Life Sciences Workpackage

Consolider E-Science

Final meeting


INB
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
IRB
BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE


CRG
Centre de Regulació Genòmica
INB


INB
cnio
Centro Nacional de Investigaciones Oncológicas


More than 100 papers
20 organized meetings
>10 MEuros secure funding


A large number of intra-program collaborative projects

LETTER

Whole-genome sequencin
mutations in chronic lymp

Xose S. Puente[1], Magda Pinyol[2], Víctor Quesada[1], Laura Conde[3],
Pedro Jares[3], Sílvia Beà[3], Marcos González-Díaz[5], Laia Bassagany
Dolors Colomer[3], José M. C. Tubío[4,8], Cristina López[3], Alba Navar
Jesús M. Hernández[5], Diana A. Puente[1], José M. P. Freije[1], Gloria V
Sara Guijarro[3], Anna Enjuanes[3], Lluís Hernández[3], Jordi Yagüe[7],
Heinz Himmelbauer[10], Ester Castillo[10], Juliane C. Dohm[10], Silvia
Jesús San Miguel[5], Romina Royo[13], Josep L. Gelpí[13], David Torren
Roderic Guigó[15], Mónica Bayés[16], Simon Heath[16], Marta Gut[16], Pe

LETTERS

nature genetics

Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia

na C Queirós[4,11], Alba Navarro[1], Guillem Clot[1],
lle Brun-Heath[2], Magda Pinyol[6], Sergio Barberán-Soler[7],
Rico[8], Simone Ecker[8], Miriam Rubio[8], Romina Royo[9],
Conde[1], Mónica López-Guerra[1], Dolors Colomer[1],
ica Bayes[2], Marta Gut[2], Josep L Gelpí[9],
se S Puente[10], David G Pisano[8], Alfonso Valencia[8],
ín[10], Elías Campo[1] & José I Martín-Subero[4]

DNA methylation is a major mechanism in cell differentiation and neoplastic transformation[4–7]. The epigenetic modifications contribut-
ing to CLL development, however, are not fully understood[8–10]. Here

LETTERS

nature genetics

Exome sequencing identi
splicing factor *SF3B1* gen

Víctor Quesada[1], Laura Conde[2], Neus Villamo
Andrew J Ramsay[1], Sílvia Beà[2], Magda Pinyol[4], A
Alba Navarro[2], Tycho Baumann[5], Marta Ayme
Jesús M Hernández[6], Marcos González-Díaz[6],
José M C Tubío[3], Romina Royo[7], Josep L Gelpí
Miguel Vázquez[8], Alfonso Valencia[8], Heinz Hi
Ivo Gut[10], Xavier Estivill[3], Armando López-Gu

Here we perform whole-exome sequencing of sample
105 individuals with chronic lymphocytic leukemia (
the most frequent leukemia in adults in Western coun

OPEN ACCESS Freely available online

PLoS one

Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells

Sarah Djebali[1], Julien Lagarde[1], Philipp Kapranov[2¤a], Vincent Lacroix[1¤b], Christelle Borel[3], Jonathan M. Mudge[4], Cédric Howald[5], Sylvain Foissac[1,2¤c], Catherine Ucla[3], Jacqueline Chrast[5], Paolo Ribeca[1], David Martin[1], Ryan R. Murray[6], Xinping Yang[6], Lila Ghamsari[6], Chenwei Lin[6], Ian Bell[2], Erica Dumais[2], Jorg Drenkow[7], Michael L. Tress[8], Josep Lluís Gelpí[9], Modesto Orozco[9], Alfonso Valencia[8], Nynke L. van Berkum[10], Bryan R. Lajoie[10], Marc Vidal[6], John Stamatoyannopoulos[11], Philippe Batut[7], Alex Dobin[7], Jennifer Harrow[4], Tim Hubbard[4], Job Dekker[10], Adam Frankish[4], Kourosh Salehi-Ashtiani[6,12], Alexandre Reymond[5], Stylianos E. Antonarakis[3]*, Roderic Guigó[1,13]*, Thomas R. Gingeras[2,7]*

1 Bioinformatics and Genomics, Centre for Genomic Regulation and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain, **2** Affymetrix Inc., Santa Clara, California, United States of America, **3** Department of Genetic Medicine and Development, University of Geneva Medical School, University Hospitals of Geneva, Geneva, Switzerland, **4** Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom, **5** The Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland,

CSH PRESS GENOME RESEARCH

Resource

GENCODE: The reference human genome annotatic for The ENCODE Project

Jennifer Harrow,[1,9] Adam Frankish,[1] Jose M. Gonzalez,[1] Electra Tapanari,[1] Mark Diekhans,[2] Felix Kokocinski,[1] Bronwen L. Aken,[1] Daniel Barrell,[1] Amonida Zadissa,[1] Stephen Searle,[1] If Barnes,[1] Alexandra Bignell,[1] Veronika Boychenko,[1] Toby Hunt,[1] Mike Kay,[1] Gaurab Mukherjee,[1] Jeena Rajan,[1] Gloria Despacio-Reyes,[1] Gary Saunders,[1] Charles Steward,[1] Rachel Harte,[2] Michael Lin,[3] Cédric Howald,[4] Andre

Research

Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts

Milana Frenkel-Morgenstern,[1] Vincent Lacroix,[2] Iakes Ezkurdia,[1] Yishai Levin,[3] Alexandra Gabashvili,[3] Jaime Prilusky,[4] Angela del Pozo,[1] Michael Tress,[1] Rory Johnson,[5] Roderic Guigo,[5] and Alfonso Valencia[1,6]

Genome Research
www.genome.org

molecular systems biology

ARTICLE

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

artners

d Alfonso Valencia[1,*]

NIO), Madrid, Spain, [2] Structural Bioinformatics Group, Centre for
ercomputing Center, Barcelona, Spain, [4] Computational
ystems Biology Group, National Centre for Biotechnology

ore
er Research Centre (CNIO), C/Melchor Fernández Almagro 3,
ail: valencia@cnio.es

given proteome ('interactome') is the
ms Biology. Separately the prediction
is a well-established scientific area.
action partners. We provide a proof of
es representing known interactors in


,... and a large number of
inter-program
collaborative projects

BIOINFORMATICS
#1 IN MATHEMATICAL & COMPUTATIONAL BIOLOGY JOURNALS
IMPACT FACTOR: 5.468

Cell-Dock: high-performance protein–protein docking

Carles Pons[1,2], Daniel Jiménez-González[3,4,*], Cecilia
Harald Servat[3,4], Daniel Cabrera-Benítez[4], Xavier Agui
Juan Fernández-Recio[1,*]
+ Author Affiliations
*To whom correspondence should be addressed.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 15 2010, pages 1903–1904
doi:10.1093/bioinformatics/btq304

Sequence analysis

Cloud-Coffee: implementation of a parallel consistency-based multiple alignment algorithm in the T-Coffee package and its benchmarking on the Amazon Elastic-Cloud

Paolo Di Tommaso[1], Miquel Orobitg[2], Fernando Guirado[2], Fernado Cores[2], Toni Espinosa[3] and Cedric Notredame[1,*]

Barcelona,
pont, C. de
niversitad

Published online 9 May 2011 Nucleic

T-Coffee: a web server fo alignment of protein and structural information anc

Paolo Di Tommaso[1], Sebastien Moretti[2,3], Alberto Montanyola[4], Jia-Ming Chang[1], Jea

[1]Centre For Genomic Regulation (Pompeu Fabra Univ
Spain, [2]Vital-IT, Swiss Institute of Bioinformatics, Qua
Switzerland, [3]Department of Ecology and Evolution, B
Switzerland and [4]Department of Computer Science ar
Campus de Cappont, C. de Jaume II 69, E-25001 Lle

Received February 18, 2011; Revised March 23, 2011; Accepted April 5, 2011

Software

Highly accessed Open Access

A user-friendly web portal for T-Coffee on supercomputers

Josep Rius[1*], Fernando Cores[1], Francesc Solsona[1], Jano I van Hemert[2], Jos Koetsier[2] and Cedric Notredame[3]

* Corresponding author: Josep Rius jrius@diei.udl.cat

Author Affiliations

1 Department of Computer Science and Industrial Engineering, University of Lleida, C/Jaume II 69, E-25001 Lleida, Spain
2 UK National e-Science Centre, University of Edinburgh, 10 Crichton Street, EH8 9AB, Edinburgh, UK
3 Centre For Genomic Regulation (Pompeu Fabra University), C/Doctor Aiguader 88, 08003 Barcelona, Spain

For all author emails, please log on.

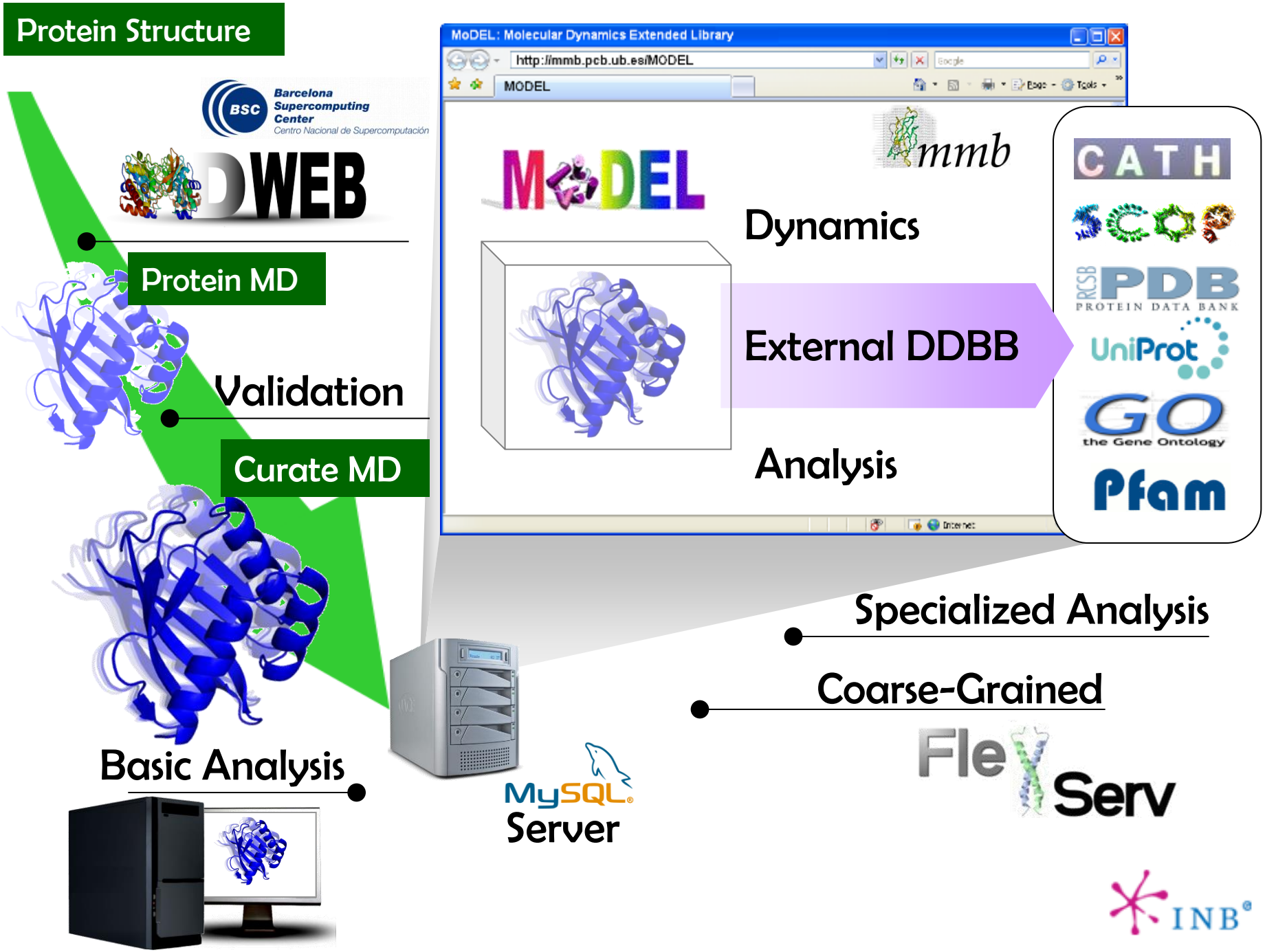
BMC Bioinformatics 2011, **12**:150 doi:10.1186/1471-2105-12-150

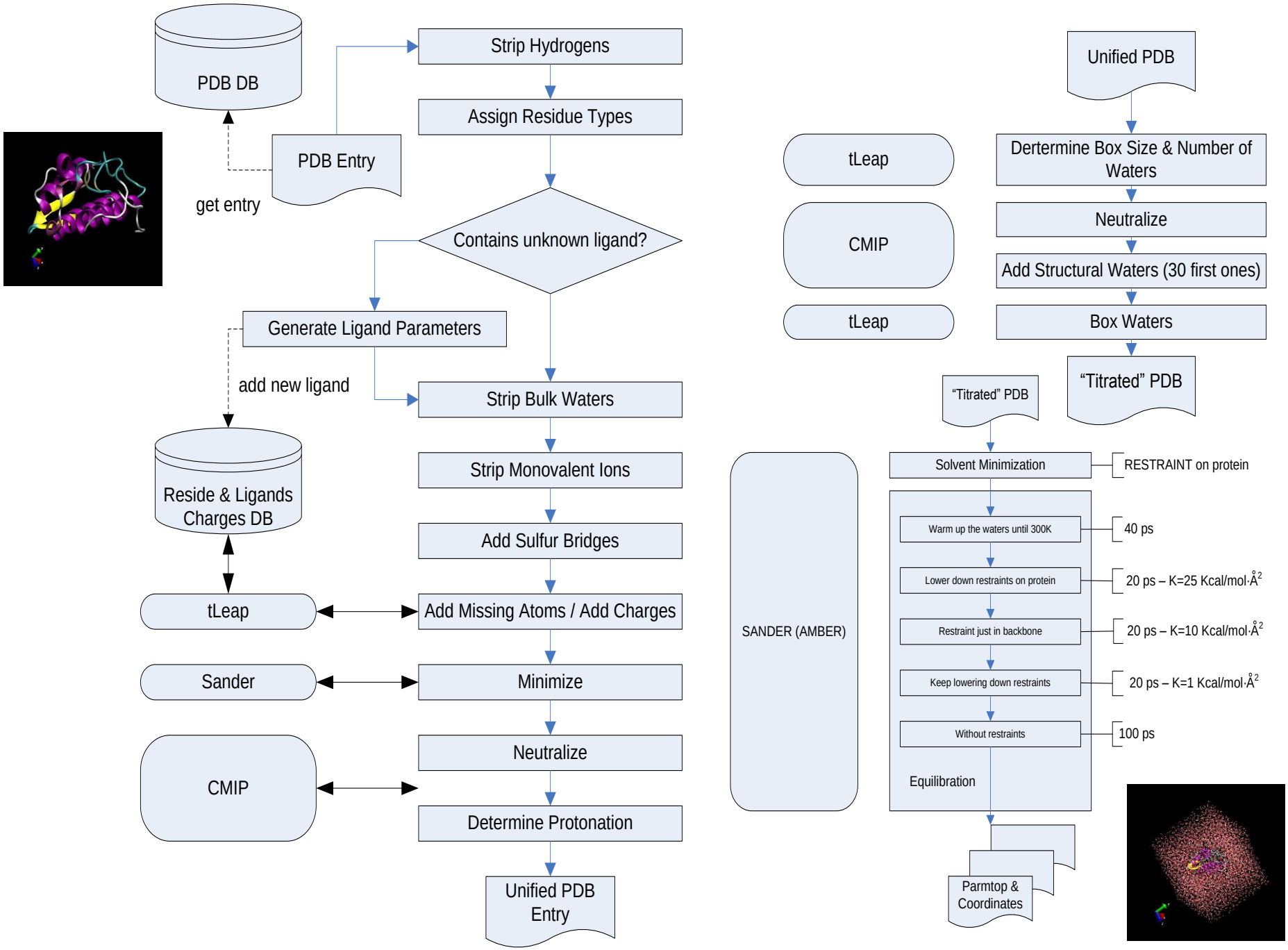
t me t
urnal
is jour
r of ev
Perm
date
rnal is

Main research lines

- Molecular Simulations
- Alternative Splicing
- Genomics from HiSeq
- Genetic basis of complex disseases

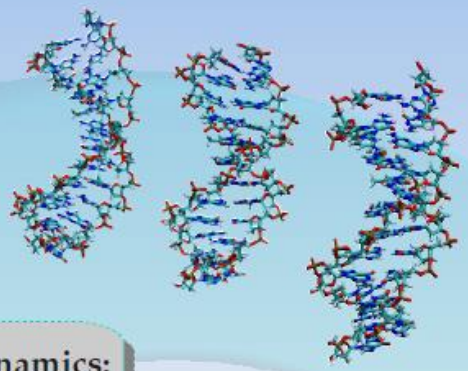
- *Software development*


Protein Structure
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
MDWEB
Protein MD
Validation
Curate MD
Basic Analysis
MoDEL: Molecular Dynamics Extended Library
http://mmb.pcb.ub.es/MODEL
MODEL
MODEL
mmb
Dynamics
External DDBB
Analysis
CATH
SCOP
RCSB PDB
PROTEIN DATA BANK
UniProt
GO
the Gene Ontology
Pfam
Specialized Analysis
Coarse-Grained
FlexServ
MySQL
Server
INB


PDB DB
PDB Entry
get entry
Strip Hydrogens
Assign Residue Types
Contains unknown ligand?
Generate Ligand Parameters
add new ligand
Strip Bulk Waters
Reside & Ligands Charges DB
Strip Monovalent Ions
Add Sulfur Bridges
tLeap
Add Missing Atoms / Add Charges
Sander
Minimize
CMIP
Neutralize
Determine Protonation
Unified PDB Entry
Unified PDB
tLeap
Dertermine Box Size & Number of Waters
CMIP
Neutralize
Add Structural Waters (30 first ones)
tLeap
Box Waters
"Titrated" PDB
"Titrated" PDB
SANDER (AMBER)
Solvent Minimization
RESTRAINT on protein
Warm up the waters until 300K
40 ps
Lower down restraints on protein
20 ps – K=25 Kcal/mol·Å2
Restraint just in backbone
20 ps – K=10 Kcal/mol·Å2
Keep lowering down restraints
20 ps – K=1 Kcal/mol·Å2
Without restraints
100 ps
Equilibration
Parmtop & Coordinates

NAFlex: *A web server for the study of Nucleic Acids Flexibility.*

Input Trajectory:
crd, dcd, netcdf, xtc, binpos



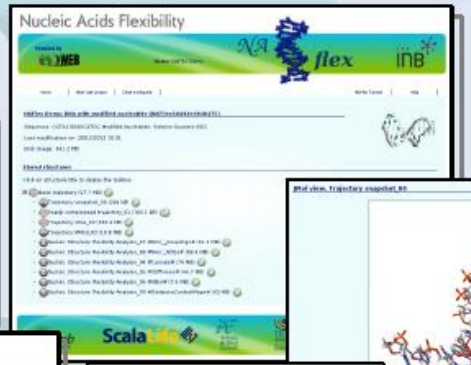
Analysis & Visualization

Cartesian Analysis
Helical Description
Principal Component Analysis
Stiffness Analysis
Energy Analysis
NMR Descriptors

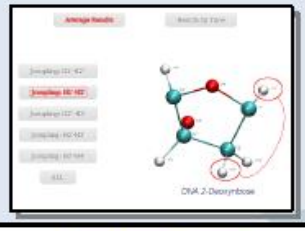
All-Atom
Molecular Dynamics

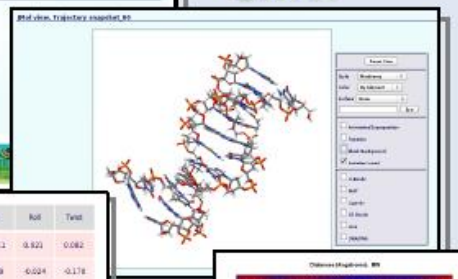
NAMD
AmberTools
GROMACS FAST. FLEXIBLE. FREE.

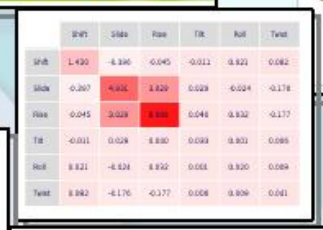
Coarse-Grained Dynamics:
Mesoscopic Elastic Model
Worm-Like Chain Model



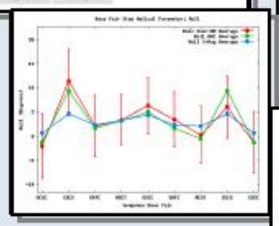
Jmol

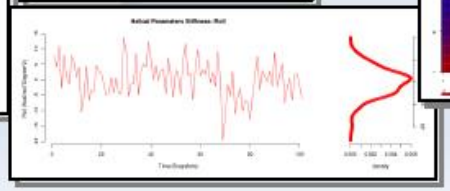


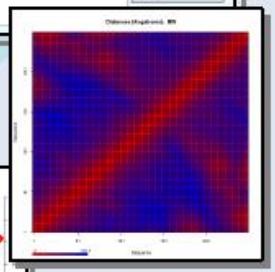




	Shift	Slide	Rise	Tilt	Roll	Twist
Shift	1.420	-0.196	-0.045	-0.011	0.021	0.082
Slide	-0.267	-0.031	0.029	0.029	-0.024	-0.178
Rise	-0.045	0.029	0.000	0.046	0.032	-0.177
Tilt	-0.011	0.029	0.030	0.093	0.001	0.086
Roll	0.021	-0.024	0.032	0.001	0.020	0.089
Twist	0.082	-0.178	-0.177	0.008	0.009	0.041

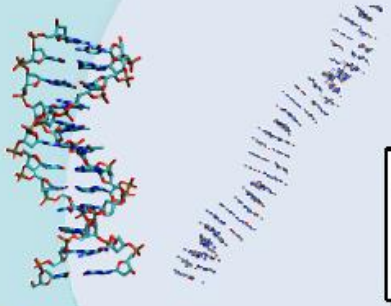






Dynamics Simulations

Input Structure:

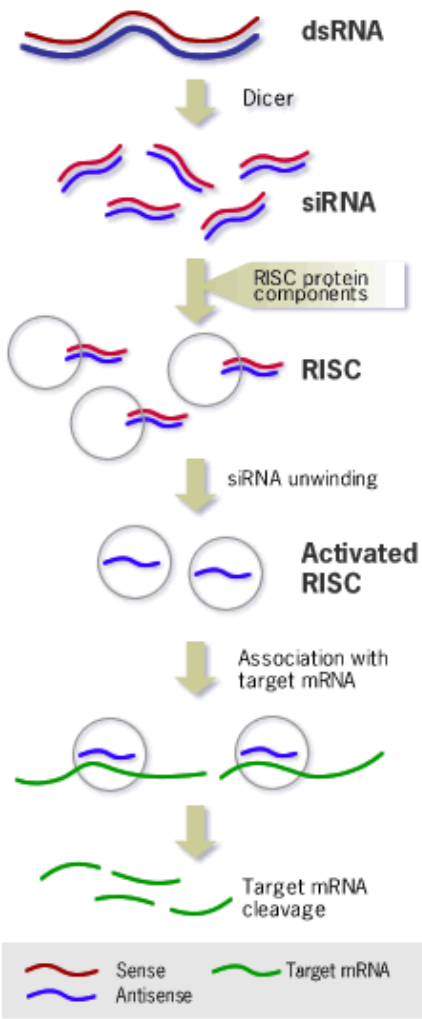


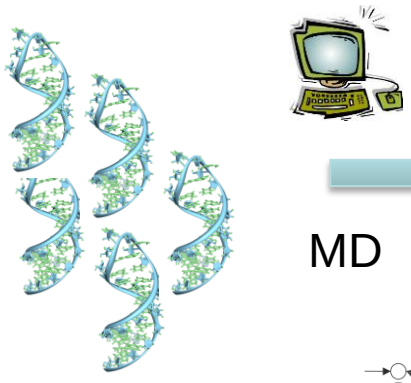
3D Structure Generation

B-DNA,A-DNA,A-RNA
User defined Nucleic Acids
NA From Helical Parameters
Coarse-Grained Models

Input Sequence:
GATTACCA

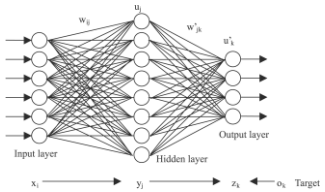
Predicting SiRNA efficiency



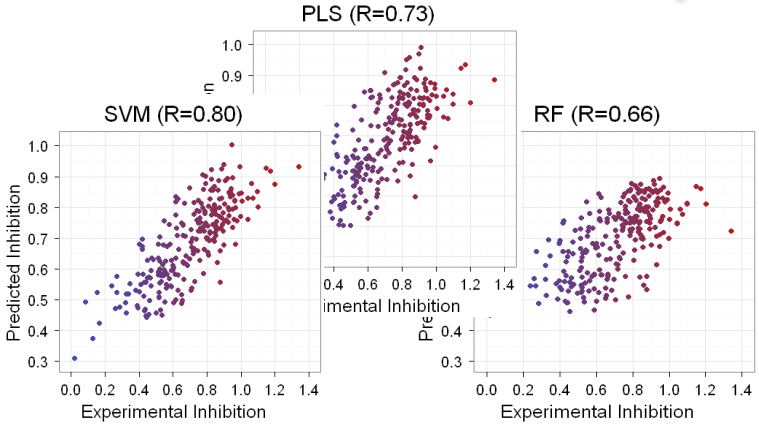


MD

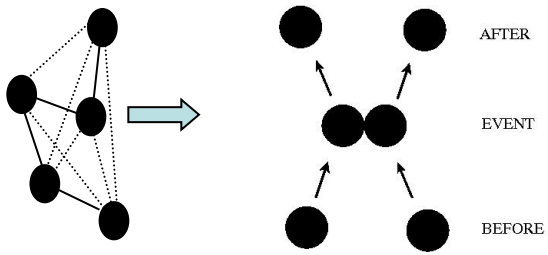
$\{RNA\}_i$, $\{X\}_i$



AI training
SVM, NN,
PLS,…



Validation



$$\vec{r}_i(t+t_c) = \vec{r}_i(t) + \vec{v}_i(t)t_c$$

Collision: $m_j\vec{v}_j + \Delta\vec{p} = m_j\vec{v}_j{}'$ $\quad m_i\vec{v}_i = m_i\vec{v}_i{}' + \Delta\vec{p}$

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{b_{ij}{}^2 - v_{ij}{}^2(r_{ij}{}^2 - d^2)}}{v_{ij}{}^2} \qquad \Delta p = \frac{2m_i m_j}{m_i + m_j}\left(v_i - v_j\right)$$

$$\vec{r}_i^{\,MD}$$

LT

$$\left\{\lambda_i^{MD}, \hat{e}_i^{MD}\right\}$$

this is our input to run the algorithm

similarity index, force constant, etc.

$$\left\{\lambda_i^{LD,BD}, \hat{e}_i^{LD,BD}\right\}$$

(ED)

$$\vec{r}_i^{\,LD,BD}$$

initial condition: all atoms in their equilibrium positions

$$r_i = r_i^0 + \tau\left(1 - e^{-\Delta t/\tau}\right)v_i^0 + \frac{\Delta t}{\tau}\left(1 - \frac{\tau}{\Delta t}\left(1 - e^{-\Delta t/\tau}\right)\right)F_i^0 + \Delta r_i^G$$

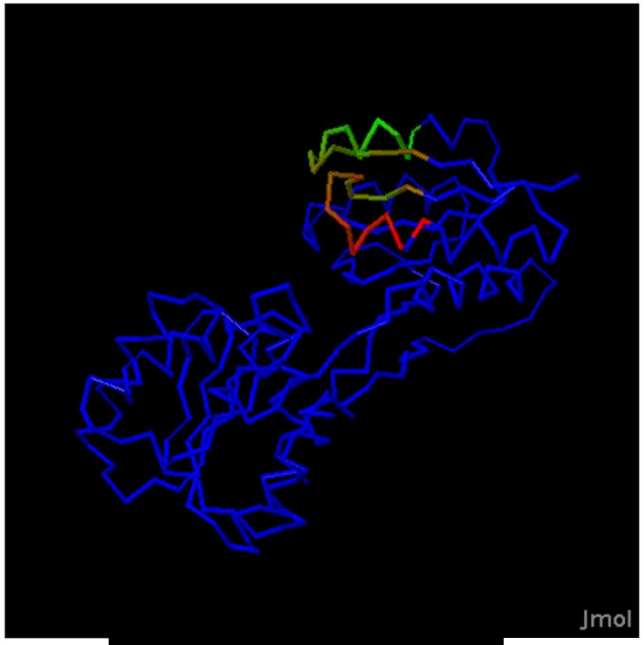
$$v_i = e^{-\Delta t/\tau}v_i^0 + \frac{1}{\gamma}\left(1 - e^{-\Delta t/\tau}\right)F_i^0 + \Delta v_i^G$$

Speed-up by a factor of 1000

1URP.pdb - NMA

Chained correlations

Residue: 50 Search width: 3 Search depth: 10 Threshold: 0.97 RMS type: Gaussian

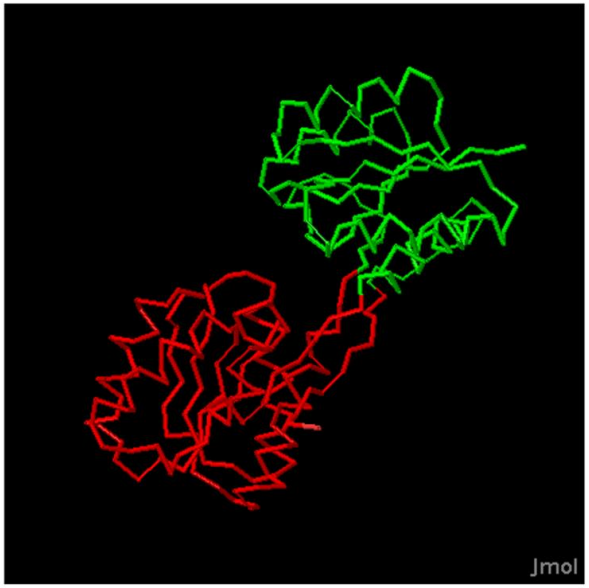


Residu

Depth	I
0	
1	
2	4
3	3
4	8,
5	9,
6	
7	
8	
9	1
10	

Dynamic domain detection method

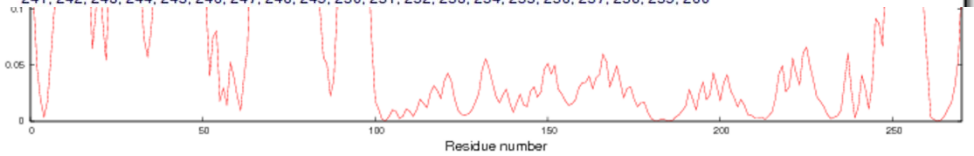
Movement by normal mode 1. Gaussian RMS applied



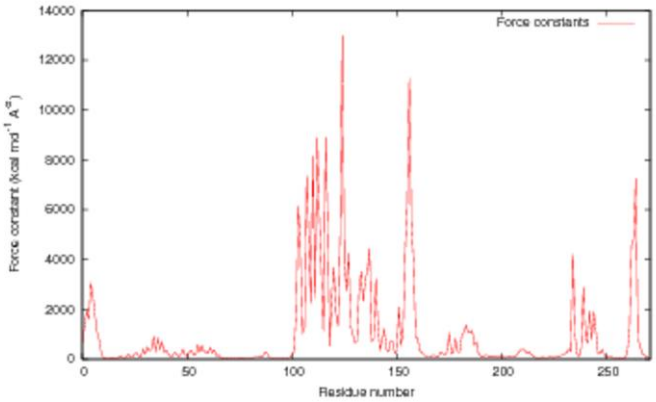
Residues in each cluster:

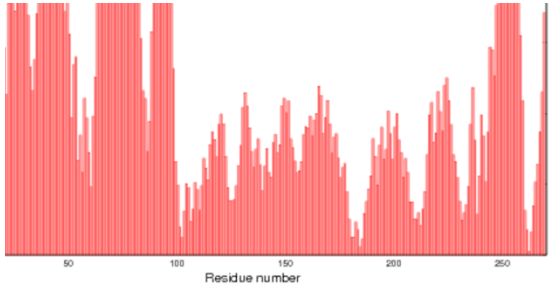
Cluster 1: 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270

Cluster 2: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260



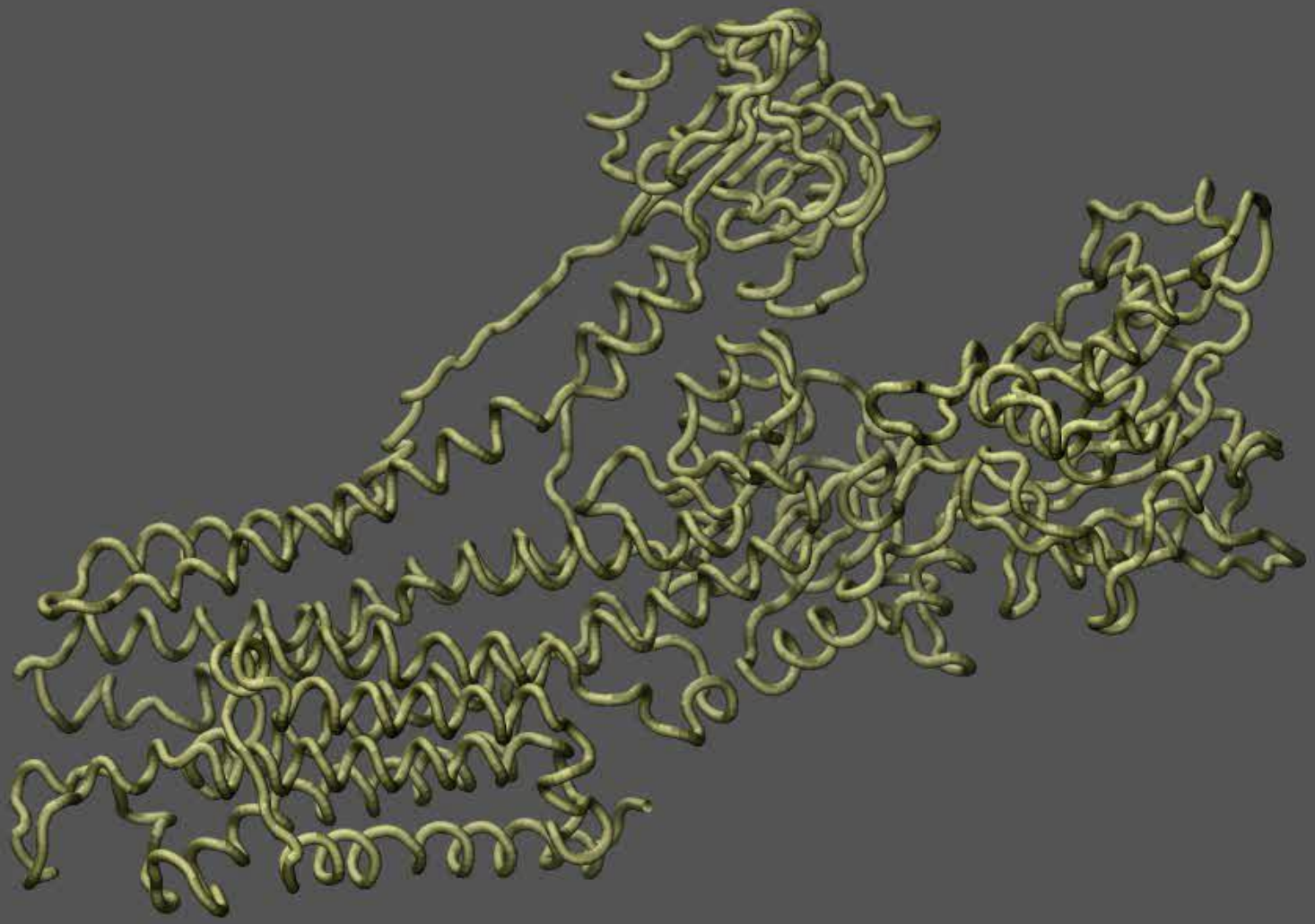
Hinge prediction

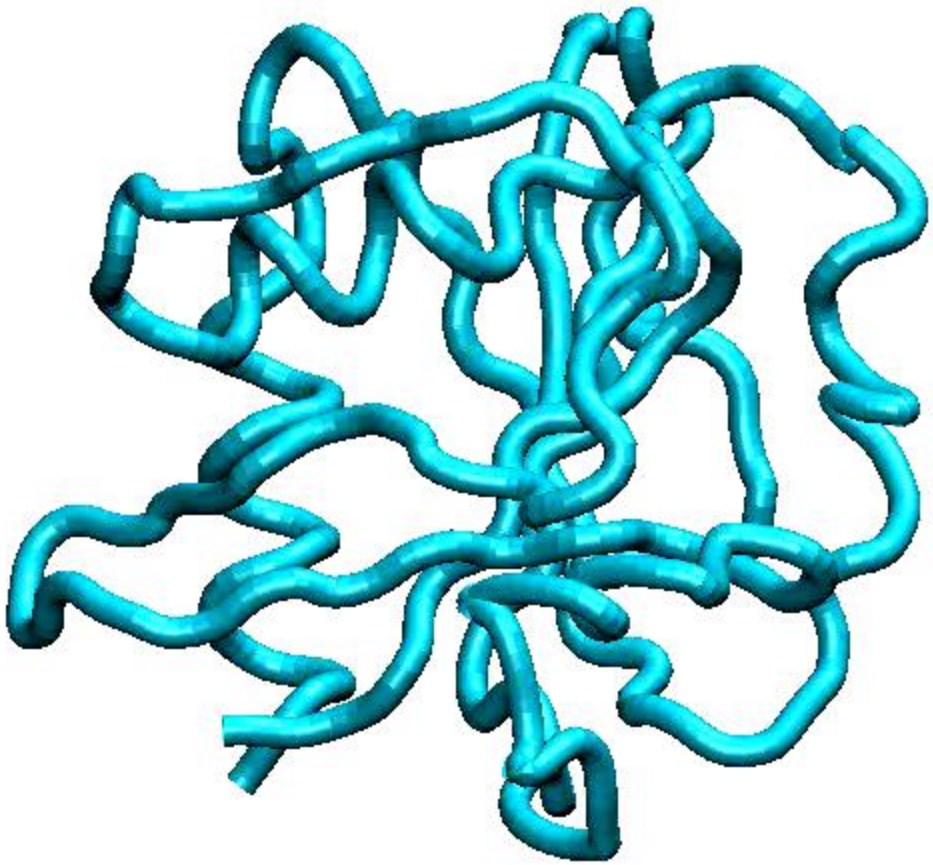


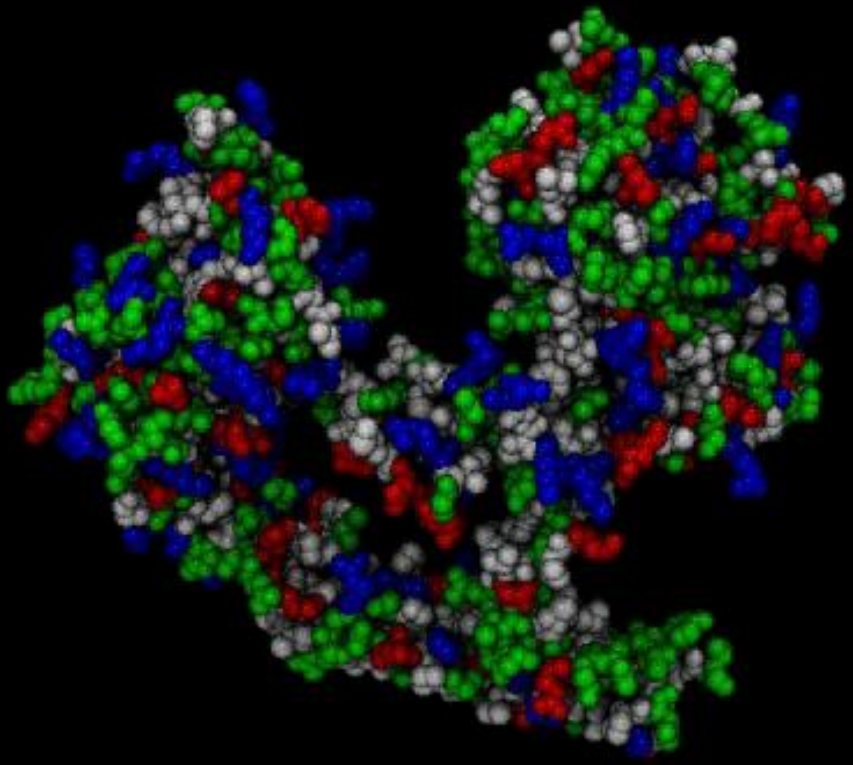


The peak constant corresponds to residue 125 with value 13002.110352

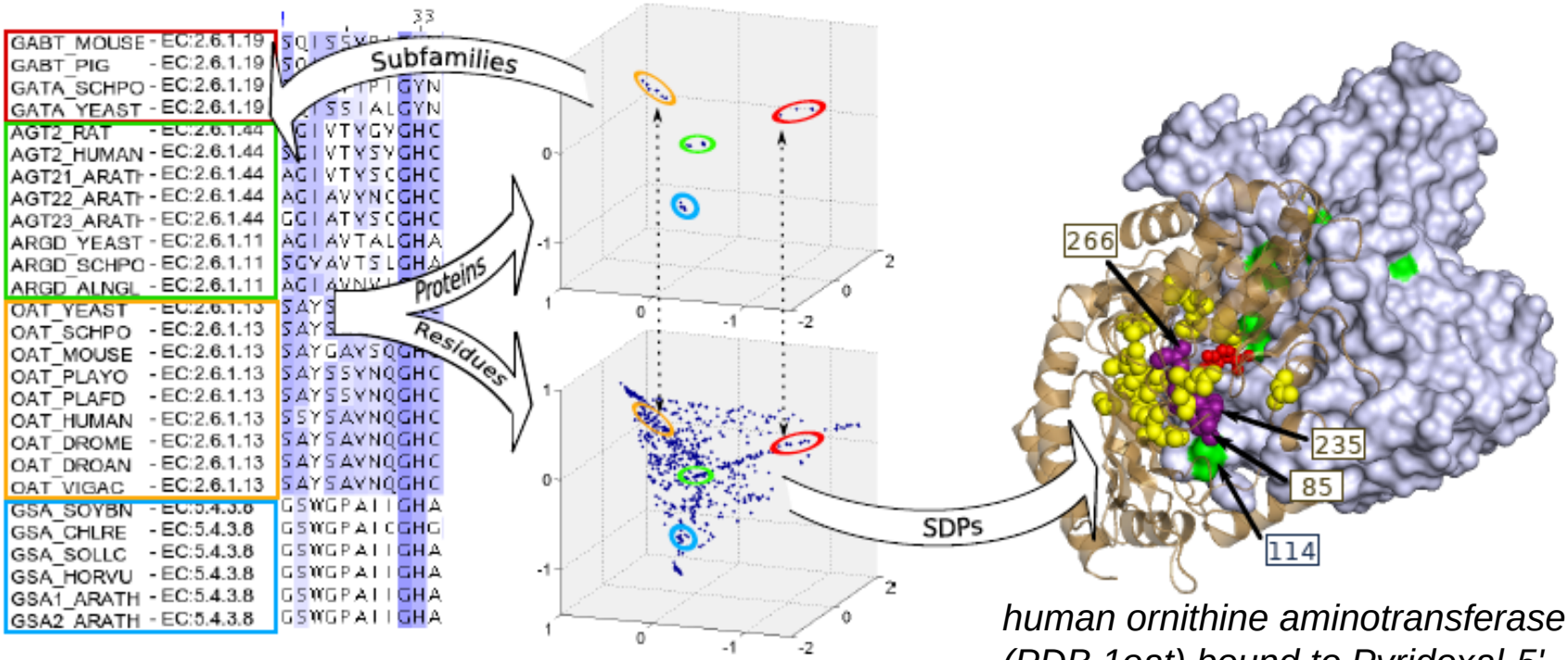
Residues in the top 20% are: 125, 157







Prediction of protein interaction sites from protein family sequence analysis



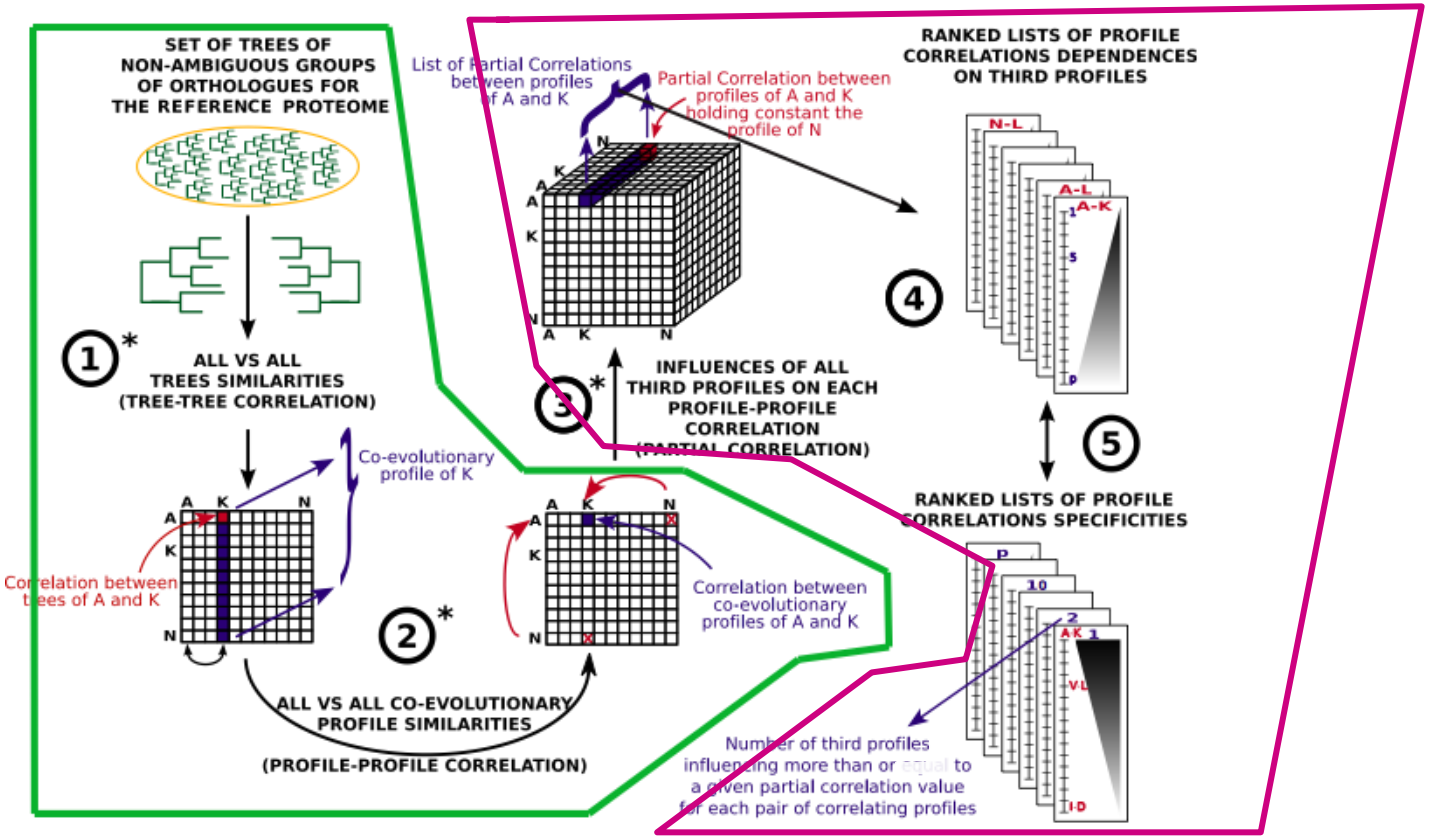
Correspondence analysis of multiple sequence alignments to identify sequence groups and associated significant residues (SDPs)

human ornithine aminotransferase (PDB 1oat) bound to Pyridoxal-5'-phosphate (red spheres). The two subunits of the complex are shown in brown cartoons and grey surface. Key predicted residues (SDPs) in yellow/violet spacefill and green surface respectively.

Predicting interaction networks

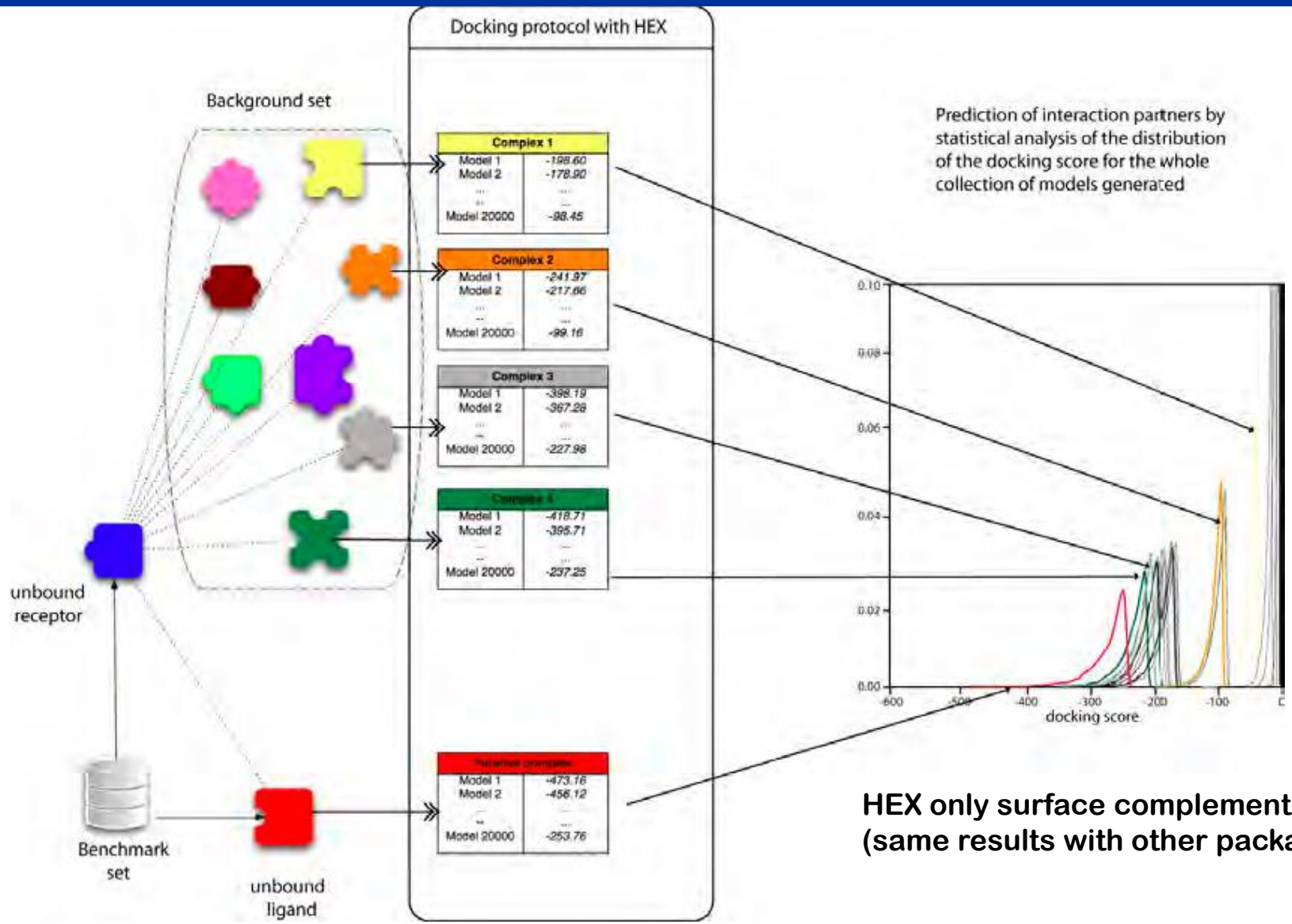
Detecting Evolutionary Similarities

Contrast function



* Correlations with p-values < 10E-5

HT docking: all against all protein docking. 56 unbound proteins against a benchmark of 633 proteins producing 20,000 models for each one of them (total of more than 700 M models)



HEX only surface complementarity (same results with other packages)

Alternative Splicing

In allowing for the generation of a diverse range of mature RNAs from the same gene, alternative splicing allows has considerable theoretical potential to expand the range of cellular proteins.

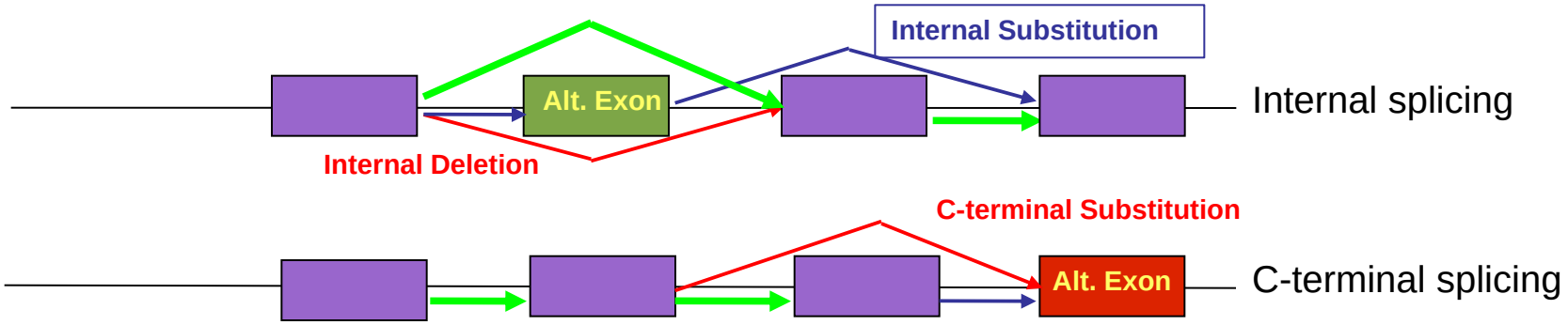
Recent studies estimate that **40–80%** of multi-exon human genes can produce differently spliced mRNAs.

There are many studies that implicate alternative transcripts in biological processes such as development.

This has lead to the hypothesis that **alternative splicing can explain the apparent lack of correlation between organism complexity and numbers of genes.**

In the past few years, it has become clear that a phenomenon called alternative splicing is <u>one reason human genomes can produce such complexity with so few genes</u>.*

*Science, July 2005

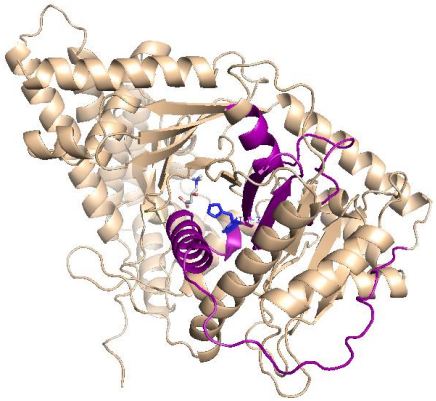


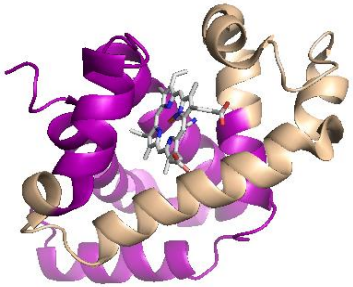
The implications of alternative splicing in the ENCODE protein complement

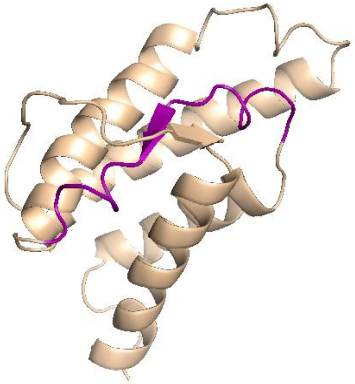
Michael L. Tress[a,b], Pier Luigi Martelli[c], Adam Frankish[d], Gabrielle A. Reeves[e], Jan Jaap Wesselink[a], Corin Yeats[f], Páll Isólfur Olason[g], Mario Albrecht[h], Hedi Hegyi[i], Alejandro Giorgetti, Domenico Raimondo[j], Julien Lagarde[k], Roman A. Laskowski[e], Gonzalo López[a], Michael I. Sadowski[l], James D. Watson[e], Piero Fariselli[c], Ivan Rossi[c], Alinda Nagy[i], Wang Kai[g], Zenia Størling[g], Massimiliano Orsini[m], Yassen Assenov[h], Hagen Blankenburg[h], Carola Huthmacher[h], Fidel Ramirez[h], Andreas Schlicker[h], France Denoued[k], Phil Jones[e], Samuel Kerrien[e], Sandra Orchard[e], Stylianos E. Antonarakis[n], Alexandre Reymond[o], Ewan Birney[e], Søren Brunak[g], Rita Casadio[c], Roderic Guigó[k,p], Jennifer Harrow[d], Henning Hermjakob[e], David T. Jones[l], Thomas Lengauer[h], Christine A. Orengo[f], László Patthy[i], Janet M. Thornton[e,q], Anna Tramontano[r], and Alfonso Valencia[a,s]

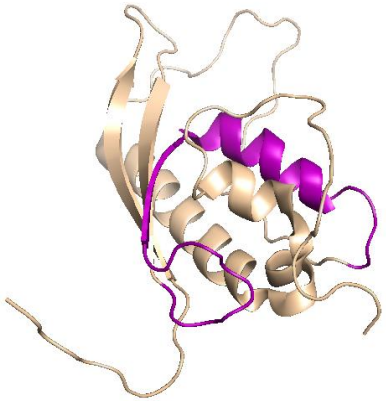
[a]Structural Computational Biology Programme, Spanish National Cancer Research Centre, E-28029 Madrid, Spain; [c]Department of Biology, University of Bologna, 33-40126 Bologna, Italy; [d]HAVANA Group, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [e]European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom; [f]Department of Biochemistry and Molecular Biology and [l]Bioinformatics Unit, University College London, London WC1E 6BT, United Kingdom; [g]Center for Biological Sequence Analysis, BioCentrum-DTU, DK-2800 Lyngby, Denmark; [h]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany; [i]Biological Research Center, Hungarian Academy of Sciences, 1113 Budapest, Hungary; [j]Department of Biochemical Sciences, University of Rome "La Sapienza," 2-00185 Rome, Italy; [k]Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica, E-8003 Barcelona, Spain; [m]Center for Advanced Studies, Research and Development in Sardinia (CRS4), 09010 Pula, Italy; [n]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; [o]Center for Integrative Genomics, Genopode
d [p]Centre de Regulació Genòmica, Universitat Pompeu Fabra, E-08003 Barcelona, Spain

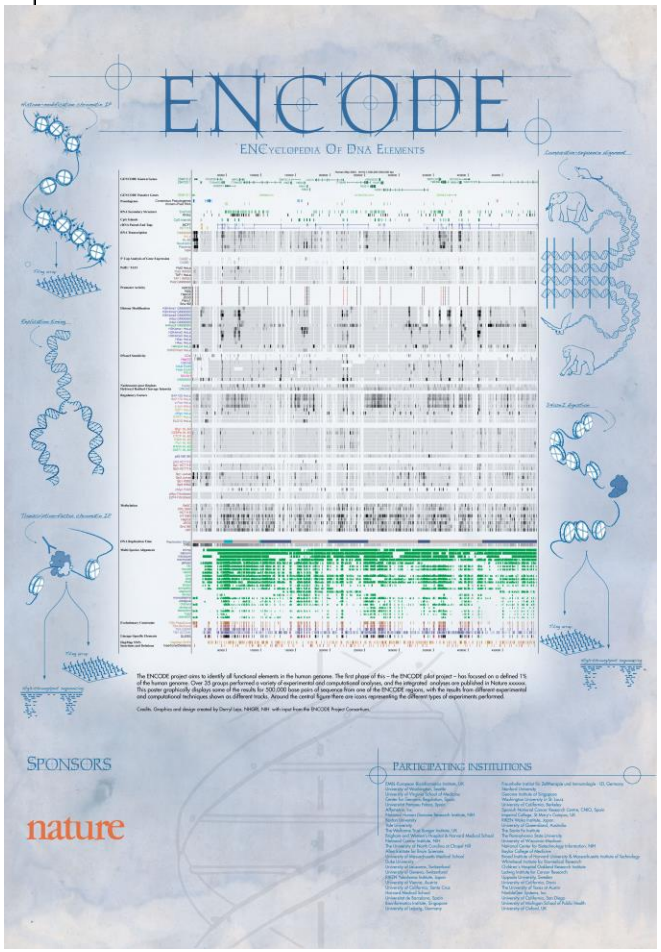
low October 20, 2006)







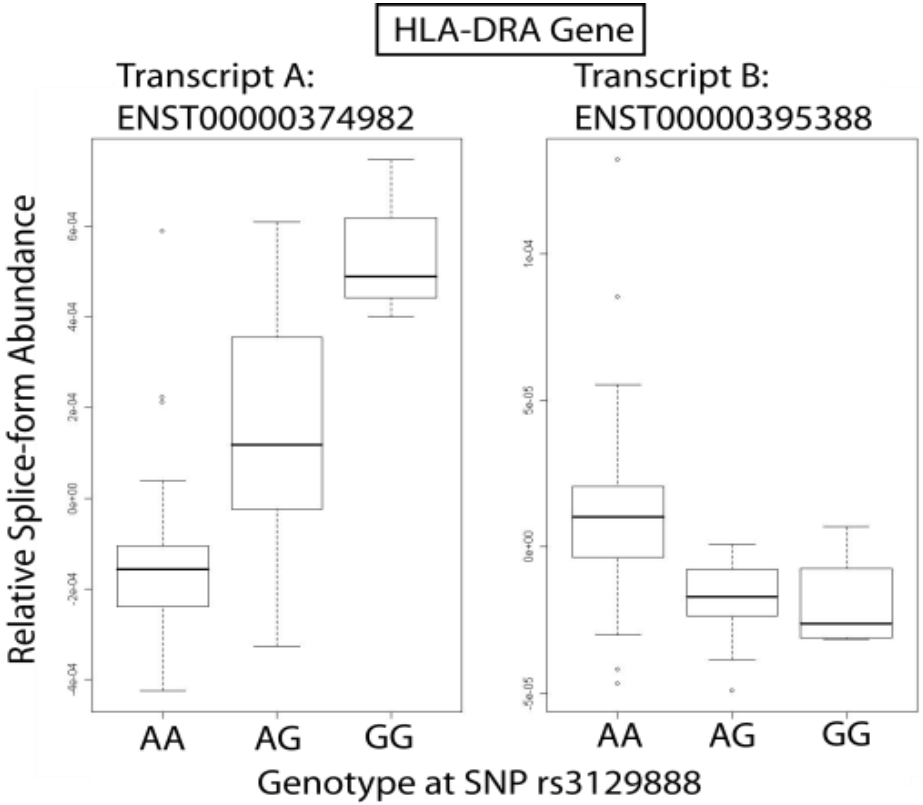




1 Transcriptome genetics using second generation sequencing in a Caucasian population

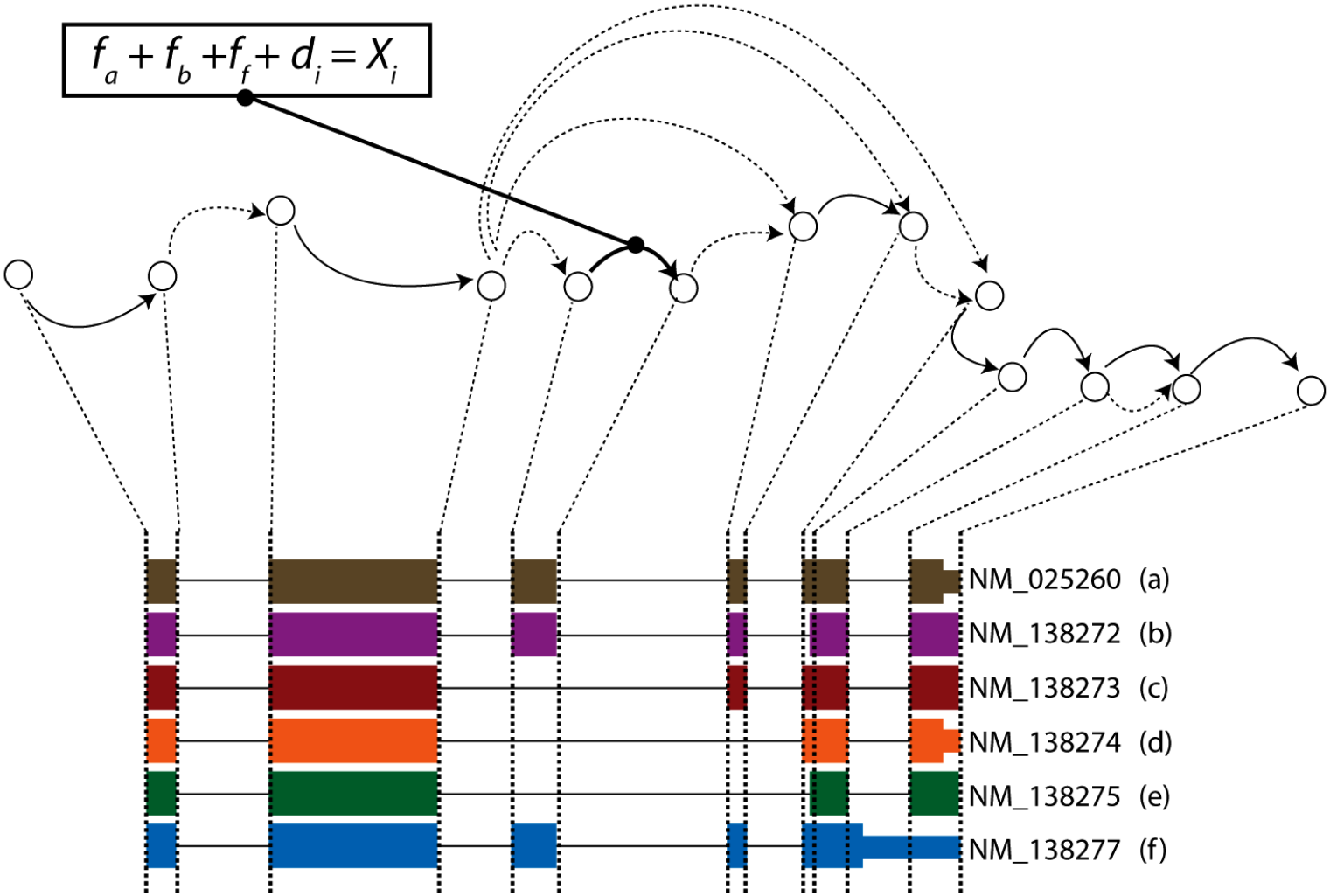
Stephen B. Montgomery[1,2*], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2*]

Using Flux, CRG group has been able to discover mutations with opposite effects on alternative transcript variants within the same gene



Montgomery et al. Nature. 464:773–777

The Flux Capacitor. Compute transcript abundance from RNASeq experiments



Bioinformatics Advance Access published February 3, 2013

BIOINFORMATICS ORIGINAL PAPER

2013, pages 1–8
doi:10.1093/bioinformatics/btt016

Gene expression

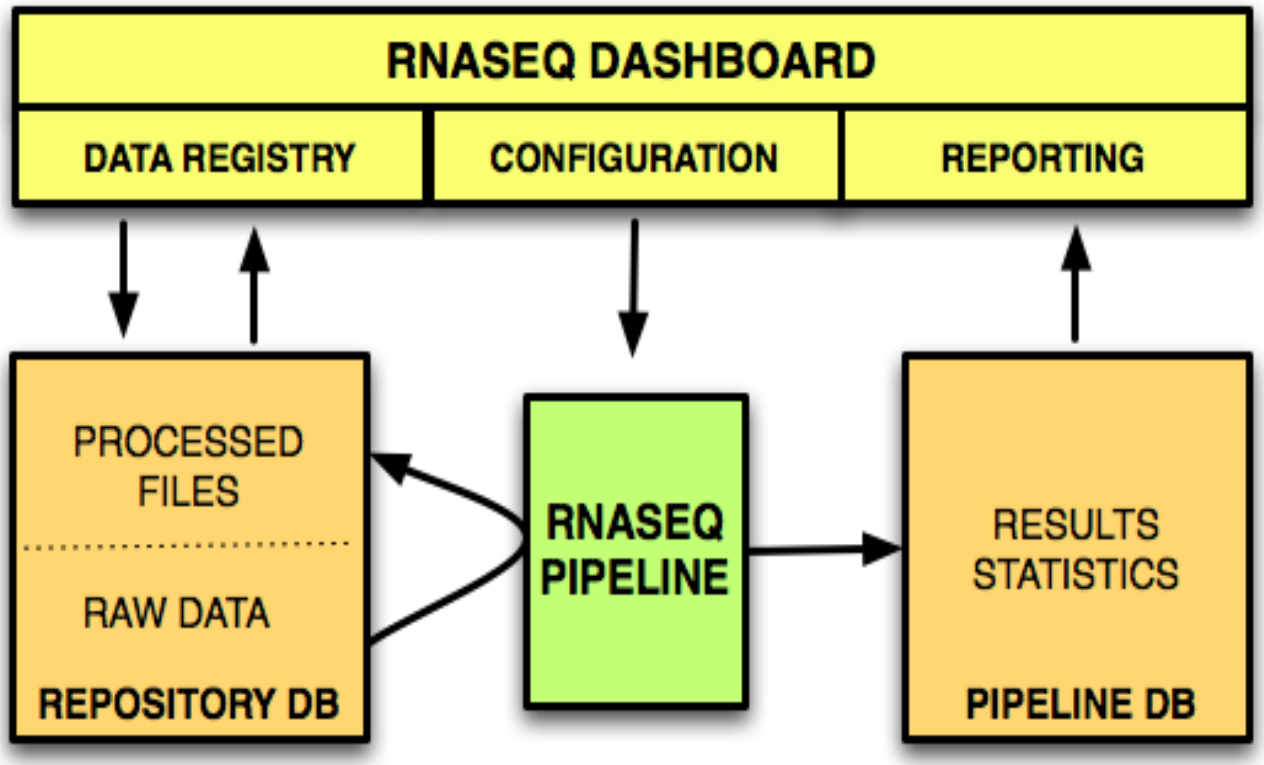
Advance Access publication January 17, 2013

Grape RNA-Seq analysis pipeline environment

David G. Knowles[1,2,*], Maik Röder[1,2], Angelika Merkel[1,2] and Roderic Guigó[1,2,*]

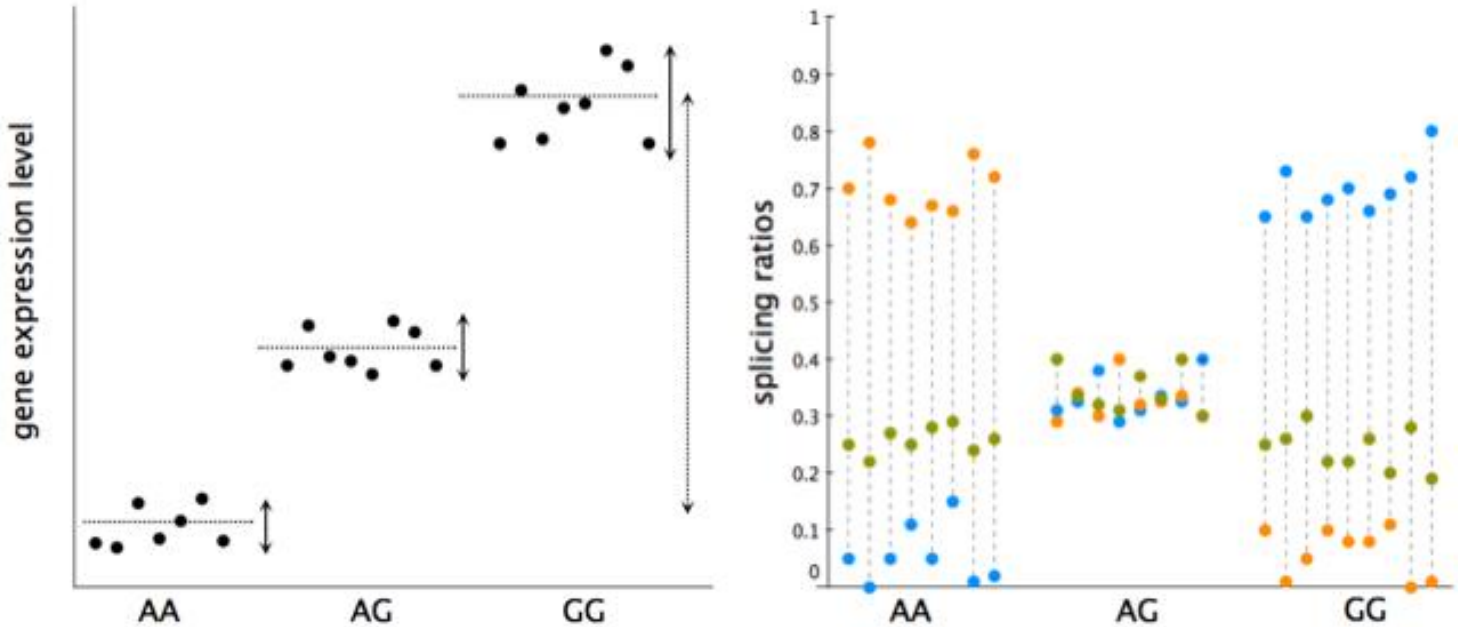
[1]Bioinformatics and Genomics Group, Centre for Genomic Regulation (CRG) and [2]Universitat Pompeu Fabra (UPF), Dr Aiguader 88, 08003 Barcelona, Spain

Associate Editor: Ivo Hofacker



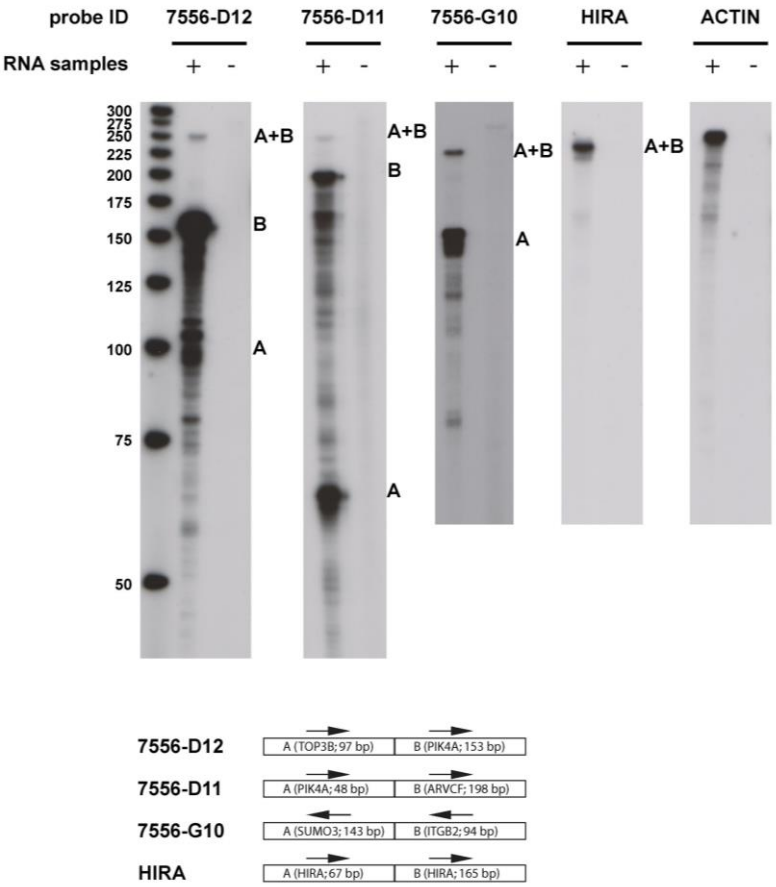
sQTLseekeR: a general method to detect genetic variants associated with alternative splicing in RNA sequencing population studies

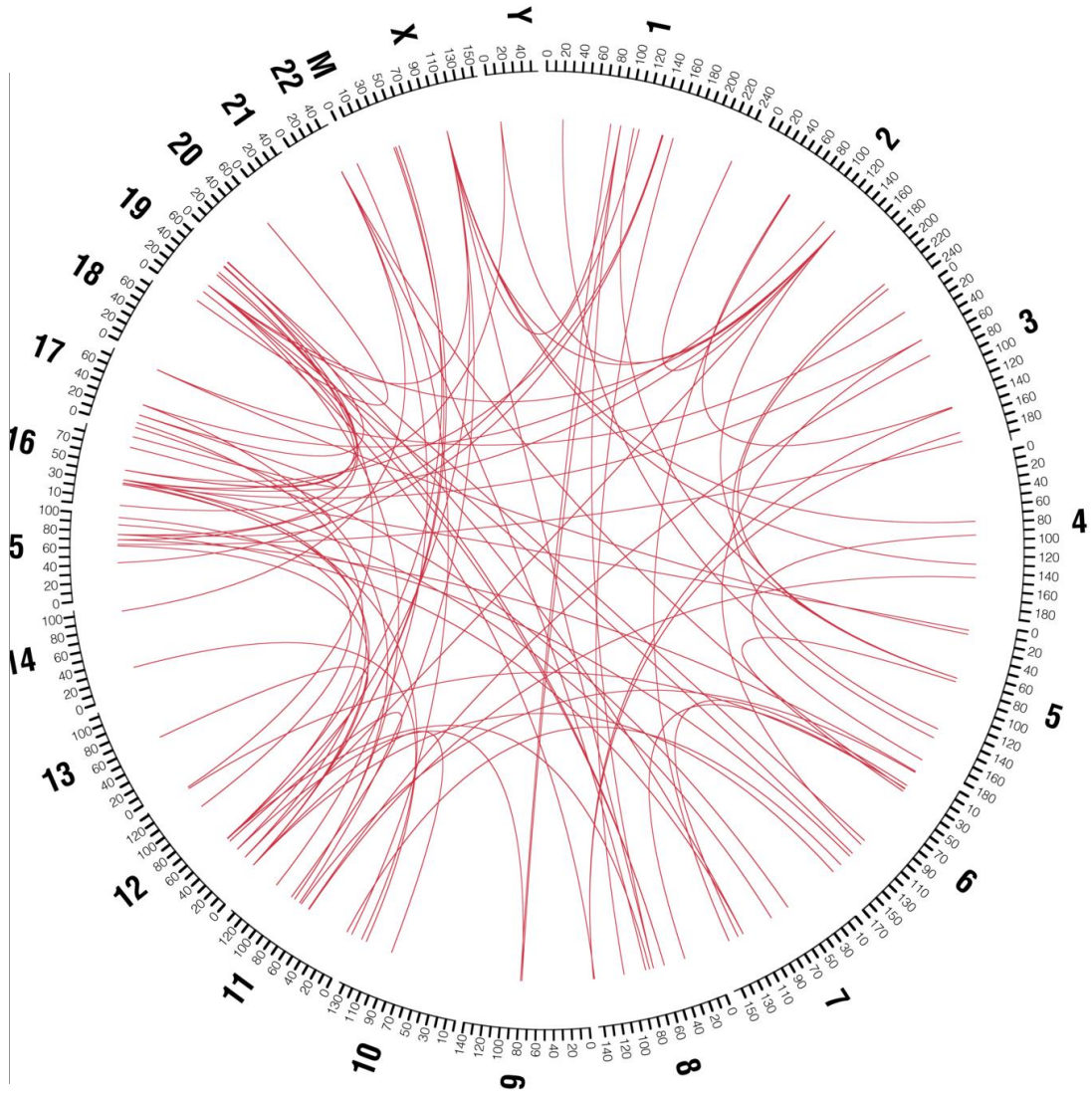
Jean Monlong[1,4], Miquel Calvo[3] and Roderic Guigó*[1,2]



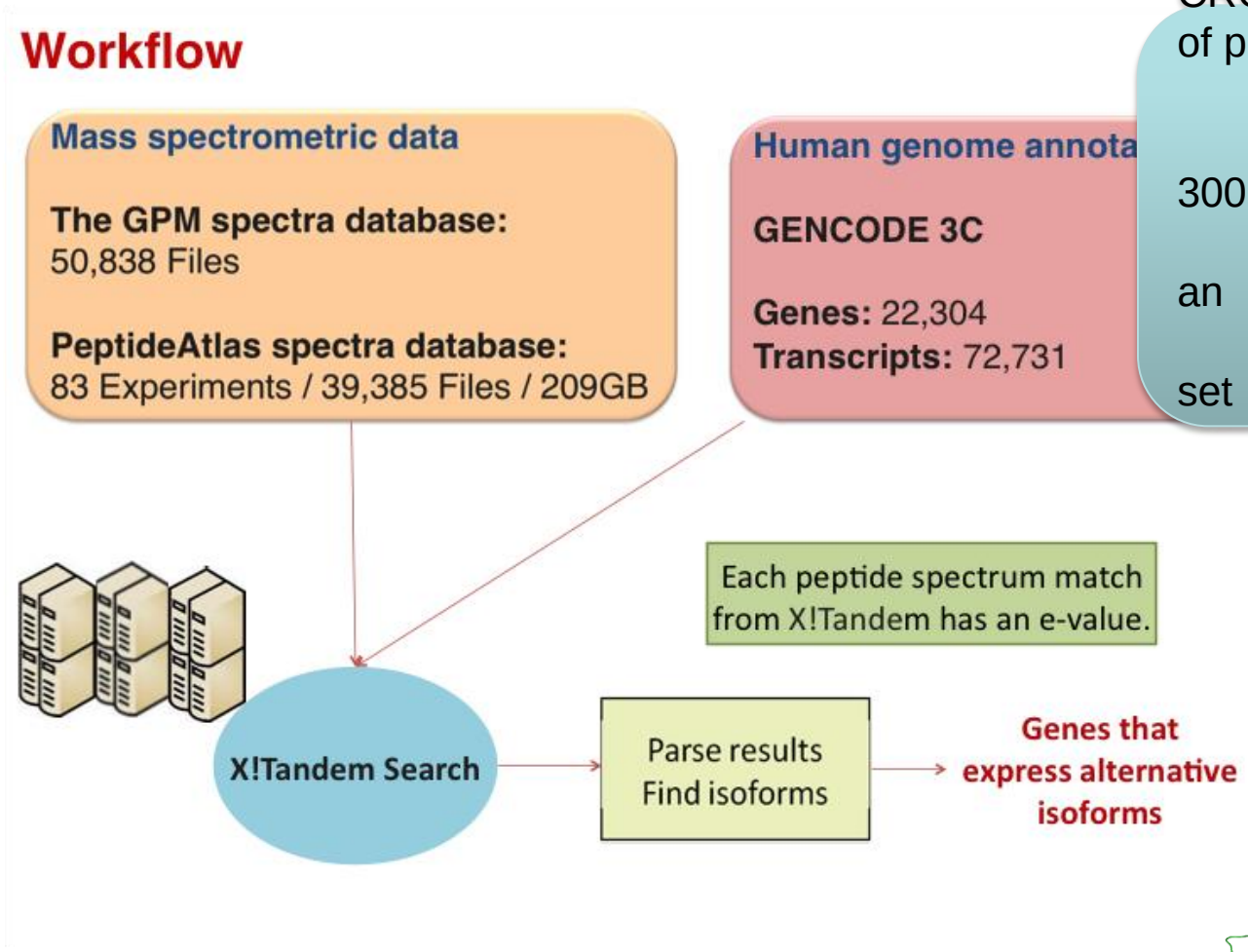
Inter-genic (trans) splicing

Experimental verification through RNAase protection assays





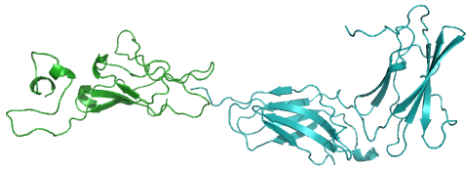
Trans-splicing and the human genome



CRG generated database of possible joined exons

Search Space: 300,370 joined peptides. plus an equivalent random set of c

We did detect a number of chimeric junction-spanning peptides, but the huge increase in search space size re-sized the spectra-peptide match e-values (which in turn influenced the combined p-values).



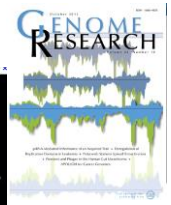
Human genome: 72h

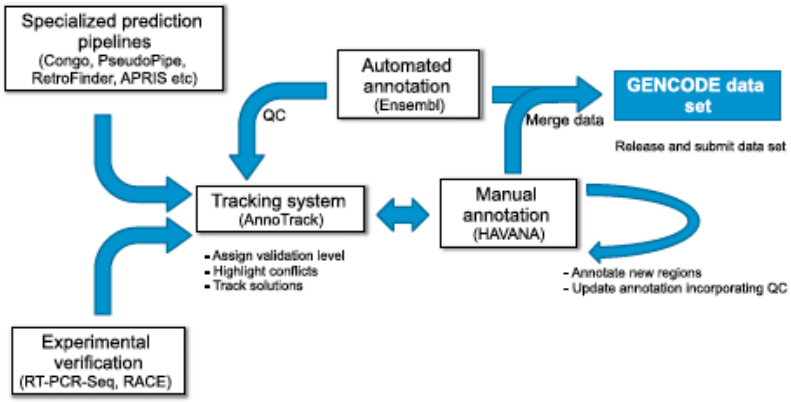
Possible exon joins: $190e^9$. A subset of 300.370 junctions (1/58 of the combinations for chromosome 21 and 22) 68h = **43000 days**

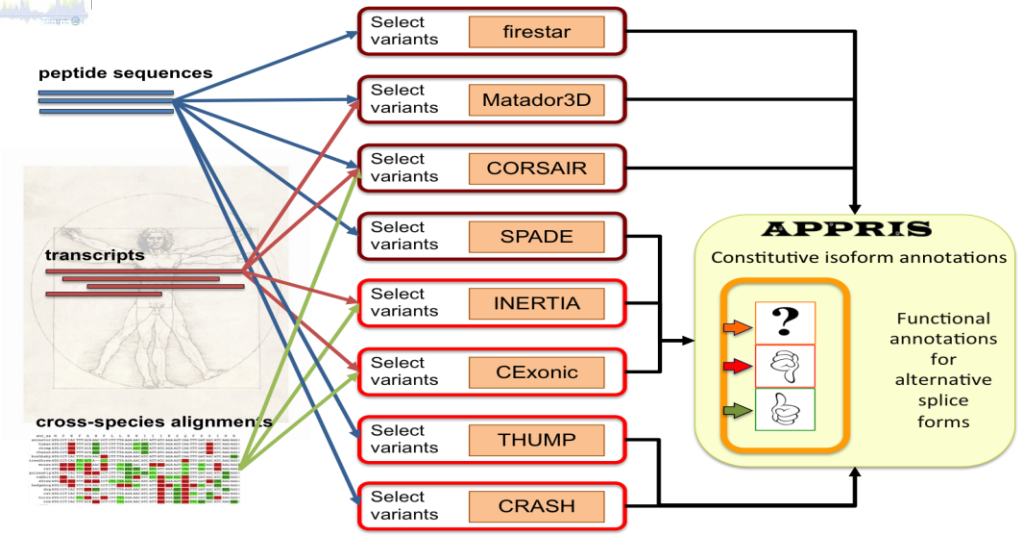
Canonical Splicing variants

APPRIS: ***analysis of functional annotations for alternative splice forms***
Rodriguez et al., ***NAR 2012***

GENCECODE/ENCODE:
Gen Res 2012, Nature 2012







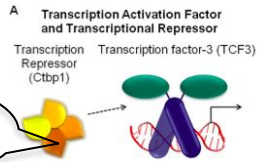
Transplicing/chimeric mRNAs

POST TRANSCRIPTIONAL REGULATION

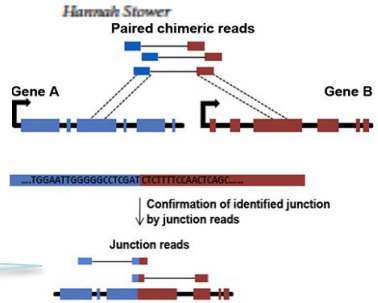
Chimeric protein production

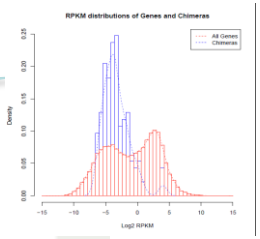
Chimeric transcripts arise from the joining of exons from two or more different genes. Reporting of such transcripts has become more widespread as data from genome-wide transcriptional analyses has increased. However, the number of known chimeric transcripts far outnumbers the reported number of chimeric proteins. Here, Frenkel-Morgenstern et al increase the number of identified translated chimeric transcripts and describe features indicative of their biological functionality.

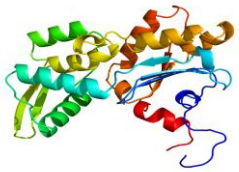
Firstly the authors analysed previously reported human tissue RNA-seq datasets for the presence of chimeric reads, that is, those that do not align to annotated transcripts and include a chimeric exon-exon junction. This approach confirmed the expression of 175 out of 7,424 previously reported chimeric transcripts from 16 human tissues. Analysis of the level of expression of these transcripts revealed that while chimeric transcripts are themselves expressed at a low level, they incorporate transcripts that are normally expressed at a high level; they are also expressed in a highly tissue-specific manner. To confirm the translation of these transcripts, the authors both searched mass spectrometry databases and generated their own shotgun mass spectrometry datasets. They initially searched for peptides that spanned a junction site of the human chimeric transcripts and they found 12 chimeric proteins. Following this,they generated targeted mass spectrometry data and this confirmed the expression of a further three novel chimeric proteins. The 175 expressed chimeras are enriched in signal and transmembrane peptides suggesting that generating chimeric transcripts is a mechanism for altering the localisation of the translated protein. Thus the authors' data suggests a potential function for chimeric proteins. The search is on to elucidate precise biological roles.

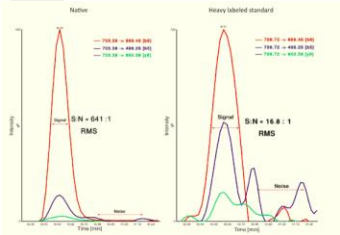


Chimeras may produce a dominant negative effect









Two overlapping mass-spec peptides, (18 experiments, Pvalue<10-7, FDR<1%)
Dr. Levin, Weizmann Institute

Nat Rev Genet

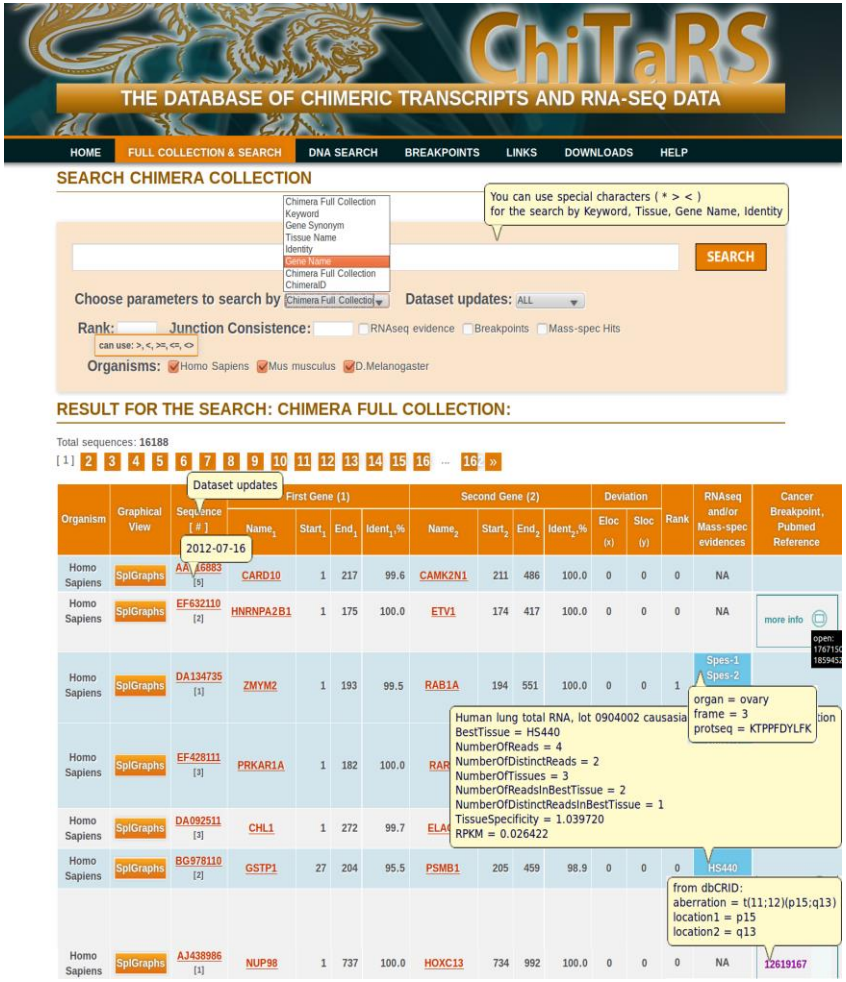
Research

Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts

Milana Frenkel-Morgenstern,[1] Vincent Lacroix,[2] Iakes Ezkurdia,[1] Yishai Levin,[3] Alexandra Gabashvili,[3] Jaime Prilusky,[4] Angela del Pozo,[1] Michael Tress,[1] Rory Johnson,[5] Roderic Guigo,[5] and Alfonso Valencia[1,6]

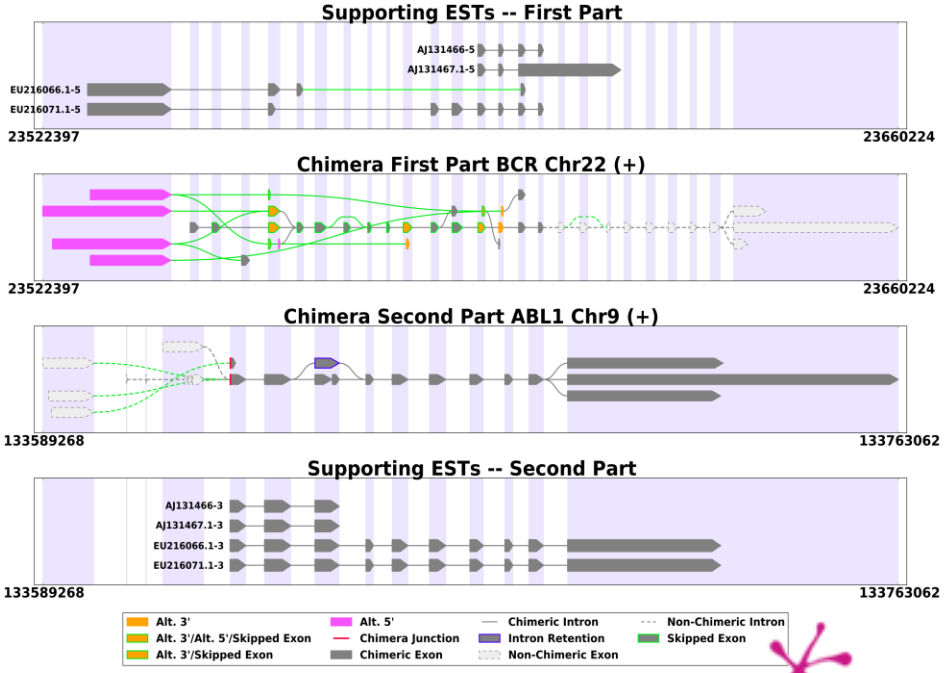
Genome Research
www.genome.org

ChiTaRS: chimeric transcripts/proteins Database

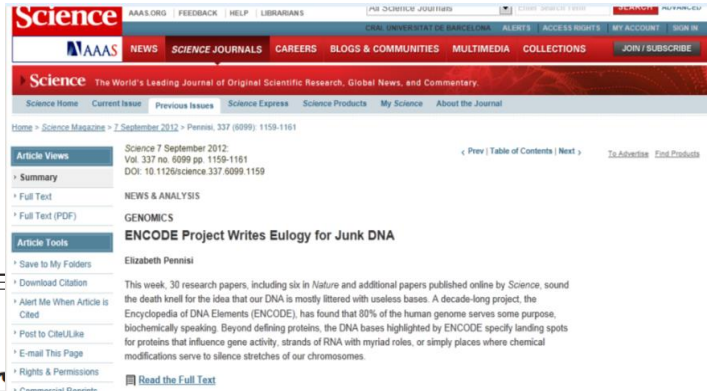


Morgenstern et al., NAR 2012

- Large scale identification
- RNAseq, ESTs, Literature, MS-
- Validation of junction sites
- Graphical output



INB



ARTICLE

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

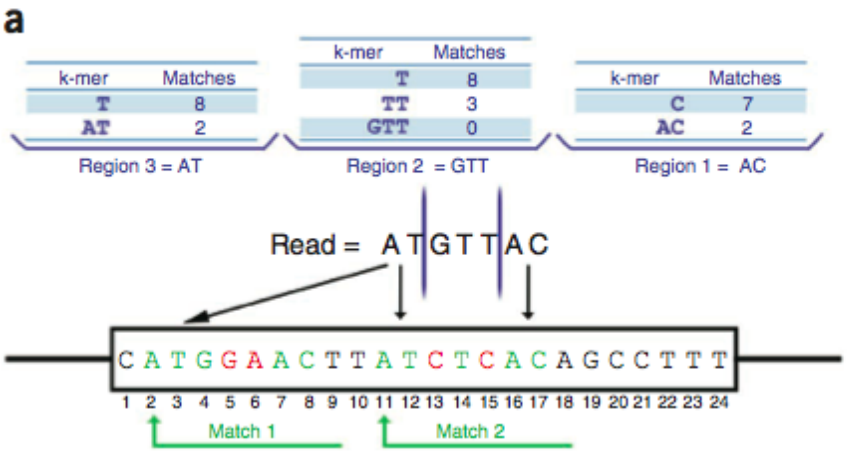
The GEM mapper: fast, accurate and versatile alignment by filtration

Santiago Marco-Sola[1], Michael Sammeth[1], Roderic Guigó[2] & Paolo Ribeca[1,3]

Because of ever-increasing throughput requirements of sequencing data, most existing short-read aligners have been designed to focus on speed at the expense of accuracy. The Genome Multitool (GEM) mapper can leverage string matching by filtration to search the alignment space more efficiently, simultaneously delivering precision (performing fully tunable exhaustive searches that return all existing matches, including gapped ones) and speed (being several times faster than comparable state-of-the-art tools).

In recent years, the superexponential increase in worldwide sequencing capacity[1] has driven a substantial amount of research into the development of efficient algorithms for the analysis of short-read sequence data. In particular, rapid alignment of reads alignment to the rest of the sequence; as few mismatches are usually allowed in the seed, this approach provides good performance but is also inflexible and incapable of returning all existing matches.

As the needs of the field are rapidly shifting, the assumptions described above are now constantly challenged. Some current biological problems (such as nonmodel-organism studies for which matching reference sequences might be incomplete, inaccurate or missing, or cross-species comparisons in evolutionary studies[2]) and new experimental protocols (data in color space, bisulfite-converted sequences or RNA sequencing[3]) require a higher tolerance to errors or more flexible alignment models[4,5]. Other applications (prediction of genomic variation or



GEM: crystal-clear DNA alignment

Gregory G Faust & Ira M Hall

The Genome Multitool (GEM) mapper rapidly and accurately provides all alignments of a read within a user-defined number of mismatches.

Over the past several years, the flood of data produced by second-generation DNA sequencers has motivated the development of a new generation of DNA alignment tools designed for mapping short reads to large reference genomes[1]. These aligners have allowed

Gregory G. Faust is in the Department of Computer Science, Ira M. Hall is at the Center for Public Health Genomics and both are in the Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, USA.
e-mail: irahall@virginia.edu

of heuristics can provide an excellent cost-performance trade-off and in general achieve high-quality results. Yet, it is inevitable that in certain situations, heuristic strategies will fail to identify biologically relevant alignments. The key advance of the GEM aligner is that it uses a strategy that is guaranteed to identify all alignments within a user-specified edit distance. Remarkably, GEM accomplishes this difficult task while also achieving a significant speed increase over other popular aligners.

GEM's impressive speed and accuracy derive from its use of a novel strategy for pruning the partitioning, each match is extended into a full alignment using dynamic programming. This latter step often dominates the run time of traditional 'seed-and-extend' aligners that rely on fixed-length seeds for initial matching. In contrast, GEM adaptively partitions queries into variably sized subsequences according to their uniqueness in the reference genome. This drastically reduces the total number of reference-genome loci requiring the computationally expensive extension step, thereby dramatically improving overall speed.

1. Li, H. & Homer, N. *Brief. Bioinform.* **11**, 473–483 (2010).
2. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. *Nat. Methods* **9**, 1185–1188 (2012).
3. Li, R. *et al. Bioinformatics* **25**, 1966–1967 (2009).
4. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
5. Langmead, B. & Salzberg, S.L. *Nat. Methods* **9**, 357–359 (2012).
6. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
7. Alkan, C. *et al. Nat. Genet.* **41**, 1061–1067 (2009).
8. Li, H. & Durbin, R. *Bioinformatics* **26**, 589–595 (2010).
9. Alkan, C., Coe, B.P. & Eichler, E.E. *Nat. Rev. Genet.* **12**, 363–376 (2011).

Genetic basis of complex disseases

Chronic Lymphocitic Leukemia

CLL-ICGC. Cancer Project

50 cancers

25000 cancer genomes

ICGC DATA PROCESSING WORKFLOW

Cancer and normal DNA from 500 CLL patients

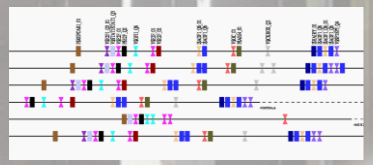
Centro Nacional de Análisis genómico (cnag)

Exome and whole-genome sequencing

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

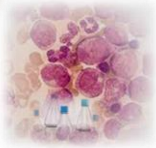
Data Storage, Management and processing

Identification of Cancer specific indels through whole-genome assembly



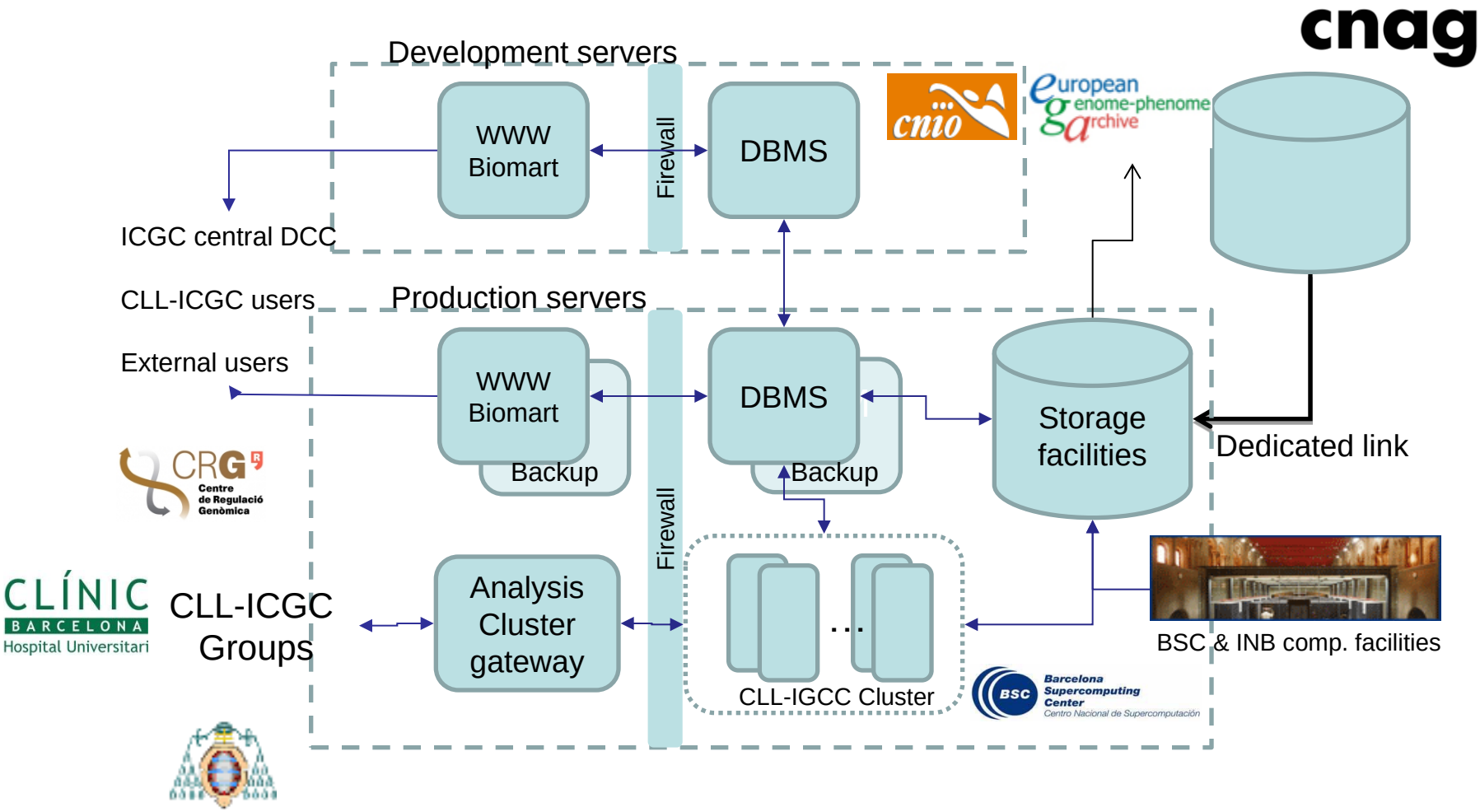
Results and raw data

International Cancer Genome Consortium



Experimental validation

CLL INFRASTRUCTURE SETUP



Col. S.Girona

Chronic Limphocytic Leukemia

Análisis de los datos de Hseq de **Chronic Lymphocytic Leukemia (CLL) Genome Project:** Objetivo entender las bases genéticas de la Leucemia

Secuenciación
Genoma

500 Paciente

BSC

Data Management
esperable 1.5 Pb

HPC Computing
Hasta 10% uso de MN

Validación
Experimental

Objectivo: Encontrar la bases genéticas de la Leucemia

Nature 2011; Nature Genetics 2012, Nature Genetics 2012

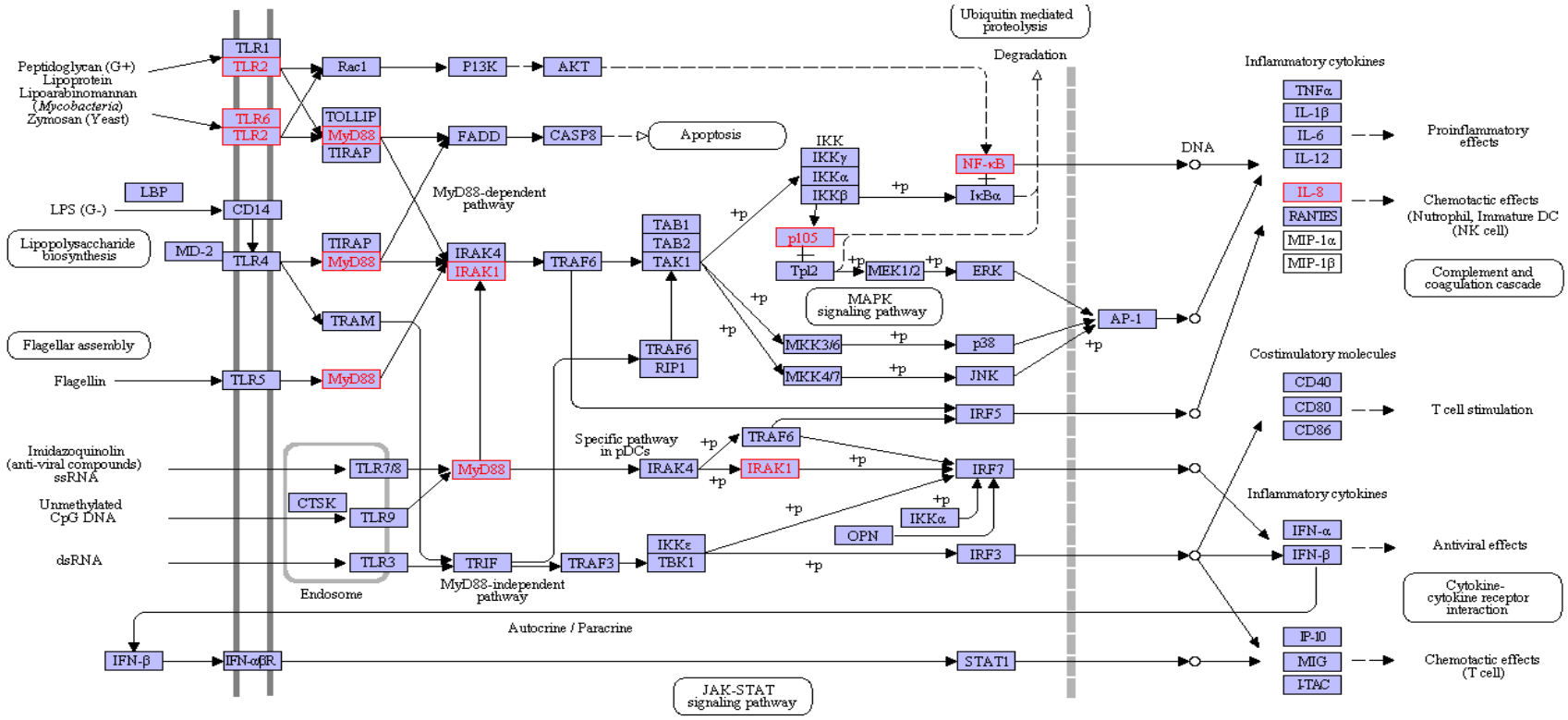
nature genetics

NATURE GENETICS VOLUME 44 | NUMBER 1 | JANUARY 2012

Toll Like Mutations

Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia

Víctor Quesada[1], Laura Conde[2], Neus Villamor[2], Gonzalo R Ordóñez[1], Pedro Jares[2], Laia Bassaganyas[3], Andrew J Ramsay[1], Sílvia Beà[2], Magda Pinyol[4], Alejandra Martínez-Trillos[5], Mónica López-Guerra[2], Dolors Colomer[2], Alba Navarro[2], Tycho Baumann[5], Marta Aymerich[2], María Rozman[2], Julio Delgado[5], Eva Giné[5], Jesús M Hernández[6], Marcos González-Díaz[6], Diana A Puente[1], Gloria Velasco[1], José M P Freije[1], José M C Tubío[3], Romina Royo[7], Josep L Gelpí[7], Modesto Orozco[7], David G Pisano[8], Jorge Zamora[8], Miguel Vázquez[8], Alfonso Valencia[8], Heinz Himmelbauer[9], Mónica Bayés[10], Simon Heath[10], Marta Gut[10], Ivo Gut[10], Xavier Estivill[3], Armando López-Guillermo[5], Xose S Puente[1], Elías Campo[2,11] & Carlos López-Otín[1,11]



LETTE

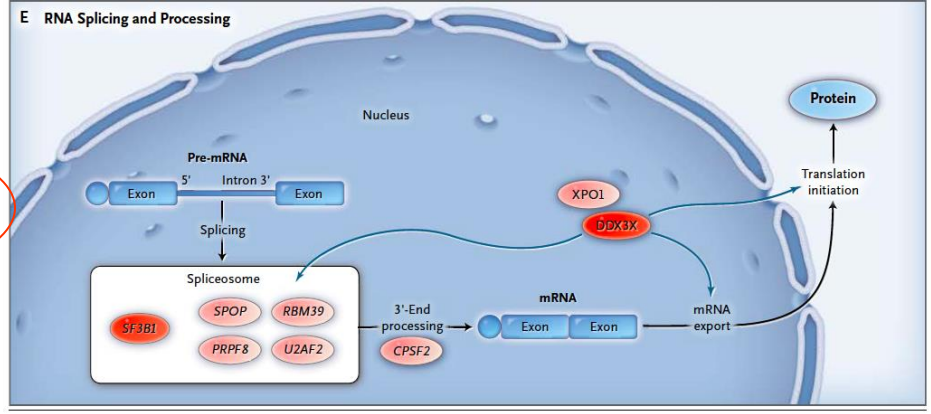
nature genetics

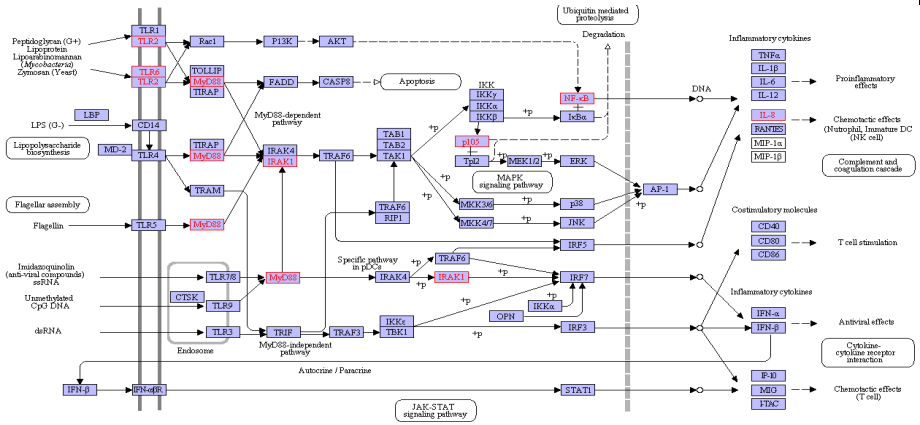
NATURE GENETICS VOLUME 44 | NUMBER 1 | JANUARY 2012

Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia

Víctor Quesada[1], Laura Conde[2], Neus Villamor[2], Gonzalo R Ordóñez[1], Pedro Jares[2], Laia Bassaganyas[3], Andrew J Ramsay[1], Sílvia Beà[2], Magda Pinyol[4], Alejandra Martínez-Trillos[5], Mónica López-Guerra[2], Dolors Colomer[2], Alba Navarro[2], Tycho Baumann[5], Marta Aymerich[2], María Rozman[2], Julio Delgado[5], Eva Giné[5], Jesús M Hernández[6], Marcos González-Díaz[6], Diana A Puente[1], Gloria Velasco[1], José M P Freije[1], José M C Tubío[3], Romina Royo[7], Josep L Gelpí[7], Modesto Orozco[7], David G Pisano[8], Jorge Zamora[8], Miguel Vázquez[8], Alfonso Valencia[8], Heinz Himmelbauer[9], Mónica Bayés[10], Simon Heath[10], Marta Gut[10], Ivo Gut[10], Xavier Estivill[3], Armando López-Guillermo[5], Xose S Puente[1], Elías Campo[2,11] & Carlos López-Otín[1,11]

Here we perform whole-exome sequencing of samples from 105 individuals with chronic lymphocytic leukemia (CLL)[1,2], the most frequent leukemia in adults in Western countries. We found 1,246 somatic mutations potentially affecting gene function and identified 78 genes with predicted functional alterations in more than one tumor sample. Among these genes, *SF3B1*, encoding a subunit of the spliceosomal U2 small nuclear ribonucleoprotein (snRNP), is somatically mutated in 9.7% of affected individuals. Further analysis in 279 individuals with CLL showed that *SF3B1* mutations were associated with faster disease progression and poor overall survival. This work provides the first comprehensive catalog of somatic mutations in CLL with relevant clinical correlates and defines a large set of new genes that may drive the development of this common form of leukemia. The results reinforce the idea that targeting several well-known genetic pathways, including mRNA splicing, could be useful in the treatment of CLL and other malignancies.





The NEW ENGLAND JOURNAL of MEDICINE

N ENGL J MED 365;26 NEJM.ORG DECEMBER 29, 2011

ORIGINAL ARTICLE

SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia

Lili Wang, M.D., Ph.D., Michael S. Lawrence, Ph.D., Youzhong Wan, Ph.D., Petar Stojanov, B.A., Carrie Sougnez, B.S., Kristen Stevenson, M.S., Lillian Werner, M.S., Andrey Sivachenko, Ph.D., David S. DeLuca, Ph.D., Li Zhang, Ph.D., Wandi Zhang, M.D., Alexander R. Vartanov, B.A., Stacey M. Fernandes, B.S., Natalie R. Goldstein, B.A., Eric G. Folco, Ph.D., Kristian Cibulskis, B.S., Bethany Tesar, M.S., Quinlan L. Sievers, B.A., Erica Shefler, B.S., Stacey Gabriel, Ph.D., Nir Hacohen, Ph.D., Robin Reed, Ph.D., Matthew Meyerson, M.D., Ph.D., Todd R. Golub, M.D., Eric S. Lander, Ph.D., Donna Neuberg, Sc.D., Jennifer R. Brown, M.D., Ph.D., Gad Getz, Ph.D., and Catherine J. Wu, M.D.


INB
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
IRB
BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE


CRG
Centre de Regulació Genòmica
INB


INB
cnio
Centro Nacional de Investigaciones Oncológicas