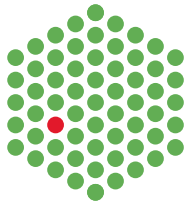
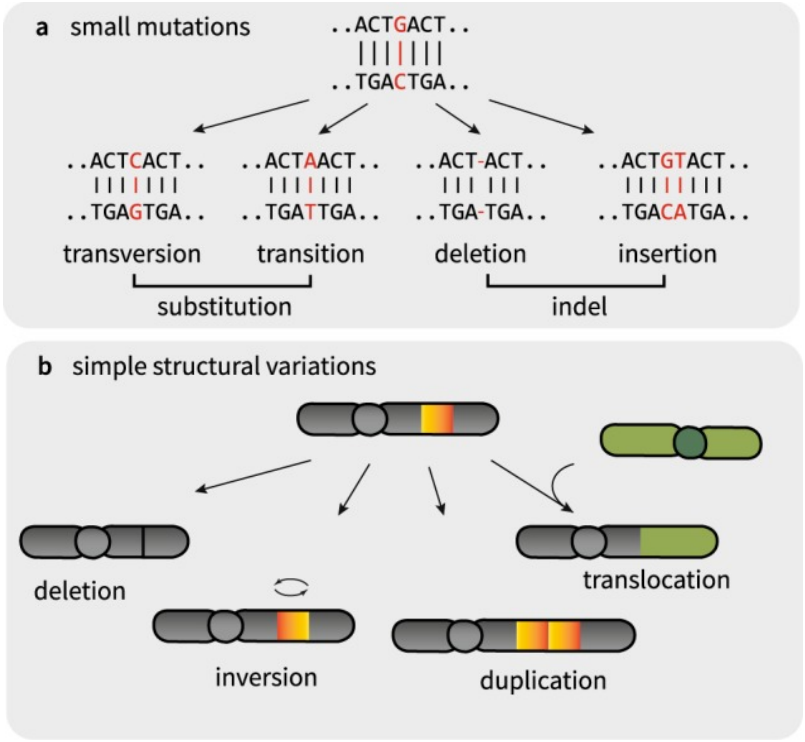
Detection of structural variations using bulk whole-genome sequencing and Delly



Structural variants are diverse, and occur commonly in the human germline



(Yi, K., & Ju, Y., S., Exp Mol Med 2018)

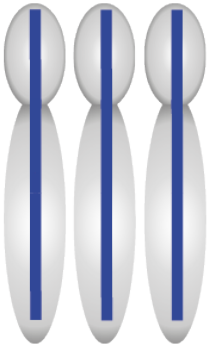
Heritable sequence differences between individuals

0.1%

0.5-1%

Germline structural variation has been linked to various syndromes

Chr21
3N

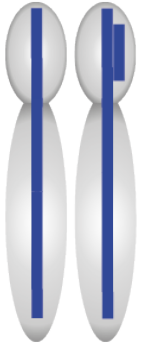


Down Syndrome

Chr17
Deletion

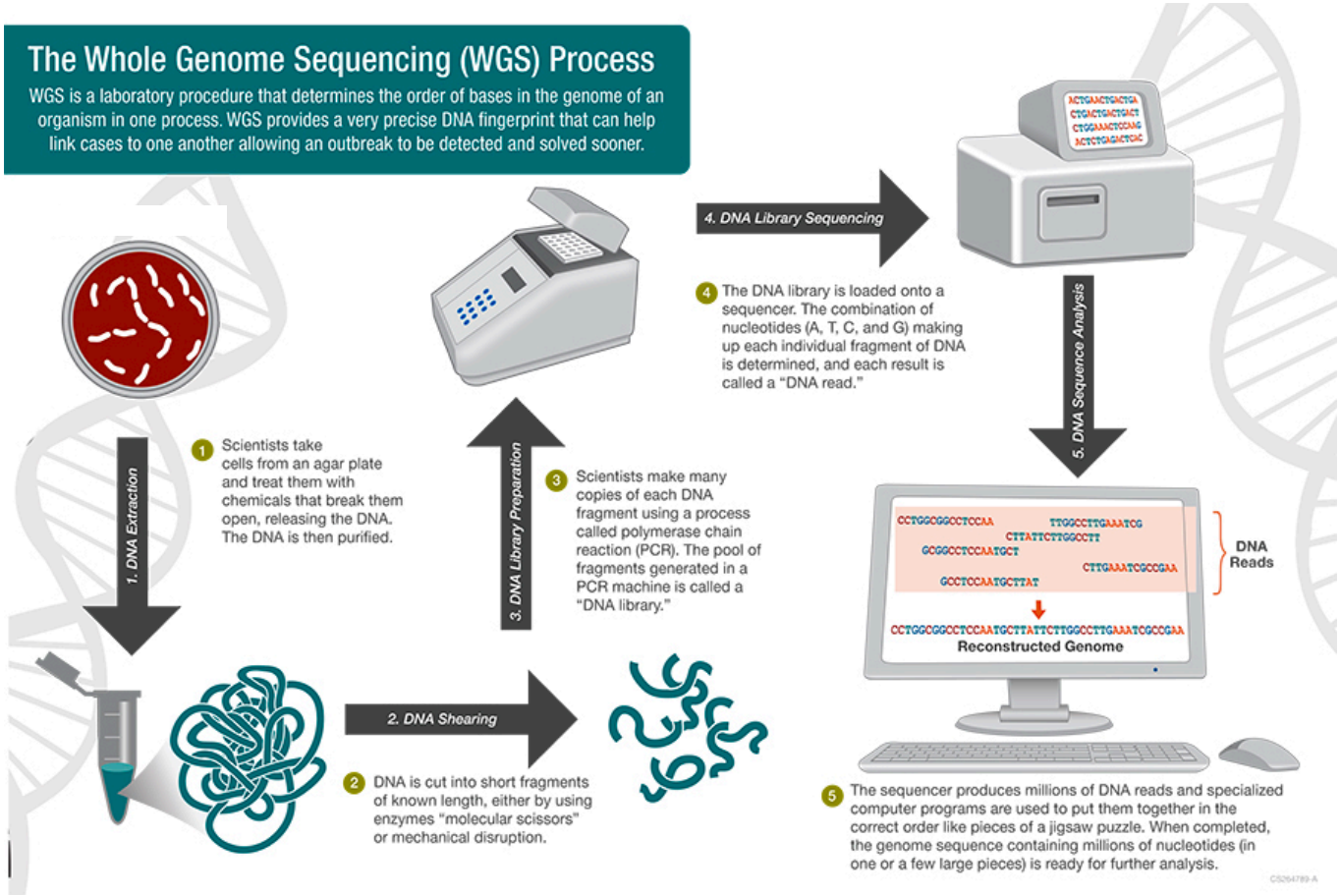
Smith-Magenis Syndrome

Chr2
Duplication



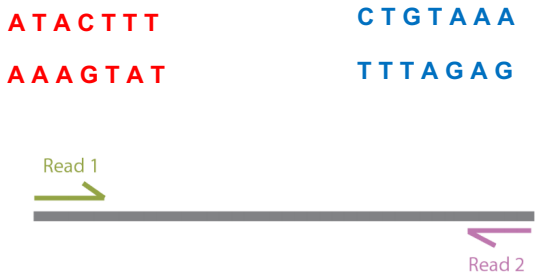
Lynch Syndrome

How do we get the genome sequence information?

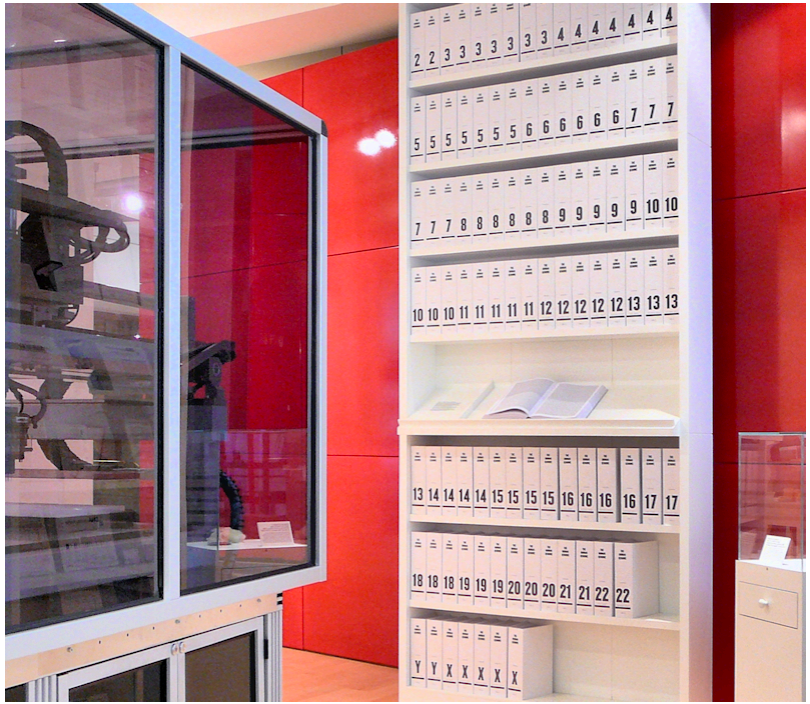


Fastq files are the starting point of genomics data analysis

Paired-end
*_1_sequence.txt.gz *_2_sequence.txt.gz



Searching for the origin of DNA fragment using reference genome



Wellcome Collection, London

The first printout of the human reference genome

A T A C T T T

A A A G T A T

BWA

bam file

A T A C T T T. Chr3:1000

A A A G T A T. Chr11:35082

Quality checking of bam file after the alignment

- Checkpoint1: How much portion of the reads were aligned to the reference genome?
- Checkpoint2: How much portion of the reads were from PCR duplicates?
- If the bam file pass the QC, we are ready to detect structural variations!

Alfred

install with bioconda | downloads 12k total | License GPLv3 | release v0.1.17 | DOI 10.1093/bioinformatics/bty1007

Alfred: BAM alignment statistics, feature counting and feature annotation

Alfred is available as a Bioconda package, as a statically linked binary from the GitHub release page or as a minimal Docker container. Please have a look at Alfred's documentation for any installation or usage questions.

Tobias Rausch
Senior Bioinformatician

Structural Variant Calling (germline & somatic)

BIOINFORMATICS

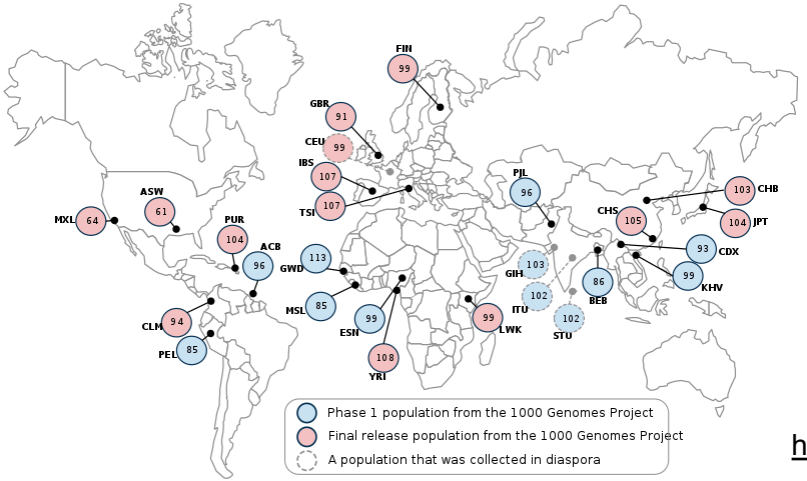
Vol. 28 ECCB 2012, pages i333–i339
doi:10.1093/bioinformatics/bts378

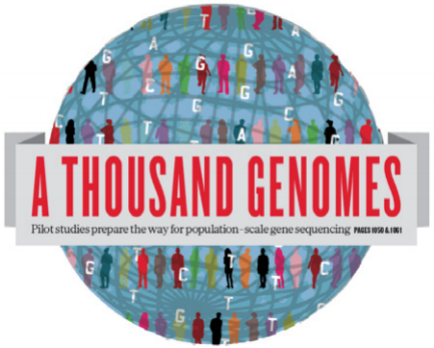
DELLY: structural variant discovery by integrated paired-end and split-read analysis

Tobias Rausch,[1,3,*], Thomas Zichner[1], Andreas Schlattl[1], Adrian M. Stütz[1], Vladimir Benes[3] and Jan O. Korbel[1,2]

[1]European Molecular Biology Laboratory (EMBL), Genome Biology, Meyerhofstr. 1, 69117 Heidelberg, Germany and [2]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]European Molecular Biology Laboratory (EMBL), Core Facilities and Services, Meyerhofstr. 1, 69117 Heidelberg, Germany

Rausch et al. 2012

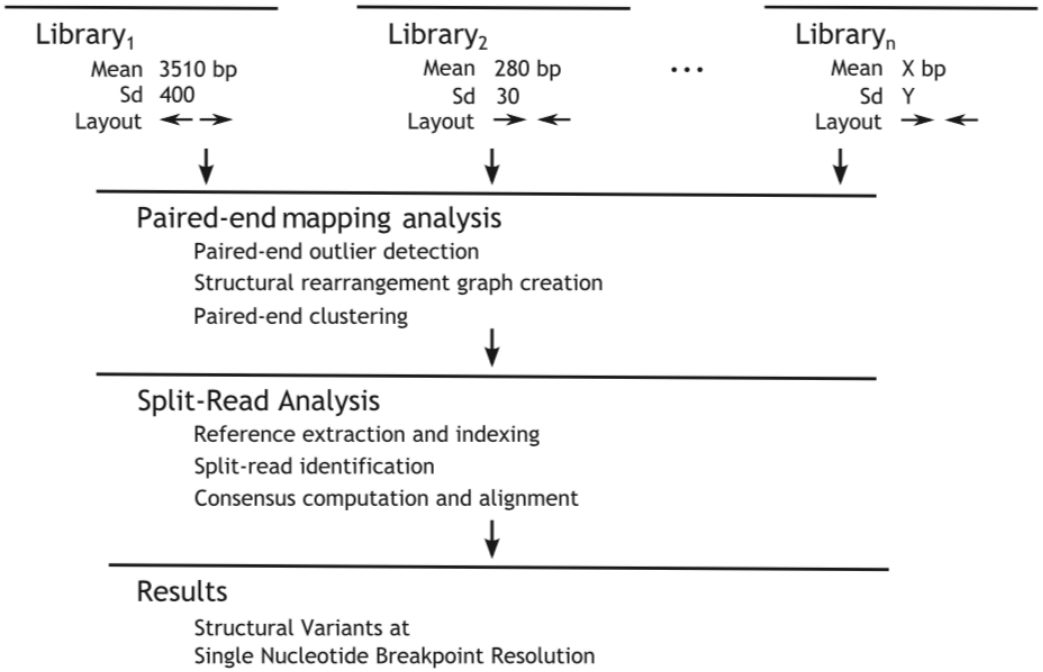




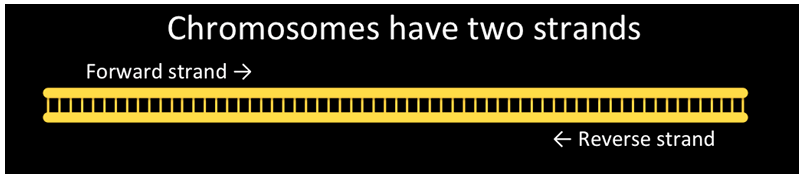
https://www.nature.com/articles/nature15393

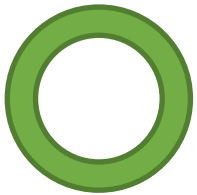
Two components of DELLY analysis

- Paired-end mapping analysis
- Split-read analysis



Paired reads have different orientation by default





Paired reads gives estimation of probable insert size of that library

Library 1
Fragment size : 500bp
Insert size : 400bp

Read1
(50bp)

Read2
(50bp)

Library 2
Fragment size : 350bp
Insert size : 250bp

Read1
(50bp)

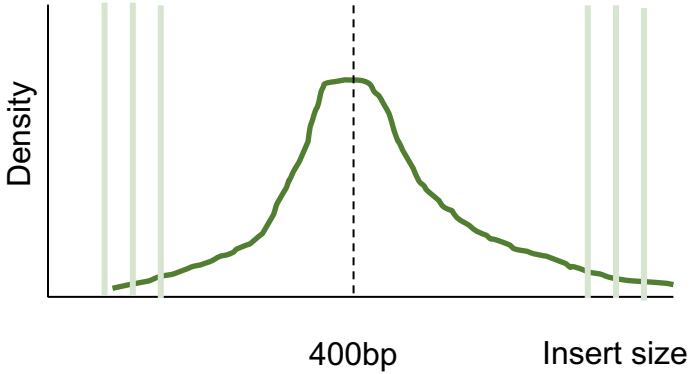
Read2
(50bp)

- Standard Illumina sequencing libraries tend to have a fragment size of 100-700bp
- Insert size is impacted by the process of cluster generation in which libraries are denatured, diluted and distributed on the two-dimensional surface of the flow cell and then amplified.

Discordantly mapped read-pairs can be identified by insert size & orientation

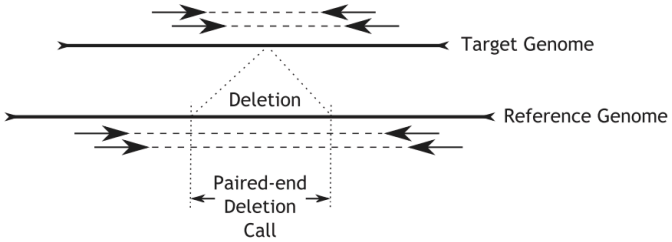
Abnormal orientation

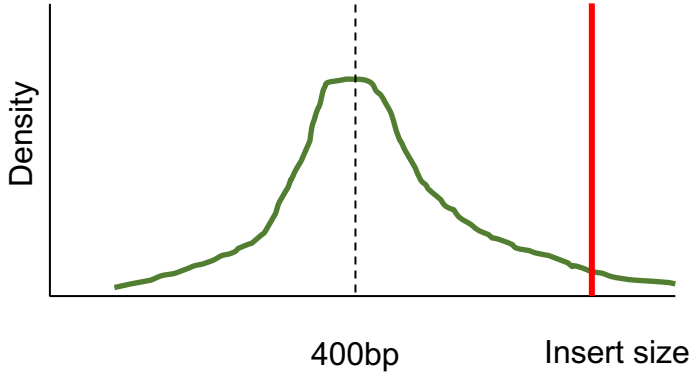
Abnormal insert size



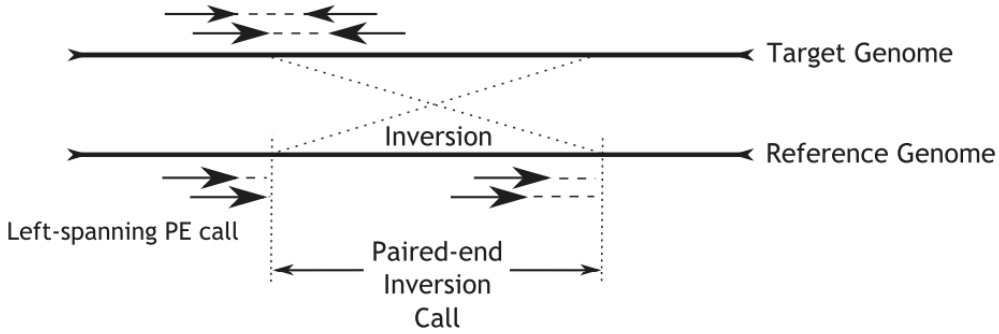
- Discordantly mapped read-pairs that either have an abnormal orientation or an insert size greater than the expected range.
- Default insert size cutoff is three standard deviations from the median insert size.

Deletion: default orientation & extremely big insert size

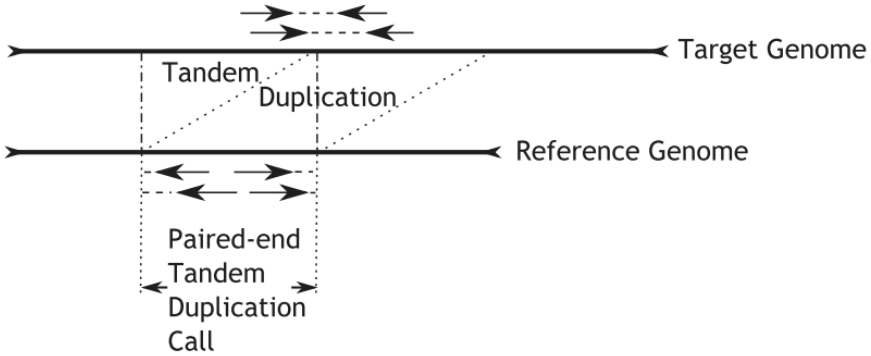




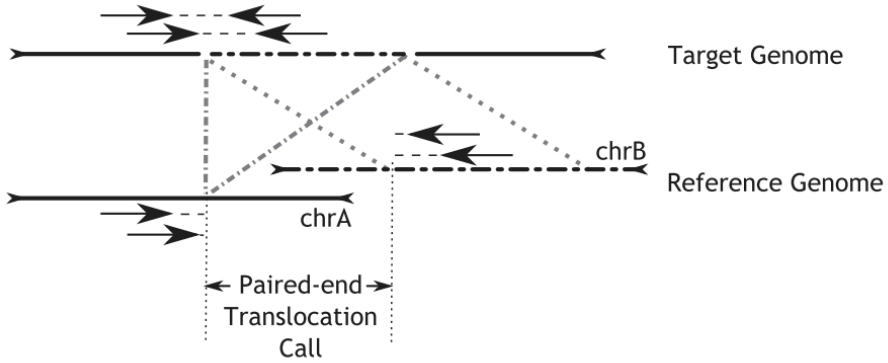
Inversion: abnormal orientation



Tandem duplications: first and second read changed their relative order

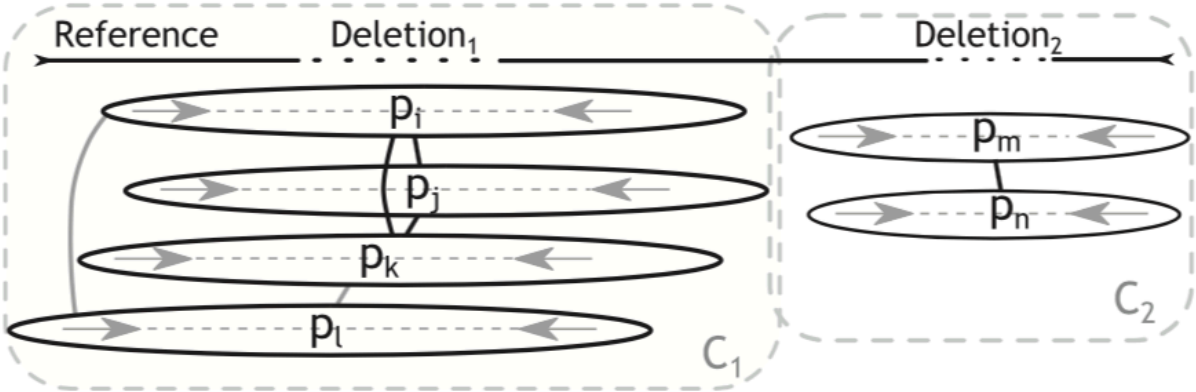


Translocation: paired-ends mapping to different chromosomes



Collect discordant pairs and make clusters

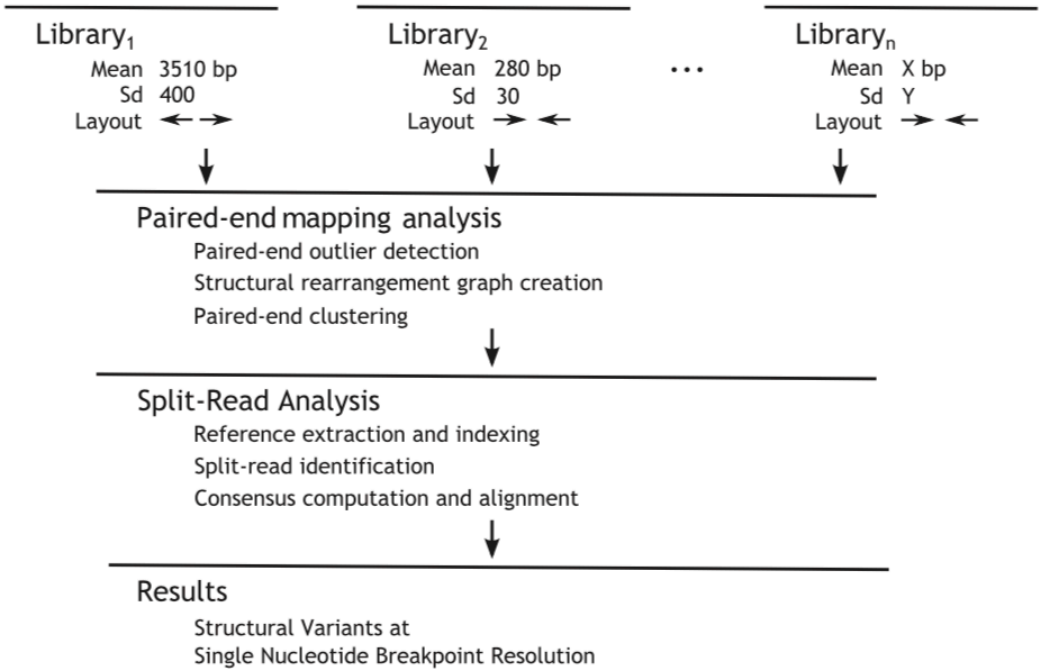
It gives breakpoint-containing genomic intervals!



Now we need Split-read analysis to determine breakpoint!

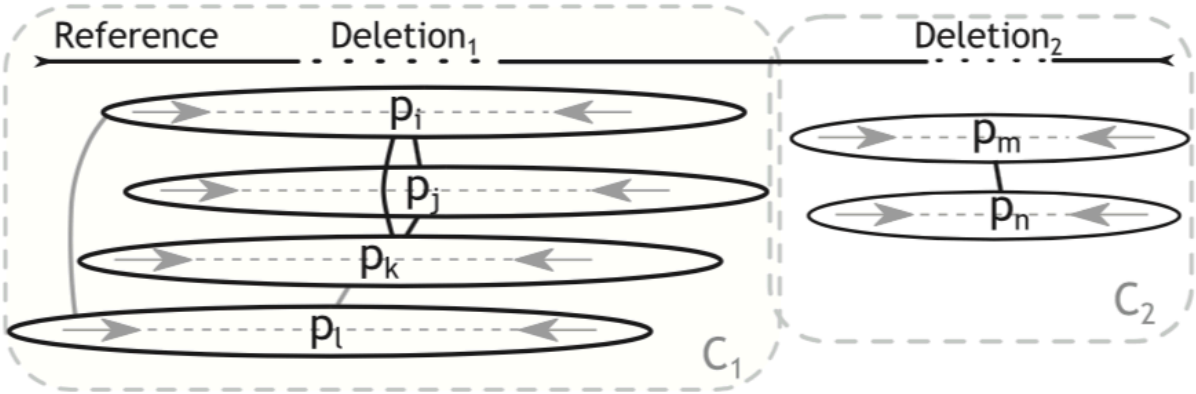
Two components of DELLY analysis

- Paired-end mapping analysis
- Split-read analysis



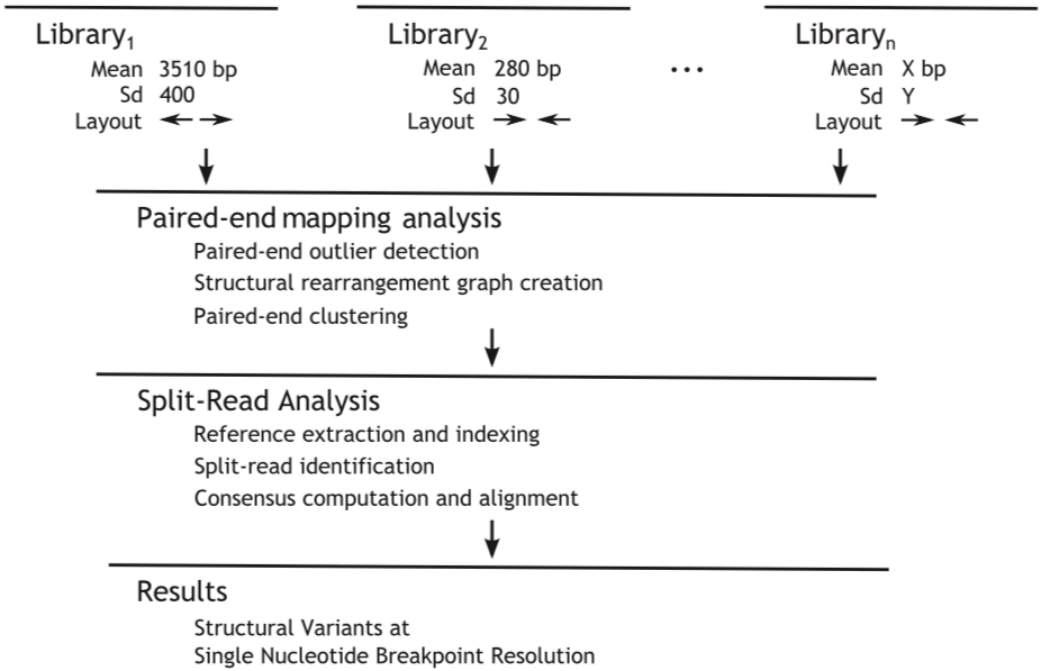
Searching for the single-anchored paired ends

A single-anchored paired-end is a read pair where one read maps to the reference and the other read is unmapped



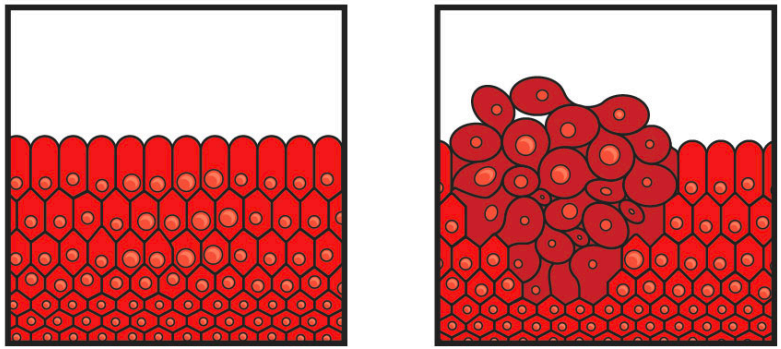
Two components of DELLY analysis gives SV call

- Paired-end mapping analysis
- Split-read analysis



Let's try Delly analysis to identify somatic SVs!

Real world



Normal cells

Cells forming a tumour

Model system

RPE1-WT

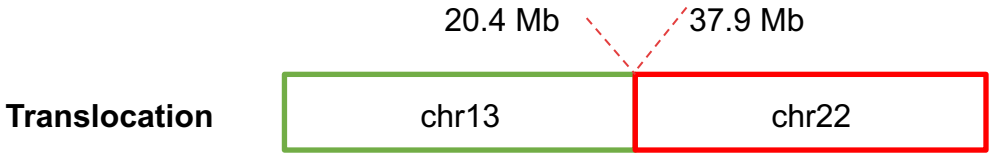
vs

RPE1-BM510

https://github.com/jeongdo801/SV_practical_computational

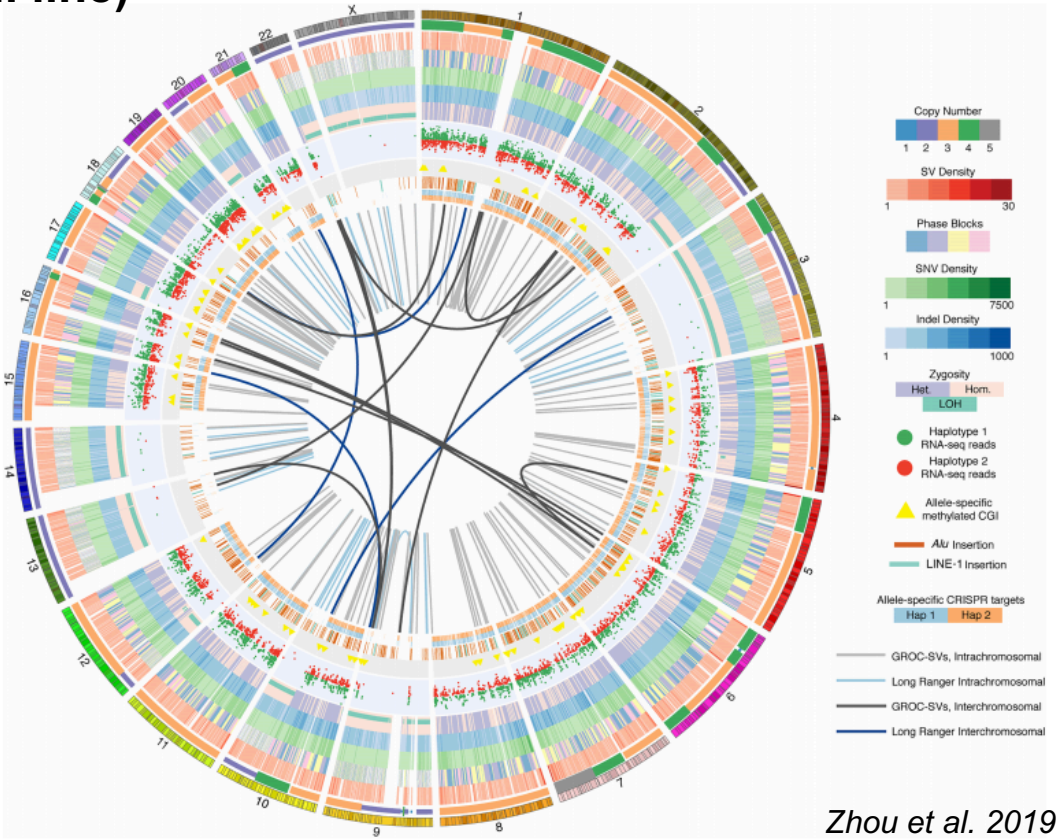
Delly example clonal somatic translocation (Pure BM510 VAF100% vs WT)

- CHROM, POS : [chr22, 37934425], [chr13, 20412250]
- MAPQ : mapping quality [60]
- GT : genotype [0/1] [0/0]
- GL : log10-scaled genotype likelihoods for RR, RA, AA genotypes [-23.6655,0,-90.8635] [0,-7.5218,-88.7961]
- GQ : Genotype Quality [10000] [75]
- FT : Per-sample genotype filter [PASS] [PASS]
- RC : Raw high-quality read counts for the SV [0] [0]
- RCL : Raw high-quality read counts for the left control region [0] [0]
- RCR : Raw high-quality read counts for the right control region [0] [0]
- CN : Read-depth based copy-number estimate for autosomal sites [-1] [-1]
- DR : # high-quality reference pairs [25] [23]
- DV : # high-quality variant pairs [5] [0]
- RR : # high-quality reference junction reads [27] [25]
- RV : # high-quality variant junction reads [9] [0]
- PRECISE/IMPRECISE :SVs refined using split-reads [PRECISE]
- PASS/LowQual : genotype quality [PASS]



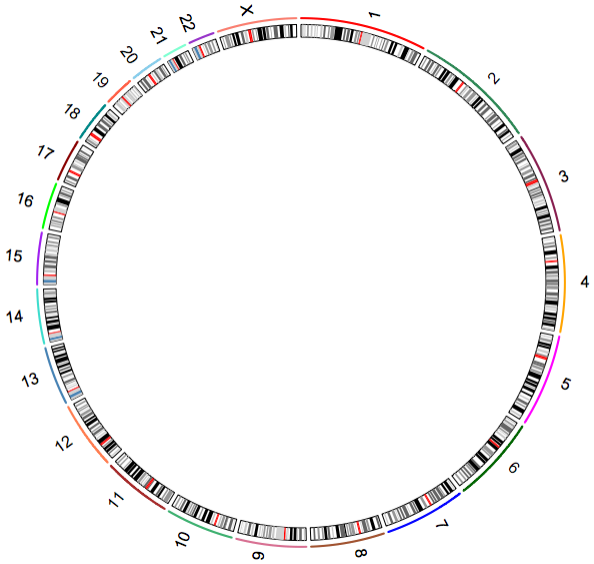
Can we visualize it with more fancy way?

Variety of SV information can be included in the circos plot (K562 cell line)

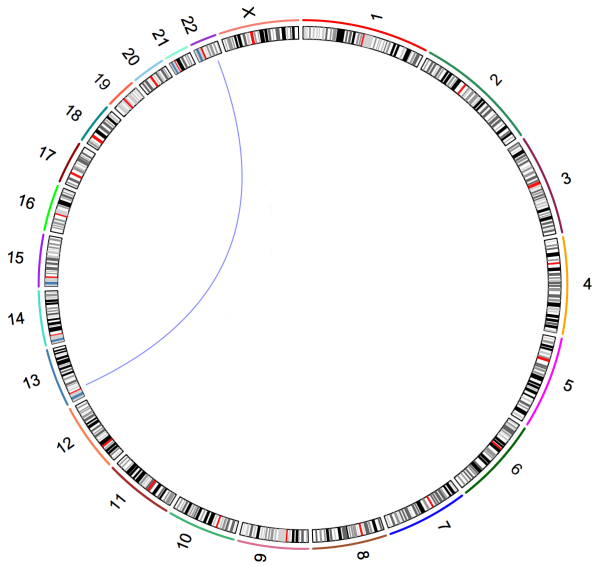


Zhou et al. 2019

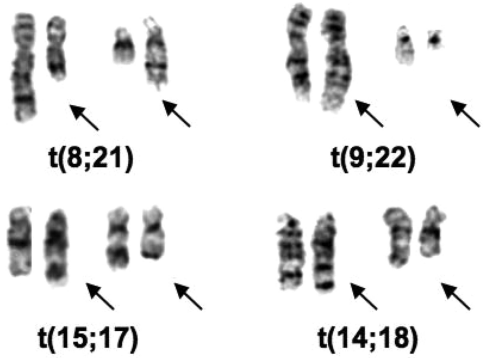
Visualize SVs using circos plot



Visualize SVs using circos plot

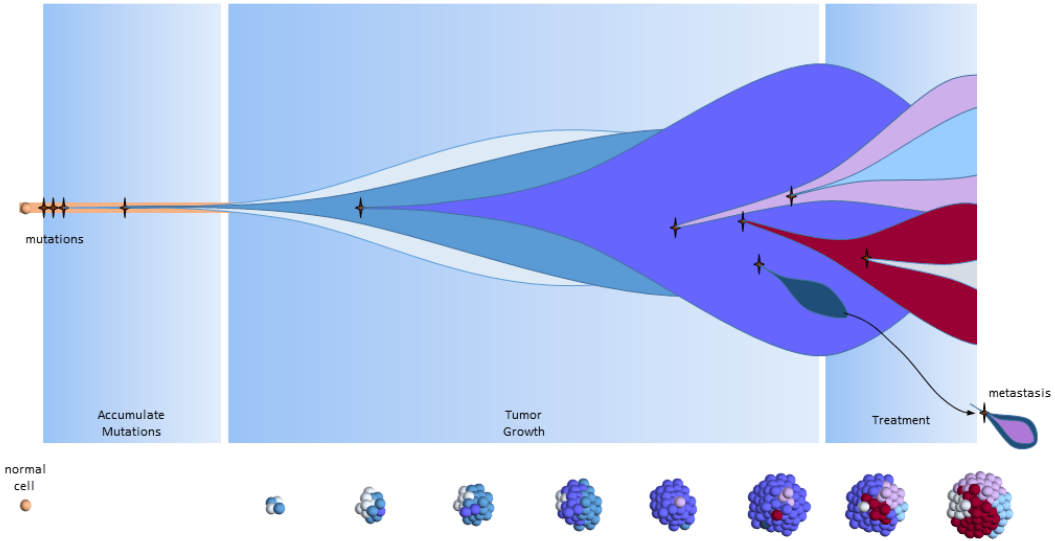


But what about sub-clonal variation?



(Rowley, J., D., Blood 2008)

10-40% of certain AML subtypes are driven by translocations

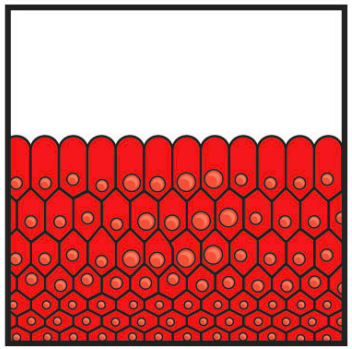


In theory, VAFs as low as 2.5% can be identified; assuming:

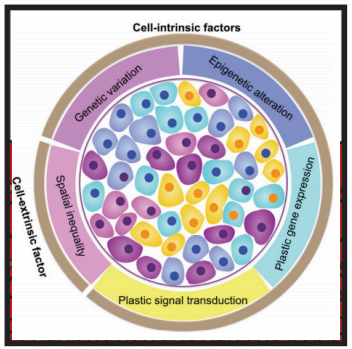
1. Known target locus/loci
2. Extremely high sequencing depth (20,000 - >100,000X)

Can we detect subclonal translocation using same method we practiced today?

Real world



Normal cells



Cells forming a tumour

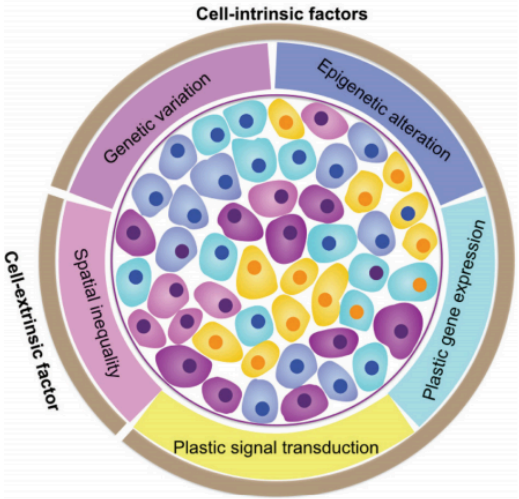
Model system

WT

vs

VAF20% BM510

What would be the best strategy to detect subclonal SVs?



- If certain variants have low VAF, it can be difficult to be confidently detected by bulk DNA sequencing approaches, what would be the reason?
- Question1: How can we more confidently detect subclonal SV event?
- Question2: How can we know which subset of cells carry SVs?

Thank you for listening

Day2 Single-cell approach