Cancer Genomics

Djordje Klisic

Bioinformatics Project Manager

SevenBridges

What is cancer?

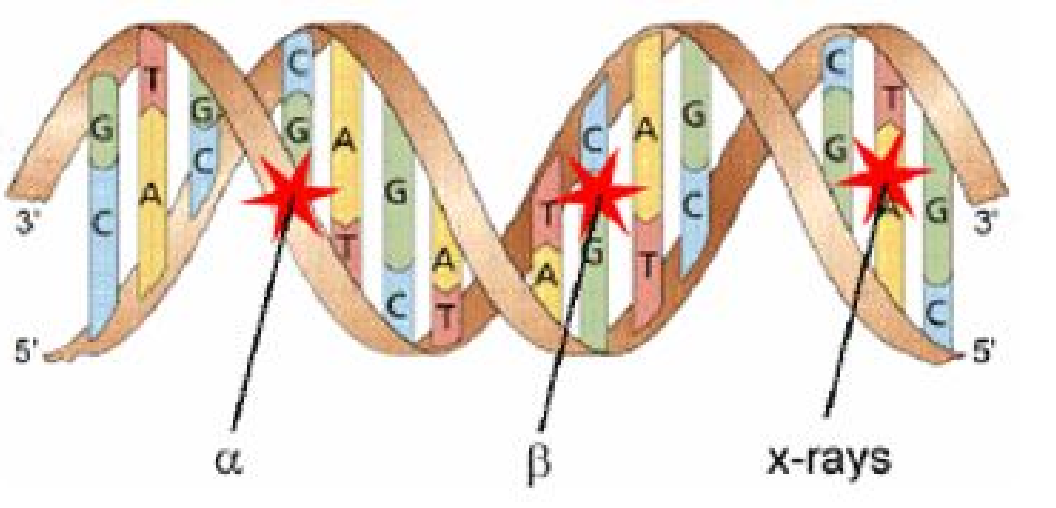
- A group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body (definition by NIH)

What causes cancer?

- EM radiation
- Chemical agents
- Free radicals
- Genetic factors
- Infections (viruses)

Radiation

- Direct damage



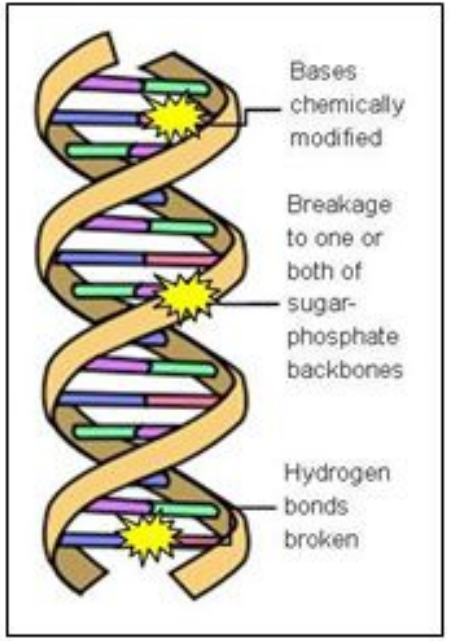
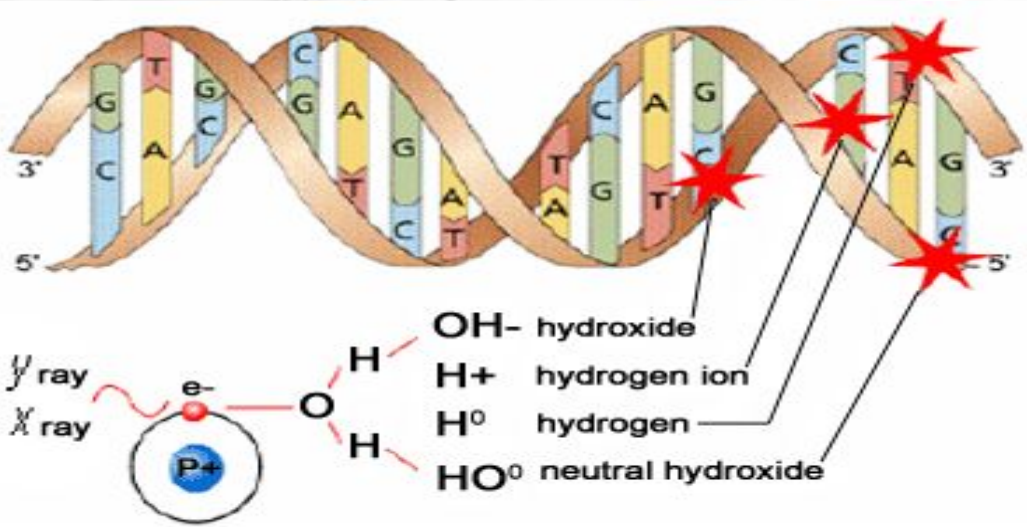


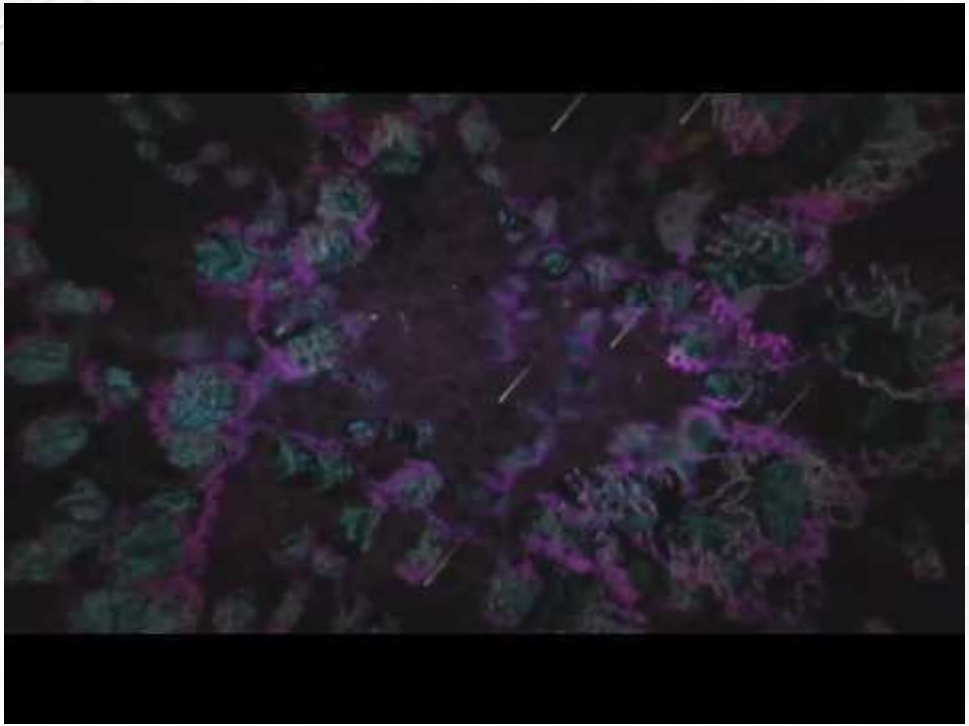
Figure 1: Potential direct effects of ionizingradiation on DNA molecules.

Radiation

- Indirect damage

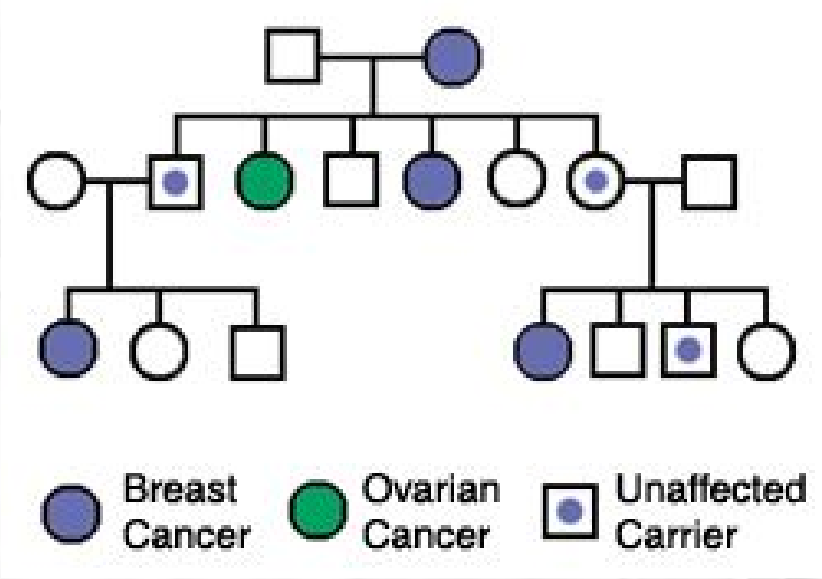


Chemical Agents and Free Radicals

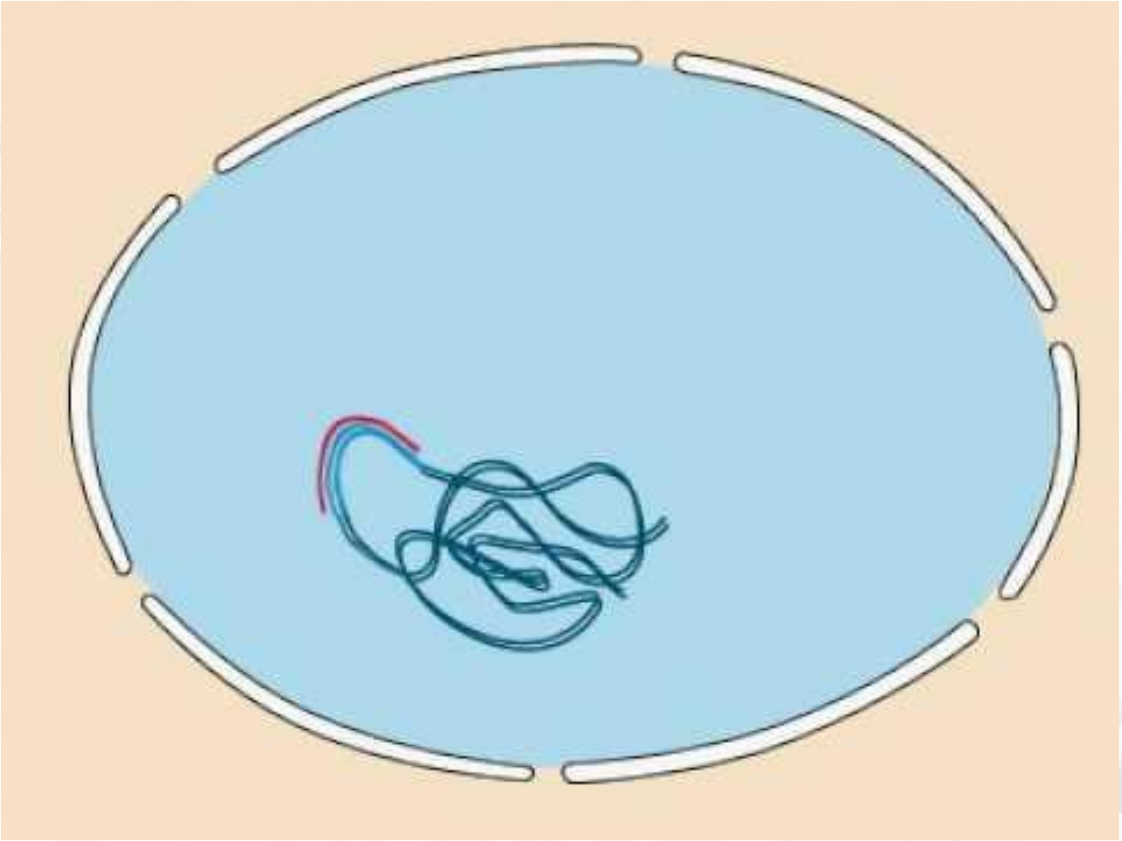


Genetic factors

- A typical pedigree from a family with a mutation in the BRCA1 gene
- Fathers can be carriers and pass the mutation onto offspring
- Not all people who inherit the mutation develop the disease, thus patterns of transmission are not always obvious



Viruses



"Drivers" of Cancer

Cancer is a genetic disease that is caused by changes to genes which control the way our cells function, especially how they grow and divide.

1. **Abnormal growth (proto-oncogenes)**

   Cellular growth mechanism is damaged and cell starts to multiply uncontrollably

2. **Damaged control mechanism (tumor suppressors)**

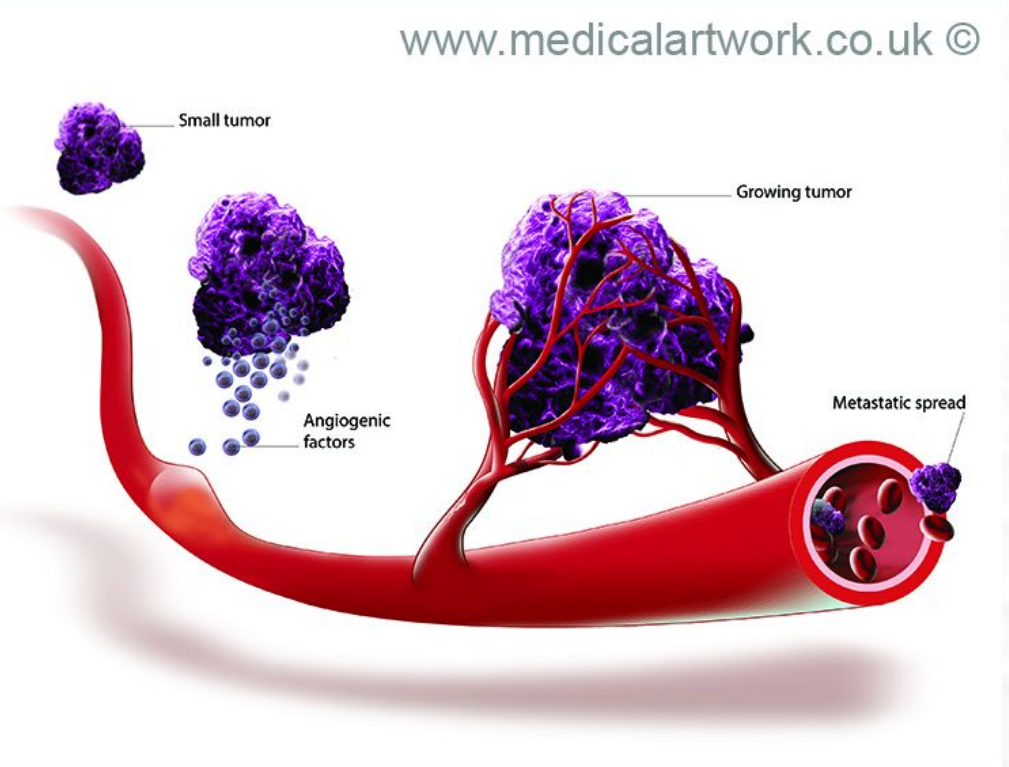
   Cells with certain alterations in tumor suppressor genes may divide in an uncontrolled manner (TP53 - Apoptosis)

3. **Damaged repair mechanism**

   Accumulated errors in this group of genes can lead to uncontrollable proliferation

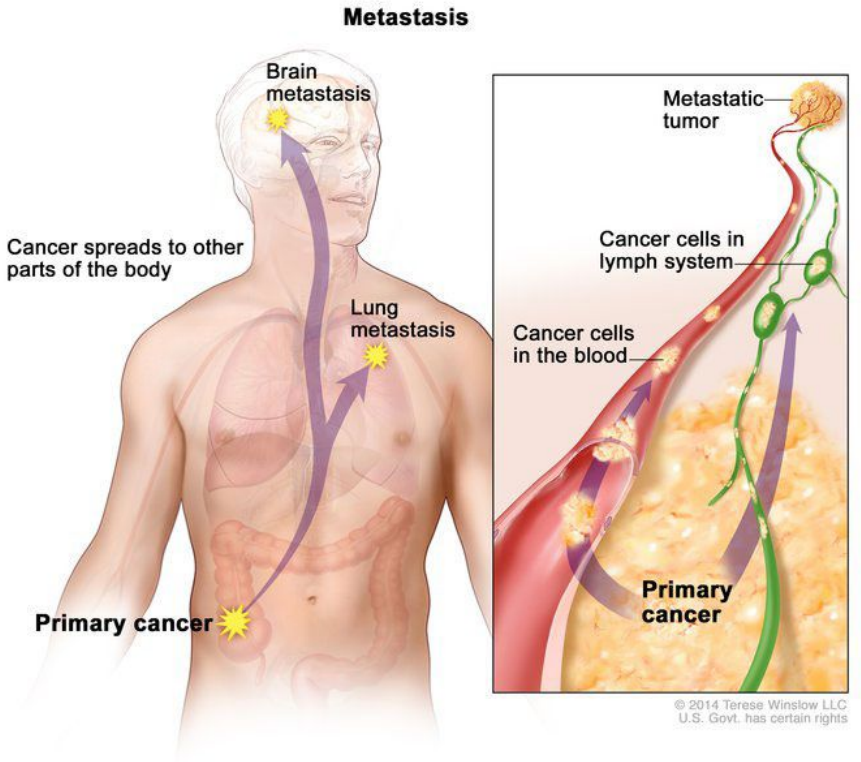
Angiogenesis

- Tumors greater than several hundred cell require active nutrient supply
- Big enough tumor cell group starts emitting angiogenic factors
- When active nutrient supply is present, tumor can continue its uncontrollable growth
- Tumor tissue contact with bloodstream can lead to metastasis



Metastasis

In metastasis, cancer cells break away from where they first formed (primary cancer), travel through the blood or lymph system, and form new tumors (metastatic tumors) in other parts of the body. The metastatic tumor is the same type of cancer as the primary tumor.



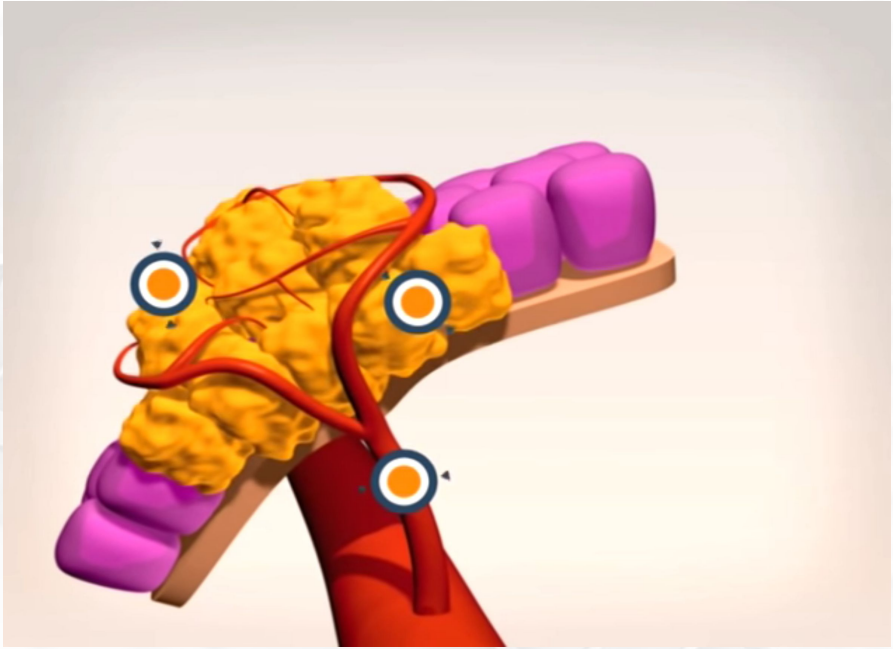
Cancer trivia (part 1):

- More than 575,000 people die of cancer, and more than 1.5 million people are diagnosed with cancer per year in the US.
- Cancer is considered to be one of the leading causes of morbidity and mortality worldwide.
- The financial costs of cancer in the US per year are an estimated $263.8 billion in medical costs and lost productivity.
- African Americans are more likely to die of cancer than people of any other race or ethnicity.
- It is believed that cancer risk can be reduced by avoiding tobacco, limiting alcohol intake, limiting UV ray exposure from the sun and tanning beds and maintaining a healthy diet, level of fitness and seeking regular medical care.

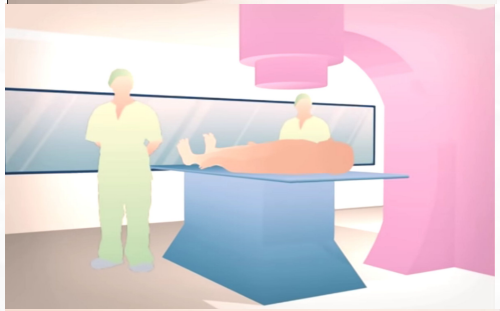
Cancer trivia (part 2):

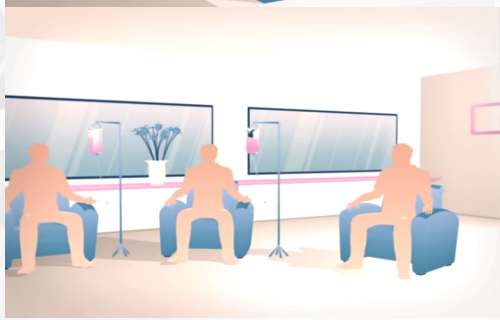
- Screening can locate cervical cancer, colorectal cancer and breast cancer at an early, treatable stage.
- Vaccines such as the human papillomavirus (HPV) vaccine assists in preventing some cervical, vaginal, vulvar, and oral cancers. A vaccine for hepatitis B can reduce liver cancer risk.
- According to the World Health Organization (WHO), the numbers of new cancer cases is expected to rise by about 70% over the next 20 years.
- The most common sites of cancer among men are lung, prostate, colon, rectum, stomach and liver.
- The most common sites of cancer among women are breast, colon, rectum, lung, cervix and stomach.

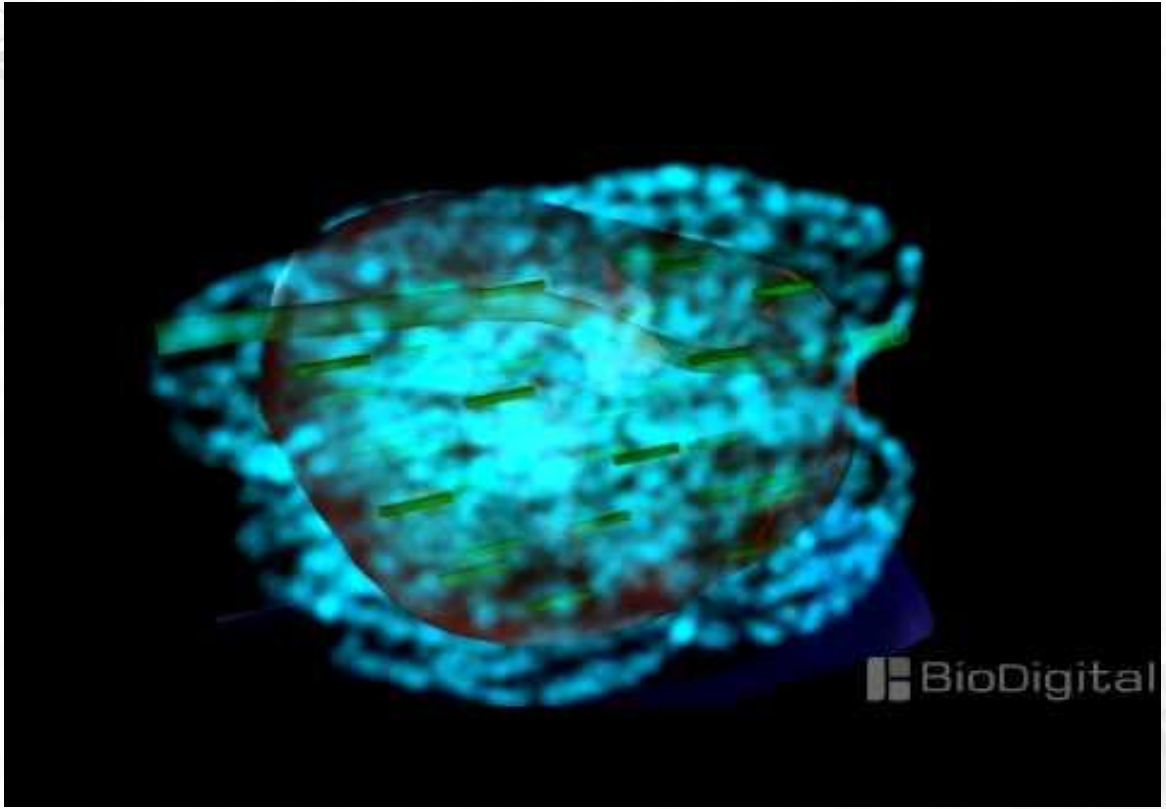
- Target cancer vascular system
- Target cell proliferation
- Provoke immune system



1. Surgery
2. Radiation therapy
3. Chemotherapy



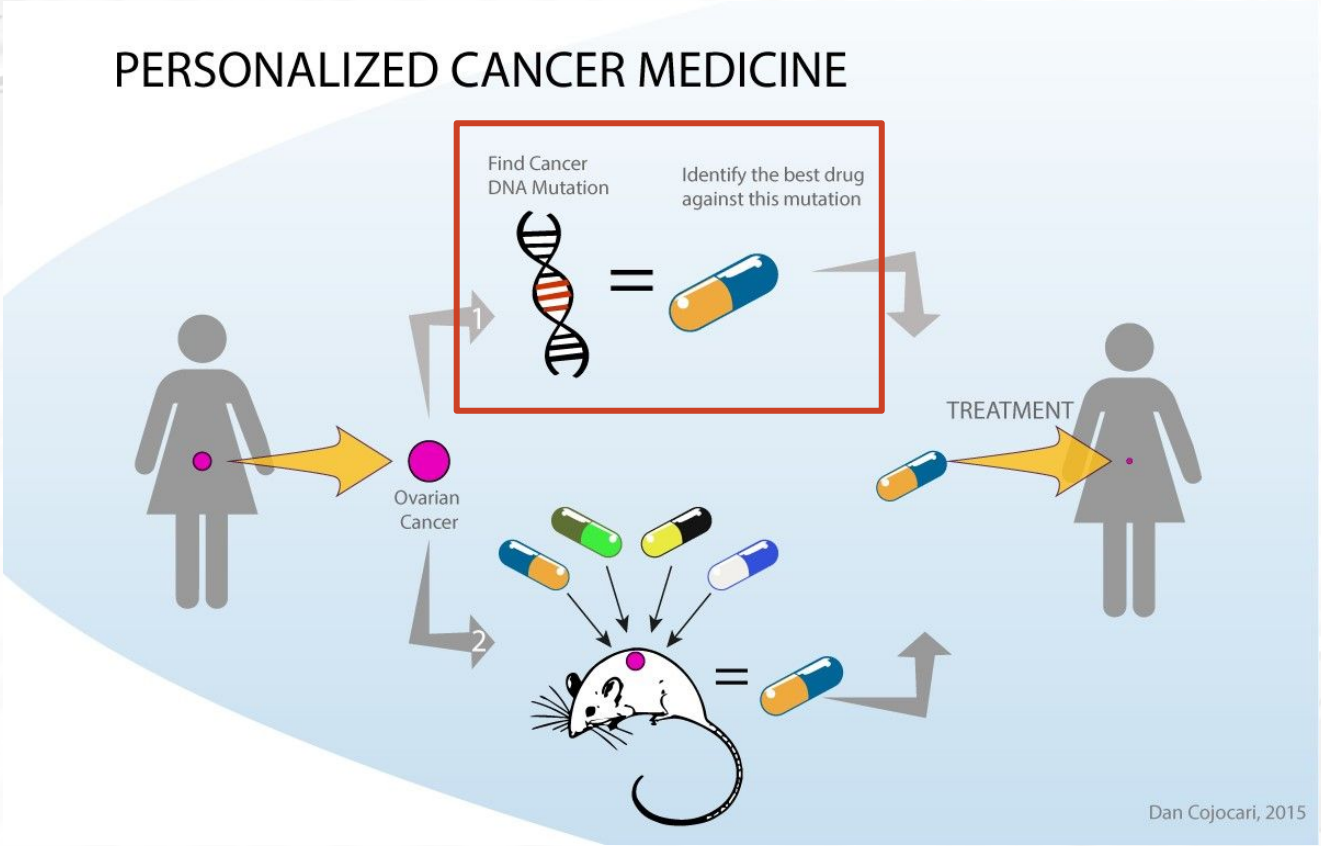



BioDigital

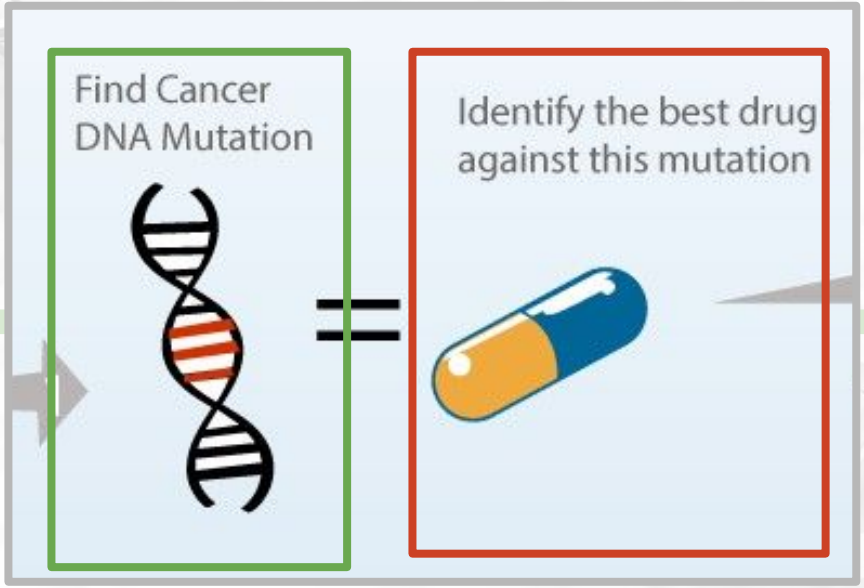
4. Photodynamic Therapy

5. Hyperthermia

6. Targeted Cancer Therapy


PERSONALIZED CANCER MEDICINE
Find Cancer DNA Mutation
Identify the best drug against this mutation
1
=
TREATMENT
Ovarian Cancer
2
=
Dan Cojocari, 2015

Sequencing → Alignment → Somatic Calling



Annotations → Prioritization

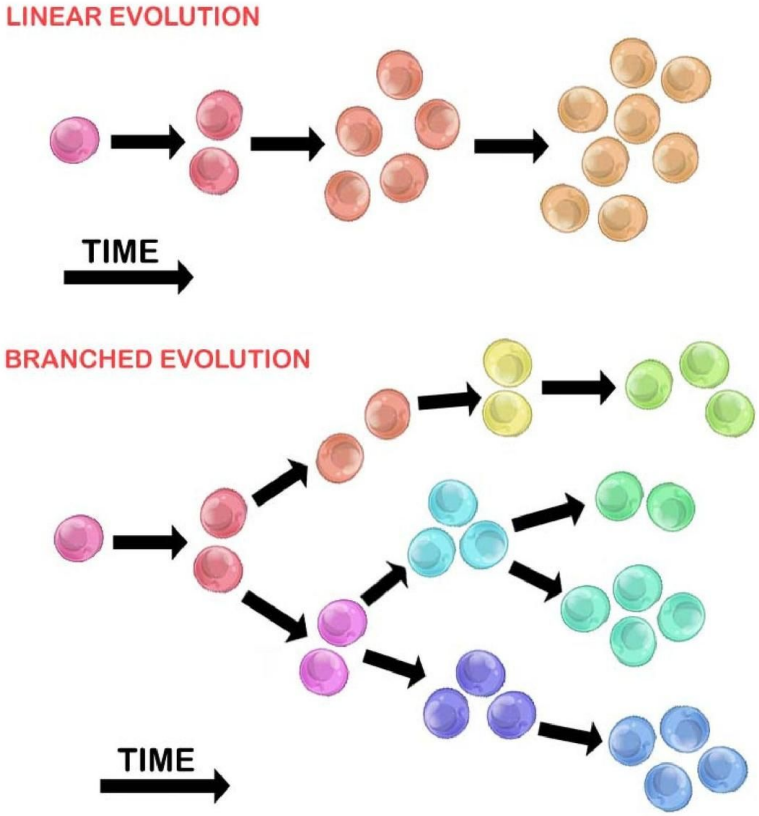
SevenBridges

- Advances in NGS and computational algorithms led to higher accuracy in somatic SNV calling

- Problems still present for true positives detection:
  - Low AF
  - Artifacts
  - Tumor contamination
  - Low coverage of high GC content regions
  - Sequencing errors

Clonal heterogeneity:

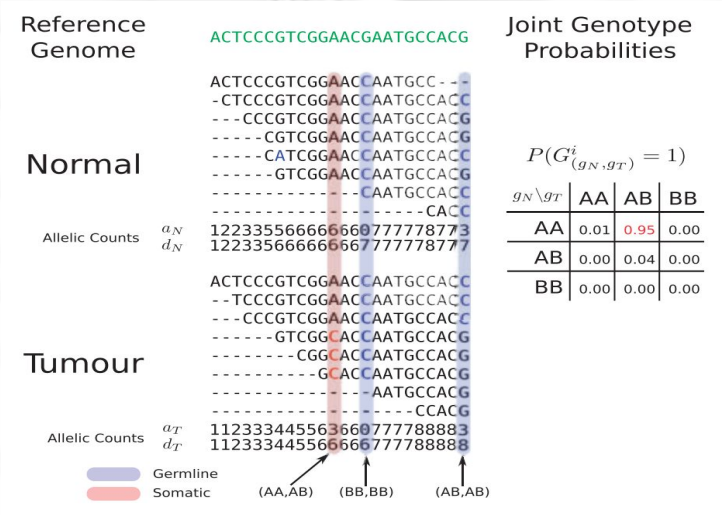
- Causes variants to be non-uniformly present in tumors
- False negative sSNVs (true sSNVs not called by the tools)
- False positives sSNVs (germline SNVs or not an SNV at all)



Cancer calling:

- Call tumor and normal separately and then compare them (early version of tools such as **SNVMix**)
- Call tumor and normal simultaneously and compare them simultaneously at each locus of a possible sSNVs (**JointSNVMix**, **SomaticSniper**, **Strelka**, **VarScan2**)
- Joint tumor - normal calling enables distinction between germline and somatic events, eliminating huge number of potential somatic false positives that are in fact germline mutations
- **MuTect** and **EBCall** are specialized in detecting low allele frequency somatic mutation

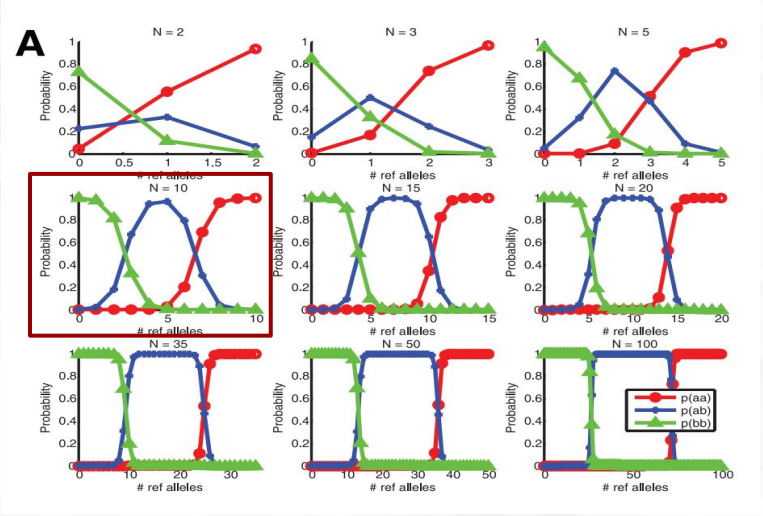
JointSNVMix



$P(G^i_{(g_N,g_T)} = 1)$

$g_N \backslash g_T$	AA	AB	BB
AA	0.01	0.95	0.00
AB	0.00	0.04	0.00
BB	0.00	0.00	0.00

Probabilistic approach based on a Binomial mixture model, called SNVMix1, which computes posterior probabilities, providing a measure of confidence on the SNV predictions



For a diploid genome, we consider all pairs of alleles that gives rise to the set, G = {AA,AB,BB},

Cartesian product of G with itself:

$G \times G = \{(g_N, g_T) : g_N, g_T \in G\}$

SomaticSniper

Bayesian comparison of the genotype likelihoods in the tumor and normal, as determined by the germline genotyping algorithm implemented using MAQ (Mapping and Assembly with Qualities) algorithm

$$S = -10\log_{10}\left(\sum_{G_i=0}^{9}\frac{P(T|G_i)P(G_i)P(N|G_i)P(G_i)}{\sum_{G_j=0}^{9}P(T|G_j)P(G_j)\sum_{G_k=0}^{9}P(N|G_k)P(G_k)}\right)$$

T - Tumor, **N** - Normal, **G** - Genotype,

D - data in either tumor or normal

P(D|Gi |) - Genotype likelihood. It is any of 10 possible diploid genotypes (i.e. AA, AC, AG, AT, CC, CG, CT, GG, GT, TT)

Genotype likelihood using MAQ (θ is the expected rate of heterozygous mutations in the population of interest and G_R is the reference base at the position of interest)

$$P(G_1) = \begin{cases} \theta \\ \frac{\theta}{2} \\ \theta^2 \\ 1 - \sum_{k=0}^{9} P(G_k)P(G_k \neq G_R) \end{cases}$$

Genotype is heterozygous but shares an allele with the reference

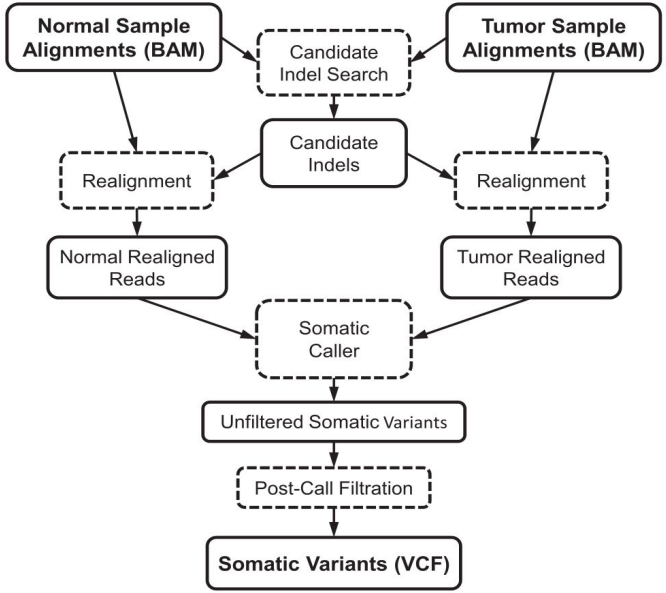
Genotype is homozygous variant

Genotype is heterozygous, but shares no alleles with the

Genotype is homozygous for the reference base

Strelka

- Strelka is essentially a mini workflow
- It performs an initial realignment around indels in the normal and cancer BAM files.
- Next, it uses a complex set of calculations, based on a Bayesian probability model, to report the most likely genotype at candidate sites along with the phred-scaled joint probability of the most likely normal sample genotype and the event of any somatic mutation in the cancer sample



MuTect

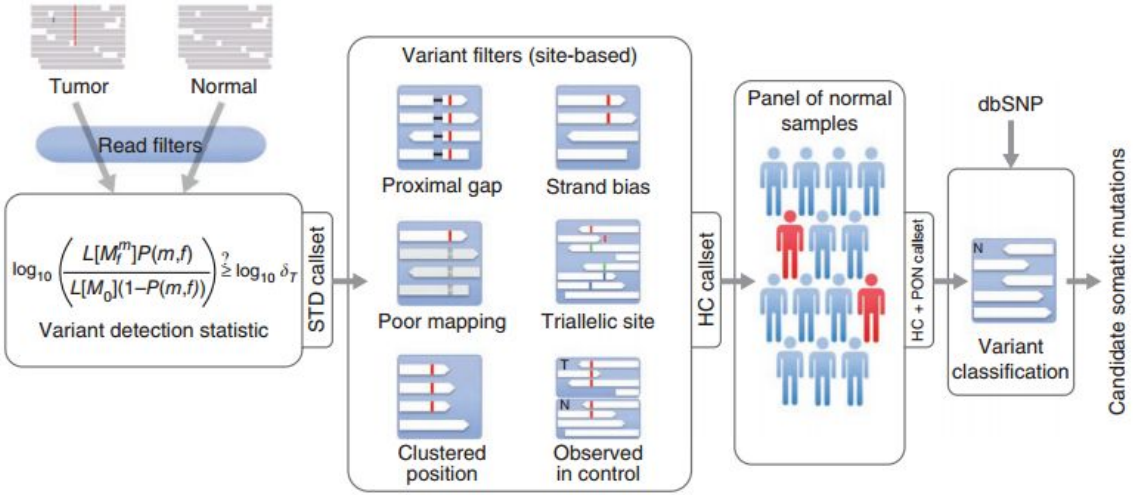
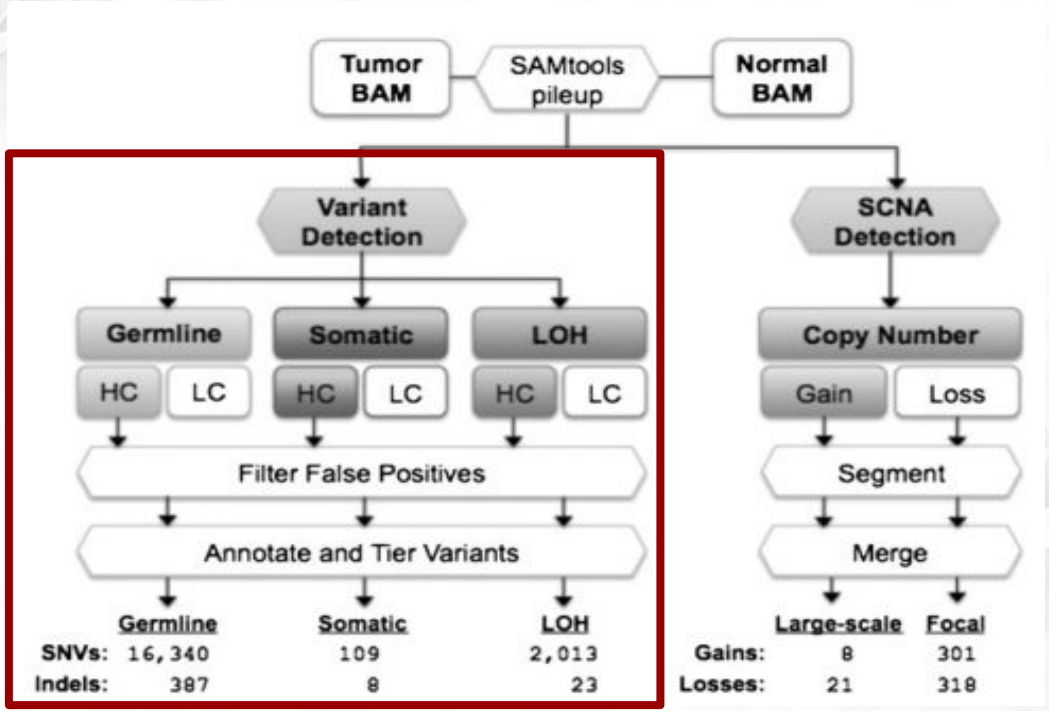


Figure 1 Overview of the detection of a somatic point mutation using MuTect. MuTect takes as input next-generation sequencing data from tumor and normal samples and, after removing low-quality reads (**Supplementary Methods**), determines whether there is evidence for a variant beyond the expected random sequencing errors. Candidate variant sites are then passed through six filters to remove artifacts (**Table 1**). Next, a panel of normal samples (PON) filter is used to screen out remaining false positives caused by rare error modes only detectable in additional samples. Finally, the somatic or germ-line status of passing variants is determined using the matched normal sample. STD, standard; HC, high confidence.

VarScan2


Tumor BAM
SAMtools pileup
Normal BAM
Variant Detection
SCNA Detection
Germline
Somatic
LOH
HC
LC
HC
LC
HC
LC
Copy Number
Gain
Loss
Filter False Positives
Segment
Annotate and Tier Variants
Merge
Germline
Somatic
LOH
SNVs: 16,340
109
2,013
Indels: 387
8
23
Large-scale
Focal
Gains: 8
301
Losses: 21
318

Mutation detection algorithm:

- First, it determines if both samples meet the minimum coverage requirement (by default, three reads with base quality ≥20
- By default, a variant allele must be supported by at least two independent reads and at least 8% of all reads
- Variants are called homozygous if supported by 75% or more of all reads at a position; otherwise they are called heterozygous.
- If the genotypes do not match, then their read counts are evaluated by one-tailed Fisher's exact test in a two-by-two table
- If the resulting P-value meets the significance threshold (default 0.10), then the variant is called somatic (if the normal matches the reference) or LOH (if the normal is heterozygous).
- If the difference does not meet the significance threshold, the variant is called germline
- If the genotypes match, the variant is called germline
- Germline, LOH, and somatic mutations are further categorized as HC or LC by the VarScan processSomatic command.

IGV comparison:

Tumor sample
Normal Sample
Mutation list

COSMIC

Website
Database (VCF)