Analysing genomes and transcriptomes using Illumina sequencing

Dr. Heinz Himmelbauer
Centre for Genomic Regulation (CRG)
Ultrasequencing Unit
Barcelona

The Sequencing Revolution

High-Throughput Sequencing
2000

Image credit: U.S. Department of Energy, JGI

96 sequences per hour

High-Throughput Sequencing
2009

2.6 million sequences per hour

Next generation sequencing platforms

- Applications
  - mRNA-Seq: expression profiling, transcript discovery
  - Small RNAs (miRNAs)
  - ChIP-Seq, RIP-Seq
  - DNA methylation
  - Re-sequencing, haplotyping (SNPs and structural variation)
  - *De novo* sequencing

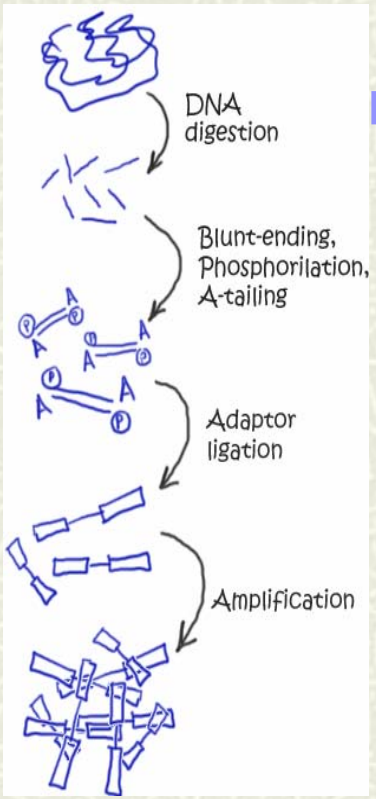
Sequencing platforms

Technology	Year	Read length (nt)	Chemistry	Template amplification
Sanger	1977	1000	SBS	cloning/PCR
454	2005	250-500	Pyro	PCR
Solexa	2006	36-75	SBS	PCR
SOLiD	2007	35	Ligation	PCR
Helicos	2008	30	SBS	no amplification
PacBio	2010 ?	1000	SBS	no amplification
VisiGen	2009 ?	1000	SBS	no amplification
Oxford Nanopore	?	?	Elco	no amplification

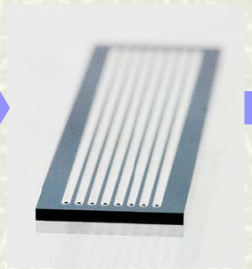
SBS = Sequencing by Synthesis
Pyro = Pyrosequencing
Ligation = Sequencing by ligation assays
Elco = Electrochemical detection

Illumina/Solexa technology

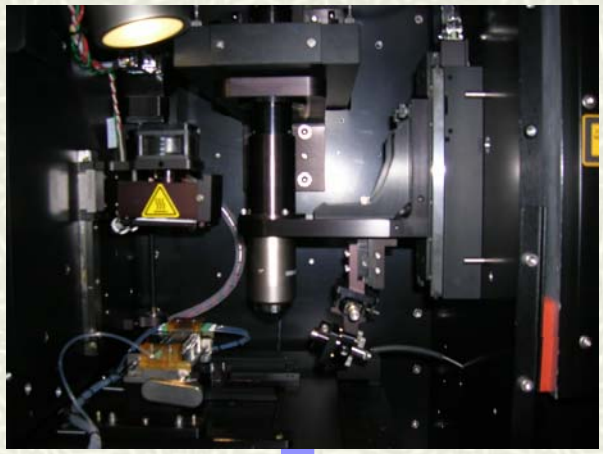
Library preparation



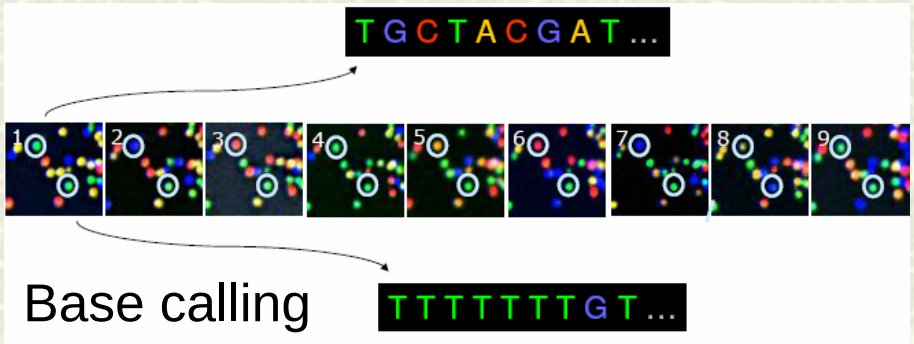
Cluster generation



Cycle sequencing



Solexa GA II run:
- 45 hours
- 1 TB image data
- 10 x 8 mio. 36-50 nt
- up to 2.9 Gbp



Summary of Solexa properties

- Reads are short (36 nt), but get progessively longer (50 nt, 75 nt)
- 8 samples can be sequenced in parallel
- Protocols in use at the CRG:
  - Shotgun sequencing (genomic DNA, cDNA, ChIP-Seq, RIP-Seq)
  - mRNA-Seq
  - Indexed RNA-Seq
  - Small RNA (miRNAs) identification and profiling
  - Gene expression profiling (Tag sequencing, similar to SAGE)
  - Paired end sequencing (long and short paired ends)
- Advantages
  - Millions of reads from your sample
  - Relatively cheap
- Disadvantages
  - Millions of reads from your sample
  - Some biases, poorly understood
  - Sequencing errors (but Illumina improves)
  - Demands on IT infrastructure (data processing, analysis, backup)

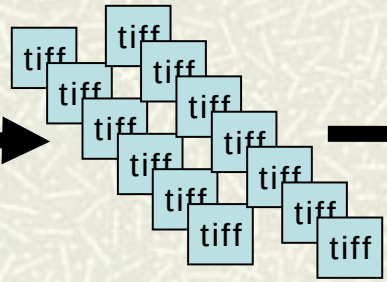
Illumina GA II workflow

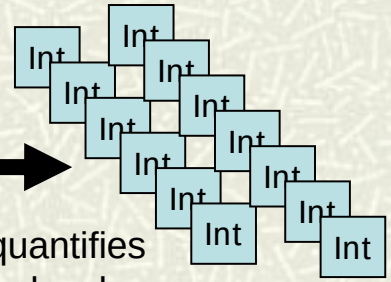
4 images/tile per cycle
~23 GB per cycle
~800GB for 36 cycles

1 intensity file per tile
~1.4 GB per cycle
~ 50 GB for 36 cycles

Illumina Genome Analyzer

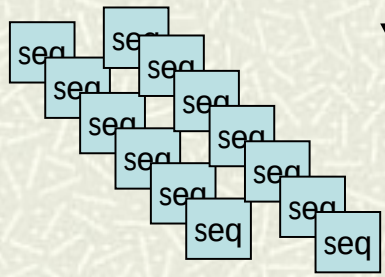
1 flowcell
8 lanes





Firecrest: quantifies the intensity levels in images.

Bustard: calls bases by comparing intensity levels



1 sequence file per tile
~1.1 GB per cycle
~40 GB for 36 cycles

Gerald: Aligns sequences with Eland

Run summary
Error report
Alignment coordinates

Several files per lane
~40 GB for 36 cycles

Computing Infrastructure

Images written to Dell Precision Workstation (2 Dual-cores 2.66 Ghz, 3GB RAM, 1TB RAID)

Images copied to HP Proliant IPAR (2 Quad cores 3 GHz, 16GB RAM, 4TB RAID). Intensity files created 'real-time'.

Images taken by GA II

Images

Images

Images

Intensity files copied to analysis server

Intensity files

Data stored on 1TB FATA discs (17TB RAID). Images written to tape and deleted.

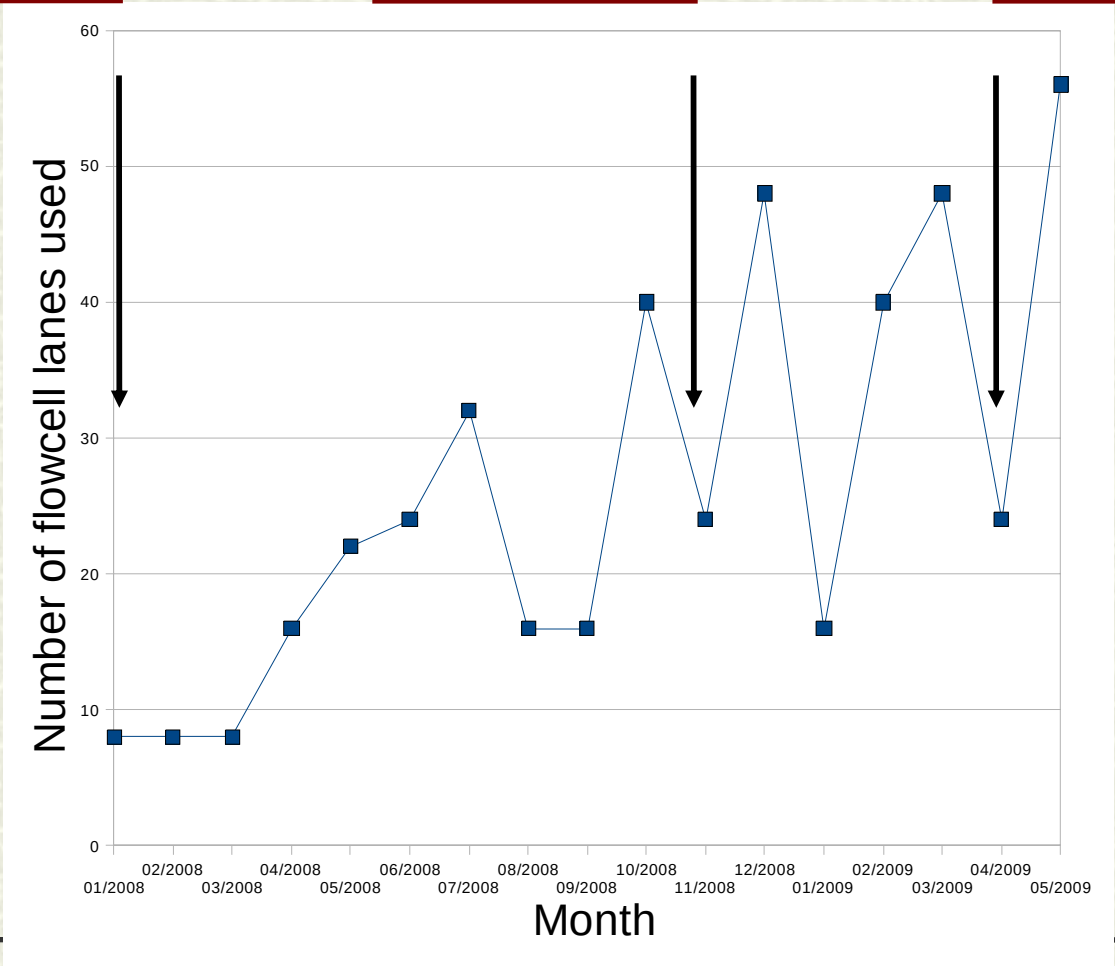
Base-calling, Gerald and post-analysis performed on HP Proliant analysis servers (2 or 4 Quad-cores ~3.2 GHz, 32GB RAM)

Solexa/Illumina sequencing at the CRG

Set up of GA I

Upgrade to GA II

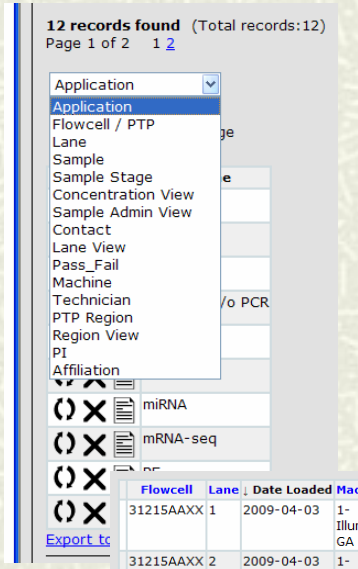
Set up of 2nd GA II

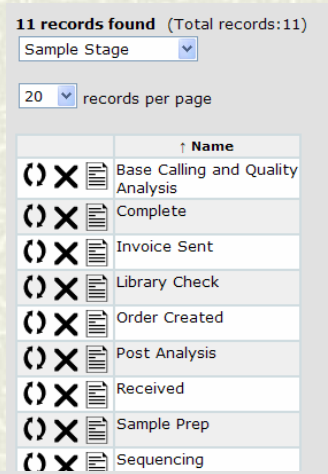


Open-source LIMS for Solexa run management

- Dadabik (database interface creator) web browser user interface
- Dadabik web browser admin interface
- SQL database
- Automated scripts which update the database and send notification e-mails
- Requires Dadabik, SQL, Apache, PHP, and a Linux web server

Open-source LIMS for Solexa run management





Flowcell	Lane	Date Loaded	Machine	Lab ID	Name	Contact	PI	Affiliation	Application	Organism	Stage	Cluster Density	Successful_Cycles	Loading Concentration	Total Reads	Percent Unique	Percent No Match
31215AAXX	1	2009-04-03	1-Illumina GA II	243	HeLa 10% INPUT	klaus fortschegger	Ramin Shiekhattar	CRG	ChIP Seq	Human	Post Analysis	124244	0	6.00	12413152	48.6%	33.9%
31215AAXX	2	2009-04-03	1-Illumina GA II	245	12 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	85549	0	3.60	8554947	61.6%	32.3%
31215AAXX	3	2009-04-03	1-Illumina GA II	246	13 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	78718	0	3.60	7871819	63.8%	31.4%
31215AAXX	4	2009-04-03	1-Illumina GA II	1000	PhiX	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	Genomic		Sequencing	92659	0	2.00	9265916	79.8%	18.9%
31215AAXX	5	2009-04-03	1-Illumina GA II	247	14 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	76231	0	3.60	7623141	78.6%	16.3%
31215AAXX	6	2009-04-03	1-Illumina GA II	248	15 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	99312	0	3.60	9931255	84.1%	8.9%
31215AAXX	7	2009-04-03	1-Illumina GA II	249	16 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	76414	0	3.60	7641463	69.8%	23.9%
31215AAXX	8	2009-04-03	1-Illumina GA II	250	17 cycles	Ultrasequencing Core Facility	Heinz Himmelbauer	CRG	miRNA	Mouse	Post Analysis	68306	0	3.60	6830696	75.3%	19.5%
3130GAAXX	1	2009-03-30	2-Illumina GA II	197	DIS8 (WT)	mercè guzman	Josep Vilardell	CRG	genomic w/o PCR	S. cerevisiae	Post Analysis	116588	36	6.00	11531489	27.2%	53.7%
3130GAAXX	2	2009-03-30	2-Illumina GA II	198	yJV31 (mut)	mercè guzman	Josep Vilardell	CRG	genomic w/o PCR	S. cerevisiae	Post Analysis	76217	36	6.00	7675649	24.3%	54.8%

Cluster densities

- 150.000 clusters per tile recommended
- 100 tiles per lane, 15 million raw sequences per lane
- Number of usable sequences not affected by further density increase
- Improved software (Illumina pipeline 1.3.2 and 1.4.0) to deal with merged clusters

Solexa sequencing:
Many millions of of short reads

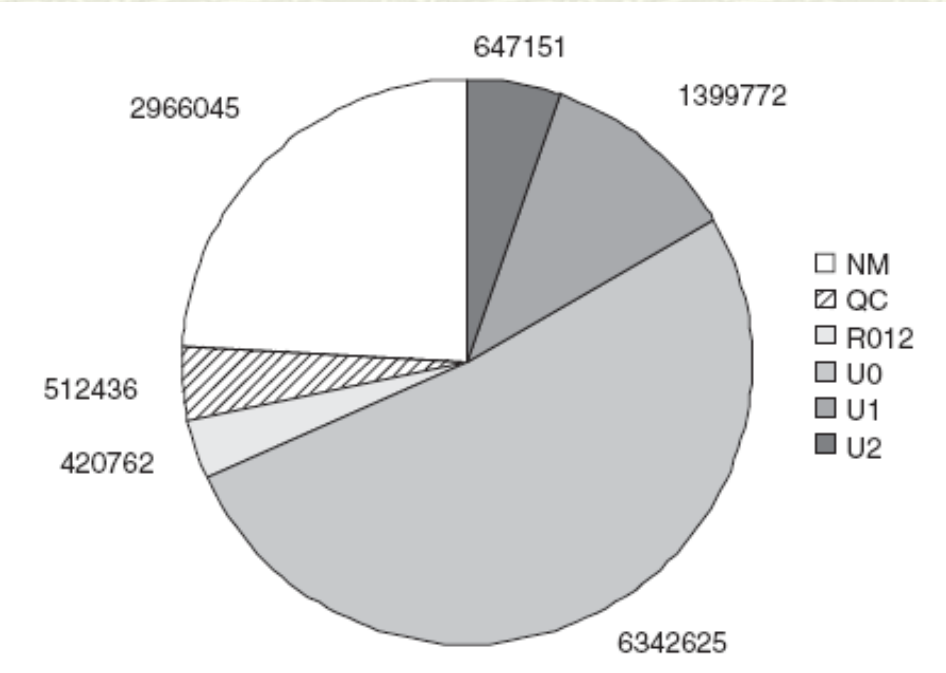
```
seq: AATTATGCCTTGATGTGTGGGCGATCGCTTCACATA
seq: ATGAAAAAATCGCGCTAGATTTATTGCCGGTGATTT
seq: AAACTCAAATTCCTATTATATCAAACAATTTTTATA
seq: AACTGACGGAGTTCCTTTTTGGTCGCTCCGCTTTGC
seq: AGCTTGAATGCCCCTAGTGCTTTTATTTTATCGCGC
seq: ATCCGCATAAAAGCATTTATCCATGTGGGTAAGCTG
seq: ATTGGGGGTGGGATTAGCGCTTGCACTAGATTTTGG
seq: AATAAAACCCCATGCGTTTTACTTTTTCTTGGGCGT
seq: AGCAAAATAAACCAATCATCTTAAGCCAGCAAAGCA
seq: GCTCTTTTGCACGATAAATTTCACCGCTTCTTTATG
seq: AAAAGCGCAATTGAAAGACGATTTACGCCAATTGAG
seq: AGCTTACCCTCTTTATCATGGTAAGTGAAAGGGGGG
seq: GTGGTCAATATTTTGGGGTTTTTTAACATTAAATTT
seq: GGGTTTAAAGCAGTCTTTTTTATAAAAAAGTGATTT
seq: AACAACAAGCTCTCTGTGGGGCTTTTTGGCGGTATC
seq: AGGGATCGTTTTTGCTCACTAACTCGCCCTTTGGCT
seq: GTGATTAAAATGGGGCGCACCCAGCTTCAAGACGCT
seq: ATGAAGCTTTGTCGCTTTGCTTAGTTATAGAGAGTT
seq: GGGTGGTAGGAGTGTGGAGCTAGTAGAGGCTGTGGT
seq: AATTTGGAATAAGTAGAGATAAGTCGTTTAGTATCC
seq: GAGTTTTAAAGTGTCTAGCCCATAAGAAGAAAAAGT
[...]
```

Analysis of Solexa data

- Comparison with reference sequence (not possible with *de novo* sequencing)
  - Genome sequence
  - Transcripts or gene predictions
  - Selected data sets (e.g. miRNA sequences)
- Positioning Solexa reads on the reference genome by alignment
  - No match
  - Match at exactly one position
  - Match at more than one position
- Mostly only reads with exactly one match in the reference sequence are considered for downstream analysis
- Fast algorithms required

Exemplary data (GA I)

- Postioning 12.3 million reads on the *Helicobacter acinonychis* reference genome
- Up to two mismatches permitted



Software to align reads against a reference genome

- Programs differ by: Speed, sensitivity, mismatch tolerance, permitted read length
- Eland (Illumina)
  - - 2 mismatches in alignment tolerated, no Indels
  - - Reads need to have identical lengths
  - - Read lengths up to 32 nt
  - + Very fast
- SeqMap (Jiang, Wong, Bioinformatics, 2008)
  - + Tolerates up to 5 mismatches and Indels
  - + Different read lengths of input sequences is possible
  - + Reads may exceed lengths of 32 nt
  - - Slow
- Soap (Li et al., Bioinformatics, 2008)
  - + Tolerates up to 5 mismatches and Indels
  - - Reads should have similar lengths
  - + Read lengths up to 60 nt
  - + Fast for entire genomes, demanding on hardware

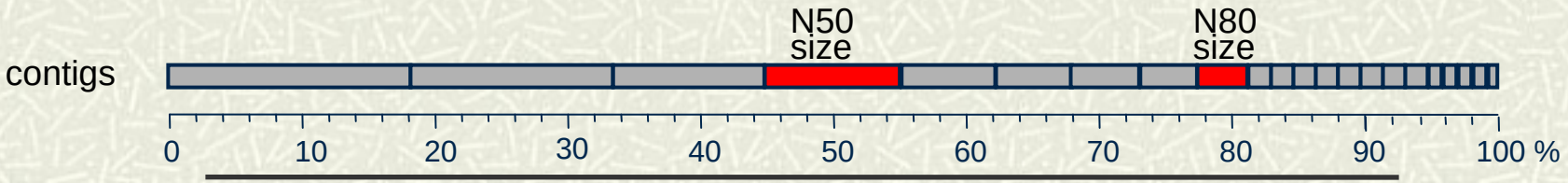
Solexa resequencing

- Alignment against reference genome
- Estimation of sequencing error rate
- Discovery of biologically relevant sequence differences
- Discovery of variation within a species, by inter-strain comparisons

De novo genome assembly from Solexa data? YES !

- Input data: 12.3 million reads (32 nt) from *Helicobacter acinonychis* str. Sheeba (1.6 Mbp)
- Previously sequenced with Sanger technology (Eppinger *et al.*, 2006)
- Identical DNA preparations used

		Reads	Contigs	NotMatched	Mean Len	Max Len	N50	N80	Seq Cov
a	Q >20	1174569	1330	5	1157	10866	2257	1015	97,39%
	Q >25	1098233	1379	5	1110	13684	2232	958	97,16%
	Q >30	1034034	1628	4	937	13673	1893	772	96,49%
	Q >35	978900	1967	4	773	13020	1558	597	96,33%
b	merged		932	5	1603	18577	3476	1364	95,21%
c	all merged		937		1636	18854	3659	1501	97,70%
d	contig overlap >5 bp		228		6762	62478	19890	8012	97,70%



De novo genome assembly from Solexa data? YES !

Resource

ABySS: A parallel assembler for short read sequence data

Jared T. Simpson,[1] Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol[2]

Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada

Widespread adoption of massively parallel deoxyribonucleic acid (DNA) sequencing instruments has prompted the recent development of de novo short read assembly algorithms. A common shortcoming of the available tools is their inability to efficiently assemble vast amounts of data generated from large-scale sequencing projects, such as the sequencing of individual human genomes to catalog natural genetic variation. To address this limitation, we developed ABySS (Assembly By Short Sequences), a parallelized sequence assembler. As a demonstration of the capability of our software, we assembled 3.5 billion paired-end reads from the genome of an African male publicly released by Illumina, Inc. Approximately 2.76 million contigs ≥100 base pairs (bp) in length were created with an N50 size of 1499 bp, representing 68% of the reference human genome. Analysis of these contigs identified polymorphic and novel sequences not present in the human reference assembly, which were validated by alignment to alternate human assemblies and to other primate genomes.

Genome Research, June 2009

Acknowledgements

CRG - Ultrasequencing Unit

Lab: **Ana Vivancos**
Ester Castillo
Anna Menoyo
Maik Zehnsdorf

Bioinformatics:
Robert Kofler
Juliane Dohm
Matt Ingham
Debayan Datta
André Minoche

Mònica Bayés

Cooperations

CRG – Systems Biology Program
Marc Güell
Luis Serrano

CRG- Genes and Disease Program
Eulàlia Martí

Universitat Autonoma de Barcelona
Miguel Pérez-Enciso