Variant Calling
(using High-throughput Sequencing Data)

Autumn 2018

ABOUT THE COURSE

Common concepts and themes across modules

- Technical / bioinformatic
  - Study design (RNA-seq, Statistical genomics)
  - Mapping principles (Genome assembly, Gene expression, Variant calling)
  - Alignment principles (Genome assembly and particularly variant calling)
  - Alternative splicing (Gene expression, variant calling)
  - Model system vs non-model system organisms
- The practice of bioinformatics / computational biology (!!! relevant for all modules !!!)
  - Reproducibility
  - Best practices
  - Testing (in the sense of trying code on a toy dataset where output is known)
  - Documenting
  - Statistics and hypothesis testing
  - Summary statistics and visualisation
  - Sanity checking / validation of results

INF-BIOX121

- The tough challenges:
  - High Throughput Sequencing
  - Biological Applications
  - Informatics
- What is the secret for 18 hours with me
  - to be fun for you?
  - to be fun for me?
  - you to learn as much as possible?
- An attitude that works well in all walks of life
  - take initiative for problem solving
  - then ask for help

Struggling with the UNIX environment?

- It’s normal!
- Do you need a buddy?
- Anyone want to be a buddy?

Our goal for the participants

- Finish the course with an understanding of all major concepts
- Know how to run a variant calling pipeline
- Ready to make informed choice about whether you may need a simpler or more complex pipeline
- Focus is on Variant Calling and functional annotation
  - Previous courses have expressed need for more downstream analysis
  - So we have included more of this
- You might experience some overlap with other modules of the course, but repetition can be beneficial

Biologist advice to a biologist

- 1st advice: DON'T PANIC
  2nd advice: DON'T PANIC
  3rd advice: DON'T

- Do the unix part of tutorial of Unix and perl primer for biologists: http://korflab.ucdavis.edu/Unix_and_Perl/ , alternatively:
  - http://linuxcommand.org/learning_the_shell.php#contents
  - http://www.tuxfiles.org/linuxhelp/cli.html

Yyou are not the first to have problems
  - search for it on the net
  - or, if you can't find the answer, ask in a forum

- Learn to write a very simple script

- Make a list of your favourite commands as you learn them.

Often when something is difficult to understand it is often either because:

- It is not clear what is being done
- Or because it is not clear why it should be done

So stop me and ask if the WHAT or the WHY is unclear!!!

Some thoughts about scientific problems vs. techniques

- Always keep in mind what the basic goal is and what the basic issues are with attaining that goal.
- Learn to switch from the broad picture to the details and back again.
- In my opinion it is mistake to focus too much on specific techniques. Become good at different techniques, but learn them because they help you solve problems.
- Don't go looking for things you can apply a technique to or at least be careful to not do only this.
    - Antony van Leeuwenhoek is a good counter-example

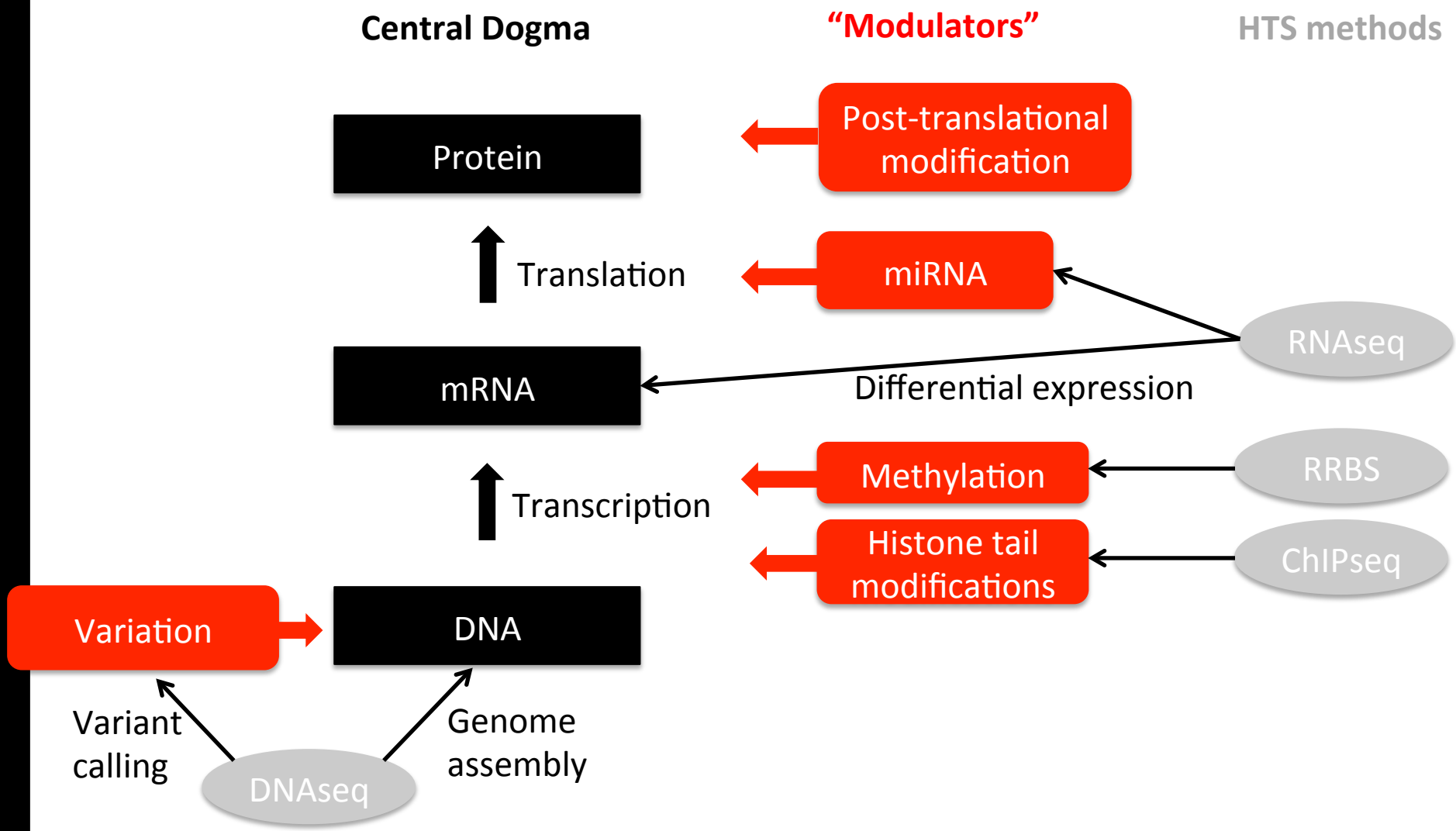
An exception to not focussing on techniques
Antony van Leeuwenhoek (1632-1723)

WHAT IS VARIANT CALLING?

What is variant calling?

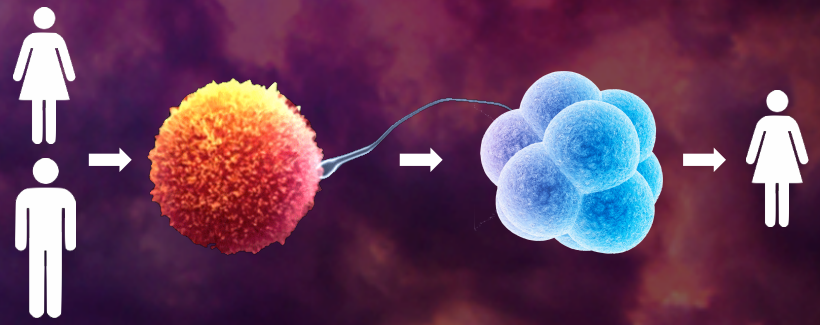
- Recall the jigsaw exercise from the mapping module
- Variant calling is determining how a DNA sample differs from the reference genome
- This gives a good approximation to the genome in the DNA sample without having to perform a genome assembly which would requires considerably more sequence data and effort.

Central dogma and HTS



WHY STUDY VARIATION?
3 GOOD REASONS

Humans and other multicellular organisms



A reproducing system which is:

* Self-assembling
* Self-repairing
* Self-operating
* Self-upgrading
* Self-aware
* Social and spiritual

REASON1: The full information underlying this system is stored in the DNA of EVERY cell

Even more amazing that what we are, is the **unlikely** story of our evolution?

An interesting digression on the greater scheme of things

REASON 2: The Fermi paradox

"Where are all the extra-terrestrials"

The Drake equation

$N = R \times f_p \times n_e \times f_l \times f_i \times f_c \times L$

N = The number of civilizations in our galaxy with which communication might be possible

R = The number of stars in our galaxy — **20 bn**

$\mathbf{f_p}$ = The fraction of those stars that have planets — **1**

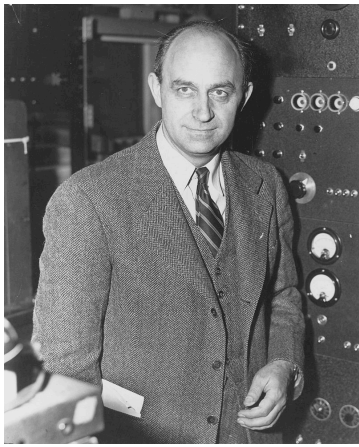
$\mathbf{n_e}$ = The average number of planets that can potentially support life per star that has planets — **0.25**

$\mathbf{f_l}$ = The fraction of planets that could support life that actually develop life at some point — **1**

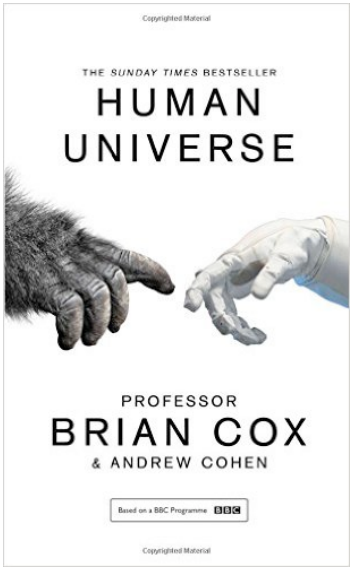
$\mathbf{f_i}$ = The fraction of planets with life that actually go on to evolve intelligent life (civilizations) — **?**

$\mathbf{f_c}$ = The fraction of civilizations that develop a technology that releases detectable signs of their existence into space — **1**

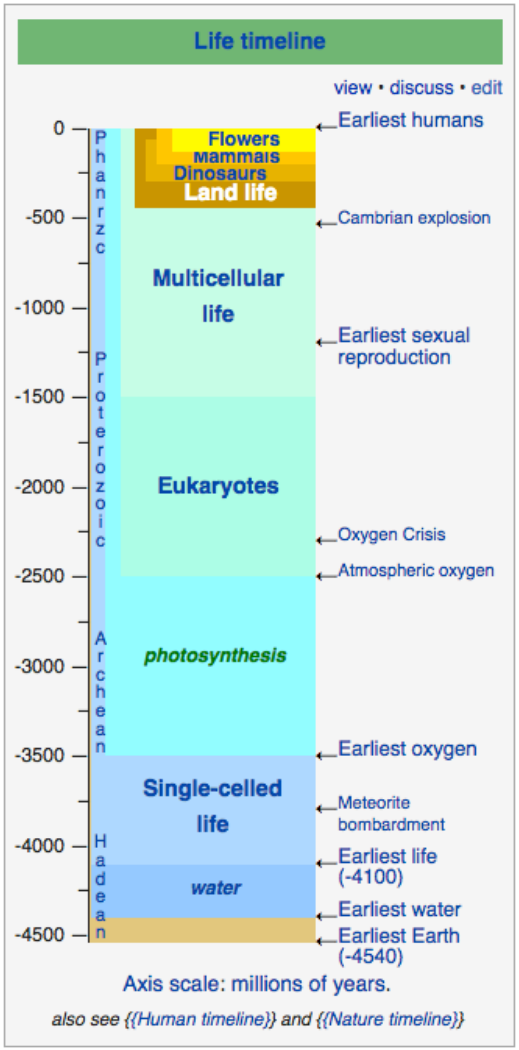
L = The length of time for which such civilisations release detectable signals into space — **?**

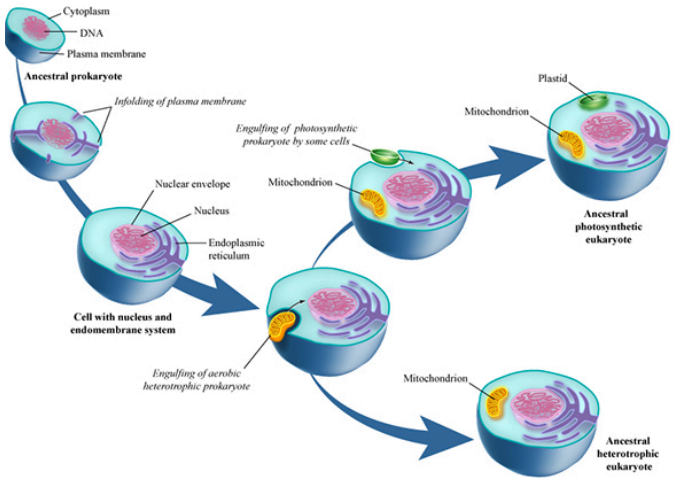


Enrico Fermi



Evolution of intelligent life: 3 kingdoms and major events





Archea

Prokaryotes

Origin of life 4.1 Bya

? ?

Endocytosis of aerobic prokaryote

Endocytosis of photosynthetic prokaryote

multi

Plants

multi

Fungi

Eukaryotes

multi

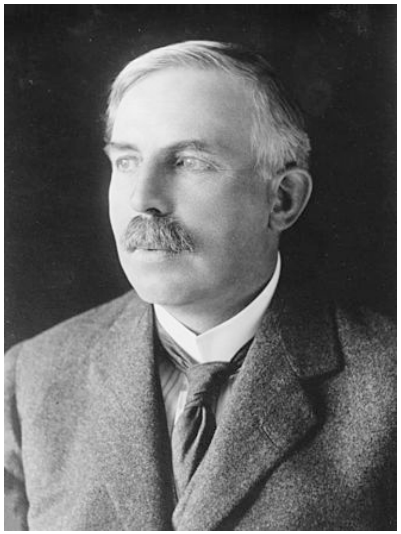
Animals

Cambrian explosion (600 Mya)

Modern humans (0.25 Mya)

Sapling Learning, 2013

Or could it be **L** that is very small?



1911 Rutherford discovers the atomic nucleus

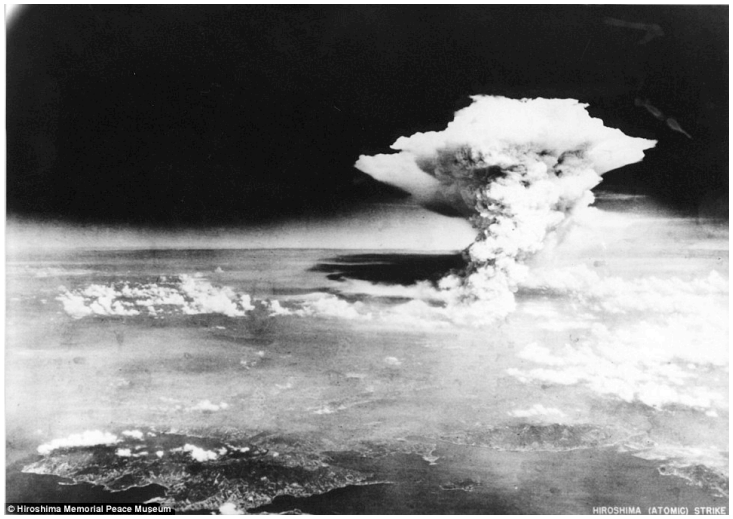
The Drake equation

$N = R \times f_p \times n_e \times f_l \times f_i \times f_c \times L$

only 34 years

Universe **12 bn** years
Life **3.5 bn** years

Active destruction



1945 Hiroshima

1903 Wright brothers complete powered flight

Or **passive** destruction: global warming

Takes only slightly longer (few hundred years)

More concretely, why study variation?

- To known the DNA sequence
- Why is the DNA sequence of interest?
- Biology
  - Understanding populations e.g. out of africa hypothesis AND population bottleneck
  - Genetic basis of behaviour
- Medicine
  - Monogenic diseases (e.g. Medical Genetics)
  - Polygenic disease (see next slides)

Natural courtship song variation caused by an intronic retroelement in an ion channel gene

Yun Ding[1], Augusto Berrocal[1]†, Tomoko Morita[1], Kit D. Longden[1] & David L. Stern[1]

Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians

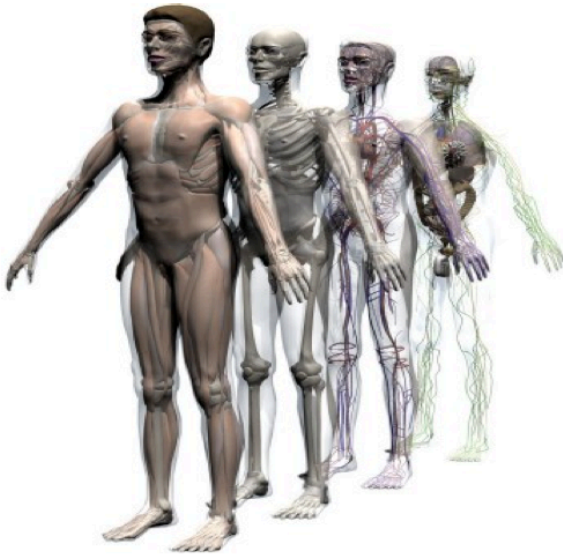
Luca Pagani, Stephan Schiffels, Deepti Gurdasani, Petr Danecek, Aylwyn Scally, Yuan Chen, Yali Xue, Marc Haber, Rosemary Ekong, Tamiru Oljira, Ephrem Mekonnen, Donata Luiselli, Neil Bradman, Endashaw Bekele, Pierre Zalloua, Richard Durbin, Toomas Kivisild, Chris Tyler-Smith

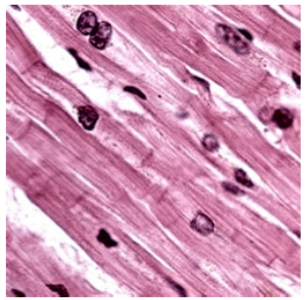
The biological “stack”

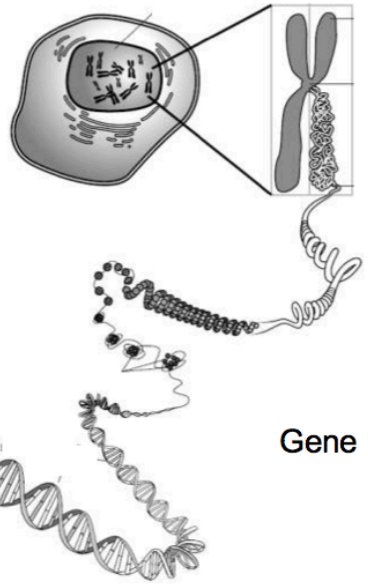
Psychology / Psychiatry

“The great unknown”

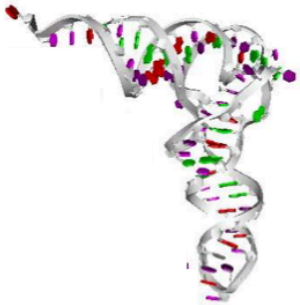
Organism | Organ system | Organ | Cells | Cell | Chromosome







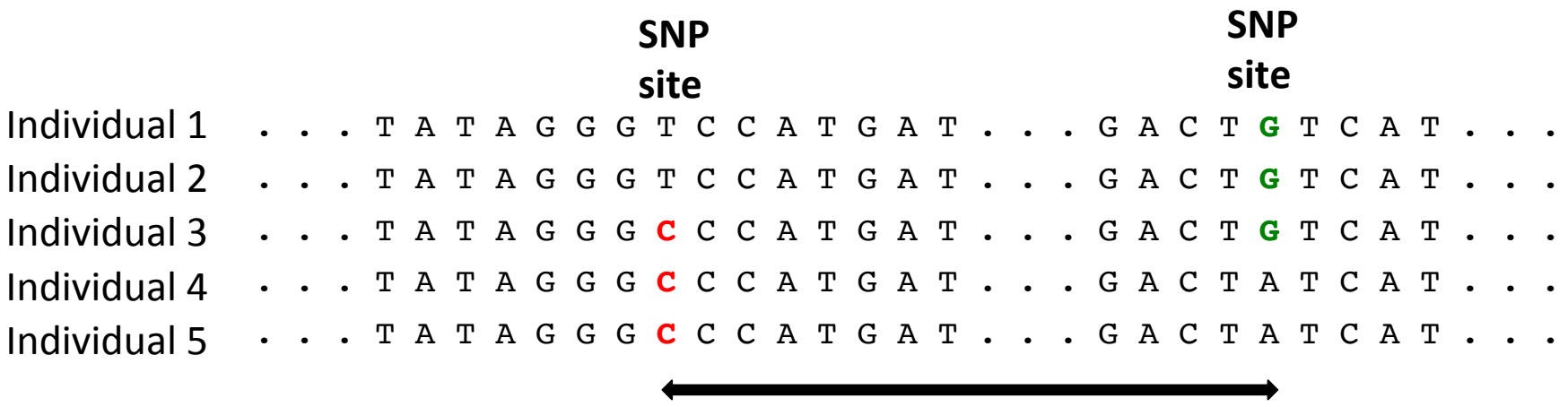
Protein



RNA

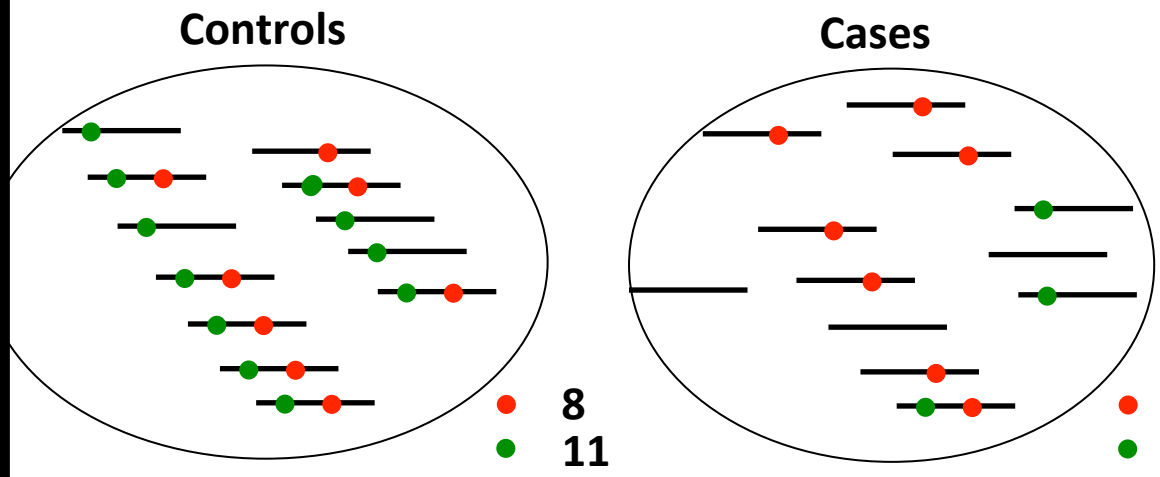
Genetics / Molecular Biology

Genome Wide Association Study



Genotyped SNPs typically 10k apart
(Genome is 3 billion base and array is 300k)

If a region of DNA has no effect on the phenotype of interest, we expect same frequency (or odds) of SNP in case and controls

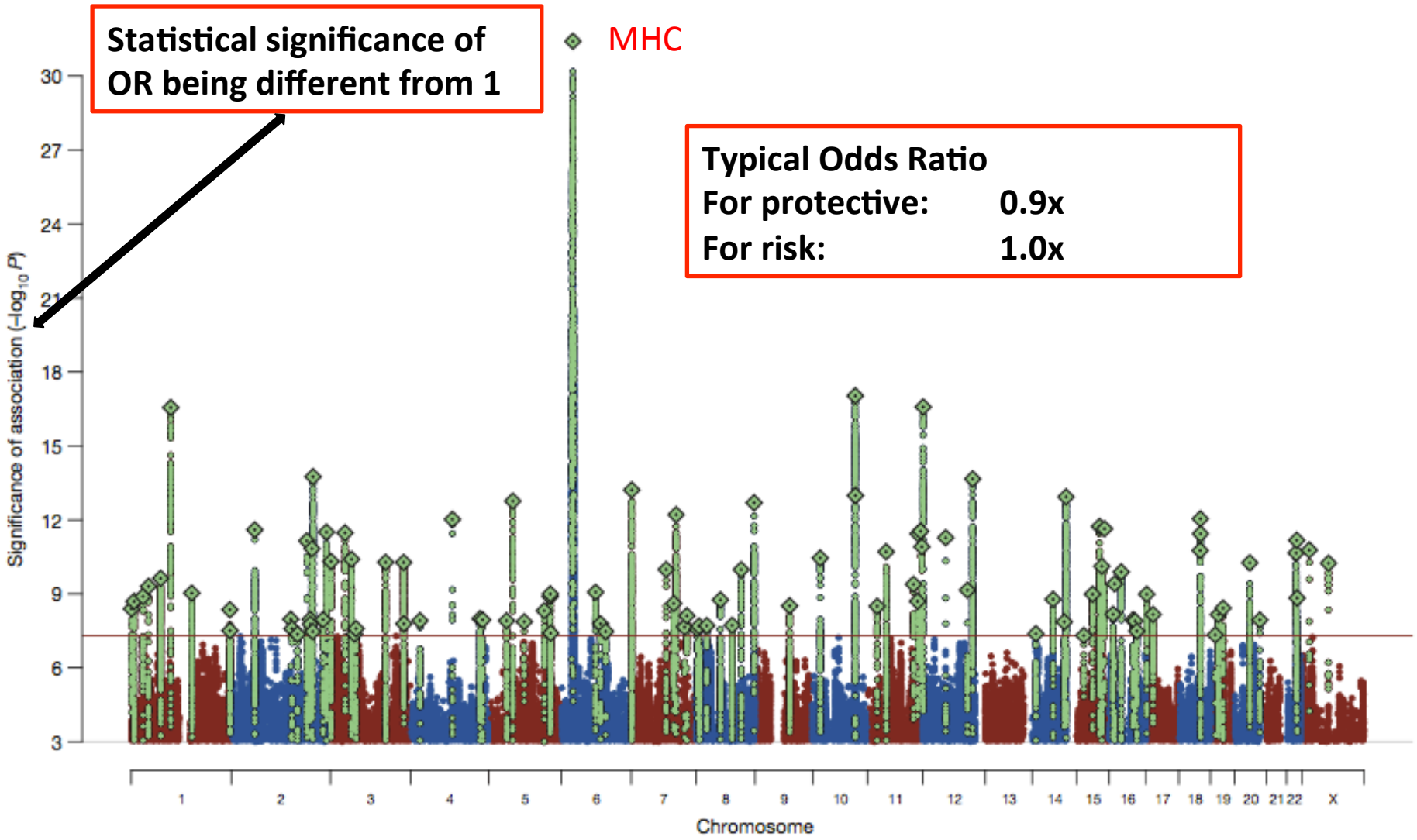


OR= Odds of the SNP in the cases / Odds of the SNP in the controls

OR > 1: risk SNP
OR < 1: protective SNP

GWAS results

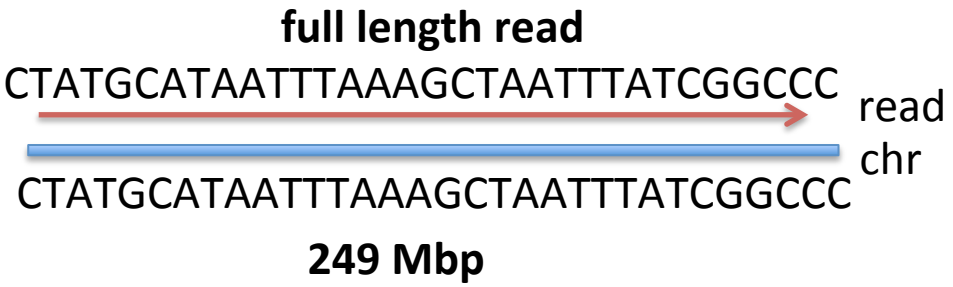
Biological insights from 108 schizophrenia-associated genetic loci **(2014)**

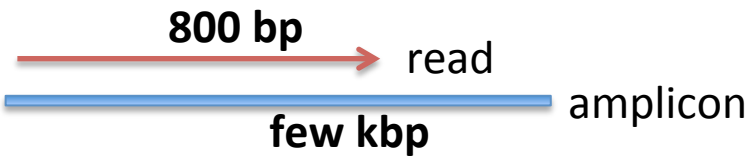


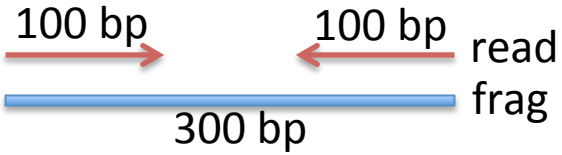
HTS: ONE GREAT STRENGTH AND ONE GREAT WEAKNESS

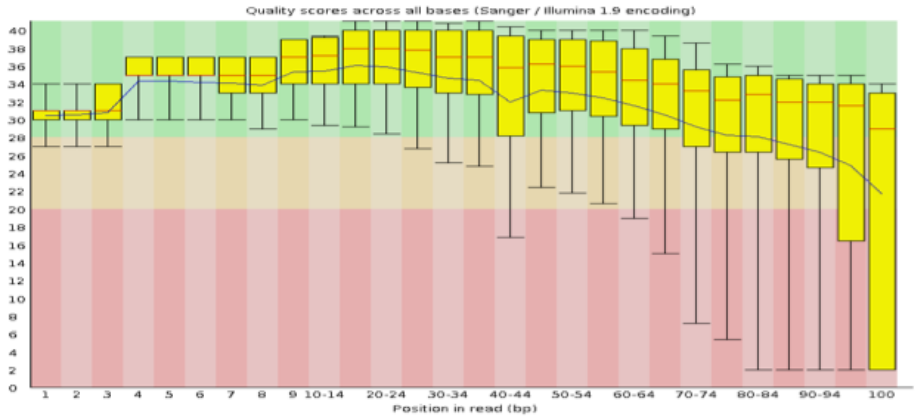
In a perfect world – Perfect sequencing

- Perfect sequencing:
  - single molecule (no PCR)
  - **full** length
  - no deterioration of quality
- The real world has improved but is not perfect:
  - Sanger
    - PCR
    - length: some kb
    - limited number of reads
    - high quality
  - HTS (Illumina)
    - PCR
    - 100-300 bp
    - Paired-end option (PE)
    - **billions** of reads
    - high quality, but deteriorating along read









Looking for a ping pong ball in a stadium

- With a hand torch!!

Or with the lights on?

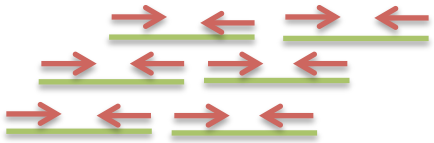
VARIANT CALLING WORKFLOW OVERVIEW

A quick overview of the HTS workflow

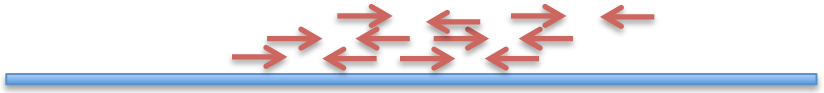
Fragment sample

Capture

Sequence

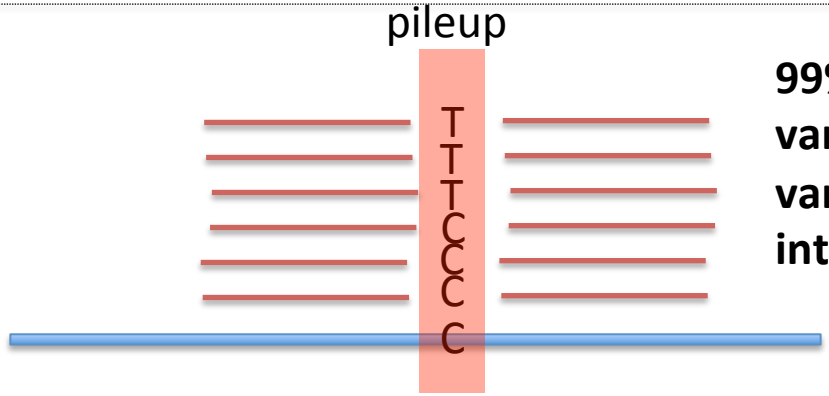


Map



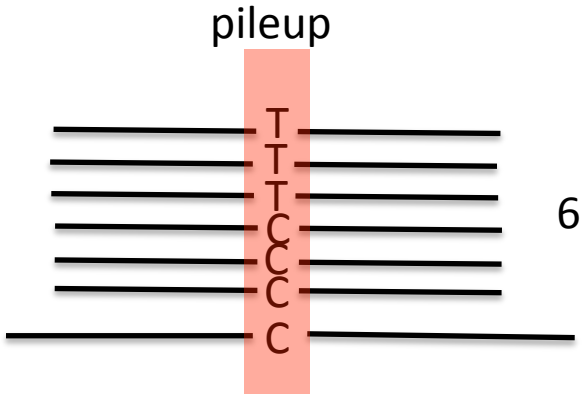
Align

Variant call



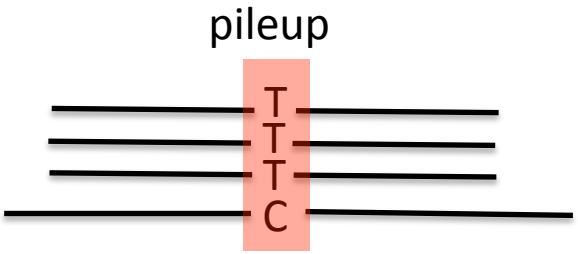
99% of sites are NOT variant, but it is the variant sites that are interesting

Variant sites



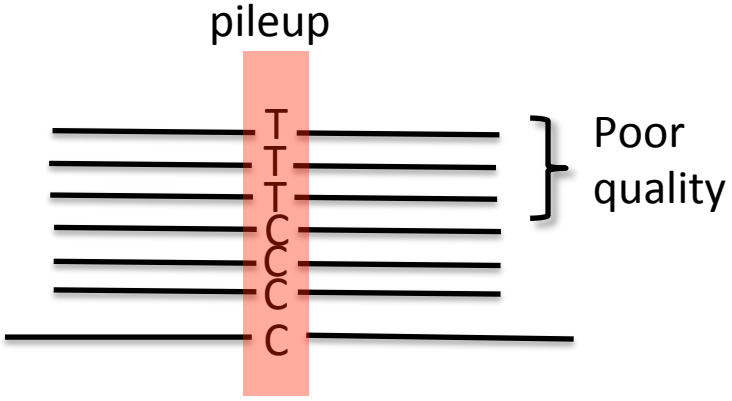
The common and easy case

- Good mapping of reads
- Good base qualities
- Good depth



Poor depth

- May not have sampled both alleles
- Could be C/T or T/T



Poor quality

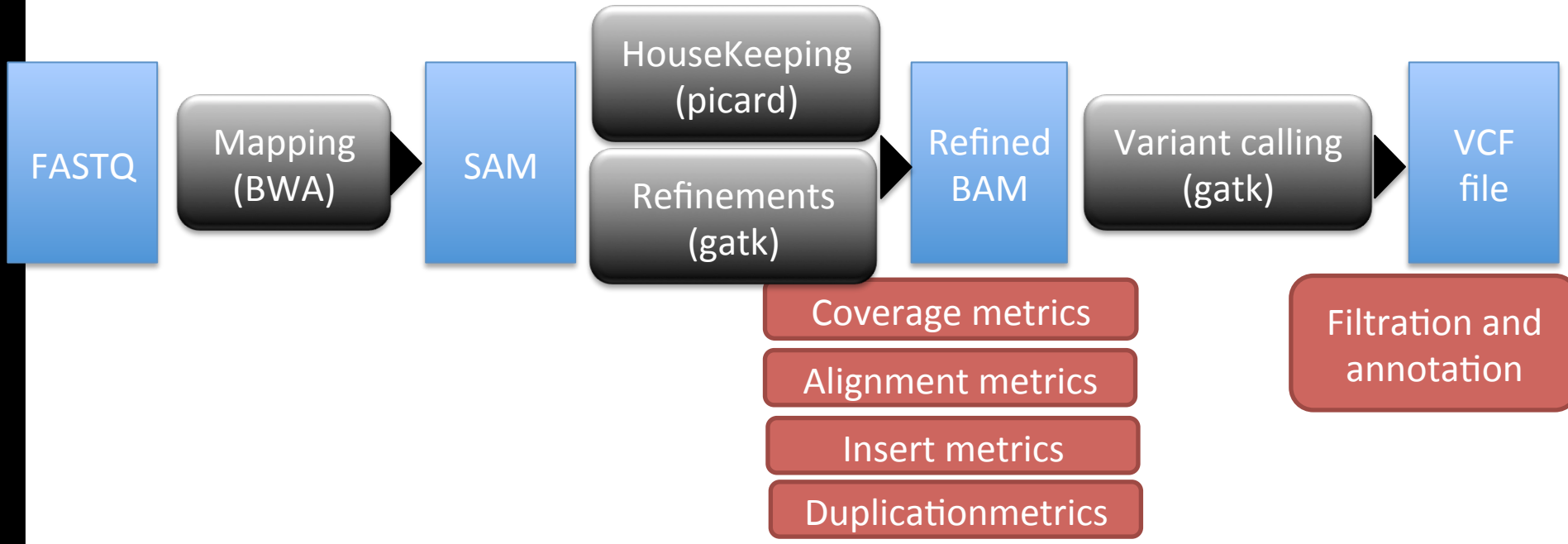
- of base calls
- of read mapping

Can lead to false variant calls: site could be ref

C/C and Ts are just base call errors

Our pedagogical approach

Multiple iterations with increasing complexity

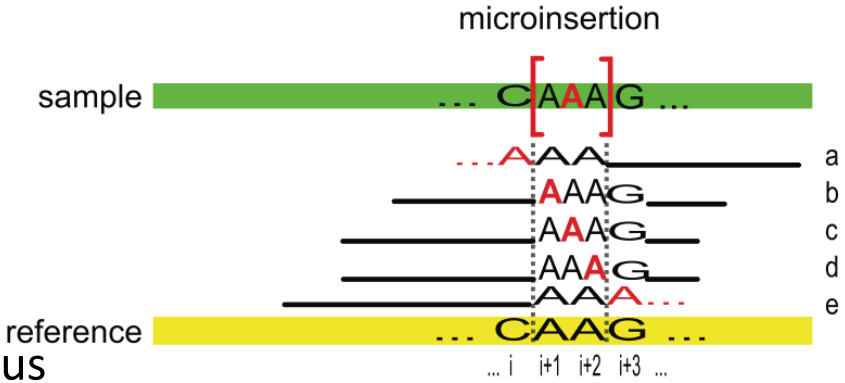


What do we aim to achieve with this approach

- Hoped for advantages of approach:
  - simple to understand most basic pipeline and easy to run
  - practice in each iteration
- Don't worry if you do not understand everything at the first iteration
- With bioinformatic pipelines it is generally a good idea to quickly create a simple version before make it more complex
  - prototyping
  - Check out Tom Wujec's TED talk "**The marshmellow challenge**"
- We will focus on single sample analysis, and will discuss more complicated setups later
  - multi-sample
  - pooled

Why do we need a “complex” pipeline?

- The biggest problem with variant calling is large numbers of FPs and FNs:
  - Mainly driven by bad alignments
  - FNs often create FPs
  - Can be systematic across samples, thus creating consistent SNPs across samples that are just machine artifacts.



- FPs and FNs, may result in:
  - the signal in our data drowning in noise ➔ no result
  - false results ➔ erroneous result

- You may find that reviewers will object to a pipeline which is not state-of-the-art, which creates trouble for you in the review process eventhough you may actually have a result.

Choice of example dataset

- Human data – arbitrary

- What if you work on other species?
  - if you have a reference genome and reads from a sample you can run the most basic pipelines
  - For more advanced pipeline you may have to spend some effort finding or generating auxiliary files in correct format
    - dbSNP
    - Genome annotation
  - You may find that they are already available for many species
    - Check out the GSA/GATK website
    - Check out http://getsatisfaction.com/gsa

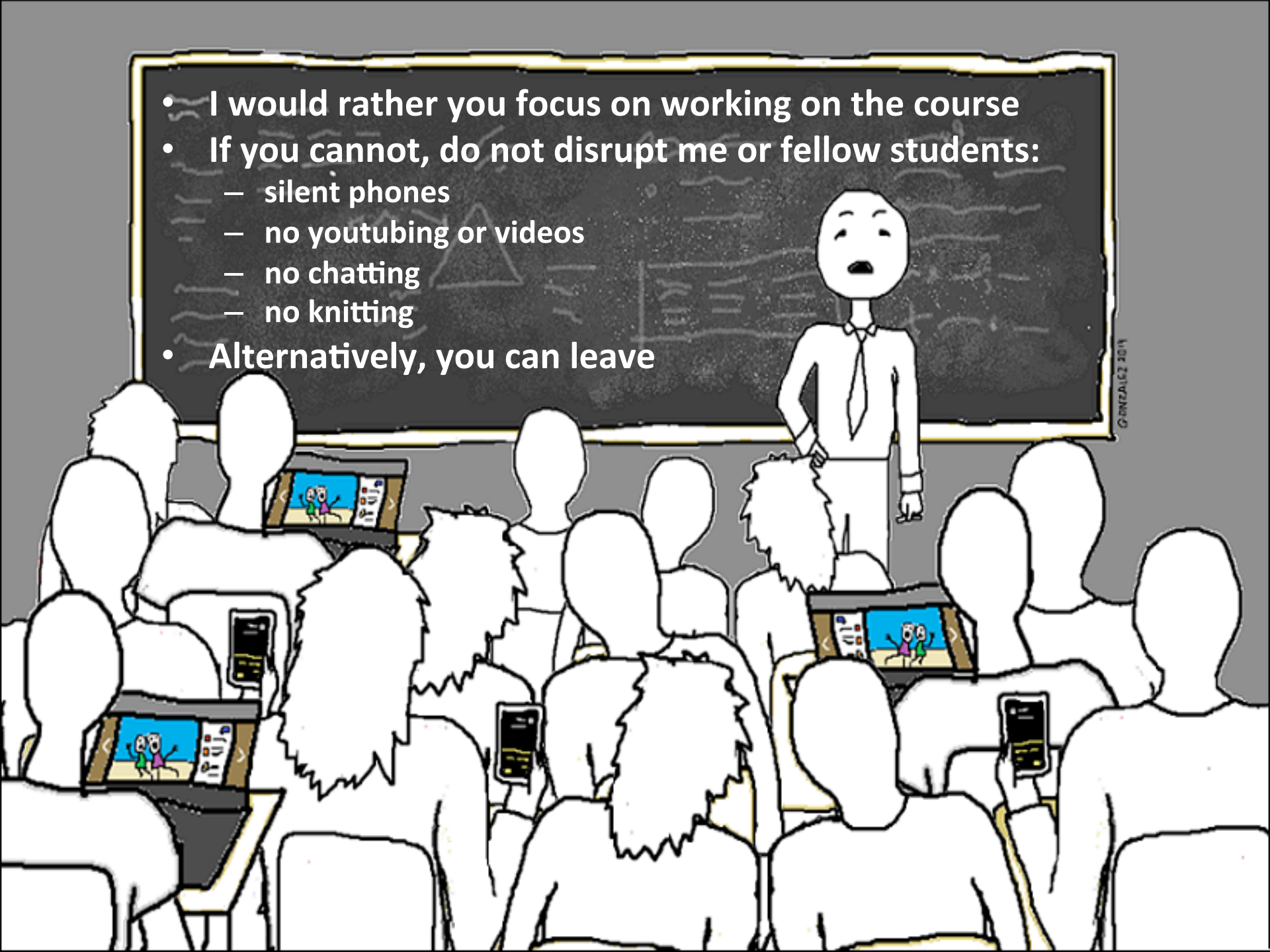
GSA team at the Broad Institute

- A large fraction of the materials and software in this course are produced by the **Genome Sequencing and Analysis Group** team at the Broad Institute
- Information sources
  - http://www.broadinstitute.org/gsa/wiki
  - http://www.getsatisfaction.com/gsa
- People
  - Mark A. DePristo, Manager of Medical and Population Genetics Analysis
  - Eric Banks, Team Lead
  - Guillermo del Angel
  - Ryan Poplin
  - Kiran Garimella, Team Lead
  - Mauricio Carneiro
  - Chris Hartl
  - Khalid Shakir, Team Lead
  - Matthew Hanna
  - David Roazen
- Others at the Broad
  - Heng Li: samtools and bwa
  - Tim Fennell: picard
  - Alec Wysoker: picard
  - Brac Chapman: bcbio.variation
- And others outside the Broad
  - sources at bottom of slides

RECOMMENDATIONS FOR PRACTICALS

For each practical

- Most key concepts explained up front
  - sometimes concepts will be first introduced in the practical (this will be the exception)
- Run through of the practical in plenum
- Read carefully through the practicals. All the information you need for doing the practical should be in the file
- Practical session
  - Try to get it running
  - If fail or computation takes time ➔ go to exercisesResults directory
- You will be doing “copy and paste”
  - that is OK
  - but you must understand what you are copying and pasting
  - if you don’t ask for explanations
- Use of labels to indicate status: “finished” and “needing help”
- It is hard!
  - new concepts
  - new environment (Unix)
  - try to understand both **and take notes!**


• I would rather you focus on working on the course
• If you cannot, do not disrupt me or fellow students:
– silent phones
– no youtubing or videos
– no chatting
– no knitting
• Alternatively, you can leave

Other dos and donts

- Do
  - Sign the attendance sheet
  - Remember that there are wide differences in competance
  - Consolidate your knowledge if you get ahead
  - Talk and get help from your neighbours

- Do **not:**
  - ask questions in plenum if you get ahead, instead consolidate what you have learnt (tool websites)
  - think that the slides are a substitute for taking notes

Unix attitude

- **Get used to having to know where you are and to tell the computer where things are**

```
java -Xmx2G -jar ${swDir}/picard-tools-1.67/MarkDuplicates.jar \
INPUT=aln.posiSrt.clean.bam \
OUTPUT=aln.posiSrt.clean.dedup.bam \
METRICS_FILE=aln.posiSrt.clean.dedup.bam_metrics.duplication.txt
```

- **Be as precise as you are in your field**

../vc/exerSandbox/03_advancedPipeline

- **Pay attention to detail and changes**

```
-rw-rw-r-- 1 timothyh timothyh  26M Oct  2 14:30 aln.posiSrt.clean.dedup.bam
-rw-rw-r-- 1 timothyh timothyh 145K Oct  2 14:30 aln.posiSrt.clean.dedup.bai
```

- **String commands together if you can**

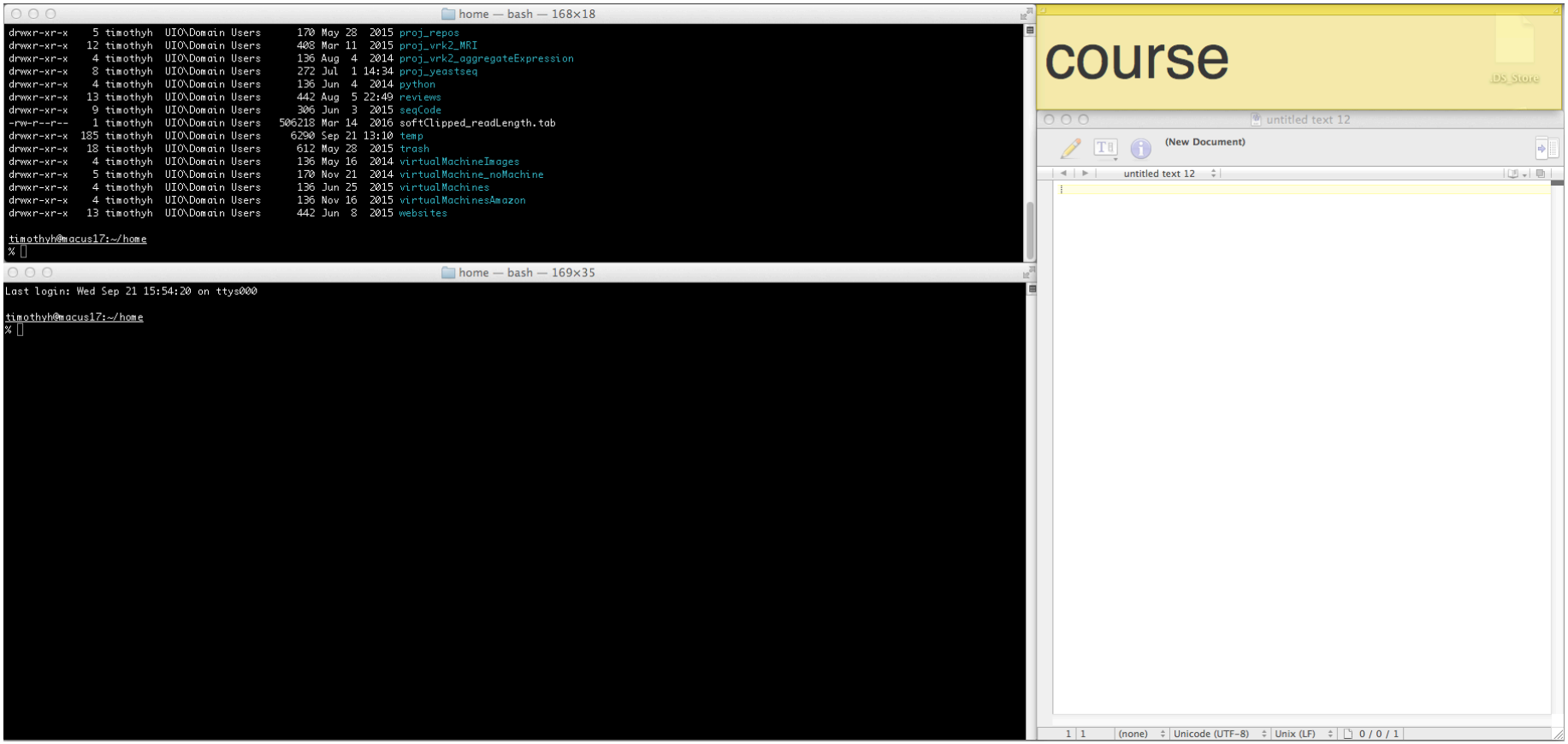
```
cat ${dataDir}/reads_agilentV1_chr5/simul_indels/deletions.vcf | grep -v "^#" | wc -l
```

Some useful unix shortcuts

- Ctrl + a: go to beginning of line
- Ctrl + e: go to end of line
- Ctrl + k: remove text after pointer
- Ctrl + u: remove text before pointer
- Arrow up and down to scroll through history
- Ctrl + r: reverse search history

Organising your desktop

- Keep an electronic notepad of file paths
- Useful to have multiple terminals
- Sometimes issues with code over several lines >> use wide terminals



Kahoot!

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

And punch in the code for the ***Unix skills and bash variables*** *quiz*

I will explain answers as we go along

Why use variables for file paths?

- There are many different files in the exercises
  - with long paths (bcse there is a directory structure)
- A lot of errors in exercises are due students getting file paths wrong AND the primary goal of this course is not to teach file paths
- If one wishes to change a file path that is used in multiple locations / exercises one needs to make multiple changes in multiple locations
- You should learn to use variables in your own scripts
- If you ever get confused, just execute the following command to see what a variable contains
  - echo $variableName