Introduction to somatic variant calling

IN-BIOS5000/IN-BIOS9000, Friday, 2020-10-30

Daniel Vodák

Outline

- General background
    - Somatic and germline variation
    - Motivation for somatic variant calling
    - Sequencing (and read mapping/alignment)
- Variant types
    - Variant-derived metrics
- Variant calling, classification and annotation
- Factors that complicate somatic variant calling
- Artifact variant calls
- General advice for somatic variant calling

Germline variation

- **Germline variant**
  - *“A variant present at the point of an individual’s conception. The variant **is present in every cell in the**ir **body** and can be passed on to the next generation [is **hereditary**].” (https://www.genomicseducation.hee.nhs.uk/glossary/germline-variant)*
- Healthy human reproductive cells are haploid (having 1 sex chromosome and 1 copy of each autosome)
  - Merging of two rep. cells at conception, subsequent division and differentiation
- Healthy human body cells are diploid (having two sets of chromosomes: sex chromosomes + autosomes)
  - Heterozygous and homozygous germline variants

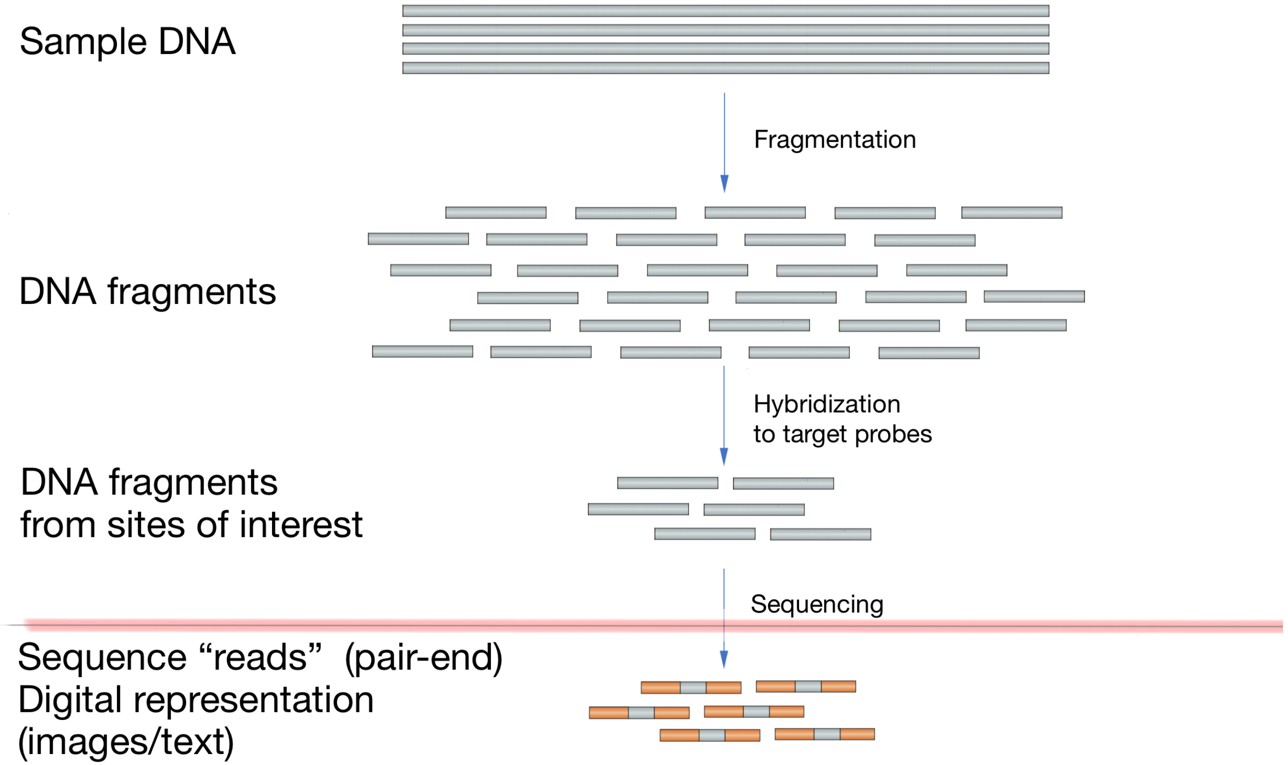
Somatic variation

- **Somatic variant/mutation**
  - *“**A variant arising in somatic cells** which can therefore not be passed on to the next generation.*
    *The DNA within somatic cells may change **during an individual’s lifetime**, but these changes will not be passed on to any children [is **not hereditary**]. Such genetic variants can, however, be **responsible for some conditions such as cancer**.”*
    (https://www.genomicseducation.hee.nhs.uk/glossary/somatic-variant)
- Cancer cells contain both, germline and somatic variants
- Cancer cells are often aneuploid – they lose or gain chromosomes or their parts
- “Somatic variant calling” is most often used in the context of cancer
  - Discovering genomic changes specific to the cancer cells

Motivation for somatic variant calling

- Definition of tumor subtypes, stratification of patients
- Some variants provide sensitivity or resistance to specific treatment
    - Personalized/precision medicine - treatment decisions based on presence/absence of specific somatic variant(s)
- Discovery of “driver variants” (as opposed to “passengers”) –mutational events that are crucial for tumor development
    - Understanding the biological mechanisms behind the various cancer types
    - Activation of oncogenes and deactivation of tumor suppressor genes
- How much is there to research?
    - Differences between two individuals with the same cancer type
    - Differences between the cells of a given tumor (tumor heterogeneity)
        - A tumor genome is evolving, sometimes in a reaction to treatment

DNA sequencing



Sample DNA

Fragmentation

DNA fragments

Hybridization
to target probes

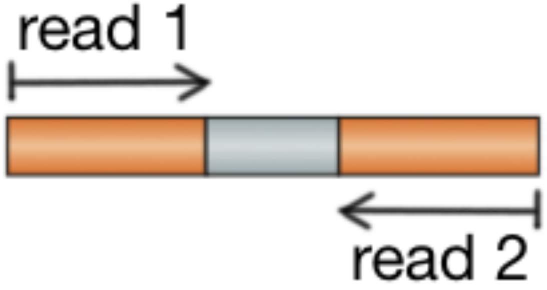
DNA fragments
from sites of interest

Sequencing

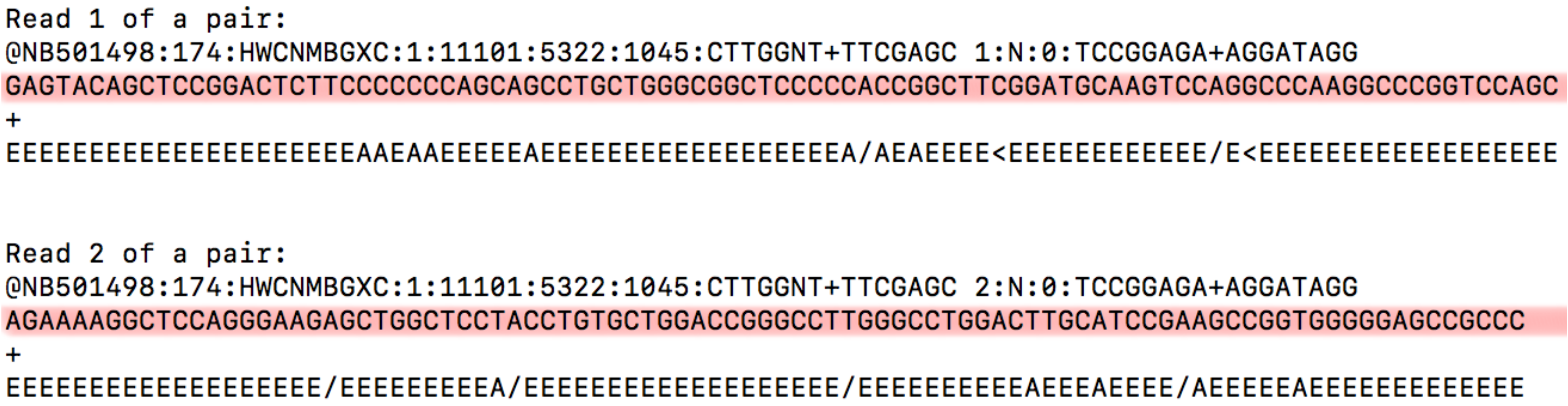
Sequence “reads” (pair-end)
Digital representation
(images/text)
of the sequenced DNA fragments

- Insert, insert size
- Short vs. long fragments

- Whole genome, targeted sequencing (e.g., exome)



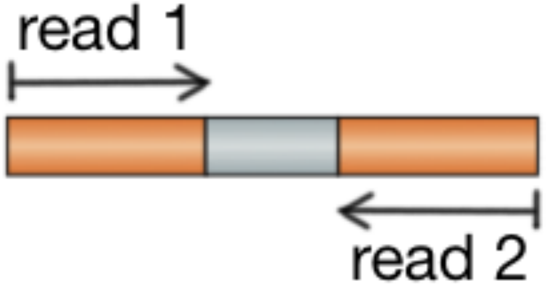
Sequencing reads – an example



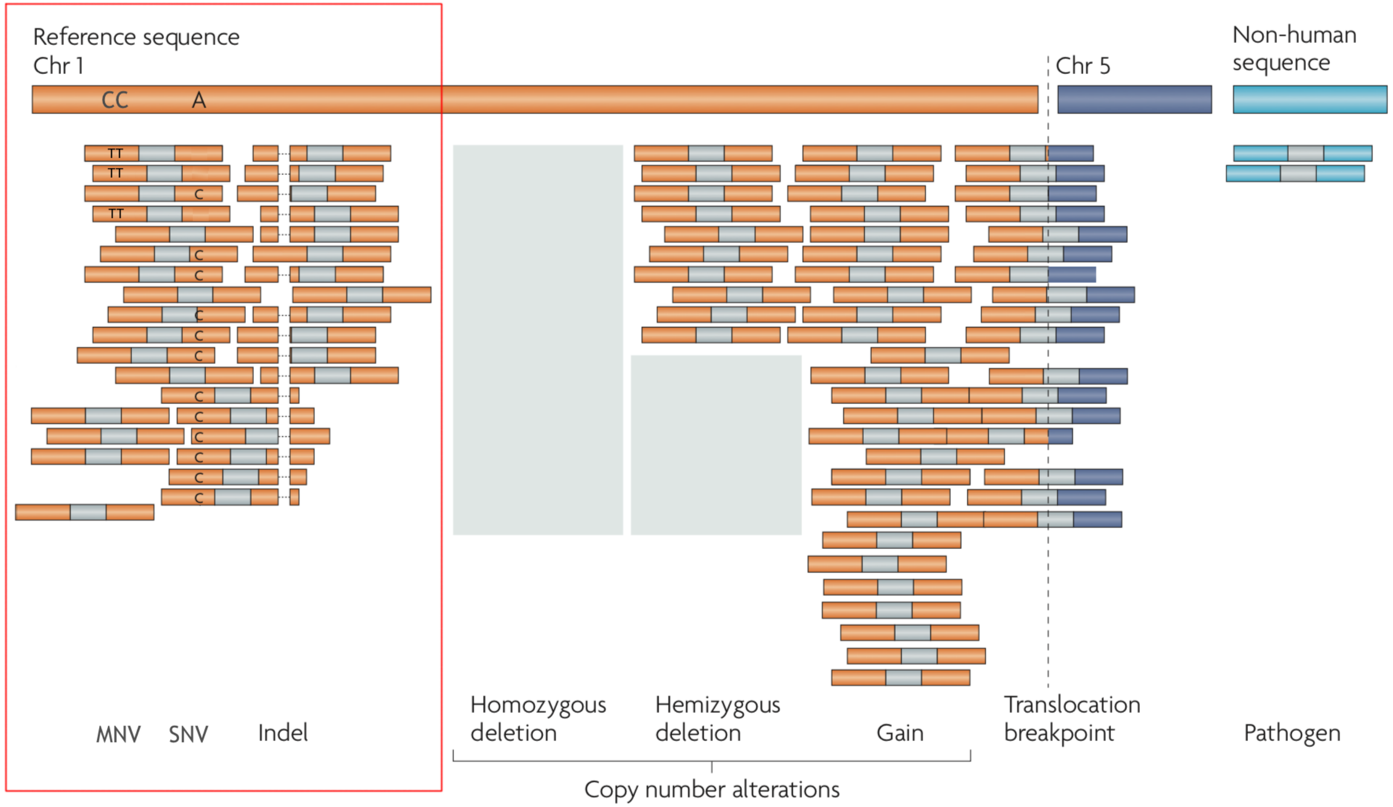
```
Read 1 of a pair:
@NB501498:174:HWCNMBGXC:1:11101:5322:1045:CTTGGNT+TTCGAGC 1:N:0:TCCGGAGA+AGGATAGG
GAGTACAGCTCCGGACTCTTCCCCCCCAGCAGCCTGCTGGGCGGCTCCCCCACCGGCTTCGGATGCAAGTCCAGGCCCAAGGCCCGGTCCAGC
+
EEEEEEEEEEEEEEEEEEEEEAAEAAEEEEEAEEEEEEEEEEEEEEEEEEA/AEAEEEE<EEEEEEEEEEEE/E<EEEEEEEEEEEEEEEEEE

Read 2 of a pair:
@NB501498:174:HWCNMBGXC:1:11101:5322:1045:CTTGGNT+TTCGAGC 2:N:0:TCCGGAGA+AGGATAGG
AGAAAAGGCTCCAGGGAAGAGCTGGCTCCTACCTGTGCTGGACCGGGCCTTGGGCCTGGACTTGCATCCGAAGCCGGTGGGGGAGCCGCCC
+
EEEEEEEEEEEEEEEEEEE/EEEEEEEEEA/EEEEEEEEEEEEEEEEEEE/EEEEEEEEEEAEEEAEEEE/AEEEEEAEEEEEEEEEEEEE
```

- >= (tens of) millions of read-pairs per sample



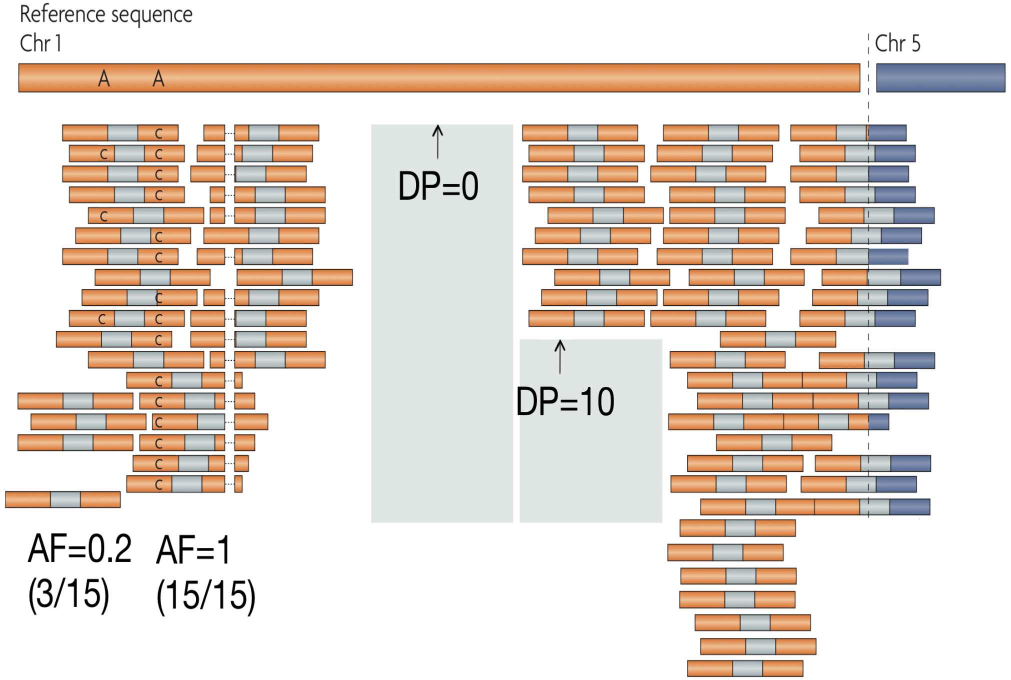
Read mapping/alignment and small variants

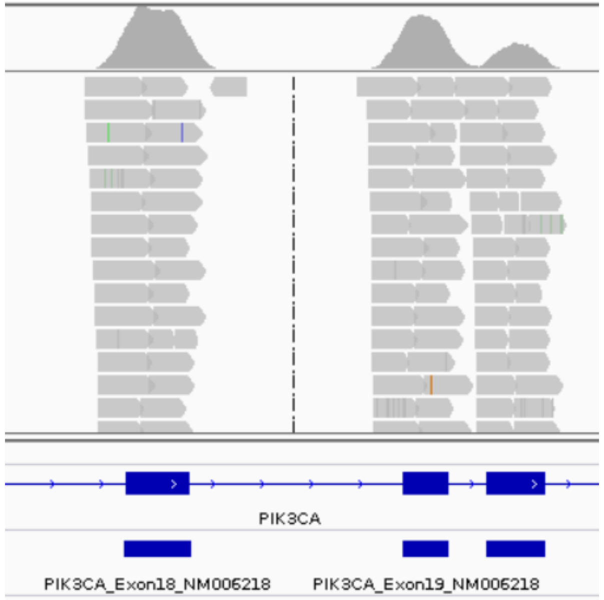


- Single Nucleotide Variants (SNVs) and Multi-Nucleotide Variants (MNVs)
  - Changes of a single codon or multiple close-by codons
    - STOP codons, functional protein domains
  - Splice-site and binding-site changes
- Insertions/Deletions (INDELs)
  - Possible "(open reading) frame-shift" changes

(edited) Image from: Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685–696 (2010).

Basic small variant metrics

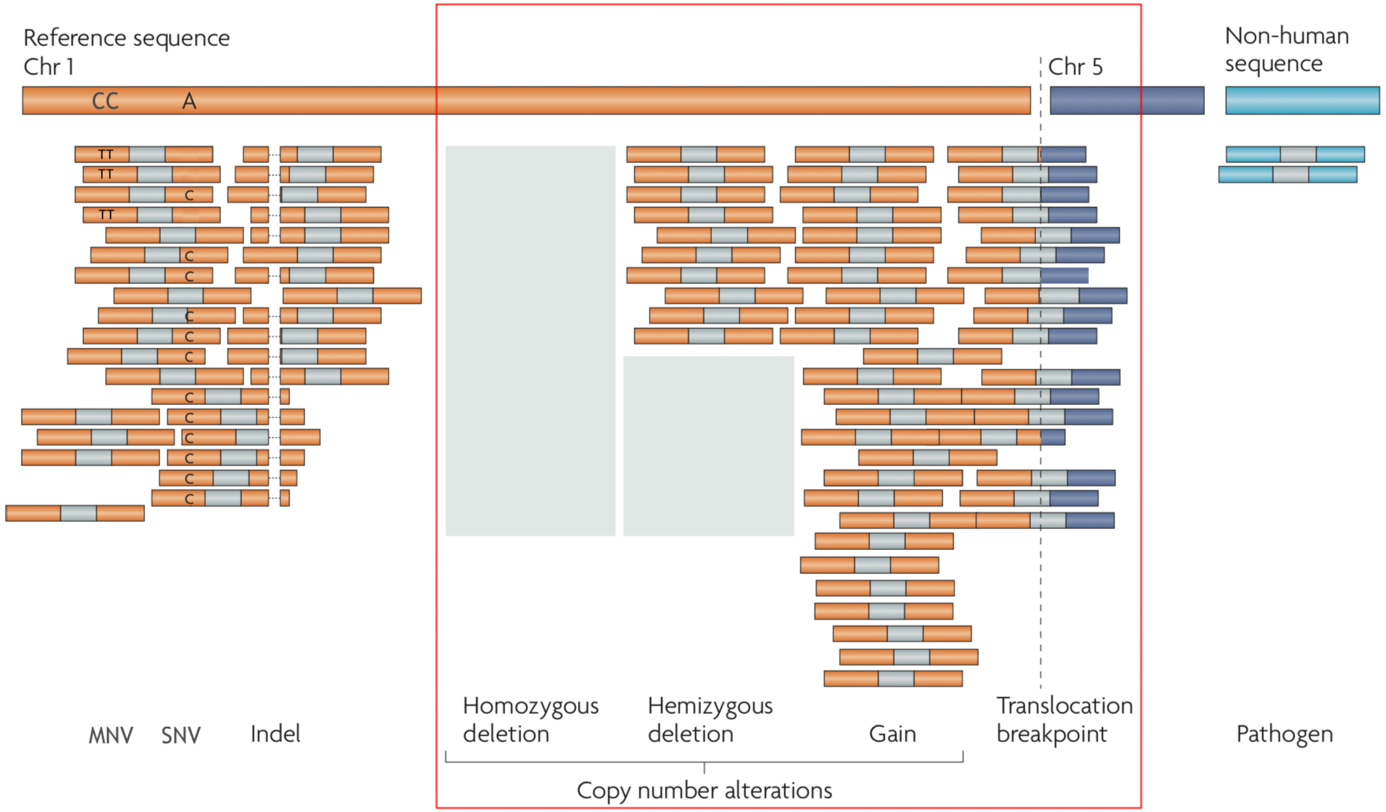




- Read depth (DP)
  - Number of reads at variant site
  - Coverage: the corresponding sample-wide metric: ~mean DP/ (e.g., 300X)
  - Can be highly variable (tumor, targeted assays)
  - Number of unique molecules from the input sample
- Variant allelic fraction (VAF, VF)
  - Fraction of reads (at the variant site) that carry/support the variant

(edited) Image from: Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685–696 (2010).

Copy number variants, genomic rearrangements

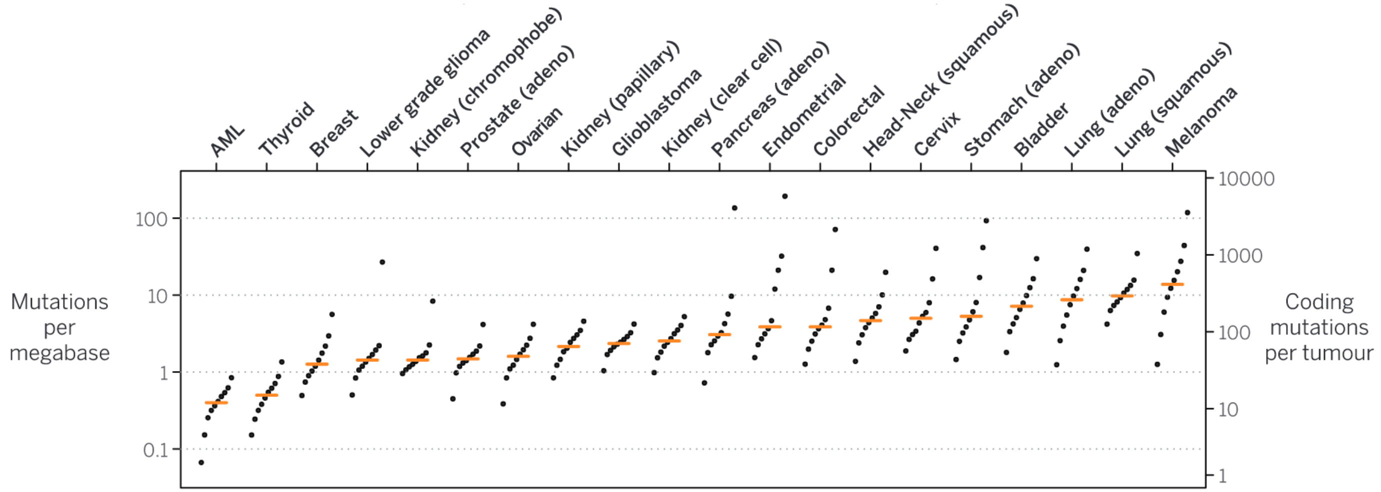


- Copy Number Variants (CNVs)
  - Protein expression levels
  - Loss of heterozygosity (LOH)
- Genomic rearrangements
  - E.g., translocations
    - Protein fusions
  - “Breakpoints”
    - Often difficult to detect with targeted assays

(edited) Image from: Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685–696 (2010).

Cancer-relevant variant-based metrics (I)

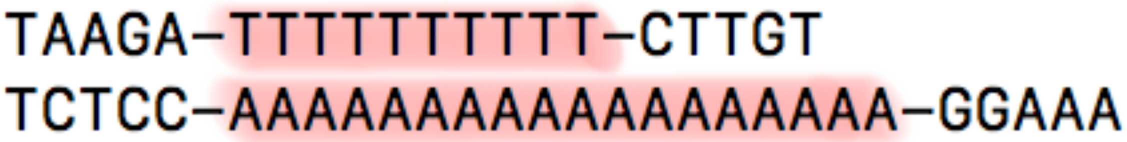
- Tumor Mutational Burden (TMB)
  - Number of (small) somatic variants seen per million base-pairs of the assay target
    - Normalized mutation rate – useful for comparing mutation loads between different assays
    - e.g., 10 variants identified in an assay covering 2.5 mega-bases gives TMB of 4
  - Variability between cancer types
  - The TMB level can be relevant for selection of treatment



(cropped) Image from: Martincorena, I, & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. Science (American Association for the Advancement of Science), 349(6255), 1483–1489.

Cancer-relevant variant-based metrics (II)

- Micro-Satellite Instability (MSI)
    - “Microsatellites” (short tandem repeats) – repeats of short DNA sequences spread across the genome (as short as mono- and di-nucleotides)



```
TAAGA–TTTTTTTTTT–CTTGT
TCTCC–AAAAAAAAAAAAAAAAAA–GGAAA
```

-
    - “MS instability” refers to high mutation rates within MS sites
        - This points to improperly functioning DNA mismatch repair
    - Common for certain tumor types (most notably colorectal cancer)
    - Relevant for selection of treatment

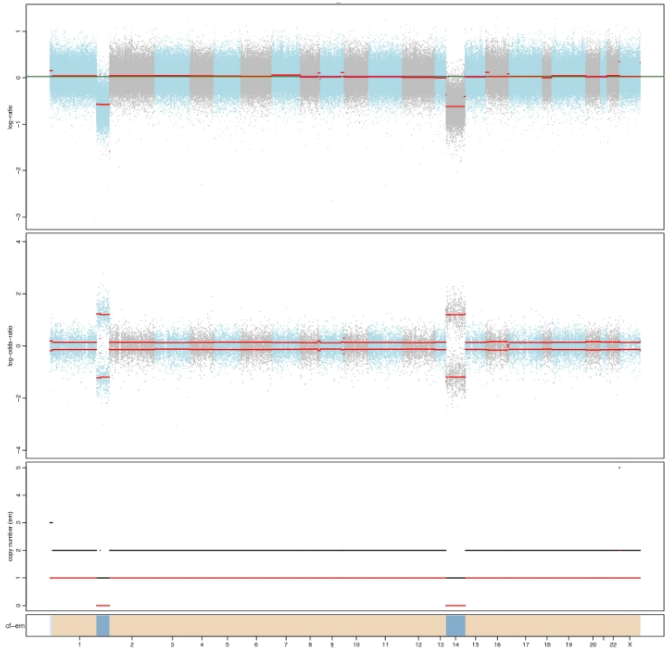
Variant calling, classification and annotation

- Variant calling
  - Determining whether a variant is present in a given sample
- Variant classification
  - Determining a found variant’s origin (artifact, germline, somatic)
- Variant annotation
  - Using information from external sources/databases in order to determine a variant’s rarity in a certain population, function/impact, significance for a specific disease, etc.

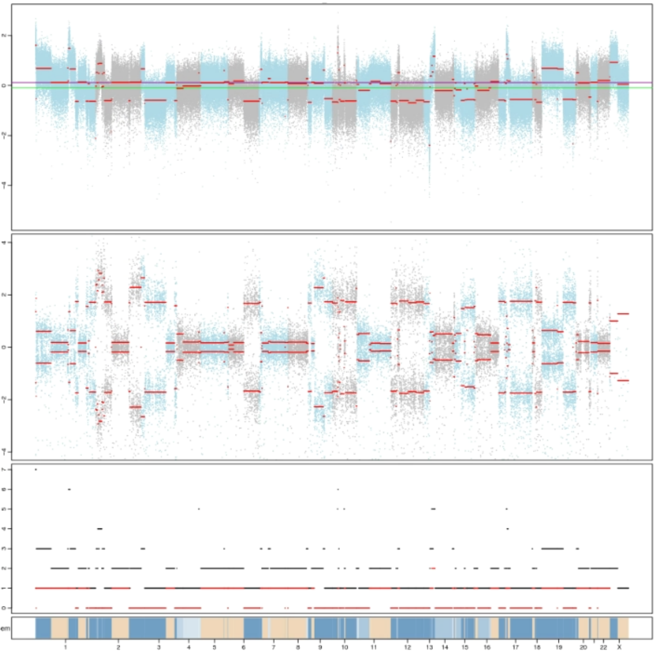
- Availability of “matched controls” (healthy tissue from the same individual as the assayed tumor tissue)
  - One person’s somatic variant can be another person’s germline variant
  - In absence of a matched control, annotation might be necessary for classification
- Private germline variants
  - Germline variants not documented in variant databases
  - Can be easily misclassified as somatic if matched control is not available

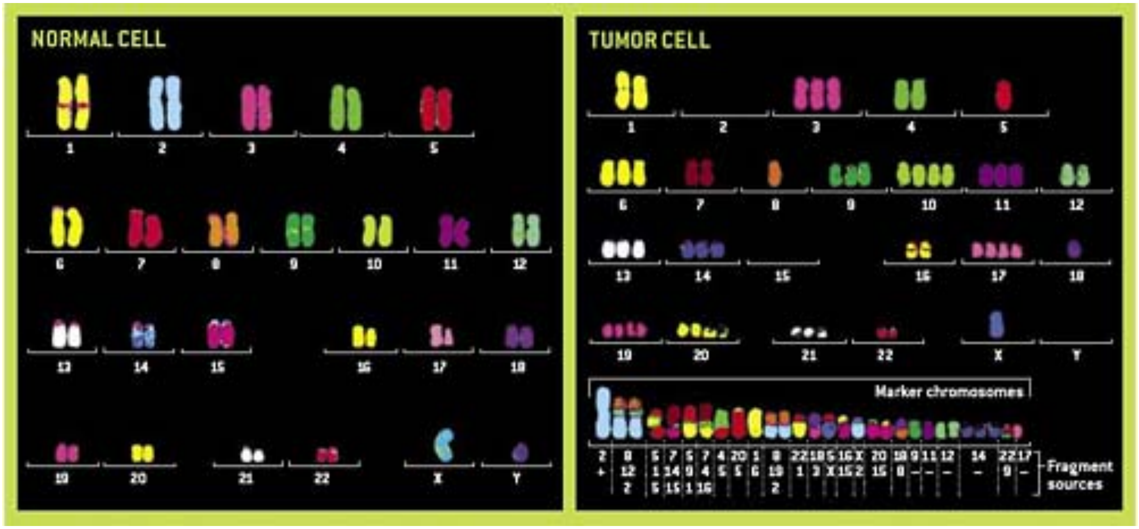
Factors that complicate somatic variant calling (I)

- Aneuploidy – copy number gains and losses



Lost part of chr. 1 and whole chr. 14





Right-hand (karyotyping) image from: https://www.berkeley.edu/news/media/releases/2007/06/26_drugresistance.shtml

Factors that complicate somatic variant calling (II)

- Tumor content/sample purity (fraction of tumor cells in the physical sample)
  - Tumor type, sample collection
- Tumor heterogeneity
- Original description text for the figure to the right: *"The gray circle represents a normal cell, and the central dot depicts the initiating somatic mutation that drives the founder clone in the tumor. The different colored circles represent subclones that have accumulated successive mutations. Note that in the primary tumor, several subclones coexist, and although some expand, others remain dormant or become extinct. Metastases can originate from either a major clone in the primary tumor (metastasis 1), or from minor clones (metastasis 2). Metastases can also undergo clonal evolution (as shown in metastasis 1)."*

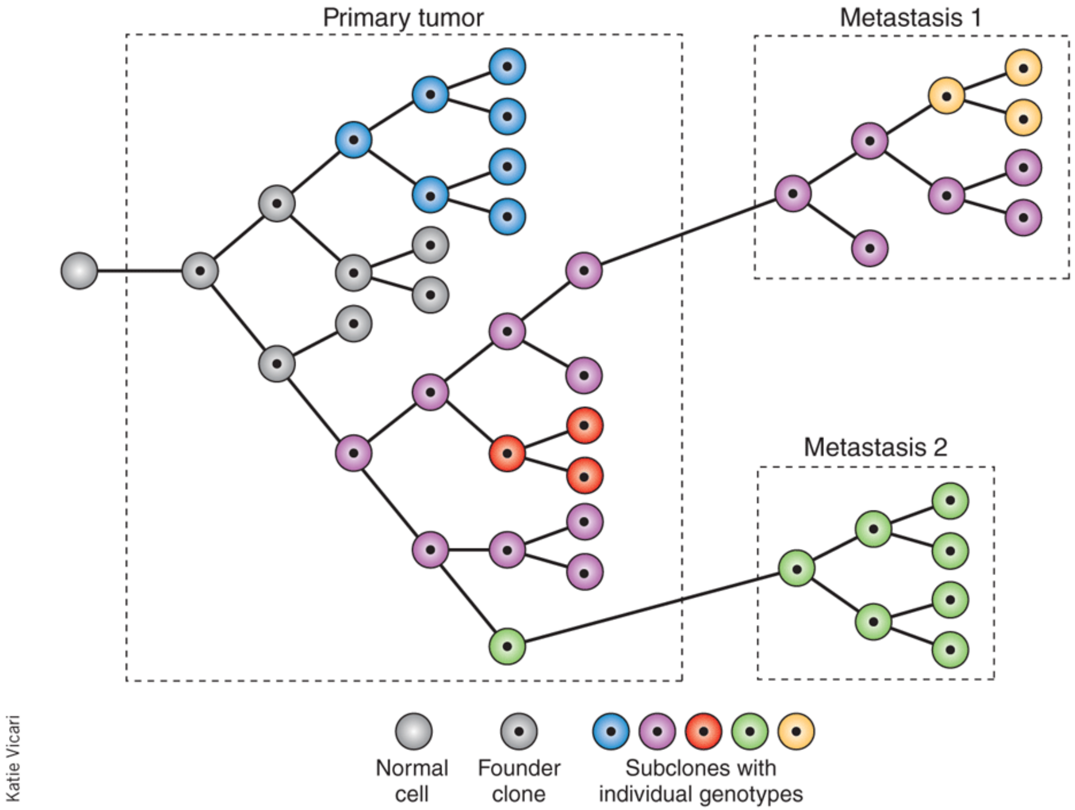


Image and quote from: Caldas, C. Cancer sequencing unravels clonal evolution. *Nat Biotechnol* **30,** 408–410 (2012).

Wide allelic fraction ranges

- In case of germline variants, allelic fractions are usually either 0.5 (heterogeneous variants) or 1.0 (homogeneous variants)
  - Except for some parts of the sex chromosomes in case of males
- Anywhere between 0.00 and 1.00 in case of somatic variants
  - Tumor content/sample purity, tumor heterogeneity/clonality, copy number changes
  - Not all tumor cells are included in the sequencing

- Please note: “variant (allelic) fraction” is not the same as “variant frequency”
  - **Variant fraction**: fraction of reads at a variant site **in a given sample** that carry the variant
  - **Variant frequency**: fraction of chromosomes **in a given population** that carry the variant

Artifact variant calls

- Sequencing and alignment errors
  - Not sample specific, but often technology- and software-specific – a matched normal sample processed and sequenced in the same manner as the tumor can greatly help mitigate these
    - Sequencing both matched samples to a similar depth is optimal
- Oxidation (C>A, G>T)/deamination (C>T, G>A)
  - Chemical changes of specific nucleotides, which appear as nucleotide substitutions (usually across the genome, but with low allelic fractions)
    - Only one strand affected – can be quantified by PicardTools/GATK tools
  - Caused *after* sample extraction (by inadequate storing or formalin fixation)
- Mismatched samples and/or sample contamination
  - Human error, sequencing instrument issues (“index jumping”)
  - Typically leads to known germline variants being classified as somatic variants

General advice for somatic variant calling

- Use software designed specifically for somatic variant calling
  - Germline variant callers will likely not take into account challenges specific to somatic variant calling
- Many new variant callers are published every year
  - Different approaches (sometimes tailored for specific types of data)
  - Sometimes dramatically different results (especially for variants other than SNVs)
    - Benchmark papers
  - Using multiple variant callers (and a visual review) can be a good idea
- When possible, having a matched normal is best
  - Generated with the same technology, ideally sequenced to similar depth
  - A panel of normals ("PON") generated with the same technology can be an alternative
    - More work, less reliable (a panel cannot account for private germline variants)