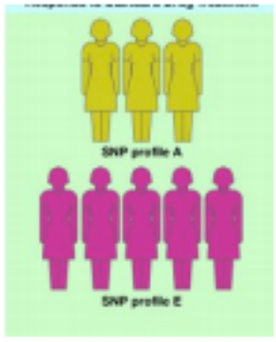
GENOMIKA

Dando boas vindas!

Marcel Caraciolo, CTO
marcel@genomika.com.br

Cenário Atual

Genes in the DNA...



...produces the final phenotype

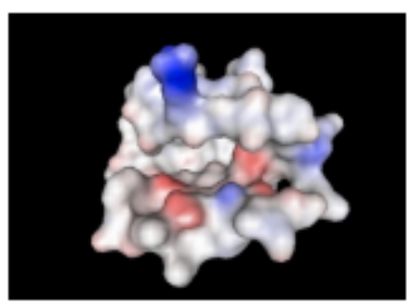
...code for proteins...

>protein kinase

acctgttgatggcgacagggactgta
tgctgatctatgctgatgcatgcatgc
tgactactgatgtgggggctattgac
ttgatgtctatc....

From genotype to phenotype.

...whose structure accounts for function...



...plus the environment...

Data is information

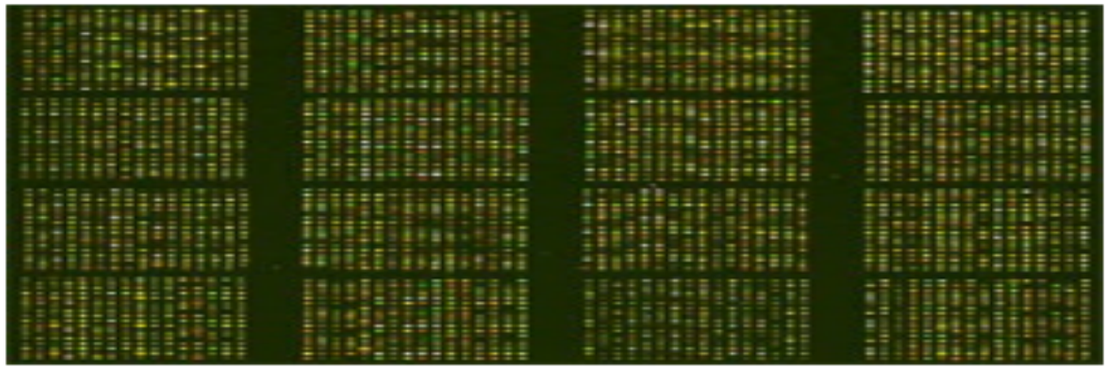
GENOMIKA

DNA sequencing

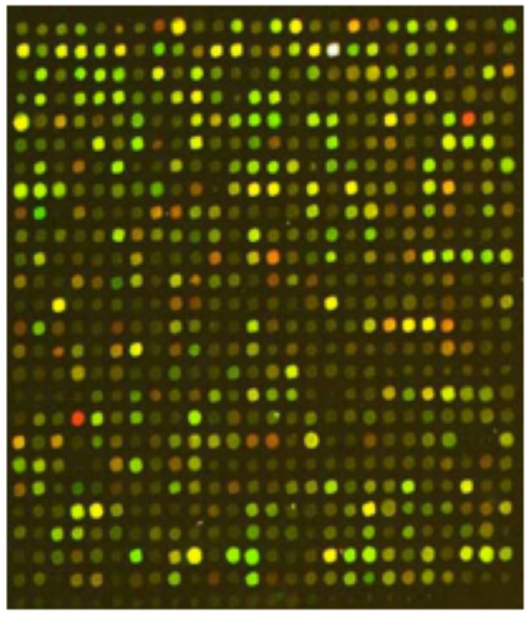
Reading the order of bases in DNA fragments

High Throughput Technologies

- **1998** arrayed DNAs were used
- **1991** oligonucleotides are synthesized on a glass slide through photolitography (Affymax Research Institutes)
- **1995** DNA Microarrays
- **1997** Genome wide Yeast Microarray



Nature Milestones DNA Technologies



GENOMIKA

High Throughput Technologies

by producing **massive** amounts of sequence data, really **fast**

GENOMIKA

High Throughput Technologies

Next Generation Sequencing
SOLID **12Gbp** per round

Genes in the DNA…

>protein kunase

acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgcatgctgactactga
tgtgggggctattgacttgatgtctatc….

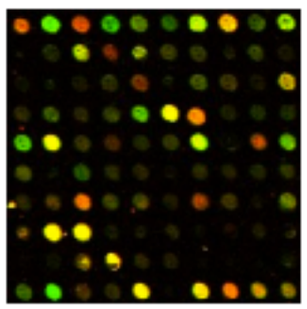
…when expressed in the proper moment and place…

…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype…

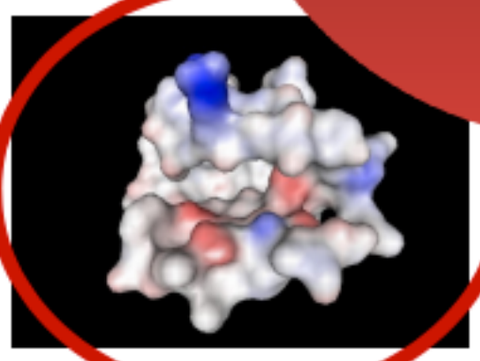
A typical tissue is expressing among **5,000 and 10,000 genes**



Data ≠ Information

…conforming complex interaction networks…

…code for proteins…



That undergo post-translational modifications, somatic recombination…

100K-500K proteins

…whose structures account for function…

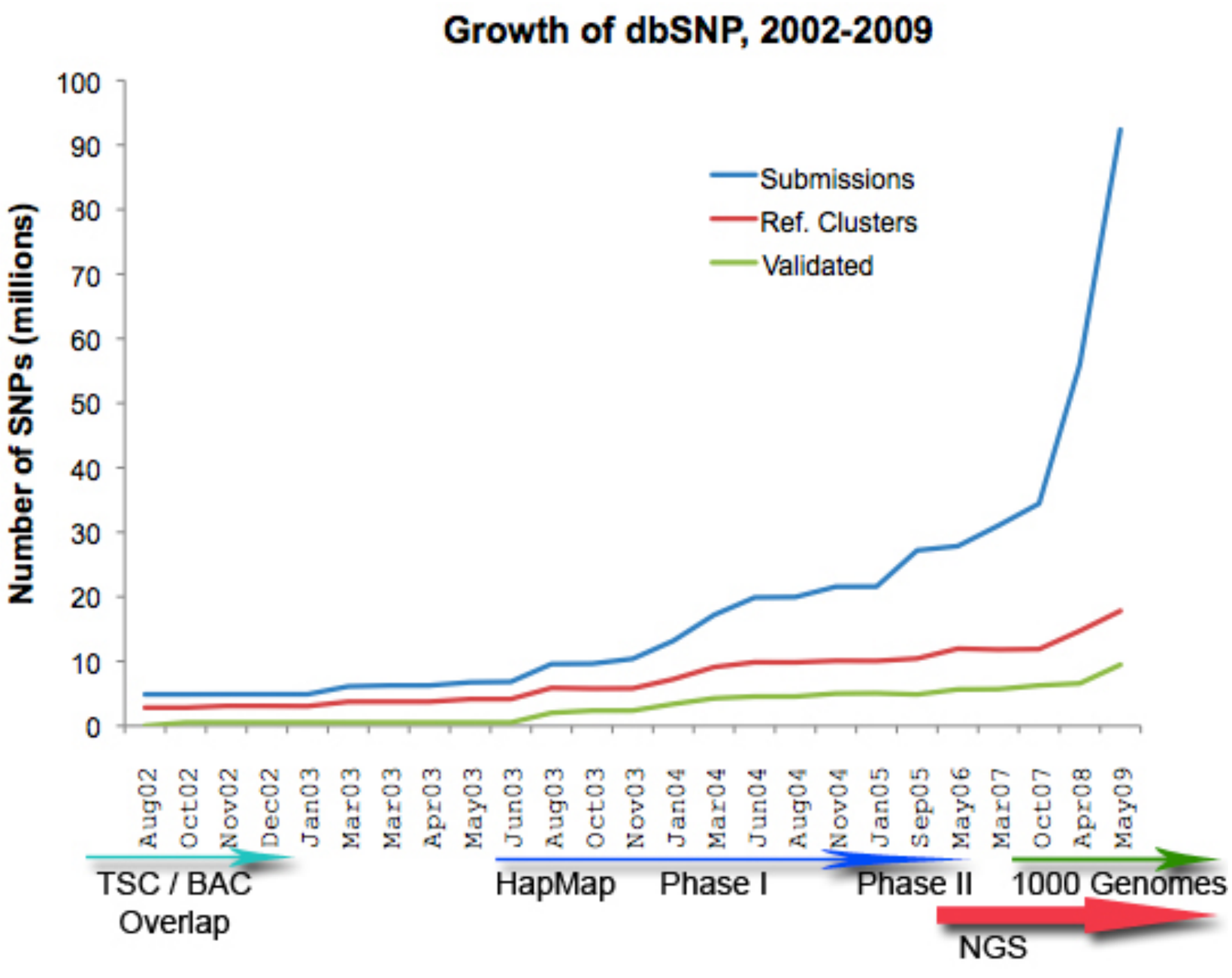
Each protein has an average of **8** interactions

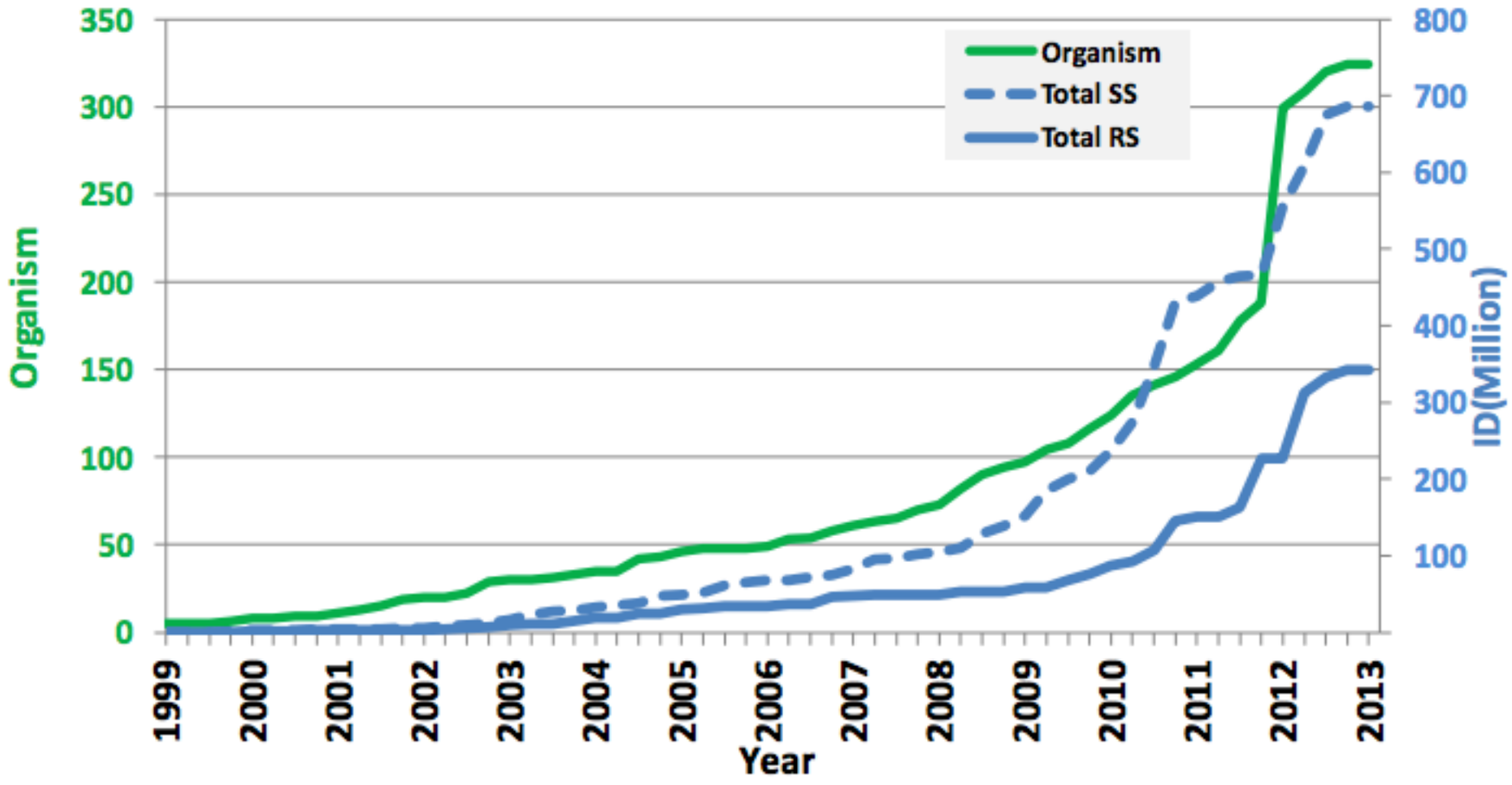
…in cooperation with other proteins…

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

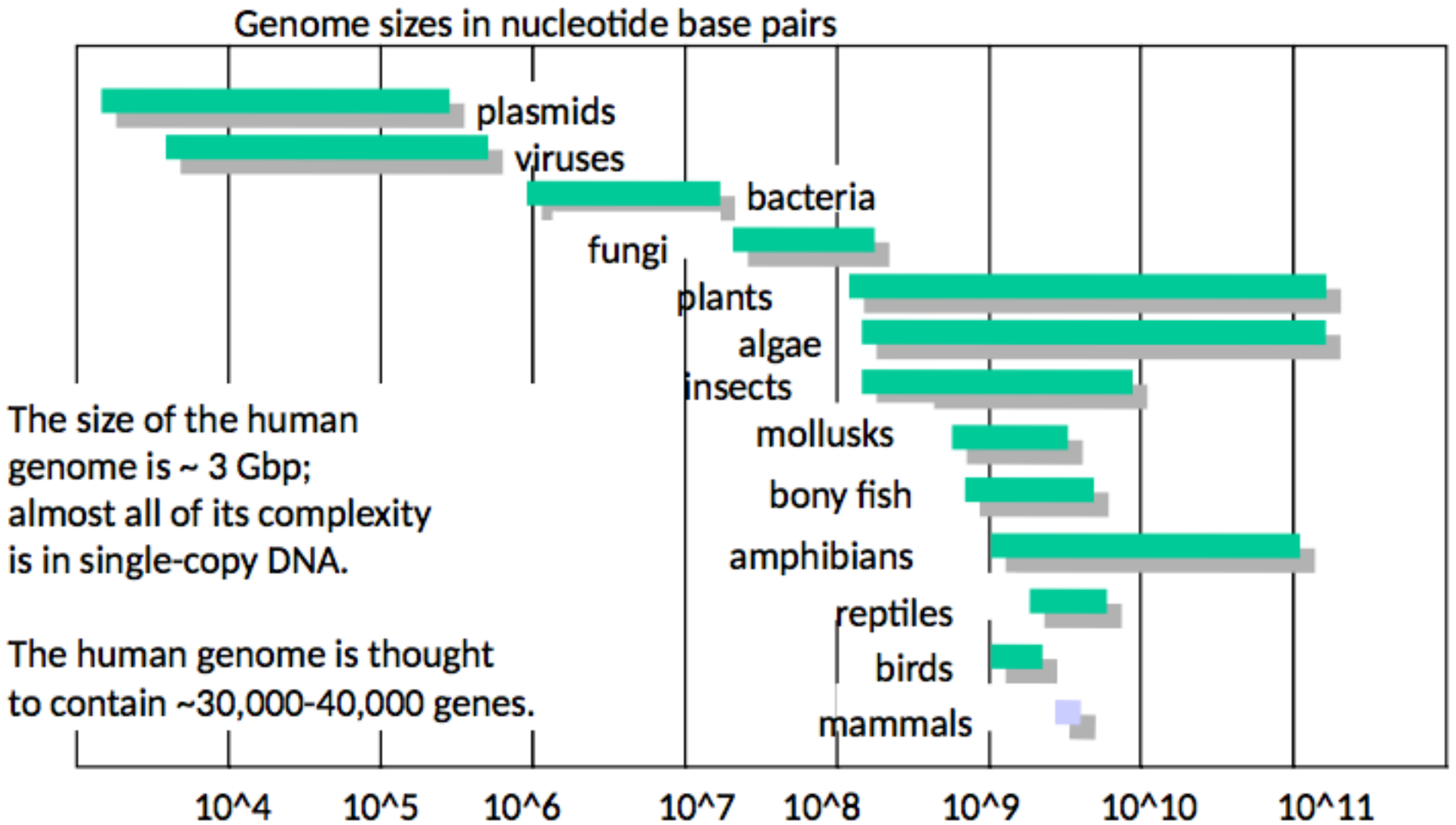
National Human Genome Research Institute

genome.gov/sequencingcosts


Growth of dbSNP, 2002-2009
Number of SNPs (millions)
0
10
20
30
40
50
60
70
80
90
100
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



- 300+ organisms including mammals, birds, fish, plants, and bacteria
- 700+ million Submitted SNP (ss)
- 300+ million Reference SNP (rs)
- ~75% are non-human



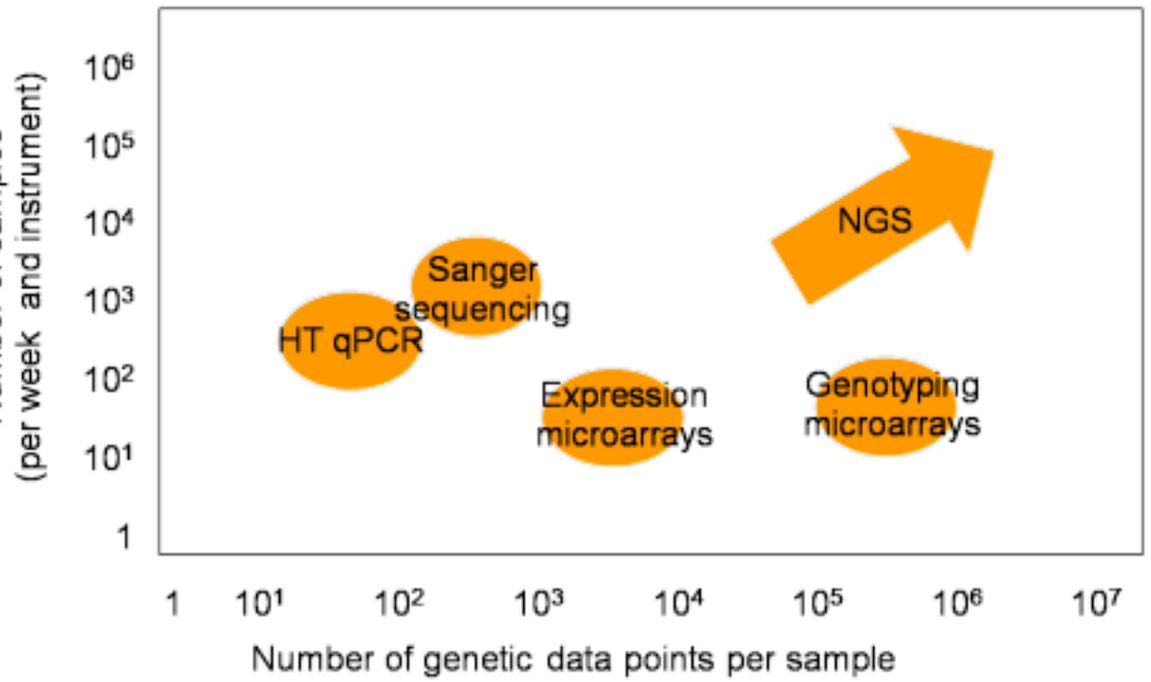
http://www3.kumc.edu/jcalvet/PowerPoint/bioc801b.ppt

Computing capabilities (*CPU power doubles in ~18-24 moths, hard drive capacity doubles in ~12 moths, network bandwidth doubles in ~20 moths*) should increase : **7-10x** in 5 years. Follows **Moors's law**

Data projection in 3-5 years: **100x** increase in sequencing volume. Still new technologies with higher throughput to come very soon !!!

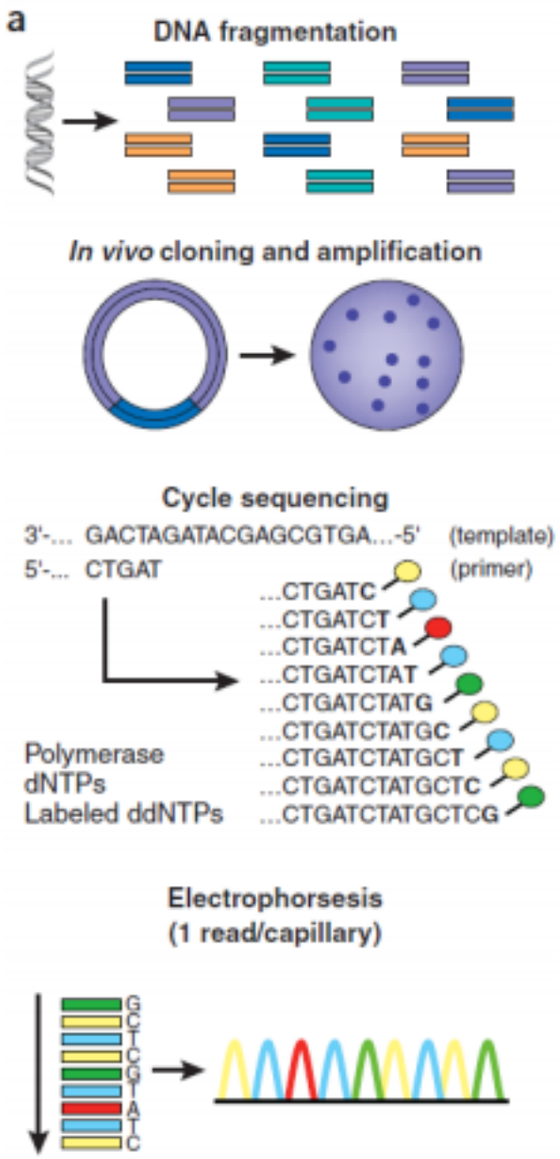
Relative throughput of the different HT technologies

NGS emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming



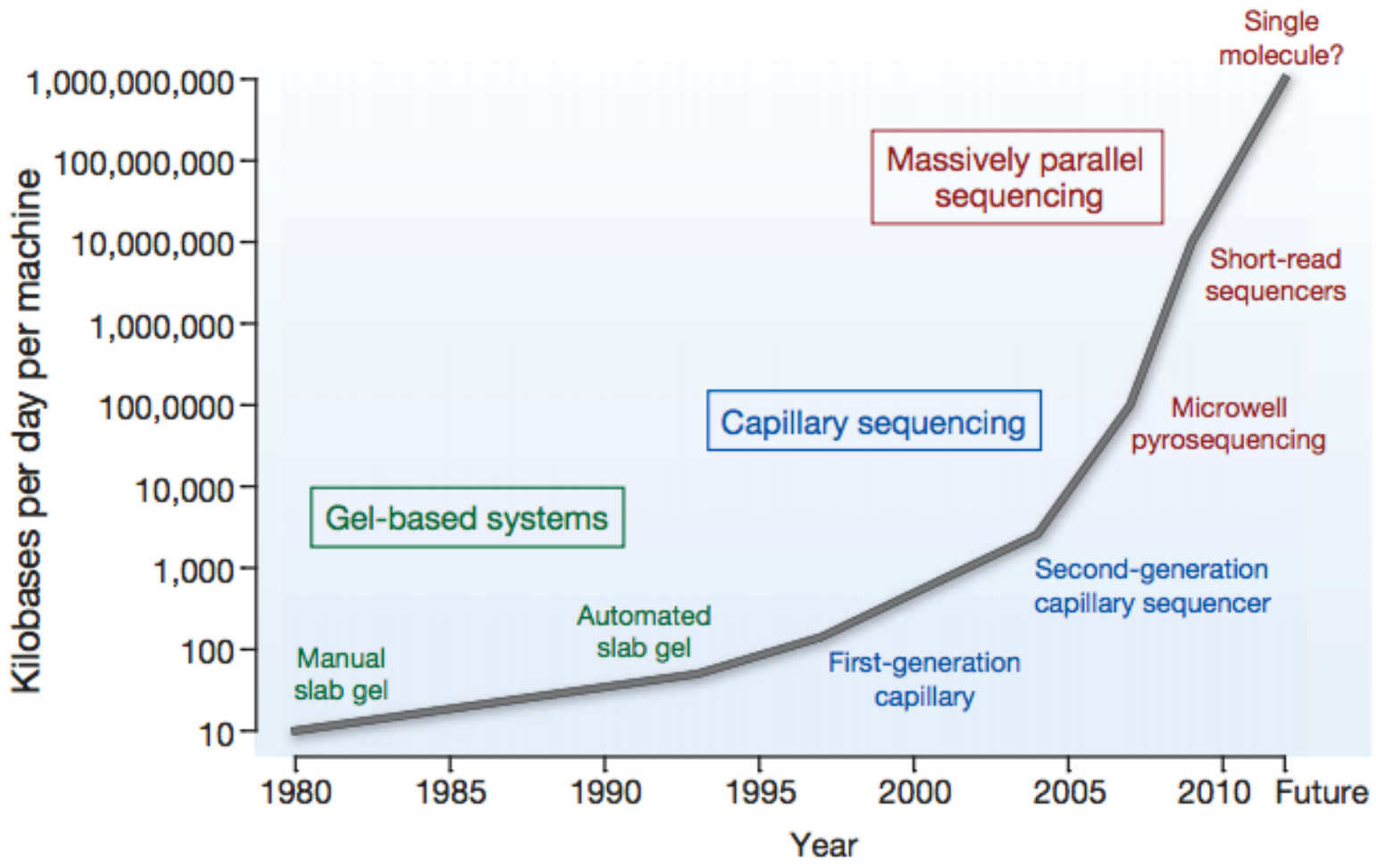
Too many sequences to be handled in a standard computer

First Generation: Sanger



- Fragment DNA
- Clone into plasmid and amplify
- Sequence using dNTP + labelled ddNTPs (stops reaction)
- Run capillary electrophoresis/ gel and "read" DNA code
- Low output, long reads (~300-1000 nt), high quality

First Generation to NGS



1977 - Sanger
Chain-termination
method

Illumina

Solid

Ion Torrent

454

Pacific Biosciences

Oxford Nanopore

GENOMIKA

Human Genome

- First draft genome of human in 2001, final 2004
- Estimated costs $3 billion, time 13 years

- Today:
- Illumina: <1 week, ~1000$
- Exome: <6 weeks*, $300-400

- The 1000$ genome is here?

GENOMIKA

Storage And Analysis

Highest cost is (almost) not the sequencing
but
storage and analysis

A standard human (30-40x) whole-genome sequencing exp. would create 150 Gb of data

BGI, based in China, is the world's largest genomics research institute, with 167 DNA sequencers producing the equivalent of 150-300 human genomes a day. (2012)

NGS & Bioinformatics

- Extreme data size causes problems
  - Just transferring and storing the data
  - Standard comparisons fail (N*N)
  - Standard tools can not be used
  - Think in fast and parallel programs

GENOMIKA

Roche 454 SEQUENCING

Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads

illumina®

life™

ion torrent
by life technologies

PACIFIC BIOSCIENCES™

GENOMIKA

Differences between the various platforms

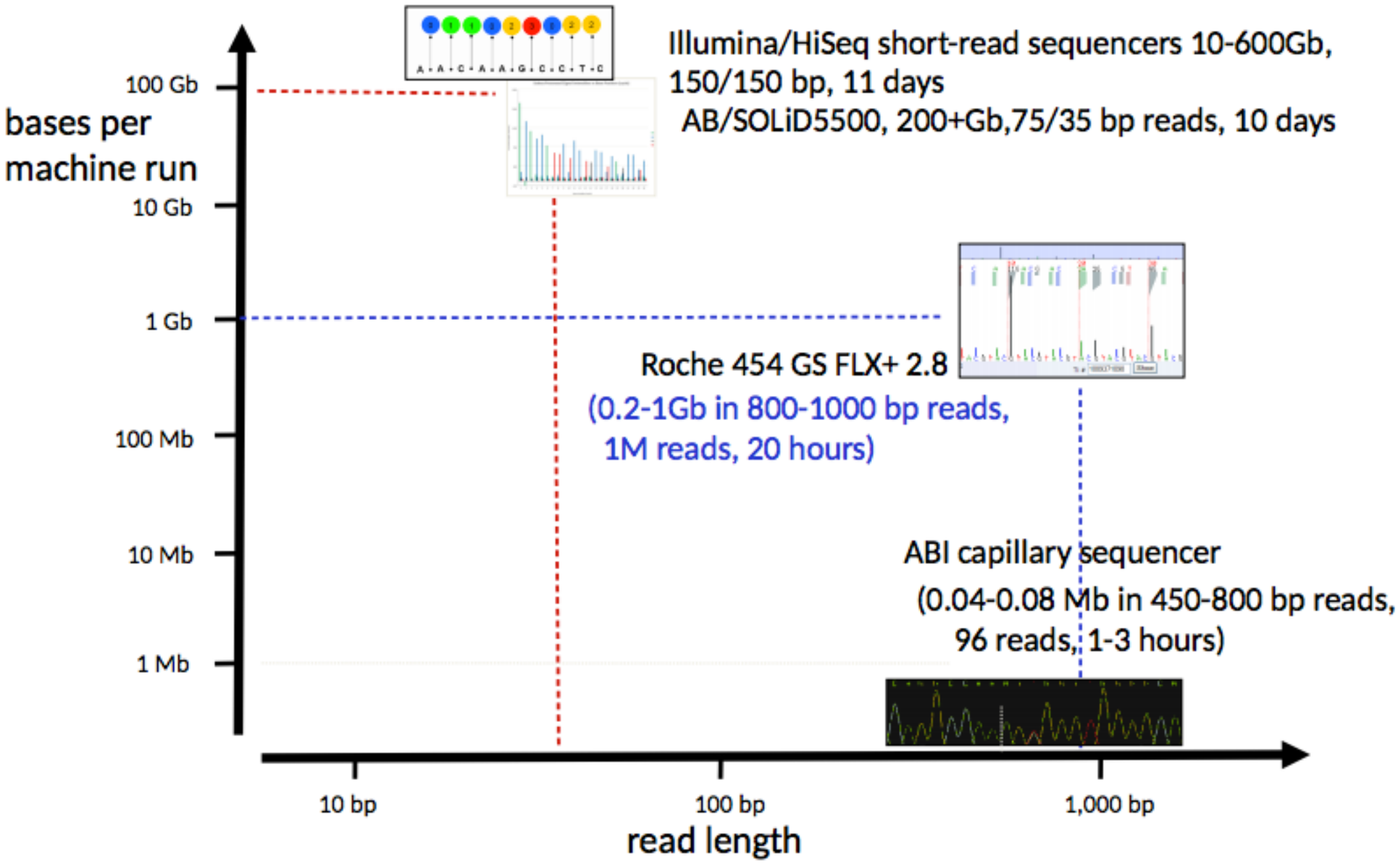
- Nanotechnology used.
- Resolution of the image analysis.
- Chemistry and enzymology.
- Signal to noise detection in the software.
- Software/images/file size/pipeline
- Cost
- Applications

GENOMIKA

Similarities - LOTS of Data
General ways of dealing at the sequences

- Assemble them and look at what you have.
- You map them (align against a known genome) and then look at what you have.
- Or a mixture of both!
- Somes you select the DNA you are sequencing or you try to sequence everything
- Depends on biological question, sequencing machine you have, and how much time and money you have.
- **NGS is relatively cheap but think what you want to answer, because analysis will not do magic.**

NextGen Sequencers



NextGen Sequencers

3 main platforms:

- **Solexa/illumina**
- **Roche 454**
- **ABI SOLiD**

- Follow an approach similar to Sanger sequencing, but do away with separation of fragments by size and "read" the sequence as the reaction occurs
- Several different "next generation" sequencing platforms developed and commercialized, more on the way.
- Simultaneously sequence entire libraries of DNA sequence fragments

Roche 454

- First next generation method to be commercially available

- Uses a "sequencing by synthesis" (SBS) approach:
    - DNA is broken into pieces of 500-1,400 bp, ligated to adaptors, and amplified on tiny beads by PCR (emulsion PCR)
    - Beads (with DNA attached) are placed into tiny wells (one bead per well) on a PicoTiter Plate that has millions of wells. Each well is connected to an optical fibre.
    - DNA is sequenced by adding polymerase and DNA bases containing pyrophosphate. The different bases (A,C,G,T) are added sequentially in a flow chamber
        - When a base complementary to the template is added, the pyrophosphate is released and a burst of light is produced
        - The light is detected and used to call the base

- Initially 100-150 bp, but they have been improved to 600-1000 bp

- >1 million, filter-passed reads per run (20 hours)

- 1 billion bases per day

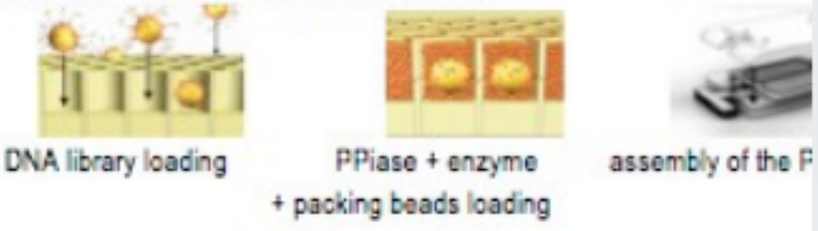
Roche 454 pyrosequencing

Principel

Preparation of the DNA includes : DNA fragmentation (nebulization), DNA size selection, Fragment end polishing, Adaptor ligation, Library immobilization, fill in reaction and ssDNA library isolation. At the end of these steps, the DNA fragments are ready for the emulsion PCR (emPCR).

emPCR include the immobilisation of the DNA fragments on ca
indirect enrichment resulting in an immobilized and amplified lib

Sequencing includes a prewash, the loading DNA library bead
end of these steps you get your data.



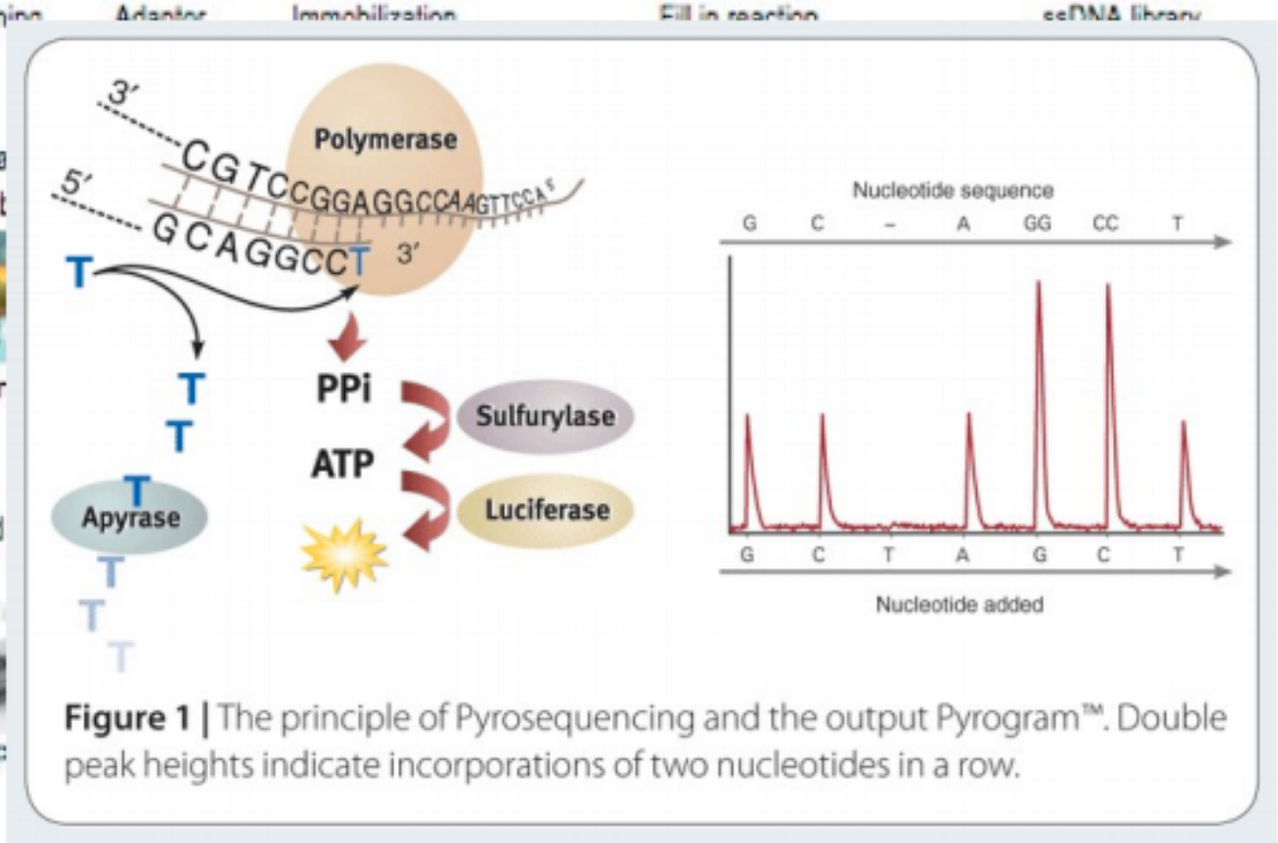
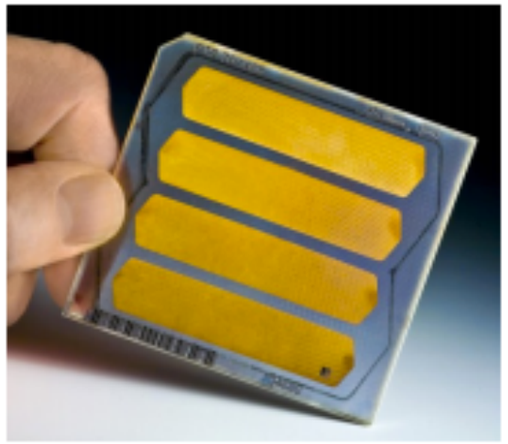


Figure 1 | The principle of Pyrosequencing and the output Pyrogram™. Double peak heights indicate incorporations of two nucleotides in a row.

Roche 454 pyrosequencing

- Good for
  - “de novo“ sequencing (longer reads).
  - Resequencing (expensive)
    - New bacterial genomes.
    - Amplicons
- Pyrosequencing. Bias with long polinucleotide streches



Roche 454

Throughput	400-600 million high-quality, filter-passed bases per run* 1 billion bases per day
Run Time	10 hours
Read Length	Average length = 400 bases
Accuracy	Q20 read length of 400 bases (99% at 400 bases and higher for prior bases)
Reads per run	>1 million high-quality reads
Data	Trace data accepted by NCBI since 2005
Computing Requirements	Cluster recommended (Roche GS FLX Titanium Cluster available)
Robustness	No complex optics or lasers; reagents have long shelf life

GENOMIKA

Roche 454 pyrosequencing

System Performance	
Throughput	35 million high-quality, filtered bases per run*
Run Time	10 hours sequencing 2 hours data processing
Avg. Read Length	400 bases*
Accuracy	Q20 read length of 400 bases (99% accuracy at 400 bases)
Reads per Run	100,000 shotgun, 70,000 amplicon
Sample Input	gDNA, amplicons, cDNA, or BACs depending on the application
Physical Dimensions	40 cm wide x 60 cm deep x 40 cm high (the size of a laser printer) Weight = 55 lbs.
Computing	Linux-based OS on HP desktop computer included. All software is point-and-click.

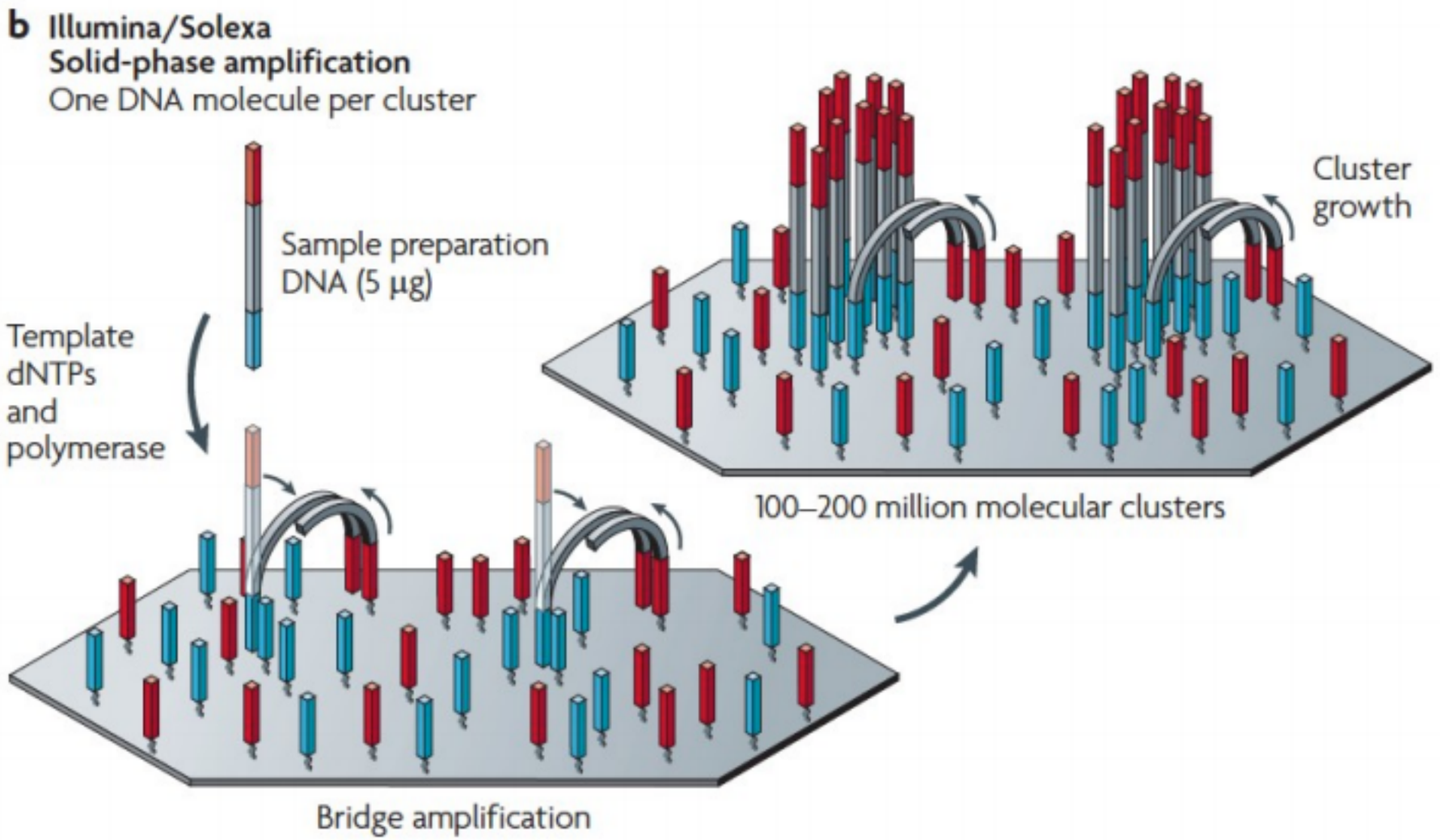
***Typical results**. Average read length and number of reads depend on specific sample and genomic characteristics*

GENOMIKA

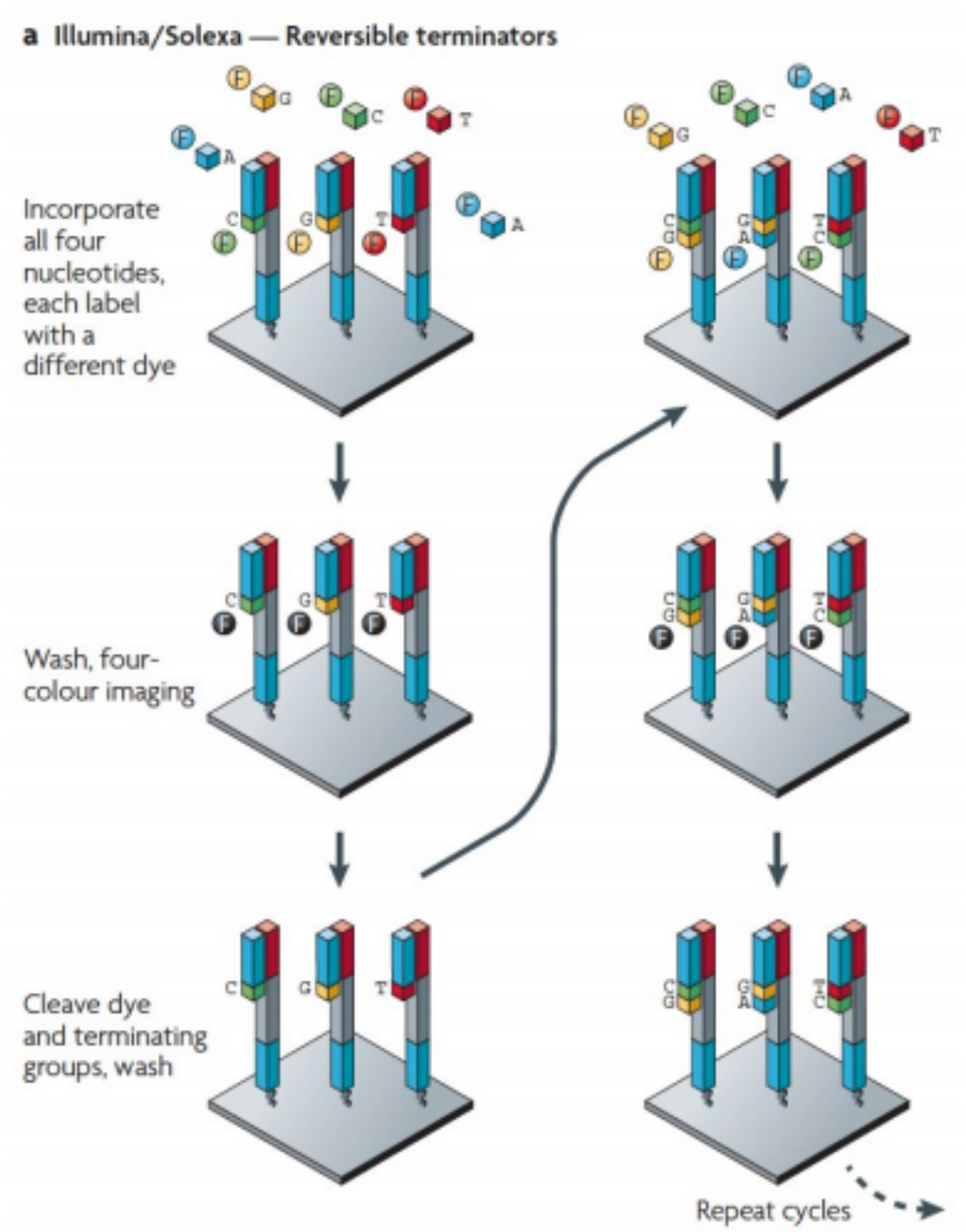
Illumina

- Over 90% of all sequencing data is produced on Illumina systems.
- Uses a "sequencing by synthesis" approach:
  - DNA is broken into small fragments and ligated to an adaptor.
  - The fragments are attached to the surface of a flow cell and amplified.
  - DNA is sequenced by adding polymerase and labeled reversible terminator nucleotides (each base with a different color).
  - The incorporated base is determined by fluorescence.
  - The fluorescent label is removed from the terminator and the 3' OH is unblocked, allowing a new base to be incorporated
- Started with 35 bp, increased now to up to 150 bp
- One run can give up to 10-600 Gb, 300-6000 million paired-end reads
- 75-85% of bases at or above Q30

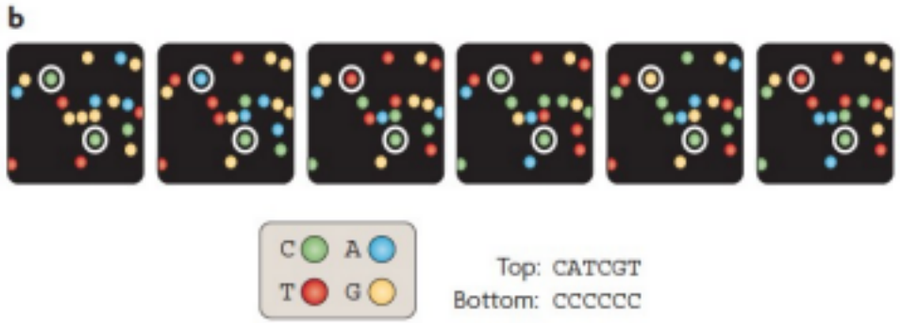
Illumina


b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation
DNA (5 µg)
Template
dNTPs
and
polymerase
Cluster
growth
100–200 million molecular clusters
Bridge amplification

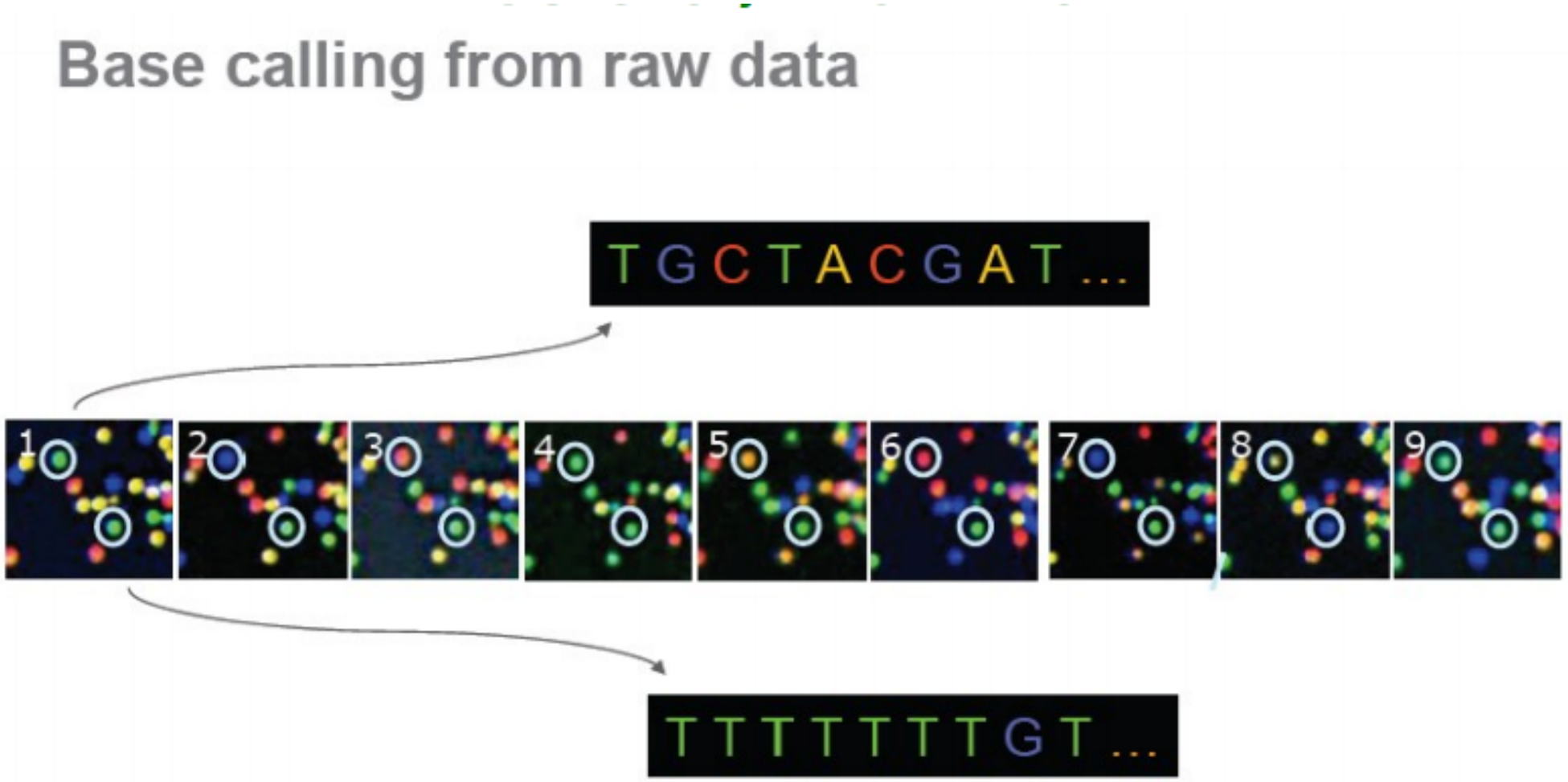
Illumina



Solexa / illumina



Illumina



From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4

The identity of each base of a cluster is read off from sequential images

Illumina

Illumina-HiSeq 2500

600 Gb/run in 11 days
2x100 bp fragments
6 billion reads per run

Illumina

Illumina-MiSeq

175-245 Mb 4h 1x 36bp

1.5-2.0 Gb 27h 2x150 bp

GENOMIKA

SOLID (ABI/ Life Technologies)

- **Colorspace**
- "**sequencing by ligation**" method
- Does not use polymerase, instead uses DNA ligase for sequencing:
  - DNA is broken into small fragments and ligated to an adaptor.
  - The fragments are attached to beads and amplified by emulsion PCR. Beads are attached to the surface of a glass slide.
  - DNA is sequenced by adding 8-mer fluorescently labelled oligonucleotides
  - If an oligo is complementary to the template, it will be ligated and 2 of the bases can be called.
  - The attached oligo is then cut to remove the label and the next set of labelled oligos are added
  - The process is repeated from different starting points (using different universal primers) so that each base is called twice
- 200 Gb, 1.8 billion reads per run, 35bp-75bp, 10 days

SOLID (ABI/ Life Technologies)

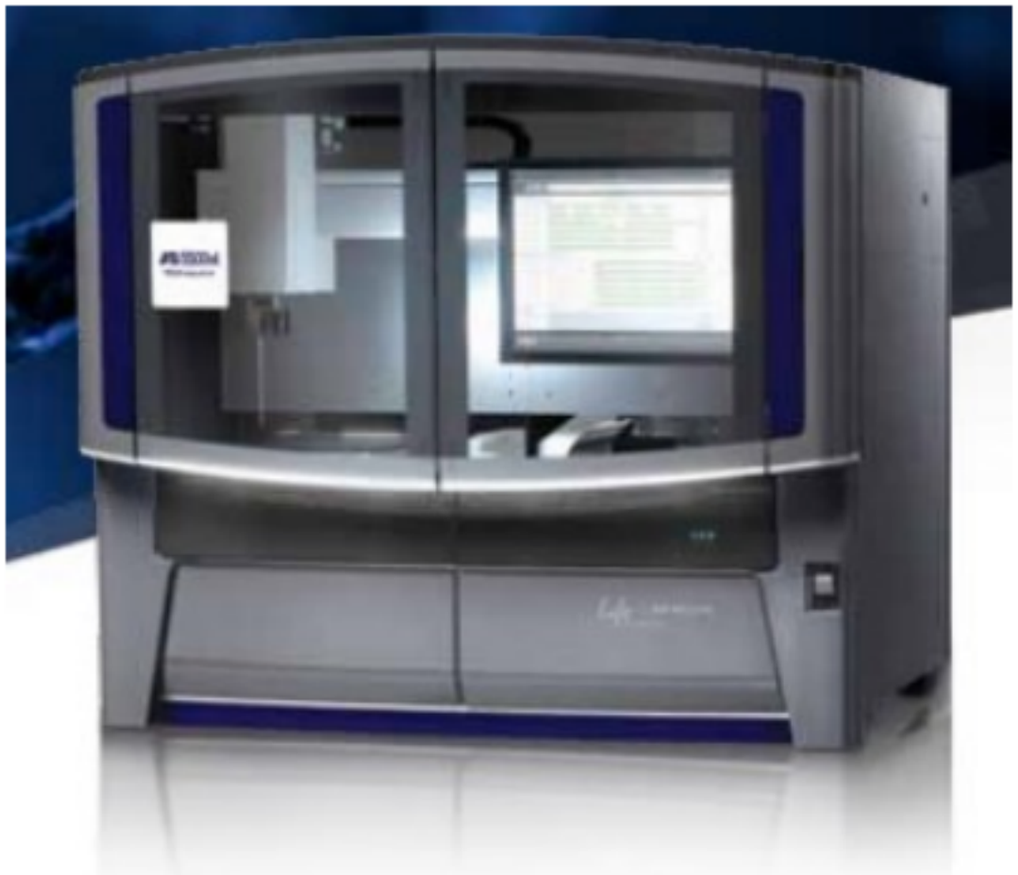
200 Gb/run (microbeads)
300 Gb/run (nanobeads)

35-75 bp fragments

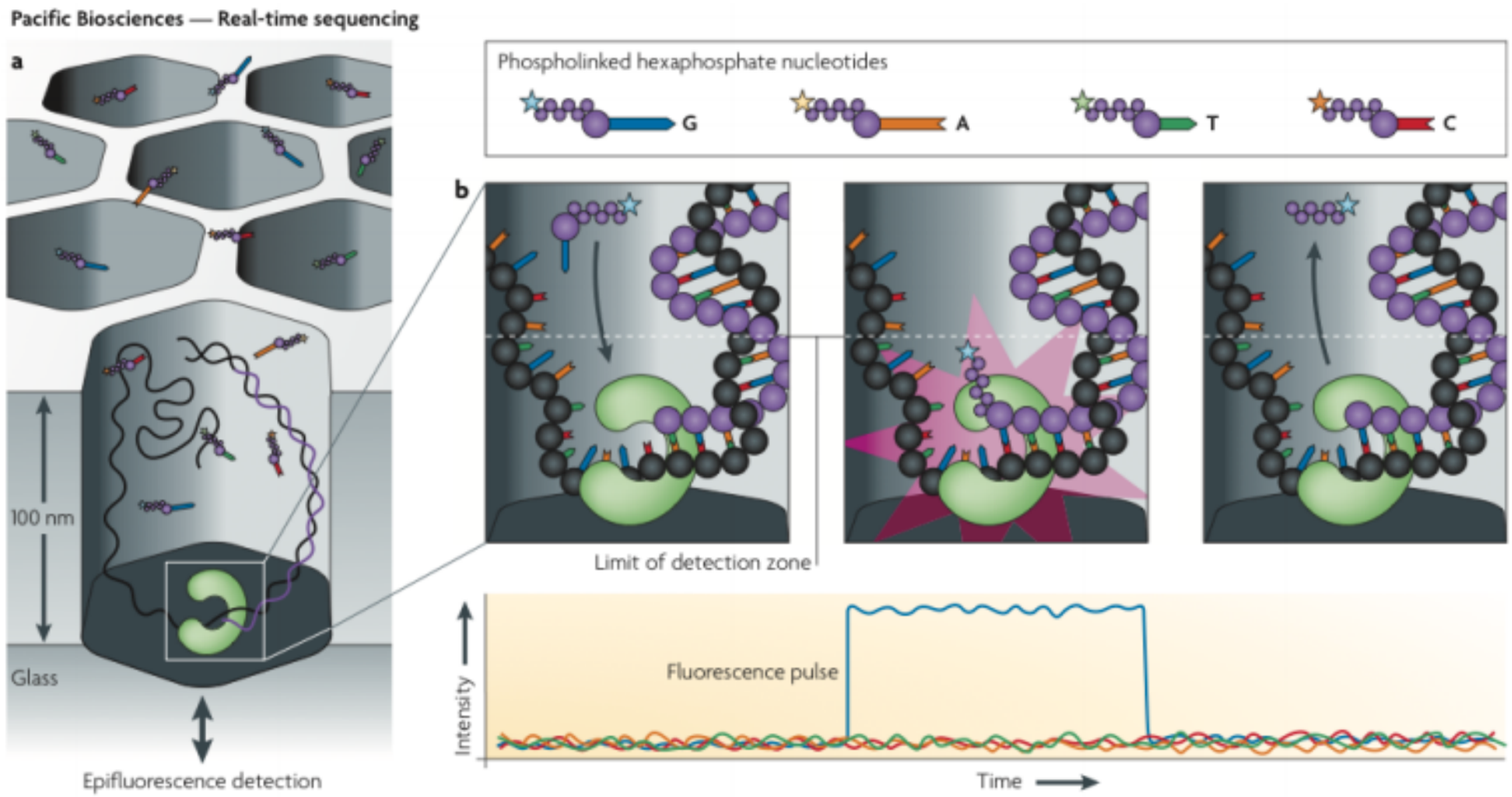
1.8 - 4.8 billion reads/run

2x6 lanes/run
96 bar-codes

ECC: 99.99% accuracy



Pacific BioScience



Slowed down DNA polymerase, measure light emission,
Long reads > 10kb, high error rate (but random)

GENOMIKA

Comparison

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

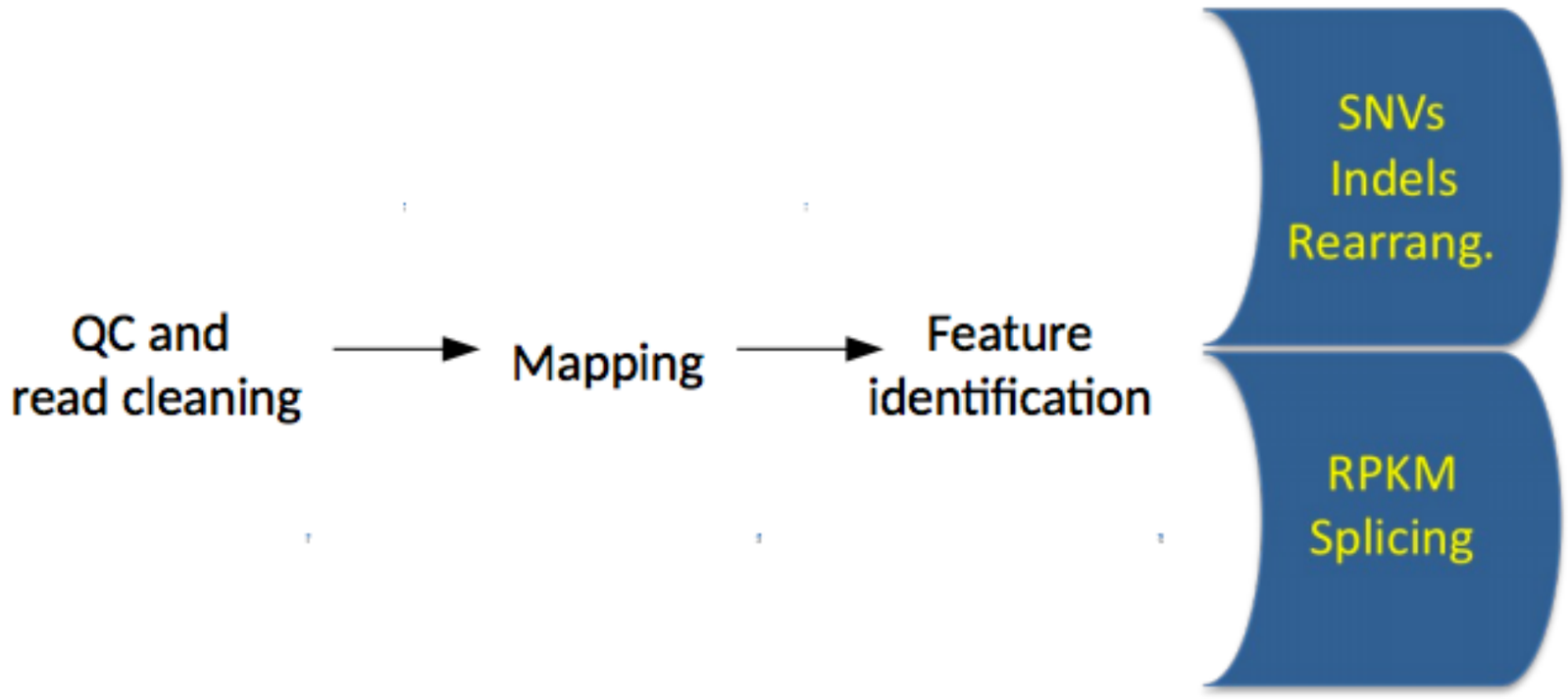
- Short fragments
- Color-space
- High throughput
- Cheap

- Resequencing
- ChipSeq
- RNASeq
- MethylSeq

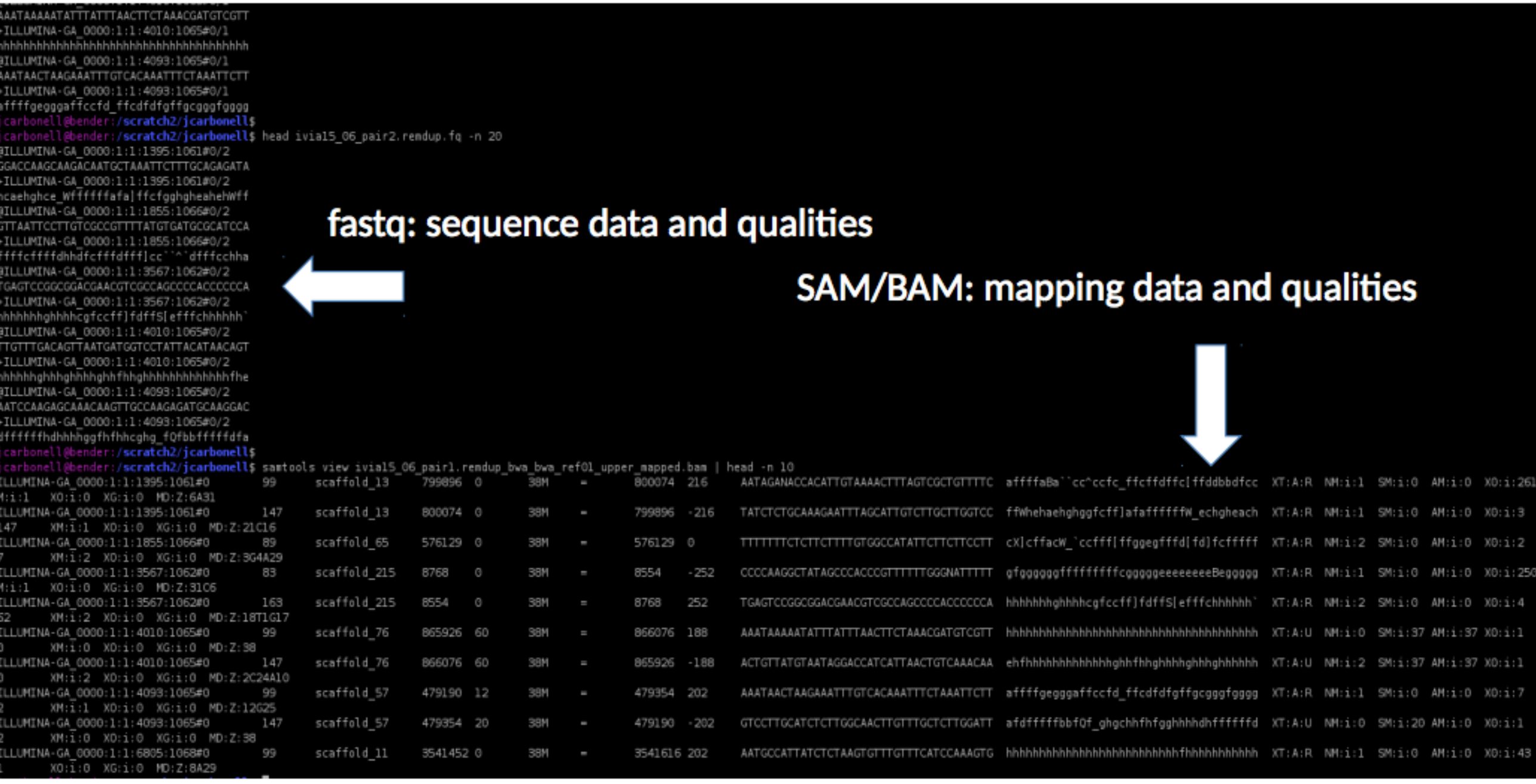
Comparison

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

Basic Steps NGS Technology



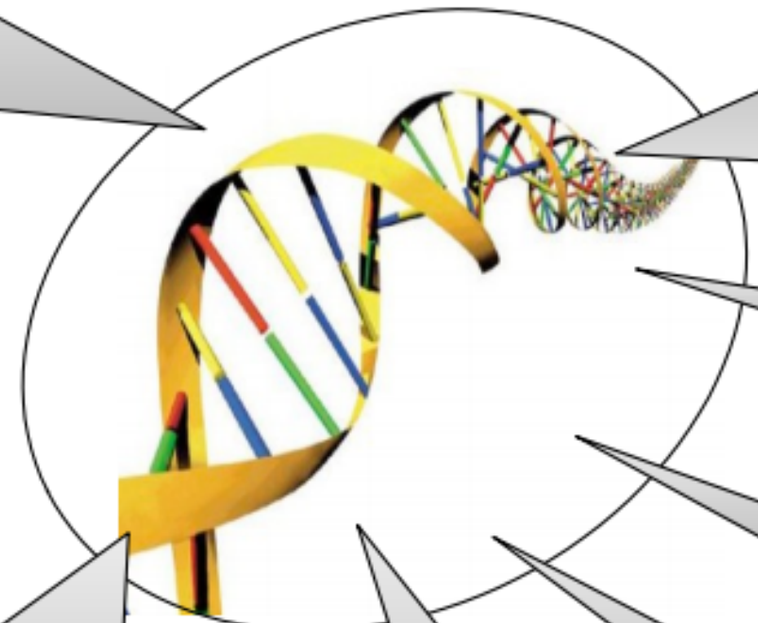
File Formats



GENOMIKA

Most Applications of NGS

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

Metagenomics
Metatranscriptomics

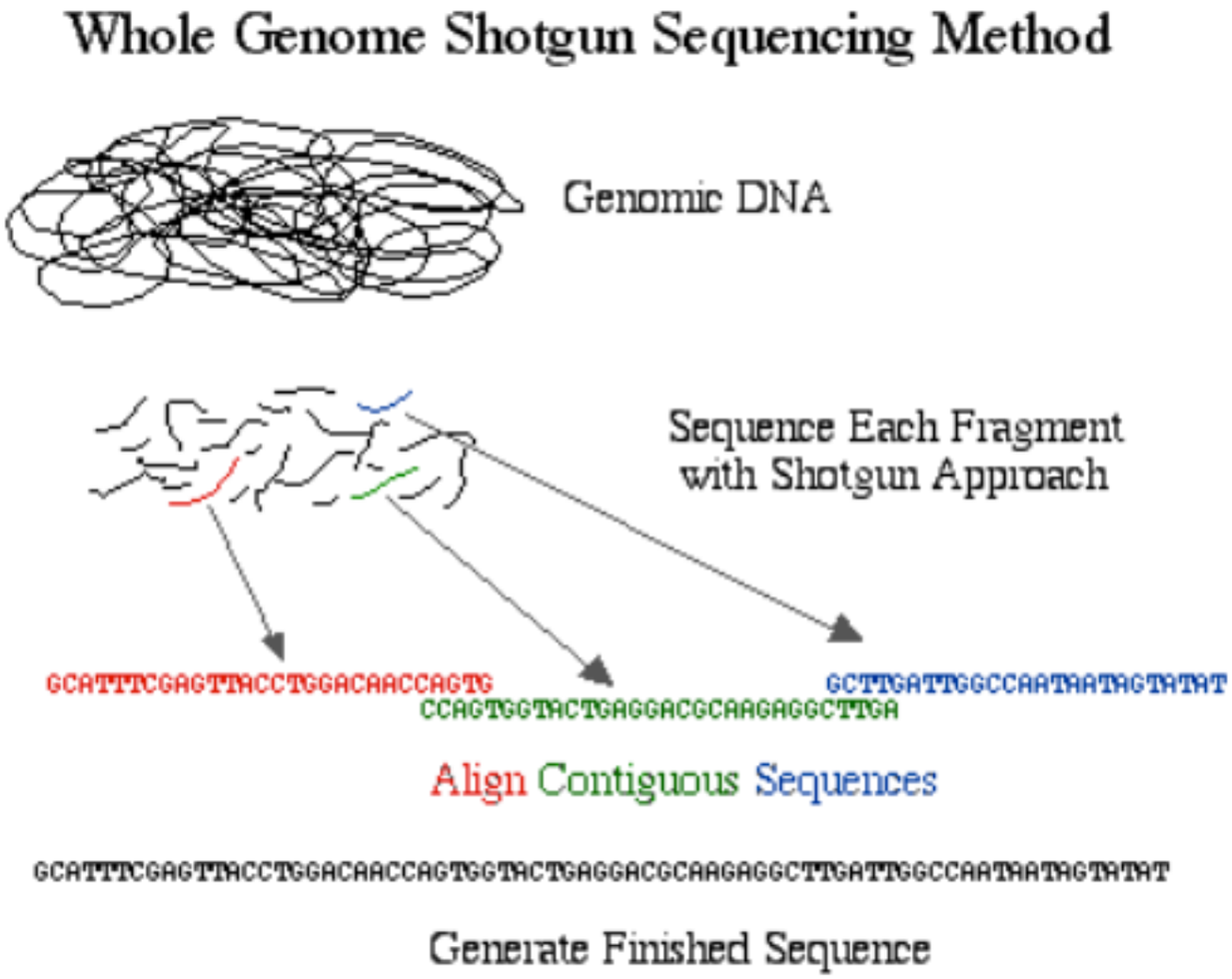
ChIP-seq /Epigenomics
- Protein-DNA interactions
- Active transcription factor binding sites
- Histone methilation

GENOMIKA

Most Applications of NGS

Whole GENOME Resequencing

- Need reference genome
- Variation discovery

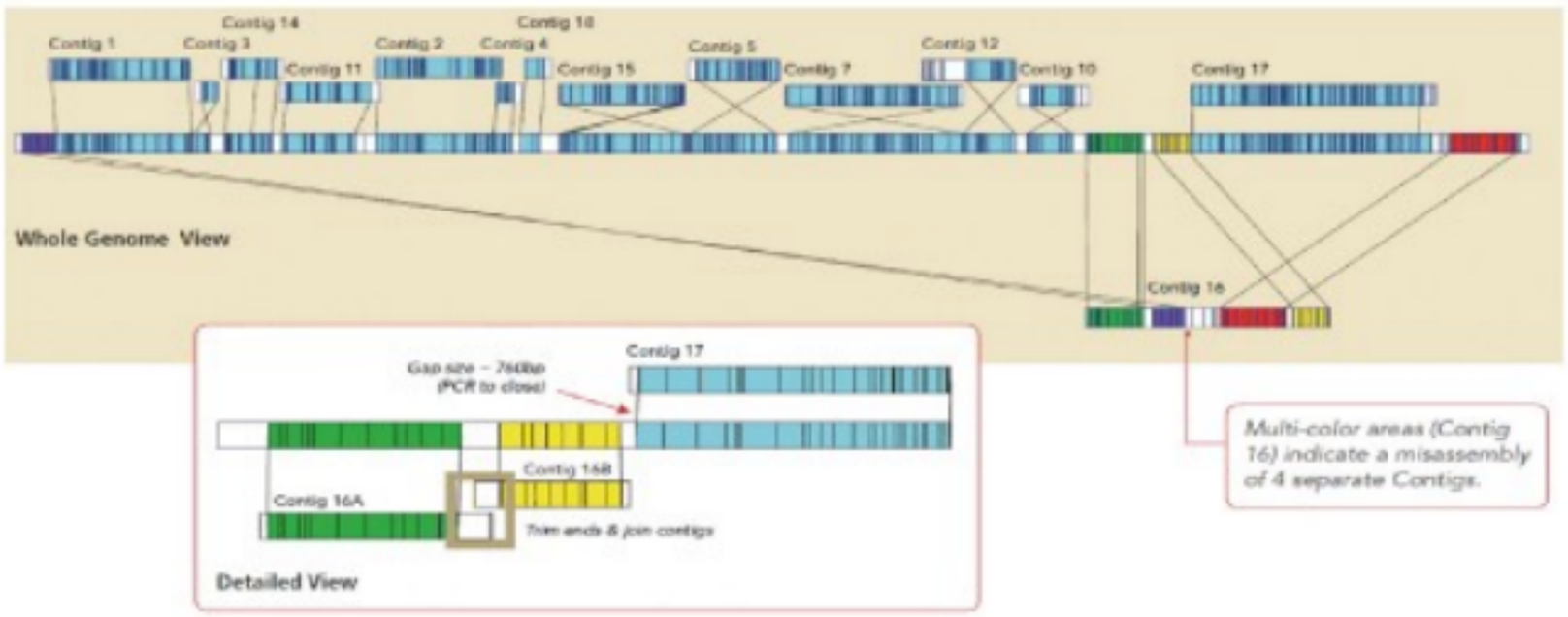


GENOMIKA

Most Applications of NGS

- **Whole GENOME "de novo" sequencing**
  - Uncharacterized genomes with no reference genome available
  - known genomes where significant structural variation is expected.
  - Long reads or mate-pair libraries. Sequencing mostly done by Roche 454 and also Illumina.
  - Assembly of reads is needed: Computational intensive

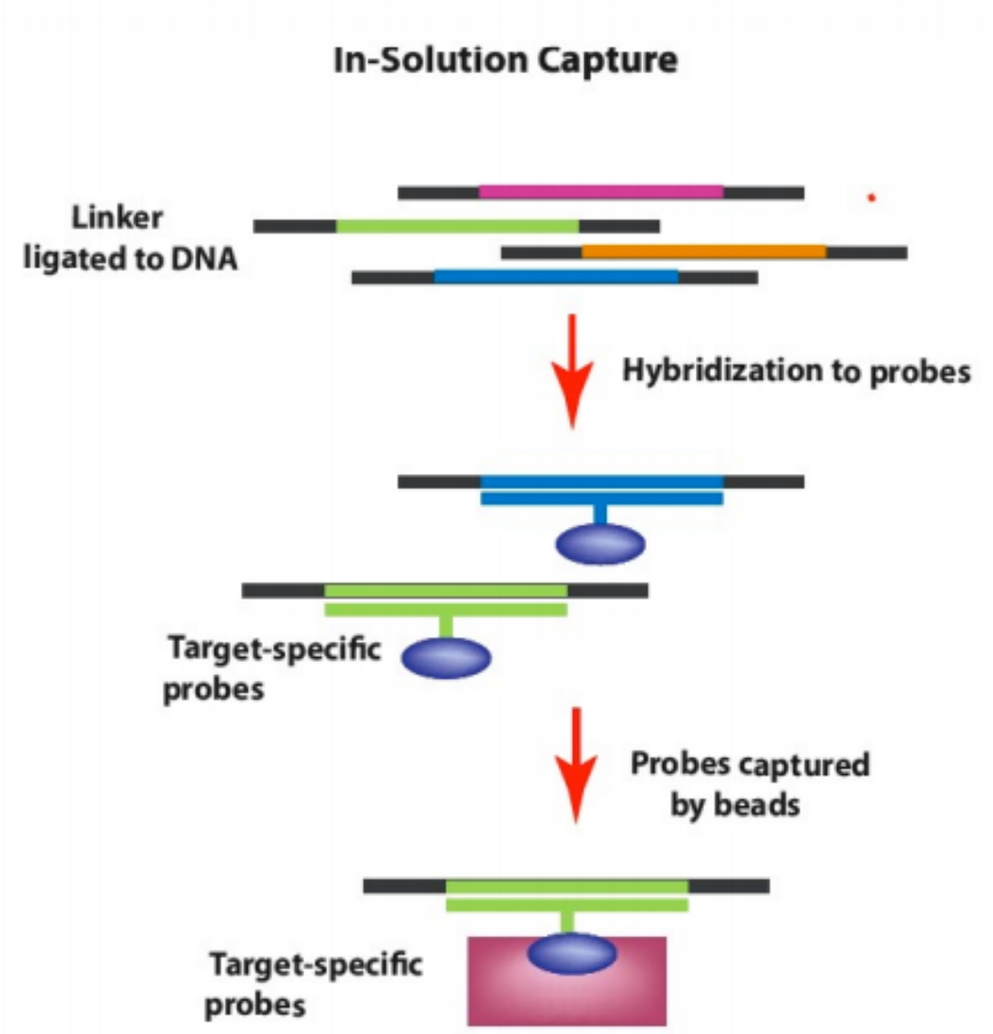
- E.g. Genome bacteria sequencing



GENOMIKA

Most Applications of NGS

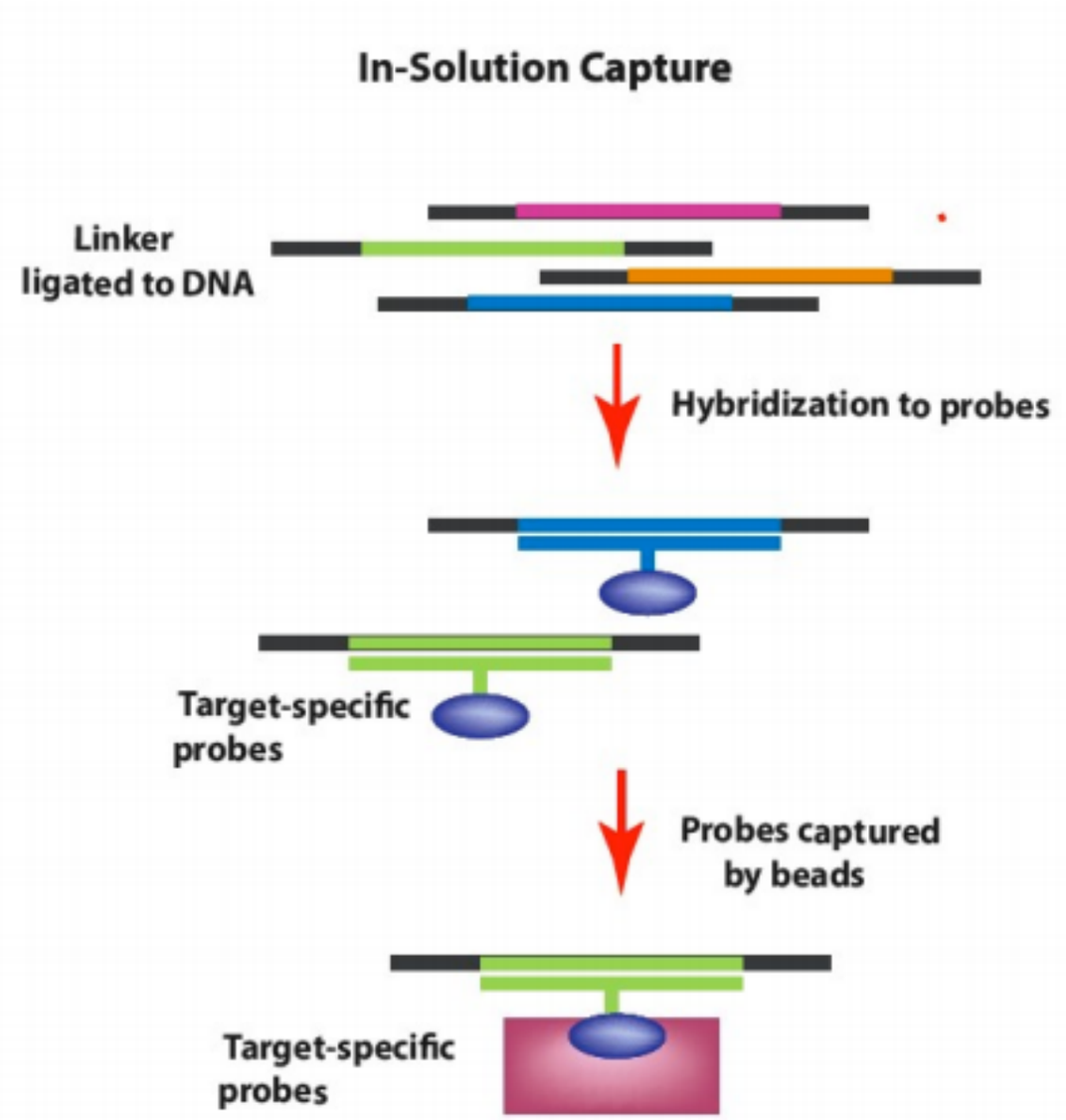
- **Whole EXOME Resequencing**
  - Need reference genome
    - Available for Human and Mouse
  - Variation discovery on ORFs
    - 2% of human genome (lower cost)
    - 85% disesease mutation are in the exome
  - Need probes complementary to exons
    - Nimblegen
    - Agilent



- E.g. Human exome

Most Applications of NGS

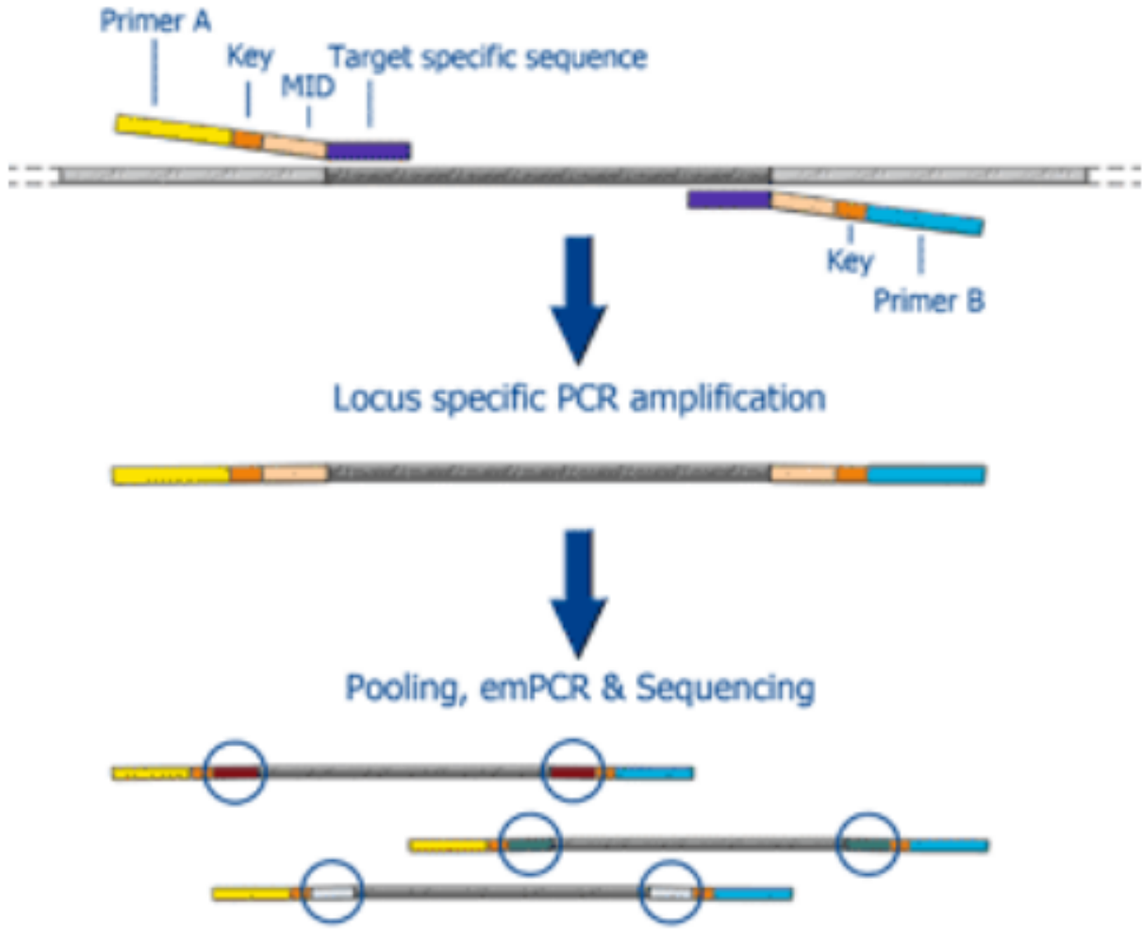
- **Targeted Resequencing**
  - Capture of specific regions in the genome
- **Custom genes panel sequencing**
  - Allows to cover high number of genes related to a disease
  - *E.g. Disease gene panel*
- Low cost and quicker than capillary sequencing
- Multiplexing is possible
- Need custom probes complementary to the genomic regions
  - Nimblegen
  - Agilent



GENOMIKA

Most Applications of NGS

- **Amplicon sequencing**
  - Sequencing of regions amplified by PCR.
  - Shorter regions to cover than targeted capture
  - No need of custom probes
  - Primer design is needed
  - High fidelity polymerase
  - Multiplexing is needed



- *E.g. P53 exon amplicon sequencing*

Most Applications of NGS

- **RNA-Seq**
  - Sequencing of mRNA
  - rRNA depleted samples
  - Very high dynamic range
  - No prior knwoledge of expressed genes
  - Gives information about (richer than microarrays)

- Differential expression of **known or unknown** transcripts during a treatment or condition
- **Isoforms** and
- New **alternative splicing** events
- **Non-coding** RNAs
- Post-transcriptional mutations or **editing**,
- **Gene fusions**.

Applications of RNAseq

Qualitative:
* Alternative splicing
* Antisense expression
* Extragenic expression
* Alternative 5' and 3' usage
* Detection of fusion transcripts

....

Tophat/Cufflinks
Scripture
Alexa

Quantitative:
* Differential expression
* Dynamic range of gene expression

....

edgeR
DESeq
baySeq
NOISeq

Histórias de uso do NGS

GENOMIKA

NATURE | ARTICLE OPEN

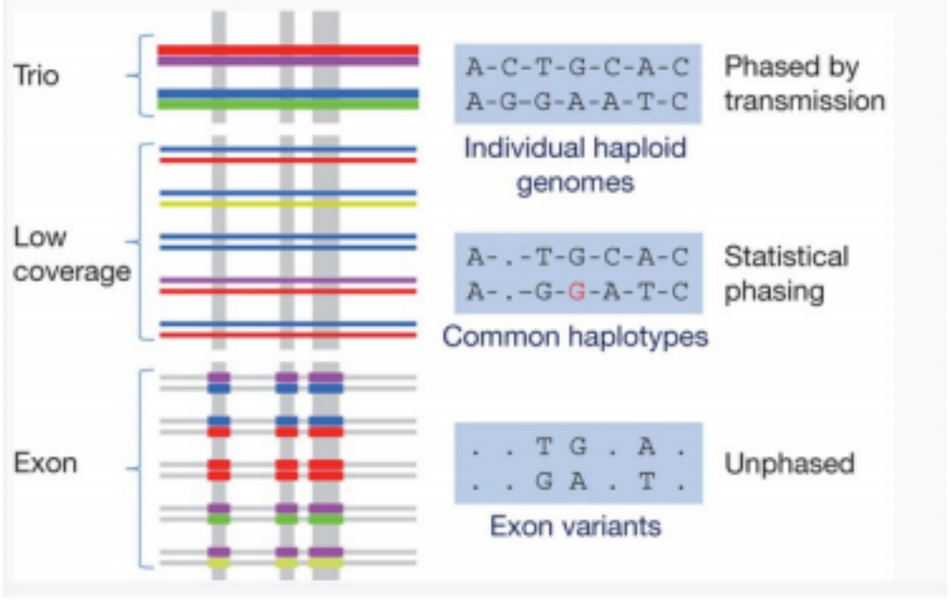
previous article next article

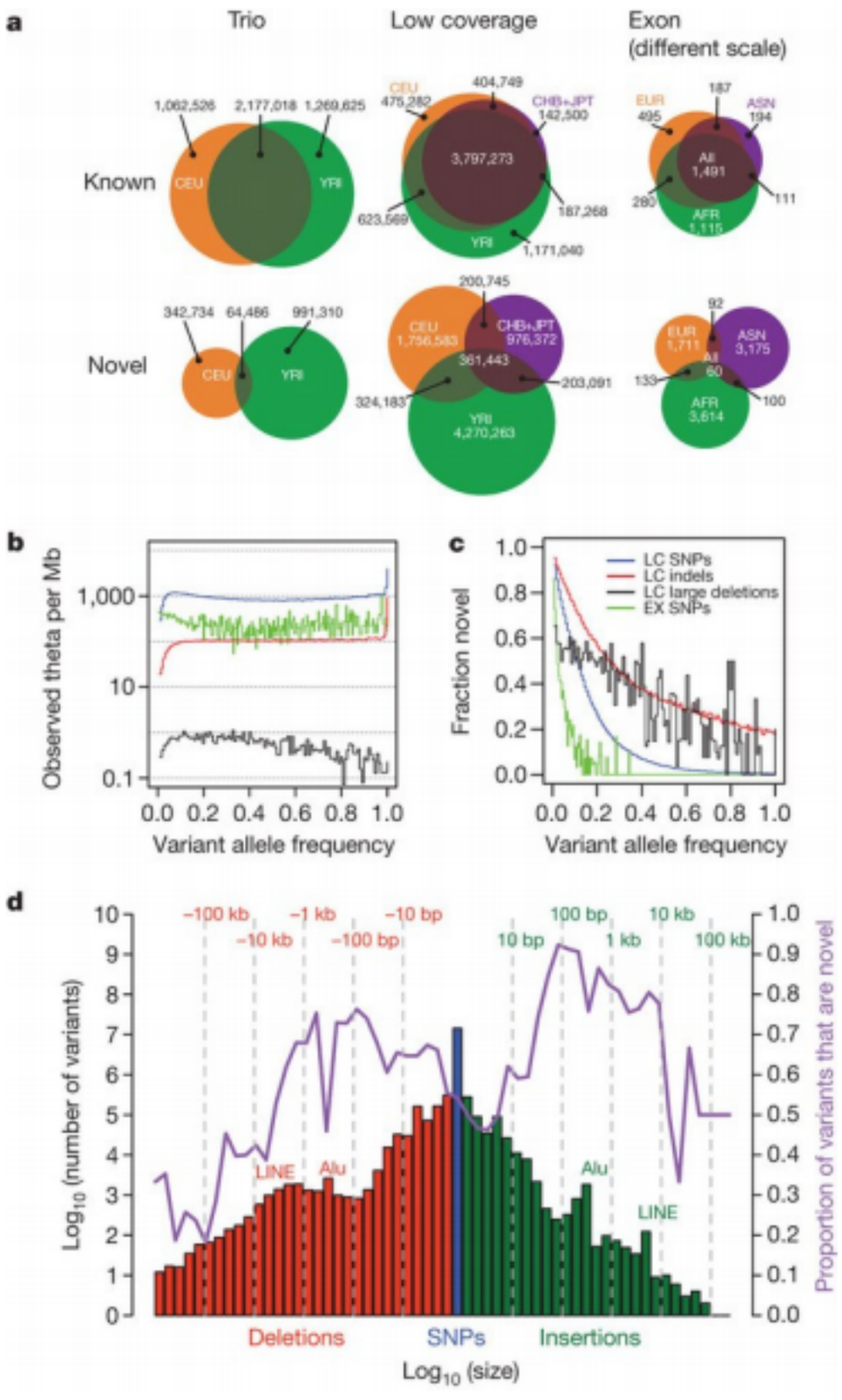
A map of human genome variation from population-scale sequencing

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

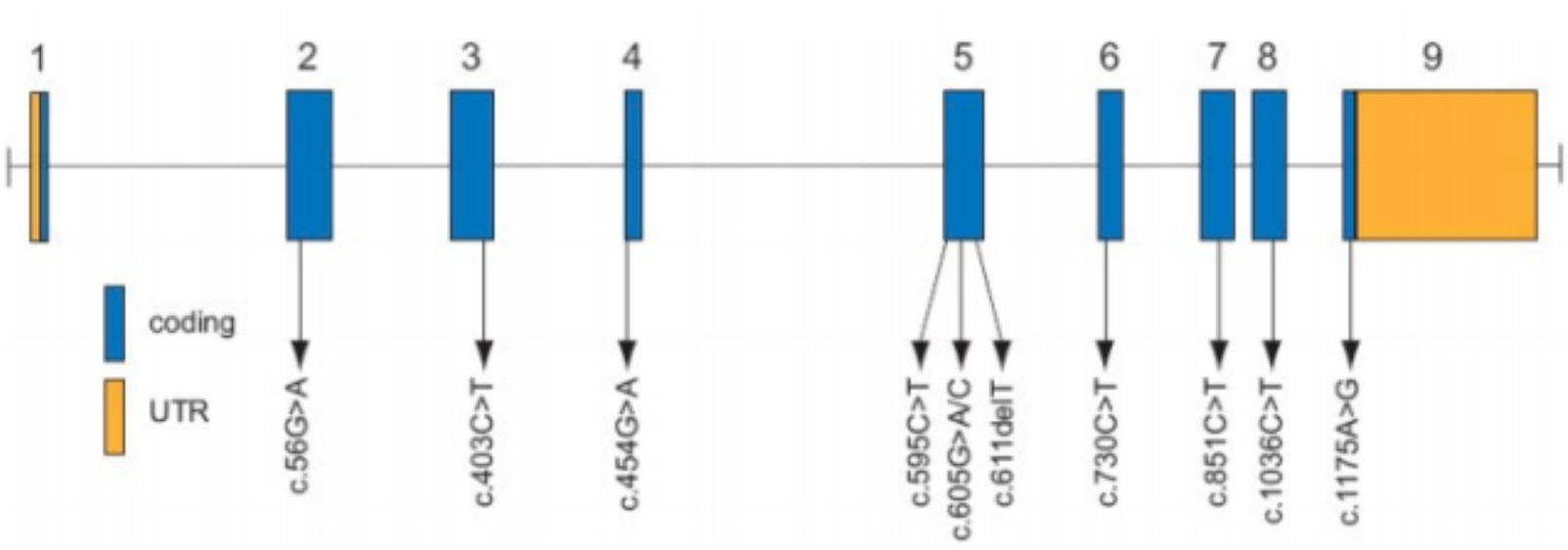


Figure 2. Genomic structure of the exons encoding the open reading frame of *DHODH*

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

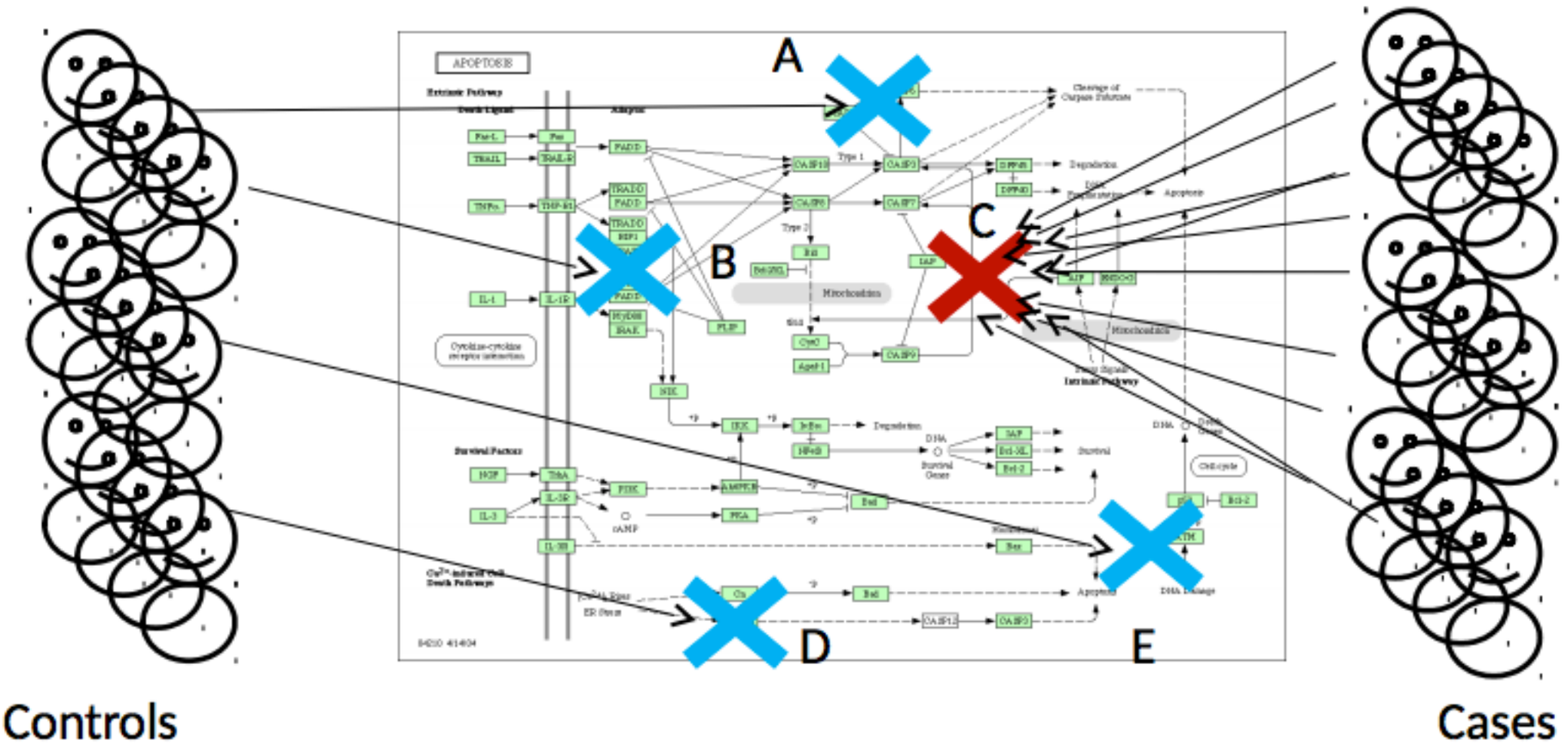
GENOMIKA

Não é tão fácil,
Desafios grandes a encontrar…

GENOMIKA

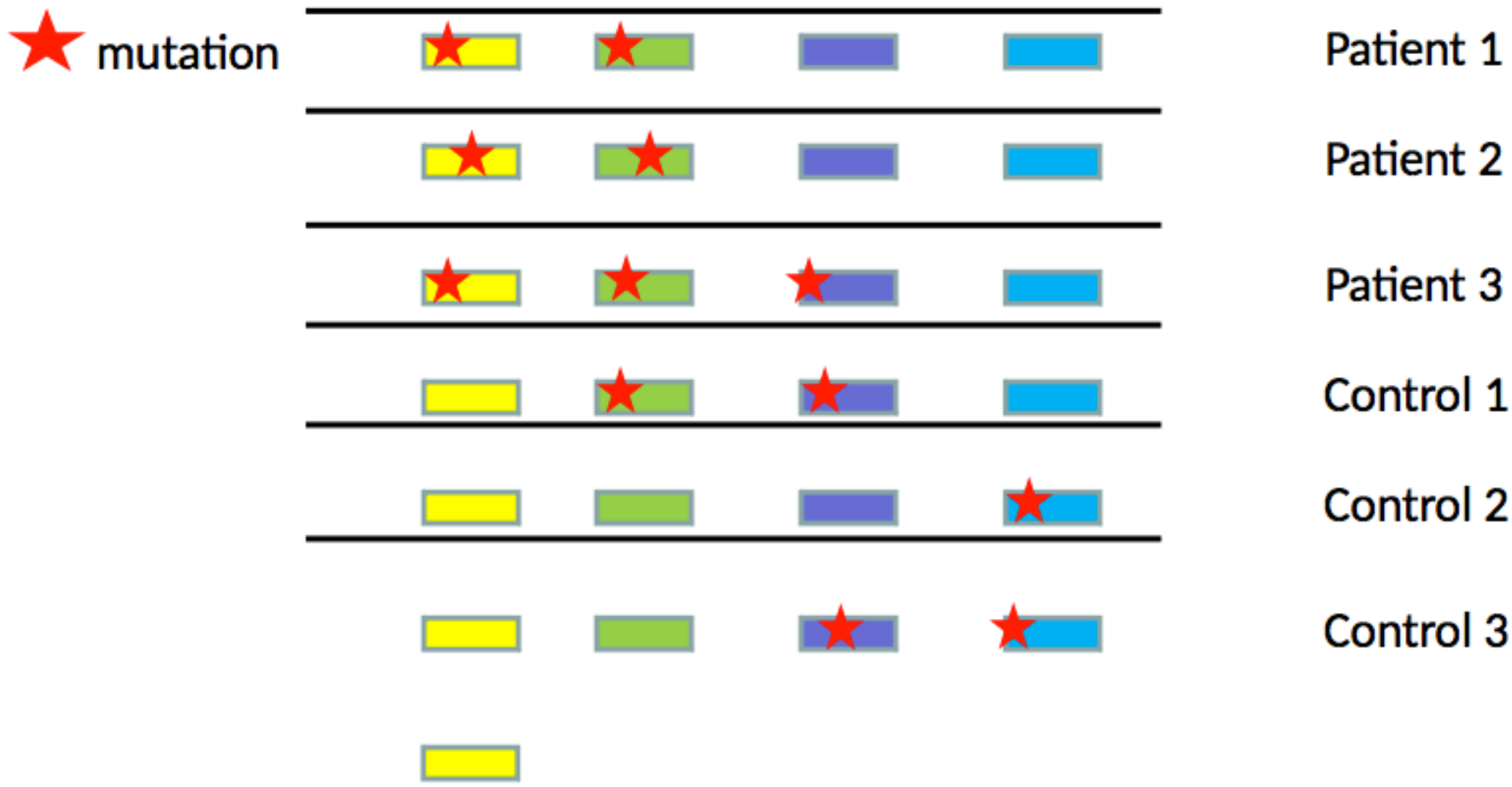
Análise Secundária
Detectar as mutações associadas à doença

- Caso mais simples: Doenças monogênicas associadas a um único gene.



GENOMIKA

Análise Secundária
Detectar as mutações associadas à doença

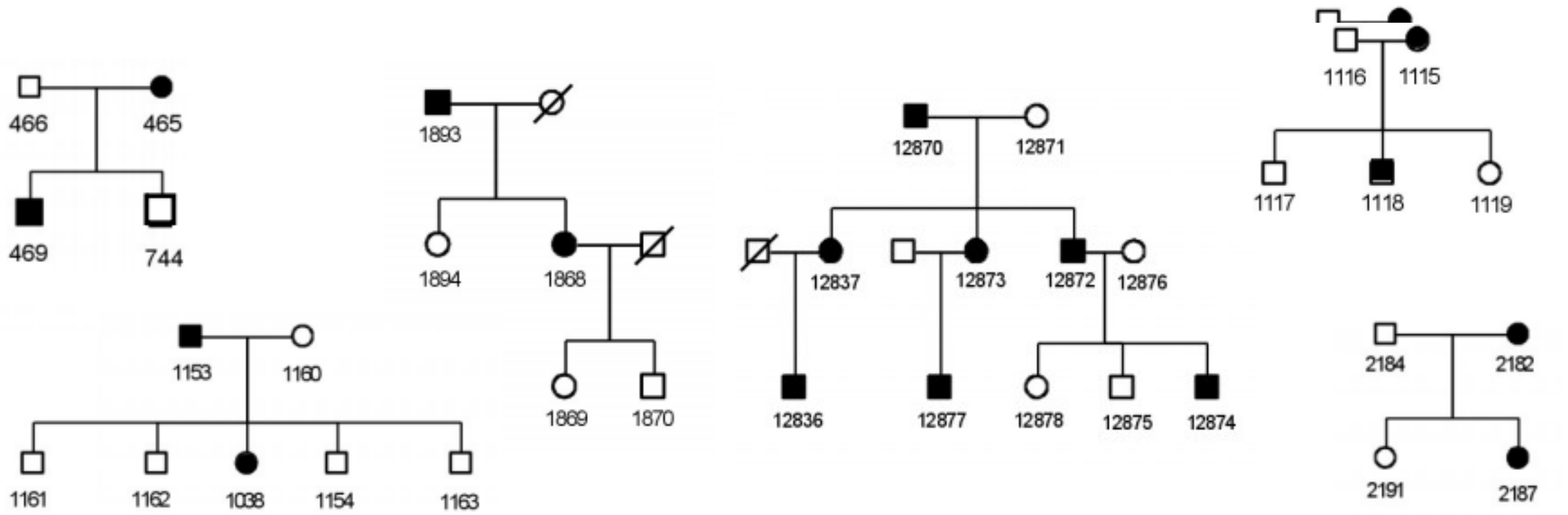


candidate gene (shares mutation for all patients but no controls)

Ainda assim, localizar variantes raras não é uma tarefa fácil…

- Interrogar 50Mb produz muitas variantes…
- Em muitos casos não estamos procurando por novas variantes e sim por já conhecidas
- Muitos fenótipos podem estar associadas a mutações diferentes e genes diferentes

Usando histórico familiar pode ajudar

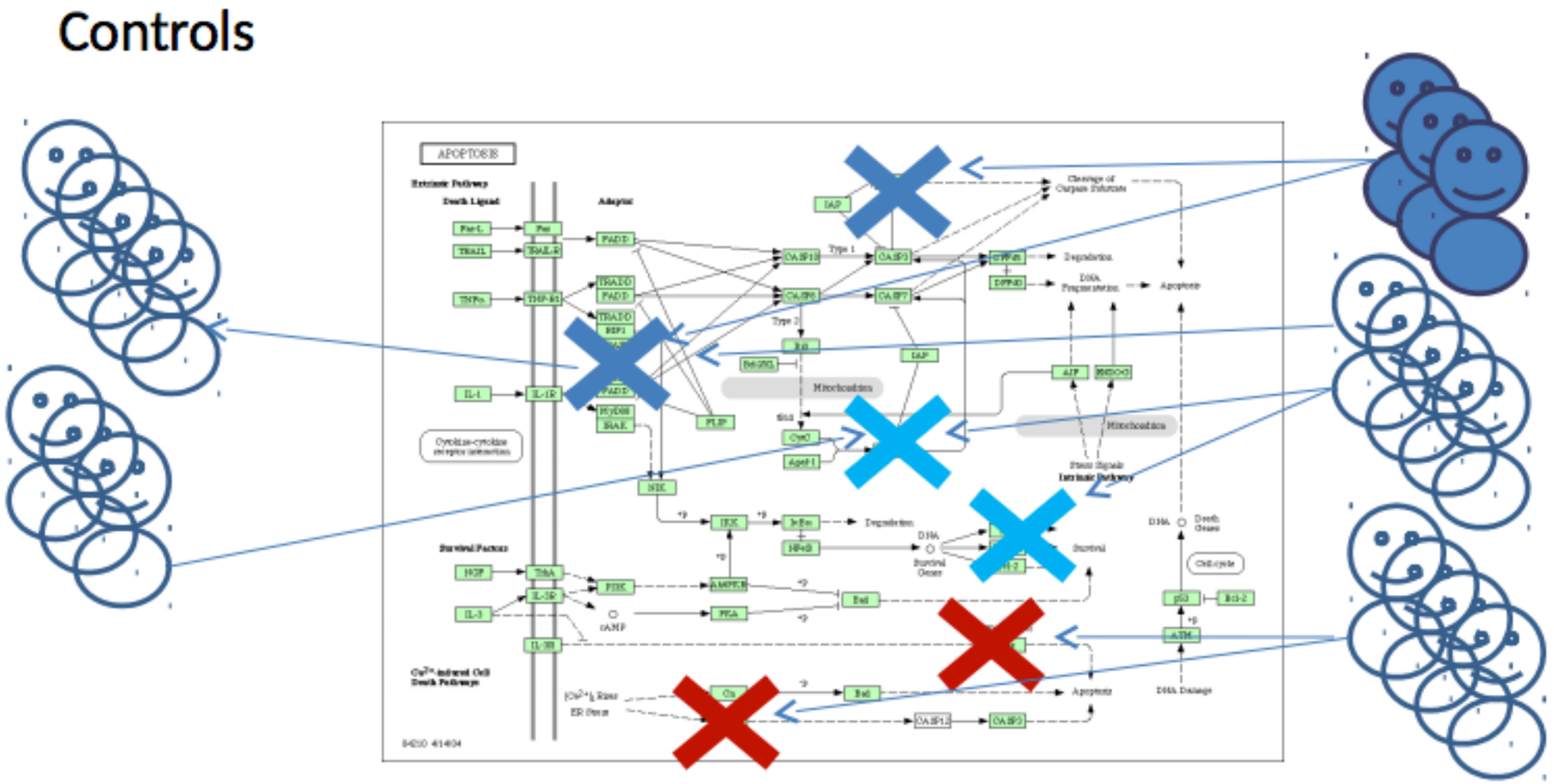


	Families					
	1	2	3	4	5	6
Variants	3403	82	4	0	0	0
Genes	2560	331	35	8	1	0

Problem: how to prioritize putative candidate genes

GENOMIKA

Localizar associações de genes para algumas doenças é uma tarefa difícil…



They can have different mutations (or combinations).

Many cases have to be used to obtain significant associations to many markers.

The only common element is the pathway (yet unknow) affected.

GENOMIKA

Fica claro que...

NGS está revolucionando a maneira
como estamos fazendo pesquisa com genoma

Mas fica o próximo passo:

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

GENOMIKA

GENOMIKA

Introdução às Tecnologias NGS

Marcel Caraciolo, CTO
marcel@genomika.com.br