Variant Calling

BCB 5200 Introduction Bioinformatics I

Fall 2017

Tae-Hyuk (Ted) Ahn

Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

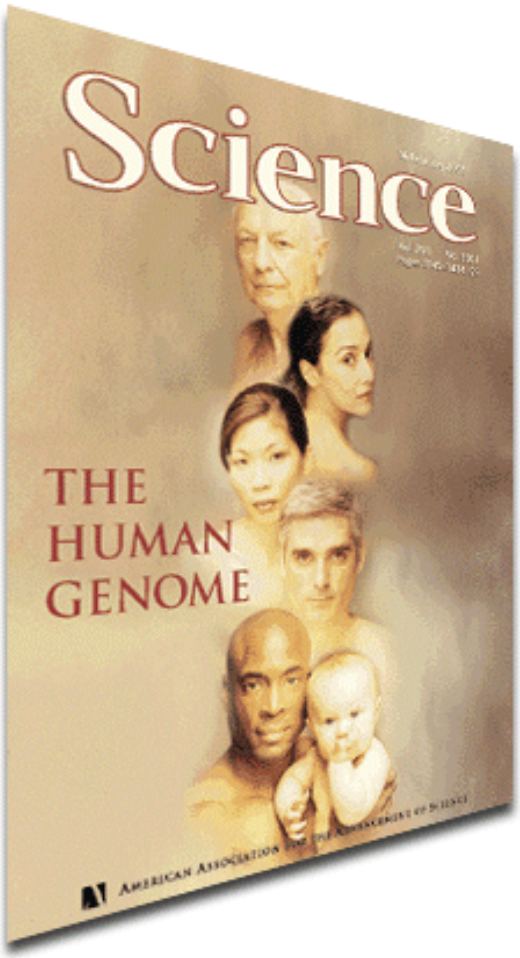
SAINT LOUIS UNIVERSITY™

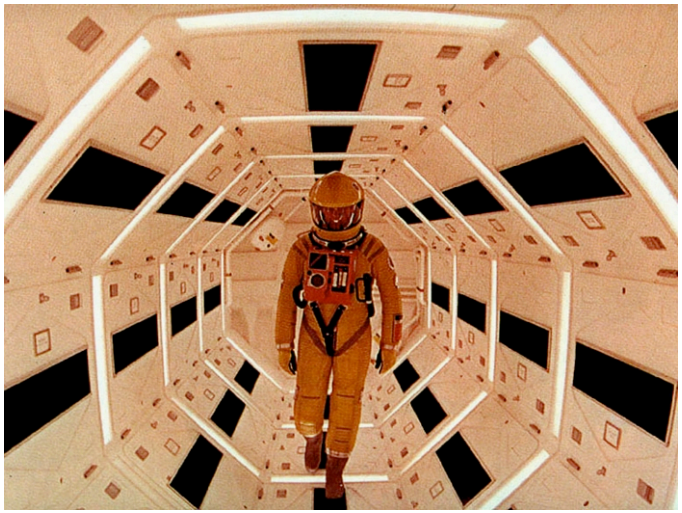
EST. 1818

The Human Genome

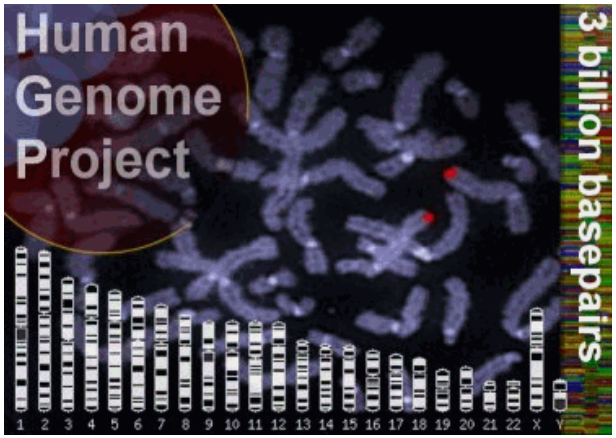
- 3 billions base pairs (ATGC)
- 20,000 protein-coding genes
- 99.6% inter-individual identity (yet 4 millions differences)
- 99% identical to chimpanzee genome (yet 6% different genes)

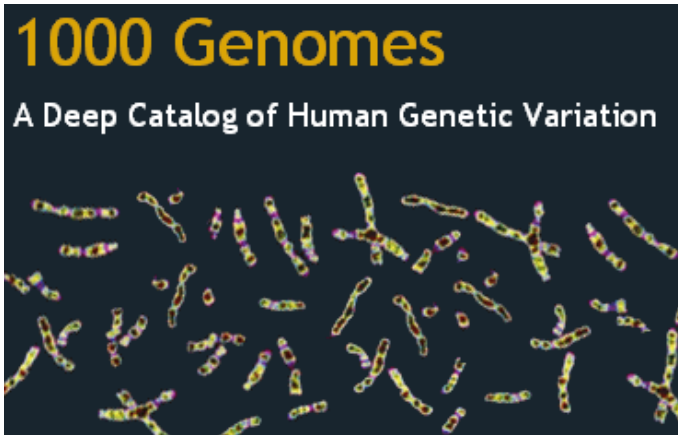
The Human Genome





Exploring the Human Genome





2002 2008

Sanger sequencing, targeted genotyping

Genome-wide genotyping (GWAS)

Exome sequencing

Genome sequencing

International HapMap Project

- The International HapMap Project (launched in 2002) was an organization that aimed to develop a haplotype map (HapMap) of the human genome, to describe the common patterns of human genetic variation.
- HapMap is used to find genetic variants affecting health, disease and responses to drugs and environmental factors.
- Canada, China, Japan, Nigeria, the United Kingdom, and the United States.

1000 Genome Project

- Whole genome sequencing and complete description of human genetic diversity in >1000 individuals from multiple world populations
- http://www.1000genomes.org

The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

What is Cancer Genomics?

- In cancer cells, small changes in the genetic letters can change what a genomic word or sentence means.
- A changed letter can cause the cell to make a protein that doesn't allow the cell to work as it should.
- These proteins can make cells grow quickly and cause damage to neighboring cells.
- By studying the cancer genome, scientists can discover what letter changes are causing a cell to become a cancer.
- The genome of a cancer cell can also be used to tell one type of cancer from another.

Short Video

How to sequence the human genome - Mark J. Kiel

http://ed.ted.com/lessons/how-to-sequence-the-human-genome-mark-j-kiel

The Cancer Genome Atlas (TCGA)

- The Cancer Genome Atlas (TCGA) is a project, begun in 2005, to catalogue genetic mutations responsible for cancer, using genome sequencing and bioinformatics.
- TCGA applies high-throughput genome analysis techniques to improve our ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.
- http://cancergenome.nih.gov/

Types of Genetic Variation

Cancer is driven by genomic alterations like:

- Single Nucleotide Aberrations
  - Single Nucleotide Polymorphisms (**SNP**s) - mutations shared amongst a population
  - Single Nucleotide Variations (**SNV**s) - private mutations
- Short Insertions or Deletions (indels)
- Copy Number Variations (CNVs)
- Larger Structural Variations (SVs)

SNPs vs. SNVs

Both are aberrations at a single nucleotide

- **SNP**
  - Aberration expected at the position for any member in the species (well-characterized)
  - Occur in population at some frequency so expected at a given locus
  - Validated in population
  - Catalogued in dbSNP (http://www.ncbi.nlm.nih.gov/snp)
- **SNV**
  - Aberration seen in only one individual (not well characterized)
  - Occur at low frequency so not common
  - Not validated in population

Really a matter of frequency of occurrence

SNVs of interest

- **Non-synonymous** mutations
  - Result in amino acid change
  - Impact protein sequence
  - Missense, nonsense, stop gained/lost mutations
- **Somatic** mutations in cancer
  - Tumor-specific mutations

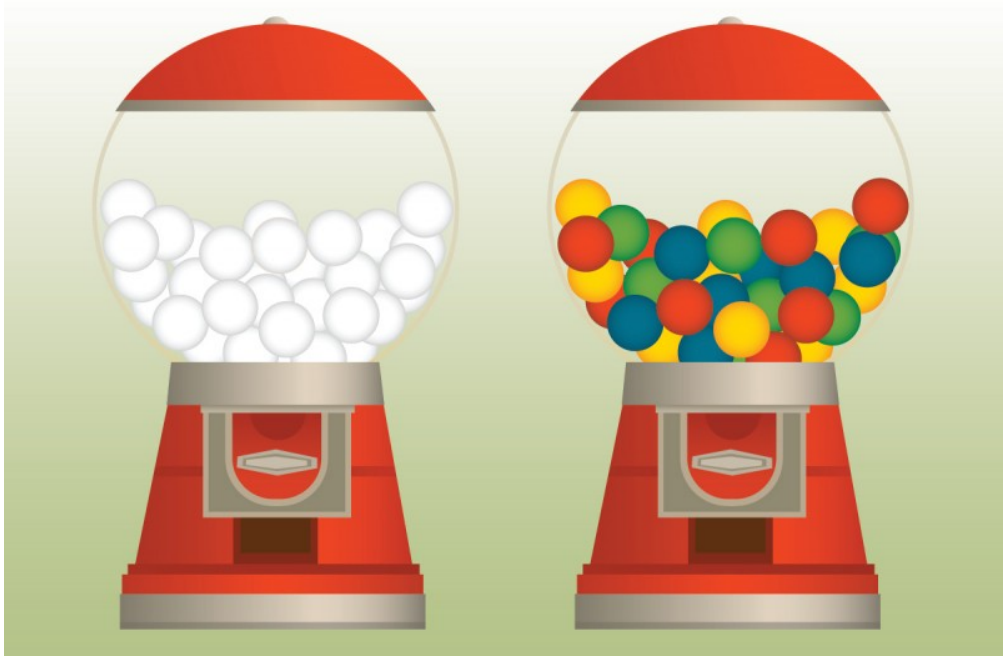
Challenges of accurate somatic variant calling

Not as simple as identifying sites with a variant allele in the tumor not present in the normal

- Artifacts from PCR amplification or targeted (exome) capture
- Machine sequencing errors
- Incorrect local alignment of reads
- Tumor heterogeneity
- Tumor-normal cross-contamination

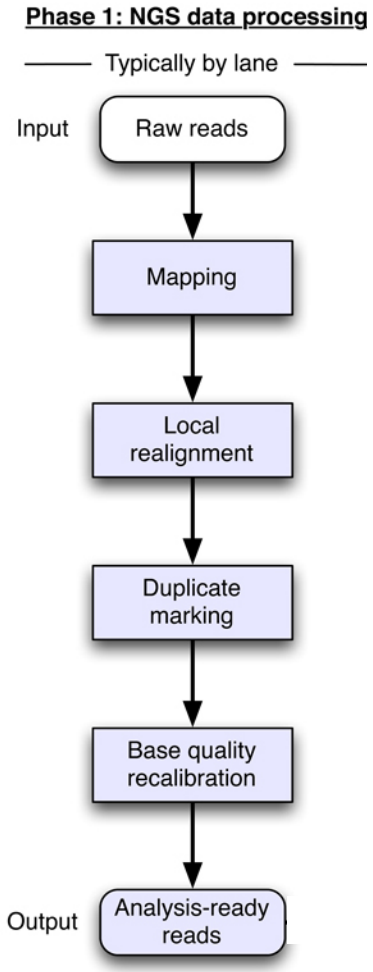
What Is Tumor Heterogeneity?

- Until recently, the cells within a tumor were thought to be similar to one another at any given stage of the cancer, like the white gumballs in the machine on the left.
- But scientists are finding that a person's tumor cells can be highly diverse. This could mean that a biopsy may capture only a fraction of tumor cells that are not representative of the whole tumor, like a ball ejected at random from the machine on the right



https://www.mskcc.org/blog/what-tumor-heterogeneity

A framework for variation discovery



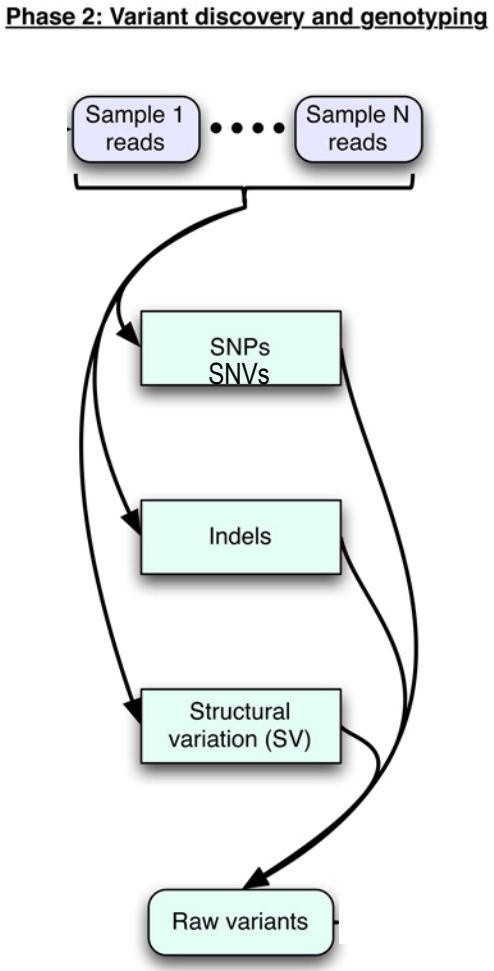
Phase 1: Mapping

- Place reads with an initial alignment on the reference genome using mapping algorithms
- Refine initial alignments
  - local realignment around indels
  - molecular duplicates are eliminated
- Generate the technology-independent SAM/BAM alignment map format

Accurate mapping crucial for variation discovery

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

A framework for variation discovery

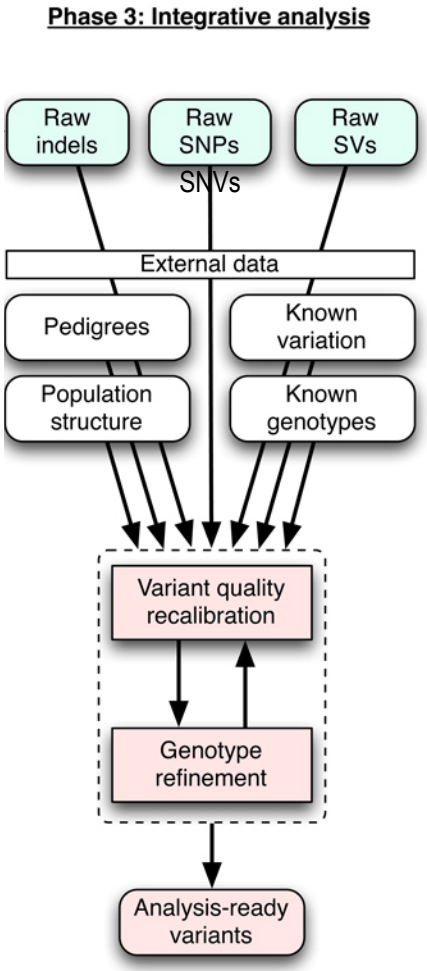


Phase 2: Discovery of raw variants

- Analysis-ready SAM/BAM files are analyzed to discover all sites with statistical evidence for an alternate allele present among the samples
- SNPs, SNVs, short indels, and SVs

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

A framework for variation discovery

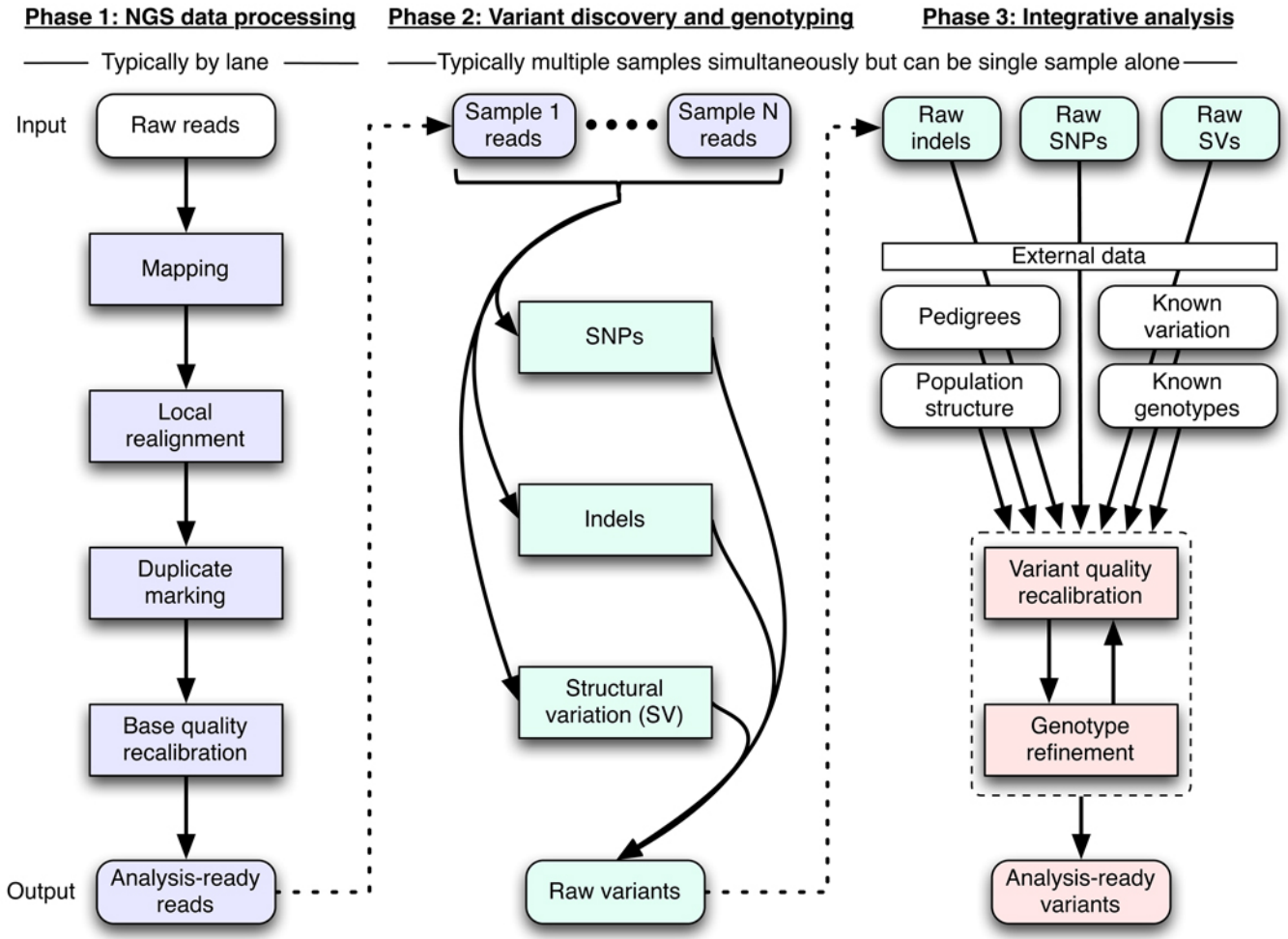


Phase 3: Discovery of analysis-ready variants

- technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium, and family and population structure are integrated with the raw variant calls from Phase 2 to separate true polymorphic sites from machine artifacts
- at these sites high-quality genotypes are determined for all samples

DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

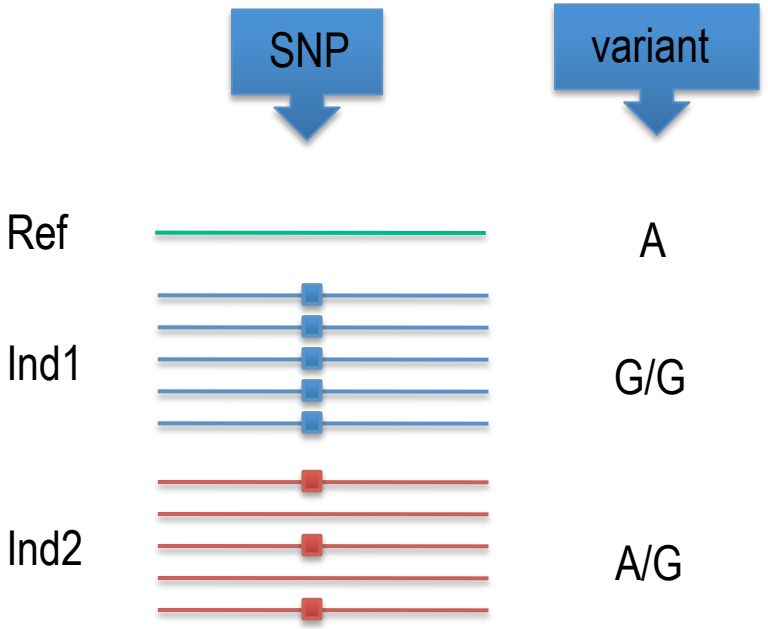
A framework for variation discovery



DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).

Variant calling methods

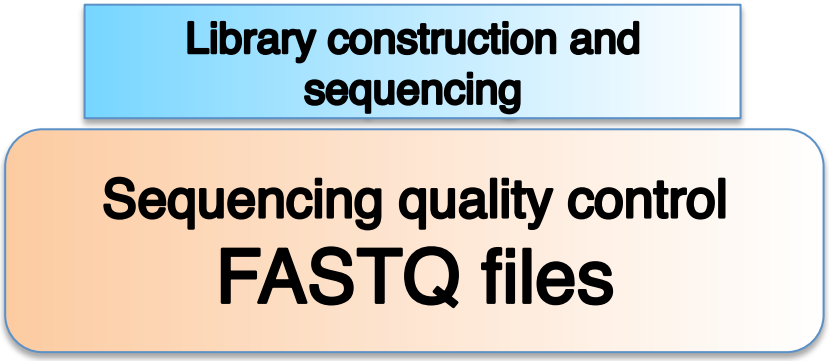
- > 15 different algorithms
- Three categories
  - Allele counting
  - Probabilistic methods, e.g. Bayesian model
    - to quantify statistical uncertainty
    - Assign priors based on observed allele frequency of multiple samples
  - Heuristic approach
    - Based on thresholds for read depth, base quality, variant allele frequency, statistical significance



Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun; 12(6):443-51. PMID: 21587300.

http://seqanswers.com/wiki/Software/list

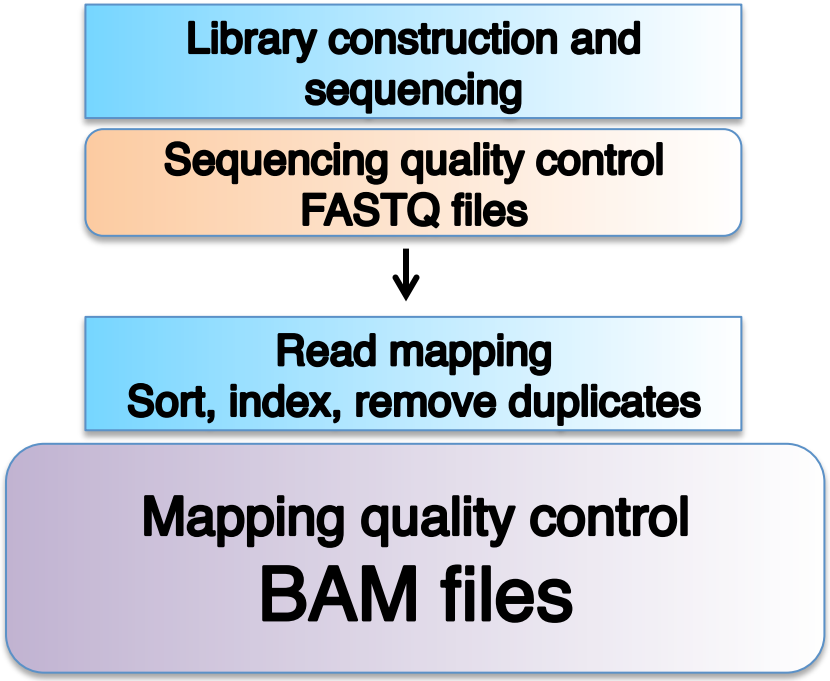
Sequencing




```
@G:1:1:11:1079#0/1
TGATTGATTCCATTCCATTCCATTCCATTTCATTCCATTGCAATCCCTTCCAATCCATTCCATTCCATTCCATTC

+G:1:1:11:1079#0/1
`Xa^YO\_^a_`__`a__^a^a^_a``^_\`\\]``[XUGXXXXXWUTWWVWUSTXXPUWYYRVWYYYXZYXYWZ
```

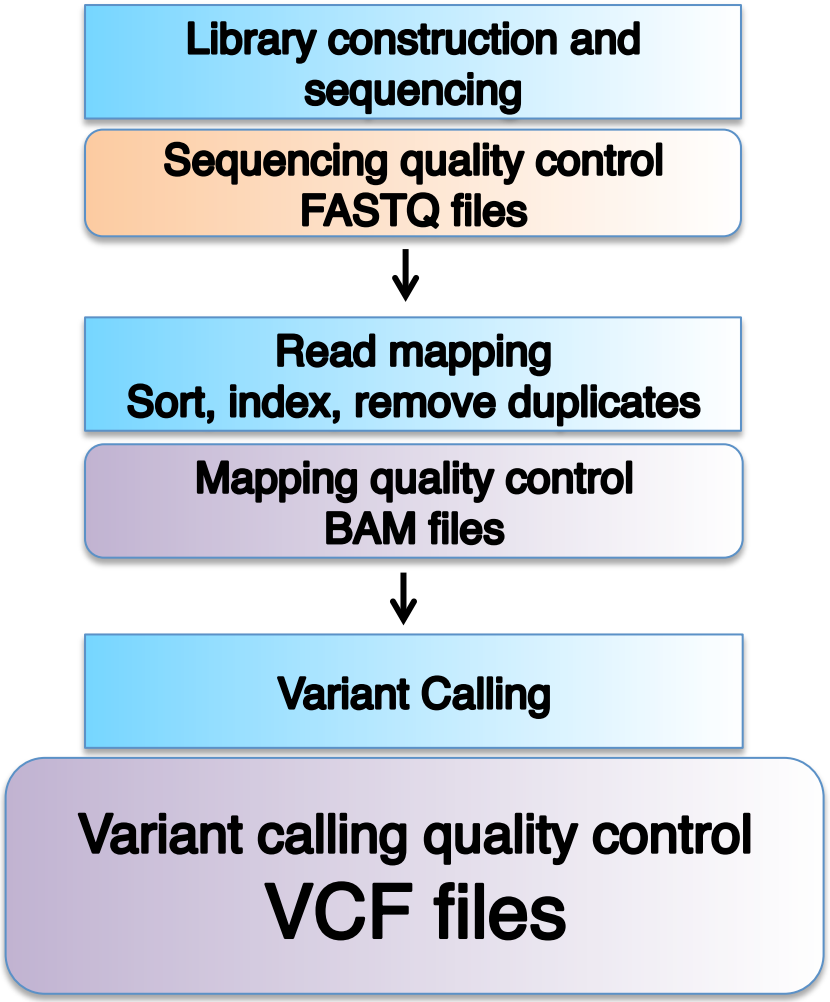
Mapping


Library construction and sequencing
Sequencing quality control
FASTQ files
Read mapping
Sort, index, remove duplicates
Mapping quality control
BAM files

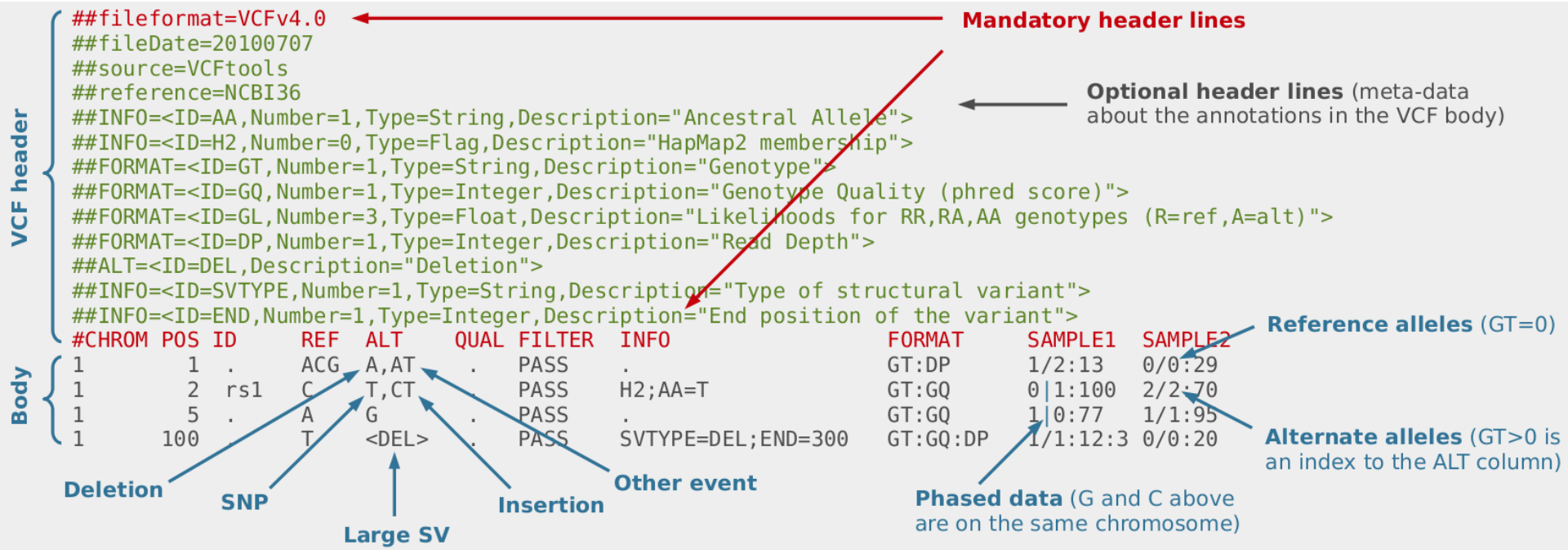
Mapping


711 721 731 741 751 761 771 781 791
GTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTGTTAGAAGCGCGTGGTCACAACGTTACCGTTATCGATCCGGTCGAAAAACTGCTGGCAGT

Variant Calling



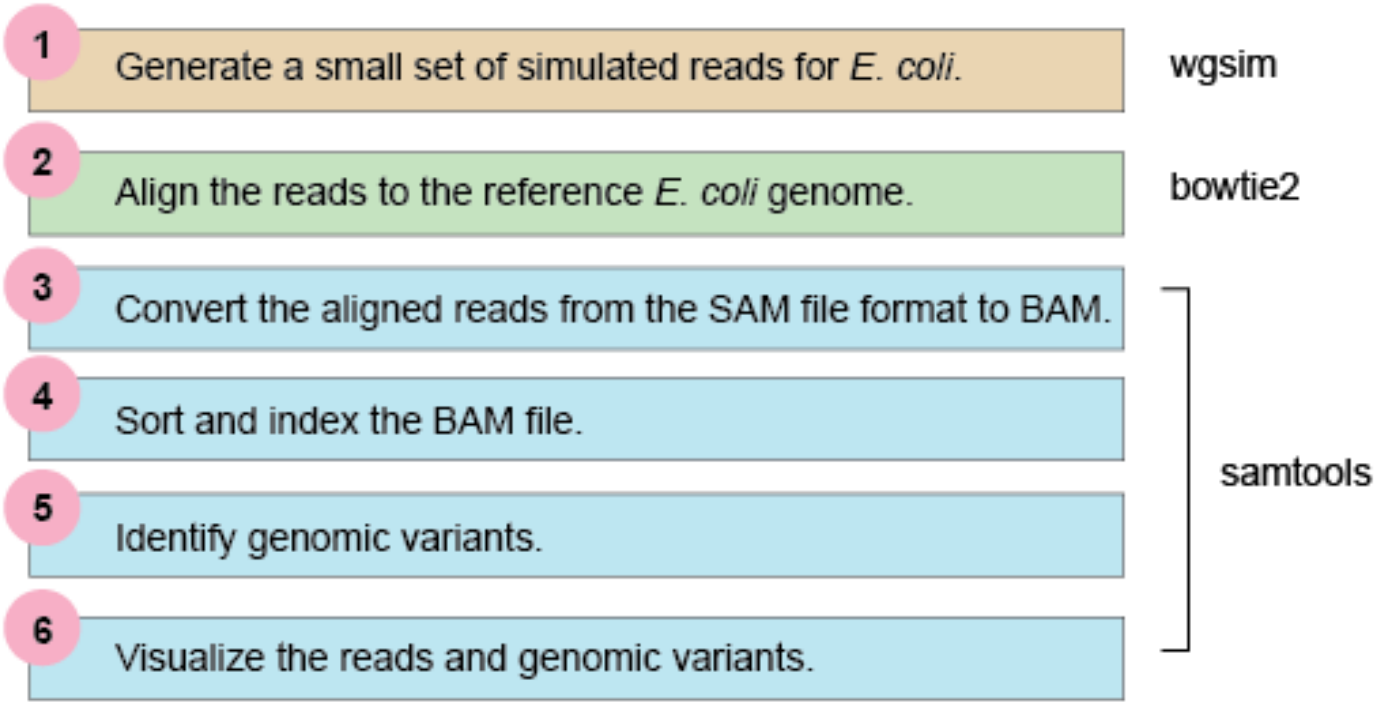
Variant Call Format (VCF)



Tools for Variant Calling

- Alignmer
  - Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/)
  - BWA (http://bio-bwa.sourceforge.net/)
- Tools for variant calling
  - SAMtools & BCFtools (http://www.htslib.org/download/)
  - VCFTools (https://vcftools.github.io/index.html)
  - GATK (https://www.broadinstitute.org/gatk/)
- Browser
  - SAMtools
  - IGV (https://www.broadinstitute.org/igv/)
  - UCSC Genome Browser (https://genome.ucsc.edu/)

Variant Calling Lab



Generating Simulated Reads

- To download wgsim (https://github.com/lh3/wgsim)
  - You can download zip file
  - You can download as

```
$ git clone https://github.com/lh3/wgsim.git
```

  - You don't need to install if the wgsim is on the system. Check it.
- How can you add the software path into your environment to run the program easily?
  - Open your ~/.bashrc profile and add the software path

```
$ pwd
/faculty/ahnt/Courses/BCB5250/software/wgsim
$ echo 'PATH=/faculty/ahnt/Courses/BCB5250/software/wgsim:$PATH' >> ~/.bashrc
$ . ~/.bashrc
```

Generating Simulated Reads

- First, we need a small set of sample read data.
- The command line usage for wgsim is:
- wgsim [options] <in.ref.fa> <out.read1.fq> <out.read2.fq>
- By default, wgsim reads in a reference genome in the **FASTA** format, and generates simulated **paired-end** reads in the **FASTQ** format.
- E. coli genome (NC_008253.fna) is at /tmp directory. Copy the file into your Lab02_VariantCalling directory.

```
$ wgsim -N 1000000 NC_008253.fna sim_reads_1M.fq
```

Aligning Reads using Bowtie2

- Download Bowtie2 (before download, check it is already installed)

```
$ wget https://downloads.sourceforge.net/project/bowtie-bio/
bowtie2/2.3.0/bowtie2-2.3.0-linux-x86_64.zip
```

- bowtie2-build to index your genome

```
$ bowtie2-build -f NC_008253.fna NC_008253.fna
```

- Run bowtie2

```
$ bowtie2 -x NC_008253.fna -U sim_reads_1M.fq -S
sim_reads_1M_aligned.sam
```

Understanding SAM Format

- As SAMtools is primarily concerned with manipulating SAM files, it is useful to take a moment to examine the sample SAM file generated by bowtie, and to dive into the details of the SAM file format itself. The first six lines from the bowtie SAM file are extracted below:

Headers

Alignments

```
1 @HD VN:1.0  SO:unsorted
2 @SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
3 @PG ID:bowtie2  PN:bowtie2  VN:2.1.0
4 gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
  008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
  ATCACCAACCATCTGGTAGCGATGAT  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
  YT:Z:UU
5 gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
  008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
  CCCCCGCCAAAATCACCAACCATCTG  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
  :UU
6 gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
  008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
  GACGCGTACAGGAAACACCGAAAAAA  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
      YT:Z:UU
```

Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

Understanding SAM Format

Field Name	Description	Example from the e_coli SAM file
QNAME	Unique identifier of the read; derived from the original FASTQ file.	gi|110640213|ref |NC_008253.1 | _418_952_1:0:0_1:0:0_0/1
FLAG	A single integer value (e.g. 16), which encodes multiple elements of meta-data regarding a read and its alignment. Elements include: whether the read is one part of a paired-end read, whether the read aligns to the genome, and whether the read aligns to the forward or reverse strand of the genome. A useful online utility decodes a single SAM flag value into plain English.	16: indicates that the read maps to the reverse strand of the genome.
RNAME	Reference genome identifier. For organisms with multiple chromosomes, the RNAME is usually the chromosome number; for example, in human, an RNAME of "chr3" indicates that the read aligns to chromosome 3. For organisms with a single chromosome, the RNAME refers to the unique identifier associated with the full genome; for example, in *E. coli*, the RNAME is set to: gi|110640213|ref|NC_008253.1|.	gi|110640213|ref|NC_008253.1|
POS	Left-most position within the reference genome where the alignment occurs.	418
MAPQ	Quality of the genome mapping. The MAPQ field uses a **Phred-scaled probability** value to indicate the probability that the mapping to the genome is incorrect. Higher values indicate increased confidence in the mapping.	42: indicates low probability (6.31e-05) that the mapping is incorrect.
CIGAR	A compact string that (partially) summarizes the alignment of the raw sequence read to the reference genome. Three core abbreviations are used: M for alignment match; I for insertion; and D for Deletion. For example, a CIGAR string of 5M2I63M indicates that the first 5 base pairs of the read align to the reference, followed by 2 base pairs, which are unique to the read, and not in the reference genome, followed by an additional 63 base pairs of alignment. Note that the CIGAR string is a partial representation of the alignment because M indicates an alignment match, but this could indicate an exact sequence match or a mismatch [2]. If you would like to determine match v. mismatch, you can consult the optional MD field, detailed below.	70M: 70 base pairs within the read match the reference genome.
RNEXT	Reference genome identifier where the mate pair aligns. Only applicable when processing paired-end sequencing data. For example, an alignment with RNAME of "chr3" and RNEXT of "chr4" indicates that the mate and its pair span two chromosomes, indicating a possible structural rearrangement. A value of * indicates that information is not available.	the *E. coli* simulated data is single read sequencing data, and all RNEXT values are therefore set to * (no information available).
PNEXT	Position with the reference genome, where the second mate pair aligns. As with RNEXT, this field is only applicable when processing paired-end sequencing data. A value of 0 indicates that information is not available.	the *E. coli* simulated data is single read sequencing data, and all RNEXT values are therefore set to 0 (no information available).
TLEN	Template Length. Only applicable for paired-end sequencing data, TLEN is the size of the original DNA or RNA fragment, determined by examining both of the **paired-mates**, and counting bases from the left-most aligned base to the right-most aligned base. A value of 0 indicates that TLEN information is not available.	the *E. coli* simulated data is single read sequencing data, and all RNEXT values are therefore set to 0 (no information available).
SEQ	the raw sequence, as originally defined in the FASTQ file.	CCAGGCAGTGGCAGGT...
QUAL	The **Phred quality score** for each base, as originally defined in the FASTQ file.	.

Converting SAM to BAM

- Download Samtools (http://www.htslib.org/)
- To convert from SAM to BAM, use the SAMtools view command:

```
$ samtools view -b -S -o sim_reads_1M_aligned.bam
sim_reads_1M_aligned.sam
```

  - -b: indicates that the output is BAM.
  - -S: indicates that the input is SAM.
  - -o: specifies the name of the output file.

- BAM files are stored in a compressed, binary format, and cannot be viewed directly. However, you can use the same view command to display all alignments. For example, running:

```
$ samtools view sim_reads_1Ms_aligned.bam | more
```

Sorting and Indexing

- The next step is to sort and index the BAM file.

```
$ samtools sort sim_reads_1M_aligned.bam
sim_reads_1M_aligned.sorted
```

- Once you have sorted your BAM file, you can then index it.

```
$ samtools index sim_reads_1M_aligned.sorted.bam
```

Identifying Genomic Variants

- The first step is to use the SAMtools mpileup command to calculate the **genotype likelihoods** supported by the aligned reads in our sample:

```
$ samtools mpileup -g -f NC_008253.fna
sim_reads_1M_aligned.sorted.bam >
sim_reads_1M_variants.bcf
```

-g: directs SAMtools to output genotype likelihoods in the **binary call format (BCF)**. This is a compressed binary format.

- -f: directs SAMtools to use the specified reference genome. A reference genome must be specified, and here we specify the reference genome for *E. coli*.

Identifying Genomic Variants

- The second step is to use bcftools:

```
$ bcftools call -c -v sim_reads_1M_variants.bcf >
sim_reads_1M_variants.vcf
```

  - -c: directs bcftools to call SNPs.
  - -v: directs bcftools to only output potential variants

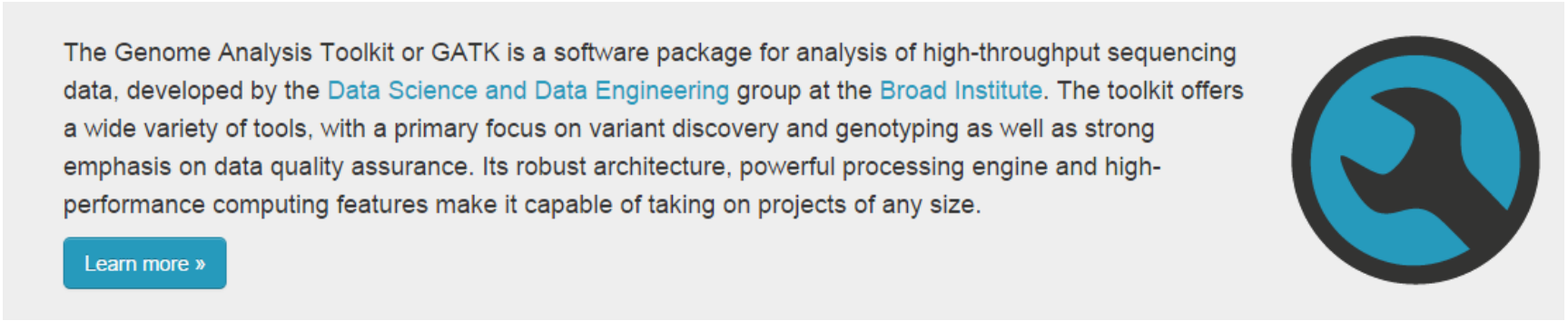
Visualizing Reads

- For the final step, we will use the SAMtools tview command to view our simulated reads and visually compare them to the *E. coli* reference genome. To do so, run:

```
$ samtools tview sim_reads_aligned.sorted.bam
NC_008253.fna
```

GATK

- https://www.broadinstitute.org/gatk/



- https://www.broadinstitute.org/gatk/guide/best-practices
- http://broadinstitute.github.io/picard/
- https://broadinstitute.github.io/picard/command-line-overview.html

Picard

- http://broadinstitute.github.io/picard/
- https://broadinstitute.github.io/picard/command-line-overview.html

- https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-A-3-GATK_Best_Practices_and_Broad_pipelines.pdf

IGV

- https://www.broadinstitute.org/igv/