Discovery and annotation of variants by exome analysis using NGS

Granada, June 2011

Javier Santoyo
jsantoyo@cipf.es
http://bioinfo.cipf.es
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

INB

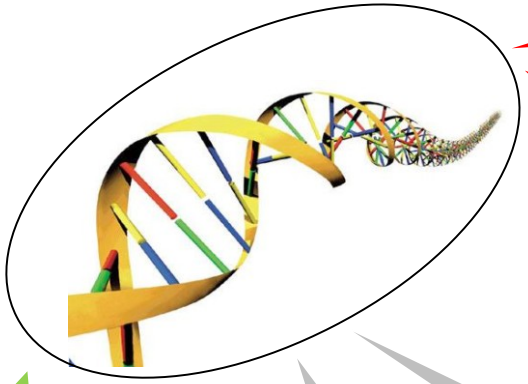
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

ciberer

Some of the most common applications of NGS.

Many array-based technologies become obsolete !!!

RNA-seq
Transcriptomics:
Quantitative
Descriptive
(alternative
splicing)
miRNA



Resequencing:
SNV and indel

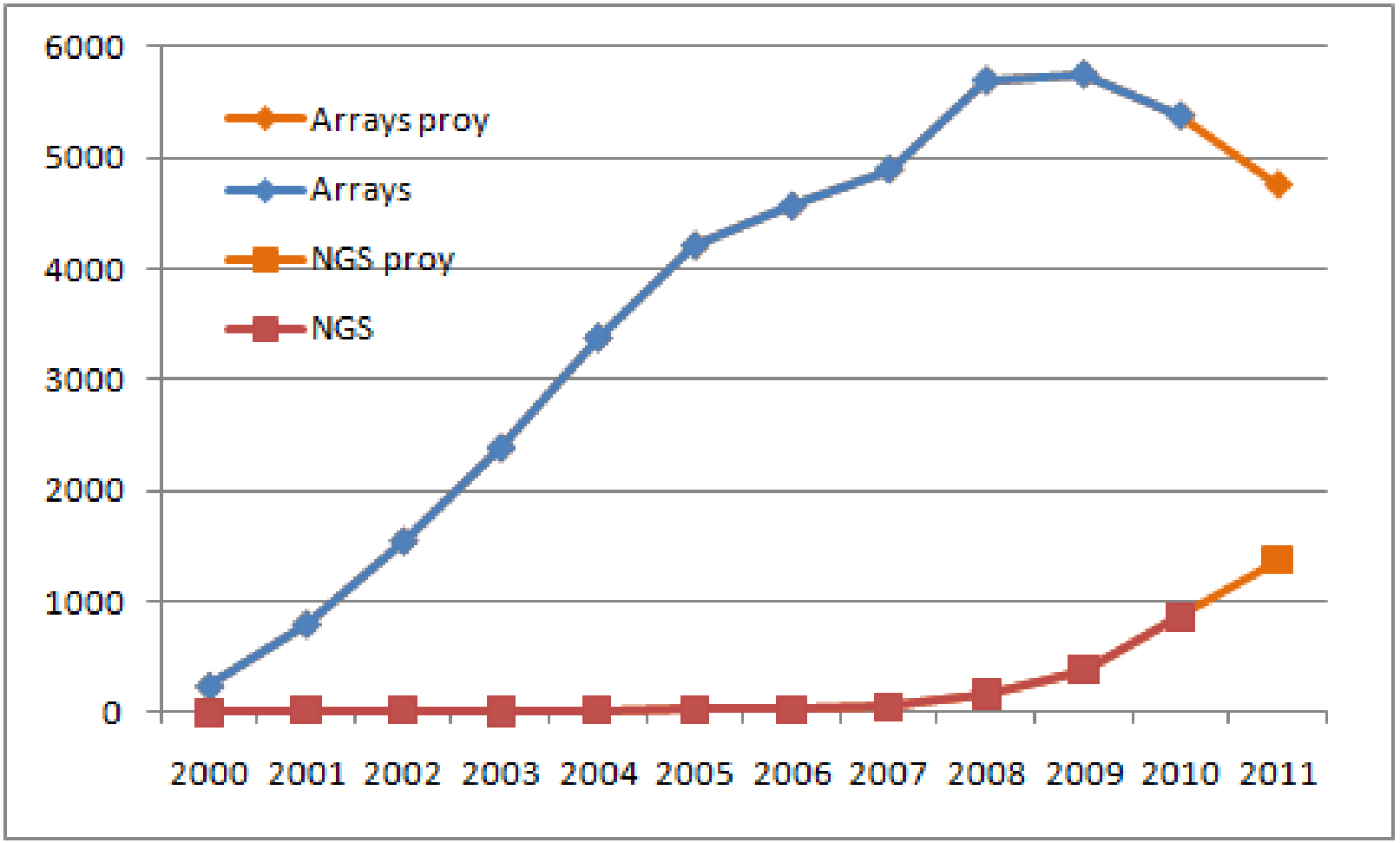
Structural
variation (CNV,
translocations,
inversions, etc.)

Chip-seq
Protein-DNA interactions
Active transcription factor
binding sites, etc.

De novo
sequencing

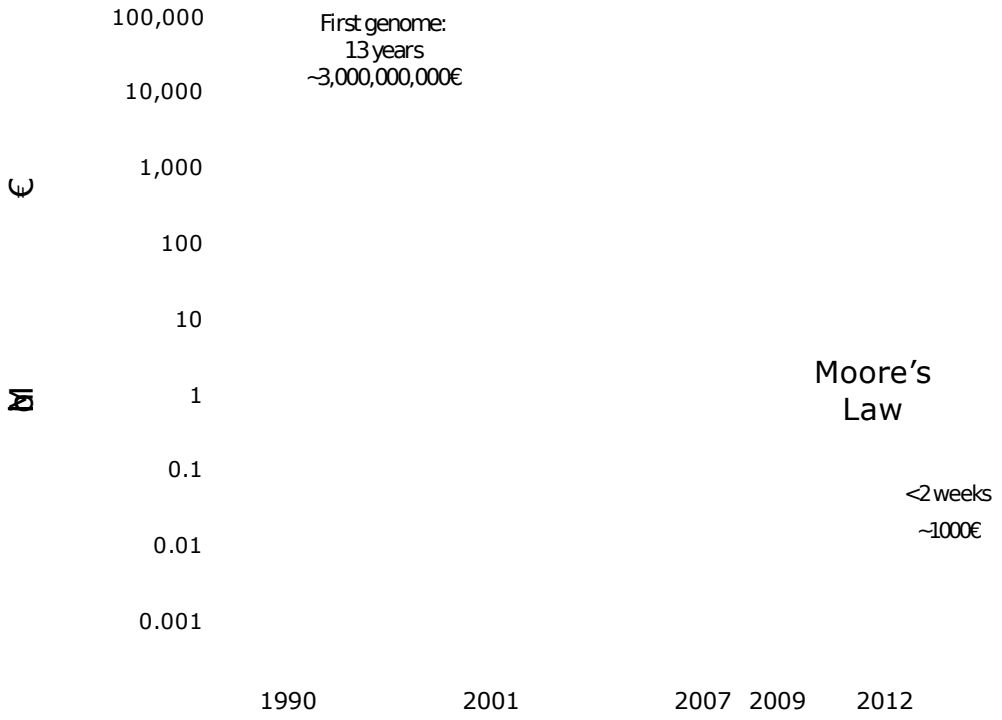
Metagenomics
Metatranscriptomics

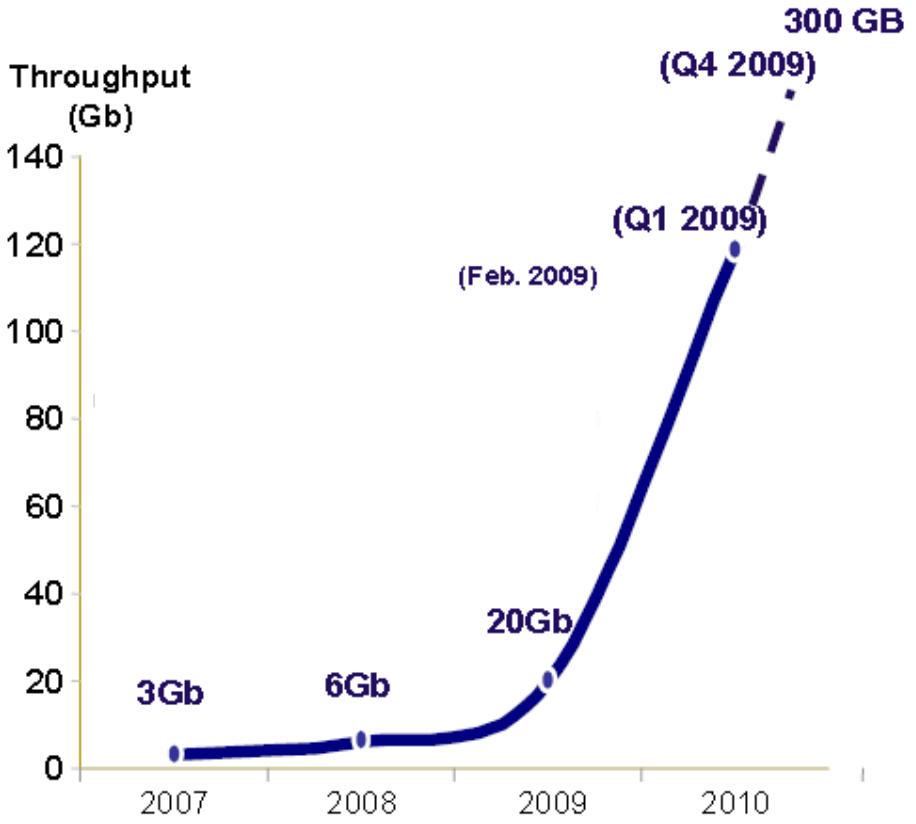
Evolution of the papers published in microarray and next gen technologies



Source Pubmed. Query: "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]
Projections 2011 based on January and February

Next generation sequencing technologies are here





While the cost goes down, the amount of data to manage and its complexity raise exponentially.

Sequence to Variation Workflow

Raw Data

FastQ

FastX Tookit

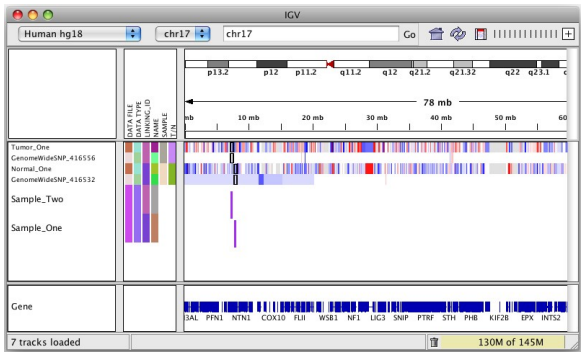
FastQ

BWA/ BWASW

BWA/ BWASW

SAM

IGV



SAM Tools

BAM

Samtools mPileup

Pileup

BCF Tools

Raw VCF

BCF Tools

Filter VCF

GFF

Raw Sequence Data Format

- Fasta, csfasta
- Fastq, csfastq
- SFF
- SRF
- The eXtensible SeQuence (XSQ)

- Others:
  - http://en.wikipedia.org/wiki/List_of_file_formats#Biology

Fasta & Fastq formats

- FastA format
  - Header line starts with “>” followed by a sequence ID
  - Sequence (string of nt).
- FastQ format
  - First is the sequence (like Fasta but starting with “@”)
  - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
- Nearly all aligners take FastA or FastQ as input sequence
- Files are flat files and are big

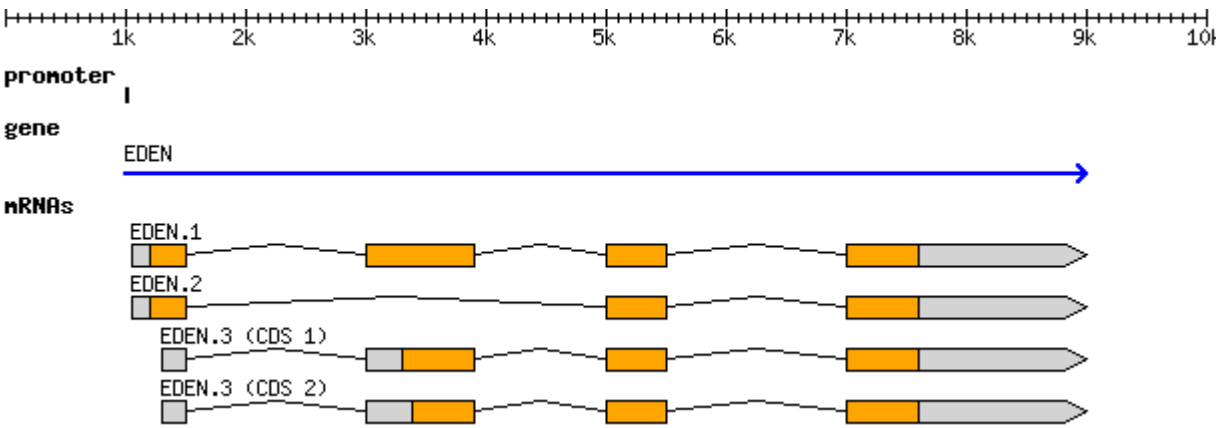
Processed Data Formats

- Column separated file format contains features and their chromosomal location. There are flat files (no compact)
  - GFF and GTF
  - BED
  - WIG
- Similar but compact formats and they can handle larger files
  - BigBED
  - BigWIG

Processed Data Formats GFF

- Column separated file format contains features located at chromosomal locations
- Not a compact format
- Several versions
  - GFF 3 most currently used
  - GFF 2.5 is also called GTF (used at Ensembl for describing gene features)

GFF structure



GFF3 can describes the representation of a protein-coding gene

GFF3 file example

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001
ctg123 . exon            1300  1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  2  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  2  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  2  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  2  ID=cds00004;Parent=mRNA00003
```

```
Column 1: "seqid"
Column 2: "source"
Column 3: "type"
Column 4: "start"
Column 5: "end"
Column 6: "score"
Column 7: "strand"
Column 8: "phase"
Column 9: "attributes"
```

BED

- Created by USCS genome team
- Contains similar information to the GFF, but optimized for viewing in the UCSC genome browser
- BIG BED, optimized for next gen data – essentially a binary version
  - It can be displayed at USCS Web browser (even several Gbs !!)

WIG

- Also created by USCS team
- Optimized for storing “levels”
- Useful for displaying “peaks” (transcriptome, ChIP-seq)
- BIG WIG is a binary WIG format
- It can also be uploaded onto USCS web browser

Short Read Aligners, just a few?

AGiLE
BFAST
BLASTN
BLAT
Bowtie
BWA
CASHX
CUDA-EC
ELAND
GNUMAP
GMAP and GSNAP
Geneious Assembler
LAST
MAQ
MOM
MOSAIK
Novoalign
PALMapper
PerM
QPalma
RazerS
RMAP
SeqMap
Shrec
SHRiMP
SLIDER
SLIM Search
SOAP and SOAP2
SOCS
SSAHA and SSAHA2
Stampy
Taipan
UGENE
XpressAlign
ZOOM
...

Adapted from http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Mapped Data: SAM specification

This specification aims to define a generic sequence alignment format, SAM, that describes the alignment of query sequencing reads to a reference sequence or assembly, and:

• Is flexible enough to store all the alignment information generated by various alignment programs;

• Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;

- SAM specification was developed for the 1000 Genome Project.
  - Contains information about the alignment of a read to a genome and keeps track of chromosomal position, quality alignment, and features of the alignment (extended cigar).
- Includes mate pair / paired end information joining distinct reads
- Quality of alignment denoted by mapping/pairing QV

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

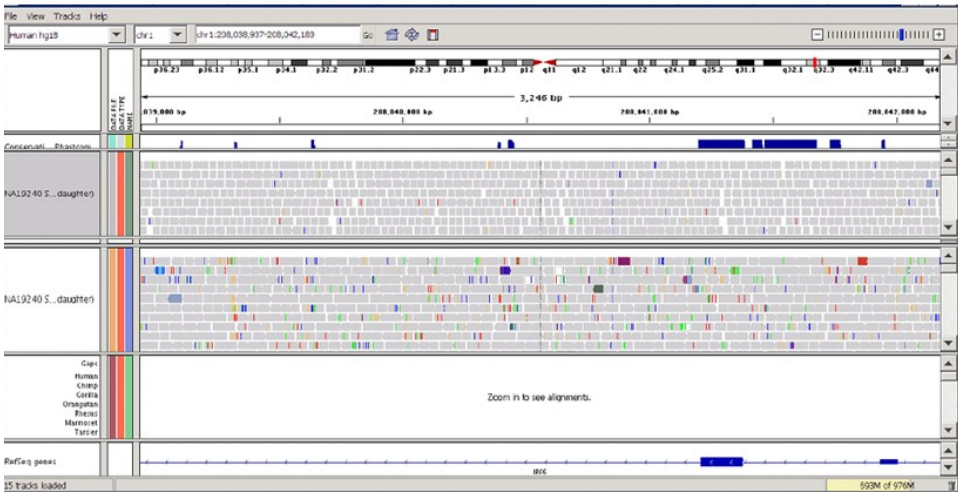
Mapped Data: SAM/BAM format

- SAM (Sequence Alignment/Map) developed to keep track of chromosomal position, quality alignments and features of sequence reads alignment.
- BAM is a binary version of SAM - This format is more compact
- Most of downstream analysis programs takes this binary format
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

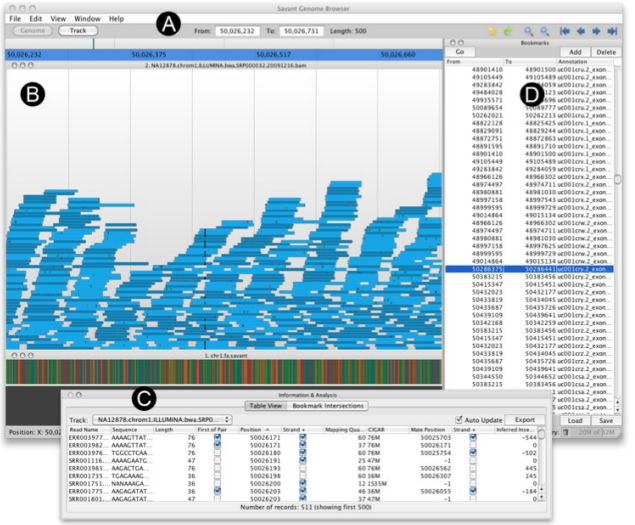
Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

BAM format

- Many tertiary analysis tools use BAM
  - BAM makes machine specific issues “transparent” e.g. colour space
  - A common format makes downstream analysis independent from the mapping program



IGV http://www.broadinstitute.org/igv



SAVANT Fiume et al

Adapted from http://www.bioinformatics.ca/files/CBW%20-%20presentations/HTSeq_2010_Module%202/HTSeq_2010_Module%202.pdf

SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

SAM format

Aligners natively generating SAM

- BFAST, `Blat-like Fast Accurate Search Tool' for Illumina and SOLiD reads.
- Bowtie. Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in samtools-C.
- BWA, Burrows-Wheeler Aligner for short and long reads.
- GEM library. Short read aligner. Convertor provided by the developers.
- Karma, the K-tuple Alignment with Rapid Matching Algorithm.
- Mosaik. The latest version support SAM output.
- Novoalign. An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.
- SNP-o-matic, short read aligner and SNP caller.
- SOLiD BaseQV Tool. Developed by Applied Biosystems for converting SOLiD output files.
- SSAHA2 (since v2.4). Classical aligner for both short and long reads.
- Stampy, by Gerton Lunter. An accurate read aligner capable of gapped alignment for Illumina short reads. Used for indel discovery on the 1000 genomes data. Not released.
- TopHat for mapping short RNA-seq reads bridging exon junctions.

VCF format

- The Variant Call Format (VCF) is the emerging standard for storing variant data.

- Originally designed for SNPs and short INDELs, it also works for structural variations.

- VCF consists of a header section and a data section.

- The header must contain a line starting with one '#', showing the name of each field, and then the sample names starting at the 10th column.

- The data section is TAB delimited with each line consisting of at least 8 mandatory fields

The FORMAT field and sample information are allowed to be absent.

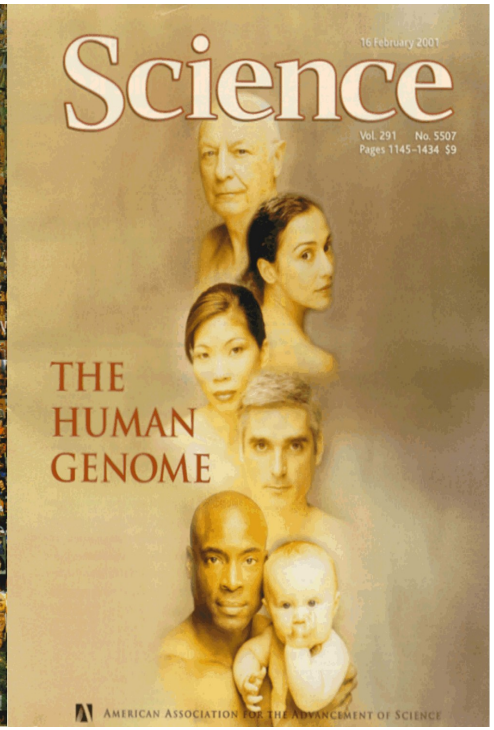
http://samtools.sourceforge.net/mpileup.shtml

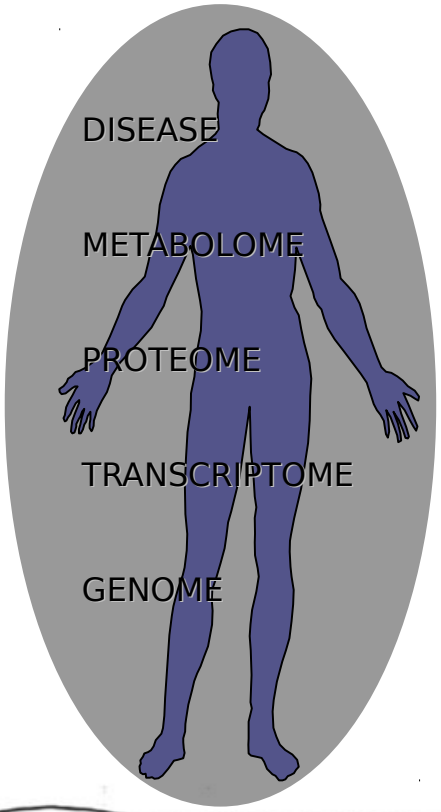
VCF data section fields

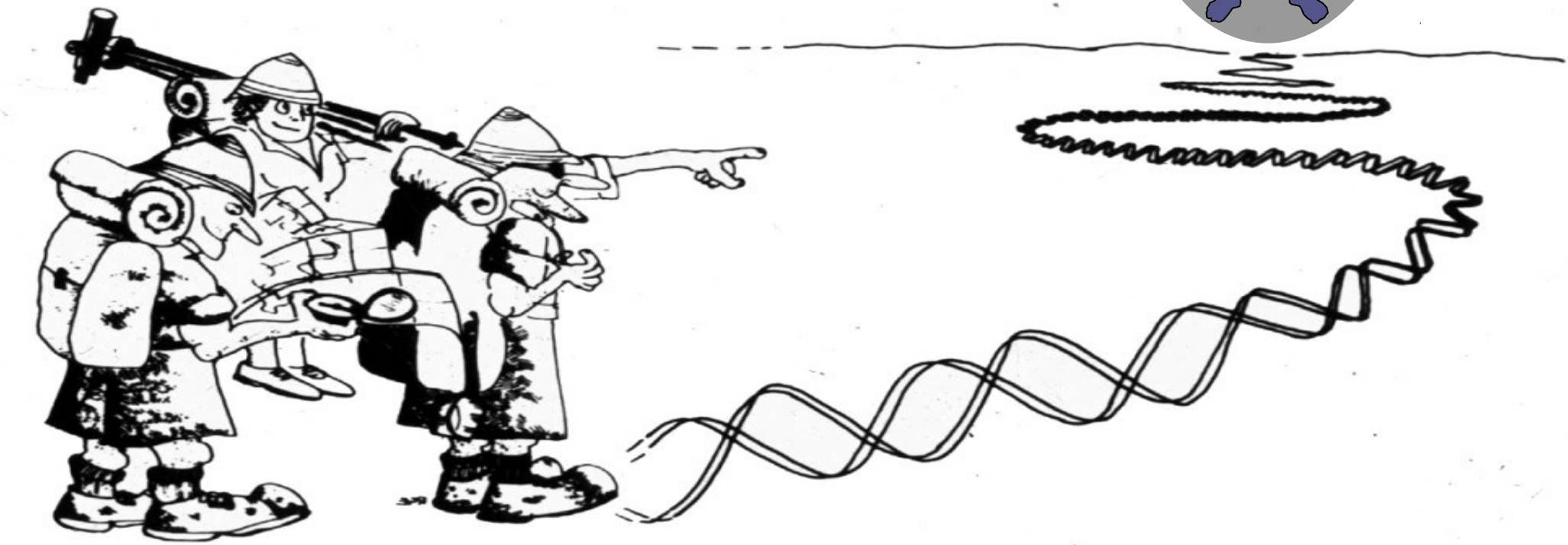
Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier. Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative seuqence(s).
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

http://samtools.sourceforge.net/mpileup.shtml

The Draft Human Genome Sequence millestone

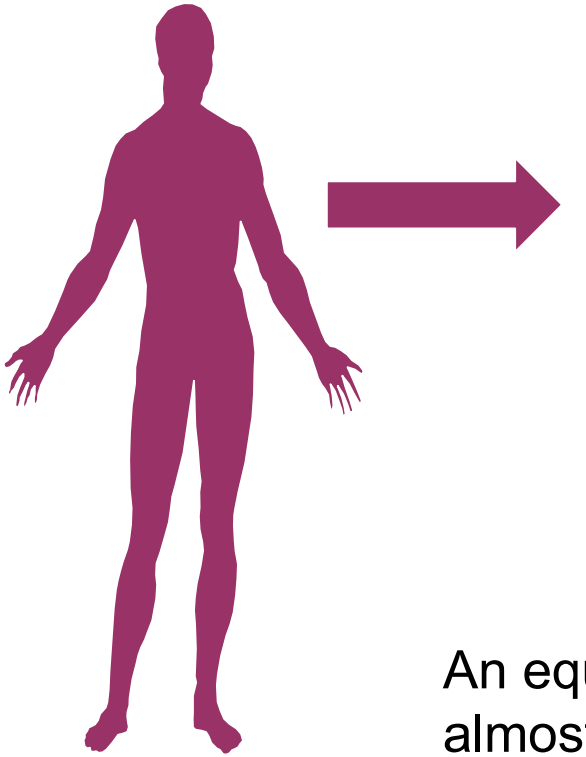






Some diseases are coded in the genome

How to find mutations associated to diseases?



An equivalent of the genome would amount almost 2000 books, containing 1.5 million letters each (average books with 200 pages).
This information is contained in any single cell of the body.

In monogenic diseases only one mutation causes the disease.

Example:
Book 1129, pag. 163, 3rd paragraph, 5th line, 27th letter should be a A instead a T

The challenge is to find this letter changed out of all the 3000 millions of letters in the 2000 books of the library

Solution:
Read it all

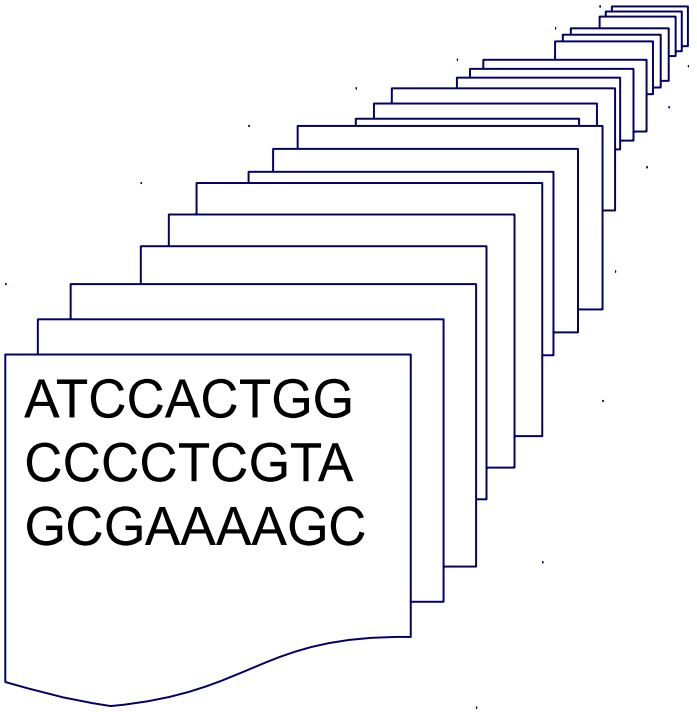
Problem:
Too much to read

The bioinfomatic challenge: finding the mutation that causes the disease.

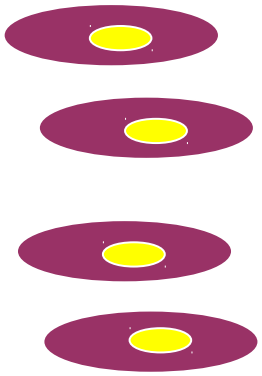
The problem is even worst: we cannot read the library complete library.

Sequencers can only read small portions. No more than 500 letters at a time.

So, the library must be inferred from fragments of pages of the books



Reading the text



En un lugar de la Mancha, de cuyo nombre no quiero acordarme…

En un lugar de la Mancha, de cuyo nombre no quiero acordarme…

En un lugar de la Mancha, de cuyo nombre ni quiero acordarme…

En un lugar de la Mancha, de cuyo nombre ni quiero acordarme…

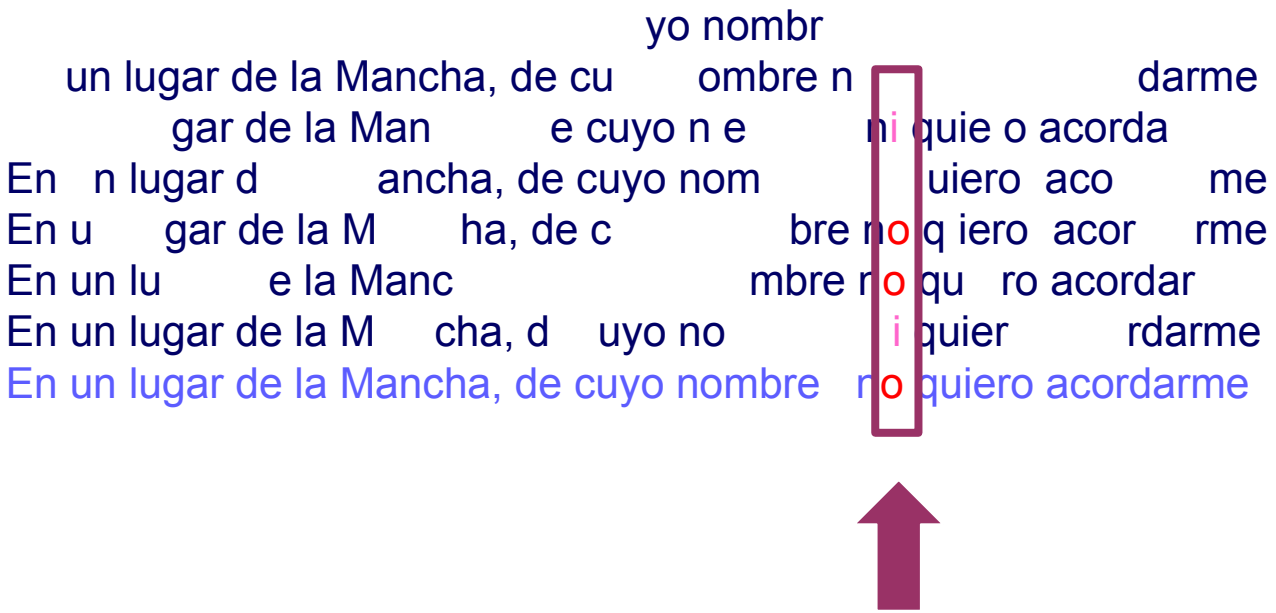
En u | n lugar d | e la Manc | ha, de c | uyo no | mbre no qu | iero acor | darme

En un lu | gar de la M | ancha, de c | uyo nom | bre no q | uiero aco | rdarme

En | un luga | r de la Ma | ncha, de cu | yo nombr | e ni quie | ro acordar | me

En un lu | gar de la Man | cha, d | e cuyo n | ombre n | i quier | o acorda | rme

Mapping fragments and detection of mutations



Space reduction: Look only for Mutations that can affect transcription and/or gene products

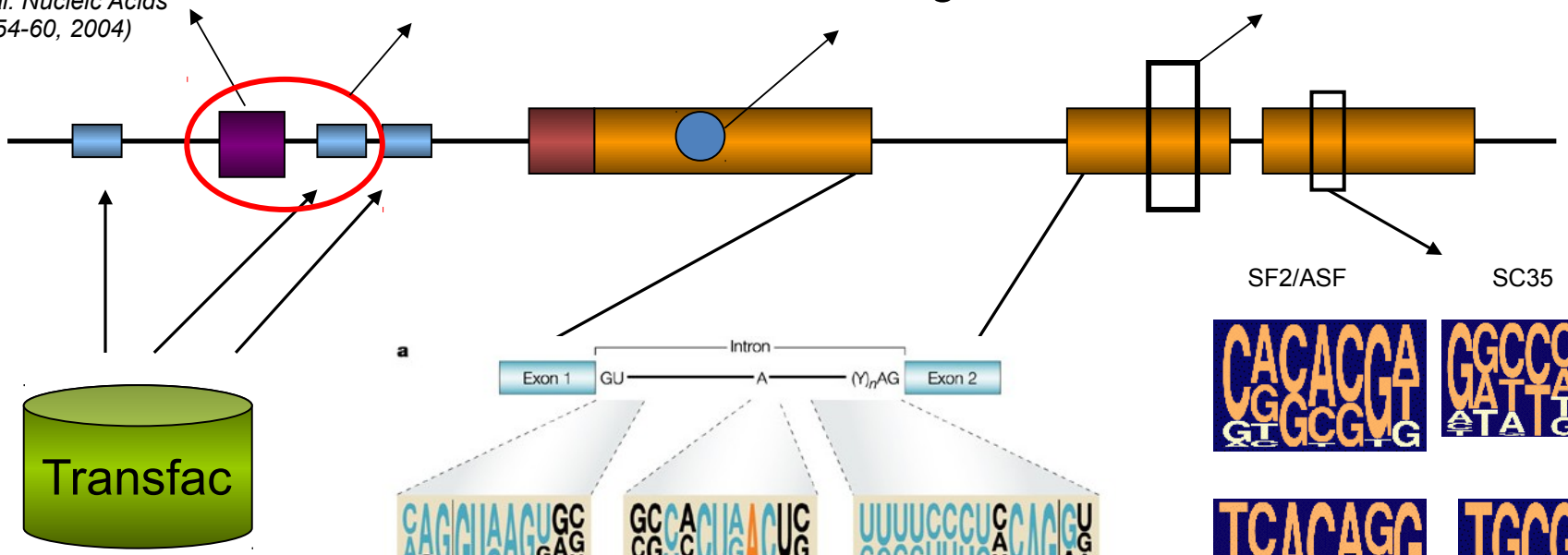
Triplex-forming sequences

(Goñi et al. Nucleic Acids Res.32:354-60, 2004)

Human-mouse conserved regions

Amino acid change

Splicing inhibitors



TFBSs

(Wingender et al., *Nucleic Acids Res*., 2000)

Intron/exon junctions

(Cartegni et al., *Nature Rev. Genet*., 2002)

SF2/ASF

SC35

SRp40

SRp55

ESE (exonic splicing enhancers) motifs recognized by SR proteins

(Cartegni et al., *Nucleic Acids Res*., 2003)

Why exome sequencing?

- Whole-genome sequencing of individual humans is increasingly practical . But cost remains a key consideration and added value of intergenic mutations is not cost-effective.

- Alternative approach: targeted resequencing of all protein-coding subsequences (exome sequencing, ~1% of human genome)

- Linkage analysis/positional cloning studies that focused on protein coding sequences were highly successful at identification of variants underlying monogenic diseases (when adequately powered)

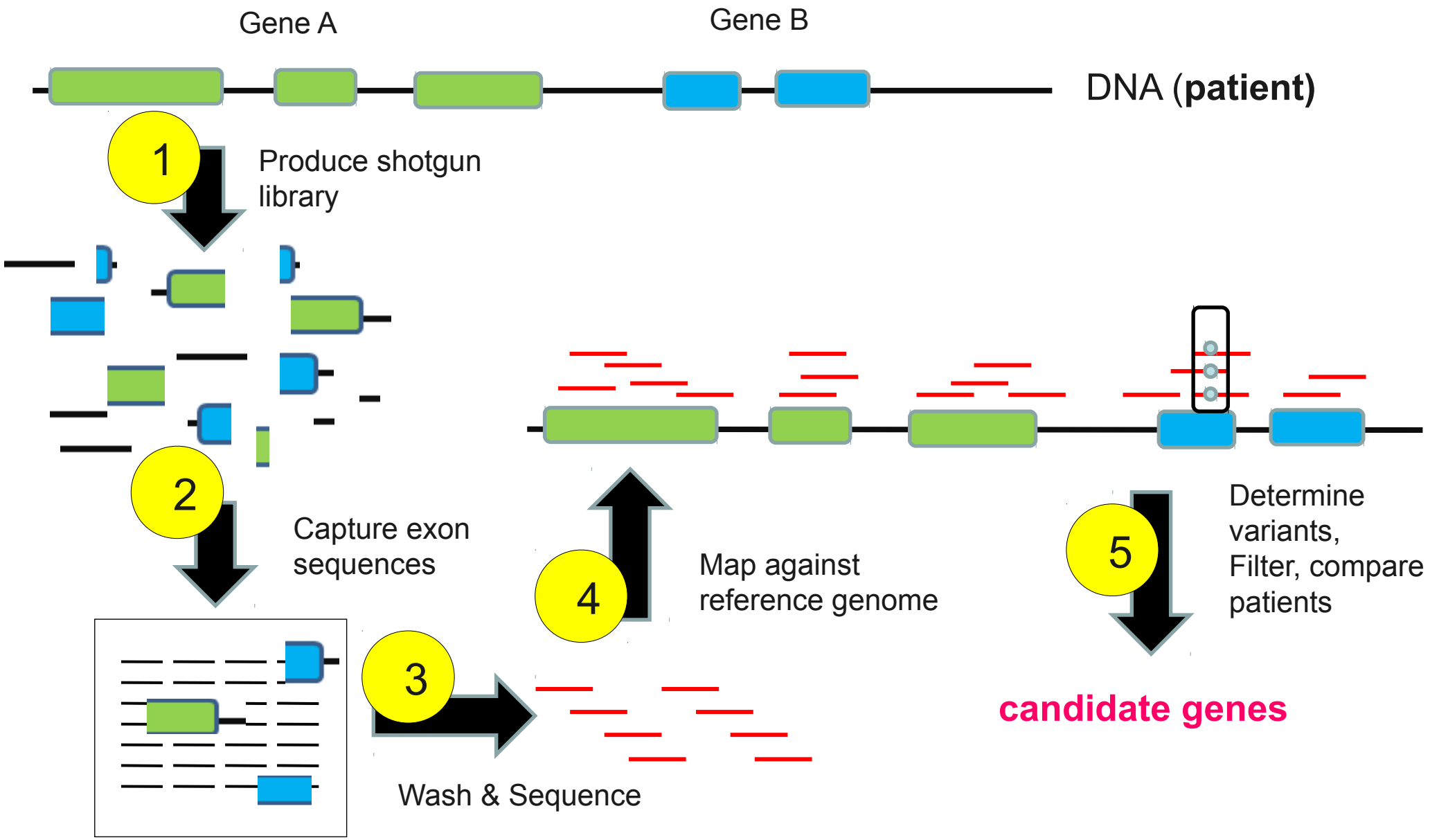
- Known allelic variants known to underlie Mendelian disorders disrupt protein-coding sequences

- Large fraction of rare non-synonymous variants in human genome are predicted to be deleterious

- Splice acceptor and donor sites are also enriched for highly functional variation and are therefore targeted as well

The exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes

How does exome sequencing work?



Exome sequencing makes medical genomics a reality

Leslie G Biesecker

Massively parallel sequencing of the exomes of four individuals with Miller syndrome, combined with filtering to exclude benign and unrelated variants, has identified causative mutations in *DHODH*. This approach will accelerate discovery of the genetic bases of hundreds of other rare mendelian disorders.

The genes underlying mendelian disorders have for the past several decades been identified through positional cloning, a process of meiotic mapping, physical mapping and candidate-gene sequencing[1]. Recently, whole-exome sequencing combined with a filtering methodology was demonstrated as an approach to identify the gene underlying a mendelian disorder using a small number of affected individuals, with a proof-of-concept study that correctly identified the gene previously known to underlie Freeman-Sheldon syndrome[2]. Now, on page 30 of this issue, Michael Bamshad and colleagues[3] report the gene underlying an uncharacterized mendelian disorder, Miller syndrome, using the same strategy. Miller syndrome, also known as postaxial acrofacial dysostosis (MIM#263750), is a rare malformation syndrome that comprises anomalies including cleft palate, absent digits, ocular anomalies and others. The identification of the gene mutated in this disorder will allow improved diagnosis and a starting

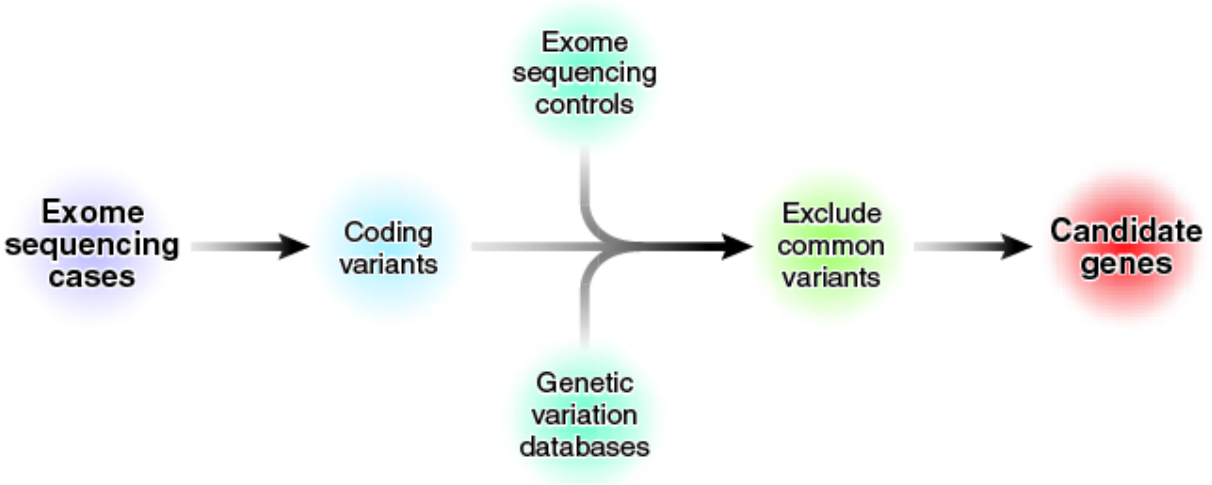


Figure 1 Exome sequencing and filtering strategy. In Ng *et al.*[3], the list of variants from the exome sequences of four individuals with Miller syndrome was first screened to select for genes found to have two nonsynonymous, splice site or indel sequence variants in each of the individuals. This list was then compared to the exome sequences of eight healthy controls[2] and dbSNP to exclude common variation and combined with a filtering strategy used to narrow the list of likely candidate genes underlying this rare disorder.

in a massively parallel short-read sequencer to sequence at ~40-fold coverage. They used a stepwise filtering approach to screen the iden-

the brute-force approach of exome sequencing combined with filtering that identified the disease-causing gene. This new approach

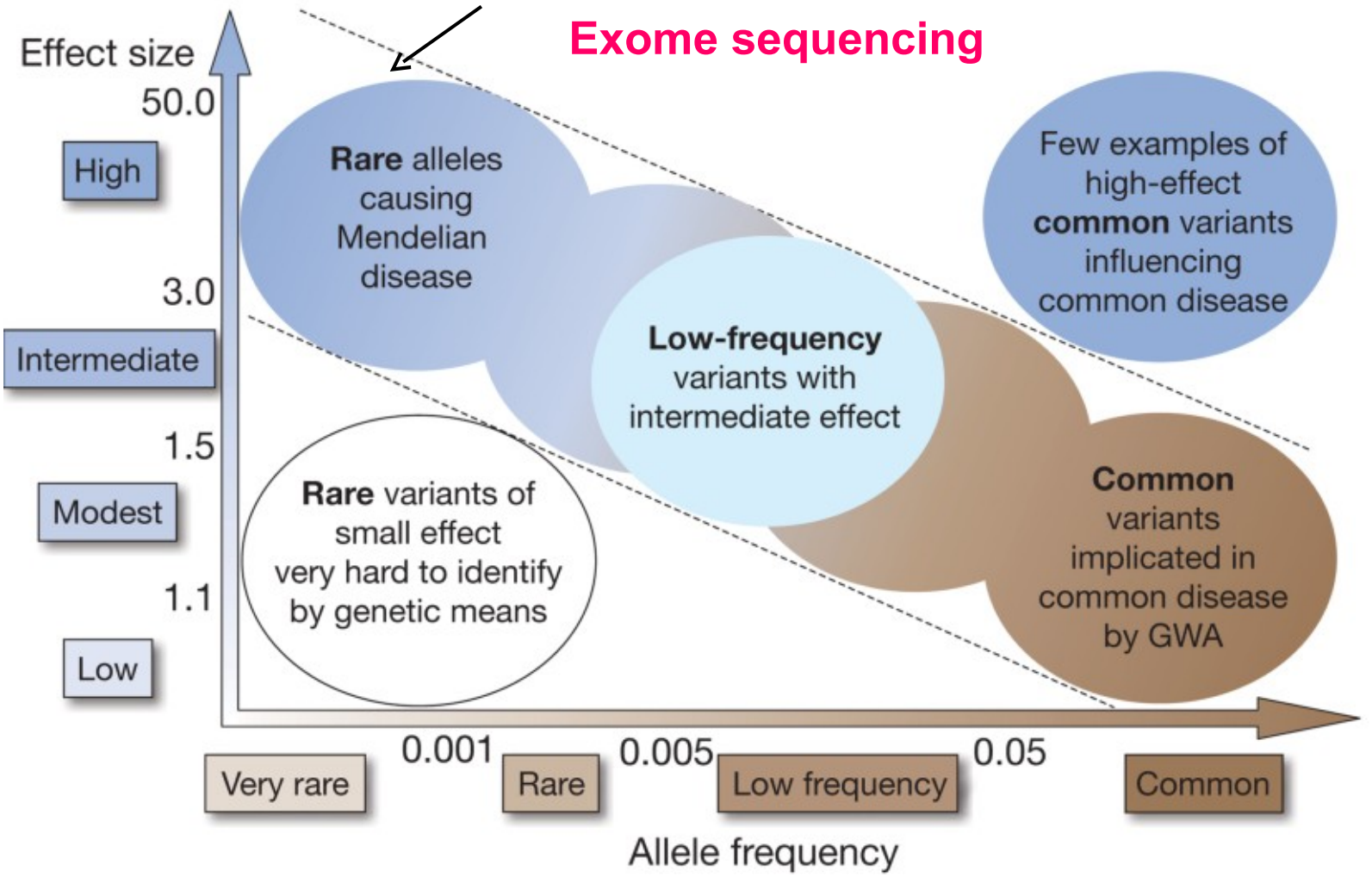
Exome sequencing
Common Research Goals

- ✓Identify novel genes responsible for monogenic diseases
- ✓Use the results of genetic research to discover new drugs acting on new targets (new genes associated with human disease pathways)
- ✓Identify susceptibility genes for common diseases

Challenges

To develop innovative bioinformatics tools for the detection and characterisation of mutations using genomic information.

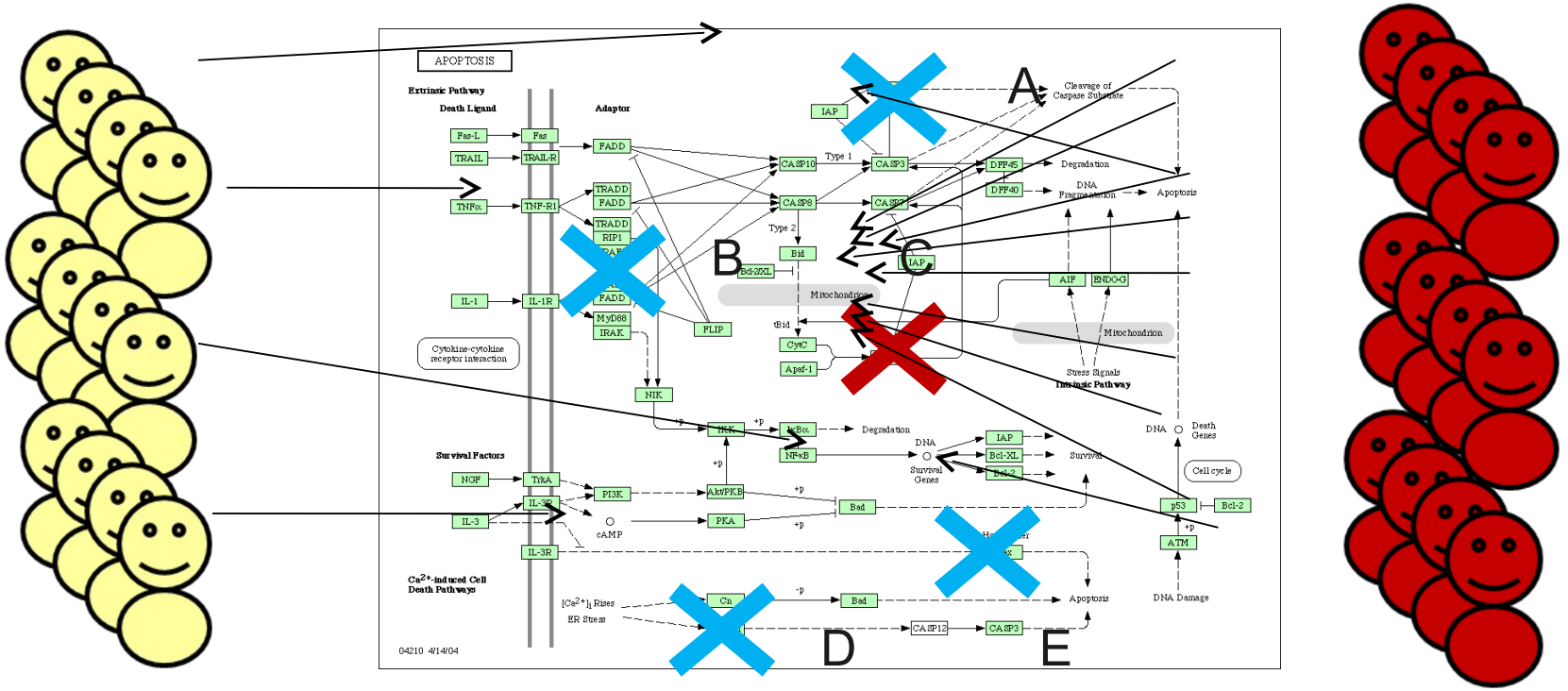
Rare and common disorders



Manolio TA, et al (2009). Finding the missing heritability of complex diseases. Nature., 461(7265), 747

Finding mutations associated to diseases

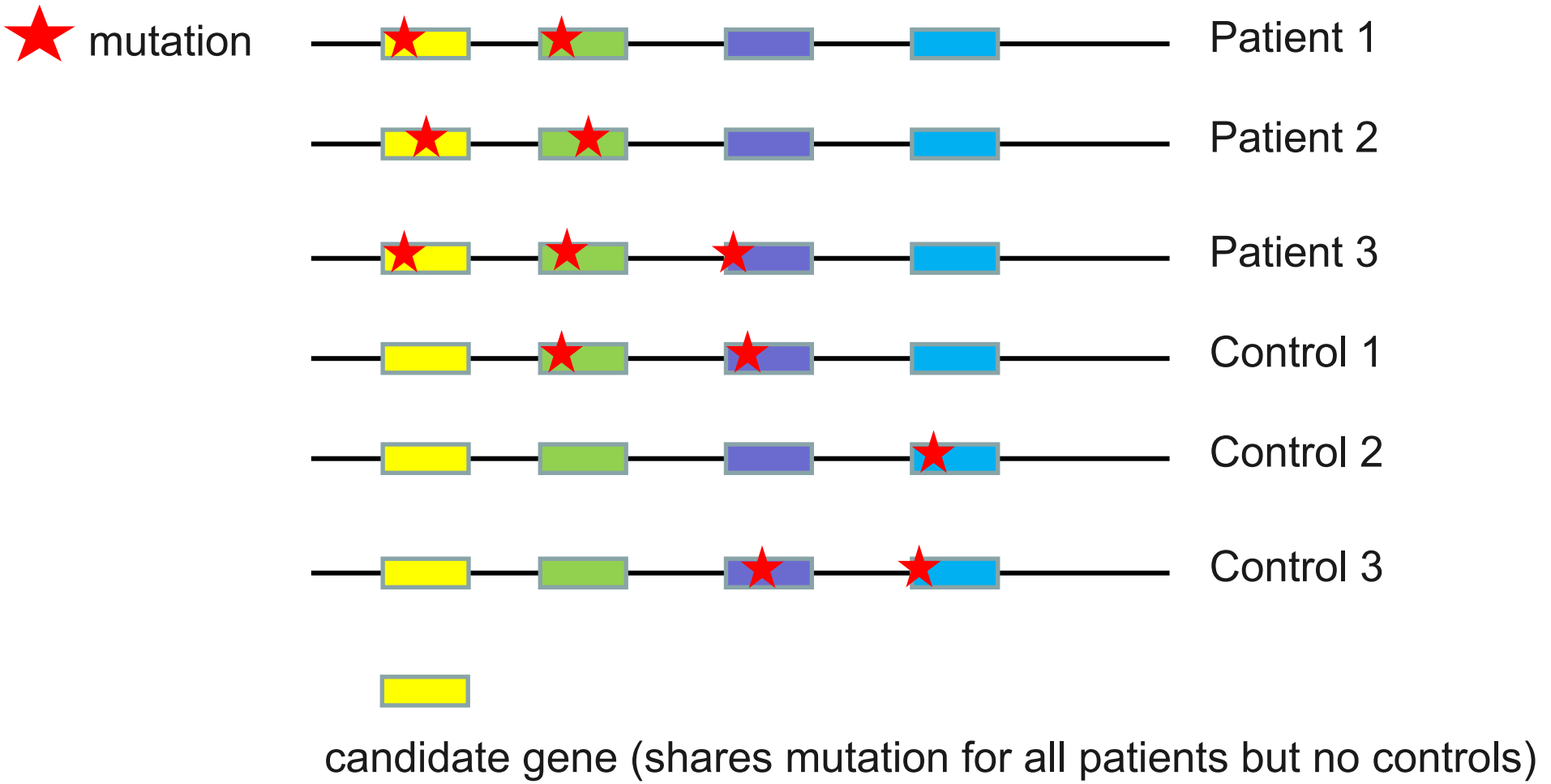
The simplest case: dominant monogenic disease



Controls

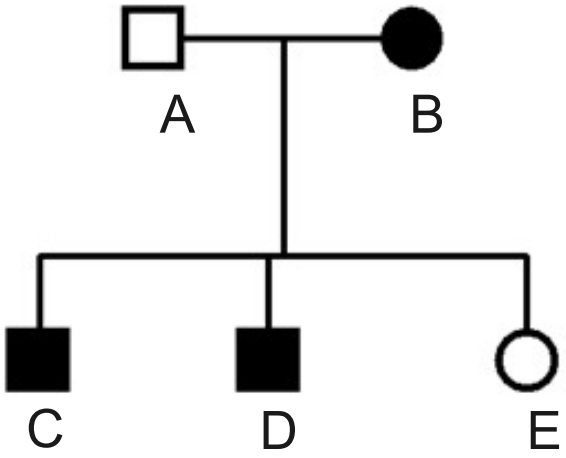
Cases

The principle: comparison of patients and reference controls



Different levels of complexity

- Diseases can be dominant or recessive
- Diseases can have incomplete penetrancy
- Patients can be sporadic or familiar
- Controls can or cannot be available

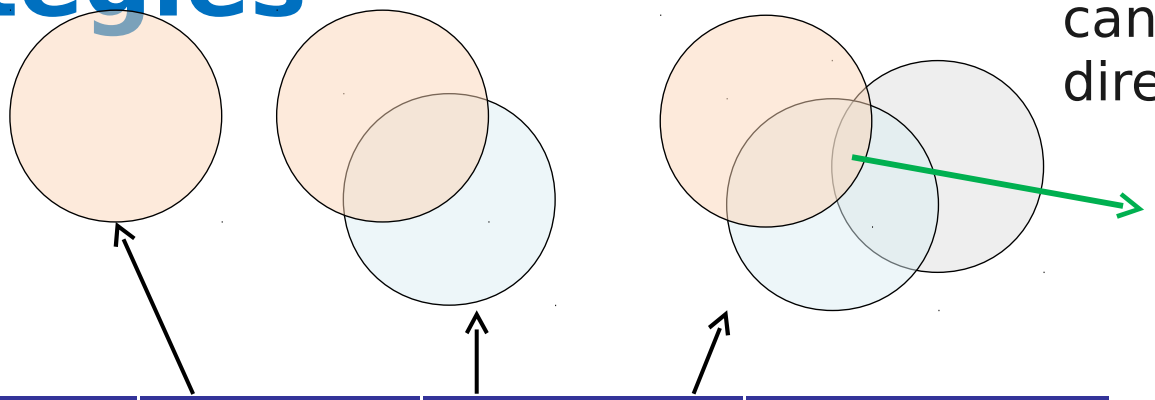


Dominant: (hetero in B, C and D) AND (no in A and E) AND no in controls

Recessive: (homo in B, C and D) hetero in A and D AND NO homo in controls

Ad-hoc strategies are needed for the analysis

Filtering Strategies

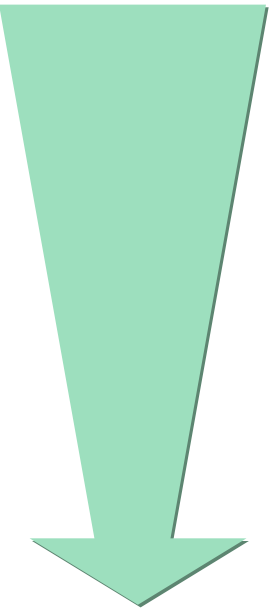


Reducing the number of candidate genes in two directions

Share Genes with variations

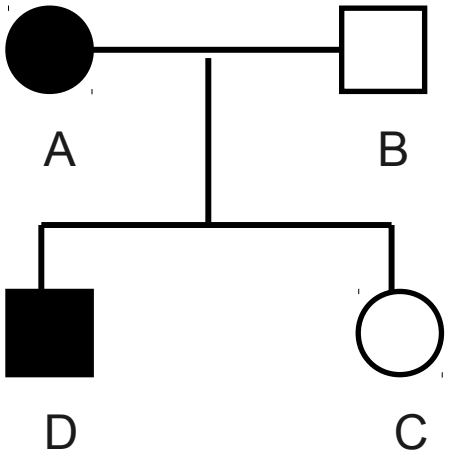
	Patient 1	Patient 2	Patient 3
No filtering	genes with variant	genes with variant	genes with variant
Remove known variants	genes with variant	genes with variant	genes with variant
Remove synonymous variants	genes with variant	genes with variant	genes with variant
...			

Several shared Genes



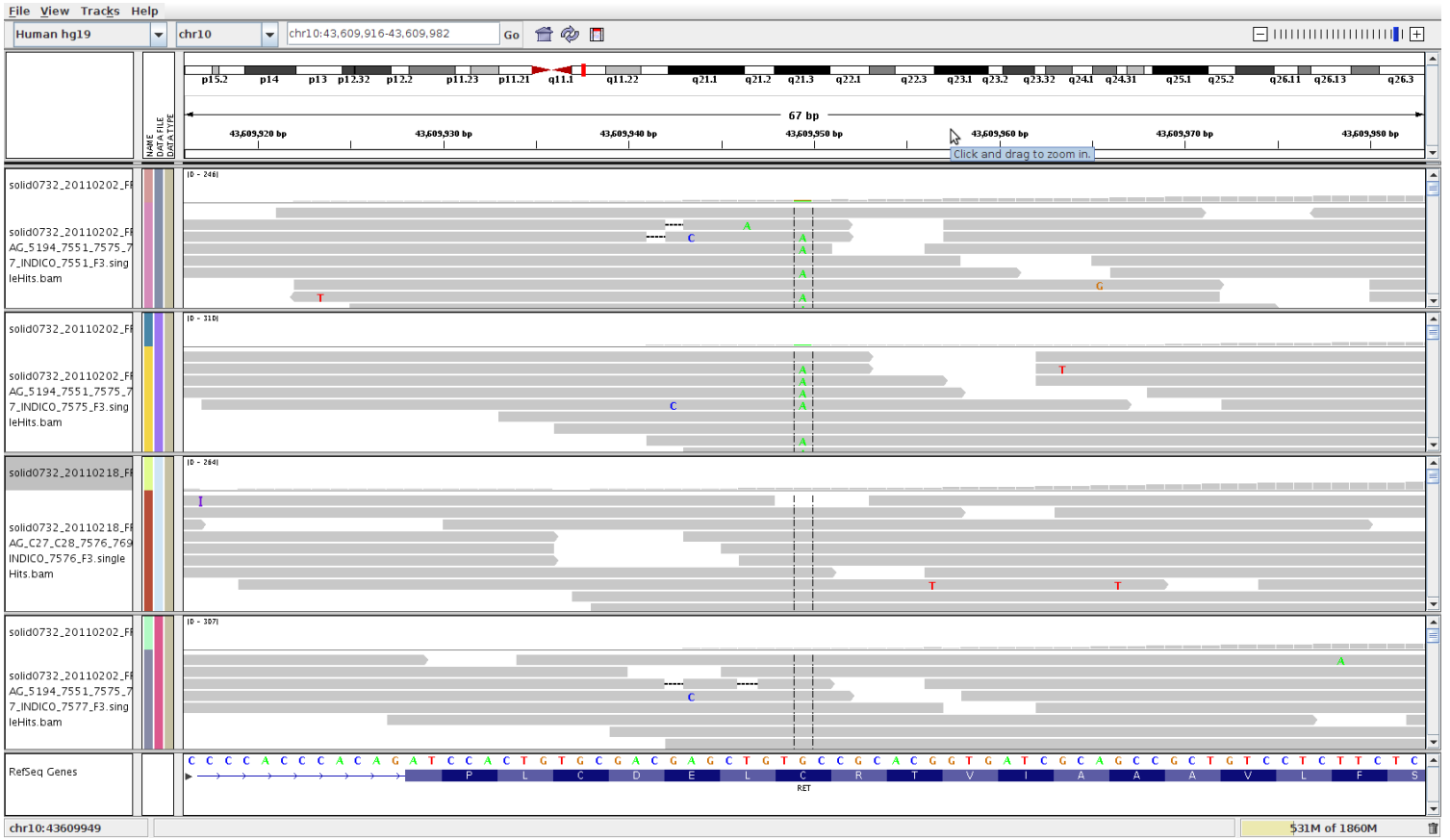
Few shared Genes

An example with MTC



Dominant:
Heterozygotic in A and D
Homozygotic reference allele in B and C
Homozygotic reference allele in controls

RET, codon 634 mutation



A

D

B

C

The Pursuit of Better and more Efficient Healthcare as well as Clinical Innovation through Genetic and Genomic Research

Public-Private partnership

- ✓**Autonomous Government of Andalusia**
- ✓**Spanish Ministry of Innovation**
- ✓**Pharma and Biotech Companies**

MGP Specific objectives

- ✓ To sequence the genomes of clinically well characterized patients with potential mutations in novel genes.
- ✓ To generate and validate a database of genomes of phenotyped control individuals.
- ✓ To develop bioinformatics tools for the detection and characterisation of mutations

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

SAMPLES + UPDATED AND COMPREHENSIVE HEALTH RECORD

Currently14,000 *Phenotyped* DNA Samples

from patients and control individuals.

Prospective Healthcare: linking research & genomic information to health record databases

Patient health & sample record real time automatic update and comprehensive data mining system

JUNTA DE ANDALUCIA
Consejería de Salud

Hospitales Universitarios Virgen del Rocío

Estación Clínica (SIDCA) /Clinical Work Station

From Information Management to Clinical Knowledge Management

SIDCA Bio e-Bank

Andalusian DNA Bank

Direct link to the health care system

MPG roadmap is based on the availability of >14.000 well-characterized samples with a permanent updated sample information & PHR that will be used as the first steps of the implementation of personalized medicine in the Andalusian HCS

>14.000 phenotyped samples

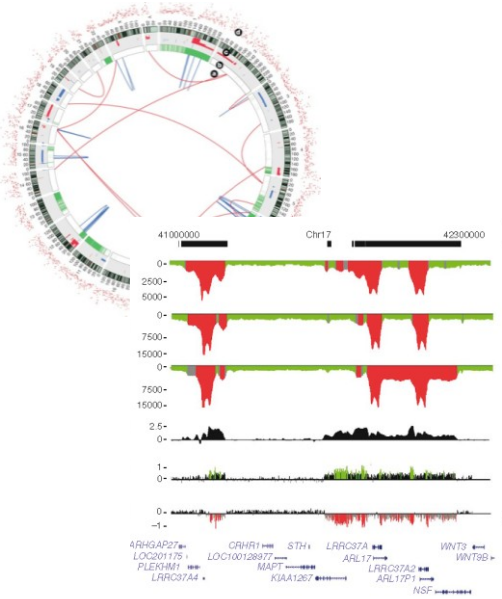
- Cancer
- Congenital anomalies (heart, gut, CNS,…)
- Mental retardation
- MCA/MR syndromes
- Diabetes
- Neurodegenerative diseases
- Stroke (familiar)
- Endometriosis
- Control Individuals

with

- Unknown genes
- Known genes discarded
- Responsible genes known but unknown modifier genes
- Susceptibility Genes
- …

Two technologies to scan for variations

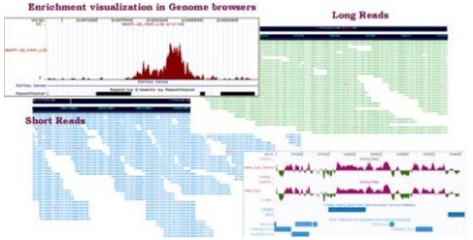
454 Roche
Longer reads
Lower coverage



Structural variation

- Amplifications
- Deletions
- CNV
- Inversions
- Translocations

SOLiD ABI
Shorter reads
Higher coverage



Variants

- SNPs
- Mutations
- indels

Analysis Pipeline

Raw Data Quality Control QC

Reads Mapping & QC

Variant Calling

Variant Filtering

Variant annotation

Disease/Control comparison

FastQC & in house software

BWA, Bowtie, BFAST
QC in house software

GATK, SAMTOOLS, FreeBayes

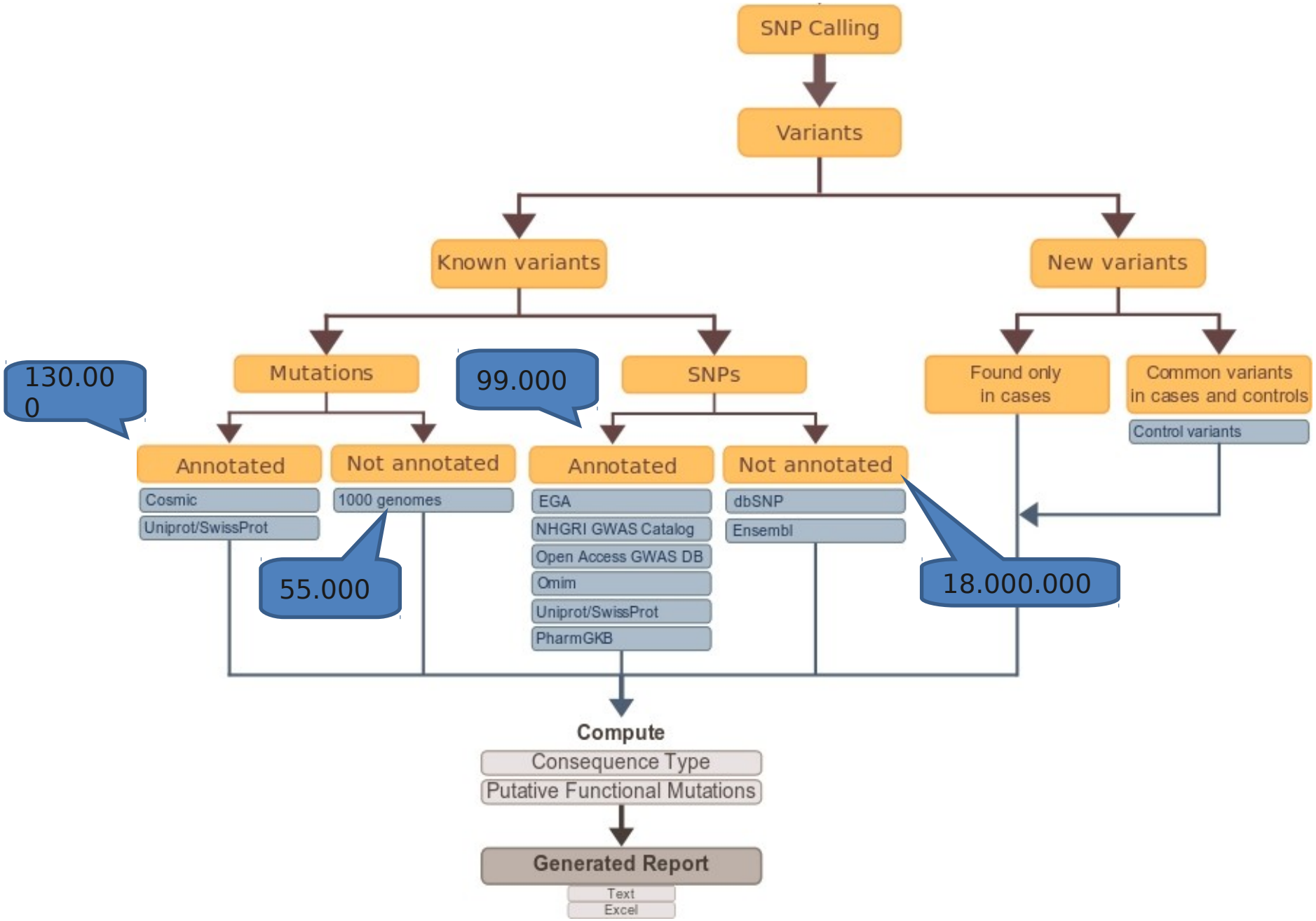
dbSNP, 1000 Genomes

Annovar & in house software

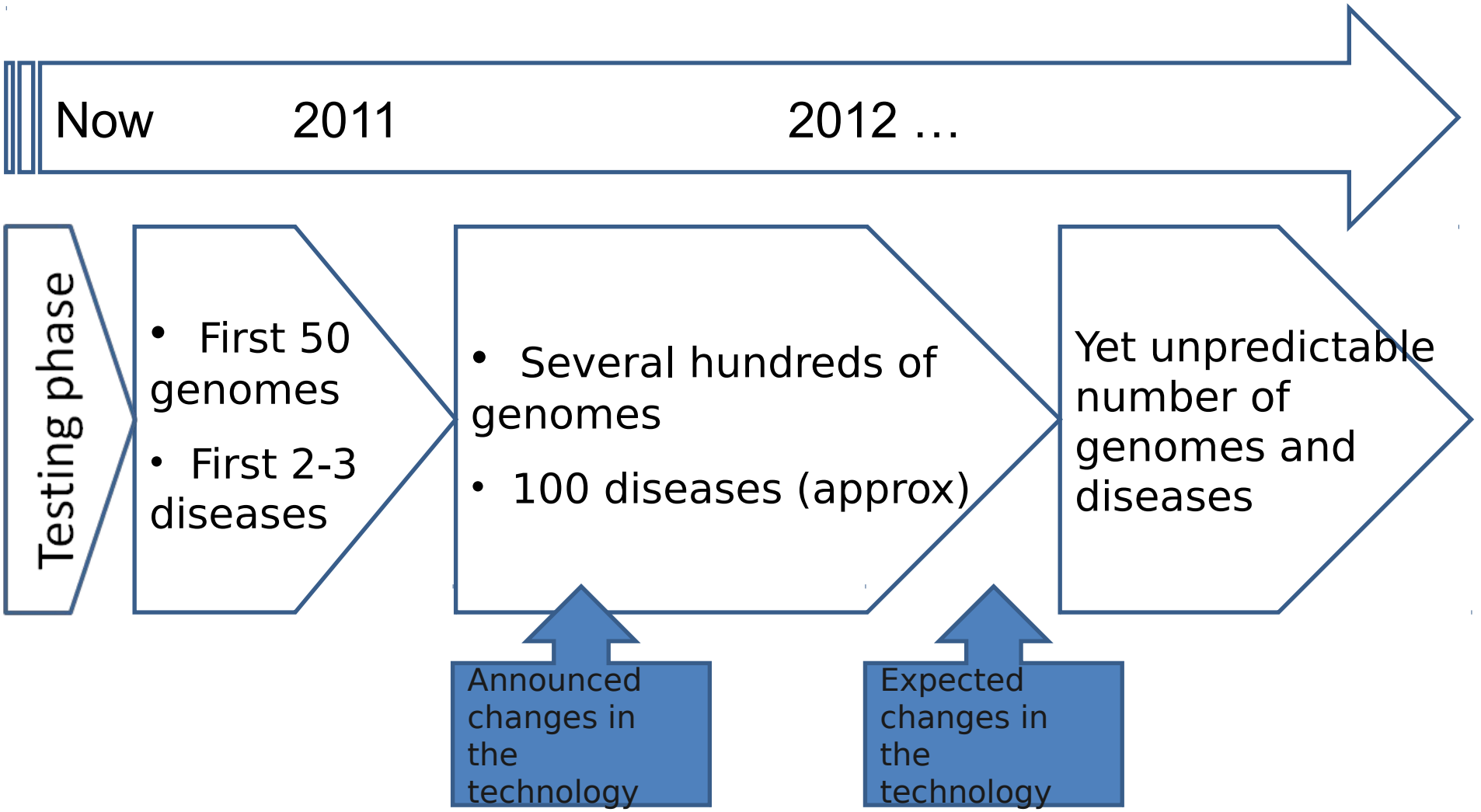
Family & healthy controls

Bioscope

Approach to discovery rare or novel variants



Timeline of the MGP




SOLiD 4 Analyzer

Nimblegen capture arrays

- SeqCap EZ Human Exome Library **v1.0** / v2.0

- Gene and exon annotations (v2.0):
  - RefSeq (Jan 2010), CCDS (Sept 2009), miRBase (v.14, Sept 2009).

- Total of ~30,000 coding genes (theoretically)
  - ~300,000 exons;
  - 36.5 Mb are targeted by the design.

- 2.1 million long oligo probes to cover the target regions.
  - Because some flanking regions are also covered by probes, the total size of regions covered by probes is 44.1 Mb

- Real coverage:
  - Coding genes included: 18897
  - miRNAs 720
  - Coding genes not captured: 3865

Sequencing at MGP

By the end of June 2011 there are 72 exomes sequenced so far.

4 SOLiD can produce 20 exomes per week

The facilities of the CASEGH can carry out the MGP and other collaborative projects at the same time

Samples

Samples Results

Edit

Sample	Platform	Sequencer	FlowCell	Date in Sequencer	Date in cluster	Reads QC	Phenotype	status	Relationship	Family (mother's id)
C3	solid	M(editerraneo)	f1	20101112	2011-02-04	Pass	control	Healthy		
C11	solid	M	f1	20101112	2011-02-04	Pass	control	Healthy		
C14	solid	M	f1	20101112	2011-02-04	Pass	control	Healthy		
C37	solid	M	f1	20101112	2011-02-04	Pass	control	Healthy		
C22	solid	M	f2	20101112	2011-02-04	Fail	control	Healthy		
C23	solid	M	f2	20101112	2011-02-04	Fail	control	Healthy		
C24	solid	M	f2	20101112	2011-02-04	Fail	control	Healthy		
C25	solid	M	f2	20101112	2011-02-04	Fail	control	Healthy		
C15	solid	A(tlantico)	f1	20110118	2011-01-11	Pass	control	Healthy		
C16	solid	A	f1	20110118	2011-01-11	Pass	control	Healthy		
C28	solid	A	f1	20110118	2011-01-11	Pass	control	Healthy		
C29	solid	A	f1	20110118	2011-01-11	Pass	control	Healthy		
C33	solid	A	f2	20110118	2011-01-11	Pass	control	Healthy		
C40	solid	A	f2	20110118	2011-01-11	Pass	control	Healthy		
C41	solid	A	f2	20110118	2011-01-11	Pass	control	Healthy		
C42	solid	A	f2	20110118	2011-01-11	Pass	control	Healthy		
C38	solid	I(ndico)	f1	20110202	2011-02-11	Pass	control	Healthy		
C43	solid	I	f1	20110202	2011-02-11	Pass	control	Healthy		
4128	solid	I	f1	20110202	2011-02-11	Pass	RP	Affected	daughter	4125
5193	solid	I	f1	20110202	2011-02-11	Pass	RP	Affected	daughter	3865
5194	solid	I	f2	20110202	2011-02-11	Pass	RP	Affected	son	3865
7551	solid	I	f2	20110202	2011-02-11	Pass	MTC	Affected	Mother	7551
7575	solid	I	f2	20110202	2011-02-11	Pass	MTC	Affected	son	7551
7577	solid	I	f2	20110202	2011-02-11	Pass	MTC	Healthy	daughter	7551
7445	solid	A	f1	20110204	2011-02-11	Pass	FQ	Affected	daughter	7447
7446	solid	A	f1	20110204	2011-02-11	Pass	FQ	Healthy	Father	7447

Real coverage and some figures

Sequencing

Enrichment +Sequence run: ~2 weeks
400,000,000 sequences/flowcell
20,000,000,000 bases/flowcell
Short 50bp sequences

Exome Coverage

SeqCap EZ Human Exome Library v1.0 / v2.0
Total of ~30,000 coding genes (theoretically)

~300,000 exons;
36.5 Mb are targeted by the design (2.1 million long oligo probes).

Real coverage:

Coding genes included:	18,897
miRNAs	720
Coding genes not captured:	3,865
Genes of the exome with coverage >10x:	**85%**

Data Simulation & analysis

- Data simulated for 80K mutations and 60x coverage

				SNPs			INDELS		
SAMPLE	Number of Reads	Bases	Number of Mutations	Total	Homozygous	Heterozygous	Total	Homozygous	Heterozygous
80K60x2	100,000,071	5.00E+09	82328	73383	24431	48952	8945	2917	6028

	SNP									
	TOTAL	True Positives			False Positives			False Negatives		
		Total	Homozygous	Heterozygous	Total	Homozygous	Heterozygous	Total	Homozygous	Heterozygous
Bfast + GATK	73981	68815	22666	46149	5166			4568	1765	2803
Bowtie + GATK	86118	70768	23924	46844	15350			2615	507	2108
Bwa + GATK	70266	69949	23619	46330	317			3434	812	2622

	INDEL									
	TOTAL	True Positives			False Positives			False Negatives		
		Total	Homozygous	Heterozygous	Total	Homozygous	Heterozygous	Total	Homozygous	Heterozygous
Bfast + GATK	5780	4904	1810	3094	876			4041	1107	2934
Bowtie + GATK	0	0	0	0	0			8945	2917	6028
Bwa + GATK	5108	3484	1417	2067	1624			5461	1500	3961

- BWA finds less false positives in simulated data

Real Data & analysis

- Analiysis data is compared to genotyped data
  - BFAST higher number of variants and 95% agreement

			Bwa			Bowtie		Bfast		
Date	Sample	Reads	Coverage	Snps	indels	Coverage	Snps	Coverage	Snps	indels
2011/02/03	C3	95×10^6	36.36x	17.061	203	49.08x	23.776	54.09x	49.385	531
2011/02/03	C37	97×10^6	35.34x	27.858	225	48.90x	44.725	52.53x	103.784	1.379
2011/02/03	C11	66×10^6	26.38x	15.260	162	37.36x	22.836	41.05x	63.500	598
2011/02/03	C14	100×10^6	39.68x	18.032	207	51.76x	24.313	56.63x	45.497	589
2011/02/03	C22	97×10^6	28.26x	13.386	140	44.26X	22.387	50.19x	49.416	536
2011/02/03	C23	101×10^6	34.12x	14.144	180	54.23x	24.111	62.60x	53.314	538
2011/02/03	C24	71×10^6	18.27x	12.044	119	28.81x	20.859	34.74x	49.354	522
2011/02/03	C25	98×10^6	29.72x	18.988	172	50.21x	28.710	55.87x	76.945	1.108
2011-03-10	465	97,792,819	35.29x	23295	337	51.04x	35180	53.02x	113170	1493
2011-03-10	466	108,152,596	46.95x	19705	223	62.03x	26135	65.04x	75922	595
2011-03-10	469	90,286,940	41.67x	18939	251	53.45x	25474	55.53x	60596	576
2011-03-10	C12	101,223,049	33.15x	16120	194	46.56x	22580	51.56x	63104	579

And this is what we find in the variant calling pipeline

Coverage > 50x
Variants (SNV): 60.000 – 80.000
Variants (indels): 600-1000
100 known variants associated to disease risk

none	1	2116429	C	missense	0	PRKCZ,
none	1	2116429	C	missense	0	PRKCZ,
none	1	2116429	C	missense	0	PRKCZ,
none	1	2116429	C	utr-3	0	PRKCZ,
none	1	2318893	C	missense	0	MORN1
none	1	2452167	C	missense	0	PANK4
dbSNP_1000Genomes	1	2452569	T	coding-synonymous	2985862	PANK4
none	1	3680294	A	missense	0	CCDC27
none	1	3745852	T	missense	0	KIAA05
none	1	3746432	G	missense	0	KIAA05
dbSNP_1000Genomes	1	3755675	T	coding-synonymous	1891941	KIAA05
none	1	6029181	G	missense	0	NPHP4
none	1	6101899	A	missense	0	KCNAB
none	1	6101899	A	intron	0	KCNAB
none	1	6132842	C	coding-synonymous	0	KCNAB
none	1	6132842	C	coding-synonymous	0	KCNAB
none	1	6535559	T	missense	0	PLEKH
none	1	6535559	T	missense	0	PLEKH
none	1	6535559	T	missense	0	PLEKH
none	1	6535559	T	missense	0	PLEKH
none	1	6535559	T	missense	0	PLEKH
none	1	6647590	A	missense	0	ZBTB48
none	1	6694129	T	missense	0	THAP3
none	1	6695719	T	utr-3	0	DNAJC1
none	1	6704720	C	missense	0	DNAJC1
none	1	6711636	C	coding-synonymous	0	DNAJC1
dbSNP_1000Genomes	1	7889941	C	coding-synonymous	2640908	PER3
dbSNP_1000Genomes	1	7890117	T	missense	2640909	PER3
dbSNP_1000Genomes	1	8425900	T	coding-synonymous	3753275	RERE
dbSNP_1000Genomes	1	8425900	T	utr-5	3753275	RERE
dbSNP_1000Genomes	1	8425900	T	coding-synonymous	3753275	RERE
none	1	9086361	C	missense	0	SLC2A7
none	1	9117600	A	missense	0	SLC2A5
none	1	9117600	A	missense	0	SLC2A5
none	1	9129619	C	utr-5	0	SLC2A5
none	1	9129619	C	utr-5	0	SLC2A5
none	1	9770594	C	coding-synonymous	0	PIK3CD
none	1	10042458	A	missense	0	NMNAT

```
Known snps phenotypic effect
-----------------------------------------

Hits     Description
8        Amyotrophic Lateral Sclerosis (ALS)
7        Parkinson's disease
6        Rheumatoid Arthritis
5        common polymorphism
4        Multiple complex diseases-Crohn's disease, combined control
      dataset
3        Alzheimer's Disease
3        LDL cholesterol
3        Skin pigmentation
3        Type 1 diabetes
2        Coronary Artery Disease
2        in allele DQB1*0501 and allele DQB1*0502
2        Multiple complex diseases-Bipolar disorder
2        Multiple complex diseases-Coronary Artery Disease, gender
     differentiated
2        Multiple complex diseases-Crohn's disease, combined control
     dataset, gender differentiated
2        Multiple complex diseases-Type I Diabetes, combined control
     dataset
2        Multiple complex diseases-Type II Diabetes Mellitus,
    combined control dataset
2        Systemic Lupus Erythematosus (SLE), gender differentiated
    in women
2        Triglycerides
2        Type I Diabetes
1        893Ser-expressing (ABCB1:2677G>T (Ala893Ser)) cells showed
```

From data to knowledge
Some considerations

Obvious: huge datasets need to be managed by computers
Important: bioinformatics is necessary to properly analyze the data.
Even more important and not so obvious: hypotheses must be tested from the perspective provided by the bioinformatics

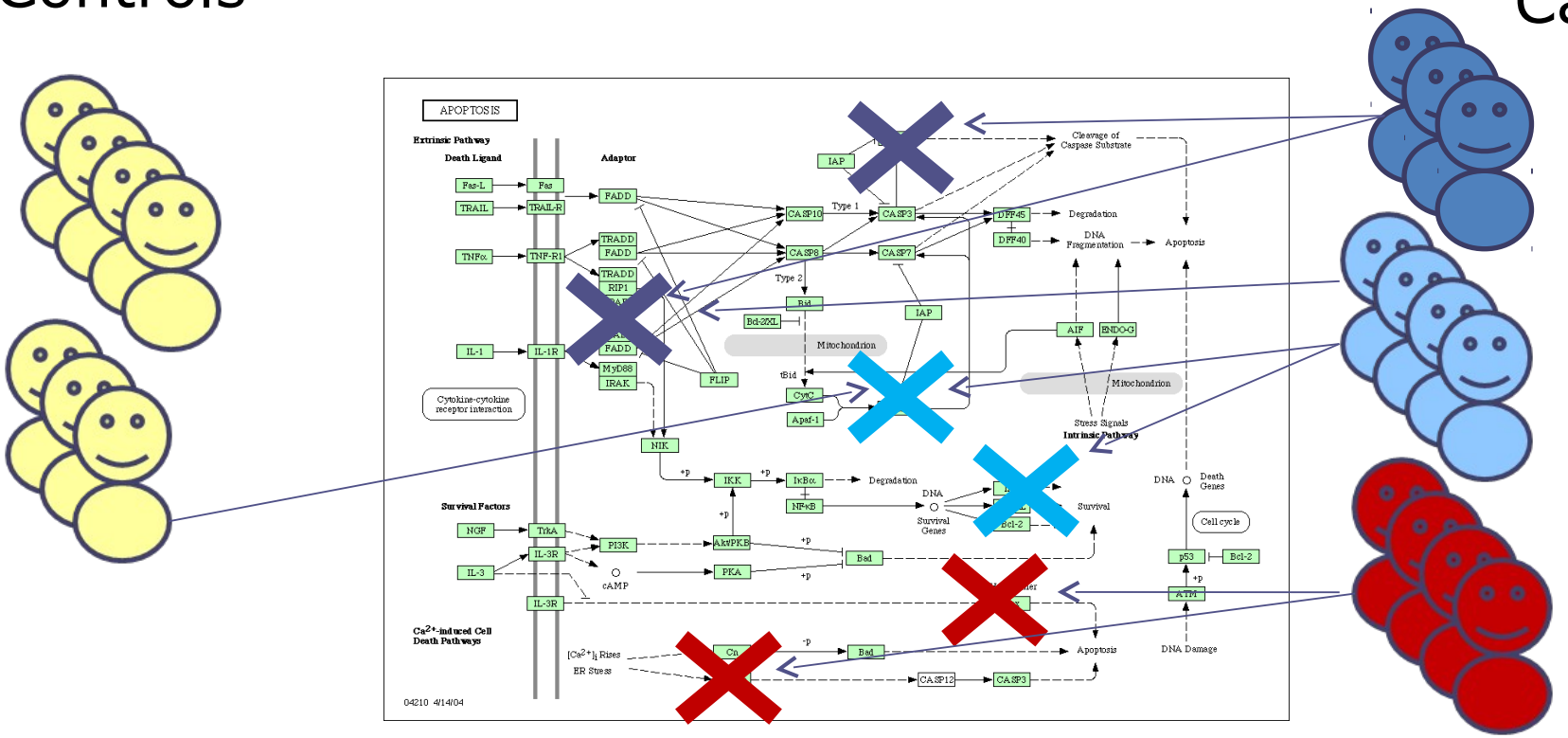
The science is generated where the data reside.
Yesterday's "one-bite" experiments fit into a laboratory notebook. Today, terabite data from genomic experiments reside in computers, the new scientist's notebook

And It gets more complicated

Context and cooperation between genes cannot be ignored

Controls

Cases



The cases of the multifactorial disease will 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.

The Bioinformatics and Genomics Department at the Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, and…

Joaquín Dopazo
Eva Alloza
Roberto Alonso
Alicia Amadoz
Davide Baù
Jose Carbonell
Ana Conesa
Alejandro de Maria
Hernán Dopazo
Pablo Escobar
Fernando García
Francisco García
Luz García
Stefan Goetz
Carles Llacer
Martina Marbà
Marc Martí
Ignacio Medina
David Montaner
Luis Pulido
Rubén Sánchez
Javier Santoyo
Patricia Sebastian
François Serra
Sonia Tarazona
Joaquín Tárraga
Enrique Vidal
Adriana Cucchi

...the INB, National Institute of Bioinformatics (Functional Genomics Node) and the CIBERER Network of Centers for Rare Diseases

CAG

Área de Genómica

Dr. Rosario Fernández Godino

Dr. Alicia Vela Boza

Dr. Slaven Erceg

Dr. Sandra Pérez Buira

María Sánchez León

Javier Escalante Martín

Ana Isabel López Pérez

Beatriz Fuente Bermúdez

Área Bioinformática

Daniel Navarro Gómez

Pablo Arce García

Juan Miguel Cruz

Secretaría/Administración

Inmaculada Guillén Baena

HOSPITAL UNIVERSITARIO VIRGEN DEL ROCÍO

Dr. Macarena Ruiz Ferrer

Nerea Matamala Zamarro

Prof. Guillermo Antiñolo Gil

Director de la UGC de Genética, Reproducción y Medicina Fetal Director del Plan de Genética de Andalucía

CABIMER

Director de CABIMER y del Departamento de Terapia Celular y Medicina Regenerativa

Prof. Shom Shanker Bhattacharya,

CENTRO DE INVESTIGACIÓN PRÍNCIPE FELIPE

Responsable de la Unidad de Bioinformática y Genómica y Director científico asociado para Bioinformática del Plan de Genética de Andalucía

Dr. Joaquín Dopazo Blázquez