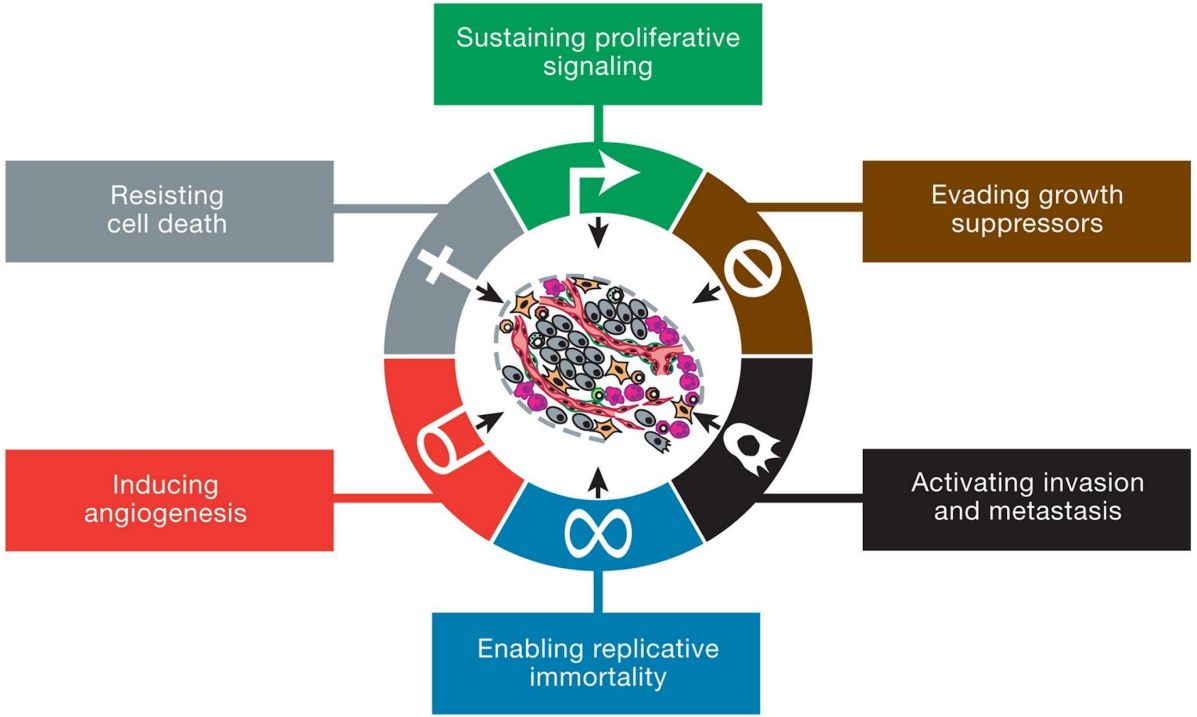
SevenBridges

Bioinformatics of cancer

11.05.2023

Boris Majić

What is cancer?

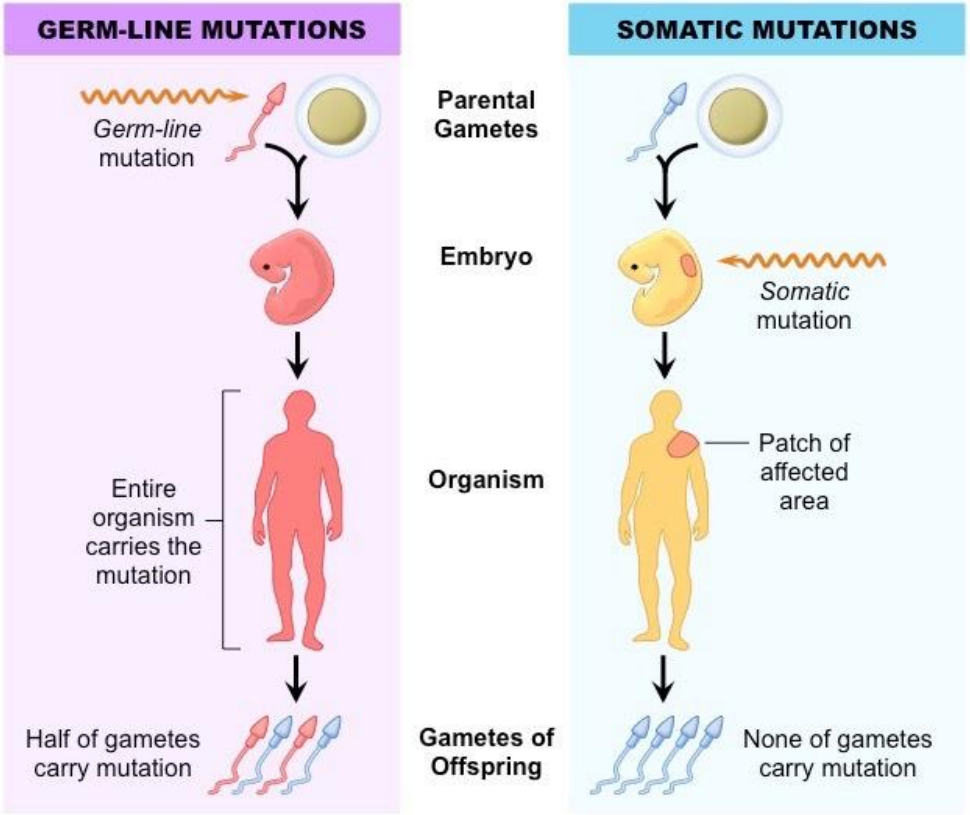


Definition(s)

Cancer Fact sheet WHO (2014):

Cancer is a disease caused by an uncontrolled division of abnormal cells in a part of the body. These cells can invade nearby tissues and spread to other parts of the body through a process called metastasis

Variants (mutations)



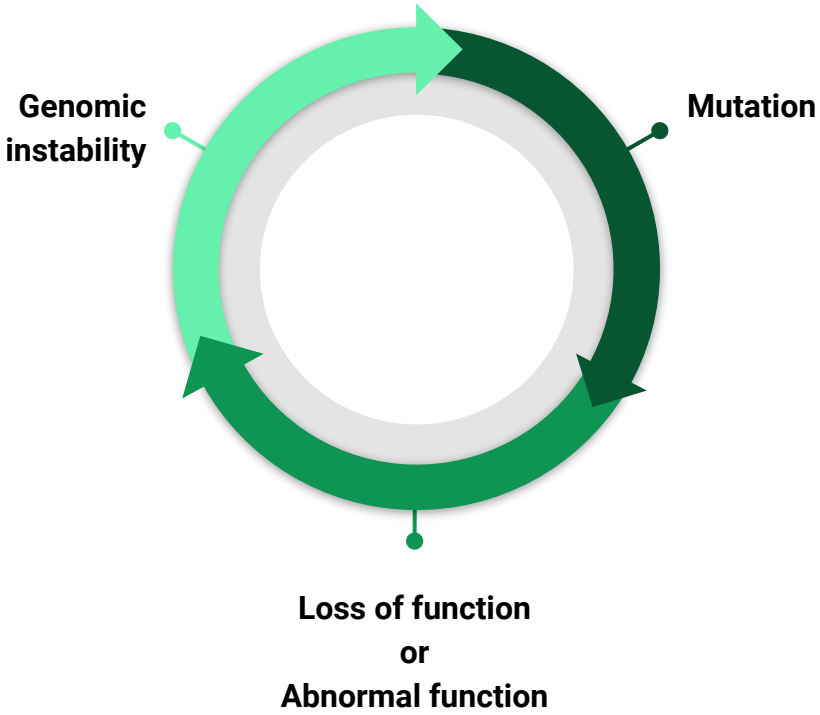
Cell immortality

Cells are programmed to die in specific circumstances:

- After a number of cell divisions
- When exposed to stress
- When sustaining damage in structures or exhibiting strange behaviour - **apoptosis**

Genes responsible for apoptosis
tumor-suppressors

Mutation consequences



Tumors - benign or malignant

Immortality ≠ cancer

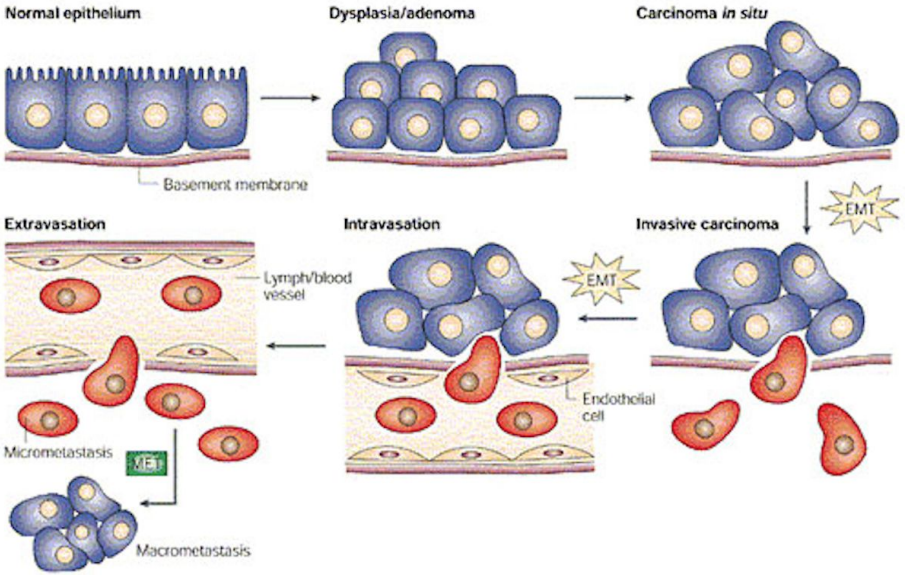
Benign tumor:

- Uncontrolled growth
- Immortal cells
- Some abnormal functionality
- Energy intensive
- Compressing neighboring tissues

Maligni tumori: ?

Malign tumors

- Invasivity and metastasis characterise malign tumors
- Use of blood and lymph flows to migrate to other tissues



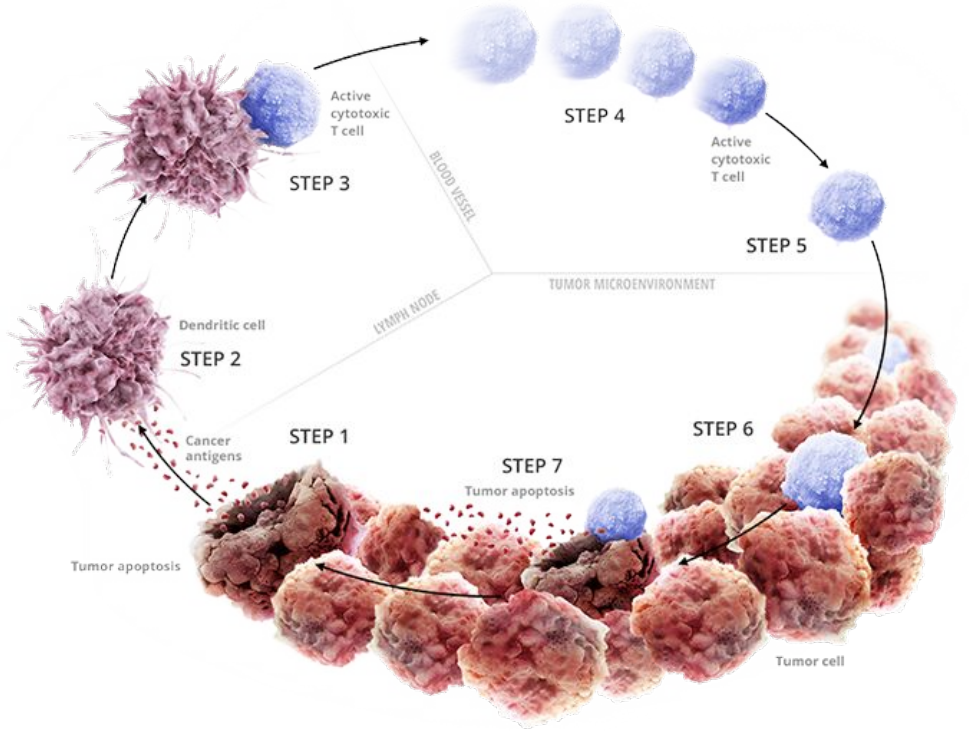
Immune response to tumors

Antigens - molecules activating the immune response

Mutations cause changes in the cell surface – **neoantigens**.

More mutations - stronger immune response

More mutations - more aggressive tumor



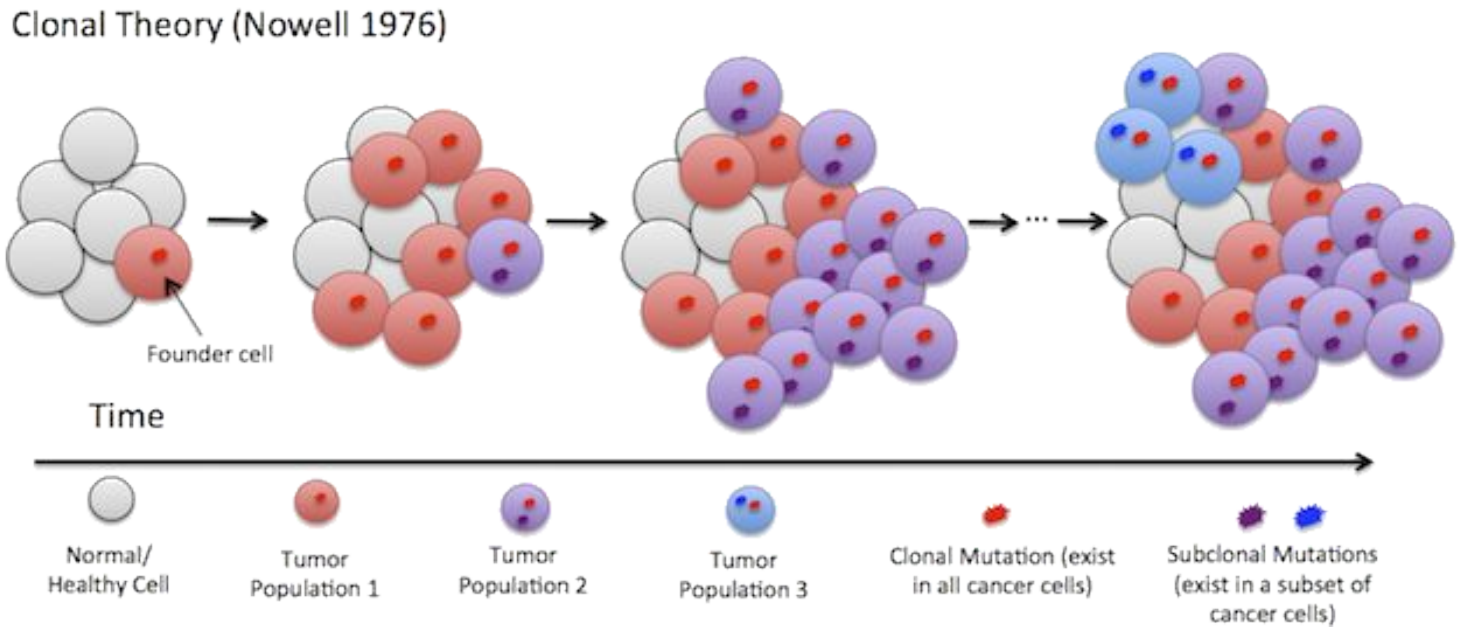
Characteristics of cancer

Cancer cells are defined by:

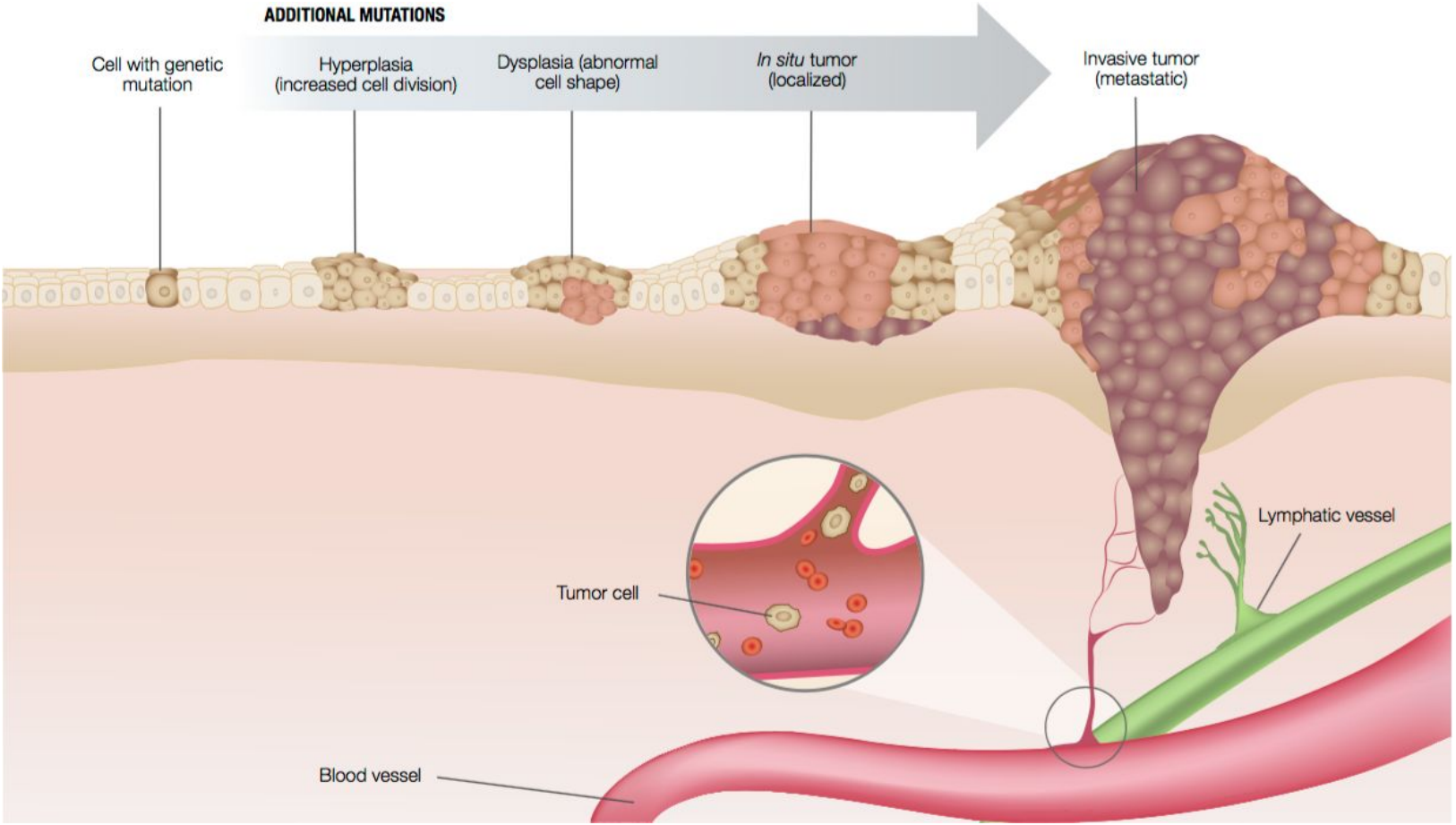
- Immortal
- Always dividing
- Loss of function
- Energy demanding causing angiogenesis

Malign - invading and metastasis
Avoiding cellular defense mechanisms

Genomic instability - tumor clones



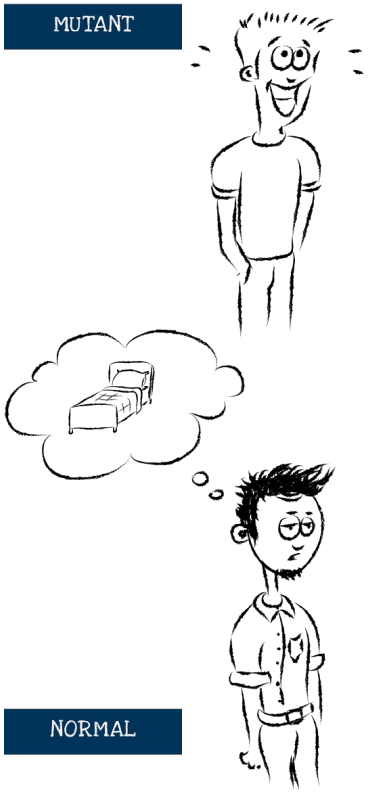
Evolution of tumor


ADDITIONAL MUTATIONS
Cell with genetic mutation
Hyperplasia (increased cell division)
Dysplasia (abnormal cell shape)
In situ tumor (localized)
Invasive tumor (metastatic)
Lymphatic vessel
Tumor cell
Blood vessel

Variant Calling

IT'S LIKE A PUZZLE...

SLEEPLESS MUTANT



chromosome 12 gene BHLHE41

reference genome

...GCGGCTGCCGCCCGTTCCCGCTGCTATACCCC...

pieces of DNA produced by a sequencer

GCTGCCGCCCGGTTC

GCCGCCCGGTTCCCG

GCCCGGTTCCCGCTG

CGGTTCCCGCTGCTA

chromosome 12 gene BHLHE41

reference genome

...GCGGCTGCCGCCCGTTCCCGCTGCTATACCCC...

pieces of DNA produced by a sequencer

GCTGCCGCCCGTTC

naah... just a sequencing error...

GCCGCCCGGTTCCCG

GCCCGTTCCCGCTG

CGTTCCCGCTGCTA

Variant calling concept

- Checking if all reads at a tested position support the reference
- Reference supporting reads – REF
- Variant (Alternative) supporting reads – ALT
- Depth/Coverage = REF + ALT (number of reads covering that position)
- Variant Allele Frequency = ALT / (REF + ALT)
  - Coverage 30 – 20 REF reads, 10 ALT reads
    - VAF = 0,33 or 33%
- Genotypes
  - 0/0 = Both alleles match the reference (homozygous)
  - 0/1 = One allele matches reference and one does not (heterozygous)
  - 1/1 = Both alleles do not match reference (homozygous)
  - 1/2 = One allele contains one variant and the other another one (heterozygous)

Variant calling concept

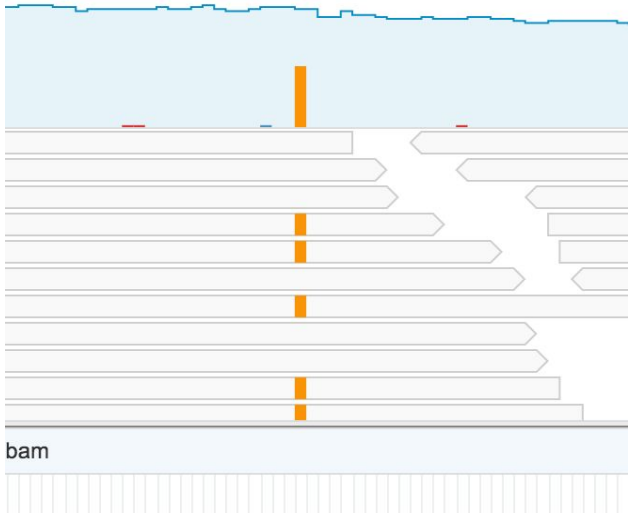
- **Multisample and joint calling**
  - Multisample does not take other samples into consideration when deciding if there is a variant at the position in the currently tested sample
  - Joint calling takes into consideration variants at the position in question in other samples than the one currently being tested
- **Somatic calling – a type of multisample calling**
  - Two samples are compared for differences
  - Parameters are adjusted to accommodate cancer properties

Multisample VCF example

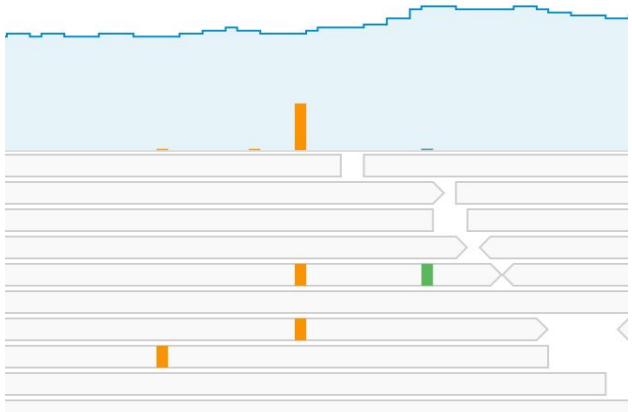
CHR	POS	REF	ALT	FORMAT	NA12877	NA12878
1	14125	T	A	GT, VAF	0/1, 0.43	0/0, 0.1
5	14125	A	G	GT, VAF	0/0, 0.0	0/1, 0.4

Germline vs Somatic calling

- Germline Variant Calling
  - Difference of a tested sample to a reference genome
  - Each sample compared to a reference genome (single sample or joint calling)
- Somatic Variant Calling
  - Difference of both Normal tissue sample (Normal) and Tumor tissue sample (Tumor) to a reference genome
  - Finding mutations (usually by applying statistical methods) that occur in Tumor but are not present in Normal



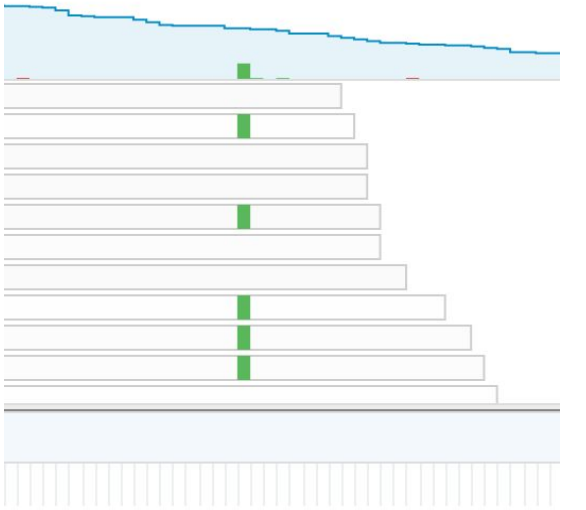
GERMLINE MUTATION



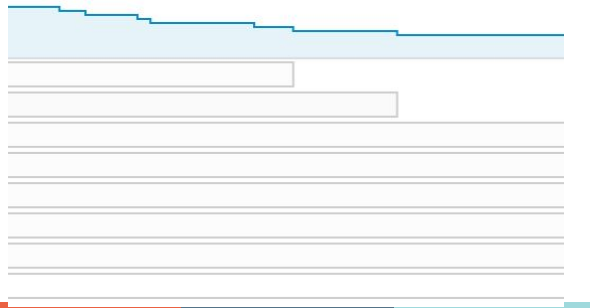
TUMOR SAMPLE

Somatic Calling

NORMAL SAMPLE



SOMATIC MUTATION



Variant calling results

- The result of Variant Calling is a file in VCF format, which contains mutations
- A plain text file format for storing variant data
- A number of line starting with ## -the header
- Main header line:
  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
- This is followed by the actual variant data, one entry per line
  22 10001 . A C 40 PASS DP=14 GT 0/1
- More than one sample can be in one line
- For details: VCF specification

Variant calling results

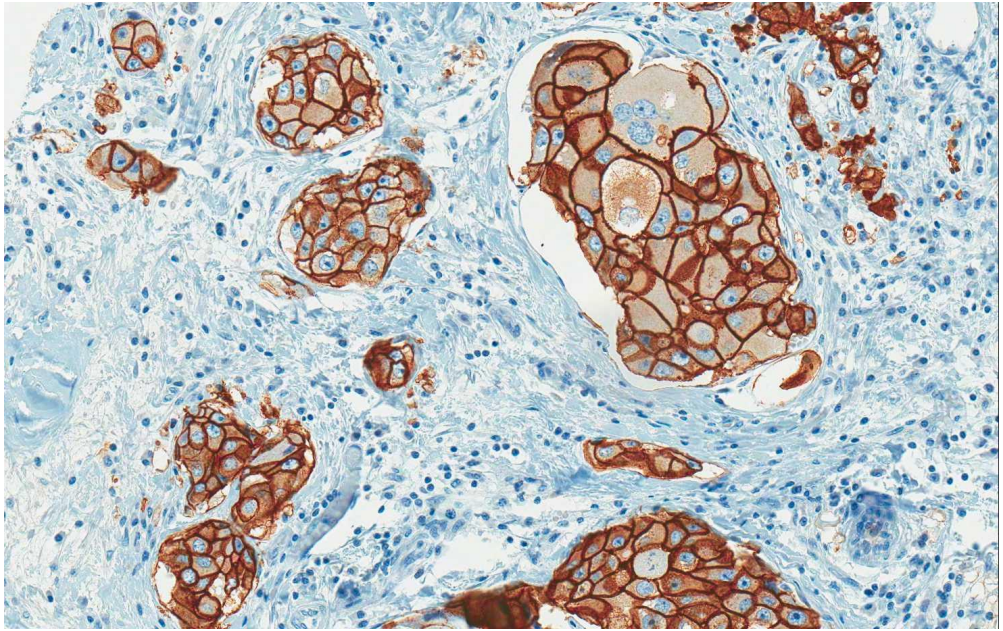
- Example of VCF format
- Each row represents one mutation

CHR	POS	REF	ALT	FORMAT	NA12878
1	14300	A	G	GT, VAF	0/1, 0.4
2	15367	A	C	GT, VAF	1/1, 0.9
3	25612	C	G,A	GT, VAF	1/2, ?
5	5632	TA	T	GT, VAF	0/1, 0.5
7	7824	T	TA	GT, VAF	1/1, 0.8

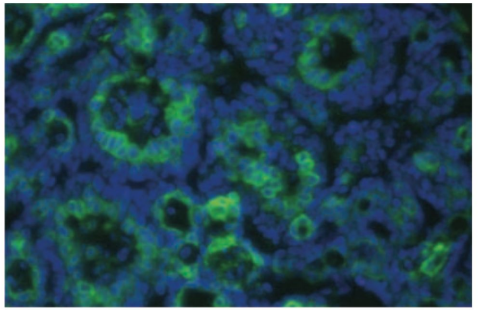
Somatic Calling Goals

- Finding variations that occur in tumor tissue but do not occur in normal tissue
  - Possible cause of cancer occurrence
  - Variations frequently occurring in the same tumor type (early diagnosis)
  - Possible targets for cancer treatment
- Engineering vaccines and medications that interact with the tumor tissue but do not damage the normal tissue

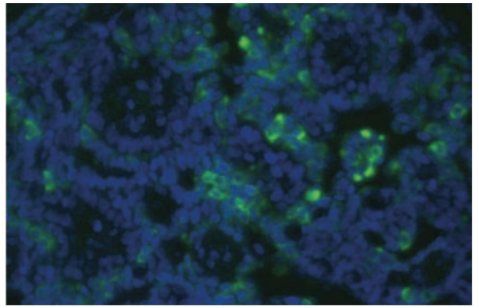
Tumor purity and heterogeneity



Tumor (brown) interlaced with normal (blue)



SMA



Keratin 8

Different markers shown by different cells

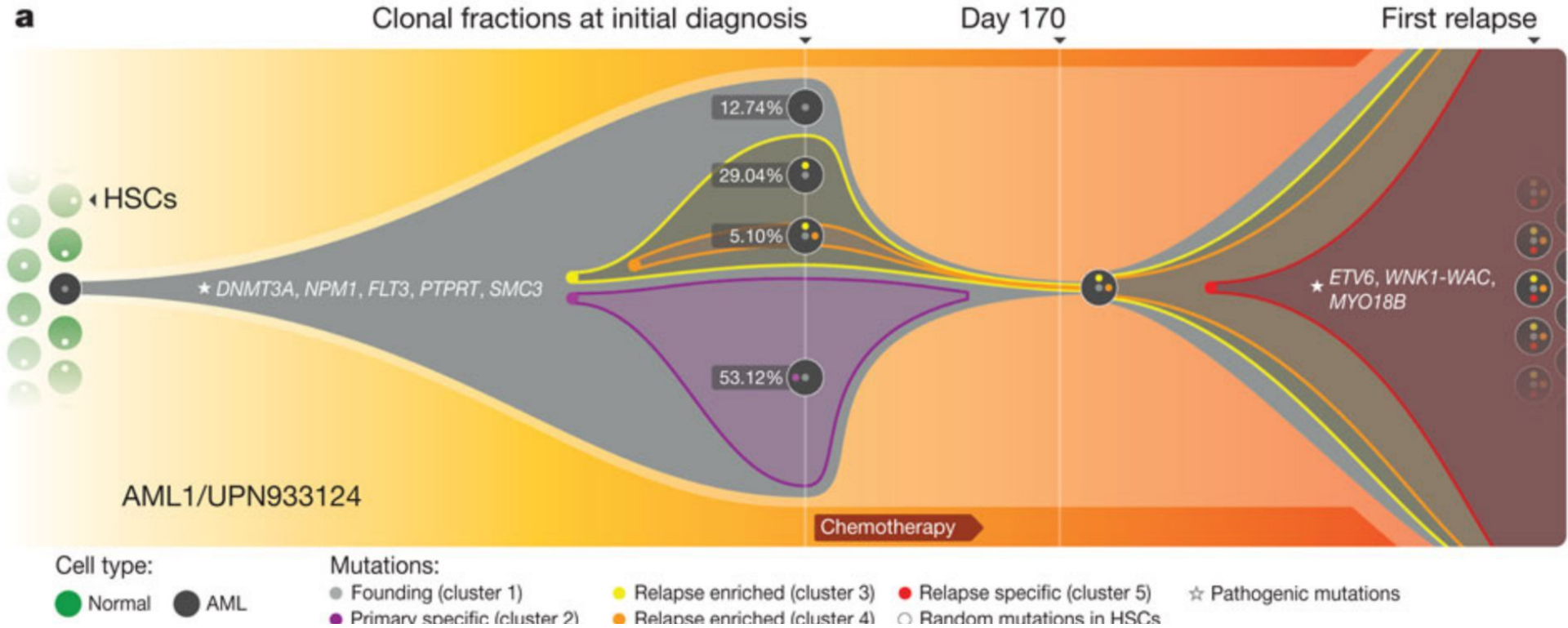
Somatic Variant Calling Challenges

- Tumor heterogeneity
  - Different biopsy locations yield different variations and/or different variant allele frequencies
- Tumor purity
  - Hard to isolate tumor cells from healthy cells when preparing for sequencing
- Low variant allele frequency (directly related to above mentioned) in combination with sequencing errors
  - Hard to conclude that something is a somatic variant
- No datasets for benchmarking (Truth VCF)
  - Hard to benchmark available tools and analysis methods

Somatic Variant Calling

- A vast number of callers
- There is no “star” among somatic variant callers, they perform based on the issue at hand
- Types of Somatic Variant callers
  - Position based e.g. Varscan
  - Context aware (perform realignment / reassembly) e.g. Strelka
- Different callers have different pros and cons
  - Quality of calls
  - Runtime
  - Memory consumption
  - Possibility to adjust parameters
  - Type of analysis they are suitable for (WGS, WES)

Fishplot


a
Clonal fractions at initial diagnosis
Day 170
First relapse
HSCs
12.74%
29.04%
5.10%
53.12%
★ DNMT3A, NPM1, FLT3, PTPRT, SMC3
★ ETV6, WNK1-WAC, MYO18B
AML1/UPN933124
Chemotherapy
Cell type:
Normal
AML
Mutations:
Founding (cluster 1)
Primary specific (cluster 2)
Relapse enriched (cluster 3)
Relapse enriched (cluster 4)
Relapse specific (cluster 5)
Random mutations in HSCs
☆ Pathogenic mutations

Why sequence tumors

- Find neoantigens - devise personalized therapy (i.e. mRNA vaccines)
- Detect characteristics of tumor tissue - tumor classification - mutational signatures
  - adjust therapy
  - better prognosis
- Track therapy efficiency (especially for liquid tumors)
- Look for tumor-specific transport proteins
- Early detection - ctDNA sequencing

Other applications of bioinformatics

- Monitor epigenetic characteristics of tumor tissue - personalized medicine

- Computer vision - Biomedical image processing - tumor tracking, location and size determination - help with surgeries
  - MRI, CT scans, RT

- Association testing - Statistical testing

Conclusions

- Personalized cancer therapy is coming
- Understanding the genomic cancer landscape is becoming practically useful
- Somatic analysis is not a direct extension of germline analysis
- Many concepts are purely abstract
- Computational processing limited by sequencing technology

Ispit

milan.kovacevic@sbgenomics.com

Sledeće nedelje - priprema za ispit sa Milanom

Ispit - predrok - poslednja nedelja maja