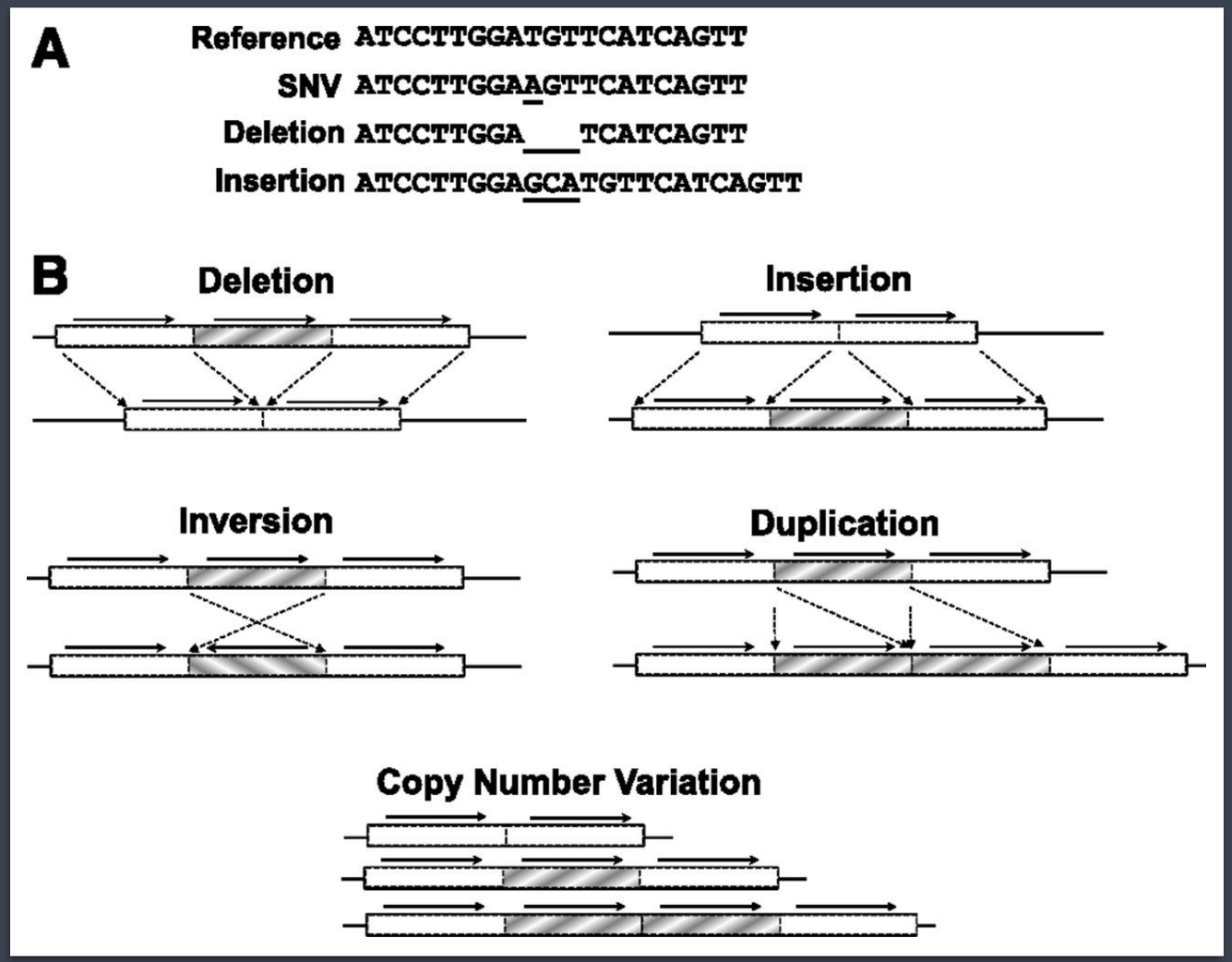
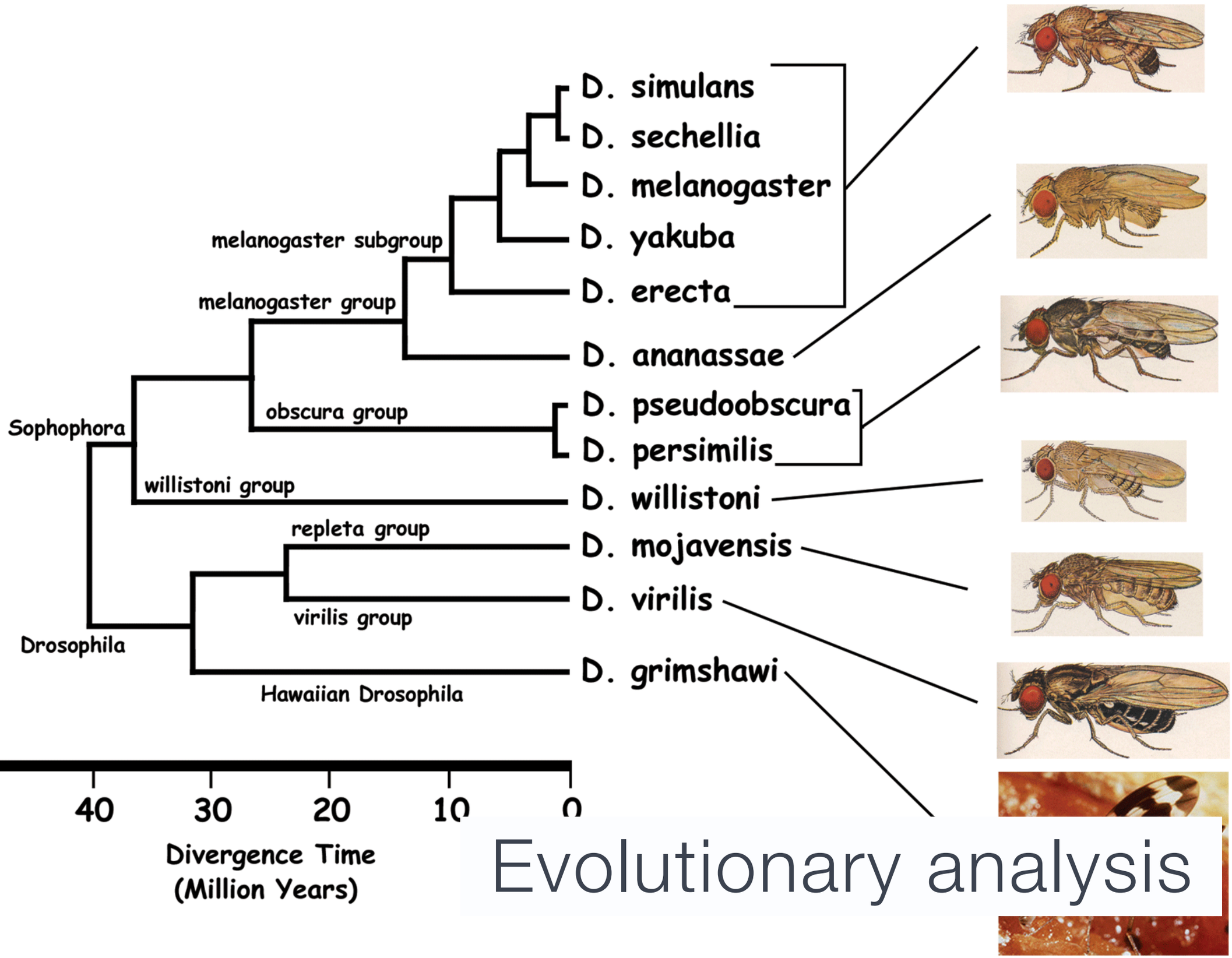
Introduction to Variant Detection

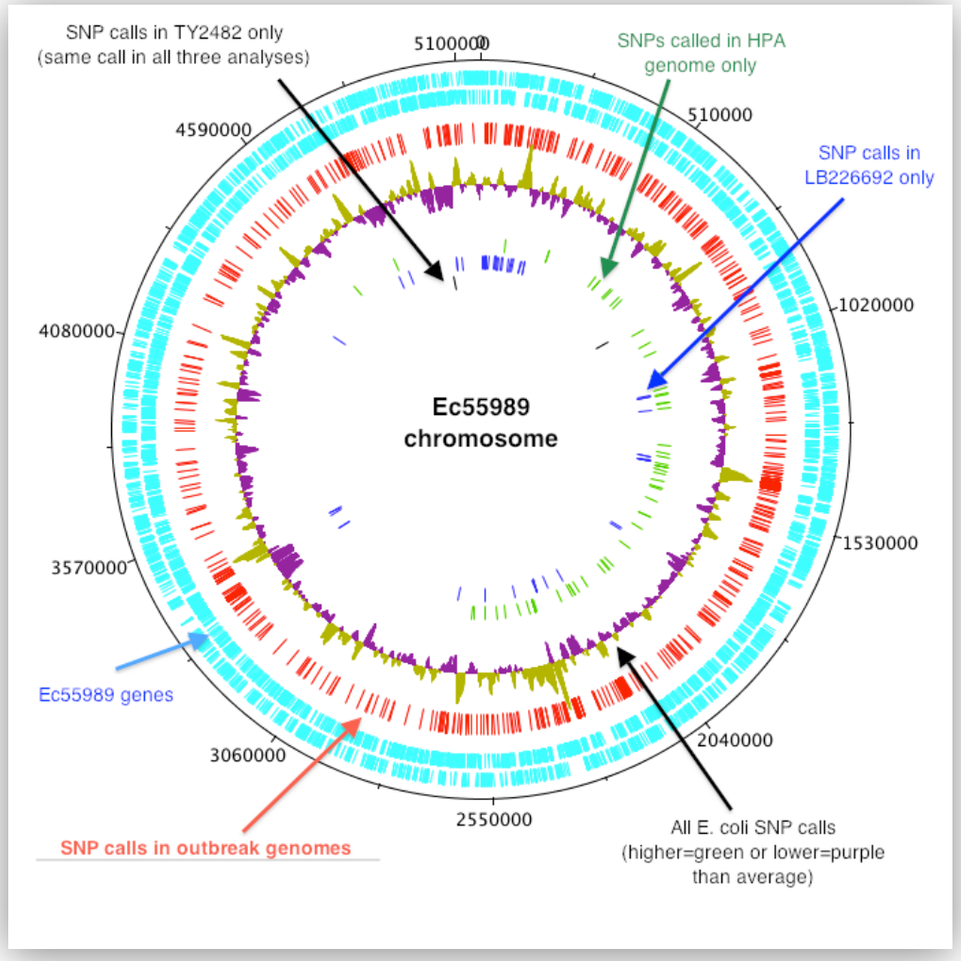


Johnson et al., Blood, 2013, 122(19)

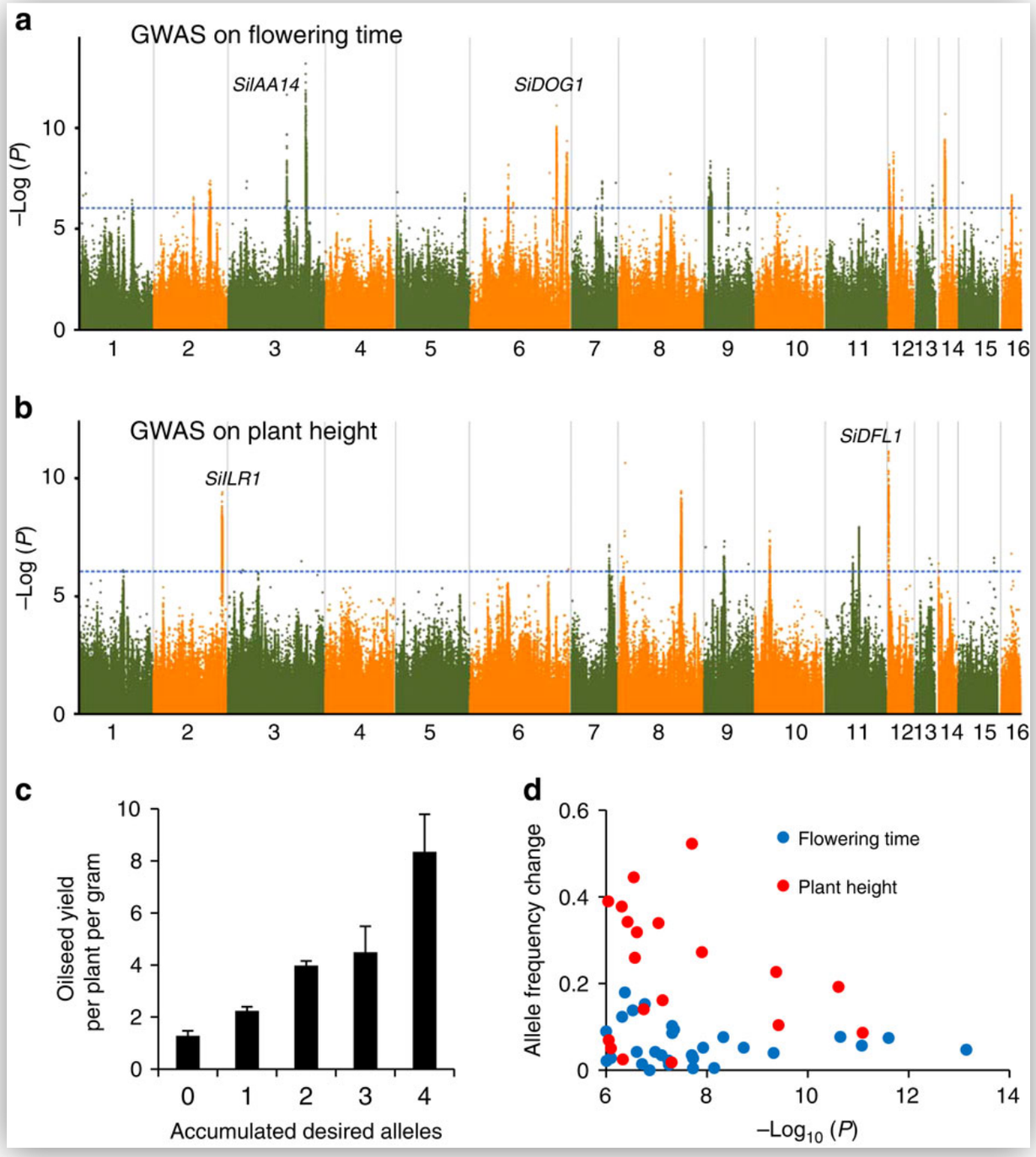


Evolutionary analysis

http://insects.eugenes.org/DroSpeGe/

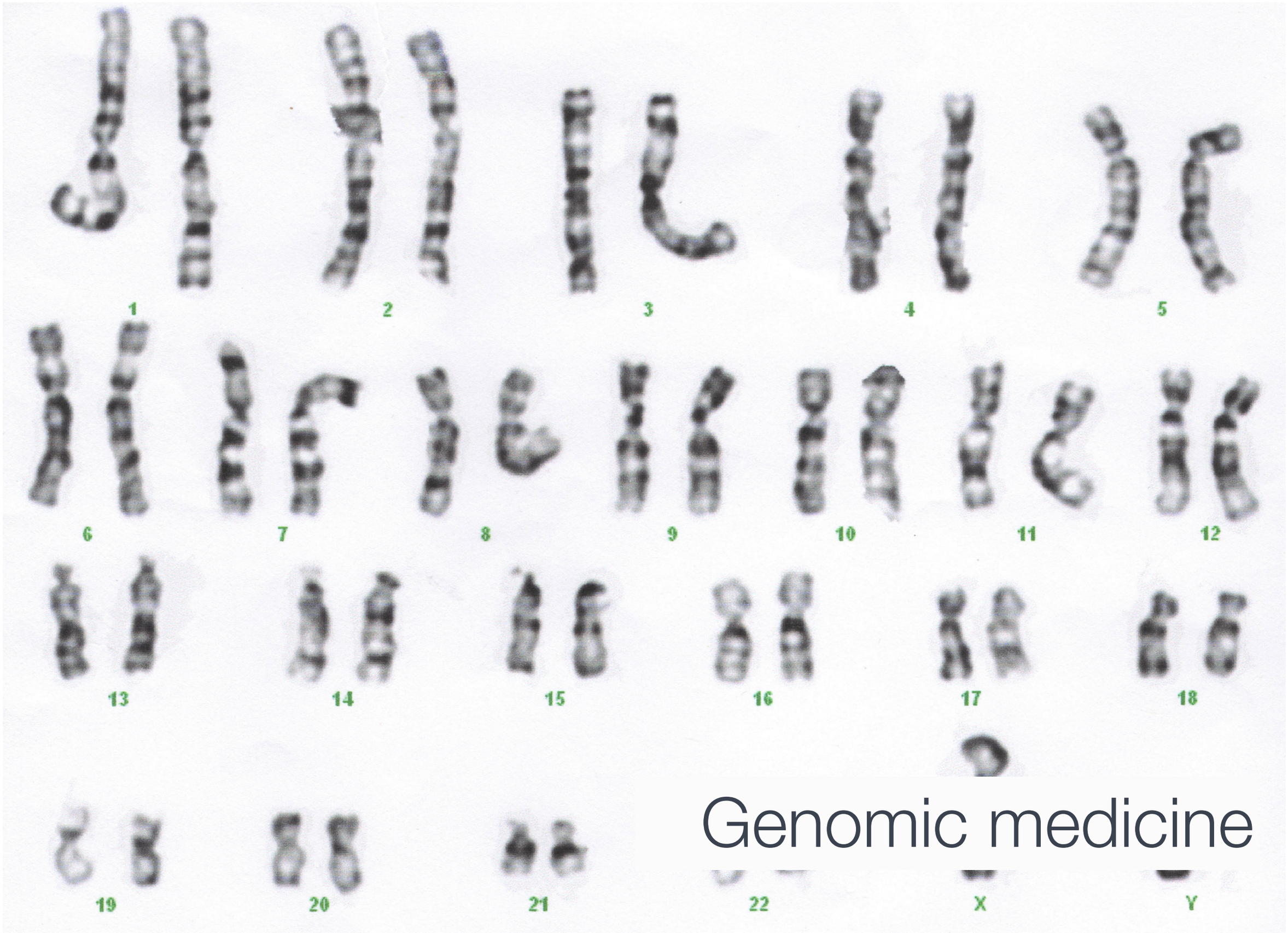


https://bacpathgenomics.wordpress.com/tag/snp/



http://www.nature.com/ncomms/2015/151019/ncomms9609/fig_tab/ncomms9609_F6.html

Medicine and Agriculture



Genomic medicine

Overview

- Human variations
  - Germline
  - Somatic
- Types of Variations
- Sequencing strategies to identify variants
- Generalized analysis workflow (GATK best practice guidelines)

» Any heritable “mutation” is considered a germline variant.

- found in populations, discovered by large-scale population analyses, and contained in databases like dbSNP, HapMap
- most are not deleterious

» A somatic variant is any mutation that arises in a single cell of an individual and is only present in the descendants of that cell, not all the cells of an individual.

- found in rapidly growing cancer cells
- can be silent or pathogenic

Germline vs Somatic mutations

Nature 491, 56–65 (01 November 2012)

- Most human genomic variants have no phenotypic impacts
  - Most of those that do are either positively selected (i.e. they confer a reproductive advantage) or are neutral, and typically, they affect traits like height, facial features, hair or skin color, often associated with ethnic origin
- Some genomic variants have effects that are deleterious to health
  - Most of these are recessive: their effect is observed only if both alleles are affected; these recessive alleles are often associated with specific ethnic groups
  - Those that are dominant will either be selected against and disappear, or have effects that minimally impact reproductive fitness (e.g., adult cancer)

This implies that the vast majority of alleles commonly found in the population do not directly cause disease

Types of variations

Single Nucleotide Polymorphisms (SNPs)

Copy Number Variations (CNVs)

Structural Variations (SVs)

For SNPs, many different methods have been used:

- Hybridization based, primarily SNP arrays
- Enzyme-based methods, primarily oligonucleotide ligation and RFLP
- Methods measuring physical properties of DNA

For CNVs, the main methods were hybridization based

For SVs the most reliable ones used partial sequencing of large clones (e.g. fosmids)

NGS can detect all types of variants

(Paired-end data preferred!)

How to assess genomic diversity?

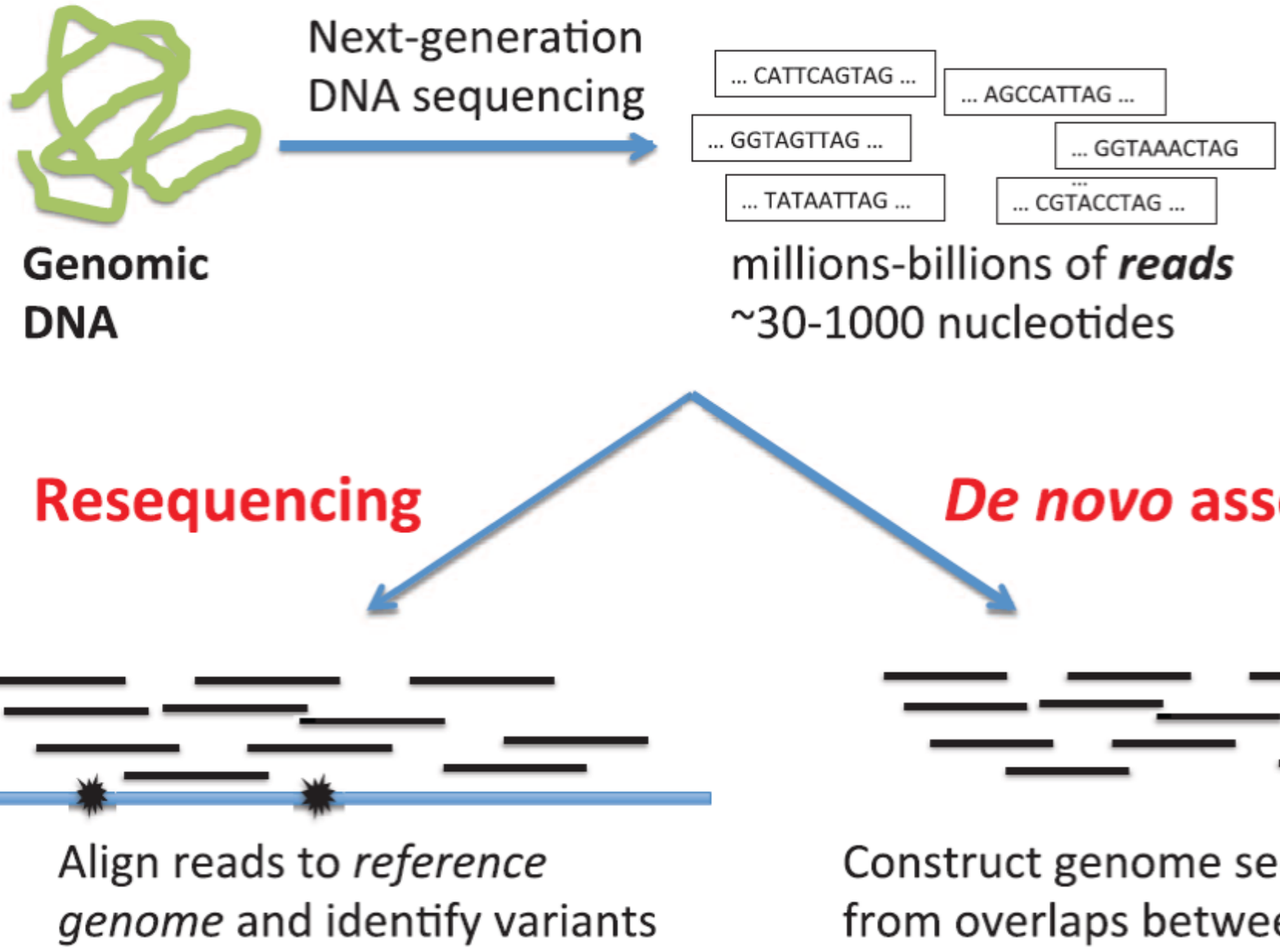
Sequencing strategies

Whole Genome Sequencing (WGS)

(for SNPs/small indels, CNVs and SVs)

Exome Sequencing (for SNPs/small indels)

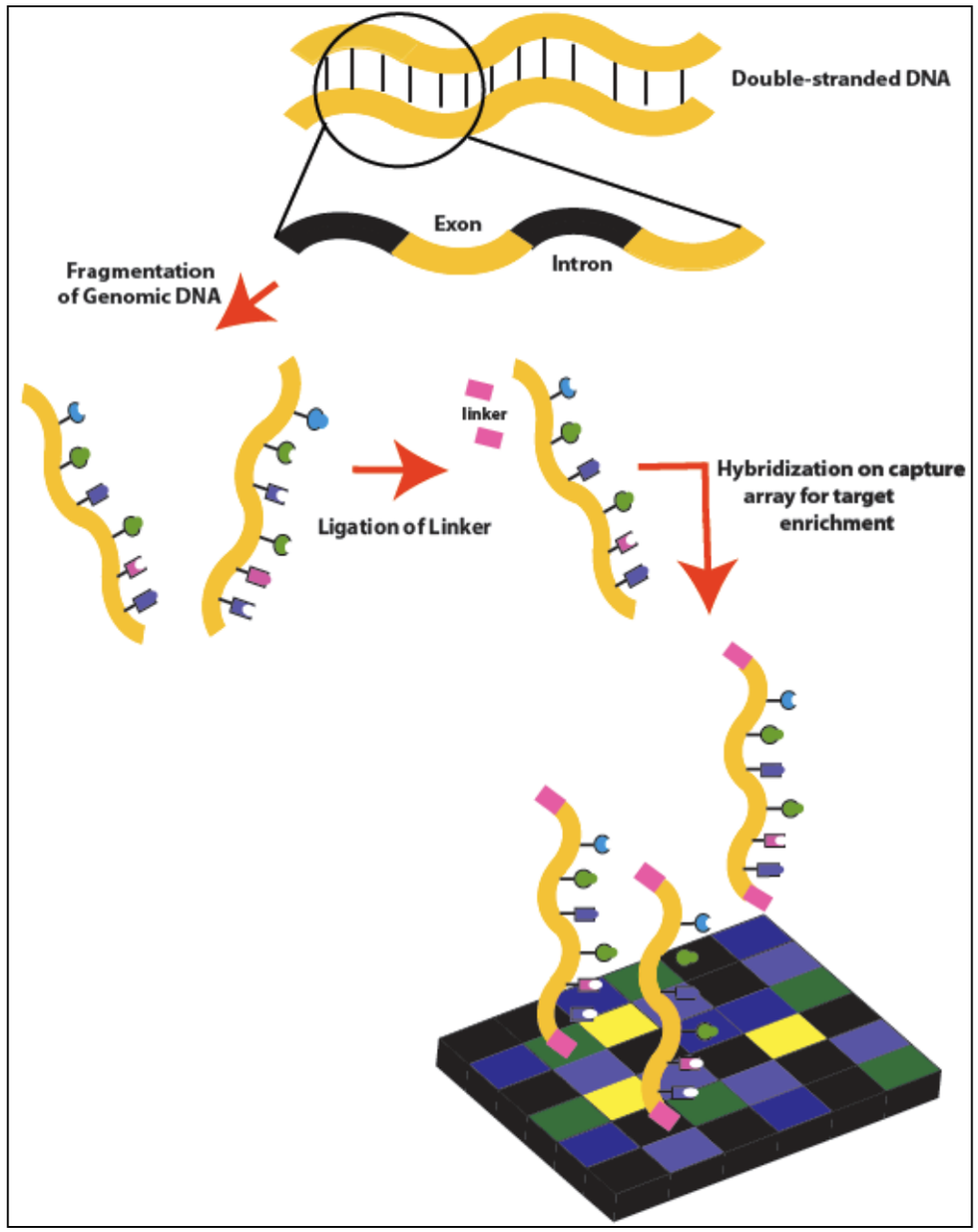
Gene Panels (for SNPs/small indels)

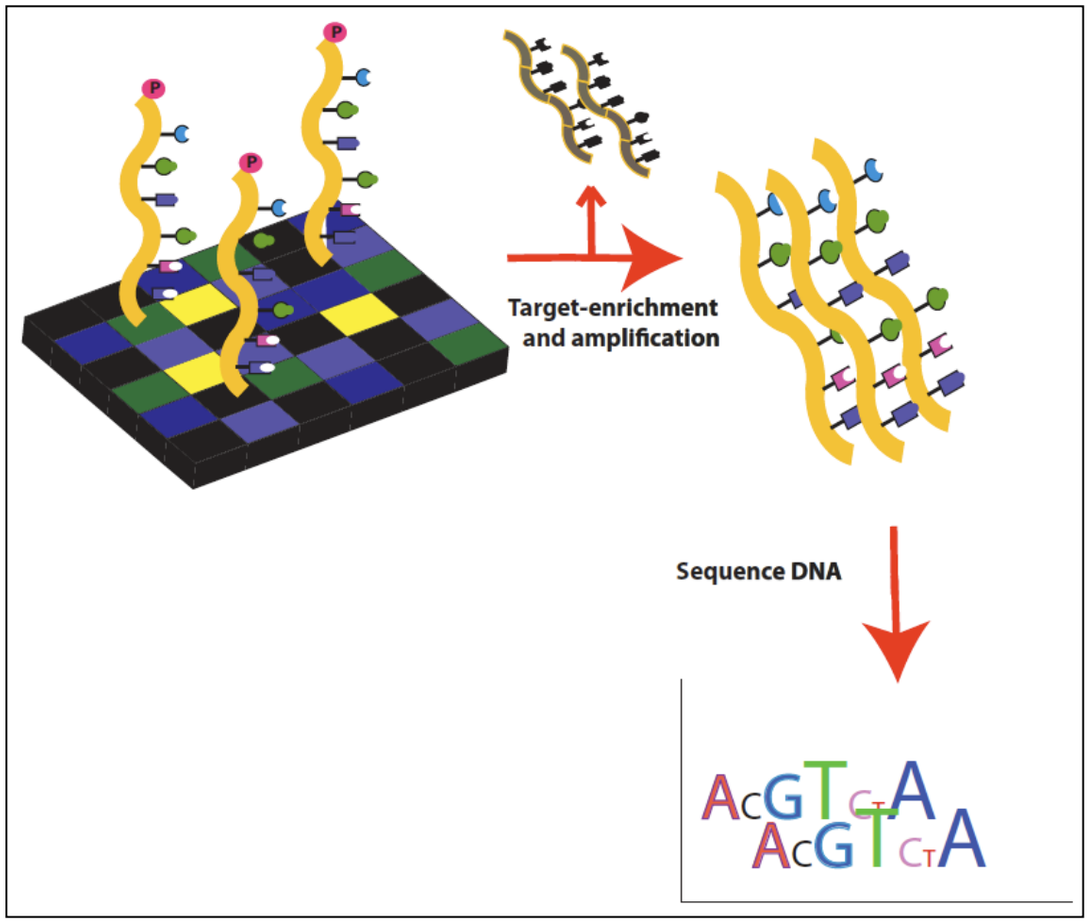


Benjamin J. Raphael*

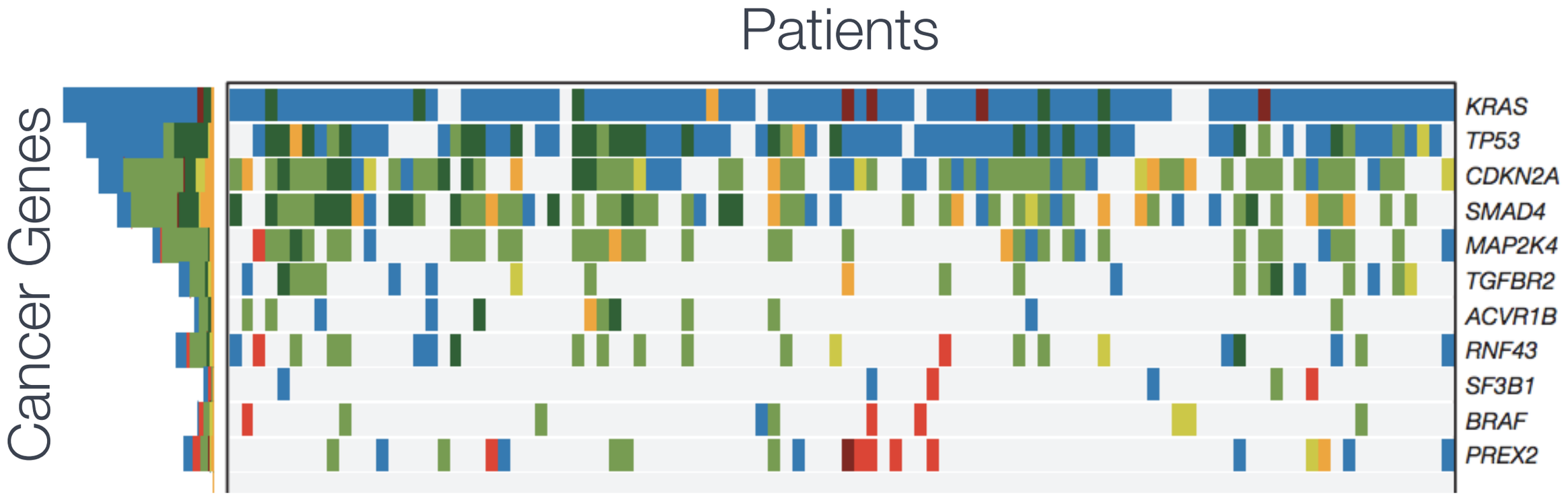
Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America

Whole genome sequencing





Exome sequencing

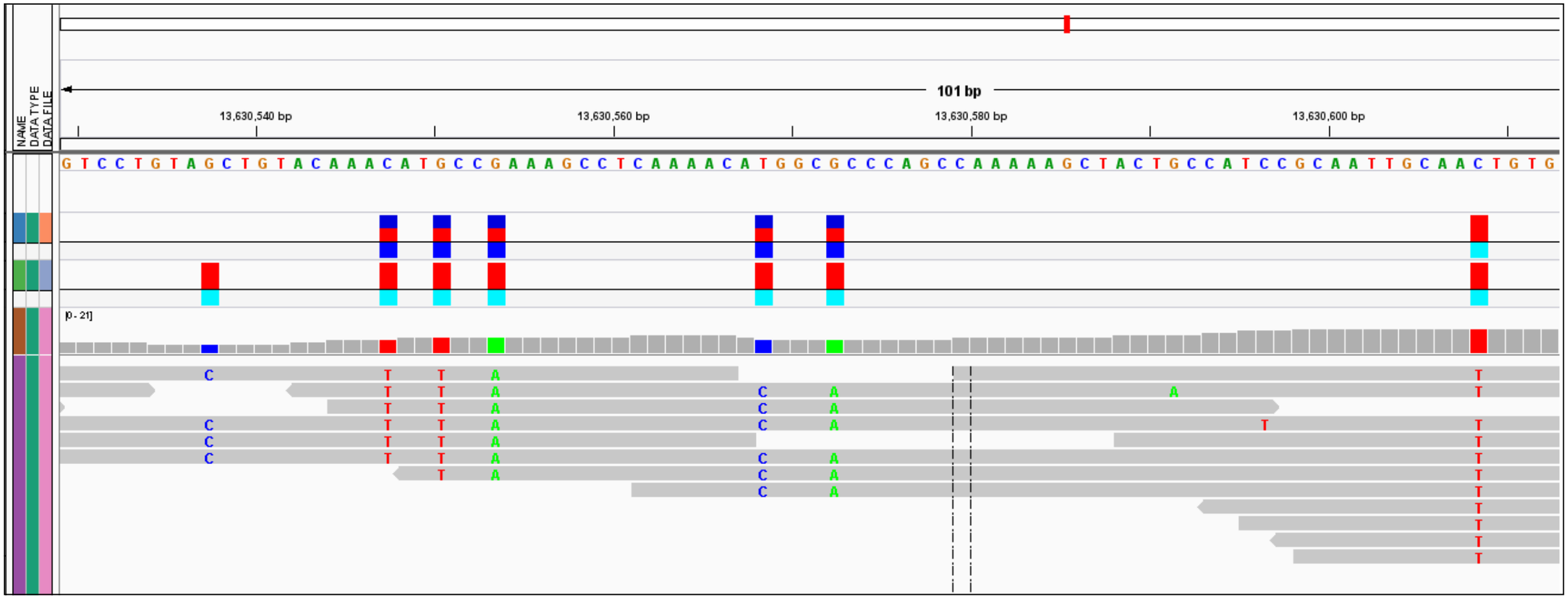


A visualization of an analysis using a panel of known cancer genes

Gene panel sequencing for diagnostics

- Targeted gene panels are the most commonly used for diagnostics
- Coverage:cost considerations for various methods, based on number of samples
- Variants in un-targeted or non-exonic regions will be missed

Gene panels or ES or WGS: Which one is “better”?



Sequencing depth and cost

For WGS

- Haploid genome size => 3.2 Giga base pairs (3.2 billion)
- Minimum 30x for WGS

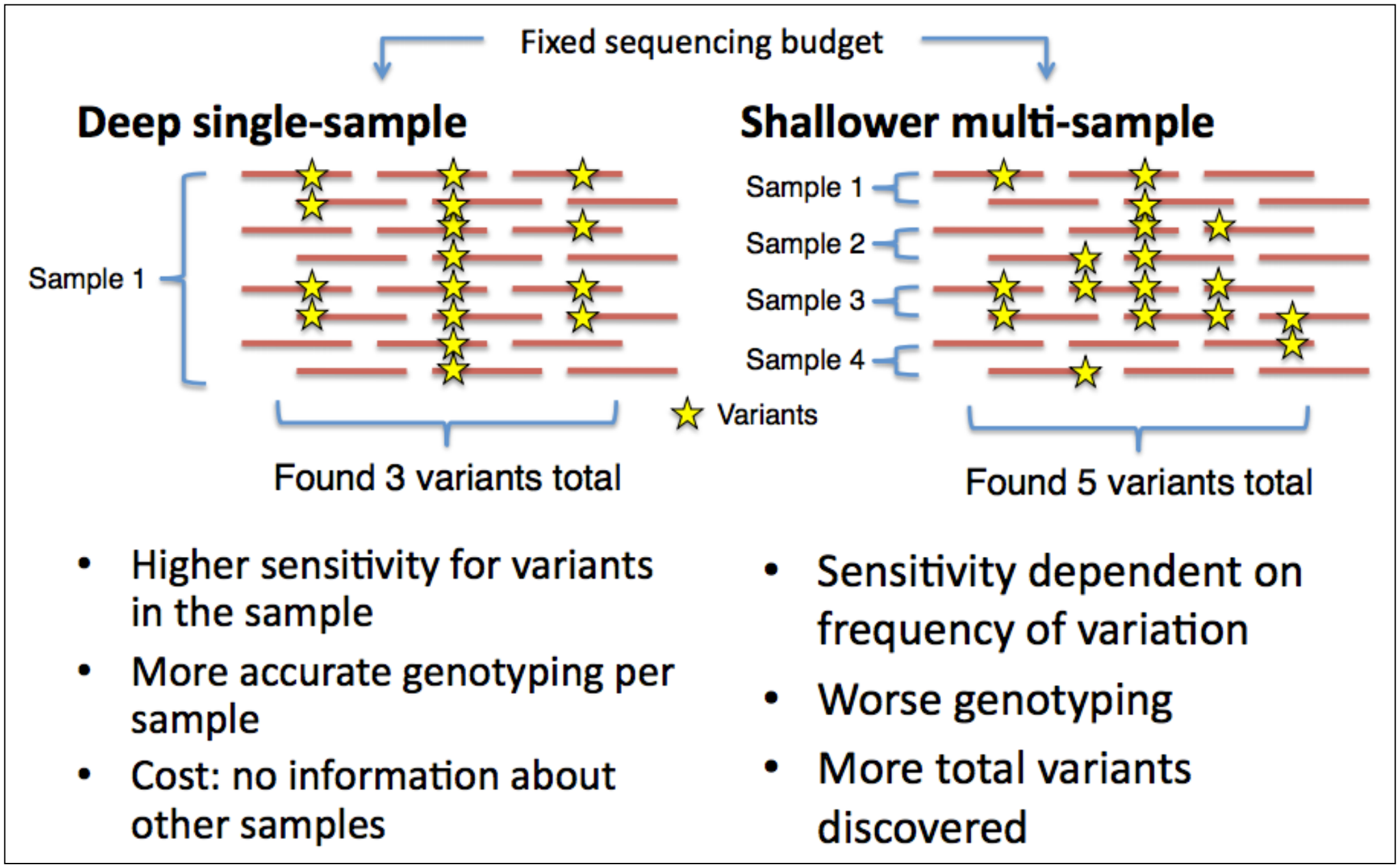
For Exome Sequencing

- Exome size => 33 Mega base pairs (33 million bases)
- About 100 times smaller than WGS
- 70x-100x for ES, with additional considerations for unevenness of coverage

For Gene Panels

- 10x-20x coverage for gene panels for heterozygous germline variants

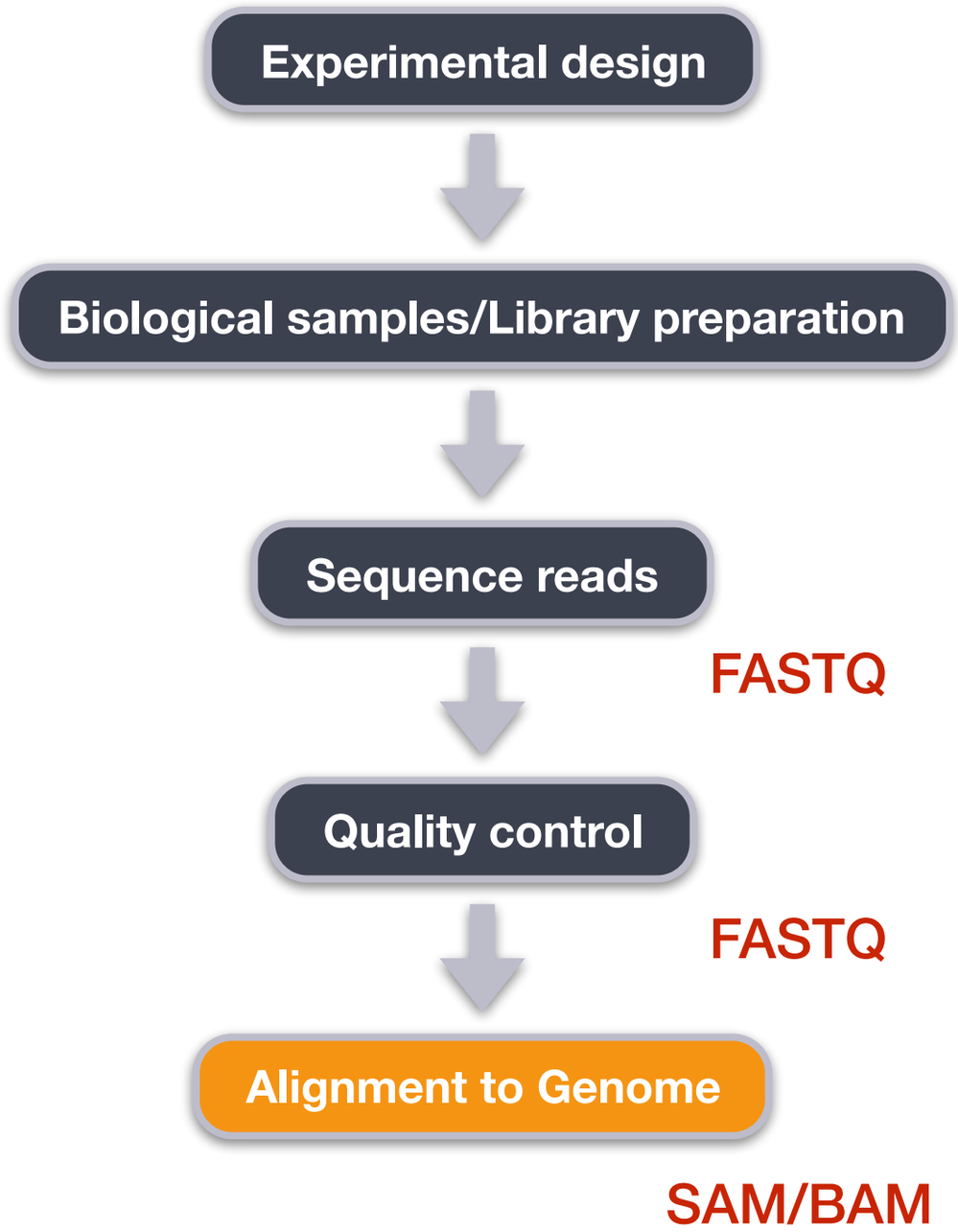
Sequencing depth?

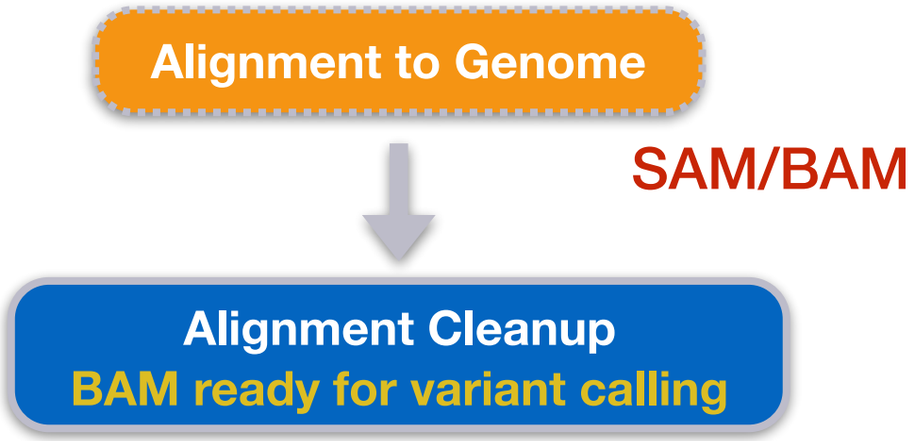


Sequencing depth and cost

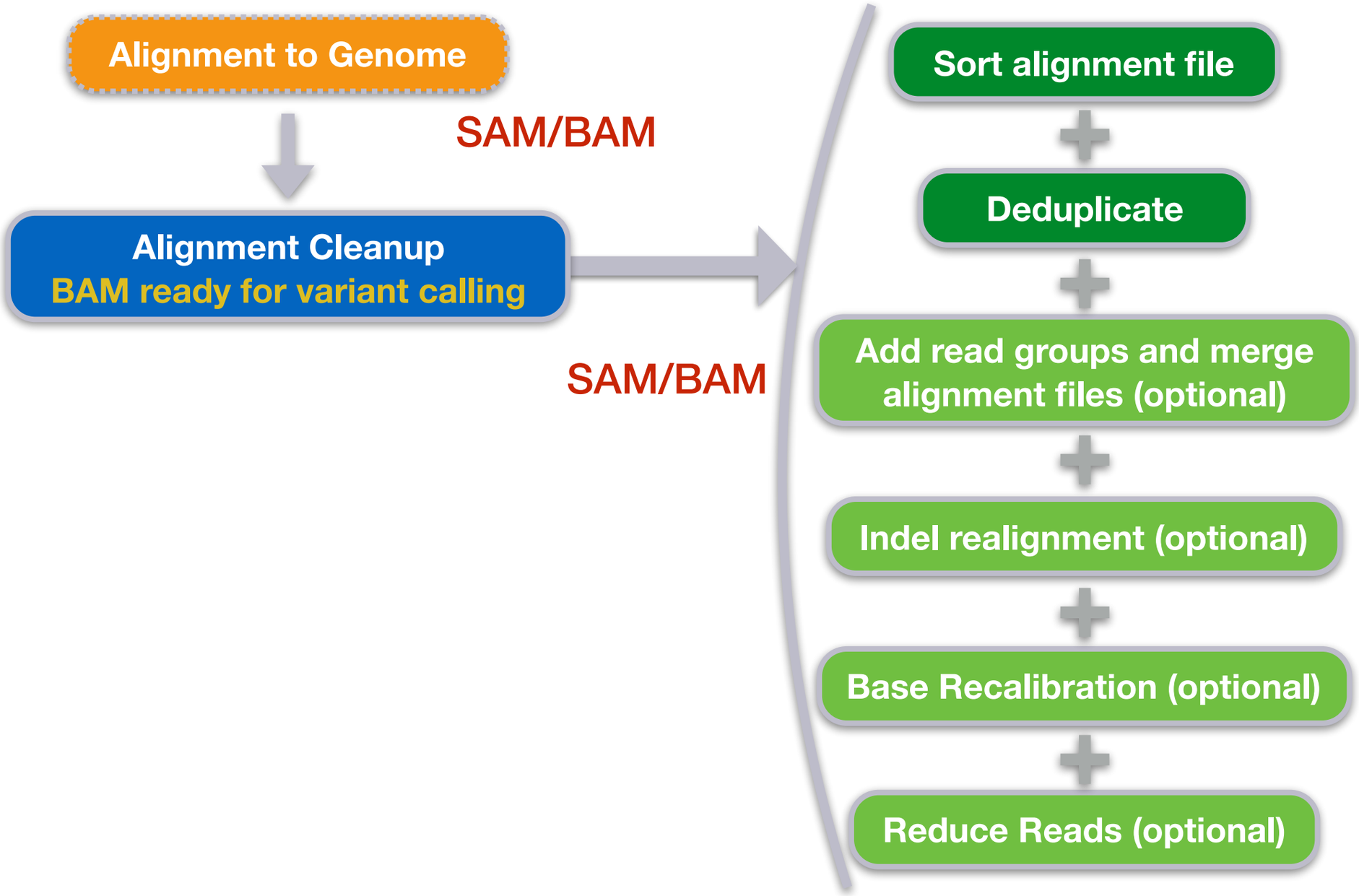
Adapted from GATK best practices guidelines (2012)

Generalized Variant Calling Workflow

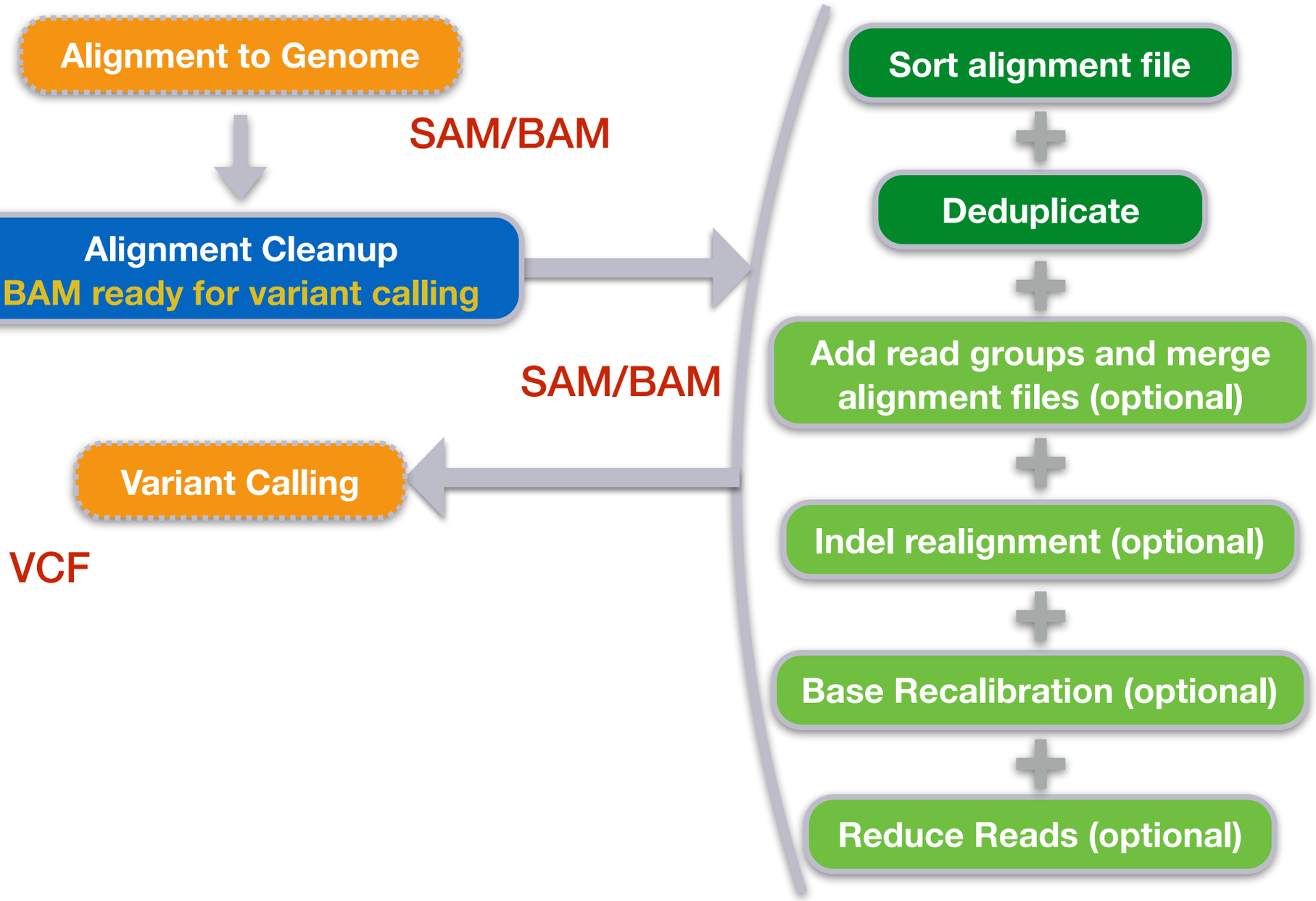




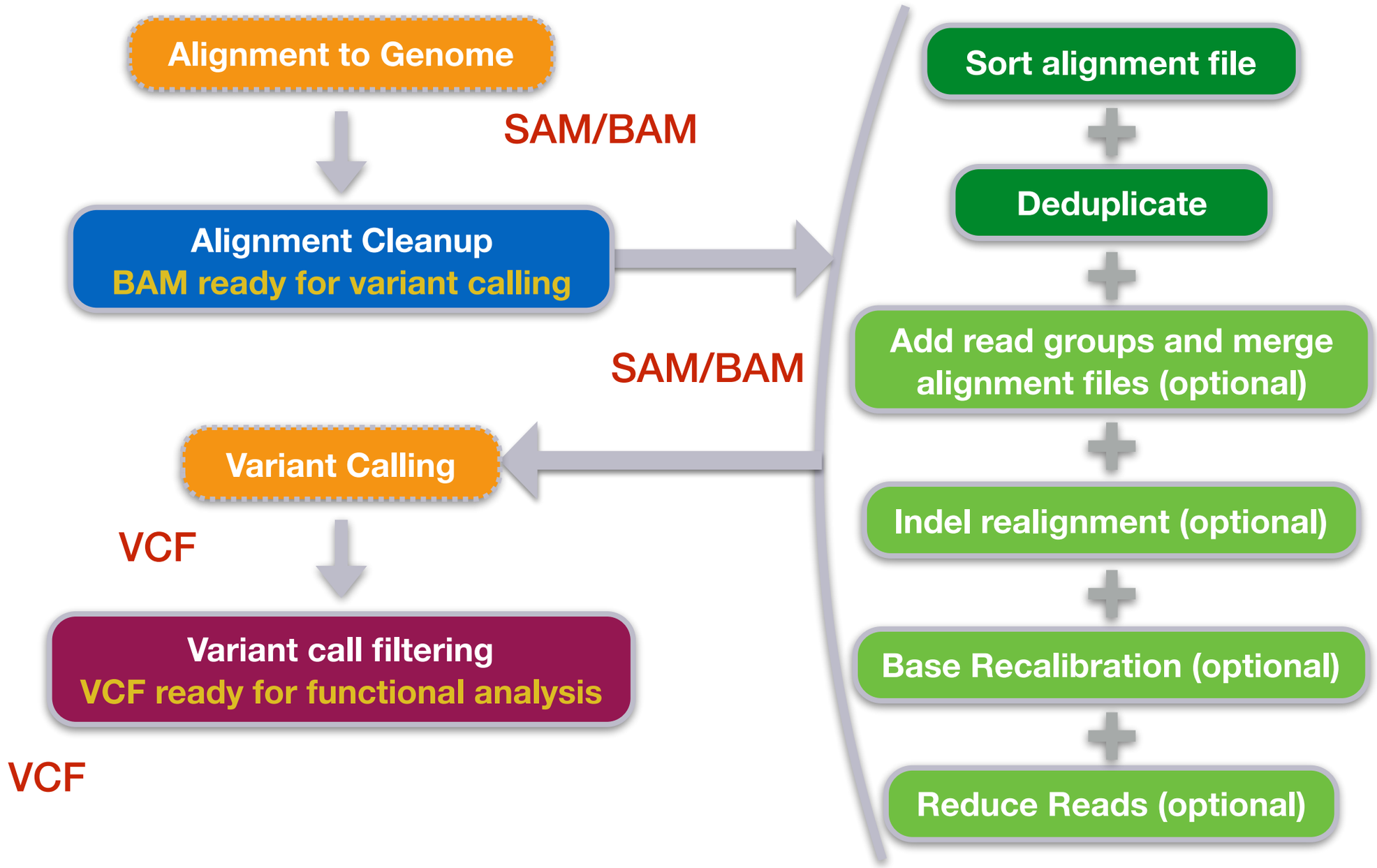
Generalized Variant Calling Workflow



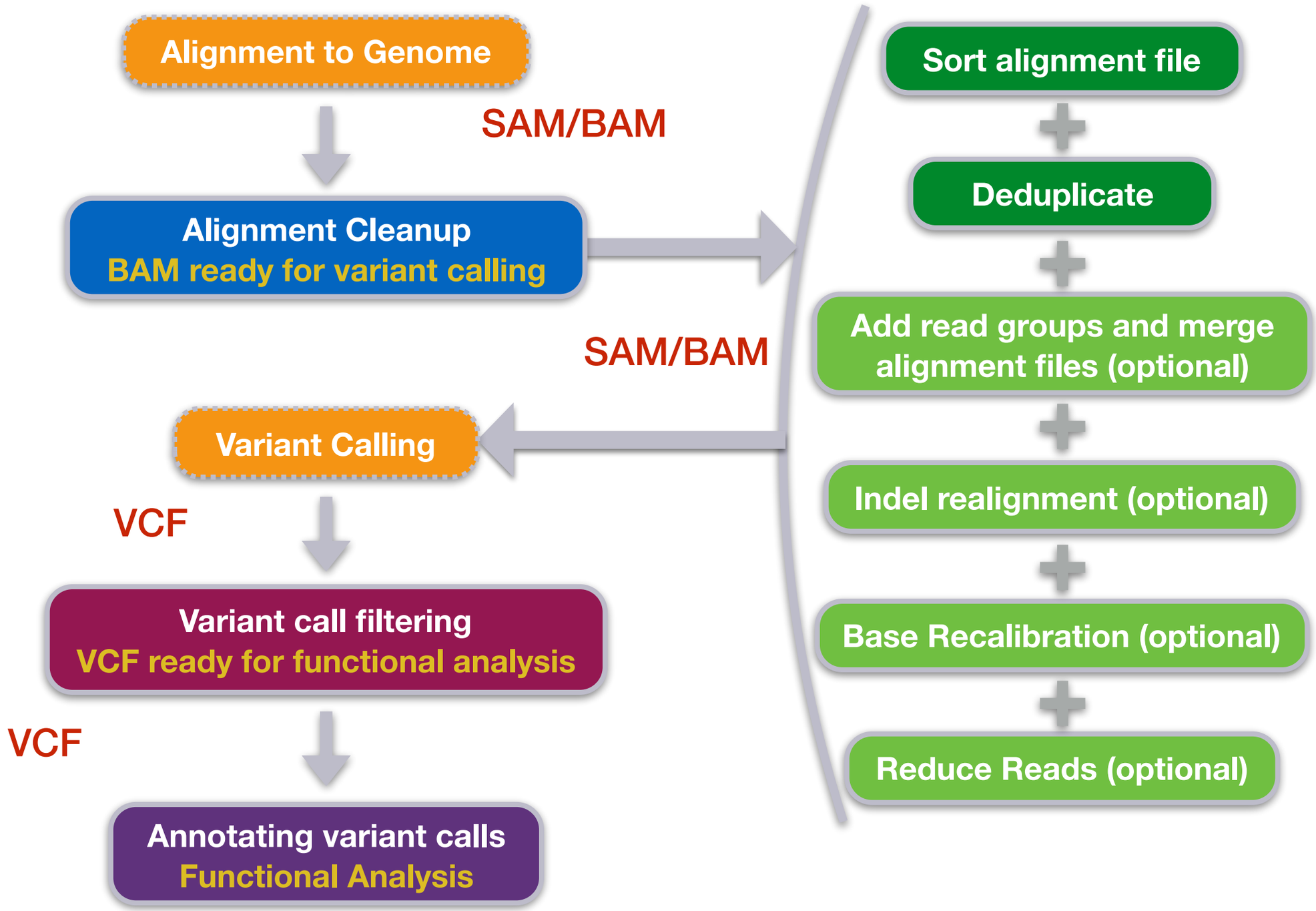
Generalized Variant Calling Workflow



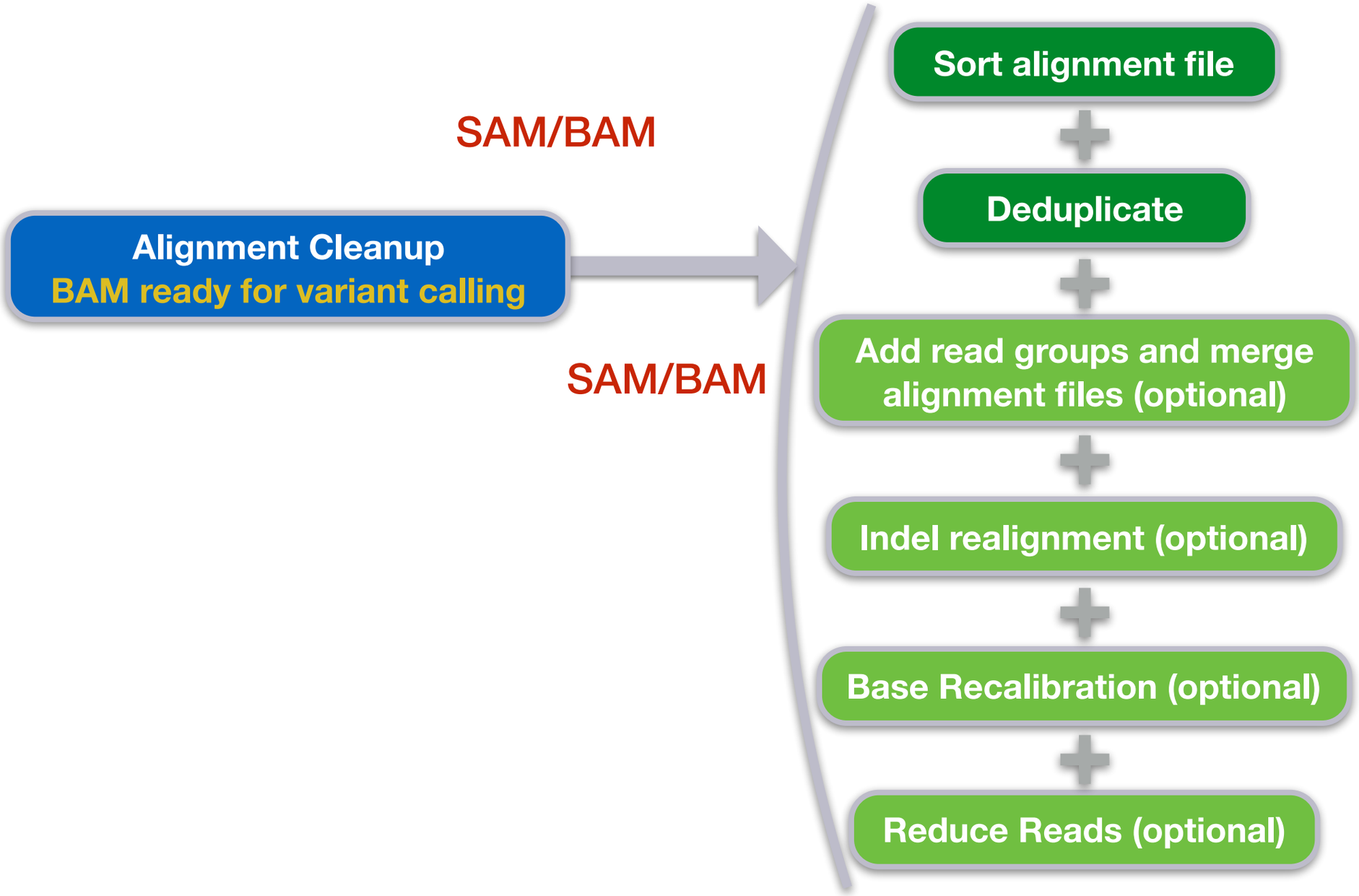
Generalized Variant Calling Workflow



Generalized Variant Calling Workflow

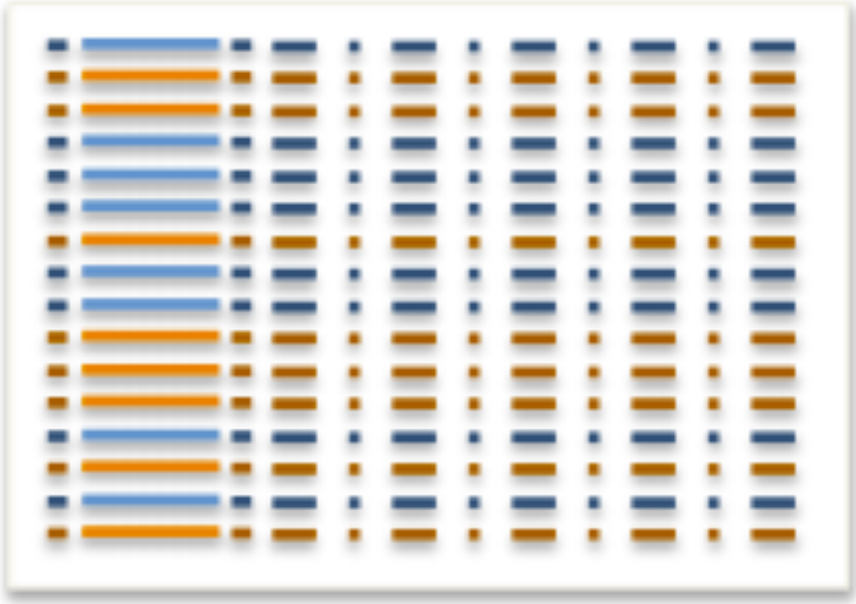


Generalized Variant Calling Workflow

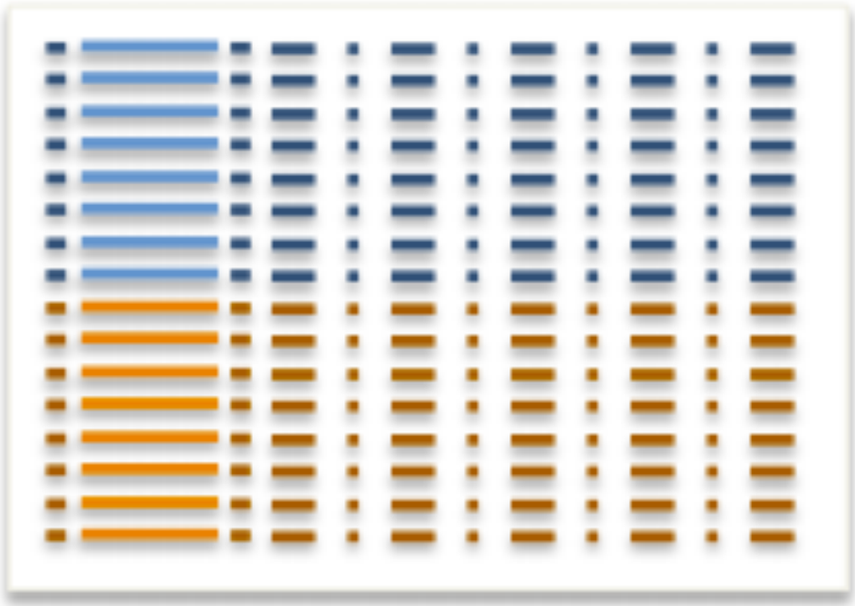


Generalized Variant Calling Workflow

Sort alignment file



The reads are in no particular order…

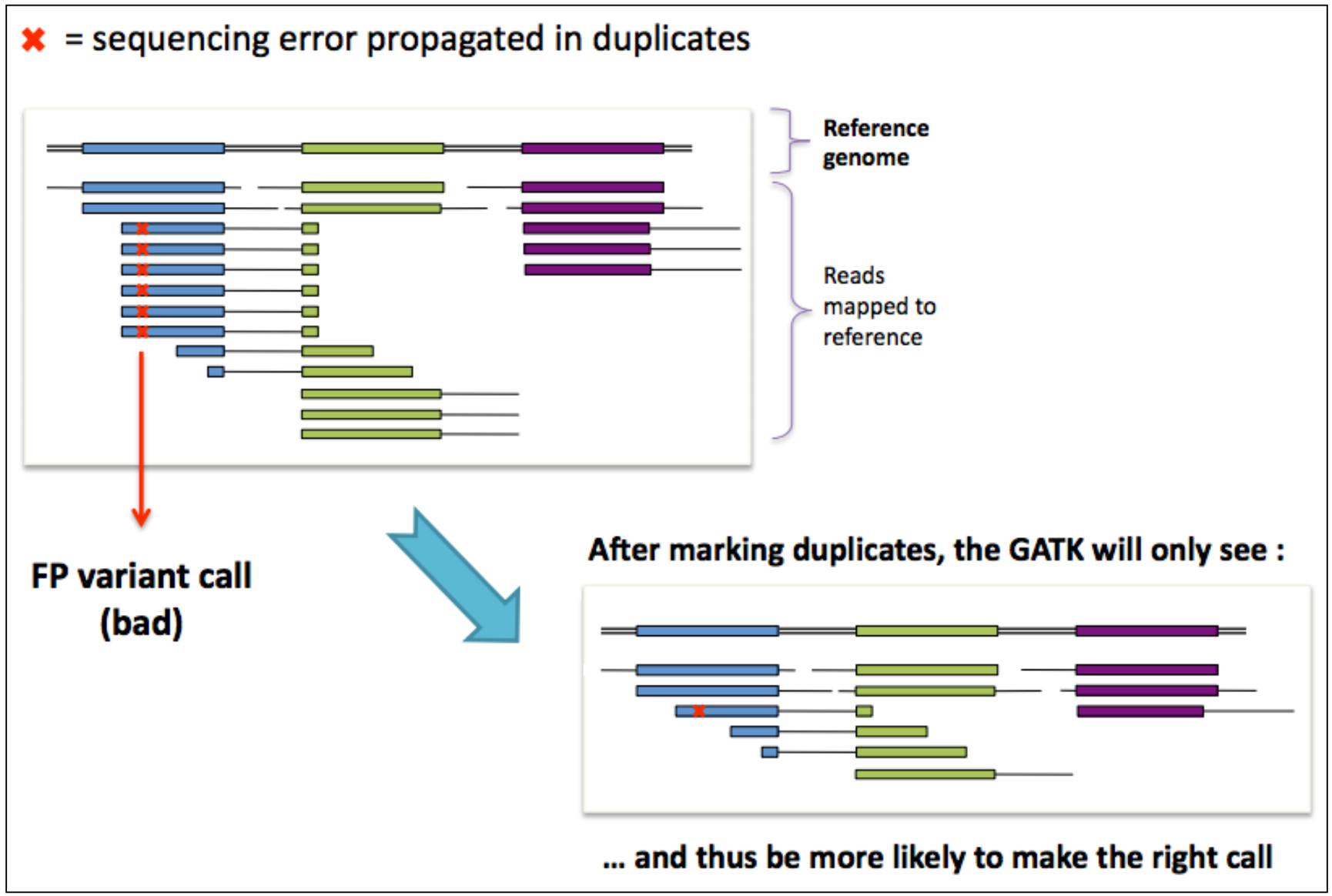


So we need to explicitly sort the SAM file…

Generalized Variant Calling Workflow

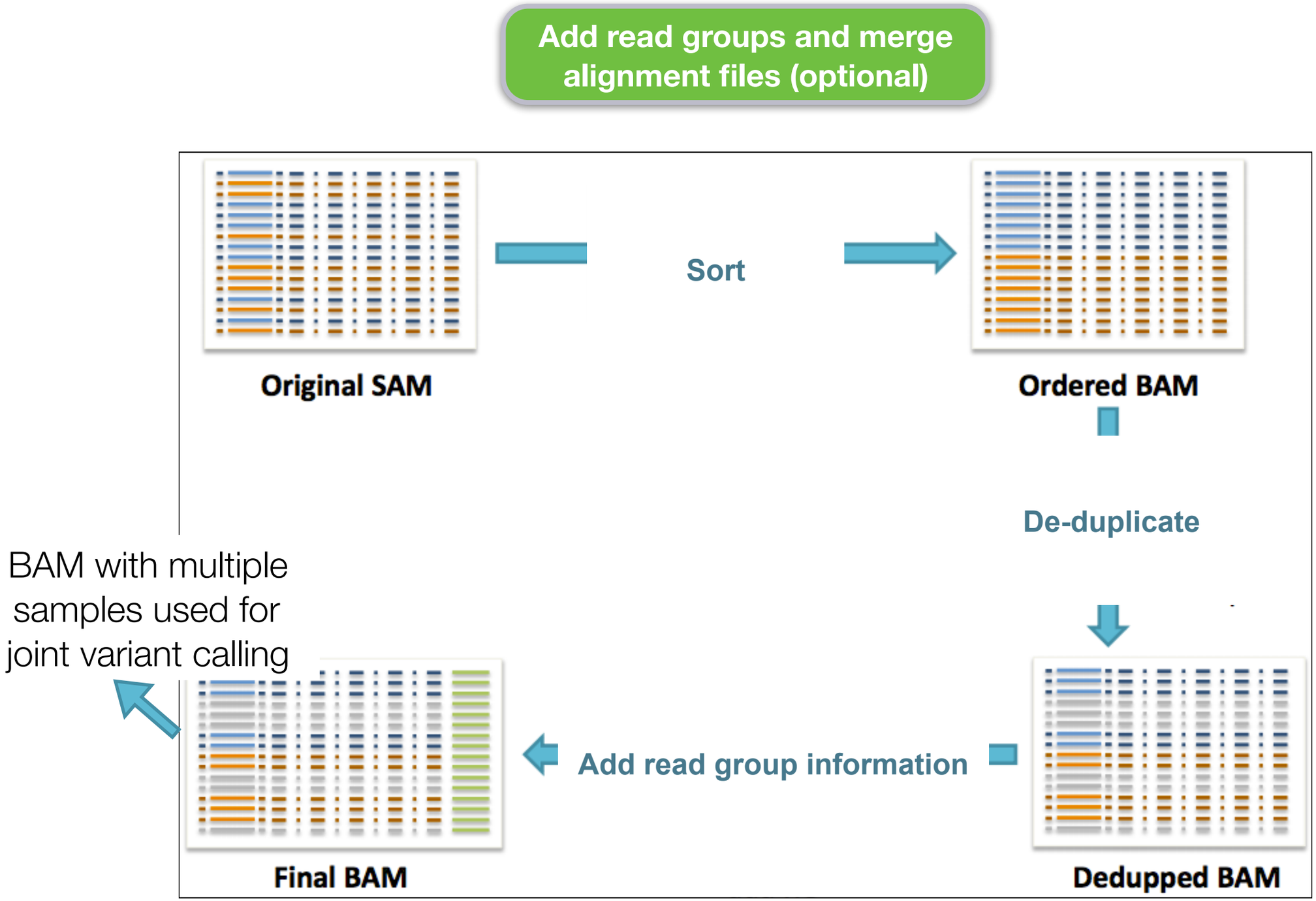
https://www.broadinstitute.org/gatk/guide/presentations?id=3391

Deduplicate



Generalized Variant Calling Workflow

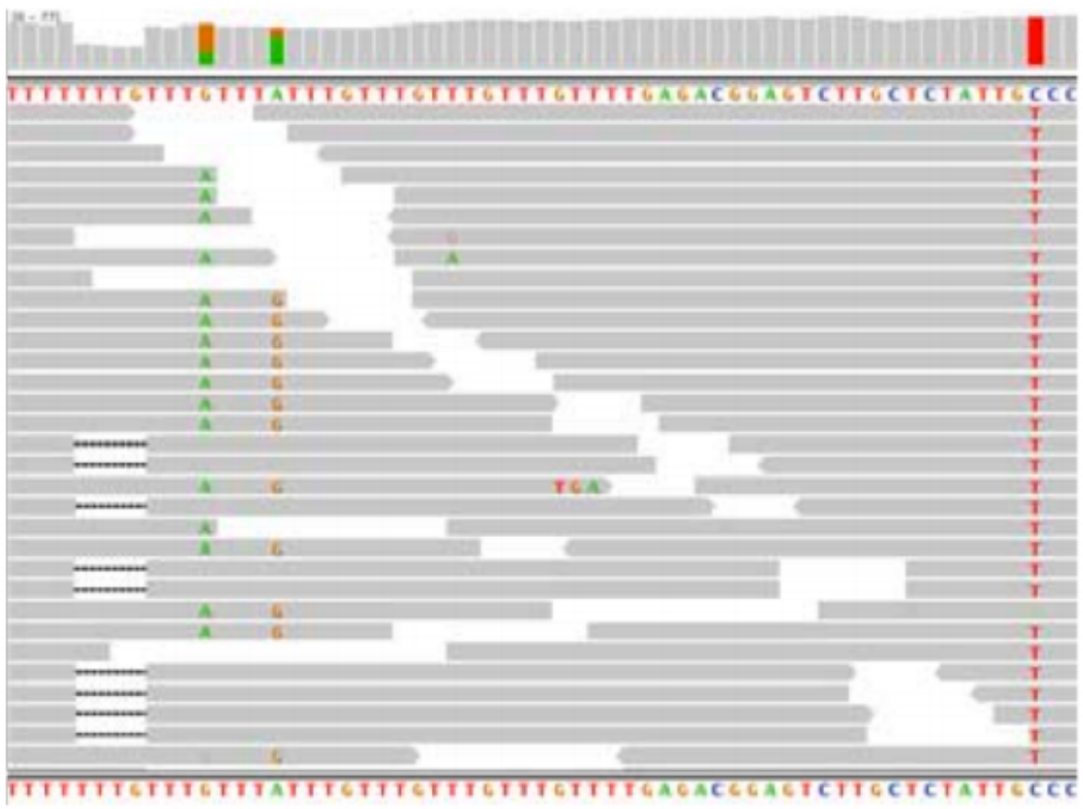
https://www.broadinstitute.org/gatk/guide/presentations?id=3391



Generalized Variant Calling Workflow

https://www.broadinstitute.org/gatk/guide/presentations?id=3391

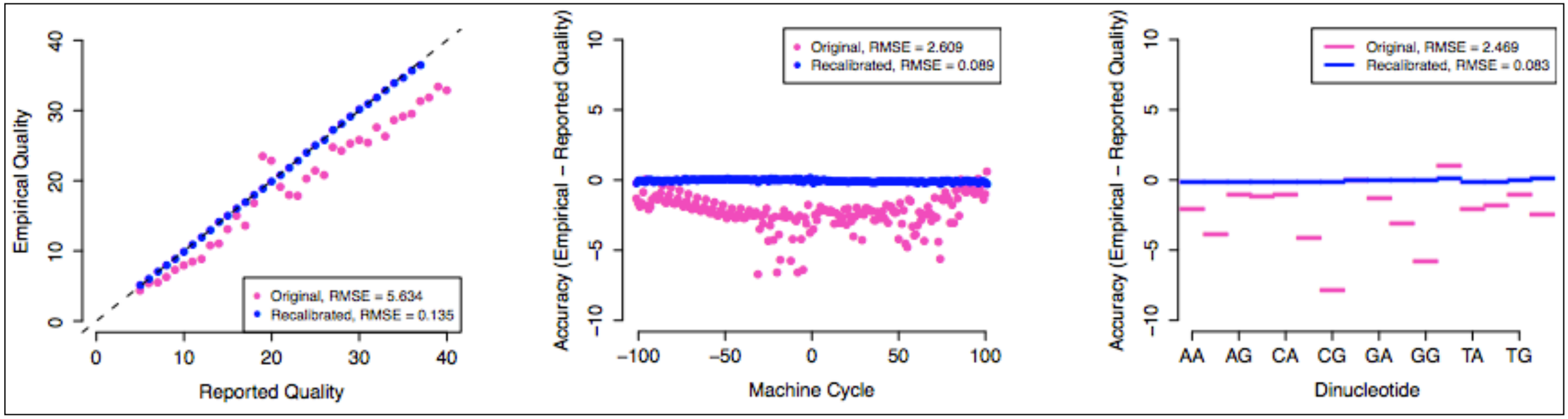
Indel realignment (optional)



Generalized Variant Calling Workflow

Base Recalibration (optional)

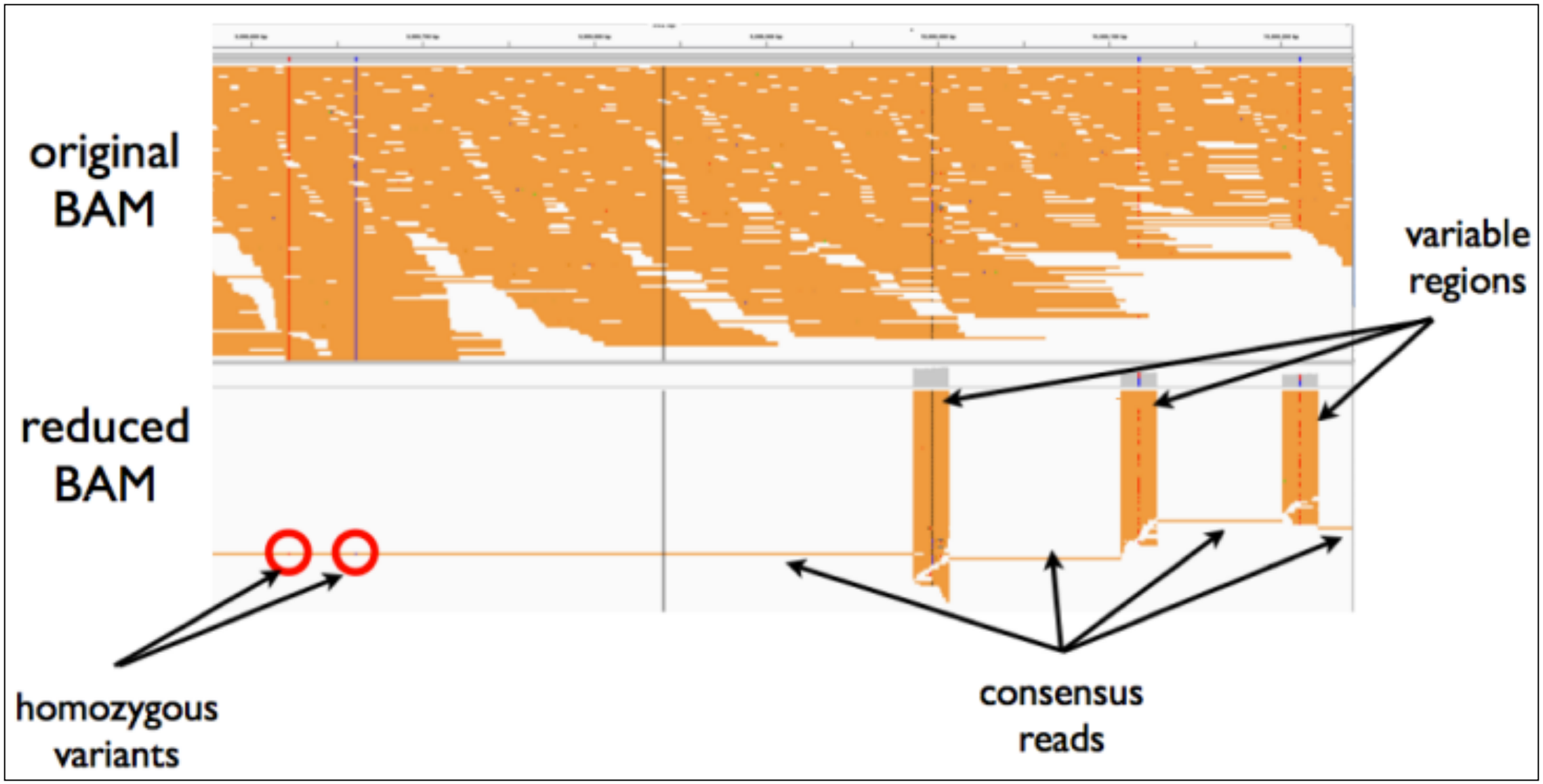
This step removes any systematic biases the creep in during sequencing



Generalized Variant Calling Workflow

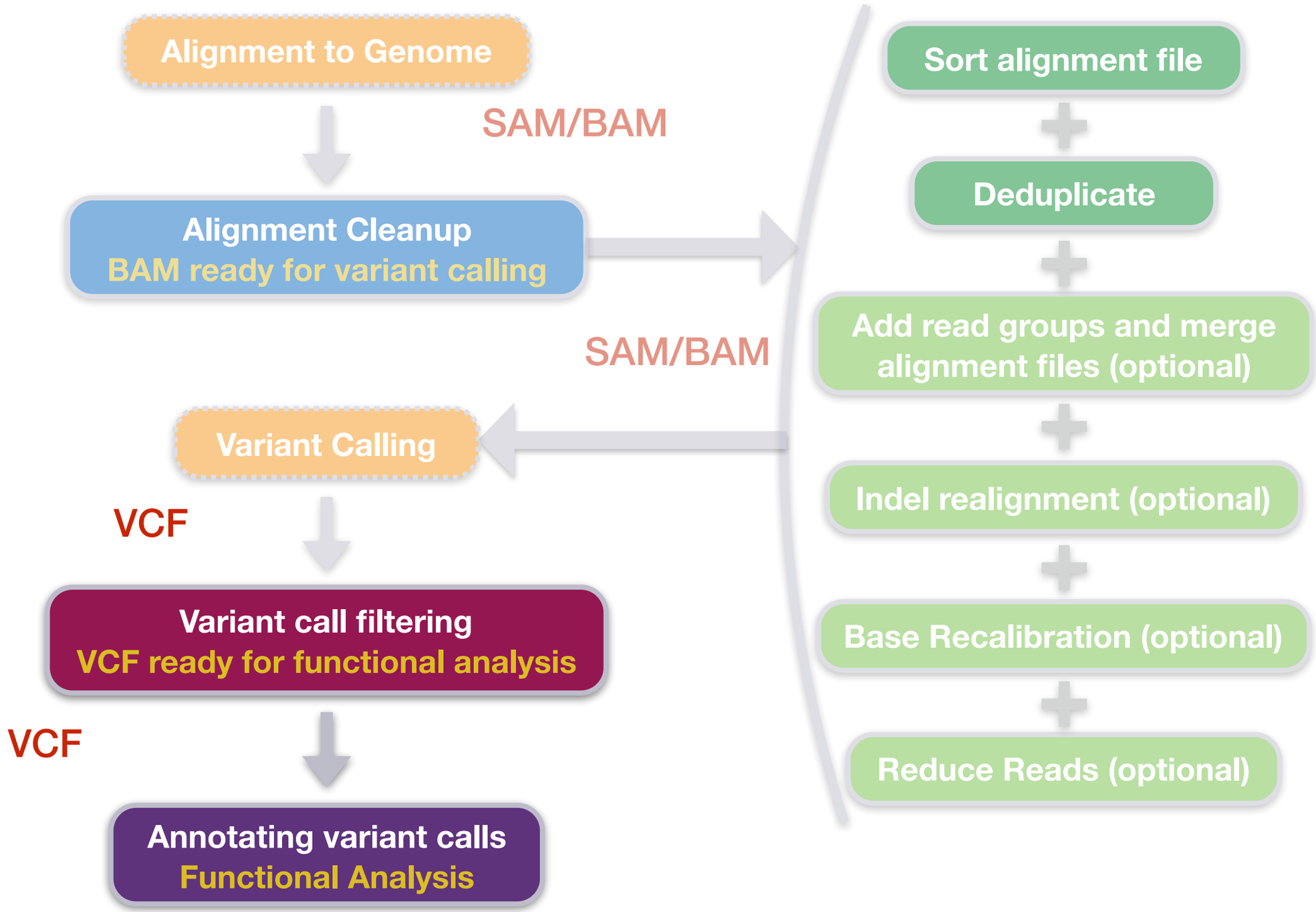
https://www.broadinstitute.org/gatk/guide/presentations?id=3391

Reduce Reads (optional)

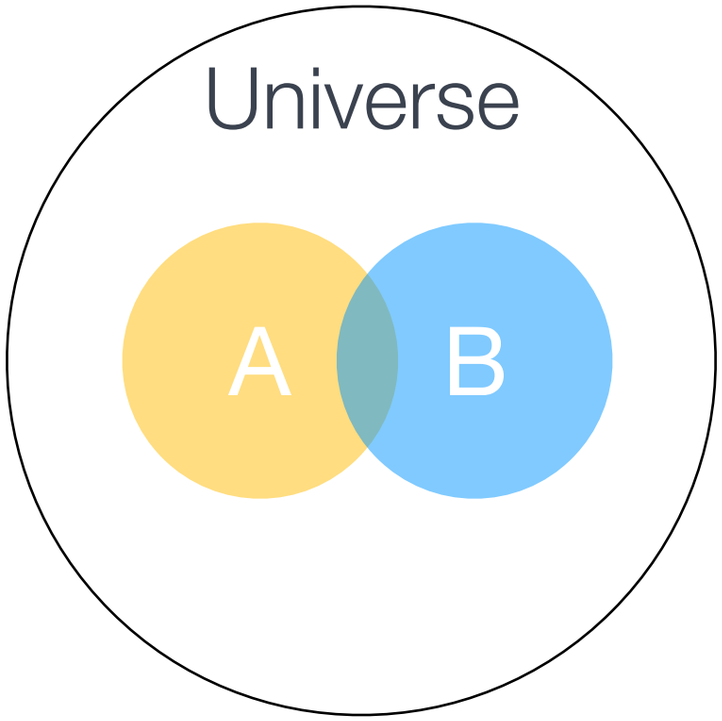


Generalized Variant Calling Workflow

https://www.broadinstitute.org/gatk/guide/presentations?id=3391



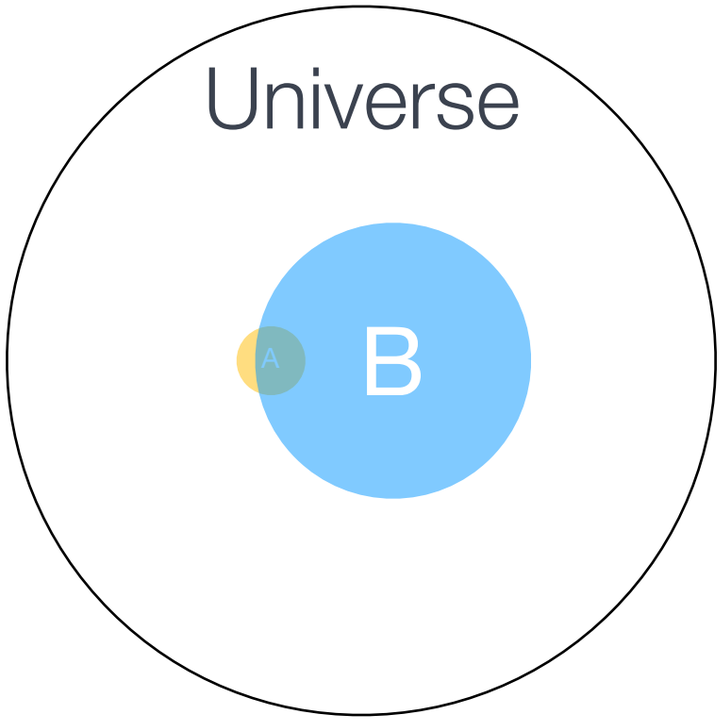
Generalized Variant Calling Workflow



Probability of real variant “A” given observed variants “B” in alignments

Variant detection

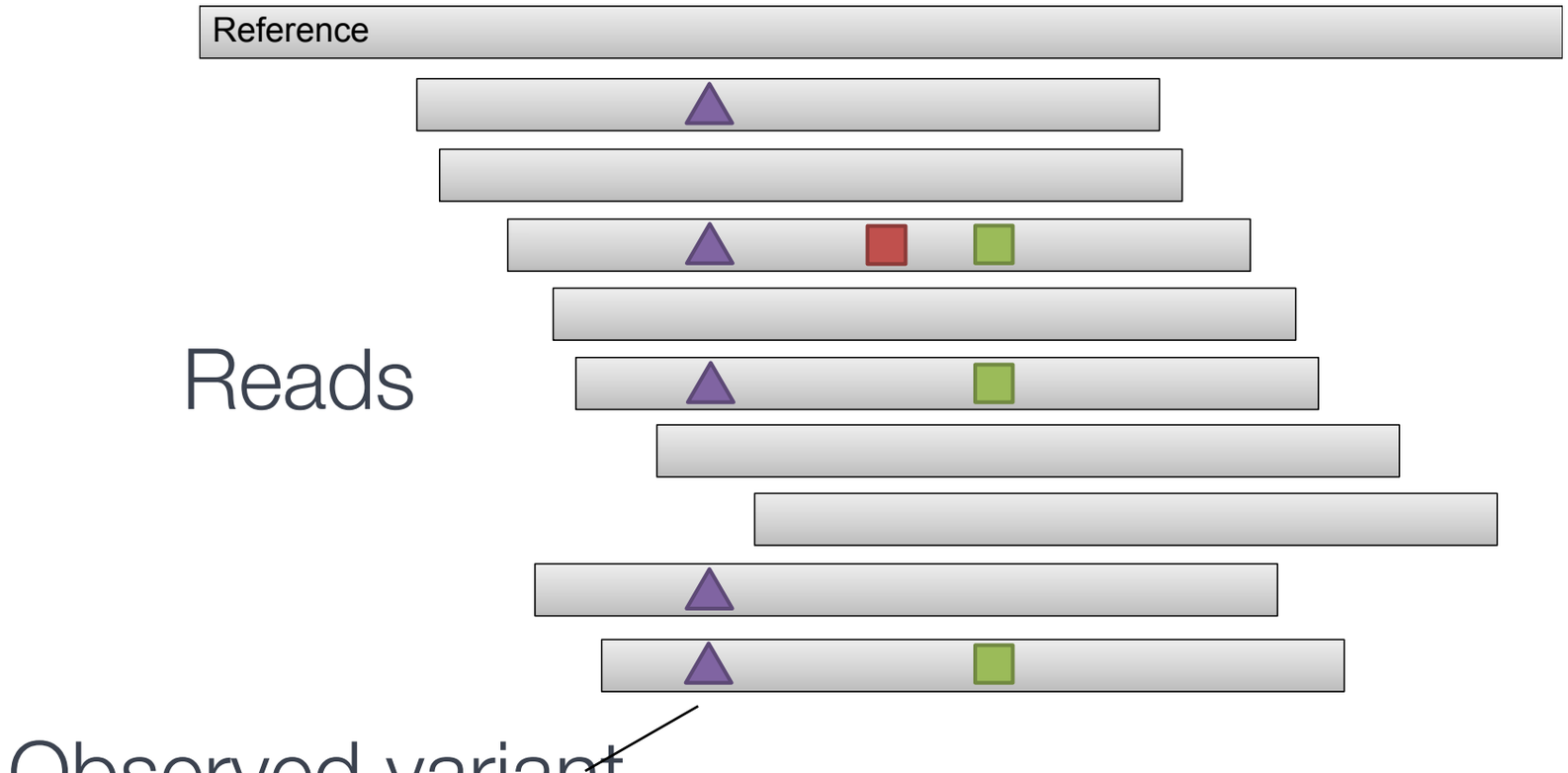
From Erik Garrison’s CSHL Advanced Sequencing Tutorial, http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html



Probability of real variant “A” given observed variants “B” in alignments

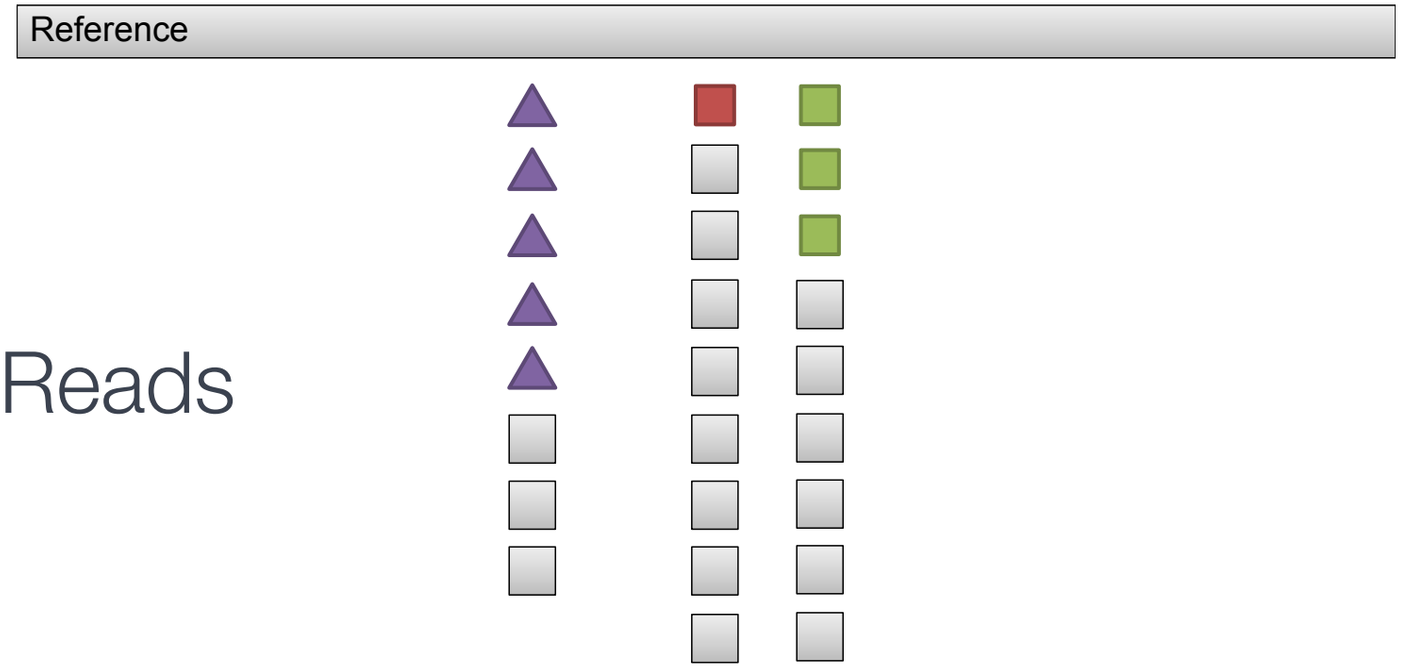
Variant detection

From Erik Garrison’s CSHL Advanced Sequencing Tutorial, http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html



Finding variants one at a time

From Erik Garrison's CSHL Advanced Sequencing Tutorial, http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html



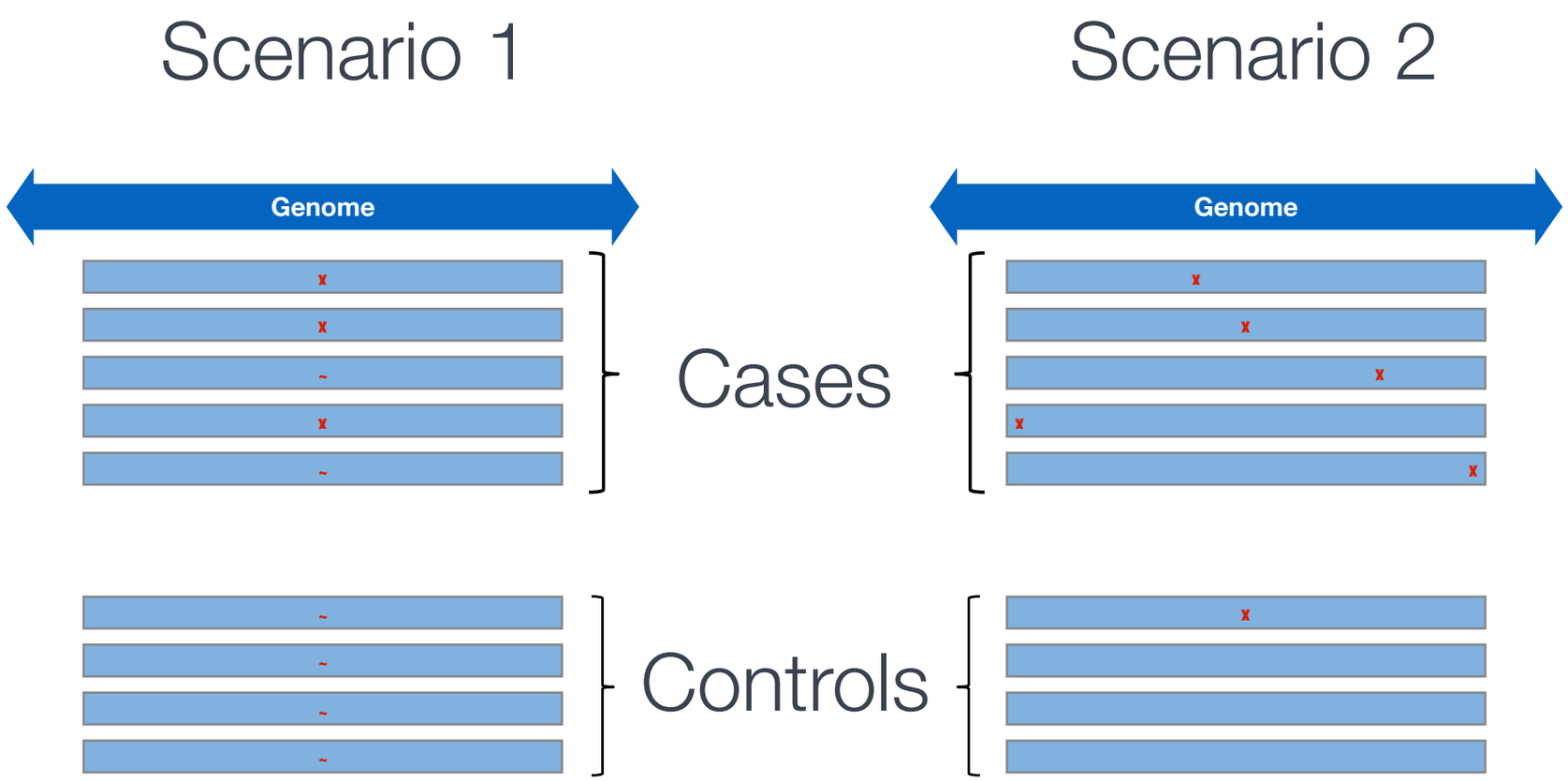
Haplotype information is lost

Finding variants one at a time

From Erik Garrison's CSHL Advanced Sequencing Tutorial, http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

Finding variants the FreeBayes way

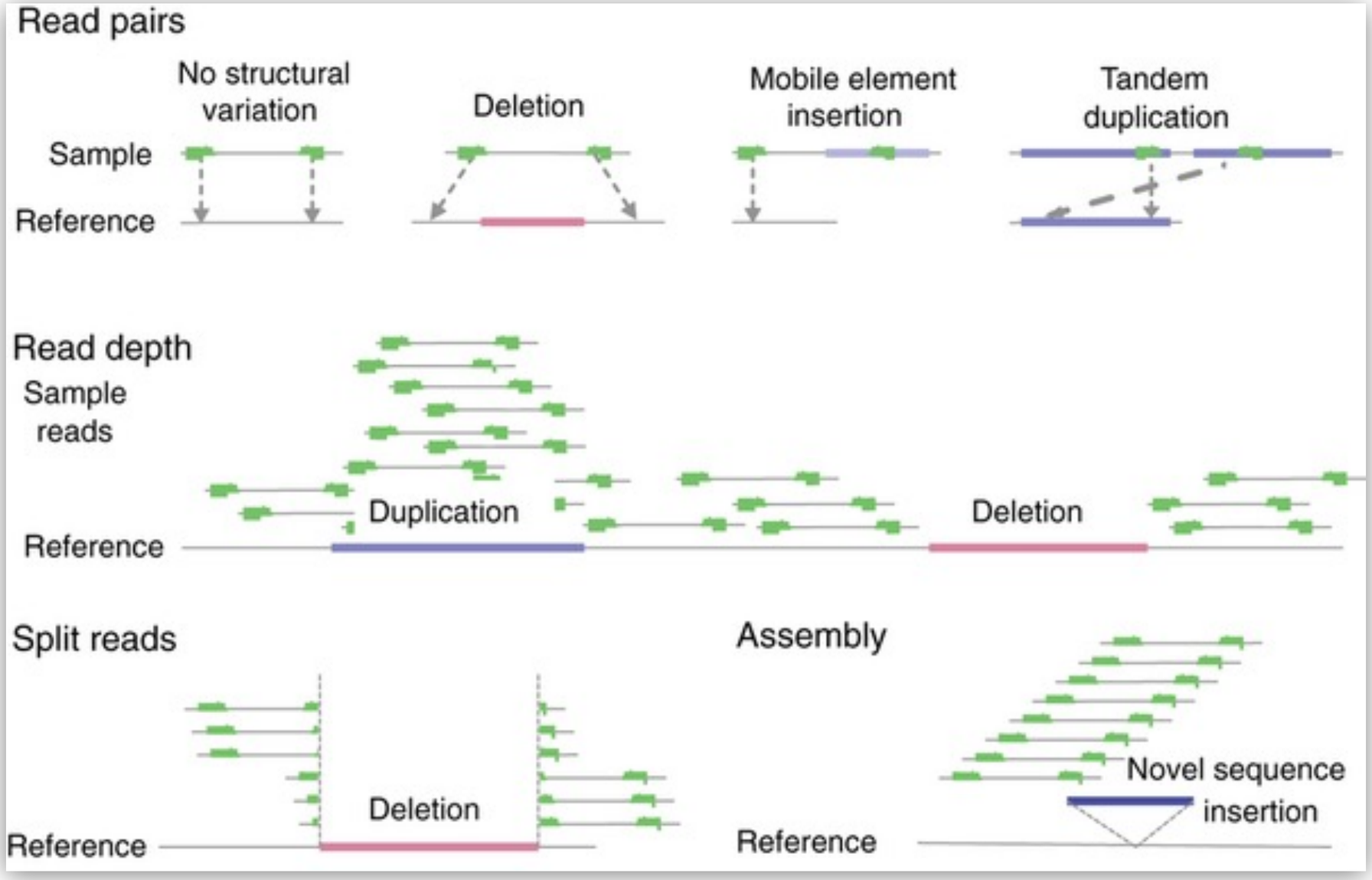
From Erik Garrison's CSHL Advanced Sequencing Tutorial, http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html



How many samples does one need for a given population study?

Experimental considerations

CNVs and SVs



CNV and SV detection is more complex

Monya Baker, Nature Methods 9, 133–137 (2012)