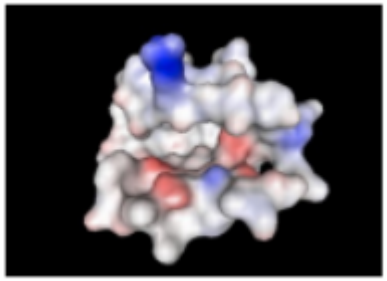
Introdução ao NGS

Rodrigo Bertollo
rodrigo@genomika.com.br

Cenário Atual

Sequencias de DNA formam genes



Genes podem ser traduzidos em proteínas

Genótipo vs Fenótipo

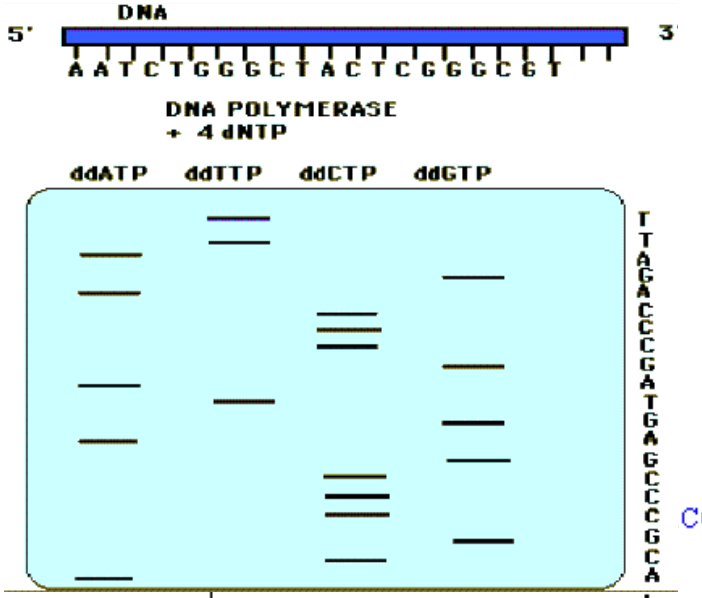
Pressão ambiental

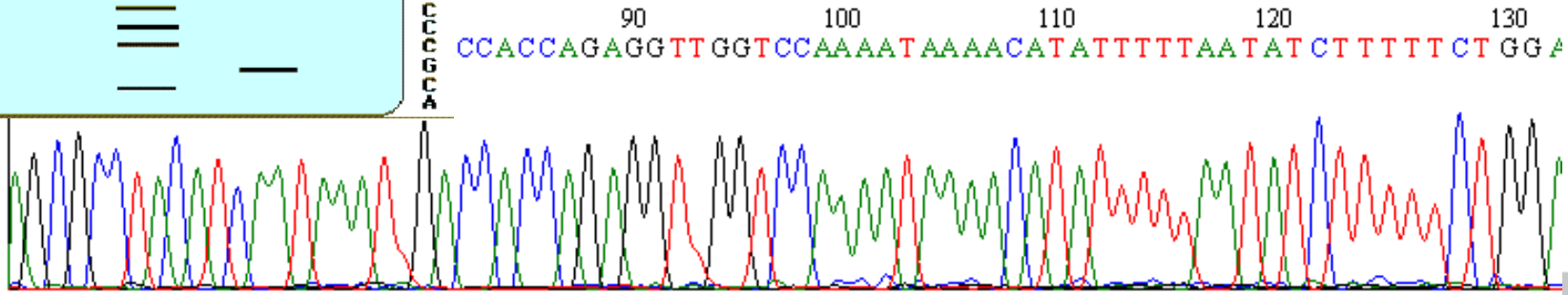
GENOMIKA

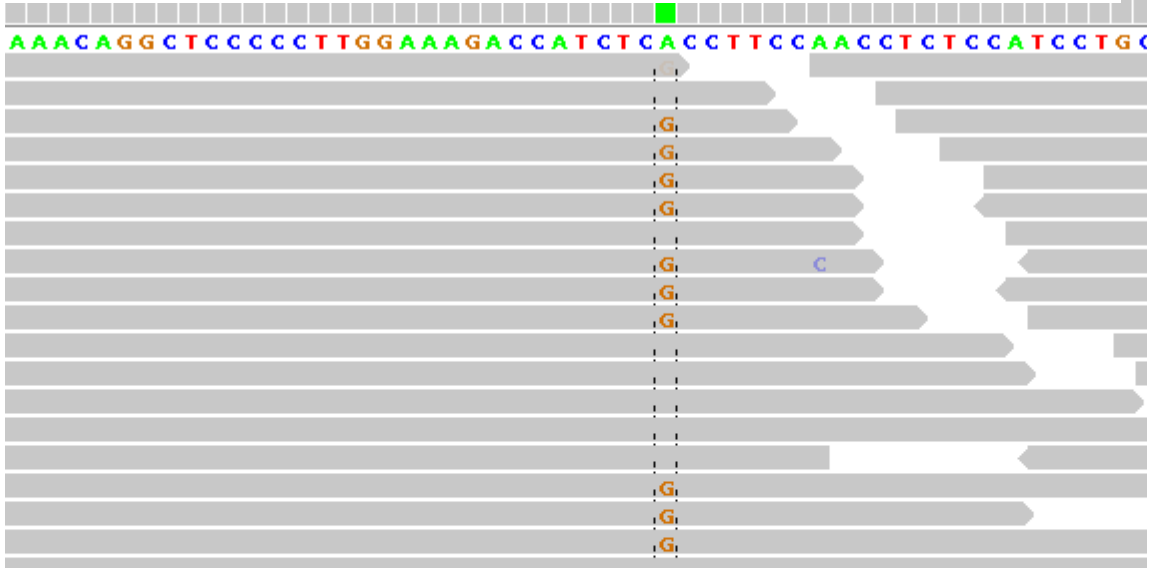
Sequenciamento de DNA

Sequenciando fragmentos de bases nitrogenadas do DNA

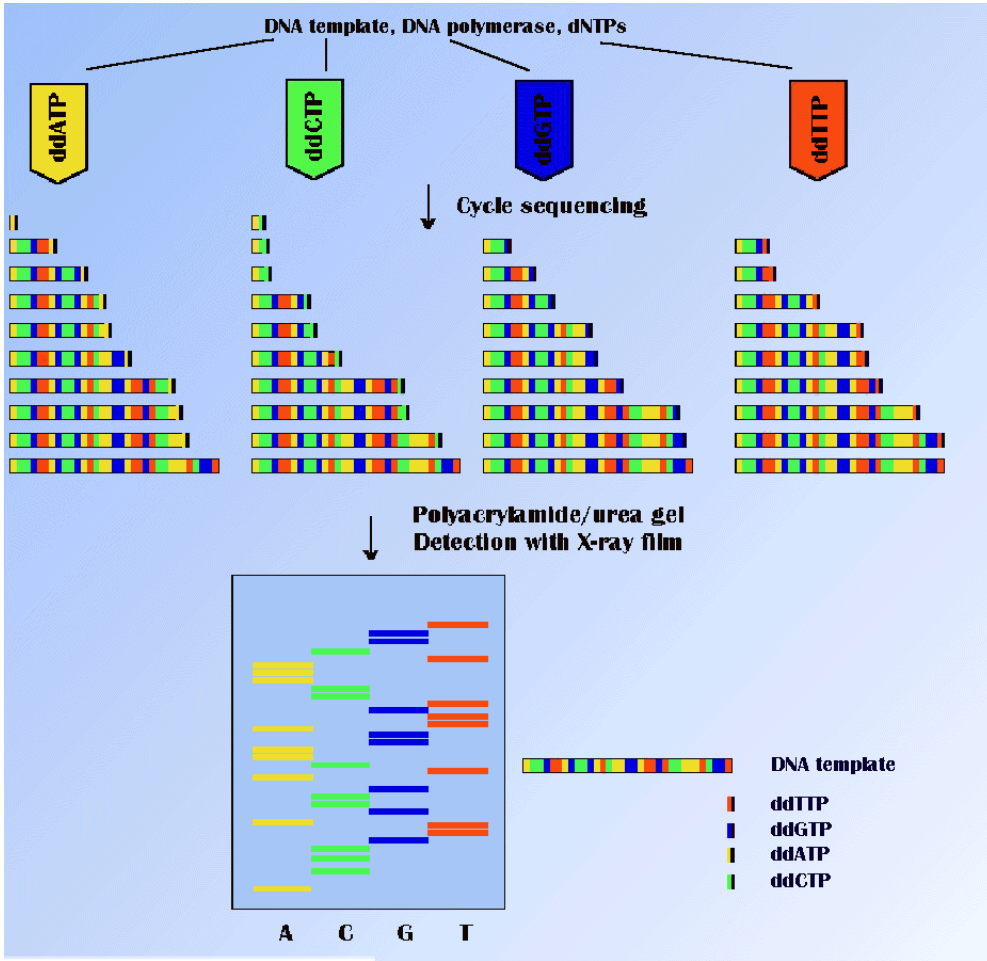
História

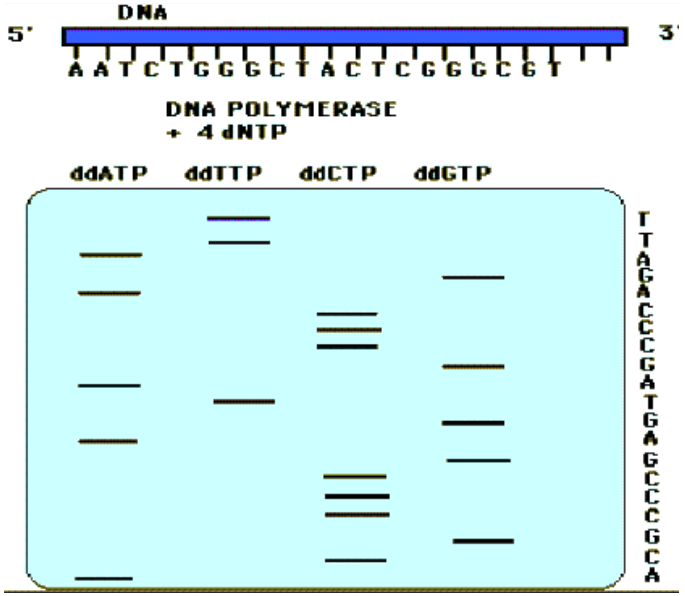




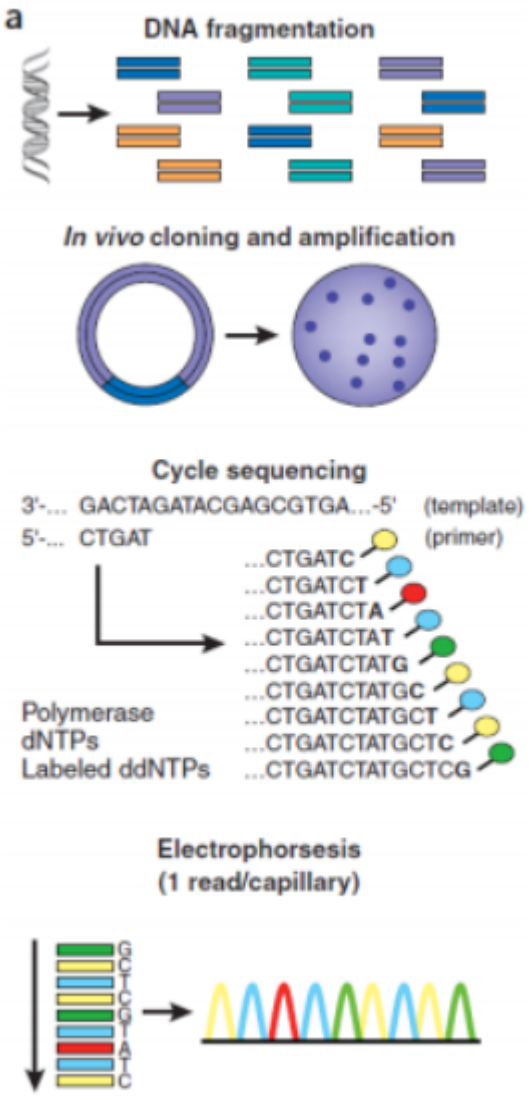


GENOMIKA

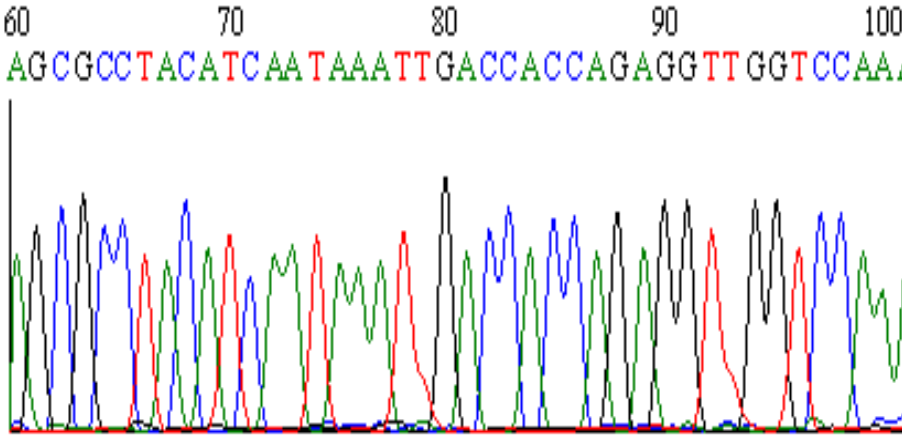

DNA template, DNA polymerase, dNTPs
ddATP
ddCTP
ddGTP
ddTTP
Cycle sequencing
Polyacrylamide/urea gel
Detection with X-ray film
A C G T
DNA template
ddTTP
ddGTP
ddATP
ddCTP


5'
DNA
3'
A A T C T G G G C T A C T C G G G C G T
DNA POLYMERASE
+ 4 dNTP
ddATP
ddTTP
ddCTP
ddGTP
T T A G A C C C G A T G A G C C C G C A

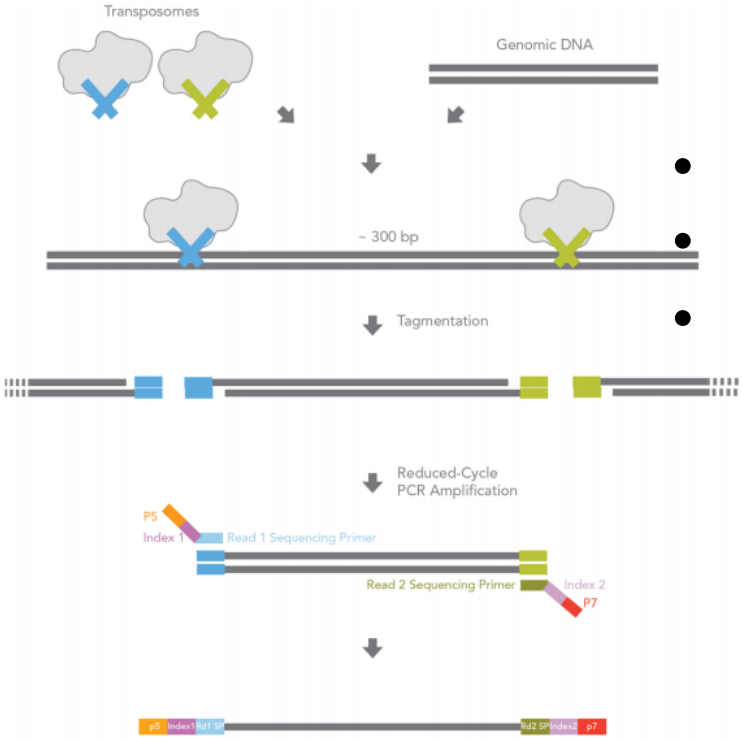
GENOMIKA



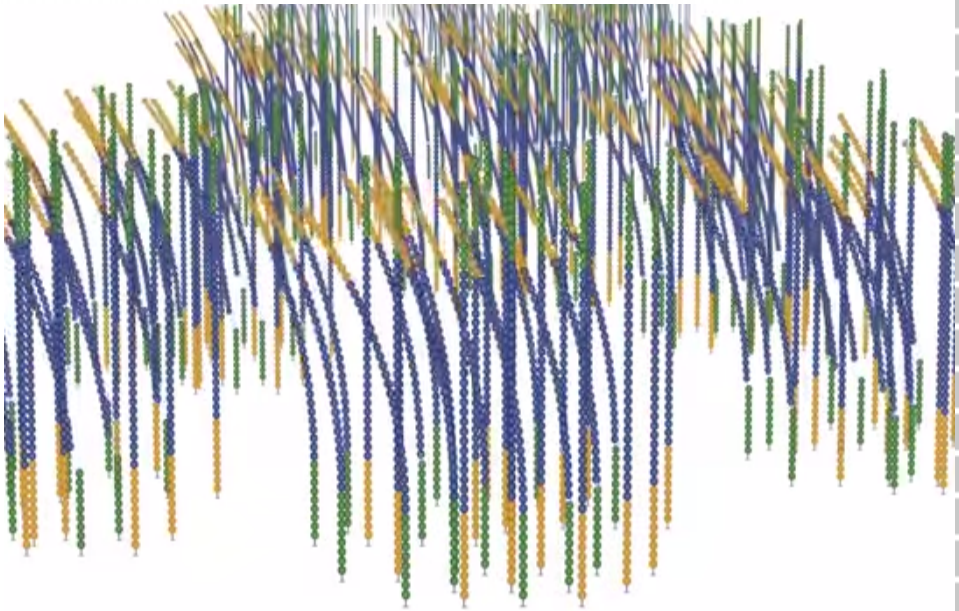
- Limitações te tamanho do fragmento
- Análise de 1 indivíduo por vez
- Preço (custo/sequencia)

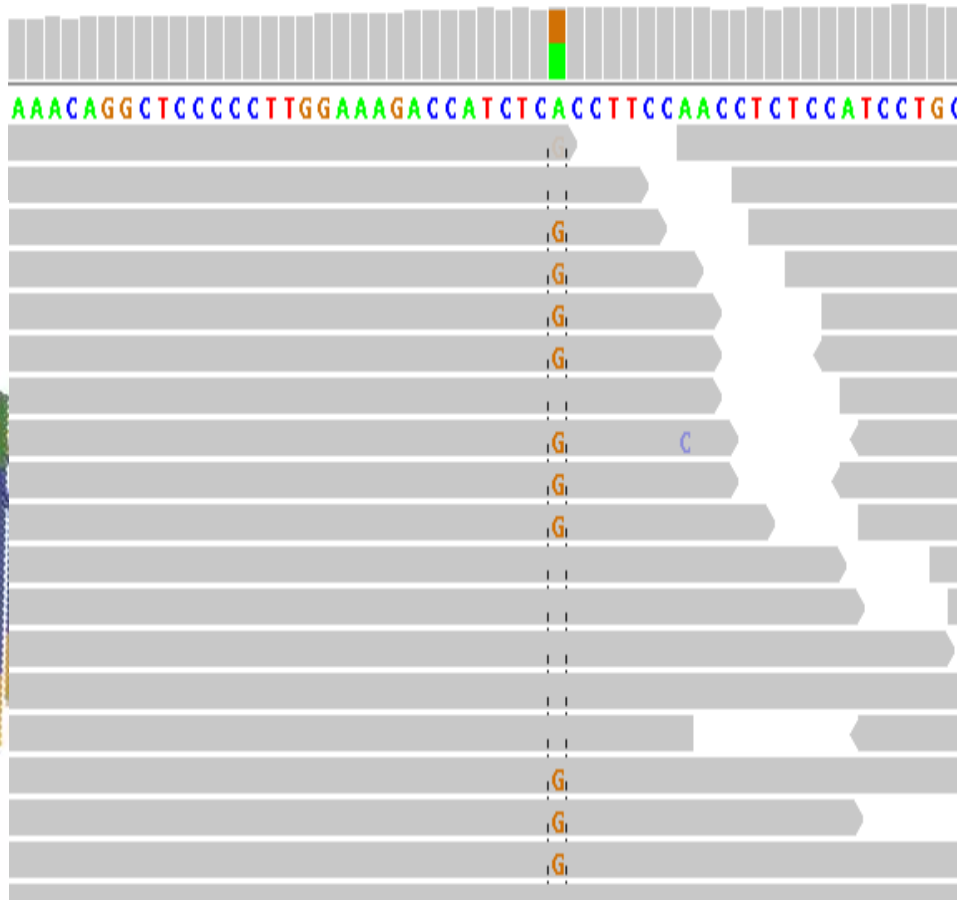


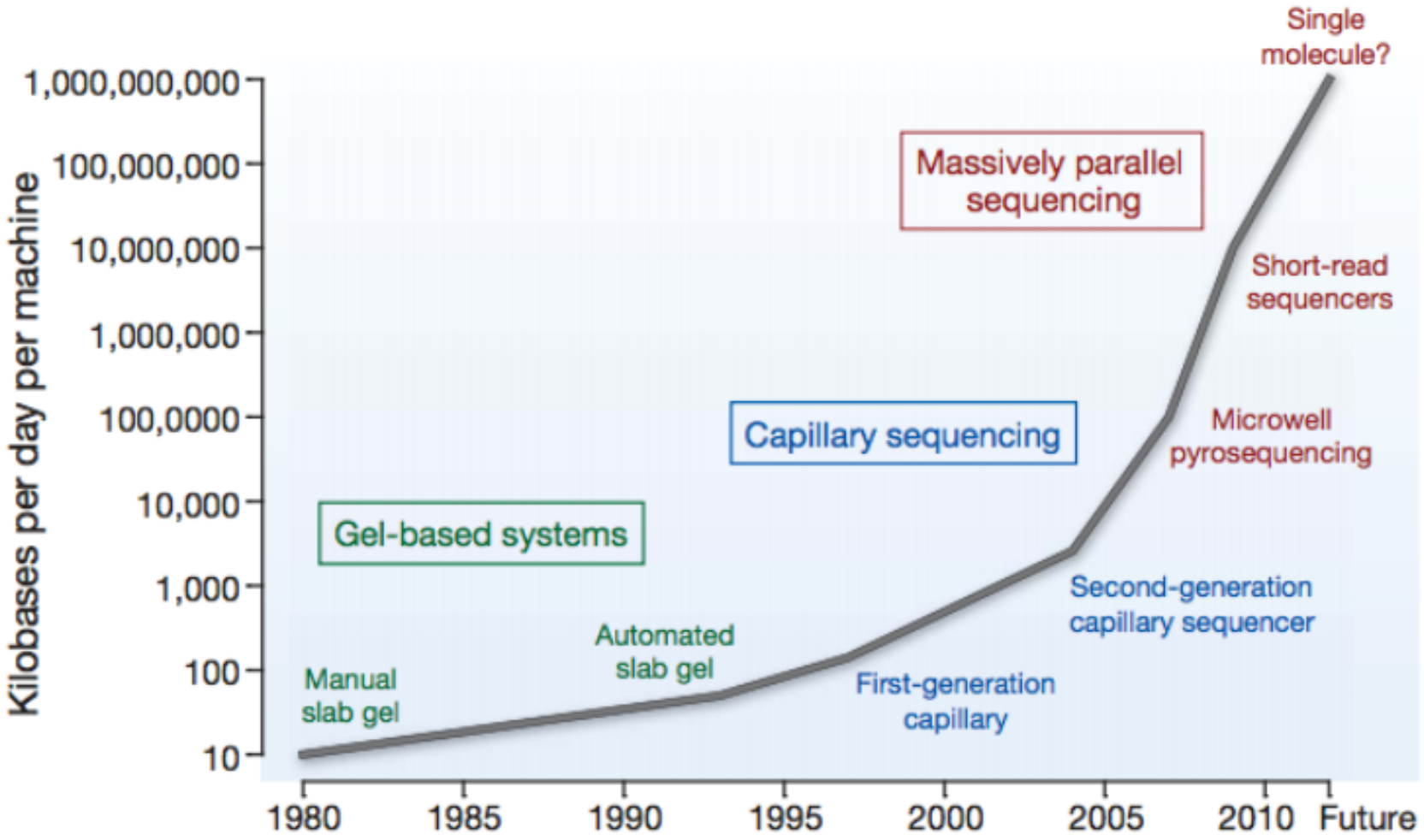
GENOMIKA

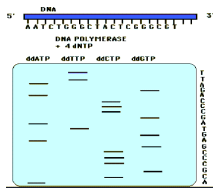


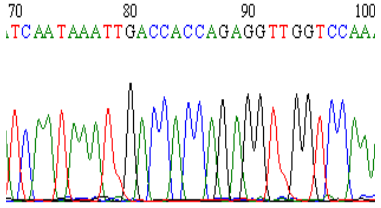
- Limitações te tamanho do fragmento (???)
- Análise de N indivíduos por vez
- Preço (custo/sequencia)

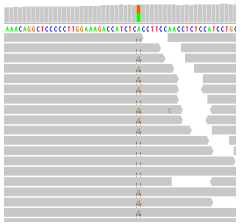





Kilobases per day per machine
1,000,000,000
100,000,000
10,000,000
1,000,000
100,0000
10,000
1,000
100
10
1980
1985
1990
1995
2000
2005
2010
Future
Single molecule?
Massively parallel sequencing
Short-read sequencers
Capillary sequencing
Microwell pyrosequencing
Gel-based systems
Second-generation capillary sequencer
Automated slab gel
Manual slab gel
First-generation capillary







GENOMIKA

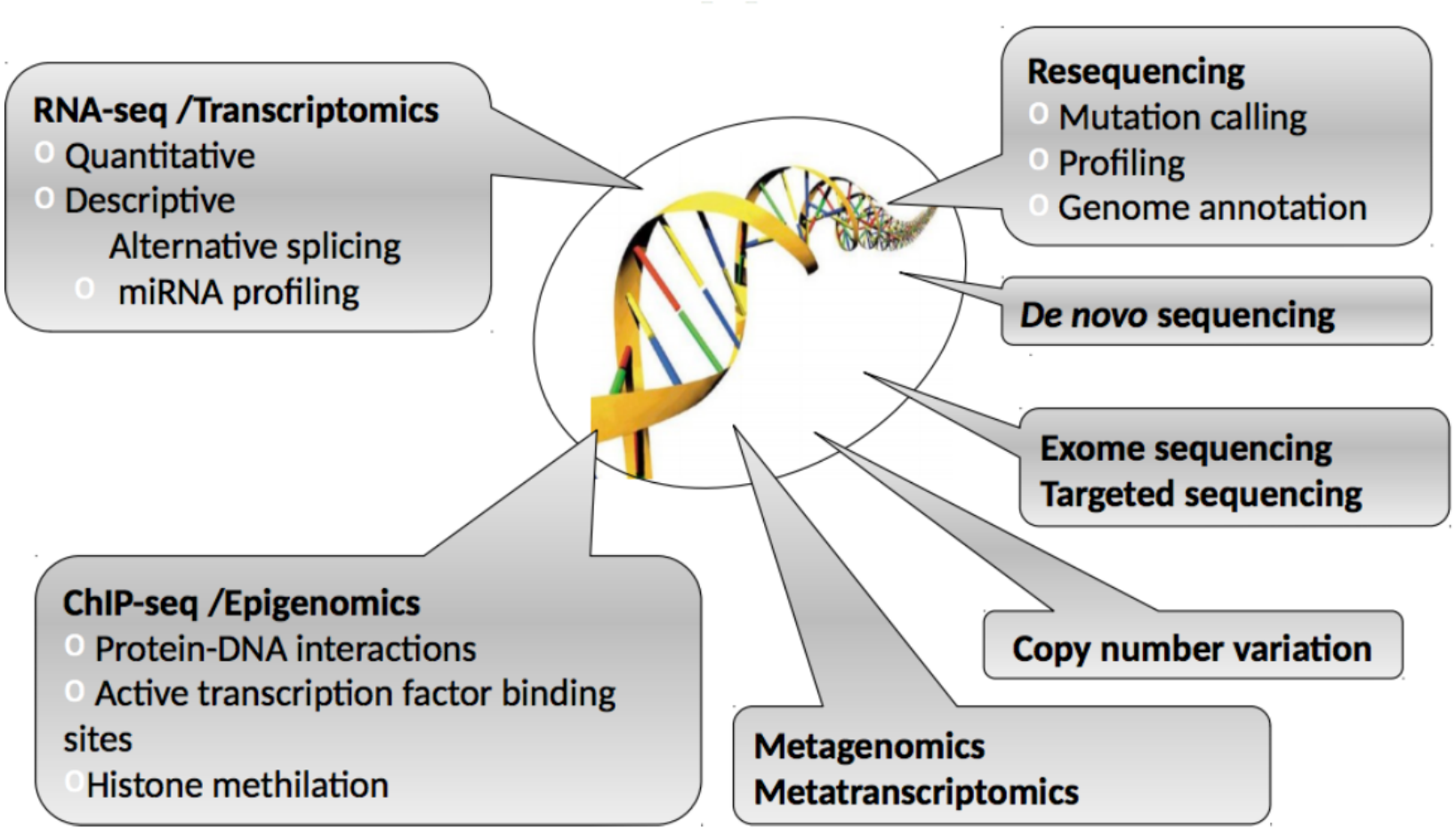
Plataformas

GENOMIKA

Plataformas

Platform	3730XL	454 FLX	454 GS JR	HiSeq 2000	MiSeq	SOLiD 5500	IonTorrent	PacBio RS
Method of amplification	Clonal plasmid amplification	emRCR on beads	emRCR on beads	Bridge PCR amplification	Bridge PCR amplification	emPCR on bead	emPCR on bead	None
Chemistry	Chain termination	Synthesis (Pyro-sequencing)	Synthesis (Pyro-sequencing)	Synthesis (Reversible termination)	Synthesis (Reversible termination)	Ligation (dual-base encoding)	Synthesis (H^+ detection)	Synthesis
Instrument Cost	$376k	$500k	$108k	$690k	$125k	$595k	$67.5k	$695k
Yield per Run	60 kb	900 Mb	50 Mb	600 Gb	1 Gb	155 Gb	1 Gb	20-80 Mb
Read Length (bases)	650	750	400	100	150	75 + 35	200 (318 chip)	<1,800 - >5,000
Reagent Cost (library + run)	$96	$6 200	$1 100	$23 610	$1 035	$10 503	$925	$272
Cost per Mb	$1600	$7	$22	$0.039	$1	$0.068	$0.93	$3.4-13.6
Primary error & error rate	Substitution 0.1-1 %	Indel 1%	Indel 1%	Substitution >0.1%	Substitution >0.1%	indel >0.01%	Indel ~1%	Indel ~15%
Primary Advantage	Low cost for small study	Long read length	Long read length	Most output at lowest cost	Easy workflow & fast run	Each lane can be run independently & ability to rescue failed cycle	Fast run, low cost, and trajectory to longer read	Longest read length, single molecule real-time seq
Primary Disadvantage	High cost for large study	Unreliable for homopolymer region; High cost NGS	High cost per Mb	High capital cost & computation need	Few reads & higher cost per Mb	Relatively short read, more gap in assemblies	Unreliable for long homopolymer region	High error rates, Low output, expensive

Aplicações



GENOMIKA

Plataformas

Roche 454

- Long fragments
- Errors: poly nts
- Low throughput
- Expensive

- De novo sequencing
- Amplicon sequencing
- RNASeq

Illumina

- Short fragments
- Errors: Hexamer bias
- High throughput
- Cheap

- Resequencing
- De novo sequencing
- ChipSeq
- RNASeq
- MethylSeq

SOLiD

- Short fragments
- Color-space
- High throughput
- Cheap

- Resequencing
- ChipSeq
- RNASeq
- MethylSeq

Date	Cost per Mb	Cost per Genome
Sep-01	$5,292.39	$95,263,072
Mar-02	$3,898.64	$70,175,437
Sep-02	$3,413.80	$61,448,422
Mar-03	$2,986.20	$53,751,684
Oct-03	$2,230.98	$40,157,554
Jan-04	$1,598.91	$28,780,376
Apr-04	$1,135.70	$20,442,576
Jul-04	$1,107.46	$19,934,346
Oct-04	$1,028.85	$18,519,312
Jan-05	$974.16	$17,534,970
Apr-05	$897.76	$16,159,699
Jul-05	$898.90	$16,180,224
Oct-05	$766.73	$13,801,124
Jan-06	$699.20	$12,585,659
Apr-06	$651.81	$11,732,535
Jul-06	$636.41	$11,455,315
Oct-06	$581.92	$10,474,556
Jan-07	$522.71	$9,408,739
Apr-07	$502.61	$9,047,003
Jul-07	$495.96	$8,927,342
Oct-07	$397.09	$7,147,571
Jan-08	$102.13	$3,063,820
Apr-08	$15.03	$1,352,982
Jul-08	$8.36	$752,080
Oct-08	$3.81	$342,502
Jan-09	$2.59	$232,735
Apr-09	$1.72	$154,714
Jul-09	$1.20	$108,065
Oct-09	$0.78	$70,333
Jan-10	$0.52	$46,774
Apr-10	$0.35	$31,512
Jul-10	$0.35	$31,125
Oct-10	$0.32	$29,092
Jan-11	$0.23	$20,963
Apr-11	$0.19	$16,712
Jul-11	$0.12	$10,497
Oct-11	$0.09	$7,743
Jan-12	$0.09	$7,666

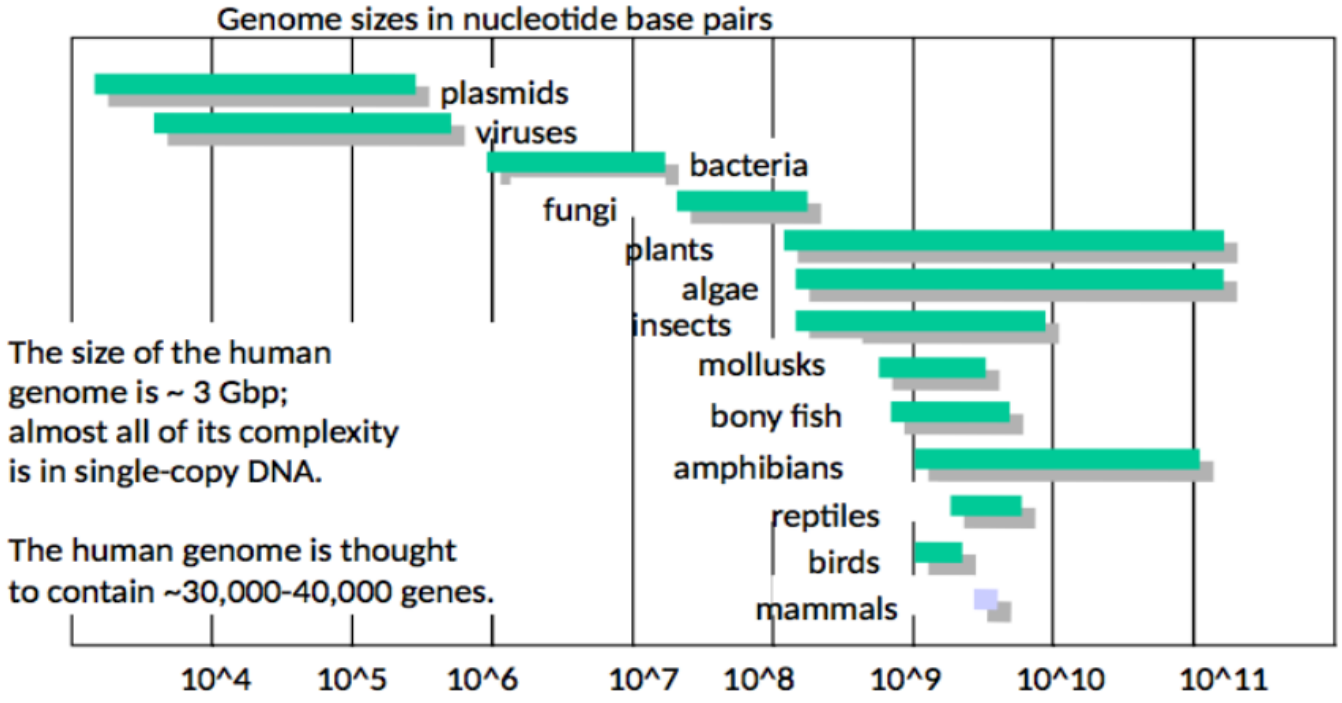
National Human Genome Research Institute

genome.gov/sequencingcosts

Genoma Humano

Antes:

- Primeiro projeto do Genoma Humano foi finalizado em 2004.
- Custo estimado de 3 bilhões de U$.
- 13 anos para ser completado.



Genoma Humano

Hoje:

- Sequenciamento pela plataforma Illumina que surgiu em 2012.
- Custo estimado de mil dollares
- < 1 semana para ser completo.

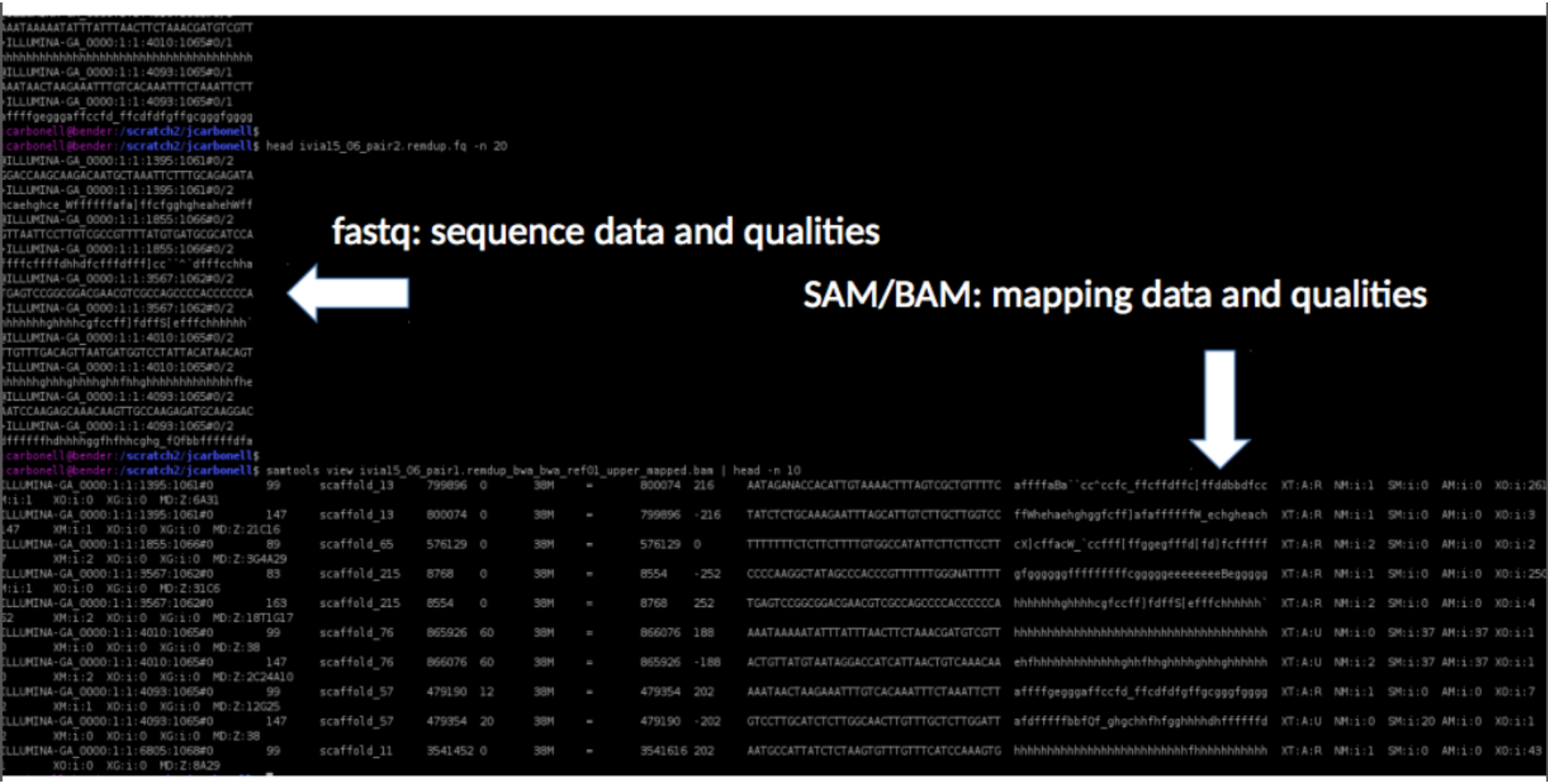
GENOMIKA

Armazenamento de dados e informação

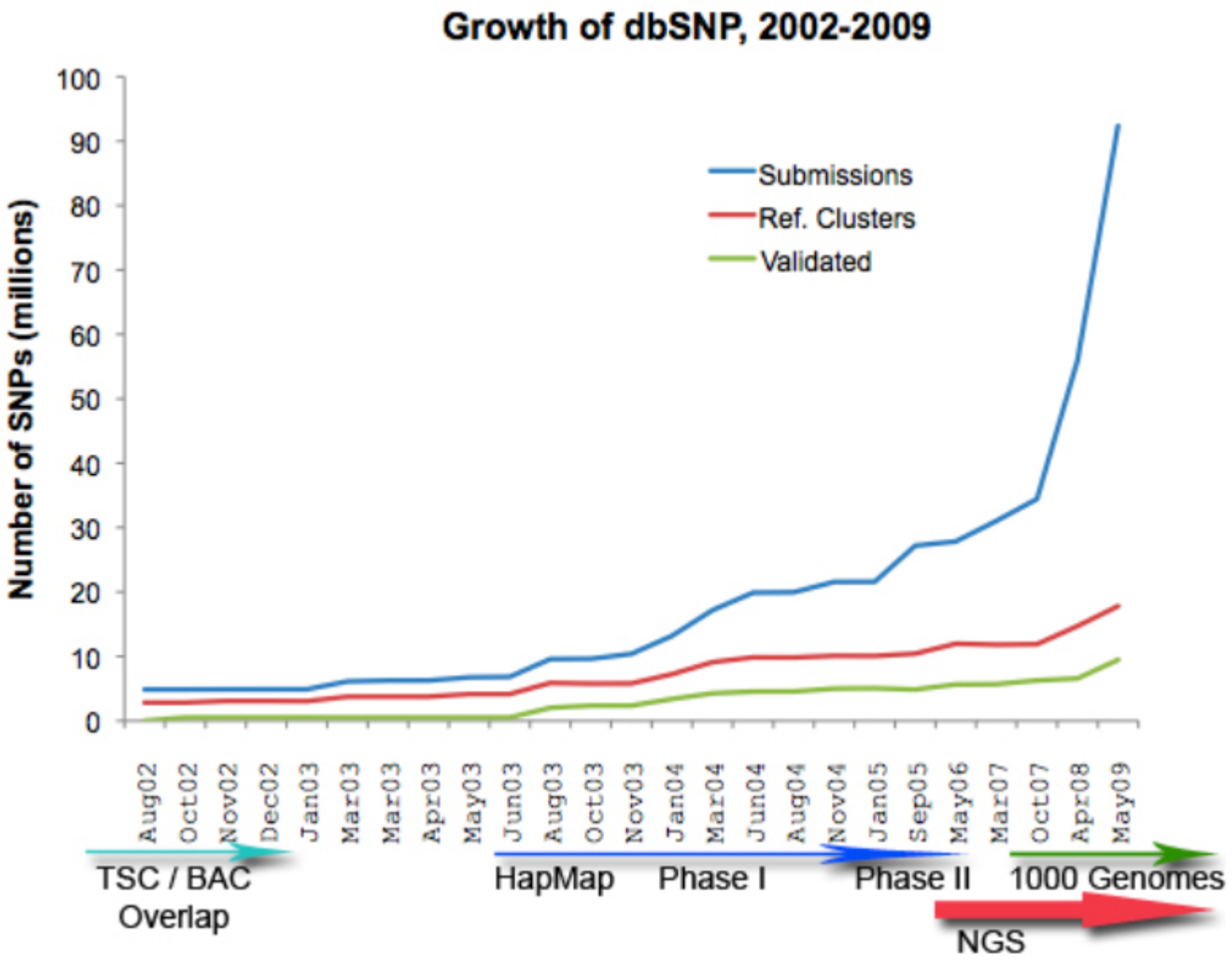
- Um sequenciamento de genoma de um humano, com cobertura de 30-40 leituras por sequencia, gera em média 150 Gbase de dados.
- Parcialmente o custo maior do sequenciamento está no armazenamento do que no próprio sequenciamento
- BGI (China), é o maior centro de sequenciamento do planeta com 167 sequenciadores, produzindo analises de 150-300 genomas por dia (2012)

GENOMIKA

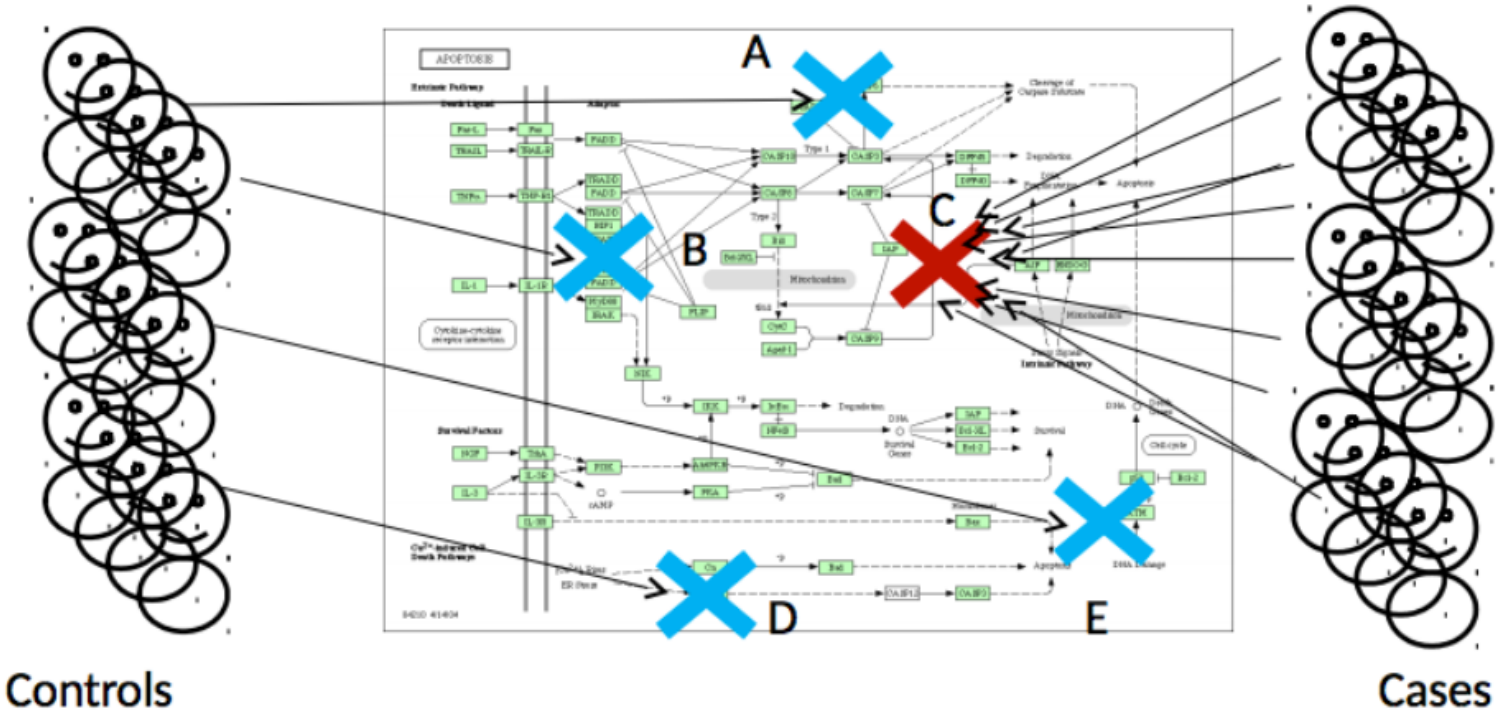
Formato de arquivos



GENOMIKA

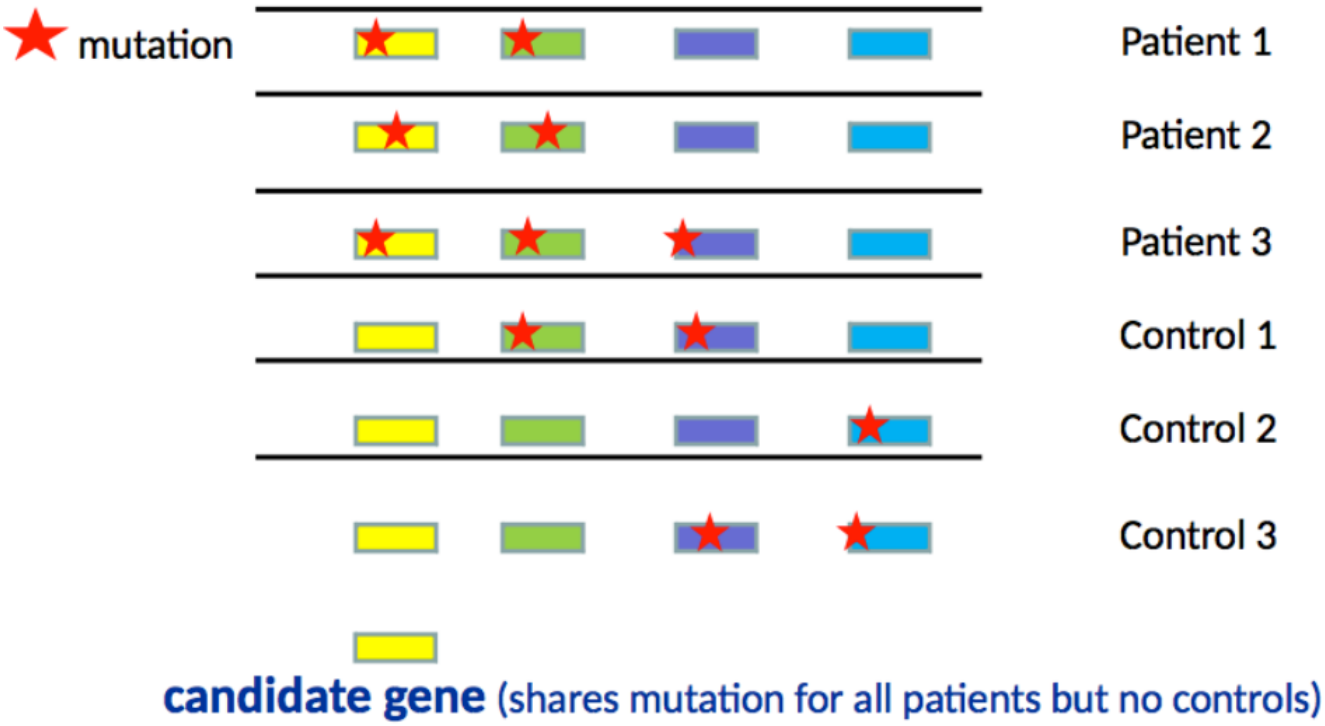

Growth of dbSNP, 2002-2009
Number of SNPs (millions)
100
90
80
70
60
50
40
30
20
10
0
Submissions
Ref. Clusters
Validated
Aug02
Oct02
Nov02
Dec02
Jan03
Mar03
Mar03
Apr03
May03
Jun03
Aug03
Oct03
Nov03
Jan04
Mar04
Jun04
Aug04
Nov04
Jan05
Sep05
May06
Mar07
Oct07
Apr08
May09
TSC / BAC Overlap
HapMap
Phase I
Phase II
1000 Genomes
NGS

Análise Secundária
Detectar as mutações associadas à doença



- Uma única mutação / combinação de alterações
- Significância estatística
- Um gene / uma via metabólica / um cromossomo (InDel)

Análise Secundária
Detectar as mutações associadas à doença



GENOMIKA

NATURE | ARTICLE OPEN

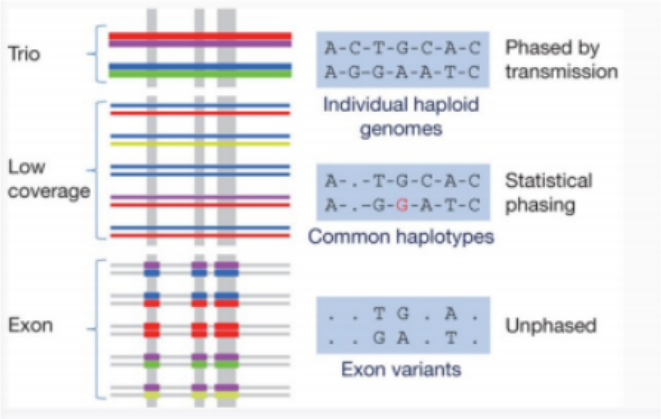
previous article next article

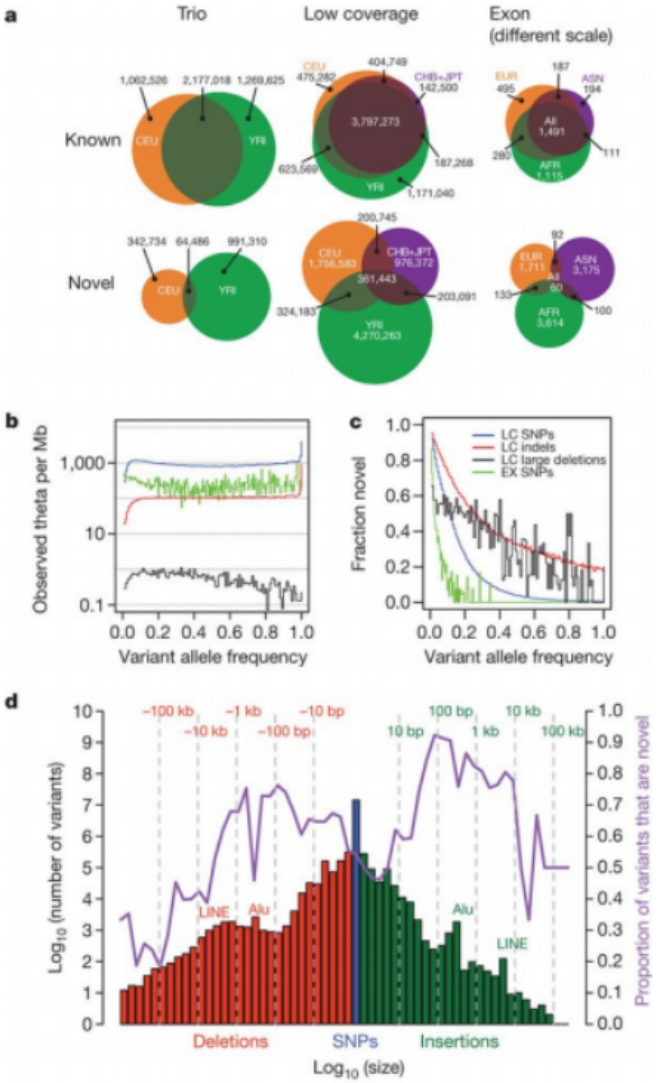
A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature **467**, 1061–1073 (28 October 2010) | doi:10.1038/nature09534
Received 20 July 2010 | Accepted 30 September 2010 | Published online 27 October 2010





GENOMIKA

Published in final edited form as:
Nat Genet. 2010 January ; 42(1): 30–35. doi:10.1038/ng.499.

Exome sequencing identifies the cause of a Mendelian disorder

Sarah B. Ng[1,*], **Kati J. Buckingham**[2,*], **Choli Lee**[1], **Abigail W. Bigham**[2], **Holly K. Tabor**[2], **Karin M. Dent**[3], **Chad D. Huff**[4], **Paul T. Shannon**[5], **Ethylin Wang Jabs**[6,7], **Deborah A. Nickerson**[1], **Jay Shendure**[1,†], and **Michael J. Bamshad**[1,2,8,†]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA [2]Department of Pediatrics, University of Washington, Seattle, Washington, USA [3]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA [5]Institute of Systems Biology, Seattle WA, USA [6]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA [7]Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland [8]Seattle Children's Hospital, Seattle, Washington, USA

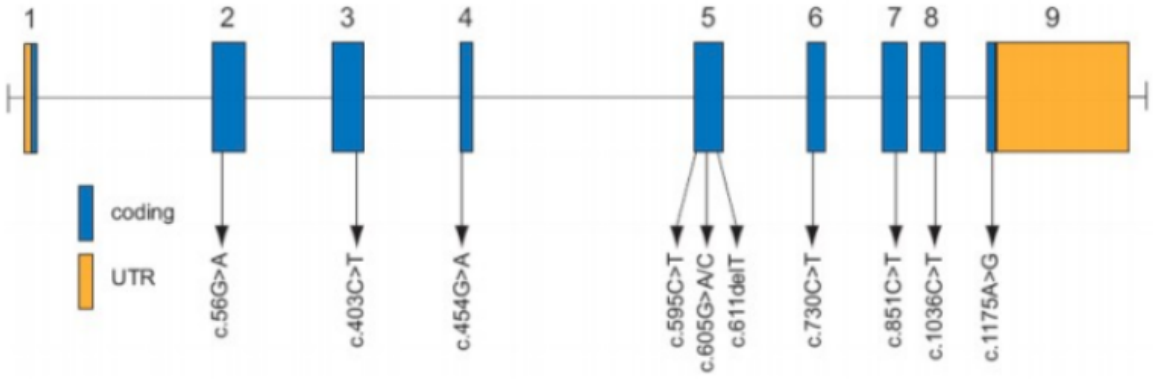


Figure 2. Genomic structure of the exons encoding the open reading frame of *DHODH*

DHODH is composed of 9 exons that encode untranslated regions (orange) and protein coding sequence (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.

Miller syndrome

GENOMIKA

Fica claro que…

- NGS está revolucionando a maneira como estamos fazendo pesquisa com genoma.
- Novas metodologias de análises.
- Revolução das técnicas estatísticas.
- Especialistas para vias metabólicas vs. especialistas para doenças

Mas fica o próximo passo:

Como revolucionar as nossas vidas quando formos capazes de processar **TODOS** os **DADOS**

GENOMIKA

Introdução ao NGS

Rodrigo Bertollo
rodrigo@genomika.com.br