Somatic variant calling and interpretation in the context of cancer

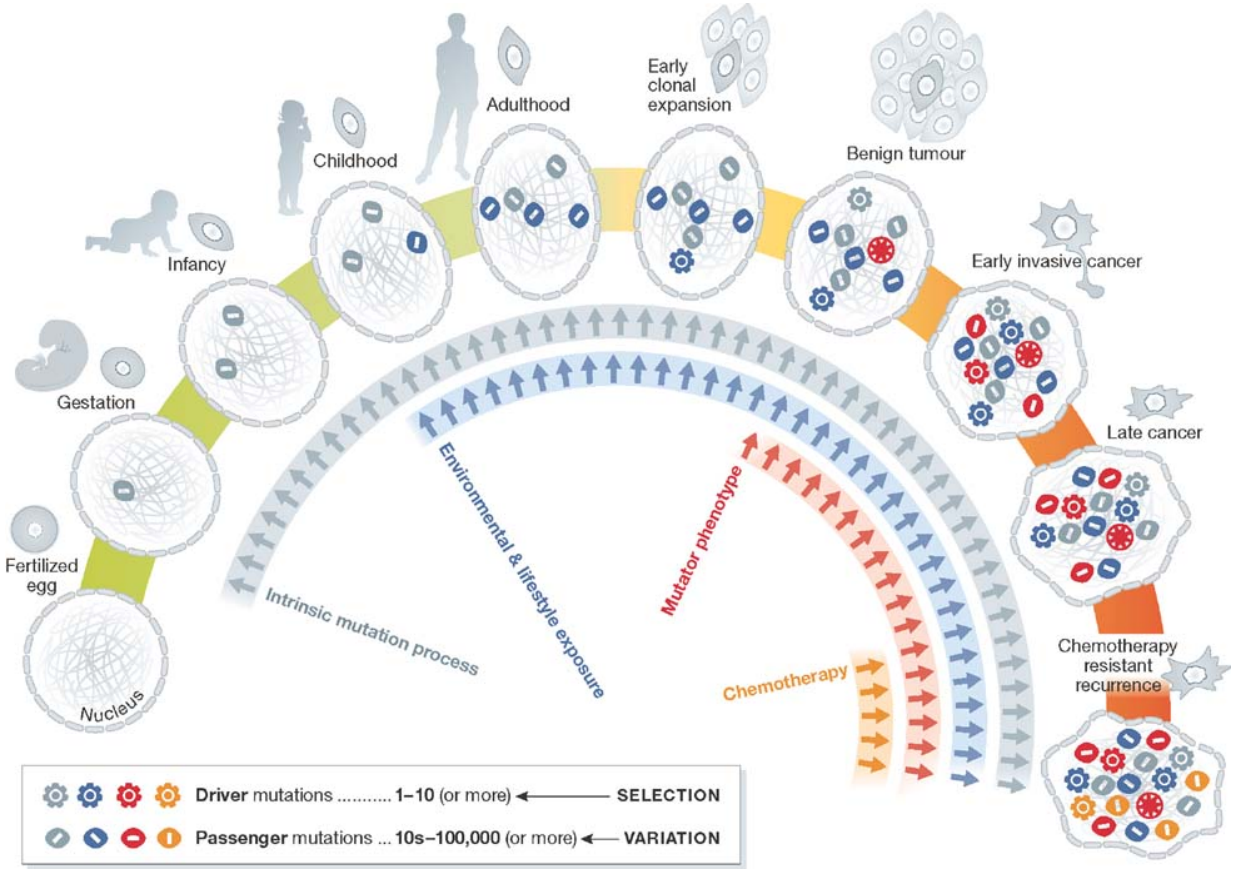
Sigve Nakken
Centre for Cancer Cell Reprogramming (CanCell)
Dept. of Tumor Biology
Institute for Cancer Research
Oslo University Hospital

Centre for Bioinformatics
Department of Informatics
University of Oslo

IN-BIOS 5000/9000 – Fall 2022

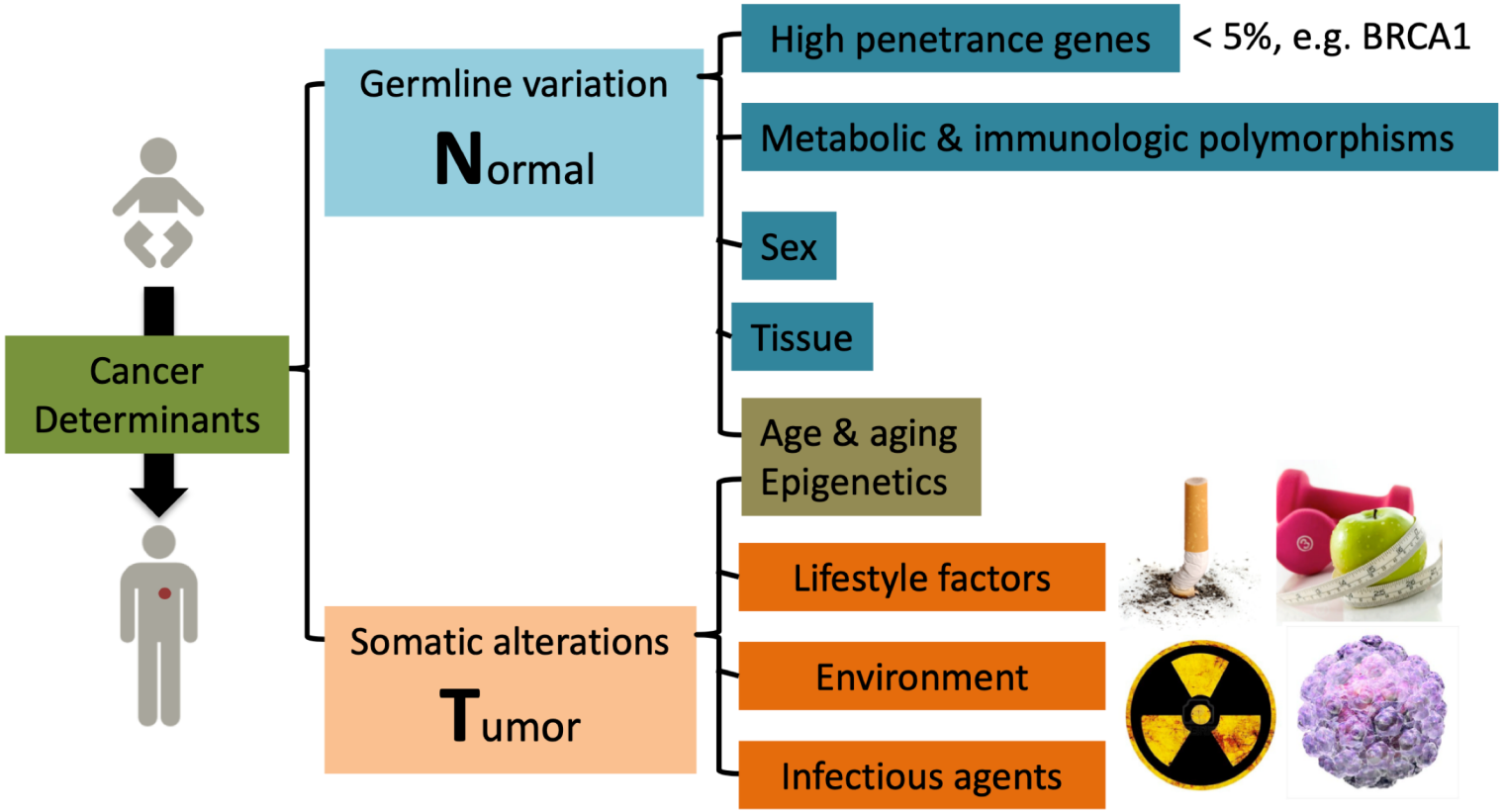
Why cancer?

Cancer – a disease of the genome



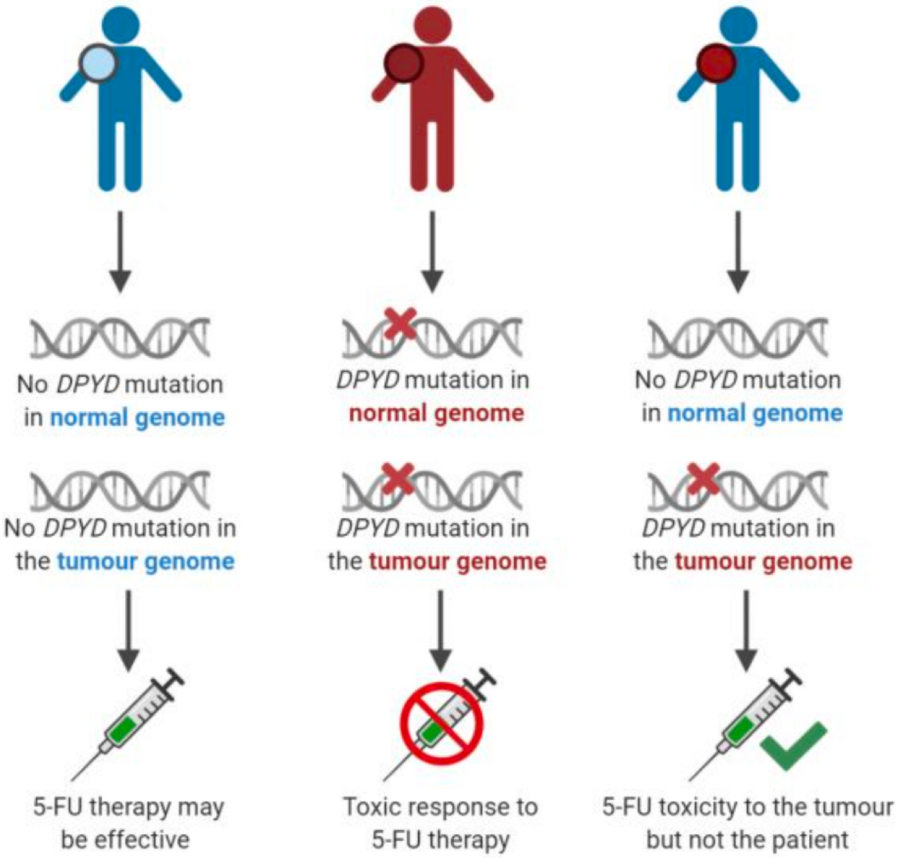
Stratton et al., EMBO Mol Med, 2013

Cancer – a disease of the genome



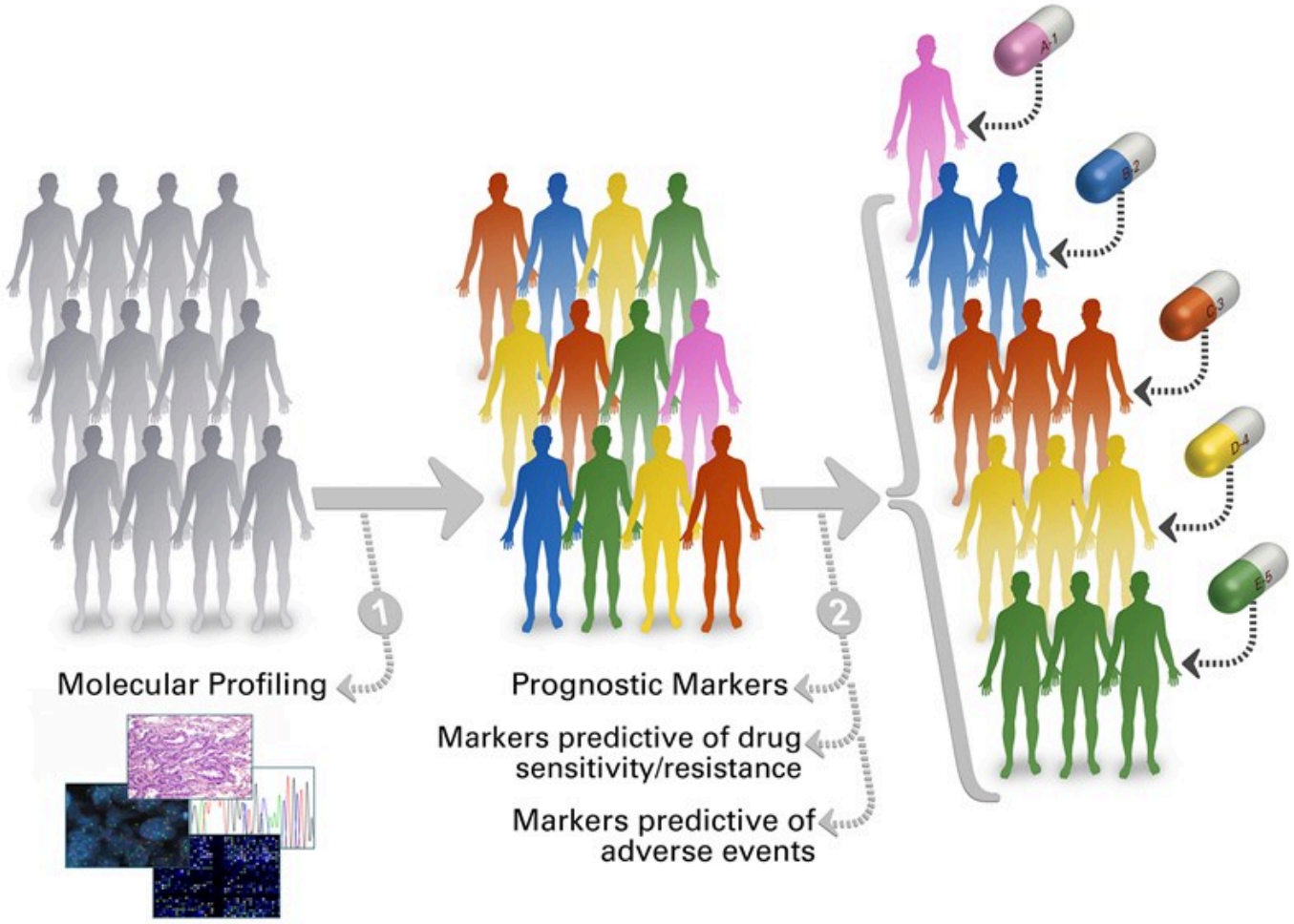
GATK: Introduction to Somatic Variant Discovery

Cancer sequencing informs on treatment options

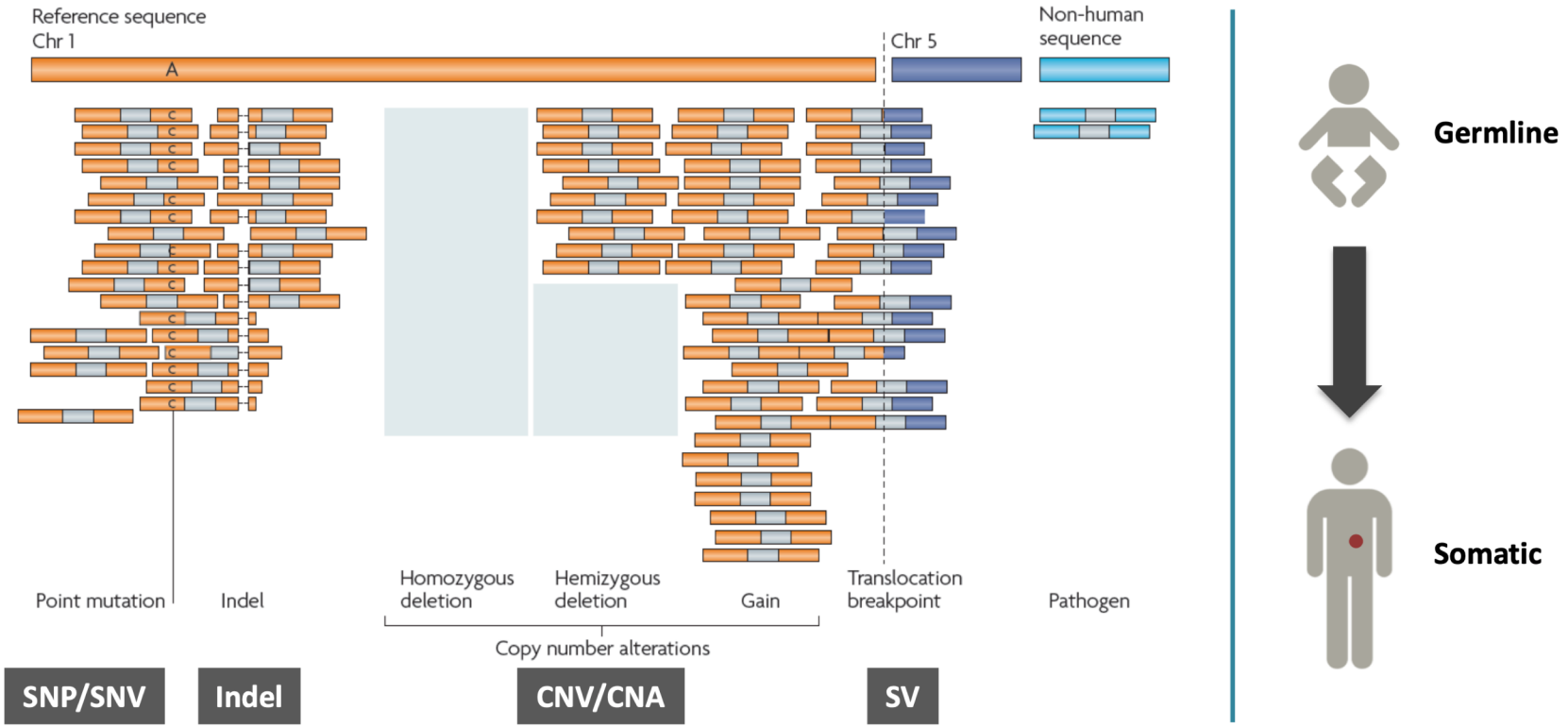


https://www.bcgsc.ca/news/genome-sequencing-helps-prioritize-cancer-treatment-options

Precision cancer medicine



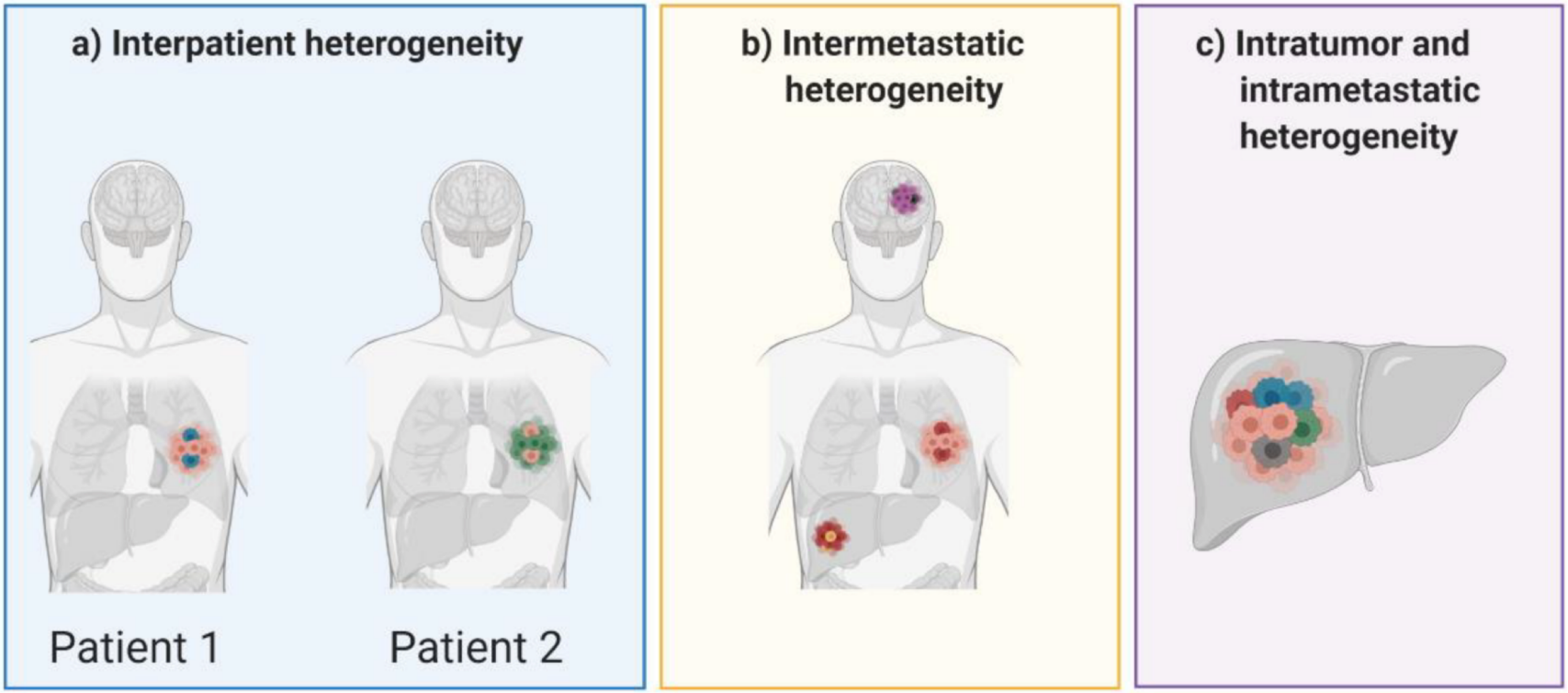
Cancer: multiple types of DNA aberrations



GATK: Introduction to Somatic Variant Discovery

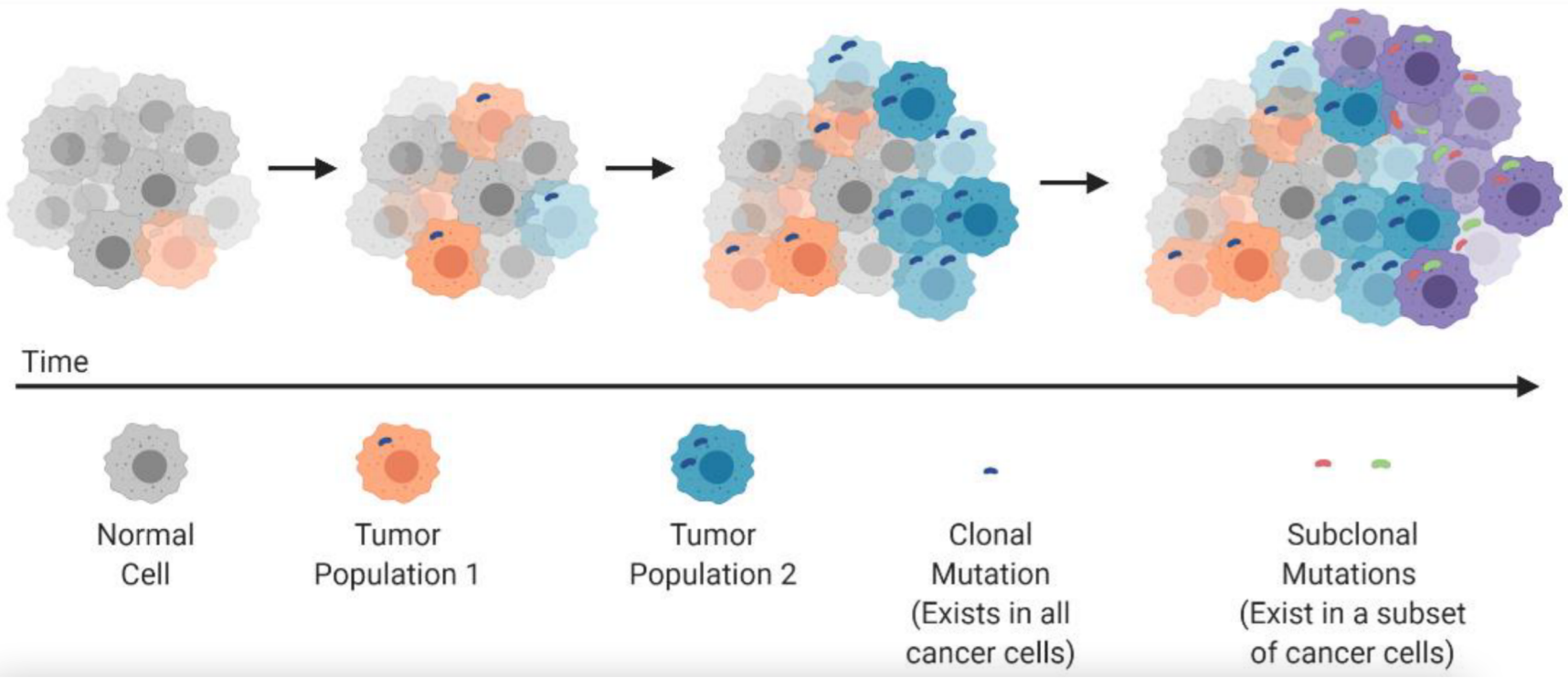
Why is somatic variant calling so challenging?

Cancer complexity



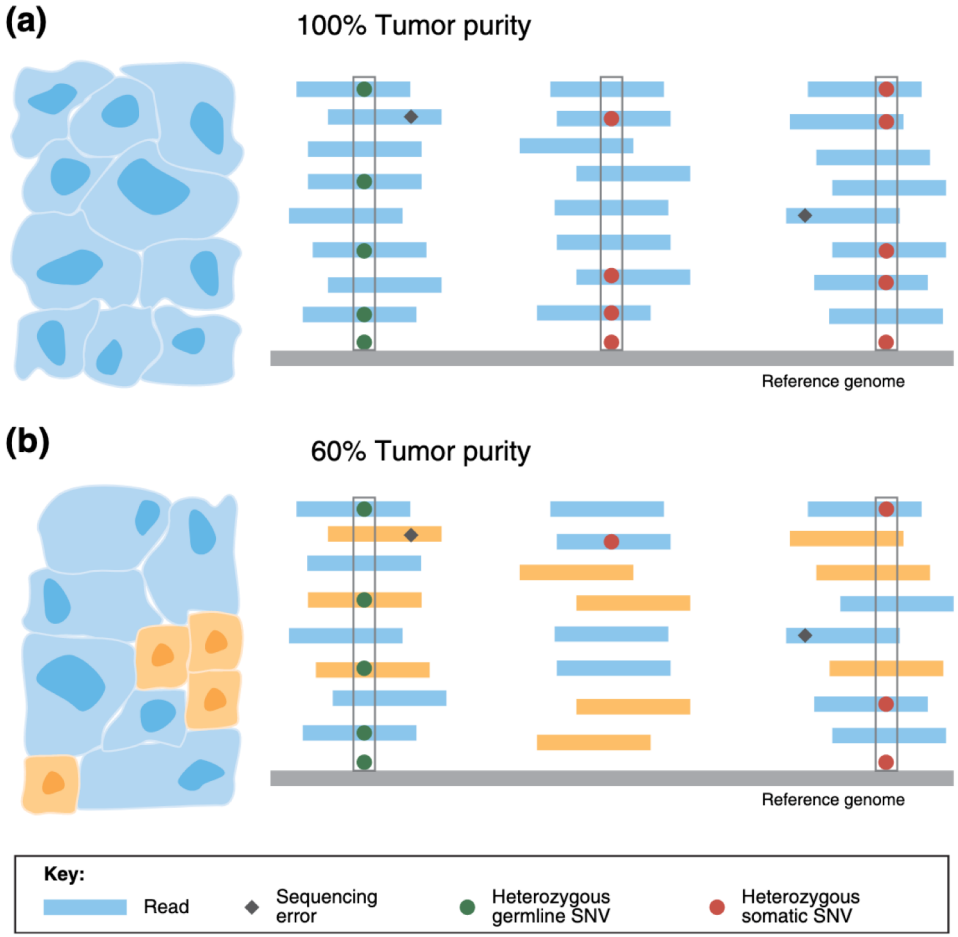
El-Sayes et al., Cancers, 2021

Tumor purity and heterogenity (I)



El-Sayes et al., Cancers, 2021

Tumor purity and heterogenity (II)



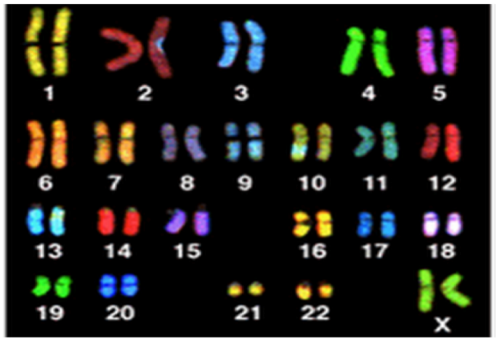
$$Tumor\ purity = \frac{tumor\ cells}{(normal + tumor\ cells)}$$

- Deep sequencing
- Implications for targeted sequencing coverage
- Purity is traditionally assessed manually by pathologists, but can also be inferred computationally

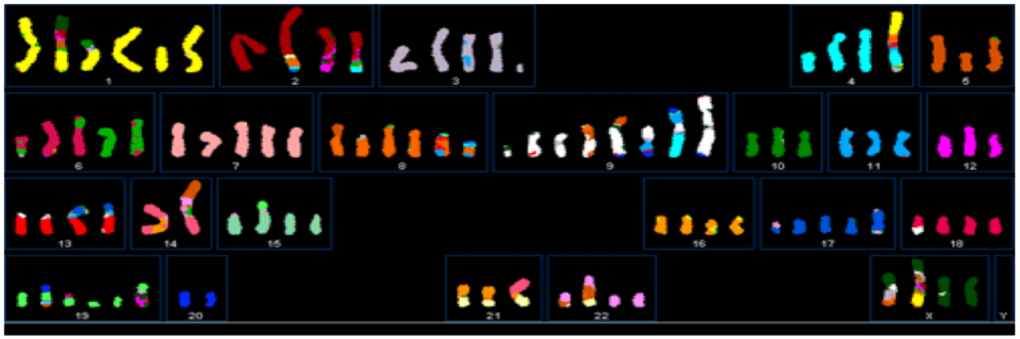
Raphael et al., Genome Med, 2014

Tumor ploidy

Normal Cell

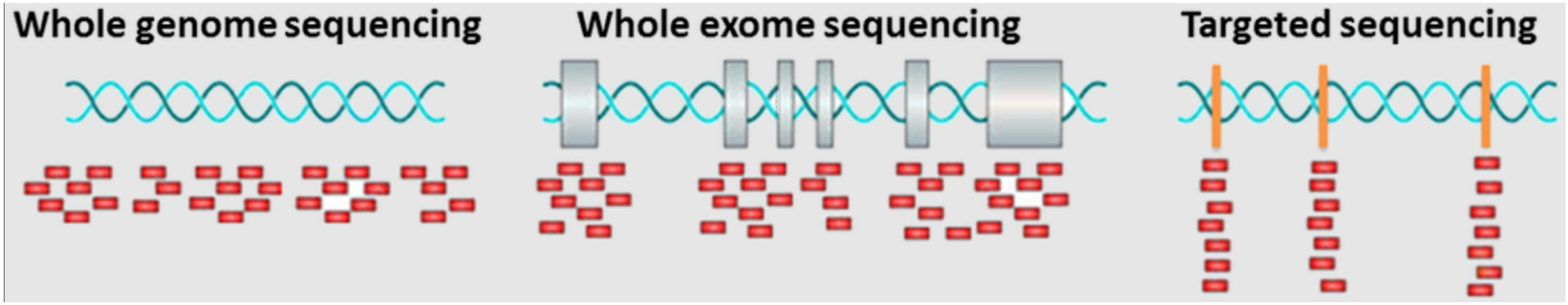


Cancer Cell Line HCC1954



- Somatic variant calling: make no ploidy assumption!

Cancer sequencing: assay design



- Typically 30-40x coverage
- More even coverage than WES
- Covers coding and non-coding/regulatory variation
- All types of variants (reliable detection of SVs)

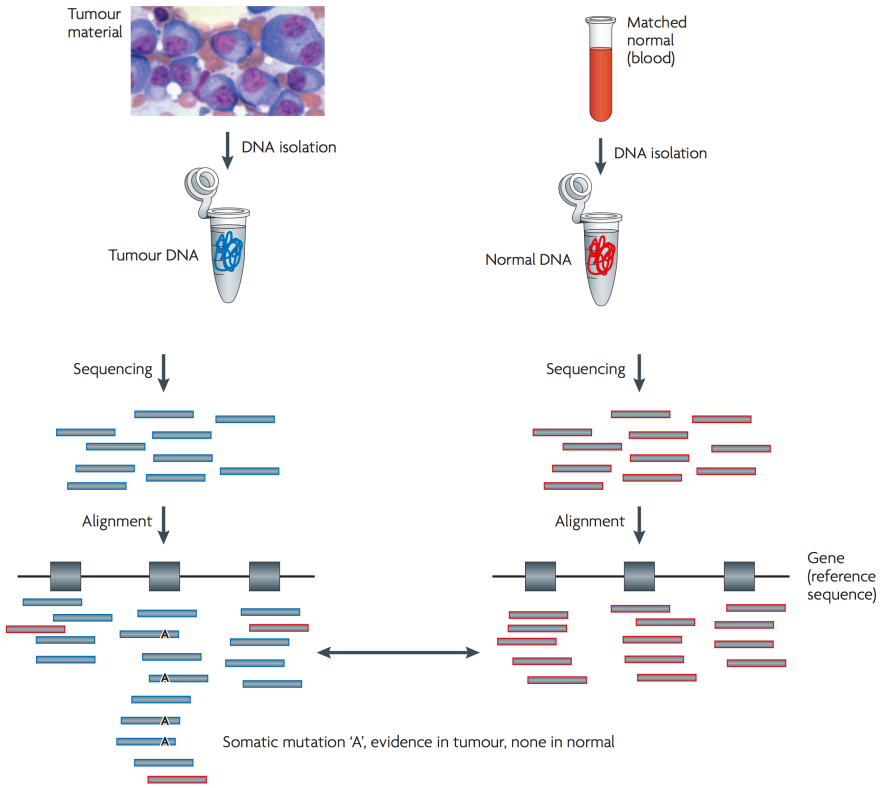
- Typically 80-100x coverage
- Coding regions only
- Cost-effective

- Typically > 300X coverage – captures subclonal variants at low allele frequencies
- Targets custom genes/regions – e.g. clinically actionable genes
- Most cost-effective

Research ← – – – – – – – – – → *Clinical applications*

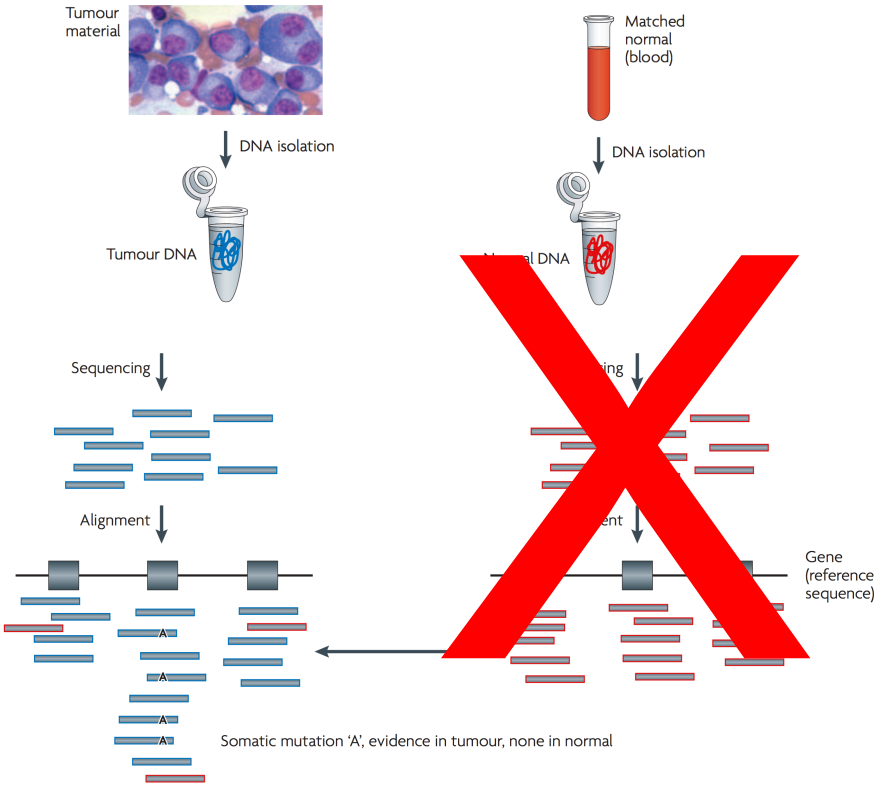
Cancer sequencing: calling design

- Two typical sequencing designs for detection of somatic variants
  - Tumor-control (**T + N**): most accurate

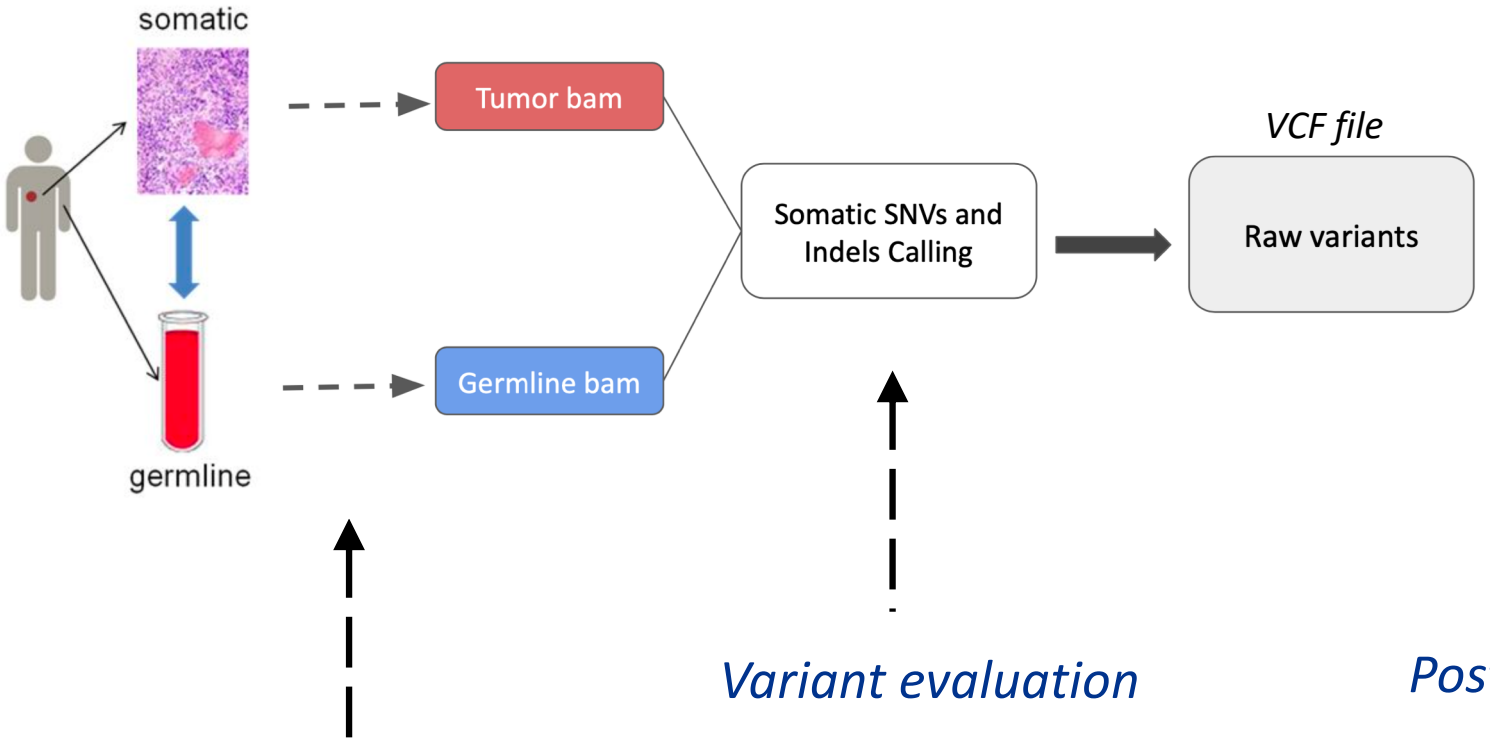


Cancer sequencing: calling design

- Two typical sequencing designs for detection of somatic variants
  - Tumor-control (**T + N**): most accurate
  - Tumor-only: most cost-effective



Somatic variant calling



GATK: Introduction to Somatic Variant Discovery

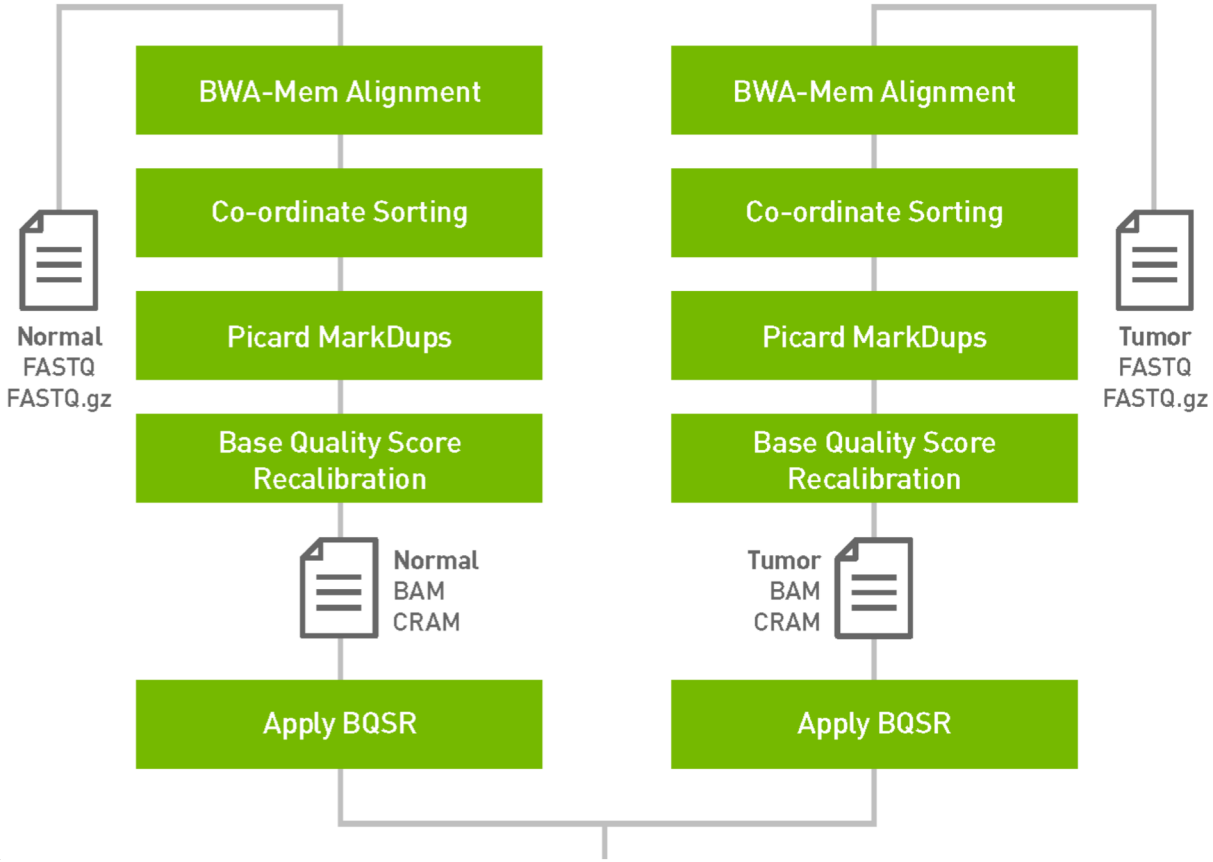
Somatic variant calling: pre-processing

Preprocessing

Quality control

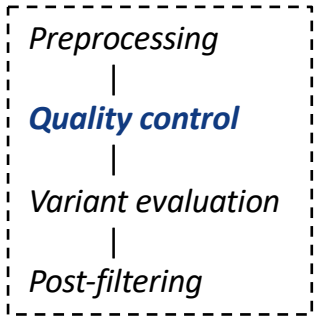
Variant evaluation

Post-filtering



https://docs.nvidia.com/clara/parabricks/v3.0/

Quality control (I)

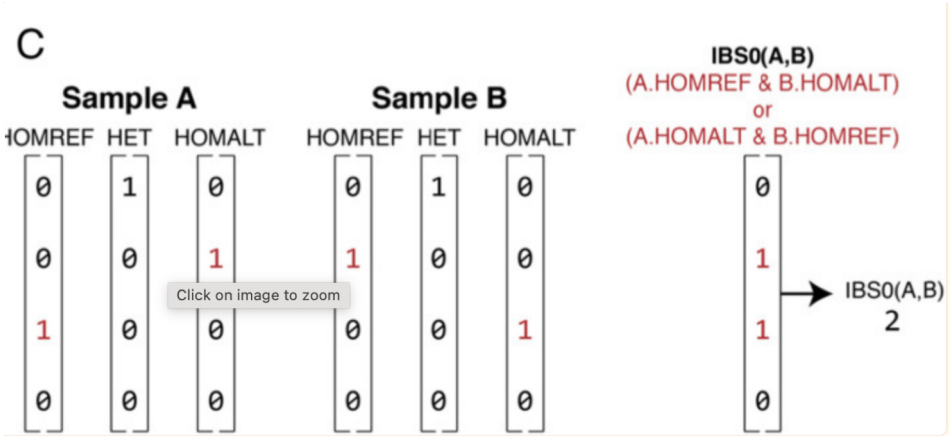


- Tumor samples subject to oxidative DNA damage during sample preparation could confound variant identification
  - Oxidation-induced C>A:G>T variants

- Detection?
  - Imbalance between complementary nucleotide substitutions
  - Tools: **GATK**

Quality control (II)

- Cross-sample contamination and sample relatedness
  - Different samples are frequently handled/sequenced together
    - Cross-individual contamination may occur, even small levels of contamination will have an impact on somatic variant detection
  - **T + N:** Check that tumor and normal sample come from the same individual!
  - Tools: **Conpair/Somalier**



Pedersen et al., Genome Med, 2020

Somatic variant calling (T + N)

Preprocessing
|
Quality control
|
Variant evaluation
|
Post-filtering

- Logic for somatic variant calling algorithms using tumor-normal design: **"substract" the germline background**
  - For a given candidate site, is the difference between tumor and normal significant?

Somatic variant calling (T + N)

Preprocessing
|
Quality control
|
Variant evaluation
|
Post-filtering

- ***First generation***: call somatic candidates through heuristic rules/ad-hoc filters
  - Rule out sequencing artefacts by thresholds (number of supporting reads etc)
  - Statistical test of difference between tumor and normal
  - Callers: **VarScan2, VarDict**
- ***Second generation***: probabilistic modeling of allele frequencies
  - What is the likelihood of non-reference base being somatic, and not sequencing noise (considering base quality, sequence context etc.)?
  - Callers: **MuTect2, Strelka2**
- ***Post-filtering***: add additional quality control on call set (read support, minimum coverage, support on both strands etc.)

Somatic variant calling (T + N)

- How to choose variant calling algorithm for a particular sequencing project?
  - Check out benchmarking results
    - A few benchmarking datasets are available – providing “gold sets” of somatic mutations
    - **Precision vs. recall**
    - Benchmarking results are often misleading
      - Which calling parameter values should be used?
  - Check whether the algorithm is designed for your assay and technology
    - E.g. has it shown good performance for detection of subclonal variants at low allele frequencies?

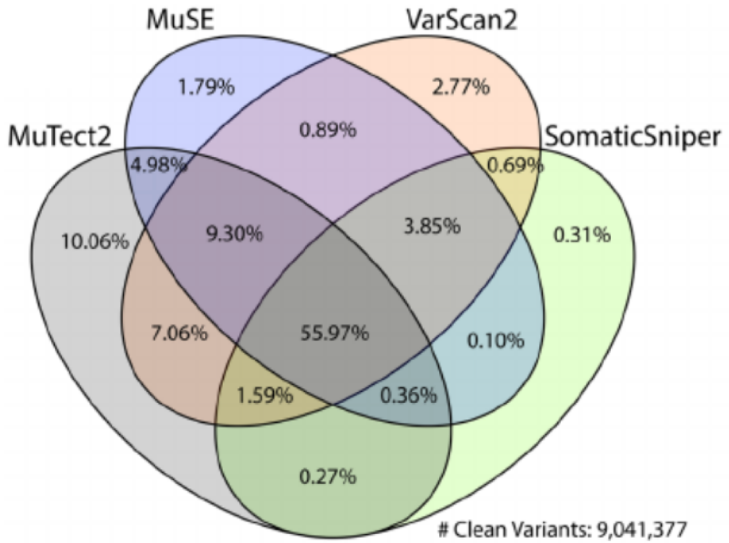
nature COMMUNICATIONS

A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing

Tyler S. Alioto, Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D. Eldridge, Eivind Hovig, Lawrence E. Heisler, Timothy A. Beck, Jared T. Simpson, Laurie Tonon, Anne-Sophie Sertier, Ann-Marie Patch, Natalie Jäger, Philip Ginsbach, Ruben Drews, Nagarajan Paramasivam, Rolf Kabbe, Sasithorn Chotewutmontri, Nicolle Diessl, Christopher Previti, Sabine Schmidt, Benedikt Brors, Lars Feuerbach, Michael Heinold, Susanne Gröbner, Andrey Korshunov, Patrick S. Tarpey, Adam P. Butler, Jonathan Hinton, David Jones, Andrew Menzies, Keiran Raine, Rebecca Shepherd, Lucy Stebbings, Jon W. Teague, Paolo Ribeca, Francesc Castro Giner, Sergi Beltran, Emanuele Raineri, Marc Dabad, Simon C. Heath, Marta Gut, Robert E. Denroche, Nicholas J. Harding, Takafumi N. Yamaguchi, Akihiro Fujimoto, Hidewaki Nakagawa, Víctor Quesada, Rafael Valdés-Mas, Sigve Nakken, Daniel Vodák, Lawrence Bower, Andrew G. Lynch, Charlotte L. Anderson, Nicola Waddell, John V. Pearson, Sean M. Grimmond, Myron Peto, Paul Spellman, Minghui He, Cyriac Kandoth, Semin Lee, John Zhang, Louis Létourneau, Singer Ma, Sahil Seth, David Torrents, Liu Xi, David A. Wheeler, Carlos López-Otín, Elías Campo, Peter J. Campbell, Paul C. Boutros, Xose S. Puente, Daniela S. Gerhard, Stefan M. Pfister, John D. McPherson, Thomas J. Hudson, Matthias Schlesner, Peter Lichter, Roland Eils, David T. W. Jones & Ivo G. Gut - Show fewer authors

Somatic variant calling (T + N)

- How to choose variant calling algorithm for a particular sequencing project?
  - Each caller typically has some strengths and weaknesses
    - a common strategy is now to **apply multiple callers and combine the variant sets**
    - "The wisdom of crowds"
    - Consensus? Majority vote? Machine learning?
    - Combining information from VCF files/callers are frequently challenging in practice



Zhang et al., Nat Commun, 2021

Somatic variant calling: VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CPCT02080287R CPCT 02080287T

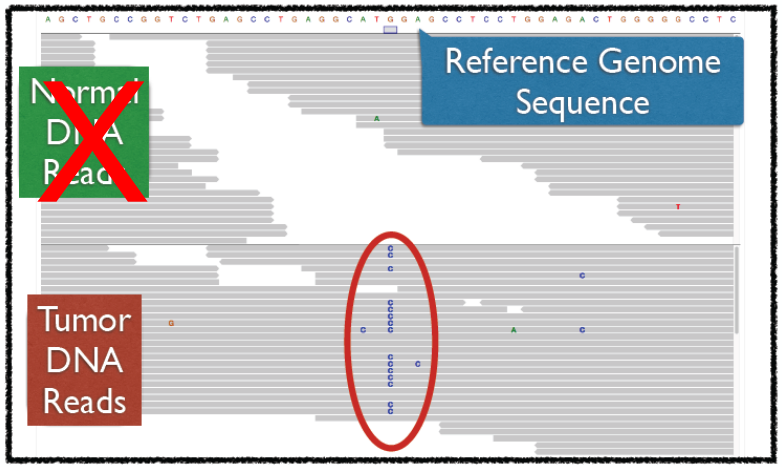
1 854389 . G A 590 **PASS** IMPACT=LINC02593,ENST00000609207,non_coding_transcript_exon_variant,NONE,false,n.2008C>T,,,NONE,1

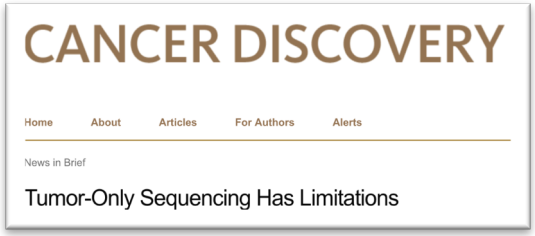
GT:**AD:AF:DP**:RABQ:RAD:RC_CNT:RC_IPC:RC_JIT:RC_QUAL:RDP:SB 0/0:**42,0**:**0**:42:1528,0:42,0:0,0,0,0,0,42,42:0:0,0,0:0,0,0,0,0,1097,1097:42:0 0/1:**43,20**:**0.317**:63:1626,741:46,21:17,3,0,0,0,43,63:0:0,0,0:533,57,0,0,0,1318,1908:67:0.35

Allelic support – tumor sample

Allelic support – normal sample

Somatic variant calling: Tumor-only





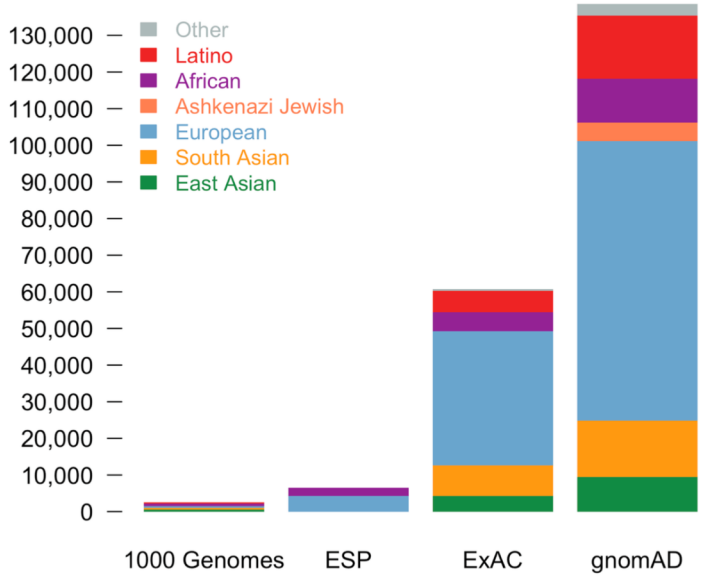
- Cost-effective strategy for identification of somatic variants – much used in the clinic
- Main challenge: robust subtraction of the germline background
  - Approach: use other sources of germline variation (databases)
  - Each individual is estimated to carry an extensive set of rare variants (i.e. *singletons*)
  - Ethnic subpopulations are under-represented in germline variant databases

Tumor-only variant filtering: **gnomAD**

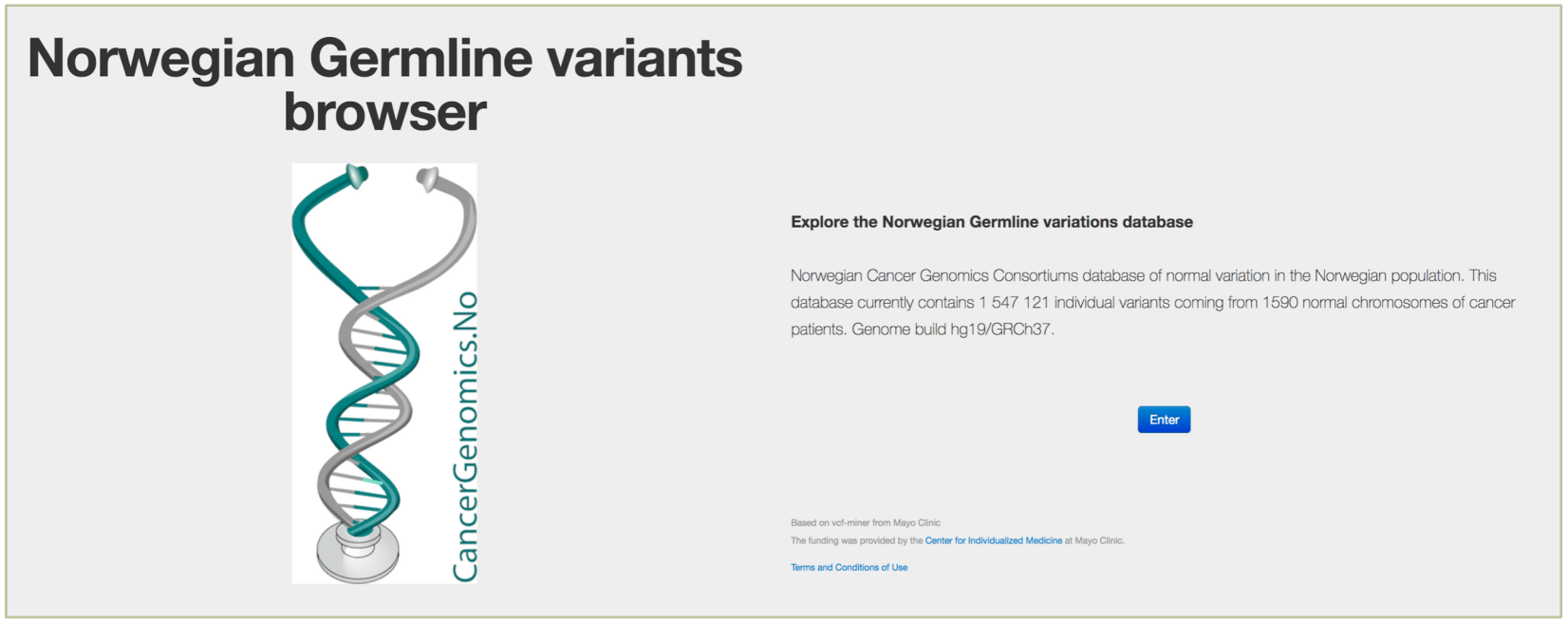
- genome **A**ggregation **D**atabase
- **Harmonizes** germline variant both exome and genome sequencing data from a wide variety of large-scale sequencing projects
- **Freely available** to the scientific community
- ~125,000 WES samples
- ~16,000 WGS samples

gnomAD

genome aggregation database



Tumor-only variant filtering: **norgene**



norgene.no

Tumor-only variant filtering: **panel-of-normals**

- What is a «panel-of-normals (PON)»?
  - Variant calls made from a set of unrelated “normal” samples
- Purpose of PON?
  - Eliminate common/recurring technical artifacts
    - should use normals made using the same data generation techniques (e.g. same capture kit for exomes, same sequencing platform etc.)
  - Secondary purpose: also eliminates germline variants not called in the matched normal (or approximates the normal if none is available)

Somatic variant calling: summary

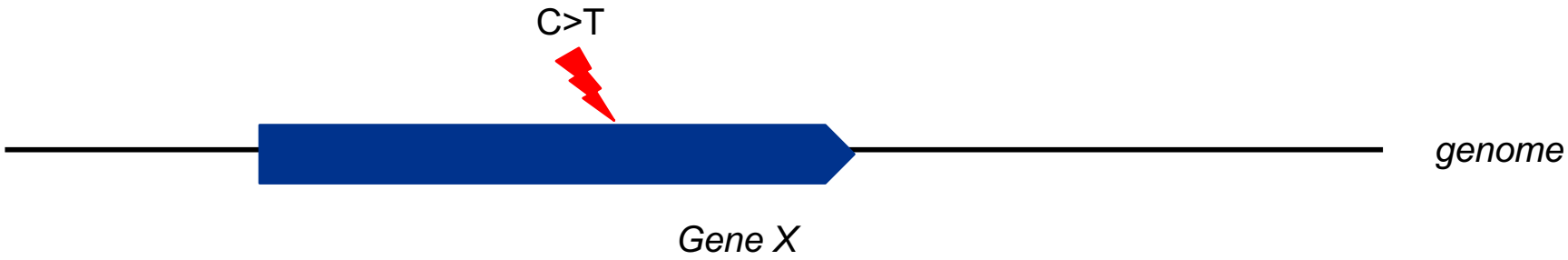
- The complexity of tumors pose challenges for variant identification – intratumor heterogeneity, tumor purity, ploidy
- WGS – WES – Targeted sequencing (research → clinic)
- Two fundamental sequencing designs: Tumor-control and tumor-only
- Multiple calling algorithms exist – each with strengths and weaknesses - a common strategy is to combine output from several callers
- Benchmarking results exist – can they be generalized?
- Understand the nature of your data/tumor and the priorities of the variant identification procedure when choosing a calling strategy

Variants have been found – now what?

ACTG**C**CTACGTCTCACCGTCGACTTCAAATCG**C**TTAACCCGTACTCCCATGCTACTGC
ATCTCGGGTTAACTCGACGTTTT**T**CATGCATGTGTGCACCCCAATATATATGCA**A**CTT
TTGTGCACCTCTGTCACGCGCGAGTTGGCACTGTCGCCCCTGTGTGCATGTGCACTGT
CTC**T**CGCTGCACTGCCTACGTCTCACCGTCGACTTCAAATCG**C**TTAACCCGTACTCCC
ATGCTACTGCATCTCGGGTTAACTCGACGTTTT**G**CATGCATGTGTGCACCCCAATATA
TATGCA**A**CTTTTGTGCACCTCTGTCACGCGCGAGTTGGCACTGTCGCCCCTGTGTGCA
TGTGCACTGTCTC**T**CGAGTTTT**G**CATGCATGTGTGCACTGTGCACCTCTGTTACGTCT

QUESTIONS/BREAK

Variant interpretation - general

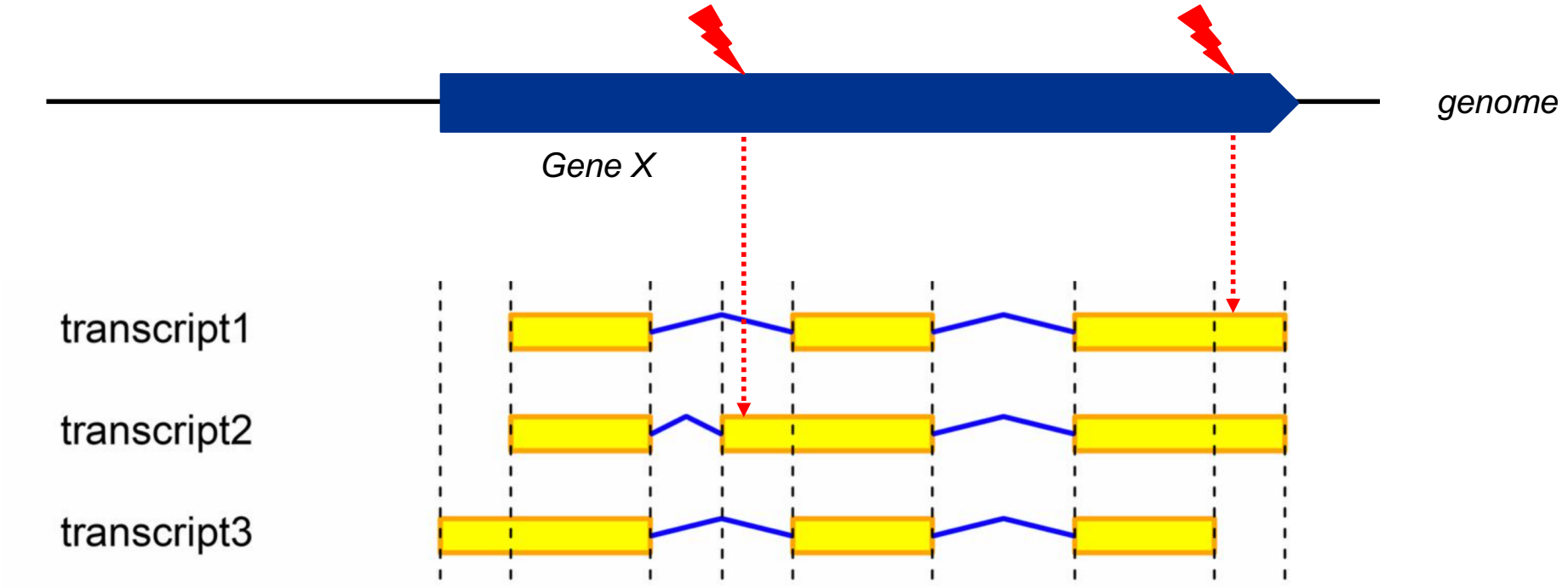


1. Which genes are affected by variants?

2. For a given gene variant, what is the consequence for the encoded protein?
   - Loss-of-function?

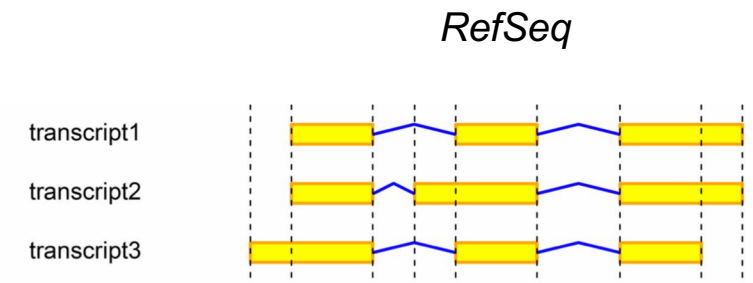
Variant interpretation - general

- A gene consists of multiple transcript isoforms

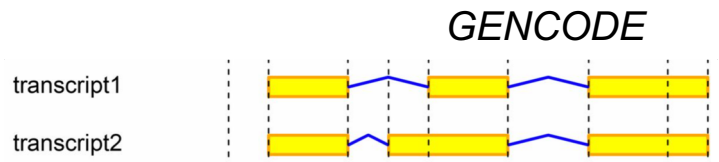


Variant interpretation – general (II)

- Several transcript databases
  - RefSeq
  - Ensembl
  - GENCODE



- Choice of transcript database impacts variant consequence/ annotation



- Frequent strategy: Report variant consequence in most commonly expressed isoform (i.e. ***principal*** isoform)

Variant interpretation - cancer

- Variant interpretation for cancer precision medicine
  - Where are the mutations located (**which genes** are mutated, and which variants are most relevant)?
    - **Therapeutic markers** (diagnosis and prognosis)
    - Germline (predisposing) + somatic
  - **What types** of mutations are found?
    - **Mutational signatures**
      - Tumor etiology, therapeutic and diagnostic markers
  - **How many** mutations are found?
    - **Tumor mutational burden** - immunotherapy

Personal Cancer Genome Reporter

Cancer Predisposition Sequencing Reporter

Which genes are mutated? (I)

- Specific genetic aberrations indicate clinical actionability
  - **Drug sensitivity**
  - Prognosis / Diagnosis
  - Drug resistance

- Multiple initiatives curate clinical variant associations in cancer

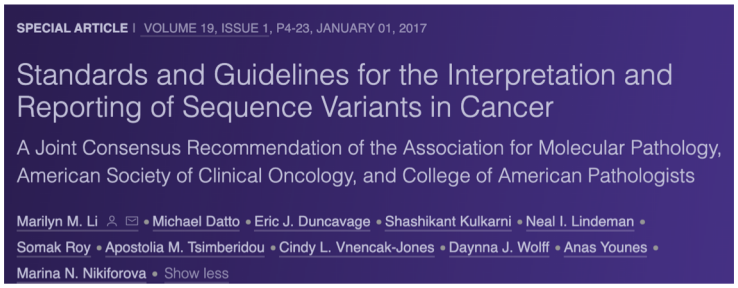
  - **Variant X** in **phenotype Y** indicates sensitivity to **drug Z**

  - **Challenge:** harmonization of knowledge databases
    - VICC (Variant Interpretation for Cancer Consortium)

- *HER2 amplification – Trastuzumab – Breast Cancer*
- *BRAF V600E – Vemurafenib – Melanoma*
- *BRAF V600E – Trametinib + Dabrafenib – NSCLC*
- *IDH1/2 mutations – Ivosidenib – AML*
- ..
- ..

Which genes are mutated? (II)

- Which somatic aberrations are most relevant in my tumor sample (actionability)?
  - Ranking and standardization frameworks - tiers
  - Key: **Strength of evidence**
  - Tumor type (**on-label** vs. **off-label**)



- **TIER 1** – strong evidence for clinical impact, same tumor type as query
- **TIER 2** – strong evidence for clinical impact in other tumor type or weak evidence for clinical impact in query tumor type
- **TIER 3** – uncertain clinical significance; coding variants in tumor suppressor genes/proto-oncogenes (mutation hotspots etc)
- **TIER 4** – other coding variants

How many mutations are found?

- **Tumor mutational burden (TMB)** - number of somatic mutations per megabase of interrogated genomic sequence

- A key driver in the generation of immunogenic neopeptides – influences **response to immune checkpoint inhibitors (ICIs)**

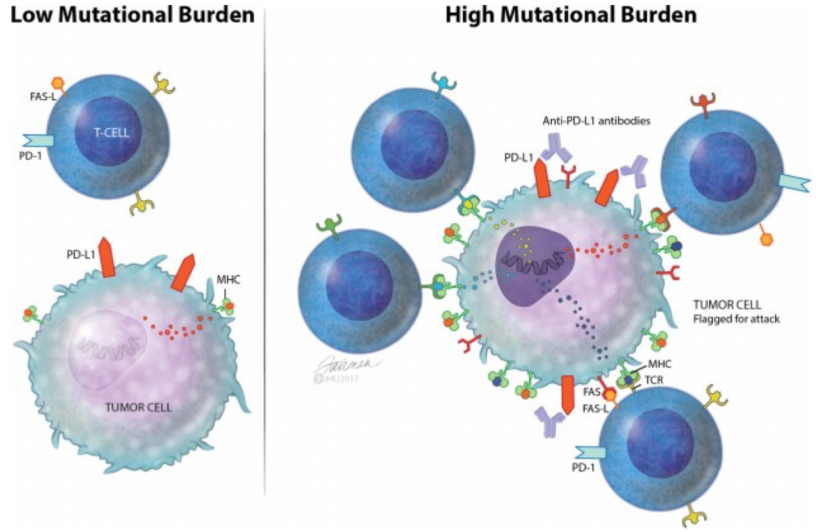


Illustration from Sharabi et al., The Oncologist (2017)

What types of mutations are found?

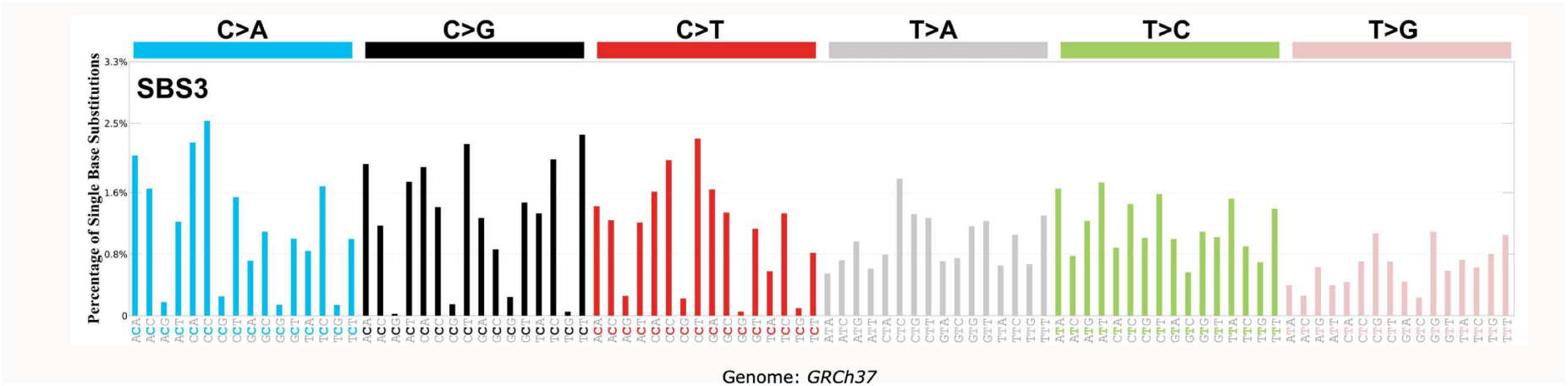
- **Mutational signatures:** characteristic mutation patterns (types and sequence context) that arise from a specific mutational process

- **Premise**: mutational processes are context-dependent (occur non-randomly in DNA)

- **Footprint:** The global set of mutations harvested from NGS **reveals a «historical footprint»** of the mutational processes that have shaped a given tumor
  - Environmental mutagens
  - Endogenous mutation processes (e.g. DNA repair defects)
  - Treatment effects
  - **Approximatly 50 established mutational signatures**

Mutational signatures (I)

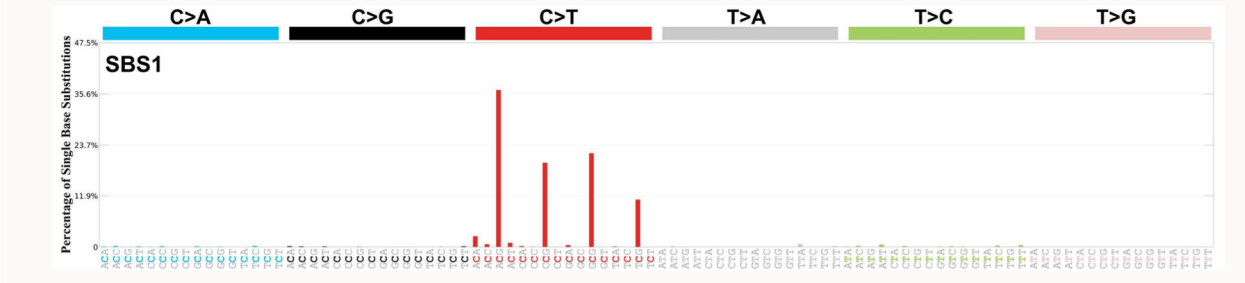
- Mutational signatures are most commonly presented through the **96-channel** approach (single base substitutions, SBS)
  - Mutation type + flanking bases



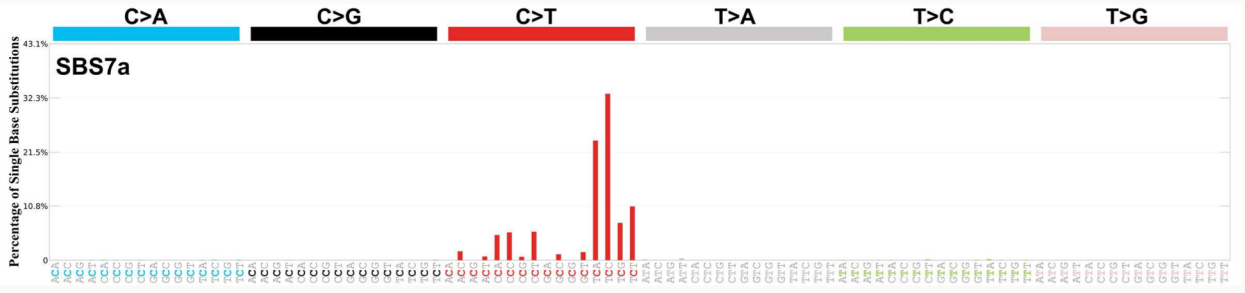
  - A single signature (attributed to a given process) is thus characterized as the **relative frequency of 96 different channels**

Mutational signatures (II)

- Aging
  - spontaneous or enzymatic deamination of 5-methylcytosine to thymine (clock-like signature)

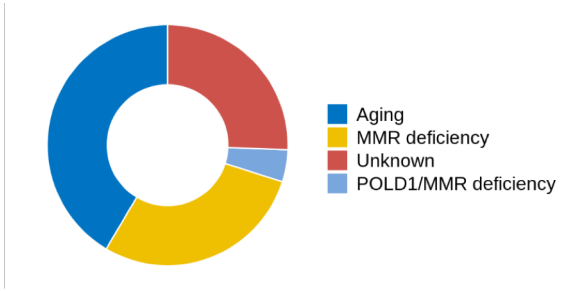


- Exposure to UV light
  - cyclobutane pyrimidine dimers or 6-4 photoproducts



Mutational signatures (III)

- Tools can «deconstruct» the profile of somatic mutations in a tumor towards contribution of known signatures
- Signatures are emerging as an important biomarker for drug response
- Often considered in combination with other markers
- Challenge: **confidence**

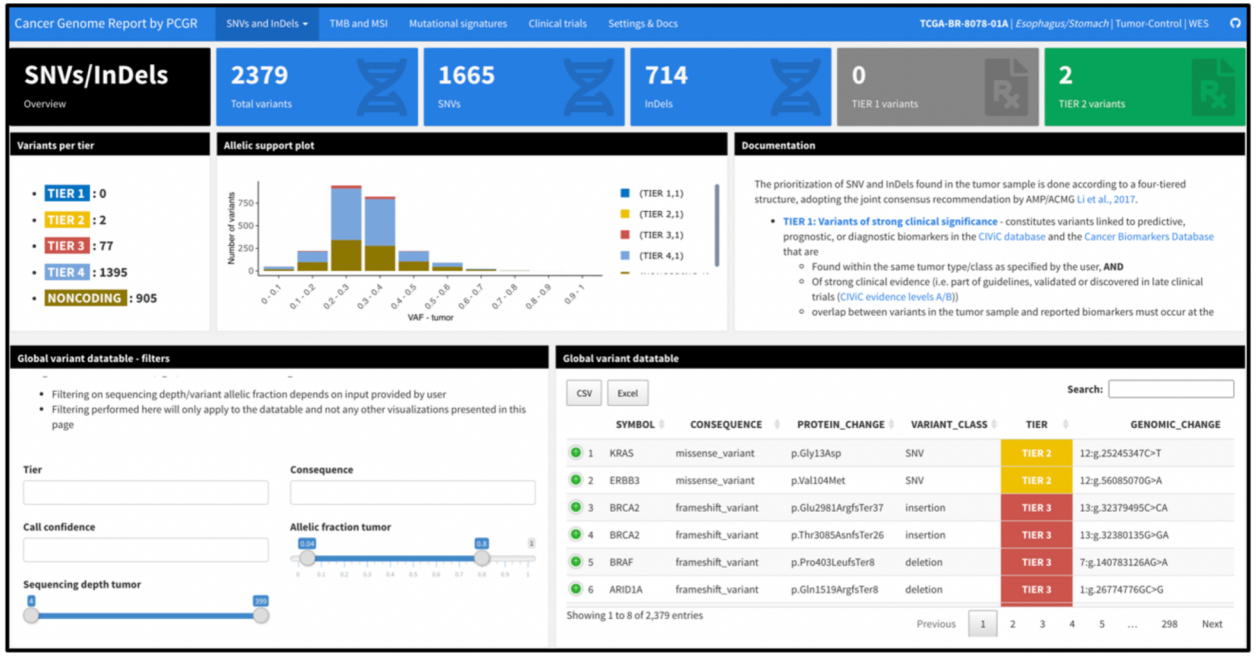


		CS-3	CS-8	Homologous Recombination Repair Deficiency	PARP inhibition[32-34], Platinum-based chemotherapy[35-37]
CS-6	CS-15	CS-20	CS-26	Mismatch Repair Deficiency	PD1-immunotherapy[48-49,52]
		CS-5	CS-8	Nucleotide Excision Repair Deficiency	Cisplatin[63-65]
		CS-18	CS-30	Base excision Repair Deficiency	
			CS-10	Deficient DNA polymerase proofreading activity	PD1-immunotherapy[48-49,52]

Adopted and modified from Van Hoeck et al., BMC Cancer, 2019

Personal Cancer Genome Reporter (I)

- In brief: A reporting engine for clinical interpretation of tumor genomes (VCF → interactive report)



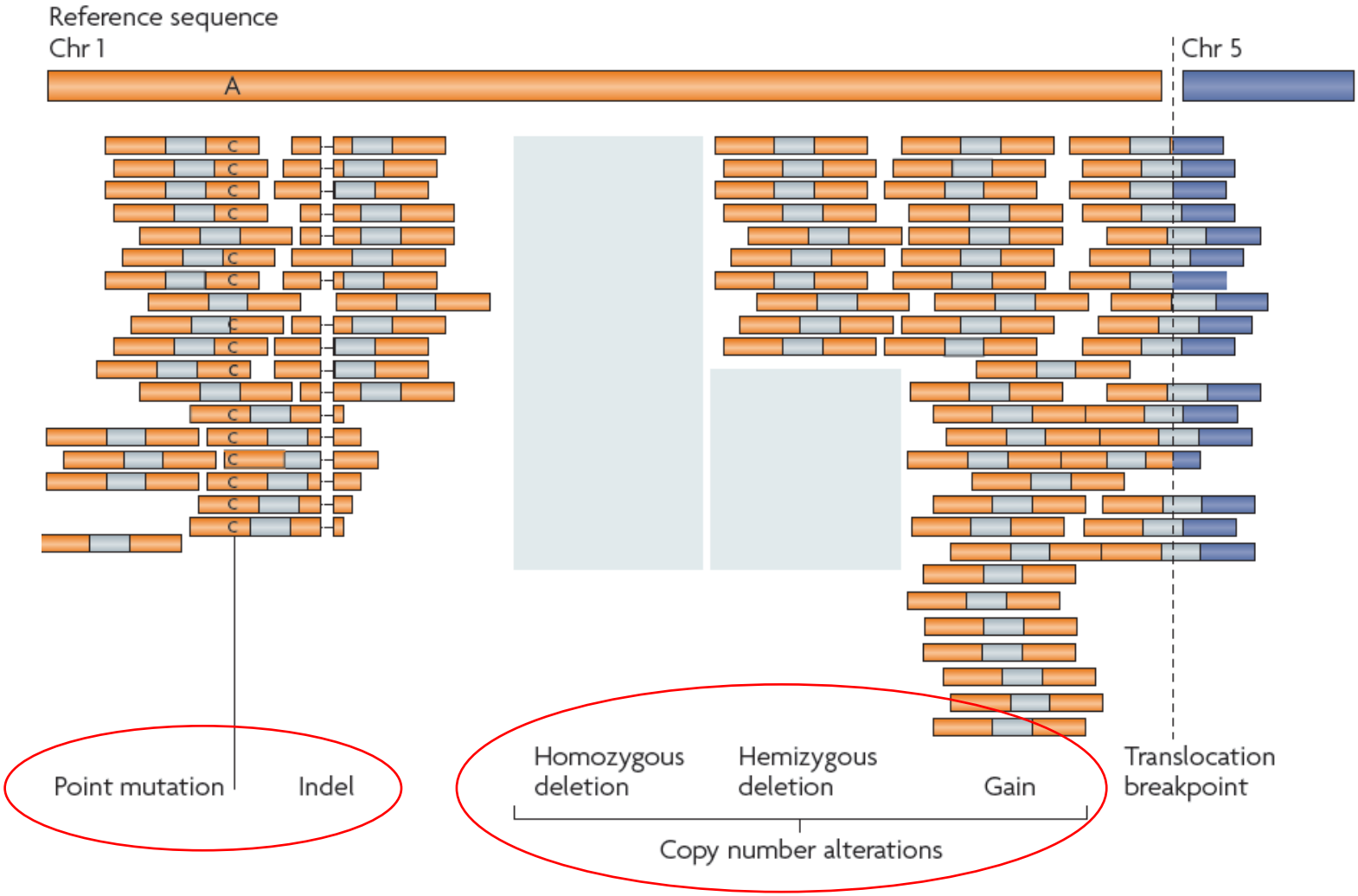
Nakken et al., Bioinformatics, 2017

https://github.com/sigven/pcgr

Personal Cancer Genome Reporter (II)

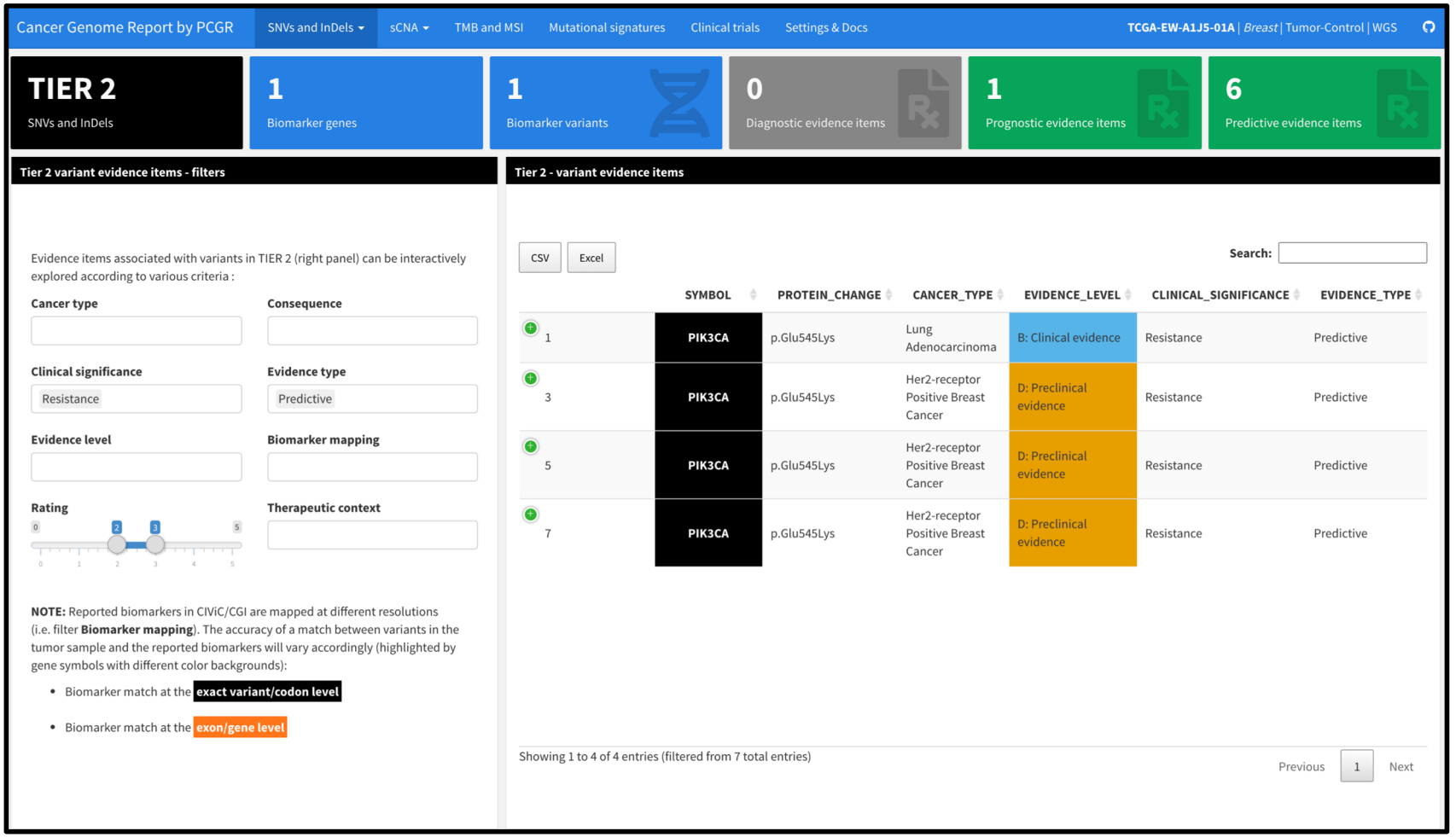
- PCGR captures current knowledge on cancer precision medicine through **data integration** of publicly available databases
  - Targeted cancer drugs
  - Known biomarkers for prognosis and diagnosis
  - Known biomarkers for drug sensitivity/resistance
  - Mutational hotspots in cancer
  - ++

Personal Cancer Genome Reporter (III)

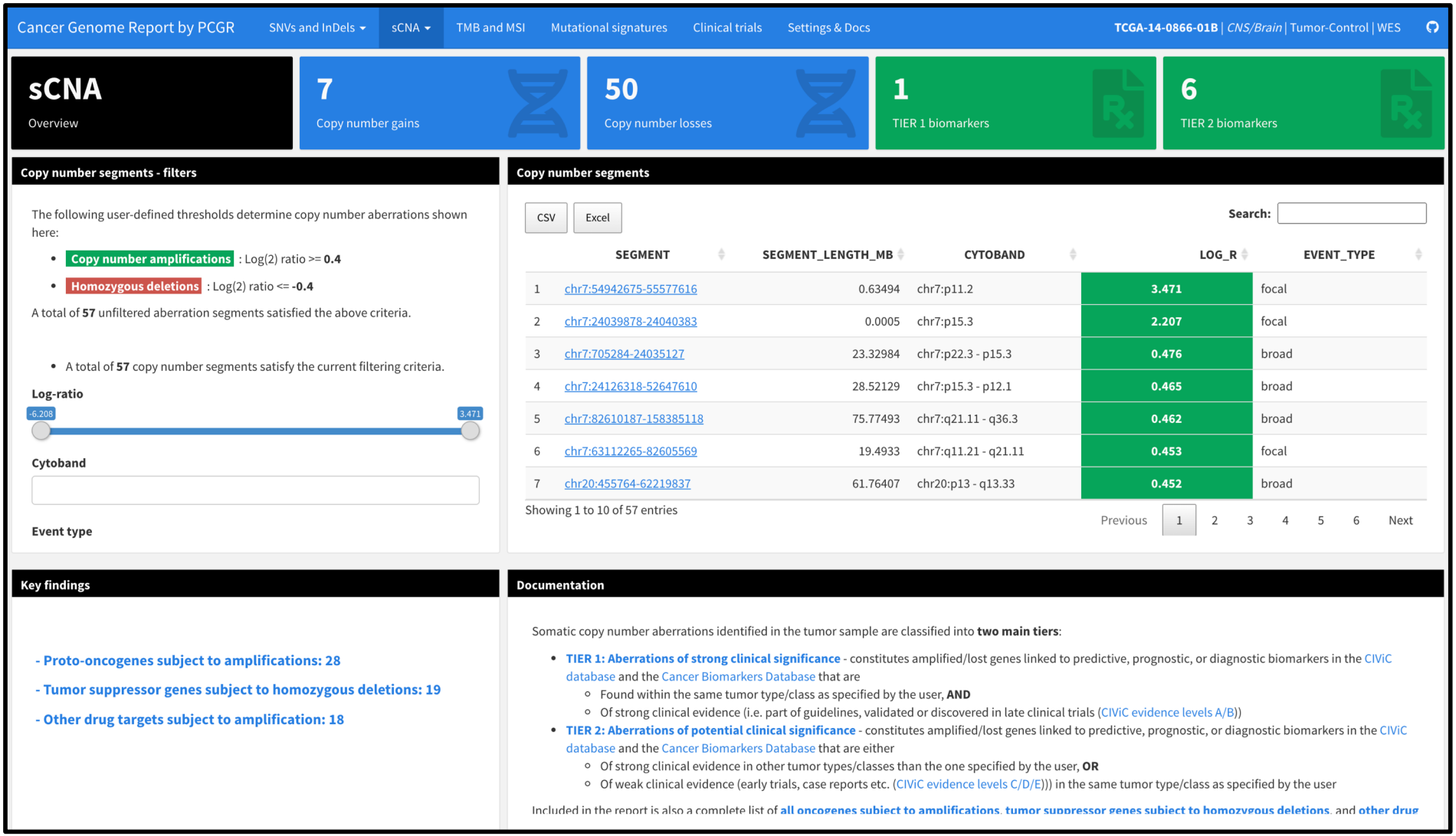


Meyerson et al. Nat Rev Genet 2010

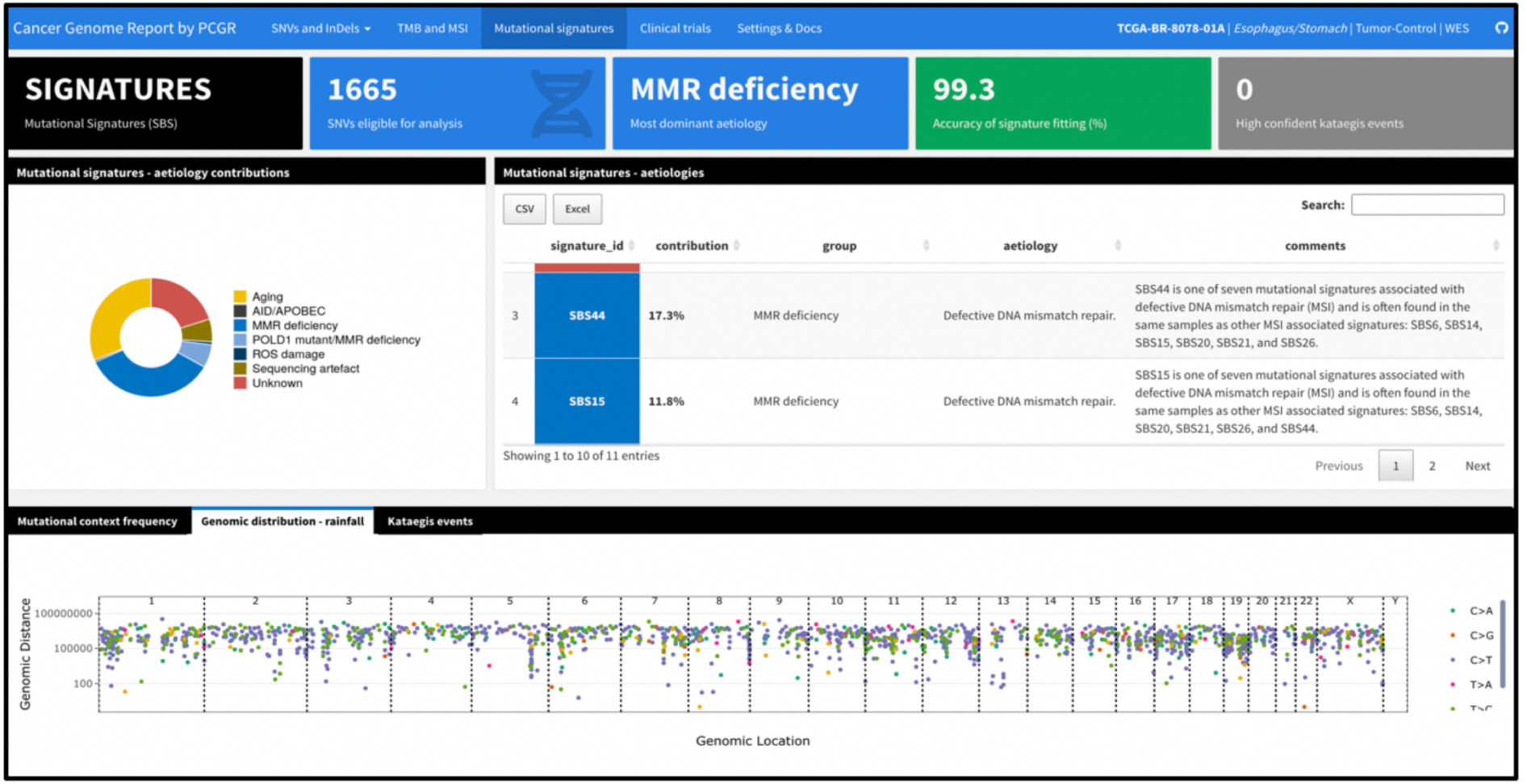
Personal Cancer Genome Reporter (IV)



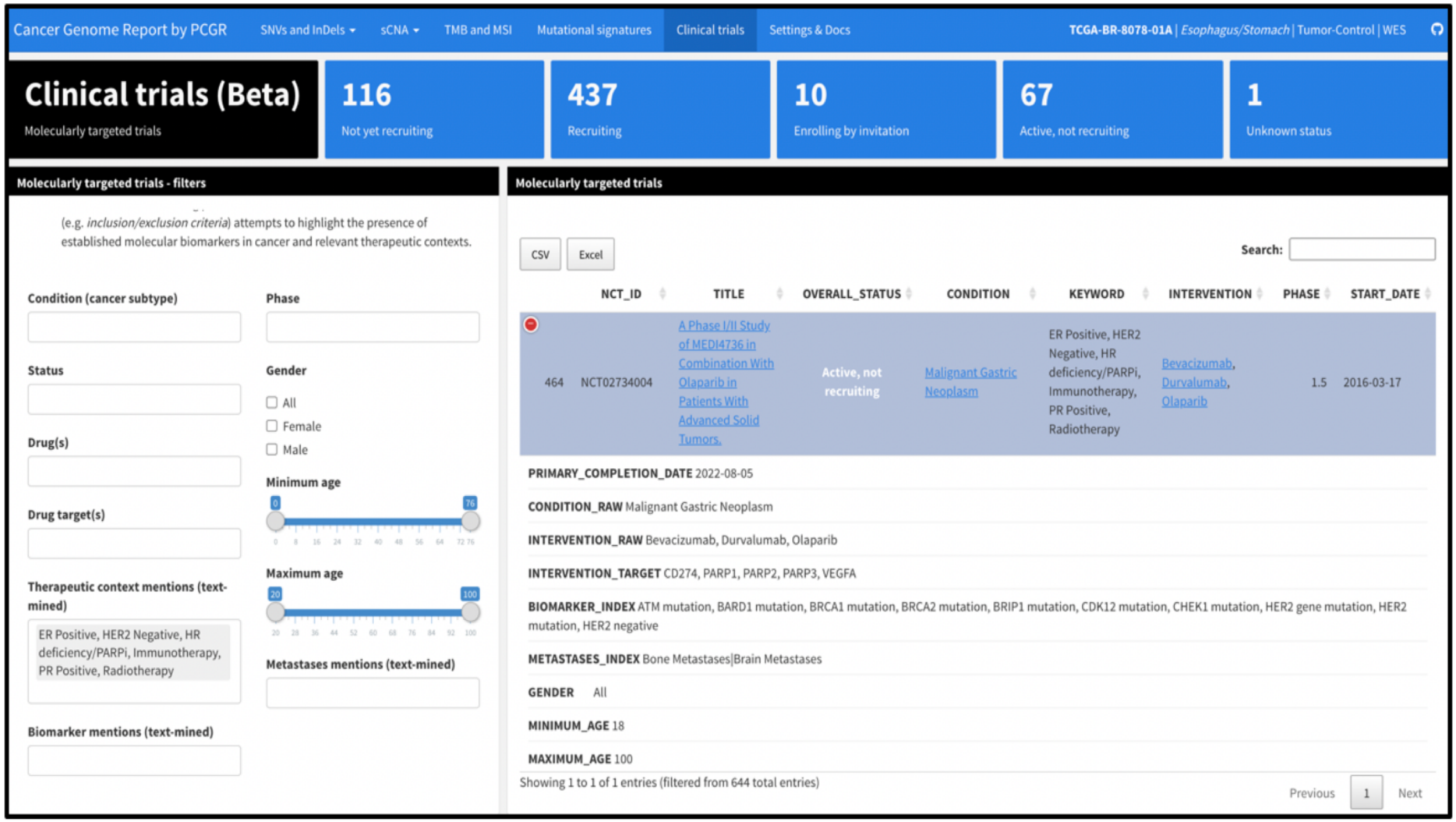
Personal Cancer Genome Reporter (IV)



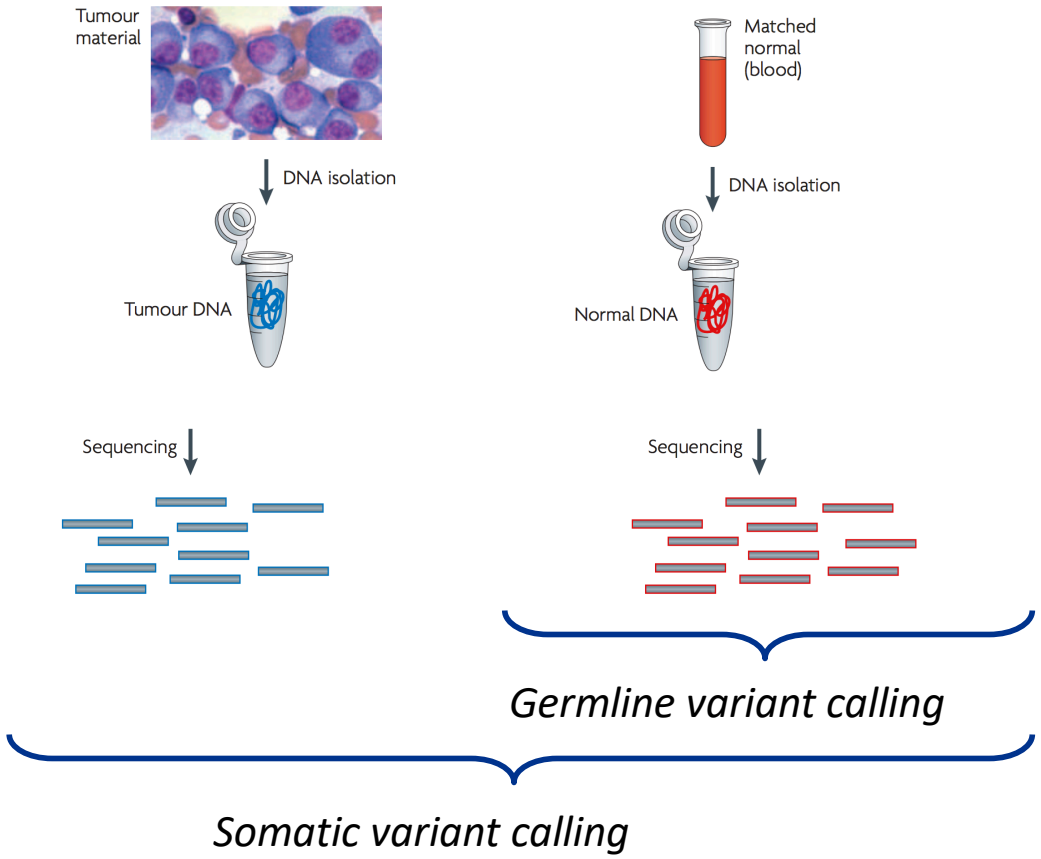
Personal Cancer Genome Reporter (IV)



Personal Cancer Genome Reporter (IV)

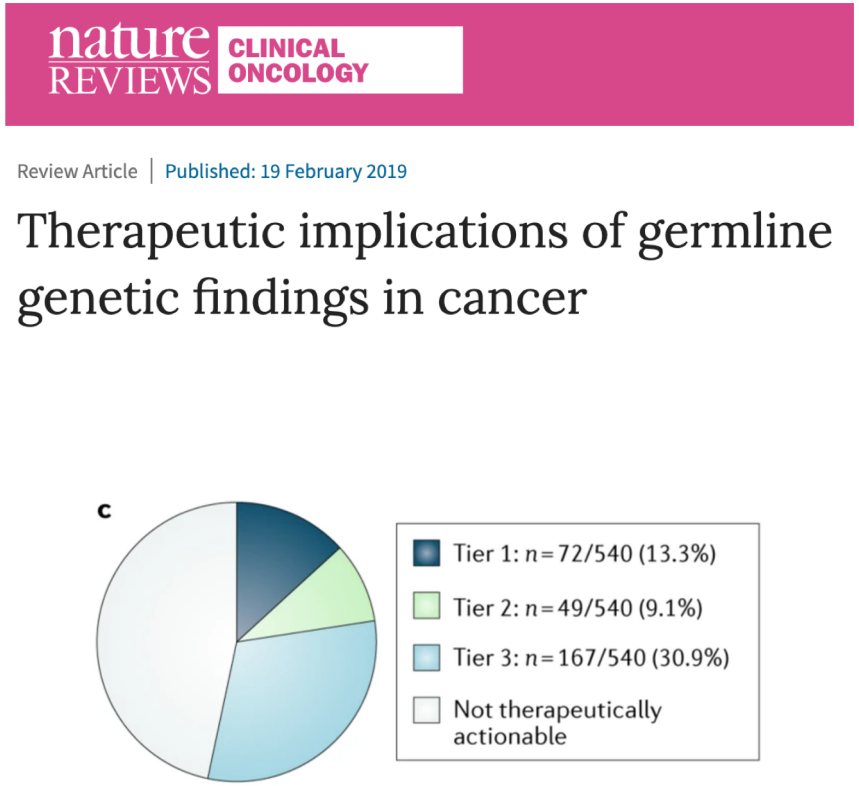


Cancer patients - germline background



Cancer predisposition interpretation

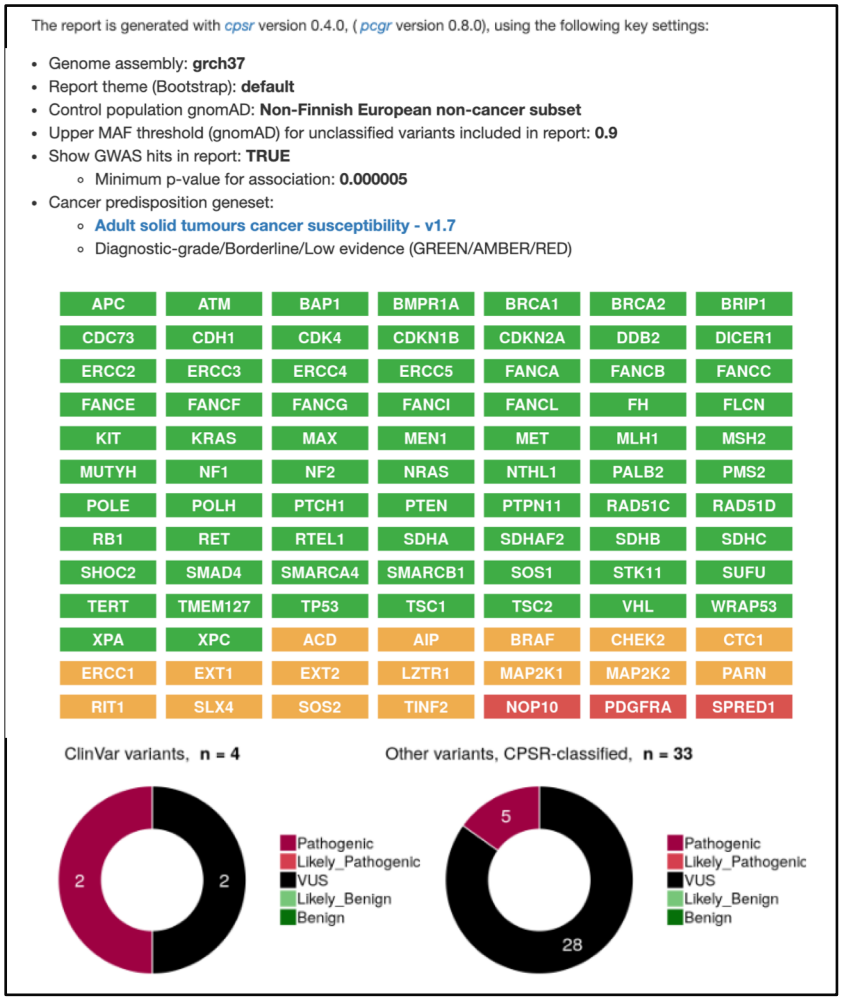
- Goal: Identify pathogenic variants (germline) conferring increased risk of tumor development
- Why important?
  - Implement surveillance and risk-reducing interventions
  - May impact type of surgery (radical /conservative)
  - Targeted therapy implications
    - BRCA (PARP)



Clinical actionability - TCGA

Cancer Predisposition Sequencing Reporter

- **CPSR:** Flexible reporting tool for interpretation of sequencing screens for cancer predisposition
- Which germline variants confer risk of tumor development? Tier structure
  - Pathogenic
  - Likely pathogenic
  - Unclassified variants
  - Likely Benign
  - Benign
- Automated pathogenicity classification
  - Predicted loss-of-function
  - population allele frequency
  - ++
- Incidental findings can also be reported

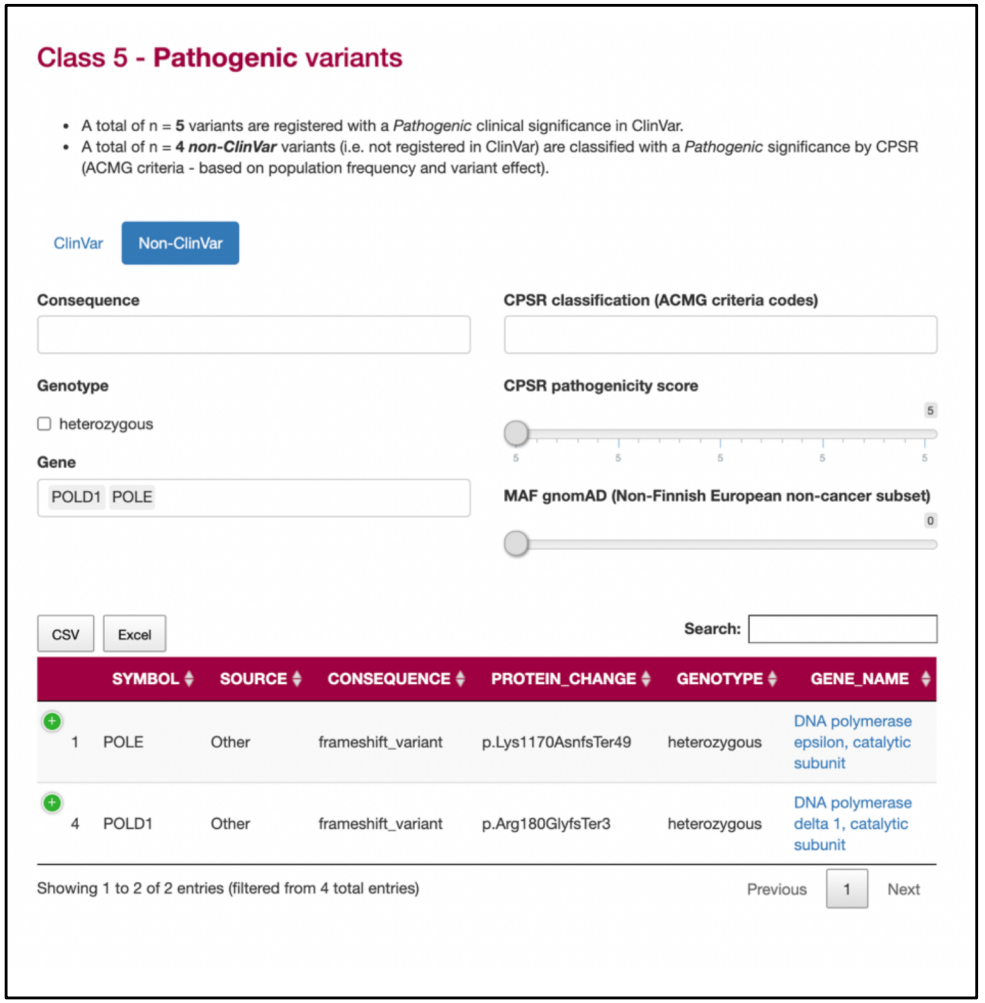


Nakken et al., Int J Cancer, 2021

https://github.com/sigven/cpsr

Cancer Predisposition Sequencing Reporter

- Flexible reporting tool for interpretation of sequencing screens for cancer predisposition
- Which germline variants confer risk of tumor development? Tier structure
  - Pathogenic
  - Likely pathogenic
  - Unclassified variants
  - Likely Benign
  - Benign
- Automated pathogenicity classification
  - Predicted loss-of-function
  - population allele frequency
  - ++
- Incidental findings can also be reported

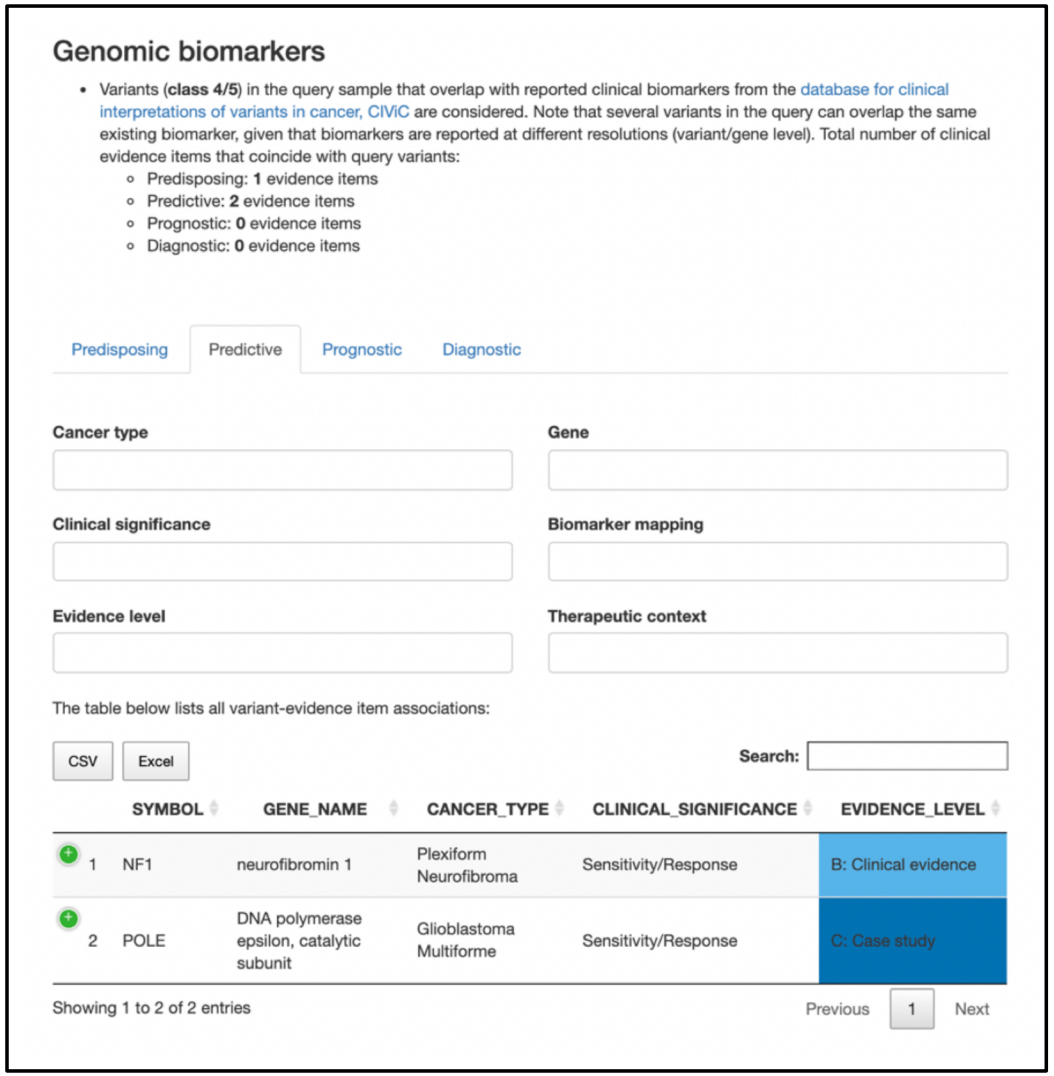


Nakken et al., Int J Cancer, 2021

https://github.com/sigven/cpsr

Cancer Predisposition Sequencing Reporter

- Flexible reporting tool for interpretation of sequencing screens for cancer predisposition
- Which germline variants confer risk of tumor development? Tier structure
  - Pathogenic
  - Likely pathogenic
  - Unclassified variants
  - Likely Benign
  - Benign
- Automated pathogenicity classification
  - Predicted loss-of-function
  - population allele frequency
  - ++
- Incidental findings can also be reported

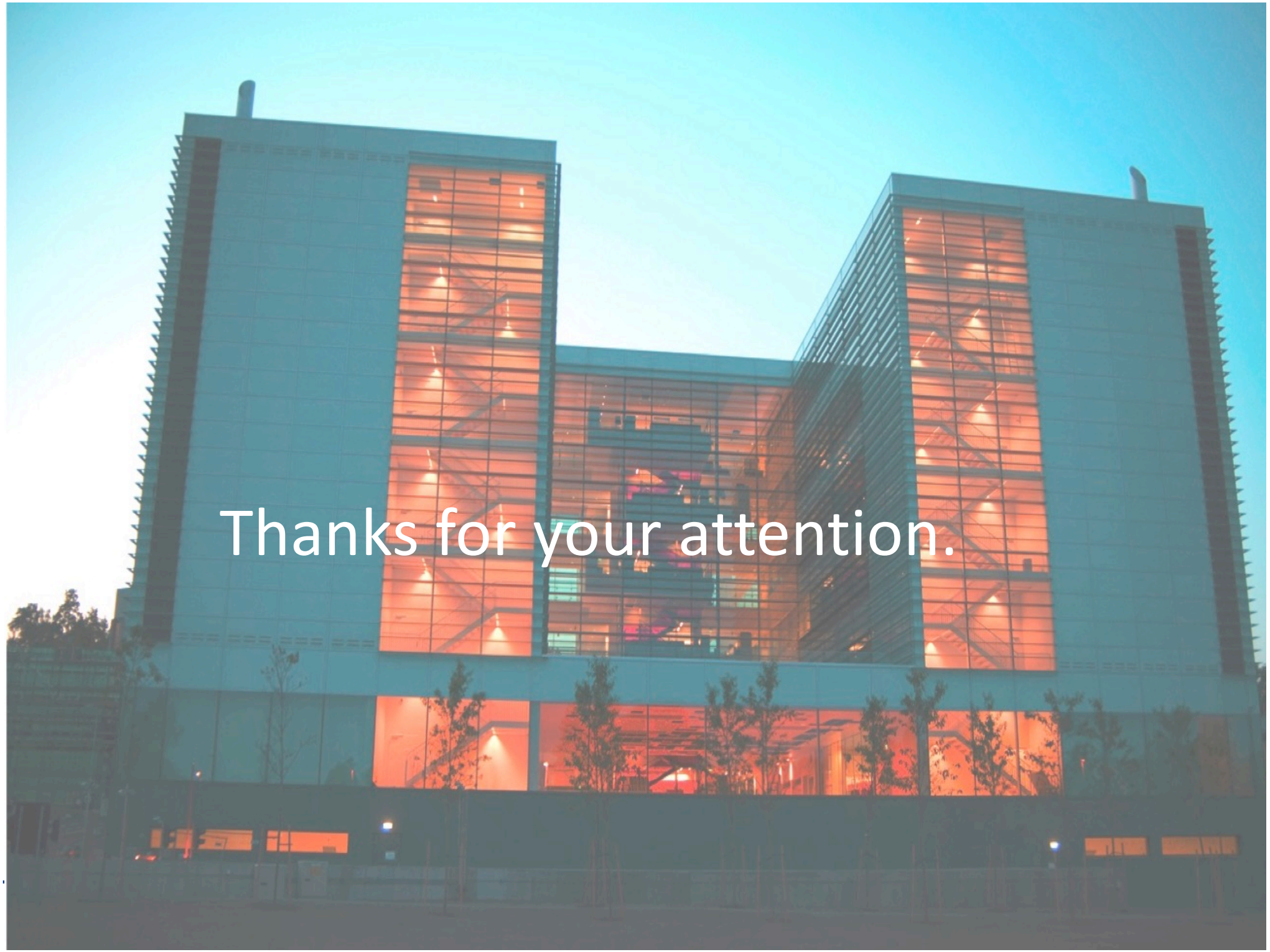


Nakken et al., Int J Cancer, 2021

https://github.com/sigven/cpsr

Variant interpretation in cancer: summary

- Comprehensive DNA variant interpretation is critical for implementation of precision cancer medicine
- Types of mutations, number of mutations, mutation locations – all may have therapeutic implications
- Variant consequences are transcript-specific
- A large number of resources have been erected to facilitate clinical interpretation of cancer genomes
- Variant prioritizaton: tier structure
- Interpretation of the germline background of cancer patients adds an additional dimension for clinical interpretation


Thanks for your attention.