Whole exome resequencing reveals an unexpected amount of variability with possible functional consequences in human miRNAs

Javier Santoyo-Lopez

Andalusian Human Genome Sequencing Centre (CASEGH)
Seville, Spain
&
Department of Bioinformatics and Genomics,
Prince Felipe Research Centre (CIPF), Valencia, Spain

http://www.medicalgenomeproject.es
Http://bioinfo.cipf.es

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

CAMDA,Vienna - July 15th, 2011

ciberer

CASEGH

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

CASEGH, Jan 2010

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

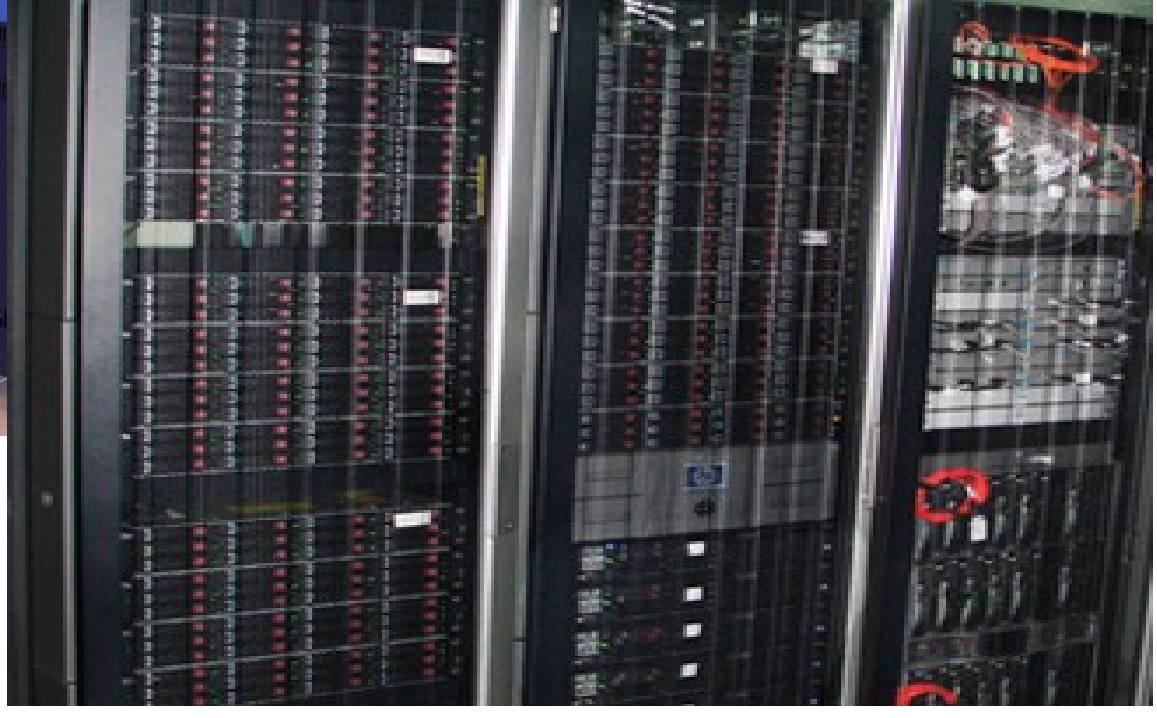
MGP Medical Genome Project

CASEGH, October 2010

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

CASEGH, October 2010



JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

The Medical Genome Project

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 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

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

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

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

New high throughput sequencing technologies will enable researchers to study human diseases, accounting for genetic variability of individual patients and the heterogeneity of their diseases. **However, a major limitation remains the ability to link clinical information to high quality sample collection**.

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

COMPREHENSIVE HEALTH IT SYSTEM

UPDATED AND COMPREHENSIVE PHR LINKED TO SI
Currently14,000 *Phenotyped* DNA Samples from patients and control individuals.

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

+

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

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

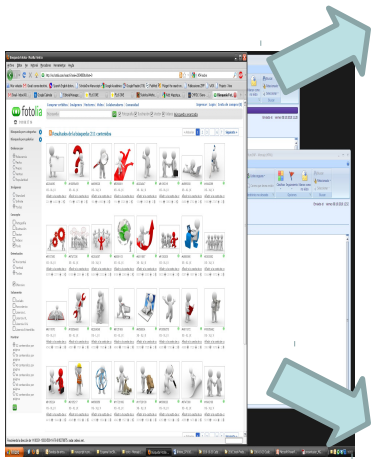
Hospitales Universitarios Virgen del Rocío

Estación Clínica (SIDCA) /Clinical Work Station
From Information Management to Clinical Knowledge Management

SIDCA Bio e-Bank

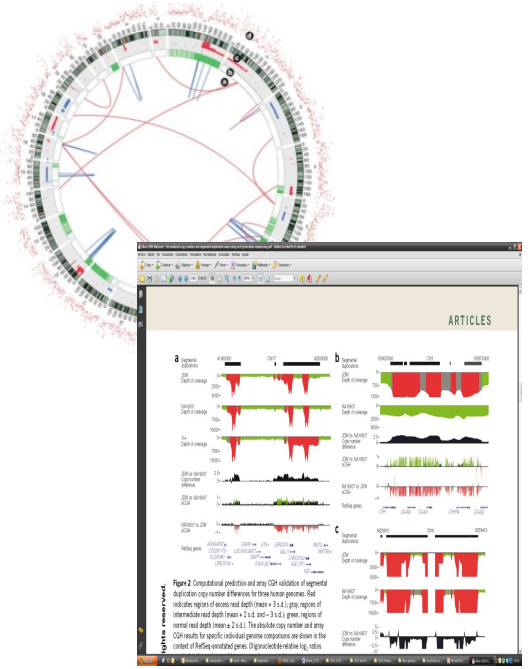
Andalusian DNA Bank

Two Technologies to scan for variations



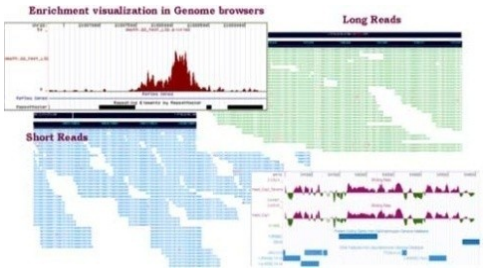
454 Roche
Longer reads
Lower coverage

SOLiD ABI
Shorter reads
Higher coverage



Structural variation

·Amplifications
·Deletions
·CNV
·Inversions
·Translocations



Variants

·SNPs
·Mutations
·indels

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

Bioinformatics analysis pipeline

FastQ
1Mreads

FastQ
Color space
200Mreads

Automatic QC

Preprocess

Position / quality

Base composition

Length statistics

Nucleotide QC

Sequence cleansing

Structural variation detection

Mapping

SNP Calling

SAM to BAM
Sort BAM
Clean BAM
Index BAM

Annotation

Structural variation

Variation

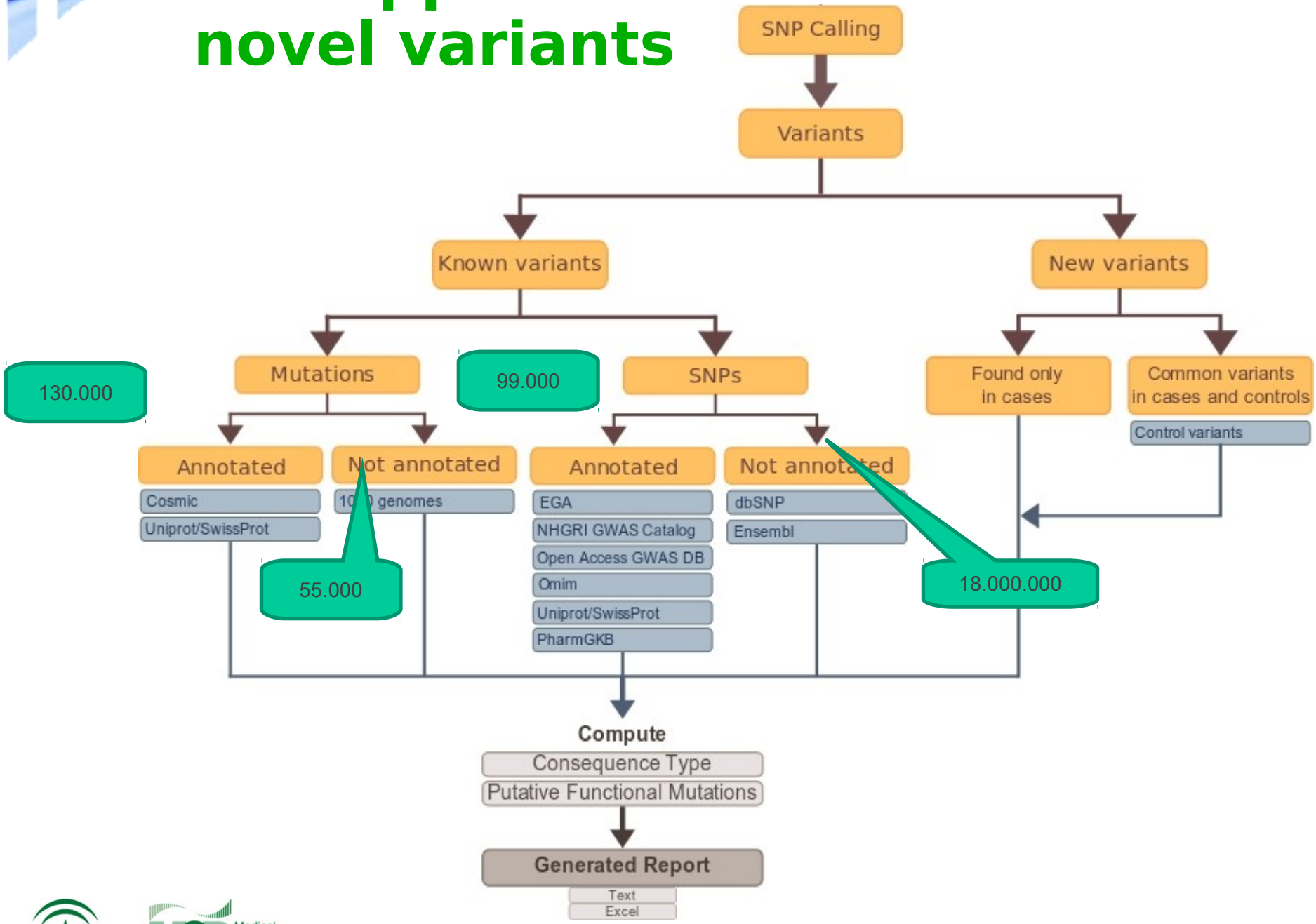
4-6 hours* | 1-2 days* | 1-2 days* | minutes

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

** 8CPUs 200Mreads (>25-30 Exomes per week)*

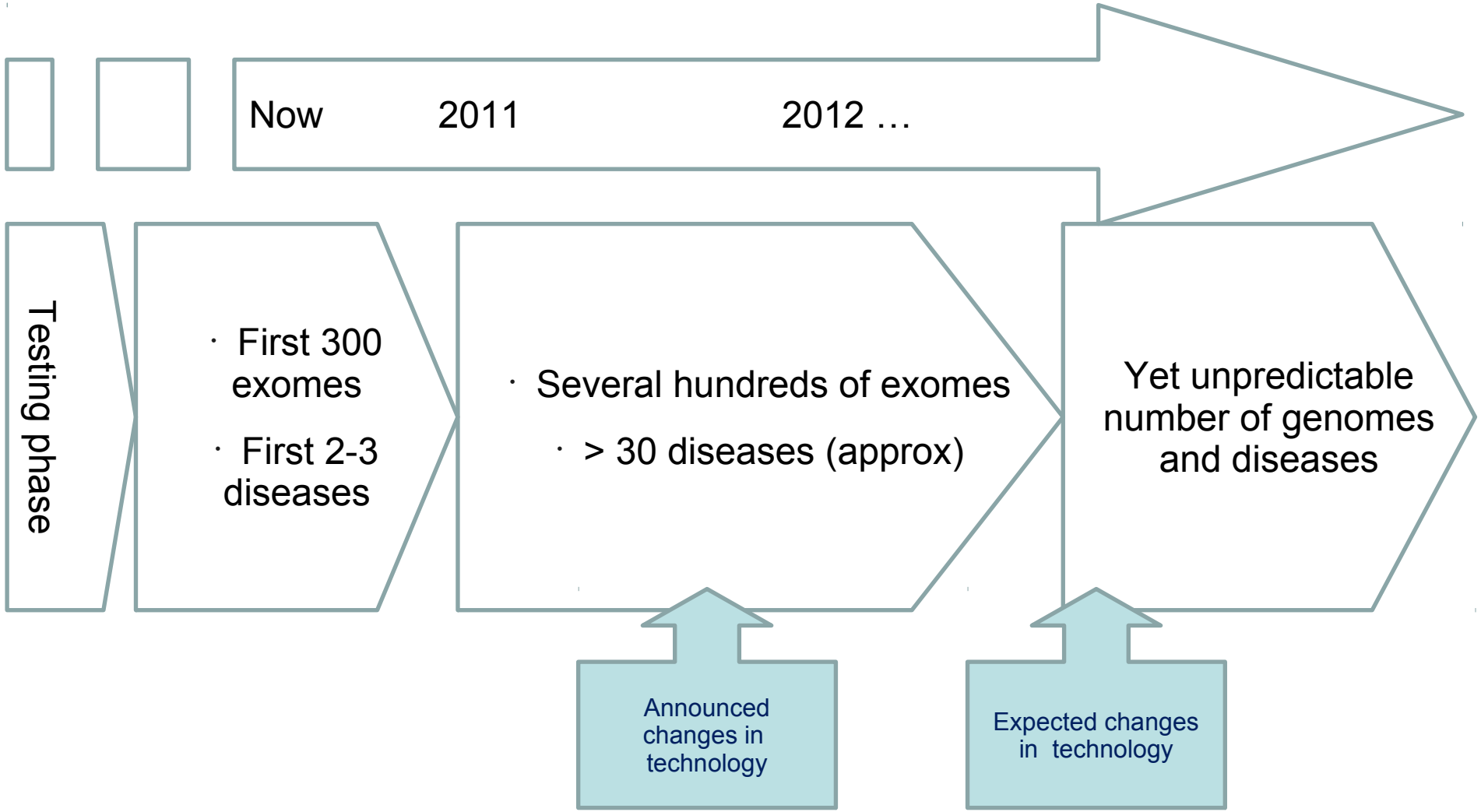
Our approach to discover rare or novel variants



JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

MGP timeline



JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

Sequenced Exomes

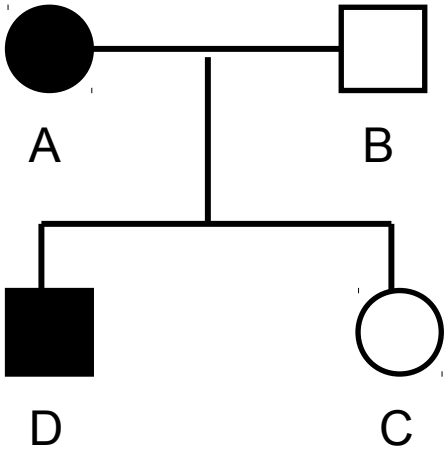
The MGP's goal is to characterize a great number of genetic diseases by means of exome sequencing from affected individuals

- So far using NimbleGen as exome capture platform we have sequenced using SOLiD 4:
  - **163 exomes** corresponding to:
    - 16 samples from individuals with known mutation (FQ, MTC and RPAD) in order to validate the sequencing platform and the bioinformatics analysis pipeline
    - 27 samples from 24 phenotyped healthy controls.
    - 120 samples corresponding to healthy and affected individuals from families that have a disease in study (RP, FQ, HSCR, MTC)
- Reaching production of 24 exomes (60x coverage) per week.

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

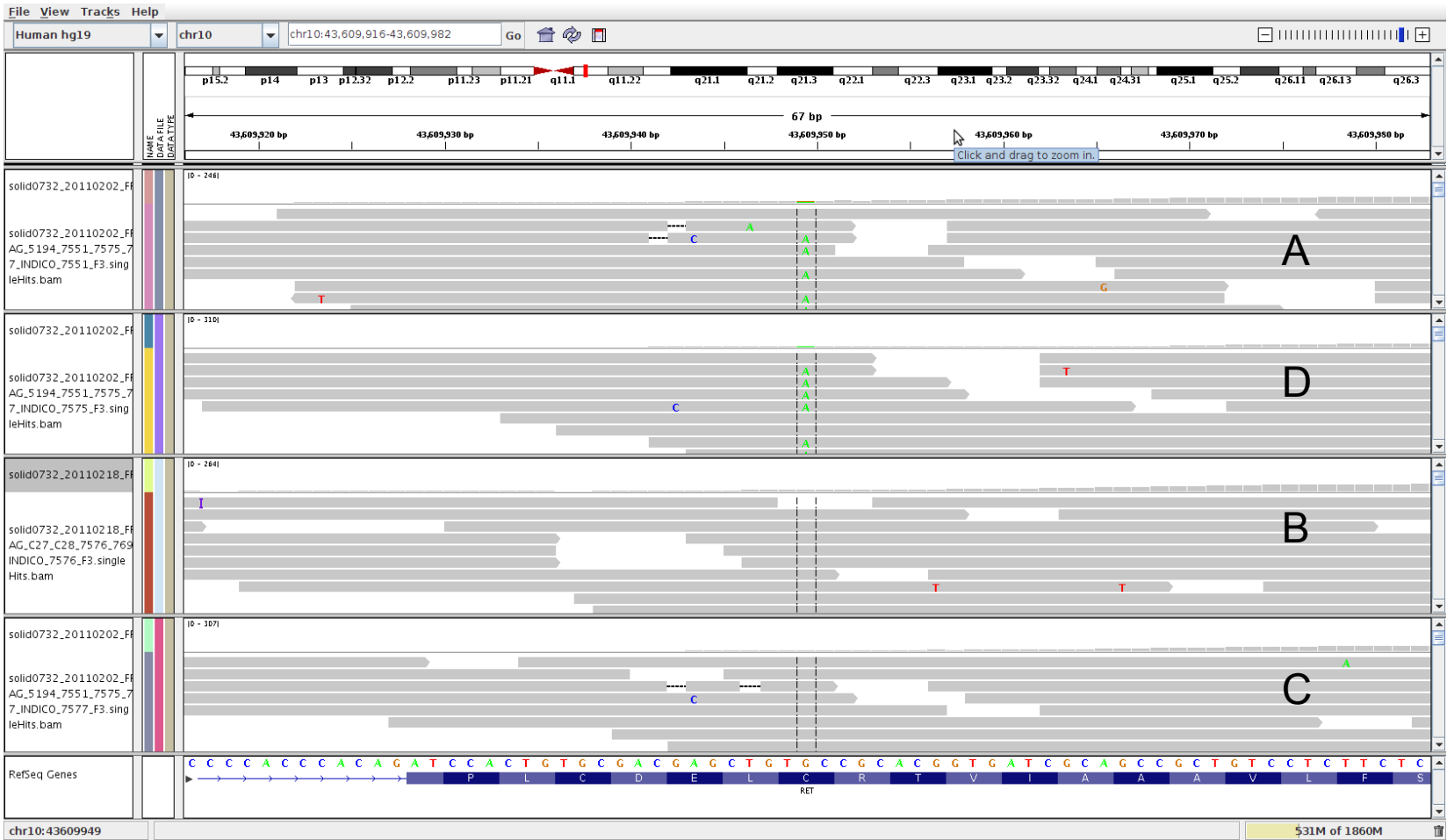
MGP Medical Genome Project

An example with MTC

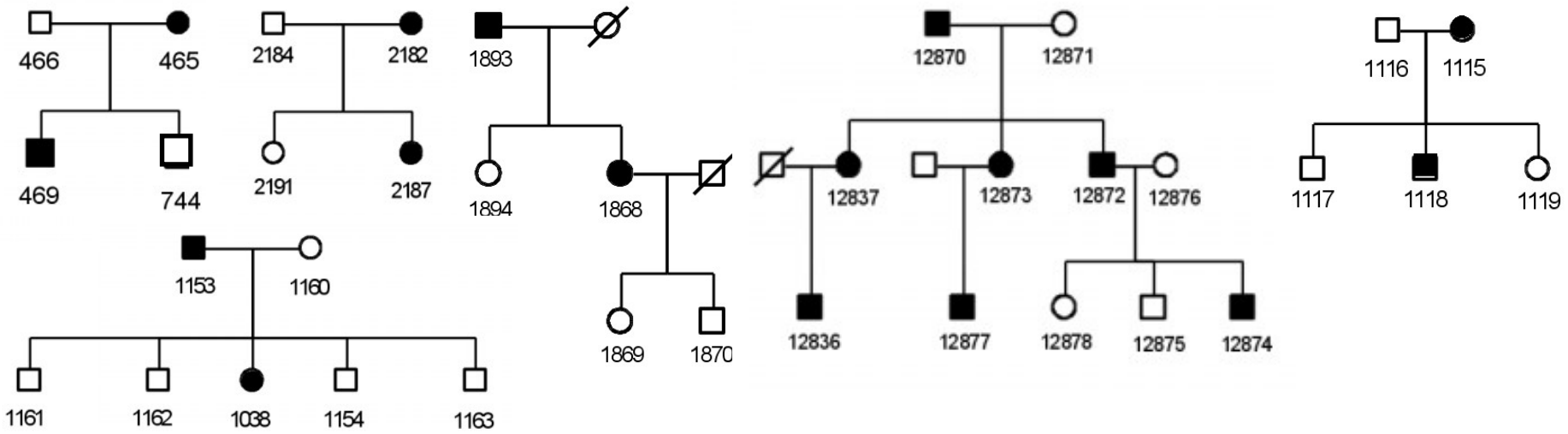


Dominant:
Heterozygotic in A and D
Homozygotic reference allele in B and C
Homozygotic reference allele in controls

The codon 634 mutation



Finding disease genes



	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

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

Real coverage and some figures

Sequencing

Enrichment +Sequence run: ~2 weeks
500,000,000 sequences
25,000,000,000 bases
Short 50bp (SOLiD) sequences

File size per exome: 300 GB

Coverage

SeqCap EZ Human Exome Library 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:

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

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

SOLiD 4 exome statistics

Samples	Total Reads	Maped Reads	Mean Coverage	nt with Cov >=5	nt with Cov >=10	nt with Cov >=20	nt with Cov >=40
3358	75571883	56,173,120 (74.33%)	50.08x	94,07%	88,84%	80,88%	68,60%
3361	56619558	43,472,037 (76.78%)	38.26x	92,82%	85,82%	75,26%	59,89%
3406	57939709	44,519,498 (76.84%)	38.80x	93,67%	87,43%	77,45%	62,08%
3411	63498579	44,277,476 (69.73%)	39.47x	93,69%	87,12%	76,66%	61,28%
4217	96865353	69,694,453 (71.95%)	60.06x	93,96%	89,09%	82,12%	71,81%
4218	101472493	69,820,897 (68.81%)	60.60x	94,66%	90,05%	83,14%	72,62%
4219	100752849	70,563,596 (70.04%)	60.95x	94,83%	90,19%	83,22%	72,71%
5296	102205294	67,274,569 (65.82%)	59.40x	94,66%	89,78%	82,40%	71,44%
5298	105902896	71,944,512 (67.93%)	65.29x	93,74%	88,46%	81,04%	70,64%
5299	102501588	73,434,877 (71.64%)	64.12x	94,34%	89,86%	83,34%	73,56%
4236	109184670	74,065,922 (67.84%)	64.56x	95,18%	91,08%	84,80%	75,04%
4239	112120064	68,945,106 (61.49%)	58.19x	94,77%	89,83%	82,24%	71,01%
6504	110840136	84,762,604 (76.47%)	74.29x	94,38%	89,96%	83,94%	75,23%
6528	103305803	75,286,220 (72.88%)	56.55x	94,53%	88,95%	80,68%	68,64%
4027	115809124	88,168,970 (76.13%)	81.93x	95,29%	91,39%	85,76%	77,40%
4026	110460098	81,873,428 (74.12%)	70.89x	94,75%	90,12%	83,40%	73,69%
4240	112919072	55,238,102 (48.92%)	42.49x	92,76%	85,74%	75,31%	60,41%
4255	106567735	75,712,011 (71.05%)	68.73x	94,97%	90,78%	84,49%	75,00%
4257	107735341	76,750,847 (71.24%)	63.68x	95,30%	91,03%	84,45%	74,32%
4258	113964751	77,335,536 (67.86%)	66.20x	95,01%	90,44%	83,55%	73,49%

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

And this is what we get from the variant calling pipeline

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

none	1	2116429	C	missense	0	PRKCZ,LOC1
none	1	2116429	C	missense	0	PRKCZ,LOC1
none	1	2116429	C	missense	0	PRKCZ,LOC1
none	1	2116429	C	utr-3	0	PRKCZ,LOC1
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	KIAA0562
none	1	3746432	G	missense	0	KIAA0562
dbSNP_1000Genomes	1	3755675	T	coding-synonymous	1891941	KIAA0562
none	1	6029181	G	missense	0	NPHP4
none	1	6101899	A	missense	0	KCNAB2
none	1	6101899	A	intron	0	KCNAB2
none	1	6132842	C	coding-synonymous	0	KCNAB2
none	1	6132842	C	coding-synonymous	0	KCNAB2
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6535559	T	missense	0	PLEKHG5
none	1	6647590	A	missense	0	ZBTB48
none	1	6694129	T	missense	0	THAP3
none	1	6695719	T	utr-3	0	DNAJC11
none	1	6704720	C	missense	0	DNAJC11
none	1	6711636	C	coding-synonymous	0	DNAJC11
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

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 47% lower intracellular digo...
1	A study in 336 recipients of hematopoietic–cell transplants

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

Exome sequencing holds much more potential information

More than 30 millions of bases are interrogated with the exome capture systems and typically only one of these bases is useful. The rest of the information is overlooked in many places but….

… we can re-analyse it looking to other functional elements:

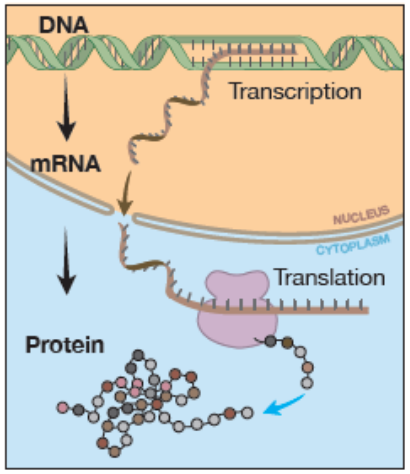
- An enormous amount of data can be used for different population, evolutionary, functional and many other types of studies.

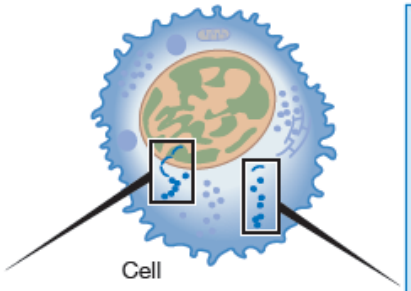
miRNAs

- They are included in targeted exome re-sequencing designs

RNA interference

1. The central dogma

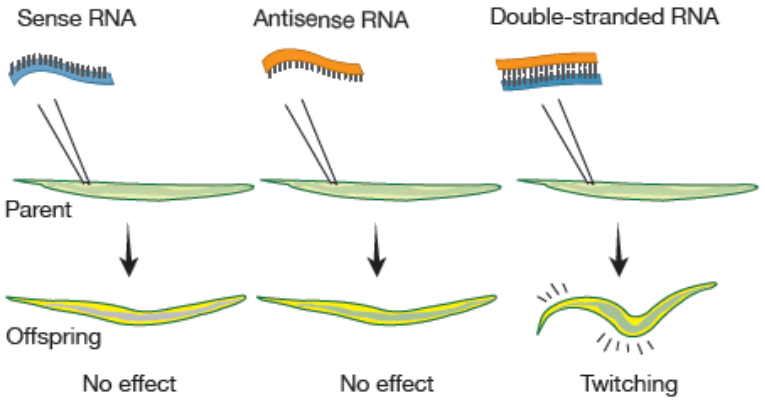




Our genome operates by sending information from double-stranded DNA in the nucleus, via single-stranded mRNA, to guide the synthesis of proteins in the cytoplasm.

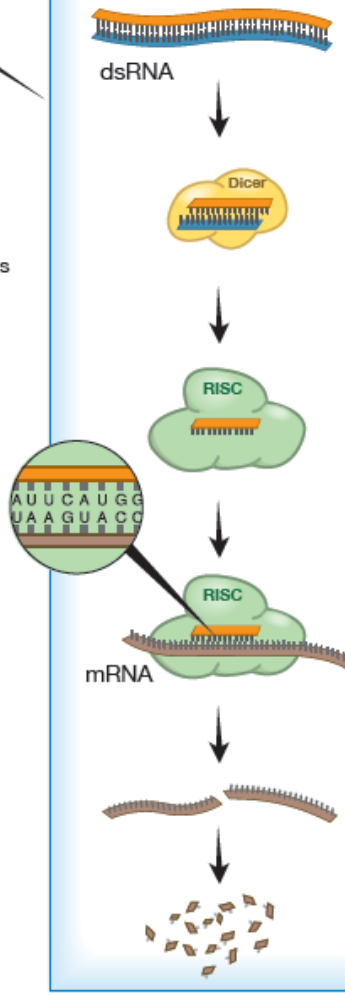
2. The experiment

RNA carrying the code for a muscle protein is injected into the worm *C. elegans*. Single-stranded RNA has no effect. But when double-stranded RNA is injected, the worm starts twitching in a similar way to worms carrying a defective gene for the muscle protein.



3. The RNAi mechanism

RNA interference (RNAi) is an important biological mechanism in the regulation of gene expression.



Double-stranded RNA (dsRNA) binds to the protein Dicer …

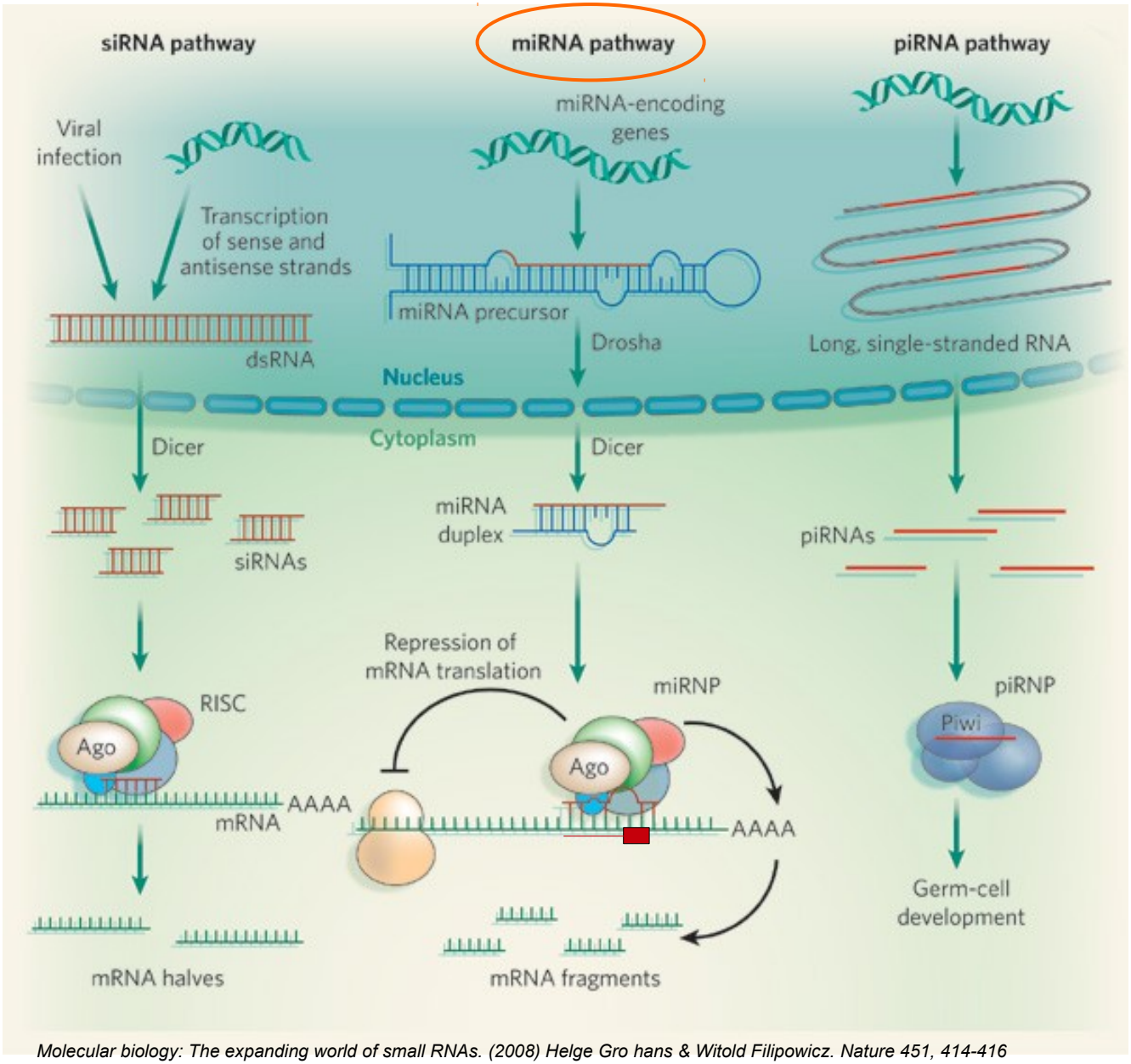
… which cleaves dsRNA into smaller fragments.

One of the RNA strands is loaded into a RISC complex…

…and links the complex to the mRNA strand by basepairing.

mRNA is cleaved and destroyed. No protein can be synthesized.

RNA interference types



Molecular biology: The expanding world of small RNAs. (2008) Helge Gro hans & Witold Filipowicz. Nature 451, 414-416

miRNAi - Features

- Bind mRNA 3'UTR
- Seed region (2-8 nt) complementarity, incomplete complementarity the remaining
- Promote repression of protein translation
- 21-24 nt

miRNAi - Functions in the cell

- Development regulation
- Tissue identity
- Neuronal plasticity...

Previous ideas on miRNAs

- Many studies on miRNA expression levels

- miRNAs have been related to different diseases

- miRNAs are thought to be highly conserved regions of the genome

Variations in the protein coding genes in the human genome

Variants predicted to severely affect the function of human protein coding genes known as loss-of-function (LOF) variants were thought:

- To have a potential deleterious effect
- To be associated to severe mendelian disease

The 1000 genomes project has revealed an enormous amount of variation a the genome level, much higher than expected

An unexpectedly large number of LOF variants have been found in the genomes of apparently healthy individuals

- A conservative estimation suggests 100 LOF variants per genome including more than 30 in a homozygous state

Previously unnoticed level of variation with putative functional consequences in protein coding regions in humans.

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

miRNAs are thought to be highly conserved regions of the genome

- Some studies suggest that most of the mutations are expected to have adverse effects in the functionality or biogenesis of miRNAs

- Early studies on variability using SNPs reported a very low level of varitation
  - Absence of polymorphims in more than 90% of human pre-miRNAs and most of them were no in the seed region
  - Strong selective constraint

- To date dbSNP reports 519 variants in miRNAs

miRNA exome data analysis and study of miRNA variability

Preliminary study for sequencing data corresponding to 23 exomes from the Medical Genome project

The aim was to survey the actual level of variability at these genomic elements an to check if:

- A restrictive scenario for miRNA variants was confirmed

  or

- A situation similar to the occurrence of LOF variants in protein coding genes existed for the miRNAs

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

Pilot study of variability in miRNAs in control population

The analysis of the information available on 23 exomes from healthy southern Spain population has uncovered an unexpected amount of variability in microRNAs.

558 variants in total were found in 291 different miRNAs

- 131 of these miRNAs are known to be involved in almost 200 diseases according the human miRNA Disease database.

487 of the variants (87%) were described for the first time in this study.

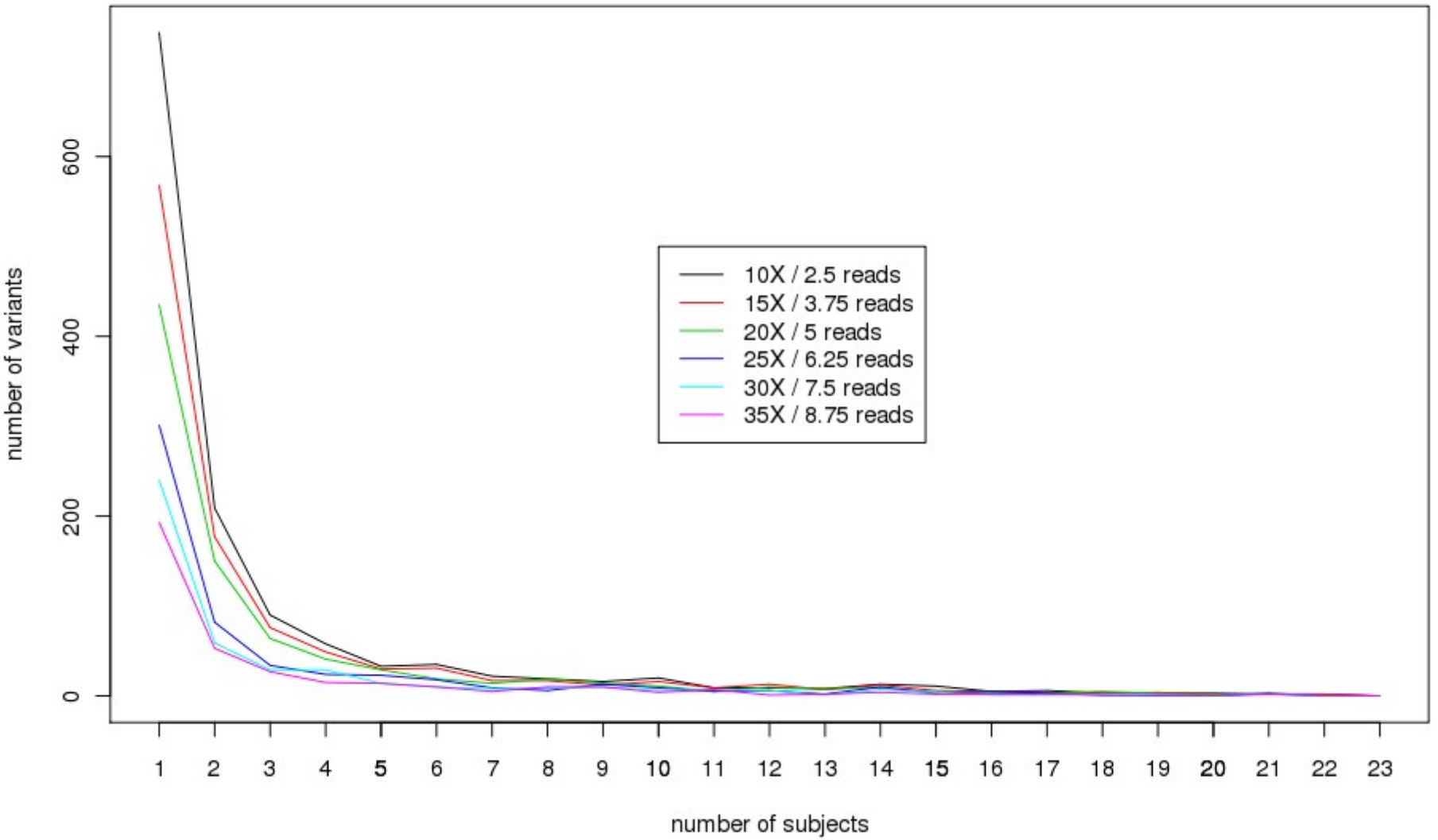
- This figure almost doubles the number of known variants (519) in miRNAs and constitutes a remarkably high ratio of discovery.

The average number of SNVs per individual affecting to miRNAs was 118

miRNA variant calling in exomas

- Exome capture system used contains 720 miRNAs
- Only high quality sequence reads with unique mapping positions to the reference human genome were used for calling variants.
- Average coverage observed in these regions was 40x and a minimum of 25x is needed to call a variant
- 30% of the reads have to contain the change to call the variation

Number of individuals supporting new variants found



JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

Variants distribution among miRNAs

Distribution of variants among the different miRNAs is not uniform.

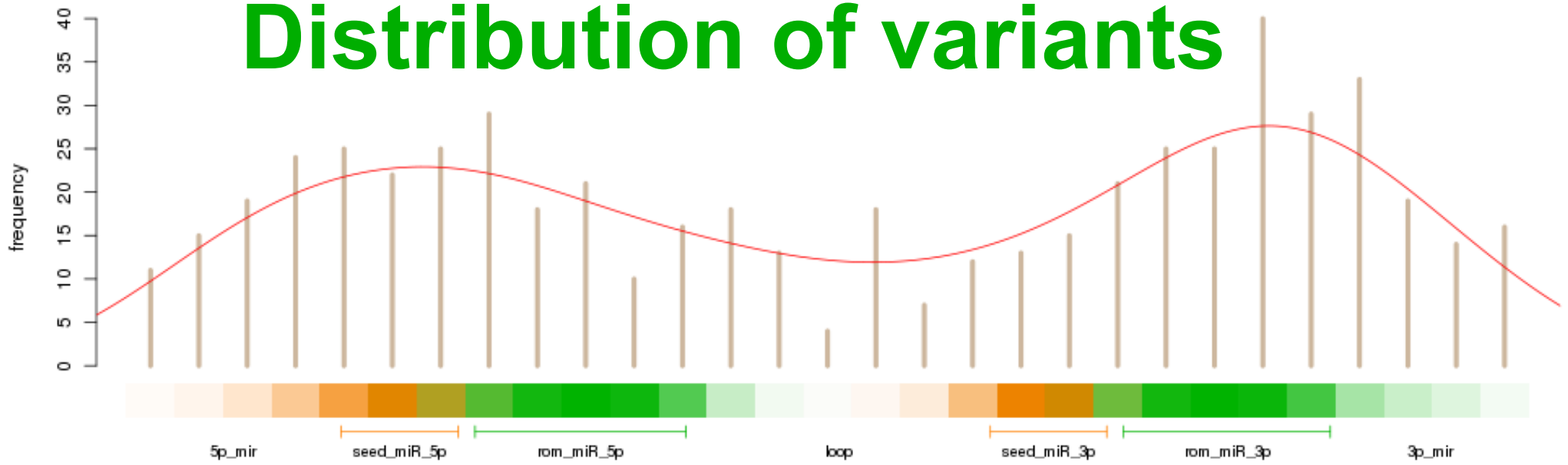
Most of them were affected by only one SNV

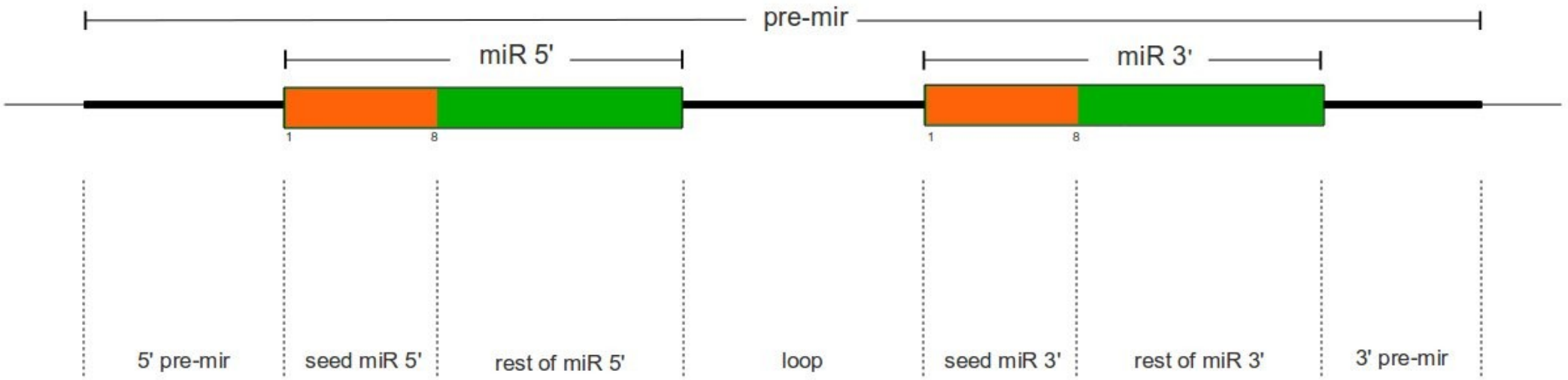
A few mature miRNAs have more variations

- One miRNA contains 6 variant positions in the mature miRNA and 4 variant positions have been found in another 3 miRNAs.

Variants were not homogeneously distributed across the structure of the pre-miRNA.

Distribution of variants





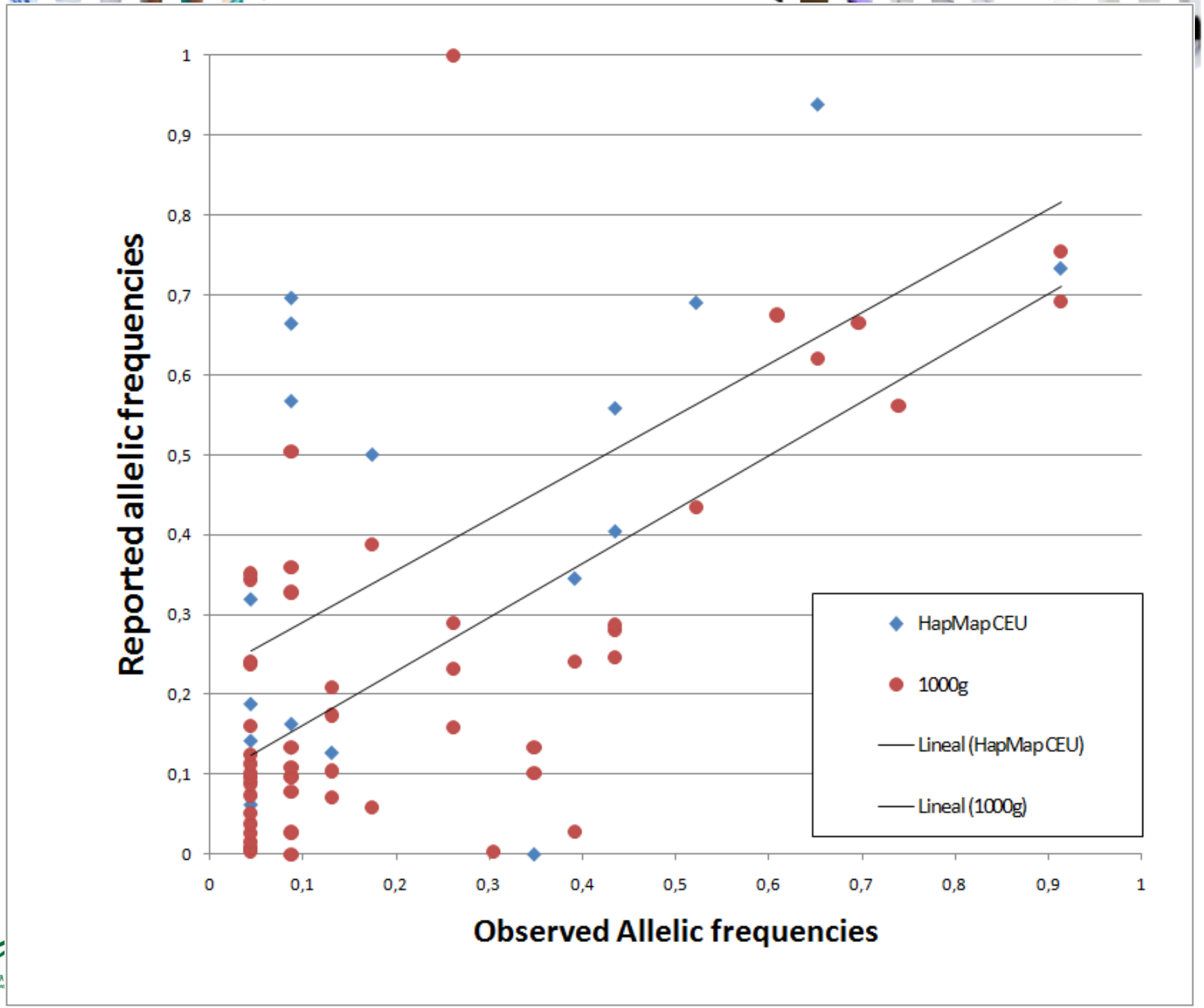
JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

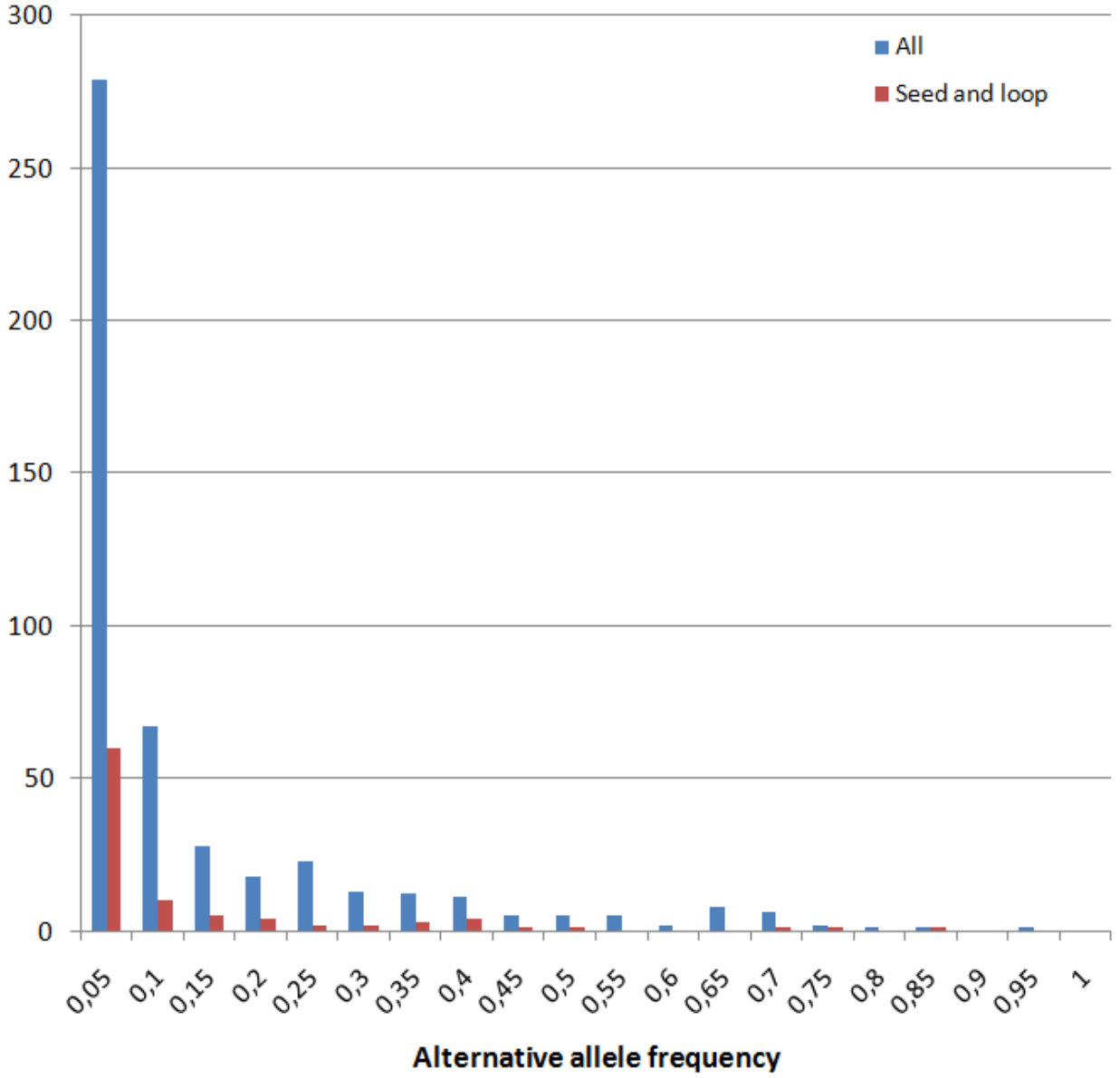
Population frequencies

- Conclusions based on frequencies estimated from a population with this small sample size must be taken with caution
  - However when the allelic frequencies in the studied sample are compared to the corresponding ones reported both in dbSNP or the 1000 genomes project. There is a correlation among them

- Many of the newly discovered variants occurred in only one individual and consequently they are at no very high frequency
  - However there are still a considerable number of variants that appear at higher frequencies.


Reported allelic frequencies
Observed Allelic frequencies
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
HapMap CEU
1000g
Lineal (HapMap CEU)
Lineal (1000g)

Alternative allele frequencies



JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

Pathogenic effect of variants

Most of the variants found in the studied samples (98.7%) were in heterozygosis.

A small number of them displayed the alternative allele in homozygosis

- None of them seem to be a rare variant according to the estimated population frequencies
- Deviations of the Hardy-Weinberg equilibrium (HWE) show that this variants are under negative selection (pathogenic). This fact would also suggest that the role of miRNAs in disease could be bigger that previously suspected.

Disease-miRNA associations

Disease	# miRNAs	pre-mir ids
Heart Failure	6	hsa-mir-10b,hsa-mir-204,hsa-mir-296,hsa-mir-300,hsa-mir-340,hsa-mir-381
Breast Neoplasms	4	hsa-mir-10b,hsa-mir-204,hsa-mir-296,hsa-mir-340
Adenocarcinoma	3	hsa-mir-106a,hsa-mir-10b,hsa-mir-204
Melanoma	3	hsa-mir-216a,hsa-mir-217,hsa-mir-296
Neoplasms	3	hsa-mir-106a,hsa-mir-10b,hsa-mir-204
Pancreatic Neoplasms	3	hsa-mir-106a,hsa-mir-204,hsa-mir-217
Adenoviridae Infections	2	hsa-mir-1274b,hsa-mir-627
Autistic Disorder	2	hsa-mir-106a,hsa-mir-381
Carcinoma, Squamous Cell	2	hsa-mir-10b,hsa-mir-296
Hepatitis C	2	hsa-mir-296,hsa-mir-448
Leukemia, Lymphocytic, Chronic, B-Cell	2	hsa-mir-16-2,hsa-mir-640
Lung Neoplasms	2	hsa-mir-106a,hsa-mir-216a
Lupus Vulgaris	2	hsa-mir-296,hsa-mir-557
Prostatic Neoplasms	2	hsa-mir-106a,hsa-mir-296
Stomach Neoplasms	2	hsa-mir-106a,hsa-mir-340
Astrocytoma	1	hsa-mir-106a
Atherosclerosis	1	hsa-mir-296
Carcinoma	1	hsa-mir-10b
Carcinoma, Hepatocellular	1	hsa-mir-106a
Carcinoma, Non-Small-Cell Lung	1	hsa-mir-16-2
Colonic Neoplasms	1	hsa-mir-106a
Colorectal Neoplasms	1	hsa-mir-492
Endometrial Neoplasms	1	hsa-mir-204
Glioma	1	hsa-mir-106a
Head and Neck Neoplasms	1	hsa-mir-204
Hepatoblastoma	1	hsa-mir-492
Hypertension	1	hsa-mir-204
Liver Neoplasms	1	hsa-mir-10b
Medulloblastoma	1	hsa-mir-10b
Mesothelioma	1	hsa-mir-106a
Muscular Disorders, Atrophic	1	hsa-mir-381
Myocardial Infarction	1	hsa-mir-10b
Nasopharyngeal Neoplasms	1	hsa-mir-10b
Neuroblastoma	1	hsa-mir-10b
Ovarian Neoplasms	1	hsa-mir-296
Patau Syndrome	1	hsa-mir-16-2
Retinal Degeneration	1	hsa-mir-204
Retinal Neovascularization	1	hsa-mir-106a
Toxoplasmosis	1	hsa-mir-106a
Urinary Bladder Neoplasms	1	hsa-mir-300

JUNTA DE ANDALUCIA
CONSEJERIA DE SALUD

MGP Medical Genome Project

CASEGH

Genomics

Dr. Rosario Fernández Godino

Dr. Alicia Vela Boza

Dr. Sandra Pérez Buira

María Sánchez León

Javier Escalante Martín

Ana Isabel López Pérez

Beatriz Fuente Bermúdez

Bioinformatics

Daniel Navarro Gómez

Pablo Arce García

Juan Miguel Cruz

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

Dr. Javier Santoyo (CASEGH's Scientific Coordinator)

José Carbonell

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

The INB, National Institute of Bioinformatics (Functional Genomics Node)

INB

The CIBERER Network of Centers for Rare Diseases

ciberer

MINISTERIO DE CIENCIA E INNOVACIÓN
F.E.D.E.R.
ProyectoDeméter
Rticc
JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD
MGP Medical Genome Project
iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
AFFILIATED REGIONAL INSTITUTE