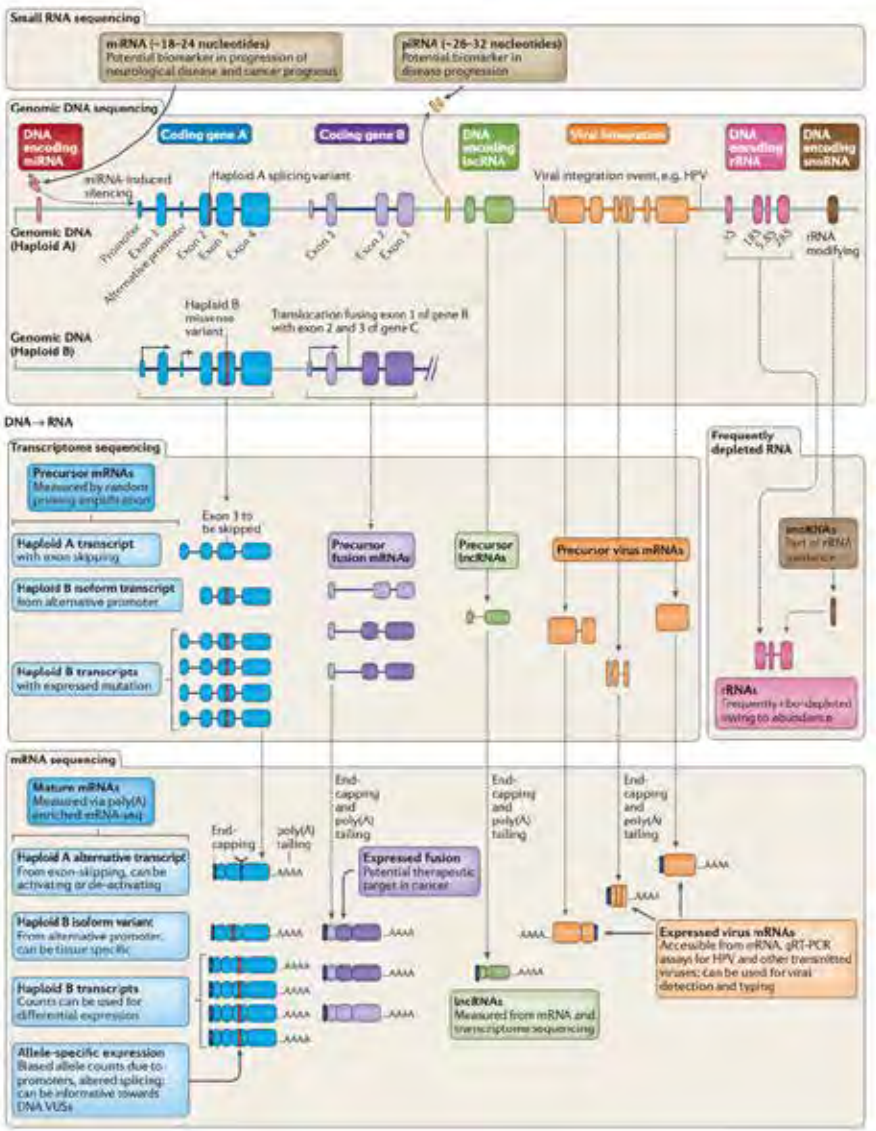


NGS VARIANTS



BEYOND THE CODING

Human Variation At the RNA Level

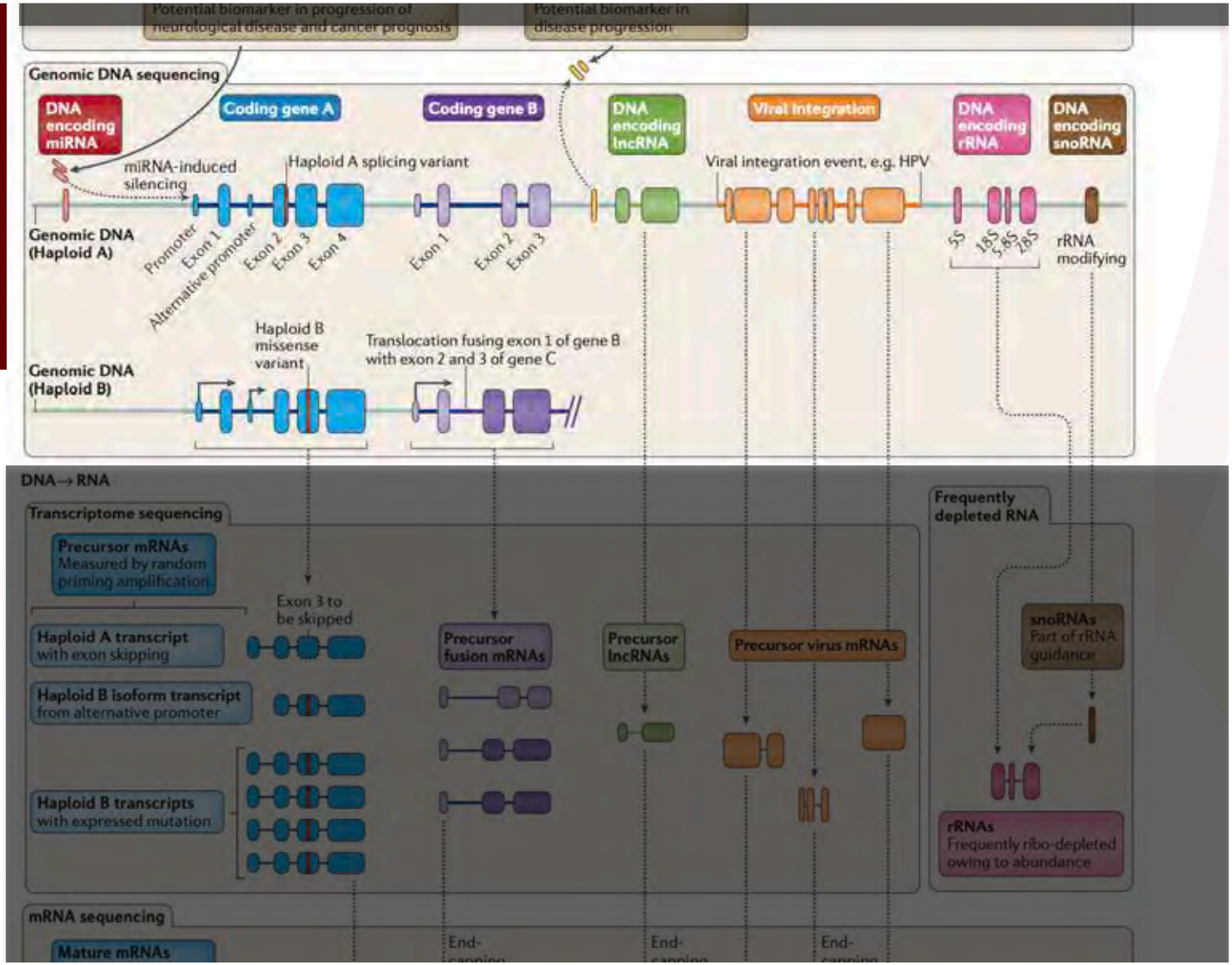


APPLICATIONS OF NEXT-GENERATION SEQUENCING

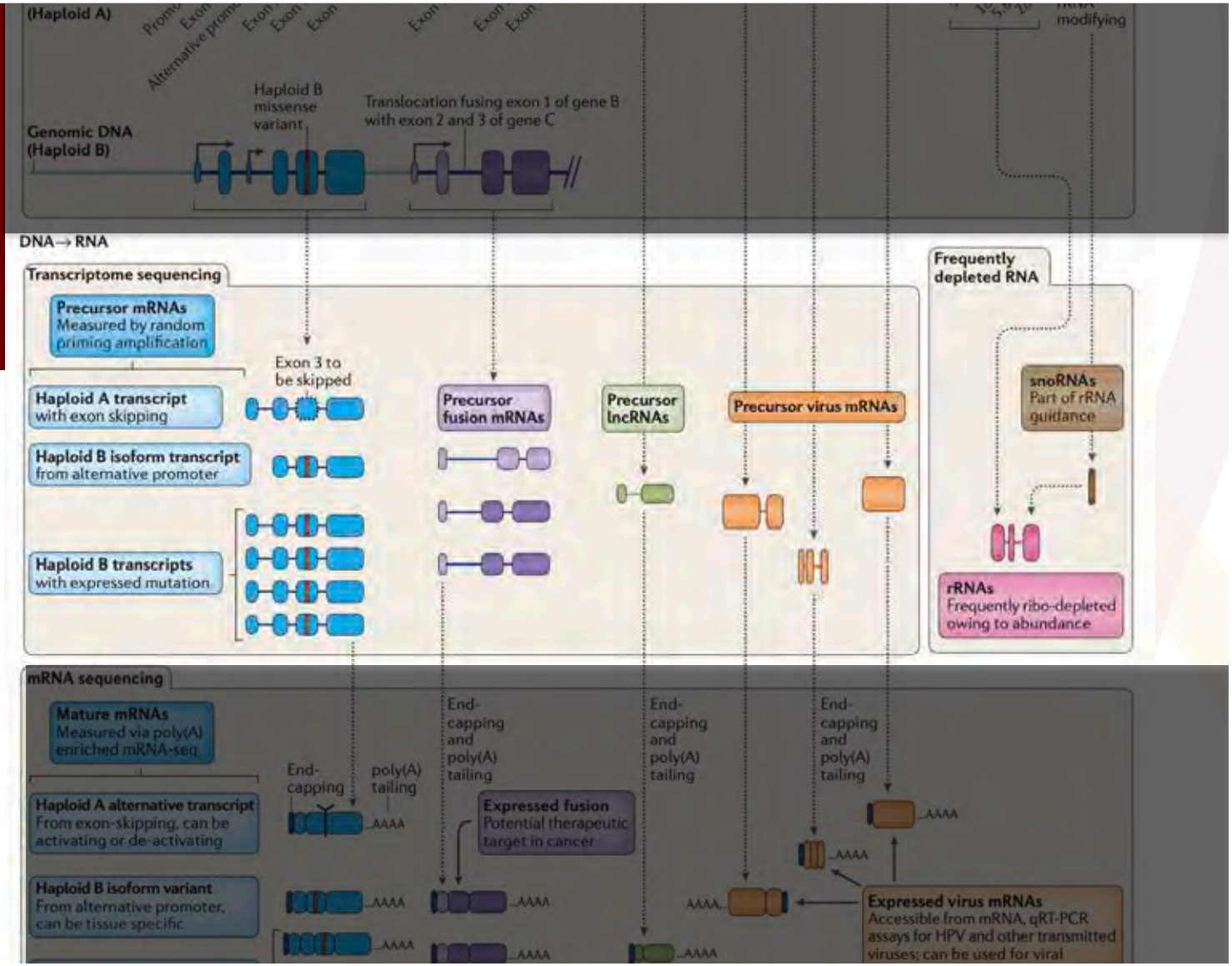
Translating RNA sequencing into clinical diagnostics: opportunities and challenges

Sara A. Byron[1], Kenneth R. Van Keuren-Jensen[2], David M. Engelthaler[2], John P. Carey and David W. Craig[1]

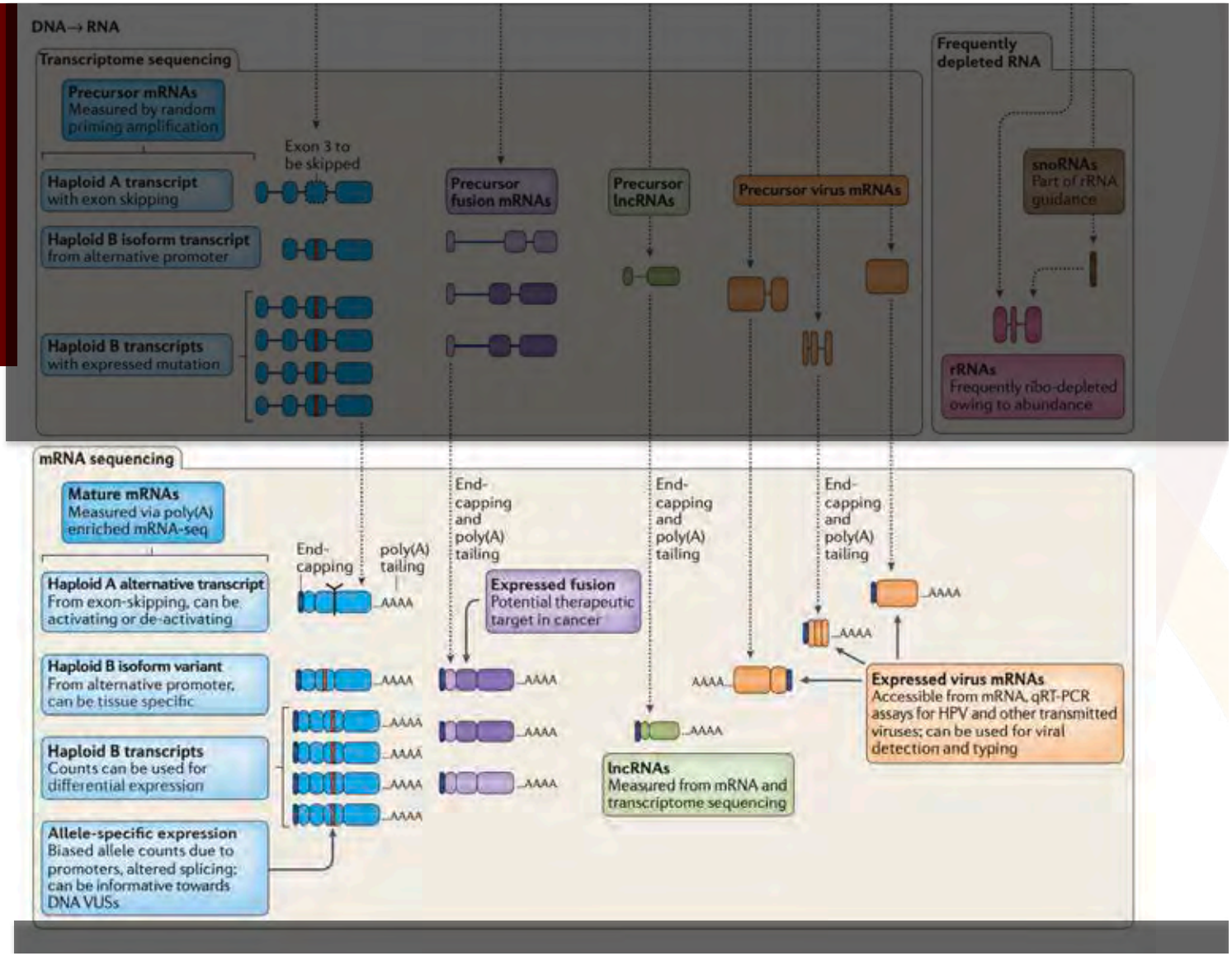
Human Variation At the RNA Level



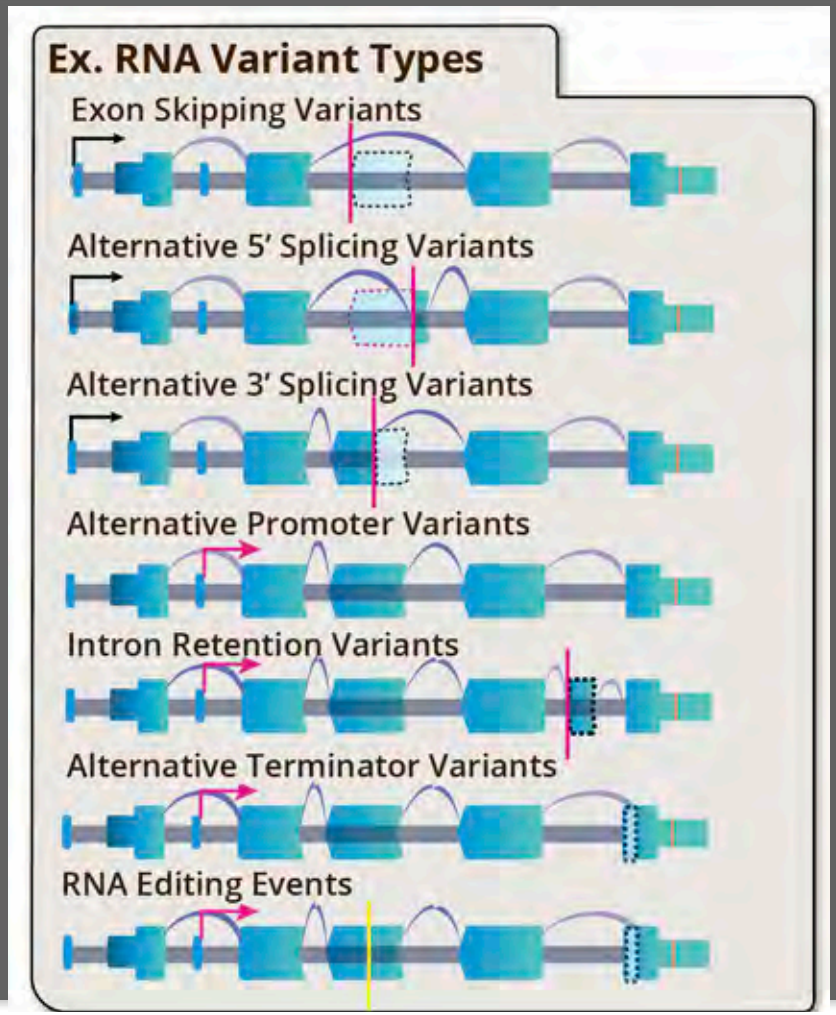
Human Variation At the RNA Level



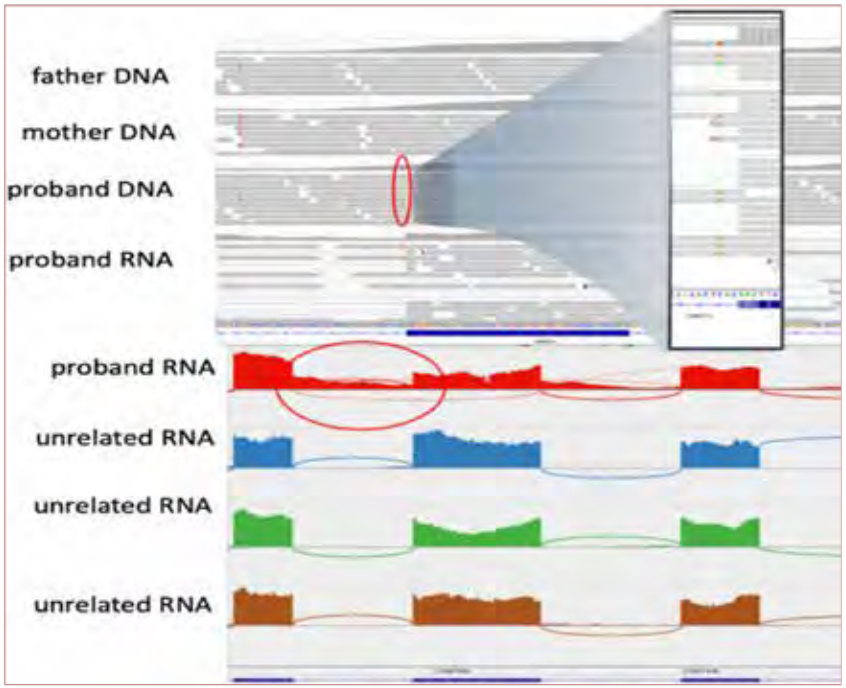
Human Variation At the RNA Level



TRANSCRIPTOME: SNPS



INTRON INCLUSION



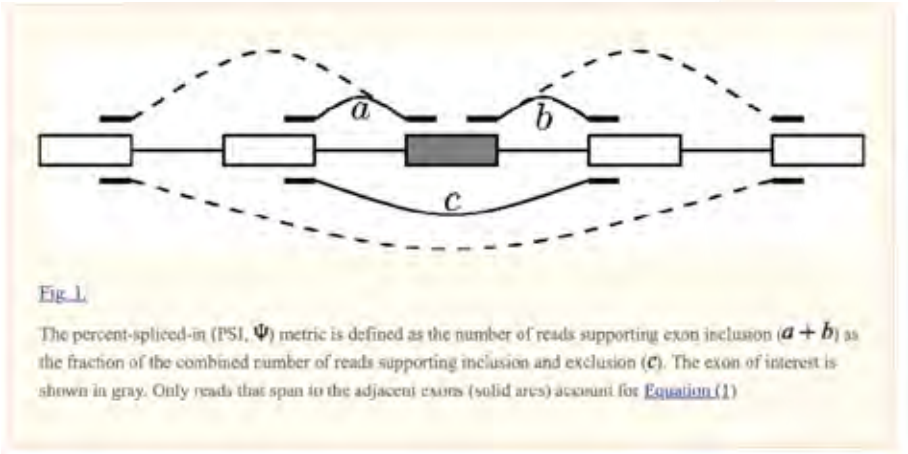


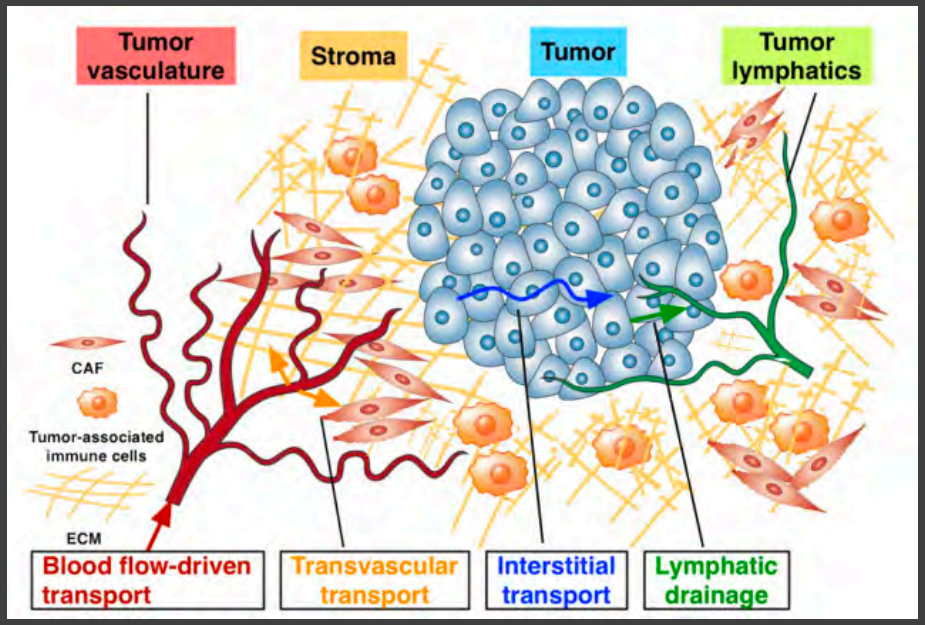
Fig 1.

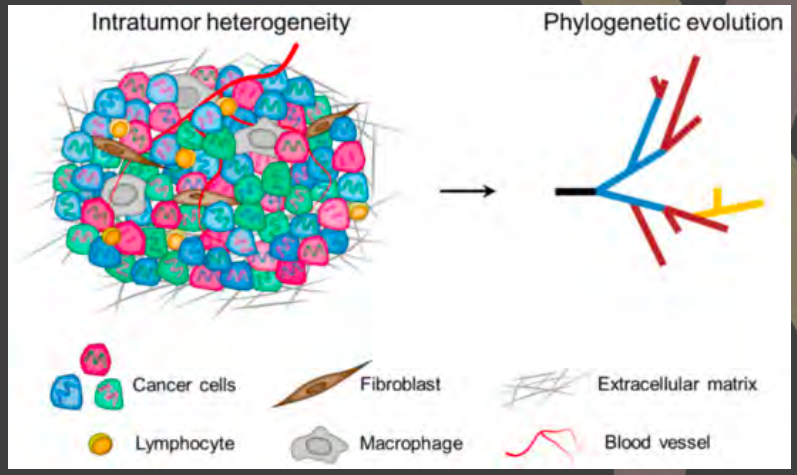
The percent-spliced-in (PSI, Ψ) metric is defined as the number of reads supporting exon inclusion ($a + b$) as the fraction of the combined number of reads supporting inclusion and exclusion (c). The exon of interest is shown in gray. Only reads that span to the adjacent exons (solid arcs) account for Equation (1)

$$\Psi = \frac{a+b}{a+b+2c}$$

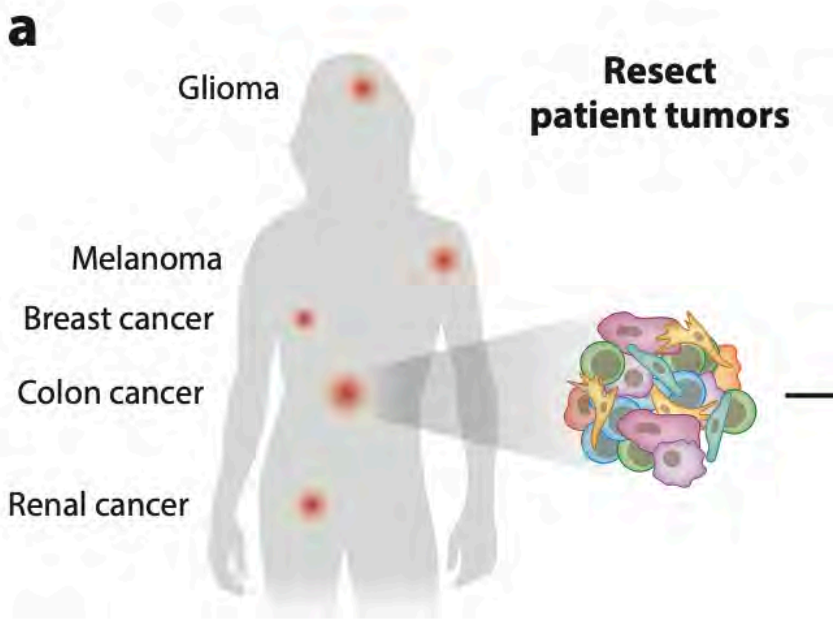
Bioinformatics. 2013 Jan 15; 29(2): 273–274

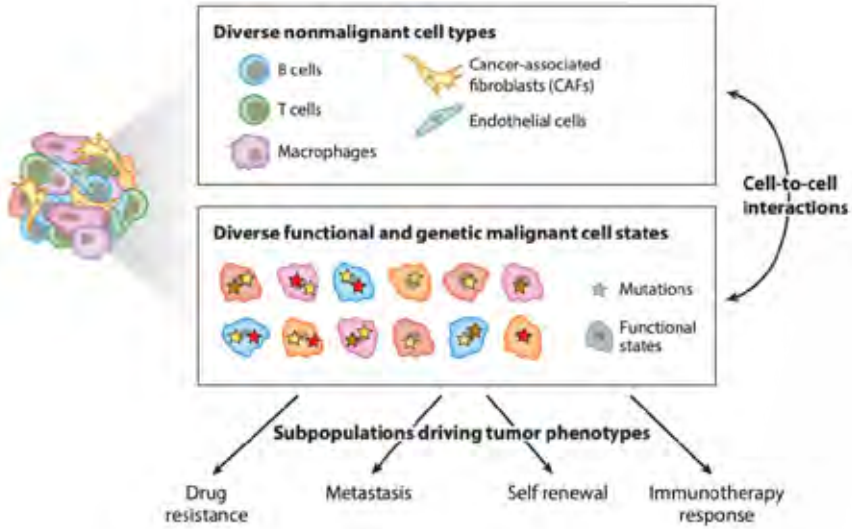
CANCER TRANSCRIPTOMICS



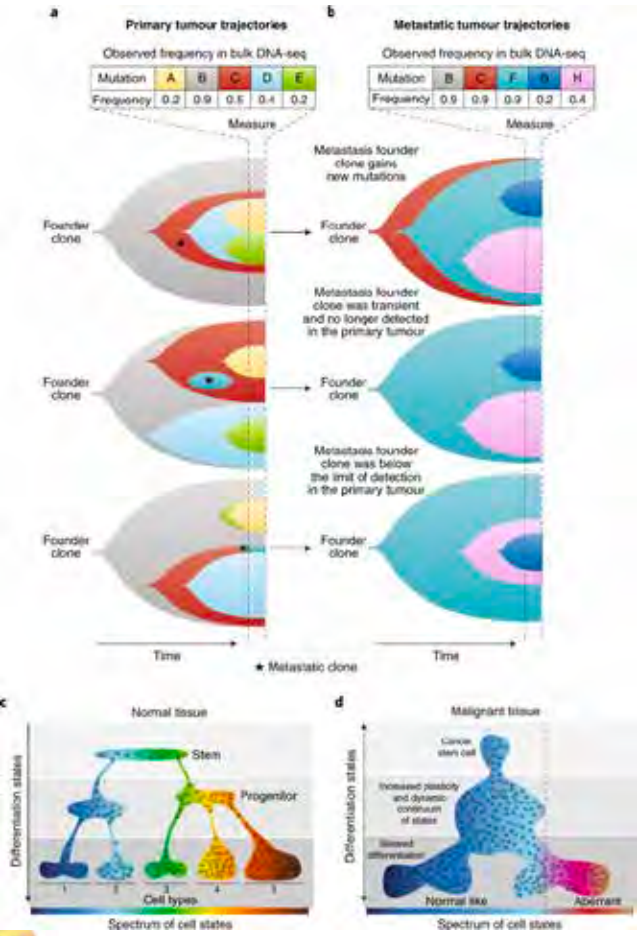


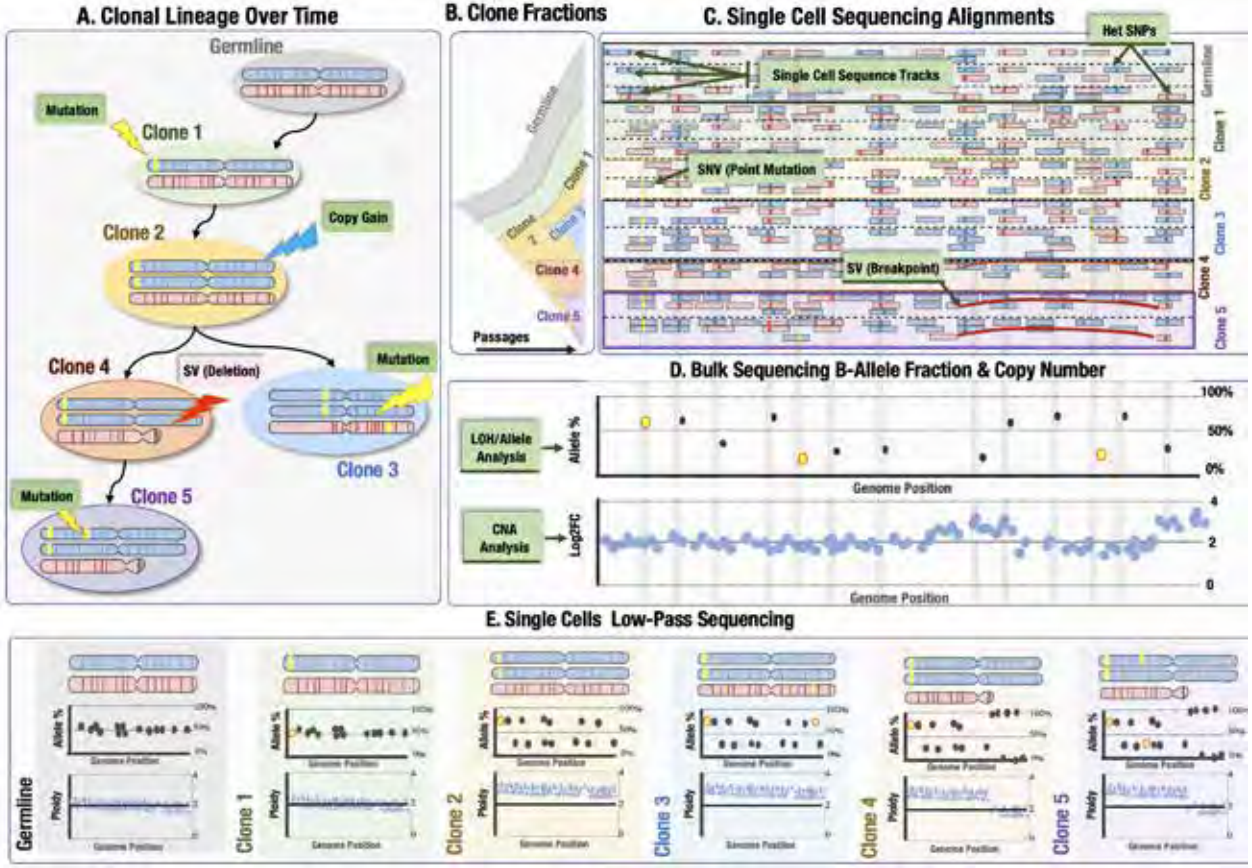
TUMOR BIOLOGY IS DRIVEN BY SUB-POPULATIONS AND HETEROGENEIGHTY





CANCER IS A MIXTURE OF TUMOR AND HEALTHY CELLS





EASY EXAMPLE: FUSIONS

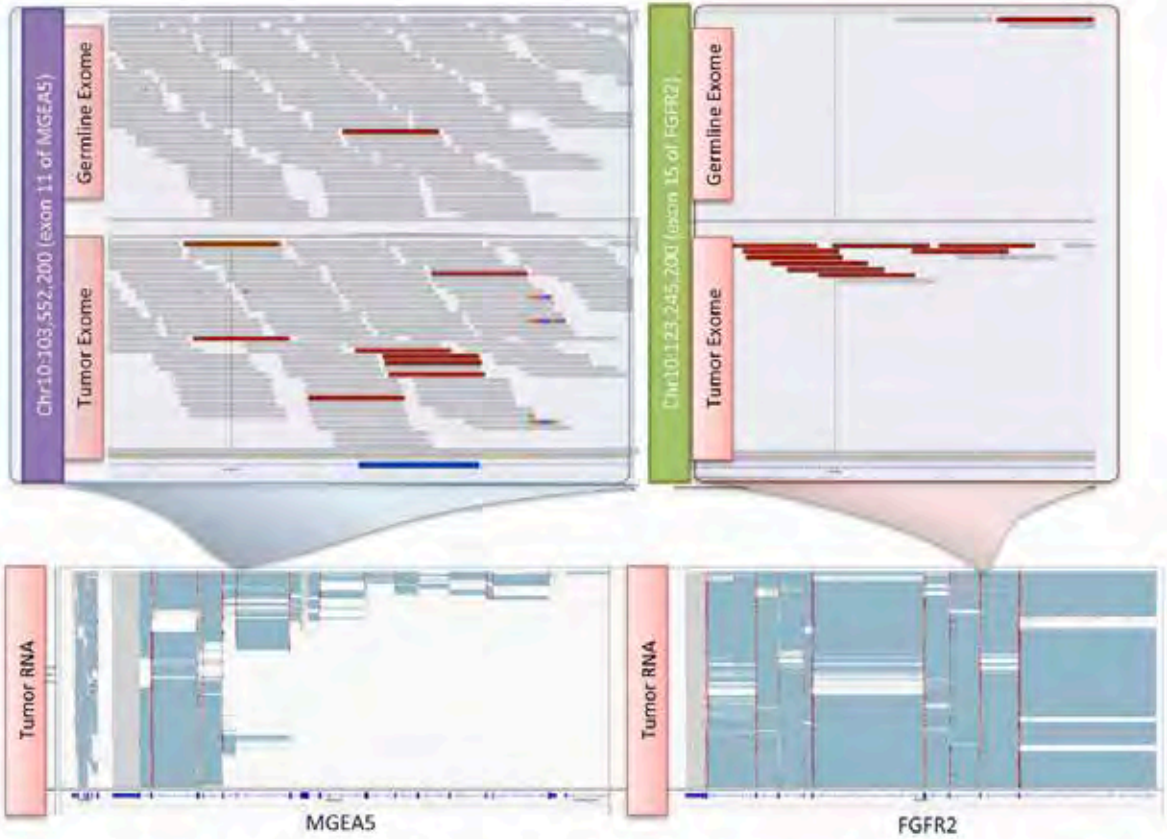


Figure 4. Visualisation of FGFR2-MGEA5 fusion in the Integrative Genomics Viewer (IGV).

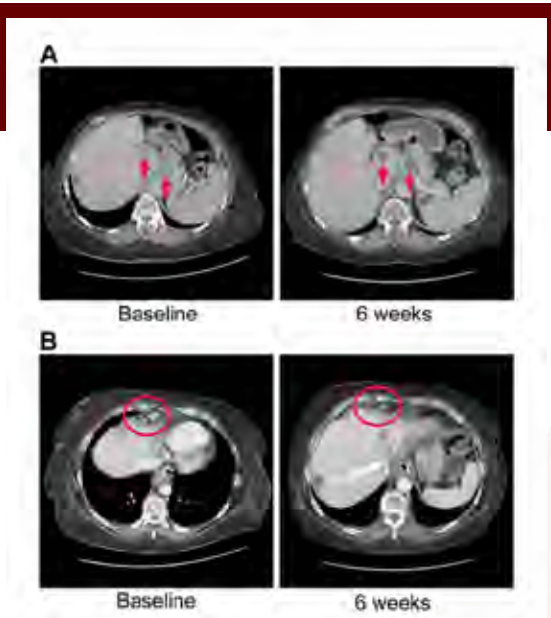


Figure 7. Anti-tumor activity in Patient 4 harboring an *FGFR2-MGEA5* fusion, to FGFR inhibitors. A) CT images of patient 4, whose tumor possessed an *FGFR2-MGEA5* fusion, at baseline and 6 weeks demonstrate central necrosis of a caudate liver lobe mass (left arrow), 2.6 cm at baseline and 6 weeks, and shrinkage of a metastatic supraceliac axis lymph node (right arrow), 3.1 cm and 2.9 cm at baseline and 6 weeks respectively. **B)** CT images of patient 4 showing shrinkage of metastatic lymph nodes involving the right cardiophrenic angle (red circles), 1.3 cm and 0.5 cm at baseline and 6 weeks respectively.
doi:10.1371/journal.pgen.1004135.g007

OPEN **Clinical Implementation of Integrated Genomic Profiling in Patients with Advanced Cancers**

Mitesh J. Borad[1,2], Jan B. Egan[1], Rachel M. Condjella[1], Winnie S. Liang[1], Rafael Fonseca[1,2], Nicola P. Ritacco[1], Ann E. McCullough[1], Michael T. Barrett[1,3], Katherine S. Hunt[1], Mia D. Champion[1,2], Maitray D. Patel[1], Scott W. Young[1], Alvin C. Silva[1], Thai H. Ho[1,2], Thorvardur R. Halfdanarson[1,2], Robert R. McWilliams[1,2], Konstantinos N. Lazaridis[1], Ramesh K. Ramanathan[1,2], Angela Baker[1], Jessica Aldrich[1], Ahmet Kurdoglu[1], Tyler Izatt[1], Alexis Christoforides[1], Irena Cherni[1], Sara Nasser[1], Rebecca Reiman[1], Lori Cuyugan[1], Jacqueryn McDonald[1], Jonathan Adkins[1], Stephen O. Mastrian[1], Riccardo Valdez[1], Dawn E. Jaroszewski[1], Daniel D. Von Hoff[1], David W. Craig[1], A. Keith Stewart[1,2,3], John D. Carpten[1] & Alan H. Bryce[1,2,3]

EXAMPLE CHOLANGIOCARCINOMA

35 somatic coding mutations

- Two COSMIC genes
- None in known commercial cancer panels
- One flagged inferred therapeutic context

One focal copy number event on chr3

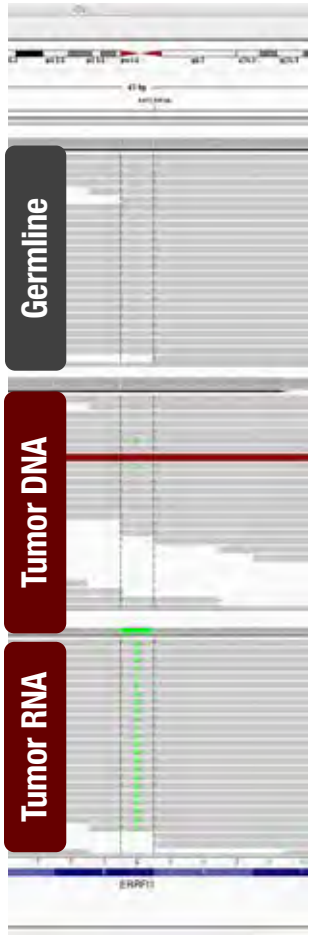
- None flagged inferred therapeutic context

RNA-seq data Differential Expression

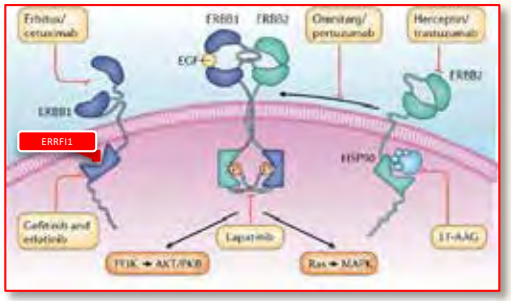
- None flagged inferred therapeutic context

Structural Variants and Fusions

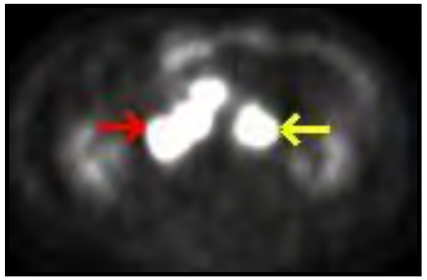
- None flagged inferred therapeutic context

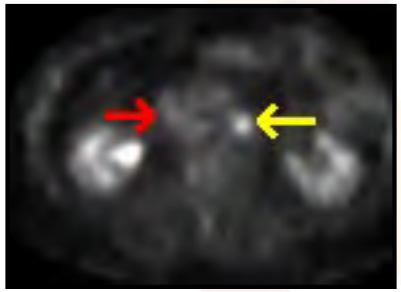


ERFFI1 – Context for EGFR Activation



- **EGFR inhibitor was recommended By Tumor Board**





- **Previously bulky much smaller and much less metabolically active at 10/25/12 on PET/CT.**

- **>60% reduction of retroperitoneal left periaortic node from 08/06/12 to 10/25/12 on PET scan.**

Sample Prep Types

Library Preparation

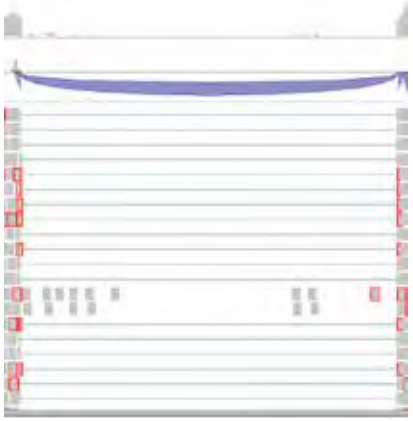
TABLE 3.1 RNA-seq Library Protocols

Library Design	Pros	Cons
Ribo-depletion	Removes ribosomal RNA from sample and provides access to mRNA and non-coding RNA	Pre-mRNA retained providing reads for junctions
Poly-A selection	Removes all non-polyadenylated RNA. Enriches for mRNA	Limited quantification of non-coding RNA
Treatment with double-stranded nuclease	Reduces expression level of highly abundant transcripts	Not a well-utilized protocol; biases/impact on global expression not characterized
MicroRNA capture	Provides access to microRNA	MicroRNA recovered depends on capture technique
mRNA capture	Provides selective representation of specific transcripts and potentially unknown fusion partners	Potentially biases quantification; reduces representation of untargeted genes, which may be of interest
Multiplexed using genetic barcoding	Many samples processed together	Some read pairs might be erroneously swapped if run cluster density is too high

FFPE Considerations

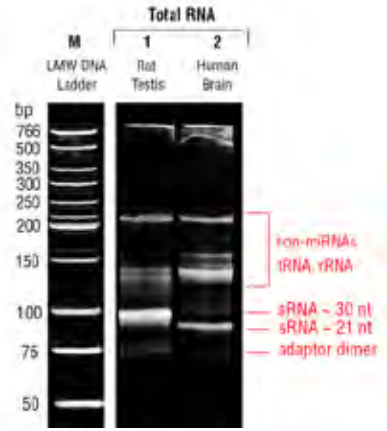
Fresh/Frozen/Cryo

Poly-A



Random Priming

Size Selection – smallRNA



A fraction of types of assays

EXAMPLE: CERVICAL CARCINOMA

Integration of HPV (Detailed)

HPV18 DNA is evident as is expression of RNA

HPV 18 Genome (NC_001357)

PCAWG

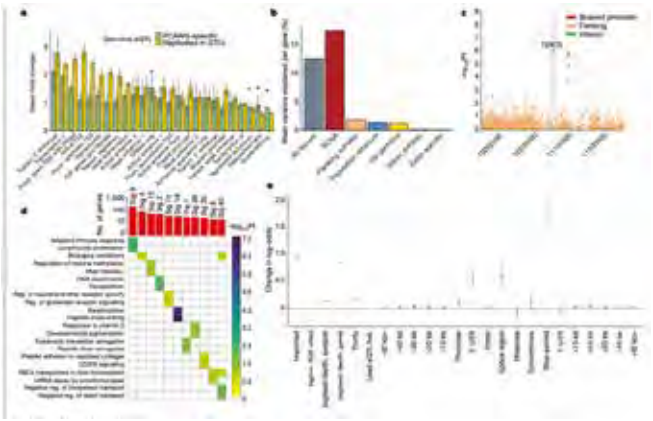
Article

Genomic basis for RNA alterations in cancer

https://doi.org/10.1038/s41586-020-1970-0
Received: 29 March 2018
Accepted: 11 December 2019
Published online: 5 February 2020
Open access

PCAWG Transcriptome Core Group[1,35], Claudia Calabrese[2,35], Natalie R. Davidson[3,4,5,6,7,35], Deniz Demircioğlu[8,9,10], Nuno A. Fonseca[2,35], Yao He[10,35], André Kahles[3,4,6,7,35], Kjong-Van Lehmann[3,4,6,7,35], Fenglin Liu[10,35], Yuichi Shiraishi[11,35], Cameron M. Soulette[12,35], Lara Urban[2,35], Liliana Greger[2], Siliang Li[13,14], Dongbing Liu[13,14], Marc D. Perry[15,16], Qian Xiang[13], Fan Zhang[10], Junjun Zhang[10], Peter Bailey[17], Serap Erkek[18], Katherine A. Hoadley[19], Yong Hou[13,14], Matthew R. Huska[20], Helena Kilpinen[21], Jan O. Korbel[18], Maximillian G. Marin[11], Julia Markowski[210], Tannistha Nandi[9], Qiang Pan-Hammarström[13,22], Chandra Sekhar Pedamallu[23,28,29], Reiner Siebert[24], Stefan G. Stark[3,4,6,7], Hong Su[13,14], Patrick Tan[9,25], Sebastian M. Waszak[18], Christina Yung[10], Shida Zhu[13,14], Philip Awadalla[15,26], Chad J. Creighton[27], Matthew Meyerson[21,28,29], B. F. Francis Ouellette[30], Kui Wu[13,14], Huanming Yang[13], PCAWG Transcriptome Working Group[5], Alvis Brazma[2,36]*, Angela N. Brooks[12,23,28,36]*, Jonathan Göke[9,31,36], Gunnar Rätsch[3,4,6,6,7,36]*, Roland F. Schwarz[2,20,32,33,36], Oliver Stegle[2,10,33,36], Zemin Zhang[10,36] & PCAWG Consortium[34]

Transcript alterations often result from somatic changes in cancer genomes[1]. Various forms of RNA alterations have been described in cancer, including overexpression[2], altered splicing[3] and gene fusions[4]; however, it is difficult to attribute these to underlying genomic changes owing to heterogeneity among patients and tumour types, and the relatively small cohorts of patients for whom samples have been analysed by both transcriptome and whole-genome sequencing. Here we present, to our knowledge, the most comprehensive catalogue of cancer-associated gene alterations to date, obtained by characterizing tumour transcriptomes from 1,188 donors of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)[5]. Using matched whole-genome sequencing data, we associated several categories of RNA alterations with germline and somatic DNA alterations, and identified probable genetic mechanisms. Somatic copy-number alterations were the major drivers of variations in total gene and allele-specific expression. We identified 649 associations of somatic single-nucleotide variants with gene expression in *cis*, of which 68.4% involved associations with flanking non-coding regions of the gene. We found 1,900 splicing alterations associated with somatic mutations, including the formation of exons within introns in proximity to Alu elements. In addition, 82% of gene fusions were associated with structural variants, including 75 of a new class, termed 'bridged' fusions, in which a third genomic location bridges two genes. We observed transcriptomic alteration signatures that differ between cancer types and have associations with variations in DNA mutational signatures. This compendium of RNA alterations in the genomic context provides a rich resource for identifying genes and mechanisms that are functionally implicated in cancer.



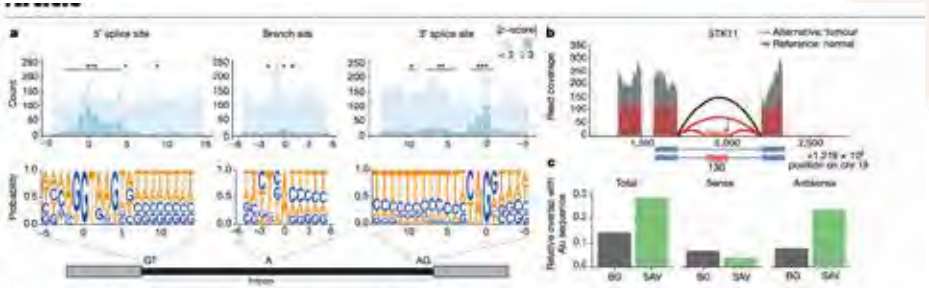


Fig. 2 | Position-specific effect of somatic mutations on alternative splicing. **a**, Top, proportion of mutations near exon–intron junctions and at branch sites that are associated with exon-skipping events. Mutations with associated splicing changes are those in which the percentage spliced in-derived (z-score) is ≥ 3 (dark blue). Asterisks denote intron positions significantly enriched for splicing changes relative to background based on a permutation test. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. Bottom, sequence motifs of regions. **b**, Example of an exonization event in the tumour-suppressor gene *STK11*. The RNA-seq read coverage for a part of the gene is shown in red for a donor carrying the alternative allele, and in grey for a random donor with reference allele. The cassette exon event is shown as a schematic below. **c**, Enrichment of SINE elements in SAVs compared to sequence background (BG). Shown for SINE elements overlapping in sense (middle) and antisense (right) directions.

PCAWG

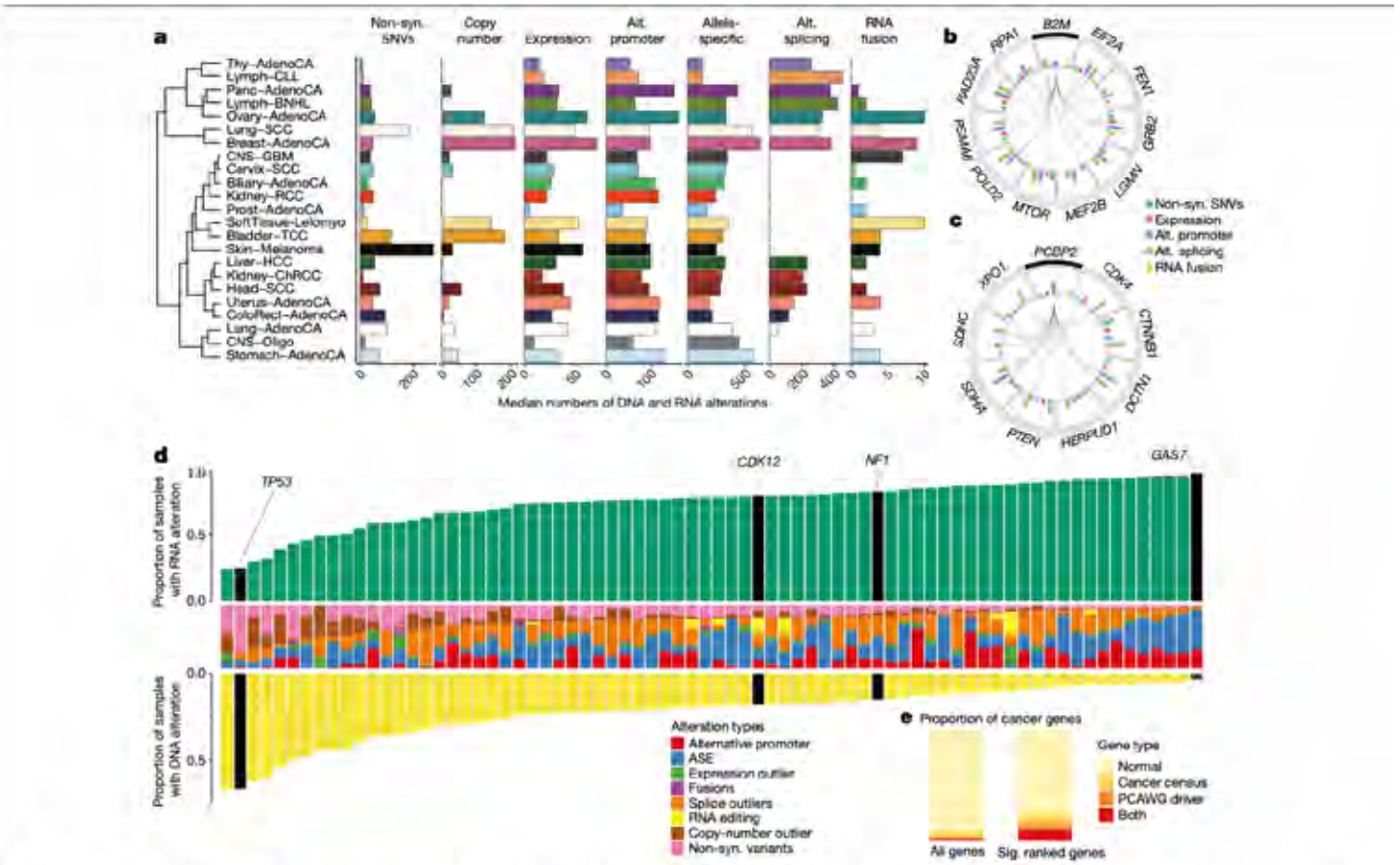
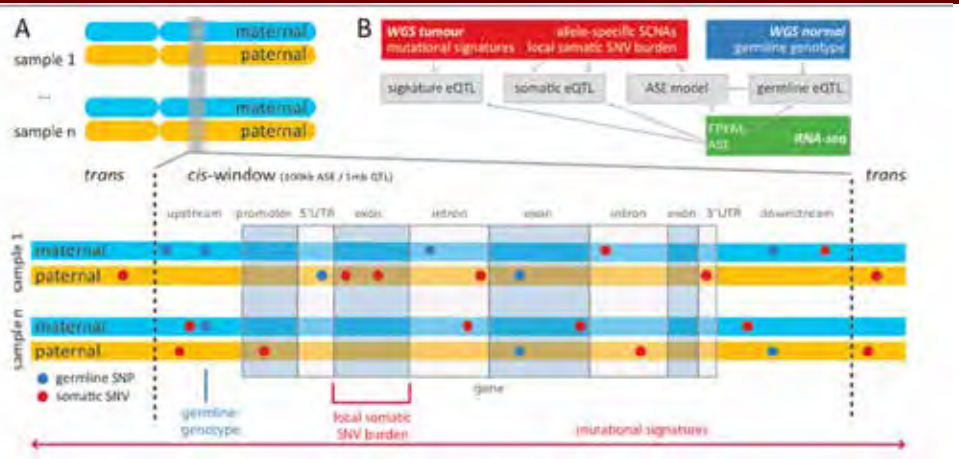
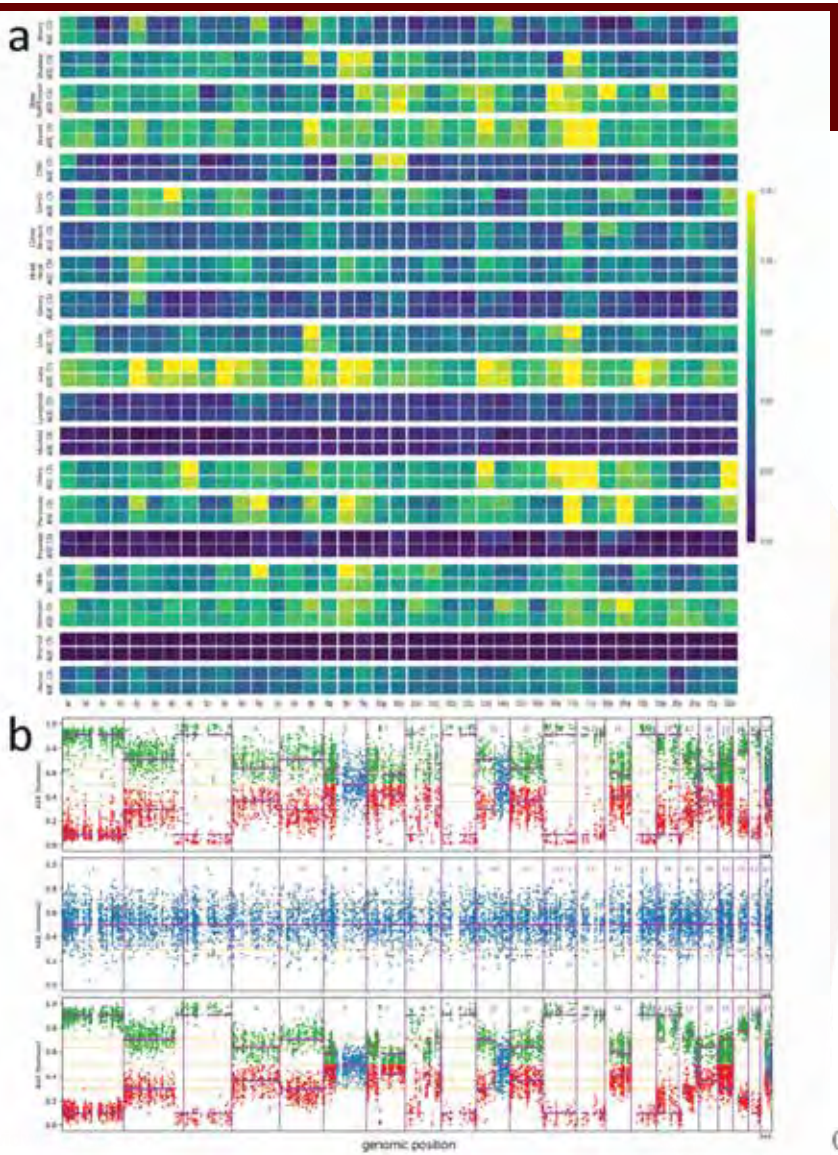
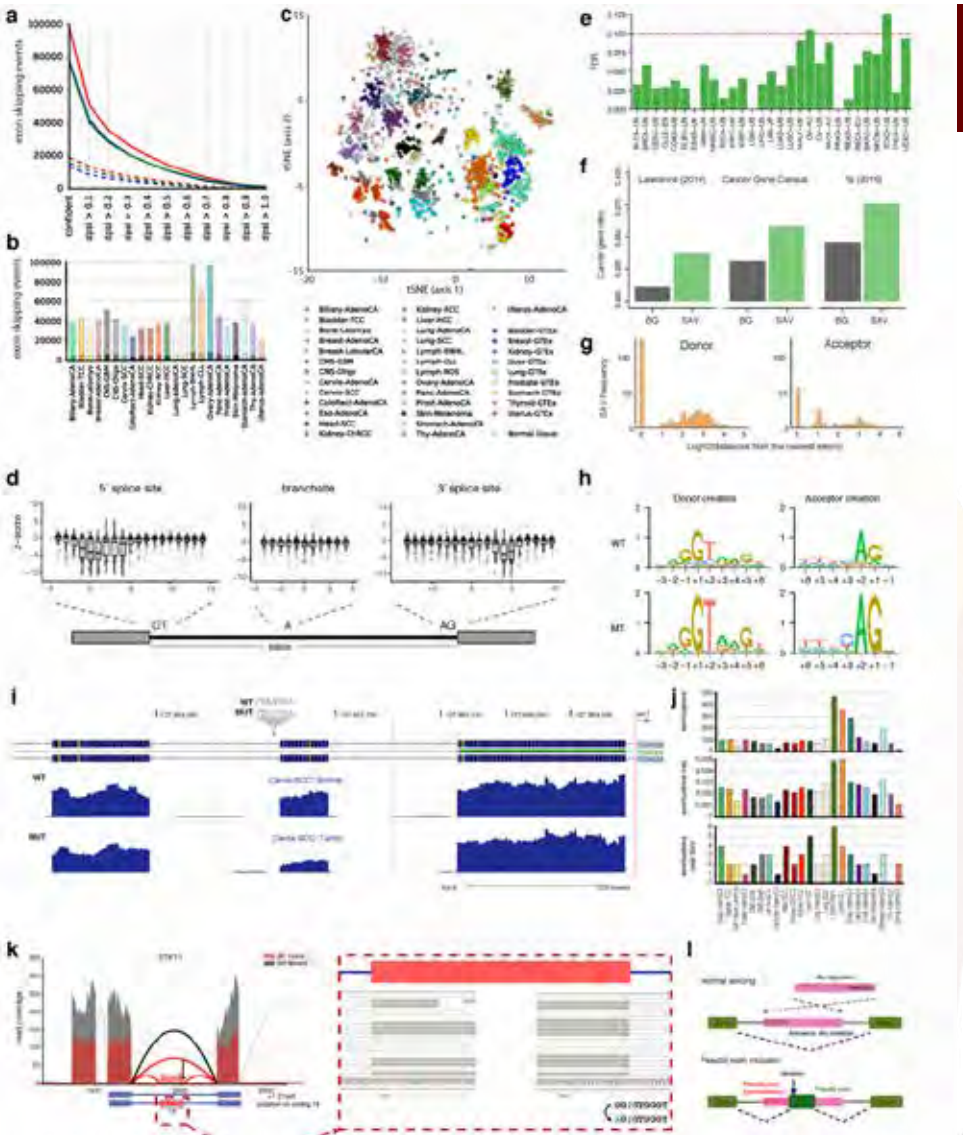


Fig. 4 | Global view of DNA and RNA alterations that affect tumours. a, The median numbers of different alterations across histotypes. Histotypes are ordered by hierarchical clustering based on the pattern of different types of alteration. Only histotypes with more than 10 donors are shown. Alt., alternative; non-syn, non-synonymous. Cancer-type abbreviations are listed in Supplementary Table 23. **b**, **c**, Circular representations of the selected genes significantly co-occurred with *B2M* (**b**) and *PCBP2* (**c**). Connecting lines indicate the specific types of co-occurrence of alteration pairs. The inner histograms indicate the frequencies of incidences of different alteration types shown in different colours. **d**, All 74 Catalogue of Somatic Mutations in Cancer (COSMIC) cancer census genes or PCAWG driver genes that are both frequently and heterogeneously altered across both RNA- and DNA-level alterations. Yellow bars indicate the proportion of samples that had DNA-level alterations, and green bars indicate the proportion of samples with RNA-level alterations. Middle column is the proportion of each alteration type observed for that gene. **e**, The enrichment of cancer genes within our list of significantly recurrent genes.


A
sample 1
maternal
paternal
sample n
maternal
paternal
B
WGS tumour
mutational signatures
allele-specific SCNAs
local somatic SNV burden
WGS normal
germline genotype
signature eQTL
somatic eQTL
ASE model
germline eQTL
RNA-seq
trans
cis-window
trans
upstream
promoter
5'UTR
exon
intron
exon
intron
exon
3'UTR
downstream
sample 1
maternal
paternal
sample n
maternal
paternal
germline SNP
somatic SNV
gene
germline genotype
local somatic SNV burden
mutational signatures


a
b
genomic position

ALT SPLICING IN PCAWG



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Translating RNA sequencing into clinical diagnostics: opportunities and challenges

Sara A. Byron, Kendall R. Van Keuren-Jensen, David M. Engelthaler, Julie C. Carpten and Nathan W. Craig

BIOMARKERS

Table 1 | **Selected examples of current RNA-based clinical tests**

RNA biomolecule	Method	Examples	Use
Viral RNA	qRT-PCR	• Influenza virus[36] • Dengue virus[38] • HIV[39] • Ebola virus[71]	Viral detection and typing
mRNA	qRT-PCR	• AlloMap (CareDx; heart transplant)[14,15] • Cancer Type ID (BioTheranostics)[111]	Diagnosis
	Microarray	Afirma Thyroid Nodule Assessment (Veracyte)[112]	Diagnosis
	qRT-PCR	• OncotypeDx (Genome Health; breast, prostate and colon cancer)[109,110] • Breast Cancer Index (BioTheranostics)[113] • Prolaris (Myriad; prostate cancer)[114]	Prognosis
	Digital barcoded mRNA analysis	Prosigna Breast Cancer Prognostic Gene Signature (Nanostring)[115]	Prognosis
	Microarray	• MammaPrint (Agendia; breast cancer)[116] • ColoPrint (Agendia; colon cancer)[116] • Decipher (Genome Dx; prostate cancer)[117]	Prognosis
miRNA	Microarray	Cancer Origin (Rosetta Genomics)[118]	Diagnosis
Fusion transcript	qRT-PCR	AML (*RUNX1–RUNX1T1*)[119]	Diagnosis
	qRT-PCR	BCR–ABL1 (BCR 21)	Monitoring molecular response during therapy
	qRT-PCR (exosomal RNA)	ExoDx Lung (ALK) (Exosome Dx)[120]	Fusion detection
	RNA-seq	FoundationOne Heme[121]	Fusion detection

Analytical validity
Accuracy and reliability of a test to measure a specific biomarker

Clinical validity
The accuracy of how well a test detects or predicts clinical diagnosis or outcome

Clinical utility
The likelihood the test is to inform clinical decisions and improve outcome

Analytical sensitivity
How often is the test positive when the biomarker is present?

Analytical specificity
How often is the test negative when the biomarker is not present?

Robustness
Repeatability and reproducibility of the assay within and across laboratories.

Limits of detection
Lowest level of reliable detection of transcripts.

Stability
Collection, handling, transport of sample and impact on robustness.

Gold standards
Reference sets for assessing sensitivity and specificity.

Clinical sensitivity
How often is the test positive in patients with the disease or clinical outcome?

Clinical specificity
How often is the test negative in patients without the disease or clinical outcome?

Prevalence
The proportion of individuals that will have a disease or outcome.

Positive predictive value
Given prevalence, the probability that subjects with a positive test result for a disorder or outcome will have the disease or outcome.

Negative predictive value
For negative tests, the probability that subjects truly will not have the disease or outcome.

Penetrance
The proportion of subjects with the biomarker that have the predicted outcome or diagnosis.

Appropriate intervention
Assessment of test impact on patient care, publishing of clinical trials.

Quality assurance
Quality control measures for tests, reagents and/or facilities.

Monitoring
Long-term monitoring of patients and establishment of guidelines for performance.

Economics
Financial costs and economic benefits associated with test.

Education
Educational materials and informed consent requirements.

ELSI
Assessment of ethical, legal and societal implications that arise in the context of the test.

Analysis Common Features

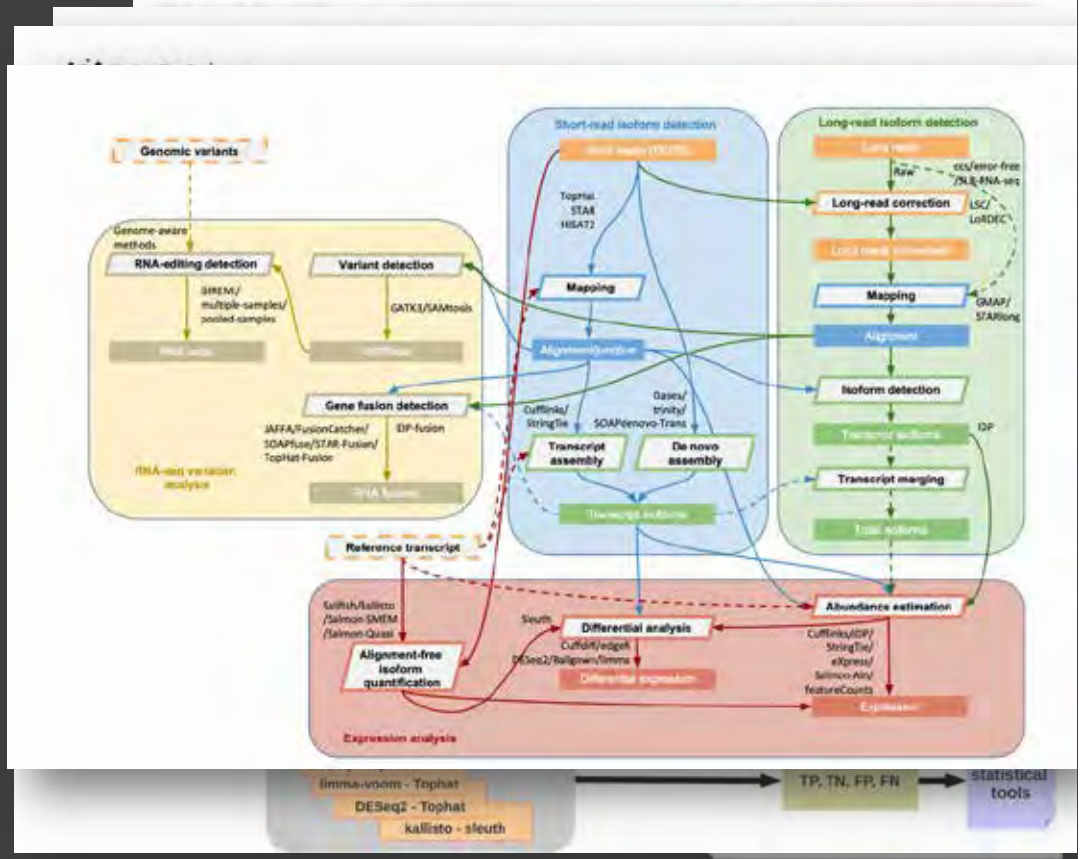
Quality Control

Alignment/Assembly

Detection & Abundance

Normalization

Significance Of Model/Hypothesis



RNA-Seq 2020

- RNA-Seq
- CaptureSeq
- RASL-Seq
- ClickSeq
- 3Seq
- cP-RNA
- 3P-Seq
- 2P-Seq
- 3'-Seq
- TIF-Seq
- PEAT
- SMORE-Seq
- TL-Seq
- TATL-Seq
- RARseq
- TAIL-Seq
- PAL-Seq
- ChIRP
- CHART
- RAP
- GRO-seq
- Bru-Seq
- 3'NT Method
- NET-Seq
- mNET-Seq
- PARE-Seq
- GMUCT
- Ribo-Seq or ARTSeq
- RIP-Seq
- CLIP-Seq or HITS-CLIP
- Pol II CLIP
- PIP-Seq
- hiCLIP
- RBNS
- TRIBE
- HiTS-RAP
- TRAP-Seq
- DLAF
- miTRAP
- CLASH
- 71 RNA Modifications
- icSHAPE
- CIRS-Seq
- SHAPE-MaP
- Structure-Seq/DMS-Seq
- SPARE
- PARS-Seq
- Cap-Seq
- CIP-TAP
- scRNA-Seq
- SUPeR-Seq
- FRISCR
- CEL-Seq
- STRT-Seq
- TCR Chain Pairing
- TCR-LA-MC PCR
- CirSeq
- TIVA
- PAIR
- CLaP
- CytoSeq
- Drop-Seq
- Hi-SCL
- InDrop
- snRNA-Seq
- Nuc-Seq
- Div-Seq
- SCRB-Seq
- G&T-Seq
- scM&T-Seq
- scTrio-seq

Analysis Involves:

- Domain Specific Knowledge (Biology/Data)
- Often home-brew software, layered over standard pipelines
- R/Bioconductor Key Tools

https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/rna-sequencing-methods-review-web.pdf

UTILIZING IDEP FRAMEWORK FOR LEARNING

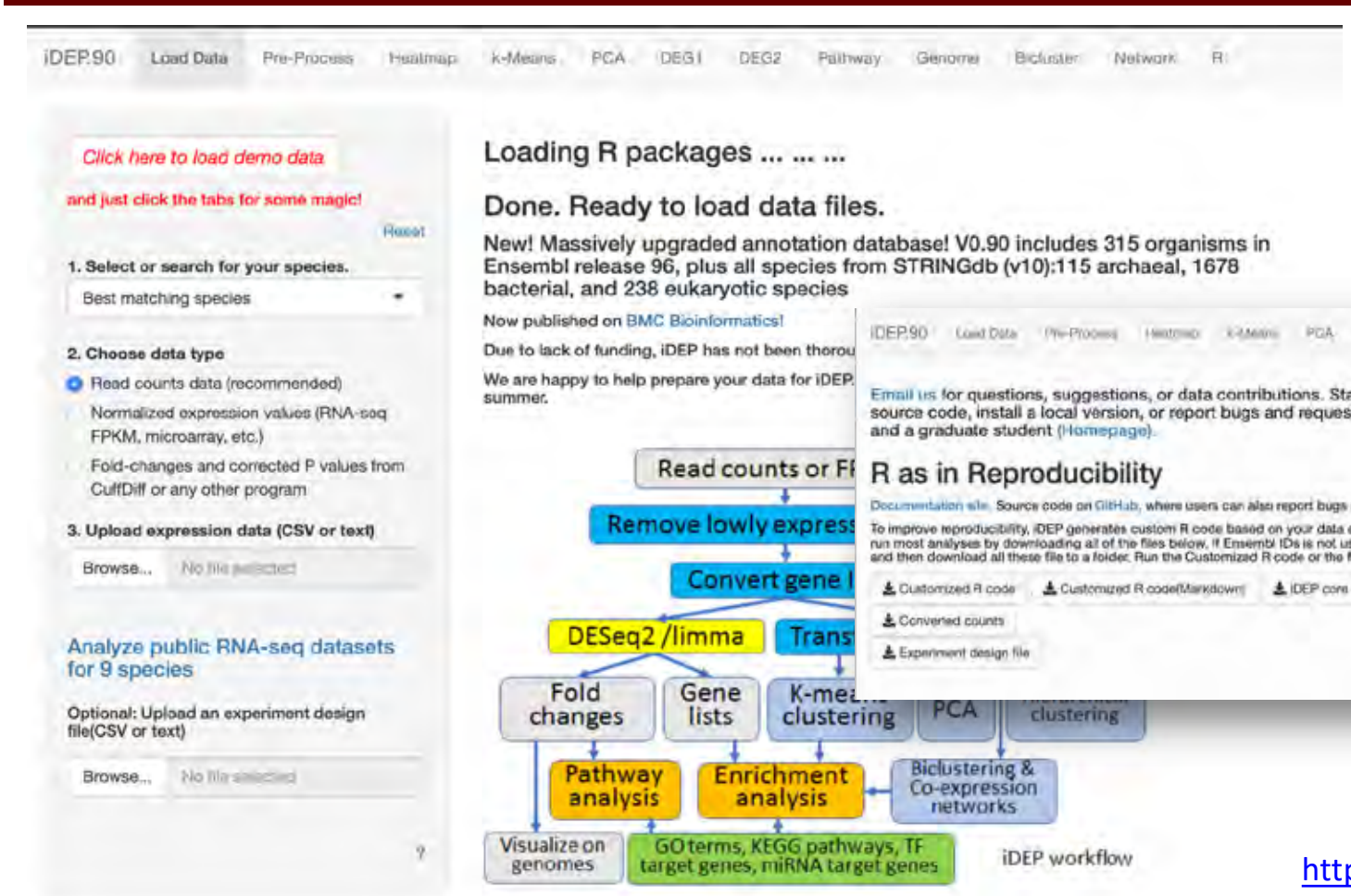
iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data

Steven Xijin Ge ✉, Eun Wo Son & Runan Yao

BMC Bioinformatics **19**, Article number: 534 (2018) | Cite this article

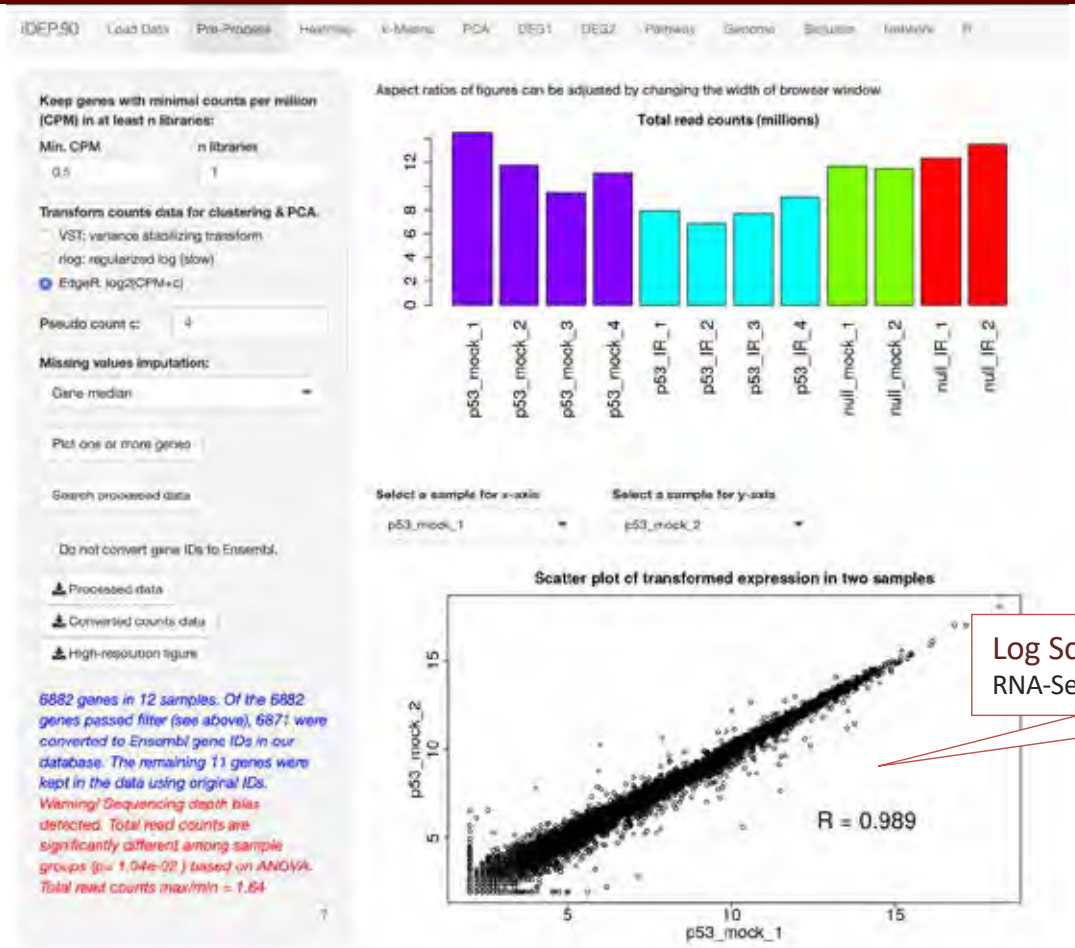
12k Accesses | **26** Citations | **23** Altmetric | Metrics

LEARNING BY EXAMPLE



http://bioinformatics.sdstate.edu/idep/

PRE-PROCESSING

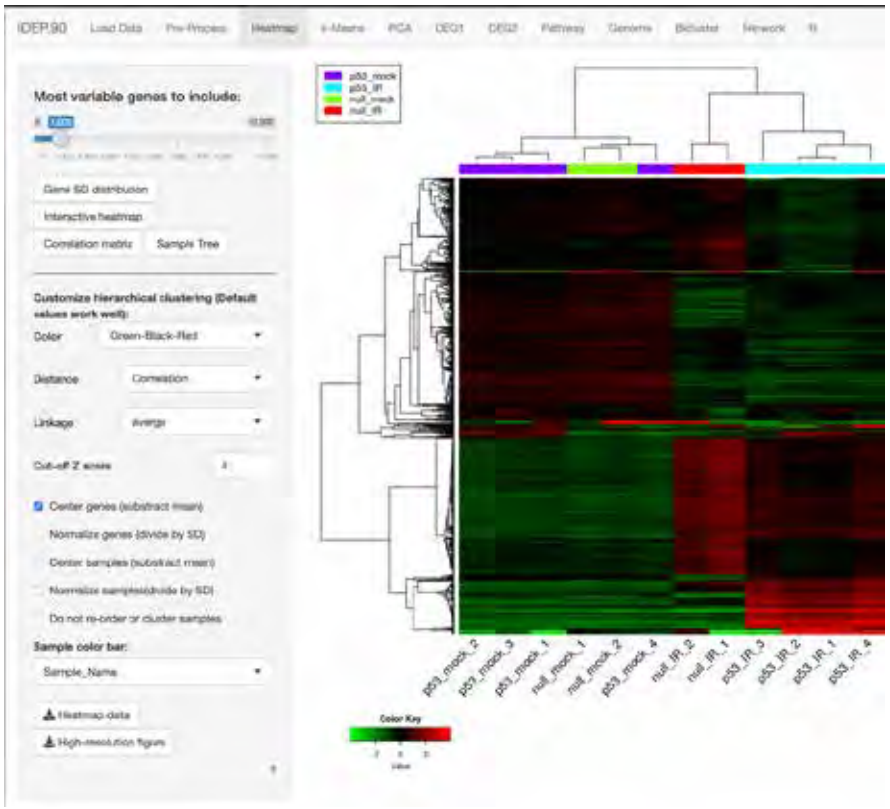


Log Scales

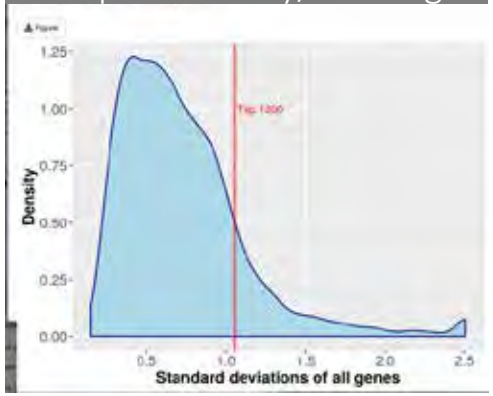
RNA-Seq spans many orders of magnitude

UNSUPERVISED: HEATMAPS, PCA, HIERARCHAL CLUSTERING

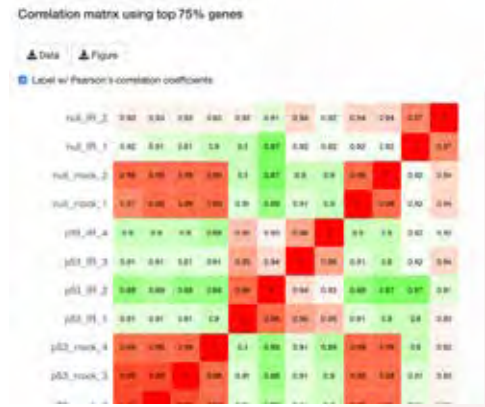
Heatmaps



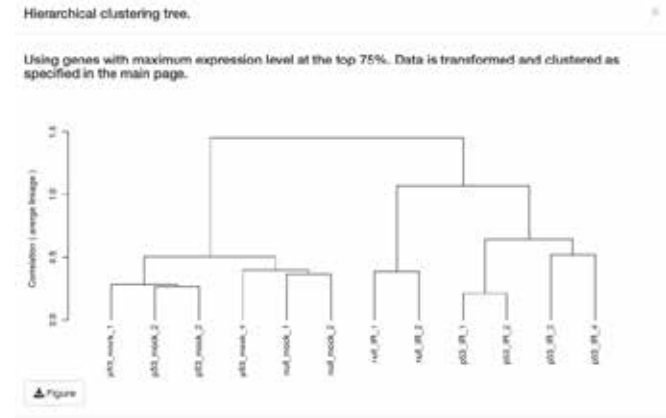
Reproducibility/Filtering



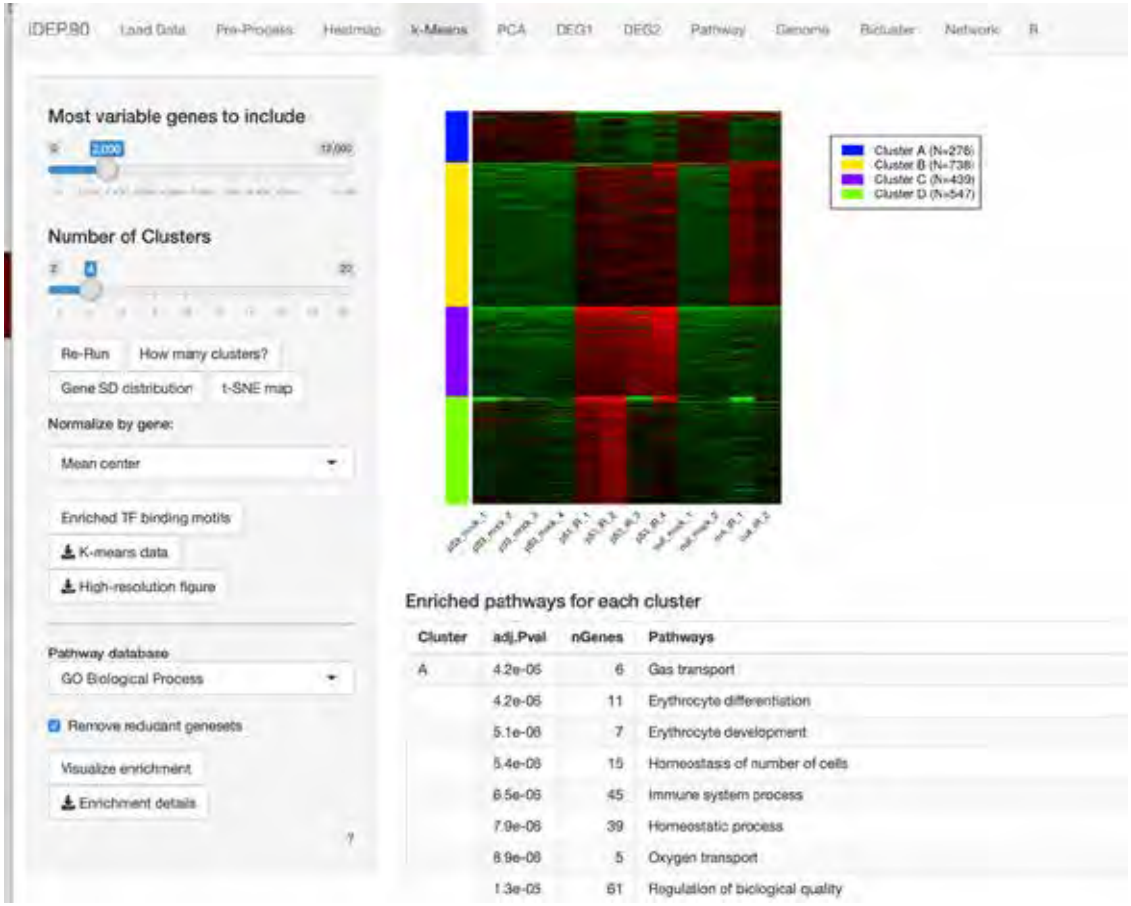
Correlation



Hierarchal Clustering

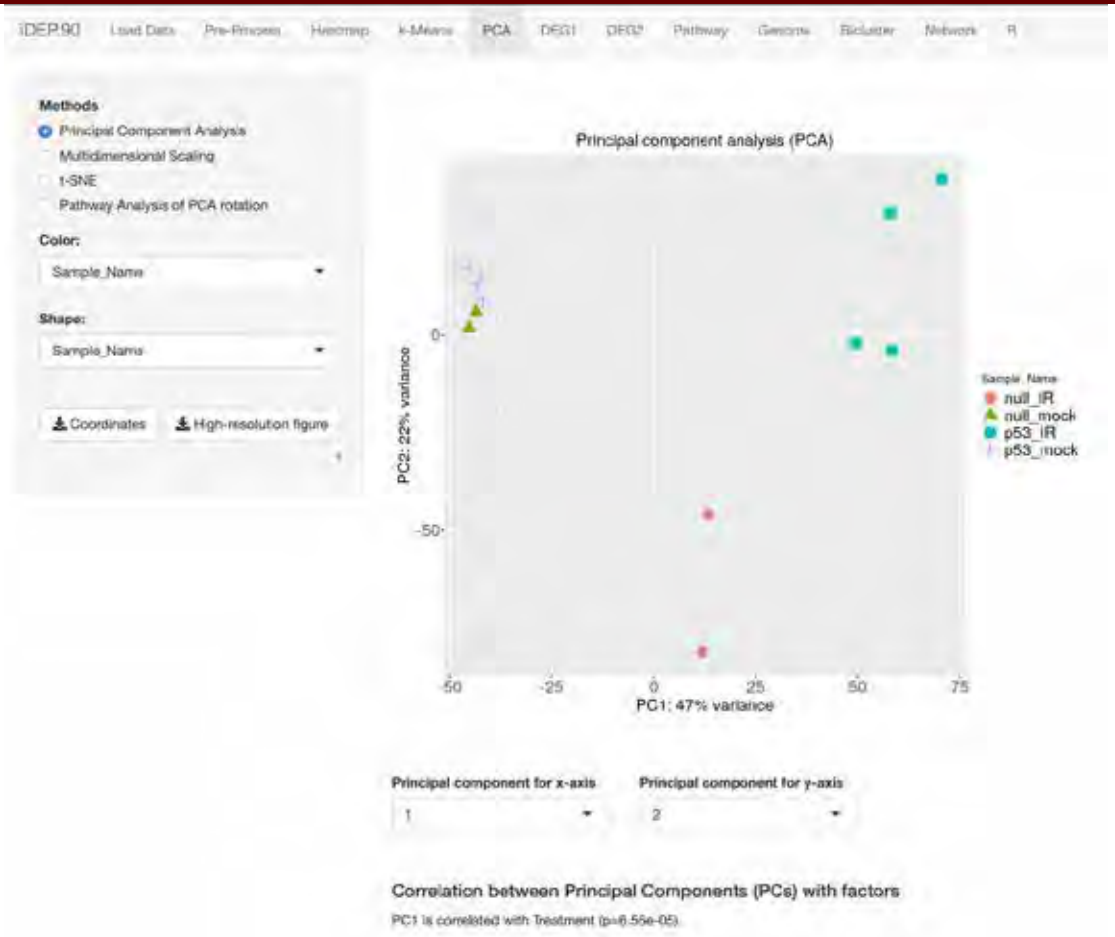


CLUSTERING

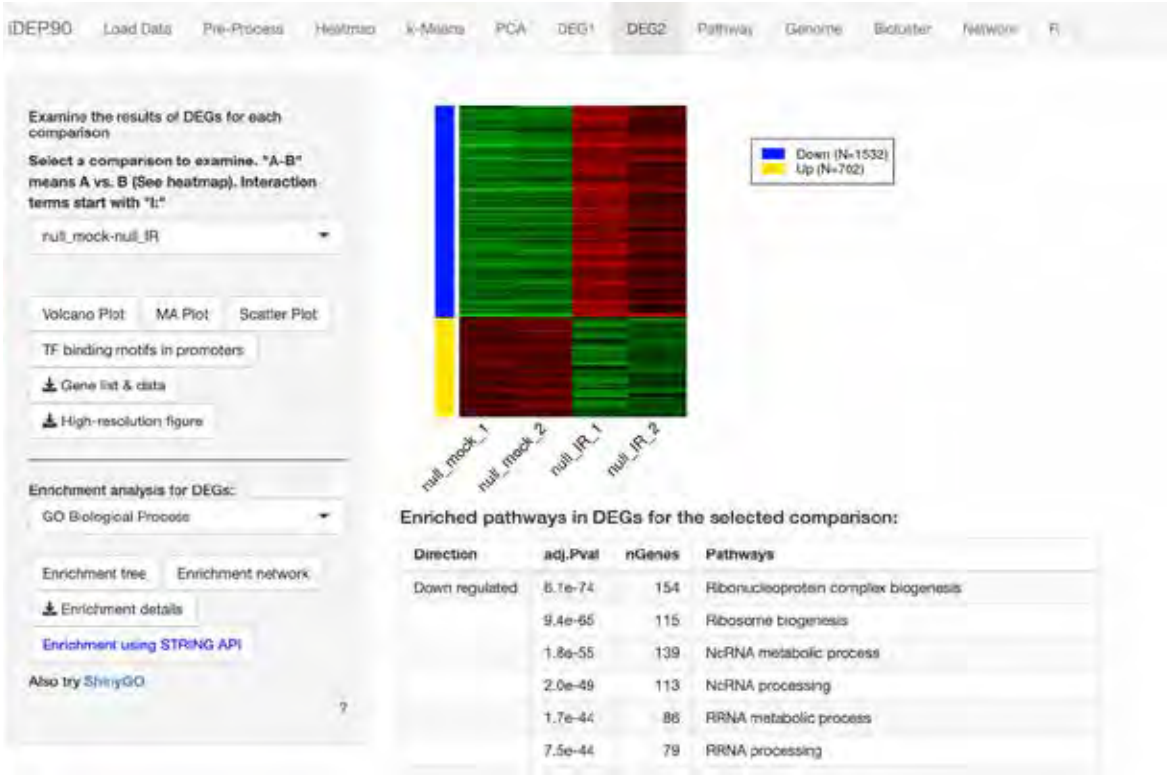


Cluster	adj.Pval	nGenes	Pathways
A	4.2e-06	6	Gas transport
	4.2e-06	11	Erythrocyte differentiation
	5.1e-06	7	Erythrocyte development
	5.4e-06	15	Homeostasis of number of cells
	6.5e-06	45	Immune system process
	7.9e-06	39	Homeostatic process
	8.9e-06	5	Oxygen transport
	1.3e-05	61	Regulation of biological quality

DIMENSION REDUCTION: PCA, T-SNE, MDS

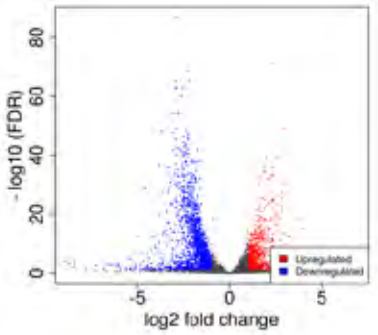


DIFFERENTIAL EXPRESSION

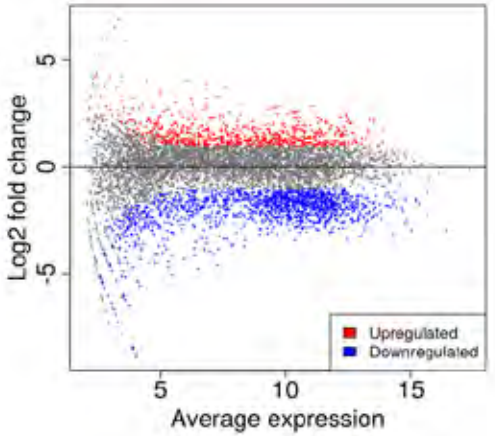


Direction	adj.Pval	nGenes	Pathways
Down regulated	8.1e-74	154	Ribonucleoprotein complex biogenesis
	9.4e-65	115	Ribosome biogenesis
	1.8e-55	139	NcRNA metabolic process
	2.0e-49	113	NcRNA processing
	1.7e-44	86	RRNA metabolic process
	7.5e-44	79	RRNA processing

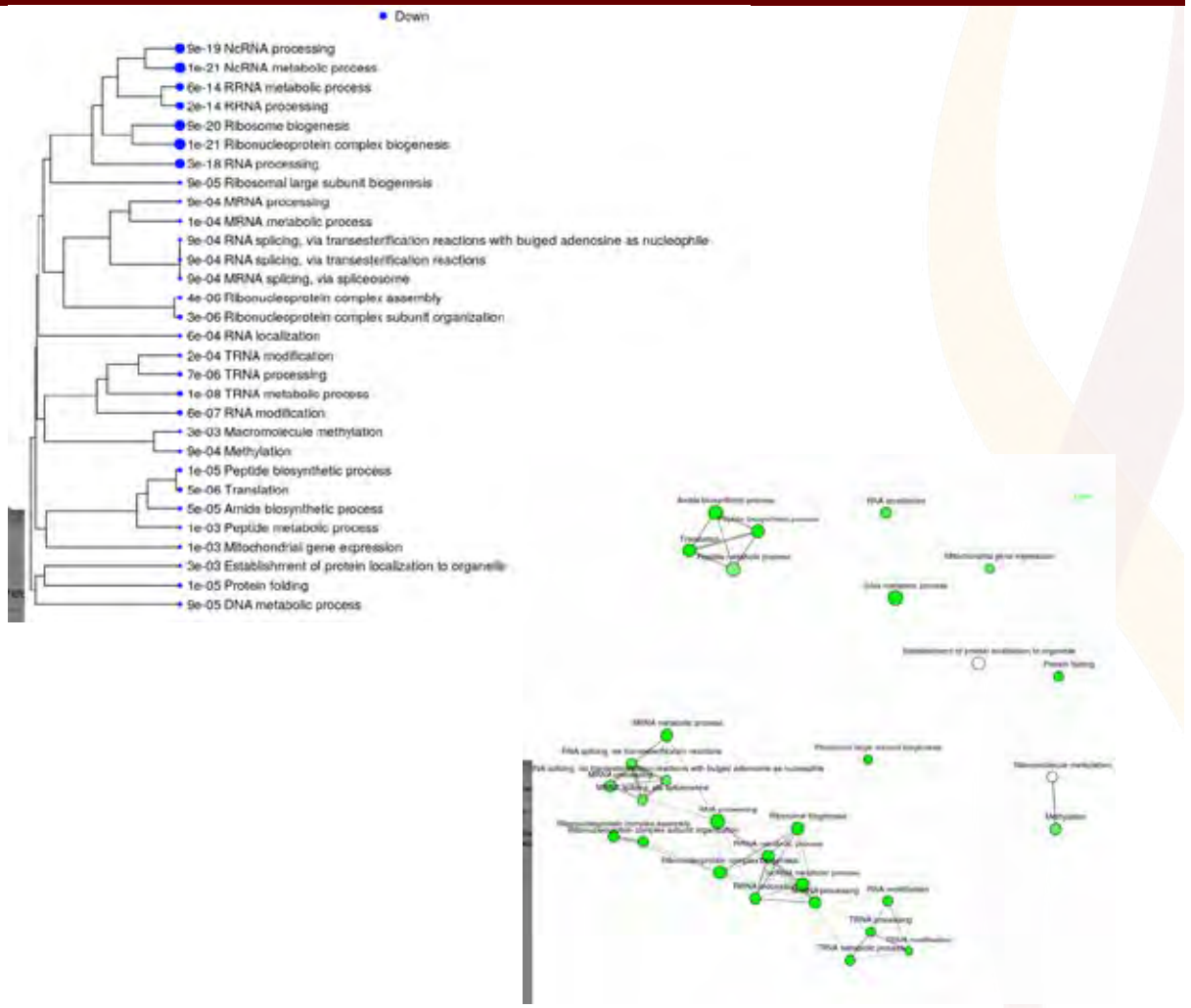
Volcano Plots



MA-Plots



PATHWAY ANALYSIS



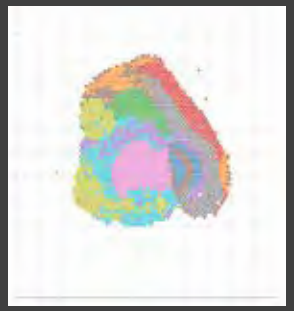
Databases, Analysis, and Software Tools

Resource	Description	URL	Refs
Annotation			
RefSeq	Curated reference sequence database (transcriptome-centric, that is, defined by transcript sequence)	https://www.ncbi.nlm.nih.gov/refseq/	75
GENCODE	Curated reference gene annotation (genome-centric, that is, defined by alignment to reference genome)	http://www.gencodegenes.org/	272
MiTranscriptome	Automated reference transcriptome based on sequence assembly, includes long non-coding RNAs	http://mitranscriptome.org/	76
Reference data			
MSigDB	Collection of experimental and curated gene sets (signatures)	http://software.broadinstitute.org/gsea/msigdb	179
Human Protein Atlas	Compendium of proteomic and transcriptomic data in diverse normal tissues	http://www.proteinatlas.org/	63
CCLE	Genomic and transcriptomic data on hundreds of cancer cell lines	https://portals.broadinstitute.org/ccle/home	60
GTEx	Transcriptomic data (RNA-seq) from normal human tissues from a large number of individuals	https://gtexportal.org/home/	62
Mitelman	Database of gene fusions and chromosomal aberrations	https://cgap.nci.nih.gov/Chromosomes/Mitelman	288
COSMIC	Catalogue of somatic mutations in cancer patients and cell lines, including gene fusions	http://cancer.sanger.ac.uk/cosmic/classic#fus	289
Tool			
QoRTs	Comprehensive collection of RNA-seq quality control functions	http://hartleys.github.io/QoRTs/index.html	290
STAR	Fast and accurate splice-aware sequence aligner	https://github.com/alexdobin/STAR	282
featureCounts	Fast read counting for gene-level or exon-level expression estimates	http://bioinf.wehi.edu.au/featureCounts/	291
Kallisto	Pseudo-alignment-based quantification at the transcript level	https://pachterlab.github.io/kallisto/	292
EdgeR	Differential expression using the negative binomial distribution (see also DESeq2)	http://bioconductor.org/packages/release/bioc/html/edgeR.html	170
Limma	Flexible linear modelling and empirical Bayes moderation to assess differential expression by use of precision weights for RNA-seq data (Voom)	http://bioconductor.org/packages/release/bioc/html/limma.html	167, 168
CIBERSORT	*In silico* transcriptome deconvolution into relative abundances of different immune cell types	https://cibersort.stanford.edu/	260
MiXCR	T cell and B cell CDR3 sequences assembler; enables repertoire profiling from RNA-seq data	https://milaboratory.com/software/mixcr/	261
GSEA	Gene set enrichment analysis	http://www.broad.mit.edu/GSEA	273
PARADIGM	Computational tool for the inference of patient-specific pathway activities	https://sbenz.github.io/Paradigm	186
FusionCatcher	A sensitive and specific tool for the detection of gene fusions	https://github.com/ndaniel/fusioncatcher	293
TopHat-Fusion	A very sensitive tool for the detection of gene fusions	http://ccb.jhu.edu/software/tophat/fusion_index.shtml	294
Analysis			
Oncomine	Web application for user-friendly analysis and exploration of cancer transcriptomes	https://www.oncomine.org/resource/login.html	180
Xena	UCSC Xena: versatile genomic data mining and analysis portal	https://xenabrowser.net/	267
Data warehouse			
ENCODE	Repository of diverse functional genomics data, including RNA-seq, from the ENCODE project	https://www.encodeproject.org/	59
GDC	Genomic Data Commons: provides access to raw and harmonized data for multiple genomic projects, including RNA-seq data processed using a standard pipeline	https://portal.gdc.cancer.gov/	295
FANTOM5	Repository of CAGE data from the FANTOM5 project	http://fantom.gsc.riken.jp/5/	271
ArrayExpress	Standard repositories of functional genomic and transcriptome profiling data	http://www.ebi.ac.uk/arrayexpress/	76
GEO		https://www.ncbi.nlm.nih.gov/geo/	77

CAGE, cap analysis of gene expression; CCLE, Cancer Cell Line Encyclopedia; ENCODE, Encyclopedia of DNA Elements; FANTOM5, Functional Annotation of the Mammalian Genome 5; GENCODE, the genome annotation project of ENCODE; GEO, Gene Expression Omnibus; GTEx, Genotype–Tissue Expression Project; Limma, Linear Models for Microarray Data; MSigDB, Molecular Signatures Database; PARADIGM, Pathway Recognition Algorithm using Data Integration on Genomic Models; QoRTs, Quality of RNA-seq Toolset; RNA-seq, RNA sequencing; STAR, Spliced Transcripts Alignment to a Reference; UCSC, University of California, Santa Cruz.

Cutting Edge: RNA-Seq Empowering Analysis of Heterogeneity



Single cell RNA-seq

Spatial Transcriptomes

SINGLE-CELL RNA-SEQ

Functional Studies w/ snRNA-seq

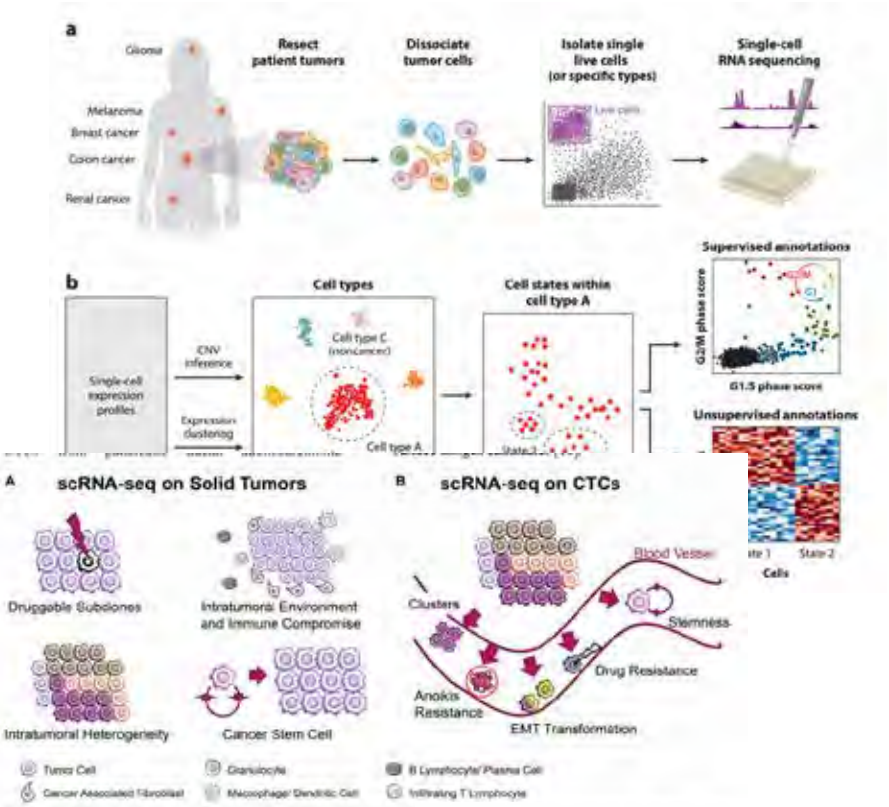
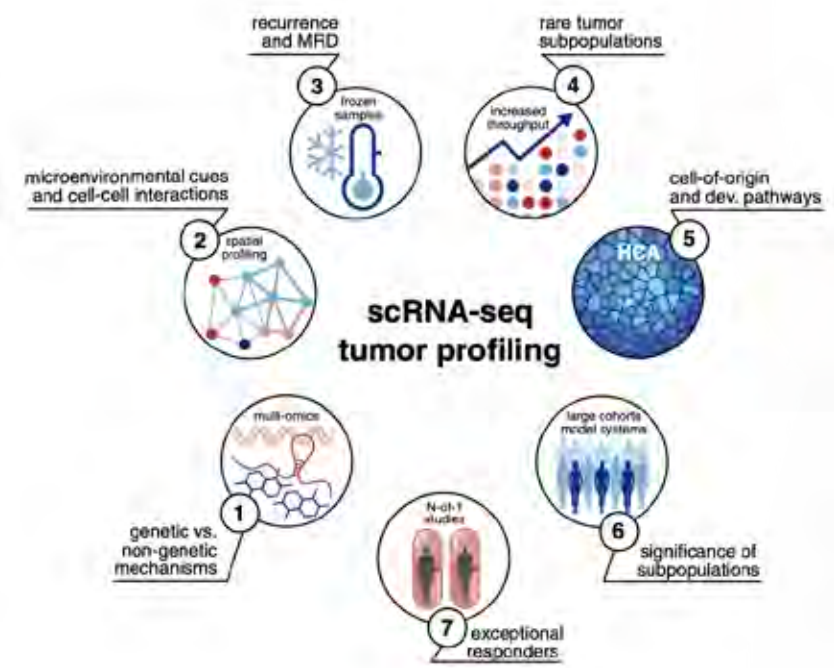


Figure 1: scRNA-seq technology facilitates cancer research when co *www.annualreviews.org • Single-Cell Expression Profiling in Cancer*
cells. (A) Findings of abnormal cell-to-cell interaction, drug resistance, and sequenced to explain the principle underlying

Study Types



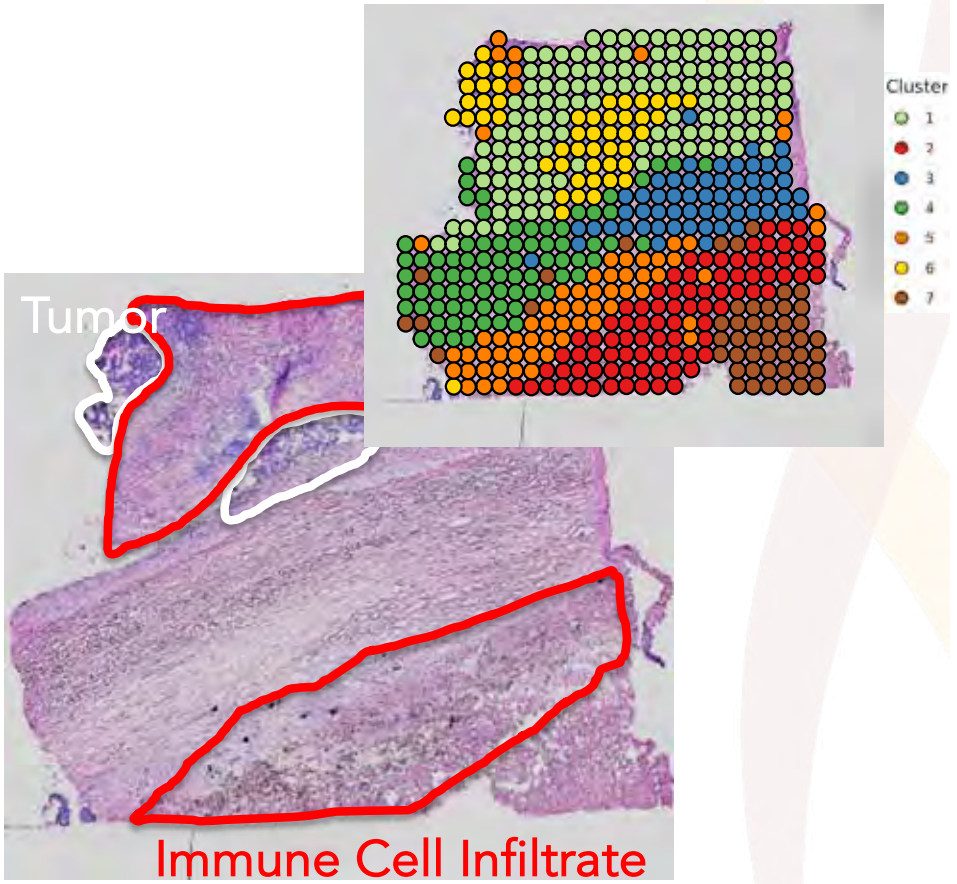
Assessing Tumor Heterogeneity

Molecular Annotation of Cluster Data

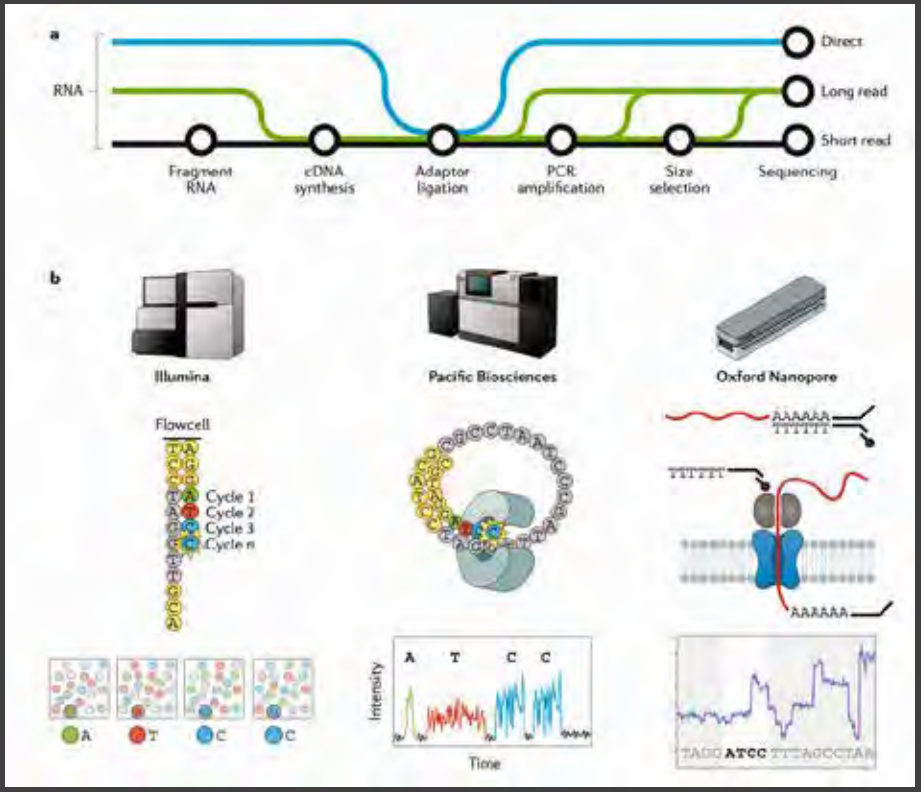
- Immune cluster profiling
- Spatial Gene Expression Maps
  - ESTIMATE Yoshihara et al., Nature Communications; 4 (2013): 2612
- Cibersort
  - Newman et al., Nature Methods. 2015; 12:453–457 (2015)
- xCell
  - Aran et al., Genome Biol. 2017;18(1):220.
- Inflammation and Immune Scoring
  - Ayers et al., J Clinical Investigation. 2017; 127(8):2930-2940.

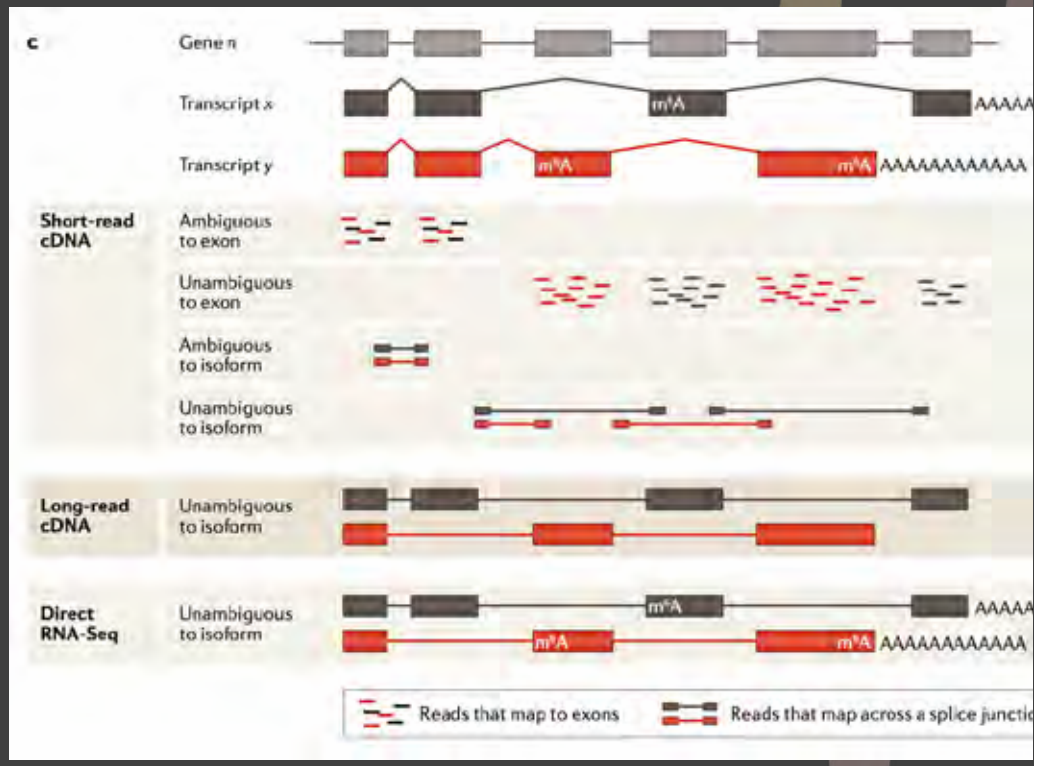
Tumor cluster profiling

- GSEA
  - Subramanian, Tamayo, et al. PNAS. 2005; 102, 15545-15550.



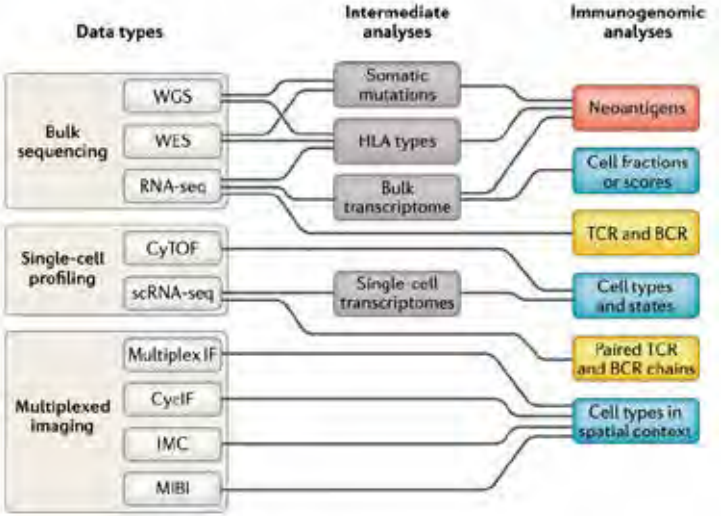
BLEEDING EDGE LONG-READS REALTIME

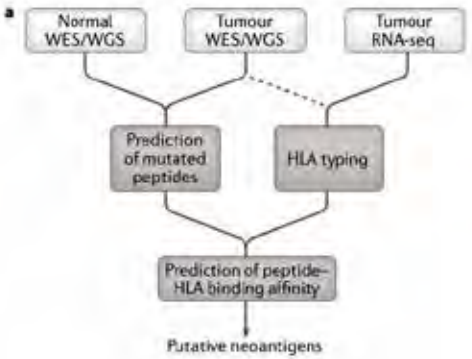


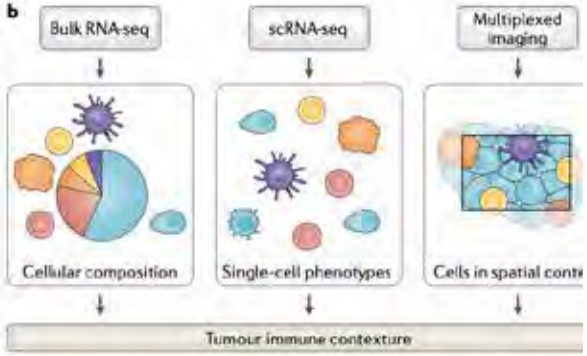


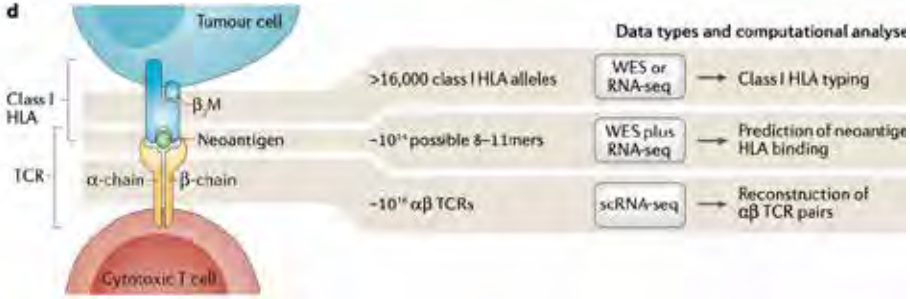
RNA sequencing: the teenage

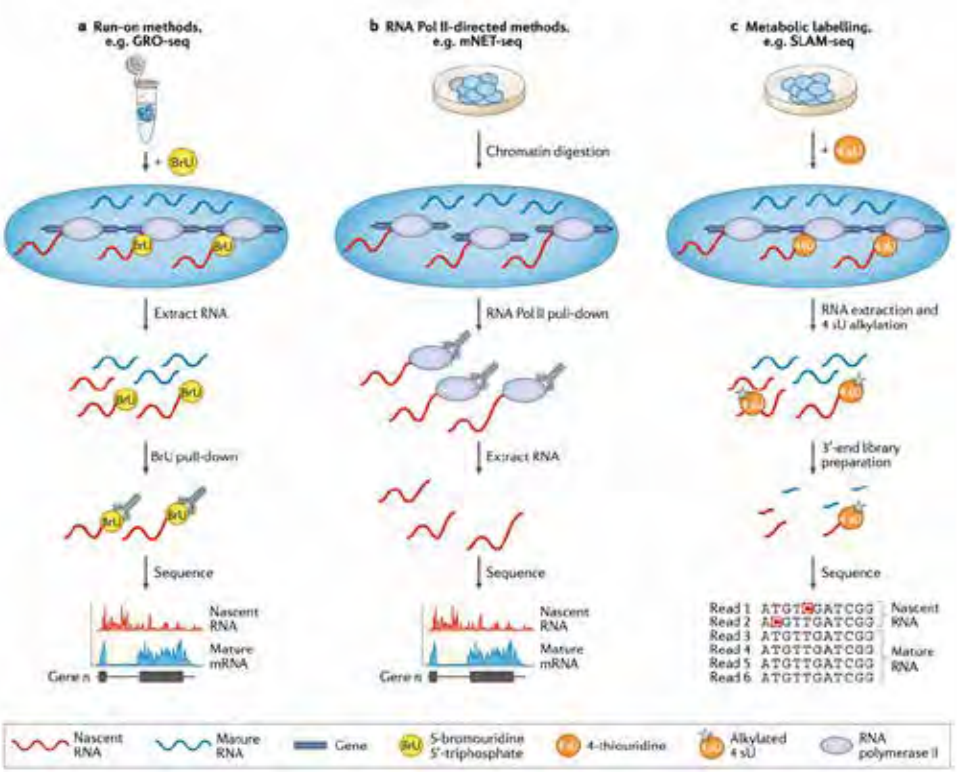
BLEEDING EDGE IMMUNE SINGLE CELL

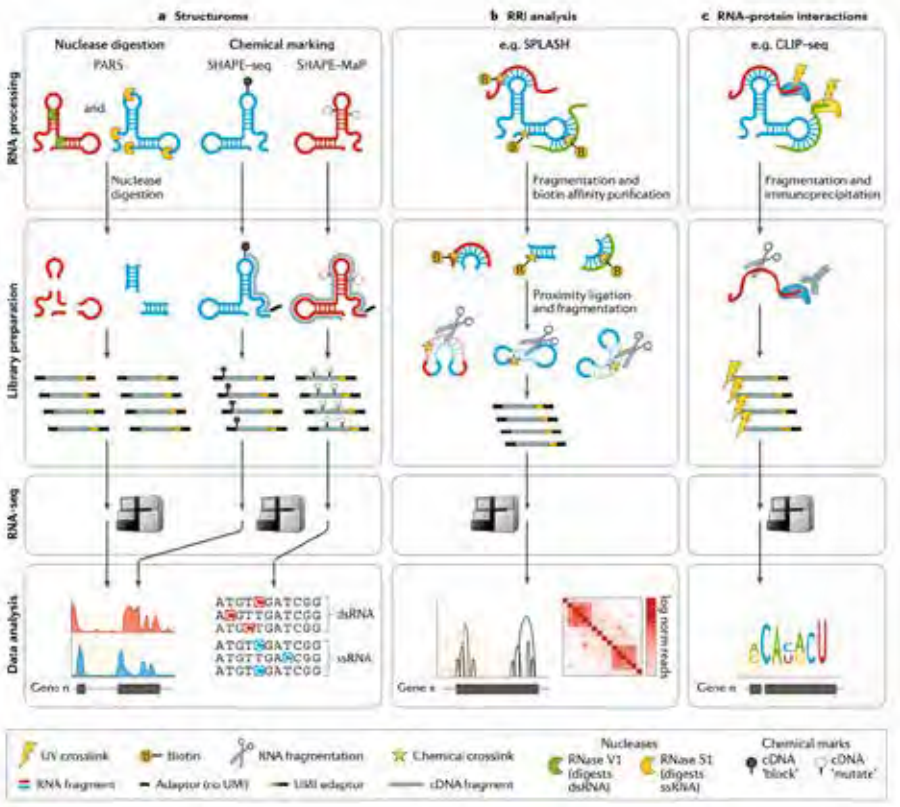









a Run-on methods, e.g. GRO-seq
b RNA Pol II-directed methods, e.g. mNET-seq
c Metabolic labelling, e.g. SLAM-seq
BrU
Chromatin digestion
4sU
Extract RNA
RNA Pol II pull-down
RNA extraction and 4 sU alkylation
BrU pull-down
Extract RNA
3′-end library preparation
Sequence
Sequence
Sequence
Nascent RNA
Mature mRNA
Gene n
Nascent RNA
Mature mRNA
Gene n
Read 1 ATGTCGATCGG
Read 2 ACGTTGATCGG
Read 3 ATGTTGATCGG
Read 4 ATGTTGATCGG
Read 5 ATGTTGATCGG
Read 6 ATGTTGATCGG
Nascent RNA
Mature RNA
Nascent RNA
Mature RNA
Gene
5-bromouridine 5′-triphosphate
4-thiouridine
Alkylated 4 sU
RNA polymerase II


a Structurome
b RRI analysis
c RNA–protein interactions
Nuclease digestion
Chemical marking
PARS
SHAPE-seq
SHAPE-MaP
e.g. SPLASH
e.g. CLIP-seq
RNA processing
Nuclease digestion
Fragmentation and biotin affinity purification
Fragmentation and immunoprecipitation
Library preparation
Proximity ligation and fragmentation
RNA-seq
Data analysis
dsRNA
ssRNA
Gene n
log norm reads
UV crosslink
Biotin
RNA fragmentation
Chemical crosslink
Nucleases
RNase V1 (digests dsRNA)
RNase S1 (digests ssRNA)
Chemical marks
cDNA 'block'
cDNA 'mutate'
RNA fragment
Adaptor (no UMI)
UMI adaptor
cDNA fragment

IN CLASS ACTIVITY