National Cancer Institute

Precision Oncology and Next Generation Sequencing

September 2014

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Overview

- Challenges in Cancer Informatics
- Disruptive Technologies
- NCI Genomics Data Commons
- NCI Cloud Pilots
- NGS and adaptive trial design
- Building a national learning health system for cancer clinical genomics

- The NCI has several large multi-technology projects underway. Two of them, TCGA and TARGET, are well established resources with rich information on tissue imaging, pathology imaging, radiological imaging, somatic variation, genome structural rearrangements, miRNA sequencing, mRNA sequencing, and protein expression measurements. The talk will highlight some of the scientific findings from TCGA and provide information about the creation of the Genomics Data Commons by the NCI Center for Cancer Genomics, as well as the role of the NCI Cloud Pilots in providing broad community access to these very important information resources. A discussion of the new molecular trials, MPACT and MATCH, that incorporate biomarker evidence into the arm selection process for the assignment of treatment will be given. These resources and trials will provide additional evidence to enable the optimal treatment of cancer. Also included is a discussion of the informatics design and architectural decisions, information workflows and tools used by these projects.

Cancer relevance

- Challenges for Cancer Informatics
- Disruptive technologies
- Drivers for social benefit in cancer research

National Challenges in Cancer Informatics

- **Lowering barriers** to data access, analysis and modeling for cancer research
- Integration of data and learning from basic and clinical research with cancer care *that* ***enable prediction*** *and improved outcomes*

We need:

- **Open Science** (Open Access, Open Data, Open Source) and Data Liquidity for the cancer community
- **Interoperability** through CDEs and Case Report Forms mapped to standards
- **Semantic interoperability** by exposing existing knowledge through **proper integration of ontology**
- **Sustainable models** for informatics infrastructure, services, data

Where we are

Disruptive technologies

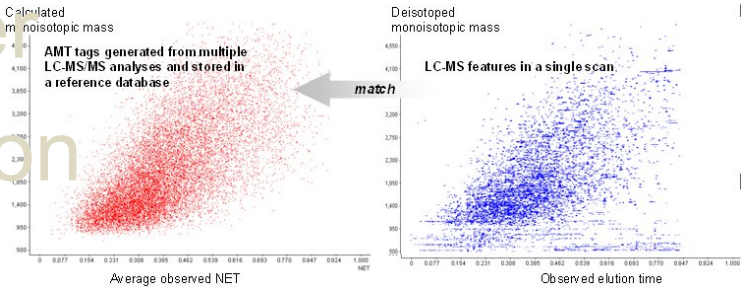
Getting social

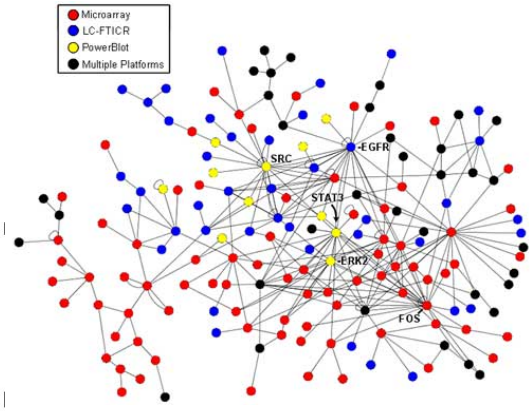
Open access to data

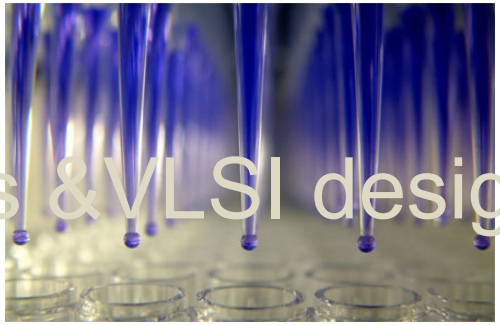
Disruptive Technologies

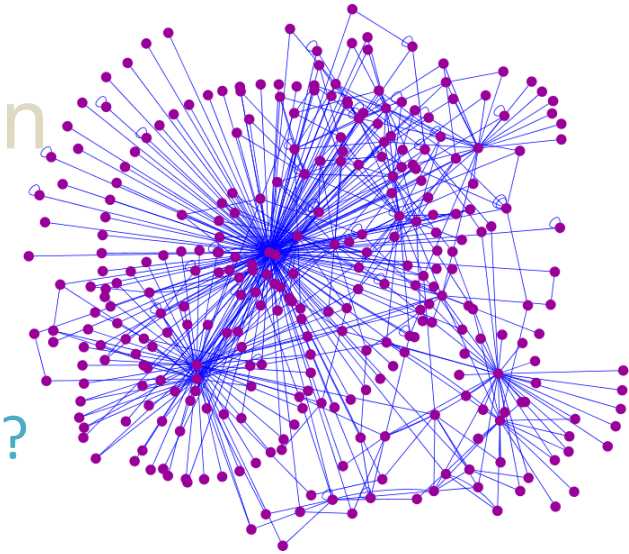
- Printing
- Steam power
- Transportation
- Electricity
- Antibiotics
- Semiconductors &VLSI design
- http
- High throughput biology

Systems view - end of reductionism?









Precision Oncology

- The era of precision medicine and precision oncology is *predicated* on the integration of research, care, and molecular medicine and the *availability of data* for modeling, risk analysis, and optimal care

How do we re-engineer translational research policies that will enable a true ***learning healthcare system****?*

Precision Oncology

- The era of precision medicine and precision oncology is *predicated* on the integration of research, care, and molecular medicine and the *availability of data* for modeling, risk analysis, and optimal care

How do we re-engineer translational research policies that will enable a true ***learning healthcare system****?*

Disruptive Technologies

- Printing
- Steam pov
- Transporta
- Electricity
- Antibiotics
- Semiconductors &VLSI d
- http
- High throughput biology
- Ubiquitous computing

World:
6.6B active mobile contracts
1.9B smart phone contracts
1.1B land lines
World population 7.1B

US:
345M active mobile contracts
287M smart phone contracts
US population 313M

Everyone is a data provider
Data immersion

What about social media?

- Social media may be one avenue for **modifying behaviors** that result in cancer
- Properly orchestrated, social media can have dramatic **impact on quality of life** for patients and survivors
- It can **reach into all segments** of our society, including underserved populations

Public Health

- These three modifiable factors - **infectious disease, smoking, and poor nutrition** and lack of exercise **contribute to at least 50%** of our current cancer burden. And the cost from loss of quality of life, pain and suffering is incalculable.

Some NCI NGS activities

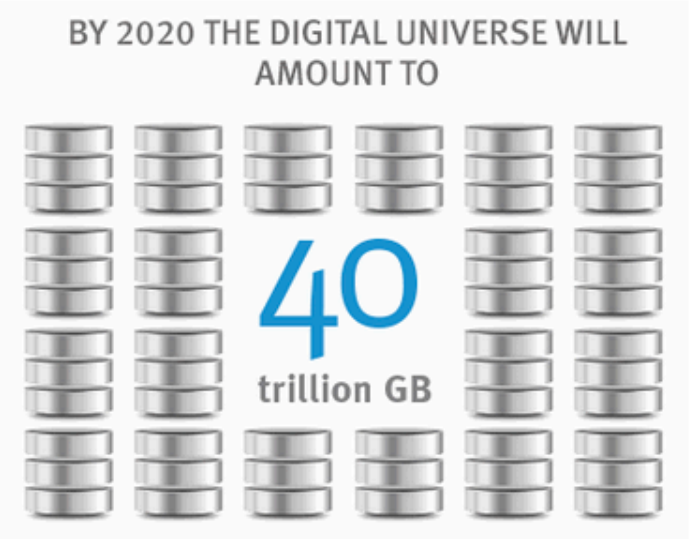
- TCGA, TARGET and ICGC
  - **Cancer Genomics Data Commons**
  - **NCI Cloud Pilots**
- Molecular Clinical Trials:
  - MPACT, MATCH, Exceptional Responders

Data are accumulating!

DATA KEEPS GROWING

The volume of digital data worldwide is growing rapidly, as the annual IDC Digital Universe study reveals. From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). From now until 2020, the digital universe will about double every two years.

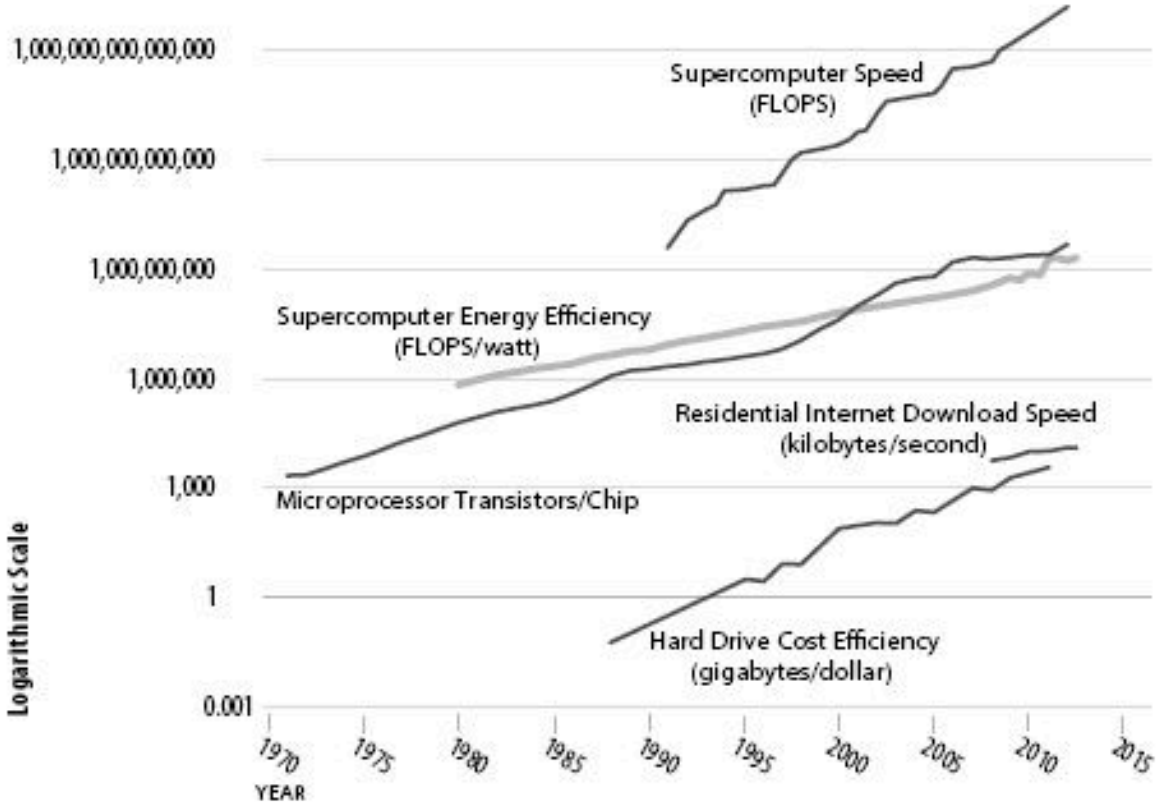
The majority of information in the digital universe, 68% in 2012, is created and consumed by consumers watching digital TV, interacting with social media, sending camera phone images and videos between devices and around the Internet, and so on.



INTERACTIVE: IDC Report: The Digital Universe in 2020.
Explore the IDC report, watch videos, and more.

From the Second Machine Age

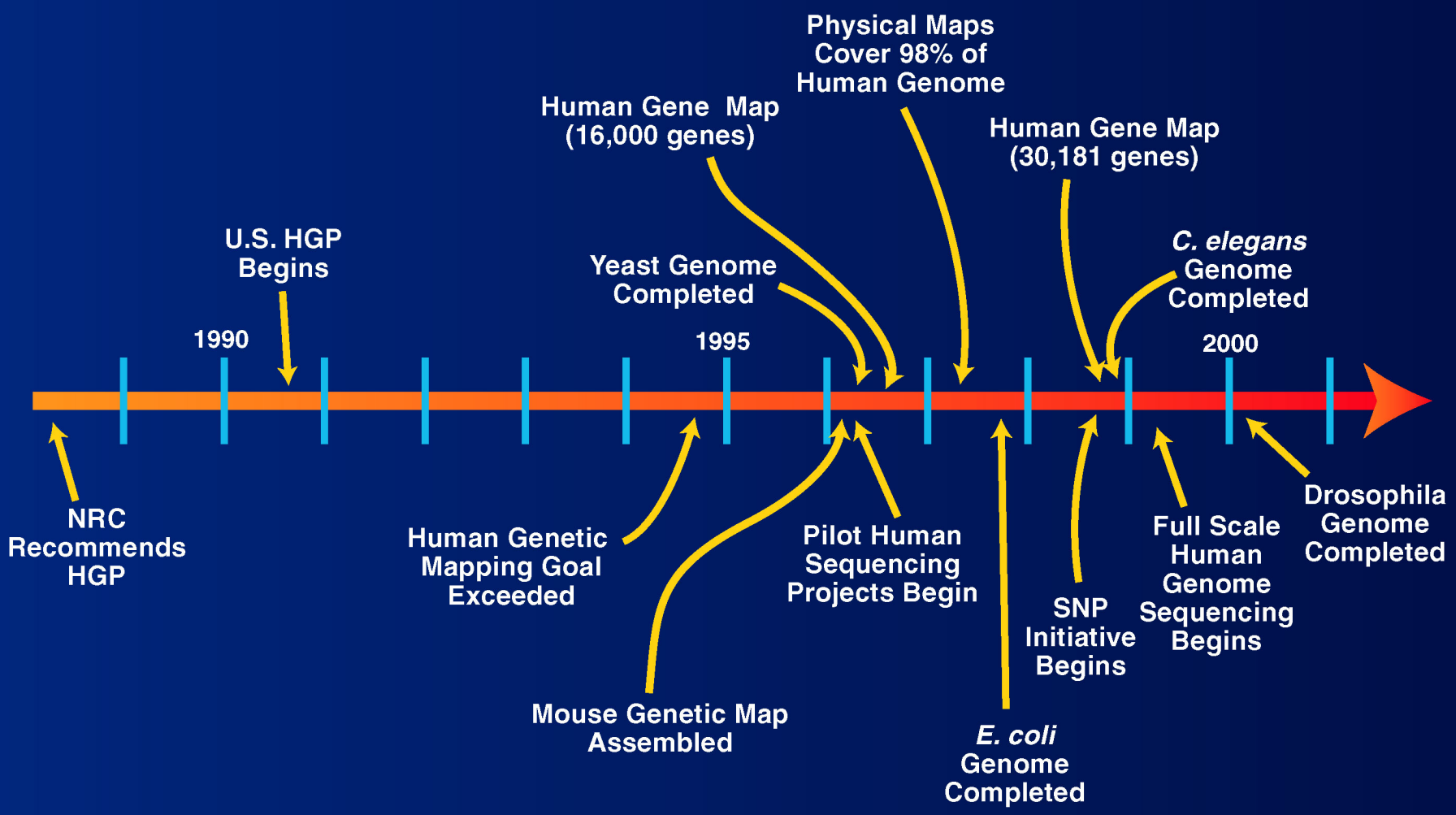
FIGURE 3.3 The Many Dimensions of Moore's Law

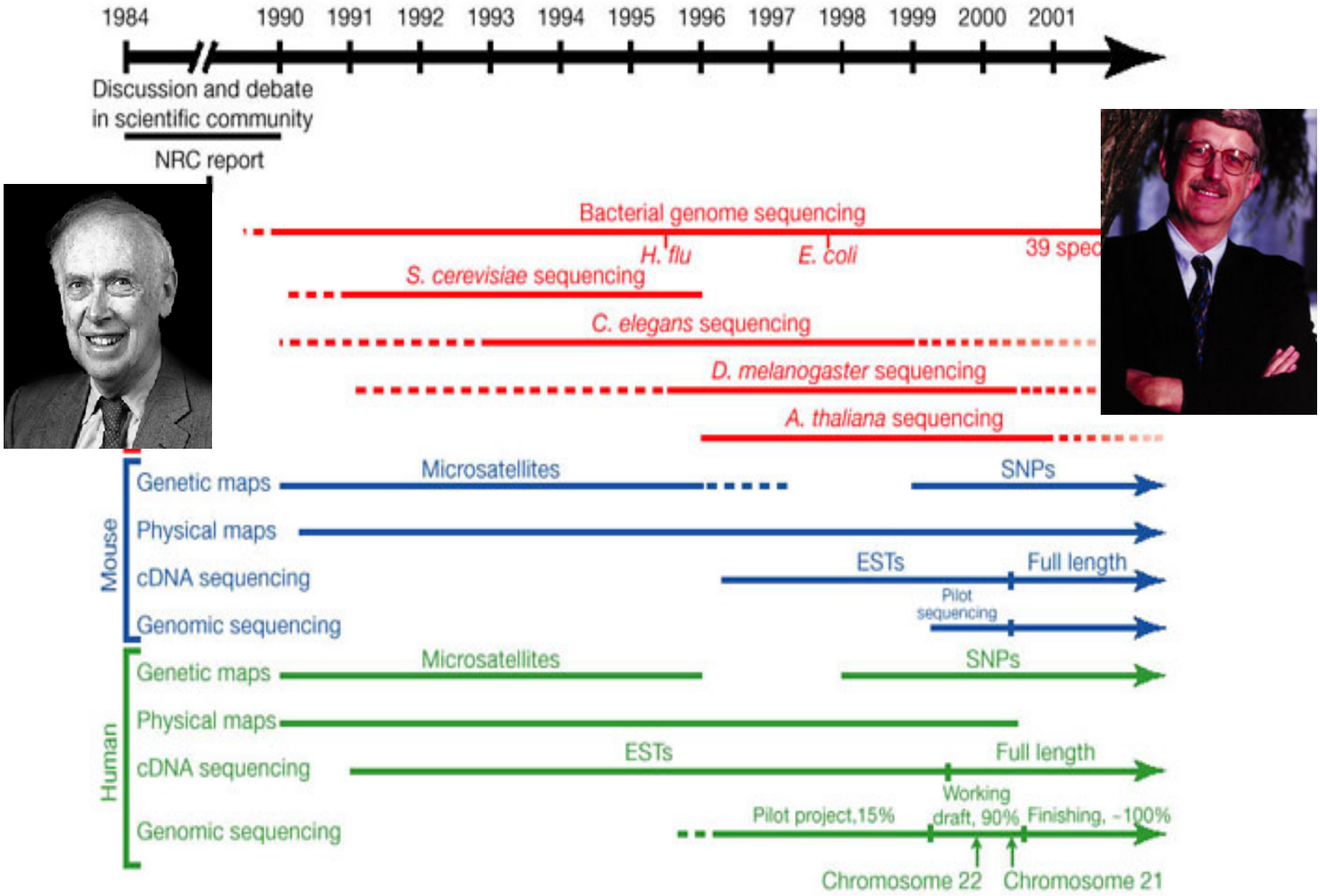


From: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies by Erik Brynjolfsson & Andrew McAfee

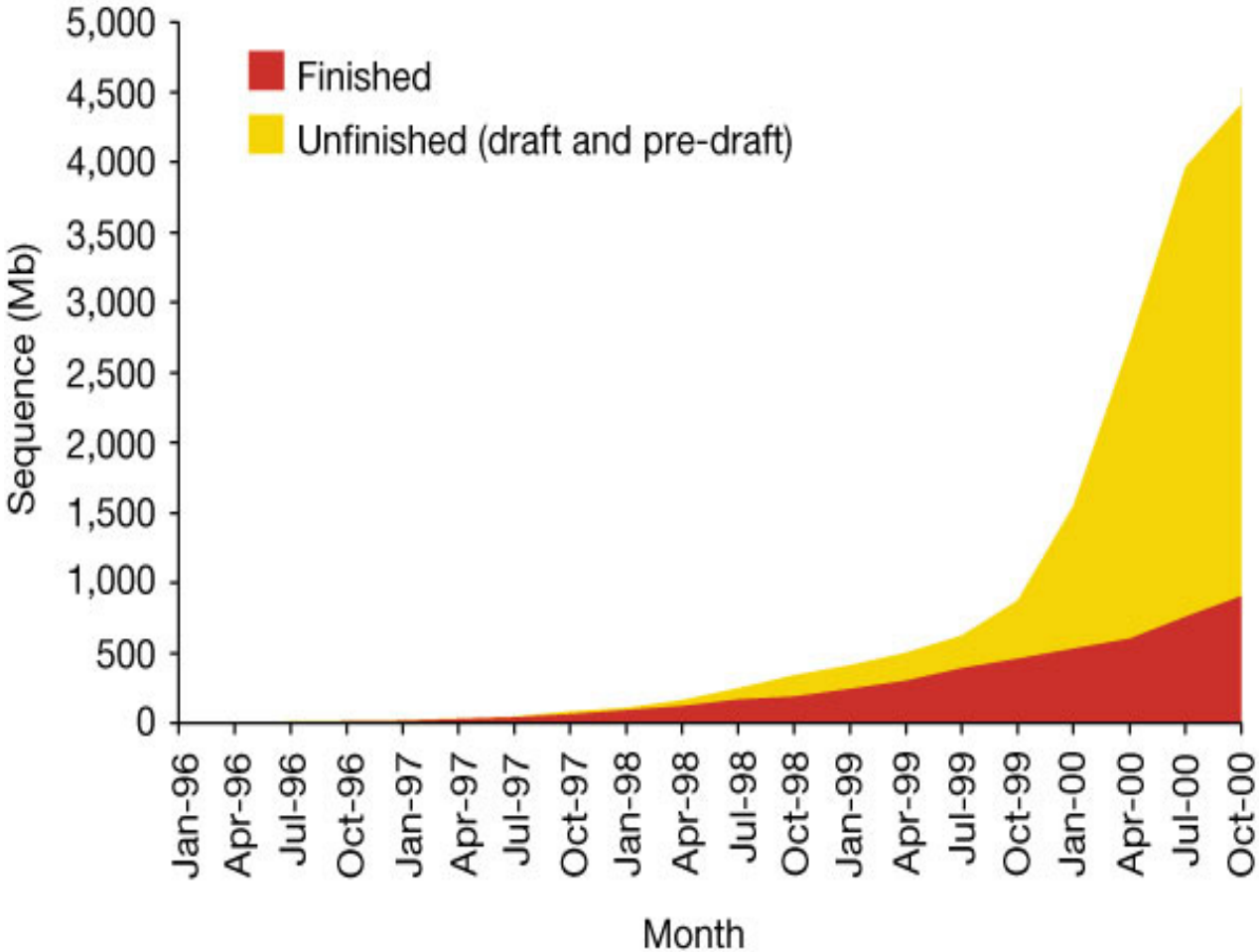
Molecular data is Big Data

- Brief trip down memory lane
- Sequencing and the Human Genome Project

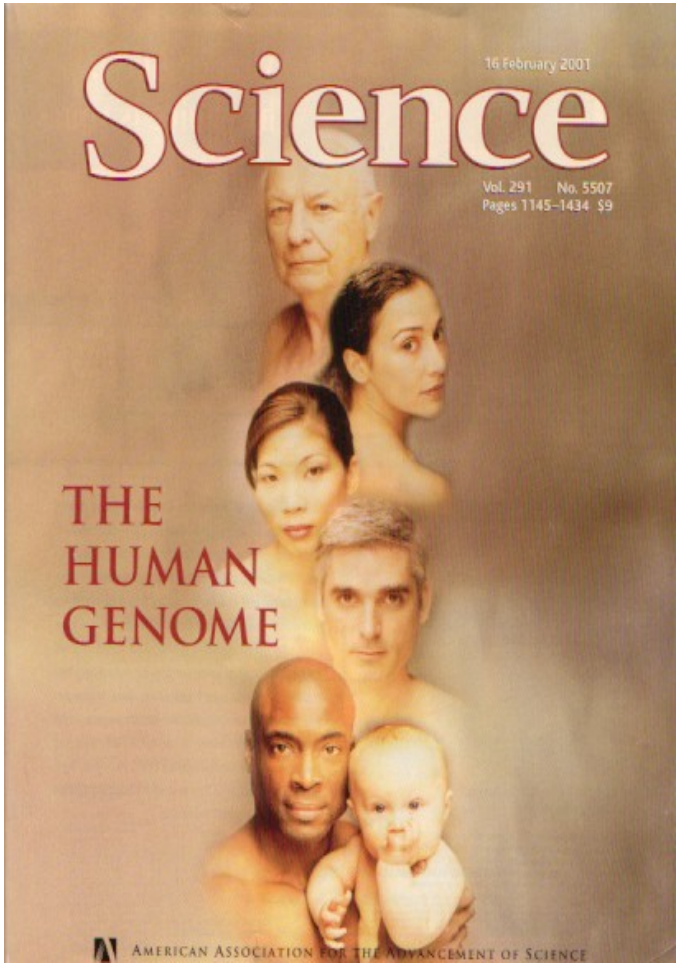

Physical Maps Cover 98% of Human Genome
Human Gene Map (16,000 genes)
Human Gene Map (30,181 genes)
U.S. HGP Begins
Yeast Genome Completed
C. elegans Genome Completed
1990
1995
2000
NRC Recommends HGP
Human Genetic Mapping Goal Exceeded
Pilot Human Sequencing Projects Begin
Full Scale Human Genome Sequencing Begins
Drosophila Genome Completed
SNP Initiative Begins
Mouse Genetic Map Assembled
E. coli Genome Completed


1984
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
Discussion and debate in scientific community
NRC report
Bacterial genome sequencing
H. flu
E. coli
39 spec
S. cerevisiae sequencing
C. elegans sequencing
D. melanogaster sequencing
A. thaliana sequencing
Mouse
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Pilot sequencing
Genomic sequencing
Human
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Genomic sequencing
Pilot project,15%
Working draft, 90%
Finishing, ~100%
Chromosome 22
Chromosome 21

GenBank High Throughput Genome Sequence (HTGS)



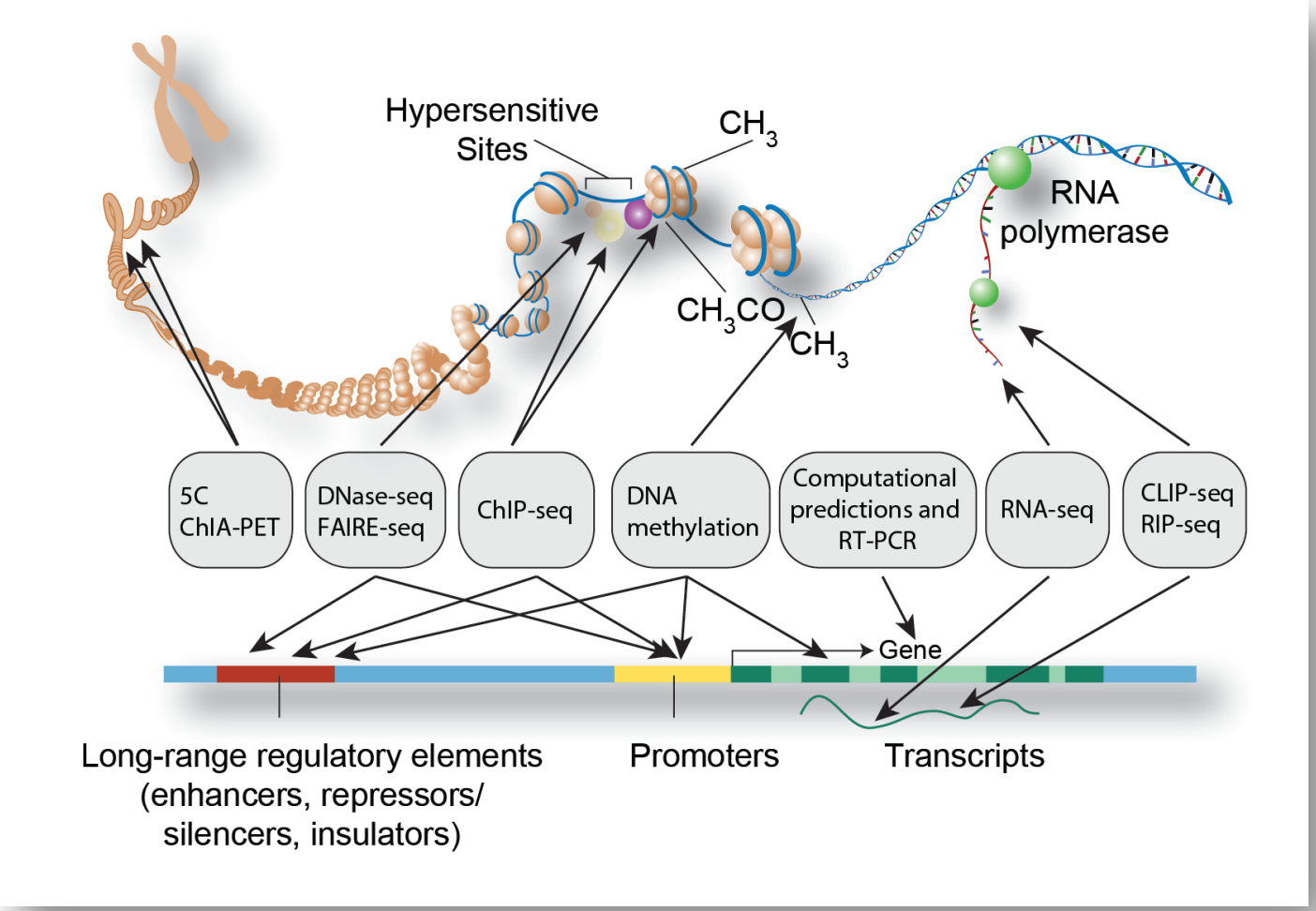
February 12, 2001



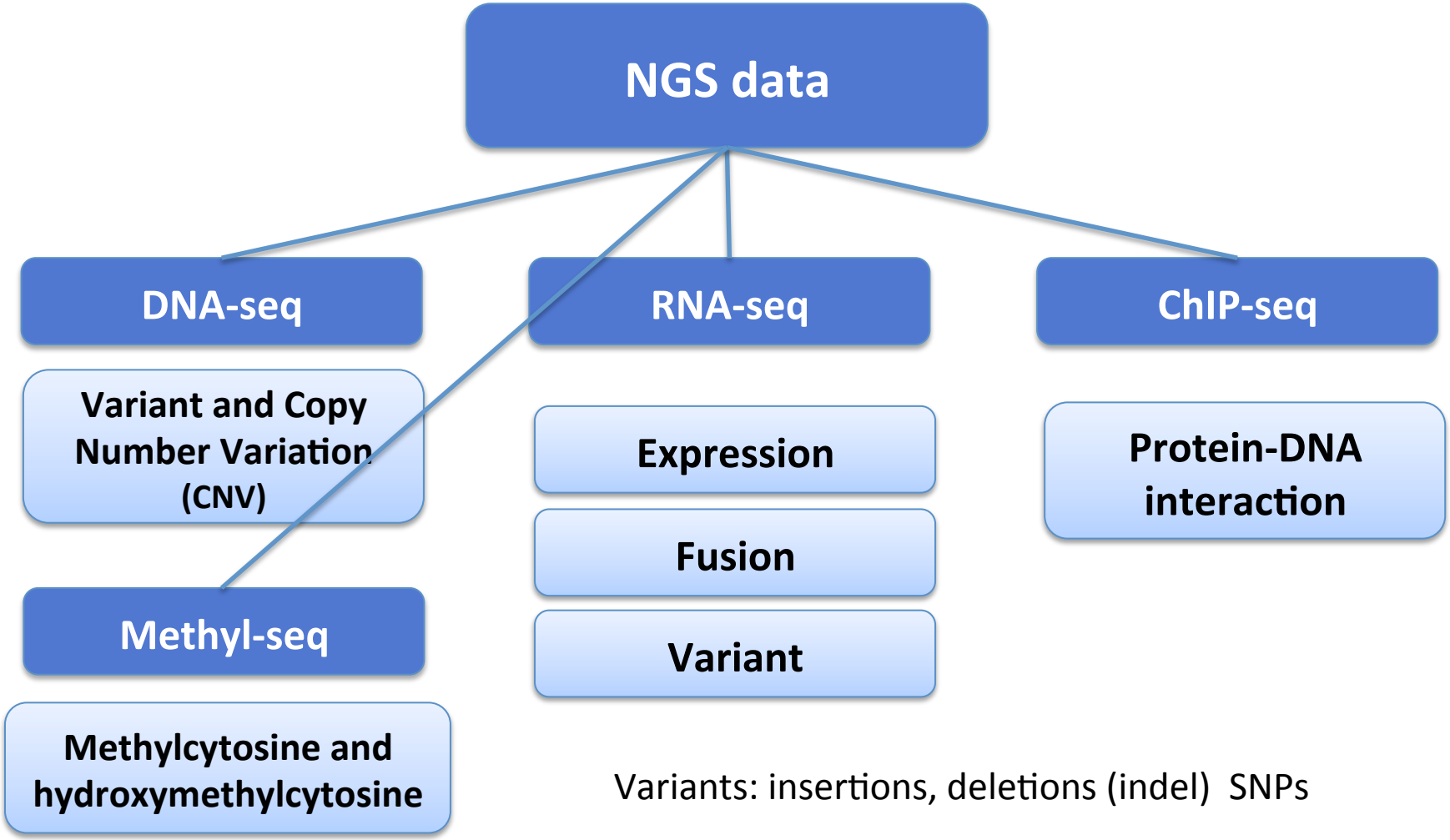
HGP outcomes

- $5.6B investment in 2010 dollars
- $800B economic development
- Enabled many basic discoveries, clinical therapies and diagnostics, and applied technologies

Assays and Data Types

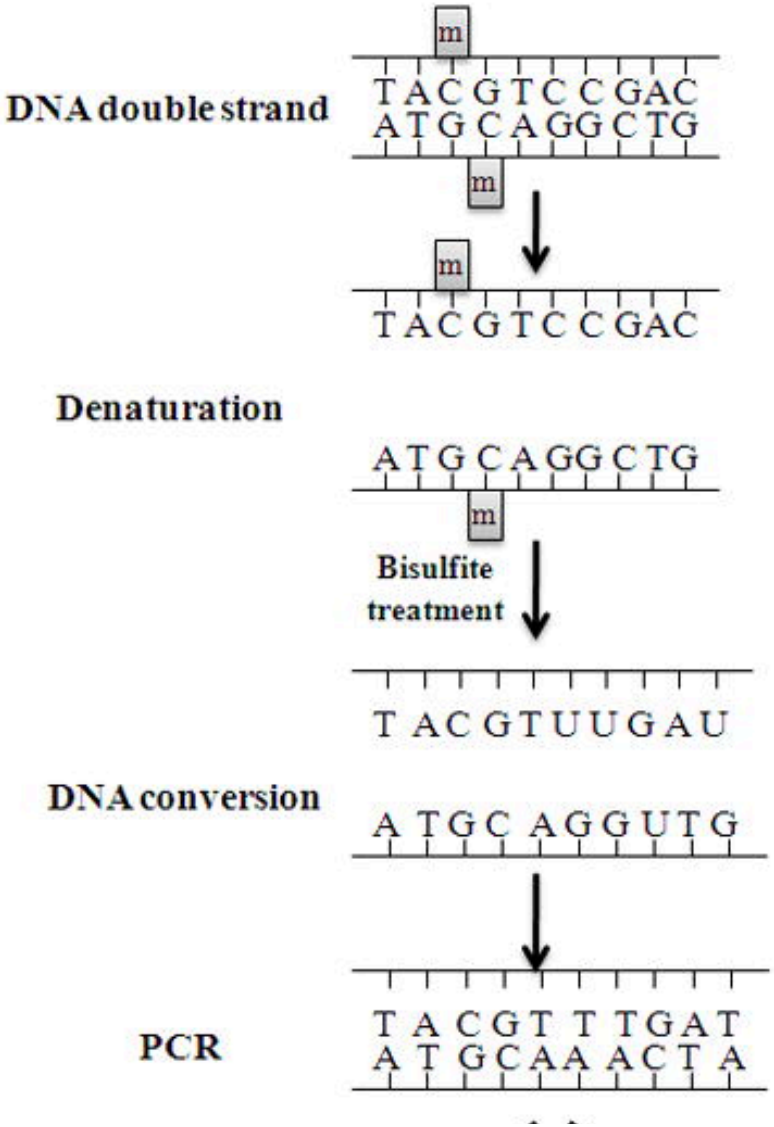


Next Gen Sequencing technologies



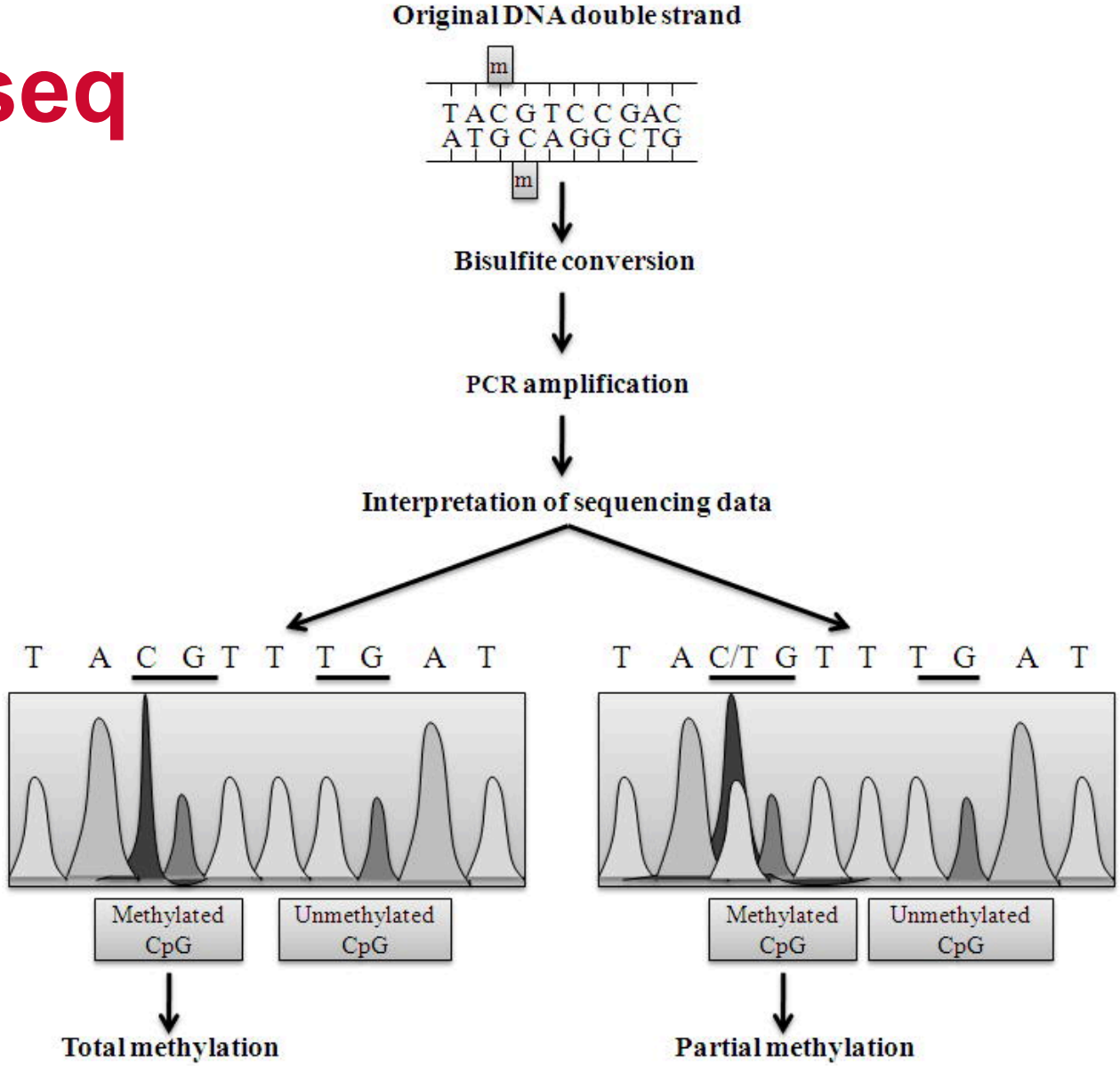
Variants: insertions, deletions (indel) SNPs

Methyl-seq



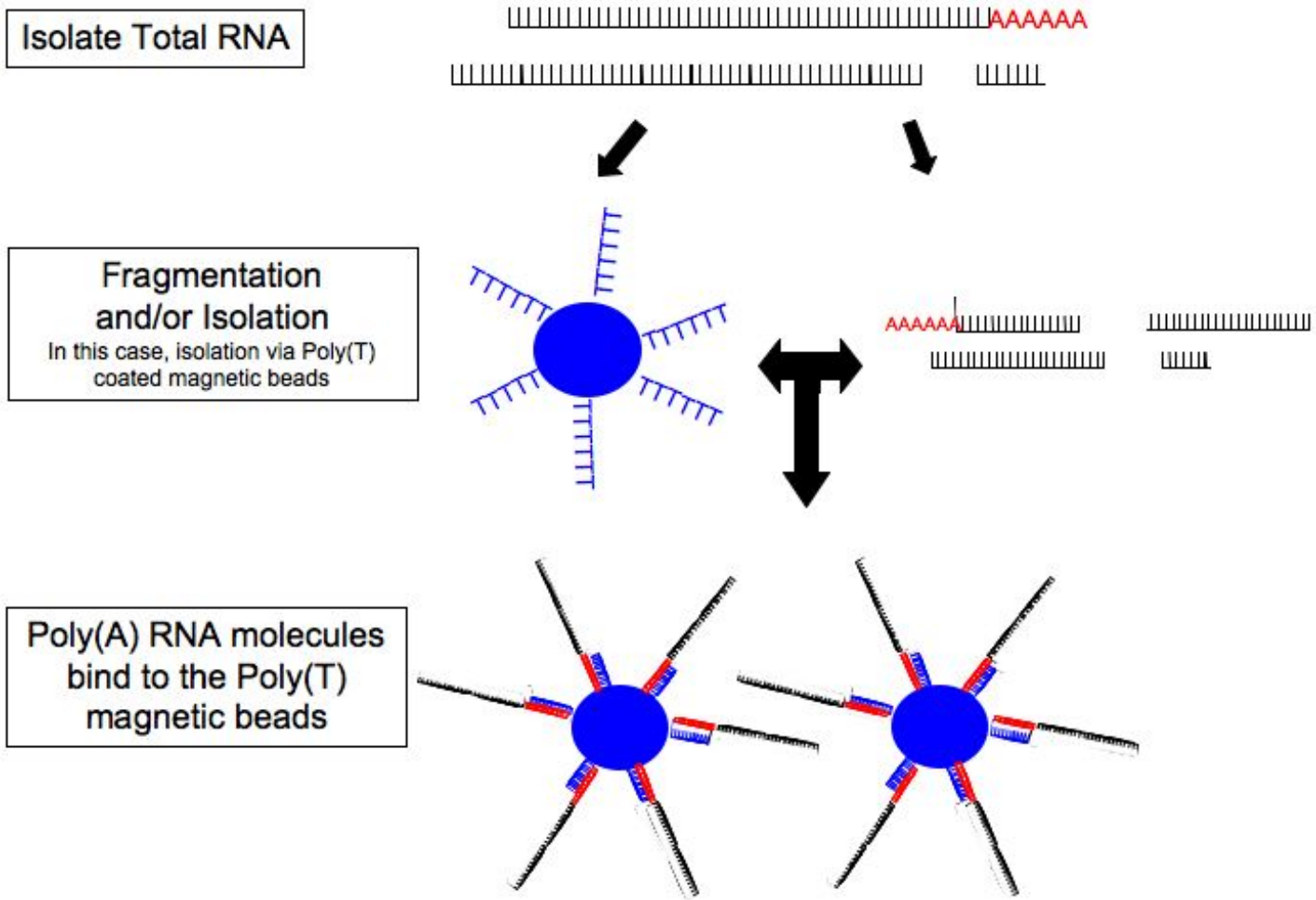
Li and Tollefsbol, Methods Mol Biol. 2011;
791: 11–21. doi: 10.1007/978-1-61779-316-5_2

Methyl-seq



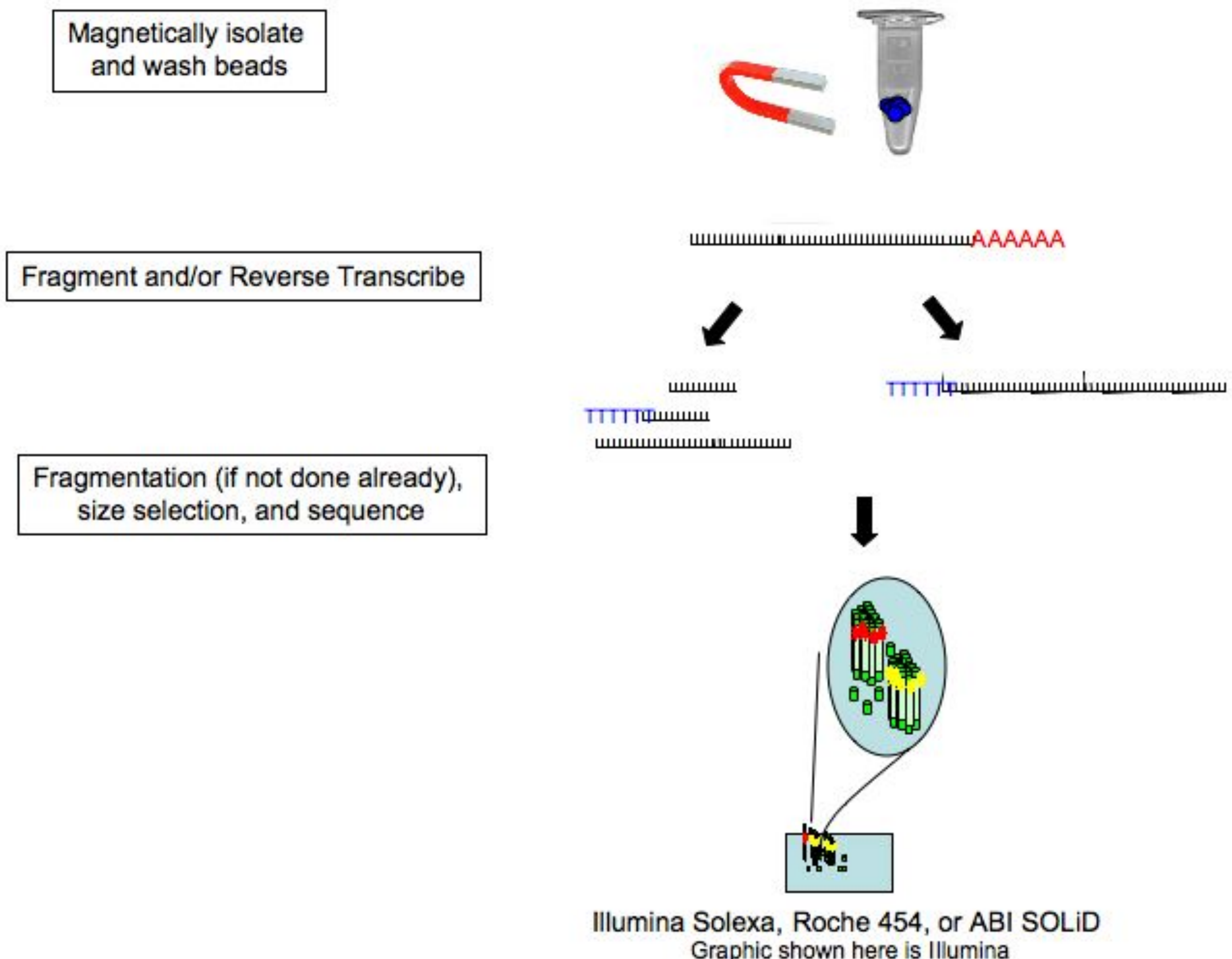
Li and Tollefsbol, Methods Mol Biol. 2011; 791: 11–21. doi: 10.1007/978-1-61779-316-5_2

RNA-seq



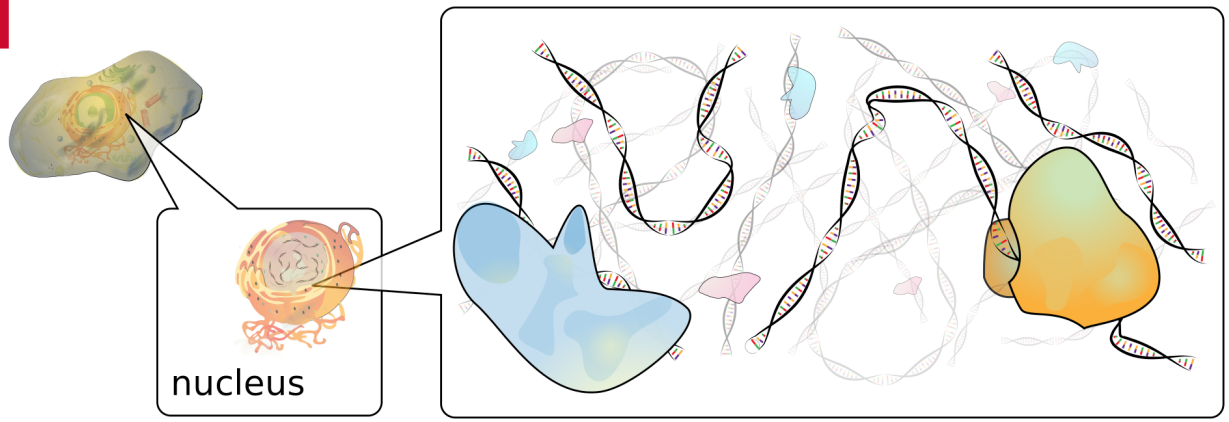
Courtesy of Vincent Montoya

RNA-seq

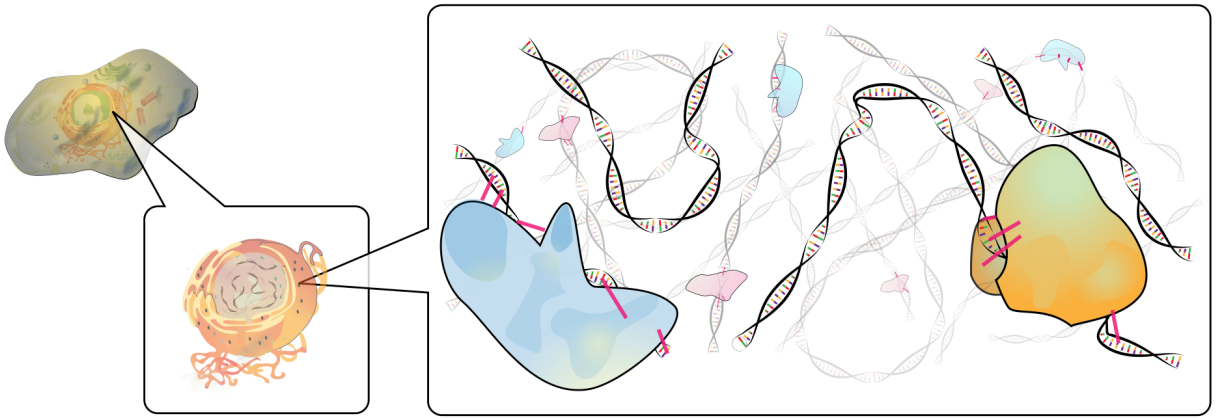


Courtesy of Vincent Montoya

ChIP-seq

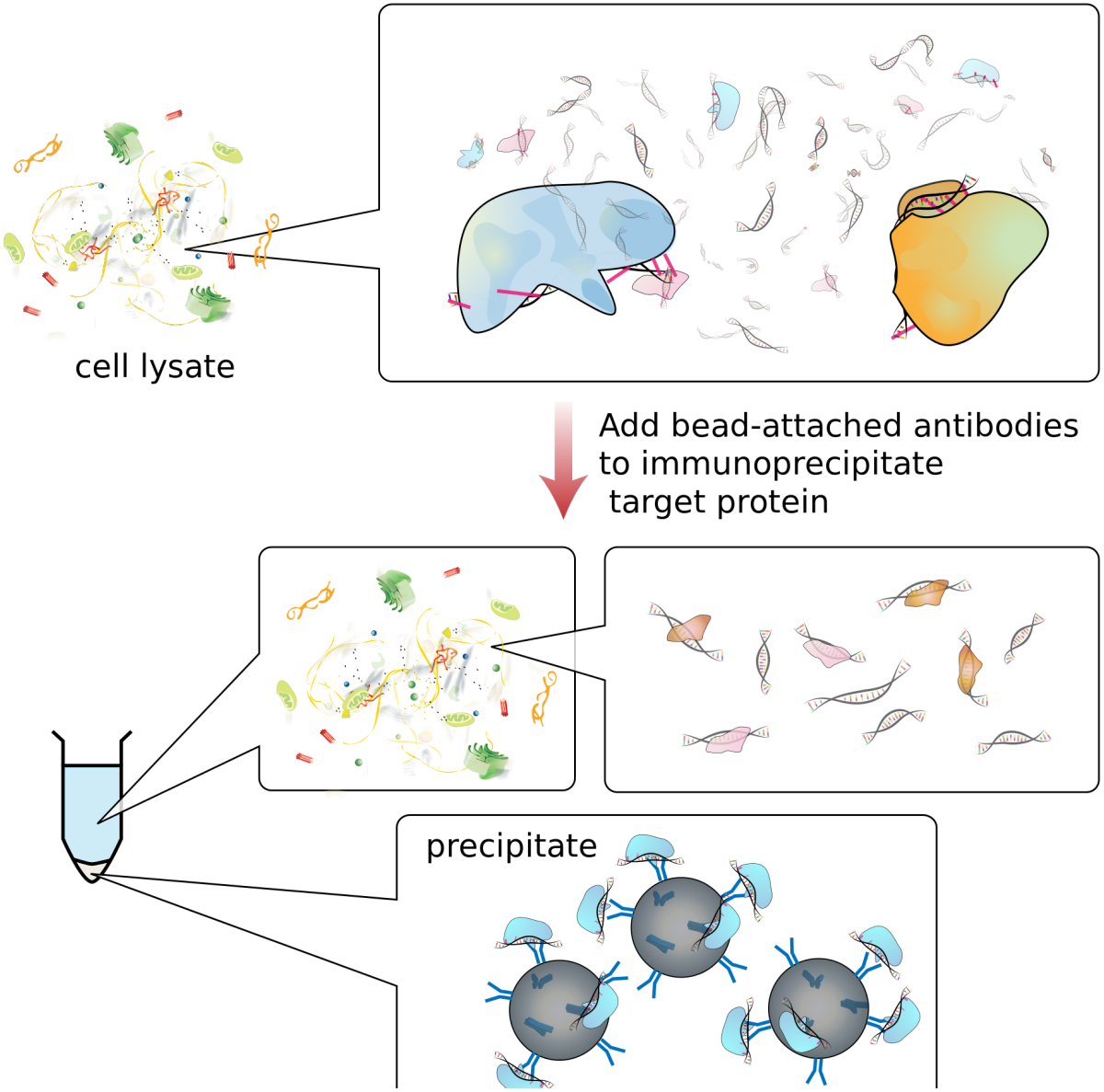


Cross-link protein to DNA

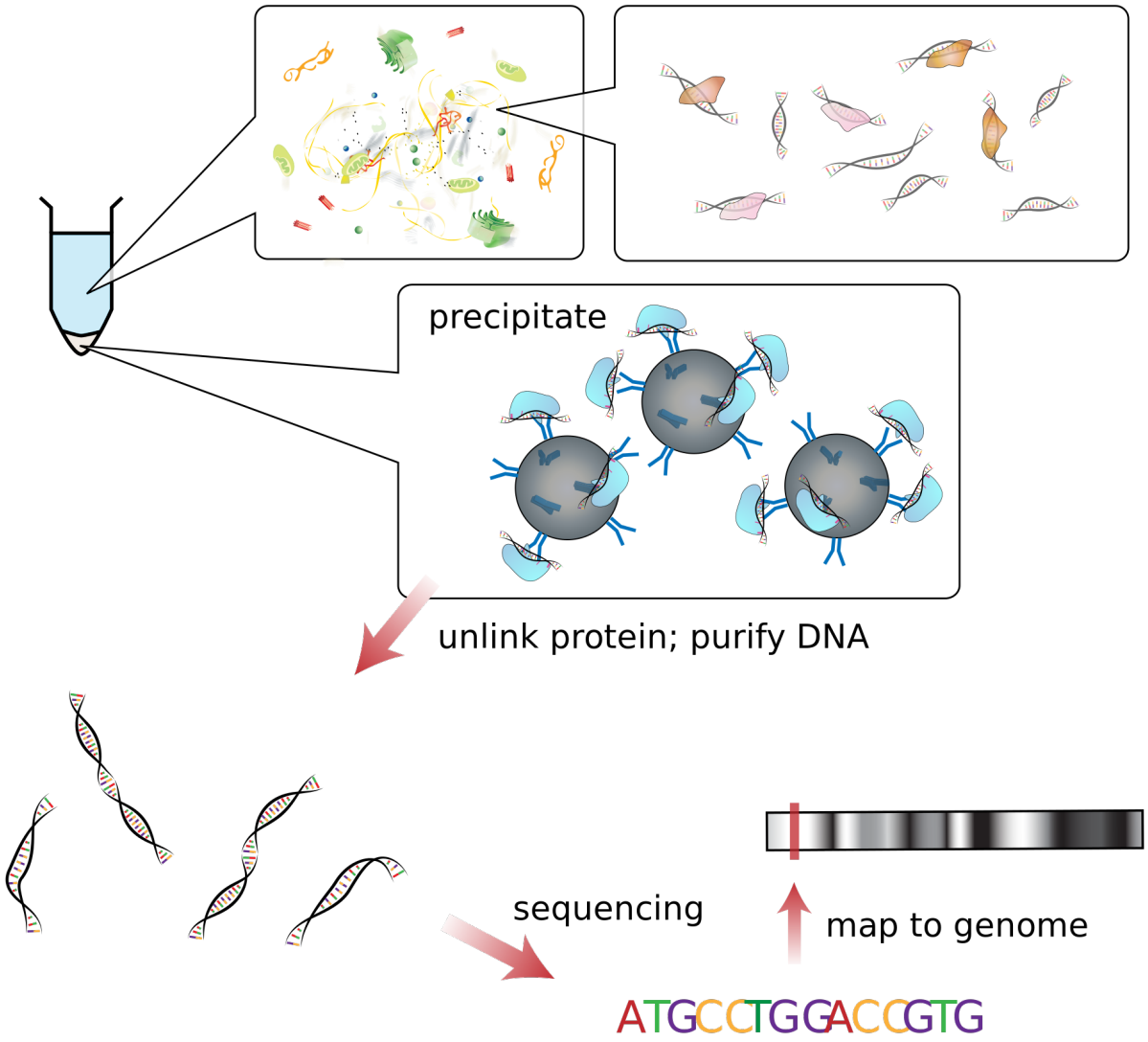


Shear DNA strands
by sonicating

ChIP-seq


cell lysate
Add bead-attached antibodies
to immunoprecipitate
target protein
precipitate

ChIP-seq


precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

TCGA history

- About three years post-HGP
- Initiated in 2005
- Collaboration of NHGRI and NCI to examine GBM, Lung and Ovarian cancer using genomic techniques in 2006.
- Expanded to 20+ tumor types.

TCGA drivers

- Providing high quality reference sets for 20+ tissue types
- Providing a platform for systems biology and hypothesis generation
- Providing a test bed for understanding the real world implications of consent and data access policies on genomic and clinical data.

Highly Recurrent *TERT* Promoter Mutations in Human Melanoma

Franklin
Lynda C

Systema
in the p
two ind
the gen
(71%) i
for E-tw
transcri
cell line
with pre
Thus, so
tumorig

TERT Promoter Mutations in Familial and Sporadic Melanoma

Susan
Antje
Kari H

Cutane
melan
diseas
(*TERT*)
bindin
start a
screen
somatic mut
melanomas,
primary mel
TERT promot

TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal

Pat
Alla
Yipi
B. A
Mic
Rog
and

A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer

Julius Gudmun
Kristrun R Bene
Droplaug N Mag
Stefania B Olafs
Fernando Fuerte
Inge M van Oor
Kin-Mang Lau[17]
Gudmundur V
Lambertus A Ki

Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer

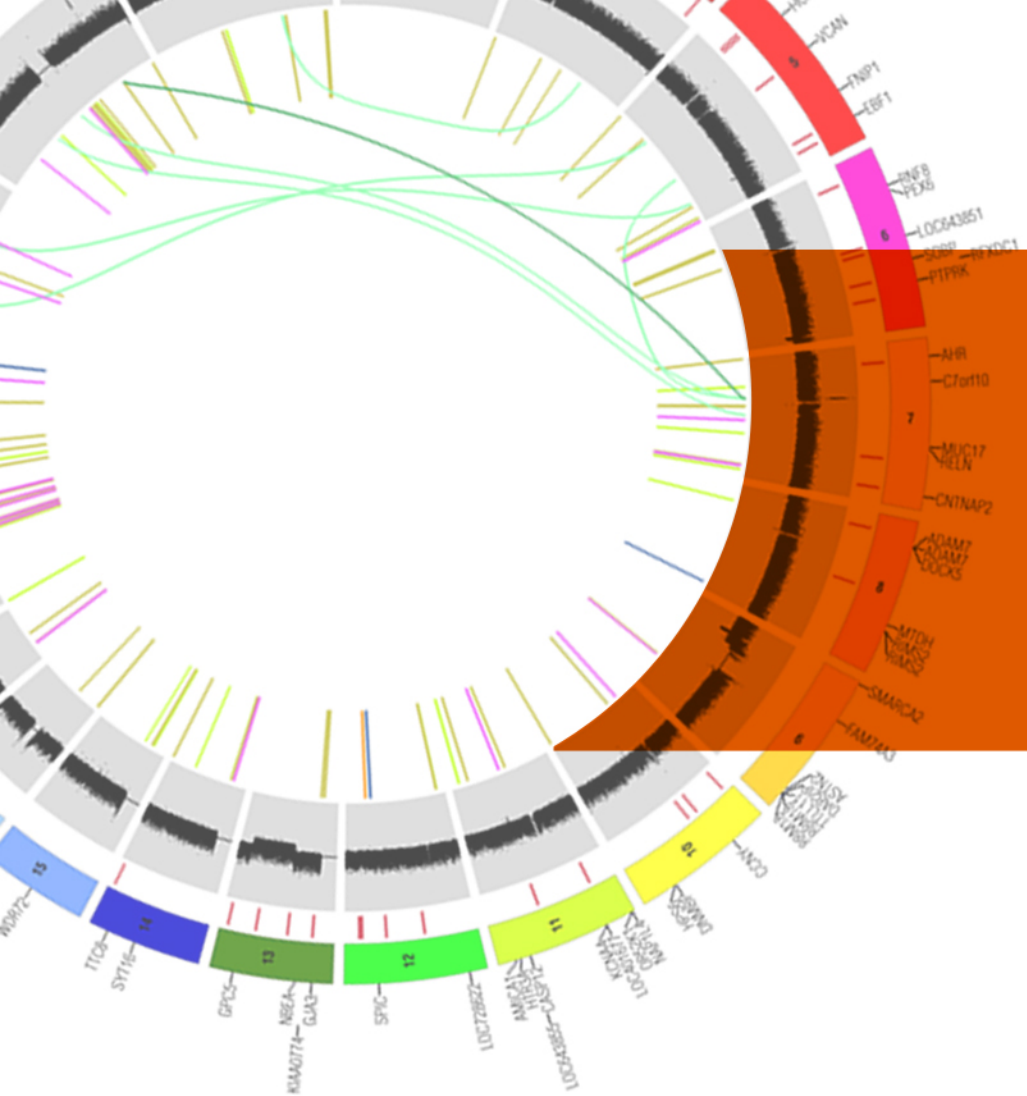
Batool Akhtar-Zaid
Cynthia F. Bartels,
Awad Jarrar,[5] Matt
Thomas Laframbois

Cancer is characteri
coding sequences and promoters, even though distal regulatory elements play a central role

DNA Methylation of Transcriptional Enhancers and Cancer Predisposition

Focus on TCGA

- TCGA consortium slides
- Thanks to Lou Staudt and Jean Claude Zenklusen



The Cancer Genome Atlas

TCGA – Lessons from structural genomics

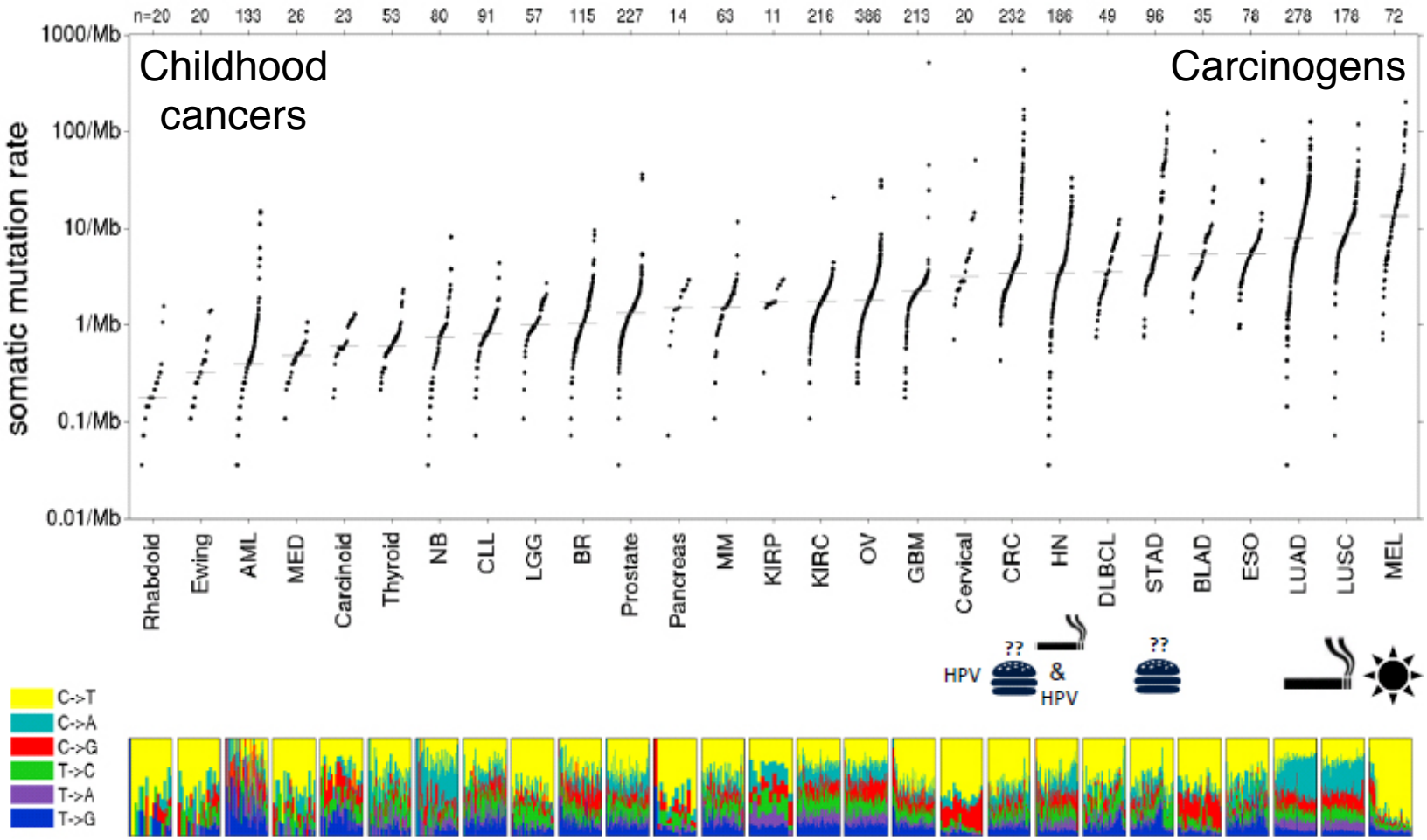
Jean Claude Zenklusen, Ph.D.

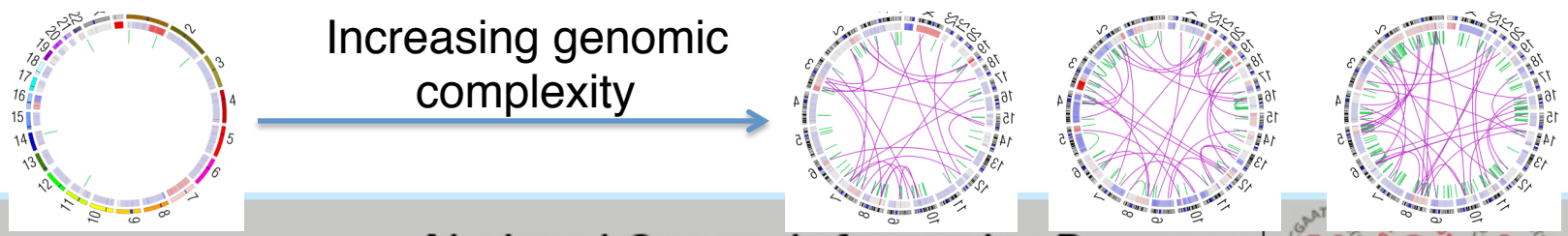
Director

TCGA Program Office

National Cancer Institute

The Mutational Burden of Human Cancer





ARTICLE

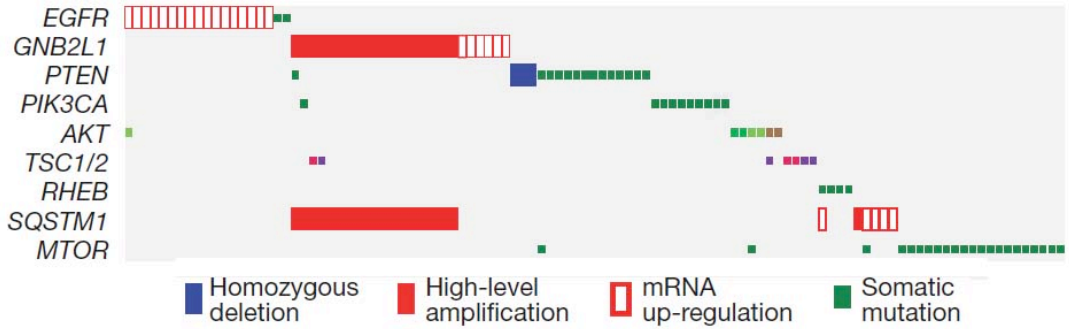
OPEN
doi:10.1038/nature12222

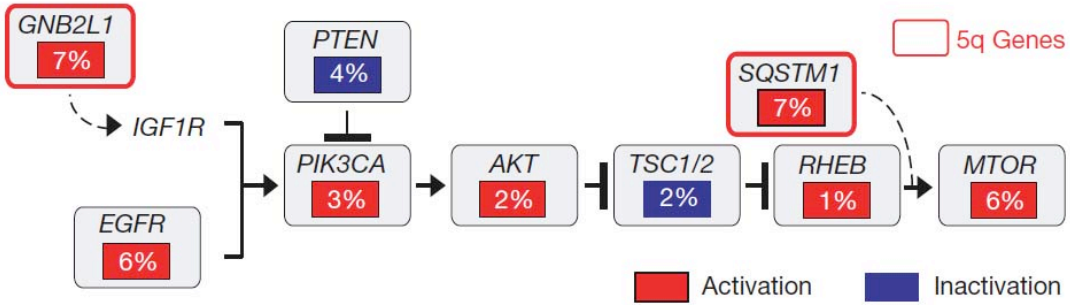
Comprehensive molecular characterization of clear cell renal cell carcinoma

The Cancer Genome Atlas Research Network*

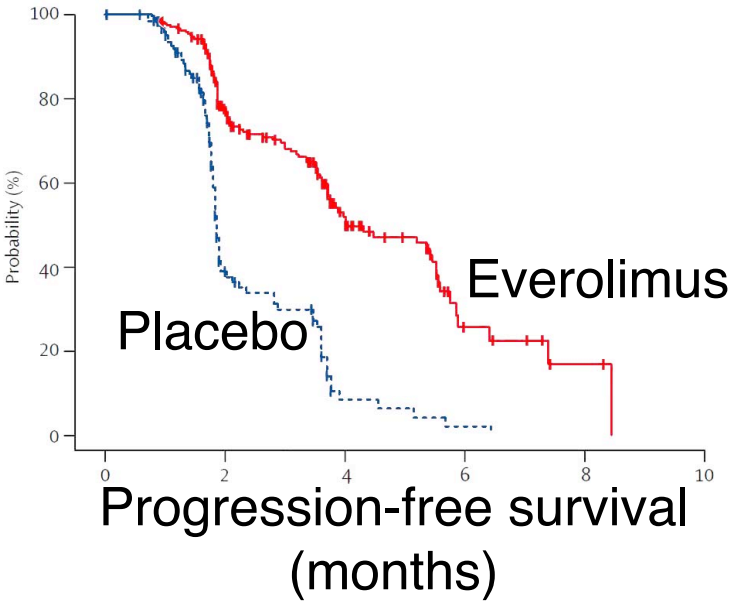
Frequent Activation of the PI(3)K Pathway in Clear Cell Renal Carcinoma

PI(3)K aberrations (28% of cases)

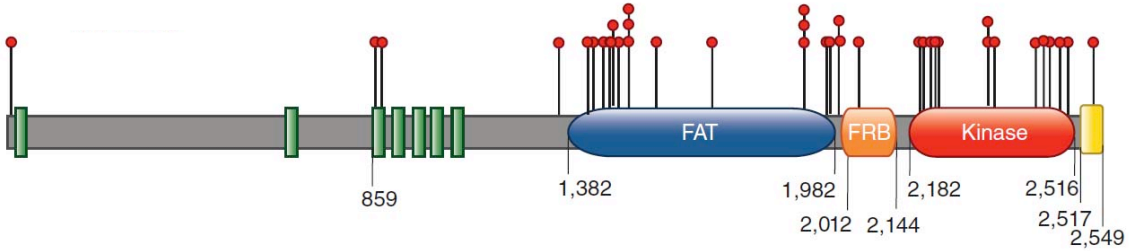




Response of RCC To Everolimus



mTOR mutations



TCGA Nature 499:45 (2013)

Sato et al Nat Gen 45:860 (2013)

Hakimi et al Nat Gen 45:849 (2013)

Motzer et al Lancet 372:449 (2008)

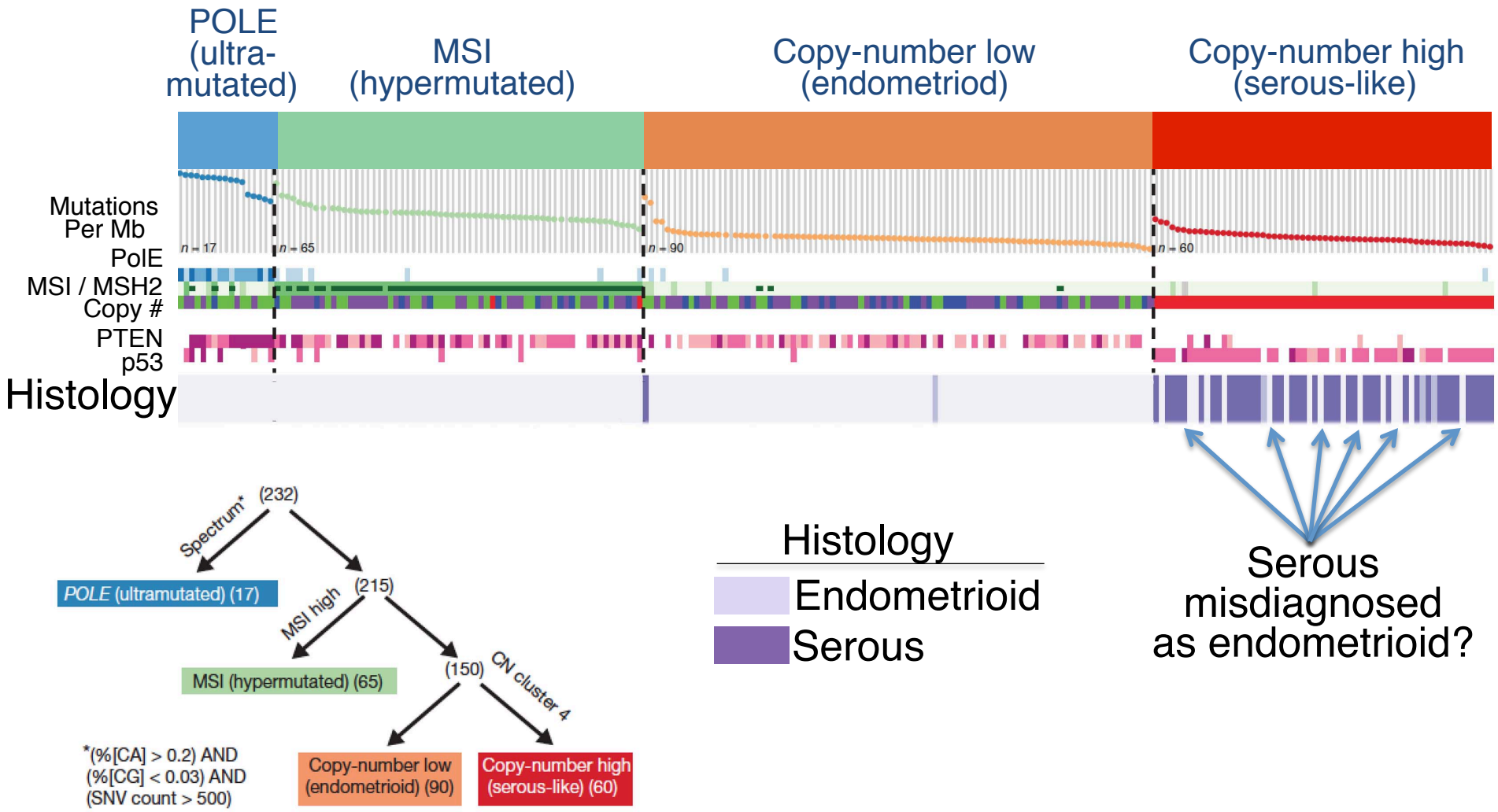
ARTICLE

OPEN
doi:10.1038/nature12113

Integrated genomic characterization of endometrial carcinoma

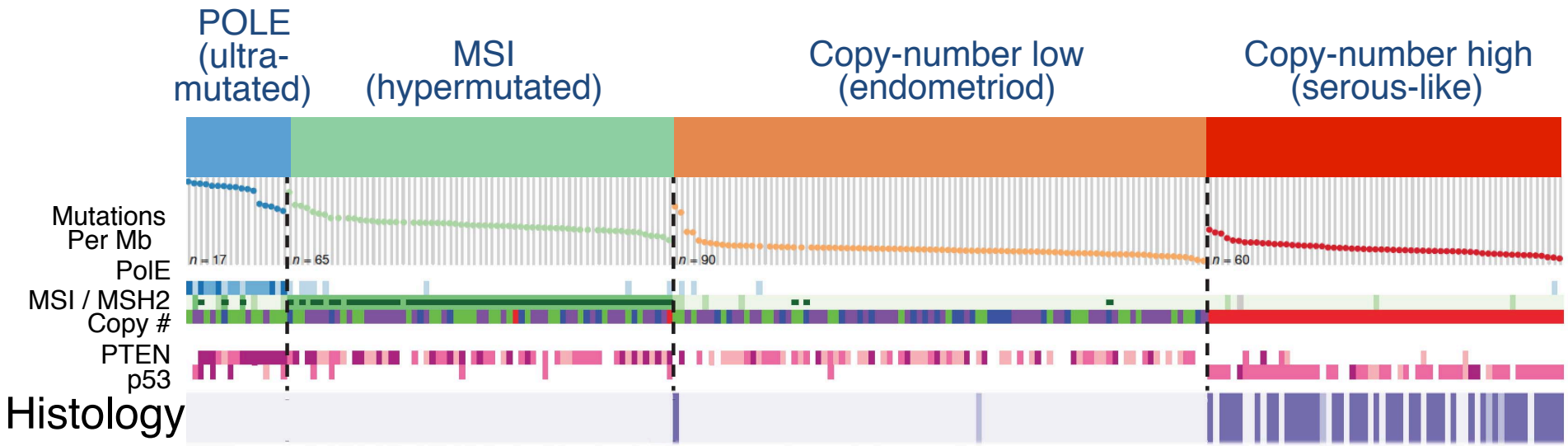
The Cancer Genome Atlas Research Network*

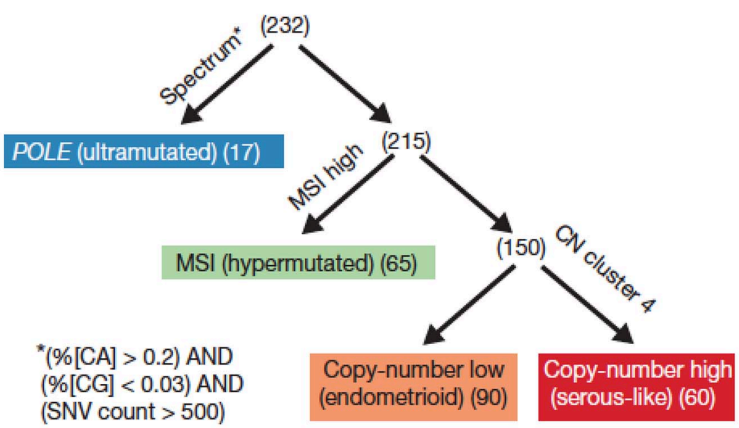
Molecular Subgroups Refine Histological Diagnosis Of Endometrial Carcinoma

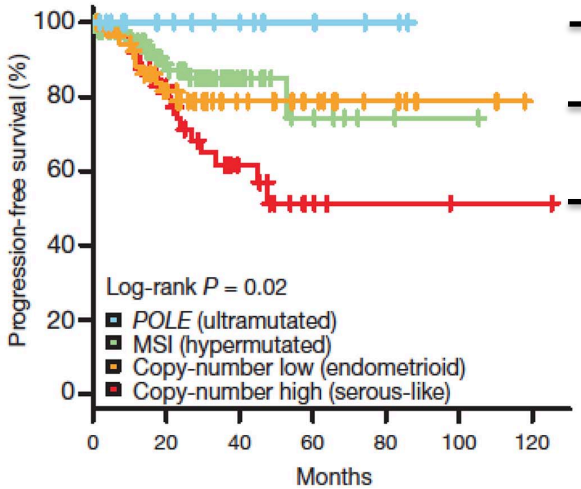


TCGA Nature 497:67 (2013)

Molecular Diagnosis of Endometrial Cancer May Influence Choice of Therapy







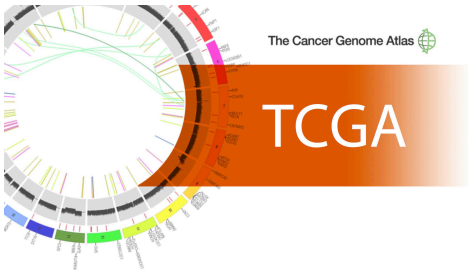
→ Surgery only?

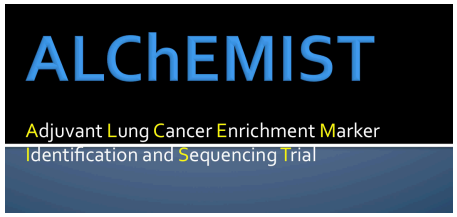
→ Adjuvant radiotherapy?

→ Adjuvant chemotherapy?

TCGA Nature 497:67 (2013)

NCI Cancer Genomics Data Commons

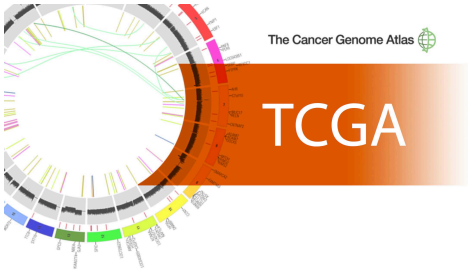


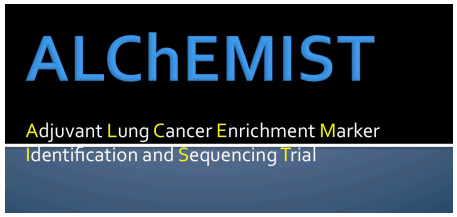


Genomic + clinical data

NCI Genomics Data Commons

NCI Cancer Genomics Data Commons





Genomic + clinical data

Cancer information donor

NCI Genomics Data Commons

Utility of a Cancer Knowledge Base

Cancer information donor

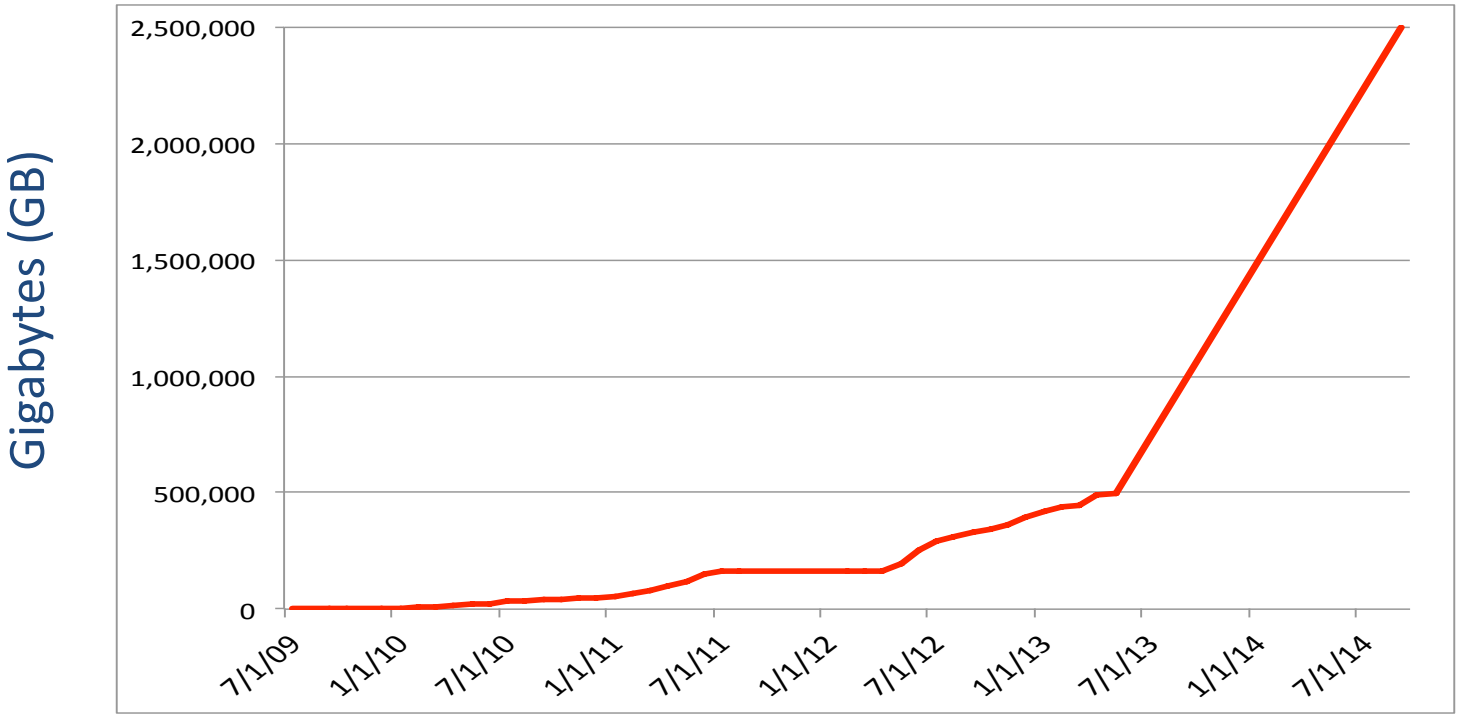
Identify low-frequency cancer drivers

Define genomic determinants of response to therapy

Compose clinical trial cohorts sharing Targeted genetic lesions

Driver for the Cloud Pilots

- An inflection point for TCGA is looming



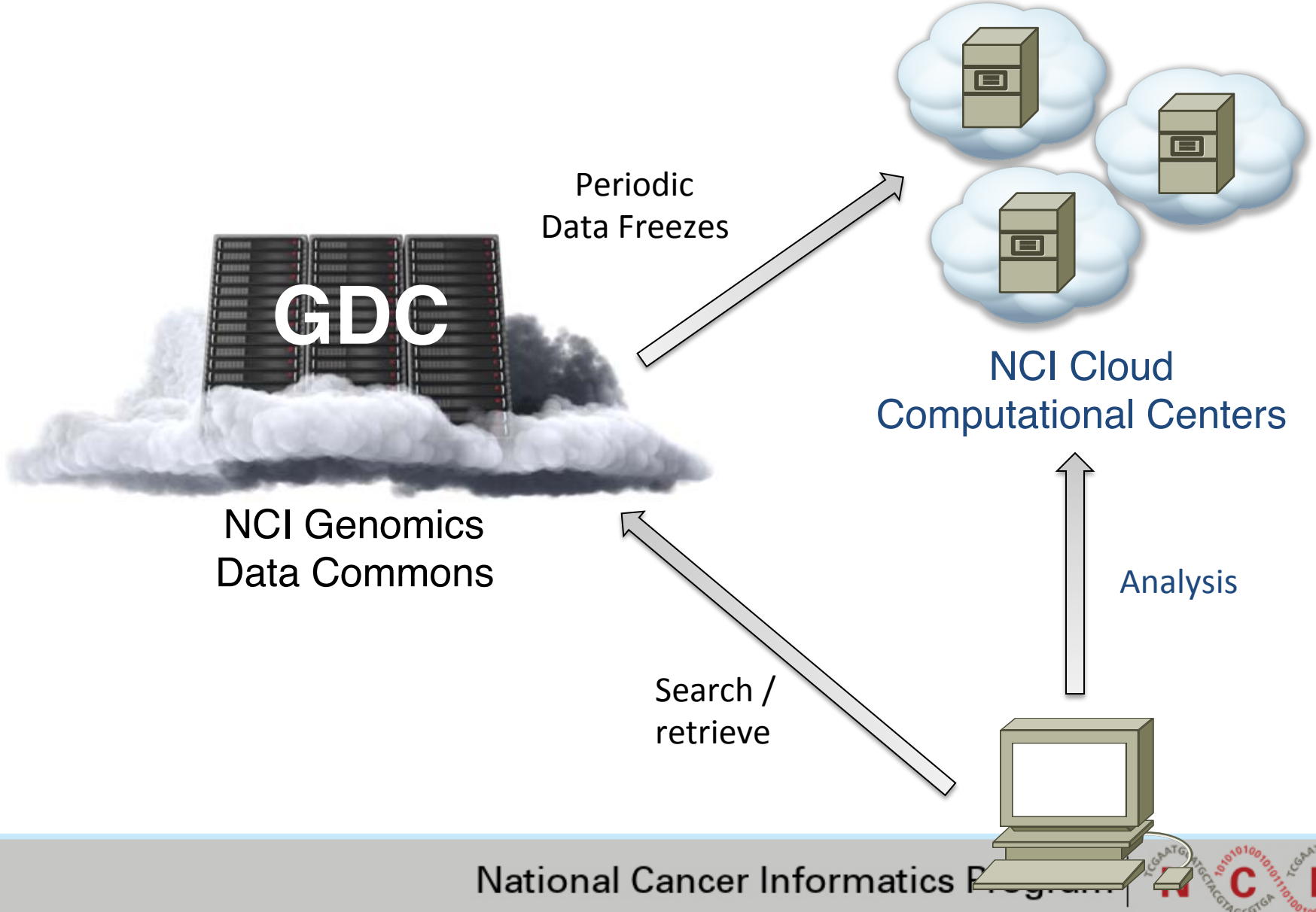
NCI Cloud Pilots

- Funding for up to 3 cloud pilots - 24 month pilots that are meant to inform the Cancer Genomics Data Commons
  - Explore models for cancer genomics APIs
  - Explore cloud models for data+analysis
- Announced this week: The Institute for Systems Biology, The Broad Institute, and Seven Bridges will be the initial consortium

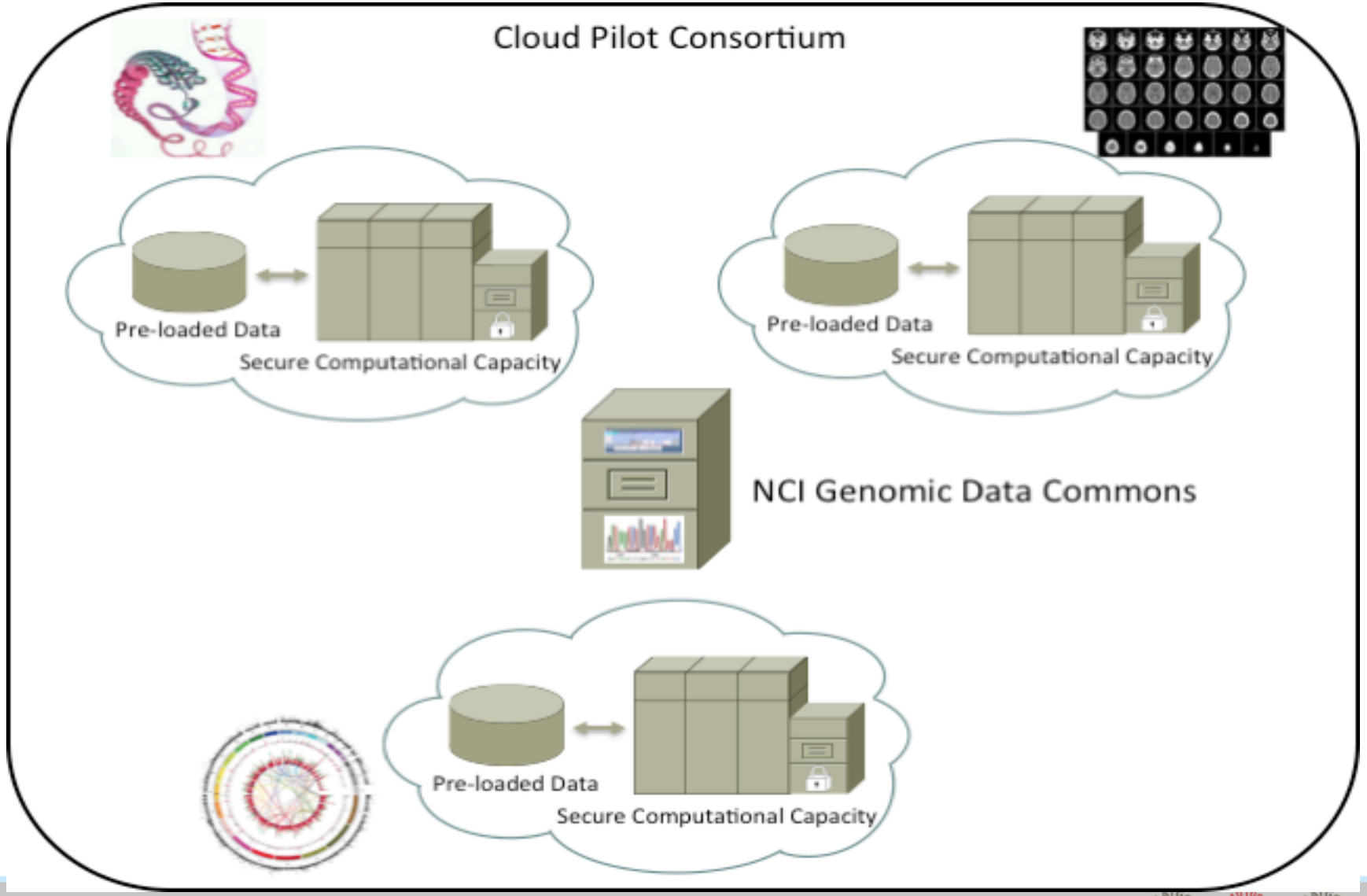
NCI Cloud Pilots

- A way to move computation to the data
- Sustainable models for providing access to data
- Reproducible pipelines for QA, variant calling, knowledge sharing
- Define genomics/phenomics APIs for discovering new variants contributing to cancer, enhancing response, modulating risk

Relationship of the Cancer Genomics Data Commons and NCI Cloud Pilots



Cancer Genomics Cloud Pilots



Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing

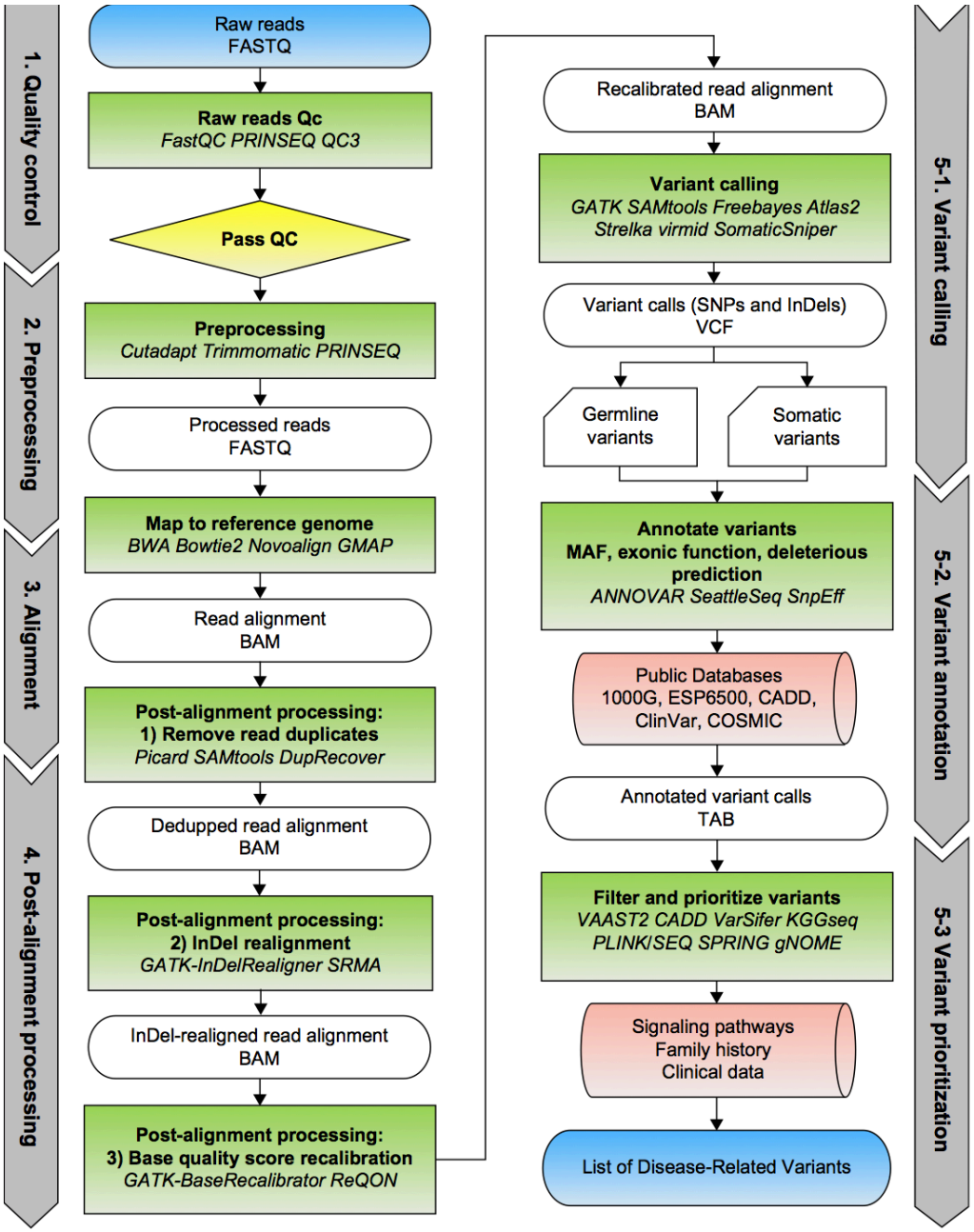
Riyue Bao[1,*], Lei Huang[1,*], Jorge Andrade[1,*], Wei Tan[2,*], Warren A. Kibbe[3], Hongmei Jiang[4,§] and Gang Feng[3,§]

[1]Center for Research Informatics, The University of Chicago, Chicago, IL, USA. [2]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. [3]Biomedical Informatics Center (NUBIC), Clinical and Translational Sciences Institute (NUCATS), Northwestern University, Chicago, IL, USA. [4]Department of Statistics, Northwestern University, Evanston, IL, USA. *These authors contributed equally to this work. [§]Co-corresponding authors.

CITATION: Bao et al. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Cancer Informatics* 2014:13(s2) 67–82 doi: 10.4137/CIn.s13779.

WES workflow

figure 1. A general framework of WES data analysis. Five major steps are shown: raw reads QC, preprocessing, alignment, post-processing, and variant analysis (variant calling, annotation, and prioritization). **Notes:** FASTQ, BAM, variant call format (VCF), and TAB (tab-delimited) refer to the standard file format of raw data, alignment, variant calls, and annotated variants, respectively. A selection of tools supporting each analysis step is shown in italic.



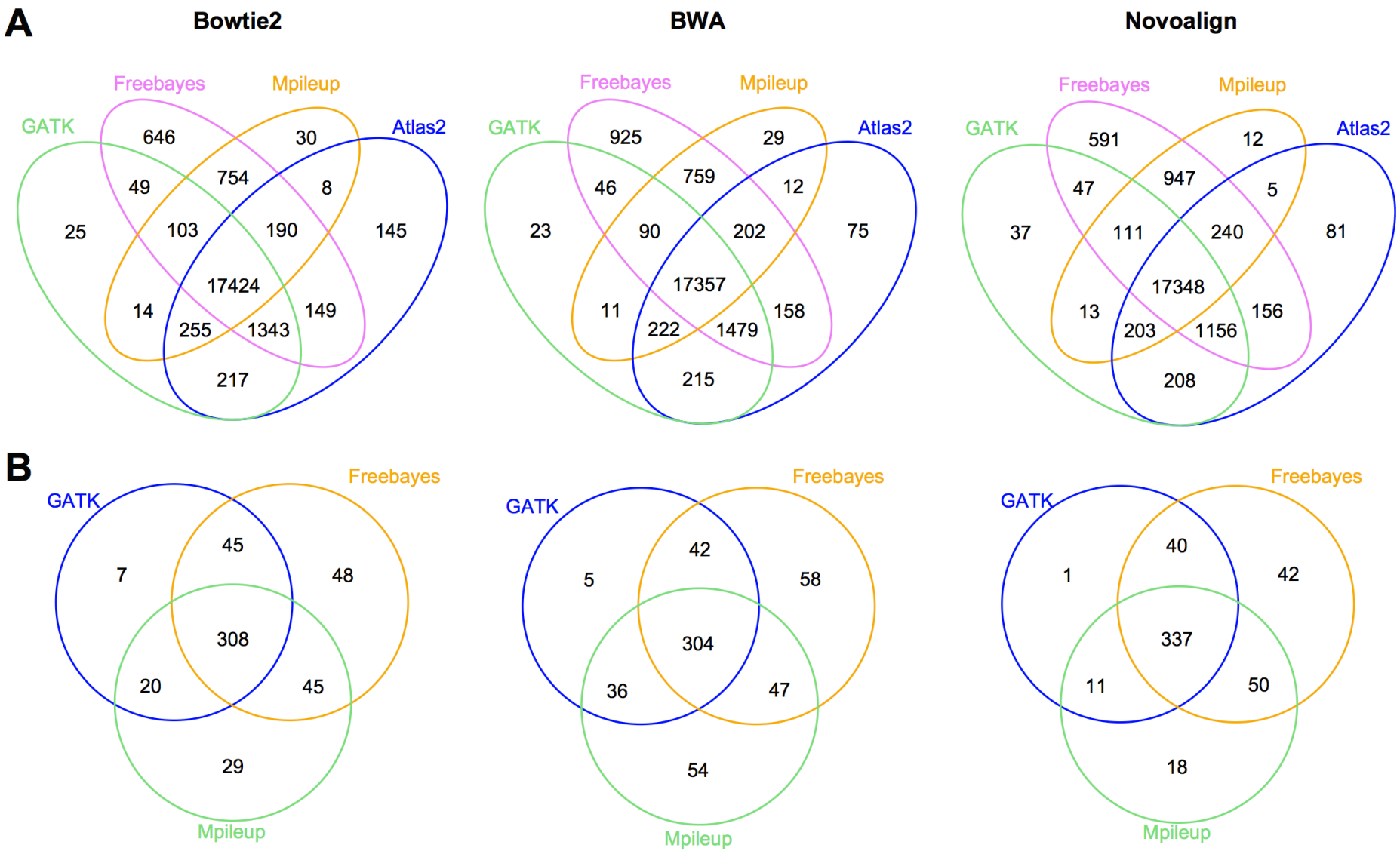
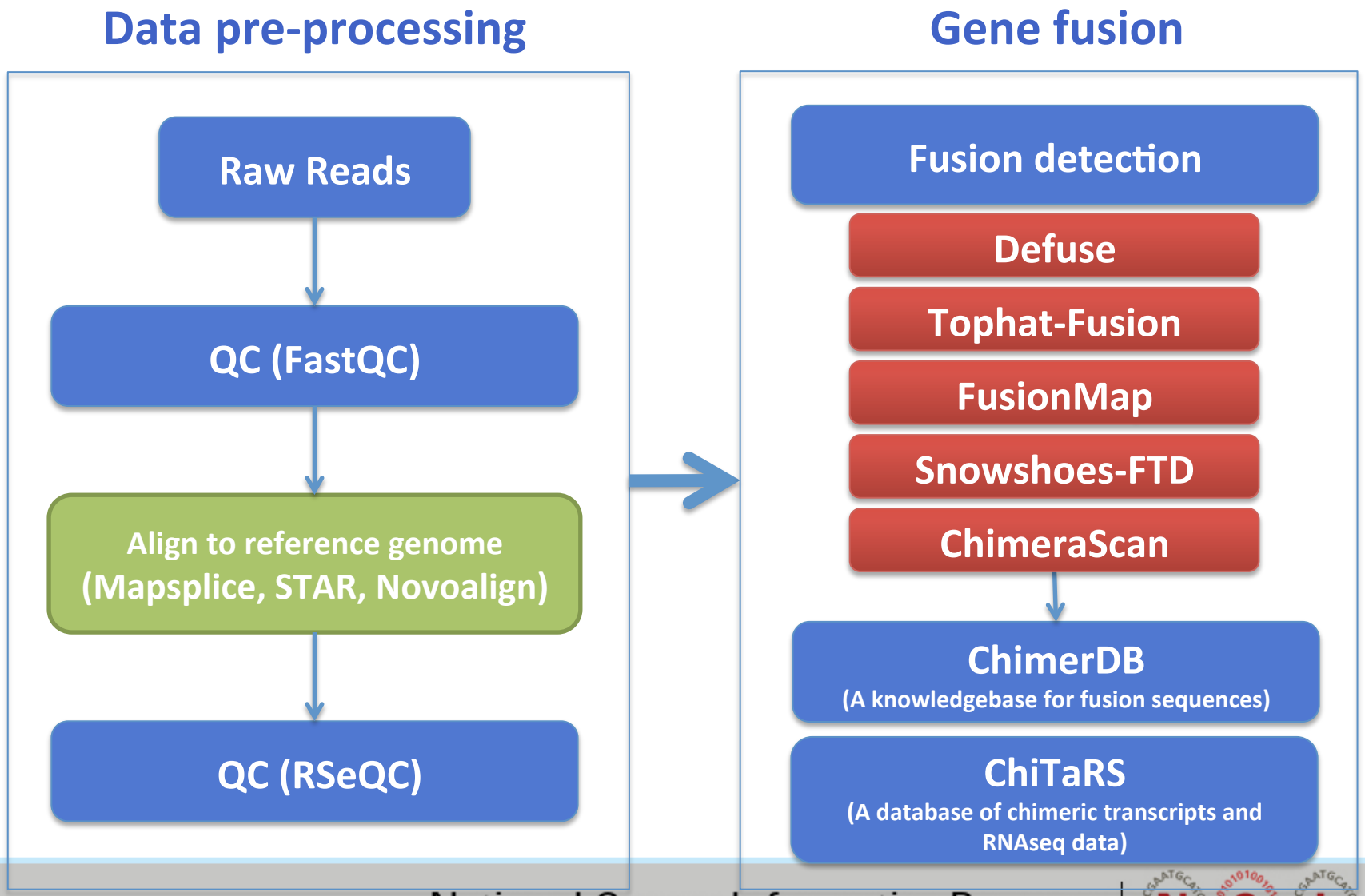
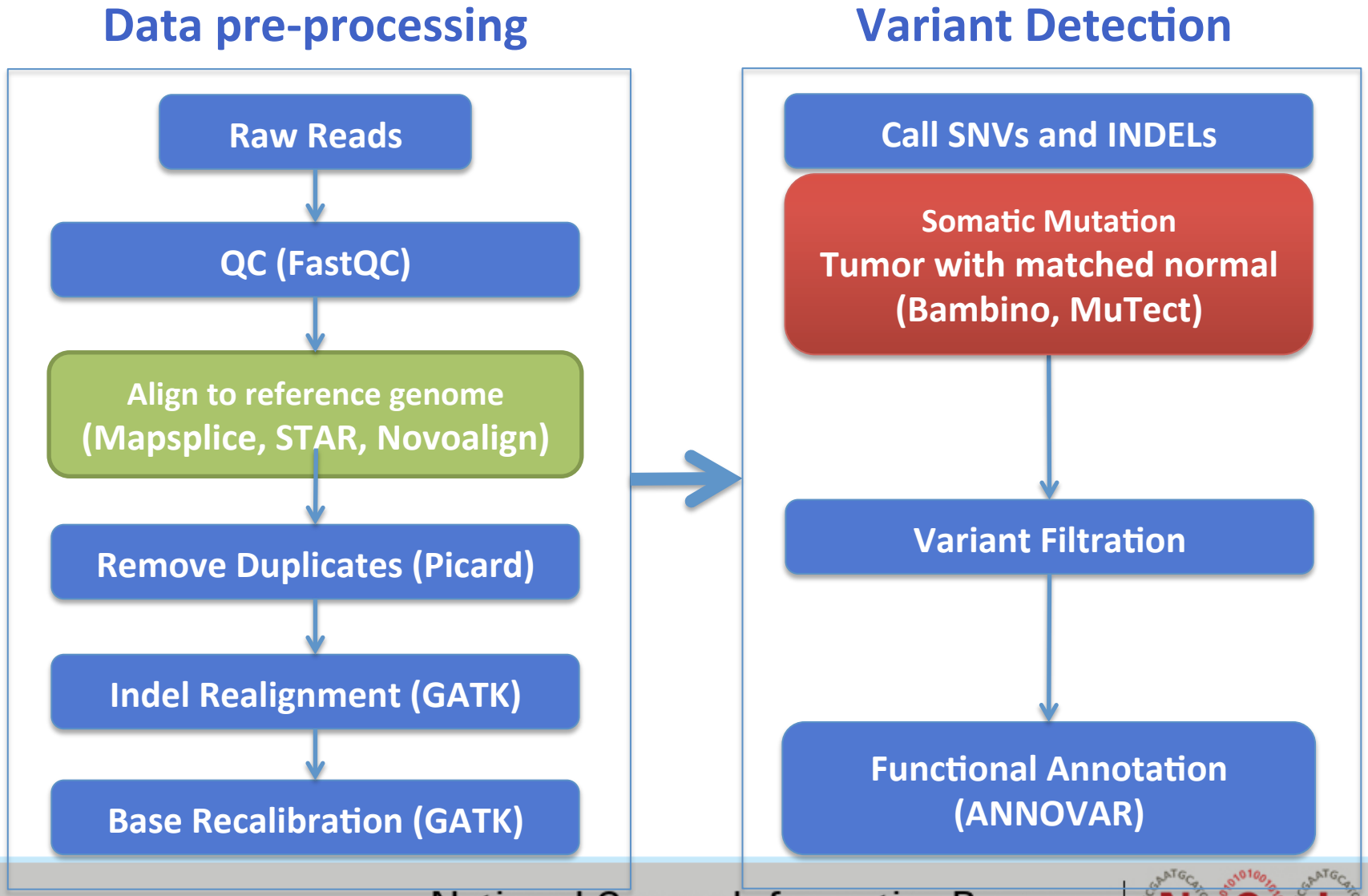


Figure 4. Counts of variants detected by four callers with alignment generated by three aligners as shown in Venn diagram of (**A**) SNVs and (**B**) indels.

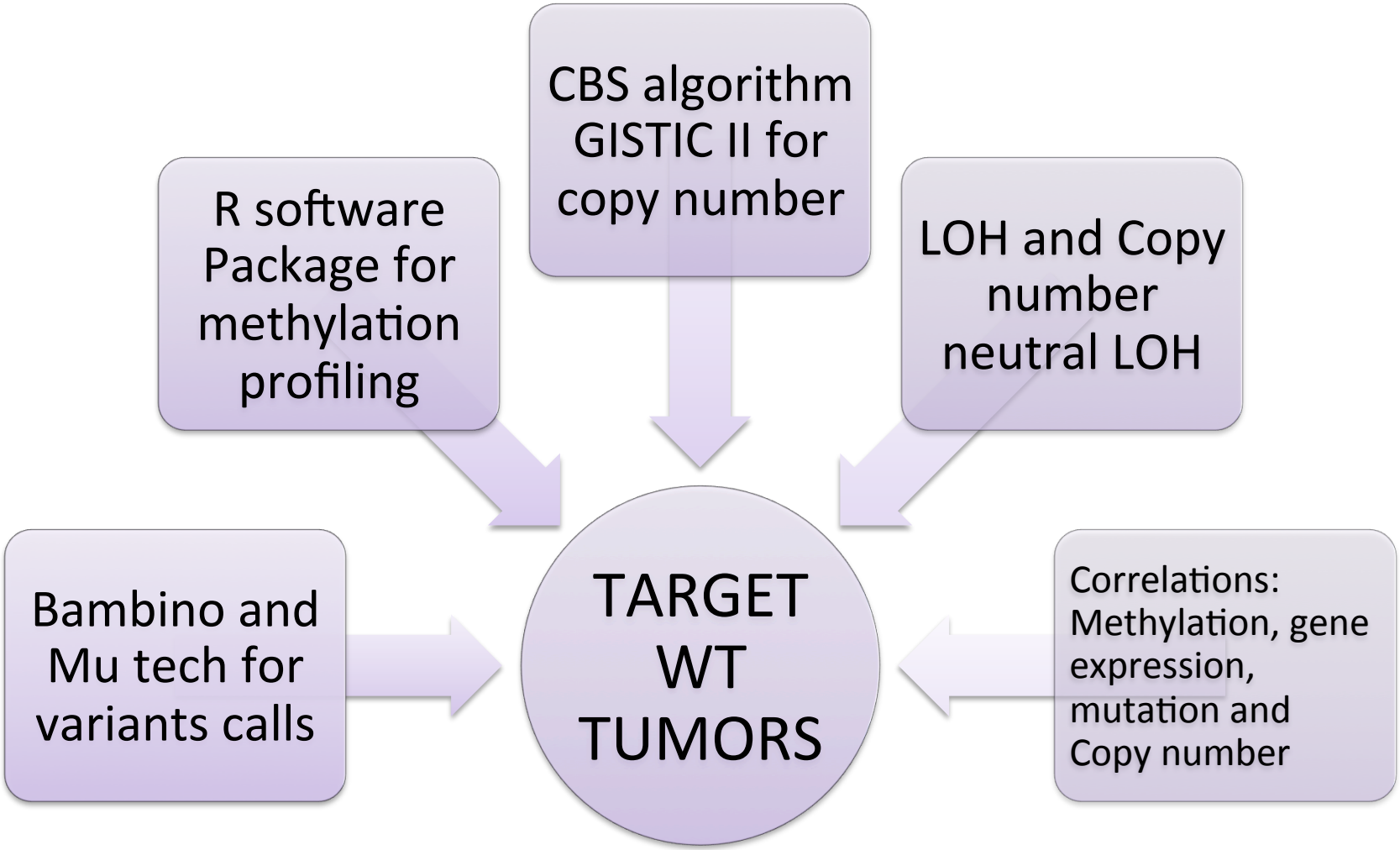
RNAseq: fusion analysis workflow



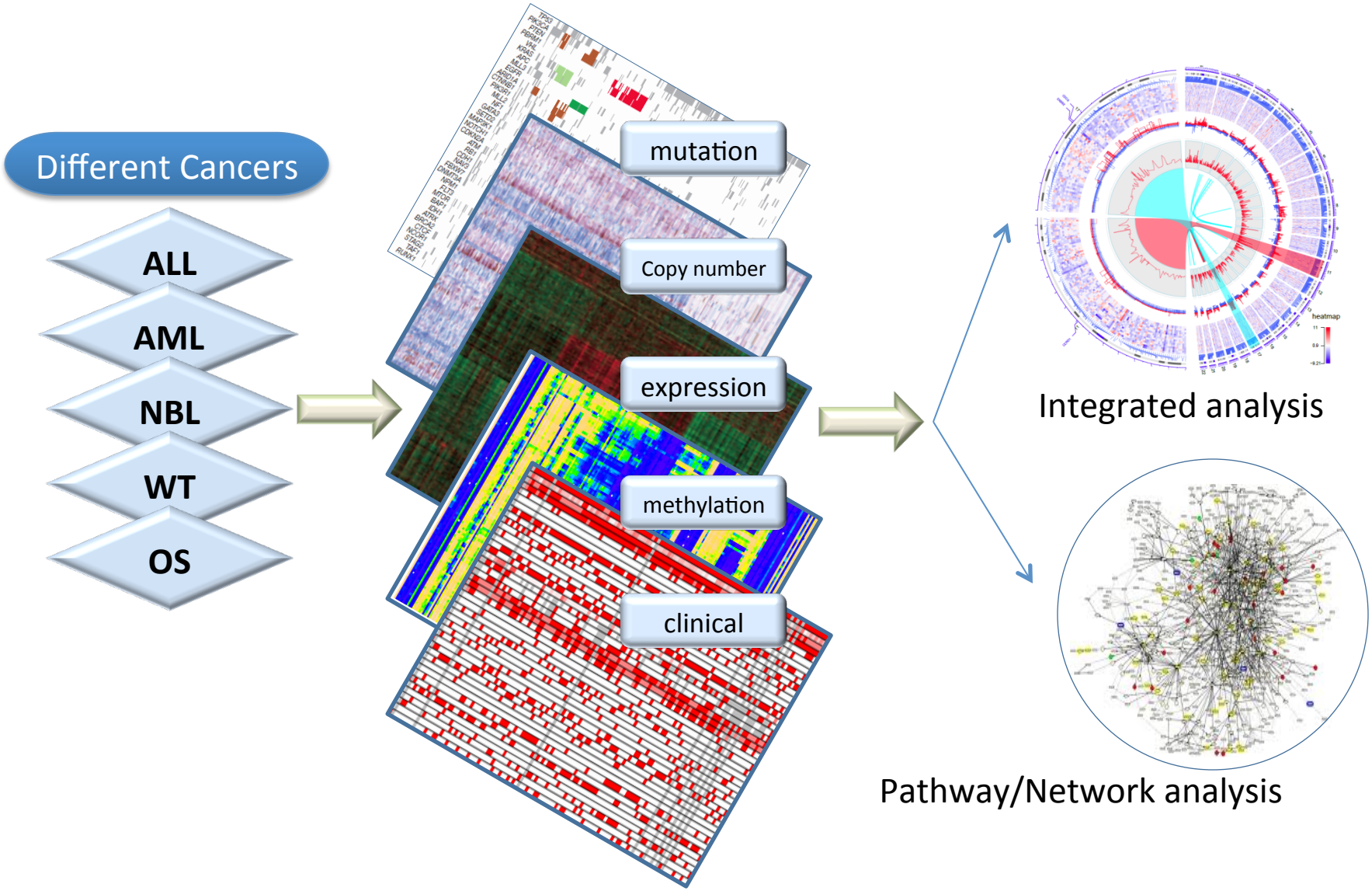
RNAseq: variant analysis workflow



TARGET Wilms Tumor (WT) Project



Integrated analysis



Institute of Medicine Report

Sept 10, 2013

Delivering High-Quality Cancer Care: Charting a New Course for System in Crisis

1. Understanding the outcomes of individual cancer patients as well as groups of similar patients

2. Capturing data from real-world settings that researchers can then analyze to generate new knowledge

3. A "Learning" healthcare IT system that learns routinely and iteratively by analyzing captured data, generating evidence, and implementing new insights into subsequent care.

"Learning IT System"

IOM Report on Cancer Care

1. **Search Prior Knowledge:** Enable clinicians to use previous patients' experiences to guide future care.

2. **Care Team Collaboration:** Facilitate a coordinated cancer care workforce & mechanisms for easily sharing information with each other.

3. **Cancer Research:** Improve the evidence base for quality cancer care by utilizing all of the data captured during real-world clinical encounters and integrating it with data captured from other sources.

What’s next?

1. Searching
2. Mining
3. Prediction

Can searching prior knowledge help future patients?

Can we make a Cinematch for cancer patients?

Netflix's **Cinematch** software analyzes each customer's film-viewing habits and recommends other movies.

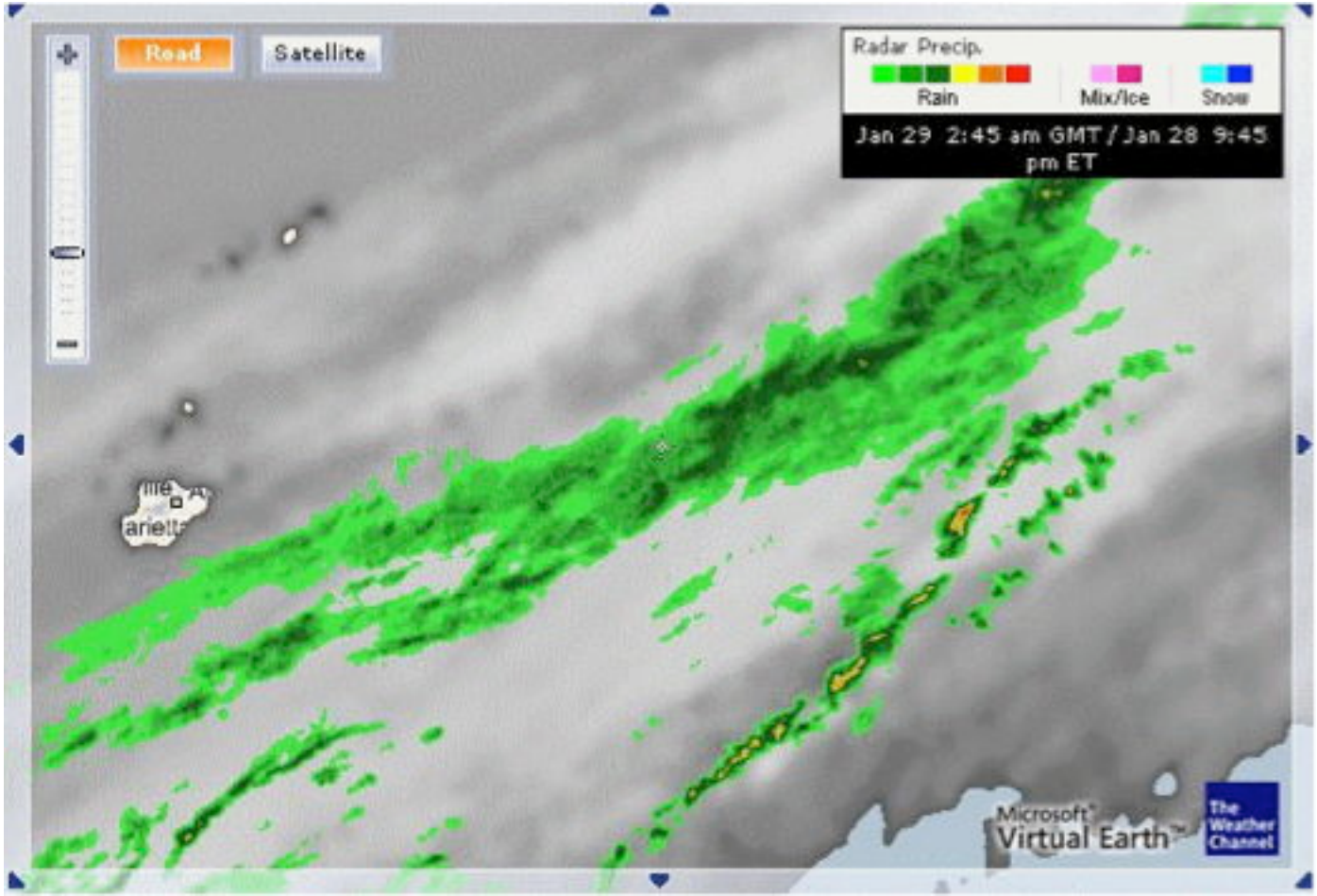
Patients like me

- Patients with diagnoses, symptoms and labs like yours are eligible for these trials…

- Patient-centered resources…

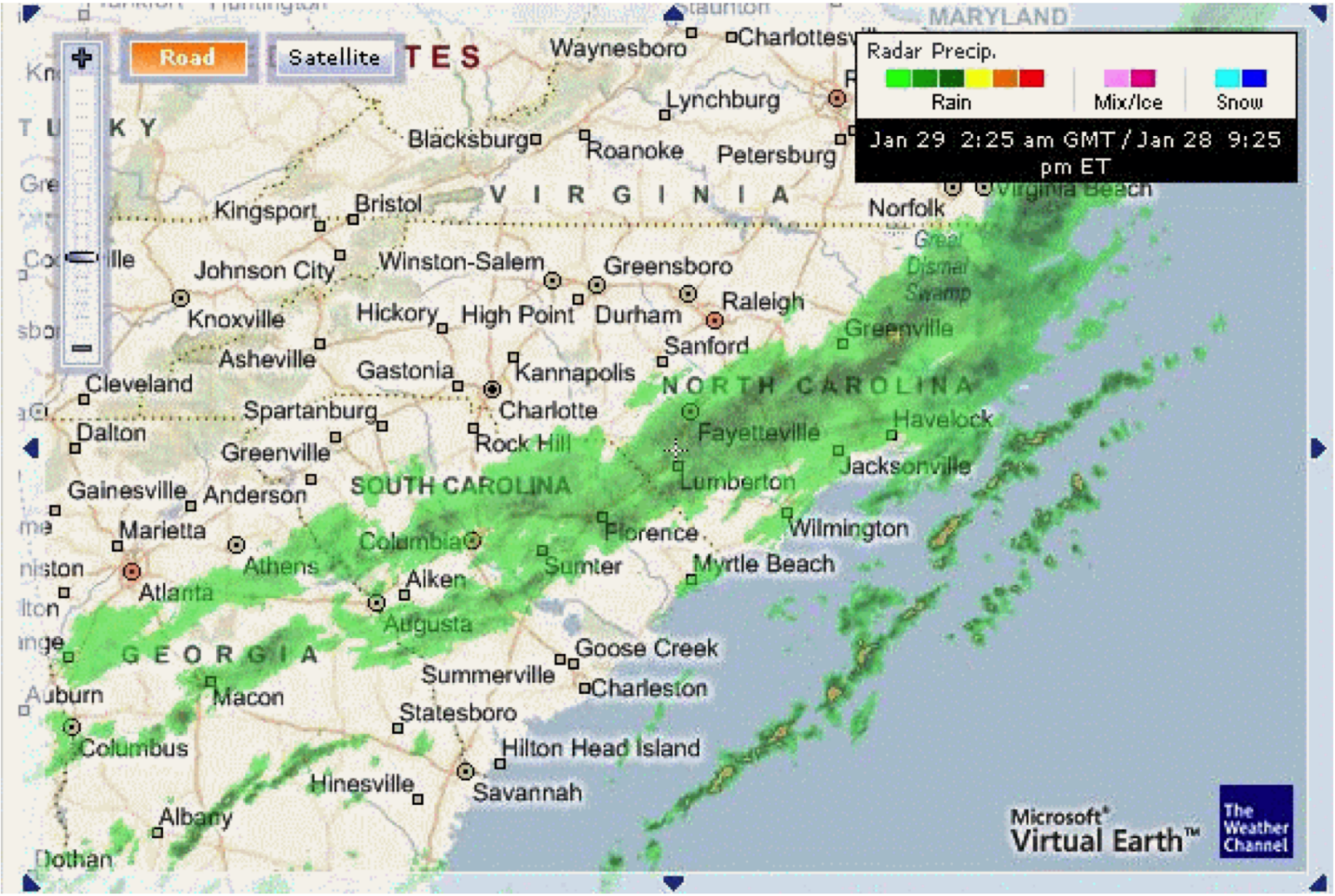
If we can forecast the weather, can we forecast cancer?

Where is the weather moving?



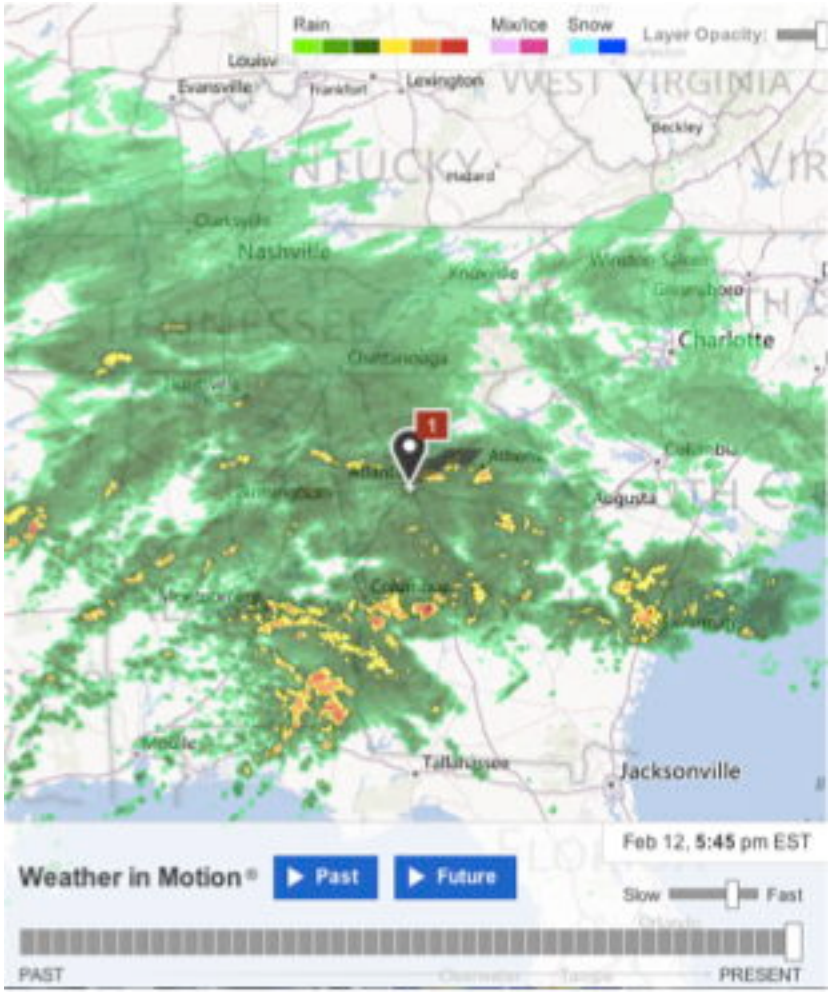
Doppler & Map Fusion

Animating the Weather

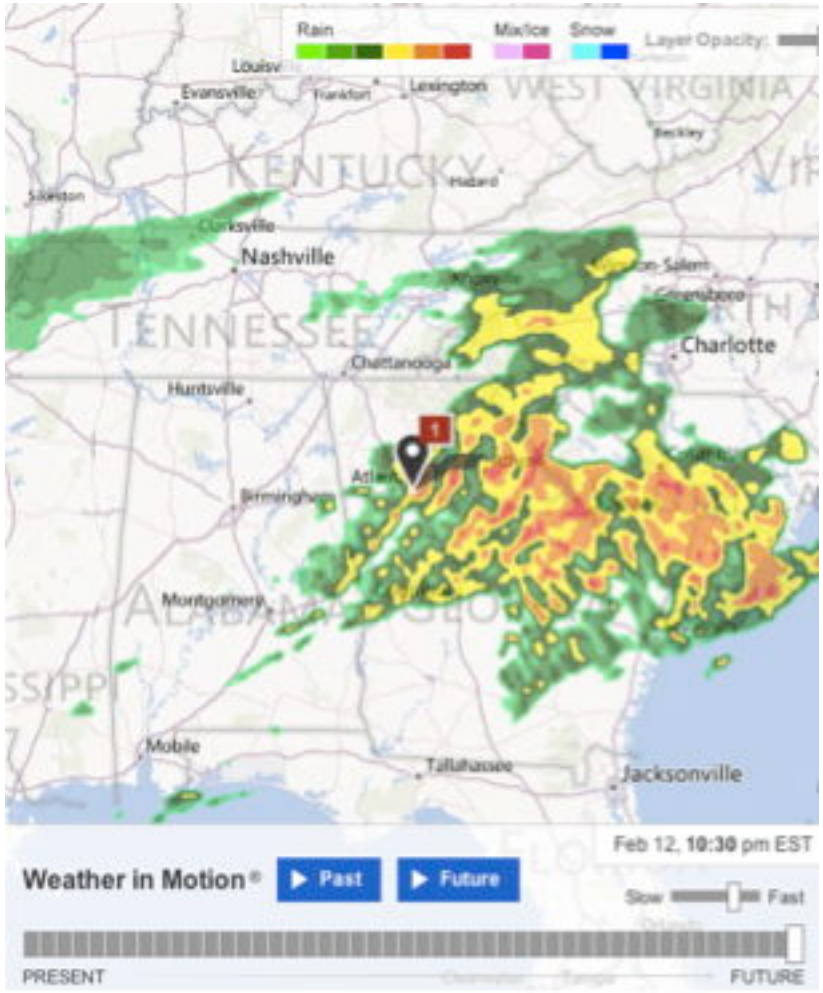


Dimension of time assists in decision making.

What about the future?



Present



5 Hours into Future

What changed?

1 **Equations**

2 **Satellite data**

3 **Computers**

Available online at www.sciencedirect.com

ScienceDirect

Journal of Computational Physics 227 (2008) 3431–3444

JOURNAL OF COMPUTATIONAL PHYSICS

www.elsevier.com/locate/jcp

The origins of computer weather prediction and climate modeling

Peter Lynch

Meteorology and Climate Centre, School of Mathematical Sciences, University College Dublin, Belfield, Ireland

Received 22 December 2006; received in revised form 26 February 2007; accepted 26 February 2007
Available online 19 March 2007

Abstract

Numerical simulation of an ever-increasing range of geophysical phenomena is adding enormously to our understanding of complex processes in the Earth system. The consequences for mankind of ongoing climate change will be far-reaching. Earth System Models are capable of replicating climate regimes of past millennia and are the best means we have of predicting the future of our climate.

The basic ideas of numerical forecasting and climate modeling were developed about a century ago, long before the first electronic computer was constructed. There were several major practical obstacles to be overcome before numerical prediction could be put into practice. A fuller understanding of atmospheric dynamics allowed the development of simplified systems of equations; regular radiosonde observations of the free atmosphere and, later, satellite data, provided the initial conditions; stable finite difference schemes were developed; and powerful electronic computers provided a practical means of carrying out the prodigious calculations required to predict the changes in the weather.

Progress in weather forecasting and in climate modeling over the past 50 years has been dramatic. In this presentation, we will trace the history of computer forecasting through the ENIAC integrations to the present day. The useful range of deterministic prediction is increasing by about one day each decade, and our understanding of climate change is growing rapidly as Earth System Models of ever-increasing sophistication are developed.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Numerical weather prediction; Climate modeling; History of NWP

1. Introduction

Among the most significant scientific advances of the past century is our ability to simulate complex physical systems using numerical models and thereby to predict their evolution. One outstanding example is the development of general circulation models (GCMs) of the atmosphere and oceans, which have brought two great advantages: We can now predict the weather for several days in advance with a high degree of

E-mail address: Peter.Lynch@ucd.ie
URL: http://maths.ucd.ie/~plynch

0021-9991/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2007.02.034

Newsweek
03.28.2014
SOLVING CANCER
YOU CAN'T CURE WHAT YOU DON'T UNDERSTAND
C
(X + Y = -C) (X + Y = -C) (X + Y = -C) (X + Y = -C)

Modeling Tumor Growth

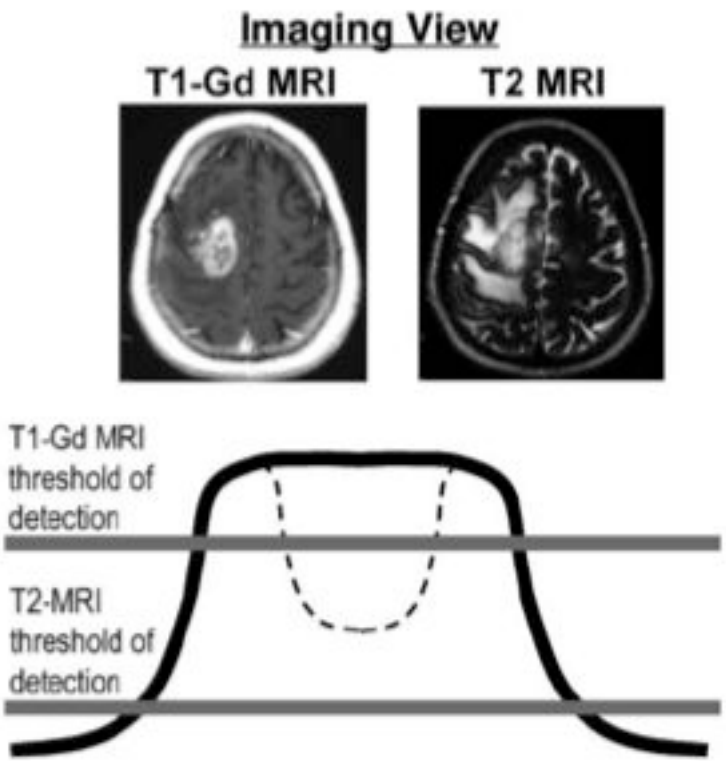
Mathematical model: proliferation of cells with the potential for invasion and metastasis

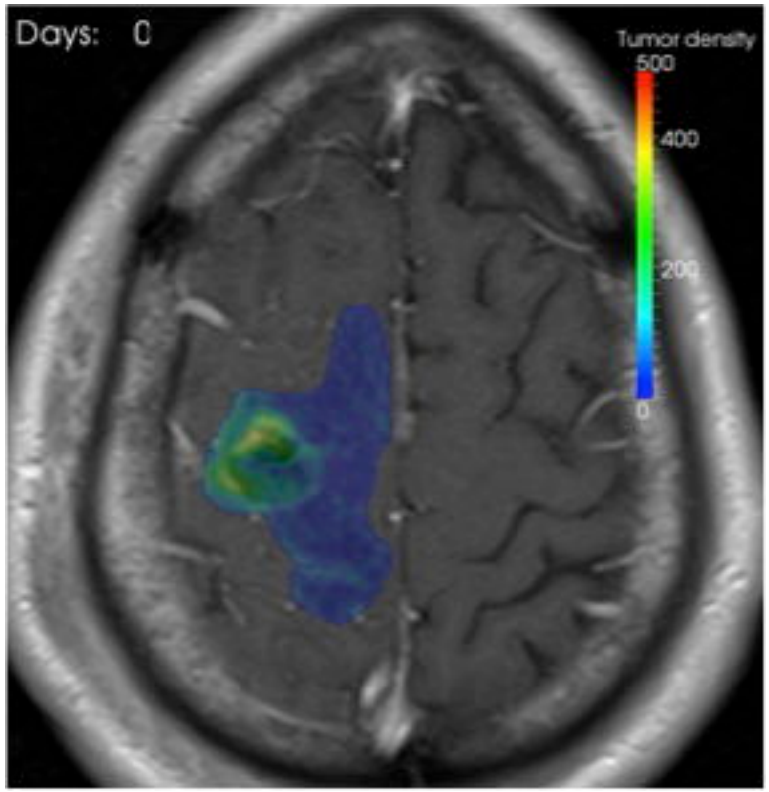
$$\overbrace{\frac{\partial c}{\partial t}}^{\text{rate of change of glioma cell concentration}} = \overbrace{\nabla \cdot (D(x)\nabla c)}^{\text{net dispersal of glioma cells}} + \overbrace{\rho c}^{\text{net proliferation of glioma cells}} - \overbrace{R(x,t)c}^{\text{loss due to therapy}}$$

Swanson et al., British Journal of Cancer, 2007: 1-7.

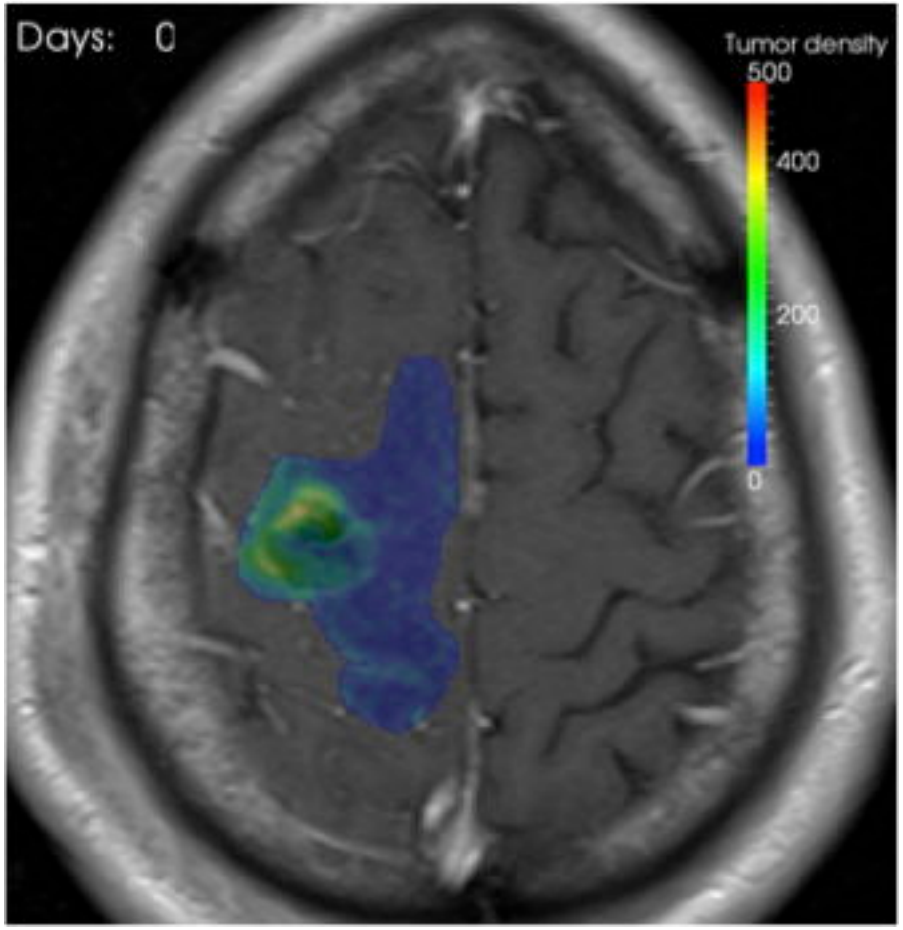
Personalized Tumor Model

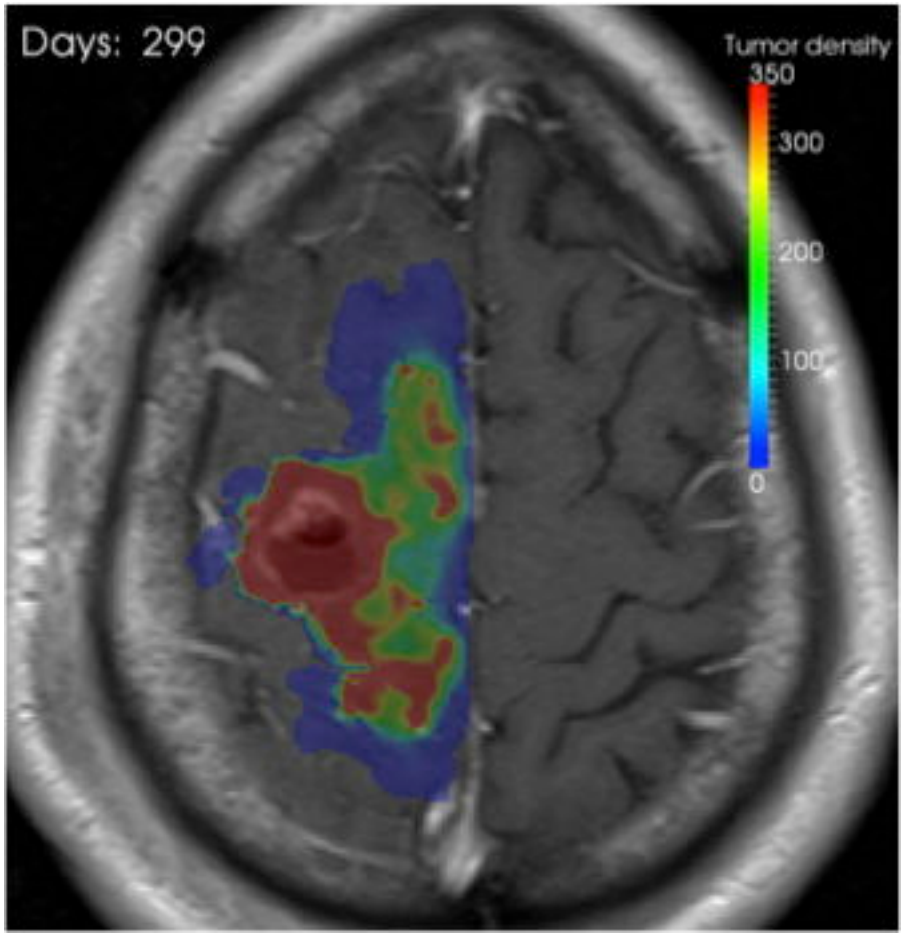
Imaging used to seed the model





Personalized Tumor Model





Today **Future**

Radiation Treatment Effects

$$\overbrace{\frac{\partial c}{\partial t}}^{\text{rate of change of glioma cell concentration}} = \overbrace{\nabla \cdot (D(x)\nabla c)}^{\text{net dispersal of glioma cells}} + \overbrace{\rho c}^{\text{net proliferation of glioma cells}} - \overbrace{R(x,t)c}^{\text{loss due to therapy}} ,$$

New term defines cell killing

$$R(\alpha, d(x,t)) = \begin{cases} 0, & \text{for } t \notin \text{therapy} \\ 1 - S(\alpha, d(x,t)), & \text{for } t \in \text{therapy.} \end{cases}$$

$$S(\alpha, d(x,t)) = e^{-\alpha BED}.$$

$$\text{BED} = nd\left(1 + \frac{d}{\alpha/\beta}\right),$$

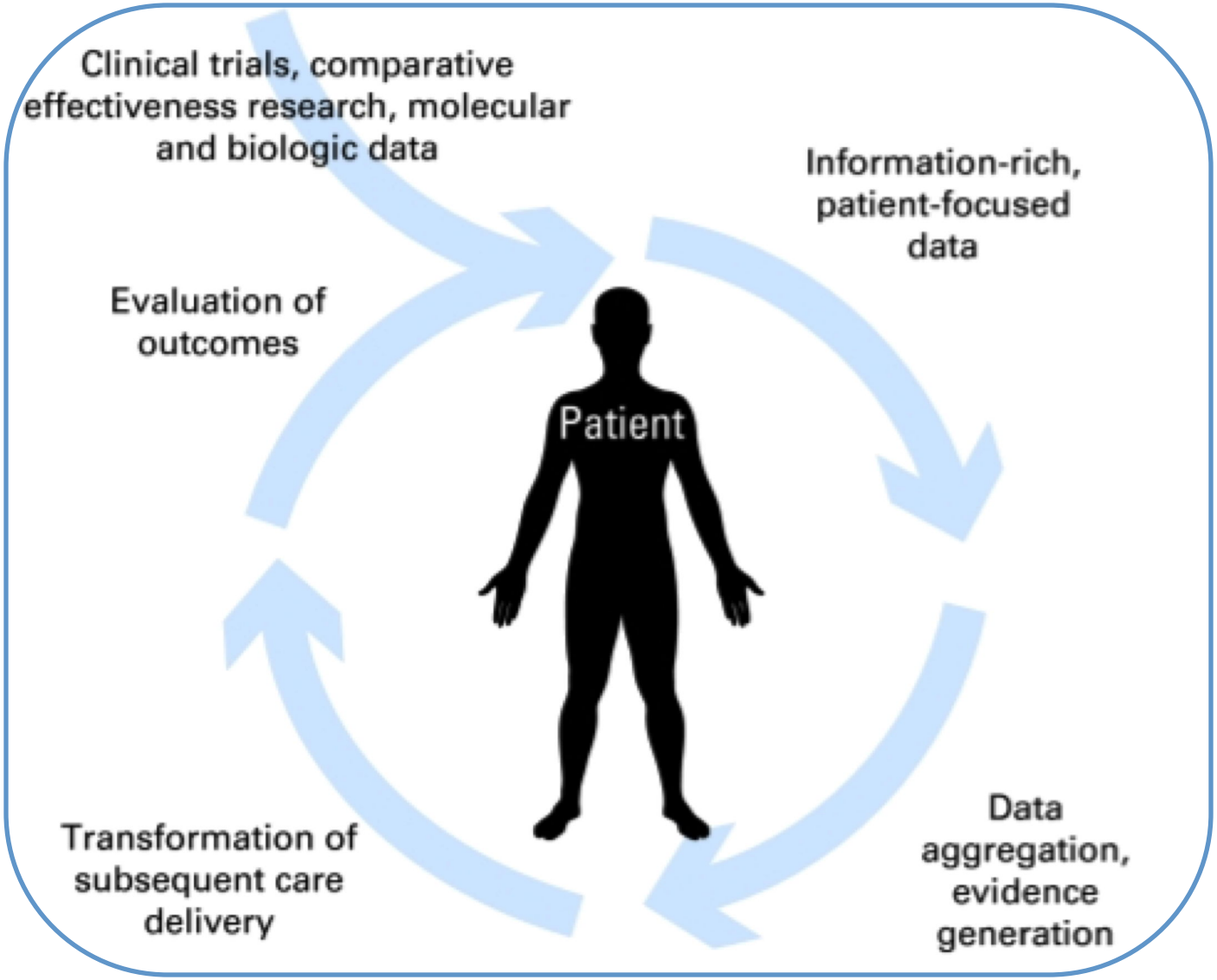
L-Q model used to describe cell killing

Population Decision Support

Rapid Learning Systems

Patient-level data are aggregated to achieve population-based change, and results are applied to care of individual patients.

Predict outcomes



Precision Oncology

- The era of precision medicine and precision oncology is *predicated* on the integration of research, care, and molecular medicine and the *availability of data* for modeling, risk analysis, and optimal care

How do we re-engineer translational research policies that will enable a true ***learning healthcare system****?*

The future

- Elastic computing ‘clouds’
- Social networks
- Big Data analytics

Semantic and synoptic data

- Precision medicine
- Measuring health
- Practicing protective medicine

Intervening **before** health is compromised

Learning systems that enable learning from **every** cancer patient

Thank you

Warren A. Kibbe
warren.kibbe@nih.gov

NATIONAL® CANCER INSTITUTE