The BreakSeq Project

Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library

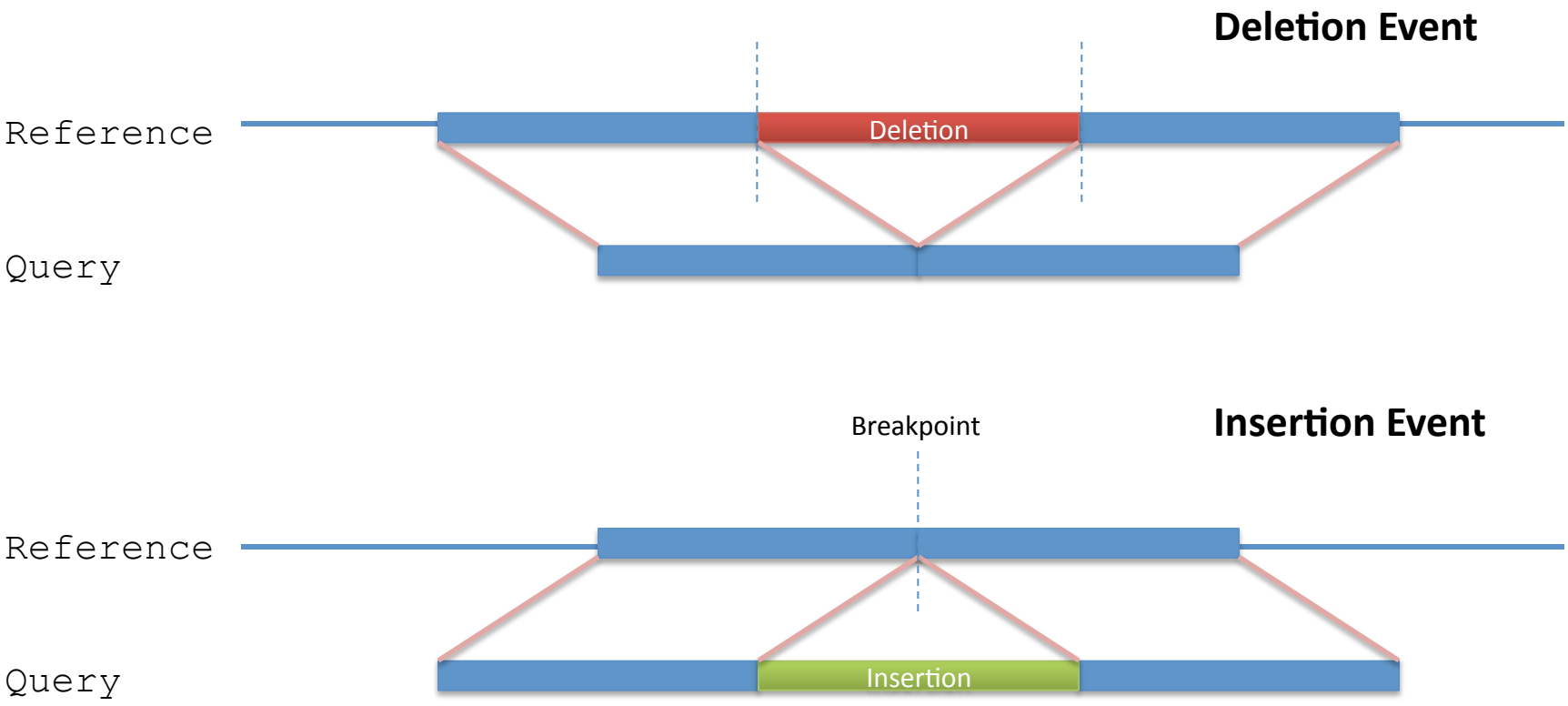
Mark Gerstein

Overview

- **Introduction**
  - SV, event type, and formation mechanism
- **The BreakSeq Analysis**
  - Analysis of SVs using a breakpoint library
- **The BreakSeq Pipeline**
  - The SV Annotation and Identification Pipeline

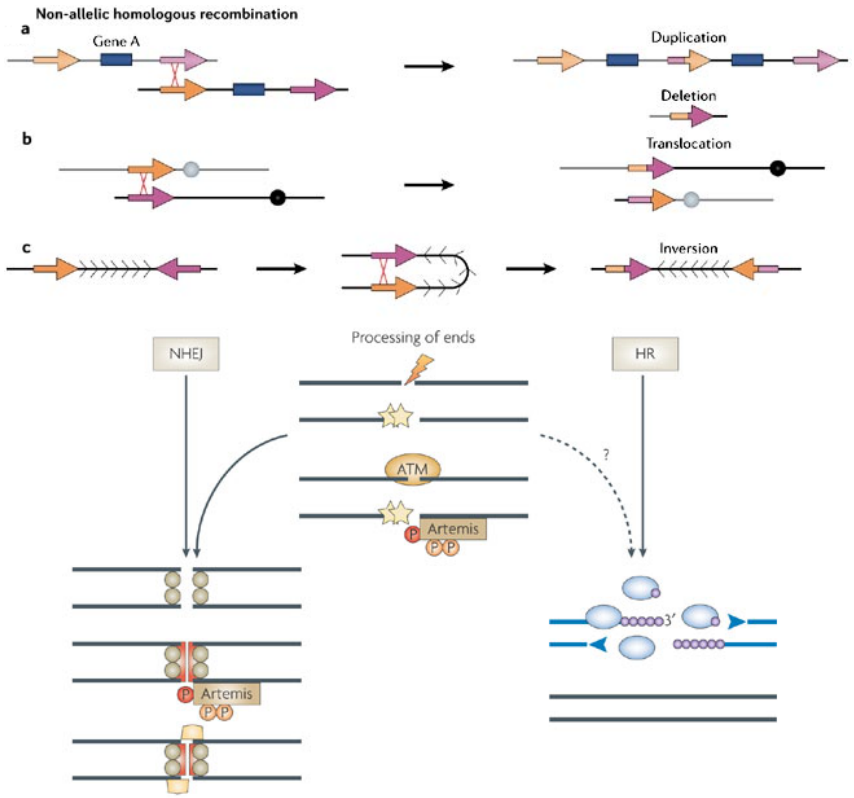
[Lam et al. Nat. Biotech. ('10)]

SV Event Type



SV formation mechanism

- Non-Allelic Homologous Recombination (**NAHR**)
- Non-homologous Recombination(**NHR**)
  - Non-homologous end joining (**NHEJ**)
  - Fork Stalling and Template Switching (**FoSTeS**)
- Transposable Element Insertion (**TEI**)
- Variable Number of Tandem Repeats (**VNTR**)



Some Issues

- Limited resolution of recent SV surveys (e.g., microarray based)
  - Prevented from intersecting with exons of genes or analyzing gene fusion events.
  - Prevented systematic deduction of the SV formation process.
  - Prevented from inferring the ancestral states of the SV events.
  - Prevented estimation of the physical properties of the SVs.

Analysis of SVs using a breakpoint library

THE BREAKSEQ ANALYSIS

Lam HY, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. “Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library”. *Nature Biotechnology* 2010 Jan;**28**(1):47-55.

SV Breakpoint Library

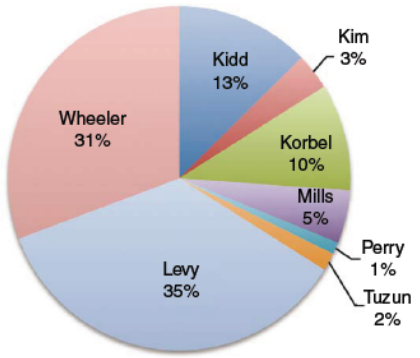
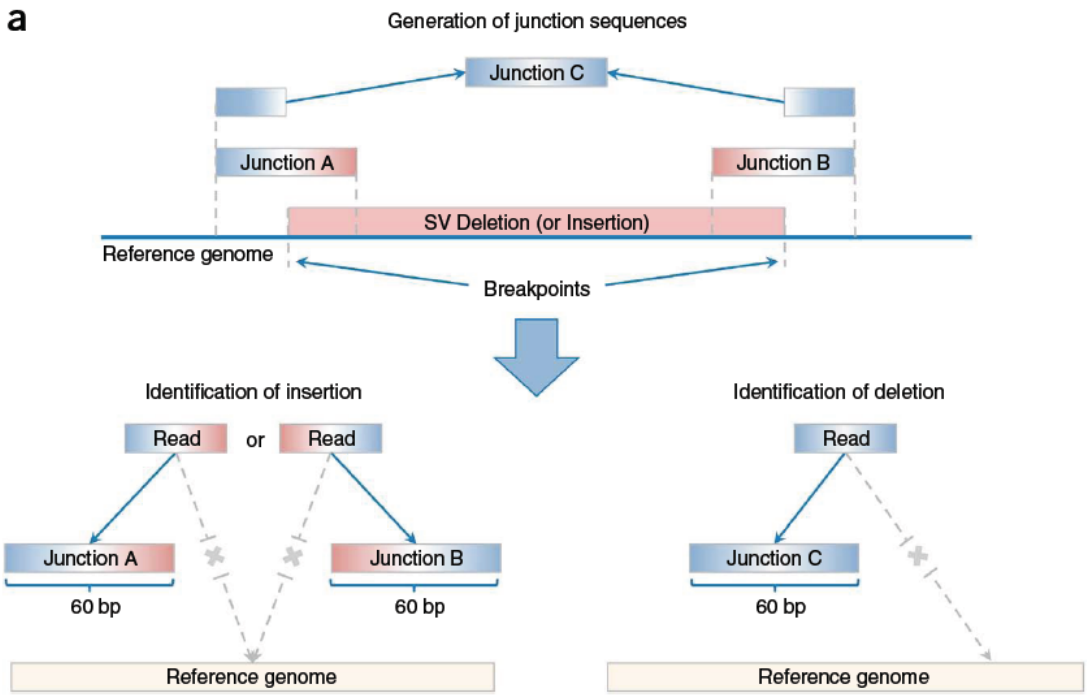


Figure 1 Composition of the SV breakpoint library. SVs in the library were based on different SV-mapping and breakpoint-sequencing strategies. A large fraction (44%) of the breakpoints were based on data generated using 454/Roche sequencing, including resequencing of an individual human genome (Wheeler[21], 602 SVs) and sequencing of breakpoints in two individuals after high-resolution and massive paired-end mapping (Korbel[5] and Kim[16], 264 SVs). The remaining 56% of the breakpoints were identified using other approaches, including Sanger capillary sequencing of breakpoints identified by whole-genome shotgun sequencing and assembly of an individual human genome (Levy[44], 694 SVs), fosmid-paired-end sequencing carried out in multiple individuals (Tuzun[3] and Kidd[6], 281 SVs), breakpoints mined from SNP discovery DNA resequencing traces (Mills[45], 98 SVs), and tiling-array-based comparative genomic hybridization followed by breakpoint sequencing (Perry[25], 22 SVs). Fewer than five breakpoints were reported in two genomes sequenced using short 36-bp reads (Illumina/Solexa)[22,23], presumably owing to the complex DNA sequence patterns frequently associated with breakpoints[5,6,25].

[Lam et al. Nat. Biotech. ('10)]

SV Junction and Identification



Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

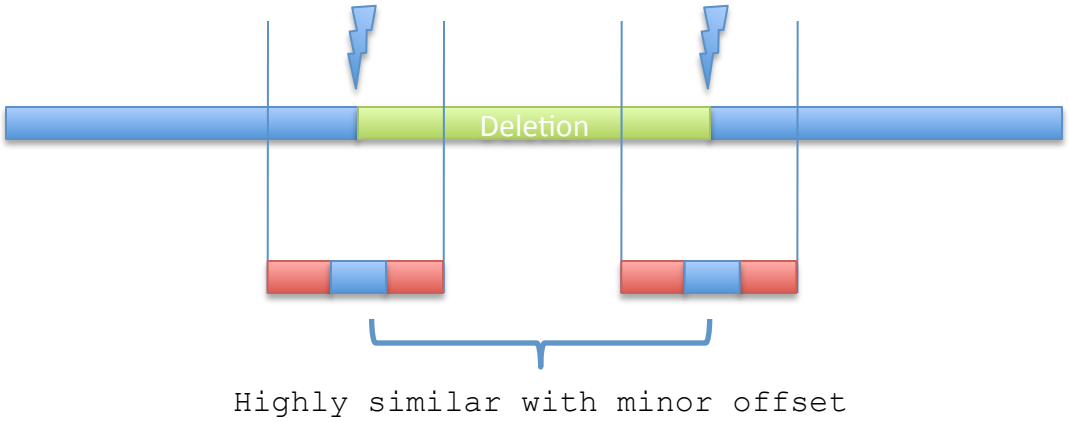
Figure 2 Mapping breakpoints using the library. (**a**) Overview of the BreakSeq approach. Breakpoints are used to generate junction sequences spanning breakpoints (upper)—the 30 bp of sequence flanking each side of the breakpoint (60 bp total). Then, DNA reads are aligned to the junction sequences (lower). Alignment results are interpreted as follows. In the case of insertions relative to the reference genome (left), sequences A and B represent the left and right breakpoint junction sequences of the nonreference SV allele, respectively. In the case of deletions (right), sequence C represents the junction sequence of the nonreference SV allele. Solid lines with arrows, successful alignments. Dashed lines with crosses, no proper alignment.

For the HCH, CEPH (NA12891) and YRI (NA18507) genomes, we identified 158, 219 and 179 SVs, respectively. 57 SVs were shared between the YRI and HCH genomes, 62 between the YRI and NA12891 genomes, 52 between the HCH and NA12891 genomes, and 42 were common to all three genomes.

[Lam et al. Nat. Biotech. ('10)]

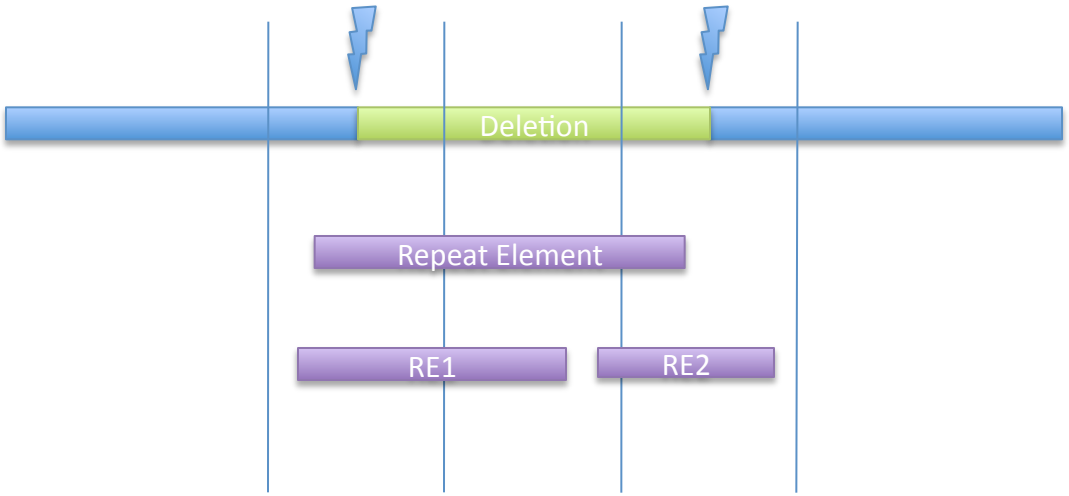
Mechanism Classification

NAHR



Single RETRO

Multiple RETRO



[Lam et al. Nat. Biotech. ('10)]

SV Mechanism Classification

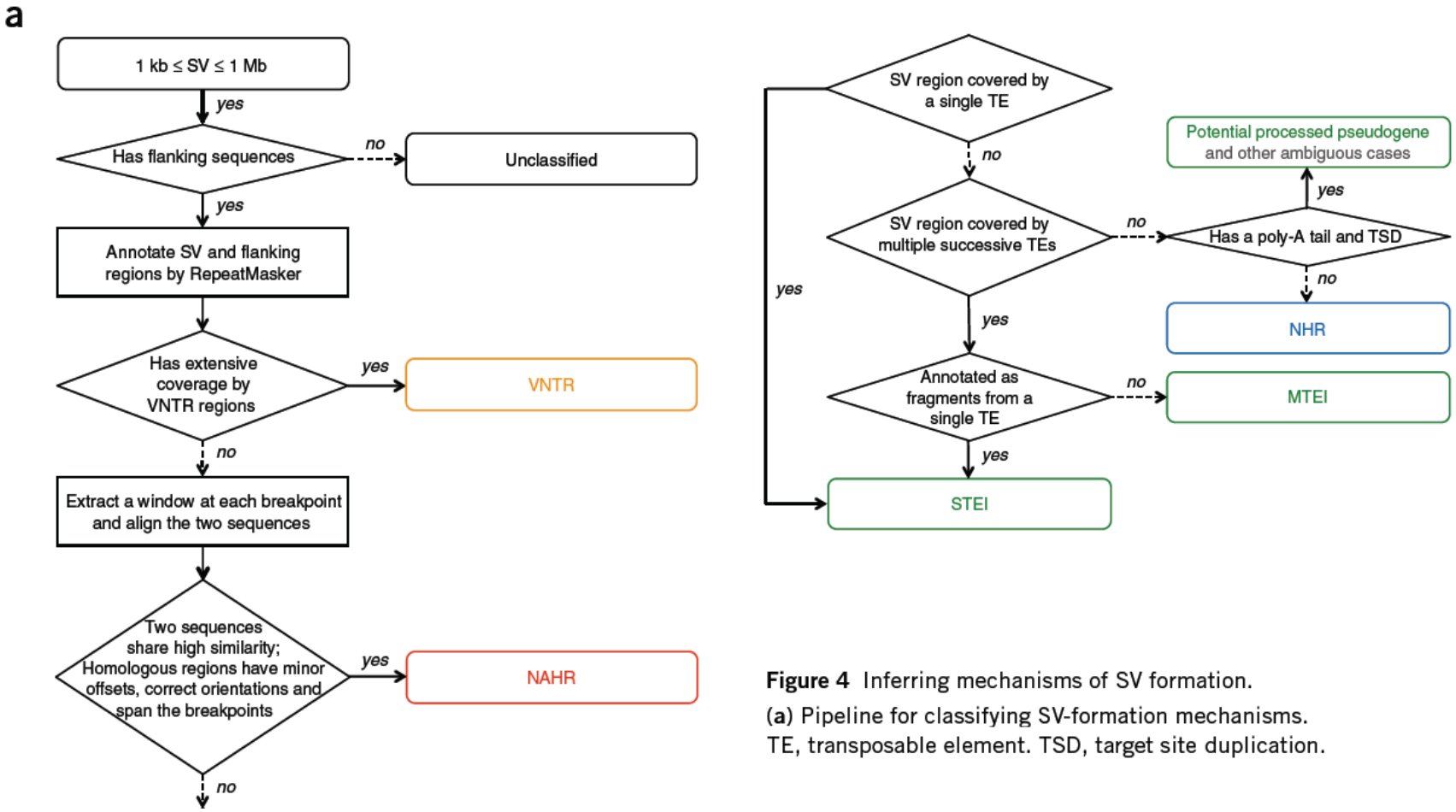
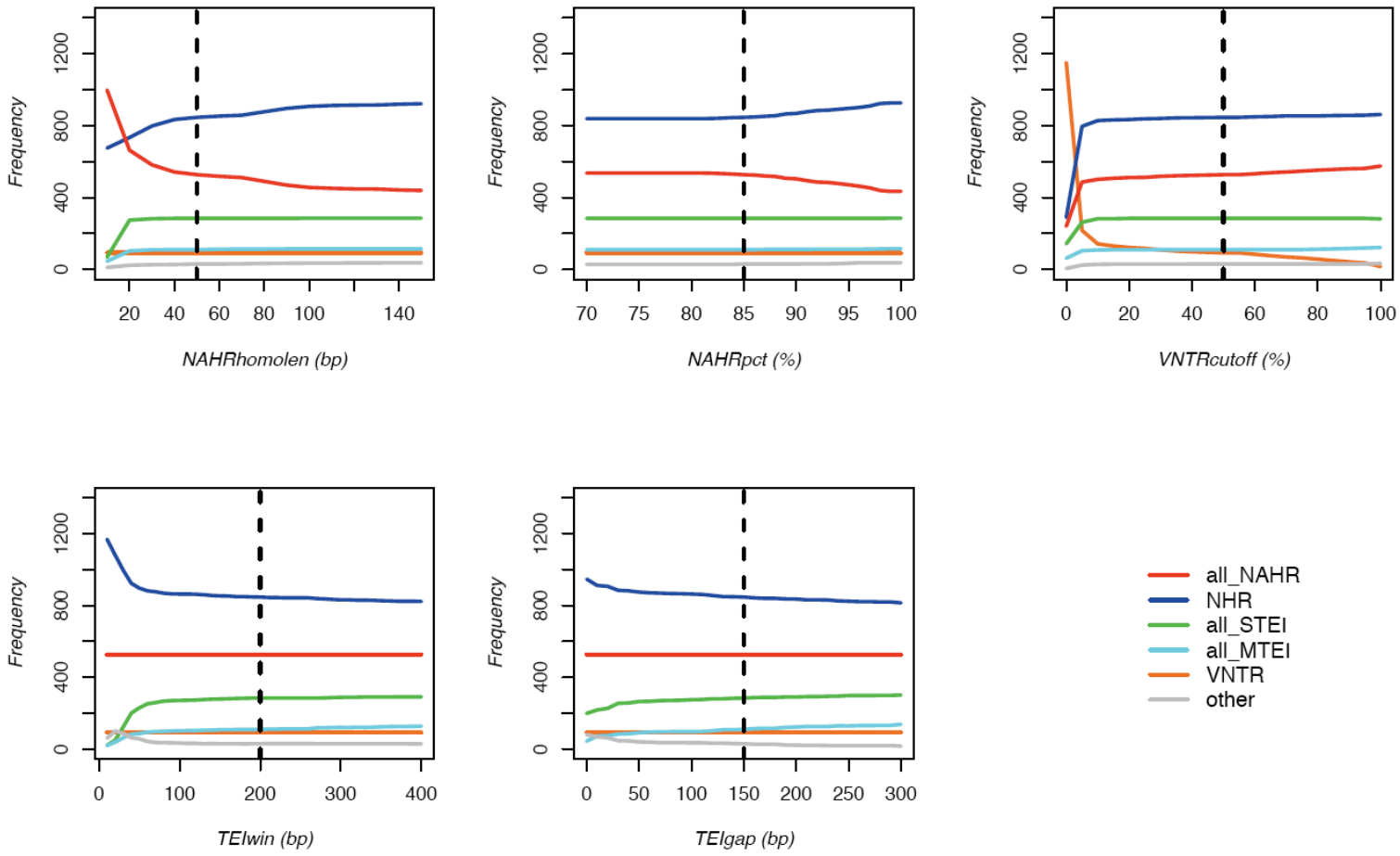


Figure 4 Inferring mechanisms of SV formation.
(**a**) Pipeline for classifying SV-formation mechanisms.
TE, transposable element. TSD, target site duplication.

[Lam et al. Nat. Biotech. ('10)]

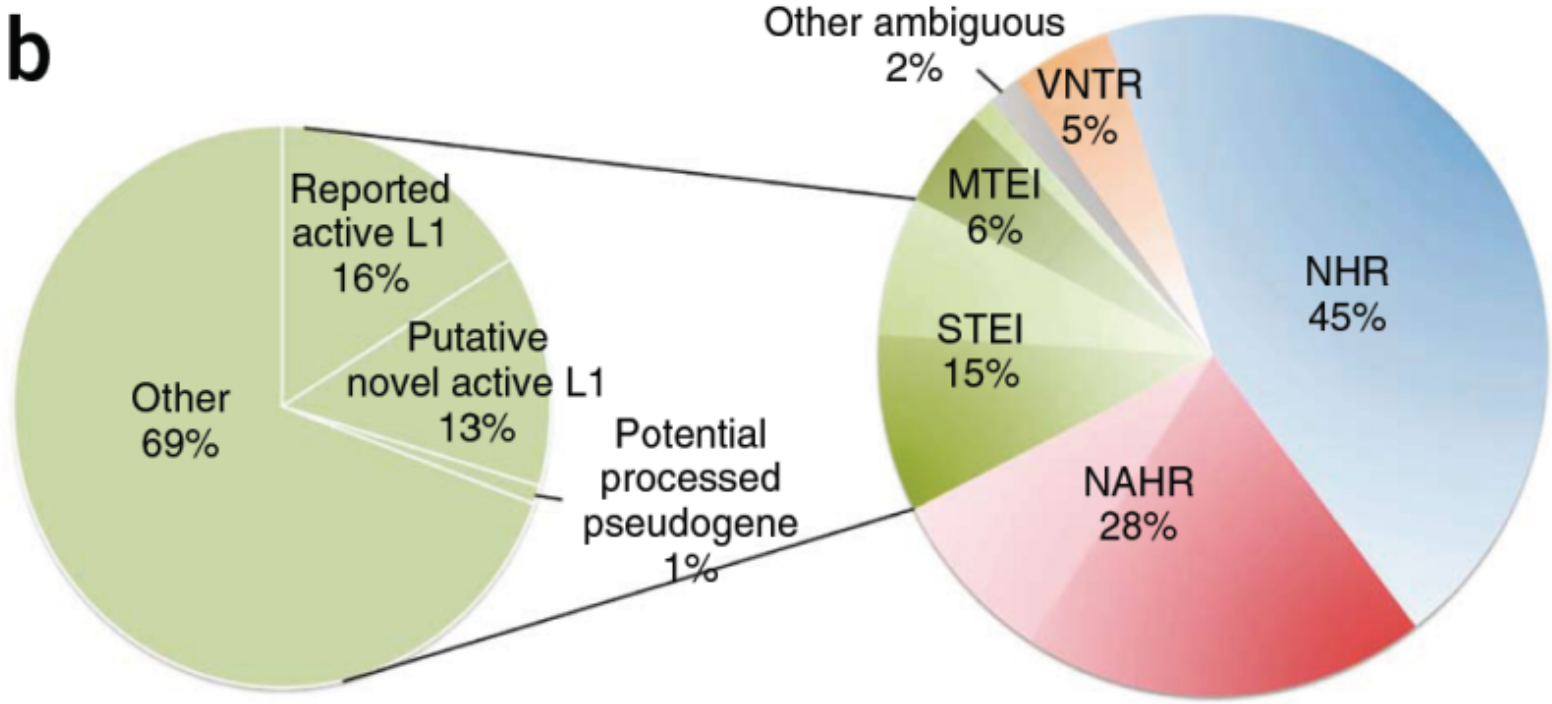
Sensitivity analysis of the classification pipeline



[Lam et al. Nat. Biotech. ('10)]

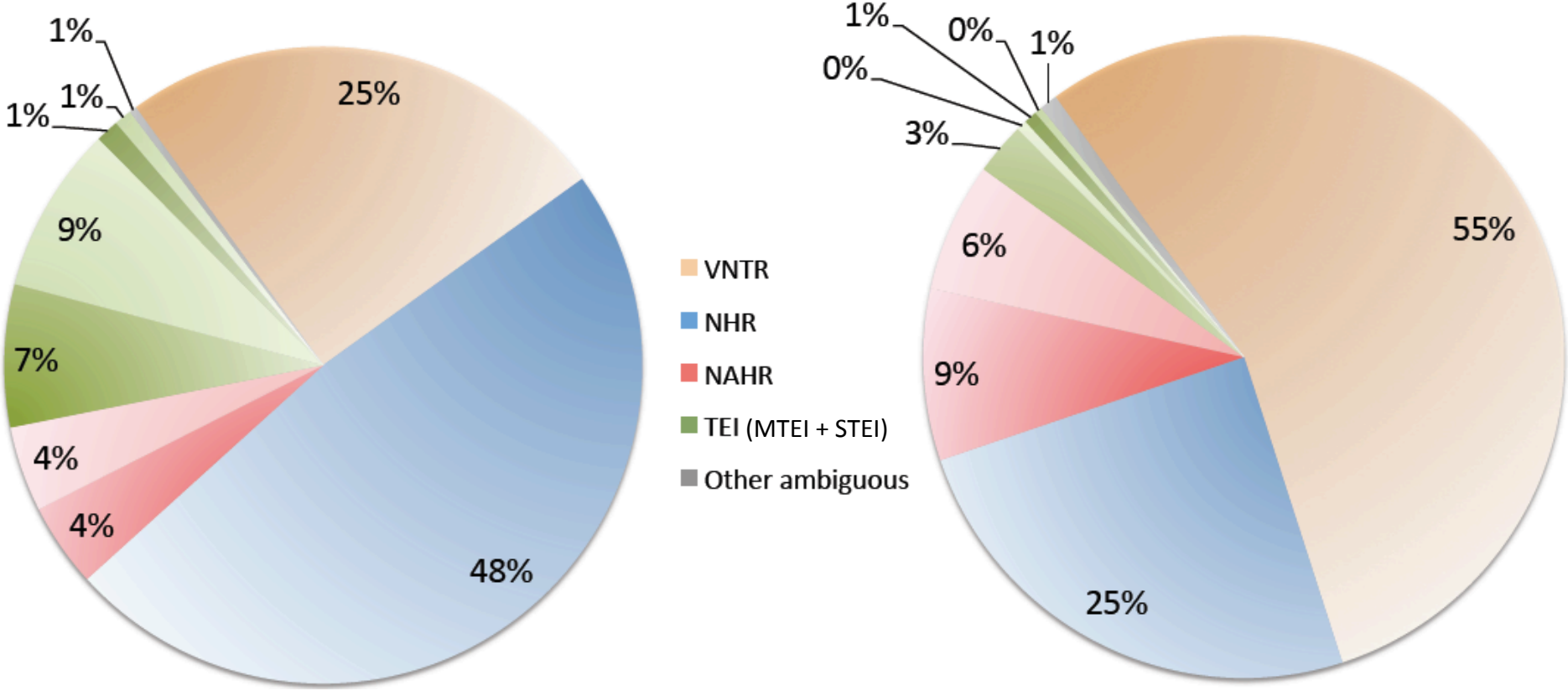
x-axis is the parameter space. y-axis is the number of SVs of different formation mechanisms classified by the pipeline using corresponding value of the varied parameter and default values of other parameters. Dotted vertical lines indicate the default parameters.

SV Formation Analysis



[Lam et al. Nat. Biotech. ('10)]

Formation mechanisms of SVs identified in the 1000 genomes project: split reads



16128 Yale SR from Zhengdong Zhang, NA12878, Aug 2009 version, >=200bp

4285 Yale SR from Zhengdong Zhang, NA12878, Aug 2009 version, >=1kb

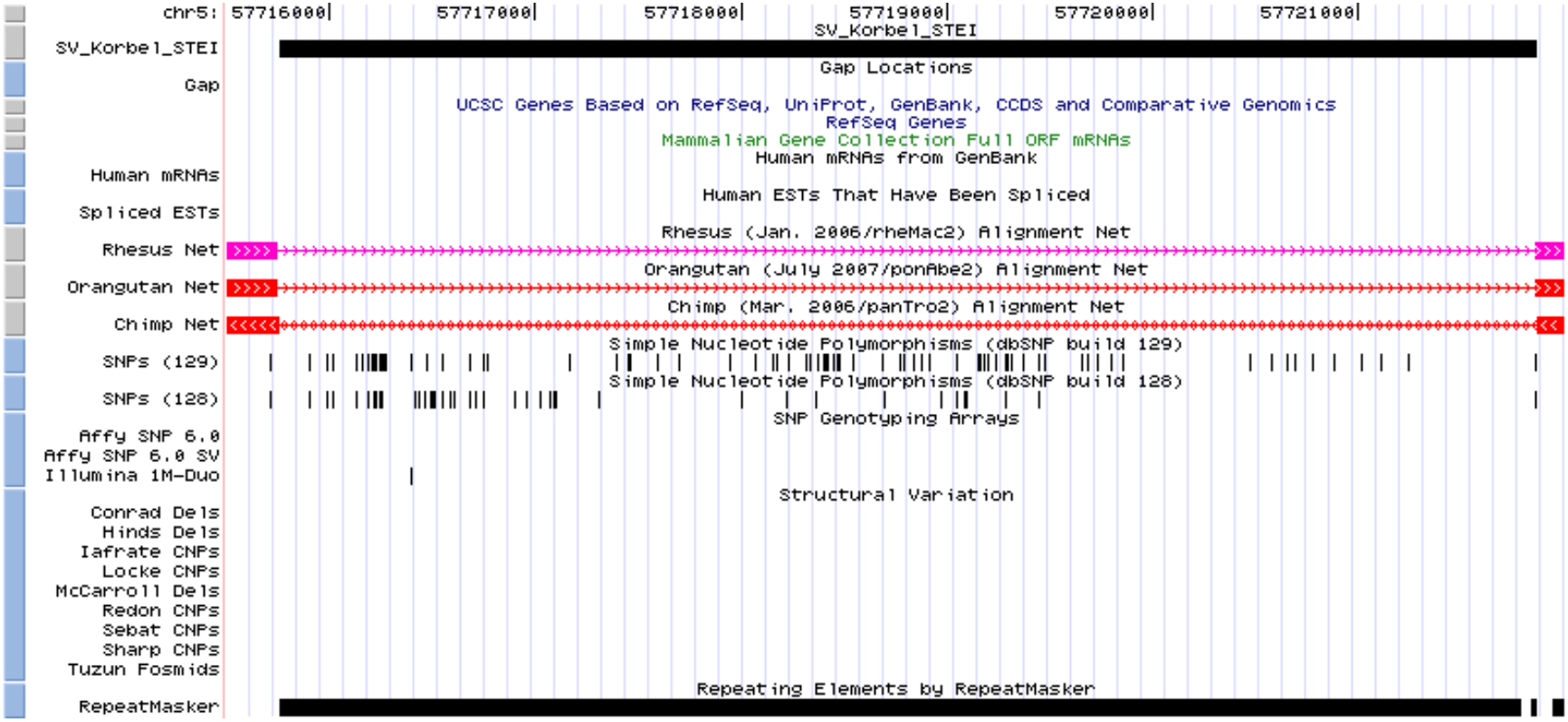
Active L1 Transposition

431 fully rectifiables overlapped with 147 Active L1s by Mills et al. 2007 consolidated from Brouha et al. 2003 and Mills et al. 2006

Chr	Source	Event	Start	End	Size	Mech	Active L1	Supported
chr1	Korbel	Insertion	84290516	84297219	6703	Mech "MTEI"; Rectified "2:2:2"	chr1:84290591-84296677['L1HS', 'Ta-1d']	2
chr1	Korbel	Insertion	245917096	245923148	6052	Mech "STEI"; Rectified "2:2:2"	chr1:245917098-245923129['L1HS', 'Ta-0']	3
chr10	Korbel	Insertion	5277306	5283354	6048	Mech "UNSURE"; Rectified "2:2:2"	chr10:5277317-5283348['L1HS', 'Ta-1dn(g)']	1
chr11	Korbel	Insertion	24306070	24312135	6065	Mech "STEI"; Rectified "2:2:2"	chr11:24306073-24312103['L1HS', 'Ta-1d']	1
chr11	Korbel	Deletion	92791150	92800593	9443	Mech "NAHR"; Rectified "1:1:1"	chr11:92793800-92799845['L1HS', 'Ta-1d']	1
chr11	Venter	Insertion	92793799	92799859	6060	Mech "STEI"; Rectified "2:2:2"	chr11:92793800-92799845['L1HS', 'Ta-1d']	1
chr11	Watson	Insertion	94809017	94815068	6051	Mech "UNSURE"; Rectified "2:2:2"	chr11:94809028-94815058['L1HS', 'Ta-1d']	1
chr15	Venter	Insertion	53005523	53011731	6208	Mech "MTEI"; Rectified "2:2:2"	chr15:53005558-53011589['L1HS', 'Ta-0']	3
chr15	Kim	Insertion	68808908	68814562	5654	Mech "MTEI"; Rectified "2:2:2"	chr15:68809138-68814556['L1HS', 'L1HS']	2
chr18	Korbel	Insertion	46124318	46130363	6045	Mech "STEI"; Rectified "2:2:2"	chr18:46124336-46130355['L1HS', 'Pre-Ta (ACG/G)']	1
chr2	Venter	Insertion	176054929	176060981	6052	Mech "STEI"; Rectified "2:2:2"	chr2:176054939-176060968['L1HS', 'Ta-1d']	1
chr20	Venter	Insertion	7044794	7050858	6064	Mech "STEI"; Rectified "2:2:2"	chr20:7044828-7050846['L1HS', 'Ta-0']	4
chr4	Watson	Insertion	59627149	59633191	6042	Mech "UNSURE"; Rectified "2:2:2"	chr4:59627160-59633190['L1HS', 'L1HS']	1
chr5	Venter	Insertion	57715759	57721867	6108	Mech "STEI"; Rectified "2:2:2"	chr5:57715758-57721790['L1HS', 'Ta-0']	3
chr5	Venter	Insertion	103882188	103888239	6051	Mech "STEI"; Rectified "2:2:2"	chr5:103882187-103888216['L1HS', 'Ta-1d']	3
chr5	Watson	Insertion	108622973	108629020	6047	Mech "UNSURE"; Rectified "2:2:2"	chr5:108622987-108629018['L1HS', 'Ta-1d']	1
chr6	Venter	Insertion	133383514	133389578	6064	Mech "STEI"; Rectified "2:2:2"	chr6:133383548-133389578['L1HS', 'Ta-1d']	3
chr7	Venter	Insertion	113203413	113209458	6045	Mech "STEI"; Rectified "2:2:2"	chr7:113203413-113209443['L1HS', 'Ta-1d']	4
chr8	Venter	Insertion	73950330	73956387	6057	Mech "STEI"; Rectified "2:2:2"	chr8:73950346-73956377['L1HS', 'Ta-1d']	4
chr8	Venter	Insertion	126664312	126670324	6012	Mech "STEI"; Rectified "2:2:2"	chr8:126664312-126670315['L1HS', 'Ta-1d']	5
chr8	Korbel	Insertion	135152107	135158208	6101	Mech "STEI"; Rectified "2:2:2"	chr8:135152168-135158198['L1HS', 'L1HS']	3
chrX	Venter	Insertion	11863121	11869370	6249	Mech "STEI"; Rectified "2:2:2"	chrX:11863128-11869354['L1HS', 'Ta-1d']	2
chrX	Venter	Insertion	95199436	95205519	6083	Mech "STEI"; Rectified "2:2:2"	chrX:95199466-95205497['L1HS', 'Ta-0']	1

[Lam et al. Nat. Biotech. ('10)]

Active L1 Transposition Example

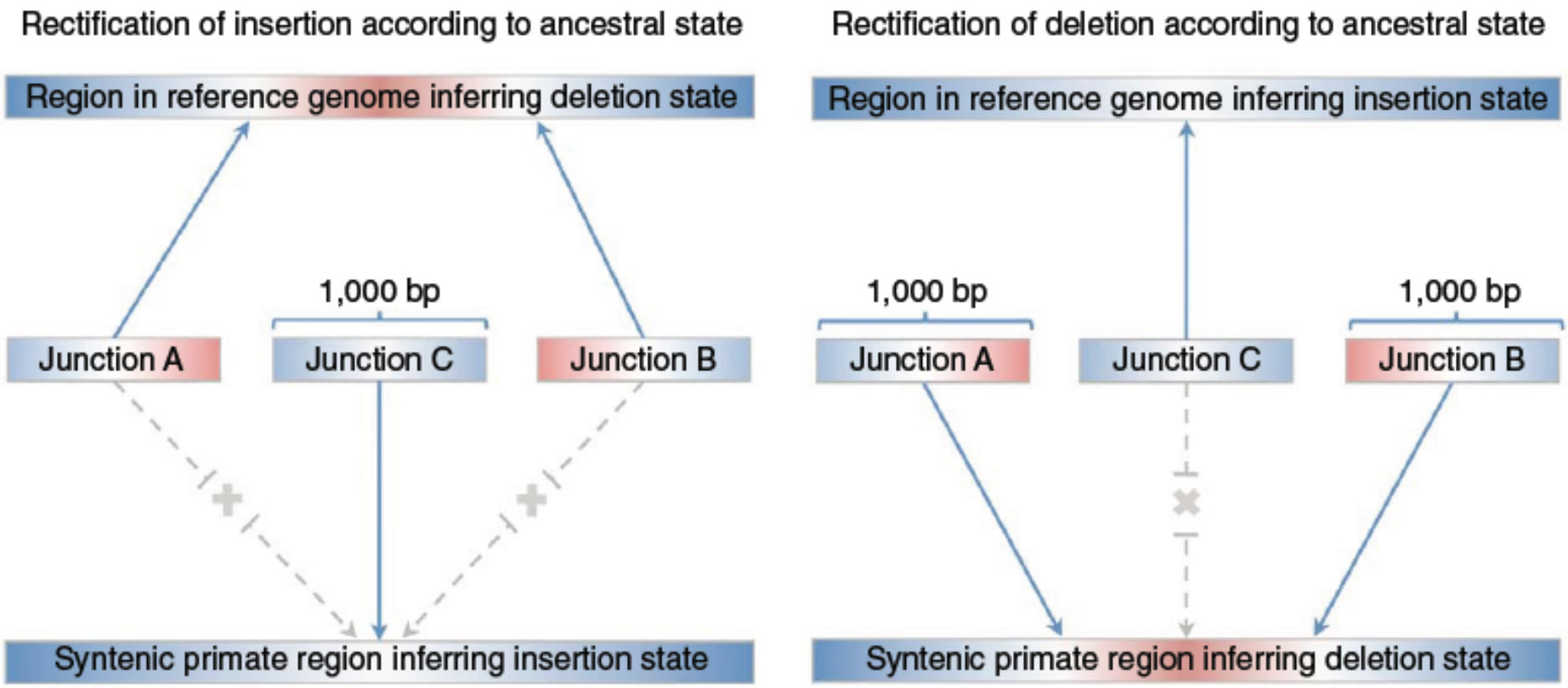


Pseudogene Number Variation

431 fully rectifiables overlapped with 13,453 duplicated and processed pseudogenes identified by PseudoPipe based on Ensembl 48

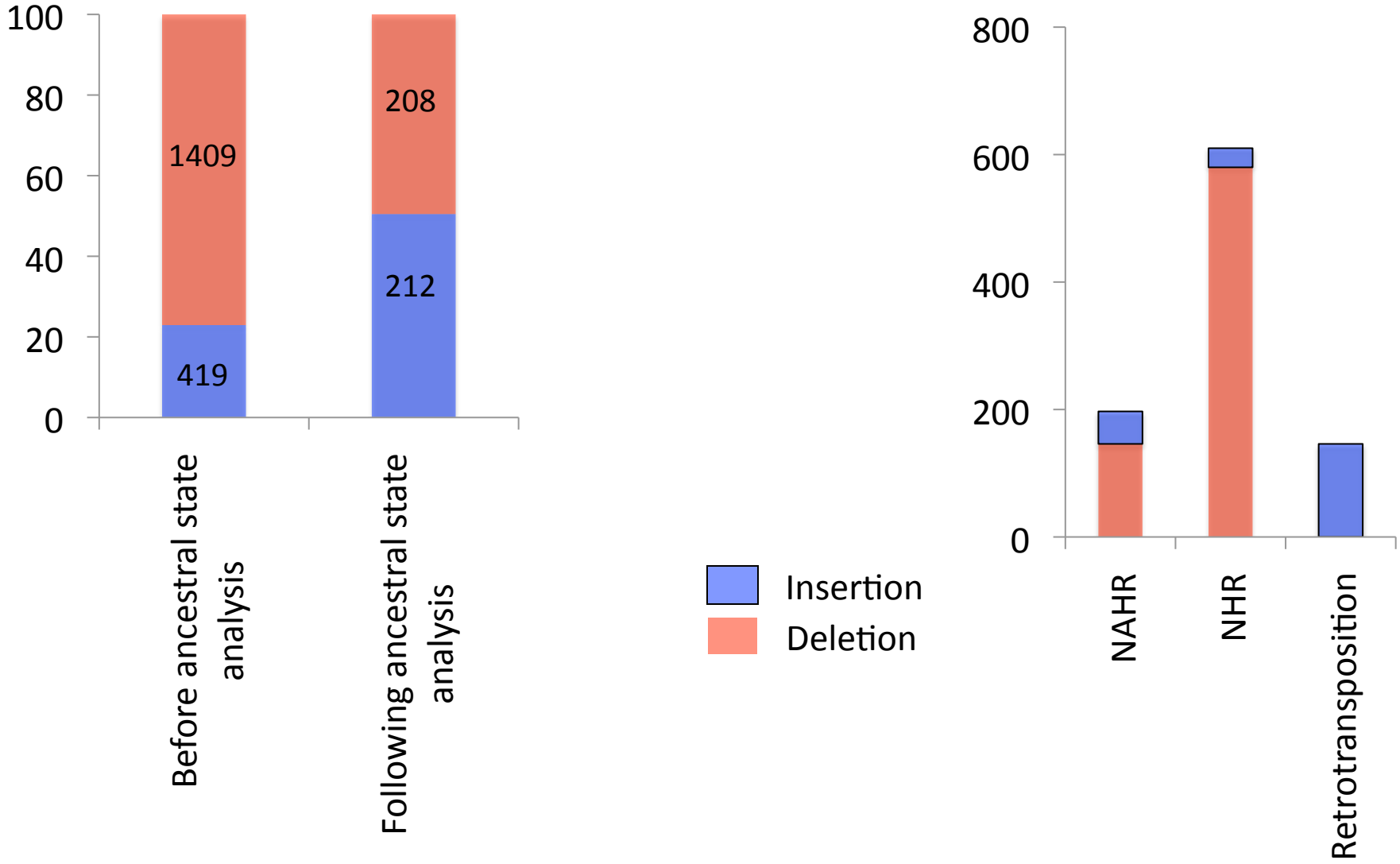
Chr	Source	Event	Start	End	Size	Mech	Pgene Type
chr10	Kidd	Deletion	100678090	100692331	14241	Mech "NAHR"; Rectified "1:1:1"	PSSD
chr12	Venter	Deletion	22467006	22473645	6639	Mech "NAHR"; Rectified "1:1:1"	PSSD
chr17	Kidd	Deletion	65603123	65859003	255880	Mech "NHR"; Rectified "1:1:1"	PSSD
chr20	Kidd	Deletion	1503149	1536176	33027	Mech "NAHR"; Rectified "1:1:1"	PSSD
chr3	Korbel	Deletion	74230280	74237487	7207	Mech "NHR"; Rectified "1:1:1"	PSSD
chr5	Watson	Deletion	64538468	64548395	9927	Mech "NHR"; Rectified "1:1:1"	DUP
chr5	Kidd	Insertion	69544715	69817387	272672	Mech "NAHR"; Rectified "2:2:2"	DUP/PSSD
chrX	Kidd	Deletion	47752047	47874915	122868	Mech "NAHR"; Rectified "1:1:1"	PSSD

SV Ancestral State Analysis



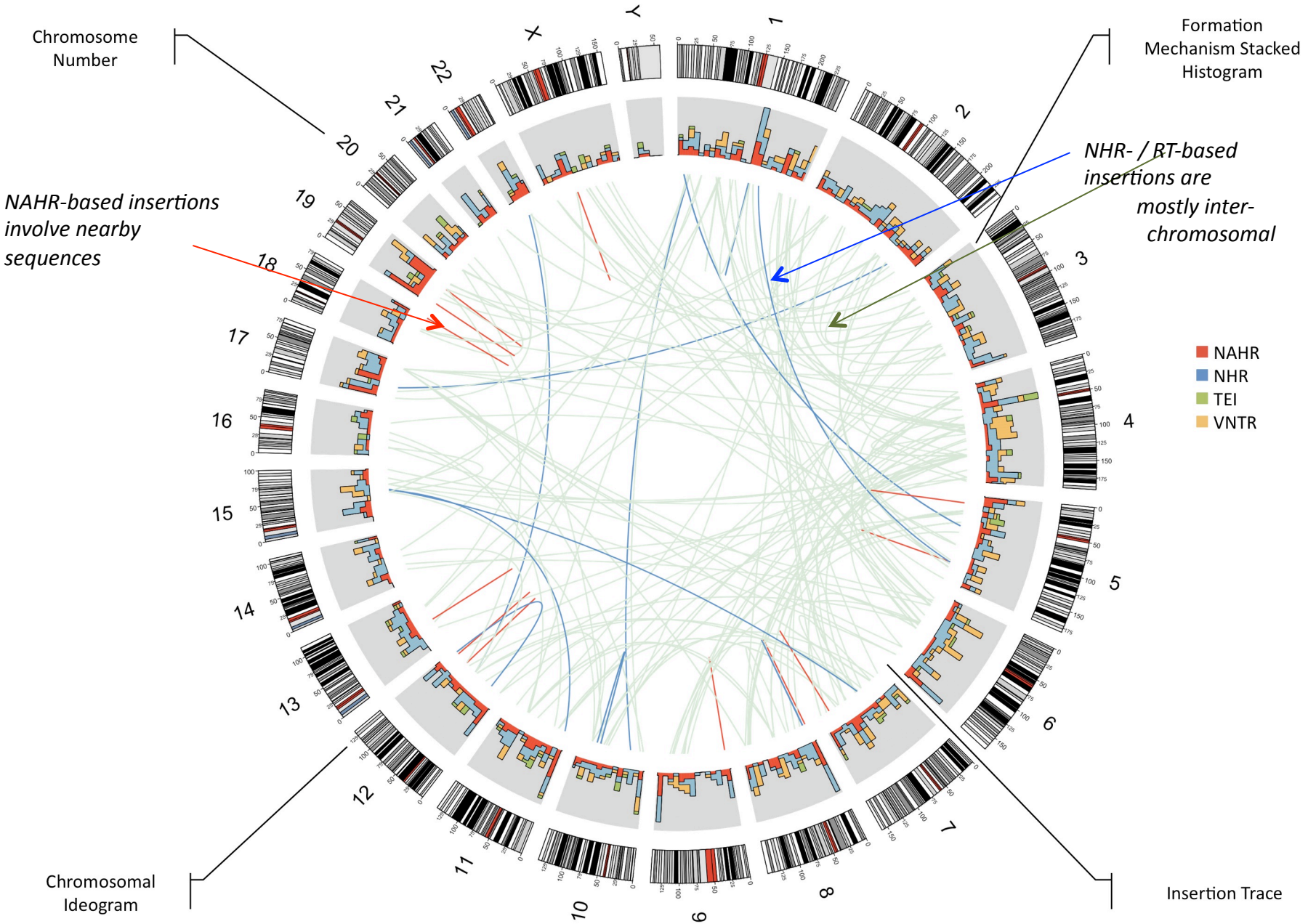
[Lam et al. Nat. Biotech. ('10)]

Ancestral state analysis reveals balance of insertions and deletions, and biases in formation mechanisms



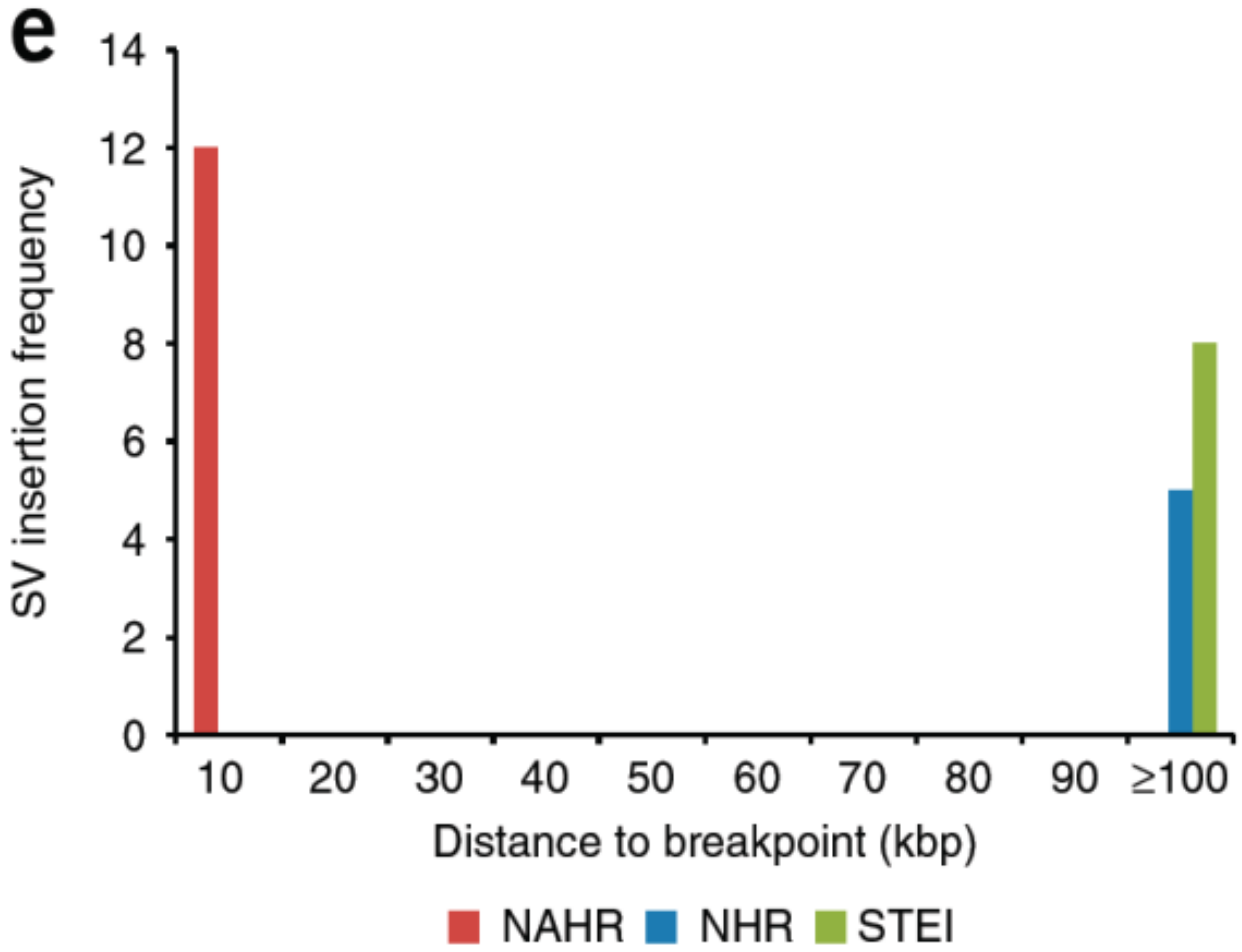
[Lam et al. Nat. Biotech. ('10)]

Tracing the origin of recent human insertions



[Lam et al. Nat. Biotech. ('10)]

Relative location of Inserted Sequence

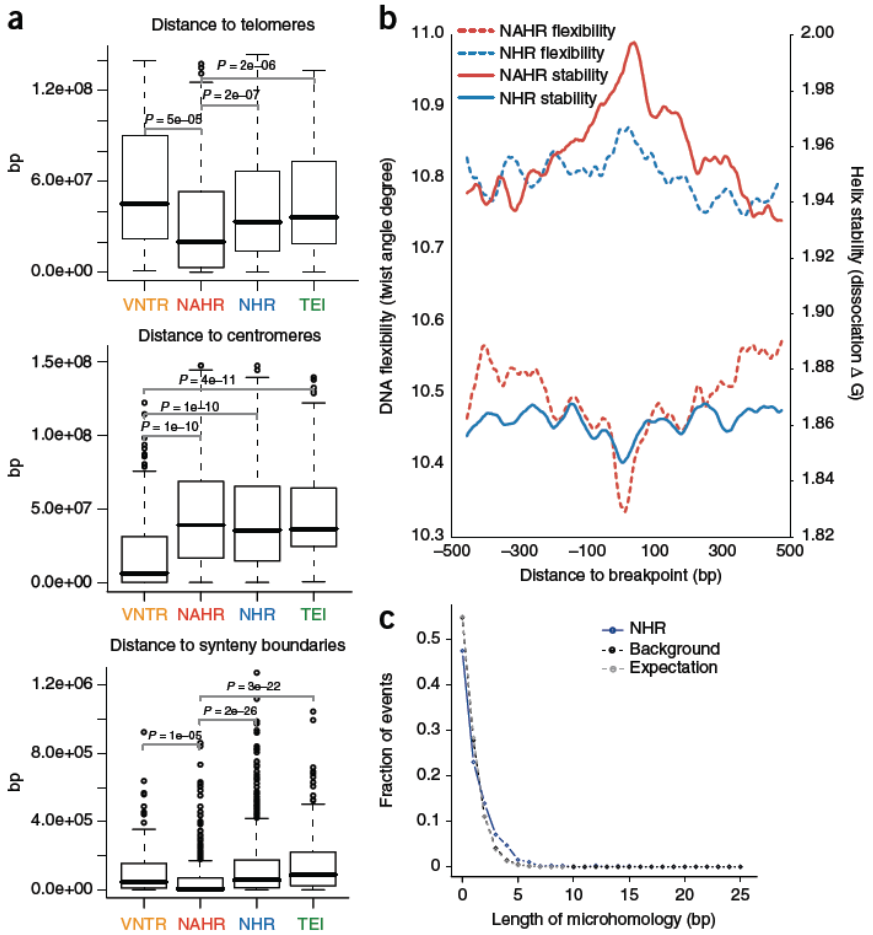


[Lam et al. Nat. Biotech. ('10)]

Breakpoint Features Analysis

Figure 5 Analysis of breakpoint features. (**a**) Distance to chromosomal landmarks. Brackets indicate significantly different classes ($P < 0.05$ in Wilcoxon rank sum test after multiple hypothesis test correction by the Holm method). NAHR events are found to be significantly closer to telomeres and human-chimpanzee synteny block boundaries than the other mechanistic classes; VNTRs are significantly enriched in centromeric and pericentromeric regions. (**b**) DNA flexibility (dashed lines and left *y*-axis) and helix stability (solid lines and right *y*-axis) around NAHR and NHR breakpoints. (**c**) Distribution of NHR events with different lengths of microhomologies at the breakpoints. Microhomologies are significantly enriched in NHR breakpoints compared to a random background (KS test, $P = 2.43E\text{-}11$).

We analyzed various features related to the physical properties of DNA at SV breakpoint junctions. In contrast to NHR, NAHR events were found to be biased toward GC-rich regions (**Supplementary Table** 5). A possible explanation for this bias is the known GC-richness of recombination hotspots[30], which we found to be significantly (P = 2.96E–03) enriched for NAHR events. Further, our results may indicate SV formation biases owing to DNA duplex stability. We thus extended our analyses by two additional features: DNA helix stability predicted by calculating the average of the dissociation free energy of each overlapping dinucleotide[31], and DNA flexibility based on the calculation of the average of the twist angle among each overlapping dinucleotide[32]. Our results indicate that in contrast to NAHR, NHR events are associated with high DNA flexibility and low helix stability, both of which are believed to be markers of fragility[33].

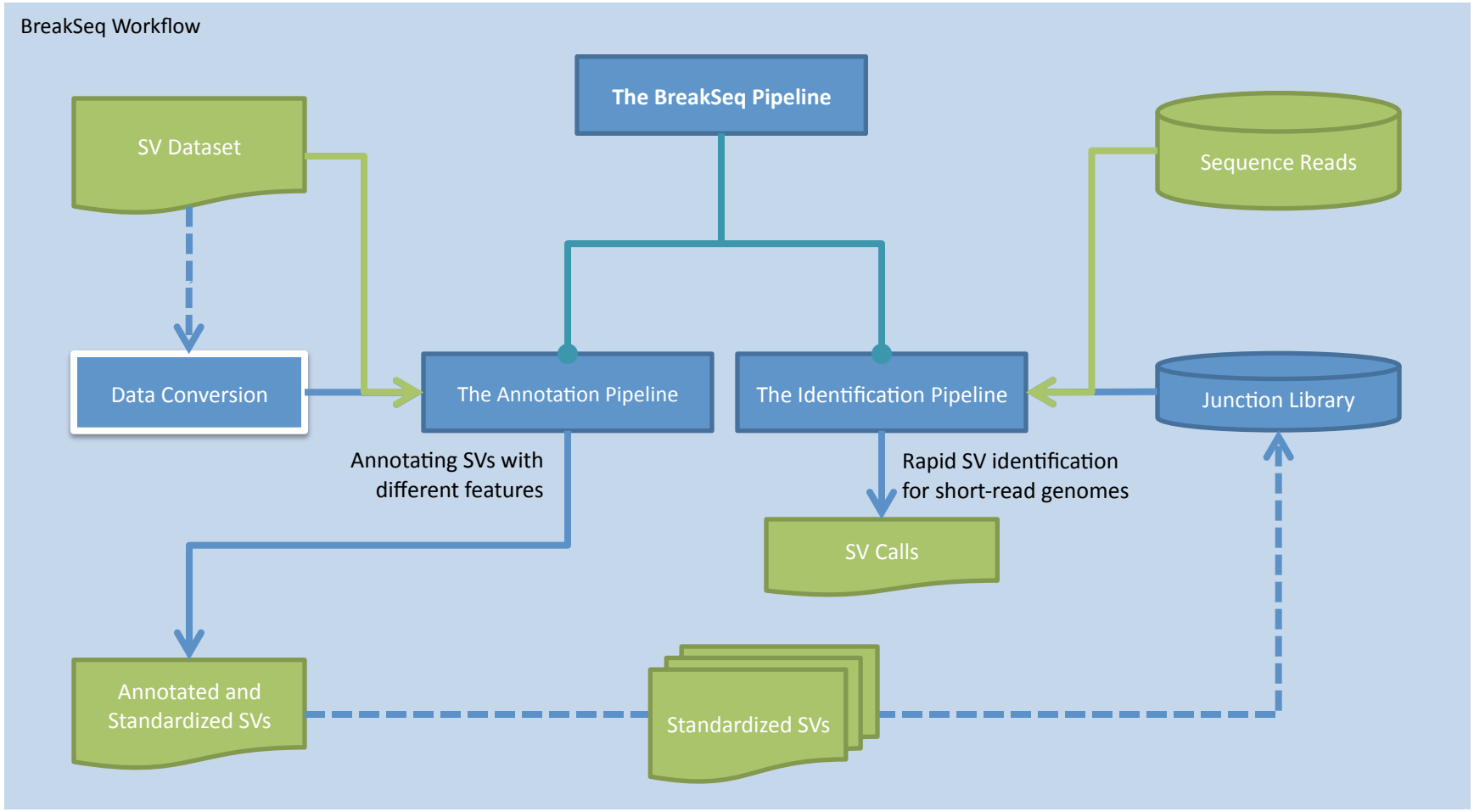


[Lam et al. Nat. Biotech. ('10)]

The SV Annotation and Identification Pipeline

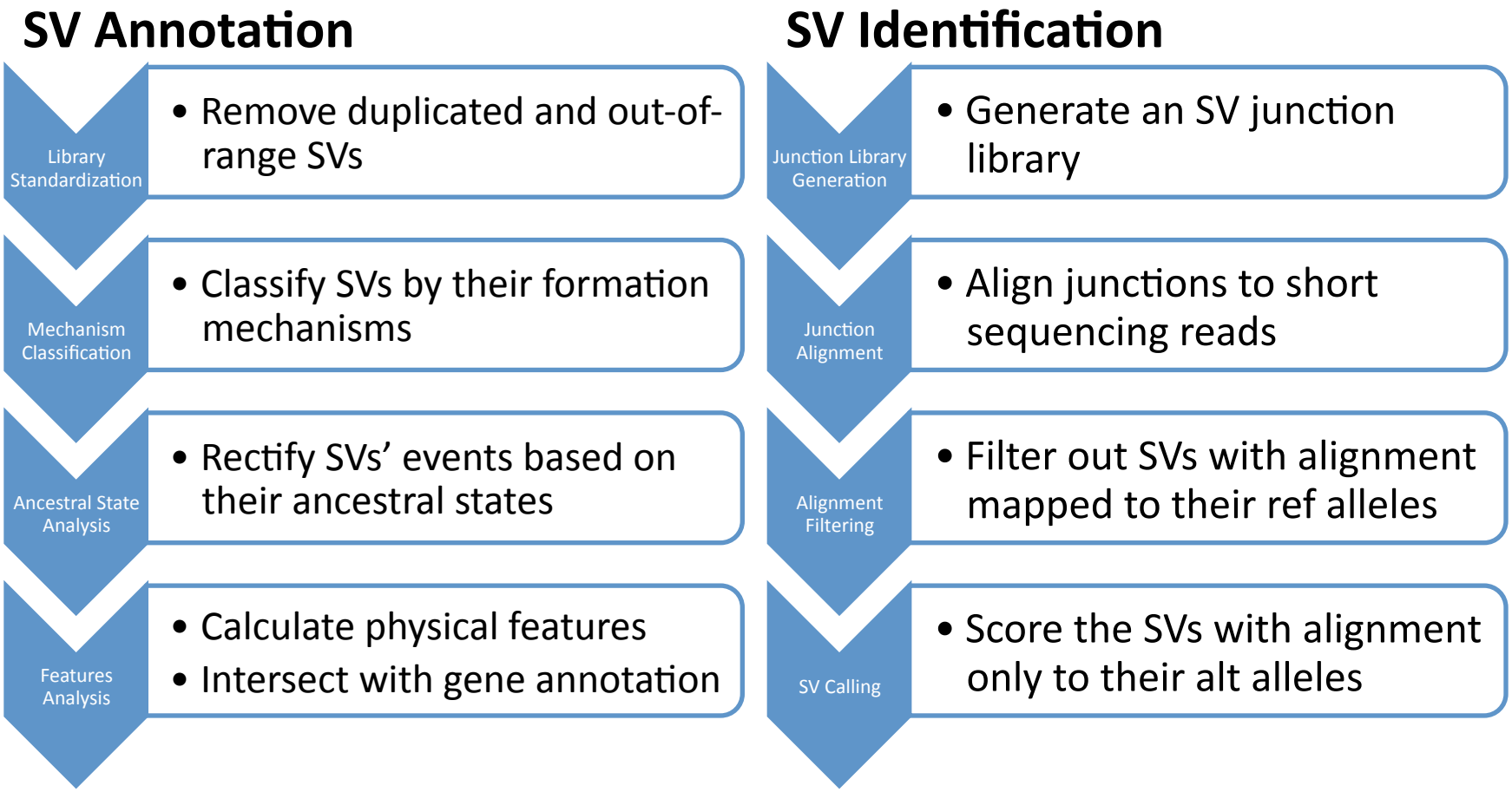
THE BREAKSEQ PIPELINE

The Pipeline Workflow



[Lam et al. Nat. Biotech. ('10)]

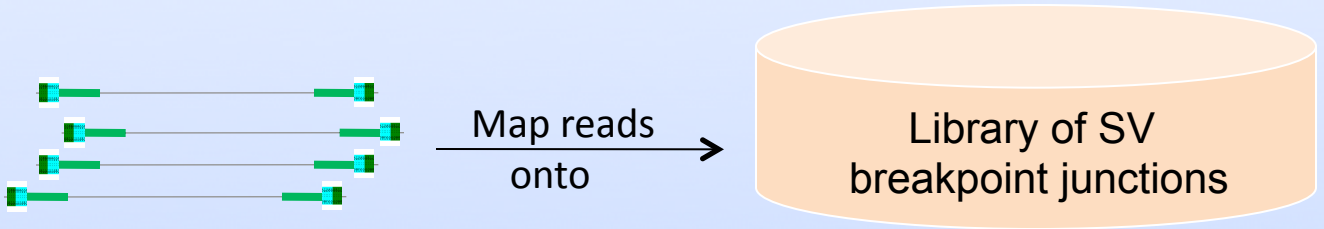
The Pipeline Modules

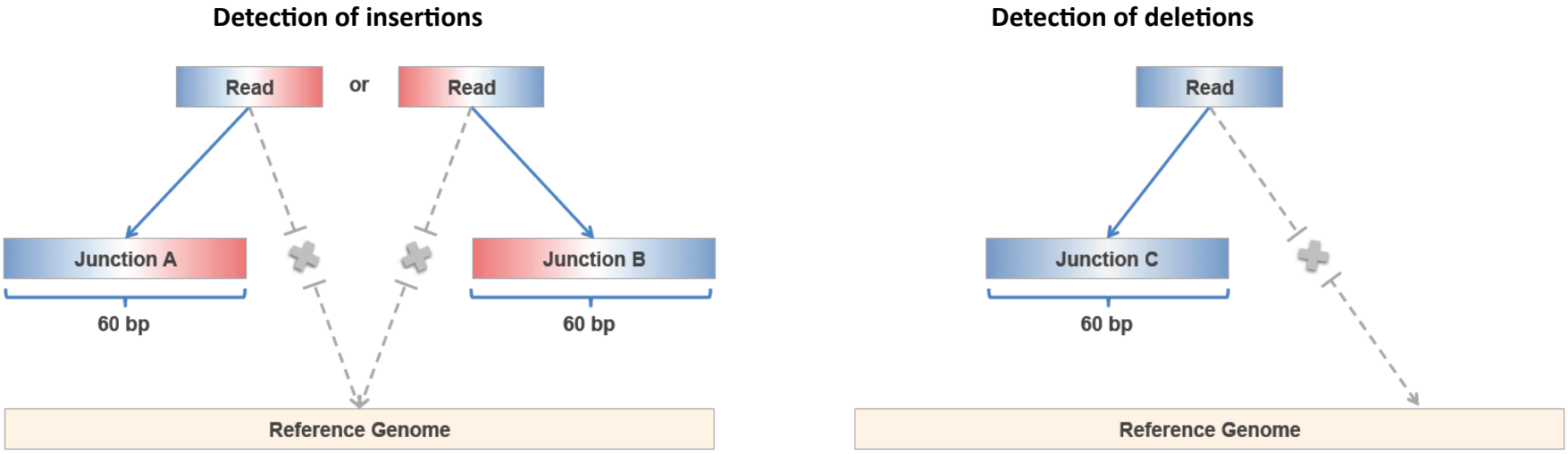


[Lam et al. Nat. Biotech. ('10)]

BreakSeq enables detecting SVs in Next-Gen Sequencing data based on breakpoint junctions

Leveraging read data to identify previously known SVs ("Break-Seq")





* Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

[Lam et al. Nat. Biotech. ('10)]

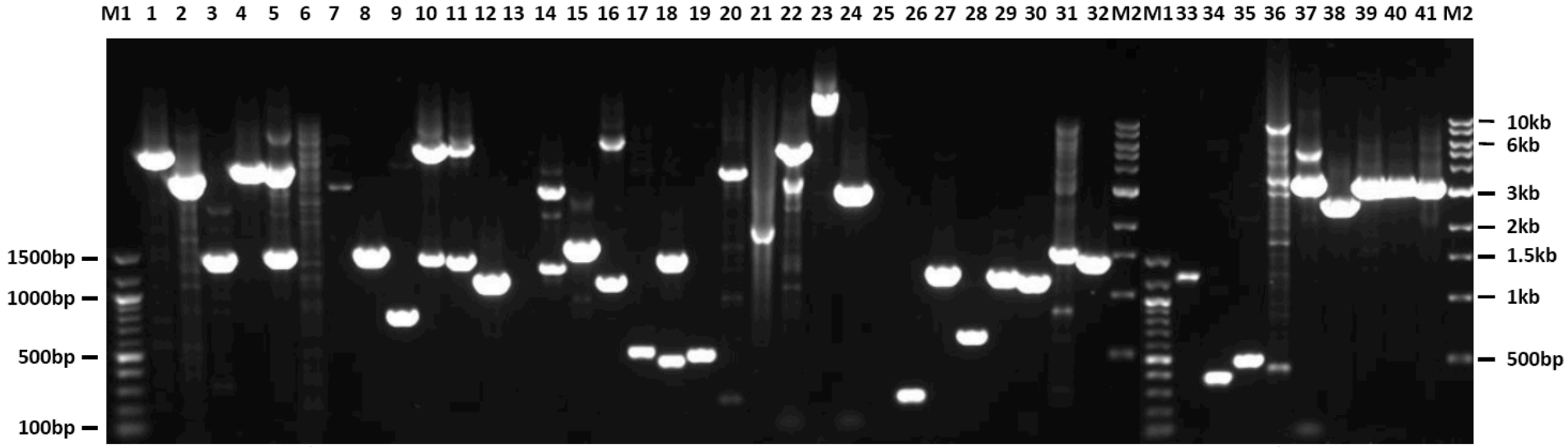
Applying BreakSeq to short-read based personal genomes boosts numbers of bp-level SVs by ~50-fold

Personal genome (ID)	Ancestry	High support hits (>4 supporting hits)	Total hits (incl. low support)
NA18507*	Yoruba	105	179
YH*	East Asian	81	158
NA12891 [1000 Genomes Project, CEU trio]	European	113	219

***According to the operational definition we used in our analysis (>1kb events) less than 5 SVs were previously reported in these genomes …**

[Lam et al. Nat. Biotech. ('10)]

PCR validations in NA12891 demonstrate high accuracy of BreakSeq and add 48 validated calls to the CEU trio



48 positive outcomes out of 49 PCRs that were scored in NA12891: 98% PCR validation rate (for low and high-support events)

12 amplicons sequenced in NA12891: all breakpoints confirmed

Adrian Stütz

[Lam et al. Nat. Biotech. ('10)]

Acknowledgement

- **Yale University**
  - Jasmine Mu
  - Hugo Lam
- **Stanford U.**
  - M Snyder
- **University of Toronto**
  - Philip Kim
- **EMBL**
  - Jan Korbel
  - Adrian Stuetz
- **University of Vienna**
  - Andrea Tanzer

More Information on this Talk

SUBJECT: Assembly

DESCRIPTION:
Computational Biology Center, IBM T J Watson Research Center, Yorktown Heights, NY, 2010.02.11, 11:00-12:00; [I:**IBM**] (Takes 25' with many questions questions.)

MORE DESCRIPTION:
Talk works equally well on mac or PC. Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers "ID" on the site. For instance,
the topic **pubnet*** can be looked up at
http://papers.gersteinlab.org/papers/pubnet)

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: **http://www.flickr.com/photos/mbgmbg/tags/kwpotppt** .