INTRODUCTION

CHIP-SEQ AND ATAC-SEQ TRAINING, OCTOBER 2019

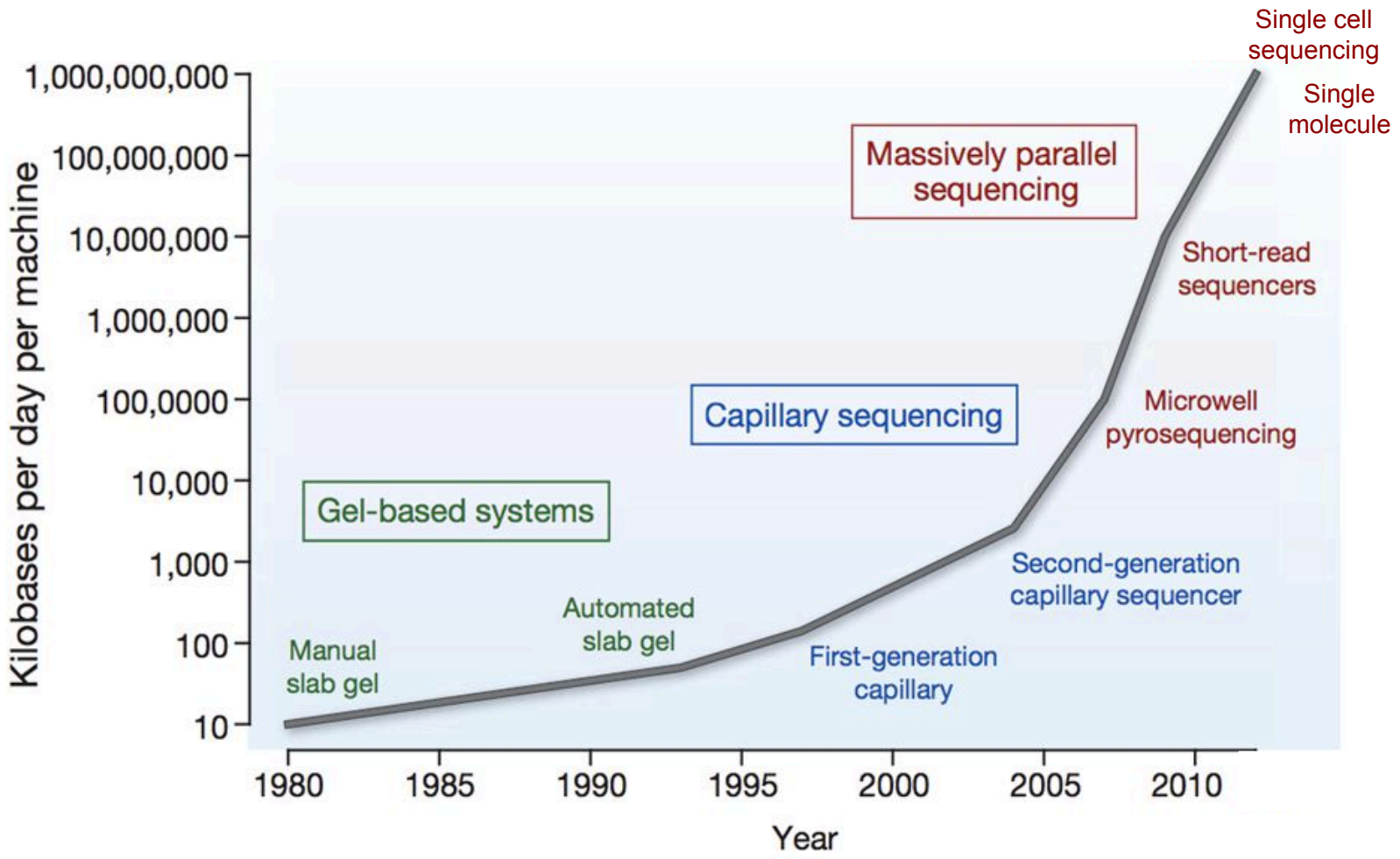
Sandra Cortijo, SLCU (sandra.cortijo@slcu.cam.ac.uk) Sergio Martinez Cuesta (Sergio.MartinezCuesta@cruk.cam.ac.uk)
Sankari Nagarajan (Sankari.Nagarajan@cruk.cam.ac.uk) Ashley Sawle (Ashley.Sawle@cruk.cam.ac.uk)

Timetable

Day 1	Topics
09:30 - 10:30	Lecture: Introduction to ChIP-seq, data analysis and QC metrics
10:30 - 11:00	Practical: Introduction to ChIP-seq, data analysis and QC metrics
11:00 - 11:15	Tea/coffee break
11:15 - 12:30	Practical: Introduction to ChIP-seq, data analysis and QC metrics (continued)
12:30 - 13:30	Lunch (not provided)
13:30 - 15:15	Lecture/practical: ChIP-seq data analysis part 1
15:15 - 15:30	Tea/coffee break
15:30 - 17:30	Lecture/practical: ChIP-seq data analysis part 2
Day 2	
09:30 - 10:15	Lecture: Recap and biological replicates
10:15 - 10:30	Tea/coffee break
10:30 - 11:15	Practicals: ChIP-seq part 3 (cont.) & working with biological replicates
11:15 - 12:00	Lecture: Differential binding analysis
12:00 - 13:00	Practical: Differential binding analysis
13:00 - 14:00	Lunch (not provided)
14:00 - 14:45	Lecture: Introduction to ATAC-seq data analysis
14:45 - 15:00	Tea/coffee break
15:00 - 17:00	Practical: ATAC-seq data analysis

A BRIEF HISTORY OF SEQUENCING

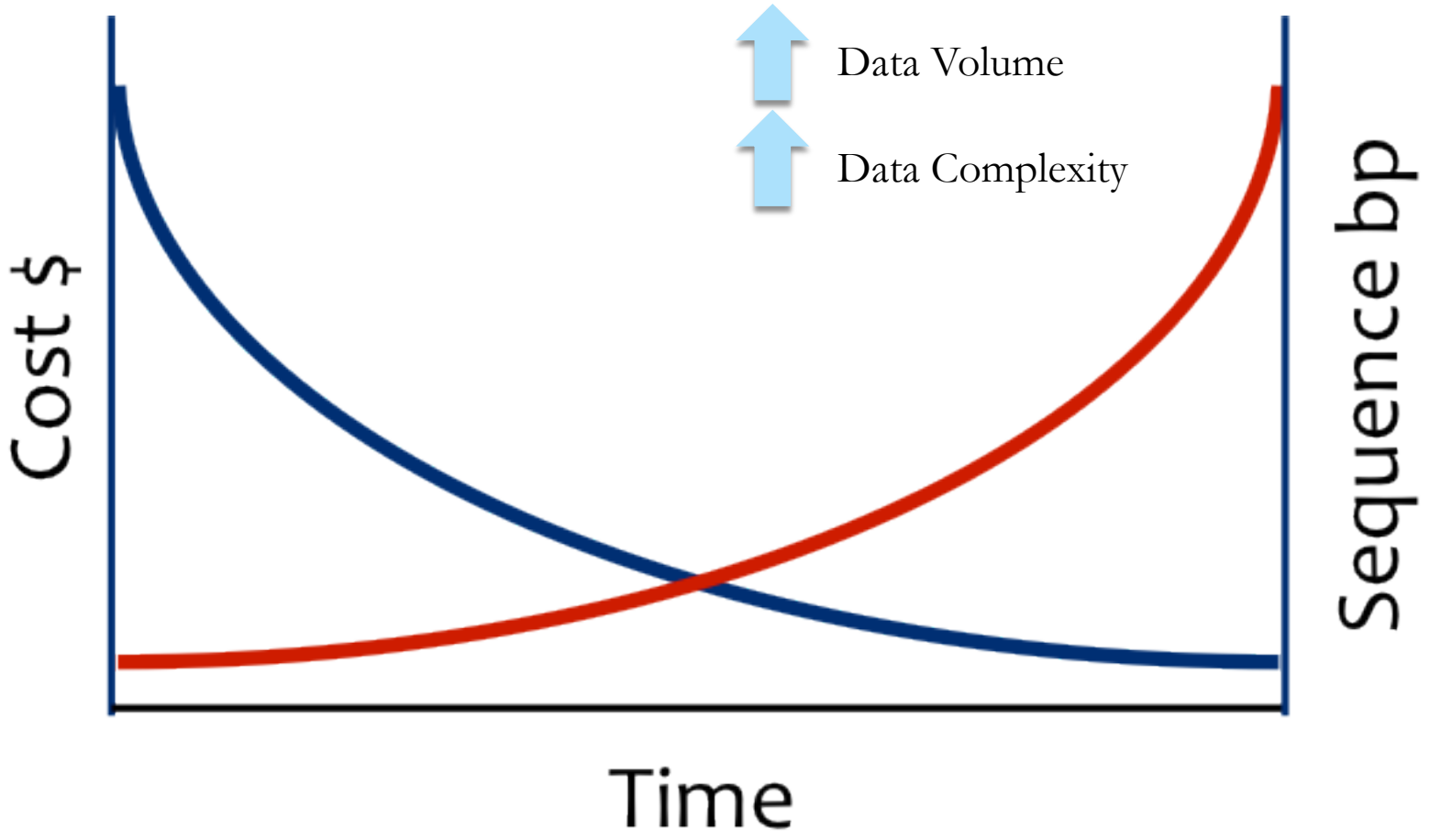
High Throughput sequencing (HTP)



HTS TECHNOLOGIES

Sequencing by synthesis

https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8


Data Volume
Data Complexity
Cost $
Sequence bp
Time

DATA VOLUME

- Scientists are expecting as many as 1 billion people to have their genomes sequenced by 2025
- The amount of data being produced daily is doubling every seven months
- Primary data measured in terabytes
  - Includes raw data from sequence machine
  - Not usually required
- Sequence data
  - FASTQ files 10-100GB
  - Analysis data are smaller (10%)
- Minimum that needs to be kept
- Movement of data (once only)

DATA COMPLEXITY

- Multiple samples
- Multiple runs
- Multiple platforms
- Sample details not evident from data
  - 120424_H183_0157_AC0KP3ACXX@HWI-HI83:157:C0KP3ACXX:6:1101:1210:1974 1:N:0:AGTCAA
- Metadata is important
- Describes the data

HTS APPLICATIONS

What do we study?

GENOME

EPIGENOME

TRANSCRIPTOME

EPIGENOME

TRANSCRIPTOME

Table 1 Applications of next-generation DNA sequencing

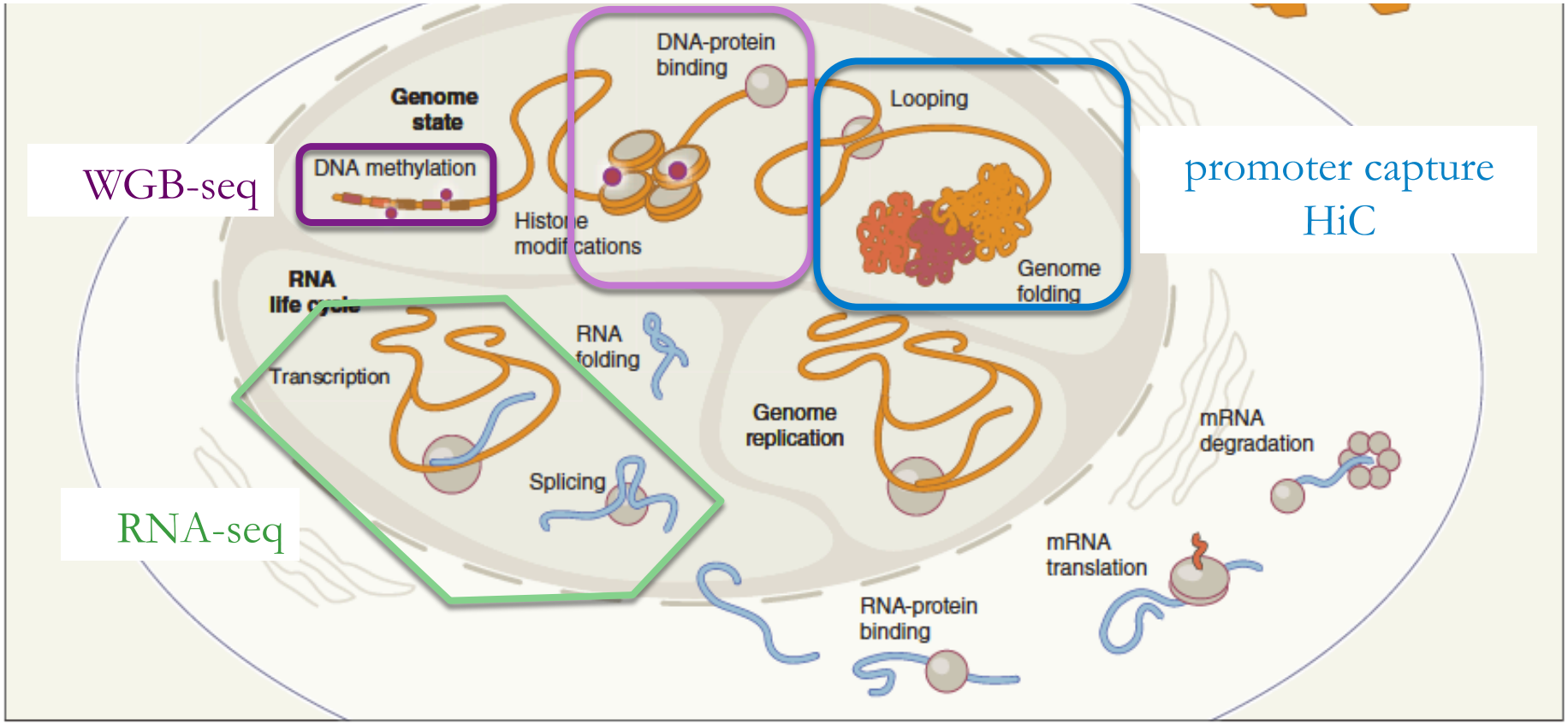
Method	Sequencing to determine:	Example reference	'Subway' route as defined in Figure 3
DNA-Seq	A genome sequence	57	Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	20	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	34	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	129	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	113	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome positioning)	130	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	131	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	46	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	39	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	119	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	41	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	71	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing
Chia-PET	Long-range interactions mediated by a protein	73	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	48	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	132	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	42	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	93	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	86	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	95	Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	92	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

FAIRE-seq, formaldehyde-assisted isolation of regulatory elements–sequencing. MAINE-Seq, MNase-assisted isolation of nucleosomes-sequencing; RIP-Seq, RNA-binding protein immunoprecipitation-sequencing; CLIP-Seq, cross-linking immunoprecipitation-sequencing; HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation; FRT-Seq, on-flowcell reverse transcription–sequencing. NET-Seq, native elongating transcript sequencing. TRAP, translating ribosome affinity purification. PhIT-Seq, phenotypic interrogation via tag sequencing.

Shendure J. et al., Nat Biotechnol. 2012 Nov;30(11):1084-94

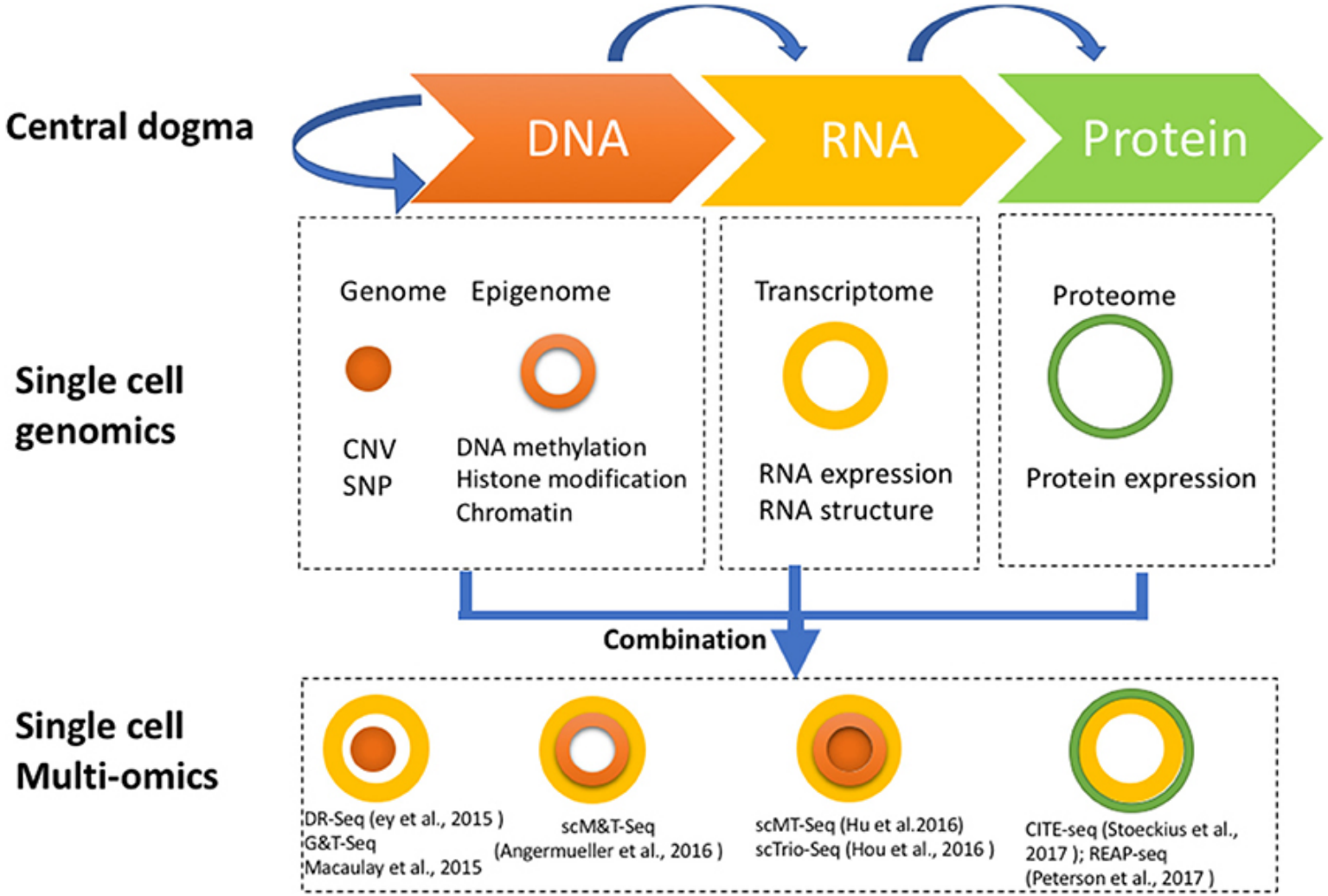
Epigenomic and Transcriptomic Assays

ChIP-seq & ATAC-seq



Shendure J. et al., Nat Biotechnol. 2012 Nov;30(11):1084-94

Single-cell multi-omics



Hu et al., Front. Cell Dev. Biol. 2018 Apr 20;6:28

HTS experimental design

Coverage

Batch effect

COVERAGE

Coverage: the number of reads representing a given nucleotide in the reconstructed sequence

Formula: N*L/G, where

G = length of haploid genome

N = the number of reads

L = average read length

e.g human genome = 3,000,000,000 (3Gb)

coverage = (100 bp)*(189×10^6)/(3×10^9 bp) = 6.3

each base in the genome will be sequenced between 6 and seven times on average

OF NOTE

- coverage is non-uniform across the genome
- read length, region mappability and GC content influence the coverage
- "callability" ≈ accessible portion of the genome
- "generating 50x mapped coverage (60x before read mapping/filtering are applied) renders ~95% of the genome and ~81% of the exome callable"

A Guide for Deep Sequencing of Human Genomes

Subramanian et al. 2011

COVERAGE

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing or Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
DNA Target-Based Sequencing	ChIP-Seq	10-14M (sharp peaks); 20-40M (broad marks)	Rozowsky et al., 2009; ENCODE 2011 Genome; Landt et al., 2012
	Hi-C	100M	Belton, J.M et al., 2012
	DNAse 1-Seq	25-55M	Landt et al., 2012
DNA Methylation Sequencing	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	Bisulfite-Seq	5-15X; 30X	Ziller, M.J et al., 2015; Epigenomics Road Map

From: genohub.com recommemded coverage for mammalian genome (3Gb)

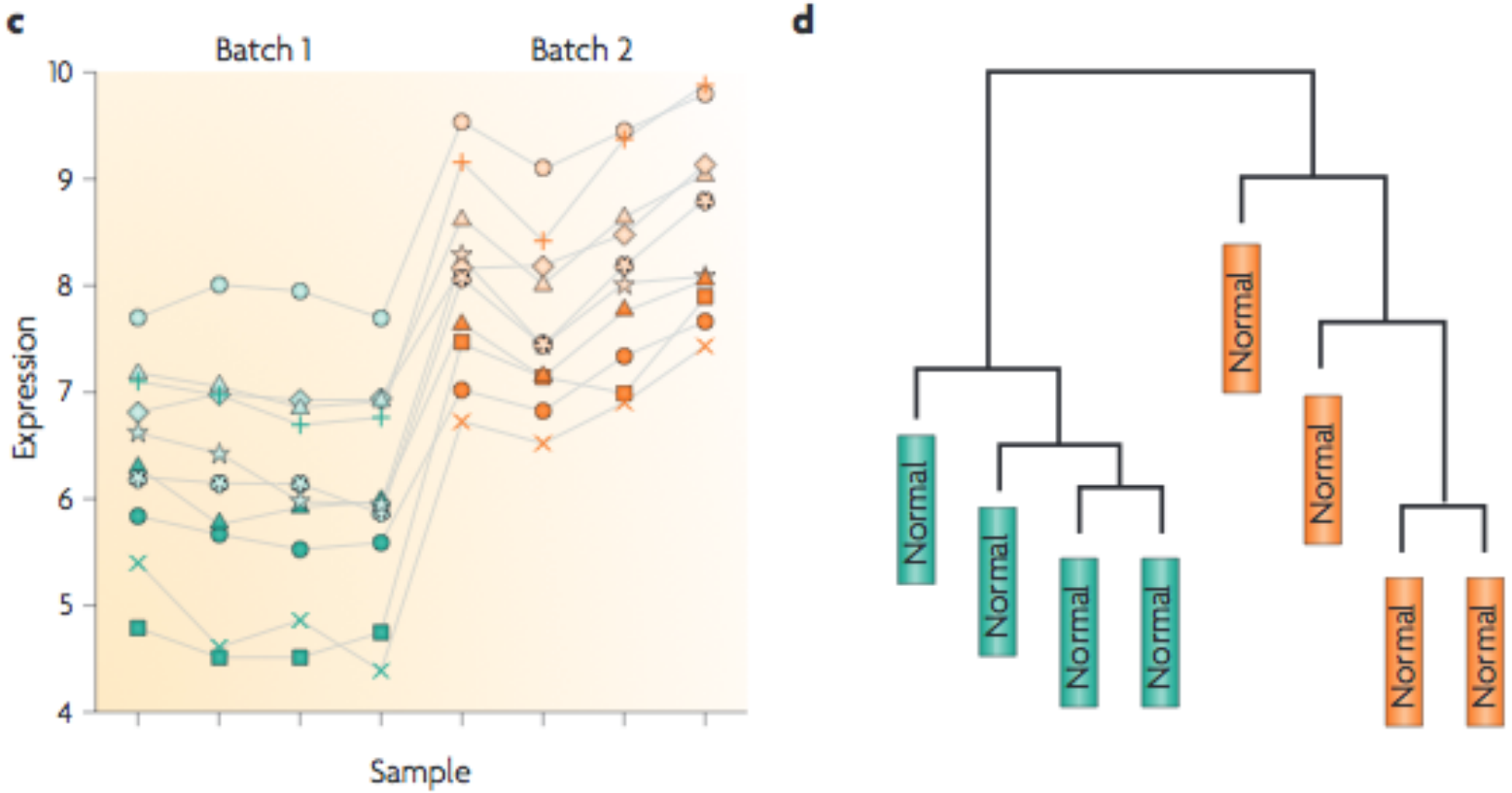
ATAC-seq: More than 50 million reads for human samples

From: https://informatics.fas.harvard.edu/atac-seq-guidelines.html

WHEN TO SEQUENCE MORE

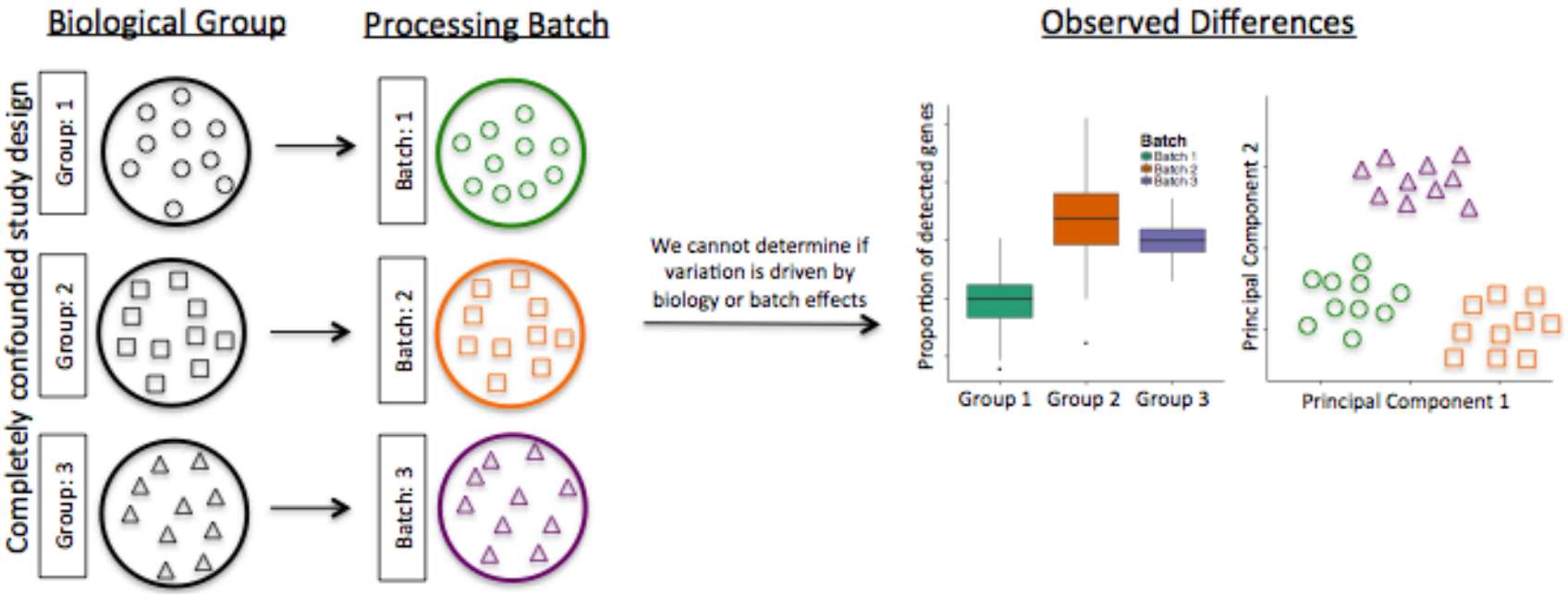
- Researching rare events
  - low binding activities in ChIP-seq
- Certain genomes need more sequencing
  - certain regions may be hard to sequence

BATCH EFFECT



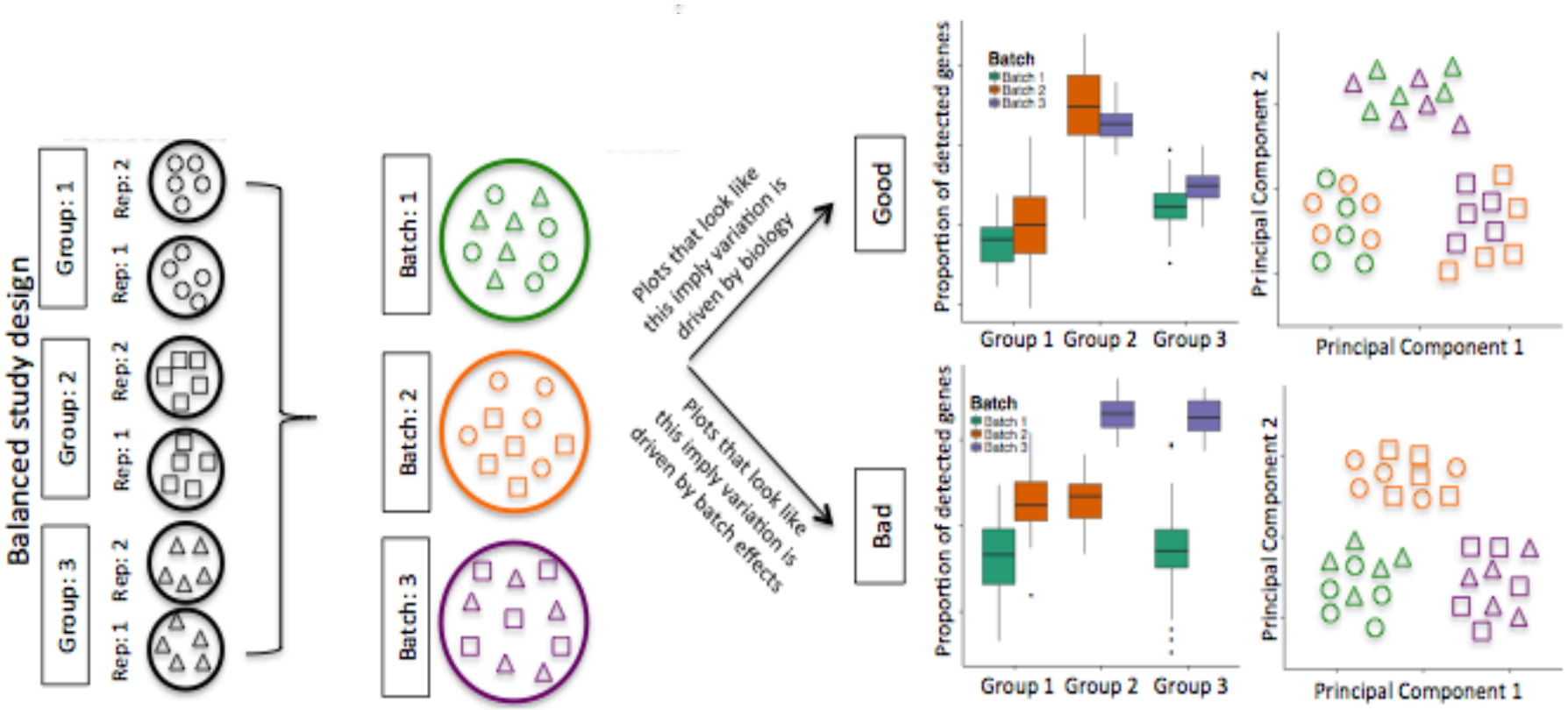
Leek et al. 2010

CONFOUNDING BIOLOGICAL VARIATION AND BATCH EFFECT



Hicks et al. 2015

CONFOUNDING BIOLOGICAL VARIATION AND BATCH EFFECT



Hicks et al. 2015