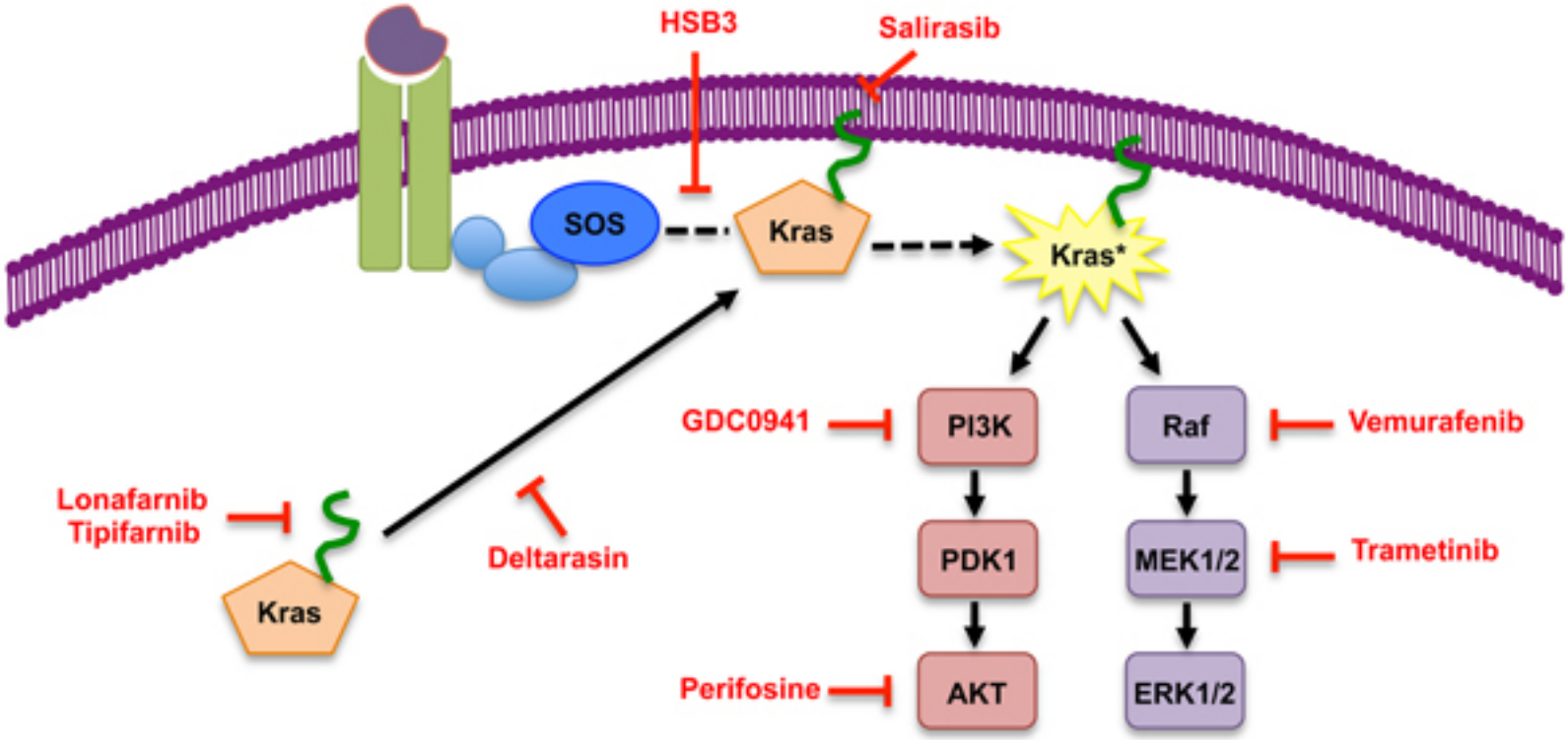
Read Mapping and Variant Calling

Johannes Starlinger

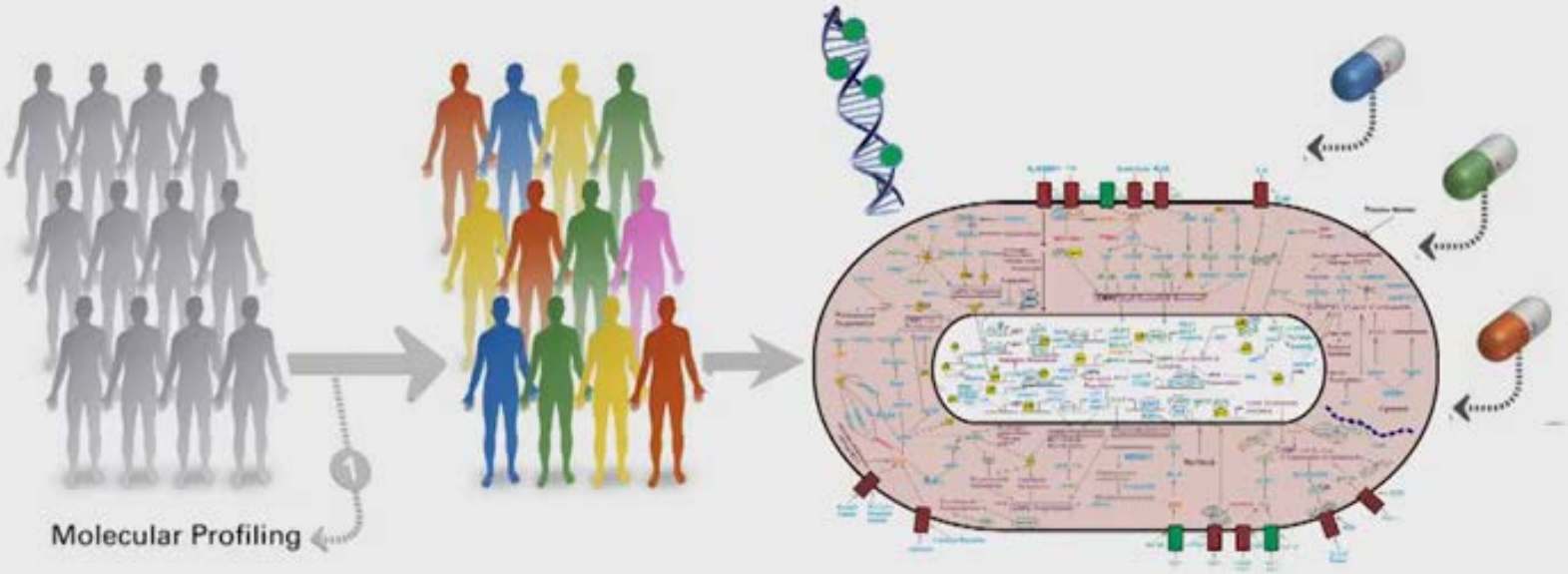
Application Scenario: Personalized Cancer Therapy

Different mutations require different therapy



Collins, Meredith A., and Marina Pasca di Magliano. "Kras as a key oncogene and therapeutic target in pancreatic cancer." *Frontiers in physiology* 4 (2013).

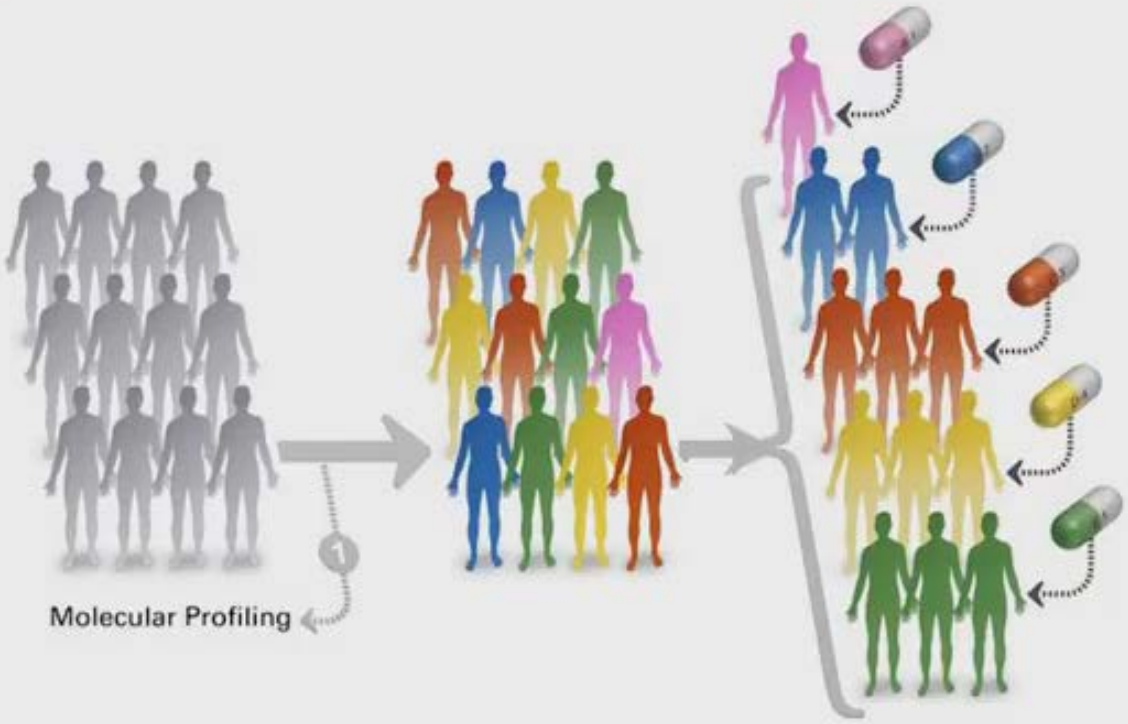
Application Scenario: Personalized Cancer Therapy



Figures: MD Anderson Cancer Center (PCT), codi.beltanenetwork.org/25-days-codi-day-9/dna-helix/ , bioinformatics.org

© Madeleine Kittner @ WBI

Application Scenario: Personalized Cancer Therapy



Figures: MD Anderson Cancer Center (PCT)

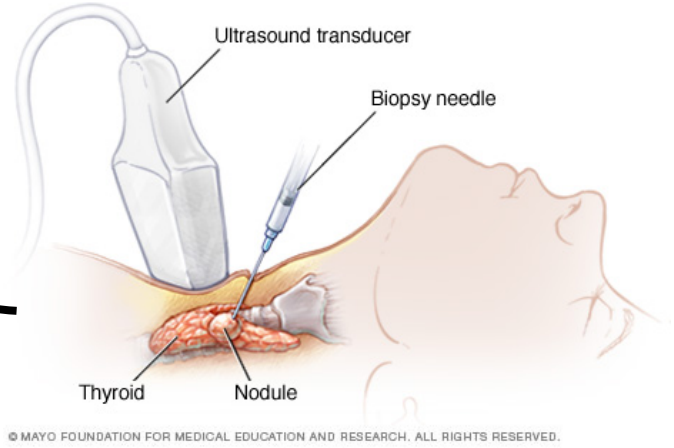
© Madeleine Kittner @ WBI

Application Scenario: Personalized Cancer Therapy

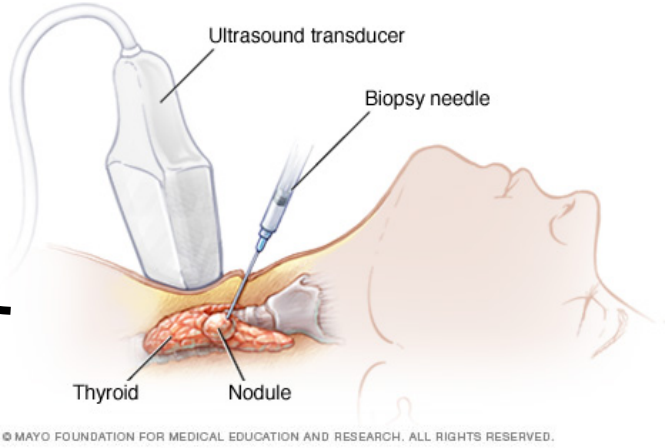
Figures: MD Anderson Cancer Center (PCT)

© Madeleine Kittner @ WBI

Molecular Profiling Workflow



Molecular Profiling Workflow



Sequencing

Reads

Read Mapping

Genome, Exome, Panel

Variant Calling

Genes & Mutations

Filter & Rank

Filtered List w/ Top-X Candidates

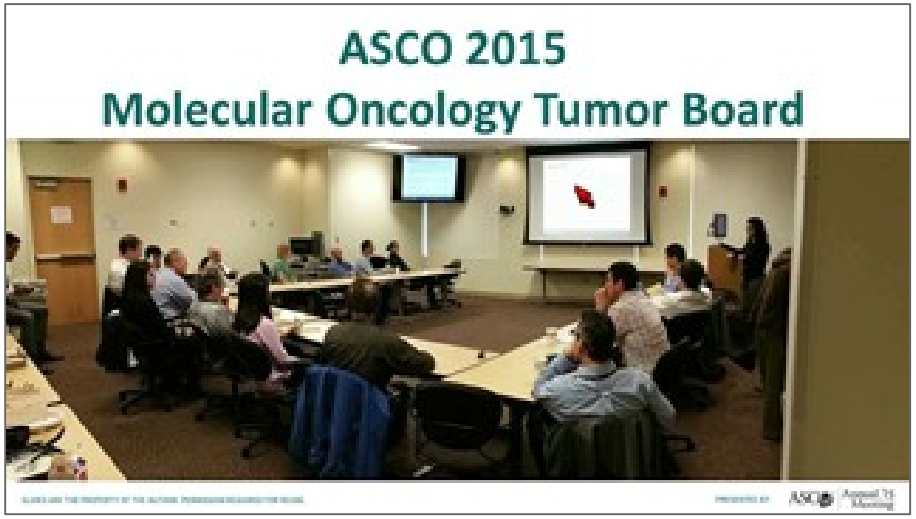
Find known Targets & Evidences

Target Candidates

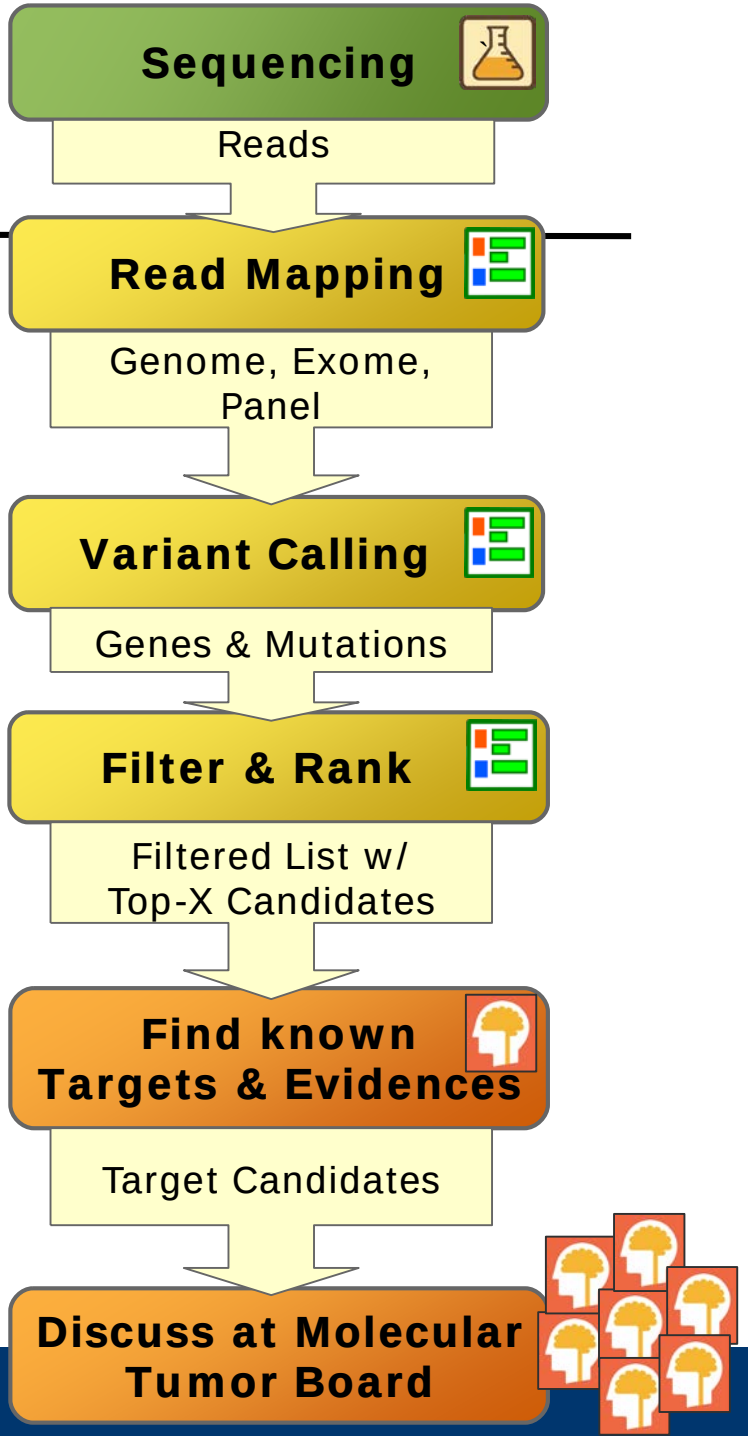
Discuss at Molecular Tumor Board

Molecular Profiling Workflow

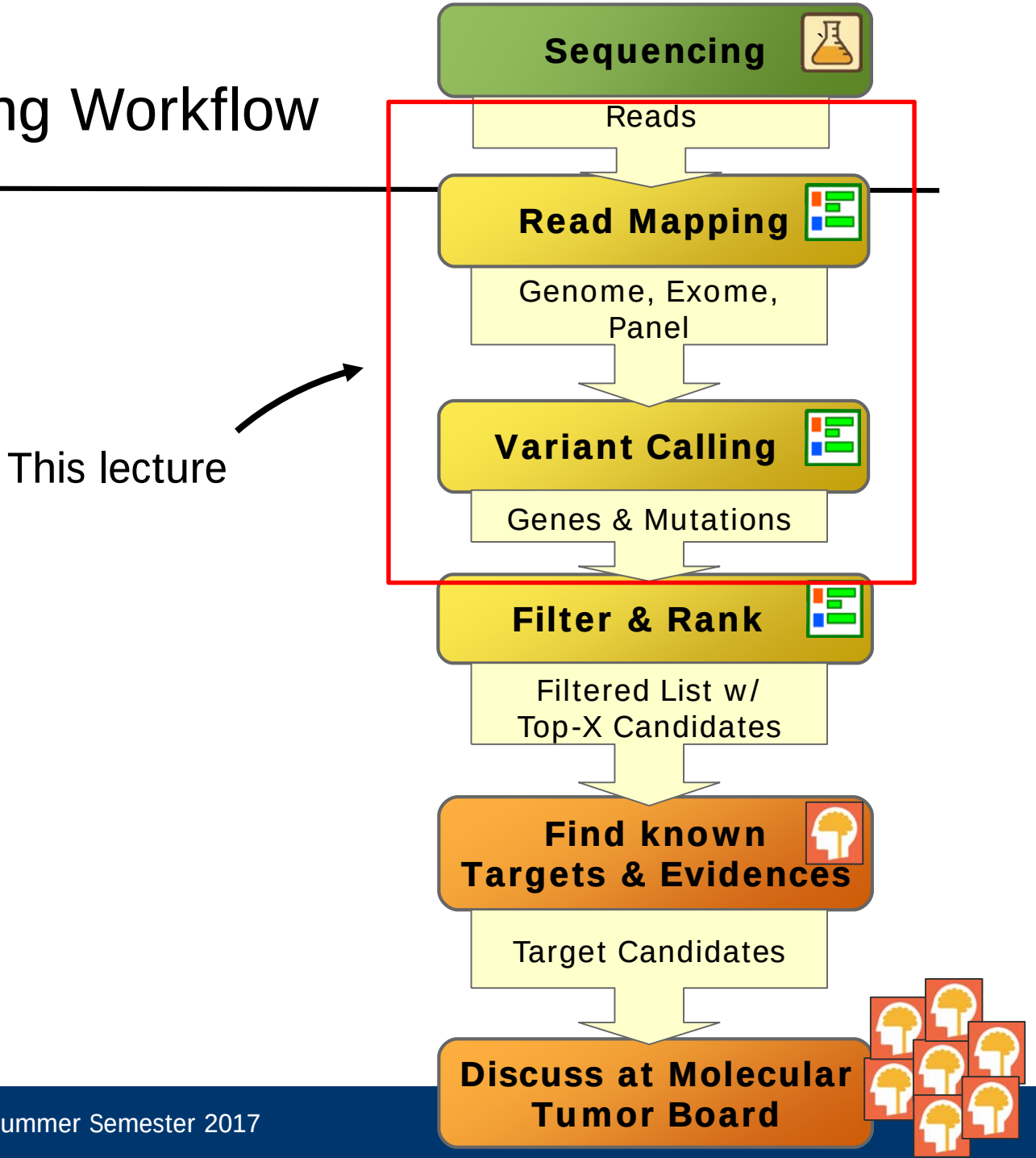
- Target candidates:
  - Mutated genes, for which a specific drug exists*



* Approved, in clinical/preclinical evaluation, known to work in other cancer type, ...



Molecular Profiling Workflow



This Lecture

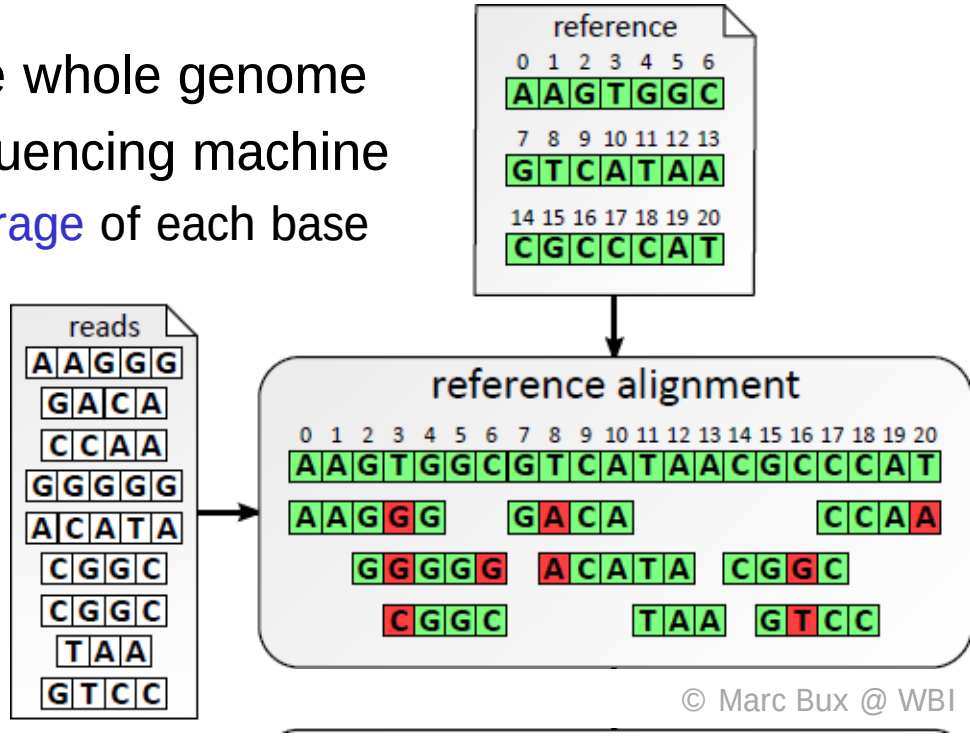
- Read Mapping
- Variant Calling

Read Mapping

- A sequencing machine outputs short sequence reads
  - Not whole genome or chromosome as one long sequence
- Need to reconstruct to whole sequence from the reads

Read Mapping

- A sequencing machine outputs short sequence reads
  - Not whole genome or chromosome as one long sequence
- Need to reconstruct to whole sequence from the reads
- General Approach:
  - Given a reference build of the whole genome
  - Given the reads from the sequencing machine
    - With a certain depth of coverage of each base
  - Find the best alignment for each read within the reference sequence
  - Together with information about matches, mismatches indels
    - Similar to an edit script



© Marc Bux @ WBI

Read Mapping: the Problem

- Wanted: fast and accurate method of mapping all reads to the reference
  - Performing an approximate matching for up to 3 billion reads
  - BLAST is fast but still too slow here
- Need to account for different sources of errors
  - Sequencing errors in the reads
  - Errors introduced by the cloning process (PCR)
  - Errors in the reference assembly
- Need strategy to handle multiple best mappings for a read
  - Generating a quality score for each read including information other than just the alignment score

Many Read Mappers available

- Key element: indexing of reference
  - Similar to BLAST q-gram index
- > 100 tools for read mapping available (cmp SEQwiki)
  - e.g., Blat, BWA, Bowtie, GSNAP, Maq, RMAP, Stampy, SHRiMP
  - Some faster, some more accurate, some more memory efficient, some better to parallelize, some better for long/short reads etc
- BWA (Burrows-Wheeler Aligner) ist very popular
  - Uses *Burrows-Wheeler transform* for indexing and compression (also used by bzip2 – and by many other read mappers)
  - 3 algorithms:
    - BWA-backtrack – for reads $\leq$ 100 bp
    - BWA-SW – for reads $\geq$ 70 bp
    - BWA-MEM – for reads $\geq$ 70 bp, faster & more accurate than BWA-SW

Different Sequencing Techniques and Read Lengths

Quelle: Wikipedia/DNA Sequencing 5/17

Comparison of high-throughput sequencing methods[60][61]

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	10,000 bp to 15,000 bp avg (14,000 bp N50); maximum read length >40,000 bases[62][63][64]	87% single-read accuracy[65]	50,000 per SMRT cell, or 500–1000 megabases[66][67]	30 minutes to 4 hours[68]	$0.13–$0.60	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[69]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 400 bp	98%	up to 80 million	2 hours	$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length[70]	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLiD sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences.[71]
Nanopore Sequencing[72])	Dependent on library prep, not the device, so user chooses read length. (up to 500 kb reported)	~92–97% single read (up to 99.96% consensus)	dependent on read length selected by user	data streamed in real time. Choose 1 min to 48 hrs	$500–999 per Flow Cell, base cost dependent on expt	Very long reads, Portable (Palm sized)	Lower throughput than other machines, Single read accuracy in 90s.
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2400	Long individual reads. Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time consuming step of plasmid cloning or PCR.

Different Types of Genomic Sequencing and Variants

- Different types of sequencing substrates:
  - Whole genome
  - Whole exome: all protein coding DNA regions
  - Panel: Sequence only a defined list of genomic sites / genes
    - Various different panels exist, e.g., for different tumor types
  - RNASeq: Sequence the mRNA present in a cell at a given point in time

- Different types of genomic variantions
  - single nucleotide variants (SNV) / polymorphisms (SNP)
  - Multi nucleotide polymorphisms (MNP)
  - Fusion, deletion, copy number variation, translocation, inversion
  - ...

This Lecture

- Read Mapping
- Variant Calling

Variant Calling: Problem definition

- Input for each position
  - A column of bases (cmp *coverage*)
  - Mapping quality score for each read
  - Base call quality for each position in the read (from sequencing)

- Output for each position:
  - Whether the genomic position is
    - Homozygous wildtype (as per reference)
    - Heterozygous
    - Homozygous variant

Quelle: Wikipedia

Sources of Mismatches

- True variants

- Errors introduced by the cloning process (PCR)
- Errors from sequencing in the reads
- Errors in the reference assembly
- Errors in the read mapping

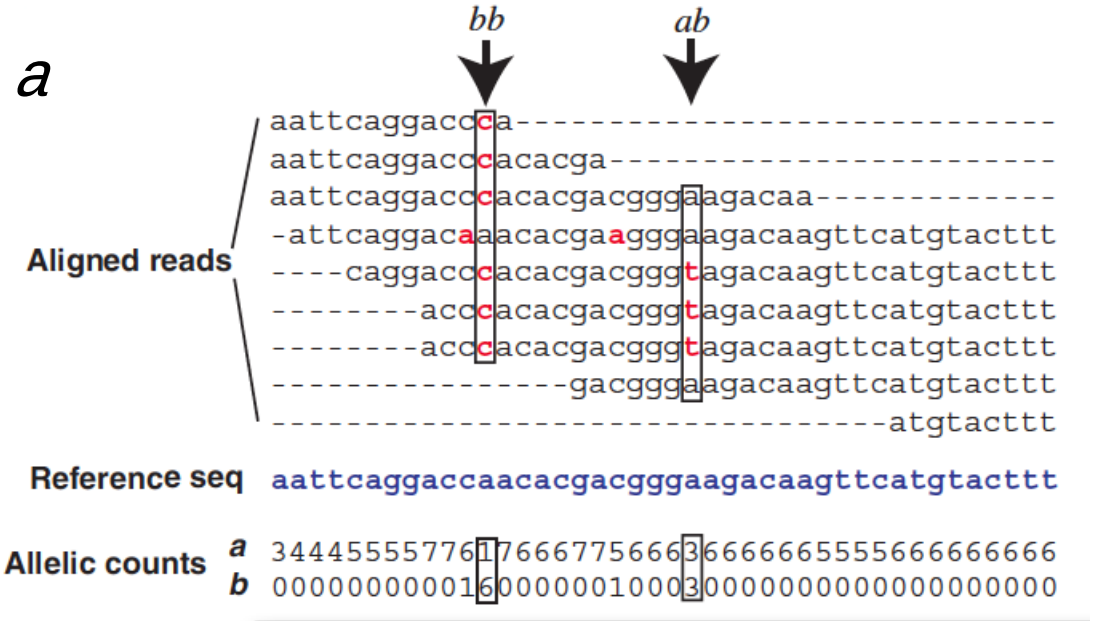
- Higher depth of coverage helps deal with the errors in the variant calling process

PHRED Quality Score Q

- Originally used by the PHRED variant calling program
  - Now widely used for characterizing error probability for bases
- PHRED score Q is logarithmically related to error probability P at which a base is wrong
  - Error rate P = 10 – (PHRED score Q / 10)
  - $Q = -10 \log_{10} P$
- Example: PHRED score Q30
  means error rate $P = 10^{-3}$
  means 1 in 1000 bases will be wrong

Error Estimation

- Given k wildtype bases *a*
- n reads
- n – k bases *b*
- with *a, b* ∈ {A,C,G,T}
- *a* ≠ *b*



- **True genotype** | **Number of Errors**

Homozygous *a,a*	n – k
Homozygous *b,b*	k
Heterozygous *a,b*	$\approx \quad \mathrm{dbinom}(n,k,p=0.5) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \frac{1}{2^n}$

From: http://www.mi.fu-berlin.de/wiki/pub/ABI/Genomics12/varcall.pdf

Naïve Approach

- Filter base calls by quality scores
- Apply frequency filter by column
- Use frequency thresholds *f(b)* for variant base *b* (PHRED Q20):

$f(b)$	genotype call
$[0, 0.2)$	homozygous reference
$[0.2, 0.8]$	heterozygous
$(0.8, 1]$	homozygous variant

- The heterozygous region is the tricky one to characterize
- Works well with high coverage

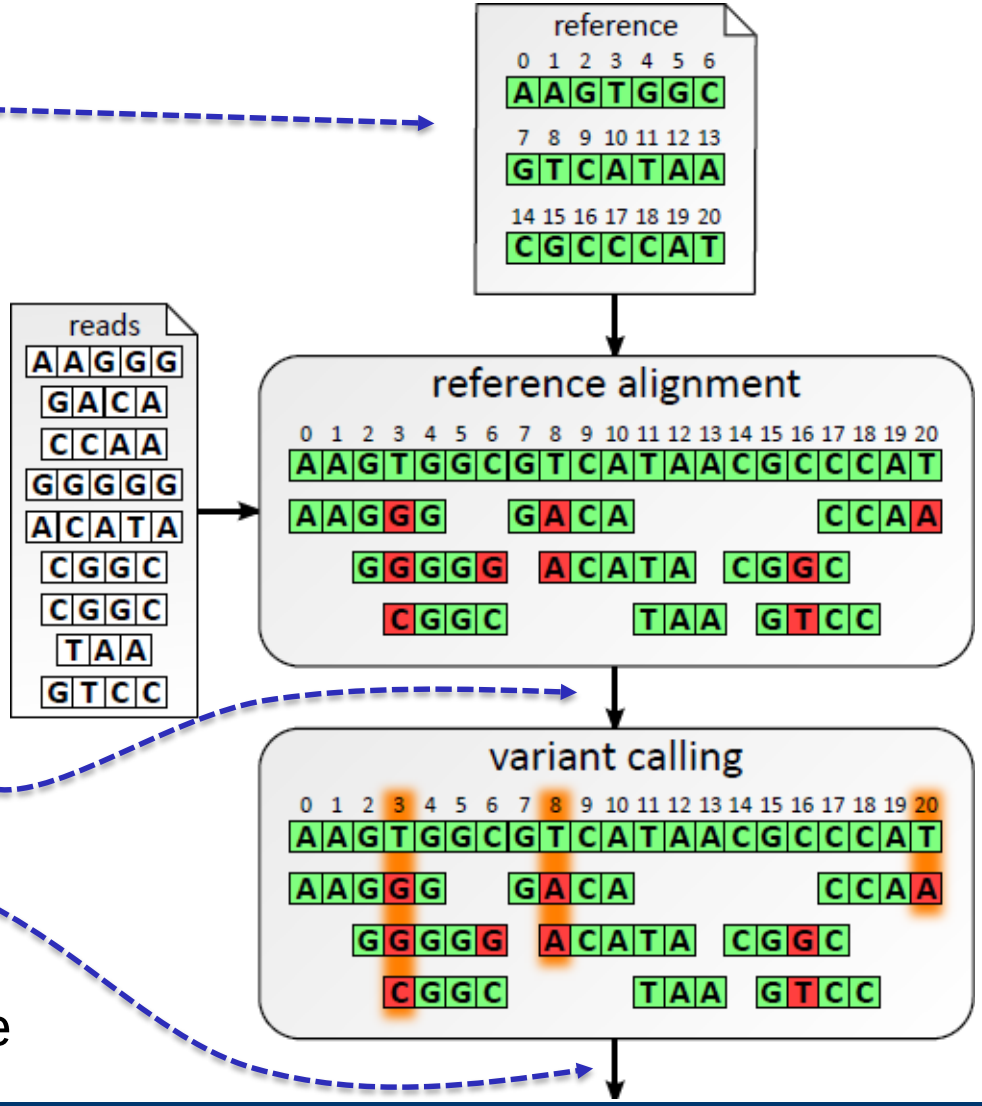
Naïve Approach Problems

- Quality threshold leads to loss of information on individual read/base qualities
- Low sequencing depth (low coverage) leads to undercalling of heterozygous genotypes
- Does not provide a measure of confidence in the call
  - No way to jugde the quality of a specific call

- Has been replaced by probabilistic methods
  - MAQ (integrates base and mapping quality scores, provides measure of reliability of the call, uses a Bayesian model)
  - SNVmix, SAMtools, GATK, VarScan2, Freebayes
  - others

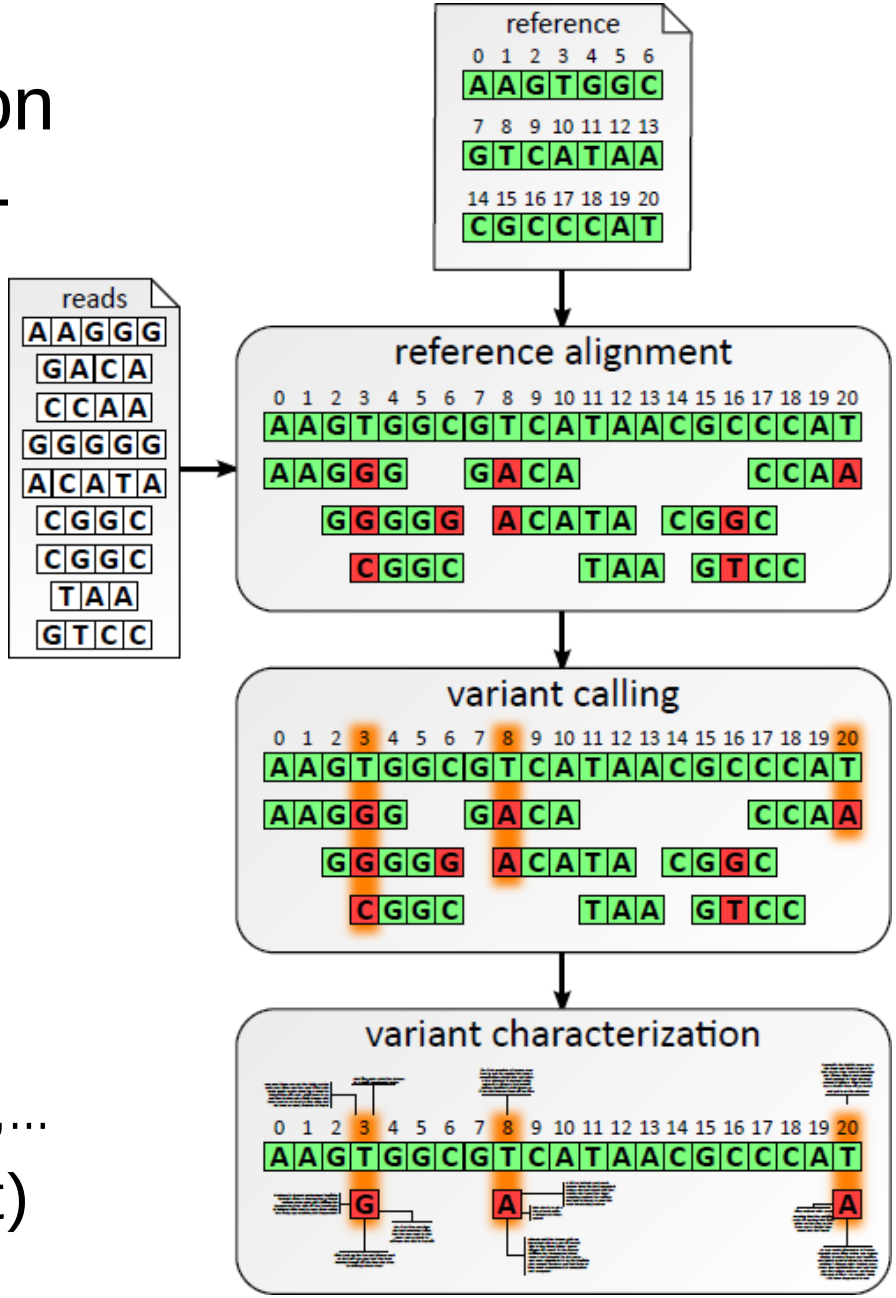
Data Formats

- FASTA
  - For sequence data
- FASTQ
  - For sequence data with quality information
- SAM
  - Sequence Alignment Map
  - Alignment details for every read
- VCF
  - Variant Call Format
  - Variants found wrt reference



Last Step: Variant Annotation

- For each identified variant from the VCF file, find
  - Coding region? Which gene?
  - Protein domain affected
  - Pathway affected
  - Population statistics (1000 genomes, TCGA, etc)
  - Predicted effect (pathogenic?)
  - Druggable? Evidence Level?
- Using numerous public databases
  - dbSNP, CIViC, ClinVar, COSMIC,...
- Using specialized tools (effect)
  - SIFT, PolyPhen, ClinGen,...



Variant Annotation

- Tools exist for subsets of information
  - ANNOVAR is very popular
- Typical data integration problem
  - Get data from many different public databases
  - Map data elements onto each other
    - Different schemata, different names for elements
  - Map identification schemes onto each other
    - Genes can have IDs from RefSeq, HGVS, Entres, HUGO, UniProt, Ensembl etc
  - Map between reference builds!
- Information constantly updated