Big data and new models needed to study DNA variation in evolution and cancer

David Haussler, UC Santa Cruz

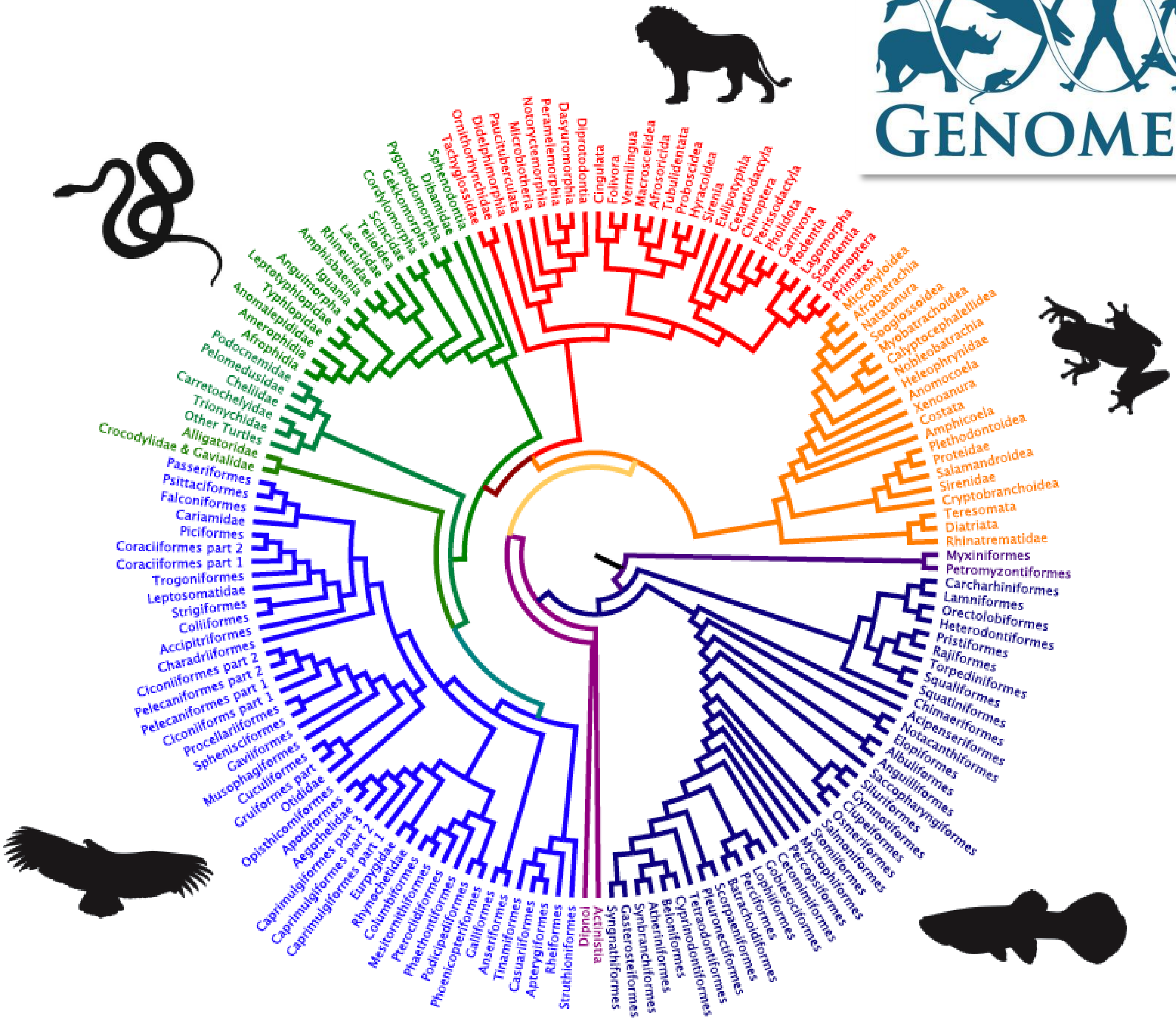
UC SANTA CRUZ

NATIONAL CANCER INSTITUTE

National Human Genome Research Institute

NATIONAL INSTITUTES OF HEALTH

GENOME 10K©



The G10K Community of Scientists

The Genome 10K Community Goal:

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

- Collect samples and sequence at least 10,000 different vertebrate species, bank fibroblast cell lines and make iPS lines for > 1,000 species. Currently ~350 genomes and dozens of iPS lines from various labs.

- Annotate genomes, map and interpret genetic differences between species, and compute the evolutionary record of genetic changes on each lineage

- Correlate with ecologic, biologic and geologic data for deep study of vertebrate diversity, biology, evolution, and for species conservation

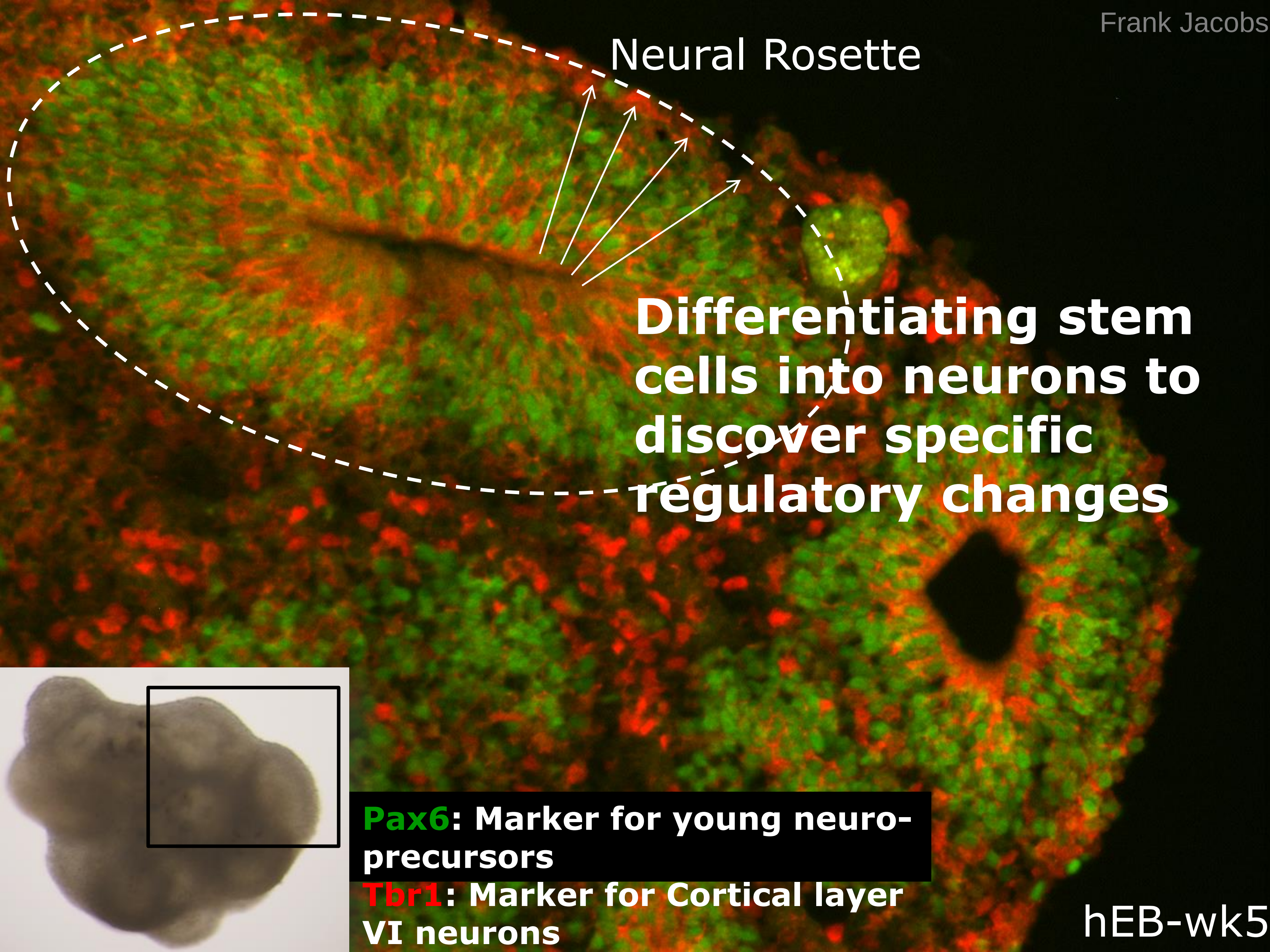
Grand scientific challenge of vertebrate molecular evolution

Reconstruct the evolutionary history of each base in the genomes of the living species

- Recognize functional elements from patterns of negative and positive selection
- Find the origins of evolutionary innovations specific to the human and other lineages

Early look at some evolutionary differences in human neurodevelopment

Frank Jacobs

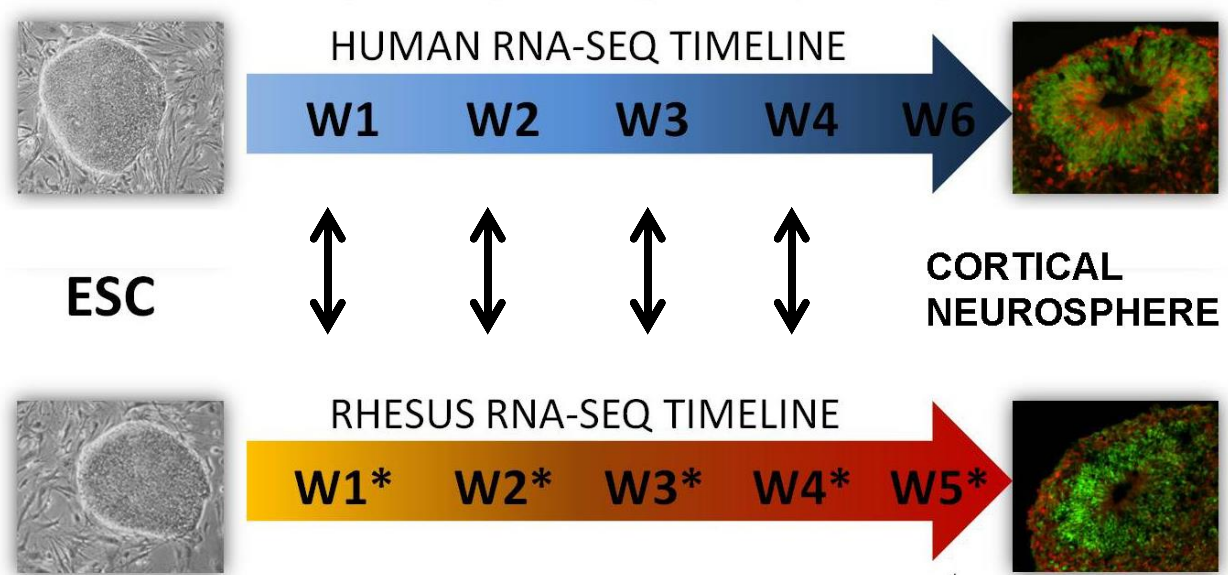


Differentiating stem cells into neurons to discover specific regulatory changes

Pax6: Marker for young neuro-precursors

Tbr1: Marker for Cortical layer VI neurons

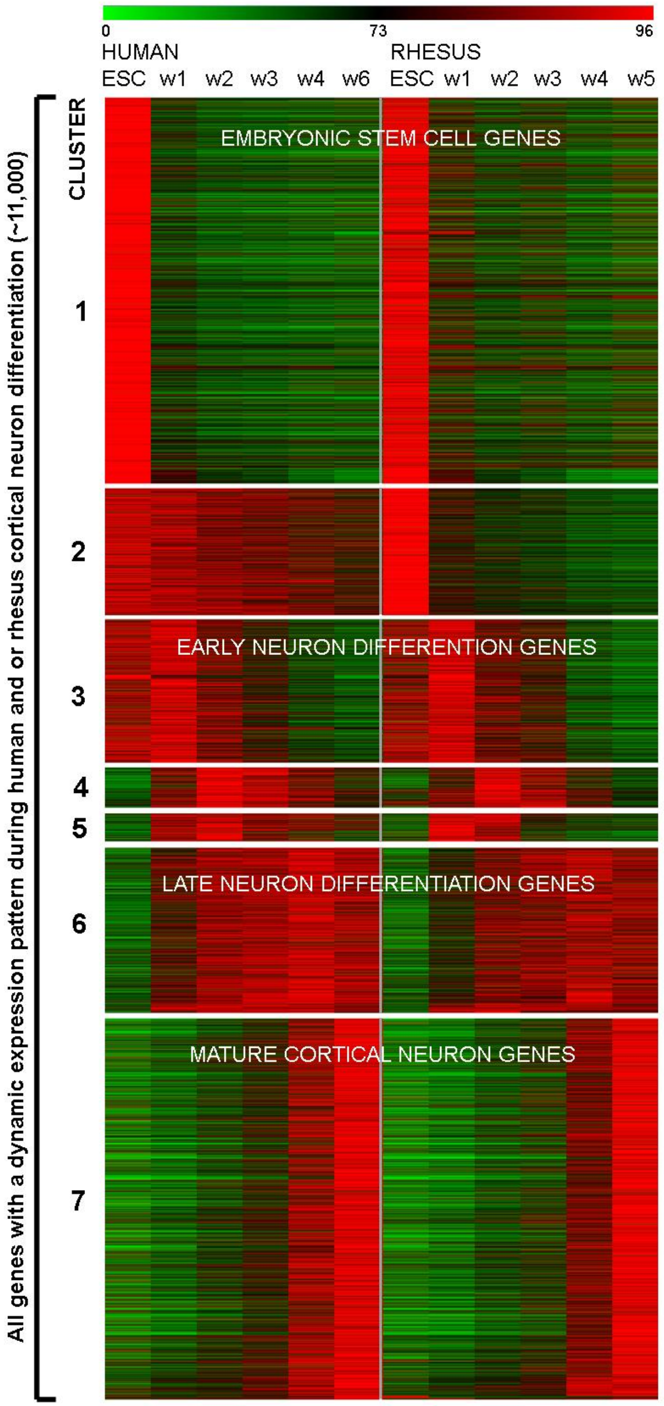
Differences in gene expression during early neural development between rhesus and human

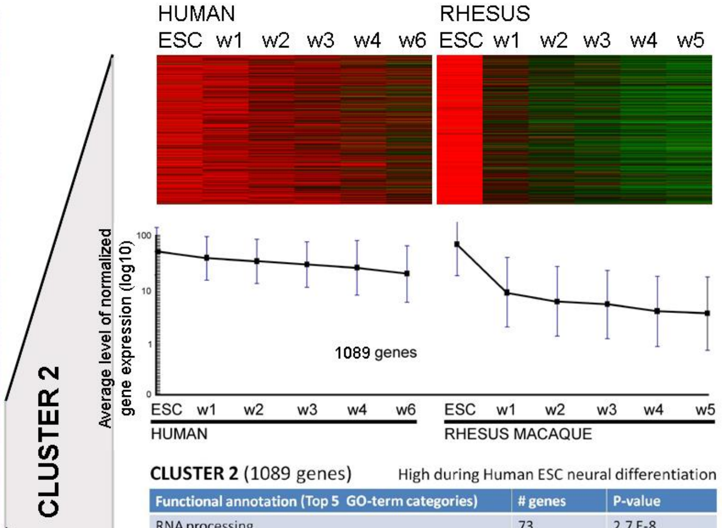


- Neural genes are defined as genes having 5 fold higher expression after neural differentiation compared to their expression in embryonic stem cells

- Between 160-300 genes are >2-fold differentially expressed between human and rhesus for each week of development

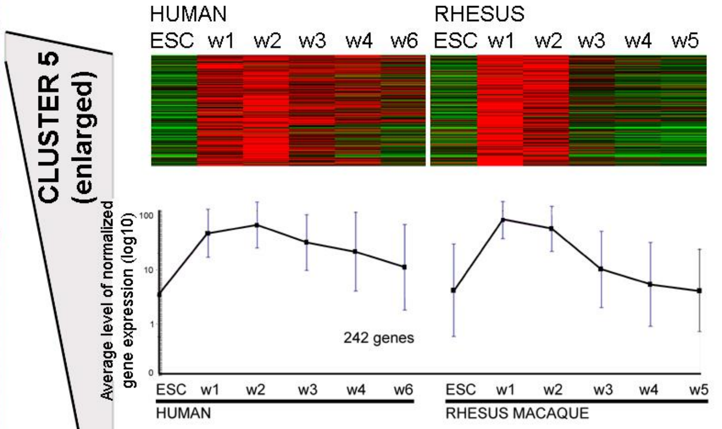
Frank Jacobs





CLUSTER 2 (1089 genes) High during Human ESC neural differentiation

Functional annotation (Top 5 GO-term categories)	# genes	P-value
RNA processing	73	2.7 E-8
Translation	49	1.8 E-6
mRNA processing	45	3.2 E-5
RNA splicing	41	4.8 E-5
mRNA metabolic process	48	7.1 E-5

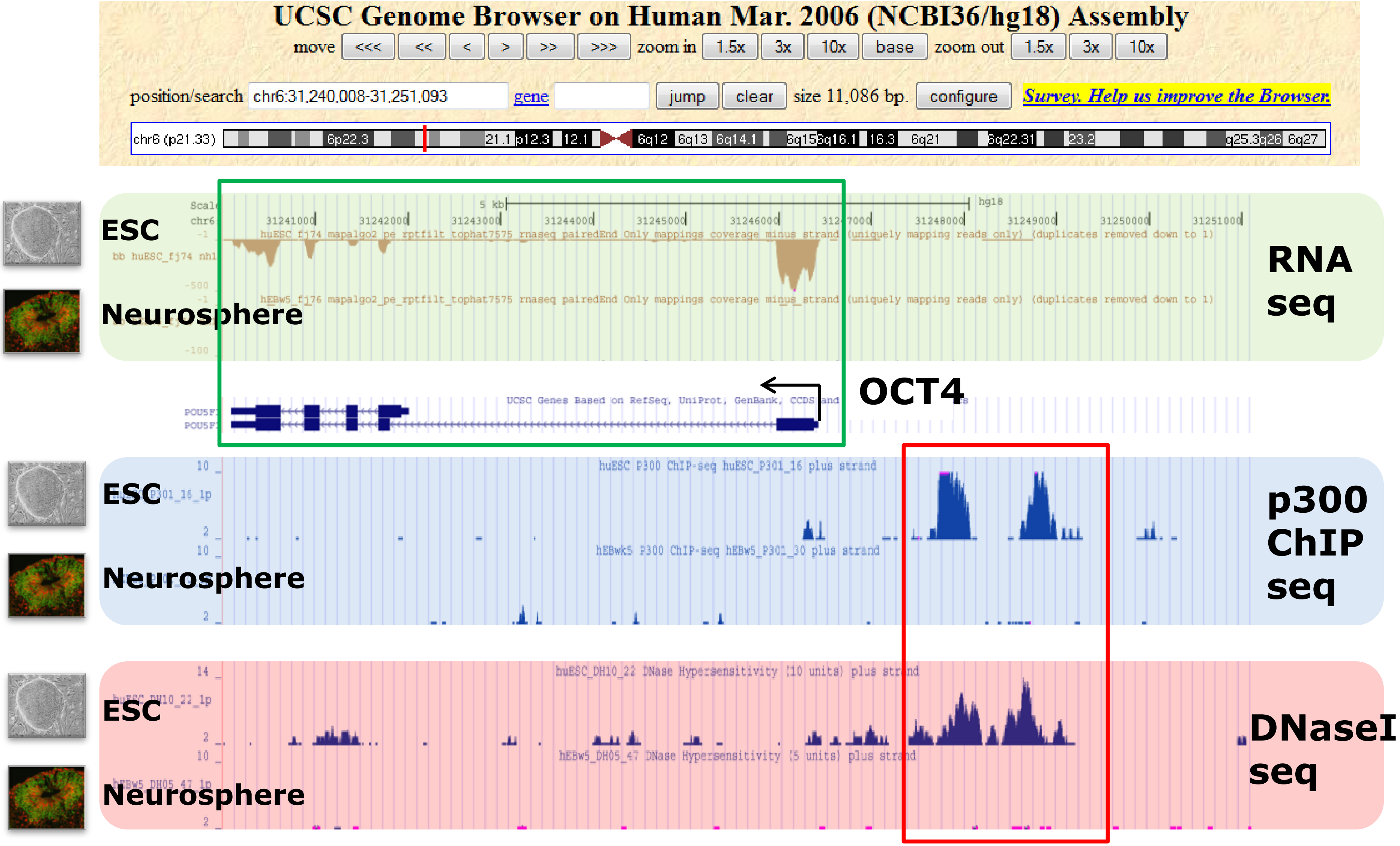


CLUSTER 5 (242 genes) Prolonged expression in human neurospheres

Functional annotation (Top 5 GO-term categories)	# genes	P-value
Regulation of cell development	11	1.8 E-4
Regulation of Nervous System Development	10	5.1 E-4
Regulation of Neurogenesis	9	8.7 E-4
Regulation of Neuron Differentiation	8	1.1 E-3
Negative Regulation of Cell Differentiation	10	1.2 E-3

Frank Jacobs

Genome-wide gene profiling by RNA-seq, ChIP-seq & DNaseI-seq



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

position/search chr6:31,240,008-31,251,093

size 11,086 bp.

huESC_fj74 mapalgo2 pe rptfilt tophat7575 rnaseq pairedEnd Only mappings coverage minus strand (uniquely mapping reads only) (duplicates removed down to 1)

hEBw5_fj76 mapalgo2 pe rptfilt tophat7575 rnaseq pairedEnd Only mappings coverage minus strand (uniquely mapping reads only) (duplicates removed down to 1)

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and

huESC P300 ChIP-seq huESC_P301_16 plus strand

hEBwk5 P300 ChIP-seq hEBw5_P301_30 plus strand

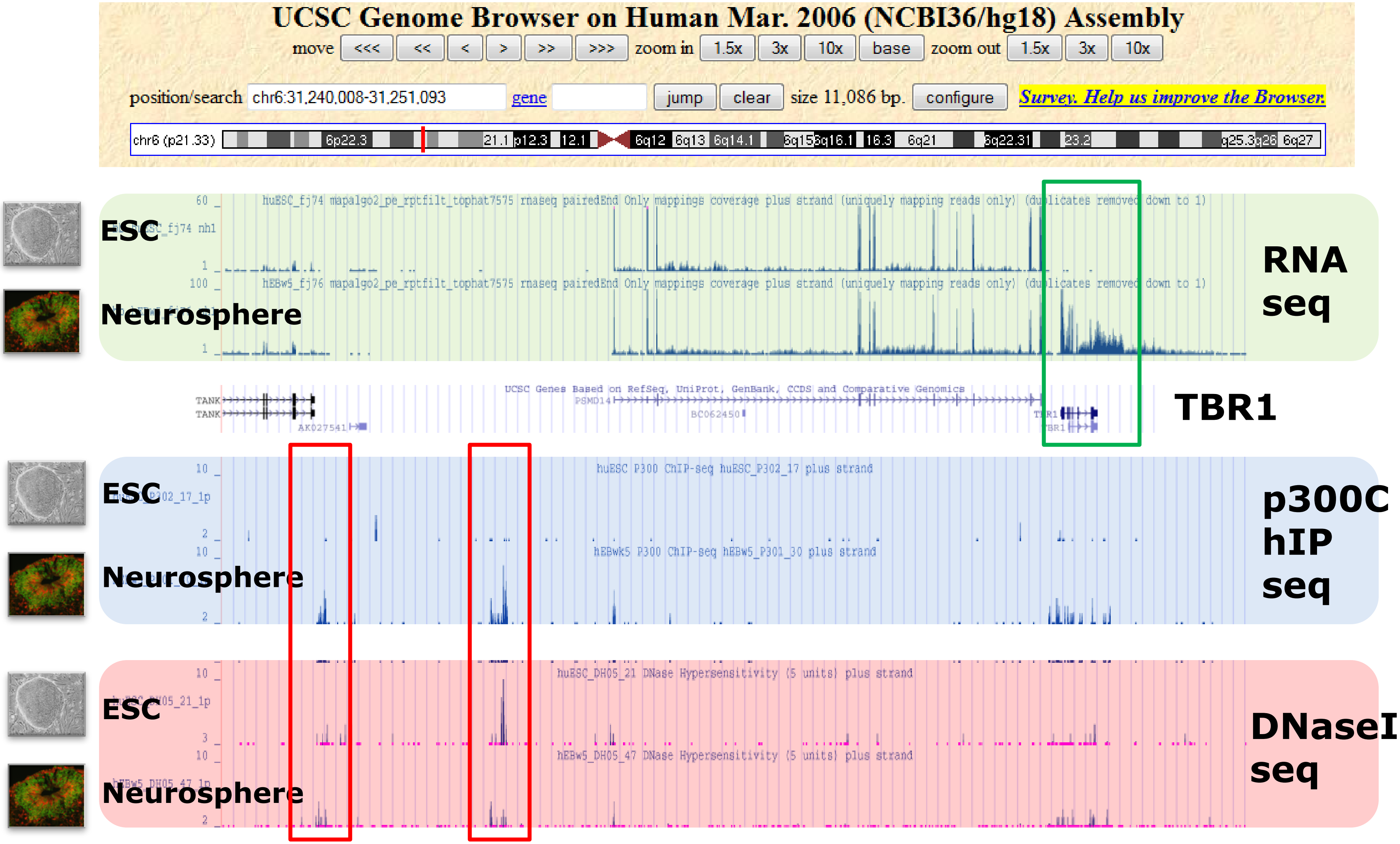
huESC_DH10_22 DNase Hypersensitivity (10 units) plus strand

hEBw5_DH05_47 DNase Hypersensitivity (5 units) plus strand

OCT4: An Embryonic Stem Cell-specific enhancer

Frank Jacobs

Genome-wide gene profiling by RNA-seq, ChIP-seq & DNaseI-seq



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

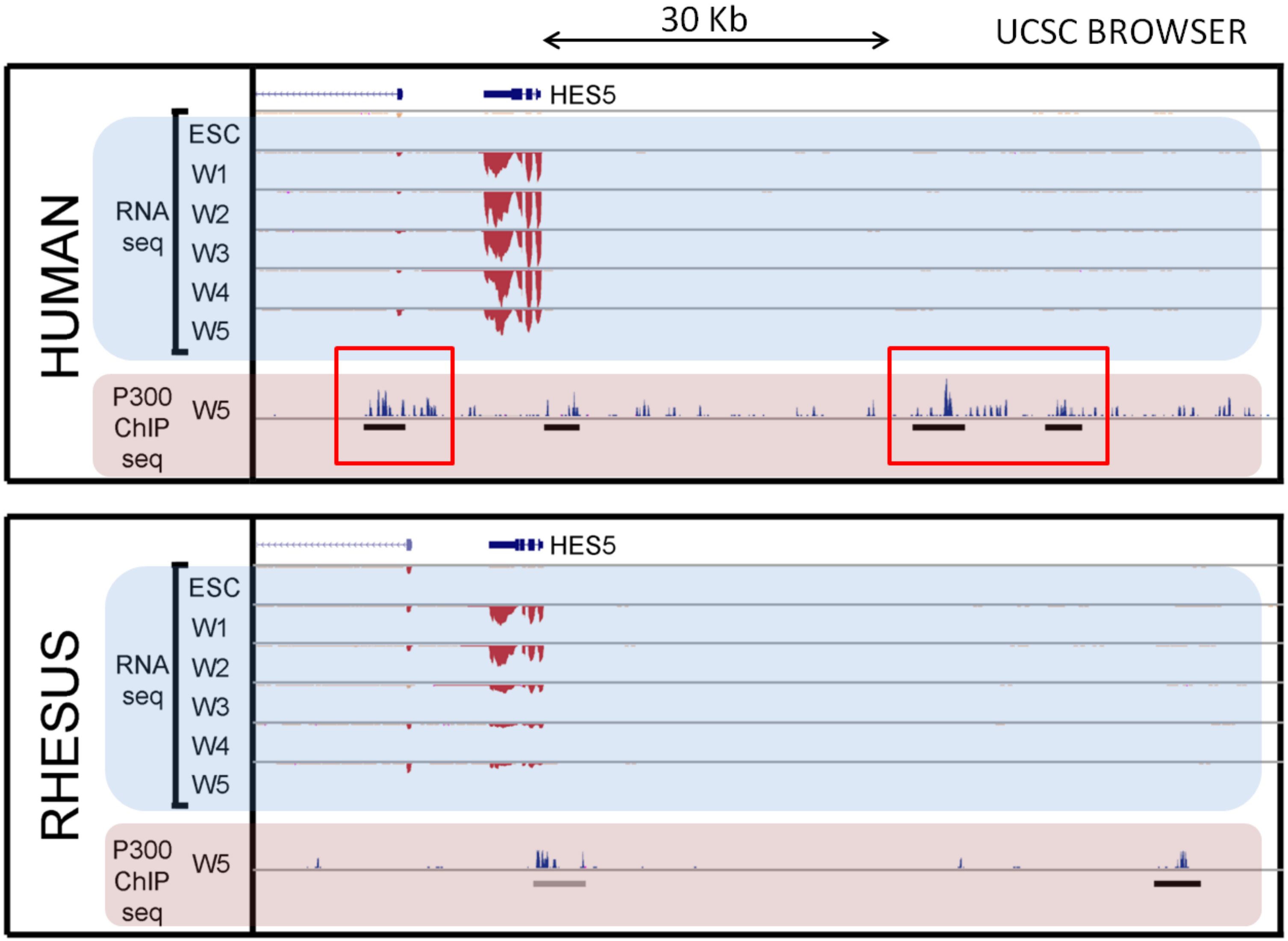
position/search chr6:31,240,008-31,251,093 gene jump clear size 11,086 bp. configure *Survey. Help us improve the Browser.*

chr6 (p21.33)
6p22.3
21.1
p12.3
12.1
6q12
6q13
6q14.1
6q15
6q16.1
16.3
6q21
6q22.31
23.2
q25.3
q26
6q27
ESC
Neurosphere
60
1
100
huESC_fj74 mapalgo2_pe_rptfilt_tophat7575 rnaseq pairedEnd Only mappings coverage plus strand (uniquely mapping reads only) (duplicates removed down to 1)
hEBw5_fj76 mapalgo2_pe_rptfilt_tophat7575 rnaseq pairedEnd Only mappings coverage plus strand (uniquely mapping reads only) (duplicates removed down to 1)
RNA seq
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
TANK
TANK
AK027541
PSMD14
BC062450
TBR1
TBR1
TBR1
ESC
Neurosphere
10
2
huESC_P300 ChIP-seq huESC_P302_17 plus strand
hEBwk5 P300 ChIP-seq hEBw5_P301_30 plus strand
p300ChIP seq
ESC
Neurosphere
10
3
2
huESC_DH05_21 DNase Hypersensitivity (5 units) plus strand
hEBw5_DH05_47 DNase Hypersensitivity (5 units) plus strand
DNaseI seq

TBR1: A Long range Cortical Neuron-specific enhancer

Frank Jacobs

HES5: differentially expressed & regulated



Frank Jacobs

General differences observed

- Increased expression of genes involved in cell proliferation during early human neurodevelopment
- Genes associated with neural differentiation are delayed in human relative to rhesus, prolonging process
- Challenging to find specific substitutions and rearrangements that account for the differences
- Once we find them, using new technology we can make selective changes in the genomes of the cells in cell culture and study the effects

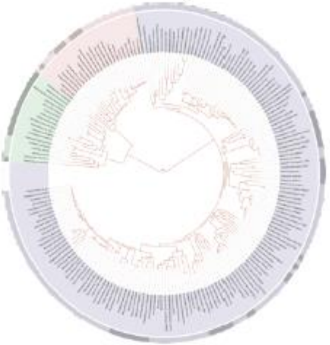
Mathematical Foundations for Comparative Genomics

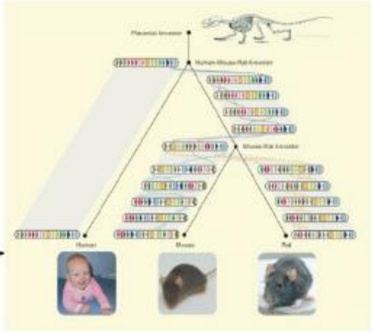
One kind of graph unifies key data structures in comparative genomics

- A graph theoretic model that allows for the trinity of **duplications**, **substitutions** and **rearrangements,** generalising many parsimony problems

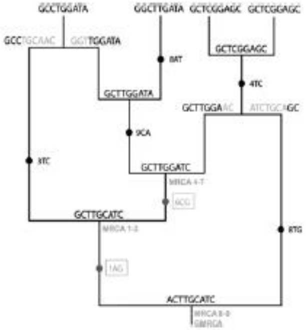
Sequence Graphs

Phylogenetic tree estimation



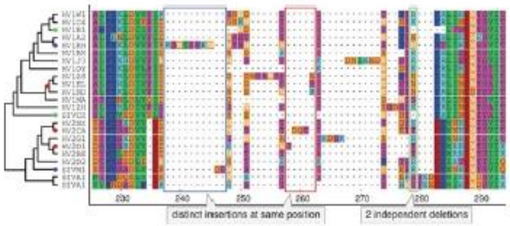


Genome rearrangement theory



Ancestral recombination graphs

Multiple sequence alignment with insertions and deletions



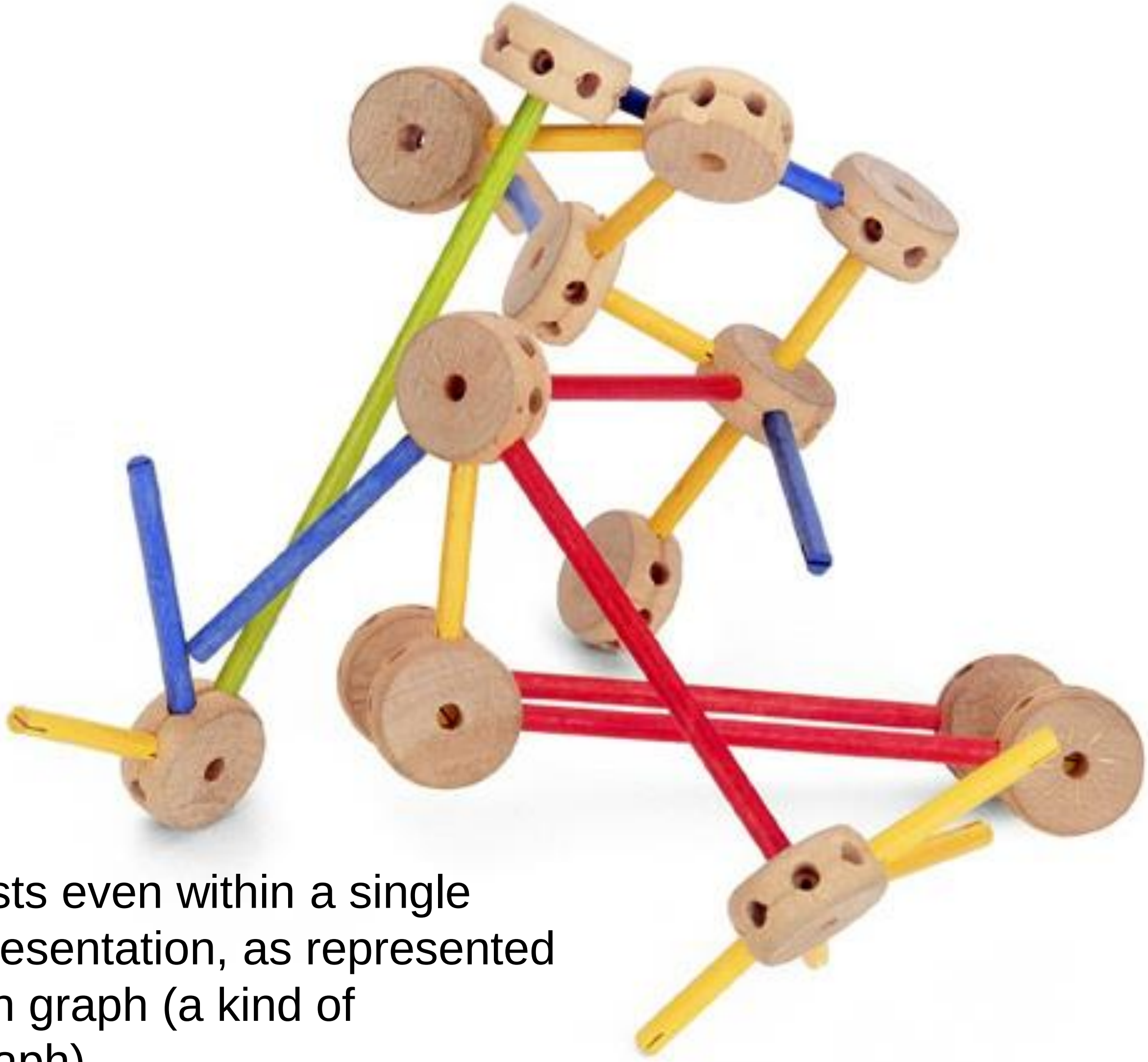
CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING
promoting discovery and invention for human health and well-being

Preprint: http://arxiv.org/abs/1303.2246

Benedict Paten

Sequence graphs are a simple construction kit to describe genome variation

Segments of DNA are attached in different ways in different genomes

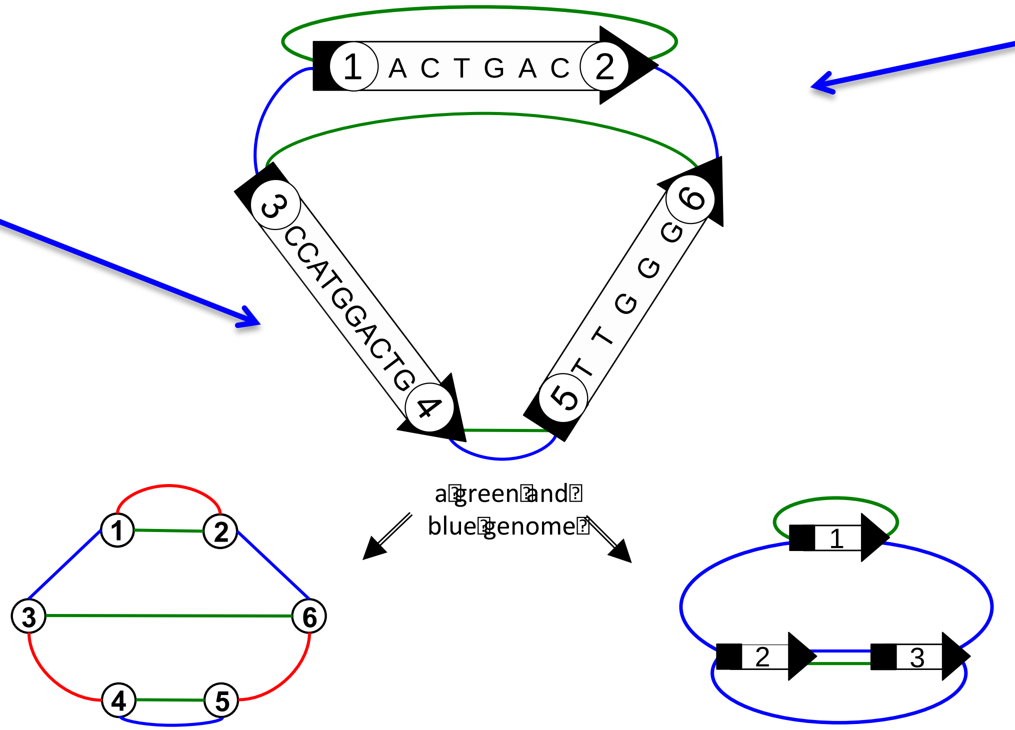


Variation exists even within a single genome representation, as represented in a De Bruijn graph (a kind of sequence graph)

Sequence graphs include both the breakpoint graph and bi-directed graph formalisms

DNA-labeled arrows are sequences

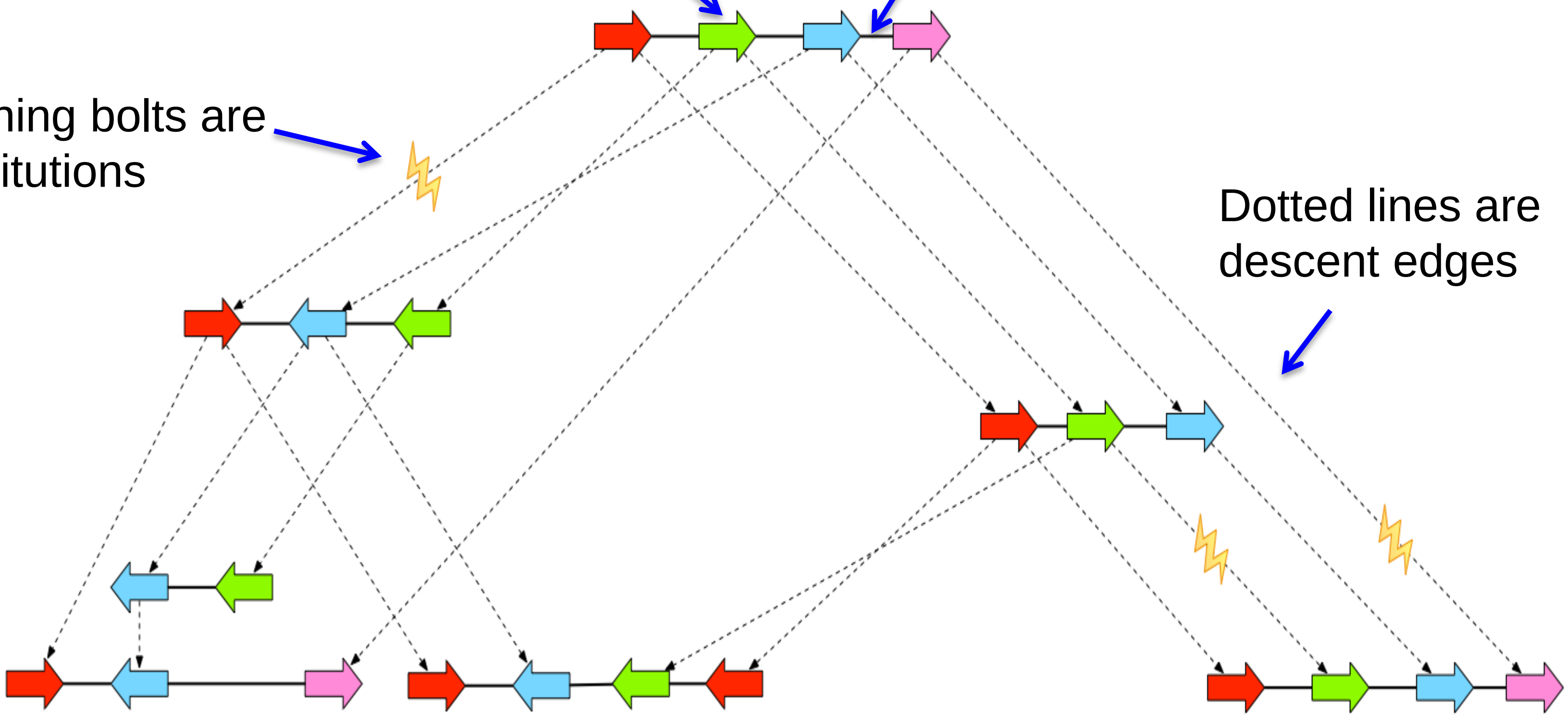
Colored lines are bonds



break point graph

bi-directed sequence graph for green and blue genomes

History graphs add descent edges to sequence graphs



Colored arrows are DNA sequences

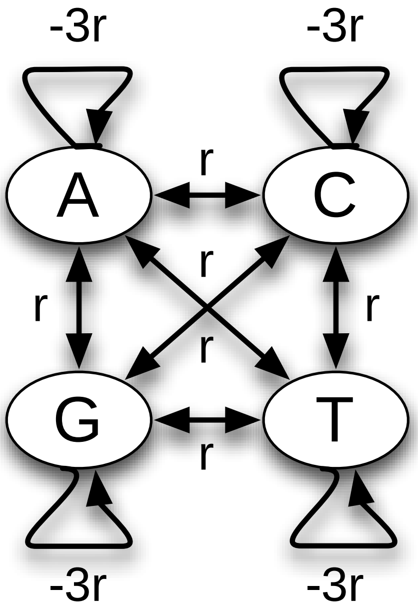
Horizontal black lines are bonds

Lightning bolts are substitutions

Dotted lines are descent edges

Benedict Paten, Preprint: http://arxiv.org/abs/1303.2246

Stochastic Models of Genome Evolution: the Jukes-Cantor model of base substitution



rate matrix

$$R = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix}$$

The probabilities of specific substitutions in time t

$$P^t = e^{Rt} = I + Rt + \frac{(Rt)^2}{2} + \frac{(Rt)^3}{6} + \ldots$$

The spectral decomposition of the rate matrix is

$$R = \beta_0 E_0 + \beta_1 E_1 + \cdots + \beta_{N-1} E_{N-1}$$

where the betas are the eigenvalues and

$$E_0, \ldots, E_{N-1}$$

are mutually orthogonal projection matrices. The probabilities of specific state changes in time t are given by the matrix

$$\begin{aligned} P^t = e^{Rt} &= I + Rt + \frac{(Rt)^2}{2} + \frac{(Rt)^3}{6} + \ldots \\ &= \sum_{d=0}^{N-1} E_d + \sum_{d=0}^{N-1} t\beta_d E_d + \sum_{d=0}^{N-1} \frac{(t\beta_d)^2}{2} E_d + \sum_{d=0}^{N-1} \frac{(t\beta_d)^3}{6} E_d + \ldots \\ &= e^{t\beta_0} E_0 + e^{t\beta_1} E_1 + \cdots + e^{t\beta_{N-1}} E_{N-1} \end{aligned}$$

For Jukes-Cantor, the eigenvalues are 0 and -4r, and the (integer-valued !) projection matrices are

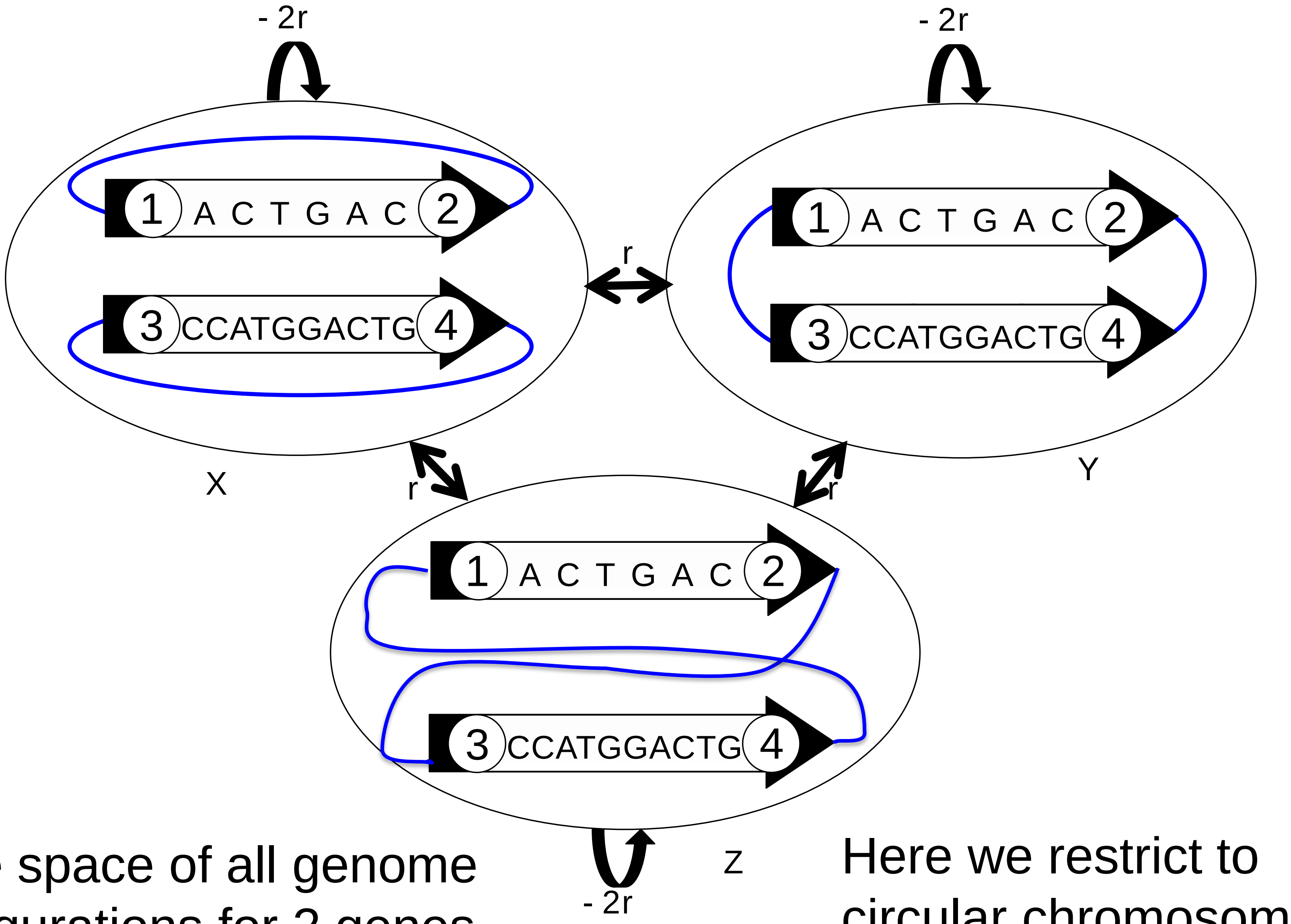
$$E_0 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad E_1 = \frac{1}{4}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

Plugging these into the general formula we get

$$P^t = E_0 + e^{-4rt}E_1$$

$$= \frac{1}{4}\begin{pmatrix} 1+3e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1+3e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1-e^{-4rt} & 1+3e^{-4rt} & 1-e^{-4rt} \\ 1-e^{-4rt} & 1-e^{-4rt} & 1-e^{-4rt} & 1+3e^{-4rt} \end{pmatrix}$$

Whole genomes change by 2-break rearrangements



State space of all genome configurations for 2 genes

Here we restrict to circular chromosomes

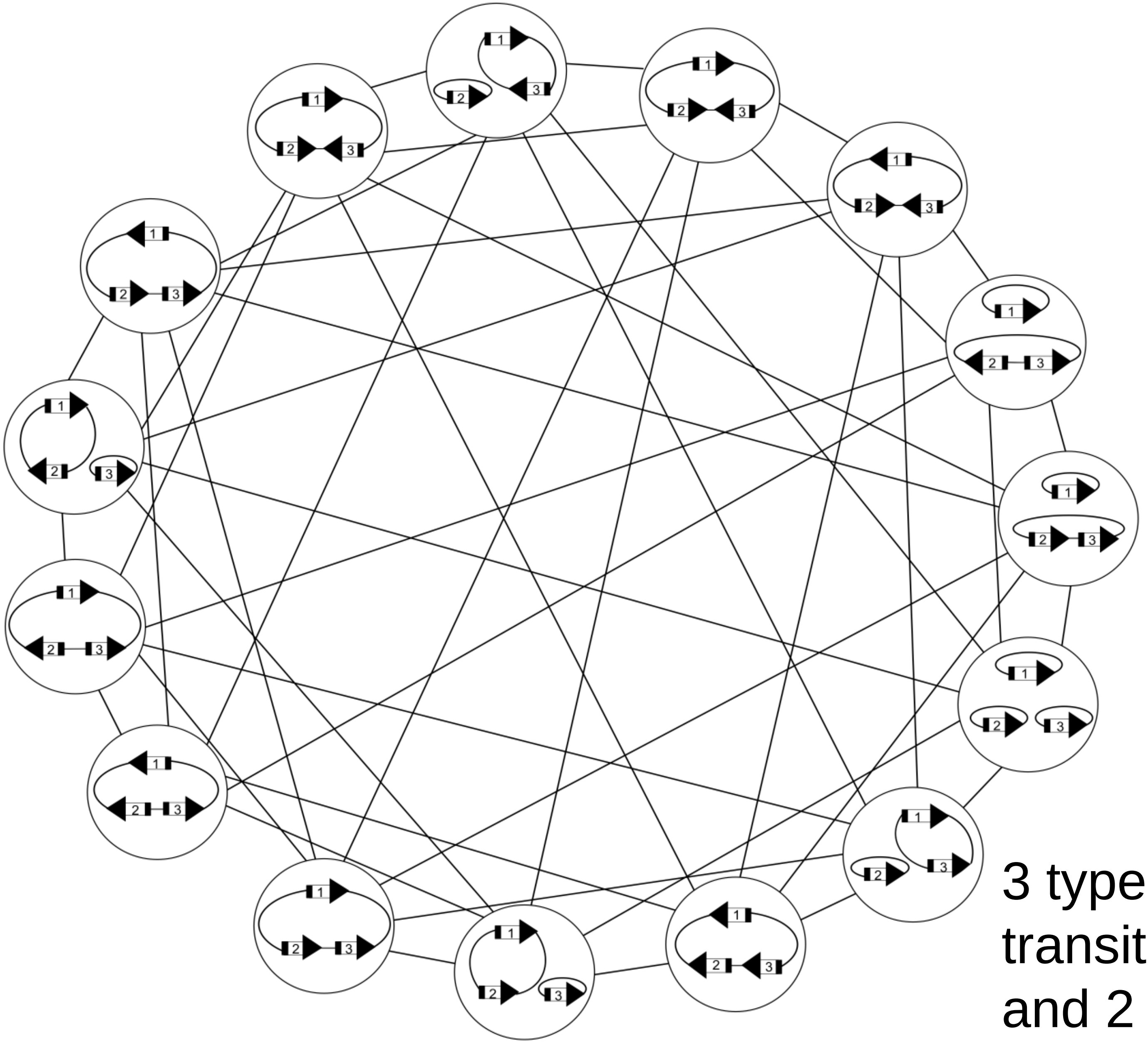
For this case of 2-gene genomes, the rate matrix for 2-break rearrangements is

$$R = \begin{pmatrix} -2r & r & r \\ r & -2r & r \\ r & r & -2r \end{pmatrix}$$

The spectral decomposition has integer-valued projection matrices like the Jukes-Cantor model, and gives

$$P_2^t = e^{Rt} = \frac{1}{3}\begin{pmatrix} 1+2e^{-3rt} & 1-e^{-3rt} & 1-e^{-3rt} \\ 1-e^{-3rt} & 1+2e^{-3rt} & 1-e^{-3rt} \\ 1-e^{-3rt} & 1-e^{-3rt} & 1+2e^{-3rt} \end{pmatrix}$$

For 3-gene genomes, there are 15 states



3 types of transitions: 0, 1 and 2 ops

For *n*-gene genomes, there are (2*n*-1)(2*n*-3) … (1) states. The general model of evolution of *n*-gene genomes by 2-break rearrangements is a random processes on matchings, explored in many areas:

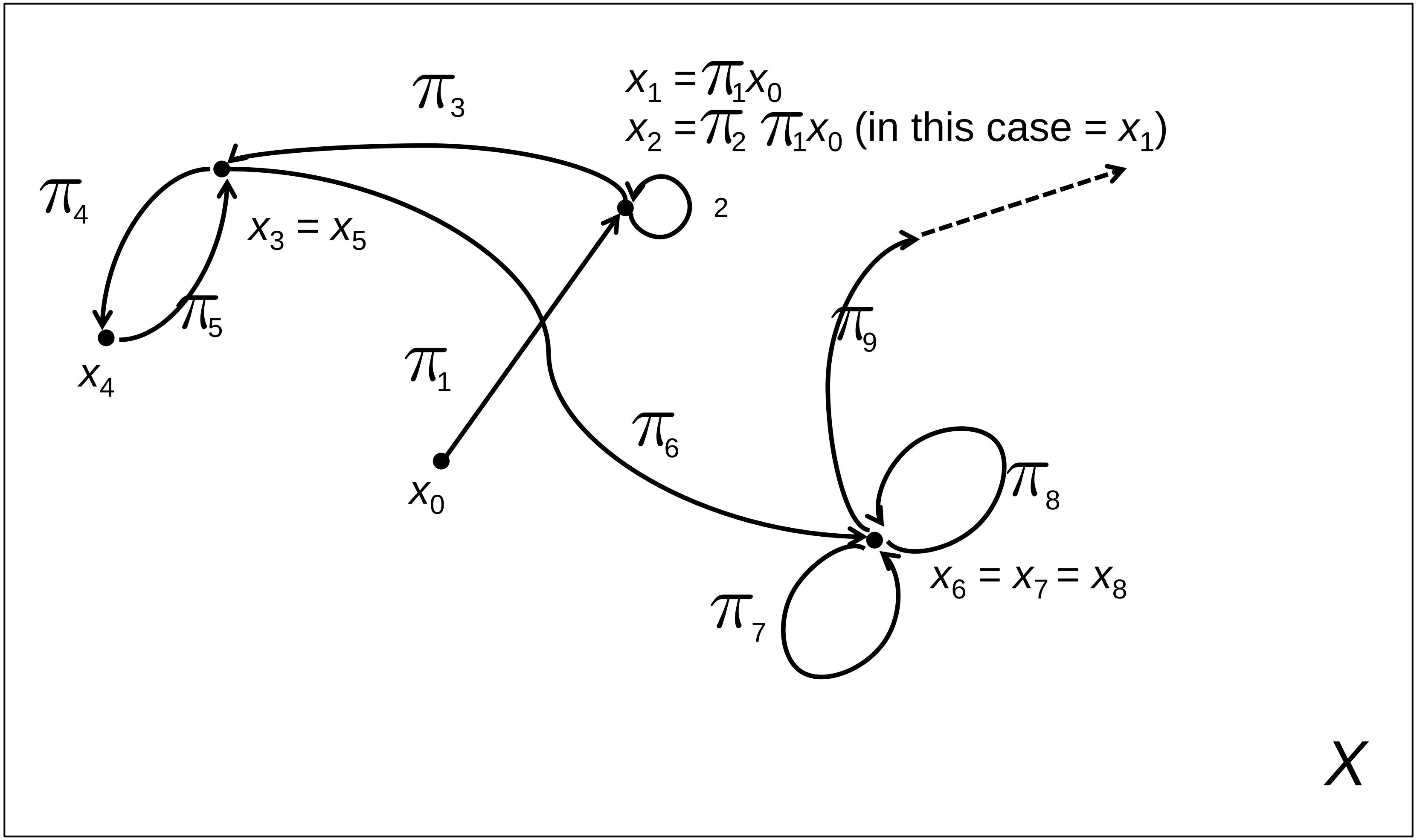
1. Diaconis and Holmes (mixing times),
2. Saxl (group representation theory),
3. MacDonald and James (symmetric functions and zonal polynomials),
4. Chillag (generalized circulants),
5. Saw and Takemura (multivariate statistics, Wishart distributions),
6. Godsil (association schemes),
7. Krieg, Bump (Hecke algebras),
8. Thrall (Lie groups).

A **homogeneous space** is a set *X* (e.g. the state space of a Markov process) and a group *G* that acts on *X*. When states are matchings on {1,2, …, 2*n*} (i.e. *n*-gene genomes), *G* is naturally the group S_{2n} of permutations of {1,2, …, 2*n*}. For a permutation π and state

$$x = \{\{i_1, i_2\}, \ldots, \{i_{2n-1}, i_{2n}\}\}$$

the action of π changes x to

$$\pi x = \{\{\pi(i_1), \pi(i_2)\}, \ldots, \{\pi(i_{2n-1}), \pi(i_{2n})\}\}$$



random walk on X by action of group G

Let the state x_0 be an arbitrary origin. The **stabilizer subgroup** $H = H_n$ is the subgroup of actions in G that leave x_0 fixed. For matchings, H is the hyperoctahedral group of symmetries of the n-cube. States in X are cosets of $G = S_{2n}$ w.r.t. H.

We write $X = G/H$. This is why

$$|X| = \frac{|S_{2n}|}{|H_n|} = \frac{(2n)!}{n!2^n} = (2n-1)(2n-3)\cdots(1)$$

In homogeneous space *X* = *G*/*H*, the group *G* acts naturally on pairs of states

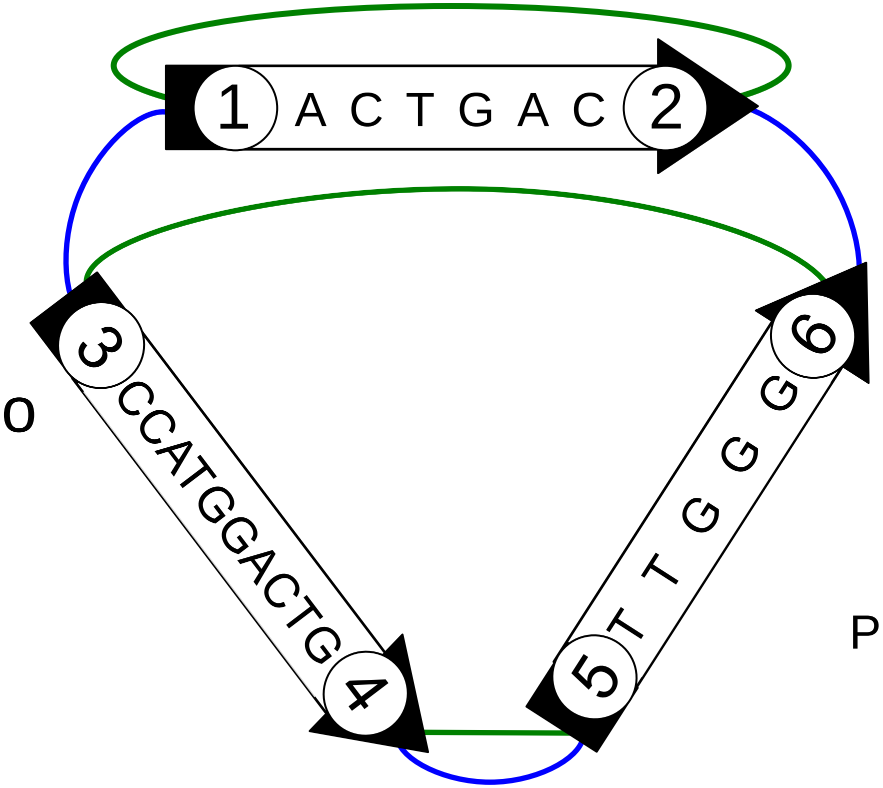
$$\pi(x, y) = (\pi x, \pi y)$$

The orbital of (x, y) is $\{(\pi x, \pi y) : \pi \in G\}$

All state pairs in the same orbital are said to have the same **difference**. Thus, each orbital defines a difference in a **difference set** *D.* In the case of the discrete Fourier space,

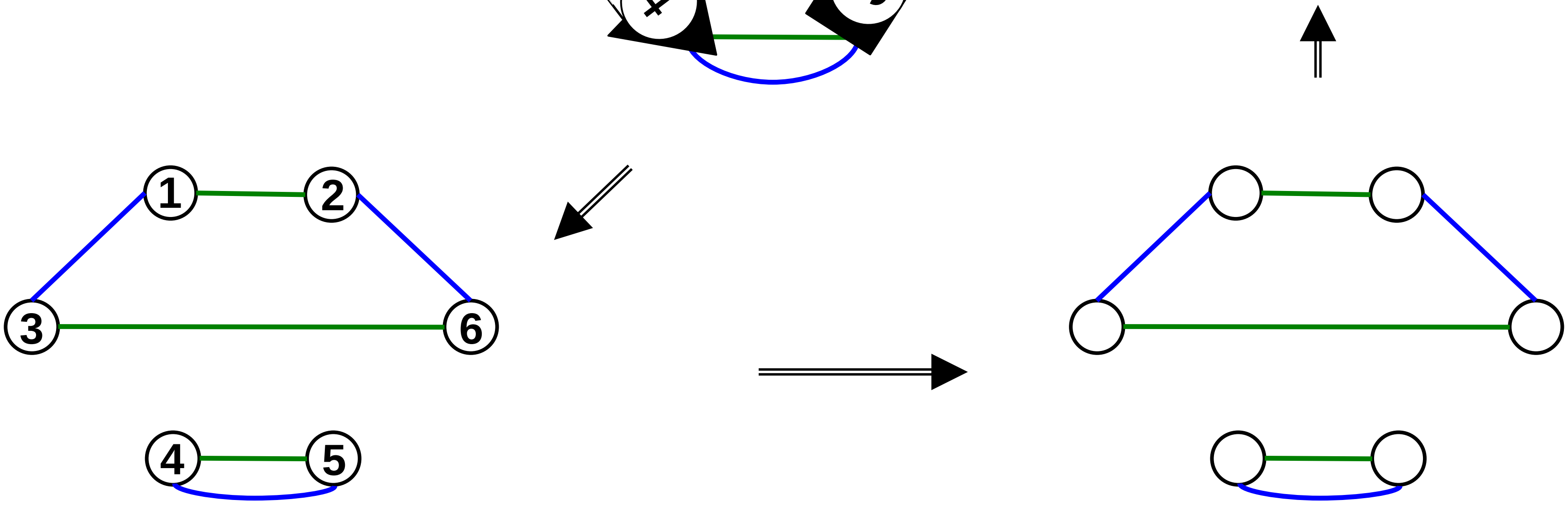
D ={-(*n*-1), -(*n*-2), …, -1, 0, 1, …, *n*-1}.

The difference between two *n*-gene genomes is a partition of the integer *n.* So *D* = set of partitions of *n*.



For example, if *n* = 3, then *D* = {(1,1,1),(2,1),(3)}.

Partition: d = (2,1)



break point graph

unlabeled breakpoint graph

In a **symmetric random walk** on *X* the probability is the same for all transitions with the same difference. The dynamics are defined by a function on the difference set *D*. The theory can be generalized to all complex functions on *D*. We call these **radial functions.** A radial function on *D* induces a unique function on *X* and *G*.

For radial functions *f* and *g*, here viewed as functions on the group *G*, we define their **convolution** as

$$(f * g)(\gamma) = \sum_{(\alpha,\beta):\gamma=\alpha\beta} f(\alpha)g(\beta)$$

This becomes the usual notion of convolving the effect of one random action followed by another when *f* and *g* are probability distributions.

A homogeneous space *X* = *G*/*H* is a **Gelfand space** if convolution of radial functions is commutative, i.e.

$$f * g = g * f$$

In this case (*G*,*H*) is said to be a **Gelfand pair**. (Same Israel Gelfand that Bernard quoted.)

The Jukes-Cantor space, the discrete Fourier space {0, …, *n*-1}, and the space of *n*-gene genomes are all Gelfand spaces.

The **Fourier Transform** is a linear mapping that Φ converts convolution into multiplication.

Think of a radial function as a |*D*|-dimensional vector. Then the Fourier transform Φ is defined by a matrix whose rows are a special orthogonal set of radial functions $\{\phi_d : d \in D\}$ called **normalized spherical functions**. The Fourier transform is written

$$\hat{f} = \Phi \bar{f}$$

where $\hat{f}$ is the Fourier transform of *f* and $\bar{f}$ is the complex conjugate of *f*. For the Fourier state space

$$\phi_d(k) = e^{i2\pi kd/n}$$

We say that the Fourier transform converts convolution into multiplication because for any radial functions *f* and *g*,

$$f * g = \sum_{d \in D} \hat{f}_d \hat{g}_d \phi_d$$

Gelfand spaces are precisely the homogneous spaces where there is a well-defined Fourier transform of the simple type we have described. There are only a few infinite families of discrete Gelfand pairs on the permutation group, so we are lucky to get one for genome rearrangements.

The spectral decomposition is associated with the inverse Fourier transform

$$f = \sum_{d \in D} \hat{f}_d \phi_d$$

The radial functions *f* and $\{\phi_d : d \in D\}$ are represented as matrices, and the Fourier coefficients $\hat{f}_d$ play the role of eigenvalues.

As an example, for the Jukes Cantor case, as |*D*|-dimensional vectors (functions on *D*), the normalized spherical functions are $(1,1)^T$ and $(3, -1)^T$. Equivalently, these can be represented by |*X*|-by-|*X*| matricies, which turn out to be the projection matrices in the spectral decomposition.

$$E_0 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad E_1 = \frac{1}{4}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

Because of the conversion of convolution to multiplication, if you convolve *f* with itself *i* times, you get

$$f * f * \cdots * f = \sum_{d \in D} \hat{f}_d^i \phi_d$$

By Taylor expansion you can get any analytical function of convolution powers, e.g. an exponential.

Thus, if *f* is taken from a radial rate matrix *R* (i.e. rate depending only on differences in *D)* and *t* is any amount of time, the matrix of probabilities of state changes over various differences is

$$P^t = \sum_{d \in D} e^{t \hat{f}_d} \phi_d$$

This generalizes the spectral decomposition method for Jukes-Cantor to a broad set of state spaces.

The Fourier transform for a general Gelfand space can be expressed as a matrix whose columns are the unnormalized spherical functions. For example, for the Jukes Cantor case, the normalized spherical functions are (1,1)^T and (3, -1)^T so the Fourier transform matrix is

$$\Phi = \begin{pmatrix} 1 & 3 \\ 1 & -1 \end{pmatrix}$$

Wonderful thing: for a Gelfand space in which the difference is symmetric, all the coefficients of the Fourier transform are integers.

For the case of n-gene genomes (matchings), the Fourier transform has an integer-valued matrix indexed by the partitions of *n*. The first few transform matrices are:

n = 2

$$\Phi = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}$$

n = 3

$$\Phi = \begin{pmatrix} 1 & 6 & 8 \\ 1 & 1 & -2 \\ 1 & -3 & 2 \end{pmatrix}$$

n = 4

$$\Phi = \begin{pmatrix} 1 & 12 & 12 & 32 & 48 \\ 1 & 5 & -2 & 4 & -8 \\ 1 & 2 & 7 & -8 & -2 \\ 1 & -1 & -2 & -2 & 4 \\ 1 & -6 & 3 & 8 & -6 \end{pmatrix}$$

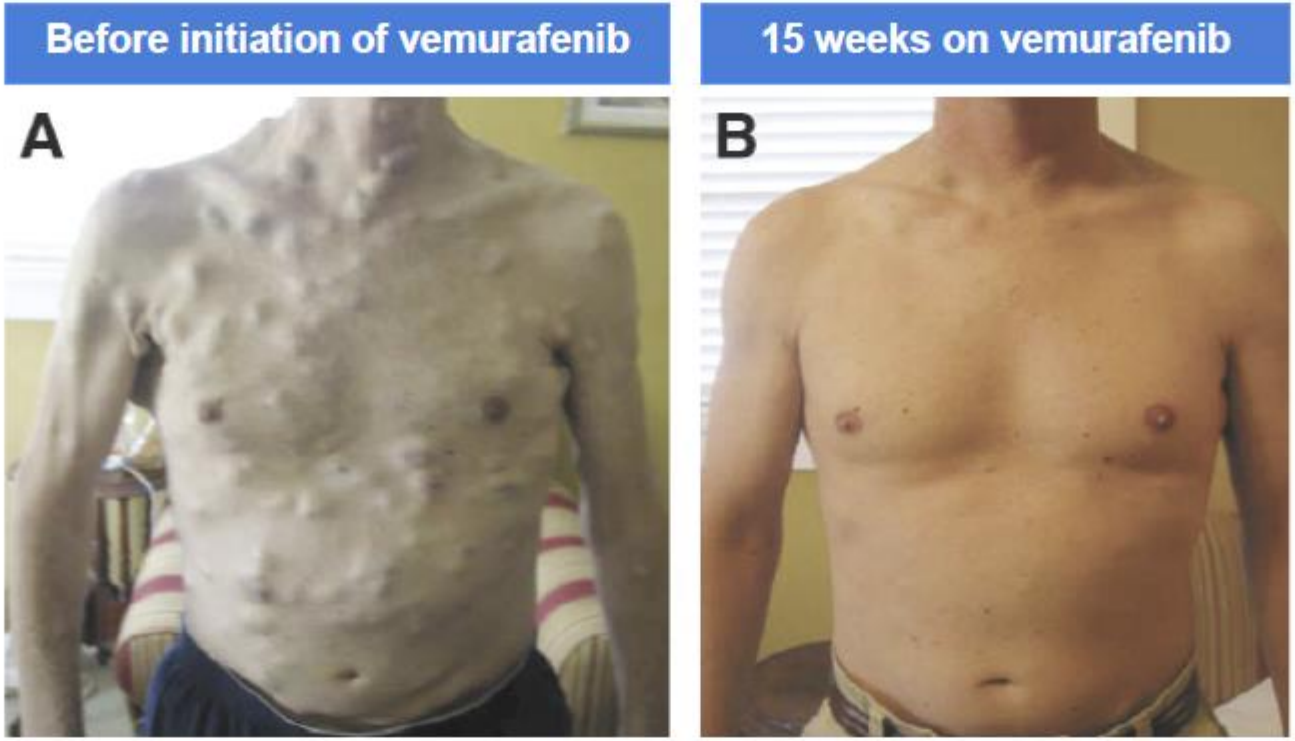
- There is no known computationally tractable closed-form formula for the integers in the Fourier transform matrix for matchings.

- Nevertheless, genome evolution by 2-break rearrangements is a special case of an extensive and beautiful theory (symmetric Gelfand spaces)

- Including duplications, gains and losses complicates the model considerably

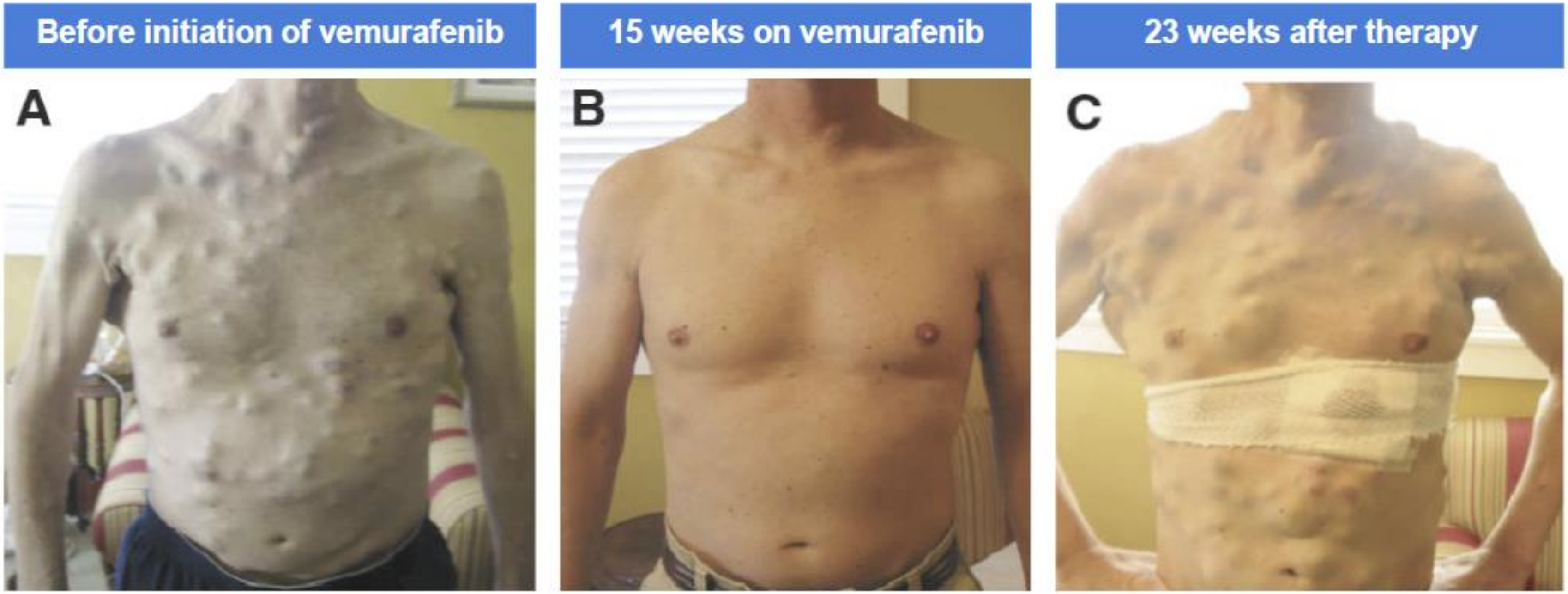
Comparative Genomics in Cancer

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



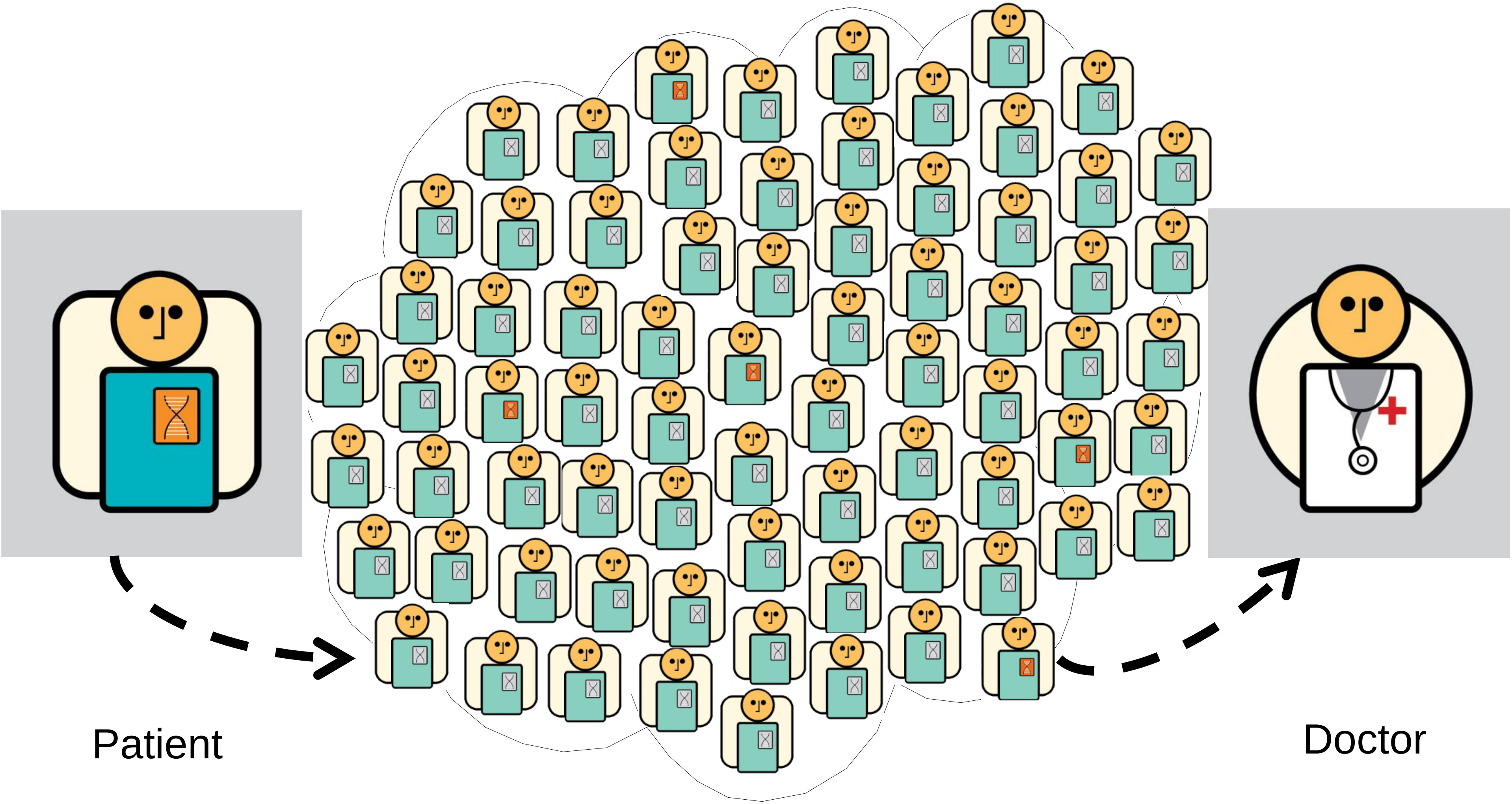
Roche

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



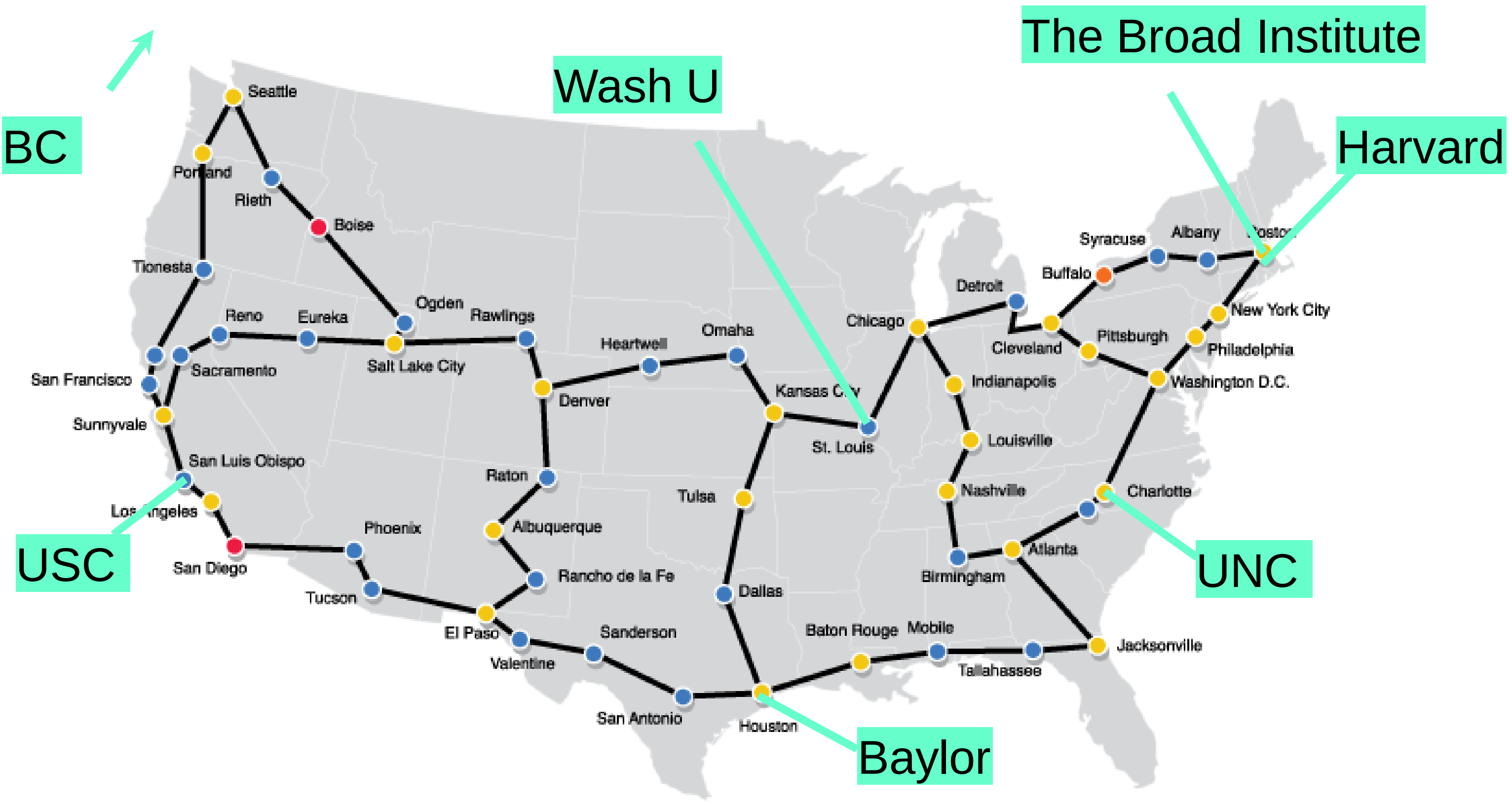
Patient

Doctor

Cancer Genome Database

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

TCGA Sequencing Centers



BC

Wash U

The Broad Institute

Harvard

USC

UNC

Baylor

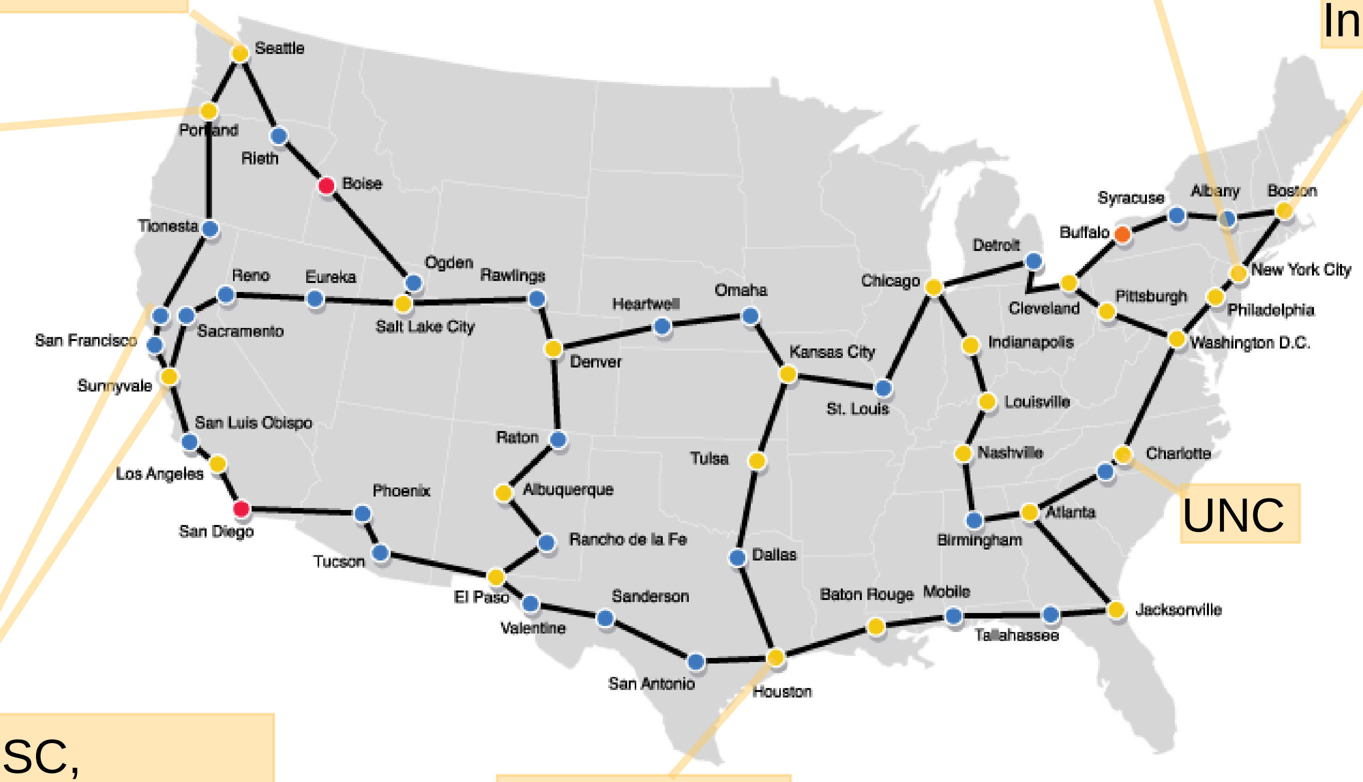
TCGA Analysis Centers

Institute for Systems Biology, Seattle

Oregon H&S U

MSK Cancer Ctr

The Broad Institute

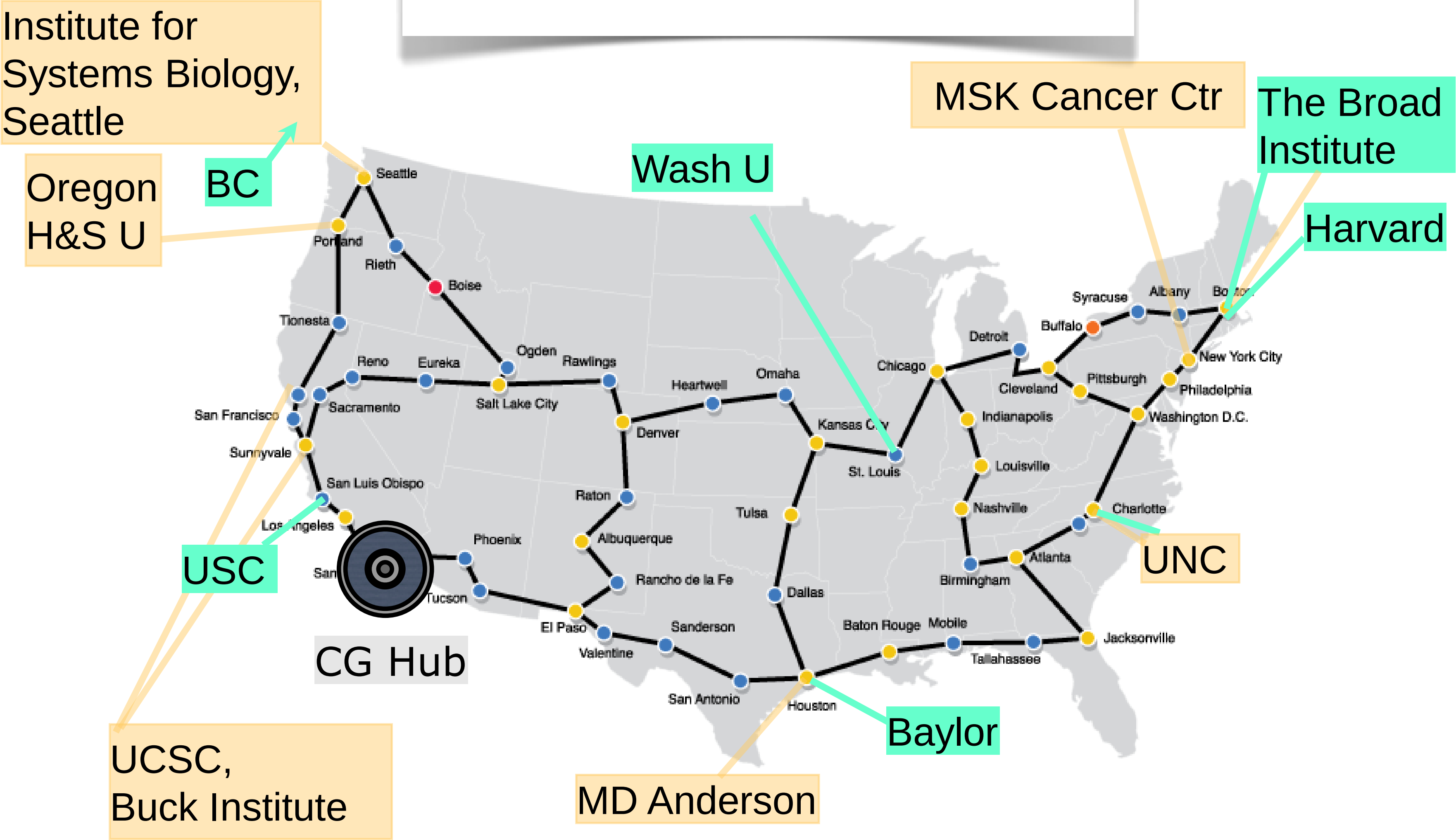


UNC

UCSC,
Buck Institute

MD Anderson

The Cancer Genomics Hub



Institute for Systems Biology, Seattle

BC

Oregon H&S U

Wash U

MSK Cancer Ctr

The Broad Institute

Harvard

USC

CG Hub

UNC

Baylor

UCSC, Buck Institute

MD Anderson

CANCER 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**

CGHub at San Diego Supercomputer Center

CGHub group

Current Stats

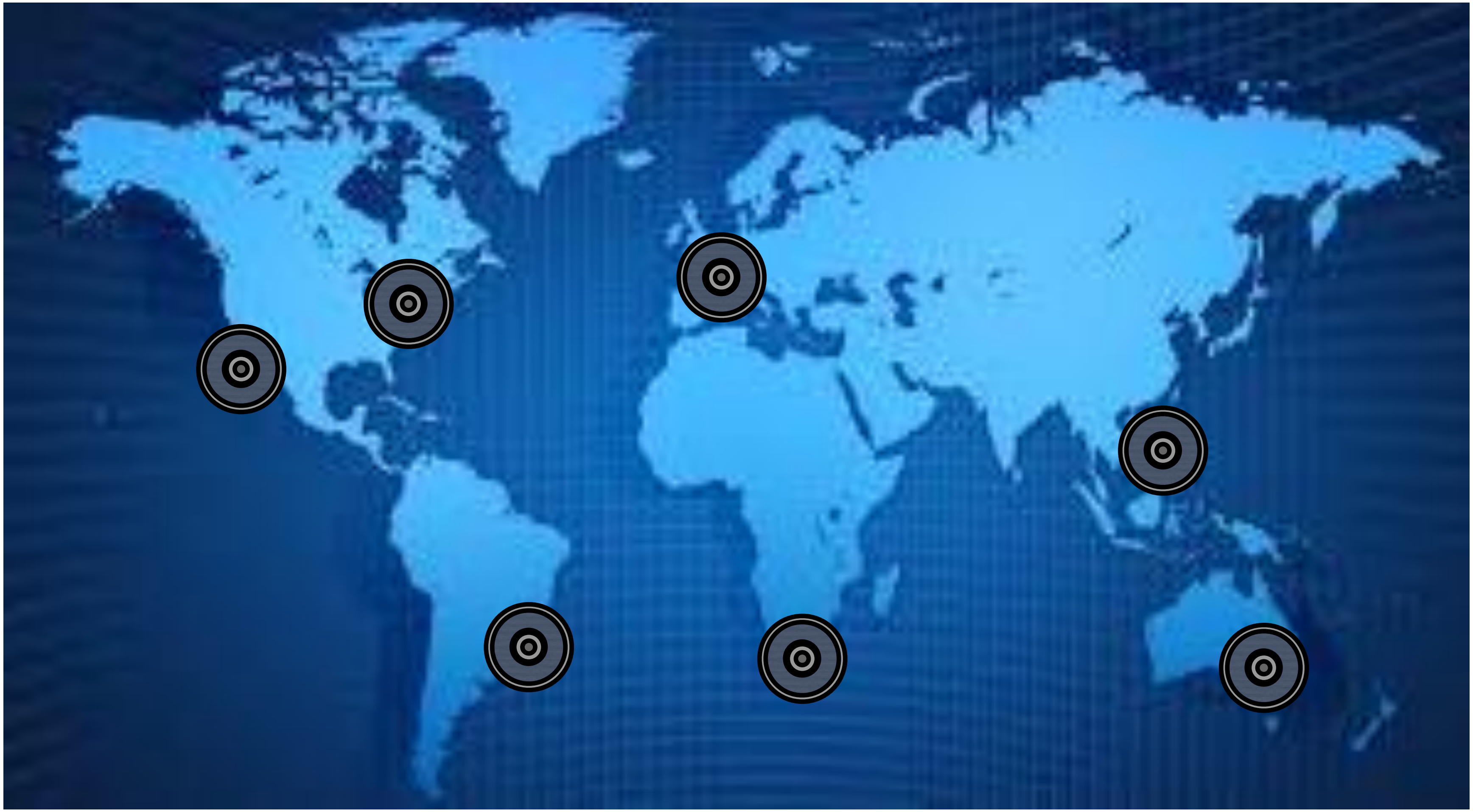
716,000 total files downloaded

10,462 TB transferred

495 TB data
43,000 files

2-4 Gb/s typical downloads in aggregate outbound from CGHub

Future Requires Global Network of Hubs

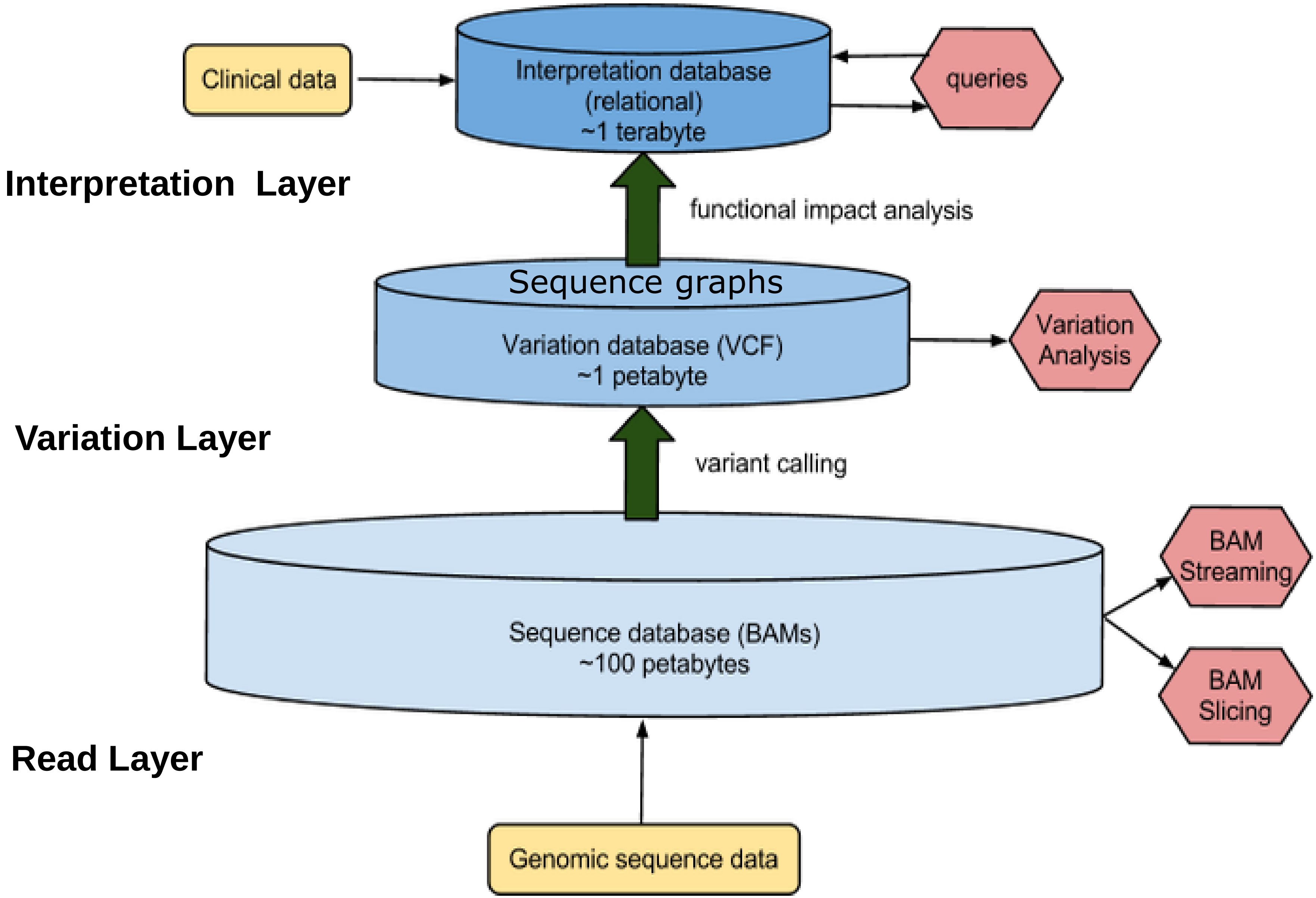


Different Requirements for 1M Genomes

- 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



Clinical data

Interpretation database (relational) ~1 terabyte

queries

Interpretation Layer

functional impact analysis

Sequence graphs

Variation database (VCF) ~1 petabyte

Variation Analysis

Variation Layer

variant calling

Sequence database (BAMs) ~100 petabytes

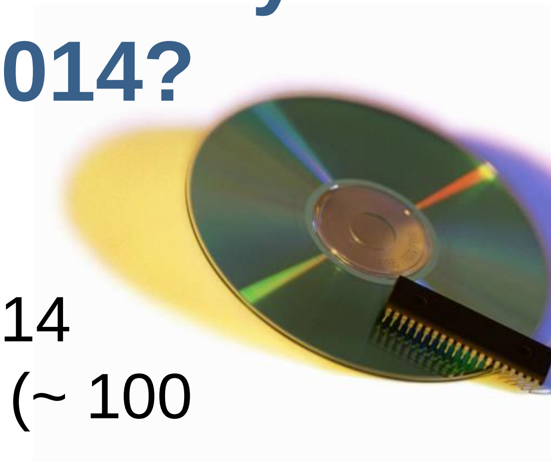
BAM Streaming

BAM Slicing

Read Layer

Genomic sequence data

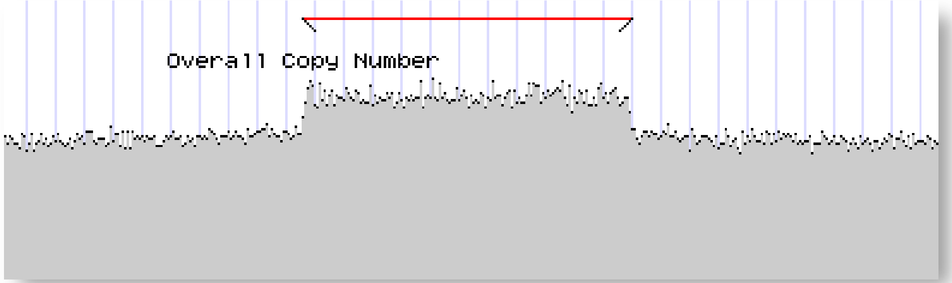
What would it cost to store and analyze 1M Cancer Genomes in 2014?



- 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

Extracting molecular state from raw DNA reads



chr2 : 29,064,107

OV-0751 Somatic Reads

```
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACAggagtattaaccccacctgatctcacgatgggagaggagacgcca

ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCC
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCG
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAAAGGAGGC
                    TATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCG
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGGCCACAGAGGTCA
    CTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCCTCTCCTCCATCTCCCATCG
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAAAGGAGGCC
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCCGAGGGCATCTCCTCCATCTCCCAC
      GGCTGGCTGCACCCTATAATGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCTCCCATCG
                   CTAGATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCCTCAGAGGGCATCACCTCCACTTCCCATCG
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCATCAGAGGGCATCTCCTCCATCCCCCATCC
             TGCACCCTATATTGTCTGAGAACAGAGTGGCTACACAGAAAATGGAGGCCCACAAAGGGCCACTTCCCCACCTCCCCTCC

               cactttctacagacgatgtcaccttccacctCACAGAAAATGGAGGCCATCAGAGGGCATCTCCtccatctcccatcg
```

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

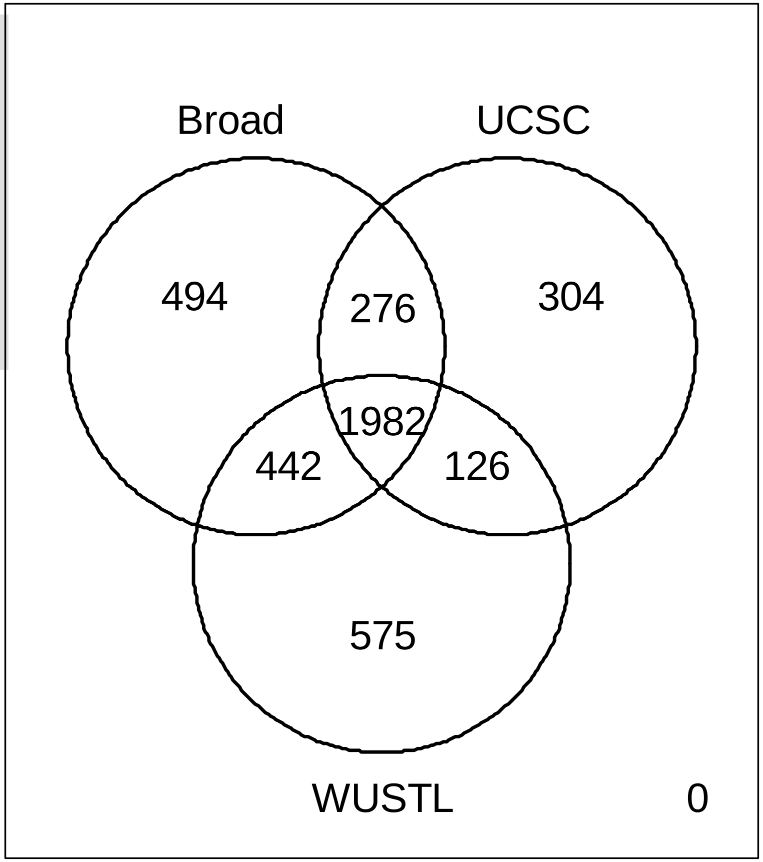
Point mutations called in tumor TCGA-13-0725

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%

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

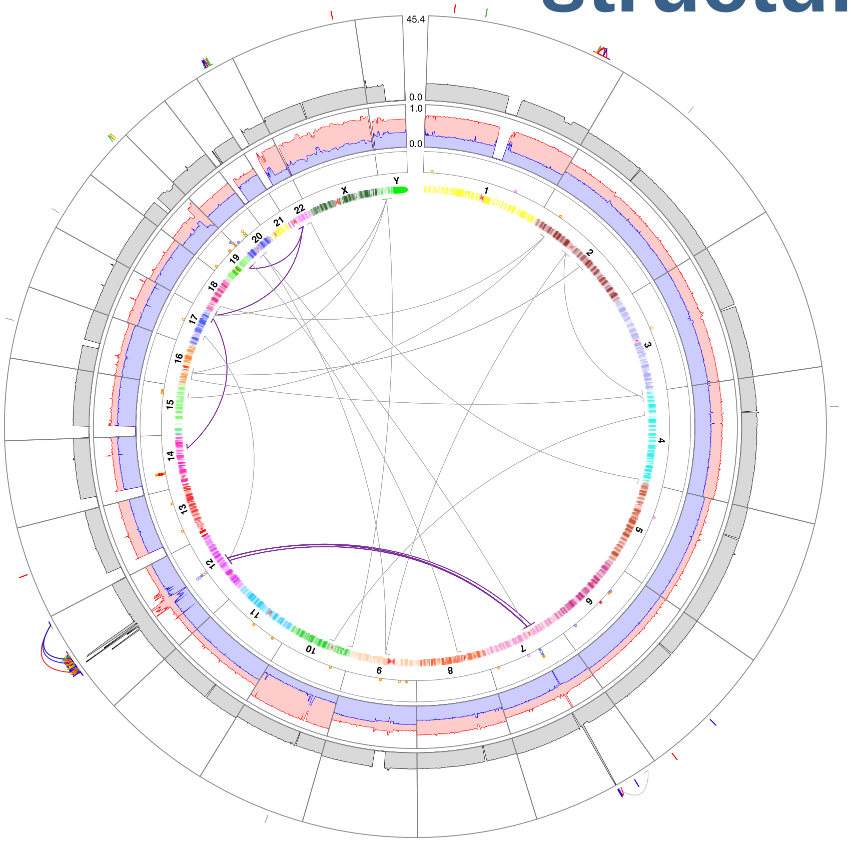
UCSC, Broad are leading a series of TCGA/ICGC international benchmark challenges. Visit cghub.ucsc.edu for TCGA Benchmark 4

TCGA-13-0725_

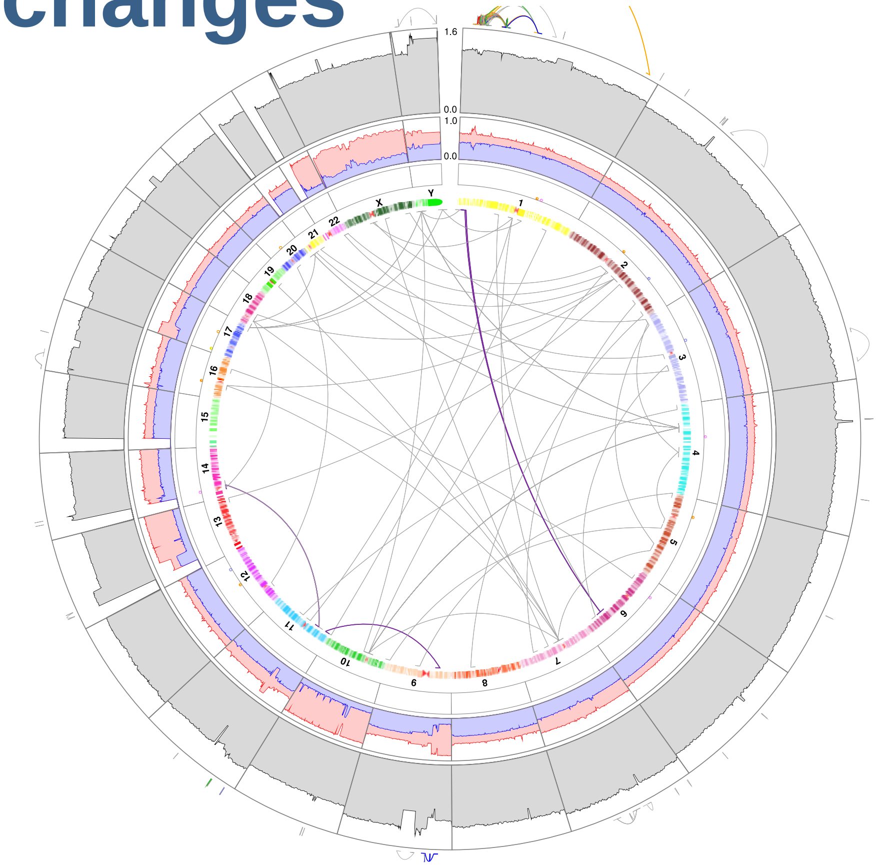


Singer Ma

Even more differences in calling structural changes



06-0152



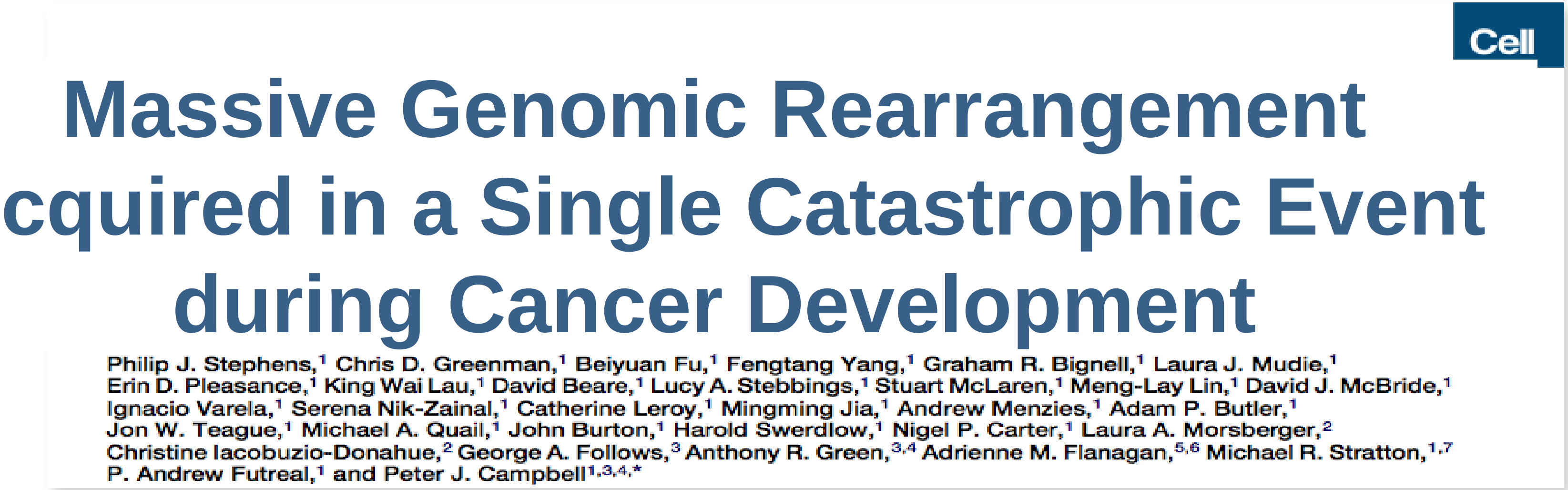
06-0188

- 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

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

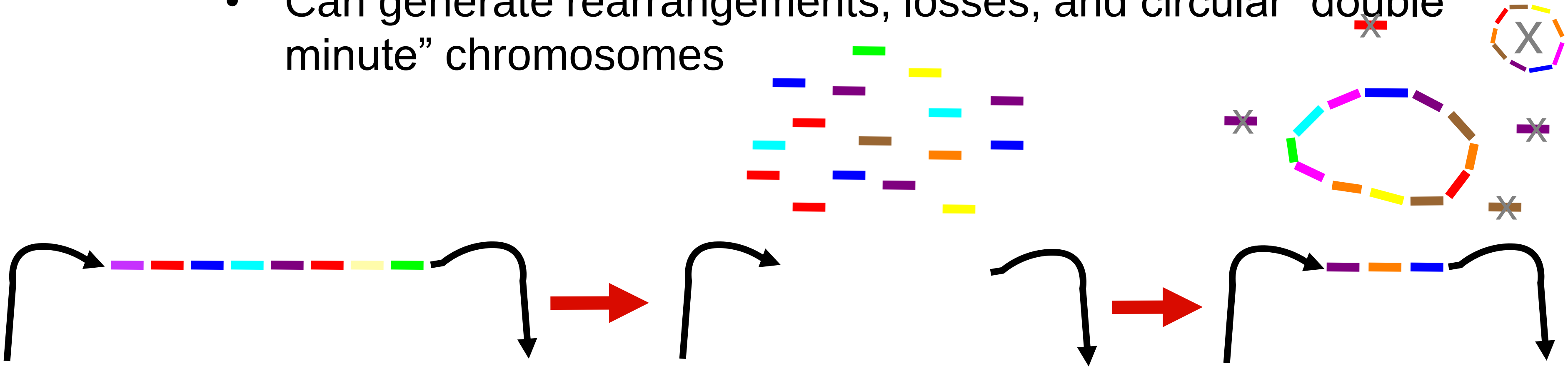


Cell

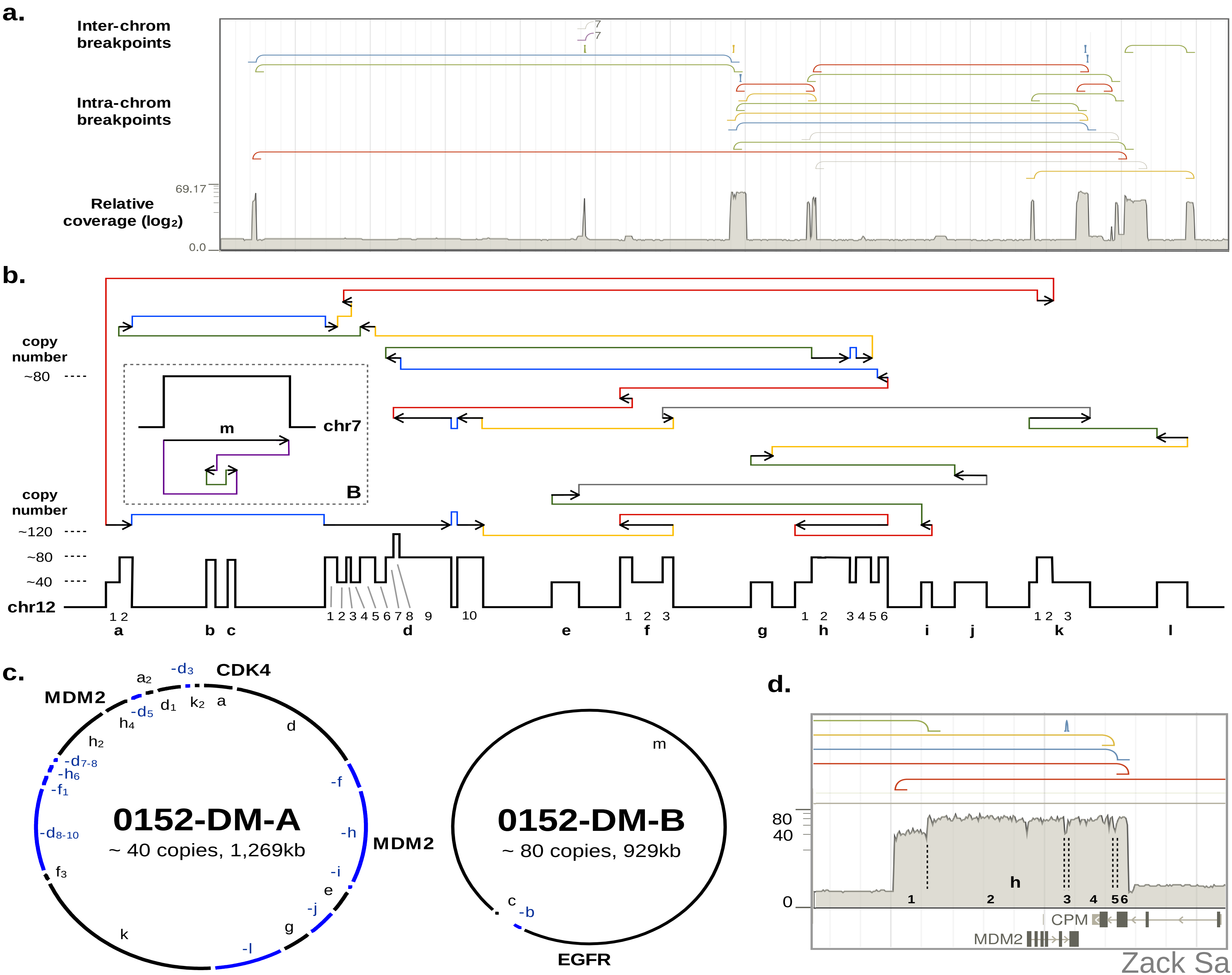
Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development

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

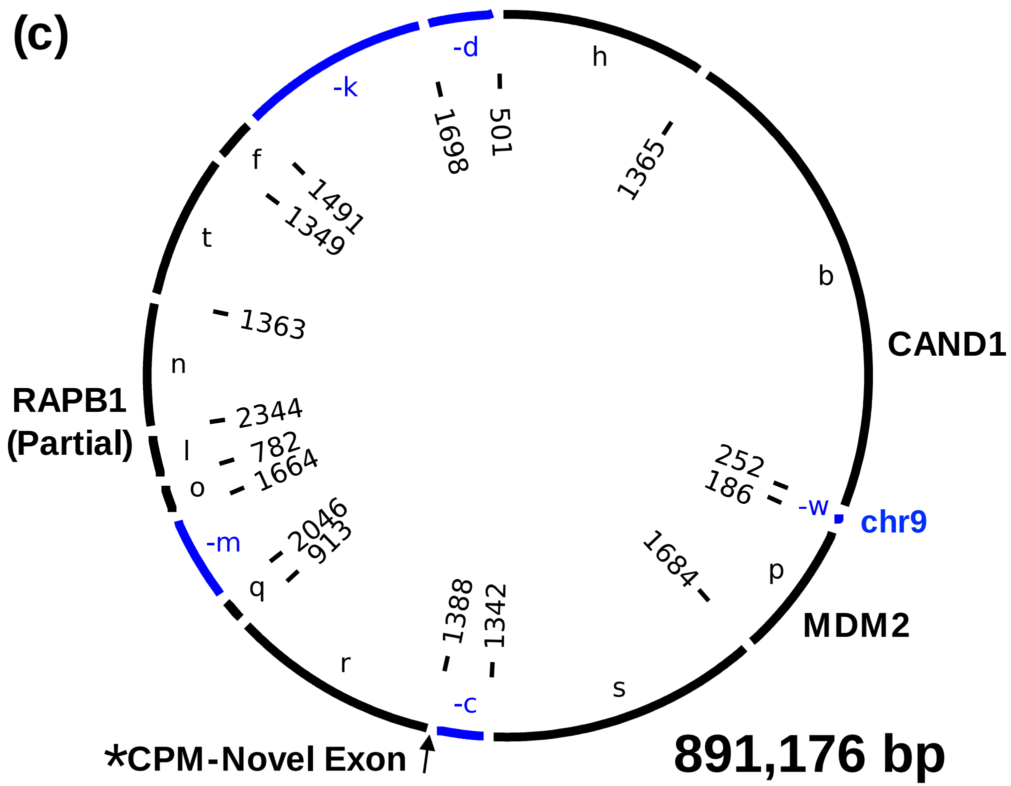


Zack Sanborn

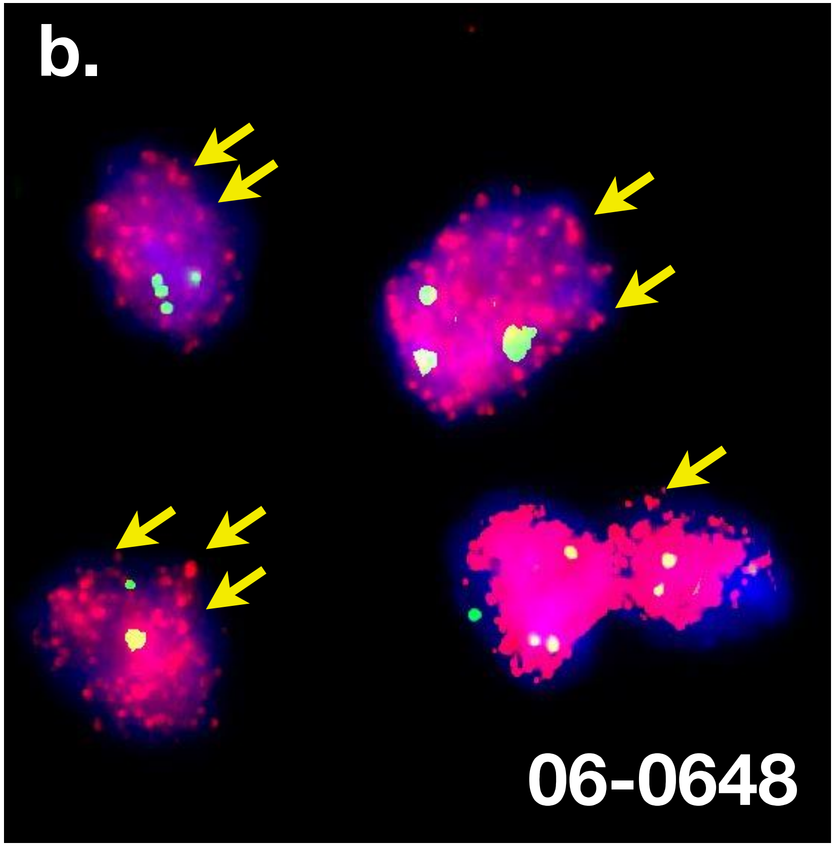


Zack Sanborn

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

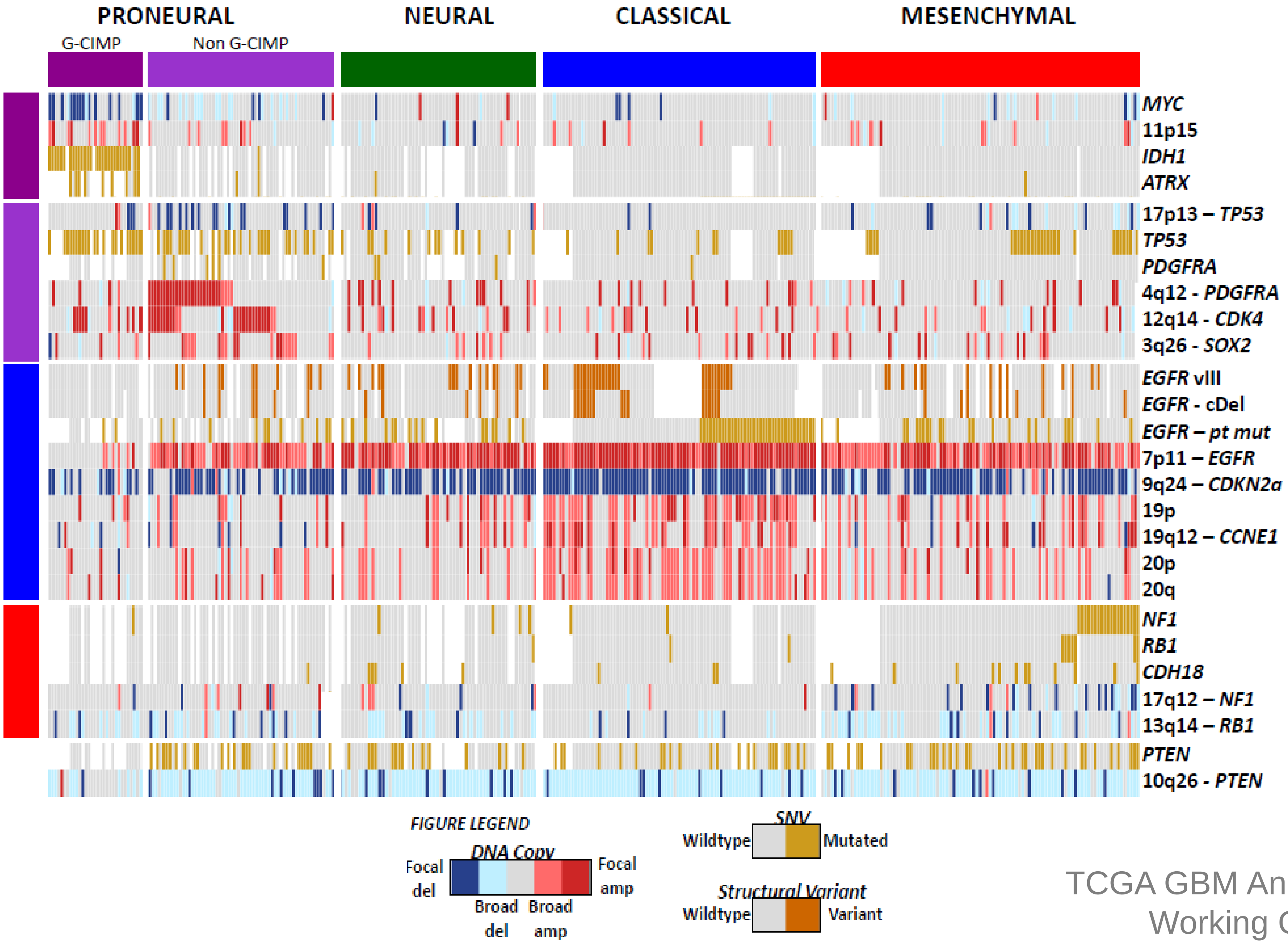


Validation by FISH



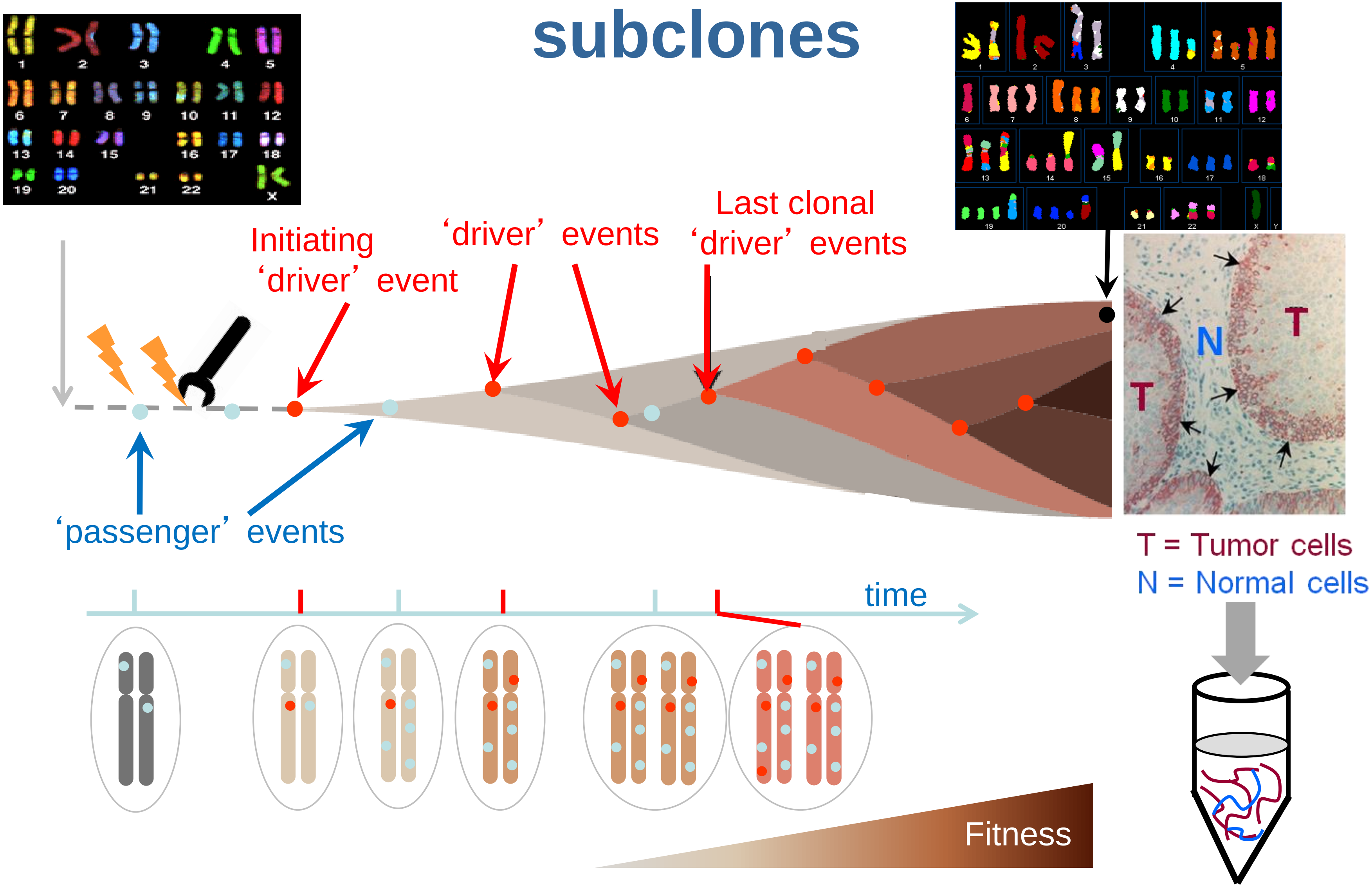
Zack Sanborn, Cameron Brennan

Highlights from analysis of 500 GBMs



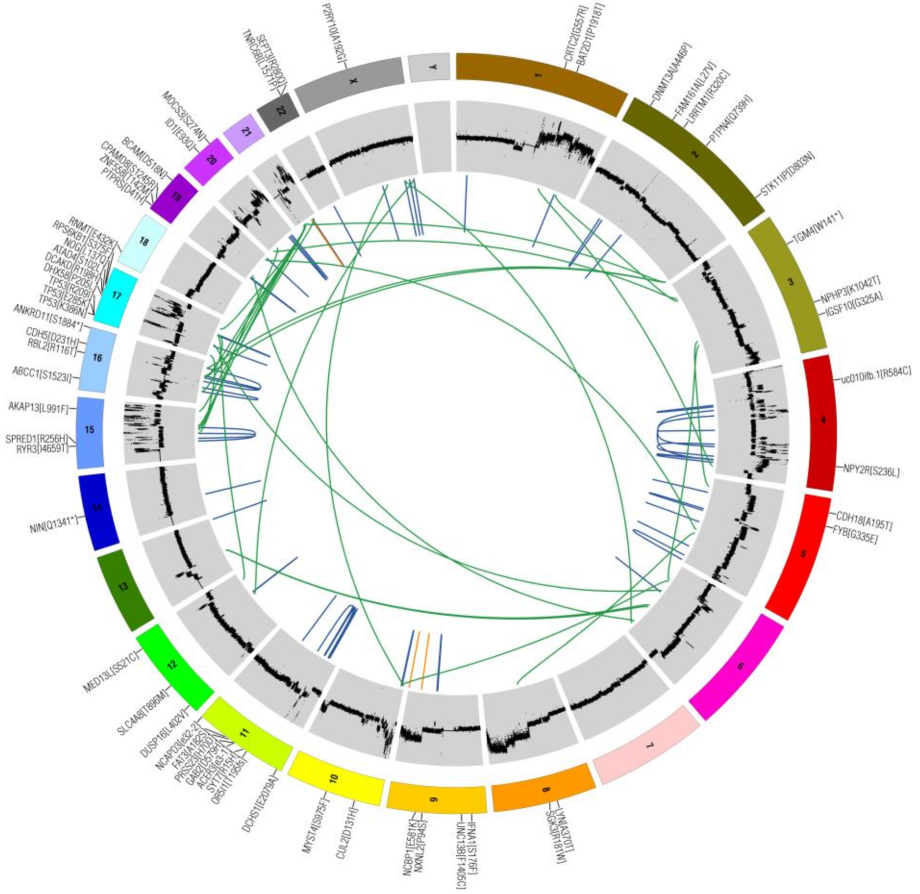
TCGA GBM Analysis Working Group

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



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

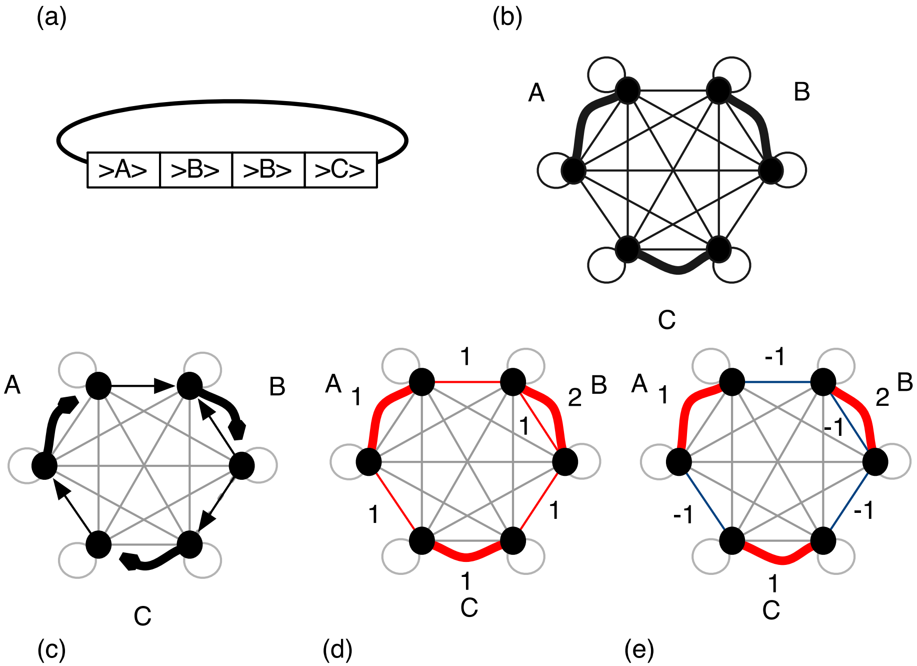
One can use sequence graphs for analysis of cancer metagenomes



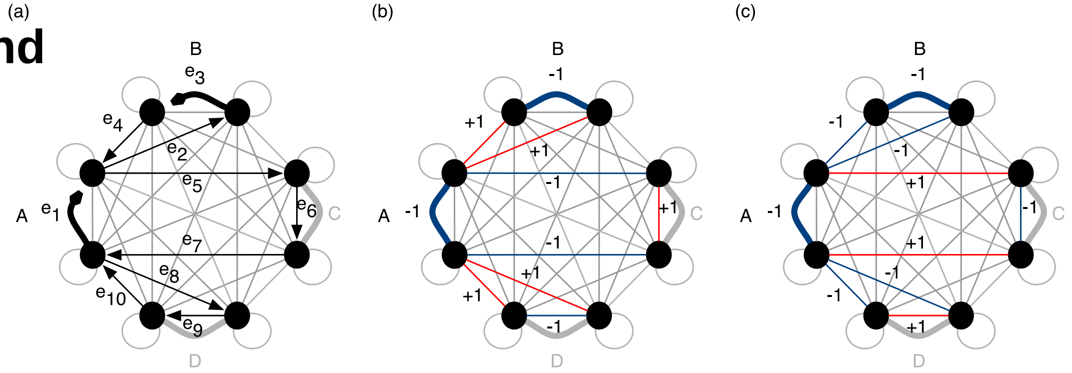
Daniel Zerbino

Algebraic/Combinatorial Approach to Comparative Metagenomics

Flows:



Alternating and simple flows:

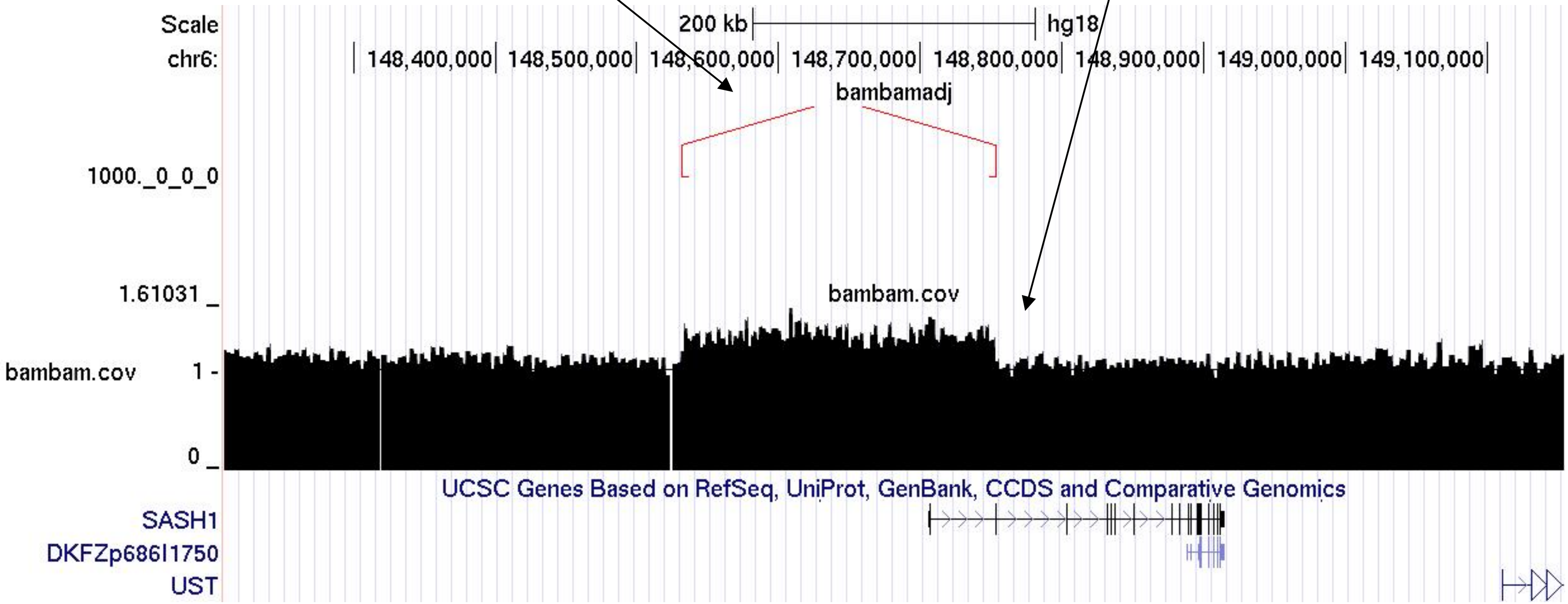


Daniel Zerbino

Duplication – raw data

Detected Breakend

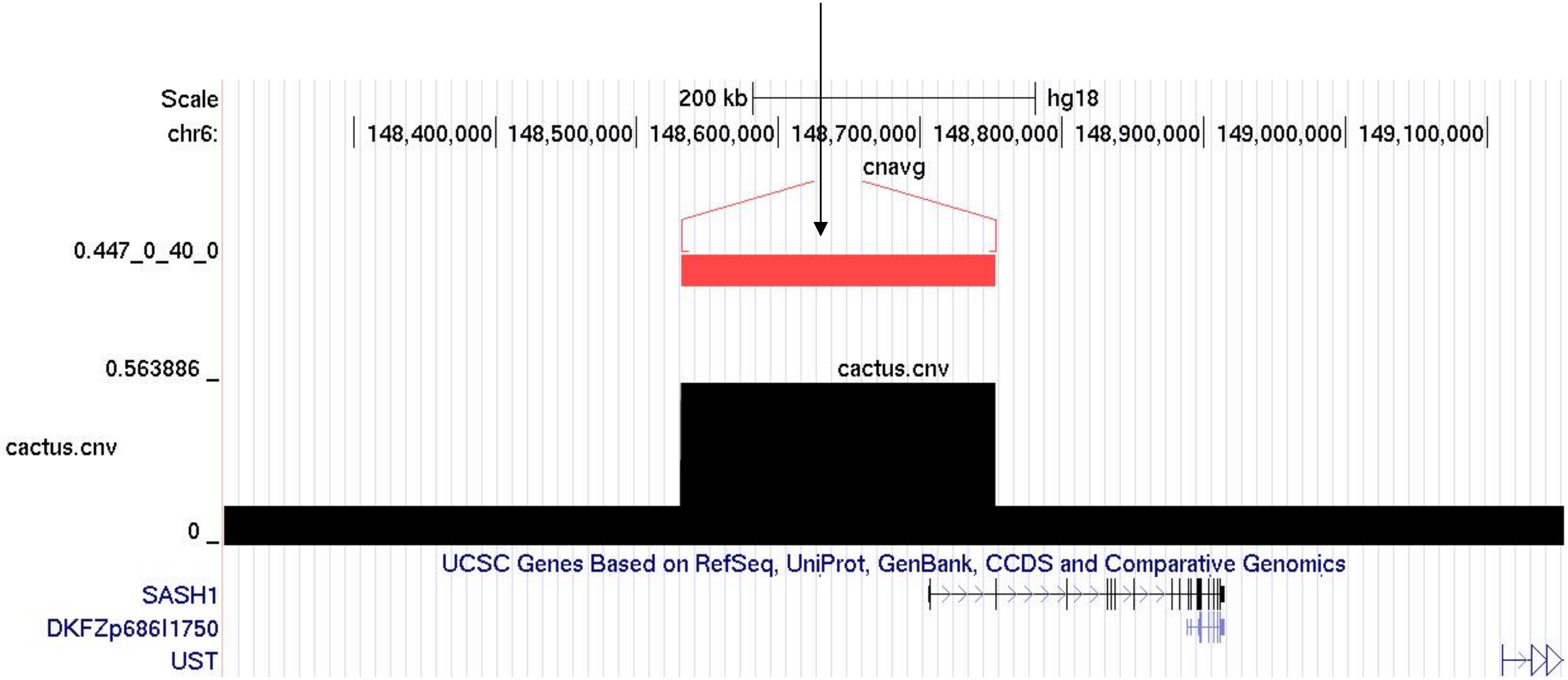
Primary Copy-Number Signal



Brian Raney

Duplication – model from data

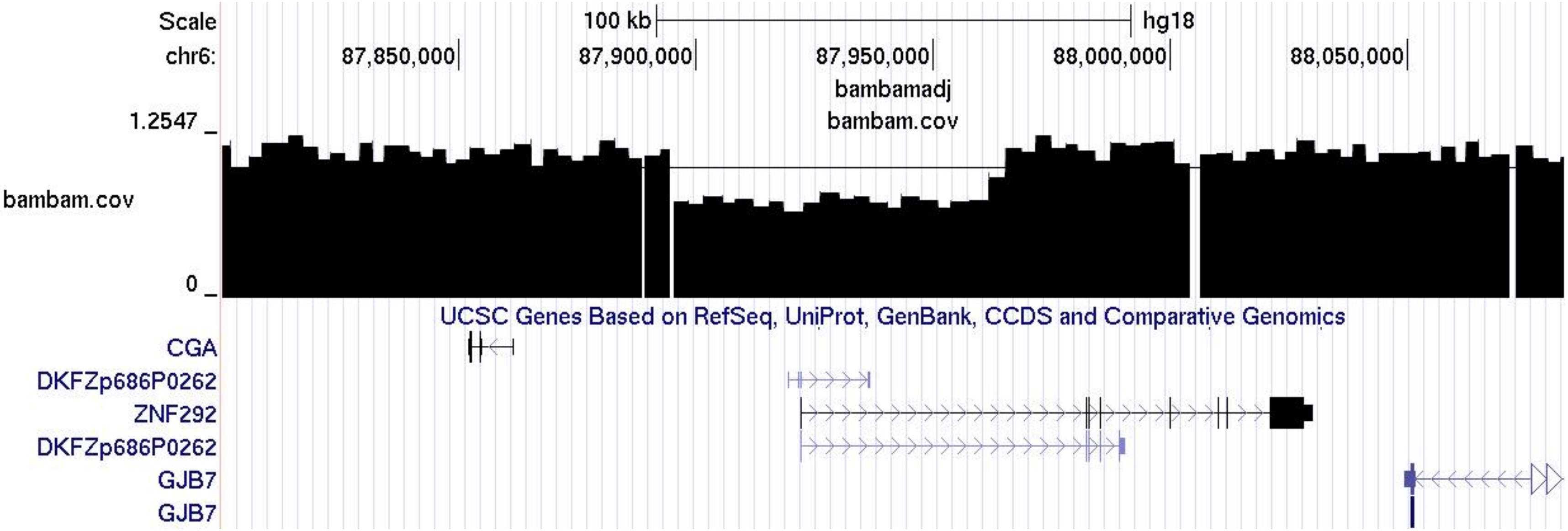
Single duplication event (Copy number change + Breakend)



Red = creation/duplication

Deletion – raw data

(No breakend detected)

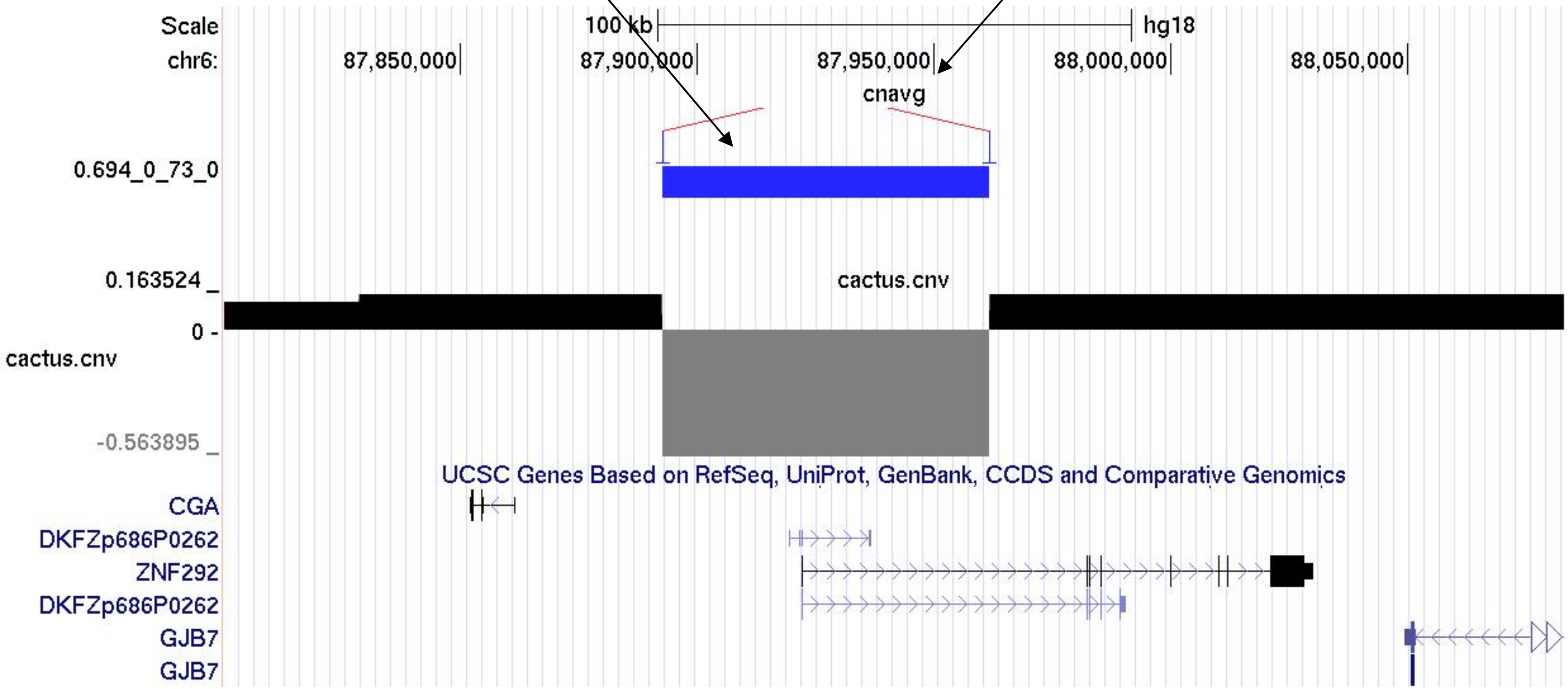


Daniel Zerbino

Deletion – model from data

Single deletion event

Suggested novel breakend creation

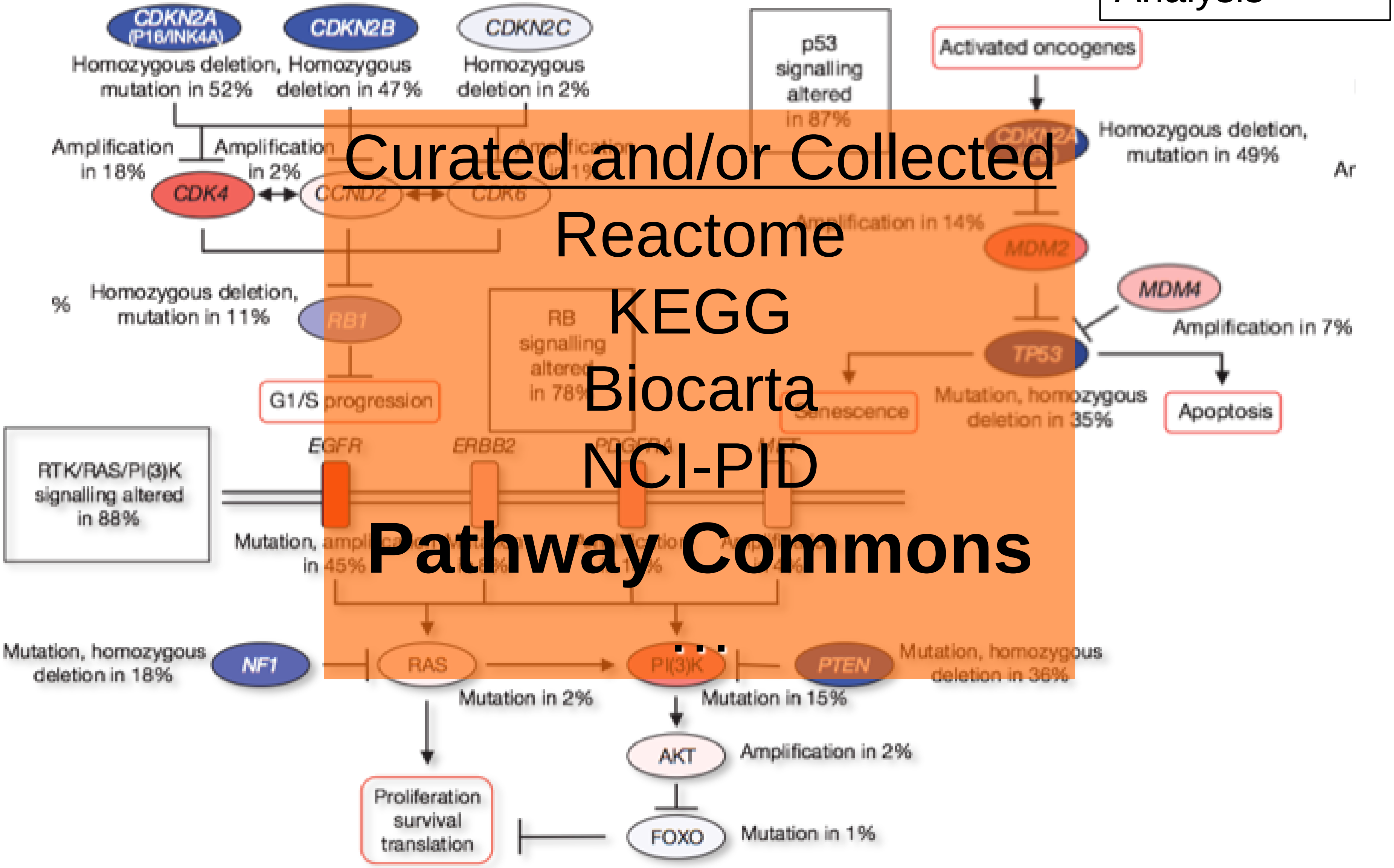


Blue = removal/deletion

Daniel Zerbino

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

TCGA
Glioblastoma
Analysis


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

Curated and/or Collected

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

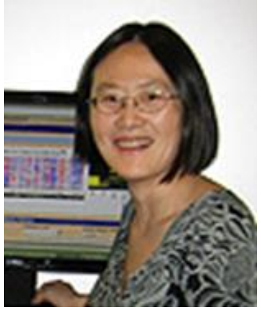
Collaborators

- 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

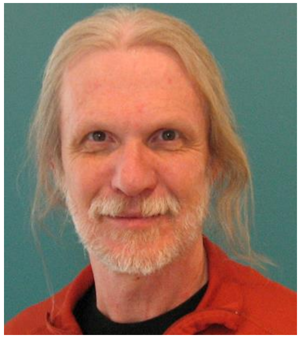


Charlie Vaske

Steve Benz

Zack Sanborn

Mark Diekhans *



Chris Benz

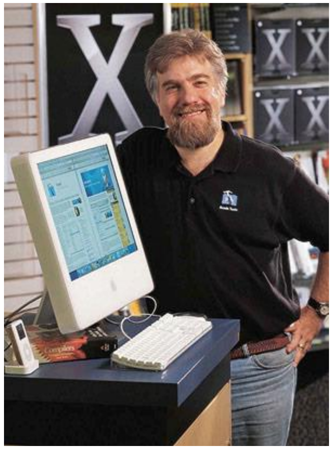
Chris Szeto

Sam Ng

Mia Grifford

James Durbin

Ted Golstein



Melissa Cline

Sofie Salama *

Chris Wilks

Amie Radenbaugh

Brian Craft

Daniel Zerbino
Kyle Elrott
Singer Ma
Artem Sokolov