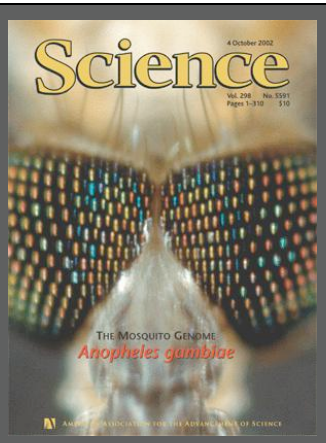
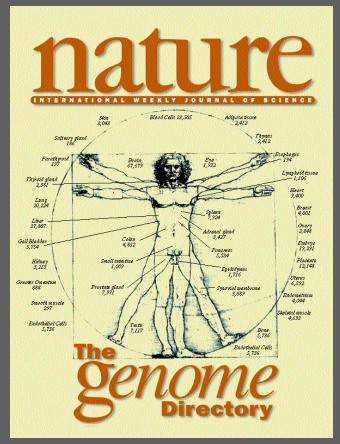

Science
THE HUMAN GENOME


nature
STAR FORMATION
A massive protostar unveiled
CANCER IMMUNOLOGY
How tumours dupe T cells
AIR POLLUTION
China's NO2 build-up seen from space
NATUREJOBS
Membrane proteomics
THE CHIMPANZEE GENOME


nature
The mouse genome


Science
4 October 2002
THE MOSQUITO GENOME
Anopheles gambiae


nature
Dictyostelium genome
Sequence of a lab classic


nature
INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
The genome Directory


nature
QUANTUM MEMORY
Controlling single photons
THE GENETICS OF NONSENSE
A cellular balancing act
IN PURSUIT OF PLEASURE
Dopamine's role revisited
THE DOG GENOME
NATUREJOBS


nature
The rat genome
Insights into mammalian evolution


Science
24 March 2000
The Drosophila Genome


nature
Citrus pathogen sequenced


nature
the human genome

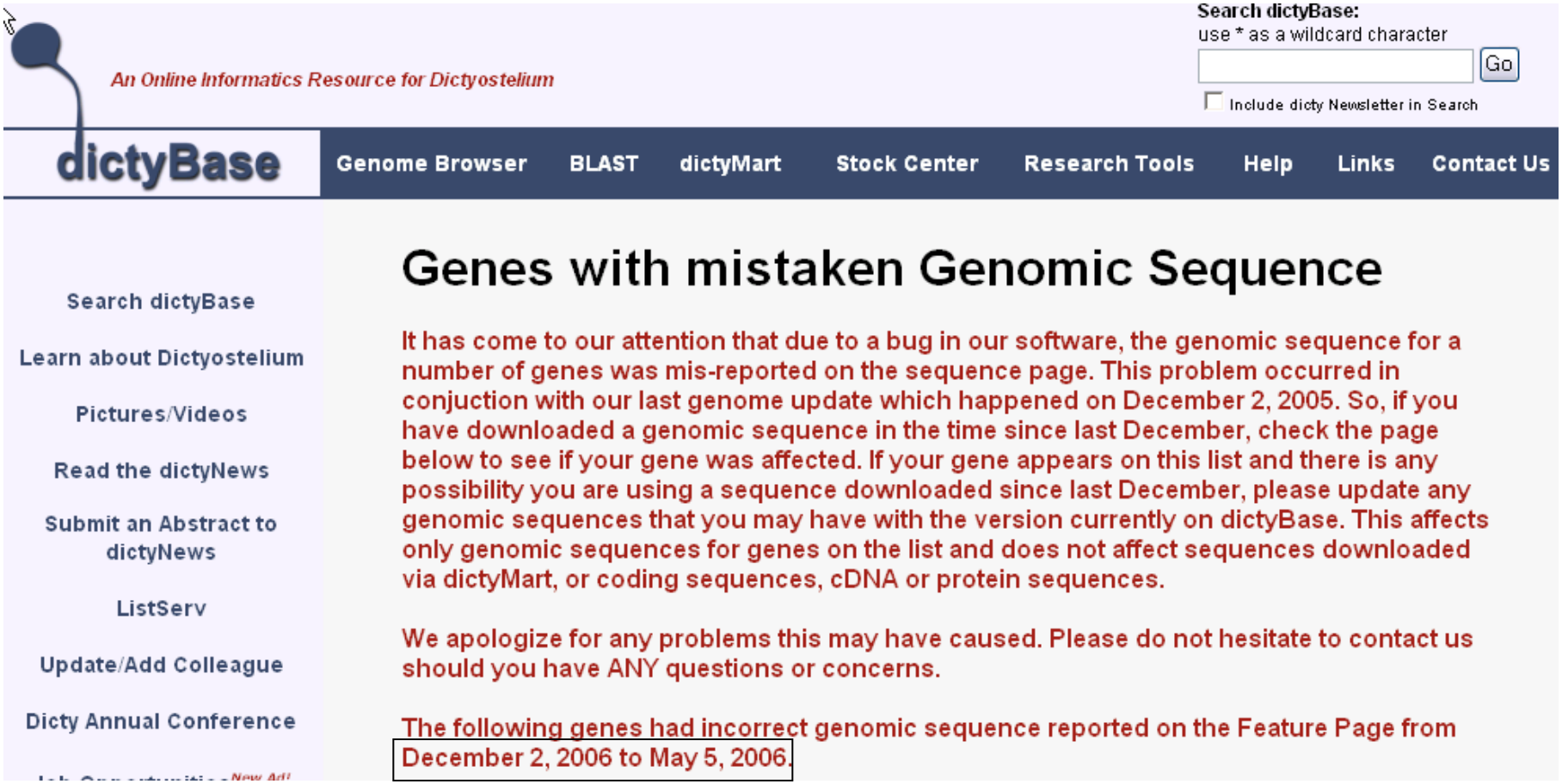

nature
The chicken genome


Science


PLoS BIOLOGY
Emergence of human genes
The spread of Ebola Zaire
Hair follicle development
Mechanisms of alternative splicing


nature
Plasmodium genomics
Genomics and proteomics pave the way for controlling malaria

Reading DNA Sequences



An Online Informatics Resource for Dictyostelium

Search dictyBase:
use * as a wildcard character
Go
Include dicty Newsletter in Search

dictyBase | Genome Browser | BLAST | dictyMart | Stock Center | Research Tools | Help | Links | Contact Us

Search dictyBase
Learn about Dictyostelium
Pictures/Videos
Read the dictyNews
Submit an Abstract to dictyNews
ListServ
Update/Add Colleague
Dicty Annual Conference

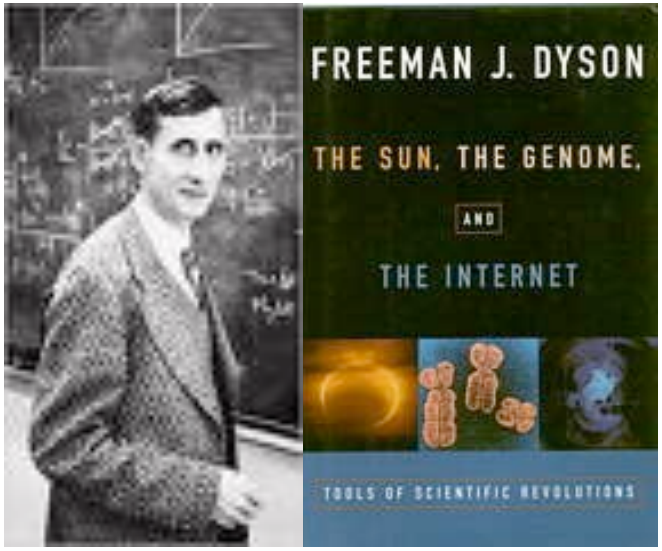
Genes with mistaken Genomic Sequence

It has come to our attention that due to a bug in our software, the genomic sequence for a number of genes was mis-reported on the sequence page. This problem occurred in conjuction with our last genome update which happened on December 2, 2005. So, if you have downloaded a genomic sequence in the time since last December, check the page below to see if your gene was affected. If your gene appears on this list and there is any possibility you are using a sequence downloaded since last December, please update any genomic sequences that you may have with the version currently on dictyBase. This affects only genomic sequences for genes on the list and does not affect sequences downloaded via dictyMart, or coding sequences, cDNA or protein sequences.

We apologize for any problems this may have caused. Please do not hesitate to contact us should you have ANY questions or concerns.

The following genes had incorrect genomic sequence reported on the Feature Page from December 2, 2006 to May 5, 2006.

Laptop Genome Sequencer



Freeman Dyson, "Pierre Teilhard de Chardin and Evolution," Marist College in Poughkeepsie, N.Y., on May 14, 2005.

*

"I am proposing now to hijack Moore's prediction and apply it to biology. ... The sequencing machines that now exist are marvels of ingenuity, but they are cumbersome and expensive.

. . .

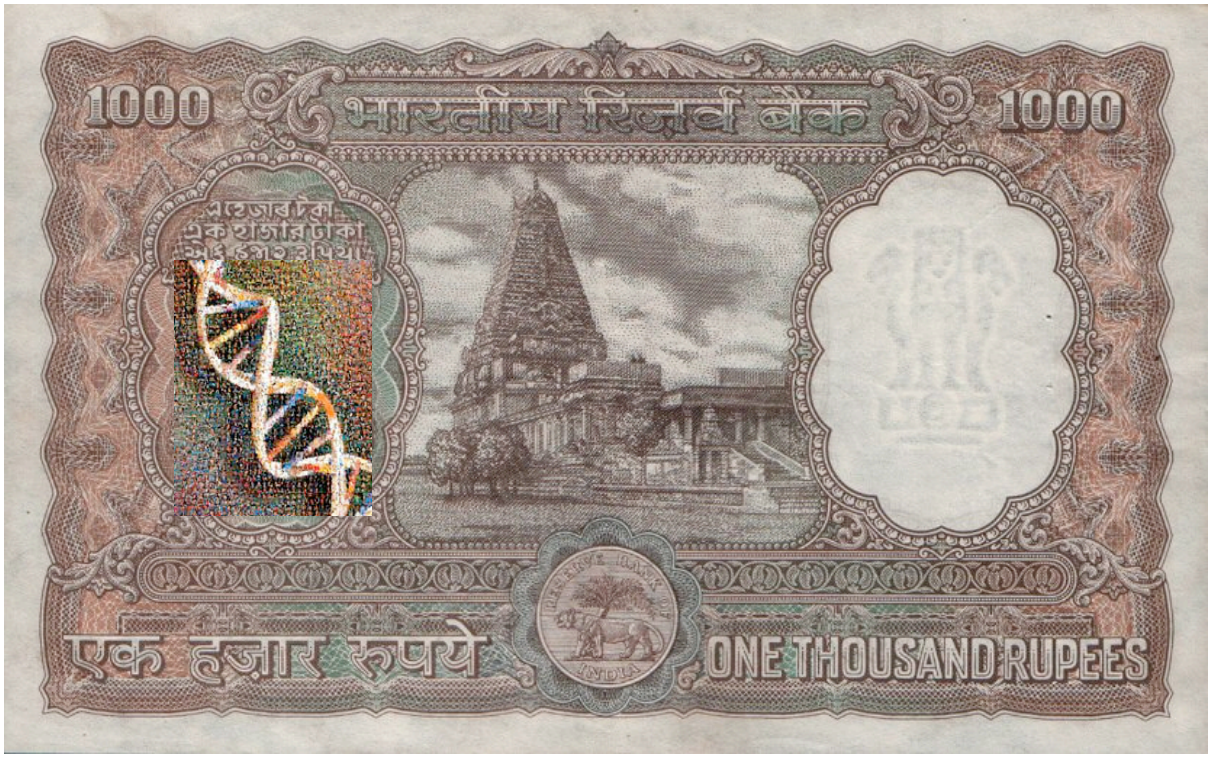
"What biology now needs is a single-molecule sequencer that can handle one molecule at a time and sequence it by physical rather than chemical methods.

. . .

"A single-molecule machine could be much cheaper as well as faster than existing machines. It might be as small and convenient as a lap-top computer…"

* * * * * * *

1000 Rupees Genome



22.67 US$ for 6 billion bases

135 billion US $ for the entire human population

Overview:
Moore's Law in Biotech

- **Miniaturization**
  - Single Molecule, Single Cell, Nano-scale, Femto-second
  - Minute amount of material: Avoid amplification
  - Non-Invasive, Asynchronous, Non-Realtime
- **Abstraction**
  - Multi-disciplinary, yet allow inter-disciplinary abstraction
- **Modularity**
  - Optimal integration of several technologies based on manipulation of single molecules on a surface.
  - Order of Emphasis: Computational, Physical, Chemical
- **Error Resilience**
  - How to build "reliable technologies" out of unreliable parts
  - 0-1 Laws and experiment design

S ¤ M ¤ A ¤ S ¤ H

Single

Molecule

Approach to

Sequencing-by-

Hybridization

Bud Mishra

Professor of Computer Science,
Mathematics and Cell Biology
⌘
Courant Institute, NYU School of Medicine, Tata Institute of Fundamental Research, and Mt. Sinai School of Medicine

Tools of the trade

Tools of the Trade

Scissors

- Type II Restriction Enzyme
  - Biochemicals capable of cutting the double-stranded DNA by breaking two -O-P-O bridges on each backbone
- Restriction Site:
  - Corresponds to specific short sequences: EcoRI GAATTC
  - Naturally occurring protein in bacteria… Defends the bacterium from invading viral DNA…Bacterium produces another enzyme that methylates the restriction sites of its own DNA

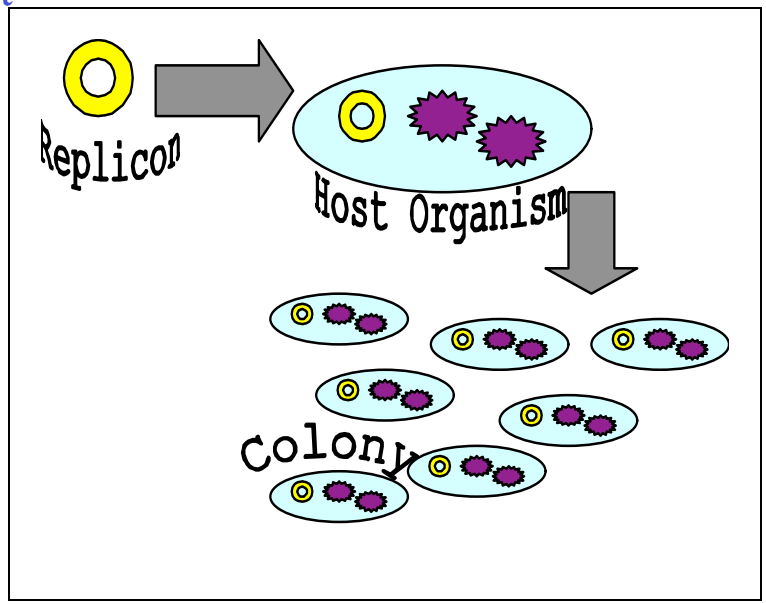
Hae III	5'…GG | CC …3' 3'…CC | GG …5'
*EcoR*I	5'…G|AATT C …3' 3'…C TTAA|G…5'
*Pst*I	5'…C TGCA|G …3' 3'…G|ACGT C …5'
*Hpa*I	5'…GTT|AAC …3' 3'…CAA| TTG …5'

Glue

- **DNA Ligase**
  - Cellular Enzyme: Joins two strands of DNA molecules by repairing phosphodiester bonds
  - T4 DNA Ligase (E. coli infected with bacteriophage T4)
- **Hybridization**
  - Hydrogen bonding between two complementary single stranded DNA fragments, or an RNA fragment and a complementary single stranded DNA fragment… results in a double stranded DNA or a DNA-RNA fragment

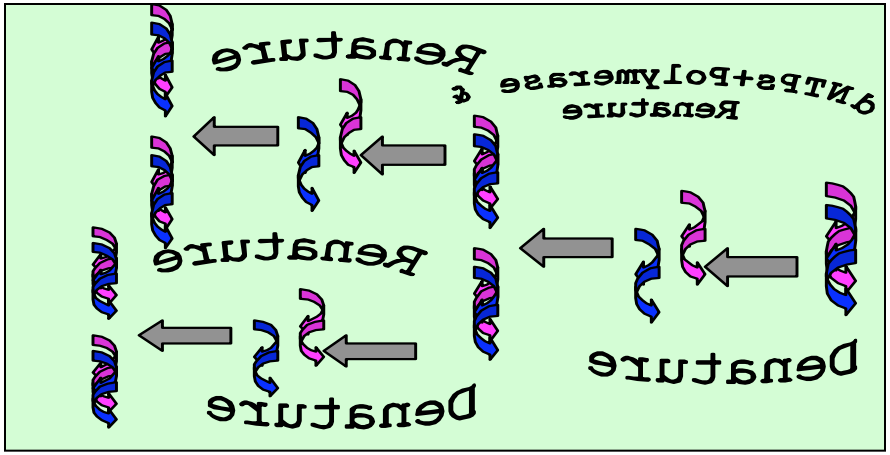
Copier

- **DNA Amplification:**
  - Main Ingredients: Insert (the DNA segment to be amplified), Vector (a cloning vector that combines with an insert to create a replicon), Host Organism (usually bacteria).

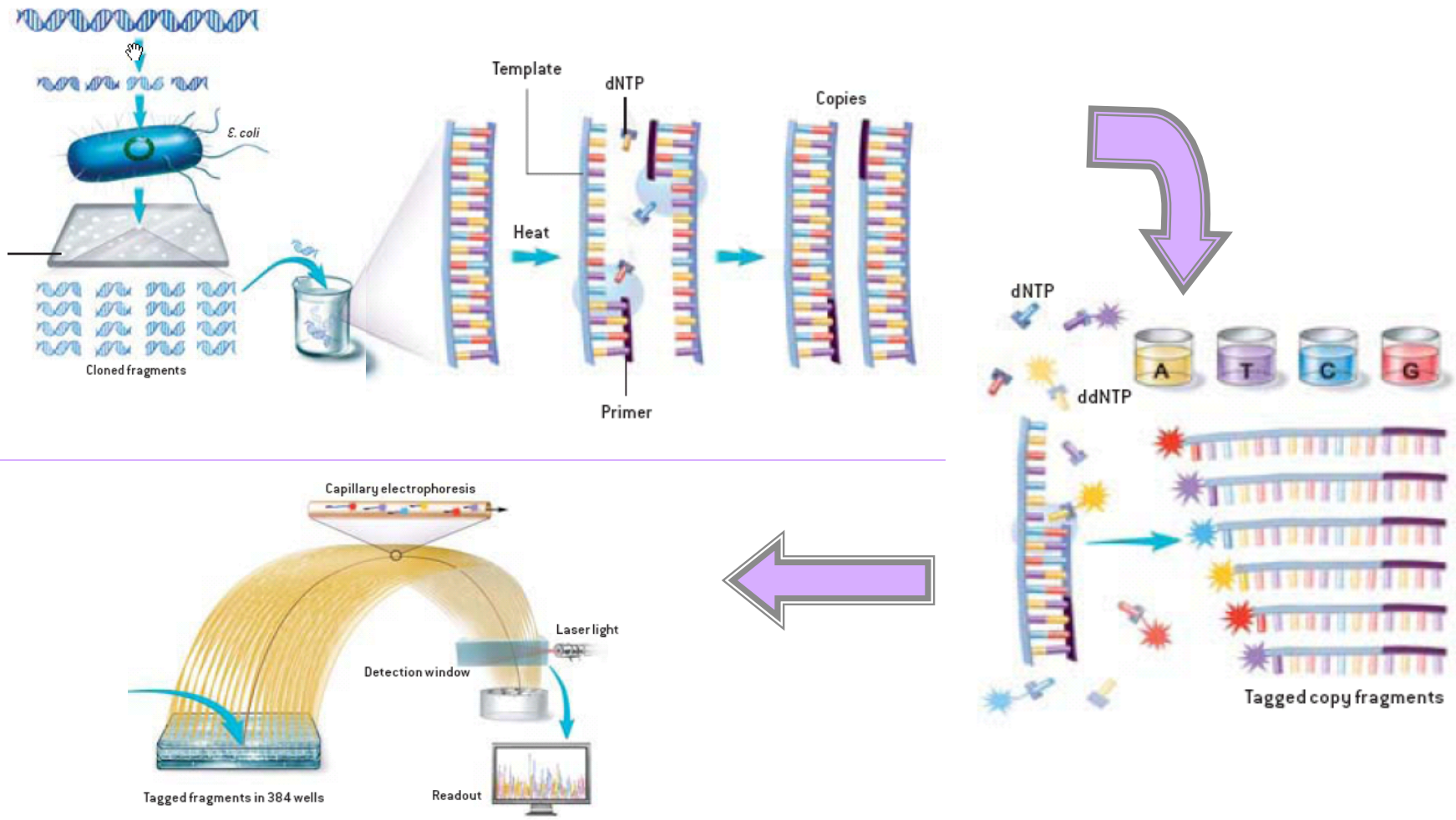


Copier

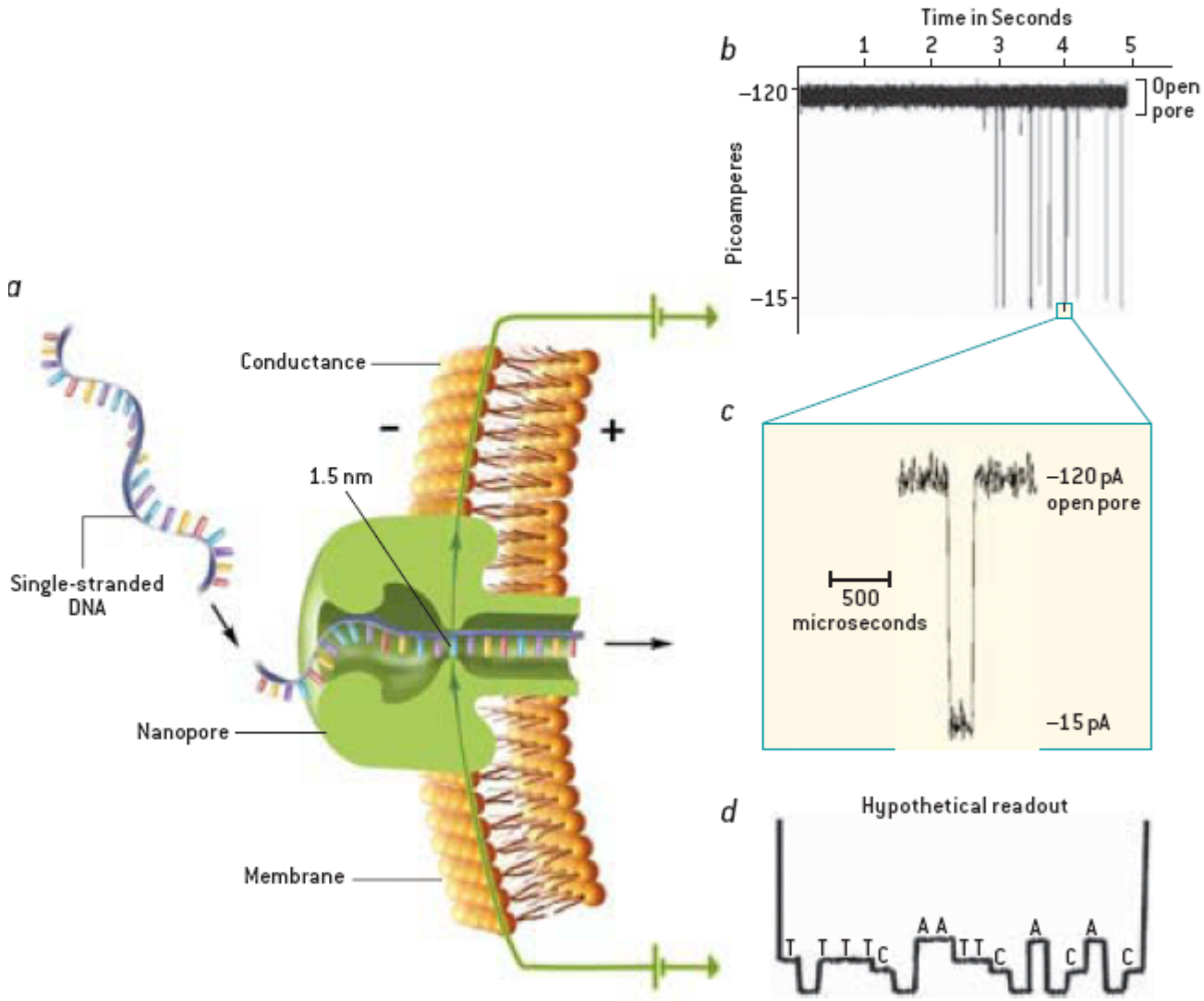
- PCR (Polymerase Chain Reaction):
- Main Ingredients: Primers, Catalysts, Templates, and the dNTPs.



Sanger Chemistry



Nanopore Sequencing



The Middle Way

- **Character: Index**
- A: 1, 11, …
- T: 2, 3, 12…
- C: 4, 5, 9, 10, 13 …
- G: 6, 7, 8, ….

ATTCCGGGCCATC

- **Sentences: w/o Index**
- ATTCCGGG…
- GGGCCATCGT…
- CGTCATTCC…

ATTCCGGGCCATC

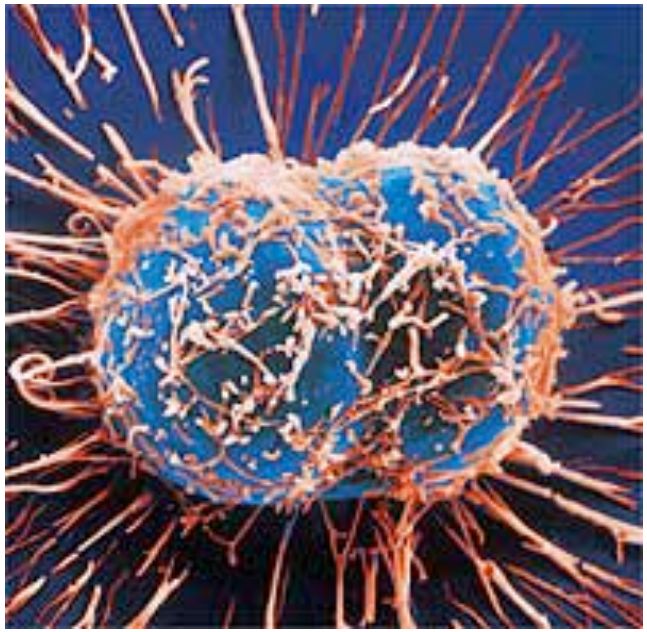
- **Words: w/ approx. Index**
- ATTC: [2..4]
- TCGG: [6..8]
- GGGC: [7..9]
- GCCA: [10..12]

ATTCCGGGCCA

S*M*A*S*H

- Sequence a human size genome of about 6 Gb—include both haplotypes.
- Integrate:
  - **Optical Mapping** (Ordered Restriction Maps)
  - **Hybridization** (with short nucleobase probes [PNA or LNA oligomers] with dsDNA on a surface, and
  - **Positional Sequencing by Hybridization** (efficient polynomial time algorithms to solve "localized versions" of the PSBH problems)

Fig 1

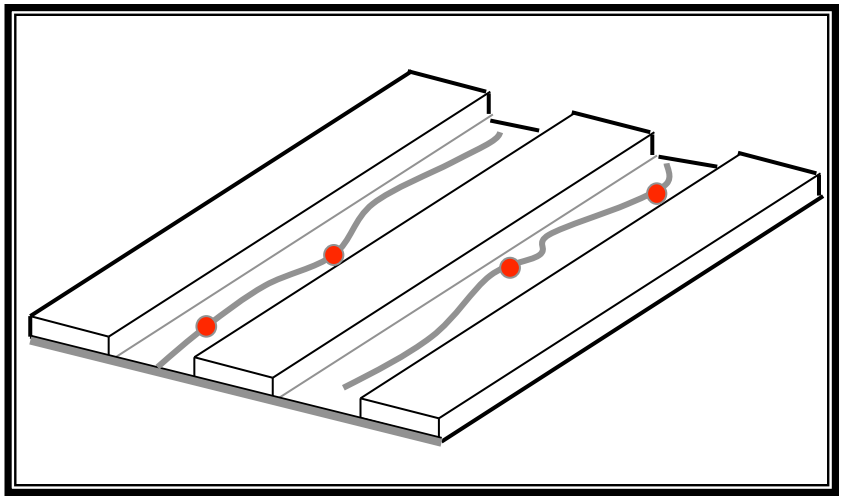


- Genomic DNA is carefully extracted

Fig 2

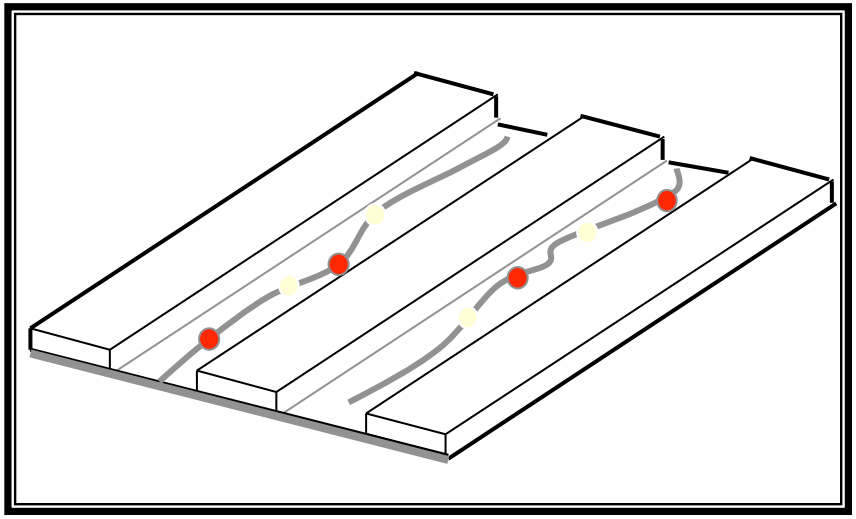
- LNA probes of length 6 – 8 nucleotides are hybridized to dsDNA (double-stranded genomic DNA)
- The modified DNA is stretched on a 1” x 1” chip.

Fig 3



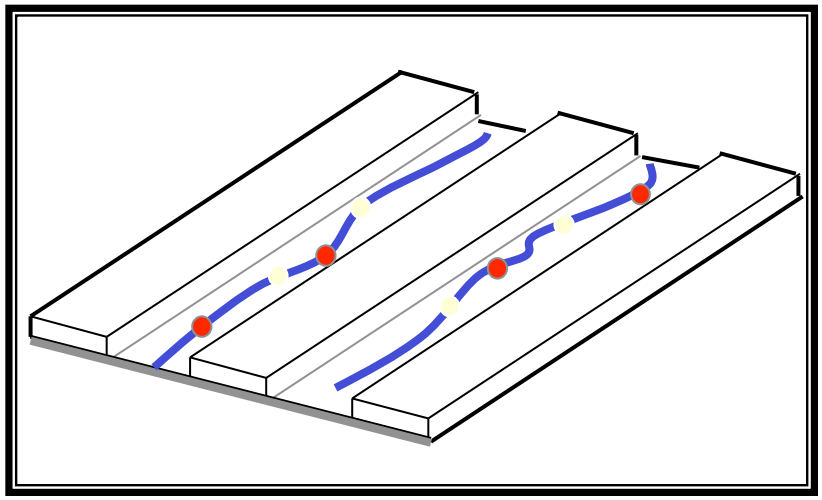
- DNA adheres to the surface along the channels and stretches out.
- Size from 0.3 – 3 million base pairs in length.
- Bright emitters are attached to the probes and imaged (Fig 3).

Fig 4



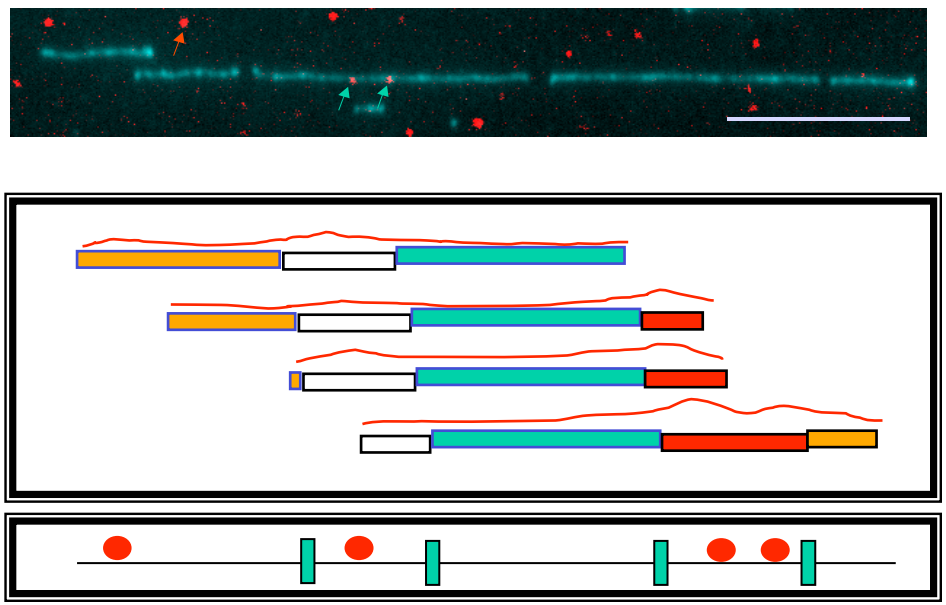
- A restriction breaks the DNA at specific sites.
- The cut fragments of DNA relax like entropic springs, leaving small visible gaps

Fig 5



- The DNA is then stained with a fluorogen (Fig 5) and reimaged.
- The two images are combined in a composite image
  - **suggesting the locations of a specific short word (e.g., probes) within the context of a pattern of restriction sites.**

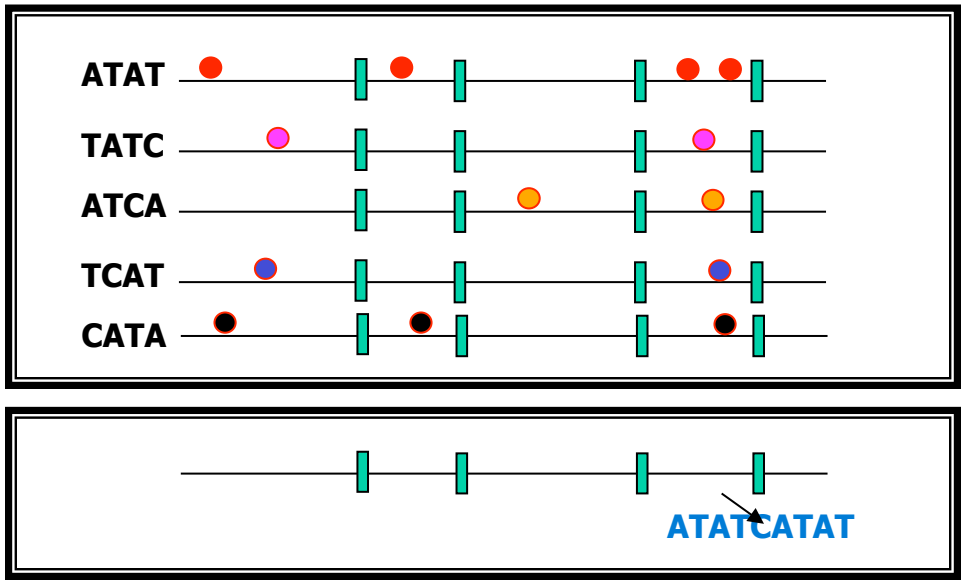
Fig 6



The *restriction sites* are represented by a tall rectangle &
The *probe sites* by small circles

- The integrated intensity measures the length of the DNA fragments.
- The bright-emitters on probes provides a profile for locations of the probes.

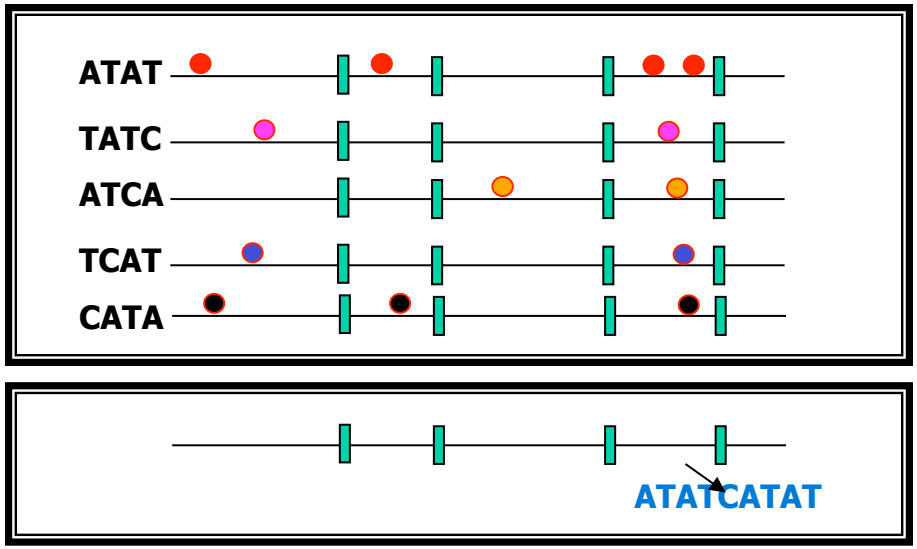
Fig 7



- These steps are repeated for all possible probe compositions
  - (modulo reverse complementarity).
- Software assembles the haplotypic ordered restriction maps with approximate probe locations superimposed on the map.

S*M*A*S*H

Fig 7



- Local clusters of overlapping words are combined by our PSBH (positional sequencing by hybridization) algorithm

Science by Stamp Collecting

Science by Coupon Collecting

Sir Ernest Rutherford

"All science is either physics or stamp collecting."

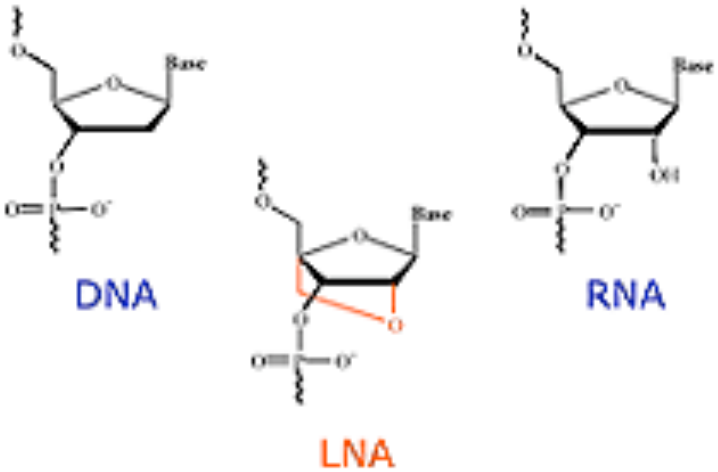
"For Mike's sake, Soddy, don't call it transmutation. They'll have our heads off as alchemists." Rutherford, winner of 1908 Nobel prize for chemistry for cataloging alpha and beta particles…

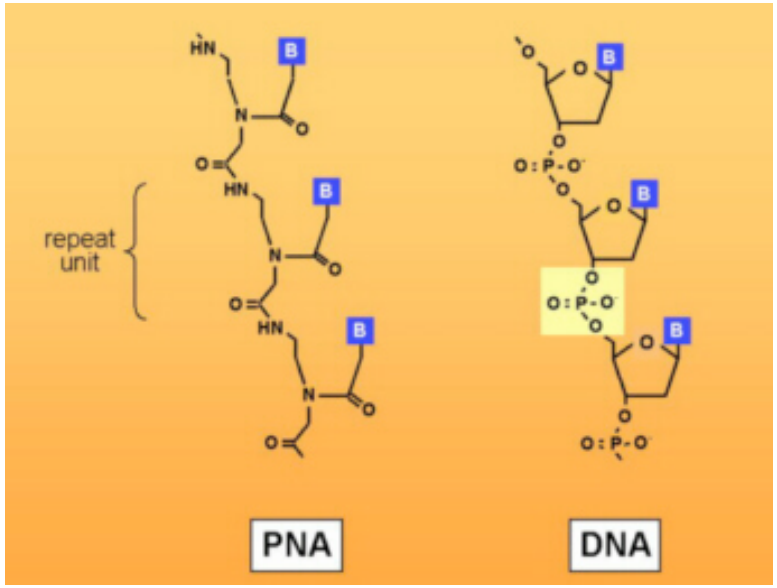
Hybridization

Probes

- LNA
  - Negative backbone with modified sugar moiety
- PNA
  - Neutral backbone made up of pseudo-peptide backbone

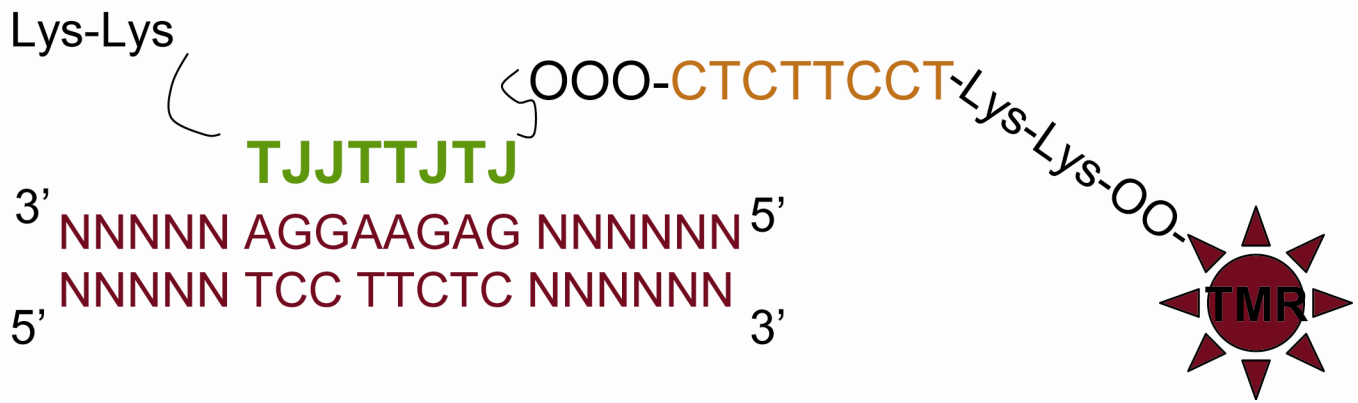
Stable complex formation at elevated temp.



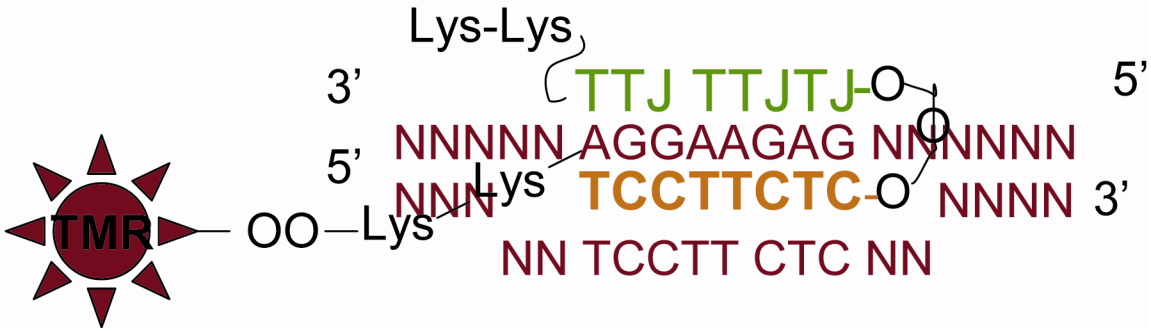


bisPNA Probe

- TMR-OO-Lys-Lys-**TCC-TTC-TC**-OOO-**JTJ-TTJ-JT**-Lys-Lys
- **Hoogsteen Binding**

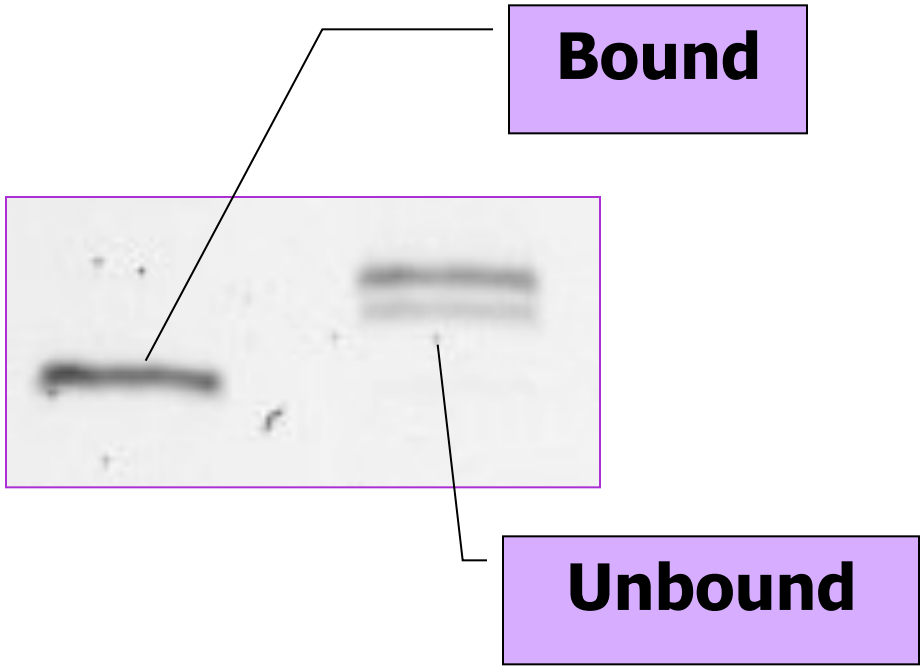


- **Watson Crick binding**: Strand Invasion



(T) Thymine; (C) Cytosine;
(J) pseudoisocytosine
(O) linkers (8-amino-3,6-dioxaoctanoic Acid. Form flexible linker

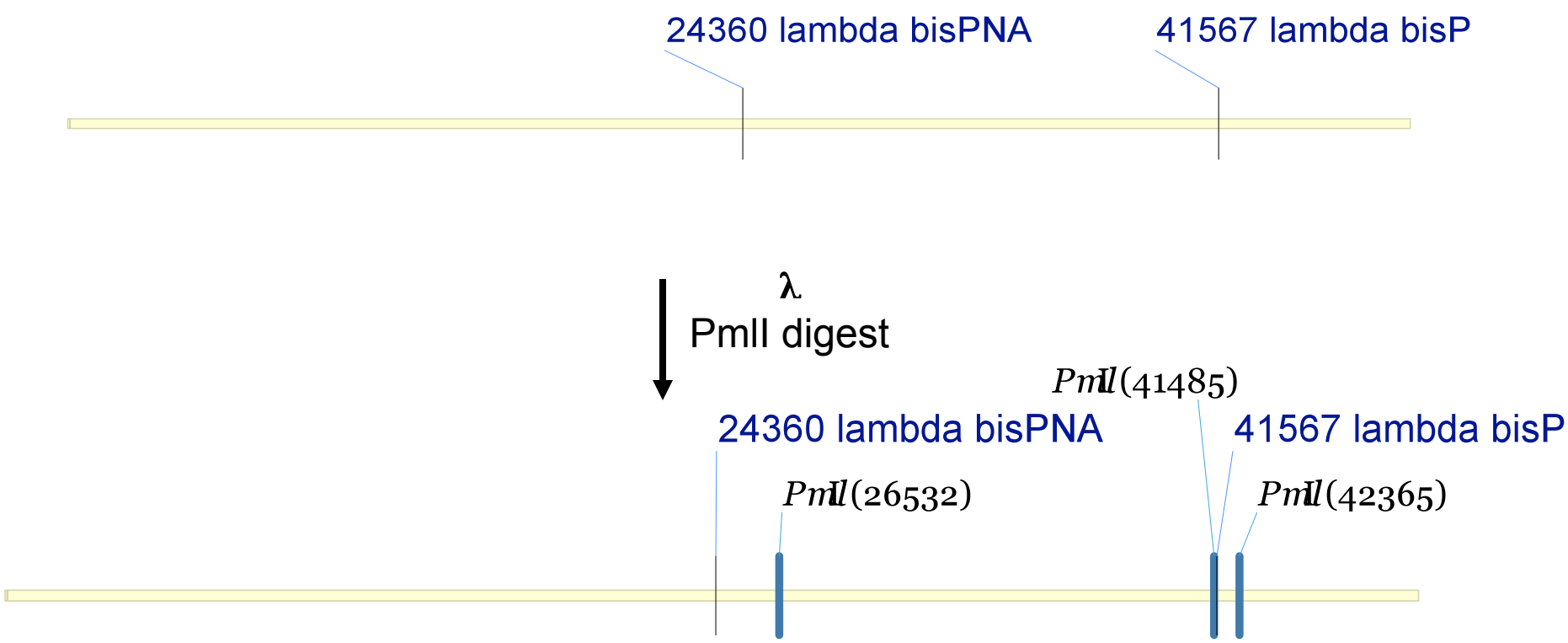
Experiments with PNA Probes



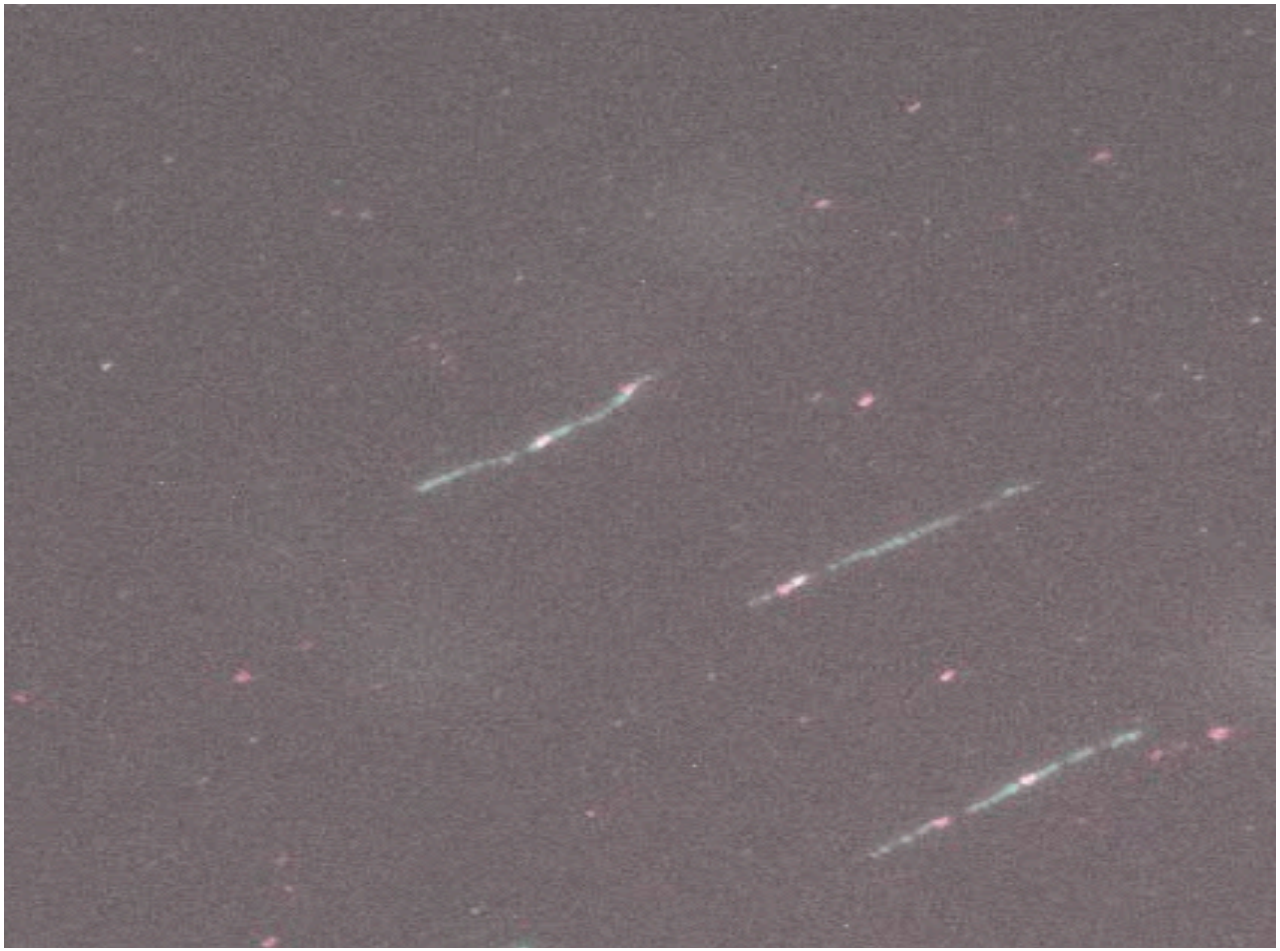
- **Calibration** using hybridization to lambda DNA molecules.
- Degree of hybridization > 90%.

bisPNA probe

bisPNA Sequence:
TMR-OO-Lys-Lys-**TCC-TTC-TC**-OOO-JTJ-TTJ-JT-Lys-Lys



Probe Map (lambda DNA)

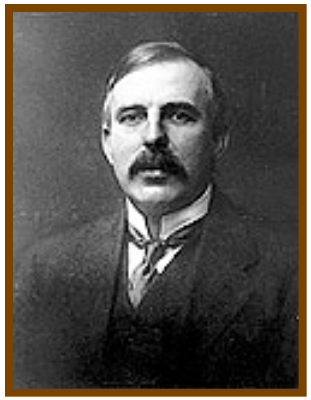


Final Probe Map

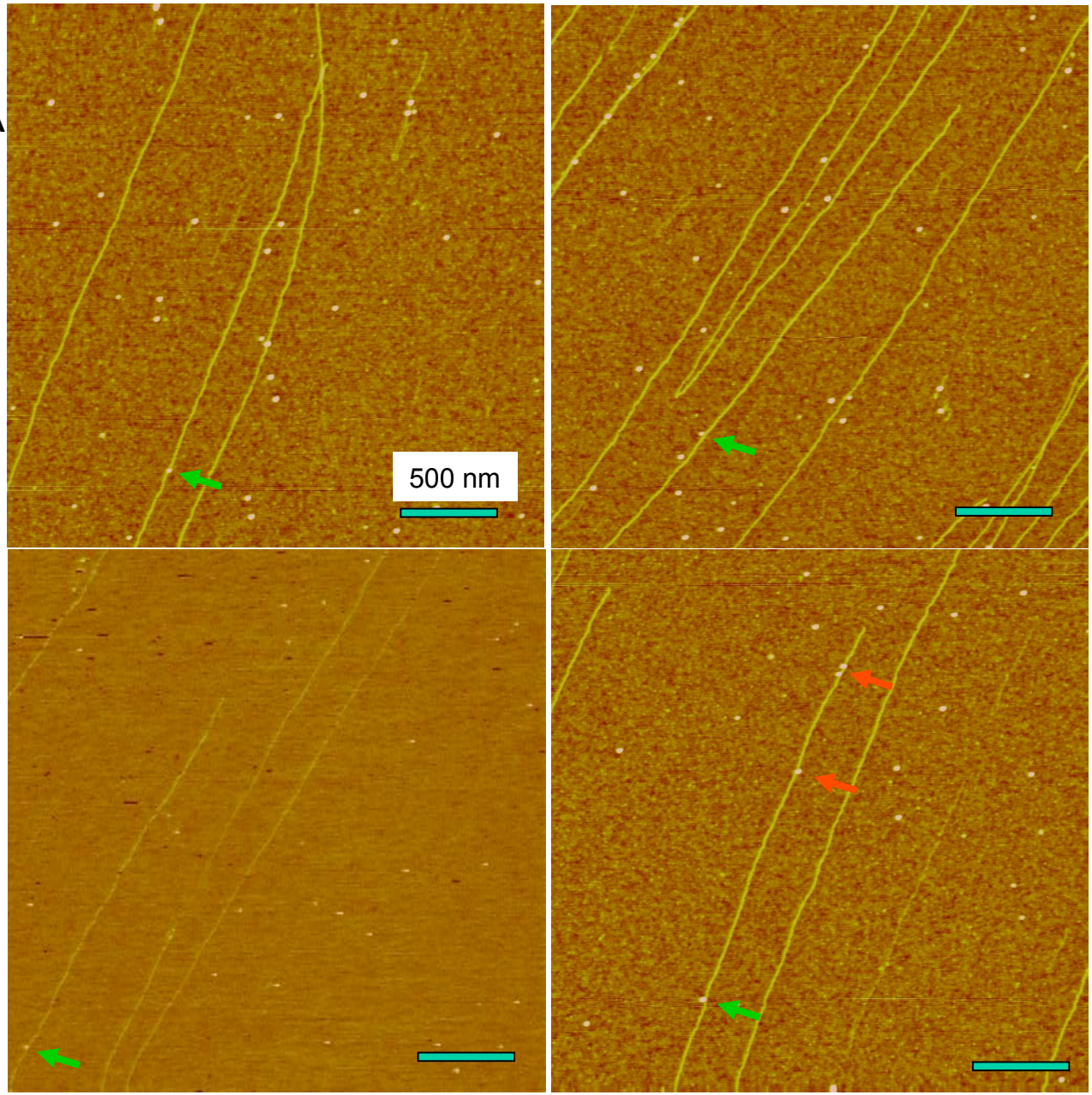
- Consensus map with 2 probe locations
  - 14.8% and 52.4% of the DNA length.
- In close agreement with the correct map
  - 50.2% and 85.7% (known from the sequence)
- Implied probe hybridization rate = 42%.
  - Significantly better than the needed 30%

Sir Ernest Rutherford

"You should never bet against anything in science at odds of more than about 10^{12} to 1."

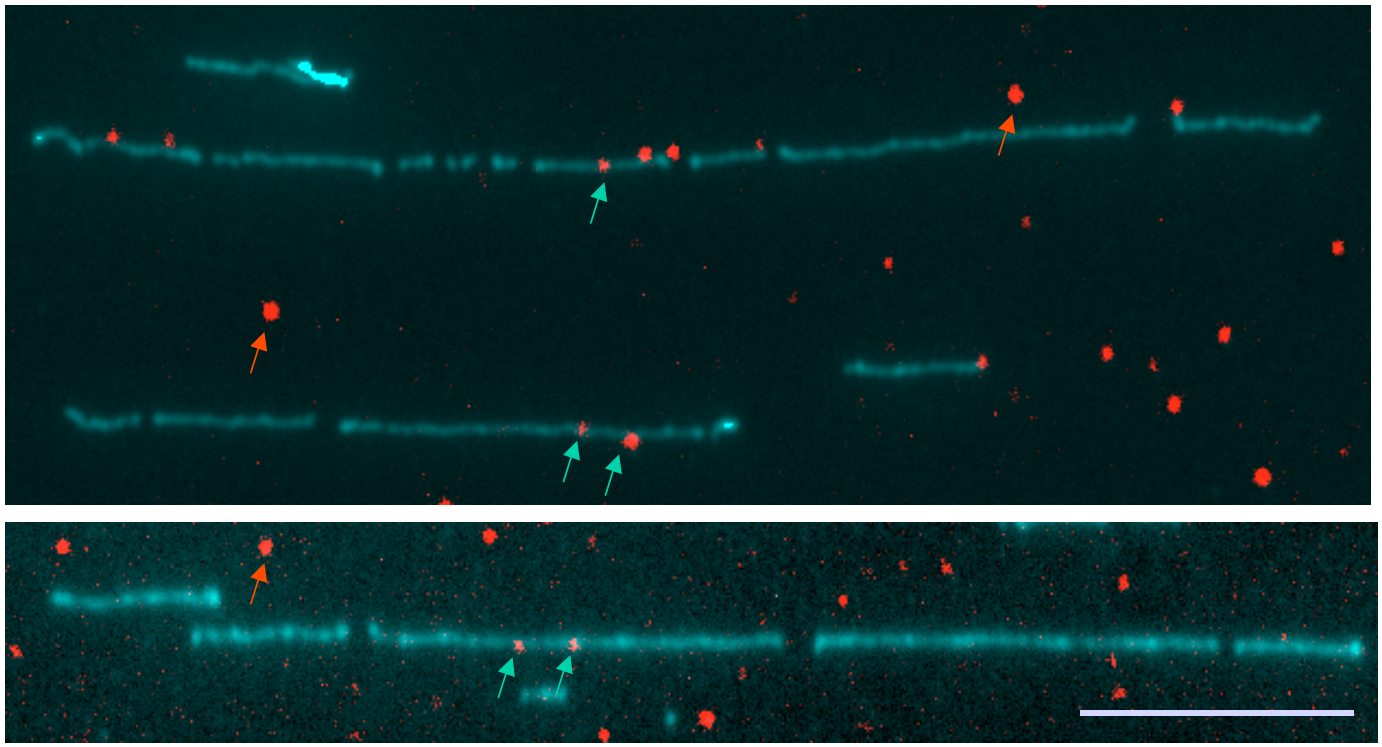


A



Four AFM images of lambda DNA with PNA probes

E. coli

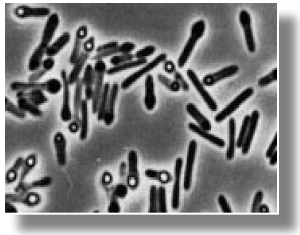


Two optical images of E coli K12 genomic DNA after restriction digestion with 6-cutter restriction enzyme Xho 1 and hybridization with an 8-mer PNA probe. Scale bar shown is 10 micron.

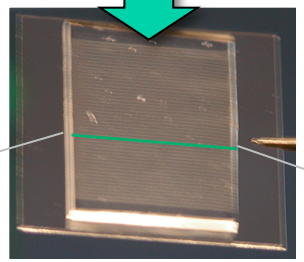
Optical Mapping

Optical Mapping

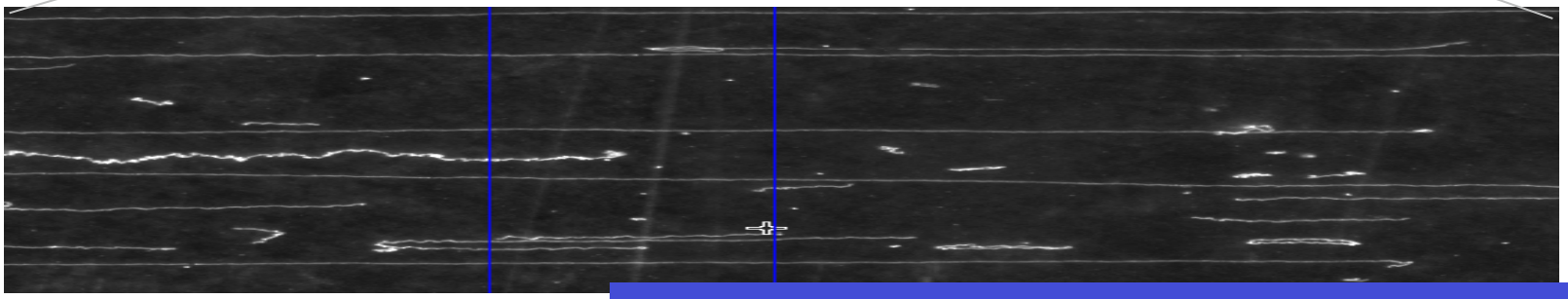
1. **Capture and immobilize whole genomes as massive collections of single DNA molecules**



Cells gently lysed to extract genomic DNA

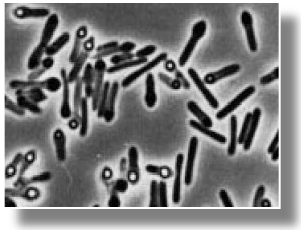


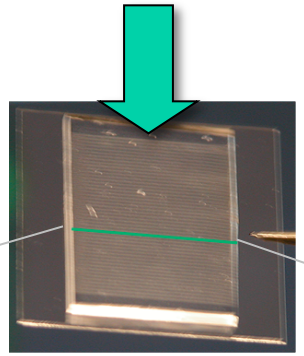
DNA captured in parallel arrays of long single DNA molecules using microfluidic device

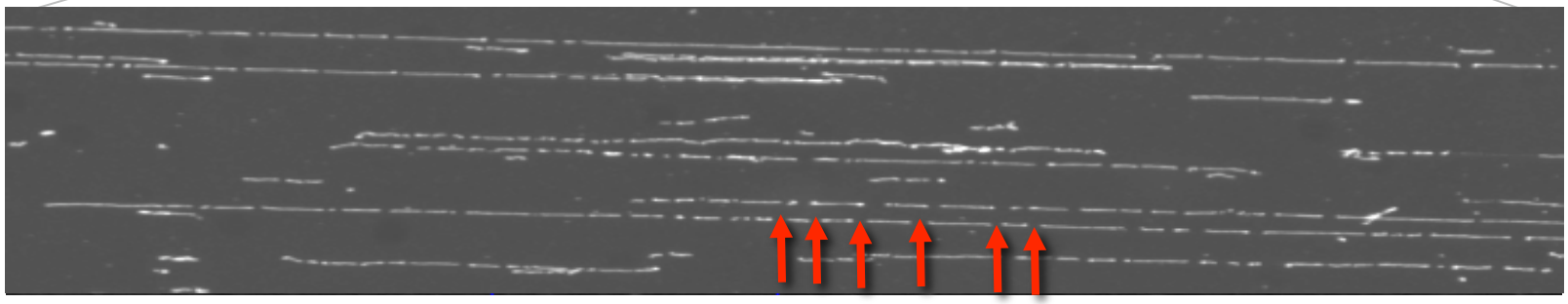


Genomic DNA, captured as single DNA molecules produced by random breakage of intact chromosomes

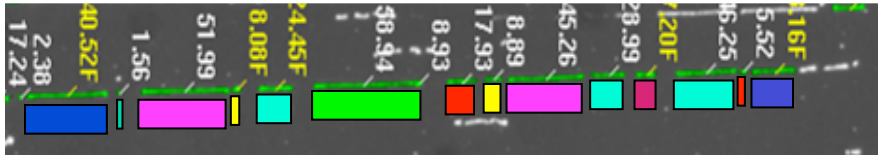
2. Interrogate with restriction endonucleases
3. Maintain order of restriction fragments in each molecule

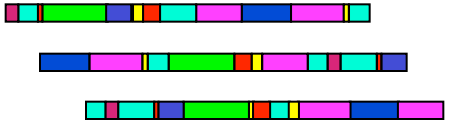




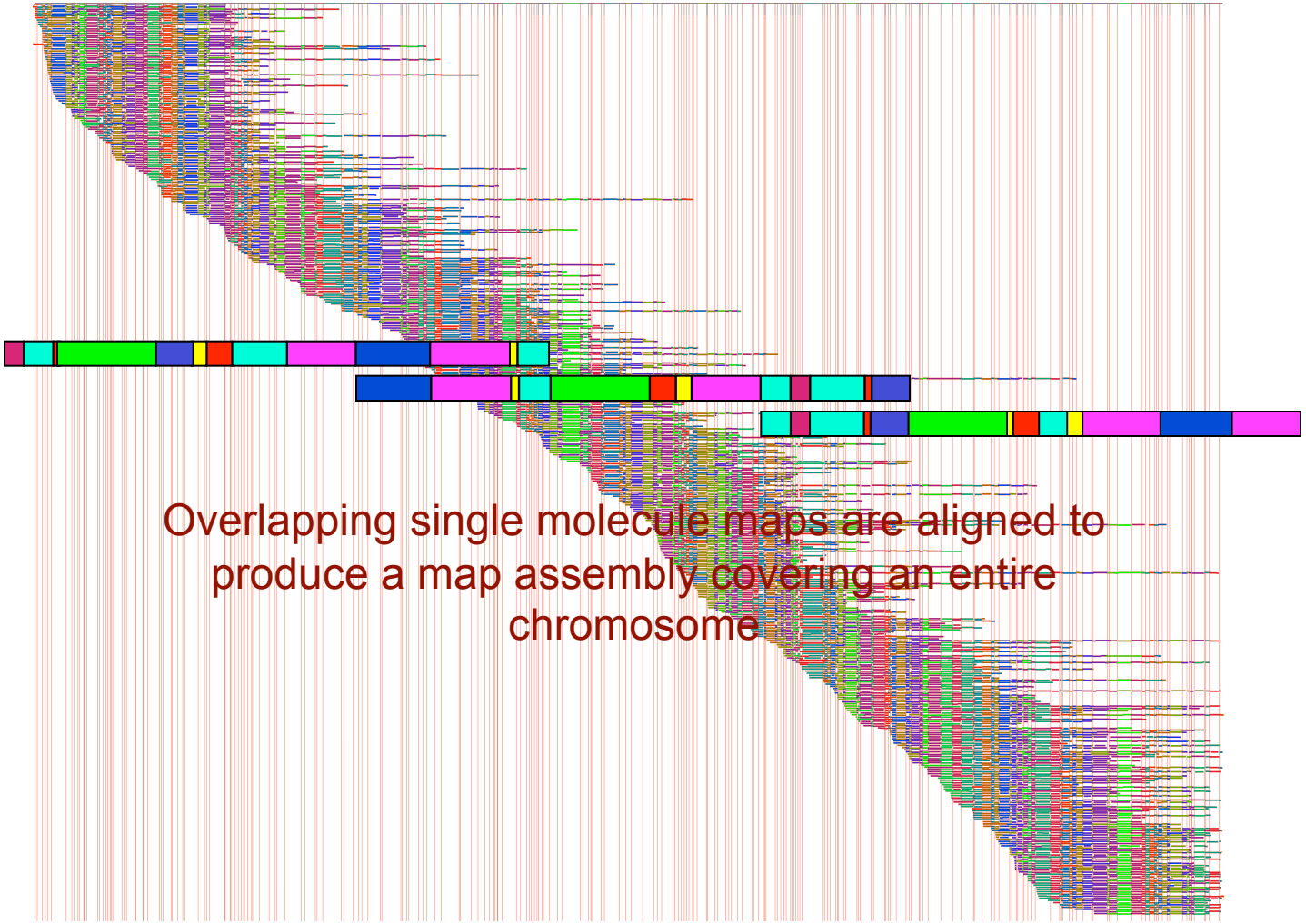


Digestion reveals 6-nucleotide cleavage sites as "gaps"

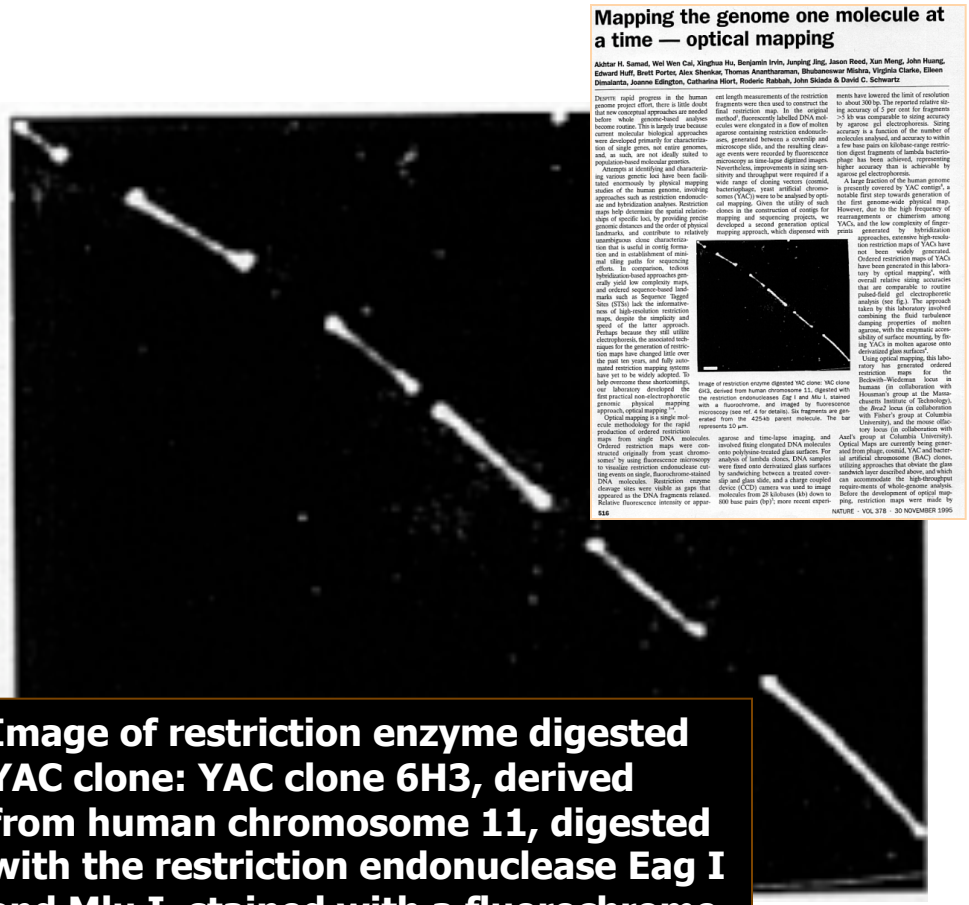




- Overlapping single molecule maps are aligned to produce a map assembly covering an entire chromosome


Overlapping single molecule maps are aligned to produce a map assembly covering an entire chromosome

Error Sources



Mapping the genome one molecule at a time — optical mapping

Akhtar H. Samad, Wei Wen Cai, Xinghua Hu, Benjamin Irvin, Junping Jing, Jason Reed, Xun Meng, John Huang, Edward Huff, Brett Porter, Alex Shenkar, Thomas Anantharaman, Bhubaneswar Mishra, Virginia Clarke, Eileen Dimalanta, Joanne Edington, Catharina Hiort, Roderic Rabbah, John Skiada & David C. Schwartz

Image of restriction enzyme digested YAC clone: YAC clone 6H3, derived from human chromosome 11, digested with the restriction endonucleases *Eag* I and *Mlu* I, stained with a fluorochrome, and imaged by fluorescence microscopy (see ref. 4 for details). Six fragments are generated from the 425-kb parent molecule. The bar represents 10 μm.

Image of restriction enzyme digested YAC clone: YAC clone 6H3, derived from human chromosome 11, digested with the restriction endonuclease Eag I and Mlu I, stained with a fluorochrome and imaged by fluorescence microscopy.

- Sizing Error
  - (Bernoulli labeling, absorption cross-section, PSF)
- Partial Digestion
- False Optical Sites
- Orientation
- Spurious molecules, Optical chimerism, Calibration

Computational Complexity & Feasibility

Complexity Issues

Various combinations of error sources lead to NP-hard Problems

Problem 1	**Partial Digestion** **Optical Cuts** **Unknown Orientation**	NP-hard Inapproximable*
Problem 2	**Partial Digestion** **Optical Cuts** **Sizing Errors**	NP-hard
Problem 3	**Partial Digestion** **Optical Cuts** **Missing Fragments**	NP-hard Inapproximable*
Problem 4	**Partial Digestion** **Optical Cuts** **Spurious Molecules**	NP-hard Inapproximable*

* No Polynomial Time Approximation Scheme (PTAS), if $P \neq NP$.

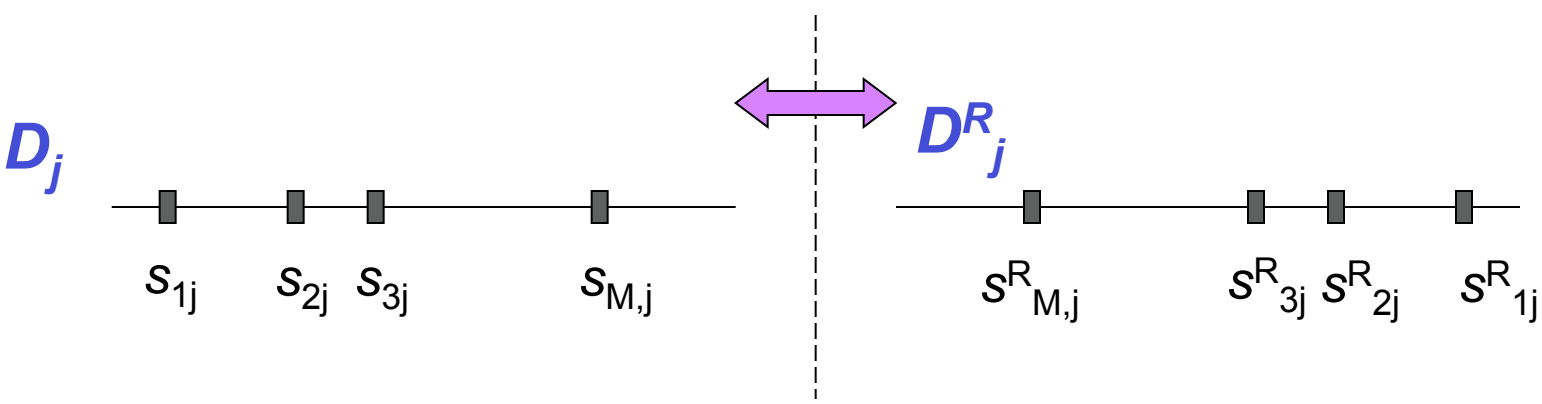
SMRM
(Single Molecule Restriction Map)

- **Data:** SMRM vectors

$$D_j = (s_{1j}, s_{2j}, \ldots, s_{M_j,j})$$
$$0 < s_{1j} < s_{2j} < \cdots < s_{M_j,j} < 1, \quad s_{ij} \in Q.$$

- **Reflection:** $s^R = 1 - s$.

$$D_j^R = (s^R_{M_j,j}, \ldots, s_{2j}, s_{1j})$$



Given: A collection of data (SMRM vectors)

$D_1, D_2, \ldots, D_m$

Compute: A final vector H

$H = (h_1, h_2, \ldots, h_n)$

"consistent" with each D_j.

- $\text{dist}(H, D_j)$ $\Leftarrow$ *Reflects "consistency requirement"*
- H <u>minimizes</u>

$$\max_j \min\left(\text{dist}(H, D_j), \text{dist}(H, D_j^R)\right) + \text{CONSTRAINTS}$$

Given: A collection of SMRM vectors

$$D_1, D_2, \ldots, D_l, D_{l+1}, \ldots, D_m$$

An *approximate* solution

$$\bar{H} = (\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_n).$$

An *approximation factor* ϵ.
A *variance upper bound* σ^2.

- An **admissible alignment** (A_k) of the data

$$D'_1, D'_2, \ldots, D'_l, D'_{l+1}, \ldots, D'_m$$

where $D'_j = \begin{cases} D_j \text{ or } D_j^R, & \text{if } 1 \le j \le l; \\ D_j, & \text{if } j > l. \end{cases}$

- **Matching Set:** Fixed A_k, given $\bar{h}_i$

$$S_{ijk} = \{s \in D'_j : |s - \bar{h}_i| \le \epsilon\}, \quad \text{and}$$

$$S_{ik} = \biguplus_j S_{ijk}.$$

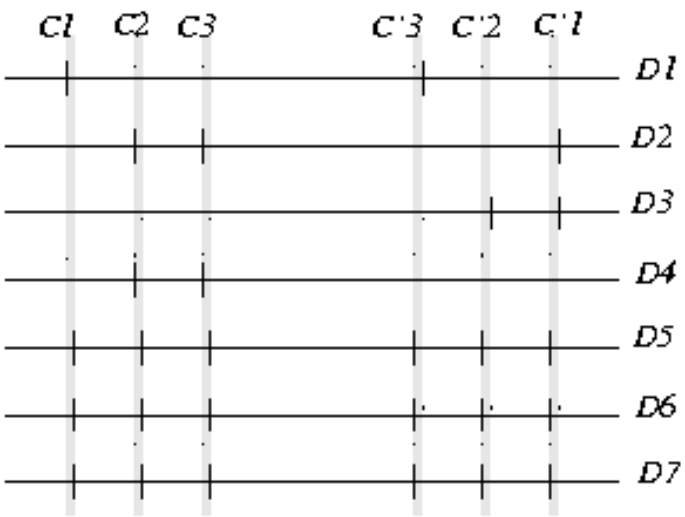
- Define $h_i = \text{mean}\,(S_{ik})$ and $\sigma_i^2 = \text{var}\,(S_{ik})$.

Determine: $\exists$? admissible alignment A_k s.t.

$$H = (h_1, h_2, \ldots, h_n)$$

with $\forall\, i\ \sigma_i^2 \le \sigma^2$.

$$C_1 \cdot C_2 \cdot C_3$$
$$\equiv \quad (x_1 + \overline{x}_2 + \overline{x}_3)\ (x_2 + \overline{x}_3 + x_4)\ (\overline{x}_1 + x_2 + x_4).$$



- Suppose that the CNF formula has a satisfying assignment such that each clause has at least one true literal and at least one false literal.
- **Then choose** $(1 \leq j \leq l)$

$$D'_j = \begin{cases} D_j & \text{if } x_j = \text{True;} \\ D_j^R, & \text{if } x_j = \text{False.} \end{cases}$$

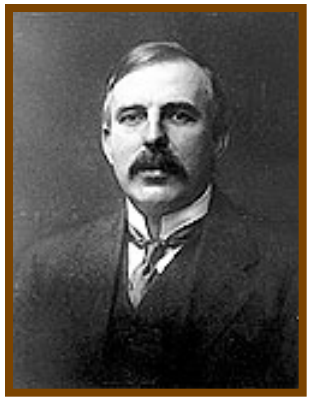
- Then for each i, variance of $S_i k$ is

$$\sigma_i^2 \leq \text{var}\ (0, 0, 0, \epsilon, \epsilon)$$
$$= \quad (2/5)(3/5)(1/5(n+1))^2 = 6/625(n+1)^2.$$

- Converse, similar.

Sir Ernest Rutherford

"If your experiment needs statistics, you ought to have done a better experiment."



Combinatorial Structure

- **Model**
  - Correct Map: $0 < h_1 < h_2 < \cdots < h_k < 1$.
  - Experimental Observations: $0 < s_1 < s_2 < \cdots < s_l < 1$.
  - Only error source $\mapsto$ Partial Digestion

$$\forall_{h_j}\ Pr[\exists_{s_i}\ h_j = s_i] = p_c.$$

- **Theorem**
  Let ϵ be a positive constant and $c \geq 1$ be so chosen that $1 - e^{-2e^{-c}} = \epsilon$. Then for

$$n \geq \frac{c}{p_c} + \frac{\ln k}{p_c} \quad (k \geq 1),$$

with probability at least $1 - \epsilon$, the correct ordered restriction map can be computed in $O(nk)$ time.

When

$$n < \frac{\ln k}{p_c(1 + p_c)} \quad (k \geq 1 \text{ and } 0 < p_c < 0.69),$$

no algorithm can compute the correct ordered restriction map with probability better than half.

- *Intuition*: Probability that all k true cut sites appear in the final map

$$[1 - (1 - p_c)^n]^k \approx e^{-ke^{-p_c n}} \approx e^{-e^{-c}} \Rightarrow p_c n \approx \ln k + c.$$

Flips & Flops

- **Model**
  – Correct Map: $0 < h_1 < h_2 < \cdots < h_k < 1$.
  No symmetric site $\forall_i \forall_{j \neq i} h_i \neq h_j^R$.

  – Experimental Observations: $0 < s_1 < s_2 < \cdots < s_l < 1$. – Only error source $\mapsto$ Partial Digestion

  $$\forall_{h_j}\ Pr[\exists_{s_i}\ h_j = s_i] = p_c.$$

- **Theorem**
  Let ϵ be a positive constant and $c \geq 1$ be so chosen that $1 - e^{-3e^{-c}} = \epsilon$. Then for

  $$n \geq \max\left[\frac{c}{p_c} + \frac{\ln k}{p_c}, \frac{1}{p_c^2} \ln\left(\frac{k}{k - \ln k - c}\right)\right],$$

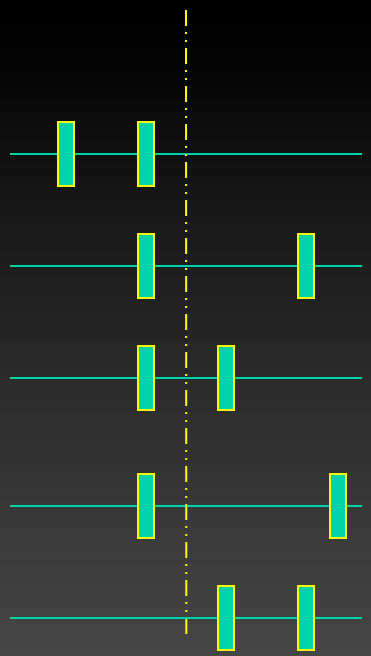
  ($k > c + \ln k$), with probability at least $1 - \epsilon$, the correct ordered restriction map can be computed.

  When

  $$n < \max\left[\frac{\ln k}{p_c(1 + p_c)}, \frac{1}{p_c^2(1 + p_c^2)} \ln \frac{k}{k - 1}\right],$$

  ($k > 1$ and $0 < p_c < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.

Intuition



- **Phase 1**

$$f \;:\; (0,1) \to (0,1/2)$$
$$:\; x \mapsto \begin{cases} x & \text{if } x \in (0,1/2); \\ x^R & \text{if } x \in (1/2,1). \end{cases}$$

- Compute $\{f(h_1), f(h_2), \ldots, f(h_k)\}$, from $\{f(s_{i1}), f(s_{i2}), \ldots, f(s_{il_i})\}$, $\quad i = 1, \ldots, n$.

- **Phase 2**

$$\hat{f} \;:\; (0,1/2) \times \{+1,-1\} \to (0,1)$$
$$:\; (f(h_j), sgn) \mapsto \begin{cases} f(h_j) & \text{if } sgn = +1; \\ f(h_j)^R & \text{if } sgn = -1. \end{cases}$$

Define a graph $G = (V, E)$,

$$\begin{aligned} V &= \{f(h_1), f(h_2), \ldots, f(h_k)\} \\ E &\subset V \times V, e = [f(h_i), f(h_j)] \in E \end{aligned}$$

if and only if

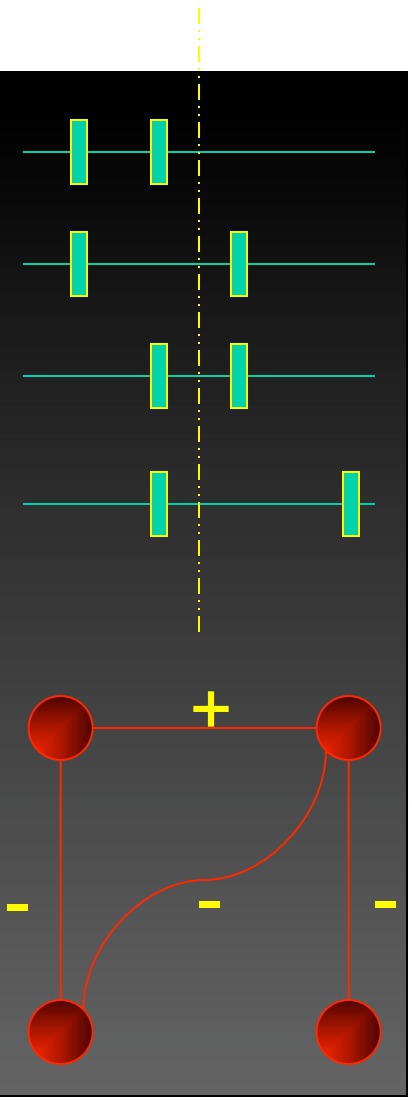
$$\exists_{s_{i'}, s_{j'}}\; f(s_{i'}) = f(h_i) \text{ and } f(s_{j'}) = f(h_j).$$

Sign Label

$$sgn(e) = sgn[(1/2 - s_{i'})(1/2 - s_{j'})].$$

- *Note*: If the graph is connected then its vertices can be labeled uniquely (up to multiplication by -1) and the correct map can be created.

$$p_e = 1 - (1 - p_c^2)^n \approx 1 - e^{-\ln(k/k - \ln k - c)} \approx \frac{\ln k}{k} + \frac{c}{k}.$$



Other Error Sources

- **Symmetric Sites**

- *Optical False Cuts*:
  –Poisson Process with parameter λ_f

$$\begin{aligned} Pr[\#\text{ false cuts} \in [x, x+\delta x] = 1] &= \lambda_f \delta x, \\ Pr[\#\text{ false cuts} \in [x, x+\delta x] \geq 2] &= o(\delta x). \end{aligned}$$

- *Oracle to distinguish two cuts*:

$$\begin{array}{lll} x & \approx_\delta & y \quad x \in [y-\delta, y+\delta], \\ x & <_\delta & y \quad x < y - \delta, \\ x & >_\delta & y \quad x > y + \delta. \end{array}$$

- Essentially the previous analysis works *mutatis mutandis*.

Discretization

- **Model Parameters**

k	=	# Cuts
m	=	# Symmetric Cuts
L	=	Length of the Clone in bps
Δ	=	Length of the Discretized Subintervals in bps
p_c	=	Partial Digestion Rate
λ_f	=	Spurious Cut, Poisson Parameter
p_E	=	Cutting Rate of the Enzyme

- **Theorem**

Assume that the sizing error $\sigma = 0$.

Let ϵ be a positive constant and $c \geq 1$ be so chosen that $1 - e^{-12e^{-c/2}} = \epsilon$. Then for

$$n \geq \frac{18}{p_c} \max\Big[c + 2\ln(k+m), \frac{c + \ln m}{p_c}, \frac{1}{p_c} \ln\left(\frac{k}{k - \ln k - c}\right) (c + \ln(L/2\Delta - k - m)), \frac{c + 2\ln k}{p_c}\Big],$$

($k > c + \ln k$, $m \geq 1$, $L > 2\Delta$ and $\lambda_f < p_c L/5\Delta$), with probability at least $1 - \epsilon$, the correct ordered restriction map can be computed in $O(nk^2)$ time.

When

$$n < \max\left[\frac{\ln(k+m)}{p_c(1+p_c)}, \frac{1}{p_c^2(1+p_c^2)} \max\left[\ln m, \ln\frac{k}{k-1}\right], \frac{\ln(L/\Delta)}{\ln(L/\lambda_f\Delta)}\right]$$

($k > 1$, $m > 1$, $L > \Delta$ and $0 < p_c < 0.69$), no algorithm can compute the correct ordered restriction map with probability better than half.

Sizing Error

- What happens when you introduce

  SIZING ERROR?

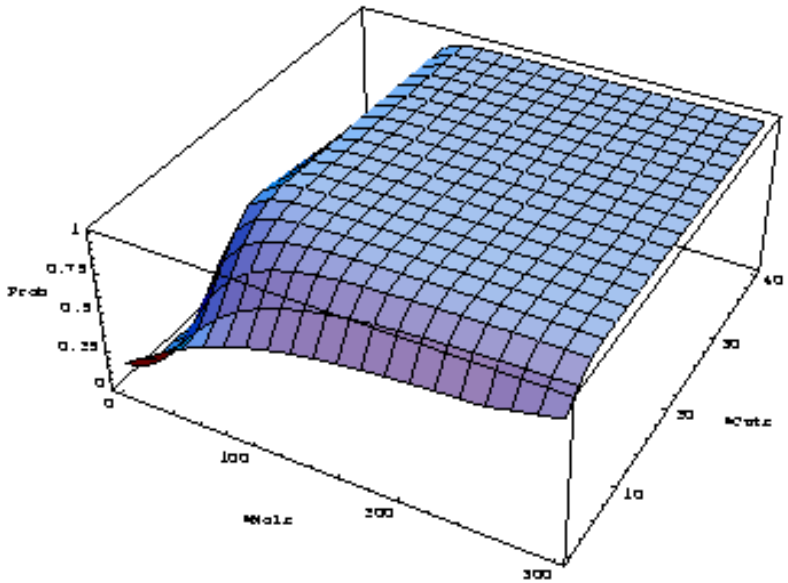
  - *Discretization*: FAILS!!!
  - *Continuous Models*

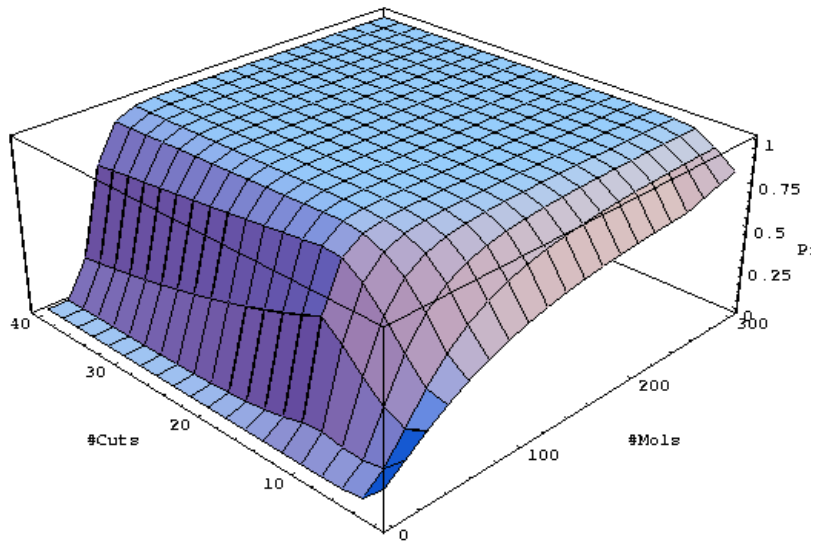
- Sizing Error
  In order to be able to compute the map with high probability, we need to satisfy the condition

$$\sigma \leq \frac{\ln 2}{2k(k-1)p_E}.$$

For BAC's $\sigma \leq 0.89\ bp$; For cosmids $\sigma \leq 60\ bp$.

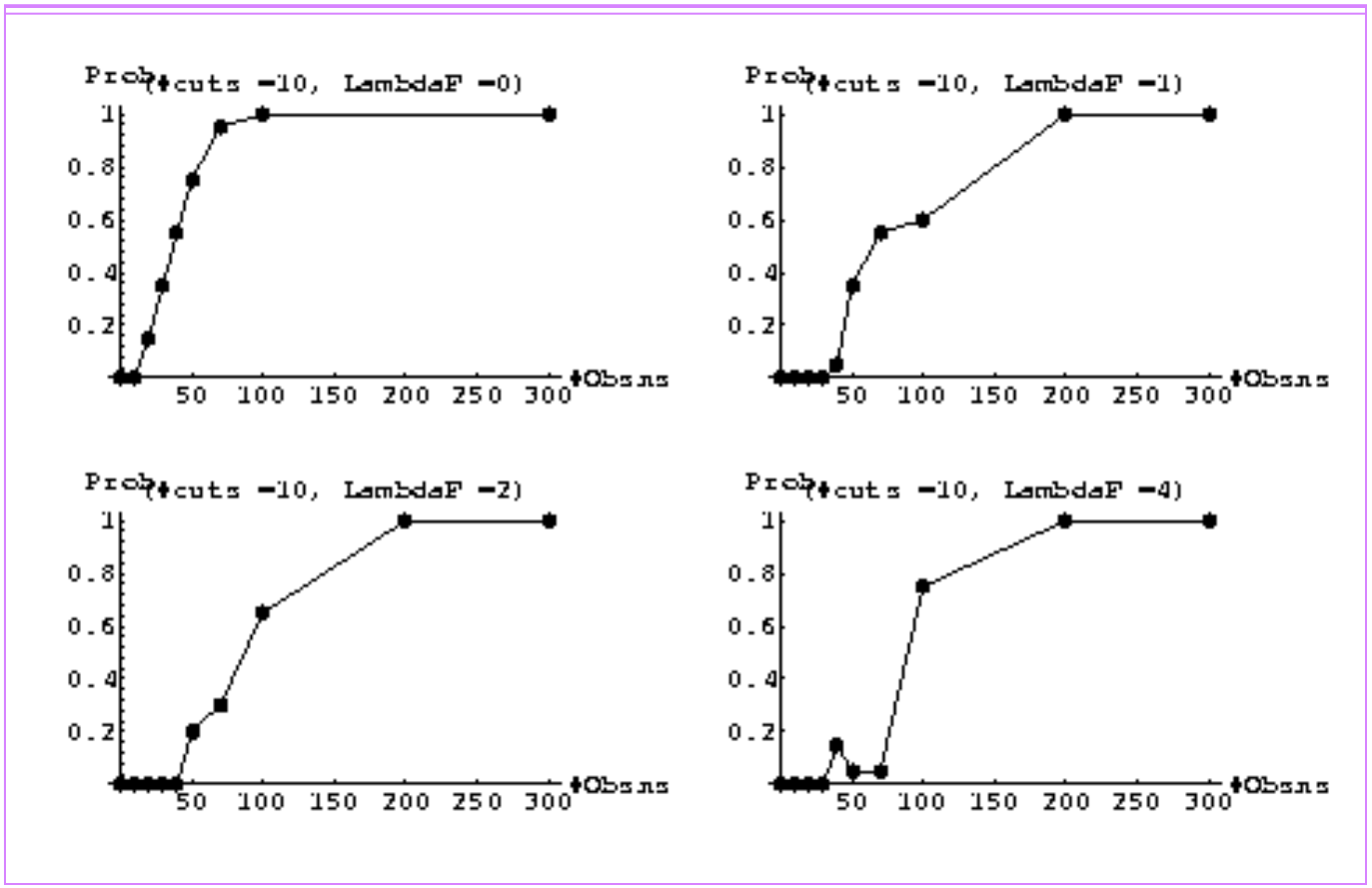
Prediction





The probability of successfully computing the correct restriction map as a function of the number of cuts in the map and number of molecules used in creating the map…

Experimental Results



Gentig: Bayesian Approach

- Model or Hypothesis H
- Prior distribution of the evidence

  $$Pr[D_i|H]$$

  Assume pair-wise conditional independence of the events D_i's

  $$Pr[D_j|D_{i_1},\ldots,D_{i_r},H] = Pr[D_j|H]$$

- Posterior distributions leads to a log-likelihood cost function

  $$\log\left(\frac{Pr[H|D_1,\ldots,D_m]}{Pr[H]}\right)$$
  $$= \text{ Bias terms} + \sum_j \log\left(\frac{Pr[D_j|H]}{Pr[D_j]}\right)$$

- Derive a cost function
- Optimize over a set of hypotheses

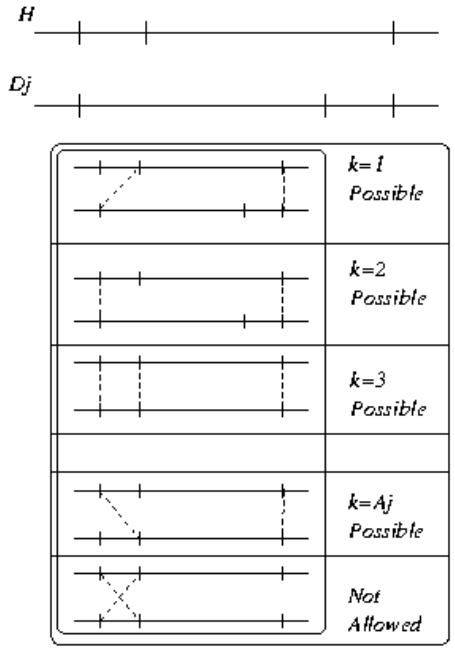
Bayesian Model

- $\mathcal{L} = \sum_j \log \left[p_b e^{-\lambda_n} \lambda_n^{M_j} + \frac{1-p_b}{2} \sum_k Pr_{jk} \right],$

- Where

$$Pr_{jk} = \left[\prod_{i=1}^{N} \left(p_{c_i} \frac{e^{-(s_{ijk}-h_i)^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i} \right)^{m_{ijk}} \right] \times \left[\prod_{i=1}^{N} (1-p_{c_i})^{(1-m_{ijk})} \right] \times e^{-\lambda_f} \lambda_f^{F_{jk}}.$$

Multiple Alignment



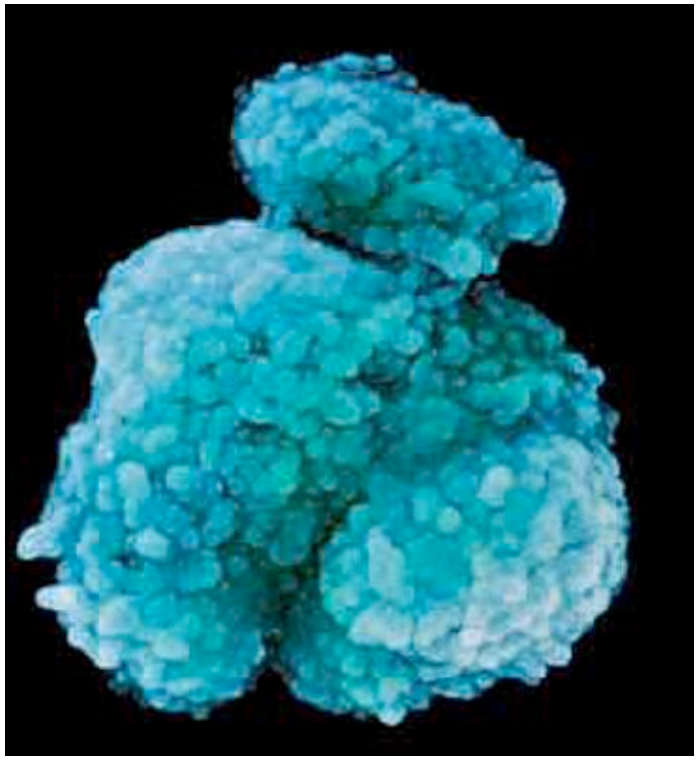
- Various alignments of cuts have to be considered.
- Fast computation is possible...
  via *Dynamic Programming* and additional heuristics

  —Key to our fast implementation.

Robustness

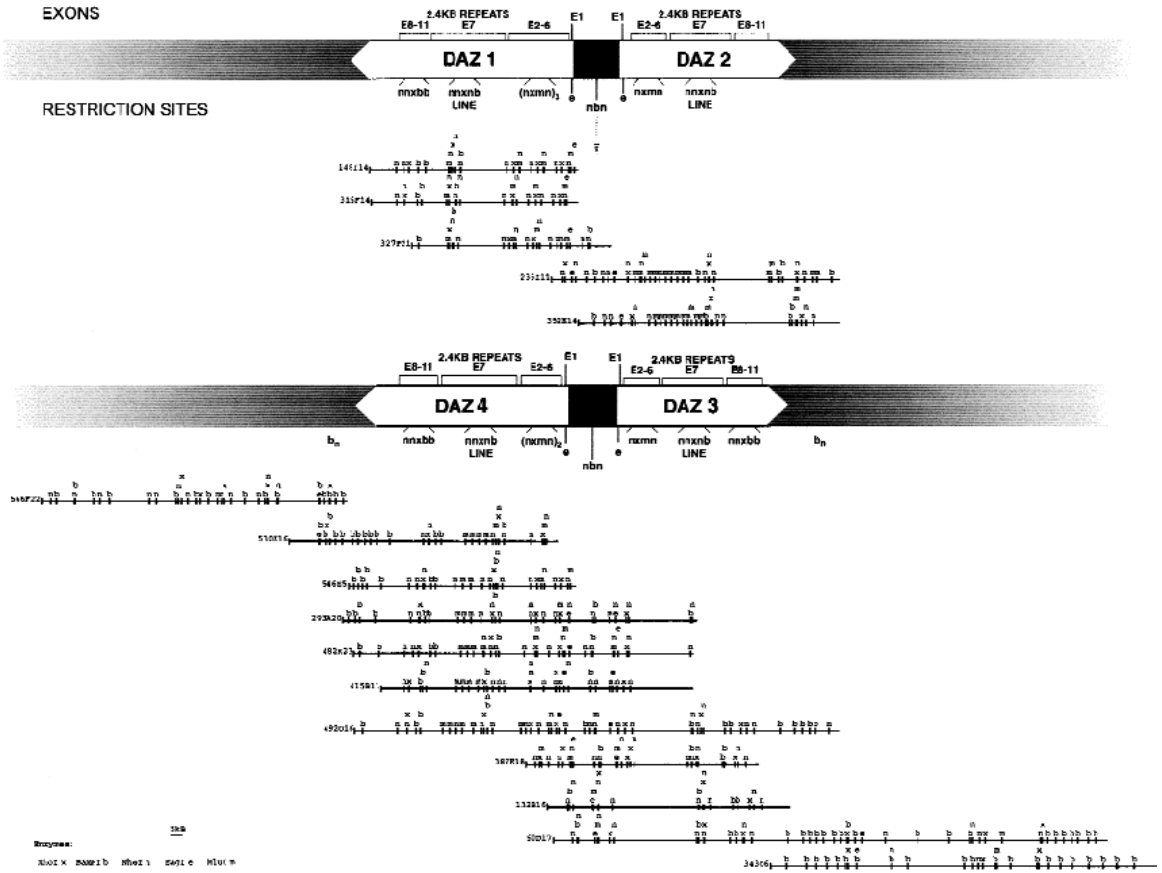
- BAC Clones with 6-cutters
  - Average Clone size = 160 Kb; Average Fragment Size = 4 Kb, & Average Number of Cutsites = 40.
- Parameters:
  - Digestion rate can be as low as 10%
  - Orientation of DNA need not be known.
  - 40% foreign DNA
  - 85% DNA partially broken
  - Relative sizing error up to 30%
  - 30% spurious randomly located cuts…

Y



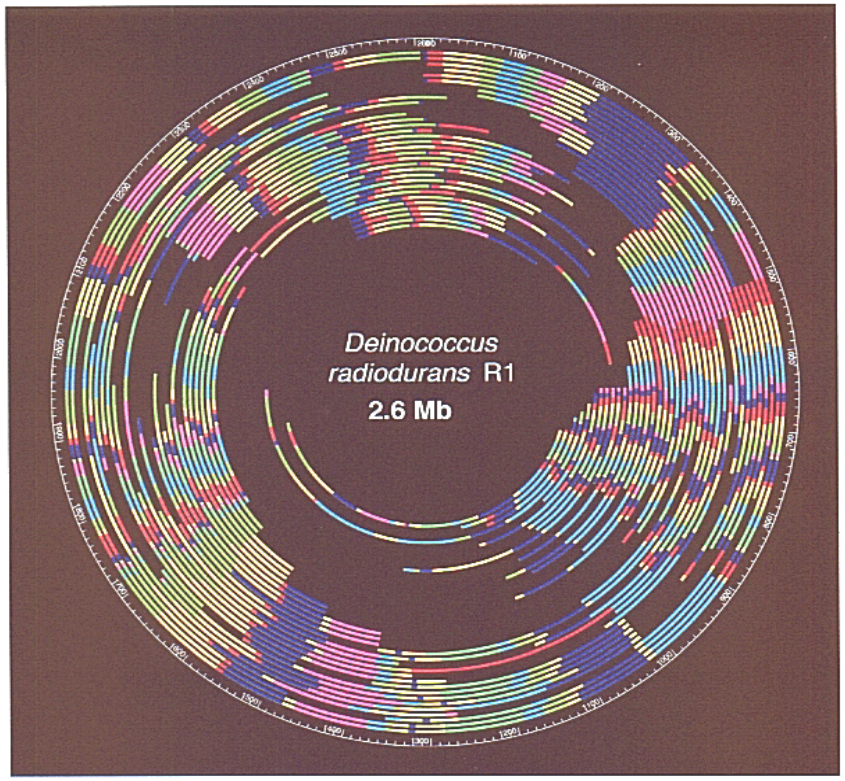
- "From a gene's point of view, reshuffling is a great restorative…
- "The Y, in its solitary state disapproves of such laxity. Apart from small parts near each tip which line up with a shared section of the X, it stands aloof from the great DNA swap. Its genes, such as they are, remain in purdah as the generations succeed. As a result, each Y is a genetic republic, insulated from the outside world. Like most closed societies it becomes both selfish and wasteful. Every lineage evolves an identity of its own which, quite often, collapses under the weight of its own inborn weaknesses.
- "Celibacy has ruined man's chromosome."
  - Steve Jones, Y: The descent of Men, 2002.

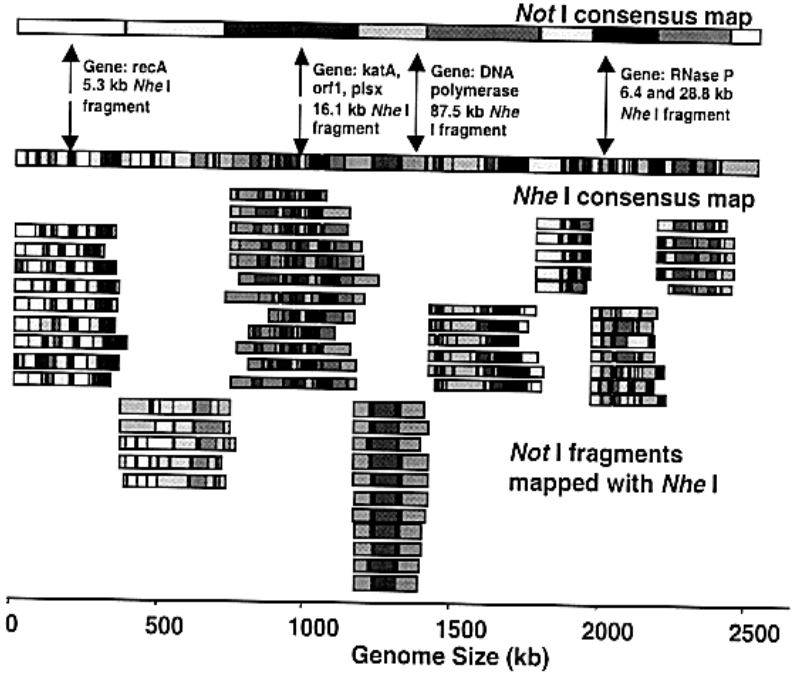
Mapping the DAZ locus on Y Chromosome



Gentig Map

Deinococcus radiodurans

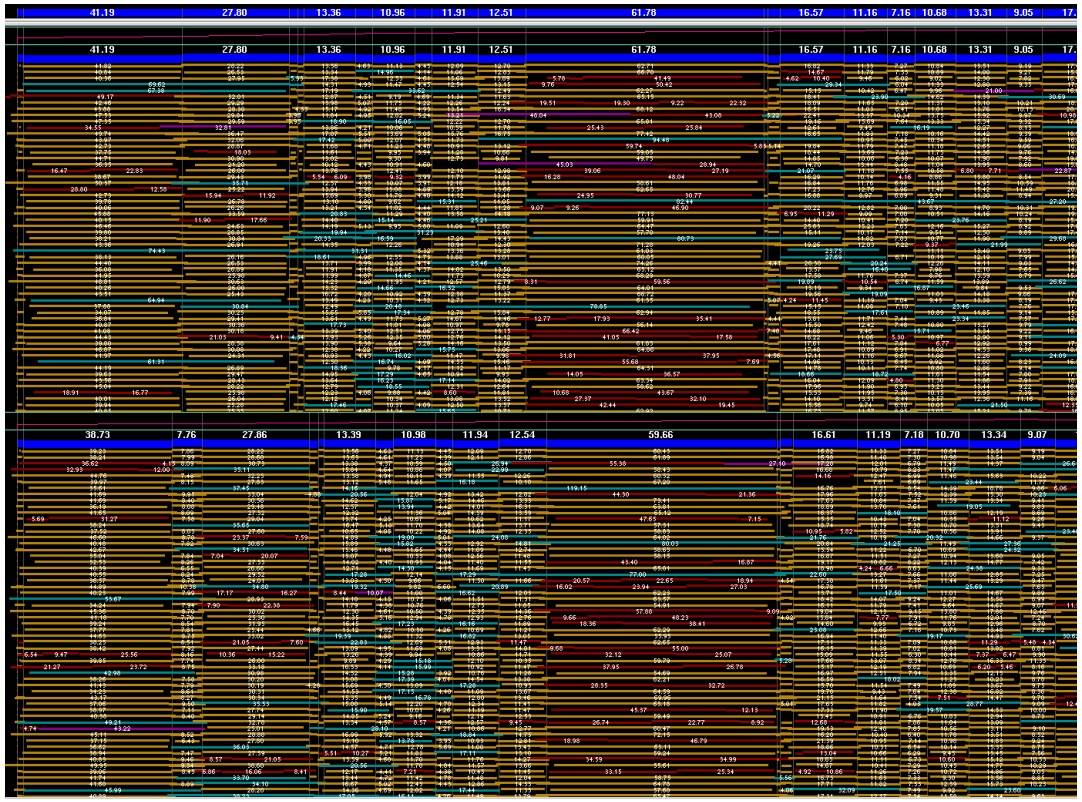




Nhe I map of *D.radiodurans* generated by **Gentig**

Single Molecule Hapoltyping:

Candida Albicans

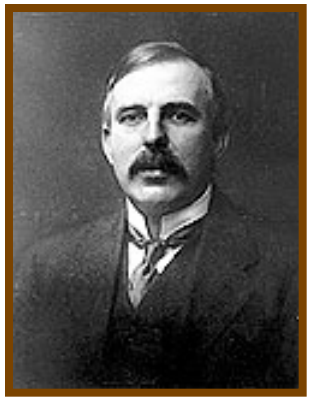


- The left end of chromsome-1 of the common fungus Candida Albicans (being sequenced by Stanford).
- Three polymorphisms:
  - (**A**) Fragment 2 is of size 41.19kb (top) vs 38.73kb (bottom).
  - (**B**) The 3rd fragment of size 7.76kb is missing from the top haplotype.
  - (**C**)The large fragment in the middle is of size 61.78kb vs 59.66kb.

Sequencing

Sir Ernest Rutherford

“We haven't the
money, so we've got
to think."



Problem to Solve…

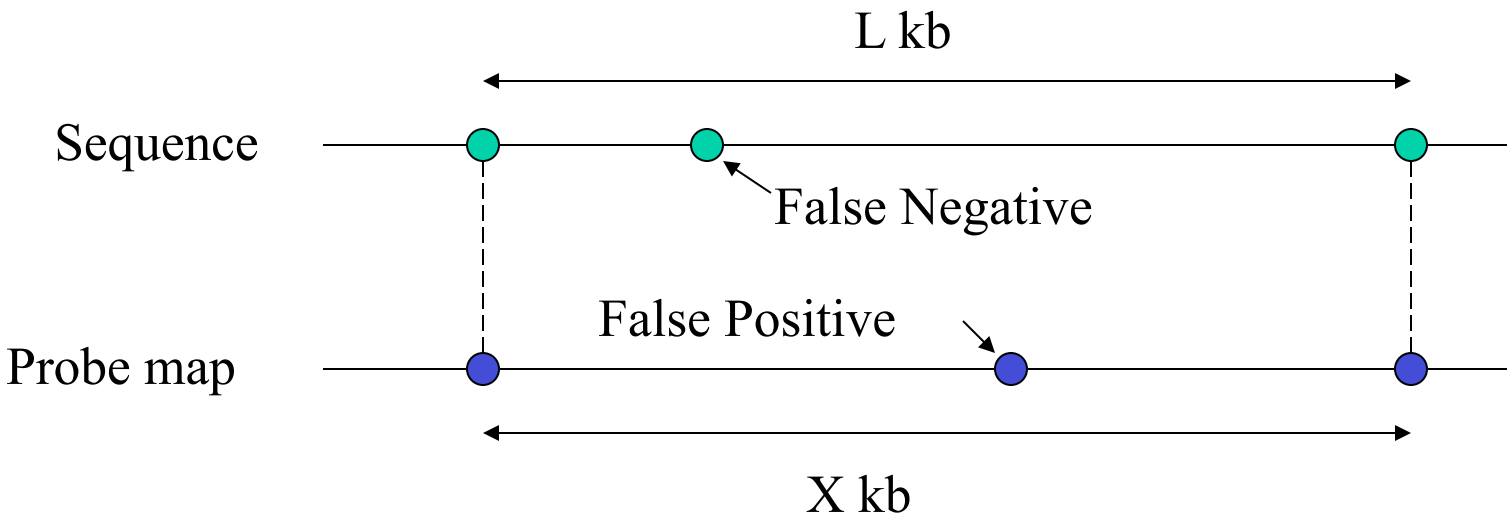
- Given probe maps of some small region of the genome for all N-bp hybridization probes (e.g. all 2080 probes of 6-bp).
- With known error rates (false positive, false negatives and sizing errors).
- **Can we reconstruct the complete sequence ?**

- Estimated Error rates for consensus probe maps from 40x data redundancy :
  - False Negative rate = 2%
  - False Positive rate = 0.006/kb (2.4% ratio for 6-bp probes)
  - Gaussian error sd = 60bp

Basic reconstruction algorithm

- Keep track of multiple sequence assemblies.
- Initialize with all possible 5-bp sequences.
- Try all 4 possible extensions of each sequence.
- Check if probe is present in corresponding map : if not add a penalty score to the sequence involved.
- Periodically delete sequences with high penalty.
- Stop when missing probe rate jumps significantly from False Negative rate (2%) to (100% - false extension rate) = 55%.
- Return highest scoring sequence.

Aligned probe pair



Likelihood computation

Let:

$Pc = 1 -$ False Negative Rate

$\lambda =$ False Positive Rate per kb

$\sigma\sqrt{L} =$ standard deviation of probe interval when sequence = L kb

Then the log-likelihood term for each aligned probe pair :

$$LL = \log(Pc) - 0.5\log\left(2\pi\sigma^2 L\right) - \frac{(X-L)^2}{2\sigma^2 L} + FP\log(\lambda) + FN\log(1 - Pc)$$

where:

$X =$ measured distance between aligned probes

$FP =$ Number of false positives between aligned probes

$FN =$ Number of false negatives between aligned probes

Anomalies

- Irresolvable Ambiguities:
  - From assemblies based on 6bp probes

```
Assembly:…tcgccCCCCTAAC ggcga…
              || || |          || | ||
Correct   :…tcgccGTTAGGGGggcga…
```

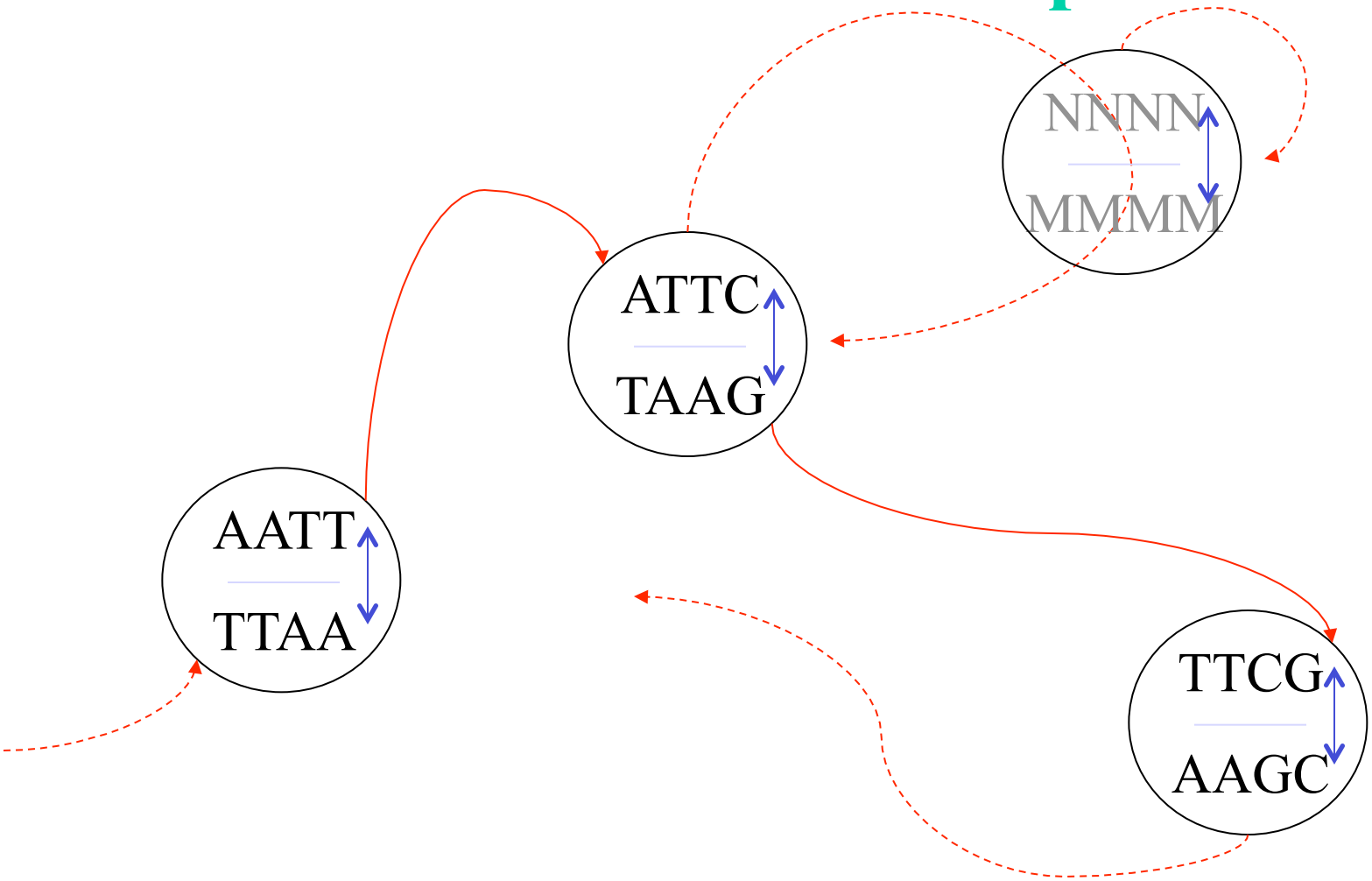
  - Error Pattern : s w s^{RC}
  - Correct Pattern : s w^{RC} s^{RC}
    - s = tcgcc (any 5 bases)
    - s^{RC}=ggcga (Reverse compliment of X)
    - w = CCCCTAAC (any short sequence under 50bp)
    - w^{RC}= GTTAGGGG (Reverse compliment of Y)

- Irresolvable Ambiguities & Unavoidable Error Patterns
  - Most common: $\sigma\ \omega\ \sigma^{RC}$ vs $\sigma\ \omega^{RC}\ \sigma^{RC}$
  - Also common: $\sigma\ \omega\ \sigma\ \tau\ \sigma$ vs. $\sigma\ \tau\ \sigma\ \omega\ \sigma$
- Many more rare/complicated patterns

  σ = any K-1 bp sequence

  ω, τ = any short sequence under 50bp
- The probabilities of such patterns can be reduced exponentially with "gapped probes" without increasing the costs.

Directed Eulerian Graph

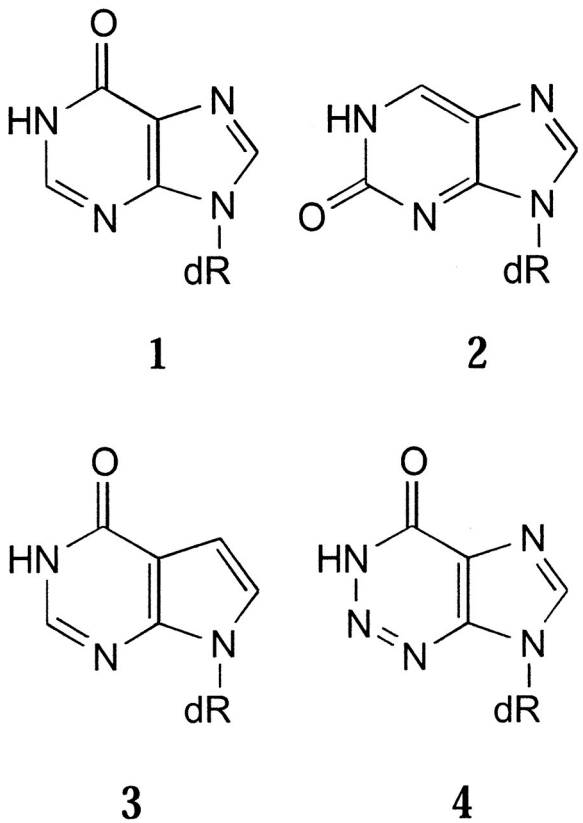


. . . .

- Mixing ‘solid’ bases with `wild-card’ bases:
  - E.g., xx-x-x-xx (9-mers) or xxx- -x- -x- -xxx (14 mers)
- An ‘inert’ base
  - Universal: In terms of its ability to form base pairs with the other natural DNA/RNA bases.
- Examples:
  - The naturally occurring base hypoxanthine, as its ribo- or 2'-deoxyribonucleoside; 2'-deoxyisoinosine; 7-deaza-2'-deoxyinosine; 2-aza-2'-deoxyinosine

2'-Deoxyinosine derivatives

- 2'-Deoxyinosine derivatives can be used as universal DNA analogues.



Loakes, D. Nucl. Acids Res. 2001 29:2437-2447; doi:10.1093/nar/29.12.2437

Nucleic Acids Research

Gapped Probes

- Gapped probes have inert wild-card bases.
- Patterns simulated include:

  xxx-xxx (6 normal, 1 gapped base)

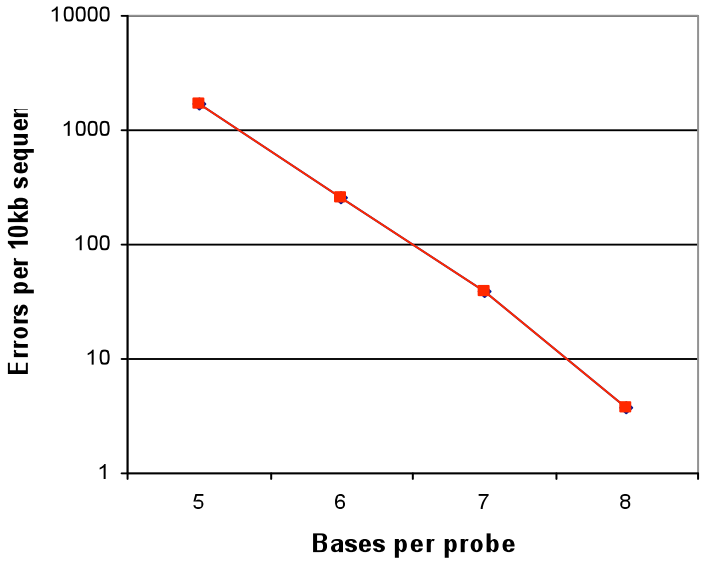
  xx-xx-xx (6 normal, 2 gapped bases)

  xx-x-x-xx (6 normal, 3 gapped bases)

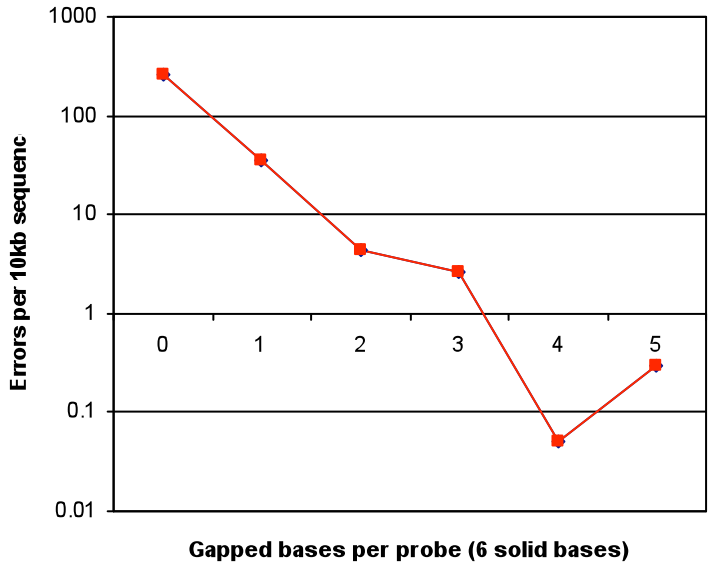
  xx-x--x-xx (6 normal, 4 gapped bases)

  xx--x-x--xx (6 normal, 5 gapped bases)

Simulation Results (Random Sequence)



UNGAPPED



GAPPED

Future

- Sequence Millions of Humans (about 0.05% of the entire Population) Haplotypically
  - with Accurate Characterization of SNP, Indel & Rearrangement Polymorphisms
- Create an Island-Coalescent Model to Characterize Genomics of Human Population
  - Mutations, Duplications, Gene Conversion, Recombination & Migration
  - Population Bottlenecks
  - Positive Selection and Gene-sweeps
  - Negative Selection and Rare Variants
- Novel Algorithms to Stochastically Model Population Structures
  - Nonparametric Models
  - diFenetti's Idea of Exchangeability & Nonparametric Models
  - Algorithms to Estimate the Stochastic Process
- Association Studies
  - Find Disease Markers
  - Origin and Progression of Diseases
  - Individualized Medicine

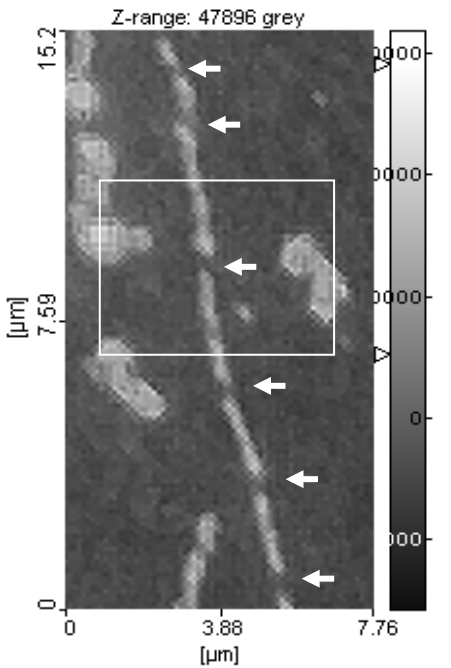
Translational Biotechnology

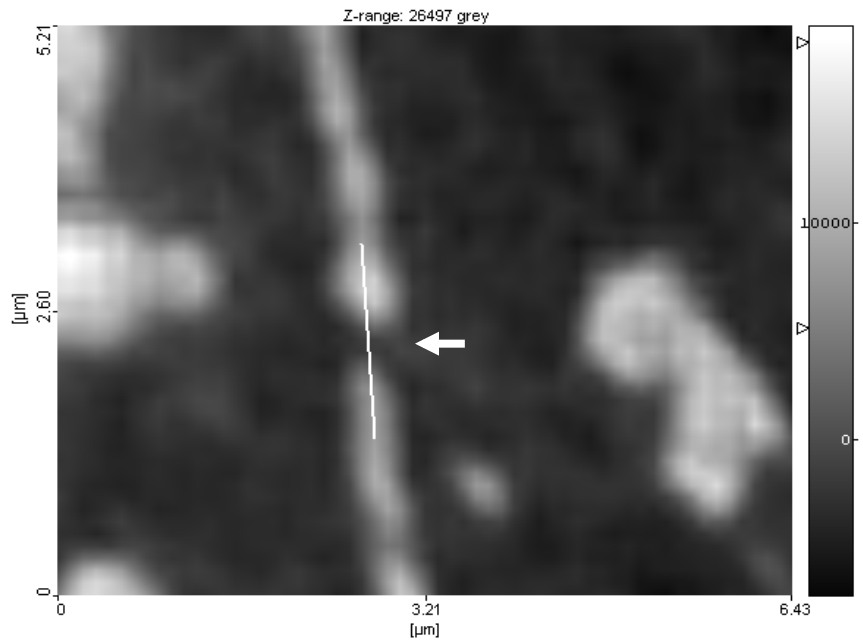
- Cheap and fast technologies for
  - Genomics
  - Epigenomics
  - Transcriptomics
  - Proteomics
- Are the currently leading technologies aiming at the correct solution?
  - Roche/454
  - Illumina/Solexa
  - ABI/Agencourt

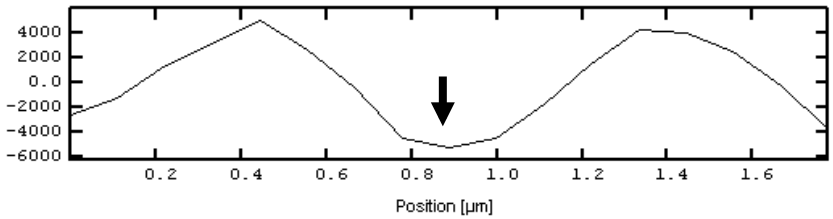
Whole Genomics Sequencing

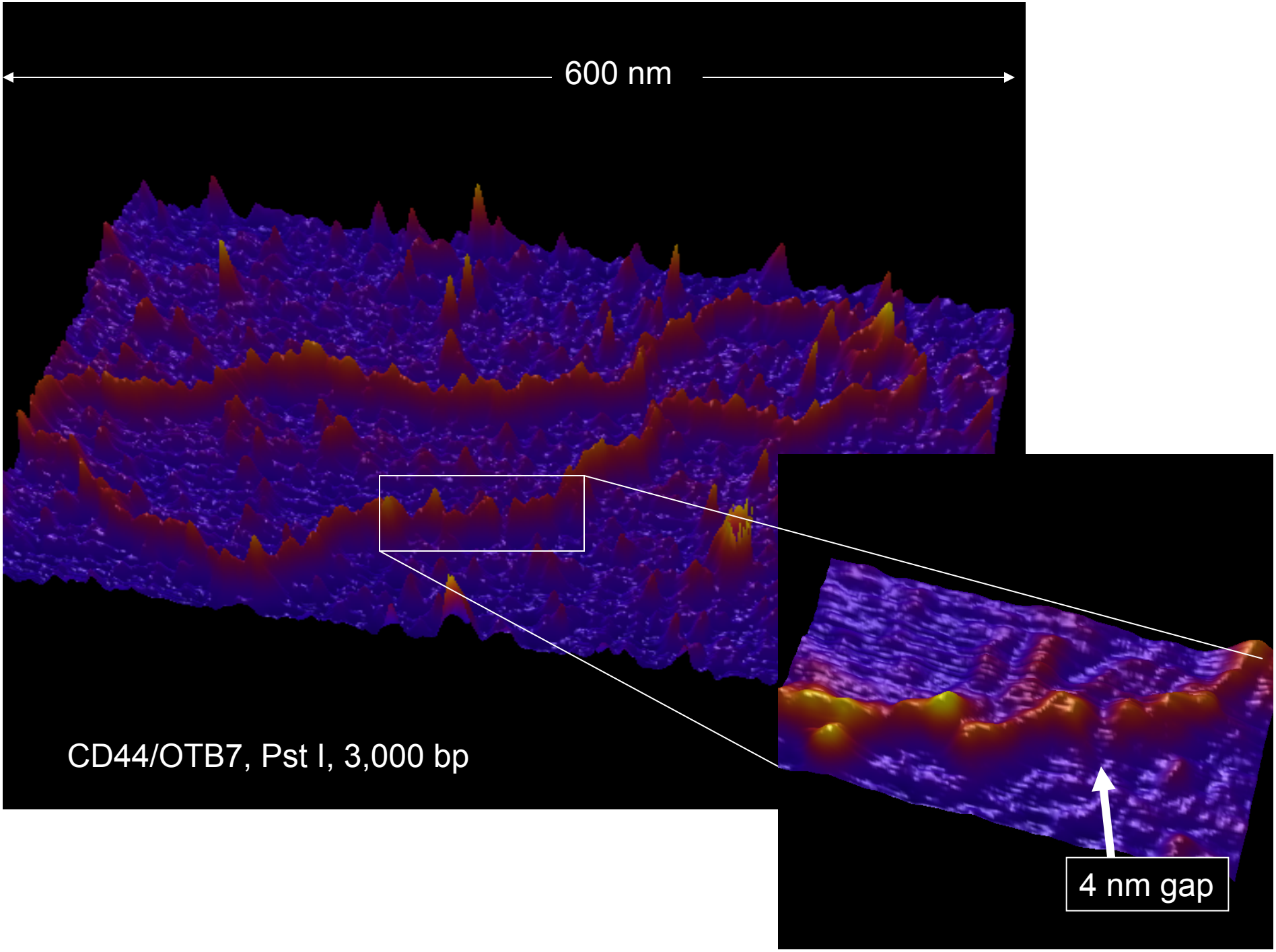
- Gap free sequences:
  - Think about rearrangements, copy-numbers, translocations, etc.
- Genotypes or Haplotypes:
  - Think about SNP's, LOH, etc.
- Short Repeats:
  - Think how to count copy number "accurately"
- Homopolymers:
  - Think about frame-shifts, etc.

Initial Experiments

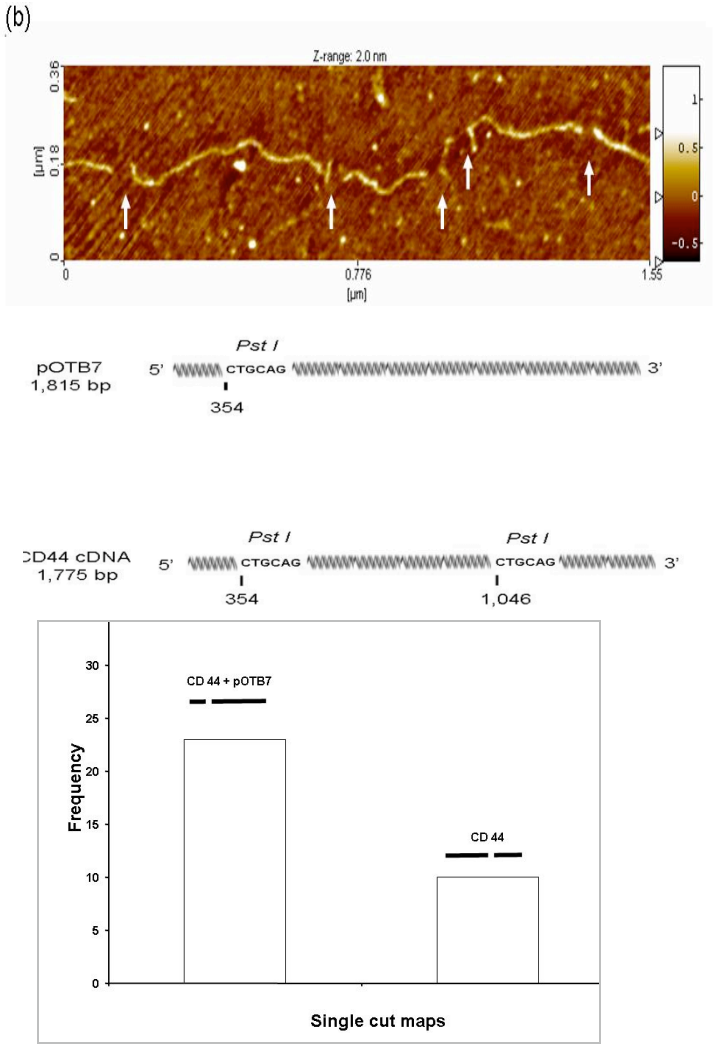




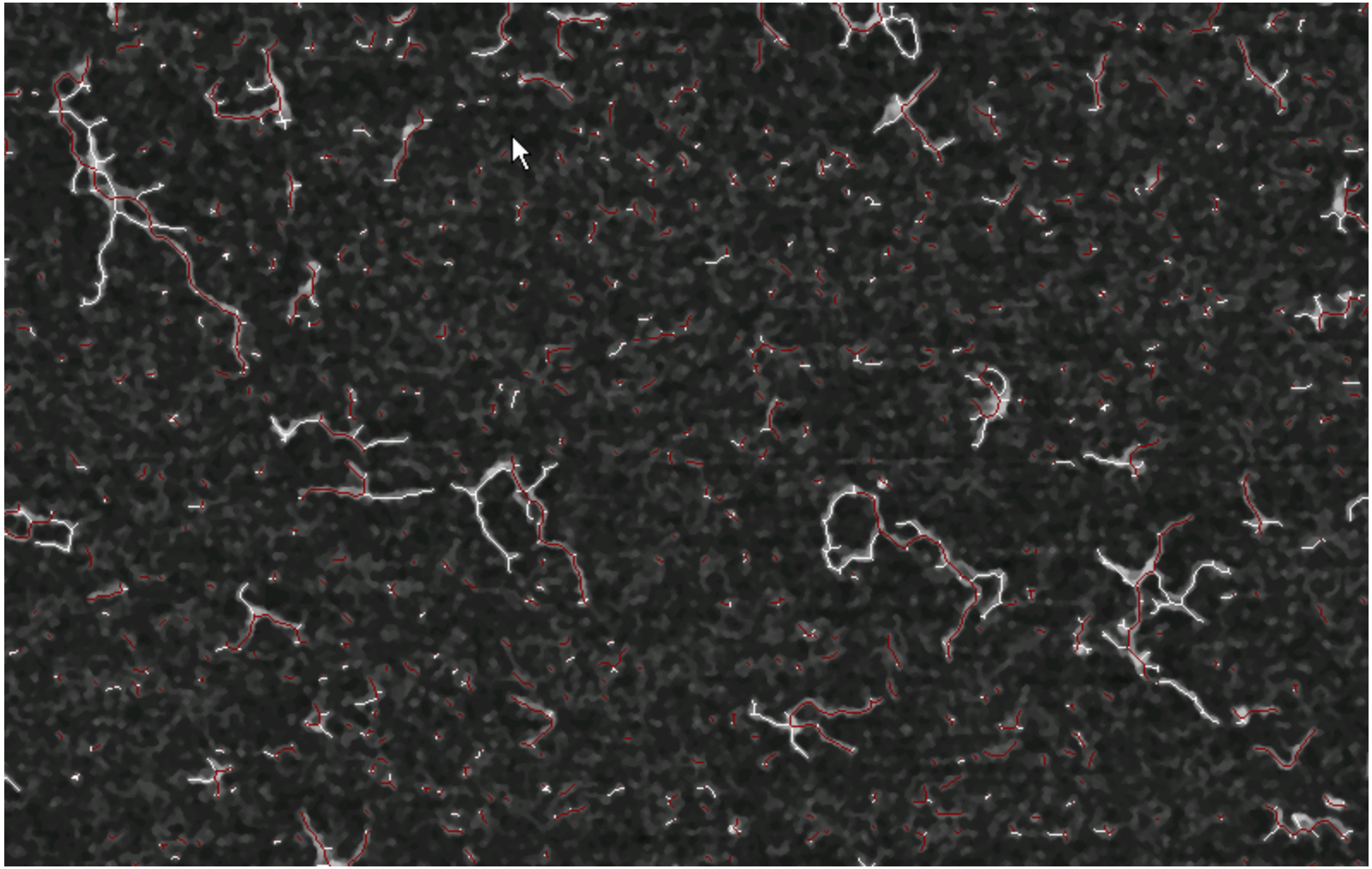



600 nm
CD44/OTB7, Pst I, 3,000 bp
4 nm gap

Single Molecule DNA Profiling



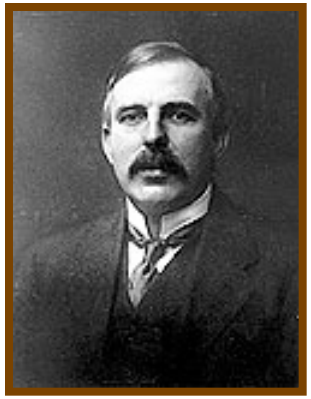
- **Applications:**
- Quantitation of a mixture of CDNAs from genes (with isoforms)
  - The human CD44 gene encodes an 80 kD, 742 amino acid cell-surface glycoprotein involved in cell-cell interaction and tumor metastasis.
  - CD44 protein is expressed as multiple isoforms in hematopoietic, lymphoid and epithelial tissues.
  - Individual mRNA isoforms of CD44 may signal tumor progression and their detection in surgical biopsies has been postulated as a biomarker for metastatic potential.
- A binary mixture containing one part cDNA from the human CD44 gene and one part DNA plasmid pOTB7.



Sir Ernest Rutherford

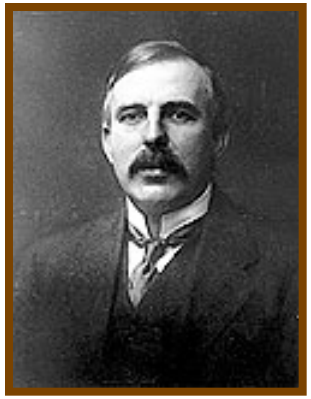
"I have become more and more impressed by the power of the scientific method of extending our knowledge of nature.

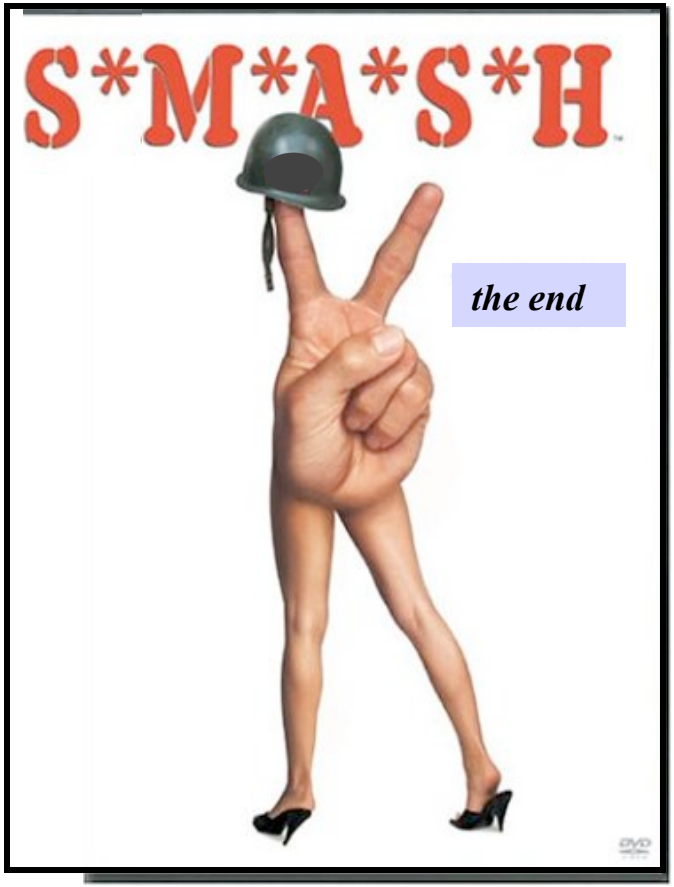
Experiment, directed by the imagination of either an individual, or still better of a group of individuals of varied mental outlook is able to achieve results which far transcend the imagination alone of the greatest natural philosopher."



Sir Ernest Rutherford

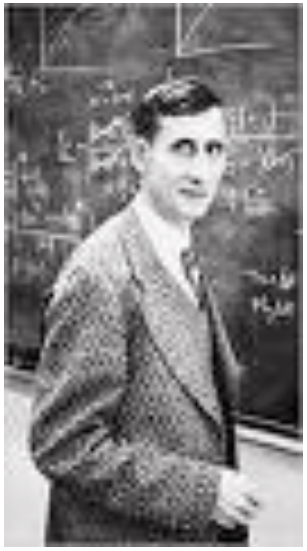
"Experiment without imagination, or imagination without recourse to experiment, can accomplish little. But for effective progress, a happy blend of these powers is necessary"





the end

Laptop Genome Sequencer



*

"What biology now needs is a single-molecule sequencer …

. . .

"A single-molecule machine could be much cheaper as well as faster than existing machines. It might be as small and convenient as a lap-top computer…"

* * * * * * *