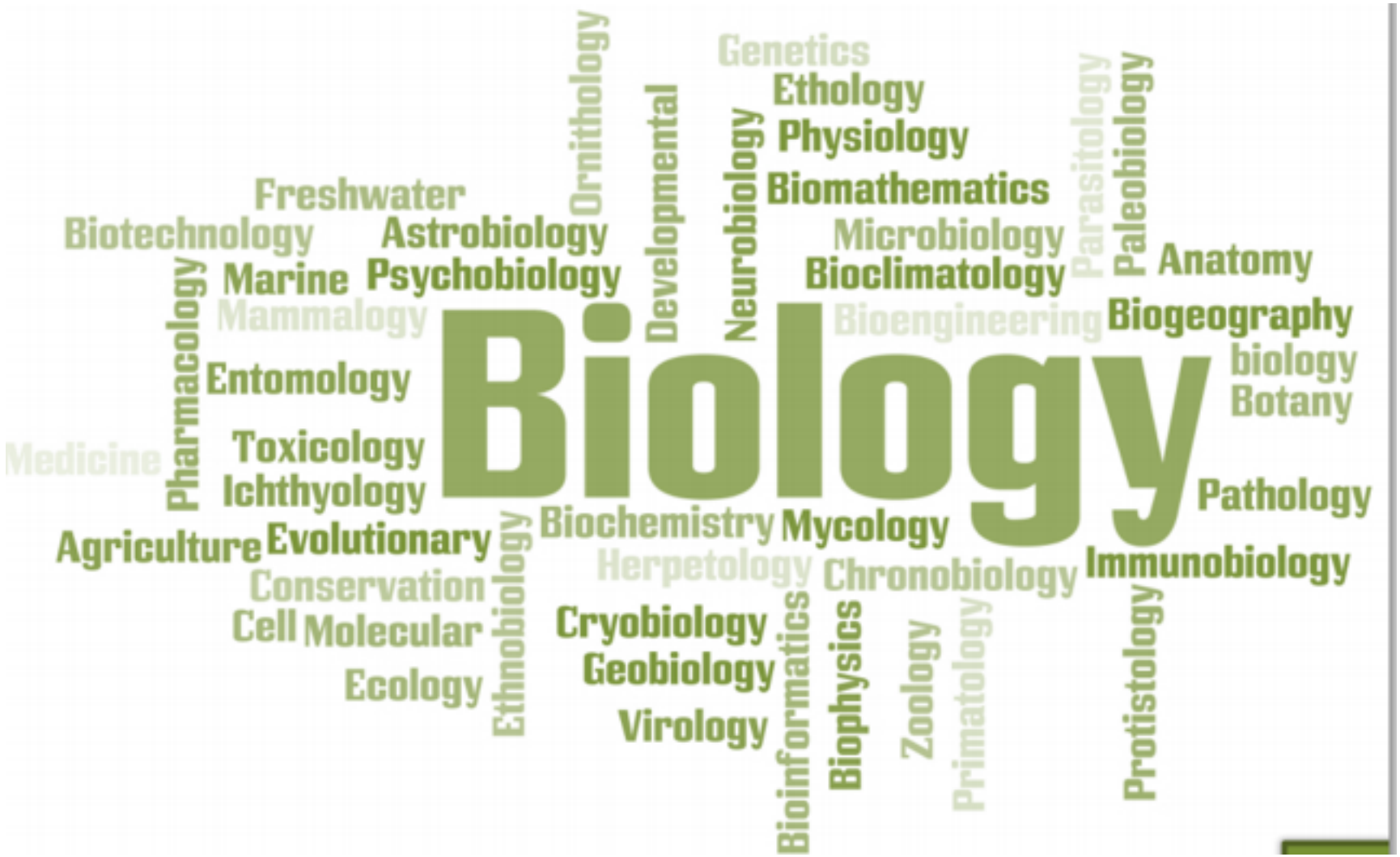
E agora para onde posso ir ?

Marcel Caraciolo, CTO
marcel@genomika.com.br

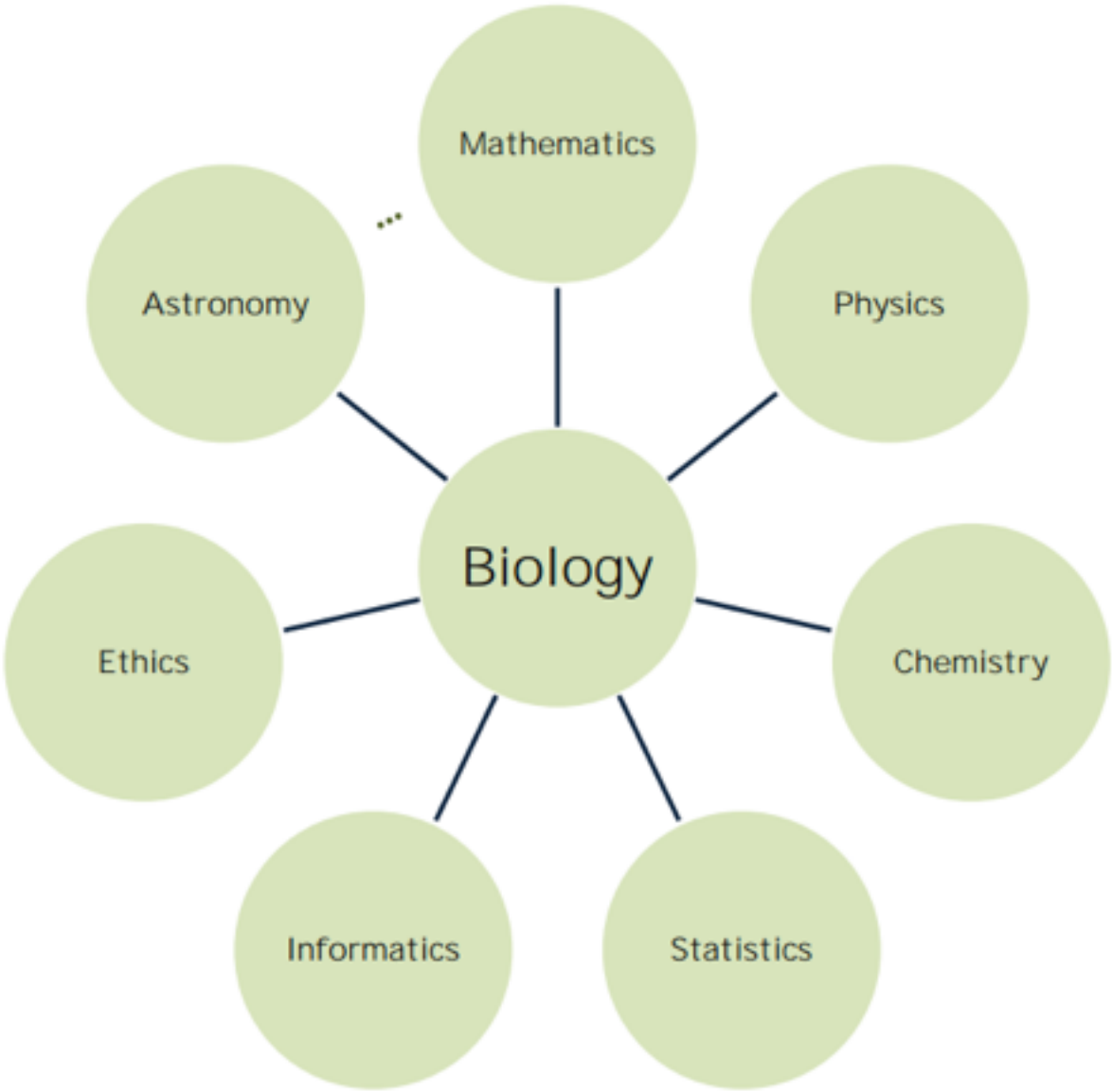
O que é bioinformática ?

Biology

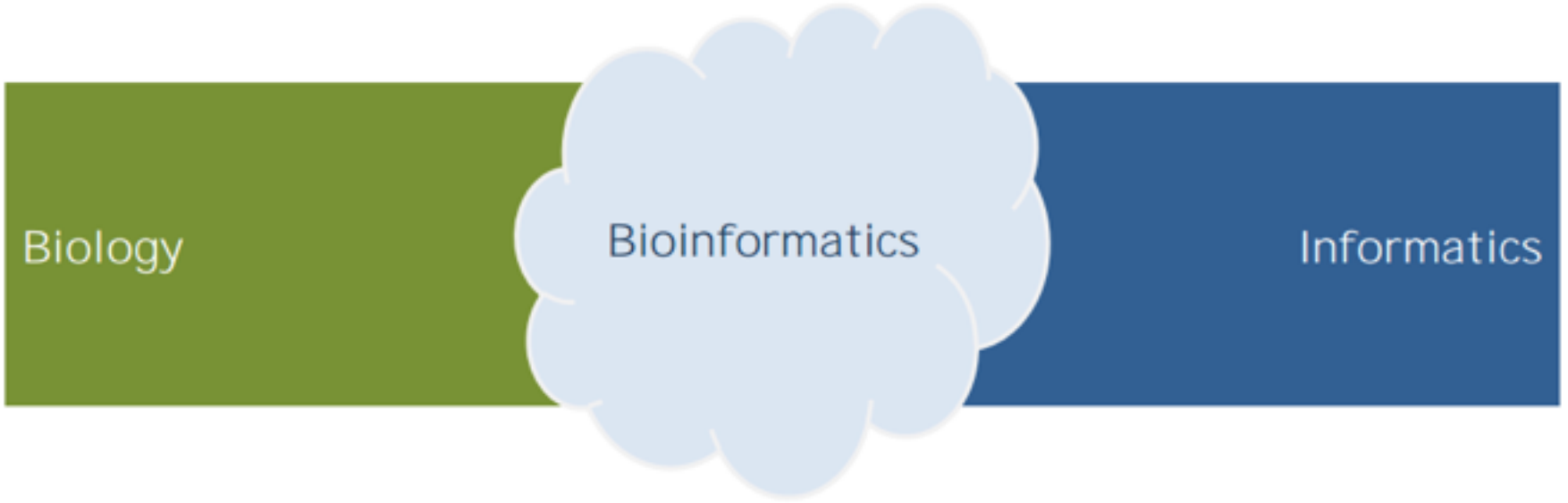
O que é bioinformática ?



O que é bioinformática ?

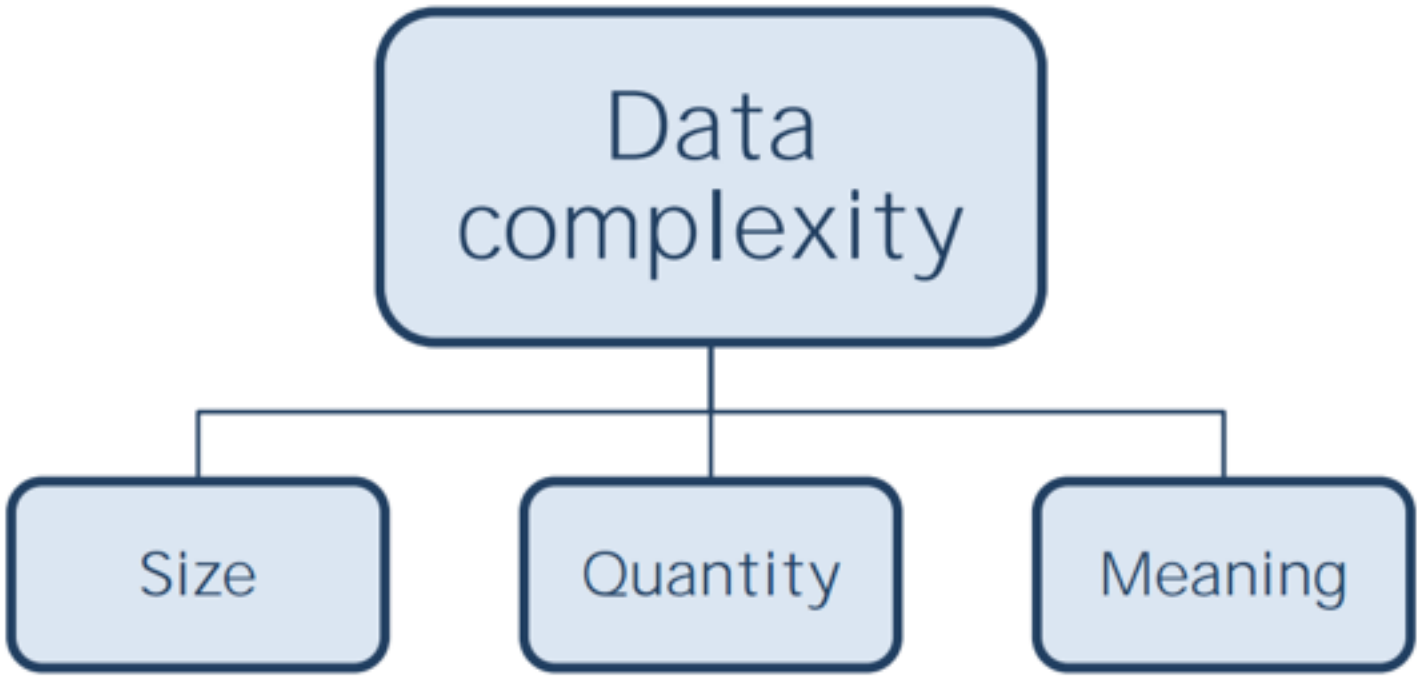


O que é bioinformática ?

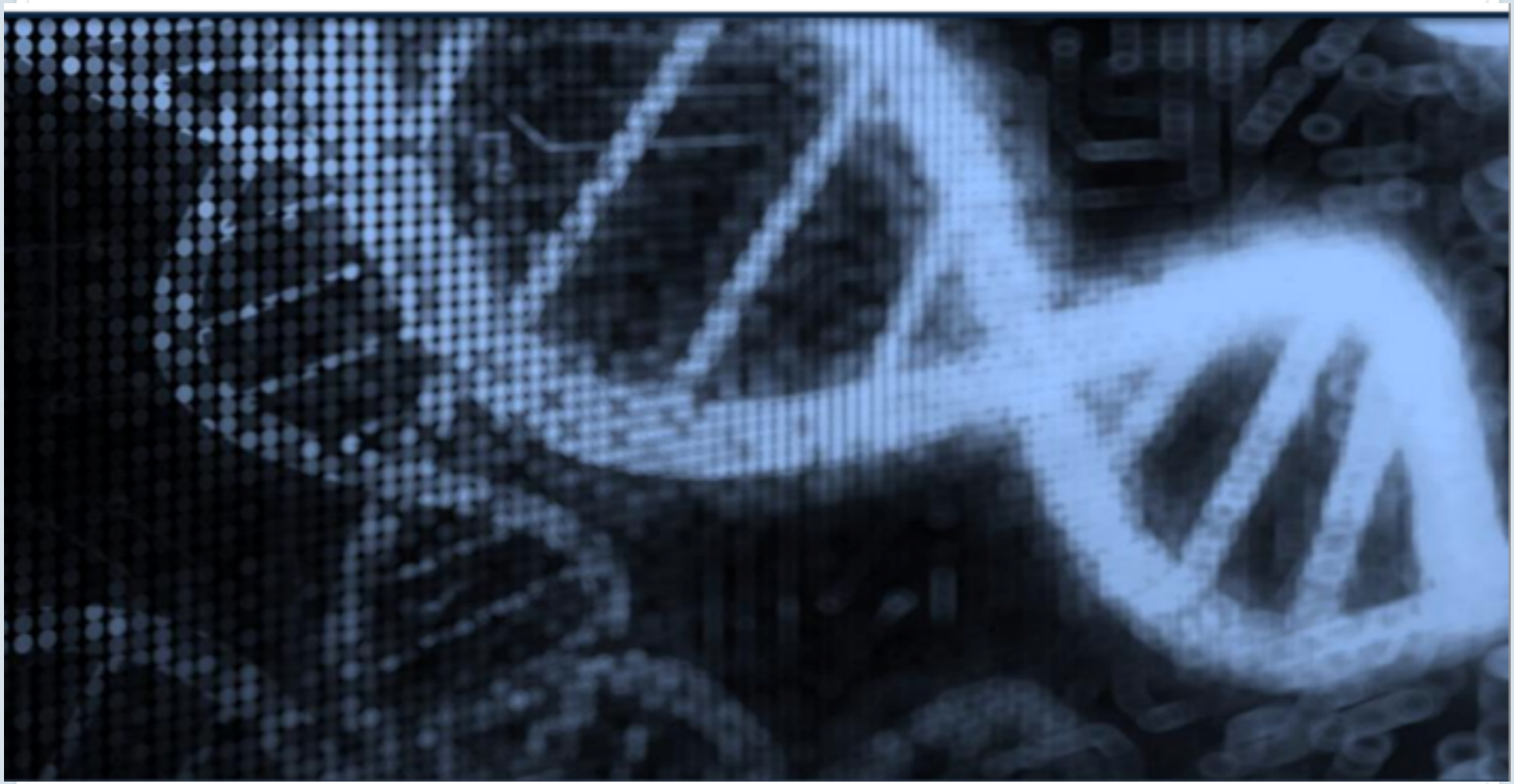


Bioinformatics is an informatics discipline for storing, retrieving, organizing and analyzing biological data.

Mas por que é tão diferente ?



Tamanho



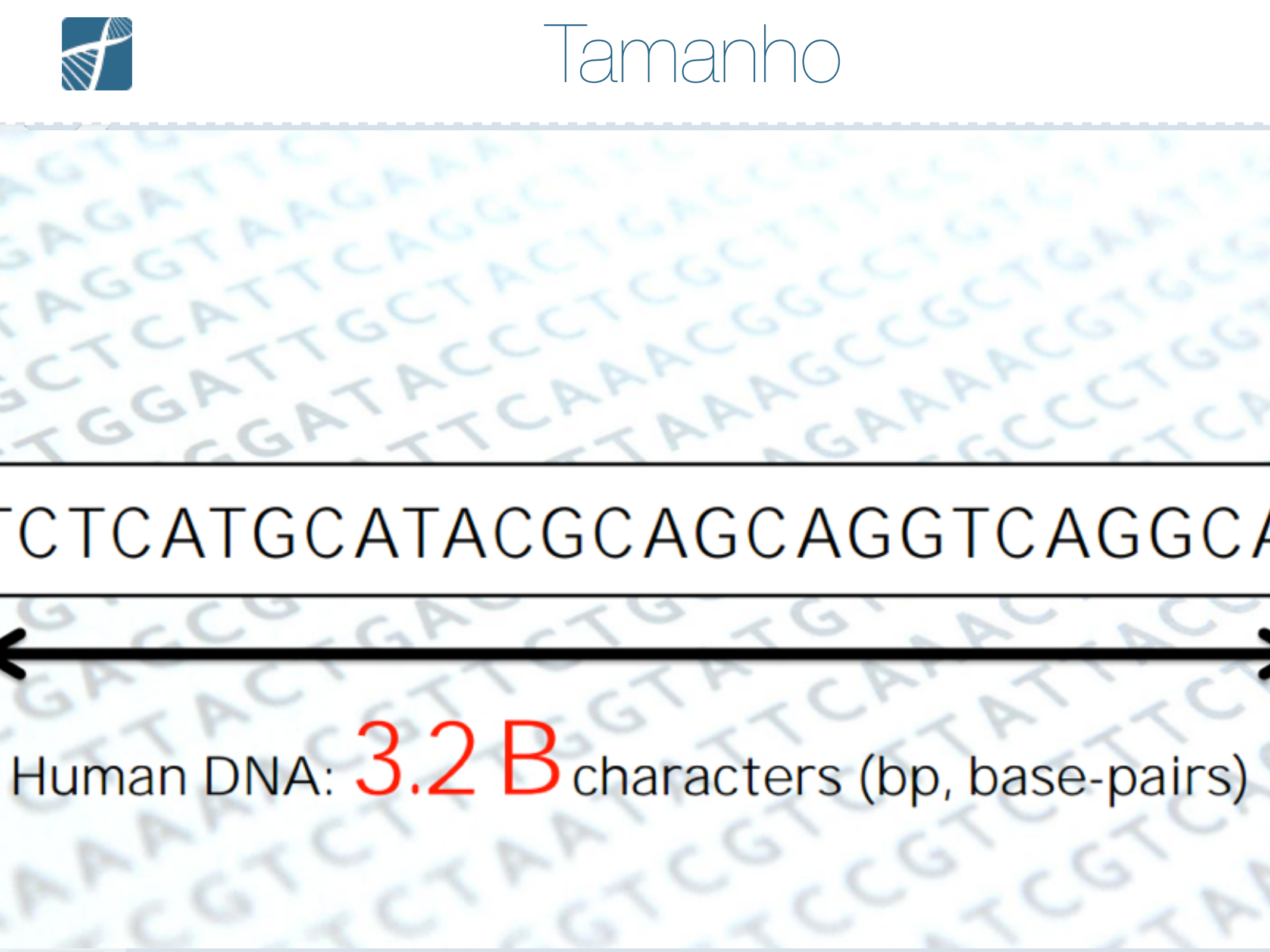
Tamanho

TCTCATGCATACGCAGCAGGTCAGGCA

DNA is:

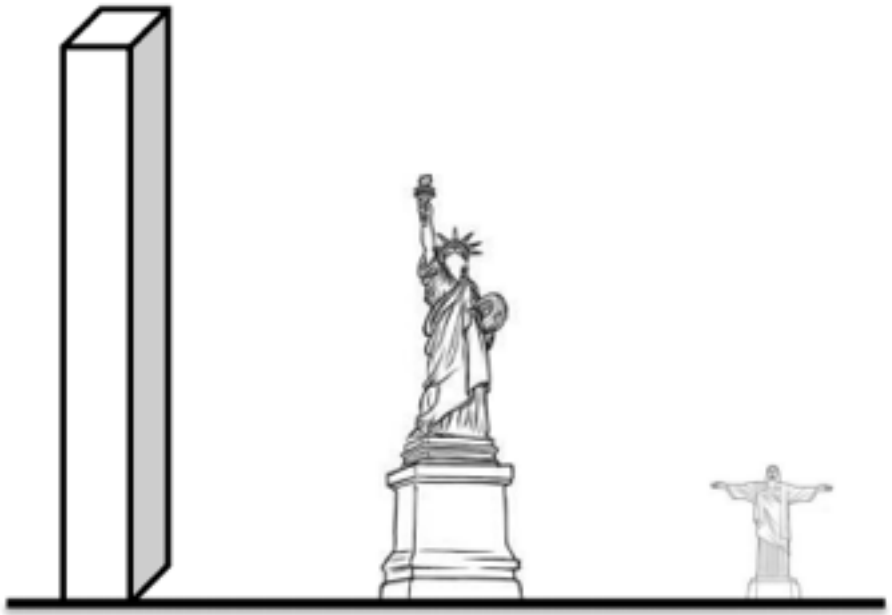
- A very large string
- Composed only by A, C, G, and T characters

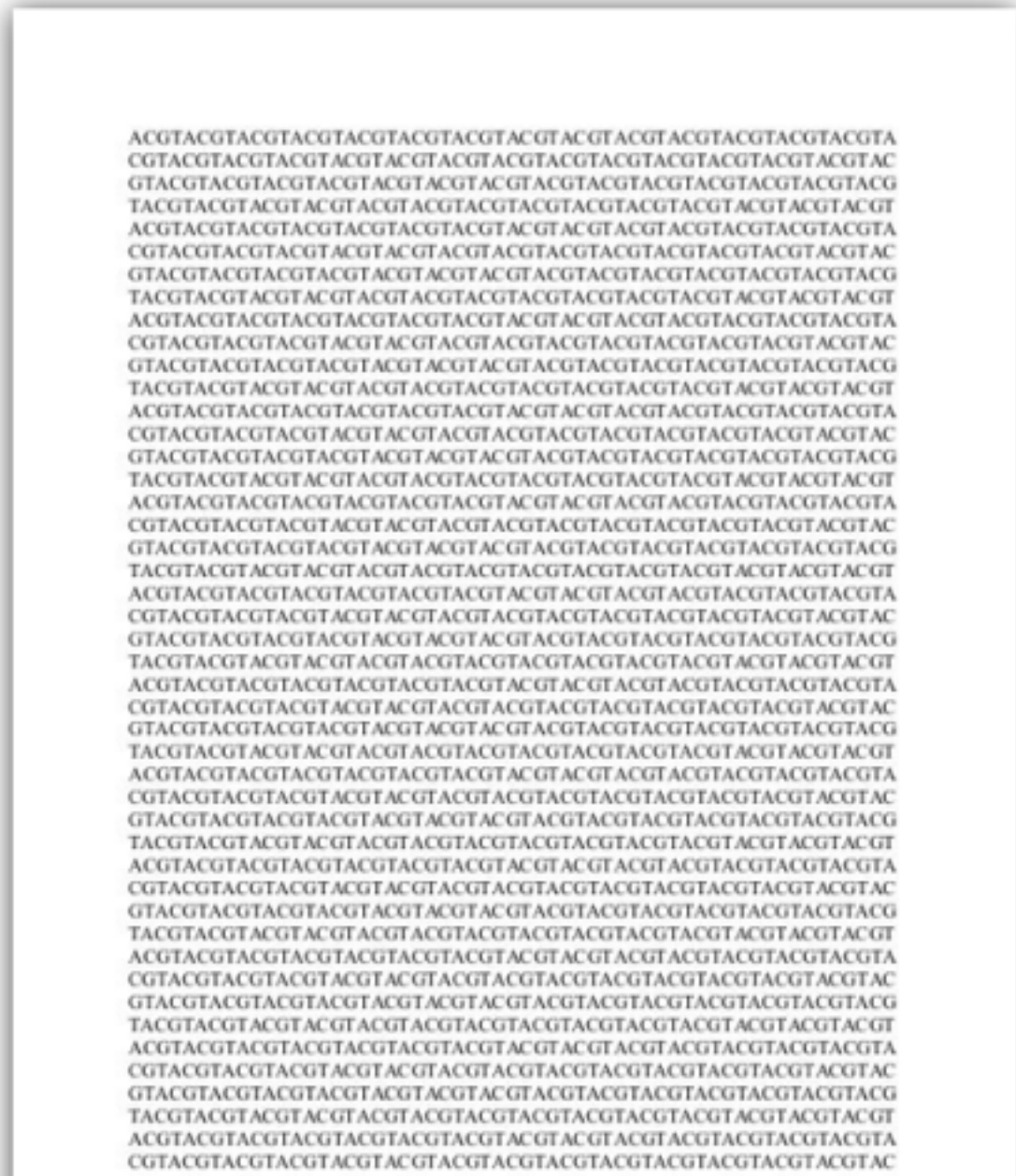
Tamanho


TCTCATGCATACGCAGCAGGTCAGGCA
Human DNA: 3.2 B characters (bp, base-pairs)

Tamanho

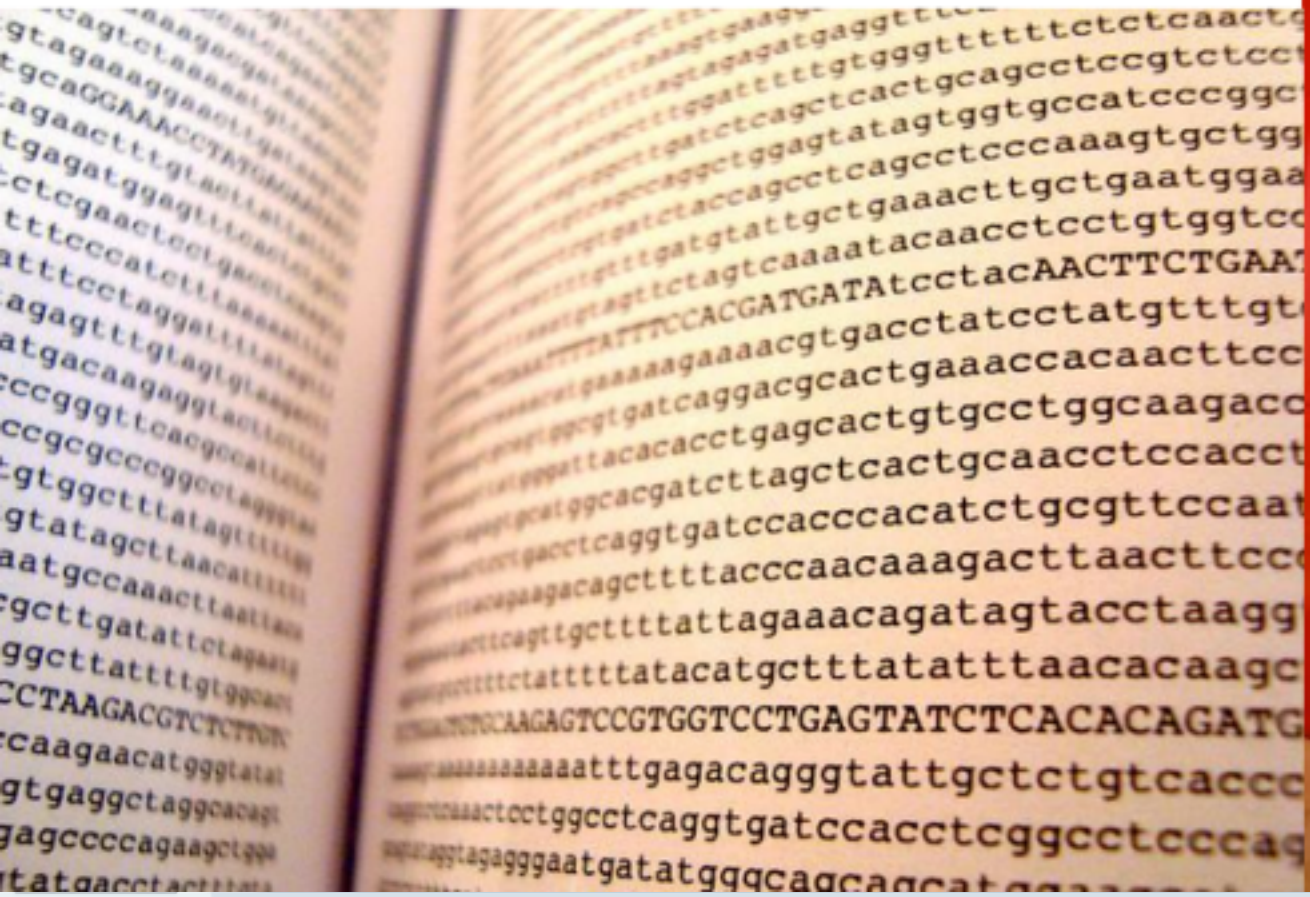
Times New Roman
12 pt
2622 bp per page
1.2 M single-sided pages
2415 reams (500 sheets)
129 meters





Tamanho

- 4.5 pt
- Double-sided pages
- 120 book volumes
- 1 bookshelf



Volume

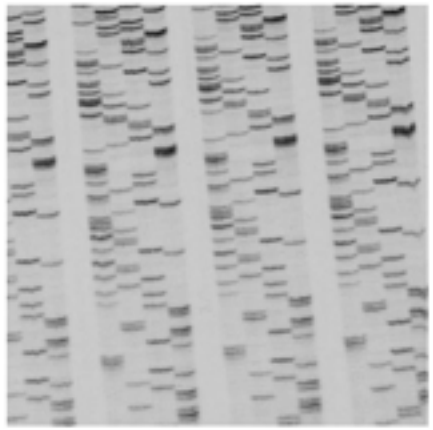
Volume

The number of sequenced genomes is exponentially increasing

Caused by the decreasing on: time and cost

	1st Human Genome	Currently
Year	1988	2014
Cost	US$ 3 billion	US$ 1 000
Time	13 years (2001)	< 1 hour

Volume

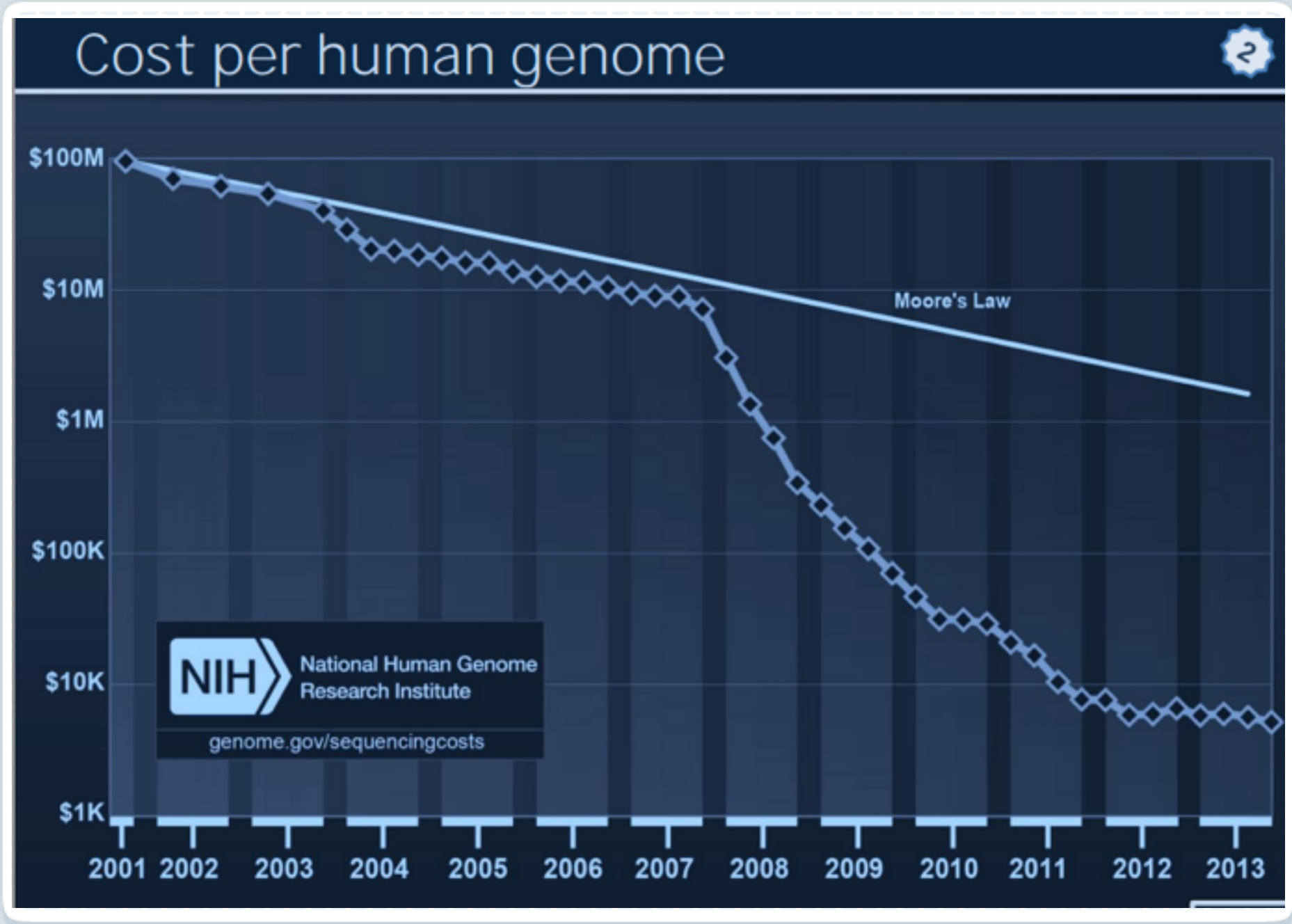

Sanger-based chemistries and capillary-based instruments
"First generation" sequencing platforms

↓

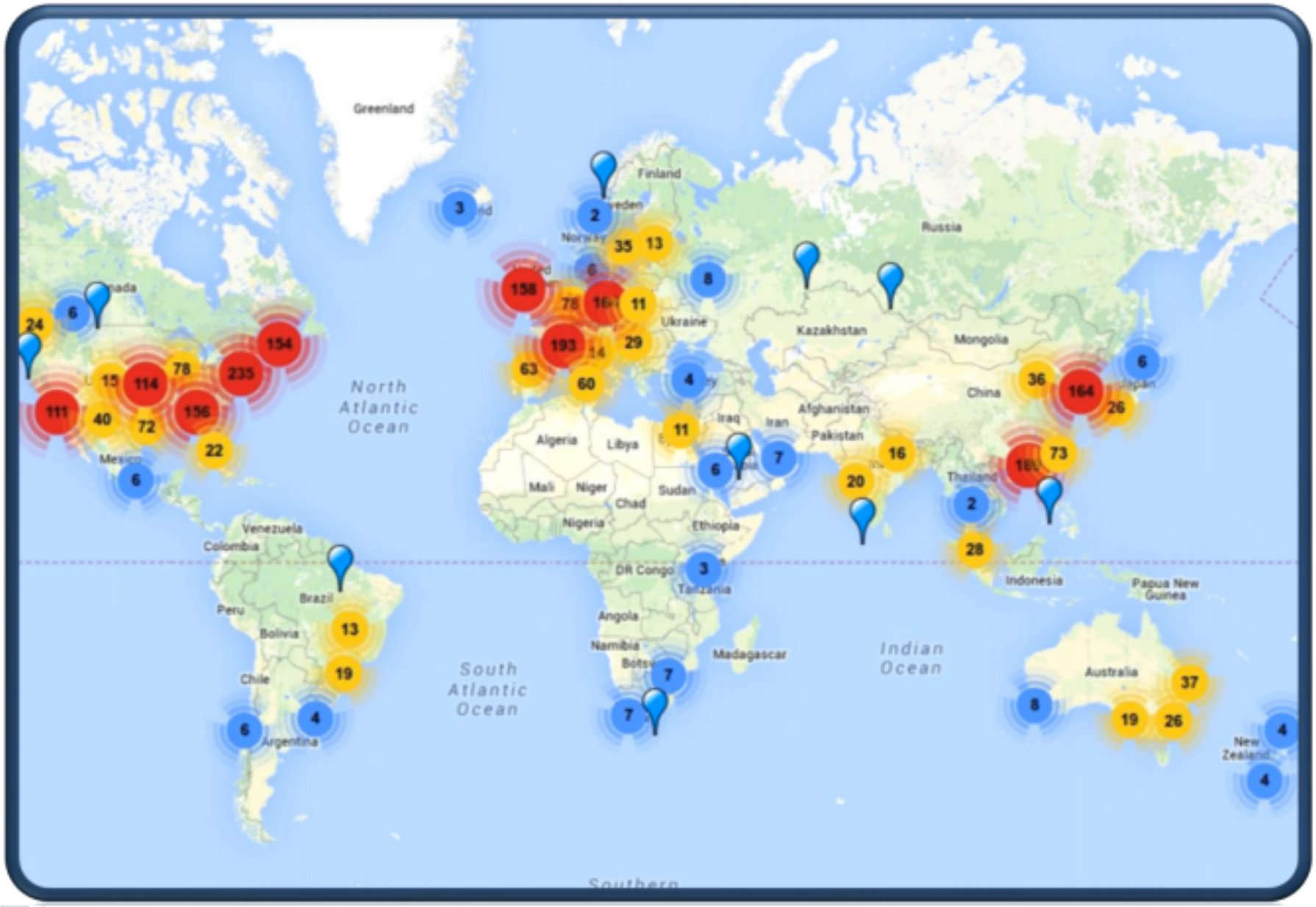
Next-Generation sequencing platforms (NGS)




Volume


Cost per human genome
2
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Volume


Greenland
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Finland
Russia
Ukraine
Kazakhstan
Mongolia
China
Afghanistan
Pakistan
Iran
Iraq
Algeria
Libya
Mali
Niger
Sudan
Chad
Nigeria
Ethiopia
DR Congo
Tanzania
Angola
Namibia
Madagascar
Venezuela
Colombia
Brazil
Peru
Bolivia
Chile
Argentina
Mexico
Thailand
Indonesia
Papua New Guinea
Australia
New Zealand
Southern
3
2
35
13
6
158
78
11
8
193
14
29
63
60
4
11
6
7
24
6
154
78
15
114
235
111
40
72
156
22
6
16
20
2
28
36
164
26
6
73
13
19
4
6
3
7
7
8
37
19
26
4
4

Volume

Complete genomics
A BGI Company

NCBI GenBank

UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

GENOME 10K

wellcome trust sanger institute

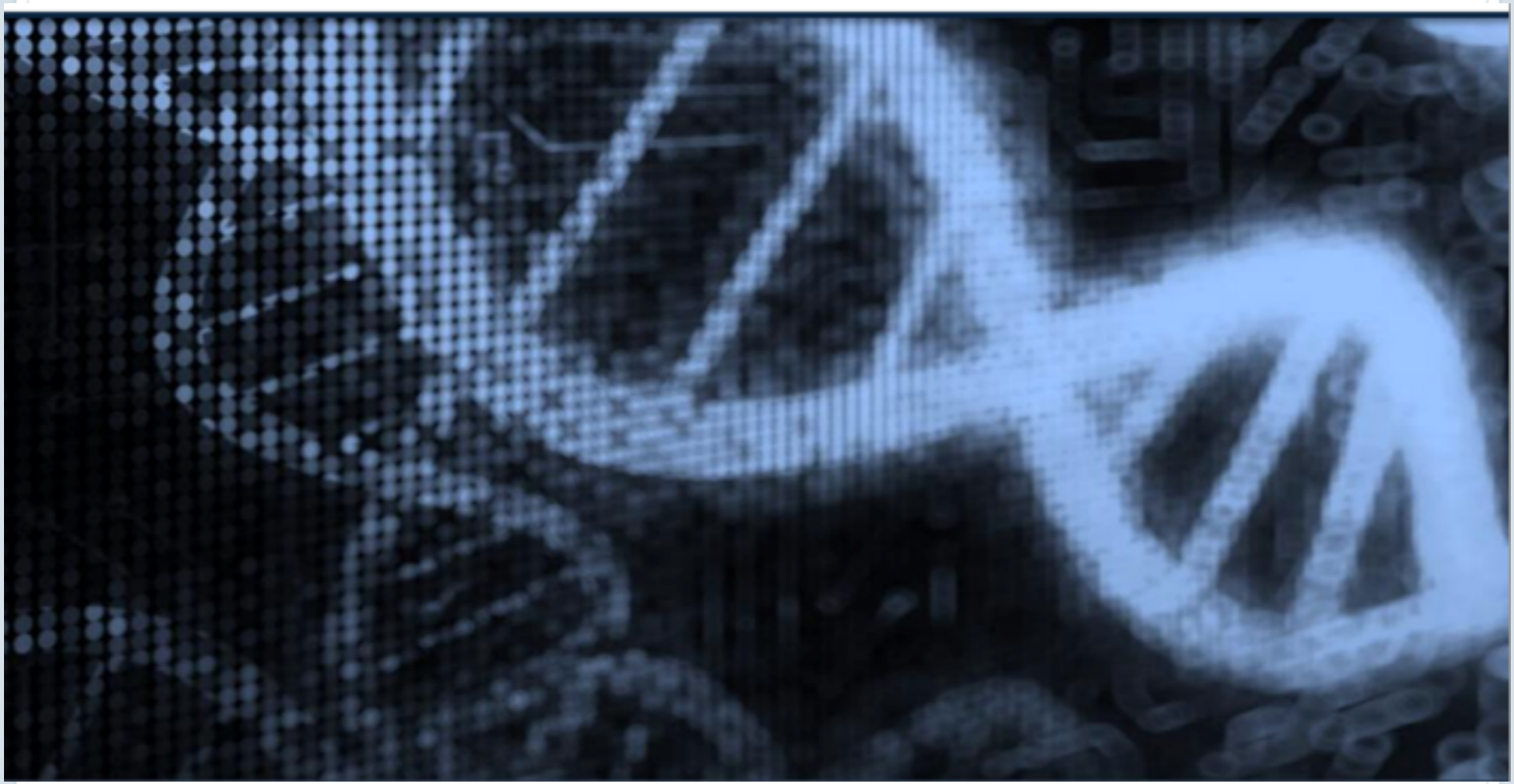
Personal Genome Project

Volume

How much disk space do we need to store all genomes of an entire population?

	Population	Compressed	Size with 2-bits	Size with 8-bits	WGS
1-person	1	7 MB	770 MB	3 GB	180 GB
Santa Maria, RS	270 K	1.8 TB	200 TB	790 TB	46 PB
Porto Alegre, RS	1.5 M	10 TB	1.1 PB	4.3 PB	260 PB
Rio Grande do Sul	10.7 M	72 TB	7.7 PB	30 PB	1.8 EB
São Paulo, SP	11.3 M	75 TB	8.1 PB	32 PB	1.9 EB
Brazil	201 M	1.3 PB	144 PB	575 PB	34 EB
Latin America	589 M	3.8 PB	422 PB	1.6 EB	100 EB
World	7 B	45 PB	5 EB	20 EB	1.1 ZB

Significado



Significado

GATTCGTACTGACTACGCAGCAG

What does it do?

What does it produces?

How does it looks like?

Is it related with a specific disease?

Does it interact with other components?

Does it interact with some medicine?

Significado

CATTCATCATGCATACGCAGCAGGT

What does it mean ...

... to me, as an individual? (Personalized medicine)

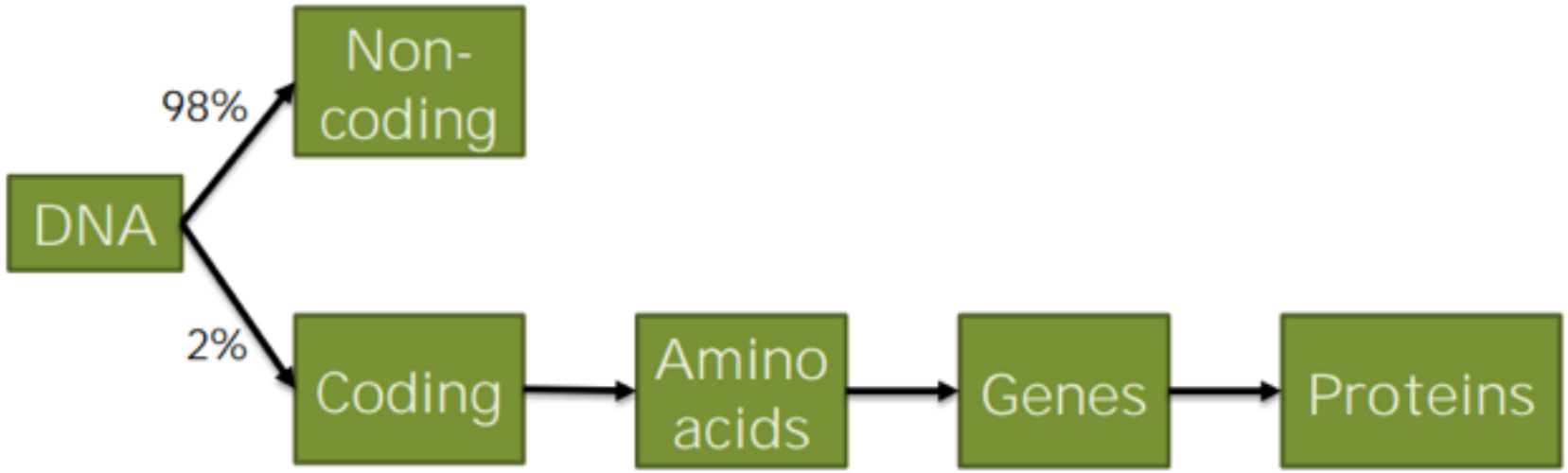
... to my population? (Public-health genomics)

... to human kind? (Science)

Significado

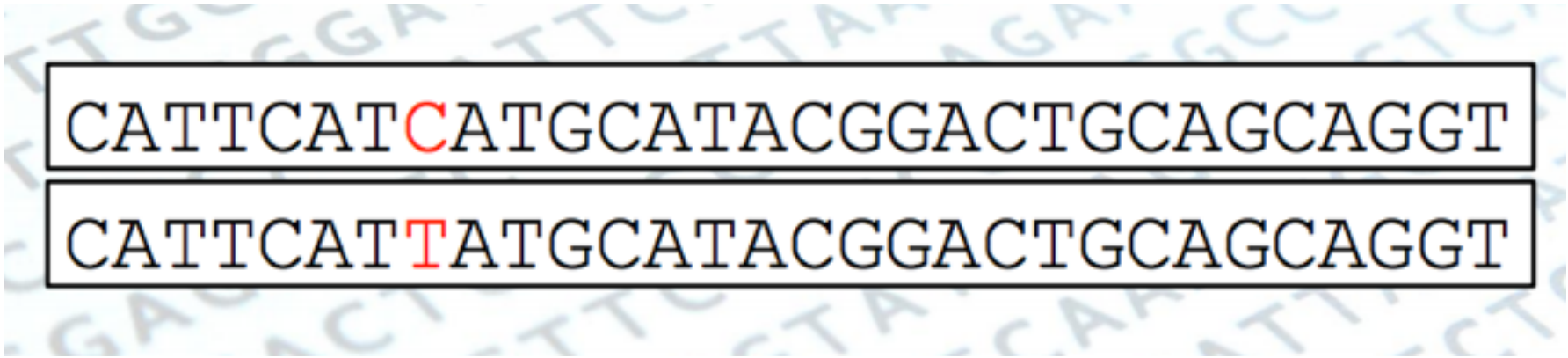
- Lowest structure present in all living organisms
- High expectation for medicine
- DNA cannot tell everything about your future
- DNA is not the only variable causing diseases
- Behavior and environment interfere
- But, DNA still plays an important role

Significado



Significado

Significado



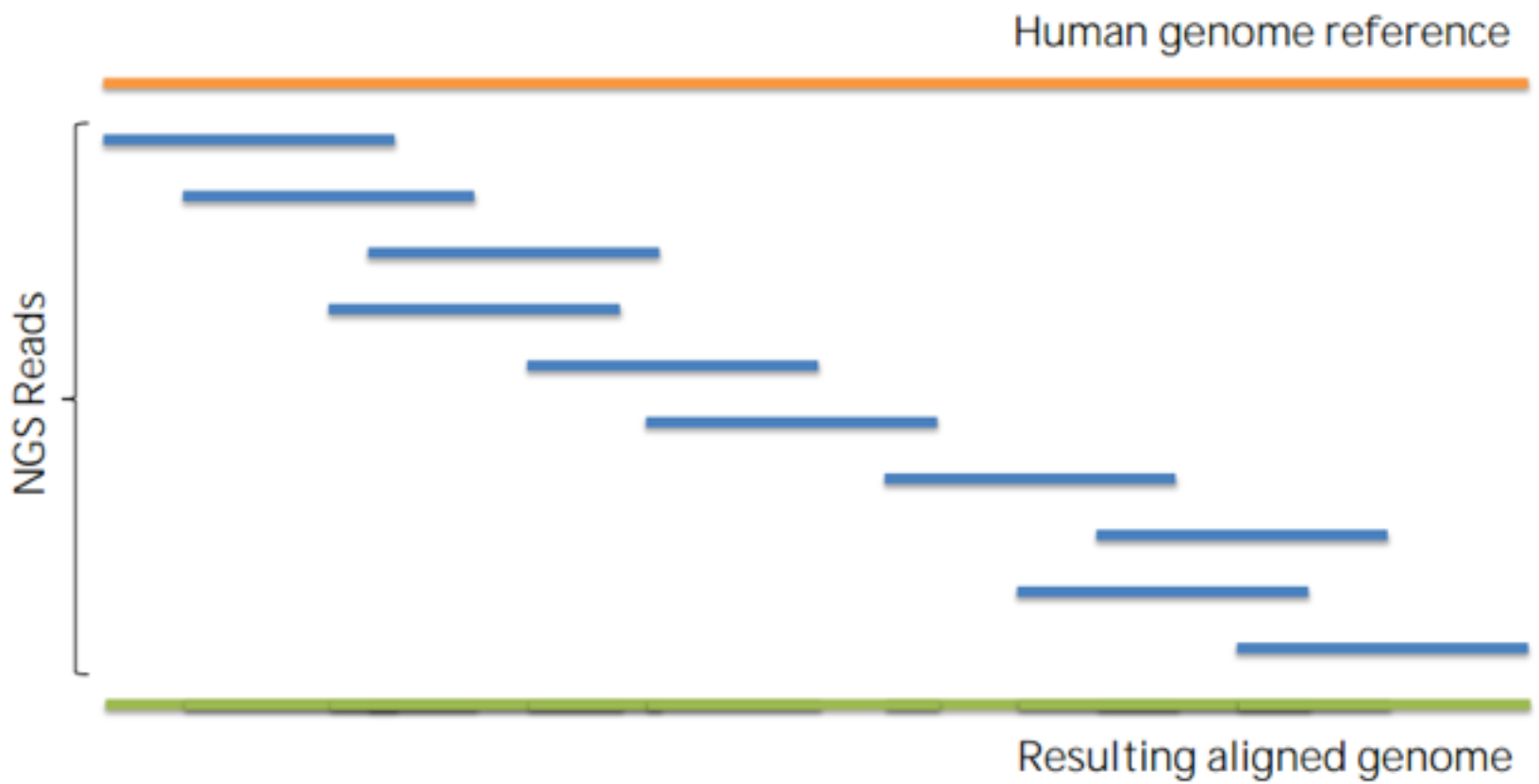
Significado

- Comparison of two or more sequences
- Similarity, similarity and similarity
- Similar sequences may have similar genes, proteins, structures and functions
- Percent identity as metric
- Considering small variations (SNPs, insertions, deletions)

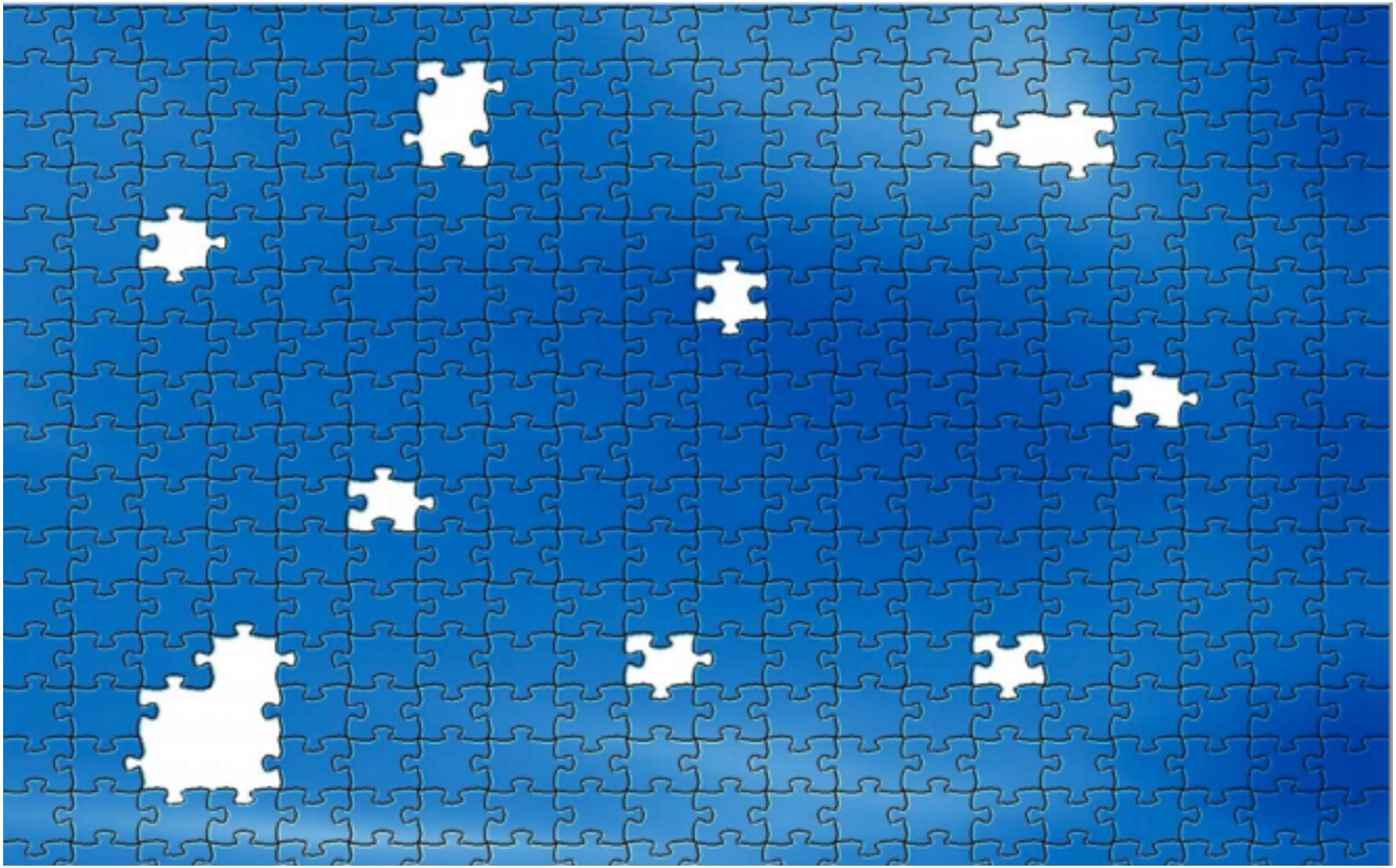
Search terms:

Hamming distance, Edit distance (Levenshtein), sequence alignment, Needleman-Wunsch algorithm, Smith and Waterman algorithm, substitution matrices, BLAST algorithm, seed-and-extend, suffix tree, homologue sequences

Significado



Significado



Significado

- Delimit parts of sequences that have a biological meaning
- Detect the begin and end of a gene
- Detect the begin and end of shapes

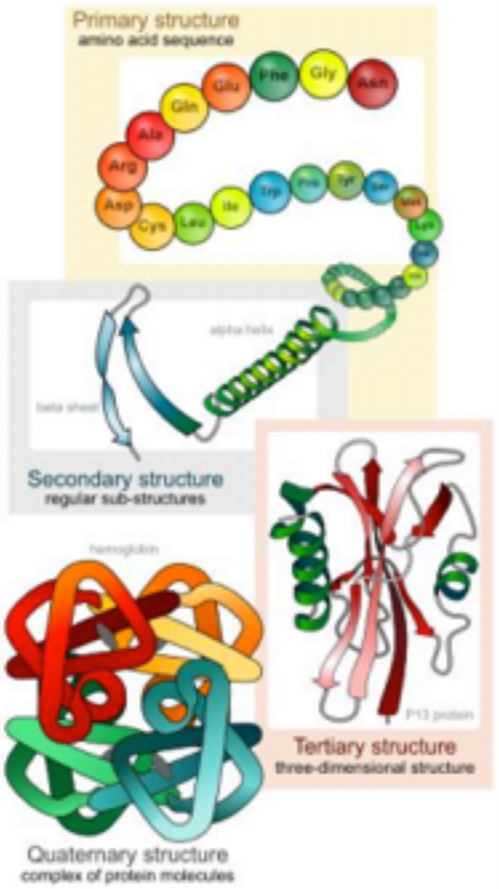
Search terms:

Conserved domains, motifs, regular expressions, Fuzzy regular expressions, Hidden Markov Models (HMMs), InterPro database, PROSITE, multiple alignments, grammars, PCR analysis

Significado

- Structure is related with function

- Structure:
  - Primary
  - Secondary
  - Tertiary
  - Quaternary



Search terms:

Conserved domains, motifs, folding, PDB, 3D transformations, protein alignment, intermolecular distances, intramolecular distance, RMSD, DALI, SSAP, graph theory, ab-initio methods, fold recognition, threading, neural networks, machine learning, support vector machines, random forests, lowest energy algorithms, ROSETTA, CASP challenge

Significado

- Genes interact with each other
- Non-coding regions of DNA are important for regulation
- Depending on geographic position of a cell and neighbors to turn on/off genes

Search terms:

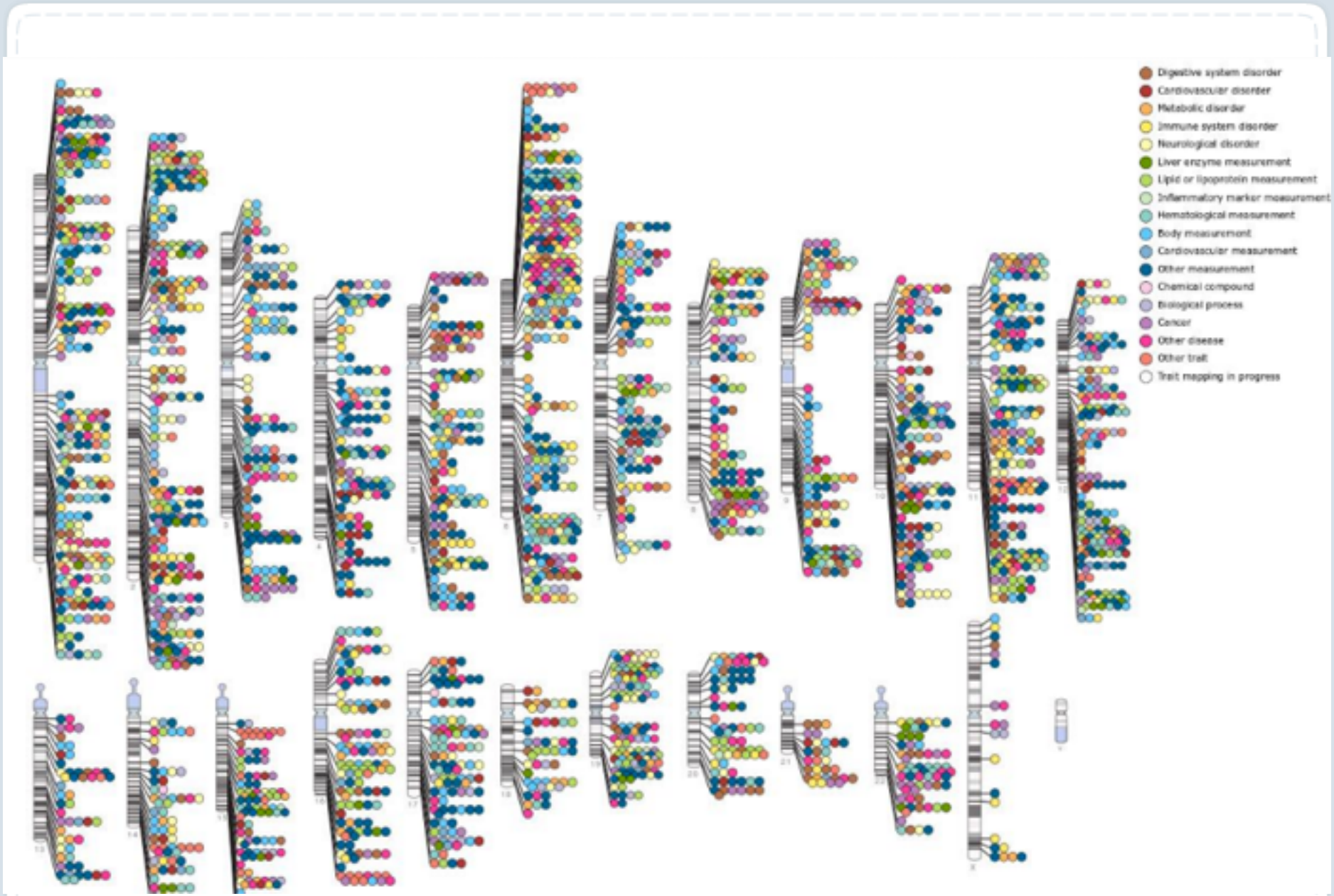
Cell regulation, junk DNA, turn on/off genes, pathways, Bayesian networks, Support vector machines, clustering algorithms, reducing dimensionality, simulation and modeling, discrete and continuous models, epigenomics

Significado

What does this sequence?
Is it related with Alzheimer's disease?
Does it define my hair color?

CATTACGGACGCATCATCATGCAGCAGGTG

Significado



Significado

- Most challenging topic
- Depend on all previous topics
- Human annotations for protein functions

Search terms:

Annotating DNA, ontologies, natural language processing (NLP), motifs, conserved domains, gene ontology, OBO, BioOntologies, UMLS, term-for-term analysis, data mining, directed acyclic graph, text mining, semantic web

Como estruturamos isto ?

Análise de Variantes

Assertion and evidence details Go to:

Clinical Assertions | Evidence

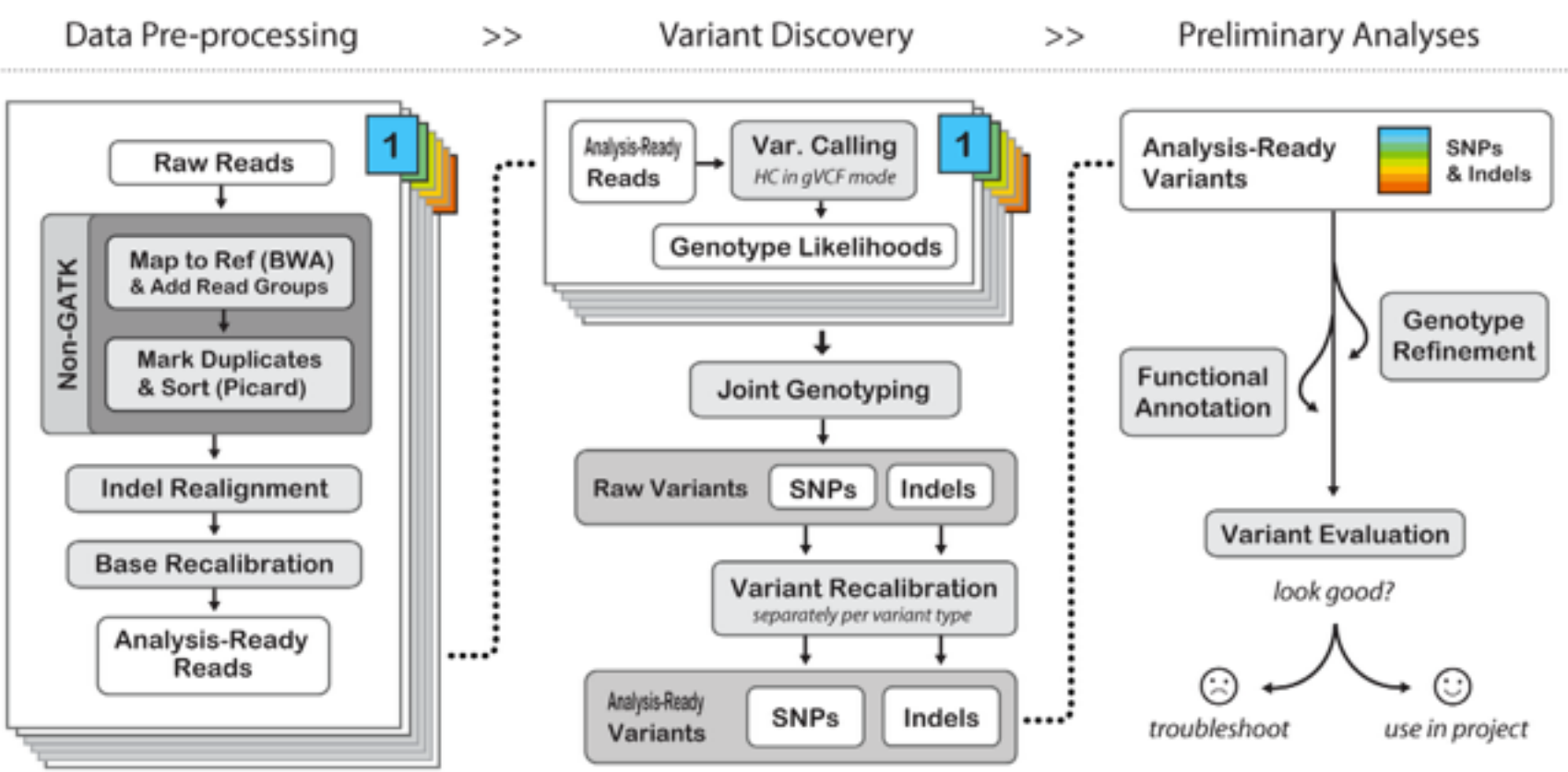
Help

Germline

Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter (Last submitted)	Submission accession
Pathogenic (Mar 3, 2004)	reviewed by professional society (evidence-based review)	literature only	Cystic fibrosis (Autosomal recessive inheritance) [MedGen | Orphanet | OMIM]	germline		American College of Medical Genetics and Genomics (ACMG) (Jun 3, 2013)	SCV000071408
Pathogenic (Mar 28, 2013)	reviewed by expert panel (evidence-based review)	literature only	Cystic fibrosis [MedGen | Orphanet | OMIM]	germline	PubMed (1)	CFTR2 (May 29, 2013)	SCV000071513
Pathogenic (Feb 13, 2013)	classified by single submitter (literature only)	literature only	Cystic fibrosis [MedGen | Orphanet | OMIM]	germline	PubMed (2)	OMIM (Dec 30, 2010)	SCV000027757
not provided (Feb 1, 2013)	not classified by submitter (literature only)	literature only	Cystic fibrosis [MedGen | Orphanet | OMIM]	germline	PubMed (2)	Invitae (Mar 30, 2013)	SCV000075041

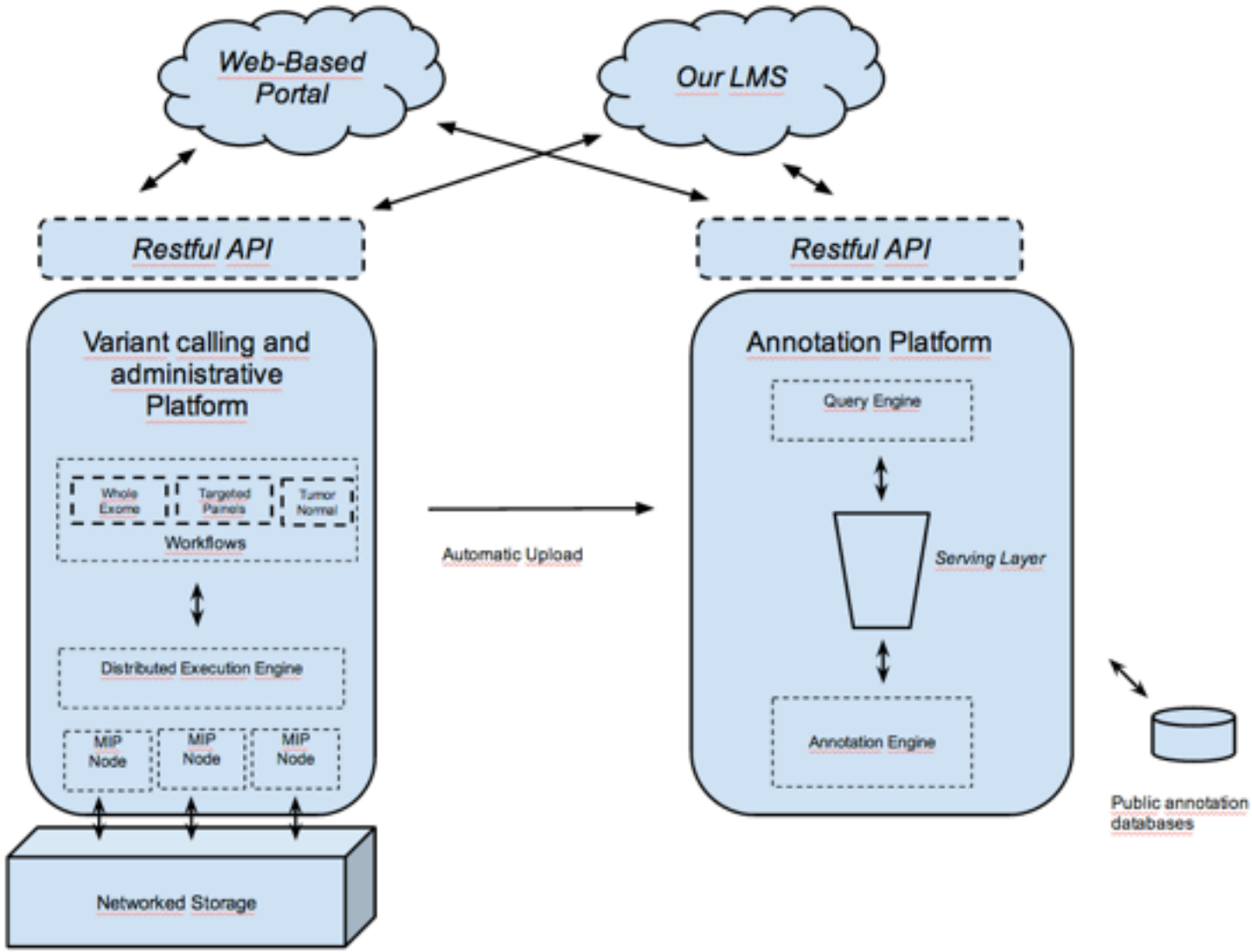
Last Updated: Jun 20, 2014

Análise de Variantes

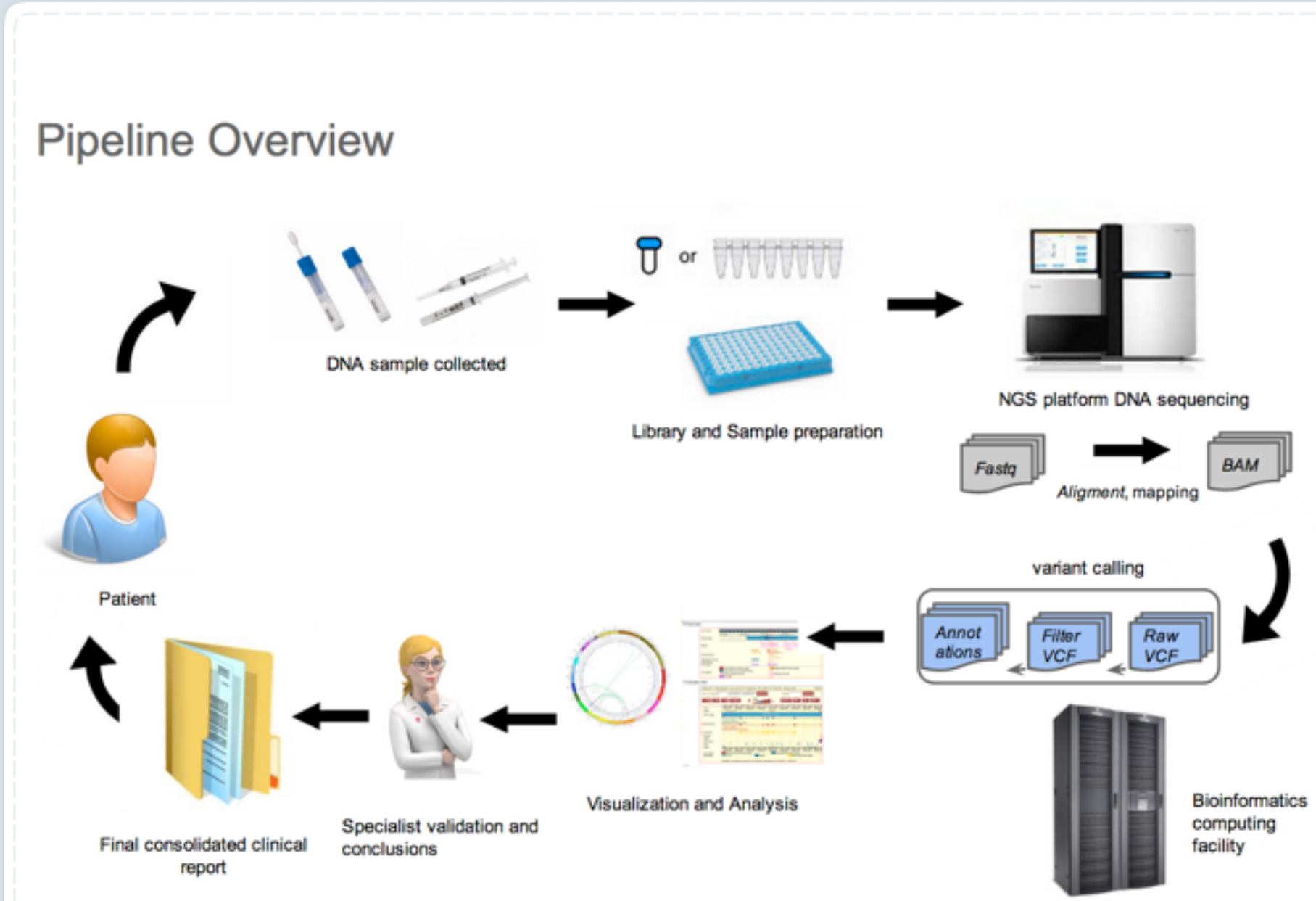


Análise de Variantes

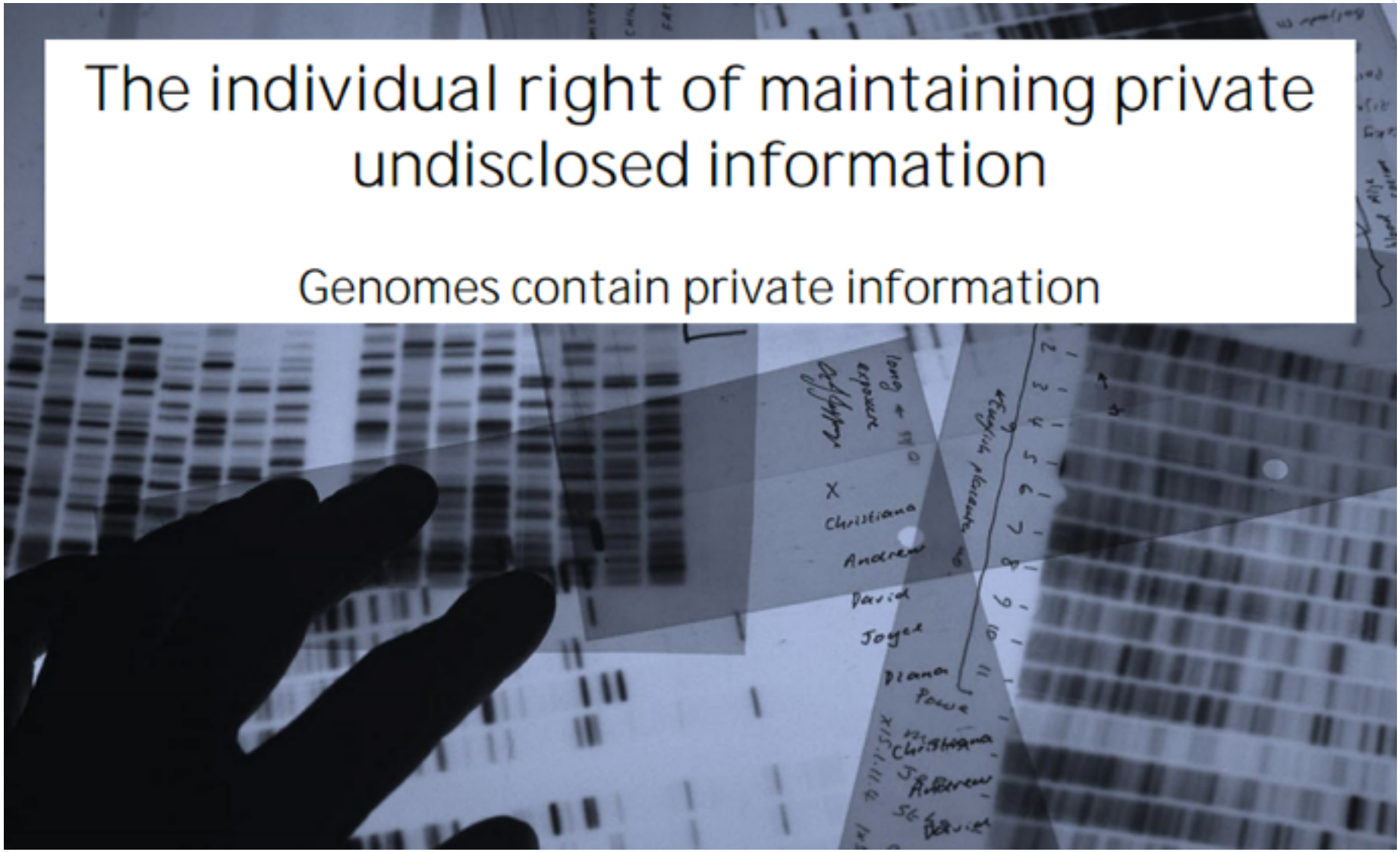
Análise de Variantes



Ciclo de vida de um exame



Segurança ?!


The individual right of maintaining private undisclosed information
Genomes contain private information

Conclusões

- Bioinformatics:
  - Interesting area
  - Many open problems
  - opportunities to acquire and apply knowledge
  - opportunities from projects
  - reachable, several courses worldwide

Conclusões

- Health Informatics
  - improve health informatics system
  - Several technologies still to improve
  - Processes, workflows and user experience
  - Information architecture
  - Lack of good solutions

Como posso aprender mais?

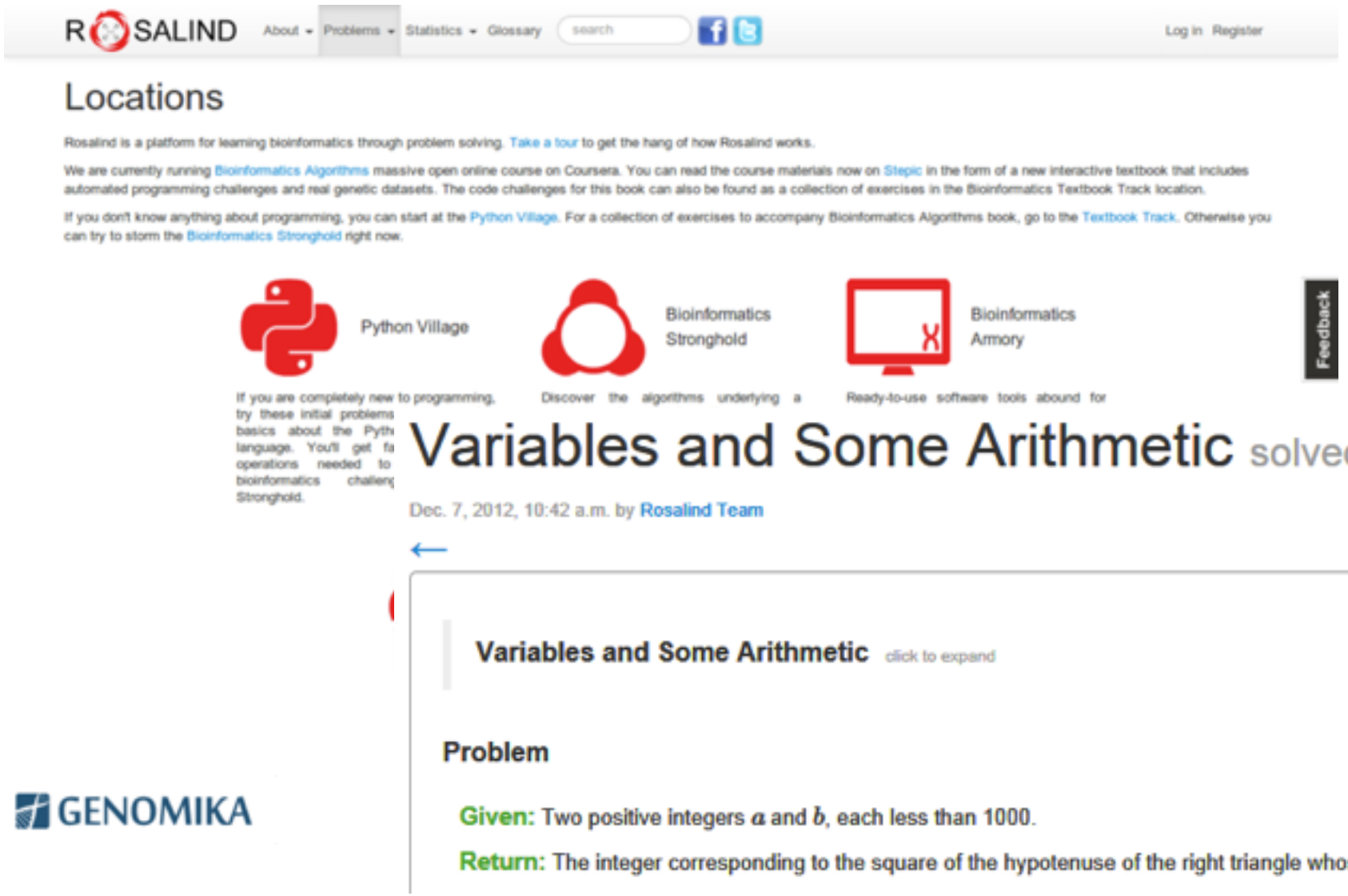
Tales of Genome (Udacity)

Curso On-line gratuito sobre Genética (bem completo!)

Rosalind

rosalind.info/

Desafios de Python na área de bioinformática



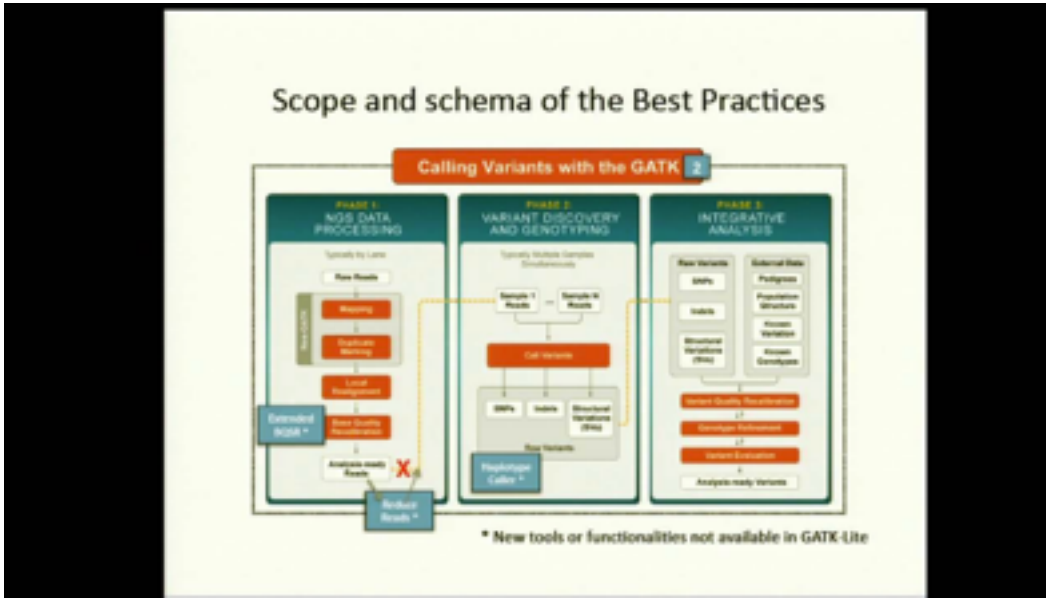
GENOMIKA

Broad workshops

https://www.broadinstitute.org/partnerships/education/broade/broad-workshops/

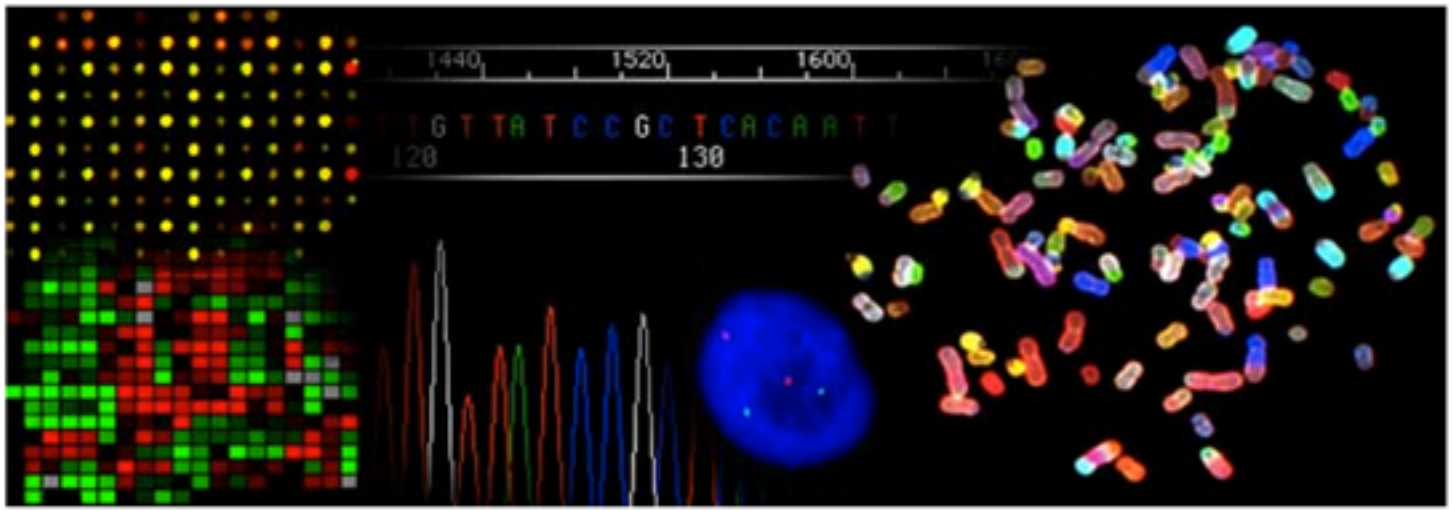
Variant analysis; sequencing pipelines, etc.

BROAD INSTITUTE



GENOMIKA

https://www.edx.org/



Genomic Medicine Gets Personal

This course will provide an introduction to genomic medicine and a better understanding of the issues associated with personal genomic information.

School:	GeorgetownX
Course Code:	MEDX202-01x
Classes Start:	4 Jun 2014
Course Length:	8 weeks
Estimated effort:	(8 weeks)

Prerequisites:

None. This is an introductory course.

GENOMIKA

https://www.coursera.org/

UC San Diego

Bioinformatics Algorithms (Part 2)

This is the second course in a two-part series on bioinformatics algorithms, covering the following topics: evolutionary tree reconstruction, applications of combinatorial pattern matching for read mapping, gene regulatory analysis, protein classification, computational proteomics, and computational aspects of human genetics.

About the Course

This course is the second in a two-part series that begins with Bioinformatics Algorithms (Part 1). It will build upon the biological and computational material covered in the first course to cover additional topics in modern computational biology.

The format for this course will be the same as that of Part 1. Each chapter of course material will cover a single biological question and slowly build the algorithmic knowledge required to address this challenge. Along the way, coding challenges and exercises (many of which ask you to apply your skills to real genetic data) will be directly integrated into the text at the exact moment they are needed.

Course Syllabus

Sessions

Feb 16, 2015 - May 10th 2015

Join for Free!

Earn a Verified Certificate

Eligible for

Verified Certificate
Statement of Accomplishment

GENOMIKA

Materiais diversos

http://ged.msu.edu/angus/bioinformatics-courses.html

http://angus.readthedocs.org/en/2014/

https://wikis.utexas.edu/display/bioiteam/NGS+Course+Resources#NGSCourseResources-Technologyvideos

Trabalhe conosco!

github.com/genomika/jobs

GENOMIKA

Obrigado!

Todos com 75% de participação receberão um certificado de participação on-line

Respondam nossa pesquisa de satisfação, é muito importante para nossas próximas edições!

GENOMIKA


RISCO
BIOLÓGICO
freedom evo


GENOMIKA


ALFRED NOBEL

"Biology easily has 500 years of exciting problems to work on."

Donald Knuth, 1993

E agora para onde posso ir ?

Marcel Caraciolo, CTO
marcel@genomika.com.br