Cancer genomincs

David Haussler
University of California, Santa Cruz

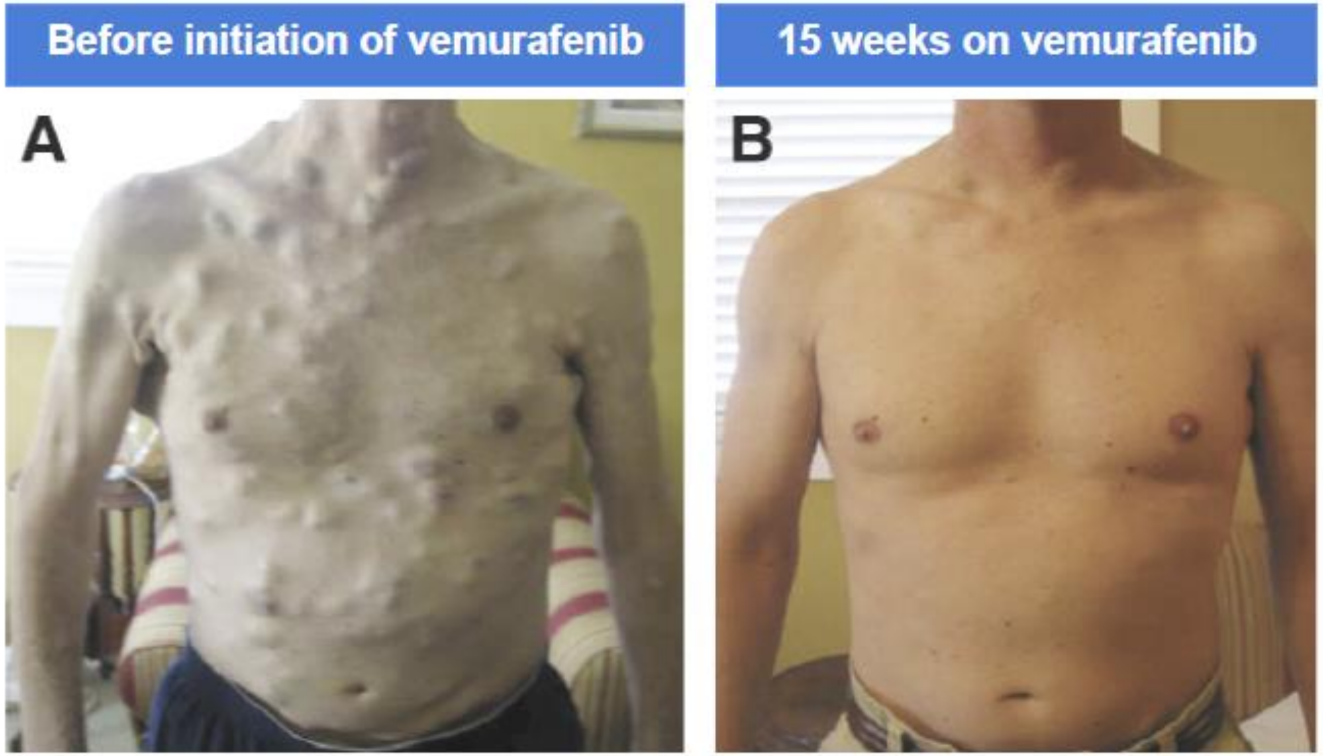
UC SANTA CRUZ

NATIONAL CANCER INSTITUTE

National Human Genome Research Institute

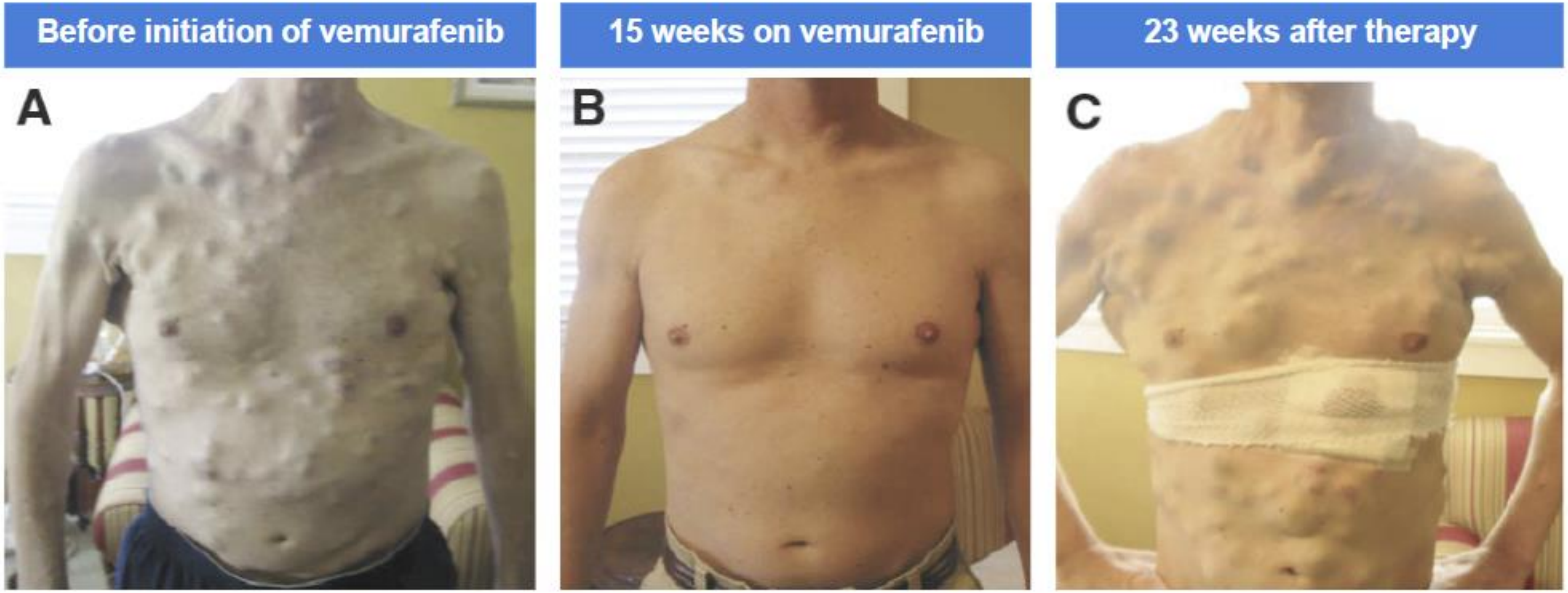
NATIONAL INSTITUTES OF HEALTH

In cancers driven by a single mutation, like BRAF V600 in metastatic melanoma, targeted drugs can give spectacular results



Roche

But combination or immunotherapies will be required to prevent relapse, just as in the treatment of HIV AIDS



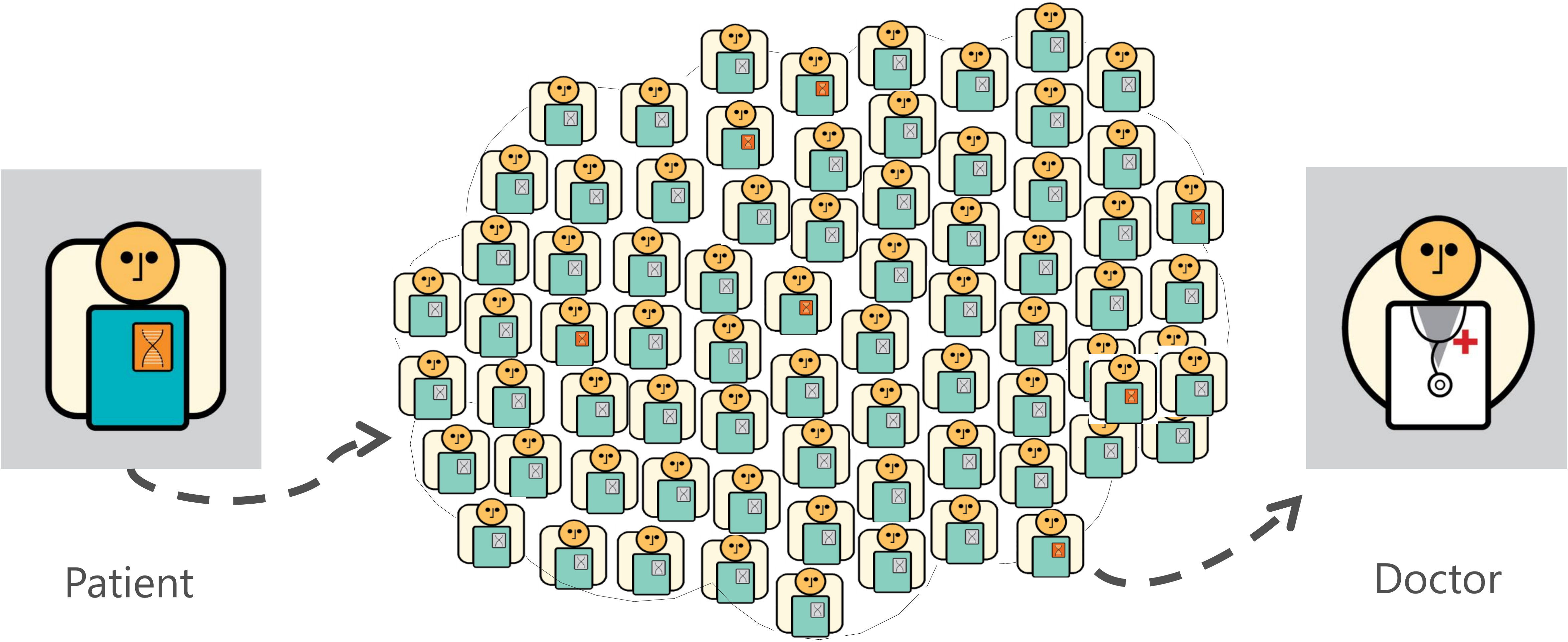
Roche

Some motivations for large-scale application of comparative genomics in cancer

- Bring data to research and insights to clinical practice
- Learn to link phenotypes, including clinical outcomes, to underlying molecular aberrations
- Create the infrastructure to select patient populations for targeted clinical trials, and to enable a new kind of global rapid learning cycle that complements targeted trials
- Gain a mechanistic, molecular level understanding of the etiology of disease and mechanisms of resistance to treatment

All these require statistical power

Genomes are the key to the future of cancer treatment



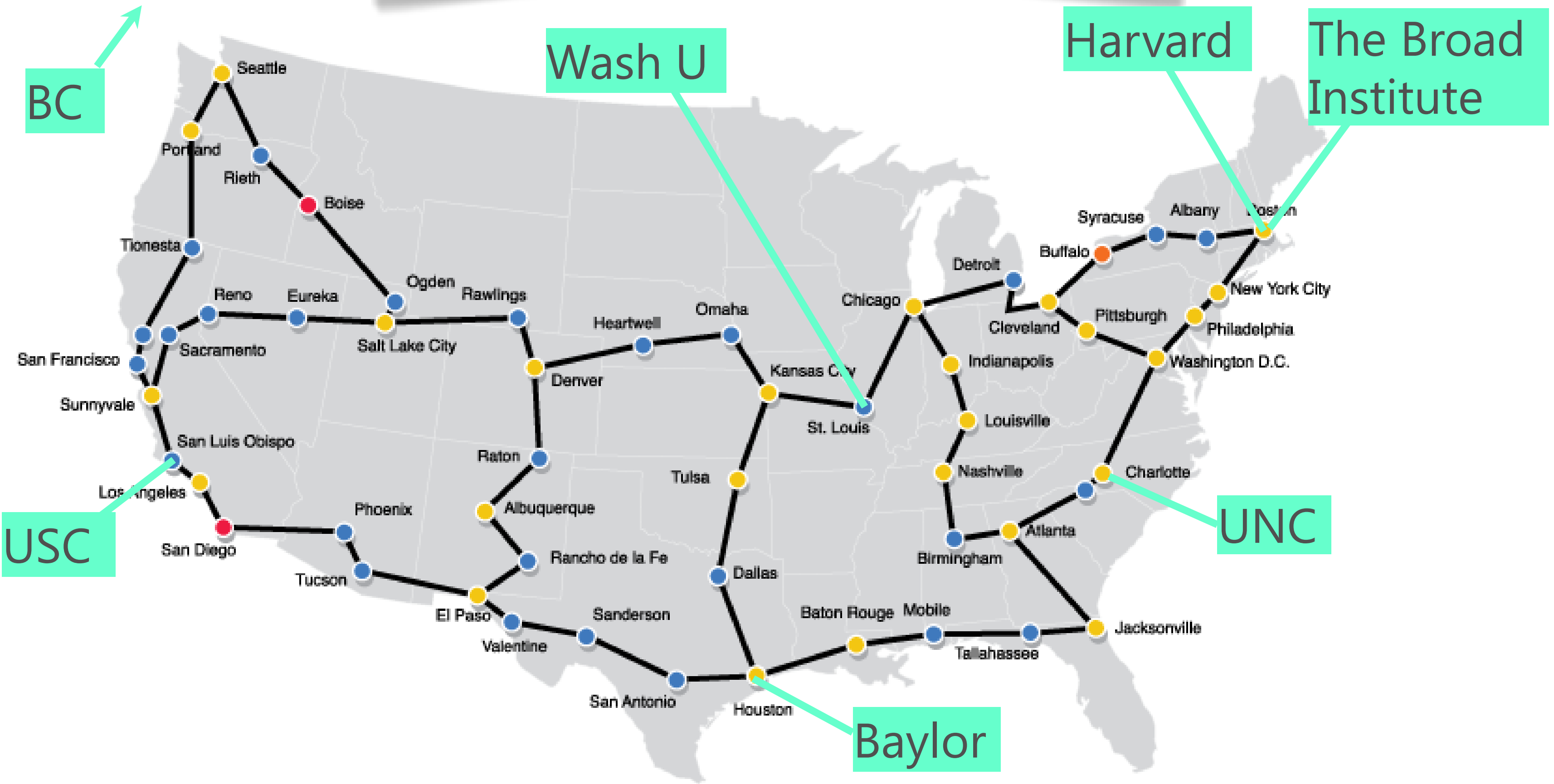
Patient

Cancer Genome Database

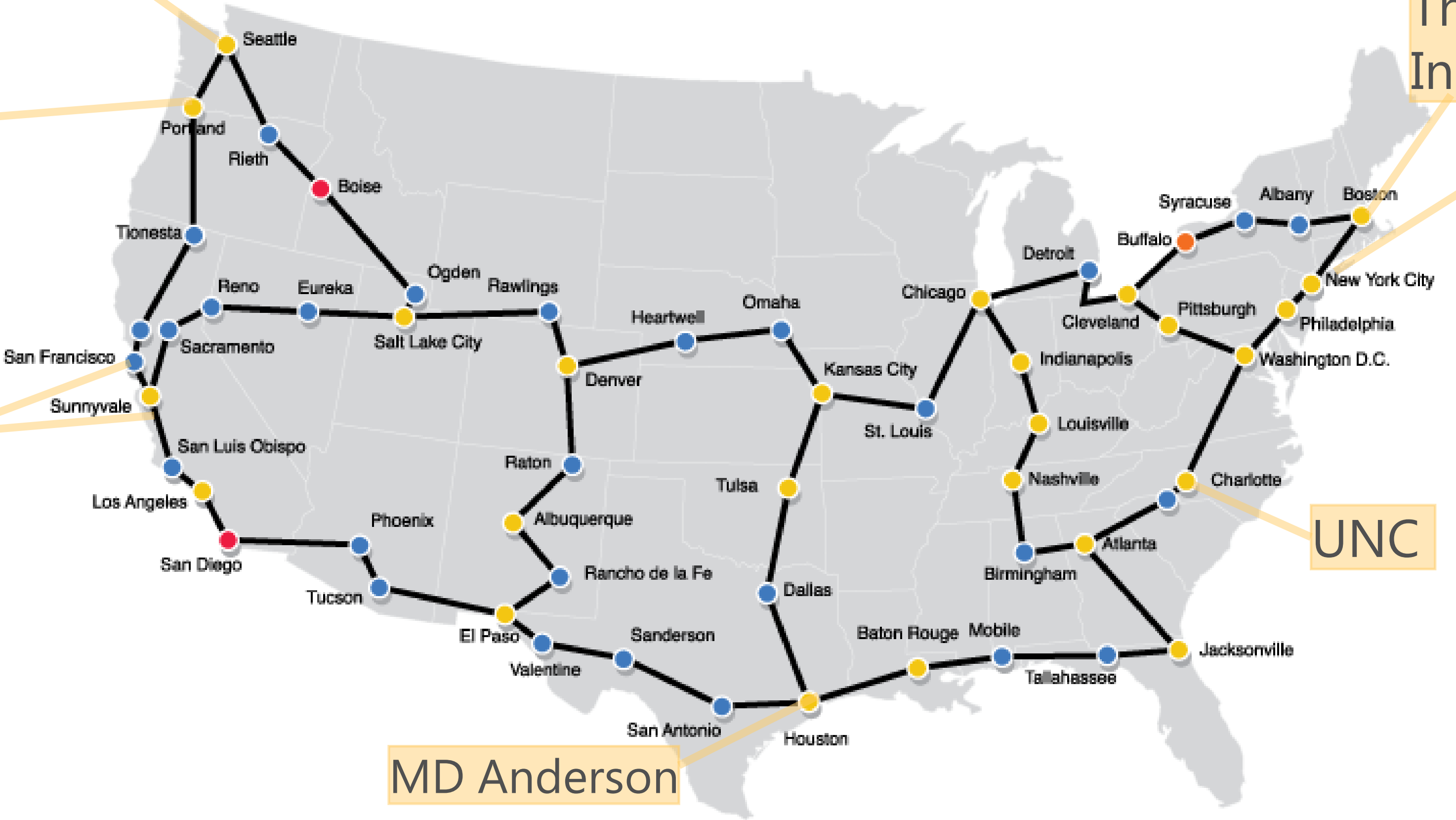
Doctor

The Cancer Genome Atlas:10,000 tumors from 20 adult cancers

TCGA Sequencing Centers



TCGA Analysis Centers



Institute for Systems Biology, Seattle

Oregon H&S U

UCSC,
Buck Institute

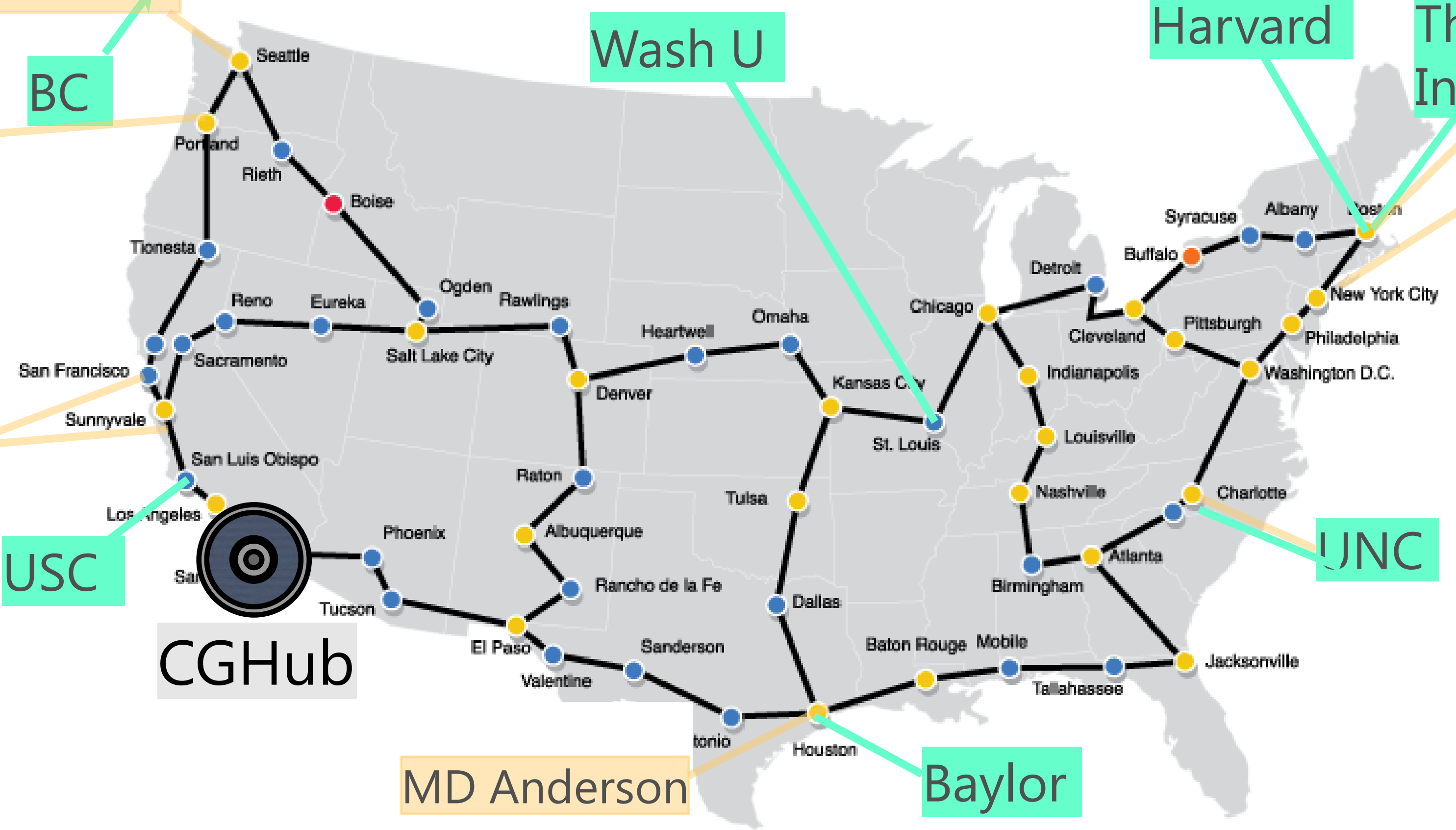
MD Anderson

The Broad Institute

MSK Cancer Ctr

UNC

The Cancer Genomic Hub



Institute for Systems Biology, Seattle

Oregon H&S U

BC

UCSC,
Buck Institute

USC

CGHub

MD Anderson

Baylor

Wash U

Harvard

The Broad Institute

MSK Cancer Ctr

UNC

CENCER GENOMICS HUB

- Total Cost ~ $100/year/genome at 50K genomes
- Houses genomes from all major NCI projects
- Planned 5 PB, Scalable to 20 PB

- FISMA compliant
- 1st NIH Trusted Partner
- COTS hardware
- High availability
- CentOS, standard linux tools
- General Parallel Filesystem
- Dual RAID 6
- Co-location opportunities

University of California
Santa Cruz Cancer Genomics Hub
Supporting the cancer research community

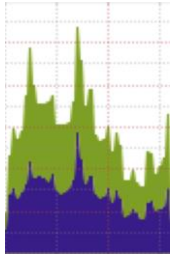
CGHub at San Diego Supercomputer Center

Current Stats (as of July 4, 2013)

490,516 total files downloaded

5,910 TB transferred

544 TB data
46,212 files



3 Gb/s typical downloads in aggregate outbound from CGHub

Future Requires Global Network of Hubs

Different Requirements for 1M Genomes

- Different types of data interactions:
  - Support both research and clinical practice
  - Compute within a provided cloud
  - Separately URIed, metadata-tagged parts of a single patient file supporting 3rd party mashups and tools
- Harmonized portable consents, sample donor has fined-grained control of who can access their data parts, trusts the security provided
- APIs, not file formats. 3rd parties must be able to build on it: goal to enable research and clinical analysis, not usurp it
- Benchmarking so all can use system to improve methods, e.g. variant calling

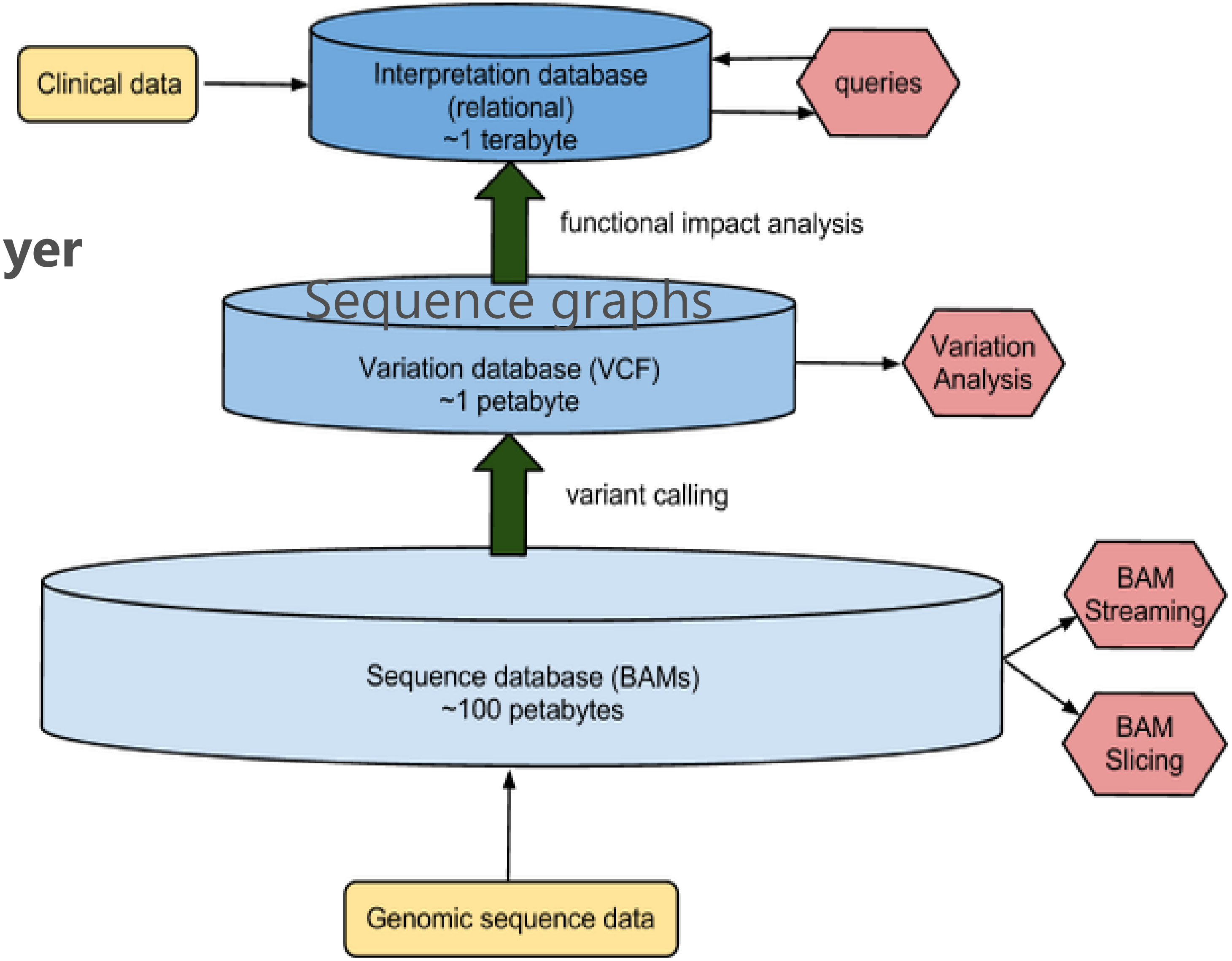
Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

Possible Genome Commons Architecture

Interpretation Layer

Variation Layer

Read Layer



What would it cost to store and analyze 1M Cancer Genomes in 2014?

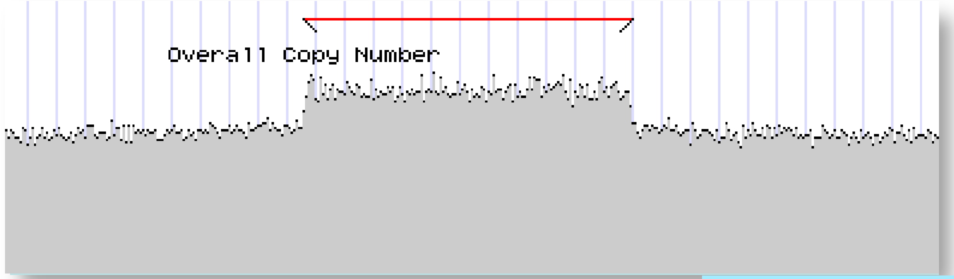
- Our estimate is ~ $50/genome/year in 2014 to store and analyze 1M whole genomes (~ 100 petabytes, 2 months of YouTube growth)
  - 25,000 disks and 100,000 processor cores
  - Including operating costs: space, electricity, operators
  - Including 2nd center to protect against disasters
- Note that cancer is the high water mark for global genome commons requirements, requirements for other diseases are smaller, less complex, assuming cancer includes full germline and somatic cell analysis

Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

- Founding partners on June 5, 2013: 70+ leading health care, research, and disease advocacy organizations from over 40 countries
- Mission: to enable rapid progress in biomedicine
- Plan:
  - create and maintain the interoperability of technology platform standards for managing and sharing genomic data in clinical samples;
  - develop guidelines and harmonizing procedures for privacy and ethics in the international regulatory context;
  - engage stakeholders across sectors to encourage the responsible and voluntary sharing of data and of methods.

Extracting molecular state from raw DNA reads



chr2 : 29,064,107

OV-0751 Somatic Reads

```
CCTTTGTCTGCATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACAggagtattaaccccacctgatctcacgatgggagaggagacgccatctgcagcagtggtggtag

CCTTTGTCTGCATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCC
               TGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCGCT
CCTTTGTCTGCATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAAAGGAGGC
                                         TATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCGCTGCCATCTGTCTCCCACC
CCTTTGTCTGCATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGGCCACAGAGGTCA
                         CTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCCTCTCCTCCATCTCCCATCGCTGTCACATATT
CCTTTGTCTGCATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAAAGGAGGCC
           ATTCTGTGGAATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCCGAGGGCATCTCCTCCATCTCCCAC
                           GGCTGGCTGCACCCTATAATGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCGCTGCCATCTGTCTC
                                        CTAGATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCCTCAGAGGGCATCACCTCCACTTCCCATCGCTGCCATCTGTCTCCCACC
                     ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCCCCCATCCCCGCCCTC
                                  TGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCCACAAAGGGCCACTTCCCCACCTCCCCTCCCTGCCCACTGGCTCCCTCC

                                    cactttctacagacgatgtcaccttccacctCACAGAAAATGGAGGCCATCAGAGGGCATCTCCtccatctcccatcgctgccatctgtctcccacc
```

chr2 : 28,500,054

Tandem Duplication Size = 564,053 bp

Zack Sanborn, now at
Five3 Genomics

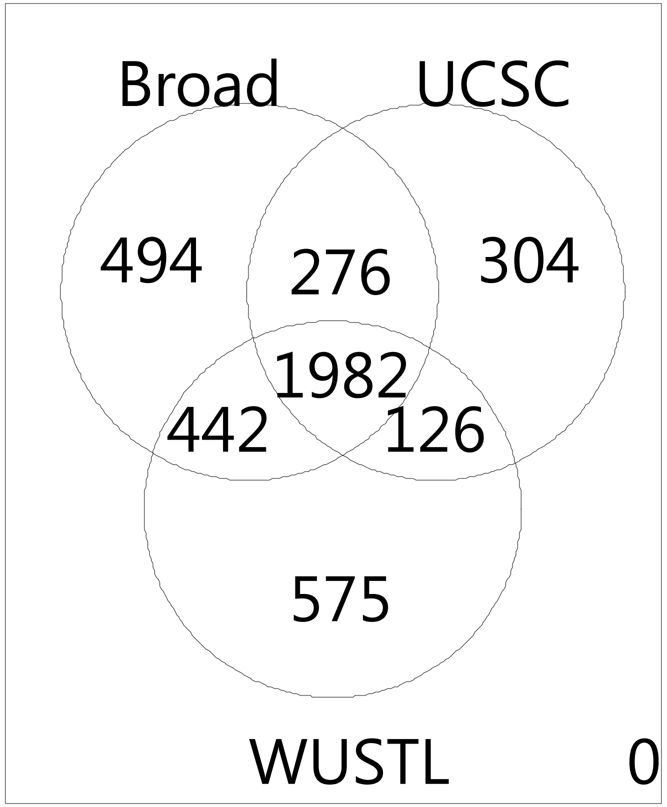
Completely solved problem? Not yet.

Given the same raw sequence (BAM) files, different mutation calling pipelines do not completely agree

Point mutations called in tumor TCGA-13-0725

Total calls:	**Called by 2 other centers**	**Called by at least 1 other**
Broad: 3,194	62%	85%
UCSC: 2,688	74%	89%
WUSTL: 3,125	63%	82%

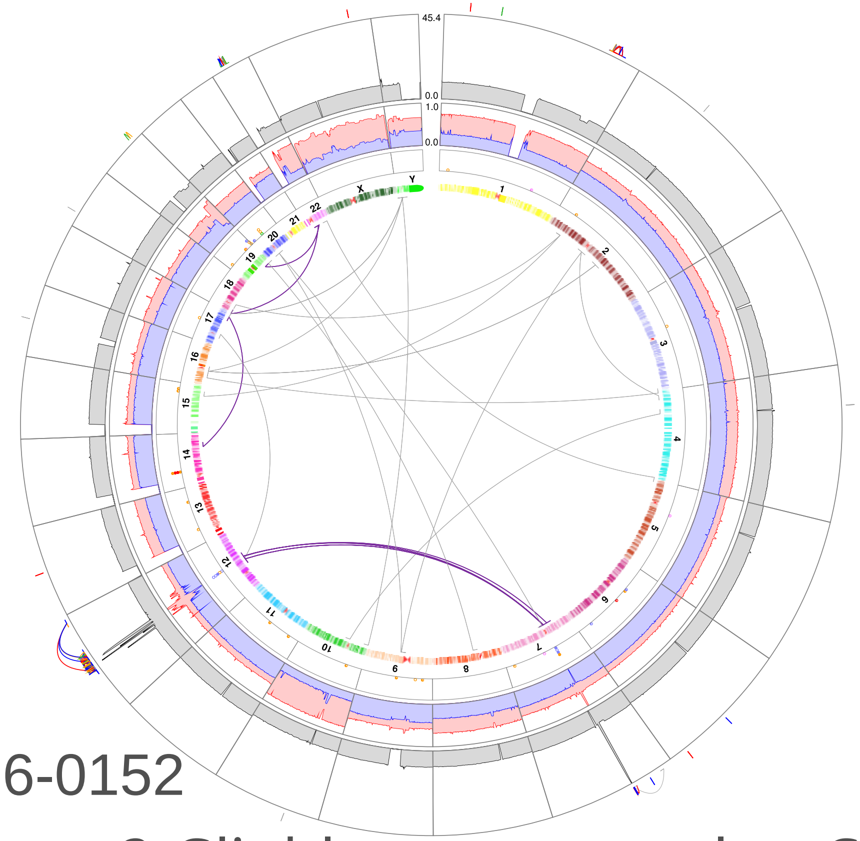
TCGA-13-0725



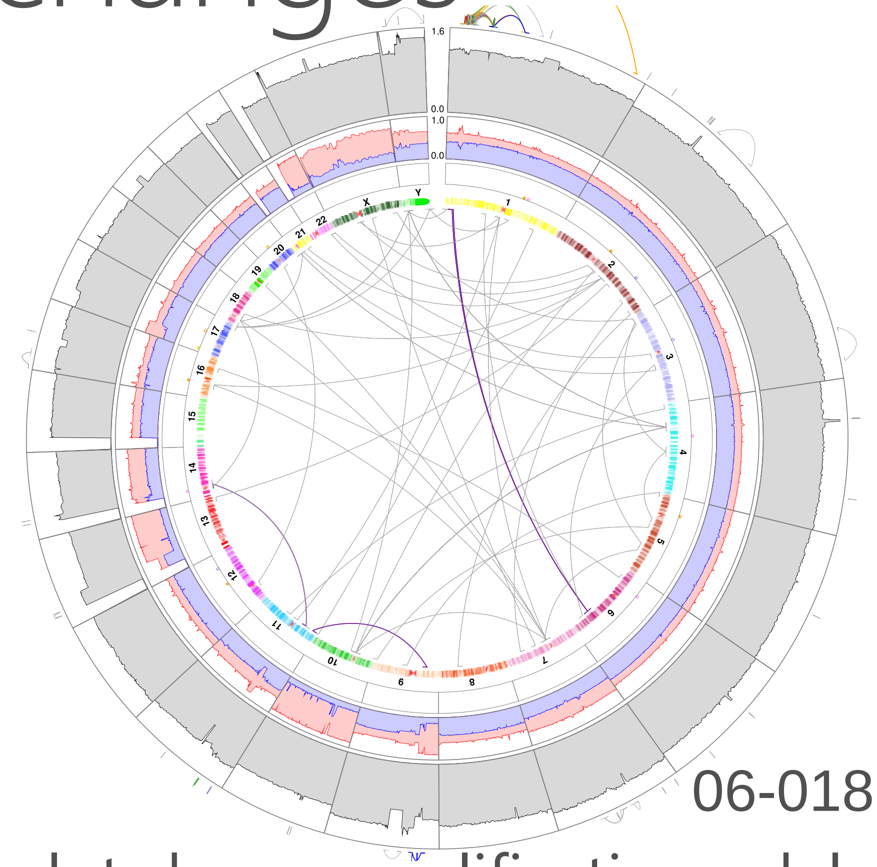
Still work to do to harden mutation-calling software, even for point mutations

UCSC, Broad are leading a series of TCGA/ICGC international benchmark challenges. Visit cghub.ucsc.edu for TCGA Benchmark 4

Even more differences in calling structural changes



06-0152



06-0188

- 2 Glioblastoma samples. Circle plot shows amplifications, deletions, inter/intra chromosomal rearrangement
- These 2 samples have 23/25 top Broad, 21/29 top UCSC events

GBM group

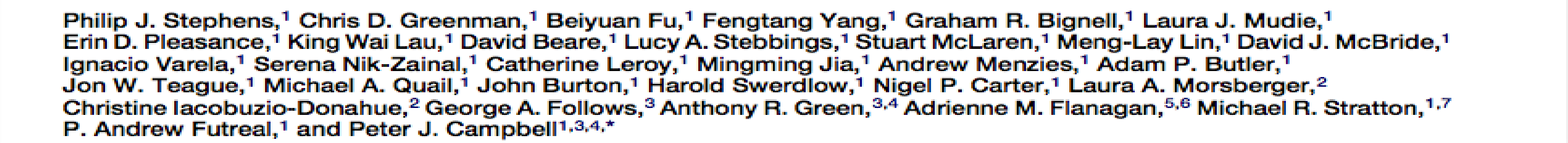
In 11/16 WGS TCGA glioblastoma cases similar events lead to homozygous loss of CDKN2A/B

	One Copy Deleted by	Other Copy Deleted by
5 GBMs	Focal Loss	Arm-Level loss of chr9p (via inter-chrom translocation)
3 GBMs	Focal Loss	Arm-Level loss of chr9p (mechanism unknown)
2 GBMs	Focal Loss	Complete loss of chr9
1 GBM	Focal Loss	Complex event
5 GBMs	*No loss detected*	*No loss detected*

Zack Sanborn

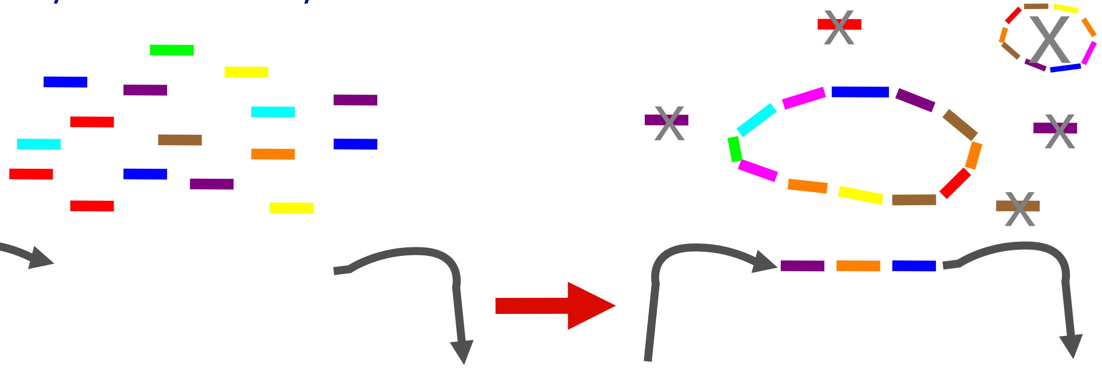
Cell

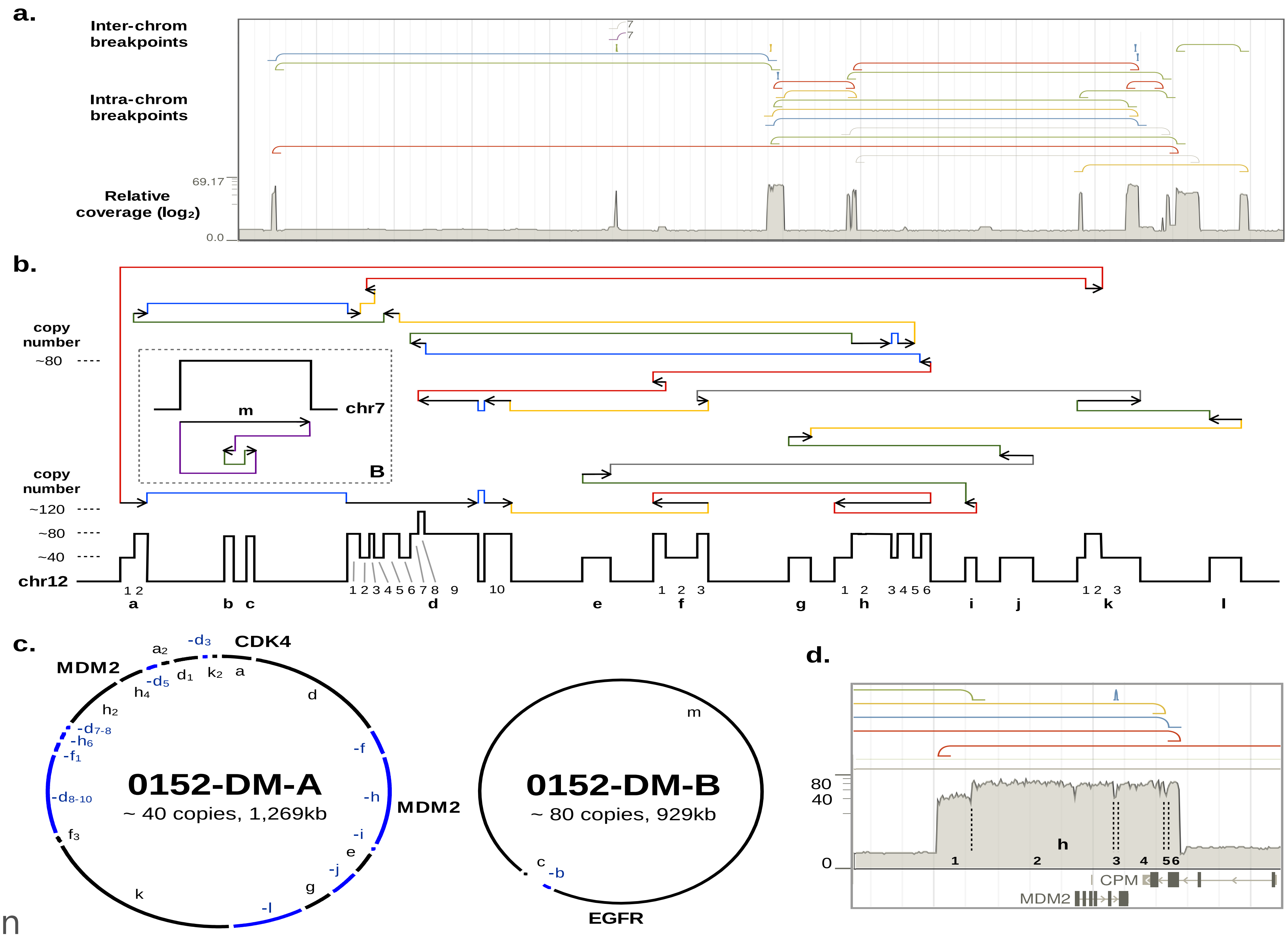
Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development



Philip J. Stephens,[1] Chris D. Greenman,[1] Beiyuan Fu,[1] Fengtang Yang,[1] Graham R. Bignell,[1] Laura J. Mudie,[1]
Erin D. Pleasance,[1] King Wai Lau,[1] David Beare,[1] Lucy A. Stebbings,[1] Stuart McLaren,[1] Meng-Lay Lin,[1] David J. McBride,[1]
Ignacio Varela,[1] Serena Nik-Zainal,[1] Catherine Leroy,[1] Mingming Jia,[1] Andrew Menzies,[1] Adam P. Butler,[1]
Jon W. Teague,[1] Michael A. Quail,[1] John Burton,[1] Harold Swerdlow,[1] Nigel P. Carter,[1] Laura A. Morsberger,[2]
Christine Iacobuzio-Donahue,[2] George A. Follows,[3] Anthony R. Green,[3,4] Adrienne M. Flanagan,[5,6] Michael R. Stratton,[1,7]
P. Andrew Futreal,[1] and Peter J. Campbell[1,3,4,*]

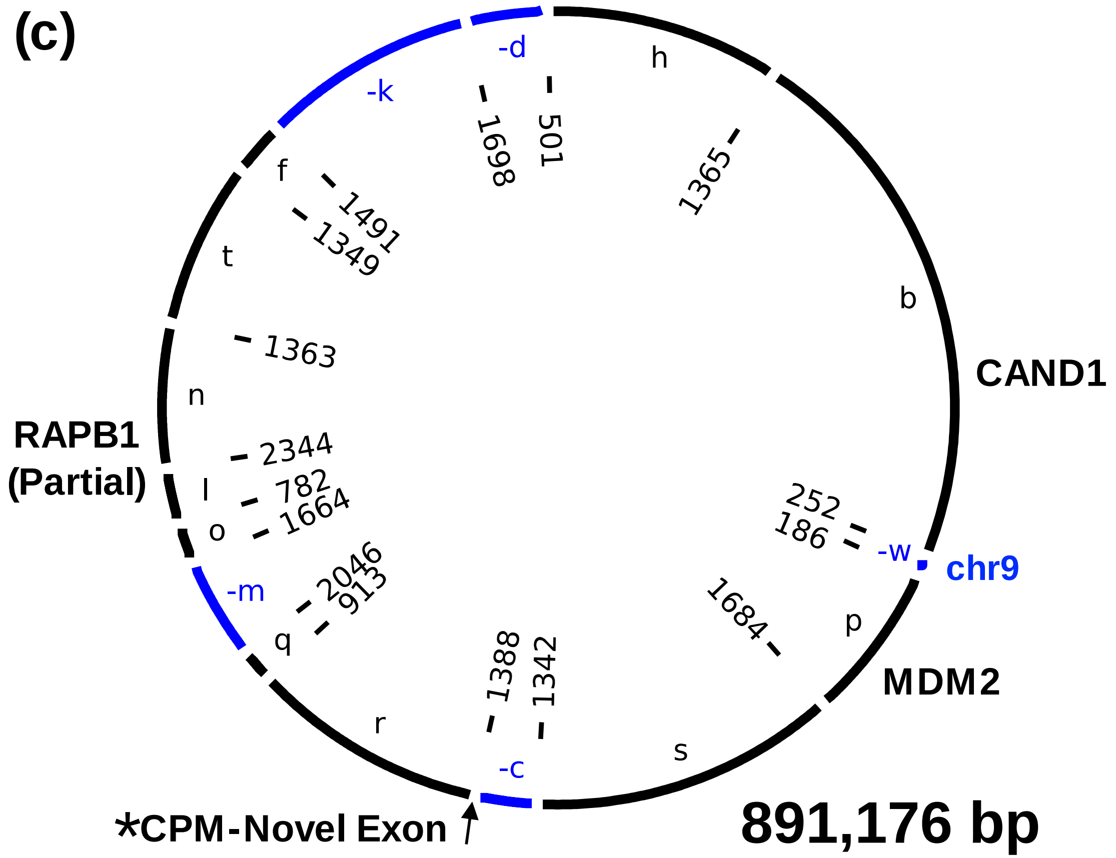
- Chromothripsis: DNA replication process get confused for a period or DNA is shattered into pieces by some high energy event when chromosome is in condensed state
- DNA repair mechanisms try to stitch genome back together
- Can generate rearrangements, losses, and circular “double minute” chromosomes



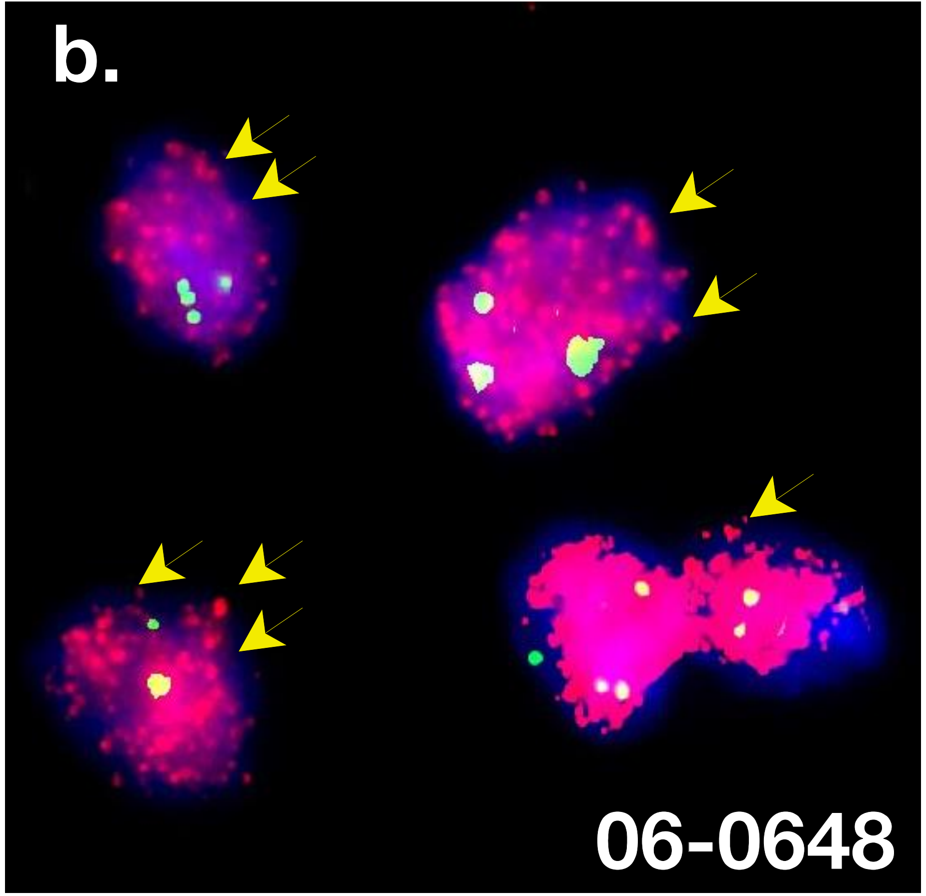


Zack Sanborn

DM from another GBM tumor. We estimate 20% of GBMs have oncogenic DMs

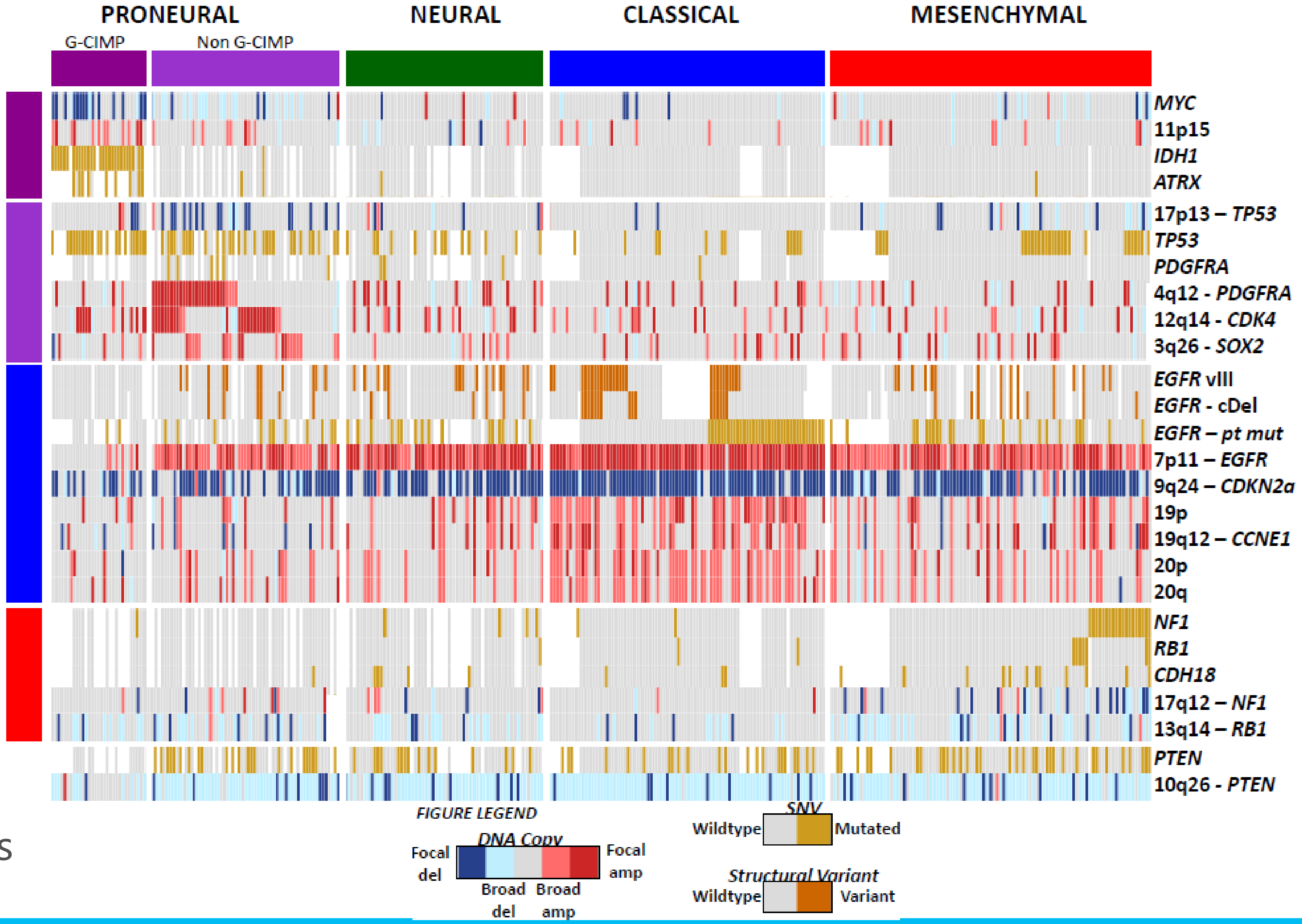


Validation by FISH



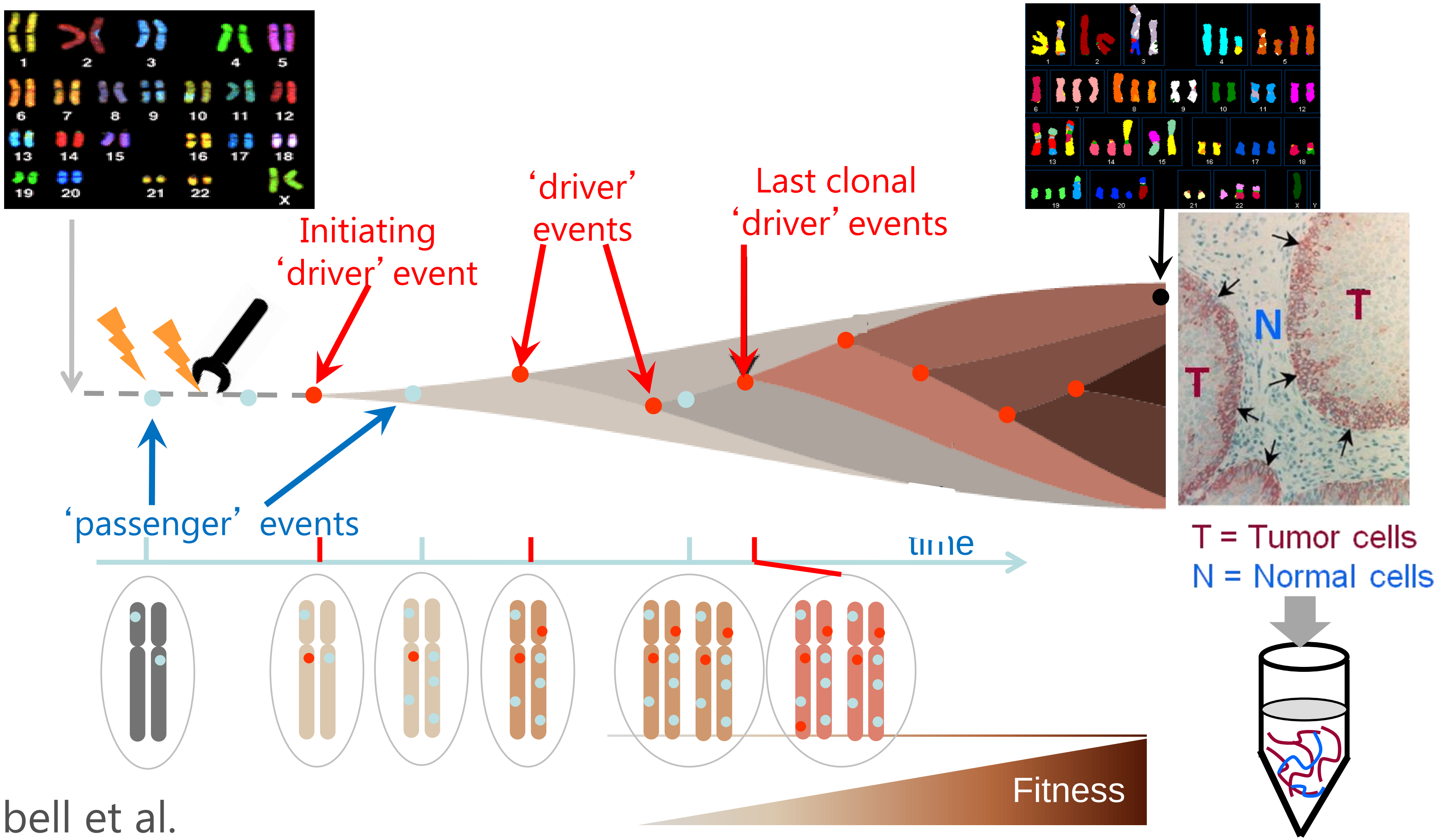
Zack Sanborn, Cameron Brennan

Highlights from analysis of 500 GBMs



TCGA GBM Analysis Working Group

Tumors have metagenomes: mixture of clones resulting from somatic selection of subclones



Initiating ‘driver’ event

‘driver’ events

Last clonal ‘driver’ events

‘passenger’ events

time

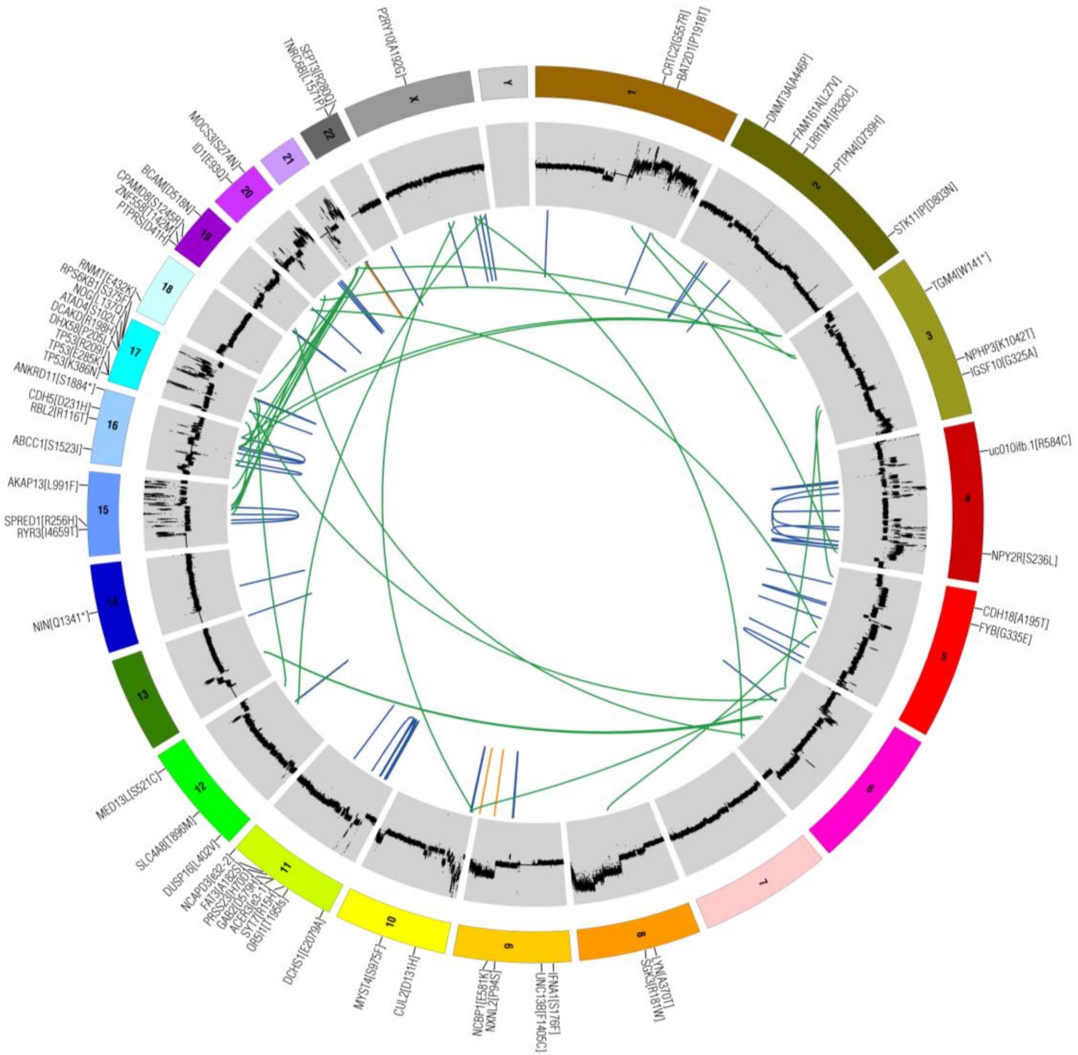
Fitness

T = Tumor cells

N = Normal cells

Adapted from Campbell et al. Nature (2010) by Gaddy Getz

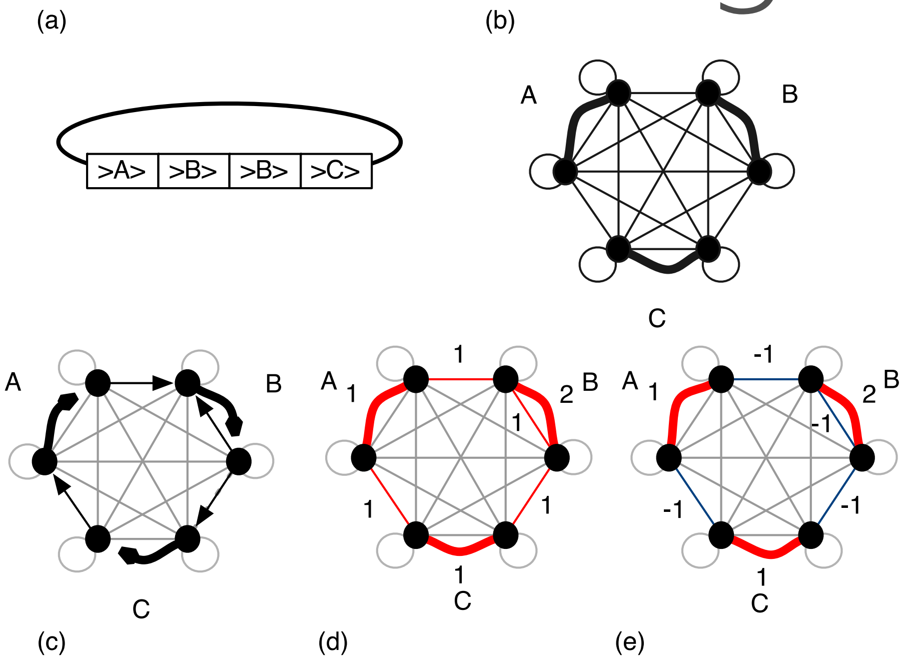
One can use sequence graphs for analysis of cancer metagenomes



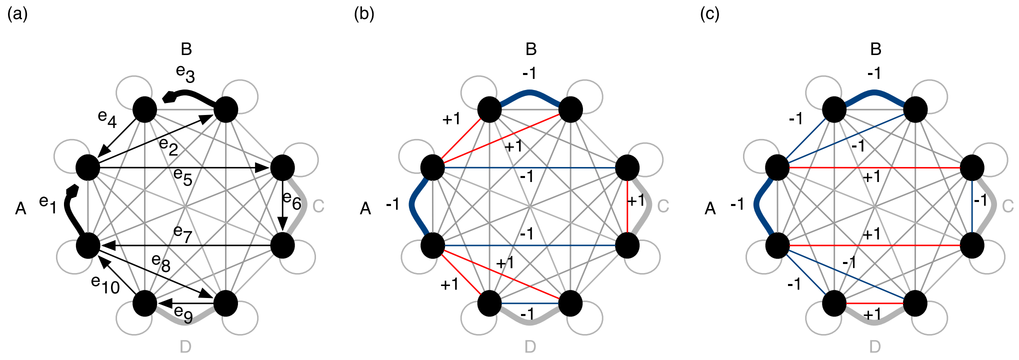
Daniel Zerbino

Algebraic/Combinatorial Approach to Comparative Metagenomics

Flows:

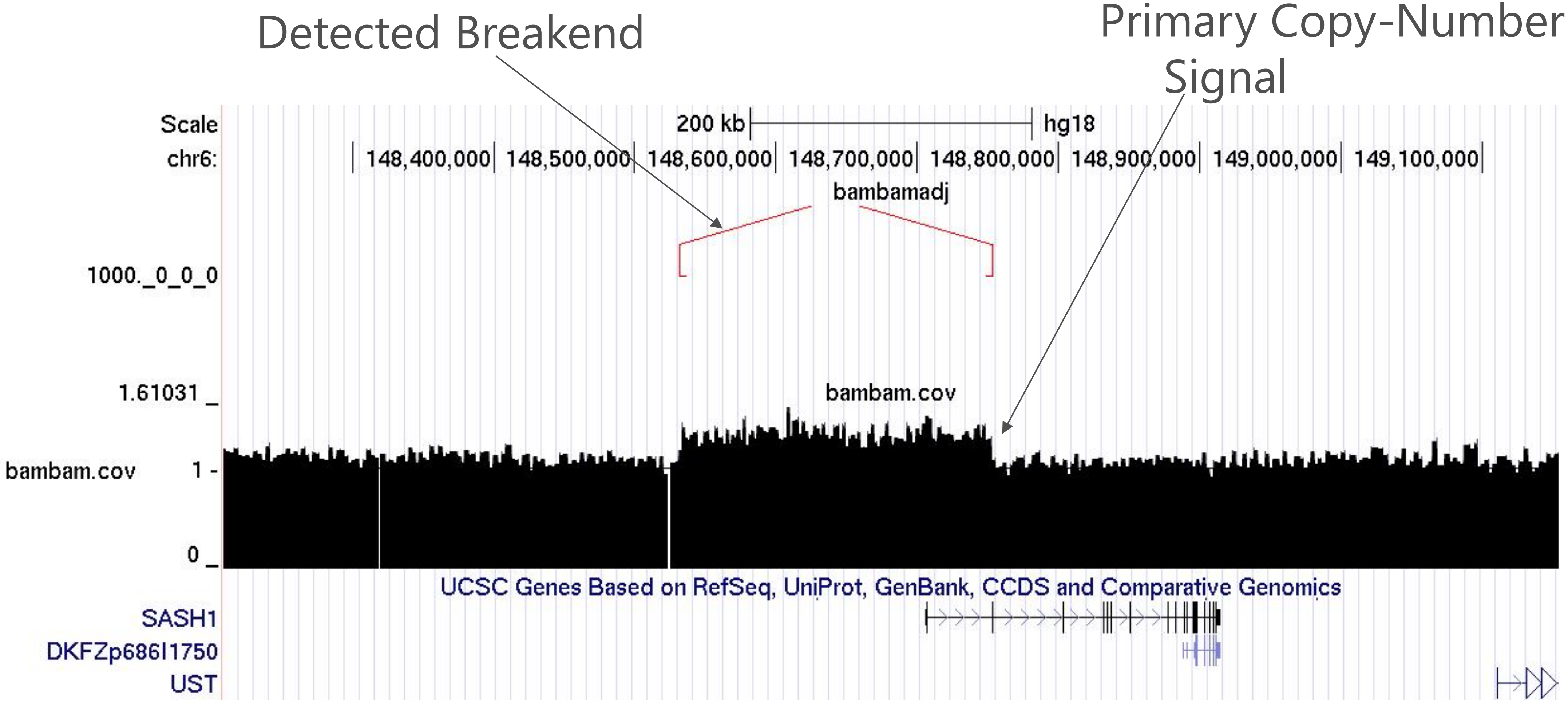


Alternating and simple flows:



Daniel Zerbino

Duplication – raw data



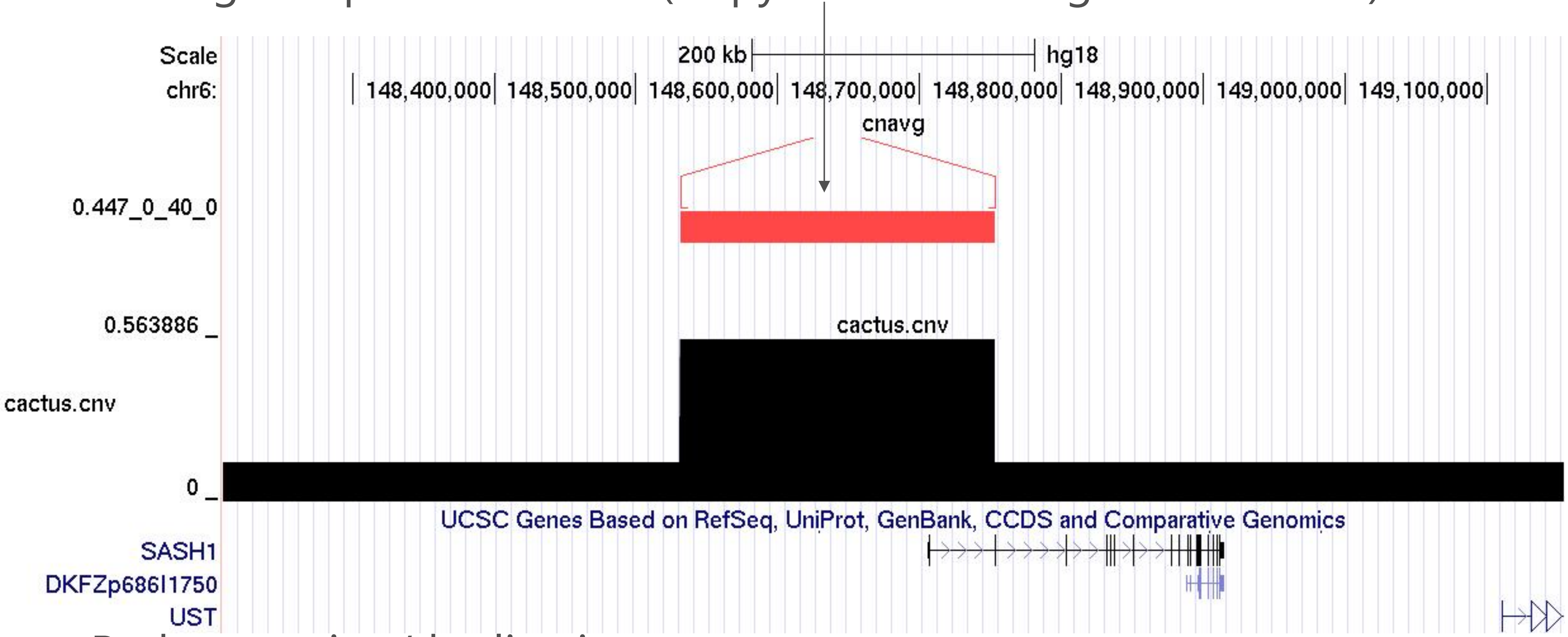
Detected Breakend

Primary Copy-Number Signal

Daniel Zerbino

Duplication – model from data

Single duplication event (Copy number change + Breakend)

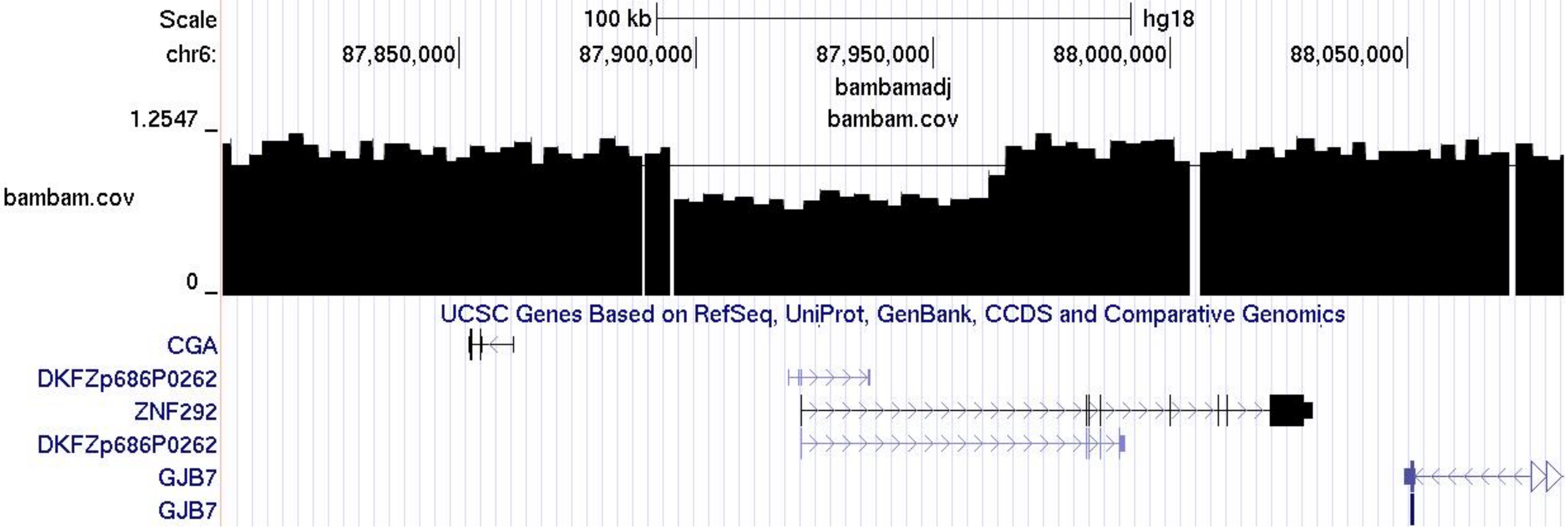


Red = creation/duplication

Daniel Zerbino

Deletion – raw data

(No breakend detected)

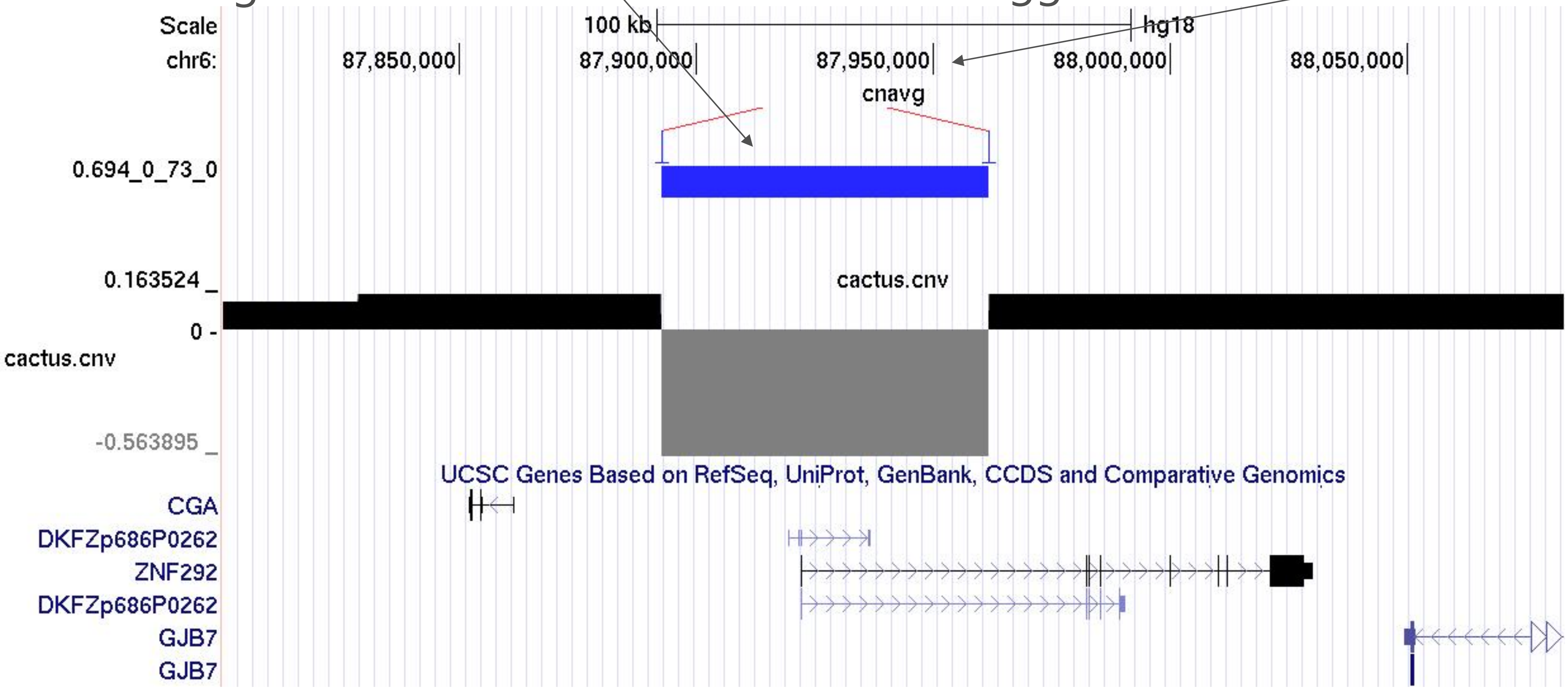


Daniel Zerbino

Deletion – model from data

Single deletion event

Suggested novel breakend creation

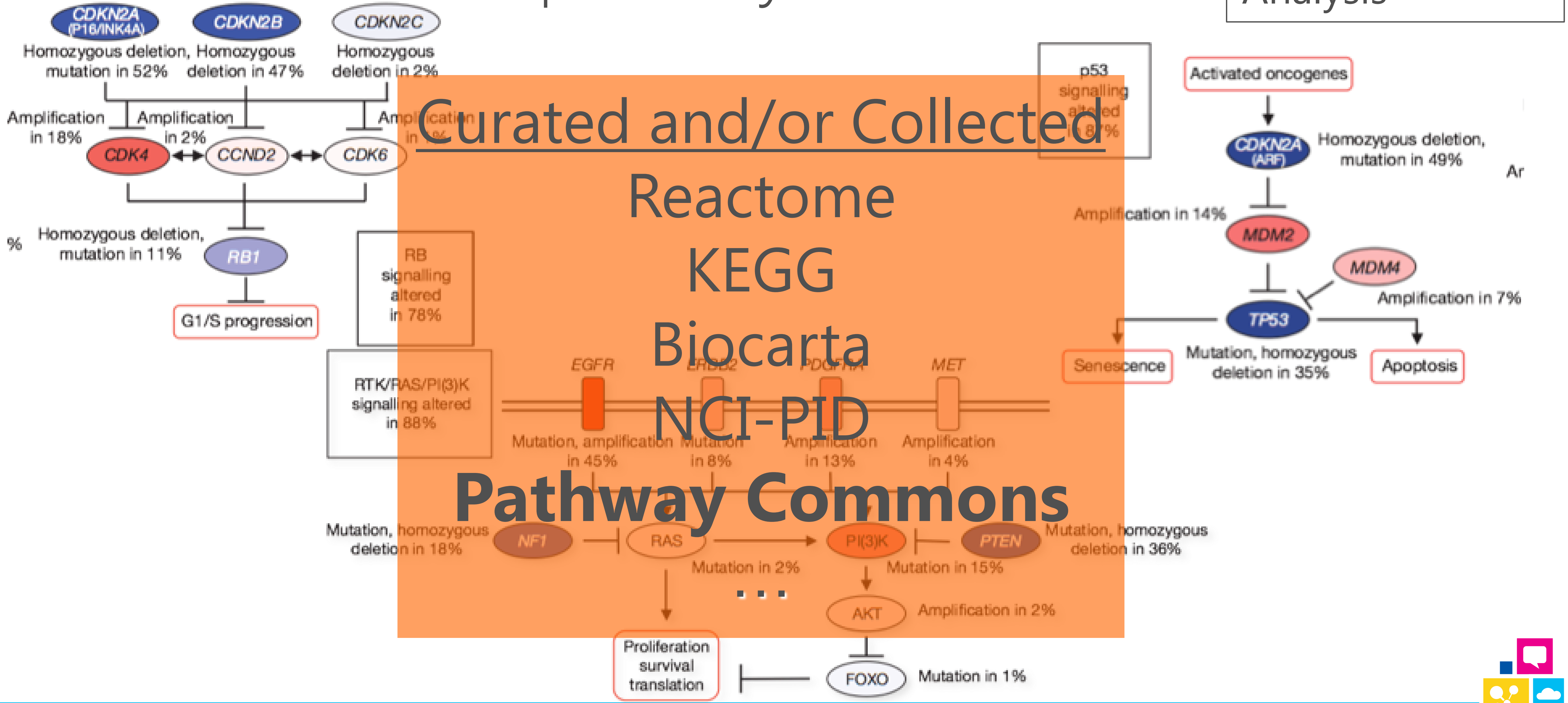


Blue = removal/deletion

Daniel Zerbino

Finally, key is interpretation of genomics data at the pathway level

TCGA Glioblastoma Analysis



CDKN2A (P16/INK4A)
CDKN2B
CDKN2C
Homozygous deletion, mutation in 52%
Homozygous deletion in 47%
Homozygous deletion in 2%
Amplification in 18%
Amplification in 2%
Amplification in 1%
CDK4
CCND2
CDK6
Homozygous deletion, mutation in 11%
RB1
G1/S progression
RB signalling altered in 78%
RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
PI(3)K
PTEN
Mutation, homozygous deletion in 36%
Mutation in 2%
Mutation in 15%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%
p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis

<u>Curated and/or Collected</u>

Reactome

KEGG

Biocarta

NCI-PID

Pathway Commons

. . .

The Age of Opportunity for the Study of Genetics and Medicine

- #1 infrastructure issue is to achieve statistical power by aggregating information. We must head off the development of genomic information silos
- #1 interpretive challenge is to accurately read a genome and model effects of genetic changes on molecular pathways and phenotypes

We must accelerate biomedical research and improve clinical practice by building new global platforms for storage, exchange and analysis of molecular and phenotypic information

Some Current Collaborators

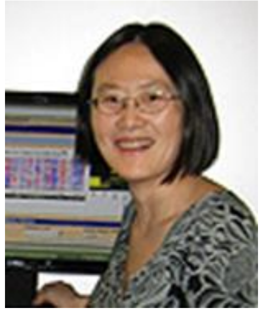
Collaborators

- Dave Patterson group, UC Berkeley
- David Altshuler, Charles Sawyers, Mike Stratton, Betsy Nabel, Brad Margus, Karen Kennedy, Tom Hudson
- Richard Durbin, Sanger Centre
- Broad Institute, Wash U., Baylor
- The Cancer Genome Atlas and its labs, esp. GBM analysis working group
- Stand Up To Cancer and its labs
- Intl. Cancer Genome Consortium and its labs
- Chris Benz, Buck Institute
- Laura Van't Veer, Laura Esserman, Joe Costello, Eric Collisson, Margaret Tempero, UCSF
- UCSC Storage Systems Group
- Joe Gray, Paul Spellman, OHSU

UCSC Cancer Integration Group

Josh Stuart, Co-PI

Jing Zhu

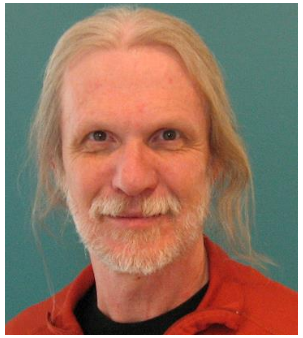


Charlie Vaske

Steve Benz

Zack Sanborn

Mark Diekhans *



Chris Benz

Chris Szeto

Sam Ng

Mia Grifford

James Durbin

Ted Goldstein

Melissa Cline

Sofie Salama *

Chris Wilks

Amie Radenbaugh

Brian Craft

Kyle Elrott
Adam Ewing
Mary Goldman
Singer Ma
Artem Sokolov
Theresa Swatloski
Daniel Zerbino

CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING
promoting discovery and invention for human health and well-being

UC SANTA CRUZ