sequencing for a
better life

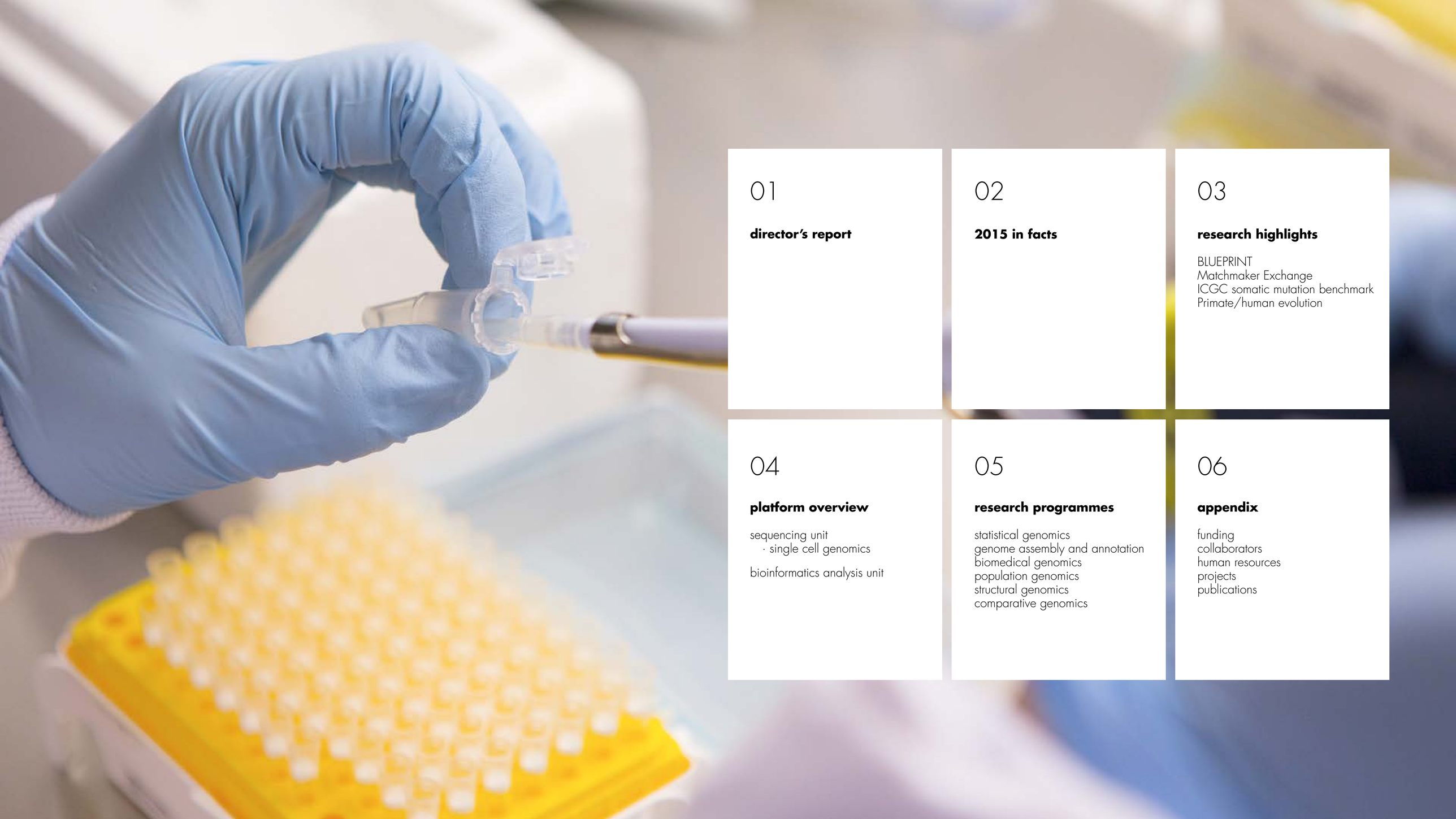
annual report
2015

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre
for Genomic
Regulation

Generalitat
de Catalunya

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD



cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

2015 has been another productive and successful year for CNAG that has also seen big changes. The administrative move to join the CRG will provide stability, and a formidable opportunity for synergy and to strengthen our research ties. Internally we have broadened our areas of research with the incorporation of the Population Genomics team led by Oscar Lao and with the arrival of Holger Heyn our Single Cell Genomics team has gained a new leader.

The CNAG has consolidated its role as a high-quality collaborator in many aspects. Several of our large-scale international projects, the EU-funded project Blueprint of the International Human Epigenome Consortium and the International Cancer Genome Consortium project are nearing their conclusion and we have been instrumental in generating the high quality data necessary for the remarkable findings of the Spanish CLL-ICGC project that have been received with a lot of acclaim by the international research community. We have taken a leading role in the ICGC publication in Nature Communications that summarizes the effort of 83 researchers from 78 institutions to create reliable standards to obtain accurate results in the detection of somatic mutations, which are a hallmark of cancer genomes.

In preparation for large-scale clinical and population-based projects we have increased our computing infrastructure to 3500 computer cores which provide 200 TFlops, and 7.6 petabyte of data storage. This capacity holds all sequencing data produced at CNAG and is used by our bioinformaticians to deliver high quality results. At the same time we have been working hard on the quality, performance, efficiency, and integration of our processes and computing. We have further developed our quality system for the entire process from sample reception, to laboratory and data analysis.

In 2015, we started B-CAST, a large-scale EU-funded project, to characterize the tumours of 10,000 breast cancer patients. This multi-year project will generate a unique opportunity to relate, background genetic profiles, cancer-specific somatic mutations with treatment outcomes.

Ivo G. Gut
Director

2015 in facts

January

The documentary The Dark Gene, partially filmed at the CNAG and with Ivo Gut among the cast, premiered at the Solothurn Film Festival.

February

Top scientists present their latest research discoveries and ideas at the 3rd CNAG Symposium on Genome Research: Rare Diseases.

March

The new research team on Population Genomics, led by Oscar Lao, starts working on in-depth genomic analyses and development of bioinformatics tools to answer questions in population genetics.

April

The CNAG participates in a project to sequence the genome of the mountain gorilla, published in *Science*.

May

The CNAG leads a study that highlights the power of mouse exome sequencing and analysis to identify mutations and genes related to certain phenotypes produced by mutagenesis.

June

A study framed in the EU funded project BLUEPRINT, with the participation of the CNAG, reveals unexpected connection between epigenetic changes associated with lymphocyte maturation and those observed in cancer.

July

The Centre for Genomic Regulation (CRG) incorporates the CNAG to help boost genome research, maximize resources and create synergies between the two institutes.

August

Last BLUEPRINT samples arrive to the centre, in total the CNAG has carried out the sequencing and analysis of over 200 reference epigenomes.

September

The CNAG participates in the 10th European Researchers Night with an interactive activity to show the role of light in sequencing and the relevance of this technique to our society.

October

Sergi Beltran co-authors an overview article about the Matchmaker Exchange platform, aimed to provide a robust and systematic approach to rare disease gene discovery and promoted by the Global Alliance for Genomics & Health (GA4GH).

4th CNAG Symposium on Genome Research: Rare Diseases II, organised in Madrid with the Instituto de Investigación Sanitaria de la Fundación Jiménez Díaz (IIS- FJD).

November

Incorporation of three new nanopore DNA sequencing systems MinIon MkI (Oxford Nanopore Technologies). These devices allow sequencing of very long stretches of DNA (~40kb).

December

A study published in *Nature Communications* and led by the CNAG reveals a high degree of heterogeneity in how cancer genome sequencing is done at different institutions across the globe that partner the International Cancer Genome Consortium (ICGC). The result lays the foundation for the coming era of cancer genomics.

annual report 2015

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

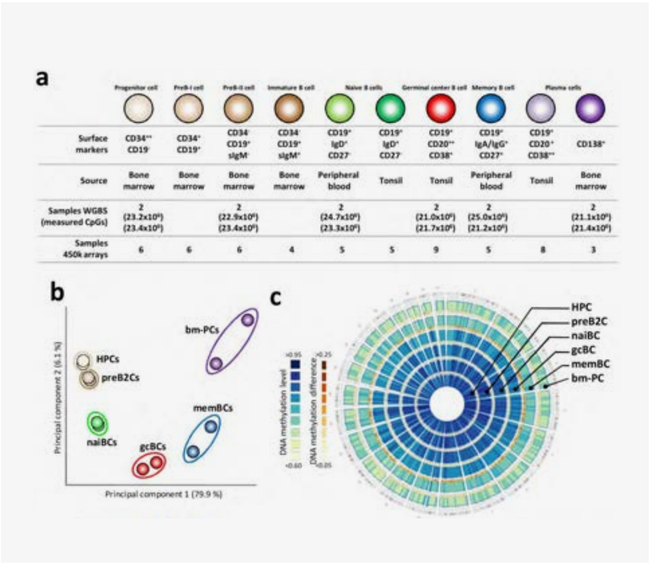
research highlights - BLUEPRINT
the BLUEPRINT study on human B-Cell differentiation

The BLUEPRINT project started in 2011 and is an ambitious initiative to provide a comprehensive set of >100 reference epigenomes from healthy and diseased cell types from the human haematopoietic system. The BLUEPRINT consortium consists of 41 partners organizations, representing 43 academic groups and 9 companies (mostly Small and Medium Enterprises (SMEs)) from 11 countries, and the project is funded until September 2016 under the European FP7 program. The CNAG generated all of the DNA methylation data for the BLUEPRINT project using whole genome bisulfite sequencing (WGBS), as well as developing and operating the primary WGBS analysis pipelines.

Working closely with the Unidad de Hematopatología of IDIBAPS in Barcelona, in conjunction with other groups from Spain, Germany, France, Holland, UK, USA and Korea, the CNAG played a leading role in the analysis and interpretation of DNA methylomes from a set of healthy samples taken from different stages of human B-cell differentiation. It was observed that CpG methylation changed extensively during B-cell maturation affecting around 30% of all measured CpGs. Early differentiation stages mainly displayed enhancer demethylation, which was associated with up-regulation of key B-cell transcription factors. Late differentiation stages, in contrast, showed extensive demethylation of heterochromatin and methylation gain of polycomb-repressed areas, and did not affect genes with apparent functional impact in B cells. This signature, previously linked to aging and cancer, was particularly widespread in mature cells with extended life span.

Work of reference

Kulis, M., et al., Whole-genome fingerprint of the *DNA methylome during human B cell differentiation*. Nat Genet, 2015. **47**(7): p. 746-56.



The figure shows (a) the B-cell populations studied; (b) a principal components analysis (PCA) of the WGBS samples and (c) a summary plot of DNA methylation levels in the WGBS samples

research highlights - Matchmaker Exchange

RD-Connect and identification of rare disease patients with similar genotype and phenotype combinations in other platforms through GA4GH Matchmaker Exchange

High-throughput genome sequencing and analysis has enabled important advances in rare disease research. These efforts, often scattered, have become a valuable source of information for many studies and researchers. RD-Connect, an EU FP7-funded project under the auspices of the International Rare Diseases Research Consortium (IRDiRC), is building a platform to integrate clinical, biosample and –omics data from a huge number of patients. The development of the platform is led by the CNAG and is hosted in Barcelona. With the aim of sharing knowledge beyond the project, RD-Connect is part of the IRDiRC/GA4GH Matchmaker Exchange (MME) project [1]. The main goal of the MME is to identify similar patients (in terms of genotype and phenotype) in other rare disease platforms harbouring pre-filtered or full genomic information.

The MME has a double challenge: defining a standard application programming interface (API), de facto, for genotype and phenotype data sharing, and the creation of a federated network that interconnects other platforms (e.g., Phenome-Central, Gene-Matcher, Café Variome, The Genesis Project, RD-Connect). The first version of the MME API [2] allows researchers to look for patients with similar phenotypic profiles and/or overlap of manually selected candidate genes. The prototype of the second version of the API (led by RD-Connect and The Genesis Project with the support of the GA4GH Data Working Group) is extending the functionality in order to allow queries on unfiltered genomic data (e.g. full panels, whole exomes or genomes) or matching even when candidate genes have not been identified (termed 1-sided and 0-sided hypothesis matching). New components and filters enable more powerful and specific questions such as "Do you have any patients similar to one with phenotypes X, Y, Z and with one rare (allele frequency < 0.01), harmful (missense or stopgain) variant in e.g. NGLY1 or TTN?" The new genome component introduces filtering options such as pathogenicity/deleteriousness score, allele frequency, gene annotation (e.g. one or more involved) or consequence. Matched results are securely returned with an internal identifier of patient and sorted by similarity (float number between zero -no similarities- and one - perfect match-) allowing the researcher to perform the query and to contact the contributor of the matched patient record.

Works of reference

1. Philippakis AA et al. *The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat.* 2015 Oct;**36**(10):915-21. doi: 10.1002/humu.22858.
2. Buske OJ et al. *The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. Hum Mutat.* 2015 Oct;**36**(10):922-7. doi: 10.1002/humu.22850.

Matchmaker Exchange participating databases and programs.

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

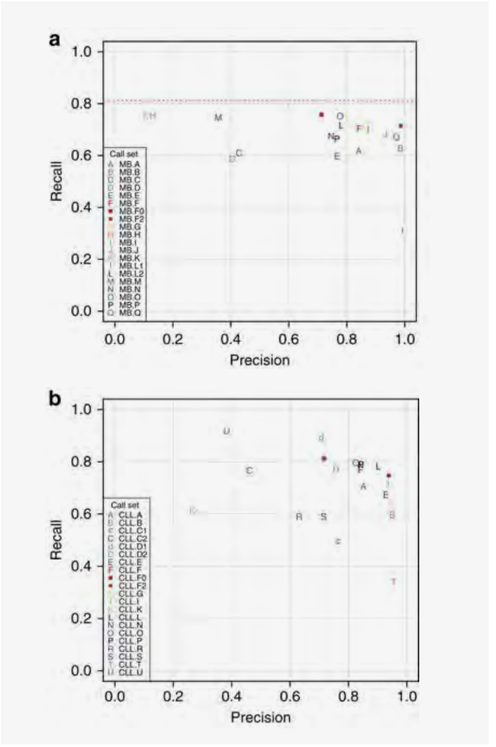
research highlights - ICGC somatic mutation calling benchmark
building the foundations for cancer genomic analysis for research and clinical diagnostics

The International Cancer Genome Consortium (ICGC) started in 2008. It is a large international initiative that joins 83 individually funded projects to generate comprehensive descriptions of 25,000 cancer genomes across most forms of cancer. Genomic analyses are done by different centres across the globe and data is provided to a central database. CNAG is involved through its work for the Spanish ICGC-CLL project, the European CageKid kidney cancer project and several of the French ICGC projects. As the data was generated and analyses were done at different places and by different researchers using different tools, it was unclear how comparable results actually are. CNAG took the lead in a benchmarking effort of the ICGC of both the sequencing and the data analysis which implicated 83 researchers from 78 institutions.

Comparing the sequencing carried out by five different large-scale sequencing centres, we found that protocols used differed substantially and because of this there were huge differences in the quality of sequences provided. The sequencing provided by the CNAG was of the highest quality. Twenty different analytical teams used the sequencing data from the CNAG for the identification of somatic mutations. The joint data from the first part of the study was used to generate a Gold set of mutations calls that were used for the assessment of the somatic mutation calls from the different participating analytical teams. The outcome was an alarmingly small degree of agreement between different analytical teams. In particular identifying somatic insertion deletion mutations was found to be very difficult. The ICGC groups have used this effort to improve their procedures. The results were published in *Nature Communications* and data has been made available to the entire scientific community to benchmark their analytical methods and improve their procedures.

Works of reference

Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nat Commun, 2015. 6: p. 10001.



Using the same whole genome sequence data dramatic differences in the ability to detect somatic mutations were observed between participating teams.

research highlights - primate / human evolution
how gene regulation has contributed to recent human and non-human primate evolution

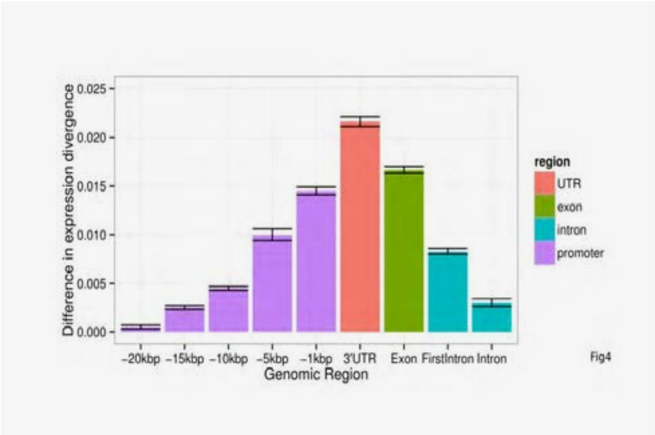
In 2015, two projects from the Comparative Genomics group provided better understanding about how gene regulation has contributed to recent human and non-human primate evolution. We described a global comparative analysis of CpG methylation patterns between humans and great apes (chimpanzee, gorilla and orangutan) using whole genome bisulfite sequencing and its comparison to patterns of SNP diversity in these species. We identified hundreds of novel regions showing an exclusive pattern of DNA methylation in blood from humans compared to great apes and estimated that ~25% of these regions were still detectable throughout several human tissues, highlighting that they most likely fixed. Moreover, these regions were enriched for specific histone modifications and, contrary to expectations, they were located distal to transcription start sites. We also reported, for the first time, a close interplay between inter-species genetic and epigenetic variation which offers a novel perspective to decipher the mechanistic basis of human-specific DNA methylation patterns and the interpretation of inter-species non-coding variation.

A second project shows for the first time VNTR (tandem repeat) diversity on a genome-wide level in human and non-human great ape populations. We demonstrated the role of VNTRs as a source for expression divergence between humans and their closest primate relatives. The presence of tandem repeats in genes, even controlling for polymorphisms, was associated with higher levels of expression divergence between human and these primates, and this association holds for genes with repeats in their 3' untranslated region, in introns, and in exons. We believe that our work supports the notion that non-coding variation has a prominent role in genome regulation.

Works of reference

Hernando-Herraez, I., et al., *The interplay between DNA methylation and sequence divergence in recent human evolution*. Nucleic Acids Res, 2015. **43**(17): p. 8204-14.

Bilgin Sonay, T., et al., *Tandem repeat variation in human and great ape populations and its impact on gene expression divergence*. Genome Res, 2015. **25**(11): p. 1591-9.



Gene expression divergence between human and chimpanzee according to the genomic position of VNTR differences. Several patterns can be noticed: VNTR variation in 3'UTR regions seems to have a higher influence in gene expression divergence than other genic regions. Also, there is a decreasing effect of gene expression in relation to the distance of VNTR to the transcription start site.

annual report 2015

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

sequencing unit

the sequencing unit of CNAG assures the sequence data production at high proficiency and quality standards necessary to enable to support at the state-of-art level the research and to be appreciated by our collaborative projects.

Head of the Unit:
Marta Gut

Biorepository:
Lidia Agueda (Manager, from January to September), Ana González (Manager, from September to December)

Sample Preparation Team:
Julie Blanc (Manager, from January to April), Marta López (Manager, from April to December), Ana González (until August), Pilar Herruzo, Maite Rico, Caterina Mata, Laetitia Casano, Amaya Alzu (until February), Beatriz Fontal (until October), Yasmina Mirassou (from February), Giulia Lunazzi (from February), Regina Antoni (from November).

Sequencing Production Team:
Katja Kahlem (Manager), Aurora Padrón, Javier Gutiérrez (until August), Nuria Aventin (from November), Glòria Plaja.

Lab Support Team:
Esther Lizano, Nicolas Boulanger (until August), Javier Gutiérrez (from September), Lidia Sevilla

The Sequencing Unit operates 12 Illumina 2nd generation sequencers and 3 Oxford Nanopores 3rd generation sequencing devices (MkI) and is supported by five teams – Biorepository, Sample Preparation, Sequencing Production, Support team and Single Cell Genomics team. The inter-team communication is assured and information tracking guaranteed by the Laboratory Information Management System.

The biorepository team manages reception, storage and quality control of all the samples received from collaborators and transfers the sample aliquot to the Sample Preparation team which is responsible for preparing sequencing ready material for the Sequencing Production team to load prepared libraries onto the sequencing instruments. The support team tests and sets up new protocols – manual and automated – and follows the correct functioning of the CNAG´s large and small technical equipment.
The Single Cell Genomics team elaborates new applications to increase the portfolio of offerings in single cell transcriptomics and genomics.

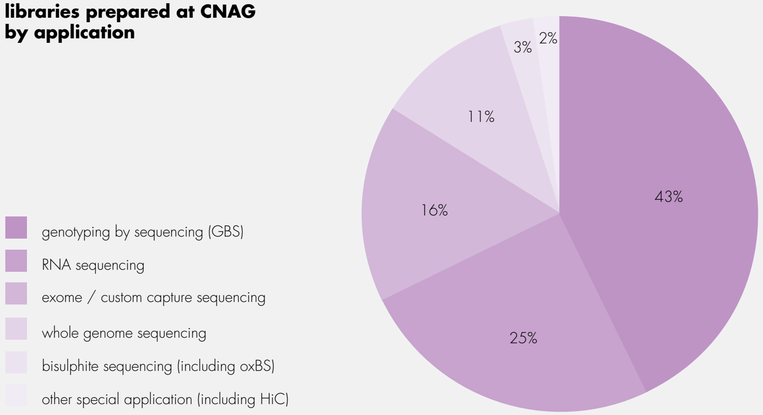
Research Projects

- Scaling up by 60% the number of sample reception, QC and storage
- FFPE samples QC improvements with introduction of qualitative and quantitative qPCR
- Implementation of automated normalization for Genotyping by Sequencing protocol (GBS)
- Reduction of the failures during sample preparation to less than 5%
- Improved version of the oxBS-WGS (DNA oxidation followed by bisulphite conversion -whole genome sequencing) protocol to detect and quantify 5-hydroxymethyl cytosine and 5-methylcytosine on single nucleotide level
- Comparison and implementation of Nimblegen, Agilent v6 and TruSeq Exome capture protocols
- Implementation of low input protocols, notably Whole Genome Bisulfite Sequencing (WGBS), Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) sample preparation
- Incorporation of 3 Oxford Nanopore Technologies MinIon MkI sequencers
- Development and implementation of sample preparation protocols for MinIon MkI
- Automation of GBS, DNA size selection and purification, WGS and stranded RNA Seq protocols
- Development of QC protocol for precision of liquid handlers pipetting

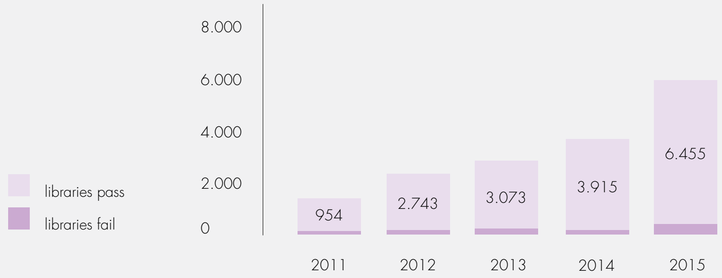
sequencing unit

the sequencing unit of CNAG assures the sequence data production at high proficiency and quality standards necessary to enable to support at the state-of-art level the research and to be appreciated by our collaborative projects.

libraries prepared at CNAG by application



libraries prepared at CNAG by year



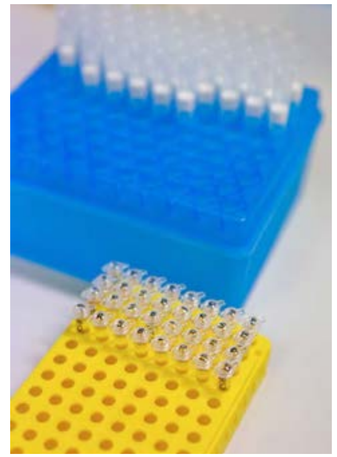
single cell genomics

Team Leader:
Maria Méndez-Lago (until November), Holger Heyn (from December)

Postdoctoral Fellows:
Amy Guillaumet, Gustavo Rodríguez-Esteban (shared with the Statistical Genomics team)

The Single Cell Genomics team is dedicated to advance genome research of single cells to enable researchers the use of single cell technology in a scalable and affordable manner.

The team implemented two independent strategies to perform transcriptomic analysis of single cells. The first technique, massively parallel single-cell RNA sequencing (MARSseq), utilizes the sequence information of short fragments of the transcripts' 3'-ends to quantify gene expression of single cells after separating them in 384 well plates. The second method, SMART sequencing, is performed within C1 Single-Cell Auto Prep system from Fluidigm and produces full-length transcript information for subsequent quantification. While the commercially available version of the C1 system allows the profiling of 96 cells per run, a beta-tested variant captures and processes up to 800 cells. Both techniques were established using cell line model systems and primary samples in close collaboration with the developers.
In experiments with external collaborators we applied single cell genomics techniques to deconvolute tissue composition, determine cell type heterogeneity, identify novel cell type defining markers or track transcriptional dynamics.



Research Projects

· Stem cell differentiation trajectories
· Dynamics during trans-differentiation and induced pluripotency
· Control of variability by post-transcriptional regulation
· Cell type composition of complex tissues
· Heterogeneity in blood cell biology

bioinformatics analysis unit

the bioinformatics analysis unit develops and operates state-of-the-art pipelines, tools and databases to manage and analyse the sequencing data generated at the CNAG.

Head of the Unit:
Sergi Beltran

Production Bioinformatics Team:
Matthew Ingham (Manager), Raul Alcántara, Colin Kingswood (until September), Eloi Casals (from October)

Data Analysis Team:
Sergi Beltran (Manager), Sophia Derdak, Steven Laurie, Raul Tonda, Davide Piscia, Joan Protasio (from June), Inés Martínez (from November), Enric Serra (until December), Anastasios Papakonstantinou (from December)

Support Team:
Jordi Camps, Jean-Rémi Trotta

The group collaborates closely with internal and external groups and delivers customised high quality user-friendly results. The group is also involved in European projects such as RD-Connect, in which it plays a central role.

The activity is carried out by a group of thirteen highly qualified data analysts, software engineers and bioinformaticians divided in three teams. The Production Bioinformatics team develops and operates the Laboratory Information Management System (LIMS) and pipelines to process, control the quality and transfer the data generated by the Laboratory. The Support team develops and operates standardised pipelines for common processes such as mapping and coverage computation. It also provides customised support to other CNAG teams. Finally, the Data Analysis team develops bioinformatics solutions to analyse sequencing data, mainly related to germinal variant and somatic mutation identification. The team analyses internal and external projects and prioritizes the findings together with the collaborators. Although most of the activity is related to clinical research (mainly on Mendelian disorders and cancer), the participation in agrogenomics and model organisms projects has been increasing steadily during the past years.

Services

- Collaborative analyses
- Experimental design
- Data processing and quality control
- Germinal variant identification
- Somatic mutation identification
- Variant annotation and filtering
- Copy Number Variant identification
- Genotyping-by-Sequencing (GBS)

Research lines

- Bioinformatics for Rare Disease Research and Clinical Genomics
- Development of pipelines and tools
- Benchmarking of data analysis methods
- Variant annotation systems
- Omics data integration
- Agrogenomics (GBS)

Selected publications

Derdak, S., et al., *Genomic characterization of mutant laboratory mouse strains by exome sequencing and annotation lift-over.* BMC Genomics, 2015. **16**: p. 351.

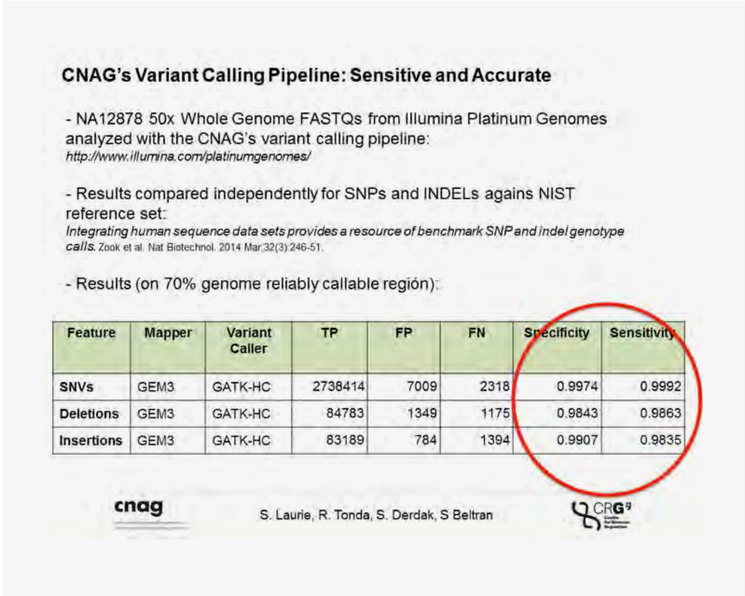
Philippakis, A.A., et al., *The Matchmaker Exchange: a platform for rare disease gene discovery.* Hum Mutat, 2015. **36**(10): p. 915-21.

Sanz-Pamplona, R., et al., *Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer.* Clin Cancer Res, 2015. **21**(20): p. 4709-18.

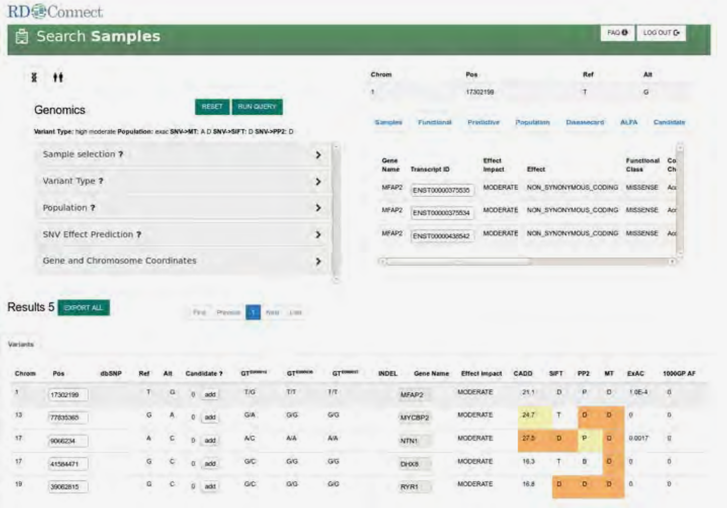
Metzger, J., et al., *Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses.* BMC Genomics, 2015. **16**: p. 764.

bioinformatics analysis unit

the bioinformatics analysis unit develops and operates state-of-the-art pipelines, tools and databases to manage and analyse the sequencing data generated at the CNAG.



CNAG's variant calling pipeline sensitivity and accuracy. The variant calling pipeline has been benchmarked against the reference set of variants from the NA12878 released by the GIAB-NIST consortium. CNAG's pipeline includes internally developed GEM3 aligner and it is extremely sensitive and accurate overall.



RD-Connect genomics platform front-end. The platform is developed by the Bioinformatics Analysis Unit and it includes genomes and exomes analysed with a standard pipeline to make them comparable. Samples can be easily analysed with the user-friendly interface selecting the available filters (variant quality, genomic and functional annotation, population frequency and pathogenicity predictions). Results are linked to a broad range of internal and external tools and resources (genotype:phenotype prioritization, clinical information, genome browsers, functional information, expression tissues, etc.). Try it out at platform.rd-connect.eu.

annual report 2015

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

statistical genomics team

Team Leader:
Simon Heath

Staff Scientist:
Emanuele Raineri

Postdoctoral Fellows:
Angelika Merkel, Ron Schuyler, Gustavo Rodríguez-Esteban (shared with the Single Cell Genomics team)

Data Analyst:
Anna Esteve

Software Engineer:
Marcos Fernández

Bioinformatics Technician:
Marc Dabad

PhD Student:
Santiago Marco

The research aim of the team is the development and implementation of statistical methods for efficient analysis of omics datasets, with a particular focus on epigenomics and transcriptomics. The main area our team has been working on recently is the analysis and interpretation of DNA methylation data from Whole Genome Bisulfite Sequencing (WGBS), with the aim of pinpointing the role of DNA methylation in normal cell differentiation as well as in pathologies such as cancer. The largest WGBS project we have been working with is BLUEPRINT, which consists of over 200 samples of mostly healthy tissue from different blood cell types, where we have been working on the changes in methylation linked with normal B-cell differentiation and in multiple myelomas.

We have also continued our work on the development of efficient methods of mapping sequence reads using the GEM mapper. The latest version, GEM3, is now 5-6 times faster than both the previous versions and leading alternative mapper BWA-Mem. More importantly, GEM3 natively produces SAM output that is fully standards compliant, allowing for simple integration of GEM3 in existing pipelines. In addition, GEM3 can now natively process WGBS data, and is by far the fastest and most accurate aligner for WGBS data that is currently available.

Research projects

- Methods for DNA methylation analysis from WGBS experiments (computational and statistical methods for calling DNA methylation and for testing for differential methylation between samples)
- Comparison of DNA methylation patterns and their relationship with nucleosome positioning across different haemopoietic cell types
- Characterization of partially methylated domains in different haemopoietic cell types and the relationship with differentiation and cancer
- Analysis pipelines for single cell RNA-Seq analysis
- DNA sequence mapping for WGBS data and for multi-kb reads from new sequencing technologies (Pacific Biosystems and Oxford Nanopore)

Selected Publications

Kulis, M., et al., *Whole-genome fingerprint of the DNA methylome during human B cell differentiation*. Nat Genet, 2015. **47**(7): p. 746-56.

Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nat Commun, 2015. **6**: p. 10001.

Agirre, X., et al., *Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers*. Genome Res, 2015. **25**(4): p. 478-87.

Lee, S.T., et al., *Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures*. Nucleic Acids Res, 2015. **43**(5): p. 2590-602.

genome assembly and annotation team

Team Leader:
Tyler Alioto

Postdoctoral Fellow:
Fernando Cruz

Technician:
Jèssica Gómez

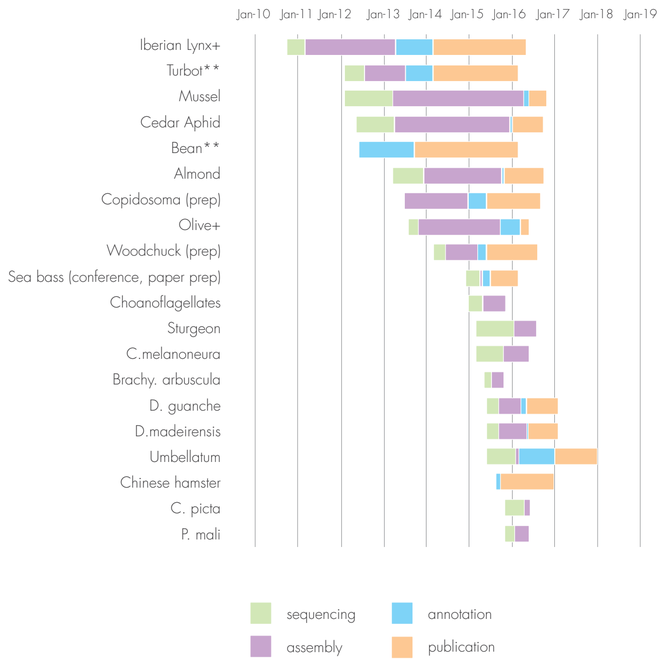
The primary responsibility of the Assembly and Annotation Team is to carry out "genome projects" in the classical sense, i.e. sequencing, assembling and annotating genomes de novo. We specialize in large eukaryotic genomes (mainly animals and plants), but we also assemble and annotate transcriptomes as well as analyze metagenomes and metatranscriptomes (ocean water, soil, etc.). Other types of genomes we analyze include those of organelles (chloroplast, mitochondria), endosymbionts (especially, those of insects), and even cancer genomes. In fact, we even led a benchmarking effort within the International Cancer Genome Consortium to identify best practices in somatic mutation calling using whole-genome sequencing (see Research Highlights).
Genome assembly is not only difficult due to the sheer size of the data and computational requirements, but also because the biology of genomes is confounded by repetitive elements, polyploidy and variation (single-nucleotide, insertions/deletions, and larger structural variants). We focus our efforts on meeting and overcoming these challenges, developing new computational protocols as each project demands. Annotation of the gene content of the newly assembled genome is key to understanding the genome, once finished. On this front we have made progress in developing a rapid robust annotation pipeline, significantly cutting down the time required to annotate a genome.

Research lines

- Genome sequence assembly
- Gene prediction/Genome annotation
- Transcriptome reconstruction
- Functional annotation
- Metagenomics
- Cancer genomics

Selected publication

Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nat Commun, 2015. **6**: p. 10001.



*Genome projects in 2015. The timeline for each project is broken down roughly by the main activity. Overlaps and iterations are not shown for clarity. **Published. +Manuscript accepted.*

biomedical genomics group

Group Leader:
Ivo Glynne Gut

Postdoctoral fellows:
Justin Whalley, Meritxell Oliva (until March), Gian-Andri Thun, Miranda Stobbe, Darek Kedra (from July)

PhD students:
Lukasz Roguski

The Biomedical Genomics Group focuses on extracting genomic information from sequencing in disease studies. We apply computational methods to determine genetic and genomic causes of disease and reversely also study the effects of the disease on the genome. For our studies we use data that is generated at the CNAG and combine it with data that we retrieve from other sources. This allows us to increase the power of our studies and ask the data questions that were not necessarily at the base of the initial study design. We have mainly been working on three classes of diseases:
1. Cancer
2. Rare diseases
3. Respiratory disorders

Research projects

- Cancer: Our main effort relates to the International Cancer Genome Consortium where we helped unearth the first recurrent somatic mutations that are not directly in protein-coding genes, but regulatory elements that induce epigenetic changes. We presented the results of benchmarking of somatic mutation analysis and calling of cancer genome sequencing (see Research Highlight). Substantial effort now is going to the PanCancer study, where we have been investigating the origin of somatic insertion and deletion mutations in cancer genomes and have been coordinating the Quality Control working group.
- Rare Disease: In rare diseases we have been concentrating on the RD-Connect project and data with the objective to mine the data for phenotype-specific modifier variants.
- Respiratory Disorders: In respiratory disorders we have been investigating the manifestation of genomic alterations and their relation to gene expression and clinical phenotypes with a particular focus on the presentation of early signs of known and potentially causal genomic events.

Selected publications

Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nat Commun, 2015. 6: p. 10001.

Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic leukaemia*. Nature, 2015. 526(7574): p. 519-24.

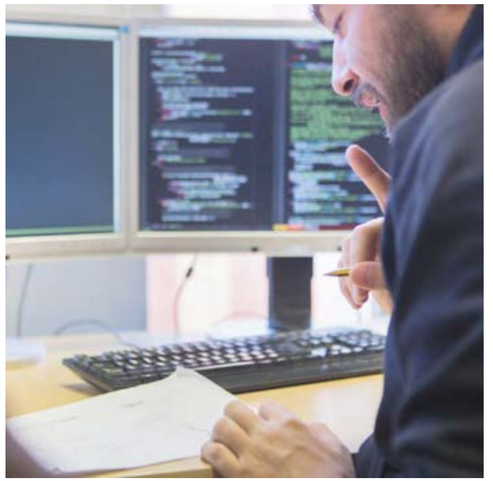
Kulis, M., et al., *Whole-genome fingerprint of the DNA methylome during human B cell differentiation*. Nat Genet, 2015. 47(7): p. 746-56.

Agirre, X., et al., *Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers*. Genome Res, 2015. 25(4): p. 478-87.

population genomics team

Team Leader:
Oscar Lao

The Population Genomics Team started in 2015. Our team focuses on describing and quantifying the genetic variation present in current populations in order to understand the micro-evolution of the given species and asses the phenotypic consequences of such genetic diversity. In particular, we address questions related to which is the genetic origin from a population point of view of a given individual, which are the demographic and selective factors that shaped the genetic variation present in a population and how ultimately this variation influences and allows us to detect the individual risk in phenotypes of interest. In order to achieve these goals, we are actively working on developing new tools and algorithms for describing population substructure in the genome and understanding the biological implications of such structure, identifying the fingerprint of polygenic adaptation in complex phenotypes and evaluating the impact of archaic introgression in phenotypes of interest. Our team focuses on human species but the universality of the proposed methods allows us to apply them to other model organisms.



Research projects

- Development of new algorithms for predicting genetic ancestry
- Identification of the fingerprint of polygenic adaptation in complex phenotypes
- Analysis of the impact of archaic introgression in the evolution of complex phenotypes

Selected publications

Medina-Gomez, C., et al., *BMD Loci Contribute to Ethnic and Developmental Differences in Skeletal Fragility across Populations: Assessment of Evolutionary Selection Pressures*. Mol Biol Evol, 2015. **32**(11): p. 2961-72.

Wollstein, A. and O. Lao, *Detecting individual ancestry in the human genome*. Investig Genet, 2015. **6**: p. 7.

structural genomics group

Group Leader:
Marc Marti-Renom

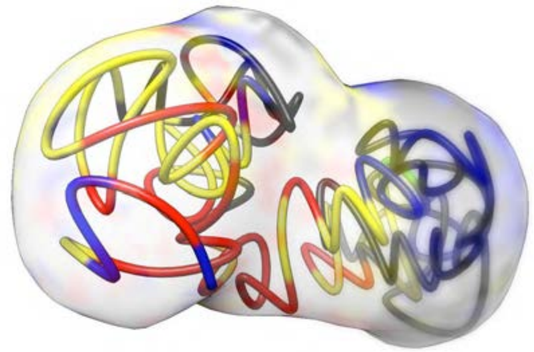
Staff Scientist:
Davide Baù

Posdoctoral Fellows:
Irene Farabella, Marco di Stefano, Yannick Spill, François Serra

PhD Students:
Silvia Galan, Paula Soler, Francisco Martínez-Jiménez, Gireesh K. Bogu, David Dufour, Carlos Baeza

Technicians:
Yasmina Cuartero, Michael Goodstadt

How biomolecules fold and function in a three-dimensional space is one of the most challenging questions in biology. For example, we have limited knowledge on how the 2-meter-long DNA molecule folds in the micro-size nucleus or how RNA, proteins and small chemical compounds fold and interact to perform their most basic functions of the cell. Our research group employ the laws of physics and the rules of evolution to develop and apply computational methods for predicting the 3D structures of macromolecules and their complexes. By doing so, we have recently contributed to understanding how the folding of the sex chromosome (III) influences the mating switch during yeast cell cycle or have identified about 3,000 functional novel long non-coding RNA molecules in Mouse. We have also developed new computational tools to predict the binding sites for chemical compounds in the surface of proteins, which have been already applied to the discovery of potential new treatments against tuberculosis.



Research projects / Research lines

- Structure determination of genomes. We develop methods for determining the 3D organization of the chromatin
- Comparative RNA structure prediction. We develop a series of tools for the alignment of RNA structures and the prediction of their structures and functions.
- Protein-Ligand interactions. We develop methods for comparative docking of small chemical compounds and their target proteins.

Selected publications

Trussart, M., et al., *Assessing the limits of restraint-based 3D modeling of genomes and genomic domains*. Nucleic Acids Res, 2015. **43**(7): p. 3465-77.

Belton, J.M., et al., *The Conformation of Yeast Chromosome III Is Mating Type Dependent and Controlled by the Recombination Enhancer*. Cell Rep, 2015. **13**(9): p. 1855-67.

Serra, F., et al., *Restraint-based three-dimensional modeling of genomes and genomic domains*. FEBS Lett, 2015. **589**(20 Pt A): p. 2987-95.

Martinez-Jimenez, F. and M.A. Marti-Renom, *Ligand-target prediction by structural network biology using nAnnoLyze*. PLoS Comput Biol, 2015. **11**(3): p. e1004157.

comparative genomics group

Group Leader:
Tomas Marques-Bonet

Postdoctoral Fellow:
Martin Kuhlwilm (DFZ fellowship)

PhD Students:
Raquel Garcia (FPU fellowship), Jessica Hernandez (FPI fellowship), Marc deManuel (FI fellowship), Lukas Kuderna (FPI fellowship), Claudia Fontsere (La Caixa fellowship) and Aitor Serres.

What make us human? This is a question still in the foundation of many disciplines. Our team analyzes a wide range of genome variants to determine processes, variants and molecular features that are intrinsic of our species. To do so, we study full genome, epigenomes and transcriptomic sequences of humans and great apes with that aim to better understand human specific features.

Research lines

- Evolution of gene regulation
- Population demography of non-human primate species
- Copy number variation in canid species

Selected publications

Hernando-Herraez, I., et al., *DNA Methylation: Insights into Human Evolution*. PLoS Genet, 2015. **11**(12): p. e1005661.

Hernando-Herraez, I., et al., *The interplay between DNA methylation and sequence divergence in recent human evolution*. Nucleic Acids Res, 2015. **43**(17): p. 8204-14.

Bilgin Sonay, T., et al., *Tandem repeat variation in human and great ape populations and its impact on gene expression divergence*. Genome Res, 2015. **25**(11): p. 1591-9.

Der Sarkissian, C., et al., *Evolutionary Genomics and Conservation of the Endangered Przewalski's Horse*. Curr Biol, 2015. **25**(19): p. 2577-83.

Xue, Y., et al., *Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding*. Science, 2015. **348**(6231): p. 242-5.

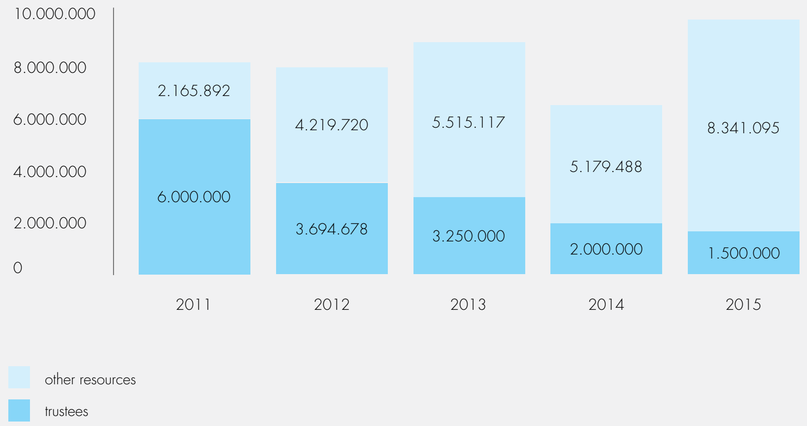
annual report 2015

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

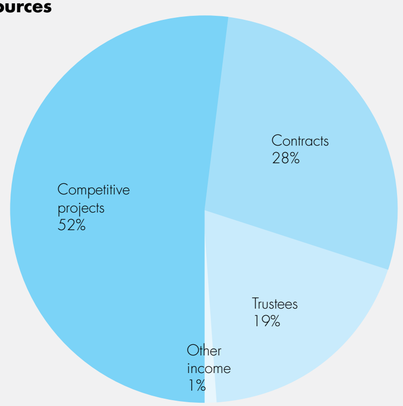
CRG
Centre for Genomic Regulation

cnag's funding

cnag's funding evolution

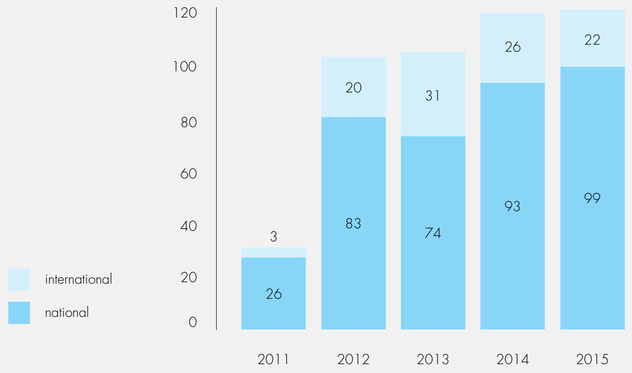


cnag's funding by sources 2015

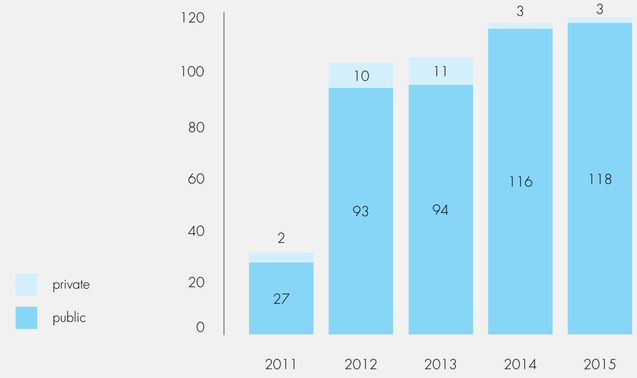


collaborators

collaborators by origin

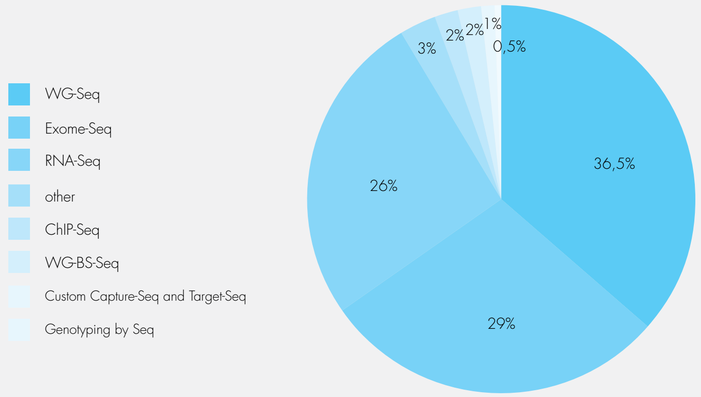


collaborators by sector

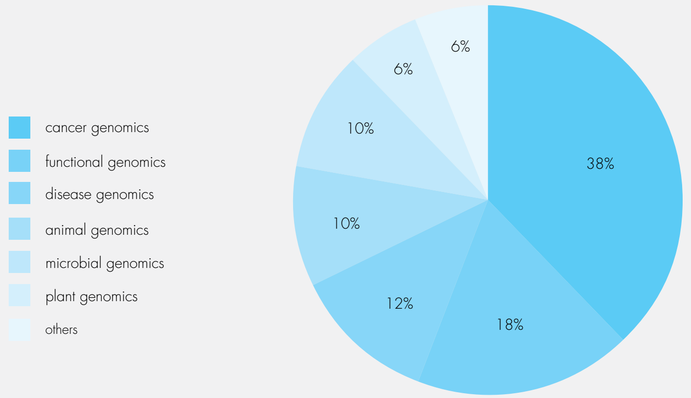


collaborators

annual activity by sequencing application



annual activity by research area



human resources
Staff evolution by area and position

	2011	2012	2013	2014	2015
SEQUENCING (LAB)	13	17	19	22	21
Unit heads and managers	5	5	5	6	5
Engineers and technicians	7	11	13	14	13
Support	1	1	1	1	1
Postdocs and staff scientists	-	-	-	1	2
BIOINFORMATICS (IT)	15	24	28	36	41
Group leaders and team leaders	5	6	5	4	6
Postdocs and staff scientists	5	9	12	14	13
Engineers and technicians	4	4	7	15	18
PhDs and Master's students	1	5	4	3	4
MANAGEMENT (ADMIN)	6	6	5	6	6
TOTAL	34	47	52	64	68

human resources
Staff evolution by area and gender

	2011	2012	2013	2014	2015
SEQUENCING (LAB)	13	17	19	22	21
Male	1	2	2	2	2
Female	12	15	17	20	19
BIOINFORMATICS (IT)	15	24	28	36	41
Male	13	20	24	28	31
Female	2	4	4	8	10
MANAGEMENT (ADMIN)	6	6	5	6	6
Male	3	3	2	2	1
Female	3	3	3	4	5
RATIO FEMALE VS TOTAL	50%	46.81%	46.15%	50%	50%

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

projects
International competitive projects in force in 2015

acronym	project type	timeline	funding
ESGI	FP7 Collaborative Project and Coordination and support Action	2011-2015	€1,145,678
CHR_CYCLE	Human Frontiers	2011-2015	€263,500
AIRPROM	FP7 Collaborative Project	2011-2016	€459,400
BLUEPRINT	FP7 Collaborative Project	2011-2016	€2,739,628
IBD-CHARAC	FP7 Collaborative Project	2012-2016	€994,600
RD-CONNECT	FP7 Collaborative Project	2012-2018	€1,490,496
BBMRI-LPC	FP7 Collaborative Project & Coordination and Support Action	2013-2017	€494,426
4DGENOME	ERC Synergy	2014-2019	€1,752,612
PHYLOCANCER	ERC Consolidator	2014-2019	€320,004
BCAST	H2020 Collaborative Project	2015-2020	€1,846,500
MUG	H2020-EINFRA	2015-2018	€408,556
ELIXIR-EXCELERATE	H2020 INFRADEV	2015-2019	€200.000
TOTAL			€12,115,400

projects
National competitive projects in force in 2015

acronym	project type	timeline	funding
RED-BIO INB	FIS/ ISCIII	2014-2017	€359,150
3D GENOMES	MINECO	2014-2016	€205,000
GENOME ANALYSIS	Suport a Grups de Recerca (Generalitat de Catalunya)	2014-2016	€77,000
BIOT	MINECO	2015-2017	€215,380
TOTAL			€856,530

publications

1. Agirre, X., et al., *Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers*. Genome Res, 2015. **25**(4): p. 478-87.
2. Alioto, T.S., et al., *A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing*. Nat Commun, 2015. **6**: p. 10001.
3. Belton, J.M., et al., *The Conformation of Yeast Chromosome III Is Mating Type Dependent and Controlled by the Recombination Enhancer*. Cell Rep, 2015. **13**(9): p. 1855-67.
4. Bilgin Sonay, T., et al., *Tandem repeat variation in human and great ape populations and its impact on gene expression divergence*. Genome Res, 2015. **25**(11): p. 1591-9.
5. Cheng, T.H., et al., Common colorectal cancer risk alleles contribute to the multiple colorectal adenoma phenotype, but do not influence colonic polyposis in FAP. Eur J Hum Genet, 2015. **23**(2): p. 260-3.
6. Der Sarkissian, C., et al., *Evolutionary Genomics and Conservation of the Endangered Przewalski's Horse*. Curr Biol, 2015. **25**(19): p. 2577-83.
7. Derdak, S., et al., *Genomic characterization of mutant laboratory mouse strains by exome sequencing and annotation lift-over*. BMC Genomics, 2015. **16**: p. 351.
8. Dobrynin, P., et al., *Genomic legacy of the African cheetah, Acinonyx jubatus*. Genome Biol, 2015. **16**: p. 277.
9. Esteban-Jurado, C., et al., *Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer*. Genet Med, 2015. **17**(2): p. 131-42.
10. Guillen, Y., et al., *Genomics of ecological adaptation in cactophilic Drosophila*. Genome Biol Evol, 2015. **7**(1): p. 349-66.
11. Haliloglu, G., et al., *Early-onset chronic axonal neuropathy, strokes, and hemolysis: inherited CD59 deficiency*. Neurology, 2015. **84**(12): p. 1220-4.
12. Hernando-Herraez, I., et al., *DNA Methylation: Insights into Human Evolution*. PLoS Genet, 2015. **11**(12): p. e1005661.
13. Hernando-Herraez, I., et al., *The interplay between DNA methylation and sequence divergence in recent human evolution*. Nucleic Acids Res, 2015. **43**(17): p. 8204-14.
14. Junier, I., et al., *On the demultiplexing of chromosome capture conformation data*. FEBS Lett, 2015. **589**(20 Pt A): p. 3005-13.
15. Kalapis, D., et al., *Evolution of Robustness to Protein Mistranslation by Accelerated Protein Turnover*. PLoS Biol, 2015. **13**(11): p. e1002291.
16. Kulis, M., et al., *Whole-genome fingerprint of the DNA methylome during human B cell differentiation*. Nat Genet, 2015. **47**(7): p. 746-56.
17. Lee, S.T., et al., *Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures*. Nucleic Acids Res, 2015. **43**(5): p. 2590-602.
18. Librado, P., et al., *Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments*. Proc Natl Acad Sci U S A, 2015. **112**(50): p. E6889-97.
19. Liu, F., et al., *Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up*. Hum Genet, 2015. **134**(8): p. 823-35.
20. Martinez-Jimenez, *F. and M.A. Marti-Renom, Ligand-target prediction by structural network biology using nAnnoLyze*. PLoS Comput Biol, 2015. **11**(3): p. e1004157.
21. Medina-Gomez, C., et al., *BMD Loci Contribute to Ethnic and Developmental Differences in Skeletal Fragility across Populations: Assessment of Evolutionary Selection Pressures*. Mol Biol Evol, 2015. **32**(11): p. 2961-72.
22. Mejlachowicz, D., et al., *Truncating Mutations of MAGEL2, a Gene within the Prader-Willi Locus, Are Responsible for Severe Arthrogryposis*. Am J Hum Genet, 2015. **97**(4): p. 616-20.
23. Metzger, J., et al., *Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses*. BMC Genomics, 2015. **16**: p. 764.
24. Morozumi, Y., et al., *Atad2 is a generalist facilitator of chromatin dynamics in embryonic stem cells*. J Mol Cell Biol, 2015.
25. Nam, K., et al., *Extreme selective sweeps independently targeted the X chromosomes of the great apes*. Proc Natl Acad Sci U S A, 2015. **112**(20): p. 6413-8
26. Navarro, J.M., et al., *Site- and allele-specific polycomb dysregulation in T-cell leukaemia*. Nat Commun, 2015. **6**: p. 6094.
27. Network and C. Pathway Analysis Subgroup of Psychiatric Genomics, Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nat Neurosci, 2015. **18**(2): p. 199-209.

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

publications

28. Philippakis, A.A., et al., *The Matchmaker Exchange: a platform for rare disease gene discovery.* Hum Mutat, 2015. **36**(10): p. 915-21.
29. Puente, X.S., et al., *Non-coding recurrent mutations in chronic lymphocytic leukaemia.* Nature, 2015. **526**(7574): p. 519-24.
30. Ramirez, O., et al., *Genome data from a sixteenth century pig illuminate modern breed relationships.* Heredity (Edinb), 2015. **114**(2): p. 175-84.
31. Rebollo-Lopez, M.J., et al., *Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery.* PLoS One, 2015. **10**(12): p. e0142293.
32. Rivas, M.A., et al., Human genomics. *Effect of predicted protein-truncating genetic variants on the human transcriptome.* Science, 2015. **348**(6235): p. 666-9.
33. Ruiz, S., et al., *Limiting replication stress during somatic cell reprogramming reduces genomic instability in induced pluripotent stem cells.* Nat Commun, 2015. **6**: p. 8036.
34. Ruiz-Orera, J., et al., *Origins of De Novo Genes in Human and Chimpanzee.* PLoS Genet, 2015. **11**(12): p. e1005721.
35. Sali, A., et al., *Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop.* Structure, 2015. **23**(7): p. 1156-67.
36. Sanchez-Mora, C., et al., *Case-control genome-wide association study of persistent attention-deficit hyperactivity disorder identifies FBXO33 as a novel susceptibility gene for the disorder.* Neuropsychopharmacology, 2015. **40**(4): p. 915-26.
37. Santpere, G., et al., *Analysis of Five Gene Sets in Chimpanzees Suggests Decoupling between the Action of Selection on Protein-Coding and on Noncoding Elements.* Genome Biol Evol, 2015. **7**(6): p. 1490-505.
38. Sanz-Pamplona, R., et al., *Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer.* Clin Cancer Res, 2015. **21**(20): p. 4709-18.
39. Segui, N., et al., *Exome sequencing identifies MUTYH mutations in a family with colorectal cancer and an atypical phenotype.* Gut, 2015. **64**(2): p. 355-6.
40. Serra, F., et al., *Restraint-based three-dimensional modeling of genomes and genomic domains.* FEBS Lett, 2015. **589**(20 Pt A): p. 2987-95.
41. Terol, J., et al., *Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation.* BMC Genomics, 2015. **16**: p. 69.
42. Trussart, M., et al., *Assessing the limits of restraint-based 3D modeling of genomes and genomic domains.* Nucleic Acids Res, 2015. **43**(7): p. 3465-77.
43. Vaque, J.P., et al., *Colorectal adenomas contain multiple somatic mutations that do not coincide with synchronous adenocarcinoma specimens.* PLoS One, 2015. **10**(3): p. e0119946.
44. Veeramah, K.R., et al., *Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach.* Genetics, 2015. **200**(1): p. 295-308.
45. Voss, K., et al., *Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II.* Transcription, 2015. **6**(5): p. 91-101.
46. Warren, W.C., et al., *The genome of the vervet (Chlorocebus aethiops sabaeus).* Genome Res, 2015. **25**(12): p. 1921-33.
47. Wollstein, A. and O. *Lao, Detecting individual ancestry in the human genome.* Investig Genet, 2015. **6**: p. 7.
48. Xue, Y., et al., *Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding.* Science, 2015. **348**(6231): p. 242-5.
49. Zacarias-Cabeza, J., et al., *Transcription-dependent generation of a specialized chromatin structure at the TCRbeta locus.* J Immunol, 2015. **194**(7): p. 3432-43.