National Cancer Institute
National Human Genome Research Institute
The Cancer Genome Atlas

Cancer and Evolutionary Genomics

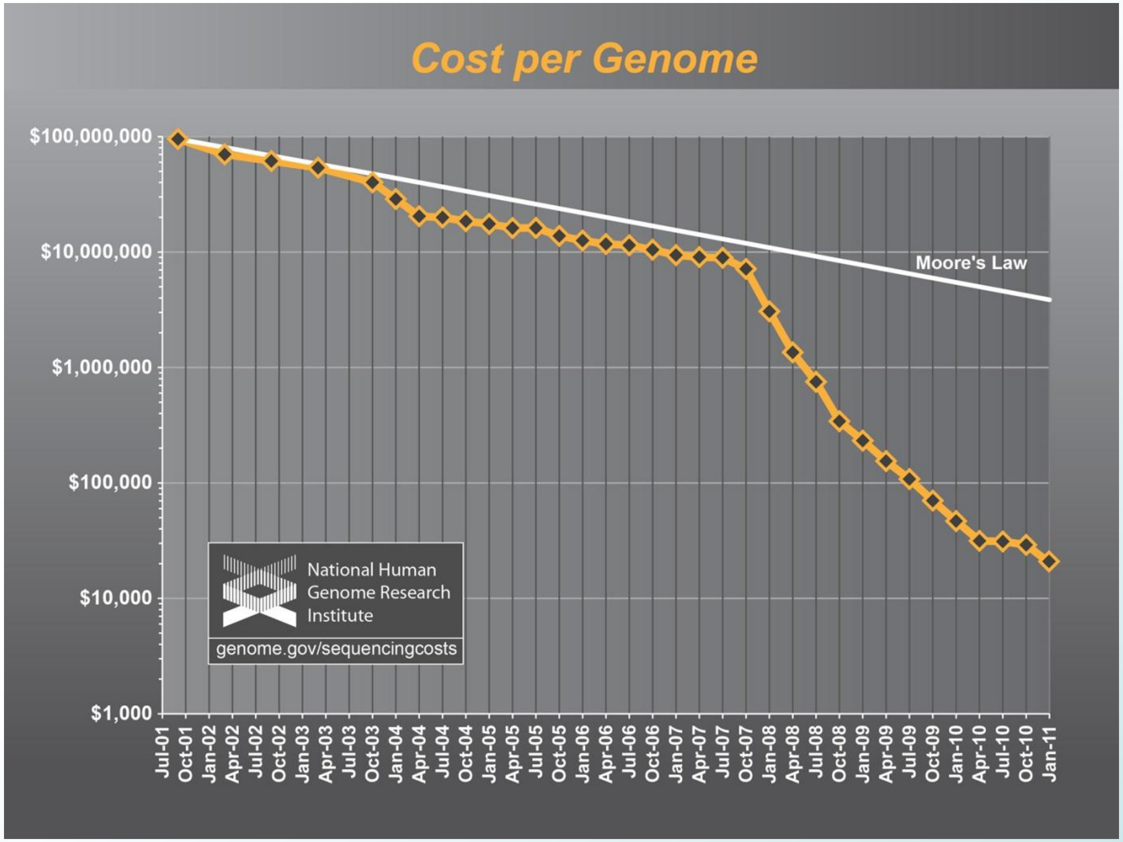
David Haussler

Center for Biomolecular Science and Engineering,

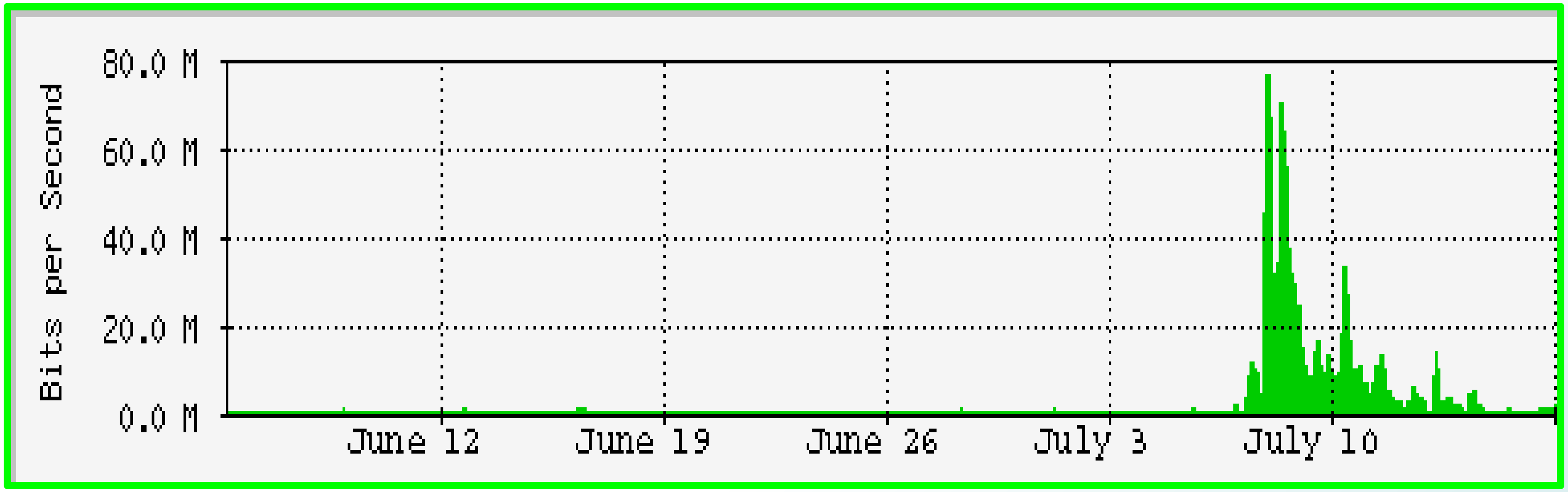
UC Santa Cruz

DNA sequencing cost reduction

- Researchers can now inexpensively sequence entire genomes.
- The cost is headed toward $1000 per genome, dramatically outpacing the Moore’s Law rate for the decreasing cost of computer processing capacity.



On July 7, 2000, UCSC posted first human genome on the web



Outgoing UCSC internet traffic for year 2000

That genome cost $300M.
Analysis was done on small cluster of Dell desktops.

Time to ramp up

Small rack of 100 Dell desktops used to assemble first public draft of human genome in 2000

Google Data

UCSC Cancer Genomics Hub (under construction)

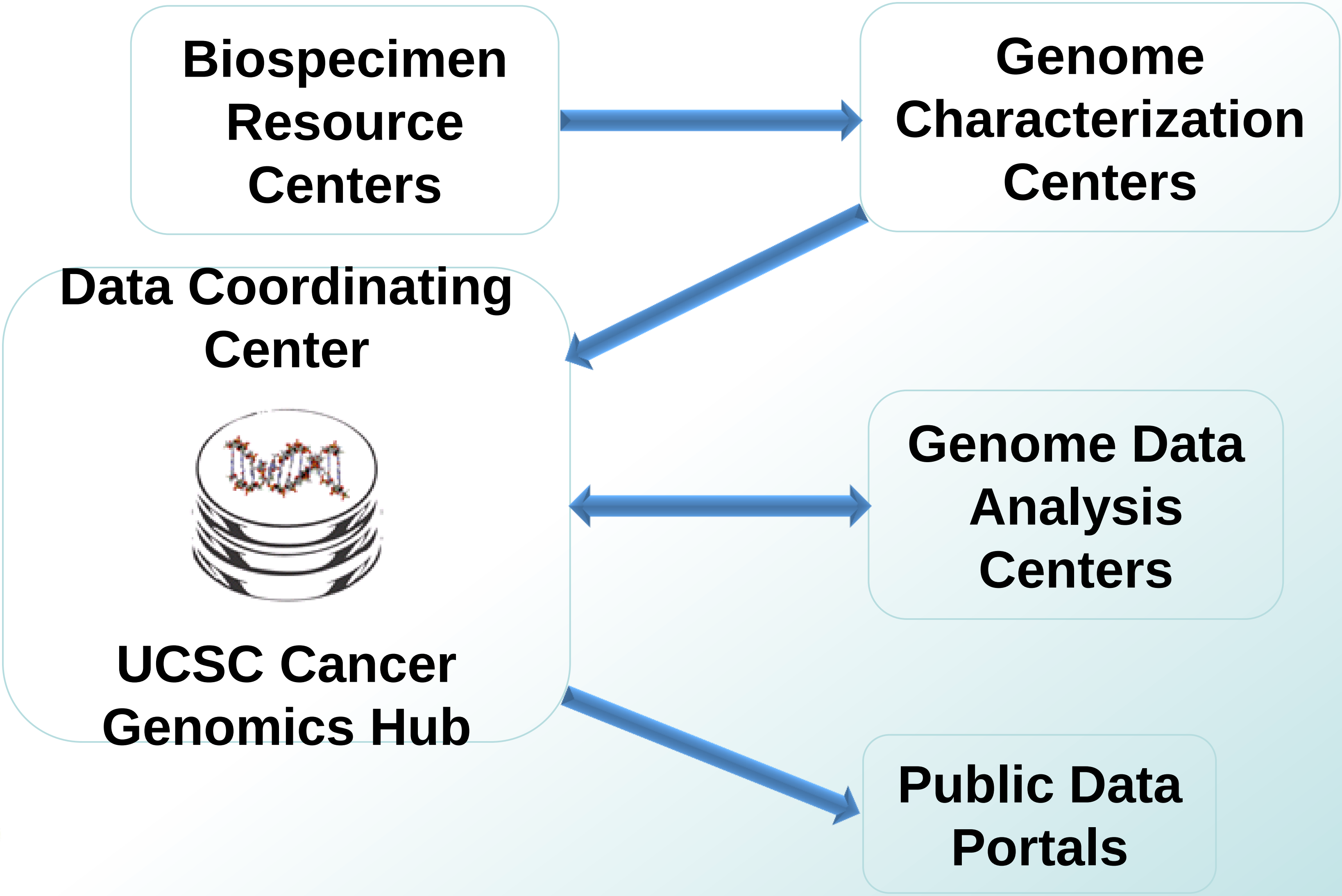
- 5 petabytes of disk storage
- 10 Gigabit dedicated connectivity between data producers, large-scale analysts and database. Modeled after Large Hadron Collider data analysis network
- Secure facility
- Co-location and cloud-computing services to be available on site for TCGA research groups

Erich Weiler and CGHub team

Cancer genomics will lead the way to personal genomics

- High mortality and cost of care make additional cost of sequencing tumor genomes relatively minor.
- Disease is caused by changes to the genome; highly individual.
- Many genes are targeted by available drugs.
- Standard of care often fails.
- Once standard of care fails, can explore strategies based on genome analysis that have not been completely validated by a traditional clinical trial (e.g., combination therapies).
- We may not need a perfectly accurate quantitative model of how most types of cancer cells work to figure out how to selectively kill them (e.g., immunology-based cancer treatments). A full genomic scan may provide a suitable “kill signature.”
- A full-genome scan will provide a means to detect a recurrence very early from a simple blood test.

The Cancer Genome Atlas (TCGA): >20 cancer types, 500 tumors from each



Biospecimen Resource Centers

Genome Characterization Centers

Data Coordinating Center

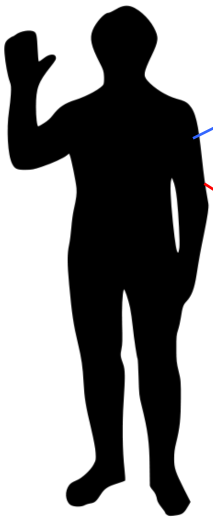
UCSC Cancer Genomics Hub

Genome Data Analysis Centers

Public Data Portals

Sequencing Cancer Genomes (Broad, Wash U., Baylor, plus Vancouver and other smaller centers)

Patient

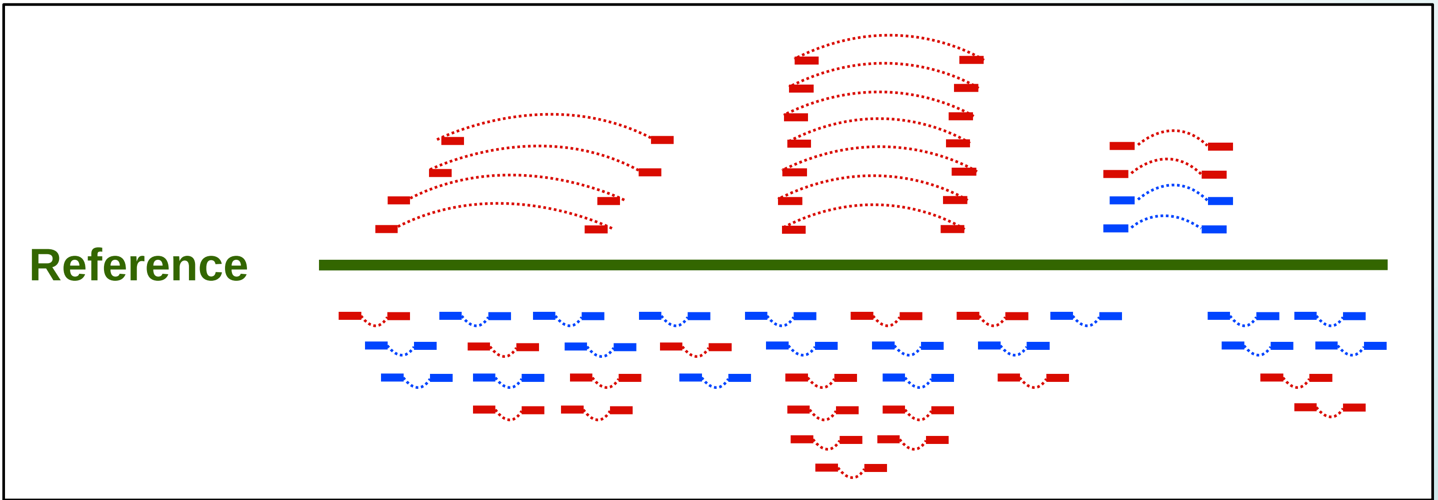


Germline DNA from blood

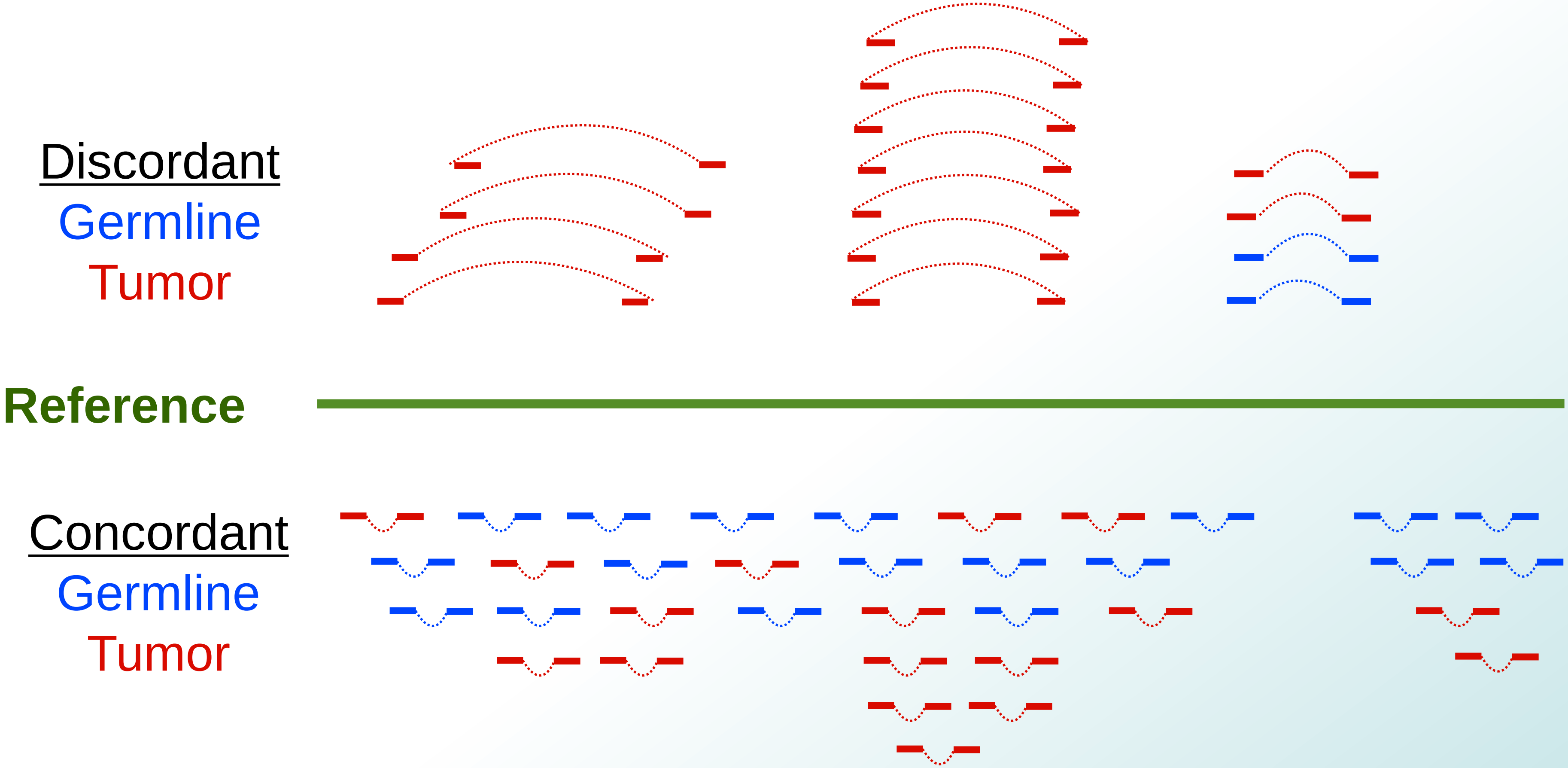
Tumor DNA

100 Millions of short read pairs

Match read pairs to the reference genome



Mapped reads:



Discordant

Germline

Tumor

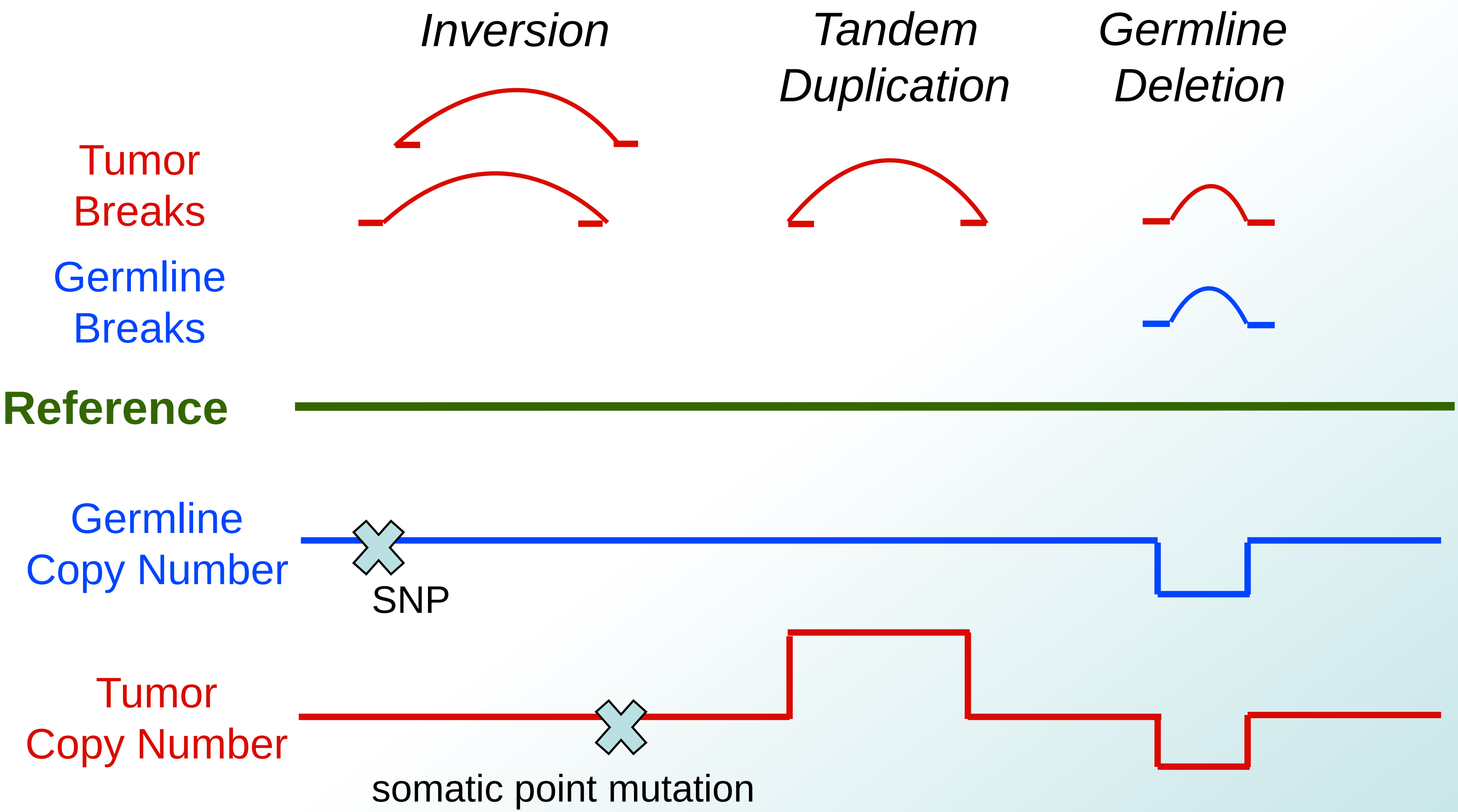
Reference

Concordant

Germline

Tumor

Low-level interpretation of mapped reads:



Inversion

Tandem Duplication

Germline Deletion

Tumor Breaks

Germline Breaks

Reference

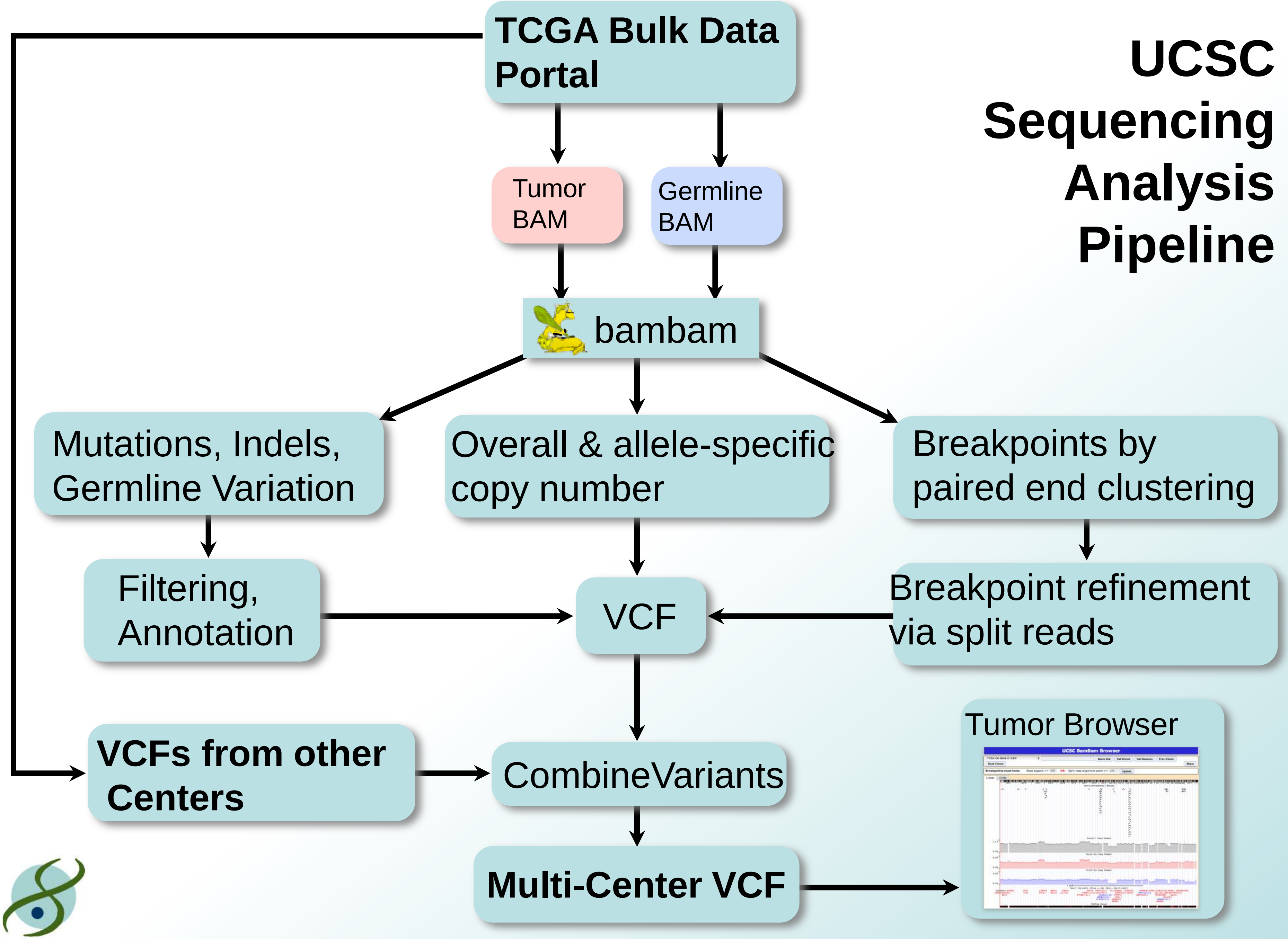
Germline Copy Number

SNP

Tumor Copy Number

somatic point mutation

UCSC Sequencing Analysis Pipeline



TCGA Bulk Data Portal

Tumor BAM

Germline BAM

bambam

Mutations, Indels, Germline Variation

Overall & allele-specific copy number

Breakpoints by paired end clustering

Filtering, Annotation

VCF

Breakpoint refinement via split reads

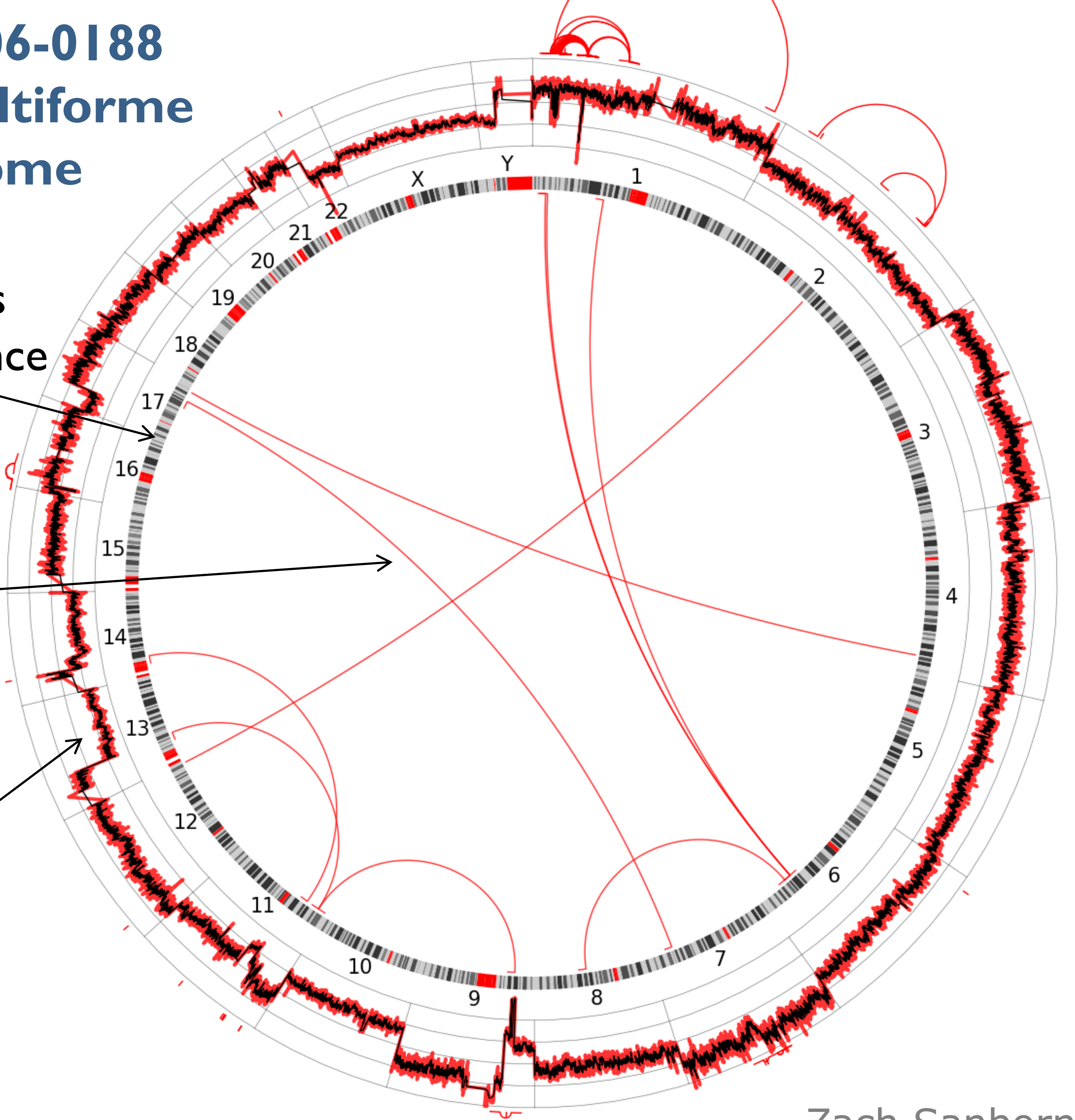
VCFs from other Centers

CombineVariants

Multi-Center VCF

Tumor Browser

Tumor TCGA-06-0188 Glioblastoma Multiforme Whole Genome

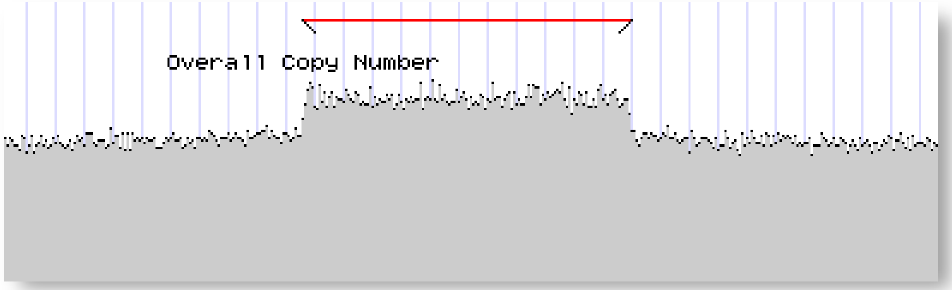


Human chromosomes 1-22, X and Y (reference genome)

Rearrangements In the tumor

Number of copies In the tumor

Zach Sanborn



chr2 : 29,064,107

OV-0751 Somatic Reads

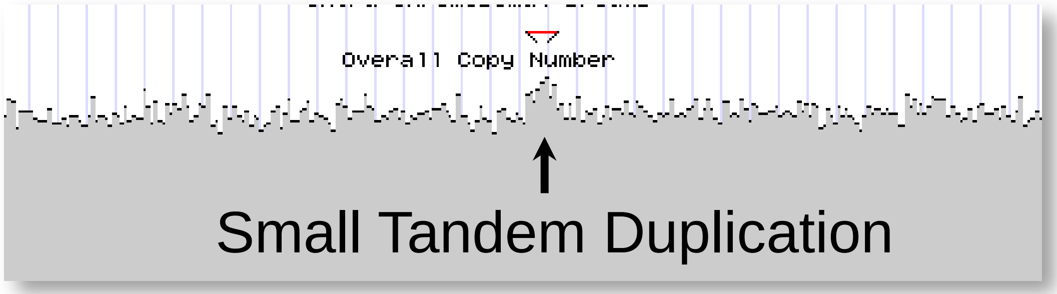
```
ATTGCTGGCTGGCTGCACCCTATATTGTCTGAGAACAGAGTGGCTACAggagtattaaccccacctgatctcacgatgggagaggagacgcca

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

               cactttctacagacgatgtcaccttccacctCACAGAAAATGGAGGCCATCAGAGGGCATCTCCtccatctcccatcg
```

chr2 : 28,500,054

Tandem Duplication Size = 564,053 bp



Small Tandem Duplication

chr2 : 169,227,553

OV-0751 Somatic Reads

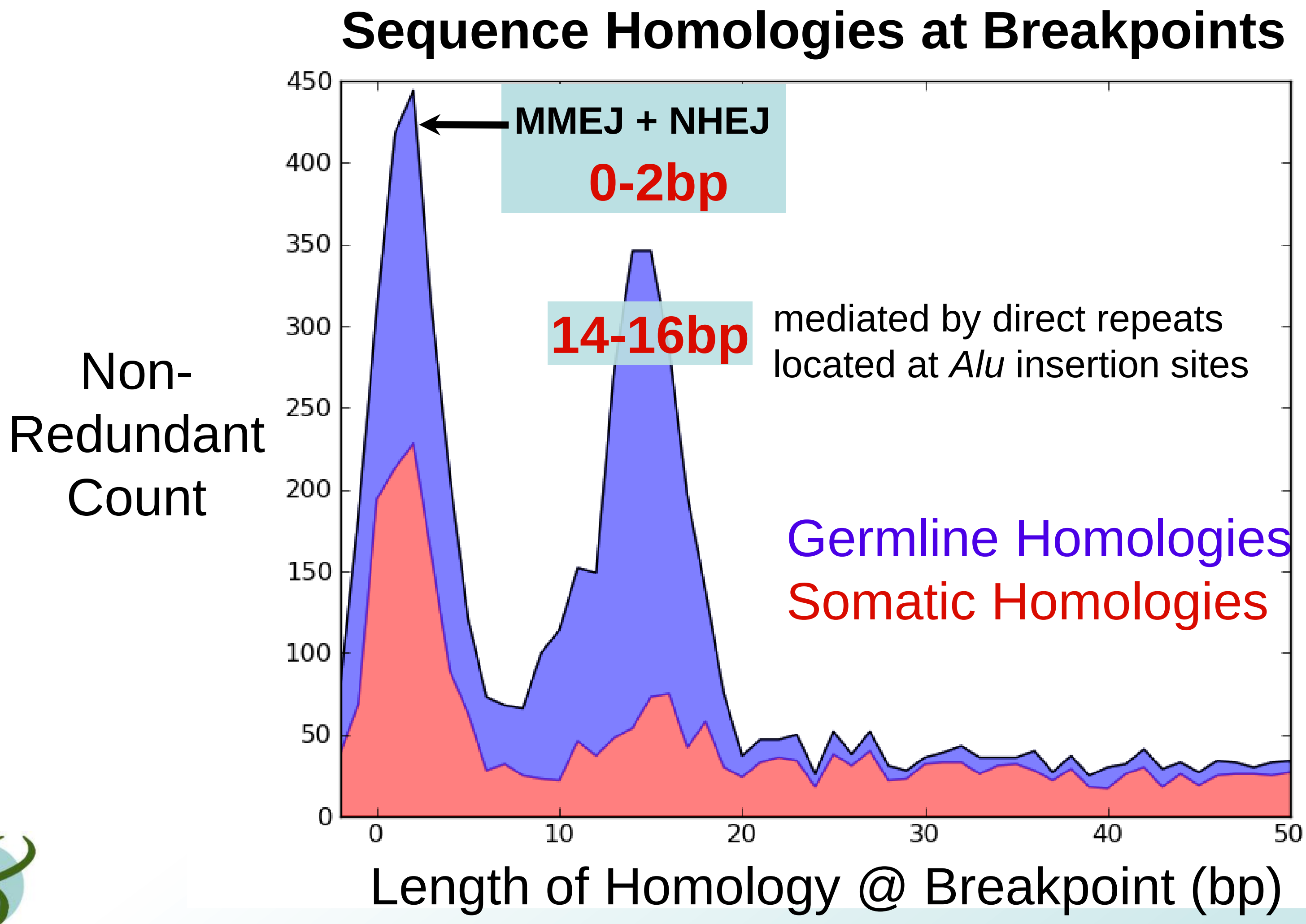
```
cccTGGAAAGAAAGCTAGACCCAGAGAGCTGGGGTCCAGCAGCCTAgcttcccacagcgtacaatatgtggtataaagtctttgcagcttcctc
   TGGAAAGAAAGCTAGACCCAGAGAGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC
CCCTGGAAAGAAAGCTAGACCCAGAGAGCTGGGGCCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAA
CCCTGGAAAGAAAGCTAGACCCAGAGAGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCCTCAACAGT
   TGGAAAGAAAGCTAGACCCAGAGAGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC
                     CAGAGAGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC
                       GAGGGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC
    GGAAAGAAAGCTAGACCCAGAGAGCTGGGGTCCAGCAGCCTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC

gggttaaatggtagttctgcttttagctgtttgaggaattgccaTACTGCTTTCCACAATGGTTTAACTAATTTGCACTCCCATCAACAGTGTC
```

chr2 : 169,210,871

Tandem Duplication Size = 16,682 bp

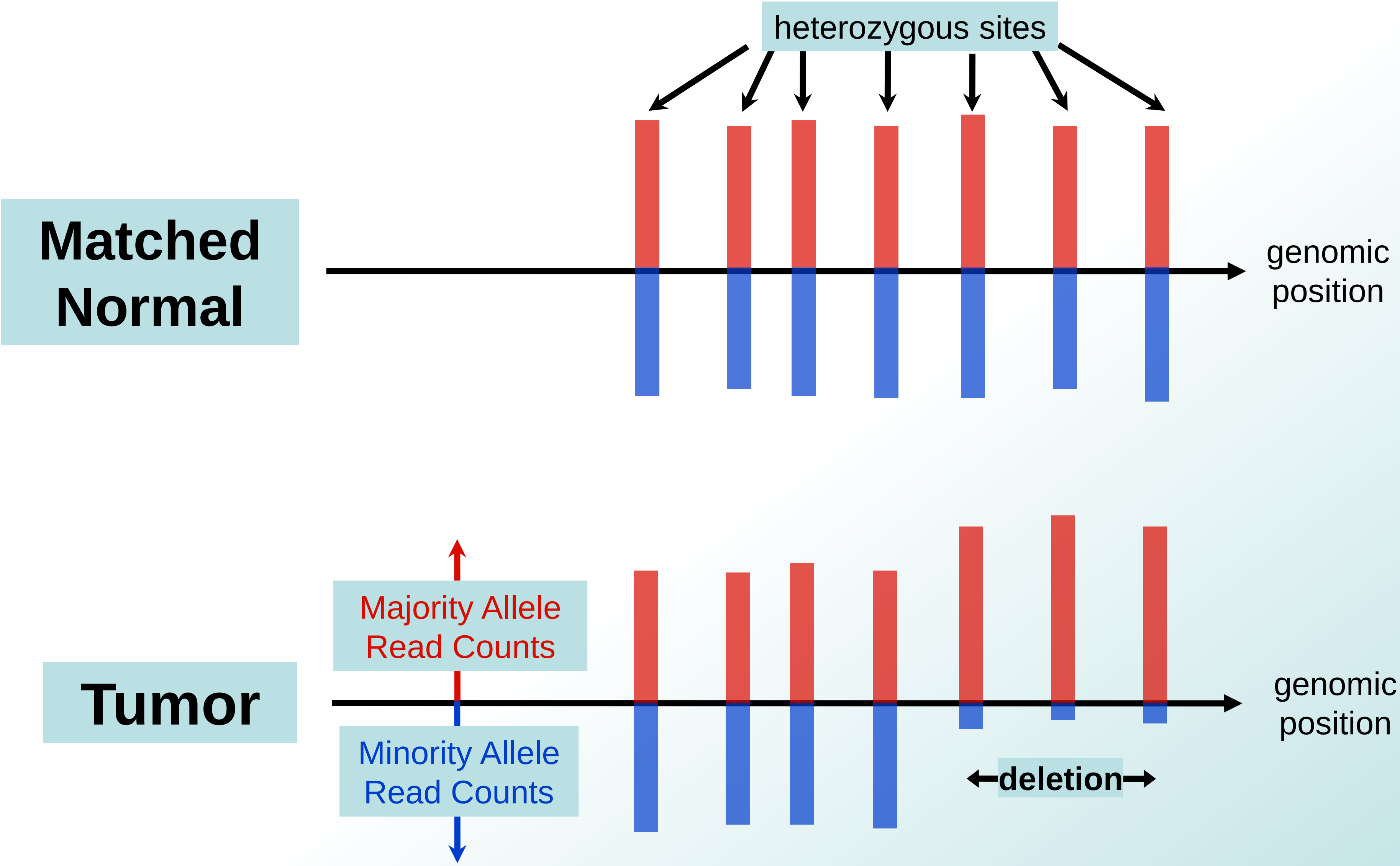
Somatic Breakpoints are enriched for Non-Homologous End Joining (NHEJ)



mediated by direct repeats located at *Alu* insertion sites

13 GBMs

Allele-Specific Copy Number



heterozygous sites

Matched Normal

genomic position

Tumor

Majority Allele Read Counts

Minority Allele Read Counts

genomic position

deletion

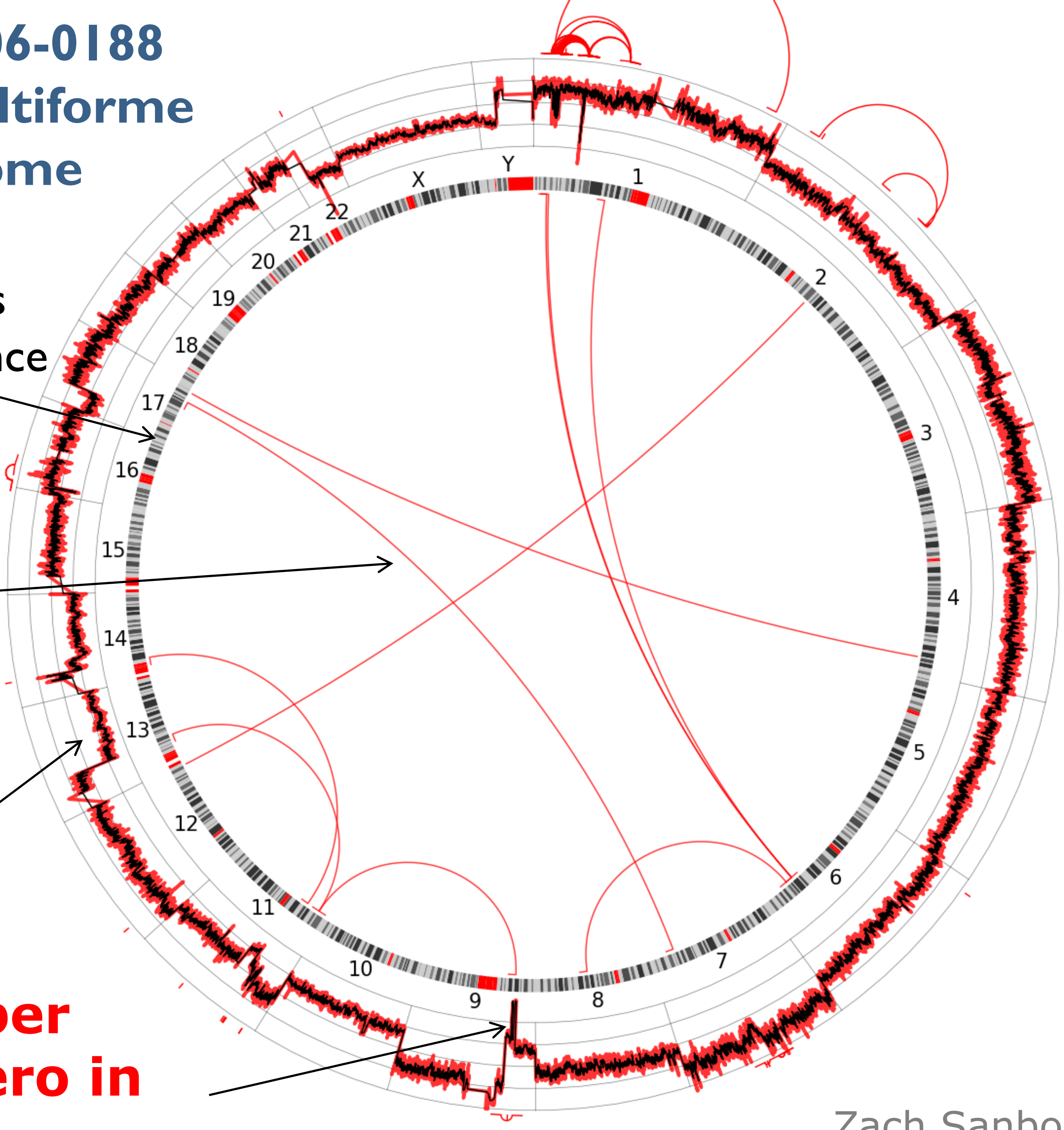
Zack Sanborn

Tumor TCGA-06-0188
Glioblastoma Multiforme
Whole Genome

Human chromosomes
1-22, X and Y (reference
genome)

Rearrangements
In the tumor

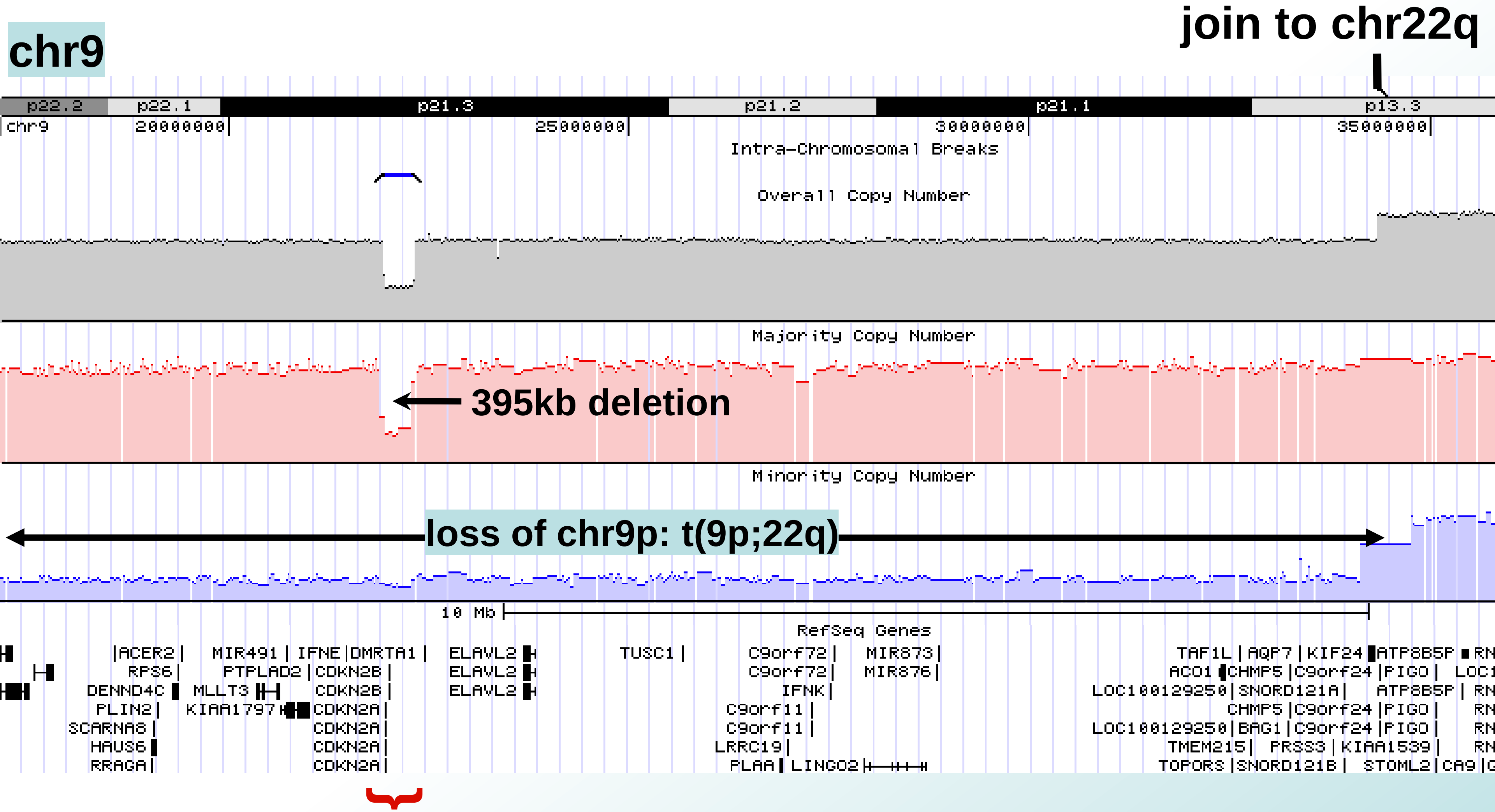
Number of copies
In the tumor



Copy number drops to zero in one place

Zach Sanborn

Glioblastoma: TCGA-06-0145


chr9

join to chr22q

395kb deletion

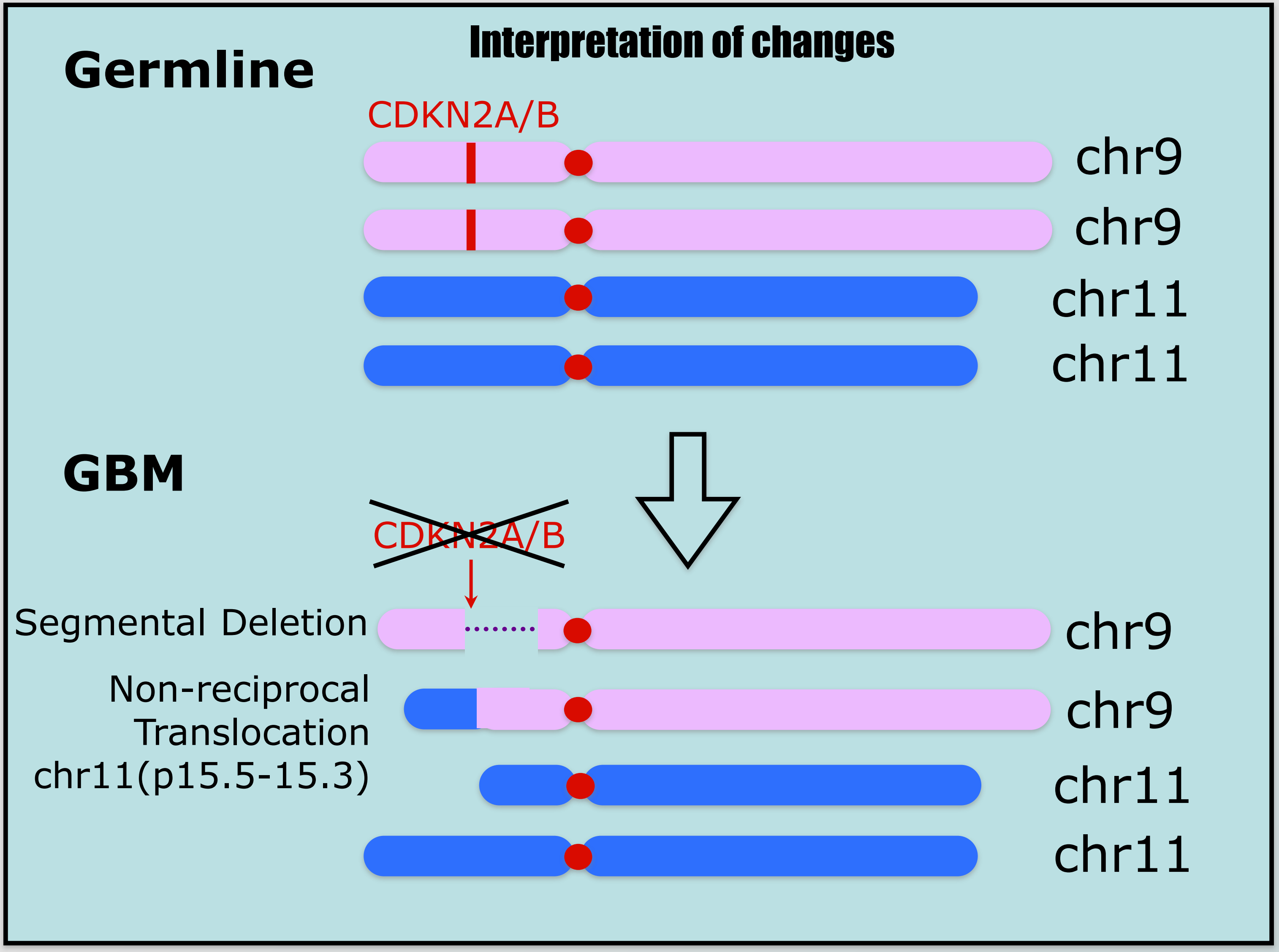
loss of chr9p: t(9p;22q)

Homozygous loss of CDKN2A/B

via inter- and intra-chromosomal rearrangements.

Zack Sanborn

Glioblastoma: TCGA-06-0145

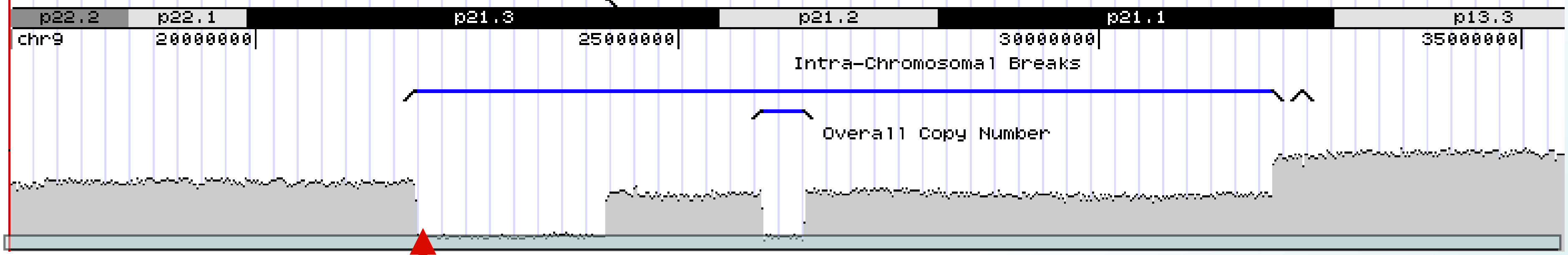


Similar double-loss motif in other GBMs

chr9

join to chr11p

GBM: TCGA-06-0188

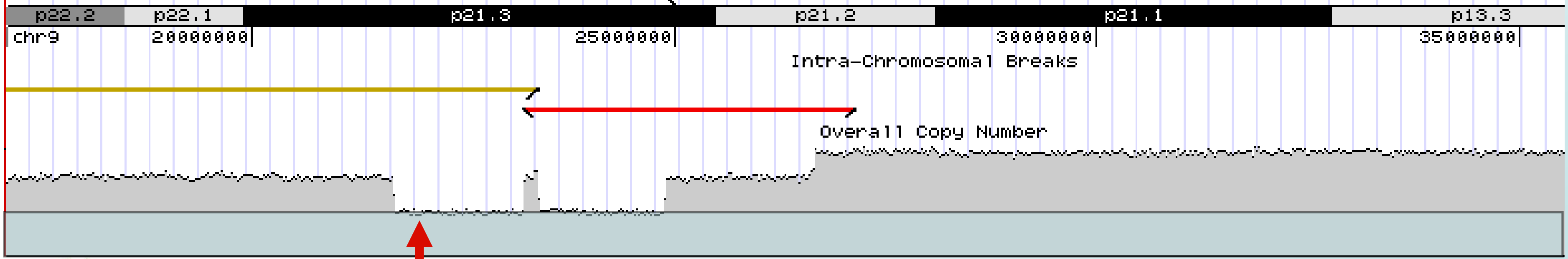


CDKN2A/B

chr9

join to chr14q

GBM: TCGA-06-0214



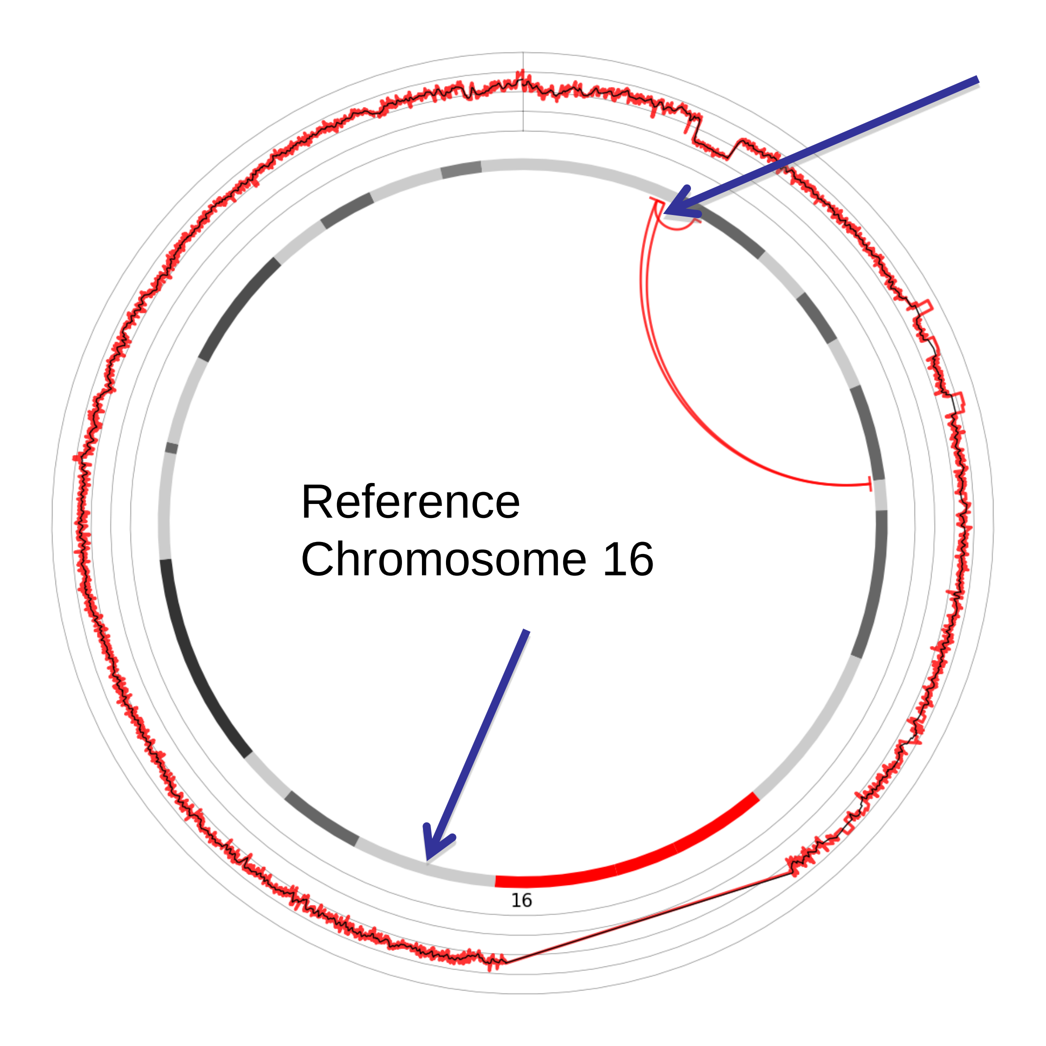
CDKN2A/B

Zack Sanborn

Similar events lead to double loss of CDKN2A/B in GBM

	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*

Simple Example of Tumor Genome reconstruction from this Glioblastoma Tumor



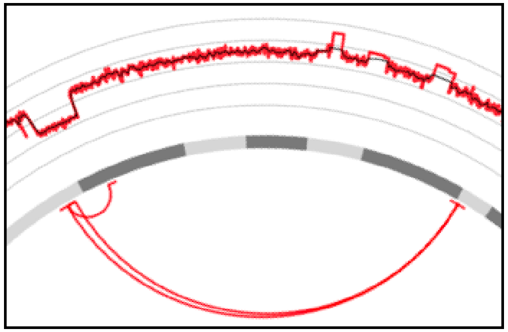
This region contains the gene A2BP1

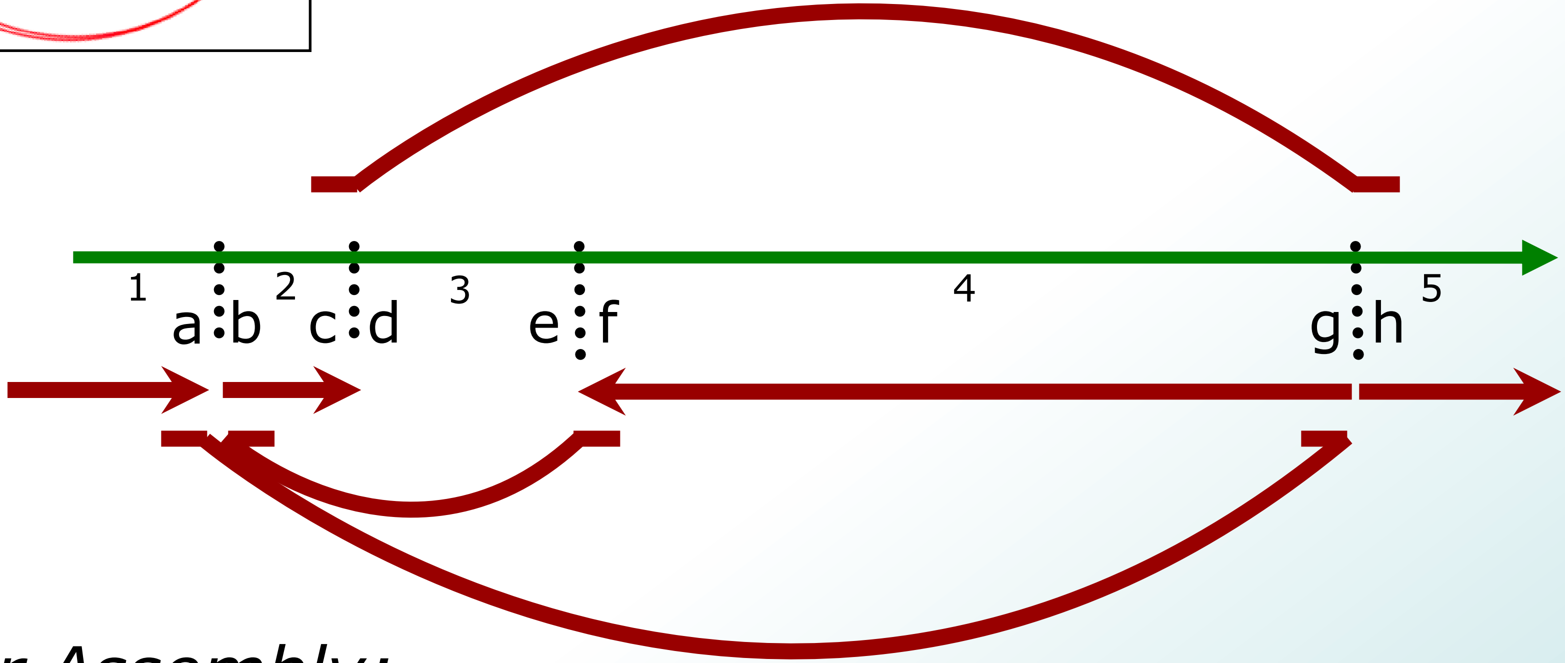
Reference Chromosome 16

The A2BP1 gene is a Tissue-specific alt-splicing regulator, important in **brain**, heart, muscle

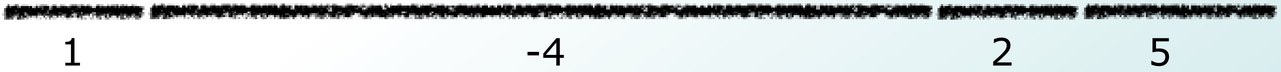
J. Zachary Sanborn

We walk the adjacencies to determine order of segments in the tumor

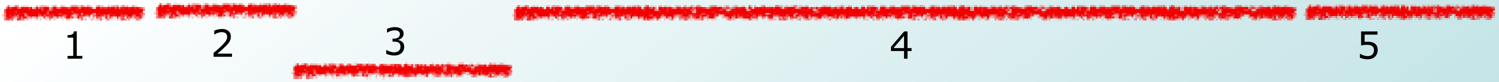




Tumor Assembly:

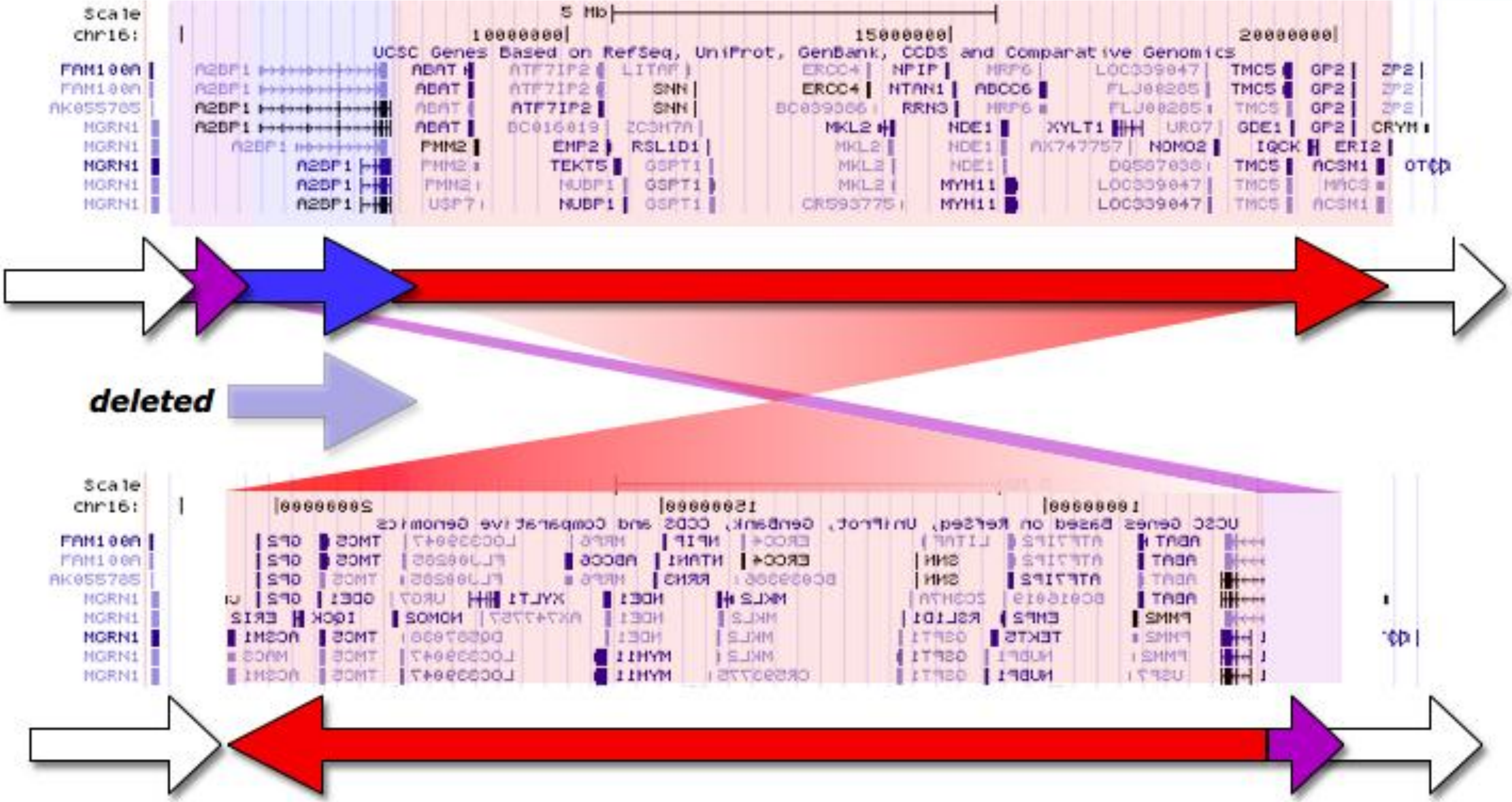


Tumor Copy Number:



Tumor Browser View of TCGA GBM

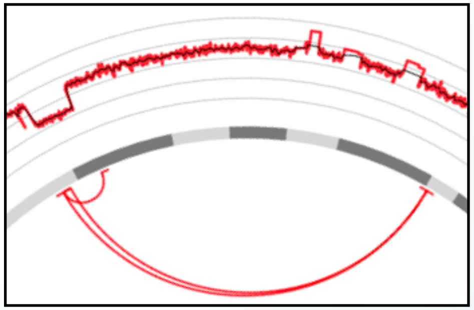
Region in Germline Genome, all on same parental haplotype

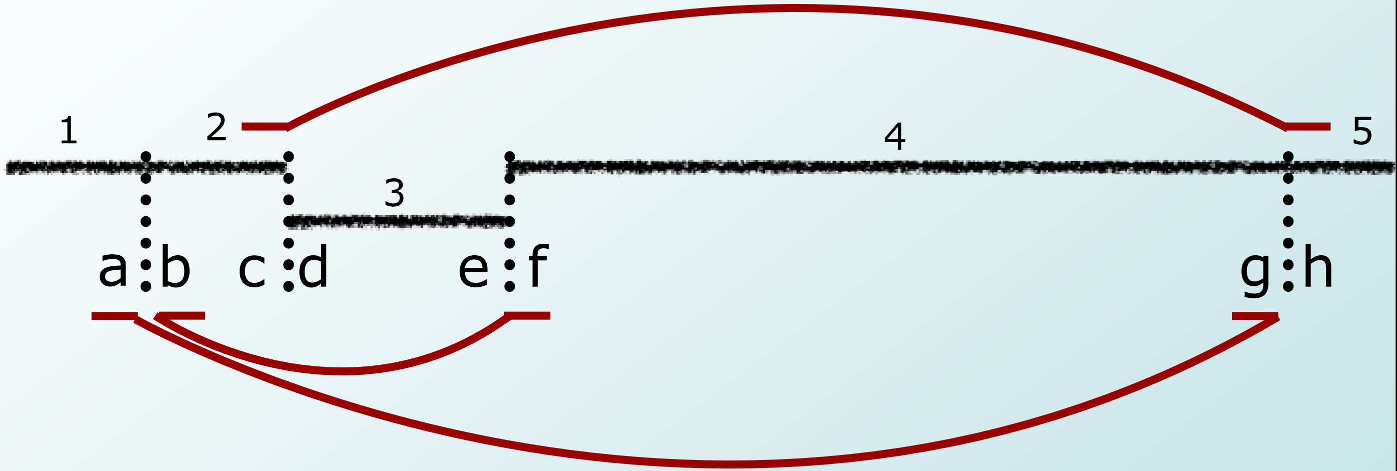


Tumor Genome Micro-Assembly

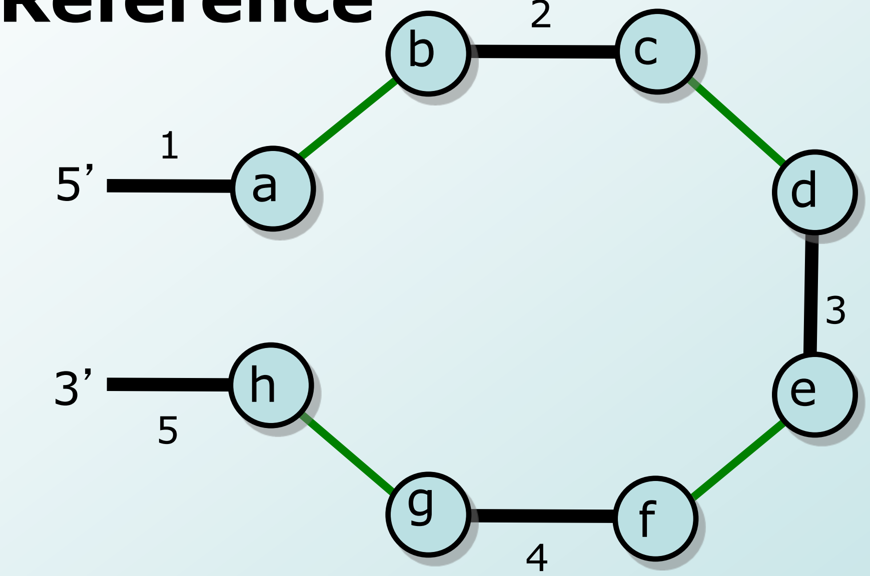
J. Zachary Sanborn

A Breakpoint graph shows how adjacencies of segment ends change



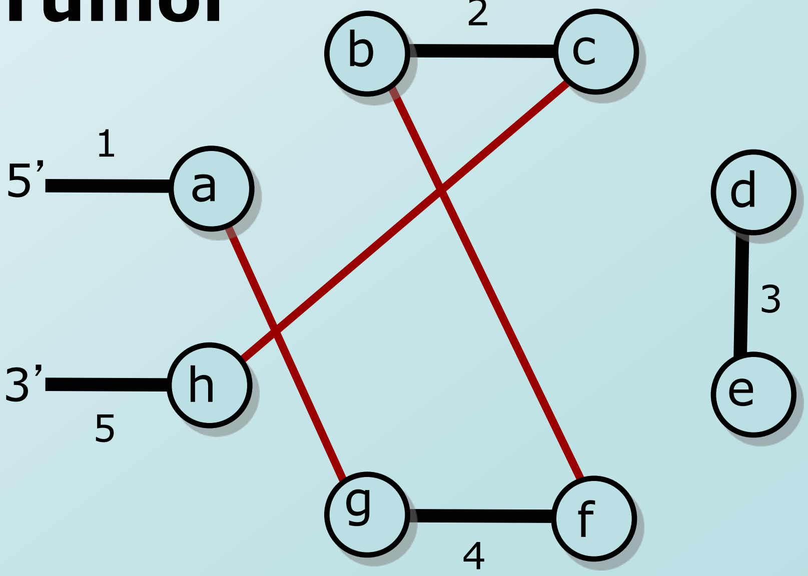


Reference



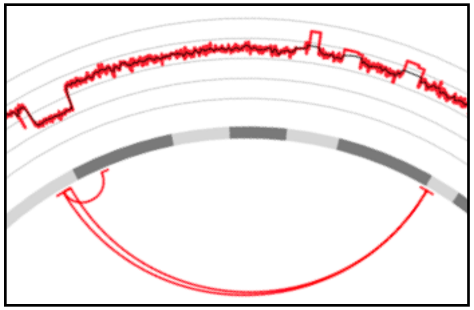
Genome = 1 2 3 4 5

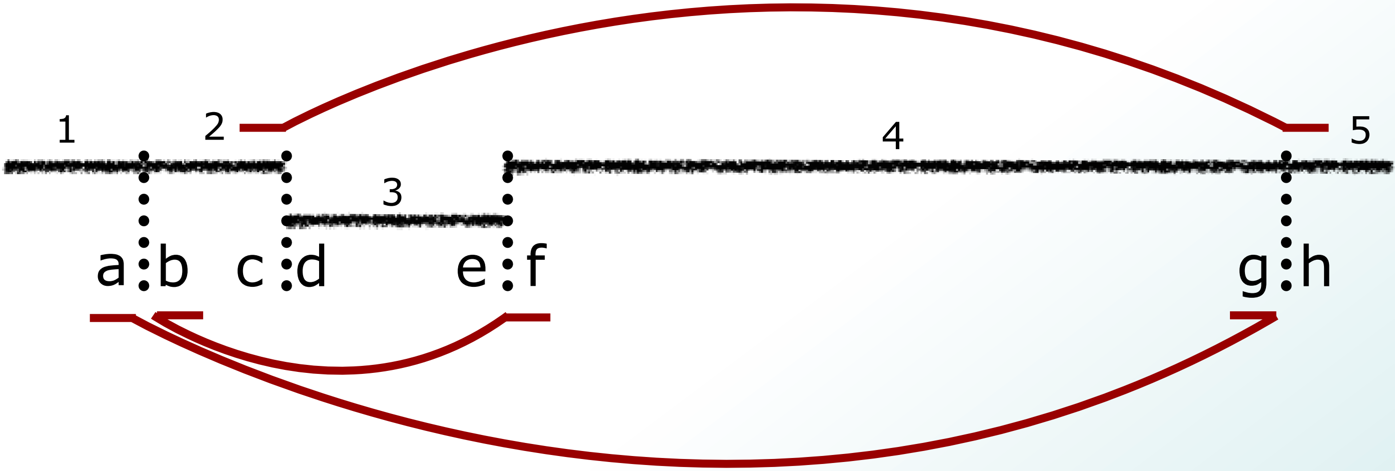
Tumor



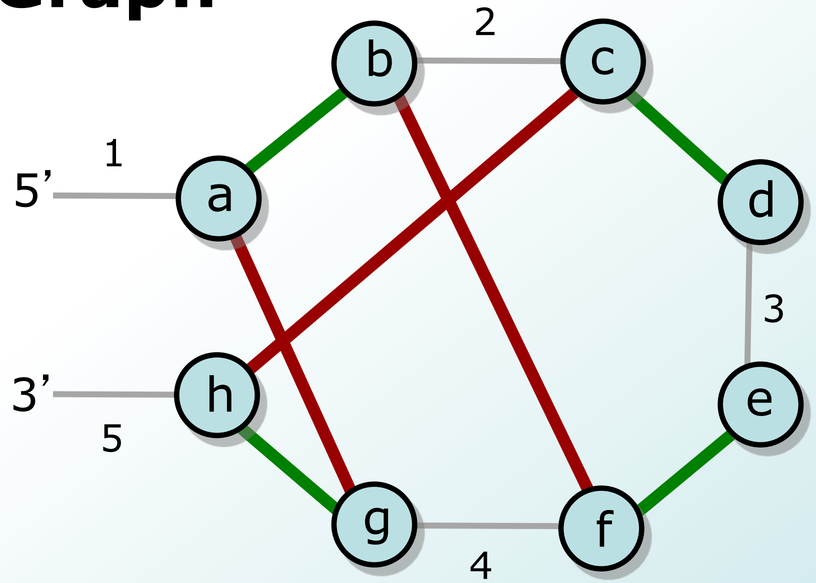
Genome = 1 -4 2 5

A Breakpoint graph shows how adjacencies of segment ends change





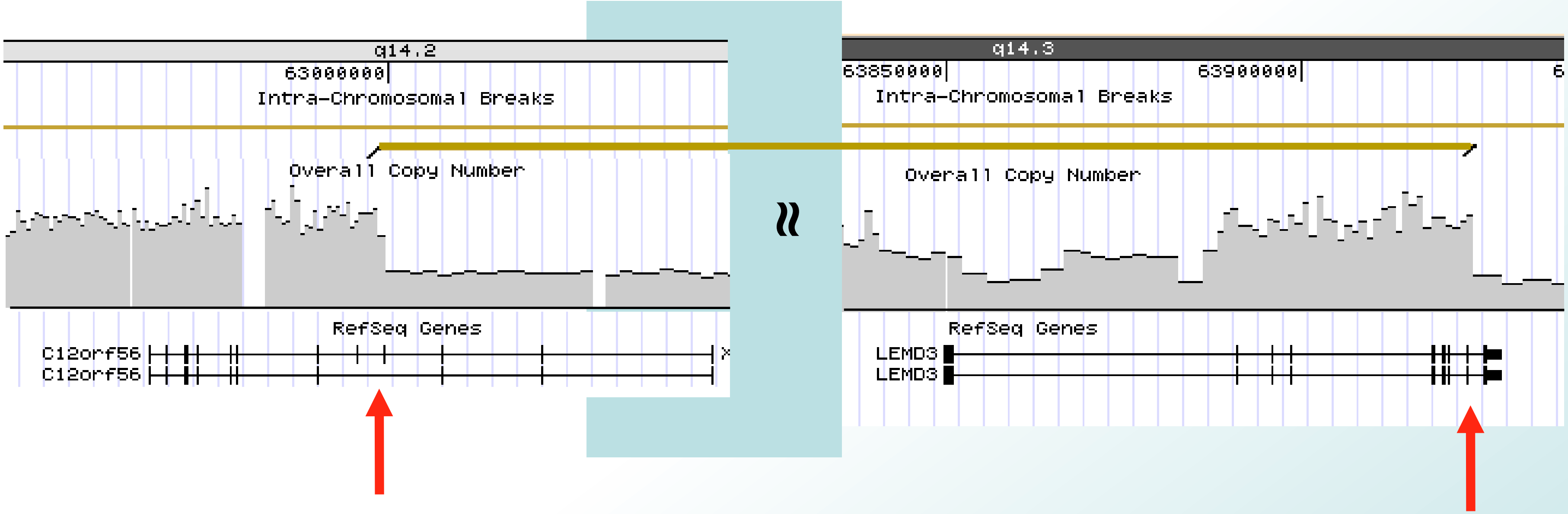
Breakpoint Graph



GBM Gene Fusions

- **Broad's dRanger** identified a set of 7 high confidence rearrangements connecting the introns of two genes, across 17 whole genome TCGA GBM datasets:
  - **3 in-frame**, potentially functional gene fusions:
    - All 3 confirmed by bambam PE clustering, CNV, and bridget split-reads
  - **4 out-of-frame**:
    - 3 confirmed by bambam, CNV, and bridget
    - 1 missed by bambam, but CNV suggests the breakpoint indeed exists

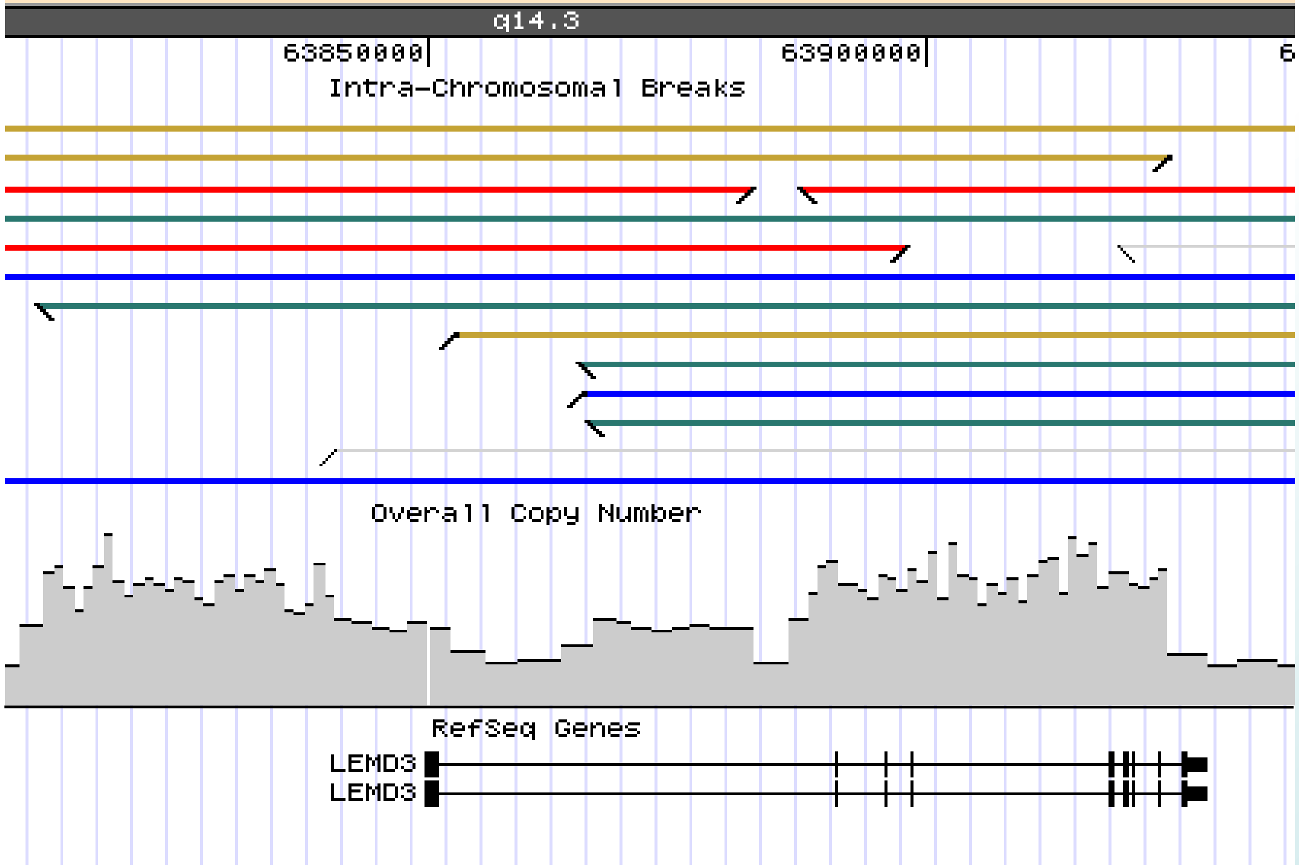
LEMD3 - c12orf56 Fusion



c12orf56 Fusion Point

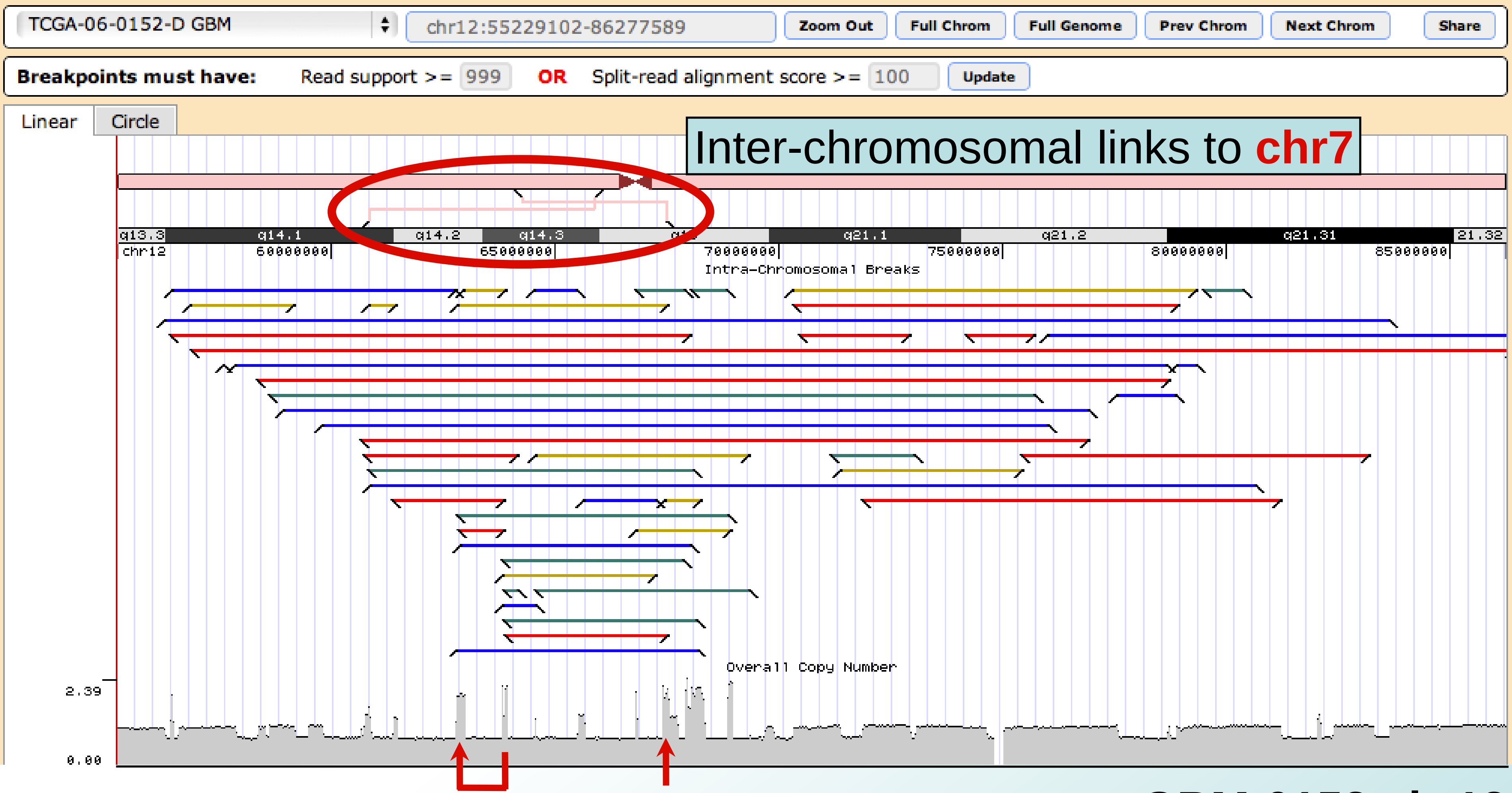
LEMD3 Fusion Point

LEMD3 - c12orf56 Fusion



LEMD3-c12orf56 Fusion Point

LEMD3-c12orf56 - Chromothripsis



TCGA-06-0152-D GBM

chr12:55229102-86277589

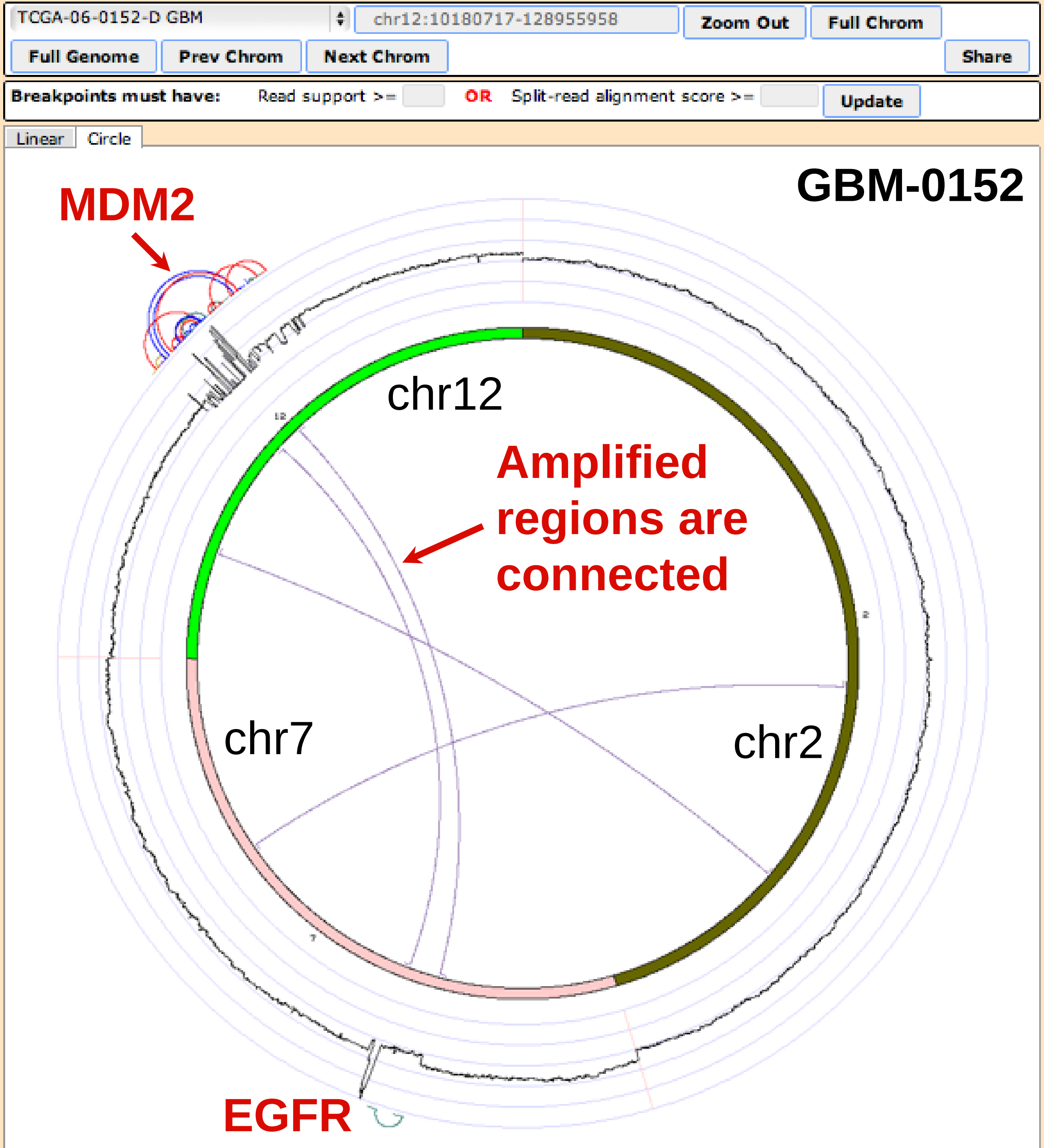
Breakpoints must have: Read support >= 999 **OR** Split-read alignment score >= 100

Inter-chromosomal links to **chr7**

LEMD3-c12orf56 Fusion

MDM2

GBM-0152 chr12



TCGA-06-0152-D GBM

chr12:10180717-128955958

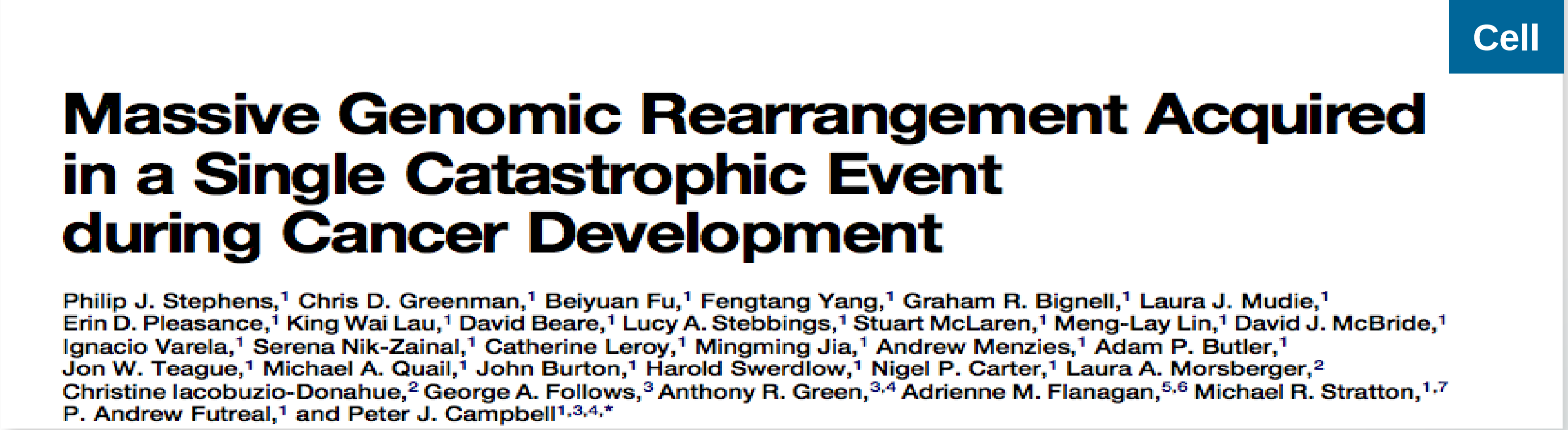
Breakpoints must have: Read support >= OR Split-read alignment score >=

GBM-0152

MDM2

Amplified regions are connected

EGFR

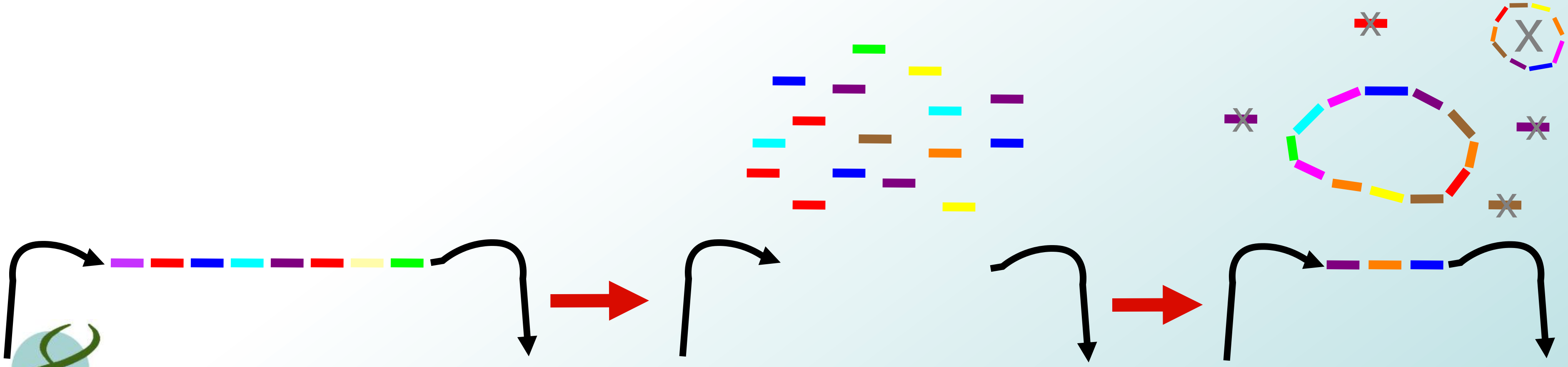


Cell

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

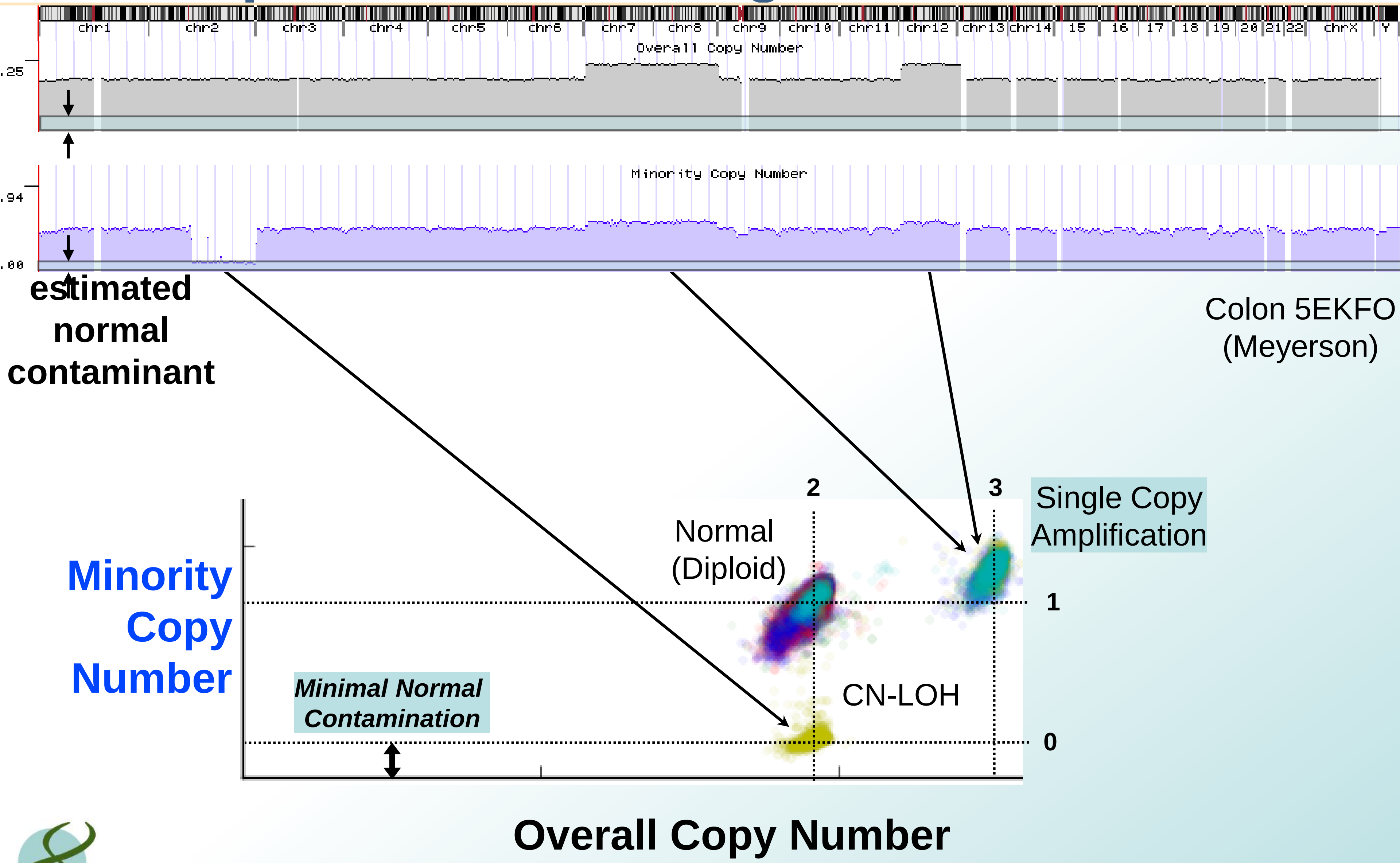
Philip J. Stephens,[1] Chris D. Greenman,[1] Beiyuan Fu,[1] Fengtang Yang,[1] Graham R. Bignell,[1] Laura J. Mudie,[1] Erin D. Pleasance,[1] King Wai Lau,[1] David Beare,[1] Lucy A. Stebbings,[1] Stuart McLaren,[1] Meng-Lay Lin,[1] David J. McBride,[1] Ignacio Varela,[1] Serena Nik-Zainal,[1] Catherine 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 shatters into pieces due to some genetic insult when chromosome is in condensed state
- DNA repair mechanisms try to stitch genome back together
- Can generate rearrangements, losses, and double minute chromosomes



Zack Sanborn

Tumors exhibit multiple rounds of duplication, rearrangement and loss



estimated normal contaminant

Colon 5EKFO (Meyerson)

Normal (Diploid)

Single Copy Amplification

CN-LOH

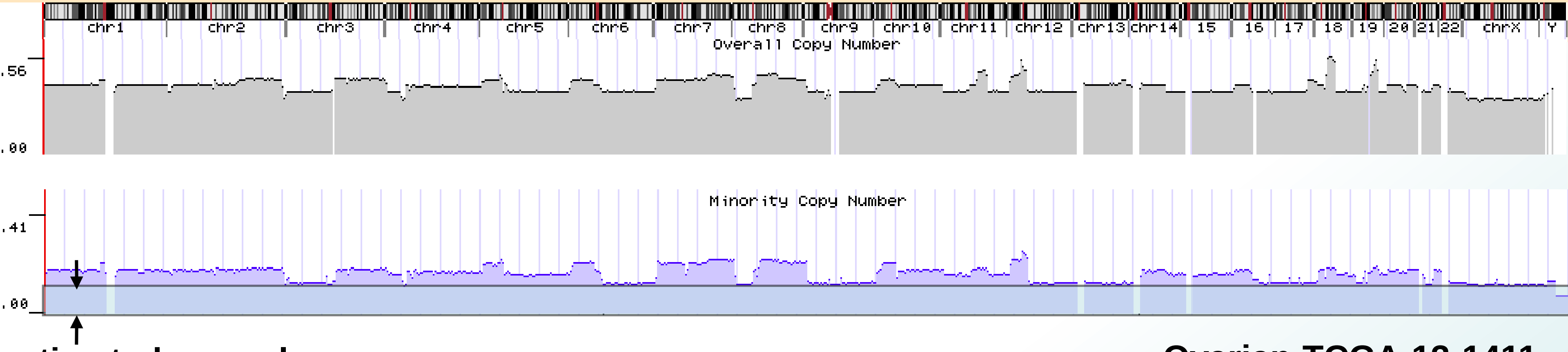
Minimal Normal Contamination

Minority Copy Number

Overall Copy Number

Zack Sanborn

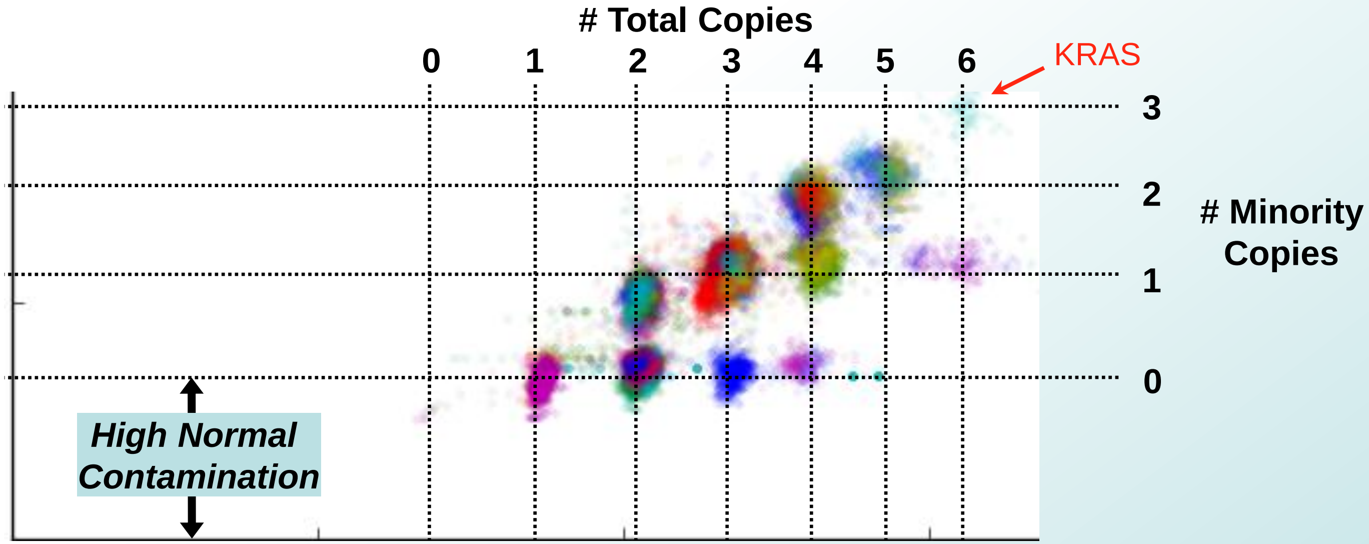
Copy Number Profile Analysis



estimated normal contaminant

Ovarian TCGA-13-1411

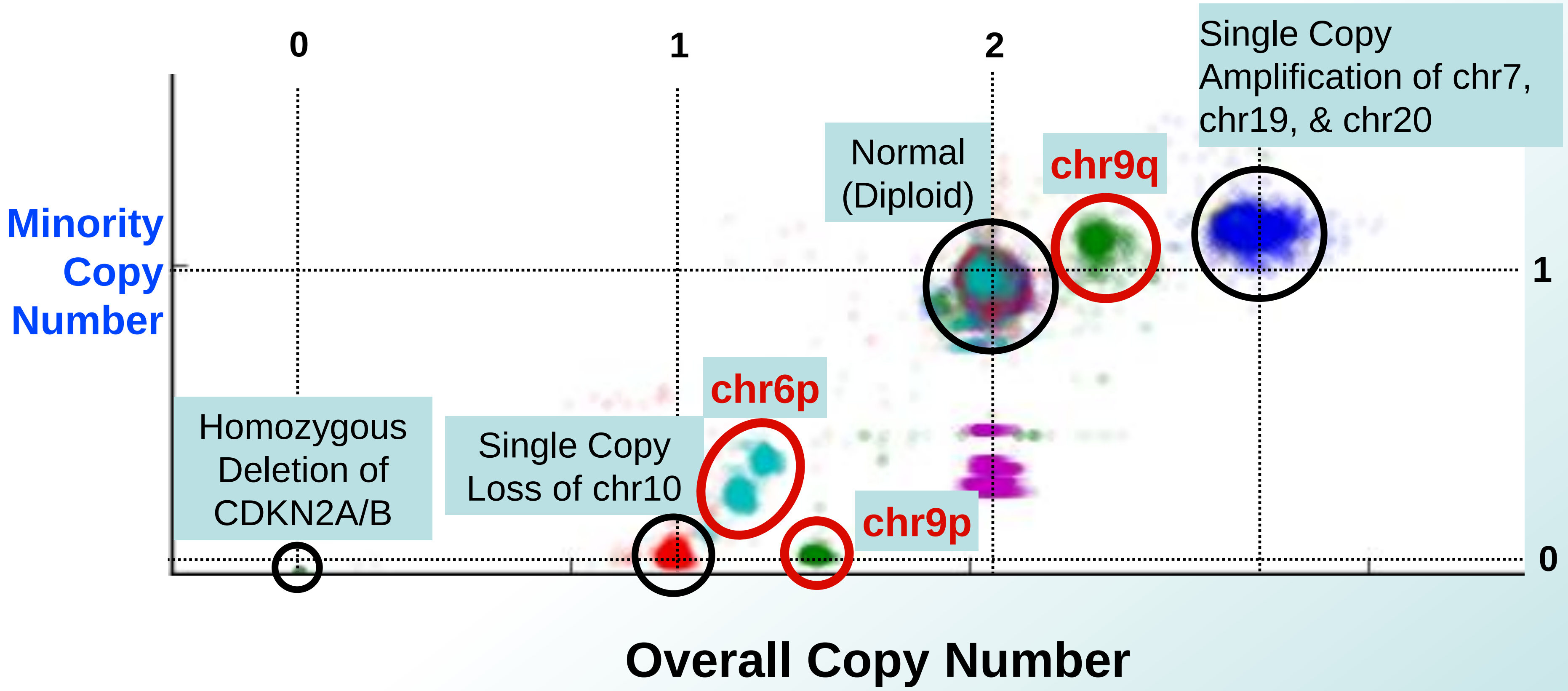
Minority Copy Number



Overall Copy Number

Zack Sanborn

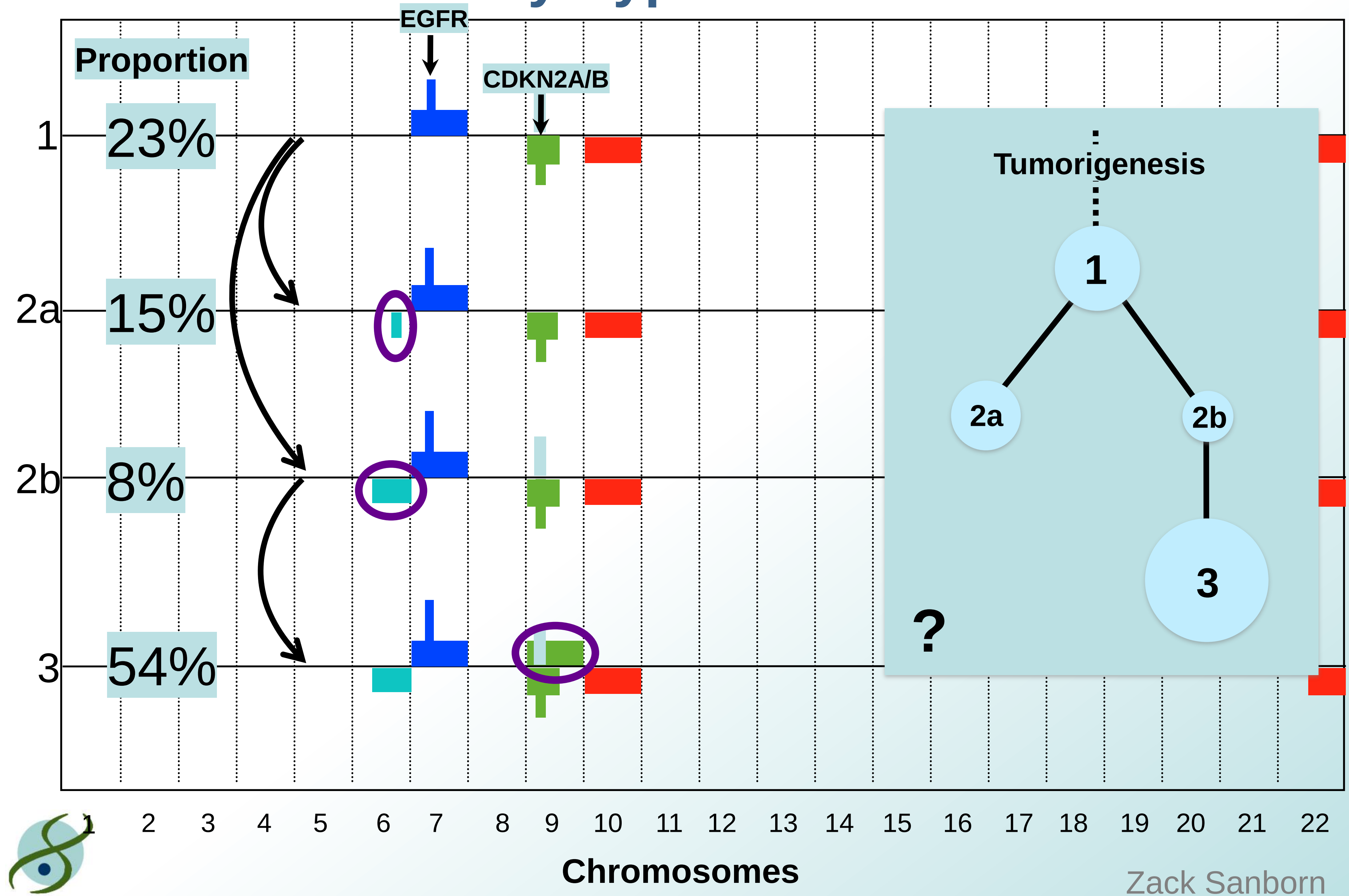
Copy Number States



GBM: TCGA-06-0185

Zack Sanborn

Simulated Progression Model to Infer Karyotype Mixture



Zack Sanborn

General model of genome evolution

A general model of structural variation should include (at least):

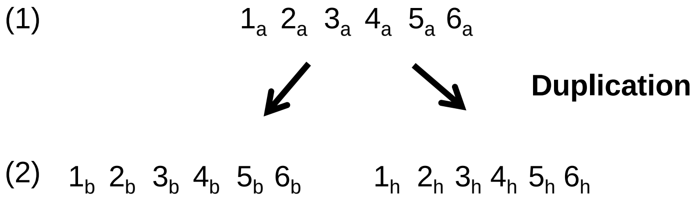
- substitution
- duplication
- rearrangement
- gain and loss

(1) 1_a 2_a 3_a 4_a 5_a 6_a

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



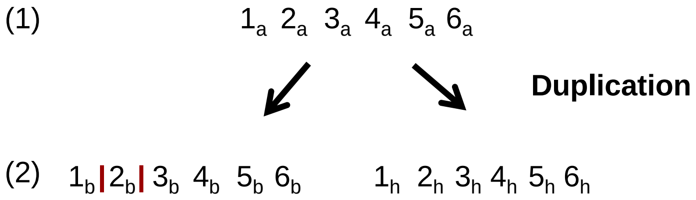
(1) 1_a 2_a 3_a 4_a 5_a 6_a

Duplication

(2) 1_b 2_b 3_b 4_b 5_b 6_b 1_h 2_h 3_h 4_h 5_h 6_h

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



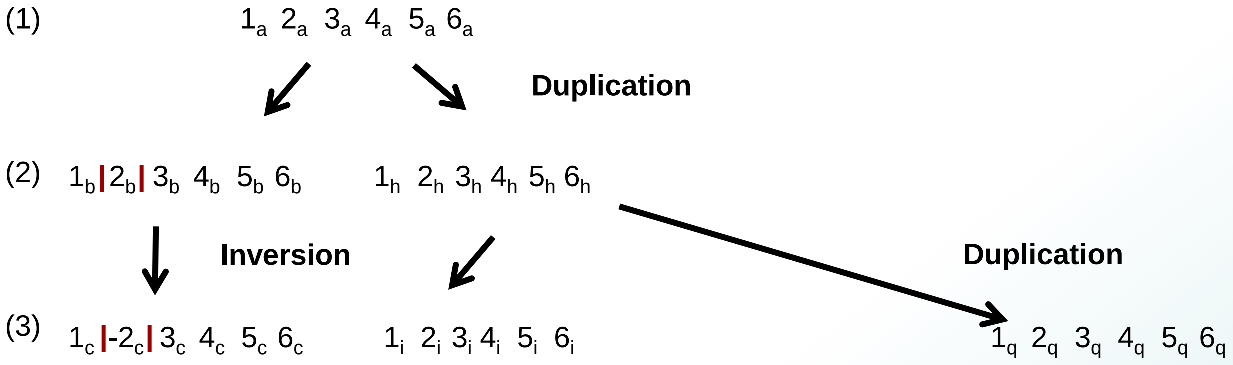
(1) 1_a 2_a 3_a 4_a 5_a 6_a

Duplication

(2) 1_b|2_b| 3_b 4_b 5_b 6_b 1_h 2_h 3_h 4_h 5_h 6_h

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b | 2_b | 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

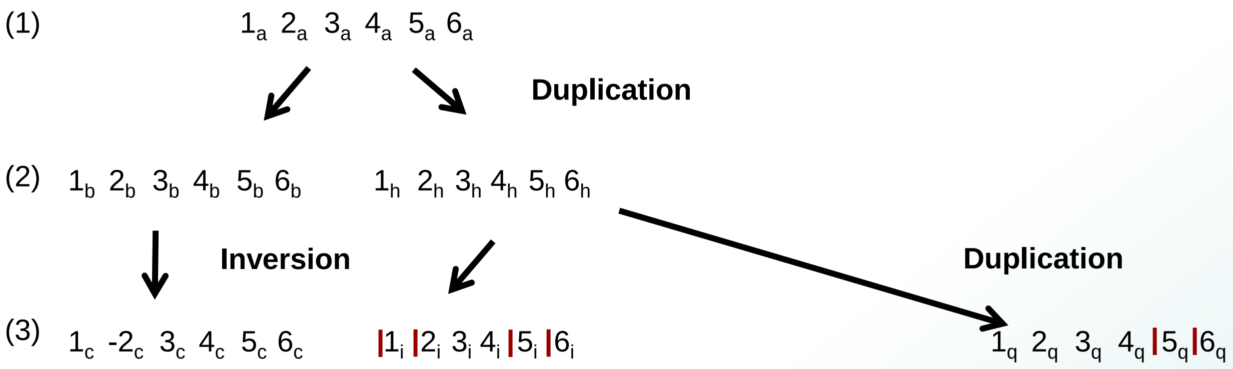
Duplication

(3) $1_c | -2_c | 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) 1_a 2_a 3_a 4_a 5_a 6_a

Duplication

(2) 1_b 2_b 3_b 4_b 5_b 6_b 1_h 2_h 3_h 4_h 5_h 6_h

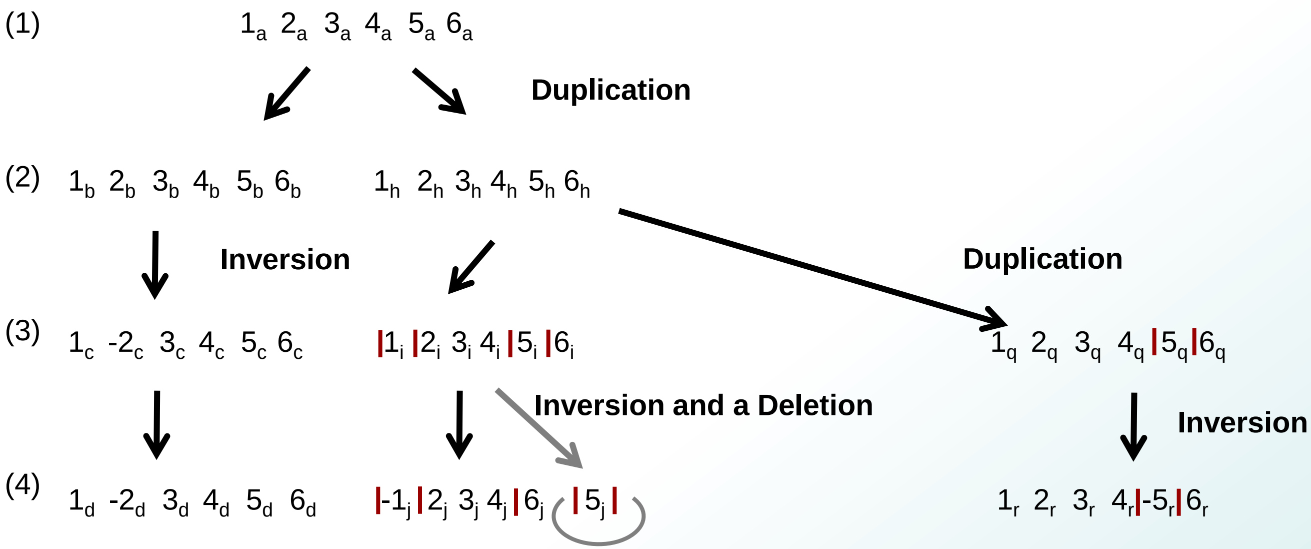
Inversion

Duplication

(3) 1_c -2_c 3_c 4_c 5_c 6_c |1_i |2_i 3_i 4_i |5_i |6_i 1_q 2_q 3_q 4_q |5_q|6_q

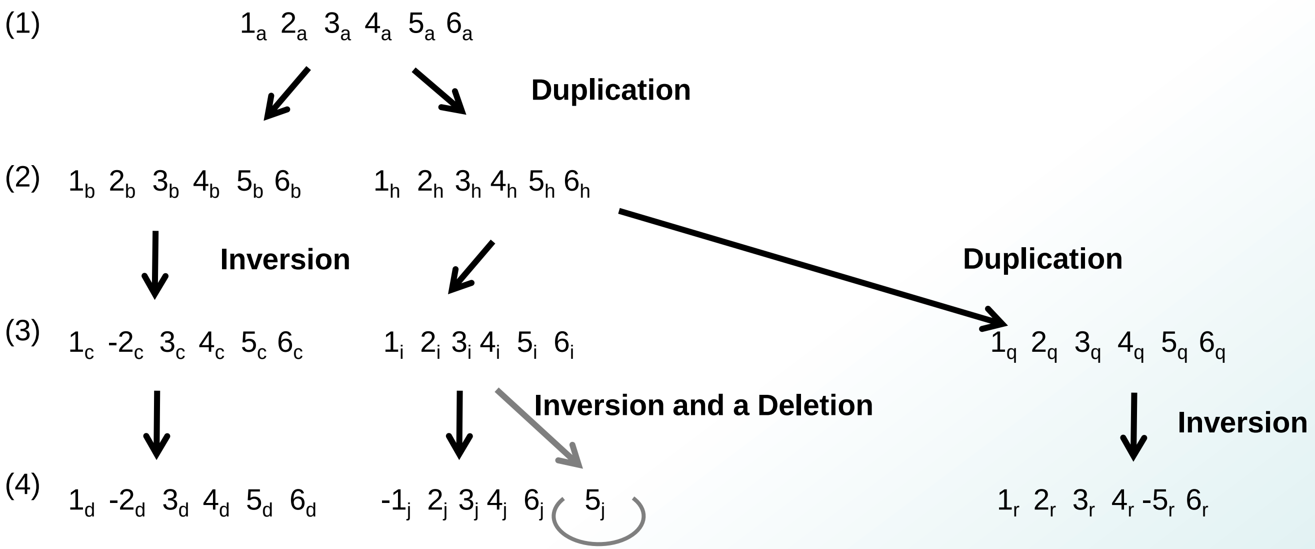
A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss


(1)
1_a 2_a 3_a 4_a 5_a 6_a
Duplication
(2)
1_b 2_b 3_b 4_b 5_b 6_b
1_h 2_h 3_h 4_h 5_h 6_h
Inversion
Duplication
(3)
1_c -2_c 3_c 4_c 5_c 6_c
$|1_i |2_i$ 3_i $4_i |5_i |6_i$
1_q 2_q 3_q $4_q |5_q |6_q$
Inversion and a Deletion
Inversion
(4)
1_d -2_d 3_d 4_d 5_d 6_d
$|-1_j |2_j$ 3_j $4_j |6_j$ $|5_j|$
1_r 2_r 3_r $4_r |-5_r| 6_r$

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b\ 2_b\ 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

Duplication

(3) $1_c\ -2_c\ 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

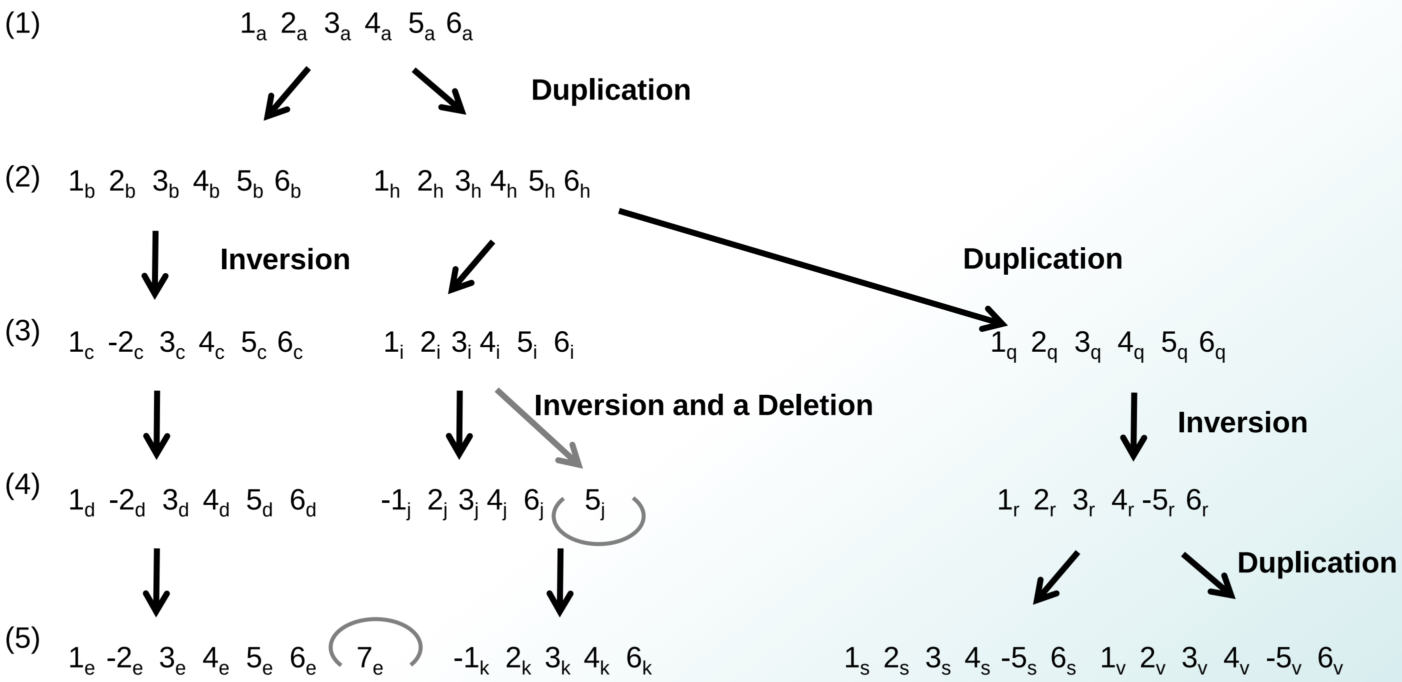
Inversion and a Deletion

Inversion

(4) $1_d\ -2_d\ 3_d\ 4_d\ 5_d\ 6_d$ $-1_j\ 2_j\ 3_j\ 4_j\ 6_j\ 5_j$ $1_r\ 2_r\ 3_r\ 4_r\ -5_r\ 6_r$

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b\ 2_b\ 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

Duplication

(3) $1_c\ -2_c\ 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

Inversion and a Deletion

Inversion

(4) $1_d\ -2_d\ 3_d\ 4_d\ 5_d\ 6_d$ $-1_j\ 2_j\ 3_j\ 4_j\ 6_j\ 5_j$ $1_r\ 2_r\ 3_r\ 4_r\ -5_r\ 6_r$

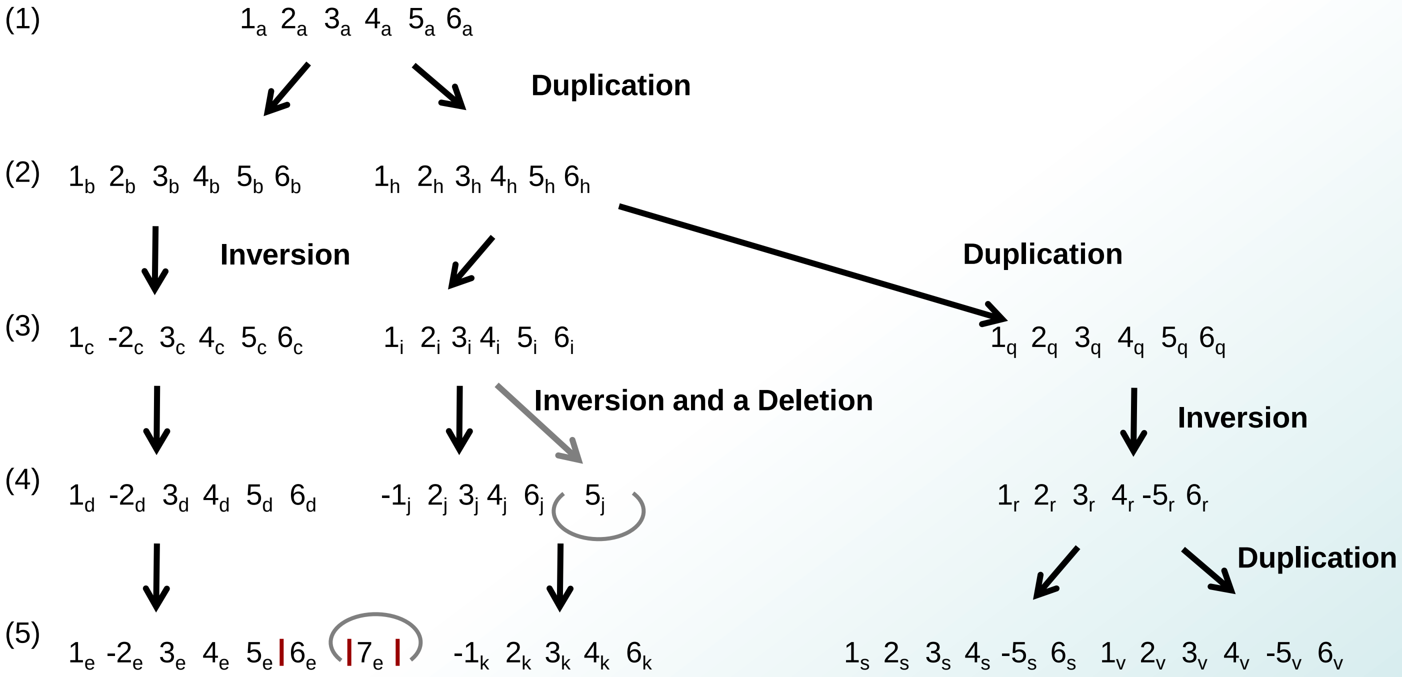
Duplication

(5) $1_e\ -2_e\ 3_e\ 4_e\ 5_e\ 6_e\ 7_e$ $-1_k\ 2_k\ 3_k\ 4_k\ 6_k$ $1_s\ 2_s\ 3_s\ 4_s\ -5_s\ 6_s$ $1_v\ 2_v\ 3_v\ 4_v\ -5_v\ 6_v$

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b\ 2_b\ 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

Duplication

(3) $1_c\ -2_c\ 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

Inversion and a Deletion

Inversion

(4) $1_d\ -2_d\ 3_d\ 4_d\ 5_d\ 6_d$ $-1_j\ 2_j\ 3_j\ 4_j\ 6_j\ 5_j$ $1_r\ 2_r\ 3_r\ 4_r\ -5_r\ 6_r$

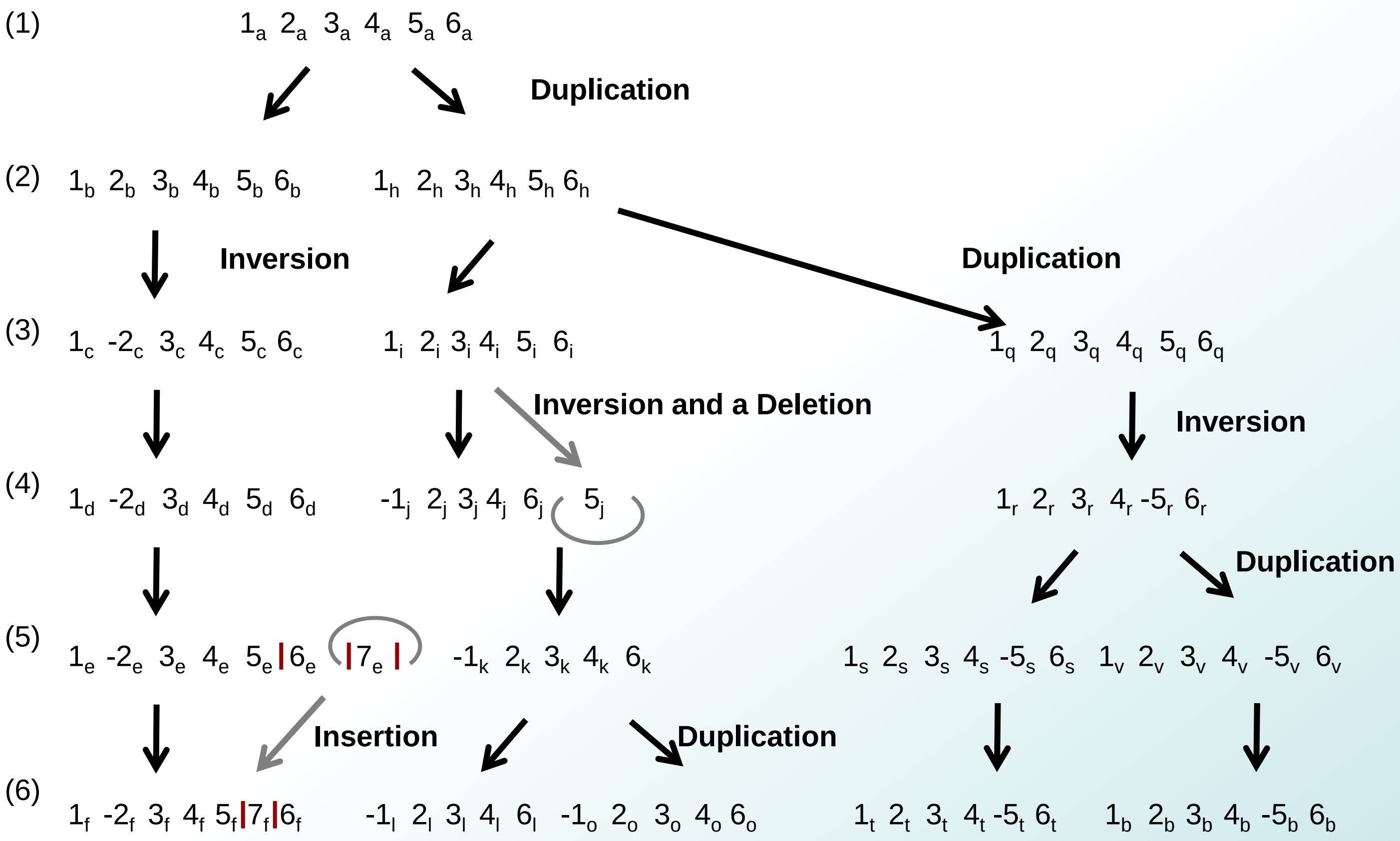
Duplication

(5) $1_e\ -2_e\ 3_e\ 4_e\ 5_e | 6_e\ |7_e |$ $-1_k\ 2_k\ 3_k\ 4_k\ 6_k$ $1_s\ 2_s\ 3_s\ 4_s\ -5_s\ 6_s\ 1_v\ 2_v\ 3_v\ 4_v\ -5_v\ 6_v$

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) 1_a 2_a 3_a 4_a 5_a 6_a

Duplication

(2) 1_b 2_b 3_b 4_b 5_b 6_b 1_h 2_h 3_h 4_h 5_h 6_h

Inversion

Duplication

(3) 1_c -2_c 3_c 4_c 5_c 6_c 1_i 2_i 3_i 4_i 5_i 6_i 1_q 2_q 3_q 4_q 5_q 6_q

Inversion and a Deletion

Inversion

(4) 1_d -2_d 3_d 4_d 5_d 6_d -1_j 2_j 3_j 4_j 6_j 5_j 1_r 2_r 3_r 4_r -5_r 6_r

Duplication

(5) 1_e -2_e 3_e 4_e 5_e|6_e |7_e| -1_k 2_k 3_k 4_k 6_k 1_s 2_s 3_s 4_s -5_s 6_s 1_v 2_v 3_v 4_v -5_v 6_v

Insertion

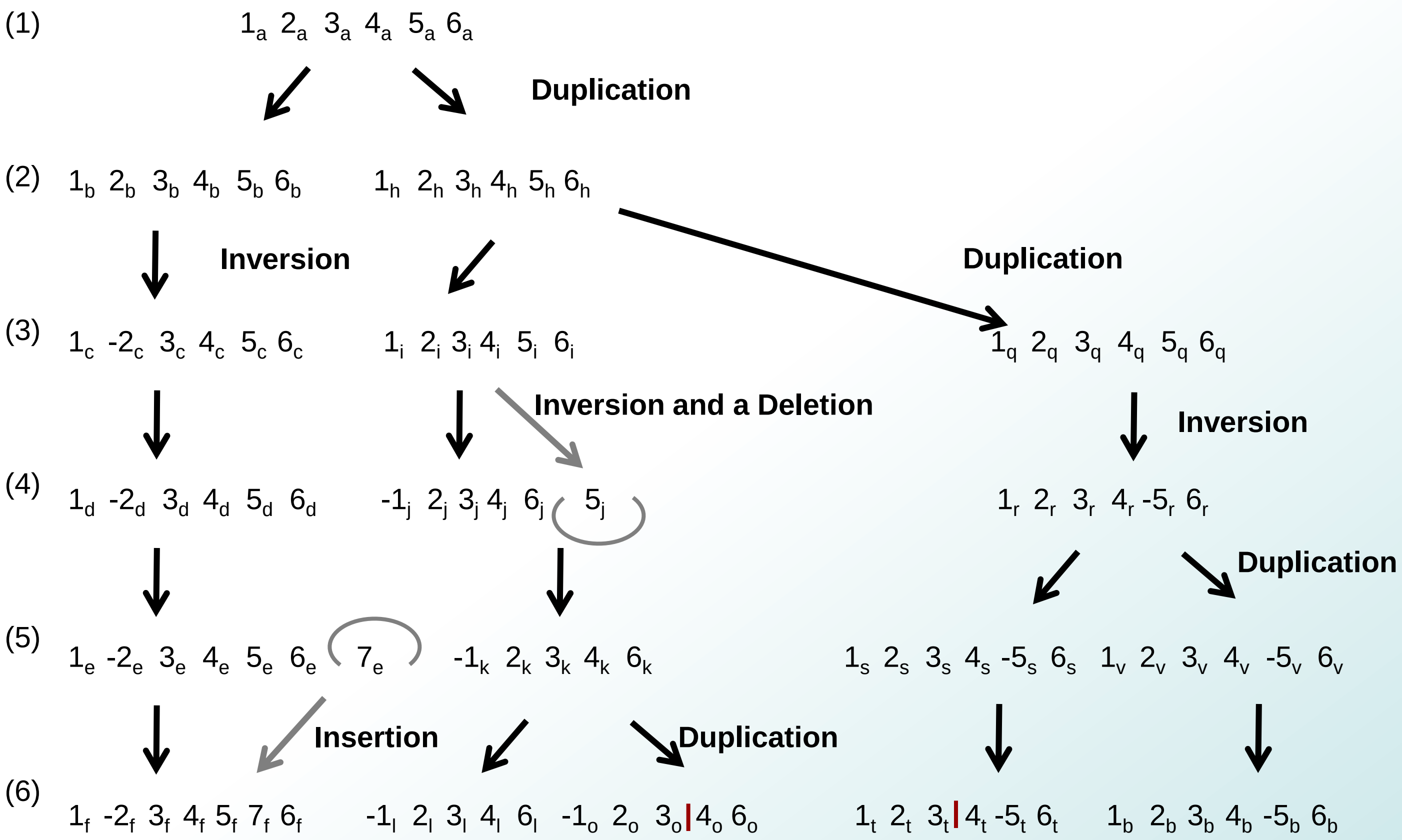
Duplication

(6) 1_f -2_f 3_f 4_f 5_f|7_f|6_f -1_l 2_l 3_l 4_l 6_l -1_o 2_o 3_o 4_o 6_o 1_t 2_t 3_t 4_t -5_t 6_t 1_b 2_b 3_b 4_b -5_b 6_b

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) 1_a 2_a 3_a 4_a 5_a 6_a

Duplication

(2) 1_b 2_b 3_b 4_b 5_b 6_b 1_h 2_h 3_h 4_h 5_h 6_h

Inversion

Duplication

(3) 1_c -2_c 3_c 4_c 5_c 6_c 1_i 2_i 3_i 4_i 5_i 6_i 1_q 2_q 3_q 4_q 5_q 6_q

Inversion and a Deletion

Inversion

(4) 1_d -2_d 3_d 4_d 5_d 6_d -1_j 2_j 3_j 4_j 6_j 5_j 1_r 2_r 3_r 4_r -5_r 6_r

Duplication

(5) 1_e -2_e 3_e 4_e 5_e 6_e 7_e -1_k 2_k 3_k 4_k 6_k 1_s 2_s 3_s 4_s -5_s 6_s 1_v 2_v 3_v 4_v -5_v 6_v

Insertion

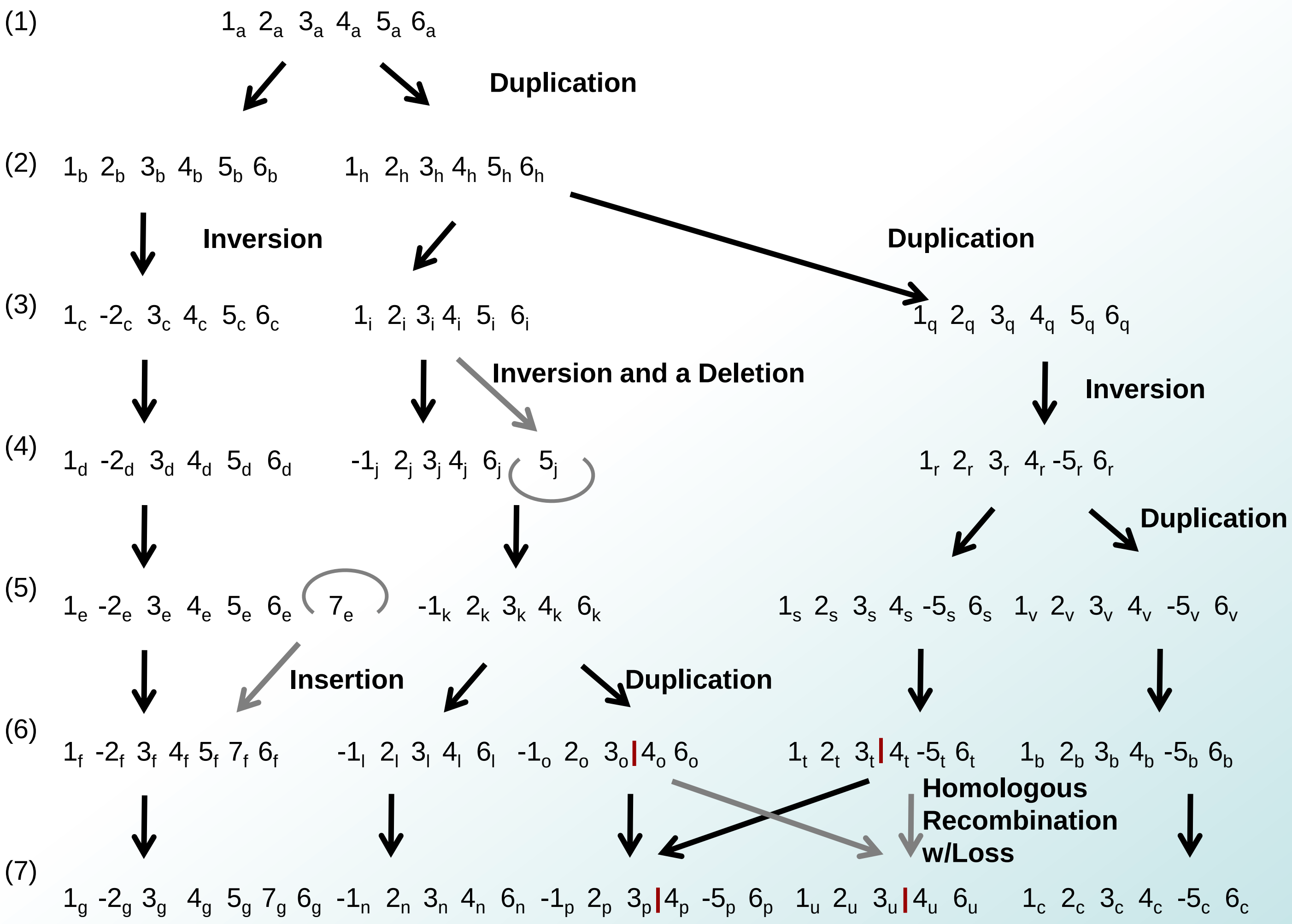
Duplication

(6) 1_f -2_f 3_f 4_f 5_f 7_f 6_f -1_l 2_l 3_l 4_l 6_l -1_o 2_o 3_o | 4_o 6_o 1_t 2_t 3_t | 4_t -5_t 6_t 1_b 2_b 3_b 4_b -5_b 6_b

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b\ 2_b\ 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

Duplication

(3) $1_c\ {-2}_c\ 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

Inversion and a Deletion

Inversion

(4) $1_d\ {-2}_d\ 3_d\ 4_d\ 5_d\ 6_d$ ${-1}_j\ 2_j\ 3_j\ 4_j\ 6_j\ 5_j$ $1_r\ 2_r\ 3_r\ 4_r\ {-5}_r\ 6_r$

Duplication

(5) $1_e\ {-2}_e\ 3_e\ 4_e\ 5_e\ 6_e\ 7_e$ ${-1}_k\ 2_k\ 3_k\ 4_k\ 6_k$ $1_s\ 2_s\ 3_s\ 4_s\ {-5}_s\ 6_s$ $1_v\ 2_v\ 3_v\ 4_v\ {-5}_v\ 6_v$

Insertion

Duplication

(6) $1_f\ {-2}_f\ 3_f\ 4_f\ 5_f\ 7_f\ 6_f$ ${-1}_l\ 2_l\ 3_l\ 4_l\ 6_l$ ${-1}_o\ 2_o\ 3_o | 4_o\ 6_o$ $1_t\ 2_t\ 3_t | 4_t\ {-5}_t\ 6_t$ $1_b\ 2_b\ 3_b\ 4_b\ {-5}_b\ 6_b$

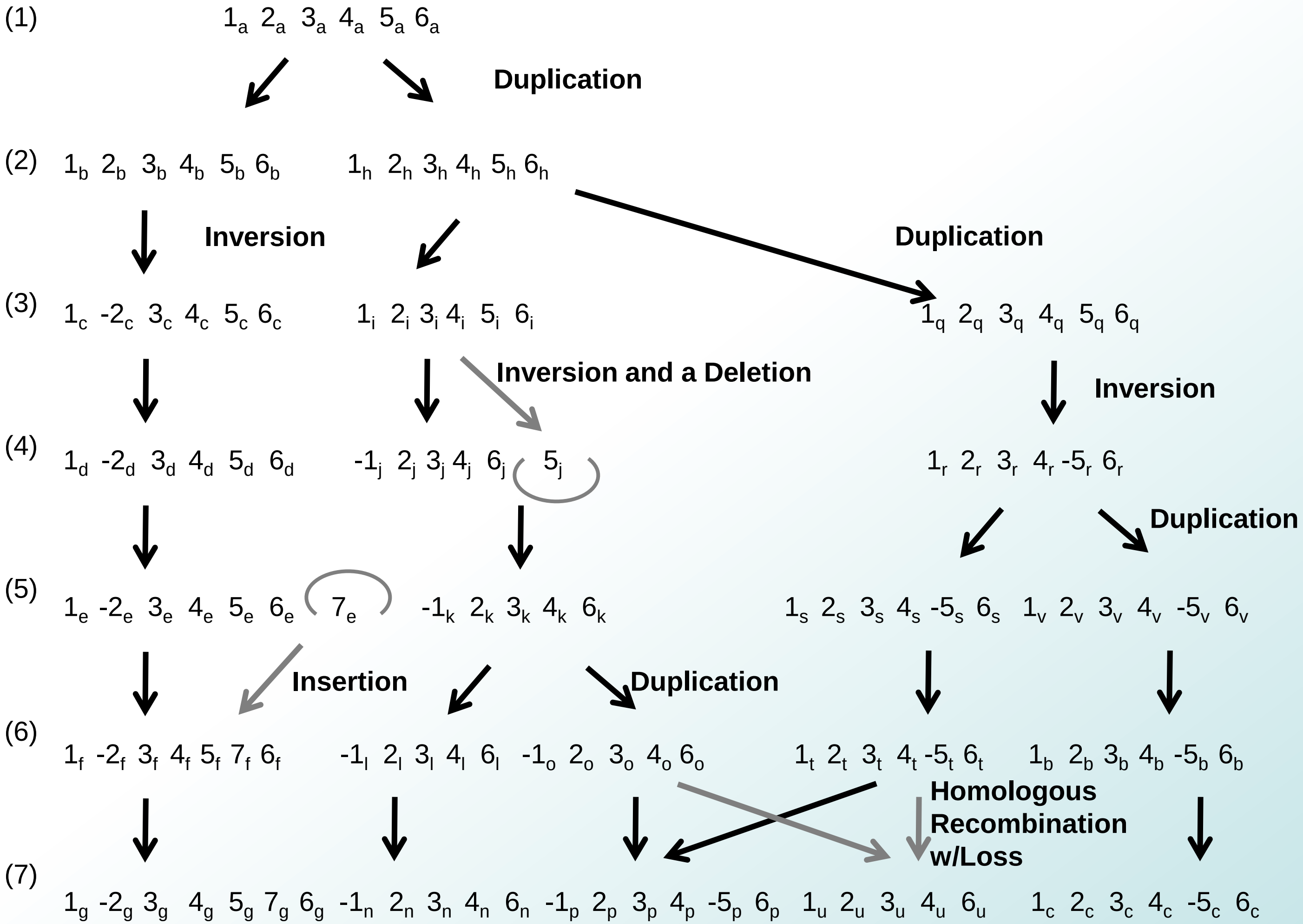
Homologous Recombination w/Loss

(7) $1_g\ {-2}_g\ 3_g\ 4_g\ 5_g\ 7_g\ 6_g$ ${-1}_n\ 2_n\ 3_n\ 4_n\ 6_n$ ${-1}_p\ 2_p\ 3_p | 4_p\ {-5}_p\ 6_p$ $1_u\ 2_u\ 3_u | 4_u\ 6_u$ $1_c\ 2_c\ 3_c\ 4_c\ {-5}_c\ 6_c$

Benedict Paten, Daniel Zerbino

A general model of structural variation should include (at least):

- substitution
- duplication
- rearrangement
- gain and loss



(1) $1_a\ 2_a\ 3_a\ 4_a\ 5_a\ 6_a$

Duplication

(2) $1_b\ 2_b\ 3_b\ 4_b\ 5_b\ 6_b$ $1_h\ 2_h\ 3_h\ 4_h\ 5_h\ 6_h$

Inversion

Duplication

(3) $1_c\ -2_c\ 3_c\ 4_c\ 5_c\ 6_c$ $1_i\ 2_i\ 3_i\ 4_i\ 5_i\ 6_i$ $1_q\ 2_q\ 3_q\ 4_q\ 5_q\ 6_q$

Inversion and a Deletion

Inversion

(4) $1_d\ -2_d\ 3_d\ 4_d\ 5_d\ 6_d$ $-1_j\ 2_j\ 3_j\ 4_j\ 6_j\ 5_j$ $1_r\ 2_r\ 3_r\ 4_r\ -5_r\ 6_r$

Duplication

(5) $1_e\ -2_e\ 3_e\ 4_e\ 5_e\ 6_e\ 7_e$ $-1_k\ 2_k\ 3_k\ 4_k\ 6_k$ $1_s\ 2_s\ 3_s\ 4_s\ -5_s\ 6_s$ $1_v\ 2_v\ 3_v\ 4_v\ -5_v\ 6_v$

Insertion

Duplication

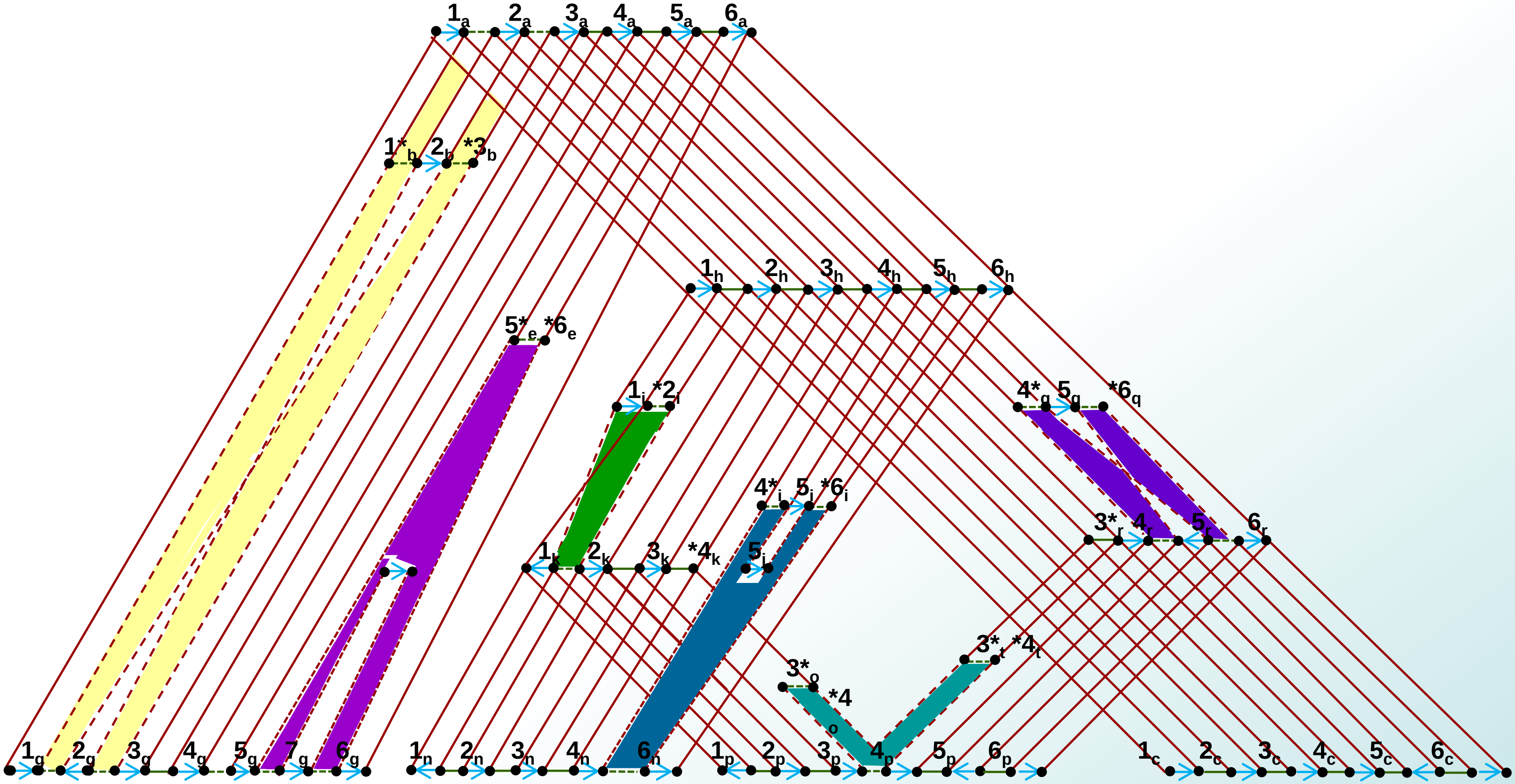
(6) $1_f\ -2_f\ 3_f\ 4_f\ 5_f\ 7_f\ 6_f$ $-1_l\ 2_l\ 3_l\ 4_l\ 6_l$ $-1_o\ 2_o\ 3_o\ 4_o\ 6_o$ $1_t\ 2_t\ 3_t\ 4_t\ -5_t\ 6_t$ $1_b\ 2_b\ 3_b\ 4_b\ -5_b\ 6_b$

Homologous Recombination w/Loss

(7) $1_g\ -2_g\ 3_g\ 4_g\ 5_g\ 7_g\ 6_g$ $-1_n\ 2_n\ 3_n\ 4_n\ 6_n$ $-1_p\ 2_p\ 3_p\ 4_p\ -5_p\ 6_p$ $1_u\ 2_u\ 3_u\ 4_u\ 6_u$ $1_c\ 2_c\ 3_c\ 4_c\ -5_c\ 6_c$

Benedict Paten, Daniel Zerbino

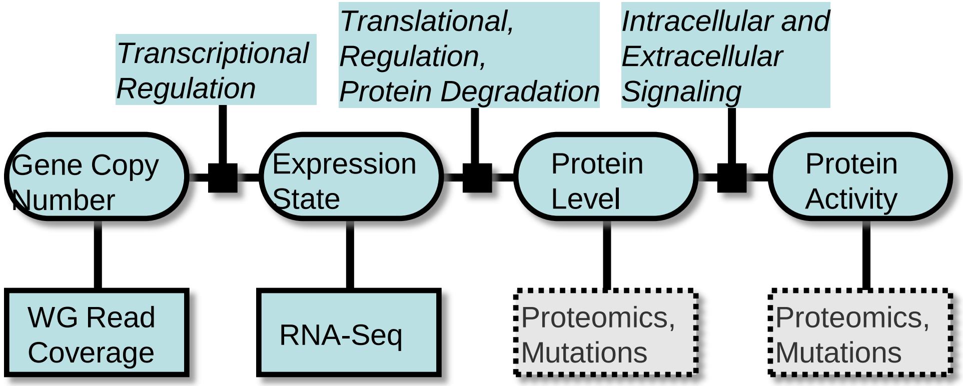
Ancestral Variation Graphs (AVGs)


1_a 2_a 3_a 4_a 5_a 6_a
$1*_b$ 2_b $*3_b$
1_h 2_h 3_h 4_h 5_h 6_h
$5*_e$ $*6_e$
1_i $*2_i$
$4*_q$ 5_q $*6_q$
$4*_j$ 5_j $*6_j$
1_k 2_k 3_k $*4_k$ 5_l
$3*_r$ 4_r 5_r 6_r
$3*_t$ $*4_t$
$3*_o$ $*4_o$
1_g 2_g 3_g 4_g 5_g 7_g 6_g
1_n 2_n 3_n 4_n 6_n
1_p 2_p 3_p 4_p 5_p 6_p
1_c 2_c 3_c 4_c 5_c 6_c

- Graph theoretic model
- Tractable framework for inference, modeling and reasoning
- Allows missing data and partial inference

Benedict Paten, Daniel Zerbino

Next: Biological and Clinical interpretation of the data



Transcriptional Regulation

Translational, Regulation, Protein Degradation

Intracellular and Extracellular Signaling

Gene Copy Number

Expression State

Protein Level

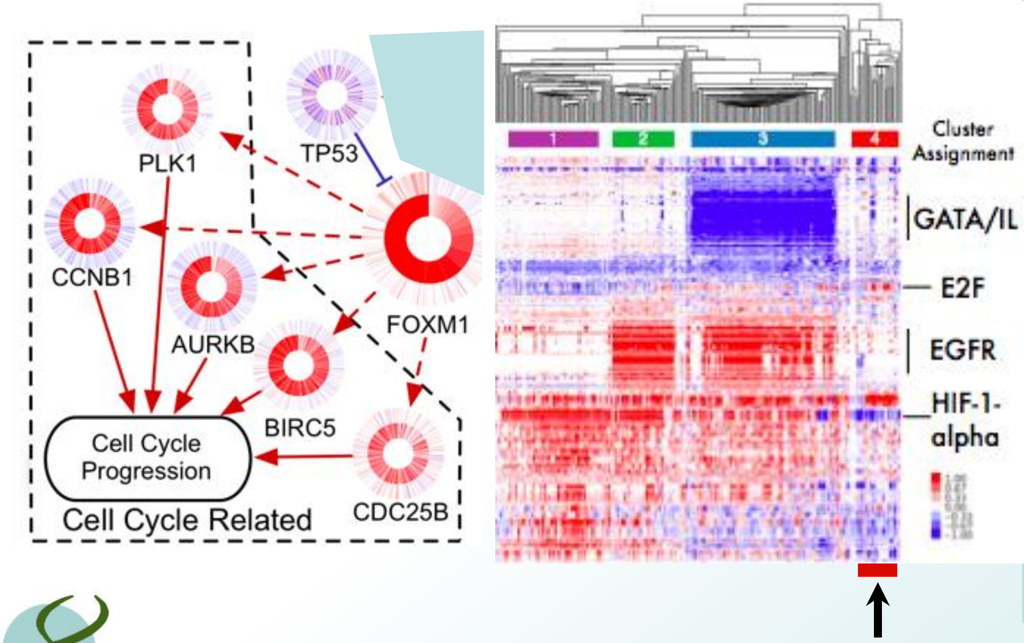
Protein Activity

WG Read Coverage

RNA-Seq

Proteomics, Mutations

Proteomics, Mutations

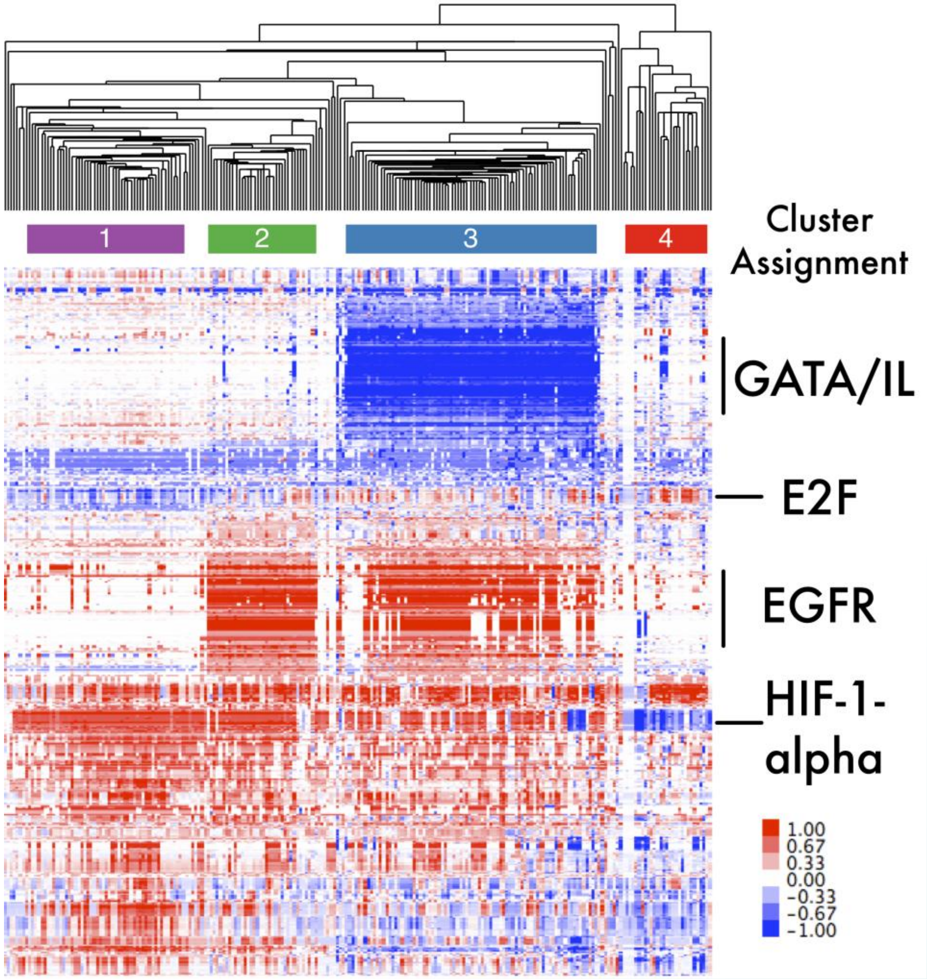


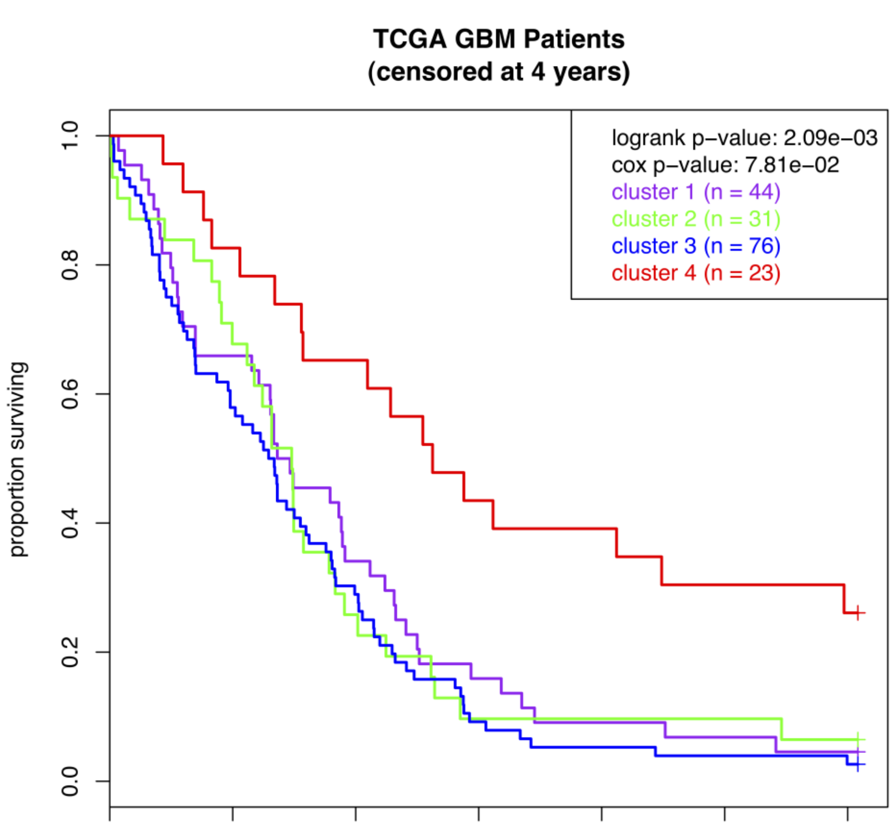
Patients with Good Prognosis

- PARADIGM is an example of a patient-specific inference model
- Identifies biological processes that are abnormally activated or suppressed in each patient
- Many types of genomics information aggregated in a biologically relevant manner
- “Central Dogma”-based graphical model
- Various gene interactions incorporated including transcriptional and post-transcriptional.

Josh Stuart group; Vaske et al. 2010

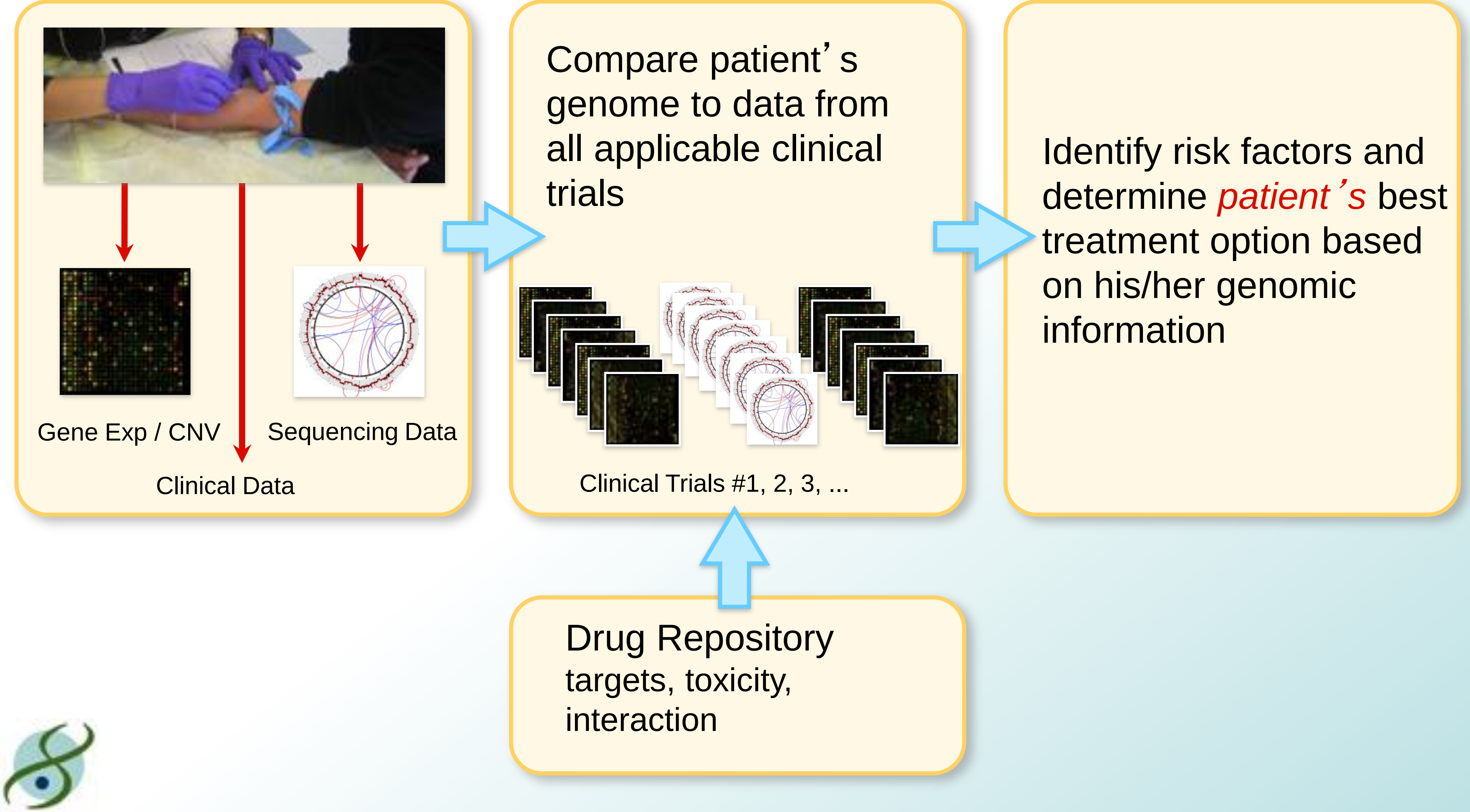
Clustering Activity Vectors Stratifies Glioblastoma Patients by Survival Time





Josh Stuart, Stephen Benz, Charles Vaske

One Goal: Targeted Cancer Treatment



Gene Exp / CNV

Sequencing Data

Clinical Data

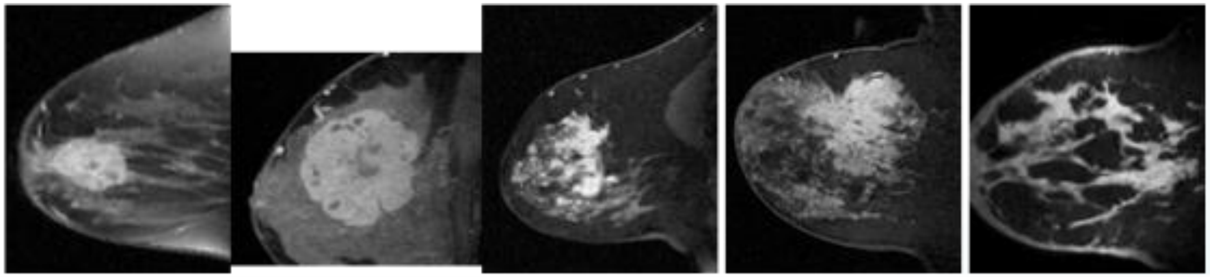
Compare patient’ s genome to data from all applicable clinical trials

Clinical Trials #1, 2, 3, ...

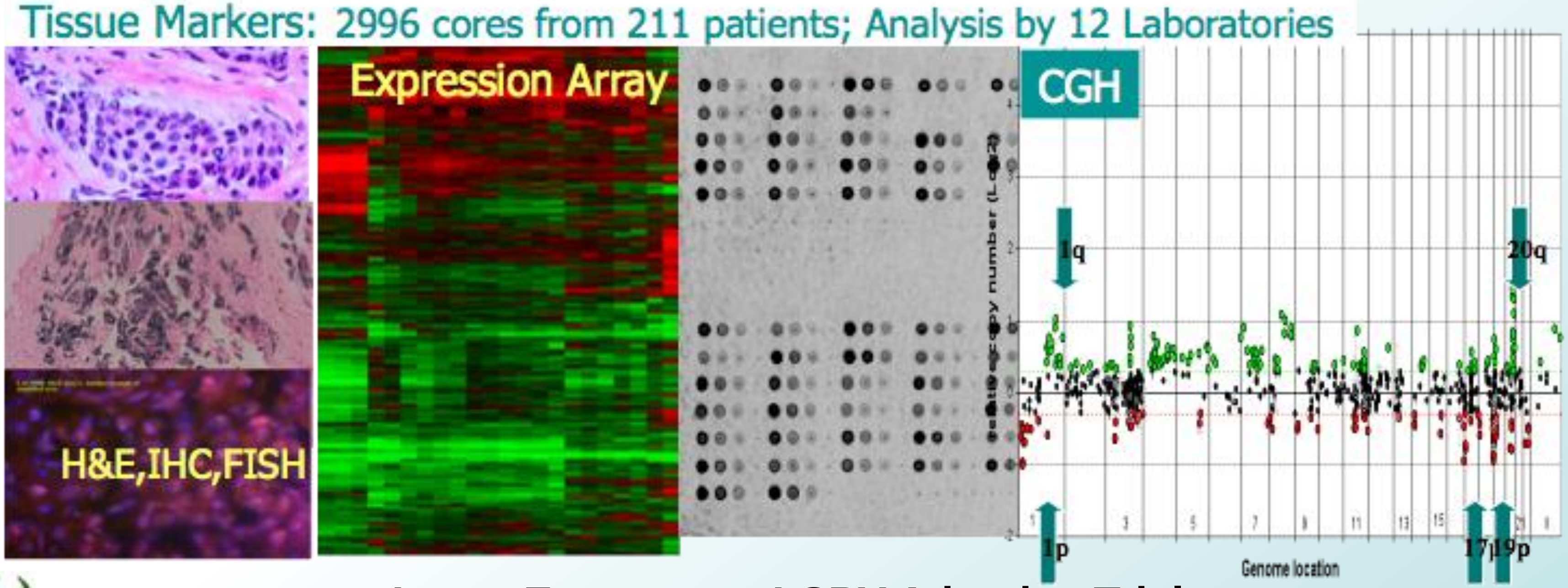
Identify risk factors and determine *patient’s* best treatment option based on his/her genomic information

Drug Repository
targets, toxicity, interaction

Too Many Biomarkers and Treatment Strategies to Test in Conventional Clinical Trials



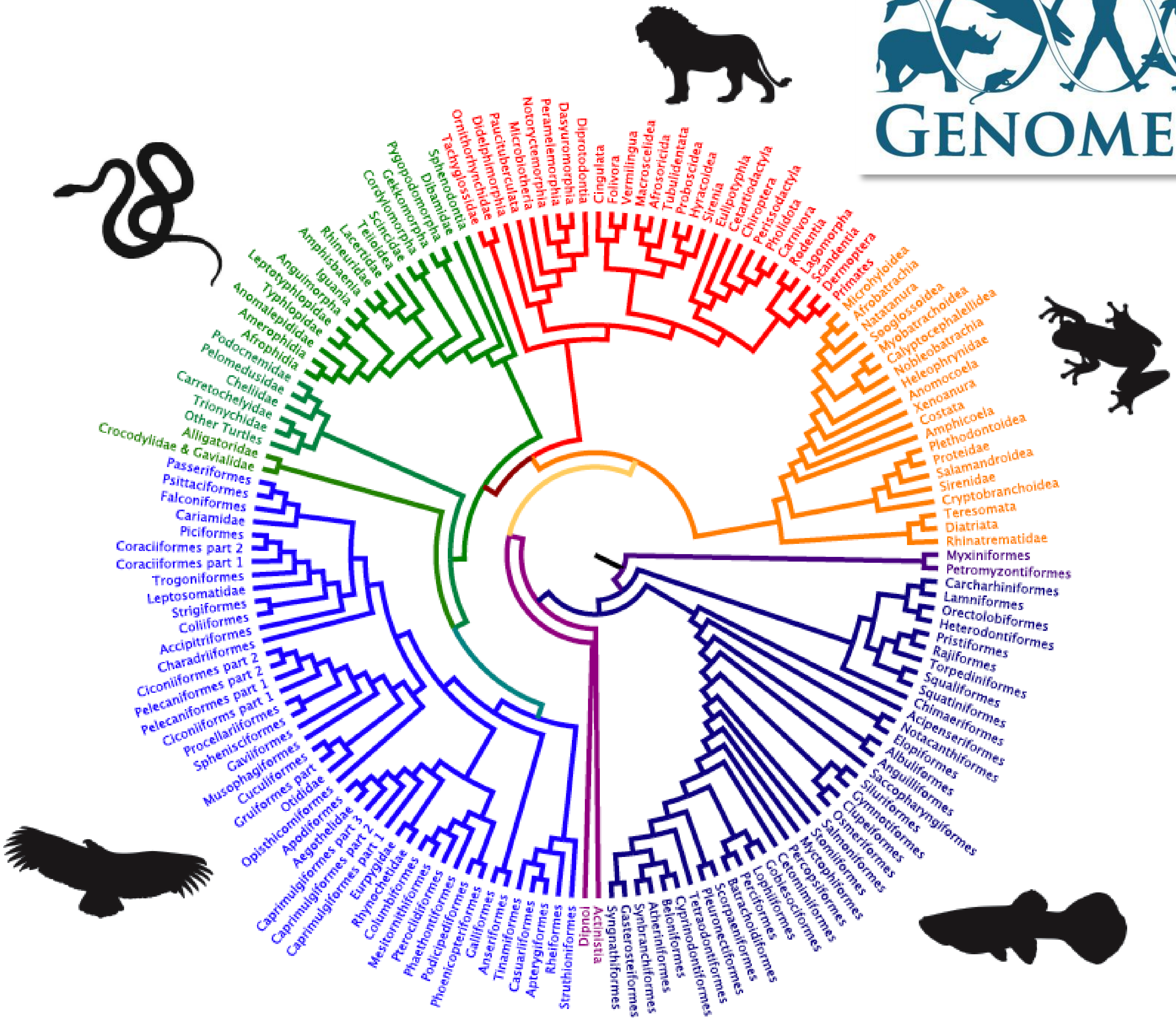
Tissue Markers: 2996 cores from 211 patients; Analysis by 12 Laboratories



Laura Esserman - I-SPY Adaptive Trial

breast surgeon and oncologist at UCSF

GENOME 10K©



The G10K Community of Scientists

The Genome 10K Project:

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

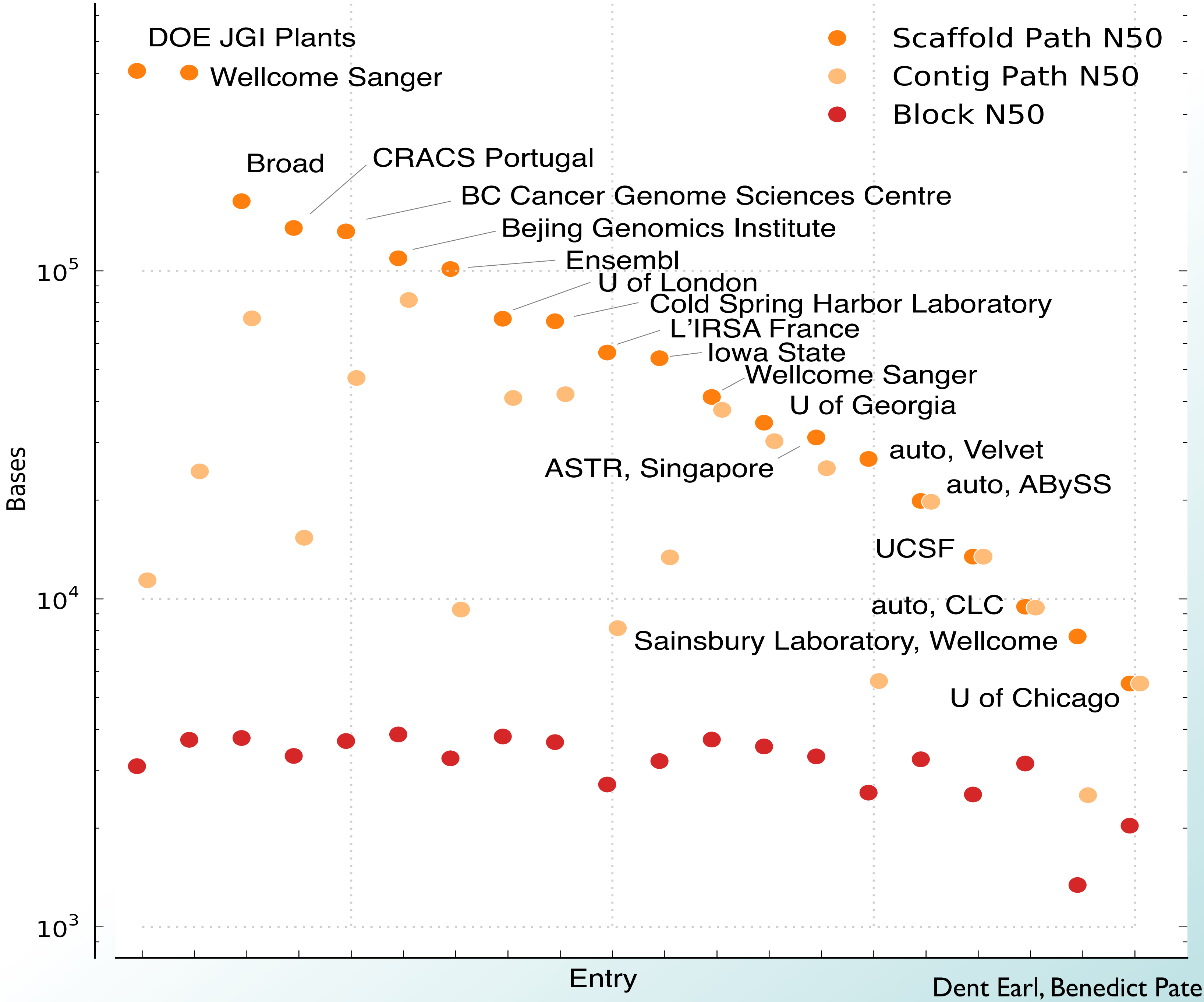
- Collect samples and sequence at least 10,000 different vertebrate species, bank fibroblast cell lines and make iPS lines for > 1,000. Currently ~200 genomes and ~20 iPS lines in progress.
- Annotate genomes, map and interpret genetic differences between species, and compute the evolutionary record 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

GENOME 10K©

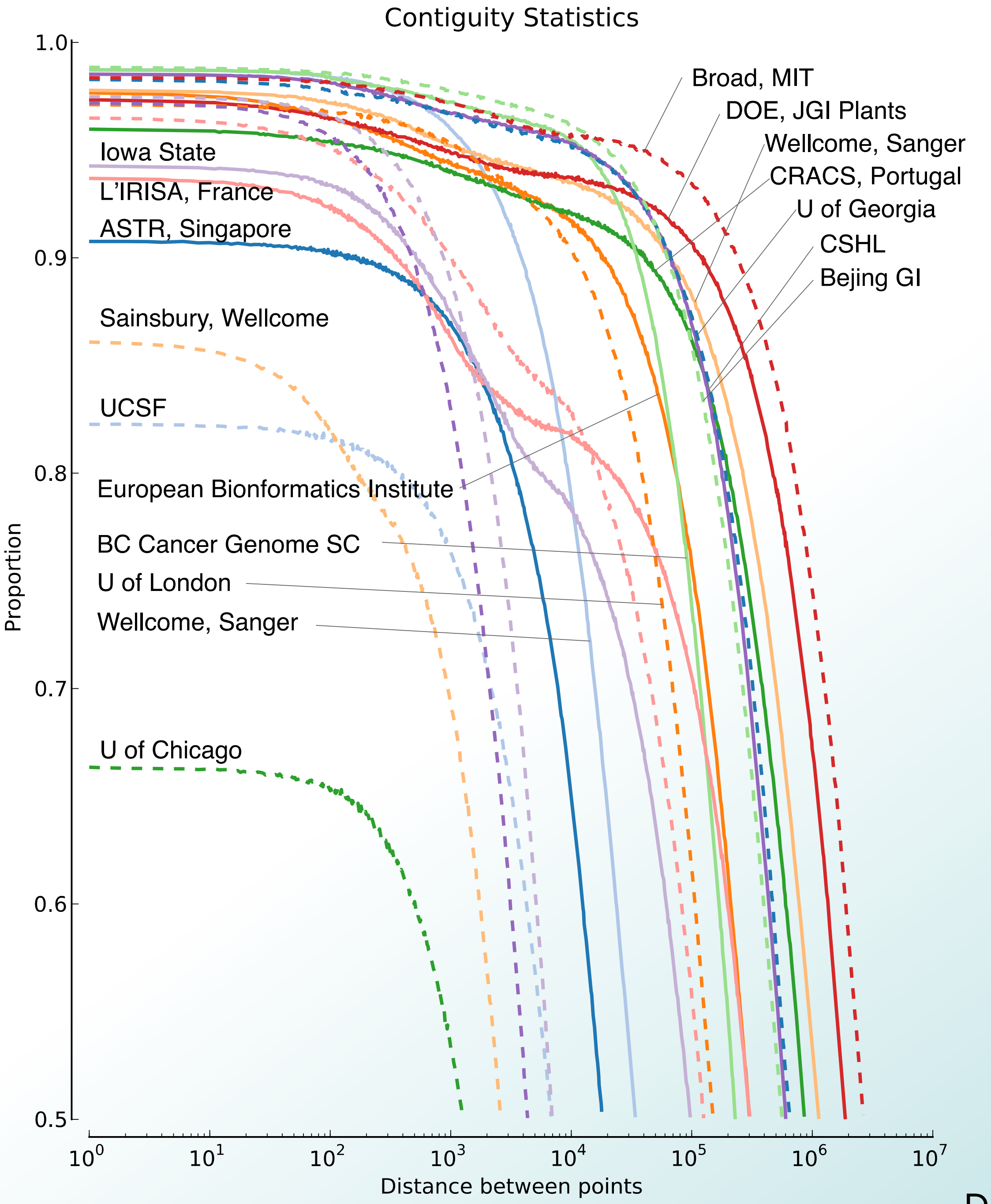
Assemblathon 1: Assembly teams competed in March to assemble a novel genome from short reads

- Organized by Joe DeRisi (UCSF), Jasper Rine(UCB) and David Haussler (UCSC), lead by Ian Korf, UCD and Benedict Paten, UCSC.
- 39 assemblies from 17 teams, ~150 attendees
- Simulated Illumina reads were used, planned follow-up challenge to include other technologies.
- Revealed limitations of short reads in genome sequencing and highlighted software challenges

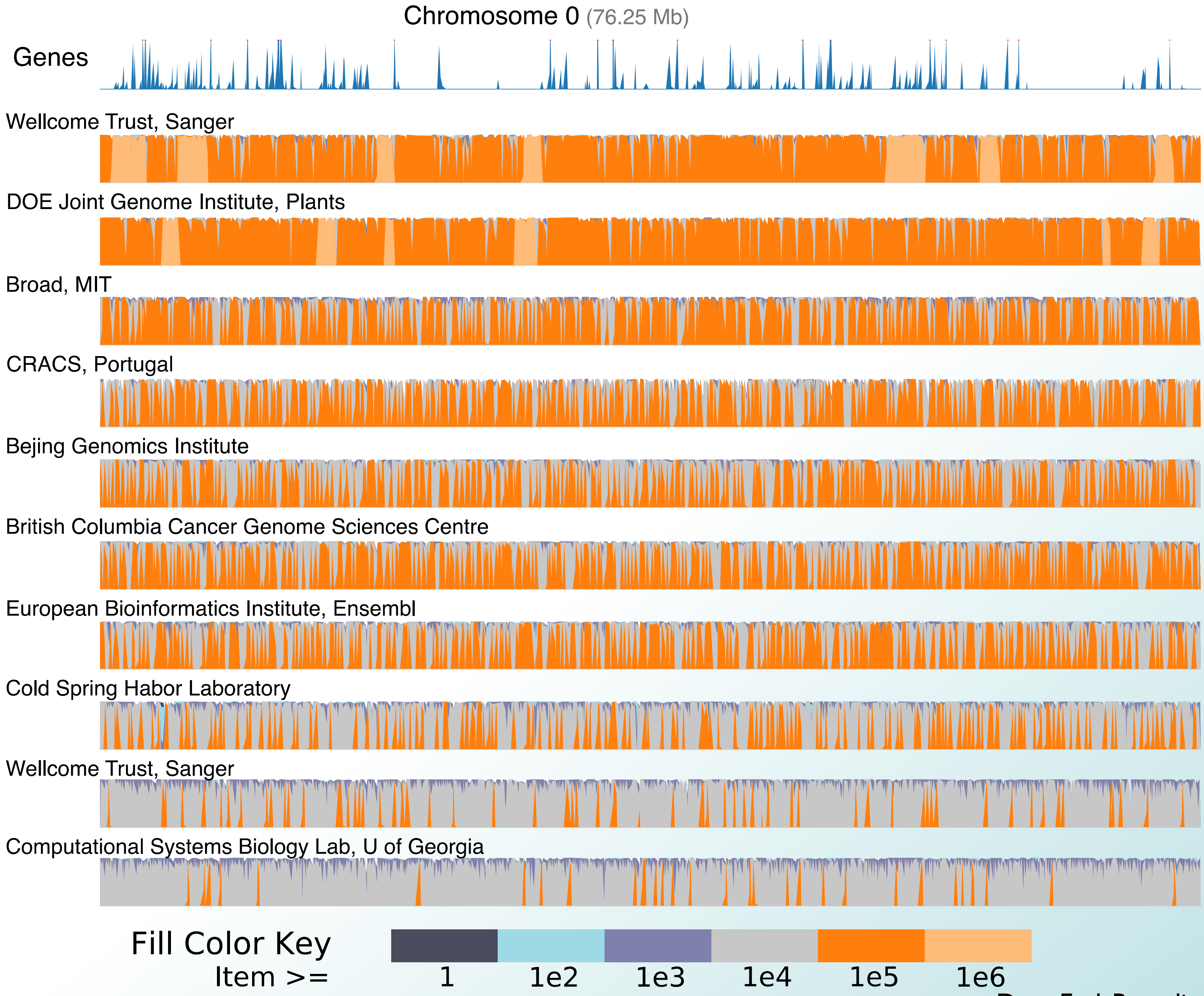
Average size of assembled piece (N50)



Dent Earl, Benedict Paten



Dent Earl, Benedict Paten



Dent Earl, Benedict Paten

THE CANCER GENOME ATLAS

STAND UP TO CANCER

NATIONAL INSTITUTES OF HEALTH

qb3 ucb-ucsc-ucsf

UCSC Cancer Genomics & Genome Analysis

- **Josh Stuart**
- Jingchun Zhu
- Zack Sanborn
- Steve Benz
- Charles Vaske
- Brian Craft
- Christopher Szeto
- Larry Meyer
- Sofie Salama
- Tracy Ballinger
- Mia Grifford
- Benedict Paten
- Daniel Zerbino
- Kord Kober
- Kyle Ellortt
- Mary Goldman
- James Durbin
- Amy Radenbaugh
- Chris Wilks
- Jim Kent
- UCSC Genome Browser Staff

Collaborators

- Stand Up To Cancer
- The Cancer Genome Atlas
- Intl. Cancer Genomics Consortium
- Christopher Benz, Buck Institute
- Laura Esserman, UCSF
- Joe Gray, LBL
- Eric Collisson, UCSF

Funding Agencies

- NCI/NIH
- NHGRI
- American Association for Cancer Research
- UCSF Comprehensive Cancer Center
- California Institute for Quantitative Biosciences (QB3)

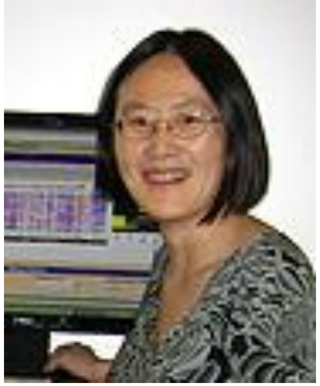
CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING
promoting discovery and invention for human health and well-being

UC SANTA CRUZ

UCSC Cancer Integration Group

Josh Stuart, Co-PI

Jing Zhu

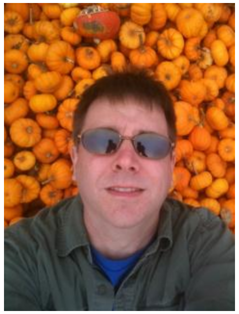


Charlie Vaske

Steve Benz

Zack Sanborn

James Durbin



Larry Meyer

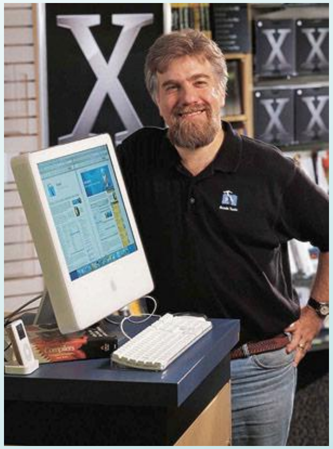
Chris Szeto

Sam Ng

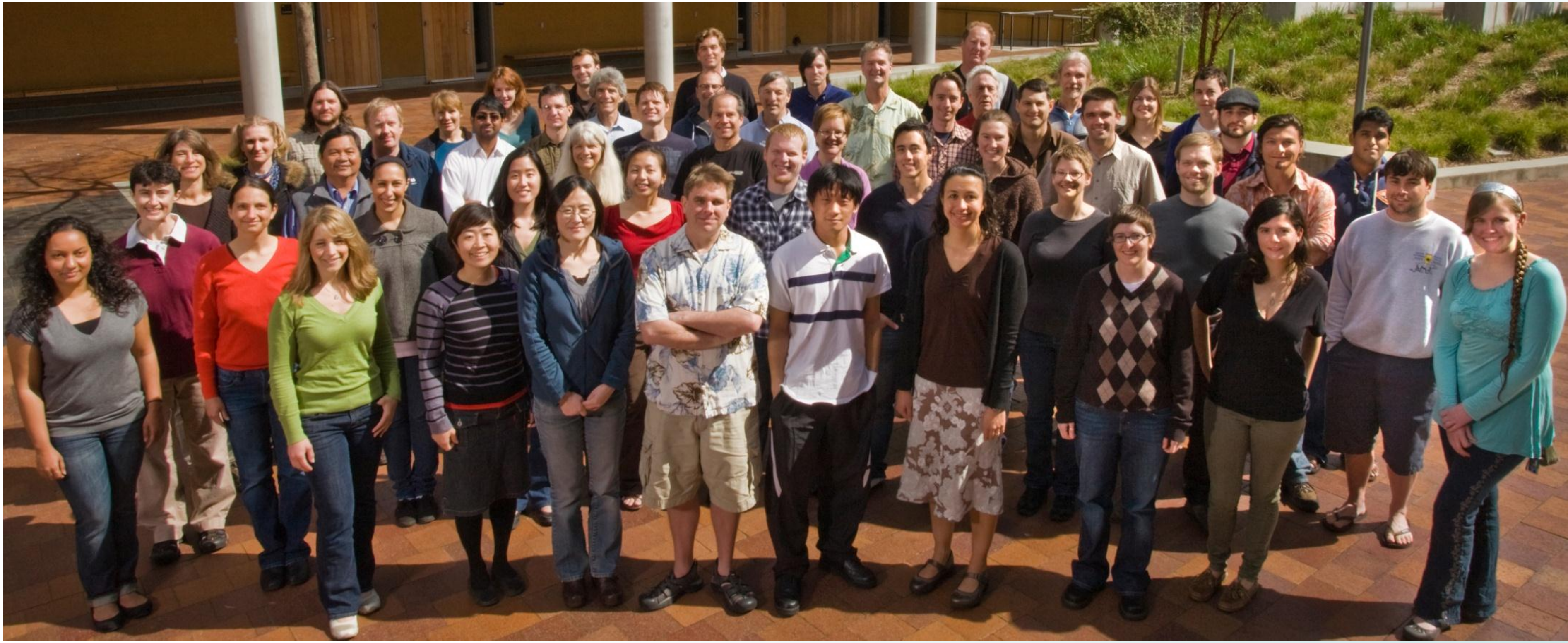
Mia Grifford

Amie Radenbaugh

Ted Golstein



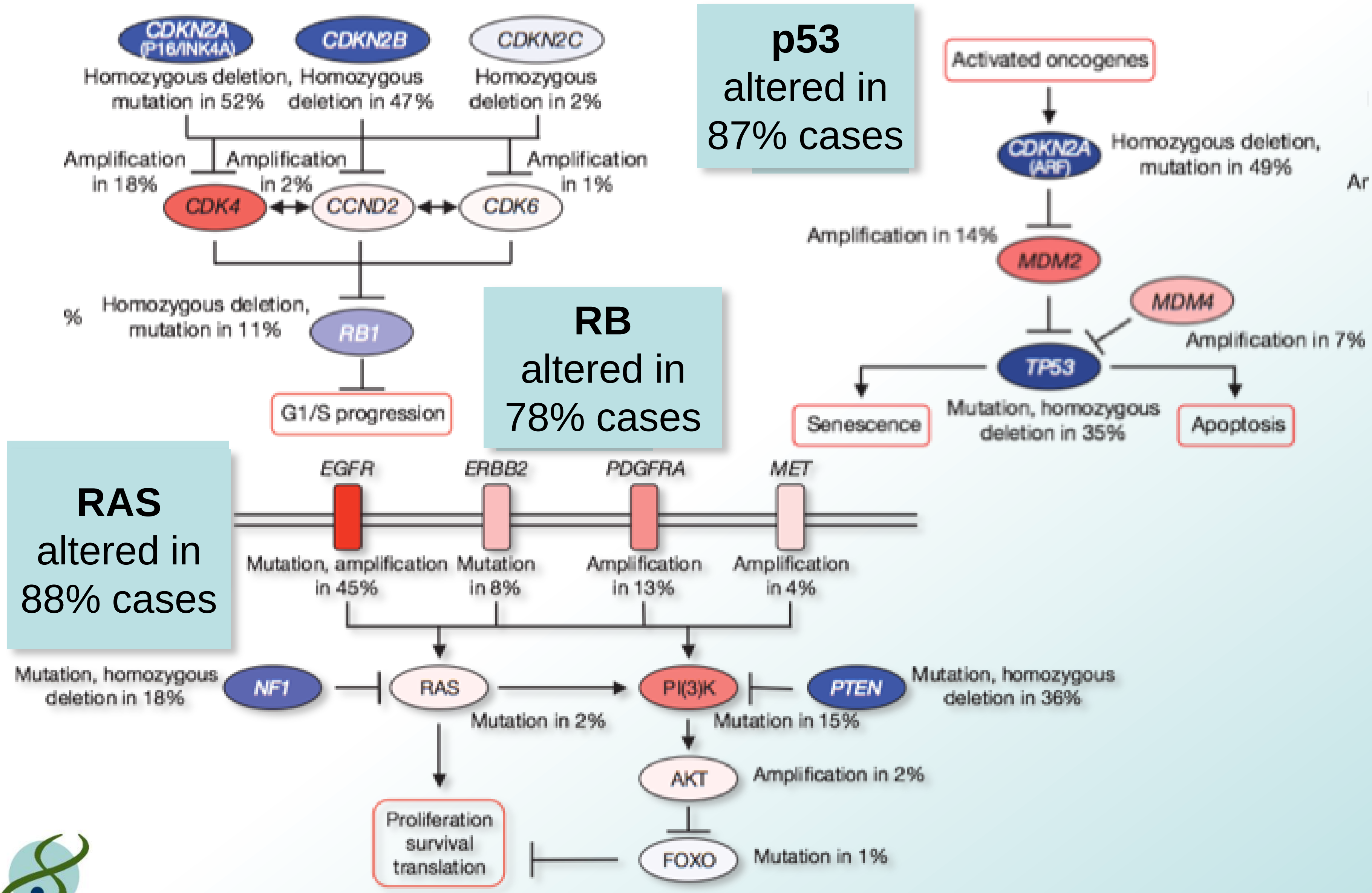
The UCSC team



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

UC SANTA CRUZ

Pathways as Genetic Units: GBM analysis



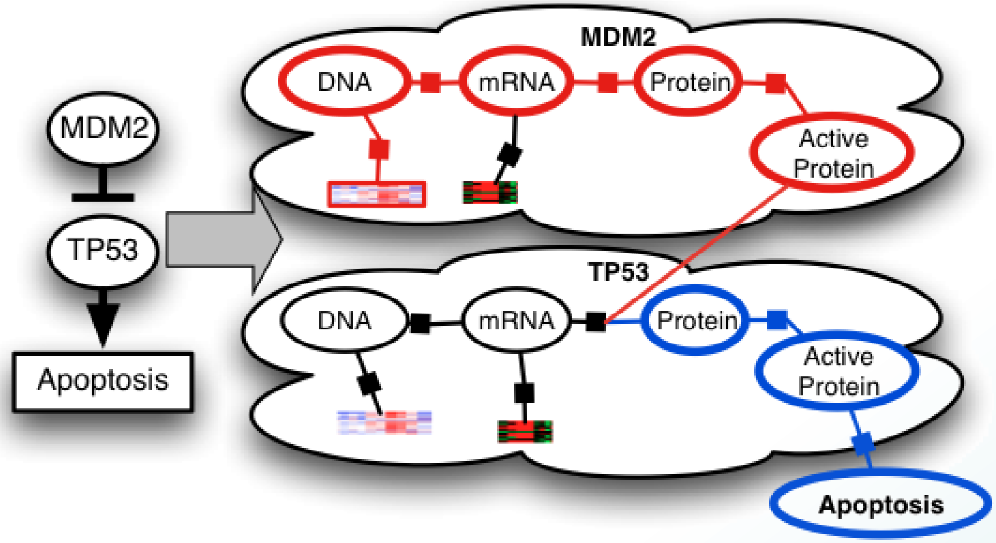
p53
altered in
87% cases

RB
altered in
78% cases

RAS
altered in
88% cases

The Cancer Genome Atlas, *Nature*, 2008

Integrated Pathway Activity (IPA)



- Abstract notion of a biological entity’s activity within a pathway context
- Calculate log-likelihood ratio (LLR) of *up (red)*, *same (white)*, and *down* (blue) states
- IPA is the largest of the three LLRs, multiplied by the sign of the state

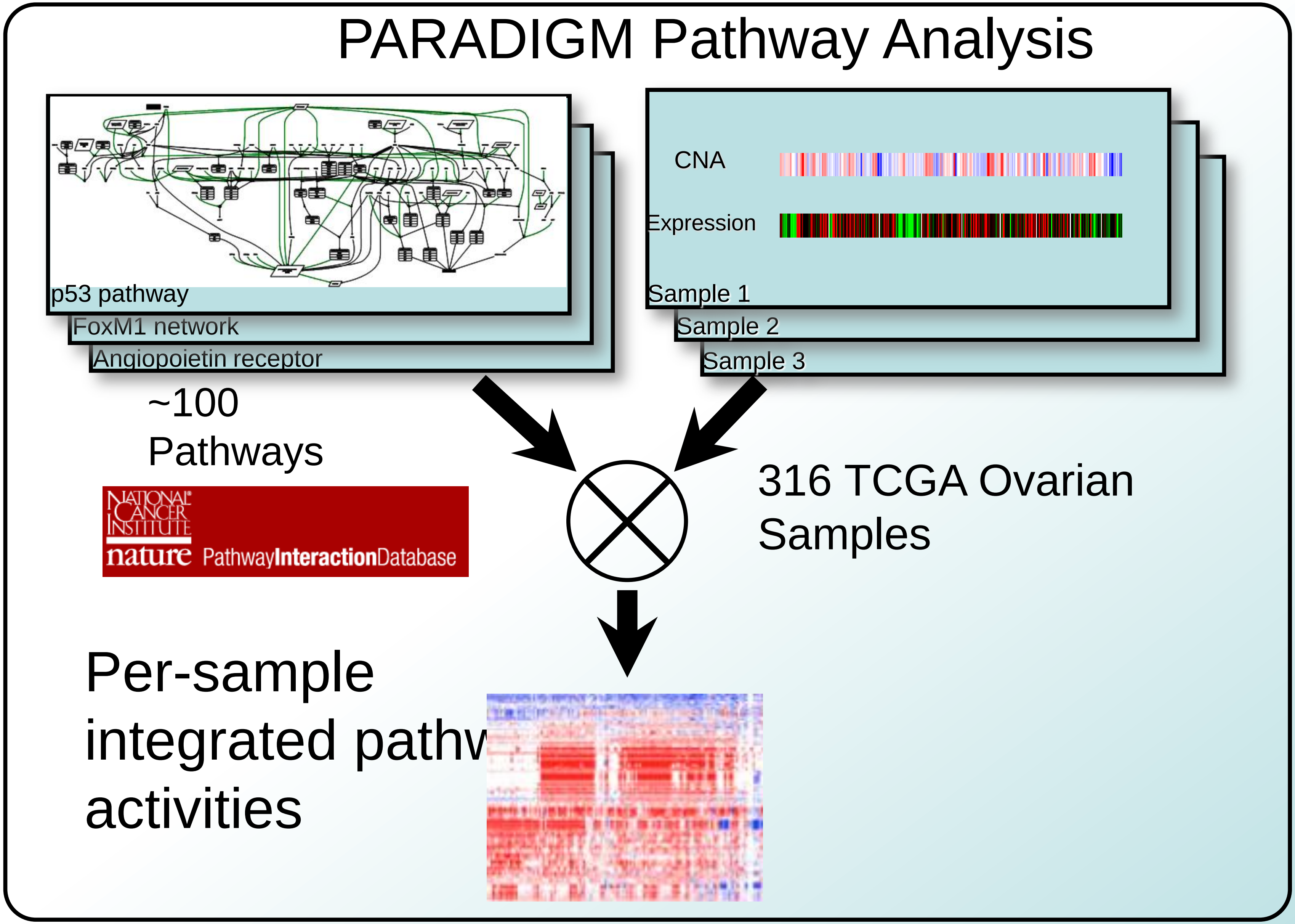
$$\log_{10}\frac{P(Data|TP53=up,\Phi)}{P(Data|TP53\neq up,\Phi)}=\log_{10}\frac{P(Data,TP53=up|\Phi)}{P(Data,TP53\neq up|\Phi)}-\log_{10}\frac{P(TP53=up|\Phi)}{P(TP53,\neq up|\Phi)}$$

log odds of state and data

prior log odds

Josh Stuart group

Pathway Interpretation of Omics Data



PARADIGM Pathway Analysis

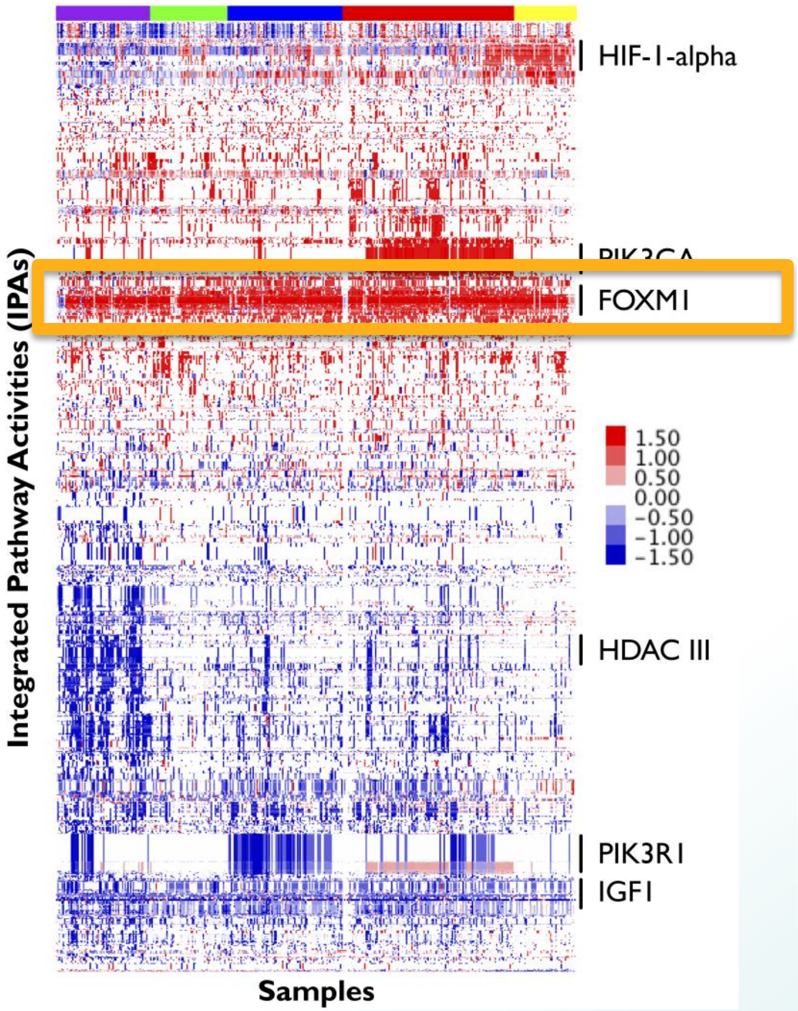
~100 Pathways

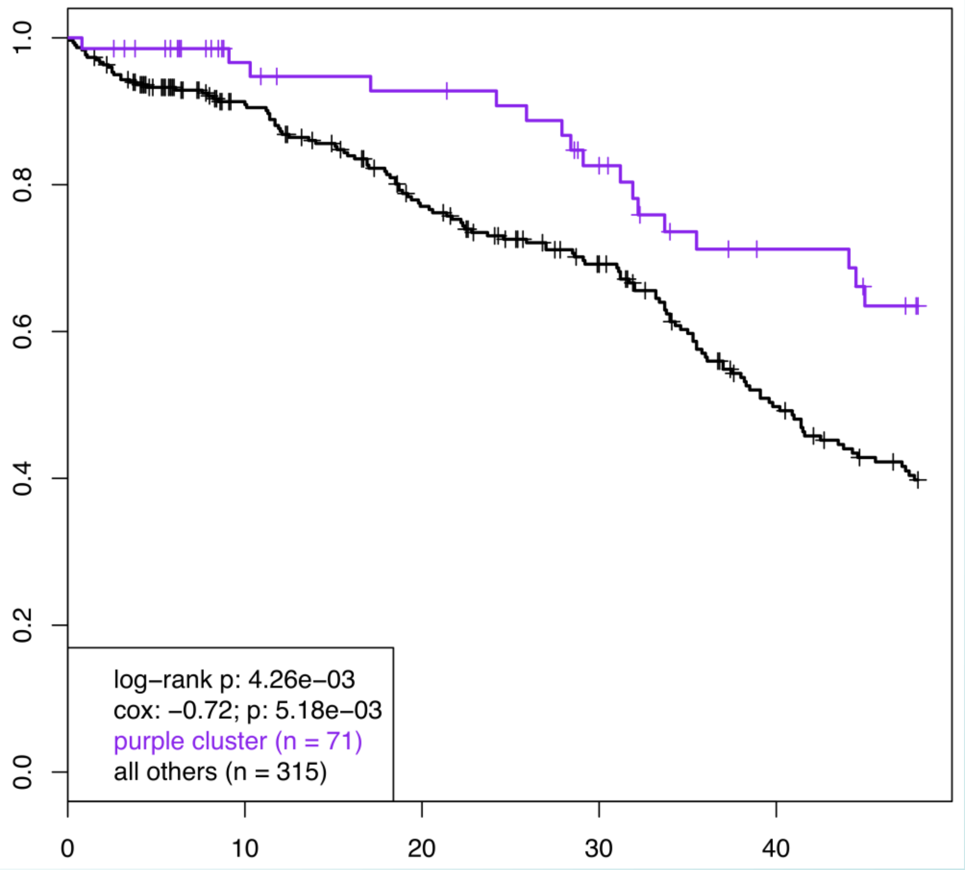
NATIONAL CANCER INSTITUTE nature PathwayInteractionDatabase

316 TCGA Ovarian Samples

Per-sample integrated pathw activities

IPAs identify FOXM1 as key player and stratify ovarian cancer patients by survival time



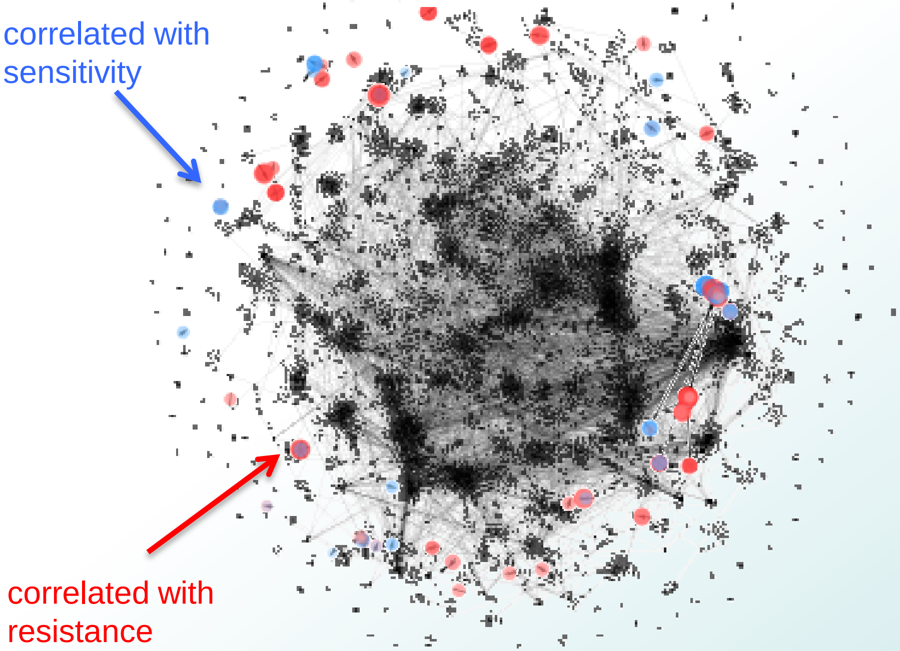


Josh Stuart, Stephen Benz, Charles Vaske

Pathway Markers of Drug Response

- Infer activities in a global “super pathway”
- Associate activities with drug sensitivity to find activity markers; similar for resistance
- Search for focused subnetworks with interconnected markers.

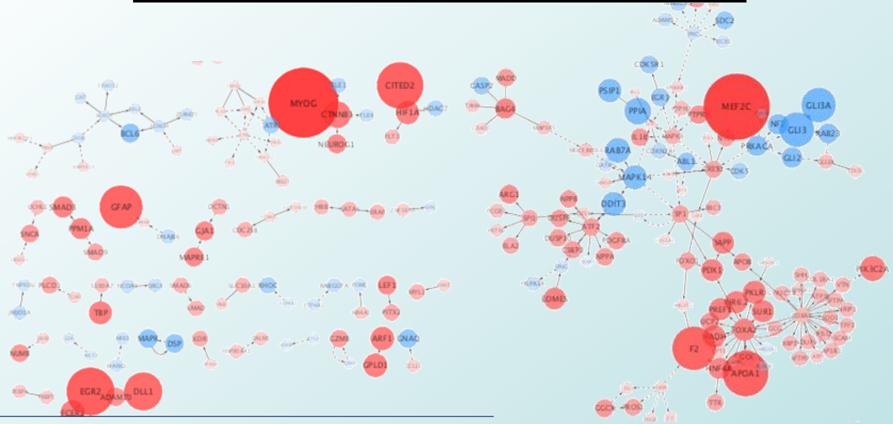
Super Pathway Overview of Sensitivity and Resistance of breast cell lines to gemcitabine



correlated with sensitivity

correlated with resistance

Gemcitabine subnet markers

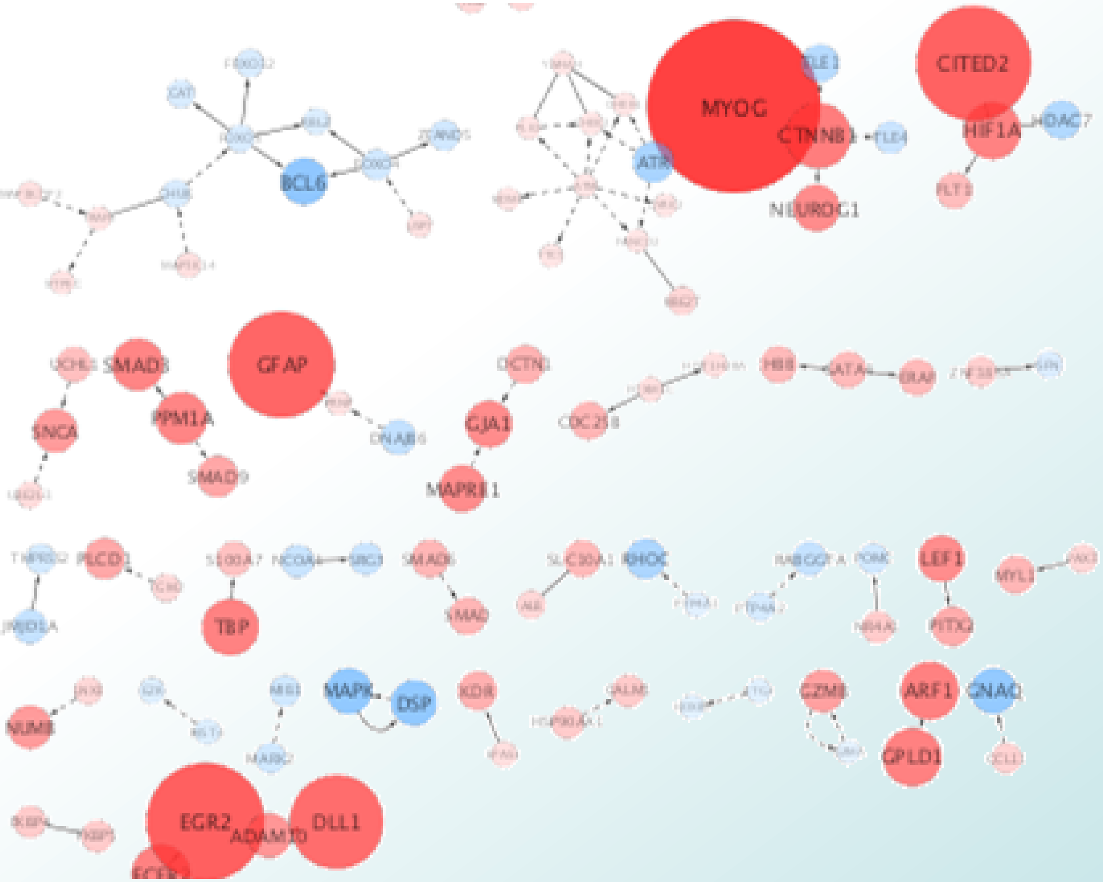


Josh Stuart group, Joe Gray group , 2011

Pathway Markers of Drug Response

- Infer activities in a global “super pathway”
- Associate activities with drug sensitivity to find activity markers; similar for resistance
- Search for focused subnetworks with interconnected markers.

Gemcitabine subnet markers

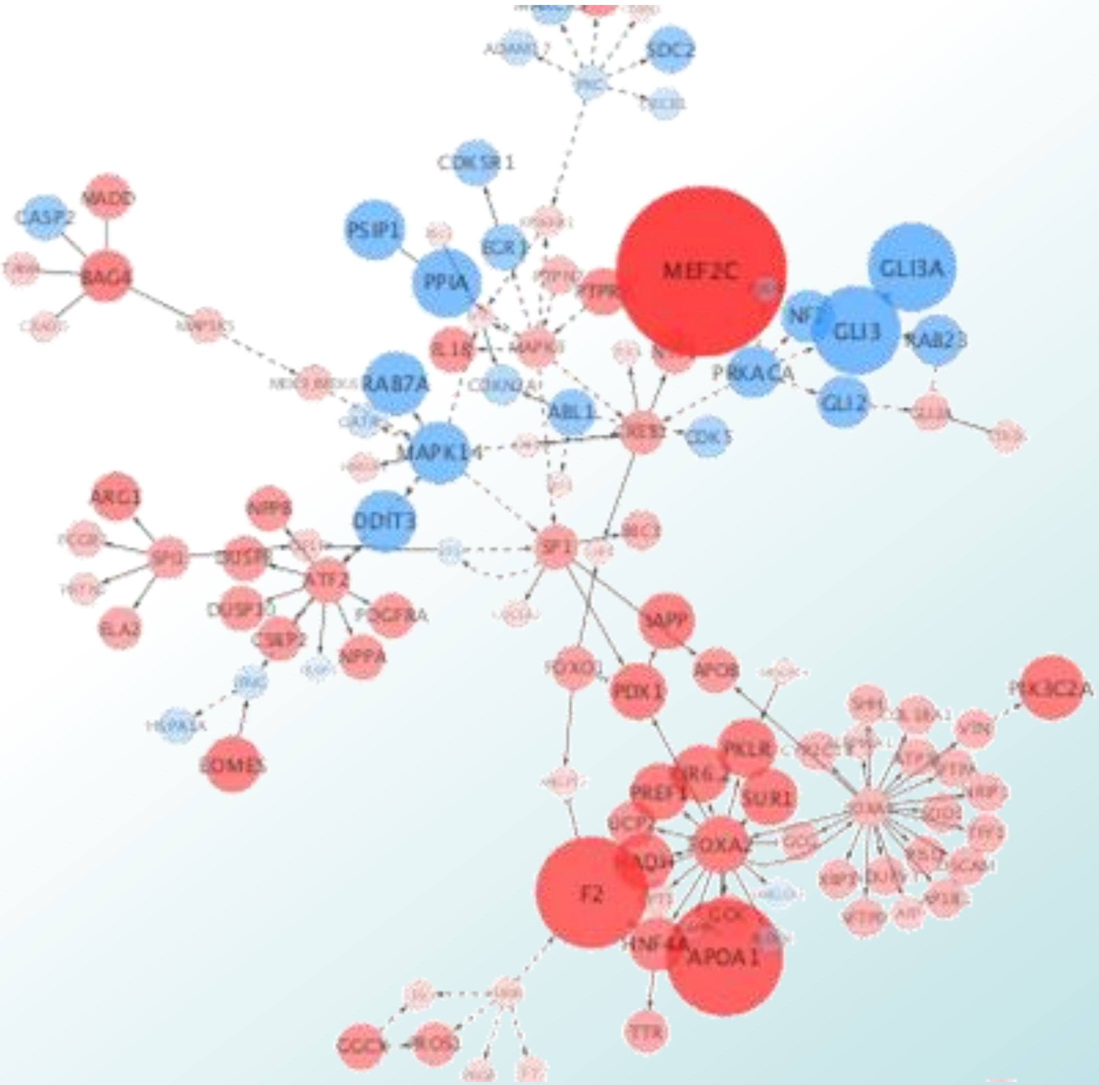


Josh Stuart group, Joe Gray group , 2011

Pathway Markers of Drug Response

- Infer activities in a global “super pathway”
- Associate activities with drug sensitivity to find activity markers; similar for resistance
- Search for focused subnetworks with interconnected markers.

Gemcitabine subnet markers



Josh Stuart group, Joe Gray group , 2011