Mining Large-scale Cancer Genomics Data Using Cloud-based Bioinformatics Approaches

Jorge Andrade, PhD

Riyue Bao, PhD

Kyle Hernandez, PhD

Samuel Volchenboum, MD, PhD

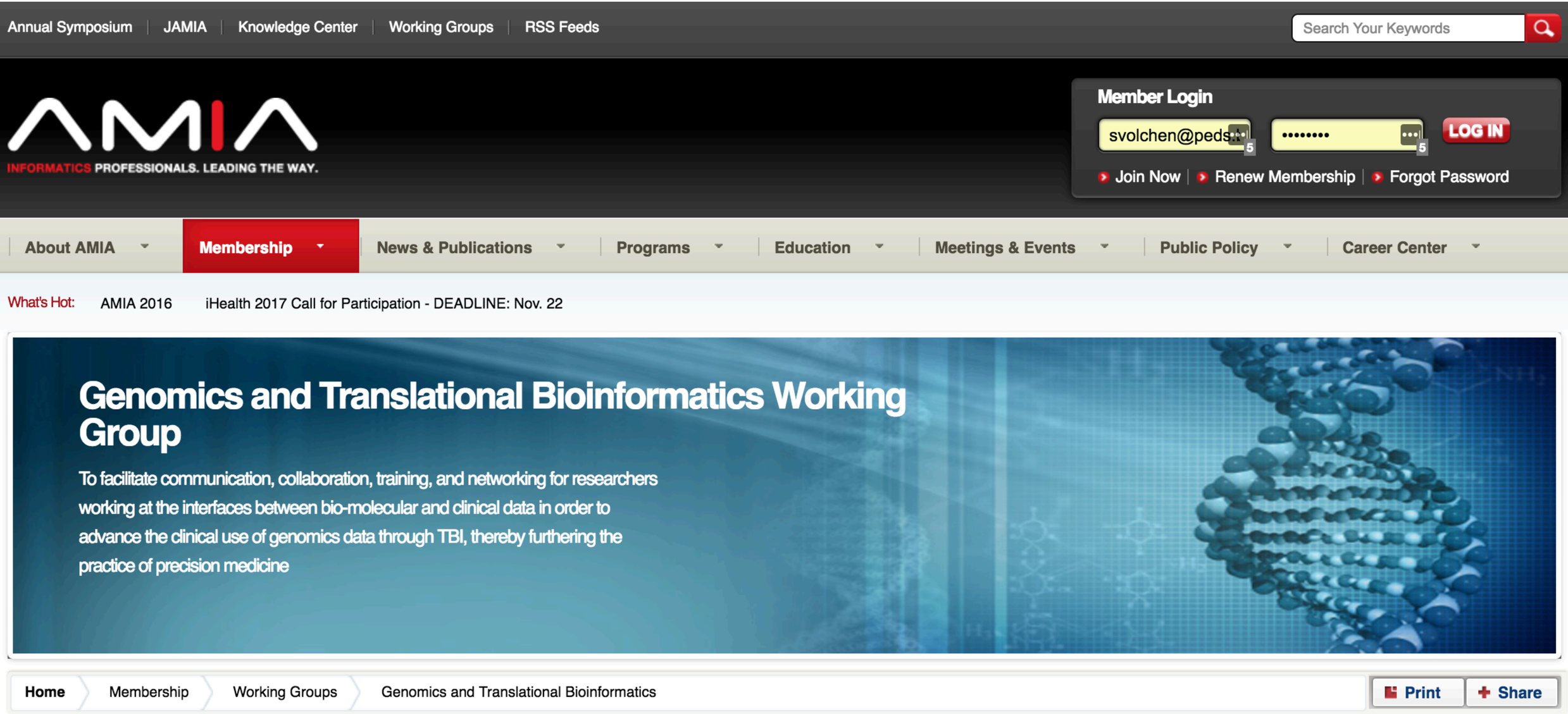
Center for Research Informatics

University of Chicago

November 13, 2016

THE UNIVERSITY OF CHICAGO

AMIA - Genomics and Translational Bioinformatics Working Group




2100 Faculty
7000 Graduate students
160,000 Alumni
3 National Laboratories
Argonne, Fermi, MBL

The University of Chicago Medical Center

The University of Chicago Medical Center

Adjacent to UChicago and Pritzker Medical School

- 570 beds
- 25,000 admits
- 492,000 outpatient visits
- 78,000 ED visits

Physicians employed by the University

One Dean for the hospital, medical school, and biological sciences division

The Setting

University of Chicago

The University of Chicago is the corporate parent and sole member of the UCMC

University of Chicago Medicine and Biological Sciences

Pritzker School of Medicine	Biological Sciences Division	University of Chicago Medical Center
Biological Sciences Division faculty teach in the Medical School	10 Basic Science Departments 13 Clinical Departments 18 PhD Programs / 1 MS	Biological Sciences Division faculty treat patients in the Medical Center


University of Chicago
Center for Research Informatics

Center for research informatics

Applications

Enterprise software

Sample tracking

Patient registries

Shared clinical data

Custom websites

REDCap support

Multi-site trials

PCORI integration

Data Warehouse

Clinical data from 2008

Retrospective studies

Quality measures

i2b2 cohort discovery

Self-service data mart

NLP over clinical notes

Data aggregation

System modeling

Systems

2 PB of Isilon storage

Backup, 30-day retention

Hosted servers (VMs)

4000-core HPC

Custom solutions

Galaxy server

Commercial analytics

Data visualization

Bioinformatics

8 PhD Bioinformaticians

Machine learning experts

Industry-grade pipelines

Custom workflows

Grant preparation

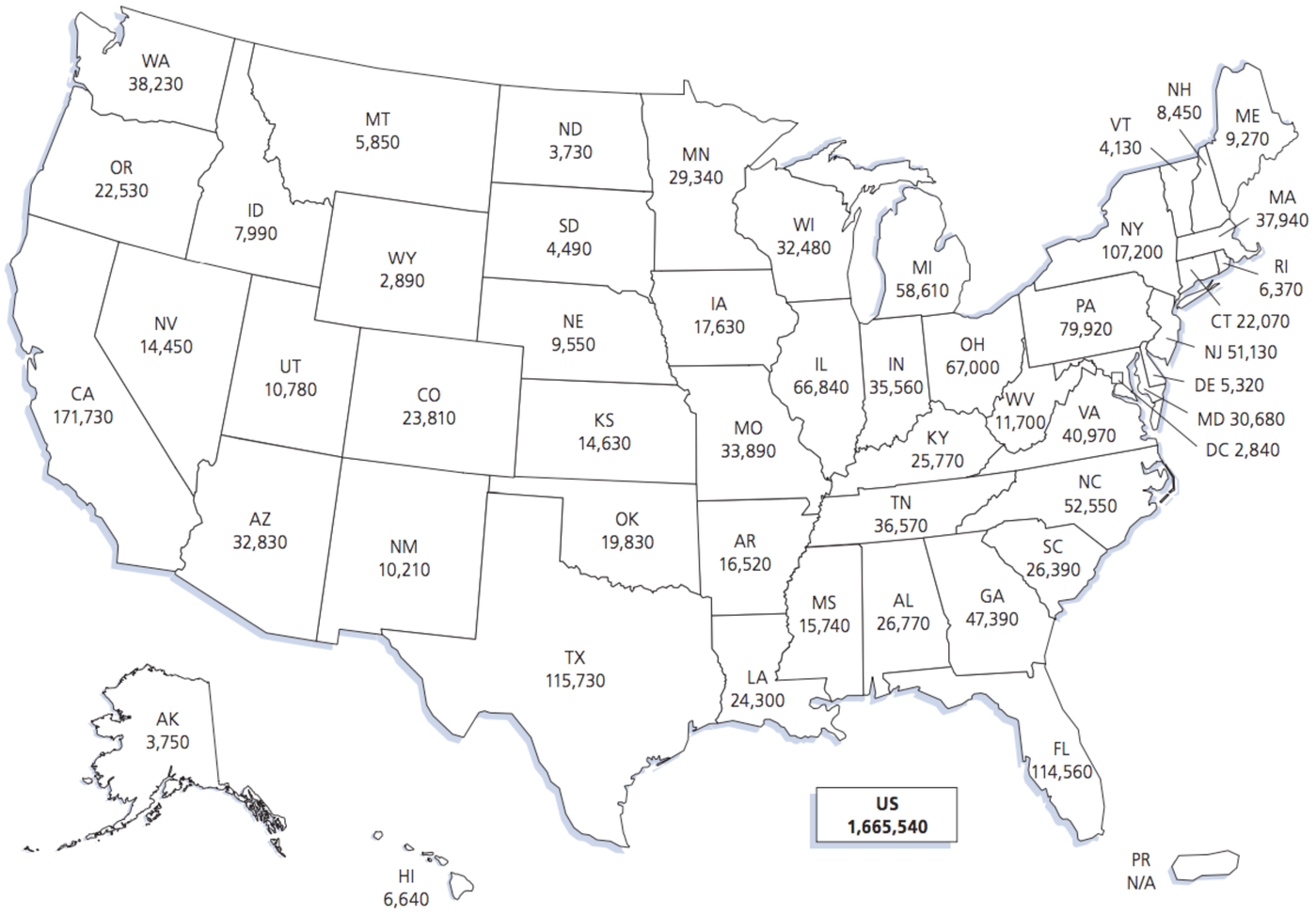
Manuscript writing

Multi-omic integration

Training and education

Cancer Statistics

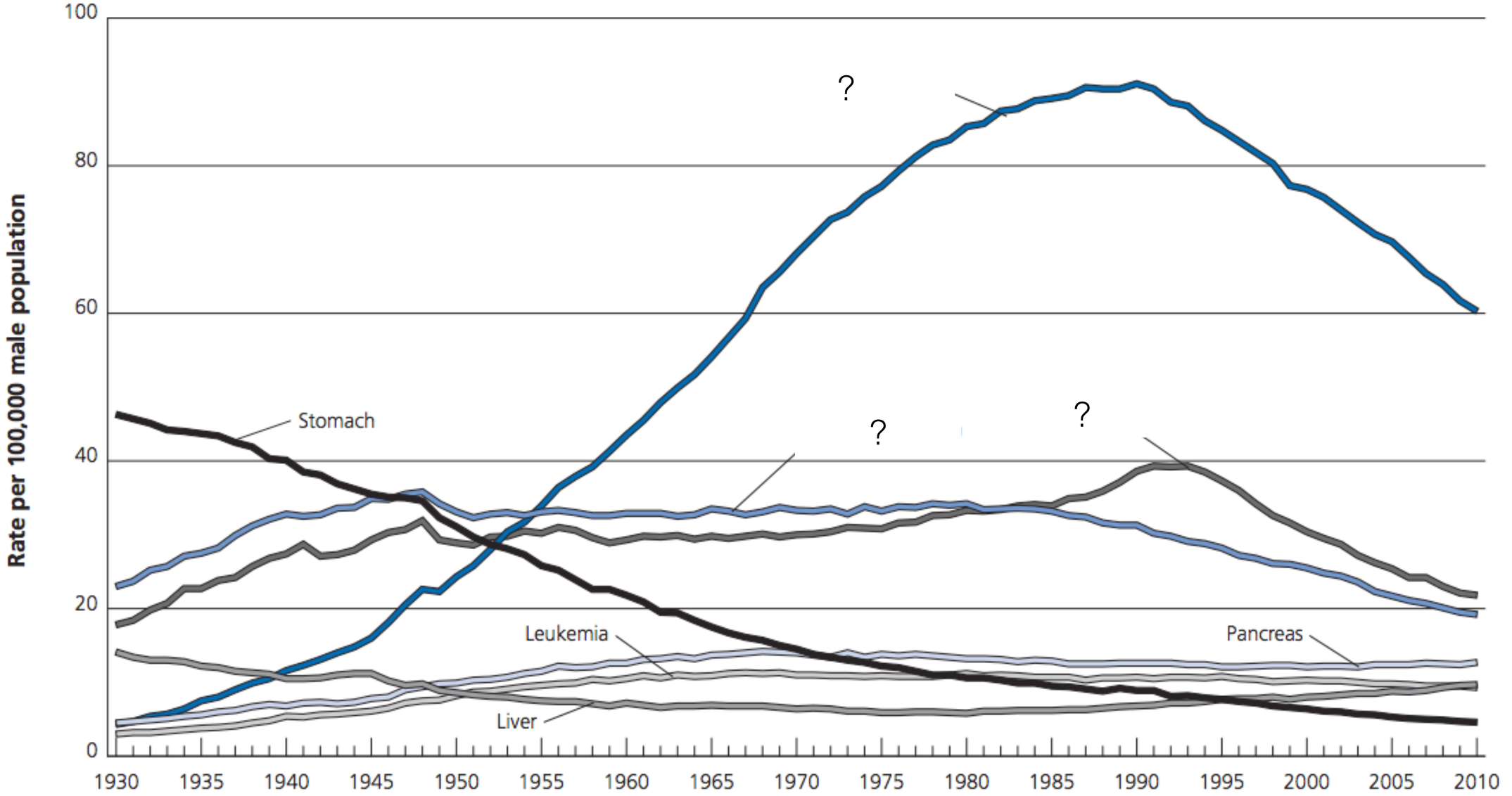
Many cases of cancer are preventable



- In 2016, over 175,000 of the ~600,000 cancer deaths will be caused by tobacco use
- Up to one third of cancers are related to obesity / poor nutrition
- Many of the 3 million cases of skin cancer are preventable
- Overall survival is around 70%, up from 50% in the 1970's

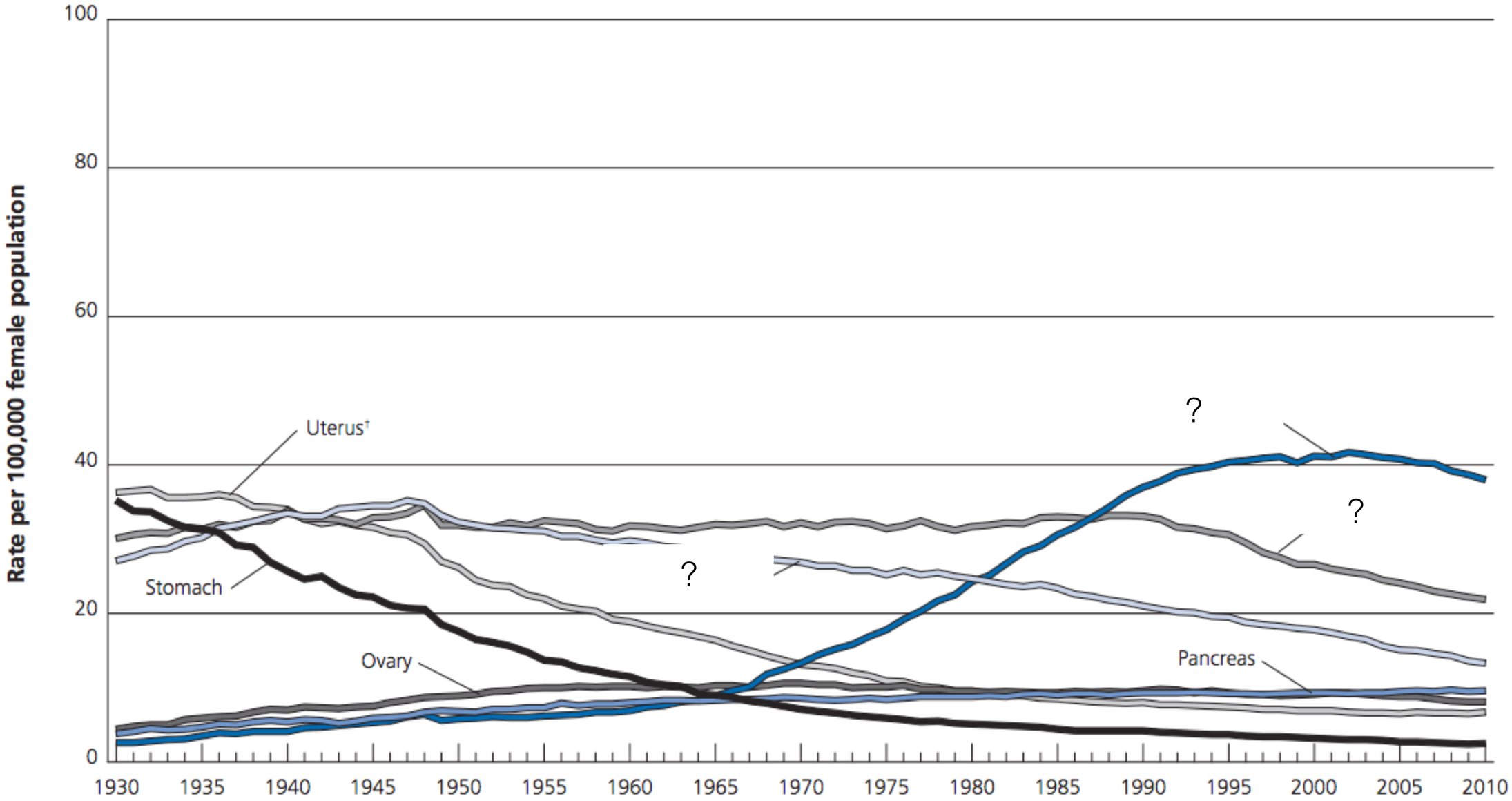
http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

Age-adjusted Cancer Death Rates*, Males by Site, US, 1930-2010



http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

Age-adjusted Cancer Death Rates*, Females by Site, US, 1930-2010



http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

Five-year Relative Survival Rates* (%) by Stage at Diagnosis, 2003-2009

	All Stages	Local	Regional	Distant
Breast (female)	89	99	84	24
Colon & rectum	65	90	70	13
Esophagus	17	39	21	4
Kidney†	72	92	64	12
Larynx	61	76	43	35
Liver‡	16	29	10	3
Lung & bronchus	17	54	26	4
Melanoma of the skin	91	98	62	16
Oral cavity & pharynx	62	83	59	36
Ovary	44	92	72	27
Pancreas	6	24	9	2
Prostate	99	100	100	28
Stomach	28	63	28	4
Testis	95	99	96	74
Thyroid	98	100	97	55
Urinary bladder§	78	70	33	5
Uterine cervix	68	91	57	16
Uterine corpus	82	95	68	17

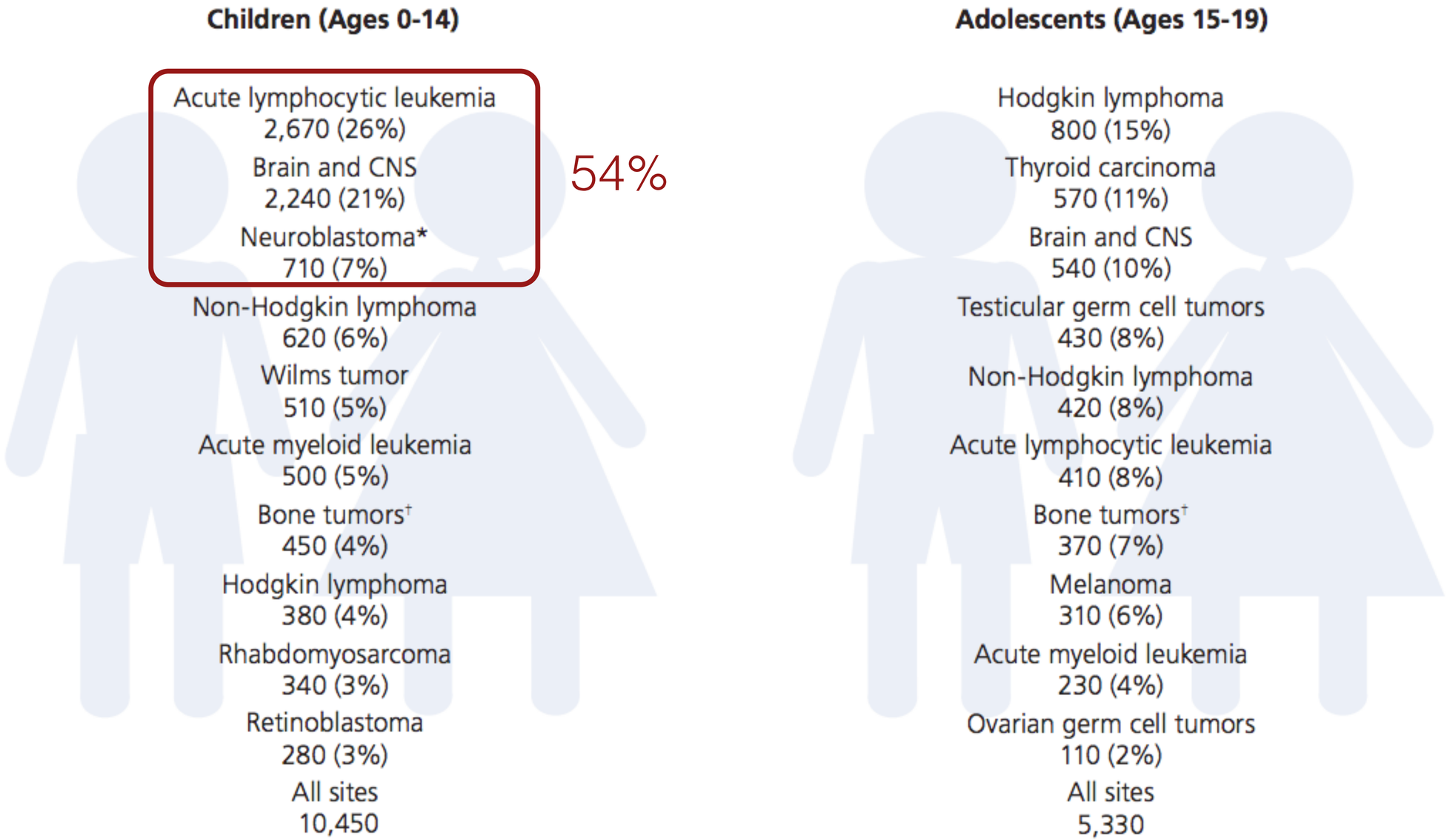
*Rates are adjusted for normal life expectancy and are based on cases diagnosed in the SEER 18 areas from 2003-2009, all followed through 2010.
†Includes renal pelvis. ‡Includes intrahepatic bile duct. §Rate for in situ cases is 96%.

Local: an invasive malignant cancer confined entirely to the organ of origin. **Regional:** a malignant cancer that 1) has extended beyond the limits of the organ of origin directly into surrounding organs or tissues; 2) involves regional lymph nodes; or 3) has both regional extension and involvement of regional lymph nodes.
Distant: a malignant cancer that has spread to parts of the body remote from the primary tumor either by direct extension or by discontinuous metastasis to distant organs, tissues, or via the lymphatic system to distant lymph nodes.

Source: Howlader N, Noone AM, Krapcho M, et al. (eds). *SEER Cancer Statistics Review, 1975-2010*, National Cancer Institute, Bethesda, MD http://seer.cancer.gov/csr/1975_2010/, based on November 2012 SEER data submission, posted to the SEER Web site, April 2013.

American Cancer Society, Surveillance Research 2014

Figure 1. Estimated Cases for Childhood and Adolescent Cancers, US, 2014



http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

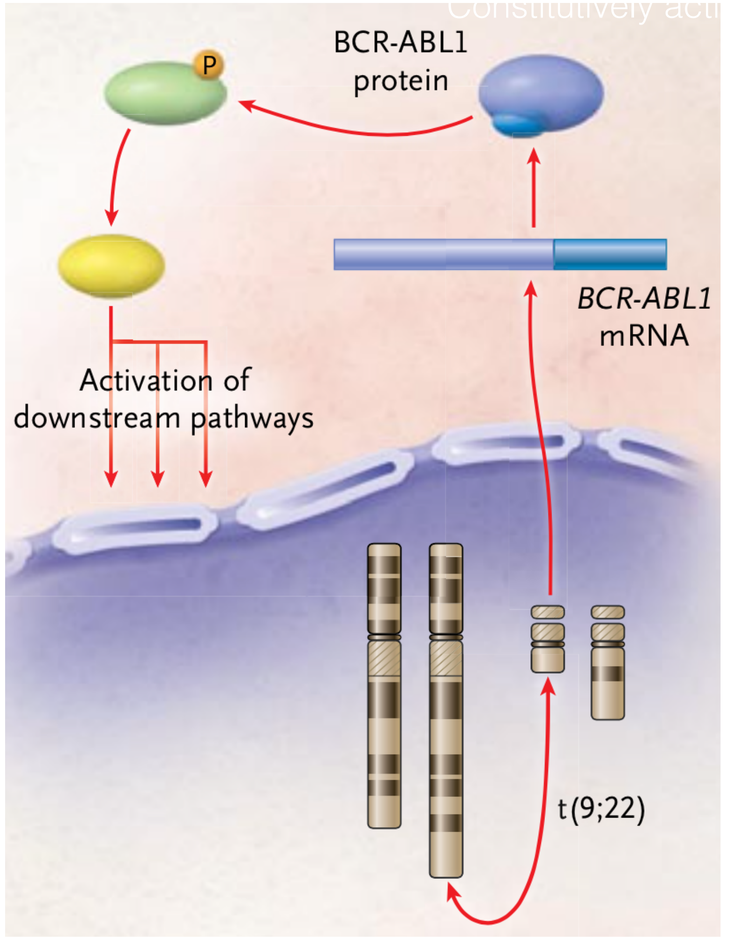
Table 1. Incidence, Mortality, and Survival Rates for Childhood and Adolescent Cancers by Sex and Race/Ethnicity

	Ages 0–14			Ages 15–19		
Characteristic	**Incidence, 2006-2010***	**Mortality, 2006-2010***	**Observed Survival (%), 2003-2009**	**Incidence, 2006-2010***	**Mortality, 2006-2010***	**Observed Survival (%), 2003-2009**
Sex						
Boys	178.0	23.3	81.3	237.7	34.5	80.0
Girls	160.1	21.1	82.0	235.5	24.7	85.4
Race/ethnicity						
Non-Hispanic White	178.2	22.4	84.2	259.4	29.0	85.9
Non-Hispanic Black	134.5	21.9	75.3	171.9	30.6	76.8
Hispanic	167.3	22.6	80.3	220.7	32.4	75.8
Asian American/ Pacific Islander	131.9	19.1	78.3	167.8	25.6	80.4
American Indian/ Alaska Native†	117.1	15.8	78.5	200.1	24.0	77.3

http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014/index

Implementation of PCM

- Genetic aberrations drive malignancies
- Highly specific anticancer agents modulate these targets

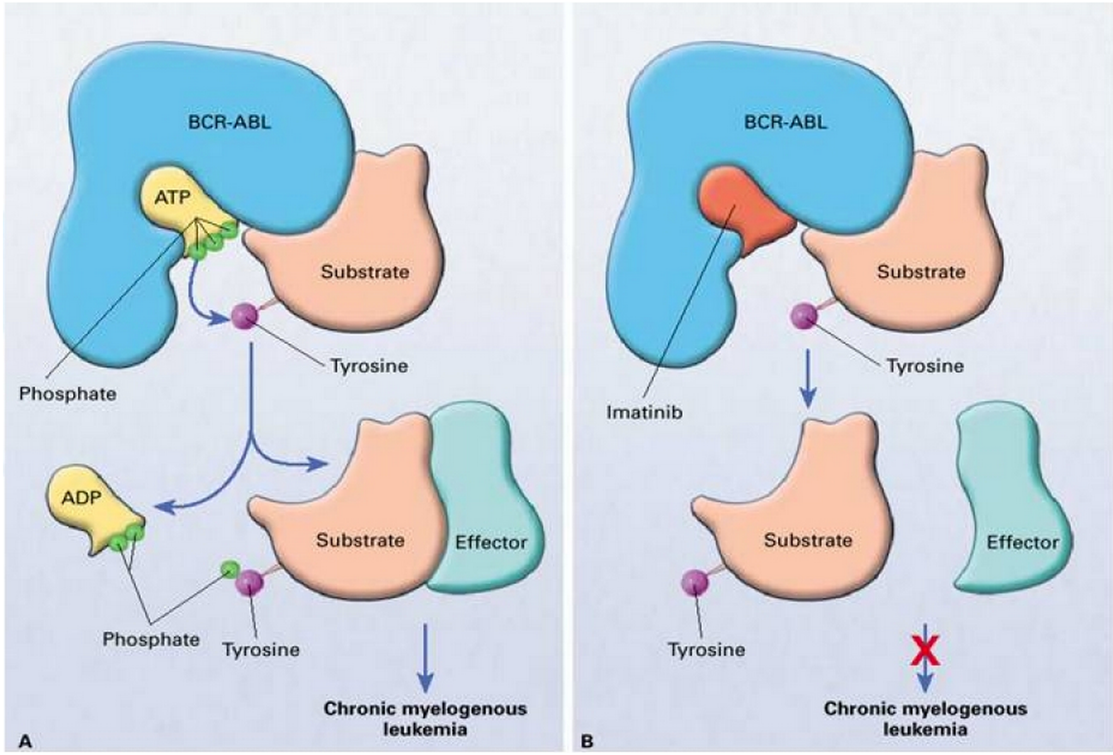


NEJM 359;7

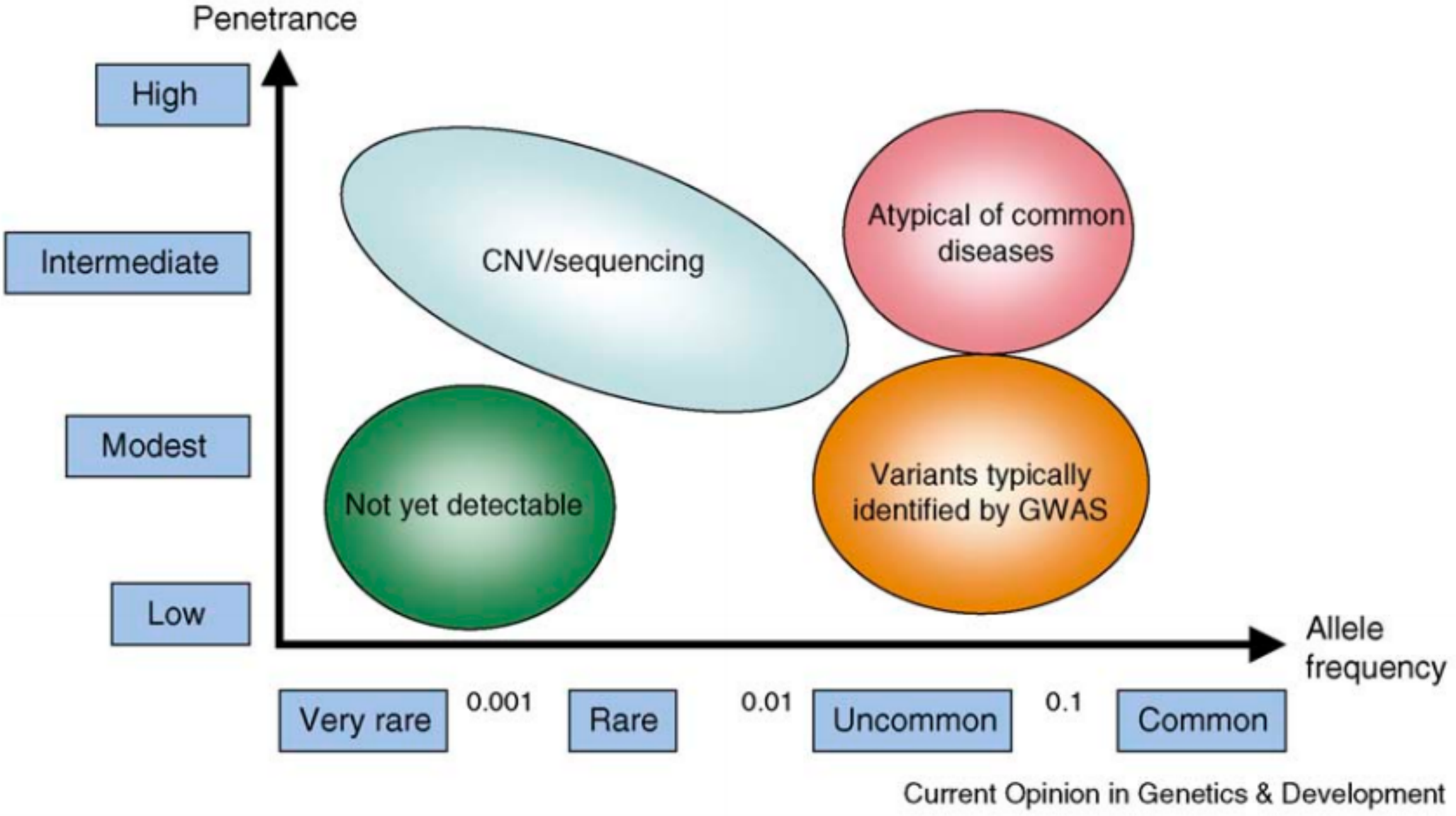
Implementation of PCM

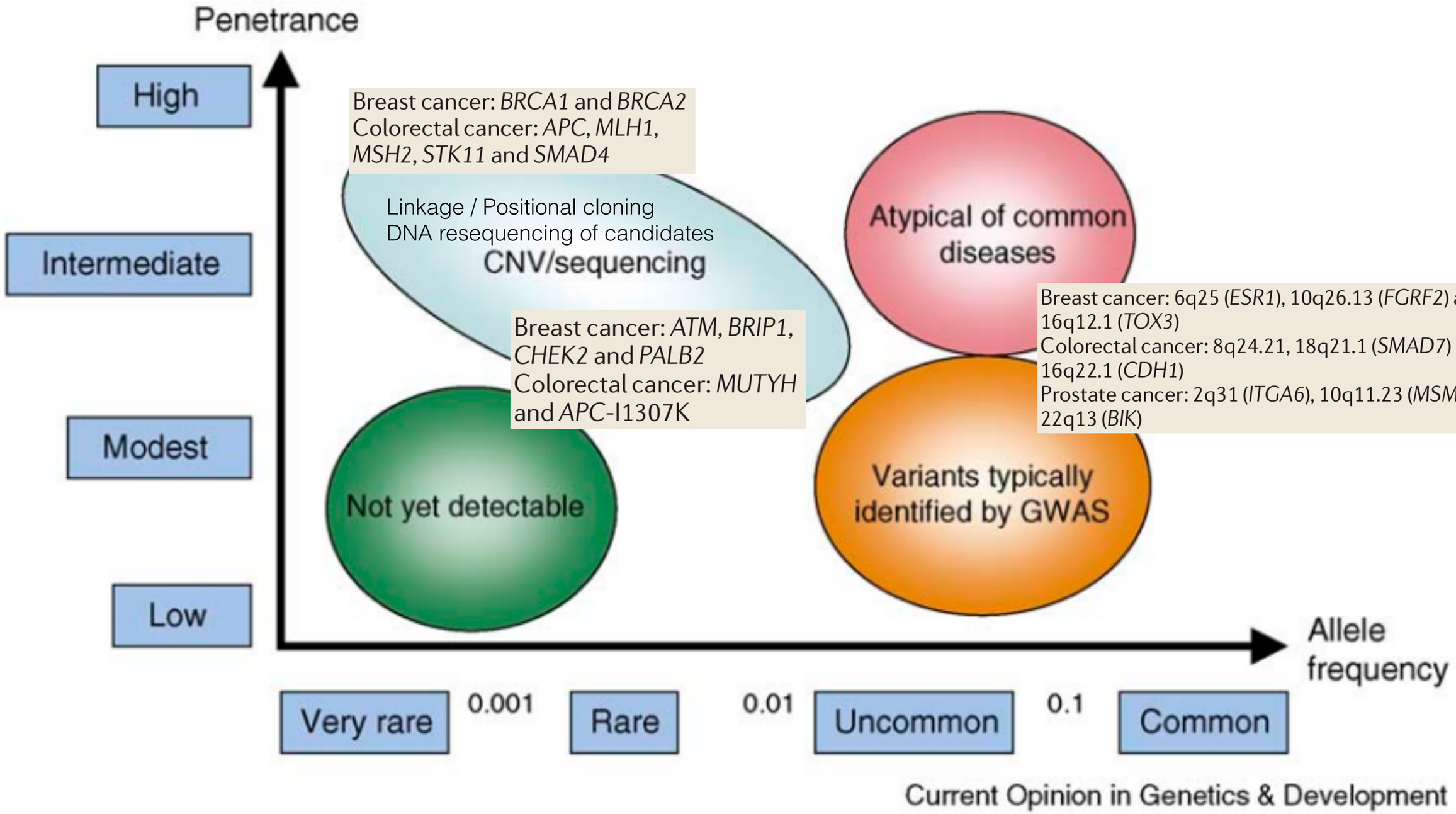
- Genetic aberrations drive malignancies
- Highly specific anticancer agents modulate these targets

Gleevec



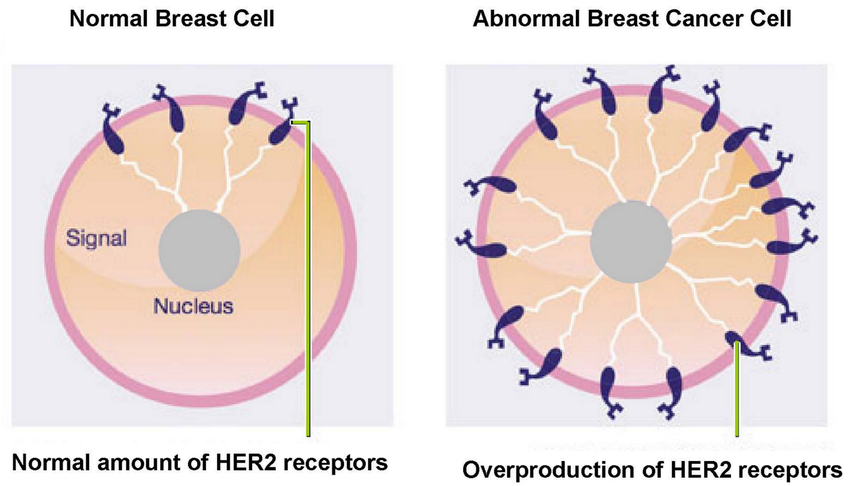
http://www.veomed.com/files/powerpoints_images/node306009/Slide9.JPG


Penetrance
High
Intermediate
Modest
Low
CNV/sequencing
Atypical of common diseases
Not yet detectable
Variants typically identified by GWAS
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common
Current Opinion in Genetics & Development


Penetrance
High
Intermediate
Modest
Low
Breast cancer: *BRCA1* and *BRCA2*
Colorectal cancer: *APC*, *MLH1*, *MSH2*, *STK11* and *SMAD4*
Linkage / Positional cloning
DNA resequencing of candidates
CNV/sequencing
Breast cancer: *ATM*, *BRIP1*, *CHEK2* and *PALB2*
Colorectal cancer: *MUTYH* and *APC*-I1307K
Not yet detectable
Atypical of common diseases
Breast cancer: 6q25 (*ESR1*), 10q26.13 (*FGRF2*) and 16q12.1 (*TOX3*)
Colorectal cancer: 8q24.21, 18q21.1 (*SMAD7*) and 16q22.1 (*CDH1*)
Prostate cancer: 2q31 (*ITGA6*), 10q11.23 (*MSMB*) and 22q13 (*BIK*)
Variants typically identified by GWAS
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common
Current Opinion in Genetics & Development

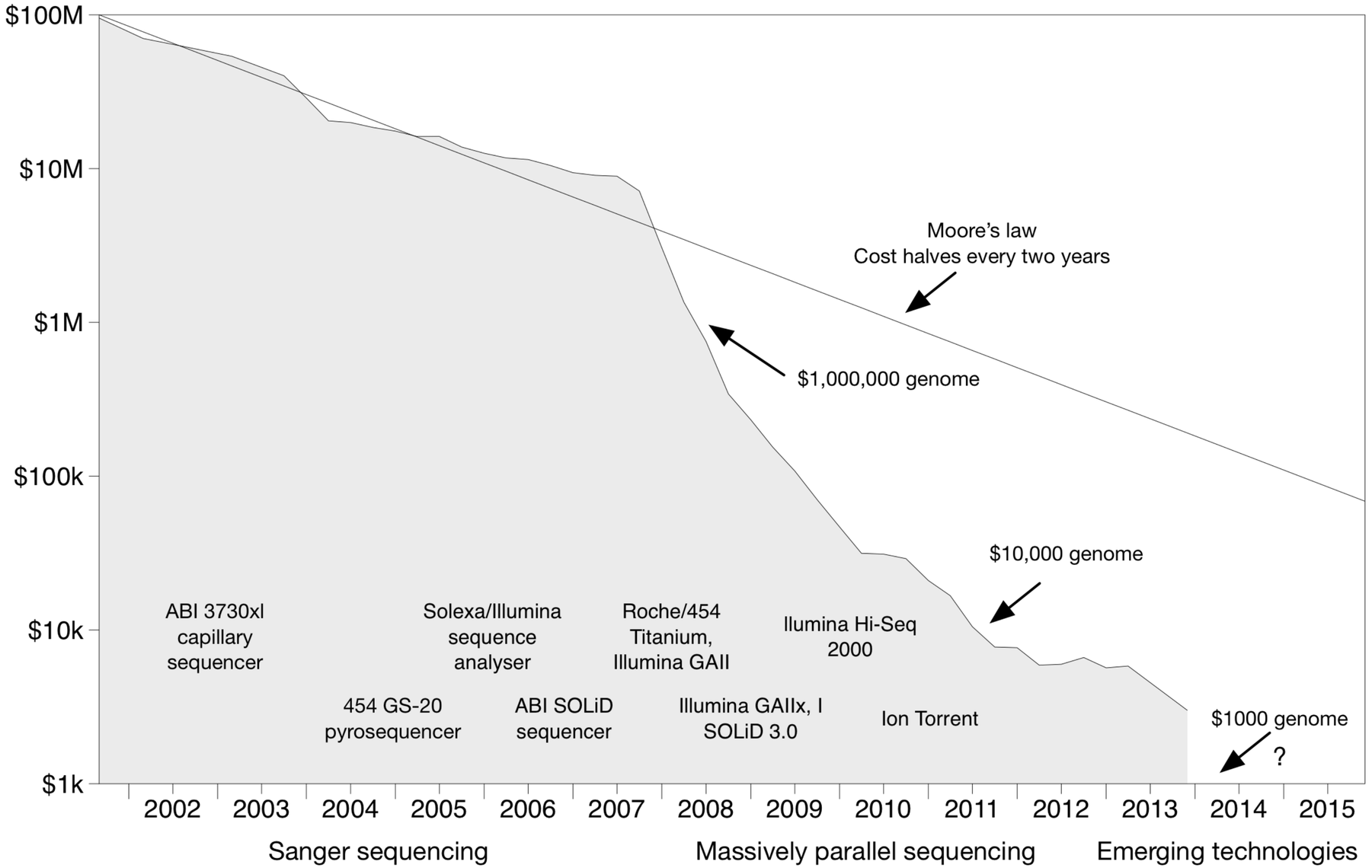
Personalized Cancer Medicine

- Using specific information about a person's tumor to help
  - Diagnose and plan treatment
  - Find out how well treatment is working
  - Make a prognosis



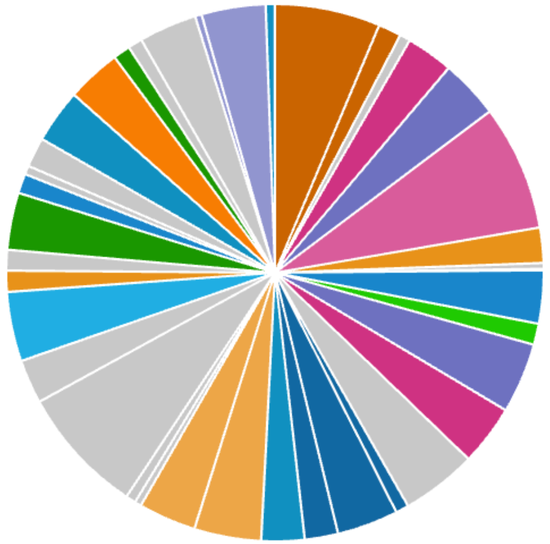
Human epidermal growth factor receptor 2

The Falling Cost of Sequencing



The Next Generation Cancer Knowledge Network

Case Distribution by Disease Type



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

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

Access Data

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

More about Accessing Data

Submit Data

The GDC provides tools to guide data submission including the **GDC Data Submission Portal**, a web-based tool for submitting clinical, biospecimen

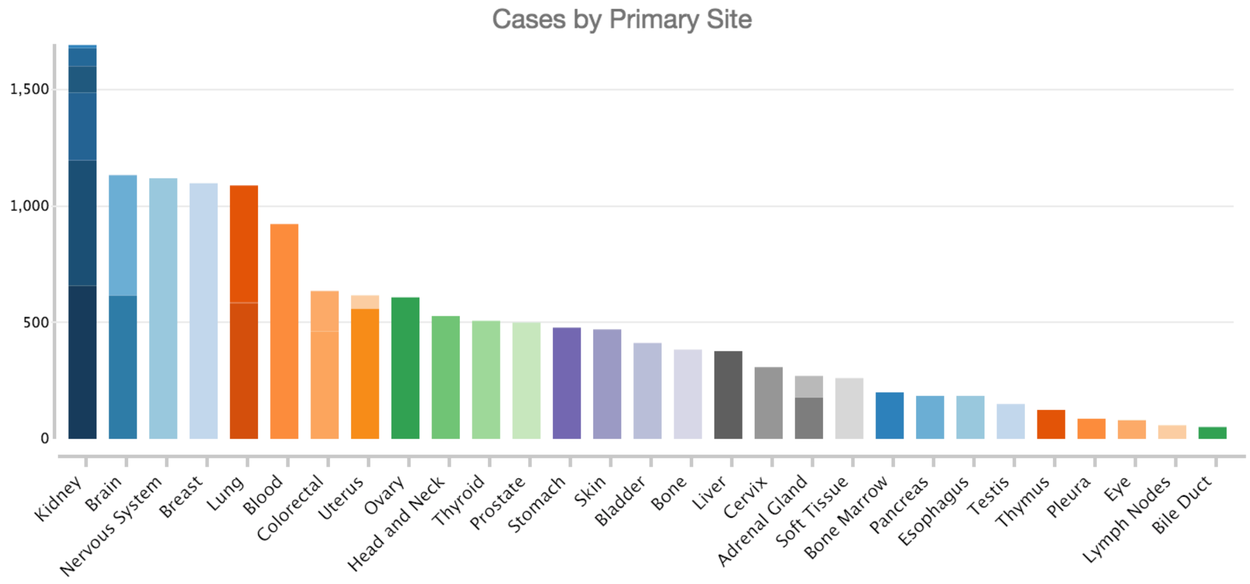
Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	182 Cases	1,514 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files



DATA PORTAL SUMMARY

Data Release 4.0 - October 31, 2016

PROJECTS	PRIMARY SITE	CASES	FILES
39	29	14,531	274,821

Infrastructure

Data is continuously being processed and harmonized by the GDC. View GDC system statistics:

Compute Infrastructure	12,800 Cores	87.96 TB RAM
Storage Infrastructure	4.98 PB Used	5.42 PB Total

View Data Download Statistics Report »

Documentation

Learn how to use the GDC Data Portal to its full potential with common topics such as:

- Browse Data using Facet Search
- Search Data with Advanced Search Technology
- Project Based Data Availability
- Controlled Access Data

Visit the Documentation Website »

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Data Portal | Website | Data Transfer Tool | API

Data Submission Portal | Documentation | Legacy Archive | GDC cBio Portal

Learning objectives

- **Understand** the principles, variables, considerations, and limitations in designing NGS experiments
- **Learn** how to analyze RNA-Seq data, ChIP-Seq data
- **Learn** standard NGS file formats and workflows
- **Accomplish** state-of-the-art data visualization
- **Be motivated!**

Jorge Andrade, PhD

Director of Bioinformatics
Center for Research Informatics
University of Chicago

1993-1998 - Bachelor's degree, Computer Science Engineering
Pontifical Catholic University of Ecuador, Quito, Ecuador

2002-2007 - PhD, Biotechnology/Bioinformatics
Royal Institute of Technology (KTH), Stockholm, Sweden

2008 - Post-Doctoral Fellowship, Genomics, Proteomics, and Bioinformatics
Karolinska Institutet, Solna, Sweden

Riyue (Sunny) Bao, PhD

2000-2004 - B.S. USTC, Hefei, China
2005-2012 - PhD, Biology, Wayne State University
2012- - University of Chicago, Senior bioinformatician

Kyle Hernandez, PhD

2002-2005 - B.S. Texas Tech University
2005-2011 - PhD, Ecology, Evolution, & Population Biology, Purdue
2011-2013 - NSF Post Doctoral Fellow
2013- - University of Chicago, Senior bioinformatician


Tzuni Garcia, PhD
Wenjun Kang, PhD