

Gene Expression Tiling Array Technology – A Steppingstone to Modern Genome Science at NASA

Viktor Stolc, PhD
2021 H. Julian Allen Award

Climactic moments in the history of DNA – INFORMATION SUBSTRATE OF LIFE

140 years



Gregor Mendel (1865)

- **Heritable** factors (*genes*), which retain their individuality, generation after generation,
-economic return on improving crops (opportunity) + curiosity = practical science

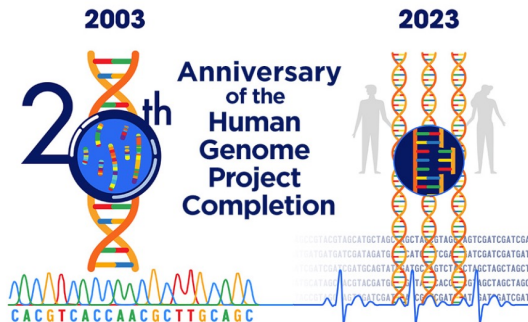


Rosalind Franklin (1952), X-ray diffraction images of DNA

- **Structure** of DNA specified the functional mechanism for copying information, encoding information in the genetic code, and production of RNA and protein catalysts for cellular biochemistry.



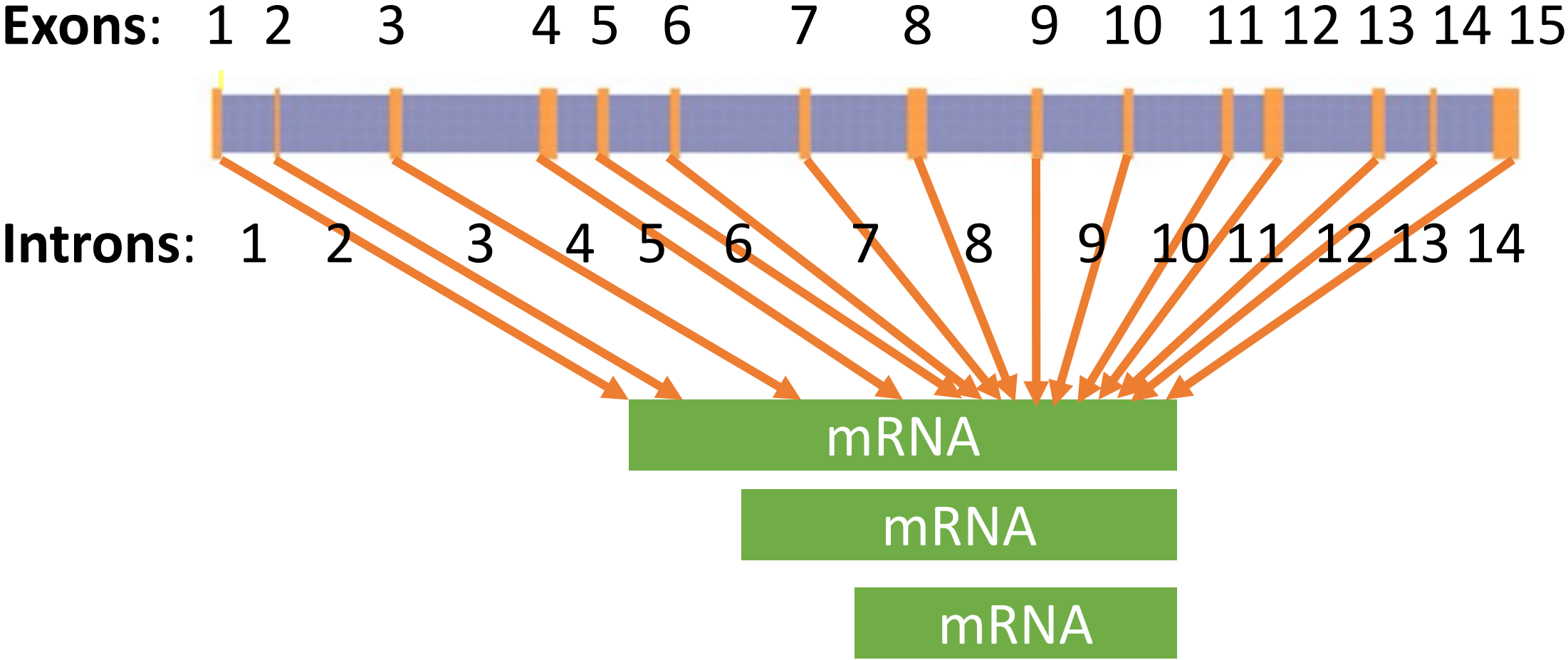
James Watson & Francis Crick (1953)



- **Sequence** of the Human Genome, the human blueprint of life required more than a decade to complete.

Craig Venter & Francis Collins (2003)

GENE STRUCTURE – DNA to mRNA



GENE STRUCTURE – DNA to mRNA

Exons: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15



Introns: 1 2 3 4 5 6 7 8 9 10 11 12 13 14



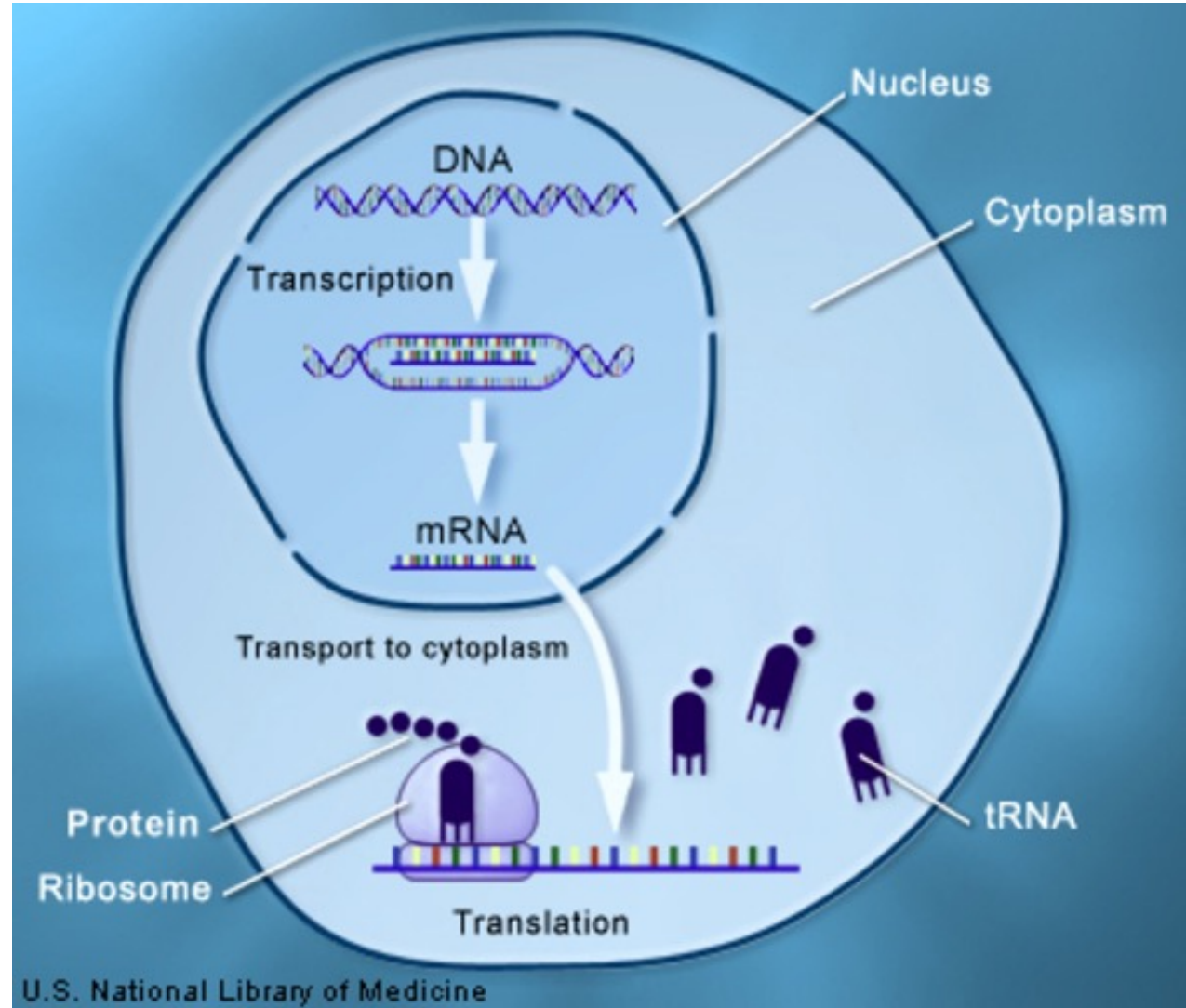
Transcription

mRNA

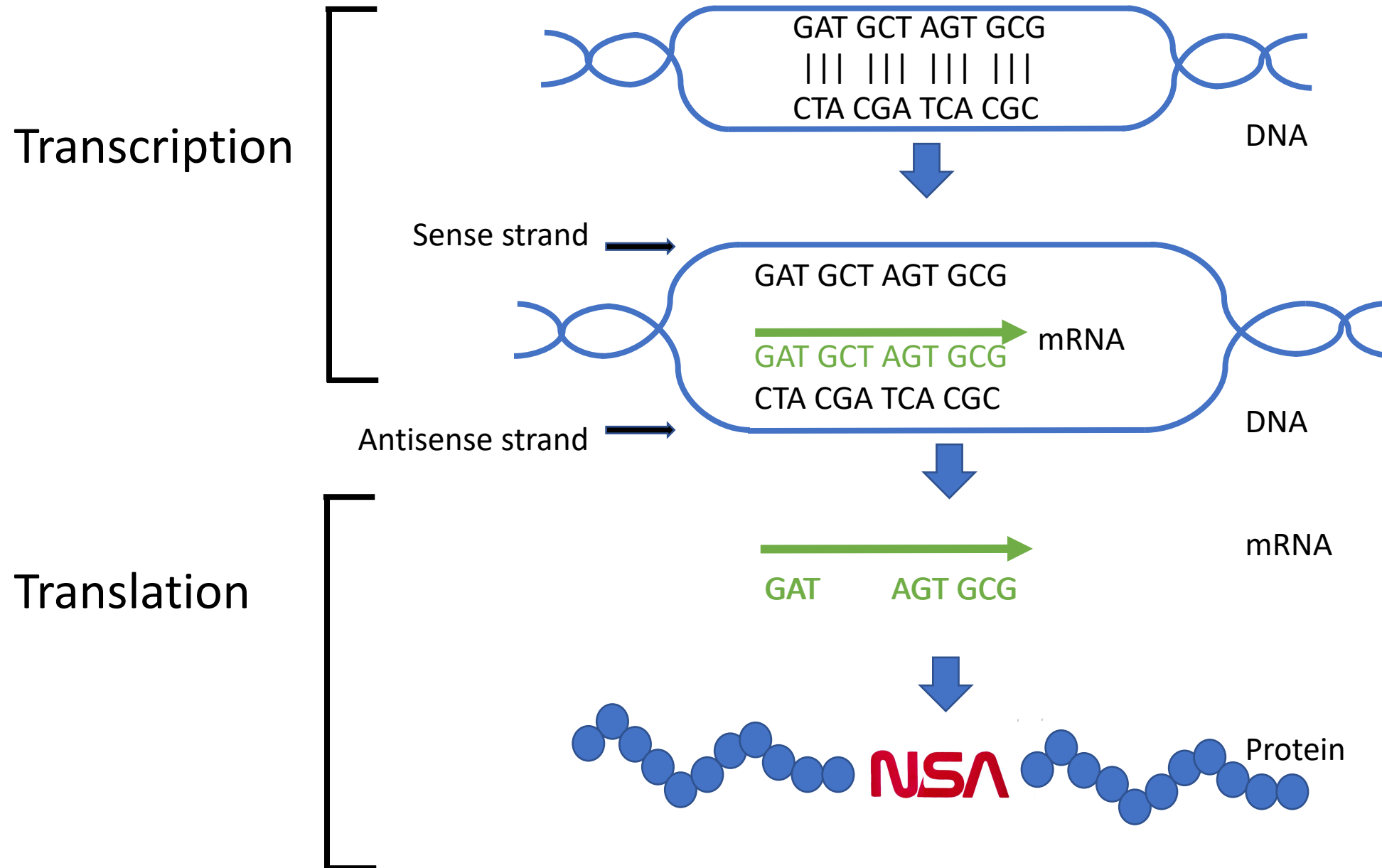
mRNA

mRNA

Protein Synthesis in a Cell



What is Transcription? What is Translation?



Information flow & processing in biological systems

The same genome specifies different cell types and functions



Measurement of mRNAs: Sifting through 6 billion nucleotides for instructions in RNA code that specify gene functions – this required big computers to attribute function to **specific sequences (oligonucleotide probes)**.

The extraordinary numbers associated with DNA and RNA

- 6,109,647,513 base pairs of information in the entire genome in each and every cell,
- All this information is somehow packed into a polymer (DNA) that is ~4 yards long,
- This 4 yards of DNA is packaged into 23 chromosomes inside a cell that is about 100 microns in diameter,
- There are about 40,000 genes in each human cell,
- Any one cell uses only a tiny fraction of the total information contained in the genome in order to make proteins that make each cell unique,
- The system is further complicated by the fact that most of the RNAs in any cell do not code for proteins! We are still trying to sort out the functions of these non-coding RNAs.

Project Goal

The goal was to identify and quantify the amounts of mRNAs being produced from the genome by the process of transcription.

Not just for few genes, not just for hundreds of genes, *but for every single gene in the human genome.*

News Headline - June 27, 2000

The New York Times

“Genetic Code of Human Life Is Cracked by Scientists”

Anticipated to lead a new era of medicine -- one that delivers the right treatment at the right time.

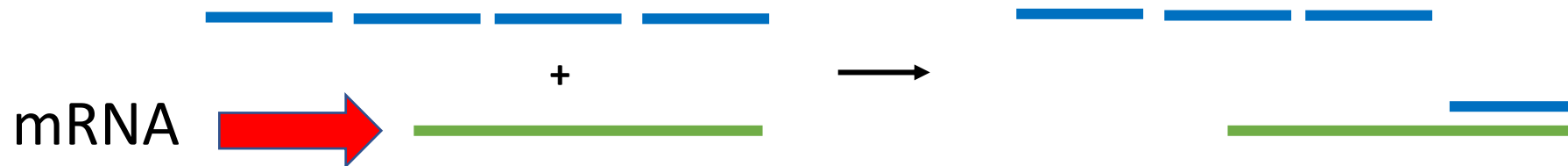
How did we go about using the newly acquired,
“complete human genome sequence” to identify all
the mRNAs synthesized
(all the transcription activity)?

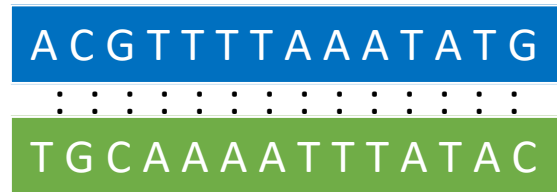
How to map the transcribed regions of the genome?



A DNA oligonucleotide can be synthesized to match a short segment of DNA in the genome.

Now, harvest all the mRNA from tissue, so we can ask the question, “do any of the mRNA strands correspond to DNA oligonucleotide probes that were made?”





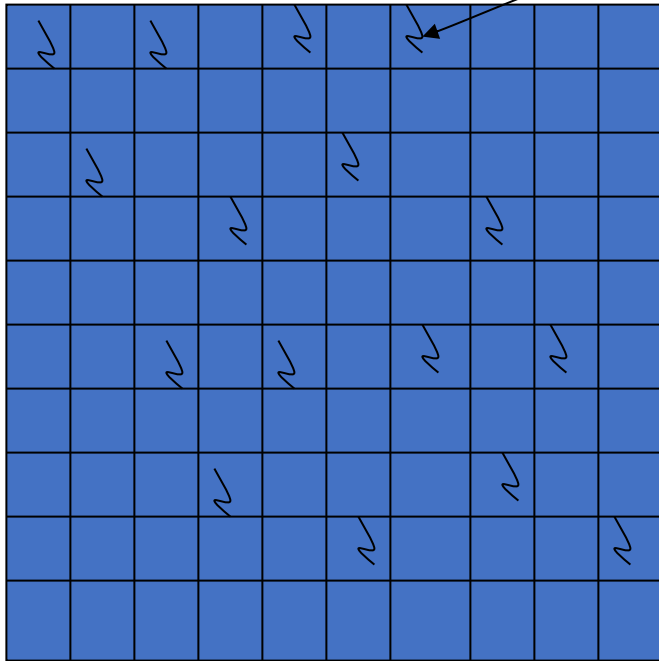
When a probe sequence is “complementary” to mRNA, a hybridization complex forms.

“Complementary” means that A’s and T’s are in the correct so that they can bind to each other, and C’s and G’s are in the correct position, so that they can bind to each other.

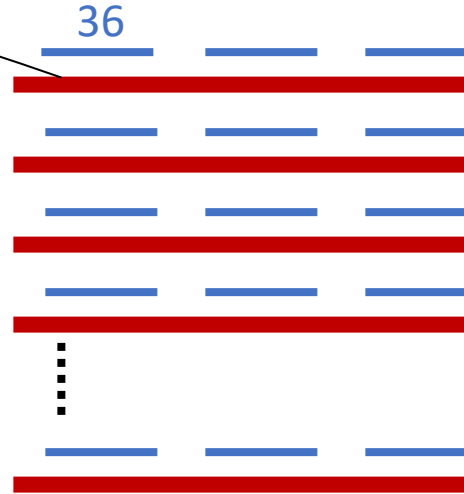
A stretch of 36 bases is the optimal probe length to ensure sequence specificity and sensitivity of hybridization.

Microarray design and capability - Whole Genome Tiling

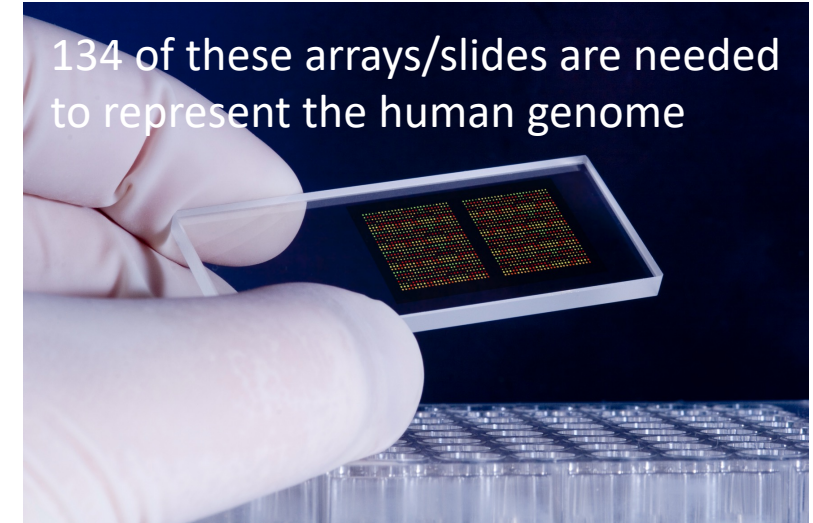
Chromosome representation: unbiased approach



780,000 probes per array



both DNA strands.



134 of these arrays/slides are needed to represent the human genome

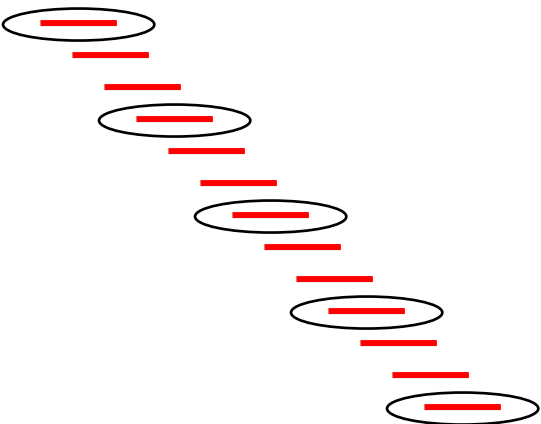
- 134 arrays/slide (~1.5Gb) *per* human genome,
- 51,874,388 36-nucleotide (nt) probes, positioned every 46 bases,
- But the situation is complicated by the fact that some sequences are only present only once in the genome (“unique”), while other sequences are present in multiple places in the genome (“non-unique”).

oligonucleotide sequence	Location	Strand	Frequency																			
ATGGGCGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	1	+	1.650000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
TGGGCGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	2	+	1.600000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GGGCGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	3	+	1.550000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GGCGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	4	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GCGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	5	+	1.450000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
CGAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	6	+	1.400000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GAACGACGGGAATTGAACCCGCGATGGTGAA	chlo	7	+	1.350000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
AACGACGGGAATTGAACCCGCGATGGTGAA	chlo	8	+	1.300000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
ACGACGGGAATTGAACCCGCGATGGTGAA	chlo	9	+	1.250000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
CGACGGGAATTGAACCCGCGATGGTGAA	chlo	10	+	1.200000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GACGGGAATTGAACCCGCGATGGTGAA	chlo	11	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
ACGGGAATTGAACCCGCGATGGTGAA	chlo	12	+	1.100000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
CGGGAAATTGAACCCGCGATGGTGAA	chlo	13	+	1.050000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GGGAATTGAACCCGCGATGGTGAA	chlo	14	+	1.000000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GGAATTGAACCCGCGATGGTGAA	chlo	15	+	1.000000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
GAATTGAACCCGCGATGGTGAA	chlo	16	+	1.000000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
AATTGAACCCGCGATGGTGAA	chlo	17	+	1.000000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
ATTGAACCCGCGATGGTGAA	chlo	18	+	1.050000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TTGAACCCGCGATGGTGAA	chlo	19	+	1.100000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TGAACCCGCGATGGTGAA	chlo	20	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GAACCCGCGATGGTGAA	chlo	21	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
AACCCGCGATGGTGAA	chlo	22	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ACCCGCGATGGTGAA	chlo	23	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CCCGCGATGGTGAA	chlo	24	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CCGCGATGGTGAA	chlo	25	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CGCGATGGTGAA	chlo	26	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GCGATGGTGAA	chlo	27	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GCATGGTGAA	chlo	28	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GATGGTGAA	chlo	29	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ATGGTGAA	chlo	30	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TGGTGAA	chlo	31	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GGTGAA	chlo	32	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GTGAATTCACAA	chlo	33	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TGAATTCACAA	chlo	34	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GAATTCACAA	chlo	35	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
AATTCACAA	chlo	36	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ATTTCACAA	chlo	37	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TTCAATTCACAA	chlo	38	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TCAATTCACAA	chlo	39	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CACAATTCACAA	chlo	40	+	1.150000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ACAATTCACAA	chlo	41	+	1.200000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CAATTCACAA	chlo	42	+	1.250000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
AATTCACAA	chlo	43	+	1.300000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ATTCACAA	chlo	44	+	1.350000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TCCACTGCCTTAATCCACTTGGCTACATCCGCCCT	chlo	45	+	1.400000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CCACTGCCTTAATCCACTTGGCTACATCCGCCCTTA	chlo	46	+	1.450000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CACTGCCTTAATCCACTTGGCTACATCCGCCCTTAC	chlo	47	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
ACTGCCTTAATCCACTTGGCTACATCCGCCCTTACG	chlo	48	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
CTGCCTTAATCCACTTGGCTACATCCGCCCTTACGC	chlo	49	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
TGCCTTAATCCACTTGGCTACATCCGCCCTTACGCT	chlo	50	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2
GCCTTAATCCACTTGGCTACATCCGCCCTTACGCTA	chlo	51	+	1.500000	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	2

How do we choose the probes so that we cover the genome completely?

In silico oligonucleotide design, with N-mer windowing, where N equals the number of bases in a probe.

This is a “4^N problem” because there are 4 possible bases in each position of a probe.



5' genomic DNA 3'

By computing the frequency of all possible probes, we were able to select only unique probes, by using the NASA Ames Supercomputer.

Step 1: NAS Computing power for optimizing oligonucleotide probes

- Large size of human genome precluded the construction of a genome-wide high-resolution tiling arrays w/o HPCs,
- In 2004, Teraflops used for weeks trillion (10^{12}) calculations per second,
- In 2022, Petaflops used for an hour thousand trillion (10^{15}) calculations per second,
- Digitize genome activity by converting the 4 letter DNA code into 0s & 1s in a computer.

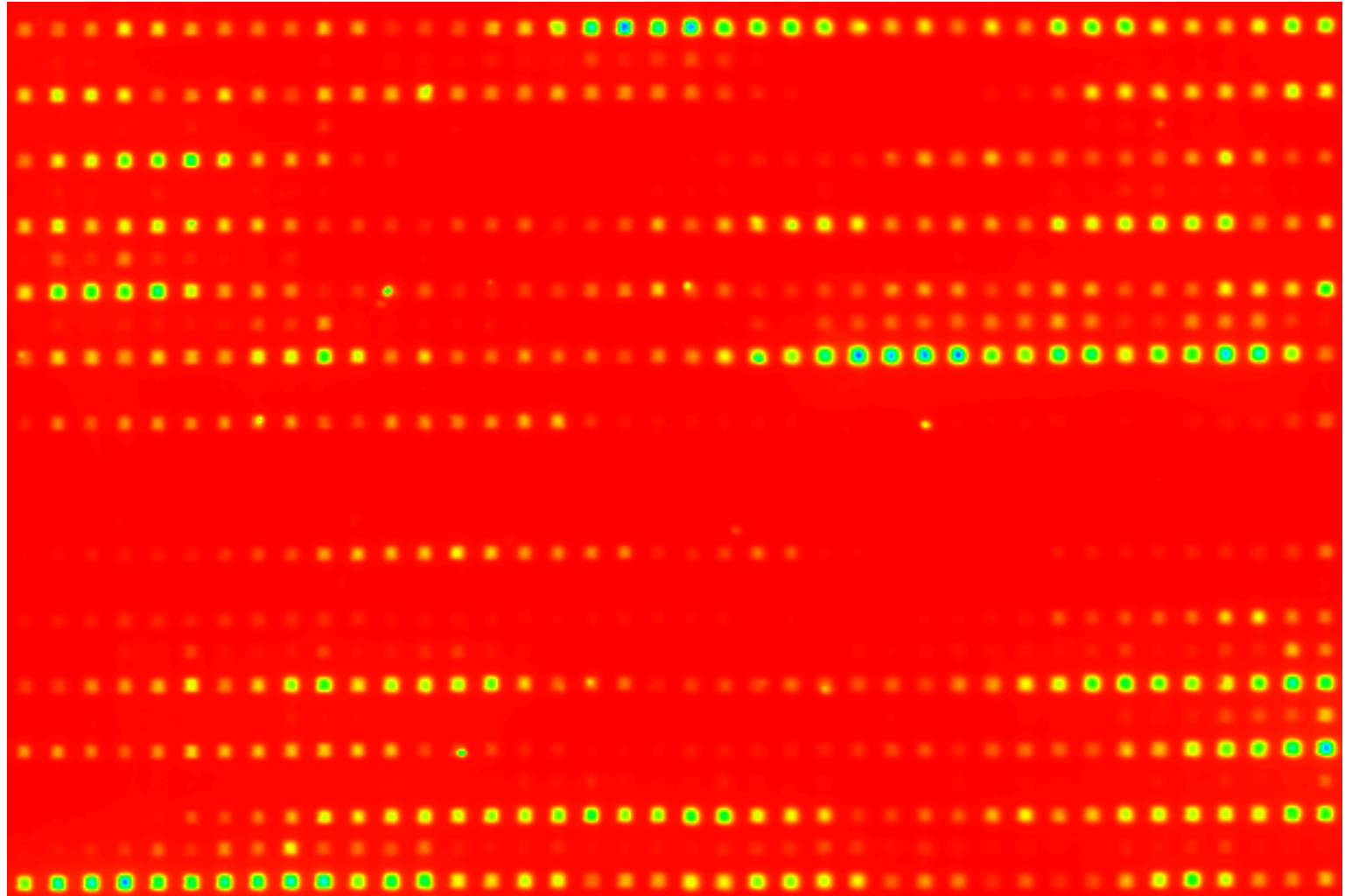


The scale keeps changing dramatically, so HPCs are essential tools of Genome Science.

Step 2: DNA Probe Empirical Optimization

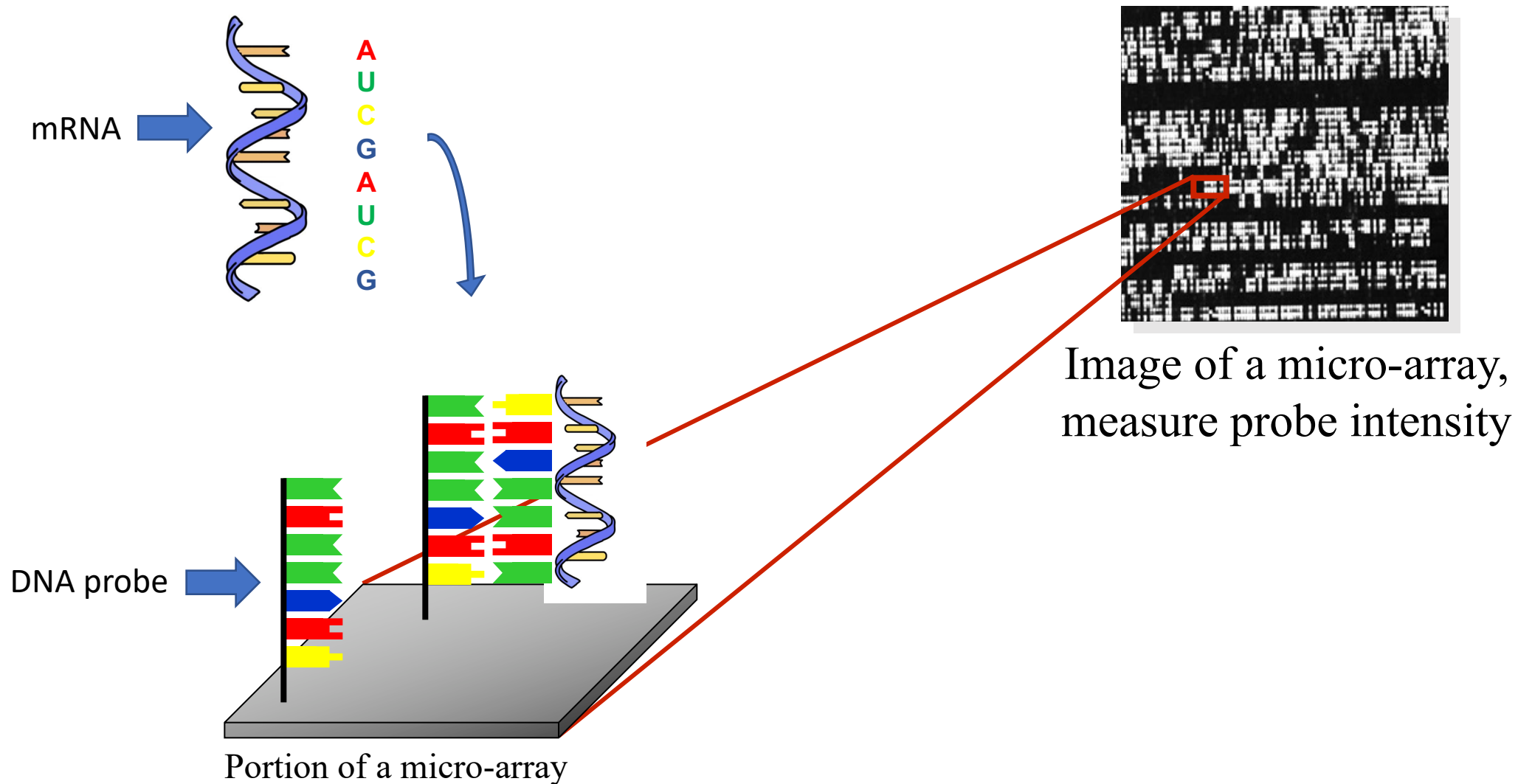
We varied 2 parameters:

1. probe length,
2. sequence uniqueness

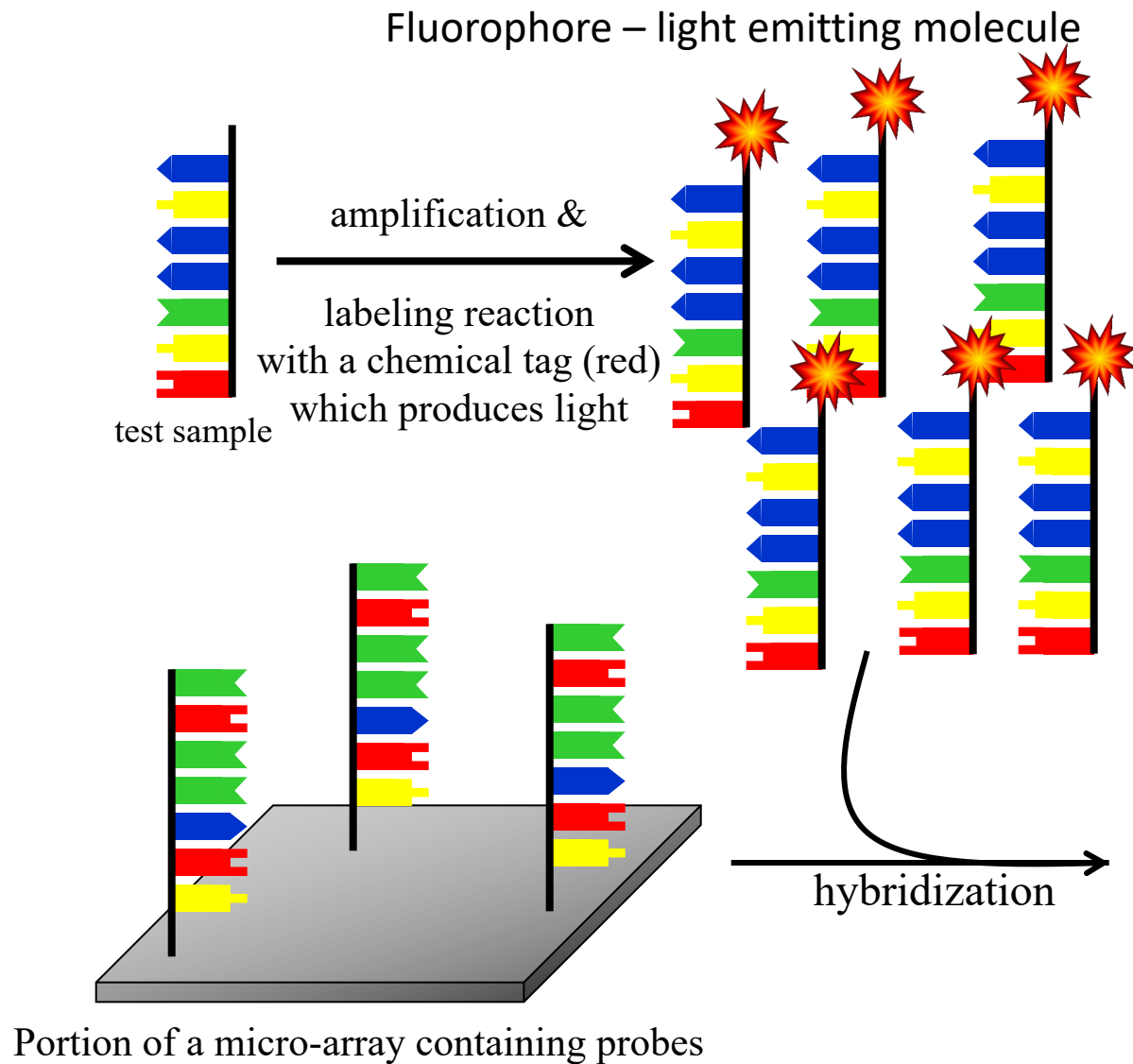


DNA probes produce different amount of *hybridization signal* based on sequence length and uniqueness.

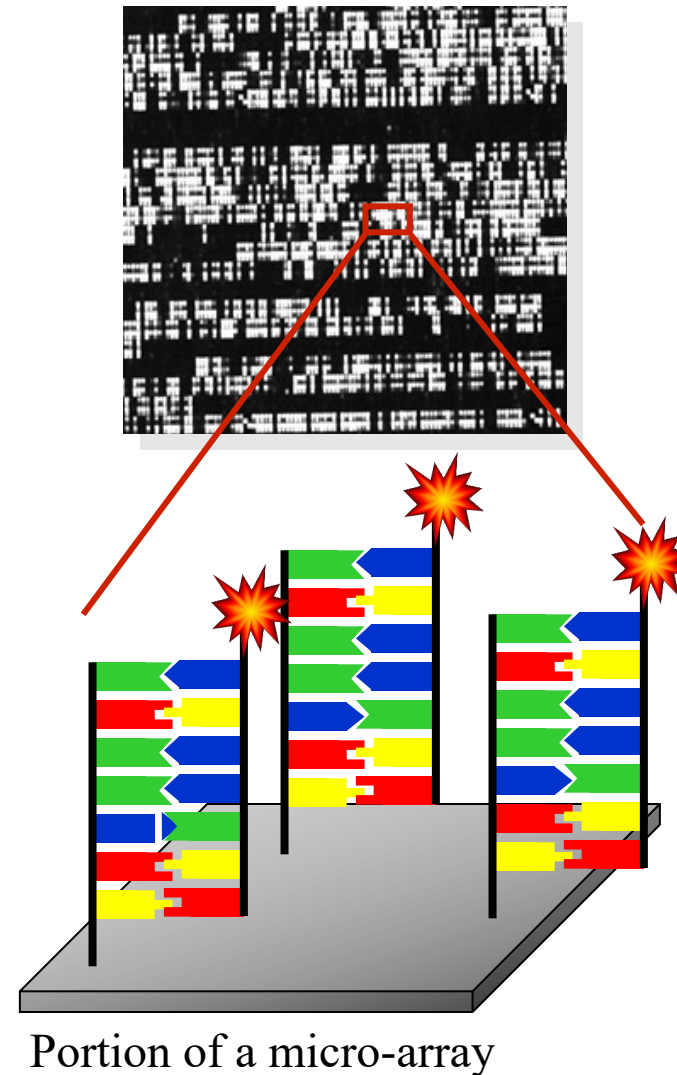
How DNA microarrays work: to measure transcription activity, hybridize arrays to mRNA



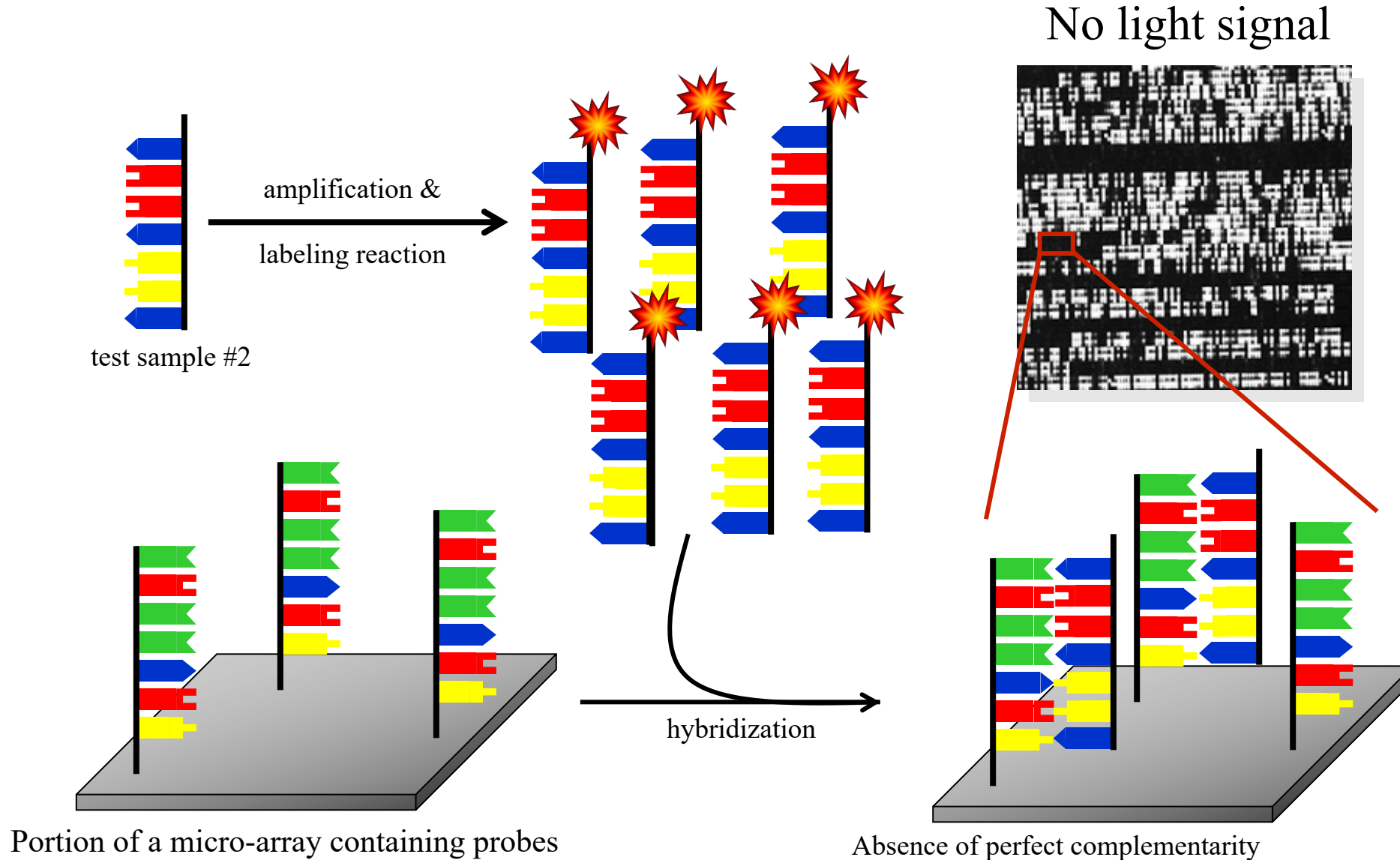
How DNA microarrays work



Positive light signal from labeled DNA



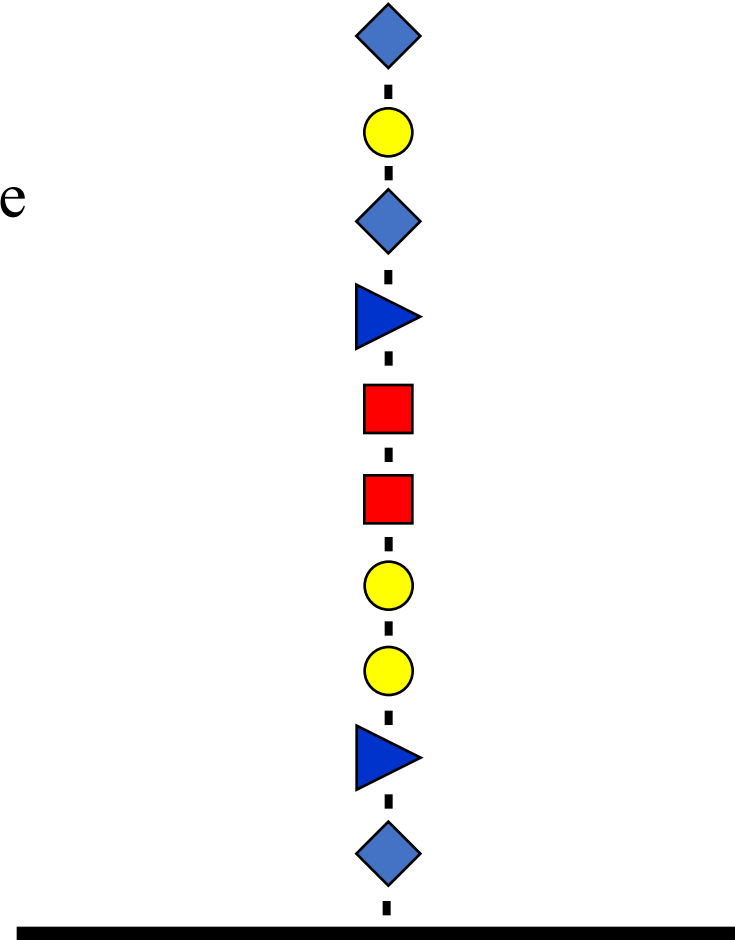
How DNA microarrays work



DNA Synthesis is a Step-Wise Process

Each monomer can be added sequentially

- Thymine
- Guanine
- ▶ Cytosine
- ◆ Adenine

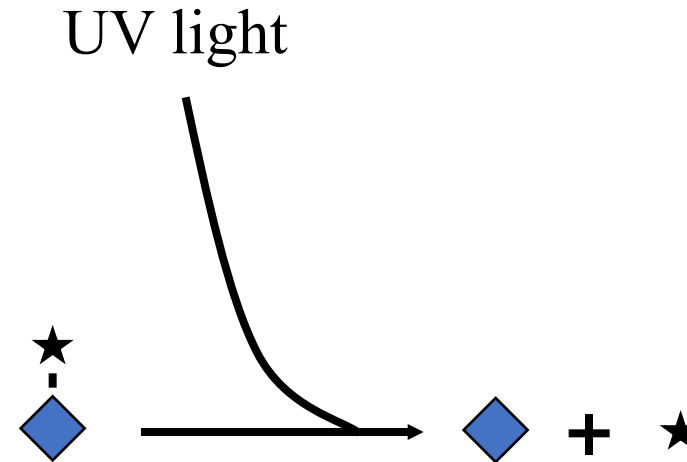


The Process can be Controlled by Light

The four monomers can be modified so that they are nonreactive until exposed to light

★ Photolabile protecting group

◆ DNA monomer

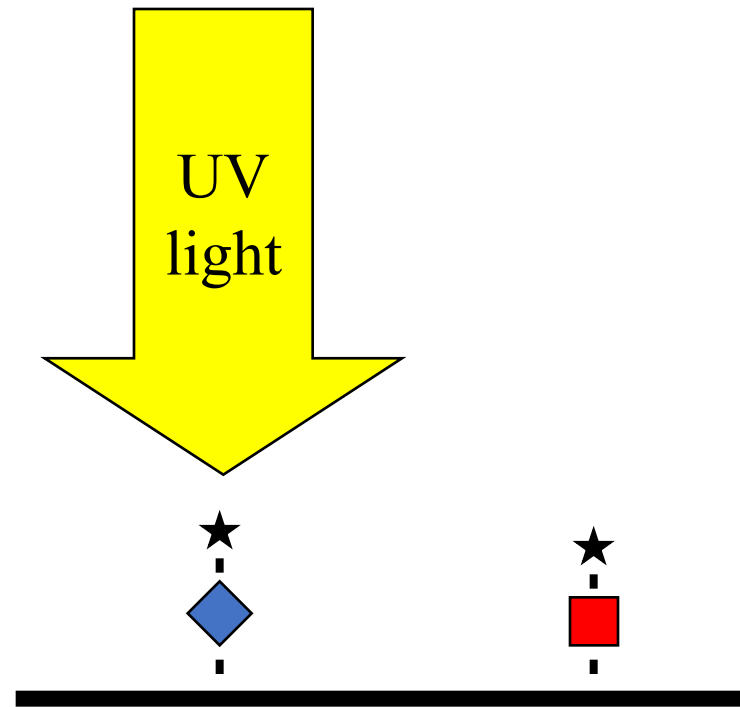


The Process can be Controlled by Light

The step-wise process now requires light to proceed.

★ Photolabile protecting group

◆ DNA monomer

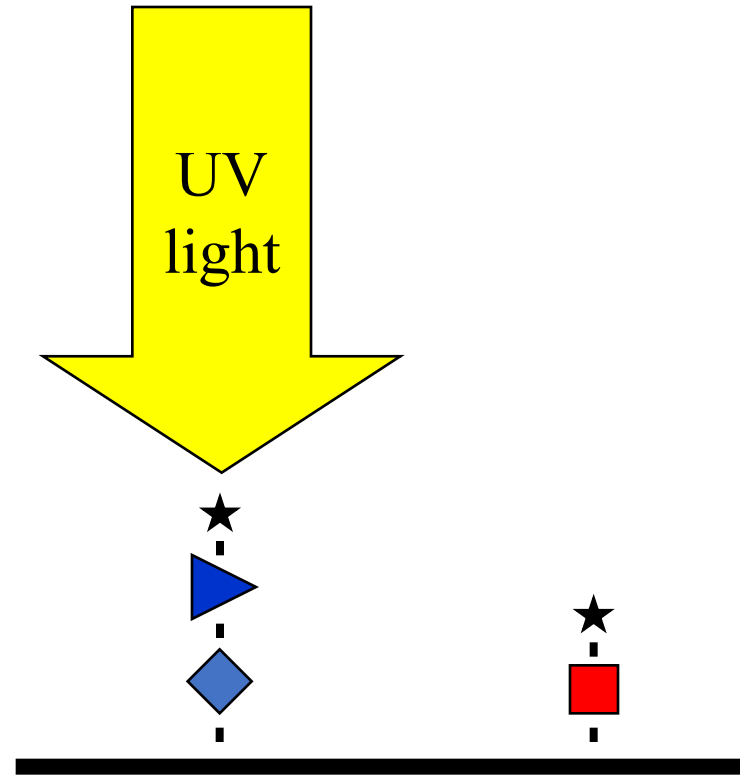


The Process can be Controlled by Light

The step-wise process now requires light to proceed.

★ Photolabile protecting group

◆ DNA monomer

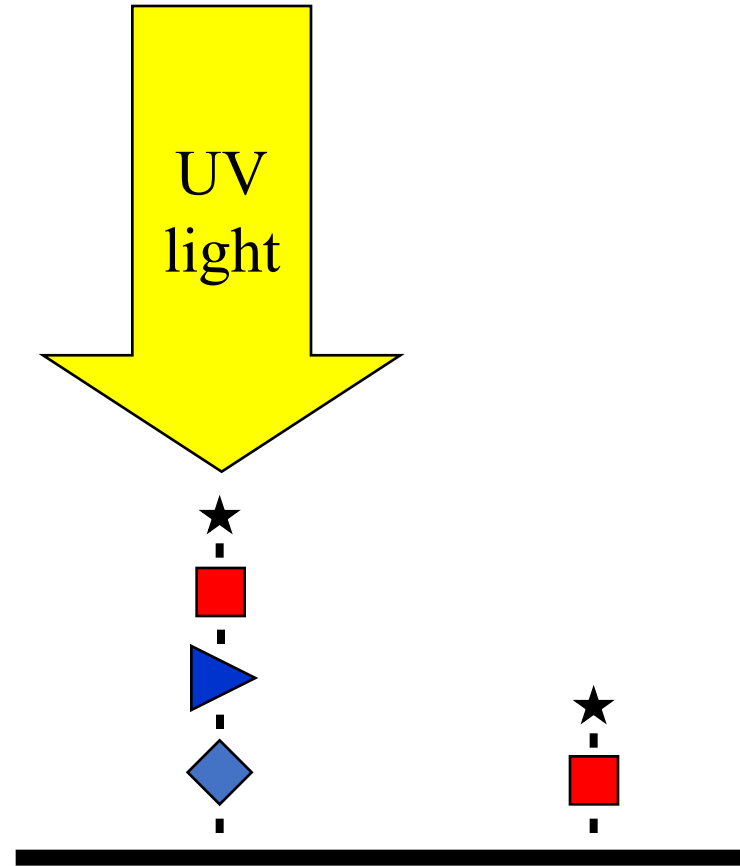


The Process can be Controlled by Light

The step-wise process now requires light to proceed.

★ Photolabile protecting group

◆ DNA monomer

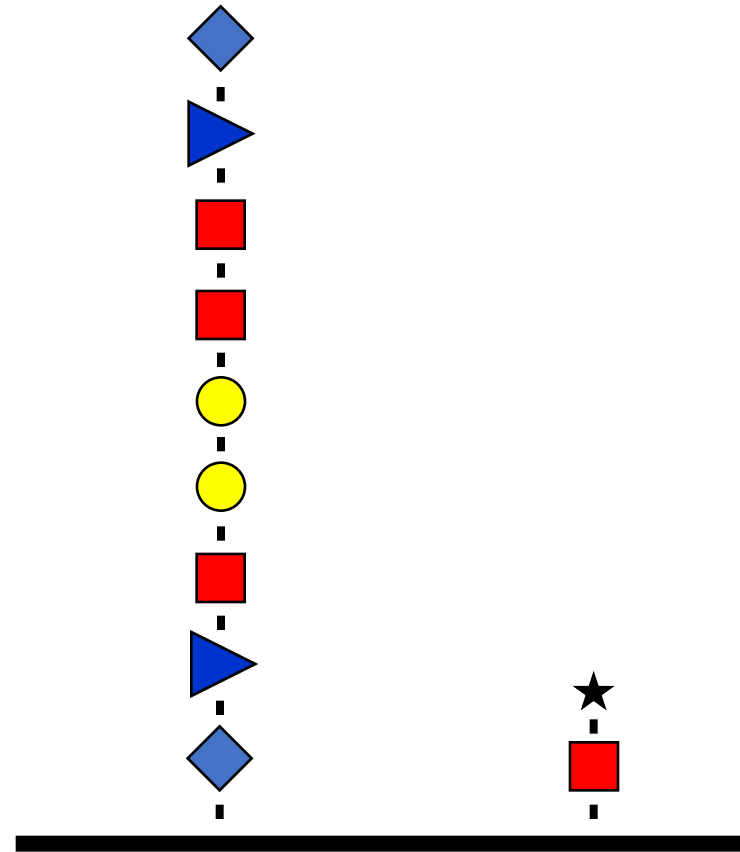


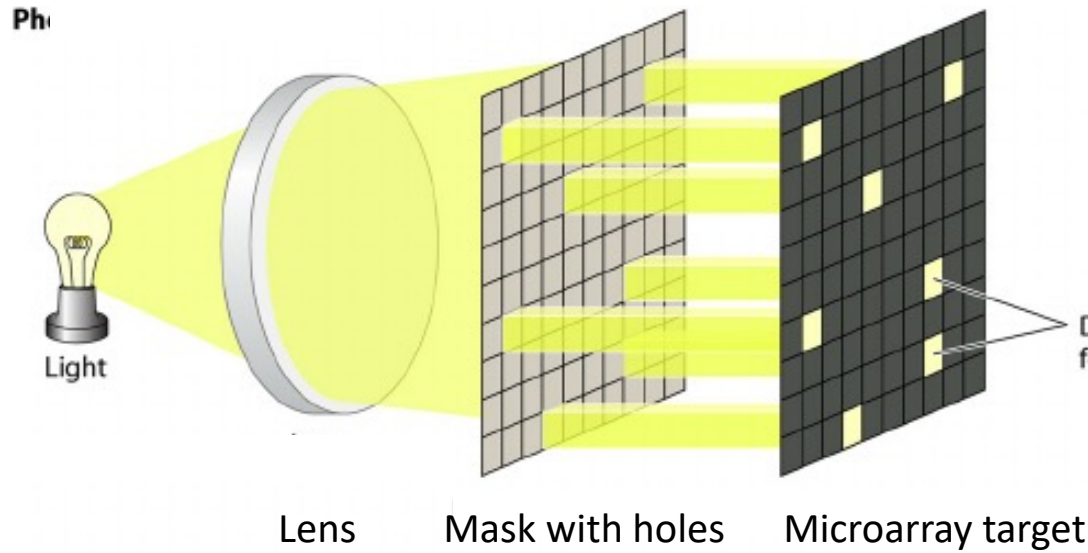
The Process can be Controlled by Light

The step-wise process now requires light to proceed.

★ Photolabile protecting group

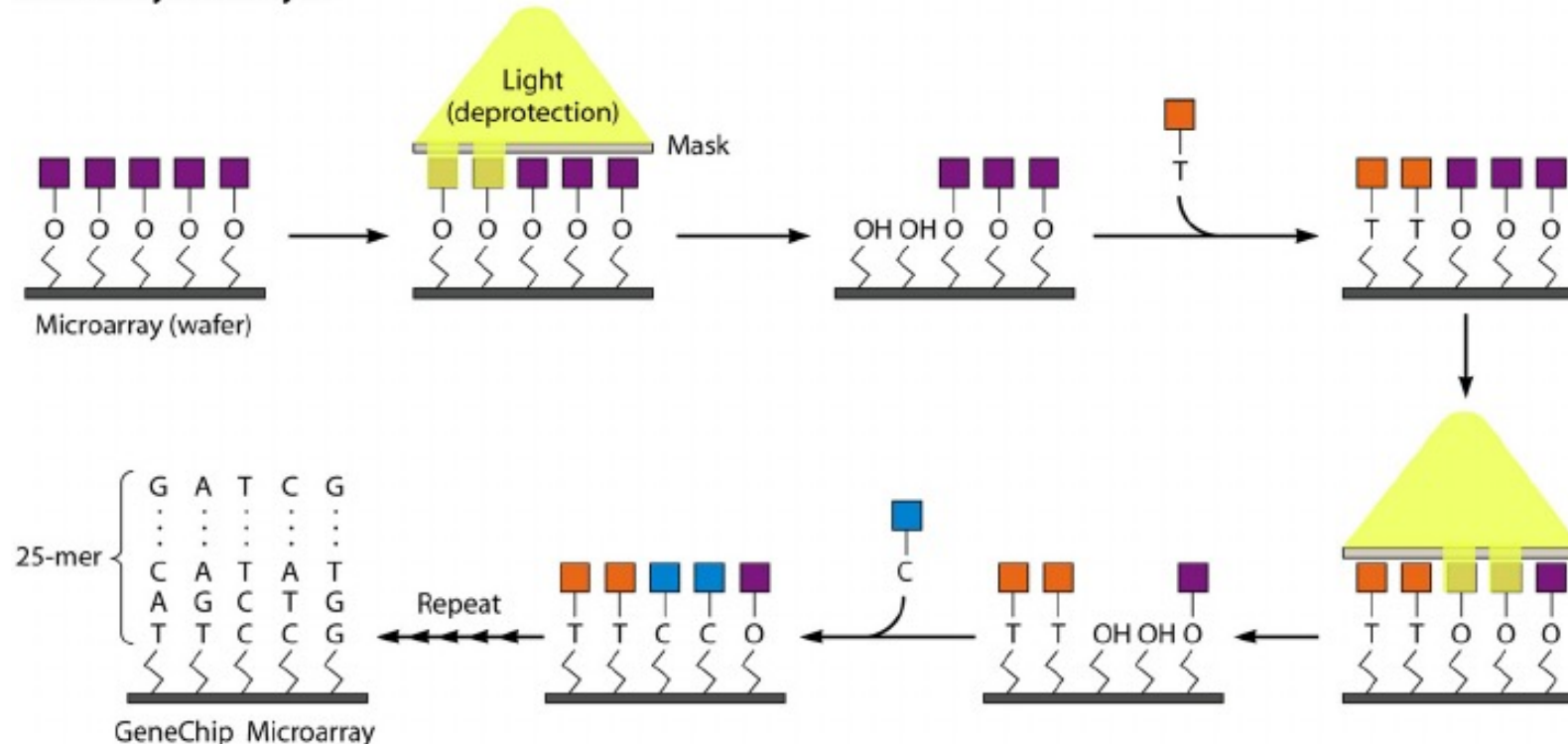
◆ DNA monomer



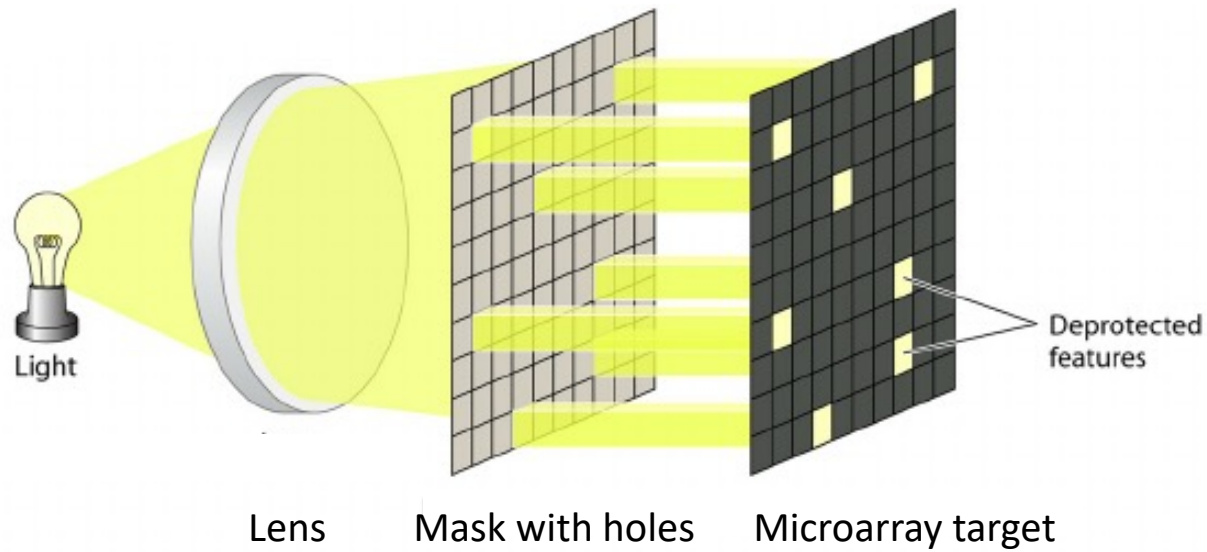


- Light is used to assemble DNA probes that are attached to a glass surface,
- This is a special type of “photolithography”.

Chemical Synthesis Cycle



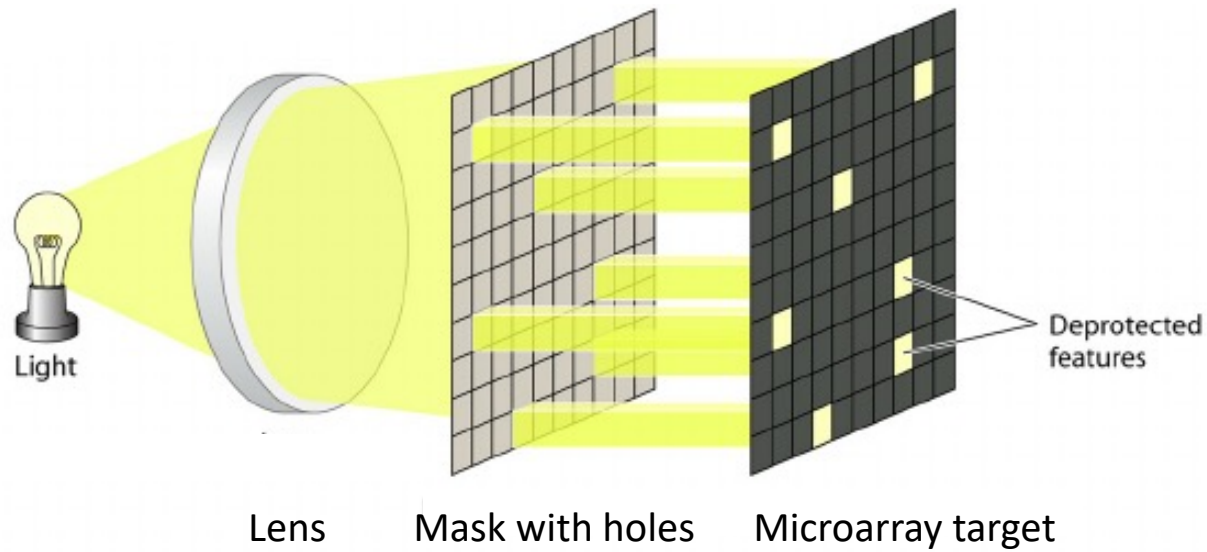
The four DNA monomers can be modified so that they are nonreactive until exposed to light.



Limitations of “Photolithography”:

1. Probe length – low specificity & sensitivity
2. Cost per custom array (>\$500K)
3. Long production time (months)

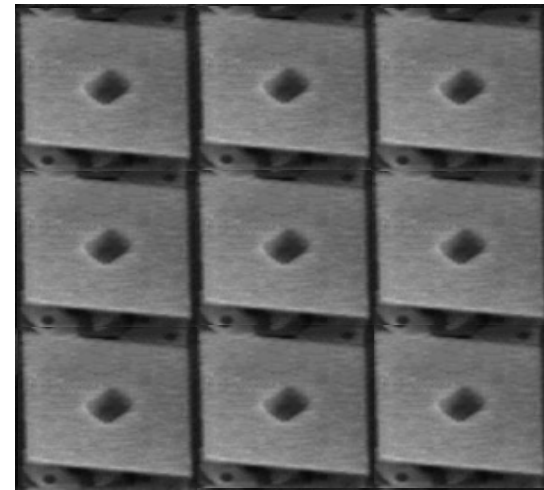
Q. How can we do better?



Limitations of “Photolithography”:

1. Probe length – low specificity & sensitivity
2. Cost per custom array (>\$500K)
3. Long production time (months)

Q. How can we do better?

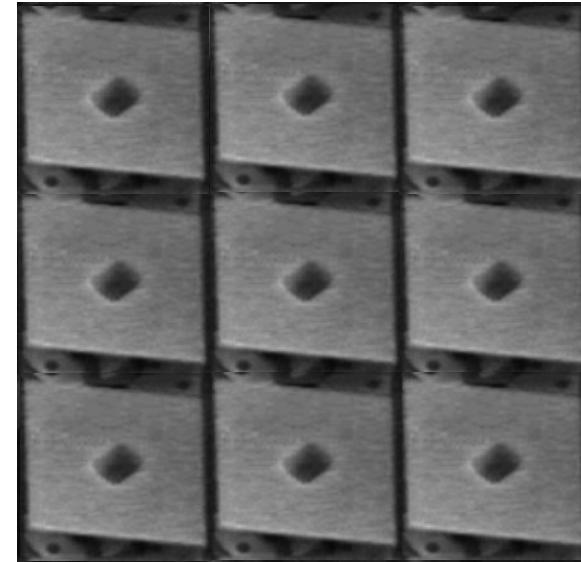


Digital Light Switch

“The Digital Light Switch”

- ◆ Mirrors spacing 17 μm
- ◆ Mirror transit time $<20 \text{ us}$
- ◆ Tilt angle ± 10 degrees
- ◆ Five mirrors = diameter human hair

Close-Up

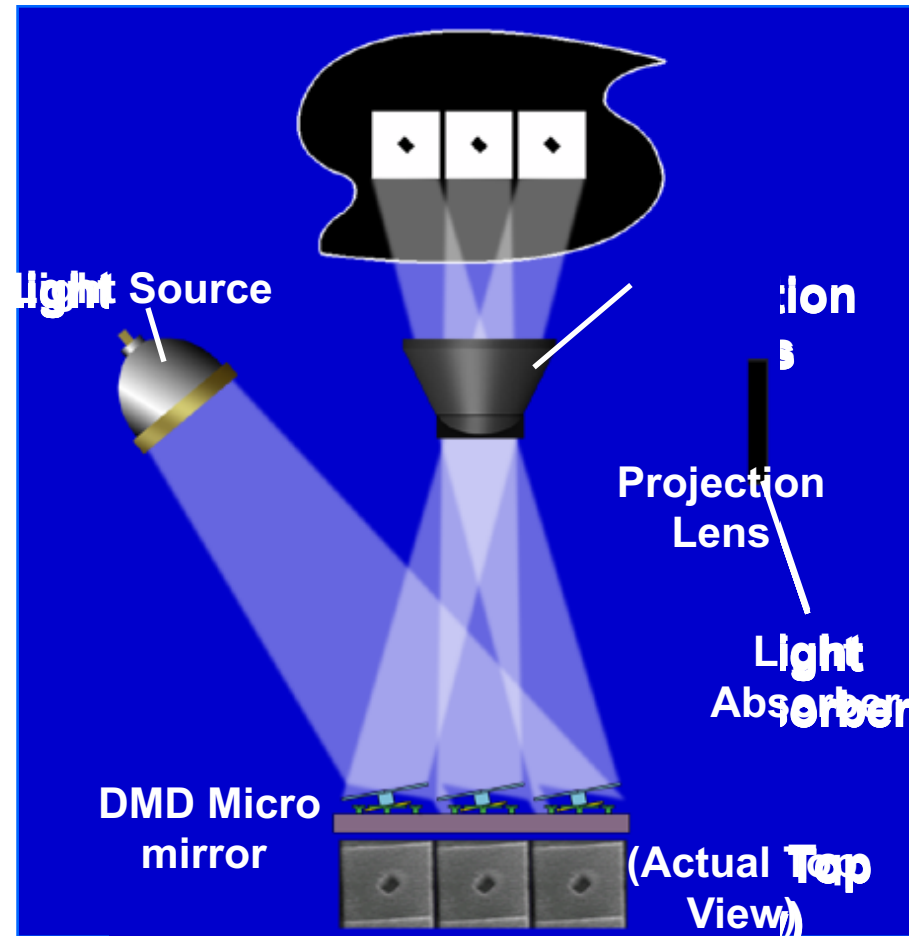


Fantastic way to shine spots of light onto a surface in a very precise and accurate way to do the chemistry needed to synthesize DNA probes.



How Spots of Light are Created

- 780,000 mirrors
- 195,000 usable features
- 4 hours/array

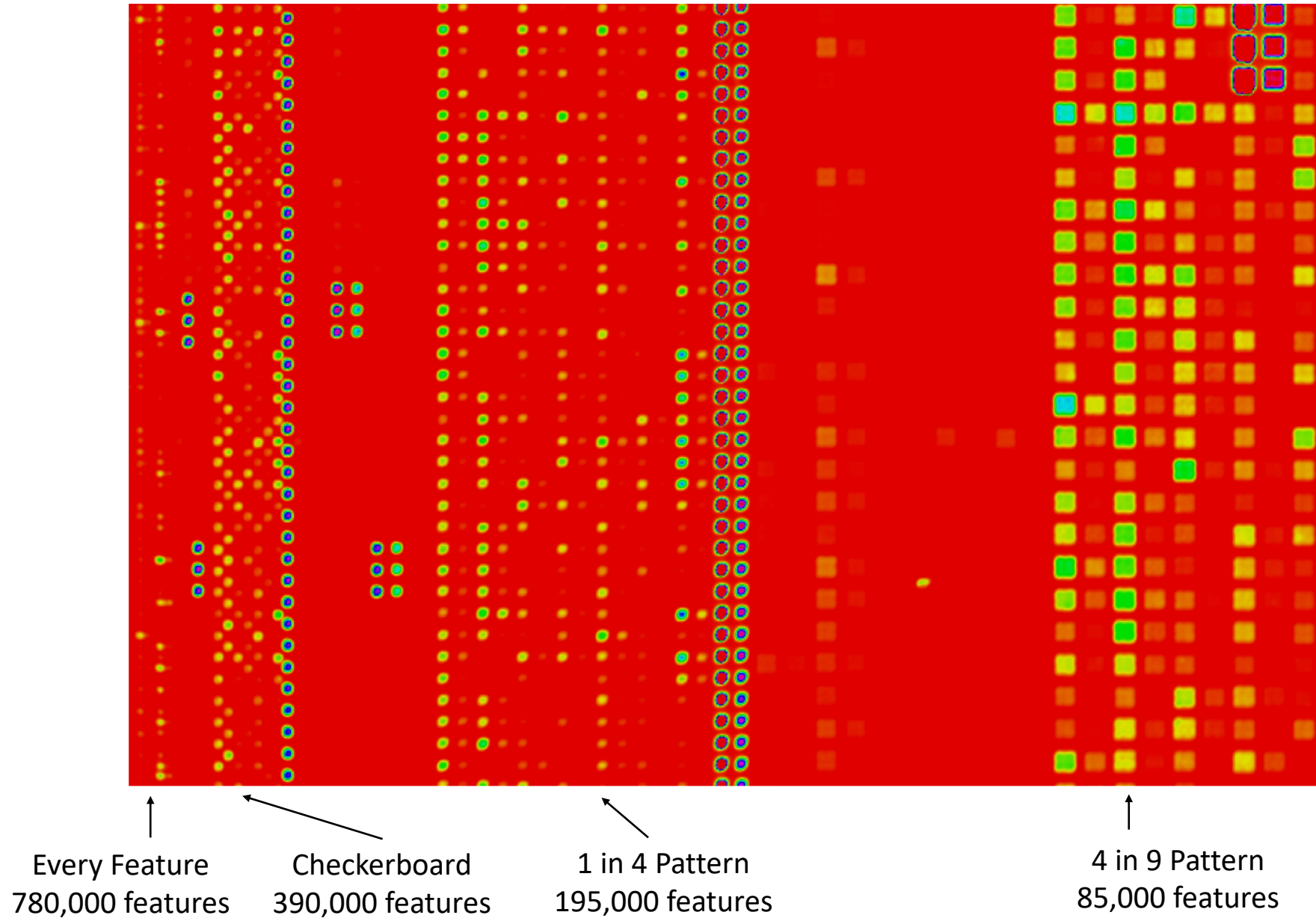


Advantages of Digital Light Processing (DLP):

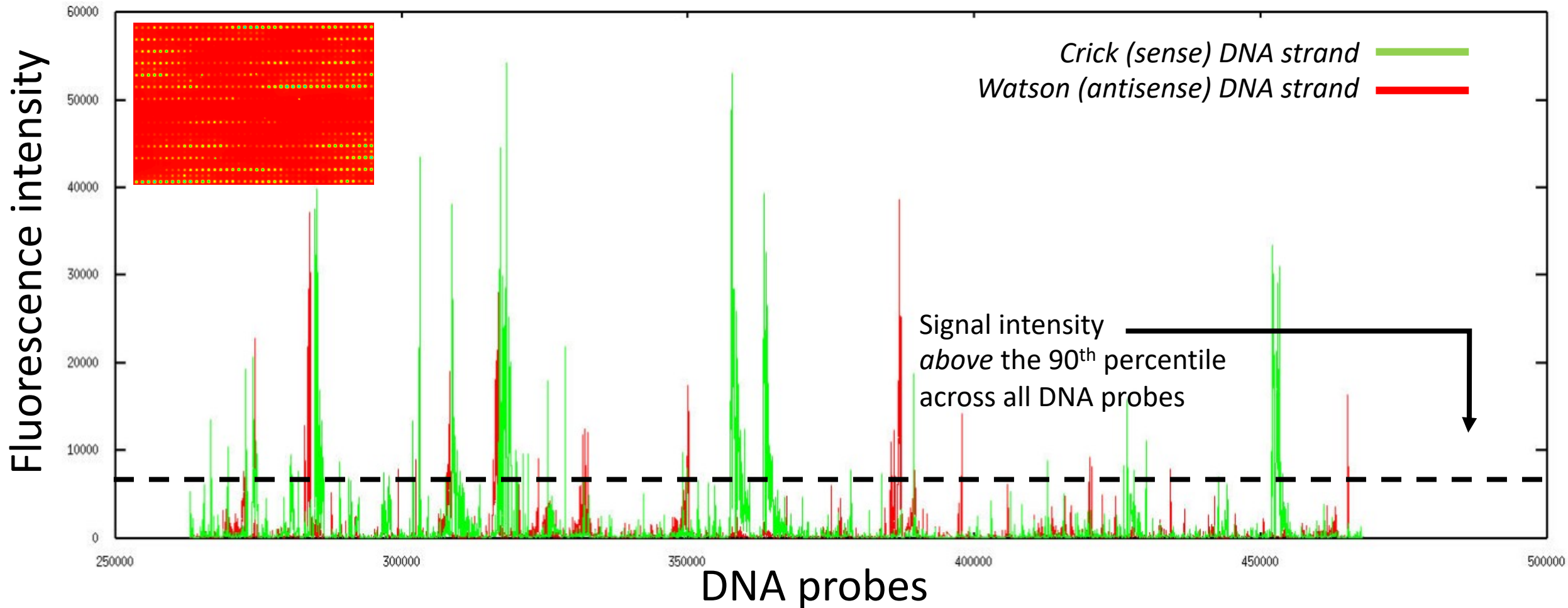
1. Probe length – high specificity & sensitivity
2. Cost per custom array (\$1000)
3. Rapid production time (hours)

This strategy for DNA probe synthesis on glass surfaces, using DLP technology, was key to enabling us to represent the entire genome on just a few glass slides.

Feature Size Options

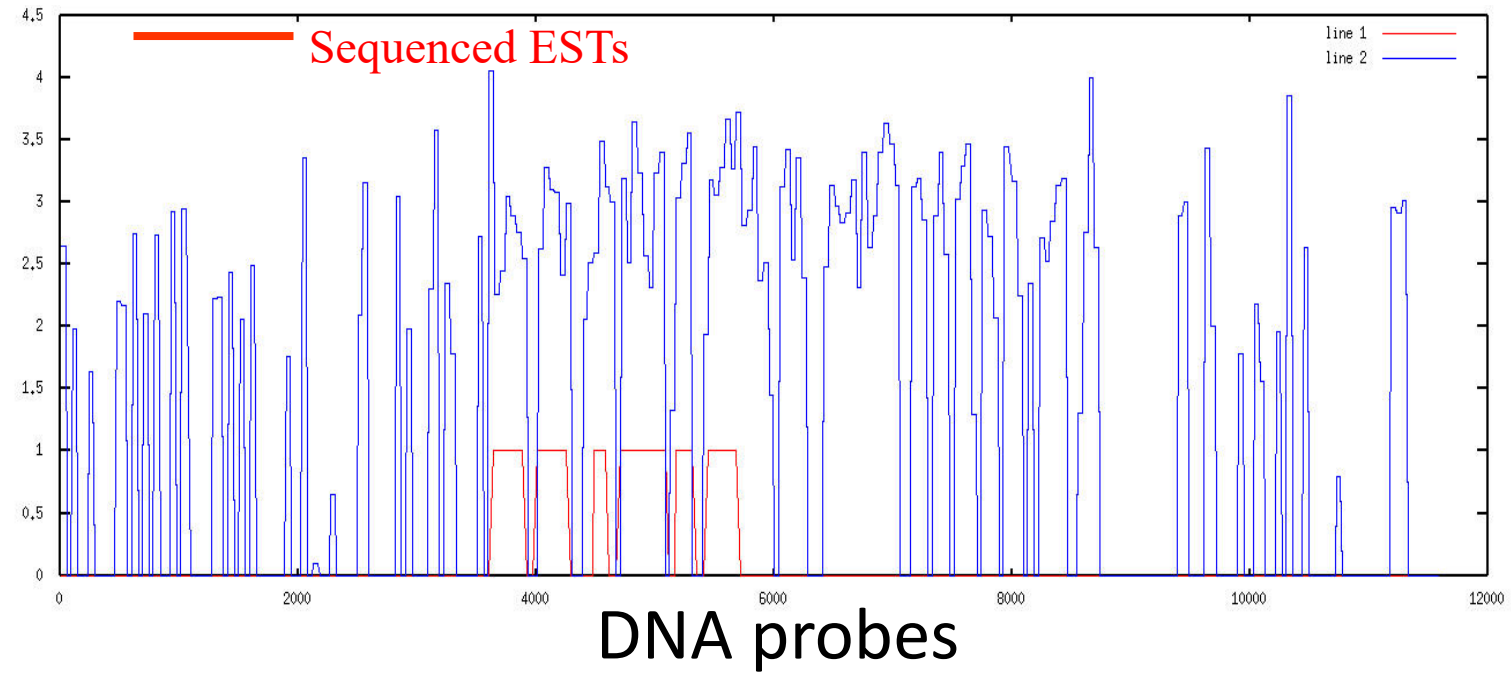


Typical transcriptome raw data



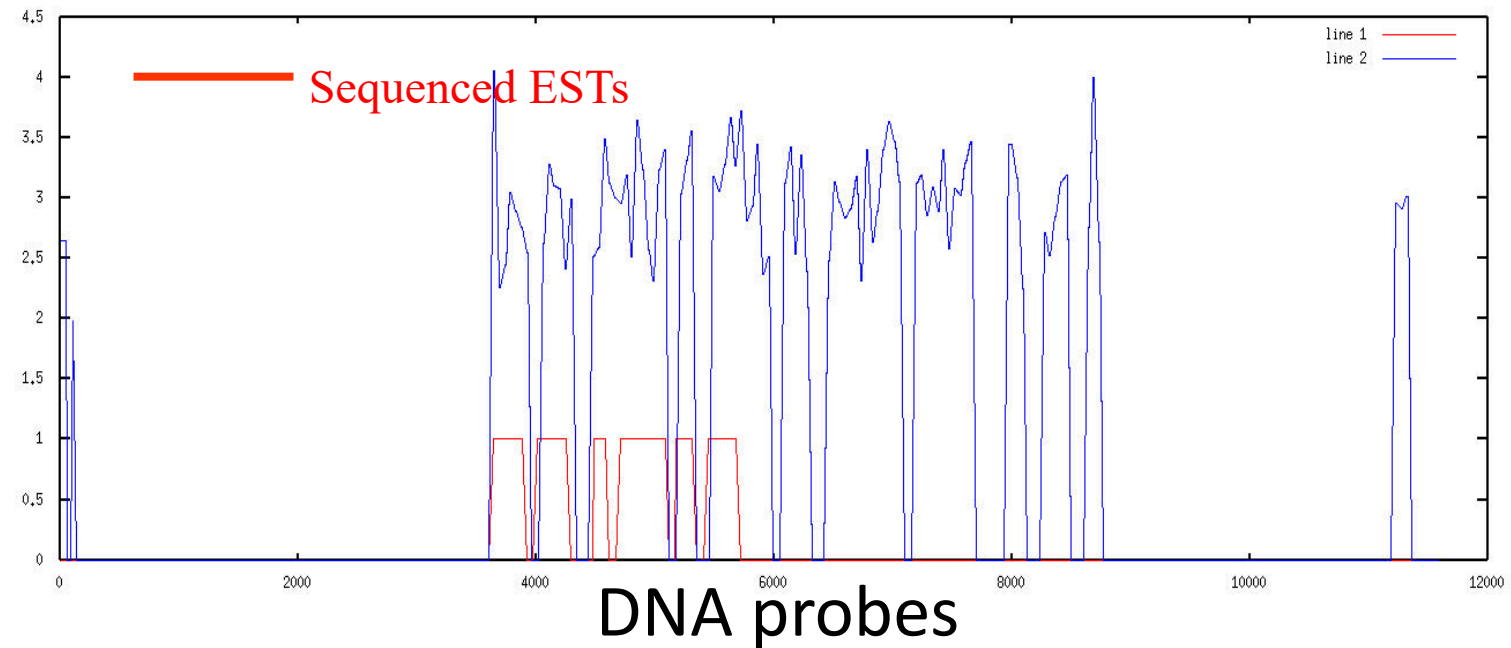
Signal to noise complexity for **Transcriptionally Active Regions (TARs)**,
Both DNA strands are co-transcribed: **anti-sense transcription, even from the same positions**
Transcription was detected in intergenic, overlapping or antisense regions to *annotated genes*

U
N
F
I
L
T
E
R
E
D

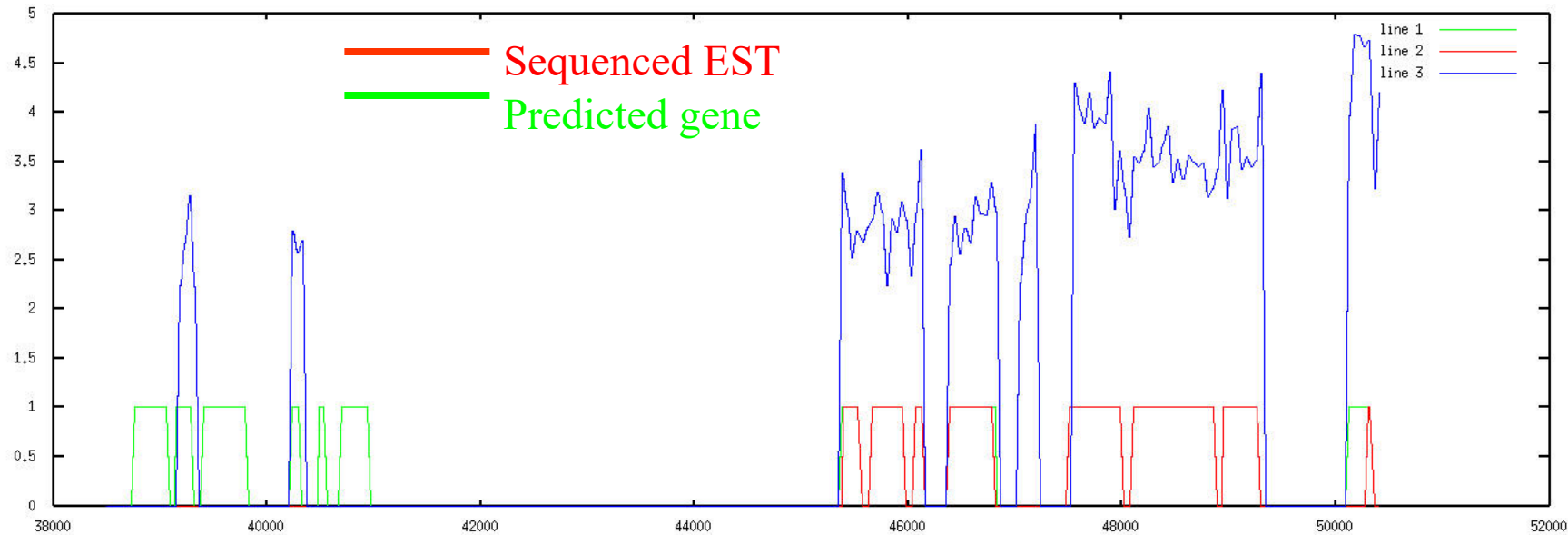


We detected many more transcribed genes than was previously known by sequencing expressed mRNAs, which are called ***Expressed Sequence Tags (ESTs)***.

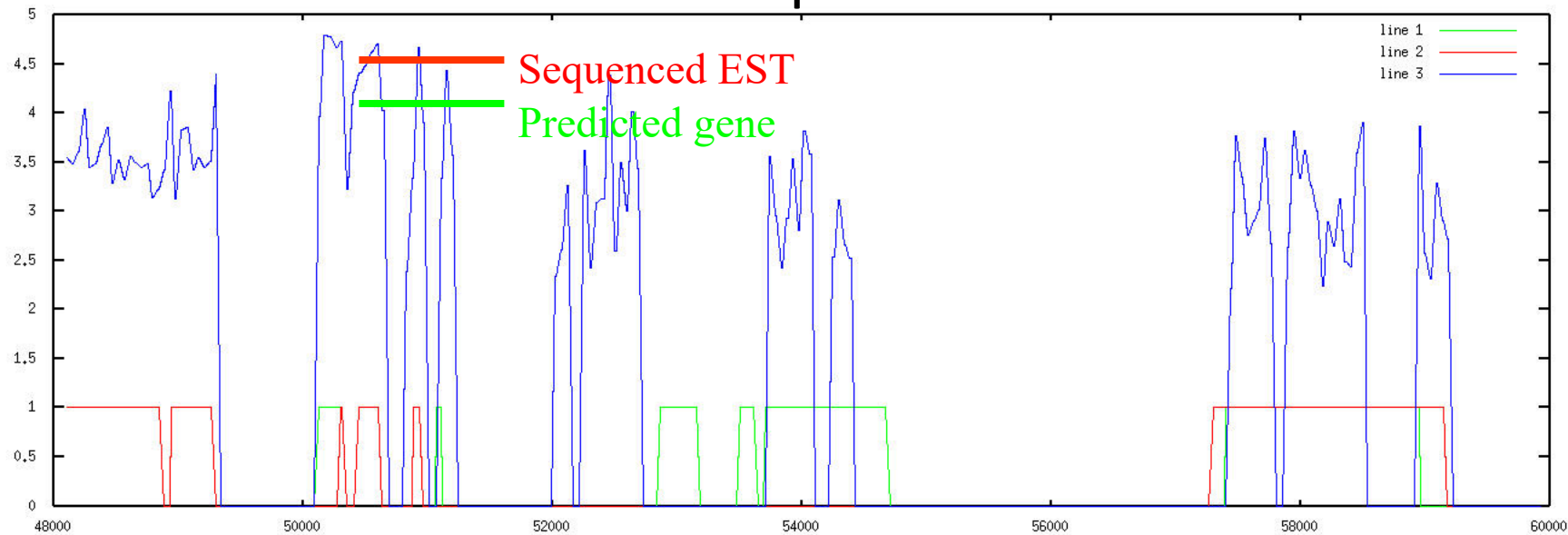
F
I
L
T
E
R
E
D



We came up with an approach to eliminate “apparent transcribed regions” of the genome by declaring that if only a few DNA probes lit up as signal (less than 10), we excluded it from consideration as “transcribed.” This was our “filtering” method.



DNA probes



DNA probes

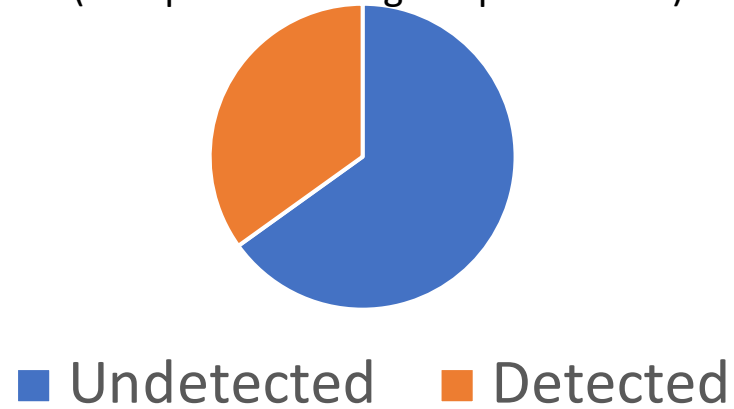
We observed many more transcribed mRNAs than previously known by conventional methods.

In addition, we observed gene expressions that corresponded with some of the *in silico* predicted genes.

This result validated the microarray method and let us know that we were on the right track.

Proportion of genes detected from four annotation sources

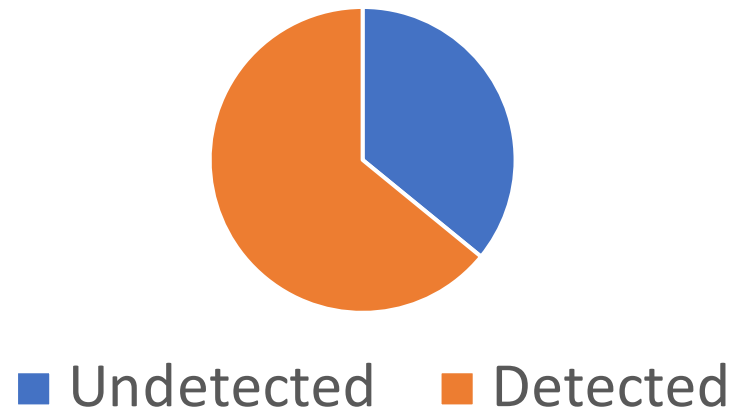
Genscan
(Computer-based gene predictions)



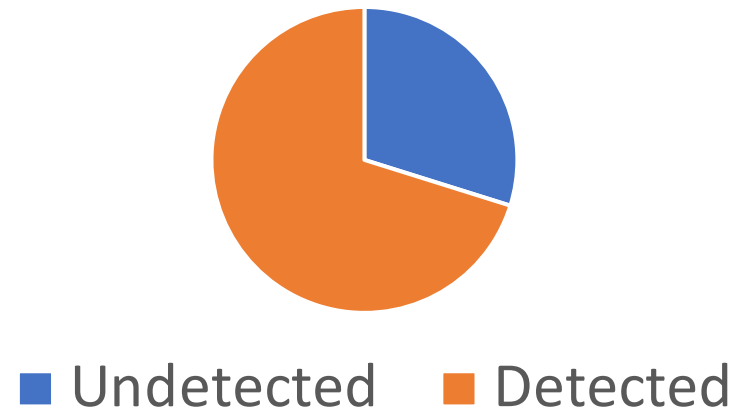
Ensembl



RefSeq



RefSeq:liver



The more orange is the pie chart, the more successful was our gene expression mapping.

In addition to detecting known and predicted genes, a primary goal was to identify novel transcribed regions.

- A total of 13,889 novel transcribed RNAs were identified;
- One third (4,931) corresponded to previously annotated exons, and
- the remaining **8,958 were novel transcribed active regions (TARs),**
from both strands of the genome.

Man on the Moon 1969



Man on a DNA chip 2002



Science 2004, 306:2242-2246

To produce this image, a photograph of an astronaut on the surface of the Moon from 1969 was digitized and converted to a four-level gray-scale image. A distinct 18mer oligonucleotide probe corresponding to perfect match sequence - or one, two, or three mismatches to achieve different shades of gray - was assigned to each of the four gray-scale values.

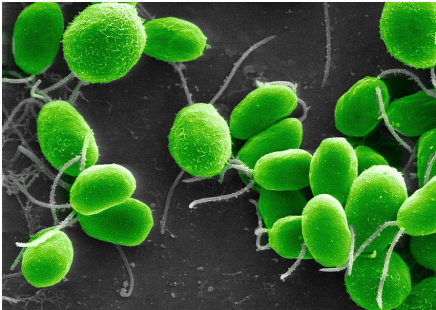
After synthesis, the array was hybridized with a single labeled oligonucleotide, which had the perfect match sequence.

SUMMARY: **DNA microarray technology enabled ...**

- A new paradigm for empirical annotation of sequenced genomes by mapping the RNA transcriptomes.
- ENCODE Project – comprehensive annotation of human genome and model organisms – this is still ongoing work (in the US and internationally).
- Identification of thousands of previously unknown protein-coding and non-protein coding genes,
- Knowledge of genome-wide *anti-sense* transcription,
- Identification of many disease-related genes,
- Diagnostic and therapeutic targets for many disease,
- Elucidation of novel functions of non-coding RNAs,
- Development of gene expression biomarkers to predict patient outcomes.

Transcriptomes of Model Organisms Mapped at ARC:

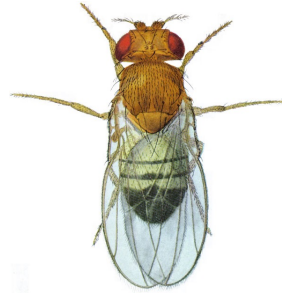
Green Algae, *Chlamydomonas reinhardtii*
PNAS 102: 3703-3707.



Sea Urchin, *Strongylocentrotus purpuratus*
Science 314, 960-962.



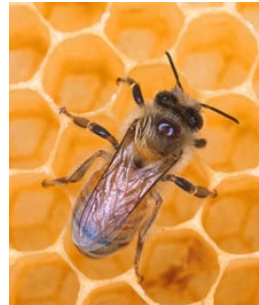
Fruit fly, *Drosophila melanogaster*
Science 306: 655-660.



Plant, *Arabidopsis thaliana*
PNAS 102: 4453-4458.



Honeybee, *Apis mellifera*.
Nature 443, 931-949



Yeast, *Sacharomyce cerevisiae*,
PNAS 101:11046-11051.
PNAS 103: 4192-4197.



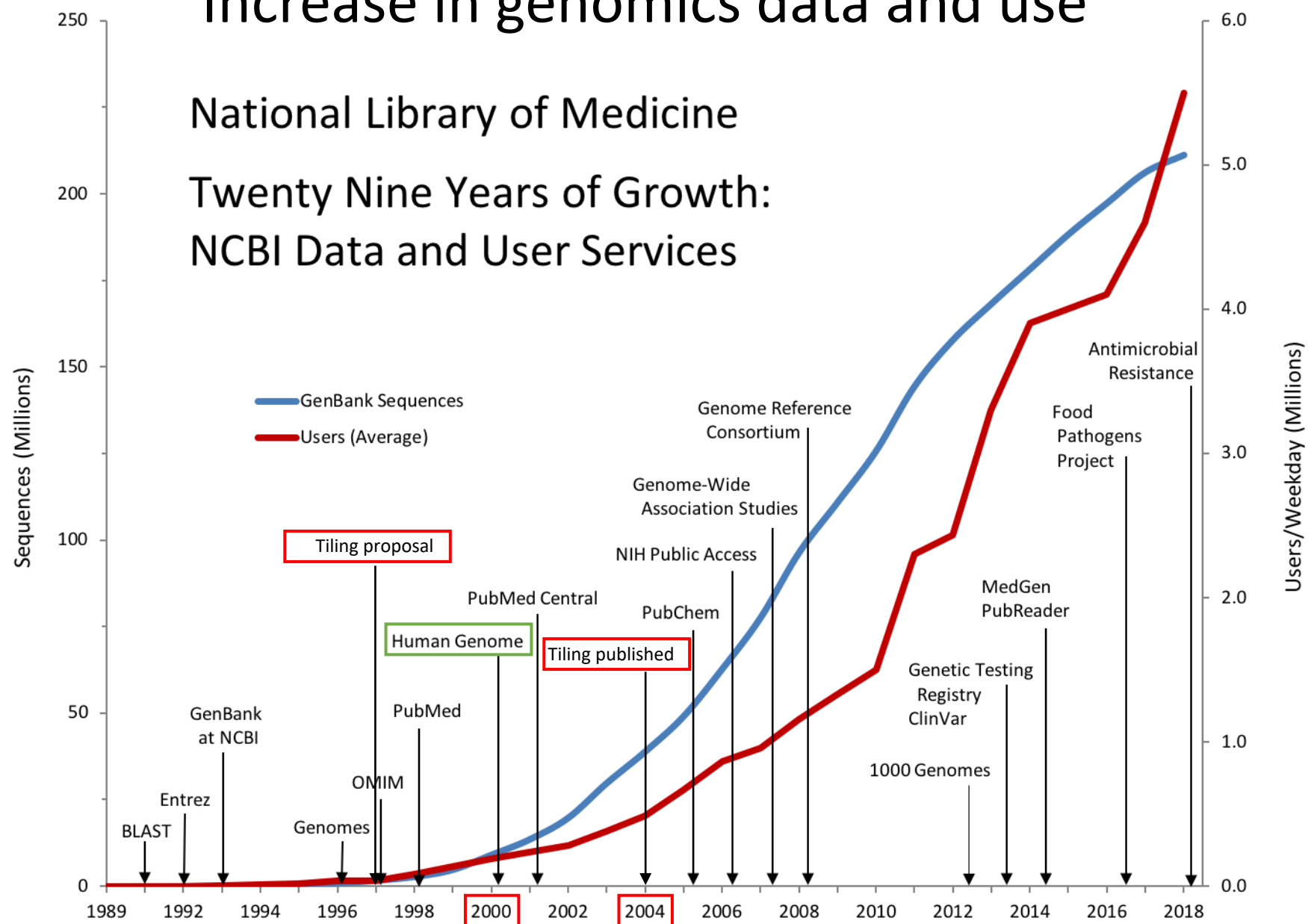
Rice, *Oryza sativa*
Nature Genetics 38: 124-129.



Increase in genomics data and use

National Library of Medicine

Twenty Nine Years of Growth:
NCBI Data and User Services



Congressional Justification FY 2020



Acknowledgements



Global identification of human transcribed sequences with genome tiling arrays.

***Science* (2004) 306:2242-2246.**

[PAUL BERTONE](#), [THOMAS E. ROYCE](#),
[JOEL S. ROZOWSKY](#), [ALEXANDER](#)
[E. URBAN](#), [XIAOWEI ZHU](#), [JOHN L. RINN](#)



Mel Averter, PhD, JD

Waraporn Tongprasit, PhD (postdoc)

Manoj Samanta, PhD

Sherman Weissman, MD

Mark Gerstein, PhD

Michael Snyder, PhD

Sidney Altman, PhD



Yale University

Dedication



Sidney Altman in 1989, after learning that he had won the Nobel Prize in Chemistry.



Meyya Meyyappan,
Chief Scientist for
Exploration
Technology
at NASA Ames



Harry Partridge,
The Ames Center
Chief Technologist



Maurice M. "Mel" Averner,
NASA Space Biology program
manager -- a visionary who
recognized and appreciated
vision in others, who introduced
the concept of terraforming
Mars.