National Cancer Institute National Human Genome Research Institute

Large-scale Cancer Genomics Data Analysis

David Haussler

Center for Biomolecular Science and Engineering,

UC Santa Cruz

National Cancer Institute National Human Genome Research Institute National Institutes of Health

Cancer Genomics Hub

- Being built to store BAM & VCF for TCGA, TARGET and CGAP/CGCI projects
- Designed for 25,000 cases with average of 200 gigabytes per case
- 5 petabytes (5 x 10^{15}) total, scalable to 20 petabytes
- General Parallel File System, Dual RAID 6 subsystems, Redundant I/O paths, 16 application processors, 12 storage controllers
- co-location opportunities

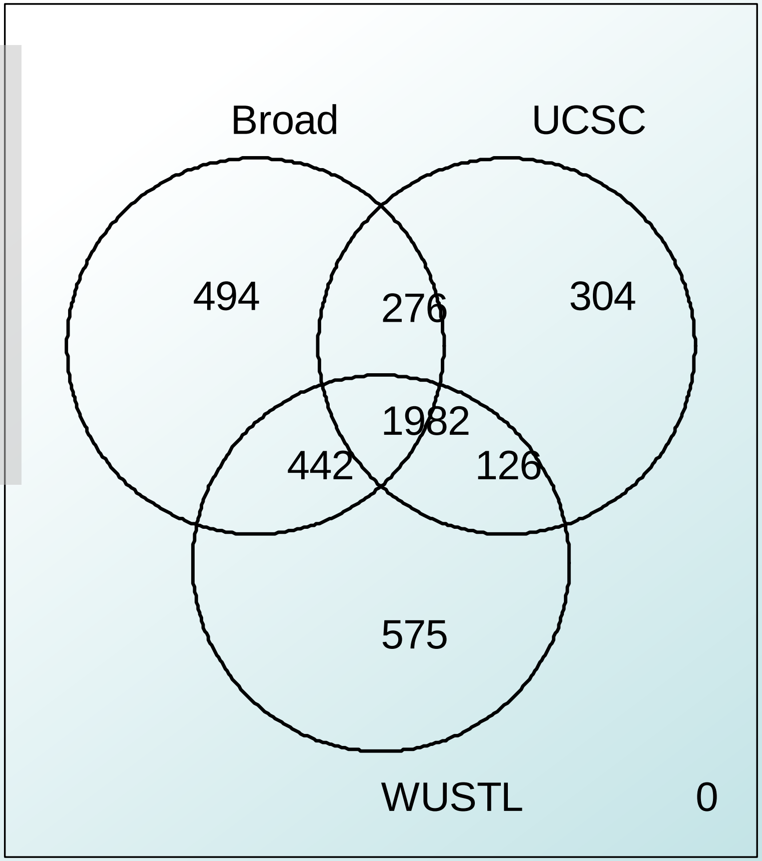
CGHub Goals

- Enable direct comparison and combined analysis of many large-scale cancer genomics datasets
- aggregate enough data to provide the statistical power to attack the full complexity of cancer mutations
- Set standards for data storage and exchange; encourage data sharing
- Maintain compatibility with EGA, dbGaP, ICGC, 1000 Genomes Project, ENCODE and other large-scale genomics efforts (e.g. VCF format, data access coordination)

Given the same BAM files, different mutation calling pipelines do not completely agree

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%

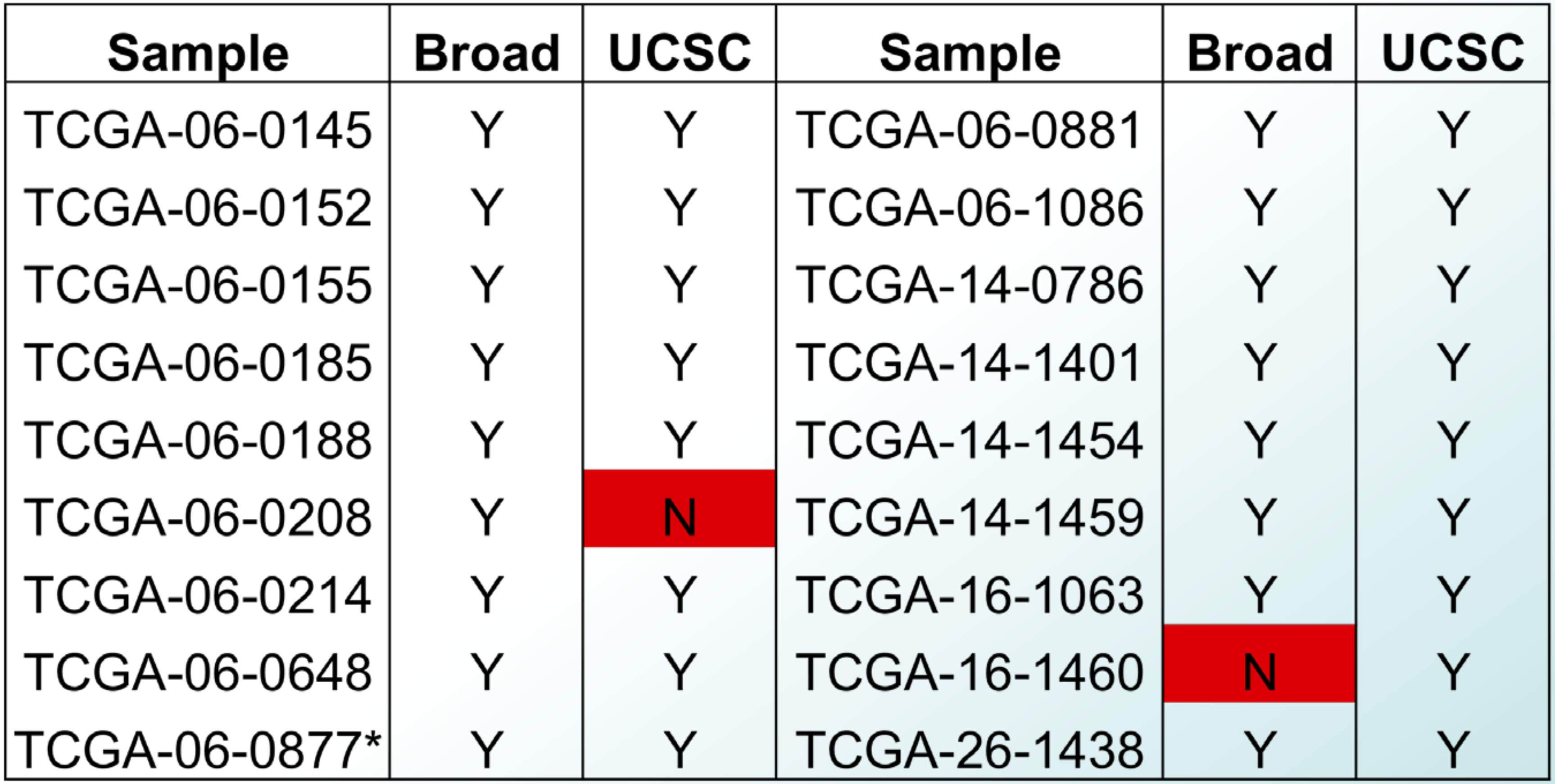


Still work to do to harden mutation-calling software

We are just beginning to look at accuracy and consistency in the detection of structural variation

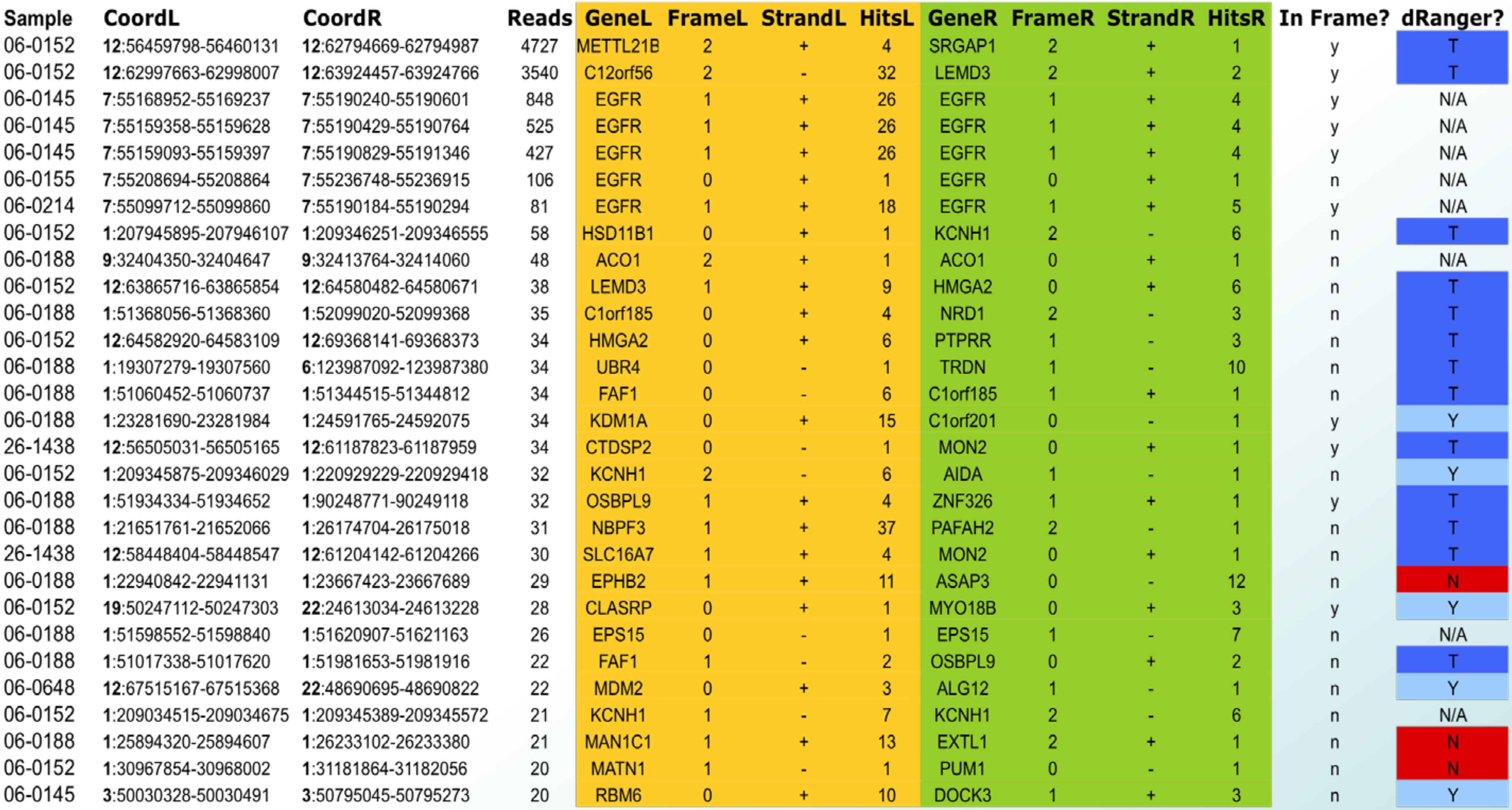
Case study: UCSC and Broad analysis of whole genome GBM data

Samples Analyzed



Sample	Broad	UCSC	Sample	Broad	UCSC
TCGA-06-0145	Y	Y	TCGA-06-0881	Y	Y
TCGA-06-0152	Y	Y	TCGA-06-1086	Y	Y
TCGA-06-0155	Y	Y	TCGA-14-0786	Y	Y
TCGA-06-0185	Y	Y	TCGA-14-1401	Y	Y
TCGA-06-0188	Y	Y	TCGA-14-1454	Y	Y
TCGA-06-0208	Y	N	TCGA-14-1459	Y	Y
TCGA-06-0214	Y	Y	TCGA-16-1063	Y	Y
TCGA-06-0648	Y	Y	TCGA-16-1460	N	Y
TCGA-06-0877*	Y	Y	TCGA-26-1438	Y	Y

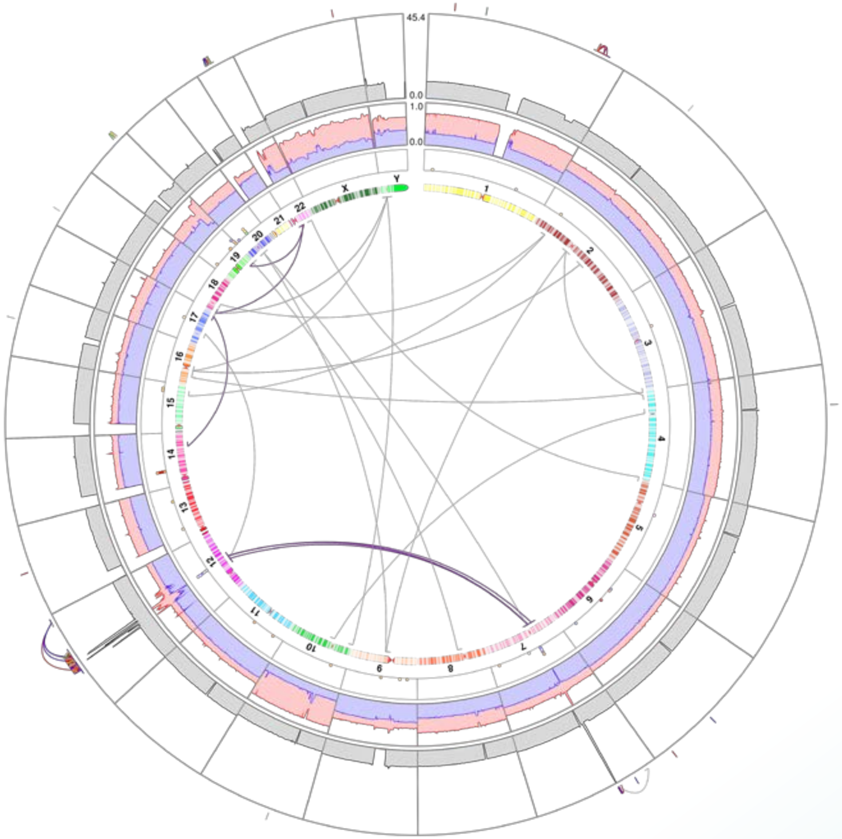
Gene fusions: BamBam 167, dRanger 188



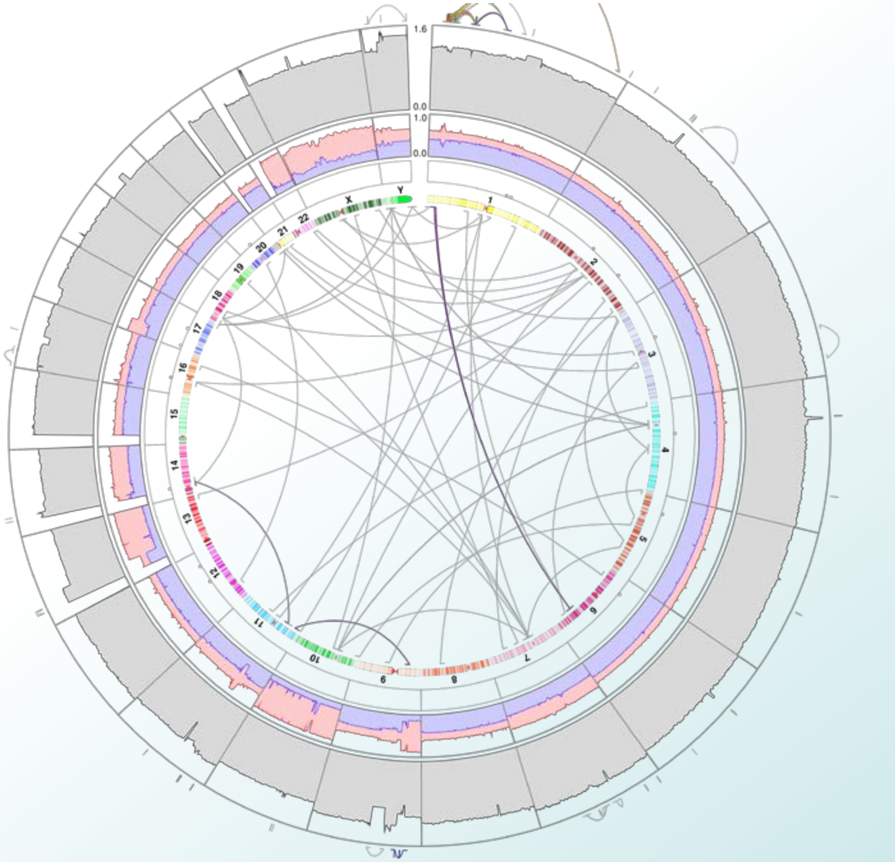
Sample	CoordL	CoordR	Reads	GeneL	FrameL	StrandL	HitsL	GeneR	FrameR	StrandR	HitsR	In Frame?	dRanger?
06-0152	12:56459798-56460131	12:62794669-62794987	4727	METTL21B	2	+	4	SRGAP1	2	+	1	y	T
06-0152	12:62997663-62998007	12:63924457-63924766	3540	C12orf56	2	-	32	LEMD3	2	+	2	y	T
06-0145	7:55168952-55169237	7:55190240-55190601	848	EGFR	1	+	26	EGFR	1	+	4	y	N/A
06-0145	7:55159358-55159628	7:55190429-55190764	525	EGFR	1	+	26	EGFR	1	+	4	y	N/A
06-0145	7:55159093-55159397	7:55190829-55191346	427	EGFR	1	+	26	EGFR	1	+	4	y	N/A
06-0155	7:55208694-55208864	7:55236748-55236915	106	EGFR	0	+	1	EGFR	0	+	1	n	N/A
06-0214	7:55099712-55099860	7:55190184-55190294	81	EGFR	1	+	18	EGFR	1	+	5	y	N/A
06-0152	1:207945895-207946107	1:209346251-209346555	58	HSD11B1	0	+	1	KCNH1	2	-	6	n	T
06-0188	9:32404350-32404647	9:32413764-32414060	48	ACO1	2	+	1	ACO1	0	+	1	n	N/A
06-0152	12:63865716-63865854	12:64580482-64580671	38	LEMD3	1	+	9	HMGA2	0	+	6	n	T
06-0188	1:51368056-51368360	1:52099020-52099368	35	C1orf185	0	+	4	NRD1	2	-	3	n	T
06-0152	12:64582920-64583109	12:69368141-69368373	34	HMGA2	0	+	6	PTPRR	1	-	3	n	T
06-0188	1:19307279-19307560	6:123987092-123987380	34	UBR4	0	-	1	TRDN	1	-	10	n	T
06-0188	1:51060452-51060737	1:51344515-51344812	34	FAF1	0	-	6	C1orf185	1	+	1	n	T
06-0188	1:23281690-23281984	1:24591765-24592075	34	KDM1A	0	+	15	C1orf201	0	-	1	y	Y
26-1438	12:56505031-56505165	12:61187823-61187959	34	CTDSP2	0	-	1	MON2	0	+	1	y	T
06-0152	1:209345875-209346029	1:220929229-220929418	32	KCNH1	2	-	6	AIDA	1	-	1	n	Y
06-0188	1:51934334-51934652	1:90248771-90249118	32	OSBPL9	1	+	4	ZNF326	1	+	1	y	T
06-0188	1:21651761-21652066	1:26174704-26175018	31	NBPF3	1	+	37	PAFAH2	2	-	1	n	T
26-1438	12:58448404-58448547	12:61204142-61204266	30	SLC16A7	1	+	4	MON2	0	+	1	n	T
06-0188	1:22940842-22941131	1:23667423-23667689	29	EPHB2	1	+	11	ASAP3	0	-	12	n	N
06-0152	19:50247112-50247303	22:24613034-24613228	28	CLASRP	0	+	1	MYO18B	0	+	3	y	Y
06-0188	1:51598552-51598840	1:51620907-51621163	26	EPS15	0	-	1	EPS15	1	-	7	n	N/A
06-0188	1:51017338-51017620	1:51981653-51981916	22	FAF1	1	-	2	OSBPL9	0	+	2	n	T
06-0648	12:67515167-67515368	22:48690695-48690822	22	MDM2	0	+	3	ALG12	1	-	1	n	Y
06-0152	1:209034515-209034675	1:209345389-209345572	21	KCNH1	1	-	7	KCNH1	2	-	6	n	N/A
06-0188	1:25894320-25894607	1:26233102-26233380	21	MAN1C1	1	+	13	EXTL1	2	+	1	n	N
06-0152	1:30967854-30968002	1:31181864-31182056	20	MATN1	1	-	1	PUM1	0	-	1	n	N
06-0145	3:50030328-50030491	3:50795045-50795273	20	RBM6	0	+	10	DOCK3	1	+	3	n	Y

136 potentially overlapping events

Whole Genome View

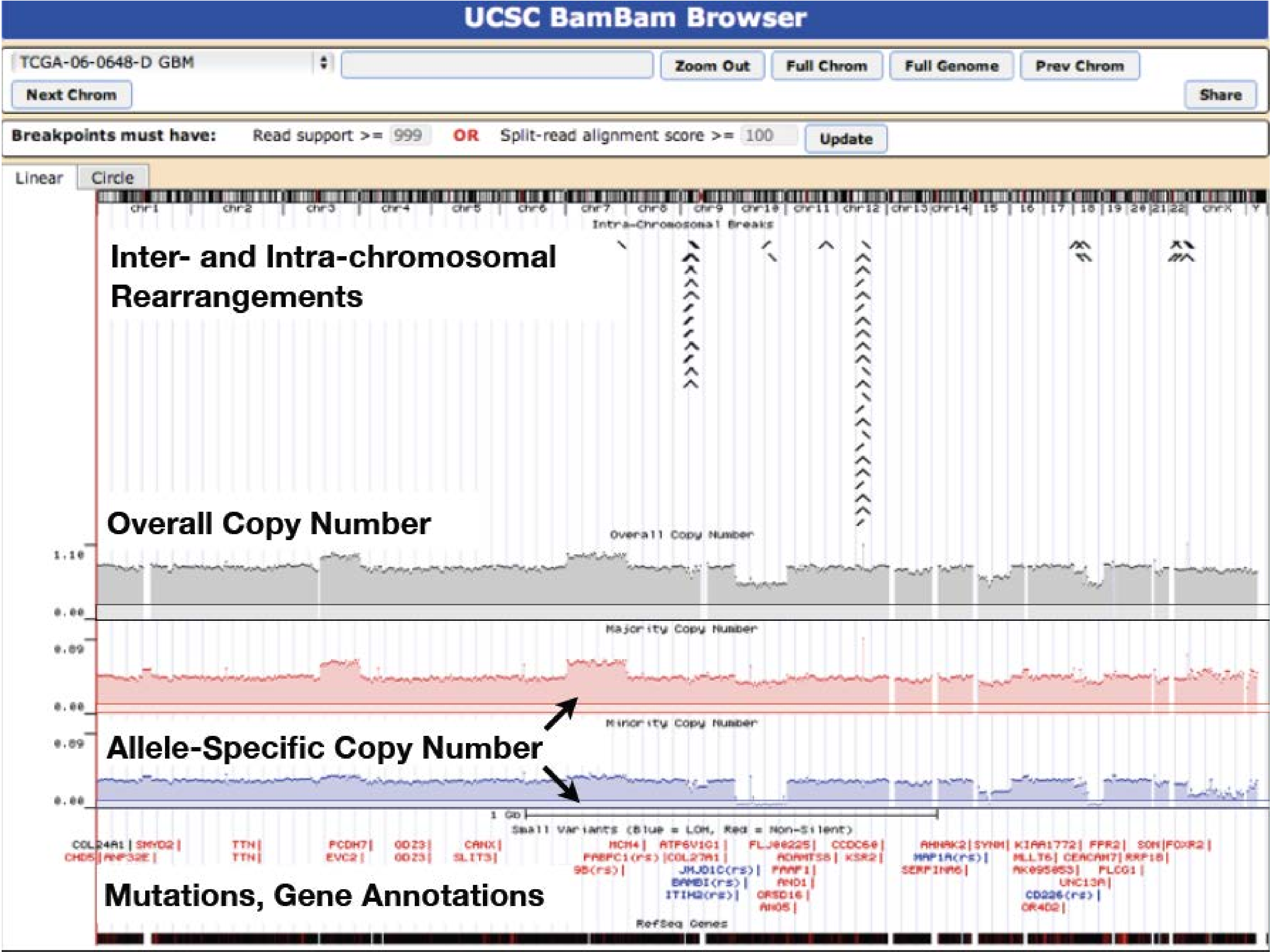


06-0152



06-0188

- Circle plot shows amplifications, deletions, inter/intra chromosomal rearrangement
- These 2 samples have 23/25 top dRanger, 21/29 top bambam events



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

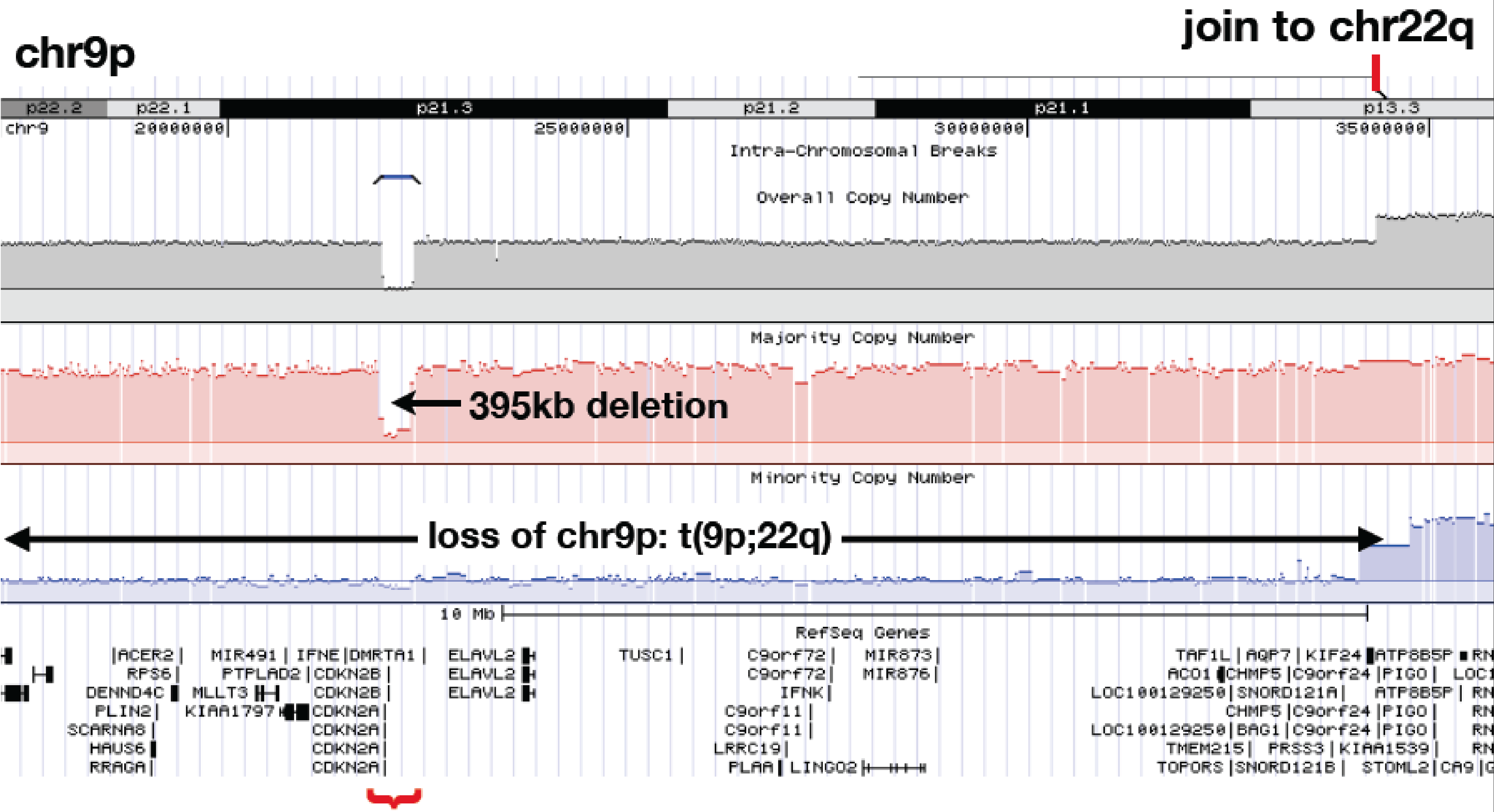
Inter- and Intra-chromosomal Rearrangements

Overall Copy Number

Allele-Specific Copy Number

Mutations, Gene Annotations

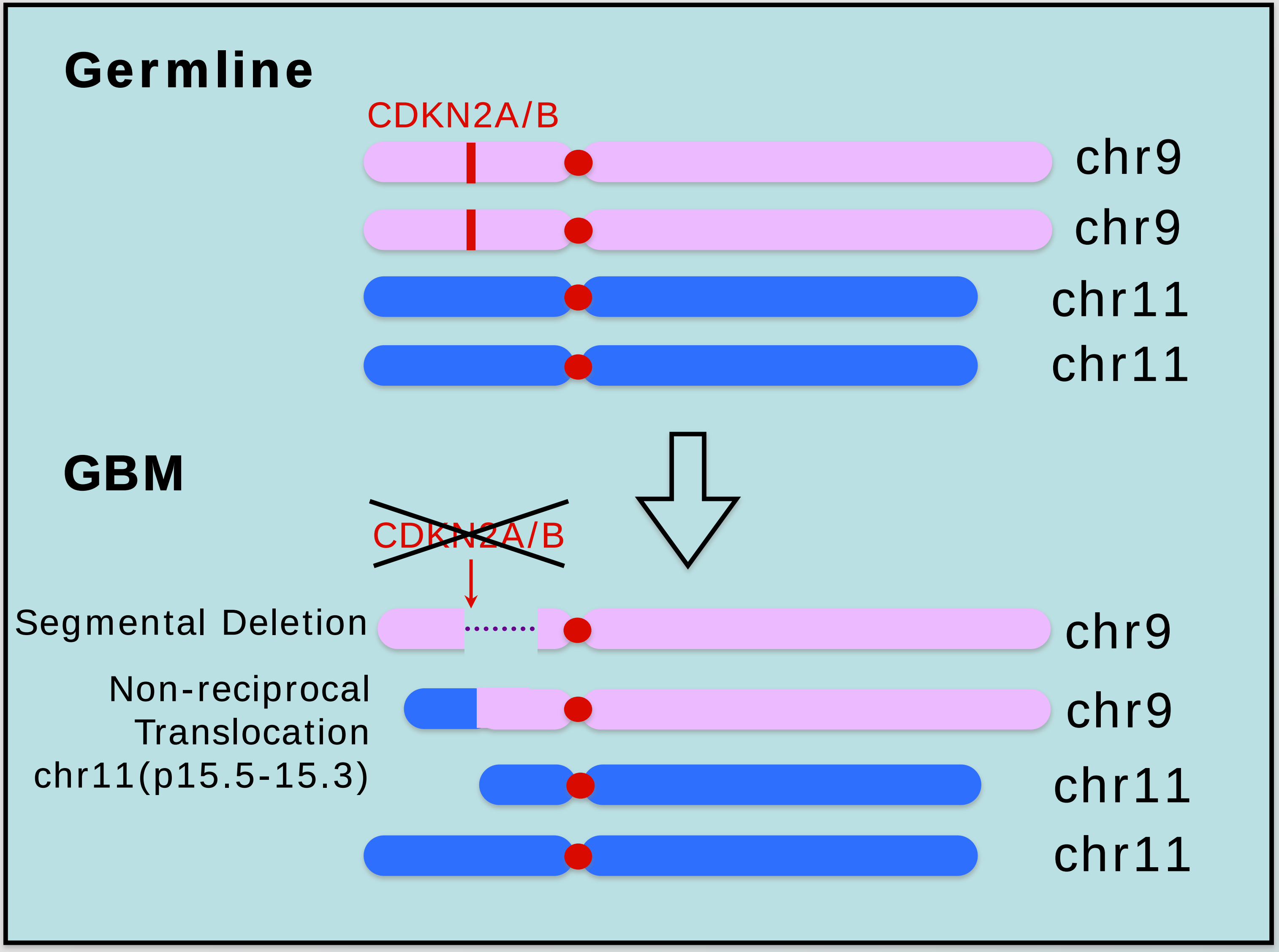
Glioblastoma: TCGA-06-0145



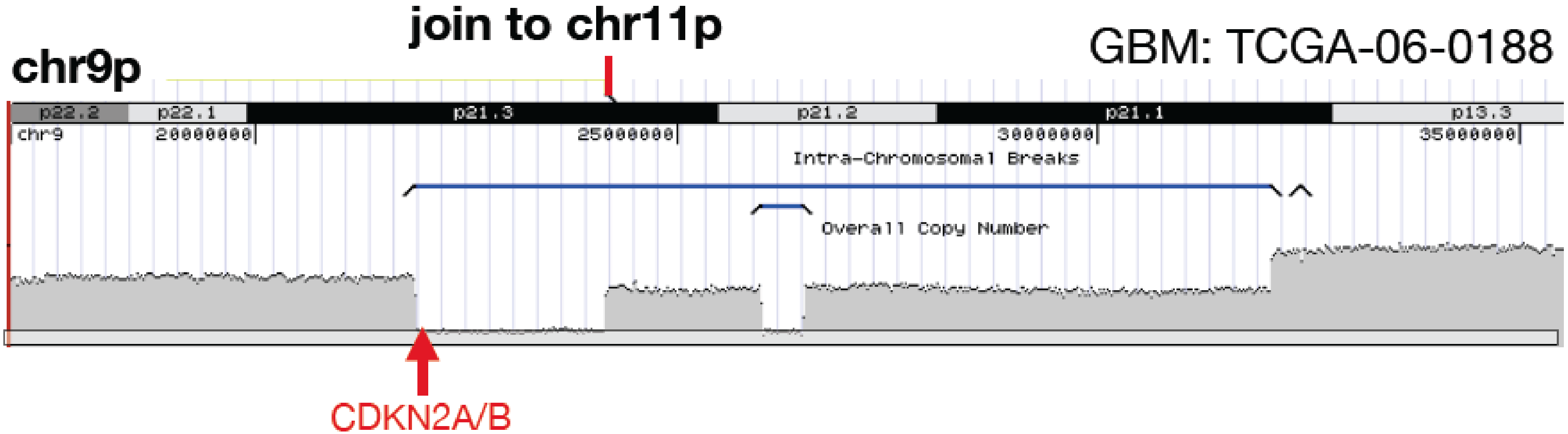
Homozygous loss of CDKN2A/B

via inter- and intra-chromosomal rearrangements.

Independent events lead to somatic homozygous loss of tumor suppressors CDKN2A/B



Similar double-loss motif in other GBMs

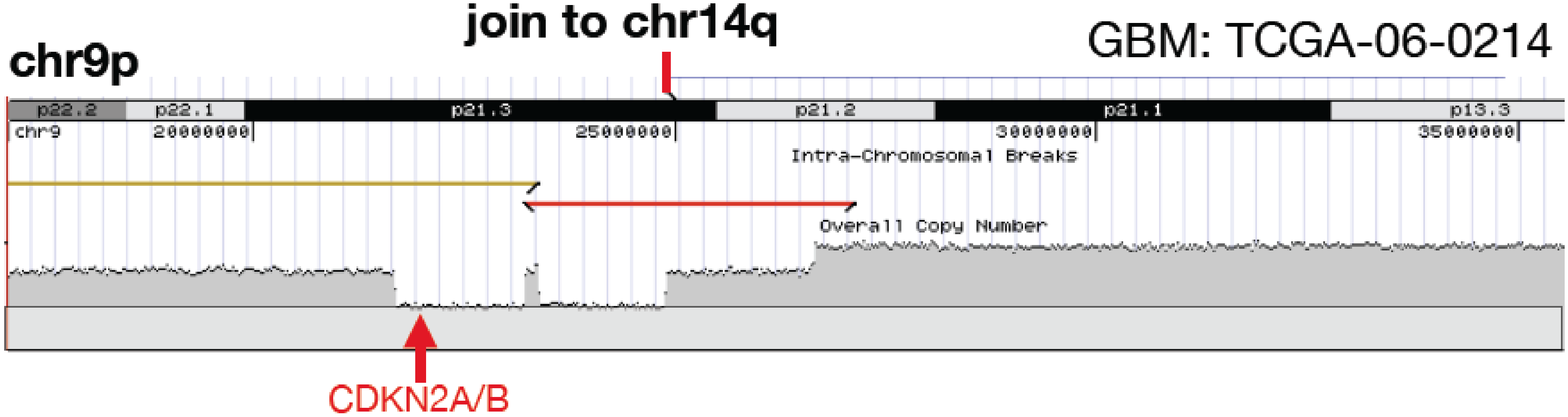


join to chr11p

chr9p

GBM: TCGA-06-0188

CDKN2A/B



join to chr14q

chr9p

GBM: TCGA-06-0214

CDKN2A/B

In 11/16 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*

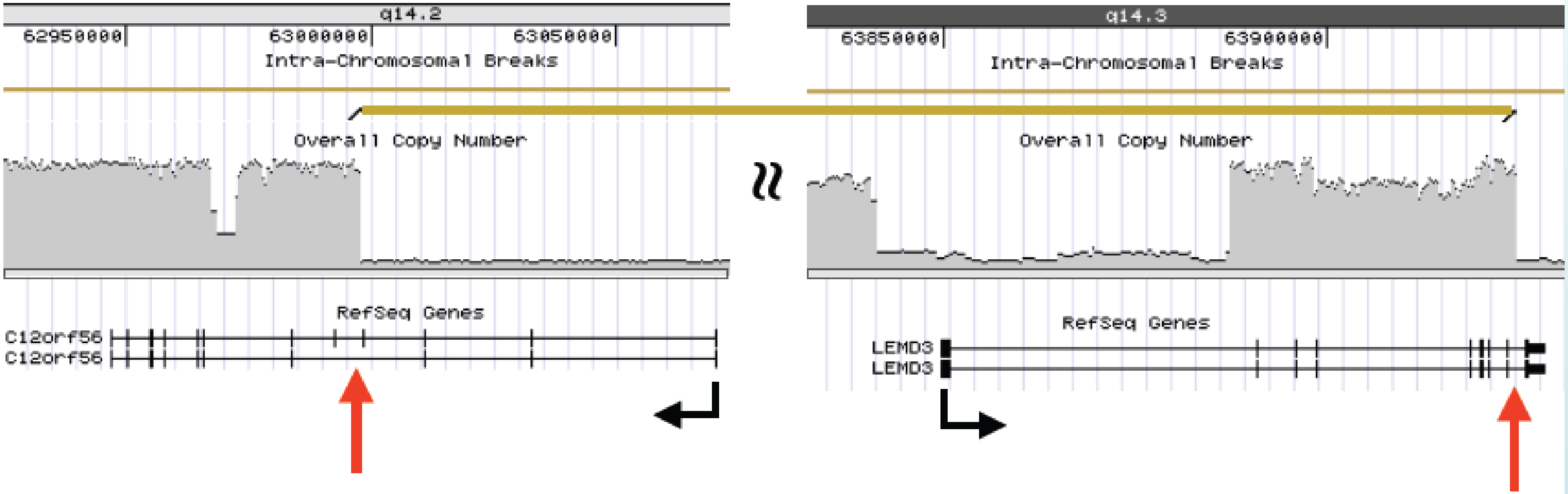
Zack Sanborn

Features of *CDKN2A/B* normal samples

Sample	Exp subtype	G-CIMP	EGFR	CDK4	MDM2	Other
TCGA-06-0152	mes		amp	amp	amp	
TCGA-06-0881	mes		amp			
TCGA-14-1454	pro					PTEN deln+FS
TCGA-16-1460	pro	+		rearr*	rearr*	IDH1 mut
TCGA-26-1438	mes			amp	amp	

LEMD3 - c12orf56 Fusion

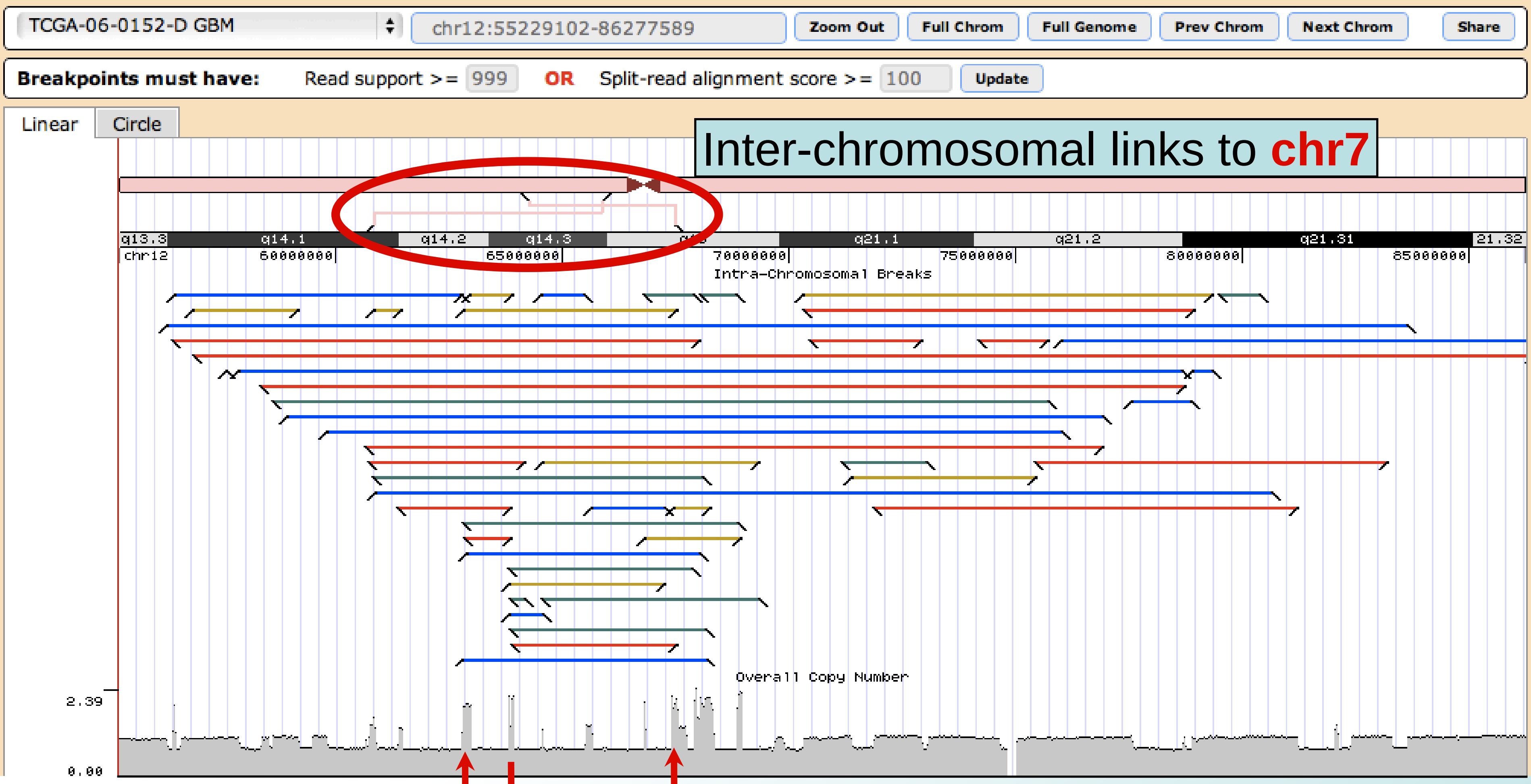
GBM: TCGA-06-0152



c12orf56 Fusion Point

LEMD3 Fusion Point

Chromothripsis in a gliblastoma



TCGA-06-0152-D GBM

chr12:55229102-86277589

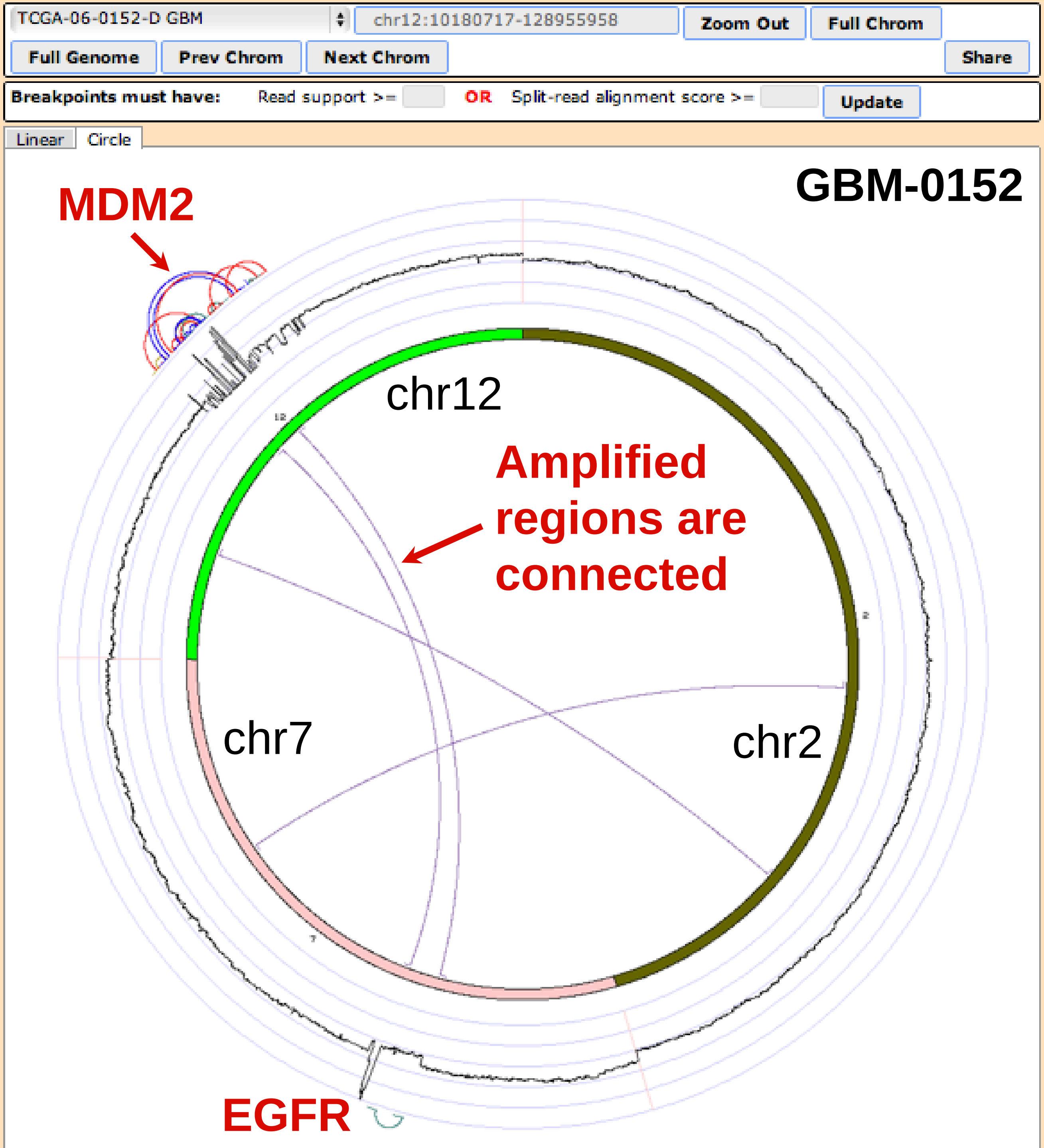
Breakpoints must have: Read support >= 999 **OR** Split-read alignment score >= 100

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

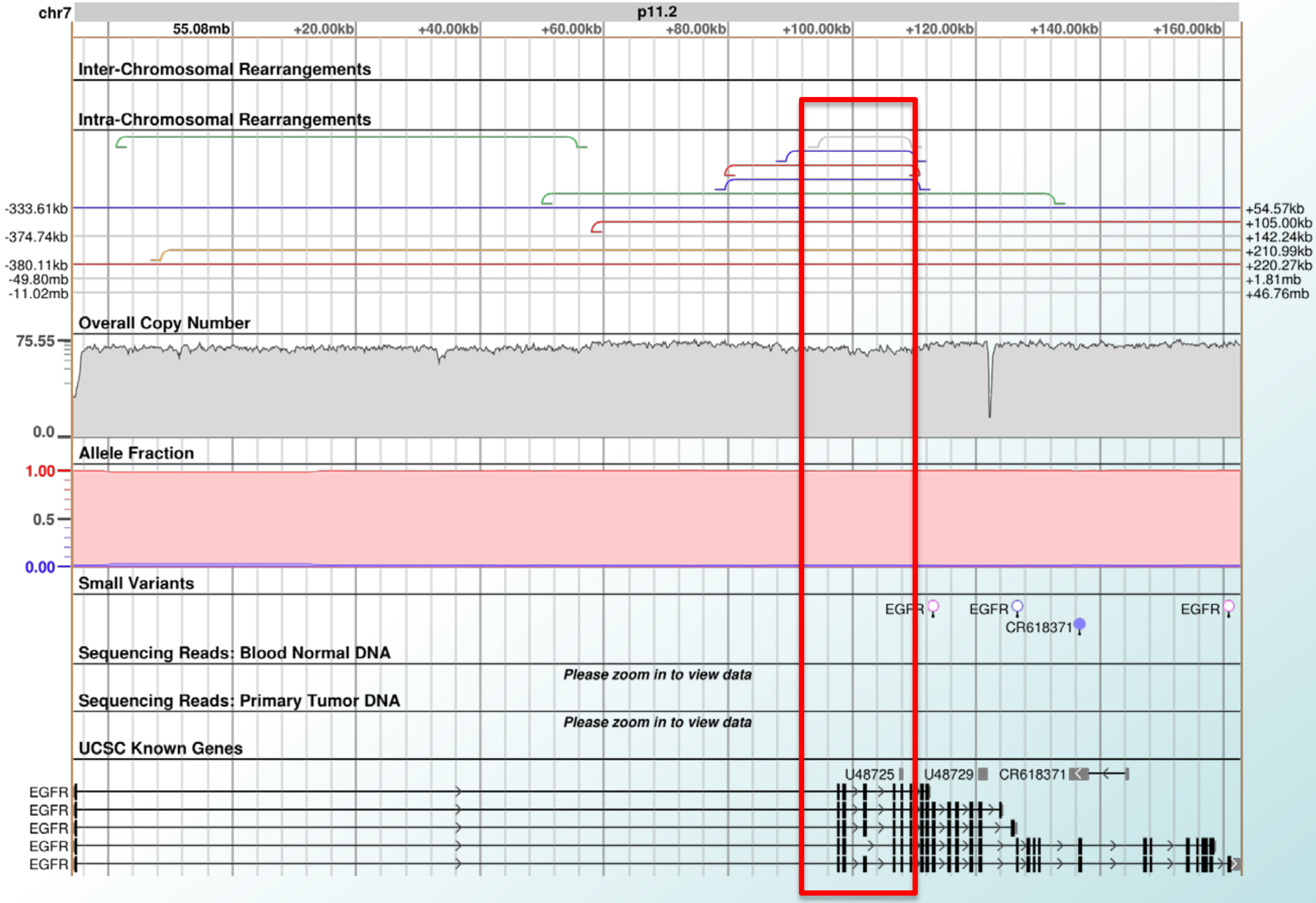
Amplified regions are connected

EGFR

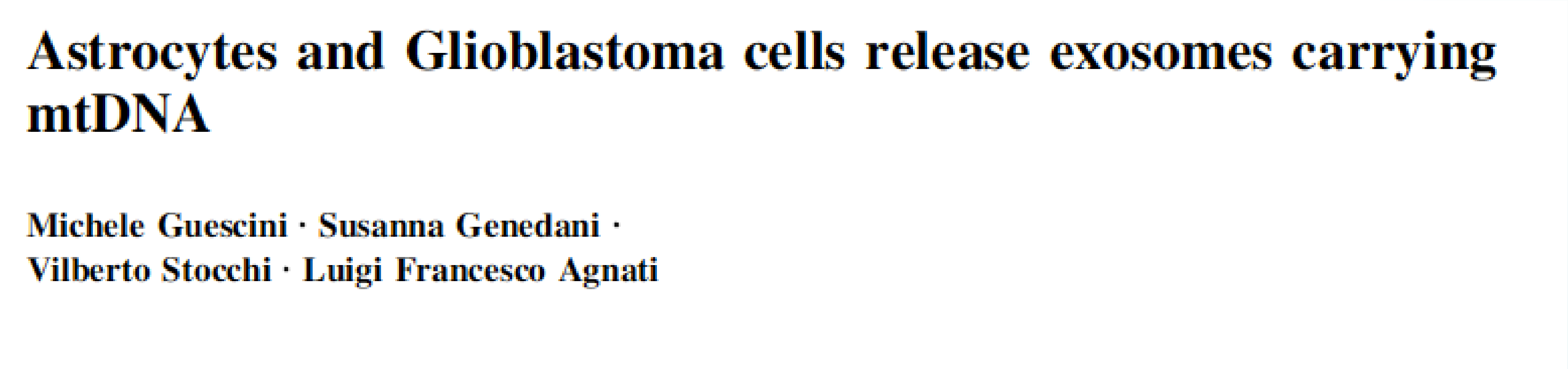
EGFR Amplifcation/Mutation

- 11/17 samples have chr7 amplifications including EGFR
- 4/11 also have EGFRviii mutations
- Exon 2-7 deletion at low copy
  - Probably happened after amplification events
  - Selection for low copy?

Example: EGFRviii mutation

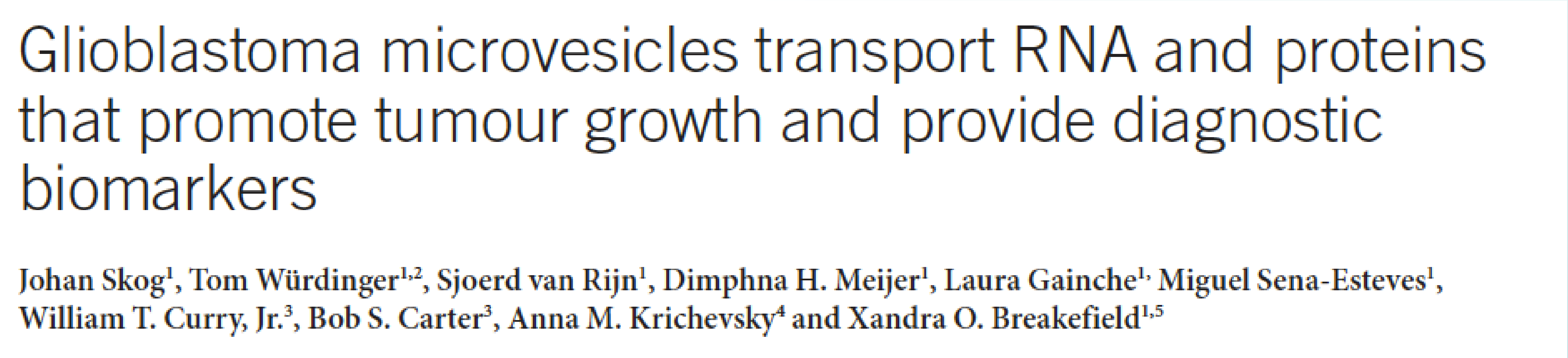


GBMs release exosomes. Could some GBM tumor DNA show up in the blood?



Astrocytes and Glioblastoma cells release exosomes carrying mtDNA

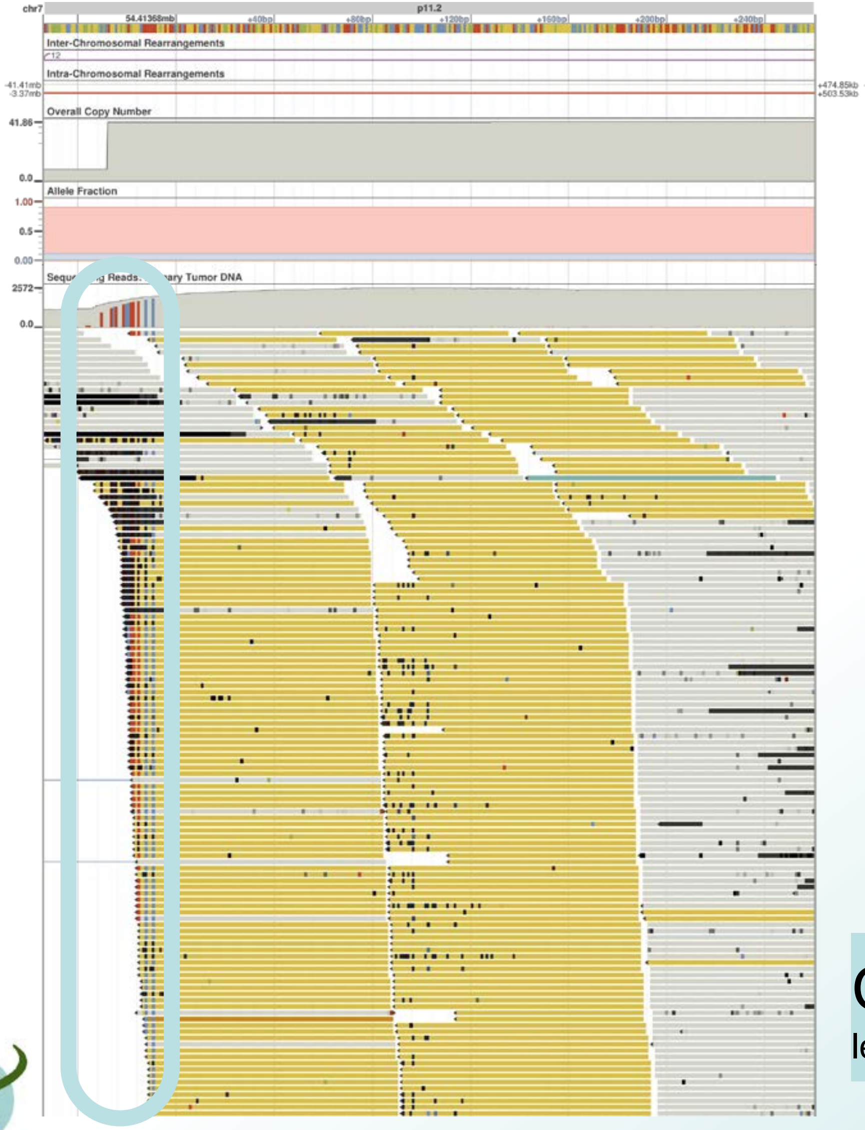
Michele Guescini · Susanna Genedani · Vilberto Stocchi · Luigi Francesco Agnati

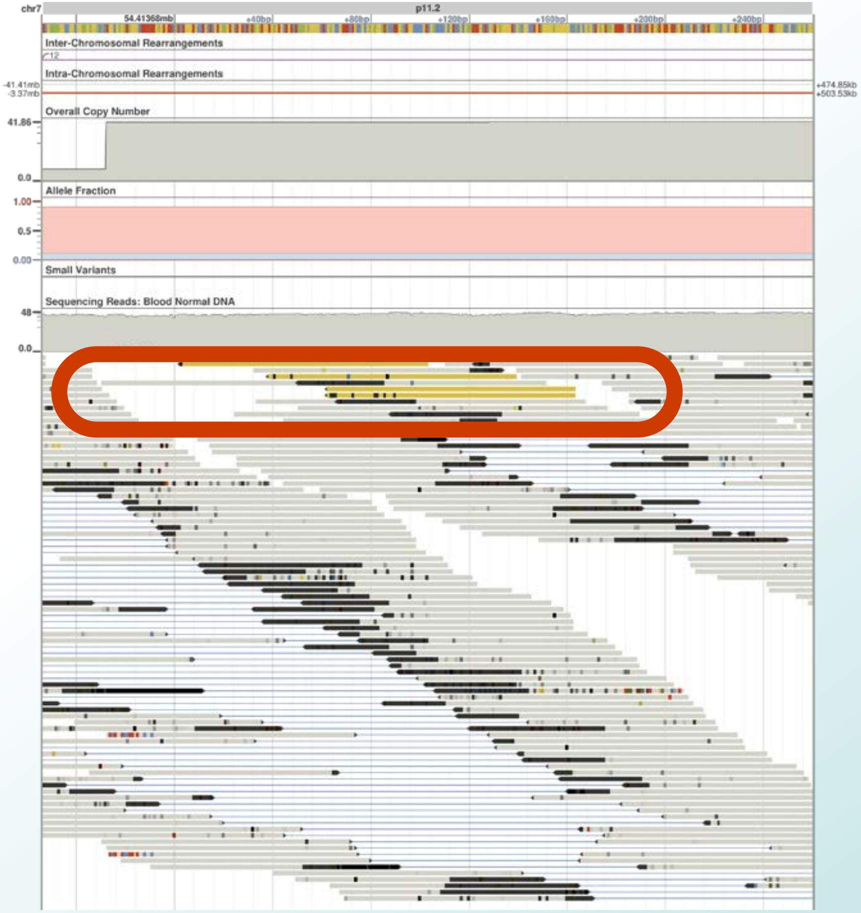


Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers

Johan Skog[1], Tom Würdinger[1,2], Sjoerd van Rijn[1], Dimphna H. Meijer[1], Laura Gainche[1]· Miguel Sena-Esteves[1], William T. Curry, Jr.[3], Bob S. Carter[3], Anna M. Krichevsky[4] and Xandra O. Breakefield[1,5]

Amplified events may provide enough reads to detect this



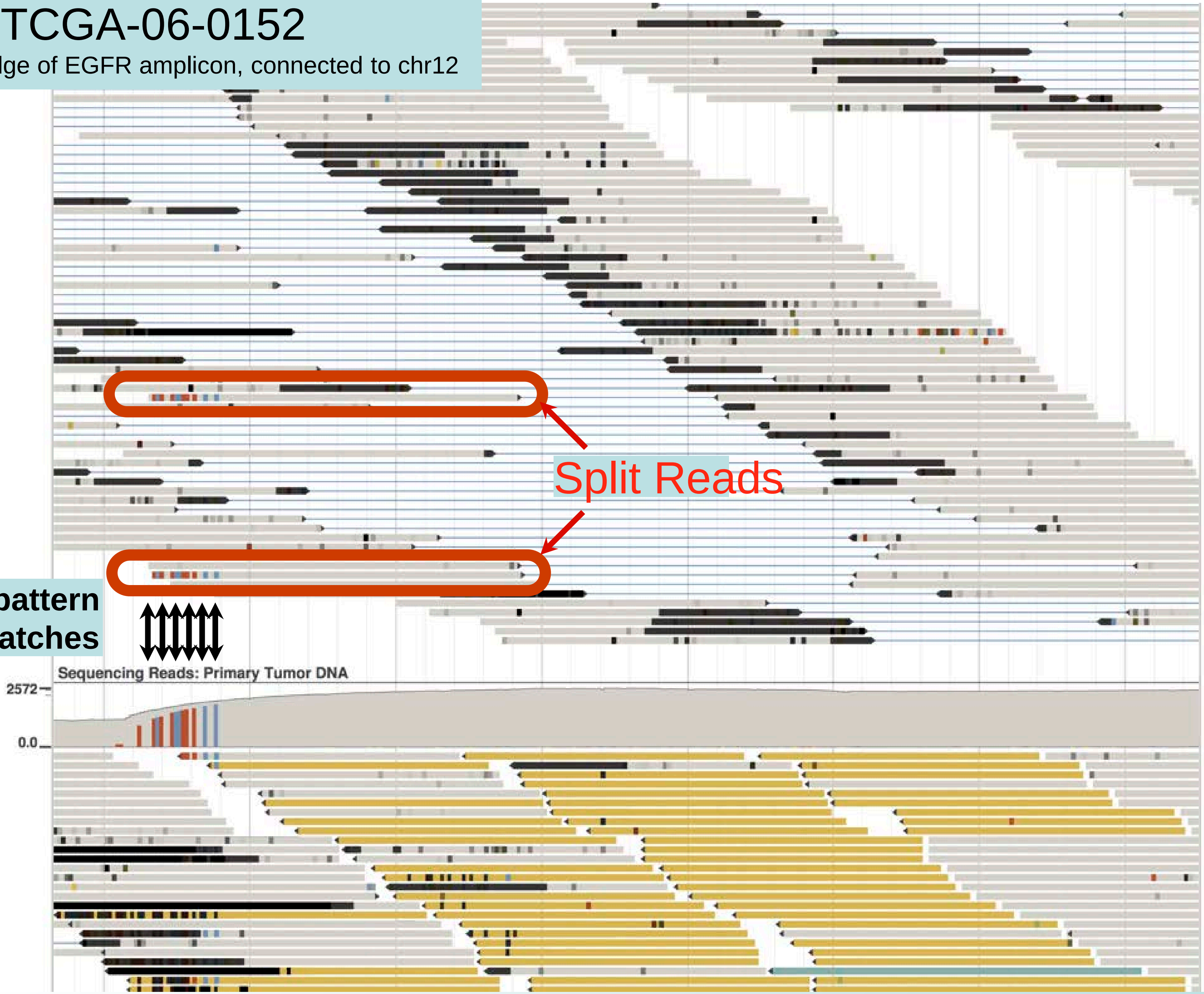


GBM: TCGA-06-0152

left-hand edge of EGFR amplicon, connected to chr12

GBM: TCGA-06-0152

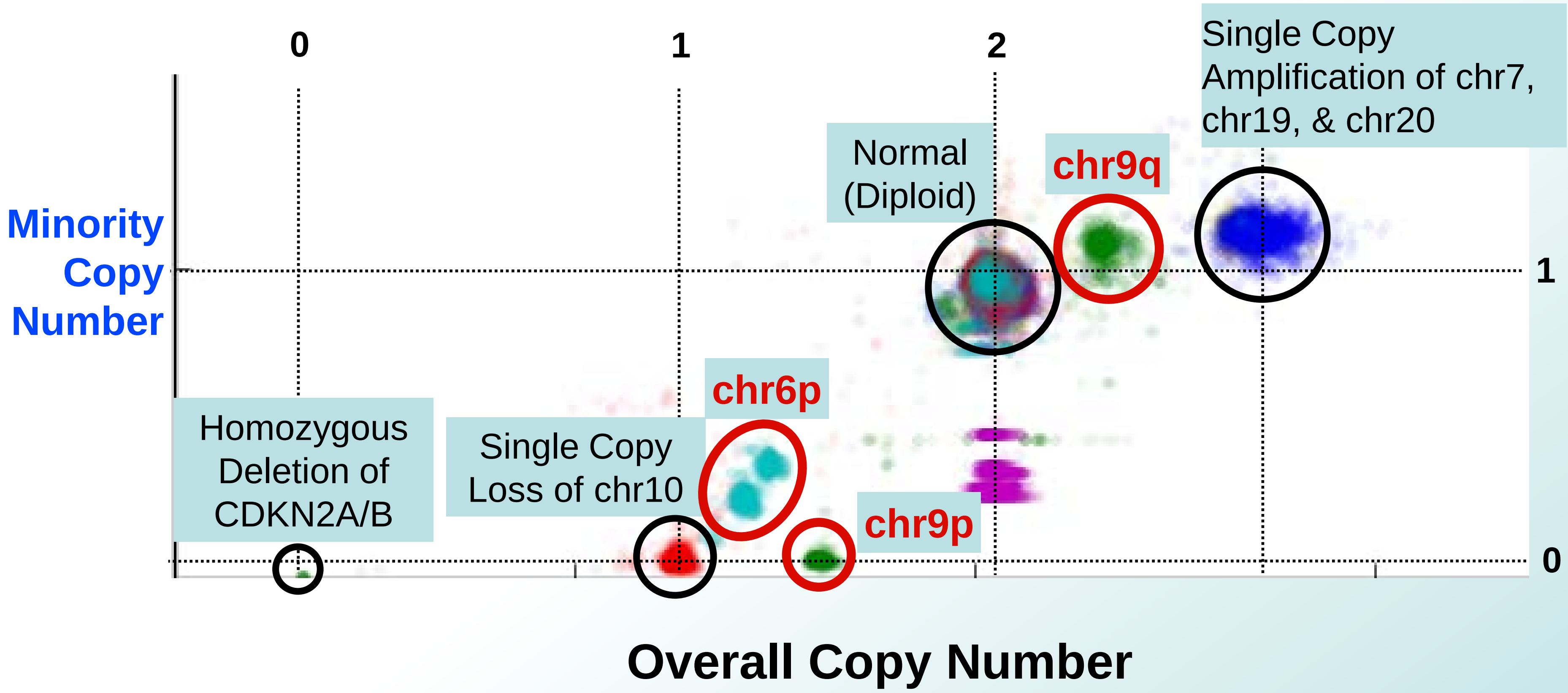
left-hand edge of EGFR amplicon, connected to chr12



Split Reads

Similar pattern of mismatches

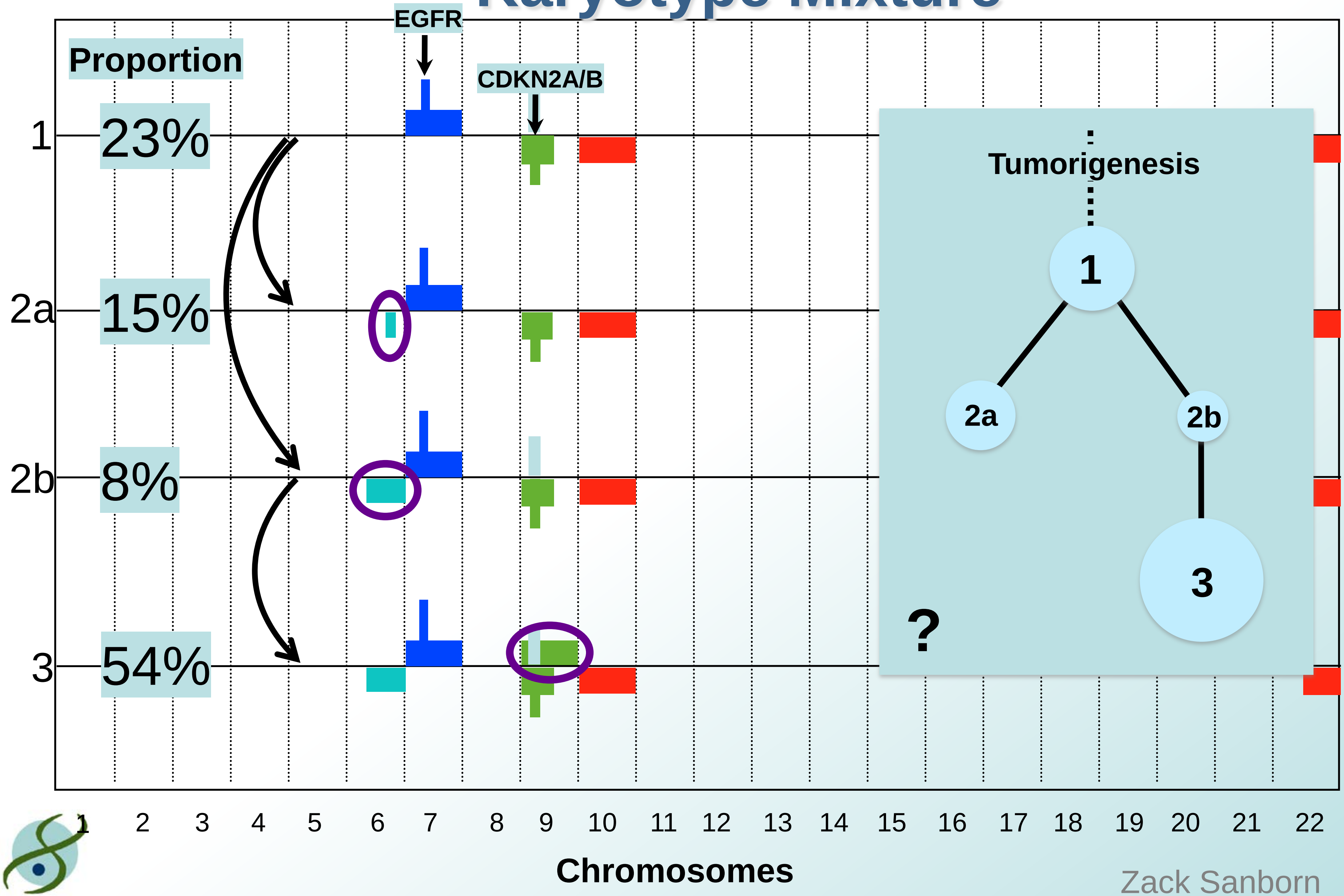
Copy Number States



GBM: TCGA-06-0185

Zack Sanborn

Simulated Progression Model to Infer Karyotype Mixture

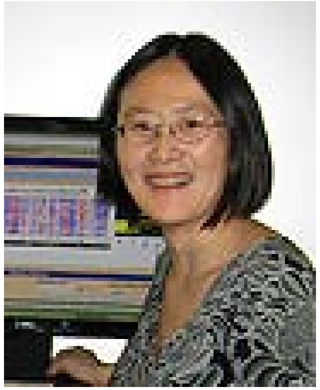


Zack Sanborn

UCSC Cancer Integration Group

Josh Stuart, Co-PI

Jing Zhu



Charlie Vaske

Steve Benz

Zack Sanborn *

James Durbin

Mark Diekhans *

Melissa Cline

Dan Carlin

Kyle Elrott

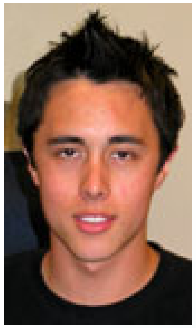
Brian Craft

Sofie Salama *

Chris Wilks

Artem Sokolov

Chris Szeto

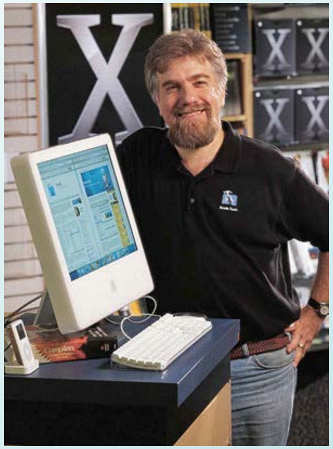


Sam Ng

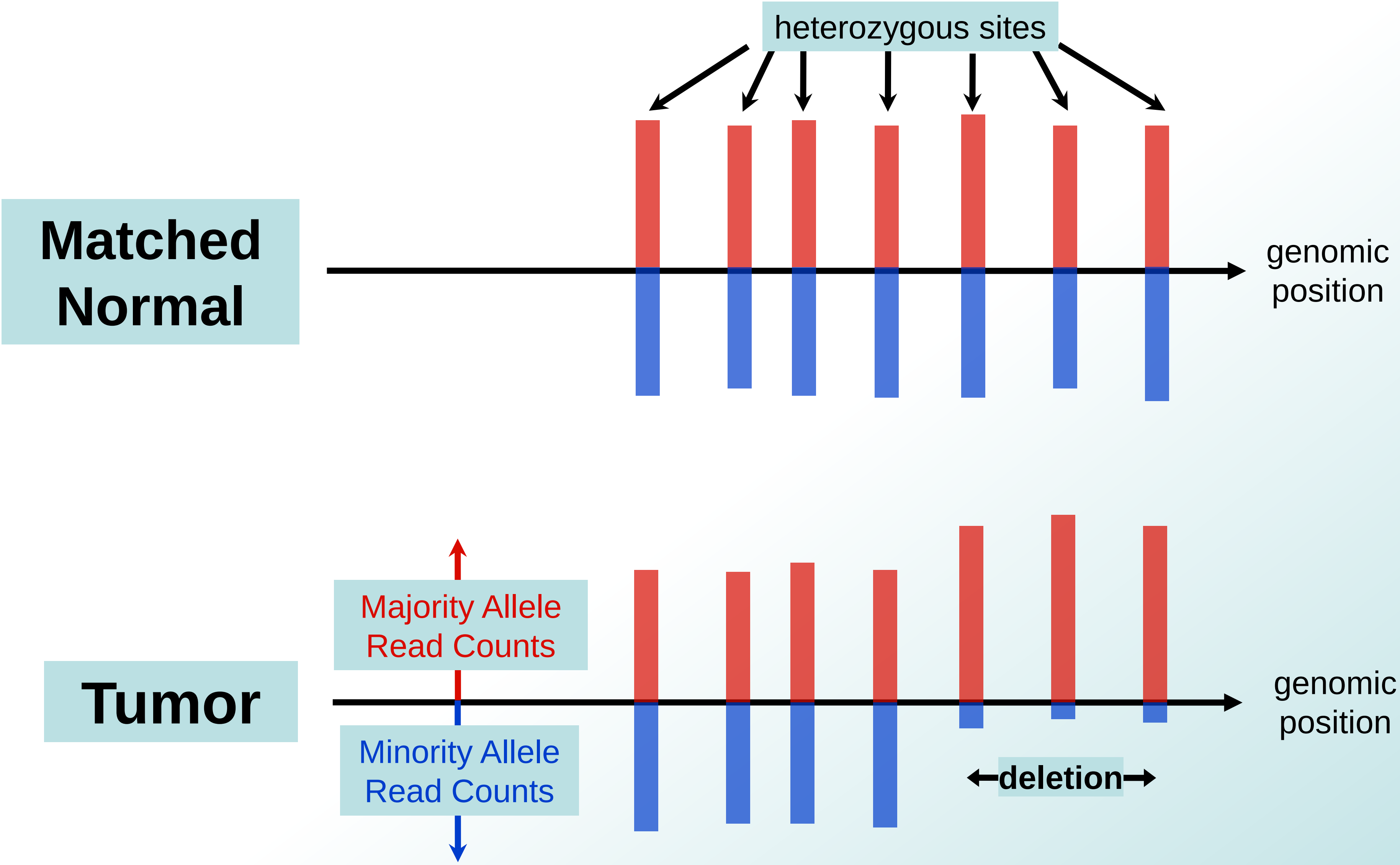
Mia Grifford

Amie Radenbaugh

Ted Golstein



Allele-Specific Copy Number



heterozygous sites

Matched Normal

genomic position

Tumor

Majority Allele Read Counts

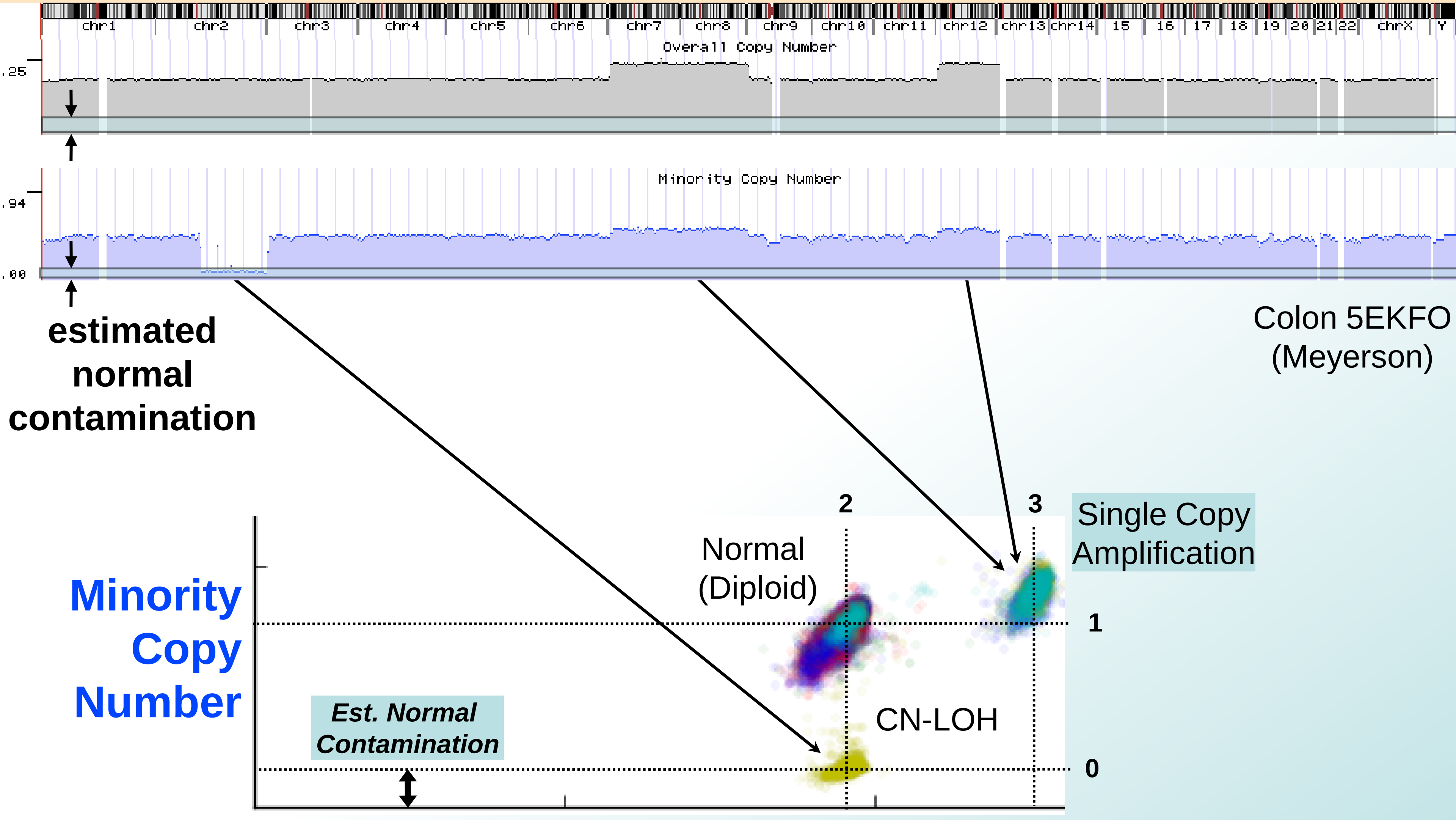
Minority Allele Read Counts

genomic position

←deletion→

Zack Sanborn

Tumors exhibit multiple rounds of duplication, rearrangement and loss



estimated normal contamination

Colon 5EKFO (Meyerson)

Normal (Diploid)

Single Copy Amplification

CN-LOH

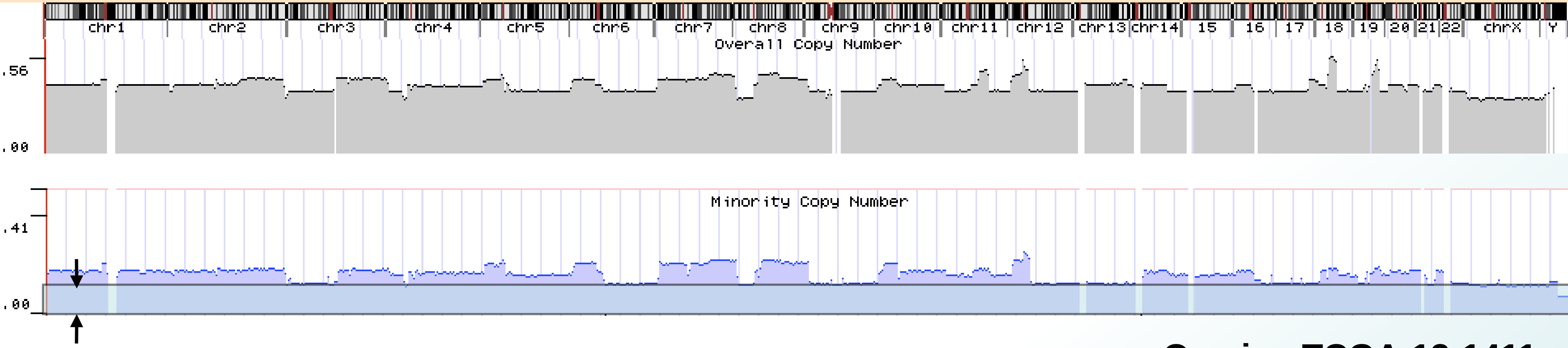
Est. Normal Contamination

Minority Copy Number

Overall Copy Number

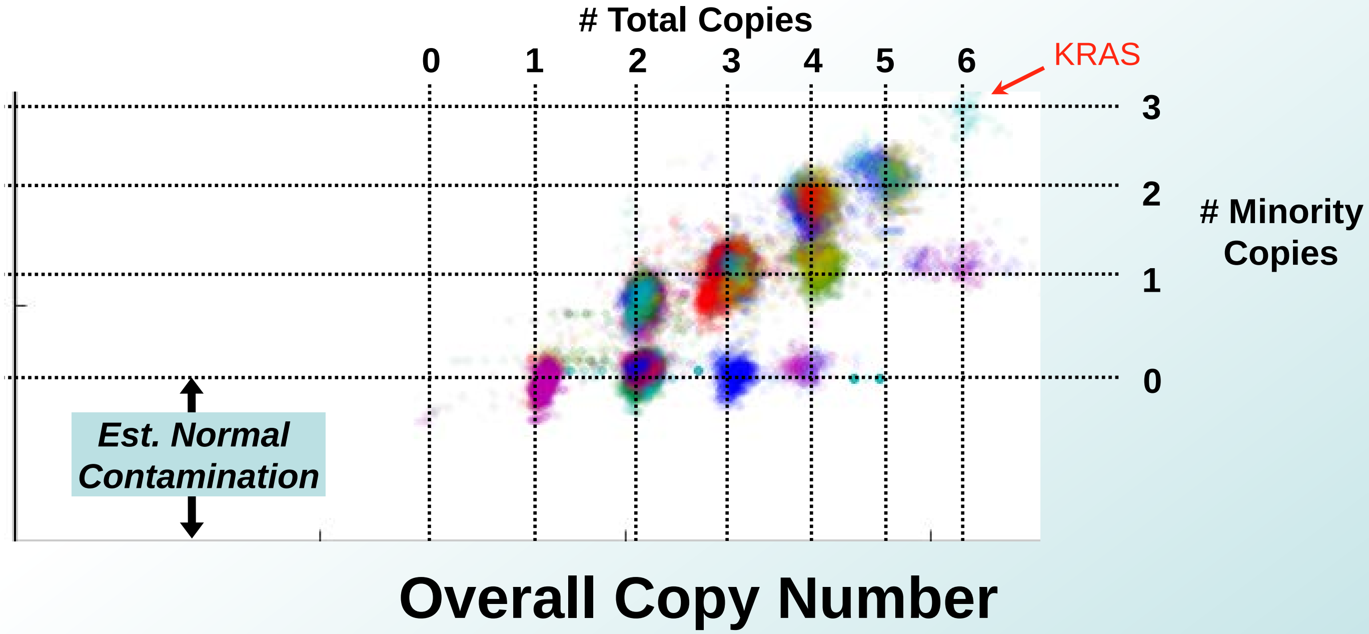
Zack Sanborn

Copy Number Profile Analysis



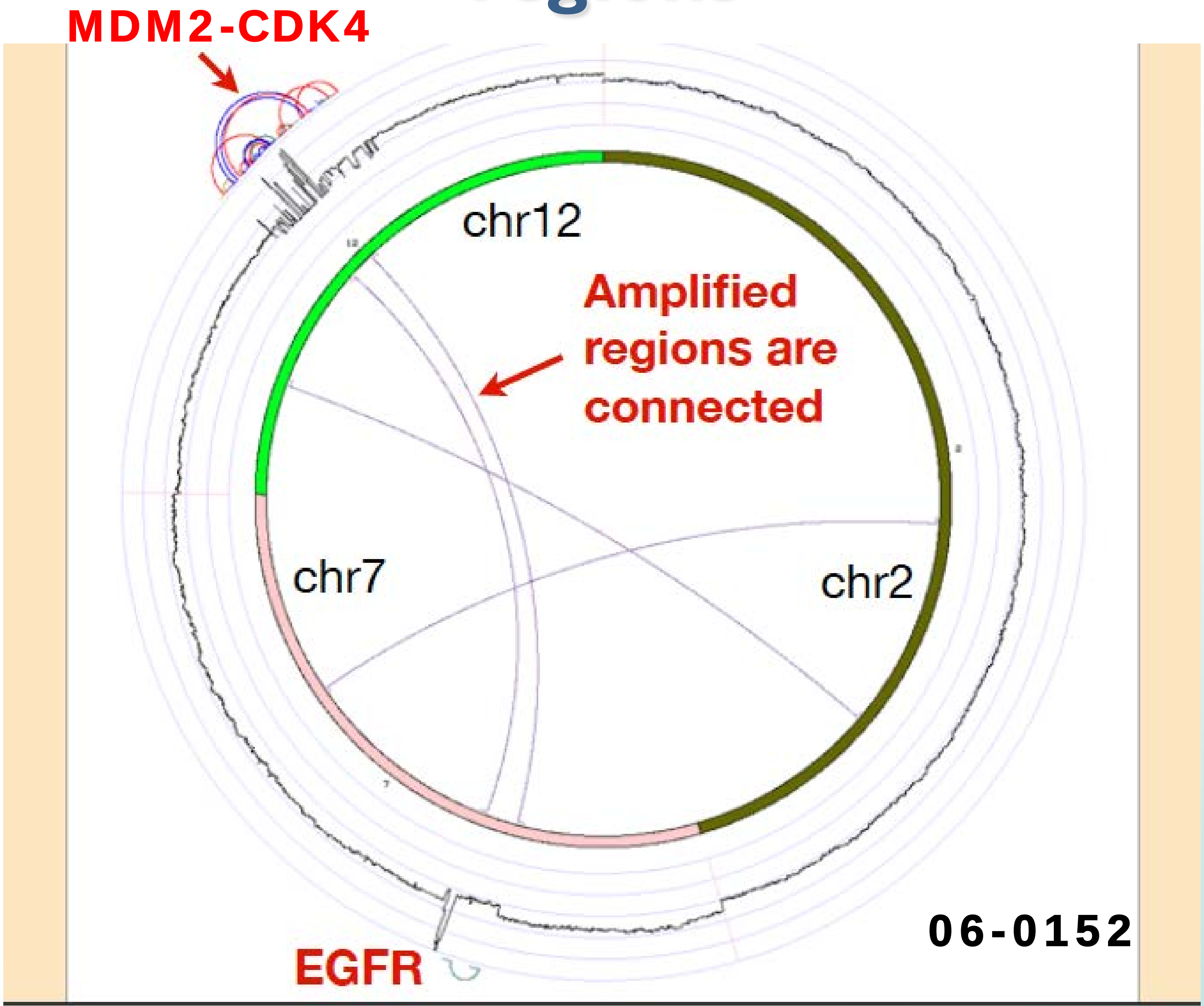
estimated normal contamination

Ovarian TCGA-13-1411



Zack Sanborn

Many rearrangements in amplified regions


MDM2-CDK4
chr12
Amplified regions are connected
chr7
chr2
EGFR