WORKFLOWS FOR BIG DATA

Dr. Angela McGaughran

Genomes and Biodiversity Workshop
Sydney; 20-22 November 2019

Image: Fullgen Technologies

Hello!

- Population genomics and rapid evolution

  → Understanding how populations respond to environmental change
- DECRA project
- More here:

  https://researchers.anu.edu.au/researchers/mcgaughran-a

  www.ang-mcgaughran.com
- Field → Lab → Computer

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

WORKFLOWS FOR BIG DATA

Dr. Angela McGaughran

Genomes and Biodiversity Workshop
Sydney; 20-22 November 2019

Image: Fullgen Technologies

Workshop outline

- Sequencing technologies
- Library preparation
- Experimental design
- Common bioinformatic problems
- Worked example in R

- *Some of the content from today's slides from:

  http://evomics.org/workshops/workshop-on-genomics/2019-workshop-on-genomics-cesky-krumlov/

ang.mcgaughran@gmail.com

Heredity material

- In the 1940s, we knew that chromosomes could be seen under a light microscope

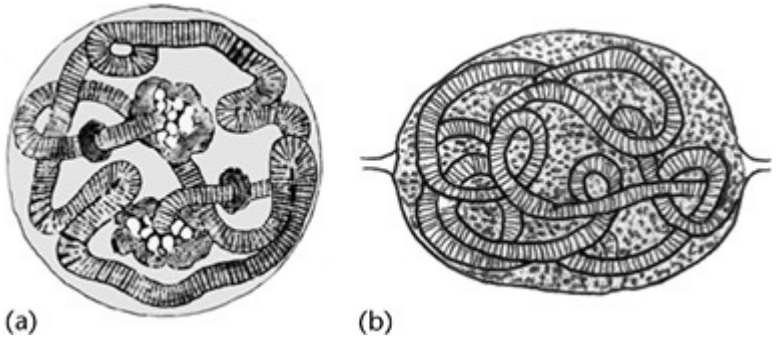


Figure 1. First drawings of polytene chromosome made by Balbiani in (1881) (a) and (1890) (b). (a) Salivary gland cells of *Chironomus plumosus* and (b) Macronucleus of *Loxophyllum meleagris.*

Image: Zhimulev and Koryakov. (2009) Polytene Chromosomes. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons.

ang.mcgaughran@gmail.com

Heredity material

- We knew that DNA was a sequence of four repeating nucleotides that provided structural support to chromosomes

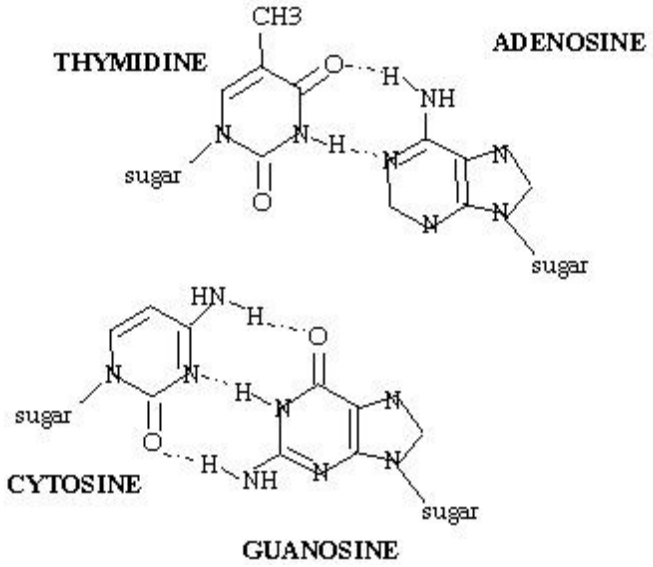


Image: http://www.madsci.org/posts/archives/1998-11/910892899.Mb.r.html

ang.mcgaughran@gmail.com

Heredity material

- In 1953, the structure of DNA was elucidated

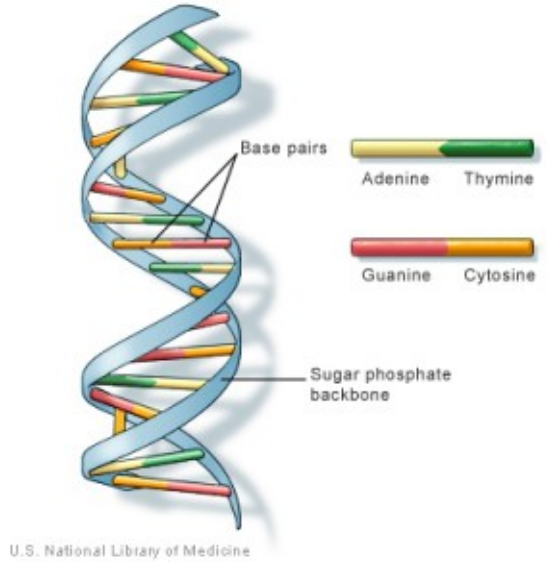


Image: U.S. National Library of Medicine.


Image: http://www.thehistoryblog.com/archives/25193

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

Sanger Sequencing

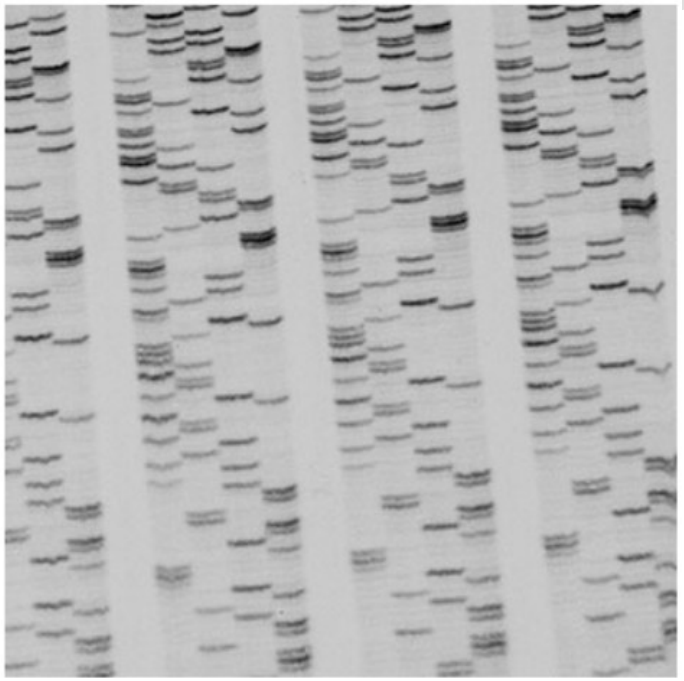


Image credit: https://unlockinglifescode.org/timeline/11

- In 1977, Sanger *et al*. described a new method of 'sequencing by synthesis'
  - Four reactions
  - Each reaction contains the four normal nucleotides plus one dideoxynucleotide (modified to prevent the addition of subsequent nucleotides → they end a nucleotide chain when incorporated)
  - Each of the reactions produces all possible fragments for its' particular base:

ang.mcgaughran@gmail.com

Sanger Sequencing

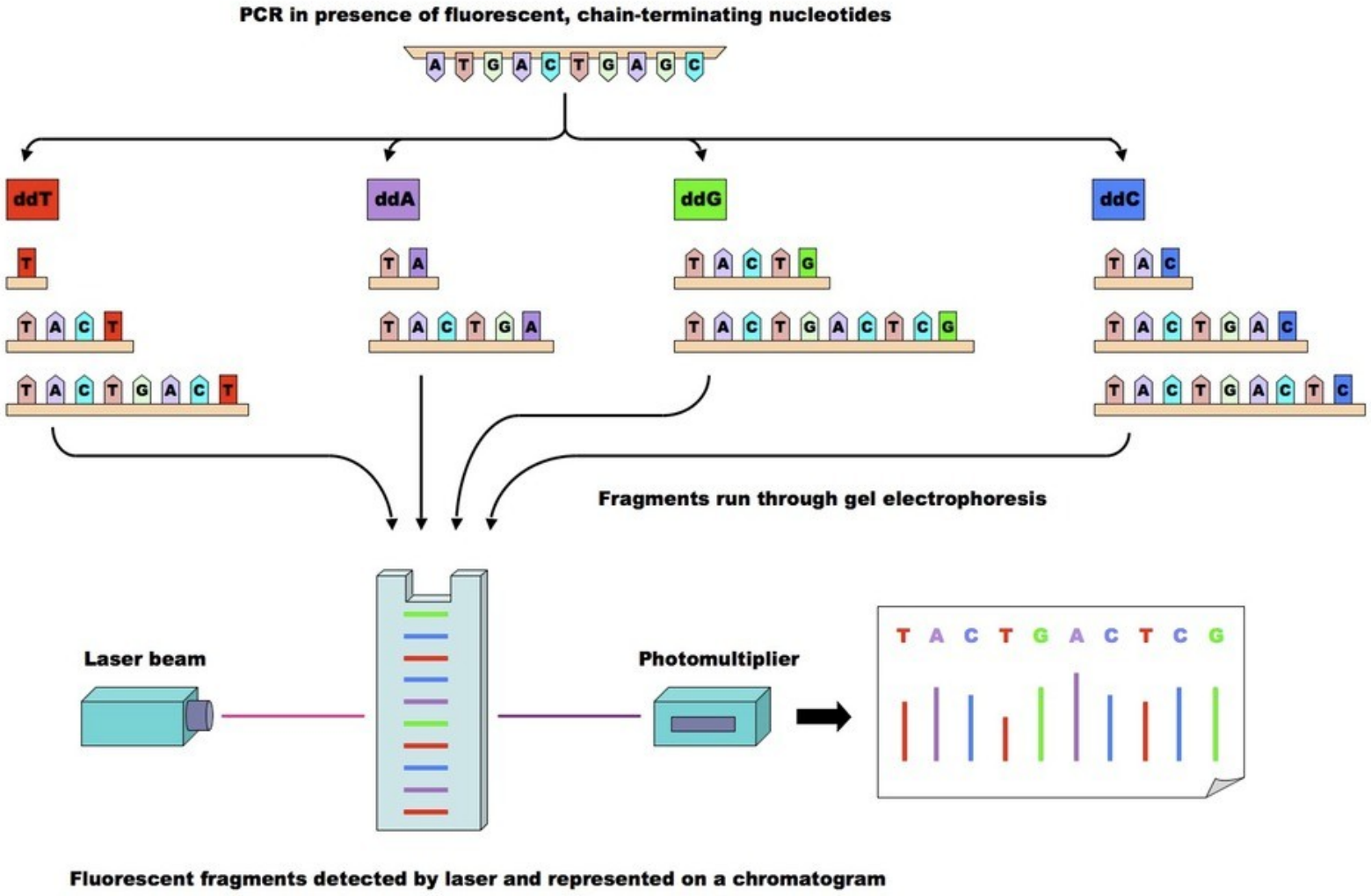


Image: BioNinja

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

Sanger Sequencing

- Dideoxynucleotides are fluorescently labelled so can be detected by a laser when run through gel electrophoresis
- Sequence represented as a chromatogram:

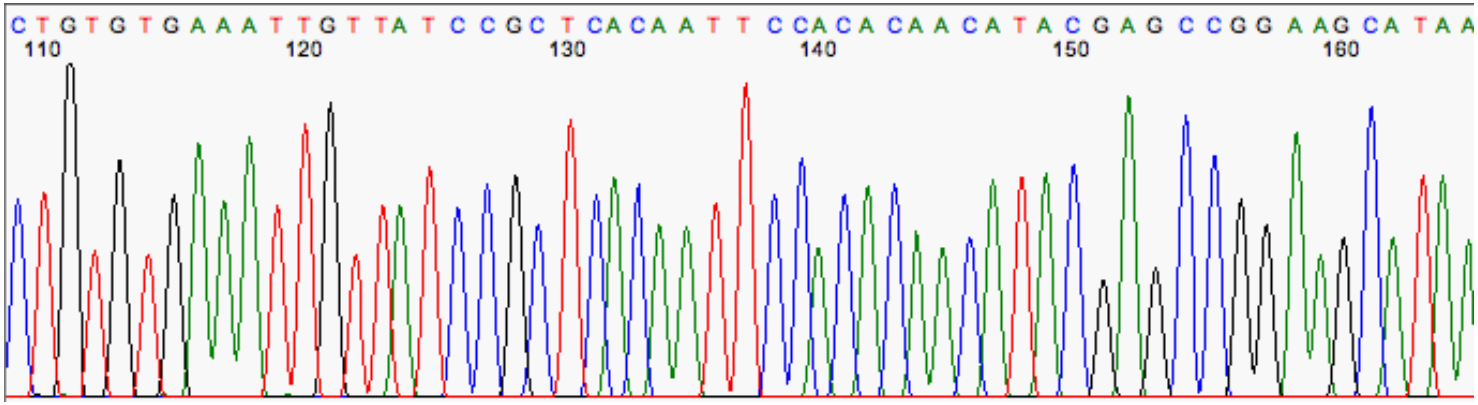


Image: hpc.ilri.cgiar.org

ang.mcgaughran@gmail.com

Strengths and limitations

Method	Pros	Cons	Applications
1st Gen (Sanger)	• Very accurate • Good for highly repetitive regions • Low technological barriers	• Very expensive ($/bp) • Limited read lengths (first 50 bp not great, starts deteriorating after about 750 bp, upper limit ~1000 bp)	• Resolving ambiguous/ repetitive regions • 5'/3' RACE • Plasmid sequencing

ang.mcgaughran@gmail.com

2nd Generation Sequencing

- Sanger = 1st generation sequencing
- 2nd = sequencing is done on clusters of molecules
  - Enabled with various methods of PCR amplification
  - Requires the construction of libraries with adapters
- Typically uses '**multiplexing**'
- Read lengths are MUCH shorter than Sanger sequencing (→'short read' technology), but the volume of data is MASSIVE

ang.mcgaughran@gmail.com

(1) Pyrosequencing

- Ion Torrent / Roche 454:
  - Sequencing by synthesis
  - Uses a silicon chip to hold the DNA
  - Rather than optics (to detect fluorescence), the chip works like a pH meter, detecting when a Hydrogen atom is released after the incorporation of a nucleotide
  - Nucleotides are washed over the chip every few seconds

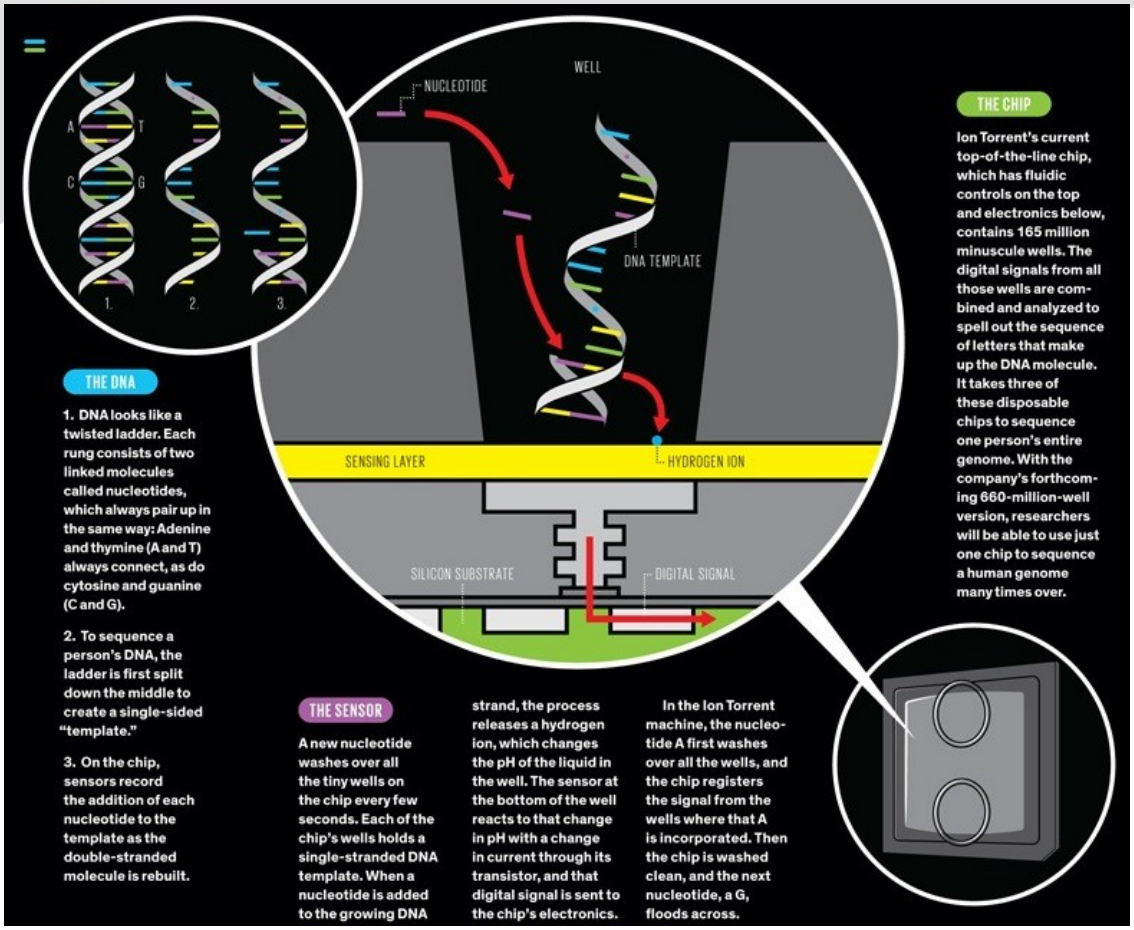


Image: https://www.topcoder.com/case-studies/ion-torrent/

ang.mcgaughran@gmail.com

(2) Illumina

- Also sequencing by synthesis
- Uses fluorescent dNTPs that block the continued binding of further nucleotides (similar to Sanger)
- A digital camera captures the fluorescence image
- The fluorophore is cleaved off and the cycle continues

ang.mcgaughran@gmail.com

The core processes

- Library preparation
- Cluster generation
- Sequencing
- Data analysis

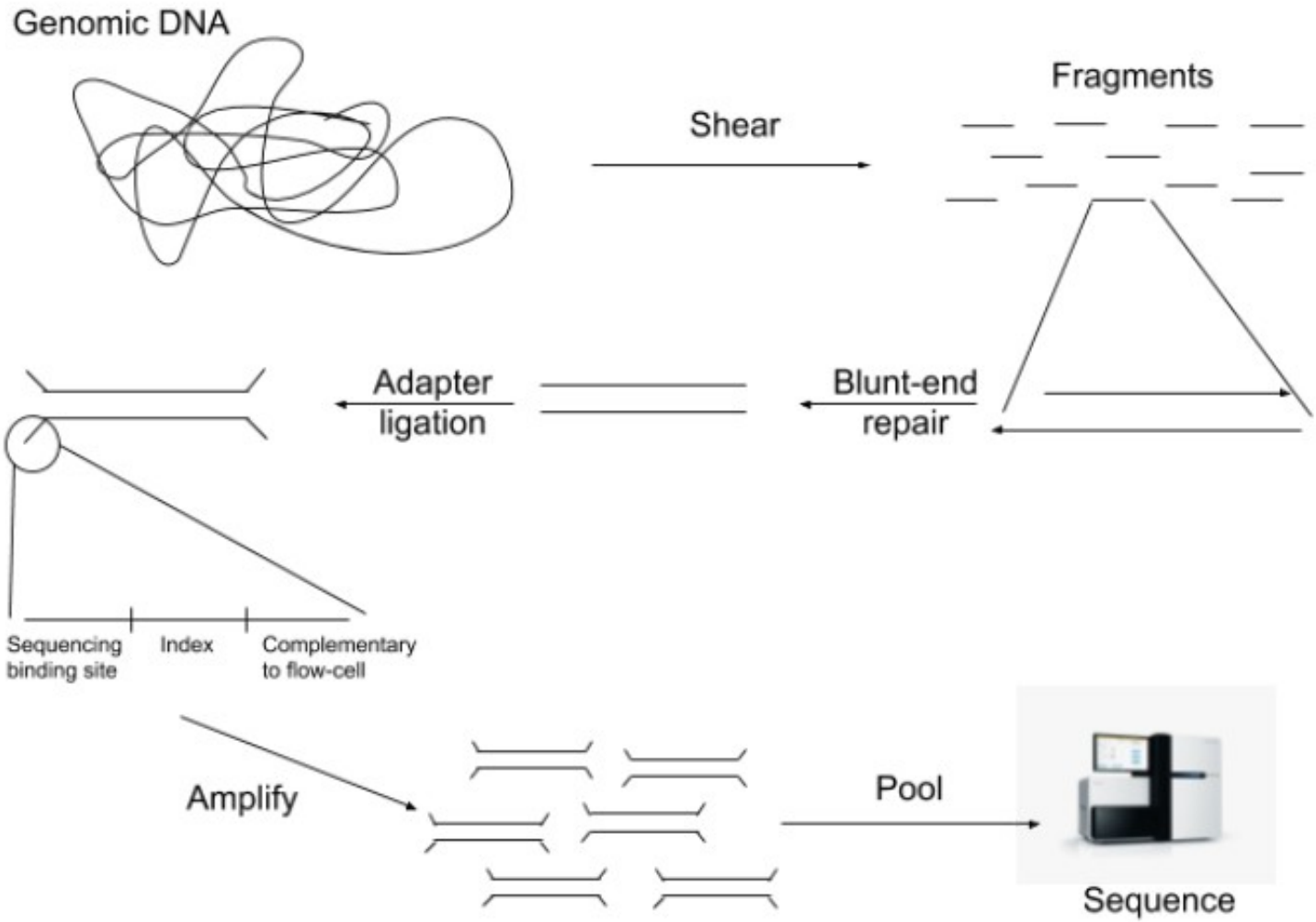
ang.mcgaughran@gmail.com

Library preparation

- Many methods; general steps:
  - Add adapters containing:
    - Barcodes (for multiplexing)
    - Sequencing primers
    - Amplification primers
    - Sequence for substrate attachment
  - Amplify fragments by universal PCR
  - Optionally pool barcoded libraries

ang.mcgaughran@gmail.com

Library preparation



ang.mcgaughran@gmail.com

Cluster generation

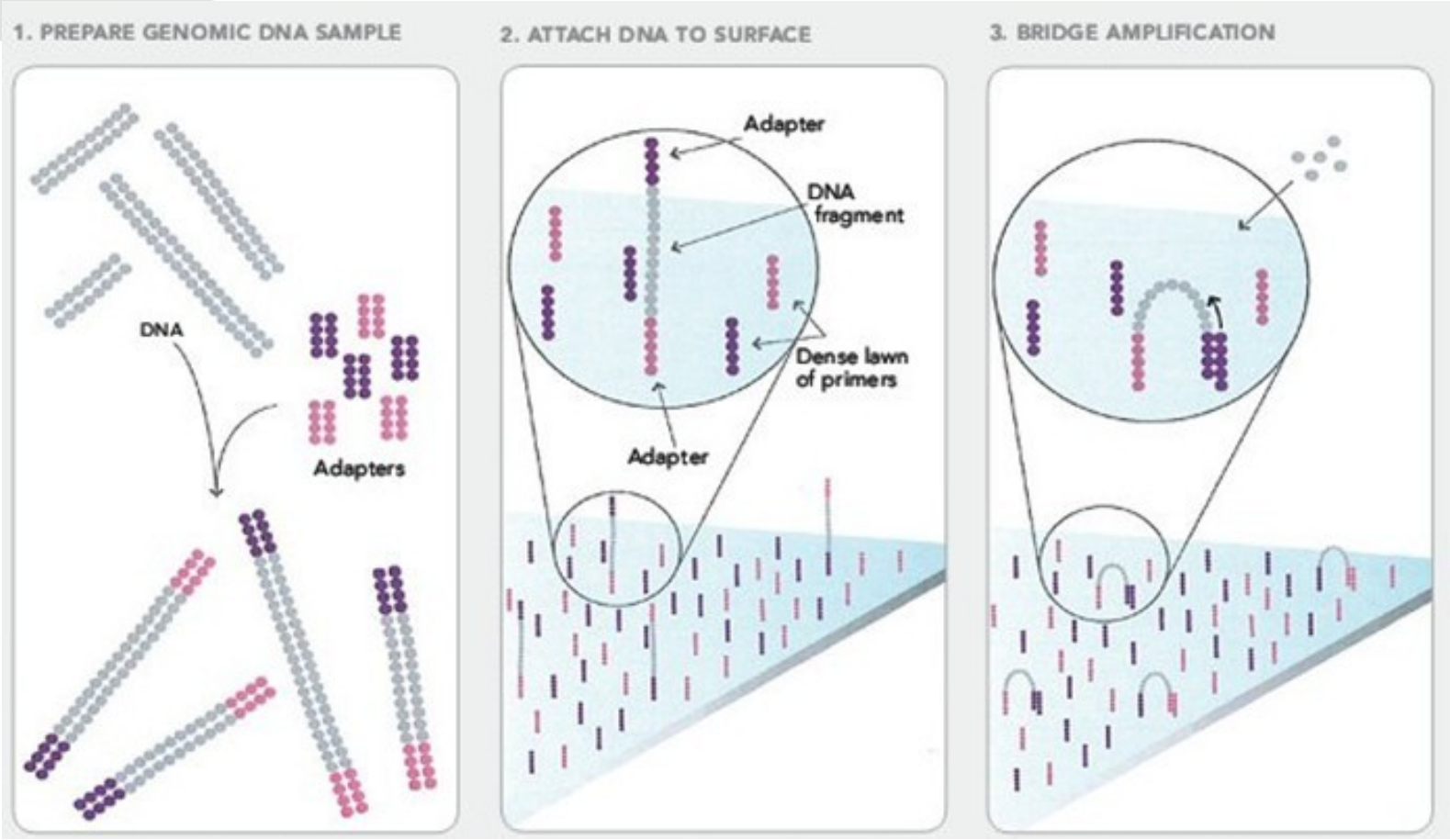


Image: www.illumina.com

ang.mcgaughran@gmail.com

Cluster generation

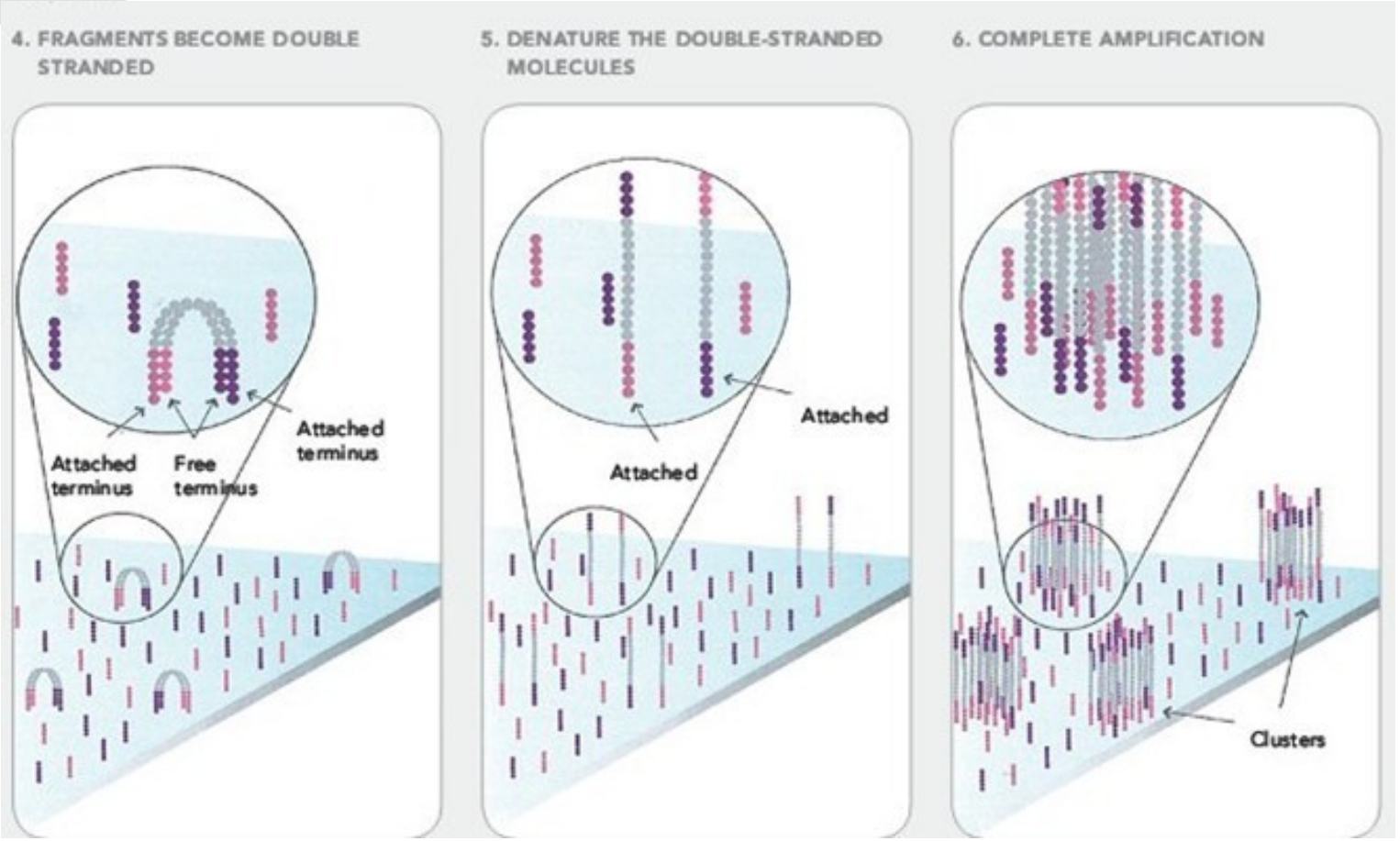


Image: www.illumina.com

ang.mcgaughran@gmail.com

Sequencing

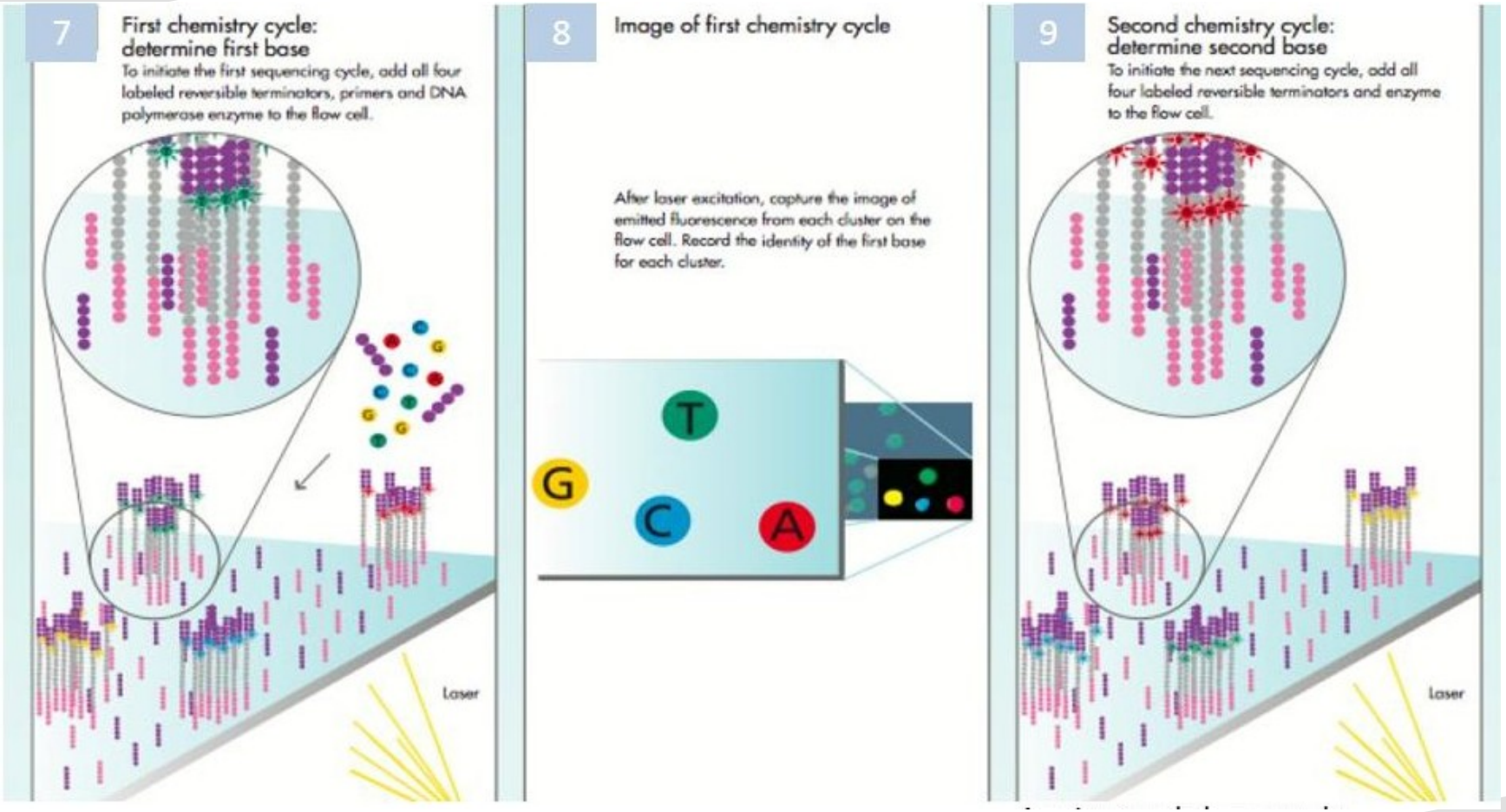


Image: www.illumina.com

ang.mcgaughran@gmail.com

(2) Illumina

- Video links

  https://www.youtube.com/watch?v=fCd6B5HRaZ8

  (0 – 3:08)

ang.mcgaughran@gmail.com

Strengths and limitations

Method	Pros	Cons	Applications
1st Gen (Sanger)	• Very accurate • Good for highly repetitive regions • Low technological barriers	• Very expensive ($/bp) • Limited read lengths (first 50 bp not great, starts deteriorating after about 750 bp, upper limit ~1000 bp)	• Resolving ambiguous/ repetitive regions • 5'/3' RACE • Plasmid sequencing
2nd Gen (Illumina short read)	• Quite accurate (0.1-1.0% error) • Cheap ($/bp)	• Unable to resolve repeat regions • Takes time for data processing	• Full genome re-sequencing • Reduced representation sequencing

ang.mcgaughran@gmail.com

Other next gen tech

- SOLiD
  - Ligation rather than polymerase-based
  - Redundant base sampling with error correction
  - Short reads
- Helicos
  - Like Illumina, but with single molecules

ang.mcgaughran@gmail.com

3rd Generation Sequencing

- Capable of sequencing SINGLE molecules
- No requirement for DNA amplification or 'sequencing by synthesis'
- 'Long read' technology (50-100,000+ bp)
- PacBio, Nanopore

ang.mcgaughran@gmail.com

3rd Generation Sequencing

- PacBio
  - Images all wells at the same time in real time with digital video
  - Interprets bases by light signal
  - Very large instrument
- Nanopore
  - Uses protein nanopores in a synthetic membrane to thread DNA through
  - Very small instrument!

ang.mcgaughran@gmail.com

Nanopore

Image: Oxford Nanopore Technologies

Image: NextBigFuture.com

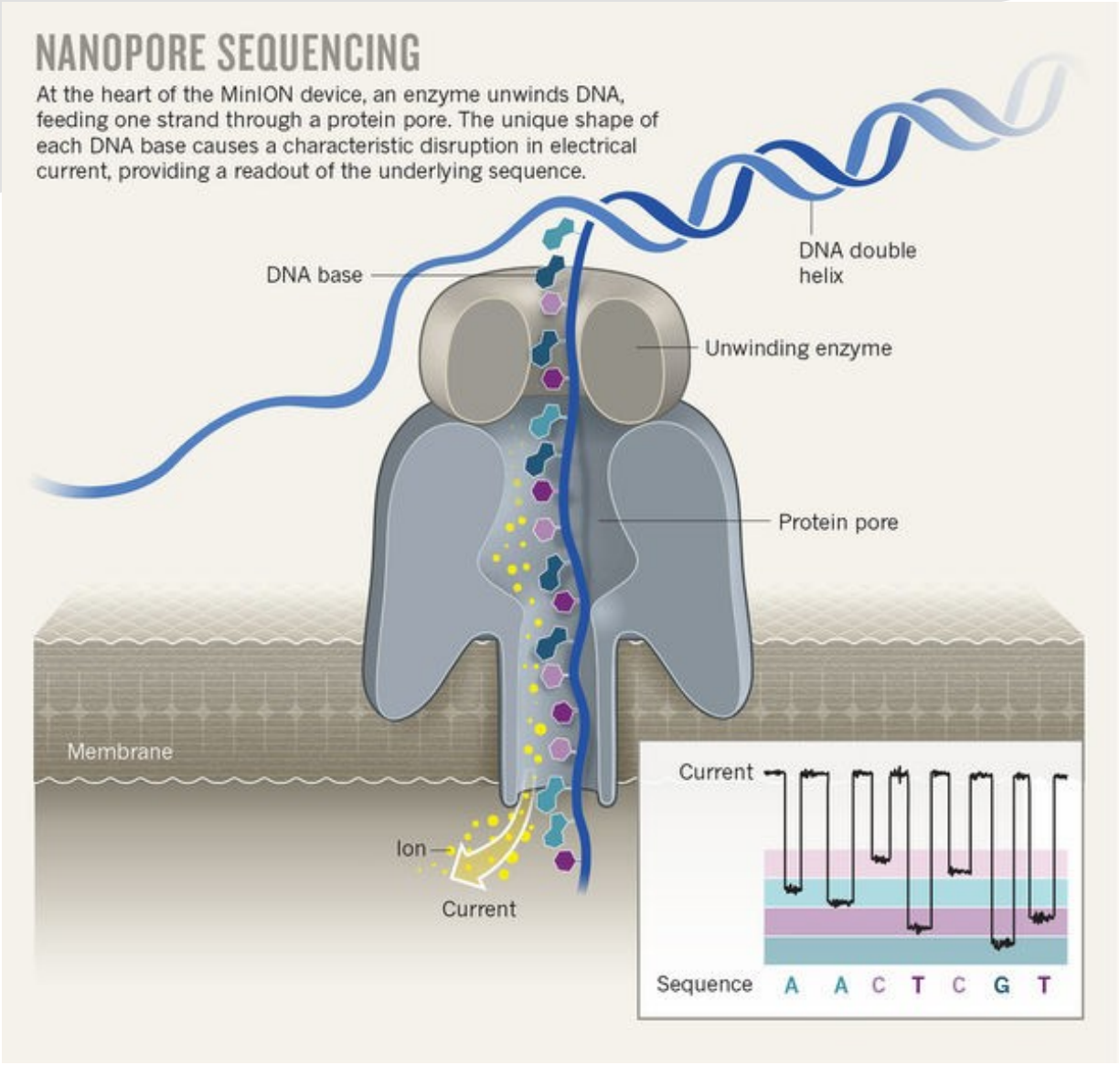


Image: Nik Spencer/Nature; https://www.nature.com/articles/550285a

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

Nanopore

Image: Sarah Johnson; https://www.nature.com/articles/550285a

Image: Oxford Nanopore Technologies

ang.mcgaughran@gmail.com

Strengths and limitations

Method	Pros	Cons	Applications
1st Gen (Sanger)	• Very accurate • Good for highly repetitive regions • Low technological barriers	• Very expensive ($/bp) • Limited read lengths (first 50 bp not great, starts deteriorating after about 750 bp, upper limit ~1000 bp)	• Resolving ambiguous/ repetitive regions • 5'/3' RACE • Plasmid sequencing
2nd Gen (Illumina short read)	• Quite accurate (0.1-1.0 % error) • Cheap ($/bp)	• Unable to resolve repeat regions • Takes time for data processing	• Full genome re-sequencing • Reduced representation sequencing
3rd Gen (PacBio, Nanopore long read)	• Far easier to assemble • Can span repeat regions • Desktop technology (minION)	• Relatively expensive • Error prone (up to 10-15% error rates)	• Genome assembly (for now), often used in concert with short-read methods

ang.mcgaughran@gmail.com

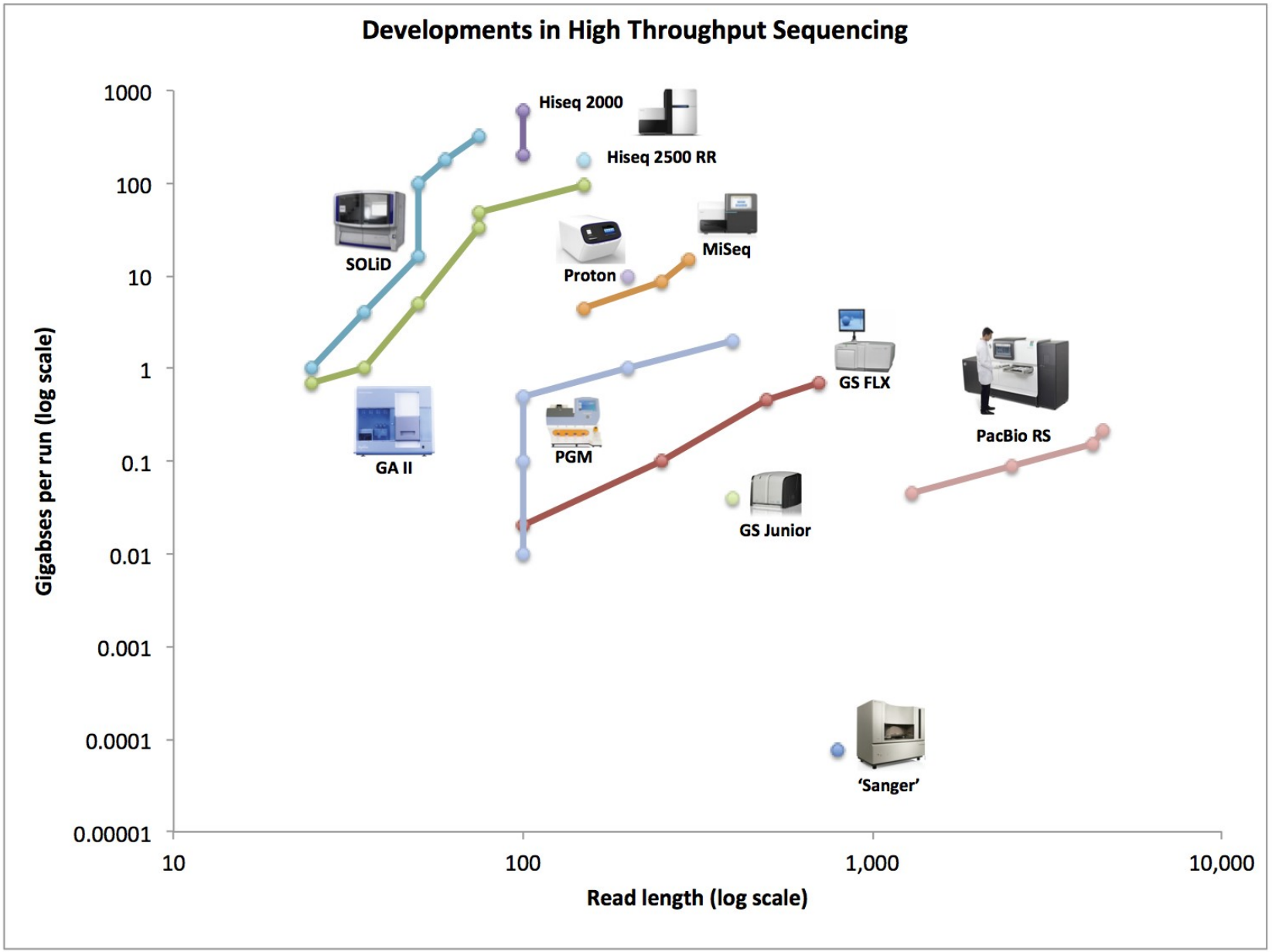


Image: https://figshare.com/articles/developments_in_NGS/100940

ang.mcgaughran@gmail.com

Experimental design

- First consideration:
  - What is your question?
- Second:
  - How much data do I need to answer it?
    - Sensitivity (e.g., no. of false negatives)
    - Specificity (e.g., no. of false positives)
    - Cost

ang.mcgaughran@gmail.com

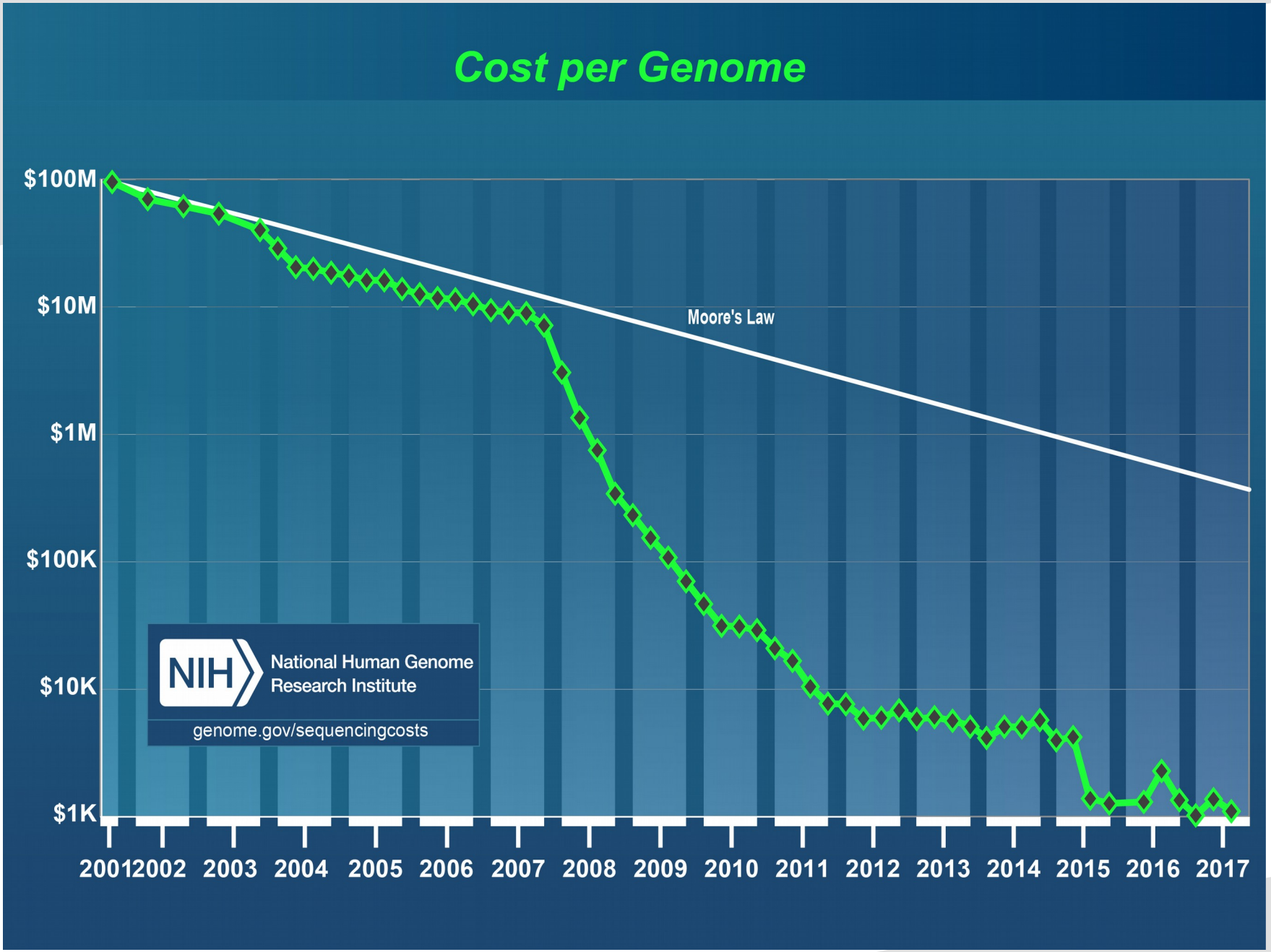


Image:https://www.genome.gov/27541954/dna-sequencing-costs-data/

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

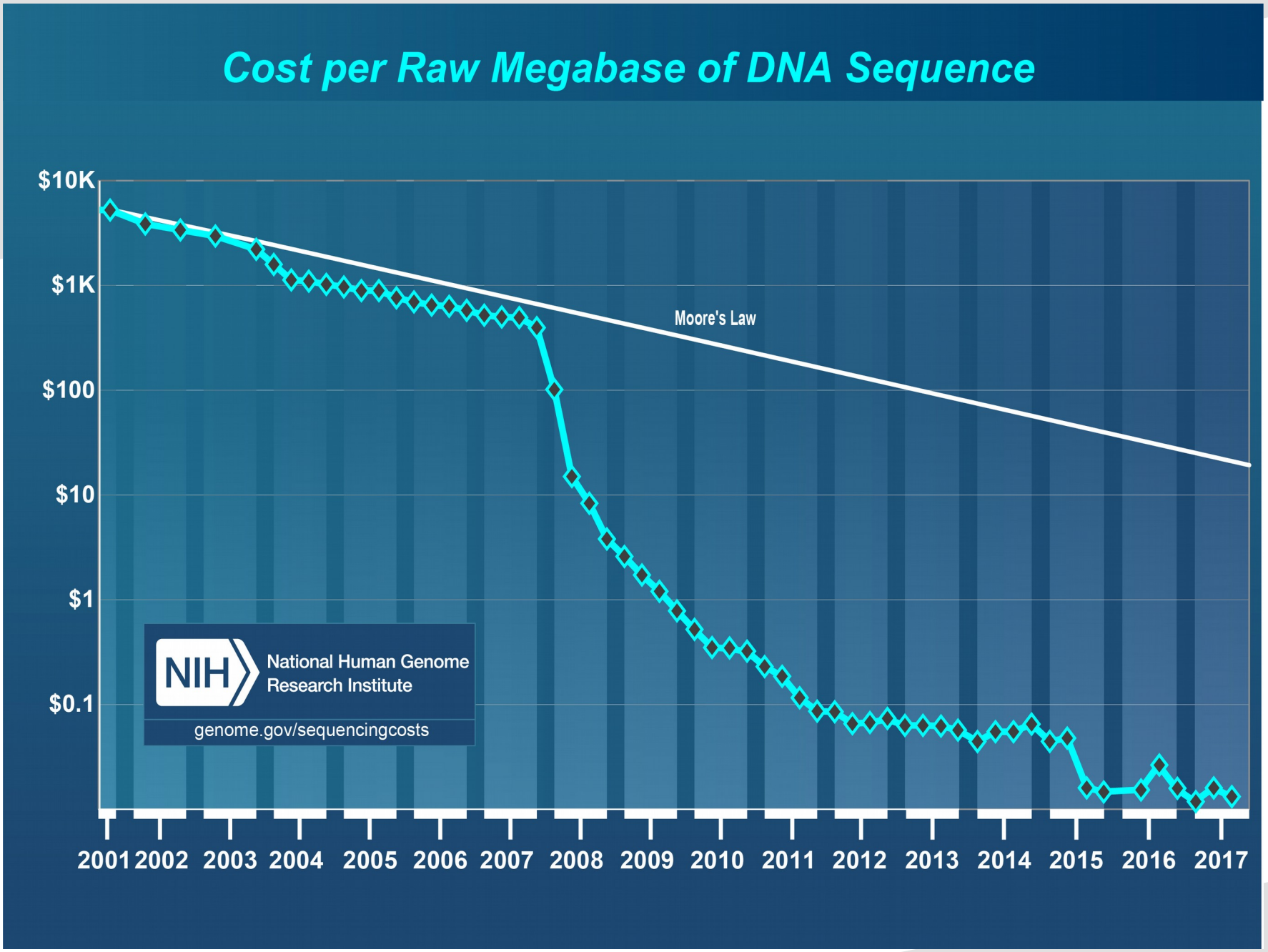


Image: https://www.genome.gov/27541954/dna-sequencing-costs-data/

Thursday, November 21, 2019

ang.mcgaughran@gmail.com

Experimental design

- Second:
  - How much data do I need to answer it?
  - Depending on the question, this may relate to sequencing coverage for a single individual or to number of samples from a population/s
  - Or, it may relate to conditions, tissues, no. of replicates, power to detect expression differences

ang.mcgaughran@gmail.com

Experimental design

- Number of samples
  - Pooling with barcoding
    - Unique tags identify samples, which can be combined to run on one lane
    - Reduces cost
    - Analysis is identical to unpooled data
    - But…
      - Some small throughput loss due to barcode fails
      - Data mis-assignment from bad barcode reads
      - Increased per-sample cost for library construction

ang.mcgaughran@gmail.com

Depth

Type of Experiment	Coverage required (for accuracy)
Haploid SNPs	> 10x
Diploid SNPs	> 30x
Rare mutations	> 50x
Genome assembly	50 – 100x short read; 20 – 50x long read

ang.mcgaughran@gmail.com

Experimental design

- Ability to call heterozygotes
- High proportion of rare/unique variants (harder to confirm)
- Genomic resources:
  - Is there a reference genome or transcriptome?
  - Is it complete? Accurate? Representative?

ang.mcgaughran@gmail.com

Experimental design

- Reference:
  - Repeat content of the genome?
  - GC content
    - Greater average coverage will be required to assemble through extreme GC regions
- Highly contiguous assemblies

ang.mcgaughran@gmail.com

Read data specs

- Number of reads:
  - Depends on the level of completeness and accuracy you require
  - Ability to identify and quantify transcripts (RNASeq)
- Length of reads:
  - Generally, the longer the reads are the better
  - But: longer poor-quality reads may not be as useful as shorter high-quality

ang.mcgaughran@gmail.com

Read data specs

- Complexity:
  - The number of distinct fragments in your library
  - Low complexity via amplification of the same initial fragment
- Single or paired end:
  - SE
  - PE

ang.mcgaughran@gmail.com

SE or PE

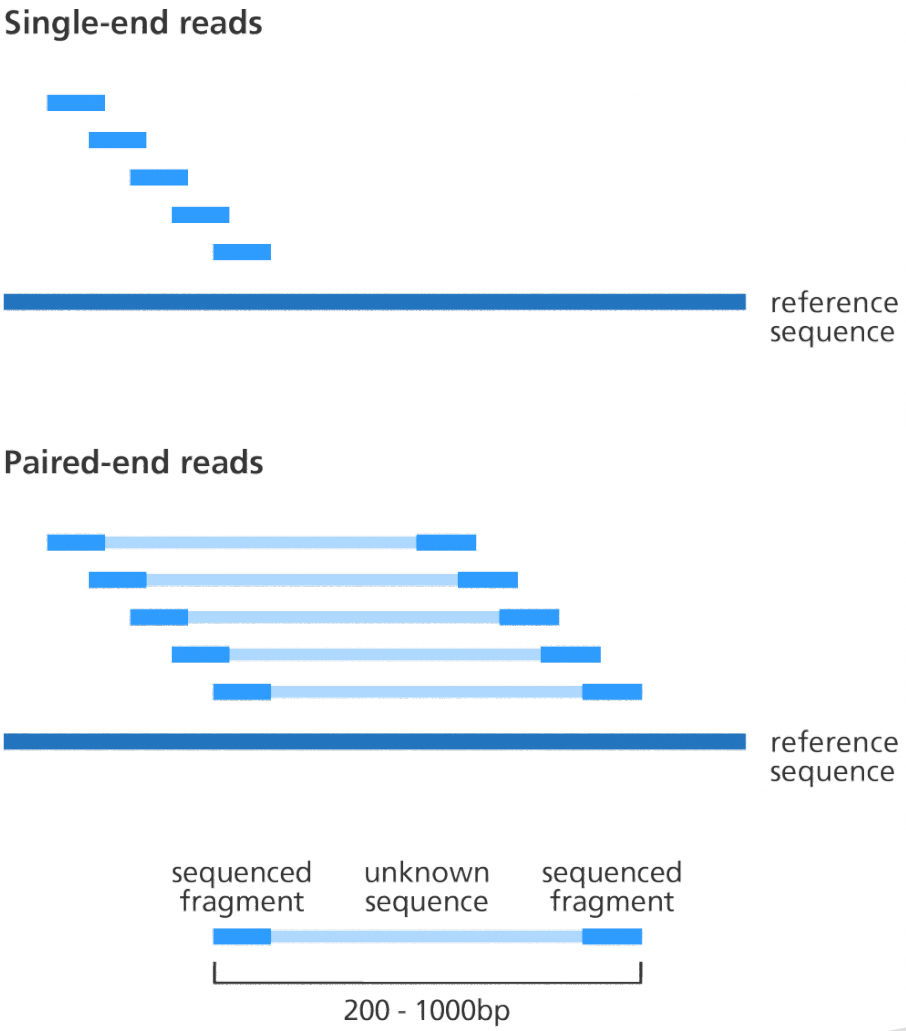


Image: Pinterest

angela.mcgaughran@anu.edu.au

SE or PE

300 bp PE with a
700 bp fragment size:
R1 = 300 bp
R2 = 300 bp
Insert = 100 bp

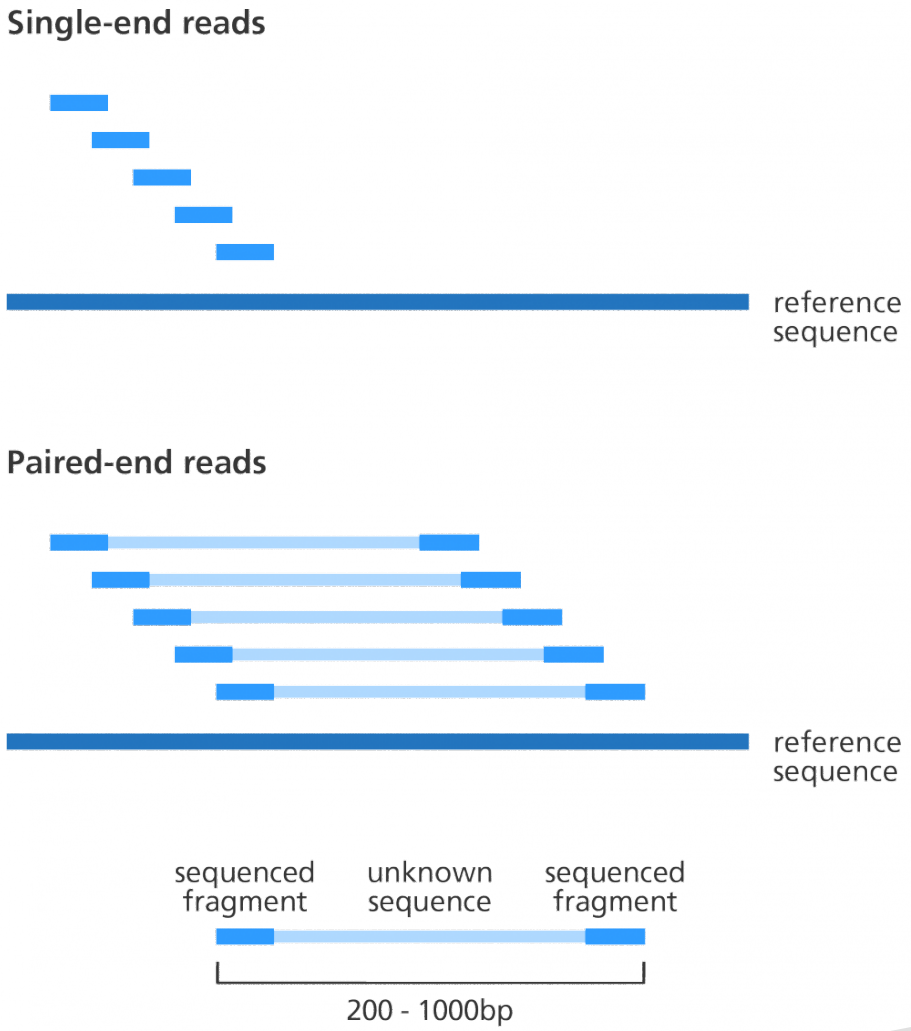


Image: Pinterest

angela.mcgaughran@anu.edu.au

Technology

- Read lengths
- Error rates

ang.mcgaughran@gmail.com

Technology

- Re-sequencing
- Genome assembly
- RNA-Seq
- Targeted sequencing (e.g., exon capture)
- RAD-Seq (GBS)
- Metagenomics

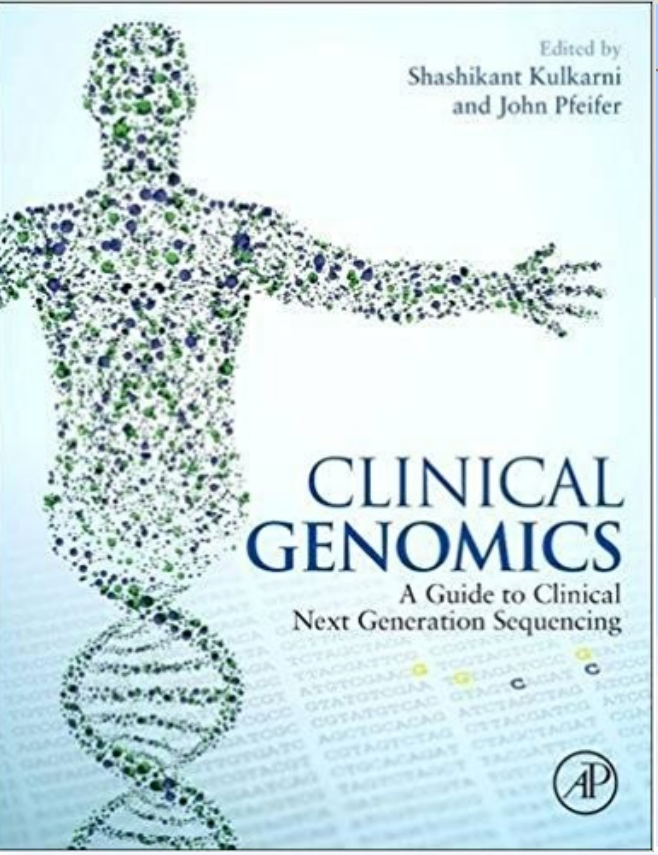
ang.mcgaughran@gmail.com

Sequencing applications


Image: Portugal Startups

- SNP discovery and genotyping
- Population sequencing
- Structural variant discovery and genotyping
- Comparative genomics
- Generating a reference genome
- Discovery of novel genes or transcripts
- Variability of isoform expression across conditions
- Characterise species present in an environment

ang.mcgaughran@gmail.com



Images: http://blog.illumina.com/agricultural-genomics

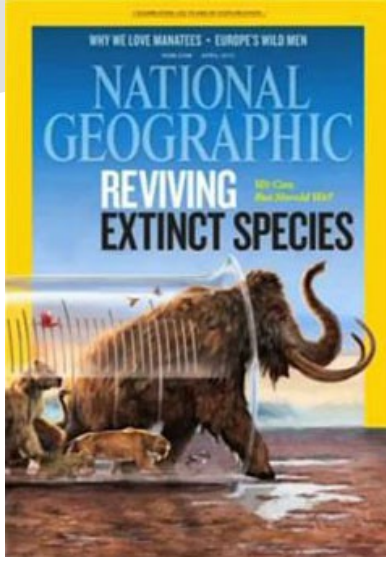


Image: CEN4GEN wildlife and conservation genomic services

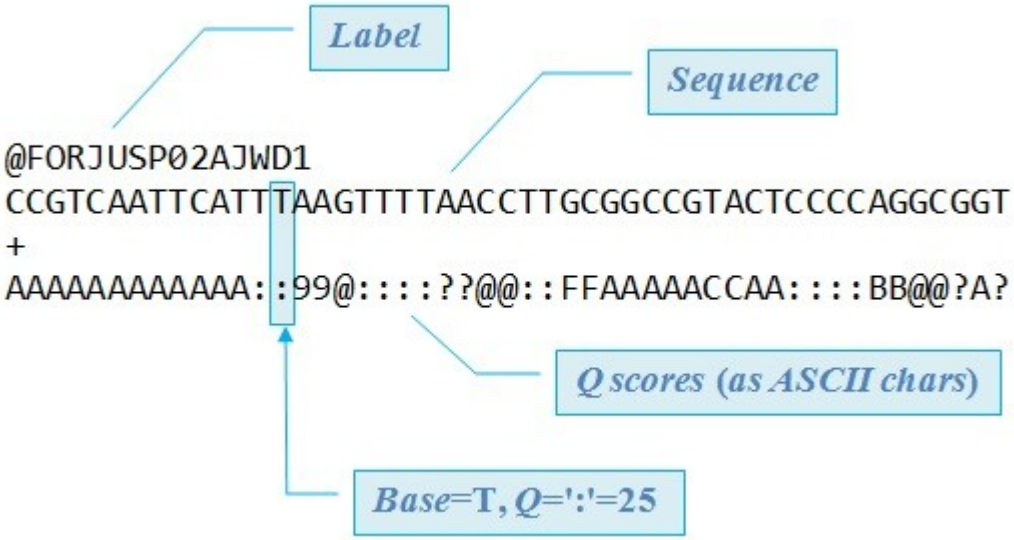
Bioinformatics

- What do we get off a sequencer?
- Fastq files:
  - A text-based format for storing biological sequence data and it's corresponding QUALITY scores
  - Uses ASCII characters to encode quality

angela.mcgaughran@anu.edu.au

Fastq files

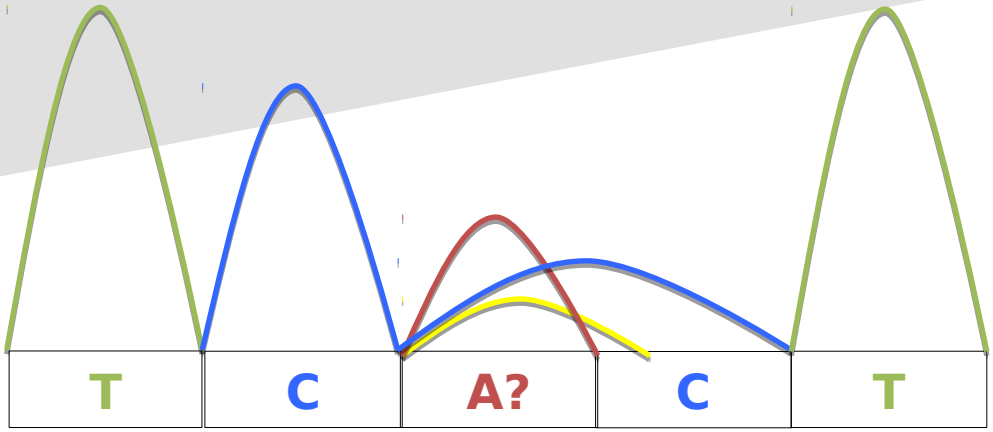
Image: https://www.drive5.com/usearch/manual/fastq_files.html



- Four lines per sequence:
  - Line 1 begins with an '@' character and a identifier
  - Line 2 is the raw sequence letters
  - Line 3 is a '+' character and sometimes the sequence identifier
  - Line 4 encodes the quality values for the sequence in Line 2

angela.mcgaughran@anu.edu.au

Quality scores



- Phred scores
- These are based on the *probability* of a base call being correct:

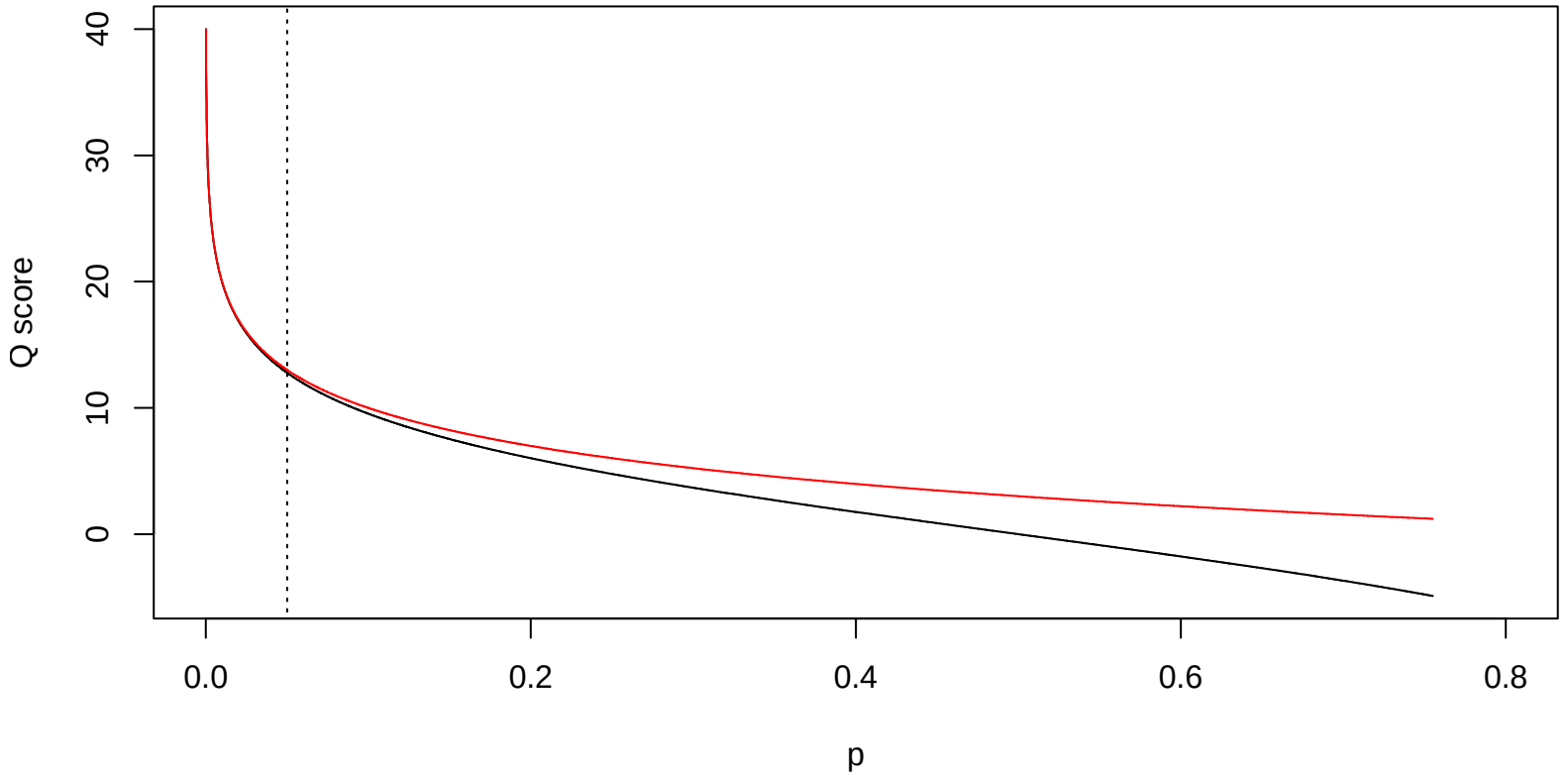
$$Q = -10 \log_{10} p$$

- Different pipelines use different characters for encoding:
  - http://en.wikipedia.org/wiki/FASTQ_format

angela.mcgaughran@anu.edu.au

Phred scores

Image: By Sealox - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11606715



E.g., $p = 0.05$ is $Q \approx 13$

angela.mcgaughran@anu.edu.au

Phred scores

- Q10 = 1 in 10 chance of incorrect base call (90%)
- Q20 = 1 in 100 chance of incorrect base call (99%)
- Q30 = 1 in 1,000 chance of incorrect base call (99.9%)
- Q40 = 1 in 10,000 chance of incorrect base call (99.99%)

- Expect to see scores between Q20 and Q40, and usually reads with Q < 20 are excluded

angela.mcgaughran@anu.edu.au

Fastq QC

- Usually involves:
  - Identifying over-represented sequences
  - Removing or trimming reads containing adaptor sequences
  - Removing or trimming reads containing low quality base calls
  - Calculating the number of reads before and after QC

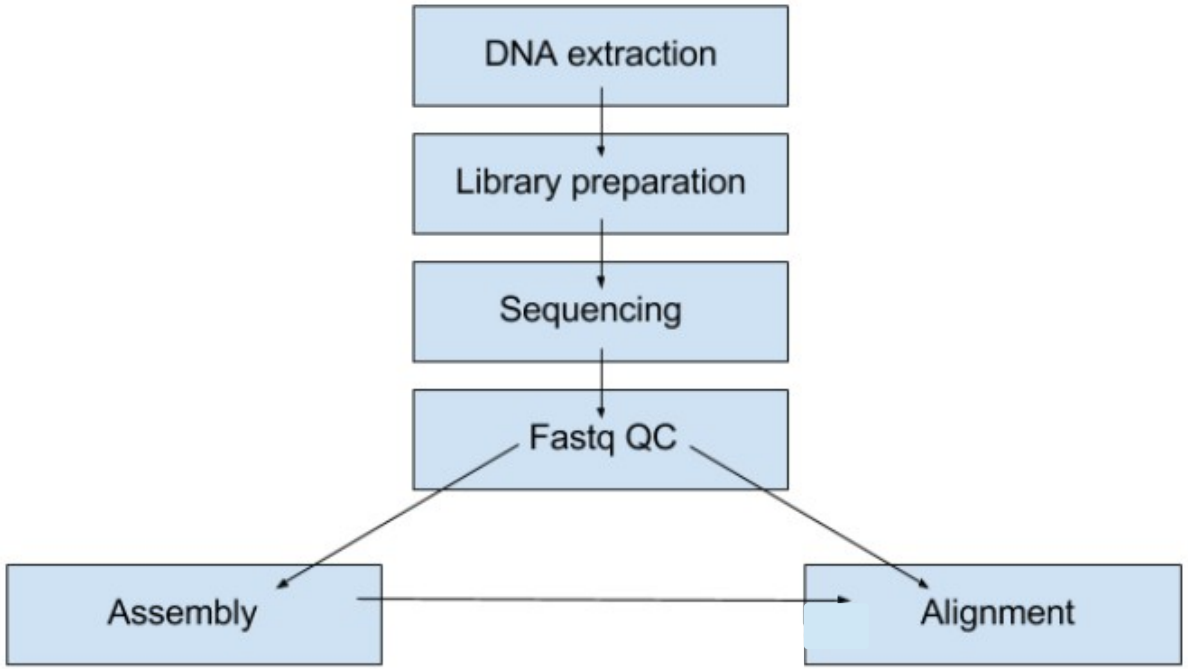
angela.mcgaughran@anu.edu.au

Fastq QC

- Benefits:
  - Reduces CPU time for next steps (assembly and alignment)
  - Reduces data storage requirements
  - Reduces potential for bias in later steps (variant calling, *de novo* assembly)
- But, not always needed (some downstream software can take quality into account)

angela.mcgaughran@anu.edu.au

Pipeline summary



angela.mcgaughran@anu.edu.au

Assembly

- So, you have a gazillion sequence reads – what next?
- Reconstruct the original DNA sequences
- *De novo*
  - e.g., when it's a new organism
- Reference-based
  - Can be done with a **reference genome**, in which case the reference assists the assembly

angela.mcgaughran@anu.edu.au

Assembly

1. Fragment DNA and sequence

Image: modified from https://www.gatc-biotech.com/en/expertise/genomics/de-novo-genome-analysis.html

angela.mcgaughran@anu.edu.au

Assembly

1. Fragment DNA and sequence

2. Find overlaps between reads

...ATAGACCTAGAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

Thursday, April 4, 2019
angela.mcgaughran@anu.edu.au

Image: modified from https://www.gatc-biotech.com/en/expertise/genomics/de-novo-genome-analysis.html

Assembly

- Align each read against every other read
- Build an overlap graph
  - de Bruijn graphs (construct a short text that contains all *k*-mers exactly once) → for Illumina, 10X data
  - String graphs (predict what character is likely to follow a given *k*-mer) → for PacBio/Nanopore data

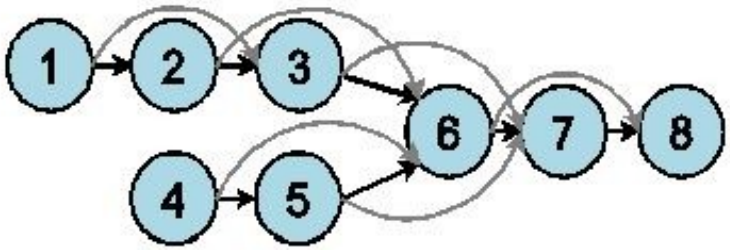
angela.mcgaughran@anu.edu.au

Assembly Method

Sequencing reads:

```
1 A C C T G A T C
2     C T G A T C A A
3       T G A T C A A T
4   A G C G A T C A
5       C G A T C A A T
6         G A T C A A T G
7             T C A A T G T G
8               C A A T G T G A
```

1. Overlap graph



ACCTG→CCTGA→CTGAT→TGATC

AGCGA→GCGAT→CGATC

GATCA→ATCAA→TCAAT→CAATG→AATGT→ATGTG→TGTGA

2. de Bruijn graph

3. String graph

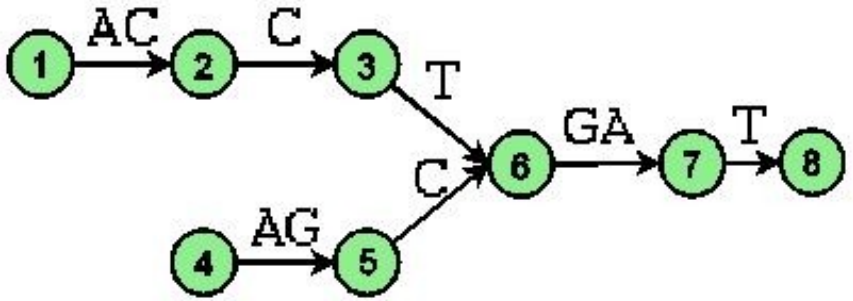


Image: http://slideplayer.com/slide/4240769/

Assembly

- Align each read against every other read
- Build an overlap graph
  - de Bruijn graphs (construct a short text that contains all *k*-mers exactly once)
  - String graphs (predict what character is likely to follow a given *k*-mer)
- Find distinct paths through the graph
- Calculate the consensus sequence for each path → CONTIGS

angela.mcgaughran@anu.edu.au

Assembly

1. Fragment DNA and sequence

2. Find overlaps between reads

...ATAGACCTAGAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

Image: modified from https://www.gatc-biotech.com/en/expertise/genomics/de-novo-genome-analysis.html

angela.mcgaughran@anu.edu.au

Assembly

1. Fragment DNA and sequence

2. Find overlaps between reads

...ATAGACCTAGAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds

Image: modified from https://www.gatc-biotech.com/en/expertise/genomics/de-novo-genome-analysis.html

angela.mcgaughran@anu.edu.au

Assembly

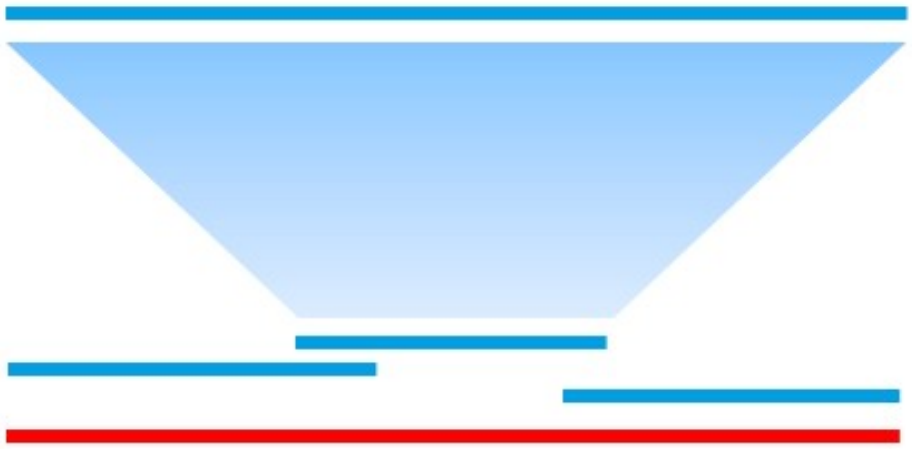
1. Fragment DNA and sequence

2. Find overlaps between reads

...ATAGACCTAGAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds



5. Finished genome hypothesis

Image: modified from https://www.gatc-biotech.com/en/expertise/genomics/de-novo-genome-analysis.html

angela.mcgaughran@anu.edu.au

Assembly

- Many different software programs available for this process, using different algorithms
  - Trinity, SPAdes, AbySS, SOAPdenovo, Canu...
- Challenges
  - Depends on sequencing technology
  - False positives via errors in reads, which make the string/de Bruijn graph more complicated
  - Repeat regions are difficult (e.g., repeats >> *k*-mer)
  - Gaps if *k*-mer size is too big and/or not all bases in the genome get sampled (e.g., low or non-uniform coverage)

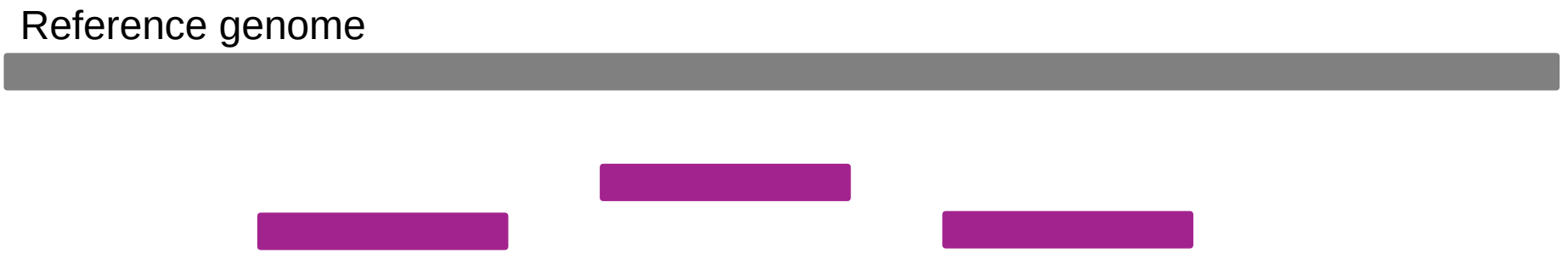
angela.mcgaughran@anu.edu.au

Alignment

- Where do the reads fit on a reference genome?
- How do our sequenced samples compare to the reference sample?

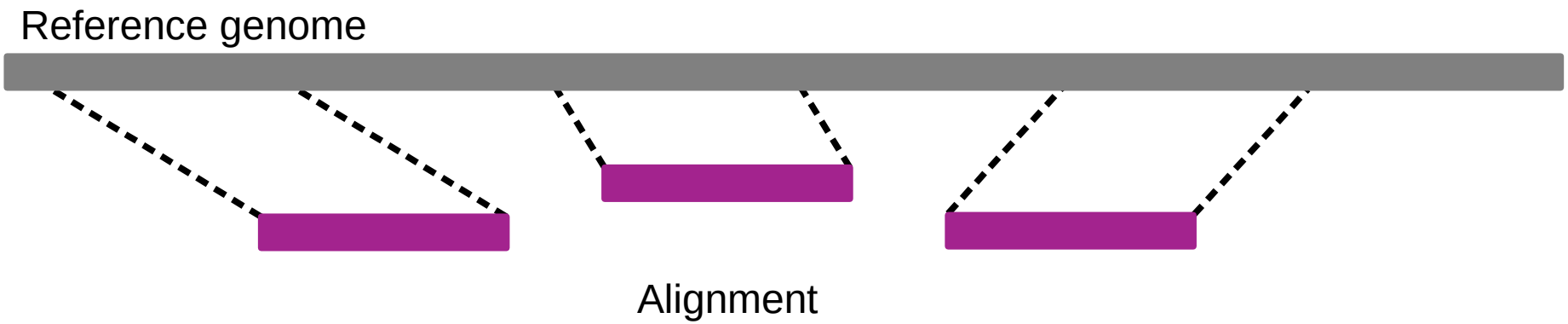
angela.mcgaughran@anu.edu.au

Alignment



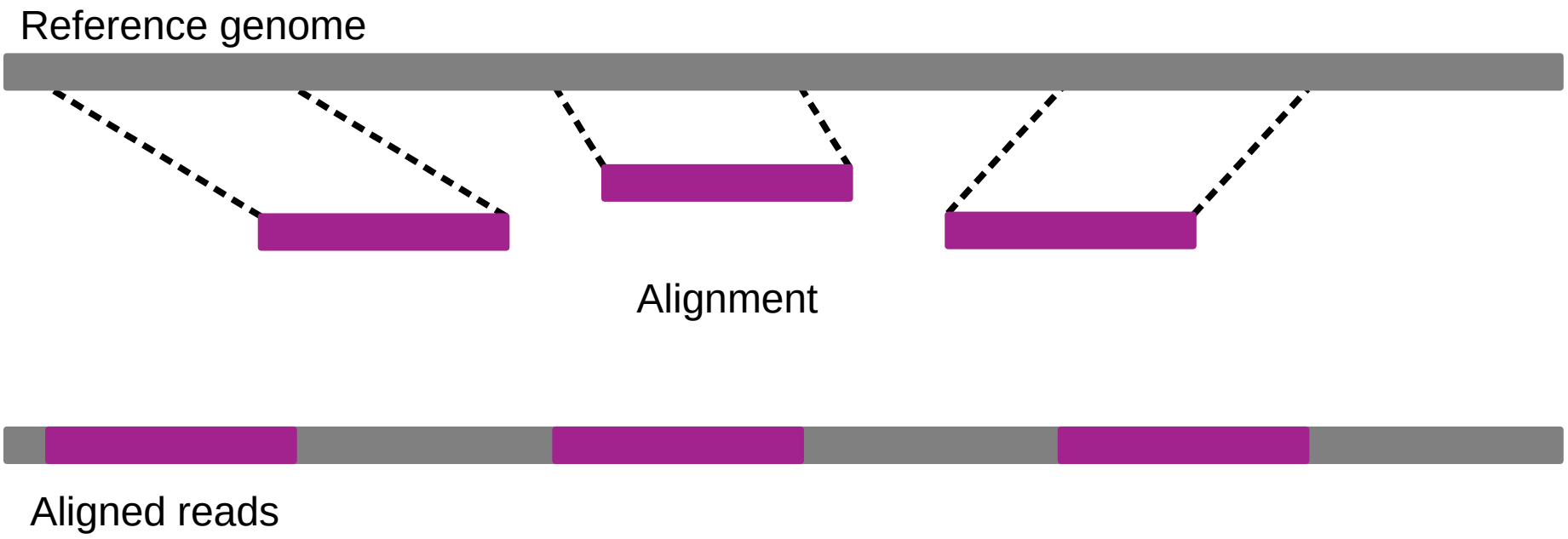
angela.mcgaughran@anu.edu.au

Alignment



angela.mcgaughran@anu.edu.au

Alignment



angela.mcgaughran@anu.edu.au

Alignment

- In practice, we go from the FASTQ file to a SAM/BAM file which lists the reads and where/if they map to the reference genome
  - SAM is a plain text file of tab-separated columns; inefficient to read and store
  - BAM is a compressed version of SAM, can be indexed and sorted

angela.mcgaughran@anu.edu.au

Alignment

- Read name [1], chromosome [3], position [4], mapping quality [5], sequence [10], quality [11]

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

Image: https://insidedna.me/tutorials/view/samtools-commands-tutorial-working-sam-bam-files

Alignment

- Many different software programs available for this process, using different algorithms
  - BWA, Botwite, BFAST, SHRiMP, MAQ…
  - https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

angela.mcgaughran@anu.edu.au

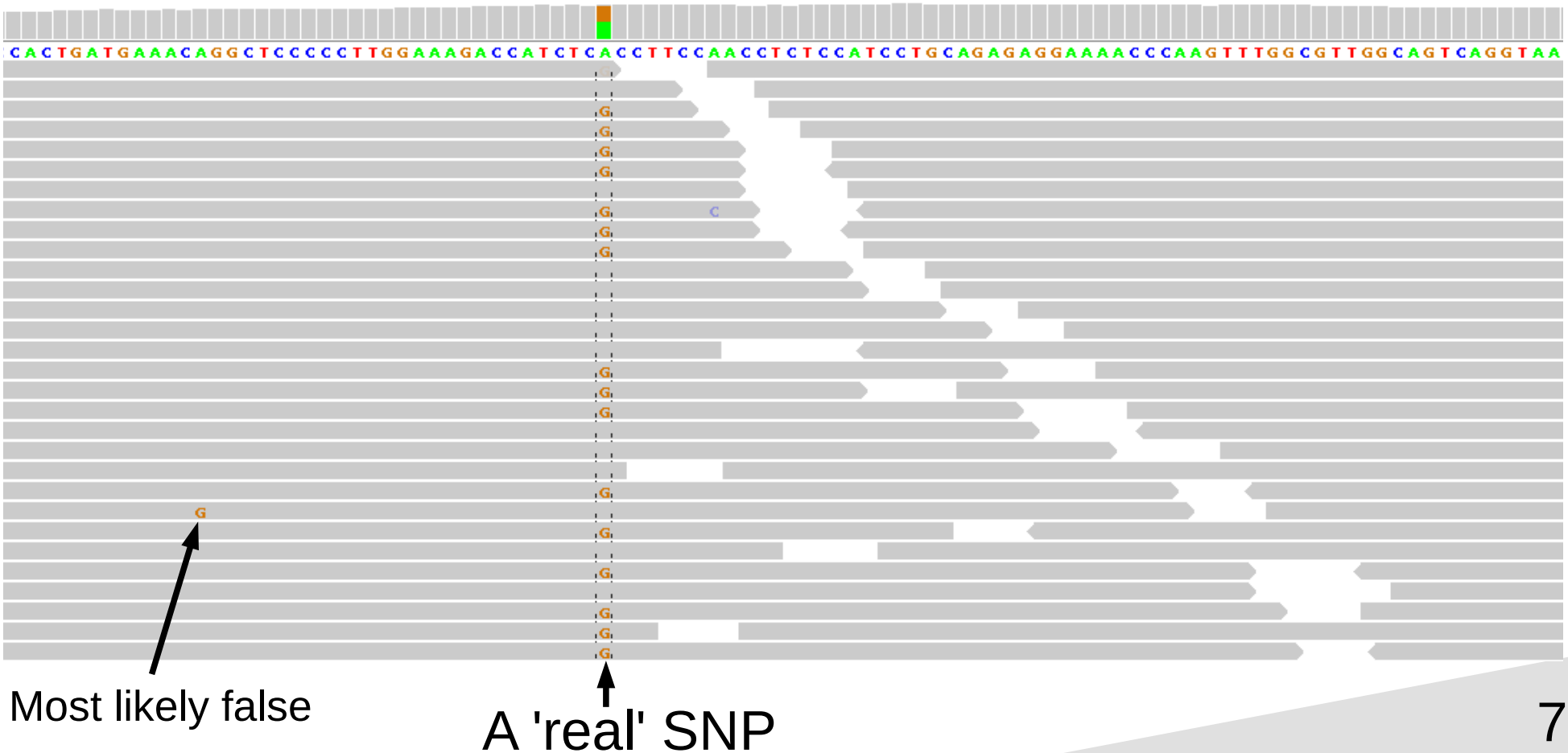
Alignment

- Remember that 2nd generation error rates are around 0.1-1%, so a 300 bp read is, on average, likely to contain 2-3 errors
  - We can often identify sequencing error as 'the odd one out' (because most reads will agree with each other):

angela.mcgaughran@anu.edu.au

Alignment

Image: http://www.ikmb.uni-kiel.de/pibase/pibase_filtering.html



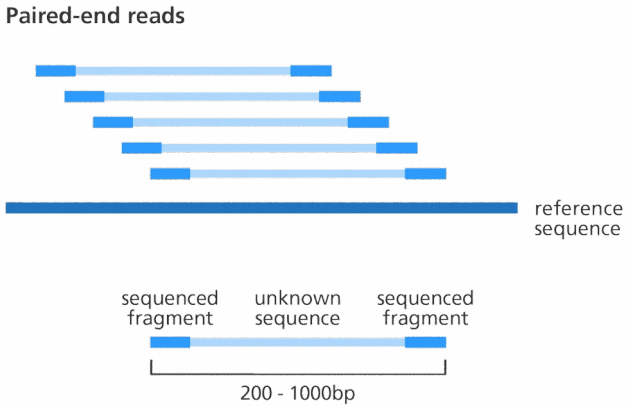
Thursday, April 4, 2019
angela.mcgaughran@anu.edu.au

Alignment

- Remember that 2nd generation error rates are around 0.1-1%, so a 300bp read is, on average, likely to contain 2-3 errors
  - We can often identify sequencing error (because most reads will agree with each other)
  - Other errors can be more difficult to identify
- The challenge is to distinguish true SNPs from false (NB: the human genome contains > 10 million SNPs)
  - Can often be removed bioinformatically

angela.mcgaughran@anu.edu.au

Alignment



- Read length matters for alignability
- Paired ends help
  - Aligning one end localises the other
  - More sensitive alignments
  - Can use to find highly variant regions and small indels
  - Necessary (for very long reads) for structural variant discovery and genotyping

angela.mcgaughran@anu.edu.au

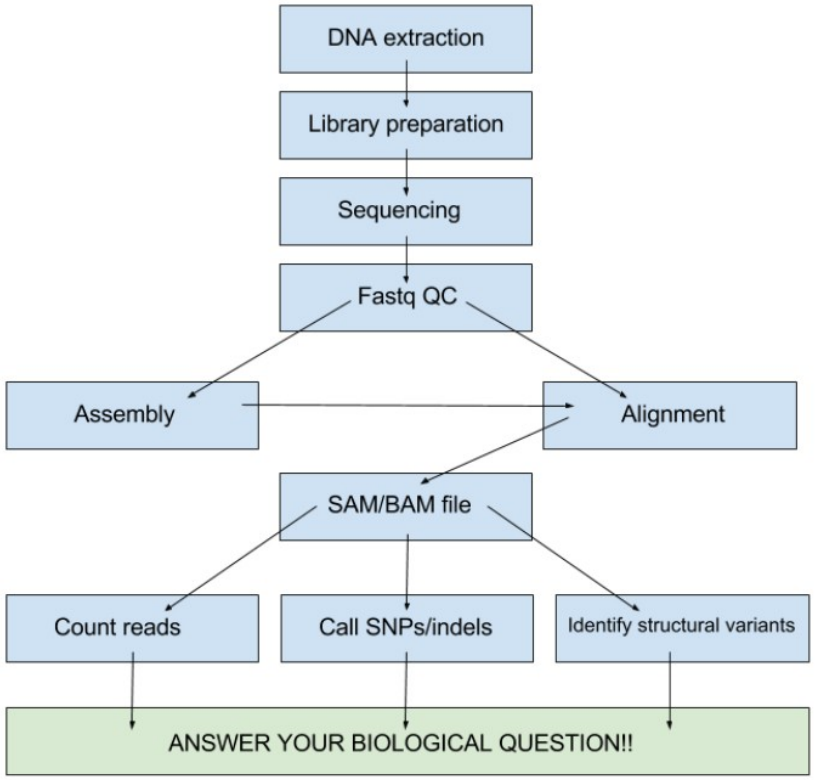
Alignment

- Reference genome considerations
  - Complete?
  - Accurate?
  - Representative?
- Challenges:
  - Sequences won’t align if they’re absent from the reference...or, if they’re too divergent
  - Not all sequences are useful (duplicates, repeats, gene families)

ang.mcgaughran@gmail.com

Alignment

- From BAMs, we can go on to do our genomic analyses:



angela.mcgaughran@anu.edu.au

Take home messages

- Sequencing technology has come a long way!
  - More and more data for less and less $$
- Different platforms have different pros and cons, including error profiles
- Important to consider the biological question when selecting your method

ang.mcgaughran@gmail.com

Take home messages

- Bioinformatics is the biggest challenge, most time-consuming
  - Approaches depend on the biological question
- This is an EVOLVING field!

angela.mcgaughran@anu.edu.au

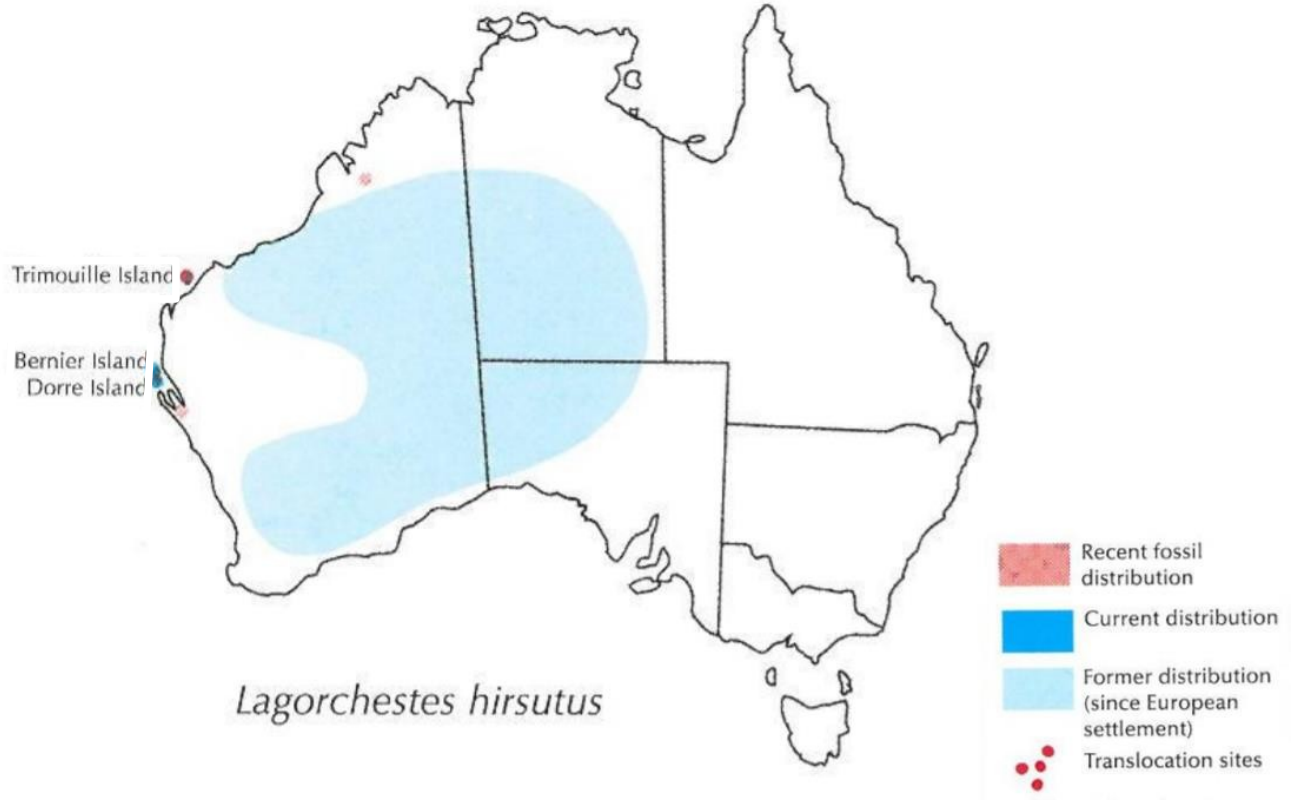
Practical demonstration

- Exploring the effects of small population size and limited migration on population structure
  - Using R
  - Using VCF files
  - Making conservation decisions

ang.mcgaughran@gmail.com

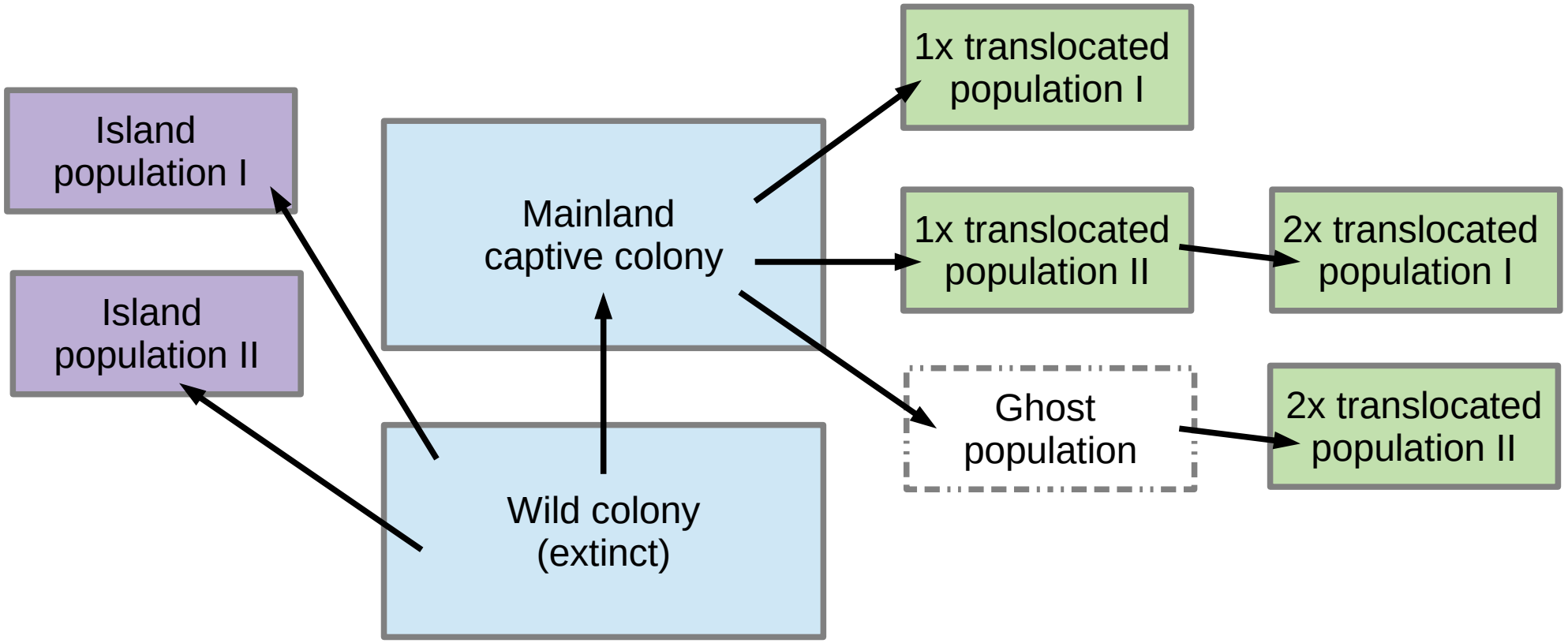
Rufous hare-wallaby

- Mainland and island populations



angela.mcgaughran@anu.edu.au

History



angela.mcgaughran@anu.edu.au

Genetic consequences

- Founder events and translocations:
  - Subset of original diversity

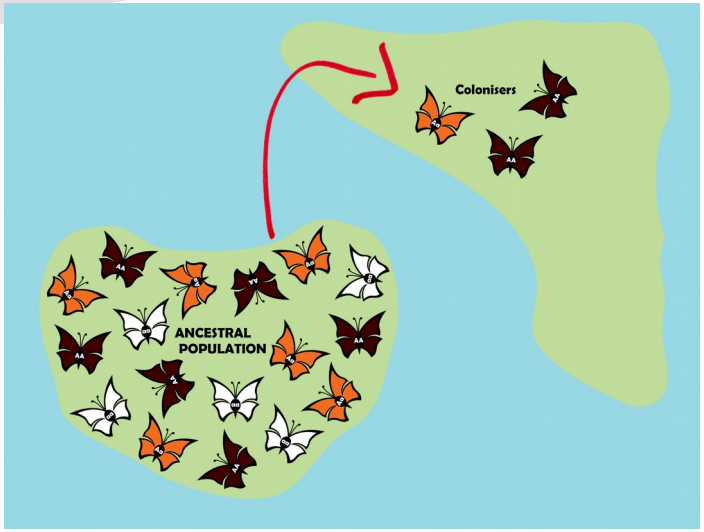


Image: https://www.pathwayz.org/Tree/Plain/FOUNDER+EFFECT

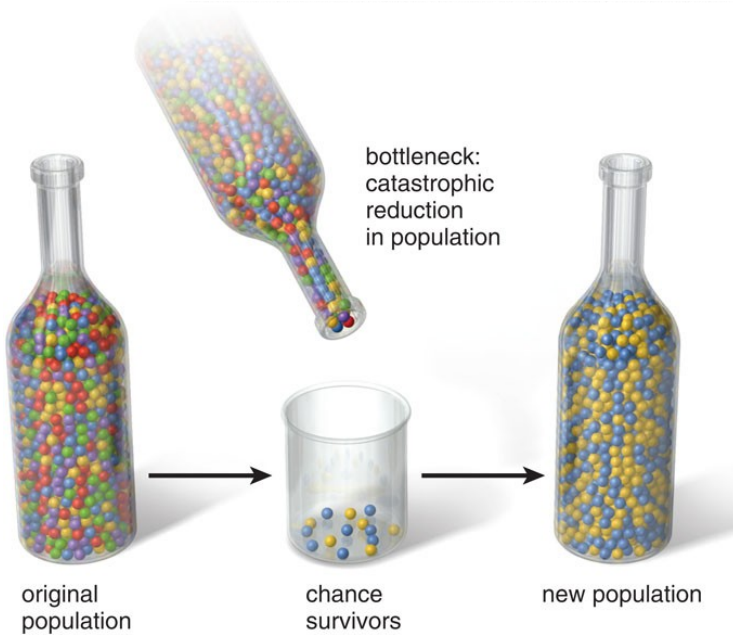


Image: https://socratic.org/questions/what-causes-genetic-drift

angela.mcgaughran@anu.edu.au

Genetic consequences

- Founder events and translocations:
  - Subset of original diversity
  - Small populations
    - Exacerbated drift

Image: http://scienceteachingarticles.blogspot.com.au/2015/02/genetic-drift.html

angela.mcgaughran@anu.edu.au

Genetic consequences

- Founder events and translocations:
  - Subset of original diversity
  - Small populations
    - Exacerbated drift

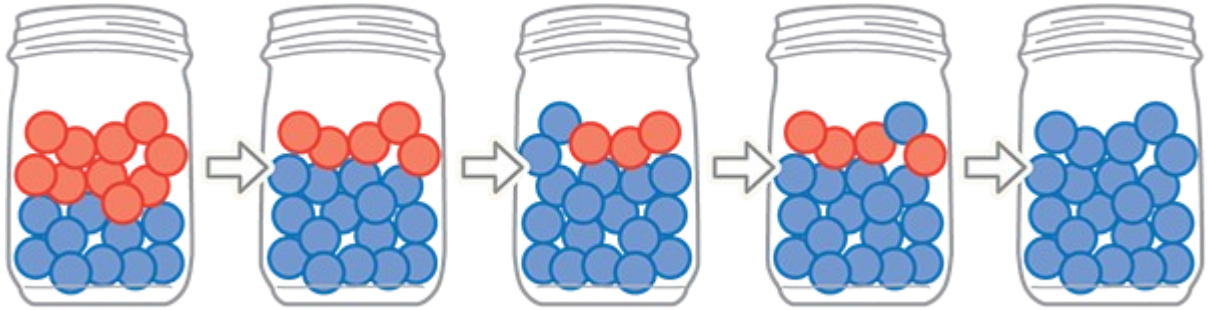
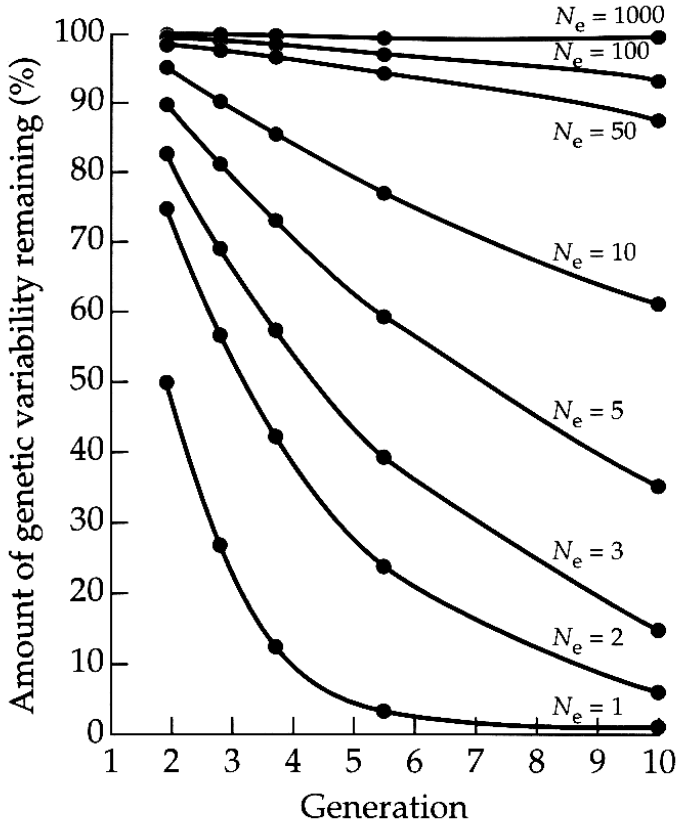


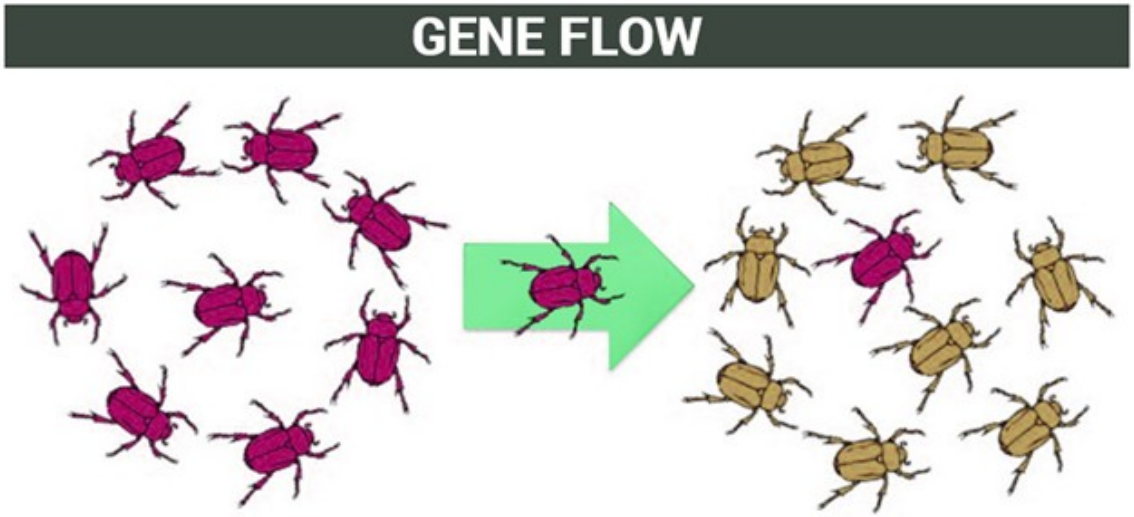
Image: http://scienceteachingarticles.blogspot.com.au/2015/02/genetic-drift.html



angela.mcgaughran@anu.edu.au

Genetic consequences

Image: https://byjus.com/biology/gene-flow/



- The exchange of genes between populations:
  - Migration, followed by breeding
  - Exchange of pollen
- Increases variation by adding new genetic material
- Absence results in increased differentiation

angela.mcgaughran@anu.edu.au

Question?

- We’re aiming to relocate a new population of Rufous hare-wallaby to a predator-free island as part of future conservation/management
- What are the effects of island transfer and translocation on diversity and differentiation?
- What is the best strategy for future translocation?

angela.mcgaughran@anu.edu.au

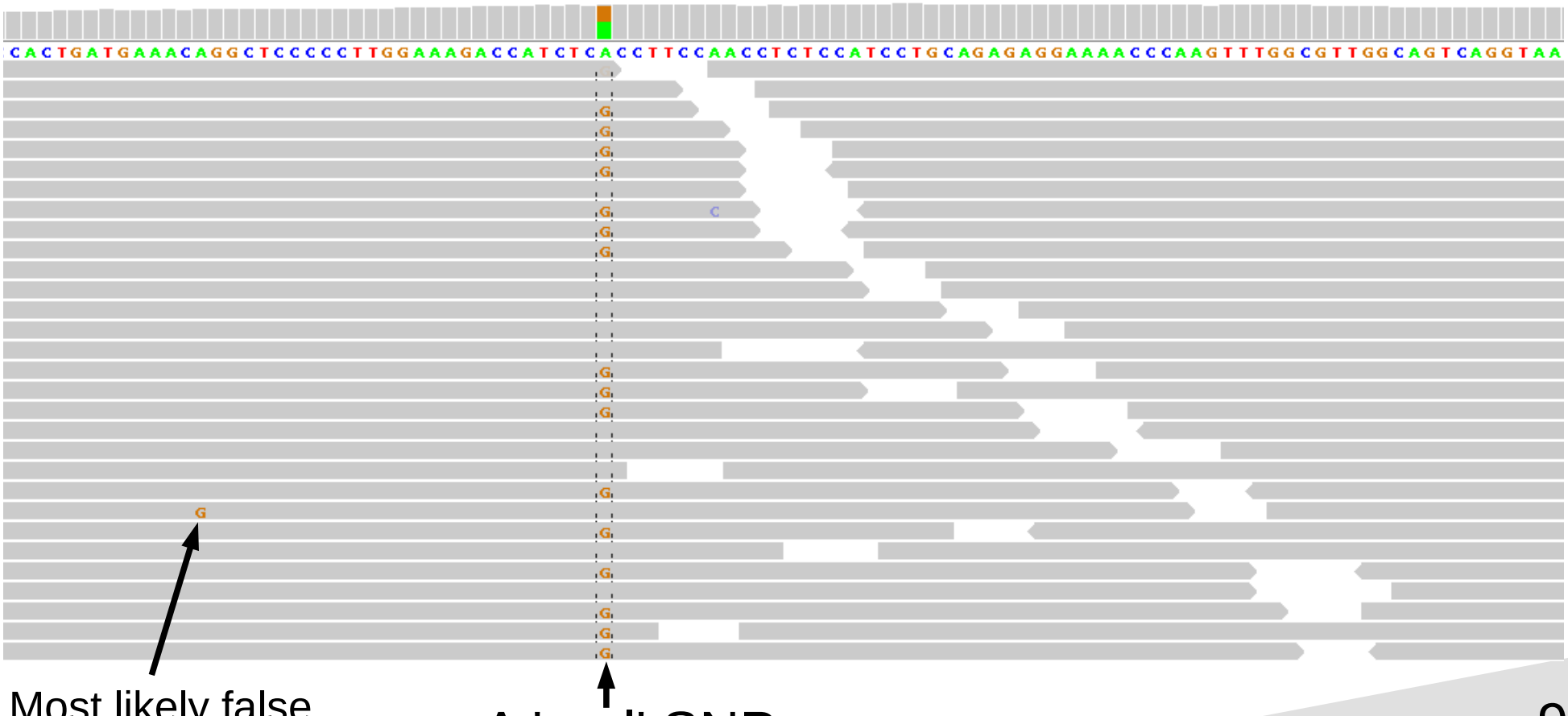
Data choice

- SNPs are good markers in population genomics:
  - Widespread across genomes – good representation
  - Older evolutionary events (slower mutation rates)
  - Can use to do a whole bunch of stuff!
- Extract SNPs from:
  - Genome / transcriptome data
  - Identify by mapping to a reference, or population pool

angela.mcgaughran@anu.edu.au

Alignment

Image: http://www.ikmb.uni-kiel.de/pibase/pibase_filtering.html



angela.mcgaughran@anu.edu.au

vcf files

- SNP file format is a variance call file (VCF):
  - https://en.wikipedia.org/wiki/Variant_Call_Format

angela.mcgaughran@anu.edu.au

R / RStudio

- R is a free, open source, powerful statistical software package
- RStudio is a wrapper program with a nicer interface than working directly with R

angela.mcgaughran@anu.edu.au

R / RStudio

- Launch RStudio
- Set your working directory
- Load the PopGenome package

angela.mcgaughran@anu.edu.au

Results

	Mainland	Isl_1	Isl_2	OTra_I	OTra_II	TTra_I
Mainland						
Island_1	0.4154					
Island_2	0.3819	0.5446				
One_TR_I	0.0433	0.5216	0.4845			
One_TR_II	0.0424	0.5211	0.4869	0.0648		
Two_TR_I	0.0510	0.5338	0.4977	0.0251	0.0728	
Two_TR_II	0.0608	0.5097	0.4767	0.1086	0.1144	0.1169

angela.mcgaughran@anu.edu.au

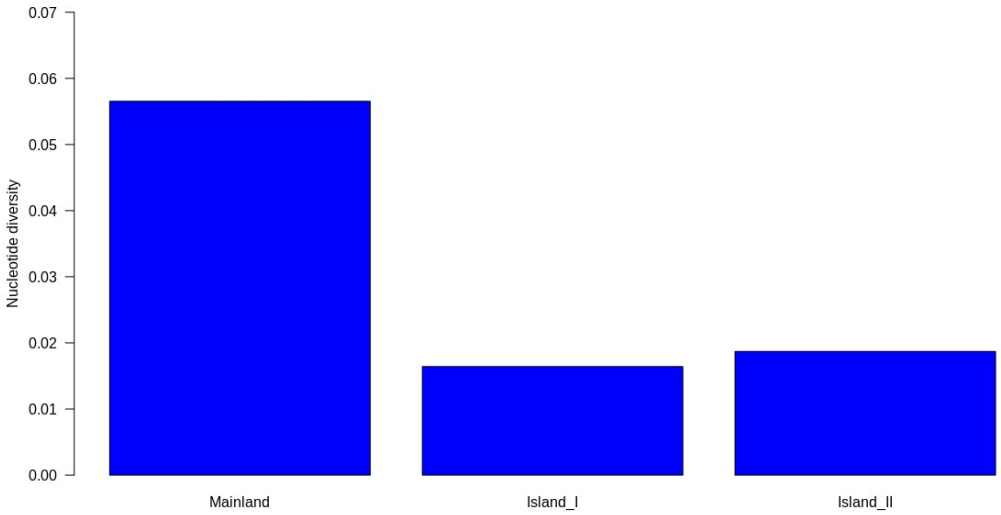
Results

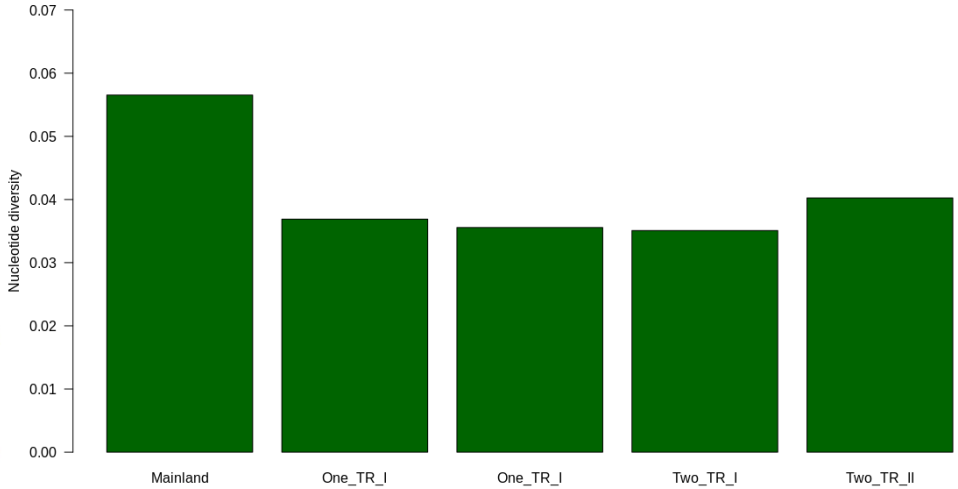
	Mainland	Isl_1	Isl_2	OTra_I	OTra_II	TTra_I
Mainland						
Island_1	0.4154					
Island_2	0.3819	0.5446				
One_TR_I	0.0433	0.5216	0.4845			
One_TR_II	0.0424	0.5211	0.4869	0.0648		
Two_TR_I	0.0510	0.5338	0.4977	0.0251	0.0728	
Two_TR_II	0.0608	0.5097	0.4767	0.1086	0.1144	0.1169

angela.mcgaughran@anu.edu.au

Advice to management

- Island pops are highly differentiated from the mainland
- Translocation reduces diversity





angela.mcgaughran@anu.edu.au

Advice to management

- To maximise the diversity of the new population:
  - Source individuals from both of the island and at least one of the mainland and/or translocated populations
- Mainland and island populations are so different → outbreeding depression?
  - Captive breeding experiments before translocation to ensure the suggested transfers are feasible

angela.mcgaughran@anu.edu.au