STRUCTURAL COMPUTATIONAL BIOLOGY GROUP

Alfonso Valencia
Group Leader

Staff Scientists
Andrea Nicole Dölker, Vera Pancaldi, Tirso Pons, Daniel Rico (until July), Michael Tress

Post-Doctoral Fellows
Simone Marsili (until October), Miguel Vázquez (until September)

Graduate Students
Maria Rigau, Juan Rodríguez, Jon Sánchez

Technicians
David A. Juan (TS)*, Martin Krallinger (TS)*, Miguel Madrid (since November) (TS)*, Filipe N. Were (TS)*

Visiting Scientists
Dimitrios Morikis (University of California, Riverside, USA), Miguel Vazquez (Norwegian University of Science and Technology, Trondheim, Norway)

*Titulado Superior (Advanced Degree)

OVERVIEW

The main interest of our Group is the study of the molecular bases of cancer by bringing an evolutionary perspective to the study of the interplay between genomics and epigenomics in tumour progression.

Our research is largely carried out in the context of large-scale genome projects, in which we develop new computational methods for the study of genome-cancer relationships.

In this general scenario, the strategic goals of the Structural Computational Biology Group are to:

- → Develop new ideas, methods and software platforms for the extraction, integration and representation of cancer data, including the analysis of molecular, genomic, epigenomic and phenotypic information in collaboration with large-scale genome projects.
- → Include new technologies for data and text mining, together with Machine Learning methods, in our cancer genome analysis framework.
- → Analyse the function, structure and specific interactions of cancer- related proteins.

> **"This year the initial phase of two large scale projects was completed; i.e. the International Human Epigenome Consortium (iHEC) and the Pancancer Analysis of Whole Genomes (PCAWG). In both cases, we have contributed to the computational analysis, including the implementation of the data analysis infrastructures and the development of new analysis methods, as well as collaborating towards the interpretation of the biological results."**

RESEARCH HIGHLIGHTS

The Group has contributed to several community efforts in different areas:

- → Epigenomics with the BLUEPRINT EU flagship project, which is part of the iHEC consortium; the results from this work were published at the end of 2016.
- → Pancancer Analysis of Whole Genomes (PCAWG), global analysis of 2500 complete cancer genomes; these results will be published in 2017.
- → The BioCreative text mining challenge in chemical compounds resulted in a number of resources and publications that appeared throughout 2016.

We have introduced a new computational method for the prediction of pairs of residues in protein interfaces. This method can help in the analysis of cancer related mutations.

We have also introduced new methods for the analysis of epigenomes at the linear two dimensional level (chromatin states) and three dimensional level (chromatin structure in the nucleus).

The cancer genome analysis system

Our Group is deeply involved in the development of a computational framework for the analysis of human genomes with specific application to the analysis of cancer genomes. Over the years, this framework has been applied to a number of collaborative cancer projects, and it has been particularly instrumental in the CLL-ICGC project.

We have now moved on to a new phase in which the framework is used for the analysis of the large set of full cancer genomes of the Pancancer Analysis of Whole Genomes (PCAWG); it is one of the four frameworks for data organisation, analysis and exploration used by the consortium.

With regards to the future, given the characteristics of the framework in terms of its modular structure, capacity of integration of new methods in working pipelines, and ease of installation (e.g. adoption of docker and cloud technologies), we consider that it can be the seed of new developments in the overarching analysis of human disease genomes.

Protein structure prediction and cancer genomes

In the context of cancer genome analysis, and as part of the Pan Cancer global effort, we have developed a set of methods to analyse the consequences of mutations in the interface of proteins. The underlying logic is that cellular functions are governed by signals transmitted via protein interactions and protein complexes. In these interactions, the amino acids located in interacting surfaces determine the intensity of the interactions and, very importantly, the specificity of the interactions. The exquisite functioning of cellular systems between proteins depends critically on the pairing of the proteins with their correct partners, and the accuracy of the interactions depends on the correct formation of pairs of residues of the 2 proteins in the interface.

We have shown that cancer associated mutations tend to accumulate in the protein interfaces to the point that, with the information available, it is possible to say that cancer related mutations specifically target protein interfaces. Therefore, understanding the nature of protein-protein interactions is important for understanding the impact of cancer mutations.

We have developed a new methodology able to predict, with high accuracy, a small set of pairs of residues located in the interface of interacting human proteins. The new methodology, based on the study of the co-evolution of the corresponding protein families, does not require any information about the corresponding structures and it is applicable to many human protein complexes for which no other information is available. Furthermore, we have shown that the pairs of residues predicted to interact are very conserved in structural terms (they occupy the same position in space over the lengthy evolutionary time), which is indicative of their importance in the organisation of the corresponding interfaces.

Based on these results, we are now exploring the use of the newly developed computational methods as an alternative approach for the interpretation of the consequences of cancer related mutations.

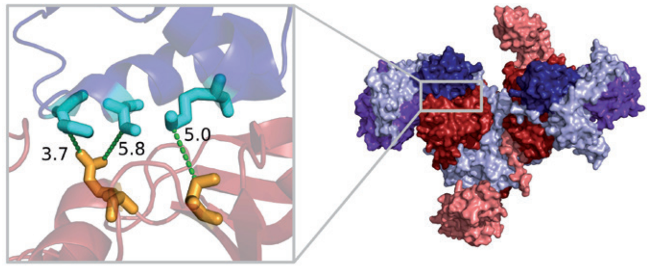

Figure 1 Co-evolution based correct prediction of pairs of residues in the interface of 2 domains of the human cytosolic phenylalanine tRNA synthetase (α subunit in dark red, B5 and B3/4 domains in β subunit in purple and dark blue, respectively); taken from Rodriguez-Rivas *et al.*, 2016.

EPIGENOME analysis infrastructure and portal

In the context of the BLUEPRINT iHEC project we have designed a system for the comparative analysis of epigenetics data (the BluePrint analysis portal http://blueprint-data.bsc.es/release_2016-08/, developed in collaboration with the BSC-CNS and EBI-EMBL). This portal is now the main point of access for the project's results (e.g. chromatin states, ChIP-Seq positions of histone modifications), enabling the direct comparison of the epigenetic structure of different cell types.

Based on the information provided by the Blueprint Analysis Portal, we have developed the methodology to compare epigenomes at the level of their organisation in functional segments (chromatin states). The initial results show that the system is not only able to reproduce the structure of the lineage differentiation during haematopoiesis, but also to detect what the main potential epigenetic driving factors of the differentiation are. The method, initially developed for the Blueprint data sets, is now being extended to other data types provided by the iHEC consortium.

Alternative splicing at the protein level

In 2016, we continued our work on alternative splicing in the context of the NIH-funded GENCODE project. Our results, summarised in a review published in *TIBS* (Tress *et al.*, 2016), show that in light of combined approaches, including protein modelling, proteomics and evolutionary analysis, there is little evidence to demonstrate that alternative isoforms are expressed at the protein level in detectable quantities. In other words, the only available evidence is that normal proteins are coded by the principal isoform of each gene and not by any of the potential alternative forms that are undoubtedly produced at the mRNA level. Even if this observation is in line with recent results of the large scale analysis of gene expression in human tissues (publications of the ENCODE/GTEx -www.gtexportal.org), it is still somewhat controversial since it indicates a big unexplained discrepancy between the results obtained at the level of gene and protein expression; a discrepancy that might have profound implications for our understanding of the role of mRNA in cells and the overall understanding of the biological function of processed RNAs.

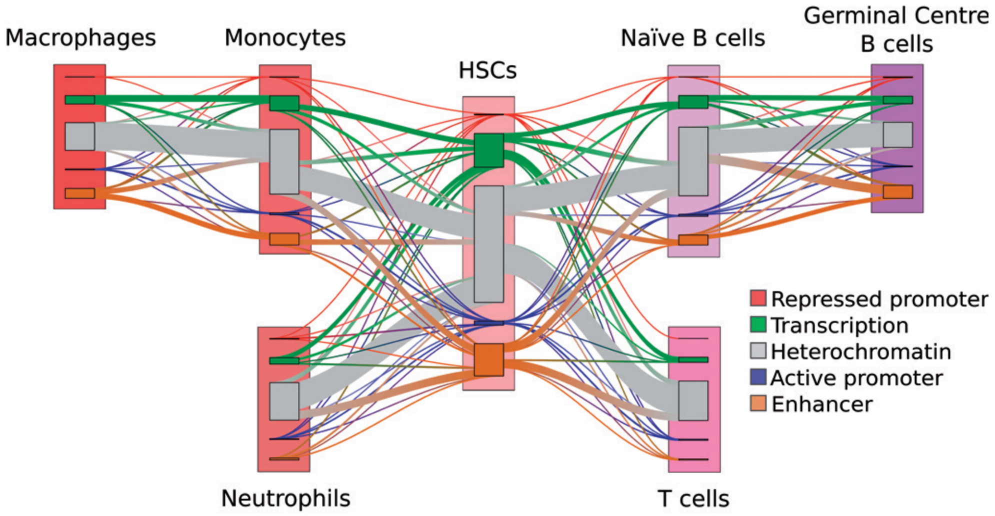

Figure 2 Representation of chromatin state transitions during haematopoietic differentiation. Boxes represent chromatin states and lines represent the observed transitions between cell types.

Biological Text Mining

Text mining, an important part of the Group's activity, has broad implications in Biomedicine. In 2016, we completed this year an exhaustive review of the application of text mining to the area of chemistry (Krallinger *et al.*, this work has been submitted to *Chem Rev*); this review was based on our experience in the analysis of text mining systems and the results in the context of the 2015 BioCreative Chemdner challenge (http://www.biocreative.org/tasks/biocreative-iv/chemdner/).

During 2016, we reached an agreement with the *Ministerio de Energía, Turismo y Agenda Digital* for the implementation of a biological text mining platform in the framework of the 'Plan de Impulso de las Tecnologías del Lenguaje'; this project is to develop tools and procedures in line with the recommendations of the European e-Infrastructure in text mining OpenMinted, in which we also participate. ■

PUBLICATIONS

- Astle WJ *et al.* (incl. Valencia A) (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167, 1415-1429.
- Chen L *et al.* (incl. Valencia A) (2016). Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167, 1398-1414.
- Stunnenberg HG; International Human Epigenome Consortium., Hirst M (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* 167, 1145-1149.
- Tress ML, Abascal F, Valencia A (2016). Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci*. PMID: 27712956.
- Rodriguez-Rivas J, Marsili S, Juan D, Valencia A (2016). Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone. *Proc Natl Acad Sci USA* 113, 15018-15023.
- Galindo-Albarrán AO *et al.* (incl. Valencia A) (2016). CD8+ T Cells from Human Neonates Are Biased toward an Innate Immune Response. *Cell Rep* 17, 2151-2160.
- Juan D, Perner J, Carrillo de Santa Pau E, Marsili S, Ochoa D, Chung HR, Vingron M, Rico D, Valencia A (2016). Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. *Cell Rep* 14, 1246-1257.
- Toll A, Fernández LC, Pons T, Groesser L, Sagrera A, Carrillo-de Santa Pau E, Vicente A, Baselga E, Vázquez M, Beltrán S, Pisano DG, Rueda D, Gut M, Pujol RM, Hafner C, Gut I, Valencia A, Real FX (2016). Somatic embryonic FGFR2 mutations in keratinocytic epidermal nevi. *J Invest Dermatol* 136, 1718-1721.
- Abascal F et al. (incl. Valencia A) (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol* 17, 251.
- Jiang Y et al. (incl. Valencia A) (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 17, 184.
- Pancaldi V, Carrillo-de-Santa-Pau E, Javierre BM, Juan D, Fraser P, Spivakov M, Valencia A, Rico D (2016). Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biol* 17, 152.
- Gurard-Levin ZA, Wilson LO, Pancaldi V, Postel-Vinay S, Sousa FG, Reyes C, Marangoni E, Gentien D, Valencia A, Pommier Y, Cottu P, Almouzni G (2016). Chromatin regulators as a guide for cancer treatment choice. *Mol Cancer Ther* 15, 1768-1777.
- Berger B, Gaasterland T, Lengauer T, Orengo C, Gaeta B, Markel S, Valencia A (2016). ISCB's Initial Reaction to The New England Journal of Medicine Editorial on Data Sharing. *PLoS Comput Biol* 12, e1004816. *Bioinformatics*. PMID:27153698. & *F1000Res 5*, pii: ISCB Comm J-157.
- Pons T, Vazquez M, Matey-Hernandez ML, Brunak S, Valencia A, Izarzugaza JM. (2016). KinMutRF: a random forest classifier of sequence variants in the human protein kinase superfamily. *Bmc Genomics* 17, 396.
- Pérez-Pérez M, Pérez-Rodríguez G, Rabal O, Vazquez M, Oyarzabal J, Fdez-Riverola F, Valencia A, Krallinger M, Lourenço A (2016). The Markyt visualisation, prediction and benchmark platform for chemical and gene entity recognition at BioCreative/CHEMDNER challenge. *Database* (Oxford), pii: baw120.
- Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, Cook C, Gasteiger E, Kim JH, Lopez R, Redaschi N, Stockinger H, Teixeira D, Valencia A (2016). Identifying ELIXIR Core Data Resources. *F1000Res* 5, pii: ELIXIR-2422.
- Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V, Fragkogianni S, Clarke L; BLUEPRINT Consortium., Flicek P, Rico D, Torrents D, Carrillo de Santa Pau E, Valencia A (2016). The BLUEPRINT Data Analysis Portal. *Cell Syst* 3, 491-495.