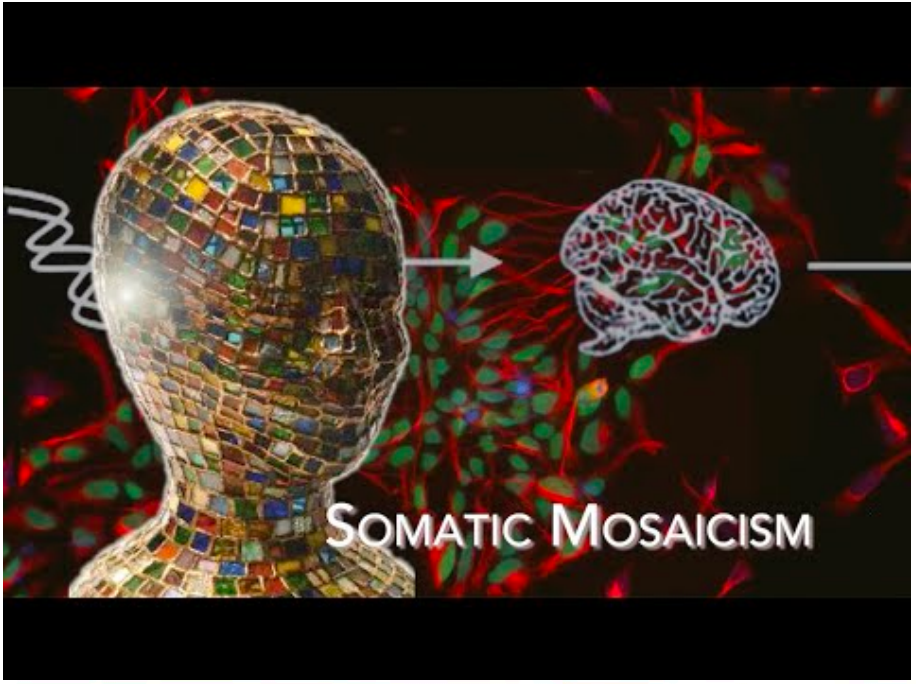
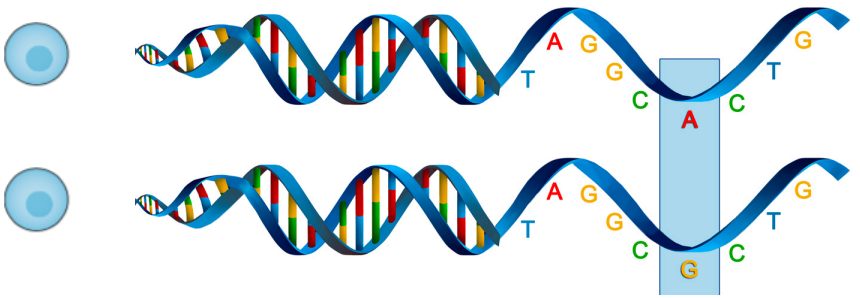
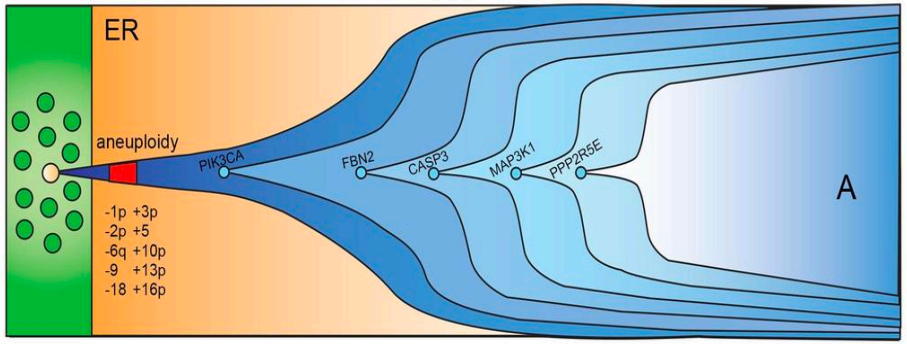
Single-cell genomics data analysis focusing on Strand-seq

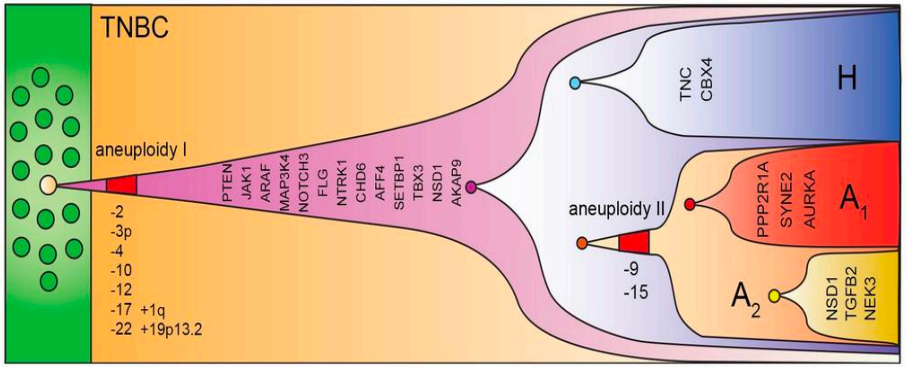
[1]David Porubsky, [2]Hyobin Jeong
[1]Dichler Lab, Department of Genome Sciences, University of Wachington
[2]Korbel Group, Genome Biology, EMBL



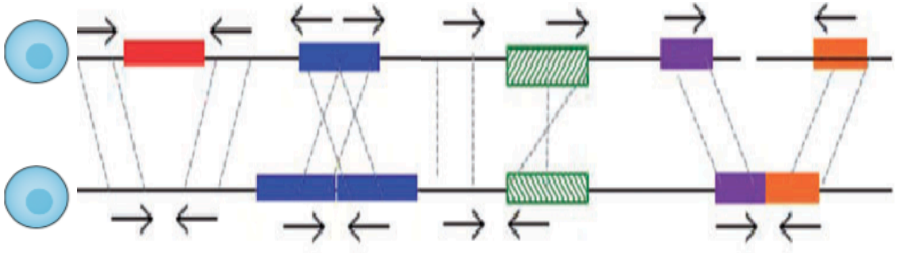
Single nucleotide variation (SNVs) - scWGS







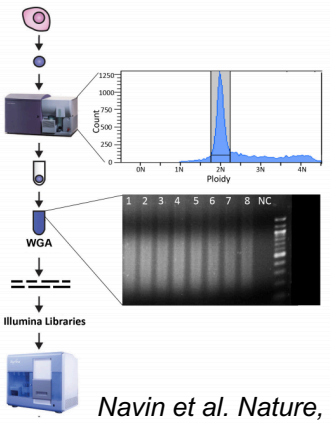
Structural Variation (SVs) – Strand-seq



Overview of the single-cell genome data analysis – scWGS

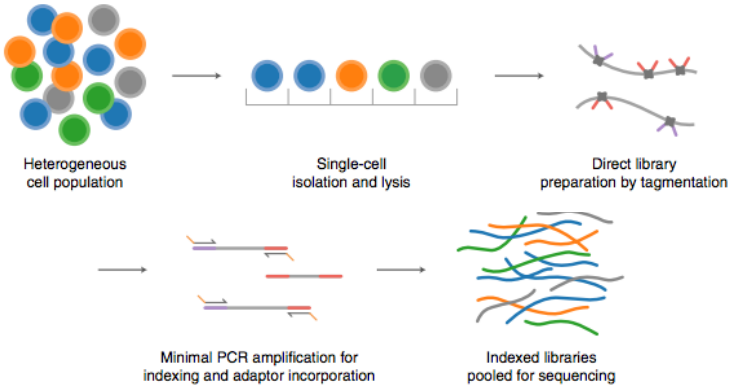
Single nucleotide variation (SNVs) - scWGS

Single-nucleus sequencing (SNS)



Navin et al. Nature, 2011

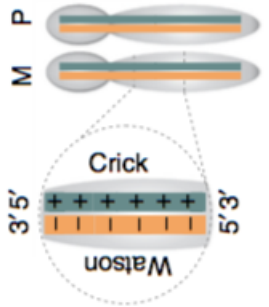
Direct library preparation (DLP)



Zahn et al. Nat Methods, 2017

Structural Variation

Strand-seq



Sanders et al. Nat protocol, 2017

Step1. Alignment - Finding a correct position of reads: Bowtie2, BWA

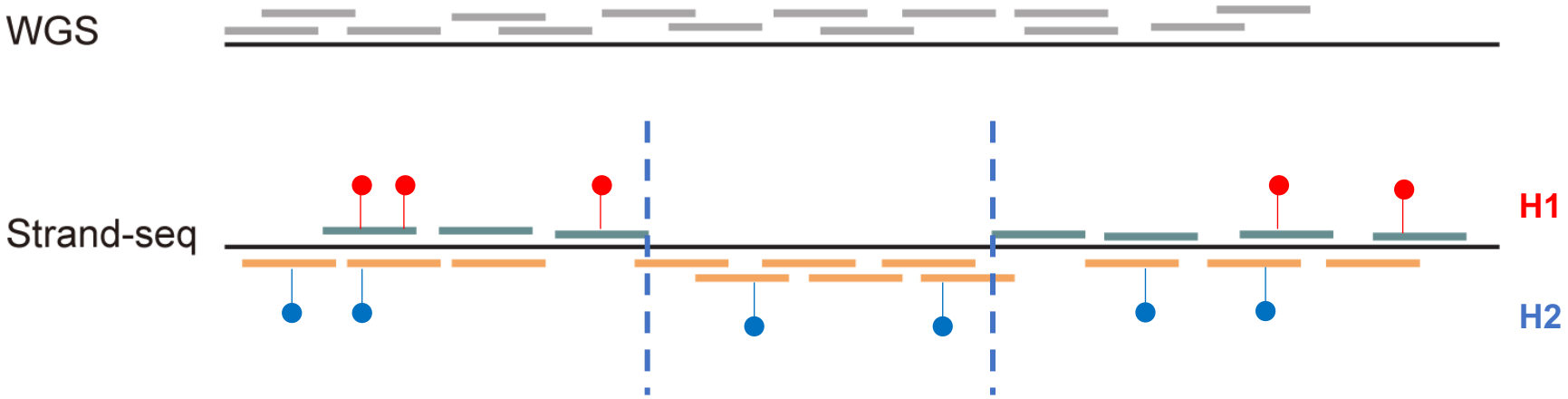
Step2. Remove PCR duplicate: Picard mark duplicate, Biobambam

Step3. Genotyping: Freebayes, GATK

Step4. Mutation calling: SCcaller, Monovar

Step5. Single-cell clustering and Phylogenetics: SCIPhi, TimeScape

Specialties and challenges of the Strand-seq data analysis

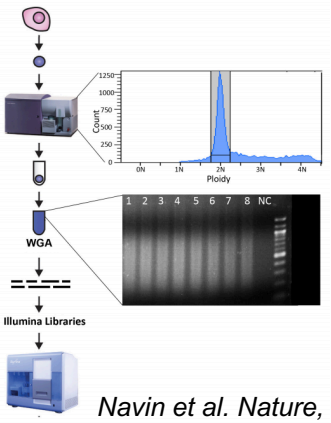


- Sequence orientation is important (Crick or Watson)
- Breakpoint needs to be detected
- Strand state and haplotypes can be assigned
- Multiple types of structural variations have their own characteristic patterns

Overview of the single-cell genome data analysis – Strand-seq

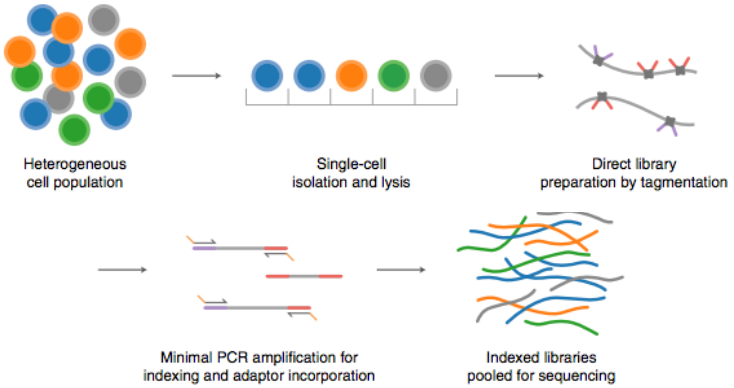
Single nucleotide variation (SNVs) - scWGS

Single-nucleus sequencing (SNS)



Navin et al. Nature, 2011

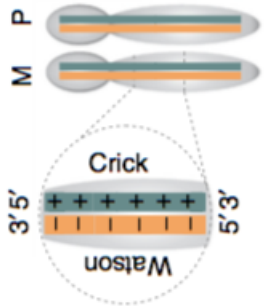
Direct library preparation (DLP)



Zahn et al. Nat Methods, 2017

Structural Variation

Strand-seq



Sanders et al. Nat protocol, 2017

Step1. Alignment - Finding a correct position of reads: BWA, sequence orientation

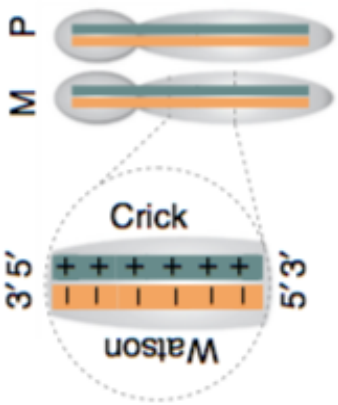
Step2. Remove PCR duplicate: Biobambam

Step3. Genotyping, Segmentation: Freebayes, StrandPhaseR, BreakpointR

Step4. Structural variation calling: MosaiCatcher

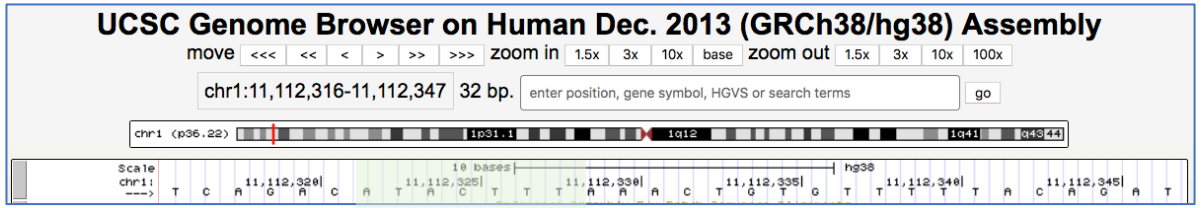
Step5. Single-cell clustering and Phylogenetics

Orientation of the reads are the key of Strand-seq data analysis



Sanders et al. Nature protocol, 2017

- Crick (C) aligns to the plus (forward) strand of the reference assembly
- Watson (W) aligns to the minus (reverse) strand



A T A C T T T

A A A G T A T

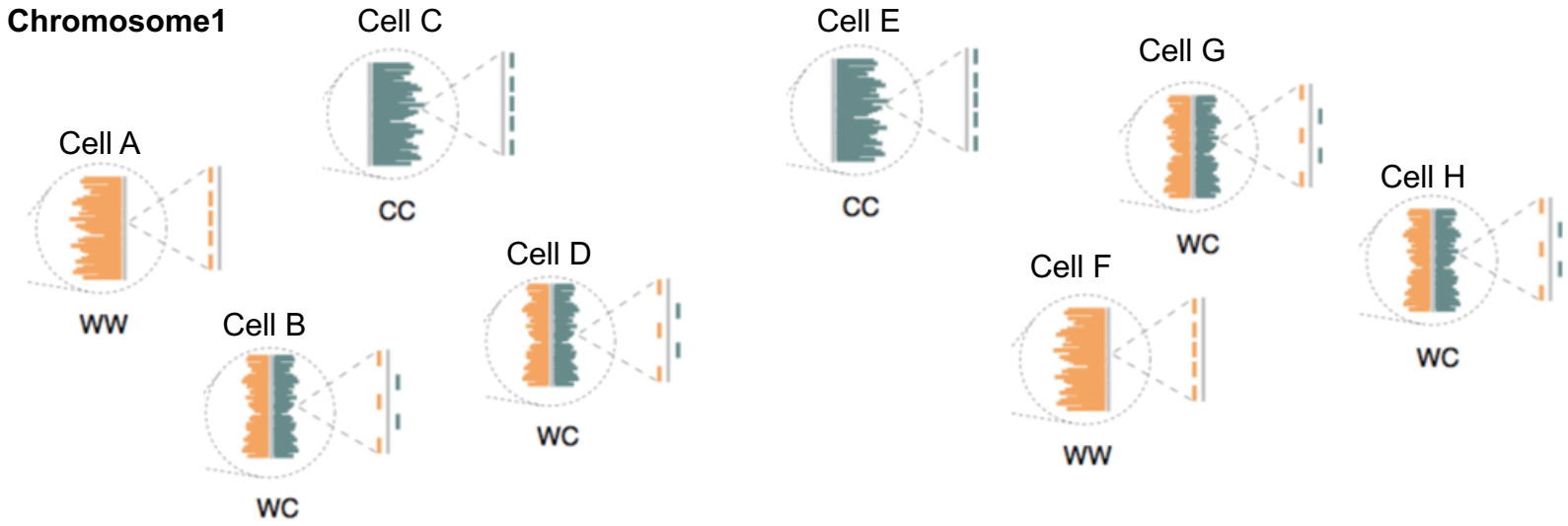
A T A C T T T

A T A C T T T

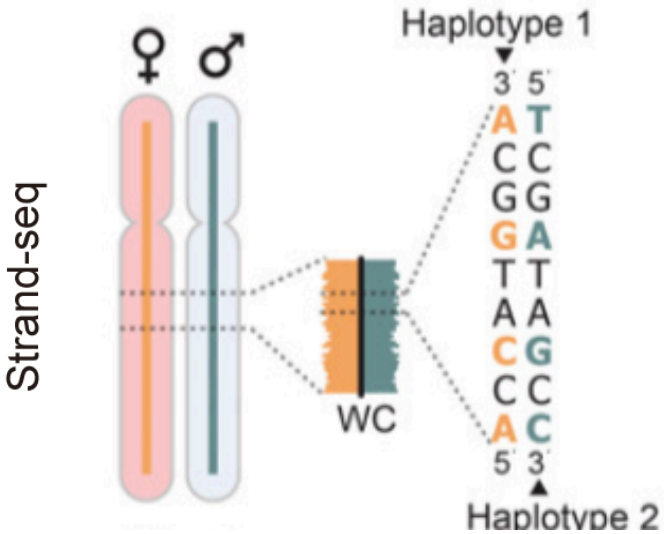
A A A G T A T

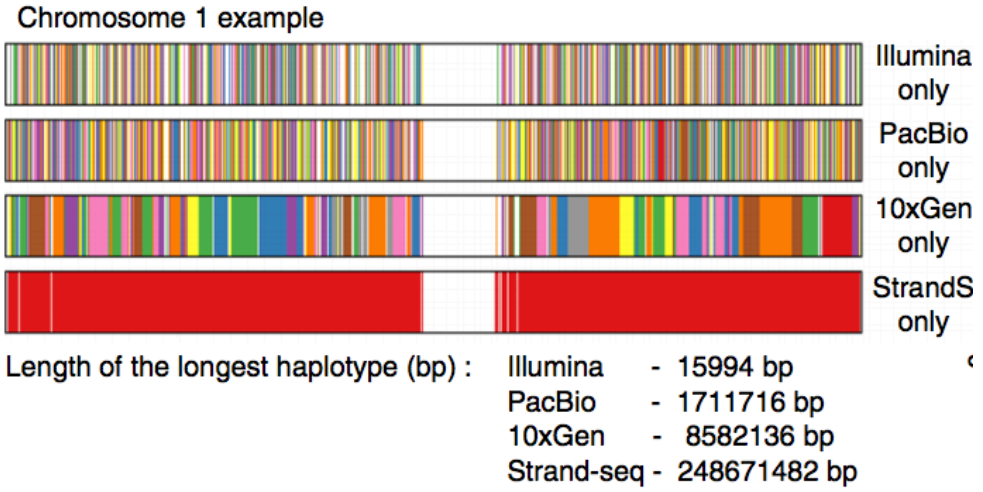
A T A C T T T Forward (+) → Crick

A A A G T A T Reverse (-) → Watson

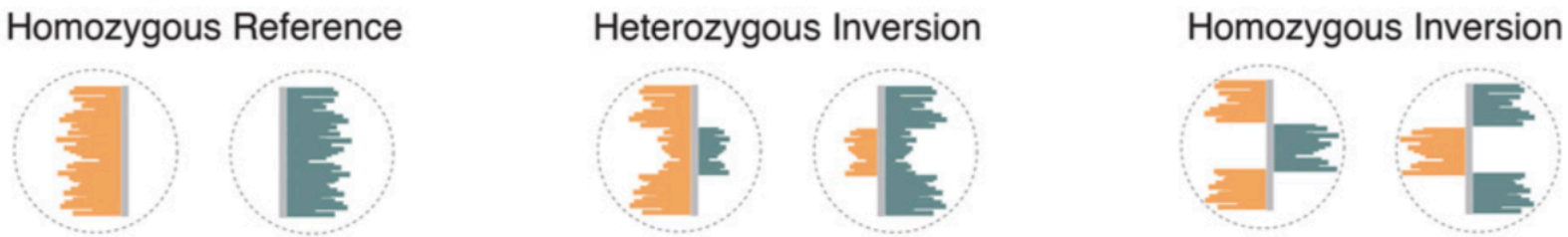


Why the orientation of the reads are important?





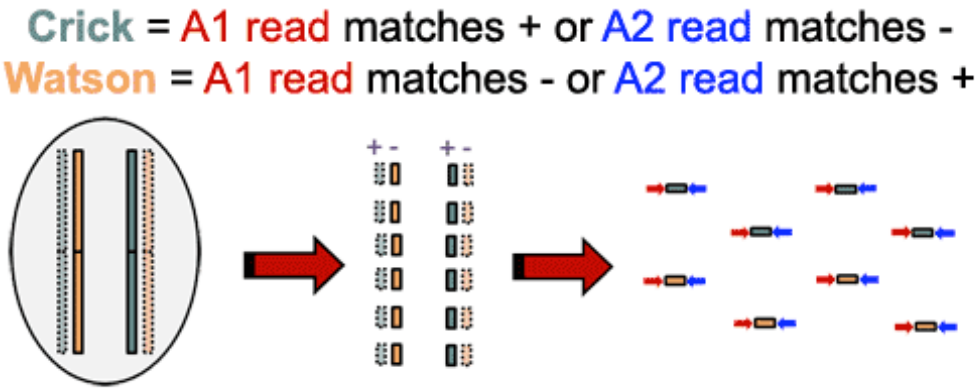
Porubsky et al. Nat comm, 2017



Sanders et al. Genome Res, 2016

How can we know the orientation of the sequencing read from the bam file?

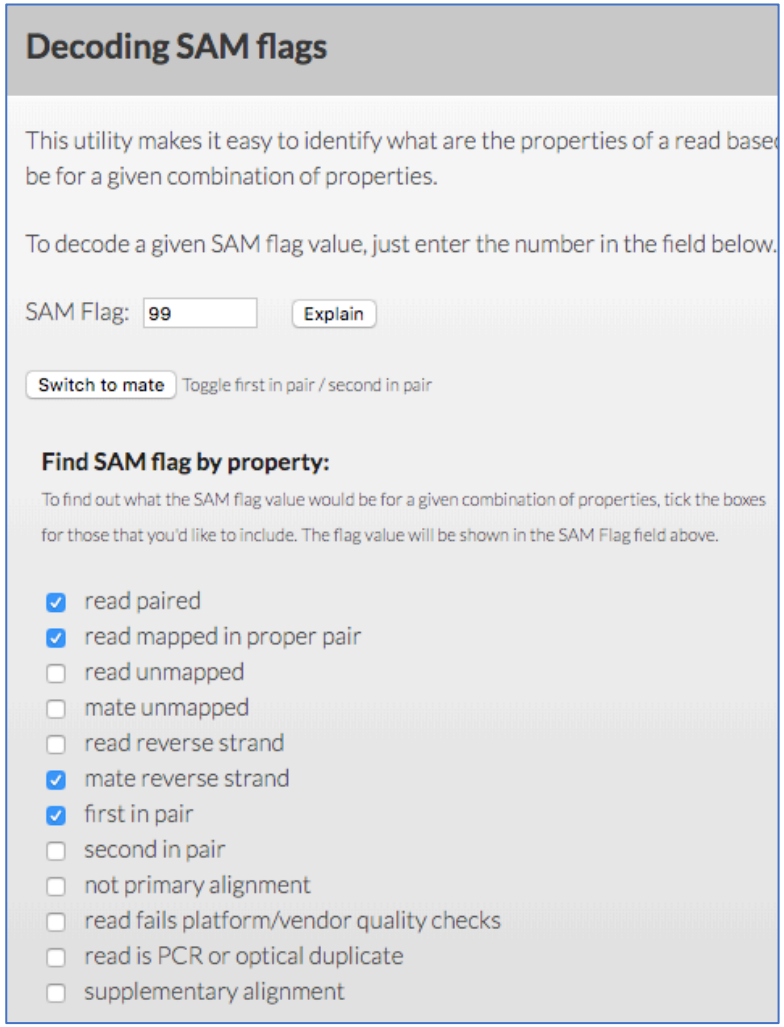
Classify reads into Crick and Watson using bamfile and samtools



https://sourceforge.net/p/bait/wiki/

##Crick reads
samtools view -f **99** output.bam > A1_C.txt
samtools view -f **147** output.bam > A2_C.txt

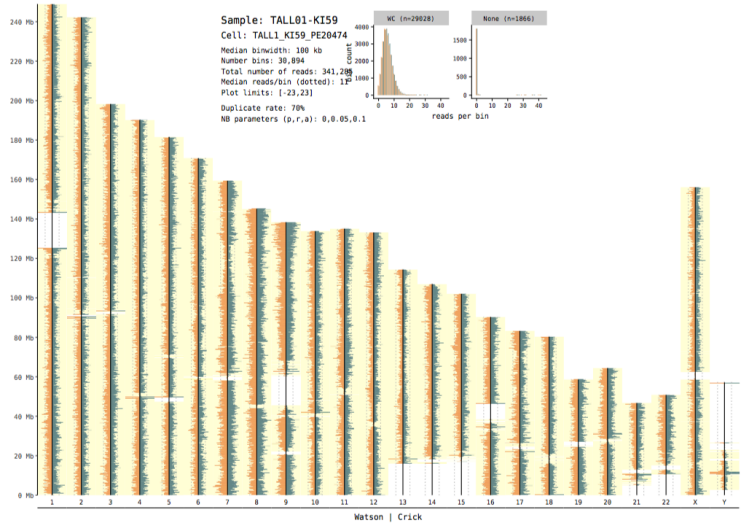
##Watson reads
samtools view -f **83** output.bam > A1_W.txt
samtools view -f **163** output.bam > A2_W.txt



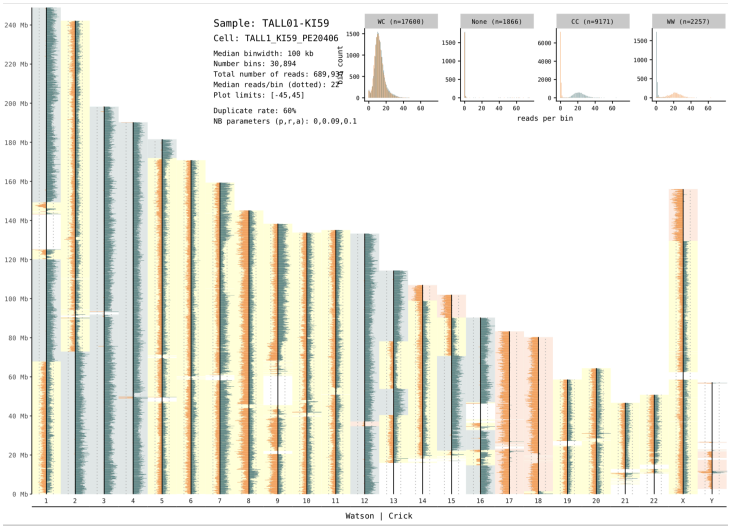
https://broadinstitute.github.io/picard/explain-flags.html

Quality check: Which one is the good quality Strand-seq library?

Whole-genome seq



Strand-seq



- In a successful Strand-seq library derived from a diploid cell, expect to see ~50% of the chromosomes exhibiting either a WW or CC state.
- If all chromosomes in the library have a WC inheritance pattern, Strand-seq was probably unsuccessful for that cell.
- By assessing the genomic coverage, level of background and proportion of WW:WC:CC chromosomes for each library, the success rate of the experiment can be assessed and successful libraries selected for further analysis

MosaiCatcher to detect abnormal pattern in the genome

MosaiCatcher to detect characteristic footprints of SVs from Strand-seq data

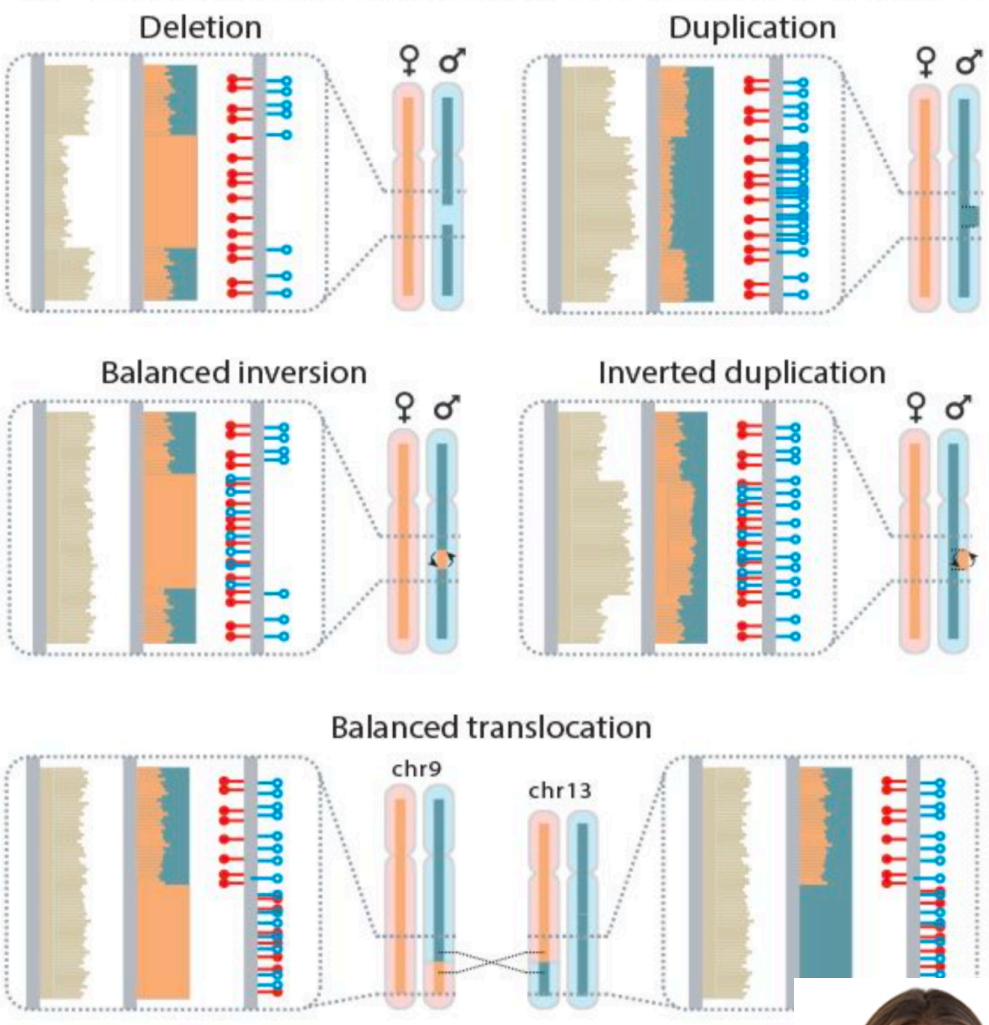


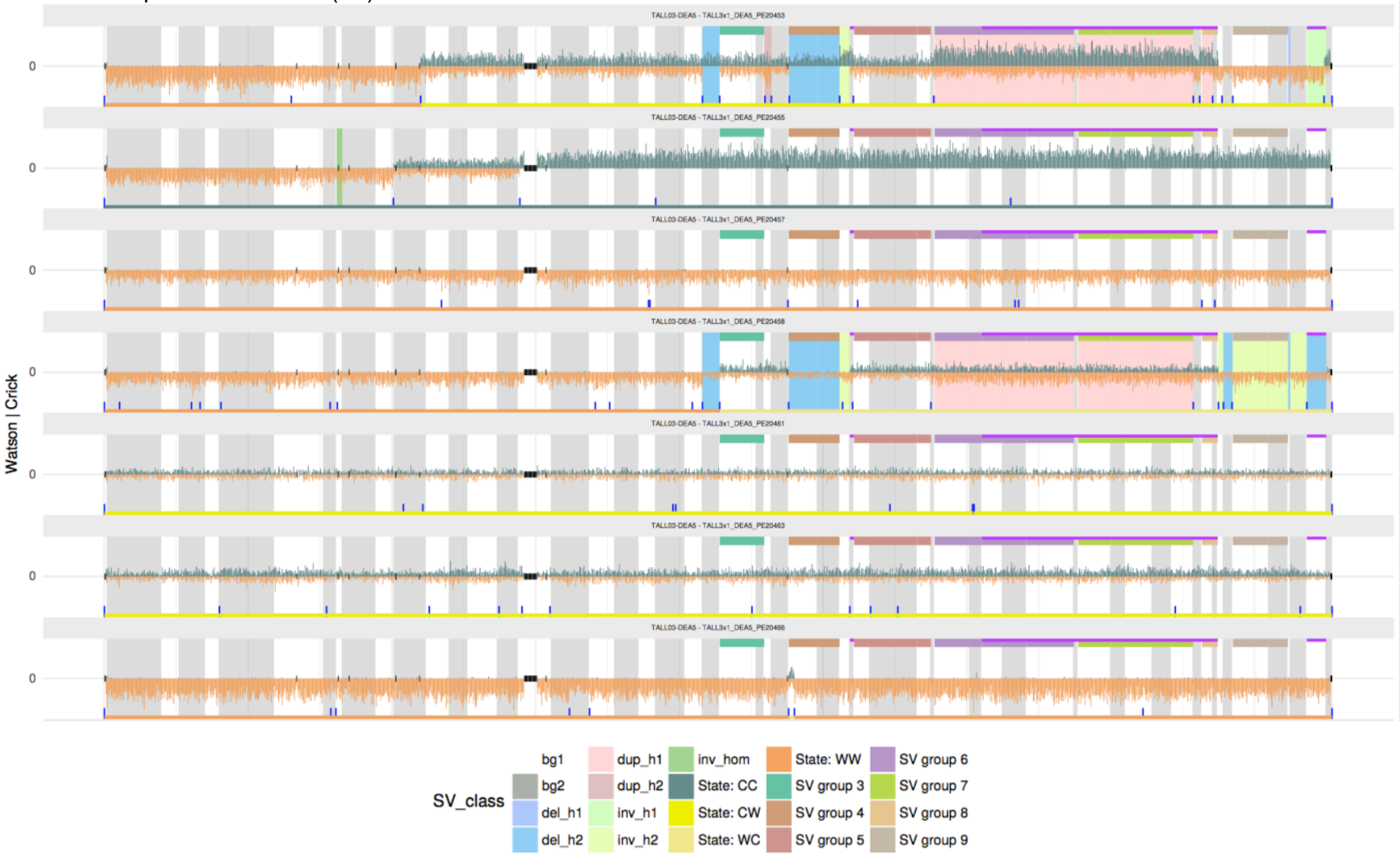
Figure from scTRIP manuscript, submitted, Sanders et al.

- Input: single-cell BAM files
- Workflow management: Snakemake

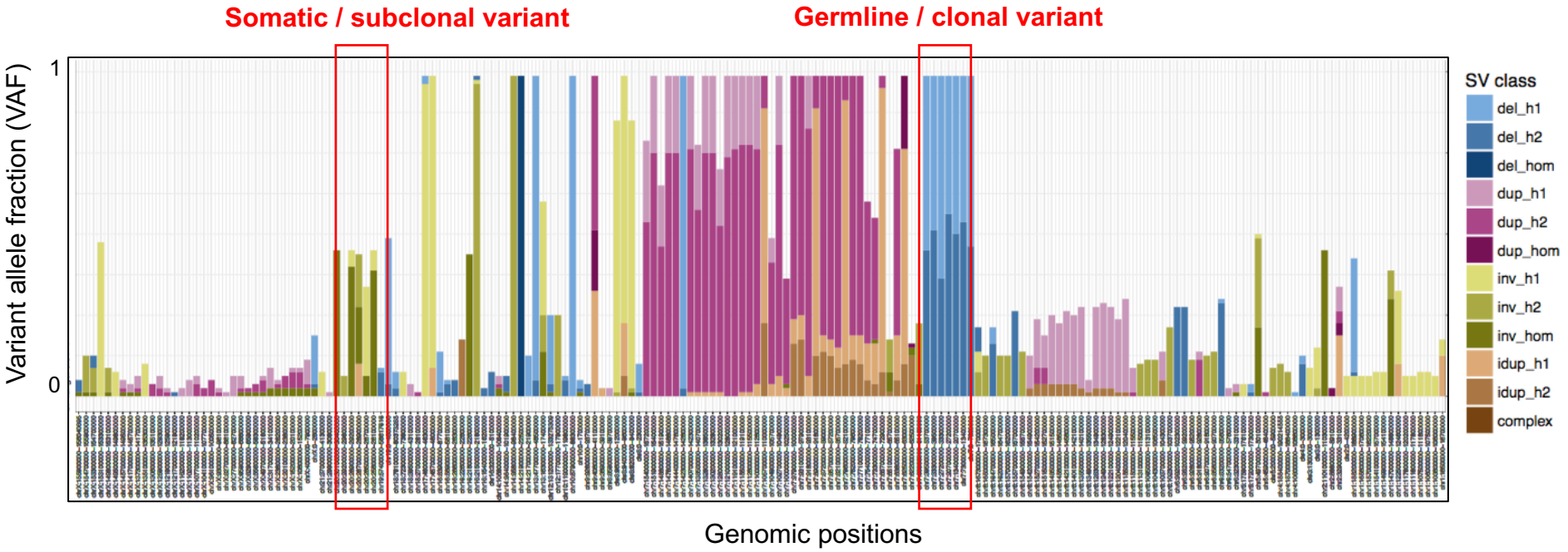
- Binned read counting (100kb) and normalization
- Assign strand-specific read data into genomic bins
- Detects and haplotype-phases heterozygous SNPs
- Segments the single cell sequence data
- Calculates genotype likelihoods for each segment and single cell using Bayesian framework

Chromosome plot with SVs called by MosaiCatcher framework

Strand-seq from T-ALL PDX (P1) Chromosome6



Variant allele fraction of SVs called by MosaiCatcher framework



- If the VAF is close to 1, the SVs are expected to be germline variant
- If the VAF is below 1, the SVs are expected to be somatic variant
- If the SVs only detected by one cell, it can be rare SV event, or an SCE (sister chromatid exchange) event
- SCEs happen independently in each single cell, and unlike SVs, SCEs are not transmitted clonally to daughter cells. Hence, changepoints resulting from SCEs are very unlikely to recur at the same position in >1 cell of a sample

Heatmap of single-cells based on SVs called by MosaiCatcher framework

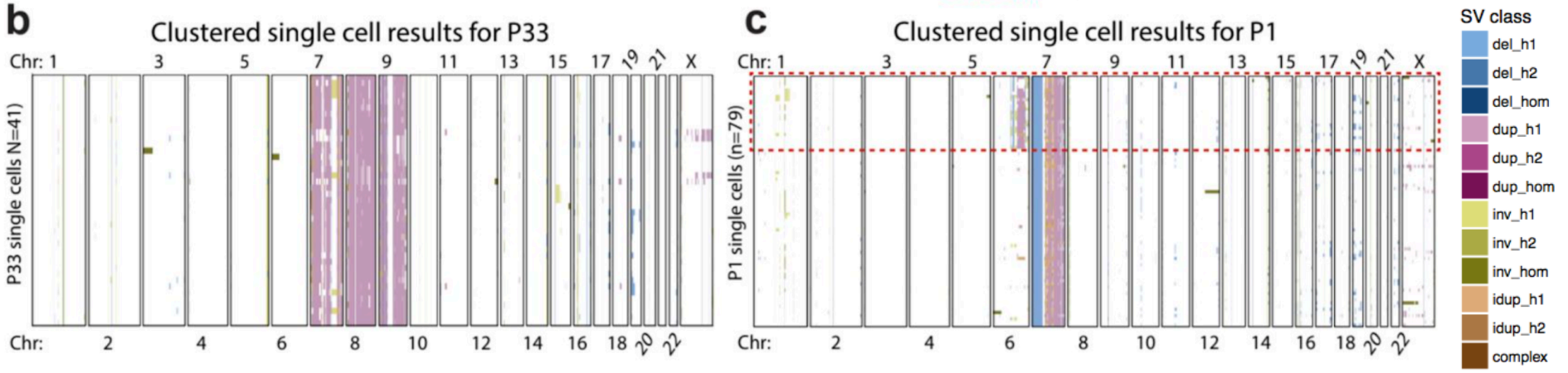
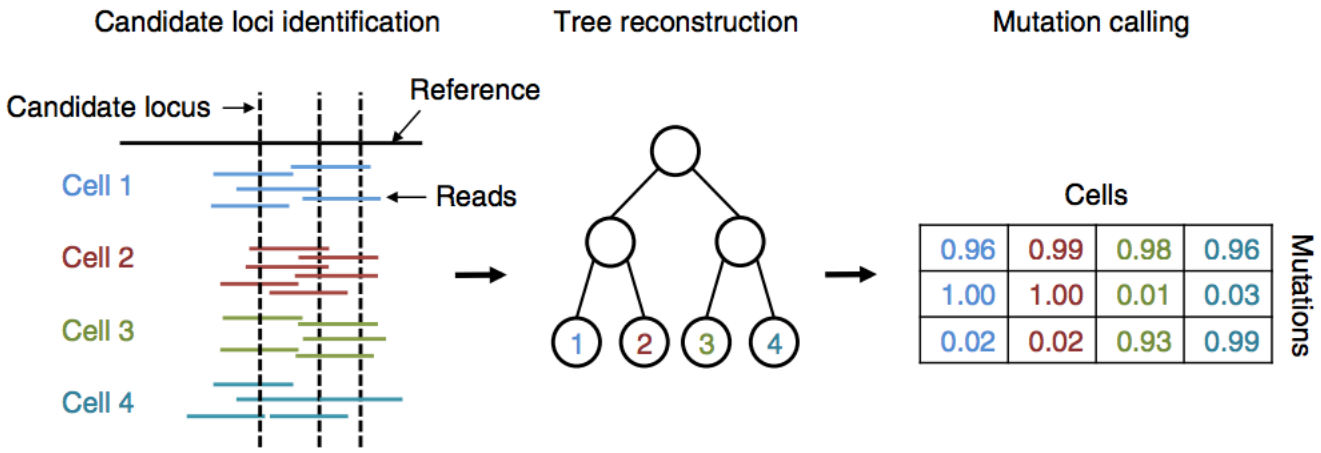
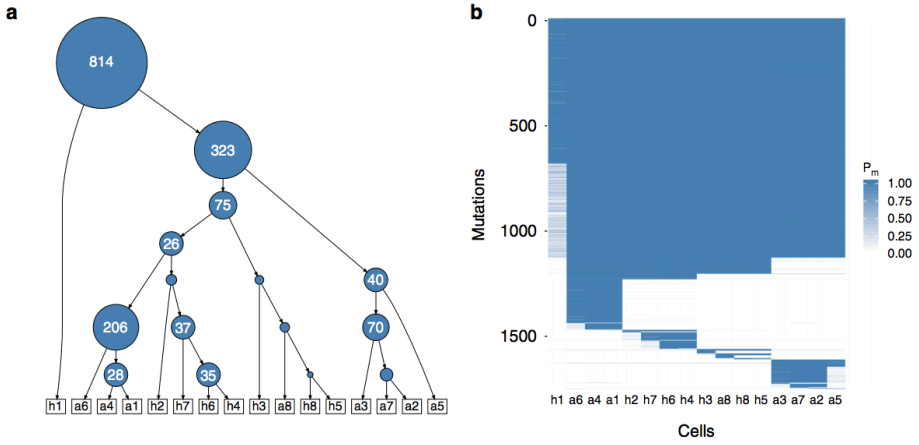


Figure from scTRIP manuscript, submitted, Sanders et al.

- This heatmap was arranged using Ward's method for hierarchical clustering of SVs genotype likelihoods in two PDX samples
- P33 shows single dominant clone but P1 shows subclonal population in the sample represented by 23 cells

More tools for the clustering and Phylogenetics of single cells - *SCIΦ*

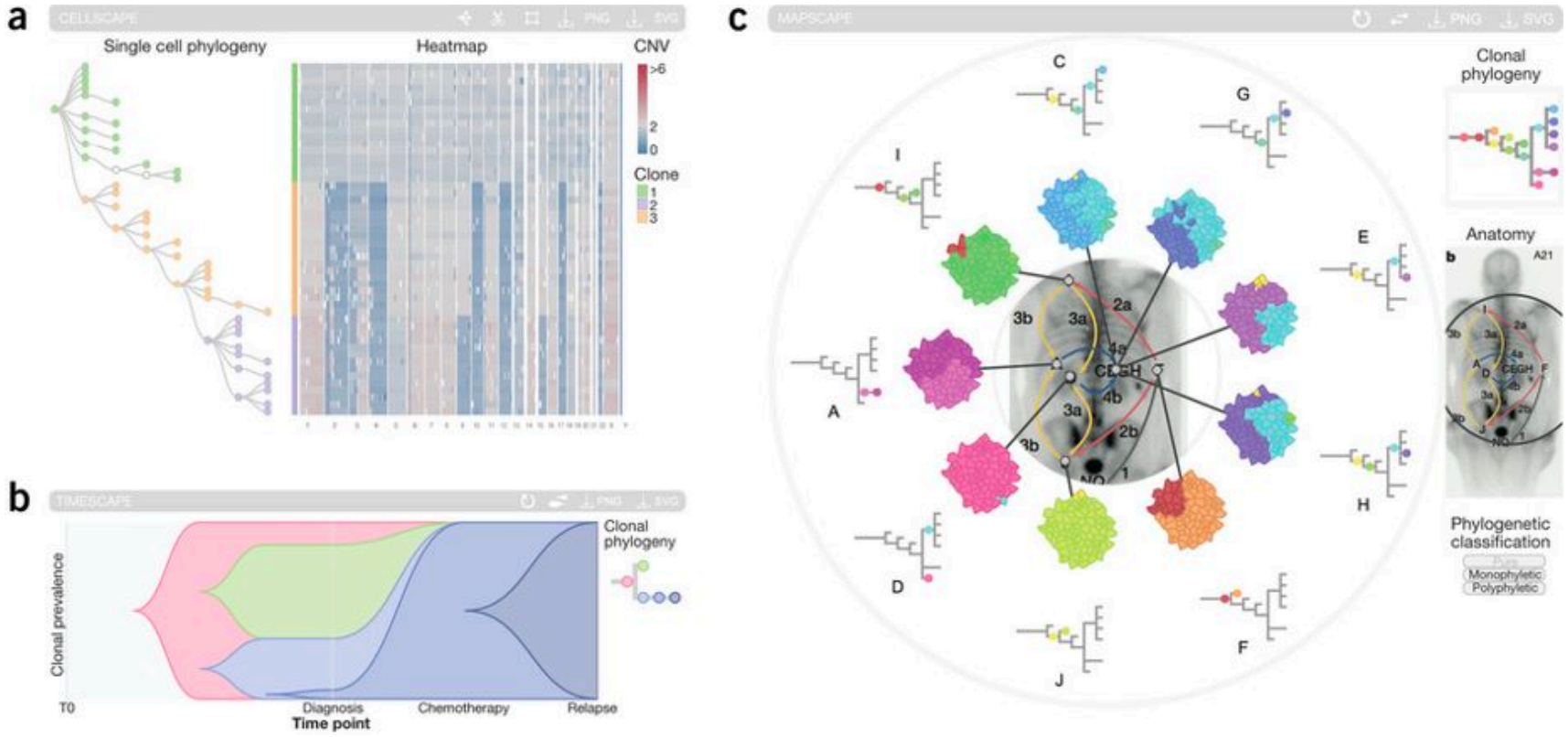




- This method jointly calls mutations in individual cells and estimates the tumor phylogeny among these cells.
- Employing a Markov Chain Monte Carlo scheme enables us to reliably call mutations in each single cell even in experiments with high drop-out rates and missing data.

Singer et al. Nat Comm, 2018
Niko Beerenwinkel group

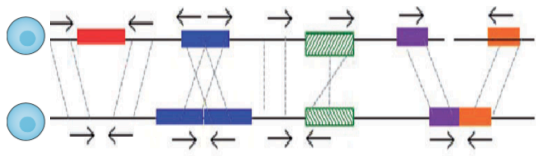
More tools for the clustering and Phylogenetics of single cells – *E-scape*



Smith et al. Nat Methods, 2017 Sohrab P Shah group

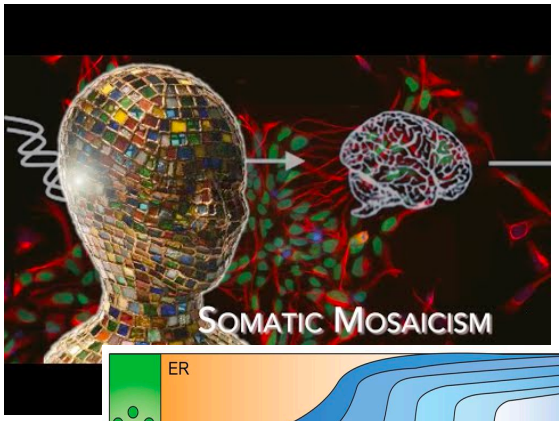
- E-scape consists of three visualization tools deployed as R html widgets
- 1) TimeScape for time series analysis
- 2) MapScape for spatial distribution analysis
- 3) CellScape for single-cell analysis, including analysis of the phylogenetic and population dynamics of cancer – the heatmap depicts genome-wide CNVs across single cells. The tree illustrates the evolutionary relationships between cells.

Summary

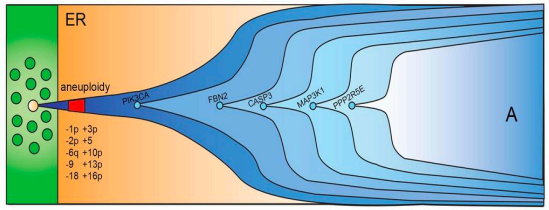


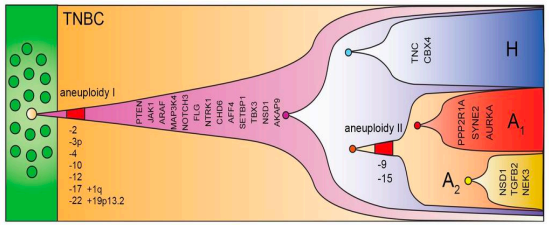
BREAKPOINTR

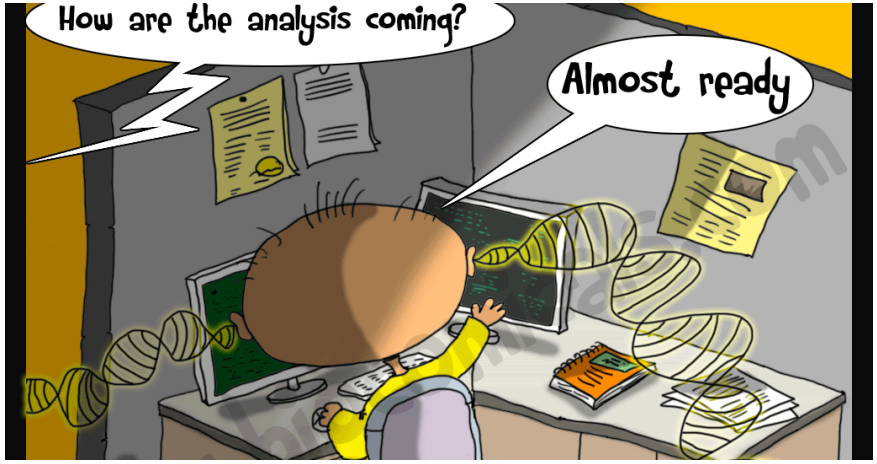
MOSAIC CATCHER



StrandPhaseR

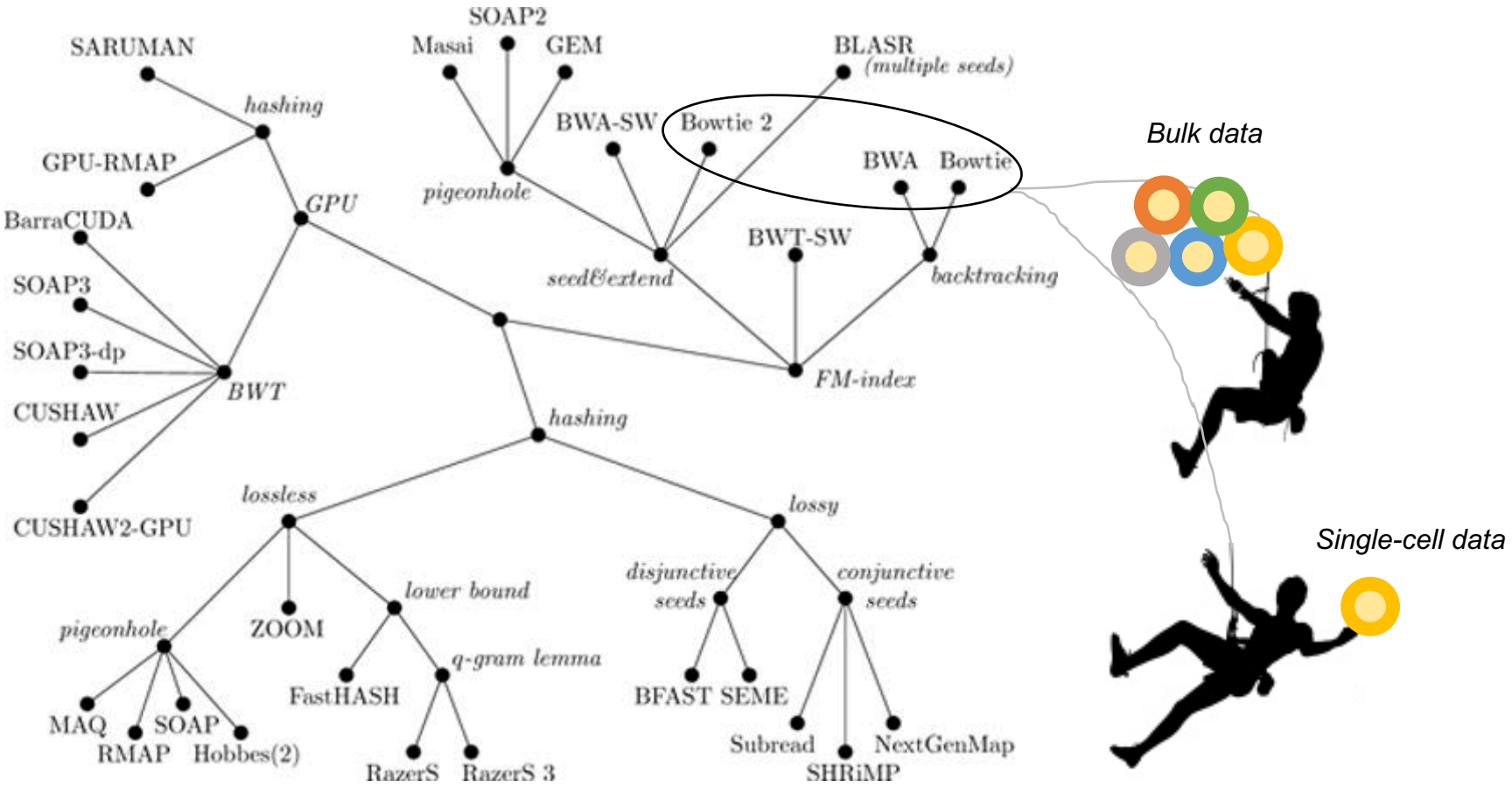






Thank you for listening

Short DNA Read Mapping: An Algorithmic Tour



Proc IEEE Inst Electr Electron Eng. 2017

Additionally, for Strand-seq, orientation of the reads need to be considered!

General challenges and approaches of the single-cell genome analysis

- Challenge1: High rates of allelic drop-out as a result of the random non-amplification of one allele at the heterozygous genotype sites
- Challenge2: False positive artifects can arise in the DNA amplification when random errors introduced early in the process
- Challenge3: Uneven amplification across the genome which results in non-uniform coverage and insufficient coverage depth for reliable base calling

- SNV caller doesn't transfer information between cells : ***SCcaller*** – it detects variants independently for each cell and accounts for local allelic amplification biases. However, it cannot recover mutations from drop-out events or loss of heterozygocity.

- SNP caller transfers information between cells : ***Monovar*** – It addresses the problem of low and uneven coverage in mutation calling by pooling sequencing information across cells while assuming that no dependencies exist across sites

How's the computational workflow different to detect structural variations (SV) ?

Translation of diagnostic footprint into expected number of copies in W and C

SV diagnostic footprints in a WC ground state

SV state	Depth	W:C	Haplotype tags W	C	W cov	C cov
Reference state	2N	50%	H1	H2	1N	1N
Deletion of H1	1N	0%	-	H2	0N	1N
Deletion (homozygous)	0N	-	-	-	0N	0N
Duplication of H1	3N	66%	2xH1	H2	2N	1N
Duplication (homozygous)	4N	50%	2xH1	2xH2	2N	2N
Inversion of H1	2N	0%	-	H1+H2	0N	2N
Inversion (homozygous)[1]	2N	50%	H2	H1	1N	1N
Inverted duplication of H1[2]	3N	33%	H1	H1+H2	1N	c

SV diagnostic footprints in a WW ground state

SV state	Depth	W:C	Haplotype tags W	C	W cov	C cov
Reference state	2N	100%	H1+H2	-	2N	0N
Deletion of H1[3]	1N	100%	H2	-	1N	0N
Deletion (homozygous)	0N	-	-	-	0N	0N
Duplication of H1[3]	3N	100%	2xH1+H2	-	3N	0N
Duplication (homozygous)	4N	100%	2xH1+2xH2	-	4N	0N
Inversion of H1[3]	2N	50%	H2	H1	1N	1N
Inversion (homozygous)	2N	0%	-	H1+H2	0N	2N
Inverted duplication of H1[3]	3N	67%	H1+H2	H1	2N	1N

SV diagnostic footprints in a CW ground state

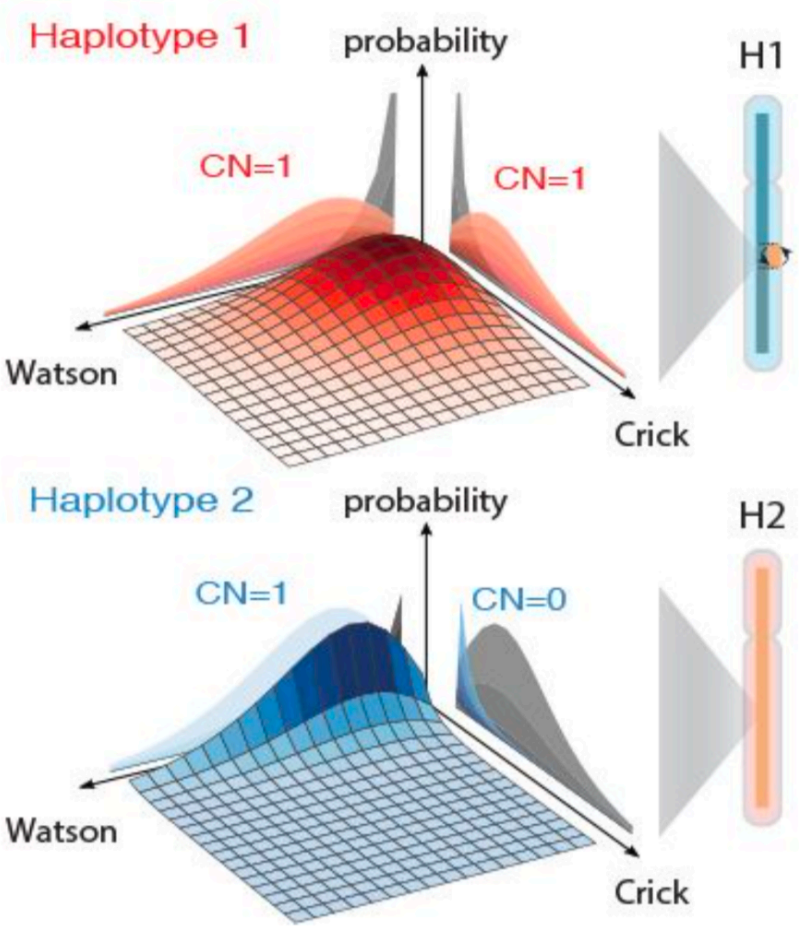
SV state	Depth	W:C	Haplotype tags W	C	W cov	C cov
Reference state	2N	50%	H2	H1	1N	1N
Deletion of H1	1N	0%	H2	-	1N	0N
Deletion (homozygous)	0N	-	-	-	0N	0N
Duplication of H1	3N	66%	H2	2xH1	1N	2N
Duplication (homozygous)	4N	50%	2xH2	2xH1	2N	2N
Inversion of H1	2N	0%	H1+H2	-	2N	0N
Inversion (homozygous)[1]	2N	50%	H1	H2	1N	1N
Inverted duplication of H1[2]	3N	33%	H1+H2	H1	c	1N

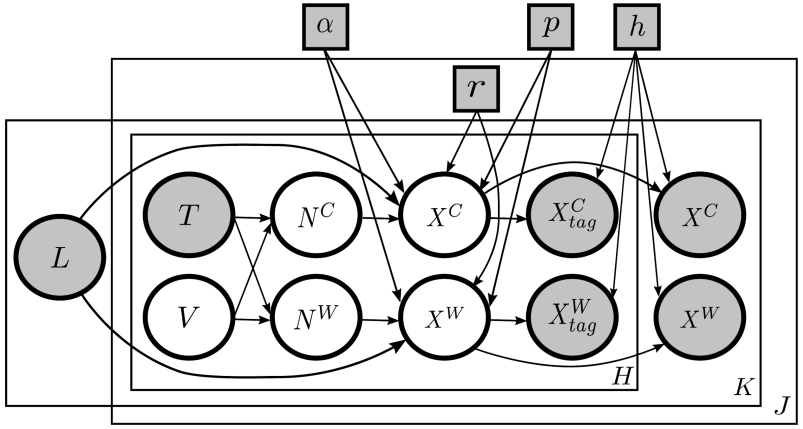
SV diagnostic footprints in a CC ground state

SV state	Depth	W:C	Haplotype tags W	C	W cov	C cov
Reference state	2N	0%	-	H1+H2	0N	2N
Deletion of H1[3]	1N	0%	-	H2	0N	1N
Deletion (homozygous)	0N	-	-	-	0N	0N
Duplication of H1[3]	3N	0%	-	2xH1+H2	0N	3N
Duplication (homozygous)	4N	0%	-	2xH1+2xH2	0N	4N
Inversion of H1[3]	2N	50%	H2	H1	1N	1N
Inversion (homozygous)	2N	100%	H1+H2	-	2N	0N
Inverted duplication of H1[3]	3N	33%	H1	H1+H2	1N	2N

- Negative binomial (NB) distribution is the basis for this Bayesian framework (p, r)
- Parameters (p, r) are estimated from the observed read counts
- Each SV diagnostic footprint can be translated into the expected number of copies sequenced in W and C orientation
- These expectations are formalized in Bayesian model

MosaiCatcher uses Bayesian model for SV genotype likelihoods in single cells





- Every haplotype-resolved SV class in a segment together with the ground state define a Watson and Crick copy number used to compute the NB likelihood of observed read counts.
- SV call is accepted if the log odds ratios (of an SV genotype vs. the reference state) was at least 4

- The probability distributions represent an InvDup on H1
- Segments on both strands are seen for haplotype 1 (H1), but H2 is represented on the W strand only

Figure from scTRIP manuscript, submitted, Sanders et al.