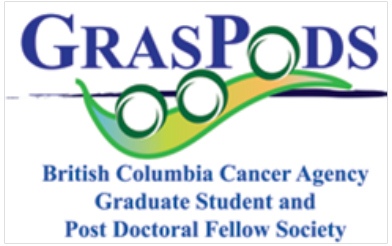
VARIANT CALLING IN CANCER GENOMES SEMINAR

Fong Chun Chan :: PhD Candidate :: Steidl and Shah Labs

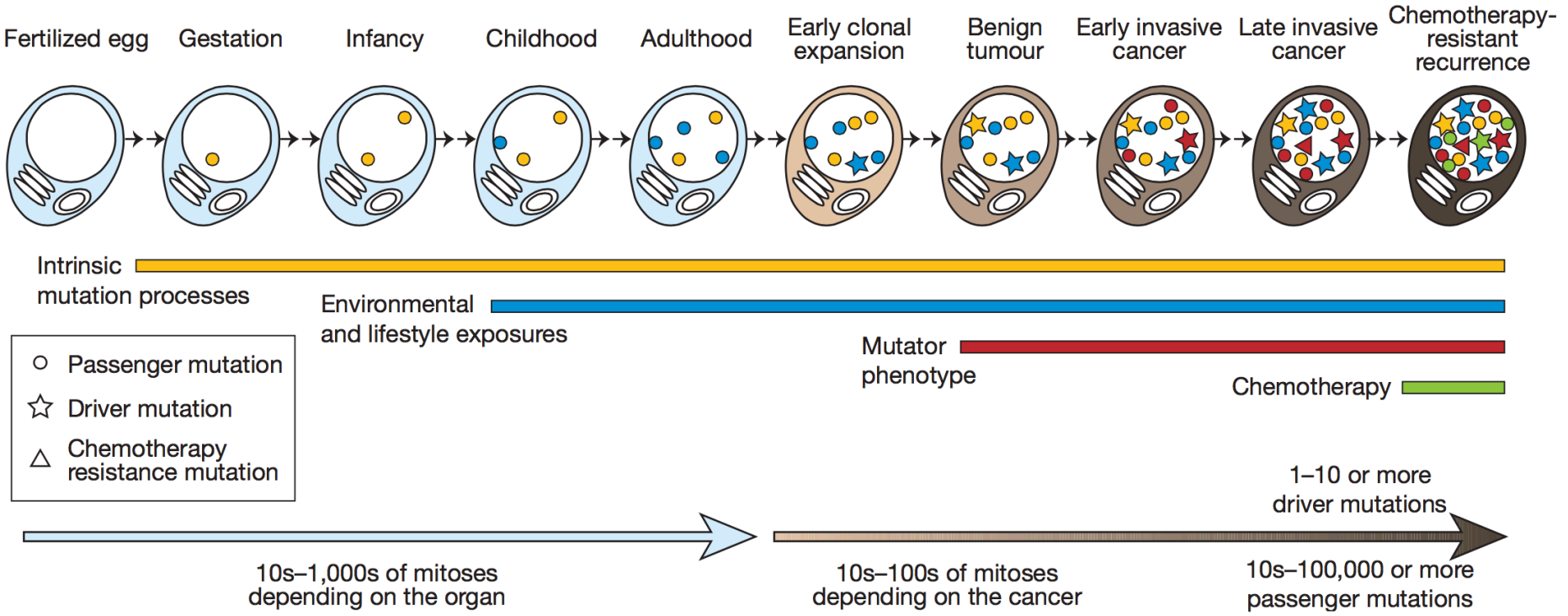
Sept 27, 2016

Graduate Program in
BIOINFORMATICS



BC Cancer Agency
CARE + RESEARCH
An agency of the Provincial Health Services Authority

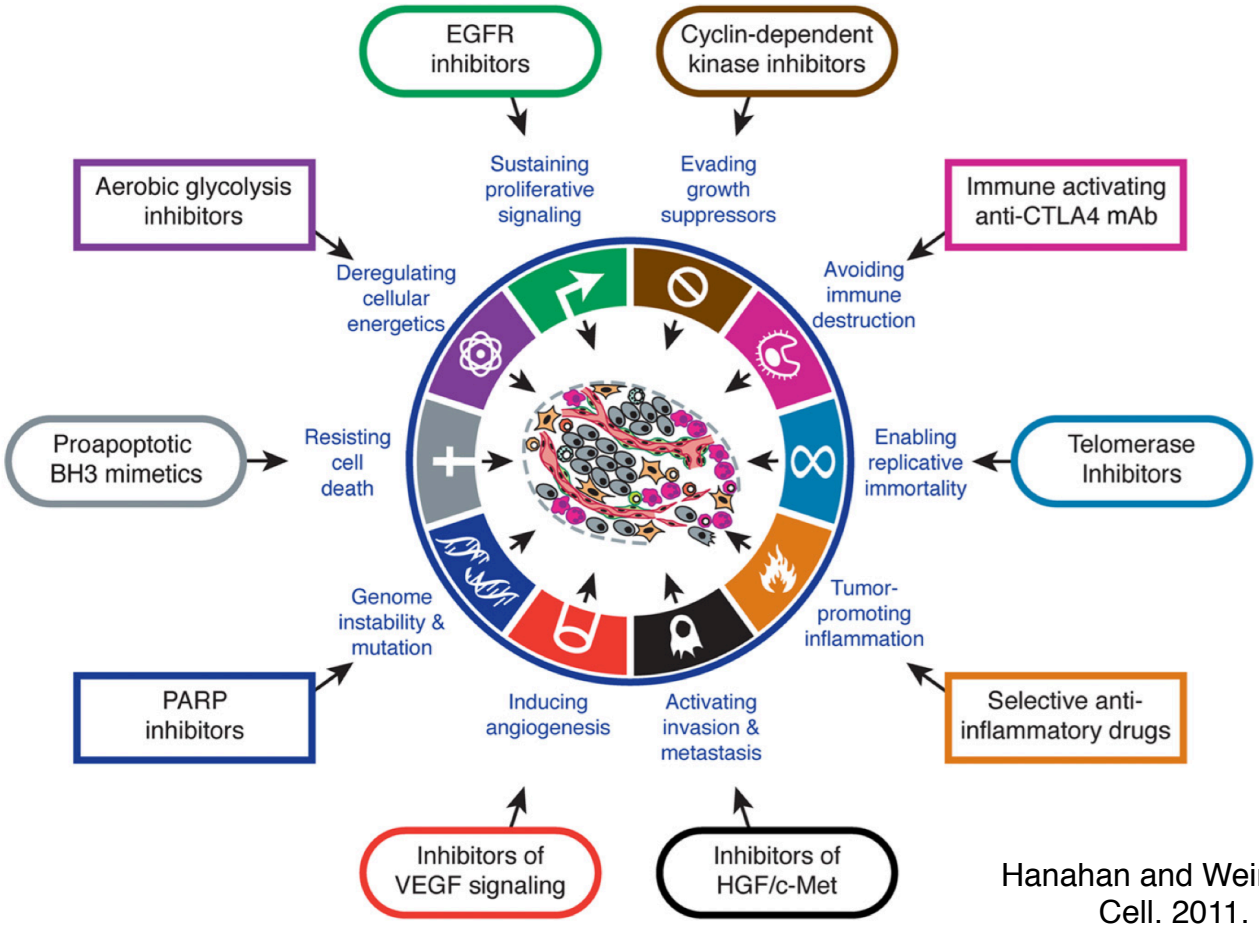
Cancer is a Genetic Disease



Stratton et al. Nature. 2009

- Somatic mutations are inherited by daughter cells

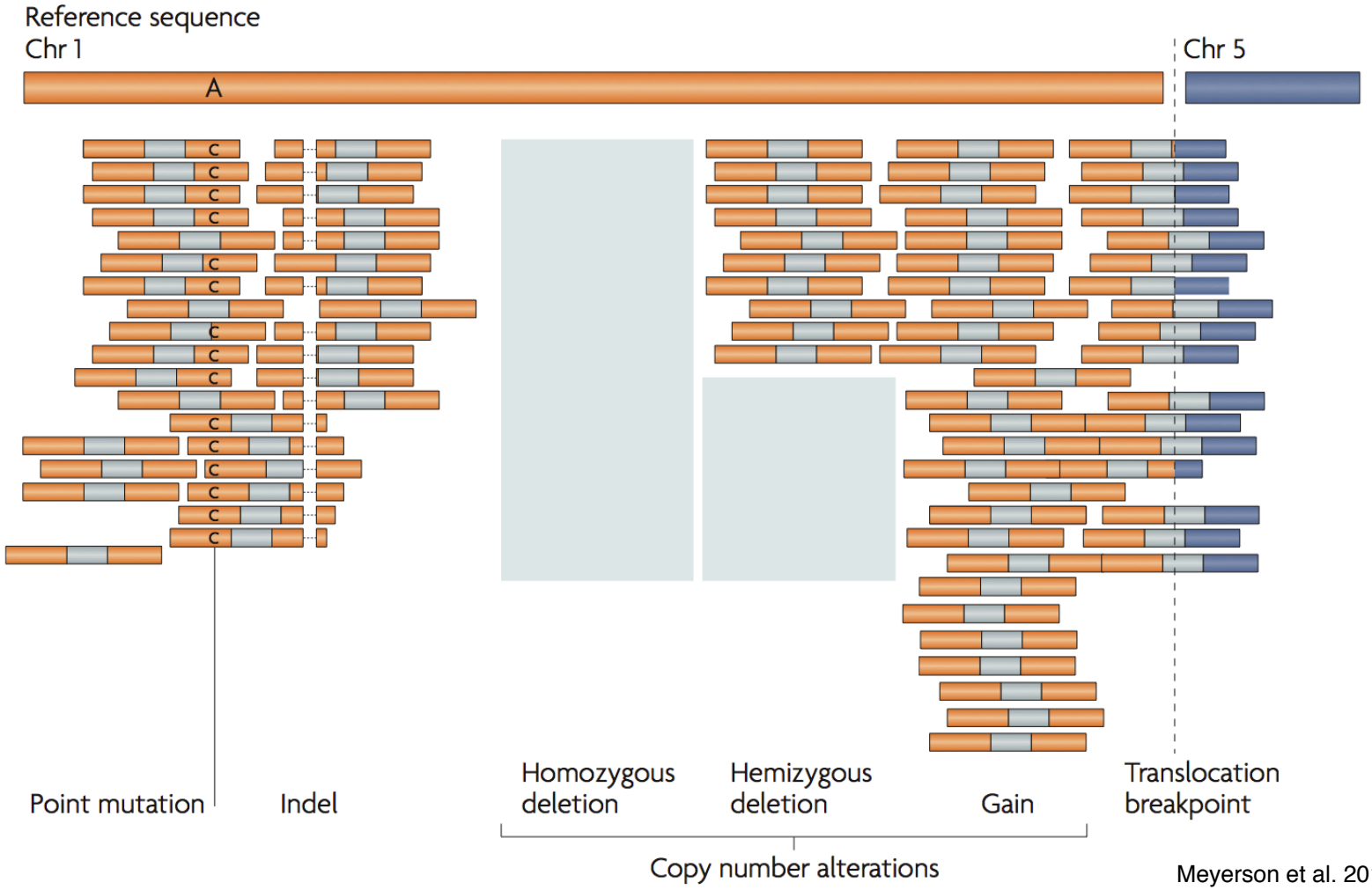
Cancer is a Genetic Disease



Hanahan and Weinberg. Cell. 2011.

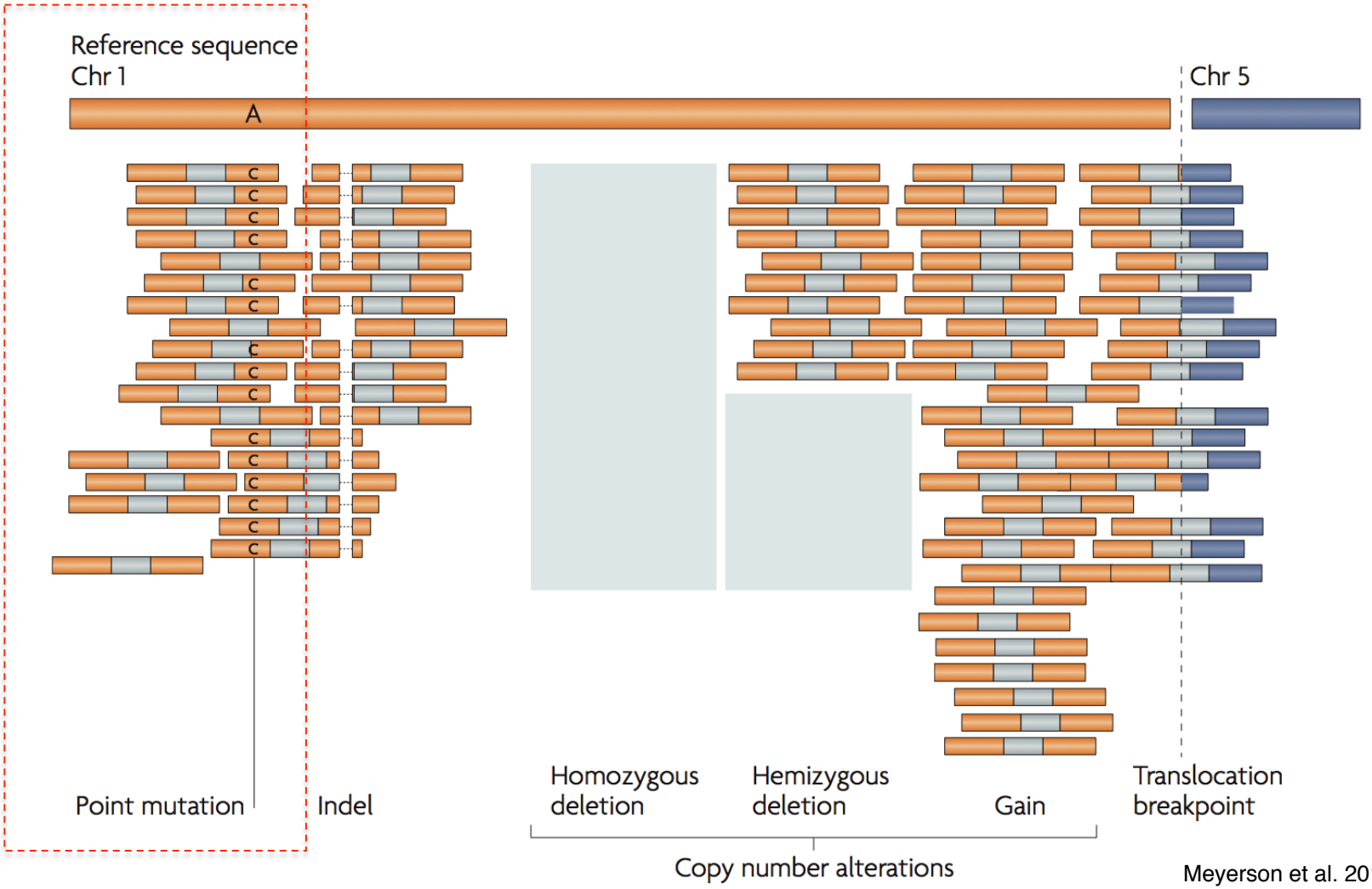
- Mutations are partially responsible for the hallmarks of cancer

High-Throughput Sequencing of Cancer Genomes



Meyerson et al. 2010

High-Throughput Sequencing of Cancer Genomes



Meyerson et al. 2010

Purpose of Today's Workshop

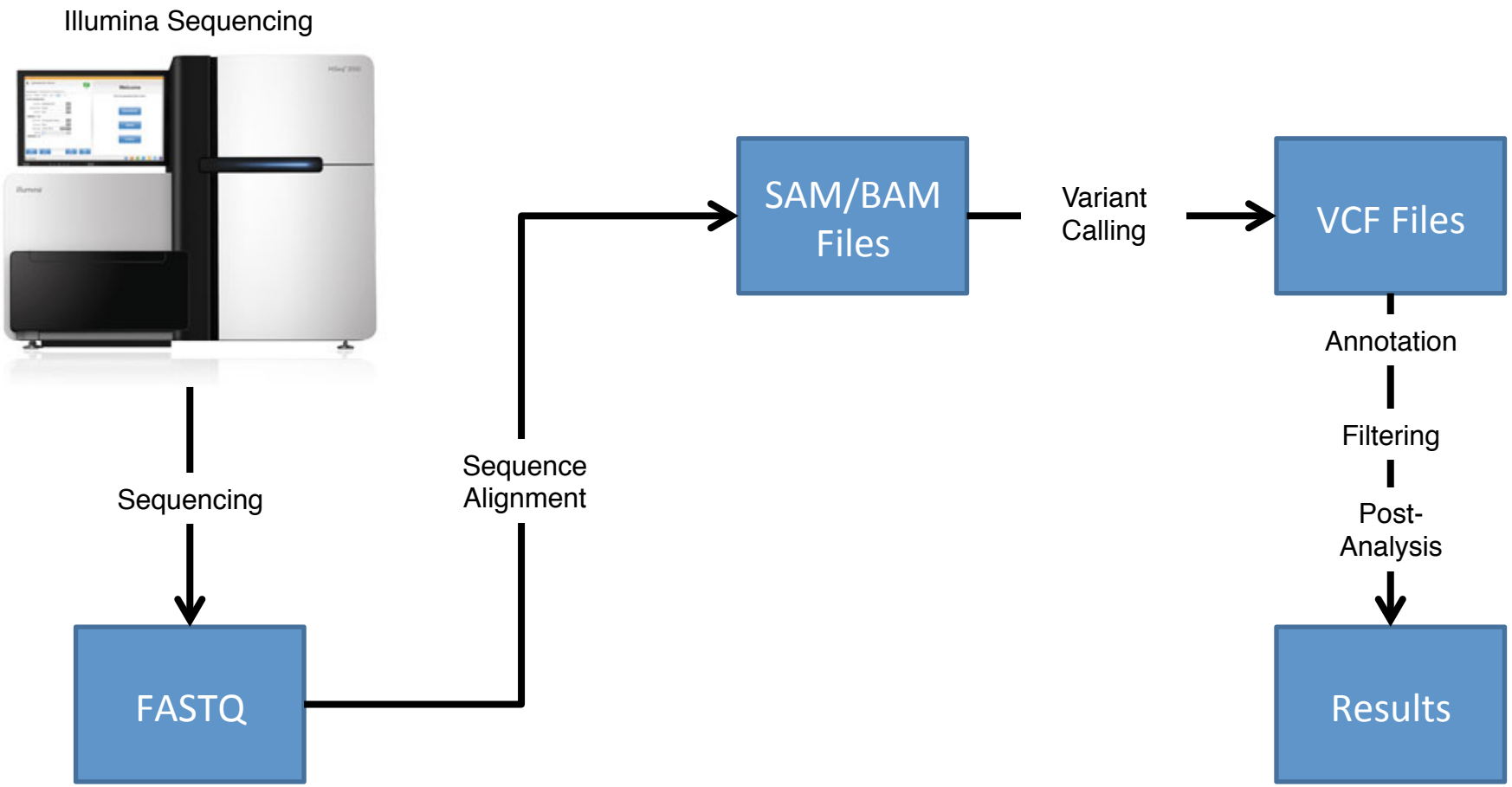
1. Understand the standard bioinformatics workflow for variant calling
   - Standard typical input/output formats

2. Understand the challenges when calling variants in cancer genomes

3. Available bioinformatic tools for variant calling

4. Analyzing single point mutation results

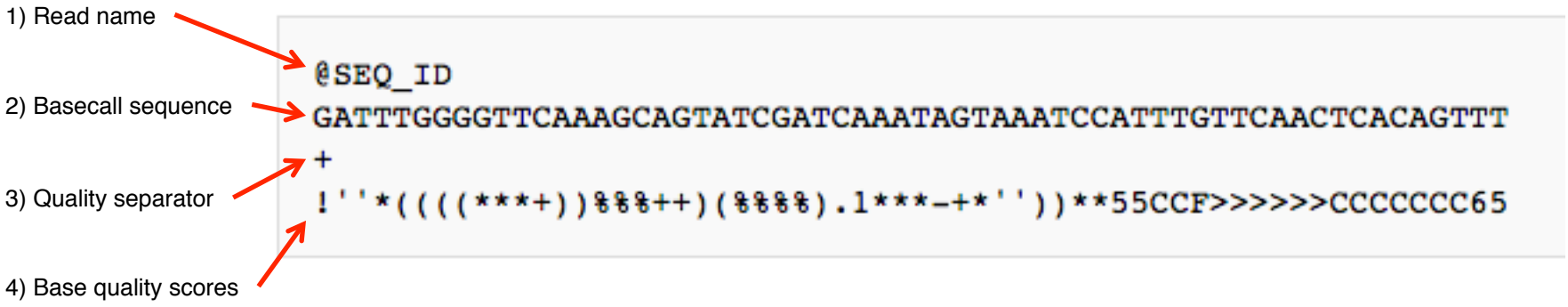
Bioinformatics Pipeline for Variant Calling



FASTQ File Format

- Text-based format for storing biological sequences

https://en.wikipedia.org/wiki/FASTQ_format



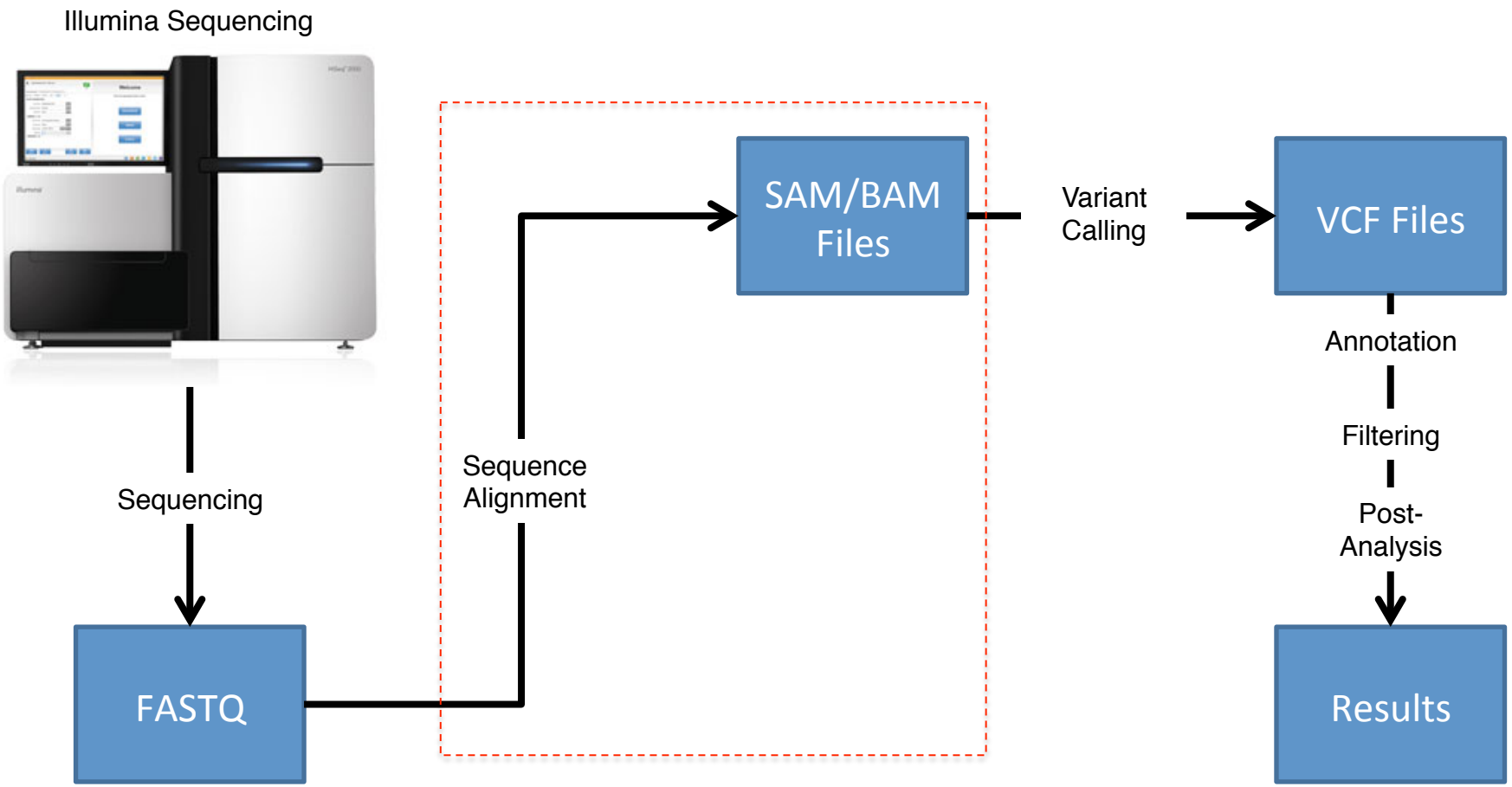
- FASTA is the same as FASTQ, but without quality scores

Base Quality Scores

- Each base quality ASCII value corresponds to a score.
  - Estimate probability of the base call being incorrect

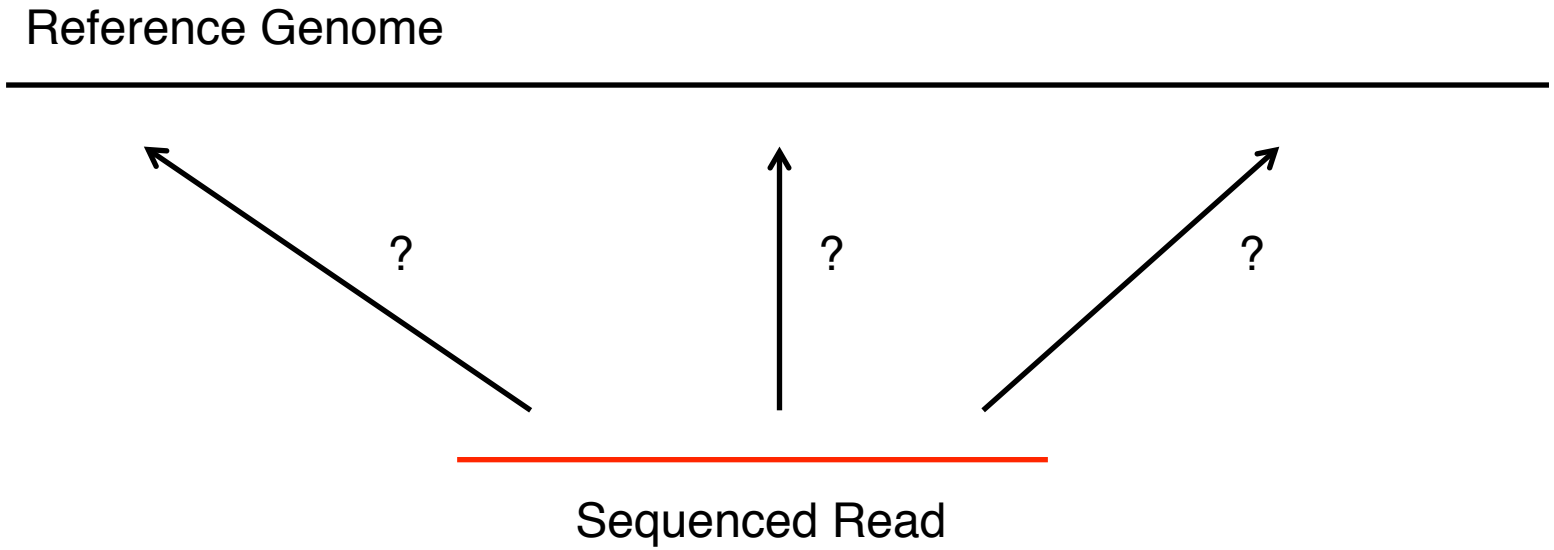
Base Quality	P_{error}(obs. Base)
3	50%
5	32%
10	10%
20	1%
30	0.1%
40	0.01%

Bioinformatics Pipeline for Variant Calling



Sequence Alignment

- Comparing sequence reads to a reference genome so that we can identify variants



Many Sequence Aligners

BIOINFORMATICS REVIEW

Vol. 28 no. 24 2012, pages 3169–3177
doi:10.1093/bioinformatics/bts605

Sequence analysis

Advance Access publication October 11, 2012

Tools for mapping high-throughput sequencing data

Nuno A. Fonseca*, Johan Rung, Alvis Brazma and John C. Marioni

EMBL Outstation, European Bioinformatics Institute (EBI), Hinxton, Cambridge CB10 ISD, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: A ubiquitous and fundamental step in high-throughput sequencing analysis is the alignment (mapping) of the generated reads to a reference sequence. To accomplish this task, numerous software tools have been proposed. Determining the mappers that are most suitable for a specific application is not trivial.

Results: This survey focuses on classifying mappers through a wide number of characteristics. The goal is to allow practitioners to compare the mappers more easily and find those that are most suitable for their specific problem.

Availability: A regularly updated compendium of mappers can be found at http://wwwdev.ebi.ac.uk/fg/hts_mappers/.

Contact: nf@ebi.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 2, 2012; revised on October 4, 2012; accepted on October 8, 2012

development of mappers that exploit read pairing information (Langmead *et al.*, 2009; Li *et al.*, 2008a; Ning *et al.*, 2001). Furthermore, the appearance of novel protocols may result in specific biases (Li *et al.*, 2008a; Malhis and Jones, 2010; Ondov *et al.*, 2008). One consequence of the increasing number of mappers is that making a suitable choice for a specific application is not easy. Resources such as the SeqAnswers forum Wiki pages (Li *et al.*, 2012) have collated information about different mappers, such as the operating system supported and different technologies that the mappers have been designed to handle. However, information about other equally important features/characteristics of the mappers is difficult to find, being still scattered through publications, source code (when available), manuals and other documentation.

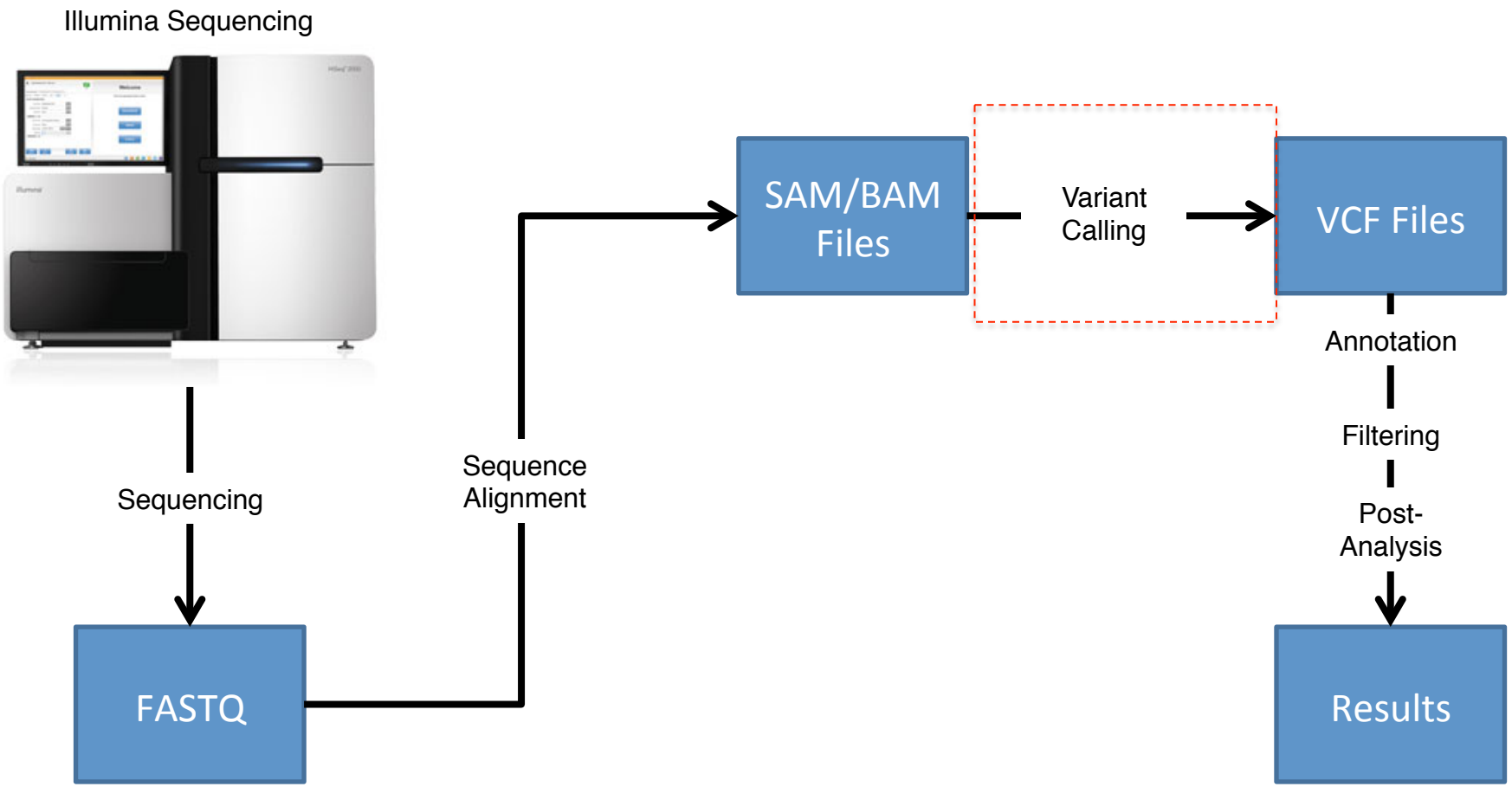
This survey aims to help overcome these challenges by allowing practitioners to compare mappers more easily and, thus, find those that are most suitable for their specific problem. It does not evaluate mappers in terms of their accuracy, but instead it presents an overview of their characteristics. It complements previ-

Sequence Alignment Map (SAM) Format

- De facto standard for storing sequence alignment information

- Binary form of SAM is BAM (Binary Alignment Format)

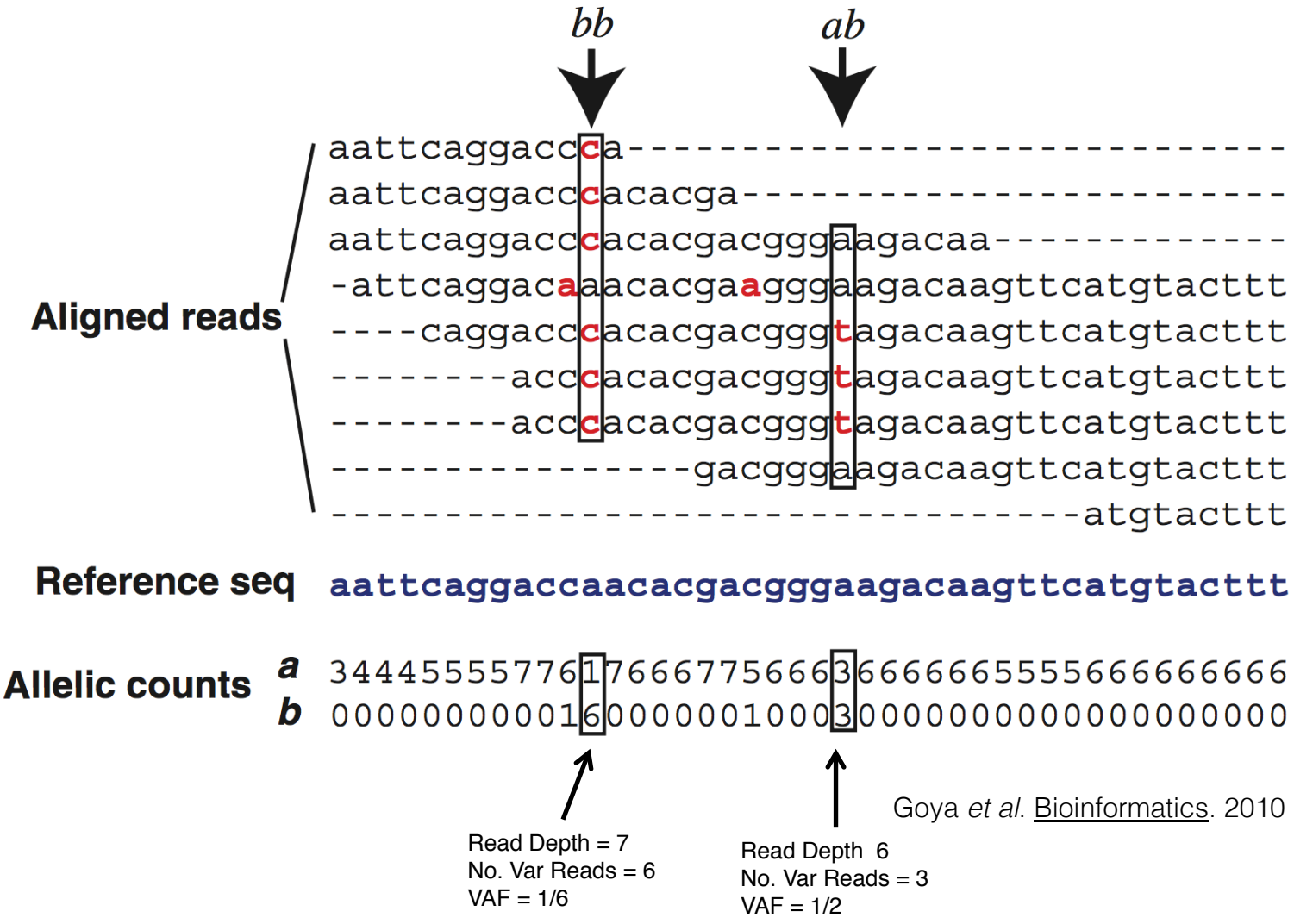
Bioinformatics Pipeline for Variant Calling



Variant Calling Nomenclature

- Read depth
  - Total sequencing number of reads covering a position
- Reference/Variant reads
  - Number of sequencing reads that support the reference/variant allele at a given position
- Variant Allelic Fraction (VAF)
  - Fraction of sequencing reads that support the variant allele at a given position

Variant Calling Nomenclature



Goya *et al*. Bioinformatics. 2010

Variant Calling in "Normal" Genomes

- In principle, variant calling in a "normal" human genome is "easy":
  - Still need to account for technical challenges (e.g. poor quality mappings, base quality, etc)

- Expected VAF:
  - Heterozygous mutations (AB) ~ 0.5 VAF
  - Homozygous mutations (BB) ~ 1 VAF

- Can be modeled with a binomial test

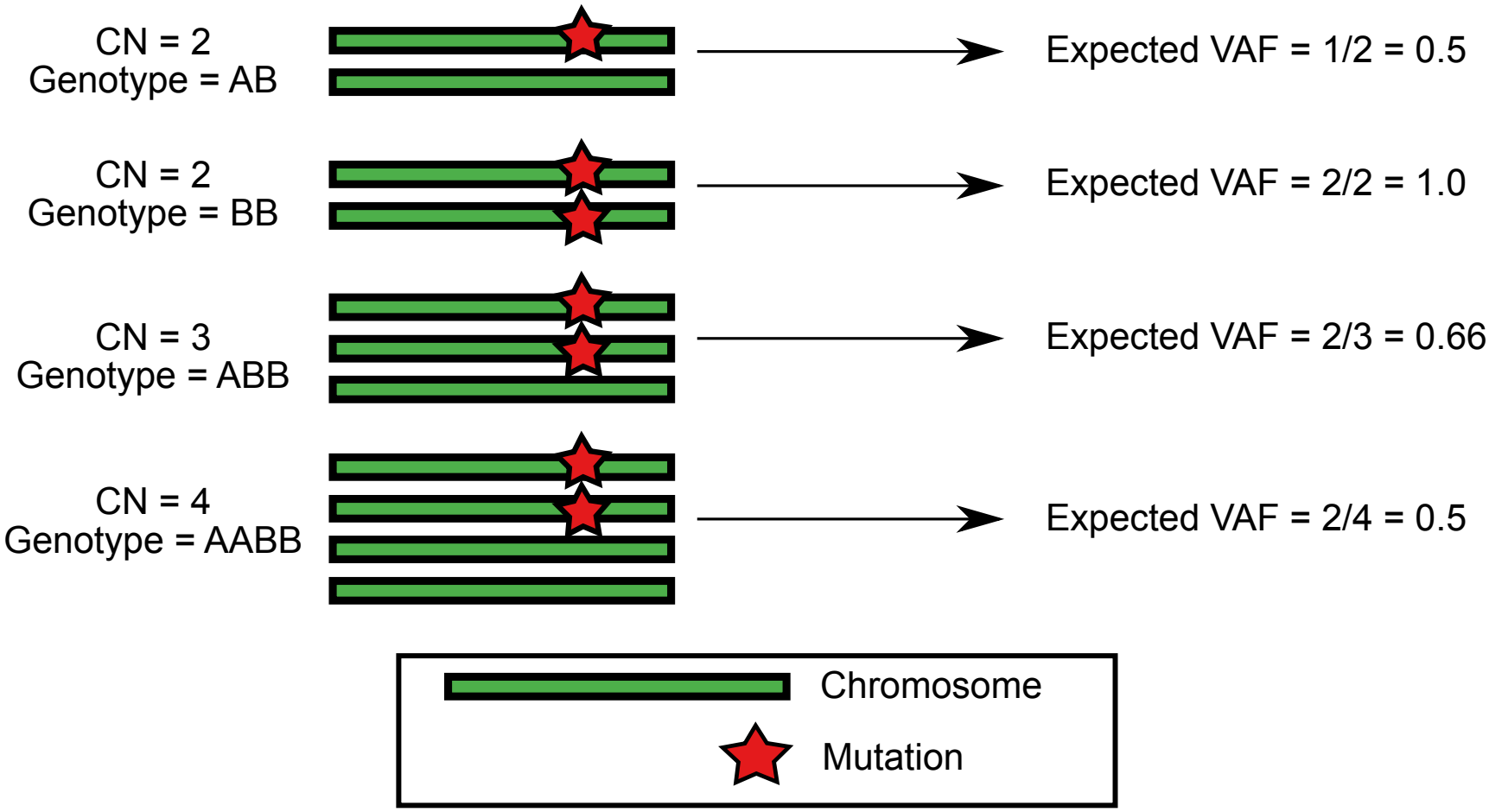
Challenges in Variant Calling in Cancer Genomes

1. Genomic instability

2. Population Structure
   - Intra-tumor heterogeneity
   - Tumor/Normal admixture

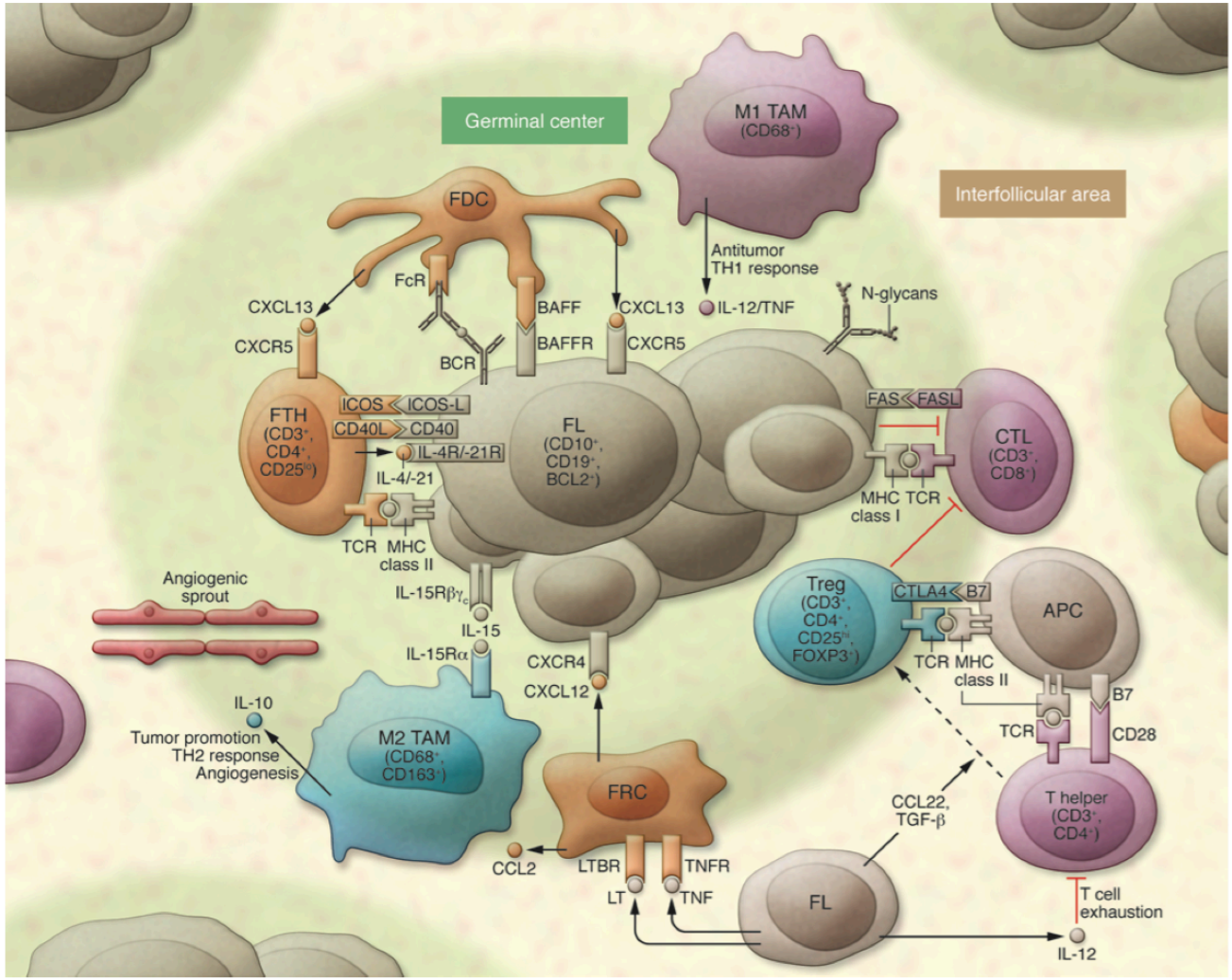
3. Technical Challenges
   - Alignment artifacts
   - Not only restricted to cancer genomes

Challenges in Variant Calling in Cancer Genomes: Genomic Instability



Slide courtesy of Andrew Roth

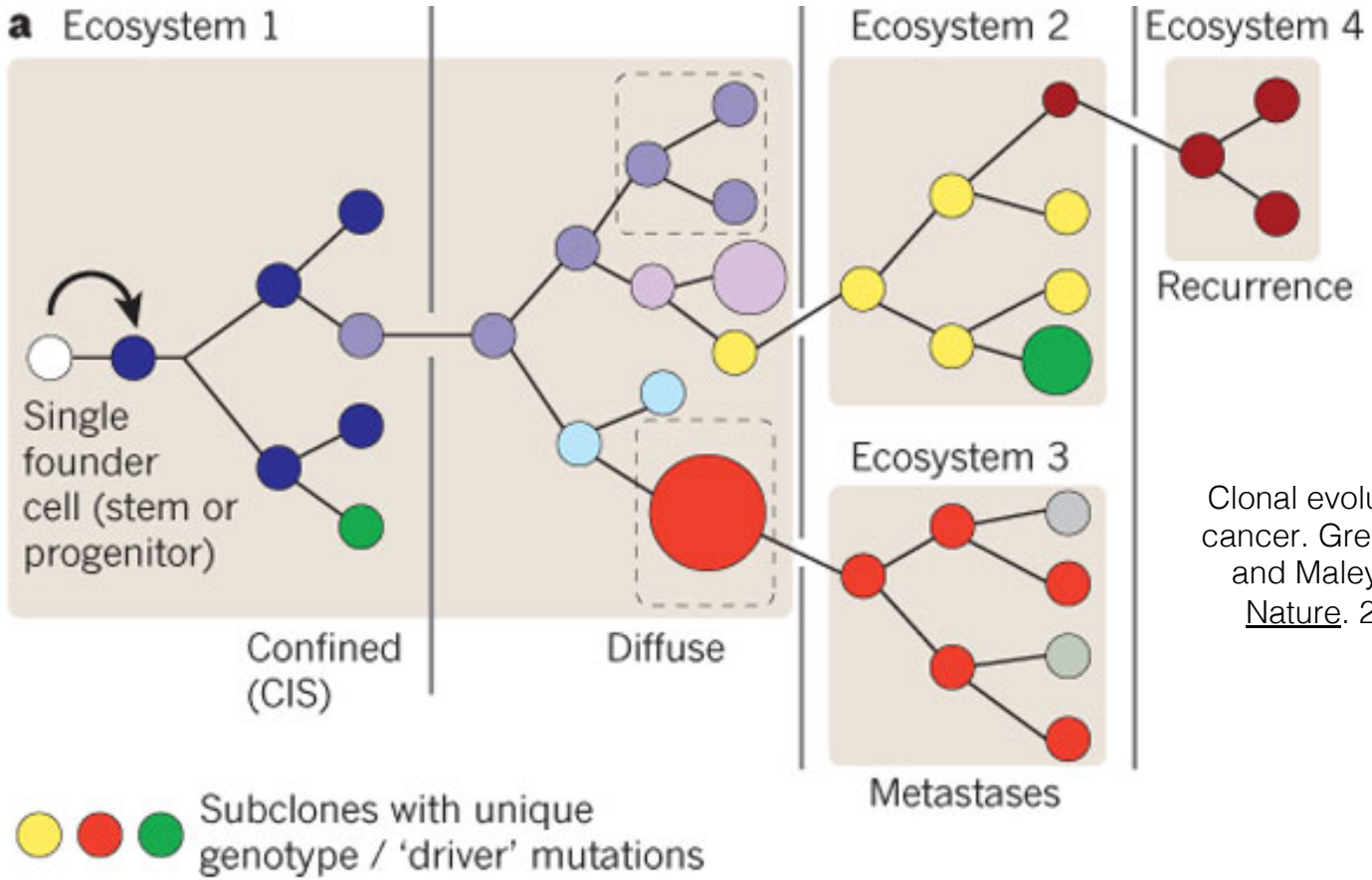
Challenges in Variant Calling in Cancer Genomes: Population Structure – Tumor Microenvironment



Kridel et al. 2012. JCI

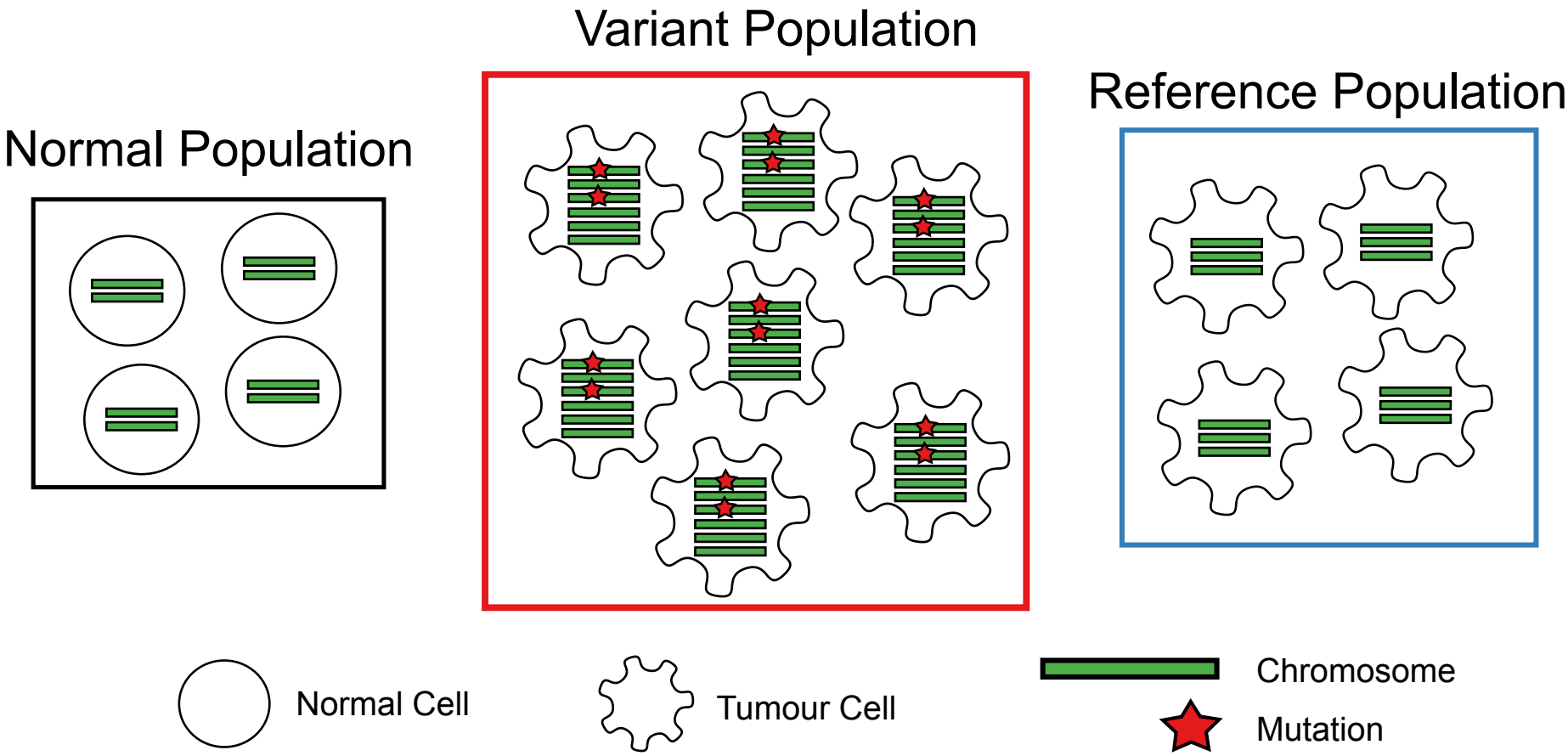
Challenges in Variant Calling in Cancer Genomes: Population Structure – Intratumor Heterogeneity

- Cancer development is an evolutionary process
- Consequence of such a process is intratumor heterogeneity



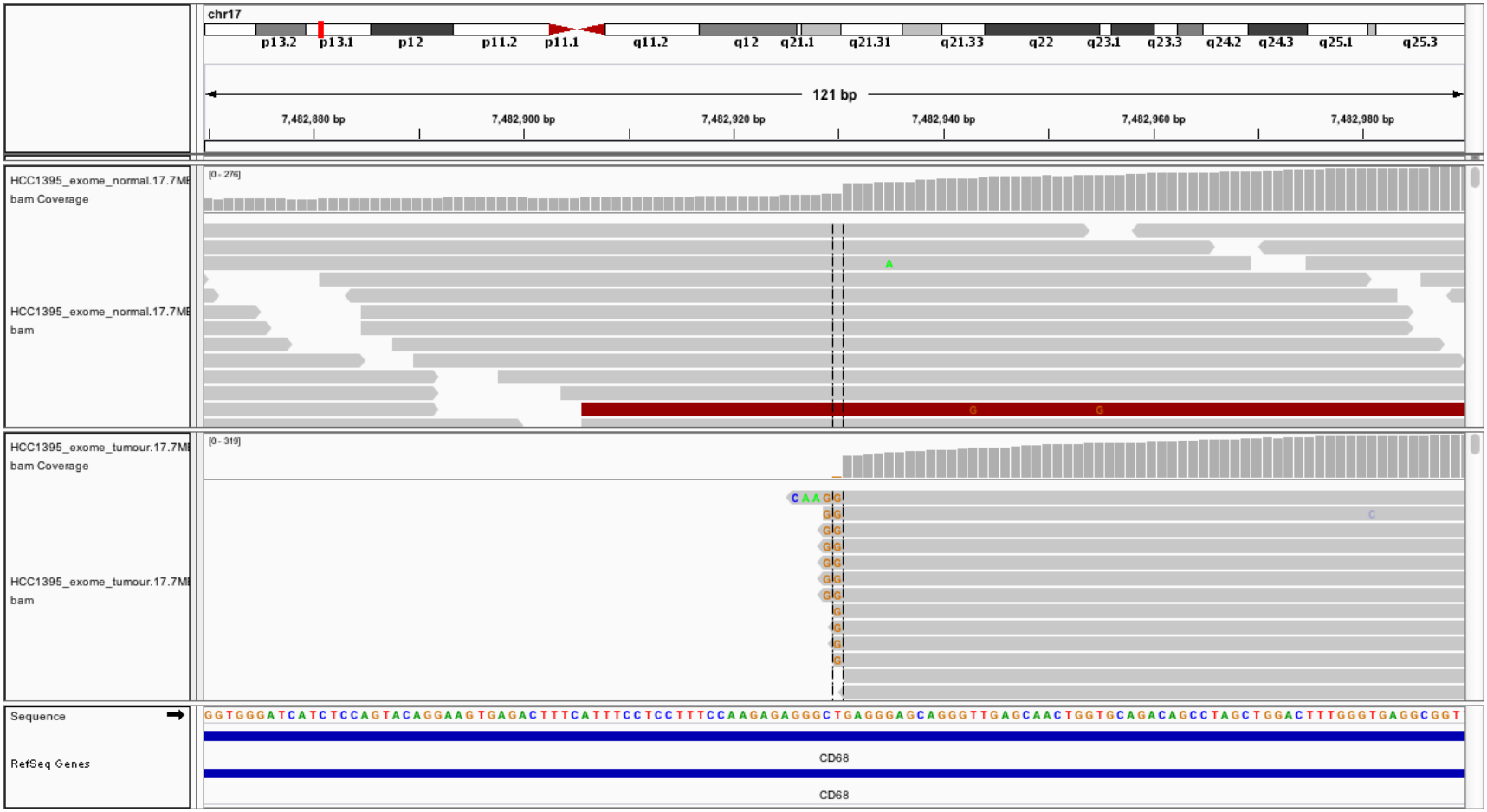
Clonal evolution in cancer. Greaves M and Maley CC. Nature. 2012

Challenges in Variant Calling in Cancer Genomes: Population Structure

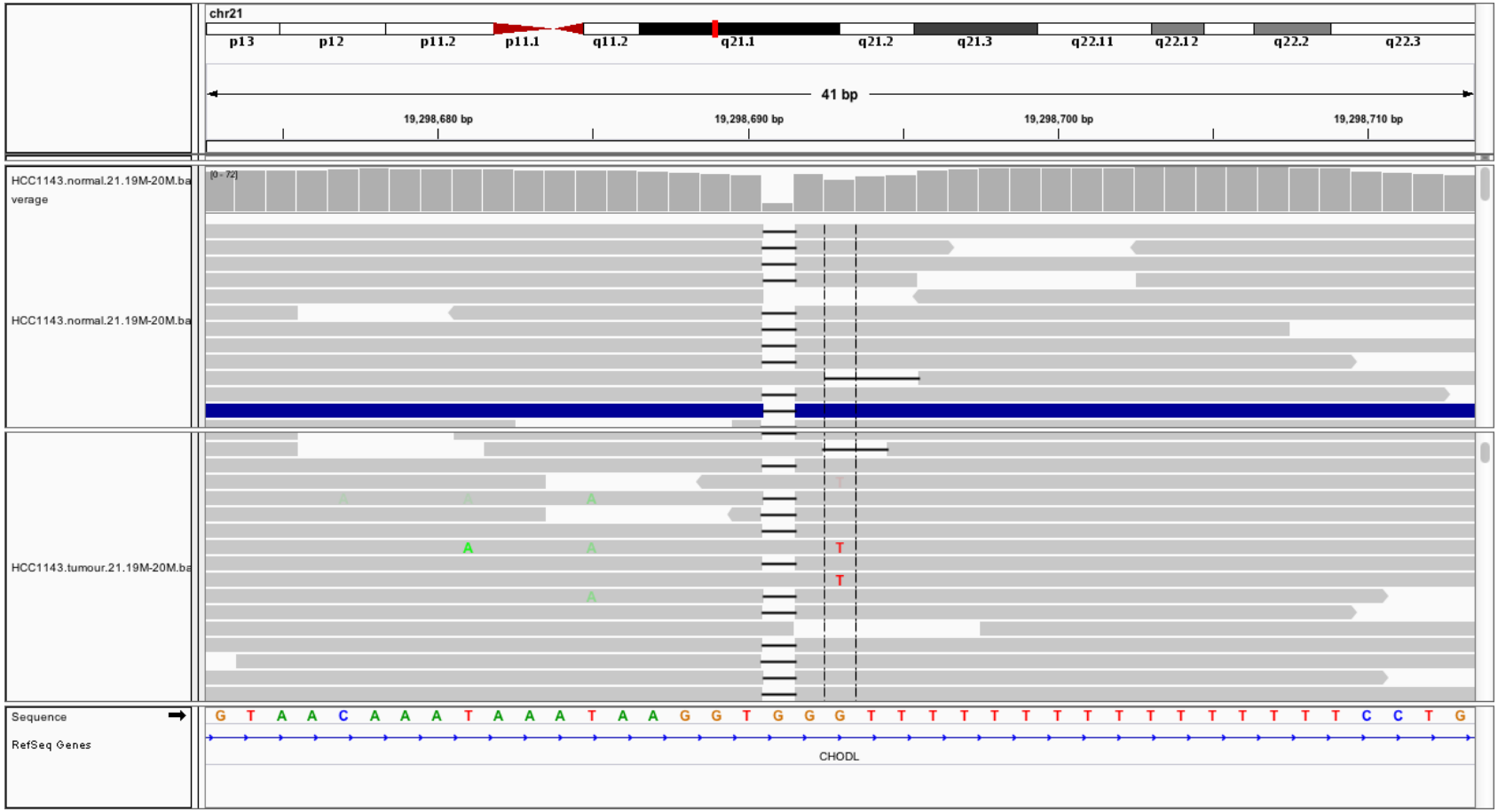


Slide courtesy of Andrew Roth

Sequencing Technical Artifacts



Sequencing Technical Artifacts



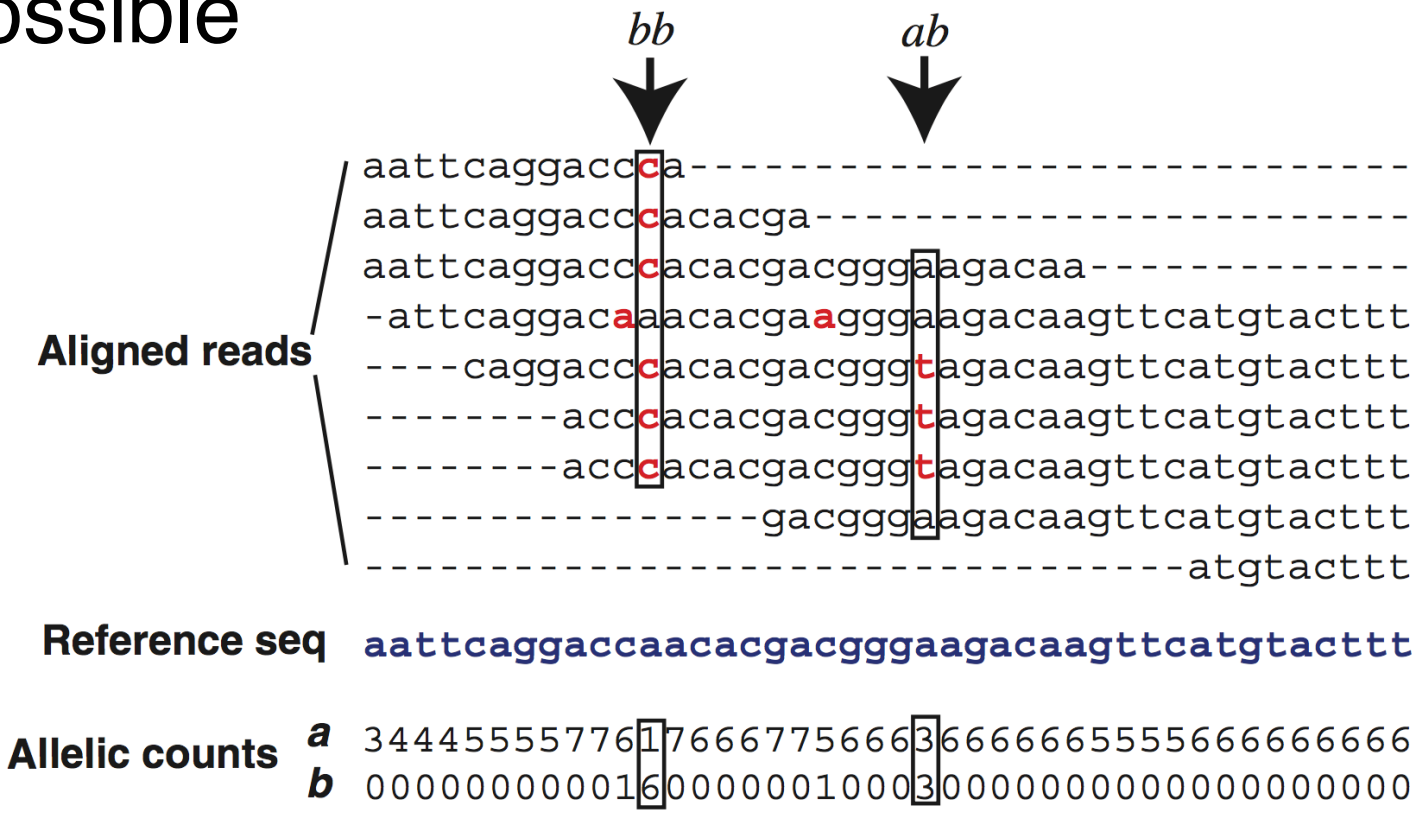
Best Practices for Variant Calling in Cancer Genomes

- One should **NOT** use germline variant callers for somatic cancer genomes (e.g. samtools, GATK)
  - Using a somatic variant callers that can account for genomic instability and population structure

- Choose a variant caller that can account for technical artifacts

Best Practices for Variant Calling in Cancer Genomes

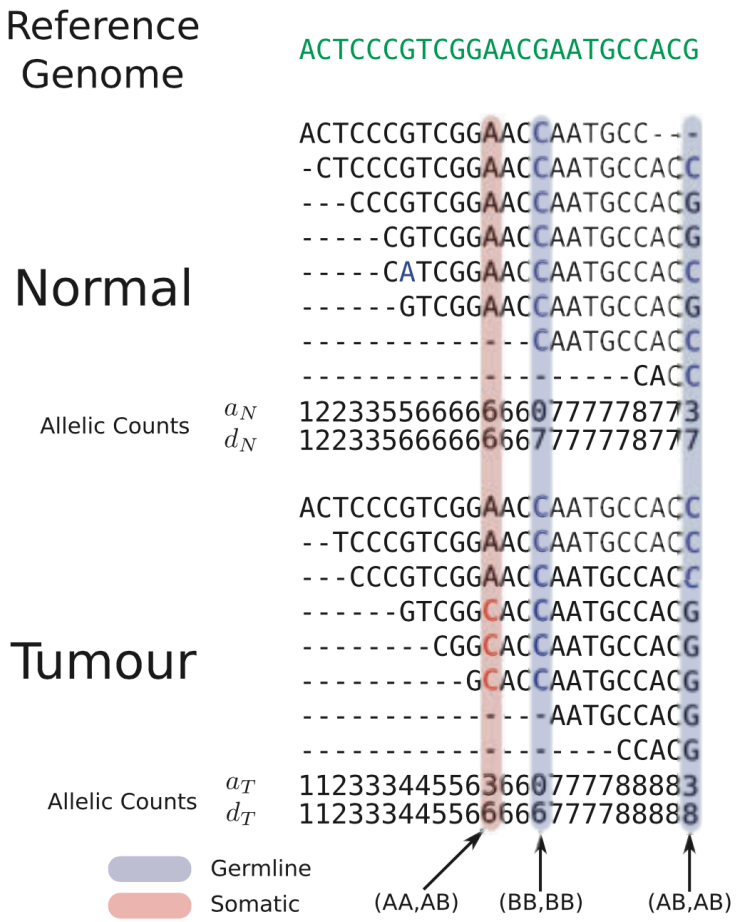
- Use paired Tumor-Normal samples when possible



Goya *et al*. Bioinformatics. 2010

Best Practices for Variant Calling in Cancer Genomes

- Use paired Tumor-Normal samples when possible



Best Practices for Variant Calling in Cancer Genomes

- Use DNA and not RNA when possible.

- Calling variants in RNA can be more challenging:
  - Non-uniformity in sequence coverage due to expression levels of the gene
  - Allelic specific expression makes genotype inference difficult (e.g. homozygous mutation or allele-specific expression?)

Available Bioinformatic Variant Callers

Question: Best Software For Detection Of Somatic Mutations From Matched Tumor:Normal Ngs Data

14

Hi all,

I know of various software solutions for this problem e.g. Somatic Sniper, JointSNVMix, etc., but having never used any of them, I am curious if any comparisons of their performance exist (peer-reviewed or otherwise).

Thanks in advance.

snv | variant | mutation | somatic | variant-calling • 23k views

4.5 years ago by Travis • **2.6k** USA

ADD COMMENT • link • Not following

modified 22 months ago by Malachi Griffith ♦ **14k** • written 4.5 years ago by Travis • **2.6k**

39

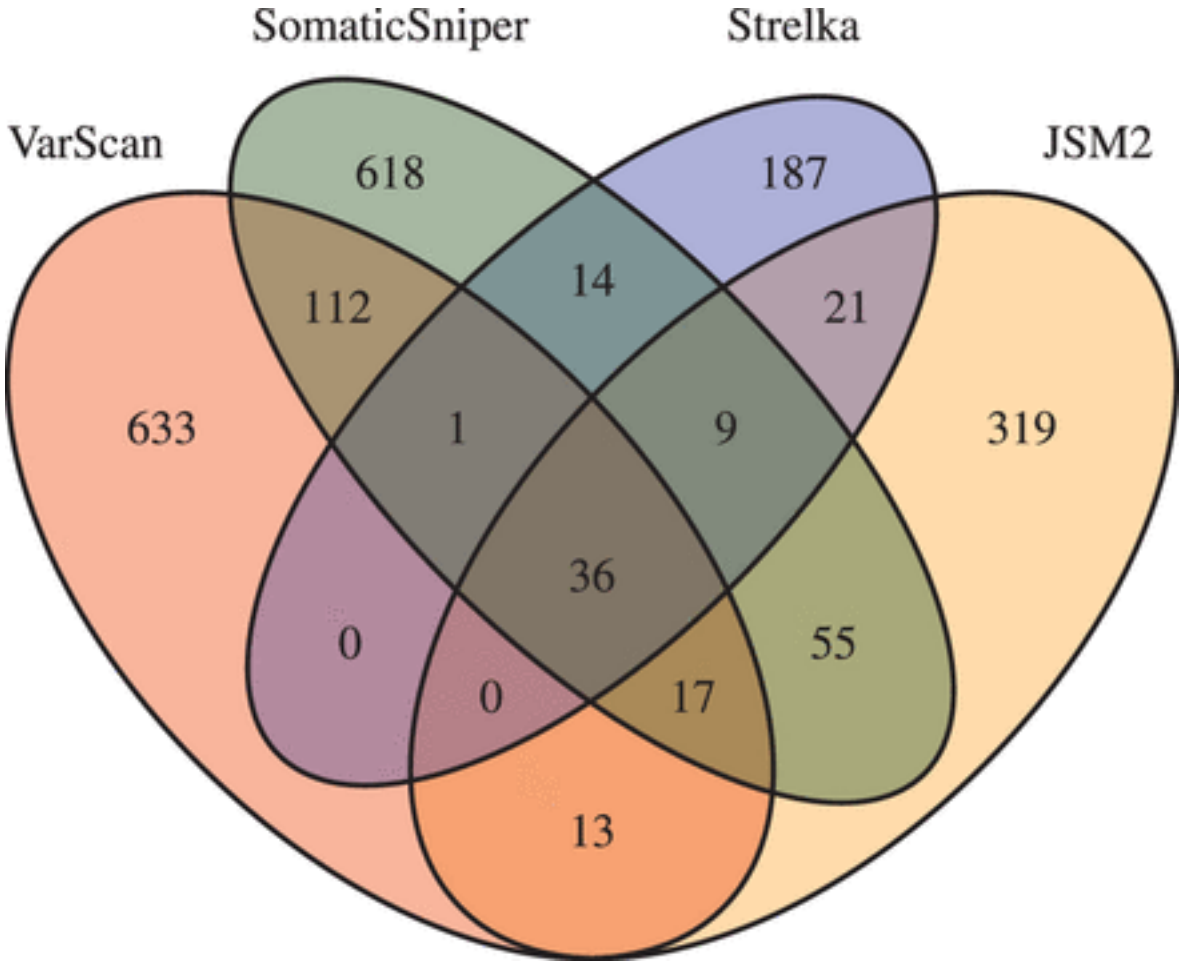
Here are a few more, a summary of the other answers, and updated links:

3.5 years ago by Malachi Griffith ♦ **14k** Washington University School of Medicine, St. Louis, USA

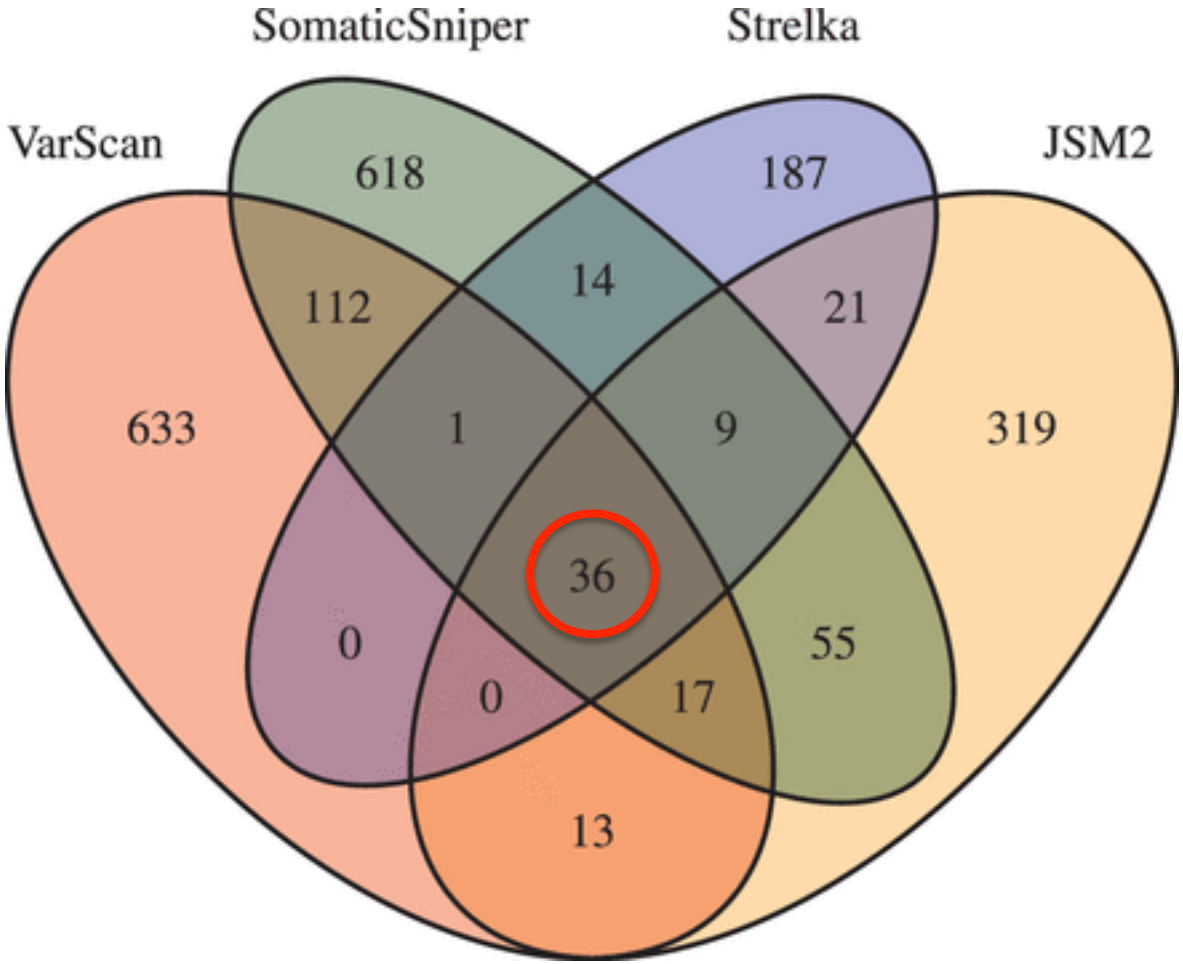
- BAYSIC (paper) - method for combining sets of genome variants
- CAKE (abstract) (paper) - pipeline that combines four other variant callers
- deepSNV (abstract) (paper)
- EBCall (abstract) (paper)
- GATK SomaticIndelDetector (note: only available after an annoying sign-up and login)
- Indelocator
- Isaac variant caller (abstract) (paper)
- joint-snv-mix (abstract) (paper)
- LoFreq (abstract) (paper) (call on tumor & normal separately and then use a filter to derive somatic events)
- MutationSeq (abstract) (paper)
- MutTect (abstract) (paper) (note: only available after an annoying sign-up and login)
- Pindel
- QuadGT (for calling single-nucleotide variants in four sequenced genomes comprising a normal-tumor pair and the two parents)
- samtools mpileup - by piping BCF format output from this to bcftools view and using the '-T pair' option
- Seurat (abstract) (paper)
- Shimmer (abstract) (paper)
- SolSNP (call on tumor & normal separately and then compare to identify somatic events)
- SNVMix (abstract) (paper)
- SOAPsnv
- SomaticCall (manual)
- SomaticSniper (abstract) (paper)
- Strelka (abstract) (paper)
- VarDict
- VarScan2 (abstract) (paper)
- Virmid (abstract) (paper)

https://www.biostars.org/p/19104/

Different Variant Callers Give Different Results



Best Practice: Use Multiple Variant Callers



DREAM Challenge

Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection

Adam D Ewing[1,2,11], Kathleen E Houlahan[3,11], Yin Hu[4,11], Kyle Ellrott[1], Cristian Caloian[3], Takafumi N Yamaguchi[3], J Christopher Bare[4], Christine P'ng[3], Daryl Waggott[3], Veronica Y Sabelnykova[3], ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants[5], Michael R Kellen[4], Thea C Norman[4], David Haussler[1], Stephen H Friend[4], Gustavo Stolovitzky[6], Adam A Margolin[4,7,8,12], Joshua M Stuart[1,12] & Paul C Boutros[3,9,10,12]

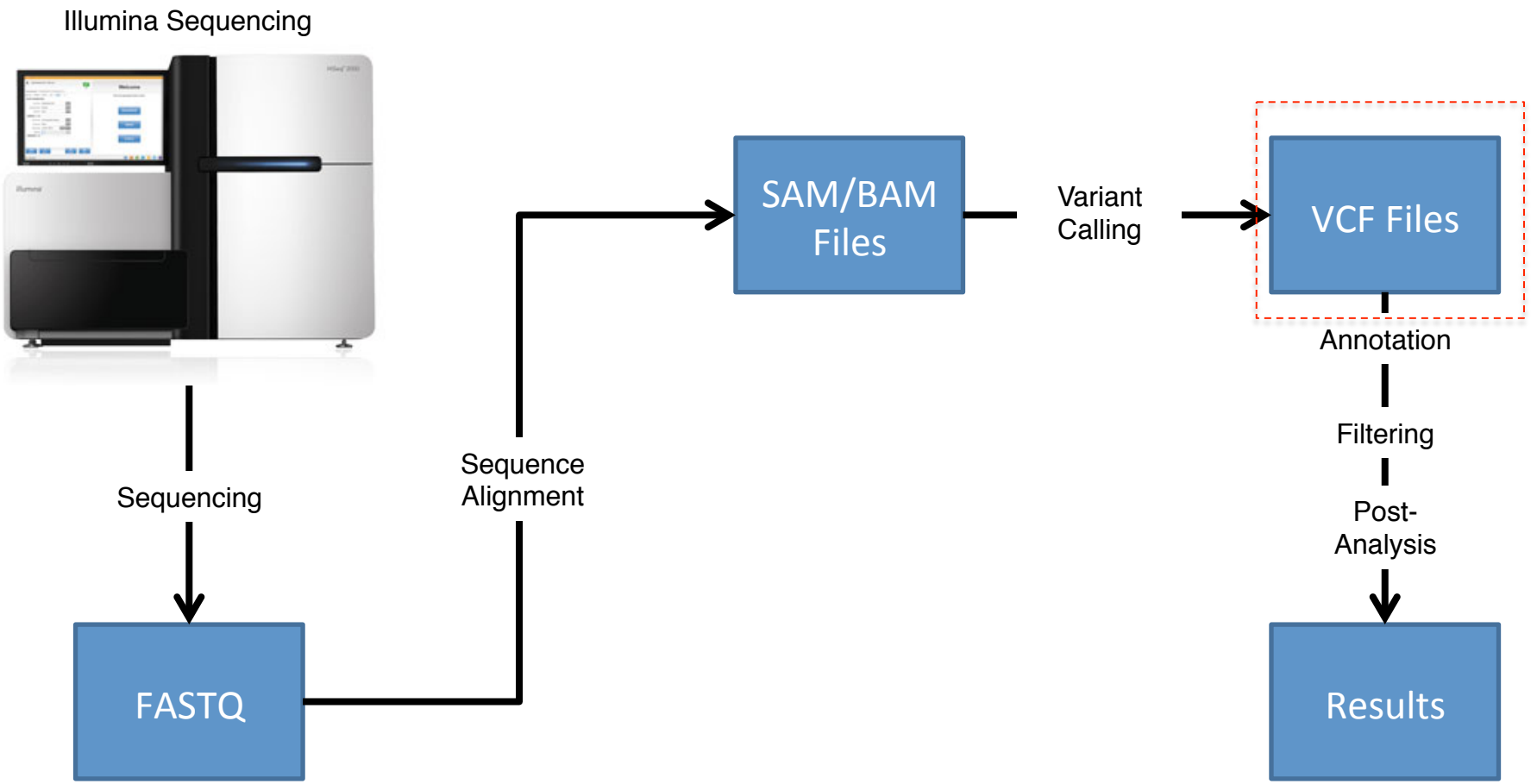
- Competition to benchmark the best somatic variant callers
  - https://www.synapse.org/#!Synapse:syn312572

Example of Variant Callers

- Strelka (Saunders et al., Bioinformatics, 2012)
  - Realignment around indels

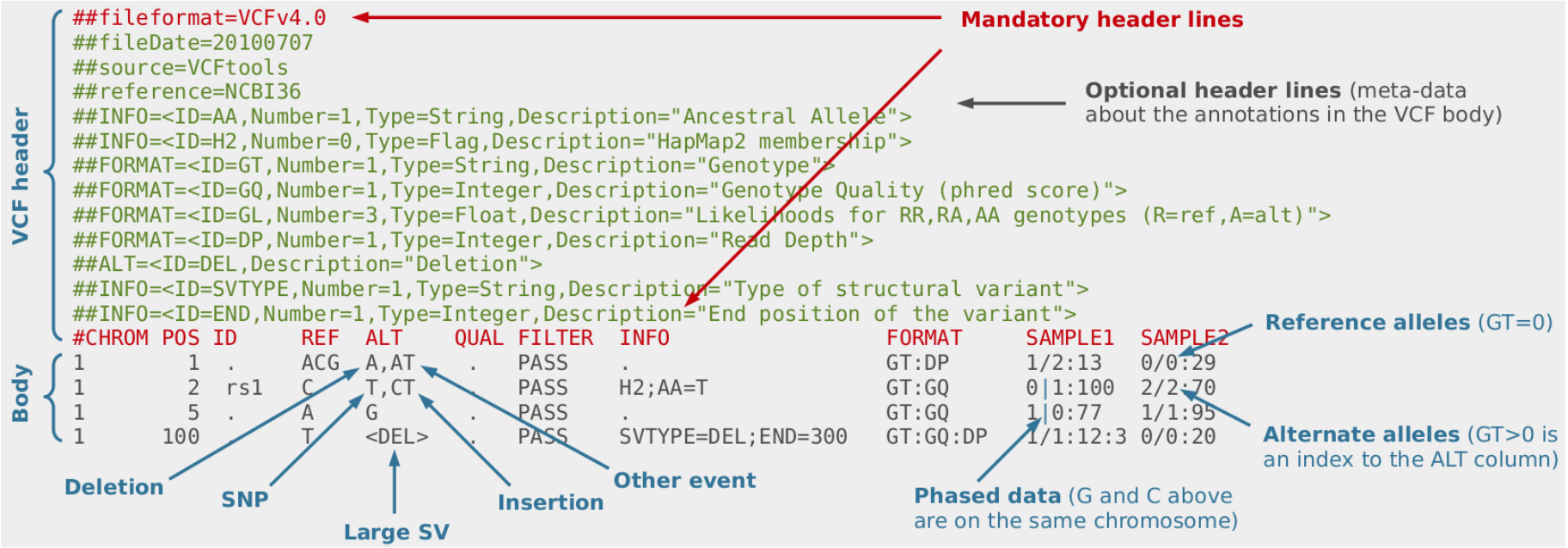
- MutationSeq (Ding et al., Bioinformatics, 2011)
  - Supervised learning based on ~1000 validated events
  - Especially effective against technical artifacts

Bioinformatics Pipeline for Variant Calling

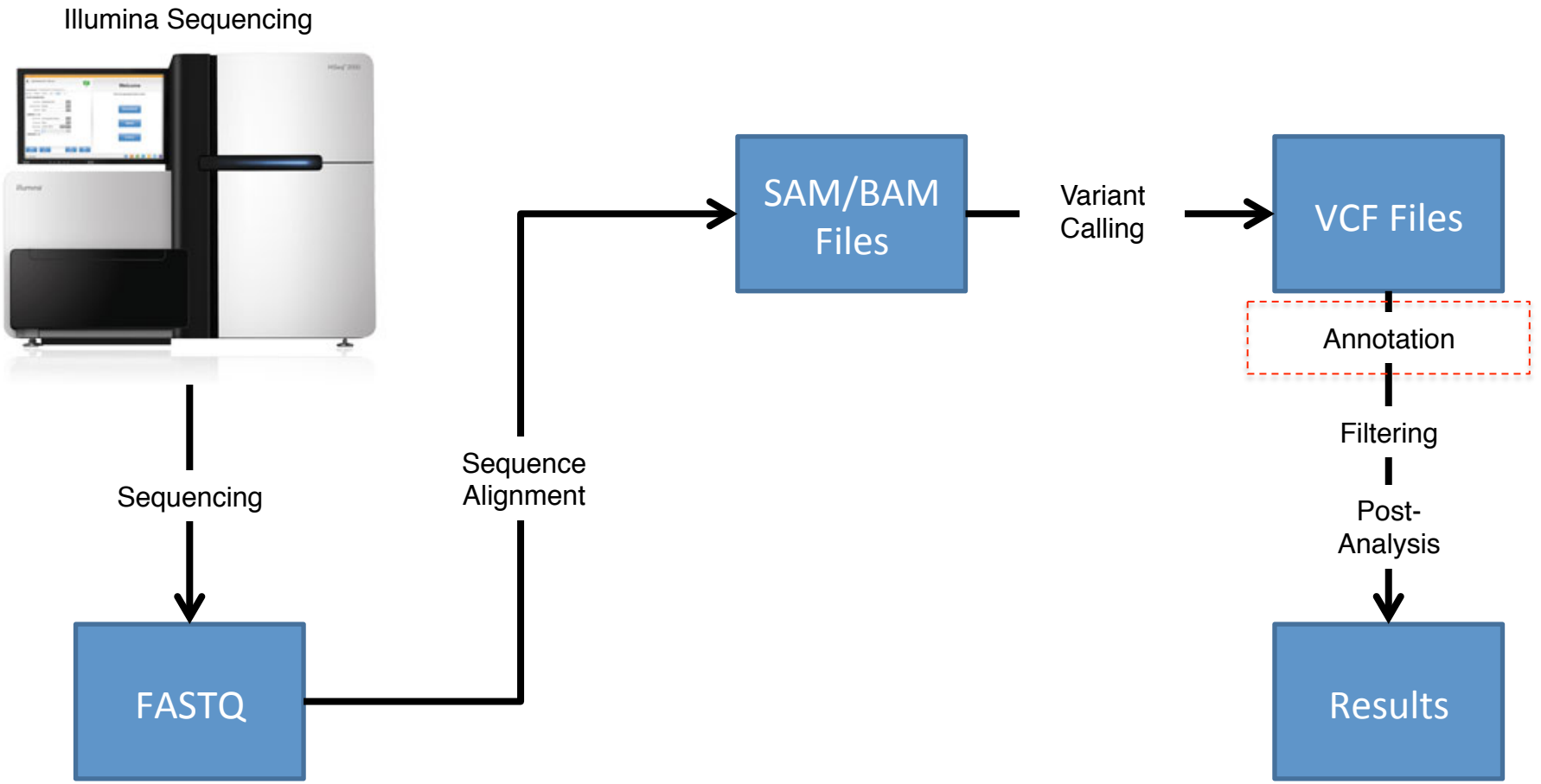


VCF Format

- De factor standard is the Variant Call Format (VCF):
  - 1 line per variant
  - Structured fields as detailed in header



Bioinformatics Pipeline for Variant Calling



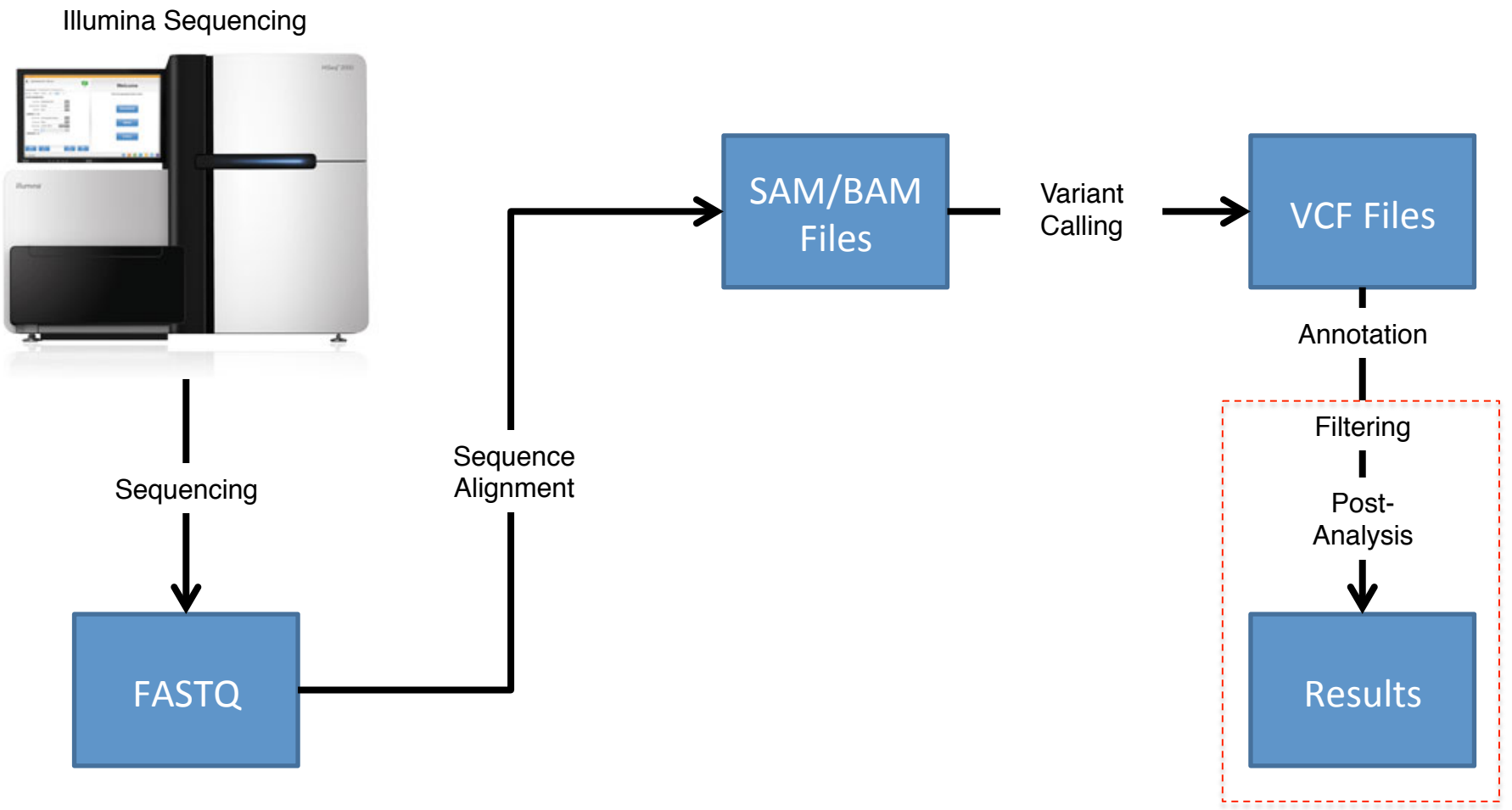
Annotation

- Annotation for SNV effects (e.g. non-synonymous, splice junction), gene annotations, etc

- At least 3 major annotation software:
  1. SnpEff (used in the lab)
  2. AnnoVar
  3. VEP

- Differences are discussed here:
  - http://blog.goldenhelix.com/ajesaitis/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/

Bioinformatics Pipeline for Variant Calling



Variant Calling Post-Analysis

- Typically convert the VCF file into a tabular format for post-analysis

- Load data into python or R for post-process
  - Distribute as an excel file to collaborators

Lab Demonstration

- Provide the setup instructions and the commands for variant calling in cancer genomes
- Specifically, the dataset is a matching tumor and normal exome from a breast cancer cell-line (HCC1395)
  - In principle, the commands/concepts are directly applicable to genome data
- Github Repo
  - https://github.com/tinyheero/variant_calling_in_cancer_genomes_seminar