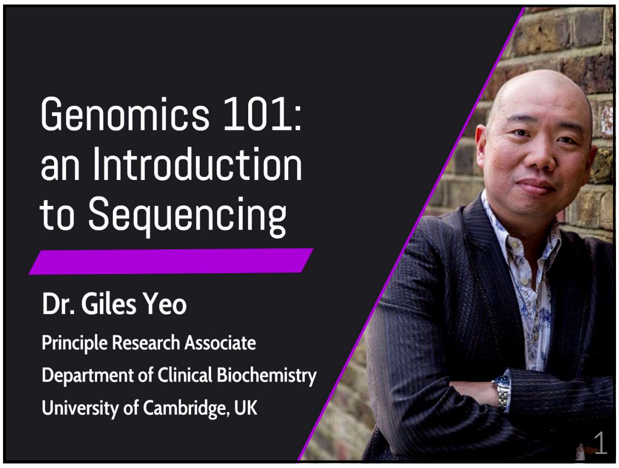
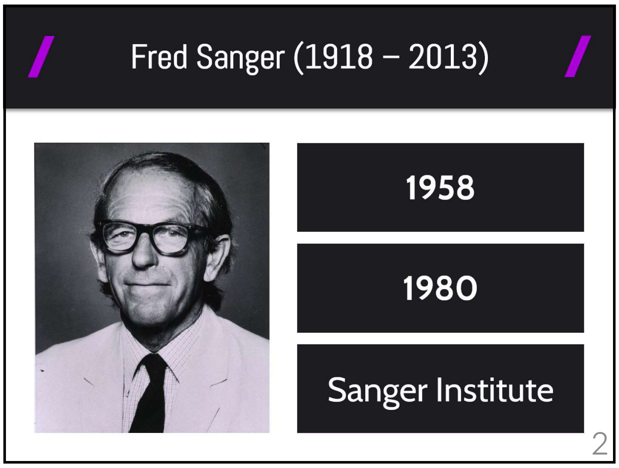


Genomics 101: an Introduction to Sequencing

Dr. Giles Yeo

Principle Research Associate
Department of Clinical Biochemistry
University of Cambridge, UK

1



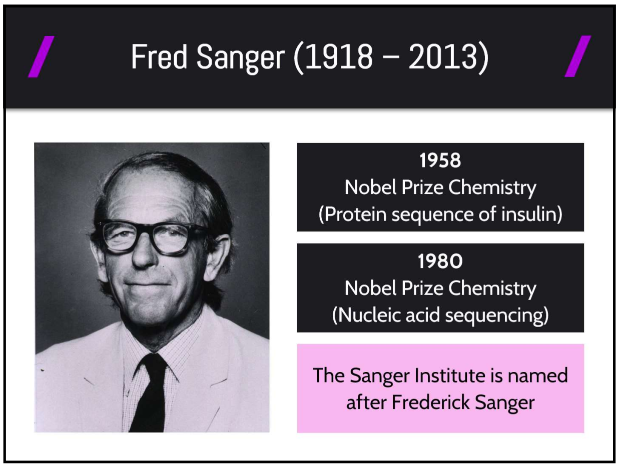
Fred Sanger (1918 – 2013)

1958

1980

Sanger Institute

2



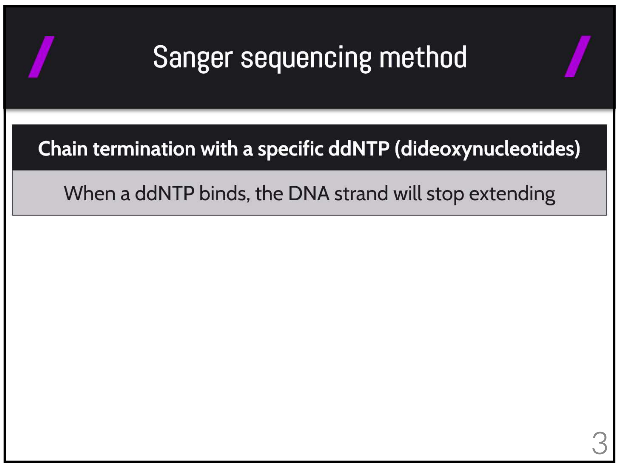
Fred Sanger (1918 – 2013)

1958
Nobel Prize Chemistry
(Protein sequence of insulin)

1980
Nobel Prize Chemistry
(Nucleic acid sequencing)

The Sanger Institute is named after Frederick Sanger

Dr. Giles Yeo – University of Cambridge, UK

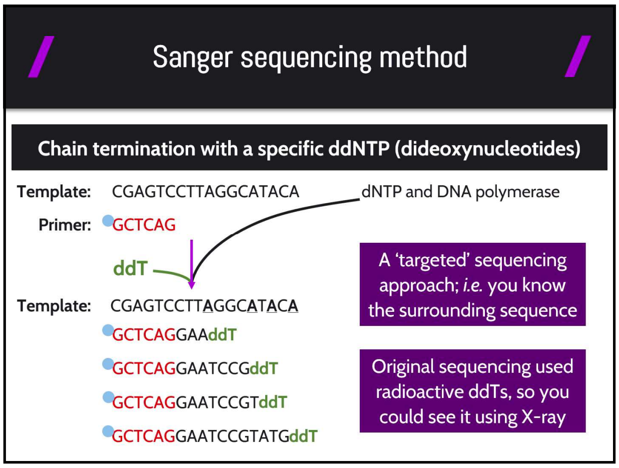


Sanger sequencing method

Chain termination with a specific ddNTP (dideoxynucleotides)

When a ddNTP binds, the DNA strand will stop extending

3



Sanger sequencing method

Chain termination with a specific ddNTP (dideoxynucleotides)

Template: CGAGTCCTTAGGCATACA

Primer: GCTCAG

dNTP and DNA polymerase

ddT

Template: CGAGTCCTTAGGCATACA

GCTCAGGAAddT

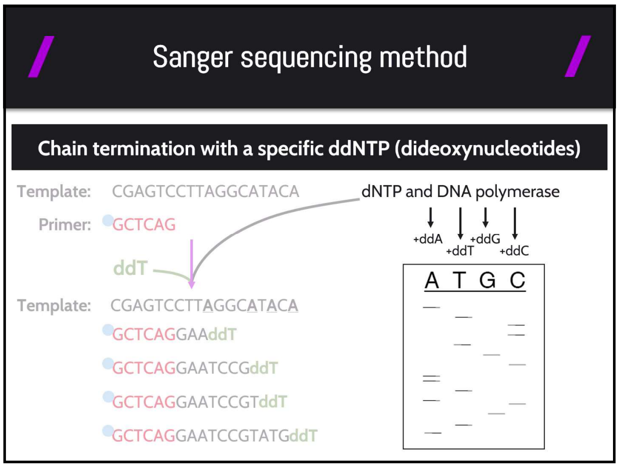
GCTCAGGAATCCGddT

GCTCAGGAATCCGTddT

GCTCAGGAATCCGTATGddT

A 'targeted' sequencing approach; *i.e.* you know the surrounding sequence

Original sequencing used radioactive ddTs, so you could see it using X-ray



Sanger sequencing method

Chain termination with a specific ddNTP (dideoxynucleotides)

Template: CGAGTCCTTAGGCATACA

Primer: GCTCAG

dNTP and DNA polymerase

ddT

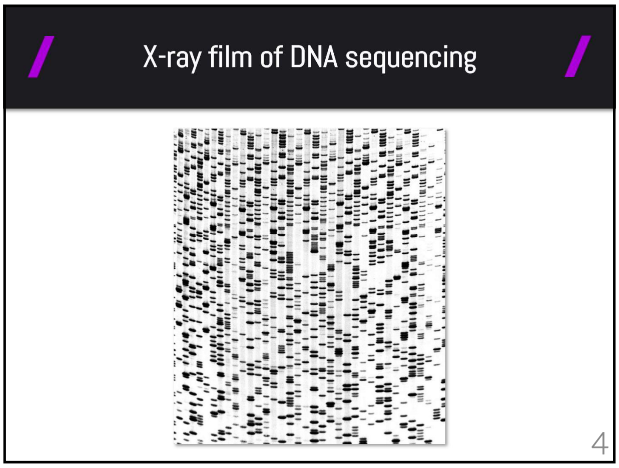
Template: CGAGTCCTTAGGCATACA

GCTCAGGAAddT

GCTCAGGAATCCGddT

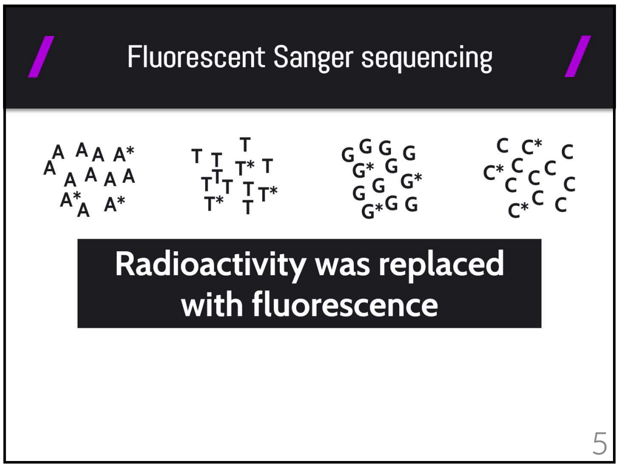
GCTCAGGAATCCGTddT

GCTCAGGAATCCGTATGddT



X-ray film of DNA sequencing

4

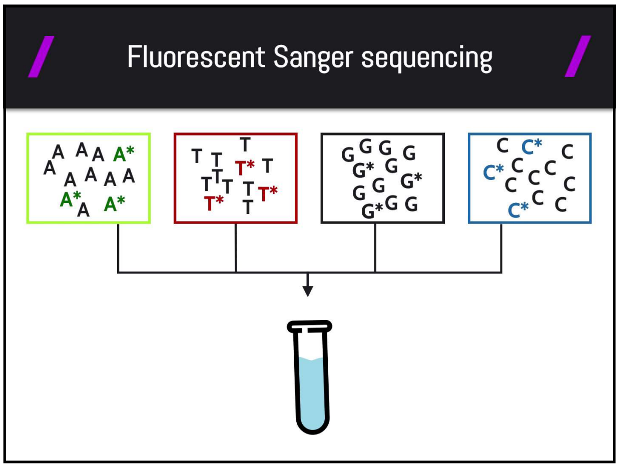


Fluorescent Sanger sequencing

A A A A A* A A A A A* A A*
T T T T* T T T T T T* T* T
G G G G G* G G* G G G* G G
C C C* C C* C C C C C C* C C

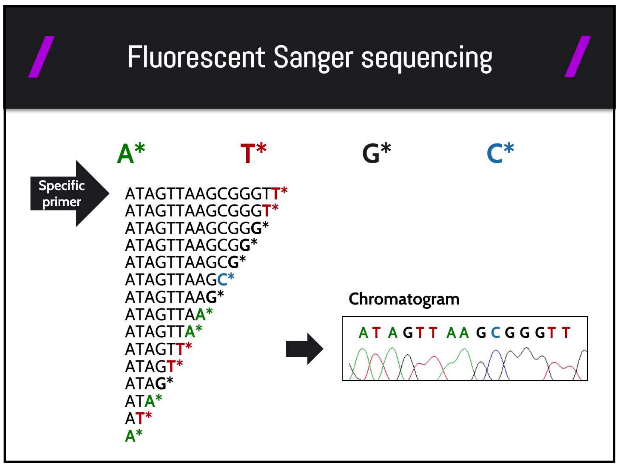
Radioactivity was replaced with fluorescence

5



Fluorescent Sanger sequencing

A A A A A* A A A A A* A A*
T T T T* T T T T T T* T* T
G G G G G* G G* G G G* G G
C C C* C C* C C C C C C* C C



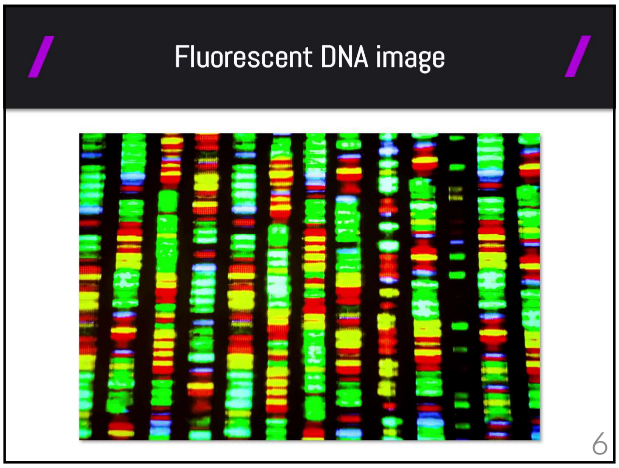
Fluorescent Sanger sequencing

A* T* G* C*

Specific primer

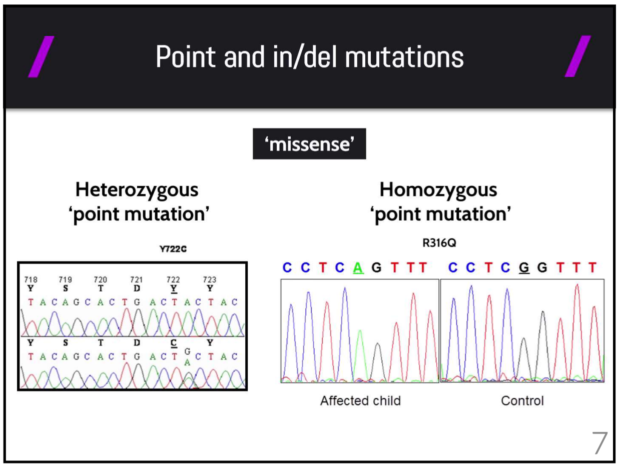
ATAGTTAAGCGGGTT*
ATAGTTAAGCGGGT*
ATAGTTAAGCGGG*
ATAGTTAAGCGG*
ATAGTTAAGCG*
ATAGTTAAGC*
ATAGTTAAG*
ATAGTTAA*
ATAGTTA*
ATAGTT*
ATAGT*
ATAG*
ATA*
AT*
A*

Chromatogram



Fluorescent DNA image

6



Point and in/del mutations

'missense'

Heterozygous 'point mutation'

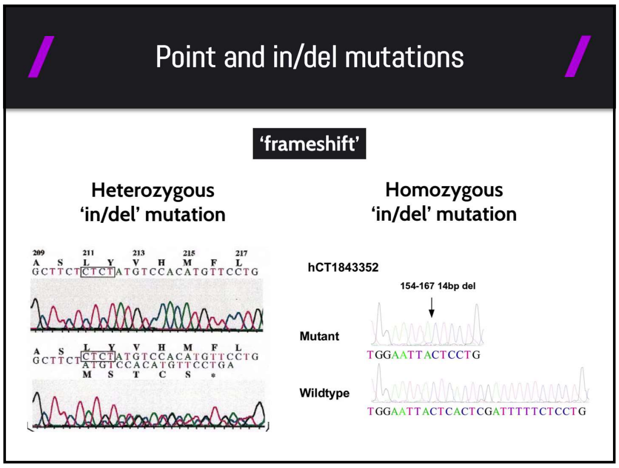
Homozygous 'point mutation'

Affected child

Control

7

Dr. Giles Yeo – University of Cambridge, UK

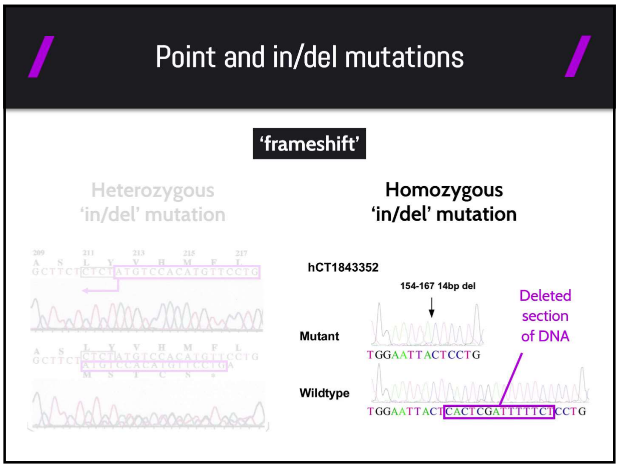


Point and in/del mutations

'frameshift'

Heterozygous 'in/del' mutation

Homozygous 'in/del' mutation

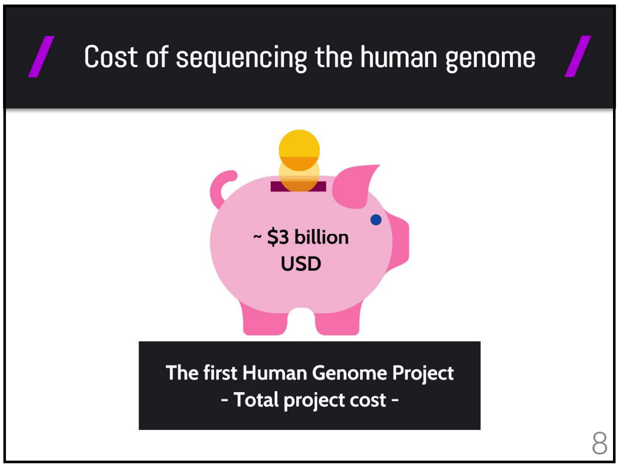


Point and in/del mutations

'frameshift'

Heterozygous 'in/del' mutation

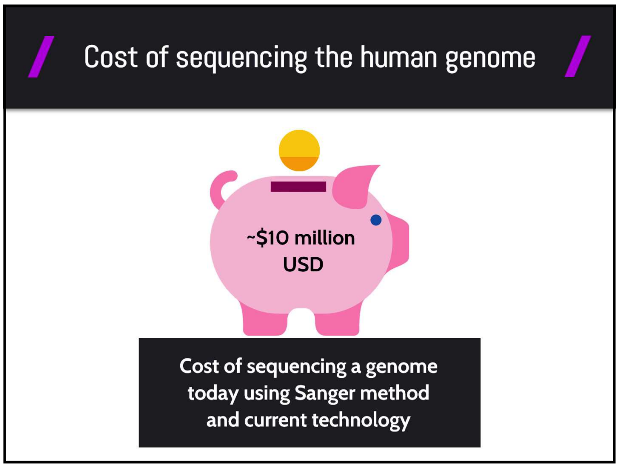
Homozygous 'in/del' mutation



Cost of sequencing the human genome

The first Human Genome Project - Total project cost -

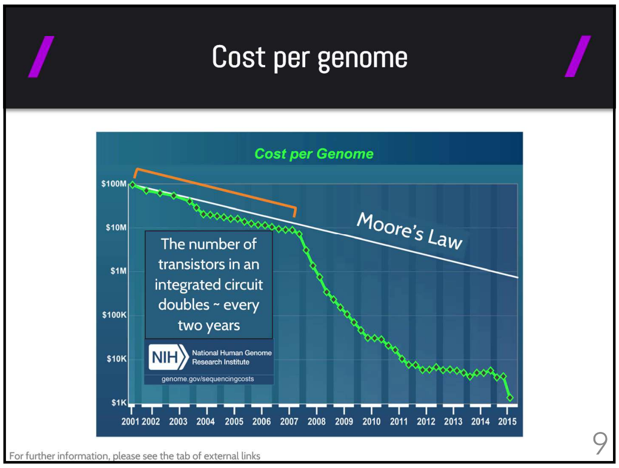
8



Cost of sequencing the human genome

~$10 million USD

Cost of sequencing a genome today using Sanger method and current technology

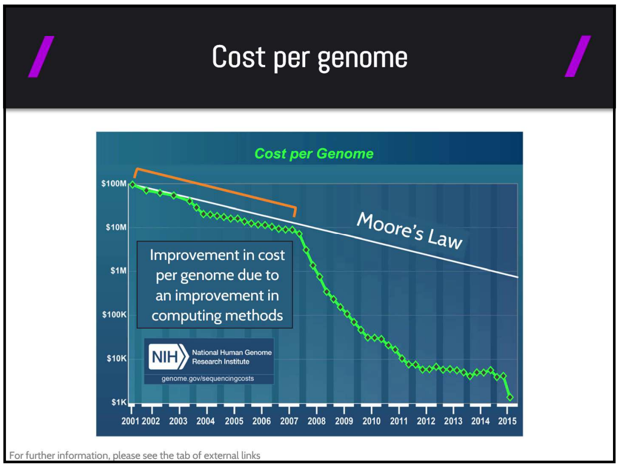


Cost per genome

The number of transistors in an integrated circuit doubles ~ every two years

9

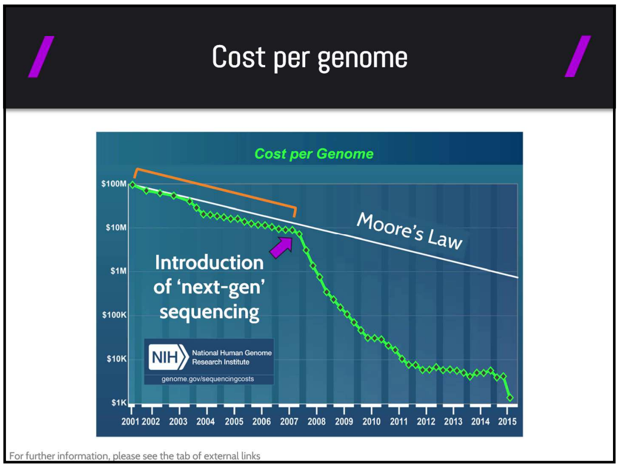
For further information, please see the tab of external links



Cost per genome

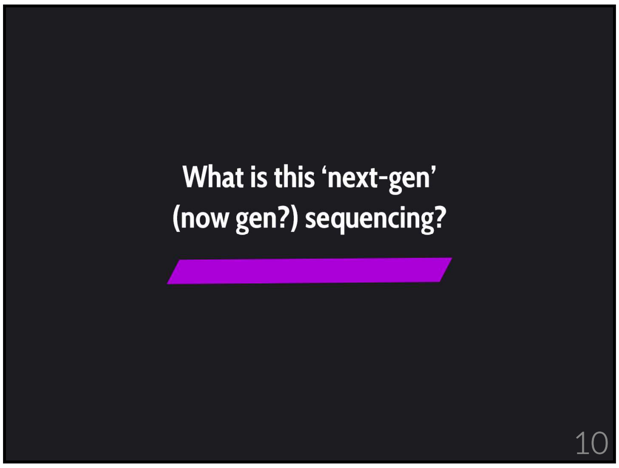
Improvement in cost per genome due to an improvement in computing methods

For further information, please see the tab of external links



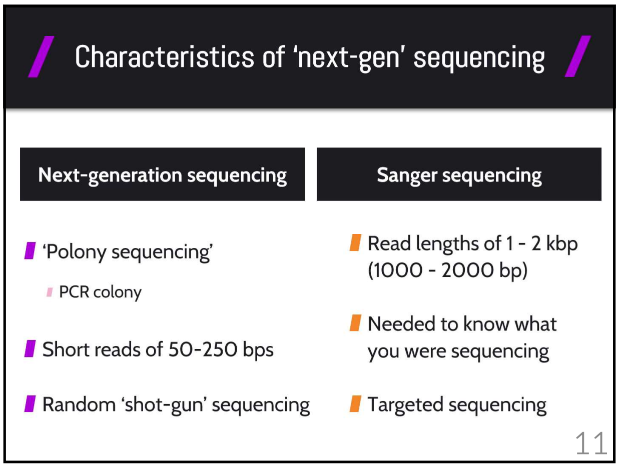
Cost per genome

For further information, please see the tab of external links



What is this 'next-gen' (now gen?) sequencing?

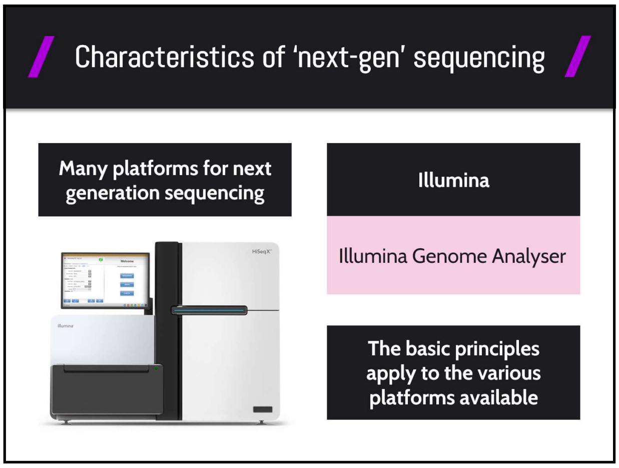
10



Characteristics of 'next-gen' sequencing

Next-generation sequencing	Sanger sequencing
'Polony sequencing' PCR colony	Read lengths of 1 - 2 kbp (1000 - 2000 bp)
Short reads of 50-250 bps	Needed to know what you were sequencing
Random 'shot-gun' sequencing	Targeted sequencing

11



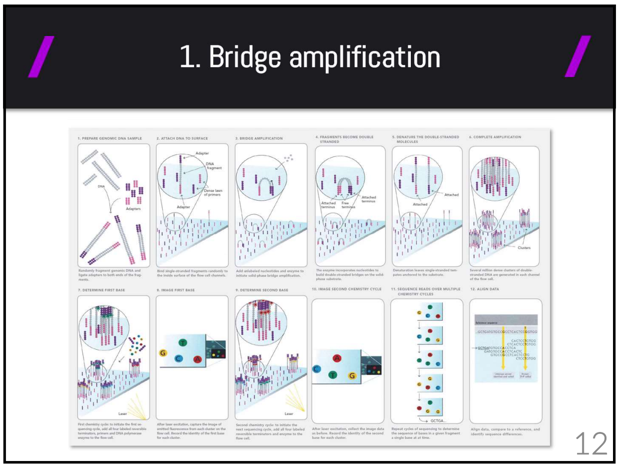
Characteristics of 'next-gen' sequencing

Many platforms for next generation sequencing

Illumina

Illumina Genome Analyser

The basic principles apply to the various platforms available



1. Bridge amplification

1. PREPARE GENOMIC DNA SAMPLE — Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE — Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION — Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED — The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES — Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION — Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE — First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE — After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

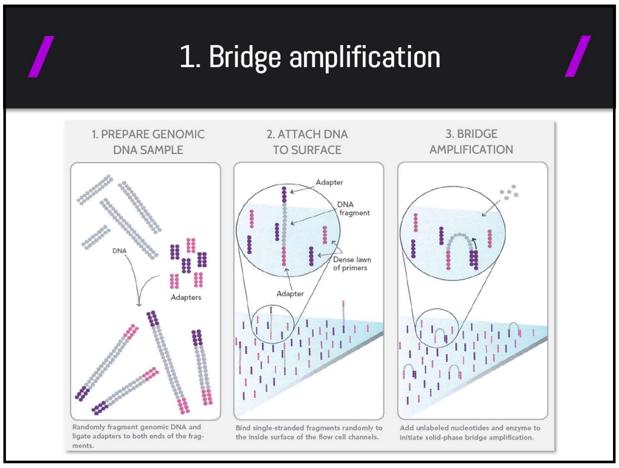
9. DETERMINE SECOND BASE — Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE — After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES — Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at at time.

12. ALIGN DATA — Align data, compare to a reference, and identify sequence differences.

12

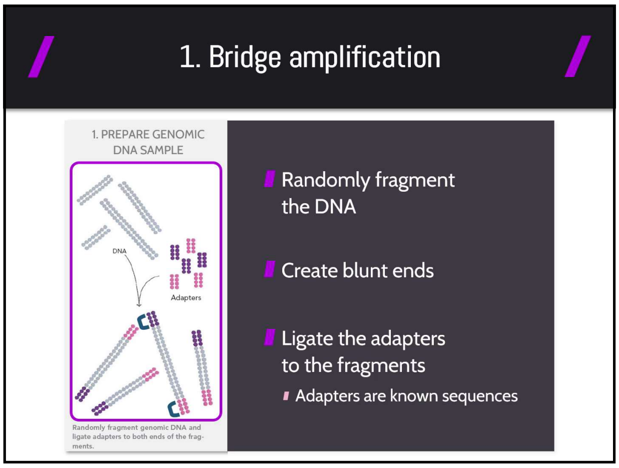


1. Bridge amplification

1. PREPARE GENOMIC DNA SAMPLE — Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE — Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

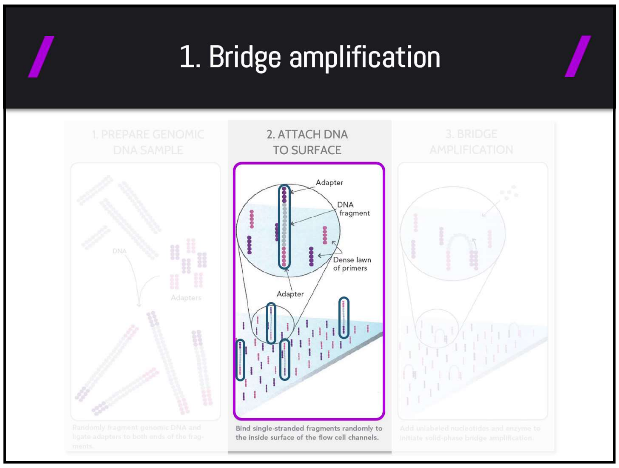
3. BRIDGE AMPLIFICATION — Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



1. Bridge amplification

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

- Randomly fragment the DNA
- Create blunt ends
- Ligate the adapters to the fragments
  - Adapters are known sequences

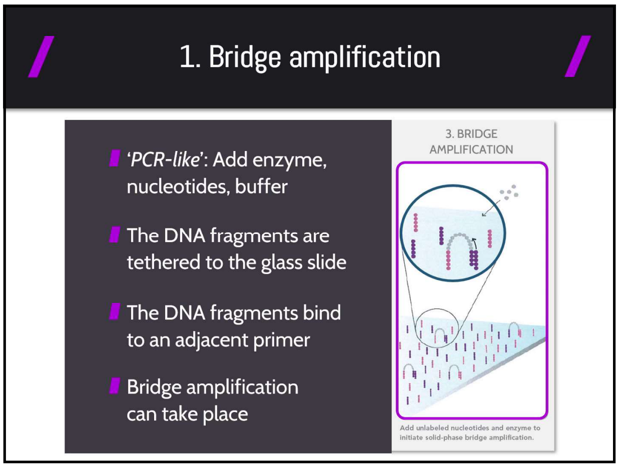


1. Bridge amplification

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

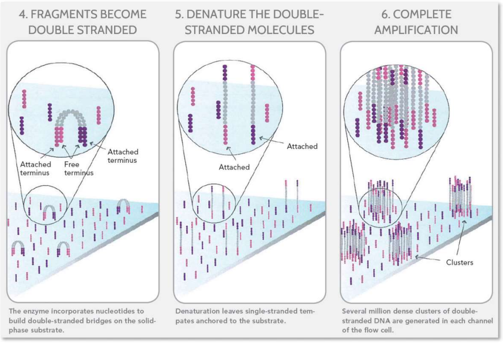


1. Bridge amplification

- *'PCR-like'*: Add enzyme, nucleotides, buffer
- The DNA fragments are tethered to the glass slide
- The DNA fragments bind to an adjacent primer
- Bridge amplification can take place

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

2. 'Polony' formation



4. FRAGMENTS BECOME DOUBLE STRANDED

5. DENATURE THE DOUBLE-STRANDED MOLECULES

6. COMPLETE AMPLIFICATION

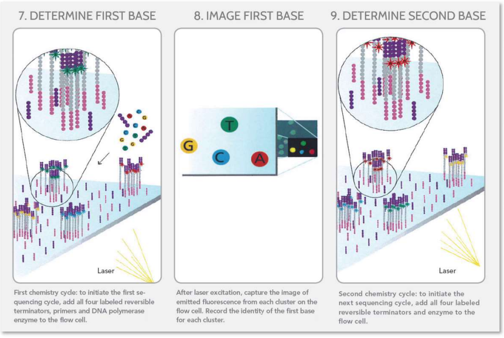
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denaturation leaves single-stranded tempates anchored to the substrate.

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

13

3. Sequencing



7. DETERMINE FIRST BASE

8. IMAGE FIRST BASE

9. DETERMINE SECOND BASE

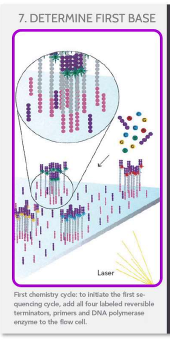
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

14

3. Sequencing



7. DETERMINE FIRST BASE

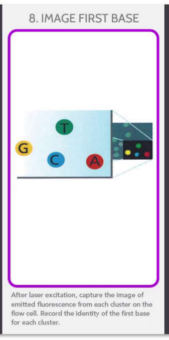
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

- Add primers
- Add fluorescent chain-terminating A, T, G, C
  - Bind after the primer

Dr. Giles Yeo – University of Cambridge, UK

3. Sequencing

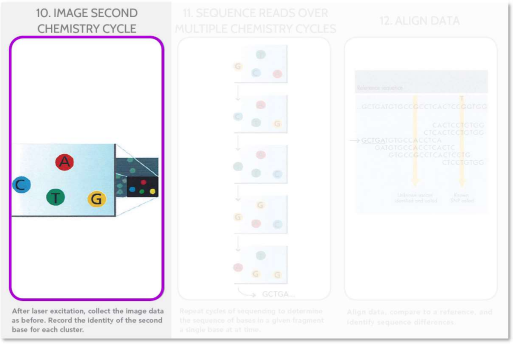
- Excite the base pair
- Take a picture of the emitted fluorescence



After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

- Deprotect the base pair to:
  - Remove the fluorescence
  - Un-terminates the terminator

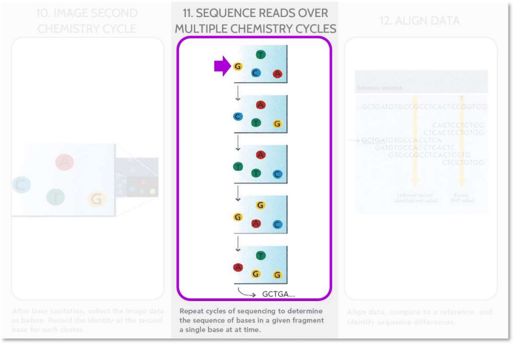
4. Sequencing over multiple cycles


10. IMAGE SECOND CHEMISTRY CYCLE
A
C
T
G
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
GCTGA...
12. ALIGN DATA
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at at time.
Align data, compare to a reference, and identify sequence differences.

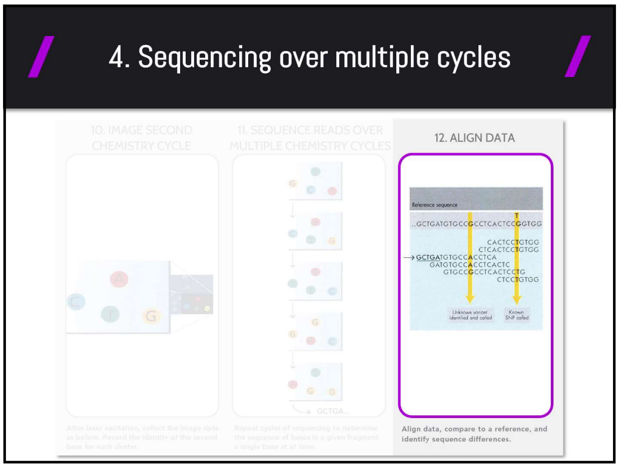
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

15

4. Sequencing over multiple cycles


10. IMAGE SECOND CHEMISTRY CYCLE
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
G
A
T
C
GCTGA...
12. ALIGN DATA
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Align data, compare to a reference, and identify sequence differences.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at at time.



4. Sequencing over multiple cycles

10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at at time.

12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

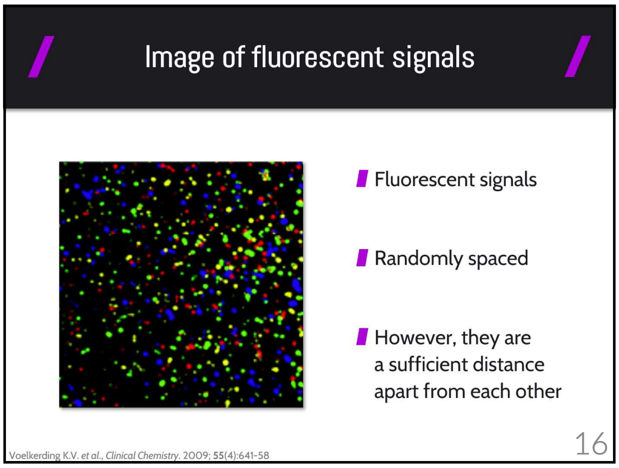
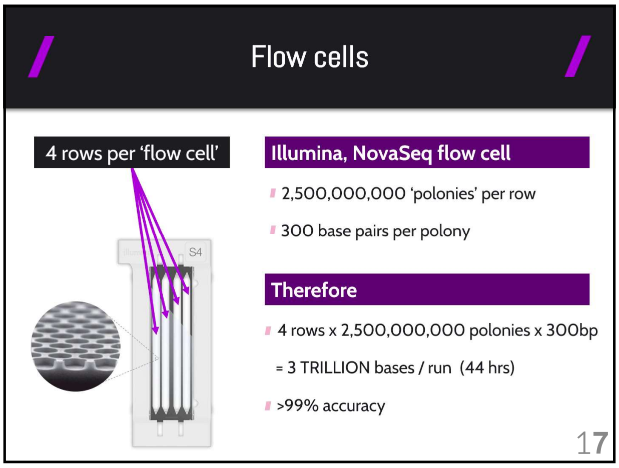


Image of fluorescent signals

- Fluorescent signals
- Randomly spaced
- However, they are a sufficient distance apart from each other

Voelkerding K.V. *et al.*, *Clinical Chemistry.* 2009; 55(4):641-58

16



Flow cells

Illumina, NovaSeq flow cell

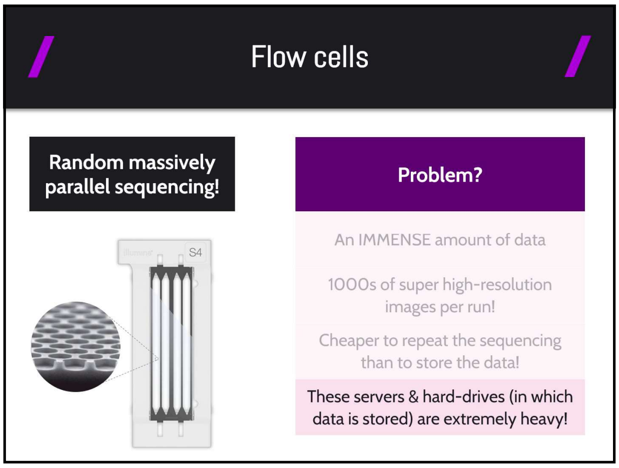
- 2,500,000,000 'polonies' per row
- 300 base pairs per polony

Therefore

- 4 rows x 2,500,000,000 polonies x 300bp

  = 3 TRILLION bases / run (44 hrs)
- >99% accuracy

17



Flow cells

Random massively parallel sequencing!

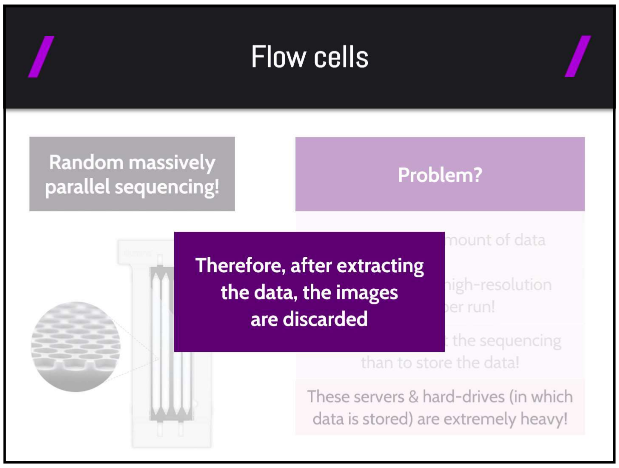
Problem?

An IMMENSE amount of data

1000s of super high-resolution images per run!

Cheaper to repeat the sequencing than to store the data!

These servers & hard-drives (in which data is stored) are extremely heavy!



Flow cells

Random massively parallel sequencing!

Problem?

mount of data

igh-resolution

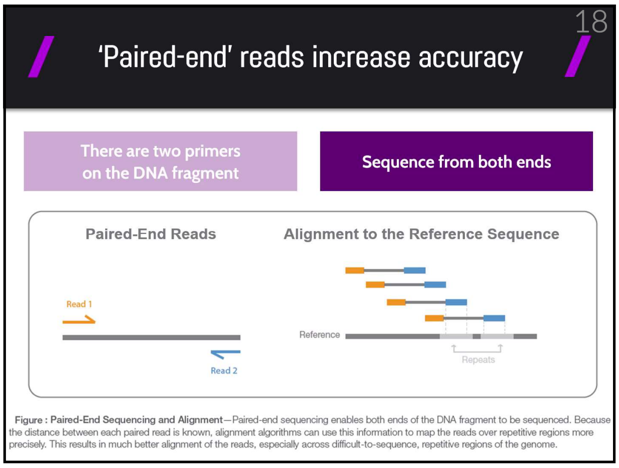
er run!

the sequencing

than to store the data!

These servers & hard-drives (in which data is stored) are extremely heavy!

Therefore, after extracting the data, the images are discarded



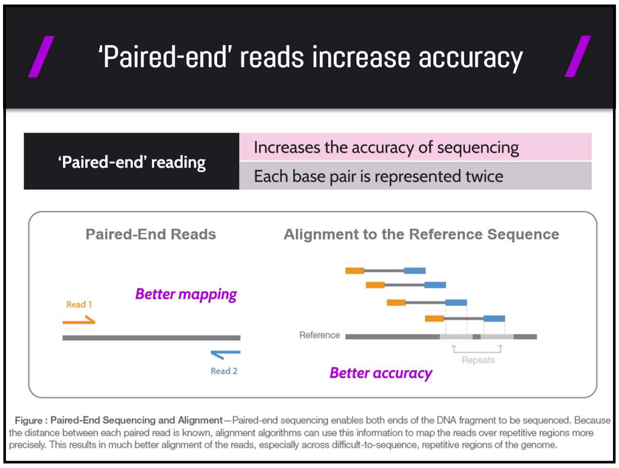
18

'Paired-end' reads increase accuracy

There are two primers on the DNA fragment

Sequence from both ends

Figure : Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.



'Paired-end' reads increase accuracy

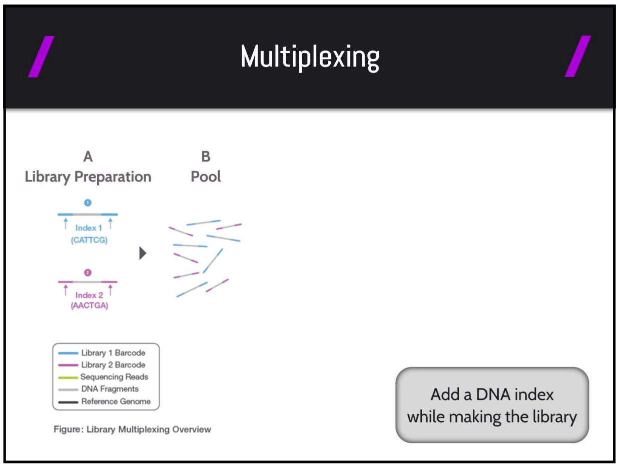
'Paired-end' reading	
	Increases the accuracy of sequencing
	Each base pair is represented twice

Figure : Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Multiplexing

Multiplexing allows for an increased number of samples

19



Multiplexing

Figure: Library Multiplexing Overview

Add a DNA index while making the library

Multiplexing

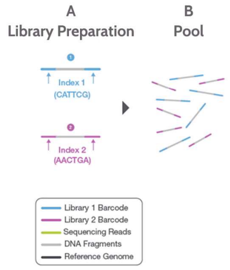


Figure: Library Multiplexing Overview

Mix the no. of samples down one flow cell

Multiplexing

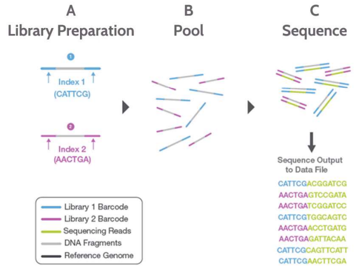


Figure: Library Multiplexing Overview

Can do all of the sequencing together

Multiplexing

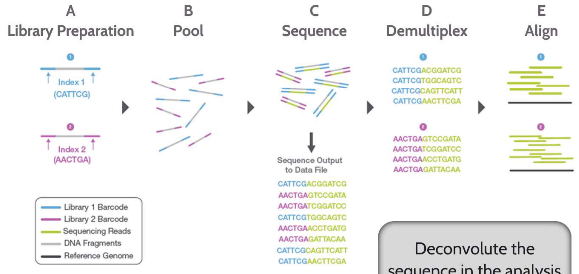
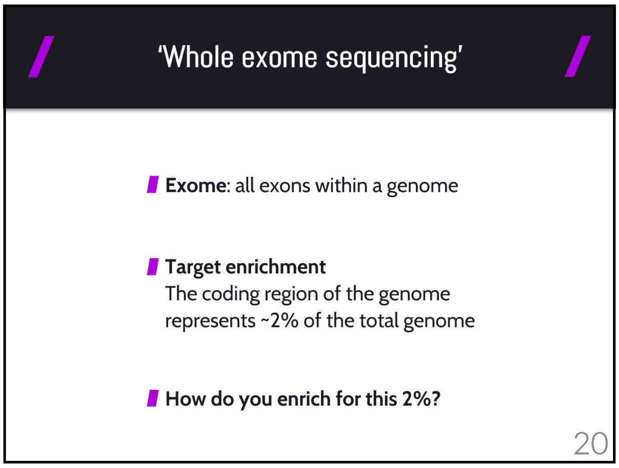


Figure: Library Multiplexing Overview

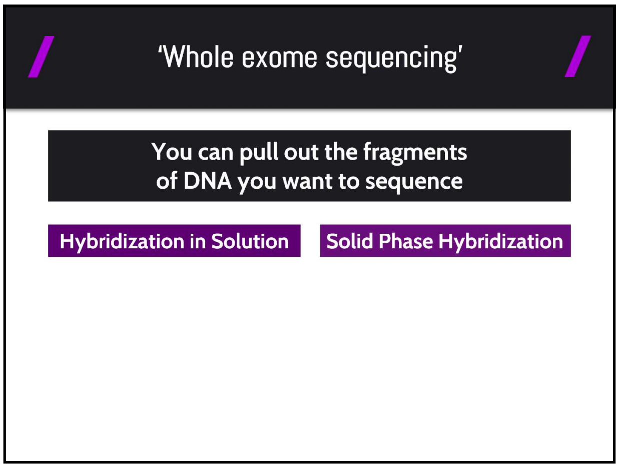
Deconvolute the sequence in the analysis



'Whole exome sequencing'

- **Exome**: all exons within a genome
- **Target enrichment**
  The coding region of the genome represents ~2% of the total genome
- **How do you enrich for this 2%?**

20

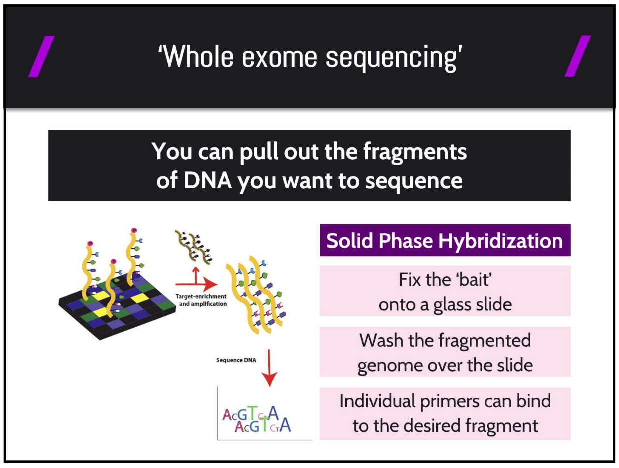


'Whole exome sequencing'

You can pull out the fragments of DNA you want to sequence

Hybridization in Solution

Solid Phase Hybridization



'Whole exome sequencing'

You can pull out the fragments of DNA you want to sequence

Solid Phase Hybridization

Fix the 'bait' onto a glass slide

Wash the fragmented genome over the slide

Individual primers can bind to the desired fragment

‘Whole exome sequencing’

You can pull out the fragments of DNA you want to sequence

Hybridization in Solution

Solid Phase Hybridization

Droplet PCR

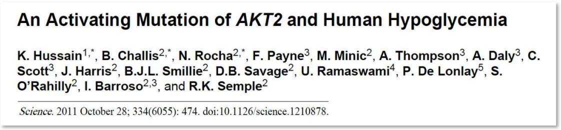
Merge the primers with bubbles of DNA

Mosaic mutations

The cost of sequencing is being reduced

21

Mosaic mutations



An Activating Mutation of *AKT2* and Human Hypoglycemia

K. Hussain[1,*], B. Challis[2,*], N. Rocha[2,*], F. Payne[3], M. Minic[2], A. Thompson[3], A. Daly[3], C. Scott[3], J. Harris[2], B.J.L. Smillie[2], D.B. Savage[2], U. Ramaswami[4], P. De Lonlay[5], S. O’Rahilly[2], I. Barroso[2,3], and R.K. Semple[2]

Science. 2011 October 28; 334(6055): 474. doi:10.1126/science.1210878.

- Digital sequencing
- You can count the copies of every fragment
  - Each of the base pairs is sequenced hundreds of times

Mosaic mutations

- Can pick up **mosaic mutations**
  - This is difficult with Sanger sequencing
  - **Next Gen Sequencing:** picks up a small percentage of mutant fragments of DNA

NOT every cell carries the mutation

Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in *PIK3CA*

Marjorie J Lindhurst[1,16], Victoria E R Parker[2,16], Felicity Payne[3], Julie C Sapp[1], Simon Rudge[4], Julie Harris[2], Alison M Witkowski[1], Qifeng Zhang[4], Matthijs P Groeneveld[2], Carol E Scott[3], Allan Daly[3], Susan M Huson[5], Laura L Tosi[6], Michael L Cunningham[7], Thomas N Darling[8], Joseph Geer[9], Zoran Gucev[10], V Reid Sutton[11], Christos Tziotzios[12], Adrian K Dixon[13], Timothy Helliwell[14], Stephen O'Rahilly[2,15], David B Savage[2,15], Michael J O Wakelam[4], Inês Barroso[2,3], Leslie G Biesecker[1] & Robert K Semple[2,15]

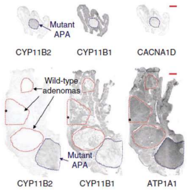
VOLUME 44 | NUMBER 8 | AUGUST 2012 NATURE GENETICS

Identifying genetic drivers in a specific cancer

Somatic mutations in *ATP1A1* and *CACNA1D* underlie a common subtype of adrenal hypertension

Elena A B Azizan[1,12], Hanne Poulsen[2,12], Petronel Tuluc[3,12], Junhua Zhou[1,12], Michael V Clausen[2], Andreas Lieb[3], Carmela Maniero[1], Sumedha Garg[4], Elena G Bochukova[4], Wanfeng Zhao[5], Lalarukh Haris Shaikh[1], Cheryl A Brighton[1], Ada E D Teo[1], Anthony P Davenport[1], Tanja Dekkers[6], Bas Tops[7], Benno Küsters[7], Jiri Ceral[8], Giles S H Yeo[4], Sudeshna Guha Neogi[9], Ian McFarlane[9], Nitzan Rosenfeld[10], Francesco Marass[10], James Hadfield[10], Wojciech Margas[11], Kanchan Chaggar[11], Miroslav Solar[8], Jaap Deinum[6], Annette C Dolphin[11], I Sadaf Farooqi[4,12], Joerg Striessnig[3,12], Poul Nissen[2,12] & Morris J Brown[1,12]

Identifying the genetic driver in a specific cancer



- Laser capture microdissection
- Compare the adenoma with the wild-type tissue *via* next generation sequencing of whole exomes

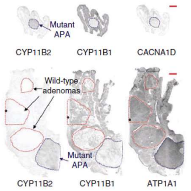
22

Identifying genetic drivers in a specific cancer

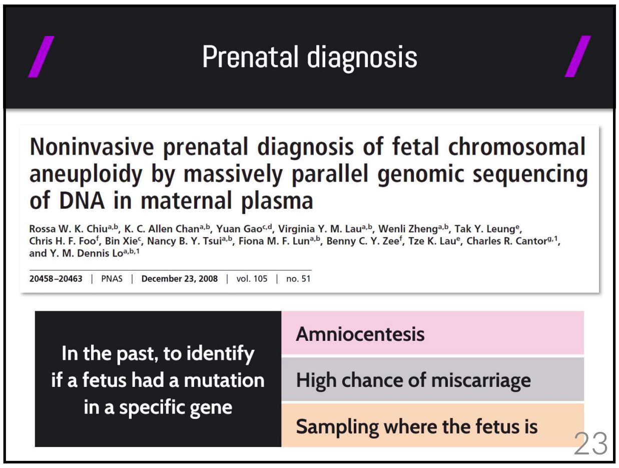
Somatic mutations in *ATP1A1* and *CACNA1D* underlie a common subtype of adrenal hypertension

Elena A B Azizan[1,12], Hanne Poulsen[2,12], Petronel Tuluc[3,12], Junhua Zhou[1,12], Michael V Clausen[2], Andreas Lieb[3], Carmela Maniero[1], Sumedha Garg[4], Elena G Bochukova[4], Wanfeng Zhao[5], Lalarukh Haris Shaikh[1], Cheryl A Brighton[1], Ada E D Teo[1], Anthony P Davenport[1], Tanja Dekkers[6], Bas Tops[7], Benno Küsters[7], Jiri Ceral[8], Giles S H Yeo[4], Sudeshna Guha Neogi[9], Ian McFarlane[9], Nitzan Rosenfeld[10], Francesco Marass[10], James Hadfield[10], Wojciech Margas[11], Kanchan Chaggar[11], Miroslav Solar[8], Jaap Deinum[6], Annette C Dolphin[11], I Sadaf Farooqi[4,12], Joerg Striessnig[3,12], Poul Nissen[2,12] & Morris J Brown[1,12]

Identifying the genetic driver in a specific cancer



Somatic mutation driven disease



Prenatal diagnosis

Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma

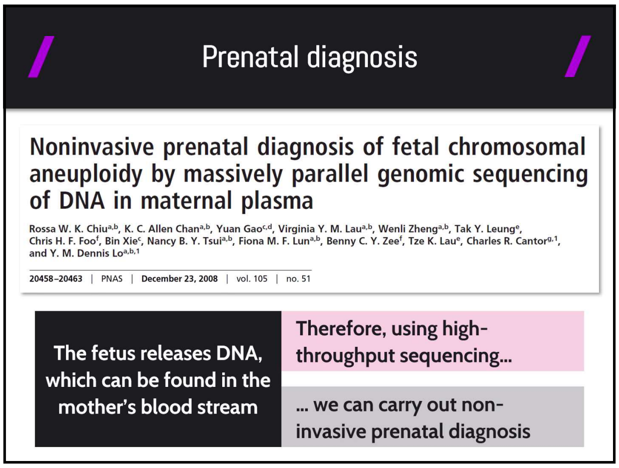
Rossa W. K. Chiu[a,b], K. C. Allen Chan[a,b], Yuan Gao[c,d], Virginia Y. M. Lau[a,b], Wenli Zheng[a,b], Tak Y. Leung[e], Chris H. F. Foo[f], Bin Xie[c], Nancy B. Y. Tsui[a,b], Fiona M. F. Lun[a,b], Benny C. Y. Zee[f], Tze K. Lau[e], Charles R. Cantor[g,1], and Y. M. Dennis Lo[a,b,1]

20458–20463 | PNAS | December 23, 2008 | vol. 105 | no. 51

In the past, to identify if a fetus had a mutation in a specific gene

- Amniocentesis
- High chance of miscarriage
- Sampling where the fetus is

23



Prenatal diagnosis

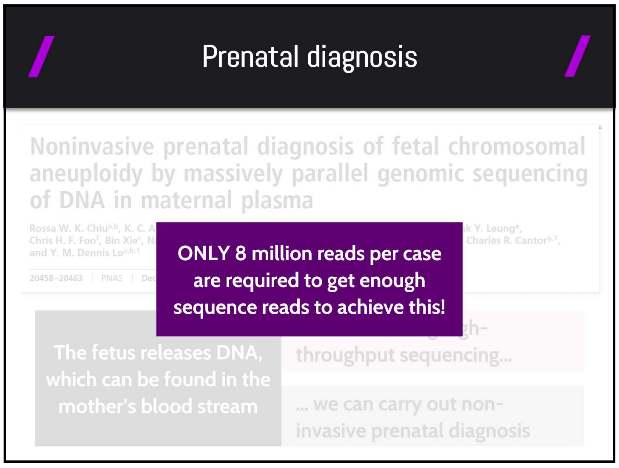
Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma

Rossa W. K. Chiu[a,b], K. C. Allen Chan[a,b], Yuan Gao[c,d], Virginia Y. M. Lau[a,b], Wenli Zheng[a,b], Tak Y. Leung[e], Chris H. F. Foo[f], Bin Xie[c], Nancy B. Y. Tsui[a,b], Fiona M. F. Lun[a,b], Benny C. Y. Zee[f], Tze K. Lau[e], Charles R. Cantor[g,1], and Y. M. Dennis Lo[a,b,1]

20458–20463 | PNAS | December 23, 2008 | vol. 105 | no. 51

The fetus releases DNA, which can be found in the mother's blood stream

- Therefore, using high-throughput sequencing…
- … we can carry out non-invasive prenatal diagnosis



Prenatal diagnosis

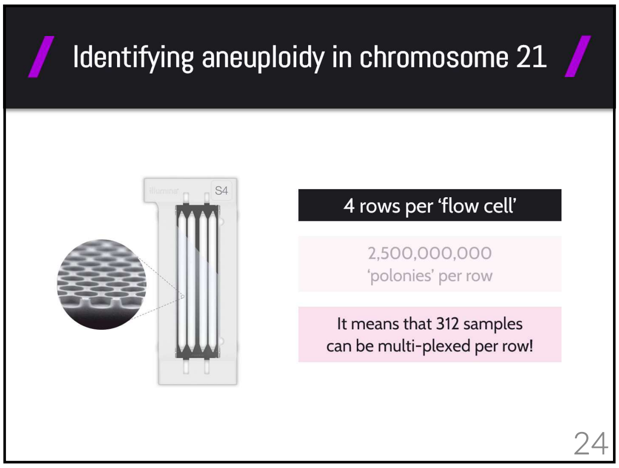
Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma

The fetus releases DNA, which can be found in the mother's blood stream

throughput sequencing…

… we can carry out non-invasive prenatal diagnosis

ONLY 8 million reads per case are required to get enough sequence reads to achieve this!



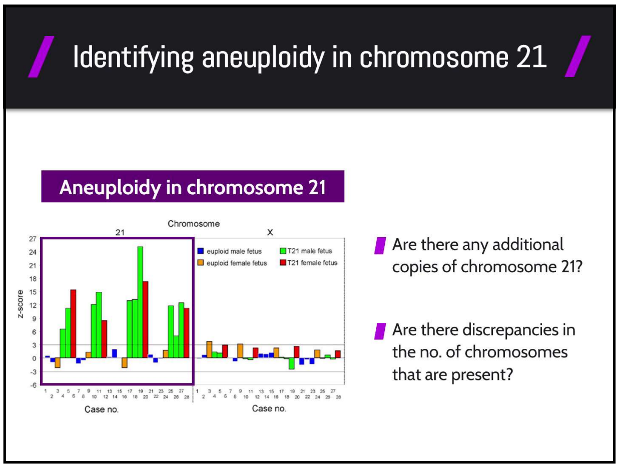
Identifying aneuploidy in chromosome 21

4 rows per 'flow cell'

2,500,000,000 'polonies' per row

It means that 312 samples can be multi-plexed per row!

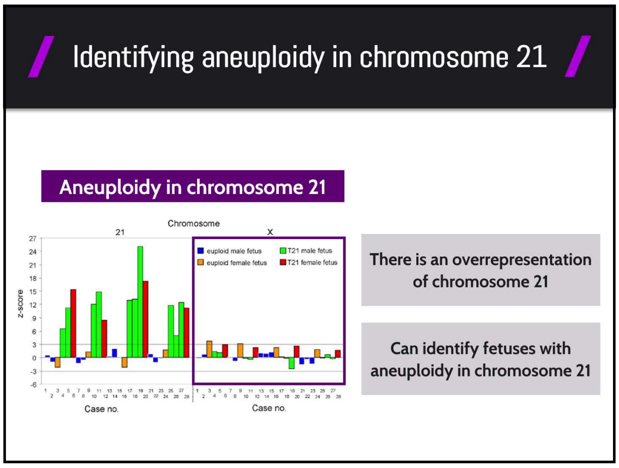
24



Identifying aneuploidy in chromosome 21

Aneuploidy in chromosome 21

- Are there any additional copies of chromosome 21?
- Are there discrepancies in the no. of chromosomes that are present?

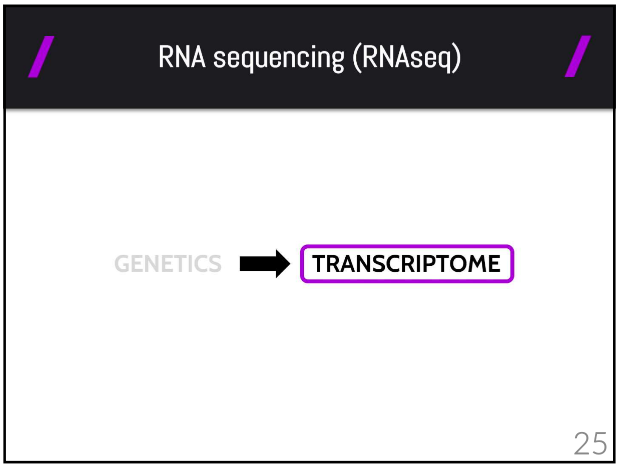


Identifying aneuploidy in chromosome 21

Aneuploidy in chromosome 21

There is an overrepresentation of chromosome 21

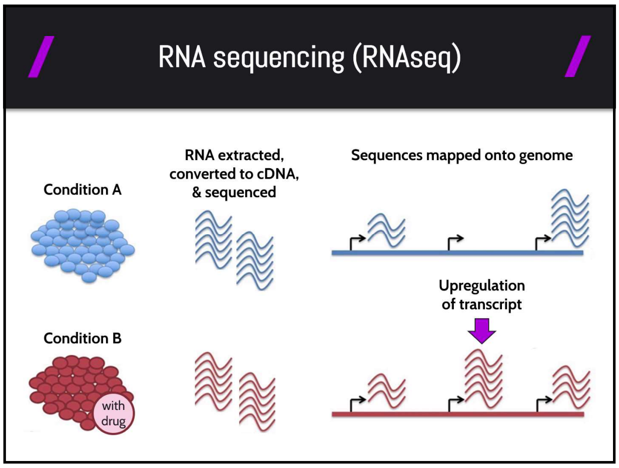
Can identify fetuses with aneuploidy in chromosome 21



RNA sequencing (RNAseq)

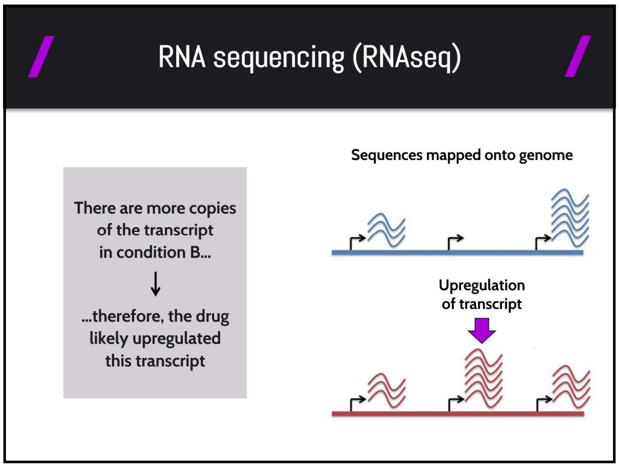
GENETICS → TRANSCRIPTOME

25



RNA sequencing (RNAseq)

Condition A
RNA extracted, converted to cDNA, & sequenced
Sequences mapped onto genome
Upregulation of transcript
Condition B
with drug



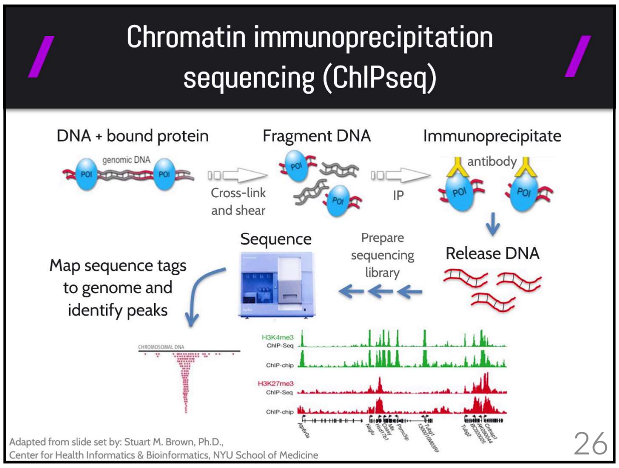
RNA sequencing (RNAseq)

There are more copies of the transcript in condition B…

↓

…therefore, the drug likely upregulated this transcript

Dr. Giles Yeo – University of Cambridge, UK



Chromatin immunoprecipitation sequencing (ChIPseq)

DNA + bound protein

Fragment DNA

Immunoprecipitate

Cross-link and shear

IP

Release DNA

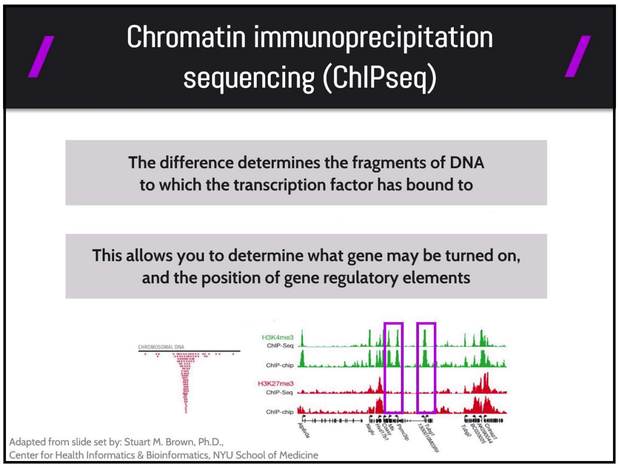
Prepare sequencing library

Sequence

Map sequence tags to genome and identify peaks

Adapted from slide set by: Stuart M. Brown, Ph.D., Center for Health Informatics & Bioinformatics, NYU School of Medicine

26

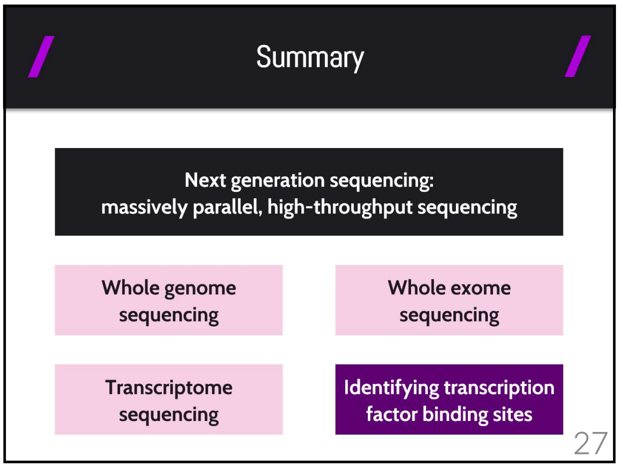


Chromatin immunoprecipitation sequencing (ChIPseq)

The difference determines the fragments of DNA to which the transcription factor has bound to

This allows you to determine what gene may be turned on, and the position of gene regulatory elements

Adapted from slide set by: Stuart M. Brown, Ph.D., Center for Health Informatics & Bioinformatics, NYU School of Medicine



Summary

Next generation sequencing: massively parallel, high-throughput sequencing

Whole genome sequencing

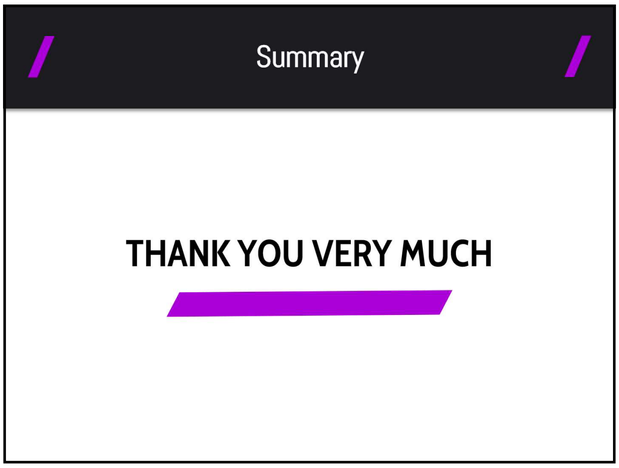
Whole exome sequencing

Transcriptome sequencing

Identifying transcription factor binding sites

27

Dr. Giles Yeo – University of Cambridge, UK



Summary

THANK YOU VERY MUCH