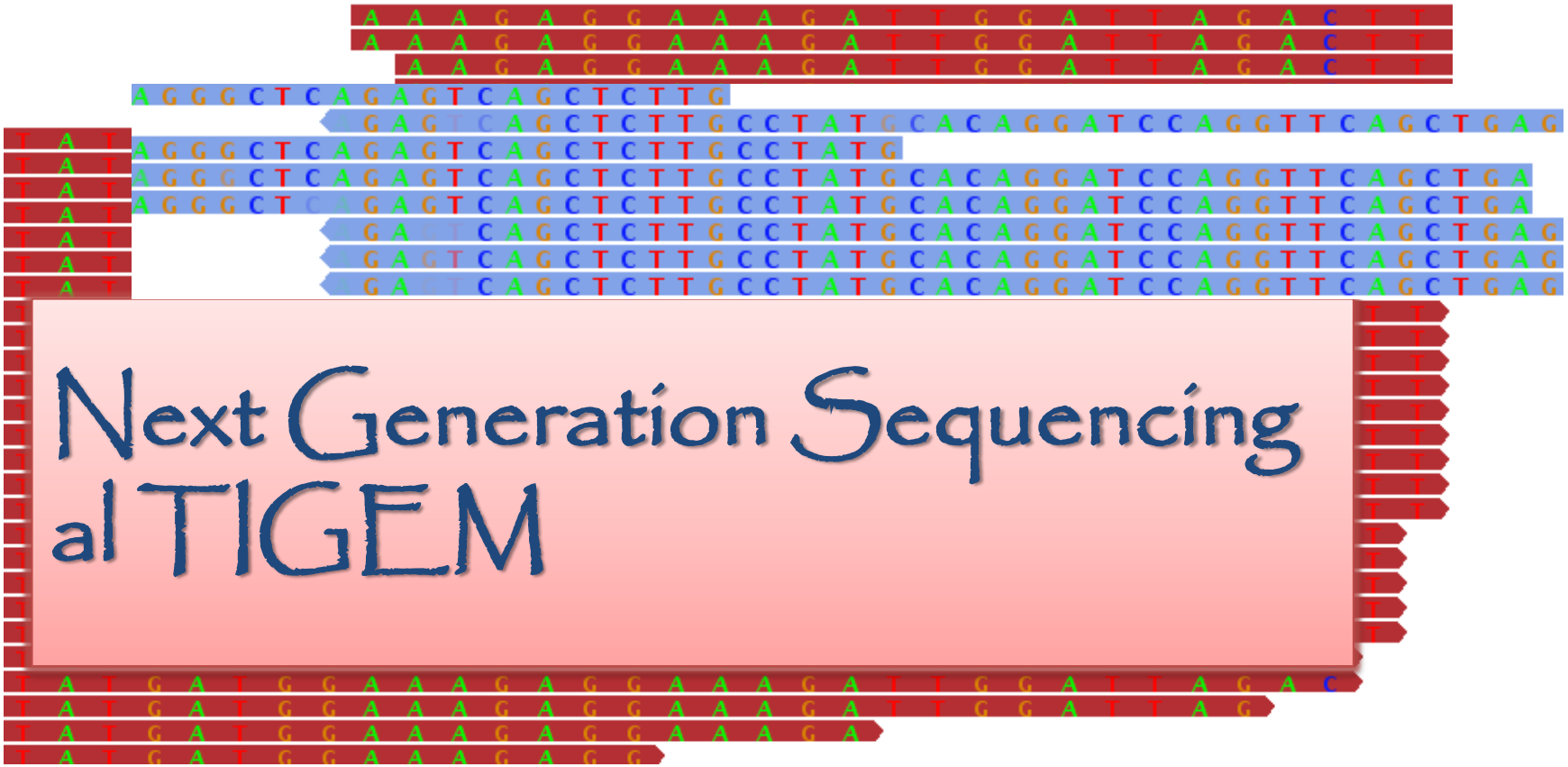


Next Generation Sequencing al TIGEM

Margherita Mutarelli

mutarelli@tigem.it

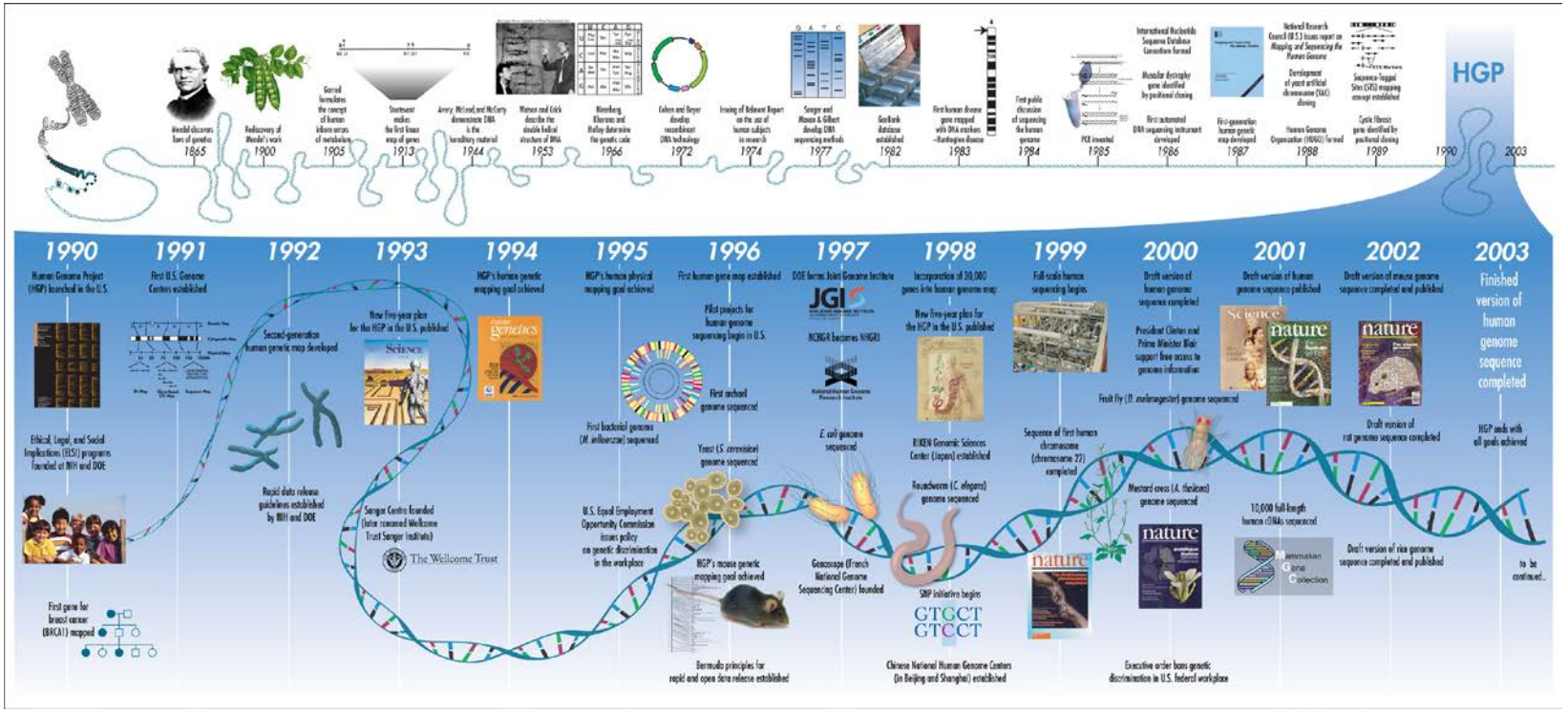
Bioinformatics Core

Next Generation Sequencing

- NGS technologies have lowered the cost of high-throughput sequencing
- Experiments once only possible for large genomic centers / consortium became feasible for the single group
- Several hundreds million sequence reads in a week at the cost of few thousand $$

igem
TELETHON INSTITUTE OF GENETICS AND MEDICINE

The Human Genome Project timeline



- Human genome sequence completion by Sanger sequencing took **13 years**, involving More than **2,000** scientists from over **20** institutes in **six** countries at the cost of **2.7 billion dollars**

The cost of an individual genome sequence

J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2[†]	250[§]	7.4	Aligned*	95[‖]	3.32 (BLAT)	234	1,000,000[¶]
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410[†]	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000[¶]
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921[†]	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000[¶]
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393[†]	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000[¶]
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647[†]	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000[¶,#]
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211[†]	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000[¶,**]
		91% MP: 600–3,500 bp	2,075[†]	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725[†]	32[§]	28	Aligned*	90	2.81 (IndexDP)	4	48,000[¶]
AML female	Illumina/ Solexa	Frag: 150–200 bp[††]	2,730[‡,††]	32	32.7	Aligned*	91	3.81[††] (MAQ)	98	1,600,000[‖]
		Frag: 150–200 bp[§§]	1,081[‡,§§]	35	13.9		83	2.92[§§] (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp[††]	1,620[‡,††]	35	23.3	Aligned*	98.5	3.46[††] (MAQ)	16.5	500,000[‖]
		MP: 200–250 bp[§§]	1,351[‡,§§]	50	21.3		97.4	3.45[§§] (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238[†]	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000[¶,‡‡]
		84% MP: 600–3,500 bp	1,211[†]	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. [†]Mappable reads for aligned assemblies. [‡]Average read-length. [‖]D. Wheeler, personal communication. [¶]Reagent cost only. [#]S.-M. Ahn, personal communication. [**]K. McKernan, personal communication. [††]Tumour sample. [§§]Normal sample. [‖‖]Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost. E. Mardis, personal communication. [‡‡]R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

Metzker, Nat Rev Genet (2010) 11:31

NGS Applications

Targeted sequencing	• Effective enrichment of selected parts of the human genome at high coverage
Whole exome NGS	• Sequencing of the whole exome or the all the exons of the X chromosome
small RNA-seq	• Identification and quantification of miRNA in tissues
RNA-seq	• 100 x 2 Identification of novel mRNA species and digital quantification
ChIP-seq	• Sequencing of chromatin immunoprecipitate
Mate Pair Sequencing	• identification of structural variants (i.e. due to abnormal insertions)

DNA resequencing at different scales

10-50 disease genes
0.1Mb
coverage @100x
(0.01Gb/sample)

RainDance Technologies

preferential exome 1-3Mb
coverage @200x
(1Gb/sample)

HaloPlex

whole exome 60Mb
coverage @60 x
(10 Gb/sample)

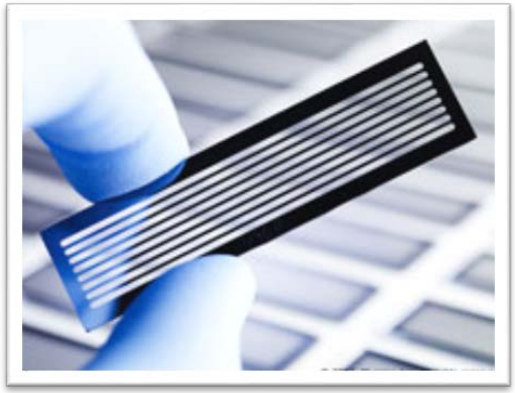
SureSelect

whole genome 3Gb
coverage @40 x
(150 Gb /sample)

igem
TELETHON INSTITUTE OF GENETICS AND MEDICINE

TIGEM NGS Facility set-up

Illumina HiSeq 1000



10 days per run

2x100nt Paired-end

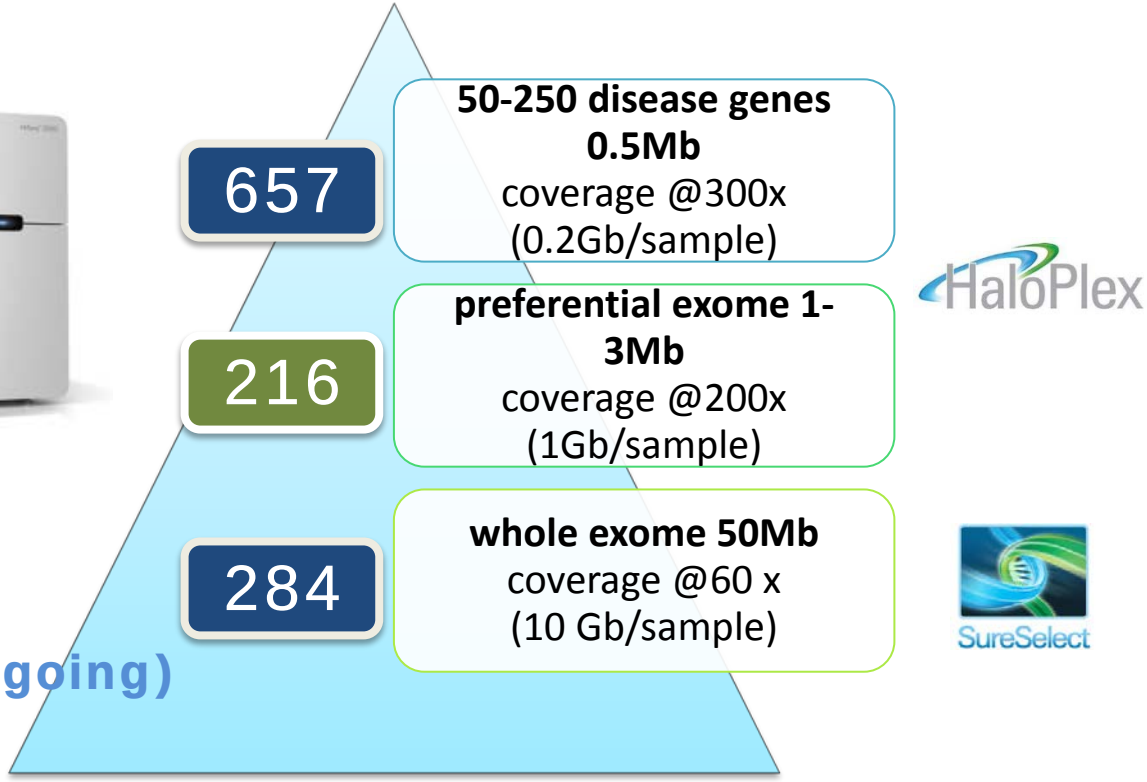
Max 300Gb sequence

The Illumina HiSeq was purchased by the Ministry of Research Telethon Foundation supports a technician salary

Consumables are from individual P.I. grants (ERC, Universities, Telethon, private, etc. ..)

DNA resequencing

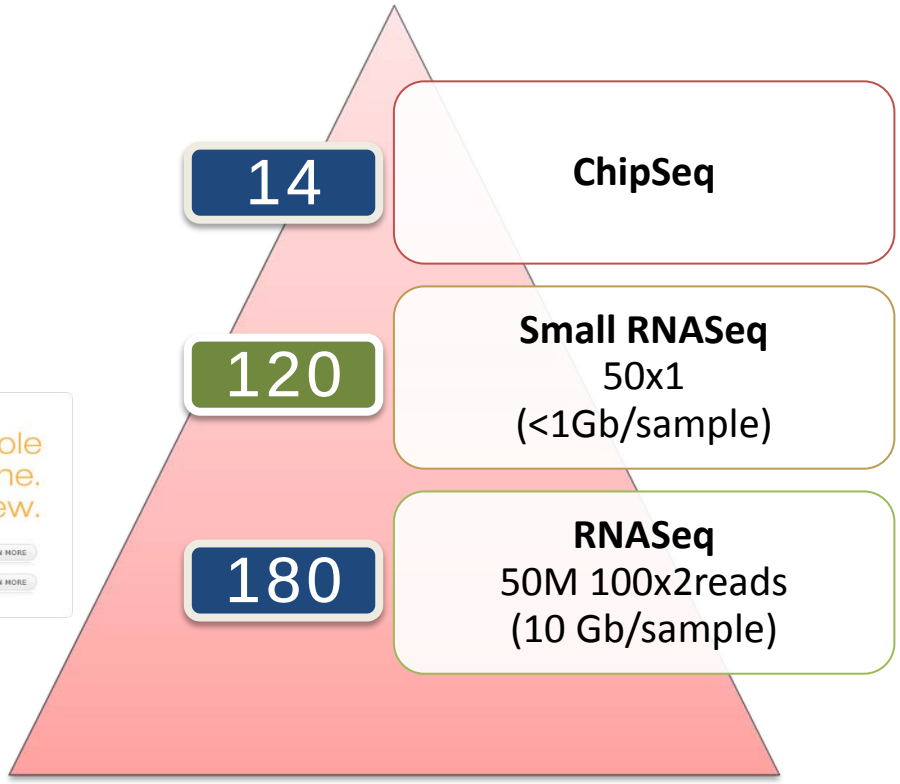
First run:
Febr. 2012
Last run:
June 2014 (ongoing)



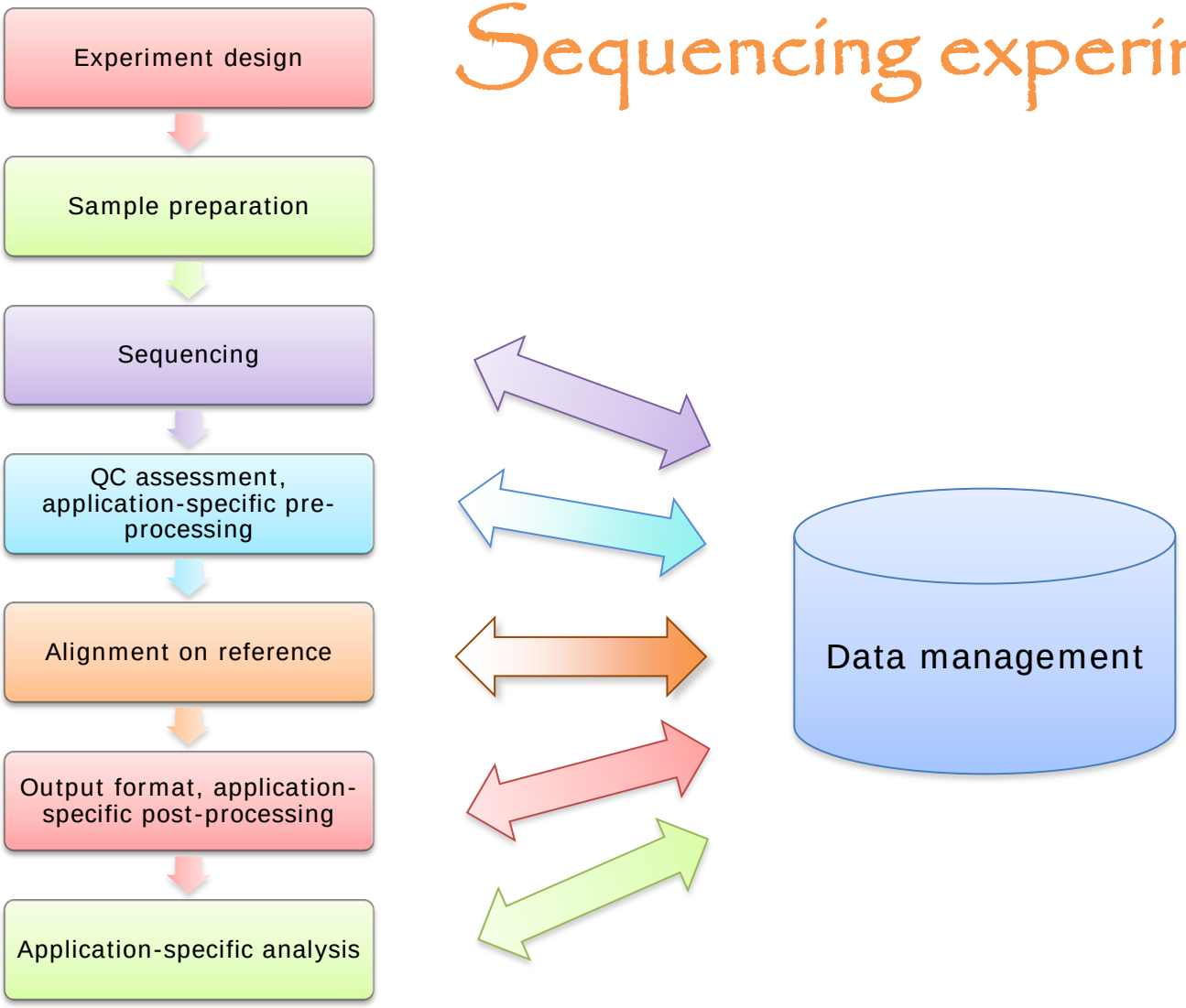
HaloPlex

SureSelect

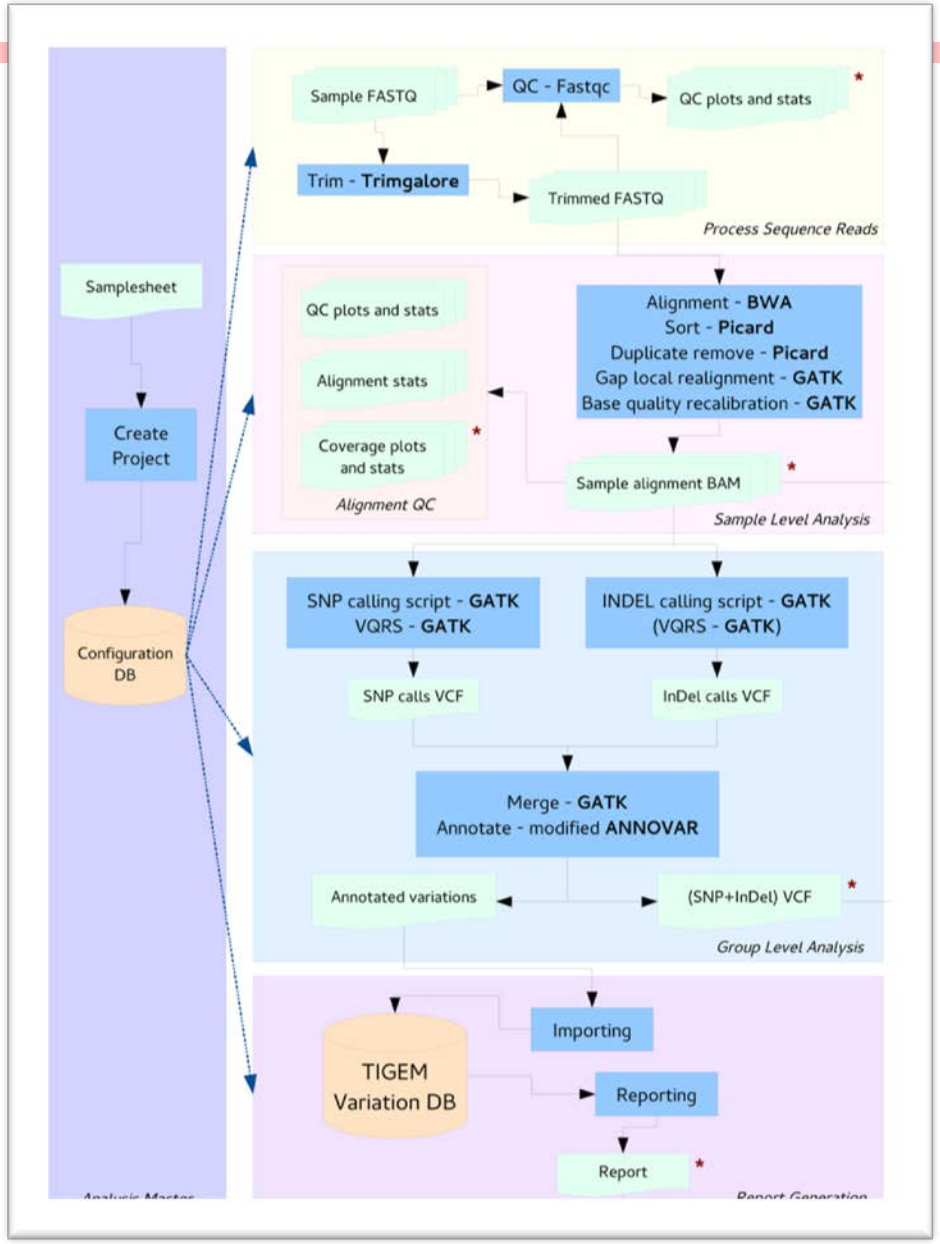
other applications



Sequencing experiment



eXSP pipeline

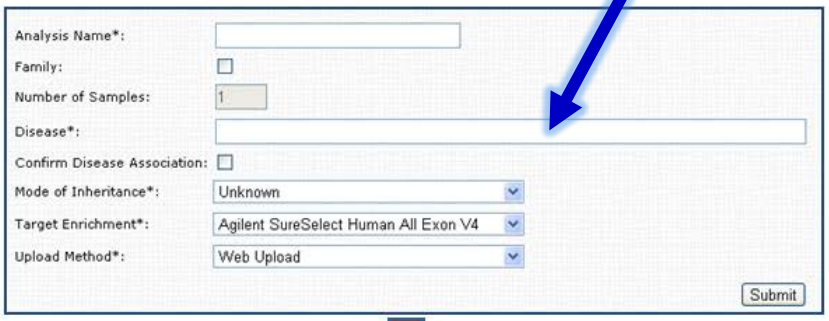


Mutarelli (2014) BMC Genomics

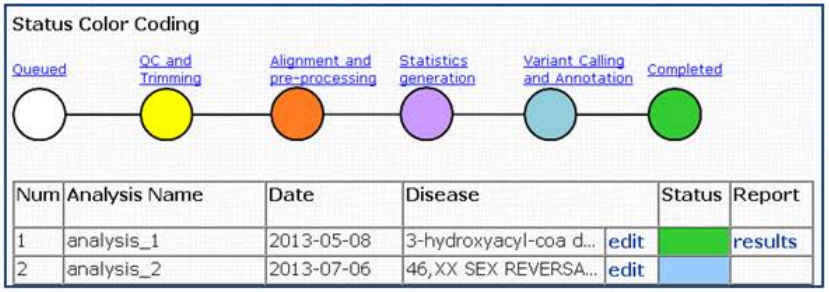
Exome Mendelian Disorder Workflow

MEDIC disease vocabulary

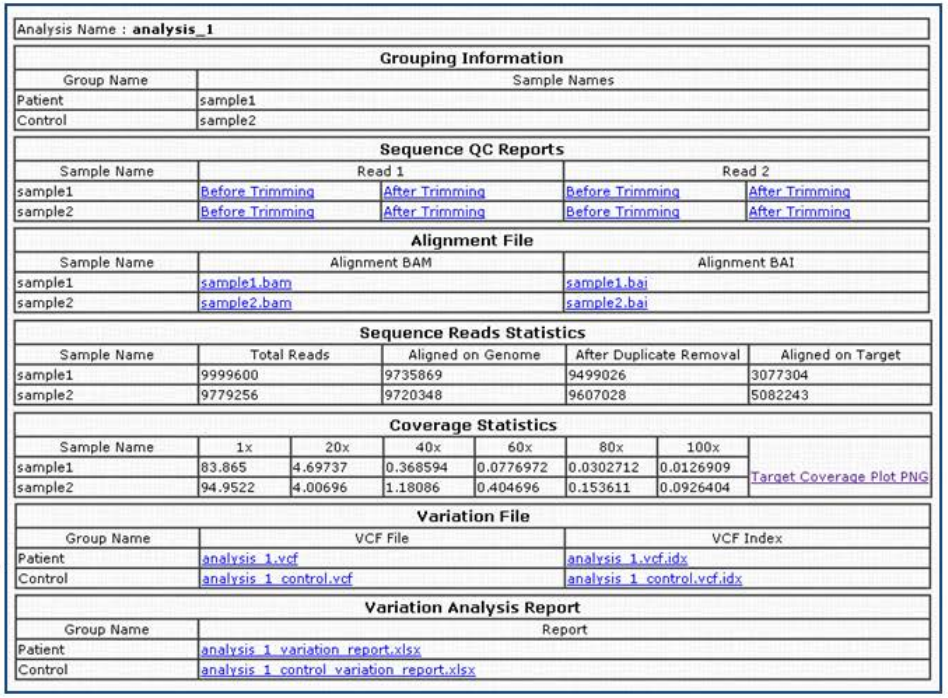
CREATE NEW ANALYSIS



ANALYSIS STATUS AND REPORTS



RESULTS



Analysis Name : **analysis_1**

Grouping Information

Group Name	Sample Names
Patient	sample1
Control	sample2

Sequence QC Reports

Sample Name	Read 1		Read 2	
sample1	Before Trimming	After Trimming	Before Trimming	After Trimming
sample2	Before Trimming	After Trimming	Before Trimming	After Trimming

Alignment File

Sample Name	Alignment BAM	Alignment BAI
sample1	sample1.bam	sample1.bai
sample2	sample2.bam	sample2.bai

Sequence Reads Statistics

Sample Name	Total Reads	Aligned on Genome	After Duplicate Removal	Aligned on Target
sample1	9999600	9735869	9499026	3077304
sample2	9779256	9720348	9607028	5082243

Coverage Statistics

Sample Name	1x	20x	40x	60x	80x	100x	
sample1	83.865	4.69737	0.368594	0.0776972	0.0302712	0.0126909	Target Coverage Plot PNG
sample2	94.9522	4.00696	1.18086	0.404696	0.153611	0.0926404	

Variation File

Group Name	VCF File	VCF Index
Patient	analysis_1.vcf	analysis_1.vcf.idx
Control	analysis_1_control.vcf	analysis_1_control.vcf.idx

Variation Analysis Report

Group Name	Report
Patient	analysis_1_variation_report.xlsx
Control	analysis_1_control_variation_report.xlsx

http://exome.tigem.it

Exome Mendelian Disorder Report

Variation Class			
Class	Frequency	Quality	Impact
I	+	+	+
II	+	+	-
III	+	-	+
IV	+	-	-
V	-	+/-	+/-

Gene Related Annotation

Symbol	Exonic Func	Gene Func	Aachange
TNPO3	stoploss SNV	exonic	NM_001191028:c.2579delA:p.X860C

General Variation Information

chrom	pos	ref	alt	type	qual	filter	vq	lod	Sample 1 Variation Class
chr7	128597309	CT	C	del	947.98	PASS			I

Genotype Information For Sample

Sample1 ZYG	Sample1 GT	Sample1 N REF reads	Sample1 N ALT reads	Sample1 N tot reads	Sample1 Perc ALT reads	Sample1 GQ
HET	0/1	26	27	52	51.9	99

Functional Predictions For Protein Damage

Avsift	LJB SIFT	LJB SIFT Pred	LJB PolyPhen2	LJB PolyPhen2 Pred	LJB LRT	LJB LRT Pred	LJB Mutation Taster	LJB Mutation Taster Pred

Functional Predictions For Phylogenetic Conservation

LJB PhyloP	LJB PhyloP Pred	LJB Gerp++	Conserved
			654;Name=lod=609

Variation Frequency From Exome Sequencing Project (ESP) & 1000 Genome Project

freq ESP6500	freq1000g 2012apr ALL	freq1000g 2012apr AFR	freq1000g 2012apr AMR	freq1000g 2012apr ASN	freq1000g 2012apr EUR

dbSNPAnnotation

dbsnp137	dbSNP137 NonFlagged	dbSNP137 Observed Allele	OMIM

Disease Group Allele Frequency From TIGEM Variant DB

freq Disease ID:X	freq Disease ID:X Controls	freq Disease ID:1	freq Disease ID ..	freq Disease ID:N

Next Generation Sequencing core

NGS Core

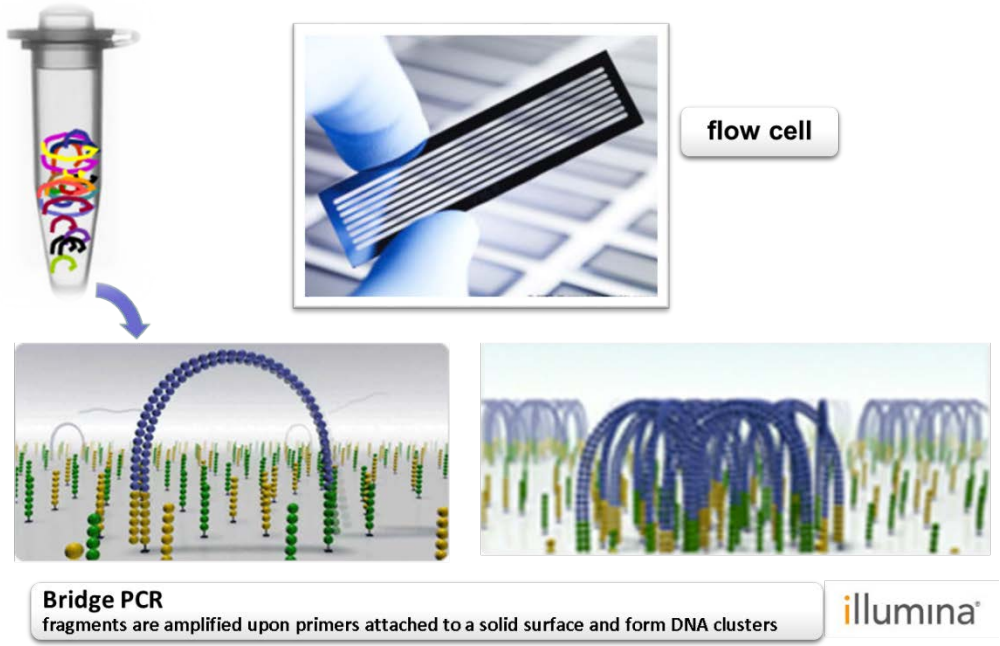
P.I.: Vincenzo Nigro

- Annalaura Torella
- Manuela Dionisi
- Marco Savarese
- Giuseppina di Fruscio

Bioinformatics Core

P.I.: Diego di Bernardo

- Margherita Mutarelli
- Veer Singh Marwah
- Diego Carrella



Unione Italiana Lotta alla Distrofia Muscolare