Detection and characterization of complex rearrangements in tumor genomes

Aaron Quinlan

quinlanlab.org

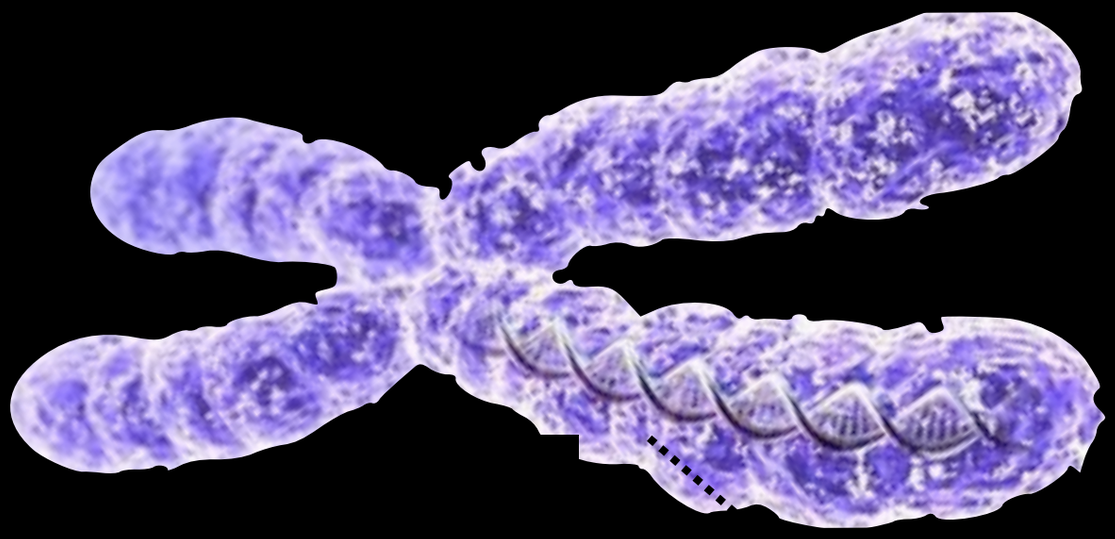
BioConductor 2013, Seattle WA, July 18, 2013

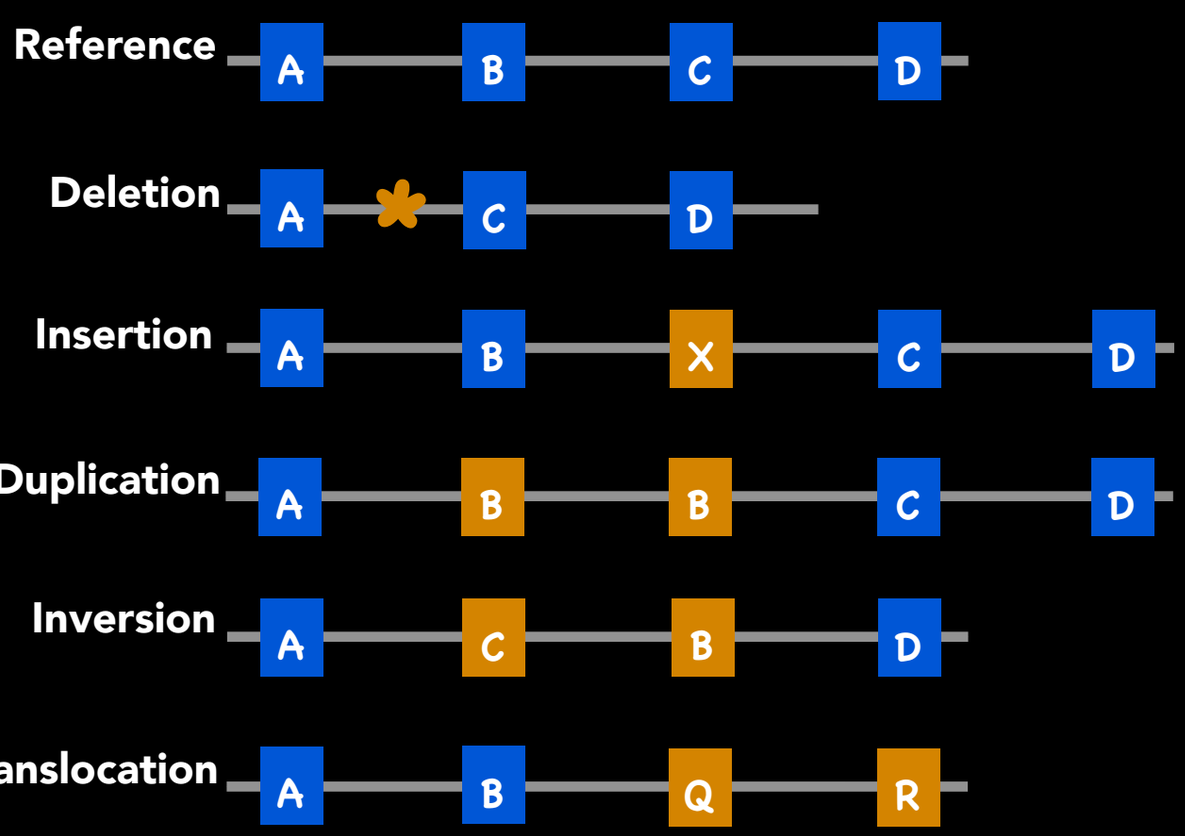
University of Virginia, Charlottesville VA

Center for Public Health Genomics

Biochemistry and Molecular Genetics

SV definitions





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

genomic rearrangement: ditto

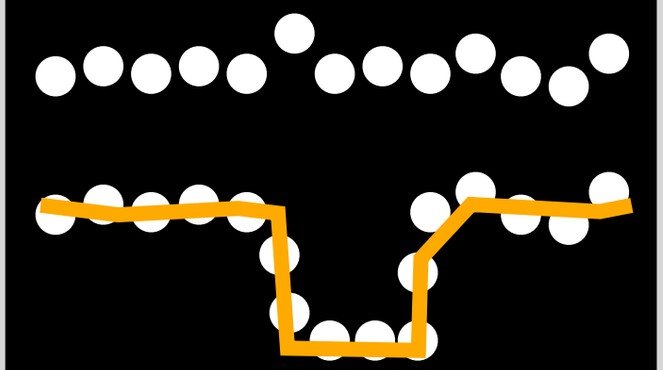
copy number variant (CNV), or alteration (CNA): an SV that alters DNA copy number

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

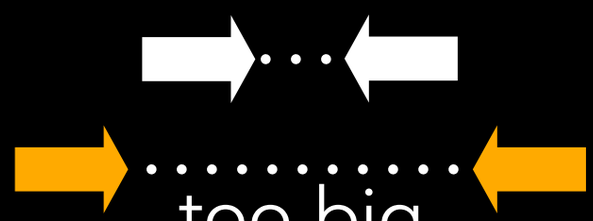
complex SV: 2 or more breakpoints that arise through a single mutational event, but cannot be explained by one DNA exchange or end-joining reaction

"Signals" for SV discovery

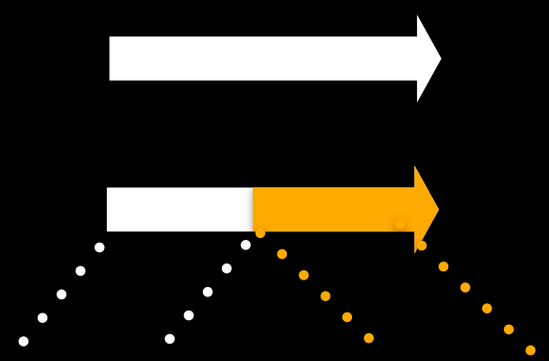
Depth of coverage



Paired-end mapping



Split-read mapping



1. Prior knowledge
2. New signals
   (e.g. positional seq.)
3. Known SV sites
4. Predictions from other tools

Most existing SV tools exploit just one signal

SV discovery is fraught with a high false negative rate.

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

SV discovery is fraught with a high false negative rate.

- 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

SV discovery is fraught with a high false negative rate.

- 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 false positives also eliminates true positives.

SV discovery is fraught with a high false negative rate.

- 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 false positives also eliminates true positives.
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)

SV discovery is fraught with a high false negative rate.

- 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 false positives also eliminates true positives.
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)
- When searching for somatic mutation in a tumor/normal comparison, a *false negative call in the normal* can cause a *false positive somatic call in the tumor*.

SV discovery is fraught with a high false negative rate.

- 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 false positives also eliminates true positives.
- The false negative rate is usually hard to measure, but is thought to be extremely high for most PEM studies (>30%)
- When searching for somatic mutation in a tumor/normal comparison, a *false negative call in the normal* can cause a *false positive somatic call in the tumor*.
- False negatives are very problematic in the context of tumor heterogeneity

lumpy

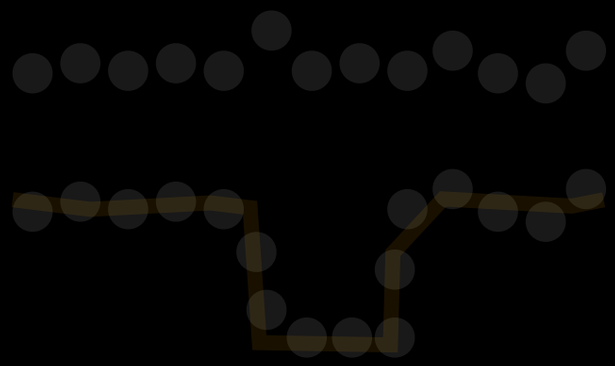
A probabilistic framework that integrates multiple alignment “signals” for SV discovery.

Improved sensitivity.

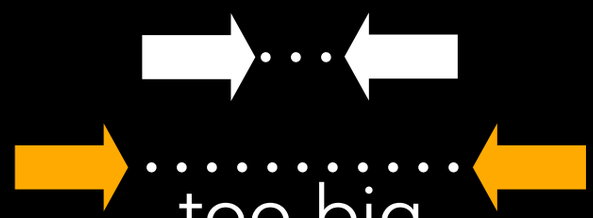
under review

DELLY: Rausch et al, 2012

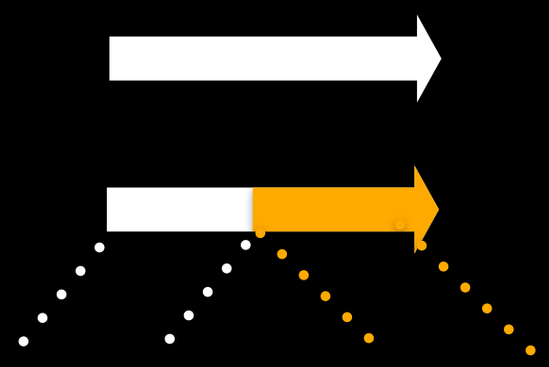
Depth of coverage



Paired-end mapping



Split-read mapping



1. *Predict*

2. *Refine*

"Stepwise"

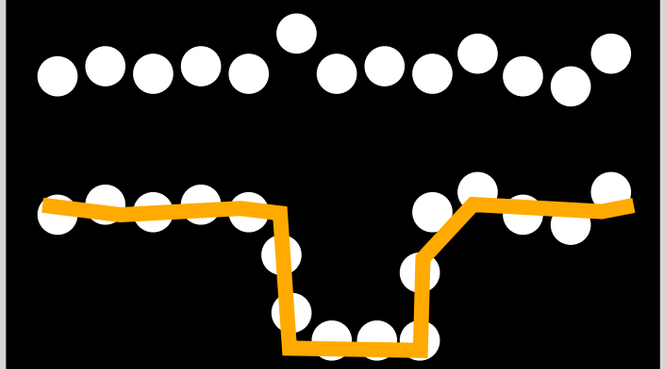
Prior knowledge

Known SV sites

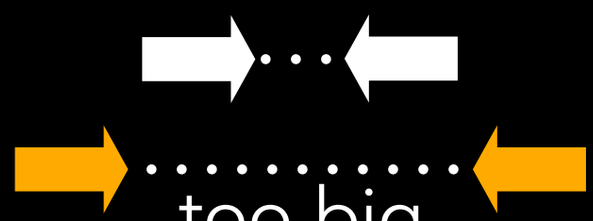
Predictions from other tools

GASVPro: Sindhi et al, 2012

Depth of coverage



Paired-end mapping

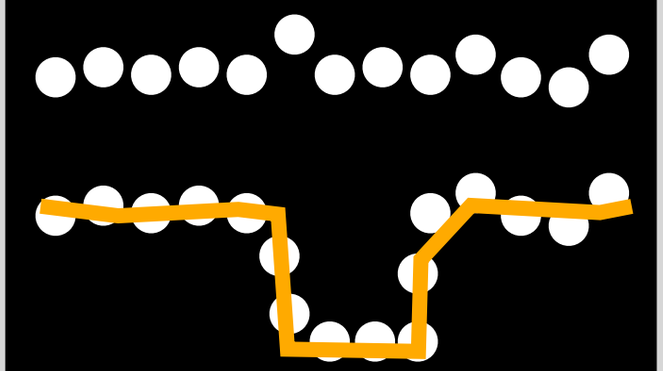


Combines DoC and PEM signals for greater specificity, especially for deletions (using DoC)

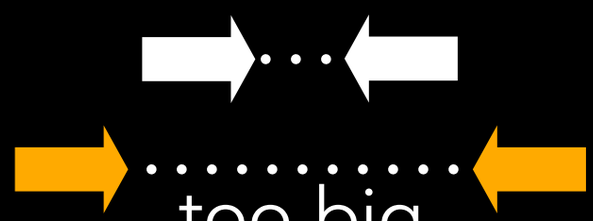
"Integrative"

LUMPY Layer et al, unpub.

Depth of coverage



Paired-end mapping

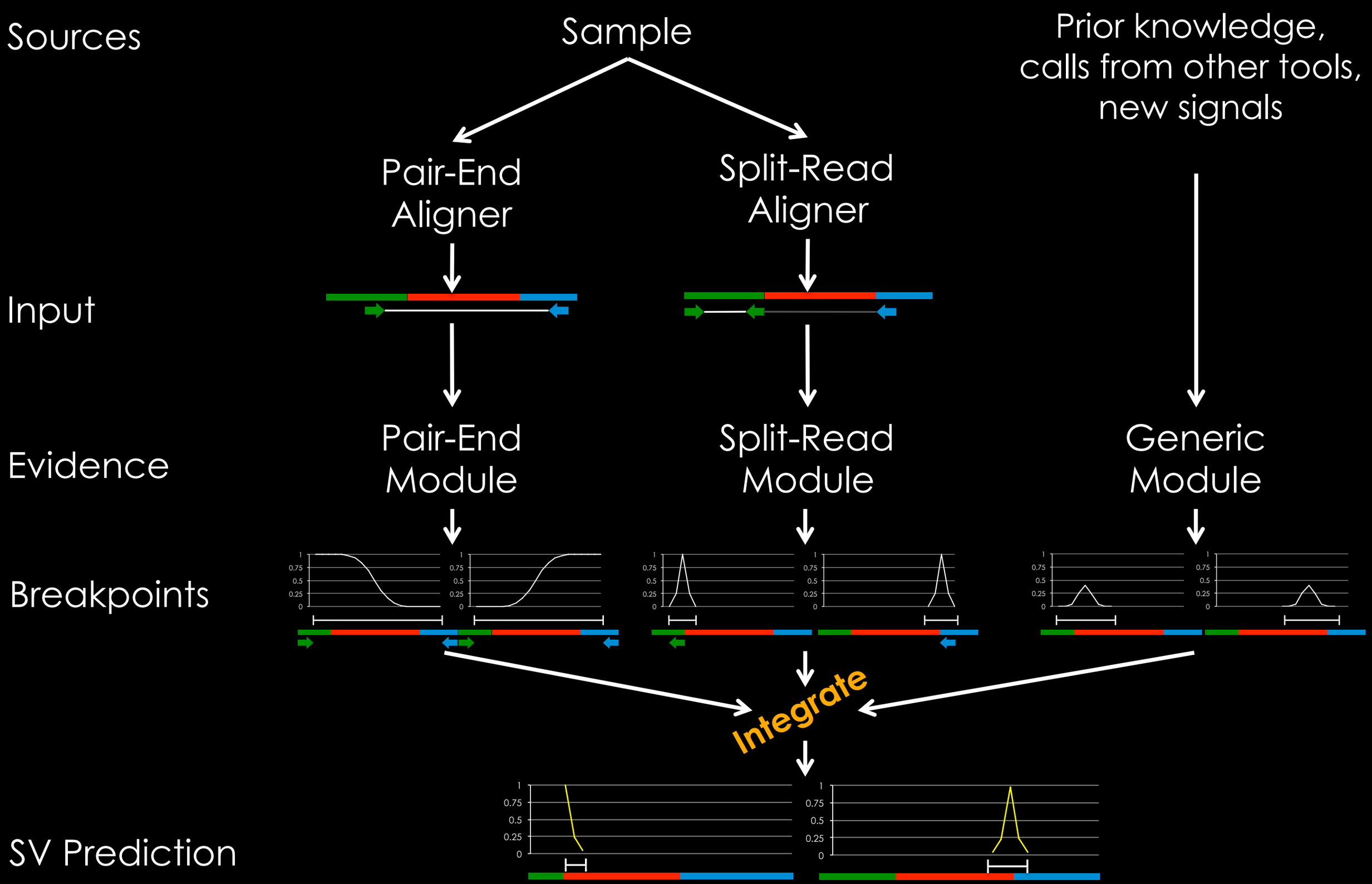


Split-read mapping

1. Prior knowledge
2. New signals
   (e.g. positional seq.)
3. Known SV sites
4. Predictions from other tools

LUMPY integrates all (and future) signals

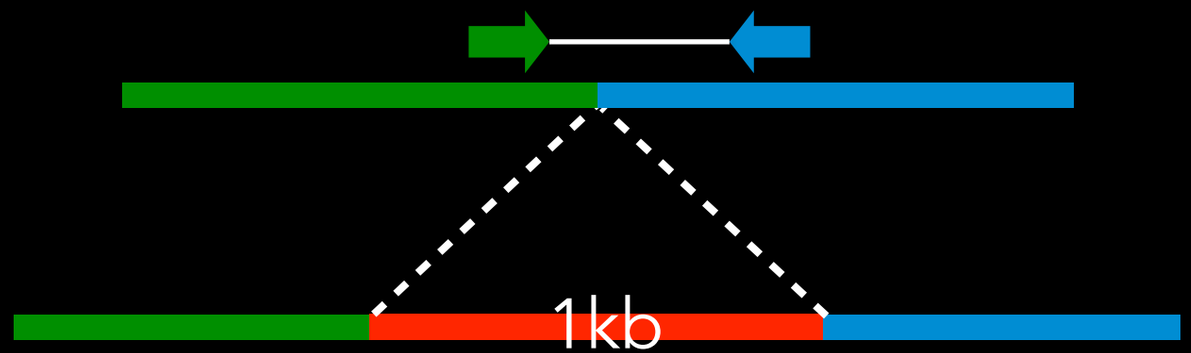
LUMPY integrates **all** SV signals



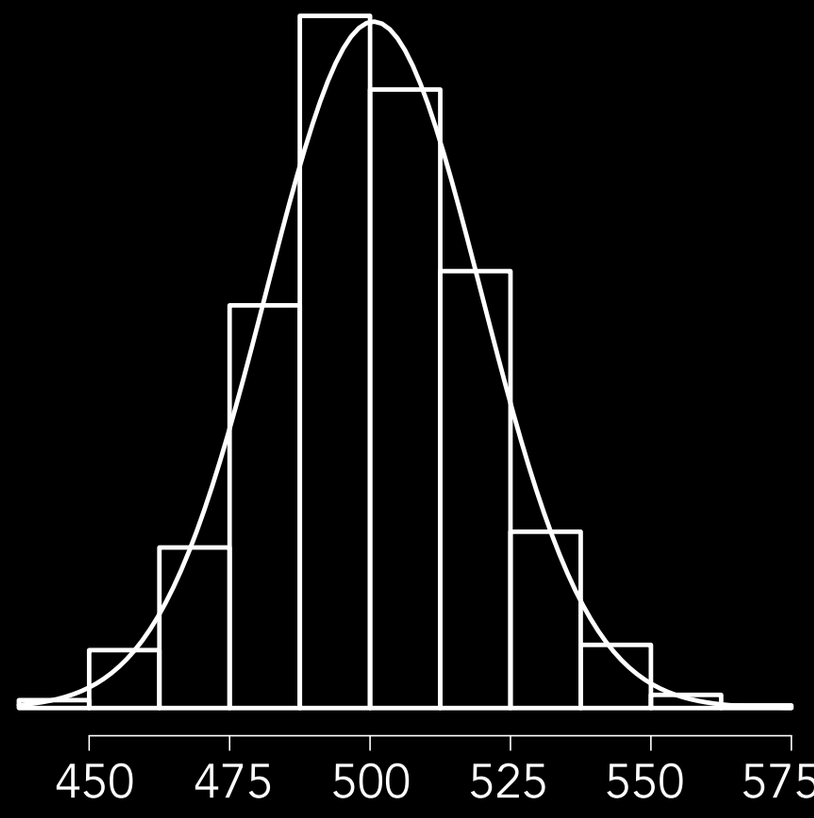
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



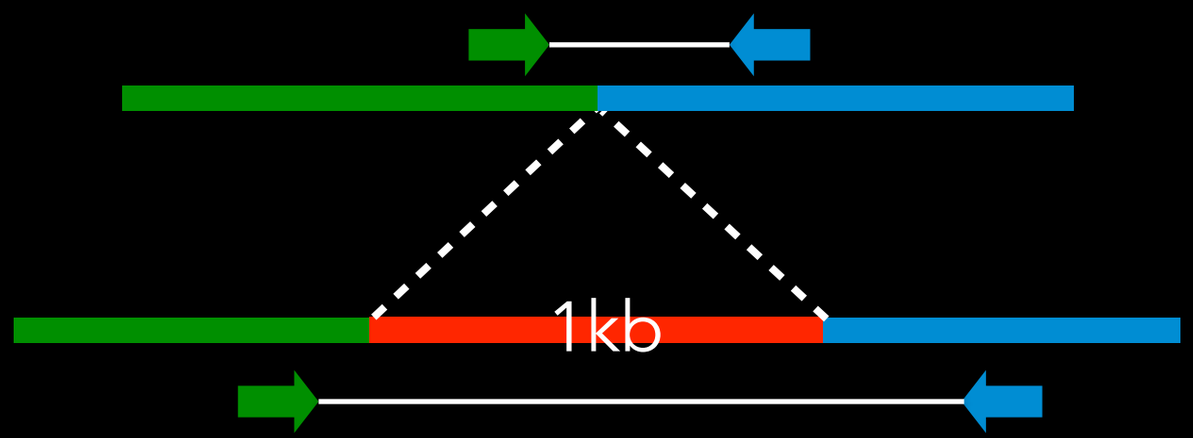
DNA library
fragment size distribution
(~500bp library)



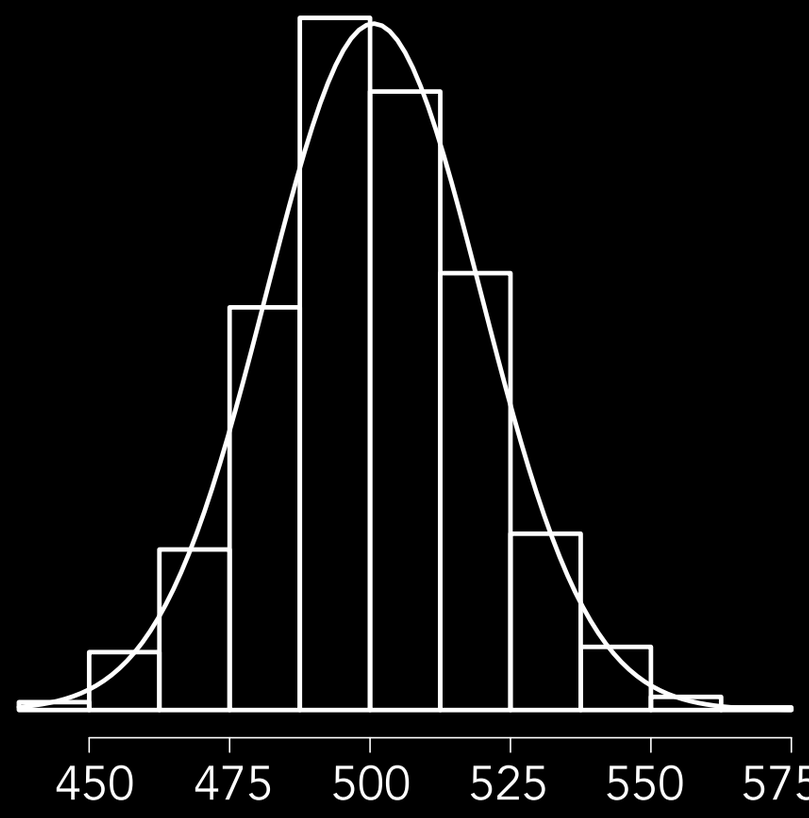
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



DNA library
fragment size distribution
(~500bp library)



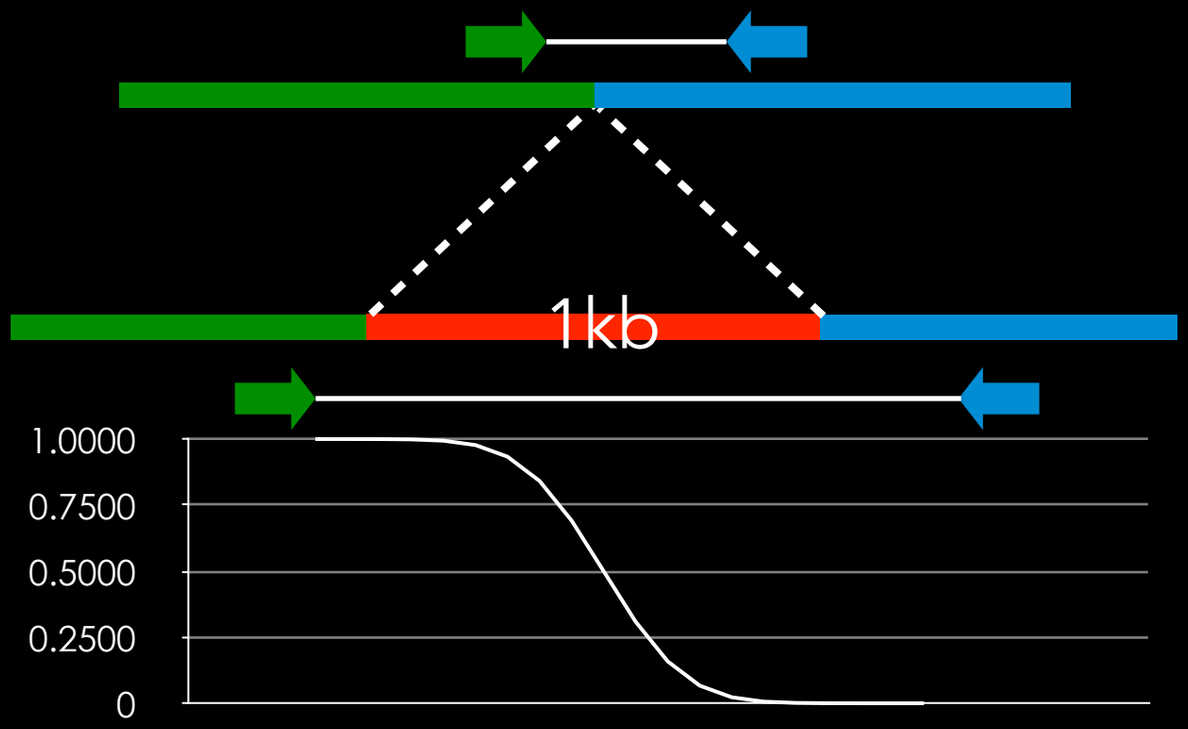
When aligned to reference, ends map ~1500bp apart.

Where are the breakpoints?

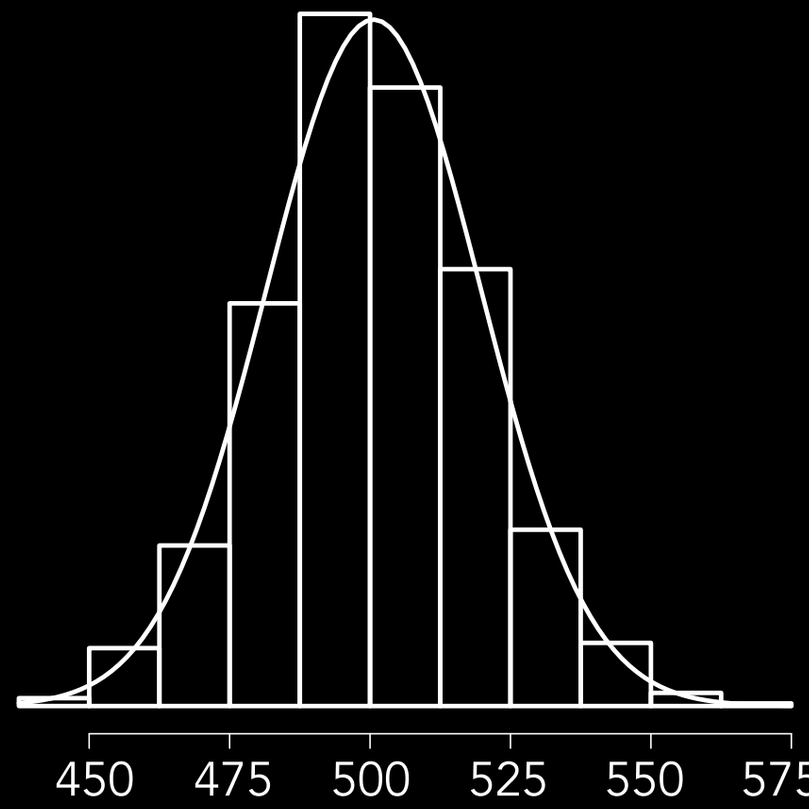
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



DNA library
fragment size distribution
(~500bp library)



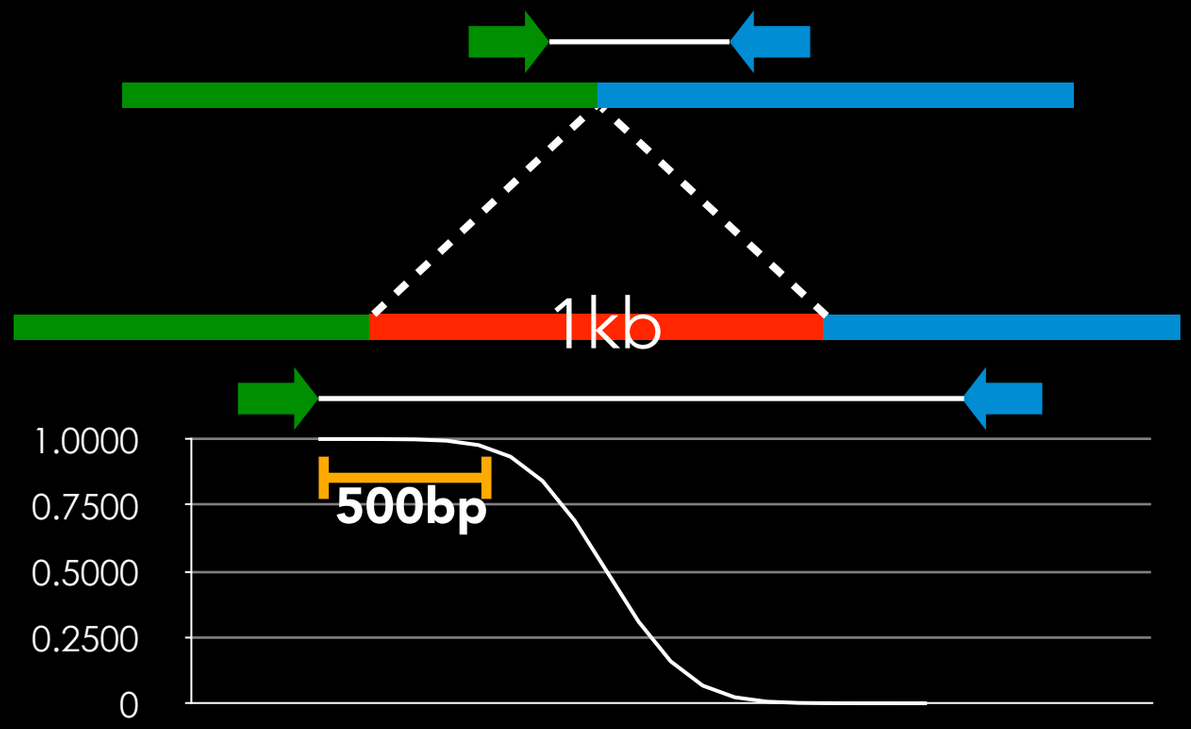
When aligned to reference, ends map ~1500bp apart.

Where are the breakpoints?

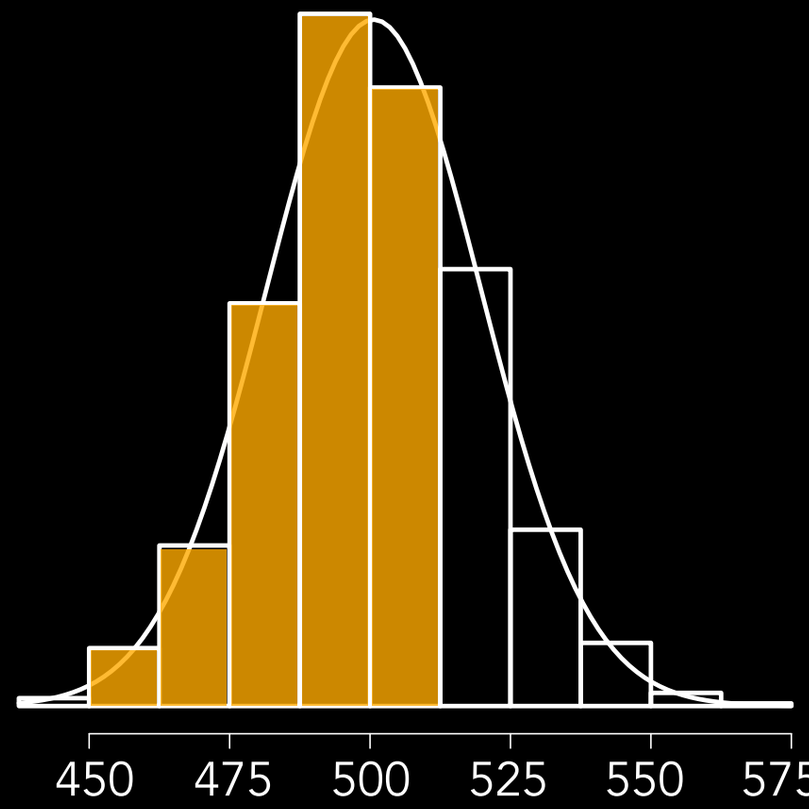
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



DNA library
fragment size distribution
(~500bp library)



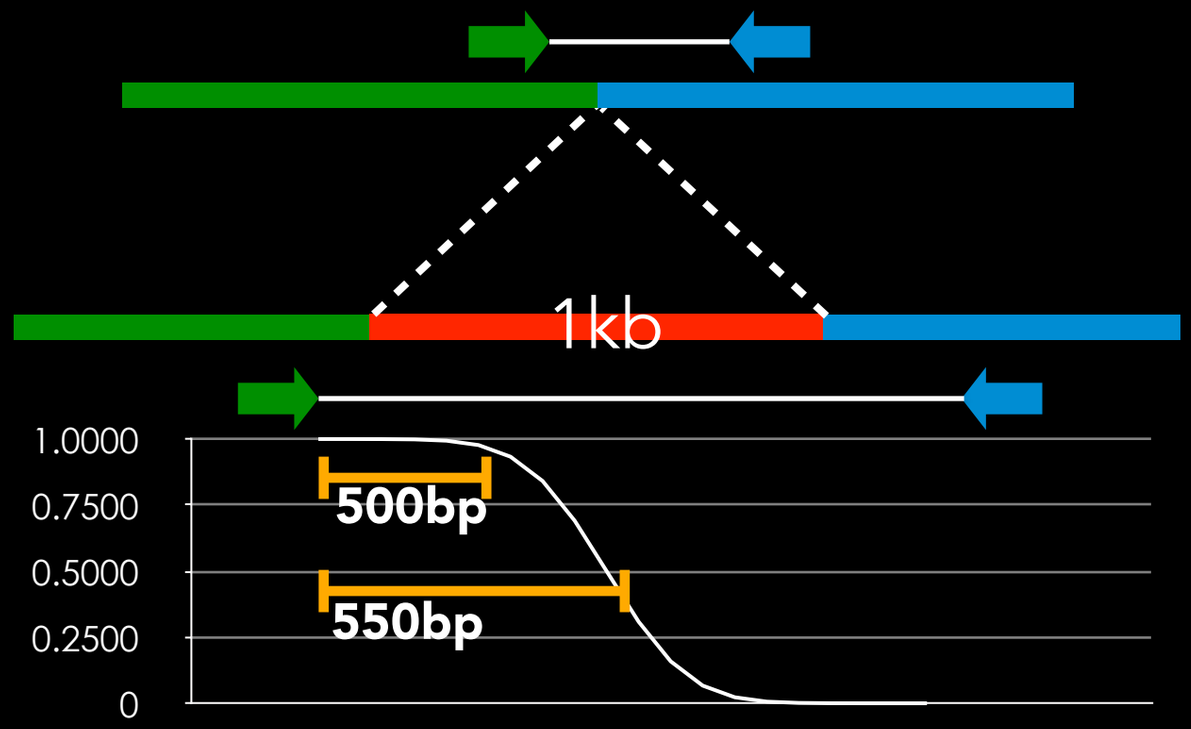
When aligned to reference, ends map ~1500bp apart.

Where are the breakpoints?

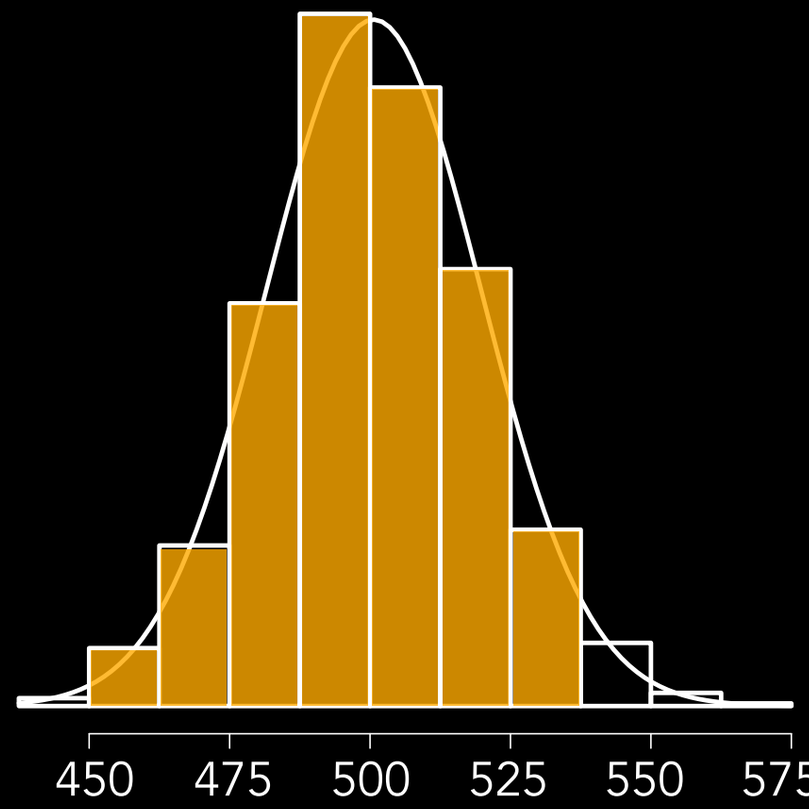
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



DNA library
fragment size distribution
(~500bp library)



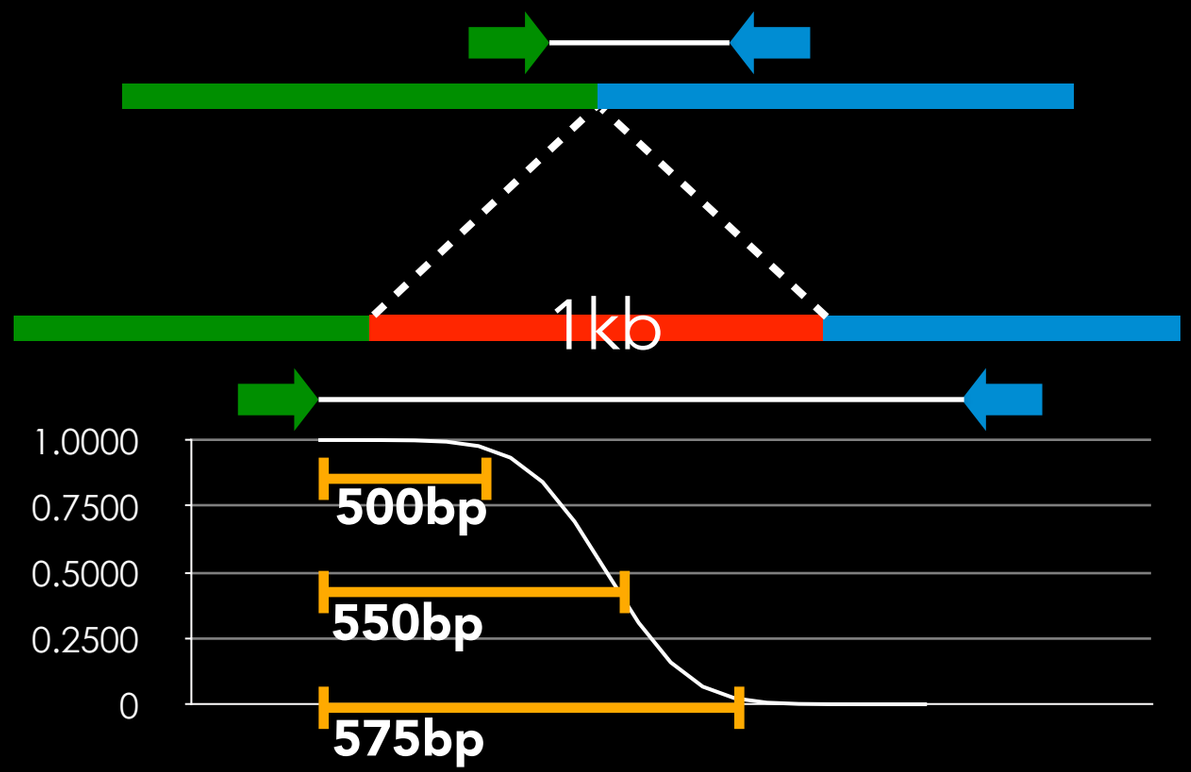
When aligned to reference, ends map ~1500bp apart.

Where are the breakpoints?

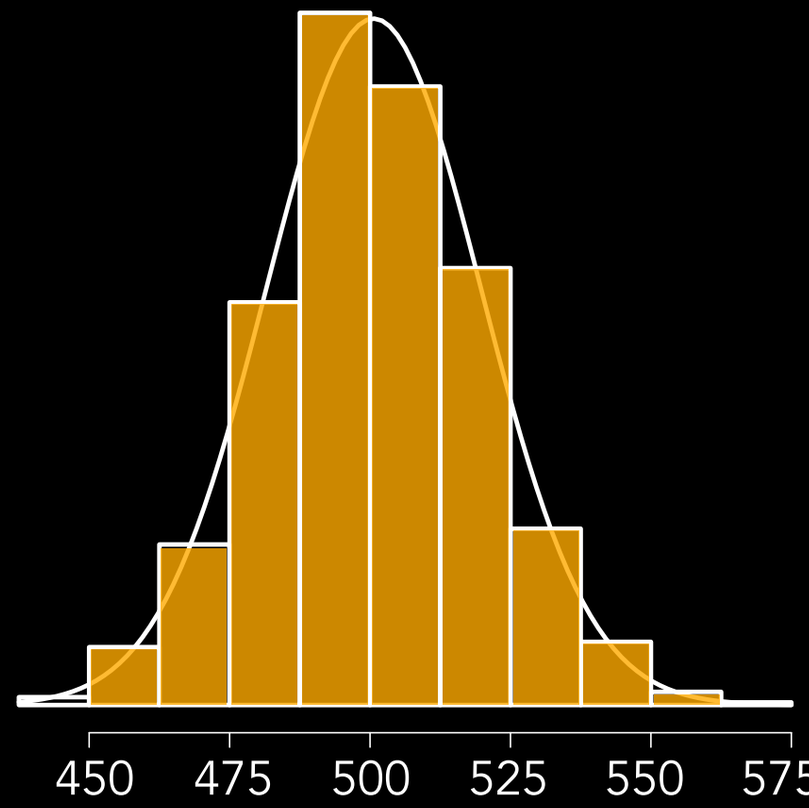
Paired-end library statistics inform SV breakpoint prediction

Sample genome

Reference genome



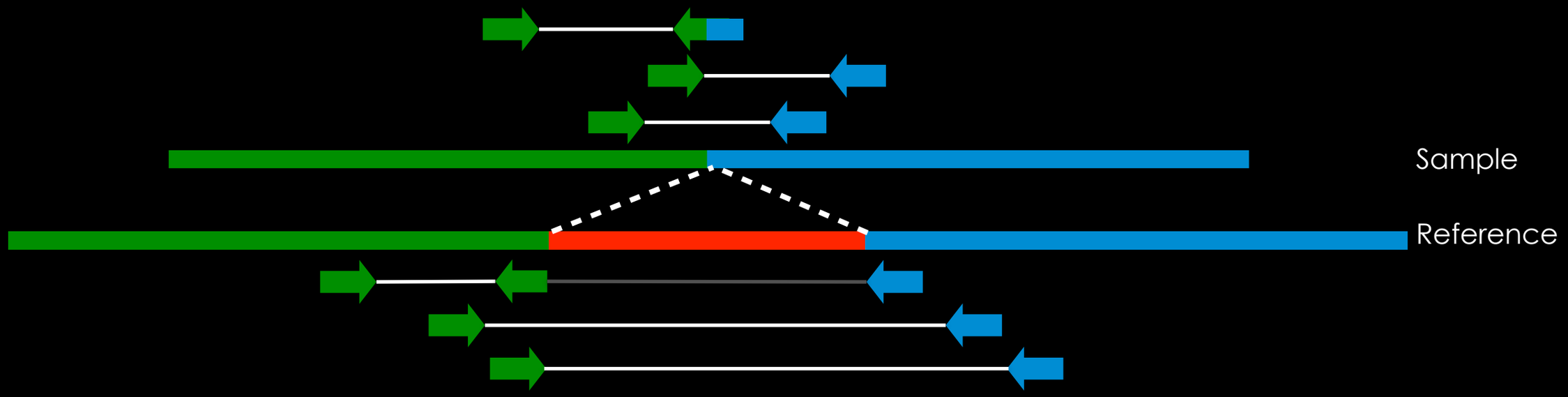
DNA library
fragment size distribution
(~500bp library)



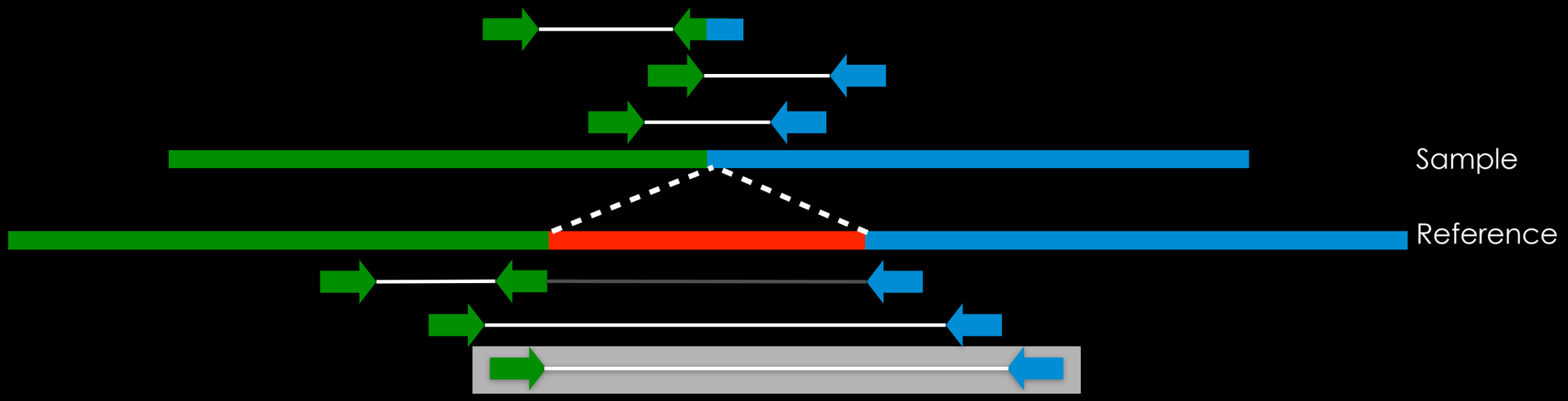
When aligned to reference, ends map ~1500bp apart.

Where are the breakpoints?

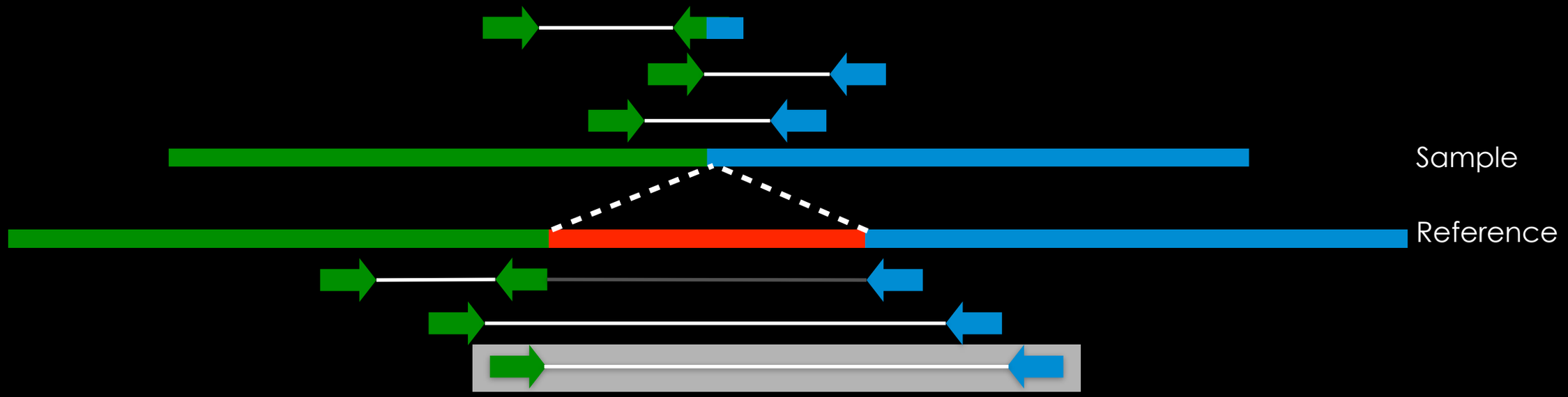
Combining SV signals

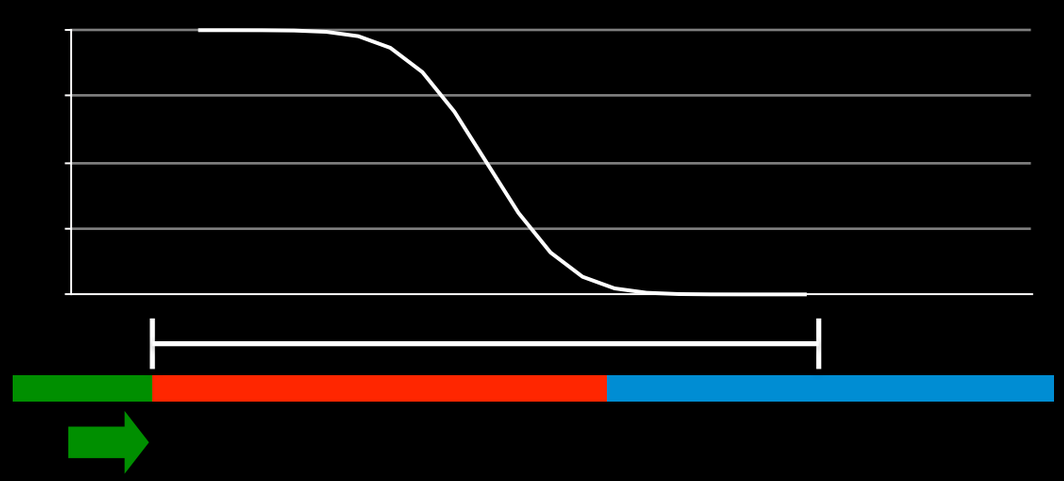


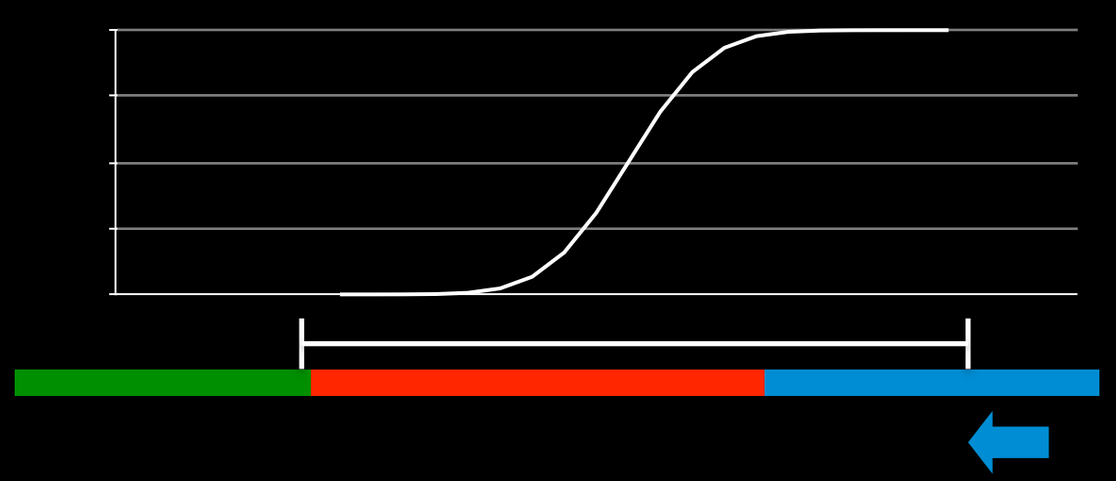
Combining SV signals



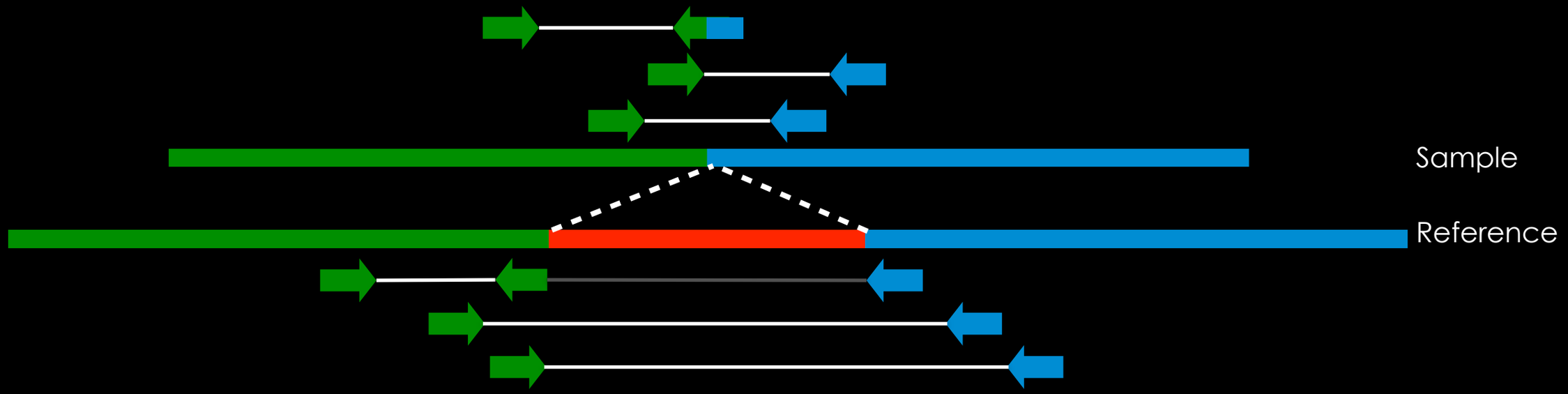
Combining SV signals

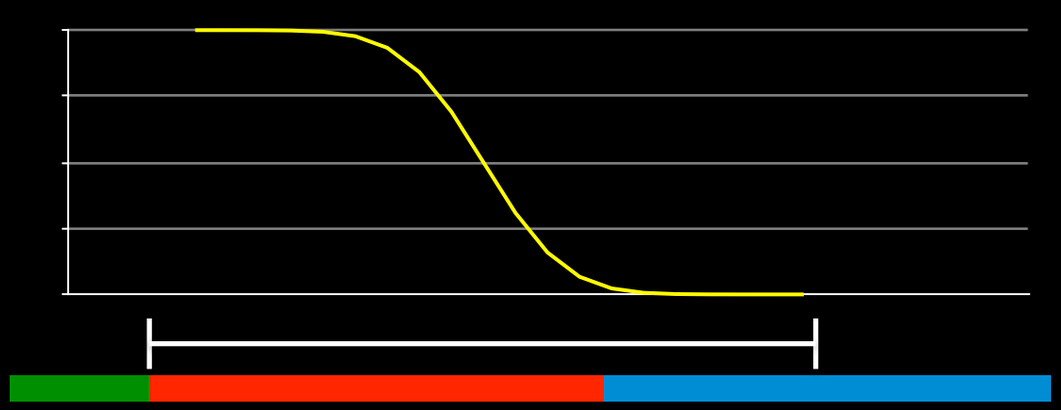


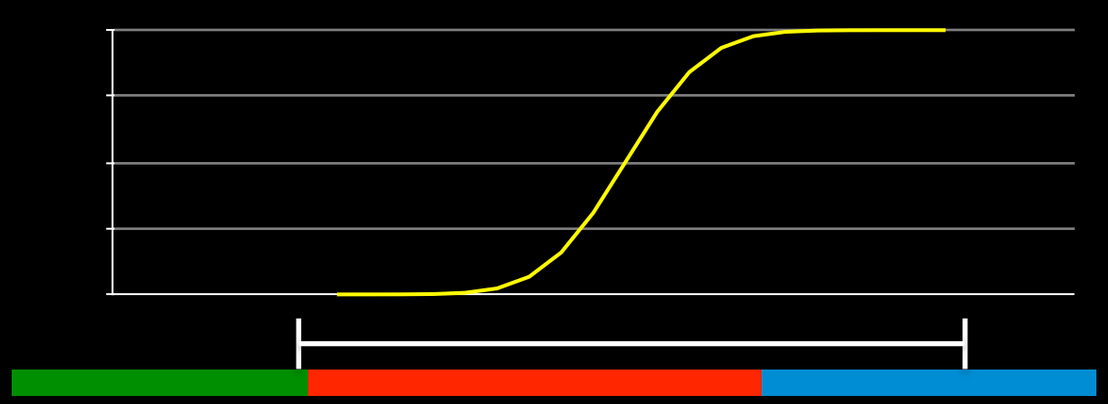




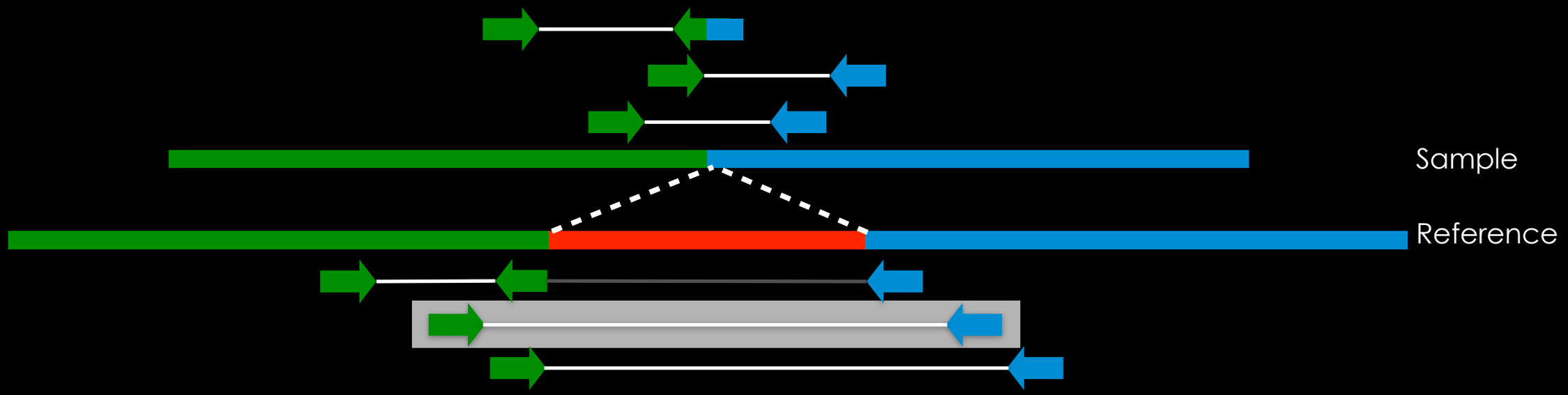
Combining SV signals

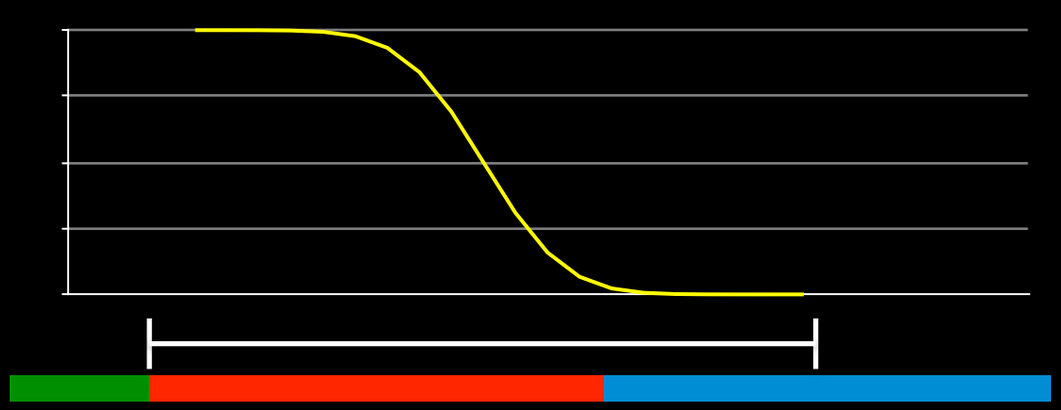


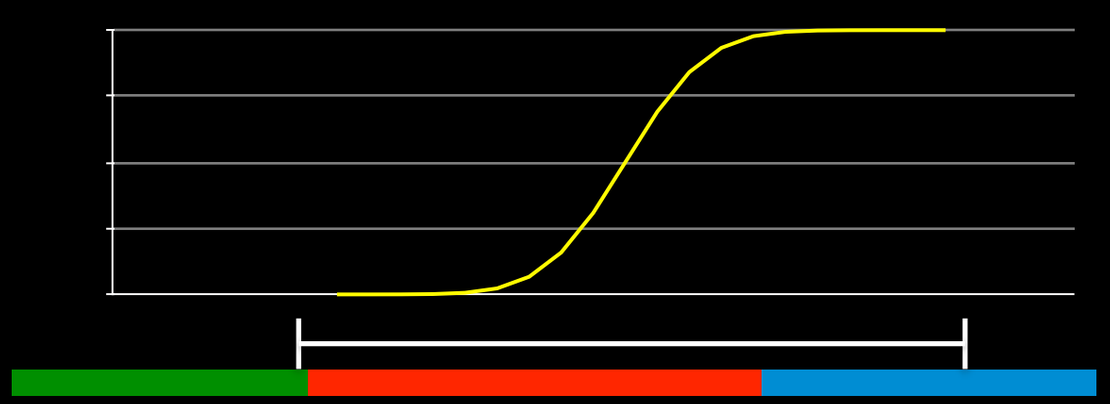




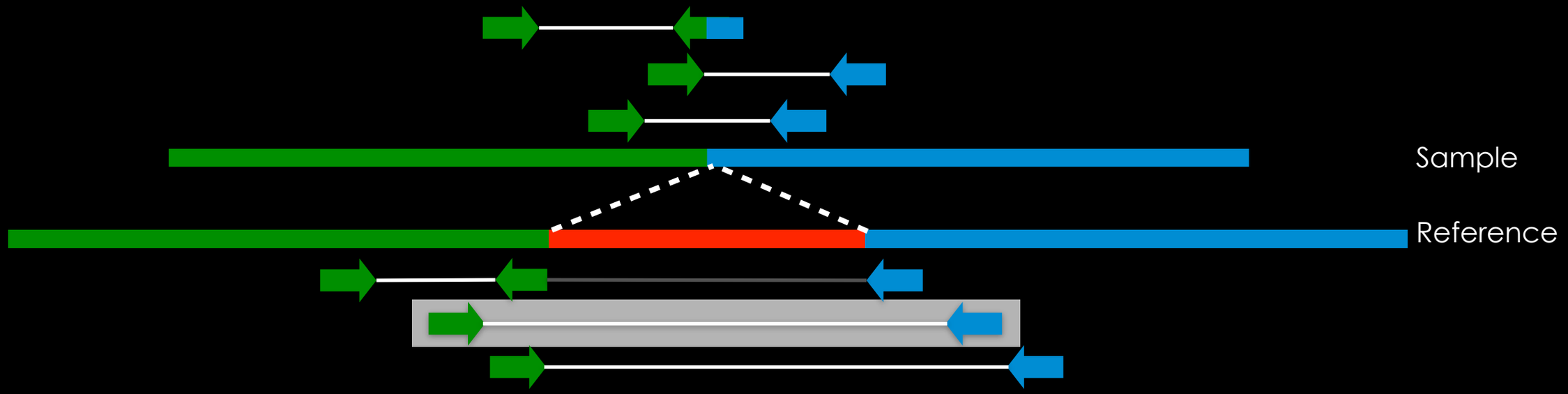
Combining SV signals

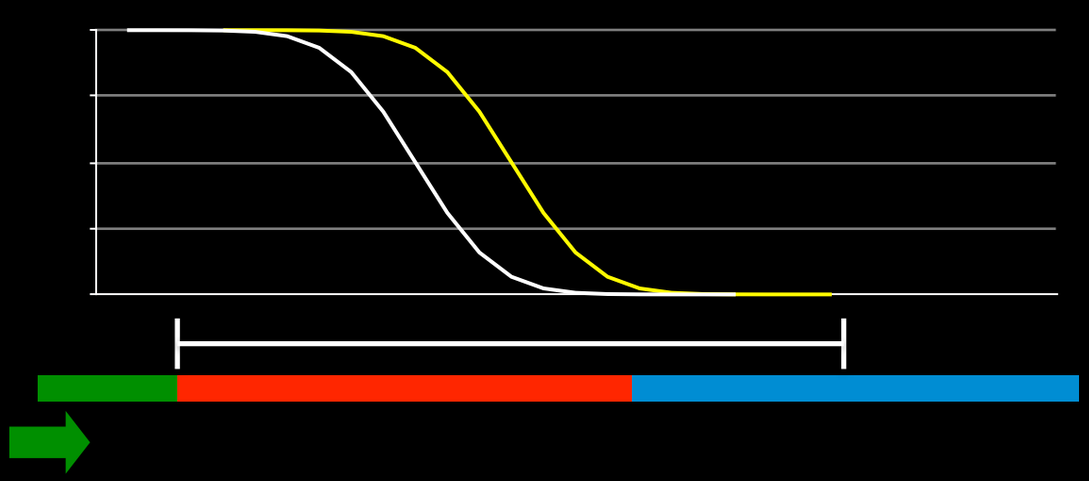


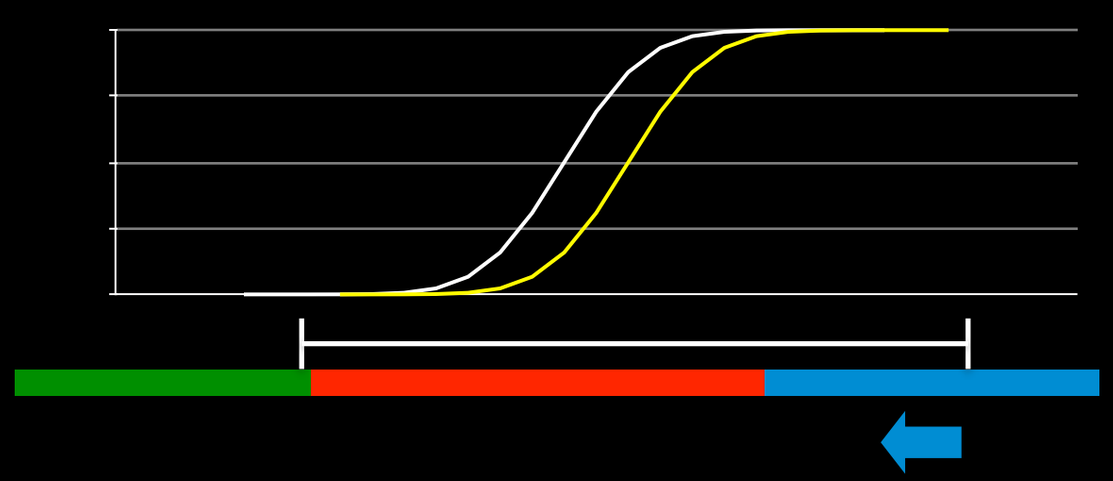




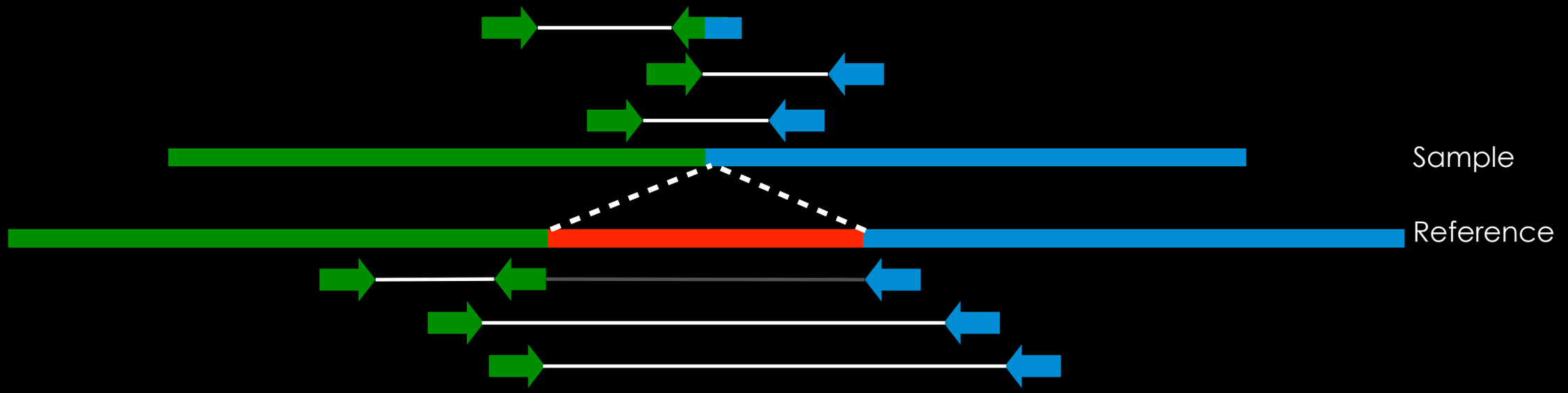
Combining SV signals

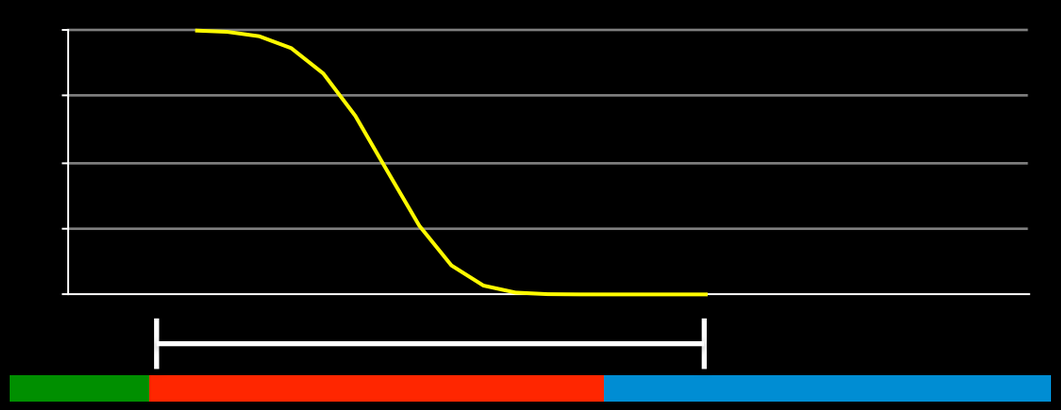


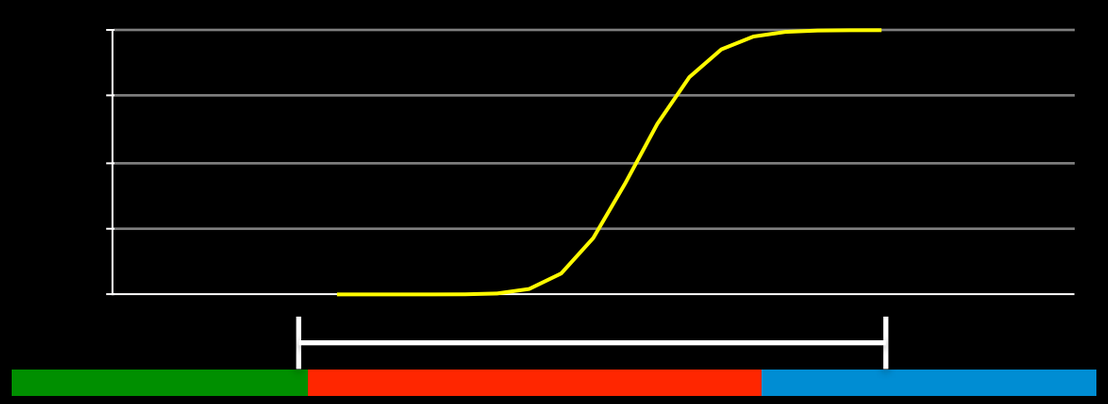




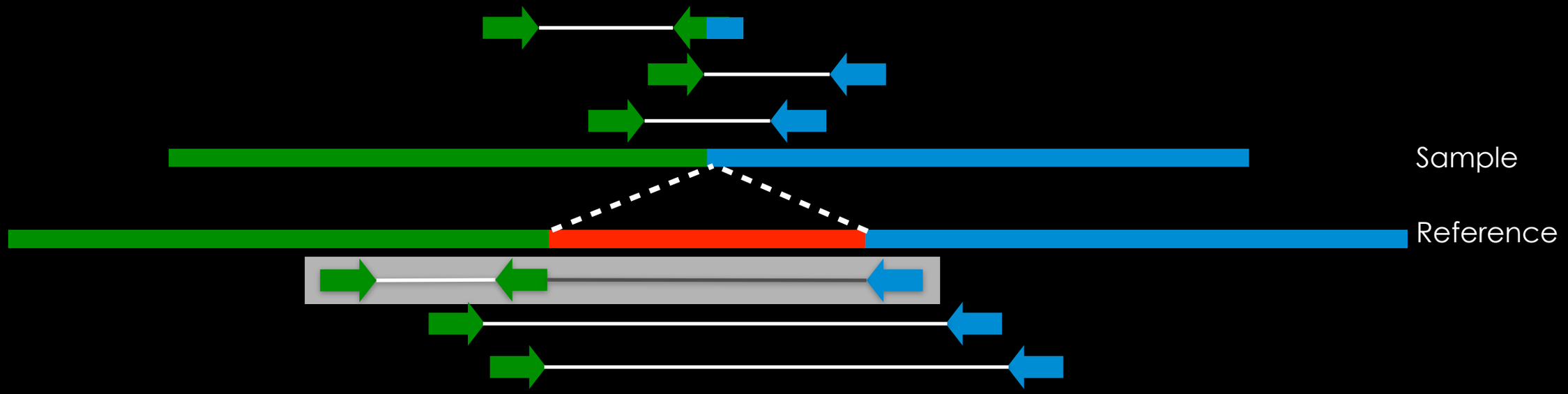
Combining SV signals

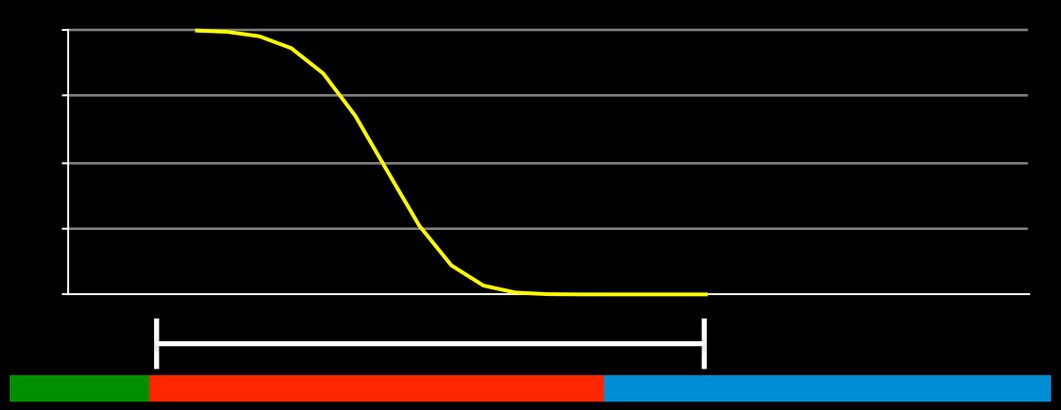


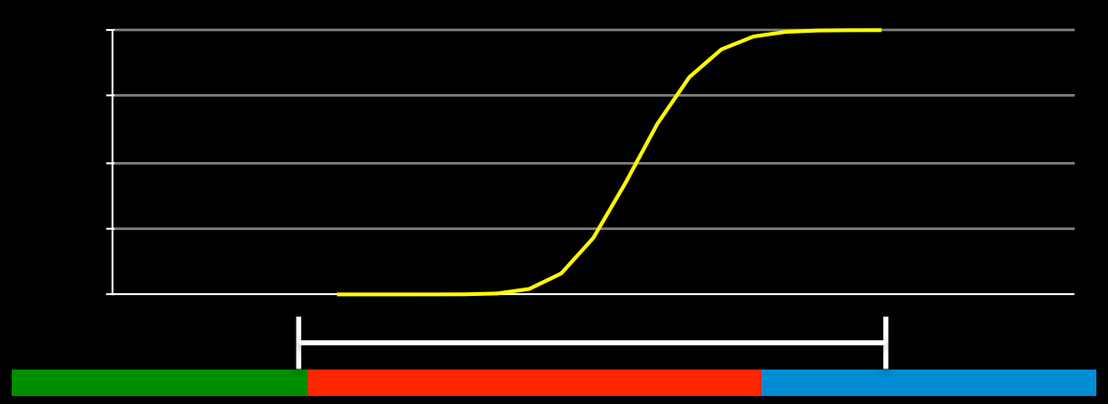




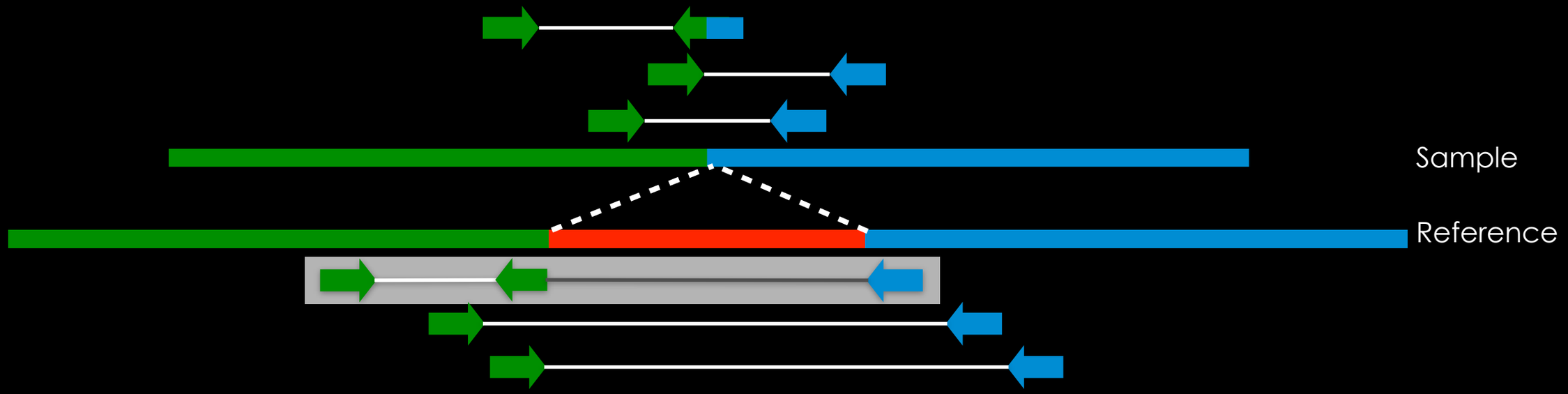
Combining SV signals

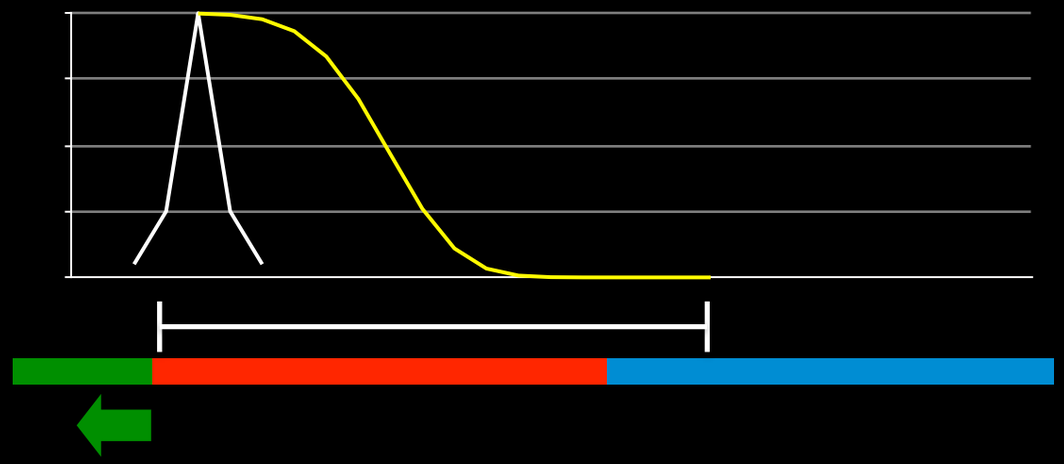


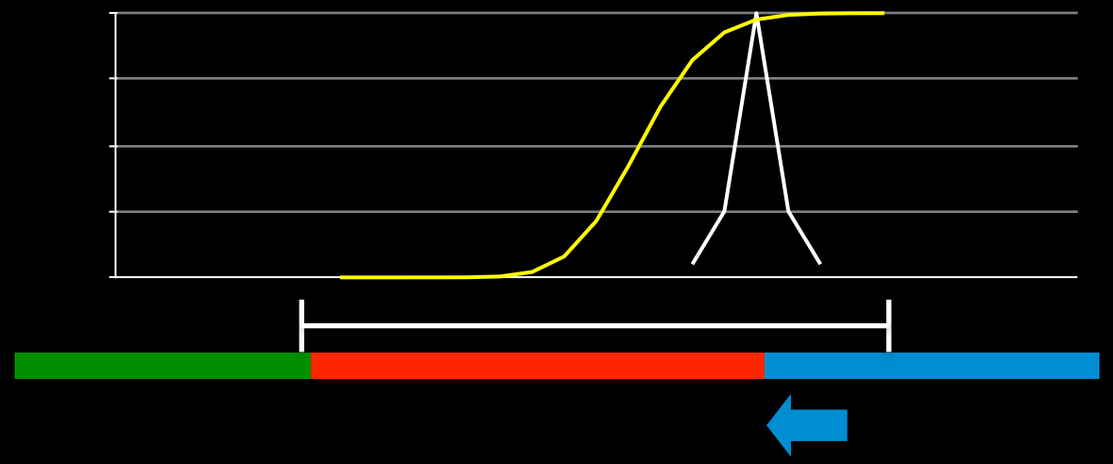




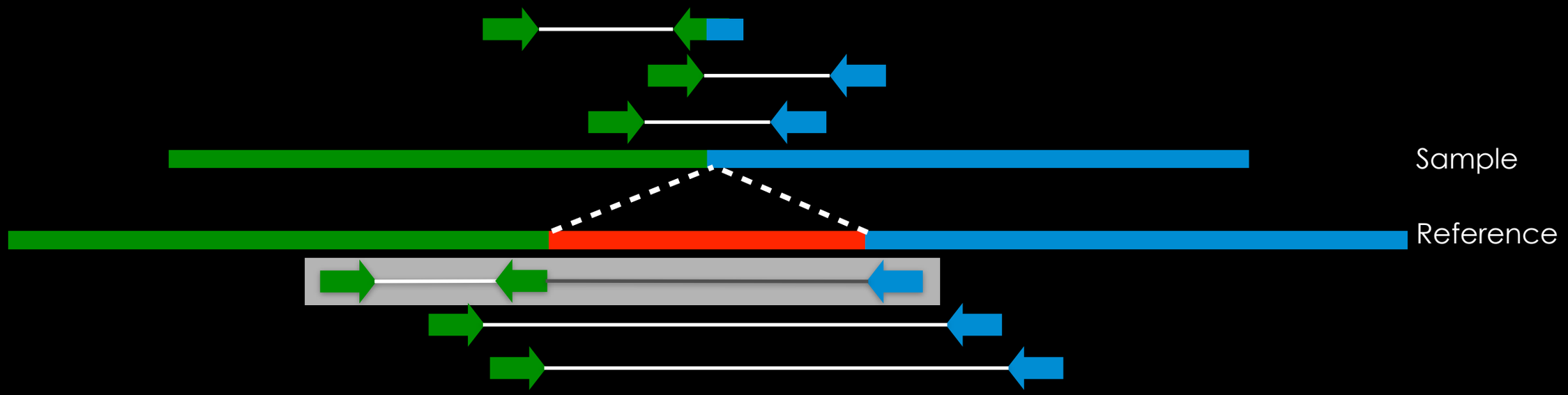
Combining SV signals

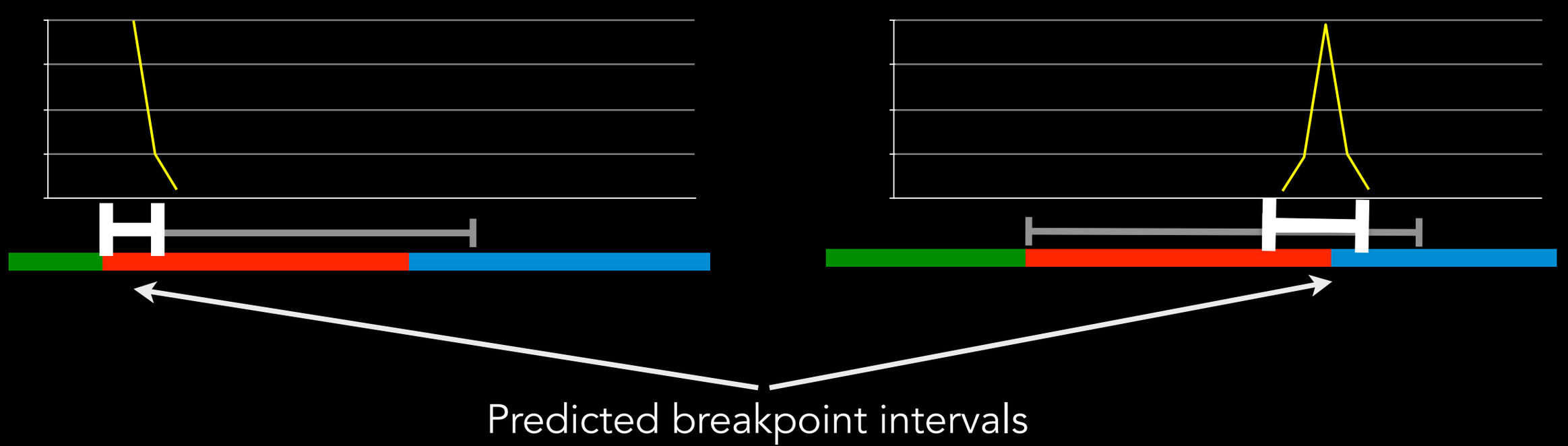






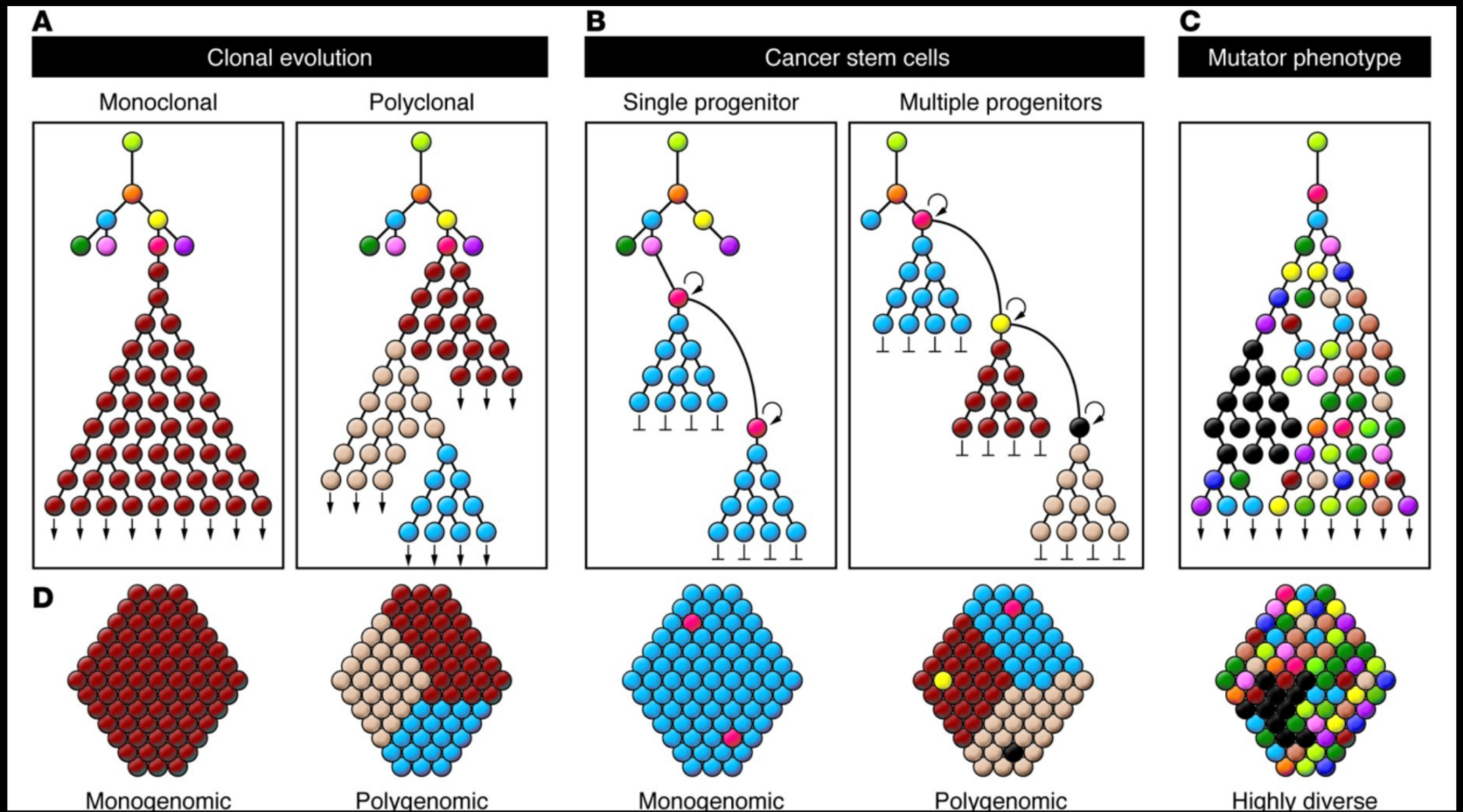
Combining SV signals





Much greater SV breakpoint resolution and sensitivity

Sensitivity is crucial in the context of tumor heterogeneity



Russnes et al, 2011

Tumor heterogeneity simulation: *an in silico* "spike in"

"tumor"

1000 Genomes
A Deep Catalog of Human Genetic Variation

hg19 + 5,516 known deletions from 1000G

+

"normal"

GRC Genome Reference Consortium

hg19 (build 37)

Tumor heterogeneity simulation: *an in silico* "spike in"

"tumor"

1000 Genomes
A Deep Catalog of Human Genetic Variation

hg19 + 5,516 known deletions from 1000G

+

"normal"

GRC Genome Reference Consortium

hg19 (build 37)

50% tumor freq.

"tumor"

FASTA

↓

wgsim (20x)

↘

40X BAM

"normal"

FASTA

↓

wgsim (20X)

↙

What fraction of the 5516 SVs can we detect?

Tumor heterogeneity simulation: *an in silico* "spike in"

"tumor"

1000 Genomes
A Deep Catalog of Human Genetic Variation

hg19 + 5,516 known deletions from 1000G

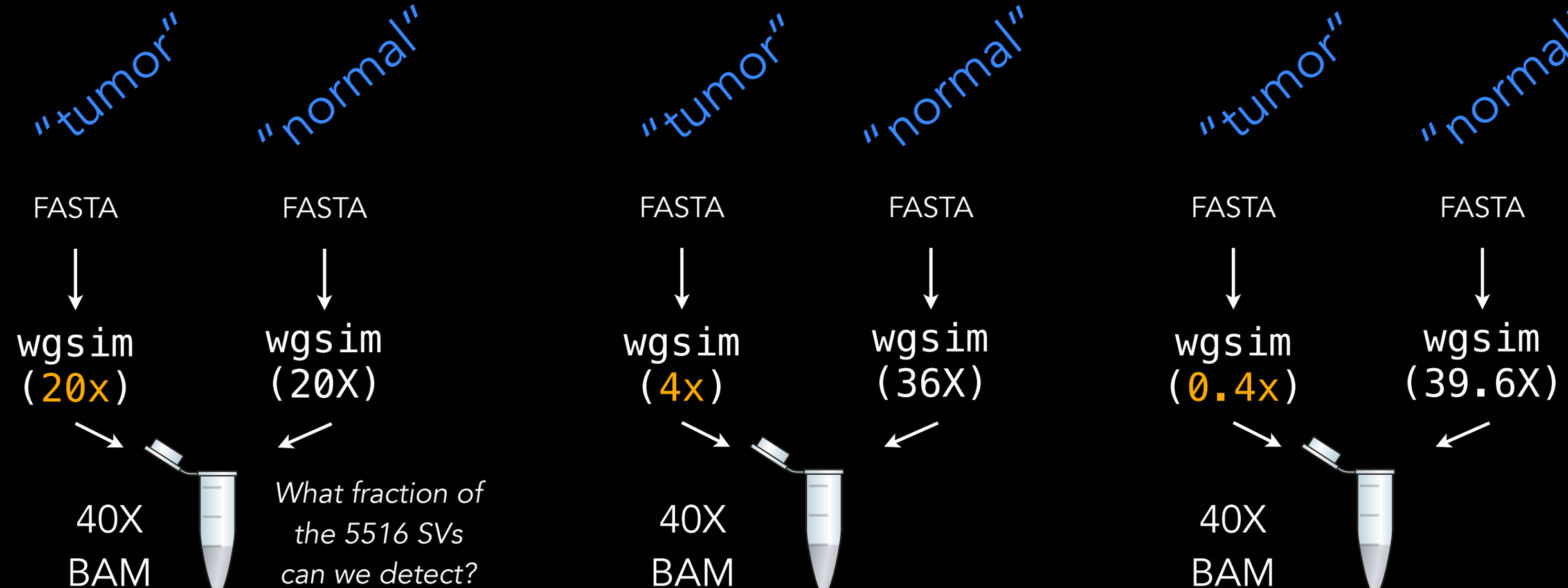
+

"normal"

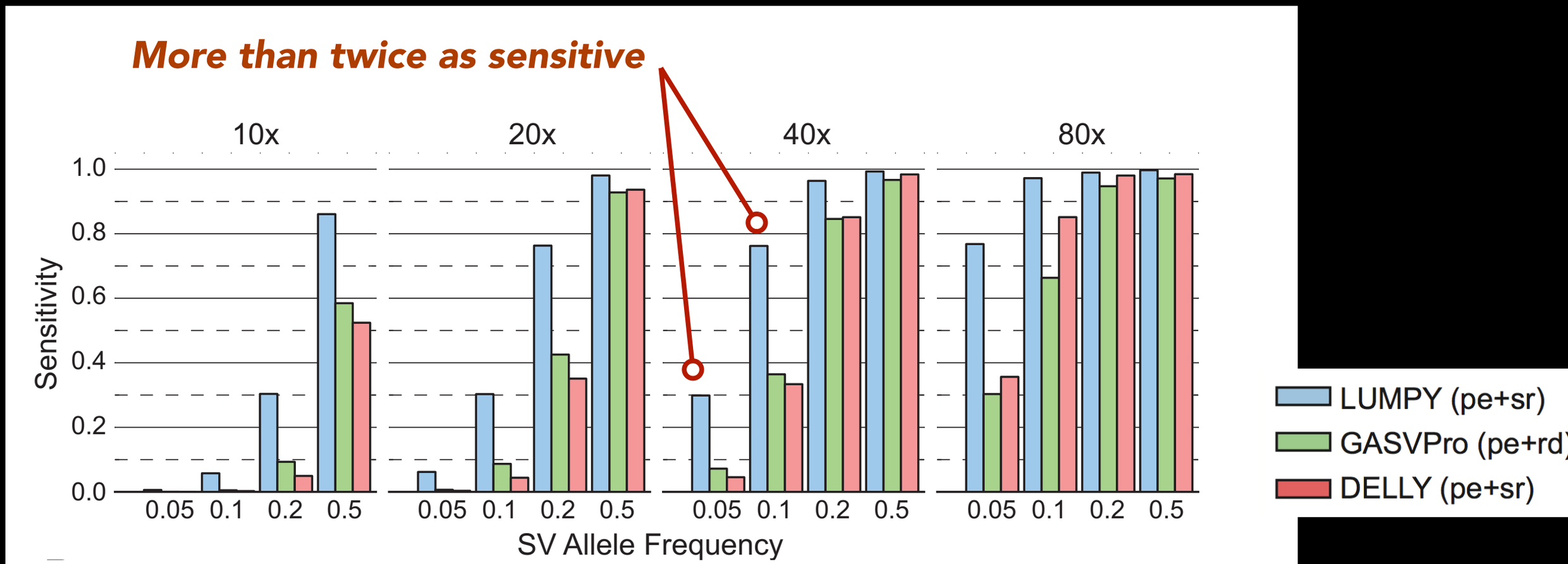
GRC Genome Reference Consortium

hg19 (build 37)

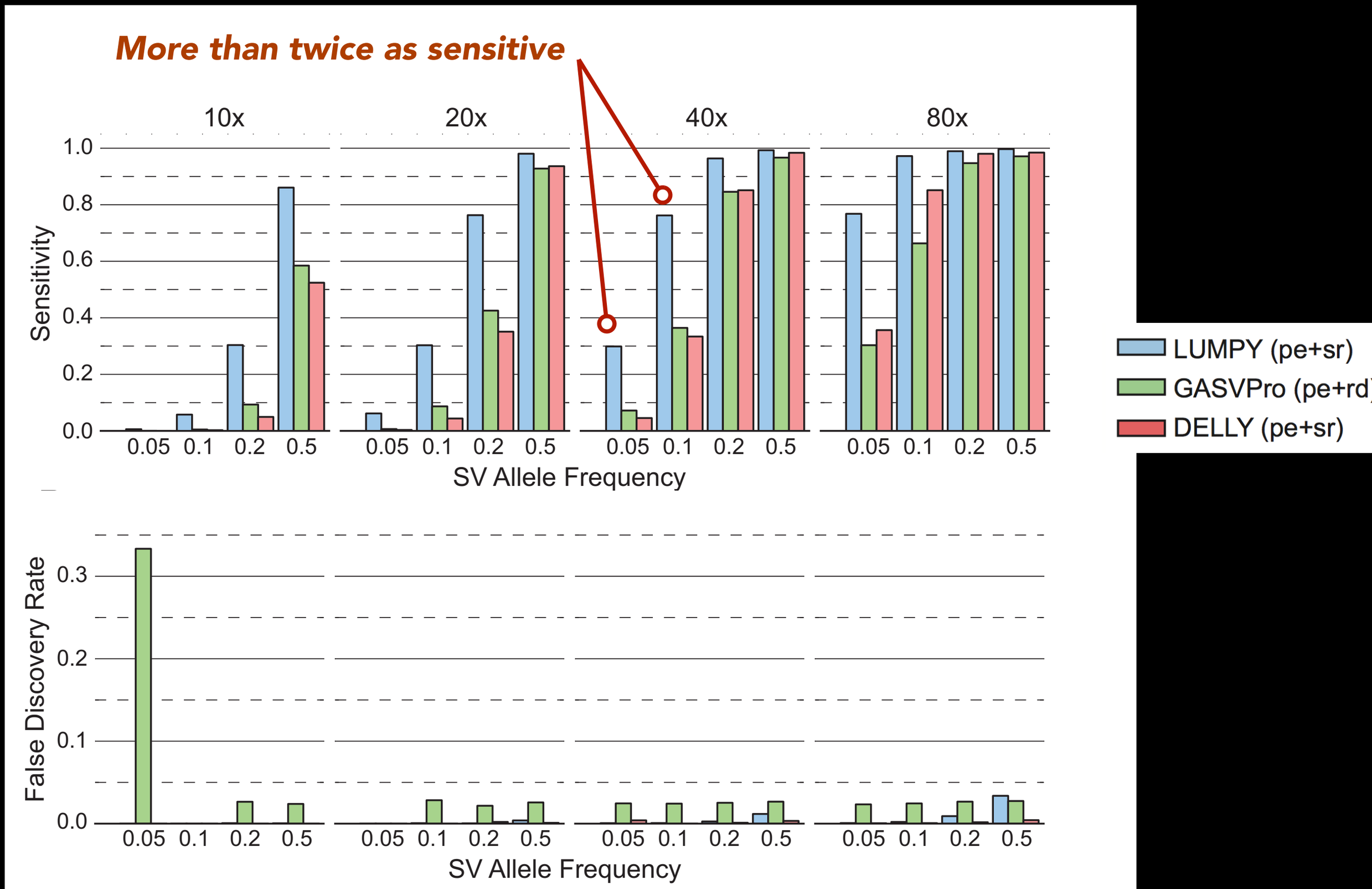
50% tumor freq. | 20% tumor freq. | . . . | 1% tumor freq.



LUMPY has highest sensitivity



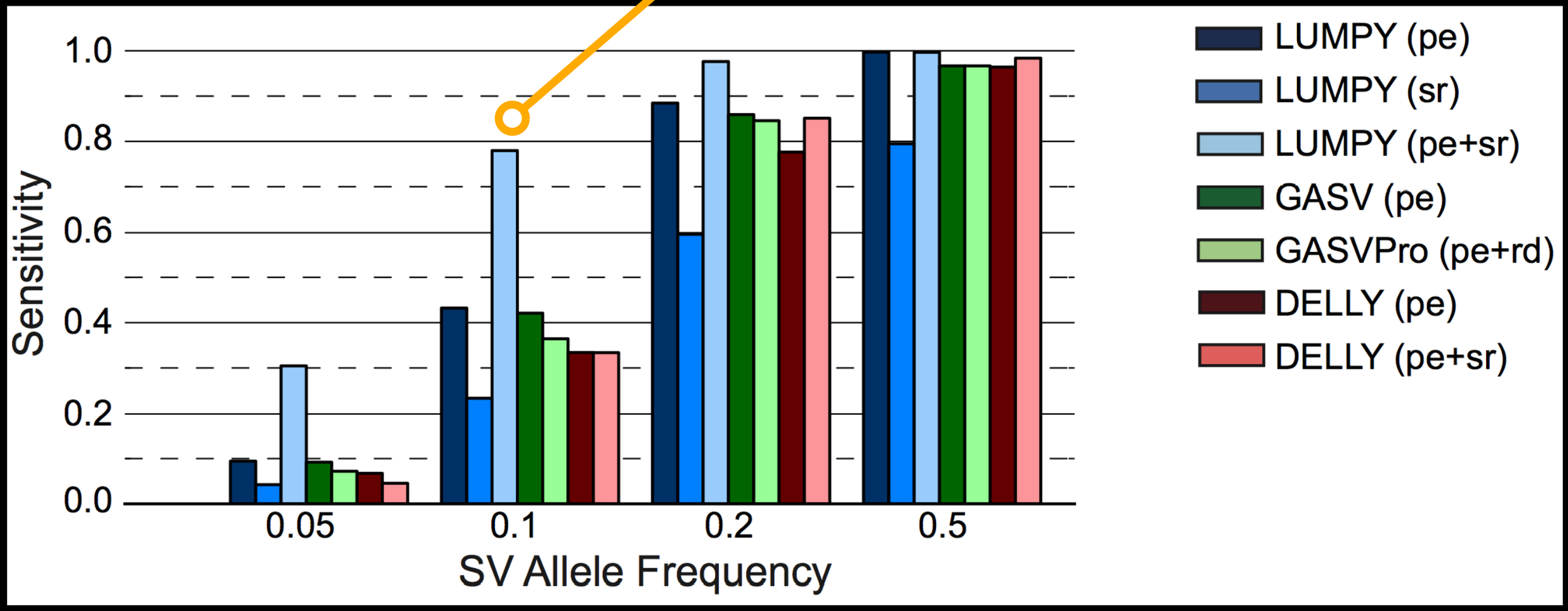
LUMPY has highest sensitivity...with minimal FDR



The impact of combining multiple SV signals

Combining paired-end and split-read signals is more sensitive than each alone

(40X coverage)

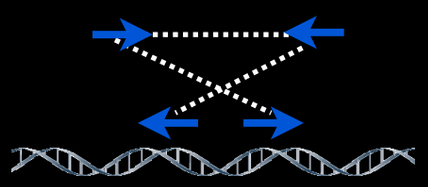


Solution 2: pool data from many samples

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

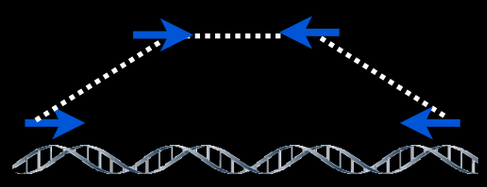
PEM clusters ***discordant*** mappings

everted orientation



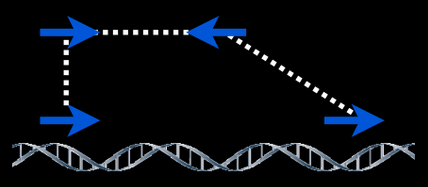
tandem duplication

too big



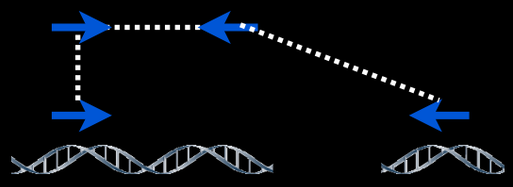
deletion

same-strand



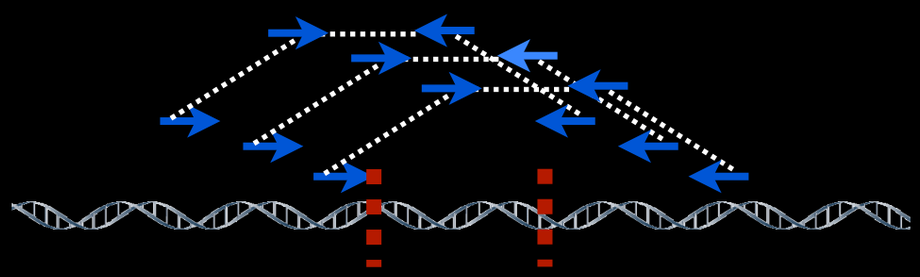
inversion

distant linkage



insertion,
retrotransposition,
translocation

Cluster to localize breakpoints



ref. genome

Deleted interval

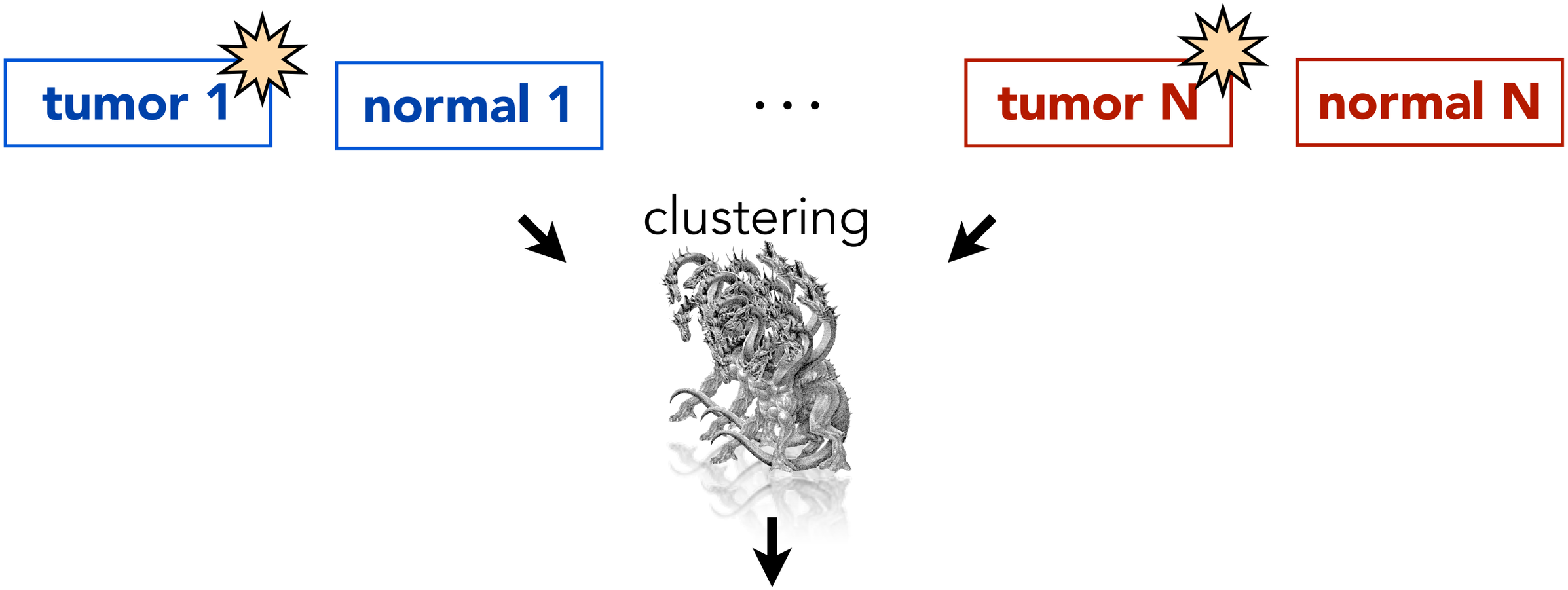
Hydra

1 signal (PEM), 1 sample

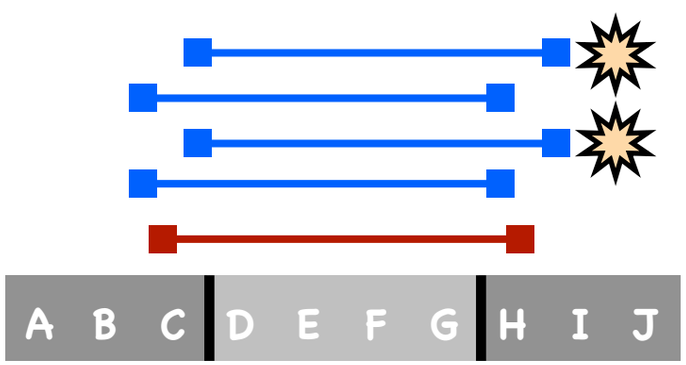
The Hydra algorithm:

- 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/

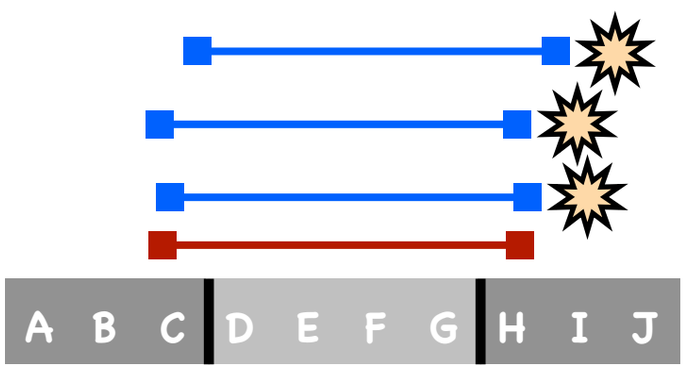
HYDRA-MULTI: Pooling prevents false somatic calls



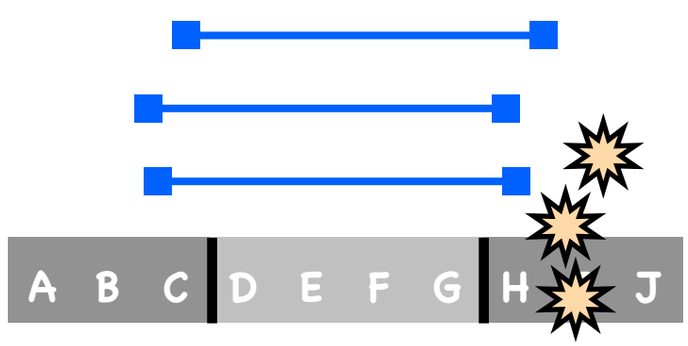
inherited mutation



inherited mutation
(FP somatic w/o pooling)



somatic mutation



Quinlan et al., Cell Stem Cell (2011);

Note: GATK pioneered population-based SNP and INDEL detection; GenomeSTRiP and VariationHunter use a similar approach

The landscape of complex variation in 64 cancer genomes.

(*using HYDRA-MULTI*)

Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms

Ankit Malhotra,[1] Michael Lindberg,[1] Gregory G. Faust,[1,2] Mitchell L. Leibowitz,[1] Royden A. Clark,[1] Ryan M. Layer,[1,2] Aaron R. Quinlan,[1,3,4,5] and Ira M. Hall[1,3,5]

[1]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22903, USA; [2]Department of Computer Science, University of Virginia, Charlottesville, Virginia 22903, USA; [3]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908, USA; [4]Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia 22908, USA

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)

Data (re++)processing

Ankit Malhotra

Michael Lindberg

129 BAM files
99.7 billion readpairs
14 TB raw data from TCGA

→

quality control (corrupt files, incorrect flags)

→

calculate library stats

↓

re-align discordant reads with NovoAlign

←

hyper-aggressive duplicate removal

←

HYDRA-MULTI

↓

Filtering
- inversion artifact
- simple repeats
- mapping quality
- number mappings
- support (3 readpairs)

→

33,218 HIGH CONFIDENCE CALLS

→

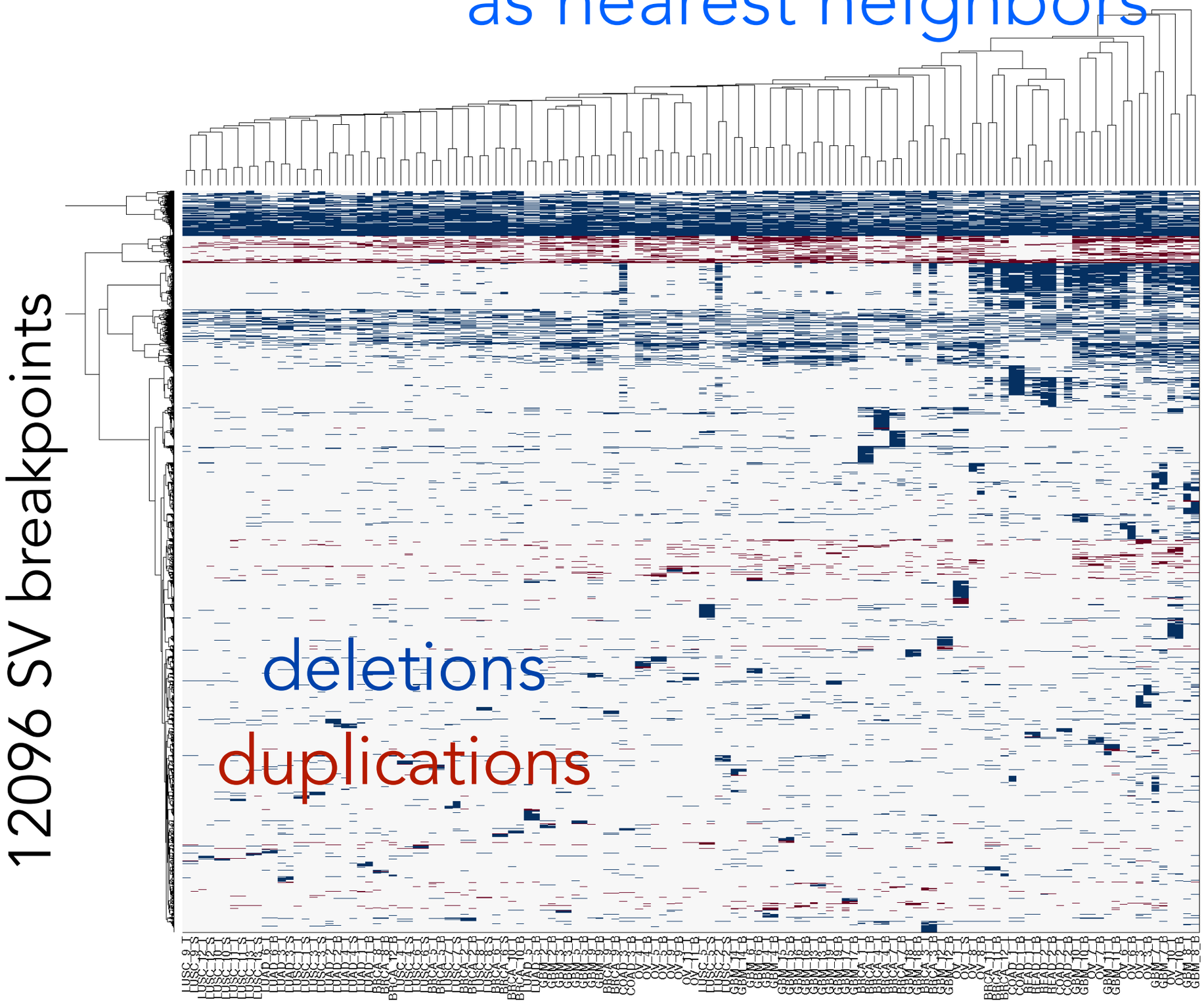
27039 inherited germline breakpoints

→

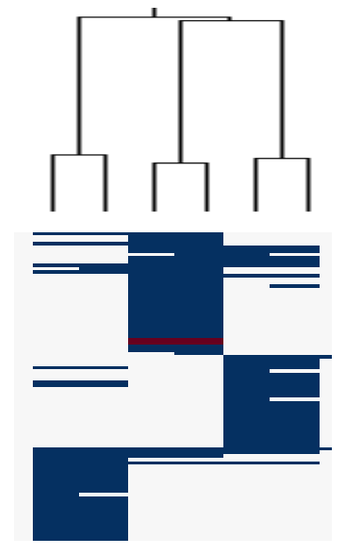
6179 de novo somatic breakpoints

How do we assess the quality of the somatic rearrangement calls?

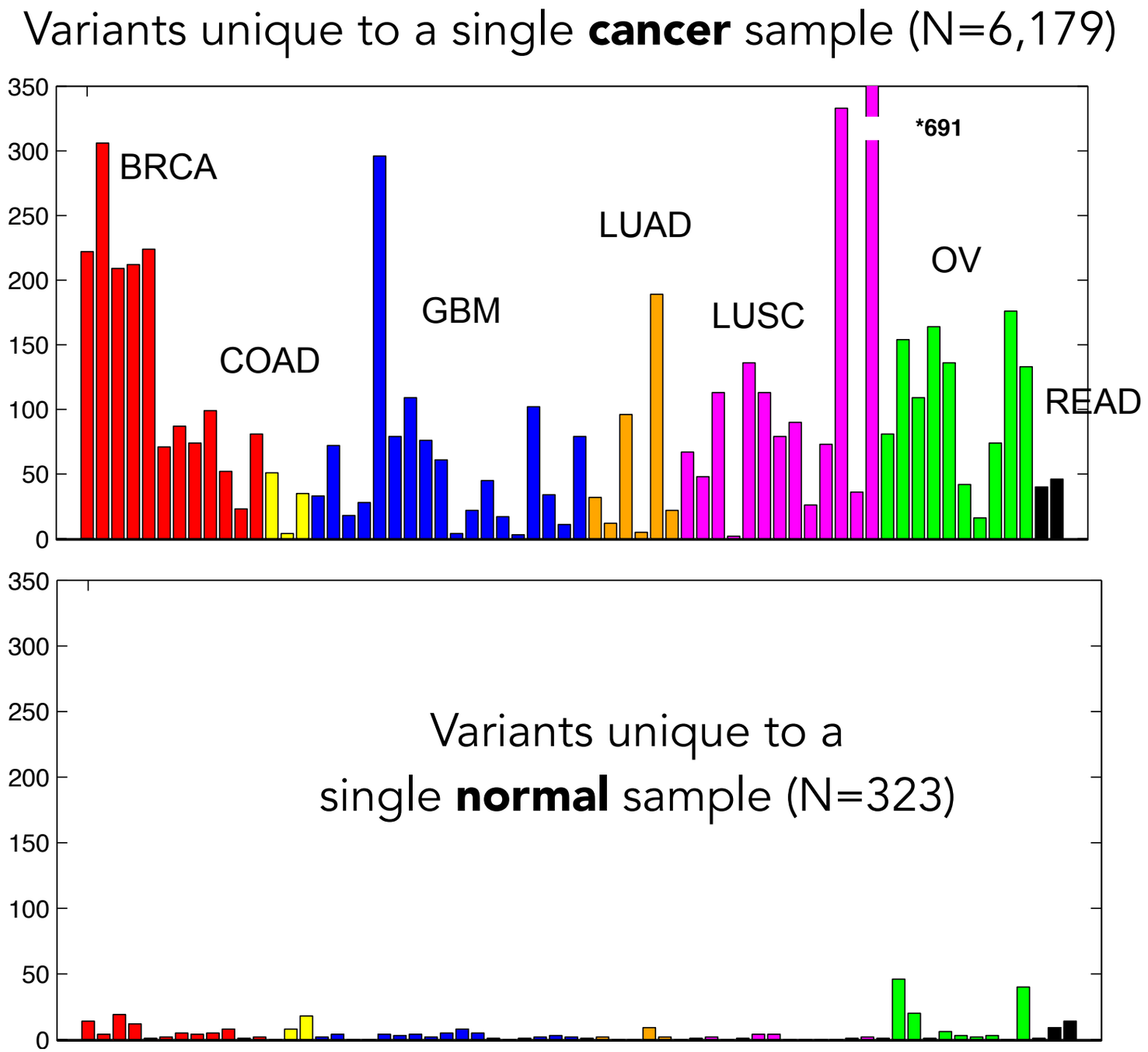
1. 64 out of 64 tumor / normal pairs cluster as nearest neighbors



3 tumor/ normal pairs



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

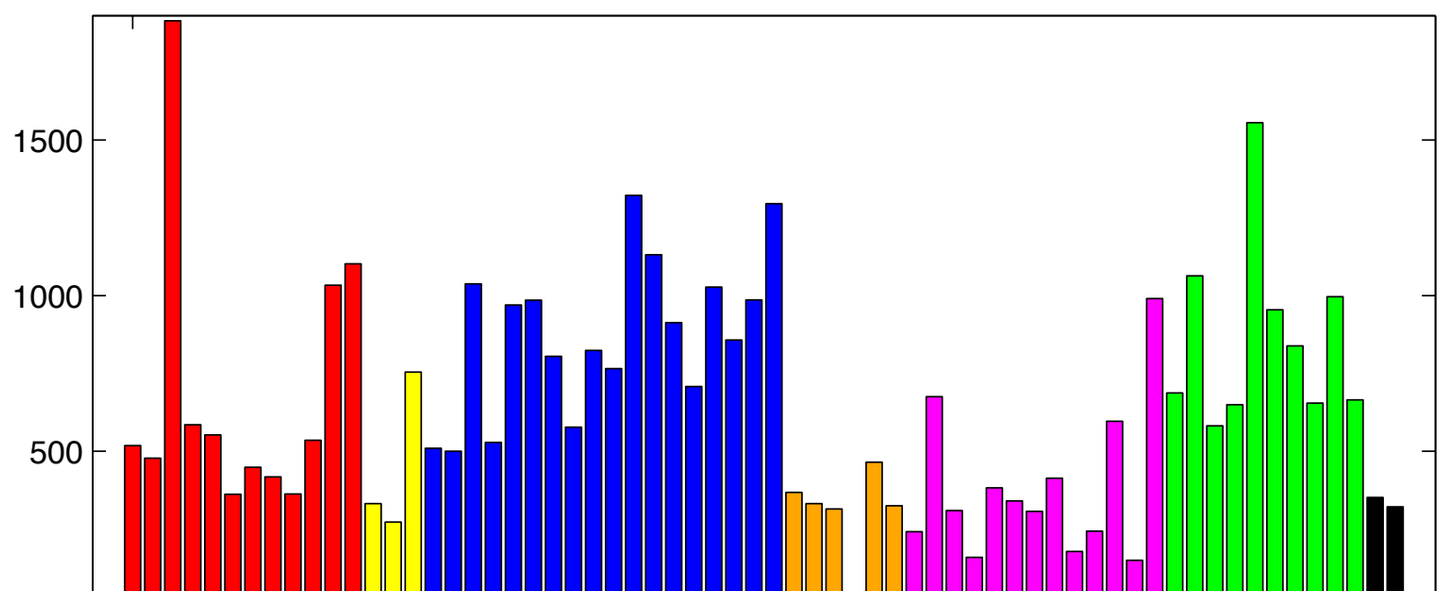


Assuming all normal-only calls are false, suggests 5% somatic prediction error rate.

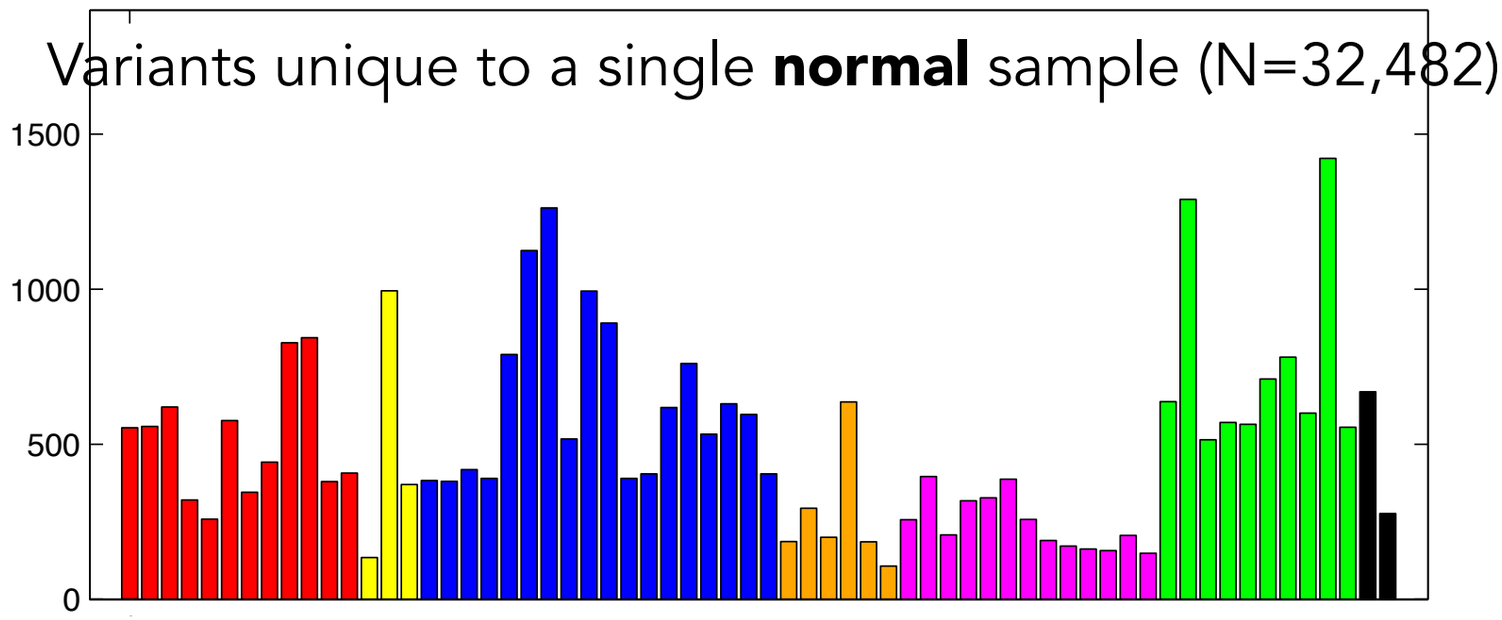
Likelihood of LOH suggests it is actually lower.

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%!

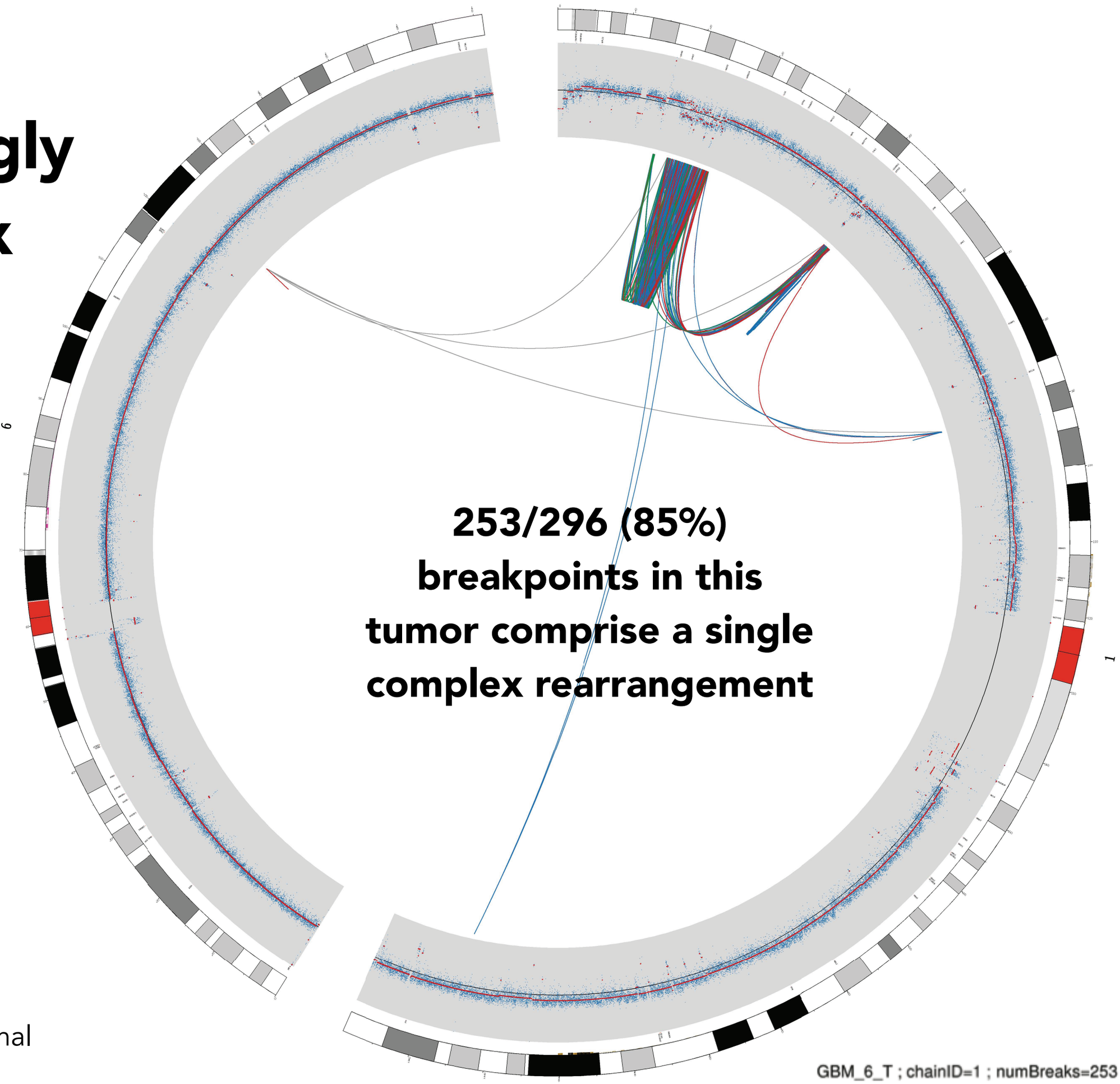
We have a high-quality set of somatic rearrangements from multiple tumors.

But what do they tell us about chromosome evolution in cancers?

Observation 1.

We immediately noticed several staggeringly complex rearrangements (CRs).

A staggeringly complex variant



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

A couple of weeks later...

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,*]

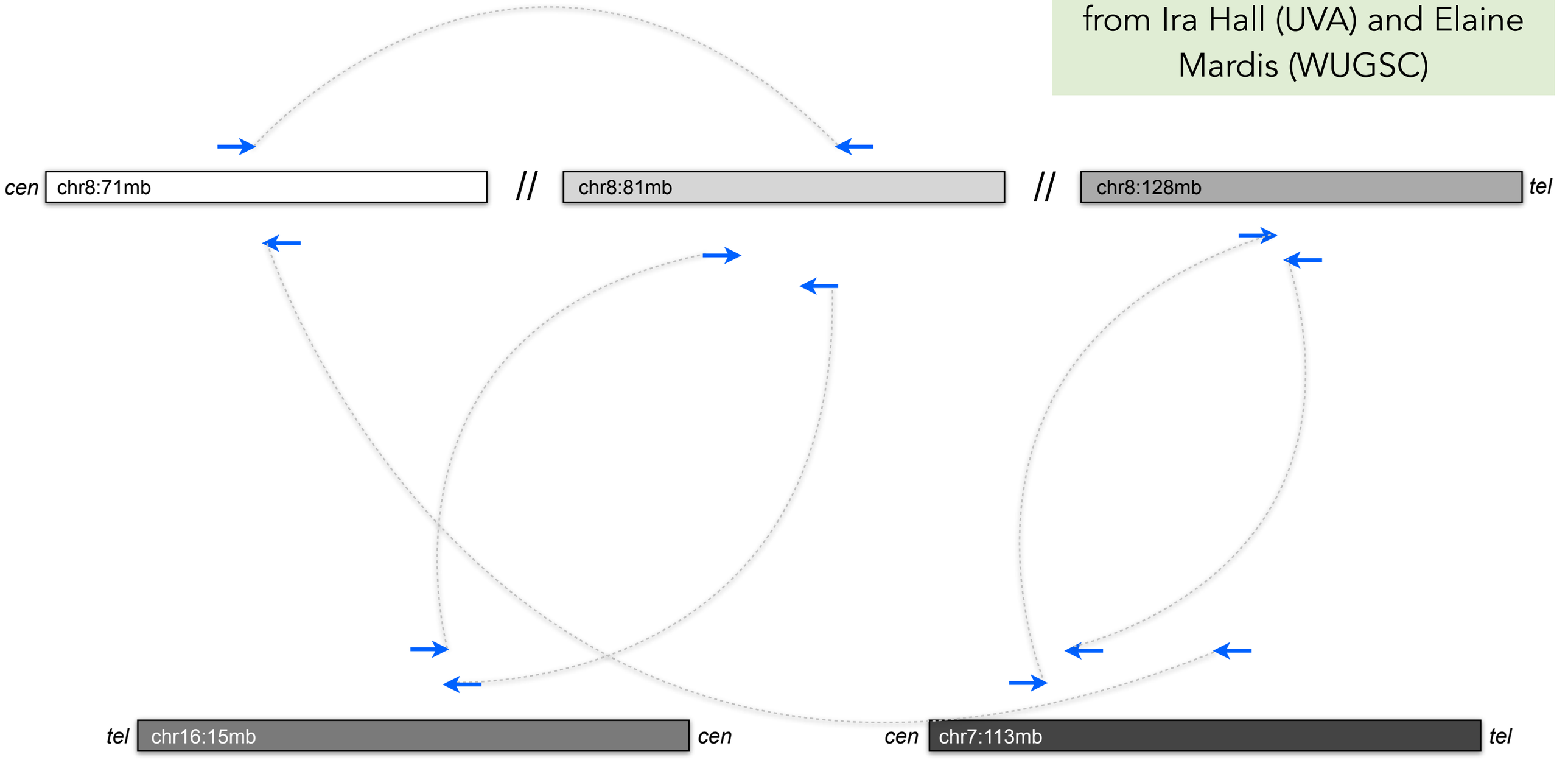
Chromothripsis: chromosome shattering in a single, catastrophic event.

Why are complex genomic rearrangements important?

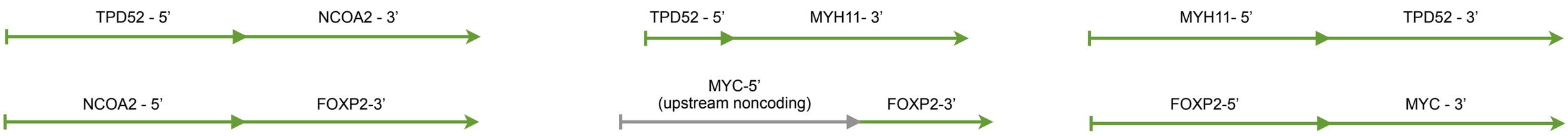
1) Punctuated genome evolution

A relatively mutation free multiple myeloma genome with 1 balanced rearrangement that produces 5 fusion genes

from Ira Hall (UVA) and Elaine Mardis (WUGSC)



fusion products

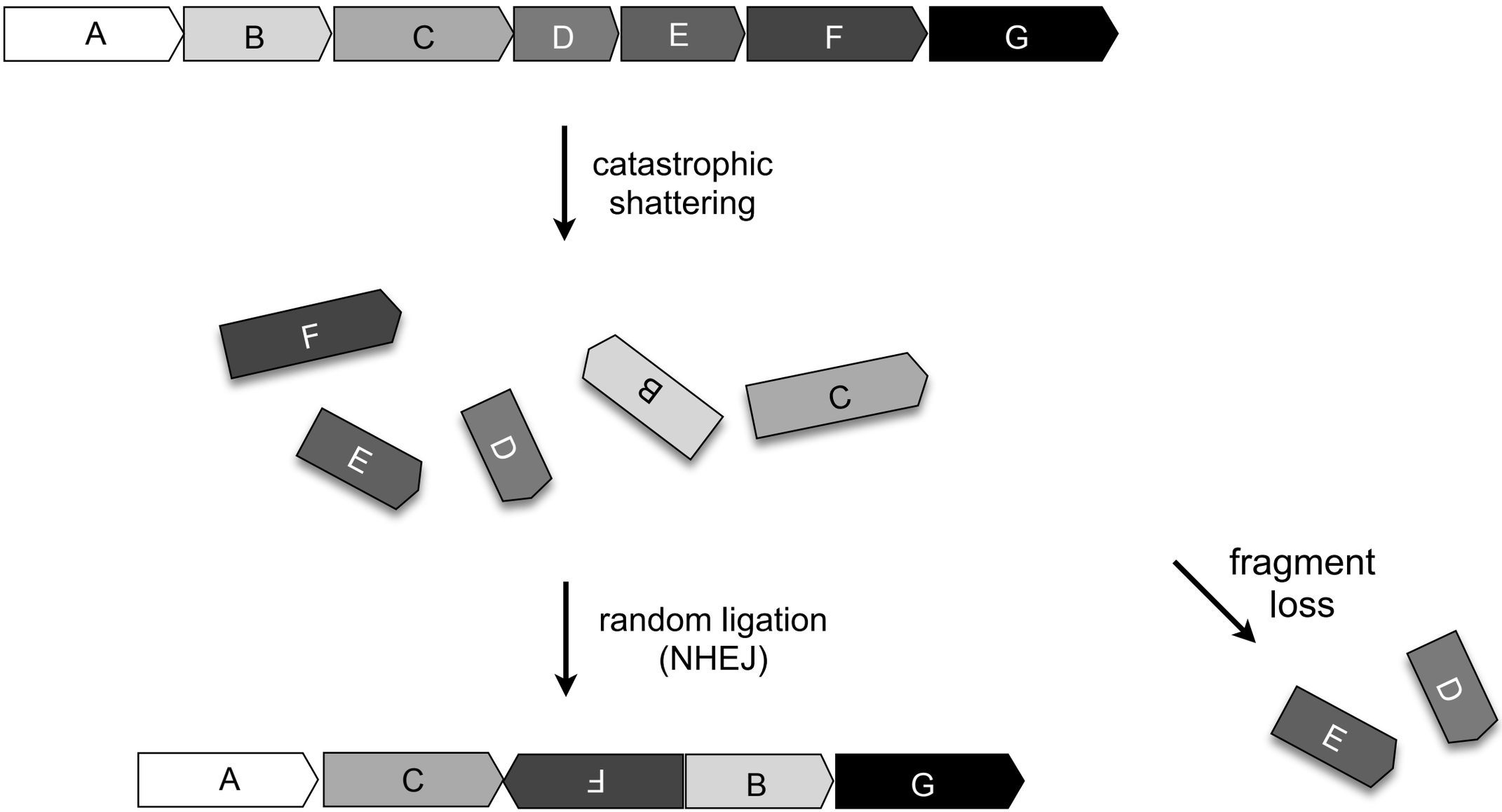


Why are complex genomic rearrangements important?

1) Punctuated genome evolution

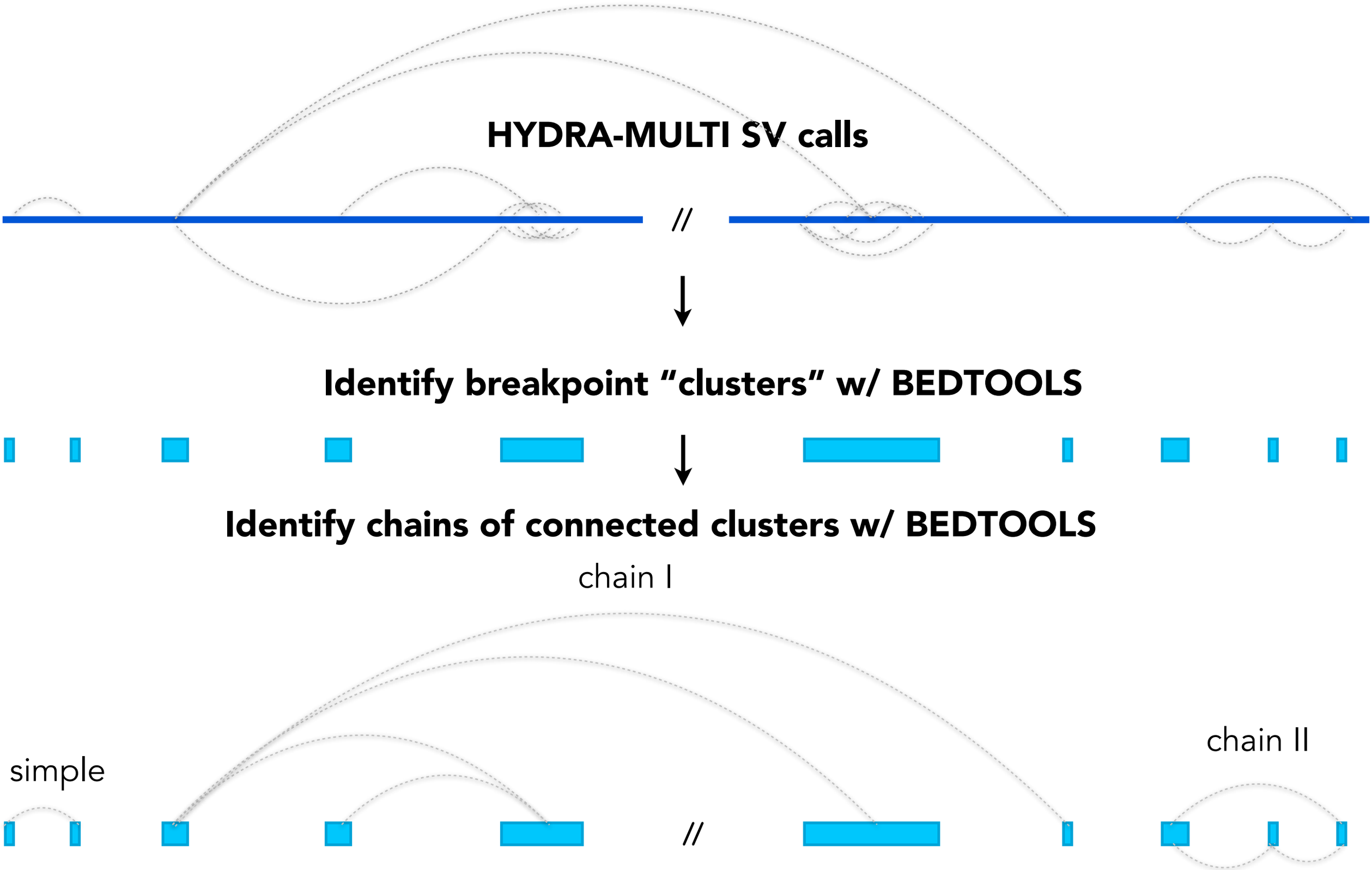
2) Mechanistically interesting

A model for chromothripis



Stephens et al., Cell, 2011

Identifying complex rearrangements



Observation 2.

Complex rearrangements are quite common in tumor genomes.

25% of all somatic breakpoints are part of complex mutations. Not random.

	Breakpoints			Complex rearrangements	
	Total (mean)	% in clusters	% in CGRs	Mild (3-9 breaks)	Extreme (>9 breaks)
BRCA (n=12)	1657 (138)	4.2%	2.1%	11	0
COAD (n=3)	90 (30)	10%	4.4%	1	0
GBM (n=18)	1088 (60)	70%	49.3%	18	9 (7)
LUAD (n=6)	356 (59)	23%	16.8%	9	2 (2)
LUSC (n=13)	1806 (139)	26.7%	7.7%	27	2 (2)
OV (n=11)	1096 (100)	11.6%	4.8%	15	0
READ (n=2)	86 (43)	11.6%	11.6%	3	0
Total	6179	25%	13.6%	84	13 (11)

Observation 3.

Complex rearrangements are very common in glioblastoma.

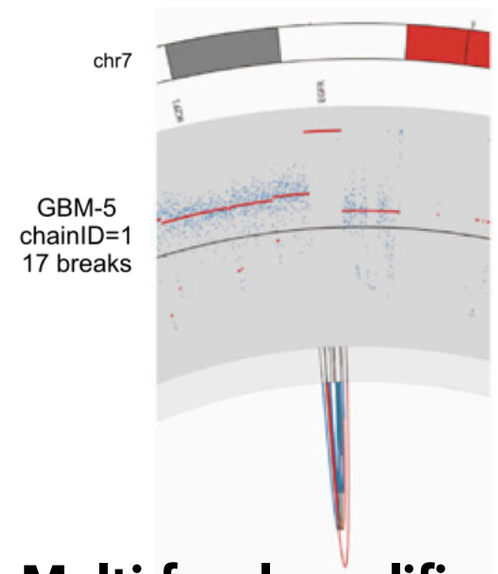
Enrichment in GBM. Compare to BRCA: more breakpoints per sample, but rarely in complex loci

	Breakpoints			Complex rearrangements	
	Total (mean)	% in clusters	% in CGRs	Mild (3-9 breaks)	Extreme (>9 breaks)
BRCA (n=12)	1657 (138)	4.2%	2.1%	11	0
COAD (n=3)	90 (30)	10%	4.4%	1	0
GBM (n=18)	1088 (60)	70%	49.3%	18	9 (7)
LUAD (n=6)	356 (59)	23%	16.8%	9	2 (2)
LUSC (n=13)	1806 (139)	26.7%	7.7%	27	2 (2)
OV (n=11)	1096 (100)	11.6%	4.8%	15	0
READ (n=2)	86 (43)	11.6%	11.6%	3	0
Total	6179	25%	13.6%	84	13 (11)

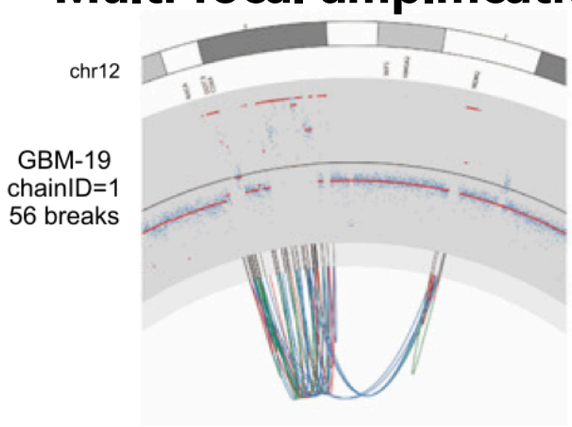
Observation 4.

Vast architectural diversity observed for complex variants

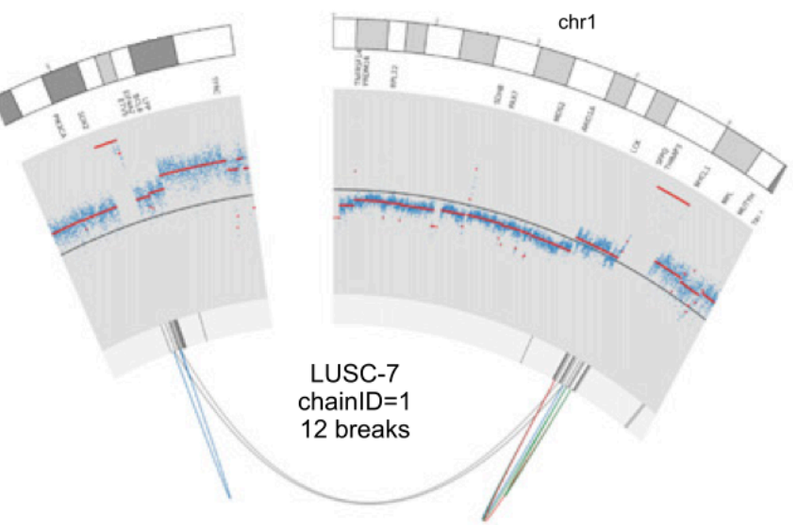
Focal amplification of EGFR



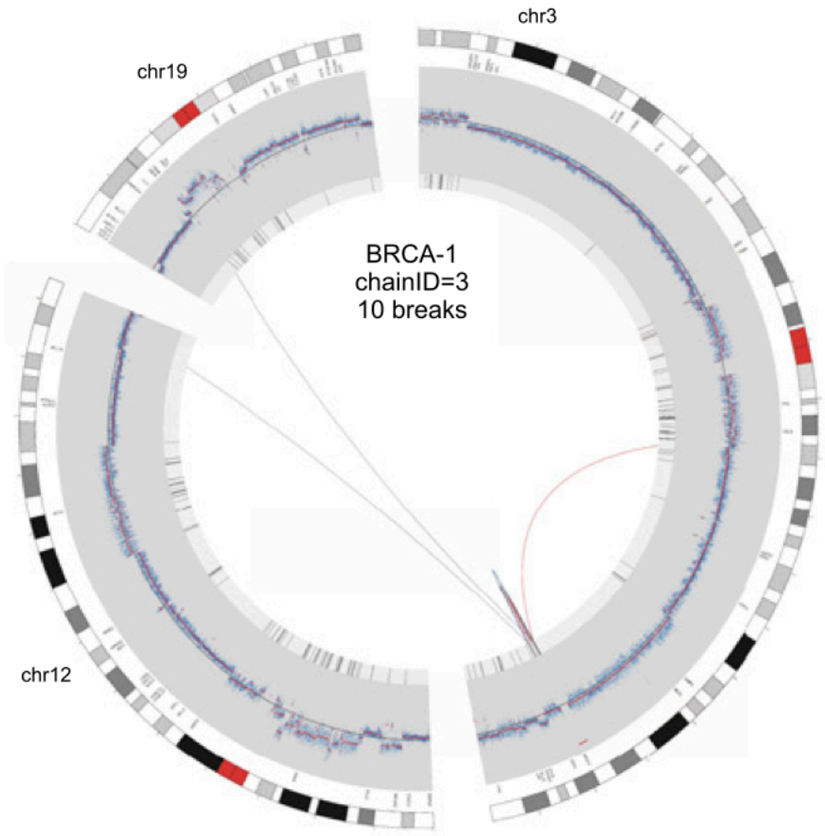
Multi-focal amplification



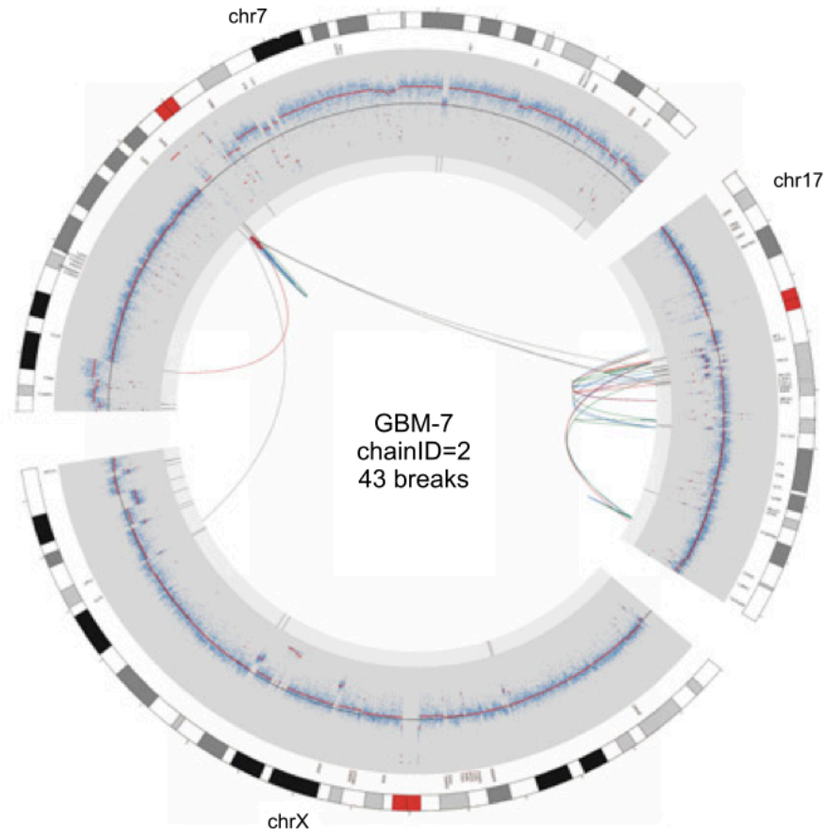
Interchrom. co-amplification



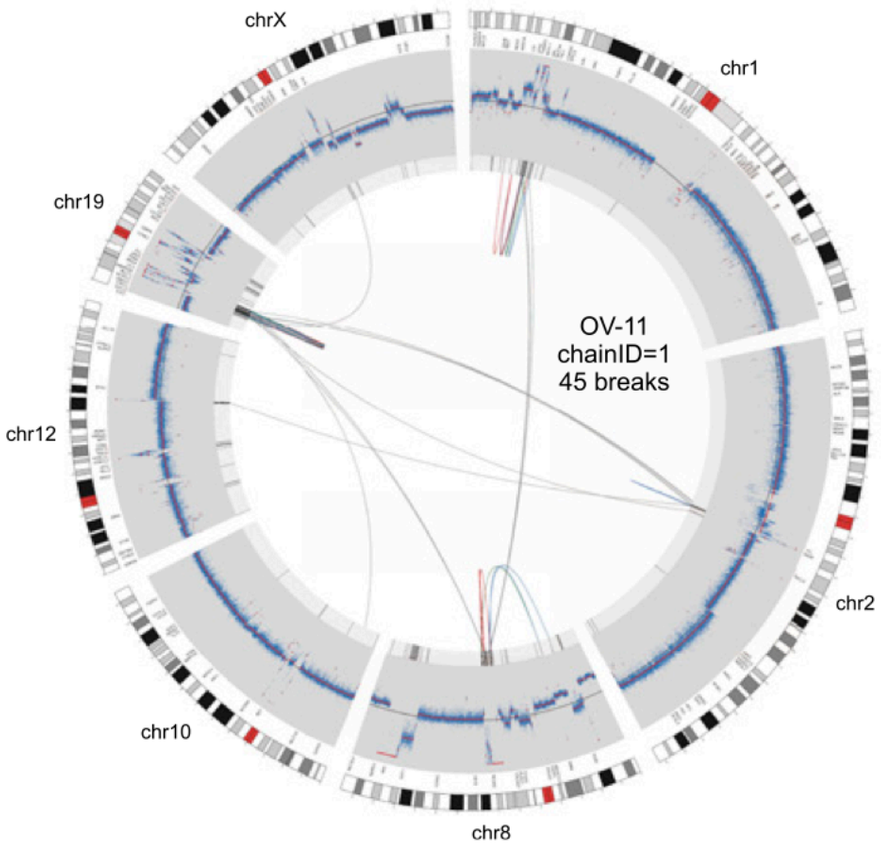
Mild



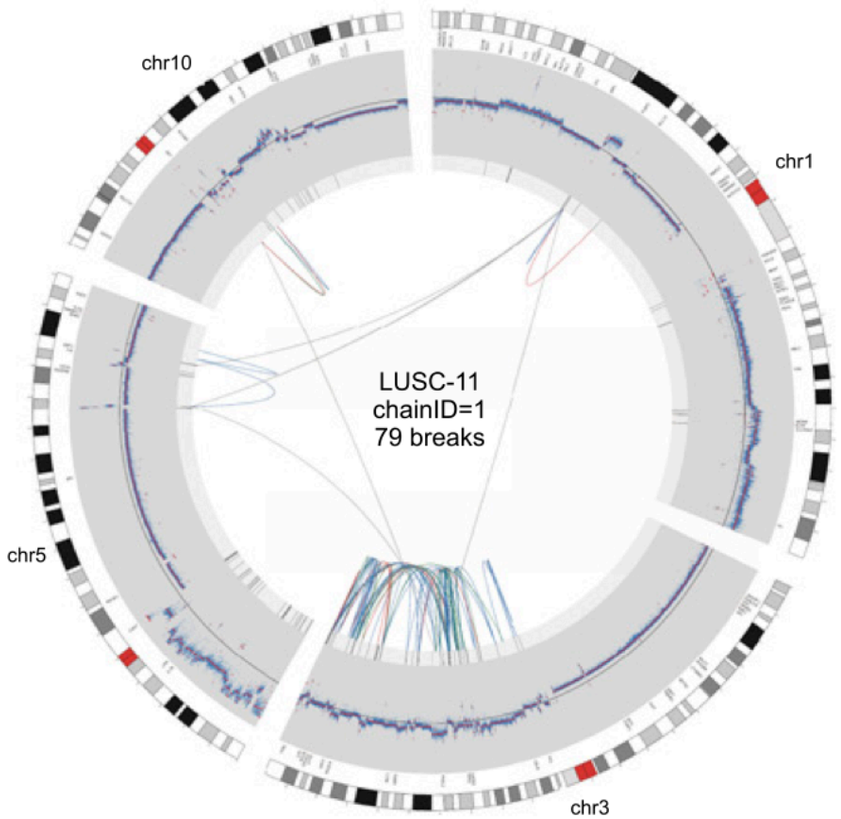
Complex



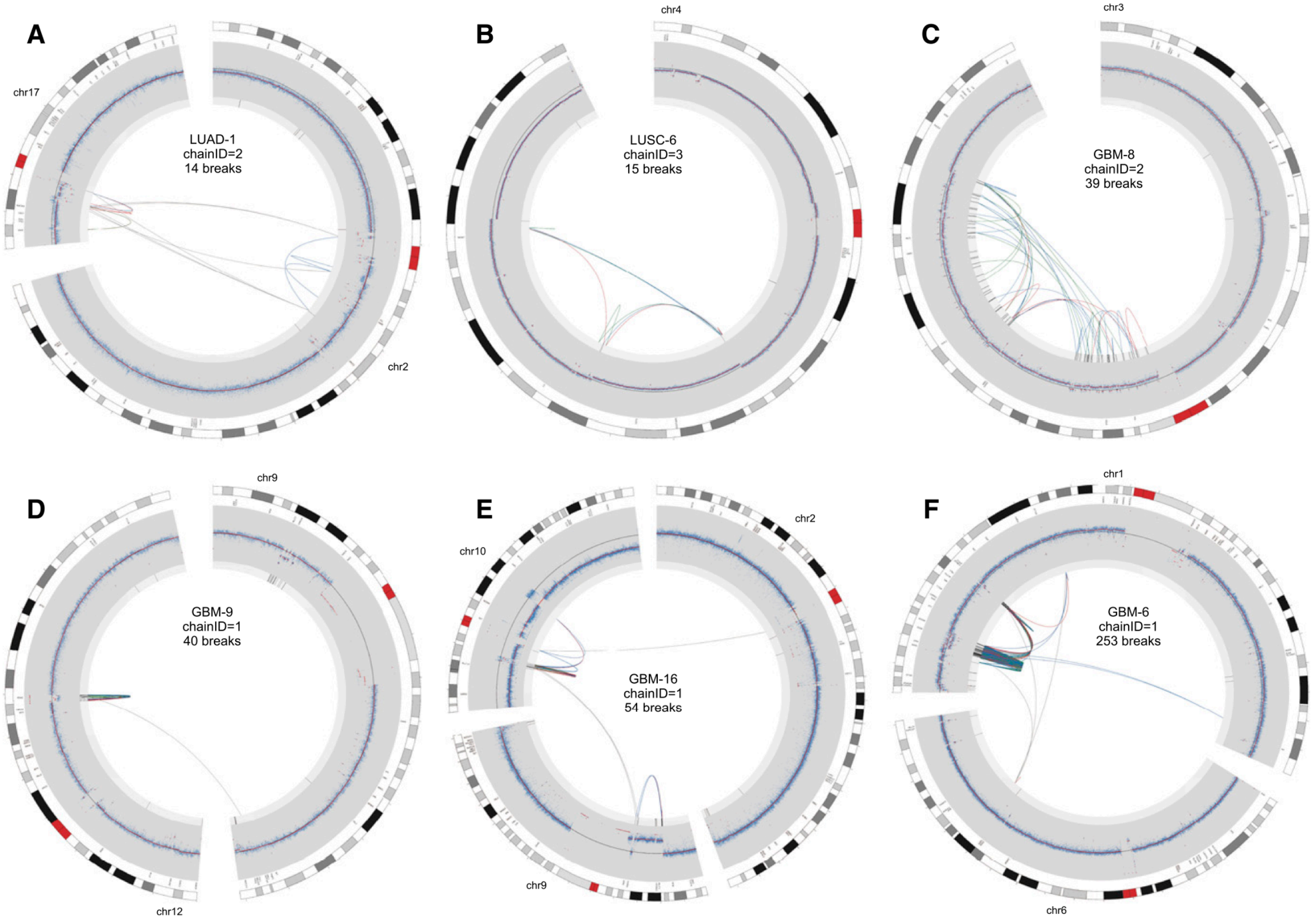
Very complex



Highly-rearranged



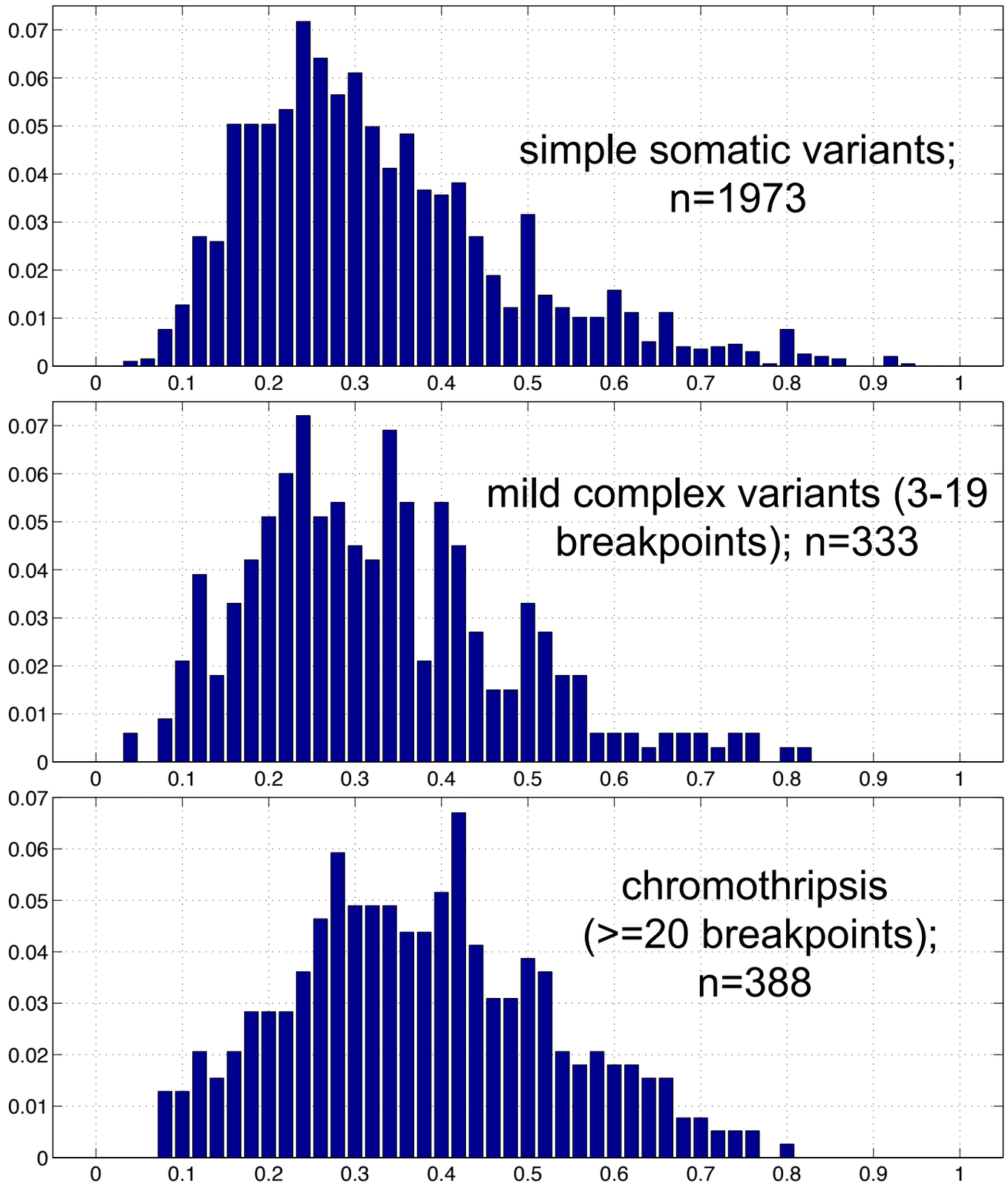
Chromothripis examples



Observation 5.

Complex rearrangements have elevated intra-tumor allele frequencies

Complex loci have higher allele frequencies



Allele Frequency

Allele Frequency

	<0.35	0.35-0.65	>0.65
simple	60.4%	34.5%	5%
complex	49.8%	45.6%	4.6%
complex <20	56.5%	39.6%	3.9%
complex >=20	44.1%	50.7%	5.2%

Why? Evidence that chromothriptic medulloblastomas form extrachromosomal circles (double minutes) containing oncogenes

Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with *TP53* Mutations

Tobias Rausch,[1,18] David T.W. Jones,[2,18] Marc Zapatka,[2,18] Adrian M. Stütz,[1,18] Thomas Zichner,[1] Joachim Weischenfeldt,[1] Natalie Jäger,[3] Marc Remke,[2,5] David Shih,[6] Paul A. Northcott,[6] Elke Pfaff,[2] Jelena Tica,[1] Qi Wang,[5] Luca Massimi,[7] Hendrik Witt,[2,5] Sebastian Bender,[2,5] Sabrina Pleier,[2,5] Huriye Cin,[2] Cynthia Hawkins,[6,8] Christian Beck,[5] Andreas von Deimling,[9] Volkmar Hans,[10] Benedikt Brors,[3] Roland Eils,[3,20] Wolfram Scheurlen,[11] Jonathon Blake,[1] Vladimir Benes,[1] Andreas E. Kulozik,[5] Olaf Witt,[5,4] Dianna Martin,[12] Cindy Zhang,[12] Rinnat Porat,[12] Diana M. Merino,[12] Jonathan Wasserman,[12] Nada Jabado,[13] Adam Fontebasso,[13] Lars Bullinger,[14] Frank G. Rücker,[14] Konstanze Döhner,[14] Hartmut Döhner,[14] Jan Koster,[15] Jan J. Molenaar,[15] Rogier Versteeg,[15] Marcel Kool,[2] Uri Tabori,[6,12] David Malkin,[12] Andrey Korshunov,[9] Michael D. Taylor,[6,16] Peter Lichter,[2,19,*] Stefan M. Pfister,[2,5,19,*] and Jan O. Korbel[1,17,19,*]

Cell

Are brain tumors particularly prone to chromothripsis?

doi:10.1038/nature10910

Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes

Jan J. Molenaar[1]*, Jan Koster[1]*, Danny A. Zwijnenburg[1], Peter van Sluis[1], Linda J. Valentijn[1], Ida van der Ploeg[1], Mohamed Hamdi[1], Johan van Nes[1], Bart A. Westerman[1], Jennemiek van Arkel[1], Marli E. Ebus[1], Franciska Haneveld[1], Arjan Lakeman[1], Linda Schild[1], Piet Molenaar[1], Peter Stroeken[1], Max M. van Noesel[2], Ingrid Øra[1,3], Evan E. Santo[1], Huib N. Caron[4], Ellen M. Westerhout[1] & Rogier Versteeg[1]

Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with *TP53* Mutations

Cell

Tobias Rausch,[1,18] David T.W. Jones,[2,18] Marc Zapatka,[2,18] Adrian M. Stütz,[1,18] Thomas Zichner,[1] Joachim Weischenfeldt,[1] Natalie Jäger,[3] Marc Remke,[2,5] David Shih,[6] Paul A. Northcott,[6] Elke Pfaff,[2] Jelena Tica,[1] Qi Wang,[5] Luca Massimi,[7] Hendrik Witt,[2,5] Sebastian Bender,[2,5] Sabrina Pleier,[2,5] Huriye Cin,[2] Cynthia Hawkins,[6,8] Christian Beck,[5] Andreas von Deimling,[9] Volkmar Hans,[10] Benedikt Brors,[3] Roland Eils,[3,20] Wolfram Scheurlen,[11] Jonathon Blake,[1] Vladimir Benes,[1] Andreas E. Kulozik,[5] Olaf Witt,[5,4] Dianna Martin,[12] Cindy Zhang,[12] Rinnat Porat,[12] Diana M. Merino,[12] Jonathan Wasserman,[12] Nada Jabado,[13] Adam Fontebasso,[13] Lars Bullinger,[14] Frank G. Rücker,[14] Konstanze Döhner,[14] Hartmut Döhner,[14] Jan Koster,[15] Jan J. Molenaar,[15] Rogier Versteeg,[15] Marcel Kool,[2] Uri Tabori,[6,12] David Malkin,[12] Andrey Korshunov,[9] Michael D. Taylor,[6,16] Peter Lichter,[2,19,*] Stefan M. Pfister,[2,5,19,*] and Jan O. Korbel[1,17,19,*]

- Stephens et al. (2011) estimated an incidence of 1.3% in all tumors, and perhaps 25% of bone cancers (by microarrays)
- Molenaar et al. (2012) estimated 11% of neuroblastoma samples (by sequencing)
- Rausch et al. (2012) estimated 13% of Medulloblastomas (by microarrays), strongly correlated with P53 loss.
- We find that 40-50% of GBM and LUSC samples have chromothripsis (by sequencing)

Summary

- Complex rearrangements are quite common in tumors.
- Many appear to be chromothripsis.
- 70% of glioblastomas have very complex rearrangements
- Fitness possibly conferred by oncogene amplification
- Origin? Prevalence? Clinical utility?

Acknowledgements

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.

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*

Funding

NHGRI: R01 HG006693-01

NIEHS: R21 ES020521-01

UVA Fund for Excellence in Science and Tech.

UVA Cancer Center Pilot Program