Exploring the structure and function of genomes.

Aaron Quinlan

quinlanlab.org

University of Virginia, Charlottesville VA

Center for Public Health Genomics

Biochemistry and Molecular Genetics

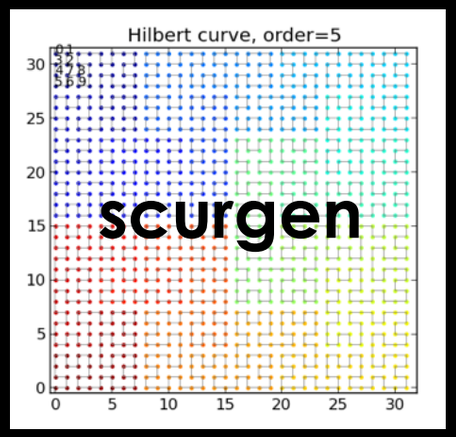
What do we work on?

Software tools for genome research

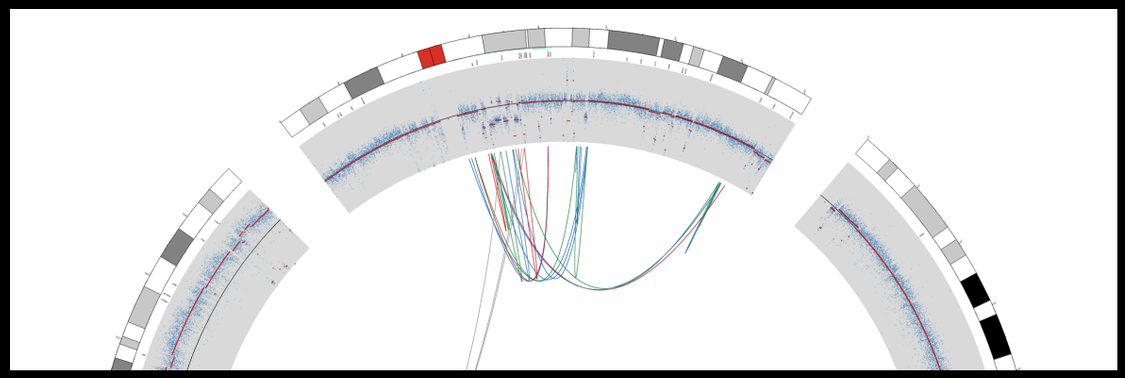
hydra

lumpy

genome query language (GQL)



Cancer genomics



Cellular and molecular origins of GBM
Initiating mutations underlying OV?
Tumor evolution and mutational mechanisms.

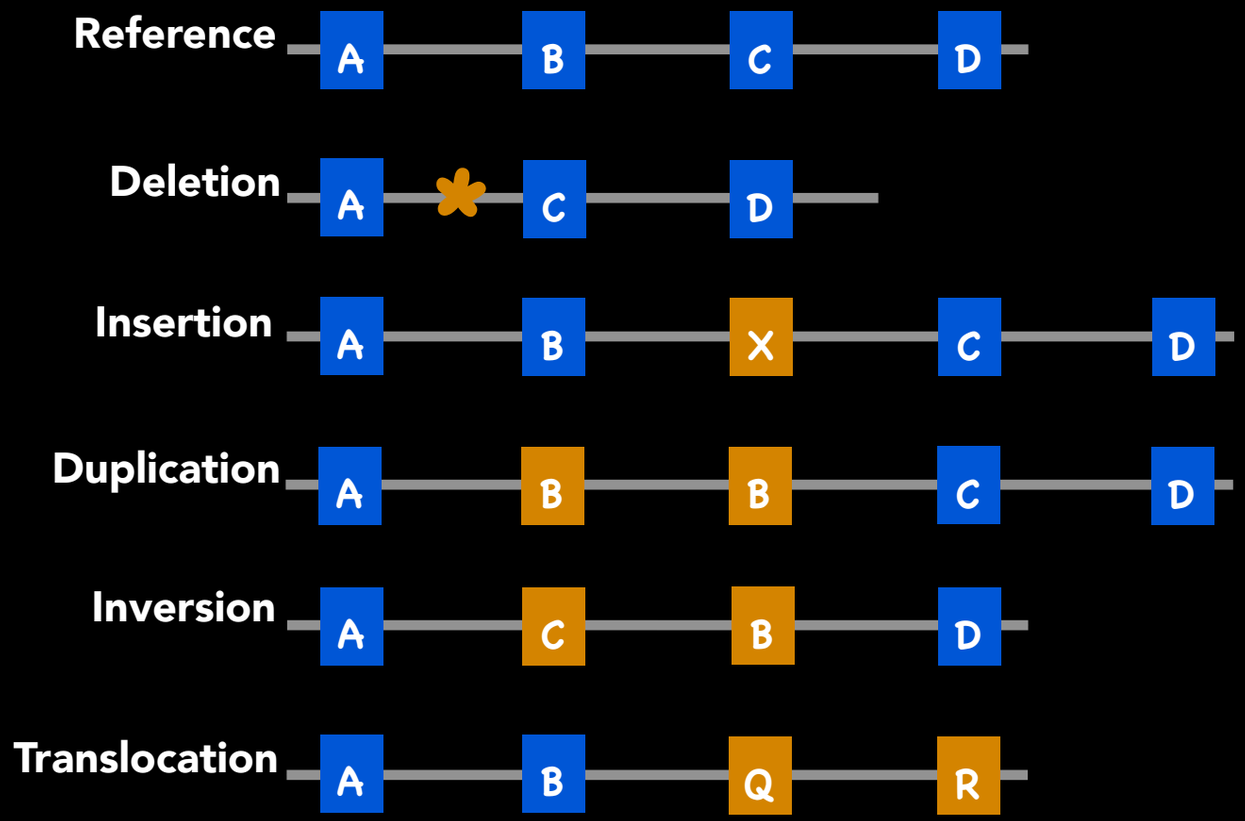
Disease genetics

Type 1 diabetes
SLE (Lupus)
Unexplained development disorders
Radiation hypersensitivity

Outline

- Discovery of structural variation (SV)
- Complex SV in 64 tumor genomes
- Inferring function from complex genomic datasets

SV definitions



structural variant (SV): a difference in the copy number, orientation or location of genomic segments >100bp

genomic rearrangement: ditto

copy number variant (CNV), or alteration (CNA): an SV that alters DNA copy number

breakpoint: The junction(s) between structurally variable genomic segments

complex SV: 2 or more breakpoints that arise through a single mutational event, but cannot be explained by one DNA exchange or end-joining reaction

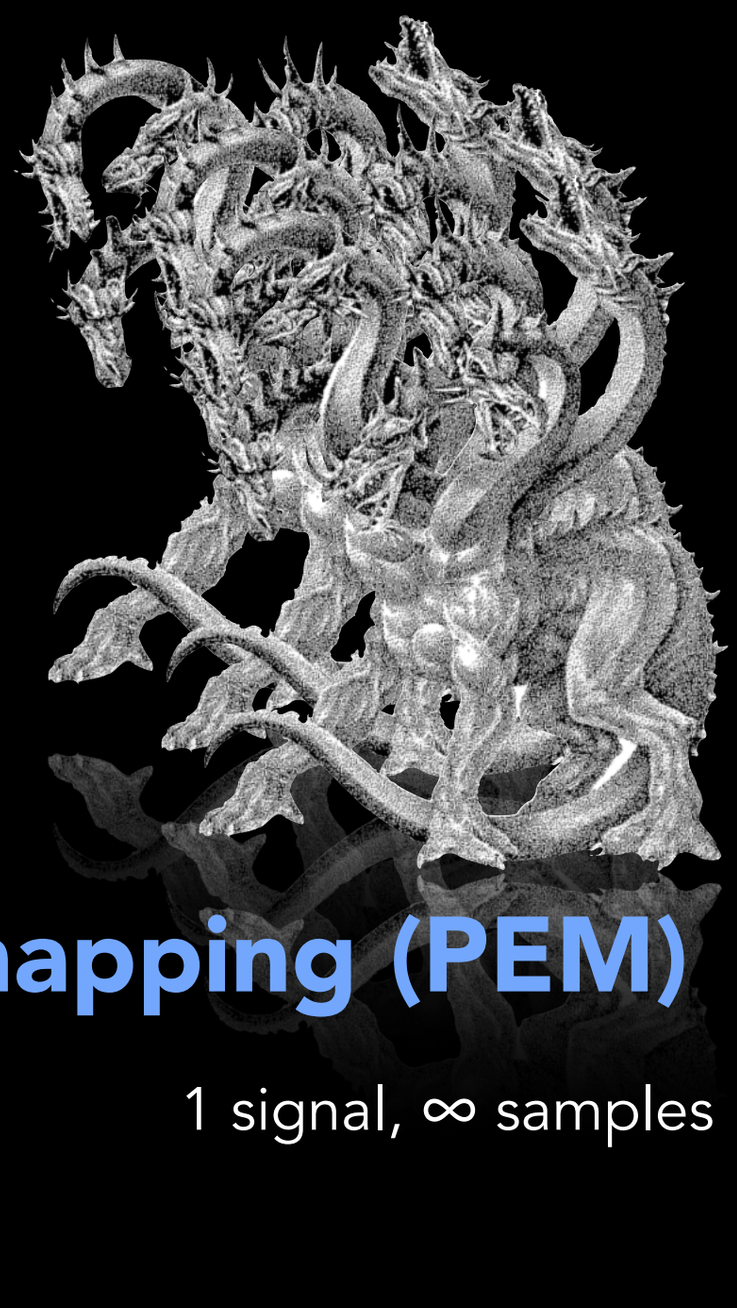
Evolution of our SV discovery tools

Hydra
(2010)

Paired-end mapping (PEM)

1 signal, 1 sample

Hydra_Multi
(2011)



1 signal, ∞ samples

LUMPY
(2012)

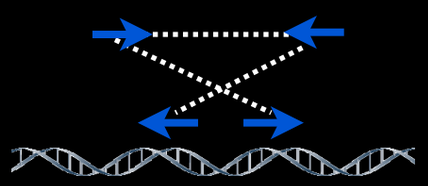
lumpy

A Probabilistic Framework for Structural Variant Discovery

many signals,
many samples

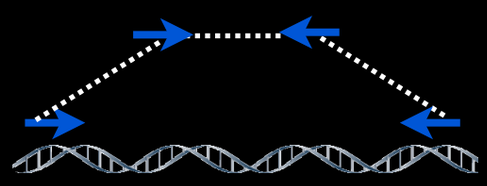
PEM clusters ***discordant*** mappings

everted orientation



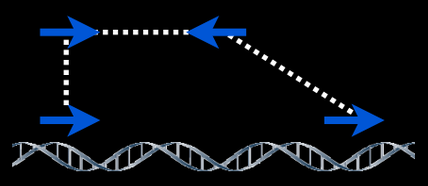
tandem duplication

too big



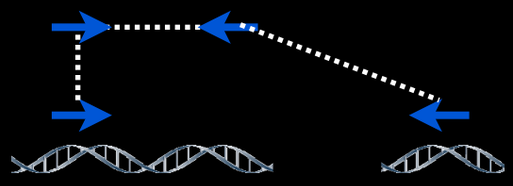
deletion

same-strand



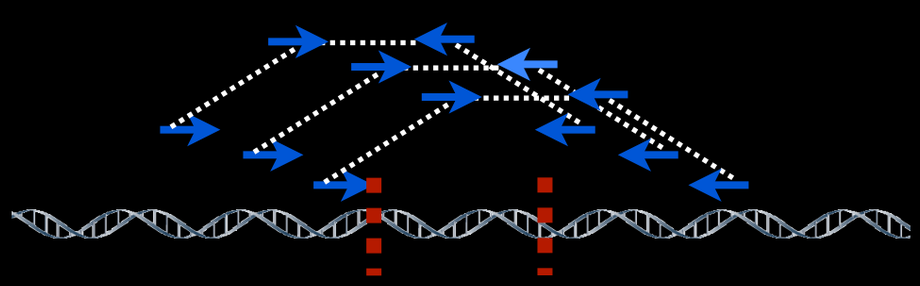
inversion

distant linkage



insertion,
retrotransposition,
translocation

Cluster to localize breakpoints



ref. genome

Deleted interval

Hydra

1 signal (PEM), 1 sample

The Hydra algorithm:

- simple & fast
- comprehensive: detects all breakpoint classes
- combinatorial: optionally uses multiple mappings to detect mobile element insertions
- Quinlan et al., 2010. *Genome Research*; http://code.google.com/p/hydra-sv/

The dirty secrets of PEM

Secret #1.

Often many false positives

- Short reads + heuristic alignment + rep. genome
  = systematic alignment artifacts (false calls)
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL** SV mapping studies use strict filters for above

Secret #2.

The false negative rate is also generally very high.

- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)

Secret #2.

The false negative rate is also generally very high.

- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment

Secret #2.

The false negative rate is also generally very high.

- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!

Secret #2.

The false negative rate is also generally very high.

- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)

Secret #2.

The false negative rate is also generally very high.

- Most current datasets have low to moderate physical coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

Solution to both FPs and FNs:

Pool data from mutiple samples.

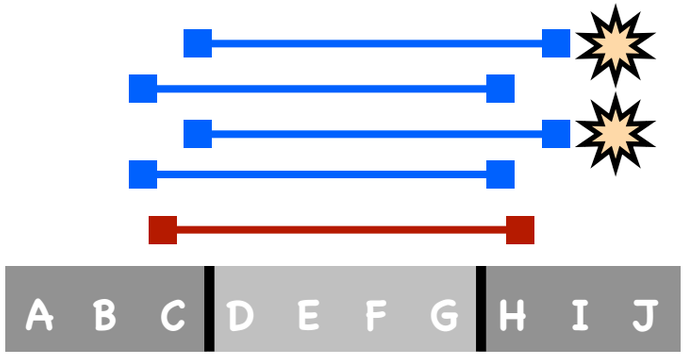
It improves SNP and INDEL calling, so why not SVs?

Pooling prevents false somatic calls

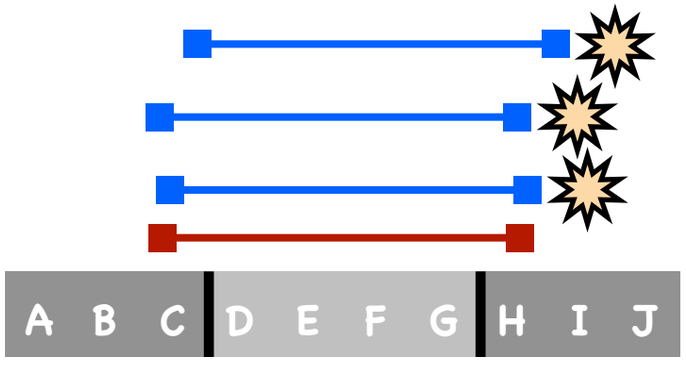
tumor 1 **normal 1** . . . **tumor 64** **normal 64**

clustering

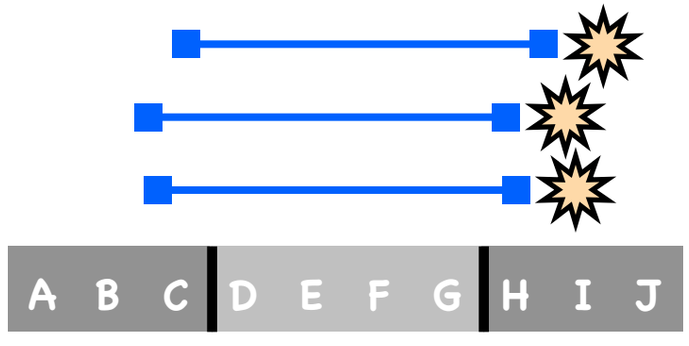
inherited mutation



inherited mutation
(FP somatic w/o pooling)



somatic mutation



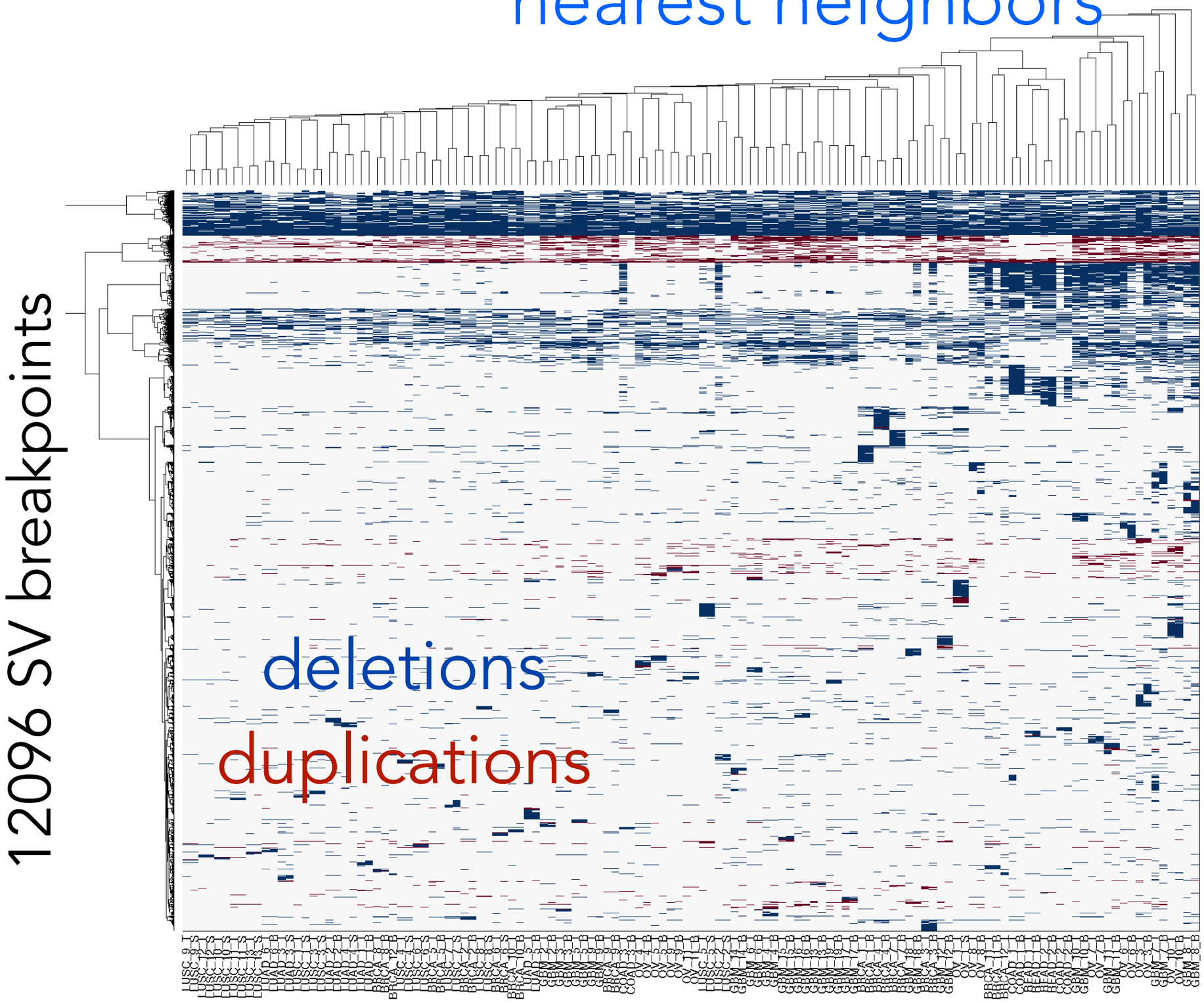
The landscape of complex variation in 64 cancer genomes.

64 Tumors and 65 matched normals (1 dup.)

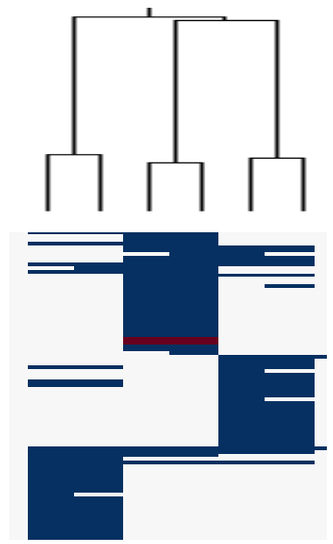
The Cancer Genome Atlas  *Understanding genomics to improve cancer care*

- 12 breast invasive carcinomas (BRCA)
- 3 colon adenocarcinomas (COAD)
- 18 glioblastoma multiforme (GBM)
- 6 lung adenocarcinoma (LUAD)
- 13 lung squamous cell carcinoma (LUSC)
- 11 ovarian serous cystadenocarcinoma (OV)
- 2 rectum adenocarcinoma (READ)

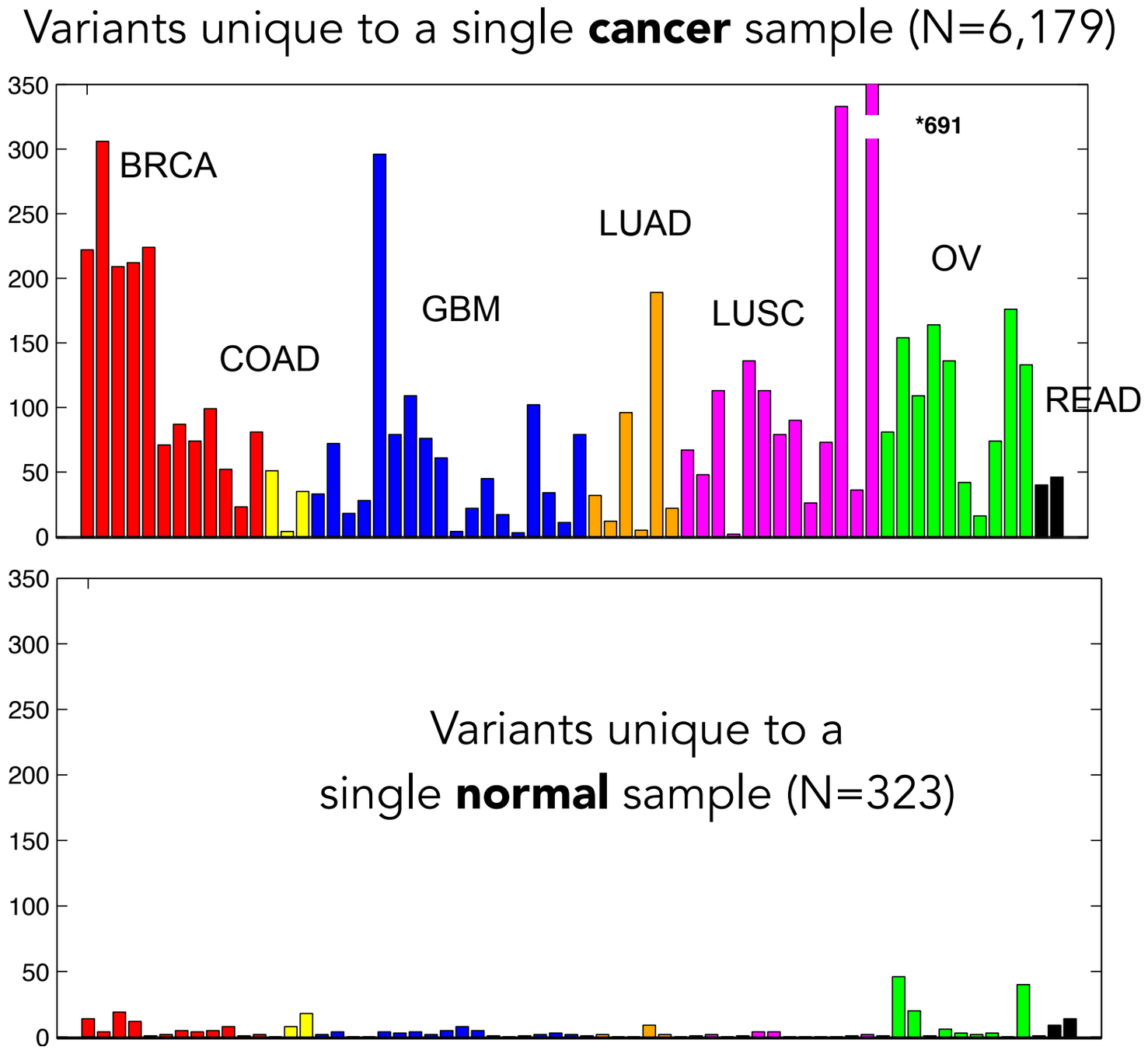
64 out of 64 tumor / normal pairs cluster as nearest neighbors



3 tumor/
normal pairs



Pooling yields accurate predictions of somatically-acquired SVs in tumors.

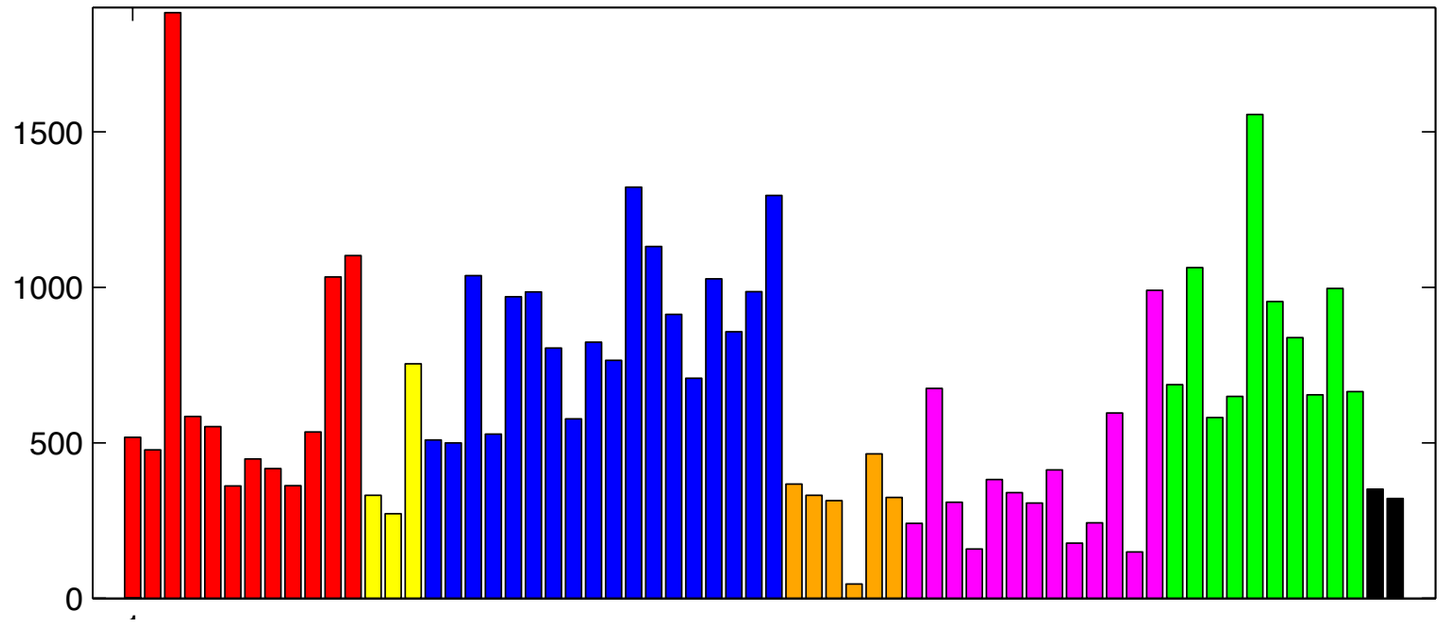


Assuming all normal-only calls are false, suggests 5% somatic prediction error rate.

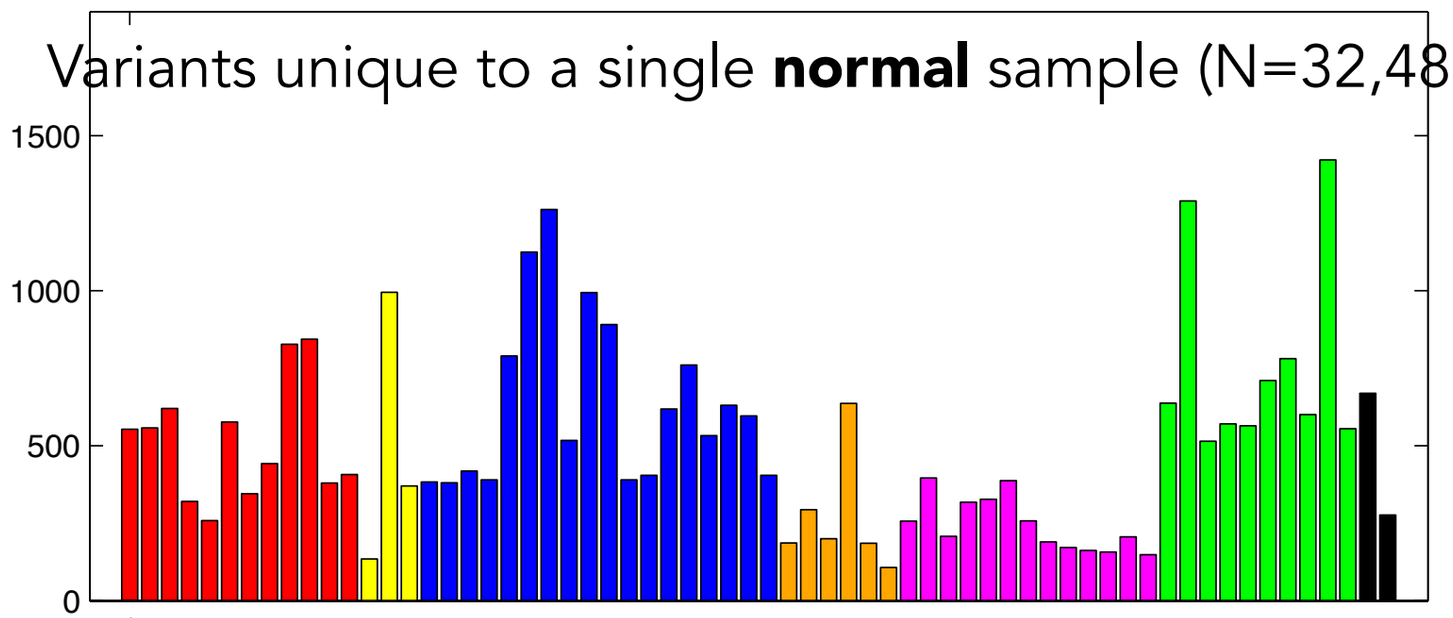
Likelihood of LOH suggests it is actually lower.

Much worse if we just did a simple tumor/normal comparison (the standard)

Variants unique to a single **cancer** sample (N=41,510)



Variants unique to a single **normal** sample (N=32,482)



Somatic misclassification rate jumps from 5% with pooling to 86%!

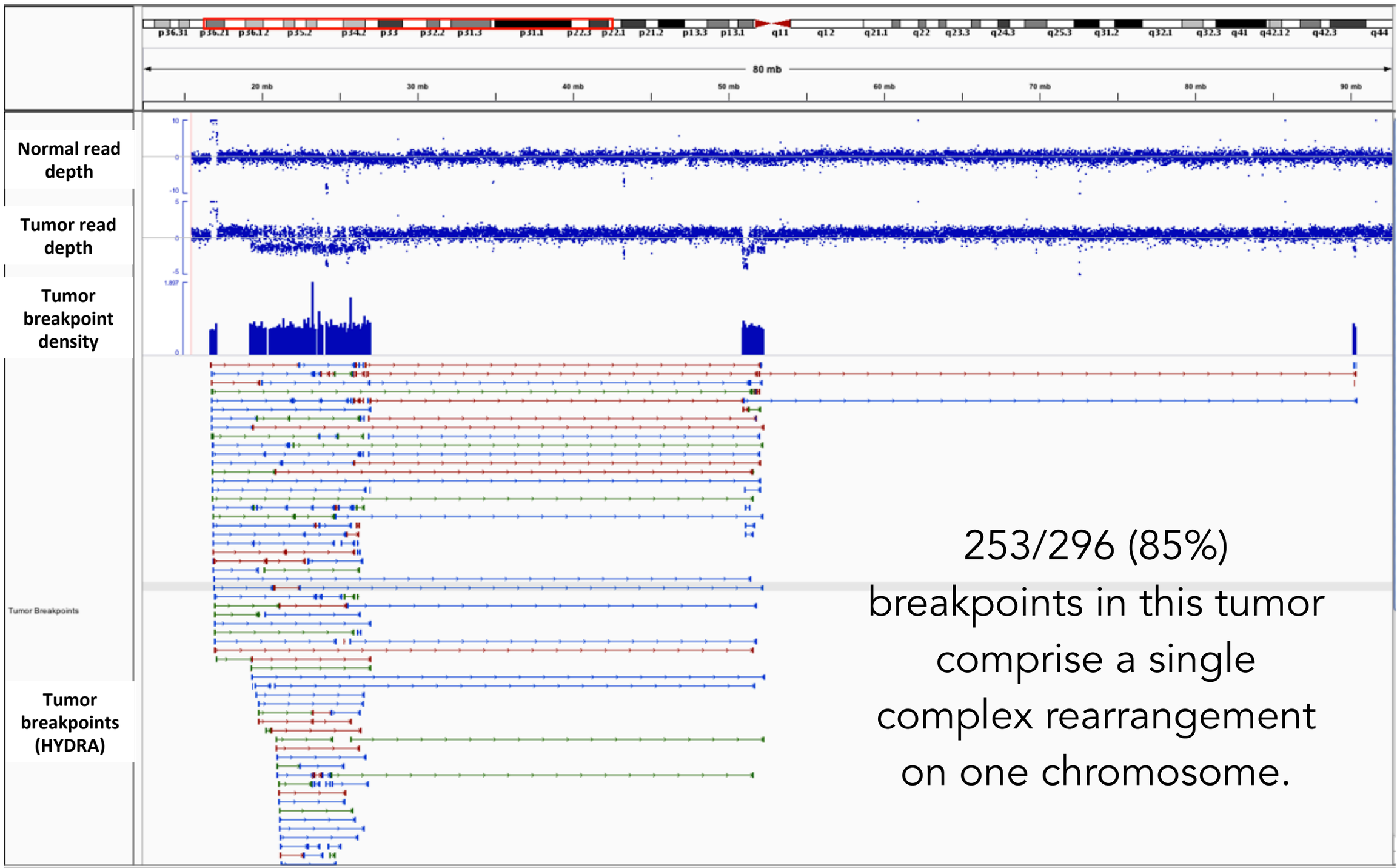
We have a high-quality set of somatic rearrangements from multiple tumors.

What do they tell us about chromosome evolution in cancers?

Observation 1.

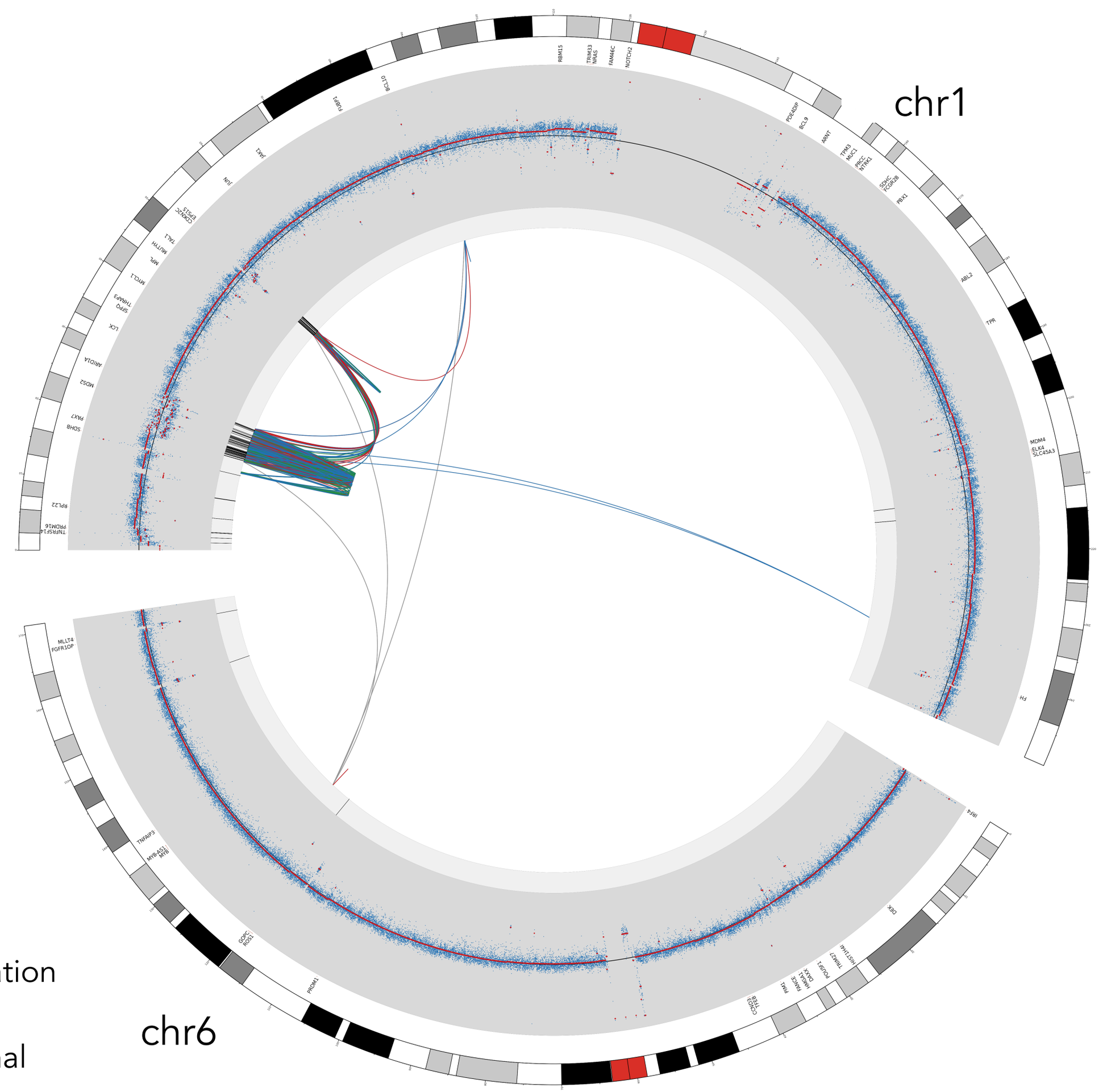
We immediately noticed a few staggeringly complex rearrangements (CRs).

A mangled tumor chromosome in IGV



This has been attributed to new mechanisms: chromothripsis (Stephens et al. 2011). The mechanism is not known.

CIRCOS plot of same event.



red = deletion
green = tandem duplication
blue = inversion
grey = inter-chromosomal

Chromothripis

Cell

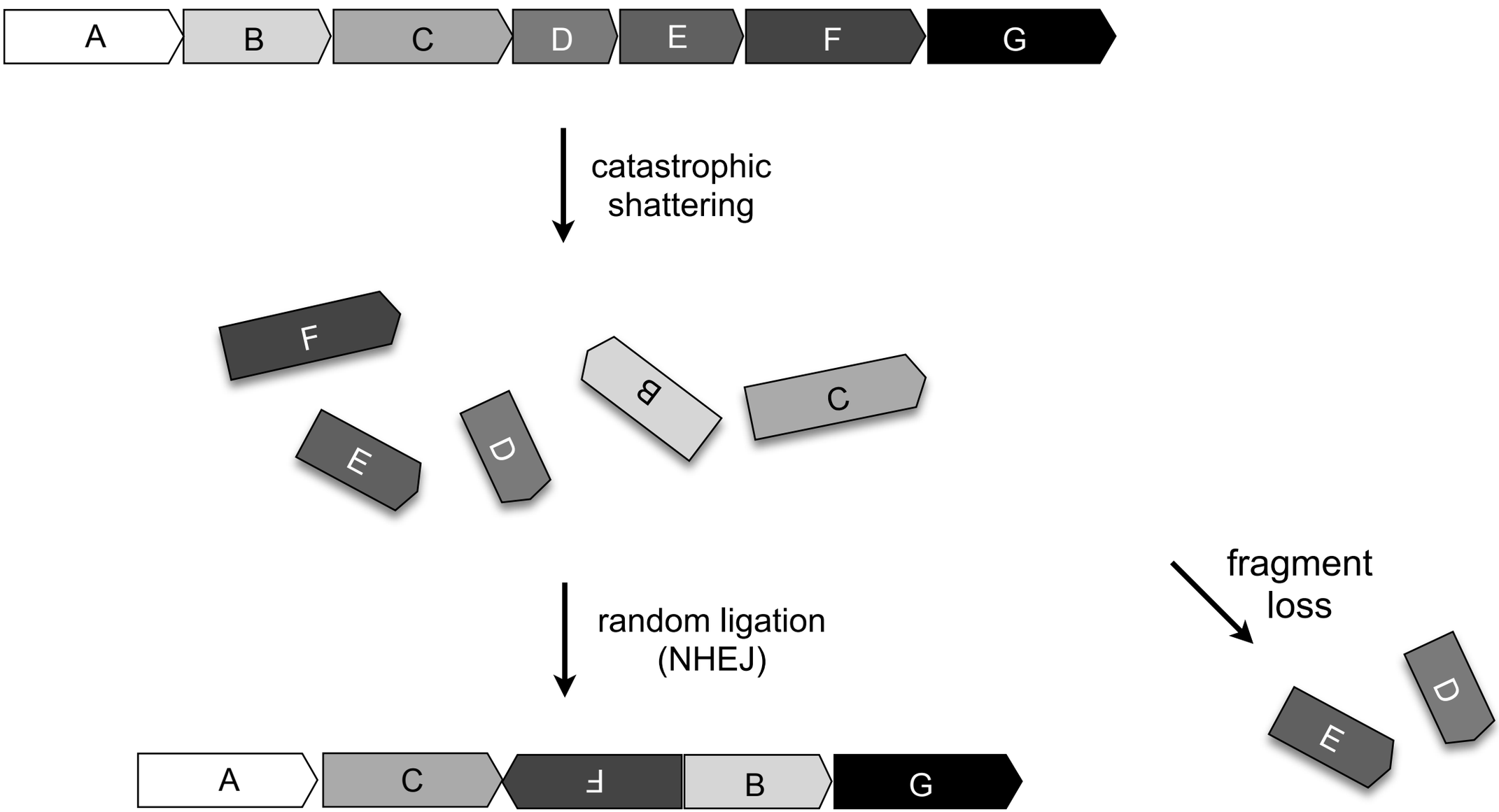
Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development

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,*]

Chromosome shattering in a single, catastrophic event.

A new, punctate form of chromosome evolution.

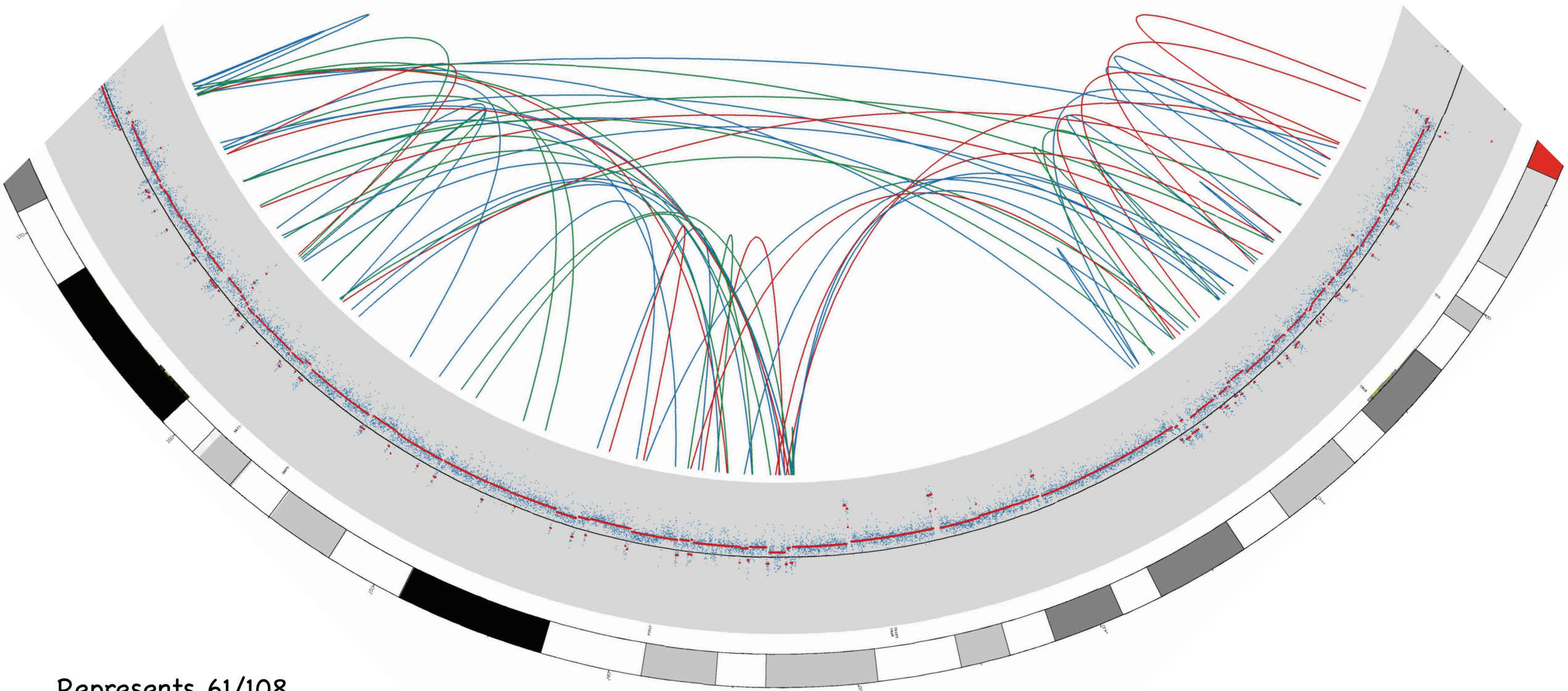
Chromothripis



Observation 2.

Complex rearrangements (CRs) are very common in tumor genomes.

Glioblastoma sample #6



Represents 61/108
breakpoints in this tumor;

1542 of 6179 breakpoints are part of CRs.
Not random.

	Experimental Data				Simulations (mean of 100 trials)			
	Tumor-specific mutations (n=6179)		Normal-specific mutations (n=323)		Tumor-specific (random shuffle)		1000 Genomes (random sample)	
class (num. breaks)	num. CGRs	total breaks	num. CGRs	total breaks	num. CGRs	total breaks	num. CGRs	total breaks
mild (3-4)	90	298	3	9	4.7	14.3	2.2	6.7
moderate (5-9)	32	204	0	0	0.2	1.1	0.1	0.7
extreme (>9)	33	1045	0	0	0	0	0	0
total	154	1542	3	9	4.9	15.4	2.3	7.4

Observation 3.

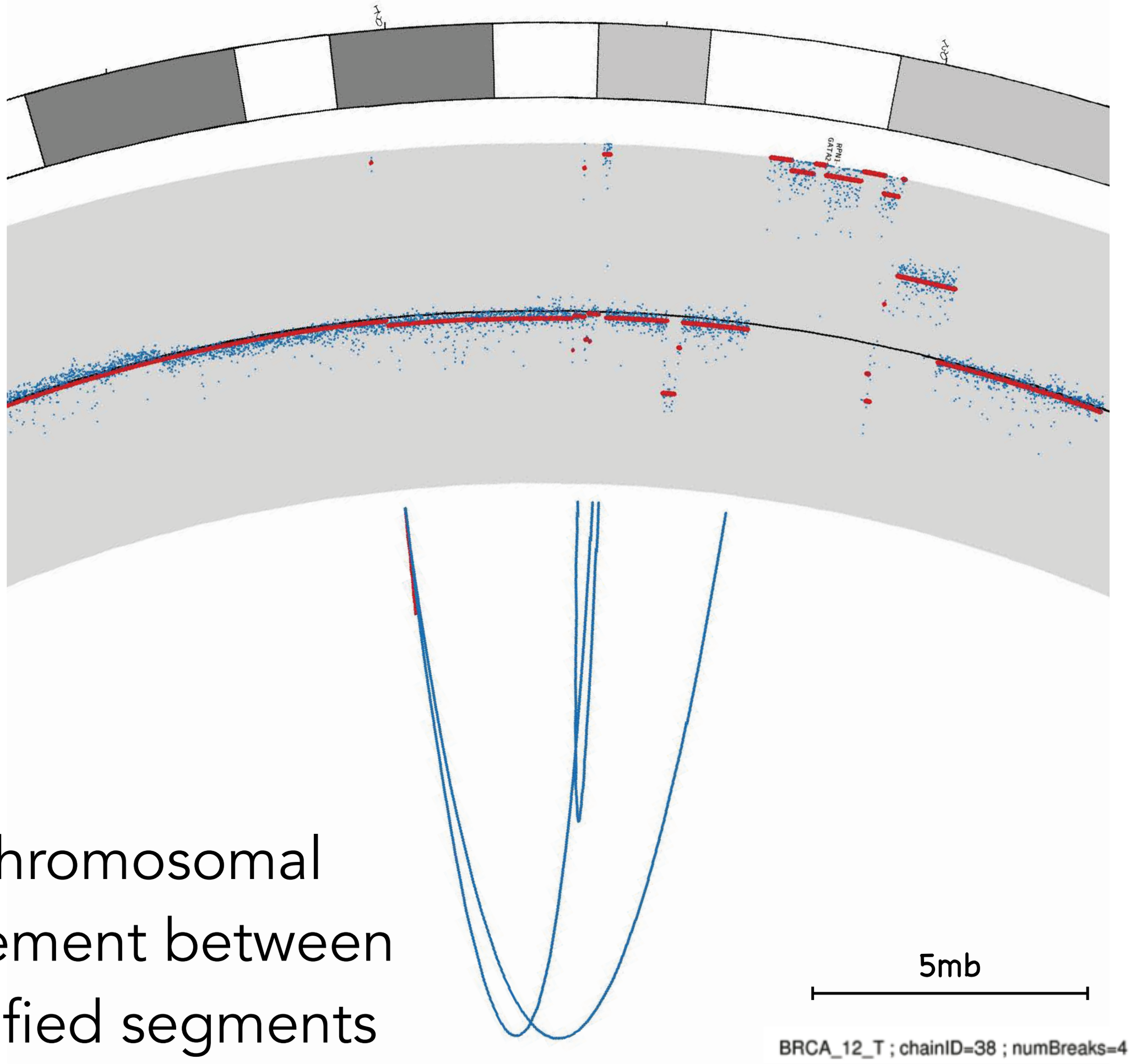
Complex rearrangements are very common in glioblastoma.

1542 of 6179 breakpoints are part of CRs.
Not random.

	total breaks (mean)	% complex
BRCA (n=12)	1657 (138)	4.2%
COAD (n=3)	90 (30)	10%
GBM (n=18)	1088 (60)	70%
LUAD (n=6)	356 (59)	23%
LUSC (n=13)	1806 (139)	26.7%
OV (n=11)	1096 (100)	11.6%
READ (n=2)	86 (43)	11.6%
total	6179	25%

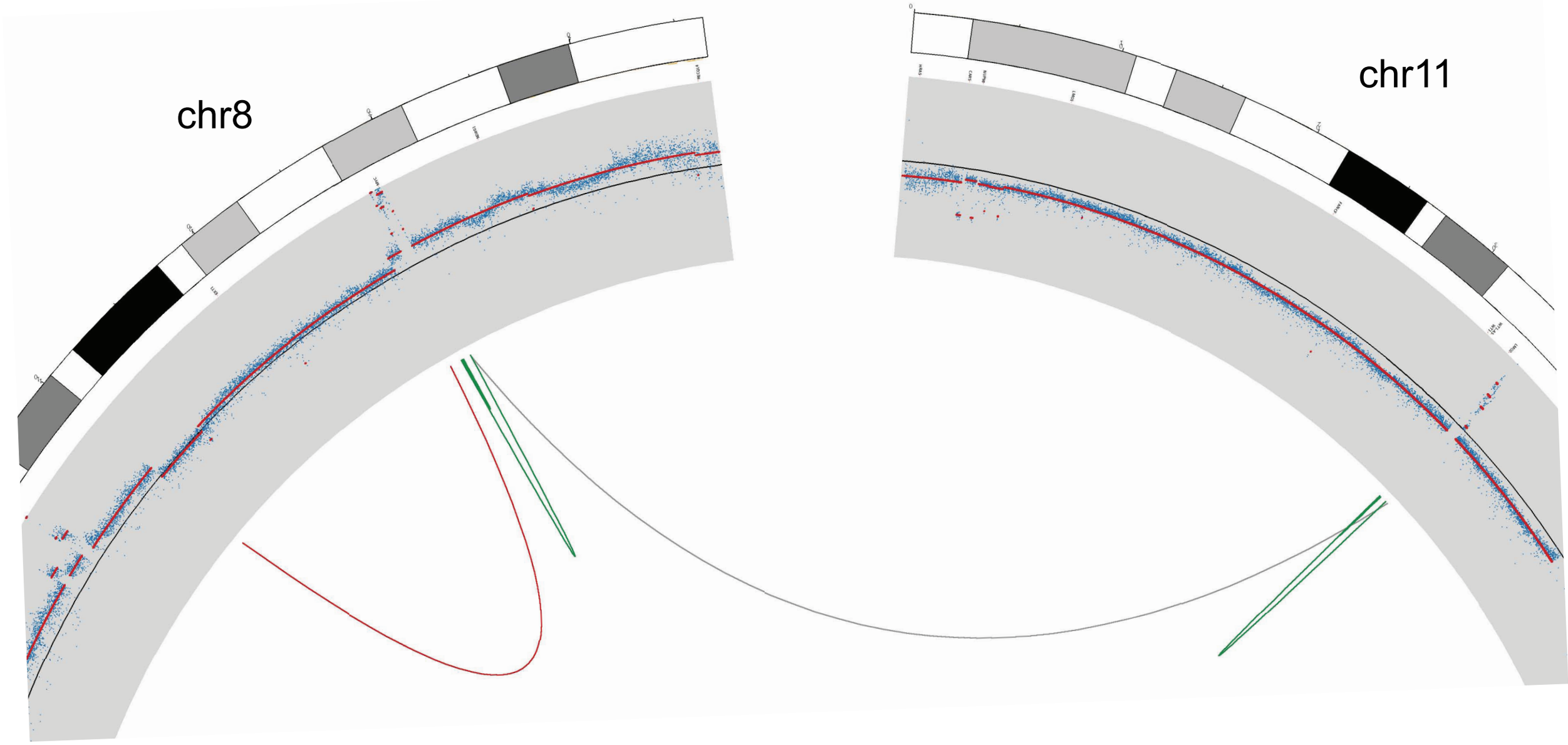
Observation 4.

There is vast architectural diversity
in non-chromothripsis
complex variants

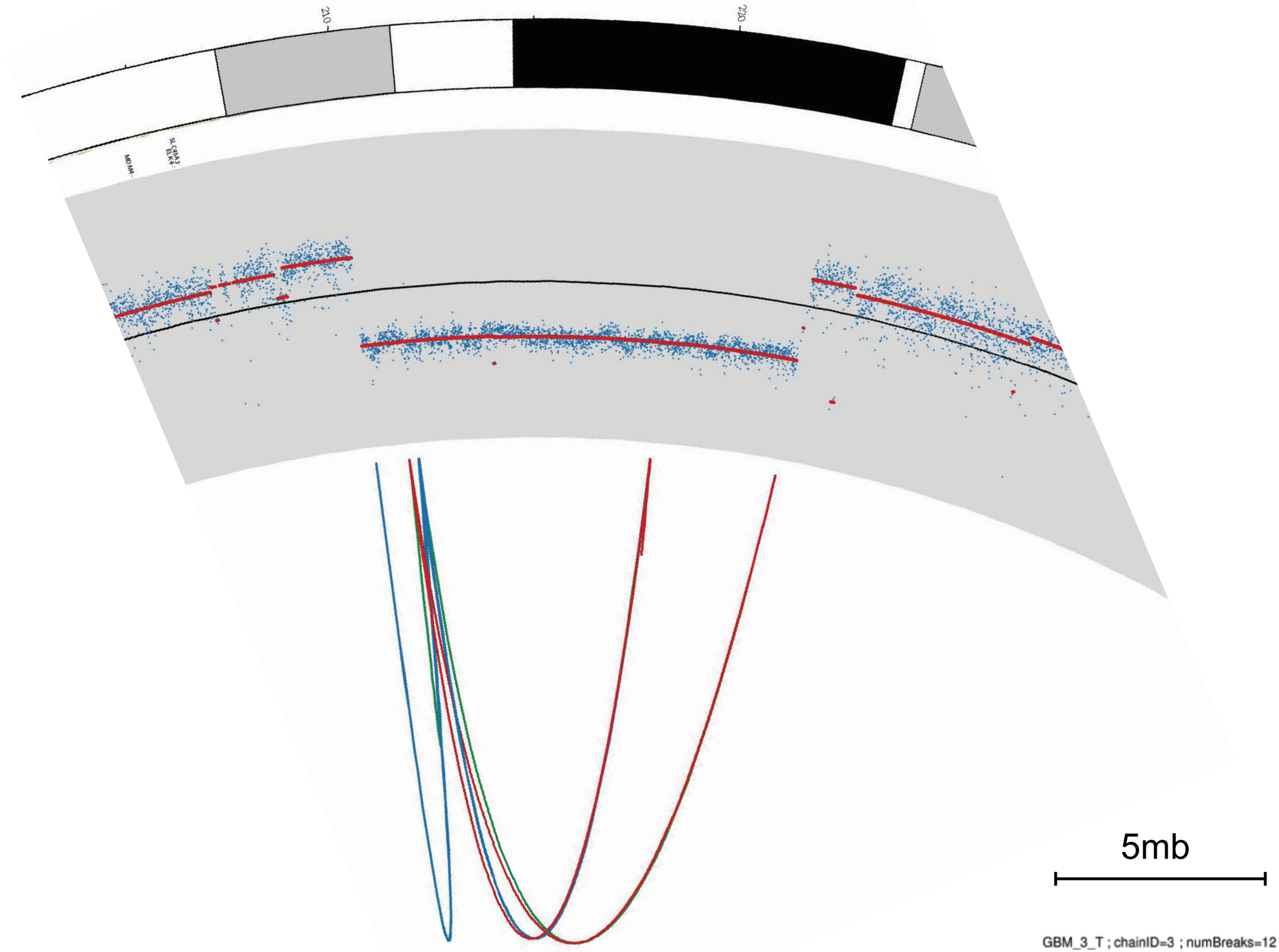


intra-chromosomal rearrangement between co-amplified segments

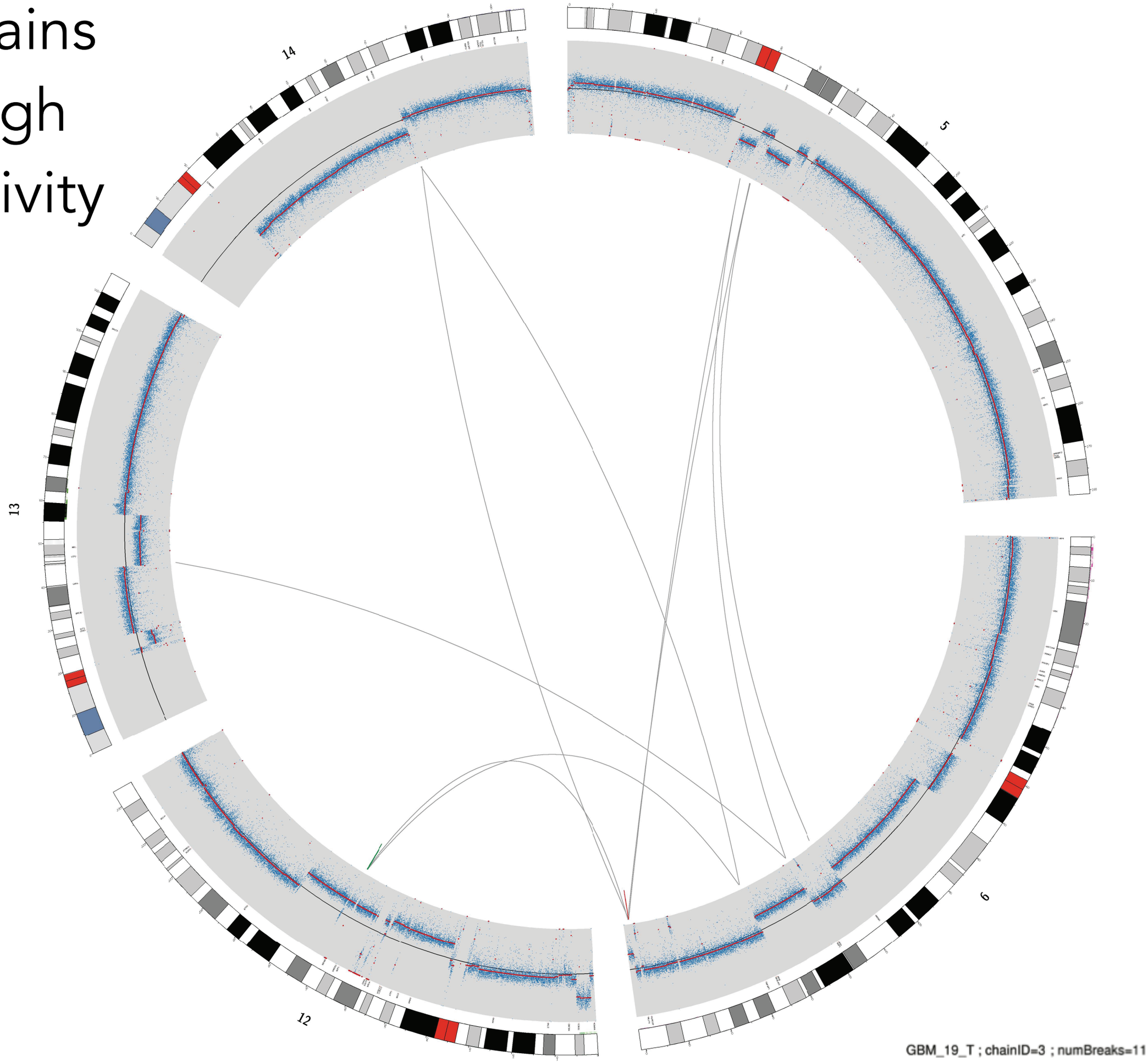
inter-chromosomal rearrangement between co-amplified segments



cryptic
rearrangements

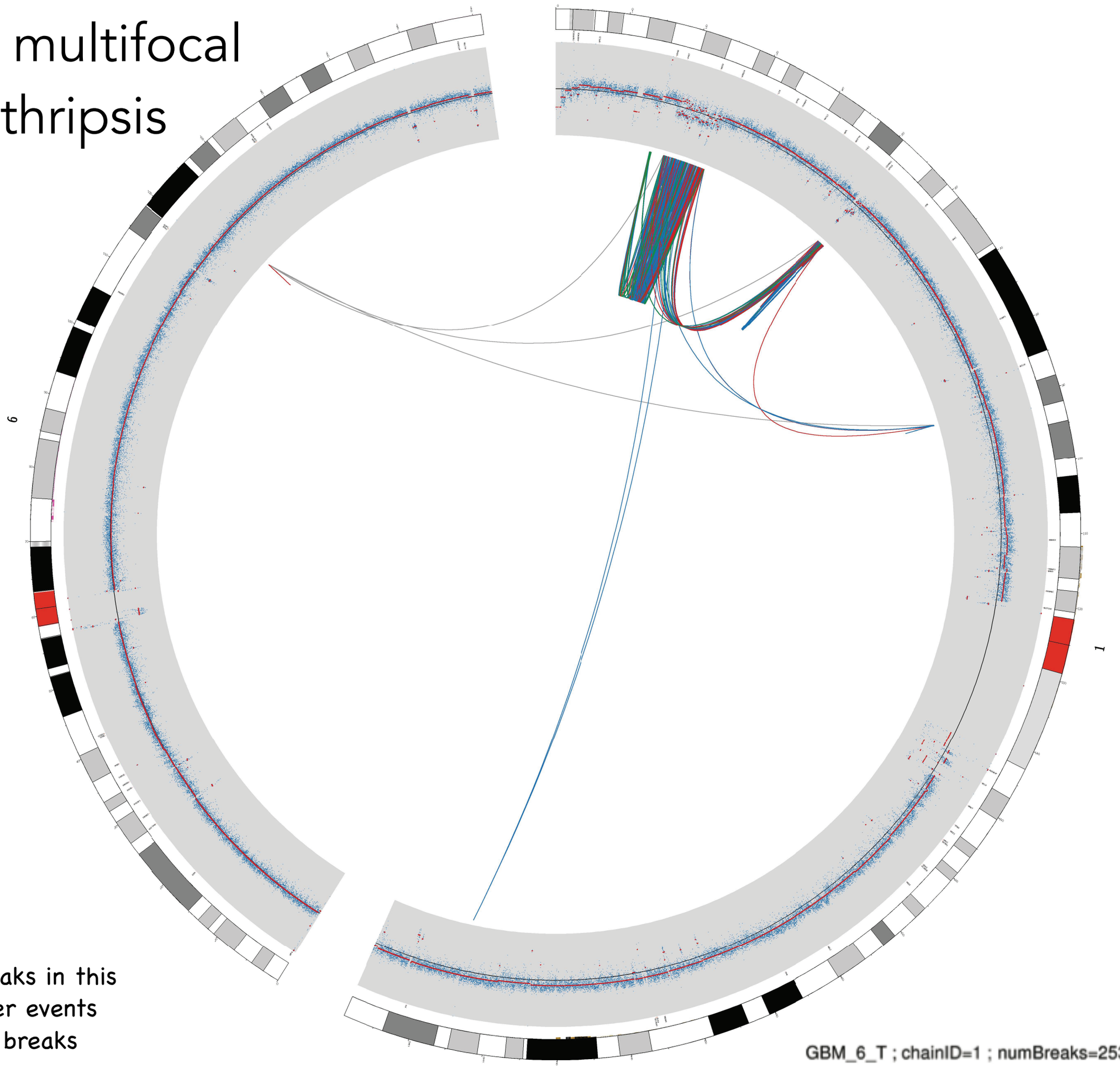


long chains with high connectivity



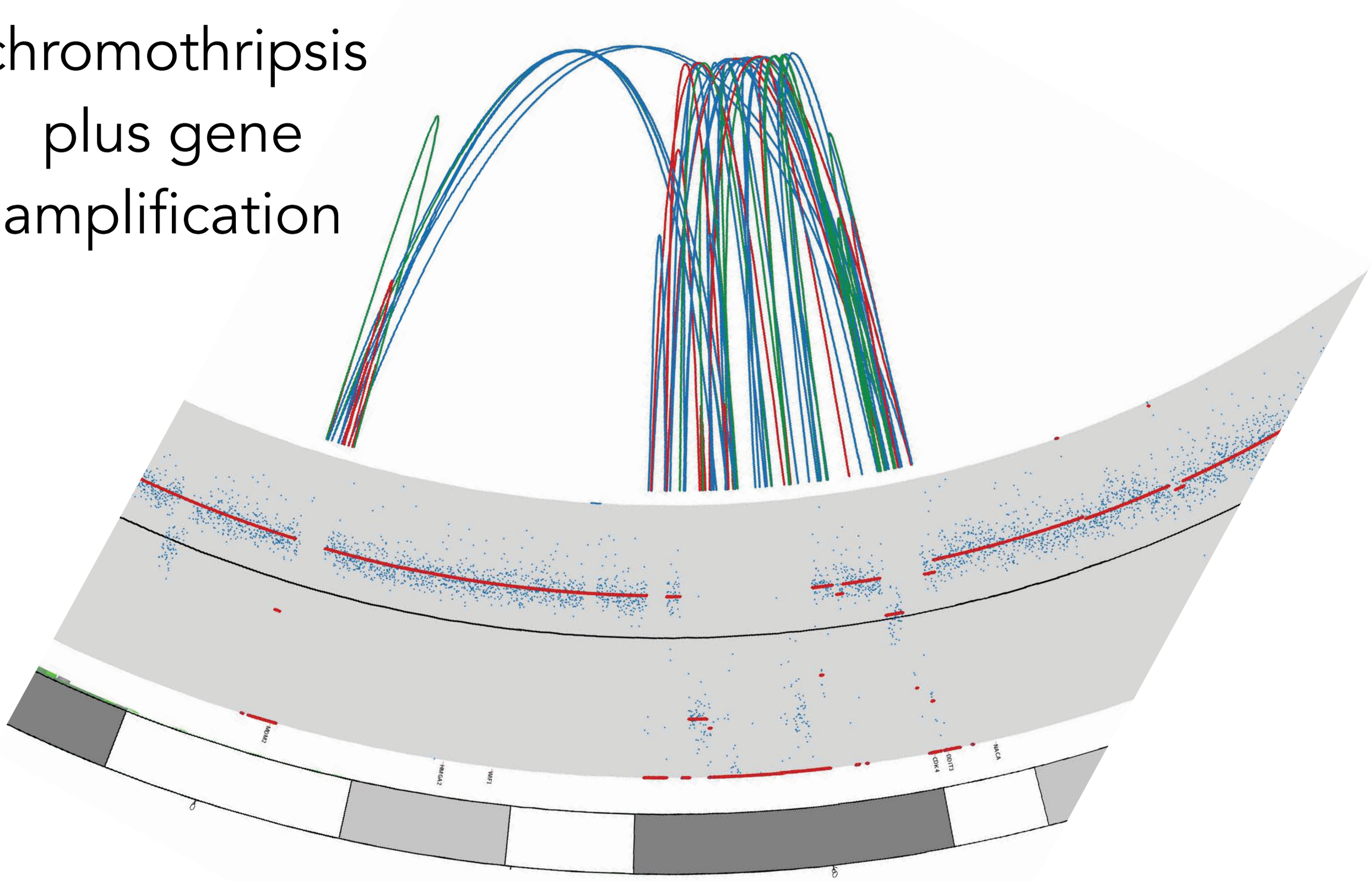
Many chromothripsis examples are horrifically complex.

The classic multifocal chromothripsis



Represents 253/296 breaks in this tumor; this and 4 other events account for 271/296 breaks

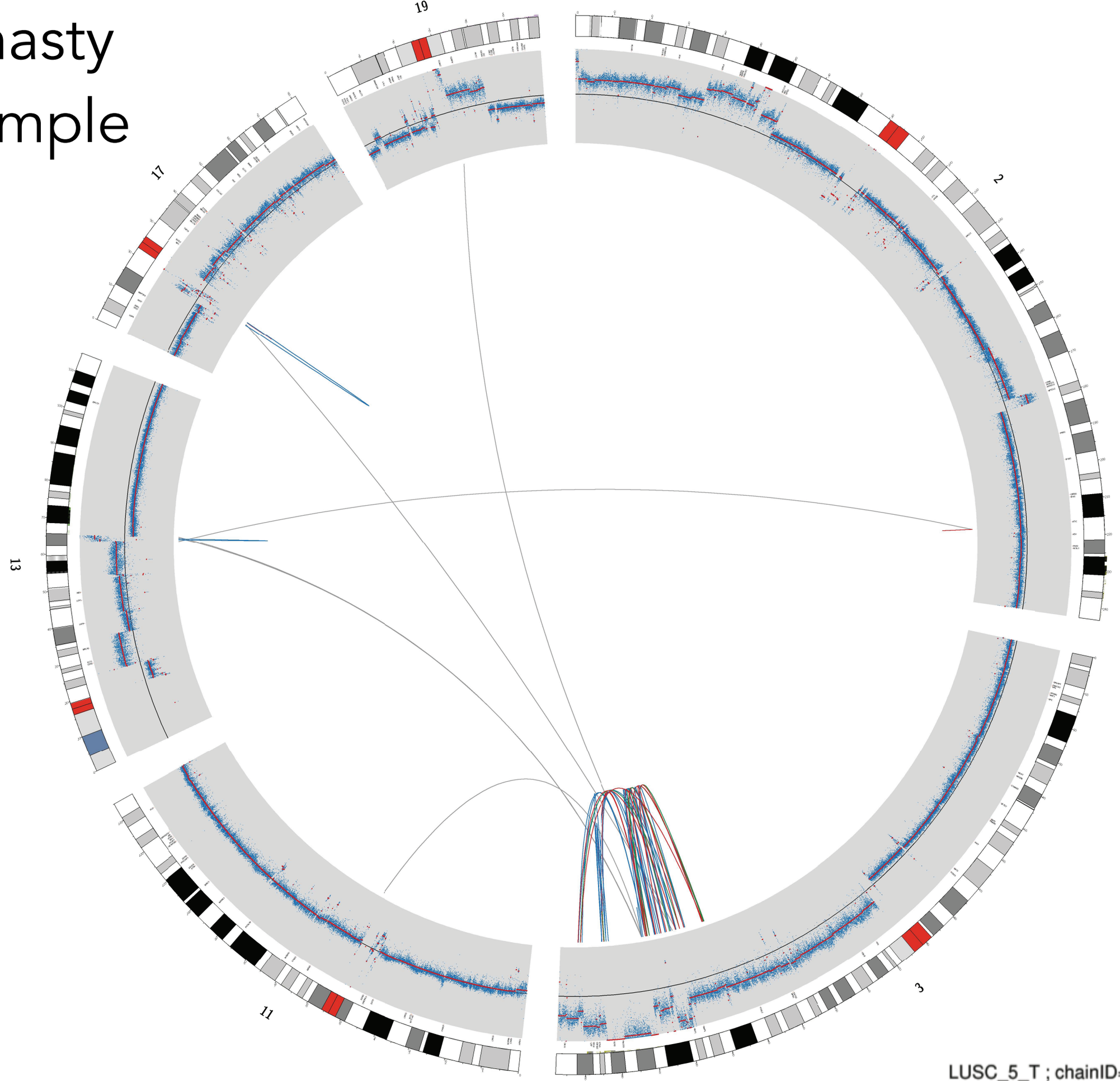
chromothripsis plus gene amplification



amplifies mdm2, cdk4, DDIT3

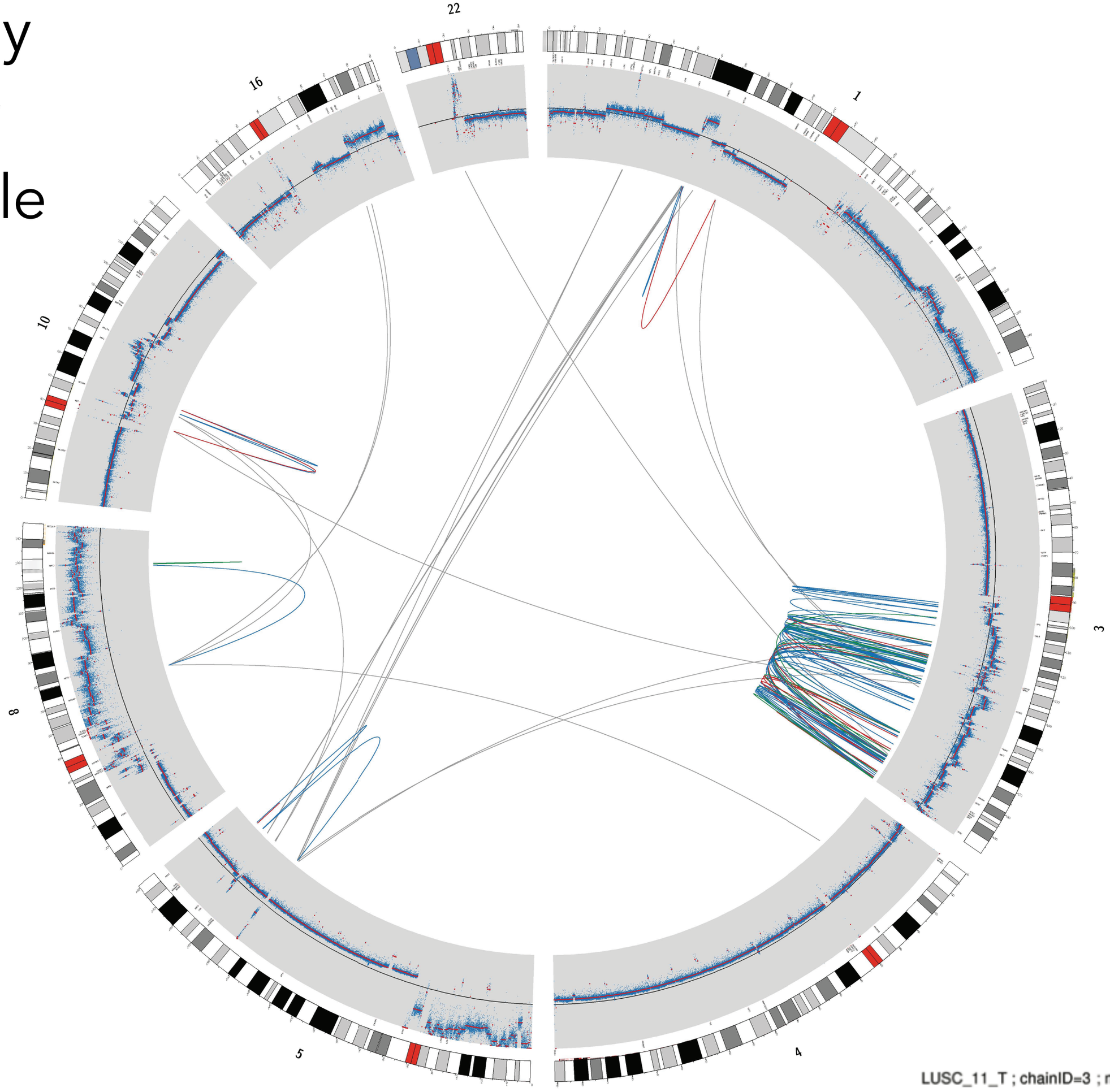
GBM_19_T ; chainID=2 ; numBreaks=56

A nasty example

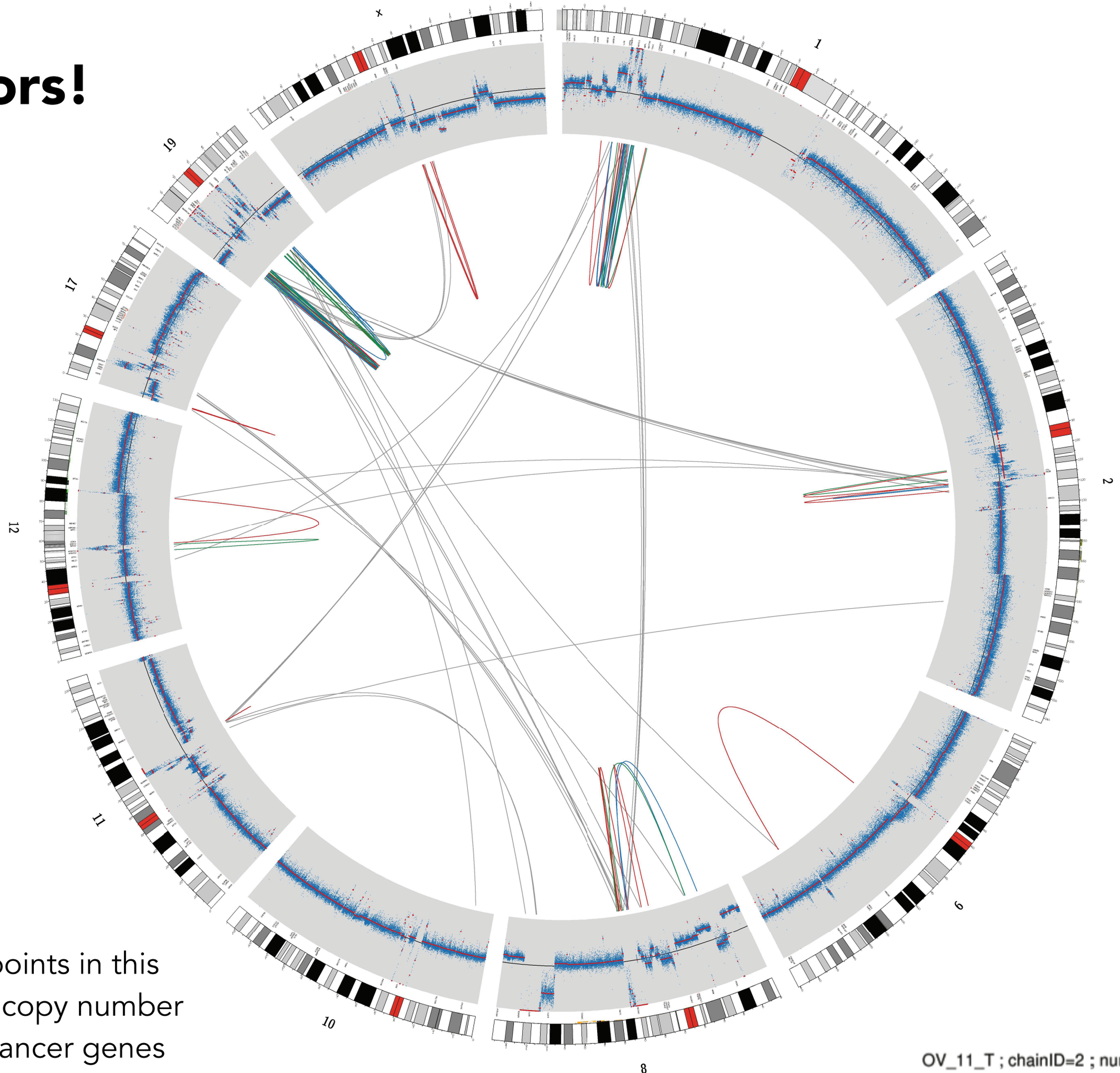


LUSC_5_T ; chainID=5 ; numBreaks=71

A really nasty example



Horrors!



93/132 breakpoints in this sample; alters copy number of numerous cancer genes

Summary

- Pooling data from multiple samples greatly improves SV discovery. Especially for identifying variants that are private to a single sample (e.g. tumors)
- Complex SVs are quite common in tumors.
- Many appear to be chromothripsis.
- 70% of glioblastomas have very complex rearrangements

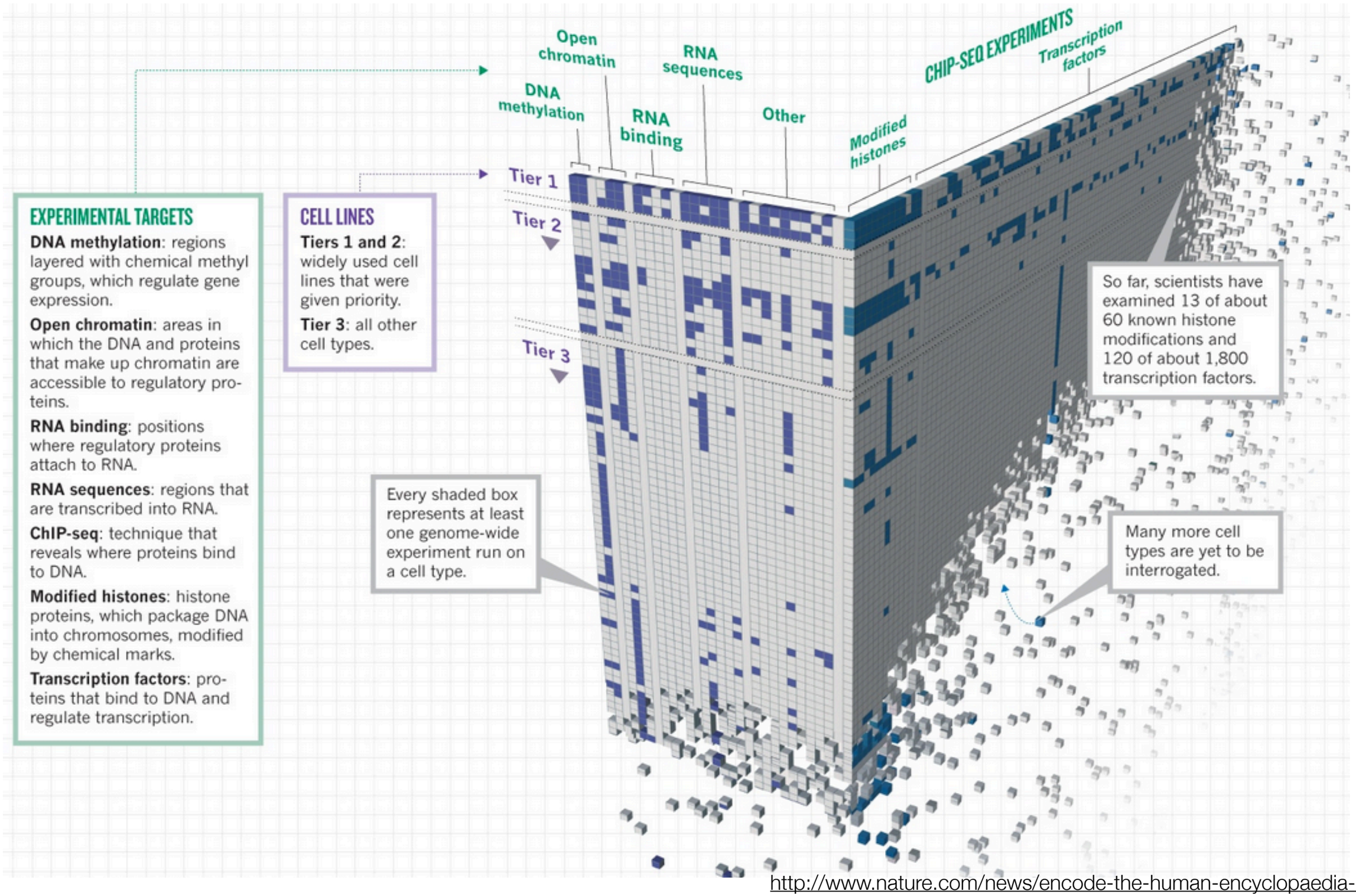
Questions we are tackling now:

1. How do complex rearrangements arise?
   - single event or stepwise accumulation?
   - look at copy number states.

2. Are complex rearrangements early events in tumorigenesis, or the tardy consequence of genomic instability?
   - look at allele frequencies and heterogeneity.

Now for something different.

How do we make sense of complex datasets to gain insights into genome biology?



http://www.nature.com/news/encode-the-human-encyclopaedia-1.11312

Scale
chr21:
200 kb
hg19
32,700,000
32,800,000
32,900,000
33,000,000
33,100,000
33,200,000
33,300,000
33,400,000
33,500,000
Uniqueness of 35bp Windows from ENCODE/OpenChrom(Duke)
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
TIAM1
TIAM1
TIAM1
TIAM1
BC014150
BC041449
SOD1
HUNK
LINC00159
SCAF4
SCAF4
SCAF4
SCAF4
SCAF4
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
GM12878 Syn Pk
H1-hESC Syn Pk
K562 Syn Pk
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
0
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
GM12878 Pk
GM12878 DS
GM12878 OS
H1hESC Pk
H1hESC DS
H1hESC OS
K562 Pk
K562 DS
K562 OS
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
GM12878 FAIRE Pk
GM12878 FAIRE DS
GM12878 FAIRE OS
H1-hESC FAIRE Pk
H1-hESC FAIRE DS
H1-hESC FAIRE OS
K562 FAIRE Pk
K562 FAIRE DS
K562 FAIRE OS
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
GM12878 Ht 1
GM12878 Pk 1
GM12878 Sg 1
GM12878 Ht 2
GM12878 Pk 2
GM12878 Sg 2
H1hESC Ht 1
H1hESC Pk 1
H1hESC Sg 1
K562 Ht 1
K562 Pk 1
K562 Sg 1
K562 Ht 2
K562 Pk 2
K562 Sg 2
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
K562 Sig
HepG2 Sig
HUVEC Sig

HeLa-S3 G2 1

HeLa-S3 Repli-seq Wavelet-smoothed Signal Rep 1 from ENCODE/UW

Replication Timing by Repli-seq from ENCODE/University of Washington

HepG2 G1b 1
HepG2 S2 1
HepG2 S3 1
HepG2 S4 1
HepG2 G2 1

HepG2 Repli-seq Wavelet-smoothed Signal Rep 1 from ENCODE/UW

Replication Timing by Repli-seq from ENCODE/University of Washington

HUVEC G1b 1
HUVEC S2 1
HUVEC S3 1
HUVEC S4 1
HUVEC G2 1

HUVEC Repli-seq Wavelet-smoothed Signal Rep 1 from ENCODE/UW

Placental Mammal Basewise Conservation by PhyloP

Mammal Cons

4 _
0 -
-4 _

Multiz Alignments of 46 Vertebrates

Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish

GERP scores for mammalian alignments

6.17 _
0 -
-12.1 _

GERP

Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples

Common SNPs(135)

Simple Nucleotide Polymorphisms (dbSNP build 131)

Duplications of >1000 Bases of Non-RepeatMasked Sequence

Repeating Elements by RepeatMasker

RepeatMasker

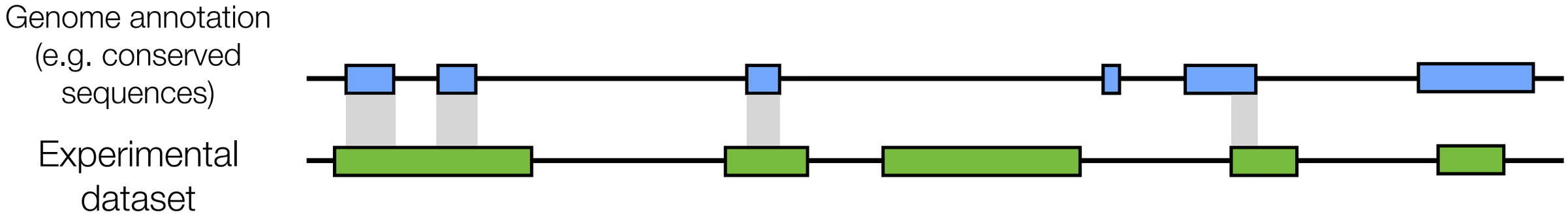
Simple Tandem Repeats by TRF

This is a hard yet important problem.

- Understand the function (or lack) of every base pair in different cell types and contexts.
- Challenges (among many):
  - Basic exploratory data analysis: slicing and dicing very large, heterogenous datasets.
  - Visualization: unbiased exploration; let the data tell its story.
  - **Testing for significant spatial relationships**

How do we detect genomic co-association?

That is, do two genomic features co-occur (overlap or have spatial consistency) more than expected?

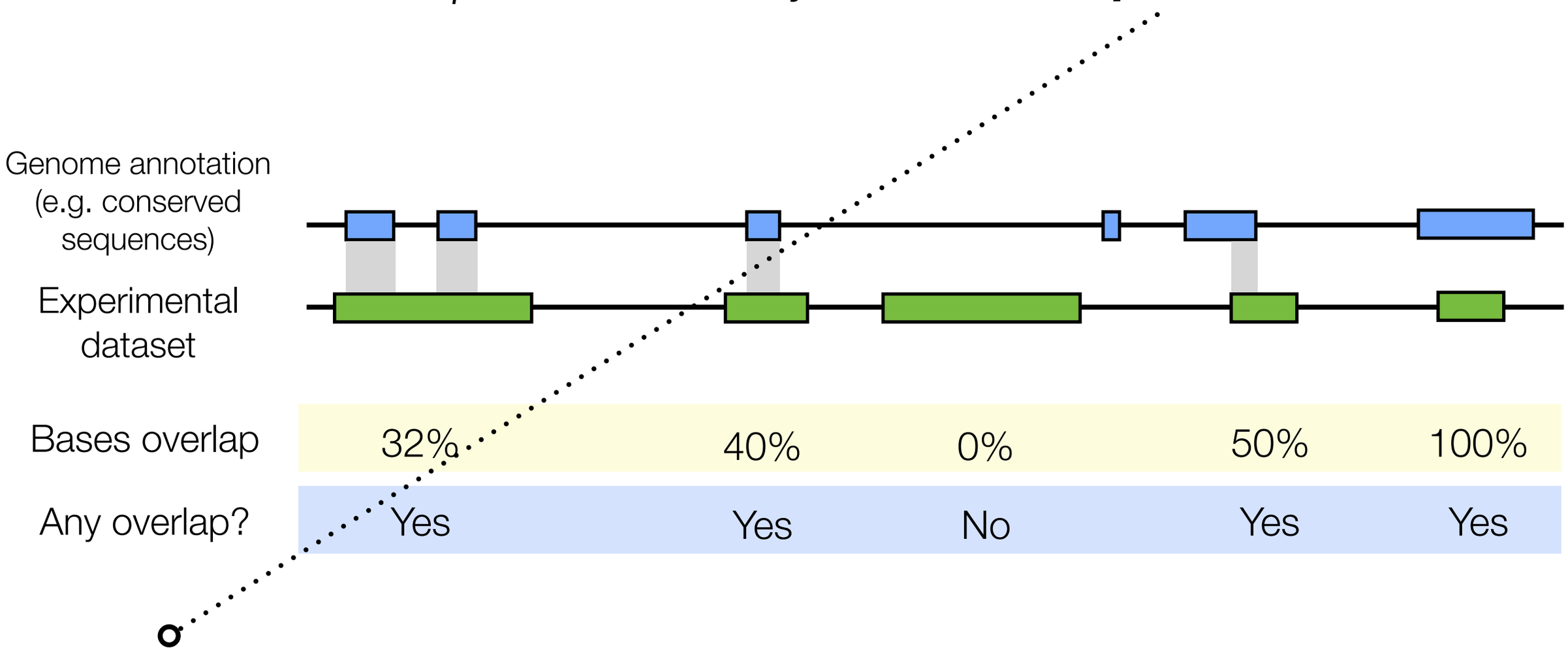


Bases overlap	32%	40%	0%	50%	100%
Any overlap?	Yes	Yes	No	Yes	Yes

Adapted from slide from Ben Brown

How do we detect genomic co-association?

That is, do two genomic features co-occur (overlap or have spatial consistency) more than ***expected****?*



How do we develop a proper null expectation in order to reduce the dimensionality of the data to an informative statistic?

Adapted from slide from Ben Brown

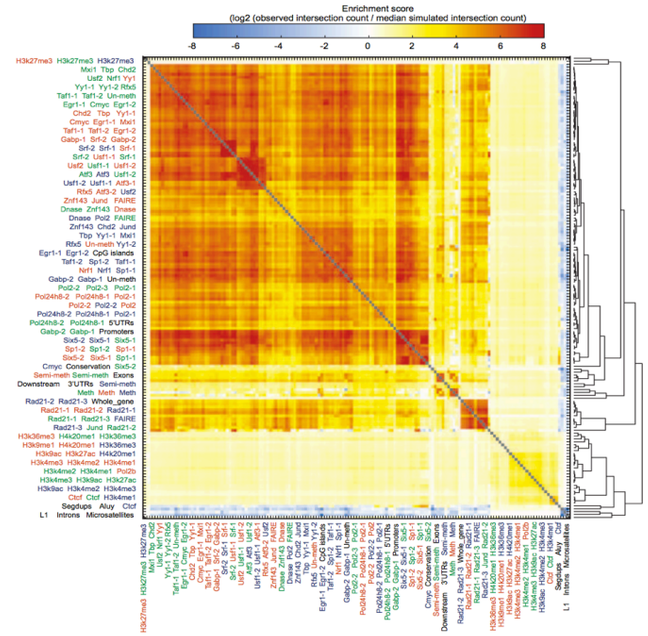
How do we develop a proper null expectation?

1. Monte-Carlo simulation

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3*]

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA



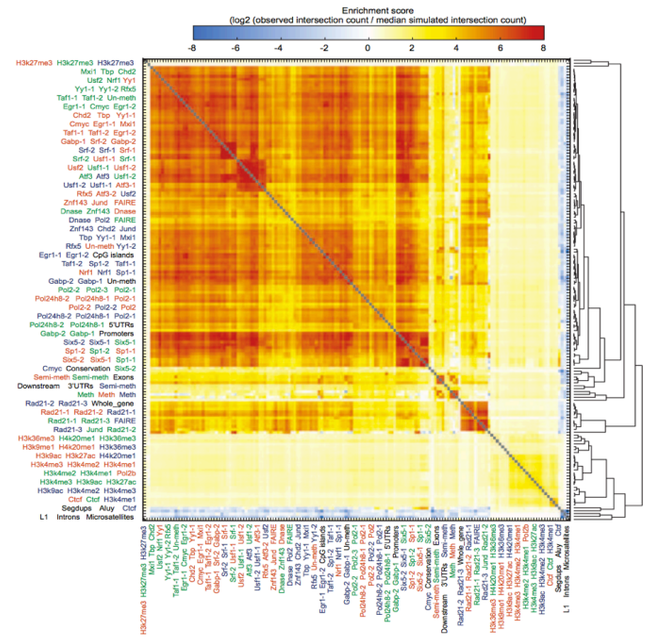
How do we develop a proper null expectation?

1. Monte-Carlo simulation

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA



2. Block bootstrap sampling (Peter Bickel)

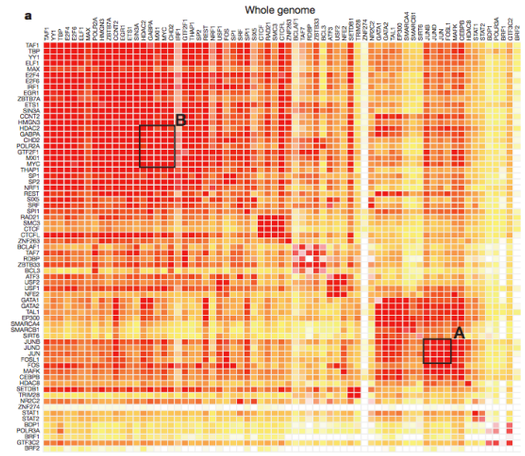
ARTICLE
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

NON PARAMETRIC METHODS FOR GENOMIC INFERENCE

By Peter J. Bickel*,†, Nathan Boley*,†, James B. Brown*,†, Haiyan Huang*,†, Nancy R. Zhang*,‡



How do we develop a proper null expectation?

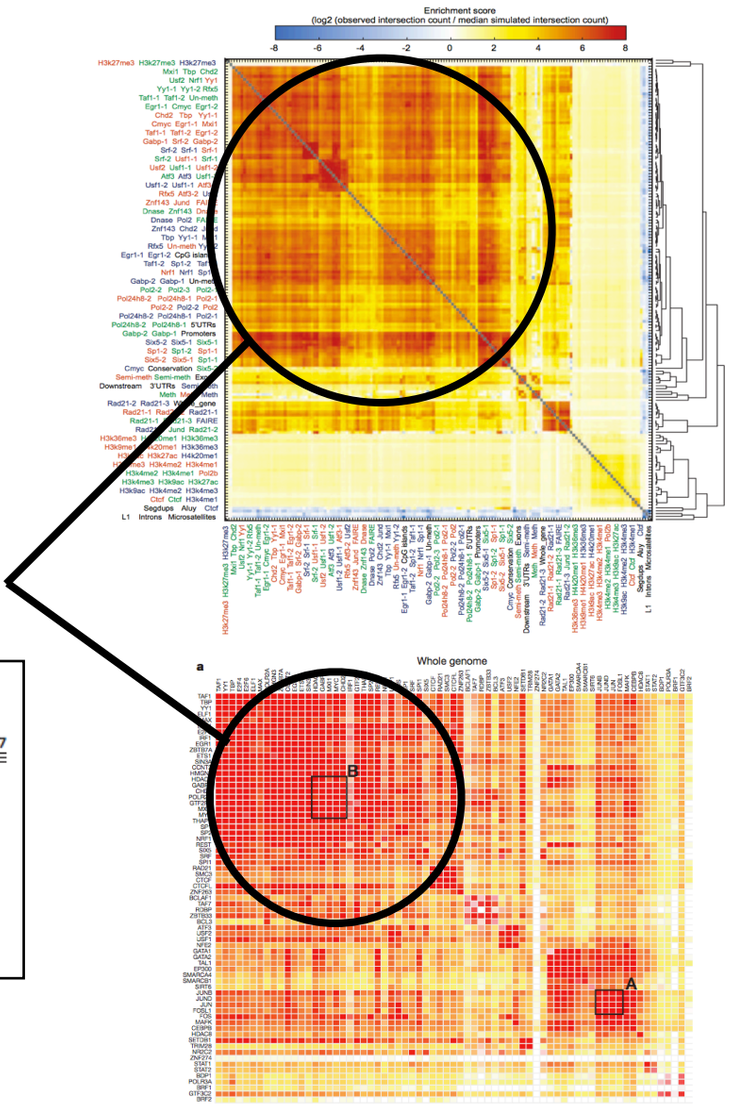
1. Monte-Carlo simulation

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

So called TF "HOT regions"
Gerstein et al; modENCODE; 2010



2. Block bootstrap sampling (Peter Bickel)

ARTICLE
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

NON PARAMETRIC METHODS FOR GENOMIC INFERENCE

By Peter J. Bickel*,†, Nathan Boley*,†, James B. Brown*,†, Haiyan Huang*,†, Nancy R. Zhang*,‡

How do we develop a proper null expectation?

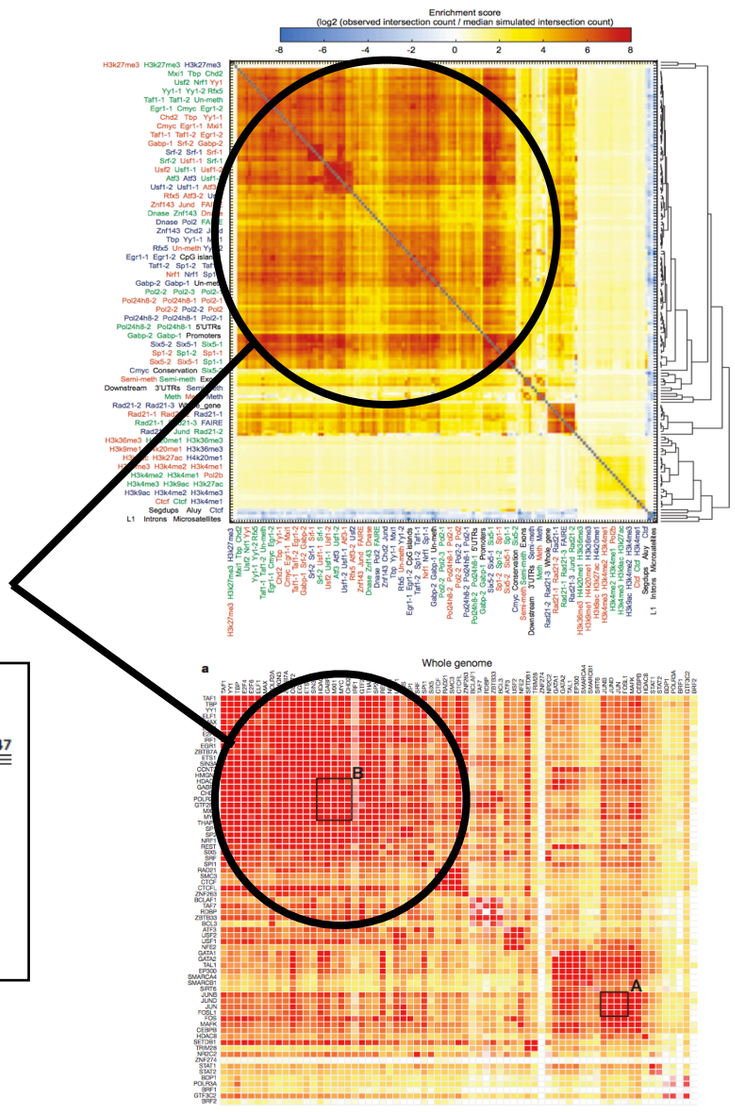
1. Monte-Carlo simulation

Binary Interval Search (BITS):
A Scalable Algorithm for Counting Interval Intersections

Ryan M. Layer[1], Kevin Skadron [1], Gabriel Robins[1], Ira M. Hall[2], and Aaron R. Quinlan[3]*

[1]Department of Computer Science, University of Virginia, Charlottesville, VA
[2]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[3]Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville, VA

So called TF "HOT regions"
Gerstein et al; modENCODE; 2010



2. Block bootstrap sampling (Peter Bickel)

ARTICLE
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

NON PARAMETRIC METHODS FOR GENOMIC INFERENCE

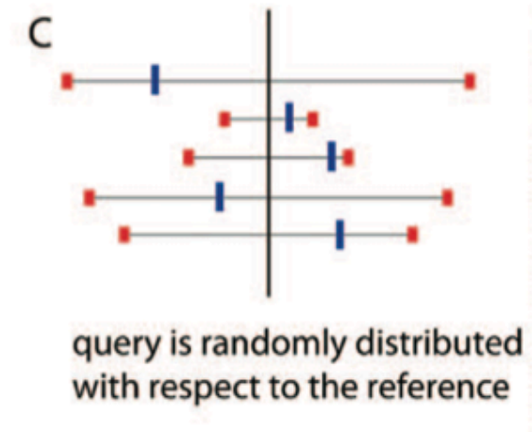
BY PETER J. BICKEL*,†, NATHAN BOLEY*,†, JAMES B. BROWN*,†, HAIYAN HUANG*,†, NANCY R. ZHANG*,‡

3. Spatial methods (Sarah Wheelan)

Exploring Massive, Genome Scale Datasets with the GenometriCorr Package

Alexander Favorov[1,2,3]*, Loris Mularoni[1]¤a, Leslie M. Cope[1], Yulia Medvedeva[2,3]¤b, Andrey A. Mironov[4,5], Vsevolod J. Makeev[2,3], Sarah J. Wheelan[1]*

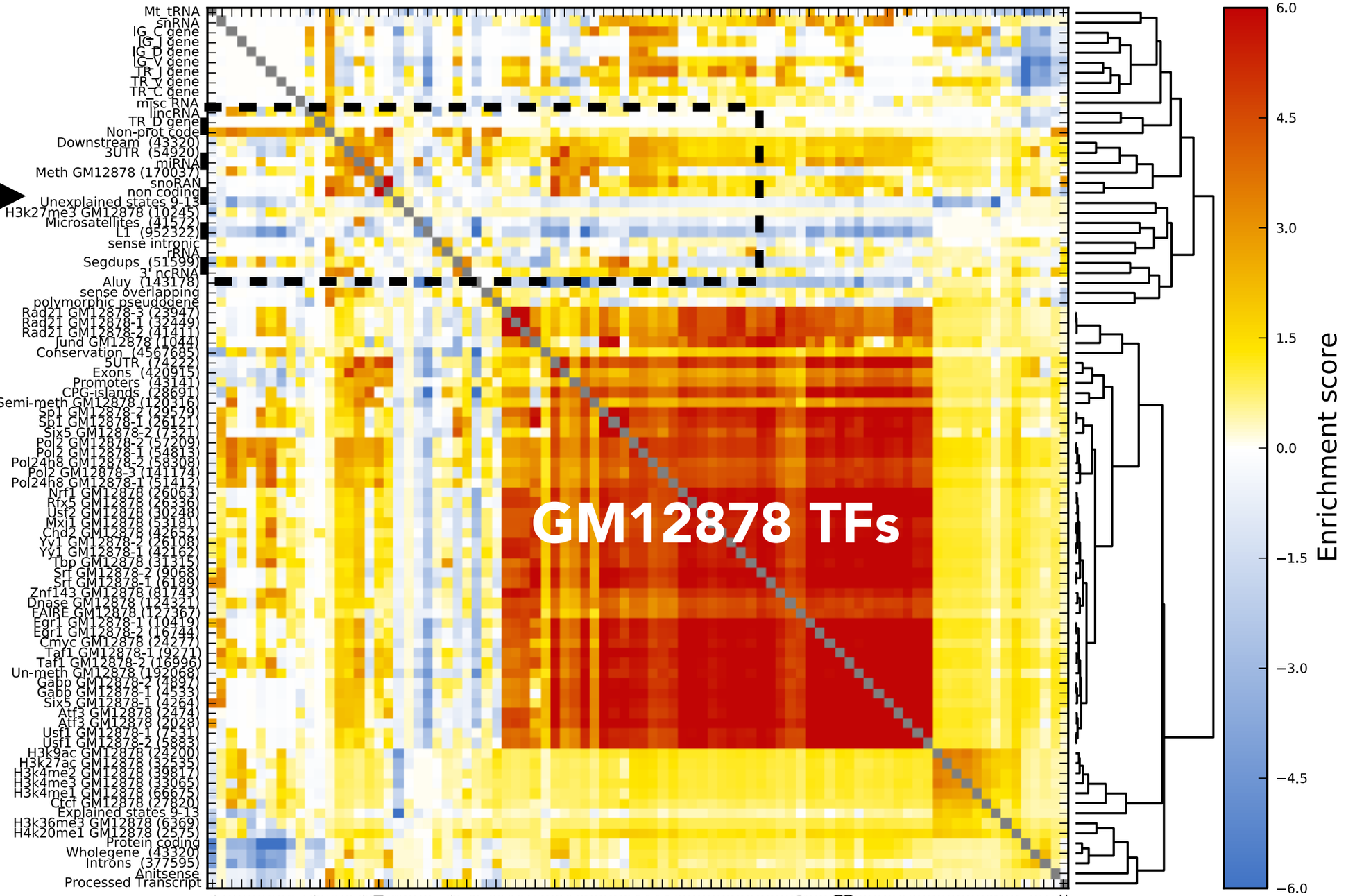
1 Department of Oncology, Division of Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **2** Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia, **3** Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia, **4** Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia, **5** Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia



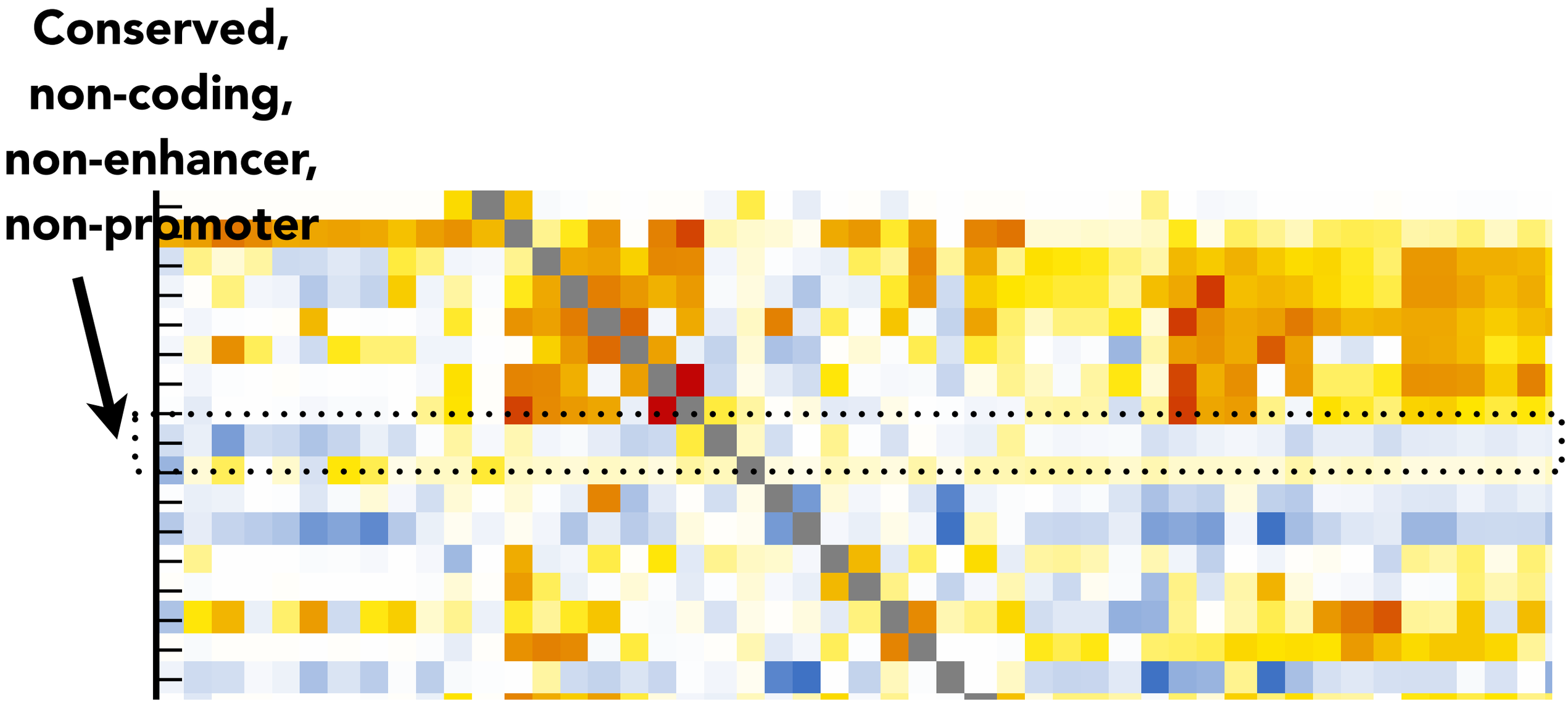
Integrating these and other, novel association statistics into ***bedtools2***

What is the role of conserved, non-coding (CNC) elements?

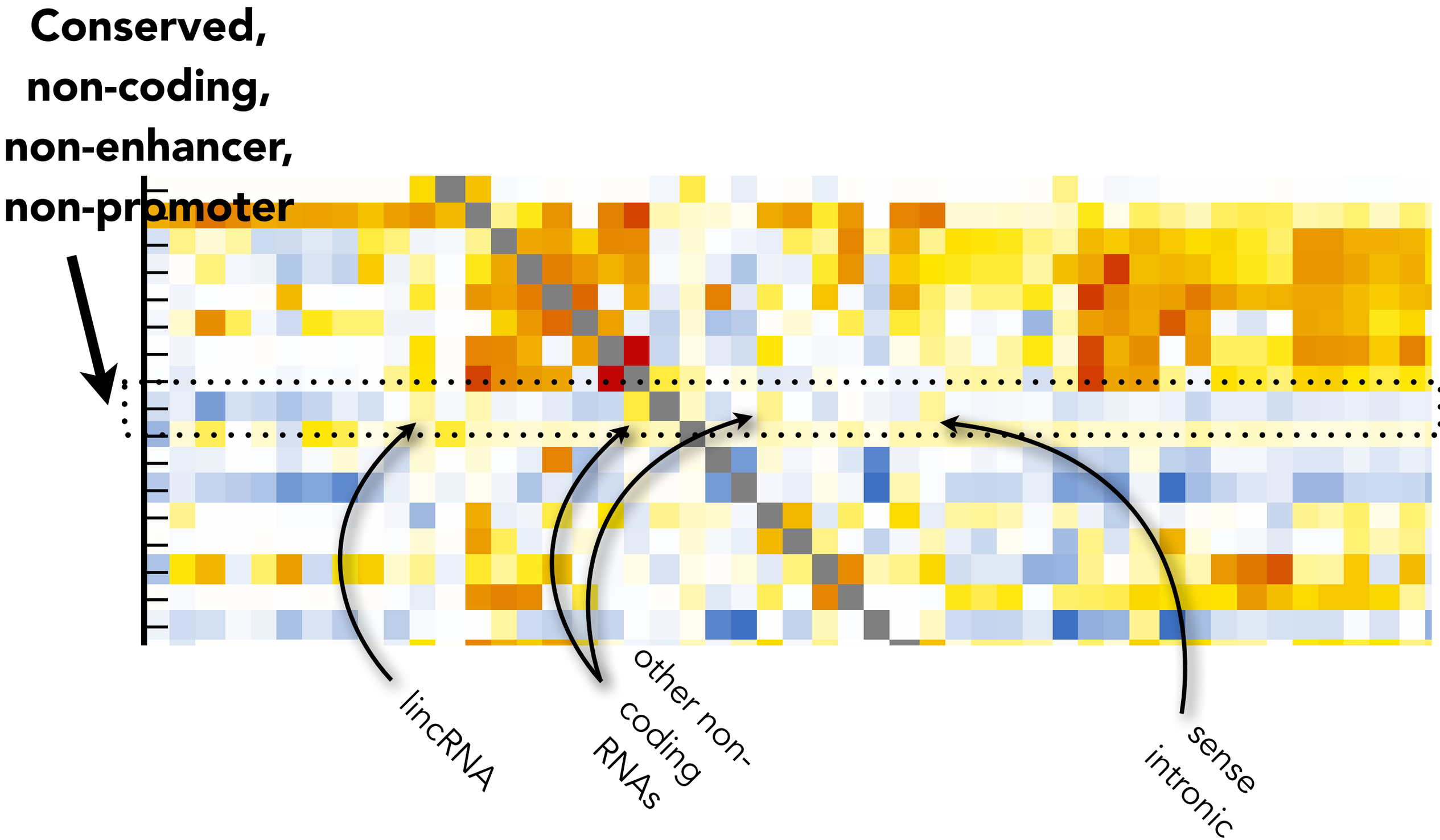
CNC elements



Making sense of conserved, yet non-coding, and non-"functional" elements



Making sense of conserved, yet non-coding, and non-"functional" elements



Genome Query Language (GQL)

(coming in 2013)

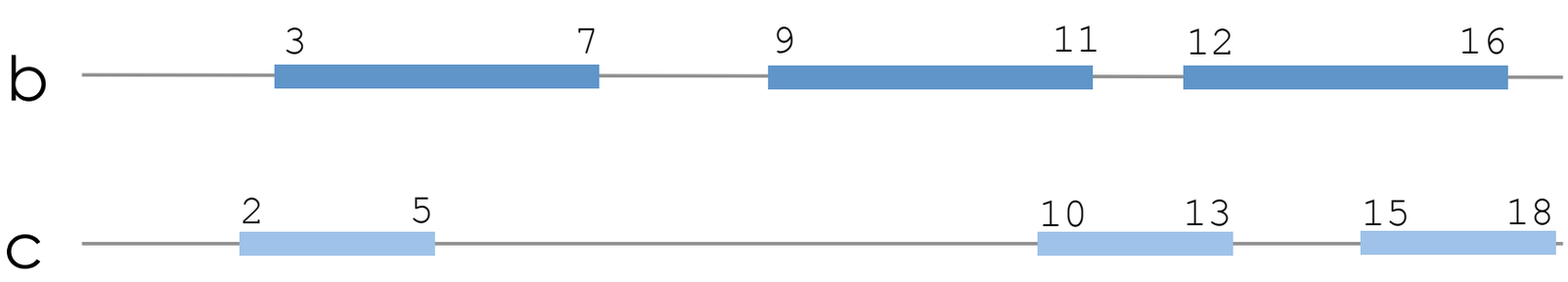
- Writing a new, simple, & expressive language for exploring complex datasets. Based on SQL.
- Set theory, statistics, visualization, data recruitment
- Abstract the language from the engine underneath
- Powerful, fast, and easy to use.
- Built upon bedtools and bedtools2

A GQL example

(find all intervals where at least 2 datasets overlap)

(e.g., find “hot regions” for TF ChIP-seq peaks)


```
chr1  1    4    a1   1    +
chr1  6    8    a2   2    +
chr1  14   17   a3   3    +
```




```
chr1  3    7    b1   1    -
chr1  9    11   b2   2    +
chr1  12   16   b3   3    +
```

```
chr1  2    5    c1   1    +
chr1  10   13   c2   2    +
chr1  15   18   c3   3    +
```

```
> result = SELECT a,b,c WHERE COUNT(>1);
> PRINT result
> PLOT result
> MEASURE result WITH(MONTE_CARLO) AND a,b,c
```

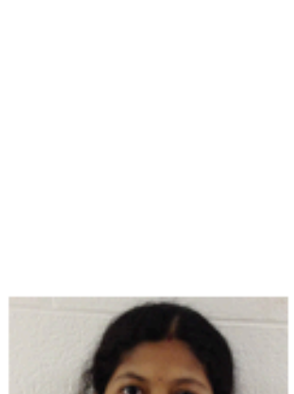
Quinlan lab

Uma Paila, Ph.D.

Postdoctoral Research Associate

udp3f @ virginia.edu

Research Projects and Interests: Investigation of the genetic basis of extreme sensitivity to ionizing radiation; development of new analytical tools for exploring genetic variation identified through next-generation sequencing projects.



Neil Kindlon, M.S.

Staff Scientist and Software Engineer

nek3d @ virginia.edu

Research Projects and Interests: Software development for genomic analysis. Structural variation discovery and interpretation using DNA sequencing technologies.

Ryan Layer

Graduate student

rl6sf @ virginia.edu

Research Interests: Scalable algorithm development for high-throughput genomic analysis; genome data mining and analysis; structural variation discovery and interpretation.

John Kubinksi
Undergraduate
(Biology)

Gift Sinthong
Undergraduate
(Comp. Science)

Acknowledgements

Ira Hall *Univ. of Virginia*

Ankit Malhotra *Univ. of Virginia*

Michael Lindberg *Univ. of Virginia*

Royden Clark *Univ. of Virginia*

Svetlana Sokolova *Univ. of Virginia*

Mitchell Leibowitz *Univ. of Virginia*

Nik Krumm *Univ. of Washington*

Evan Eichler *Univ. of Washington*

Debbie Nickerson *Univ. of Washington*

Chris Carlson *Fred Hutchinson CRC*

Mark Rieder *Univ. of Washington*

Josh Smith *Univ. of Washington*

Peter Sudmant *Univ. of Washington*

Pat Concannon Univ. of Virginia

Steve Rich Univ. of Virginia

Suna Onengut-Gumuscu Univ. of Virginia

Chris Moskaluk Univ. of Virginia

Shu-Man Fu Univ. of Virginia

Gabor Marth Boston College

Jim Robinson Broad Institute

Nick Navin MD Anderson

Kristin Baldwin Scripps

Funding

NHGRI: R01 HG006693-01

NIEHS: R21 ES020521-01

UVA Fund for Excellence in Science and Tech.

UVA Cancer Center Pilot Program