



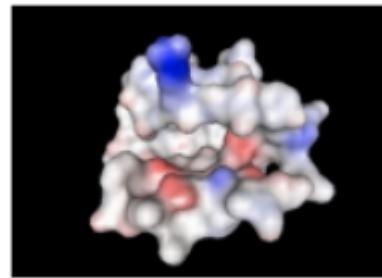
# Introdução ao NGS

Rodrigo Bertollo  
[rodrigo@genomika.com.br](mailto:rodrigo@genomika.com.br)

# Introdução a BioMol

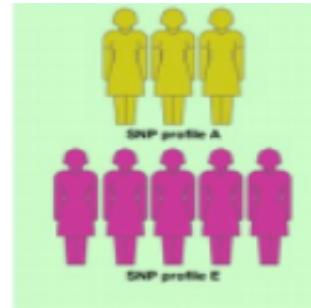


Sequencias de DNA formam genes



Genes podem ser traduzidos  
em proteínas

Genótipo vs Fenótipo

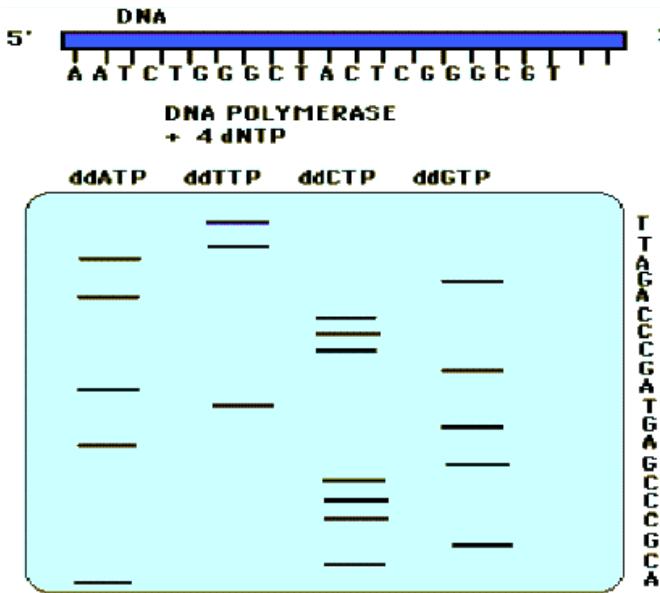


Pressão ambiental

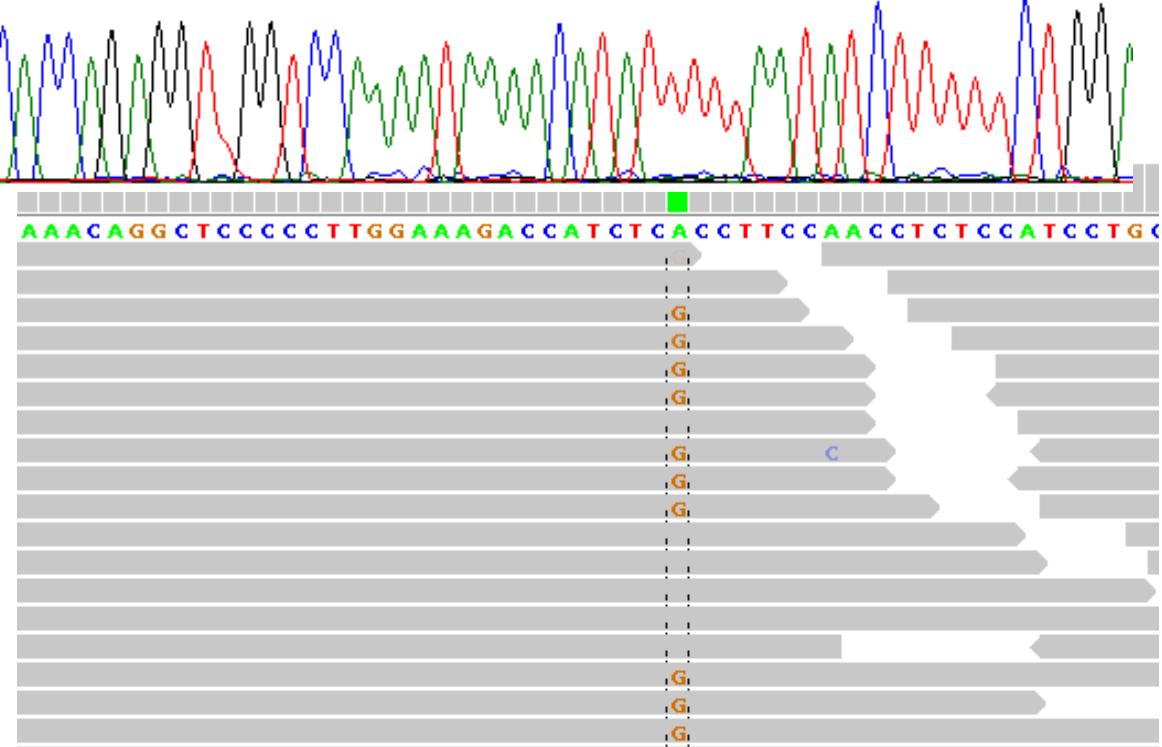
# Sequenciamento de DNA

Sequenciando fragmentos  
de bases nitrogenadas do DNA

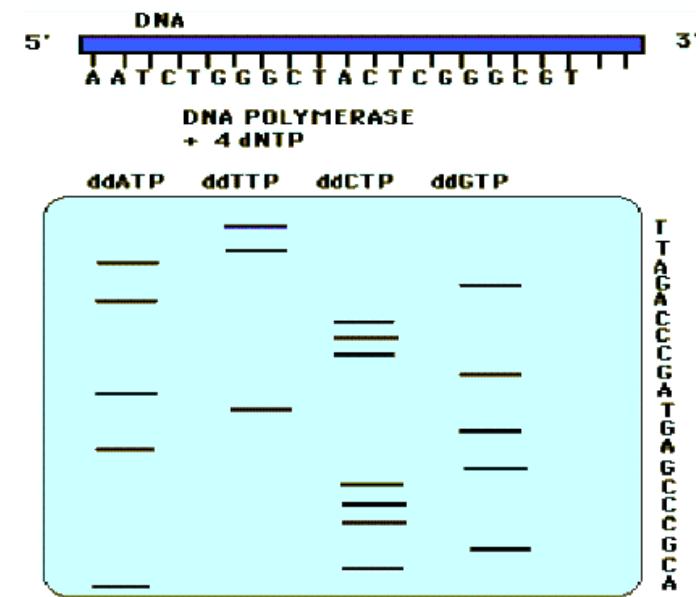
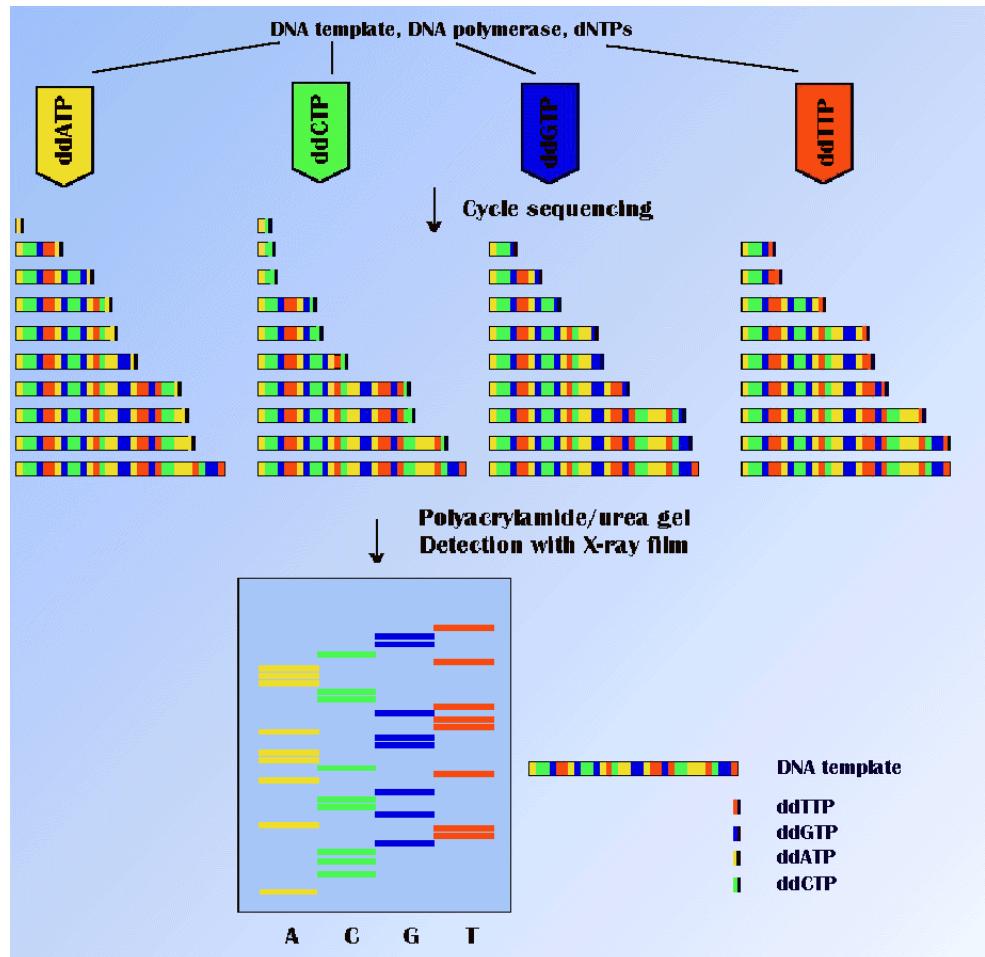
# Evolução do sequenciamento



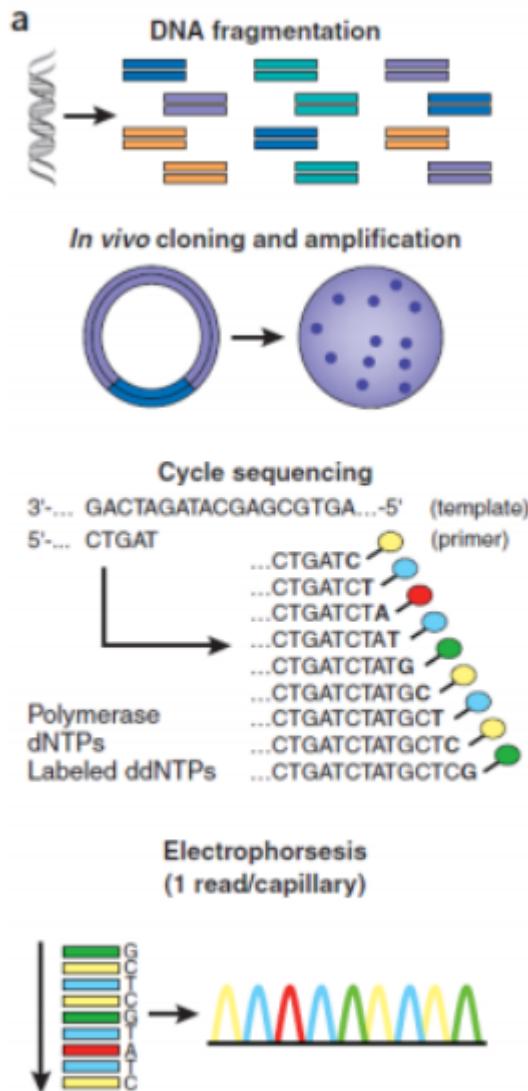
90 CCACCCAGAGGG TT GG TCC AAAA T AAAA AC AT TTTTA AT AT CTTTTCTGGAA



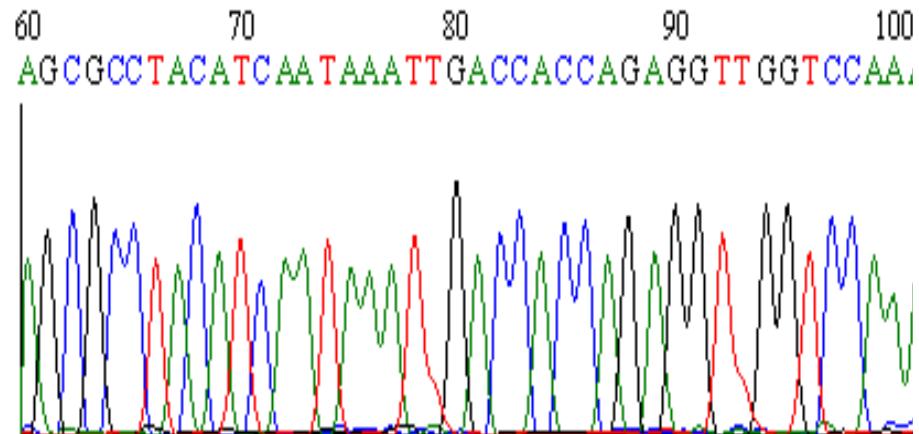
# Sequenciamento de gel



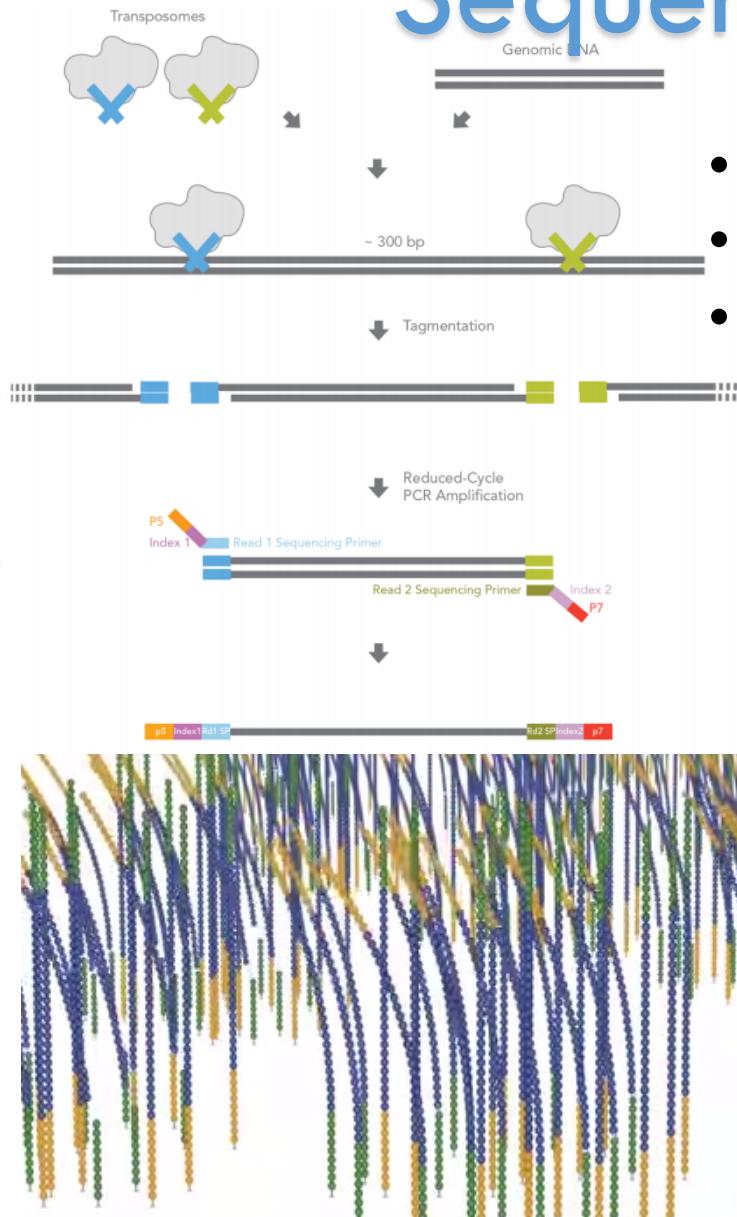
# Sequenciamento por capilar



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



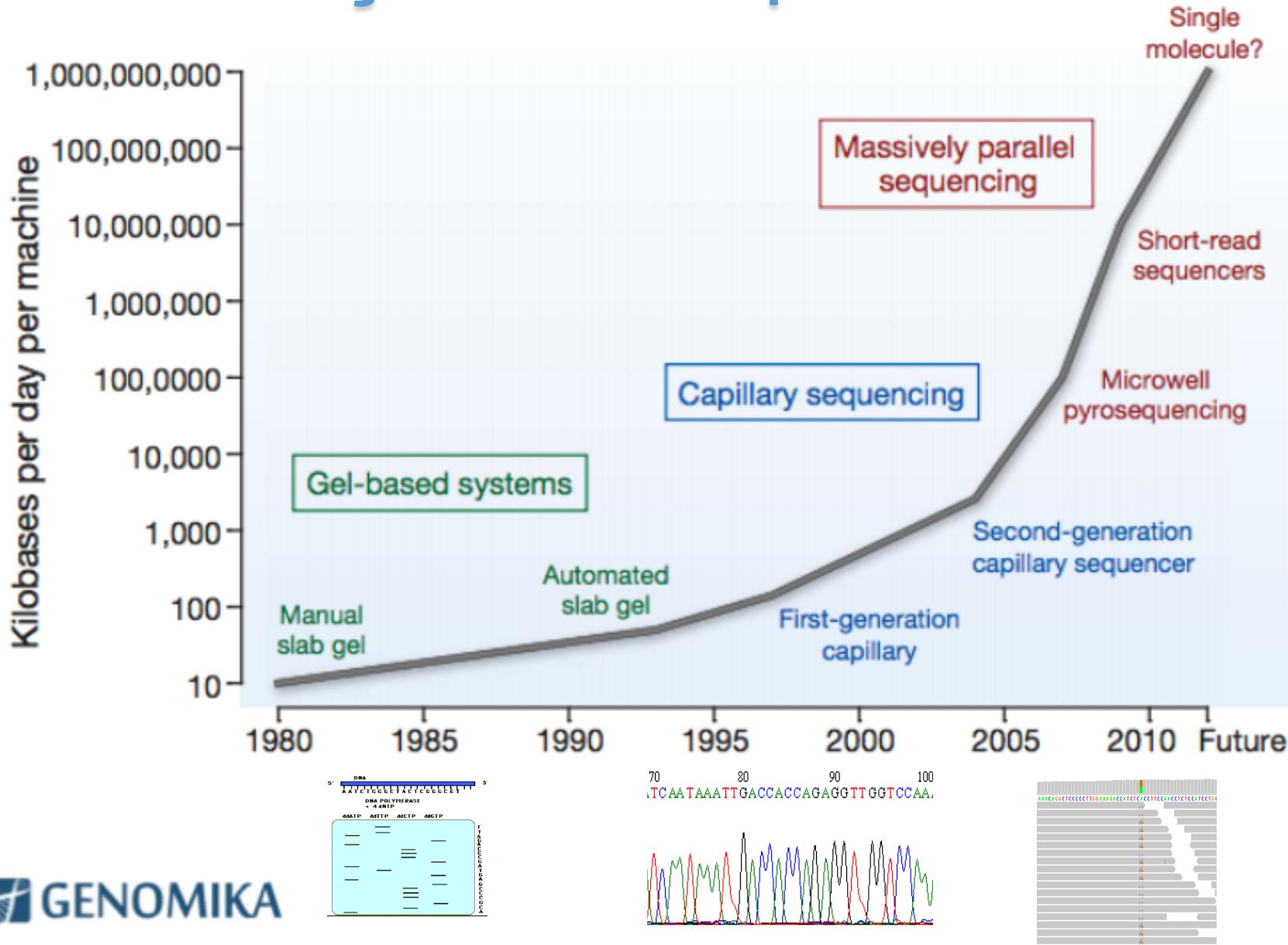
# Sequenciamento NGS



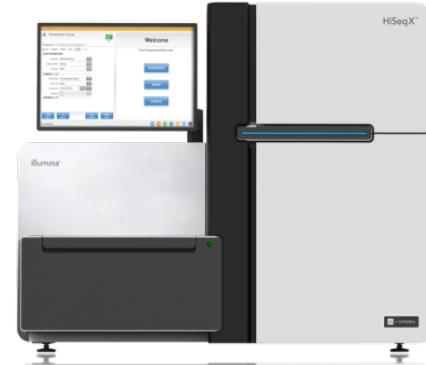
- Limitações te tamanho do fragmento (???)
- Análise de N indivíduos por vez
- Preço (custo/sequencia)



# Evolução do sequenciamento



# Plataformas



# Aplicações

## RNA-seq / Transcriptomics

- Quantitative
- Descriptive
  - Alternative splicing
  - miRNA profiling

## Resequencing

- Mutation calling
- Profiling
- Genome annotation

## *De novo* sequencing

## Exome sequencing Targeted sequencing

## Copy number variation

## ChIP-seq / Epigenomics

- Protein-DNA interactions
- Active transcription factor binding sites
- Histone methylation

## Metagenomics Metatranscriptomics

# 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

# Preço por Mbases

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



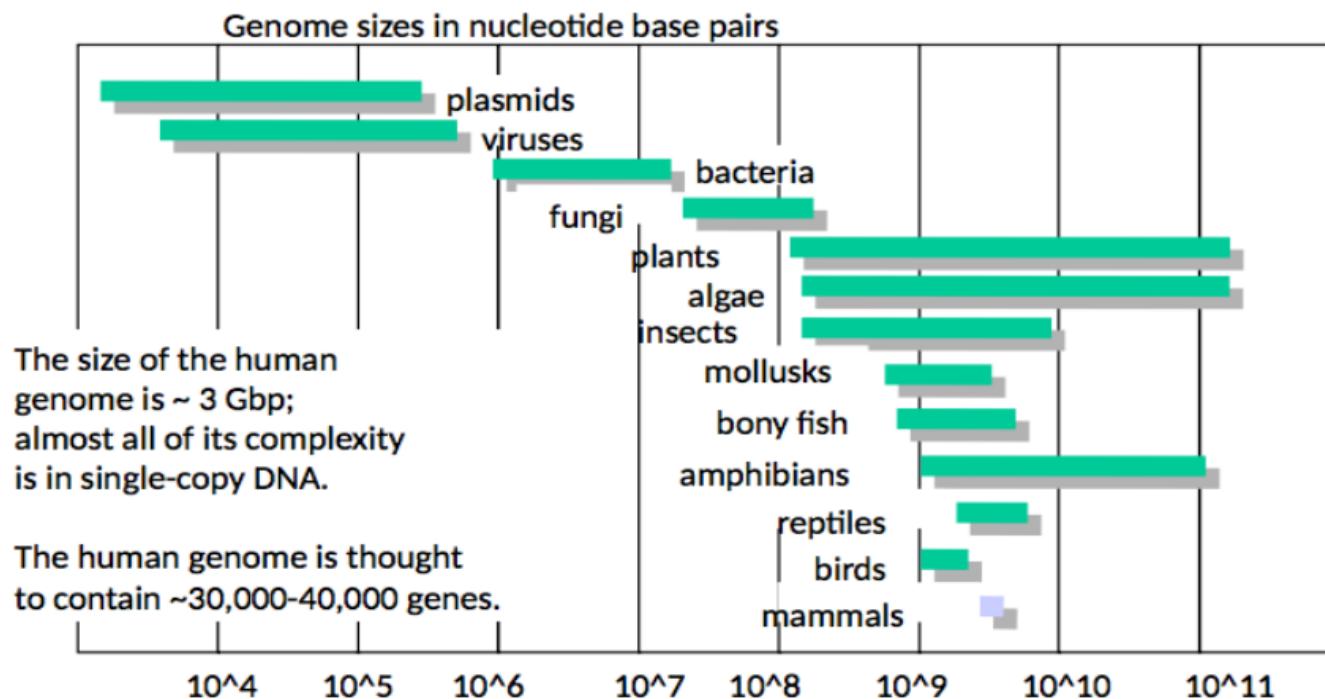
National Human  
Genome Research  
Institute

[genome.gov/sequencingcosts](http://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 dólares
- < 1 semana para ser completo.



# 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 análises de 150-300 genomas por dia (2012)



# Dados

Platform	3730XL	454 FLX	454 GS JR	HiSeq 2000	MiSeq	SOLID 5500	IonTorrent	PacBio RS
Method of amplification	Clonal plasmid amplification	emPCR on beads	emPCR 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 <sup>+</sup> 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

# Formato de arquivos

```
AAATAAAAATTTTAACTCTAACGATGTCGTT  
ILLUMINA-GA_00001:1:4010:1065#0/1  
nhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh  
ILLUMINA-GA_00001:1:4099:1065#0/1  
AAATAACTAAGAATTTCACAAATTCTAAATTCTT  
ILLUMINA-GA_00001:1:4099:1065#0/1  
fffffgegggaffccfd_ffcdffgffgcgggfggg  
carbonell@bender:/scratch2/jcarbonell$  
carbonell@bender:/scratch2/jcarbonell$ head ivia15_06_pair2.rendup.fq -n 20  
ILLUMINA-GA_00001:1:1395:1061#0/2  
GGACCAAGCAAGACAATGCTAAATTCTGCAGAGATA  
ILLUMINA-GA_00001:1:1395:1061#0/2  
caehghce_Wfffffa[ffcfgfhgeahewff  
ILLUMINA-GA_00001:1:1855:1066#0/2  
TTTAAATCCCTGTGCGTGTATGTGATGCCATOCA  
ILLUMINA-GA_00001:1:1855:1066#0/2  
ffffcffffdhhdcffffdfffflcc ``^` dfffccha  
ILLUMINA-GA_00001:1:3567:1062#0/2  
GAGTCGGCGCGAACGTCGCCAGCCCCACCCCA  
ILLUMINA-GA_00001:1:3567:1062#0/2  
nhhhhhhhhhhhcgfcff]fdffs[leffchhhh  
ILLUMINA-GA_00001:1:4010:1065#0/2  
TGTGACAGTTAATGATGGTCTATTACATAACAGT  
ILLUMINA-GA_00001:1:4010:1065#0/2  
nhhhhhhhhhhhhhhhhhhhhhhhhhhhhhfhe  
ILLUMINA-GA_00001:1:4099:1065#0/2  
ATCCAAAGACAAACAGTTCCAAGAGATGCAAGGAC  
ILLUMINA-GA_00001:1:4099:1065#0/2  
fffffdhdddhhhhggfffhhcghg_f0fbffffdf  
carbonell@bender:/scratch2/jcarbonell$  
carbonell@bender:/scratch2/jcarbonell$ samtools view ivia15_06_pair1.rendup_bwa_bwa_ref01_upper_mapped.bam | head -n 10
```

fastq: sequence data and qualities



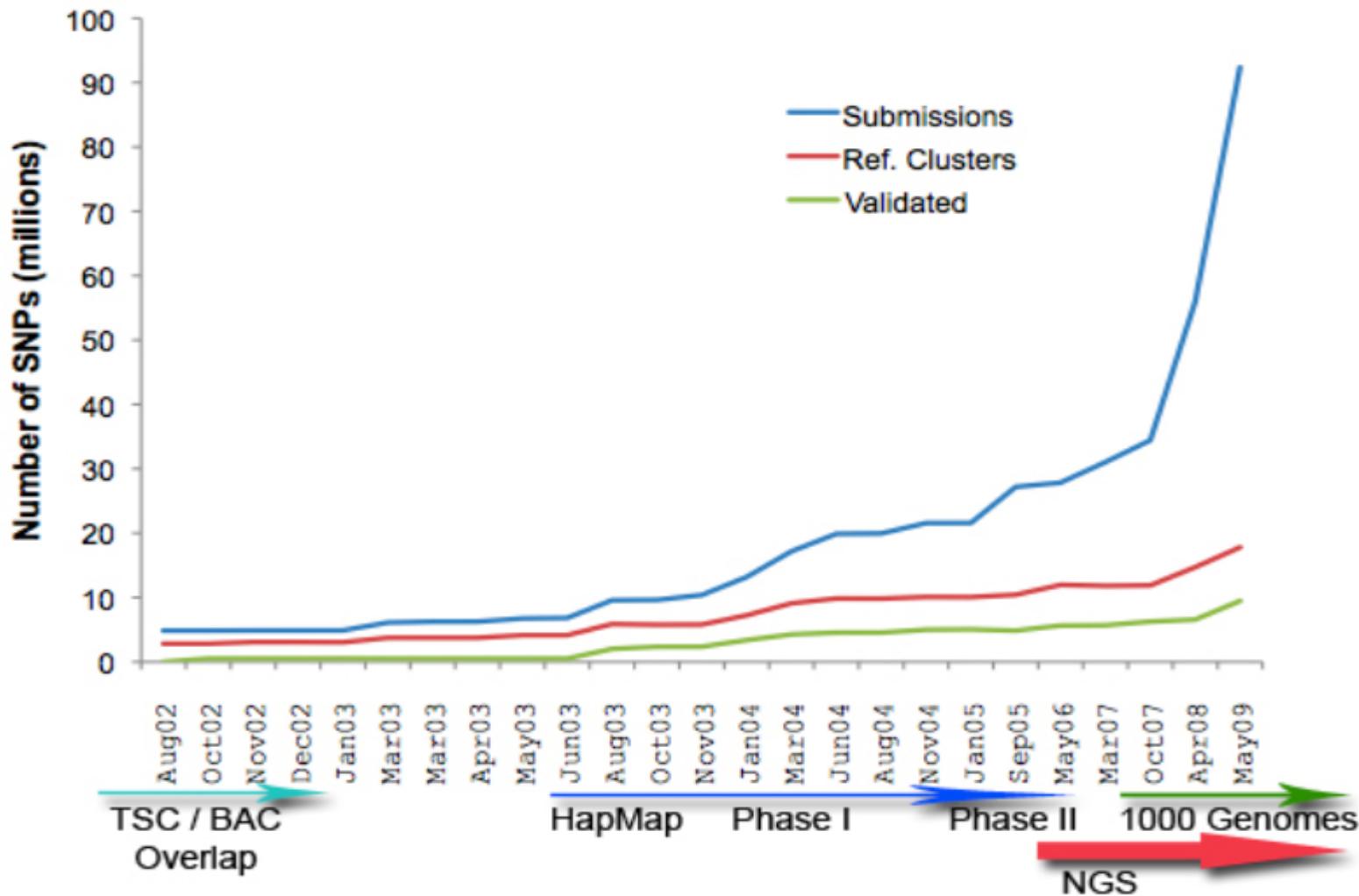
SAM/BAM: mapping data and qualities



# Análise Secundária

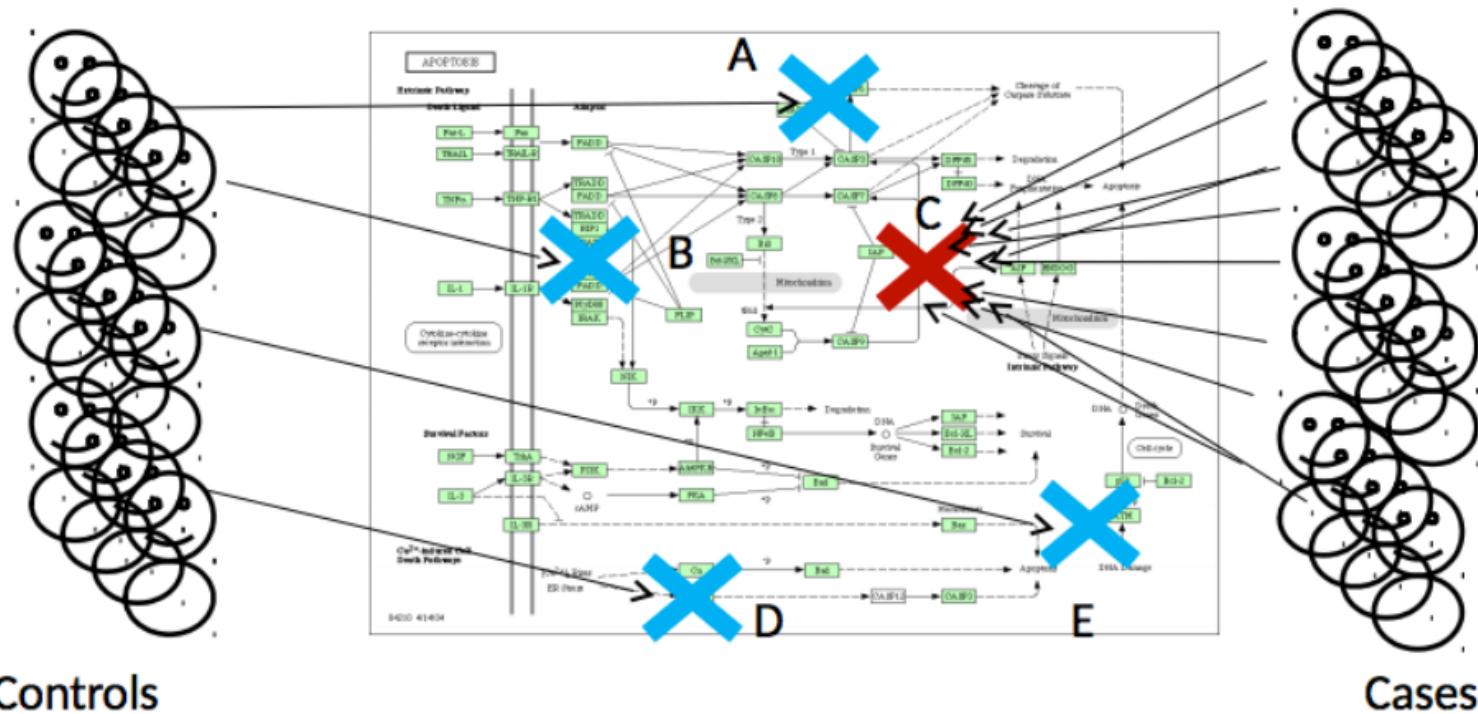
## Detectar as mutações associadas à doença

Growth of dbSNP, 2002-2009



# Análise Secundária

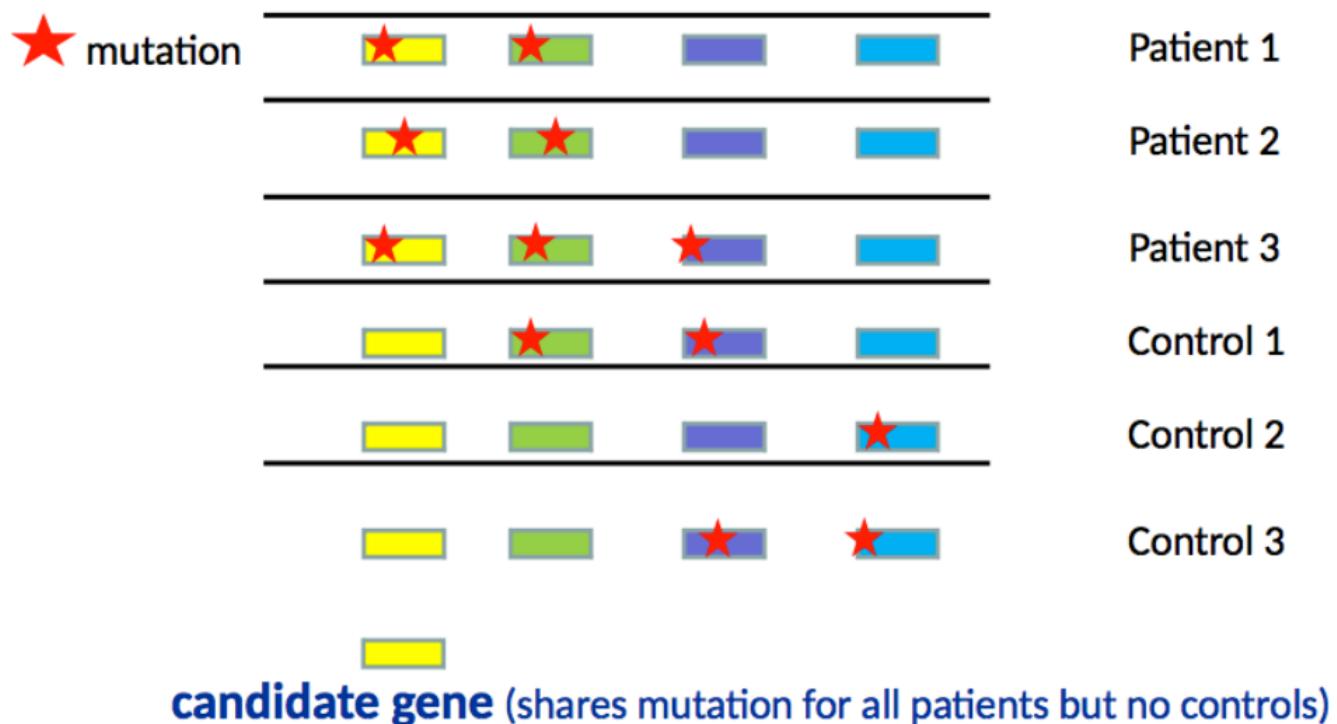
## Detectar as mutações associadas à doença



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

# Análise Secundária

## Detectar as mutações associadas à doença



# Análise Secundária

## Detectar as mutações associadas à doença

NATURE | ARTICLE OPEN

◀ previous article next article ▶

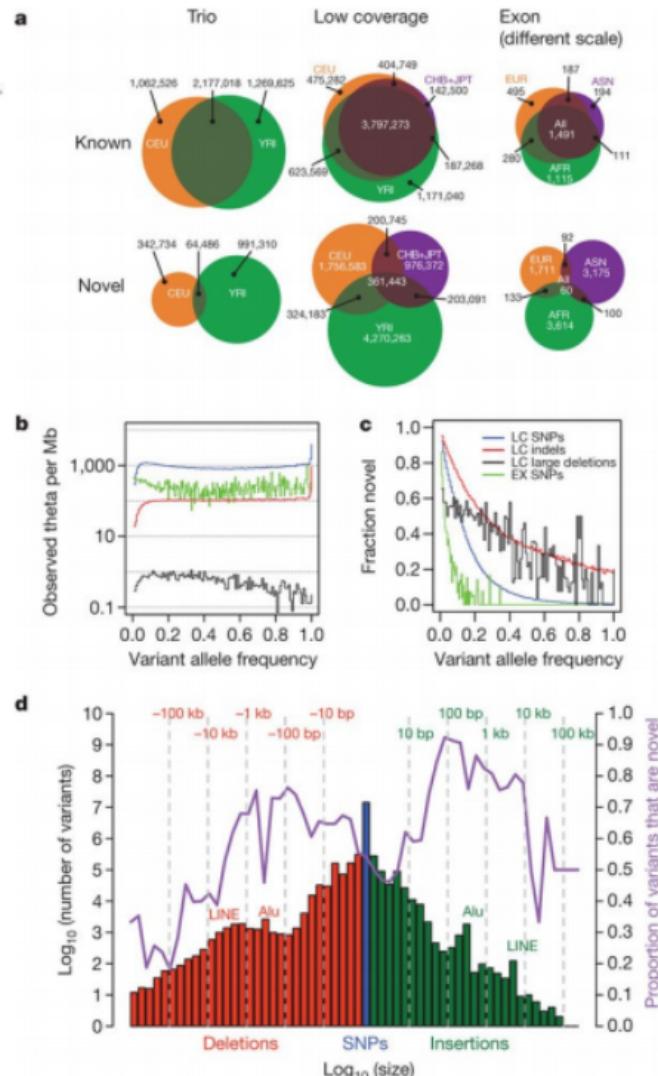
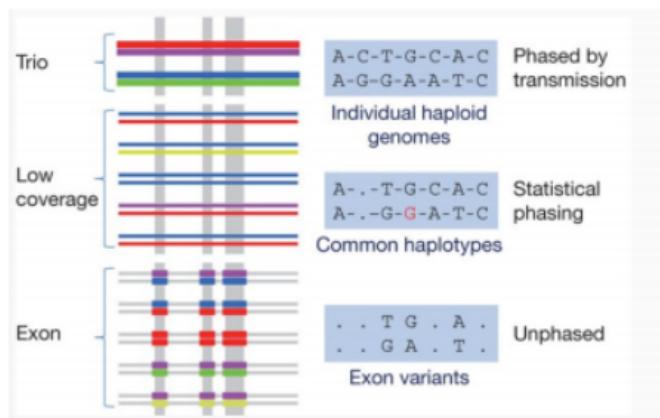
### A map of human genome variation from population-scale sequencing

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



# Análise Secundária

## Detectar as mutações associadas à doença

Published in final edited form as:

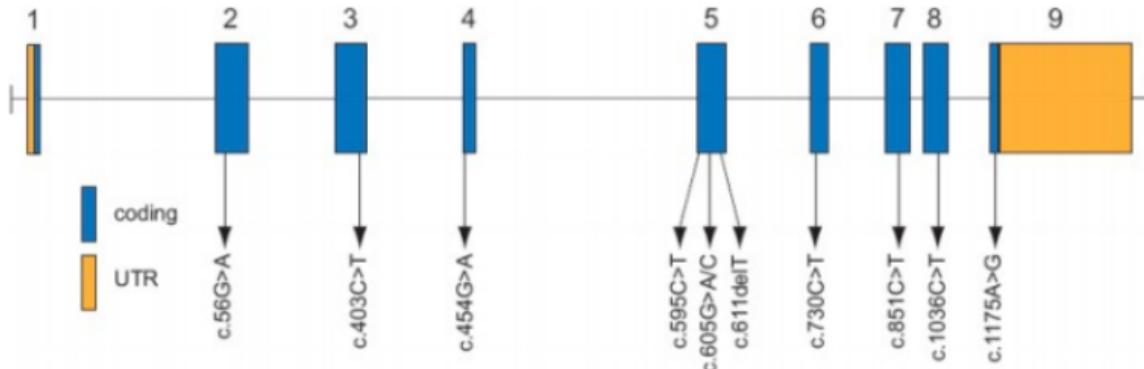
*Nat Genet.* 2010 January ; 42(1): 30–35. doi:10.1038/ng.499.

### Exome sequencing identifies the cause of a Mendelian disorder

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

<sup>1</sup>Department of Genome Sciences, University of Washington, Seattle, Washington, USA

<sup>2</sup>Department of Pediatrics, University of Washington, Seattle, Washington, USA <sup>3</sup>Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA <sup>4</sup>Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA <sup>5</sup>Institute of Systems Biology, Seattle WA, USA <sup>6</sup>Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA <sup>7</sup>Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland <sup>8</sup>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 (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.



**Miller syndrome**

# 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**



# Introdução ao NGS

Rodrigo Bertollo  
[rodrigo@genomika.com.br](mailto:rodrigo@genomika.com.br)