Genomes and Genomic Tools

Iestyn Whitehouse

What is a genome?

Telomeres
Centromeres
rDNA
Introns
Exons
Promoters
Repetitive DNA
“Junk DNA”
Transposable elements

How do you figure out the composition of a genome?

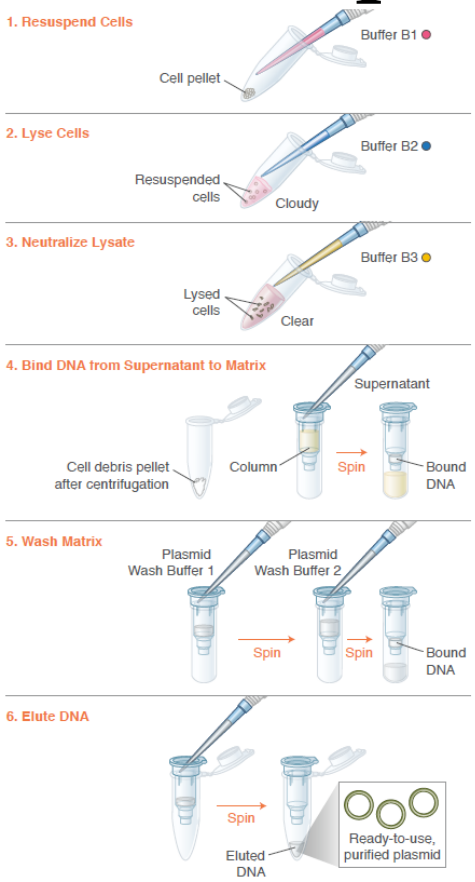
How much single copy DNA?

How much repetitive DNA?

What does a mini-prep do?

How does a mini-prep work?

How does a mini-prep work?



How does a mini-prep work?

Buffer P1

50 mM Tris-HCl pH 8.0
10 mM EDTA
100 µg/ml RNaseA

Buffer P2

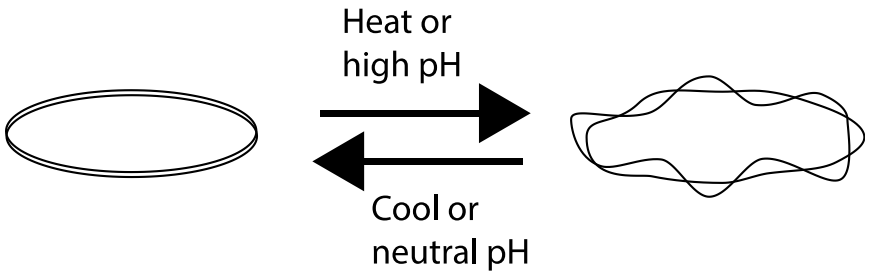
200 mM NaOH
1% SDS

Buffer P3

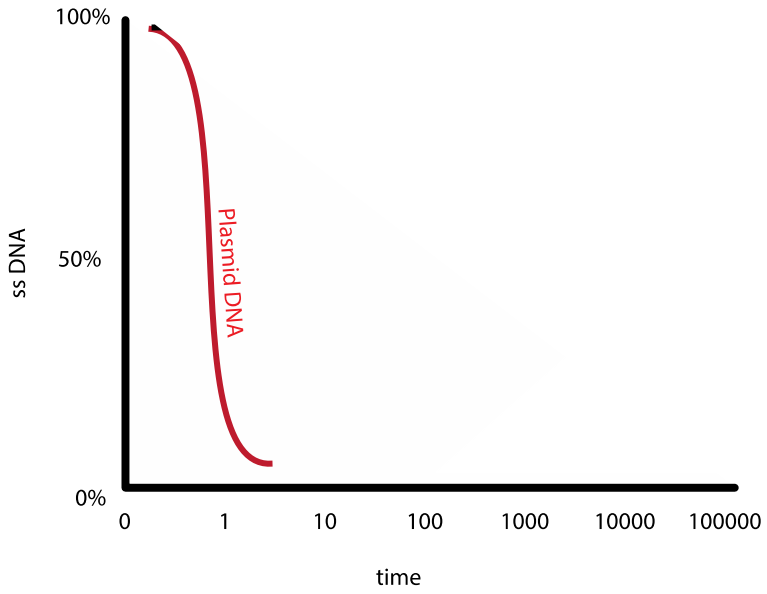
3.0 M potassium acetate pH 5.5

Spin in centrifuge

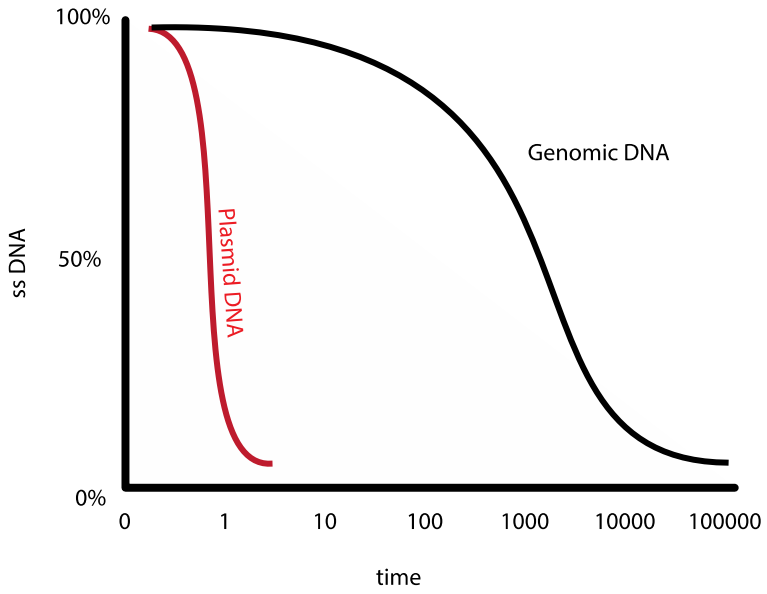
mini-prep



Reassociation Kinetics



Reassociation Kinetics



How does a mini-prep work?

Buffer P1

50 mM Tris-HCl pH 8.0
10 mM EDTA
100 µg/ml RNaseA

Buffer P2

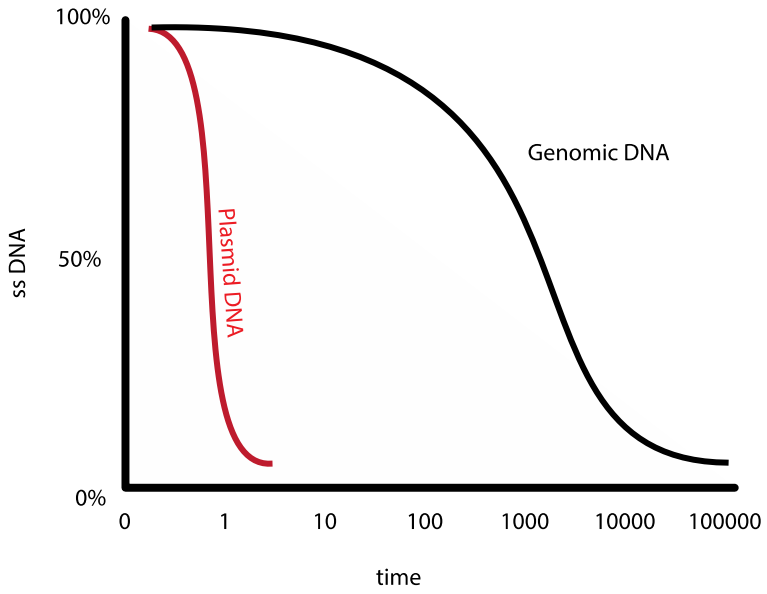
200 mM NaOH
1% SDS

Buffer P3

3.0 M potassium acetate pH 5.5

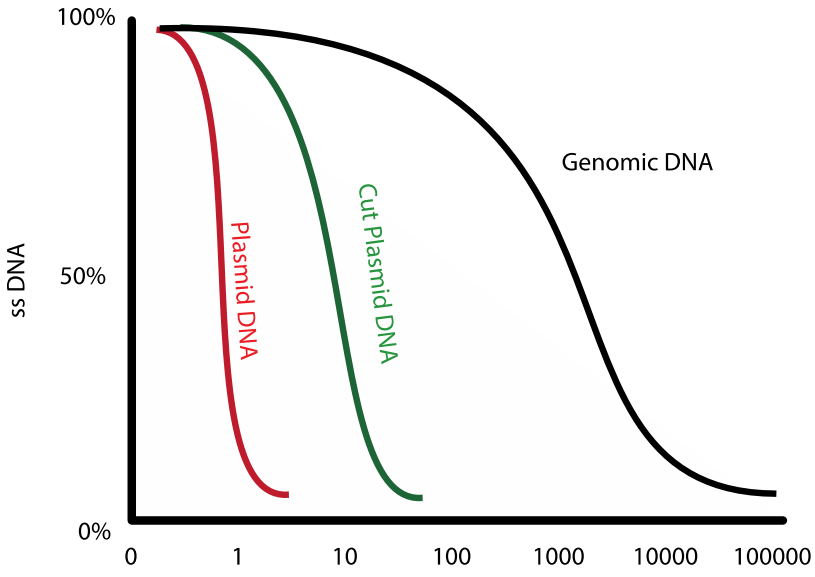
Spin in centrifuge

Reassociation Kinetics



What would happen if you cut the plasmid?

Reassociation Kinetics



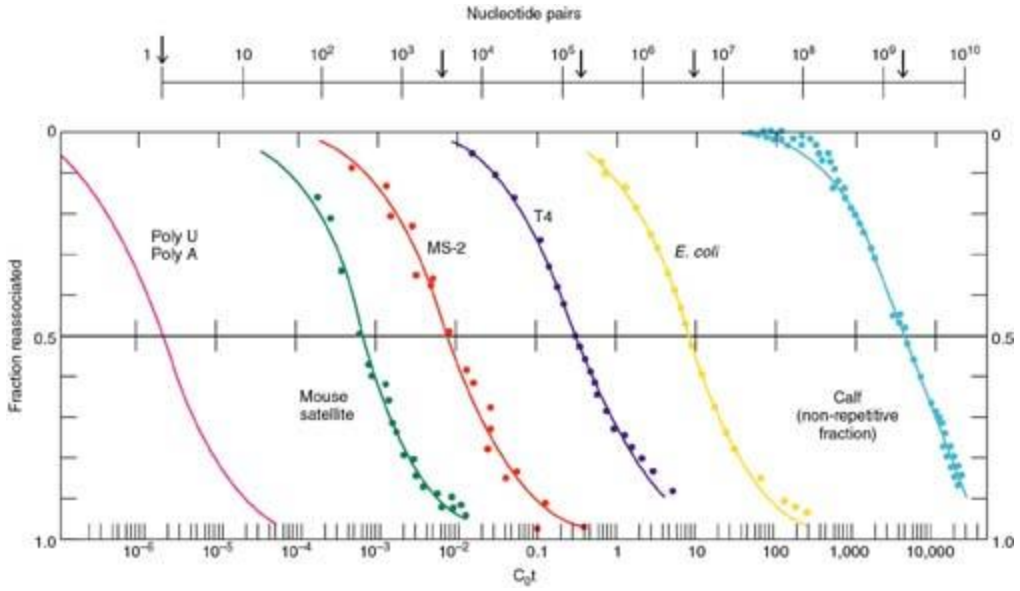
Why?

Reassociation Kinetics

Annealing is sensitive to concentration and time

$C_o t$ analysis

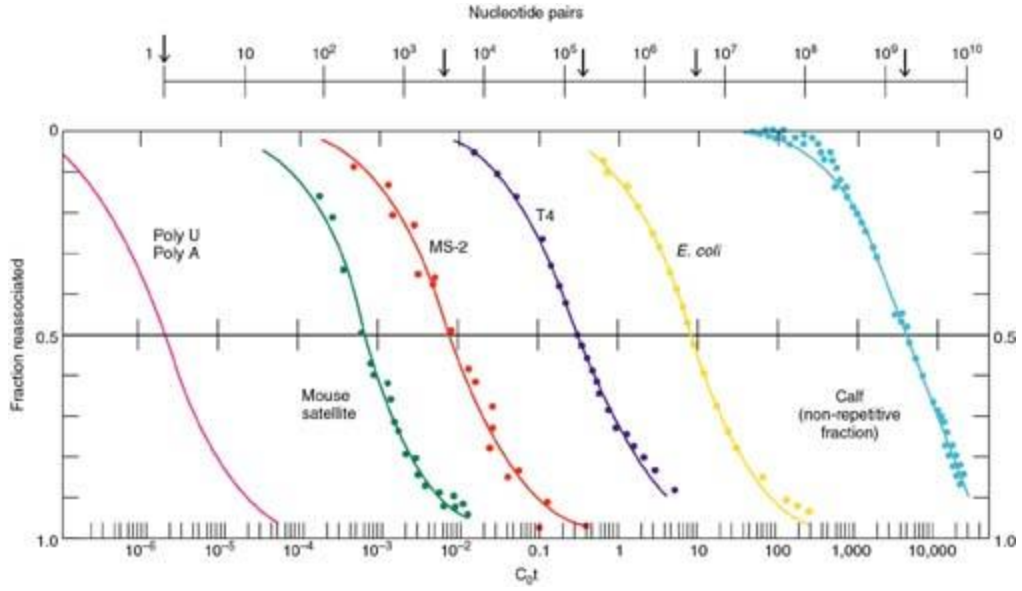
Reassociation scales according to genome size



Why?

Bigger genomes can have more combinations of sequence elements

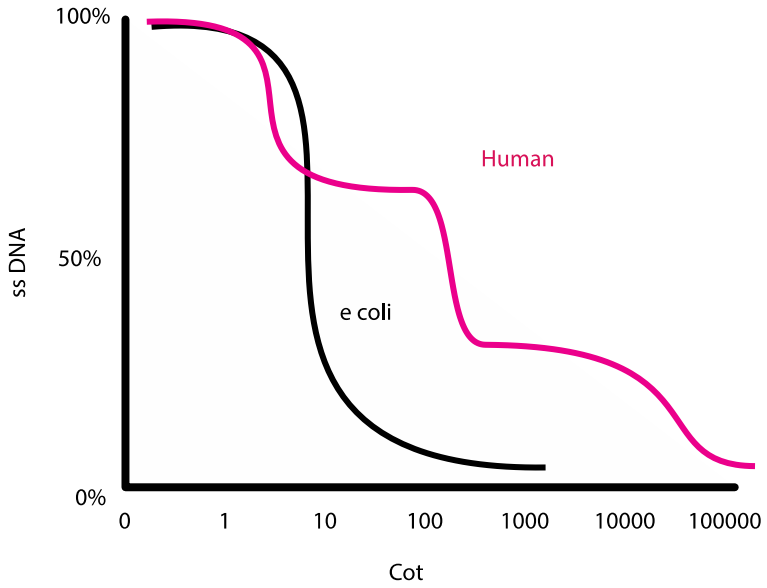
Not all DNA in the genome is equal !



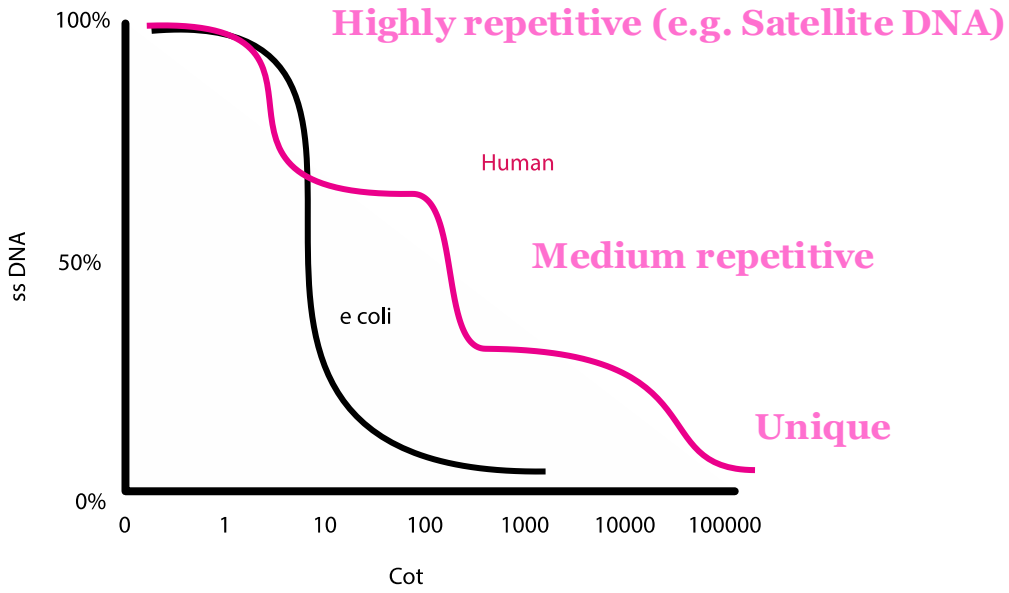
Why does mouse satellite DNA reassociate quickly?

Eukaryotic genomes often contain large quantities of repetitive DNA sequence.

Composition of eukaryotic genomes

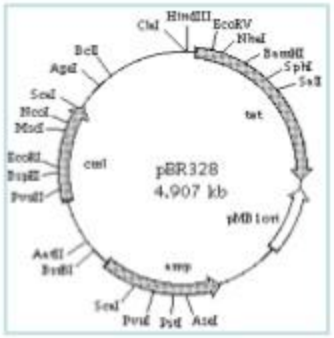


Composition of eukaryotic genomes

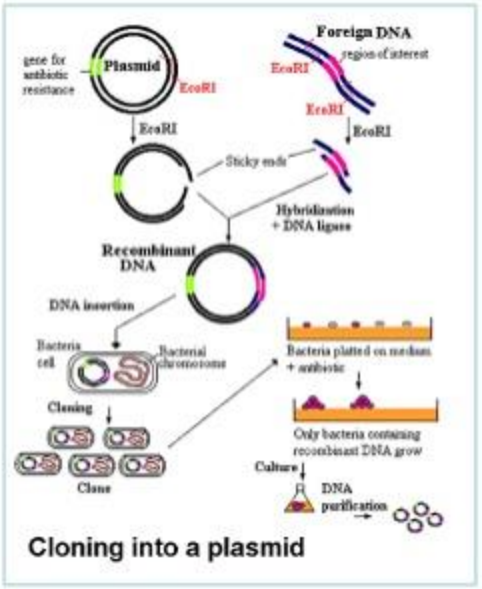


How do you sequence DNA?

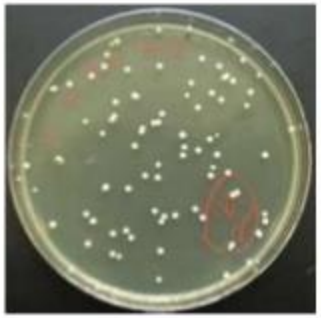
Preliminaries: Cloning



pBR322: is a vector, an engineered phage.
It can reproduce itself inside a bacterial host and do nothing else.



Cloning into a plasmid



Sub-clone DNA of interest

Transform bacteria

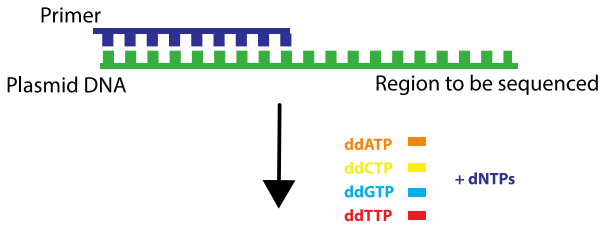
Pick colony

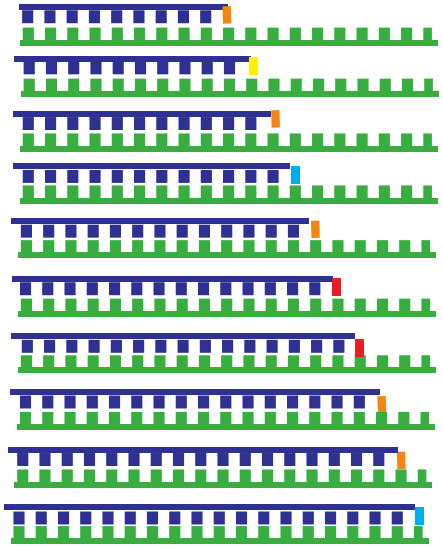
Miniprep

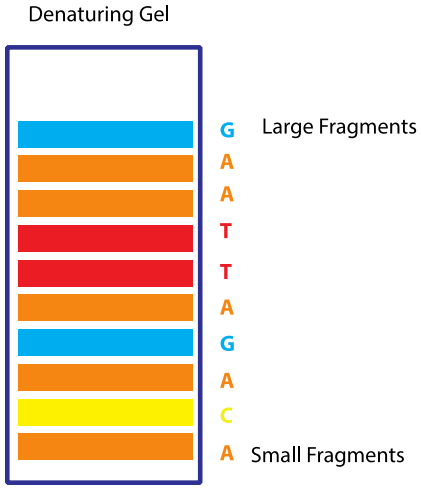
Submit purified plasmid with desired primer for sequencing

Why clone?

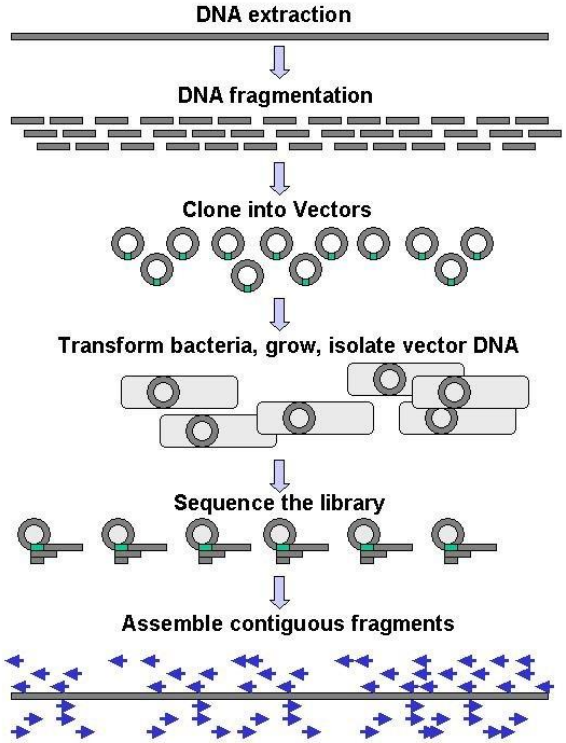
Sanger sequencing







Sequencing the human genome

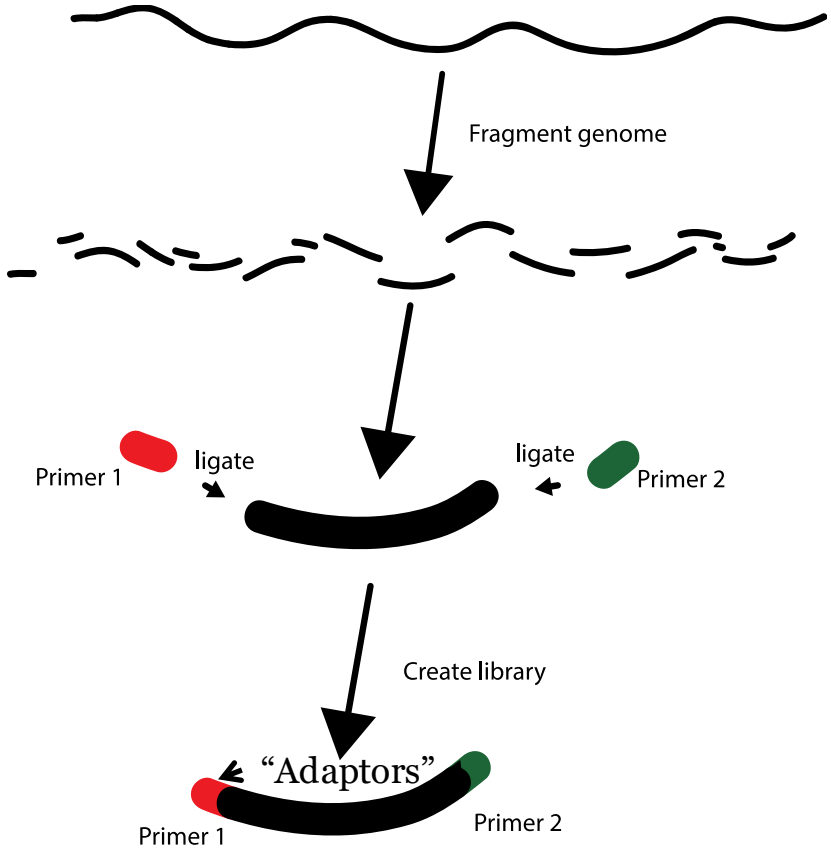


High throughput “Deep” sequencing

High throughput “Deep” sequencing

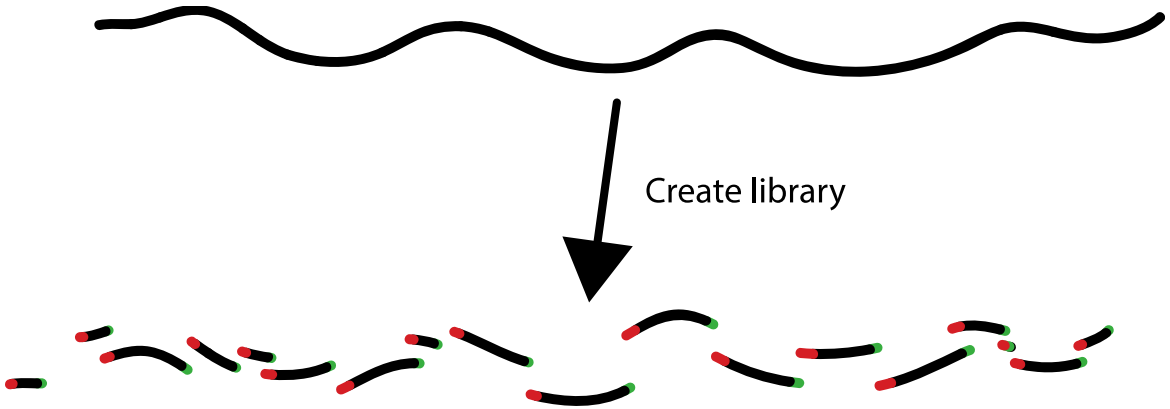
The key to deep sequencing is the generation of billions of individual “clones” without use of microorganisms!

Generating a library



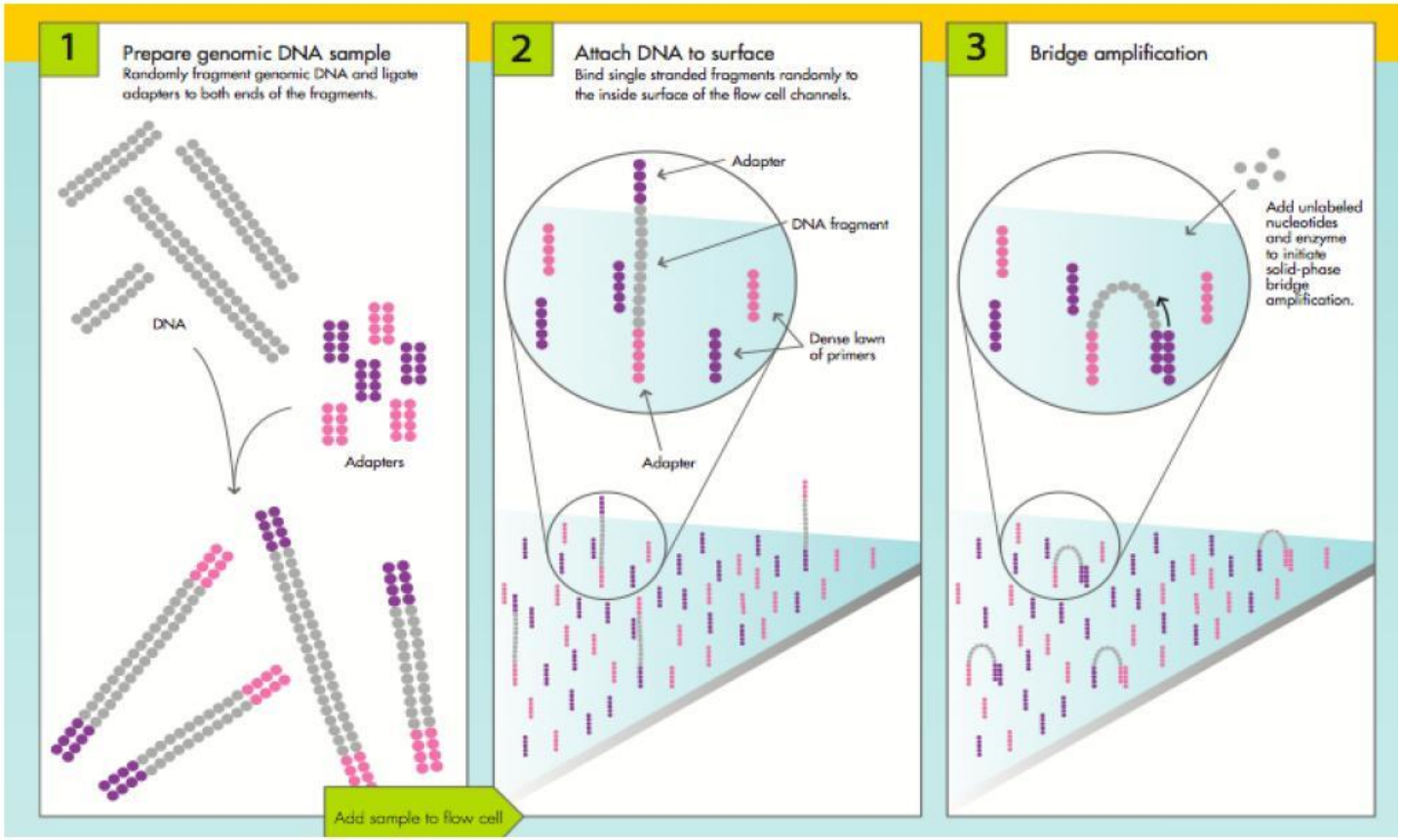
Generating a library

Generate a random mixture of sequence elements

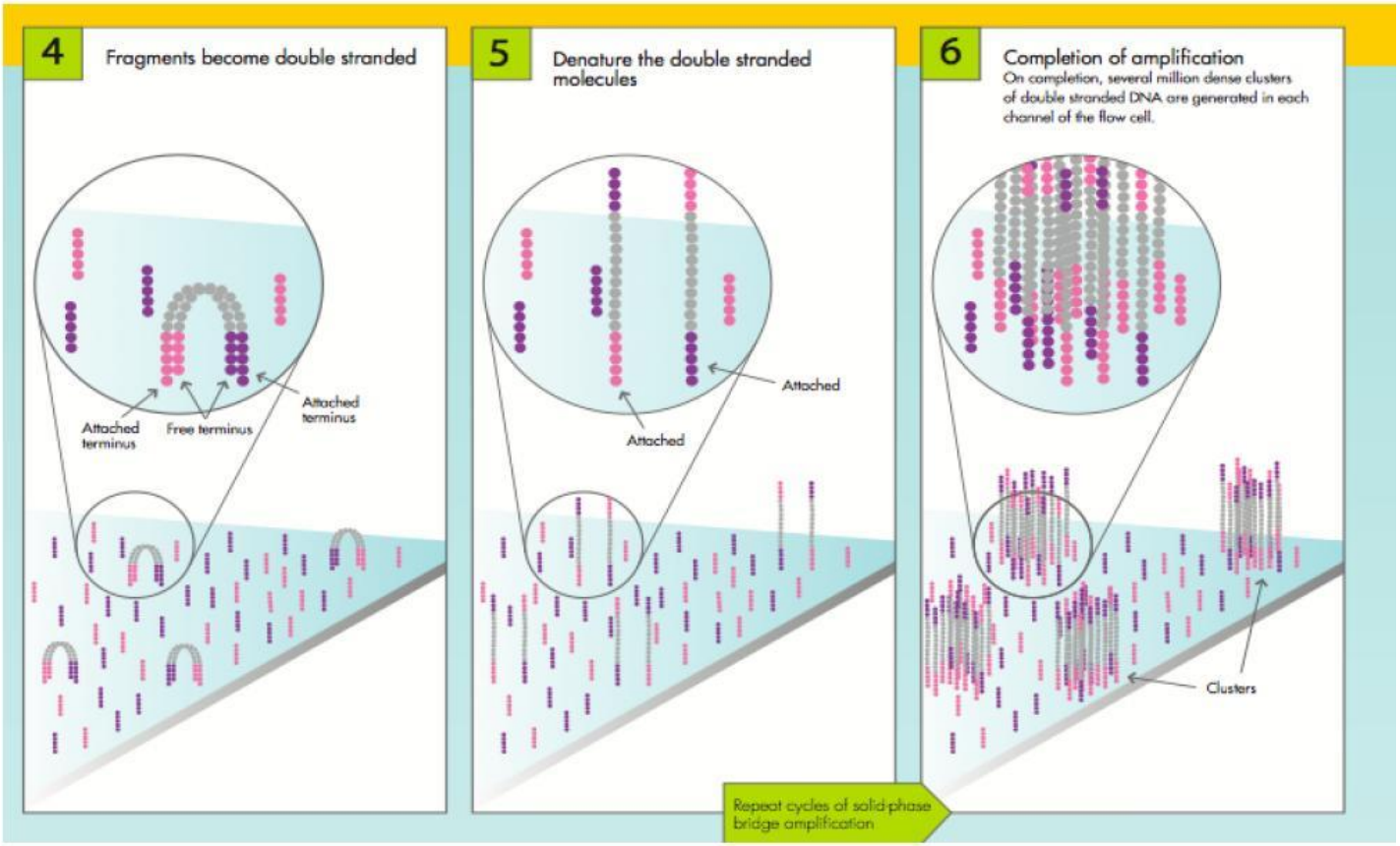


How do you clonally amplify and sequence DNA molecules from the whole population?

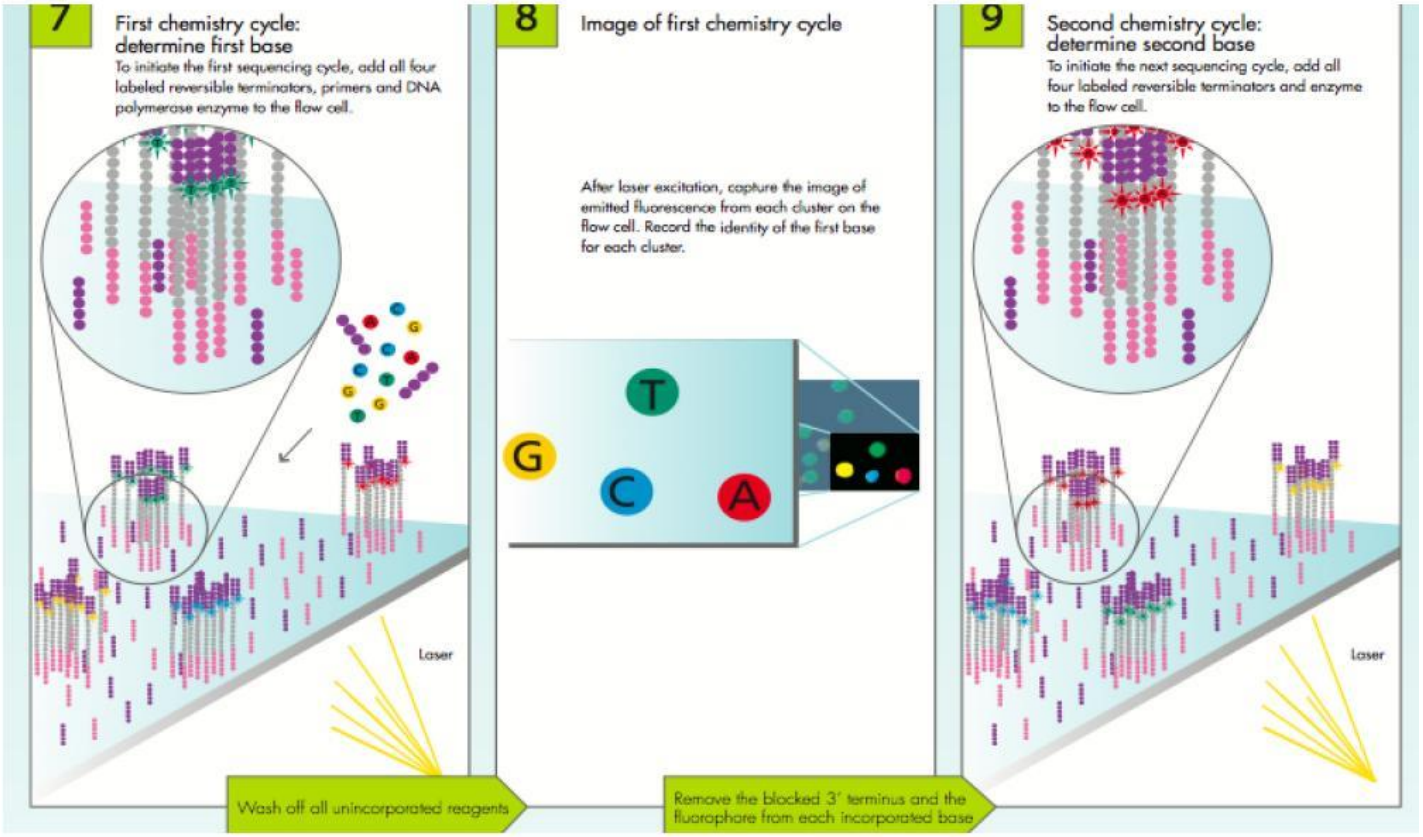
Illumina 1


1
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
DNA
Adapters
Add sample to flow cell
2
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Adapter
DNA fragment
Dense lawn of primers
Adapter
3
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

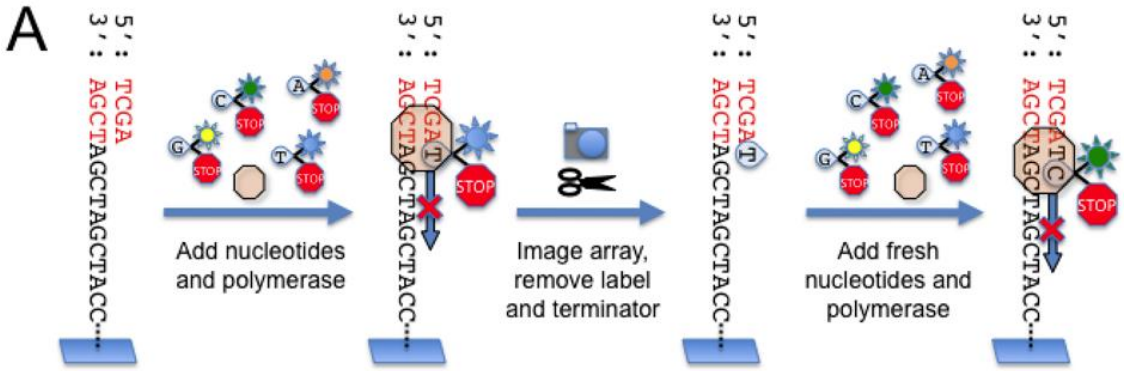
Illumina 2


4
Fragments become double stranded
Attached terminus
Free terminus
Attached terminus
Attached terminus
5
Denature the double stranded molecules
Attached
Attached
6
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Clusters
Repeat cycles of solid-phase bridge amplification

Illumina 3


7
First chemistry cycle:
determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Laser
Wash off all unincorporated reagents
8
Image of first chemistry cycle
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.
T
G
C
A
Remove the blocked 3' terminus and the fluorophore from each incorporated base
9
Second chemistry cycle:
determine second base
To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Laser

Sequence by synthesis



Deep sequencing can generate billions of short reads

Use a variety of algorithms to “map” the sequence reads to the genome.

FastQ files:

$ head N2-copy-num.R1.fastq

```
@HWI-D00269:151:C7J39ANXX:8:1101:1471:1979 1:N:0:TCTCGCGC
NGGCCACCATGGCGTATTGAATAAAATGTTGTACCTTACTTTGACAACAA
+
#<<@BGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

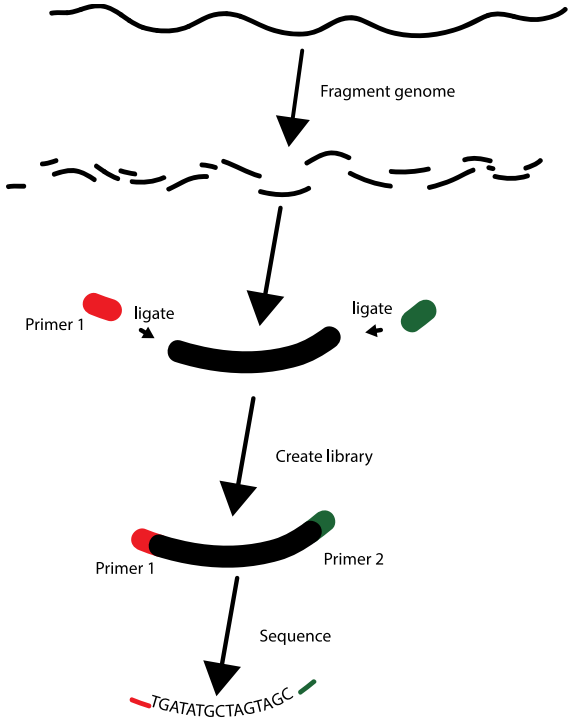
quality value characters in left-to-right increasing order of quality (ASCII):

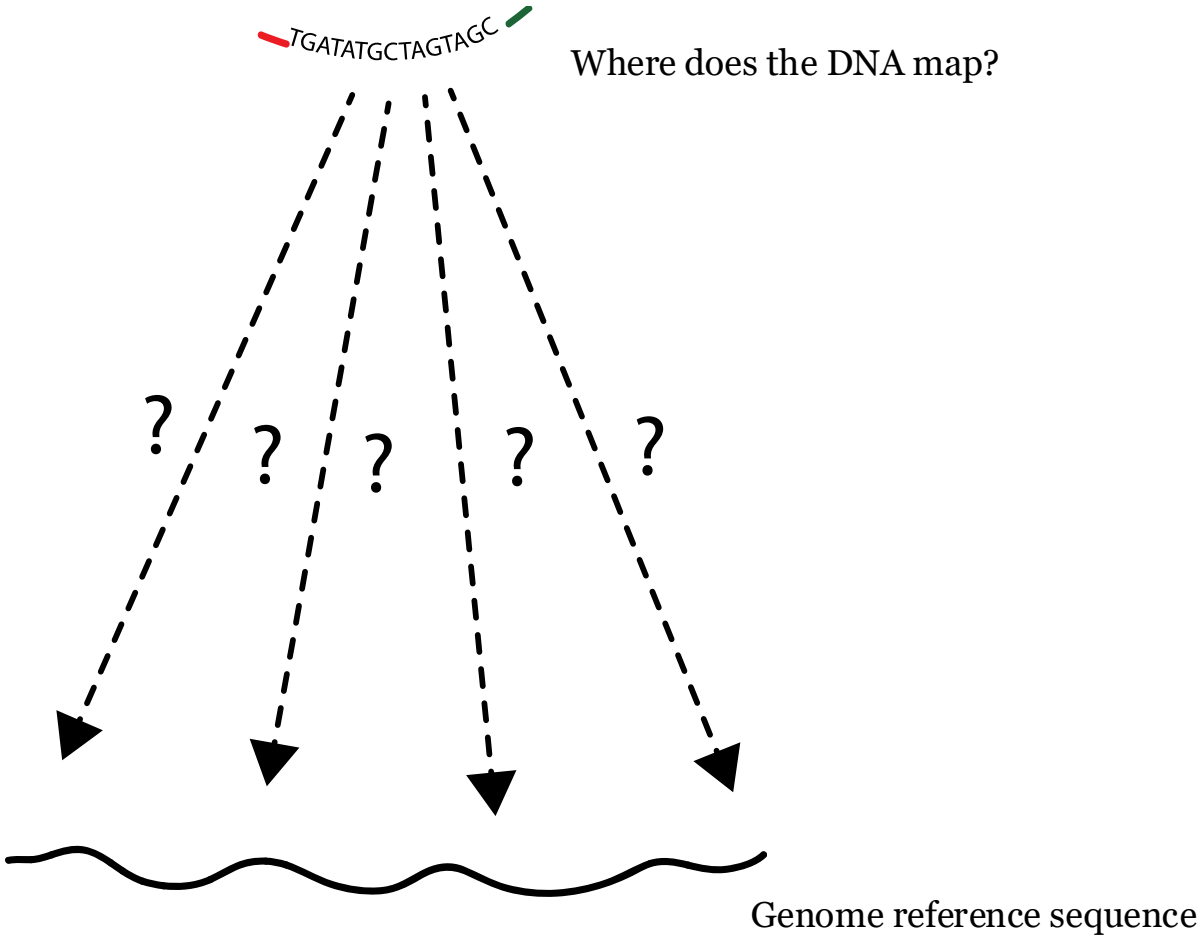
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Mapping Data

FastQ file  Bam/Sam file  Processed file for specific purpose

Read mapping:




TGATATGCTAGTAGC
Where does the DNA map?
? ? ? ? ?
Genome reference sequence

Read length matters!

Can you name a useful restriction enzyme?

EcoRI

GAATTC
CTTAAG

EcoRI

GAATTC
CTTAAG

TaqI

TCGA
AGCT

How often will you find an EcoRI site in DNA?

$4^{6} = 4096$

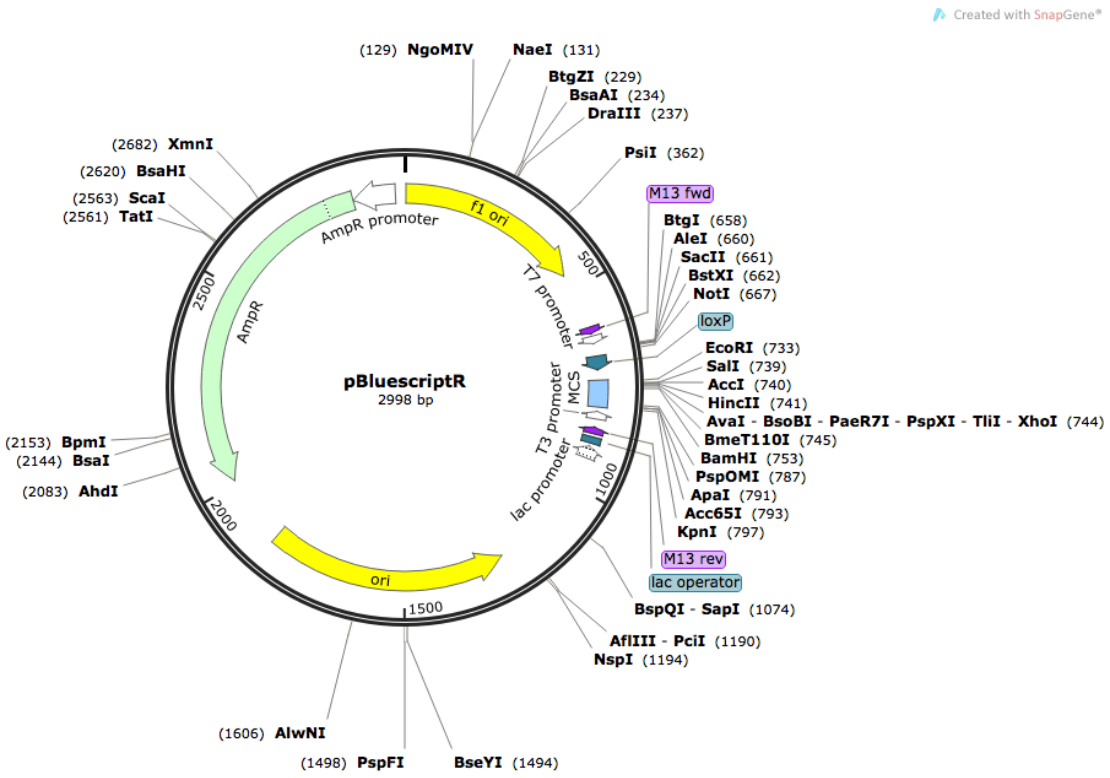
GAATTC
CTTAAG

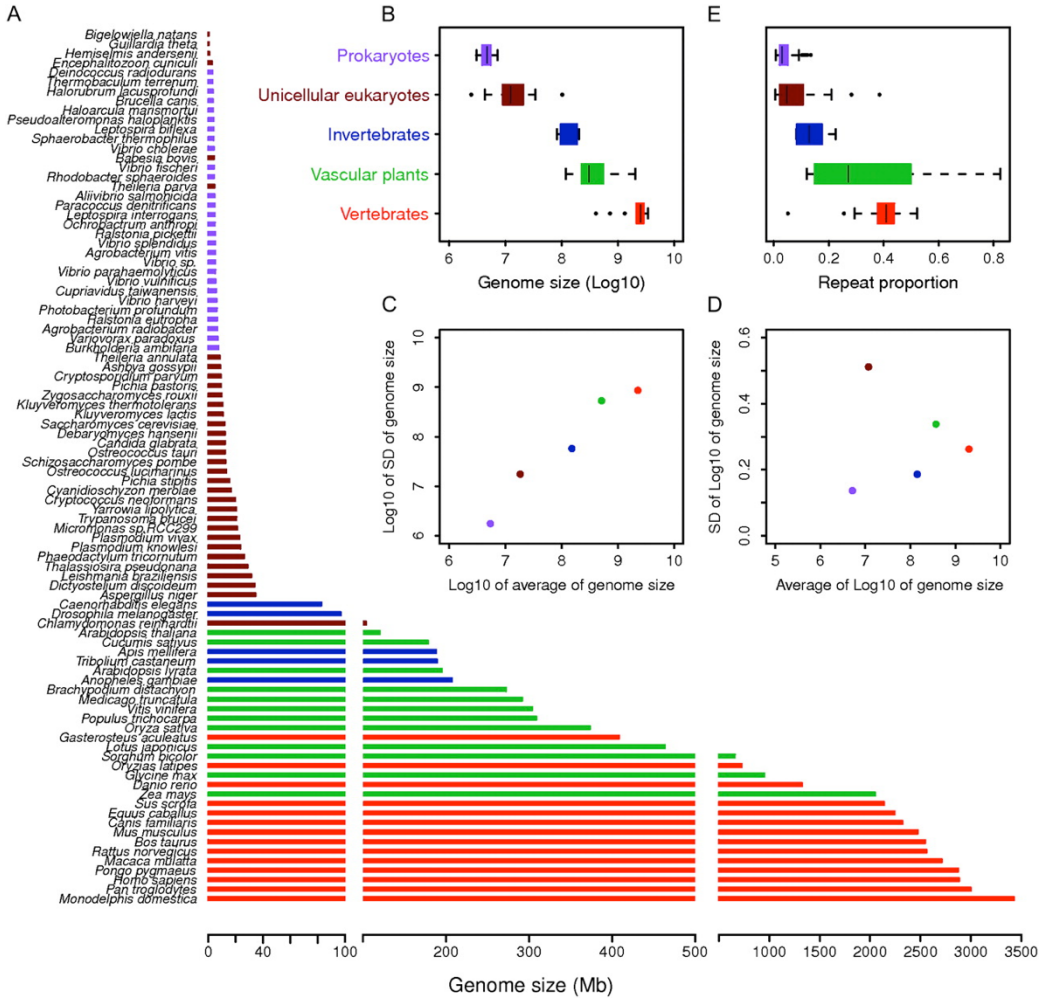
How often will you find a TaqI site?

TCGA
AGCT

$4^4 = 256$

Why is EcoRI typically more useful than TaqI?

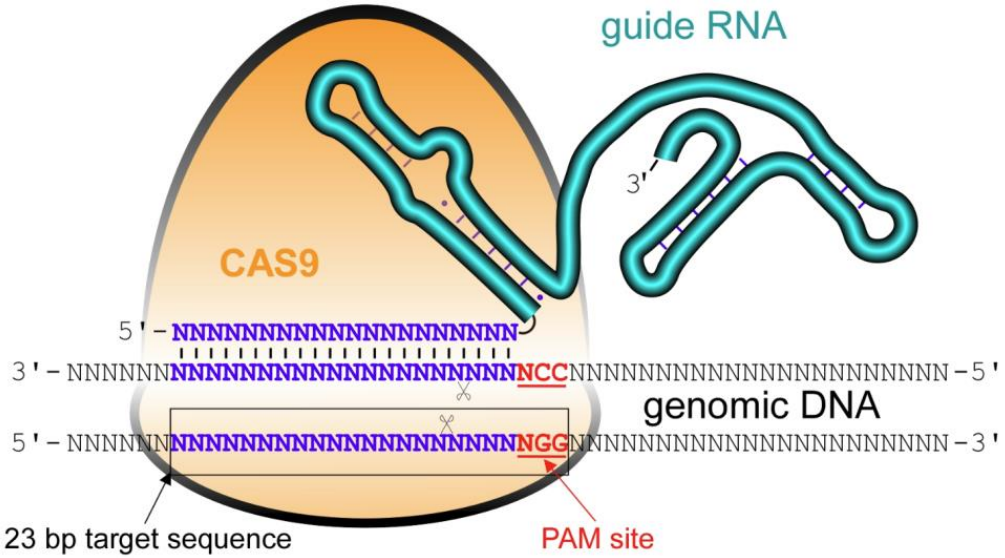



A
Bigelowiella natans
Guillardia theta
Hemiselmis andersenii
Encephalitozoon cuniculi
Deinococcus radiodurans
Thermobaculum terrenum
Halorubrum lacusprofundi
Brucella canis
Haloarcula marismortui
Pseudoalteromonas haloplanktis
Leptospira biflexa
Sphaerobacter thermophilus
Vibrio cholerae
Babesia bovis
Vibrio fischeri
Rhodobacter sphaeroides
Theileria parva
Aliivibrio salmonicida
Paracoccus denitrificans
Leptospira interrogans
Ochrobactrum anthropi
Ralstonia pickettii
Vibrio splendidus
Agrobacterium vitis
Vibrio sp.
Vibrio parahaemolyticus
Vibrio vulnificus
Cupriavidus taiwanensis
Vibrio harveyi
Photobacterium profundum
Ralstonia eutropha
Agrobacterium radiobacter
Variovorax paradoxus
Burkholderia ambifaria
Theileria annulata
Ashbya gossypii
Cryptosporidium parvum
Pichia pastoris
Zygosaccharomyces rouxii
Kluyveromyces thermotolerans
Kluyveromyces lactis
Saccharomyces cerevisiae
Debaryomyces hansenii
Candida glabrata
Ostreococcus tauri
Schizosaccharomyces pombe
Ostreococcus lucimarinus
Pichia stipitis
Cyanidioschyzon merolae
Cryptococcus neoformans
Yarrowia lipolytica
Trypanosoma brucei
Micromonas sp. RCC299
Plasmodium vivax
Plasmodium knowlesi
Phaeodactylum tricornutum
Thalassiosira pseudonana
Leishmania braziliensis
Dictyostelium discoideum
Aspergillus niger
Caenorhabditis elegans
Drosophila melanogaster
Chlamydomonas reinhardtii
Arabidopsis thaliana
Cucumis sativus
Apis mellifera
Tribolium castaneum
Arabidopsis lyrata
Anopheles gambiae
Brachypodium distachyon
Medicago truncatula
Vitis vinifera
Populus trichocarpa
Oryza sativa
Gasterosteus aculeatus
Lotus japonicus
Sorghum bicolor
Oryzias latipes
Glycine max
Danio rerio
Zea mays
Sus scrofa
Equus caballus
Canis familiaris
Mus musculus
Bos taurus
Rattus norvegicus
Macaca mulatta
Pongo pygmaeus
Homo sapiens
Pan troglodytes
Monodelphis domestica
0
40
100
200
300
400
500
1000
1500
2000
2500
3000
3500
Genome size (Mb)
B
Prokaryotes
Unicellular eukaryotes
Invertebrates
Vascular plants
Vertebrates
6
7
8
9
10
Genome size (Log10)
E
0.0
0.2
0.4
0.6
0.8
Repeat proportion
C
Log10 of SD of genome size
6
7
8
9
10
Log10 of average of genome size
D
SD of Log10 of genome size
0.0
0.2
0.4
0.6
5
6
7
8
9
10
Average of Log10 of genome size

Motif length	Possible combinations
1	4
2	16
3	64
4	256
5	1,024
6	4,096
7	16,384
8	65,536
9	262,144
10	1,048,576
11	4,194,304
12	16,777,216
13	67,108,864
14	268,435,456
15	1,073,741,824
16	4,294,967,296
17	17,179,869,184
18	68,719,476,736
19	274,877,906,944
20	1,099,511,627,776
21	4,398,046,511,104
22	17,592,186,044,416
23	70,368,744,177,664
24	281,474,976,710,656
25	1,125,899,906,842,620
26	4,503,599,627,370,500
27	18,014,398,509,482,000
28	72,057,594,037,927,900
29	288,230,376,151,712,000
30	1,152,921,504,606,850,000

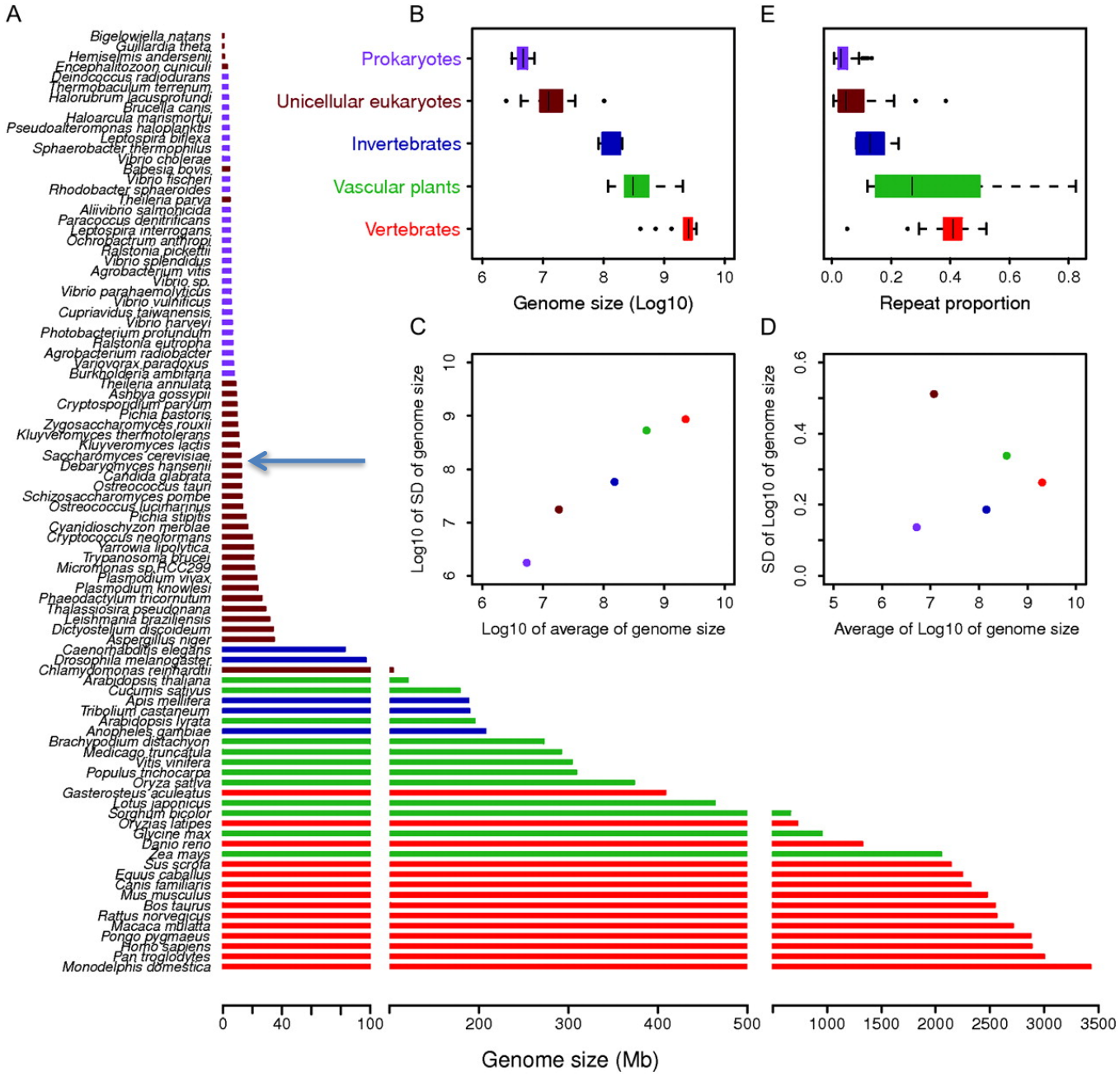
Why is CrisprCas9 such a useful enzyme?

Why is CrisprCas9 such a useful enzyme?



Sequence complexity

How long a “read” would you need to match uniquely to the budding yeast genome?


A
Bigelowiella natans
Guillardia theta
Hemiselmis andersenii
Encephalitozoon cuniculi
Deinococcus radiodurans
Thermobaculum terrenum
Halorubrum lacusprofundi
Brucella canis
Haloarcula marismortui
Pseudoalteromonas haloplanktis
Leptospira biflexa
Sphaerobacter thermophilus
Vibrio cholerae
Babesia bovis
Vibrio fischeri
Rhodobacter sphaeroides
Theileria parva
Aliivibrio salmonicida
Paracoccus denitrificans
Leptospira interrogans
Ochrobactrum anthropi
Ralstonia pickettii
Vibrio splendidus
Agrobacterium vitis
Vibrio sp.
Vibrio parahaemolyticus
Vibrio vulnificus
Cupriavidus taiwanensis
Vibrio harveyi
Photobacterium profundum
Ralstonia eutropha
Agrobacterium radiobacter
Variovorax paradoxus
Burkholderia ambifaria
Theileria annulata
Ashbya gossypii
Cryptosporidium parvum
Pichia pastoris
Zygosaccharomyces rouxii
Kluyveromyces thermotolerans
Kluyveromyces lactis
Saccharomyces cerevisiae
Debaryomyces hansenii
Candida glabrata
Ostreococcus tauri
Schizosaccharomyces pombe
Ostreococcus lucimarinus
Pichia stipitis
Cyanidioschyzon merolae
Cryptococcus neoformans
Yarrowia lipolytica
Trypanosoma brucei
Micromonas sp. RCC299
Plasmodium vivax
Plasmodium knowlesi
Phaeodactylum tricornutum
Thalassiosira pseudonana
Leishmania braziliensis
Dictyostelium discoideum
Aspergillus niger
Caenorhabditis elegans
Drosophila melanogaster
Chlamydomonas reinhardtii
Arabidopsis thaliana
Cucumis sativus
Apis mellifera
Tribolium castaneum
Arabidopsis lyrata
Anopheles gambiae
Brachypodium distachyon
Medicago truncatula
Vitis vinifera
Populus trichocarpa
Oryza sativa
Gasterosteus aculeatus
Lotus japonicus
Sorghum bicolor
Oryzias latipes
Glycine max
Danio rerio
Zea mays
Sus scrofa
Equus caballus
Canis familiaris
Mus musculus
Bos taurus
Rattus norvegicus
Macaca mulatta
Pongo pygmaeus
Homo sapiens
Pan troglodytes
Monodelphis domestica
0
40
100
200
300
400
500
1000
1500
2000
2500
3000
3500
Genome size (Mb)
B
Prokaryotes
Unicellular eukaryotes
Invertebrates
Vascular plants
Vertebrates
6
7
8
9
10
Genome size (Log10)
E
0.0
0.2
0.4
0.6
0.8
Repeat proportion
C
Log10 of SD of genome size
6
7
8
9
10
Log10 of average of genome size
D
SD of Log10 of genome size
0.0
0.2
0.4
0.6
5
6
7
8
9
10
Average of Log10 of genome size

Motif length	Possible combinations
1	4
2	16
3	64
4	256
5	1,024
6	4,096
7	16,384
8	65,536
9	262,144
10	1,048,576
11	4,194,304
12	16,777,216
13	67,108,864
14	268,435,456
15	1,073,741,824
16	4,294,967,296
17	17,179,869,184
18	68,719,476,736
19	274,877,906,944
20	1,099,511,627,776
21	4,398,046,511,104
22	17,592,186,044,416
23	70,368,744,177,664
24	281,474,976,710,656
25	1,125,899,906,842,620
26	4,503,599,627,370,500
27	18,014,398,509,482,000
28	72,057,594,037,927,900
29	288,230,376,151,712,000
30	1,152,921,504,606,850,000

12bp!

Sequence complexity

Entered nucleotide pattern : GTAGGCAATCTC
Total Hits : 1
Sequences Searched : 17
Dataset : COMPLETE S. cerevisiae Genome DNA
Strand : both strands

Sequence complexity

Entered nucleotide pattern : GGGAAAATTATC
Total Hits : 6
Sequences Searched : 17
Dataset : COMPLETE S. cerevisiae Genome DNA
Strand : both strands

Sequence complexity

Entered nucleotide pattern : AAATTTAAATTT
Total Hits : 34
Sequences Searched : 17
Dataset : COMPLETE S. cerevisiae Genome DNA
Strand : both strands

Sequence complexity

Genomes are not composed of random sequence: some regions are very biased in sequence composition so have a low complexity.

This means that you typically need sequencing reads much longer than what's predicted just from genome size.

Sequence complexity

Sequence complexity describes the number of possible “word” combinations within a given sequence

GATAGATAGATCATCA Complex sequence

AAAAAAAAAATTTTTTATTAAAAAAAAA Simple sequence

Sequence complexity

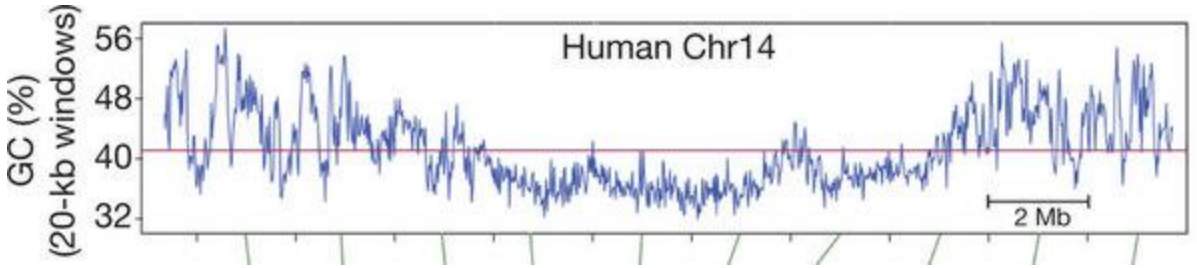
Sequence complexity describes the number of possible “word” combinations within a given sequence

10kb plasmid composed of GATC:
Need a sequence read of **7bp** for a unique match ($4^7 = 16{,}384$)

10kb plasmid composed of AT only
Need a sequence read of **14bp** for a unique match ($2^{14} = 16{,}384$)

Sequence complexity

Different regions of the genome have different base composition



“mapability”

Mapability is a measure of both sequence complexity and uniqueness.

Typically (but not always) sequences with low complexity have low mapability.

Sequence 1: TGATAGATCGATCGATCGATCGATCGA

Sequence 2: AGTCGATTCGATCGAT

Motif length	Possible combinations
1	4
2	16
3	64
4	256
5	1,024
6	4,096
7	16,384
8	65,536
9	262,144
10	1,048,576
11	4,194,304
12	16,777,216
13	67,108,864
14	268,435,456
15	1,073,741,824
16	**4,294,967,296**
17	17,179,869,184
18	68,719,476,736
19	274,877,906,944
20	1,099,511,627,776
21	4,398,046,511,104
22	17,592,186,044,416
23	70,368,744,177,664
24	281,474,976,710,656
25	1,125,899,906,842,620
26	4,503,599,627,370,500
27	18,014,398,509,482,000
28	72,057,594,037,927,900
29	288,230,376,151,712,000
30	**1,152,921,504,606,850,000**

Read length and “mappability”

Sequence 1: TGATAGATCGATCGATCGATCGATCGA

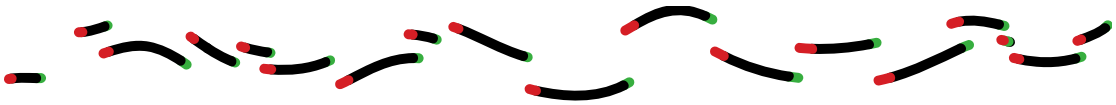
Sequence 2: AGTCGATTCGATCGAT

Unique DNA

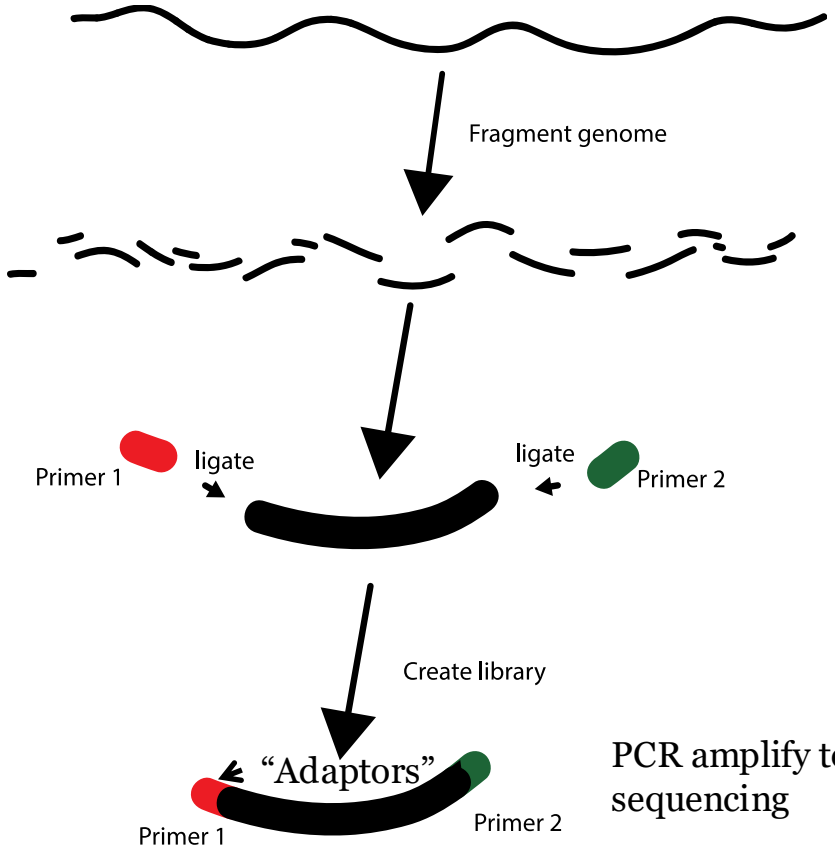
Repetitive DNA

Genomics approaches

Nearly all approaches create a “library” with adapters on either end.



Generating a library

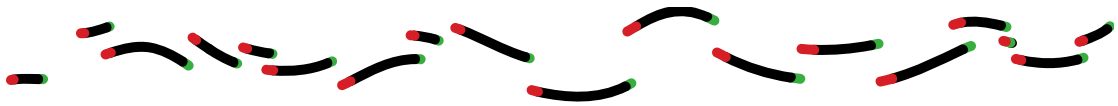


PCR amplify to generate "enough" for sequencing

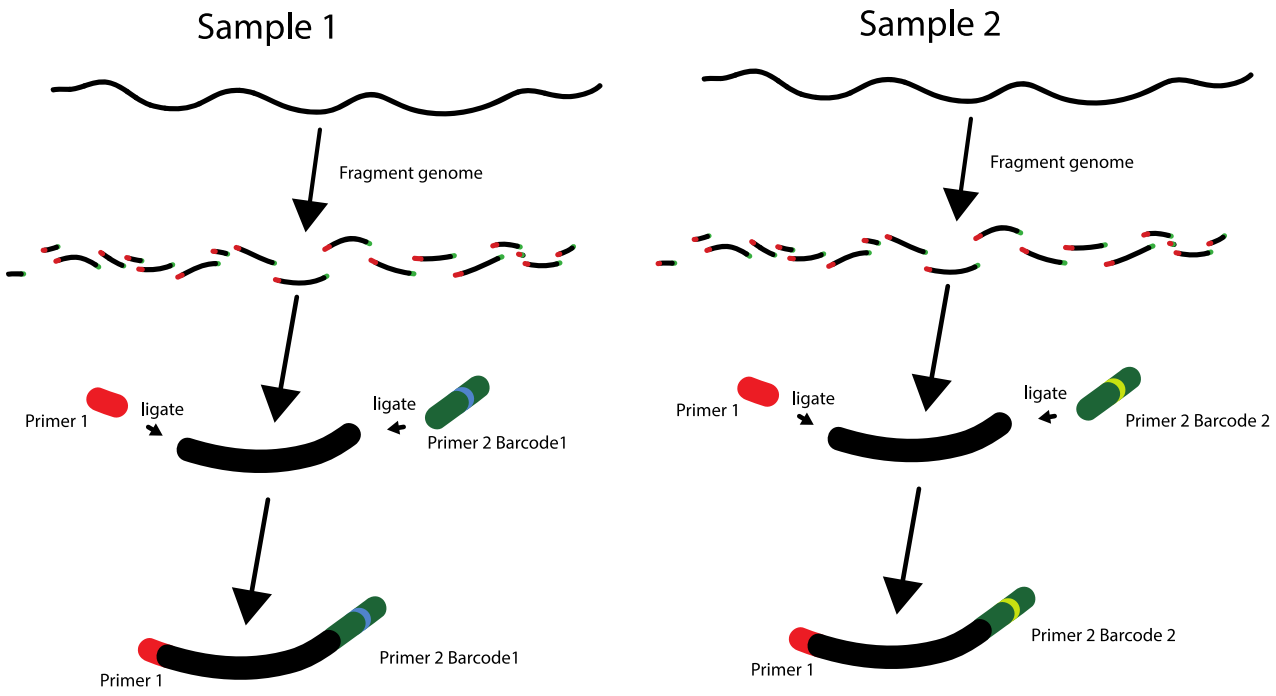
Barcoding

Most sequencing machines produce a fixed number of reads per flow cell.

What do you do if you are working with a small genome and only need a few million reads per sample?



Barcoding


Sample 1
Fragment genome
Primer 1
ligate
ligate
Primer 2 Barcode1
Primer 1
Primer 2 Barcode1
Sample 2
Fragment genome
Primer 1
ligate
ligate
Primer 2 Barcode 2
Primer 1
Primer 2 Barcode 2

Barcoding

Barcoding

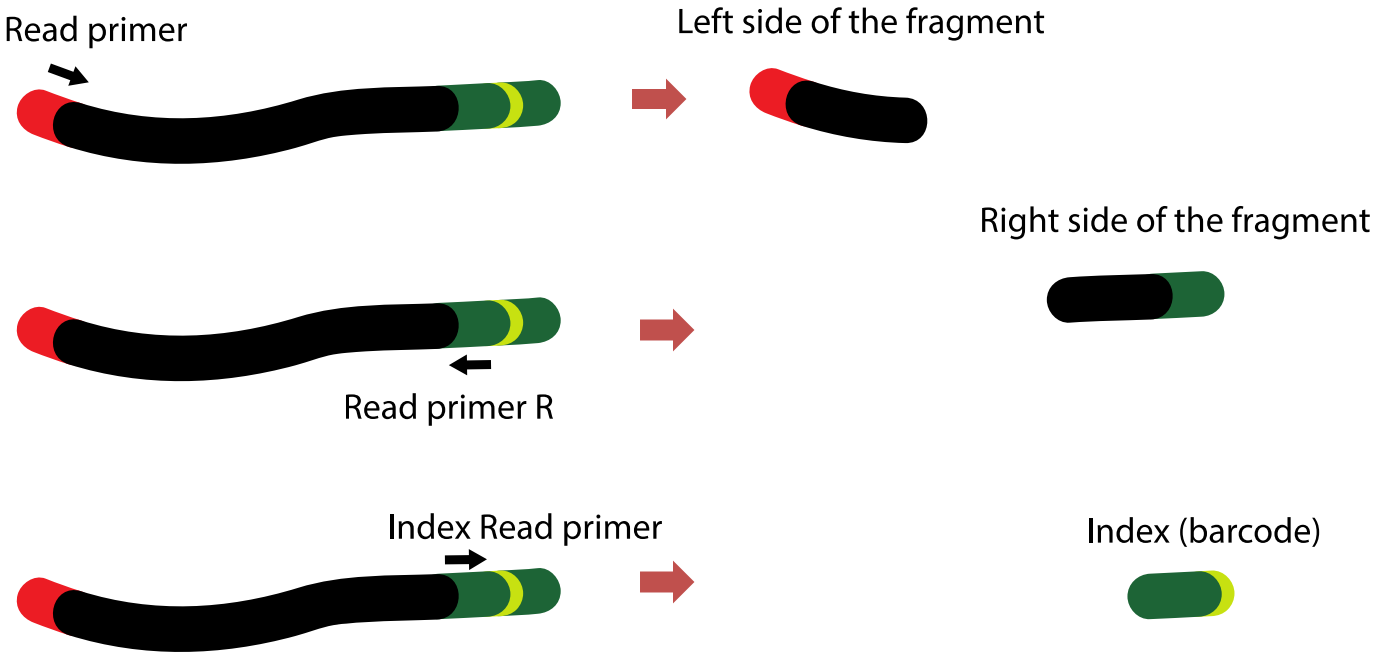

F primer
CATCTCATCCCTGCGTGTC
Multiplex index primer
5′ATCACCGACTGCCCATAGAGAGGAAAGCGG
True-seq F (30nt)
5′
AATGATACGGCGACCACCGAGATCTACACCCATCTCATCCCTGCGTGTC NNNN(.....)NNNN ATCACCGACTGCCCATAGAGAGGAAAGCGGATCACGATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTAGATGTGGGTAGAGTAGGGACGCACAG NNNN(.....)NNNN TAGTGGCTGACGGGTATCTCTCCTTTCGCCTAGTGCTAGAGCATACGGCAGAAGACGAAC
P7
5′
Barcode I7, Illumina
R701 ATCACG
R702 CGATGT
R703 TTAGGC
R704 TGACCA
R705 ACAGTG
R706 GCCAAT
R707 CAGATC
R708 ACTTGA

Paired end sequencing

Sequence both ends

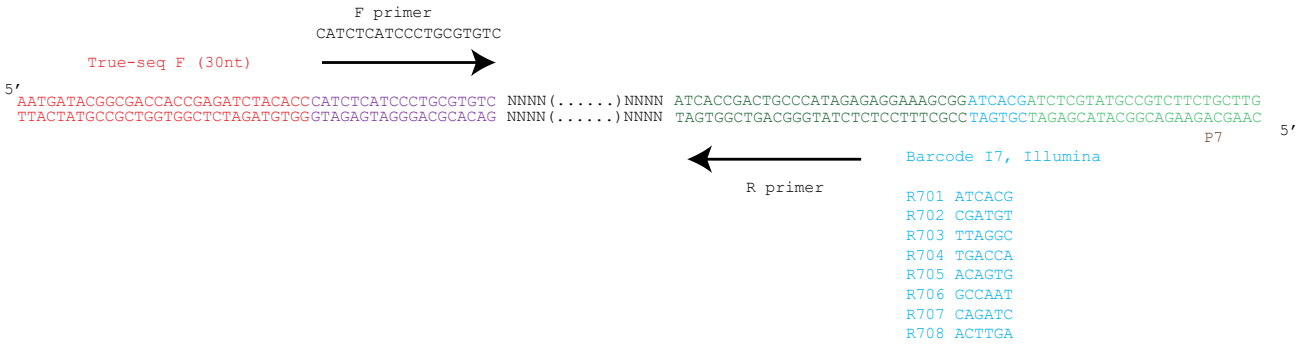
Paired end sequencing

Sequence both ends of the molecule



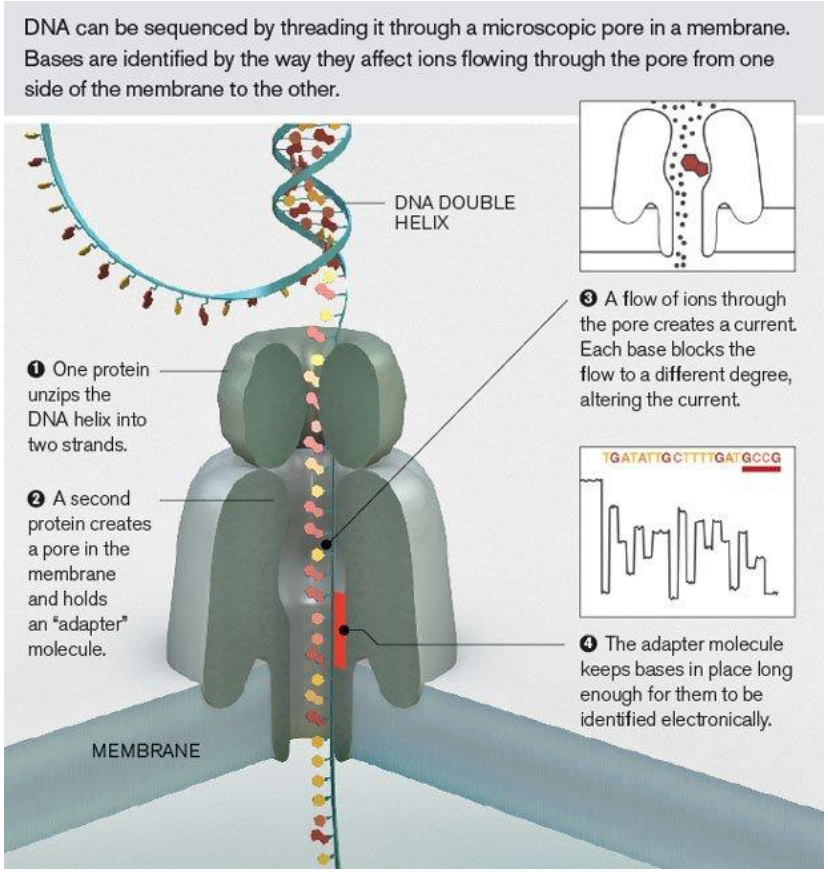
Paired end sequencing

Sequence both ends of the molecule!



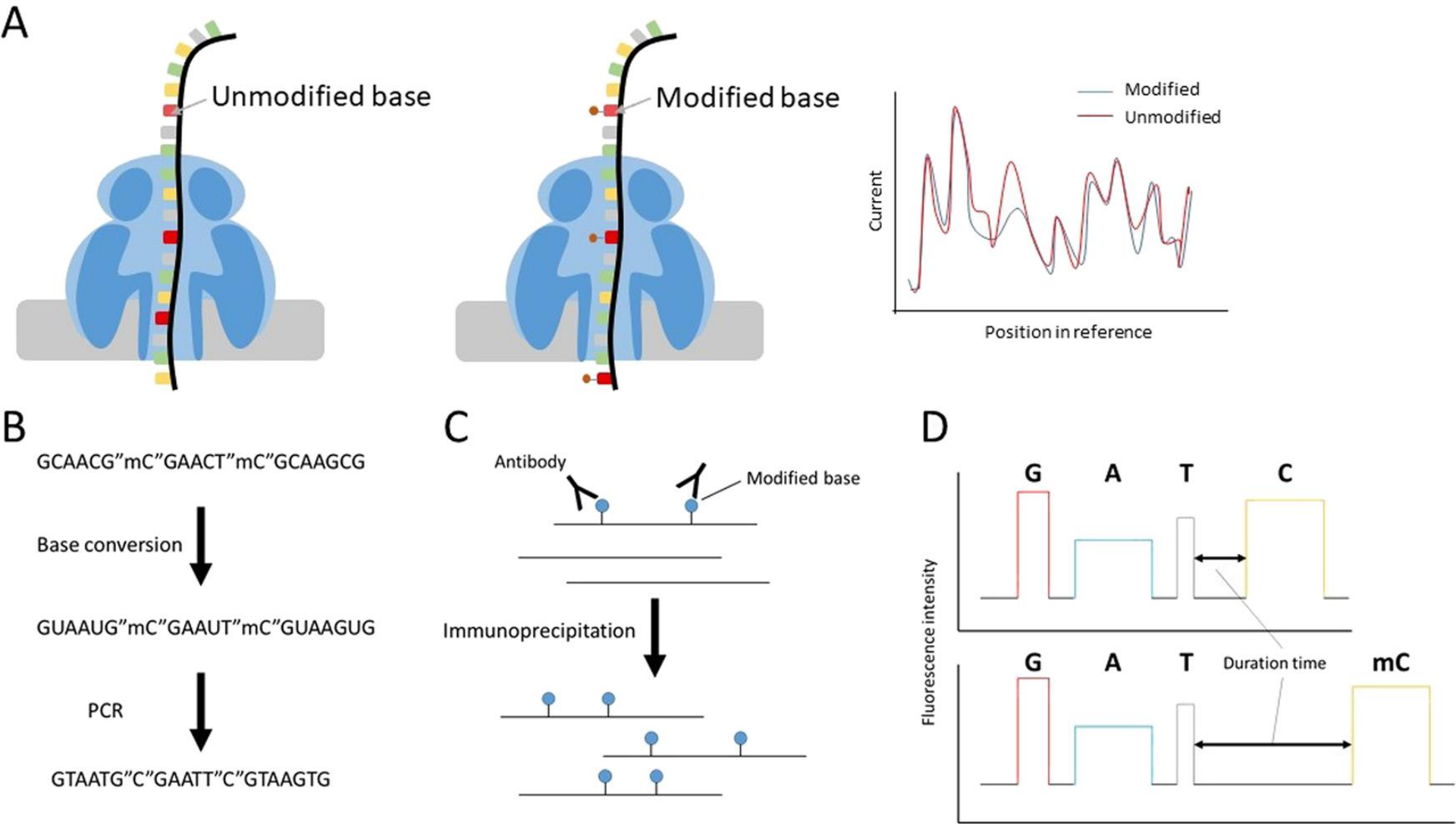
Long-read sequencing

Nanopore sequencing



Single molecule, Long reads: up to 3MB
Low accuracy, low throughput

DNA methylation



PAC Bio

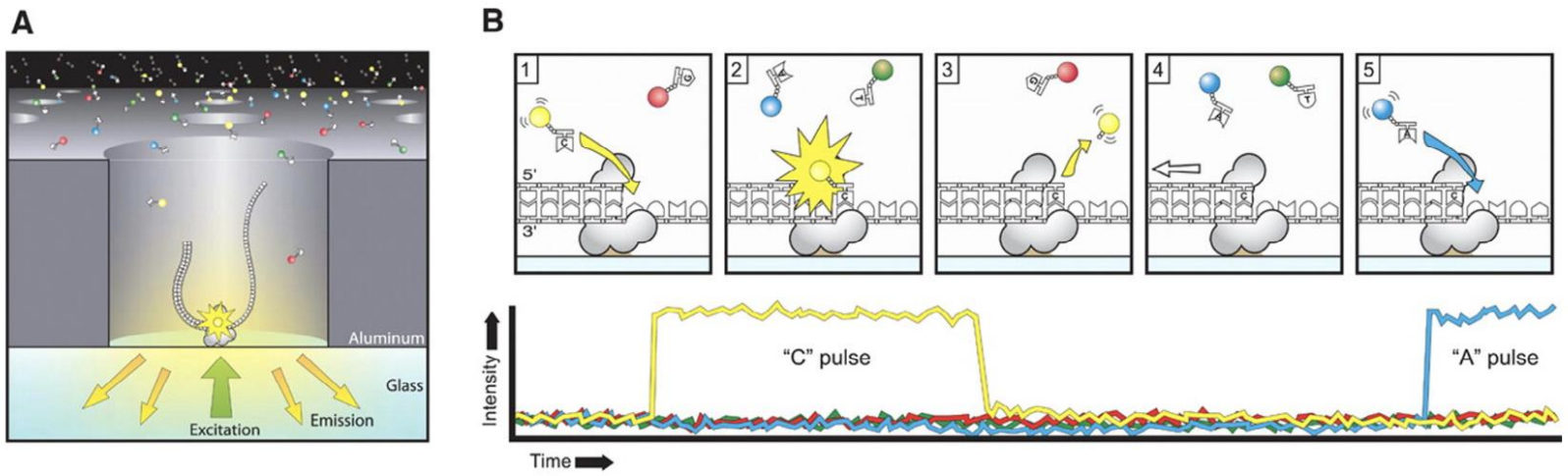
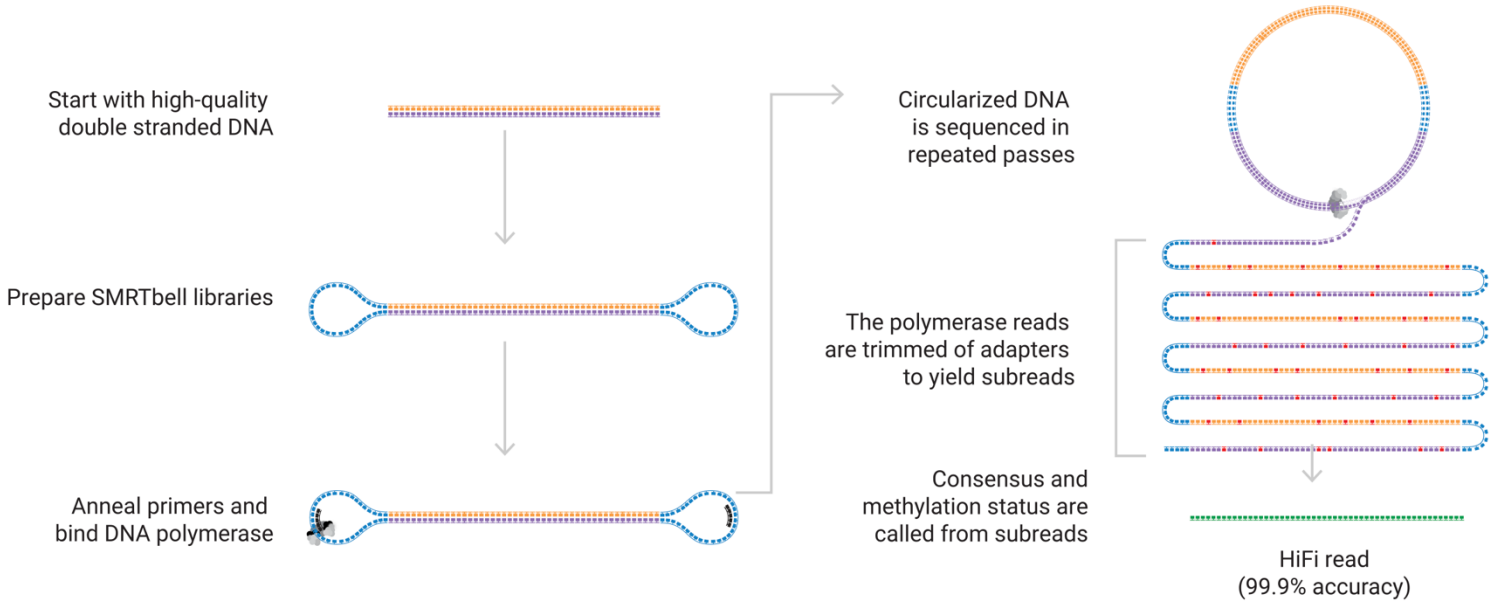


Figure 3 Sequencing via light pulses

A. A SMRTbell (gray) diffuses into a ZMW, and the adaptor binds to a polymerase immobilized at the bottom. **B.** Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue, respectively for G, C, T, and A) so that they have distinct emission spectrums. As a nucleotide is held in the detection volume by the polymerase, a light pulse is produced that identifies the base. (1) A fluorescently-labeled nucleotide associates with the template in the active site of the polymerase. (2) The fluorescence output of the color corresponding to the incorporated base (yellow for base C as an example here) is elevated. (3) The dye-linker-pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW, ending the fluorescence pulse. (4) The polymerase translocates to the next position. (5) The next nucleotide associates with the template in the active site of the polymerase, initiating the next fluorescence pulse, which corresponds to base A here. The figure is adapted from [4] with permission from The American Association for the Advancement of Science.

PAC Bio


Start with high-quality
double stranded DNA
Prepare SMRTbell libraries
Anneal primers and
bind DNA polymerase
Circularized DNA
is sequenced in
repeated passes
The polymerase reads
are trimmed of adapters
to yield subreads
Consensus and
methylation status are
called from subreads
HiFi read
(99.9% accuracy)

PAC Bio

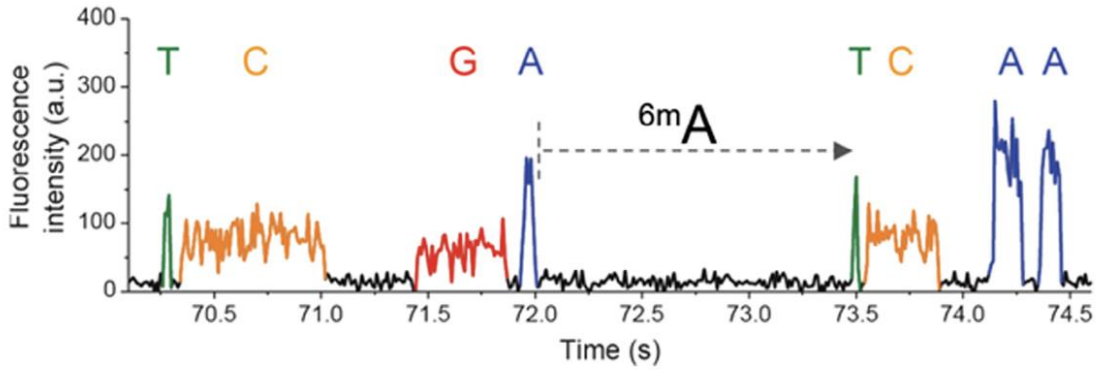
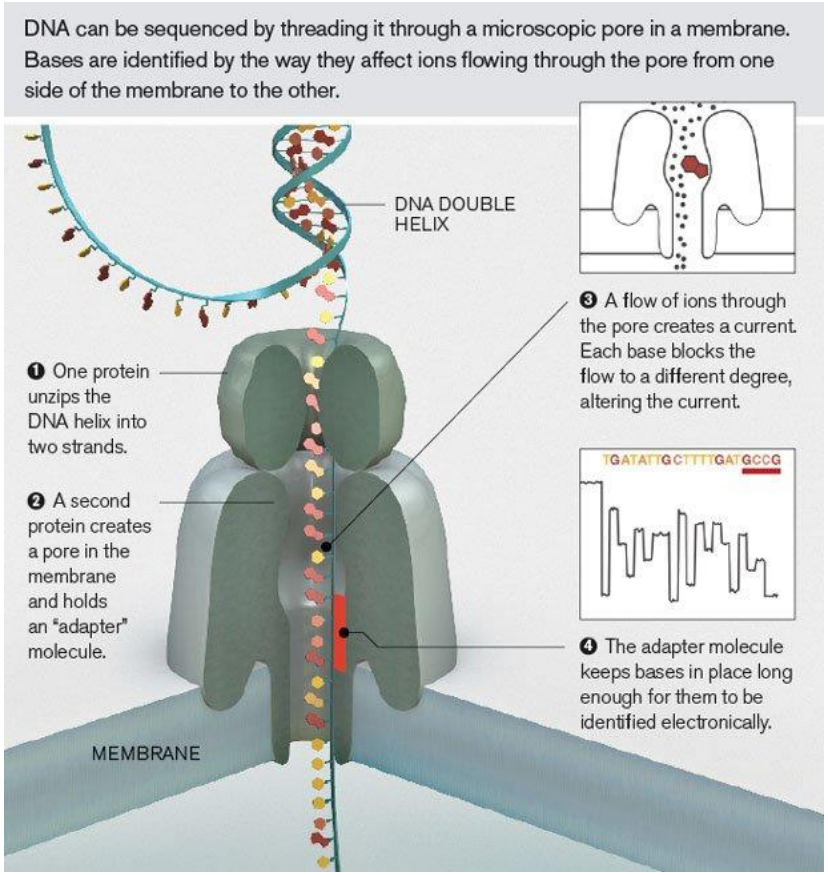


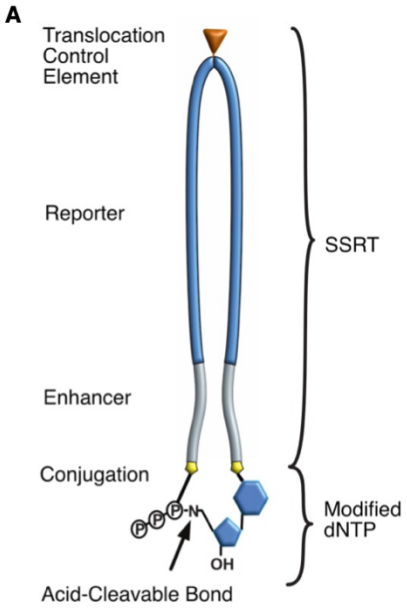
Figure 5 Detection of methylated bases using PacBio sequencing

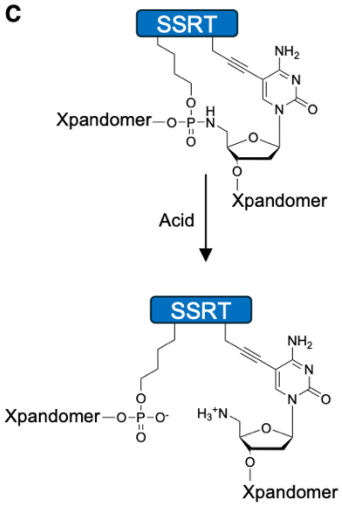
PacBio sequencing can detect modified bases, including m^6A (also known as ^{6m}A), by analyzing variation in the time between base incorporations in the read strand. The figure is adapted with permission from Pacific Biosciences [72]. a.u. stands for arbitrary unit.

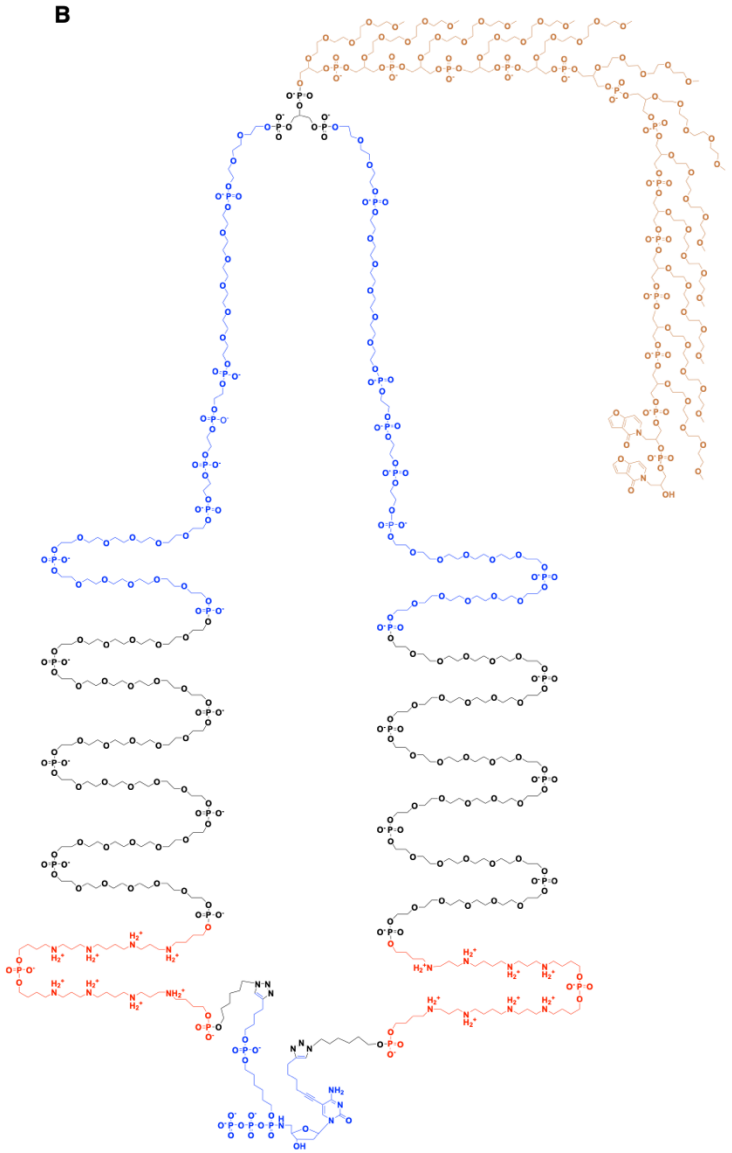
Nanopore sequencing



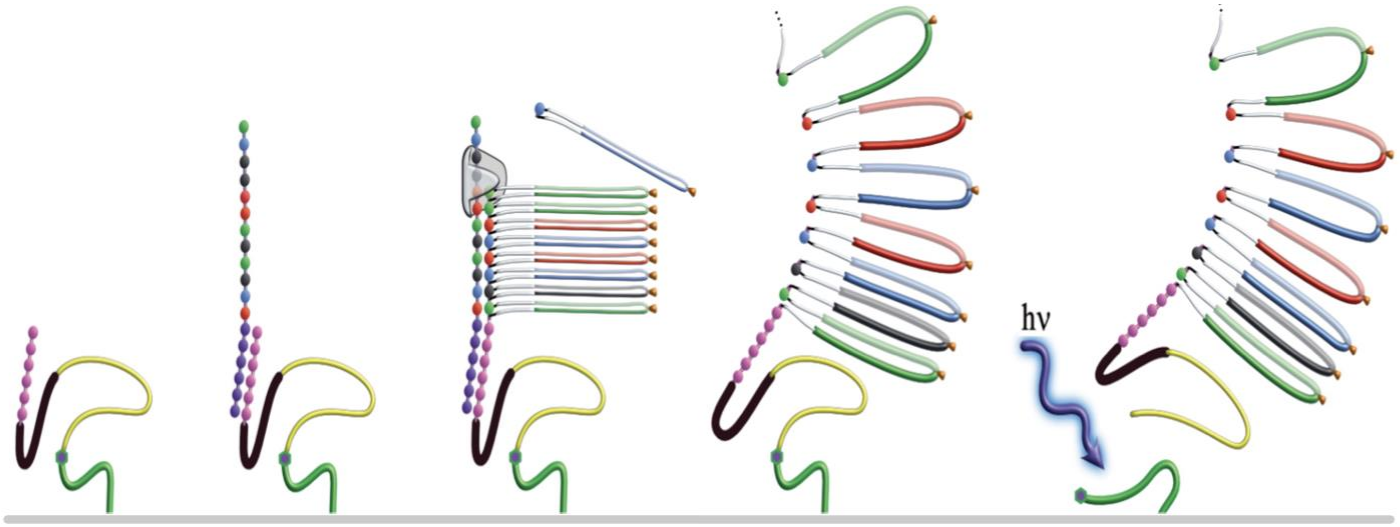
SBX Sequencing by Expandomers



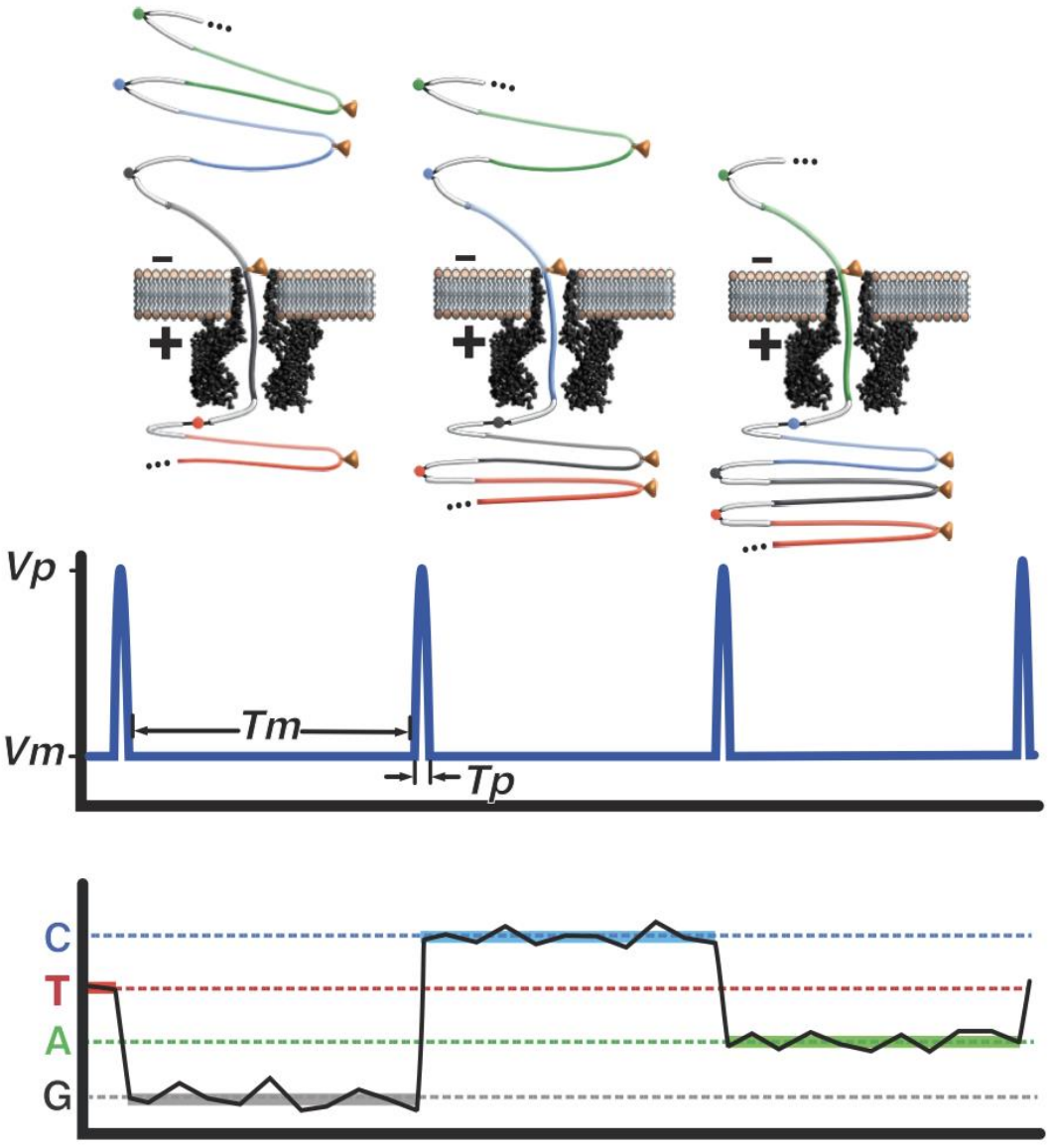




SBX Sequencing by Expandomers

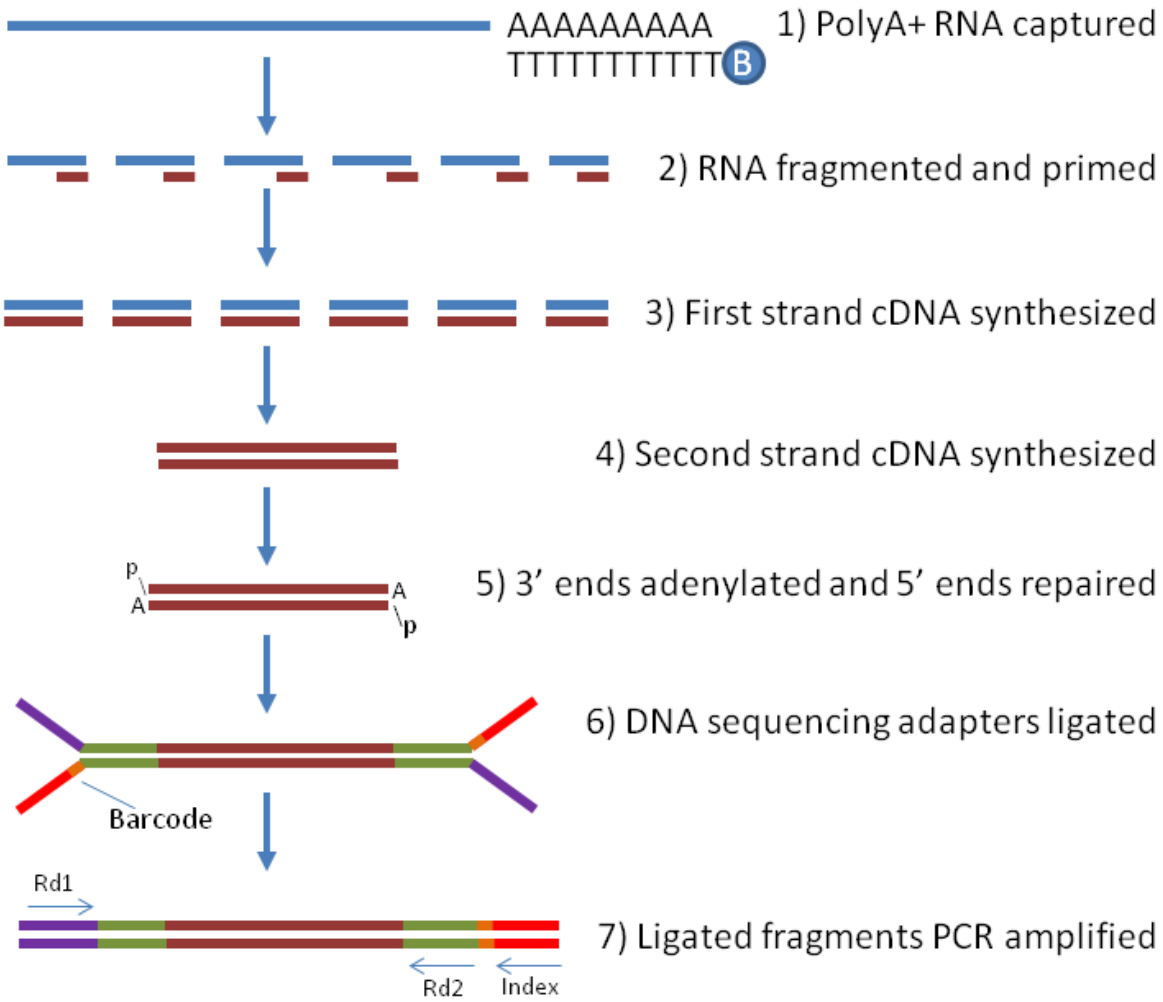


SBX Sequencing by Expandomers



Some things to consider when analyzing genomics data.

RNA-seq


AAAAAAAAA
TTTTTTTTTTB
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3’ ends adenylated and 5’ ends repaired
Barcode
6) DNA sequencing adapters ligated
Rd1
Rd2
Index
7) Ligated fragments PCR amplified

Normalization

PCR is used to amplify the library before sample submission

~20ng of library is usually sufficient for the genomics core facility to work with your sample.

20ng = 0.1 pMol of library (300bp product)

Normalization

PCR is used to amplify the library before sample submission

~20ng of library is usually sufficient for the genomics core facility to work with your sample.

20ng = 0.1 pMol of library (300bp product)

~ 60,000,000,000 molecules !

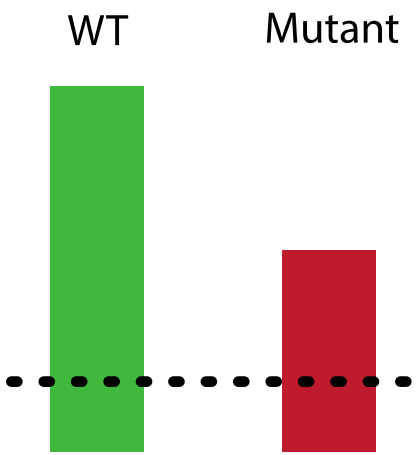
Illumina machines will now produce ~5,000,000,000 reads per flowcell

Normalization?

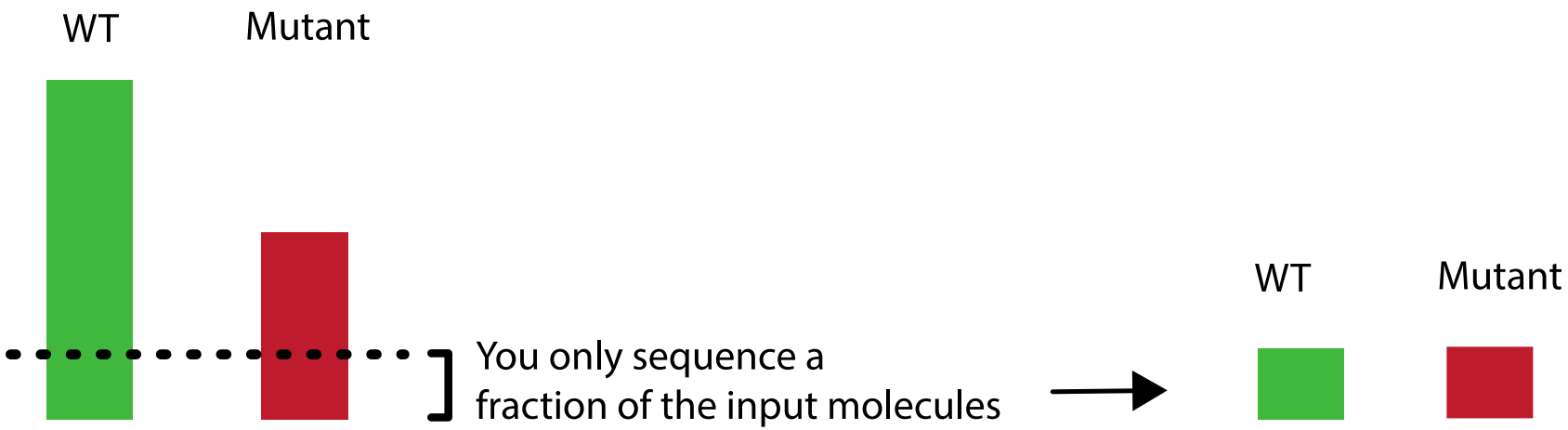
Imagine you conduct an experiment to compare mRNA in WT and a mutant of RNA polymerase II

Through other experiments you are fairly convinced that the polymerase mutant reduces all mRNA by about 2-fold.

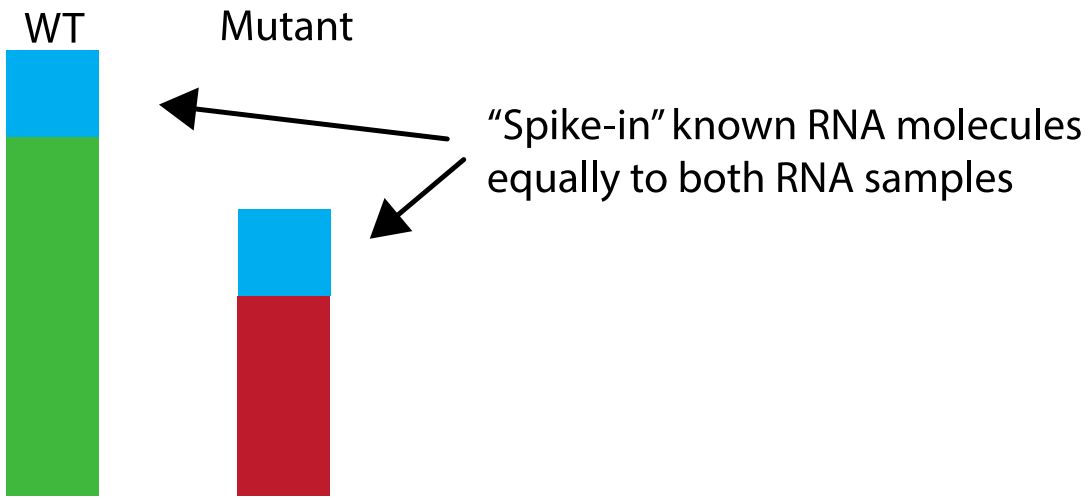
Normalization?



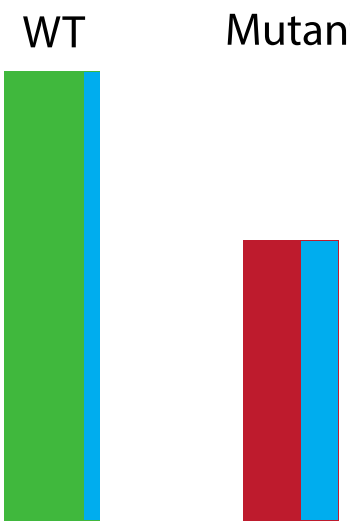
Normalization?



Normalization?

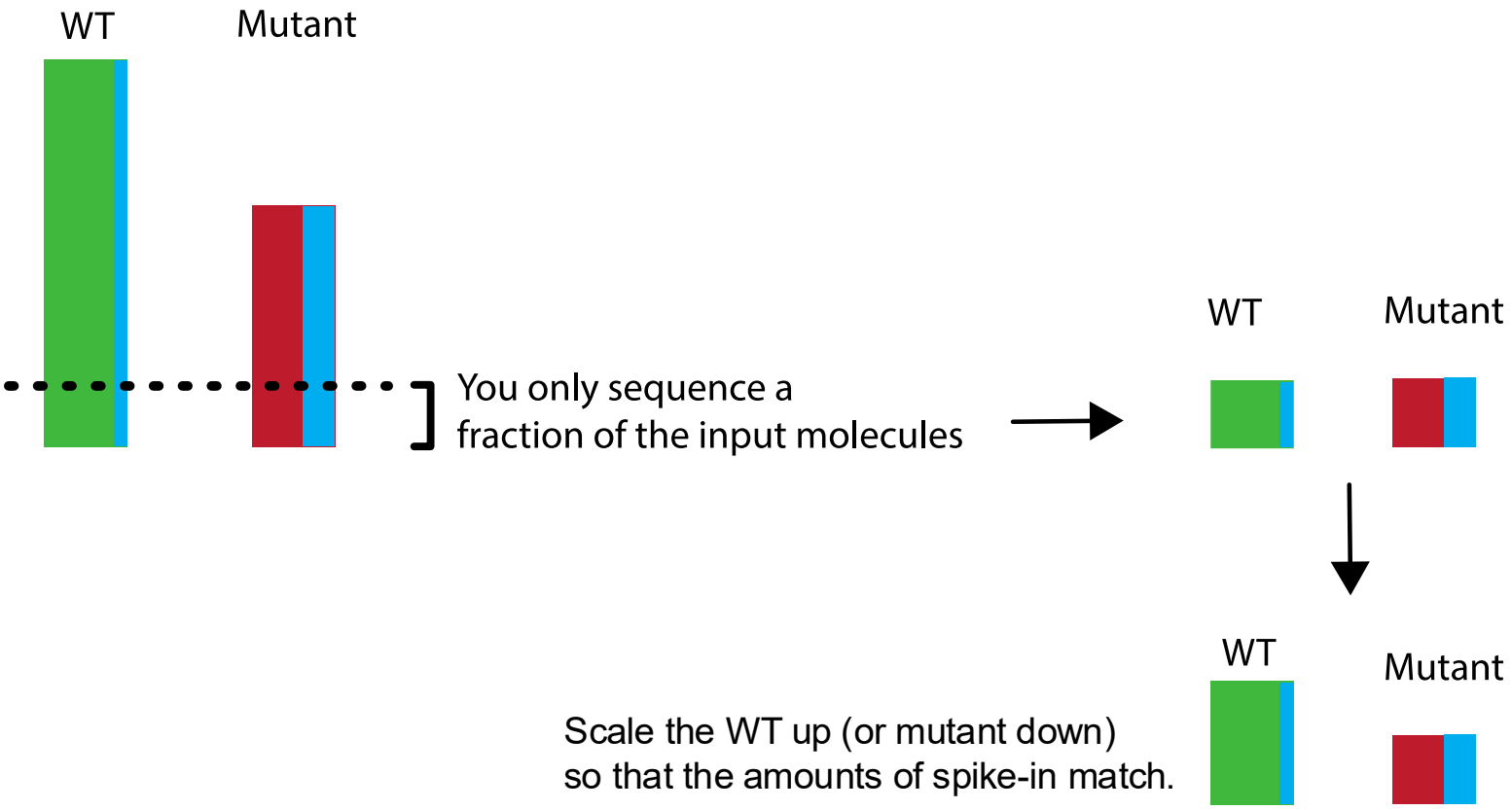


Normalization?



The “spike-in” is mixed equally in the samples

Normalization?

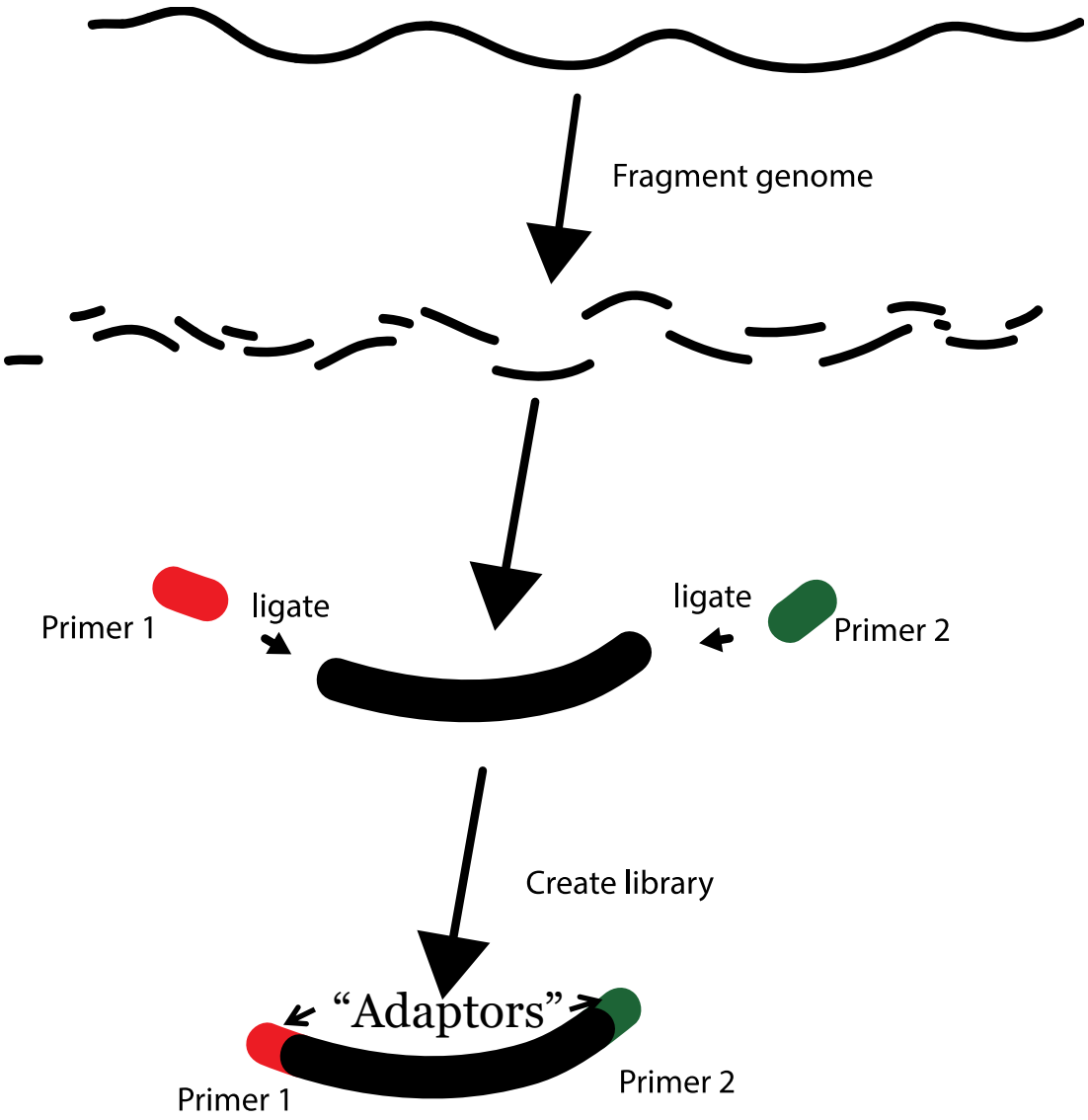


Common genomics approaches

ATAC seq

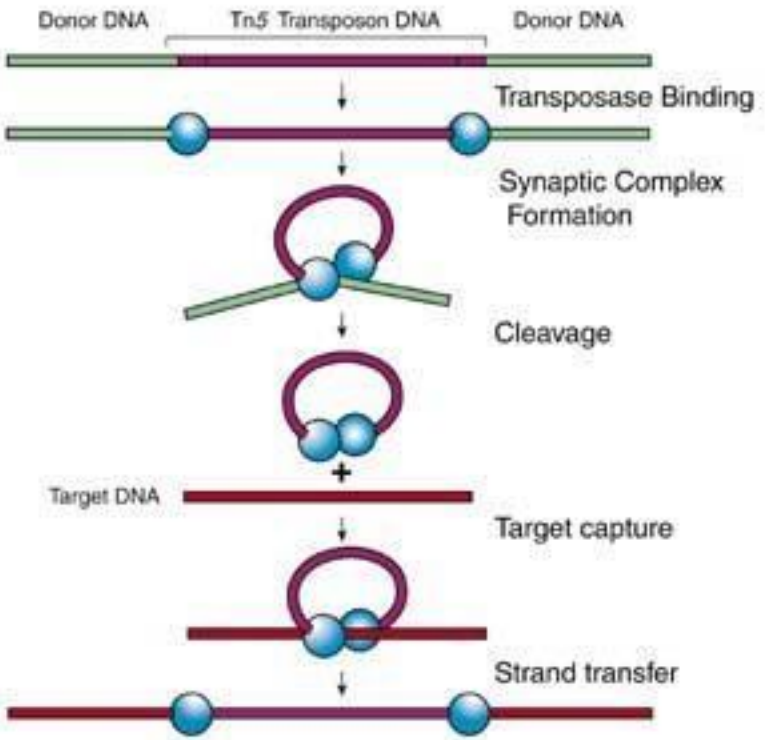
Assay for Transposase-Accessible Chromatin using sequencing

Generating a library

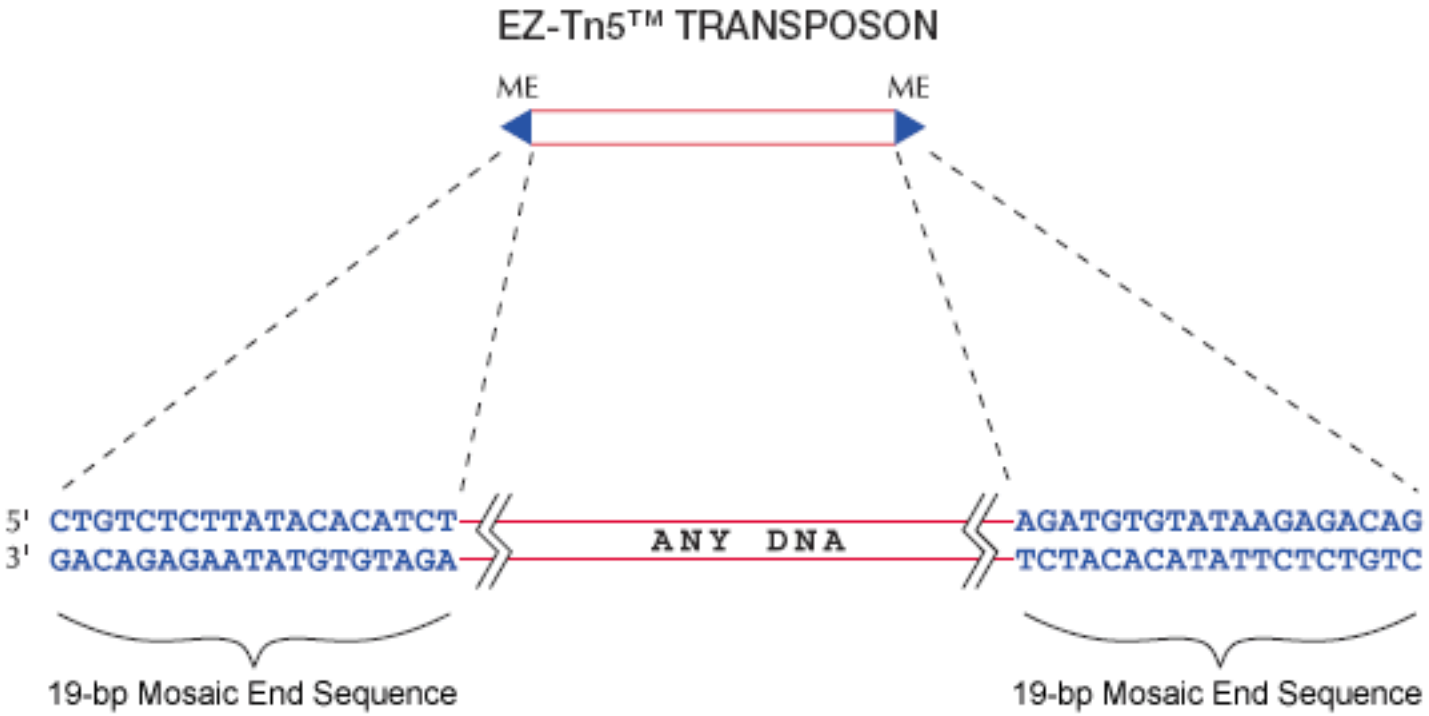


Common genomics approaches

Tn5 Transposon

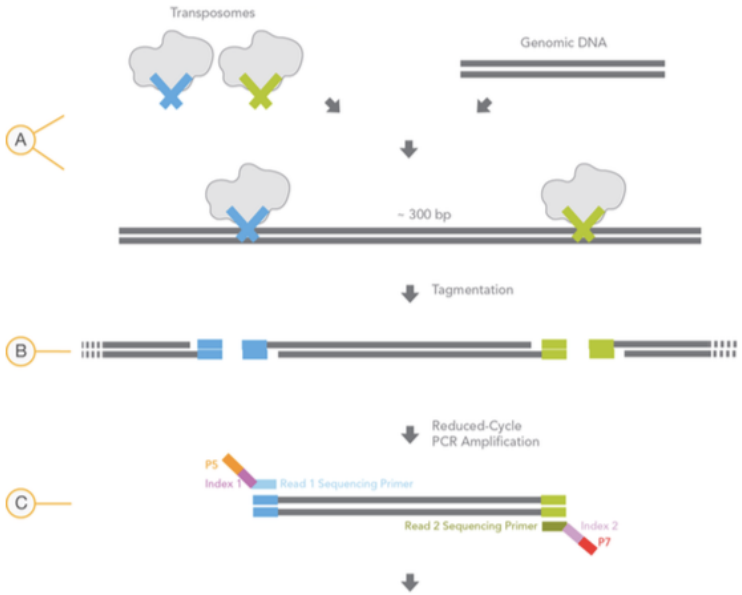


Common genomics approaches



Common genomics approaches

ATAC seq

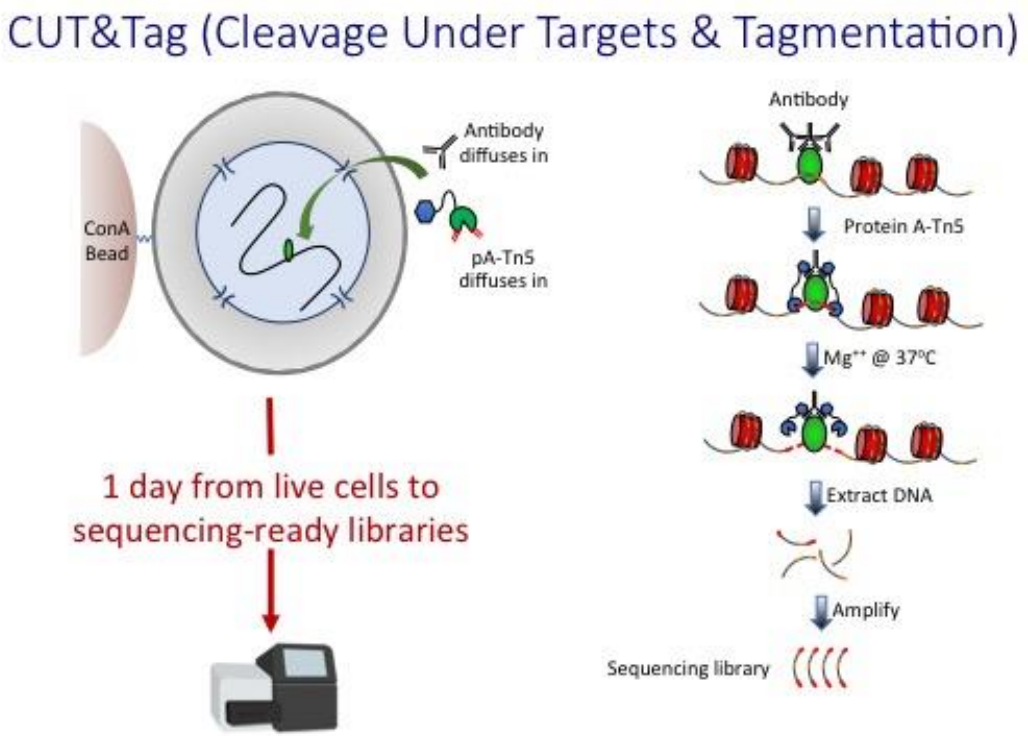


Common genomics approaches

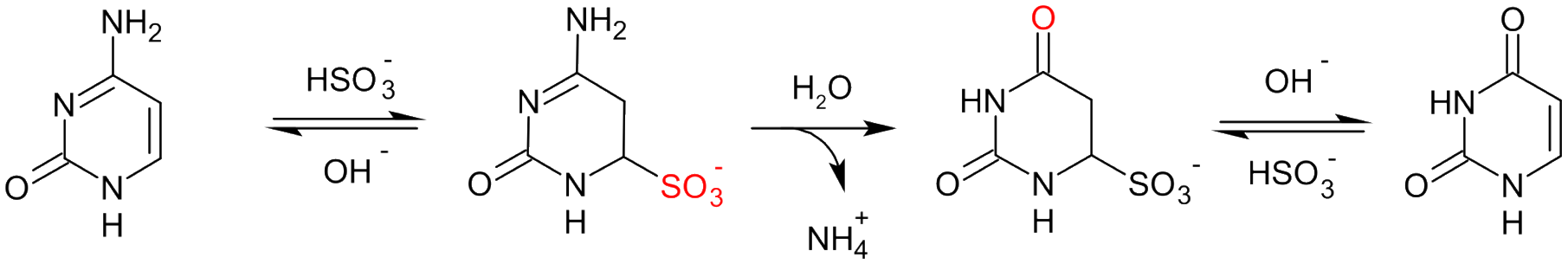
ATAC seq

Common genomics approaches

Cut and Tag
Cut and Run



Bisulphite sequencing

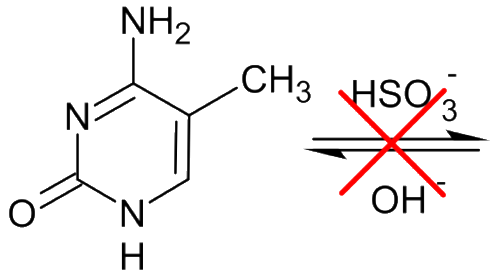


cytosine

cytosine sulphonate

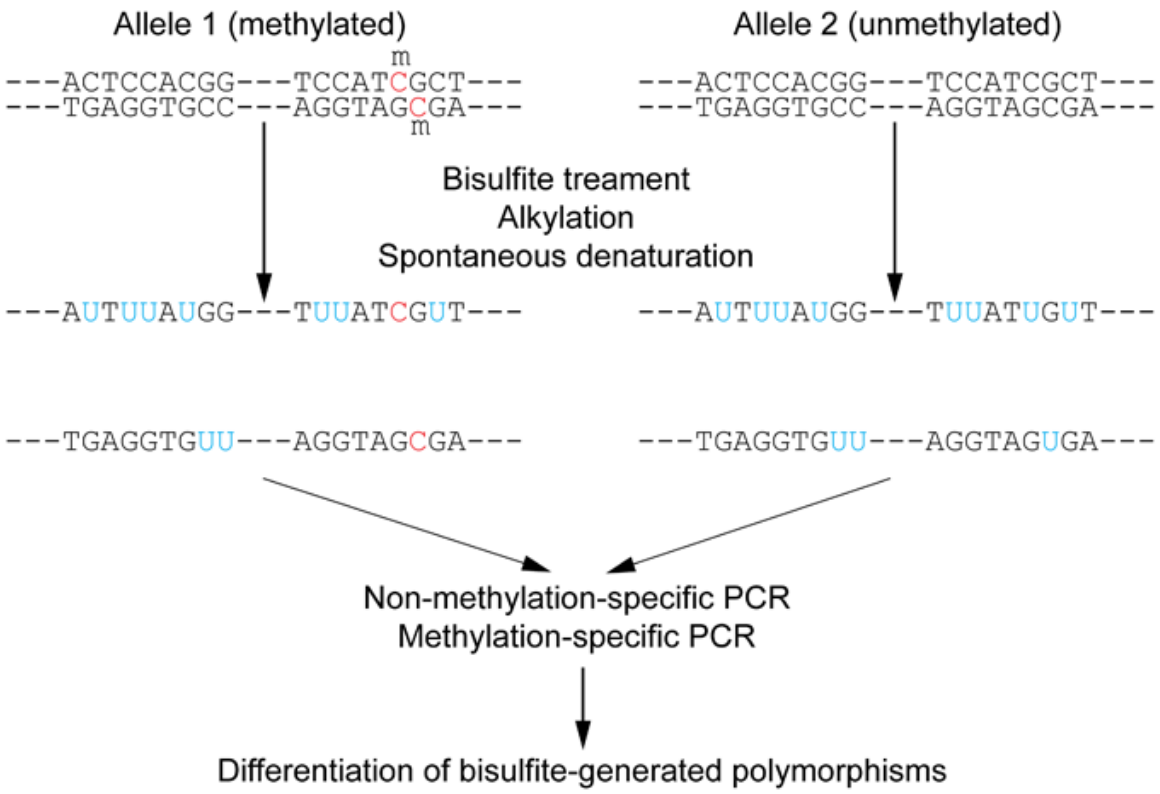
uracil sulphonate

uracil



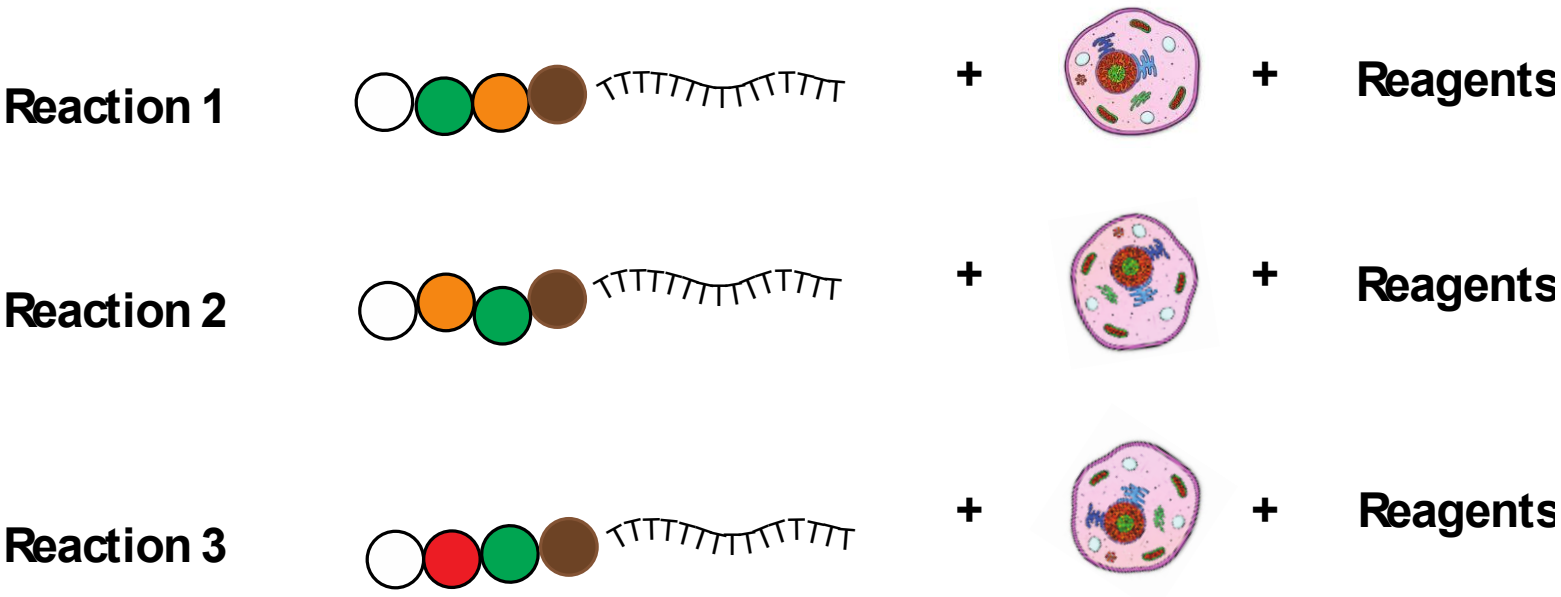
5-methylcytosine

Bisulphite sequencing



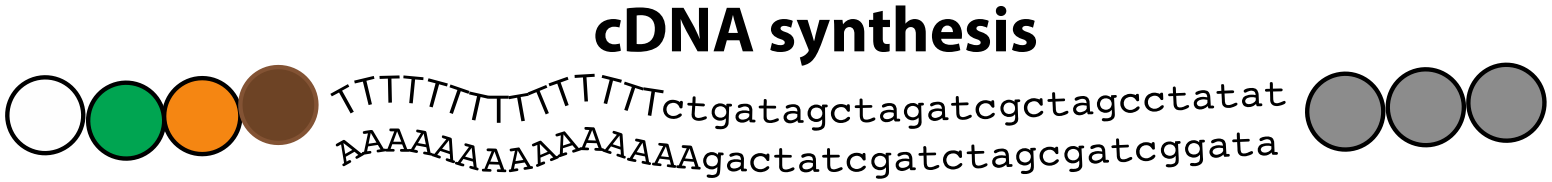
Common genomics approaches

Single cell RNA seq



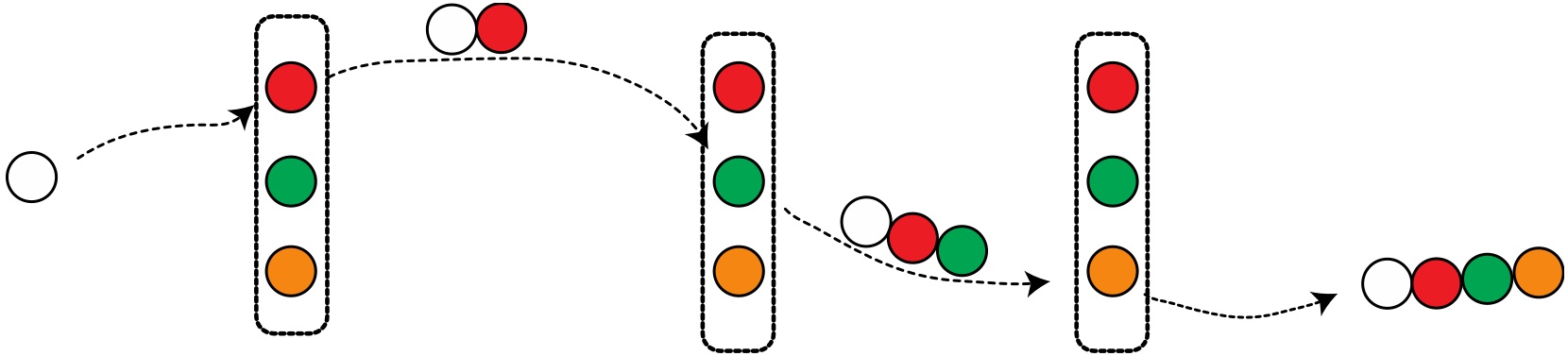
Common genomics approaches

Combinatorial indexing



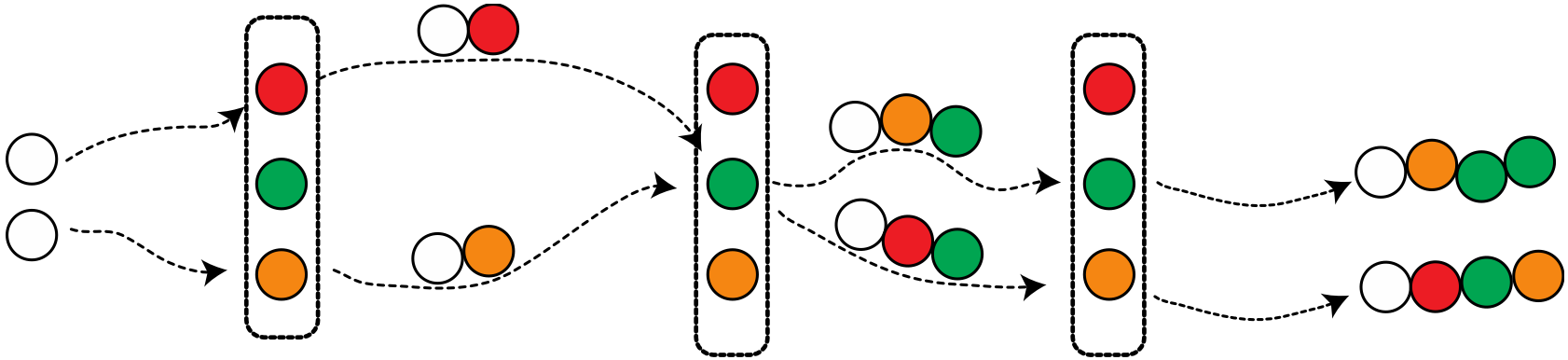
Common genomics approaches

Combinatorial indexing or Split pool barcoding



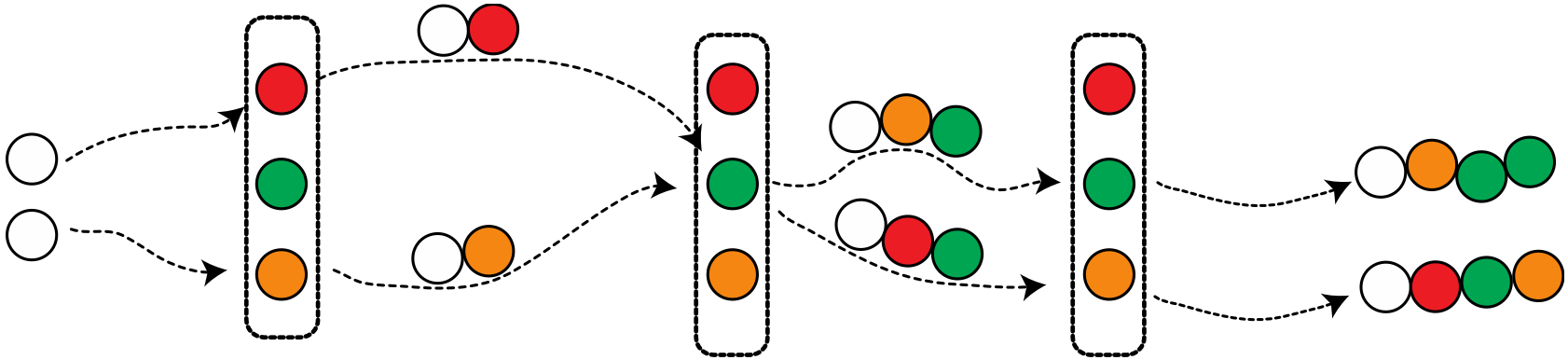
Common genomics approaches

Combinatorial indexing or Split pool barcoding



Common genomics approaches

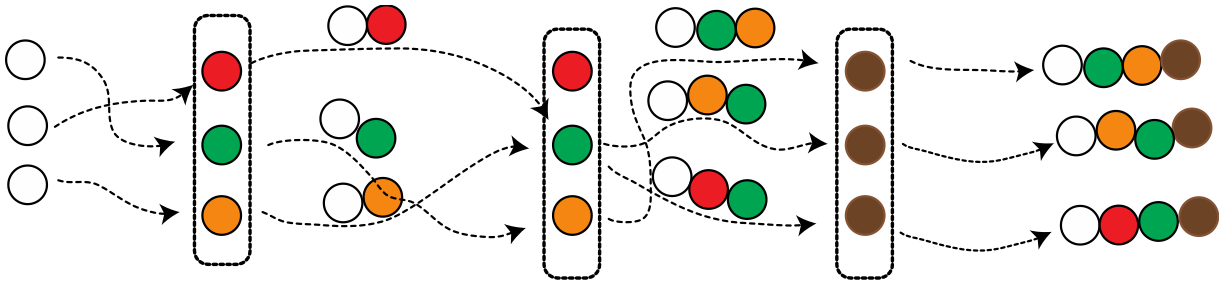
Combinatorial indexing or Split pool barcoding

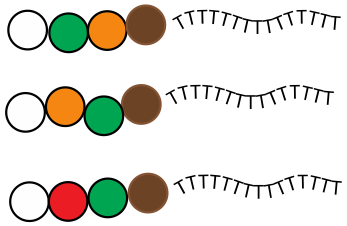


3	3	3	3*3=27 combinations
96	96	96	96*96*96=884736
384	384	384	384*384*384=56 mil

Common genomics approaches

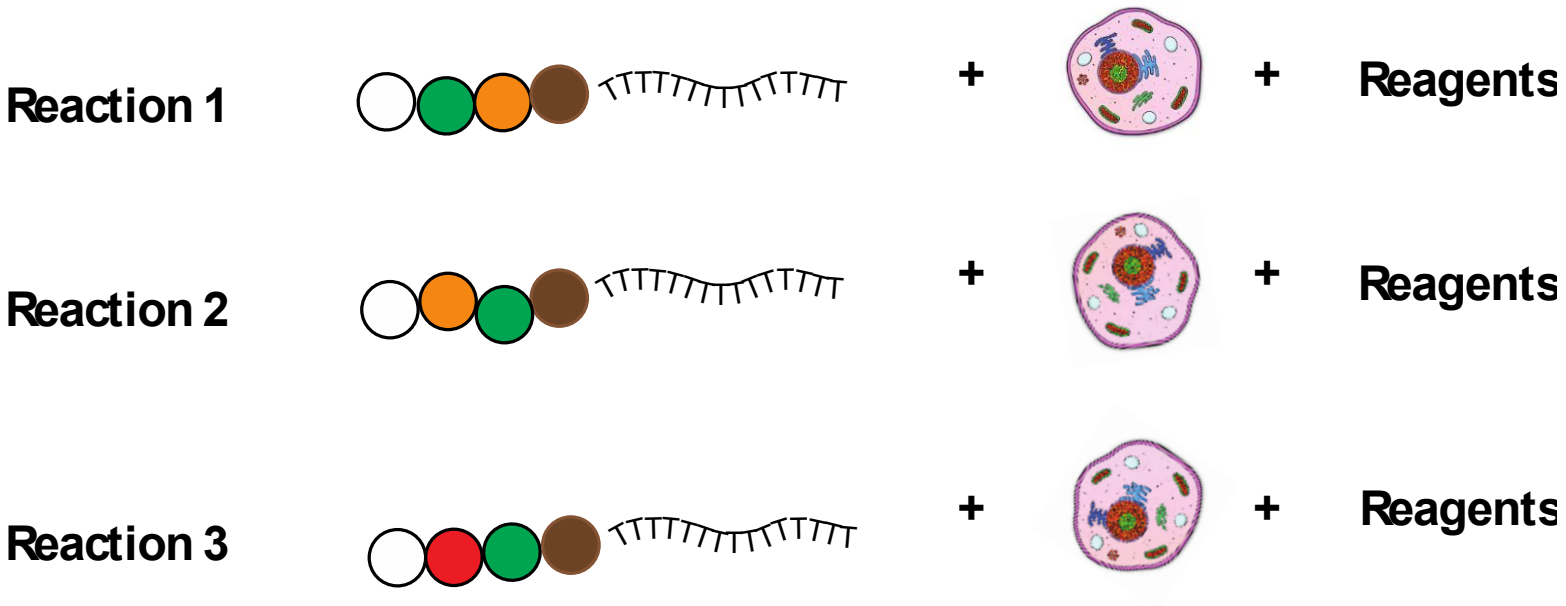
Combinatorial indexing or Split pool barcoding



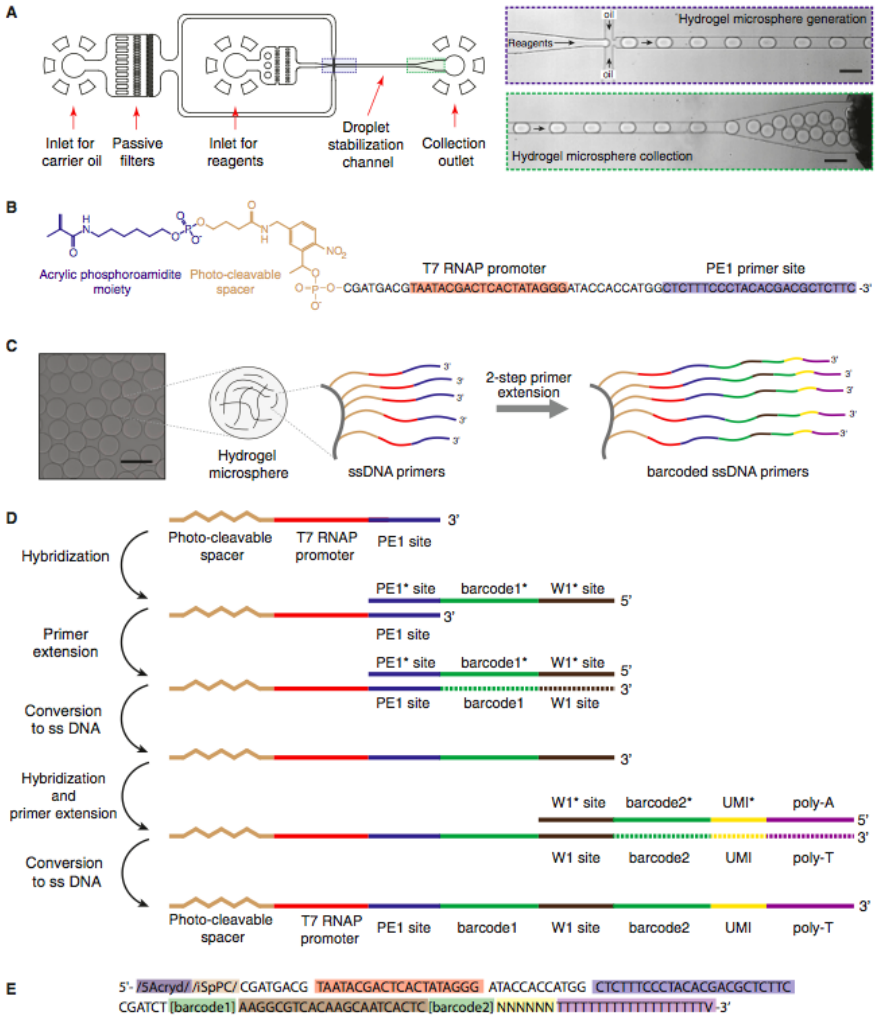


Common genomics approaches

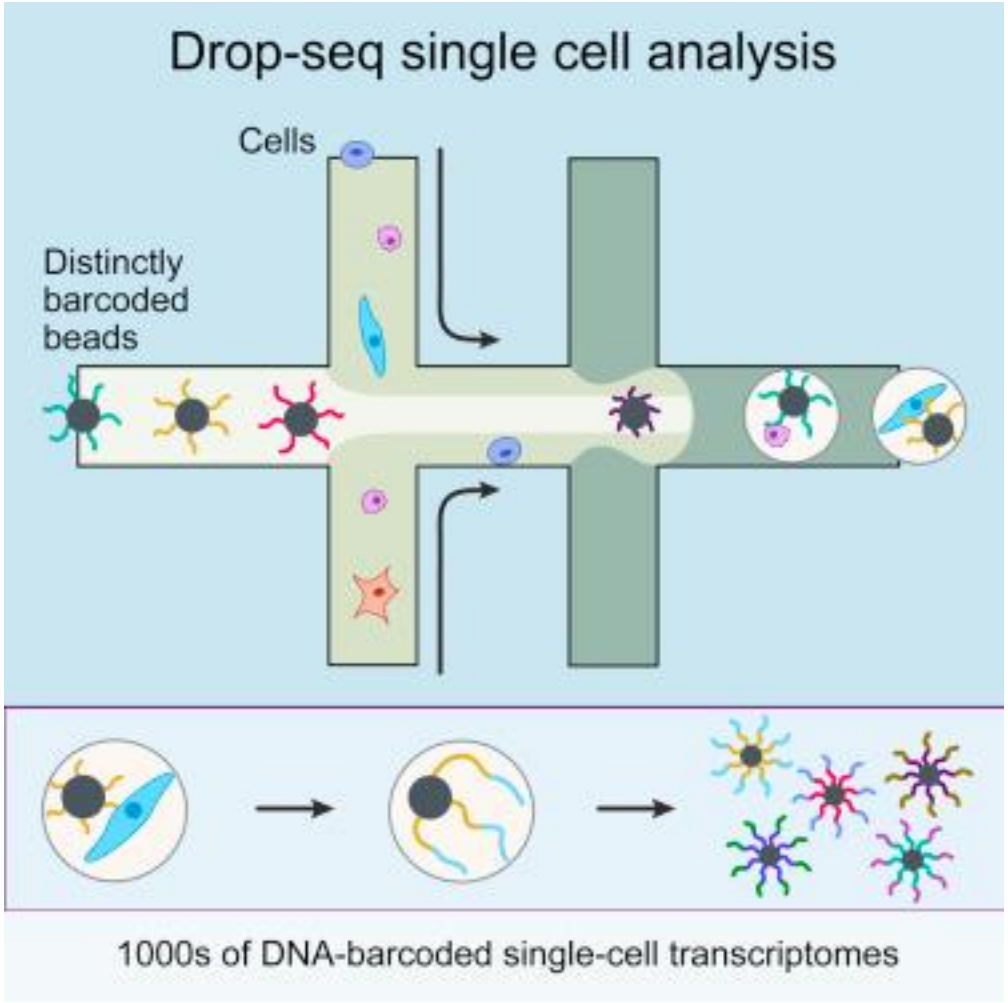
Combinatorial indexing



Common genomics approaches

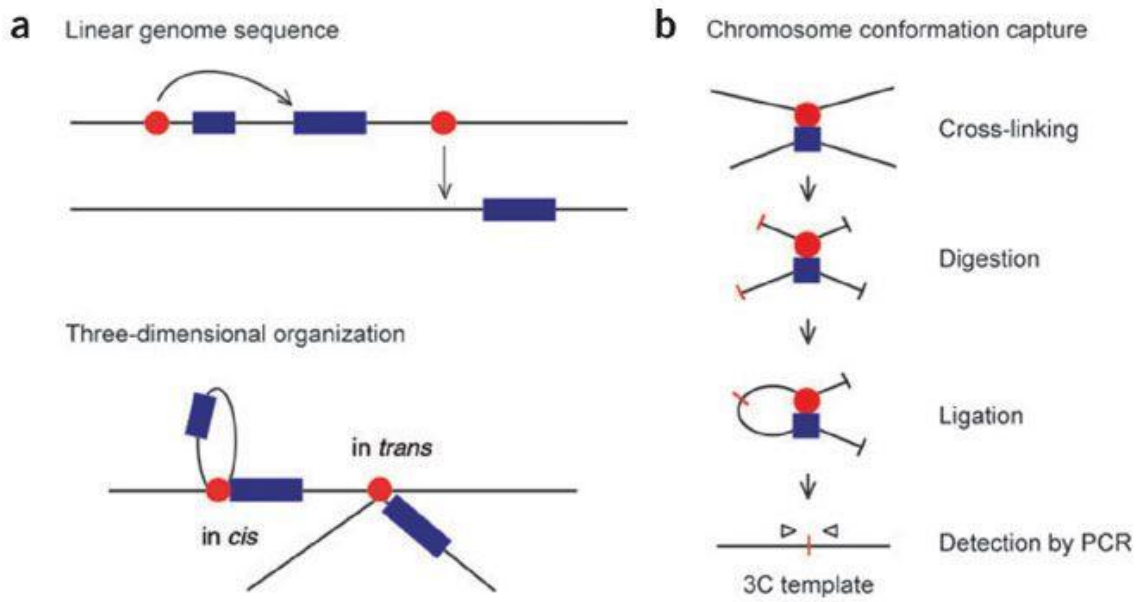


Common genomics approaches



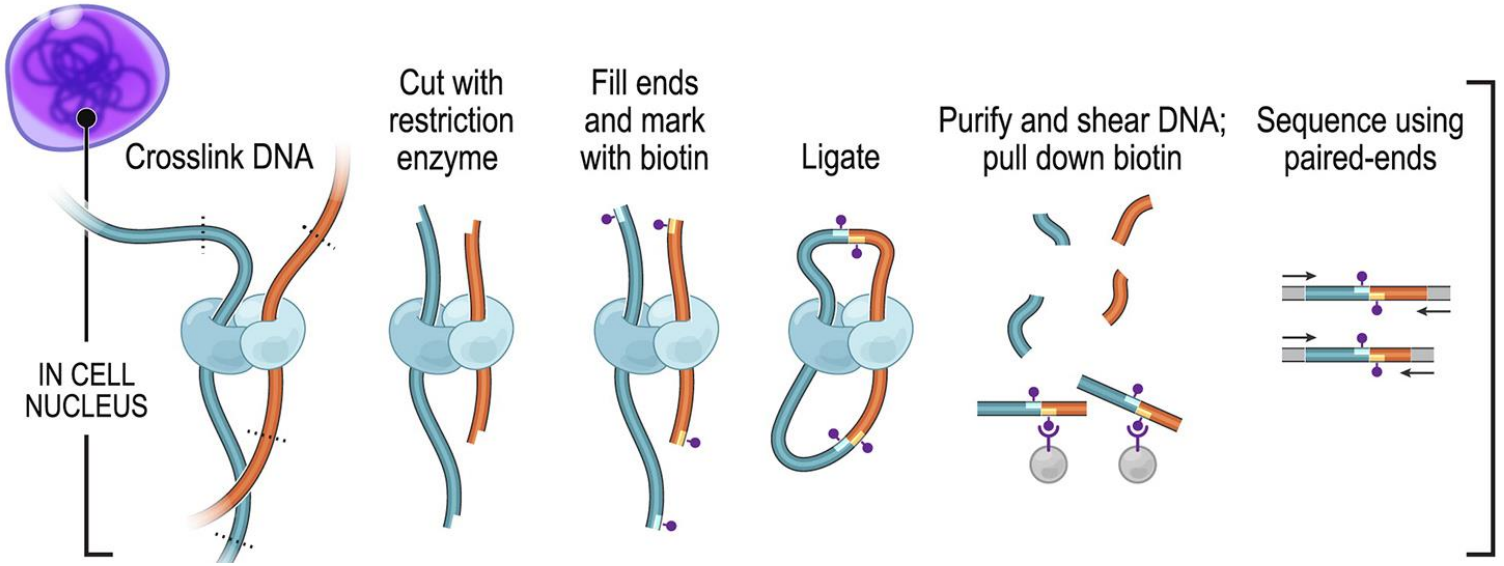
Common genomics approaches

Chromosome conformation capture (3C)



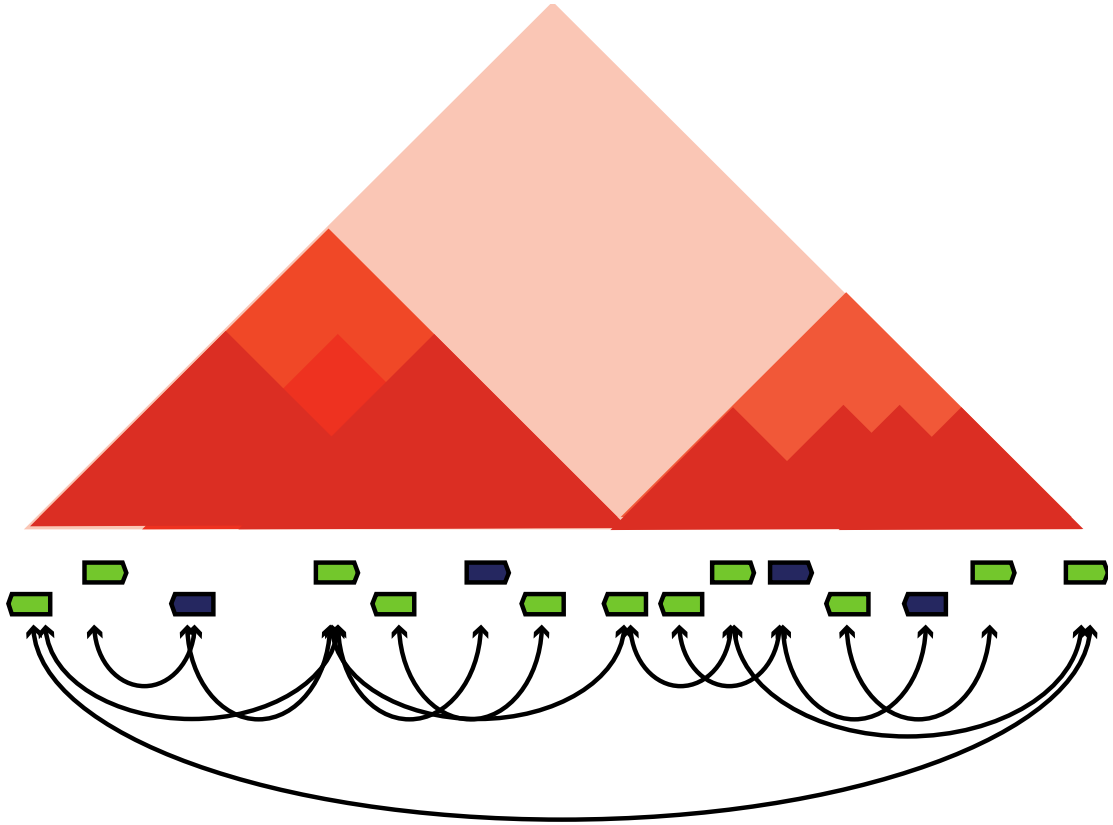
Common genomics approaches

Hi-C



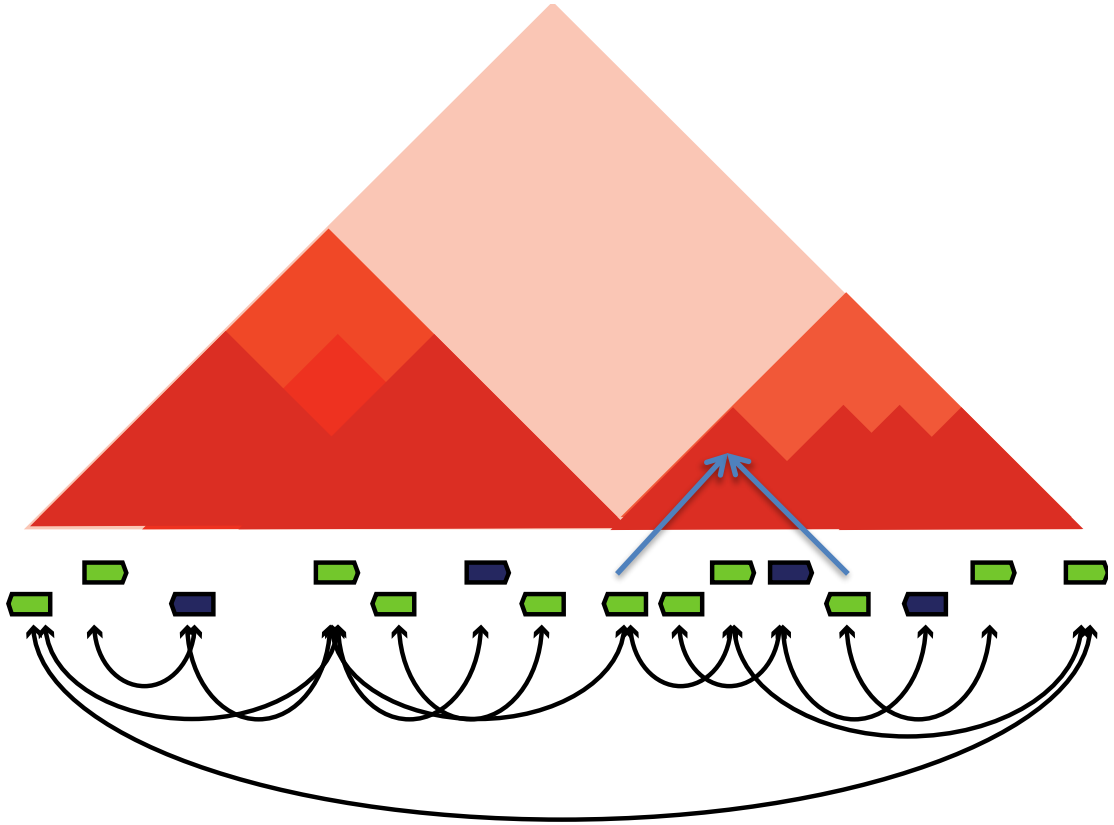
Common genomics approaches

Hi-C



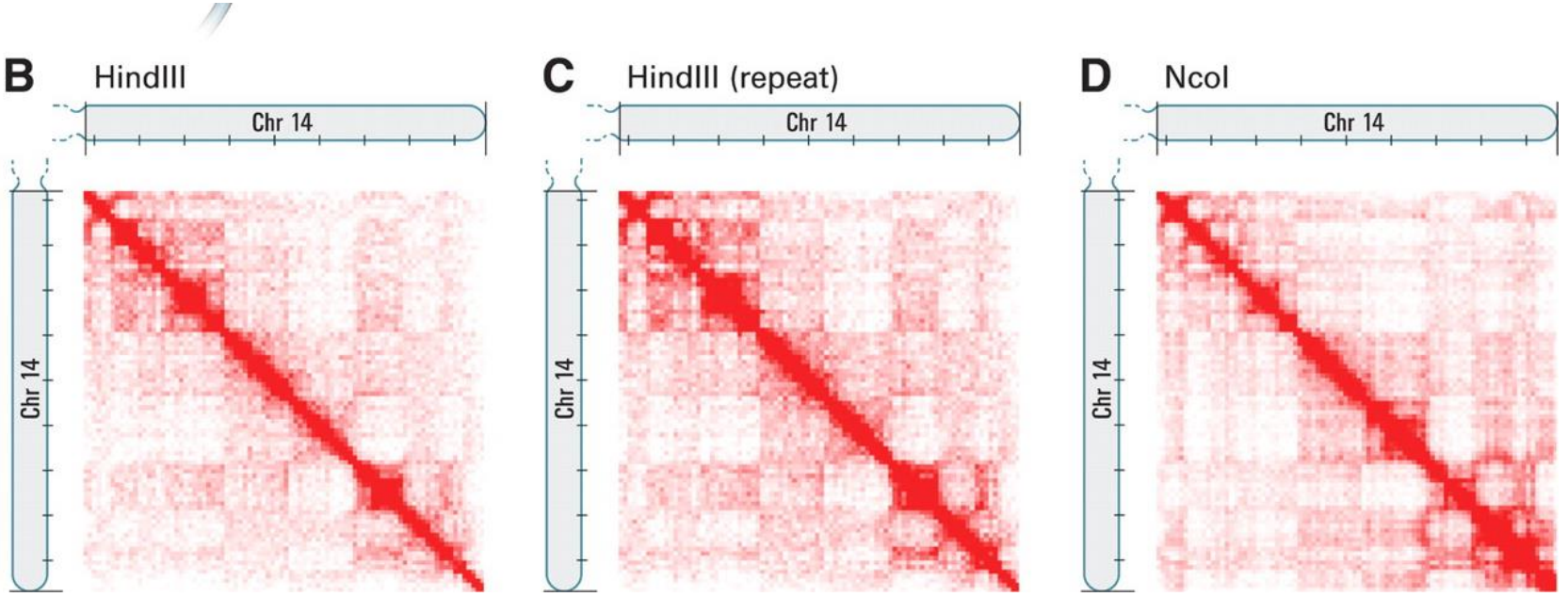
Common genomics approaches

Hi-C



Common genomics approaches

Hi-C



Ribosome-profiling