Weill Cornell Medicine
Pathology & Laboratory Medicine

Photo by Markus Spiske on Unsplash

RNA-Seq for Precision Medicine: translating research into clinical applications

Clinical and Research Genomics
Spring 2022 Course

Andrea Sboner, PhD (he/him/his)
Director of Informatics and Computational Biology

Weill Cornell Medicine
Caryl and Israel Englander
Institute for Precision Medicine

NewYork-Presbyterian

04.05.2022

What is Precision Medicine?

The right drug…

…to the right patient …

…at the right time!

Precision Medicine benefits from advancements in sequencing technology (NGS)

A very brief (and incomplete) history of genomics sequencing

- 3.1 billion nucleotides (DNA) sequenced and assembled (chr1, chr2, …, chr22, chrX, chrY) from a few individuals
- Costs: $3.8B → $796B economic impact
- 1990 – 2003 (13 years)

Economic Impact of the Human Genome Project

How a $3.8 billion investment drove $796 billion in economic impact, created 310,000 jobs and launched the genomic revolution

Simon Tripp and Martin Grueber
Battelle Memorial Institute
Technology Partnership Practice
May 2011

HapMap Project

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

1st draft Human Genome Reference Sequence

2nd draft Human Genome Reference Sequence

First Massively Parallel Sequencing Technology

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

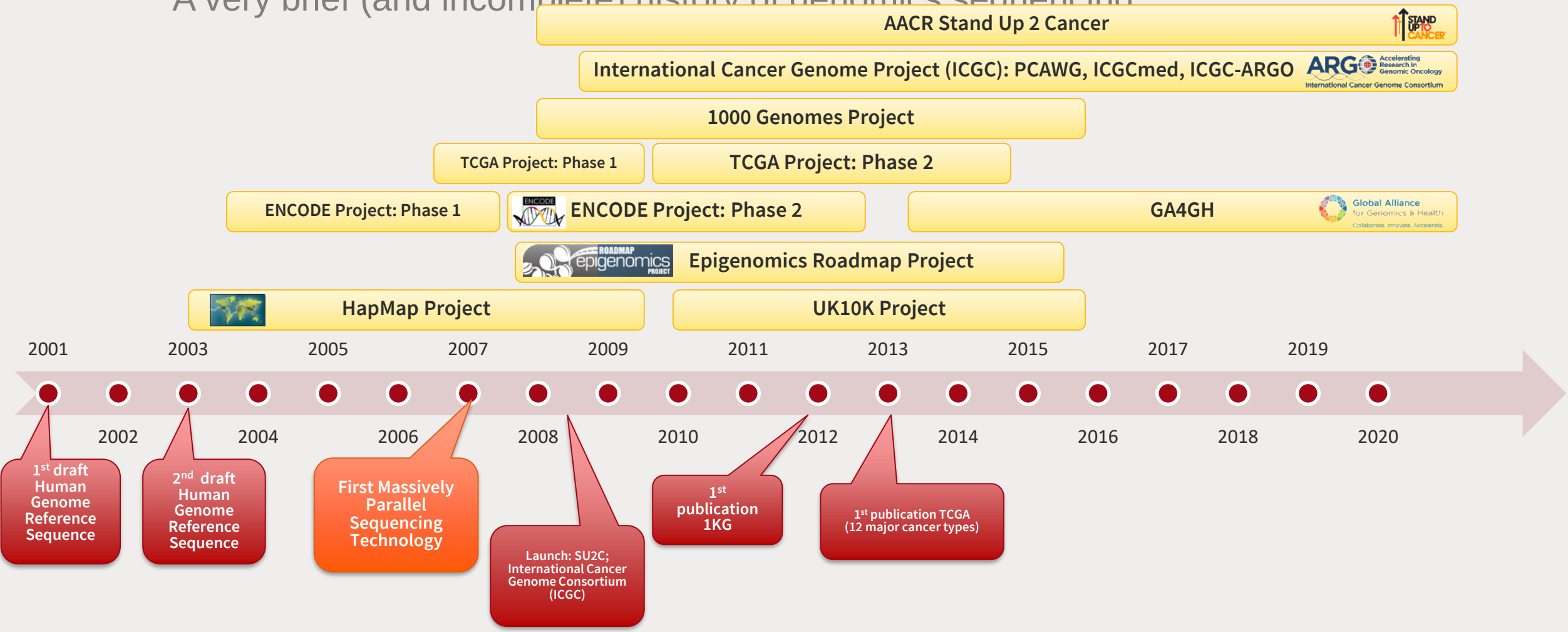
The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

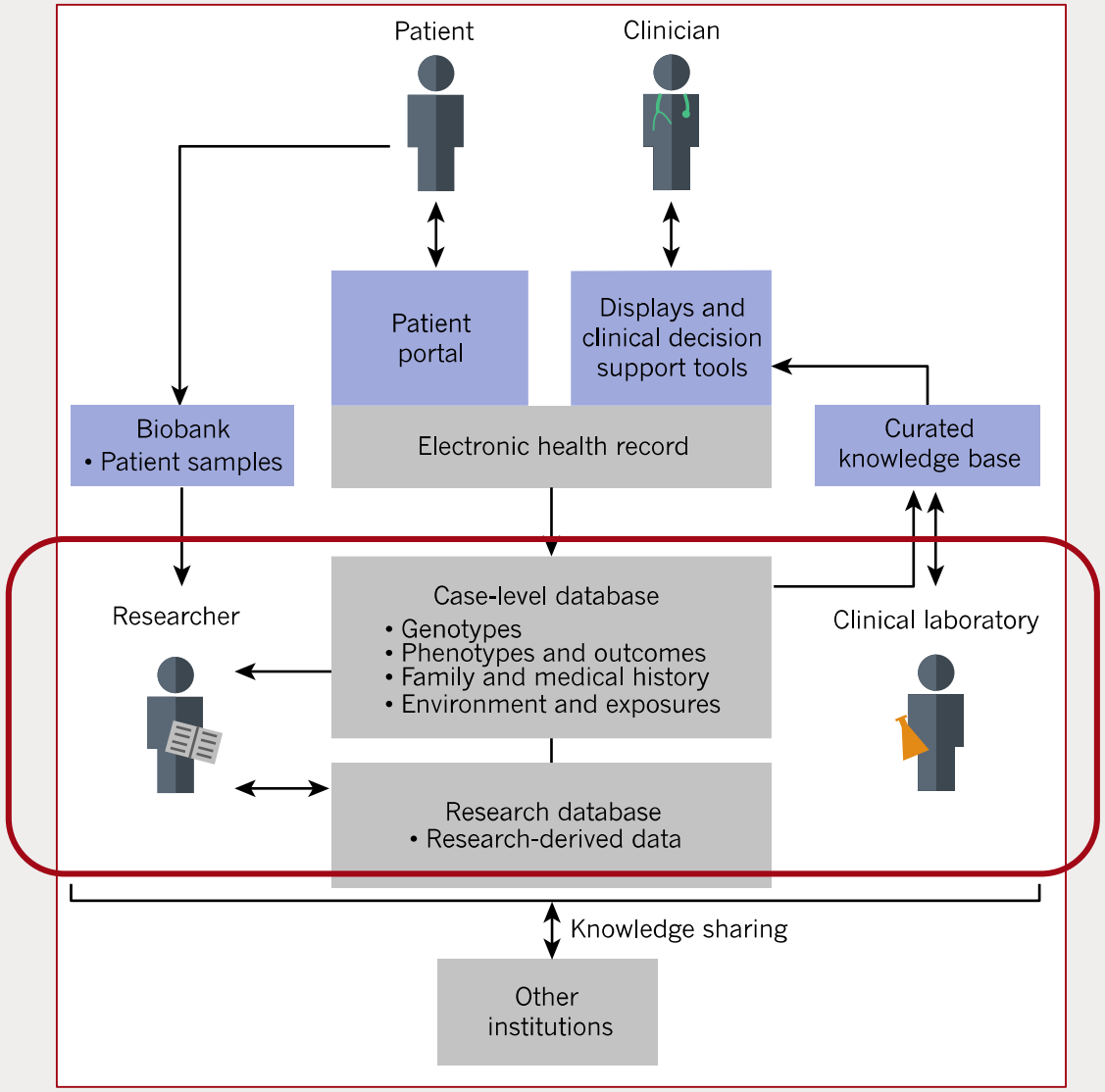
1st NGS
3.1B nucleotides from 1 individual
Jan 2007 – May 2007 (4 months)
Cost: $1.5M

Precision Medicine benefits from advancements in sequencing technology (NGS)

A very brief (and incomplete) history of genomics sequencing



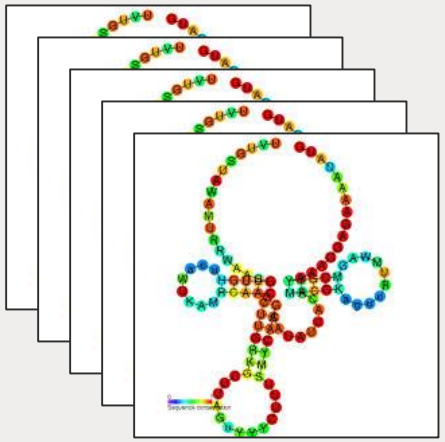
The Precision Medicine Ecosystem



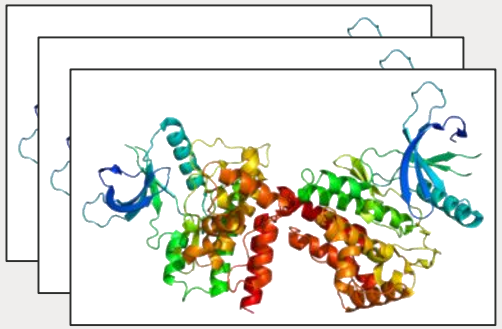
S. J. Aronson, H. L. Rehm, *Nature*. **526**, 336–342 (2015).

Transcriptome profiling

transcription →



translation →



Nucleotides
A,C,T,G

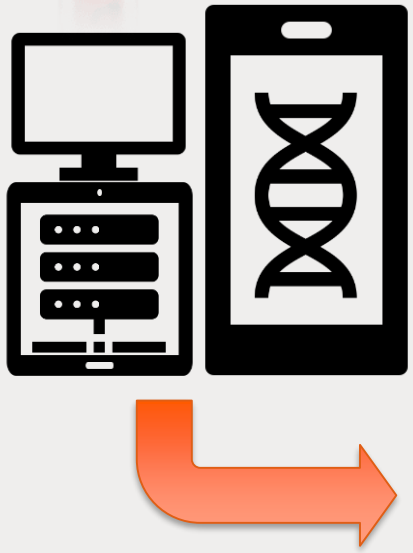
Nucleotides
A,C,U,G

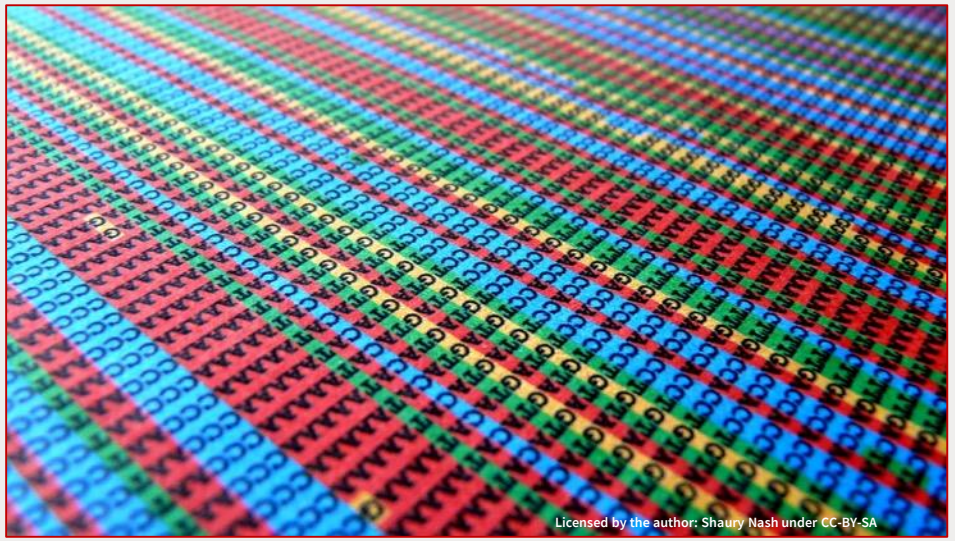
Amino Acids
A,R,N,D,C,E,Q,G,H,I,
L,K,M,F,P,S,T,W,Y,V

Transcriptome profiling goal is to characterize the RNA in a tissue or cell.

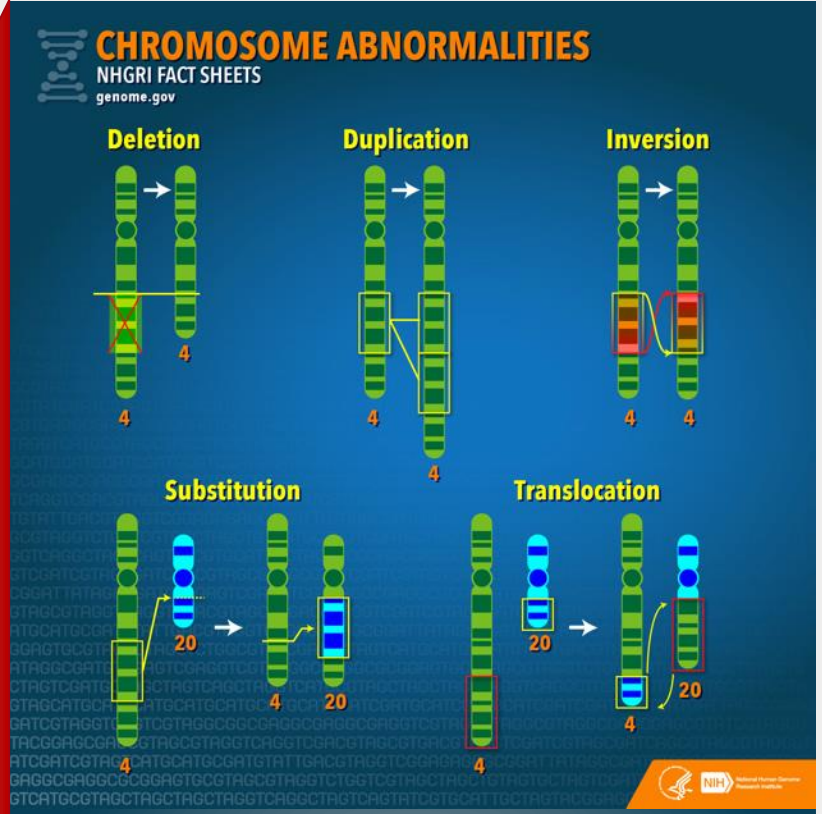
The 'simpler' structure of RNA allows to employ most techniques used for DNA analysis – hybridization, polymerase chain reaction, etc.

Sequencing & Analysis: A translation of biological material into a common language





Analysis



SNVs & InDels

```
Tumor:      ATTCGATTATTC
Reference:  ATTCGCTTATTC

Tumor:      ATTCGAACTTTATTC
Reference:  ATTCGC---TTATTC

Tumor:      ATTCG--TATTC
Reference:  ATTCGCTTATTC
```

Genome Era (1990s – 2000s)

~ 1991 Expressed Sequence Tags (ESTs) sequencing (500-800 nucleotides)

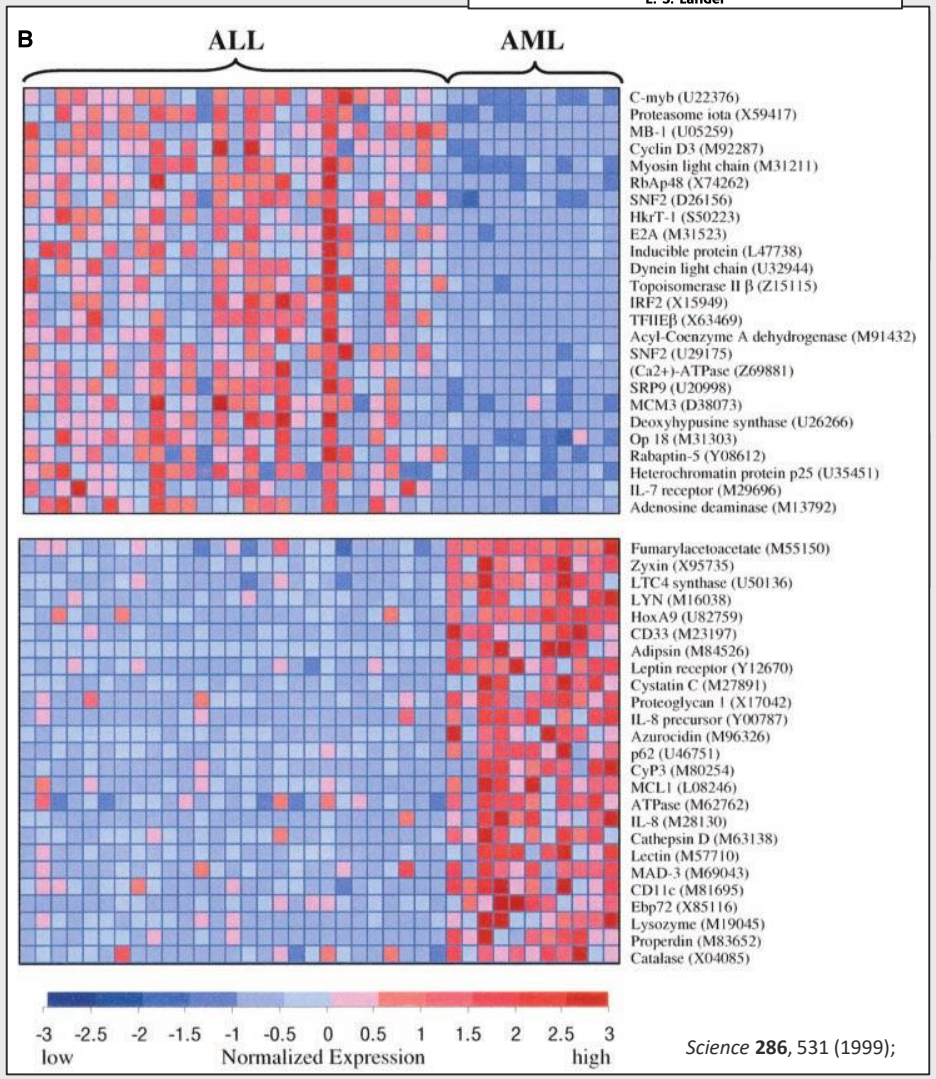
~ 1995 Series Analysis of Gene Expression (SAGE) (9-12 nucleotides)

~ 1999 Microarray

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]



Science **286**, 531 (1999);

RNA-Seq Experiment

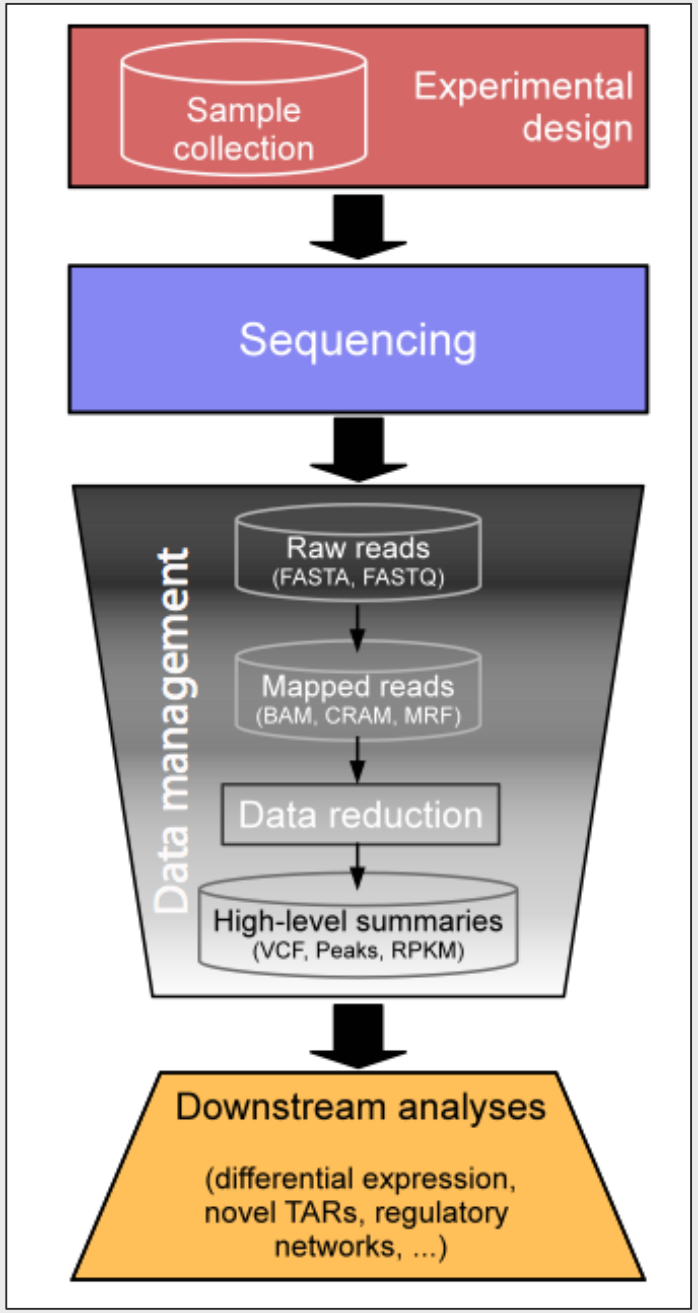
Data management:

Mapping the reads

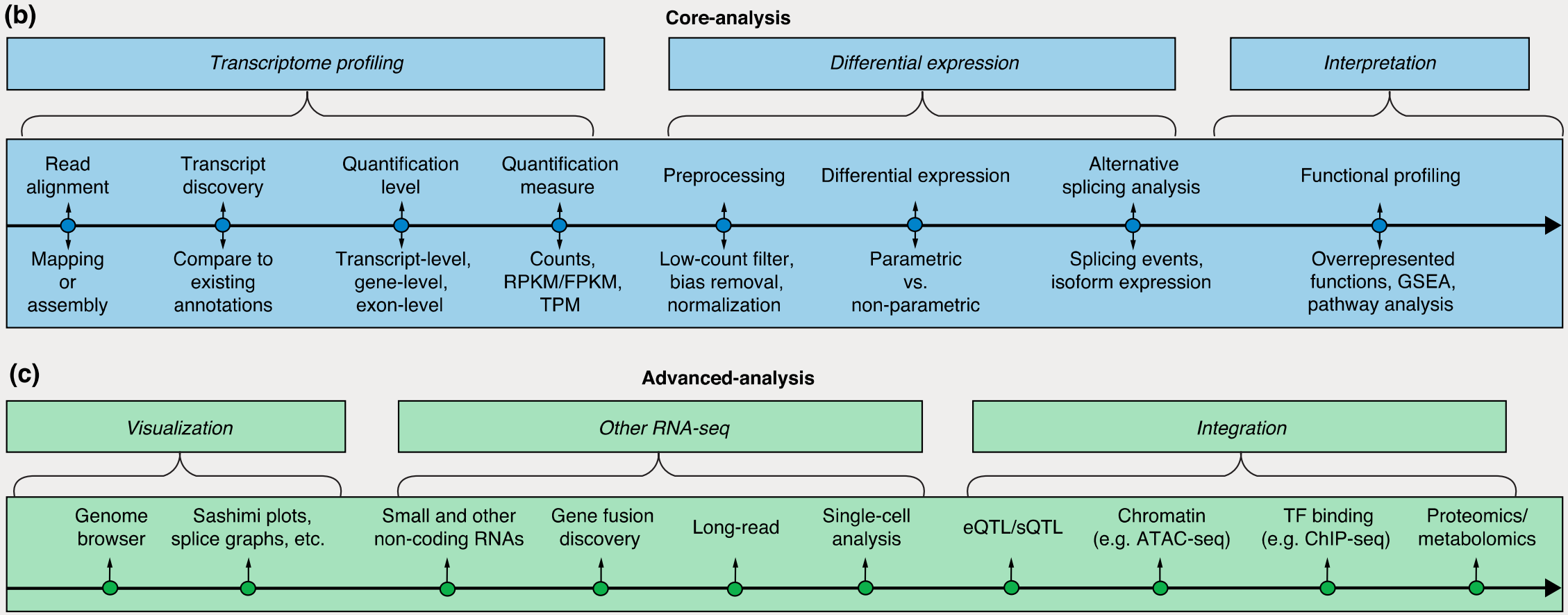
Creating summaries

Downstream analysis: *the interesting stuff*

Differential expression, chimeric transcripts, novel transcribed regions, etc.

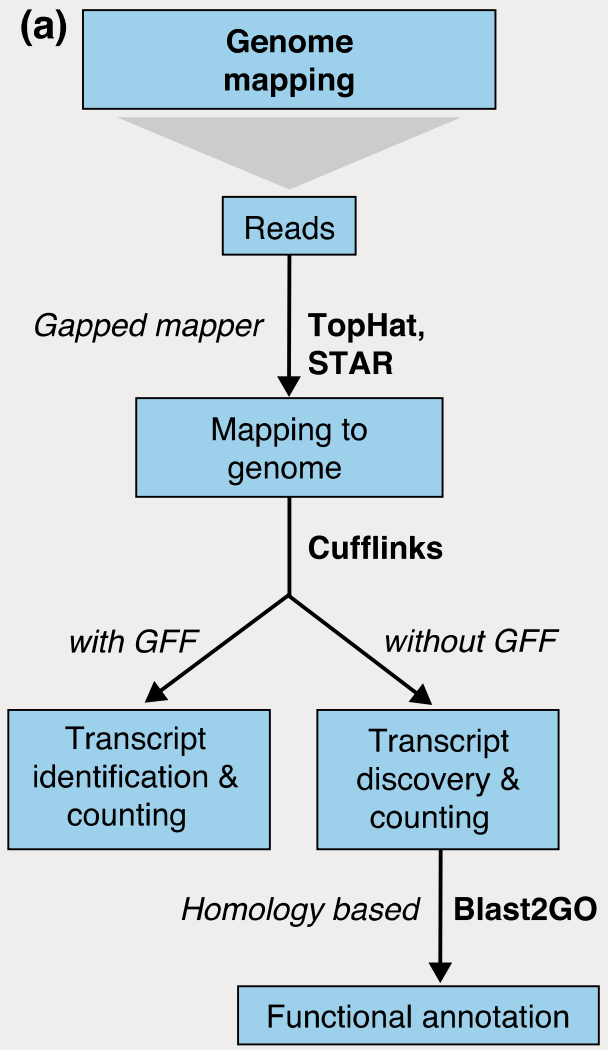


Roadmap for RNA-seq analyses



A. Conesa *et al.*, *Genome Biology*. **17**, 13 (2016).

Alignments for Transcriptomics (RNA-seq)



A. Conesa *et al.*, *Genome Biology*. **17**, 13 (2016).

Expression Quantification

FPKM/RPKM: Fragment/Reads per Kilobase of exonic region per Million of reads

Single end

Paired end

1. Divide the total # reads by 1M=scaling factor
2. Divide the read/fragment counts by the scaling factor=RPM/FPM
3. Divide RPM by the length of the genes in KB=RPKM/FPKM

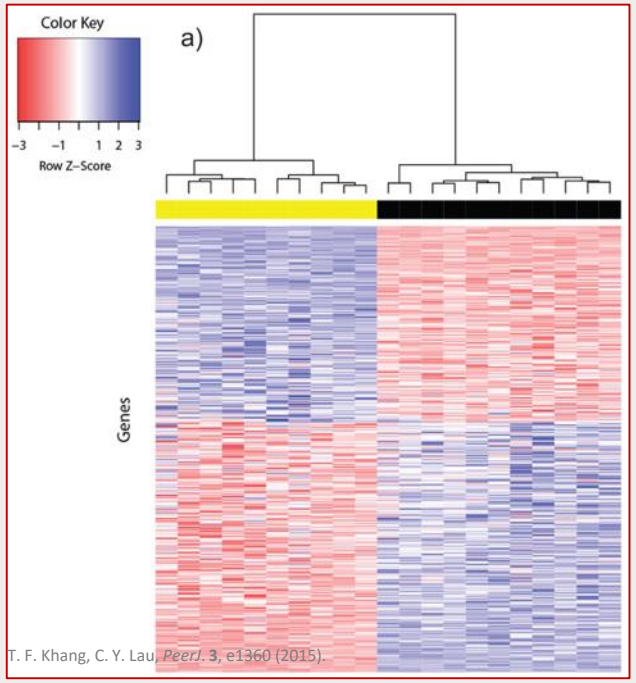
TPM: transcripts per million

1. Divide the read/fragment counts by the length of the gene=RPK
2. Sum all RPK and divide by 1M=scaling factor
3. Divide RPK by the scaling factor=TPM

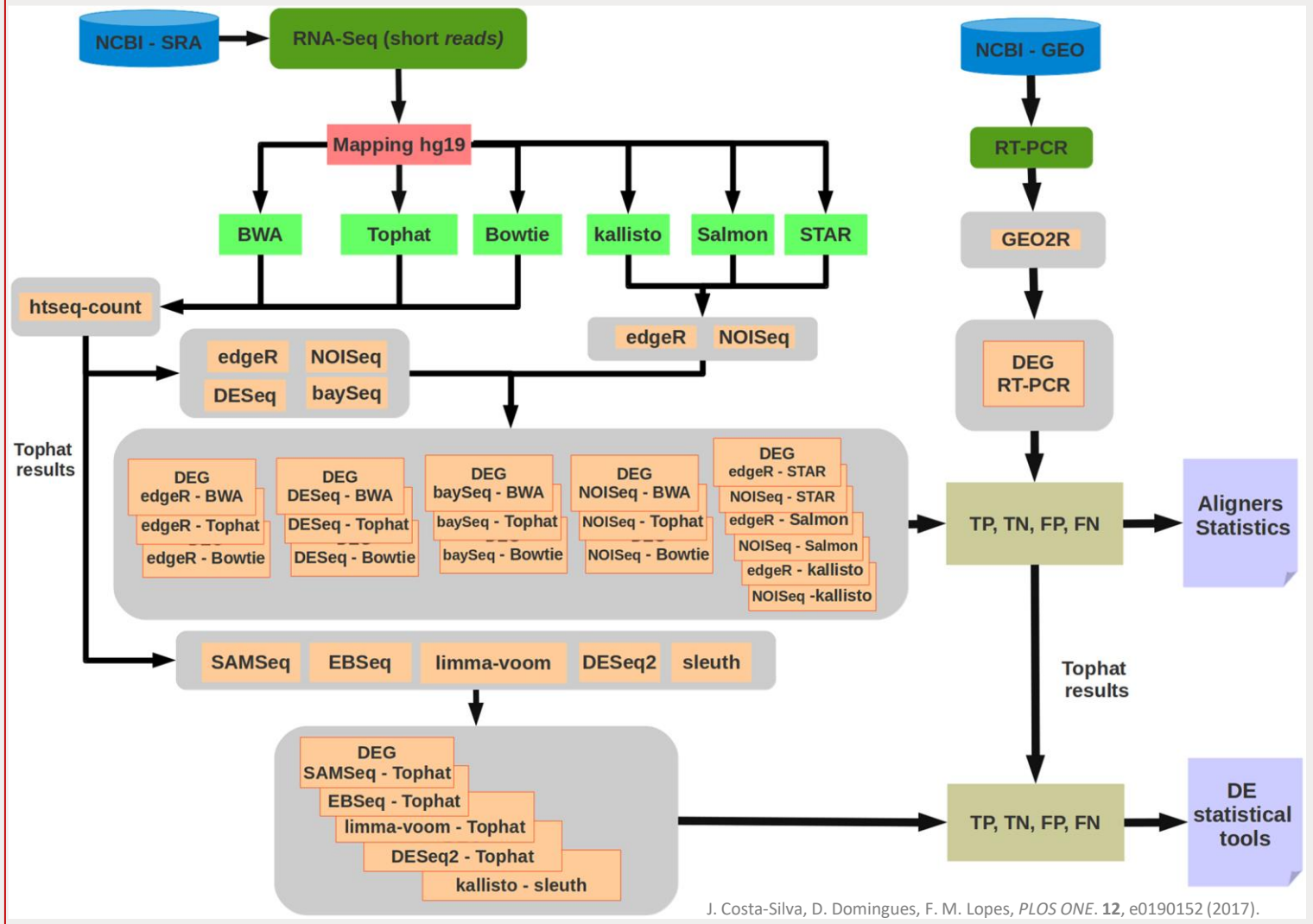
Normalization strategies affect results of comparisons (ERCC-spike-ins)

Differential Expression

Comparison of groups



T. F. Khang, C. Y. Lau, *PeerJ*. **3**, e1360 (2015).



J. Costa-Silva, D. Domingues, F. M. Lopes, *PLOS ONE*. **12**, e0190152 (2017).

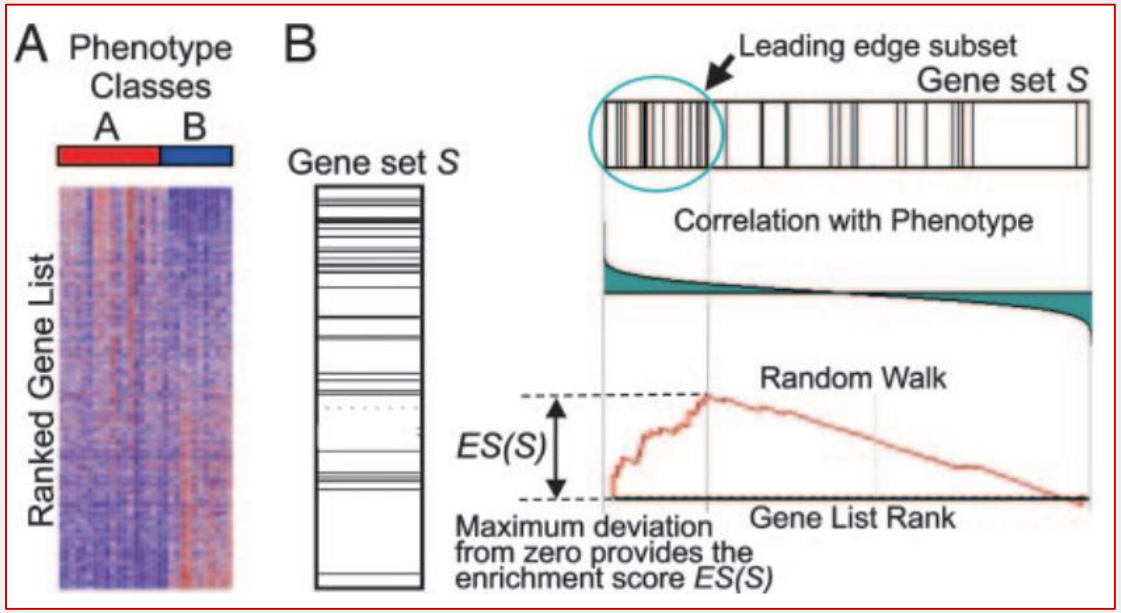
We have identified that the impact of the mapping tool on the final results is minimal, indicating the DEGs identification method is the main choice for differential expression analysis in RNA-Seq data.

We *did not* identify among the evaluated methods a tool that obtained *optimum results in all performance measures*, for the evaluated experimental conditions. The NOIseq, DESeq2 and limma+vomm methods present the best individual results with 95%, 95% and 93% of Specificity and 80%, 84% and 81% of True Positive Rate, respectively.

Pathway analysis (gene set enrichment)

Are a group of genes dys-regulated in a certain condition?

"The basic assumption is that although large changes in individual genes can have significant effects on pathways, *weaker but coordinated changes* in sets of functionally related genes (i.e., pathways) can also have significant effects. Therefore, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic (e.g. the sum of all log-fold changes), which will then be evaluated."



Introduction to differential gene expression analysis using RNA-seq Written by Friederike Dundar, Luce Skrabanek, Paul Zumbo https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

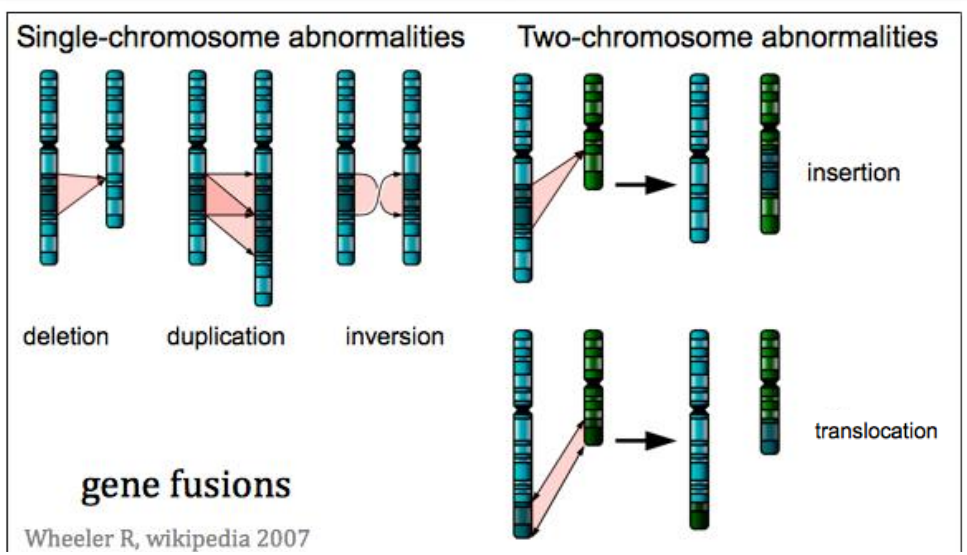
What are chimeric transcripts?

Transcripts that are *not co-linear* in the genome space

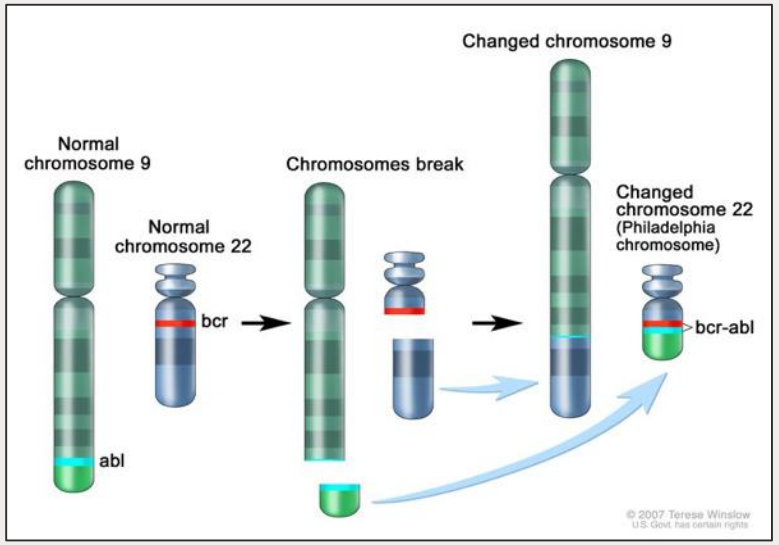
They can arise from:

genomic rearrangements, i.e. *gene fusions*

post-transcriptional events, i.e. *trans-splicing or cis-splicing*



Why are they (gene fusions) important?

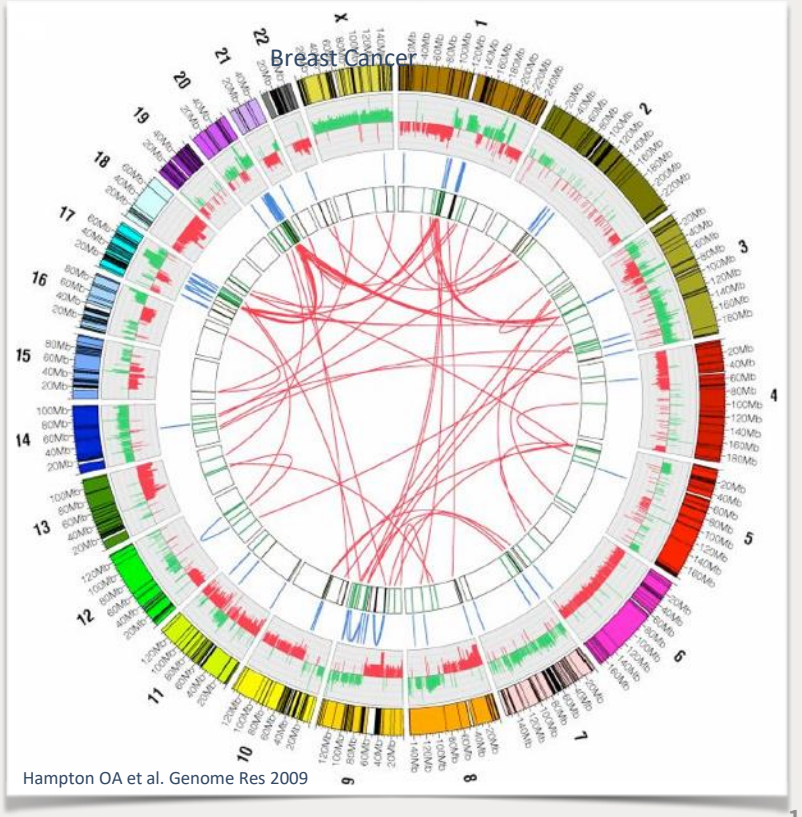


Fusion genes are often *oncogenes*

Ex: BCR-ABL1 (Philadelphia chromosome) in Chronic myelogenous leukemia (CML) and Acute Lymphoblastic leukemia (ALL) t(9;22)(q34;q11)

Fusion involving a proto-oncogene with a strong promoter resulting in *upregulation* (lymphomas)

Ex: (IgH locus)-MYC in Burkitt's lymphoma (cMYC over-expressed)



Hampton OA et al. Genome Res 2009

Why are they (trans-splicing events) important?

Trans(cis)-splicing was initially found in lower eukariotes, such as trypanosomes and worms

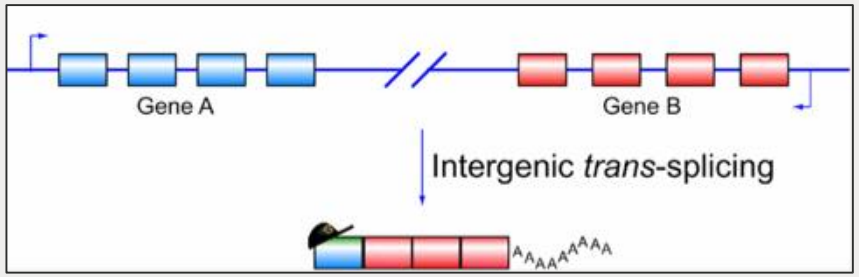
Short sequences of nucleotides are trans-spliced to distant 5' of many protein coding genes

Recently, they were found in mammalian cells:

JAZF1-SUZ12 in endometrial stroma cells (Li et al. Science 2008)

SLC45A3-ELK4 in prostate tissues (Rickman et al. Cancer Res 2009)

65% of protein-coding genes have distal 5' transcription start sites (ENCODE pilot) --> revised to ~50% the ENCODE 2012



Horiuchi, Takayuki, and Toshiro Aigaki. Biology of the Cell 98, no. 2 (January 9, 2012): 135–140.

An historical perspective of gene fusions

1914

Theodore Boveri hypothesized that cancer originates from chromosomal damage

1950s

Improved cytogenetics techniques

1960

Novel & Hungerford report a *recurrent* minute chromosome in chronic mylogenous leukemia (CML)

1970s

Chromosome banding enables the identification of specific arms and regions. Several rearrangements are identified in hematological malignancies: CML, BL, APL, follicular lymphoma

1980s-1990s

The list of rearrangements and tumor types grows including most mesenchymal tumors: e.g. Ewing sarcoma, alveolar rhabdomyo-sarcoma

Early 2000s

Most rearrangements were discovered mainly in hematological and mesenchymal neoplasms compared to epithelial

2005

The first fusion (TMPRSS2-ERG) in a common epithelial tumor (prostate cancer) was discovered (by bionformatics analysis)

2005 - Present

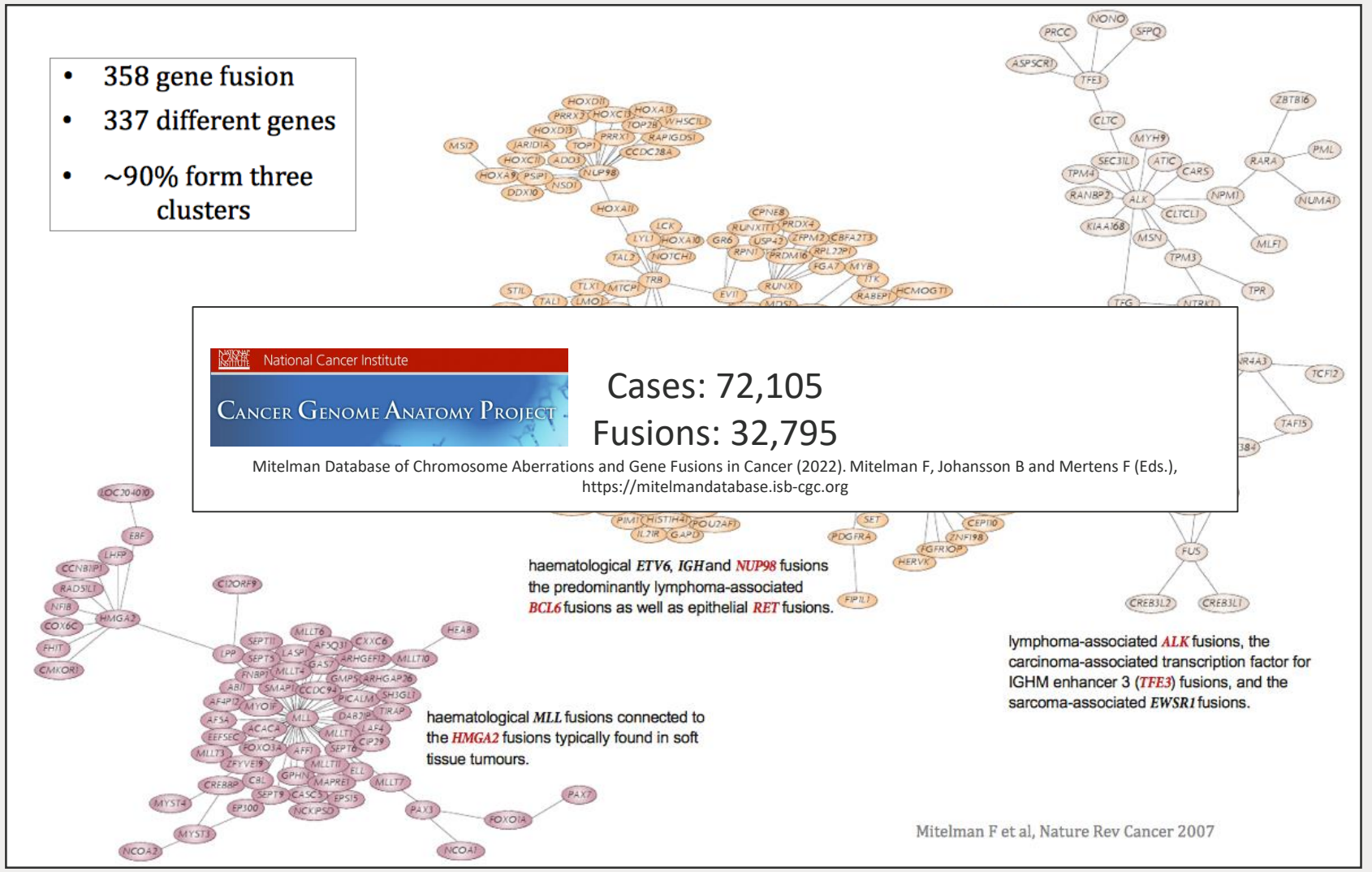
Several thousands fusion events have been characterized, mainly from NGS approaches

2018

Targeted therapy approved pan-cancer for NTRK fusions

Mitelman, F. and Heim, S. (2015) How it all began, in Cancer Cytogenetics: Chromosomal and Molecular Genetic Aberrations of Tumor Cells, Fourth Edition (eds S. Heim and F. Mitelman), John Wiley & Sons, Ltd, Chichester, UK

How many different gene fusions do we know?



Gene fusions are important for clinical treatment


Open Access Review

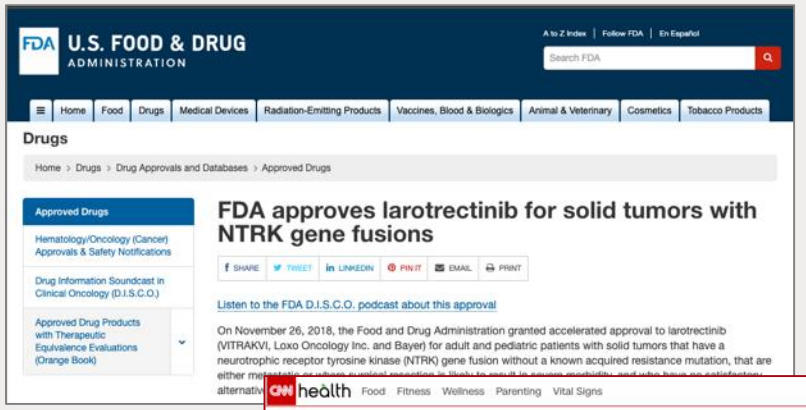
ESMO Open Cancer Horizons

CrossMark

NTRK gene fusions as novel targets of cancer therapy across multiple tumour types

Alessio Amatu,[1] Andrea Sartore-Bianchi,[1] Salvatore Siena[1,2]

Amatu A, et al. ESMO Open 2016;1:e000023. doi:10.1136/esmoopen-2015-000023



… and diagnostic/prognostic purposes

RESEARCH ARTICLE

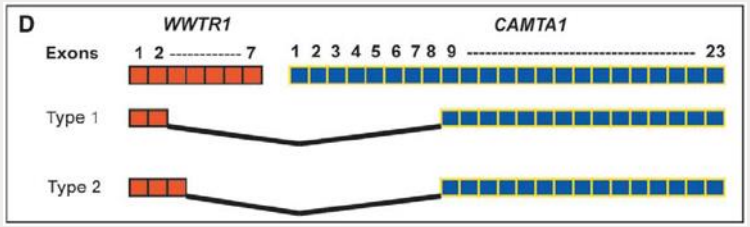
CANCER

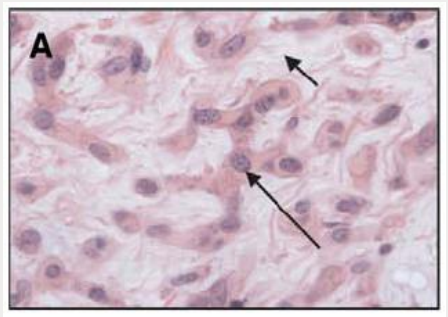
Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma

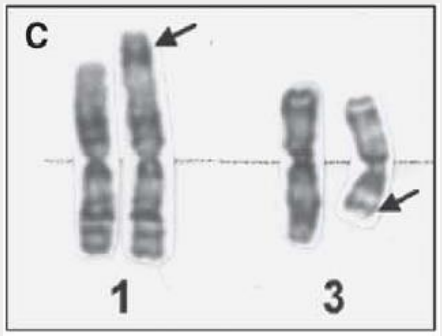
Munir R. Tanas,[1] Andrea Sboner,[2] Andre M. Oliveira,[3] Michele R. Erickson-Johnson,[3] Jessica Hespelt,[1] Philip J. Hanwright,[1] John Flanagan,[4] Yuling Luo,[4] Kerry Fenwick,[5] Rachael Natrajan,[5] Costas Mitsopoulos,[5] Marketa Zvelebil,[5] Benjamin L. Hoch,[6] Sharon W. Weiss,[7] Maria Debiec-Rychter,[8] Raf Sciot,[9] Rob B. West,[10] Alexander J. Lazar,[11] Alan Ashworth,[5] Jorge S. Reis-Filho,[5] Christopher J. Lord,[5] Mark B. Gerstein,[2,12] Mark A. Rubin,[13] Brian P. Rubin[1]*

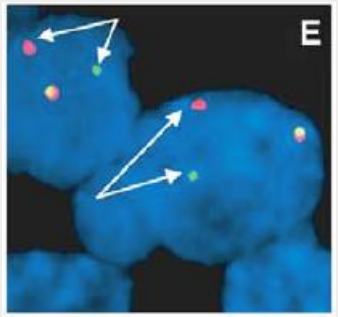
Science Translational Medicine AAAS

www.ScienceTranslationalMedicine.org 31 August 2011 Vol 3 Issue 98 98ra82





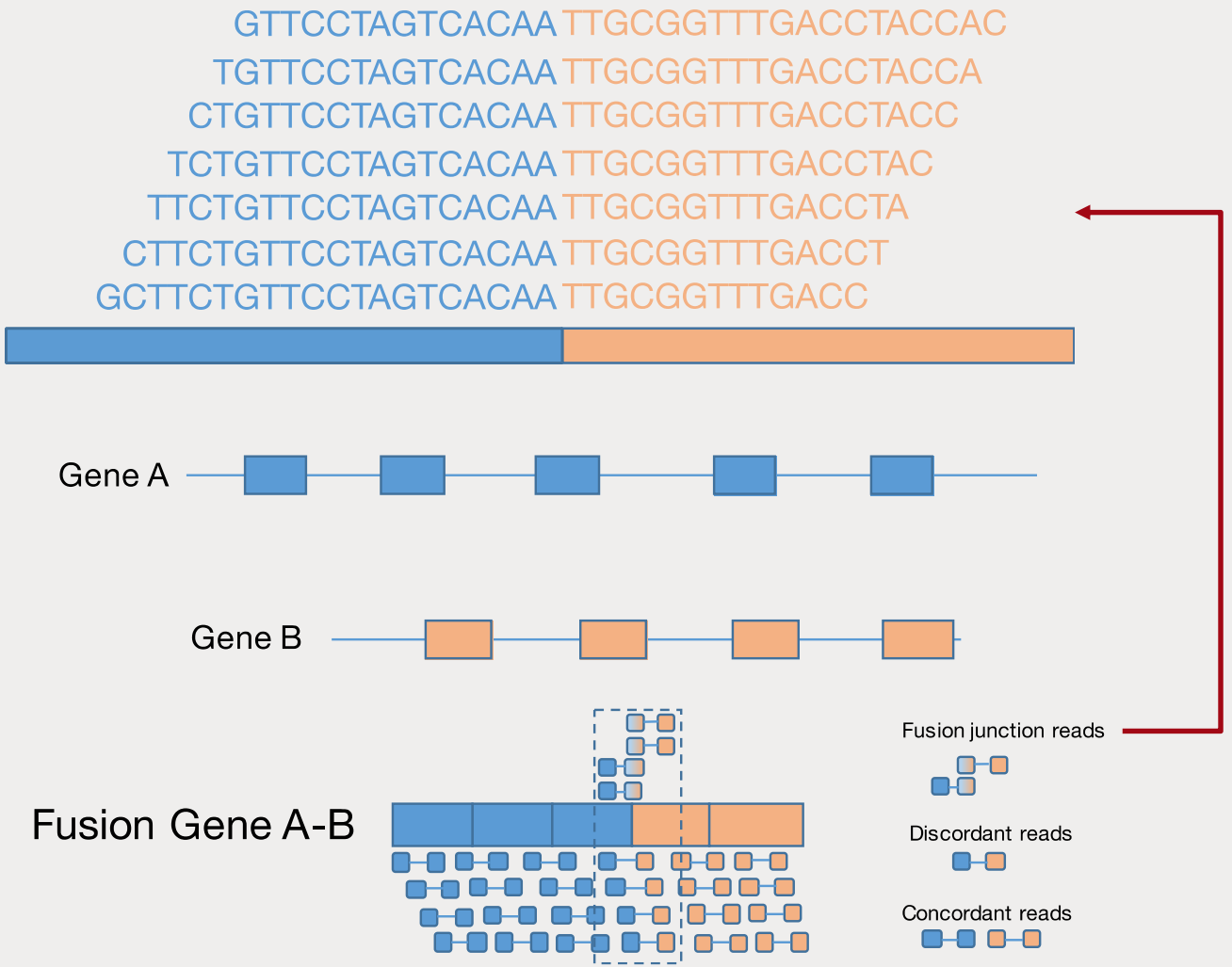




Exclusively present in *epitheliod hemangioendothelioma*

G	*WWTR1*		*CAMTA1*	
	Positive /total	%	Positive /total	%
Epithelioid hemangioendothelioma	42/47	89%	39/45	87%
Epithelioid angiosarcoma	0/7	0%	0/7	0%
Intimal sarcoma	0/5	0%	0/3	0%
Kaposi's sarcoma	0/4	0%	0/4	0%
Malignant hemangioendothelioma, NOS	0/1	0%	0/1	0%
Retiform hemangioendothelioma	0/1	0%	0/1	0%
Kaposiform hemangioendothelioma	0/3	0%	0/2	0%
Epithelioid hemangioma	0/5	0%	0/4	0%
Arteriovenous malformation	0/2	0%	0/2	0%
Angiomatosis	0/1	0%	0/1	0%
Hemangioma, NOS	0/3	0%	0/3	0%
Capillary/pyogenic hemangioma	0/5	0%	0/5	0%
Cavernous hemangioma	0/5	0%	0/5	0%
Juvenile hemangioma	0/1	0%	0/1	0%
Spindle cell hemangioma	0/4	0%	0/4	0%
Synovial hemangioma	0/1	0%	0/1	0%
Intramuscular hemangioma	0/6	0%	0/5	0%
Littoral cell hemangioma	0/6	0%	0/2	0%
Malignant hemangiopericytoma	0/1	0%	0/1	0%
Hemangiopericytoma, NOS	0/1	0%	0/1	0%
Sinonasal hemangiopericytoma	0/1	0%	0/1	0%
Glomus tumor	0/1	0%	0/1	0%
Atypical glomus tumor	0/2	0%	0/2	0%
Lymphangioma	0/7	0%	0/7	0%
Lymphangioleiomyomatosis	0/1	0%	0/1	0%
Papillary endothelial hyperplasia	0/2	0%	0/2	0%
Total cases	165		151	

Fusion supporting reads



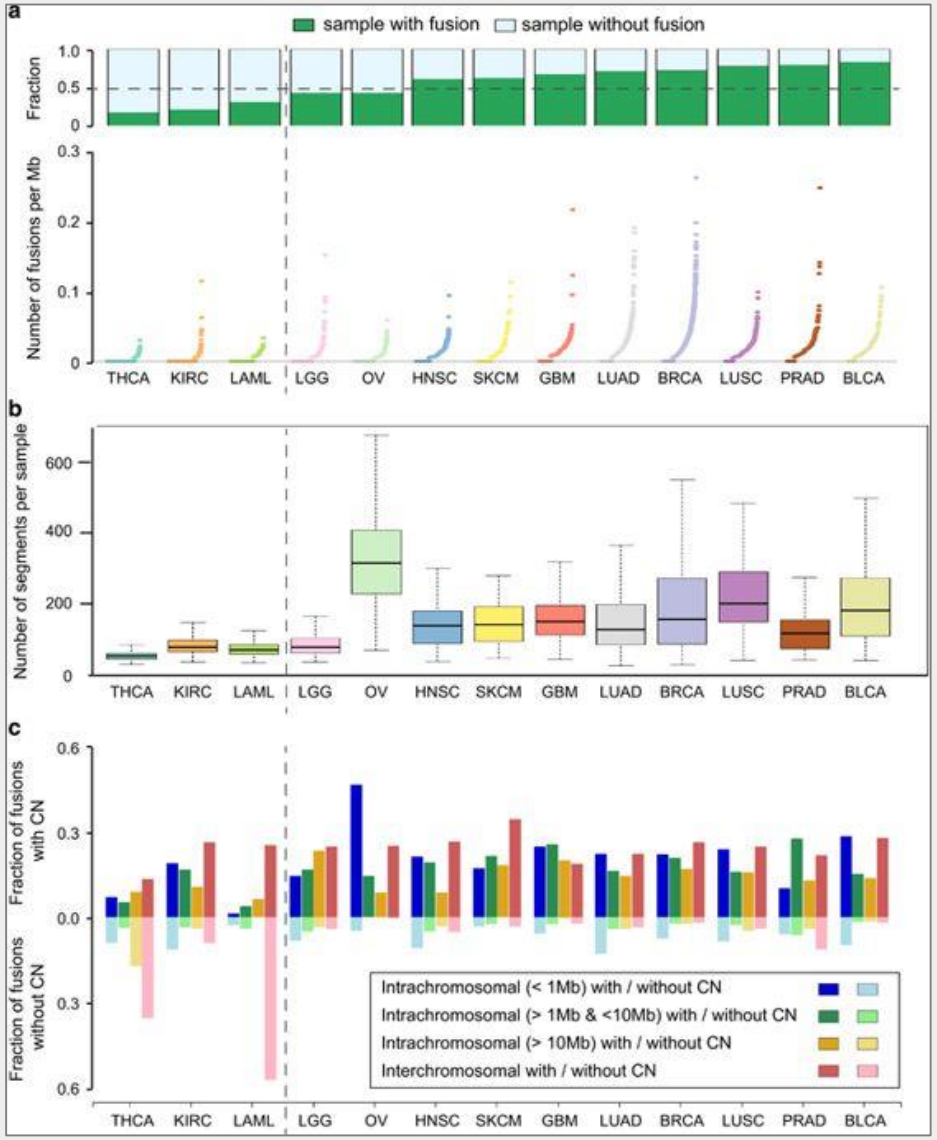
Spectrum of fusions in cancer types

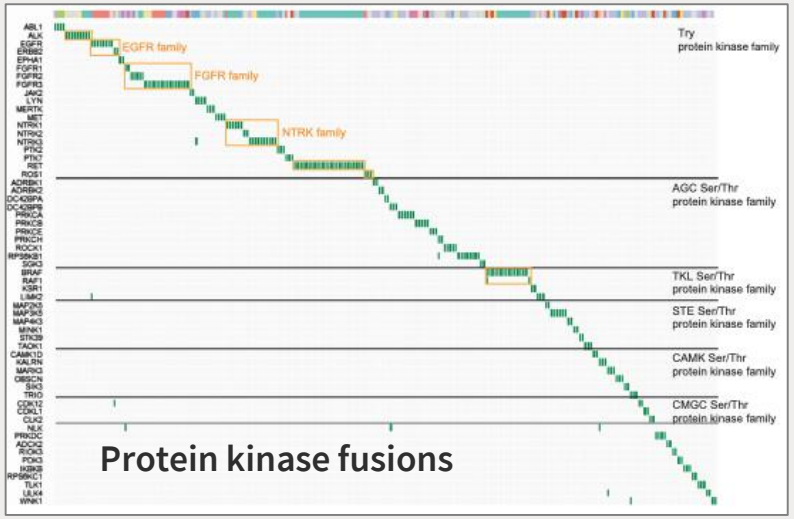
Oncogene (2015) 34, 4845–4854
© 2015 Macmillan Publishers Limited All rights reserved 0950-9232/15
www.nature.com/onc

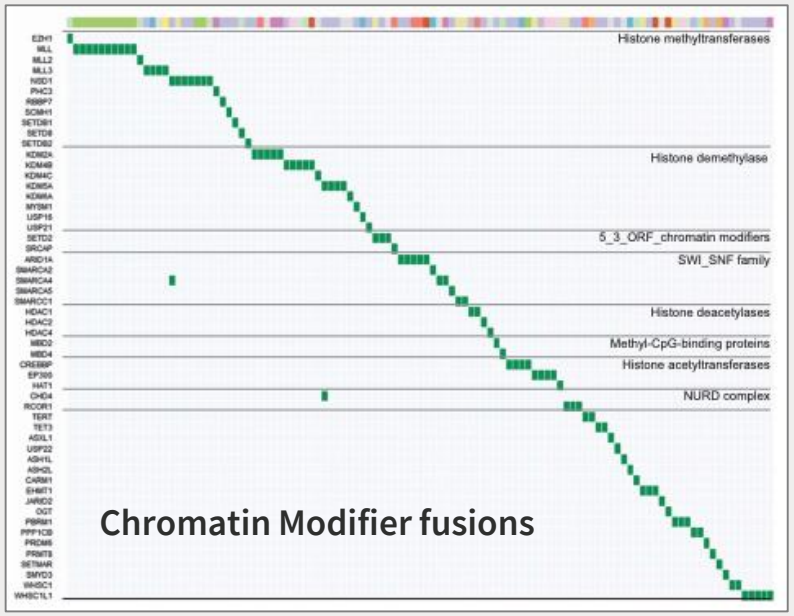
ORIGINAL ARTICLE

The landscape and therapeutic relevance of cancer-associated transcript fusions

K Yoshihara[1,2], Q Wang[1], W Torres-Garcia[1], S Zheng[1], R Vegesna[1], H Kim[1] and RGW Verhaak[1,3]



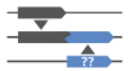




https://www.tumorfusions.org/

“Targeted” fusion detection methods

ARCHER® FUSIONPlex® NGS Assays



Novel fusions
AMP™ chemistry utilizes open-ended targeted amplification to identify gene fusions whether or not the fusion partner is known Video Blog

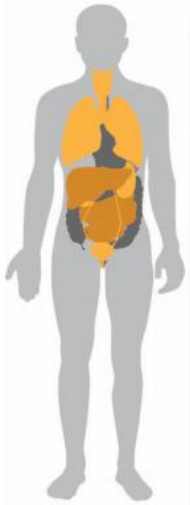
Expression
Molecular barcodes coupled with open-ended amplification allows for determination of RNA vs DNA reads Blog

nanoString

- Breast Cancer 360™ Panel
- PanCancer IO 360™ Panel
- CAR-T Characterization Panel
- Hallmarks of Cancer Collection
  - PanCancer Pathways Panel
  - PanCancer Immune Profiling Panel
  - PanCancer Progression Panel
- PlexSet™ Pre-selected Panels
- Vantage 3D™ RNA Panels
- Vantage 3D™ Gene Fusion Panels
- miRNA Panels
- Kinase Panel
- Stem Cell Panel
- nCounter Gene Fusion Panels (Ex-US only)

Prepare Library | Sequence | Analyze Data illumina®

TruSight™ Oncology 500



NTRK1, NTRK2, NTRK3 (pan-cancer) | MSI (pan-cancer)

Lung	Melanoma	Colon	Ovarian	Breast	Gastric	Bladder	Myeloid	Sarcoma
AKT1 *ALK* *BRAF* *DDR2* *EGFR* *ERBB2* *FGFR1* *FGFR3* *KRAS* *MAP2K1* *MET* *NRAS* *PIK3CA* *PTEN* *RET* *TP53* TMB	*BRAF* *CTNNB1* *GNA11* *GNAQ* *KIT* *MAP2K1* *NF1* *NRAS* *PDGFRA* *PIK3CA* *PTEN* *TP53*	*AKT1* *BRAF* *HRAS* *KRAS* *MET* *MLH1* *MSH2* *MSH6* *NRAS* *PIK3CA* *PMS2* *PTEN* *SMAD4* *TP53*	*BRAF* *BRCA1* *BRCA2* *KRAS* *PDGFRA* *FOXL2* *TP53*	*AKT1* *AR* *BRCA1* *BRCA2* *ERBB2* *FGFR1* *FGFR2* *PIK3CA* *PTEN*	*BRAF* *KIT* *KRAS* *MET* *MLH1* *PDGFRA* *TP53*	*MSH6* *PMS2* *TSC1*	*ABL1* *ASXL1* *CALR* *CEBPA* *ETV6* *EZH2* *FLT3* *GATA2* *IDH1* *IDH2* *JAK2* *KIT* *MPL* *NPM1* *RUNX1* *SF3B1* *SRSF2* *TP53*	*ALK* *APC* *BRAF* *CDK4* *CTNNB1* *ETV6* *EWSR1* *FOXO1* *GLI1* *KIT* *MDM2* *MYOD1* *NAB2* *NF1* *PAX3* *PAX7* *PDGFRA* *PDGFRB* *SDHB* *SDHC* *SMARCB1* *TFE3* *WT1*

FDA-approved drugs targeting oncogenic fusions in solid tumors

Fusion target	Therapy	Indication	FDA approval
ALK fusion	Crizotinib	Lung	August 2011
	Ceritinib	Lung	May 2017
	Alectinib	Lung	November 2017
	Brigatinib	Lung	May 2020
	Lorlatinib	Lung (second line)	November 2018
FGFR fusion	Erdafitinib	Urothelial	April 2019
	Pemigatinib	Cholangiocarcinoma	April 2020
ROS1 fusion	Crizotonib	Lung	March 2016
	Entrectinib	Lung	August 2019
RET fusion	Selpercatinib	Lung Thyroid	May 2020
	Pralsetinib	Lung	September 2020
NTRK1/2/3 fusion	Larotrectinib	Solid tumor	November 2018
	Entrectinib	Solid tumor	August 2019
PDGFB fusion	Imatinib	DFSP	November 2006
MET exon 14 skipping	Capmatinib	Lung	May 2020

Adapted from Dr. James Solomon

Tools for detecting fusion transcripts

Gene fusion detection software tools | RNA sequencing

High-throughput sequencing software tools > RNA sequencing software tools

http://omictools.com/gene-fusion-detection-category

De novo transcriptome assembly software tools | RNA sequencing

High-throughput sequencing software tools > RNA sequencing software tools

http://omictools.com/transcriptome-assembly-category

RNA-seq short-reads "only"
Bellerophontes
BreakFusion
chimeraScan
CRAC
deFuse
EricScript
FusionAnalyser
FusionCatcher
FusionFinder
FusionHunter
FusionQ
FusionSeq
Jaffa
MapSplice
PRADA
shortFuse
SnowShoes-FTD
SOAPFuse/Fusion
TopHat-Fusion
STAR-fusion

RNA-seq & DNA-seq
BreakTrans
Comrad
nFuse

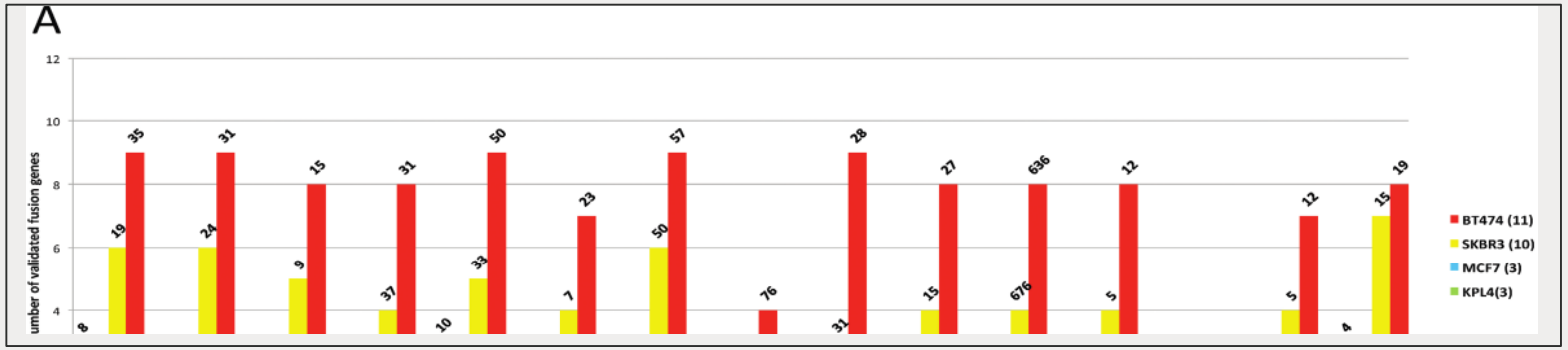
Gene fusion annotation
Chimera
Pegasus

Transcript Assembly
CuffLinks
Scripture
Trinity
Trans-Abyss

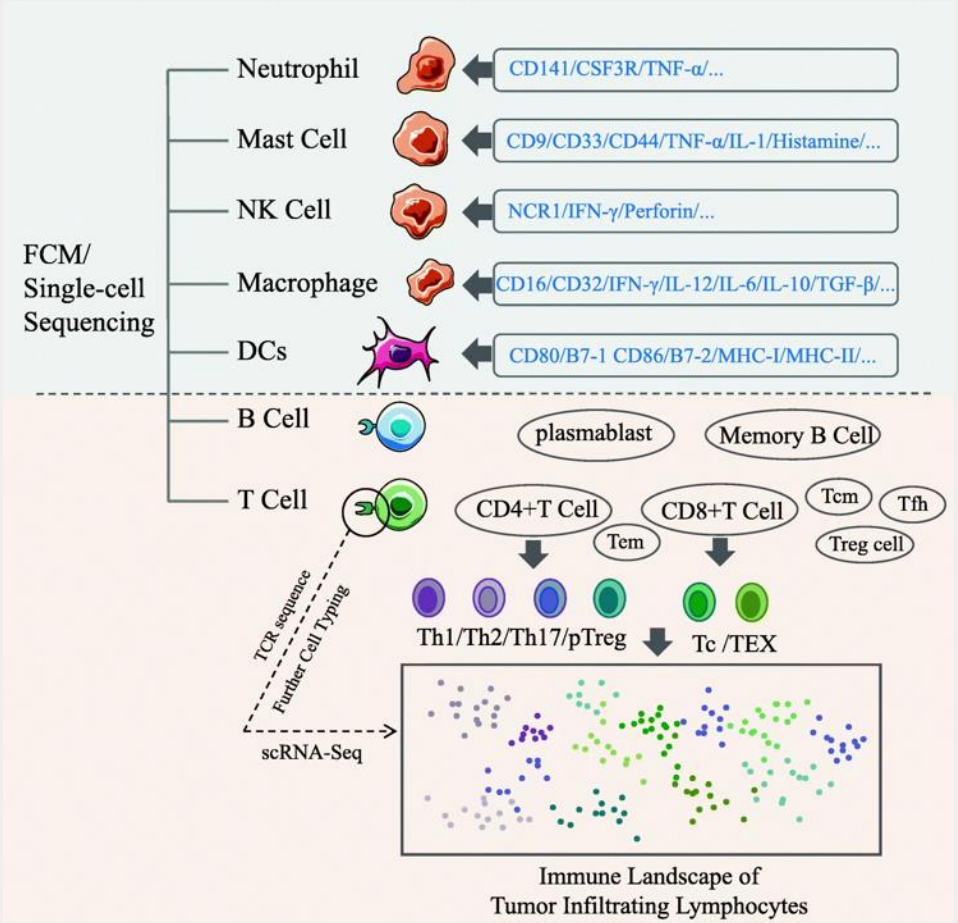
Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, No. 5 e47*
doi: 10.1093/nar/gkv1234

Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data

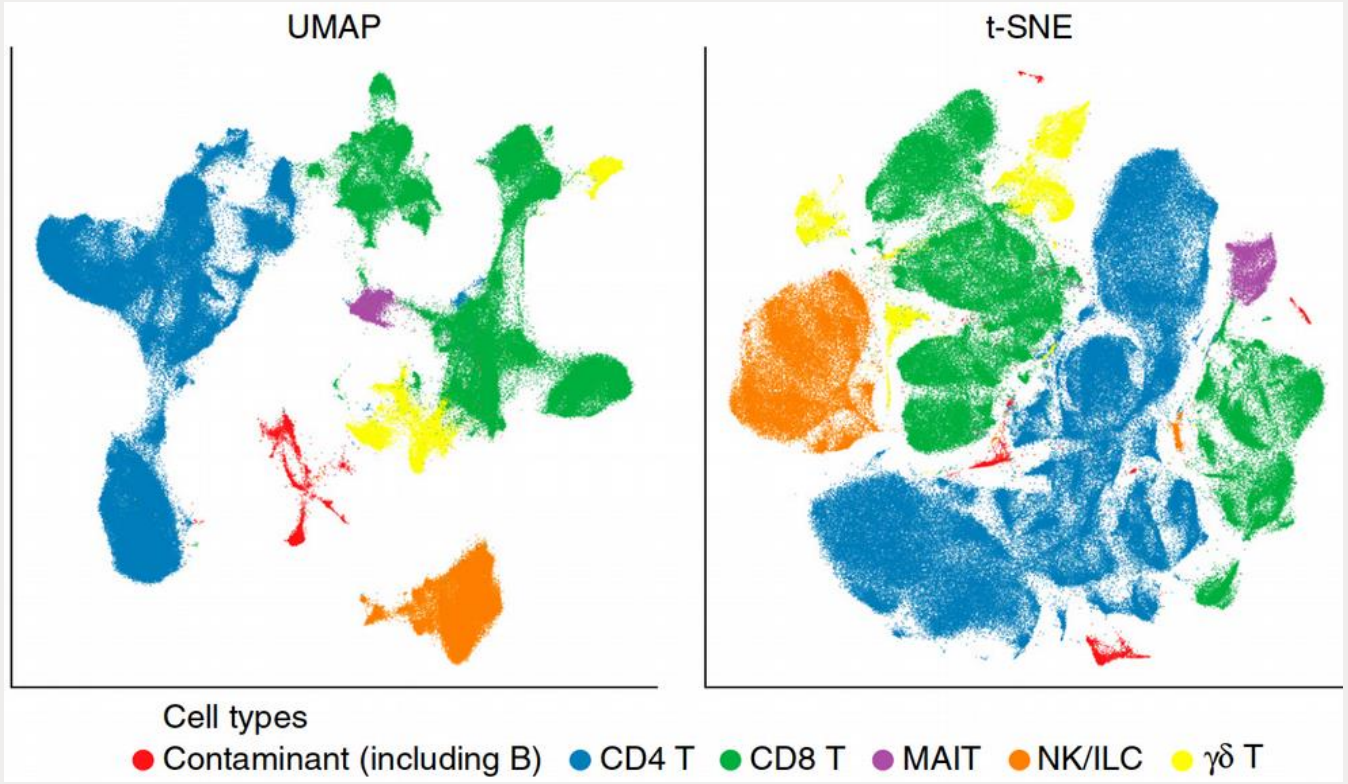
Silvia Liu[1,2,†], Wei-Hsiang Tsai[3,†], Ying Ding[1,2,†], Rui Chen[1], Zhou Fang[1], Zhiguang Huo[1], SungHwan Kim[1], Tianzhou Ma[1], Ting-Yu Chang[4], Nolan Michael Priedigkeit[5], Adrian V. Lee[6], Jianhua Luo[7], Hsei-Wei Wang[3,4,8,*], I-Fang Chung[3,8,*] and George C. Tseng[1,2,*]



Single-cell RNA sequencing (scRNA-seq): higher magnification using the NGS microscope



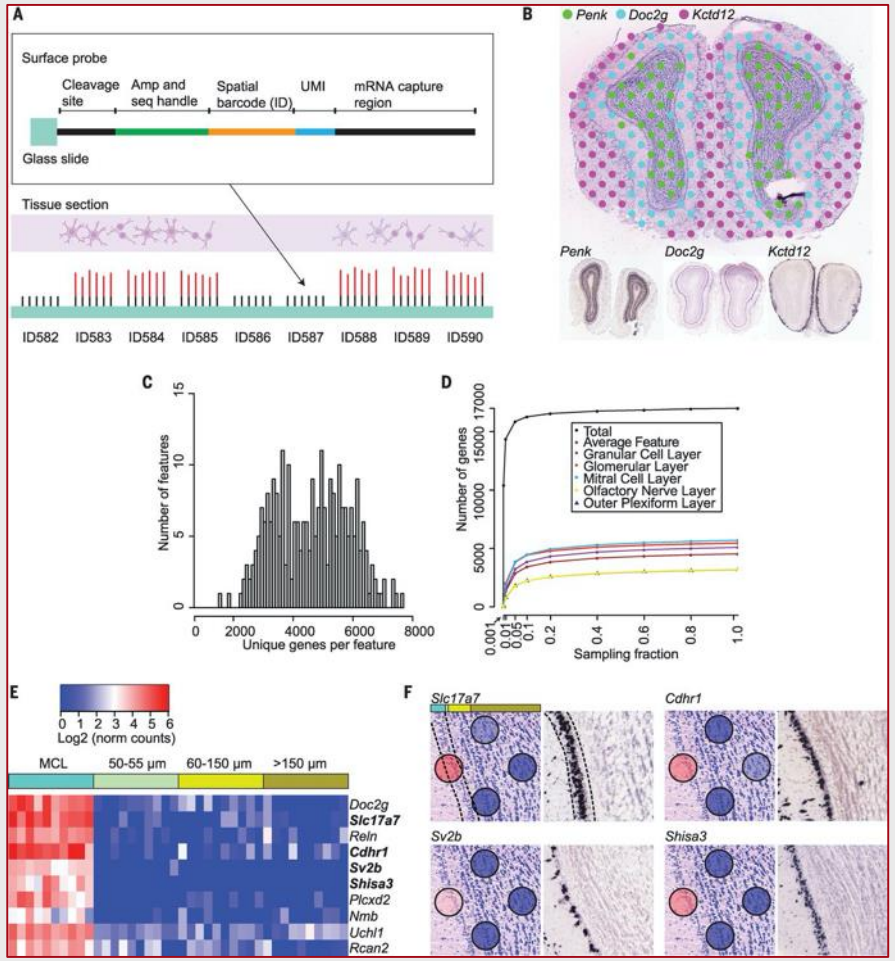
Zhang *et al. Journal of Experimental & Clinical Cancer Research* (2021) 40:81
https://doi.org/10.1186/s13046-021-01874-1



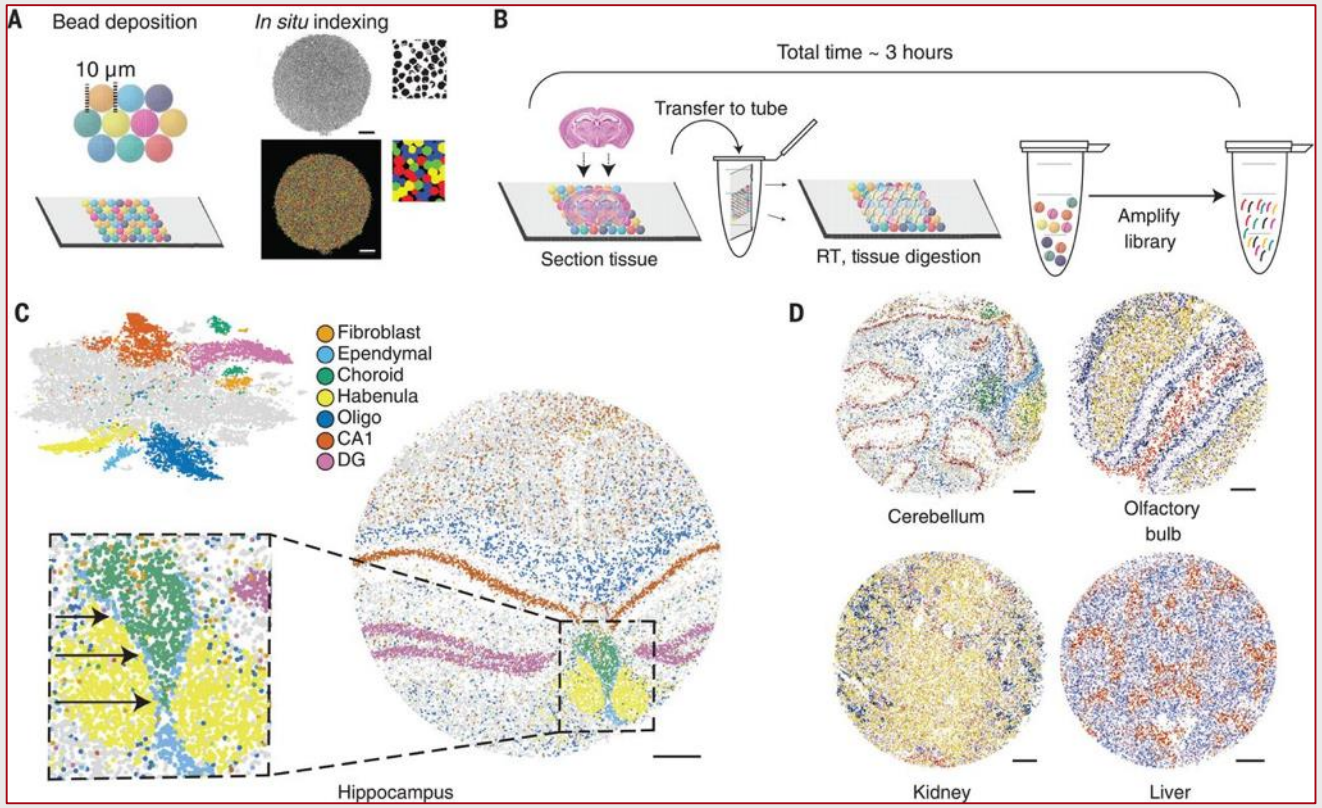
https://towardsdatascience.com/reduce-dimensions-for-single-cell-4224778a2d67

Spatial profiling

Measurements (such as gene expression) that maintain the spatial information.



P. L. Ståhl *et al.*, *Science*. **353**, 78–82 (2016).



S. G. Rodriques *et al.*, *Science*. **363**, 1463–1467 (2019).

Summary and future directions

- Massively Parallel Sequencing has enabled the discovery of additional fusion transcripts
- Specificity is the main challenge: too many false positives (FPs)!

- Longer reads: could help overcome the limitations of short reads
- Combination of tools may help further improve on the reduction of FPs

- "For the large bioinformatics community, development of a high-performing (accurate and fast) fusion detection tool or methods to combine top-performing tools remains an important and open question"

ans2077@med.cornell.edu
@asbonerw

Weill Cornell Medicine
Caryl and Israel Englander Institute for Precision Medicine

Weill Cornell Medicine
Pathology & Laboratory Medicine

Weill Cornell Medicine
Clinical Genomics Lab

NewYork-Presbyterian