

Variant Calling (using High-throughput Sequencing Data)

Autumn 2017

SOFTWARE AND DATASET – SETUP AND INTRODUCTION

Software and data installation – mini practical

- Use the terminal or a file explorer to take a look at the central repository
 - **exerDefinitions** contains all the practicals
 - **exerResults** contains pre-computed results of practicals
 - **inputData** contains all the datasets we will be working on
- Software
 - for documentation of how the tools work for a lot of commands you get a quick help by typing the following at the command line: **command** or **command -h**
 - **If this does not work, the best is to google the tool and read the online documentation.**
- Lets take a quick look at the data >> mini practical (see next slide)

Introduction of datasets in **inputData**

- reads_exomeCapt_chr5 in fastq format (**reads_agilentV1_chr5**)
 - real reads from exome capture (**real_patient**)
 - simulated: known mutations (mainly indels) and simulated reads (**simul_indels**) – same regions as real dataset
 - simulated: known mutations (mainly indels) and simulated reads (**simul_NA12878**) – same regions as real dataset
- reference data (**human_g1k_v37_chr5**)
 - **agilentV1** >> definition of capture tiles in different formats
 - **gatkBundle** >> reference data in fasta format and vcf files of known variants (dbSNP, 1000 genomes, hapmap)
- Formats >> we will return to these later

The reference genome



Or the species of your choice.

Some kind of reference is necessary for variant calling.

It is possible to use closely related species as a reference.

Naming and ordering of chromosome/contigs

	Hg18 (UCSC)	B36 (NCBI)
Contig prefix	chr	none
Mitochondrial contig	chrM	MT
Contig order	chrM, chr1, chr2,, chrX, chrY	1, 2,, X, Y, MT

- Genome references
 - Fasta file: must have .fasta extension + respect naming and order
 - Fai file (created by samtools faidx): contig, size, location, basesPerLine → for efficient random access
 - Dict file (created by Picard CreateSequenceDictionary): SAM style header describing the contents of the fasta file → for names and length of original file
- ROD (reference ordered data)
 - GATK supports several common file formats for reading ROD data: VCF, UCSC formatted dbSNP, BED
- dbSNP files
 - Must also be ROD
 - Generated by GSA from the dbSNP db using a bit of bash, awk and a perl script: sortByRef.pl. Full details: http://www.broadinstitute.org/gsa/wiki/index.php/The_DBSNP_rod
- All of the above delivered for human as part of the **GATK resource bundle**
 - Other species may also be available
 - Help on generating for another species see GATK wiki or getsatisfaction.com/gsa

The FASTQ reads: Motivating the simulation

- Why simulate in variant calling?
 - with real data we do not know the “truth”
 - in a simulation, we can define the “truth” and then “hide” it and test how well our methods can recover this “truth”
 - in a real dataset, most variants are single nucleotide polymorphisms and these are usually very easy to correctly call
- Drawbacks of simulation: usually will not accurately model all aspects of real data

The simulation

Fragment sample

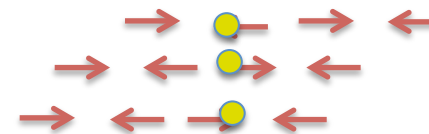
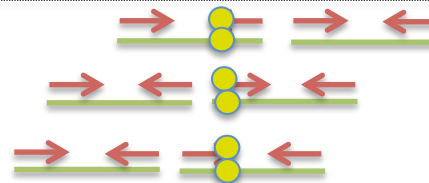
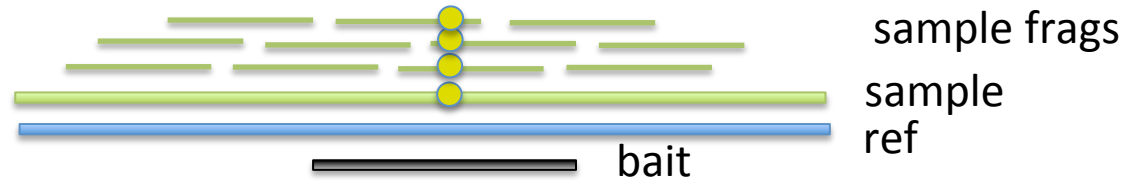
Capture

Sequence

Map

Align

Variant call



Are we able to correctly call this site as variant?

Potential challenges are:

- depth of coverage
- position in read
- size of indel

The simulated dataset

Placed in true agilent tiles on chr 5

- SNPs – 100 of each type >> 600 in total
 - 3 strands (1 hetero, 2 hetero, 3 homo)
 - 2 positions in tile

Tiles: 1-600 SNPs
Position 0-14MB)
- Deletions – 5 of each type >> 1800 in total
 - 3 strands (1 hetero, 2 hetero, 3 homo)
 - 2 positions in tile (at edge or well inside)
 - size: 1 to 60

Tiles: 601-2400
Deletions - increasing in size
in batches of 30 (14 to 72 MB)
- Insertions – 5 of each type >> 1800 in total
 - 3 strands (1 hetero, 2 hetero, 3 homo)
 - 2 positions in tile (at edge or well inside)
 - size: 1 to 60

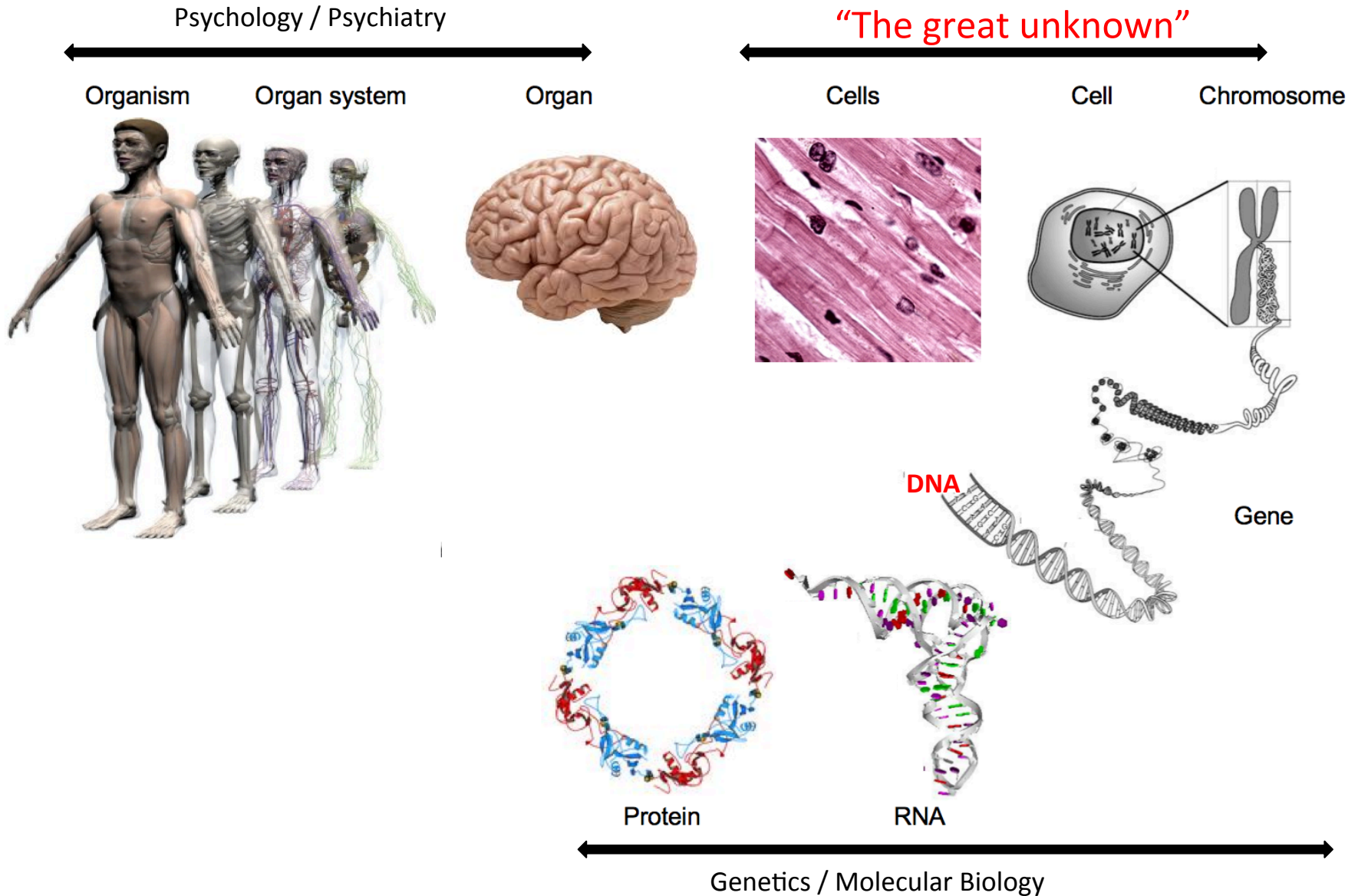
Tiles: 2401-4200
Insertions - increasing in size in
batches of 30 (72 to 131 MB)
- **A highly unrealistic set of variation but useful for studying the intricacies of variant calling**

Practical 01 – setting up your home directory

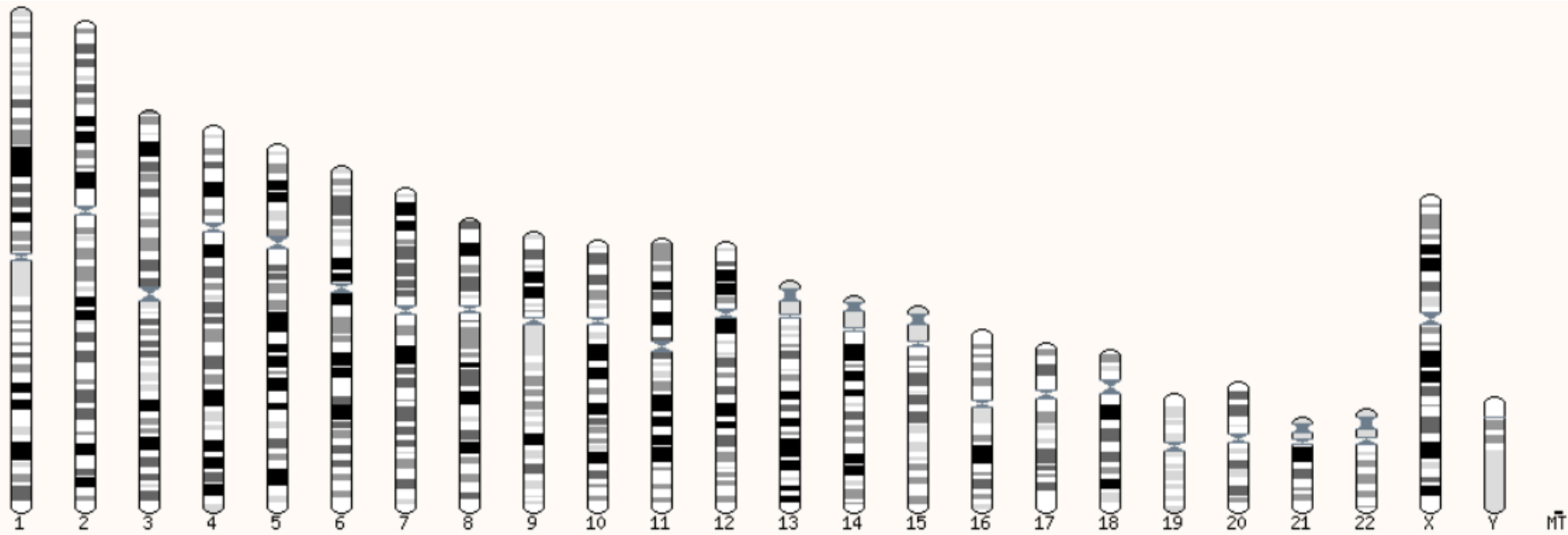
- We do this real SLOOOOOOOOW.....
- We need to create a place where you can do the exercises in your home directory
- Start two terminal windows
- First setup the environment, you need:
 - in a terminal: **source /data/vc/exerDefinitions/setupEnv.bash**
 - **Nota Bene: you should always run this when you start a new terminal**
 - test it: **echo \$localVcDir**
 - you should output something like **/share/inf-biox121/home/timothyh/vc**
- Get your copy of the files you need from the central location:
 - in a terminal: **/data/vc/exerDefinitions/copyFiles.bash**
 - The files were copied to this location: **echo \$localVcDir**
 - Check out the contents of this directory: **ls \$localVcDir**
 - You should see a list of files which are your copies of the inputData, slides and exercises (exerDefinitions) and a location where you should do all exercises (exerSandbox)

GENETICS 101

The biological “stack”



The human karyotype



Different types of mutation

- Single nucleotide polymorphism
 - A single base that is changed eg A becomes G
- Indels (involving only less than 100 bp)
 - insertion: additional bases added to the sequence
 - deletion: bases deleted from the sequence
- Large indels are referred to copy number variations
- Other types of variation include:
 - inversion
 - translocation
- Whole chromosomes can also be gained or lost (aneuploidy)
 - Most are lethal
 - Trisomy 21
 - Sex chromosomes: XXX, X0, XXY, XYY

Variation in the human genome

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

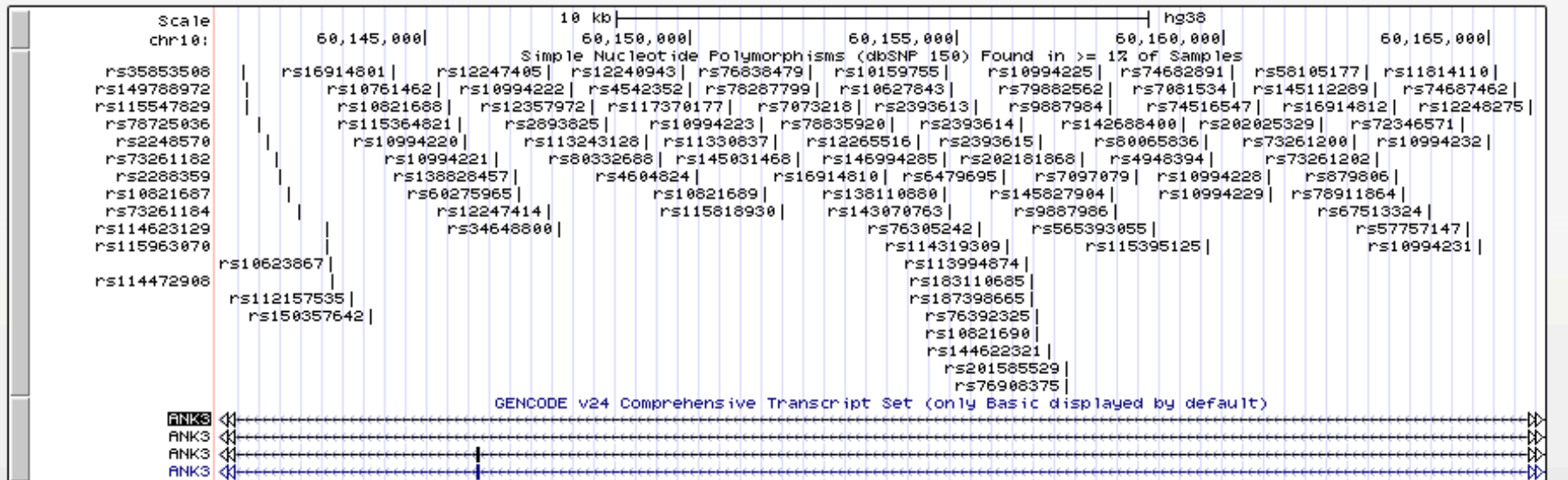
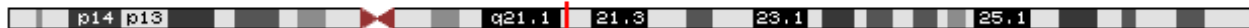
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr10:60,141,120-60,166,064 24,945 bp.

enter position, gene symbol, HGVS or search terms

go

chr10 (q21.2)



Basic concepts – Take notes

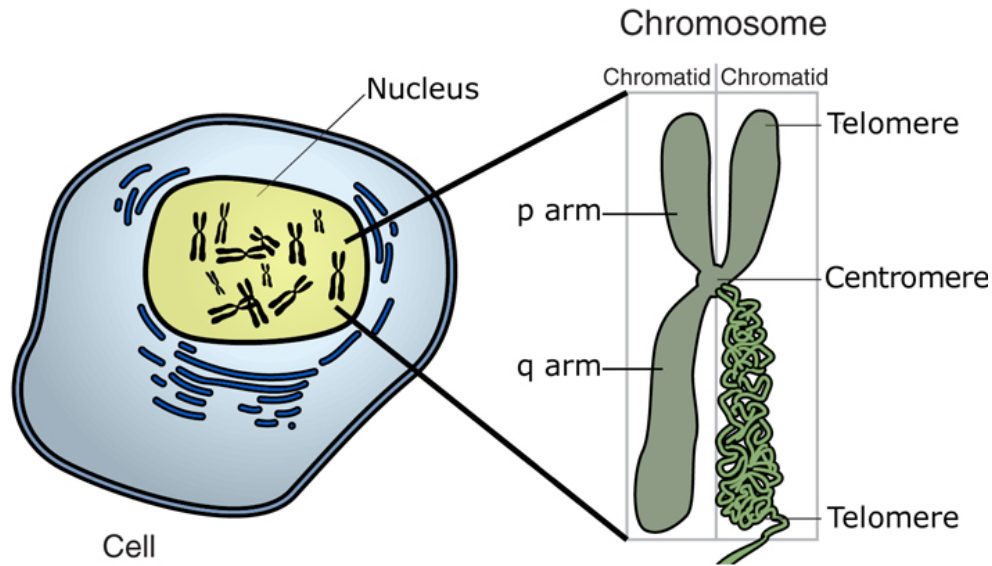
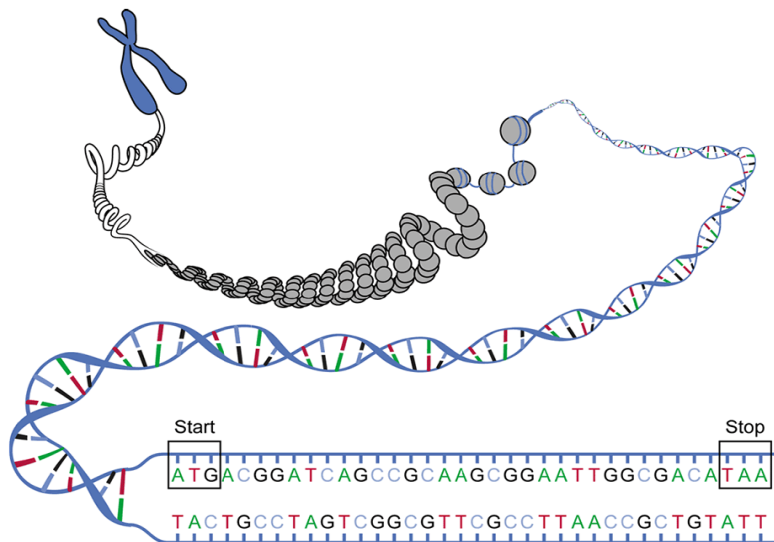
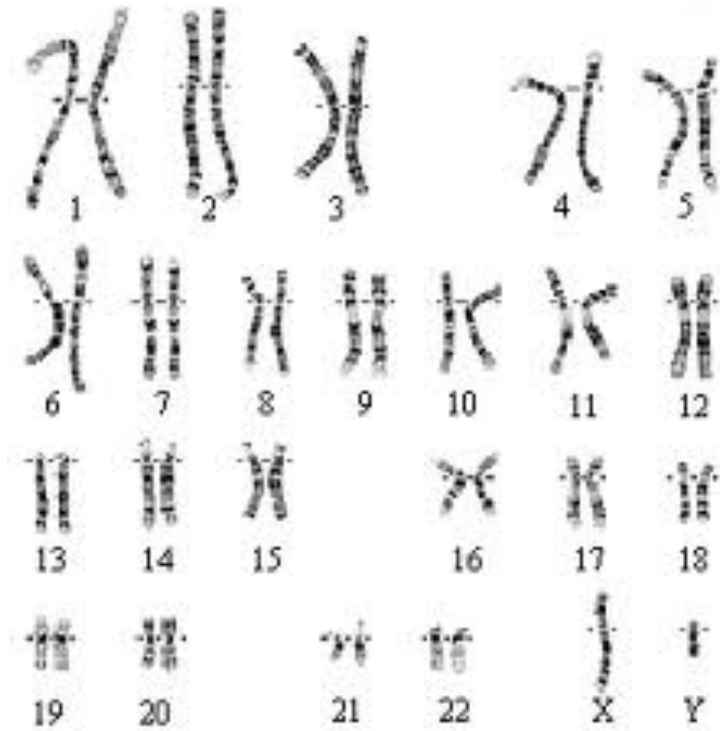


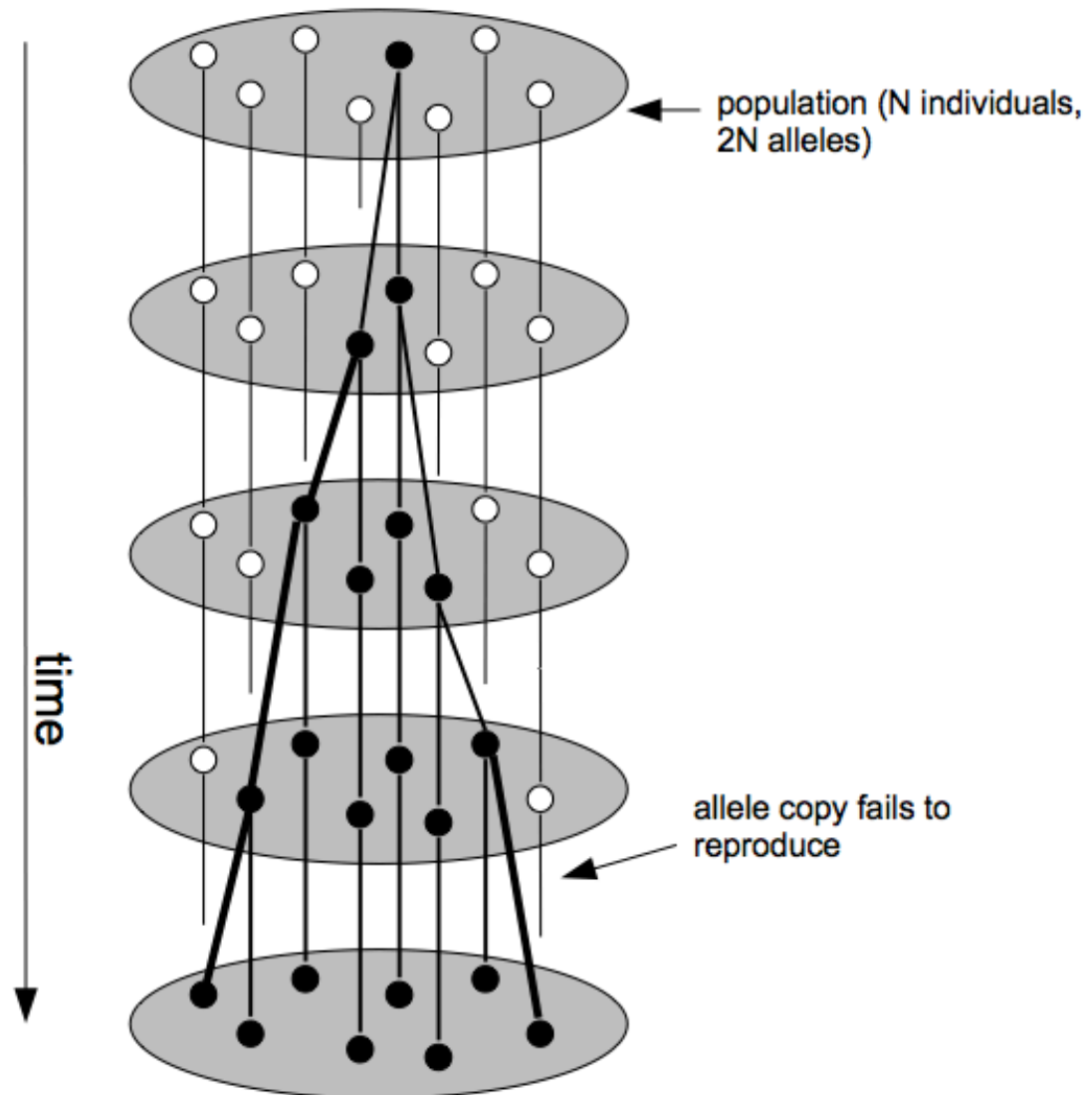
Image adapted from: National Human Genome Research Institute.



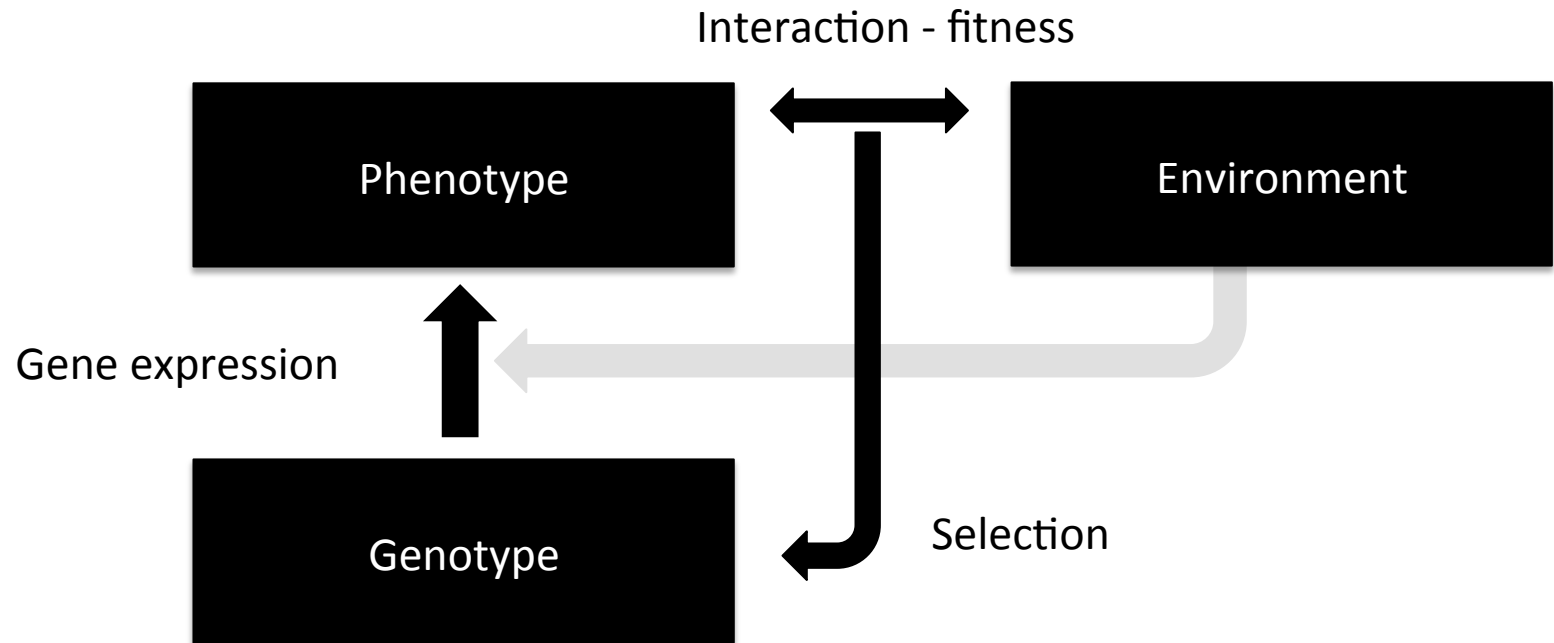
- Cells
- Chromosomes (ploidy)
- DNA complementarity and strands
- Variant site (SNPs are the most simple)
- Genotype: homozygote / heterozygote
- Haplotype

Neutral drift and selection of genomic events

New black allele arises through mutation of a white allele

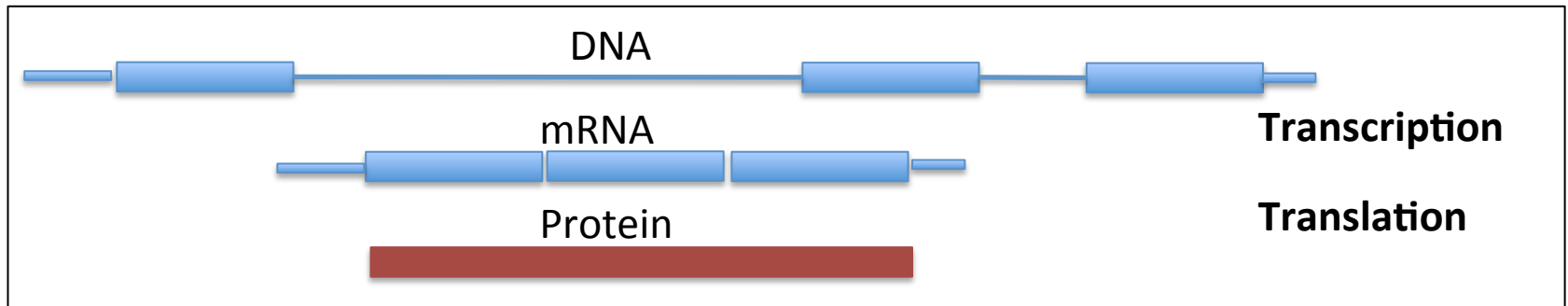


An abstract view of forces shaping evolution

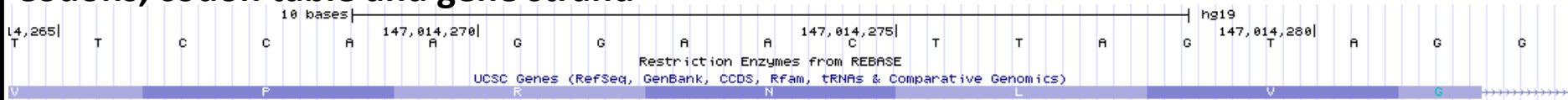


What is a gene?

Central Dogma



Codons, codon table and gene strand



- There are large numbers of genes in multicellular organisms
- Not all organism have multiple exons per gene and thus do not need splicing
- Outside genes are regulatory elements that control when and where genes are expressed

Kahoot!

On your phone or in a browser go to the webpage: kahoot.it

And punch in the code for the Genetics and genomics quiz

This time I will NOT explain as we go along, instead you will get answers in a later lecture

MAPPING AND ALIGNMENT – AN IMPORTANT DISTINCTION


Mapping

- Read mapping is the process of locating where in the reference genome a read originates from
- We typically have millions of 100-150 bp sequence reads which we need to map to a reference genome (often billions of bases)
- In addition, there will often not be an exact match between the read and the reference due to:
 - base call errors in individual reads
 - variation in the sample relative to the reference.
- One way of locating the origin of a read is to scan the whole genome looking for a match but this method is obviously very time consuming
 - later in the course we will study an algorithm which speeds this up

After mapping - Alignment

Ref	A	A	A	C	A	A	T	T	A	A	G	T				
Sample				AAAT												
Sample	A	A	A	C	A	A	A	T	A	A	T	T	A	A	G	T

Ref	A	A	A	C	-	-	-	-	A	A	T	T	A	A	G	T
Sample	A	A	A	C	A	A	A	T	A	A	T	T	A	A	G	T



Correct alignment

Sample read A A A C A A A T A A T T

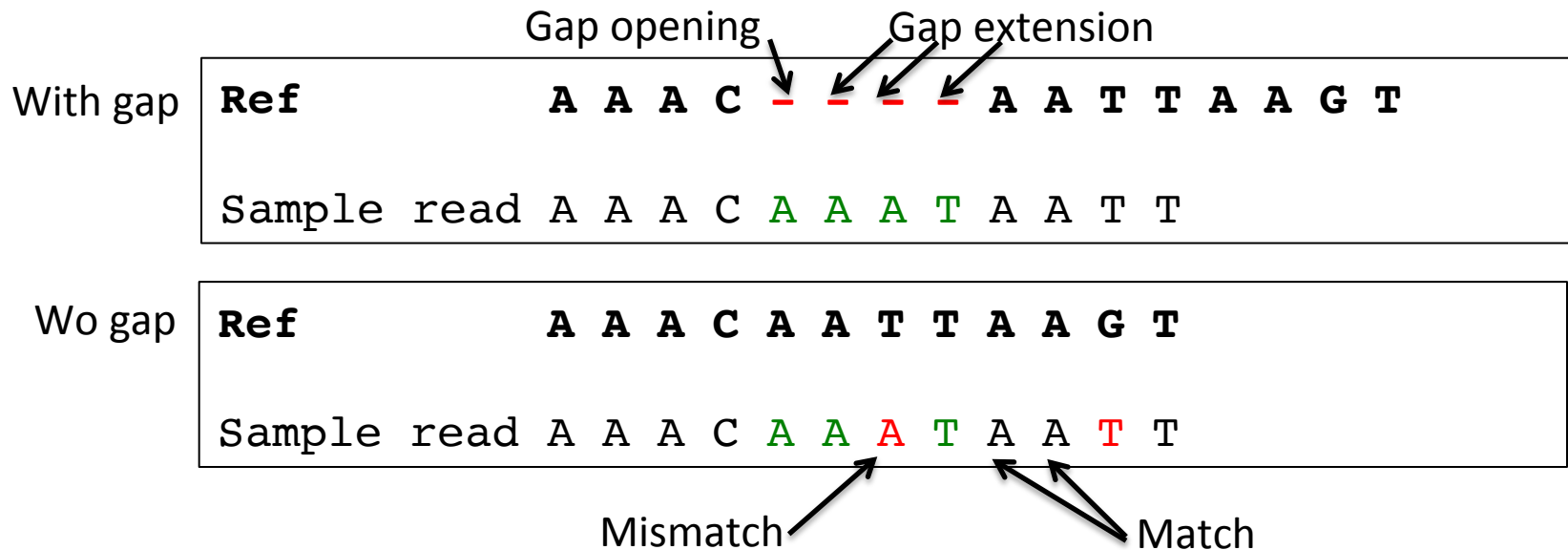
Correct alignment

Ref	A	A	A	C	-	-	-	-	A	A	T	T	A	A	G	T
Sample read	A	A	A	C	A	A	A	T	A	A	T	T				

Possible alignment

Ref	A	A	A	C	A	A	T	T	A	A	G	T
Sample read	A	A	A	C	A	A	A	T	A	A	T	T

Alignment



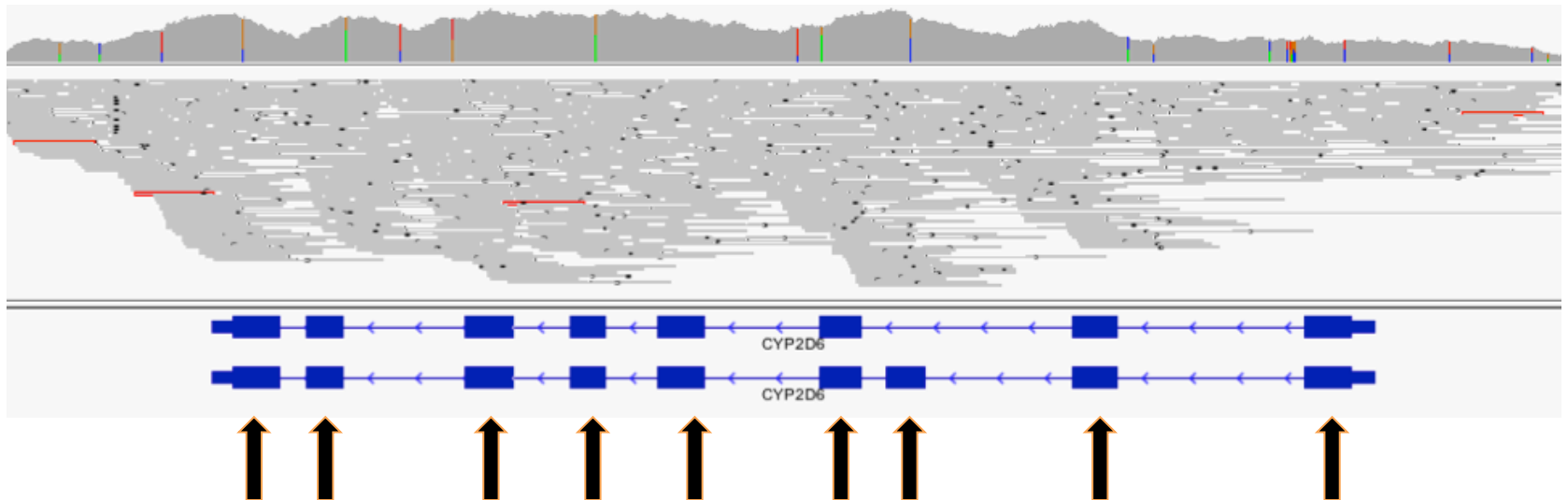
- Key component of alignment algorithm is the scoring
 - negative contribution to score
 - opening a gap
 - extending a gap
 - mismatches
 - positive contribution to score
 - matches
- The exact score contributions determine which alignment is chosen
- **Smith-Waterman** is an algorithm for finding optimal alignment given a scoring scheme without exhaustively enumerating and scoring all possible alignments



EXOME CAPTURE – ESSENTIALS

Sequencing a whole genome

- We could sequence a whole genome BUT only a small fraction is actually genes
 - in human less than 2% of the genome codes for proteins
- It saves a lot of sequencing to “capture” the regions we are interested in and sequence them



Whole genome vs. Whole exome



Option 1: Whole genome

>> sequence every fragment either single read or paired-end

BUT: Maybe we are not interested in all regions of the genome

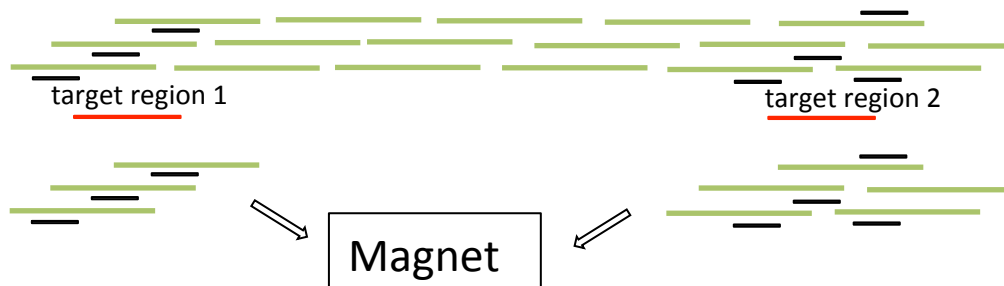
>> WG is a waste of sequencing

Option 2: Whole Exome

>> Extract the fragments in region we are interested in using hybridising oligonucleotides

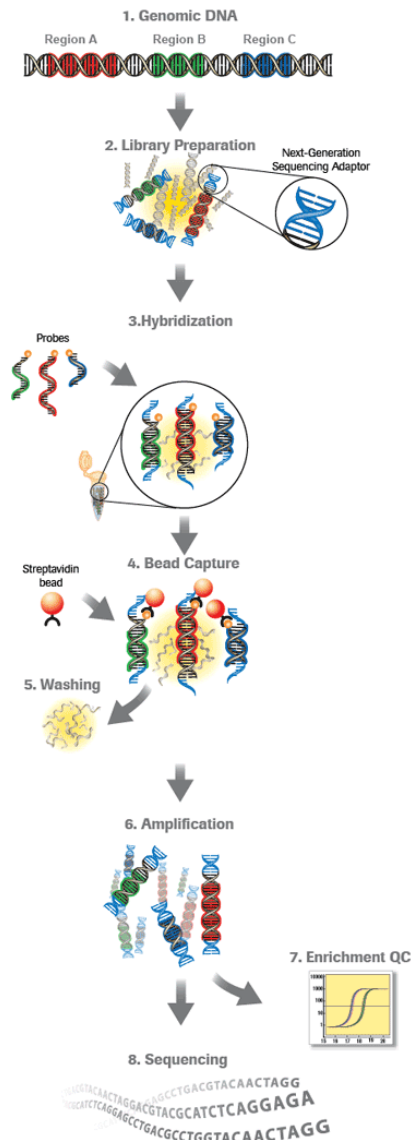


- Oligonucleotides are purchased from a manufacturer
- Several will be used to cover a target region.
- I call a group of oligos covering a target region a “tile”
- A tile will usually be a bit bigger than the target
- All the oligos together are sometimes called the “bait”
- Oligonucleotides have a biotin molecule attached to the



- The oligonucleotides bind to the DNA fragments.
- The oligonucleotides (and the hybridised DNA) are “fished” out with magnetic streptavidin beads that bind to the biotin

An overview of exome capture



Sonication

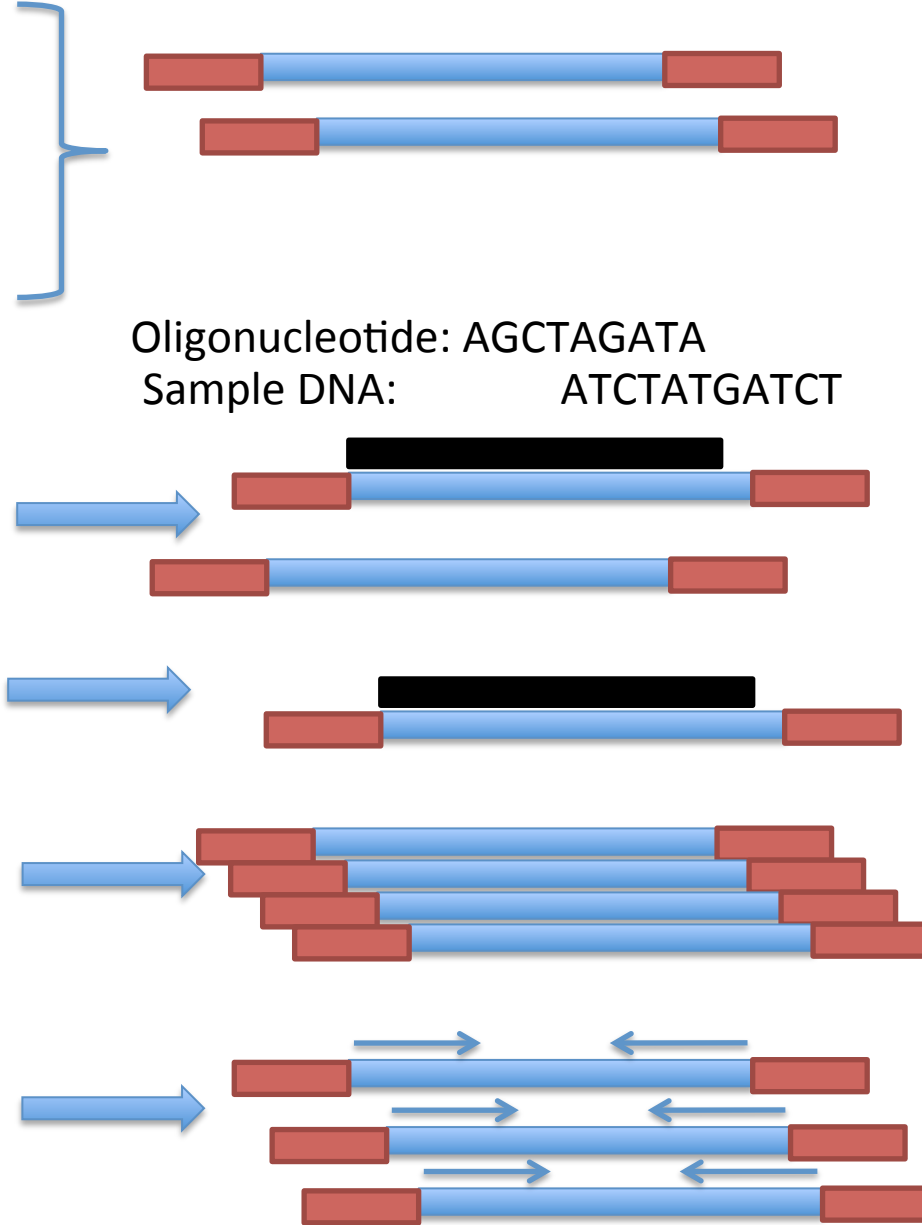
Library prep
(sequencing
adaptors on)

Hybridisation
to probes

Bead capture

Amplification

Sequencing



Practical 02 – calling variants in the blind

- We look at the read datasets in local copy
 - **Done together with instructor**
- Open script 020_basicPipeline.bash (this should be your own copy under `${HOME}/vc/exerDefinitions`)
- Explanation and execution of commands together with instructor

IGV mini practical

- Open a new terminal and start IGV
- Let us load up some data and explore some of the features of IGV
- We will load two datasets:
 - The dataset that you generated in the previous exercise
 - `${HOME}/vc/exerSandbox/02_basicPipeline`
 - Another more realistic simulated dataset
 - `${HOME}/vc/inputData/reads_agilentV1_chr5/simul_NA12878`
- Some the things we will do
 - loading files (locally or from server)
 - navigating (zoom in and out, back and forward)
 - getting access to details (on hover or click)
 - changing the way you view things (squish/expand, colours, pairs)
 - moving tracks and panels around and resizing
 - sessions
 - marking regions of interest
 - changing preferences like soft-clipped bases

Practical 02 – Variations to a vc pipeline

- `${HOME}/vc/exerDefinitions/020_variations.bash`
- Open script `020_variations.bash` (this should be your own copy under `${HOME}/vc/exerDefinitions`)