BS320 Human Genetics

Prof Leo Schalkwyk lschal@essex.ac.uk

Lecture 1
Introduction: from the human genome project to the 100,000 genomes project: strategies, achievements and prospects

objectives:

- Explain the sequencing strategies and methods used to assemble the human genome sequence
- Explain the impact of genome projects on approaches to studying human disease

Reading

Human Molecular Genetics 4, Strachan, T. and Read, A. (2010)
Garland Science, Chapters 8-9

Strachan, T., Goodship, J and Chinnery, P. (2014) Genetics and Genomics in Medicine, Garland Science

Basic texts pdfs on Moodle

Terminology in lectures:
Glossary
http://www.genome.gov/glossary/index.cfm

Via library portal:
Annual Reviews (Genetics; Genomics and Human Genetics)
Nature; Nature Reviews Genetics; Science

Online reading:
Nature resource
http://www.nature.com/scitable

Web-based resources:

- http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
- http://www.ebi.ac.uk/
- http://www.ensembl.org/index.html
- http://genome.ucsc.edu/

These resources allow the reader to access data and provide tools to analyse it, together with links between different resources.

a historical perspective

- mid-1950s cytogenetics 24 human chromosomes
- classical mapping (gene variants) almost impossible
- from late 1970s DNA-based variation and use of non-coding variation
- Human Genome Organisation (HUGO) established 1988

the Human Genome Project (HGP)

official start 1990, goals:

- development of technologies and tools: genetic and physical mapping, sequencing, databases and bioinformatics
- genome projects for 5 model organisms *E. coli, S. cerevisiae, C. elegans, D. melanogaster,* mouse
- ethical, legal and social implications

The HGP web archive

http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Cold Spring Harbor Laboratory: you can listen to interviews with the scientists involved, watch animations

- http://www.dnalc.org/home.html
- http://www.dnai.org/

a strategy to approach the HGP

- large-scale cloning methods
- mapping (physical and genetic)
- sequence based on this foundation
- data channelled into sequence databases for rapid open electronic access

originally planned as 15 year project to end 2005

5 main centres:

- Sanger Centre (Cambridge)
- 4 US centres (Whitehead Institute and MIT in Massachussetts, Washington University, DoE Joint Genome Institute and Baylor College of Medicine)

public vs private

- HGP (James Watson, Francis Collins)

- Celera Genomics (Craig Venter)

http://www.nature.com/scitable/topicpage/sequencing-human-genome-the-contributions-of-francis-686

Collins had a leading role in the HGP taking a strategic mapping based approach having succeeded in cloning some major disease genes (CF, DMD) as a determined and elegant research scientist.

Initially Venter worked within the HGP consortium, but he clashed with them over strategy and personality. There is no doubt his more high-tech and faster approach raised the game and pioneered some approaches that led to more rapid progress.

Map-based sequencing

random sequencing

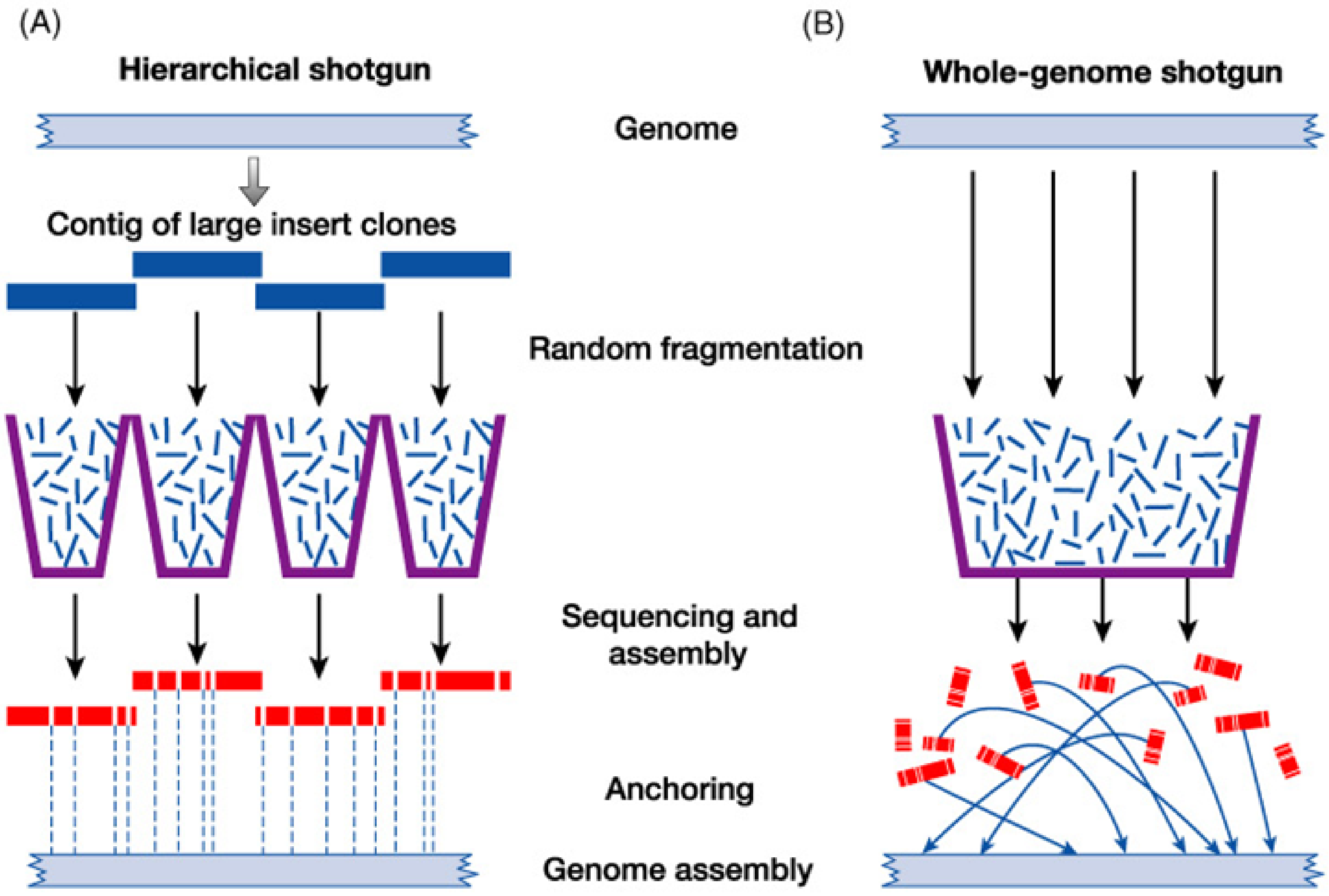


Figure 8-3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Known gene and marker sequences were essential for HGP sequence assembly

These markers locate sequences on chromosomes, in specific regions

- unique sequences, often showing genetic variation

- since they can be mapped by linkage to other markers and also identified in the DNA sequence they are a link between the physical and genetic maps

repetitive sequences caused problems with assembly

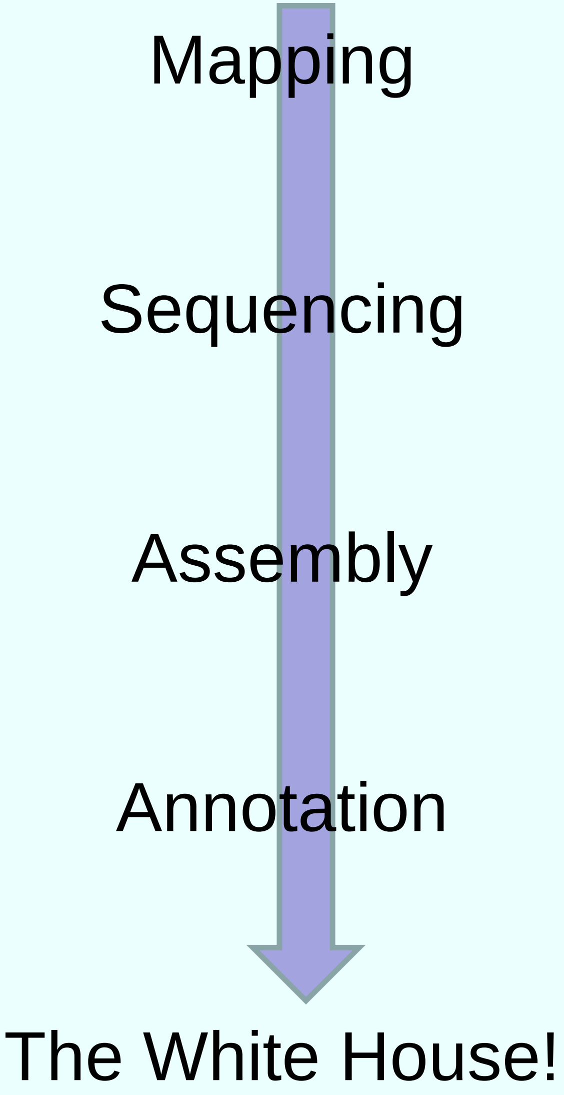
problems with identifying “real” overlaps between clones and sequences can lead to incorrect assembly of the sequence

•highly repetitive sequences (tandem repeats and transposable elements present in 1000s of copies)

•low copy number repeats (duplications of segments of chromosomes eg. by crossing over)

milestone achievements

- 1992 Généthon lab first human genetic map
- 1993 Généthon lab first human physical genome map
- 1995 Whitehead Institute/MIT first detailed physical sequence-tagged site (STS) map
- GeneMap ’98 Sanger Centre
- chromosome 22: Sanger Centre-led
- 2001 drafts from HGP and Celera
- 2003 human genome sequence complete
- May 2003: 140 genome sequences



Mapping

Sequencing

Assembly

Annotation

The White House!

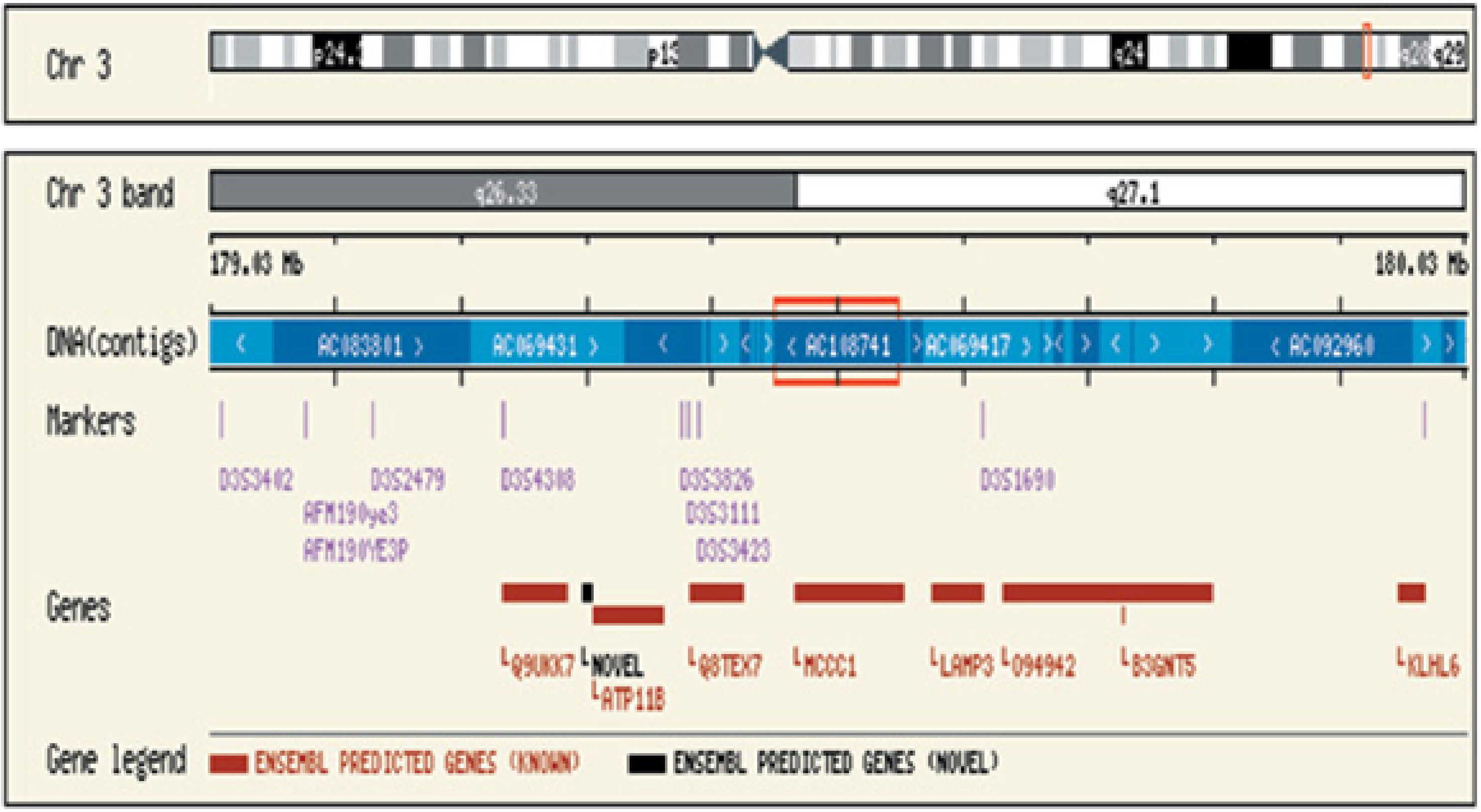


Figure 8-5 part 1 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

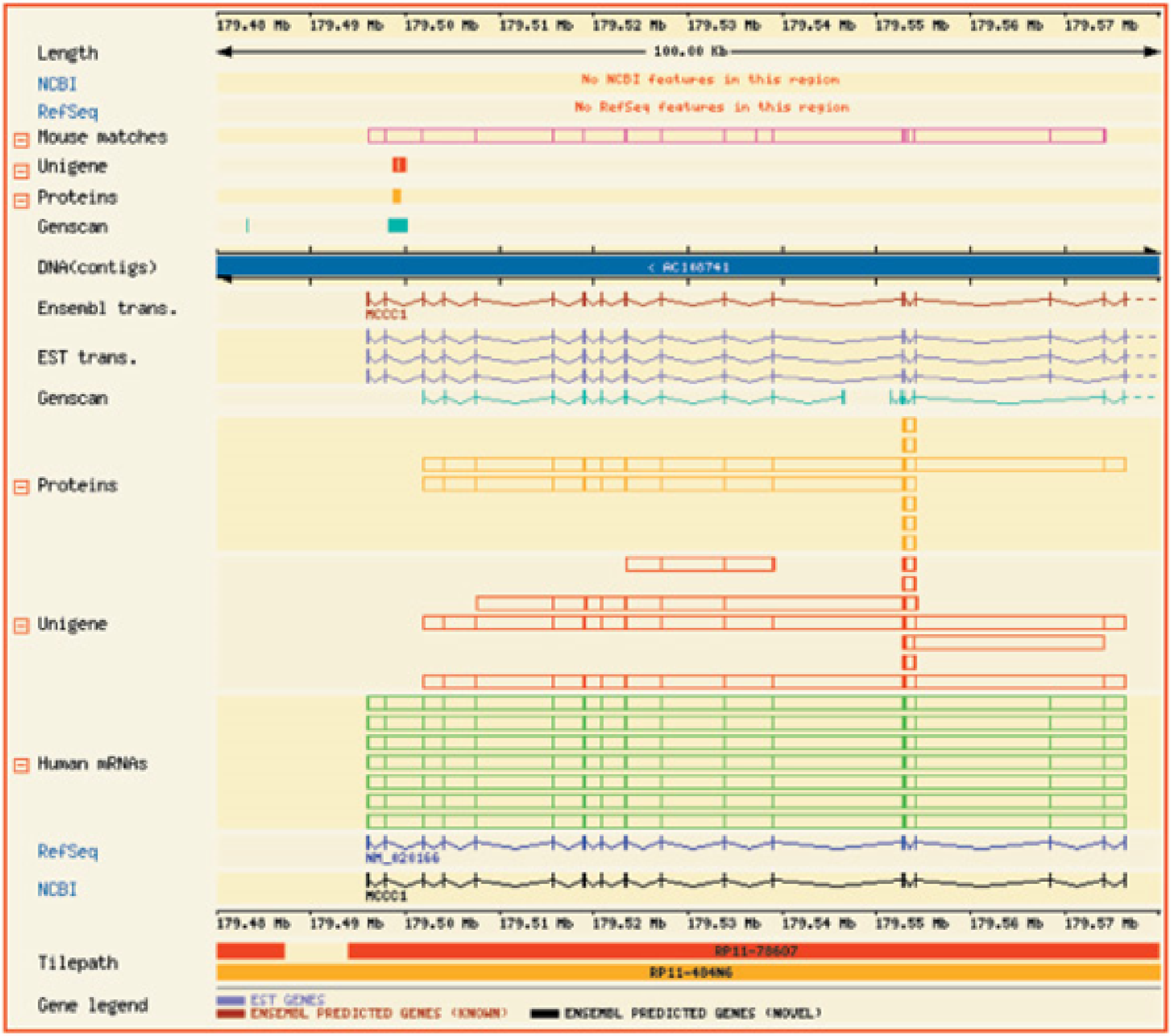


Figure 8-5 part 2 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

identifying genes in the genome sequence

- homology searches- compare the new sequence against all known sequences in the databases- to find genes already known and characterised in human/other species
- exon “prediction”- use the genetic code to find sequences that can be theoretically translated into polypeptides
- integrated gene-finding software (algorithms designed to search for homology, gene associated sequences eg. promoters)
- many small or 1 large gene?- challenge to get the correct intron/exon structure
- small exons/genes are also hard to find

how good are algorithms at predicting protein coding genes?

- estimates suggest they can detect about 90% of coding sequences
- And within the gene about 75% of exons are correctly identified
- overall the gene structure predicted is right about 50% of the time
- Genome sequences have human curators who check the computer predictions

how good are algorithms at predicting RNA genes? (answer: very good at some and poor at others)

- estimates suggest they can detect about 99.5% of tRNA sequences
- 18S, 26S rRNA: none in genome assemblies, they are located in very repetitive regions as multiple copies so sequence assembly is extremely difficult
- snRNA and snoRNA – non-coding RNAs associated with gene expression, mRNA processing etc
- miRNA, lncRNAs present new challenges and will be explored later in the module

future developments- what next once the genome sequence is known

- medical applications
  - genome diversity
  - SNP maps

- functional genomics: characterise all the functional elements, not just protein-coding genes

- comparative genomics: understanding human genome evolution

ethical, legal, social impact

Where the challenges lie for policy makers, society, clinicians and scientists:

- gene tests
- genetic discrimination
- genetic enhancement
- privacy
- patents
- cloning and gene therapy

More about some of these topics in your Issues module

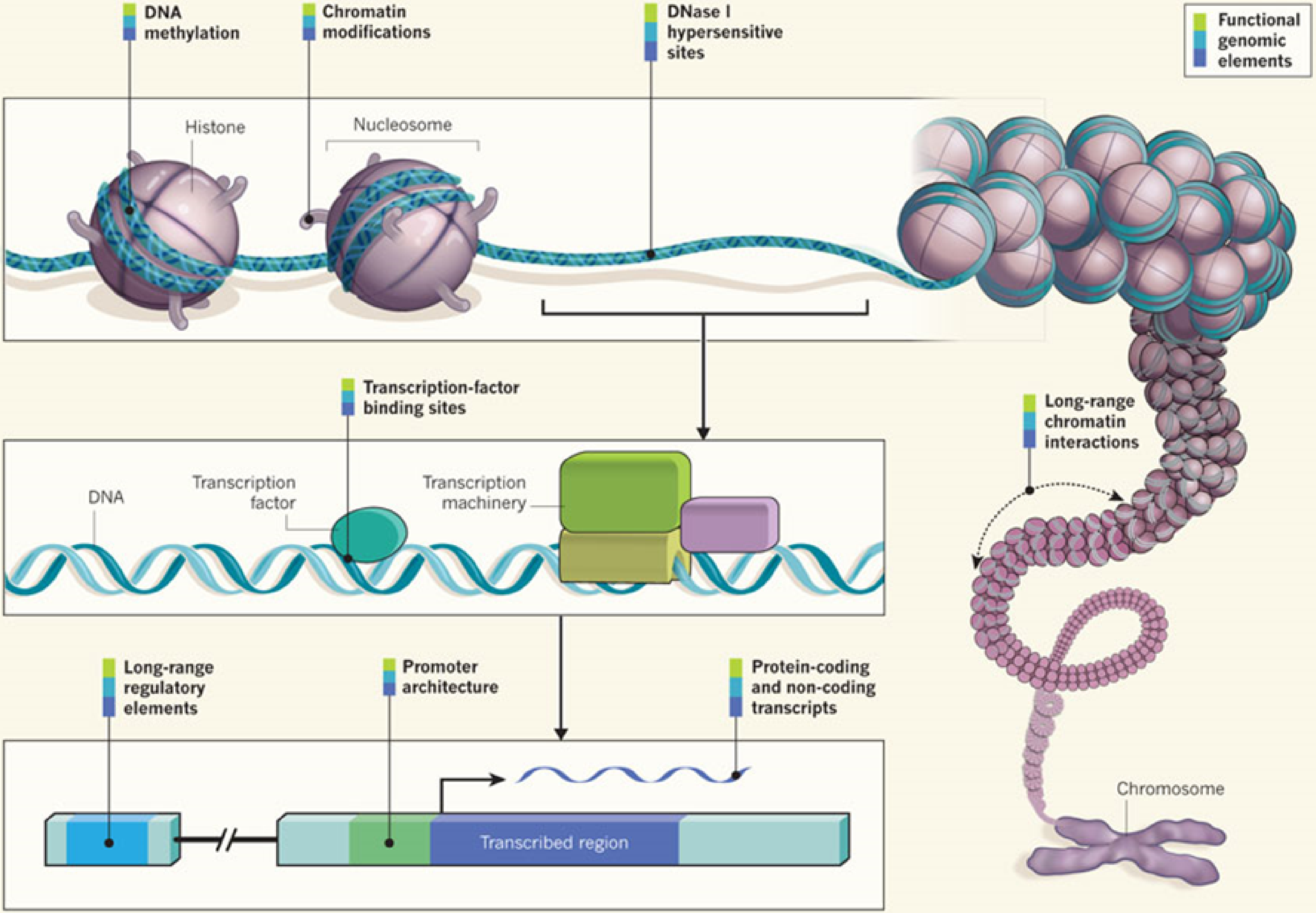
Sequencing was the easy part- it is working out what the sequence means that is the real challenge

- Functional analysis through identifying additional regions of sequence conservation
  - If DNA sequences are conserved in evolution this implies they have a function
  - However, some rapidly-evolving sequences may also be important
- Functional analysis using high-throughput methods (next slide)
  - Modifications to chromatin structure
  - Modifications to DNA bases
  - Binding sites for transcription factors

ENCODE
Encyclopedia of DNA Elements

- https://www.encodeproject.org/
- http://www.nature.com/encode/#/threads

- 1,640 genome-wide data sets prepared from 147 cell types
- Defined products or biochemical signatures


DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

100,000 genomes project

Genomics England: announced July 2013

http://www.genomicsengland.co.uk/

Up to £100 million of funding pledged by the government will:

- train a new generation of British genetic scientists to develop life-saving new drugs, treatments and scientific breakthroughs;
- train the wider healthcare community to use the technology;
- fund the initial DNA sequencing for cancer and rare and inherited diseases; and
- build the secure NHS data linkage to ensure that this new technology leads to better care for patients

Cancer (lung, paediatric, rare)

Rare diseases (diagnosis and improved testing)

Infectious diseases (sequencing pathogens: HIV, hepatitis C, TB)

100,000 genomes project

In addition to the strategic aims reports were also commissioned on two other aspects:

Ethics

- Aims
- Access
- Consent
- Public confidence
- Governance

Data

- Infrastructure
- Specification and standards
- Training and workforce

The details of these aspects are published at length here:

https://www.gov.uk/government/publications/mapping-100000-genomes-strategic-priorities-data-and-ethics