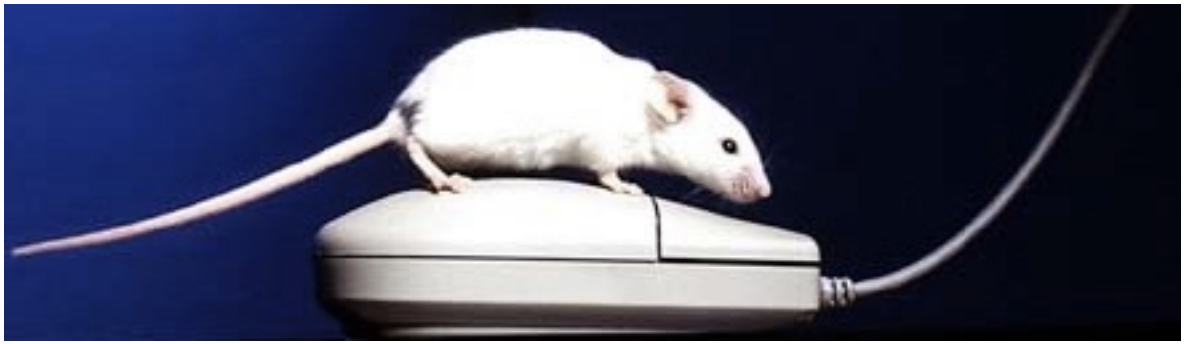
BIOINFORMATICS

Nadia Pisanti,
University of Pisa, Italy



OUTLINE OF THE TALK

- A BIT OF HISTORY:
  - Why did computer science meet biology?
  - Human Genome Project: lots of biological data turns in silico

- OPPORTUNITIES FOR BIOLOGY:
  - Analysing and Comparing Genomes
  - Understanding Genomic Regulation mechanisms and much more..

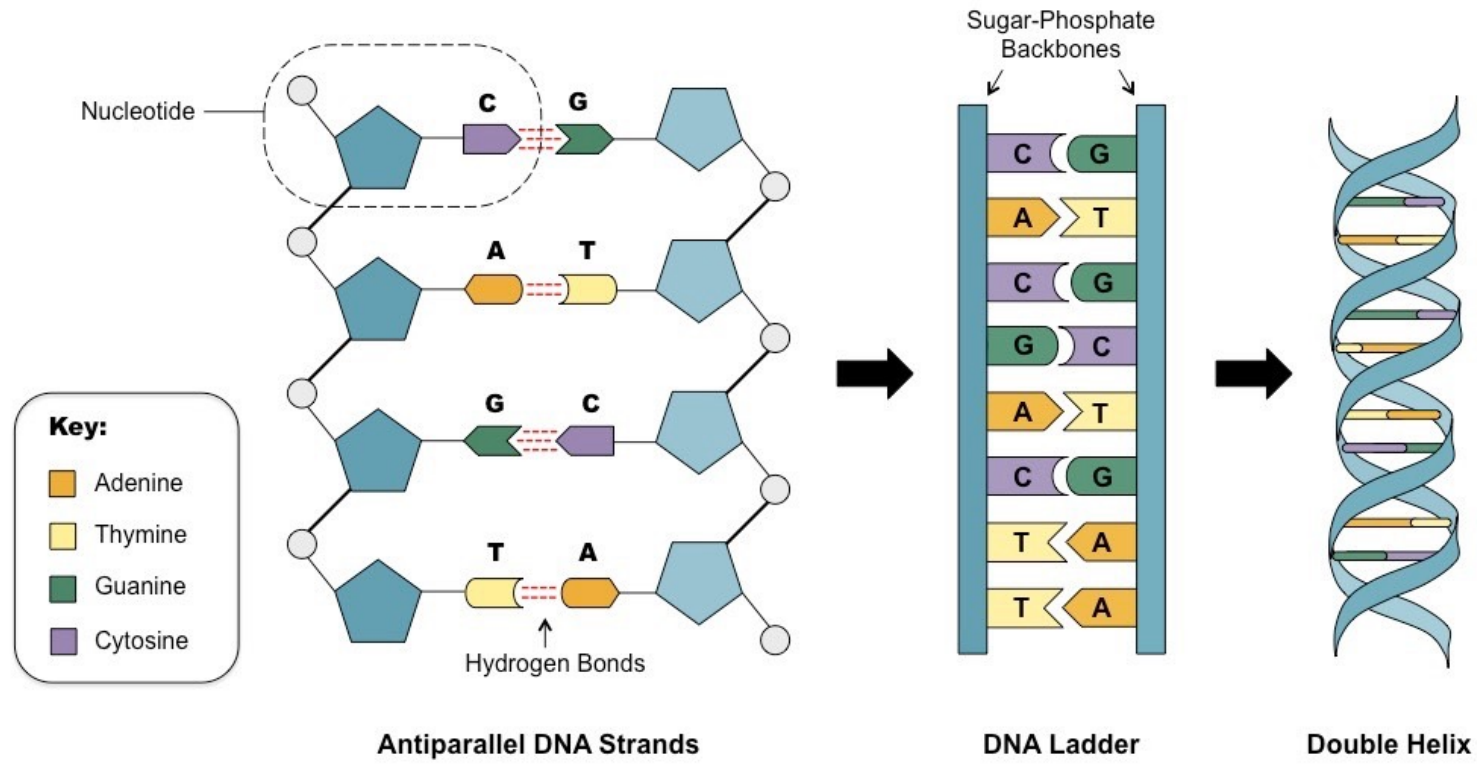
- COMPUTATIONAL CHALLENGES:
  - Assembly
  - Analysis: Efficient Algorithms and Tools
  - Storage and Browsing of Genomic Data
  - A basic algorithmic tool: alignments

Biology used to be a descriptive-only science...

Nadia Pisanti

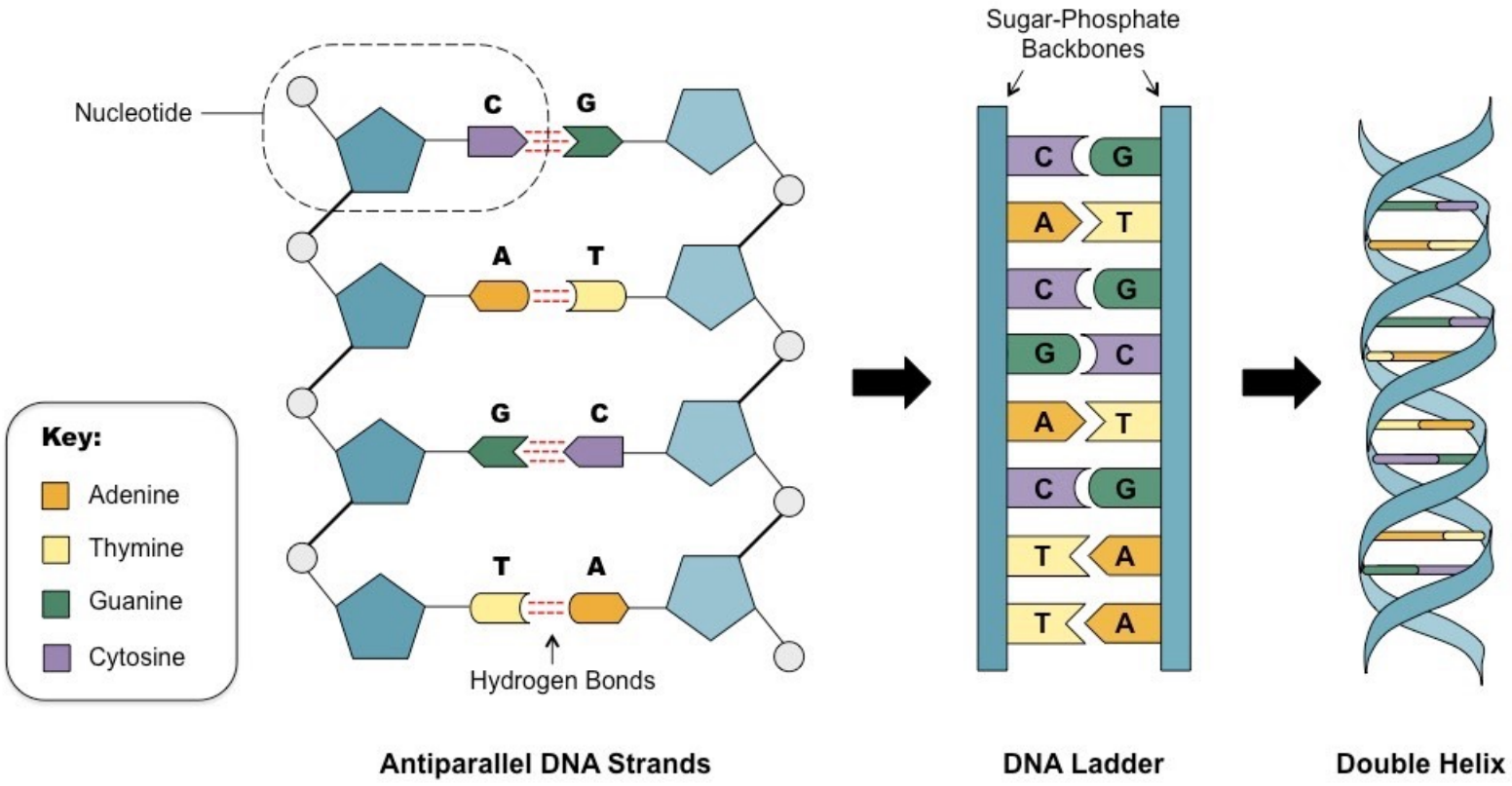
Crick, Franklin, Watson: 1953

DNA as a text



Nadia Pisanti

DNA as a text



A single sequence of letters {A,C,G,T} describes a DNA fragment

Nadia Pisanti

DNA as a text

Nadia Pisanti

From double helix to sequencing

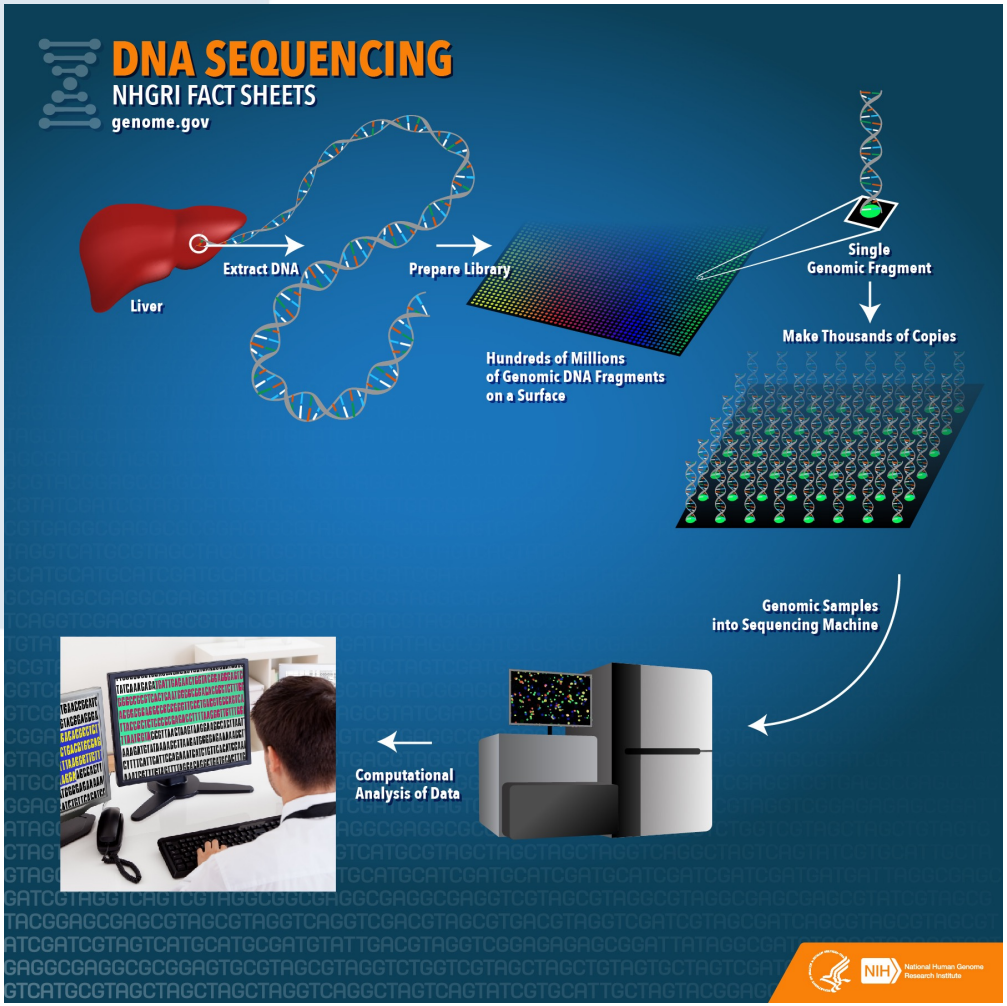
1953: F.Crick, R.Franklin, and J.Watson discover the double helix structure of DNA. [Nobel Prize 1962]

70's: The first sequencing techniques are developed (F.Sanger). [Nobel Prize 1980]

1990: The Human Genome Project begins. The goal is to identify the sequences of all the genes of the human genome (expected to be >100,000).

Still one of the biggest research projects of modern science.

DNA sequencing



Data turns *in silico*,

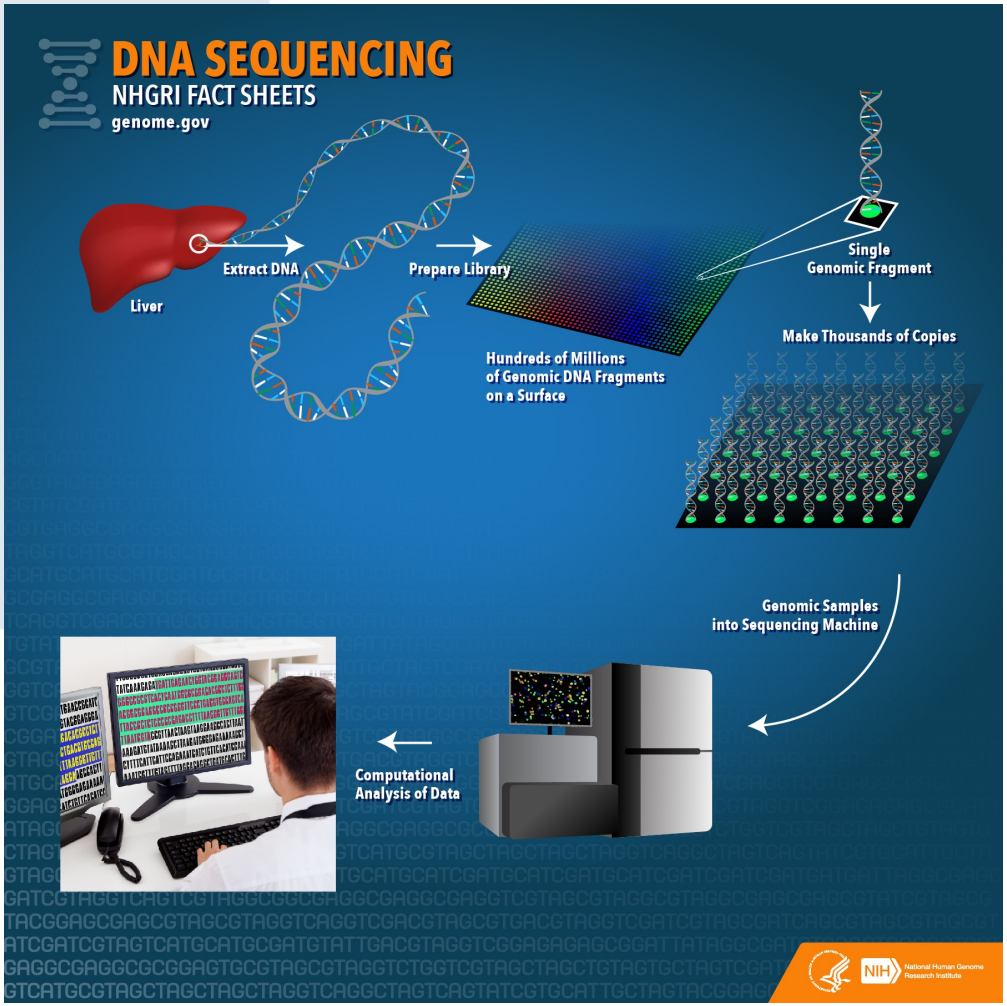
and so do:

- comparison
- classification
- analyses
- ....
- ....

... lots of biology!

Nadia Pisanti

DNA sequencing



Data turns *in silico*,

and so do:

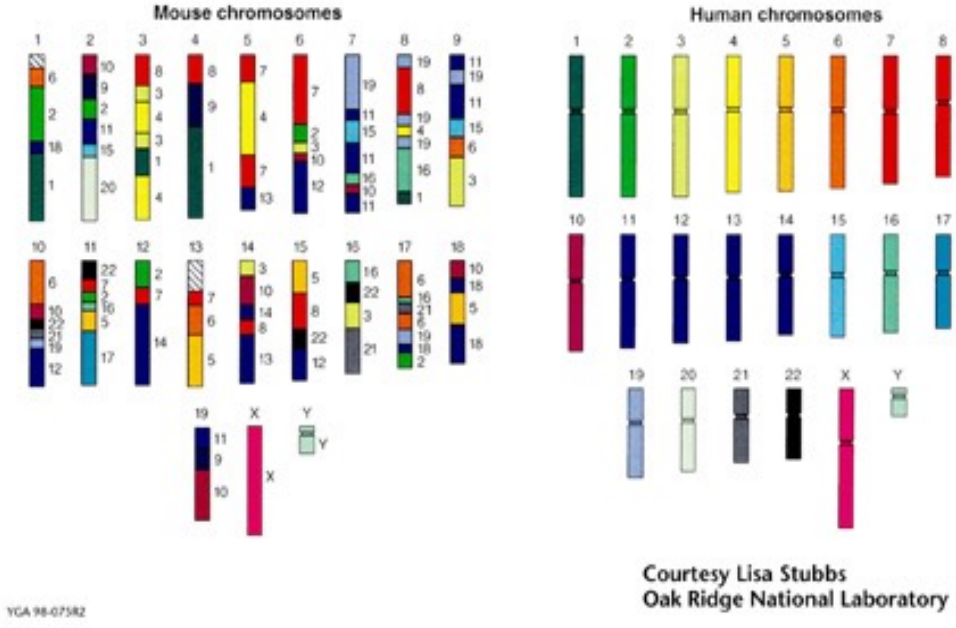
- comparison
- classification
- analyses
- ....
- ....

... lots of biology!

COMPUTATIONAL BIOLOGY

Nadia Pisanti

Mouse and Human Genetic Similarities



Nadia Pisanti

The Human Genome Project

Started in 1990.

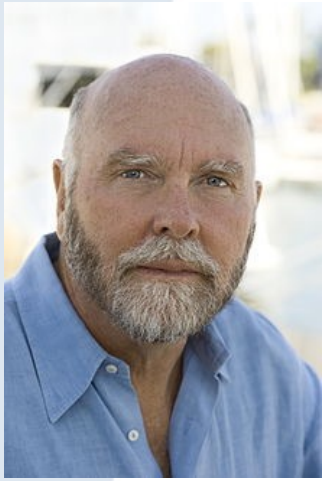
Expected ending time: 2003.

Actual ending time: 2000-2003.

Expected result: Locate and sequence and understand the function of the at least 100,000 human genes (the remaining is just *junk DNA*).

Actual result: "only" 20,000-25,000 genes were found, and "junk DNA" plays a fundamental role in gene regulation.

The Human Genome Project



1998: Craig Venter announces the creation of his company Celera Genomics, and poses a challenge to the public consortium...

1999: Celera completes the sequencing of Drosophila, exhibiting a new techniques to sequence a complex genome.

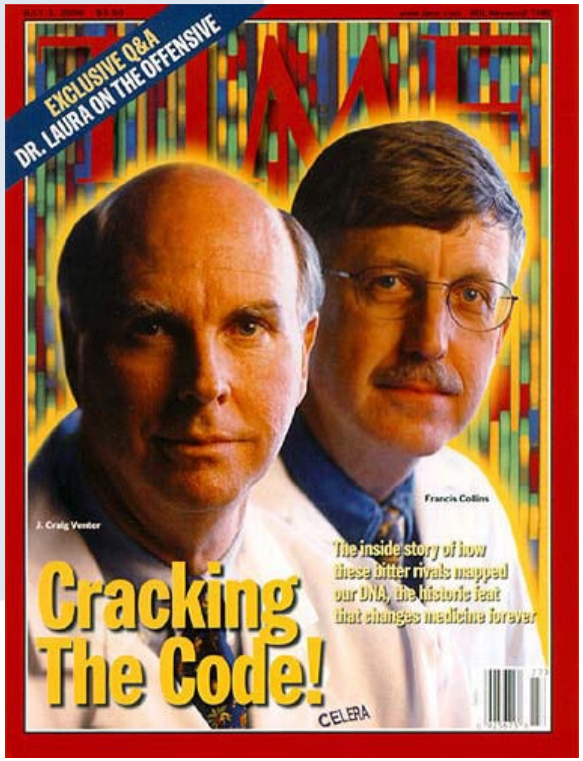
[during 1999, Celera's stock auctions grew of a 500% factor in 4 months....]

The speed of Celera, together with ethical issues*
caused a general (public) panic... and a boost!

*Celera declared its intention to patent human genomic data

Nadia Pisanti

Il progetto genoma umano



On June 26th, 2000

in a press conference at the White House,

The US president Bill Clinton and UK prime minister T.Blair (in videoconference), together with both C.Venter and F.Collins announced the

first (draft of the) human genome sequence!

13 years later:

"*If we want to make the best products, we also have to invest in the best ideas. Every dollar we invested to map the human genome returned $140 to the economy—every dollar.*"

President Barack Obama, 2013 State of the Union address.

Nadia Pisanti

Algorithmic Challenges in HGP

On the sequencing of the human genome

Robert H. Waterston*†, Eric S. Lander‡, and John E. Sulston§

*Genome Sequencing Center, Washington University, Saint Louis, MO 63108; ‡Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

On the sequencing and assembly of the human genome

Eugene W. Myers*, Granger G. Sutton, Hamilton O. Smith, Mark D. Adams, and J. Craig Venter

Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850

Algorithmic Challenges in HGP

On the sequencing of the human genome

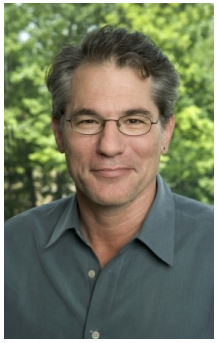
Robert H. Waterston*†, Eric S. Lander‡, and John E. Sulston§

*Genome Sequencing Center, Washington University, Saint Louis, MO 63108; ‡Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02142; and §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

On the sequencing and assembly of the human genome

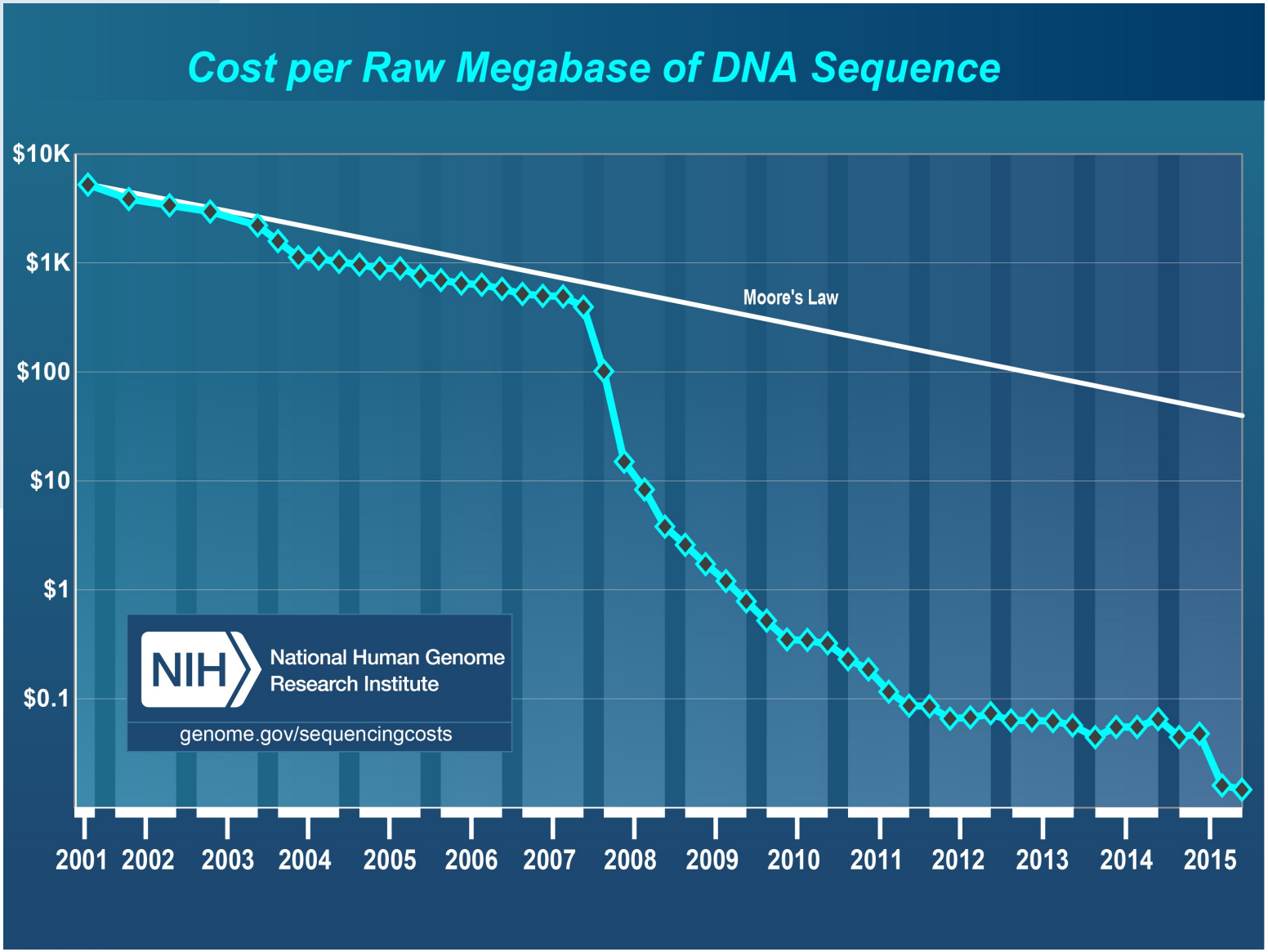
Eugene W. Myers*, Granger G. Sutton, Hamilton O. Smith, Mark D. Adams, and J. Craig Venter

Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850



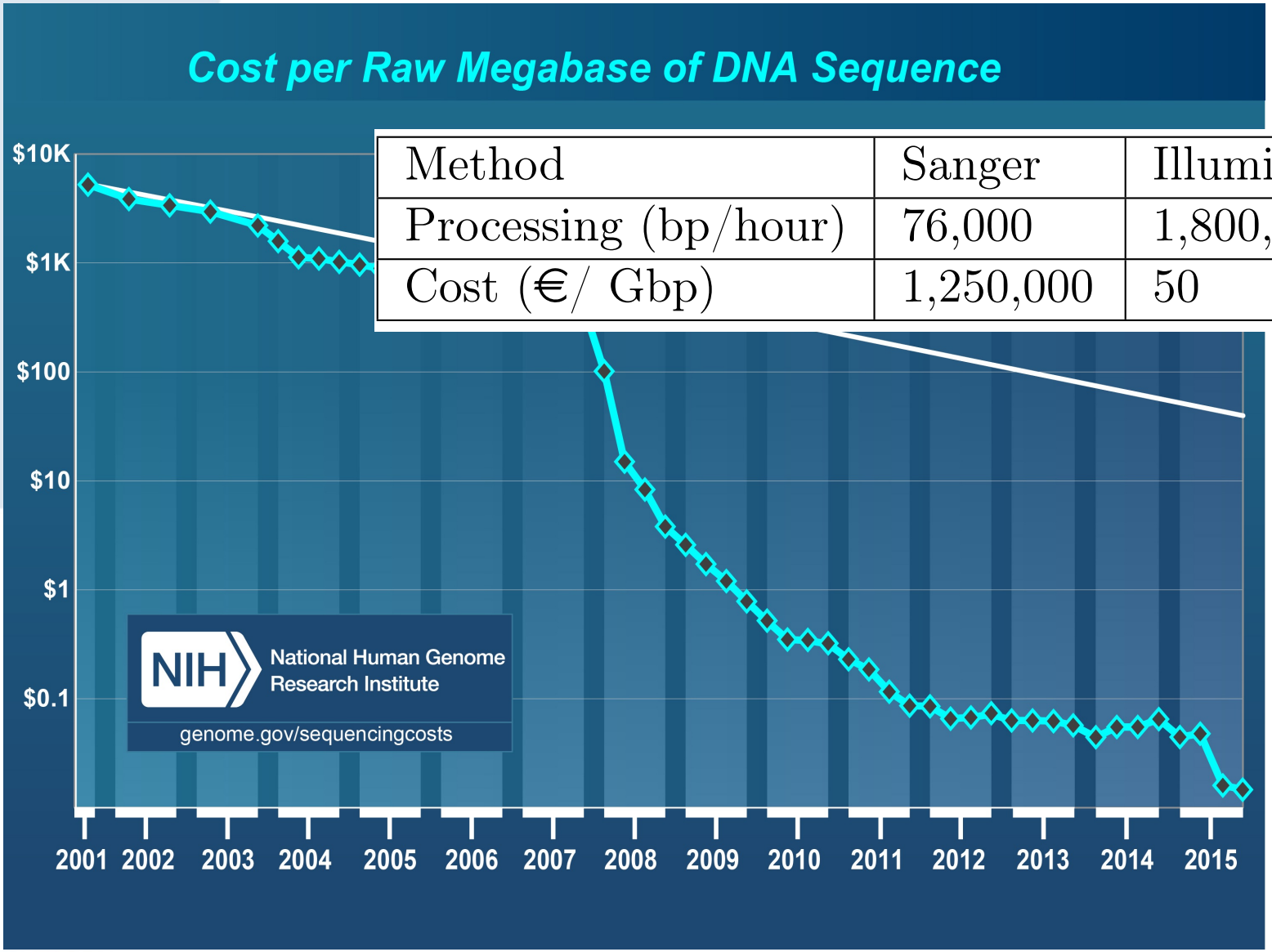
Gene Myers

New Generation Sequencing



Nadia Pisanti

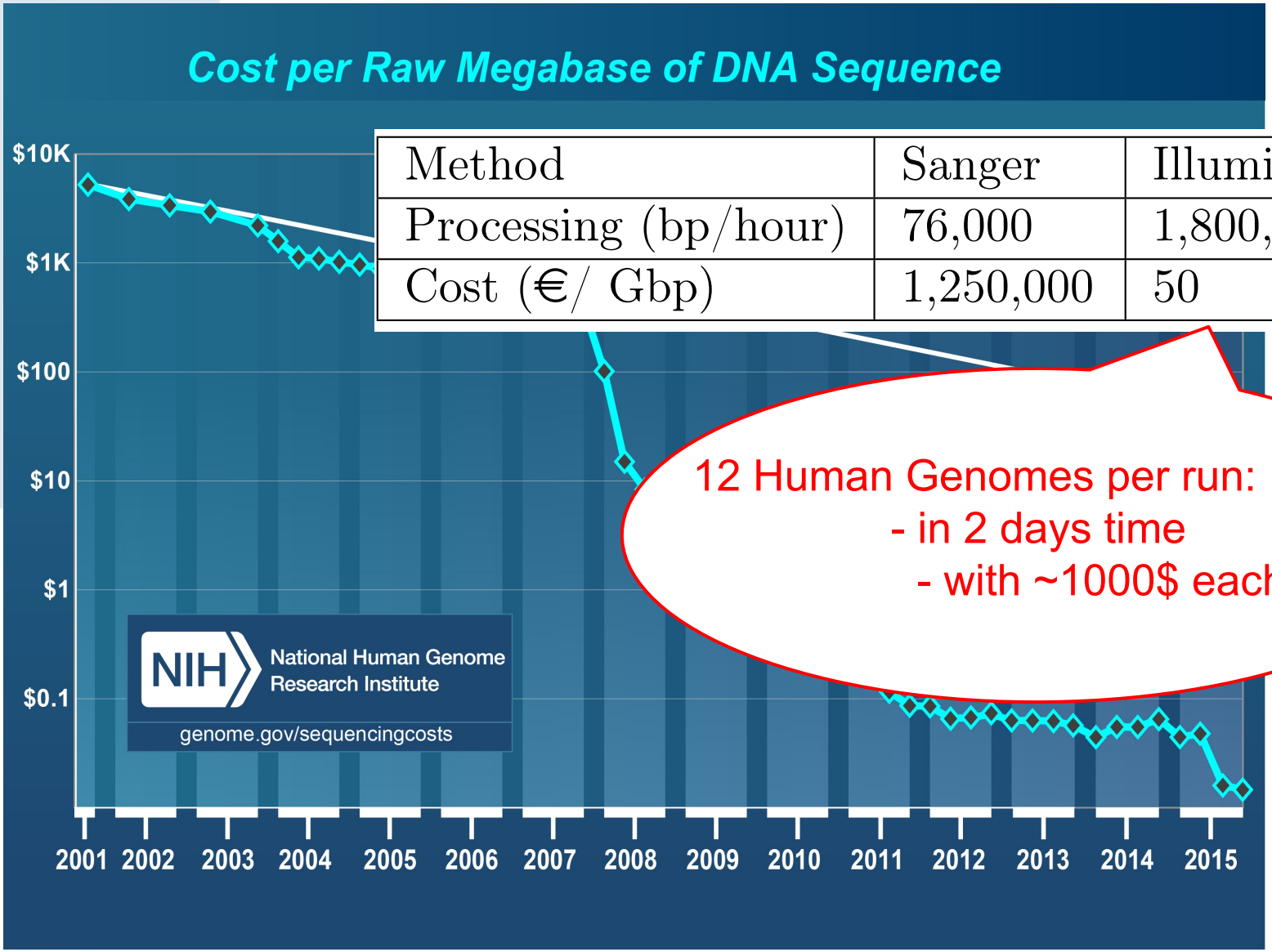
New Generation Sequencing



Method	Sanger	Illumina
Processing (bp/hour)	76,000	1,800,000,000
Cost (€/ Gbp)	1,250,000	50

Nadia Pisanti

New Generation Sequencing



Method	Sanger	Illumina
Processing (bp/hour)	76,000	1,800,000,000
Cost (€/ Gbp)	1,250,000	50

12 Human Genomes per run:
- in 2 days time
- with ~1000$ each

Nadia Pisanti

New Generation Sequencing

- **Illumina (Solexa)**
- **SOLiD**
- **ION Torrent**
- **Roche 454**
- **Pacific Bioscience**
- **Oxford Nanophore**
- **Moleculo**
- **...**





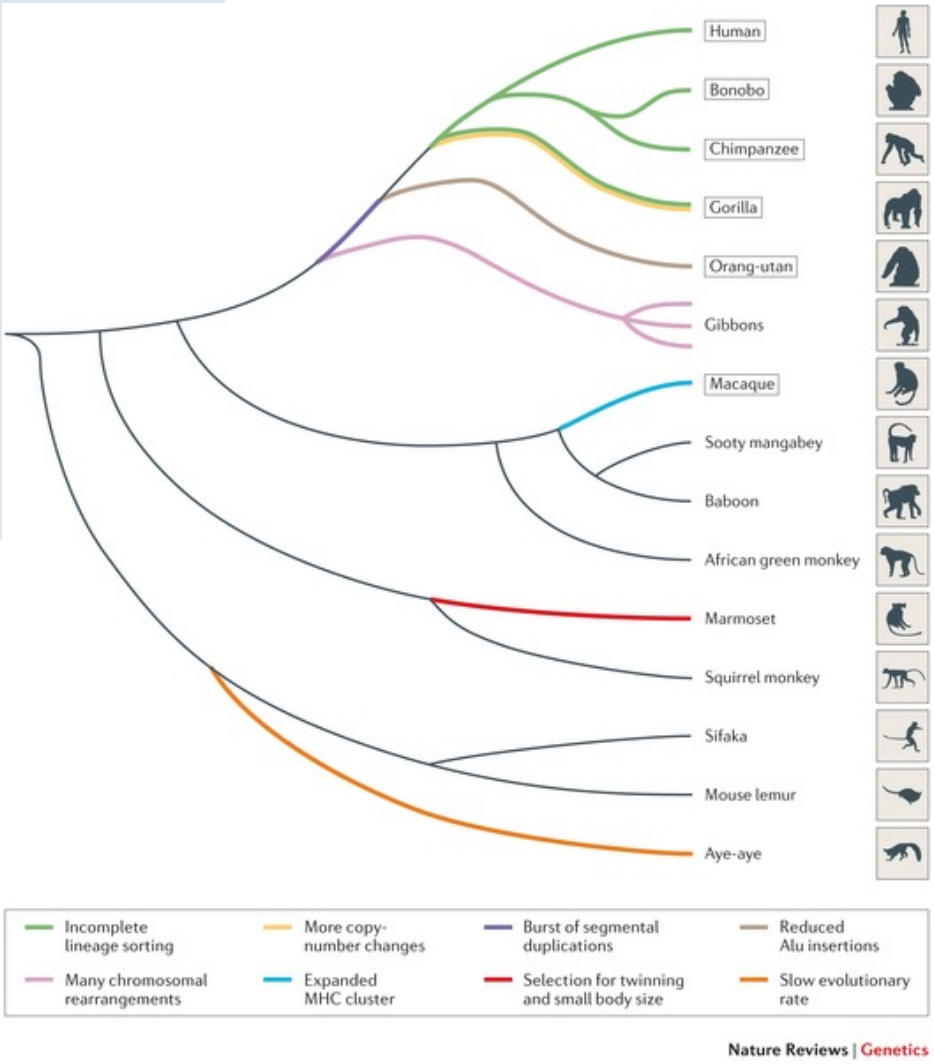
They differ in terms of costs, throughput, and performances (errors, time, read length, coverage, etc.)

They share the much lower costs and speed: millions (of millions) of fragments in a single and much cheaper run

Millions of Human Genomes

Nadia Pisanti

Millions of Genomes
WHAT CAN WE DO WITH THEM?



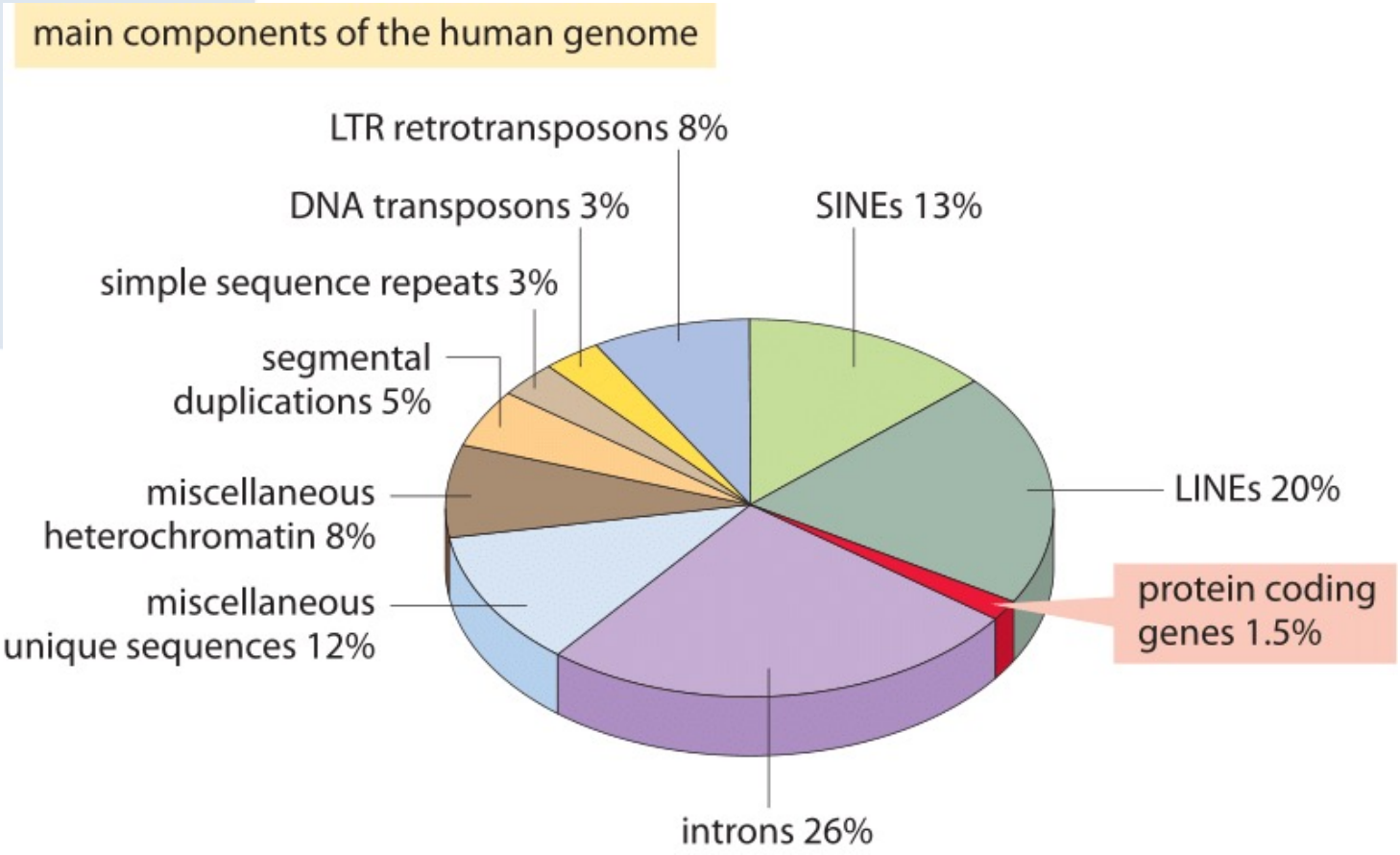
Comparing genomes of different species to study evolution processes.

Comparative Primates Genomics: Evolutionary Relationships

Nadia Pisanti

Millions of Genomes
WHAT CAN WE DO WITH THEM?

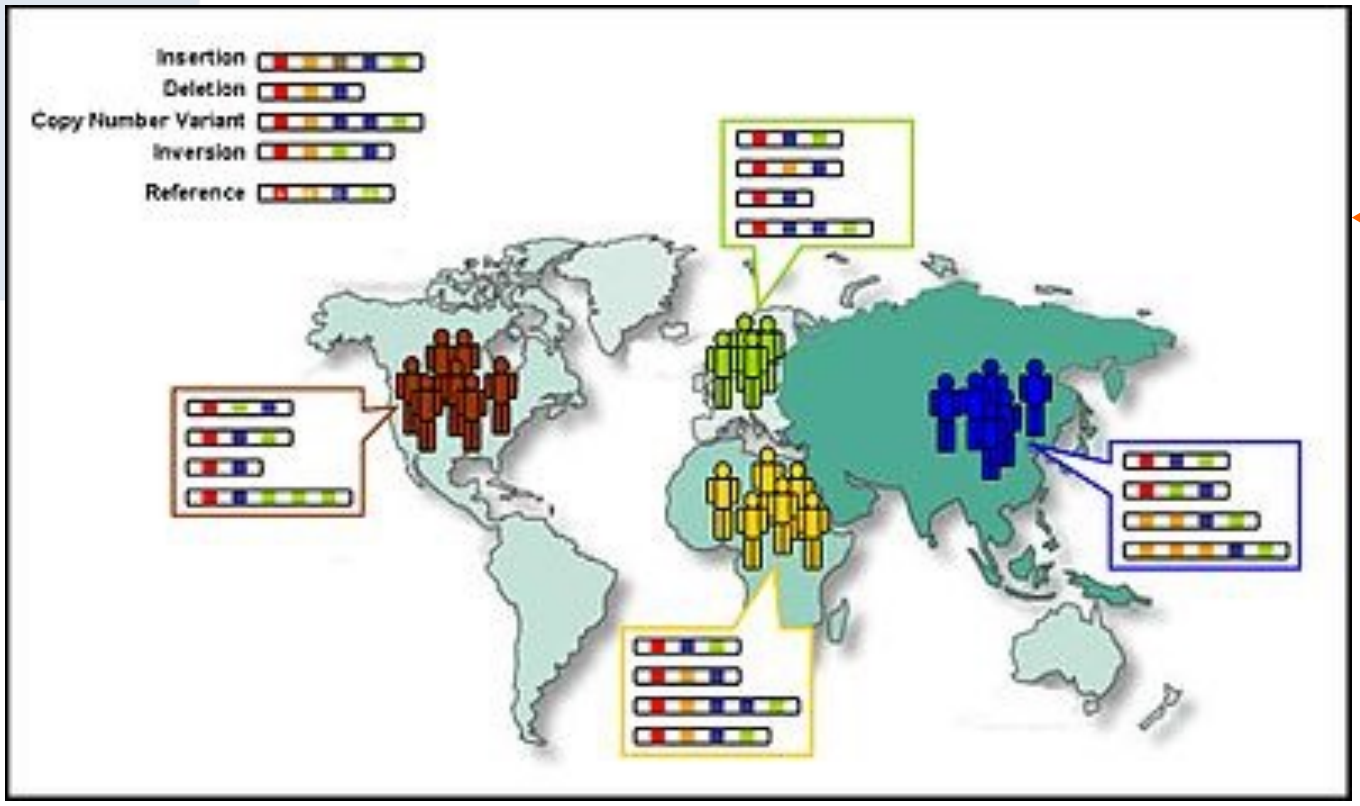
Analyse structural and functional features of genomes.



Nadia Pisanti

Millions of Genomes
WHAT CAN WE DO WITH THEM?

Study evolutionary dynamics of genomes and population genetics.

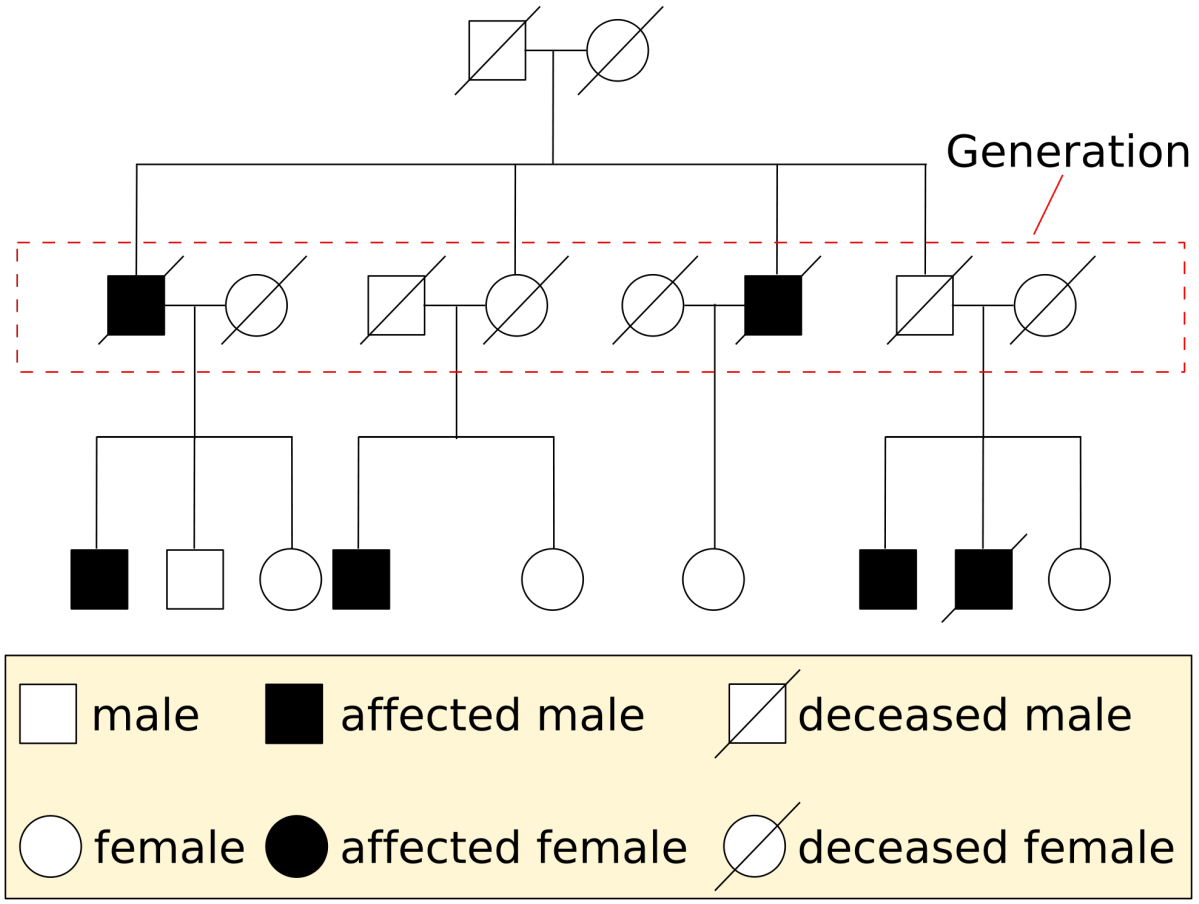


A map of genetic diversity between human populations

Nadia Pisanti

Millions of Genomes
WHAT CAN WE DO WITH THEM?

Discover molecular basis of complex traits.



Nadia Pisanti

Millions of Genomes
WHAT CAN WE DO WITH THEM?

Detect genetic variations of individuals

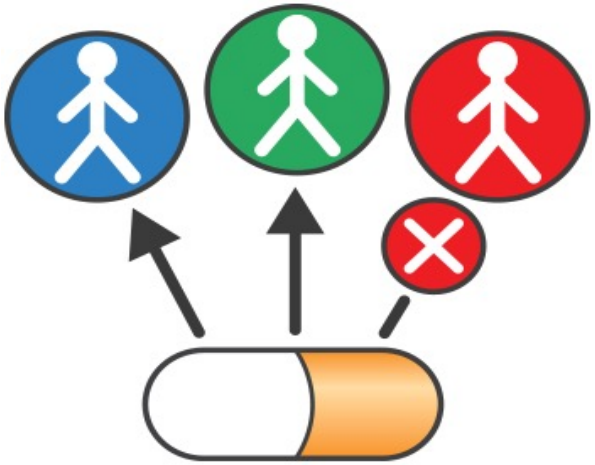
to detect and understand

the genetic causes of diseases

Or to check and understand

the genetic cause of

traitment success/failure

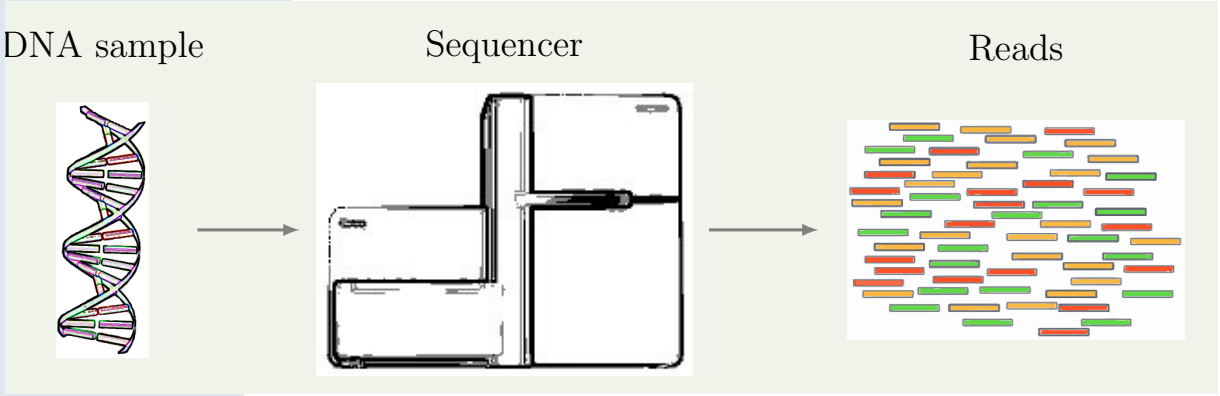


Nadia Pisanti

GENOME ASSEMBLY

Nadia Pisanti

Genome Assembly

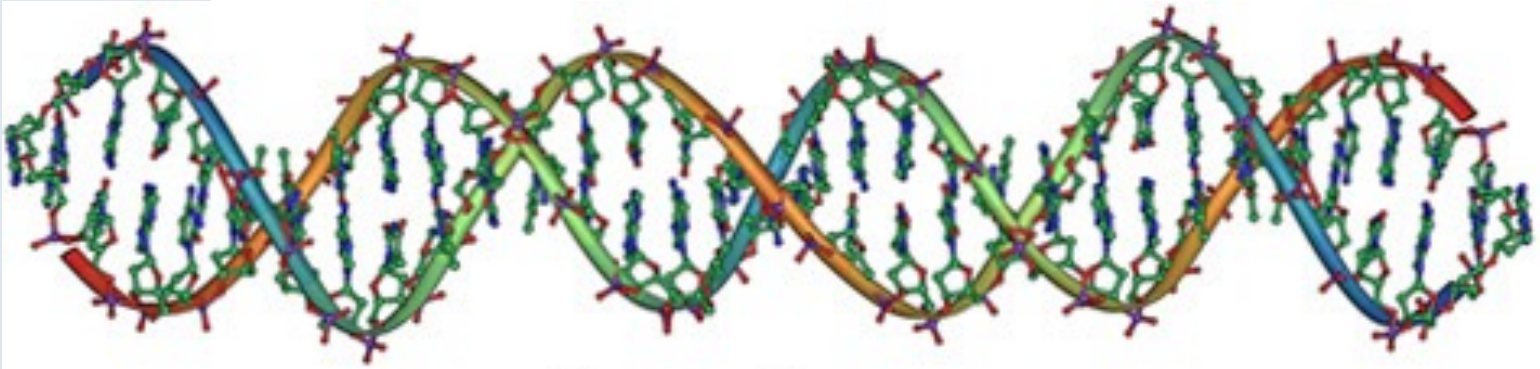


DNA is broken into million of pieces before sequencing…

"*Imagine a book cut by scissors into 10 million small pieces. Assuming that 1 million pieces are lost and the remaining 9 million are splashed with ink… try to recover the original text!*" [P.Pevzner, UCSD]

Genome Assembly

IDEA: replicate DNA before fragmenting it, and use overlap information to reconstruct the original complete sequence



Genome Sequence

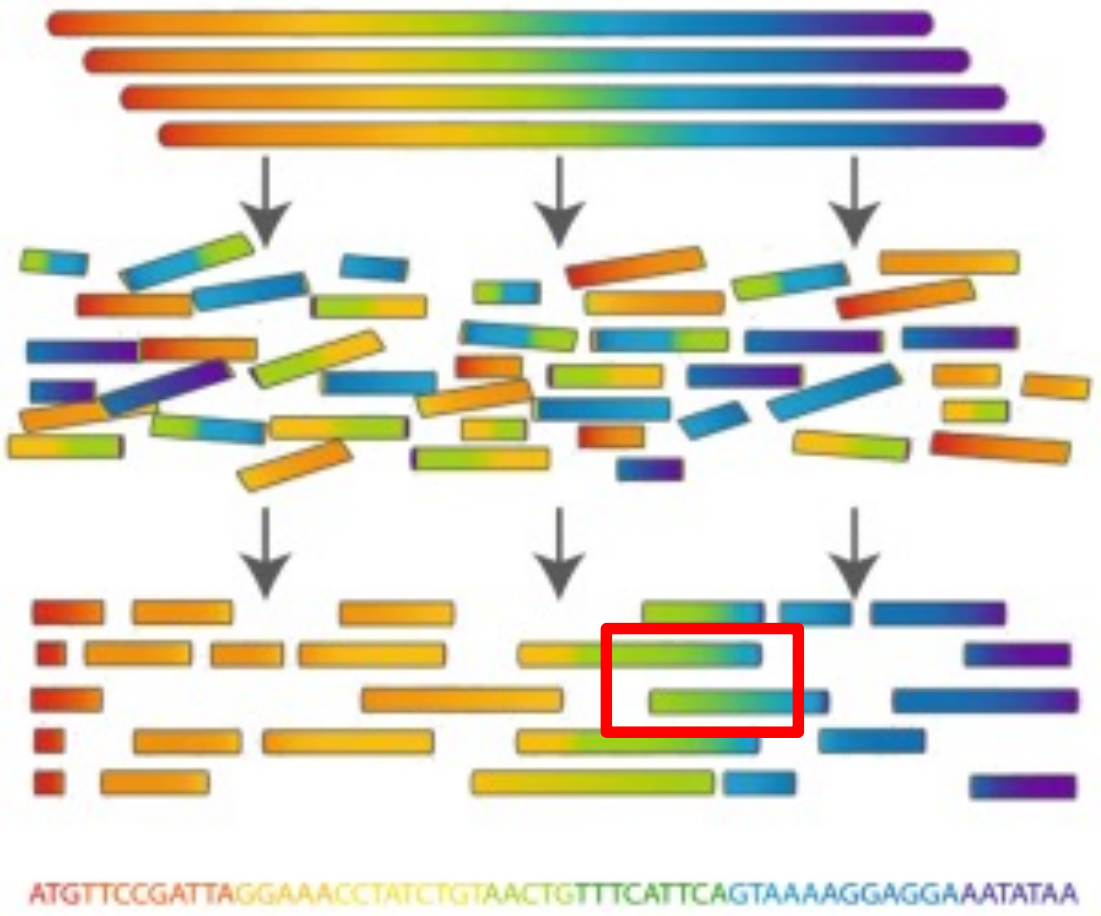
AGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGA

Fragment Reads

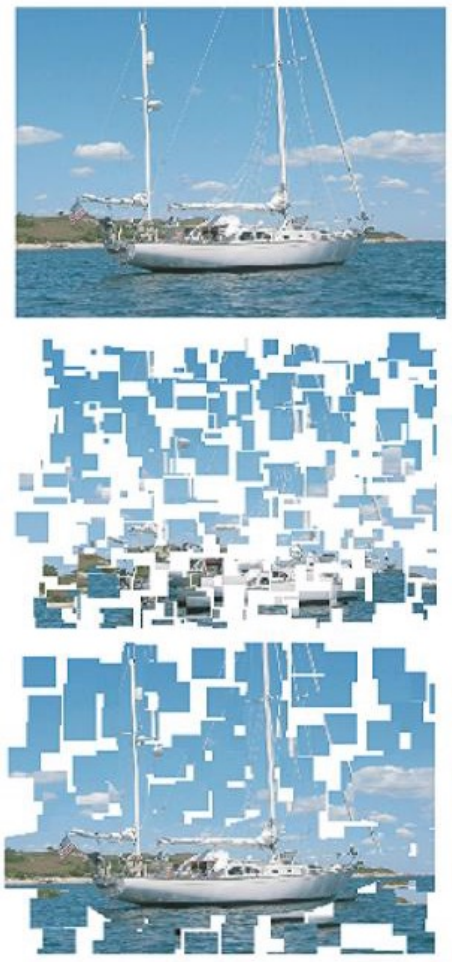
AGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACC
CTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGC
CCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGG
TGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAA
CTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGT
AGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGA

Nadia Pisanti

Genome Assembly



Genome Assembly



COVERAGE

ERROR RATE

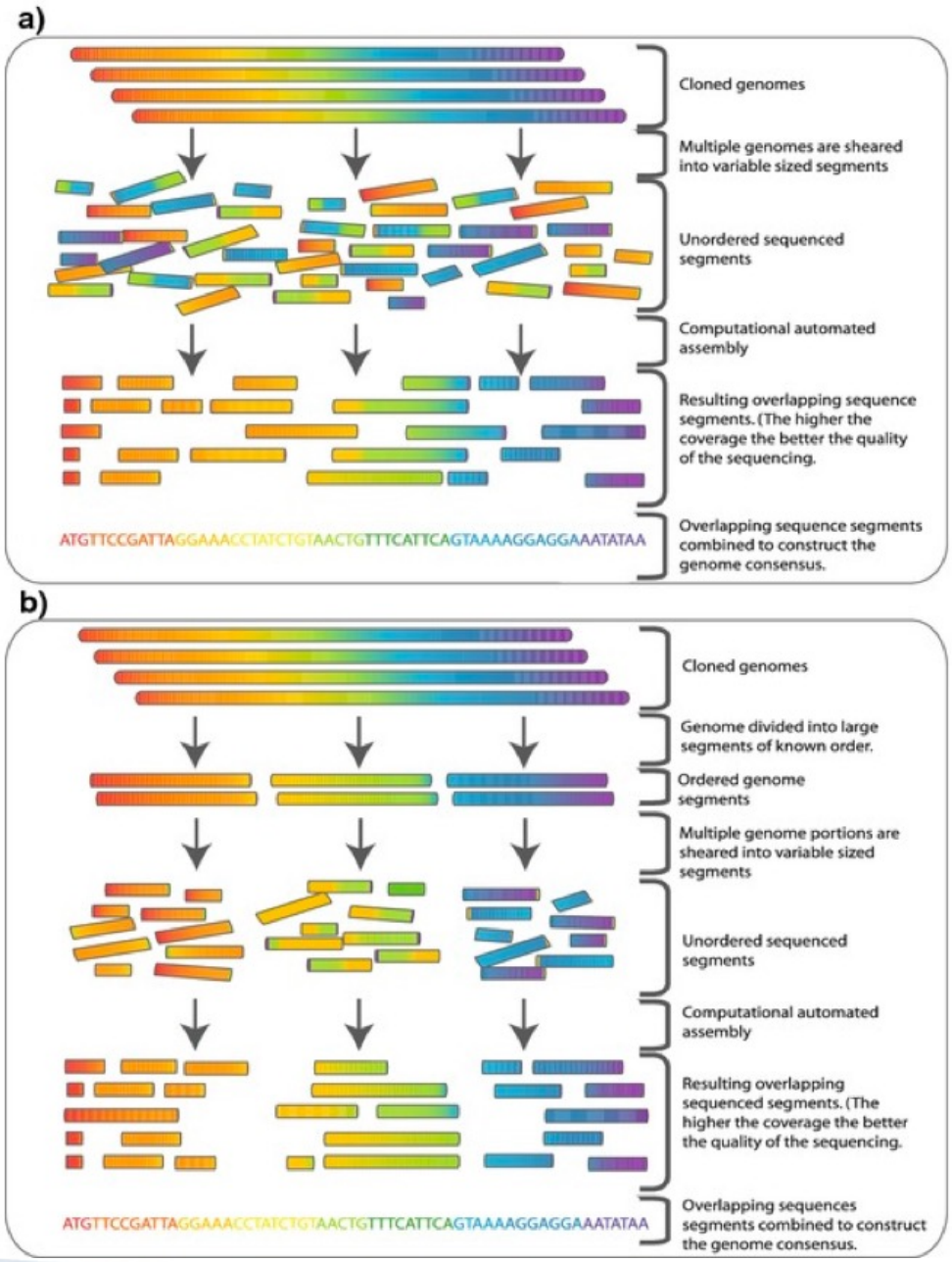
READS LENGTH

SINGLE/PAIRED READS

... in the New Generation Sequencing Scenario!

Nadia Pisanti

"shotgun" Genome Assembly

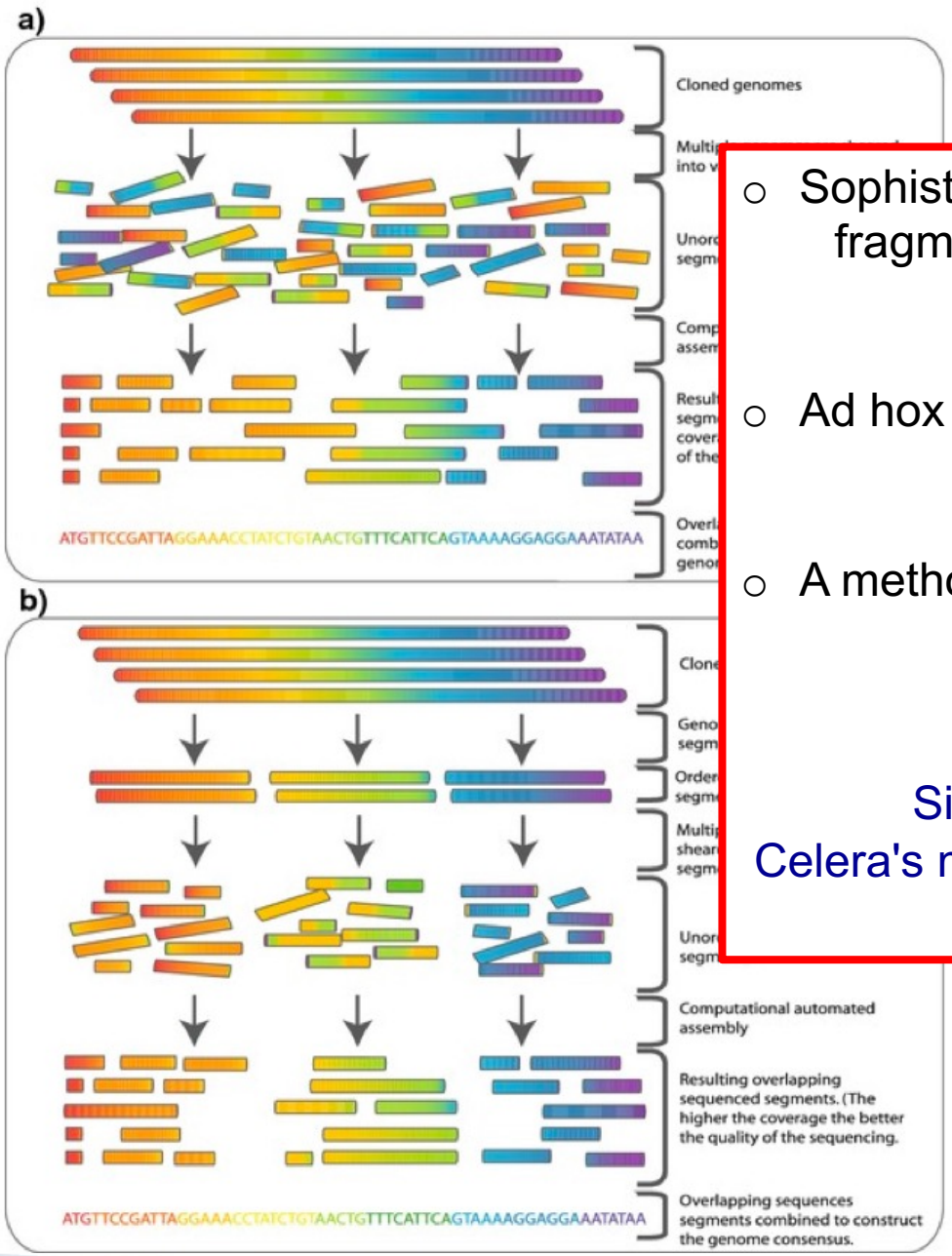


Celera

HGP

Nadia Pisanti

"shotgun" Genome Assembly



- Sophisticated data structures to index fragments and represent layout.
- Ad hox indices and algorithms to align reads
- A method to "solve" repetitions

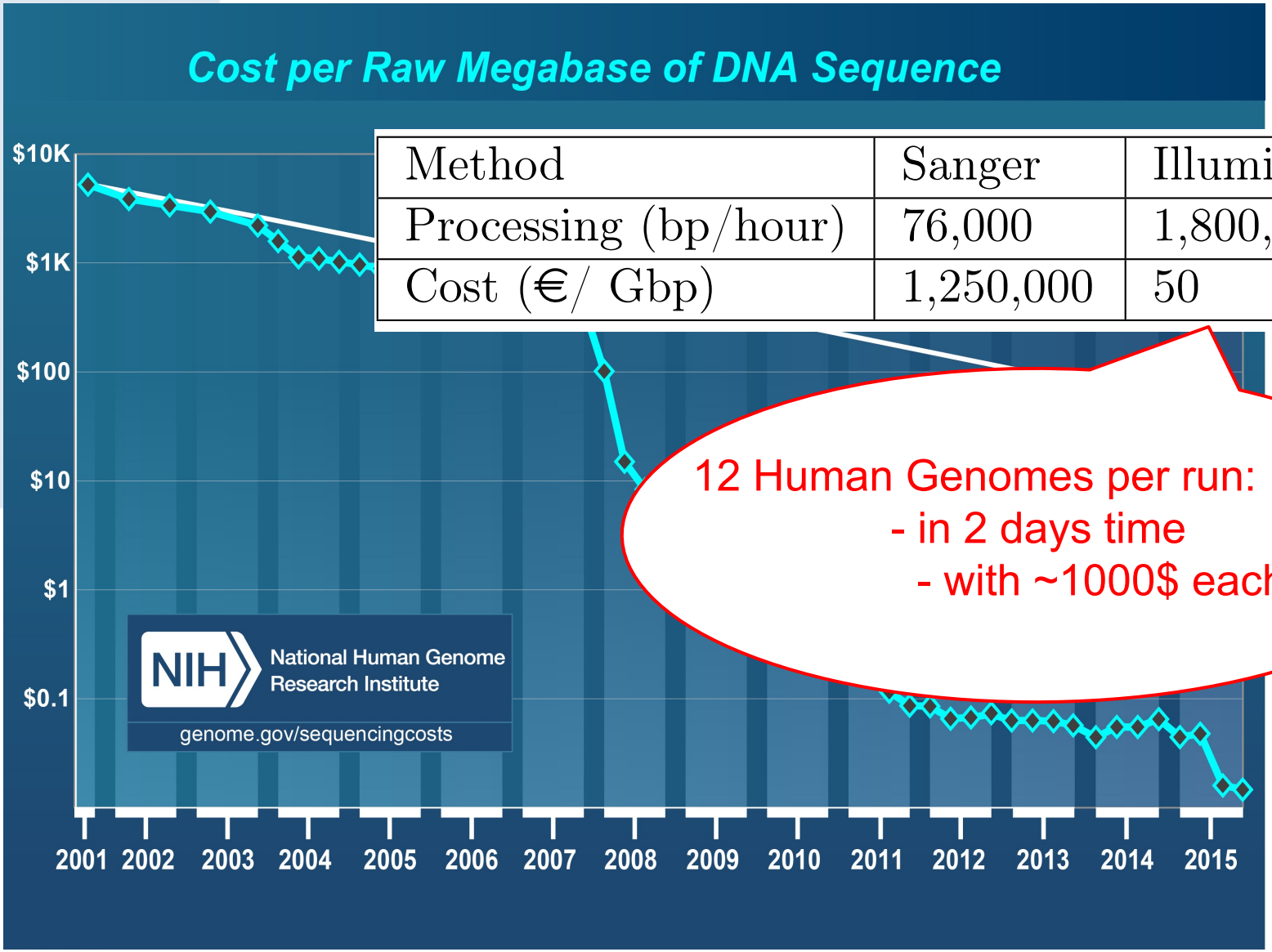
Since Human Genome Assembly,
Celera's method became the standard.. with new algorithmic insights....

Nadia Pisanti

RE-SEQUENCING

Nadia Pisanti

New Generation Sequencing



Method	Sanger	Illumina
Processing (bp/hour)	76,000	1,800,000,000
Cost (€/ Gbp)	1,250,000	50

12 Human Genomes per run:
- in 2 days time
- with ~1000$ each

Nadia Pisanti

RE-SEQUENCING

Reference genome is already available for the species
(e.g. the human genome)

- Comparing a "new" individual to the reference genome.

- First step: mapping onto reference genome (to correctly determine corresponding location in the reference genome).

Very complicated task that depends on many factors: genetic variation in the population, sequencing error, read length, and the huge volume of short reads to be mapped.

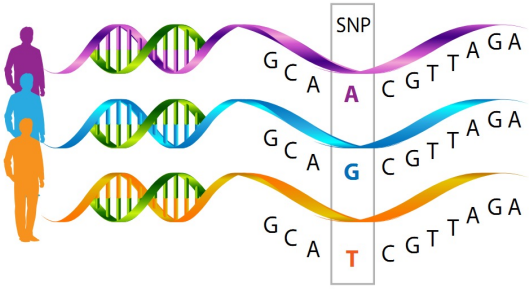
In the last years many algorithms have been developed to overcome these challenges and these algorithms have been made available to the scientific community as software packages

POLYMORPHISMS

Individual genomic variations within a specie are called **polymorphisms**

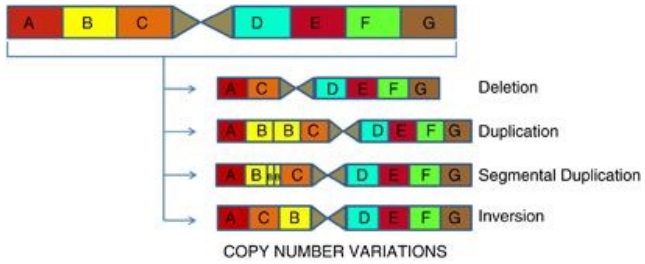
SNPs
Single Nucleotide Polymorphism

There are ~ 10M SNPs in the Human Genome, averagely one every 1000b



CNVs
Copy Number Variation

From 5 to 10% of Human Genome contributes to repeated sequences of size ranging from 50b to 3Mb



Nadia Pisanti

VARIATIONS ANALYSIS

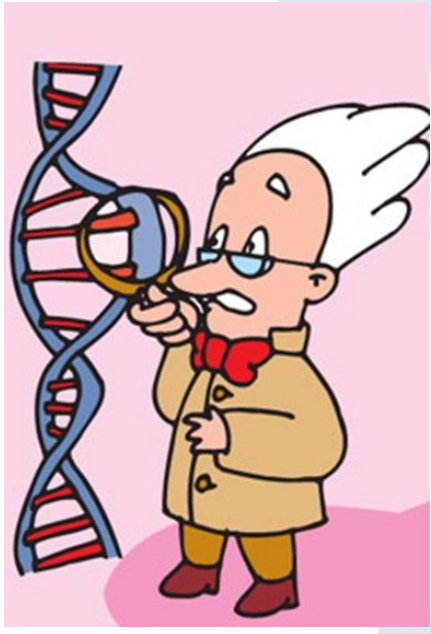
Polymorphisms are common and physiological variations

(some variations characterize a population)

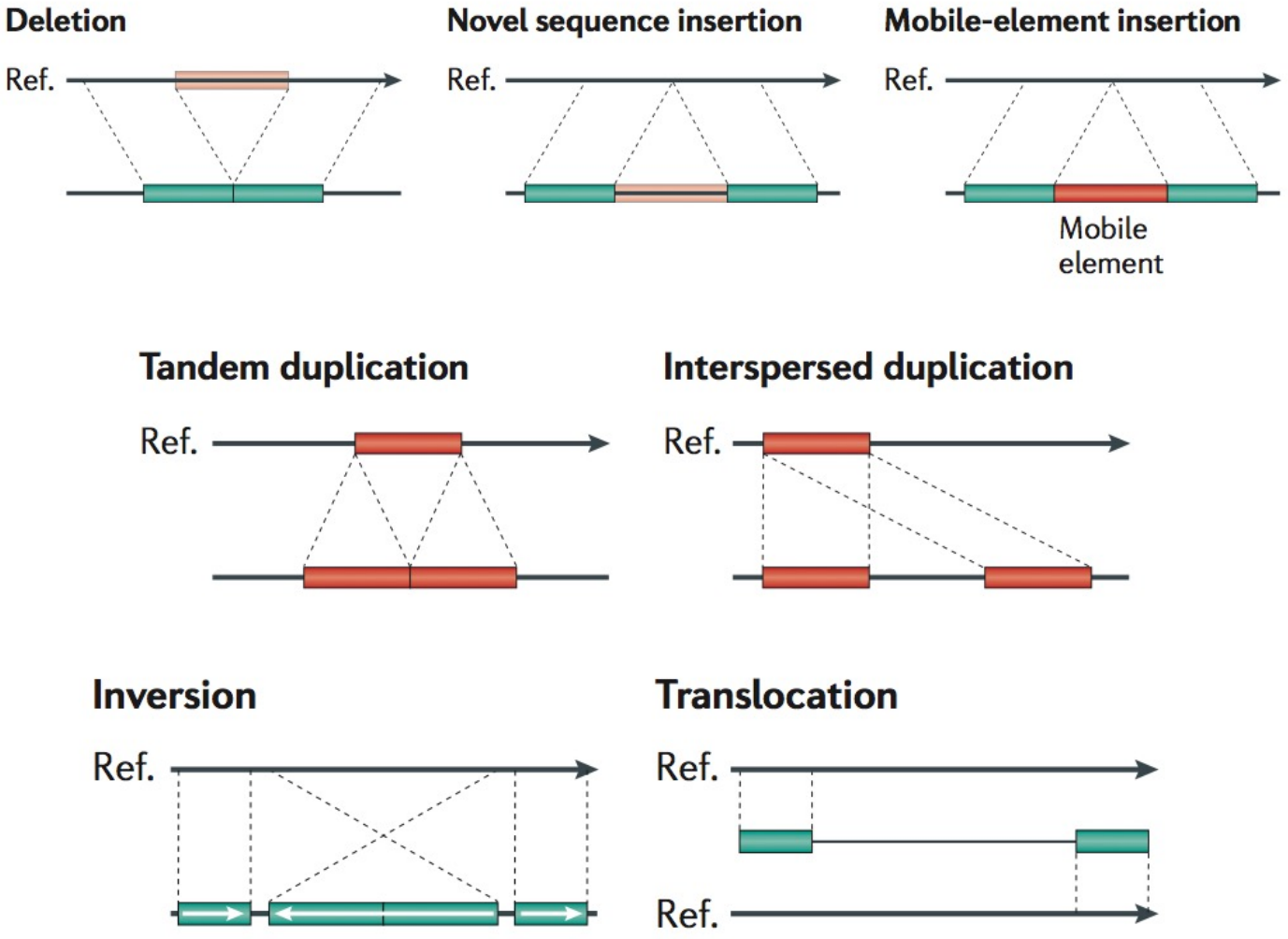
Mutations are more rare and *can be* associated to (a predisposition to) a disease

or be caused by a disease

or can be  d to drug response



STRUCTURAL VARIATIONS



Nadia Pisanti

HUMAN GENETIC VARIATIONS

Structural Variants influence gene expression

altering gene dosage  phenotypic variation

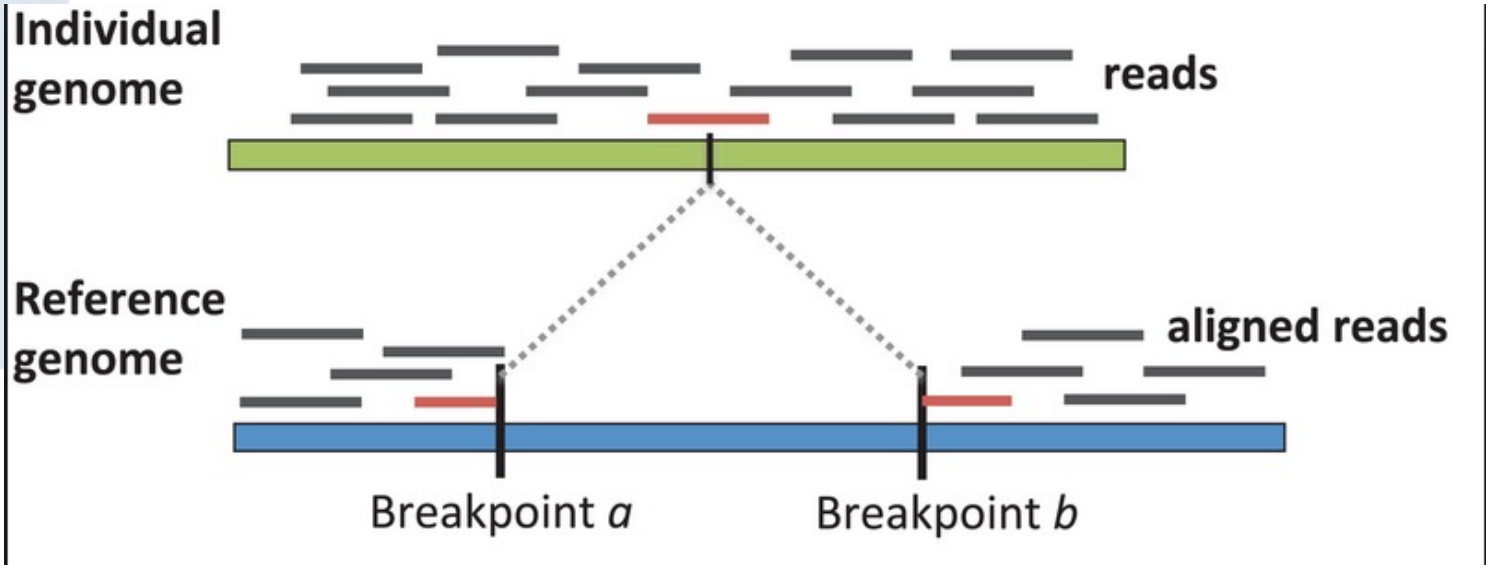
Structural Variants (SVs) are found among "normal" individuals

others occur in the course of normal process

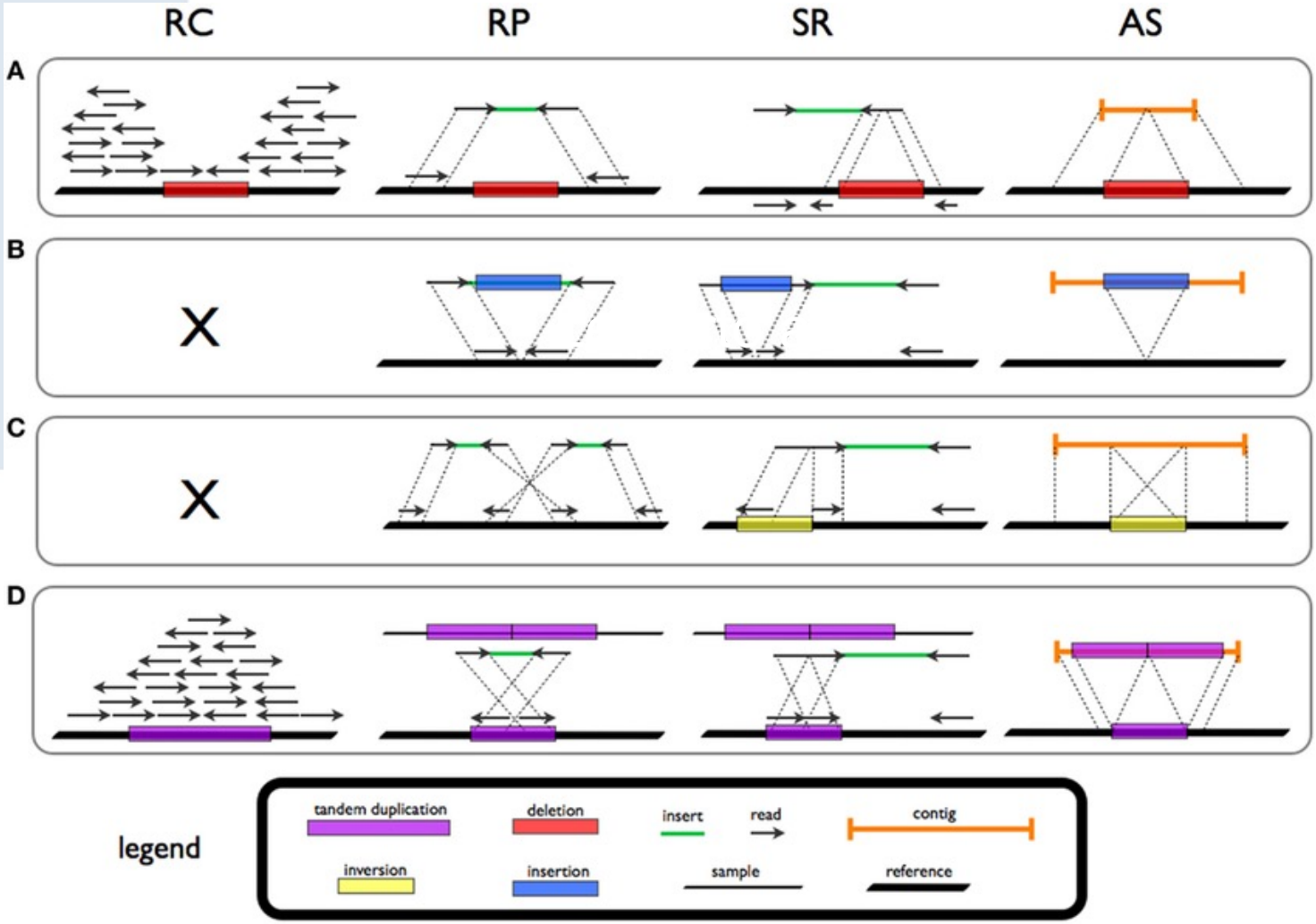
still other patecipate in causing variuos disease states

Nadia Pisanti

Detecting Structural Variants



Detecting Structural Variants



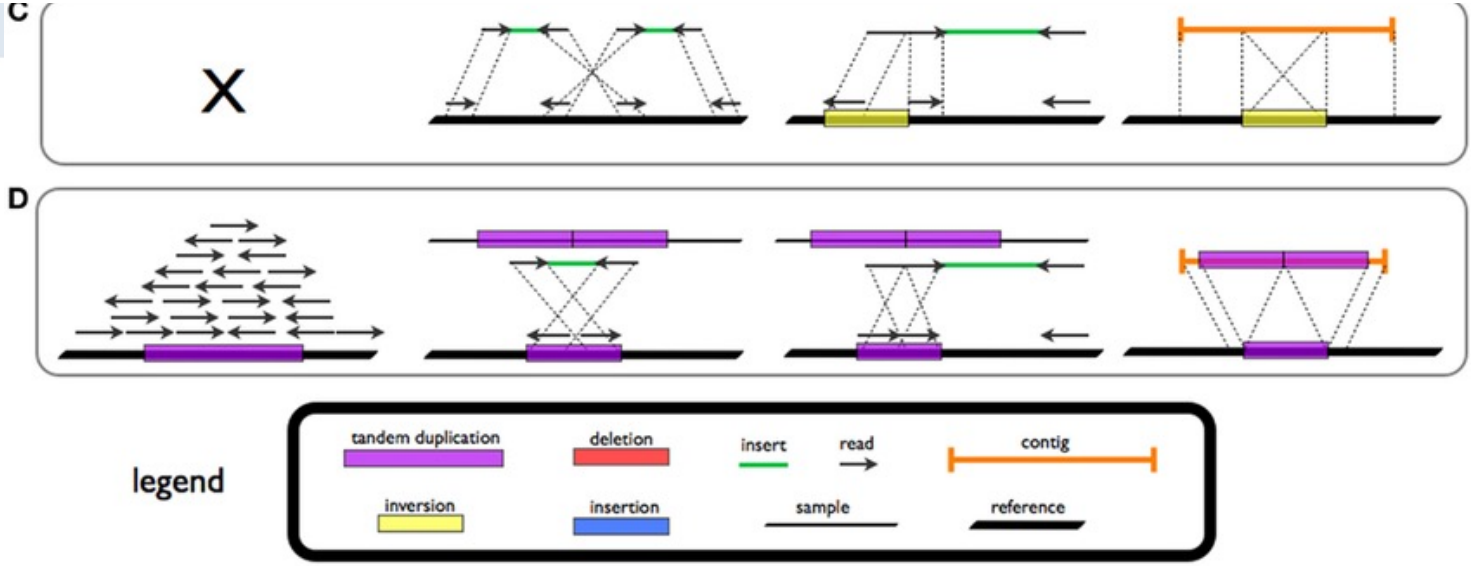
Detecting Structural Variants

RC RP SR AS

IT IS ALL MAPPING ON REFERENCE GENOME

WITH SOME SPECIFIC ALGORITHMICS

THAT DEPEND FROM THE GENOME VARIANT

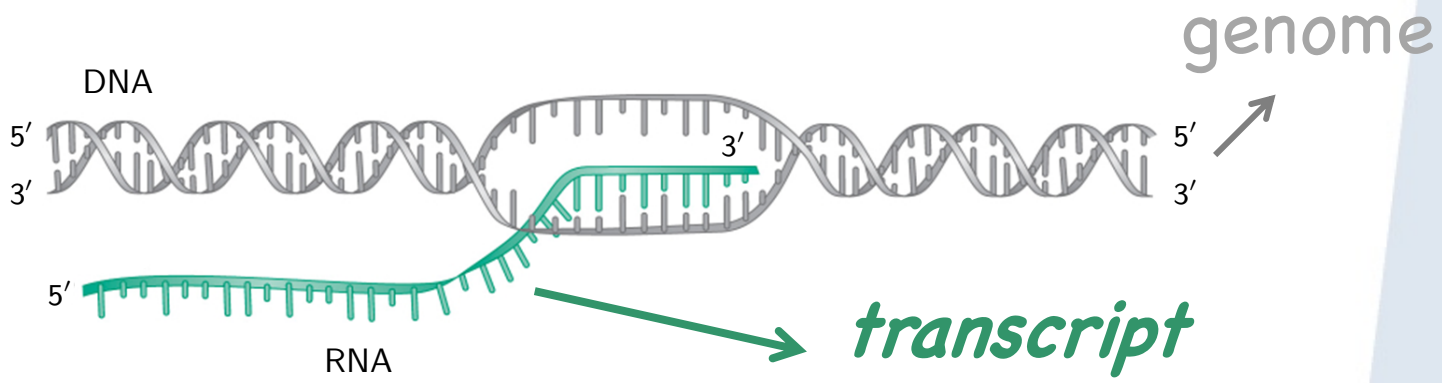


RNA-Seq

Nadia Pisanti

Genome and Transcriptome

DNA transcribes into RNA

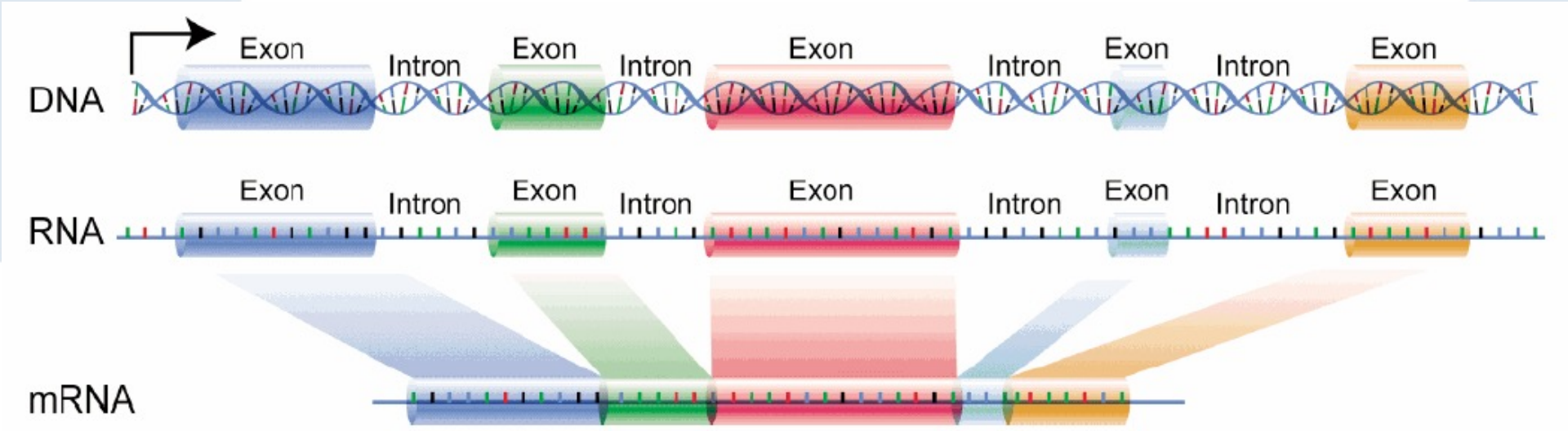


And then the RNA is translated into proteins, that actually determine what happens in our cells

Nadia Pisanti

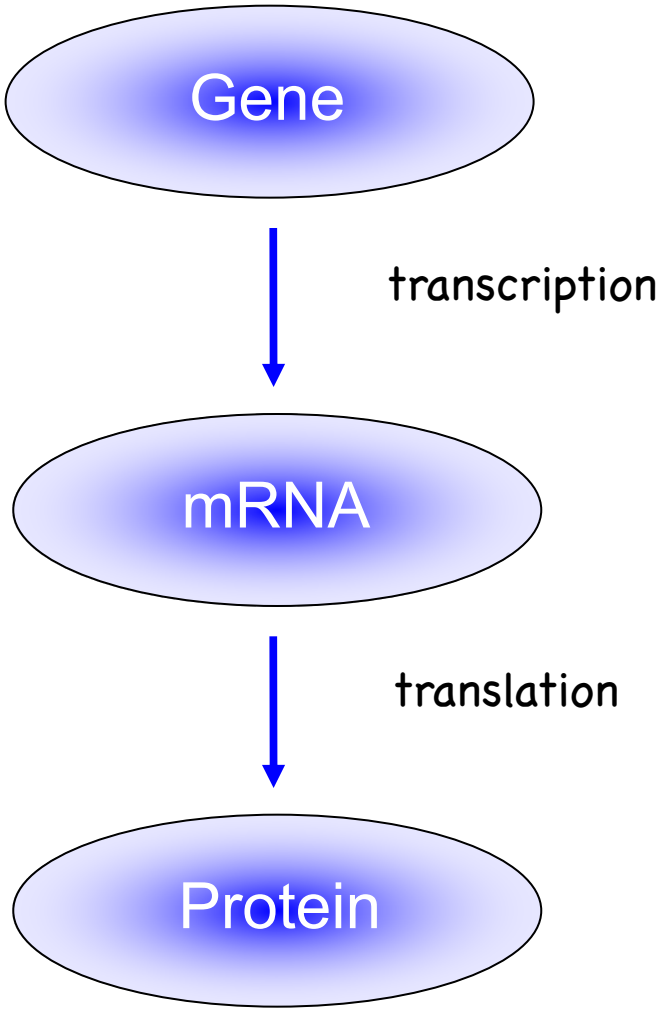
Genome and Transcriptome

DNA transcribes into RNA



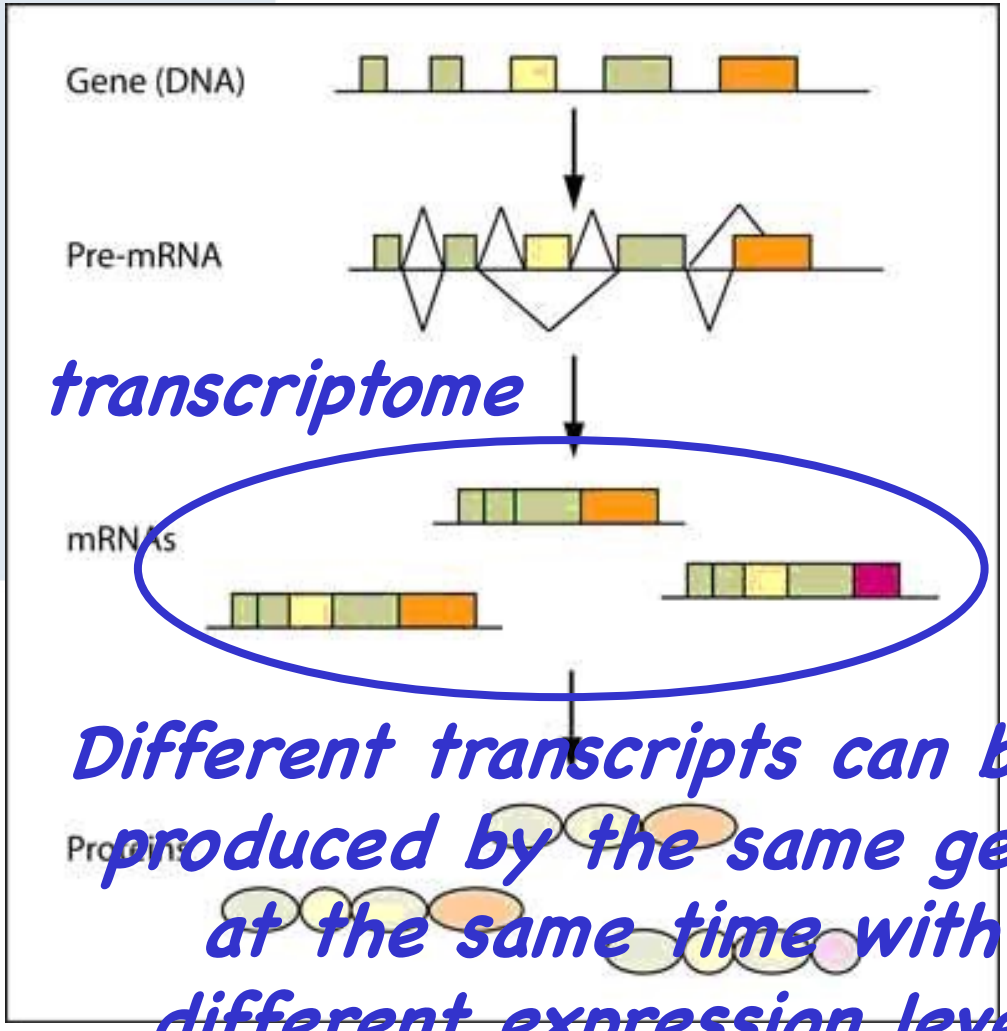
Nadia Pisanti

Genome and Transcriptome

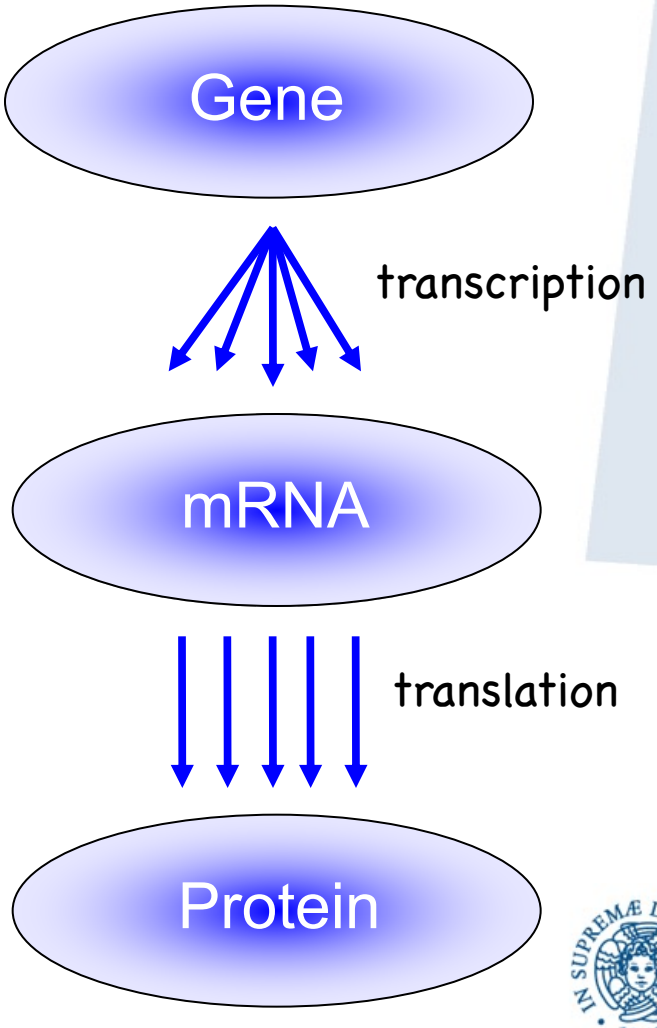


Nadia Pisanti

Genome and Transcriptome

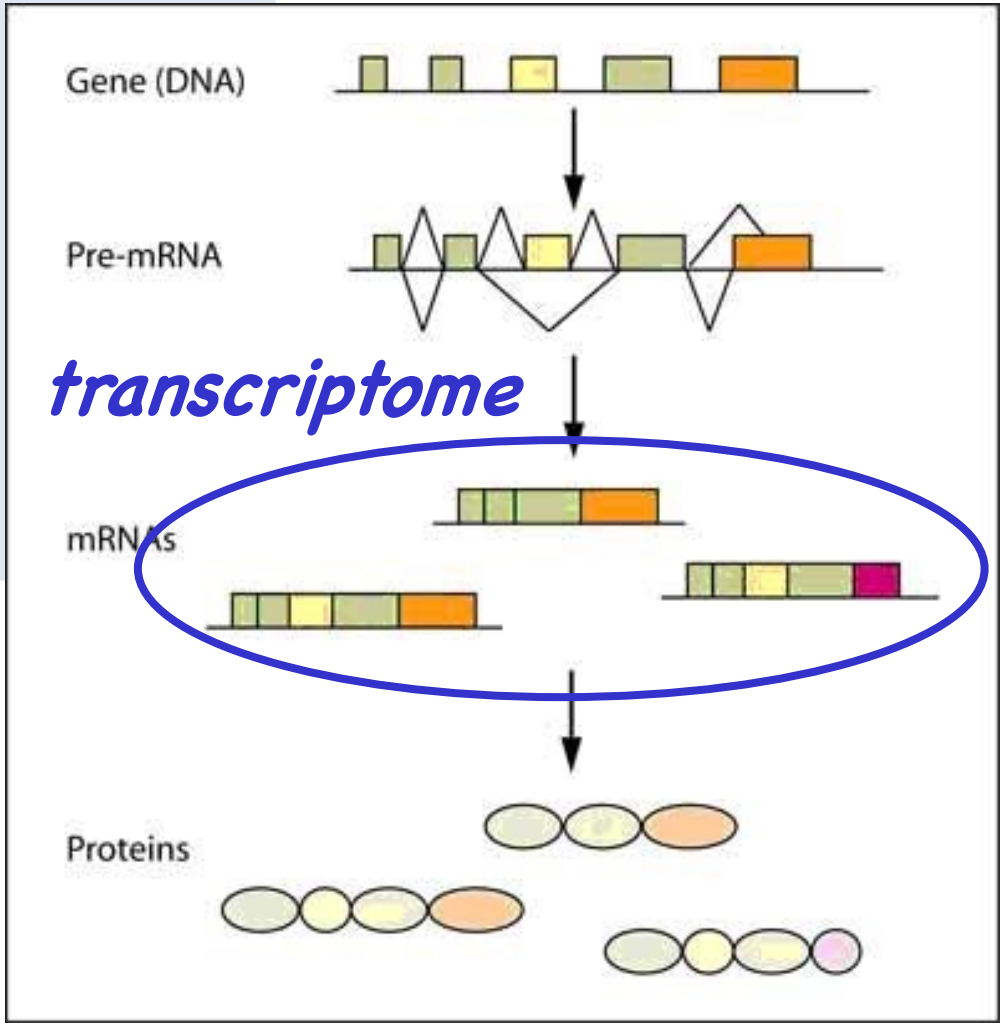


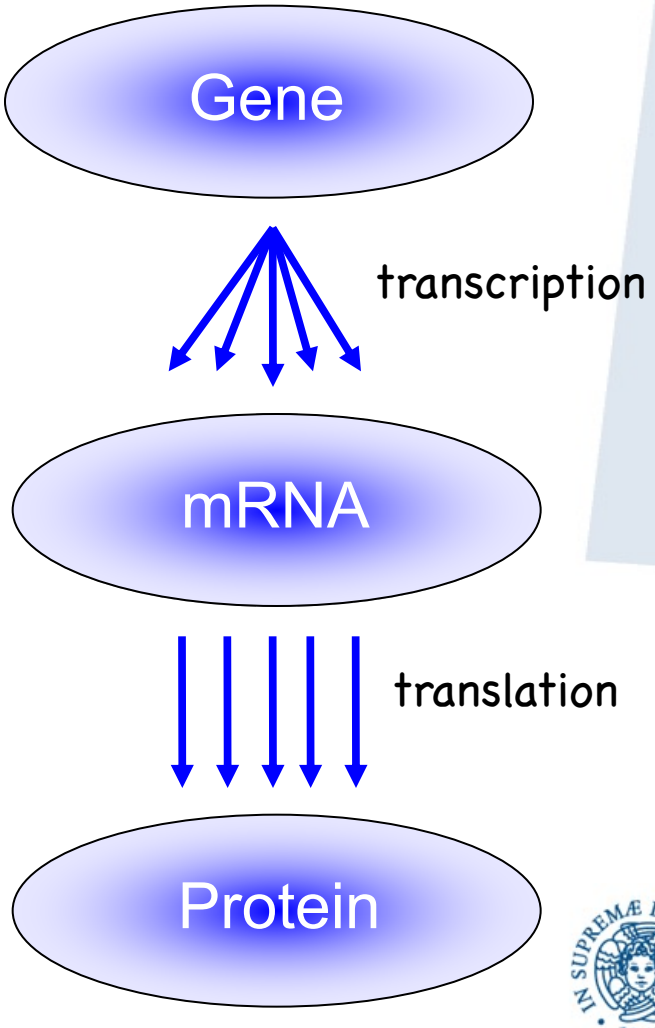
Different transcripts can be produced by the same gene at the same time with different expression level



Nadia Pisanti

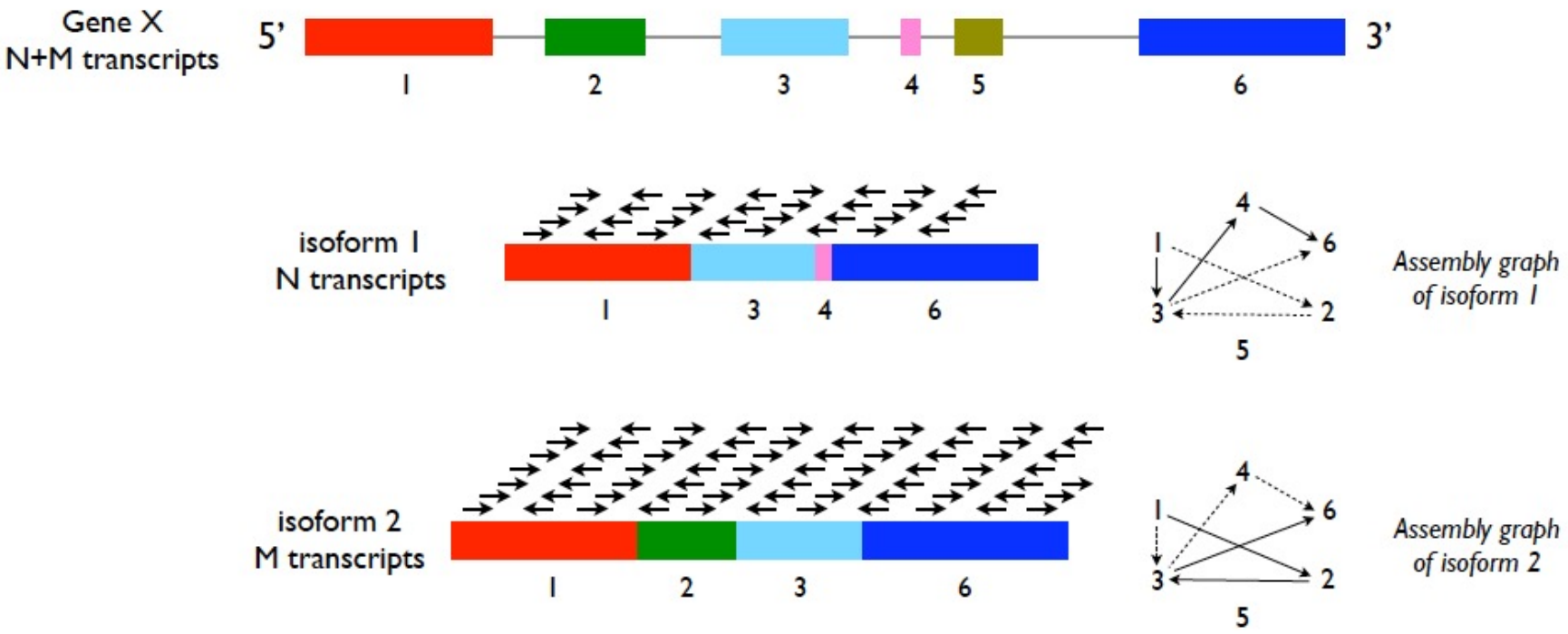
Genome and Transcriptome





Nadia Pisanti

Genome and Transcriptome



Nadia Pisanti

Transcriptome Sequencing

- Genes express differently even in the same individual:
  - In different conditions
  - In different times
  - In different tissues

- Why and How does the transcriptome change?

(much more than the genome!)

Transcriptome Sequencing

- Genes express differently even in the same individual:
  - In different conditions
  - In different times
  - In different tissues

- Why and How does the transcriptome change?

(much more than the genome!)

? regulation mechanisms ?

Nadia Pisanti

Transcriptome Sequencing

with New Generation Sequencing it is possible to sequence the transcriptome: RNA-Seq

- Genes express differently even in the same individual:
  - In different conditions
  - In different times
  - In different tissues

- When and How much does the transcriptome change?

(much more than the genome!)

Differential Expression of genes

Nadia Pisanti

Transcriptome Sequencing

with New Generation Sequencing it is possible to sequence the transcriptome: RNA-Seq

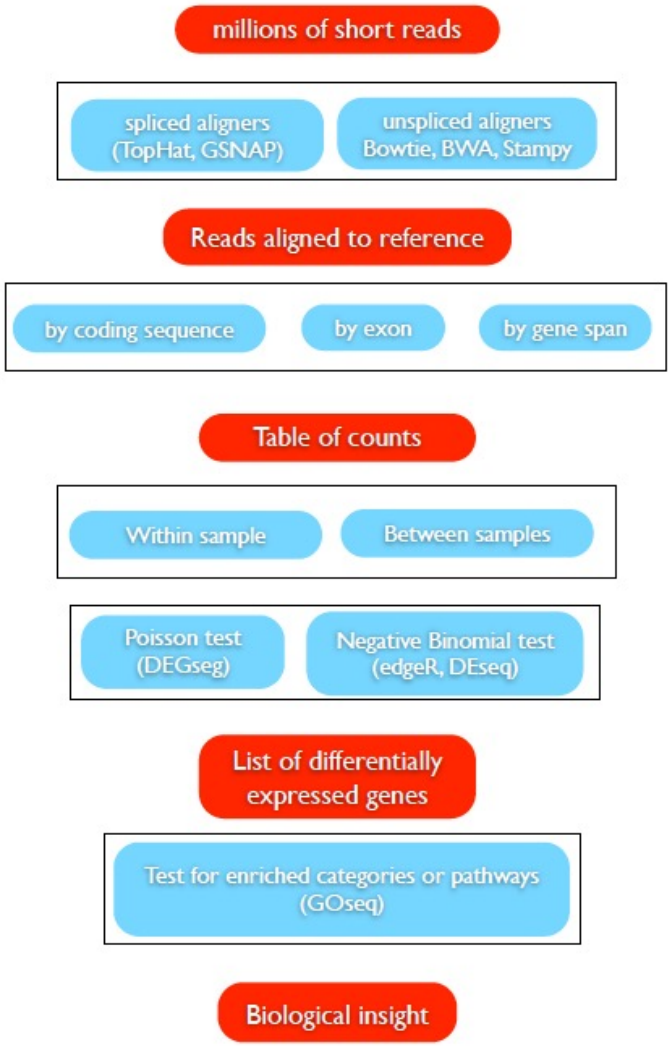
Detect all expressed RNA in a cell at a given time:

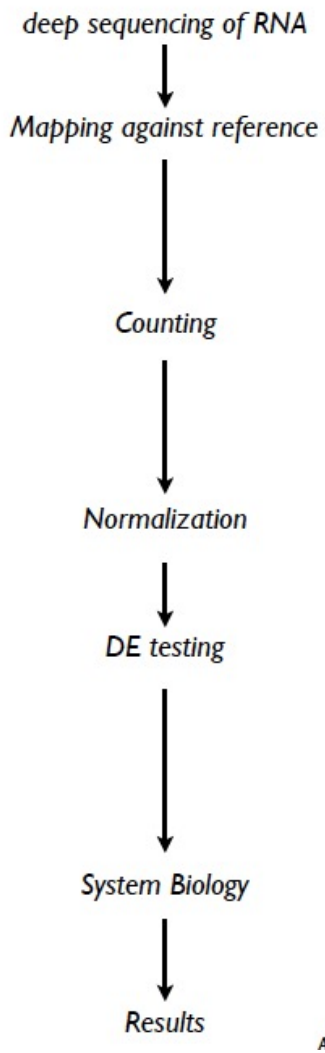
- with their genomic location, and
- with estimation of expressed transcript abundance

Nadia Pisanti

Transcriptome Sequencing

DE-seq analysis pipeline

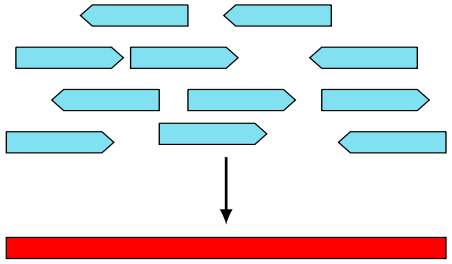




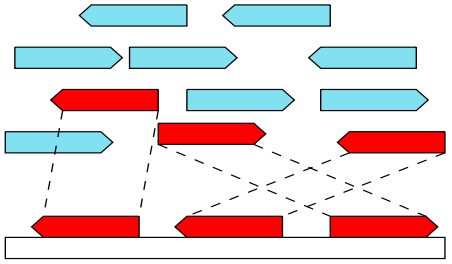
SUMMING UP

Nadia Pisanti

COMPUTATIONAL PROBLEMS ON BIOLOGICAL SEQUENCES ANALYSIS



assembly



alignment

Provide specialised tools for variant discovery and detection

Map the fragments to a **reference genome**: alignment

Map RNA-Seq transcripts onto a reference genome

reconstruct the original DNA sequence: de novo assembly

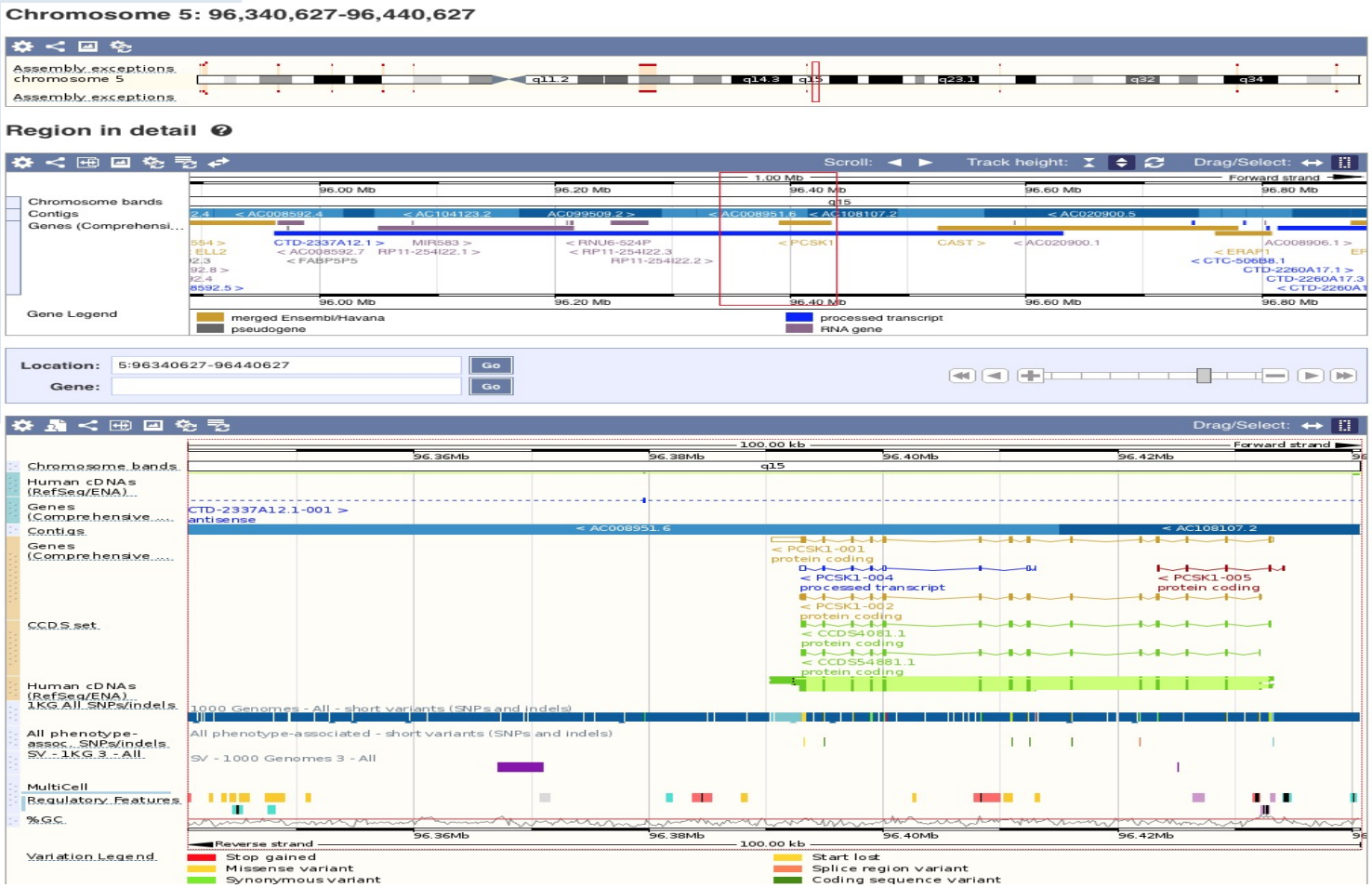
Compute transcript abundance for gene expression level estimation

Nadia Pisanti

GENOMIC DATABANKS

It is necessary to store a huge amount of data in centralised DataBanks

Develop tools for accessing, using, and visualising…



genome browser

Nadia Pisanti

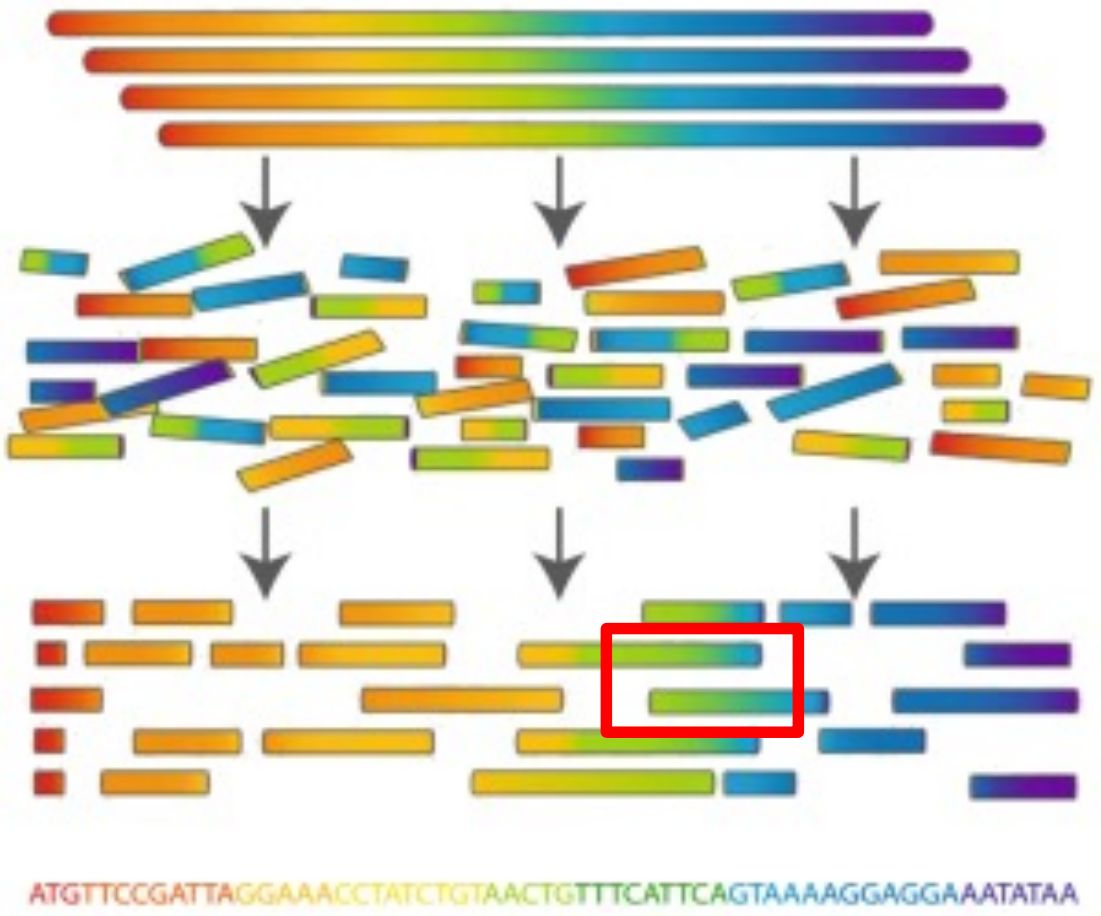
SEQUENCES COMPARISON

a.k.a.

SEQUENCES ALIGNMENT

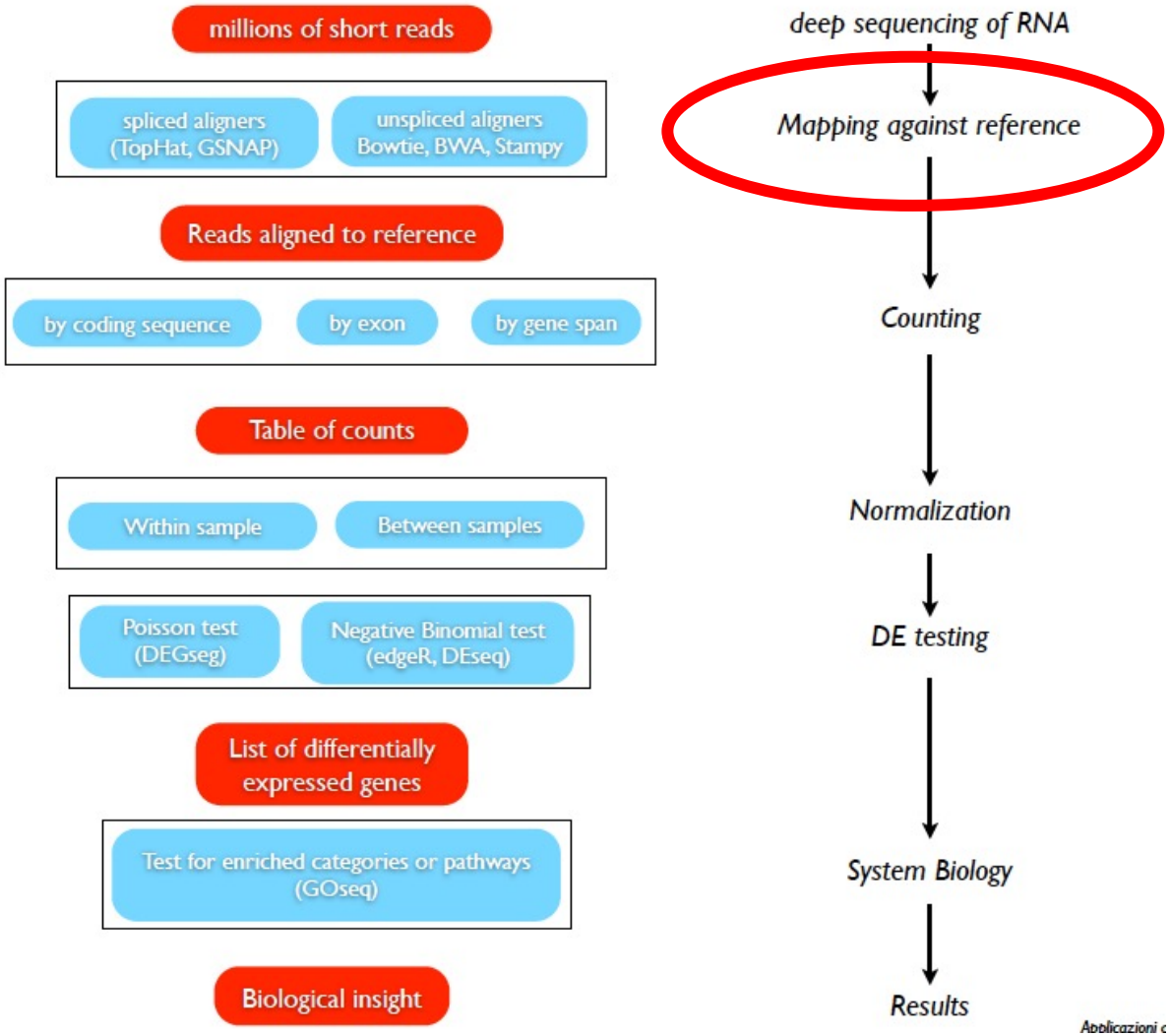
Nadia Pisanti

Genome Assembly



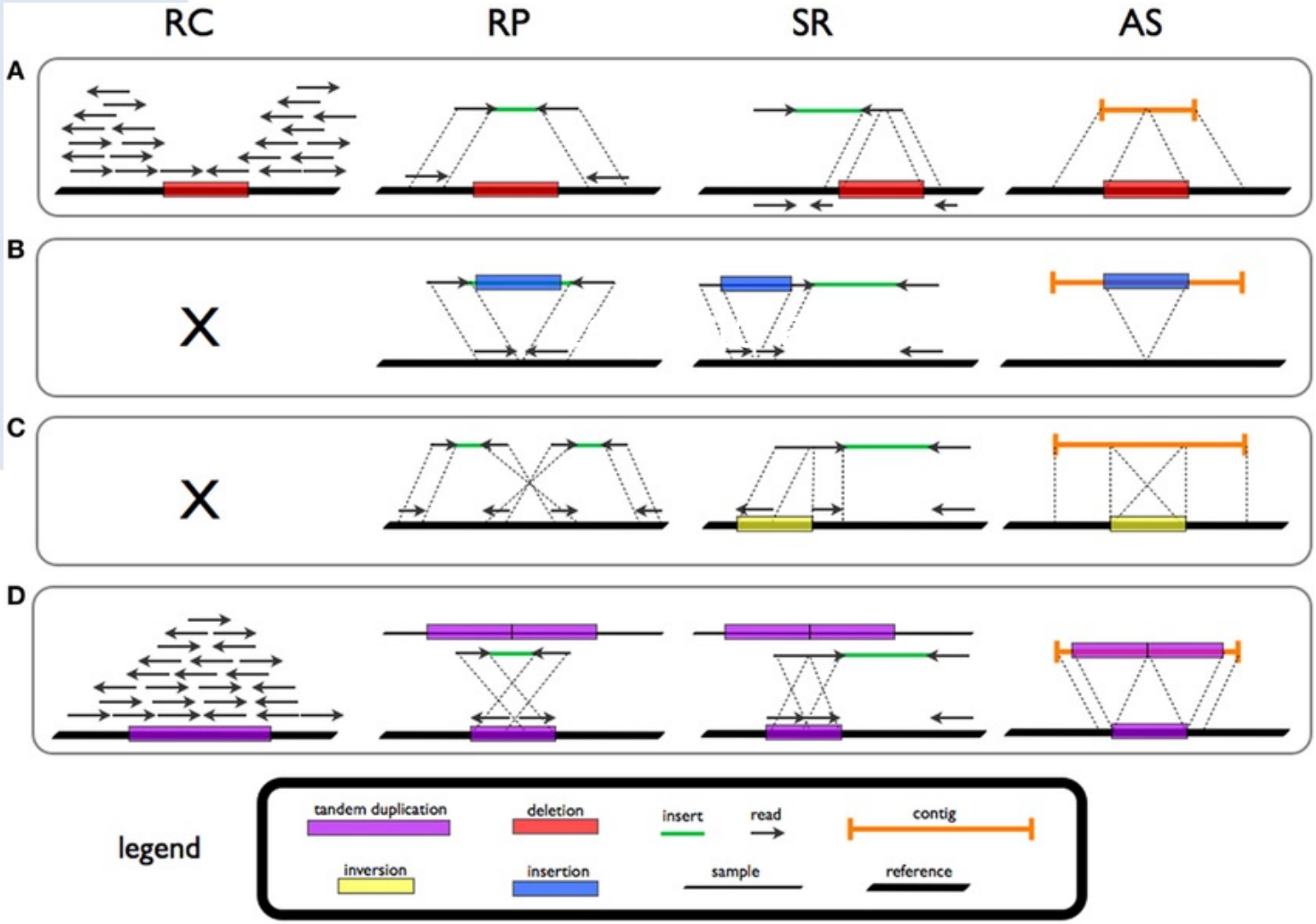
Transcriptome Sequencing

DE-seq analysis pipeline



Nadia Pisanti

Detecting Structural Variants



SEQUENCE COMPARISON

- Sequences comparison is used as a basic toolkit in many applications of bioinformatics.
- The goal is to measure how much (and how) two sequences are similar.
- There is a huge literature dating back before bioinformatics.
- In bioinformatics, allowing insertions and deletions, the problem becomes what is known as sequence alignments
- Dynamic Programming methods are a nice sample of a sophisticated algorithmic contribution to bioinformatics

SEQUENCE ALIGNMENTS

Similarity of X=GAATTCAGTTA e Y=GGATCGA ?

EDIT DISTANCE:

The minimum number – say k - of edit operations such as:

- replacing a letter
- deleting a letter
- inserting a letter

that you need to turn X into Y.

The smallest is k, the more similar X and Y are!

How to compute k?

Why "alignment"

GAATTCAGTTA
GGAT_C_G_A_

6 edits

GAATTCAGTTA
G_GA_TC_G_A

9 edits

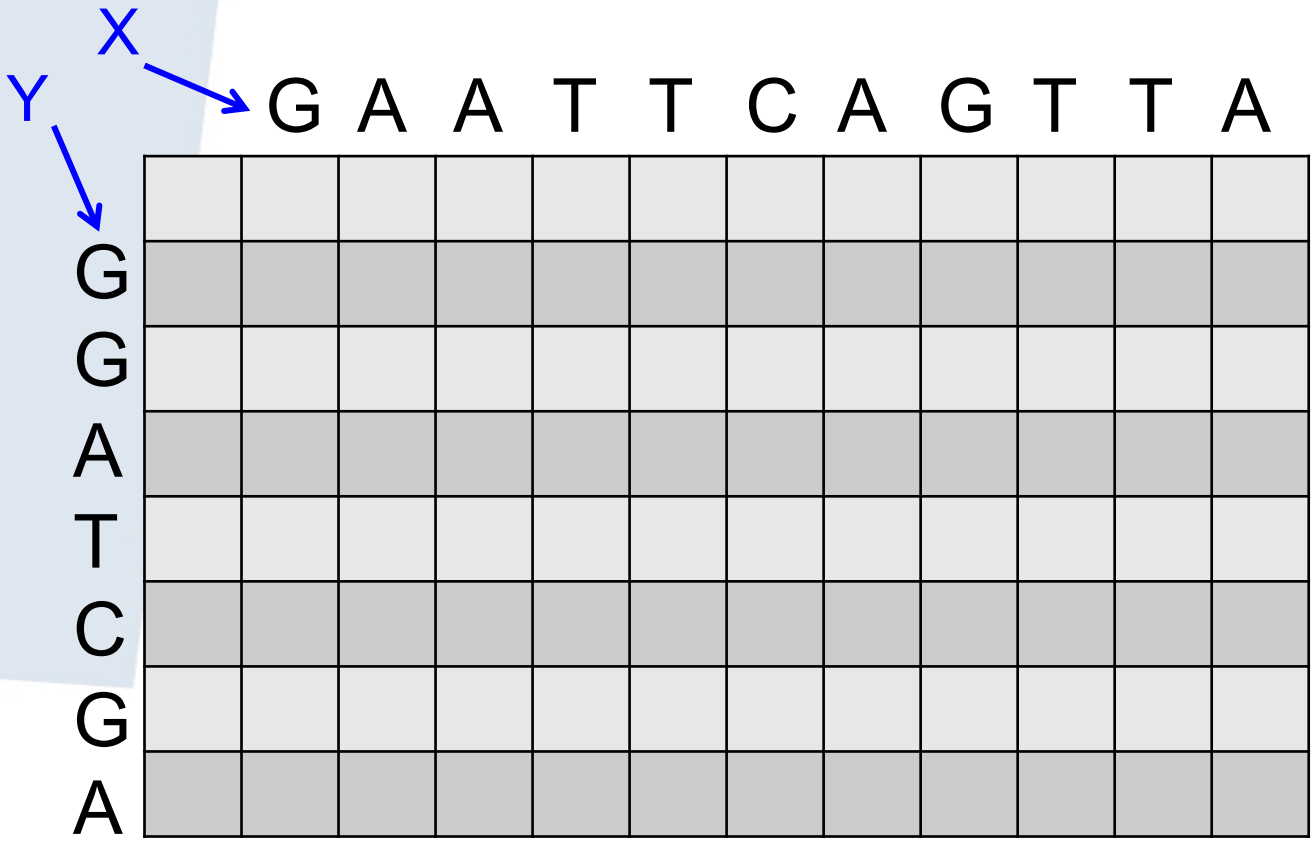
Enumerating all possible ways to *align* letters and gaps _

and pick the one with minimum number of edits...

.. is computationally untractable.

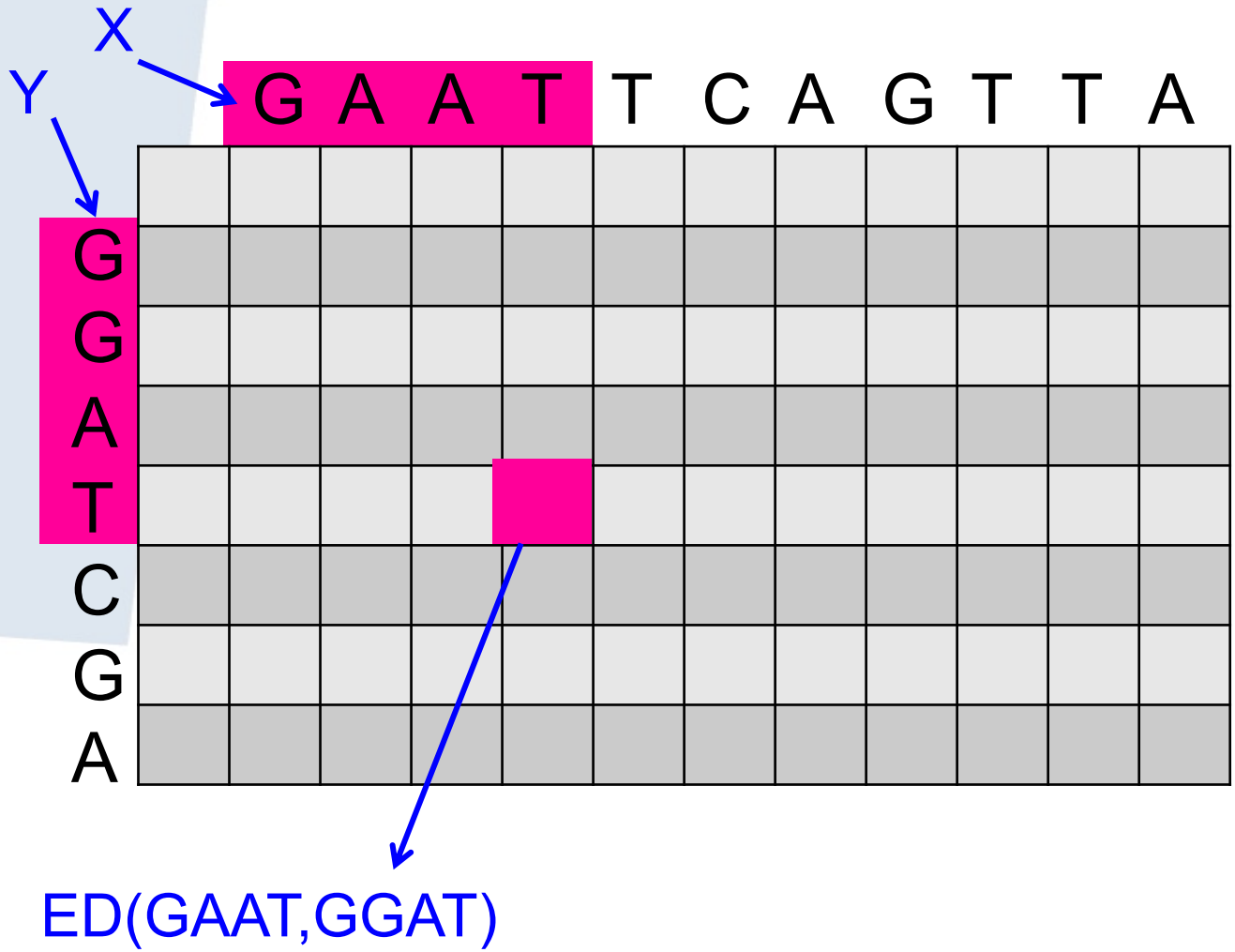
Nadia Pisanti

DYNAMIC PROGRAMMING

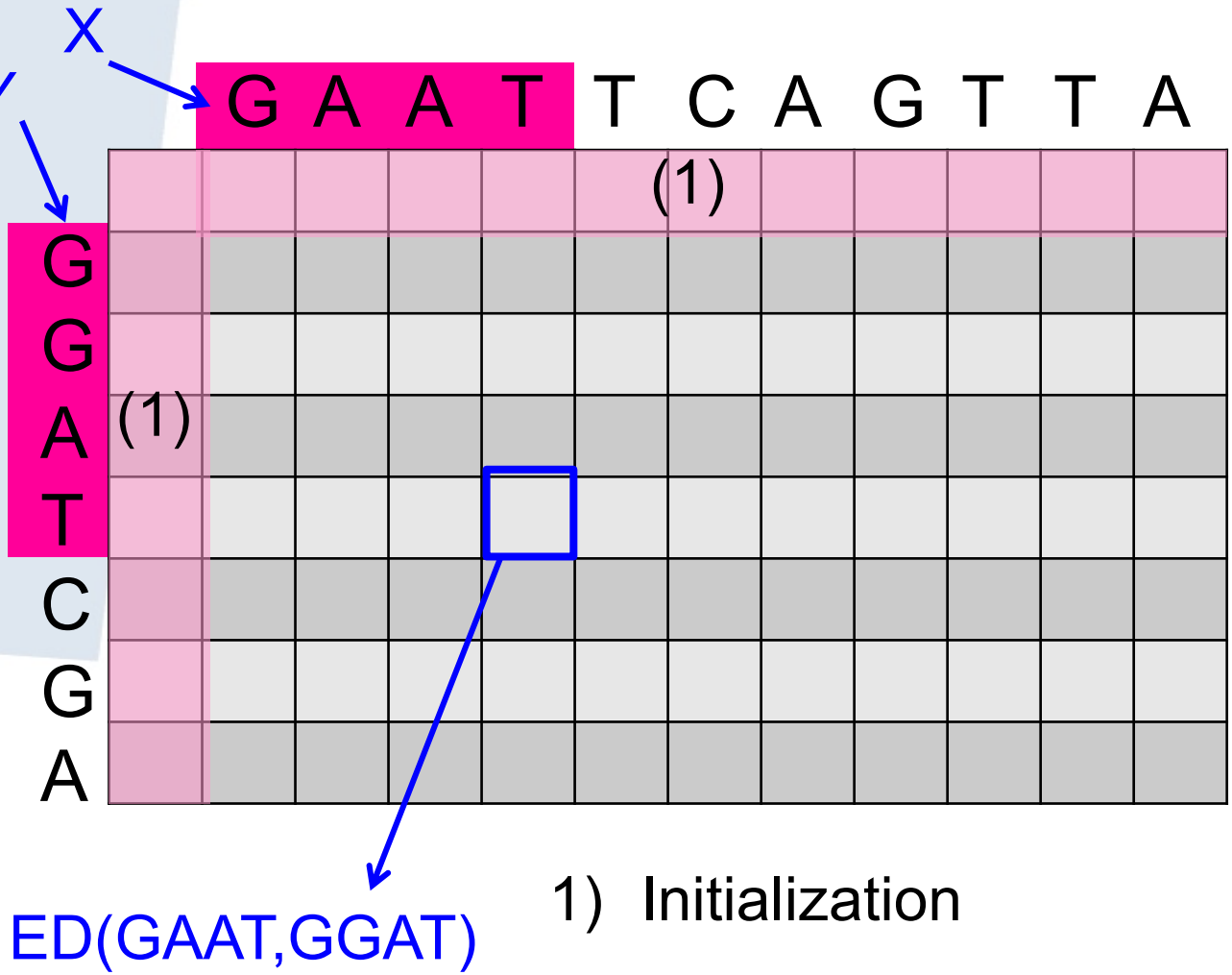


Nadia Pisanti

DYNAMIC PROGRAMMING

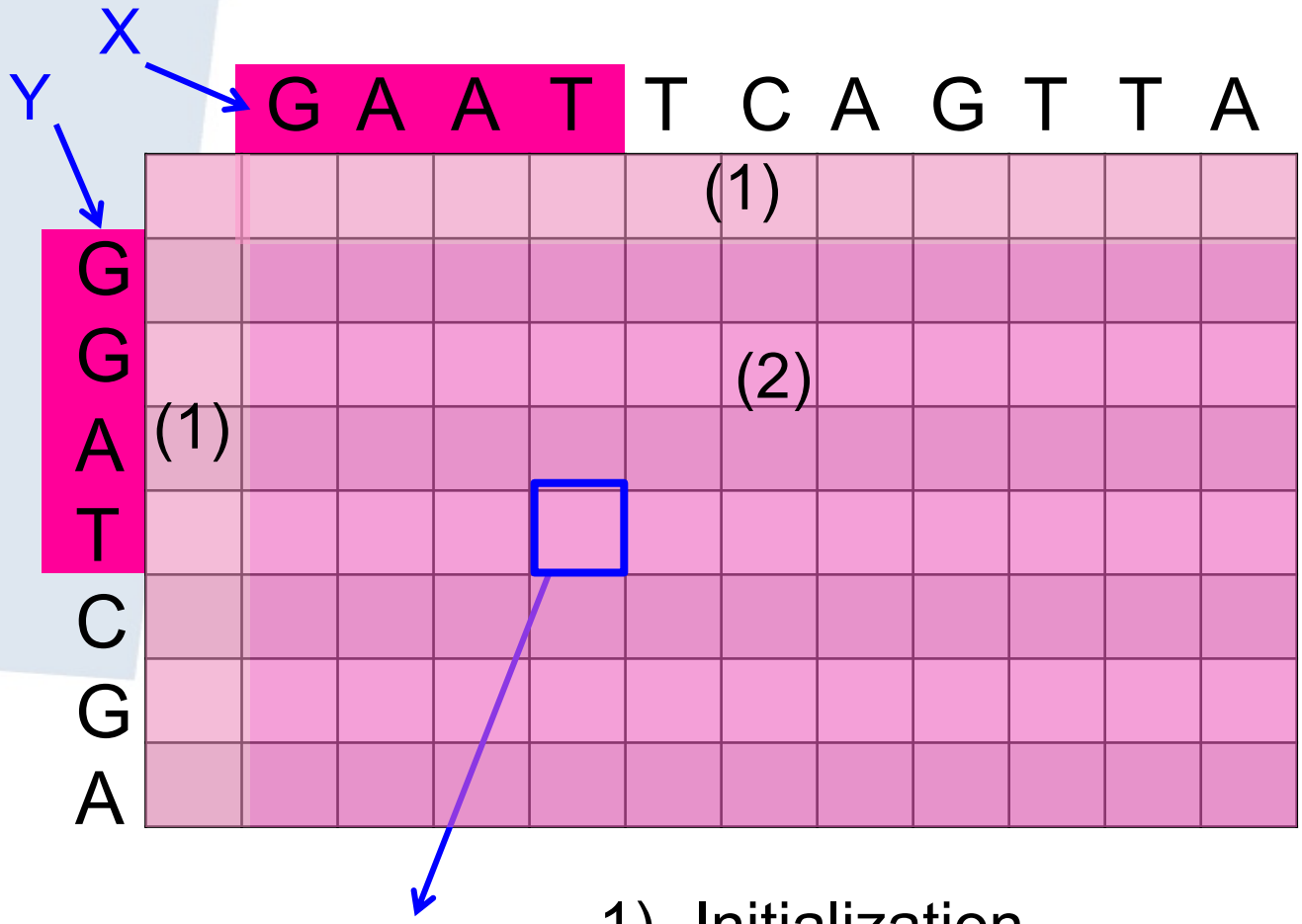


DYNAMIC PROGRAMMING



1) Initialization

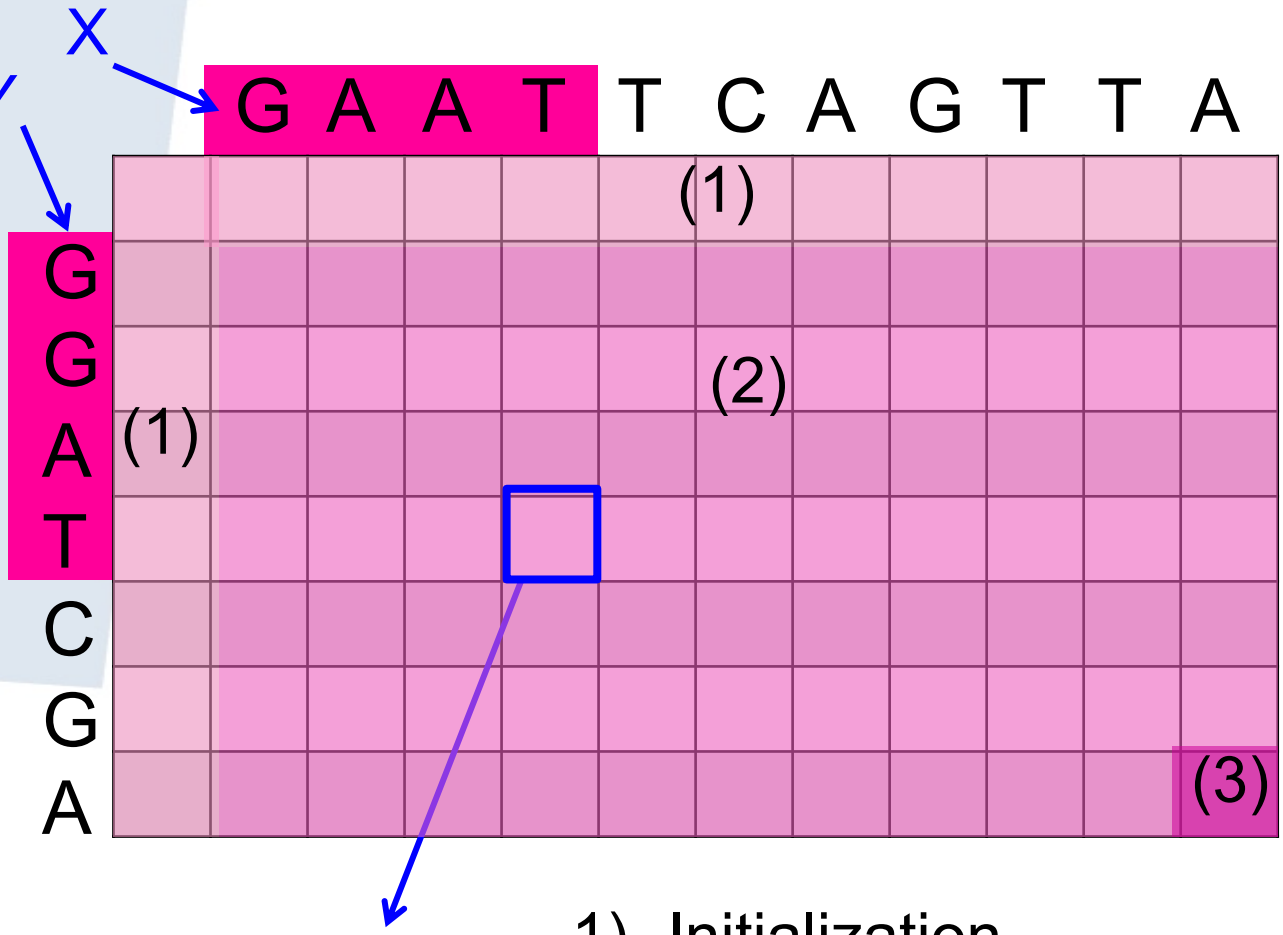
DYNAMIC PROGRAMMING



ED(GAAT,GGAT)

1) Initialization
2) Each matrix entry takes O(1)

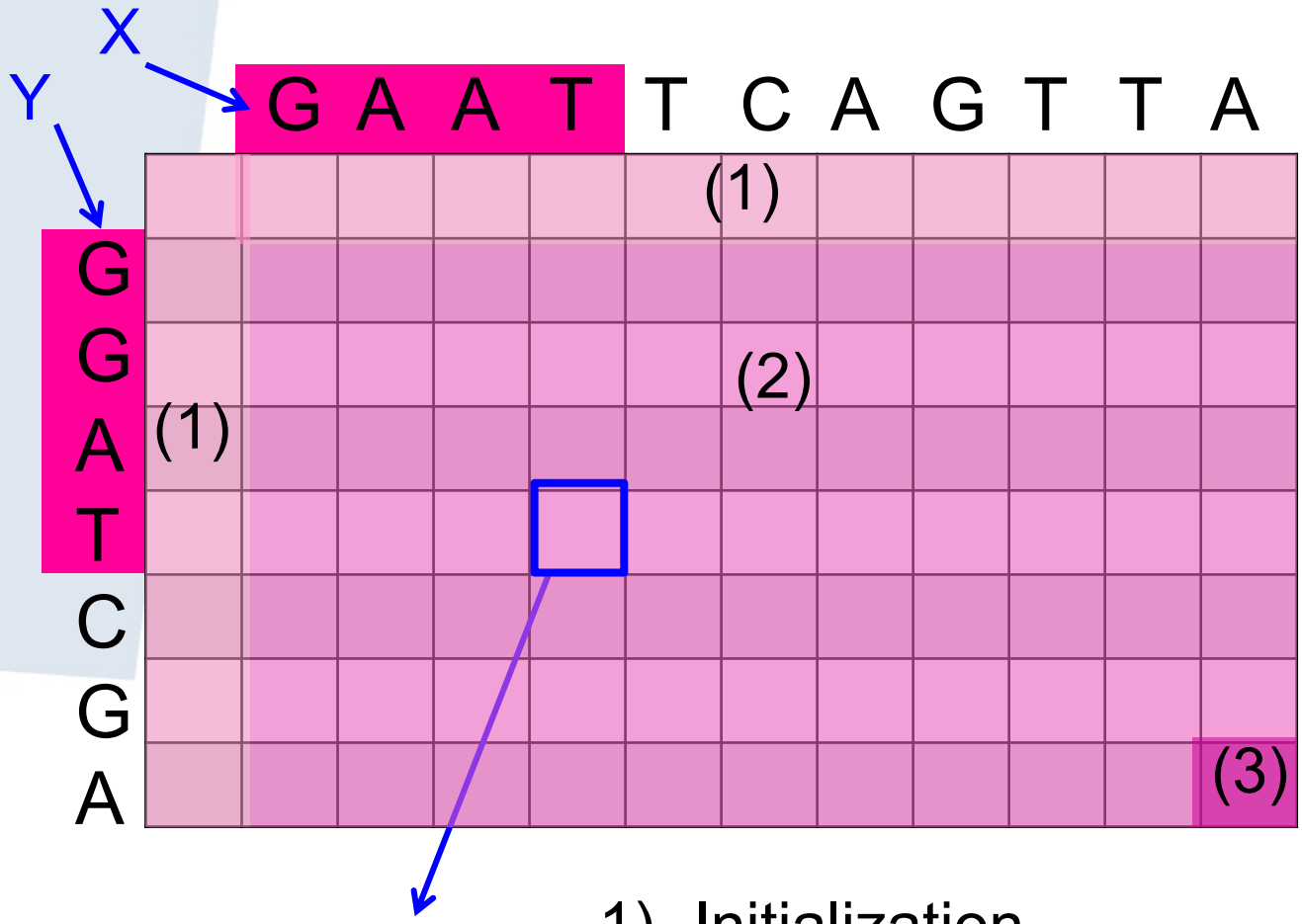
DYNAMIC PROGRAMMING



ED(GAAT,GGAT)

1) Initialization
2) Each matrix entry takes O(1)
3) Result in bottom-right entry

DYNAMIC PROGRAMMING



ED(GAAT,GGAT)

O(|X|*|Y|)
time & space

1) Initialization
2) Each matrix entry takes O(1)
3) Result in bottom-right entry

Nadia Pisanti

THANK YOU!