Weill Cornell Medicine NewYork-Presbyterian

RNA-Seq methods and gene fusions: libraries, case reports, and algorithms

Clinical and Research Genomics

Spring 2018 Course

Andrea Sboner

03.21.2018

Outline

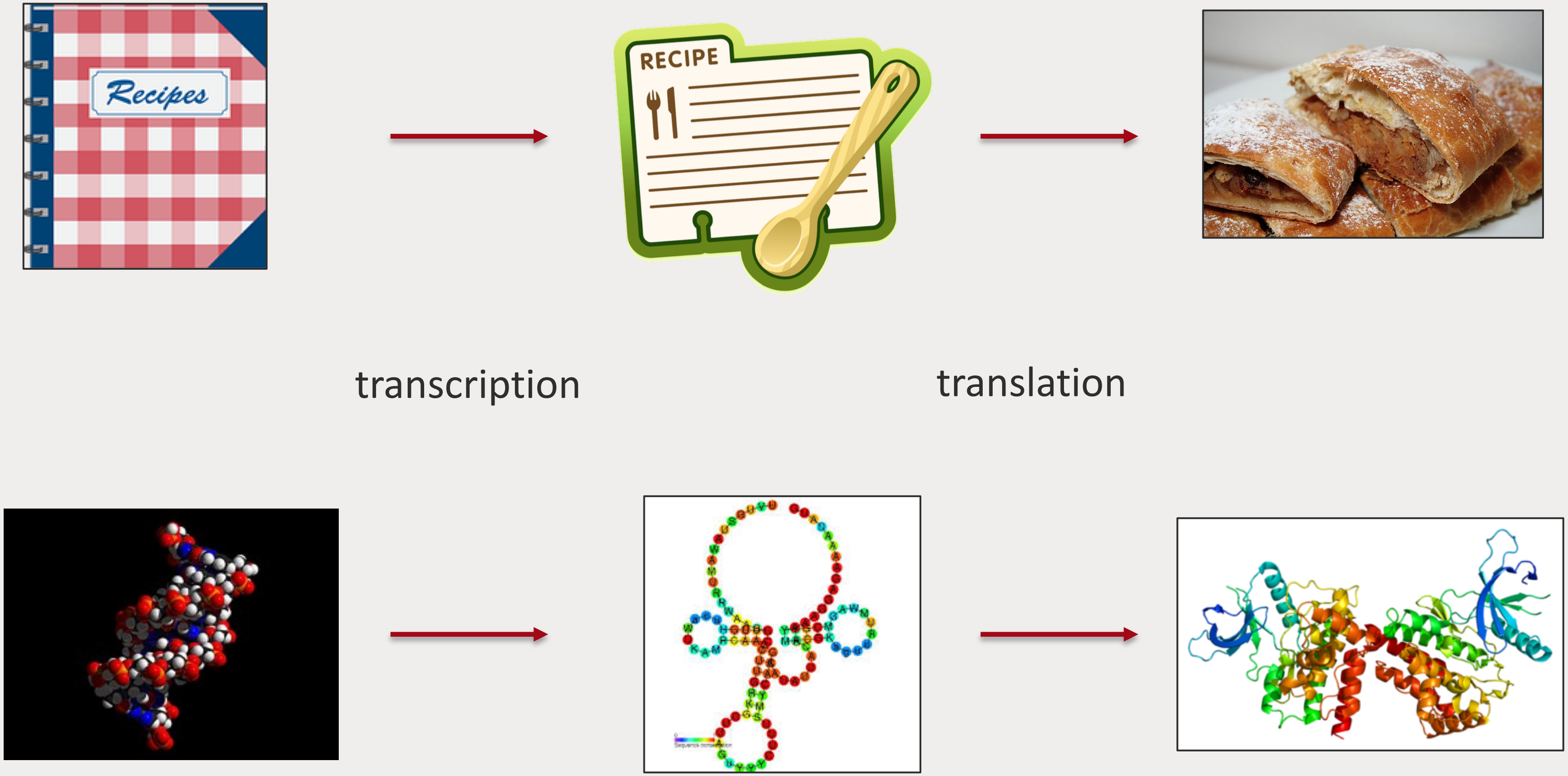
- Background of transcriptome profiling
- Next Generation Sequencing: a revolution in molecular biology
- RNA-seq application: gene fusion detection

Images throughout the presentation from pixabai.com, commons.wikimedia.org

Central dogma of molecular biology

Recipes
RECIPE

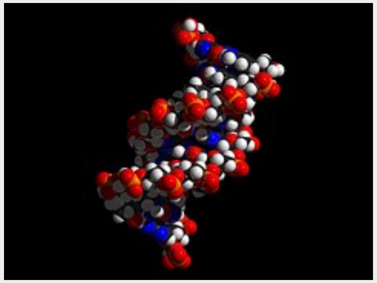
Central dogma of molecular biology



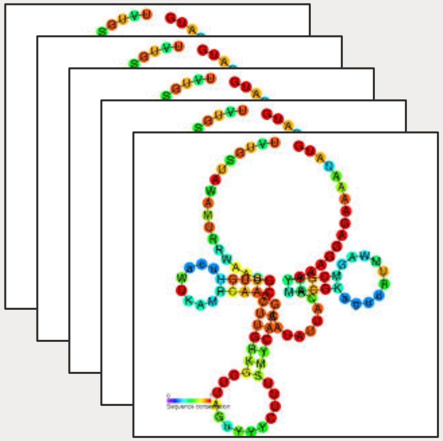
transcription

translation

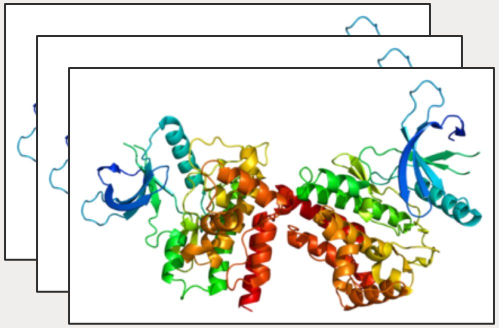
Transcriptome profiling



transcription



translation



Nucleotides
A,C,T,G

Nucleotides
A,C,U,G

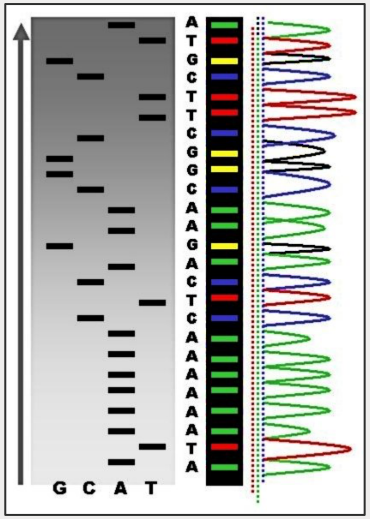
Amino Acids
A,R,N,D,C,E,Q,G,H,I,
L,K,M,F,P,S,T,W,Y,V

Transcriptome profiling goal is to characterize 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.

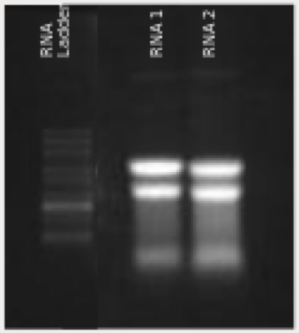
Pre-genome era (< 1990s)

~1970 Reverse Transcriptase → Allows the *reverse* transcription of RNA to DNA, generating cDNA



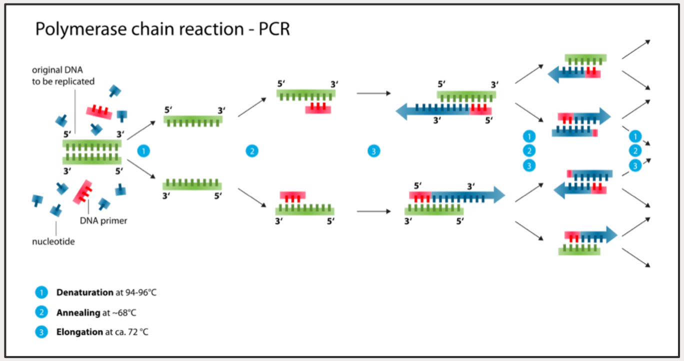
~ 1977 Sanger Sequencing → enables to 'read' the sequence of DNA

~ 1977 Northern blot → enable to measure the expression of RNA



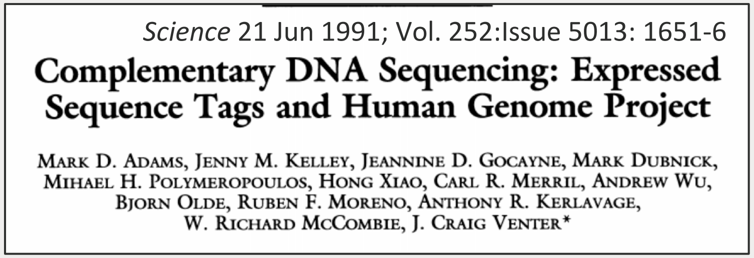
Formaldehyde gel (1%) with RNA samples run at 100V for 1 hour in 1x MOPS buffer.

~1983 Polymerase Chain Reaction (PCR) → allows the duplication/amplification of pieces of DNA



Genome Era (1990s – 2000s)

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



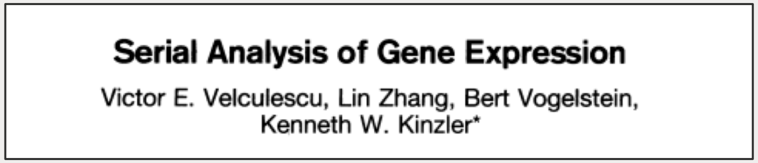
Science 21 Jun 1991; Vol. 252:Issue 5013: 1651-6

Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project

Mark D. Adams, Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R. Merril, Andrew Wu, Bjorn Olde, Ruben F. Moreno, Anthony R. Kerlavage, W. Richard McCombie, J. Craig Venter*

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

Science 20 Oct 1995:Vol. 270, Issue 5235, pp. 484-487

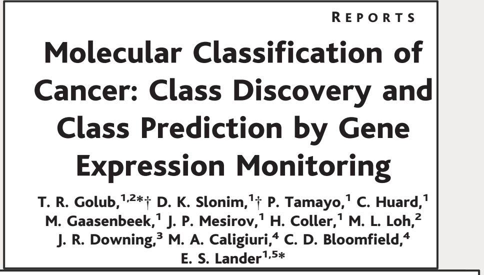


Serial Analysis of Gene Expression

Victor E. Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W. Kinzler*

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

Genome Era (1990s – 2000s)



REPORTS

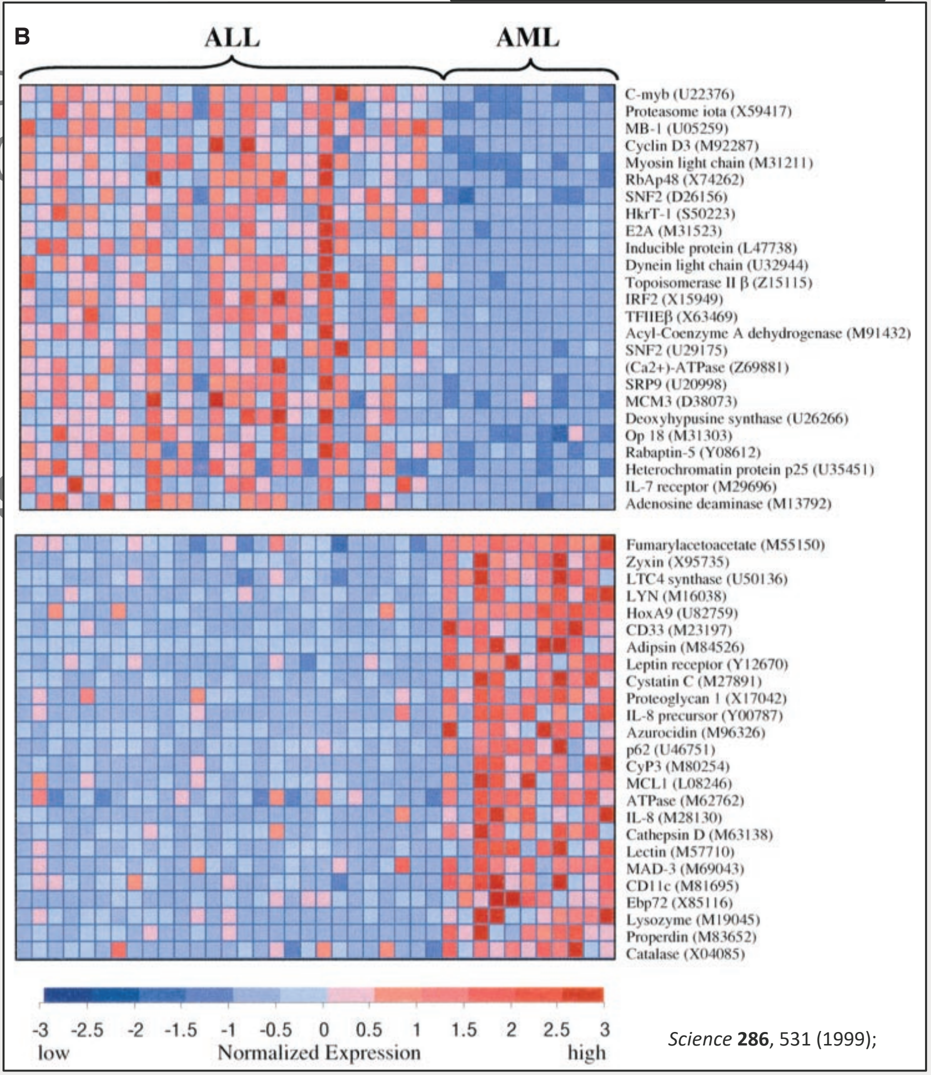
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

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*]

- ~ 1991 Expressed Sequenc
  Tags (ESTs) sequencing (50
  800 nucleotides)

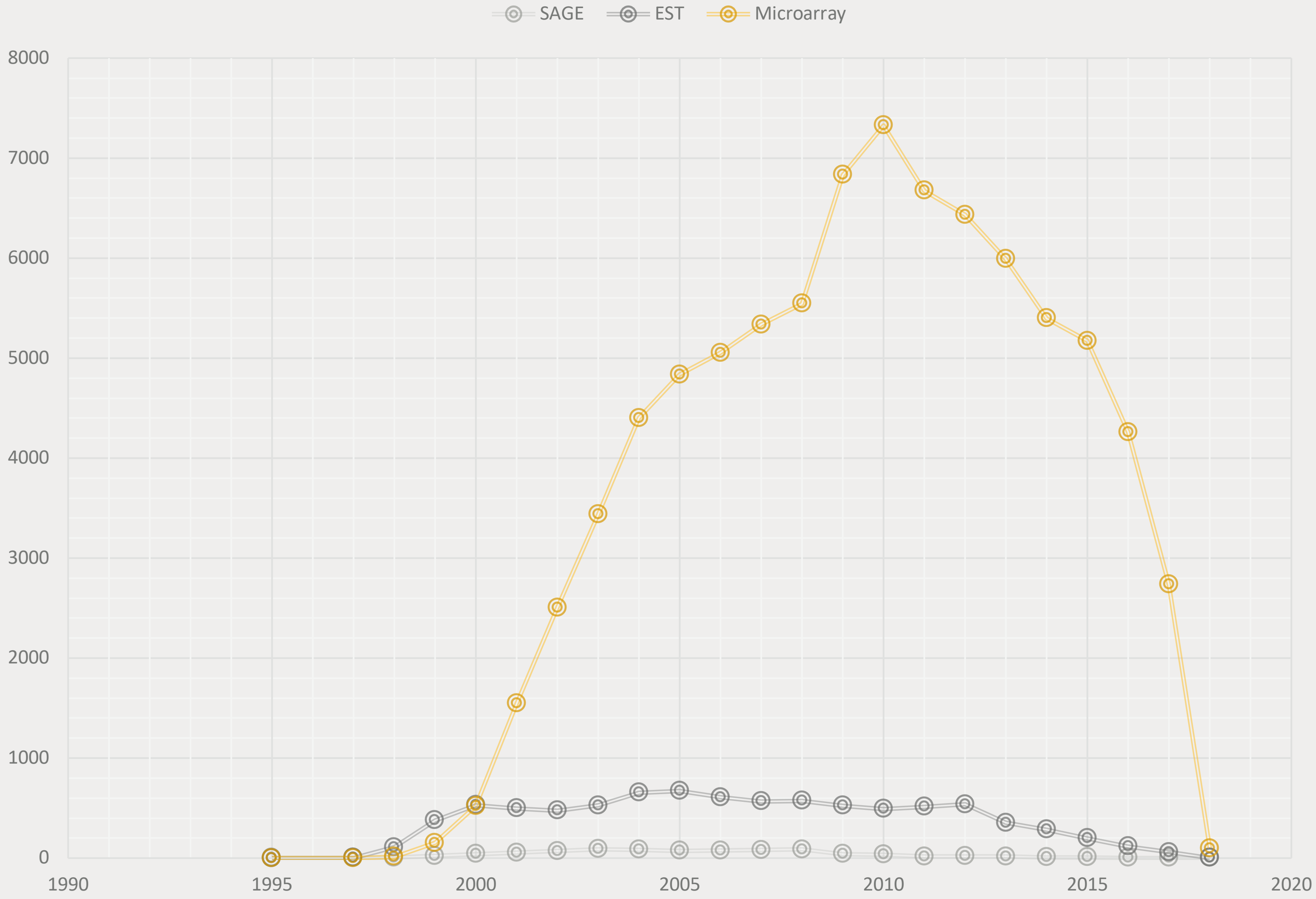
- ~ 1995 Series Analysis of
  Gene Expression (SAGE) (
  12 nucleotides)

- ~ **1999 Microarray**

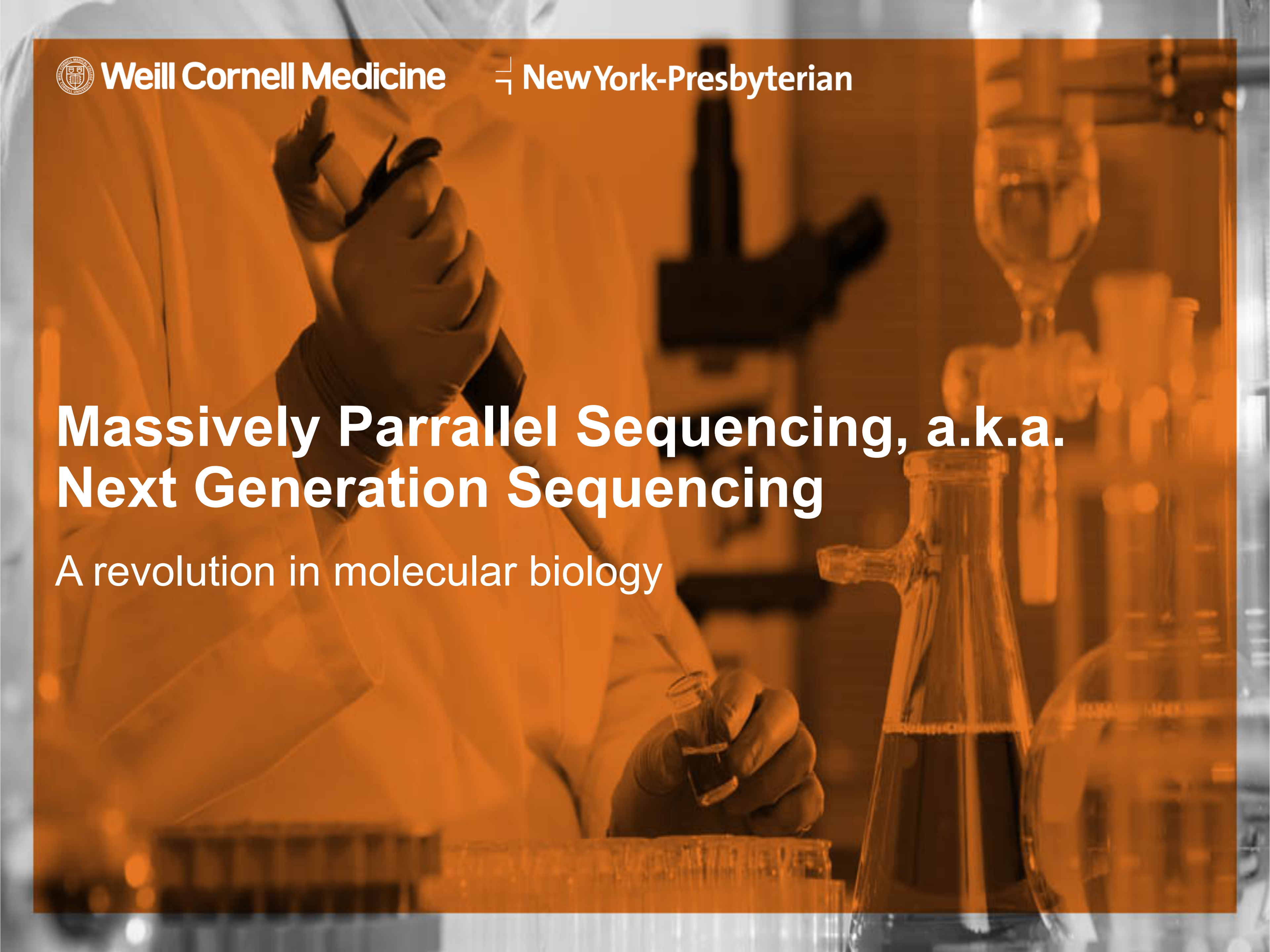


Science **286**, 531 (1999);

Number or PubMed articles



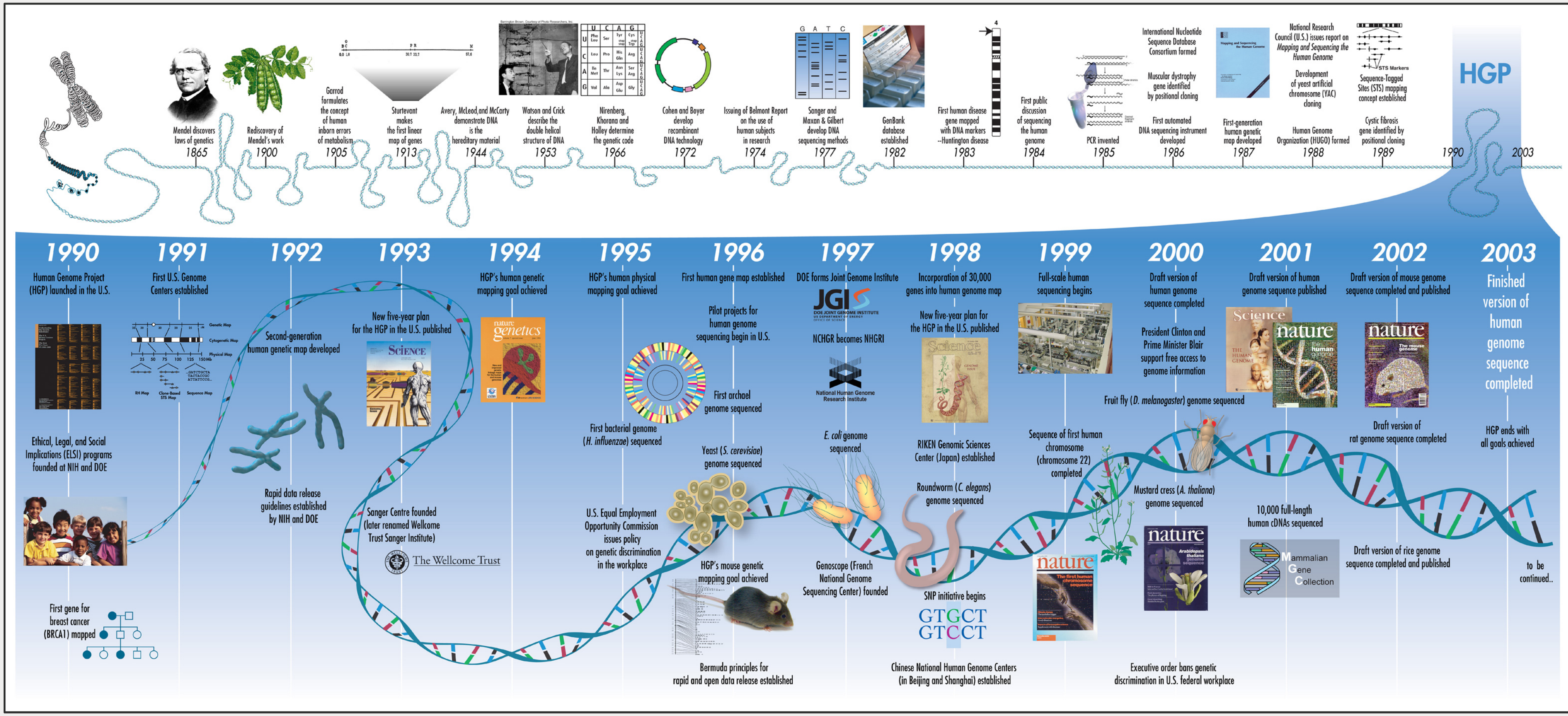
Weill Cornell Medicine NewYork-Presbyterian



Massively Parrallel Sequencing, a.k.a. Next Generation Sequencing

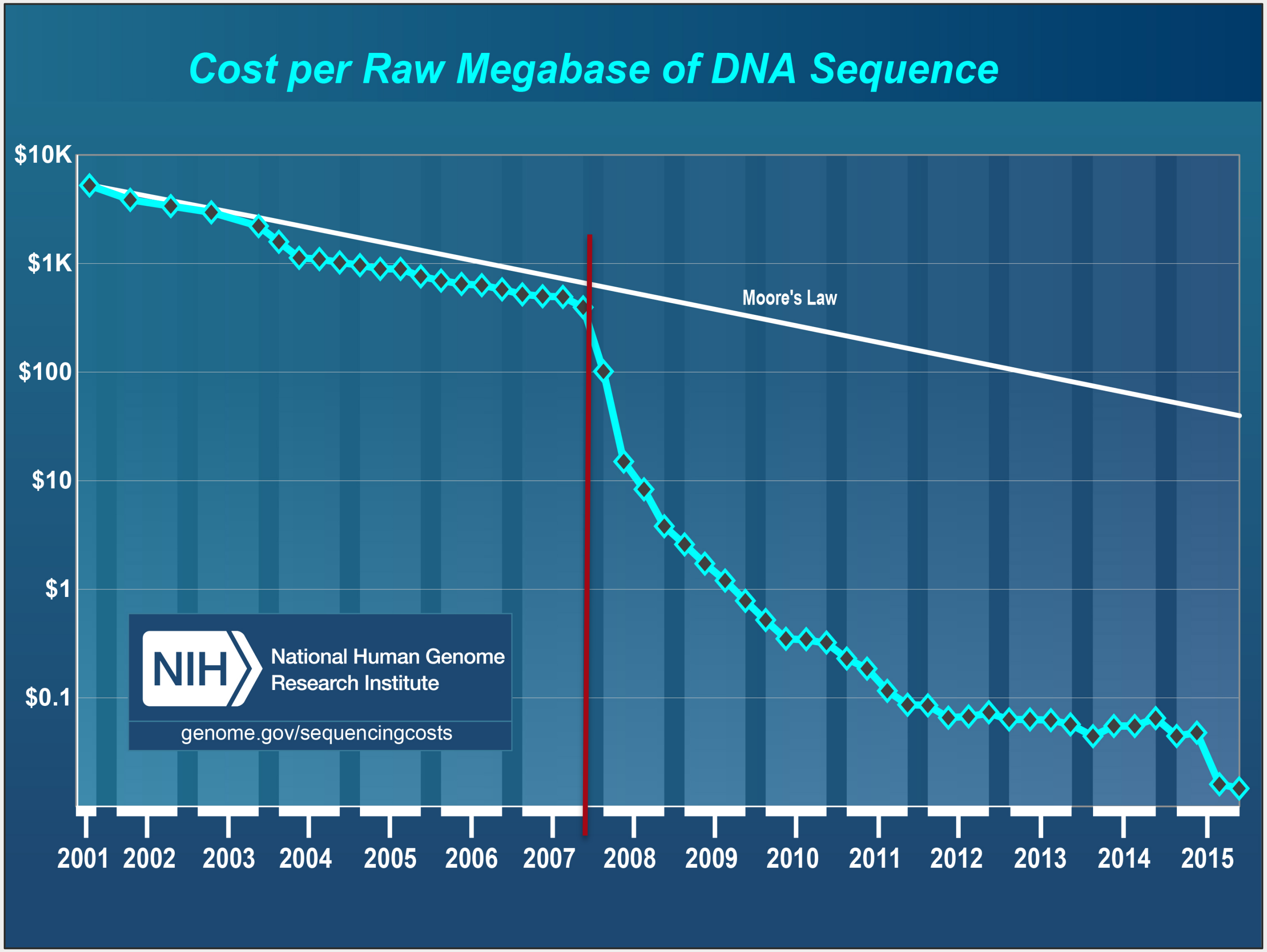
A revolution in molecular biology

The human genome reference sequence is completed in 2003


Mendel discovers laws of genetics
1865
Rediscovery of Mendel's work
1900
Garrod formulates the concept of human inborn errors of metabolism
1905
Sturtevant makes the first linear map of genes
1913
Avery, McLeod,and McCarty demonstrate DNA is the hereditary material
1944
Barrington Brown, Courtesy of Photo Researchers, Inc.
Watson and Crick describe the double helical structure of DNA
1953
Nirenberg, Khorana and Holley determine the genetic code
1966
Cohen and Boyer develop recombinant DNA technology
1972
Issuing of Belmont Report on the use of human subjects in research
1974
G A T C
Sanger and Maxan & Gilbert develop DNA sequencing methods
1977
GenBank database established
1982
First human disease gene mapped with DNA markers --Huntington disease
1983
First public discussion of sequencing the human genome
1984
PCR invented
1985
International Nucleotide Sequence Database Consortium formed
Muscular dystrophy gene identified by positional cloning
First automated DNA sequencing instrument developed
1986
First-generation human genetic map developed
1987
National Research Council (U.S.) issues report on Mapping and Sequencing the Human Genome
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed
1988
STS Markers
Sequence-Tagged Sites (STS) mapping concept established
Cystic fibrosis gene identified by positional cloning
1989
HGP
1990
2003
1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
25 50 75 100 125 150Mb
RH Map
Clone-Based STS Map
Sequence Map
...GATCTGCTA TACTACCGC ATTATTCCG...
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
SCIENCE
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
DOE forms Joint Genome Institute
JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
The first human chromosome sequence
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Arabidopsis thaliana genome sequence
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
THE HUMAN GENOME
nature
the human genome
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
The mouse genome
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

NIH National Human Genome Research Institute
Advancing human health through genomics research

Cost of sequencing decreased faster than Moore's Law



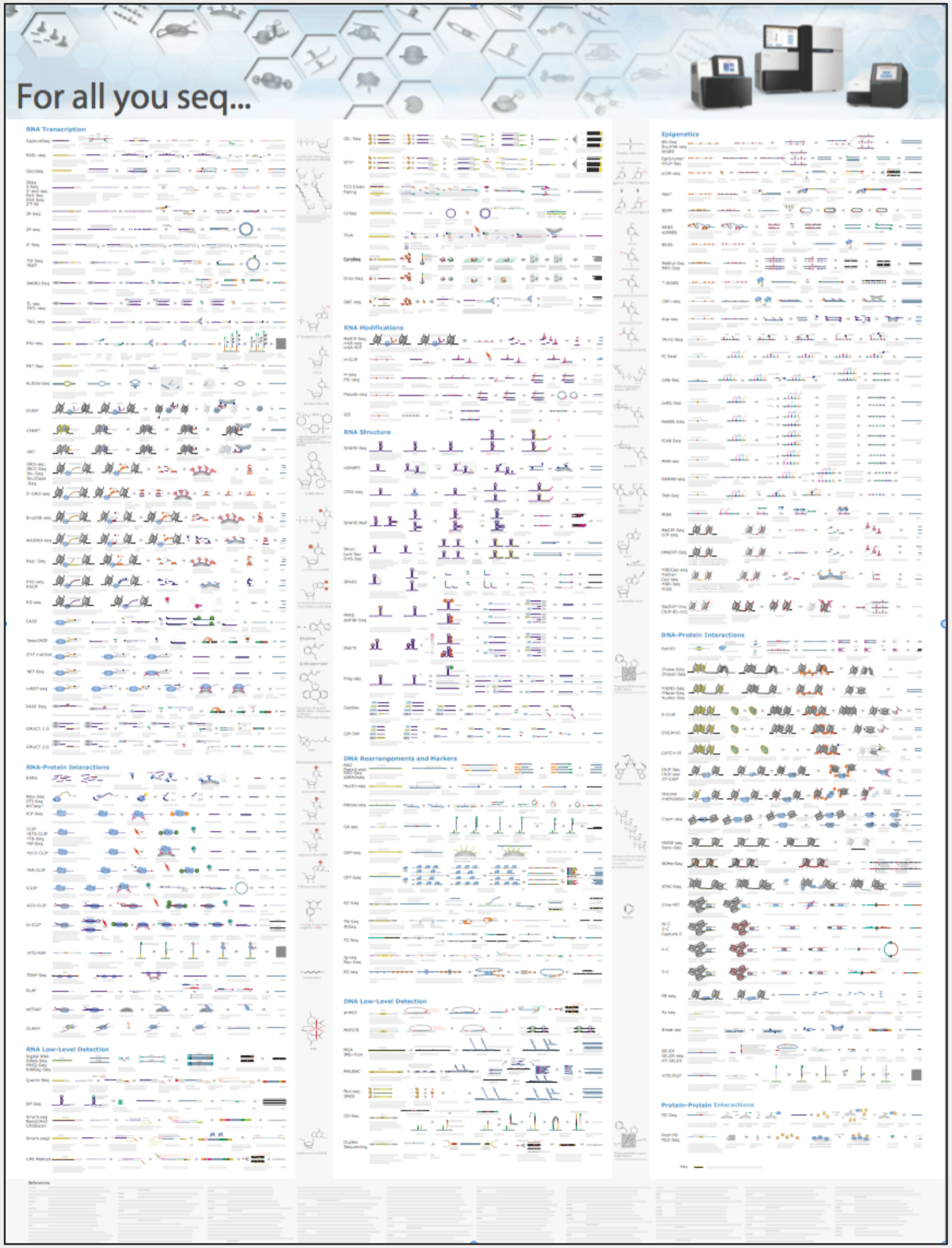
Weill Cornell Medicine NewYork-Presbyterian

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at:www.genome.gov/sequencingcosts. Accessed 3.28.2017

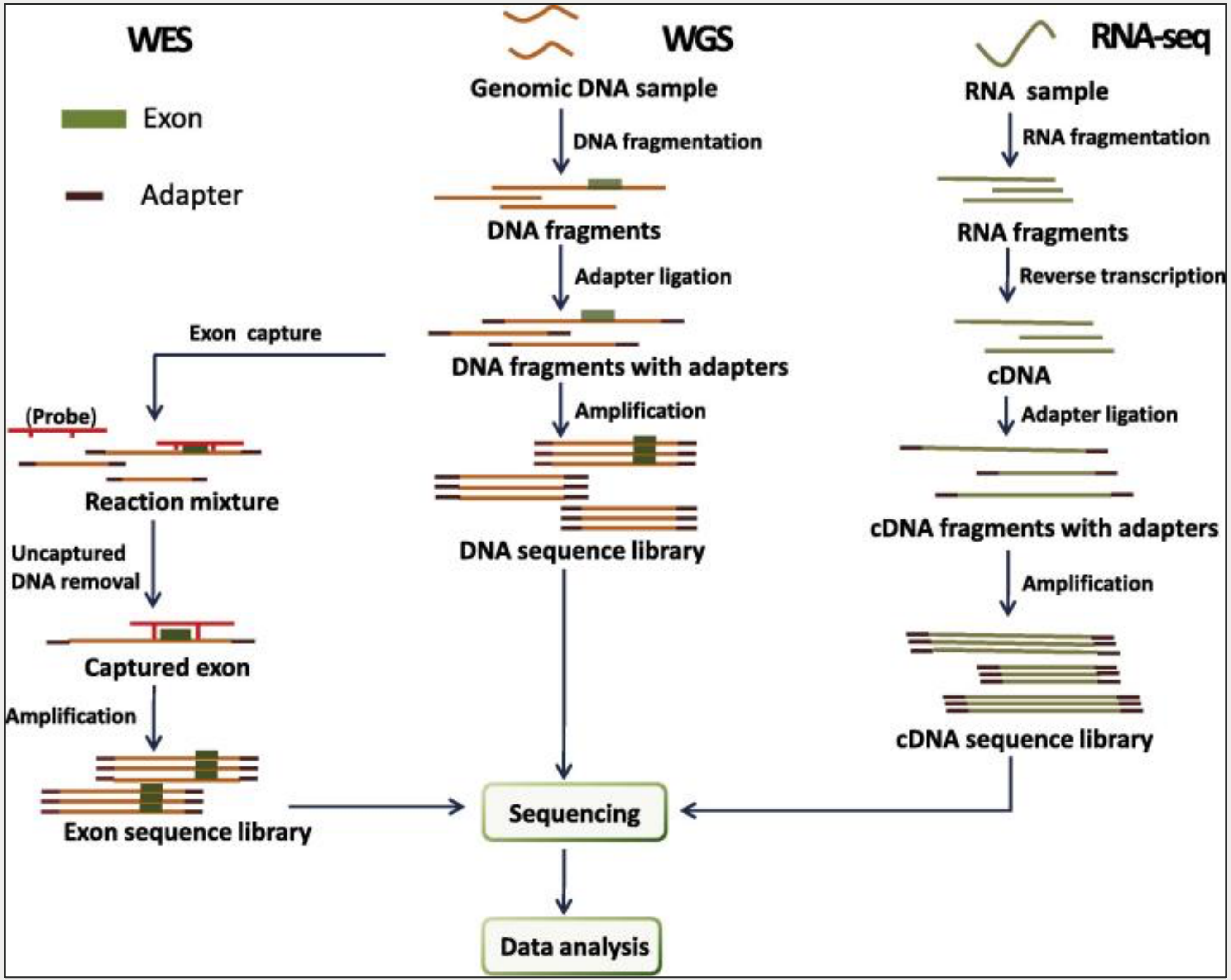
NGS allows the rapid sequencing of millions of “short” DNA/cDNA fragments

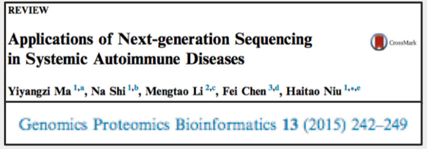
Many applications of NGS have been developed

DNA/RNA sequencing are the most common applications of NGS



Common NGS approaches




REVIEW

Applications of Next-generation Sequencing in Systemic Autoimmune Diseases

Yiyangzi Ma [1,a], Na Shi [1,b], Mengtao Li [2,c], Fei Chen [3,d], Haitao Niu [1,*,e]

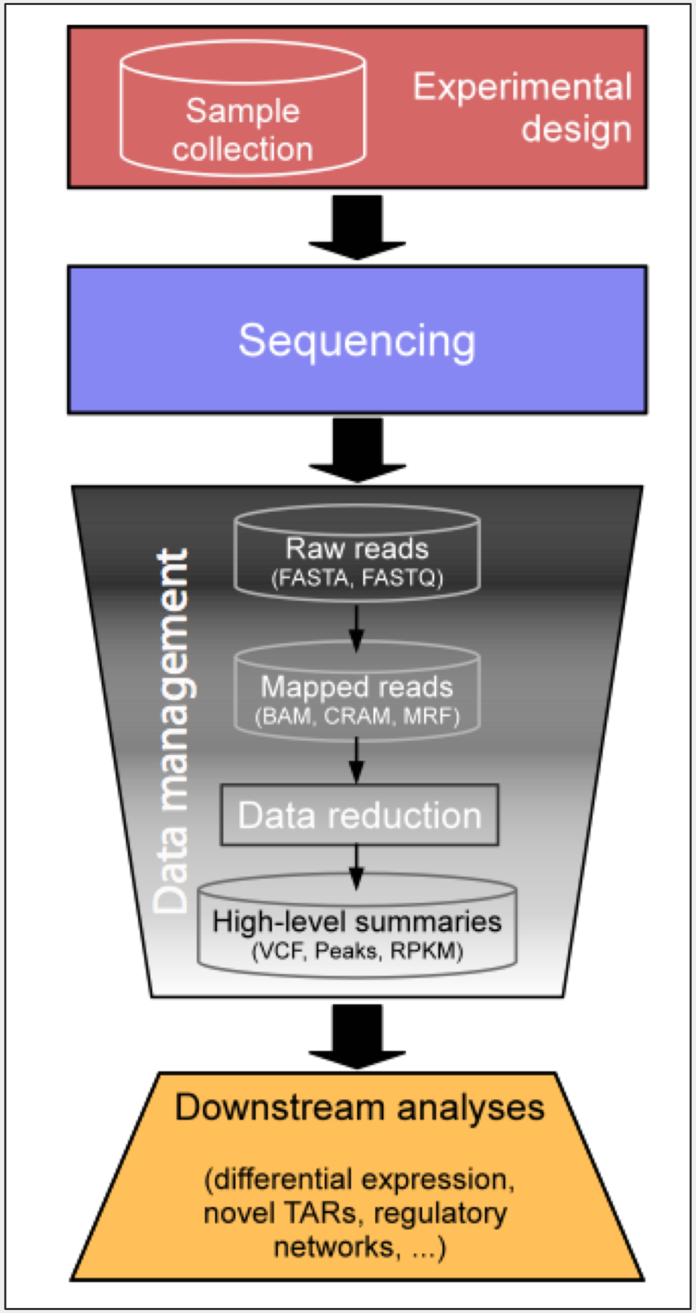
Genomics Proteomics Bioinformatics 13 (2015) 242–249

RNA-Seq Experiment

Data management:

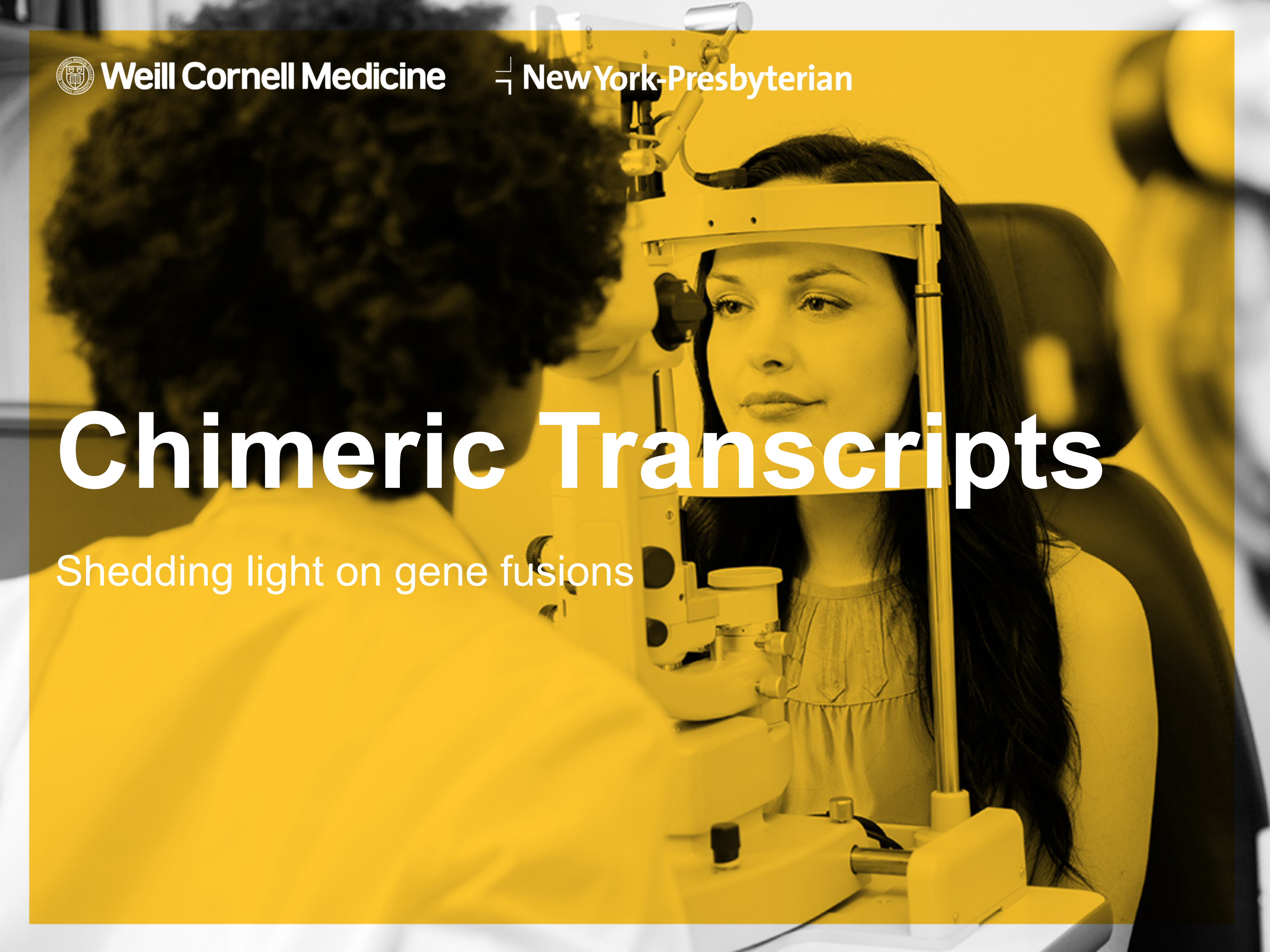
Mapping the reads
Creating summaries

Downstream analysis: *the interesting stuff*
Differential expression, chimeric transcripts, novel transcribed regions, etc.



Weill Cornell Medicine

NewYork-Presbyterian



Chimeric Transcripts

Shedding light on gene fusions

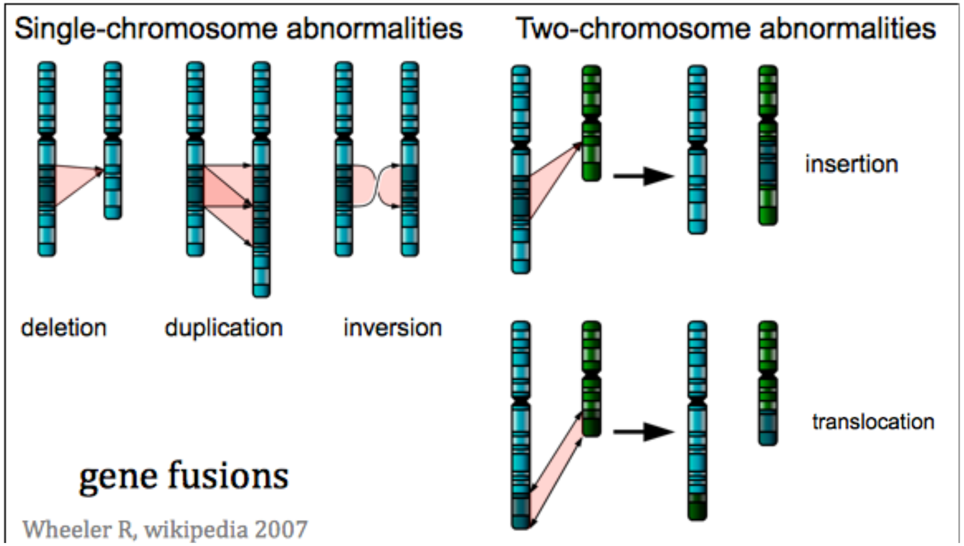
What are chimeric transcripts?

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*



gene fusions

Wheeler R, wikipedia 2007

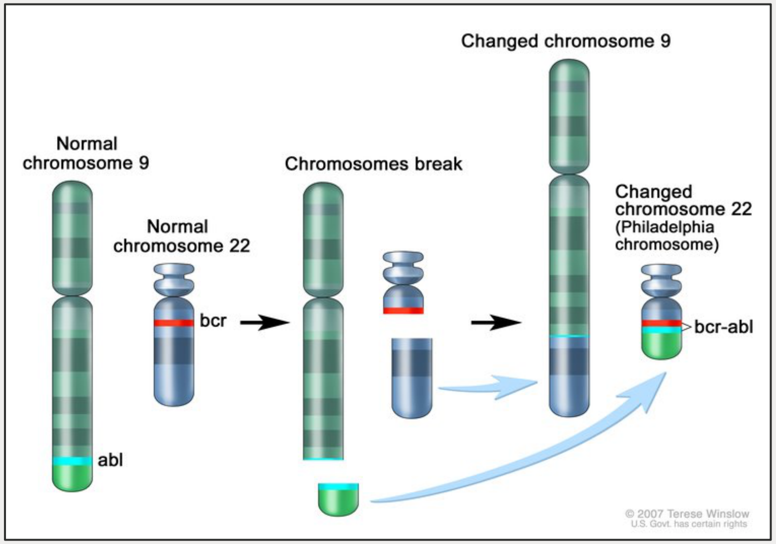
Why are they (gene fusions) important?

Fusion genes are often *oncogenes*

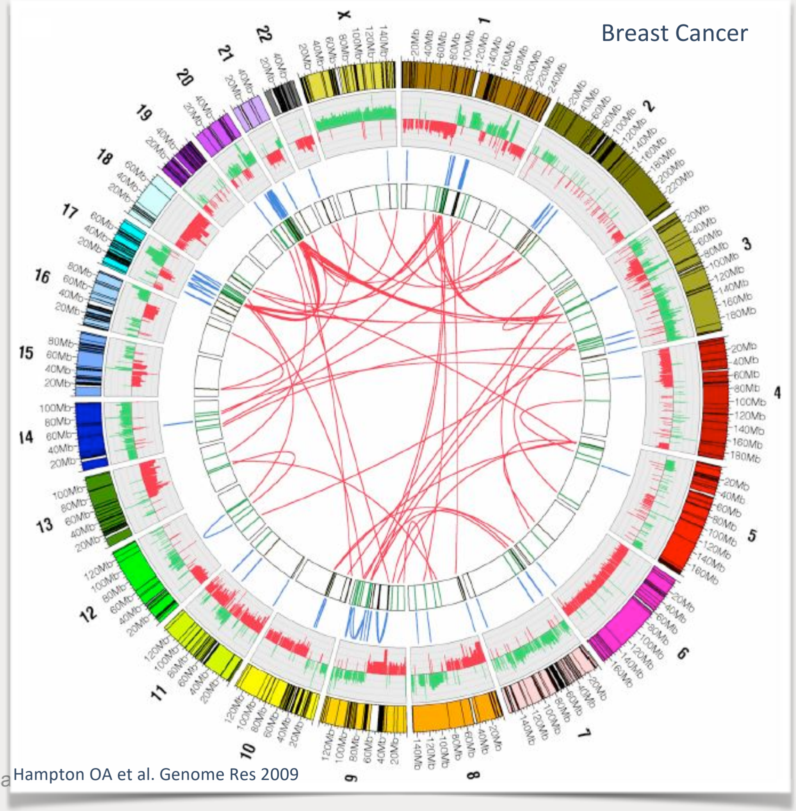
- Ex: BCR-ABL1 (Philadelphia chromosome) in Chronic myelogenous leukemia (CML) and Acute Lymphoblastic leukemia (ALL) t(9;22)(q34;q11)

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

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



© 2007 Terese Winslow U.S. Govt. has certain rights



Hampton OA et al. Genome Res 2009

Why are they (trans-splicing events) important?

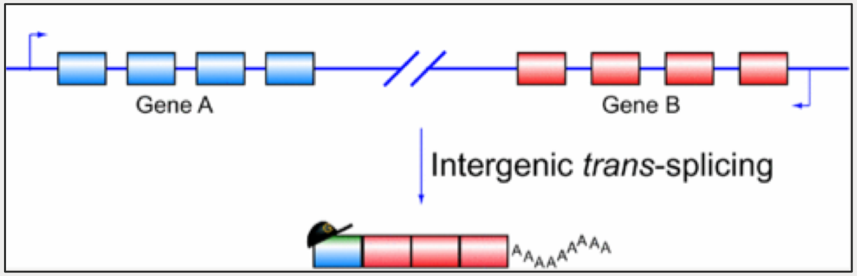
Trans(cis)-splicing was initially found in lower eukariotes, such as trypanosomes and worms

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

Recently, they were found in mammalian cells:

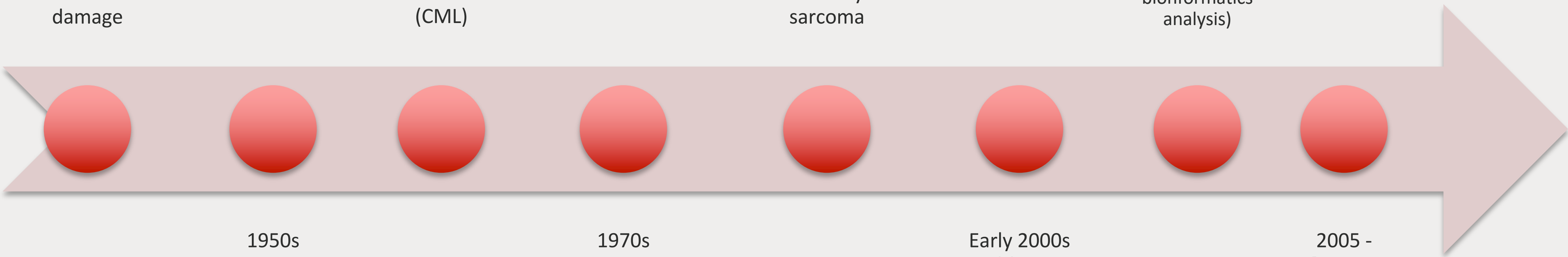
- **JAZF1-SUZ12 in endometrial stroma cells (Li et al. Science 2008)**
- **SLC45A3-ELK4 in prostate tissues (Rickman et al. Cancer Res 2009)**

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

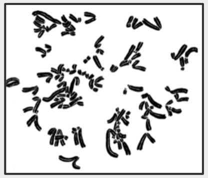


1914

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 rhabdomyosarcoma

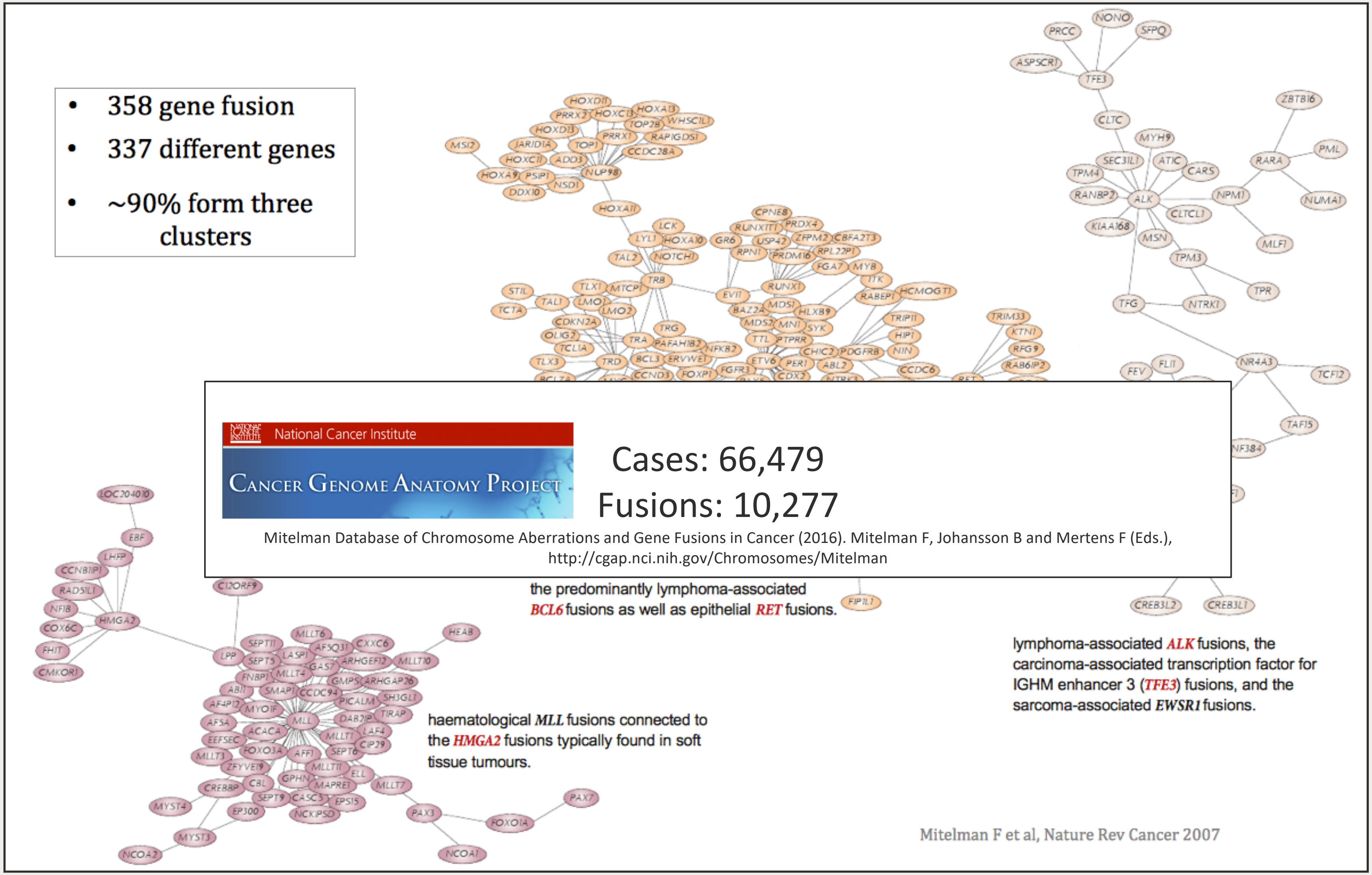
Early 2000s
Most rearrangements were discovered mainly in hematological and mesenchymal neoplasms compared to epithelial

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

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?



- 358 gene fusion
- 337 different genes
- ~90% form three clusters

National Cancer Institute

Cancer Genome Anatomy Project

Cases: 66,479

Fusions: 10,277

Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2016). Mitelman F, Johansson B and Mertens F (Eds.), http://cgap.nci.nih.gov/Chromosomes/Mitelman

the predominantly lymphoma-associated *BCL6* fusions as well as epithelial *RET* fusions.

lymphoma-associated *ALK* fusions, the carcinoma-associated transcription factor for IGHM enhancer 3 (*TFE3*) fusions, and the sarcoma-associated *EWSR1* fusions.

haematological *MLL* fusions connected to the *HMGA2* fusions typically found in soft tissue tumours.

Mitelman F et al, Nature Rev Cancer 2007

Gene fusions are important for clinical treatment…

MAY 28, 2001

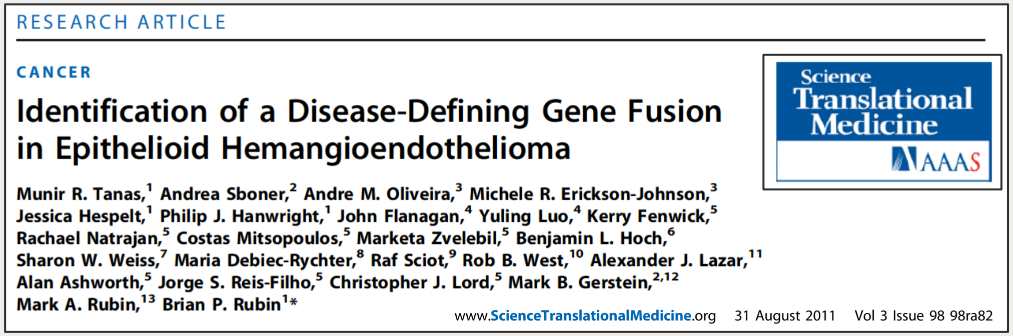
TIME

THERE IS NEW AMMUNITION IN THE WAR AGAINST CANCER. THESE ARE THE BULLETS.

Revolutionary new pills like GLIVEC combat cancer by targeting only the diseased cells. Is this the breakthrough we've been waiting for?

www.timeeurope.com AOL Keyword: TIME

… and diagnostic/prognostic purposes



RESEARCH ARTICLE

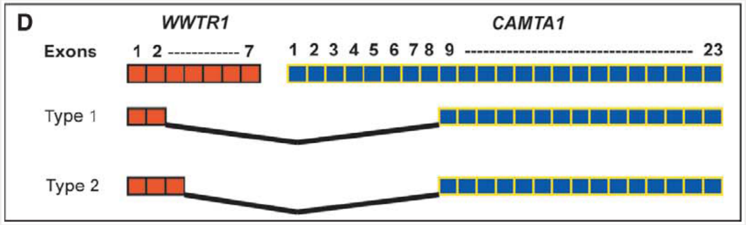
CANCER

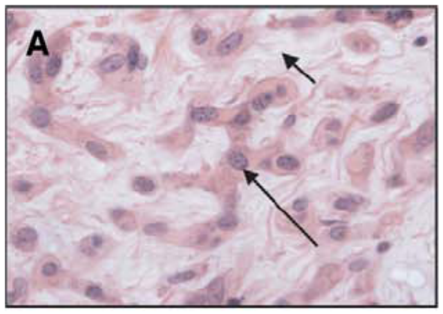
Identification of a Disease-Defining Gene Fusion in Epithelioid Hemangioendothelioma

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]*

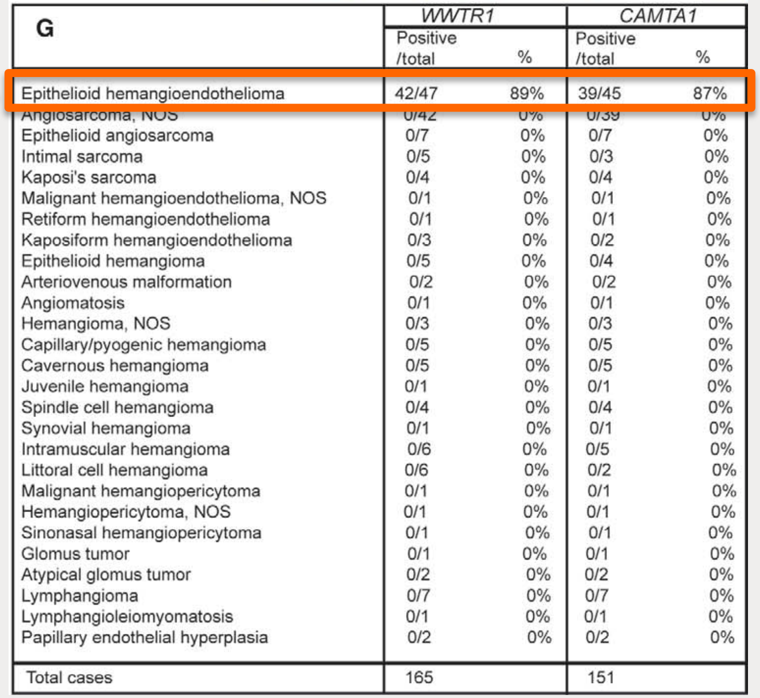
Science Translational Medicine
AAAS

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

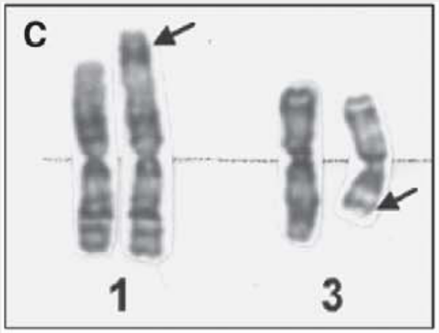


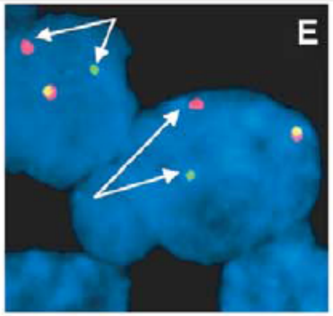


Exclusively present in *epitheliod hemangioendothelioma*



G	*WWTR1*		*CAMTA1*	
	Positive /total	%	Positive /total	%
Epithelioid hemangioendothelioma	42/47	89%	39/45	87%
Angiosarcoma, NOS	0/42	0%	0/39	0%
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	

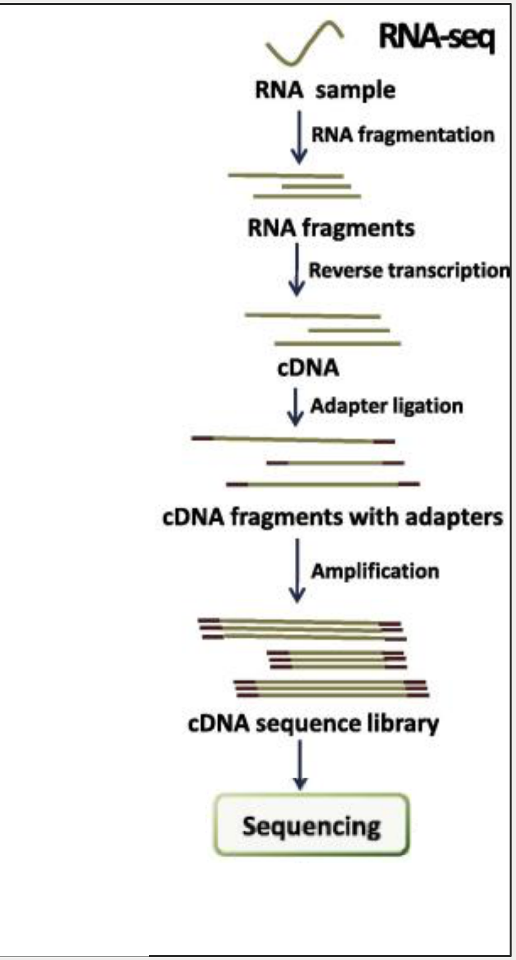




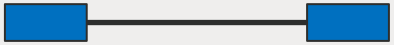
Weill Cornell Medicine NewYork-Presbyterian

Fusion Detection from paired-end RNA-Seq

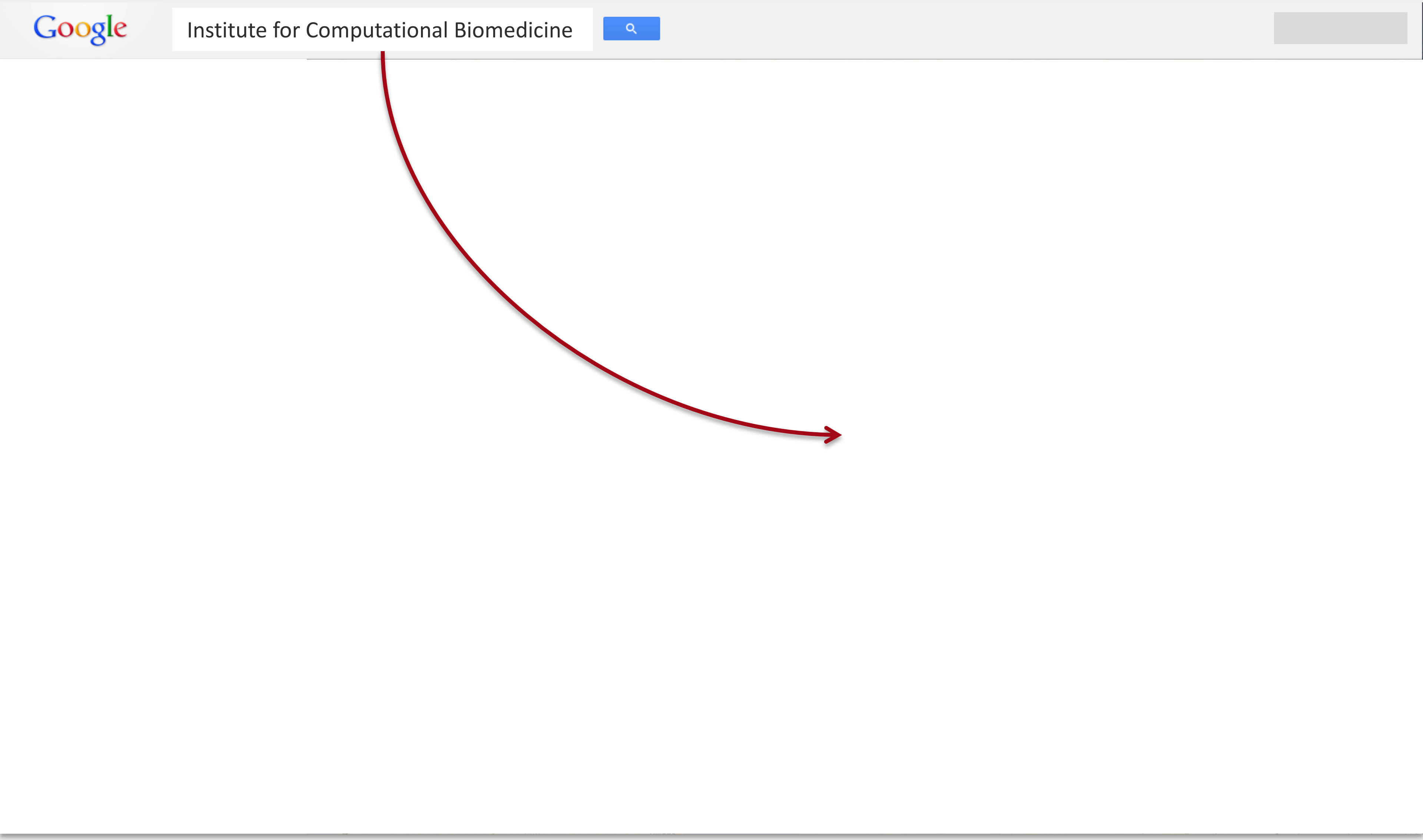
How to identify fusion transcripts from paired-end RNA-seq?



Paired-end sequencing means that we know the sequence of the two ends of a fragment

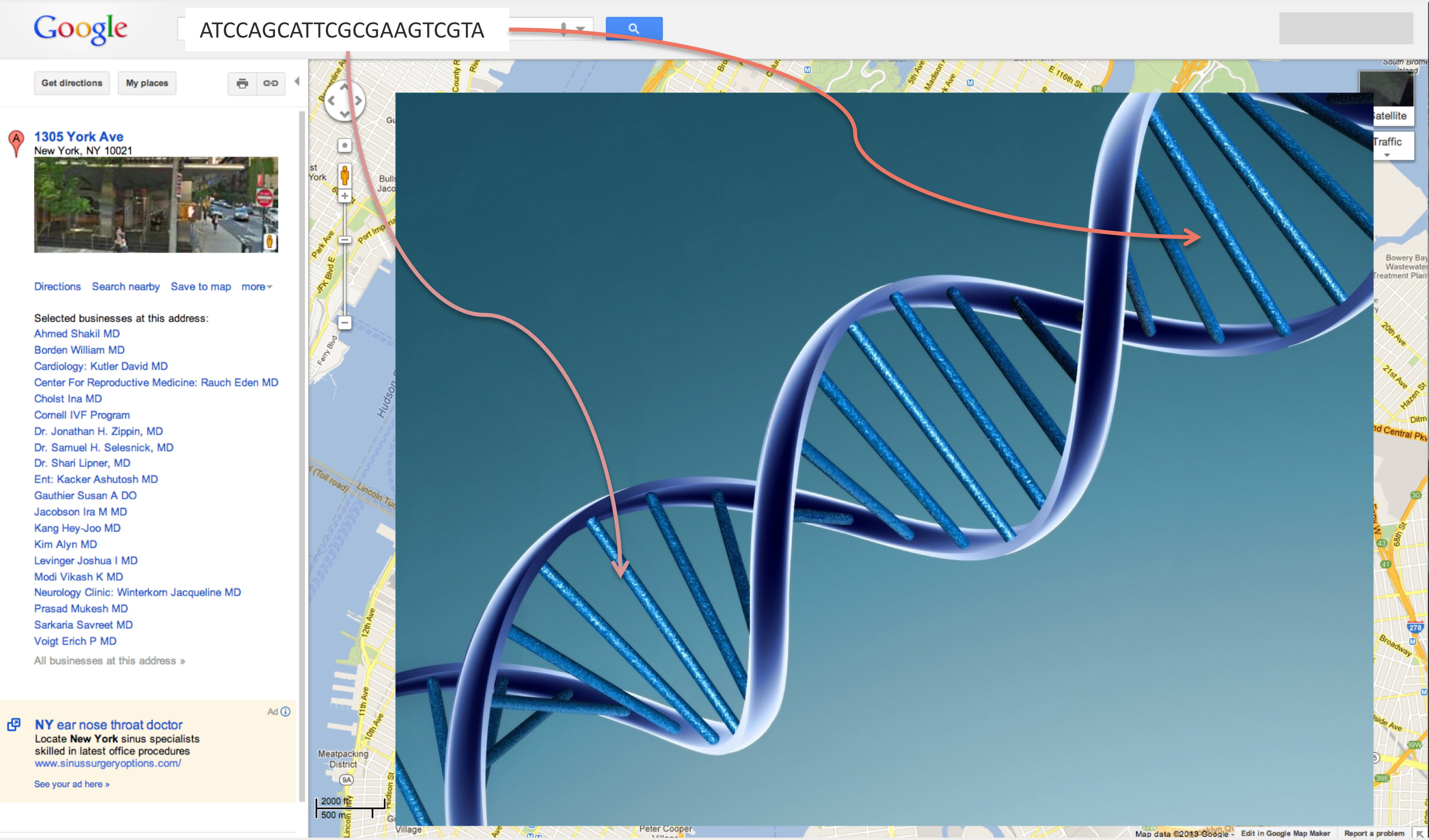


Mapping



Institute for Computational Biomedicine

Mapping



ATCCAGCATTCGCGAAGTCGTA

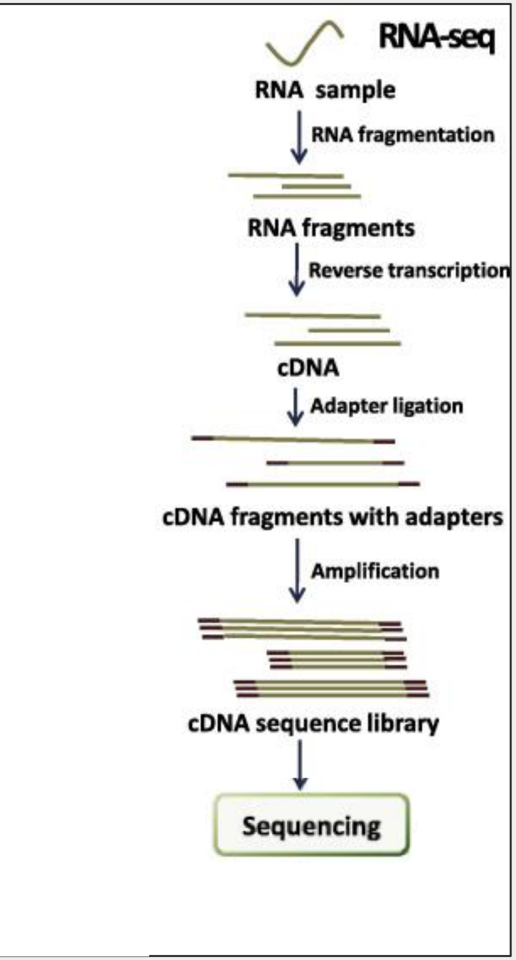
1305 York Ave
New York, NY 10021

Selected businesses at this address:

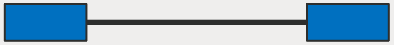
- Ahmed Shakil MD
- Borden William MD
- Cardiology: Kutler David MD
- Center For Reproductive Medicine: Rauch Eden MD
- Cholst Ina MD
- Cornell IVF Program
- Dr. Jonathan H. Zippin, MD
- Dr. Samuel H. Selesnick, MD
- Dr. Shari Lipner, MD
- Ent: Kacker Ashutosh MD
- Gauthier Susan A DO
- Jacobson Ira M MD
- Kang Hey-Joo MD
- Kim Alyn MD
- Levinger Joshua I MD
- Modi Vikash K MD
- Neurology Clinic: Winterkorn Jacqueline MD
- Prasad Mukesh MD
- Sarkaria Savreet MD
- Voigt Erich P MD

All businesses at this address »

How to identify fusion transcripts from paired-end RNA-seq?

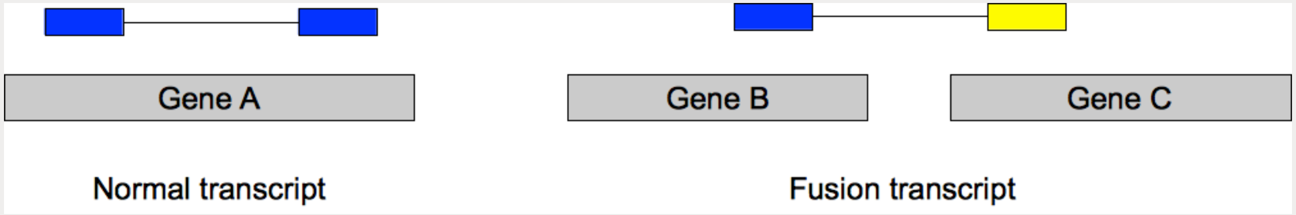


Paired-end sequencing means that we know the sequence of the two ends of a fragment

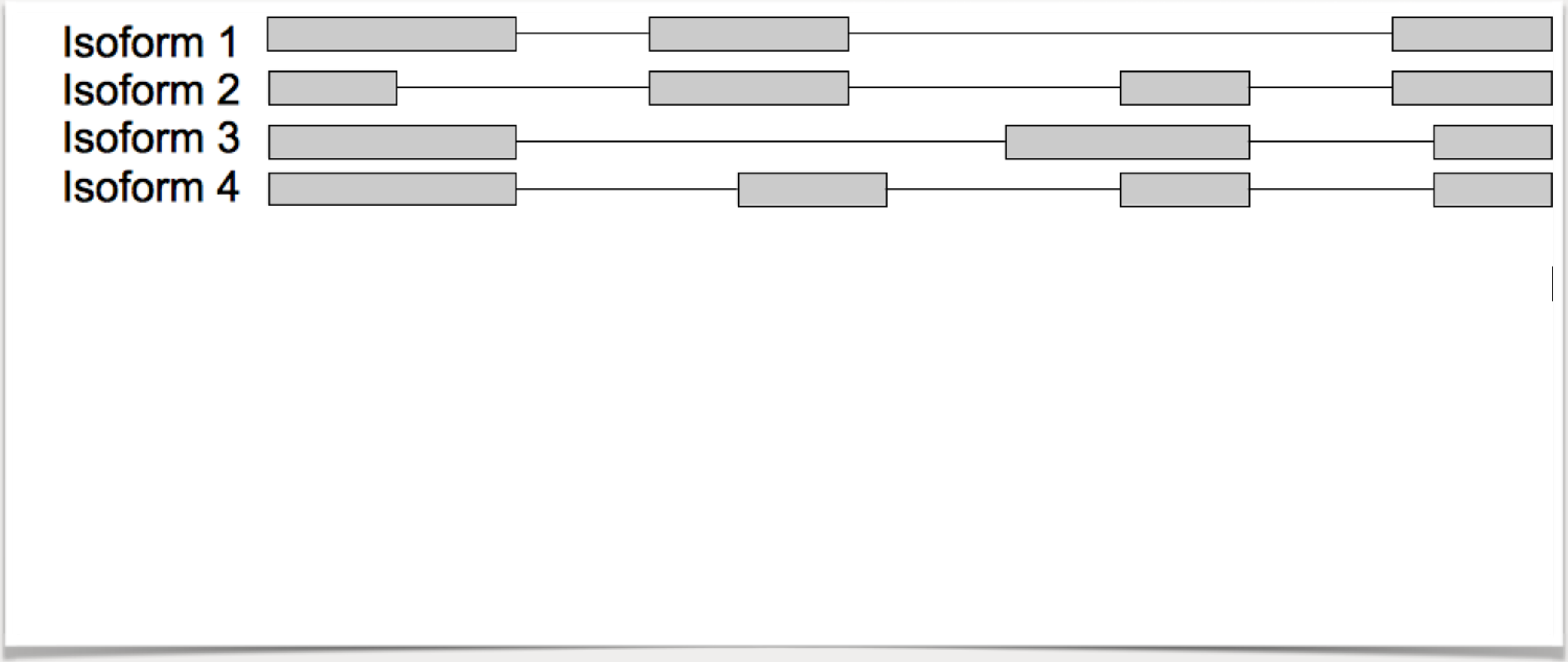


Straightforward:

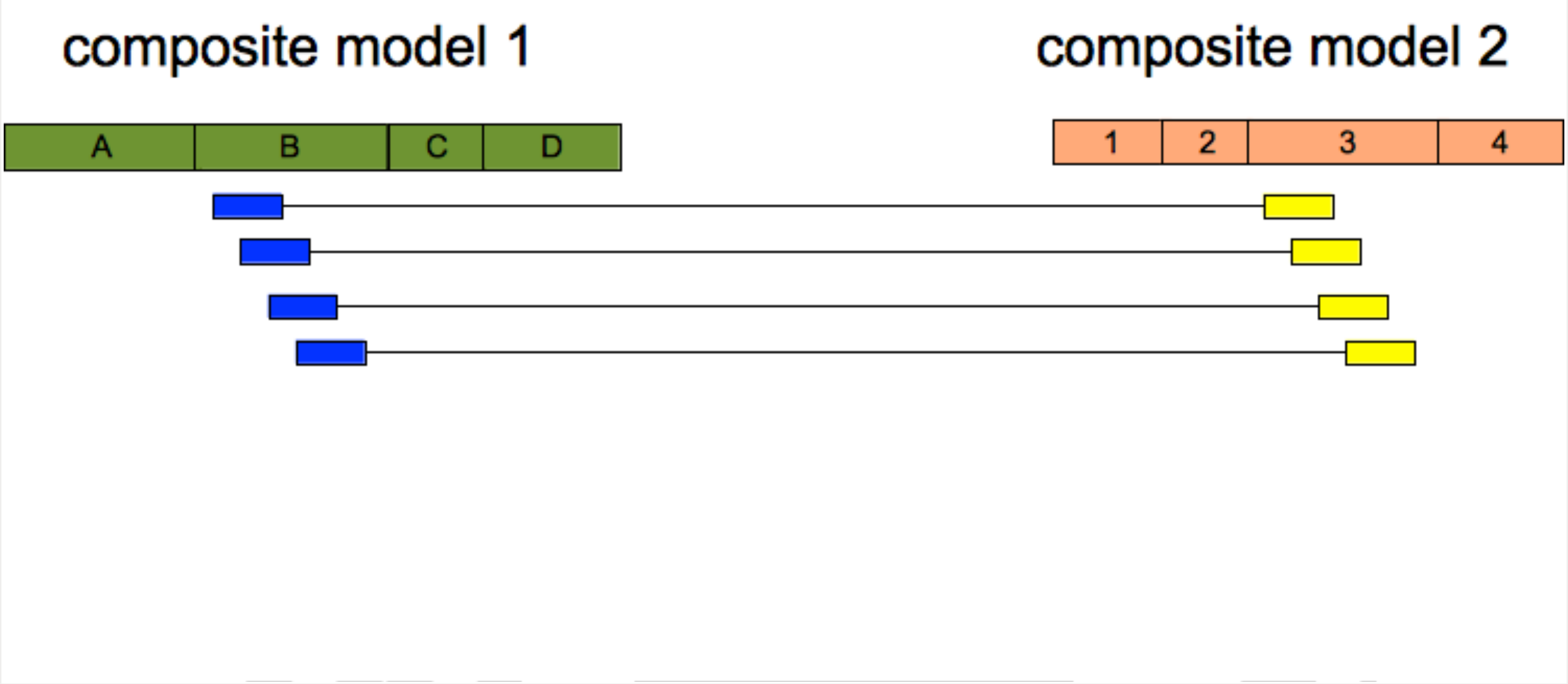
If the two ends map to different genes, then we have a potential fusion transcript



What about different isoforms?



Composite model



- Each PE read can be assigned to one “gene”
- *Potential Fusion Transcripts*: if pair belongs to different genes

Not an ideal word: sources of errors

Mis-alignments

- **Base caller error**
- **SNPs**
- **RNA editing**
- **Sequence similarity (paralogs, pseudogenes)**

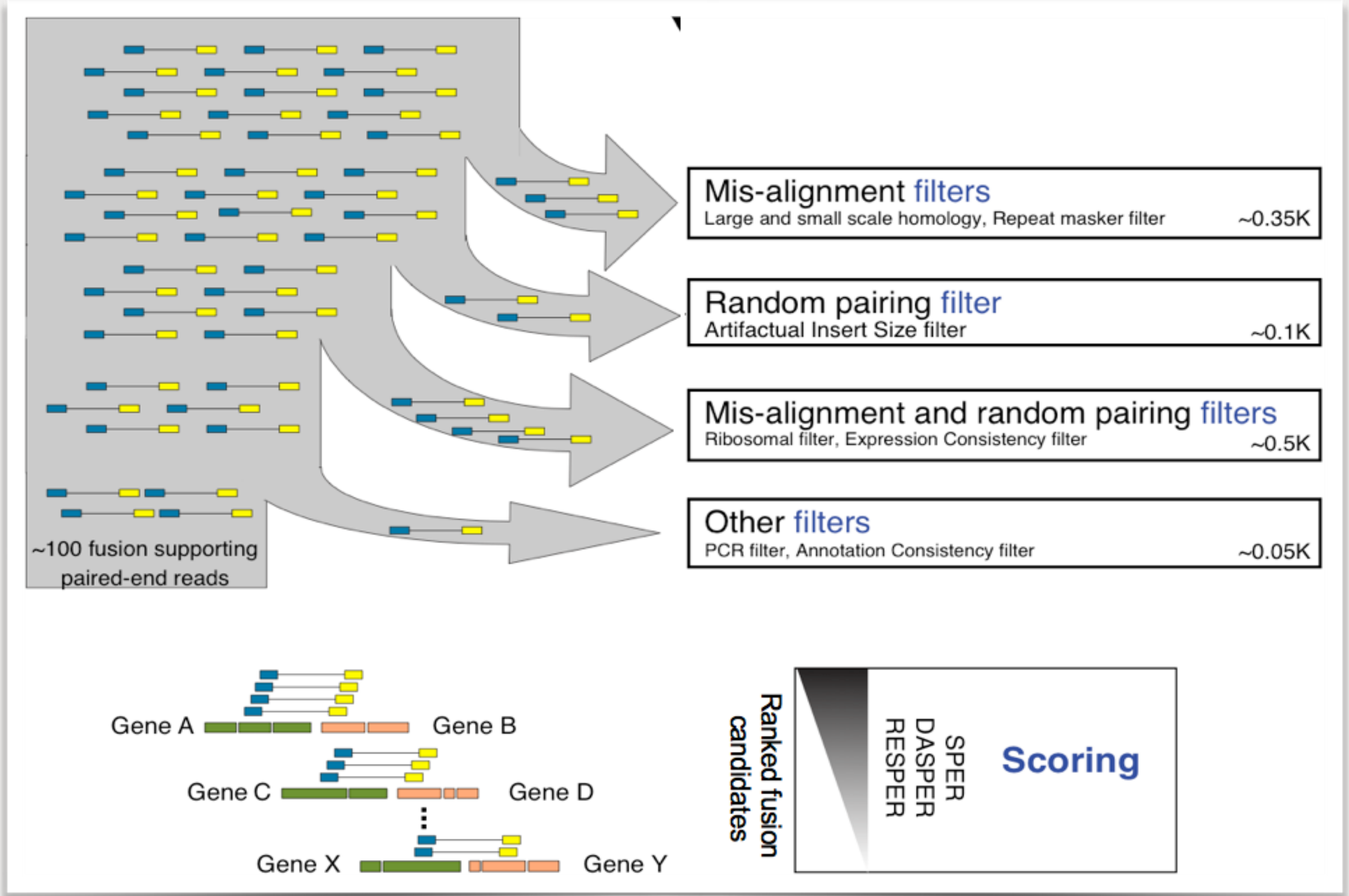
Random pairing of transcript fragments

- **Library preparation**

Combination of mis-alignment and random pairing

PCR amplification, gene annotation inconsistencies/incompleteness

Filtration Cascade Module



Mis-alignment filters
Large and small scale homology, Repeat masker filter ~0.35K

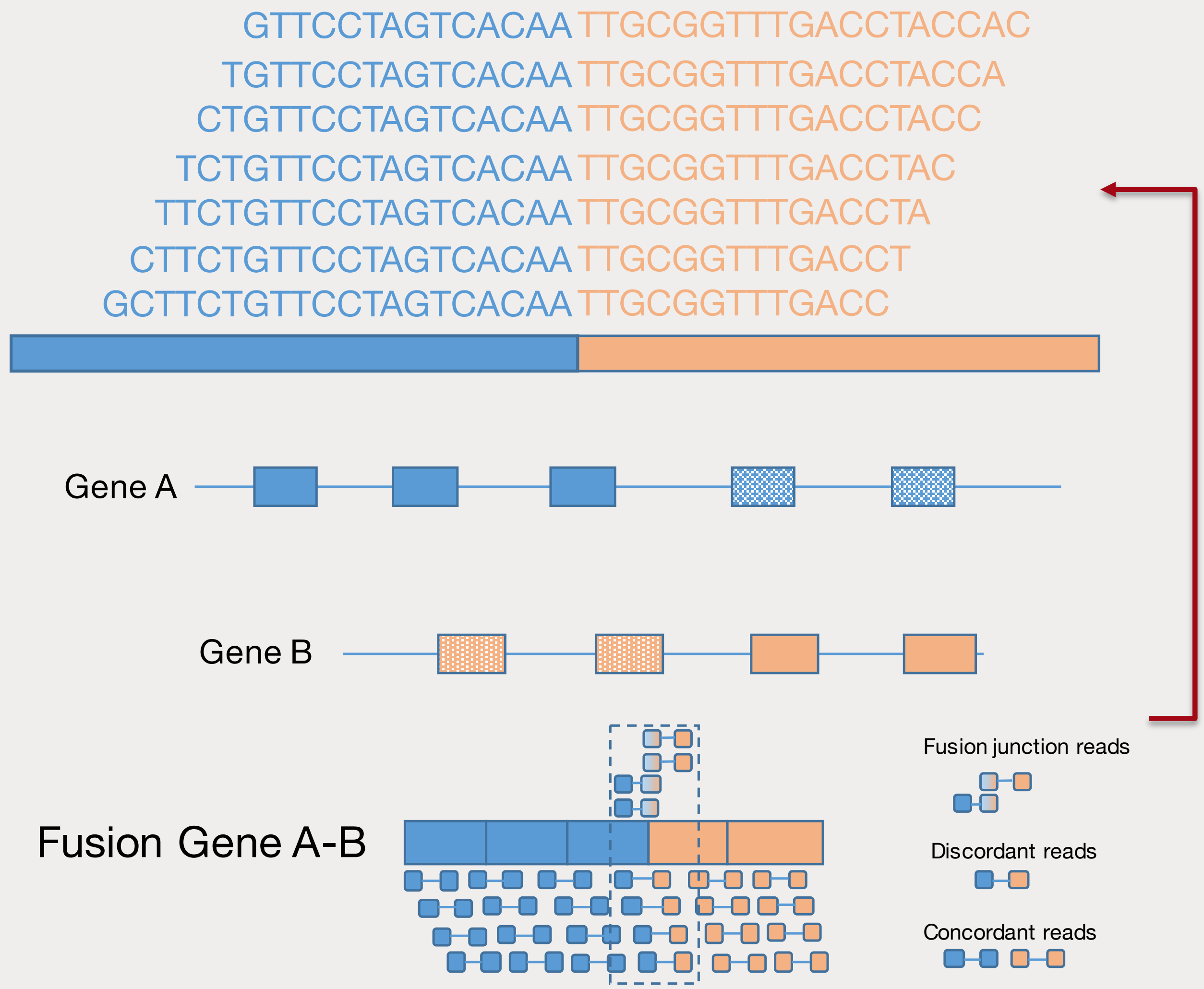
Random pairing filter
Artifactual Insert Size filter ~0.1K

Mis-alignment and random pairing filters
Ribosomal filter, Expression Consistency filter ~0.5K

Other filters
PCR filter, Annotation Consistency filter ~0.05K

Scoring
SPER
DASPER
RESPER

Augmenting the support for fusion: fusion junction reads

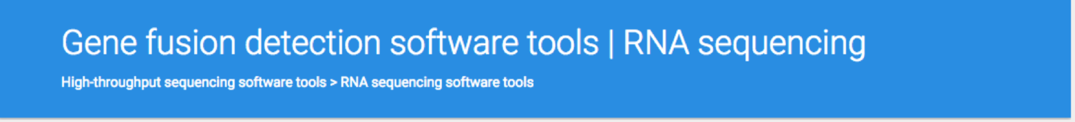


```
     GTTCCTAGTCACAA TTGCGGTTTGACCTACCAC
    TGTTCCTAGTCACAA TTGCGGTTTGACCTACCA
   CTGTTCCTAGTCACAA TTGCGGTTTGACCTACC
  TCTGTTCCTAGTCACAA TTGCGGTTTGACCTAC
 TTCTGTTCCTAGTCACAA TTGCGGTTTGACCTA
CTTCTGTTCCTAGTCACAA TTGCGGTTTGACCT
GCTTCTGTTCCTAGTCACAA TTGCGGTTTGACC
```

Weill Cornell Medicine   NewYork-Presbyterian

Tools for detecting fusion transcripts

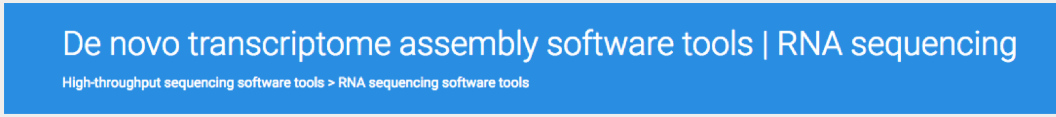
From sequencing data



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

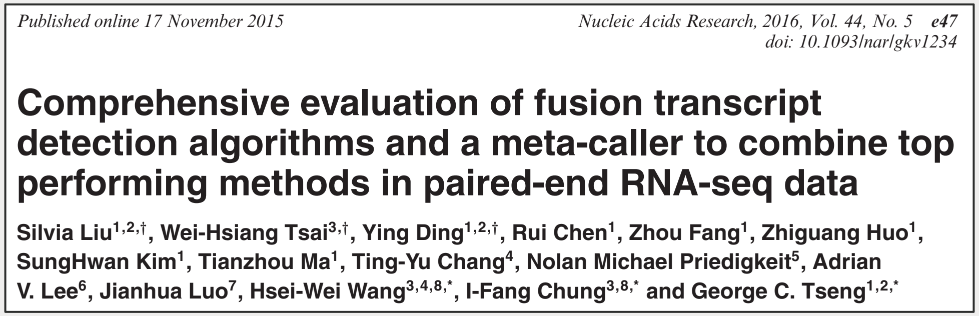
BreakTrans
Comrad
nFuse

Gene fusion annotation

Chimera
Pegasus

Transcript Assembly

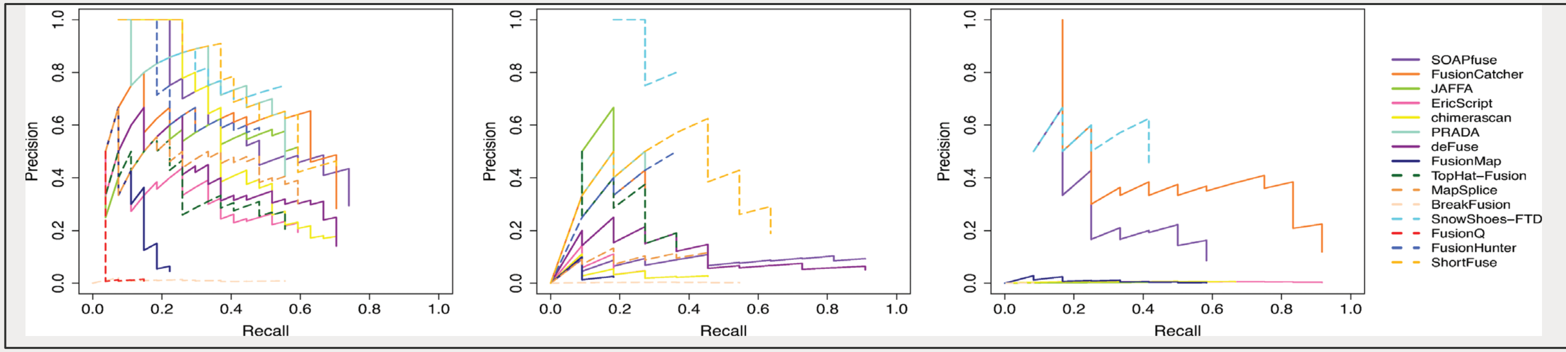
CuffLinks
Scripture
Trinity
Trans-Abyss



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,*]



Summary and Future directions

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

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