Lecture 5.

Continuation from discussion of random walks to look at models of DNA folding in chromosomes.

Later, we will look at bending of DNA and packing in very tight spaces, like viruses.

Unlike proteins, DNA molecules are extremely long (2m human, unfolded, $3.5 \times 10^9 \times 0.34\text{nm}$) and must be packed without damage + in safe topology without knots.

1. Chromosome Conformation Capture (3C). p337
   i) Take cell in state of interest
      Crosslink everything with formaldehyde
      - acts on proteins, eg histones
   ii) Restriction enzyme to cut DNA into known fragments.
   iii) Ligate to generate hybrids from both sides of contact. ⌈ shear into short strands
   iv) Amplify DNA using PCR (optional) and
   v) Sequence to see genetic location of original contact.

2. Restriction Enzyme
   Discovered in prokaryotes for use as a defense mechanism against virus infection.
   Also found in archea.
   >3000 exist >600 commercial
   EcoRI in E. Coli etc.
   Cuts DNA at a specific sequence, type I / II
   usually 4bp or 6pp long. ↗ random location ↖ in middle of sequence

Can leave either "blunt" or "sticky" ends.

Strategy: bacteria evolves its own genome to never use the restriction sequence.
Any other organism's DNA will be cut into characteristic "restriction fragments"

Eg apply to human DNA: 4kbp)

$\frac{3.5 \times 10^9 \text{ bp}}{4096 \text{ bp}} = 800{,}000$ fragments (of length

4^6 possible sequences of 6 basepairs.
One out of 4^6 will be cut.
Assuming random distribution.

3. Polymerase Chain Reaction (PCR)
Used to rapidly copy DNA strands by doubling as much as needed.
Used in forensic science to amplify tiny amounts of DNA, also dinosaurs, mamoths, dodos and ancient remains.
Based on TAQ polymerase (or engineered modifications) found in Yellowstone hot springs Thermus Aquaticus. (TAQ) lives at 72°C

add: Deoxynucleotide TP's + primers
Divalent cations Mg^{++} & K^+

slide cycle between 3 temperatures in "thermal cycler"
95°C for 20 sec denaturation melt ds → ss
55°C for 20 sec anneal primers
72°C for 1 min / 1000 base pairs elongation

Primers are short ss strands of DNA 18-20 bases long; often synthetic.

Polymerase runs 3' to 5'

so a primer is needed for reverse strand too.

4. Sequencing.

Major industry of 21st century.

Started with Sanger method in Human Genome Project. 1990 – 2003.

Overtaken by Celera Corporation under the leadership of Craig Venter.

Company takeovers every 12 months.

Cost has dropped from $2500 / Mbp Sanger to. $0.1 now.

Current market leader is Illumina with 90% of all data produced 70% of sales.

Uses very short reads 50-300bp long. 800bp Sanger.

Lower accuracy 98% made up by redundancy, but also helped by using a reference genome.

slide

Throughput 1x human genome / hour.

Innovations:

i) Sequence during synthesis.

ii) Reversible terminator bases (RTB) containing coloured dye (4 colours). After read, dye is washed away.

iii) Randomly located colonies on large 8-lane plate. Many samples mixed together instead of separate test tubes.

movie

iv) Bridge Amplification

v) Adapters like bar-codes to sort out samples retroactively.

Very clever technology shown in fast-running movie from Illumina company.

5. Magnetic beads.

Hi-C method uses one additional step: addition of biotin marker just before Ligation step.

Biotin = small molecule = vitamin H

binds very specifically (8 H-bonds $\times 8 k_B T$) to Avidin (raw eggs) } protein.
Streptavidin (bacteria) }

Attach protein to 1-5 μm diameter magnetic beads.

Use magnet to separate from rest of solution + wash contamination.

6. Hi-C method.

3C → 5C specific primers for one chromosome
→ Hi-C to chop into short strands for next-gen sequencing.

Restriction fragments ~ 4kb.
Shear down to < 300 bp.
How short can be the fragments to be uniquely recognised in genome?

$3.5 \times 10^9 = 4^M$

$M = \frac{\log 3.5 \times 10^9}{\log 4} = 16$ bp. (on average)

16 bp is enough to be unique in genome
hybrid of two sequences will give two unique loci if it has > 50 bp.
Matches next-gen sequencing.

But hybrid ends are restriction fragments
– these are all known in genome.
– limits resolution of HiC:
6 bp restriction → 4096 bp fragments
4 bp " 256 bp fragments

We will look at the results from a few recent papers:
Lieberman-Aiden et al Science (2009)
Naumova et al Science (2013)
Rao et al Cell (2014)
Lajoie et al Methods (2015)

slide
→ processing steps.
Sinkorn-Knopp balancing.
binning to lower resolution 25kb → few kb today
need ~10 events per cell of heat map.

7. cis / trans chromosomes

cis interactions (within chrom) below.

slides trans interactions indicate contacts btw chromosomes in the interphase initial state.

Associate with pattern of chr. territories.

Some chrs are often close to certain others.

The smallest, 13–22, are clustered in the center of the nucleus.

8. Genomic compartments.

Checkerboard pattern on 1–10 Mb length scale

Use princ. component analysis to factorise

A (blue) euchromatin gene-dense GC rich

B (red) heterochromatin non coding AT rich

Correlates with chromosome banding:

slides chemical labelling of GC + AT rich regions seen under microscope.

Not clear why like-with-like pattern emerges: A's associate with A's.

9. Topologically Associating Domains (TADs)

Square blocks along diagonal, no checks.

Smaller, sub Mb size.

Believed to be the functional units of gene expression + co-regulation.

TAD boundaries are associated with insulators that separate the logical units. Maybe that whole TAD block folds and unfolds to activate its genes.

10. Point interactions
Off-diagonal points where specific sequences interact. Associating genes or parts of genes?
At higher resolution: loops of DNA (below)

9.02.15

11. Bacterial TADs
Le Ting et al Science (2013).
Opposite diagonal shows circular genomes
Blocks show associating structures.

12. Probability of interaction
Prob (interaction) vs marker separation
Direct projection on diagonal of heat map.
slide Yeast has expected $L^{-3/2}$ power law
- corresponds to 3D random walk.
- prob. of return to origin (crosslink) $= \left(\frac{2}{\pi N}\right)^{3/2}$

Human chromosomes follow L^{-1}.
Confirmed for single cells (Nagano et al)
slide movie Naumova et al (2013) looked at different phases of the cell cycle:
Metaphase, when structure is most ordered:
- no TADs, continuous along diagonal
- $P(L) \sim L^{-1/2}$ in metaphase up to 10 Mb.

confirmed by other studies using Hi-C.
Beyond 10 Mb, sharp drop, independent of which chromosome: Chr 1 ≈ 250 Mb. } human.
Chr 22 ≈ 50 Mb }
Need to explain break in structure.
10 Mb may reflect "compartment" structure as seen with banding.

13. Polymer Models

First, lets look at the 10 Mb unit.

ν = 33 bp/nm our number. for 10 nm fibre

ξ_p = 30 nm (book) or 70 nm (Naumova)

L = 10 Mb / 33 = 300 μm of 10-nm fiber

$\sqrt{\langle R^2 \rangle} = \sqrt{La} = \sqrt{2L\xi_p}$ = 4.2 μm end-to-end.

$R_G = \sqrt{L\xi_p/3}$ = 1.7 μm = size of ball.

Naumova models as polymer obeying Langevin dynamics:

77 Mb of DNA ~ size of chr 17: 1 μm long 0.5 μm dia

128,000 monomers of 600 bp = 3× nucleosomes

ξ_p = 4 monomers = 2400 bp = 70 nm (above)

slide

i) linear model. A

Constrain monomers to lie along a line with σ = 120 nm standard deviation, along axis.

Forces cylinder of correct diam, but allows too much mixing along axis → flat P(s).

Correct drop at 10 Mb, defining compartment.

ii) Kink at 10^5 bp → R_G = 170 nm represents free random walk up to size of constraint.

ii) heirarchical model B.

fold within folds like a fractal. Adjust sizes to get correct overall shape, but interactions too local. 3 level heirarchy.

iii) loop models C & D work.

harmonic bonds between anchors (sequence)

C loop = 80 kb along linear scaffold

D loop = 120 kb without scaffold

movie

adjusted to give correct overall shape:

10^5 bp loop has R_G = 170 nm

exponential loop distribution.

14. Recent higher-resolution
1-5kb resolution by Rao et al (2014)
See many point interactions.
Loops of 100-200 kb in length.
CTCF = CCCTC binding factor.
binding sites at boundaries of loop
(known from sequence).
CTCF associates with cohesin in formation
of scaffold in metaphase.
So would expect to see in metaphase as well?