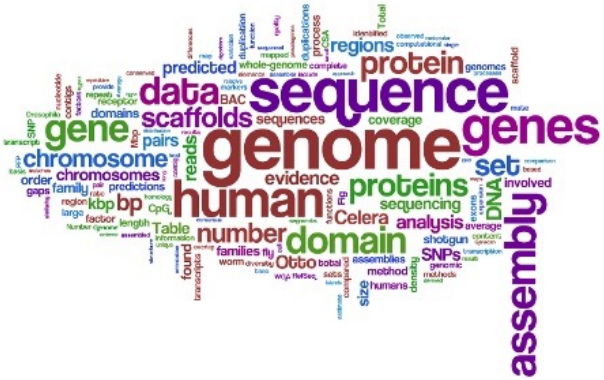
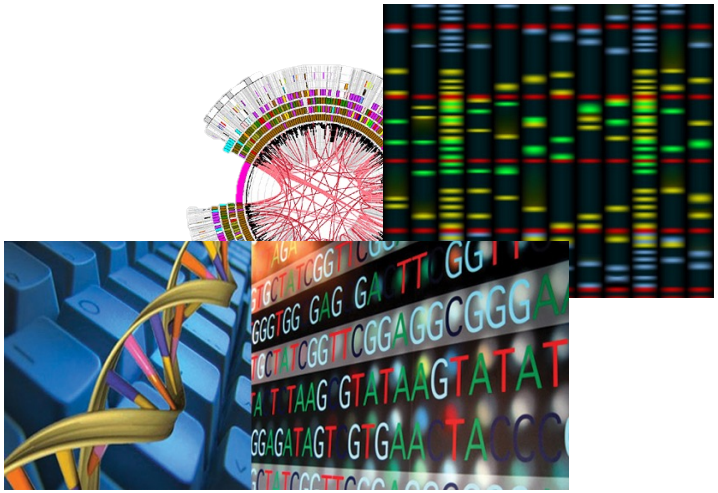


Cranfield
Environment
and Agrifood

Next Generation Sequencing Informatics

Module Intro



Prof Fady Mohareb

Head of Bioinformatics

www.cranfield.ac.uk

Attendance QR code

Module Aims

- **Module Aims**
  - To introduce the techniques that have given rise to the genomic data now available. (Practical approach)
  - Develop skills and understanding in the bioinformatics approaches that facilitate evaluation and application of these data.
  - Introduction to Next-generation DNA Sequencing (NGS), the technology that has been a huge stimulus for a lot of breakthrough discoveries in biology.
  - Overview of the application of NGS in:
    - Genomics of diseases for medical life science (Monday)
    - Agriculture sciences (Cranfield Plant Molecular Biology Group – Thursday)
    - NGS Applications in Cranfield (cross-disciplinary) + Genotyping (Friday)

KSBs (For the Bioinformatics Apprenticeship)

Knowledge (Ks):

K2: How research is conducted in bioinformatics and within the broader context of interdisciplinary life sciences (All).

K4: Details of omic-scale/big-data-driven life science making use of core platform technologies (Tuesday).

K10: Techniques to integrate, interpret, analyse and visualise biological data sets (Mon-Wed)

K11: Bioinformatics analysis methodologies and expertise in common bioinformatics software packages, tools and algorithms – including workflow management tools (All week)

K14: Licensing limitations on the use of bioinformatics software and data such as open source, commercial and academic usage restrictions. (Tue-Fri using open-source tools on Crescent)

K16: Relevant big-data and high performance computing platforms including Linux/Unix, local and remote High Performance Computing (HPC), and cloud computing (as above

KSBs (For the Bioinformatics Apprenticeship)

Skills (Ss):

S23: Work with multi-disciplinary colleagues to design life-science experiments that will generate data suitable for subsequent bioinformatics analysis (All).

S24: Provide guidance to experimental scientists on data generation methodology and handling to ensure the quality of data produced (Wed – Thurs).

S25: Recognise and critically review the format, scope and limitations of different biological data.

(Tues Practical + Assignment)

S35: Determine the best method for bioinformatics analysis, including the selection of statistical tests, considering the research question and limitations of the experimental design (Assignment).

S36: Identify and define appropriate computing infrastructure requirements for the analysis of such biological data (Assignment).

S39: Build and test analytical pipelines, or write and test new algorithms as necessary for the analysis of biological data (Assignment)

Topics Covered

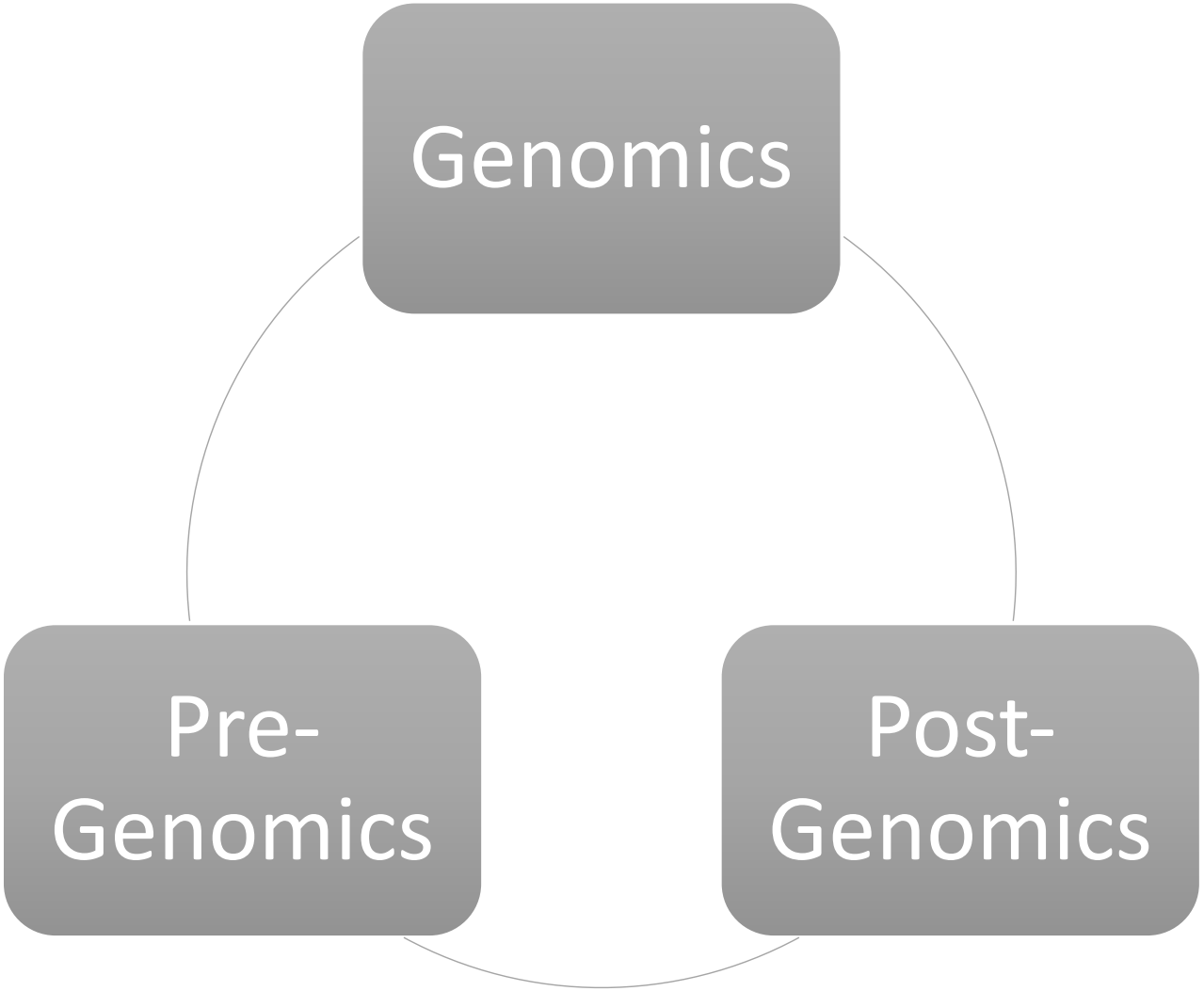
- Array Chips (Monday)
  - Microarray Technologies
  - Microarray practical (data per-processing, normalisation, downstream analysis)
- Next Generation Sequencing (Rest of the week)
  - NGS Quality Control
  - RNA-Seq (using Crescent, Cranfield's HPC, and potentially Galaxy)
  - Variant Calling (Samtools, GATK)

- Advanced NGS and genome assembly → Module 7

Timetable

Mon	Tue	Wed	Thu	Fri
11	12	13	14	15
:30 Next Generation Sequencing Informati odule Introduction C Lab 50 Computer Room [Cranf - Bldg 50]	**09:30** Next Generation Sequencing Informati Next Generation Sequencing Platforms PC Lab B50 Computer Room [Cranf - Bldg 50]	**09:30** Next Generation Sequencing Informati Introduction to RNA-Seq PC Lab B50 Computer Room [Cranf - Bldg 50]	**09:30** Next Generation Sequencing Informati NGS application in In Plant Science PC Lab B50 Computer Room [Cranf - Bldg 50]	**09:30** Next Generation Sequencing Informati Variant Calling & Genotyping PC Lab B50 Computer Room [Cranf - Bldg 50]
:30 Next Generation Sequencing Informati enomics Microarrays and Gene Expression C Lab 50 Computer Room [Cranf - Bldg 50]	**11:00** Next Generation Sequencing Informati NGS Informatics PC Lab B50 Computer Room [Cranf - Bldg 50]	**11:00** Next Generation Sequencing Informati Introduction to RNA-Seq PC Lab B50 Computer Room [Cranf - Bldg 50]	**11:00** Next Generation Sequencing Informati NGS application in In Plant Science PC Lab B50 Computer Room [Cranf - Bldg 50]	**11:00** Next Generation Sequencing Informati Practical SNP calling PC Lab B50 Computer Room [Cranf - Bldg 50]
:00 Next Generation Sequencing Informati enomics Microarrays and Gene Expression (C Lab 50 Computer Room [Cranf - Bldg 50]	**14:00** Next Generation Sequencing Informati NGS quality control & Data pre-processing PC Lab B50 Computer Room [Cranf - Bldg 50]	**12:00** Next Generation Sequencing Informati RNA-Seq Practical PC Lab B50 Computer Room [Cranf - Bldg 50]	**12:00** Next Generation Sequencing Informati DNA Mrkers & Marker assisted Selection PC Lab B50 Computer Room [Cranf - Bldg 50]	**14:00** Next Generation Sequencing Informati Variant Calling & Genotyping PC Lab B50 Computer Room [Cranf - Bldg 50]
:00 Next Generation Sequencing Informati actical: Microarray Tutorial C Lab 50 Computer Room [Cranf - Bldg 50]	**15:30** Next Generation Sequencing Informati NGS quality control & Data pre-processing PC Lab B50 Computer Room [Cranf - Bldg 50]	**14:00** Next Generation Sequencing Informati RNA-Seq Practical PC Lab B50 Computer Room [Cranf - Bldg 50]	**14:00** Next Generation Sequencing Informati Alignment-free: RNA-Seq PC Lab B50 Computer Room [Cranf - Bldg 50]	**14:45** Next Generation Sequencing Informati Practical SNP calling PC Lab B50 Computer Room [Cranf - Bldg 50]
:45 Next Generation Sequencing Informati actical: Microarray Tutorial (cont) C Lab 50 Computer Room [Cranf - Bldg 50]		**15:30** Next Generation Sequencing Informati RNA-Seq Practical PC Lab B50 Computer Room [Cranf - Bldg 50]	**15:30** Next Generation Sequencing Informati Alignment-free: RNA-Seq PC Lab B50 Computer Room [Cranf - Bldg 50]	

Is it all just “genomics”?



PCR: What is PCR?

- Polymerase used to copy DNA strands, but repetitively in cycles of amplification
- End result is the production of millions of copies of a defined sequence of DNA
- Area to be amplified is defined by oligonucleotide primers

PCR History

- Developed in the late 1980s by Kary Mullis (December 28, 1944 – August 7, 2019) who realised the potential of a thermostable DNA polymerase from *Thermus aquaticus* (Taq)
- First performed in a series of water baths
- Development of thermal cyclers
- Variants on basic technique, inc. Real time PCR

In each cycle of PCR…

- The double stranded DNA must be denatured into two single strands by heating the reaction mixture to about 95° C
- Specific primers define the sequence to be amplified, and give the polymerase a point at which to start working.
- Primers anneal when the reaction temperature drops close to their melting temperature (T_m), often around 55° C
- Extension by sequential addition of nucleotide bases by polymerase occurs at around 72° C

In each cycle of PCR…

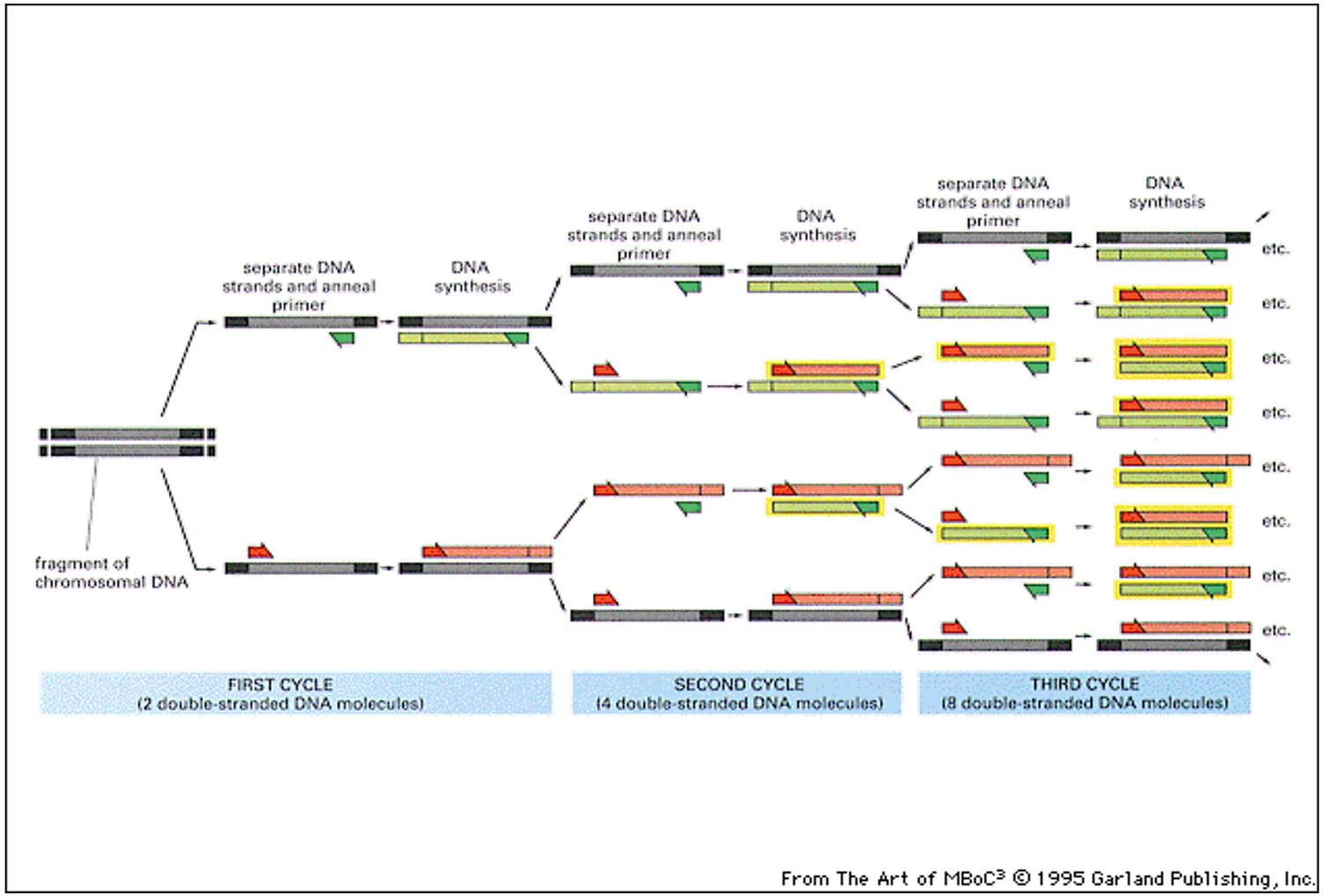
- The double stranded DNA must be denatured into two single strands by heating the reaction mixture to about 95° C
- Specific primers define the sequence to be amplified, and give the polymerase a point at which to start working.
- Primers anneal when the reaction temperature drops close to their melting temperature (T_m), often around 55° C
- Extension by sequential addition of nucleotide bases by polymerase occurs at around 72° C

So basically…

- *Denature* DNA strands
- *Anneal* primers
- *Extend* the DNA sequence with a polymerase

And repeat at least 25 times to get millions of copies (theoretical yield: 2^{25} = 33554432)

Repeating cycles = amplification:



RT-PCR or RT-PCR?

A slight tangent ...

Reverse Transcription PCR

Real Time PCR

RNA and RT(reverse transcription)-PCR

- RNA cannot be amplified directly by PCR because the difference in base composition is not compatible with the way a double stranded DNA is produced (U instead of T)
- Use of RNA as a template first requires reverse transcription to a cDNA strand, which can then be used as a template for PCR
- Reverse transcription also needs a primer in addition to those required for PCR
- RT-PCR detects gene expression and RNA viruses

PCR v RT (real time)-PCR (or qPCR)

Cranfield UNIVERSITY

Standard PCR v Real-time PCR

- Requires down-stream processing (gels)
- Qualitative/semi-quantitative
- End-point detection
- Short dynamic range <2 orders of magnitude

- No down-stream processing
- Quantitative –most accepted method compared to MA
- Exponential phase detection
- Large dynamic range <8 orders of magnitude
- Less starting quantity needed
- Rapid cycling

Real-time PCR

- Uses principles of standard PCR, and in addition fluorescence is produced & increases in proportion to the amount of PCR product in the reaction, hence the term ‘real time’

- PCR product is measured during exponential phase as opposed to end-point detection

- Detect and quantify amplified DNA product without having to run a gel

The LightCycler® 480 Real-Time PCR System

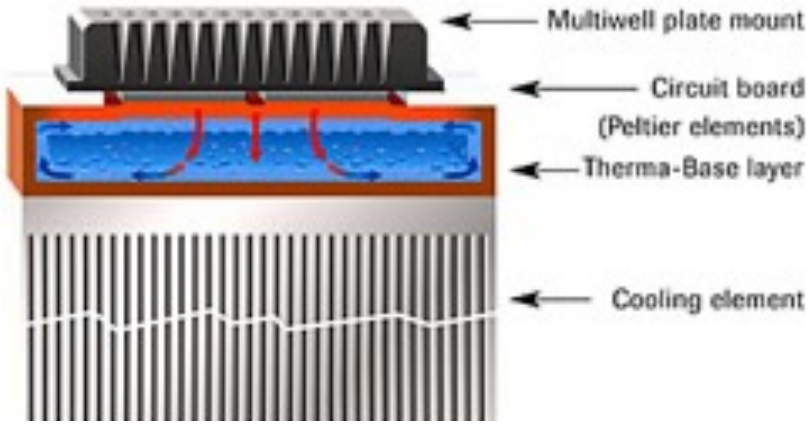
- Thermal Cycler for amplification
- Light source for excitation of fluorescent probes
- Camera for recording fluorescence
- Computer to control instrument, run software and record & analyse data

Thermal Cycler

- Heating and cooling achieved using Peltier elements: semiconductor elements, exact temperatures in the -50° to 200° C range generated through electric current

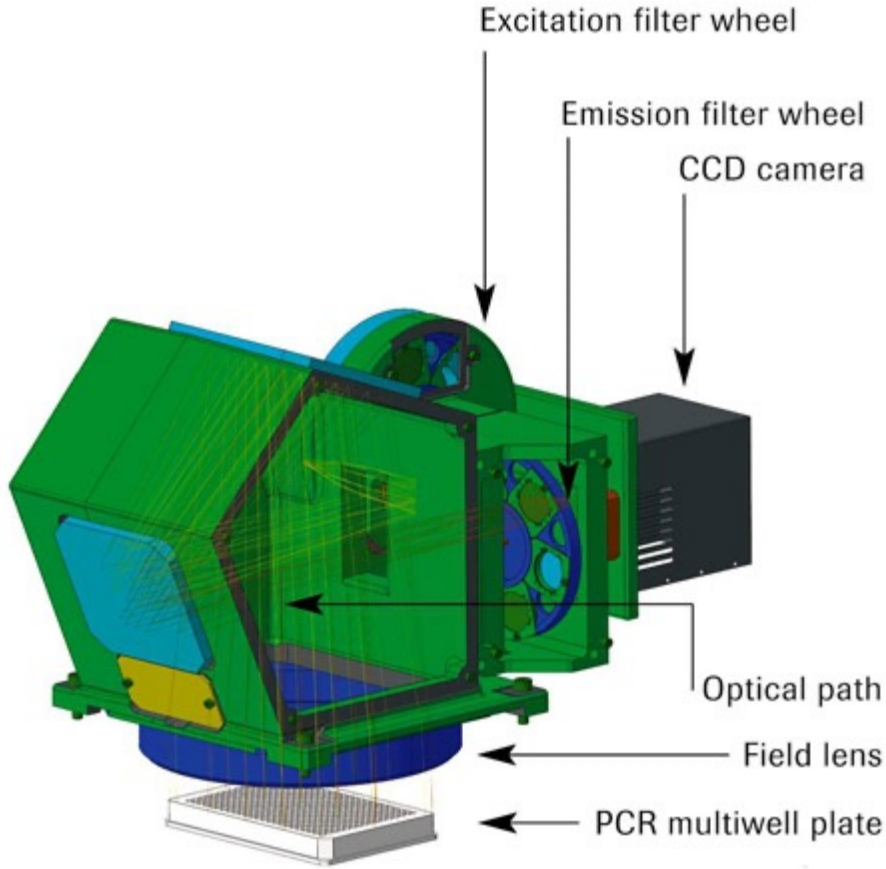
- Additionally, a thermal block ensures optimal heat transfer and distribution to all samples on multiwell plate

- The cooling element below has maximized inner surface area to facilitate rapid heat absorption



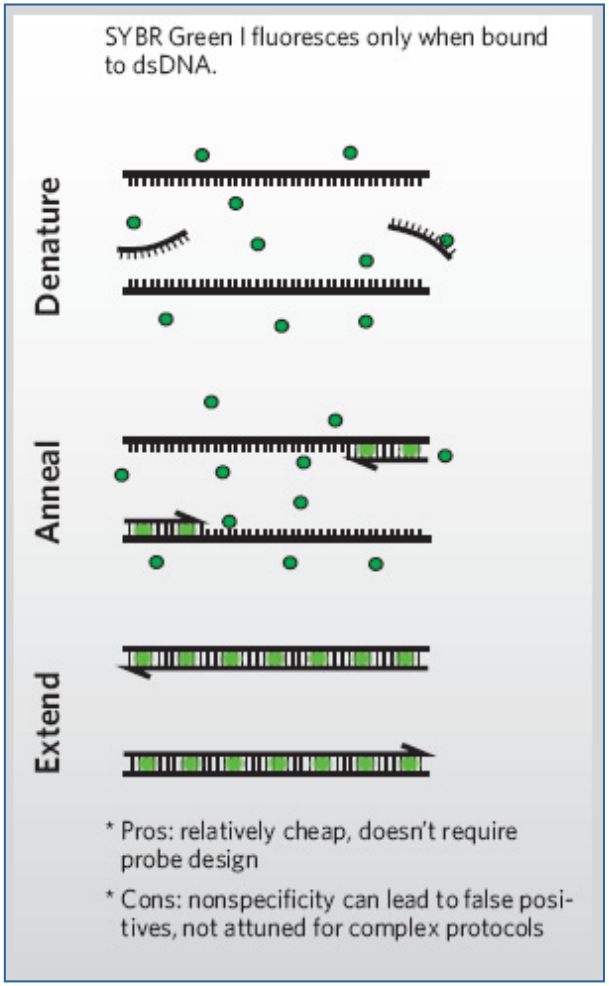
LightCycler® 480 detection unit

- CCD camera records fluorescence from behind emission filters
- A high-intensity Xenon lamp emits light over a broad wavelength range (430–630 nm). The five excitation and six emission filters (placed in filter wheels) of the system can be used in any combination
- The choice between multiple filter combinations enables optimized excitation & detection. Useful for multiplexing

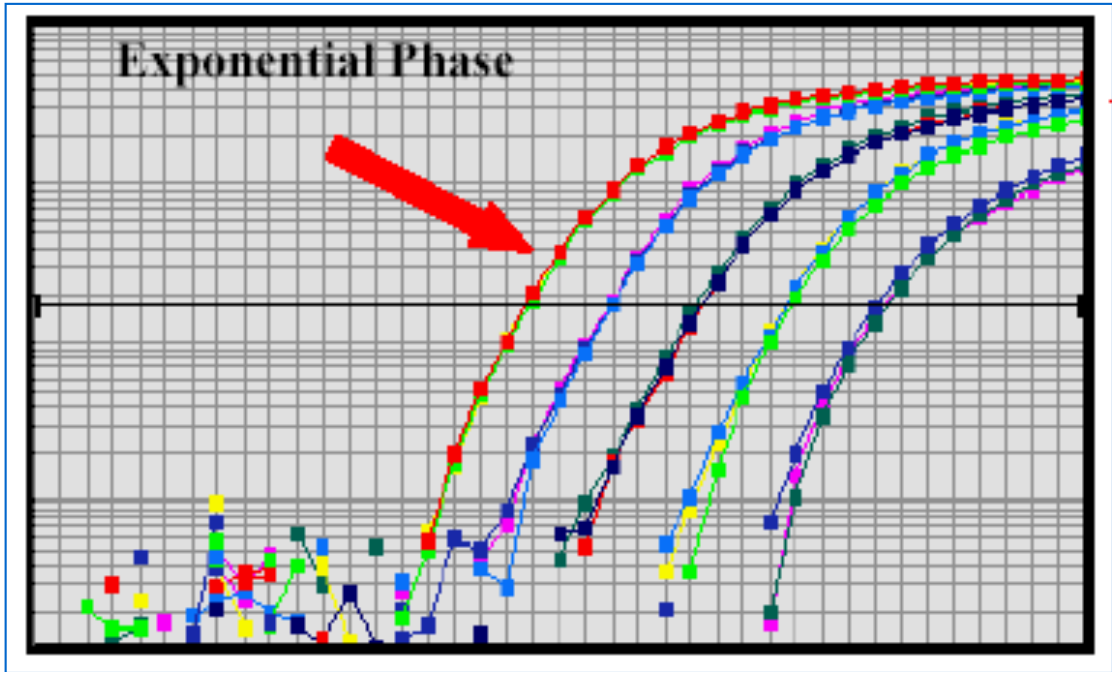


SYBR Green I technique

- SYBR Green I fluorescences when binds to double-stranded DNA during the extension phase



- A five-fold dilution series plateaus at the same place even though the exponential phase clearly shows a difference between the points along the dilution series
- Reinforces fact that using measurements taken at the plateau (end-point) phase, does not truly represent the initial amounts of starting template



Think about the Human Genome

"since the completion of the genome project ..."

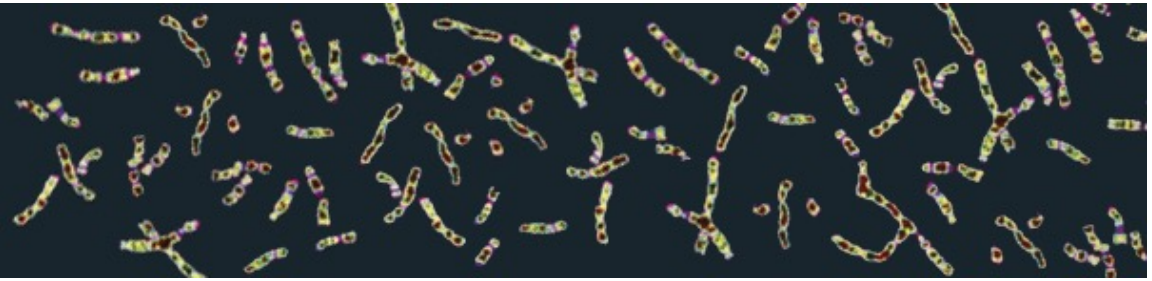
Is it "complete" ?

Approx 3Giga bases

Where - who - does the sequence really come from?

1000 Genomes

A Deep Catalog of Human Genetic Variation



Pilot Project

Three pilot studies provided data to inform the design of the full-scale project:

Pilot	Purpose	Coverage	Strategy	Status
1 – low coverage	Assess strategy of sharing data across samples	2–4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 – trios	Assess coverage and platforms and centers	20–60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 – gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

http://www.1000genomes.org/

Encode

http://www.nature.com/encode/#/threads

- 31 simultaneous publications in Nature, Genome Research, Genome Biology, and BMC Genetics

ENCODE, the Encyclopedia of DNA Elements, is a project funded by the National Human Genome Research Institute to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence. Thanks to the identification of these functional elements, 80% of the human genome now has at least one biochemical function associated with it. This expansive resource of functional annotations is already providing new insights into the organization and regulation of our genes and genome.
http://genome.ucsc.edu/ENCODE/

Cancer Genome Project

Data resources

[Cancer Gene Census:]
Mutated genes causally implicated in human cancer.

[COSMIC:]
Catalogue of somatic mutations in cancer

[CGP resequencing studies:]
Somatic mutations from systematic large scale resequencing of genes in human cancers.

[CGP cancer cell line project:]
Resequencing of known cancer genes and other analyses of human cancer cell lines.

[CGP copy number analysis in cancer:]
Analysis of copy number and loss of heterozygosity in cancer cell lines and primary tumours.

[CGP trace and genotype archive:]
Archive of sequence traces and genotype data generated by the group.

[Genomics of drug sensitivity in cancer:]
Analysis of drug sensitivity data in human cancer cell lines.

Cranfield Environment and Agrifood

Genomics england

Google Custom Search

About Us | 100,000 Genomes Project | Taking Part | For Healthcare Professionals | Research | Industry Partnerships | New

Home > The 100,000 Genomes Project

The 100,000 Genomes Project

The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

The aim is to create a new genomic medicine service for the NHS – transforming the way people are cared for. Patients may be offered a diagnosis where there wasn't one before. In time, there is the potential of new and more effective treatments.

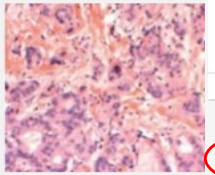
The project will also enable new medical research. Combining genomic sequence data with medical records is a ground-breaking resource. Researchers will study how best to use genomics in healthcare and how best to interpret the data to help patients. The causes, diagnosis and treatment of disease will also be investigated. We also aim to kick-start a UK genomics industry. This is currently the largest national sequencing project of its kind in the world.

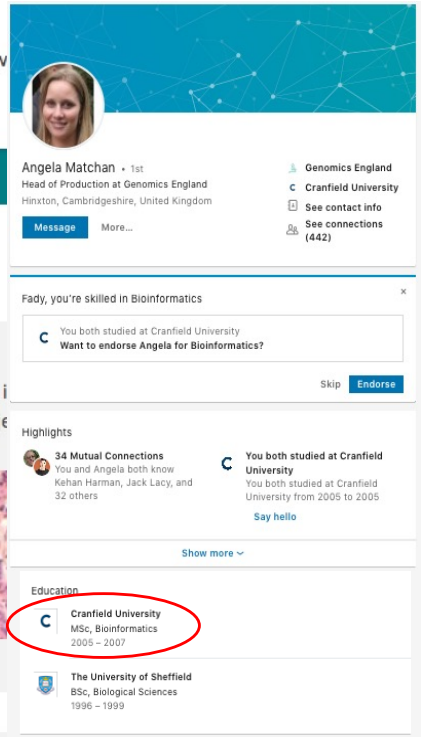
Introduction to the 100,000 Genomes Project

History of the 100,000 Genomes Project

Useful links

Cancer
Introduction to cancer i
100,000 Genomes Proje





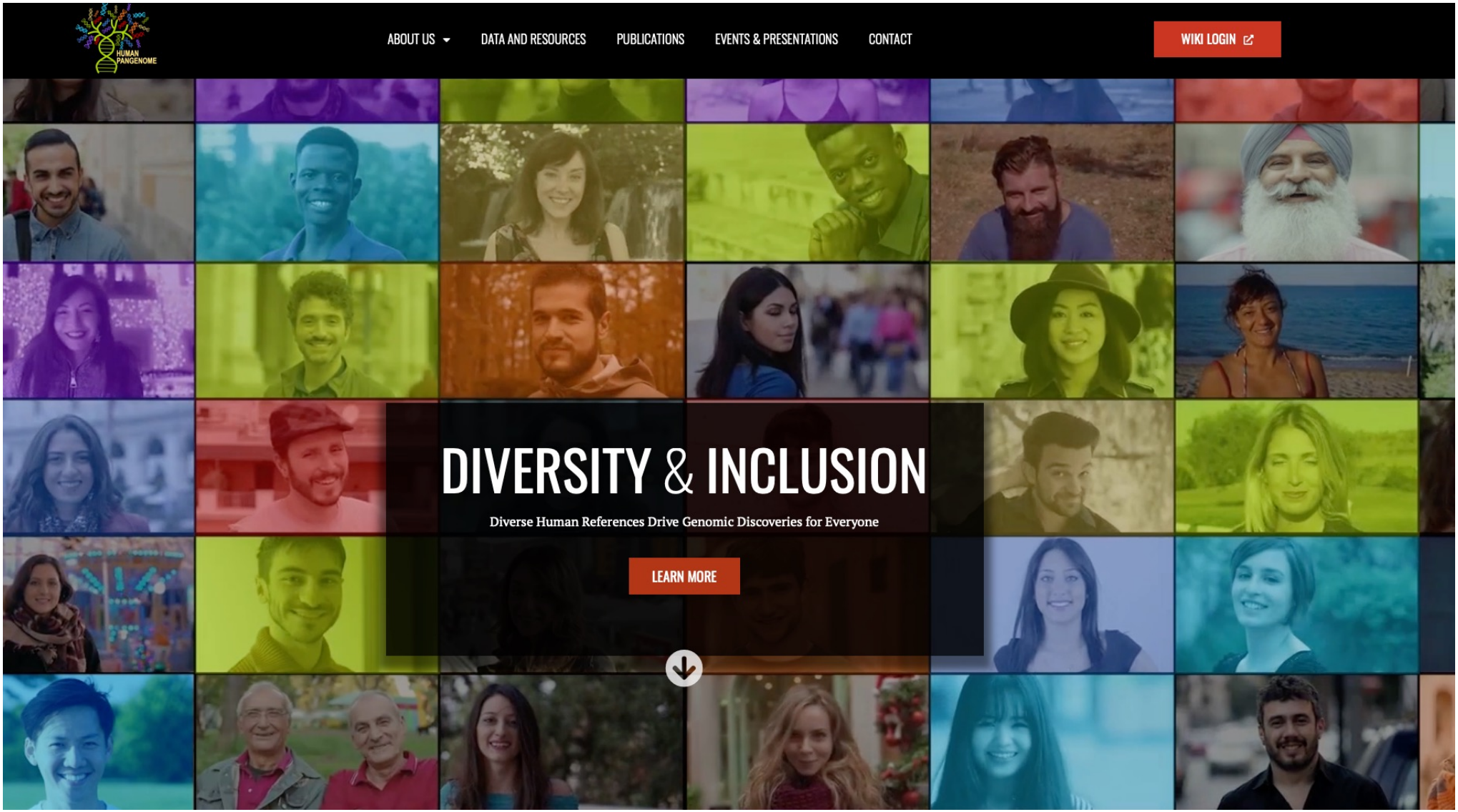
Taking part
Information about taking part in the Project

Insurance
Find out how taking part in the Project may affect insurance.

The UK Biobank

- A bit more than 500,000 volunteers in the UK (40 to 69 old).
- Volunteers are followed for at least 30 years.
- All volunteers are genotyped (using WGS and SNP Array chips)
- Data is available for free for research purposes
- Forms a fundamental basis for several ground-breaking GWAS analysis

THE HUMAN PAN-GENOME PROJECT



Coffee Break