Next generation DNA sequencing: technology and applications

Robert Lyle
Department of Medical Genetics
Ullevål University Hospital

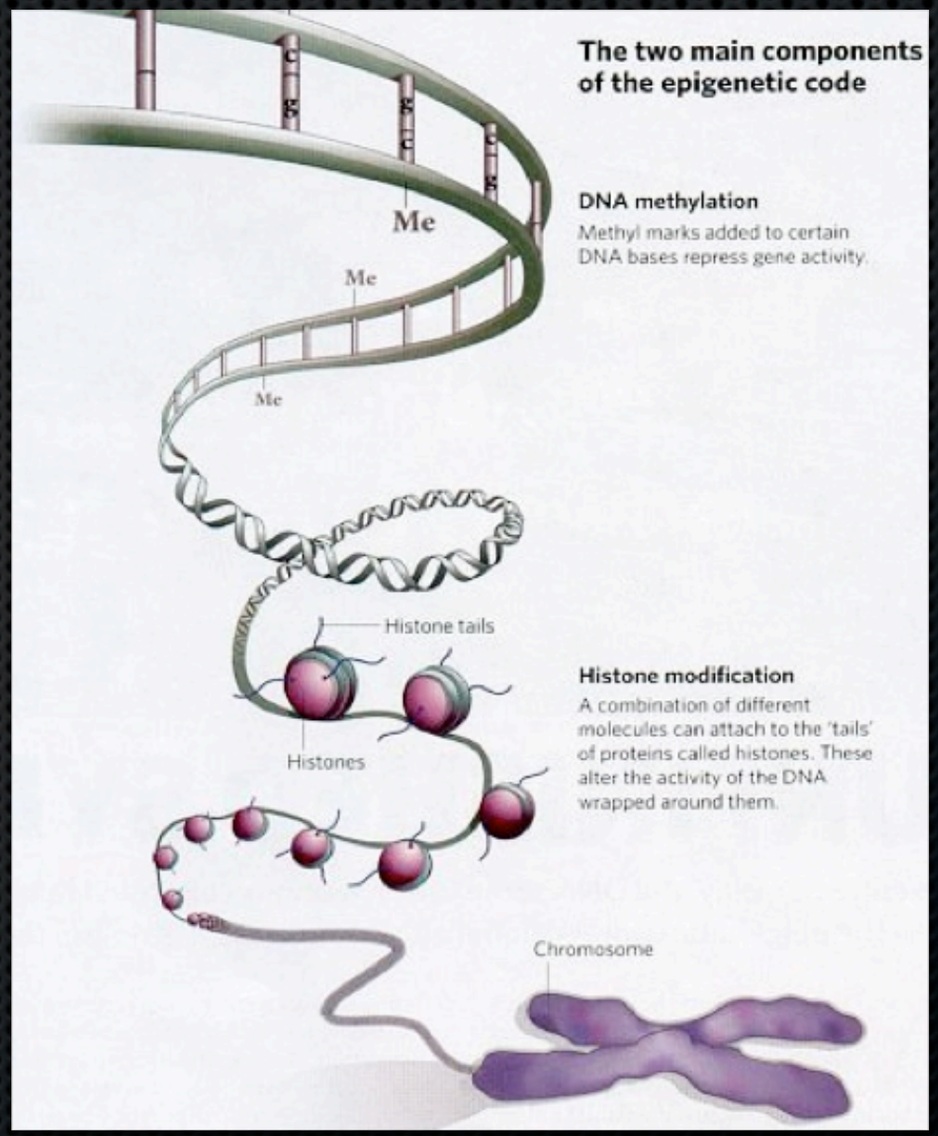
Robert.Lyle@medisin.uio.no

ULLEVÅL
University Hospital

- Project - epigenetic variation, twins and disease susceptibility
- Next generation sequencing at UUS

Epigenetics

- DNA methylation
  - CpG dinucleotides
- Histone modifications
  - acetylation
  - phosphorylation
  - methylation
  - ubiquitination

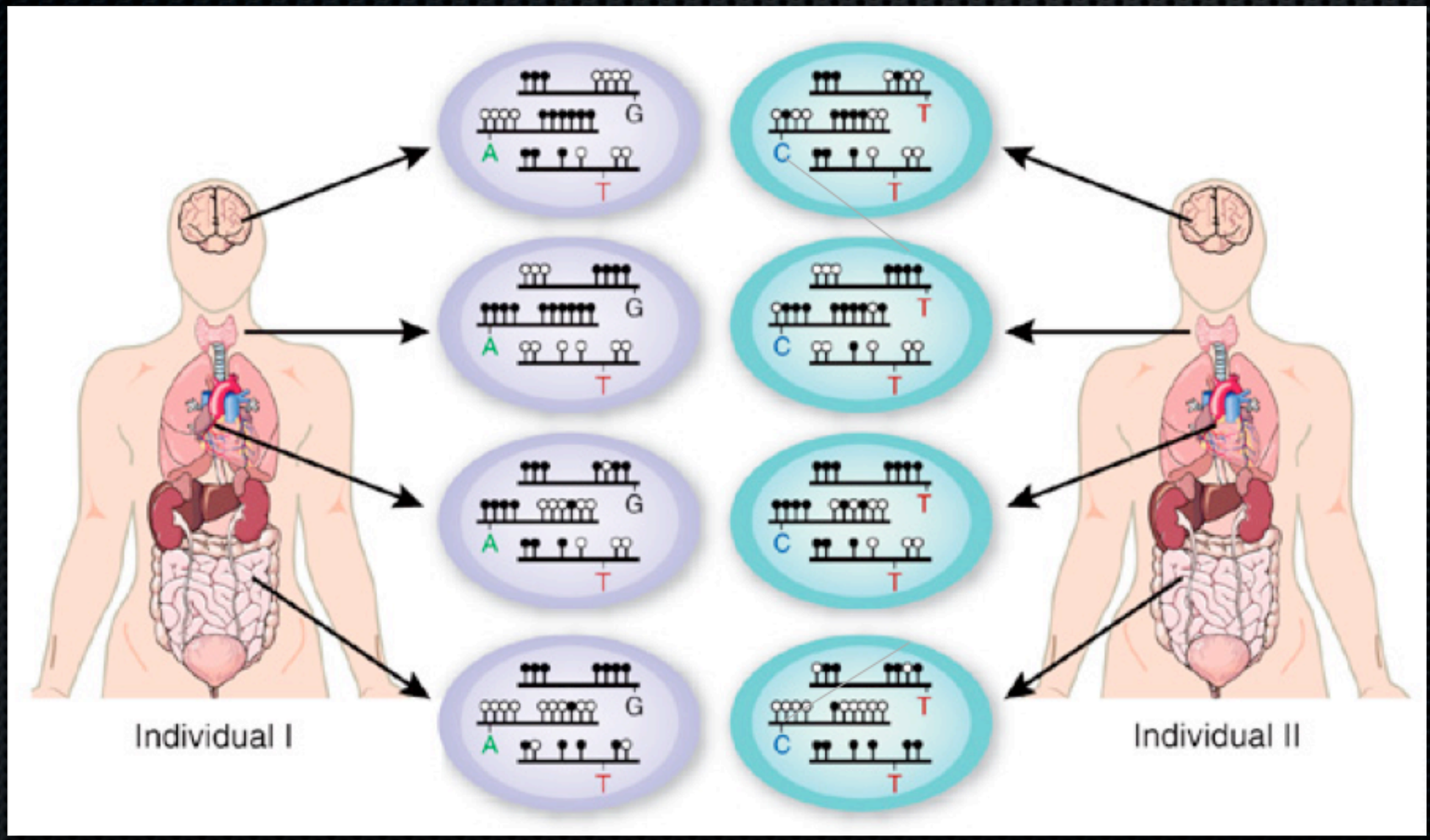


→ Control of gene expression

Broadly...

- DNA methylation
  - Long-term epigenetic silencing of specific sequences
  - transposons, imprinted genes, pluripotency genes
- Histone modifications
  - Short term, flexible epigenetic control

Epigenotypes



Potentially, we have as many epigenomes as we have cell types/devlopmental stages

Epigenetics in health and disease I

- Imprinting disorders
  - Prader-Willi/Angelman syndrome, Beckwith Wiedemann etc.
  - Uniparental disomy (UPD)
- Monogenic disorders
  - ICF syndrome - involves immunodeficiency
    - mutations in *DNMT3B* - DNA methyltransferase
  - Rett syndrome - mutations in *MeCP2*
- Cancer
  - complex DNA hypo- and/or hypermethylation

Epigenetics in health and disease II

- Environmental interactions
- Disease susceptibility
- Assisted reproductive technologies
- Cloning

Epigenetic variation and twins

Norwegain Research Council - FUGE II

Why are people different?

- Genetic
- Environmental
- ‘Stochastic’
- Epigenetic

Why are identical twins *not* identical?

- ~~Genetic?~~ Genomic sequence is the same*
- Environmental
- ‘Stochastic’
- Epigenetics

*ignoring somatic mutation: SNPs, CNVs

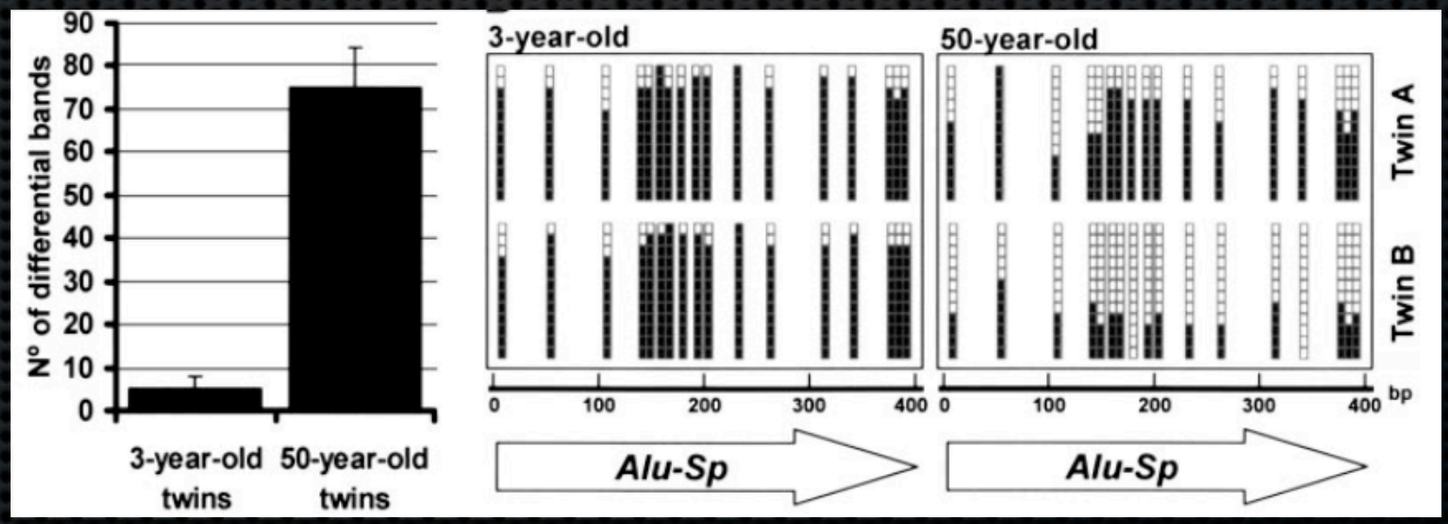
Are identical twins *(epi)genetically* identical?

- Genomic sequence is the same (except for mutations).

Epigenetic differences arise during the lifetime of monozygotic twins

Mario F. Fraga*, Esteban Ballestar*, Maria F. Paz*, Santiago Ropero*, Fernando Setien*, Maria L. Ballestar†, Damia Heine-Suñer‡, Juan C. Cigudosa§, Miguel Urioste¶, Javier Benitez¶, Manuel Boix-Chornet†, Abel Sanchez-Aguilera†, Charlotte Ling‖, Emma Carlsson‖, Pernille Poulsen**, Allan Vaag**, Zarko Stephan††, Tim D. Spector††, Yue-Zhong Wu‡‡, Christoph Plass‡‡, and Manel Esteller*§§

DNA methylation



Twins epigenetically ‘drifting apart’

Epigenotype changes over time

Project

- Discordance rates among MZ twins is ~50% for many immune-mediated diseases with large genetic component - psoriasis, asthma, IBD
- Understand the basis for this discordance rate
- Identify epigenetic differences between twins discordant for immune-mediated diseases

Strategy

- Collect monozygotic twins discordant for immune-mediated diseases
- Genome-wide epigenetic surveying (GWES)
  - DNA methylation - bisulphite sequencing
  - Histone status - chromatin immunoprecipitation
- High-throughput sequencing (>> 1 Gb per run)

Twin collection

- Protocols
- Ethics
- Consent

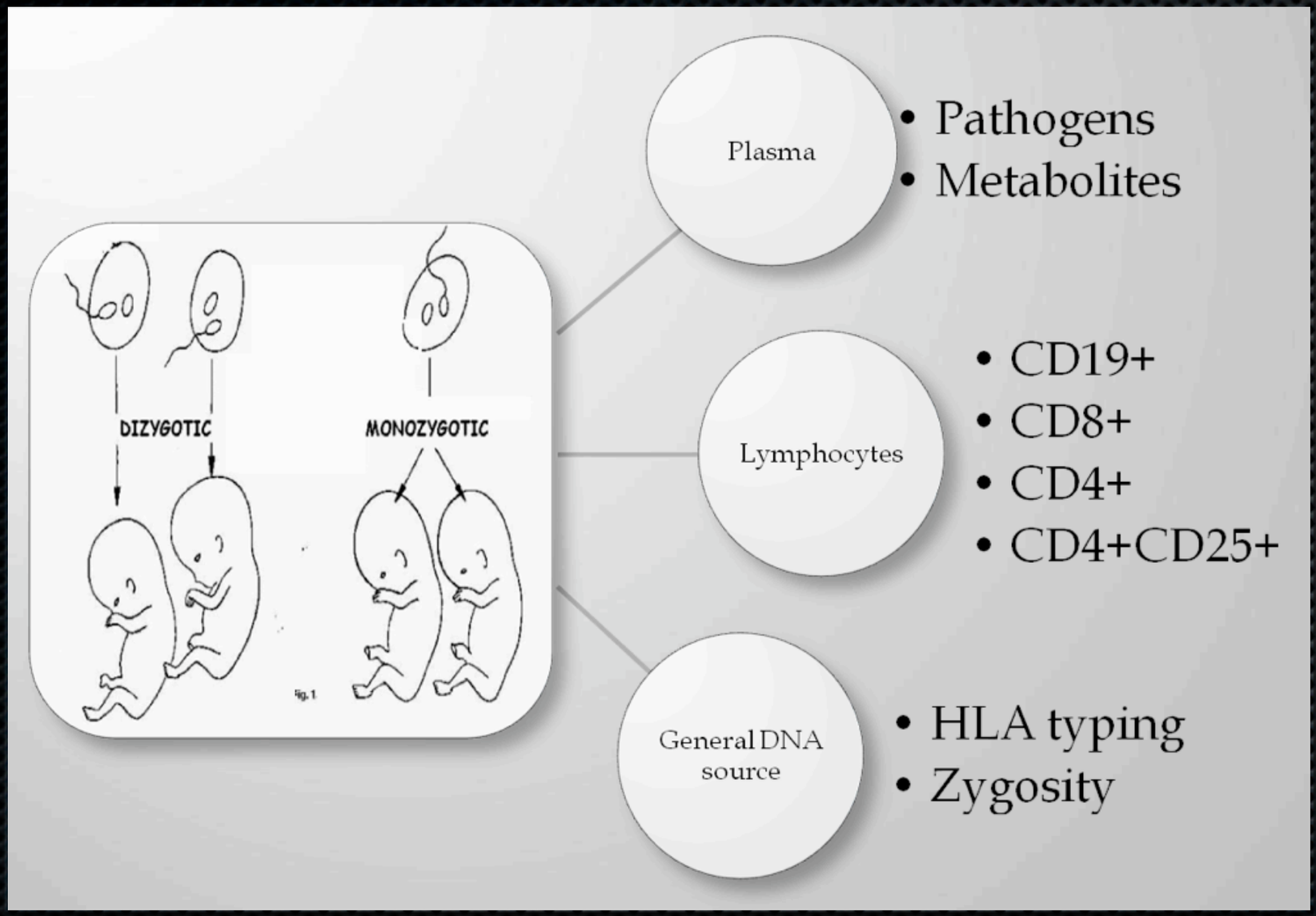
Twin group	Number, pairs	Status
Control (MZ/DZ)	lots	85 samples
Asthma	122*	consents-contact
Psoriasis	74*	consents-contact
IBD	15*	consents-contact

* discordant

Tvillingundersøkelsen ved Folkehelseinstituttet
Norwegian Institute of Public Health Twin Study

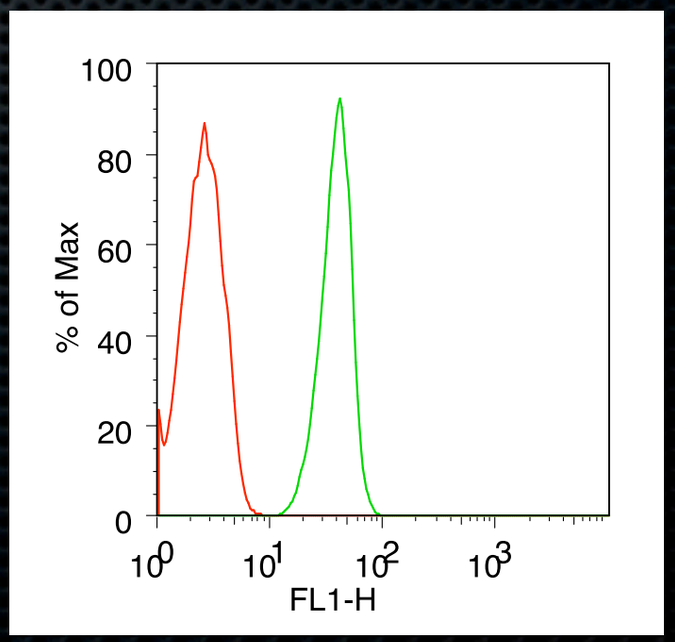
Norwegian Institute of Public Health
www.fhi.no

Sample collection



AutoMACS Pro

automated cell separation



→ cells, RNA, DNA

Processed >200 samples to date

Quantifying DNA methylation

Bisulphite sequencing (BiS)

genomic DNA

AGCTGT**m**CGATTAGCCG

methylated

AGTTGT**C**GATTAGTTG

1. bisulphite treat
2. PCR region of interest
3. sequence

AGCTGT**CG**ATTAGCCG

unmethylated

AGTTGT**T**GATTAGTTG

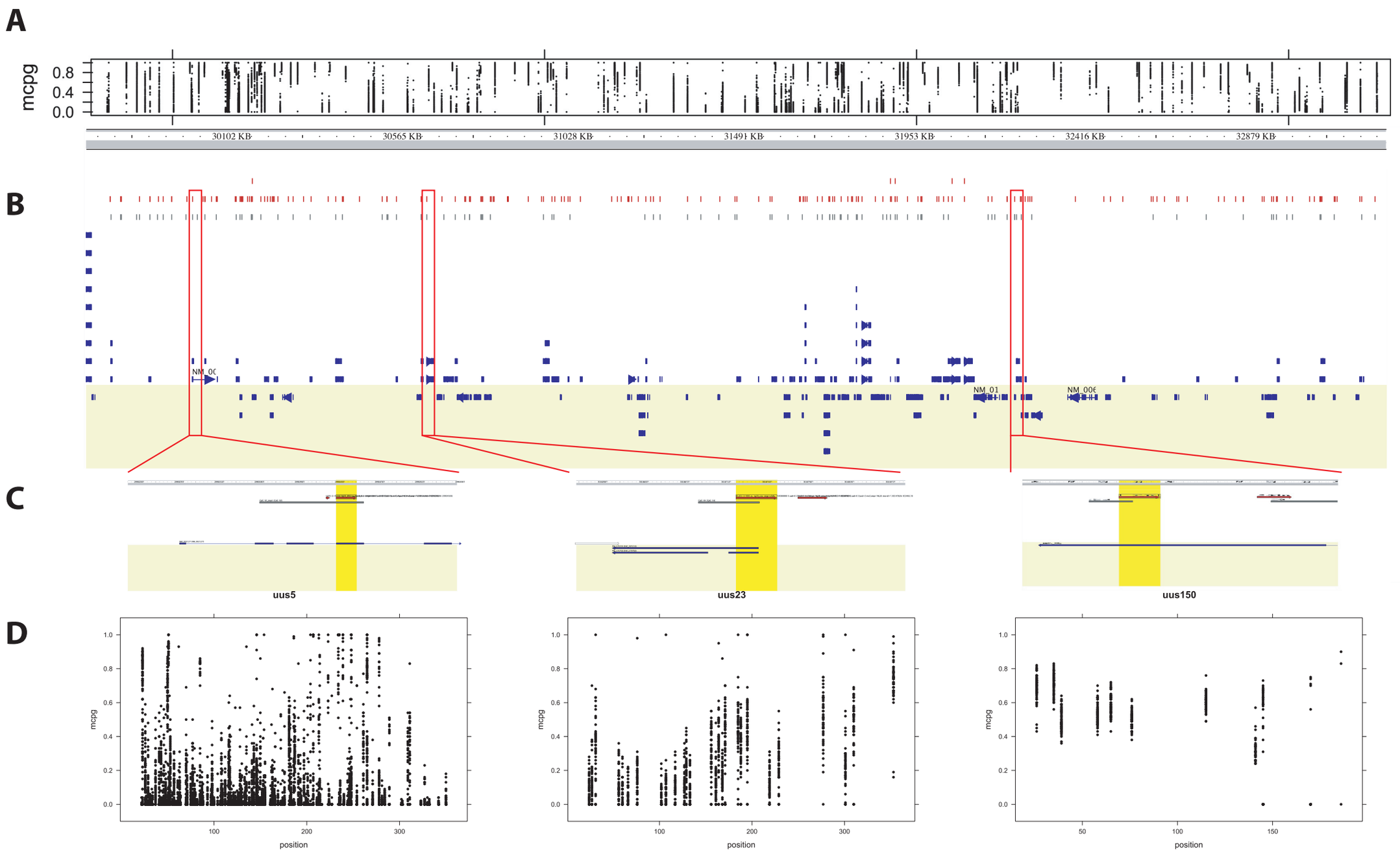
Generally low-throughput - single/several loci

High-throughput genome-wide?

DNA methylation variation

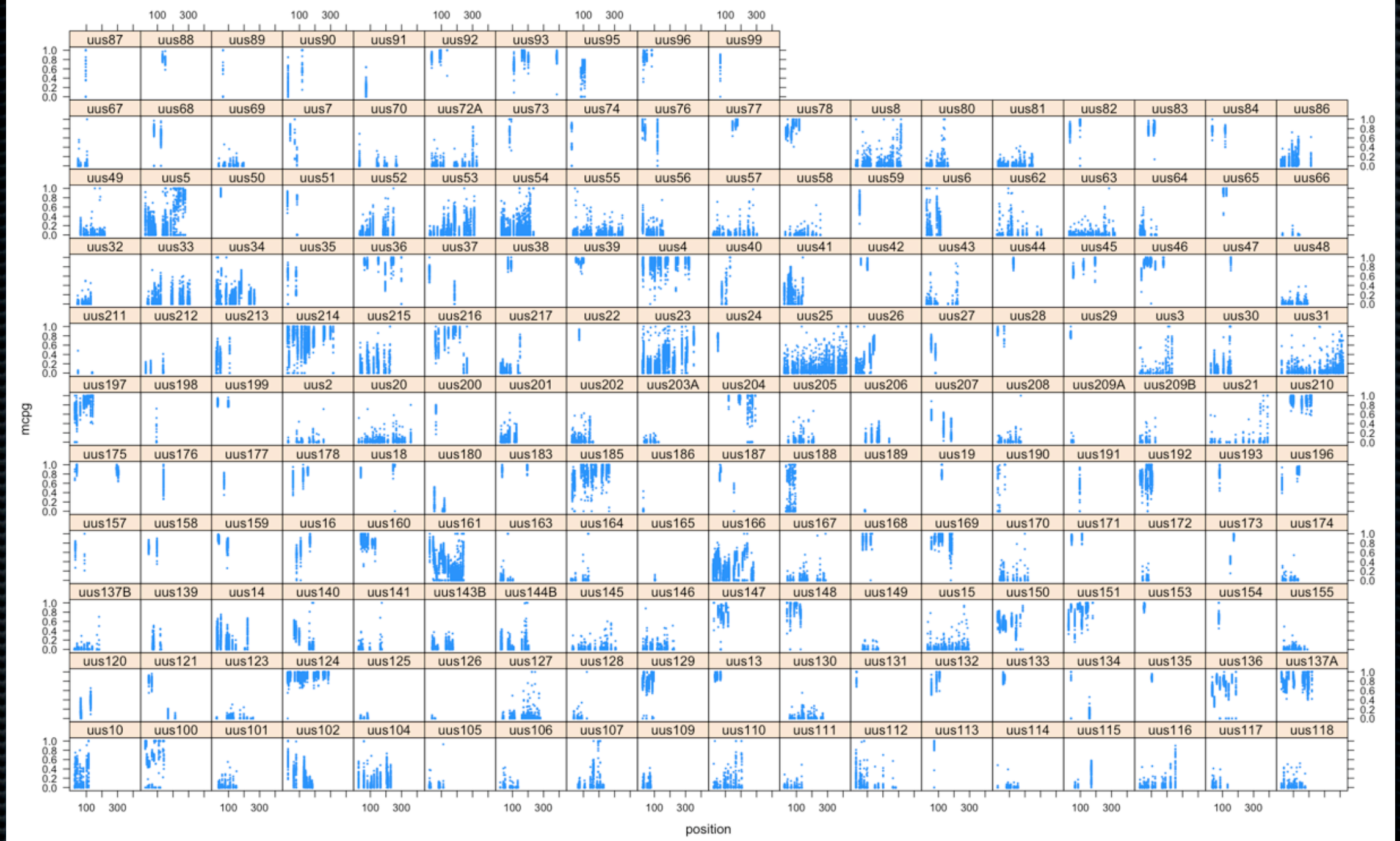
- Control MZ and DZ twins
- Regions within major histocompatability complex
- Identify variation
- How much variation under genetic control?

Patterns of DNA methylation in the MHC



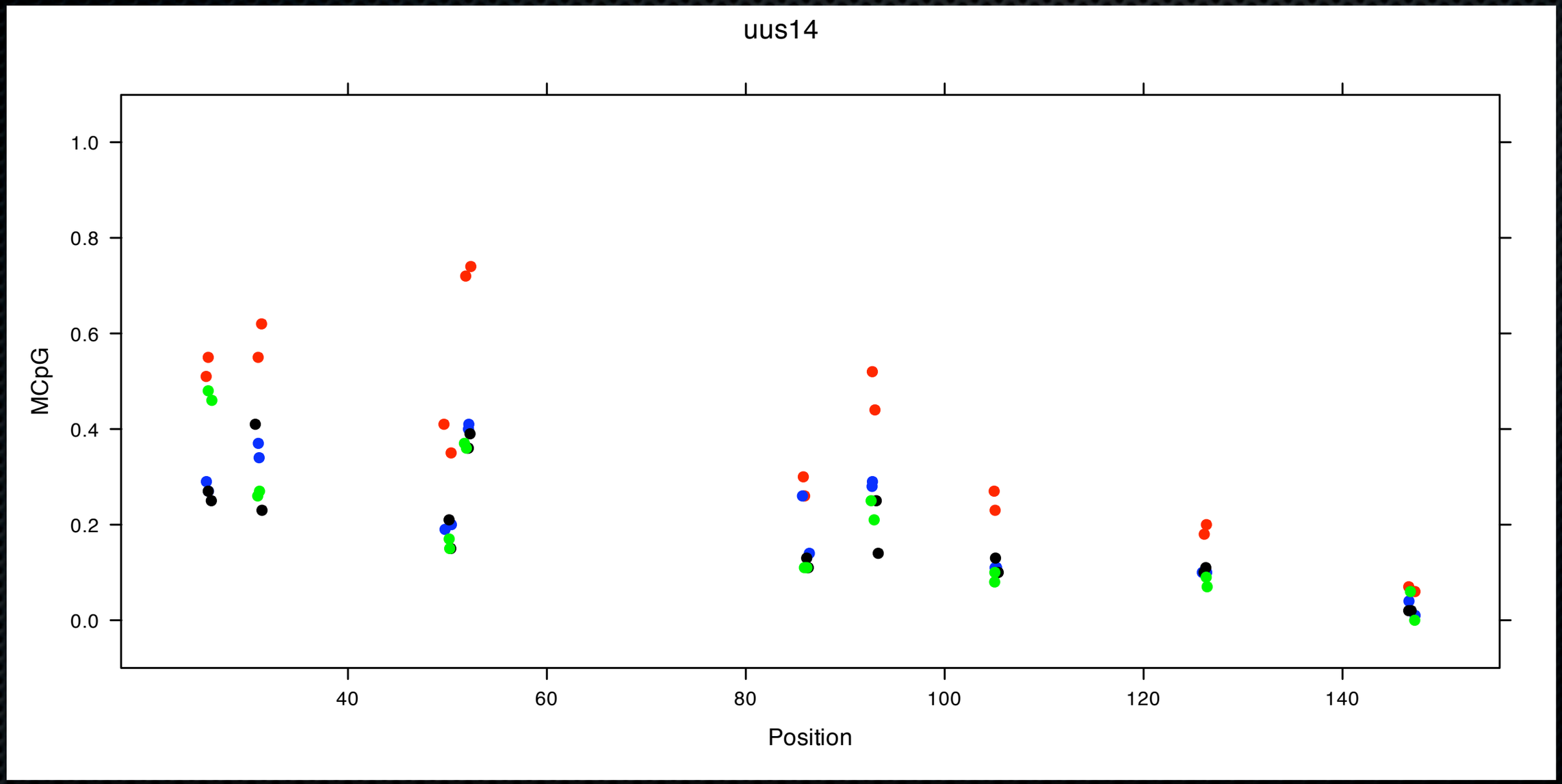
60 unrelated individuals, 190 regions, CD4+ cells

Overview of 190 MHC regions



60 unrelated individuals

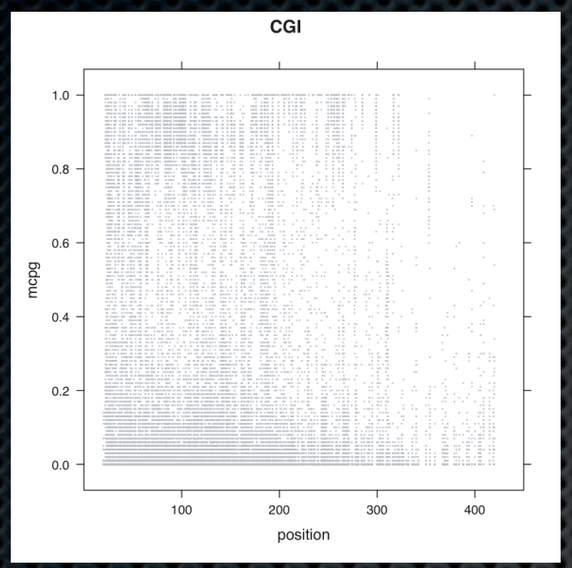
Variation in MZ twins



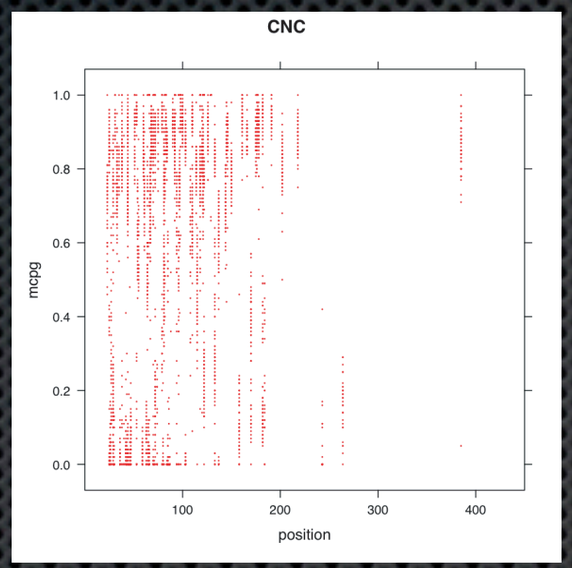
UUS14, 4 MZ pairs

Variation in region types

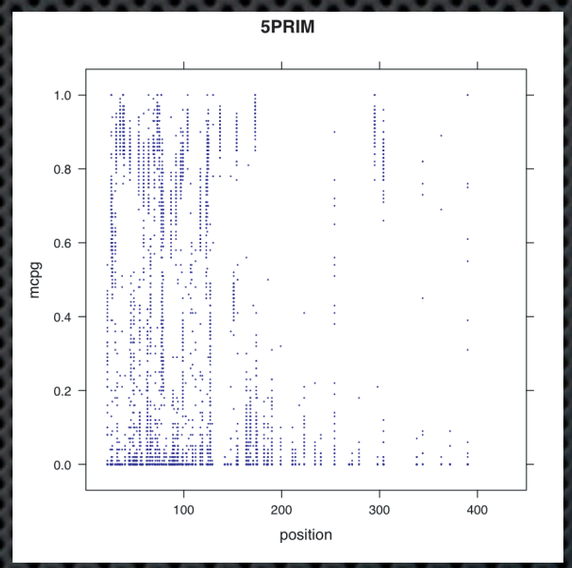
CpG islands



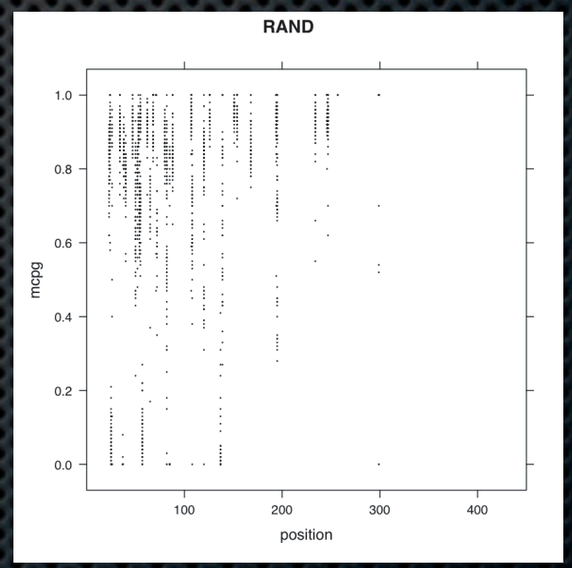
Conserved non-coding



5’ genes



Random



Different distribution of DNA methylation?

All differences are statistically significant

Summary

- Extensive variation in DNA methylation
- Can we do this genome-wide?

High-throughput sequencing

Sequencing: old and next

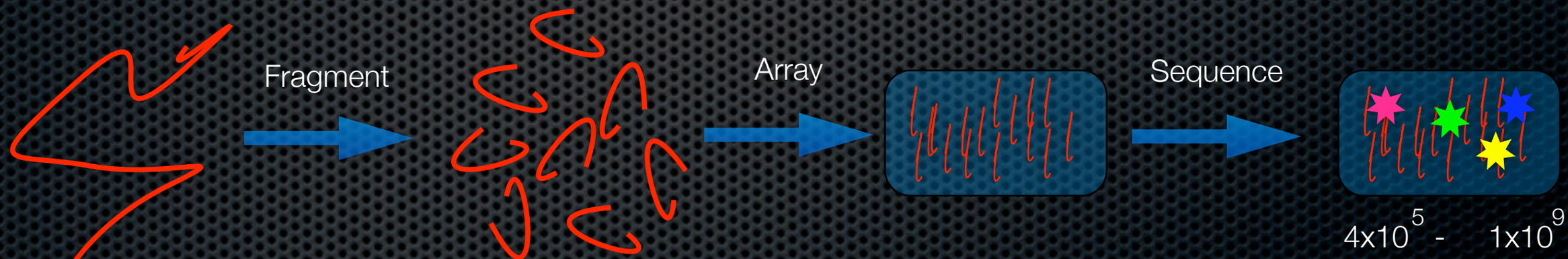
LTS

Molecules sequenced

1, 48, 96...

...unless you have a lot of machines

HTS



$4x10^5$ - $1x10^9$

...on one machine

Massively parallel

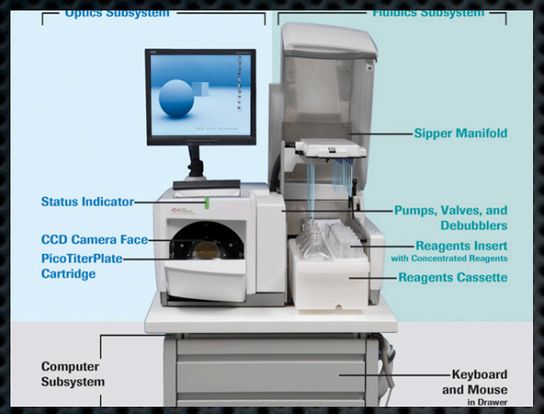
Sequencing throughput

Genome sequenced (publication year)	HGP (2003)	Venter (2007)	Watson (2008)
Time taken (start to finish)	13 years	4 years	4.5 months
Number of scientists listed as authors	> 2,800	31	27
Cost of sequencing (start to fi nish)	$2.7 billion	$100 million	< $1.5 million
Coverage	8–10 ×	7.5 ×	7.4 ×
Number of institutes involved	16	5	2
Number of countries involved	6	3	1

ABI/SOLiD - Yoruban (12x) - $60 000

HTS systems available

454



Roche

Solexa

Illumina

SOLiD

ABI

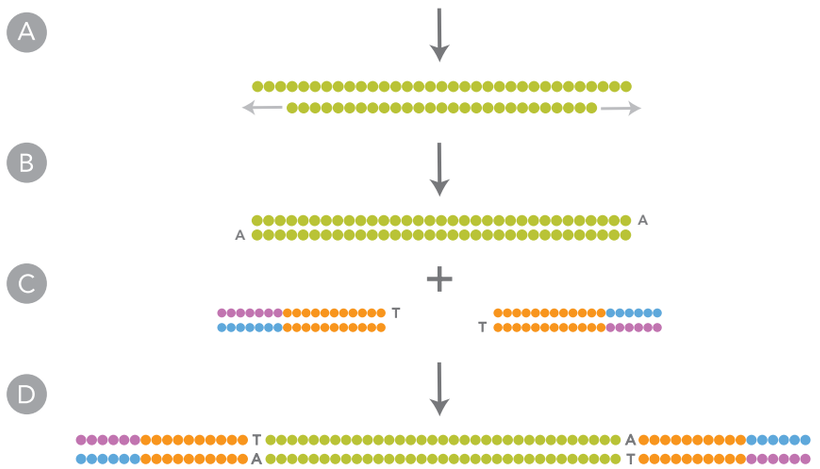
HeliScope

Helicos

Solexa (and Helicos)

1 Library Prep

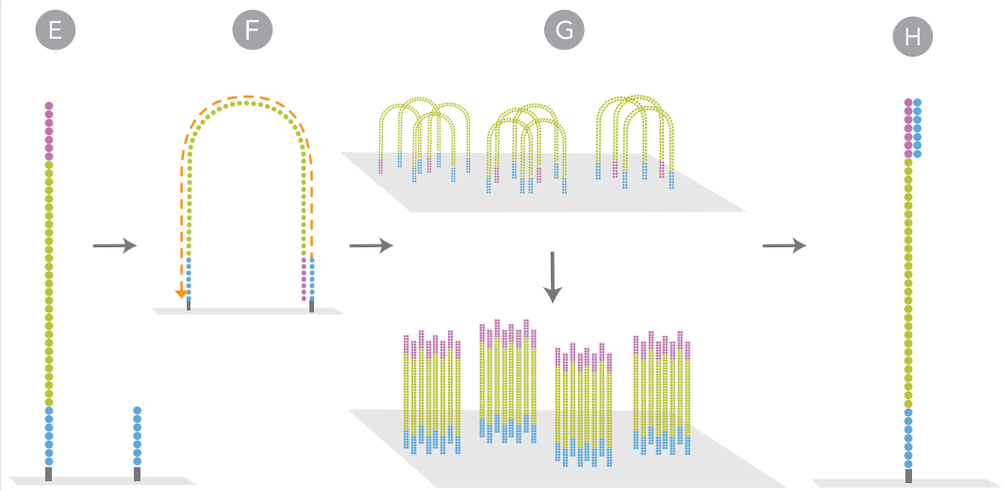
6 hours	3 hours hands-on time



- A Fragment DNA
- B Repair ends/Add A overhang
- C Ligate adapters
- D Select ligated DNA

2 Cluster Generation

5 hours	30 min. hands-on time (1–8 Samples)



- E Attach DNA to flow cell
- F Perform bridge amplification
- G Generate clusters
- H Anneal sequencing primer

454 (and SOLiD)

FIGURE 8

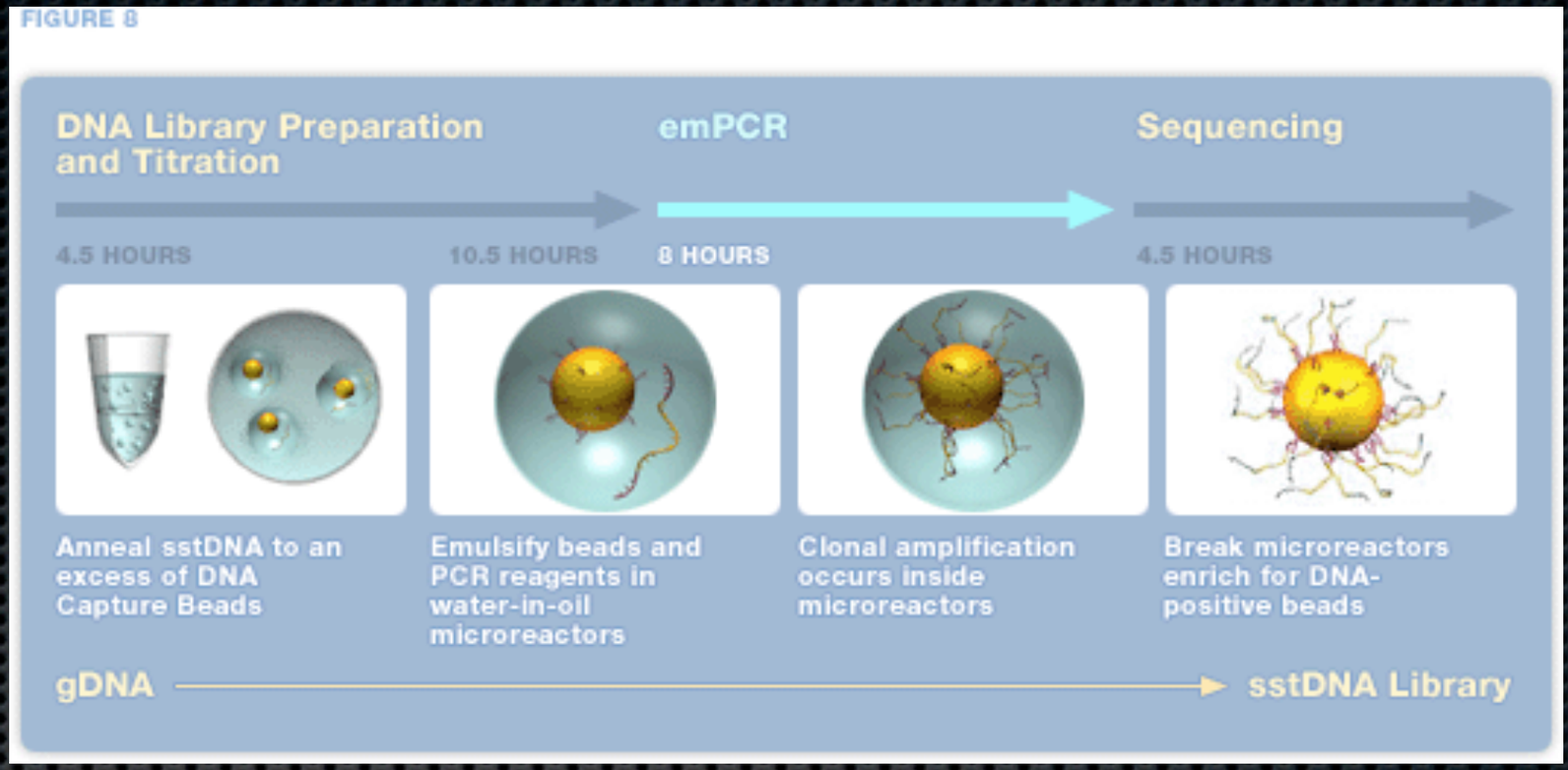
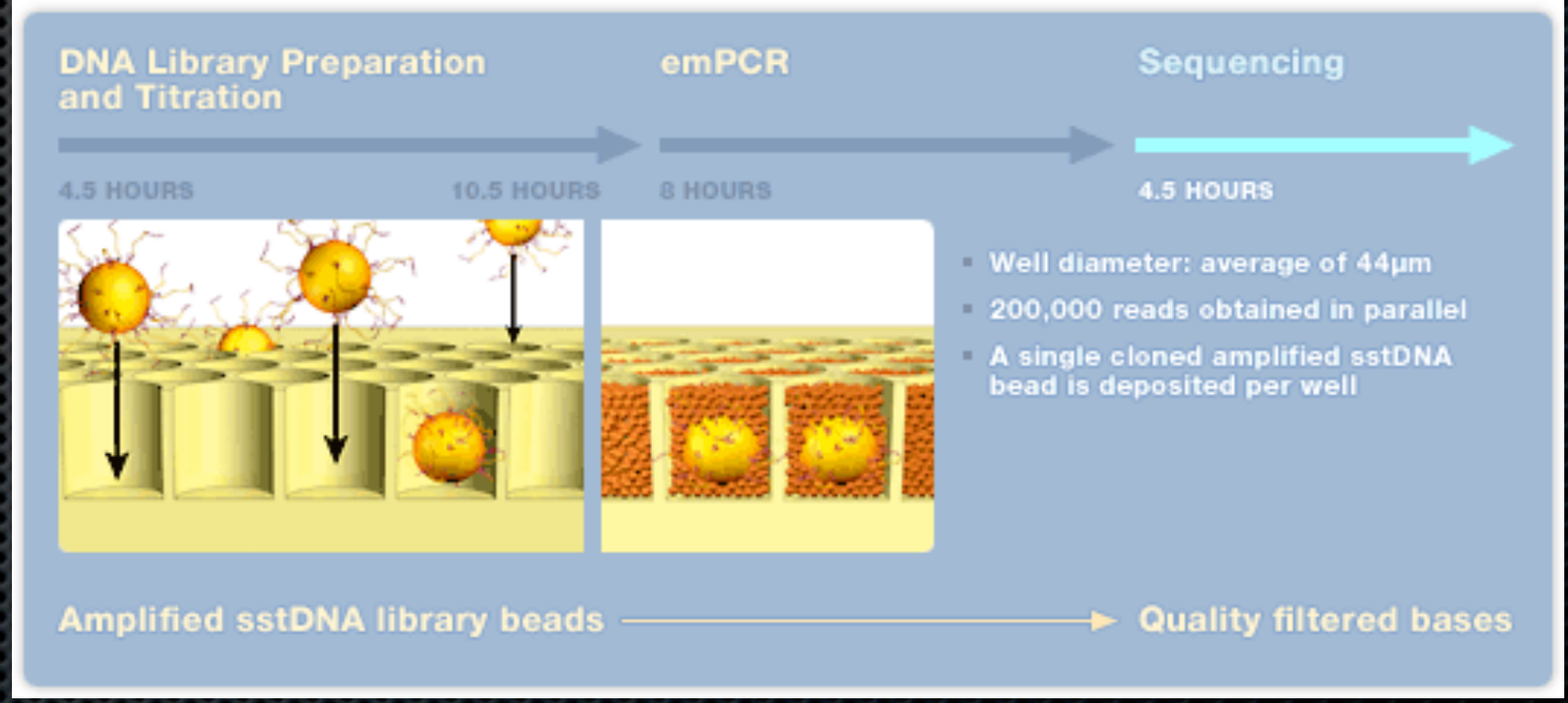


FIGURE 9



High-throughput sequencing: system comparison

COMPANY	ILLUMINA	ROCHE	APPLIED BIOSYSTEMS	HELICOS
Company	Illumina	Roche	ABI	Helicos Biosciences
Web	www.illumina.com	www.454.com	www.appliedbiosystems.com	www.helicosbio.com
System	Solexa	454	Solid	Helicos
	GenomeAnalyzer 1G	GenomeSequencer FLX	SOLiD Analyzer	HeliScope
Technology				
Sequencing method	Synthesis	Pyrosequencing	Ligation	Single molecule, synthesis
Sequence reactions	4E+07	4E+05	6E+08	1E+09
Read length, bp	50	200	35	25
Sequence per run, Mb	20000	100	10000	50000
System cost	3,321,725	4,999,000	4,501,661	9,375,000
Cost per run	24,577	62,459	53,141	119,250
Cost per Mb	1.23	625	5	2.39
Sequencing accuracy			0.9994	0.9900
Application features				
Read length	3	5	2	1
Sample prep	3	1	1	5
Sample throughput	3	1	3	5
Sequencing accuracy	3	4	5	1
DNA methylation	4	0	3	1
Cost per run	5	3	3	1
Cost per Mb	5	1	3	4
System cost	5	3	4	1
Support cost	5	4	2	1
3 year cost	5	3	3	1
Total System Rating for UUS Projects	**41**	**25**	**29**	**21**

Illumina - 20 Gb single run (~6x human genome)

Applications

Application	Project
Resequencing	whole genome linkage/association mutation detection
de novo sequencing	metagenomics new species
Expression	transcriptome SAGE miRNA
Epigenetics	DNA methylation ChIP
Variation	SNPs CNVs

Financing

Year	Source
2008	UUS UiO
2009	UiO

Important issues....

- EU tender process complete **Illumina GA II**
- Data storage (> 1 Tb per run)
- Bioinformatics
- Core facility
- Link to 454 at CEES/UiO?

People

Robert Lyle	Medical Genetics, UUS	Principal investigator
Dag Undlien	Medical Genetics, UUS/UiO	Principal investigator
Jennifer Harris	Epidemiology, NIPH/NIH	Investigator
Gregor Gilfillan	Medical Genetics, UUS	Post-doc
Kristina Gervin	Medical Genetics, UUS	PhD student
Heidi Nygård	Medical Genetics, UUS	Nurse/field worker
Ingunn Brandt	Epidemiology, NIPH	Twins DB
Hanne Akselsen	Medical Genetics, UUS	Technician
Martin Hamerø	Medical Genetics, UUS	Technician
Rune Moe	Medical Genetics, UUS	Technician
Anne Olaug Olsen	Dermatology, UUS	Clinician psoriasis
Monica Cheng Munthe-Kaas	Pediatrics/Medical Genetics, UUS	Clinician asthma
Torbjørn Rognes	Institute of Bioinformatics, UiO	Bioinformatics
Sigve Nakken	CMBN, UiO	PhD student bioinformatics
Hans-Christian Åsheim	Medical Genetics, UUS	Immunology
Thore Egeland	Medical Genetics, UUS	Statistics