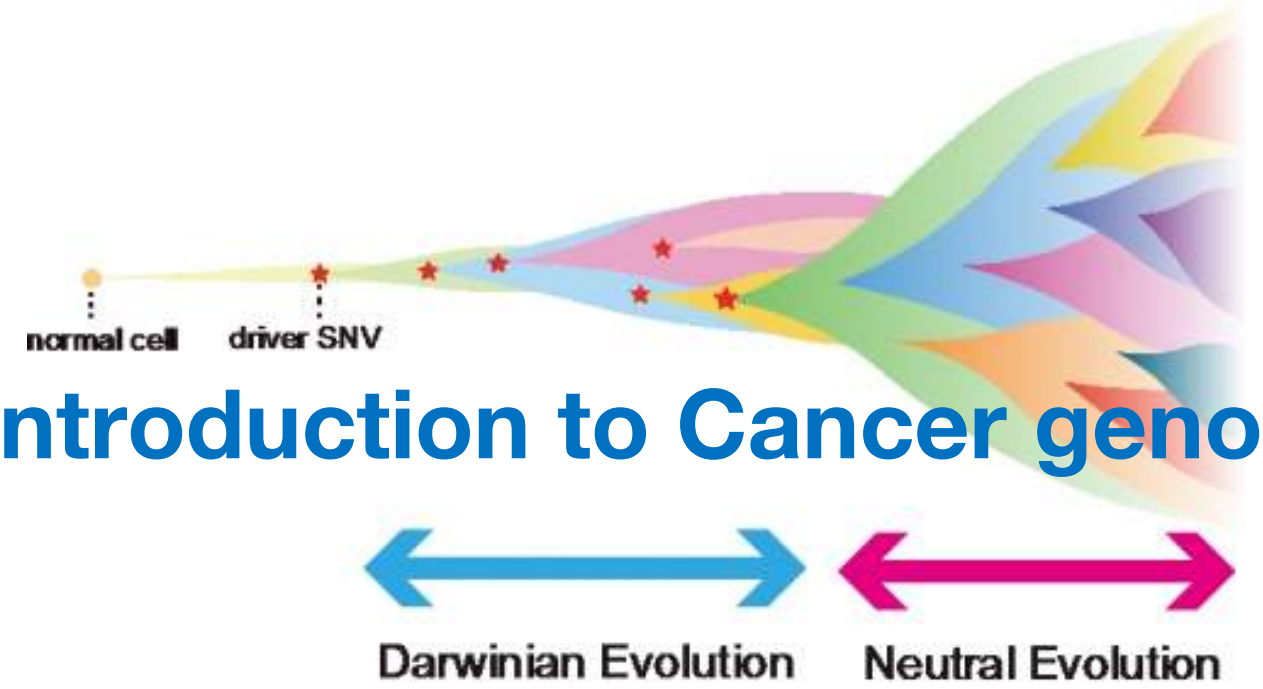


Introduction to Cancer genomics

Raheleh Rahbari

raheleh.rahbari@sanger.ac.uk

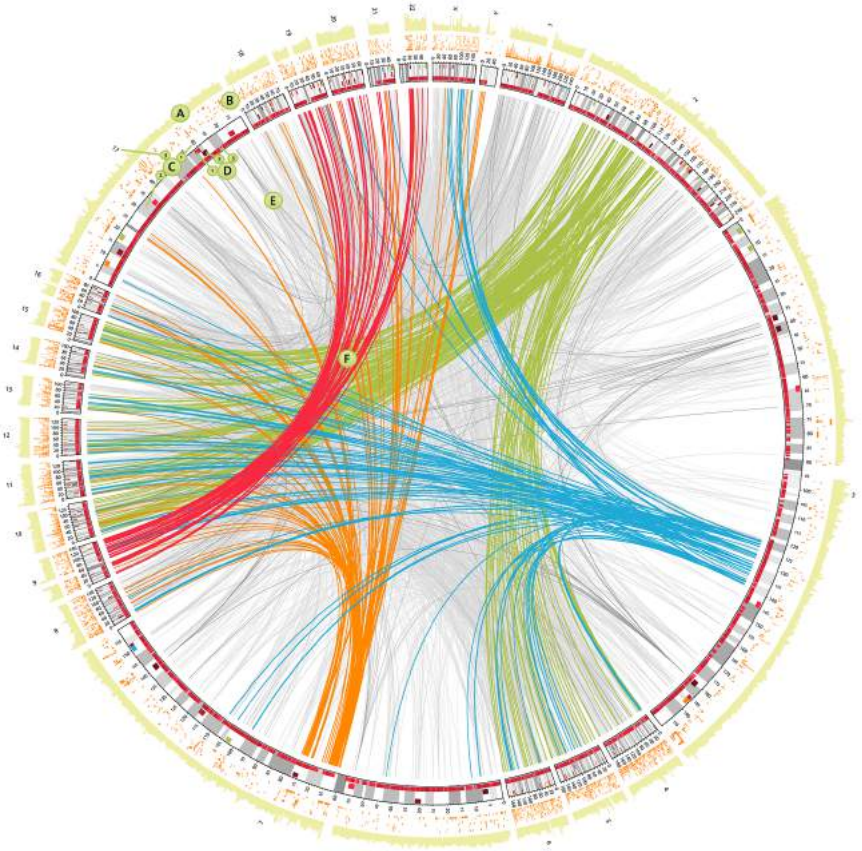
wellcome trust sanger institute

wellcome sanger institute

Learning objectives

- Cancer Vs. Normal genomes & sources of genome variation between the two
- Bioinformatic approaches to detect cancer genome variation
- How cancer genome analysis may be used to guide cancer patient management

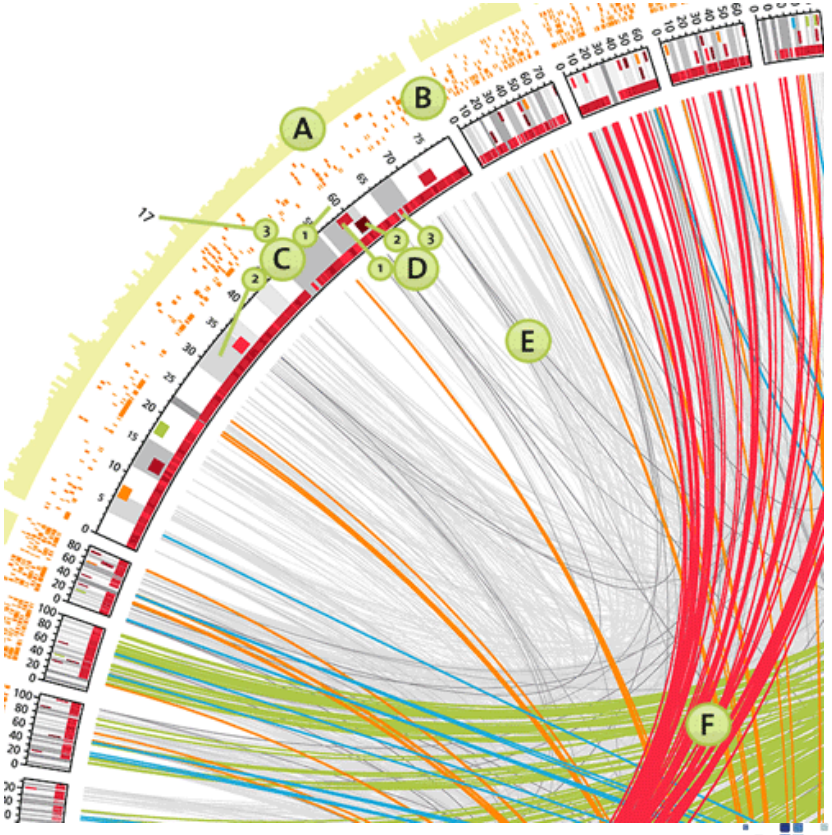
Human genome annotated with data related to genes implicated in disease



Circos, Martin Krzywinski et al., Genome Res. 2009

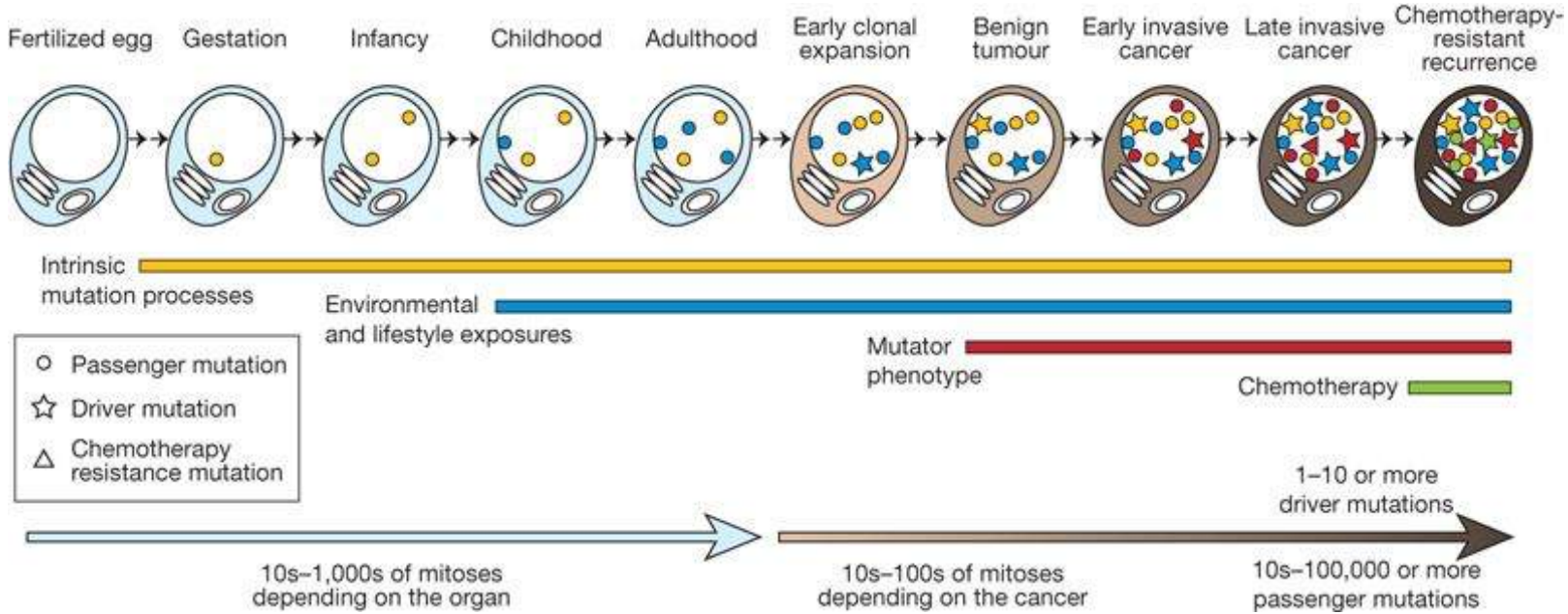
Ways that the genomes can be modified

- A Point mutations
- B Copy number alteration
- C Structural rearrangements
- D Genes & regulatory elements
- E Pathways
- F Homologous and epigenetic modifications



wellcome sanger institute

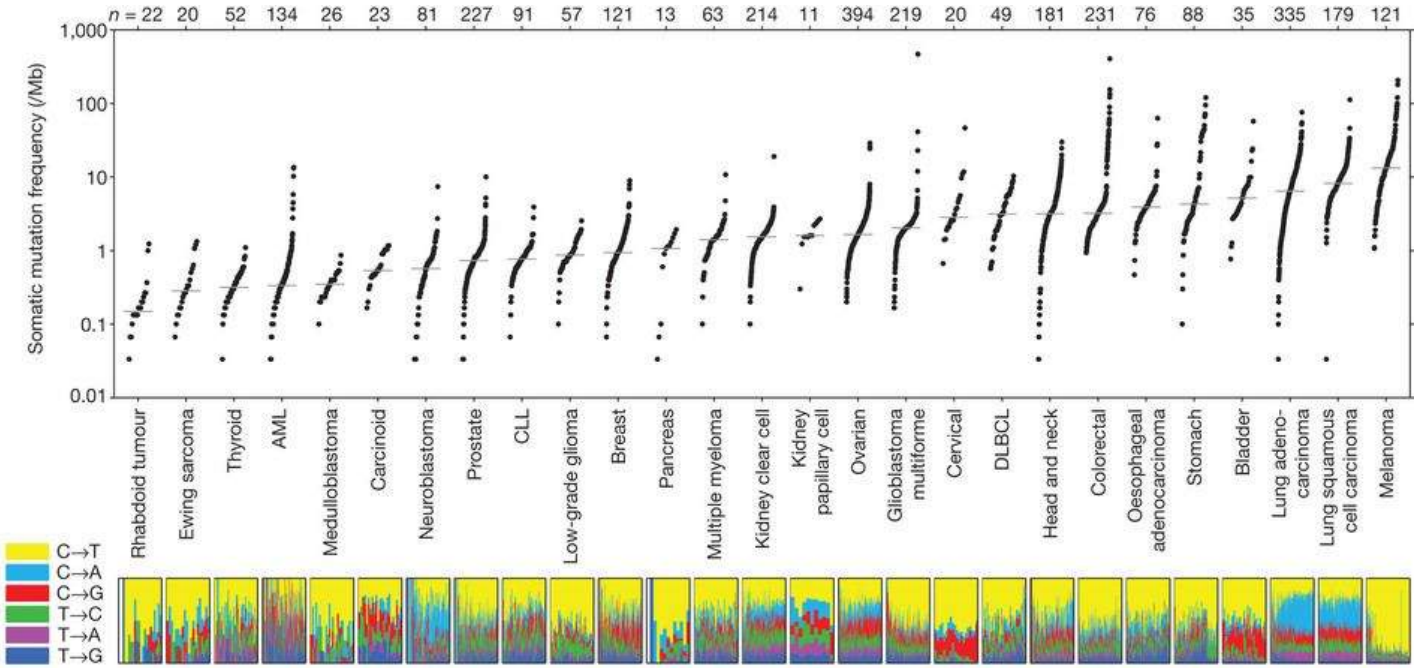
The timing of the somatic mutations



Relapse after chemotherapy can be associated with resistance mutations that often predate the initiation of treatment.

Michael Stratton, Peter J. Campbell, and P. Andrew Futreal, Nature 2009

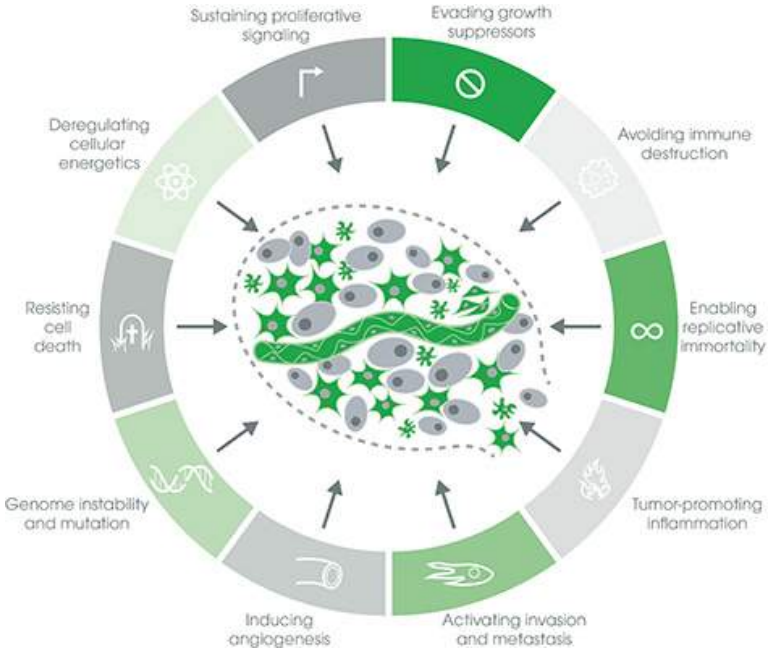
Somatic mutation frequencies observed in exomes from 3,083 tumor-normal



Lawrence et al., nature 2013

Hallmarks of cancer

Tumor calls acquire abnormal abilities by co-opting normal cell behavior

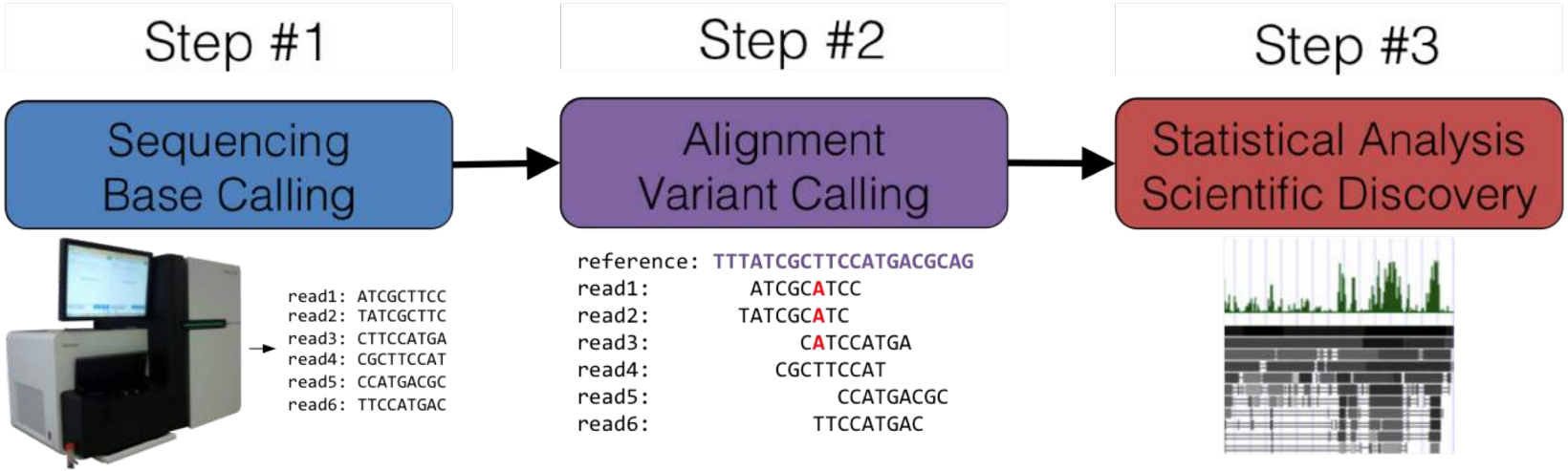


Hanahan and Weinberg, Cell 2011

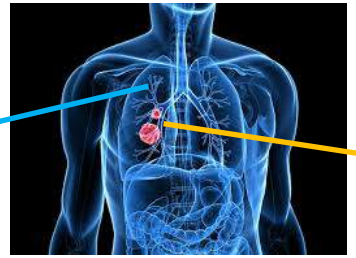
Applications of next-generation sequencing to cancer

wellcome sanger institute

General DNA sequencing workflow



Normal Sample



Cancer Sample

Gene A

Gene B

Gene C

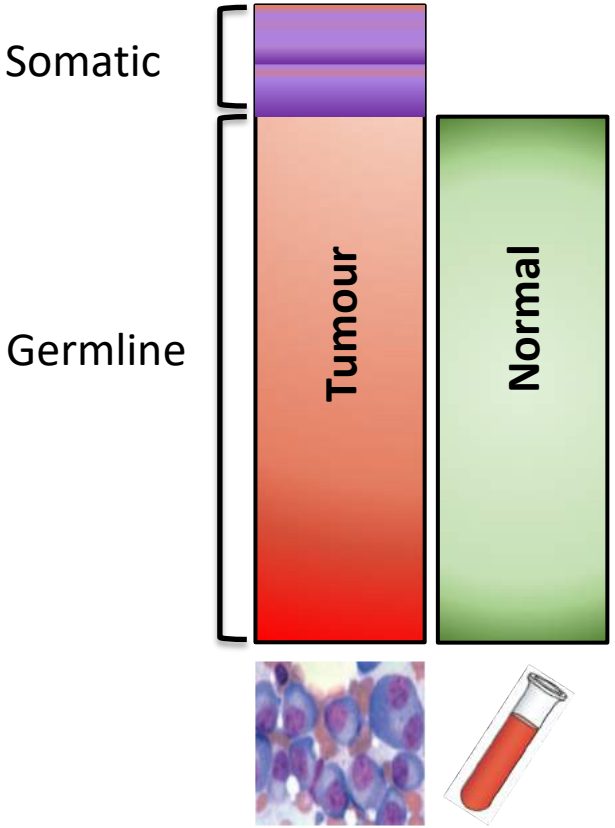
SNP c.34T>A
Germline Variant
(Inherited)

c.244C>T

Somatic Mutation (Acquired)

SNP c.34T>A

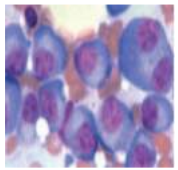
Background mutation detection



Background mutation detection

Drivers
- Implicated in oncogenesis
- Reside in cancer genes
- Inform biology
- Targets for therapy

Known cancer genes
Novel cancer genes



Passengers
- biologically inert
- reflect DNA exposures
- life History of Cancer

Signature
- spectrum of variants
- sequence context

wellcome sanger institute

Background mutation detection

Drivers

- Implicated in oncogenesi
- Reside in c
- Inform bio
- Targets for

Known cance

Novel cancer

Passengers

- biologically inert

NA exposures

y of Cancer

of variants

context

Challenges

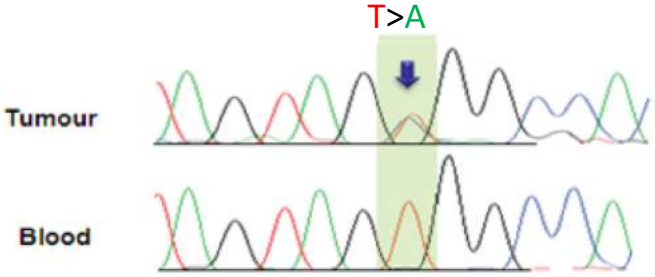
1 Need to be able to detect subtle mutations

2 Tumour contaminated with normal

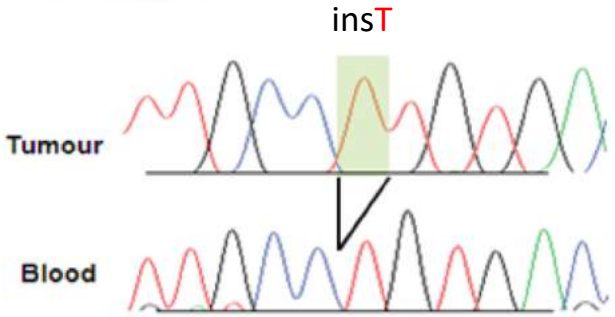
3 Large variation in ploidy

Class of mutations

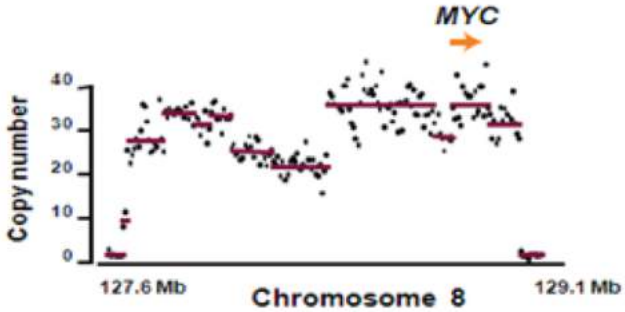
Base Substitutions



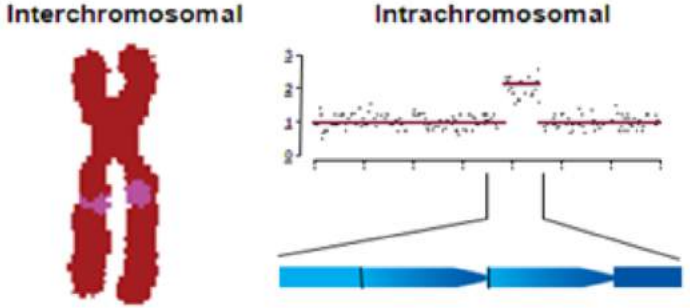
Insertions & deletions



Copy number changes



Rearrangements



wellcome sanger institute

Background: Genetics & Somatic Substitutions

- Base Pair Substitution – single base pair is changed
- Most common mutation type

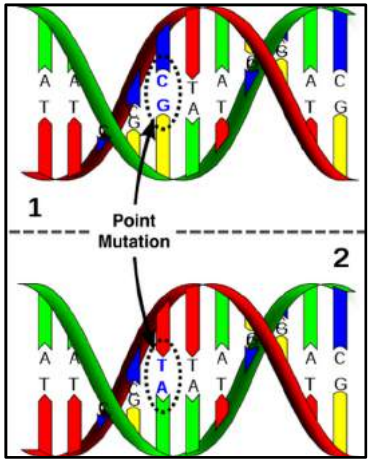
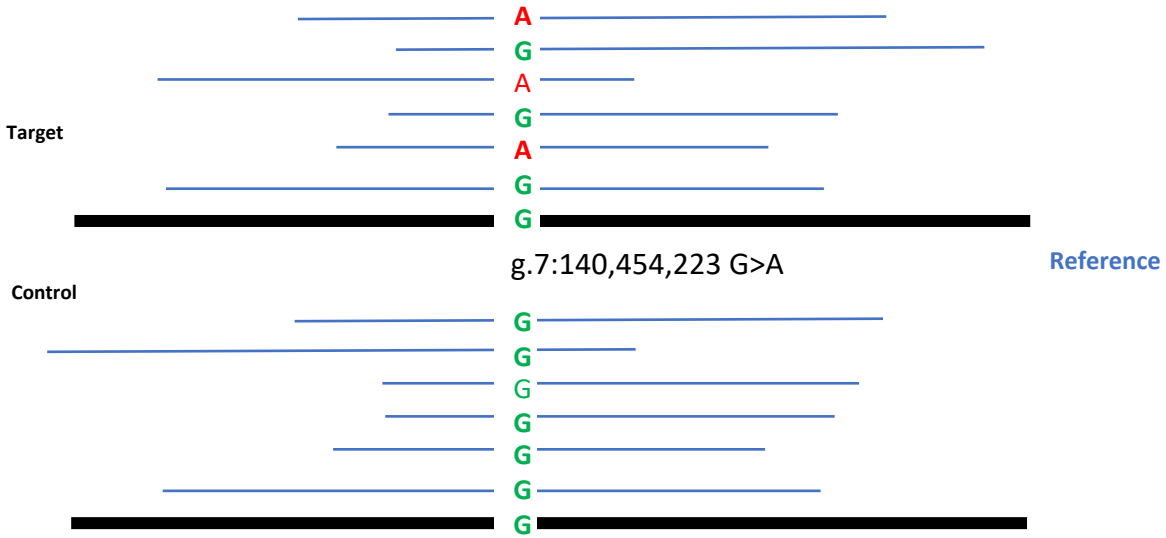


Image from http://rosalind.info/glossary/point-mutation/

- When occur in a gene
  - Silent
  - Missense
  - Nonsense



Artefact Mutations

- ❑ Sequence errors
- ❑ PCR errors
- ❑ Sample Prep errors (FFPE, 8-oxoG)

Analysis Challenges

- ❑ Clonal composition (Aberrant cell fraction)
- ❑ Changes in Copy Number
- ❑ Data Analysis (Mapping)

Sensitivity:

- ❑ Call mutations at ~10% VAF

wellcome sanger institute

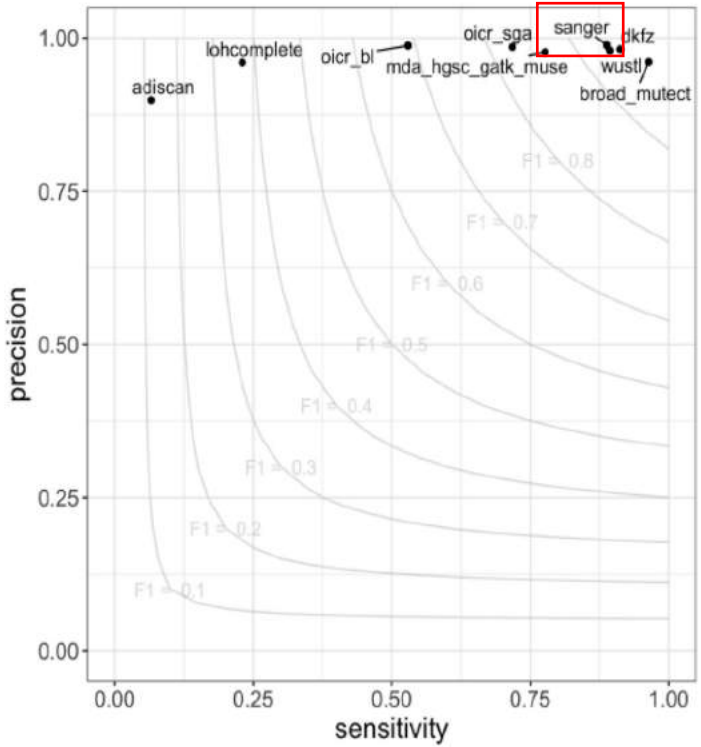
Which Somatic Mutation Substitution Caller should I use?

WGS/WES/Targeted Analysis

Caveman with

Targeted Shearwater

- Subclonal variants in deeply sequenced samples, can be unmatched
- Must ensure normal panel is optimised for your specific analysis



ICGC/TCGA Pilot-63, 250,000 of 6.4 Million mutations validated by deep sequencing on Illumina HiSeq

wellcome sanger institute

- User can specify a single global copy number estimate. Default is 2/5.
- Alternatively can use copy number inferred by ASCAT

At each position the algorithm then considers all possible genotypes that can be made from a single variant allele. The ploidy and copy number determine possible genotypes

Space of Genotypes (Normal Ploidy=N_p, Tumour Copy Number=T_p)

Let A represent the reference allele and B the variant allele (here $N_p = 2$ and $T_p = C$)

Somatic Genotypes: $G_N = AA$ and $G_T = \underbrace{A \ldots A}_{C-k} \underbrace{B \ldots B}_{k\ copies}$ where $1 < k \leq C$

SNP Genotypes: $G_N = AB\ or\ BB$ and $G_T = \underbrace{A \ldots A}_{C-k} \underbrace{B \ldots B}_{k\ copies}$ where $0 \leq k \leq$C

Where G_N and G_T are the genotypes of the Normal and Tumour respectively.

Doesn't handle:

- Somatic mutations at SNP sites
- Multi-allelic variants

wellcome sanger institute

Caveman directly estimates genotype probabilities

$$P(G_N, G_T|D) \propto P(D|G_N, G_T)P(G_N, G_T)$$

Incorporates prior probability of germline/somatic mutation.

Where D is the set of overlapping reads at position p. Here we focus on the first term

$$P(D|G_N, G_T)=\prod_{R\in Normal\ Reads} P^{(n)}(R_p|G_N, G_T, \theta)\ \prod_{R\in Tumor\ Reads} P^{(t)}(R_p|G_N, G_T, \theta)$$

Let genotype $\psi(G)$=B(G)/(A(G)+B(G)) represent the proportion* of non-ref base (B) in genotype G. Let c be the proportion of normal contamination of tumour. The covariates for read R at position p are denoted by θ and let the called base be denoted C

$\check{c} = \frac{cN_p}{cN_p+(1-c)N_T}$: Contamination is adjusted for different copy number in normal and tumour

$$P^{(n)}(R_p|G_N, G_T, \theta) = \underbrace{\psi(G_N)}_{\substack{Prob.Read\ from\ a\ chromosome\ carrying\ B\ allele}} \mathrm{P(C}|D = B, \theta) + \underbrace{(1-\psi(G_N))}_{\substack{Prob.Read\ from\ a\ chromosome\ carrying\ A\ allele}} \mathrm{P(C|D = A}, \theta)$$

Profile probability of the called base, C, given the true base D (here A) and other covariates θ e.g base quality, strand etc

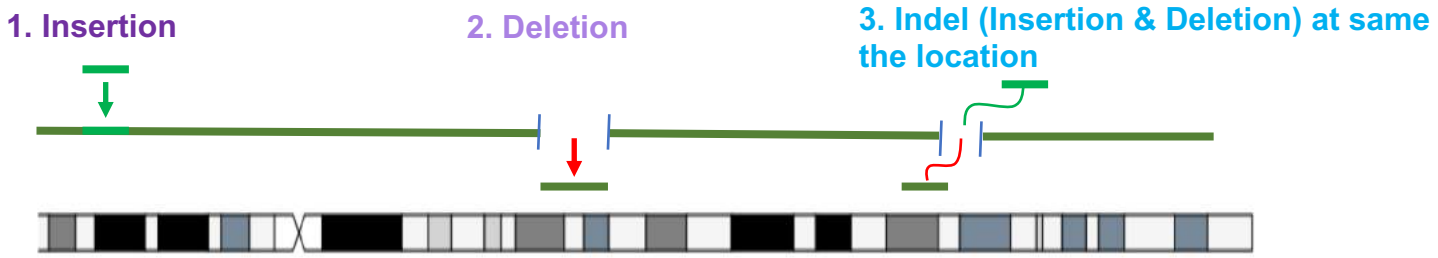
$$P^{(t)}(R_p|G_N, G_T, \theta) = \underbrace{\check{c}P^{(n)}(R_p|G_N, G_T, \theta)}_{\substack{Read\ from\ Normal\ accounting\ for\ contamination}} + \underbrace{(1-\check{c})\{\psi(G_T)\ \mathrm{P}(C|\mathrm{D = B}, \theta) + (1-\psi(G_T))\ \mathrm{P(C|D = A}, \theta)\}}_{\substack{Read\ comes\ from\ Tumour\ accounting\ for\ contamination}}$$

* Implementation also includes reference bias correction

wellcome sanger institute

Background: Genetics & somatic ins/dels

- Small Insertion & Deletions
- Less than 200base pairs
- Are known to drive cancer development
- 3 Classes:



Can be disruptive if occur in genes (switch genes on/off)

- Inframe
- Frameshift

wellcome sanger institute

Background: Genetics & somatic ins/dels

Artefact Mutations

- ❑ Sequence errors (bubbles on flow cell)
- ❑ PCR errors
- ❑ Sample Prep errors (FFPE, 8-oxoG)

Analysis Challenges

- ❑ Clonal composition (Aberrant cell fraction)
- ❑ Changes in Copy Number
- ❑ Data Analysis Mapping
  - ❑ Accuracy of mapping
  - ❑ Alignment left or right justified

Sensitivity:

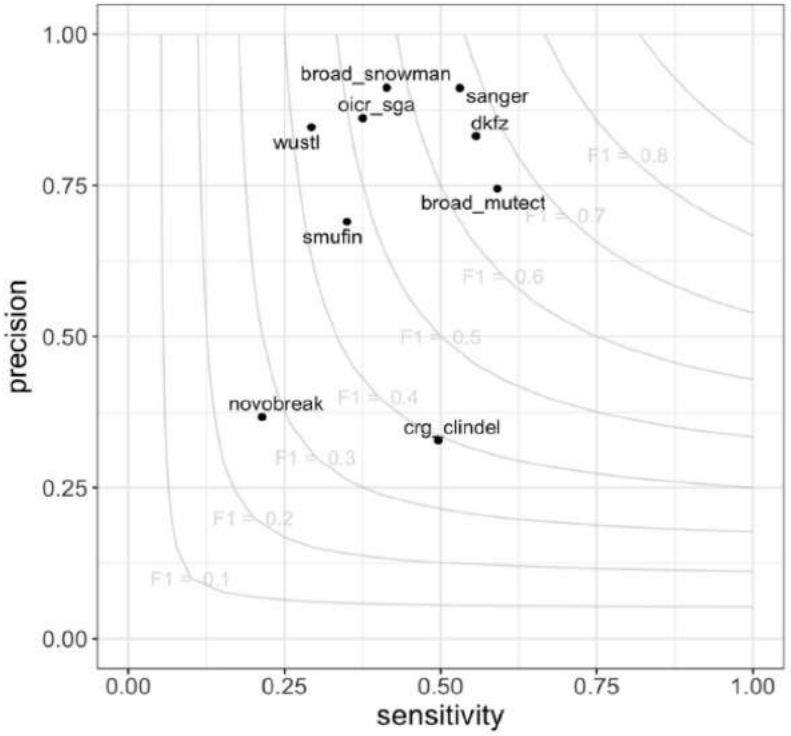
- ❑ Call mutations at ~10% VAF

wellcome sanger institute

Which Somatic Mutation Ins/dels Caller should I use?

WGS/WES/Targeted Cancer Analysis

Pindel



Pindel

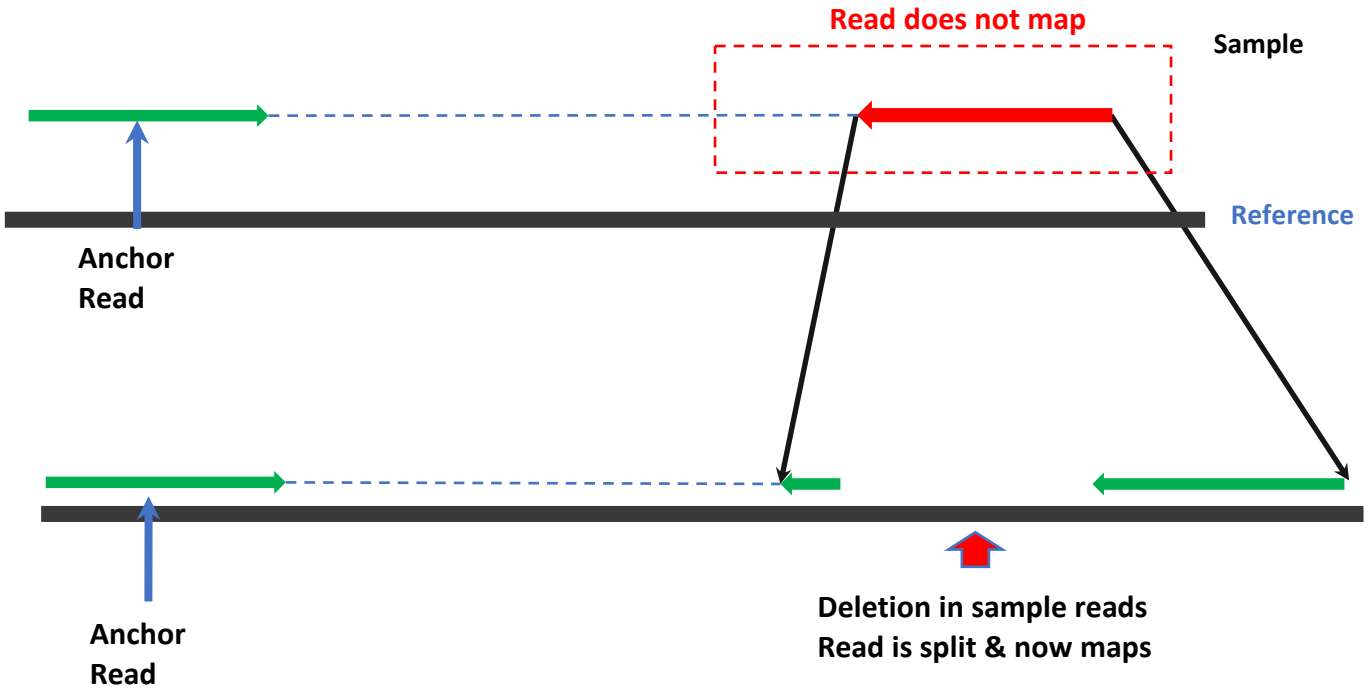
- Pattern Growth Approach (sequential pattern mining algorithm)
- Adapted to handle complex structural variants

Pindel optimised to call somatic small ins/dels in cancer samples
Requires Tumour and Matched-Normal Bam files
Detect Deletions > 10kb
Detect Insertions (factor of read length) ~50bp
Can also call tandem duplications, inversions

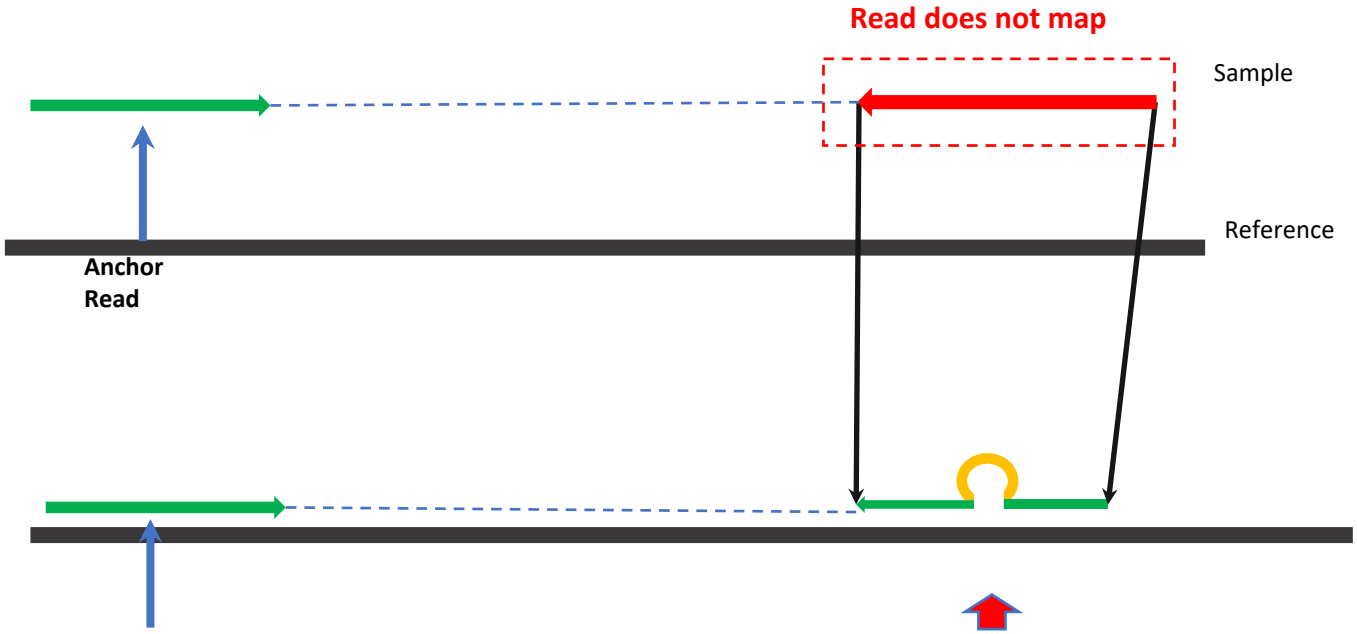
Pindel Ye K et al, Bioinformatics vol 25 no.21 2009 p2865-2871 & Raine K et al, Curr Protoc Bioinf. 52:15.7.1-15.7.12

wellcome sanger institute

Pindel: How it works for a deletion



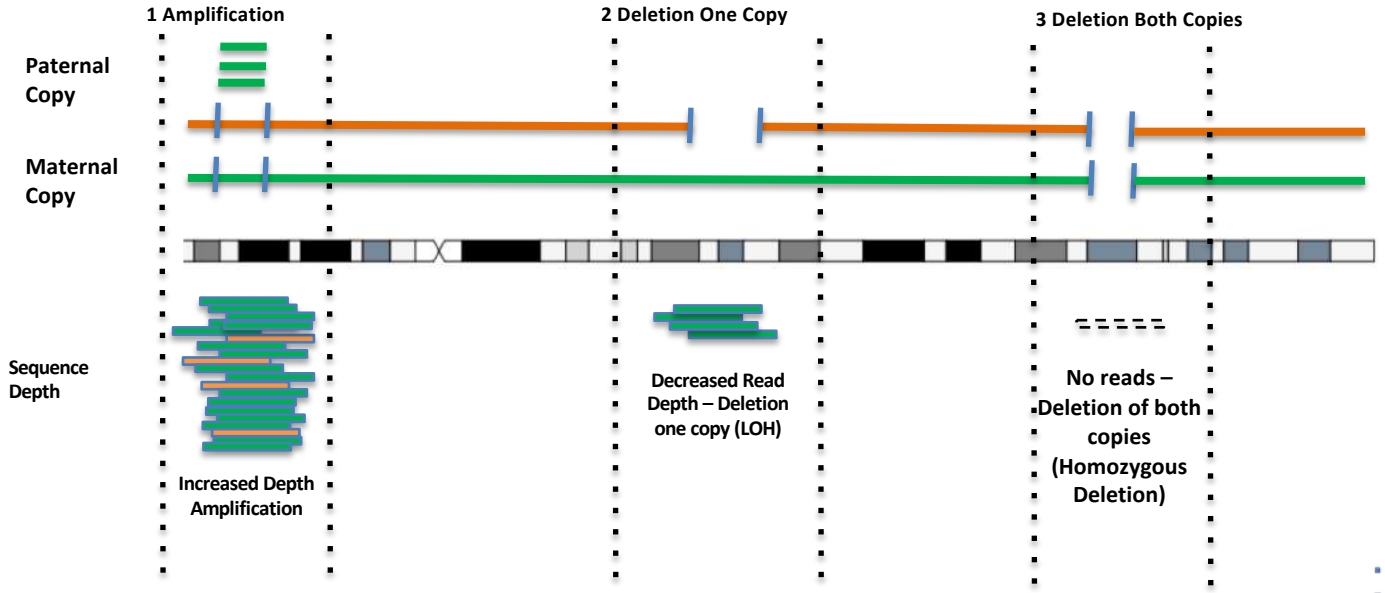
Pindel: How it works for an insertion



Background: Copy Number Algorithms

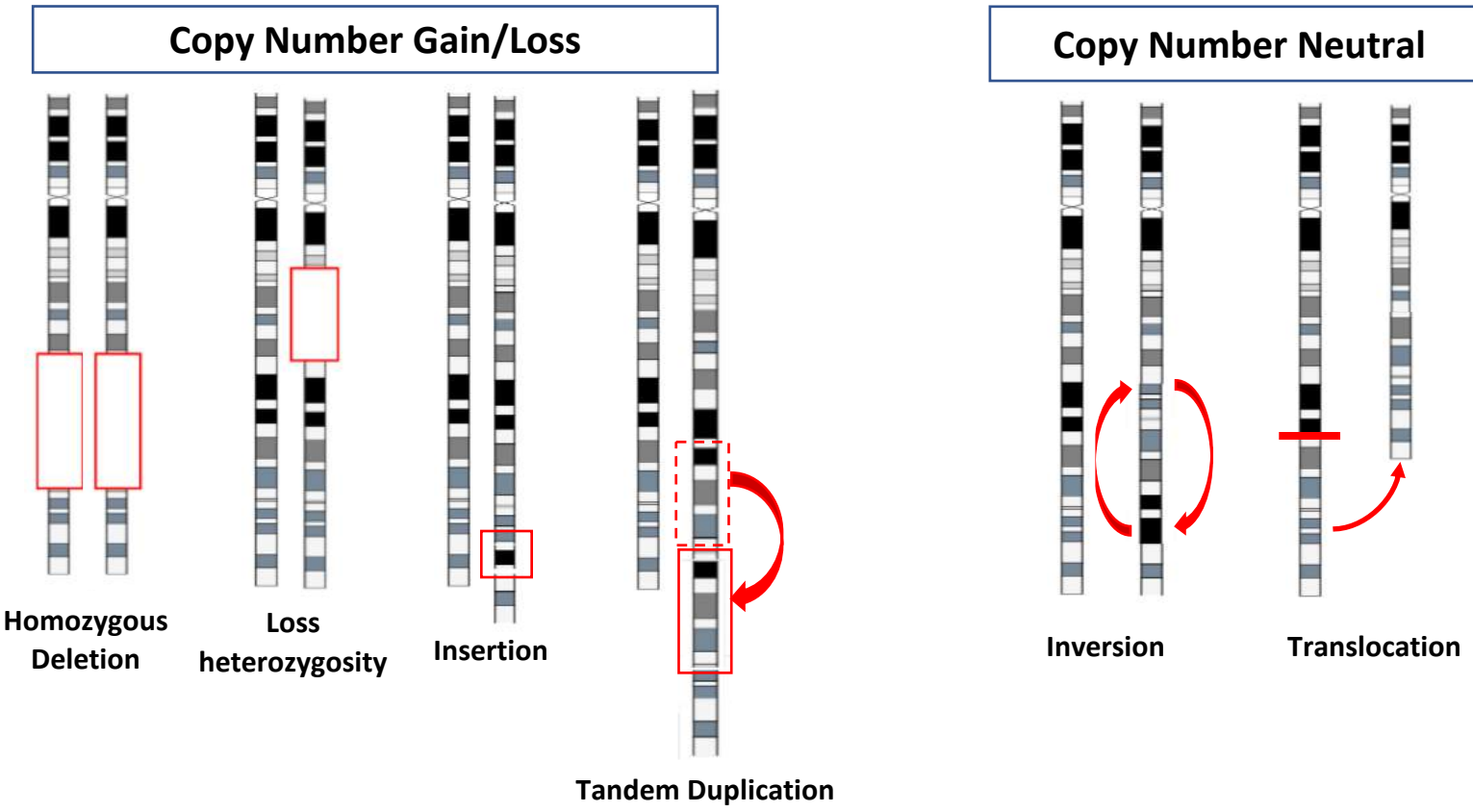
Algorithms use the read depth to detect Copy Number changes

- Increase number of reads -> Amplification
- Decreased number of reads -> Deletion One Copy (Loss of Heterozygosity, LOH)
- No reads -> Deletion Both Copies (Homozygous Deletion)



In Addition, by using Germline Information (SNPs) can also work out which allele is affected (paternal or maternal)

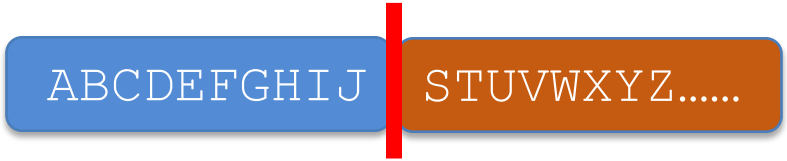
Types of Structural Variants



Scherer SW et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007 Jul;39(7 Suppl):S7-15.

DNA Breakpoint: Signposts for Structural Variants

DNA Breakpoint: All Structural Variants are characterised by two unexpected fragments of DNA being next to one another

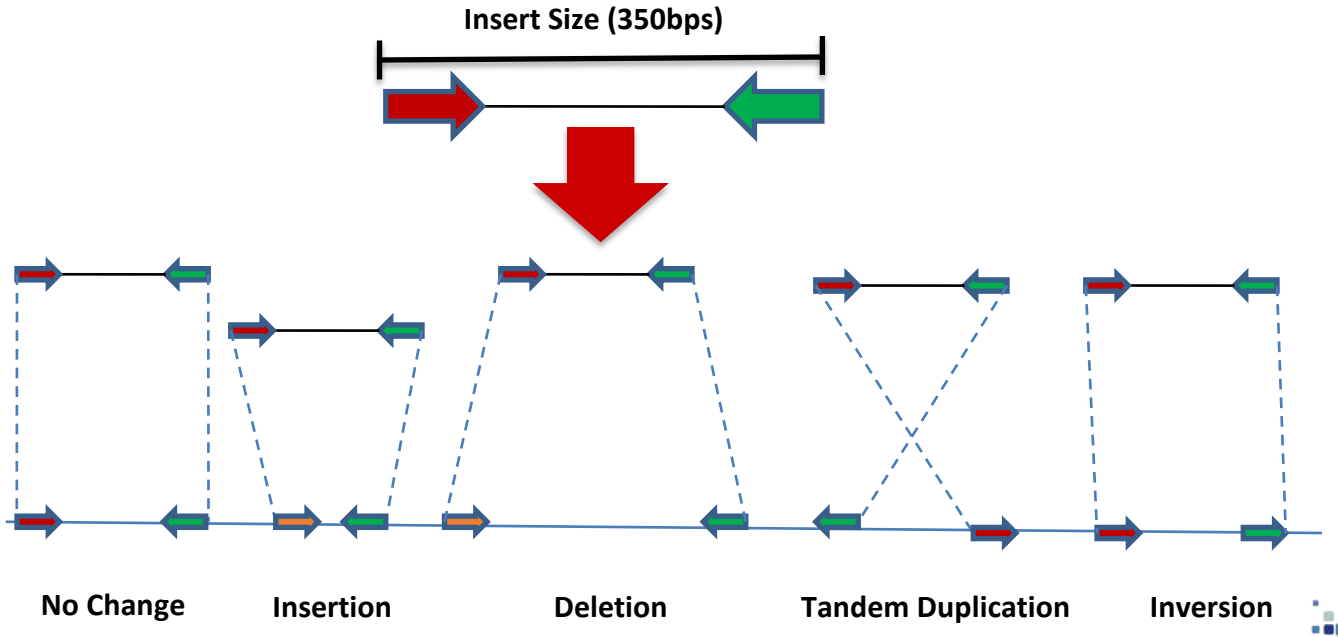


1 Breakpoint – Deletions, Tandem Duplications
2 Breakpoints – Insertions, Inversions, Translocation
Many Breakpoints – Amplicon, Chromothripsis, Chromoplexy

wellcome sanger institute

1. Paired End Analysis (aberrant read mapping)

Expect 2 reads to map with set gap (insert size), if not they could indicate structural rearrangement



2. Read depth

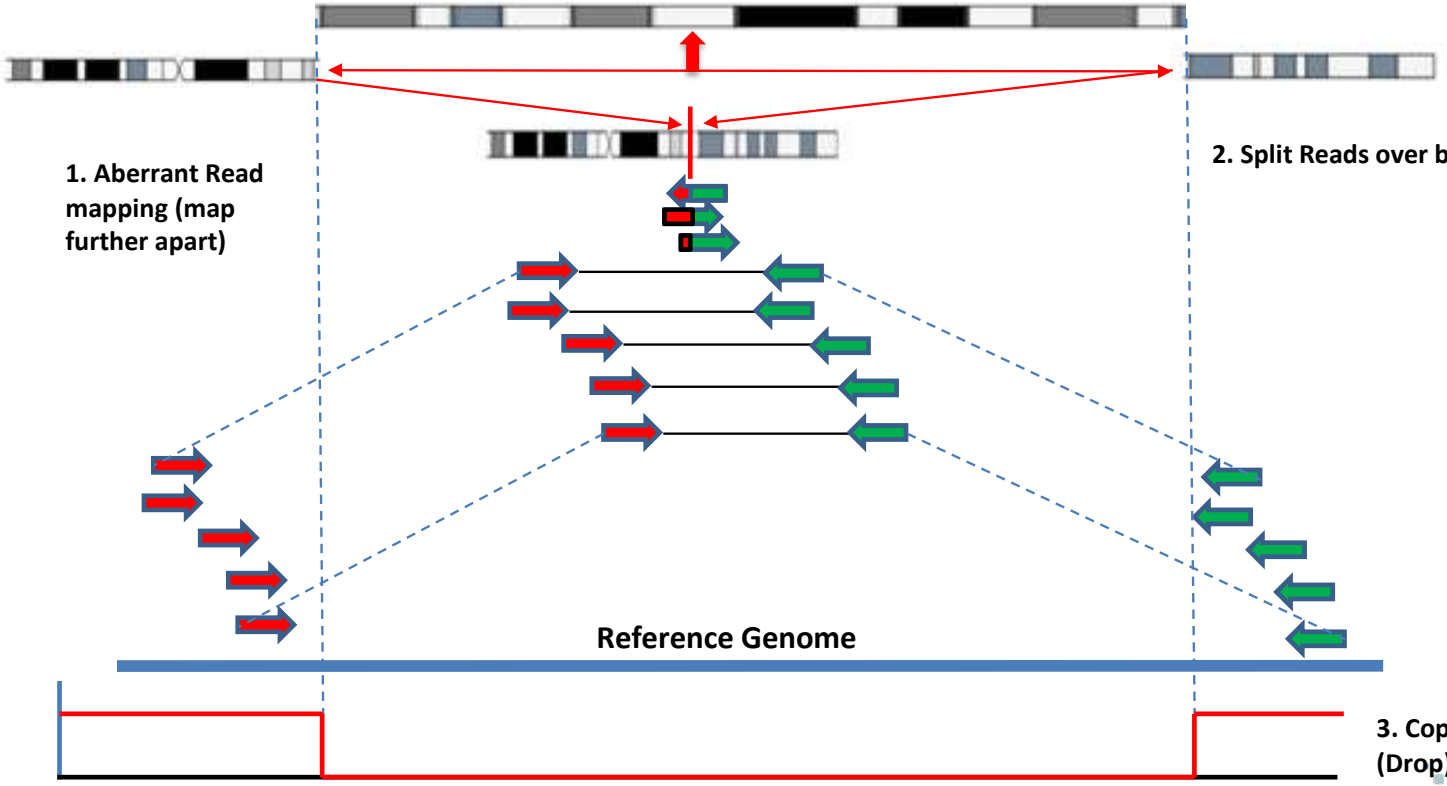
The Copy Number either side of a breakpoint can help confirm type of structural variant and validate if real

1. Copy No Change (Drop), LOH

2. Copy No Change (Drop), Homozygous Deletion

3. Copy No Change (Increase), Insertion, Tandem Duplication

Example: Signposts for a Deletion Breakpoint



Allele-Specific Copy number Analysis of Tumours

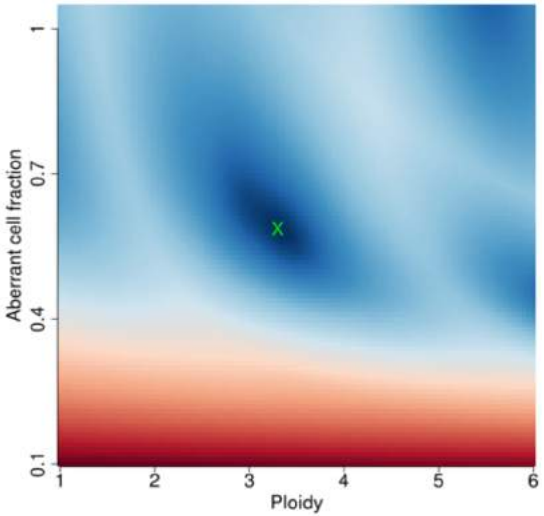
- Developed by Peter Van Loo *et al*
- Can detect Copy Number Alterations (CNAs) in Tumours
- Allelic-specific – paternal or maternal changes
  (Often cannot determine this so algorithm identifies major/minor copy number changes)
- Estimates the Absolute Copy Number
- Estimate normal & tumour composition of a sample

ASCAT, Van Loo*, Nordgard* *et al*. (2010), PNAS 107:16910-5 AscatNgs Raine K *et al (2016)* Current protocols in bioinformatics 56 (1), 15.9. 1-15.9. 17

wellcome sanger institute

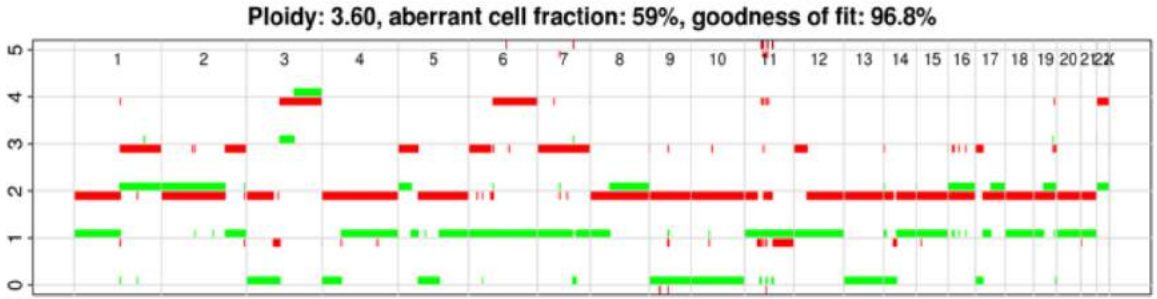
ASCAT results - Plots

Sunrise Plot



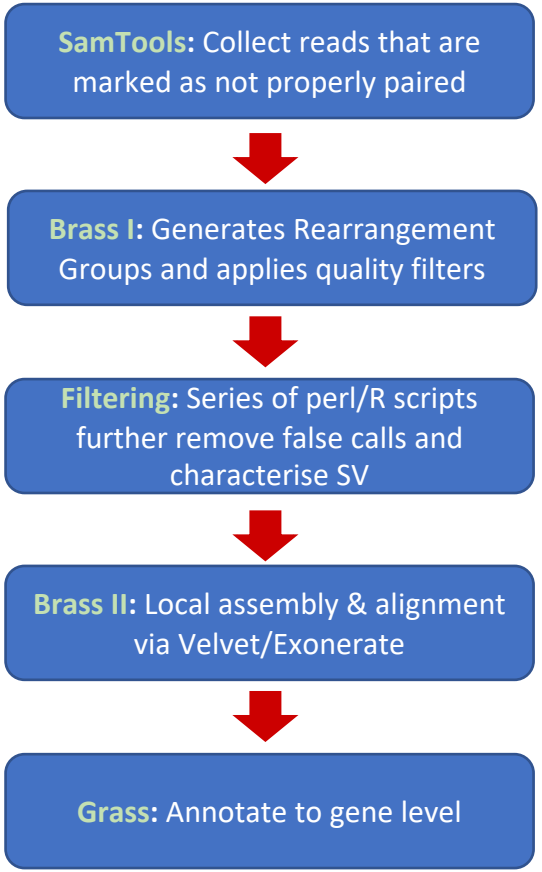
Sunrise Plot – This plot shows the goodness of fit against sample ploidy and purity. The green cross indicates the area of best fit (dark blue)

Copy Number Profile



Copy Number Profile – This is the final result with integer (clonal) copy number states. Red indicates the major and green the minor allele copy number

Brass Overall Workflow



wellcome sanger institute

Reading list

- **Somatic mutant clones colonize the human esophagus with age:**
http://science.sciencemag.org/content/362/6417/911/tab-article-info

- **Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.**
https://www.ncbi.nlm.nih.gov/pubmed/25999502

- **Tissue-specific mutation accumulation in human adult stem cells during life**
https://www.ncbi.nlm.nih.gov/pubmed/27698416

raheleh.rahbari@sanger.ac.uk

wellcome sanger institute