1. - Personalized Medicine
2. - Epigenetics and Networks

Alfonso Valencia
Spanish National Cancer Research Centre
CNIO
valencia@cnio.es
@alfons_valencia

ESHG Satellite Symposium
“Machine Learning for Personalized Medicine”
Barcelona, May 2016
Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses

Siân Jones,1* Xiaosong Zhang,1* D. Williams Parsons,1,2* Jimmy Cheng-Ho Lin,1* Rebecca J. Leary,1* Philipp Angenendt,1* Parminder Mankoo,3 Hannah Carter,3 Hirohiko Kamiyama,4 Antonio Jimeno,1 Seung-Mo Hong,4 Baojin Fu,4 Ming-Tseh Lin,4 Eric S. Calhoun,4 Mihoko Kamiyama,4 Kimberly Walter,4 Tatiana Nikolskaya,5 Yuri Nikolsky,6 James Hartigan,7 Douglas R. Smith,7 Manuel Hidalgo,7 Steven D. Leach,1,8 Alison P. Klein,1,4 Elizabeth M. Jaffee,1,4 Michael Goggins,7,8 Anindita Maitra,1,4 Christine Iacobuzio-Donahue,1,4 James R. Eshleman,1,4 Scott E. Kern,1,4 Ralph H. Hruban,1,4 Rachel Karchin,3 Nickolas Papadopoulos,1 Giovanni Parmigiani,1,9 Bert Vogelstein,1,10 Victor E. Velculescu,1† Kenneth W. Kinzler1†

Manuel Hidalgo,
Director
Clinical Programme
CNIO
PALB2
Identified as breast cancer susceptibility gene whose protein is closely associated with BRCA2 and is essential for BRCA2 anchorage to nucleus structures. [2008]

PALB2 mutations have been previously reported in patients with familial breast cancer, and the PALB2 protein is a binding partner for BRCA2. [2009]
CNIO co-clinical cancer initiative

Manuel Hidalgo,
Director
Clinical Programme
CNIO
## Genomic alterations

<table>
<thead>
<tr>
<th>#</th>
<th>PATIENT ID</th>
<th>TUMOR TYPE</th>
<th># MUTS</th>
<th>RELEVANT MUTATED GENES</th>
<th># CNV</th>
<th>RELEVANT CNV</th>
<th>PUTATIVE TARGETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PGDX4</td>
<td>Neuroendocrine tumor</td>
<td>5</td>
<td>CREB3L3, ITFPR2, MYO5B</td>
<td>0</td>
<td>0</td>
<td>CREB3L3</td>
</tr>
<tr>
<td>2</td>
<td>PGDX30</td>
<td>Glioblastoma</td>
<td>63</td>
<td>EP3A3, NF1, PTNP11, FA2, CDKN2A</td>
<td>0</td>
<td>0</td>
<td>NF1</td>
</tr>
<tr>
<td>3</td>
<td>PGDX7</td>
<td>High grade pancreatic</td>
<td>62</td>
<td>ARID1A, ARID1B, JAKMIP2, JARID2</td>
<td>6</td>
<td>GNG11</td>
<td>PI3KCA, DDR2</td>
</tr>
<tr>
<td>4</td>
<td>PGDX11</td>
<td>Pancreatic Adenocarcinoma</td>
<td>38</td>
<td>KRA55, UBA1, FAM53H, SMAD4, SLC15A2, SLC3A2, SLC22A17, TPS3.</td>
<td>10</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>PGDX61</td>
<td>Melanoma Uveal</td>
<td>5</td>
<td>GNA11, TAKO3</td>
<td>0</td>
<td>0</td>
<td>GNA11</td>
</tr>
<tr>
<td>6</td>
<td>PGDX68</td>
<td>Colon cancer</td>
<td>71</td>
<td>APC, DICER1, TPS3, CHEK1, SOS1</td>
<td>63</td>
<td>0</td>
<td>CHEK1</td>
</tr>
<tr>
<td>7</td>
<td>PGDX76</td>
<td>Melanoma</td>
<td>952</td>
<td>BRCA1, EZH2, FGFR2, FGFR3, KDR, KRAS, MET, MPL, PRKCB, PIK3CA, PIK3CB, PIK3CD, PTOR3</td>
<td>0</td>
<td>0</td>
<td>FGFR2, FGFR3, PIK3CB, PIK3CD, PIK3CB, PTOR3</td>
</tr>
<tr>
<td>8</td>
<td>PGDX135</td>
<td>Metanoma</td>
<td>29</td>
<td>BAX3, DNA5, MDN1, NRAS</td>
<td>2</td>
<td>SKT19</td>
<td>NRAS</td>
</tr>
<tr>
<td>9</td>
<td>PGDX368</td>
<td>Pancreatic Adenocarcinoma</td>
<td>18</td>
<td>SMAD4, KRAS, ERB2IP</td>
<td>0</td>
<td>0</td>
<td>ERB2IP</td>
</tr>
<tr>
<td>10</td>
<td>PGDX331</td>
<td>Pancreatic Adenocarcinoma</td>
<td>21</td>
<td>KRAS, XPC, P53</td>
<td>0</td>
<td>0</td>
<td>XPC</td>
</tr>
<tr>
<td>11</td>
<td>PGDX369</td>
<td>Pancreatic Adenocarcinoma</td>
<td>29</td>
<td>KRAS, P53, SMAD4</td>
<td>0</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>12</td>
<td>PGDX3762</td>
<td>Renal Carcinoma</td>
<td>25</td>
<td>BAP1</td>
<td>945</td>
<td>0</td>
<td>BAP1, FGFR3, NOTCH1, STK11, GNA11, ZNF668, SOCS1, IRF2</td>
</tr>
<tr>
<td>13</td>
<td>PGDX330</td>
<td>Glioblastoma</td>
<td>64</td>
<td>MLLT15, PBRM1</td>
<td>27</td>
<td>EGFR, CDKN2A, ESRF11</td>
<td>EGFR</td>
</tr>
<tr>
<td>14</td>
<td>PGDX27</td>
<td>NSCLC</td>
<td>45</td>
<td>EGFR, EZH2, TPR, TPS3, ROH</td>
<td>0</td>
<td>0</td>
<td>EGFR</td>
</tr>
<tr>
<td>15</td>
<td>PGDX17</td>
<td>NSCLC</td>
<td>69</td>
<td>ARID4B, PIK3RE, PTPRC, TPS3, TOPO2A</td>
<td>13</td>
<td>0</td>
<td>PTPRC, KIF58-RET Fusion STK11 Whole Gene Deletion</td>
</tr>
<tr>
<td>16</td>
<td>PGDX140</td>
<td>NSCLC</td>
<td>391</td>
<td>BUB3, CYLD, ESRF, EGFR, KRA55, KEAP1, KEHR, MTHF, NTRK1, NTRK3, MTHYR, RUNX1T1, TPS3</td>
<td>12</td>
<td>CDKN2A</td>
<td>EGFR</td>
</tr>
<tr>
<td>17</td>
<td>PGDX24</td>
<td>SCLC</td>
<td>32</td>
<td>PAPPA2</td>
<td>Not Done</td>
<td>Not Done</td>
<td>None</td>
</tr>
<tr>
<td>18</td>
<td>PGDX60</td>
<td>NSCLC</td>
<td>9</td>
<td>NOCT2, MML3</td>
<td>6</td>
<td>0</td>
<td>NOCT2</td>
</tr>
<tr>
<td>19</td>
<td>PGDX38</td>
<td>NSCLC</td>
<td>38</td>
<td>DLK1, IAK3</td>
<td>Not Done</td>
<td>Not Done</td>
<td>JAK3</td>
</tr>
<tr>
<td>20</td>
<td>PGDX48</td>
<td>Duodenal cancer</td>
<td>127</td>
<td>CARD11, CHL1, FANCD2, IRS3, NRAS, SMARCA4, TPS3, PIK3R1</td>
<td>25</td>
<td>0</td>
<td>NRAS, PIK3R1</td>
</tr>
<tr>
<td>21</td>
<td>PGDX39</td>
<td>Colon cancer</td>
<td>172</td>
<td>APC, EGFR, FN1, GRM1, TPS3, TSG2</td>
<td>0</td>
<td>0</td>
<td>EGFR, FN1, TSG2</td>
</tr>
<tr>
<td>22</td>
<td>PGDX372</td>
<td>Colon cancer</td>
<td>108</td>
<td>APC, AXIN2, EGFR, KDM6A, PTEN, PIK3CA, TPS3</td>
<td>92</td>
<td>0</td>
<td>EGFR, PIK3CA, PTEN</td>
</tr>
<tr>
<td>23</td>
<td>PGDX310</td>
<td>Esophageal cancer</td>
<td>96</td>
<td>PIK3R1, NF1, TPS3</td>
<td>65</td>
<td>CCND1</td>
<td>NF1, PIK3R1</td>
</tr>
</tbody>
</table>

**Garralda et al.**  
*Integrated Next Generation Sequencing and Avatar Mouse Models for Personalized Cancer Treatment.*  
Clinical Cancer Research 2014
Case 3: High grade pancreatic neuroendocrine tumor

- 44 years old male.
- High grade neuroendocrine carcinoma with disseminated lymph node disease
- Prior treatment with gemcitabine-Oxaliplatin: PR (Partial Response).
- When Progression Disease (PD): Fresh tumor specimen from lymph node was obtained for exome sequencing and for Avatar/xenograft generation.

Exome sequencing analysis:
- 64 somatic relevant mutations
- 6 Copy number variations

Point mutation in PIK3CA gene: 909F>C

PI3K inhibitors and MEK inhibitors

Point mutation in DDR2 protein: 381P>A

DDR2 mut - 4% of Squamous Cell Lung Cancer. DDR2 mut associated with sensitive to Dasatinib (Hammerman et al. Cancer Discovery 2011: )
Case 3: High grade pancreatic neuroendocrine tumor

Avatar PDX- Panc 1 test

PI3Ki, MEKi and Gemcitabine

Dasatinib + Gemcitabine

Mutation in PI3K

Patient’s treatment

Mutation in DDR2

Why it did not work?
CNIO co-clinical cancer initiative

Manuel Hidalgo,
Director
Clinical Programme
CNIO
Pancreas PDXs conserve key features of their tissue of origin

A) 

B) 

Martinez-Garcia et al., Genome Med 2014
Pancreas PDXs conserve key features of their tissue of origin

Martinez-Garcia et al., Genome Med 2014
CNIO co-clinical cancer initiative
Applying Bioinformatics to Precision Medicine

A new breed of scientist

According to Valencia, the job that Al-Shahrour does requires a very wide range of knowledge and skills; he emphasizes her "biological background, capacity to develop bioinformatics methods, deep understanding of genomics, good communication skills and proved record in team management." Also important, he adds, is her clear understanding of the limitations of the experimental and computational techniques.

Early-career scientists who wish to follow in her footsteps must be ready to embrace the training challenges. Tamayo writes: "My advice to them is to study mathematics (not only old statistics but also advanced probability), information theory, machine learning, programming, numerical methods, chemistry, physics, cellular biology and biochemistry. It is important not only to be able to talk to multiple domain experts, and develop a solid hard-core analytical mind frame to cast problems, but also to have access to a rich set of paradigms about how to deal with complexity." Cancer pharmacogenomics is "a particularly demanding field that requires a lot of flexibility and adaptability in terms of what problems one solves over time and in requiring to learn from many fields of expertise," he adds.
CNIO personalized Medicine Initiative

• Workflow of PerMed
  – Constant flow of cases
  – Short response time (days)
  – Weekly meeting with clinicians
  – Accumulating statistics to derive patterns

• *still ... this is a research project (all PerMed are research projects)*
CNIO co-clinical cancer initiative
Clinical Genetics Has a Big Problem That's Affecting People's Lives

Unreliable research can lead families to make health decisions they might regret.

Daniel MacArthur at Massachusetts General Hospital found a similar trend in a study of over 60,000 people, the results of which have been uploaded to a preprint server. On average, each of these volunteers is walking around with 53 gene variants that are classified as “pathogenic” in two widely-used databases. When the team took a closer look at 200 of these variants, they found enough evidence to classify just nine of them as pathogenic.

“Reproducibility problems in clinical genetics ... have massive and real-time consequences for thousands of families.”
Bioinformatics for personalized medicine

**Genome Analysis**

Consequences of alterations: SNPs, CNVs, miRNA, Epigenetics

**Phenotypic and pathways Analysis**

Drug relation Proteins, genes pathways

Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics

Alfonso Valencia and Manuel Hidalgo

Valencia and Hidalgo Genome Medicine 2012, 4:61
https://genomemedicine.com/content/4/7/61
In the past few years, it has become clear that a phenomenon called alternative splicing is one reason human genomes can produce such complexity with so few genes.*

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.

*Science, July 2005
The UniProt database picks one variant and clusters the others around it.

The APPRIS database select principal isoforms based on:
- protein structural information,
- functionally important residues,
- protein functional domains and
- evidence of cross-species conservation.

Assumption 1: genes have just one principal isoform
Assumption 2: the principal isoform is the oldest in evolutionary terms.

APPRIS: annotation of principal and alternative splice isoforms.

by Michael Tress
2 or more protein isoforms for 0.67% of the human genome

Isoforms with small differences are significantly over-represented.

(ketohexokinase homologous replaceable exons.)

Comparative Proteomics Reveals a Significant Bias Toward Alternative Protein Isoforms with Conserved Structure and Function

Iñaki Ezkurdia,†‡, Angela del Pozo,†‡, Adam Frankish,‡, Jose Manuel Rodriguez,† Jennifer Harrow,‡ Keith Ashman,§ Alfonso Valencia,†‡, and Michael L. Tress*†

A single isoform (principal isoform) per gene

Alterations in isoform composition in cancer

More than anticipated trans-splicing events
Chimeras confirmed at the RNA and protein levels

Chimera of actin, ACTG1 and ribosomal protein RPL13A
Overlapping unique peptides from 18 mass spectrometry experiments (Pvalue<10^{-52}) confirming transcript from ChimerDB (Kim et al., 2006, Kim et al. 2010) and a few further confirmed by SRM

*The motif 'GDGV' (a red rectangle) is the ATP-binding site missing in the chimera*
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 genomewide transcriptional analyses has increased. However, the number of known chimeric transcripts far outnumbers the reported number of chimeric proteins. Here, Frankel-Morgenstern et al increase the number of identified translated chimeric transcripts and describe features indicative of their biological functionality.

Firstly, the authors analyzed 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

Transplicling/chimeric mRNAs

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 Valencia1,6

Two overlapping mass-spec peptides, (18 experiments, Pvalue<10^-7, FDR<1%)
Dr. Levin, Weizmann Institute
**ChiTaRS: chimeric transcripts/proteins Database**

**Supporting ESTs -- First Part**

**Chimera First Part BCR Chr22 (+)**

**Chimera Second Part ABL1 Chr9 (+)**

**Supporting ESTs -- Second Part**

**Research**

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

Milana Frenkel-Morgenstern,¹ Vincent Lacroix,² lakes Ezkurdia,¹ Yishai Levin,³ Alexandra Gabashvili,⁴ Jaime Prilusky,⁴ Angela del Pozo,¹ Michael Tress,¹ Rory Johnson,⁵ Roderic Guigo,⁵ and Alfonso Valencia¹,⁶

**Milana Frenkel-Morgenstern et al., NAR 2013. 2015**
Technical Infrastructure

Operation at file system level
Developed CNIO, accessible github, will be installed at CNAG.

as in ICGC
(similar to MongoDB)

as in ICGC frontend
Use cases are divided in:

1. Data management

2. Answer research questions

The data portal provides functionalities to satisfy both categories.
Bioinformatics Core Unit

Gonzalo Gómez
Osvaldo Graña
Miriam Rubio
David G. Pisano

Spanish National Bioinformatics Institute

Victor de la Torre
Miguel Vazquez
José María Fernández-González
A Network Biology Approach to Information Processing in Epigenetic Regulation

Alfonso Valencia
Spanish National Cancer Research Centre
CNIO

valencia@cnio.es
@alfons_valencia

ESHG symposium 2016: Machine Learning for Personalized Medicine
Barna May 2016
Complex relationship between Cytosine modifications, Histone marks and Chromatin Binding Proteins

EPIGENETICS/EPIGENOMICS

Stem cells
Segmenting the Genome in Functional Regions

Mouse embryonic stem cells
“Core” epigenomic features (basic scaffold)

Histone modifications

- **H3K4me3** --- ENCODE, GSE11724, GSE12241, GSE36114
- **H3K4me2** --- ENCODE, GSE11172, GSE36114
- **H3K4me1** --- ENCODE, GSE11172, GSE36114
- **H3K27Ac** --- ENCODE, GSE36114
- **H3K9me3** --- ENCODE, GSE12241, GSE18371
- **H2Aub1** --- GSE34518
- **H3K27me3** --- GSE12241, GSE41589, GSE36114
- **H3K36me2** --- GSE41589
- **H3K36me3** --- ENCODE, GSE11724, GSE12241, GSE34518, GSE41589, GSE36114
- **H3K79me2** --- GSE11724
- **H4K20me3** --- GSE12241
- **H3K9Ac** --- ENCODE
- **H2AZ** --- GSE36114

Cytosine modifications

- **5mC** --- GSE28682
- **5hmC** --- GSE28682
- **5fC** --- GSE40148

Insulator

**CTCF** --- GSE11431, GSE25777, GSE28247
17 “Core” epigenomic features
CTCF, cytosine & histone modifications

20 Chromatin states
Combinations of core features

ChromHMM

20 Chromatin states obtained by combining ChipSeq data of 17 core features
Chromatin Related Proteins
(data from mouse ESCs)

**Polycomb**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoREST</td>
<td>GSE27841</td>
</tr>
<tr>
<td>REST</td>
<td>GSE27841</td>
</tr>
<tr>
<td>RING1B</td>
<td>GSE34518, GSE42466</td>
</tr>
<tr>
<td>SIN3A</td>
<td>GSE24843</td>
</tr>
<tr>
<td>SUZ12</td>
<td>GSE11724, GSE42466, 41589, 11431</td>
</tr>
<tr>
<td>PHF19</td>
<td>GSE41589, GSE41609</td>
</tr>
</tbody>
</table>

**ESC Transcription factors**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-MYC</td>
<td>GSE11431</td>
</tr>
<tr>
<td>E2F1</td>
<td>GSE11431</td>
</tr>
<tr>
<td>MAX</td>
<td>GSE48175</td>
</tr>
<tr>
<td>NANOG</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>N-MYC</td>
<td>GSE11431</td>
</tr>
<tr>
<td>OCT4</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>SMAD1</td>
<td>GSE11431</td>
</tr>
<tr>
<td>SOX2</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>STAT3</td>
<td>GSE11431</td>
</tr>
<tr>
<td>KLF4</td>
<td>GSE11431</td>
</tr>
<tr>
<td>Tcfcp2l1</td>
<td>GSE11431</td>
</tr>
<tr>
<td>C-MYC</td>
<td>GSE11431</td>
</tr>
<tr>
<td>E2F1</td>
<td>GSE11431</td>
</tr>
<tr>
<td>MAX</td>
<td>GSE48175</td>
</tr>
<tr>
<td>NANOG</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>N-MYC</td>
<td>GSE11431</td>
</tr>
<tr>
<td>OCT4</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>SMAD1</td>
<td>GSE11431</td>
</tr>
<tr>
<td>SOX2</td>
<td>GSE11431, GSE11724</td>
</tr>
<tr>
<td>STAT3</td>
<td>GSE11431</td>
</tr>
<tr>
<td>KLF4</td>
<td>GSE11431</td>
</tr>
<tr>
<td>Tcfcp2l1</td>
<td>GSE11431</td>
</tr>
</tbody>
</table>

**Enhancer**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPBL</td>
<td>GSE22562</td>
</tr>
<tr>
<td>SMC1</td>
<td>GSE22562</td>
</tr>
<tr>
<td>SMC3</td>
<td>GSE22562</td>
</tr>
<tr>
<td>RAD21</td>
<td>GSE25777</td>
</tr>
<tr>
<td>P300</td>
<td>GSE11431, GSE28247</td>
</tr>
</tbody>
</table>

**Cohesin complex**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPBL</td>
<td>GSE22562</td>
</tr>
<tr>
<td>SMC1</td>
<td>GSE22562</td>
</tr>
<tr>
<td>SMC3</td>
<td>GSE22562</td>
</tr>
<tr>
<td>RAD21</td>
<td>GSE25777</td>
</tr>
</tbody>
</table>

**Others**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBX3</td>
<td>GSE4424</td>
</tr>
<tr>
<td>CBX7</td>
<td>GSE42466</td>
</tr>
<tr>
<td>HCFC1</td>
<td>ENCODE</td>
</tr>
<tr>
<td>MAFK</td>
<td>ENCODE</td>
</tr>
<tr>
<td>ZC3H11A</td>
<td>ENCODE</td>
</tr>
<tr>
<td>ZNF384</td>
<td>ENCODE</td>
</tr>
<tr>
<td>ESRRB</td>
<td>GSE11431</td>
</tr>
<tr>
<td>KAP1</td>
<td>GSE41903</td>
</tr>
<tr>
<td>LAMINB</td>
<td>GSE28247</td>
</tr>
<tr>
<td>MI2B</td>
<td>GSE27841</td>
</tr>
<tr>
<td>OGT</td>
<td>GSE39154</td>
</tr>
<tr>
<td>TCF3</td>
<td>GSE11724</td>
</tr>
<tr>
<td>BRG1</td>
<td>GSE14344</td>
</tr>
<tr>
<td>POLII</td>
<td>GSE12241, GSE28247</td>
</tr>
<tr>
<td>RNAPII-8WG16</td>
<td>GSE34518</td>
</tr>
<tr>
<td>RNAPII-S2P</td>
<td>GSE34518</td>
</tr>
<tr>
<td>RNAPII-S5P</td>
<td>GSE34518</td>
</tr>
<tr>
<td>RNAPII-S7P</td>
<td>GSE34518</td>
</tr>
<tr>
<td>MED12</td>
<td>GSE22562</td>
</tr>
<tr>
<td>MED1</td>
<td>GSE22562</td>
</tr>
<tr>
<td>TAF1</td>
<td>GSE36114</td>
</tr>
</tbody>
</table>

**Chromatin remodelers**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDAC1</td>
<td>GSE27841</td>
</tr>
<tr>
<td>HDAC2</td>
<td>GSE27841</td>
</tr>
<tr>
<td>KDM2A</td>
<td>GSE40860</td>
</tr>
<tr>
<td>KDM2B</td>
<td>GSE40860</td>
</tr>
<tr>
<td>LSD1</td>
<td>GSE18515, GSE27841</td>
</tr>
<tr>
<td>TET1</td>
<td>GSE24843</td>
</tr>
<tr>
<td>MBD1A</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MBD1B</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MBD2A</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MBD2T</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MBD3A</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MBD4</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MECP2</td>
<td>GSE39610</td>
</tr>
<tr>
<td>MLL2</td>
<td>GSE48172</td>
</tr>
<tr>
<td>SETDB1</td>
<td>GSE18371</td>
</tr>
</tbody>
</table>

**Transcription related**

<table>
<thead>
<tr>
<th>Protein</th>
<th>GSE Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRG1</td>
<td>GSE14344</td>
</tr>
<tr>
<td>OGT</td>
<td>GSE39154</td>
</tr>
<tr>
<td>TCF3</td>
<td>GSE11724</td>
</tr>
<tr>
<td>MED12</td>
<td>GSE22562</td>
</tr>
<tr>
<td>MED1</td>
<td>GSE22562</td>
</tr>
<tr>
<td>TAF1</td>
<td>GSE36114</td>
</tr>
</tbody>
</table>
Epigenomic Mouse Stem Cell co-localization Network

60 Chromatin-related Proteins (CrPs)

Core epigenomic features

Distribution of CRPs / Epi features in 20 Heterochromatin states

Heterochromatin state dependent Co-localization Networks
Epigenomic Mouse Stem Cell co-localization Network

60 Chromatin-related Proteins (CrPs)

Core epigenomic features

Distribution of CRPs in 20 Heterochromatin states

Heterochromatin state dependent Co-localization Networks
Inference of chromatin **DIRECT** co-localization networks from ChIP-seq data

Sparse Partial Correlation Networks  
(Lasserre et al. 2013)

Regularized linear regression with Elastic Nets  
(Perner et al. 2014)
State 18 (Polycomb promoter)

Colors reflect known protein complexes
Global Co-localization network reflects organization in protein complexes
Epigenetics as a communication system

“Writer”

Epigenetic message (presence/absence of 5hmC, H3K4me1, ...)

“Eraser”

“Reader”

Sender  Message  Receiver

Acetylases, methylases, phosphorylases

Deacetylases, demethylases, phosphatases

Bromodomain, chromodomain, PHD finger, WD40 repeat
From “Location” to “Communication” Network

“Writer” Epigenetic signal (presence/absence of 5hmC, H3K4me1, ...)
“Eraser”

“Reader”

Sender
1.- Writer/Eraser from experiments
2.- Influences signal location (eg. KOs)

Signal
Histone marks (H3K4me1, ...). Cytosine modifications (5mC, 5hmC, 5fC).

Receiver
Not-sender signal interactors

**KNOWLEDGE FROM THE LITERATURE -> EDGE DIRECTIONALITY**

124 (52.5%) directional interactions: 56 emitter / signal, 68 receiver/signal
18 CrP nodes act as emitters or receivers of different signals
Indegree (Transcription)
Outdegree
(5hmC is the message with more receivers)
Co-localization network

**Indegree** (Transcription, 5hmC)

**Outdegree**
(5hmC is the message with more receivers)

**Betweennes/Centrality**
control that this node exerts over the interactions of other nodes in the network

\[ C_b(n) = \sum_{s \neq n \neq t} (\sigma_{st}(n) / \sigma_{st}) \]

5hmC is the most central node
Influence/popularity

- Influential nodes: information easily spreads out to the rest of the network,
- Popular nodes gather information from many regions of the network.

Effect of directionality miss-assignments and random edges removal

TET1 and LSD1, are emitters of 5hmC.
star-like chromnets: communication modules that connect different protein complexes (emitter/signal and signal/receiver interactions)

i.e. two central connectors (5fC and RYBP) connecting Polycomb, Mediator and TET1-SIN3A complexes >>> enriched in active transcription states and regulatory elements.
5hmC nucleates a star-like sub-network

5hmC indirectly connects to H3K4me1 via TET1, and with 5mC via MBD2.0.

Chromnet is enriched in regulatory elements.
Co-evolution and evolutionary interactions

Co-evolution:

- Reciprocal evolution between interacting species driven by natural selection (John N. Thompson).
- Co-evolution in about biological objects speaking evolution with each other.
- Evolution of communication requires co-adaptation of the agents and adaptation of the interaction as whole (Maynard Smith, J. & Harper, D.G.C. 2003)

Darwin (1859), *On the origin of species.*

Mirrortree Method: finding interaction partners at genome level

High-confidence prediction of global interactomes based on genome-wide coevolutionary networks

David Juan, Florencio Pazos
Nat. Rev Genet 2013

David Juan *, Florencio Pazos †, and Alfonso Valenc
+ Author Affiliations

Proc Natl Acad Sci U S A. 2008
Detecting co-evolving CrPs

13,579 metazoan (88 species) trees from eggNOG (1)

(2,3)

19,267 orthologs-only trees containing mouse proteins

58 CrPs

p value < 0.05

10,000 x 58 random proteins (background distribution)

Maximum-entropy distribution in the space of species-species distance bins \(d\) for fixed single and pair protein frequencies

\[
P(d) = \exp \left[ \sum_a h_a(d_a) + \sum_{a,b} J_{a,b}(d_a, d_b) \right]
\]

Jp,q regulates the interactions between proteins in the model.
A strong positive parameter can be interpreted as the direct symmetrical interaction between the two proteins a, b

Getting a direct co-evolutionary network

coevolutionary couplings network

\[ P(d) = \exp \left( \sum_a h_a(d_a) + \sum_{a,b} J_{a,b}(d_a, d_b) \right) \]
34 Co-evolution based connections of the 58 CrPs

CTCF/Cohesin seems to condition the evolution of most complexes

CTCF/Cohesin

Histone deacetylation

Cohesin

involved in cohesin loading

SWI/SNF complex

Histone deacetylation

Cohesin

involved in cohesin loading

SWI/SNF complex
About metazoan co-evolution
(5hmC is key in communication and co-evolution)

32% of co-evolving pairs co-localize in mESCs

P < 1E-5
About metazoan co-evolution
(5hmC is key in communication and co-evolution)

32% of co-evolving pairs co-localize in mESCs P < 1E-5

Co-evolution points to Sender-Receiver epigenetic communication

69% (background 22%) P < 1E-6

Co-evolution points to 5hmC-mediated epigenetic communication

25% (background 7%) P < 1E-6
Co-evolution between 5hmC interactors

Facilitates promoter-enhancer interactions

Part of SNF/SWI complex (required for activation of repressed genes)

MBDs inhibit binding of TET1 to 5mC

TET1: Converts 5mC into 5hmC and 5fC

MBD2

SIN3A

LSD1

OGT

TAF1

MLL2

KDM2A

KDM2B
5hmC key in communication & co-evolution in mESC
Co-evolution

Evolution of communication requires co-adaptation of the agents and adaptation of the interaction as whole (Maynard Smith, J. & Harper, D.G.C. 2003)
mESC Epigenetic Network (part 1)

Chip-Seq data – co-localization – biological complexes

Chromatin related proteins biological functions and chormatin states

5hmC, stem cells and development

Co-Evolution of Chromatin related molecular complexes
A network approach integrates 3D contacts with epigenomic data (part 2)
The 3D structure of chromatin

Z. Duan, ... W Stafford Noble, A three-dimensional model of the yeast genome, Nature 2010
Chromatin Interaction Networks (CINs)

Unravelling 3D chromatin contacts
Chromosome Conformation Capture methods

Construct networks where:
Node = chromatin fragment  Edges=contacts in 3D

What about genes? PCHi-C!

**Problem** so far: HiC contact maps dominated by interactions far from genes. Need very high coverage to pick promoters

**Solution:** Promoter-Capture HiC (PCHiC)
Add promoter capture step
Obtain promoter-centred contact maps (No pull-downs, genome-wide)

It allows to look for transcription factories (Multiple genes transcribed together)

Lyubomira Chakalova, ... Peter Fraser Replication and transcription: Shaping the landscape of the genome Nat Rev Gen 2005; Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015.
PCHi-C networks in mESCs

Interactions involving at least 1 promoter

Statistics:
Chromatin fragments (4Kb)
55,845 nodes
(19425 promoters)
69,987 edges
1 major connected component

20,523
Edges connecting
Two promoters

Coloured by modules
(highly interconnected
 Portions of the network)

Data processing with CHiCAGO
(Mikhail Spivakov)
PCHi-C networks in mESCs

Interactions involving at least 1 promoter

Statistics:
Chromatin fragments (4Kb)
55,845 nodes
(19425 promoters)
69,987 edges
1 major connected component

20,523
Edges connecting
Two promoters

Coloured by modules
(highly interconnected
Portions of the network)

Data processing with CHiCAGO
(Mikhail Spivakov)
Chromatin Assortativity (ChAs)

**Approach:**
For each chromatin fragment (node):
Calculate abundance of epigenetic feature
Eg: Proportion of fragment that has peak of EZH2 ChIP-seq
Average of proportion throughout the network (rare 0.02)

For each epigenetic feature, calculate Assortativity on the network
Eg: How much more likely are fragments covered by EZH2 peaks to interact with each other? (high 0.33)

**Data:**
PChIC networks in mouse Embryonic Stems Cells (mESCs)
(Collaboration with Peter Fraser, Babraham Institute)

78 epigenetic features (3 cytosine modification, 13 histone modifications, chromatin related proteins binding peaks)

Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs,
Juan et al. Cell Reports 2016
PCG is highly assortative in PCHi-C networks

PolyCombGroup (PCG)
Proteins and marks
EZH2
RING1b
SUZ12
PHF19

H3K27me3
H2Aub1

RNA Polymerase 2

Polycumb repressors constrain the cell state
Chromatin Assortativity in different subnetworks

In P-O subnetwork features that are only on promoter fragments have ChAs <0

HCFC1 (transcription activator complex),
SIN3A (transcriptional repressor complex), KDM2A (H3K26 demethylase),
NMYC, OGT (histone acetyl transferase complex), H3K4me2
Comparing P-P and P-O

Identify features that have different assortativities in P-P and P-O contacts

PCG on diagonal
Similar ChAs>0 in P-P and P-O
Equal importance

RNAPII:
Variable ChAs in P-O,
ChAs>0 in P-P

H3K4me3: Mark associated to active promoters
ChAs >0 in P-P
ChAs<0 in P-O
(only present in promoters)

Fragments that have this mark are more likely to interact
Preferential contacts of active gene promoters.
Assortativity of RNA Polymerase 2

5 Different RNAPII features
Binding peaks for different RNAPII variants

ChAs of RNAPII in P-O variable

Non-elongating RNAPII has low ChAs in P-P

Is expression enhancing mediated by RNAPII S2p?

Assortativity of RNAPII forms in Interactions of promoter and enhancers

Active enhancer
H3K4me1+H3K27ac

Poised Enhancer
H3K4me1+H3K27me3

Non-enhancer
No H3k4me1
Topological properties of PCG and RNAPII nodes

Apply Moduland to identify overlapping chromatin communities, measure bridgeness

<table>
<thead>
<tr>
<th></th>
<th>Bridgeness</th>
<th>Betweenness centrality</th>
<th>Degree</th>
<th>Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCG</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Very low</td>
</tr>
<tr>
<td>RNAPII general</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>RNAPII S2p</td>
<td>Low</td>
<td>Very low</td>
<td>Very low</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Model proposed:
Whereas RNAPII S5P accumulates in transcription factories, RNAPII S2p stays peripheral.

A model of transcription; gene promoters are loaded with RNAPII-Ser5P (Ser5 light gray) in factories. Elongating RNAPII-Ser2P (Ser2, dark gray) moves to the adjacent nuclear space when it becomes phosphorylated at Ser2 by CDK9.

Epigenomic Mouse Stem Cell

Distribution of CRPs in 20 Heterochromatin states

Part 1
Epigenetic + coevolutionary network

Part 2
Cromatin Capture Network
Spanish MINECO “gene sets”

Looking for students, postdocs and bioinformaticians ...

Juliane Perner  Martin Vingron  Ho-Ryun Chung  Biola-Maria Javierre  Peter Fraser  Mikhail Spivakov

Enrique Carrillo  David Juan  Simone Marsili  Daniel Rico  Vera Pancaldi

Bioinformatics Unit

Translational Bioinformatics

Spanish National Bioinformatics Institute