La genómica comparada en el estudio del desarrollo, la evolución y las enfermedades génicas

José Luis Gómez-Skarmeta

Centro Andaluz de Biología del Desarrollo
Sevilla, Spain

¿ Cómo se generan organismos tan distintos?

	Pez Fugu	Humanos
		
Nº Genes:	22.100	23.800
Cantidad DNA:	330 millones pb	3.220 millones pb

¿A qué se le llama gen?

A los marcos de lectura abierto que contienen la información necesaria para generar proteínas, esto es la región de ADN que se transcribe en ARNm

Pero el ADN también contiene regiones reguladoras necesarias para controlar cuándo, cómo y dónde se transcriben los marcos de lectura abiertos

Por ello, en un sentido estricto, un gen es el conjunto del marco de lectura abierto y todas sus regiones reguladoras

¿Cual es la influencia de las regiones reguladoras en la evolución?

Durante la evolución de vertebrados, no ha aumentado en gran medida el número de genes, pero éstos se han hecho más grandes, o lo que es lo mismo, ha aumentado el número de regiones reguladoras que actúan sobre ellos

Un mecanismo importante de la evolución es añadir nuevos elementos reguladores a genes con múltiples funciones durante el desarrollo embrionario

Estos genes suelen codificar tanto factores de transcripción (proteínas que se unen a ADN) como moléculas señalizadoras. *Genes de desarrollo*.

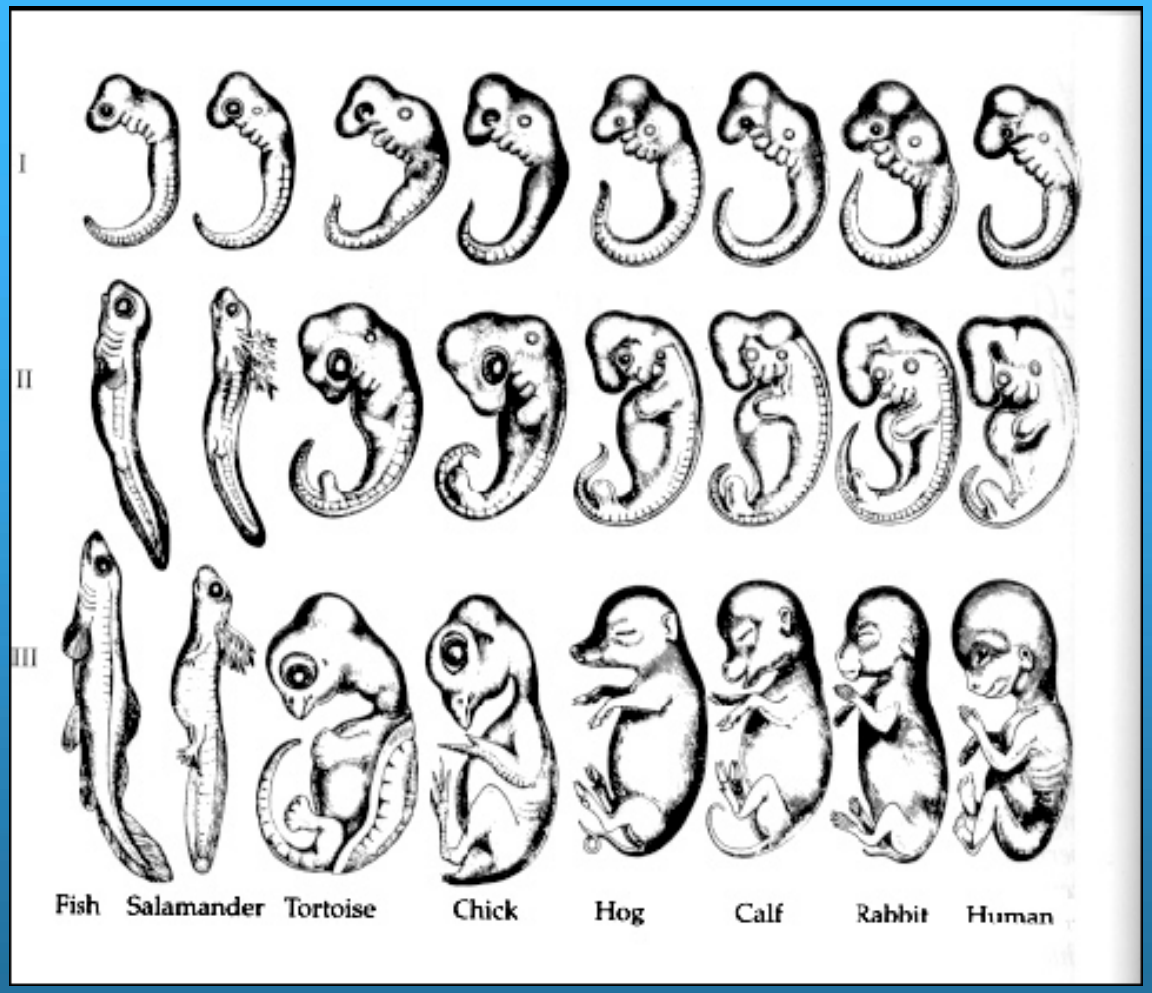
El lenguaje de los elementos reguladores es desconocido lo que previene su detección *in sílico* en el 95% del ADN que es no codificante

Genoma:

90-95% **REGULACION**

5% CODIFICANTE (proteínas)
***conservado* en Metazoos**

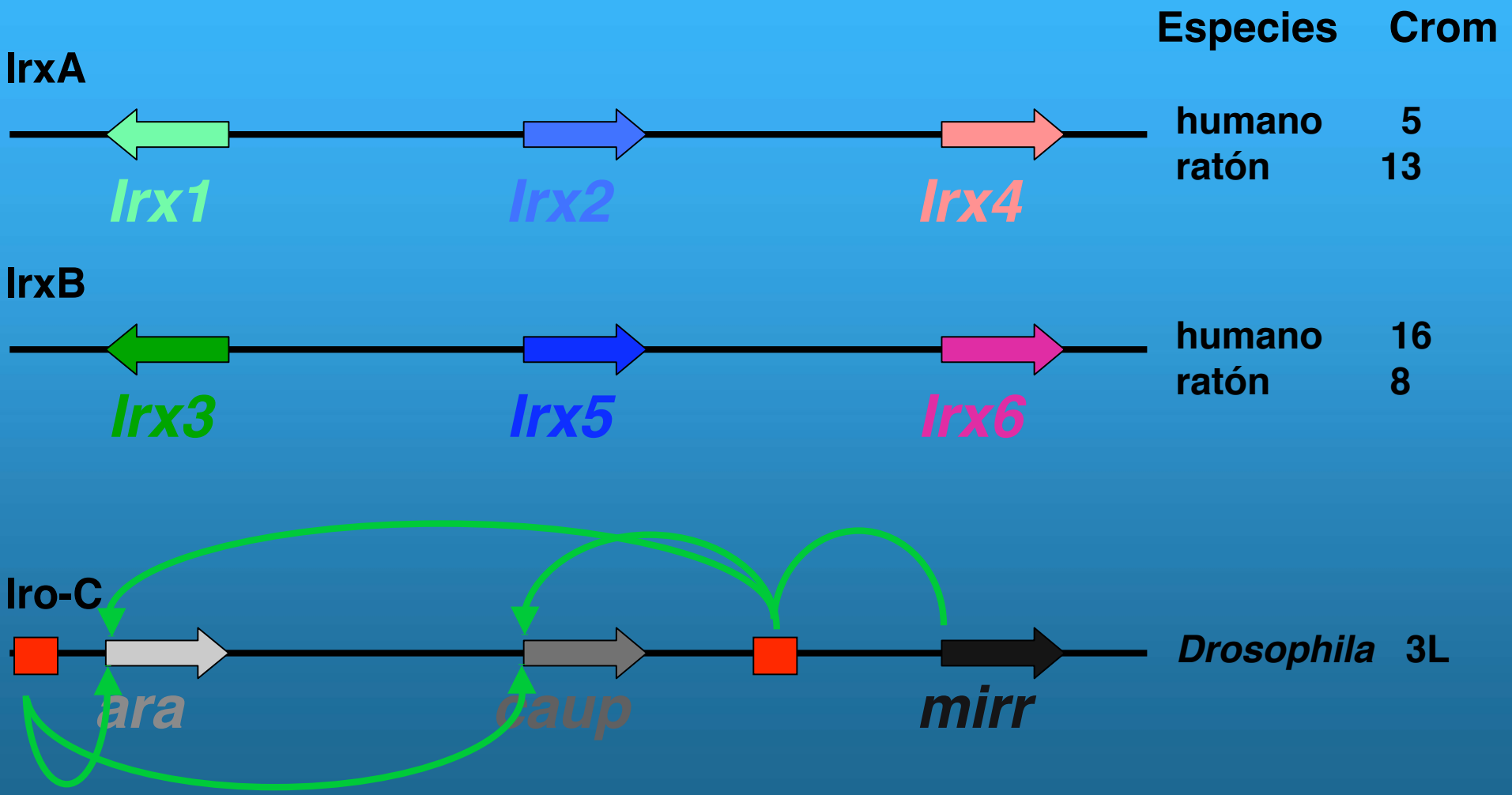
Regulación común de genes de desarrollo



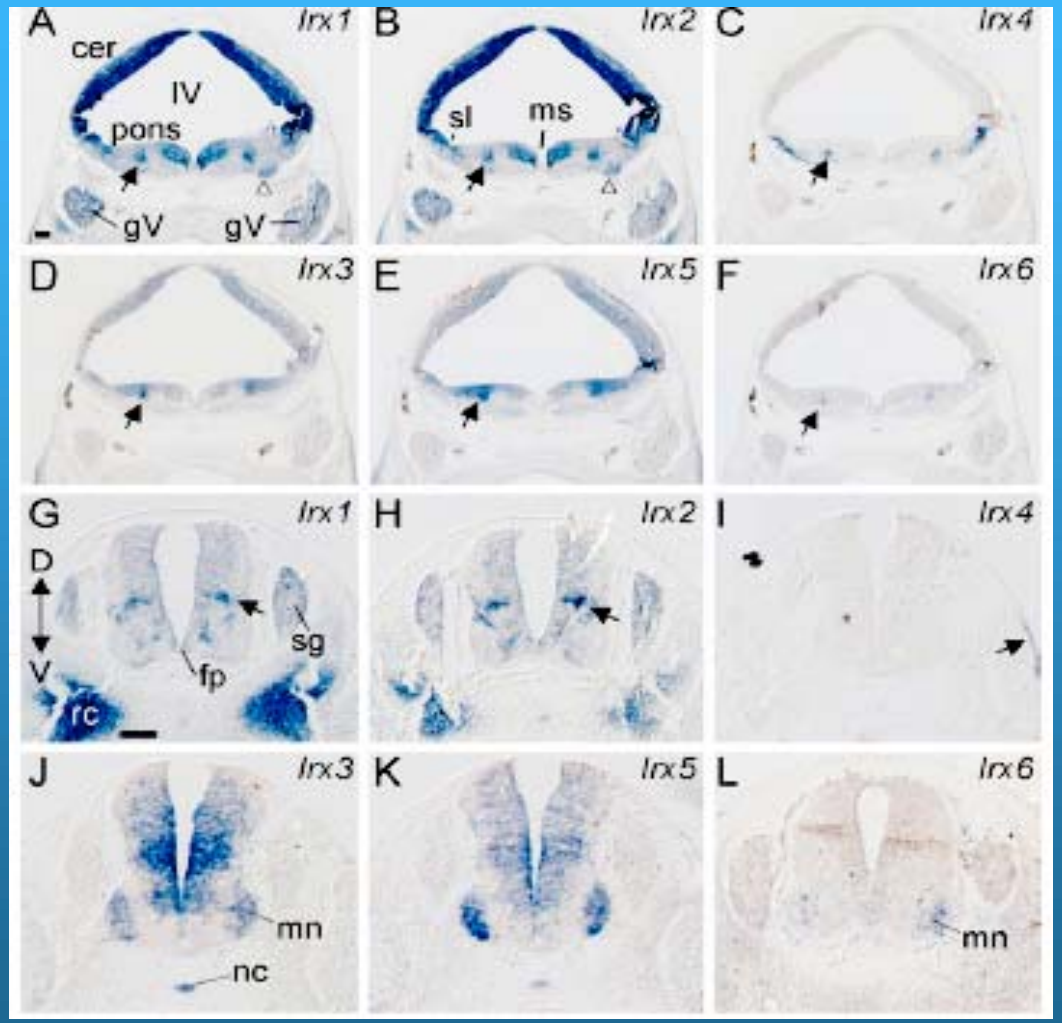
Estado filotípico donde se manifiesta el plan corporal básico

Regulación específica de taxón de genes de desarrollo

Organización genómica de los genes *Irx*



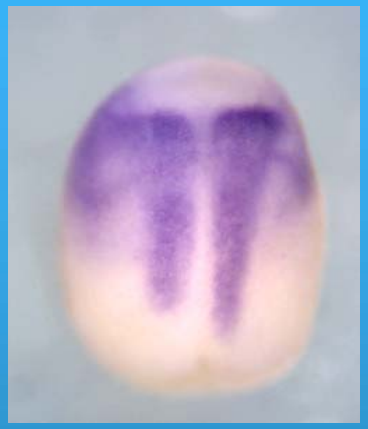
Expresión de los genes *Irx* en ratón



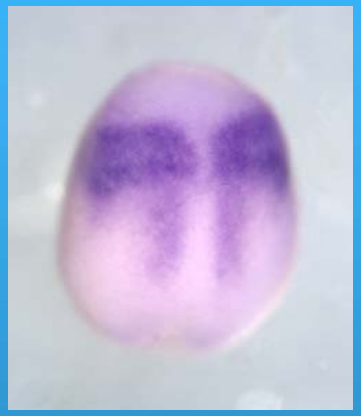
Houweling et al. Mech. Dev. 2001

Expresión de los genes *Irx* en *Xenopus*

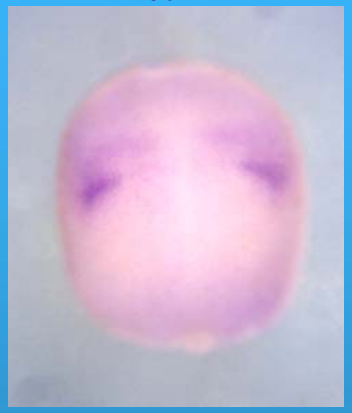
Irx1

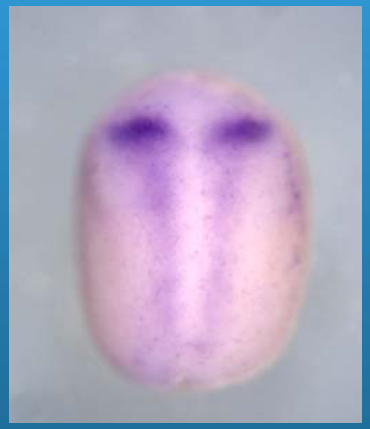


Irx2



Irx4





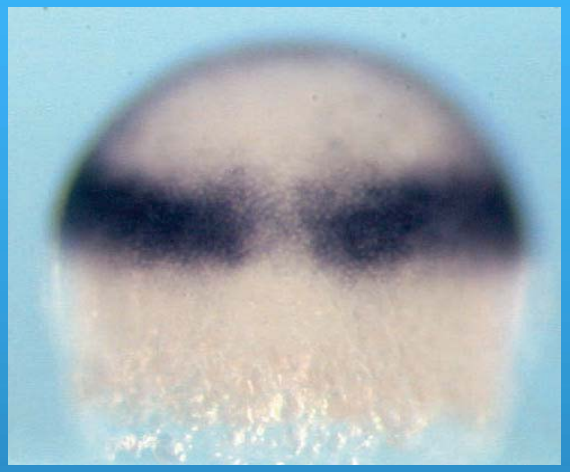
Irx3

Irx5

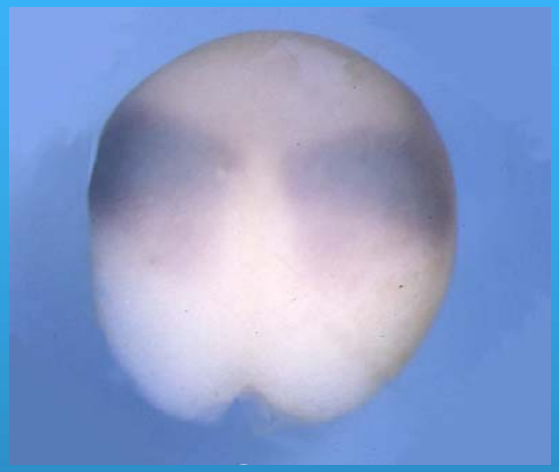
No expresado

Expresión de los genes *Irx* en diferentes vertebrados

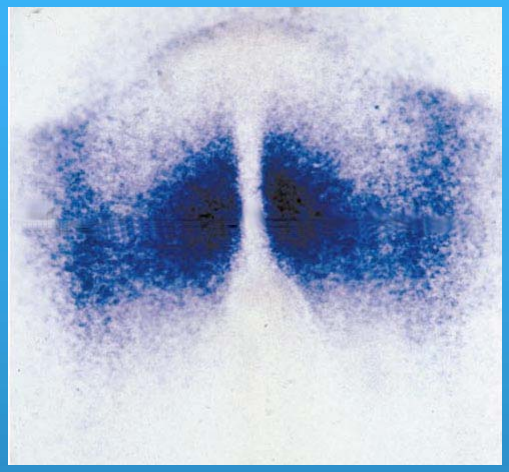
***Irx2* zebrafish**



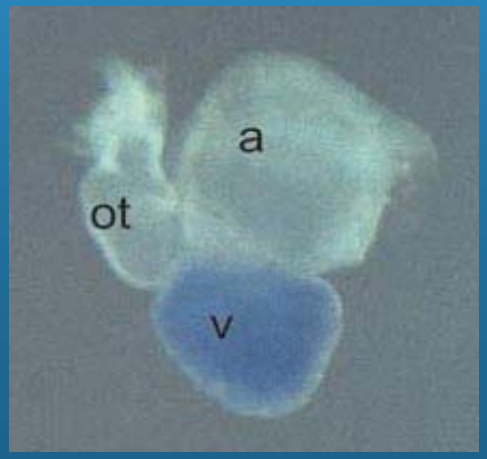
Irx2 Xenopus



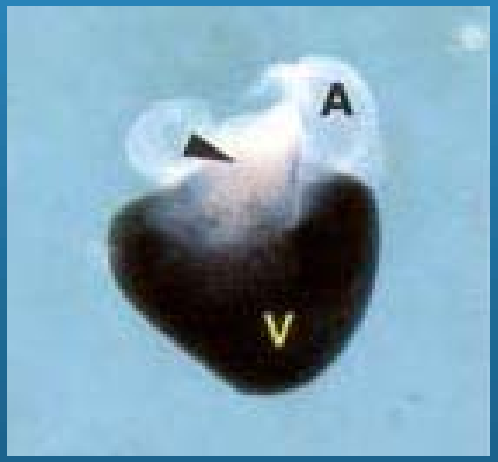
***Irx2* chick**



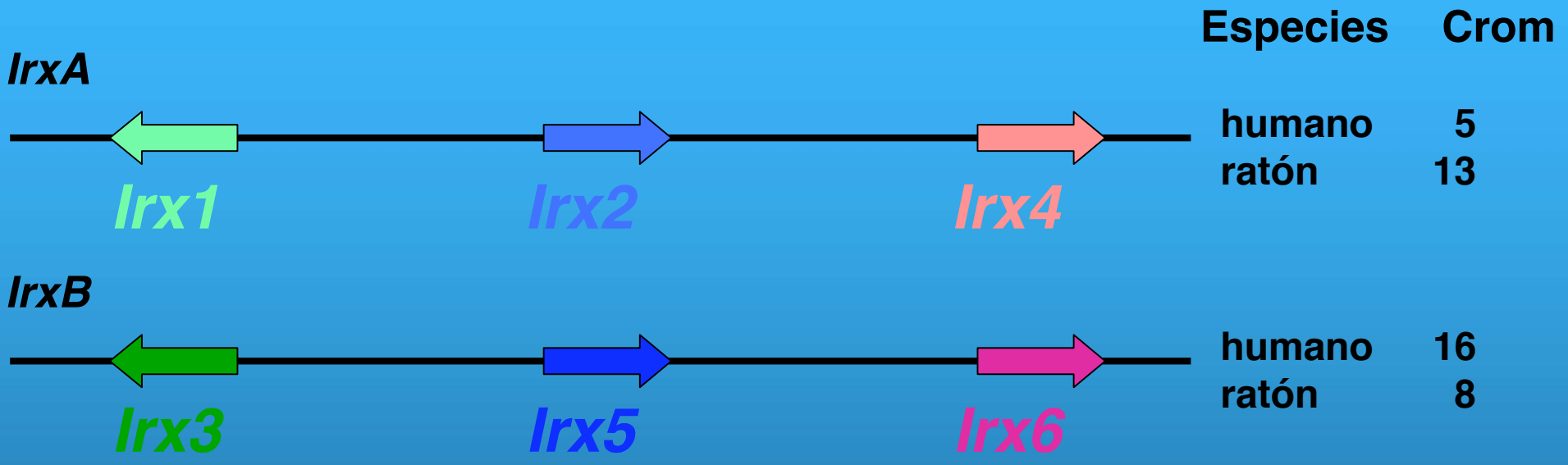
Irx4 Xenopus



***Irx4* mouse**

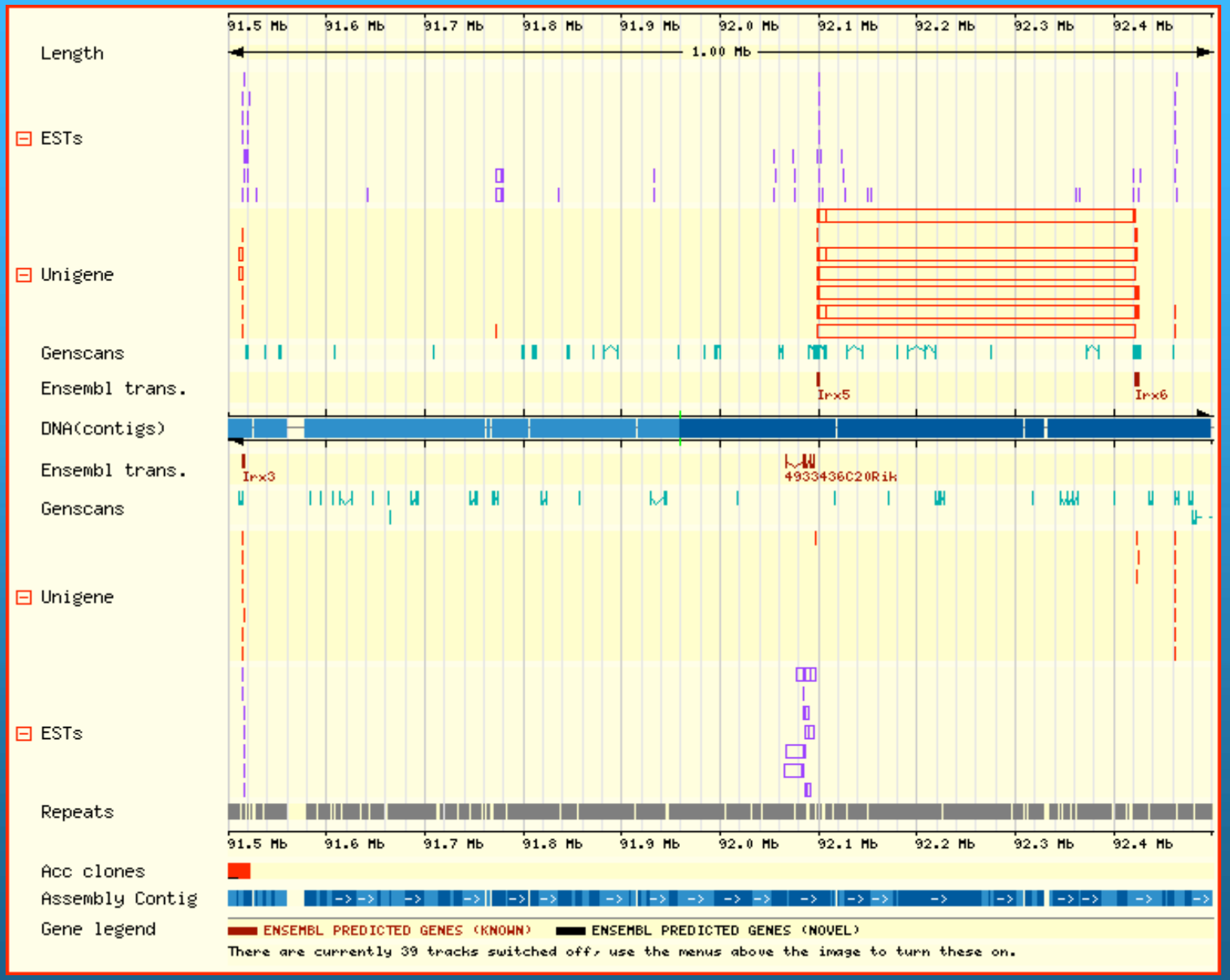


Organización genómica de los genes *Irx*

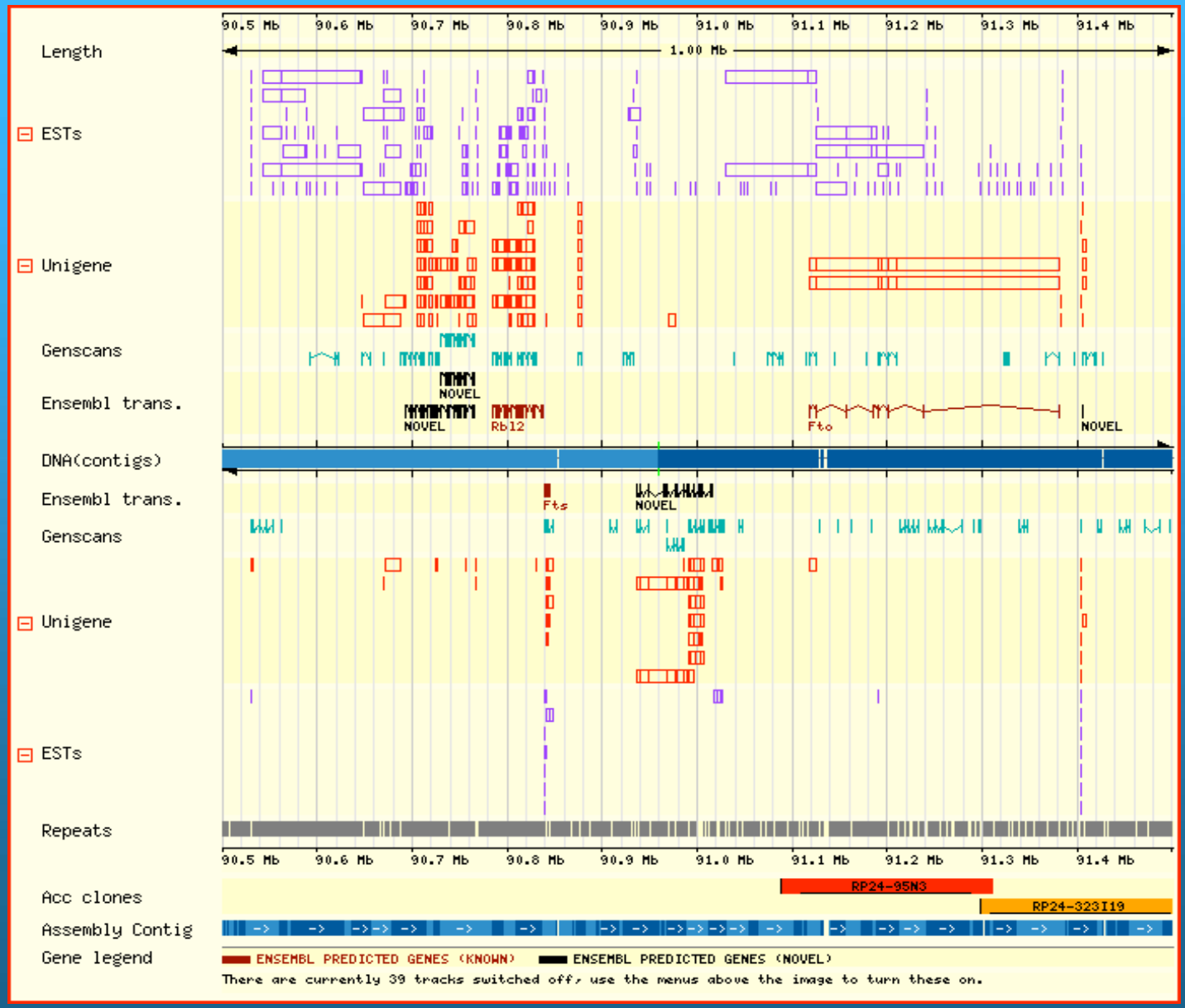


Especies	Crom
humano	5
ratón	13
humano	16
ratón	8

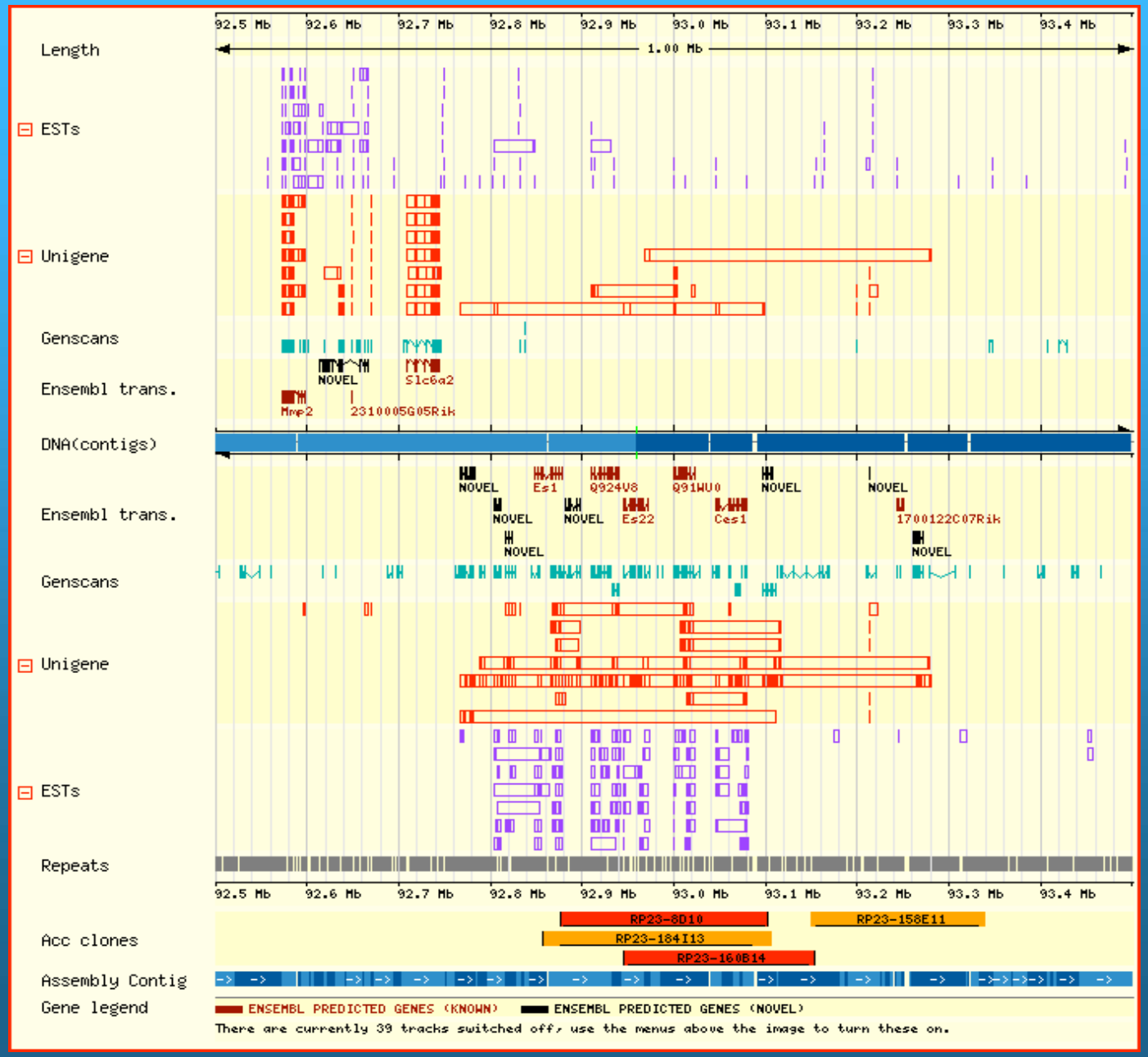
Complejo *IrxB* de ratón



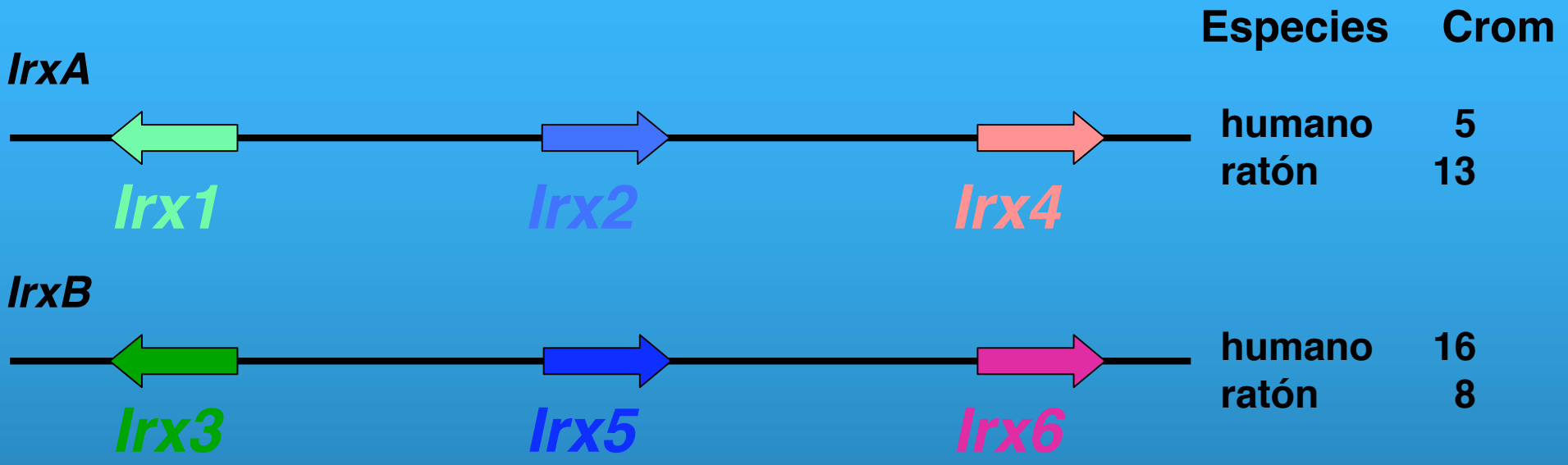
Complejo *IrxB* de ratón + 1 MB



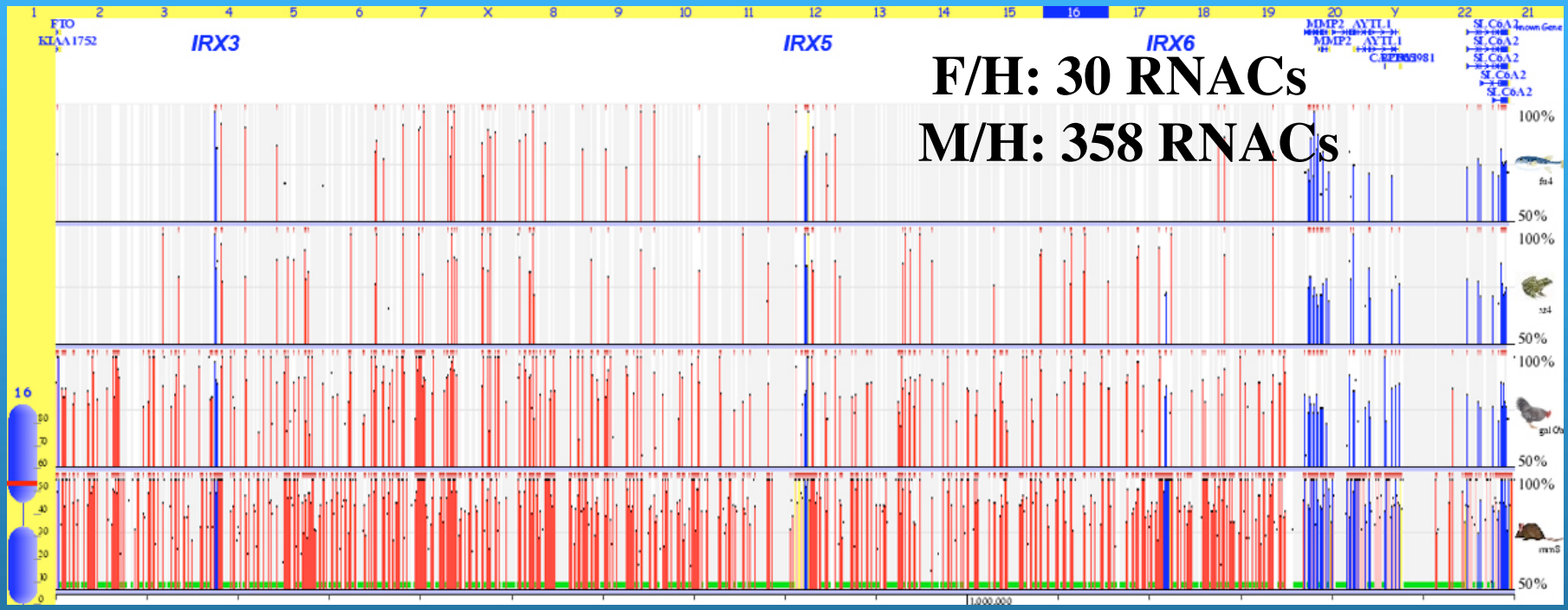
Complejo *IrxB* de ratón - 1 MB



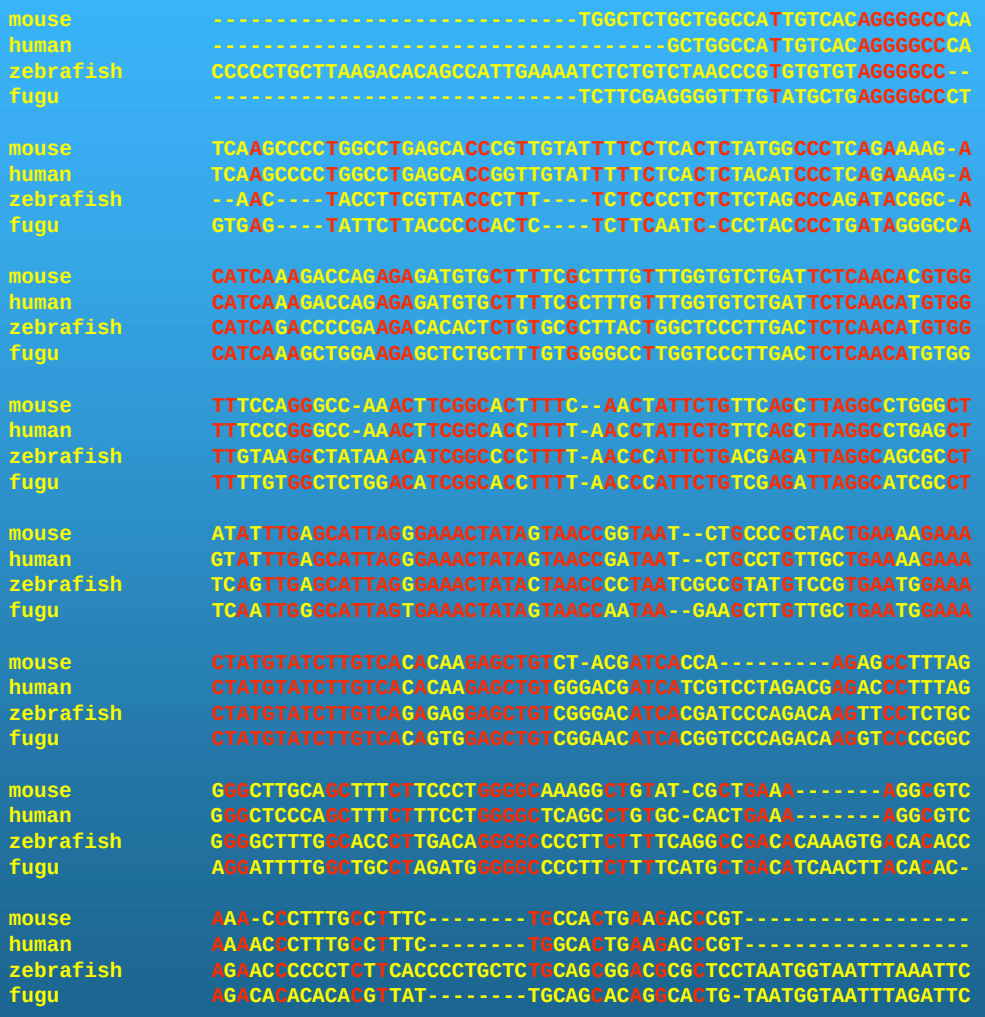
Organización genómica de los genes *Irx*



Existen multitud de regiones no codificantes altamente conservadas (RNACs) en los desiertos genómicos de los complejos *Irx*



¿Cómo son las RNACs?

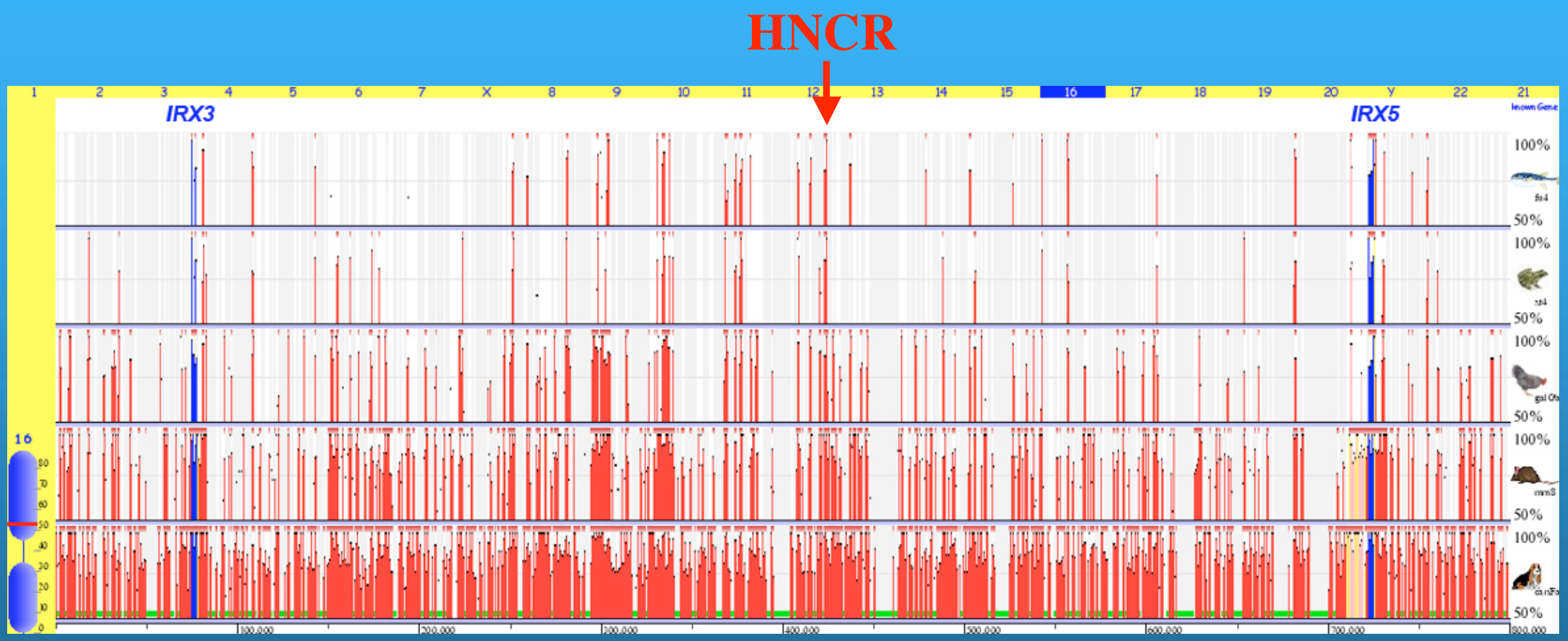


¿Son los complejos *Irx* casos singulares del genoma?

- **No, hay unas 1500-3500 RNACs entre Fugu y Humanos**
- **La mayoría están asociadas a genes esenciales para el desarrollo embrionario**
- **Muchas de ellas están localizadas en los desiertos genómicos**
- **Los complejos *Irx* contienen entre el 3-5% de todas las RNACs entre Fugu y Humanos**

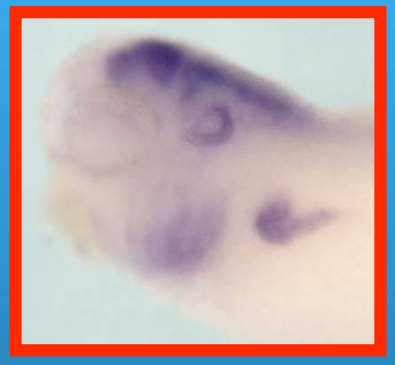
¿Cual es la función de estas RNACs?

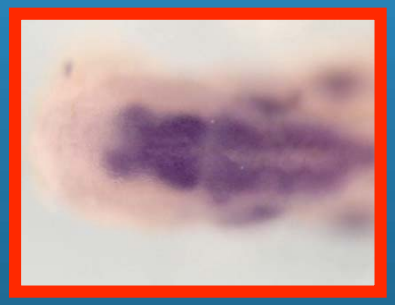
IrxB humanos/ *IrxB* ratón/ *IrxB* pollo/ *IrxB* rana/*IrxB* pez



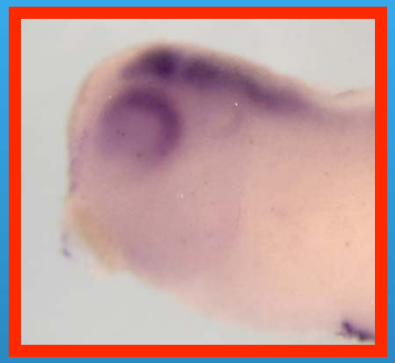
Expresión de GFP promovida por una RNAC de pez cebra en *Xenopus* y pez cebra

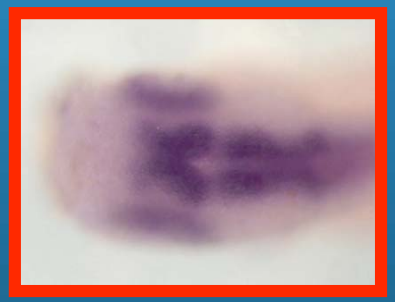
Irx3



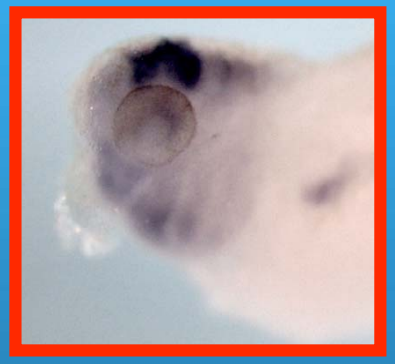


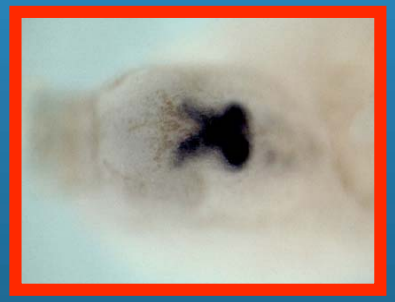
Irx5



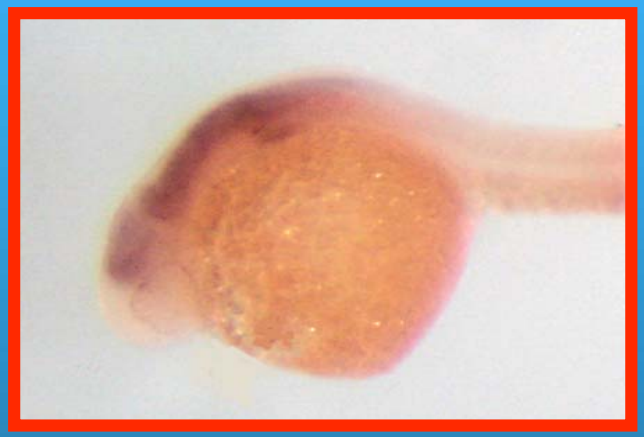


RNAC

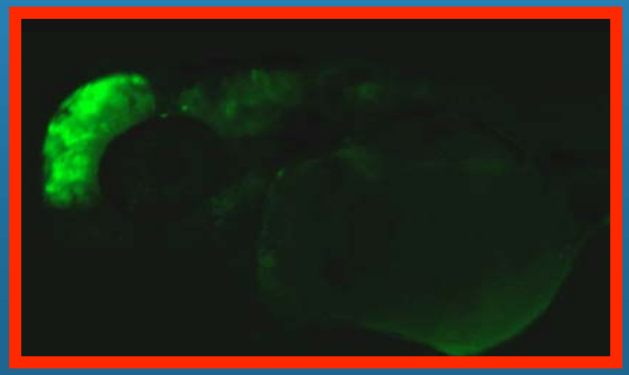




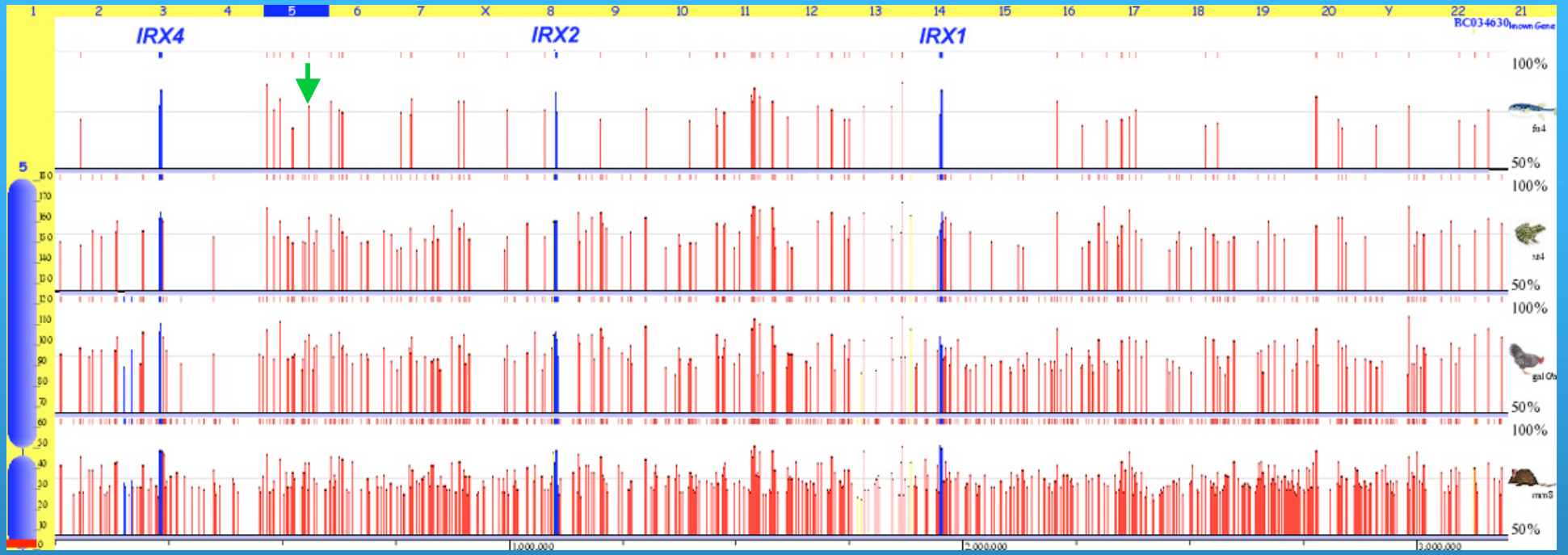
zIrx3

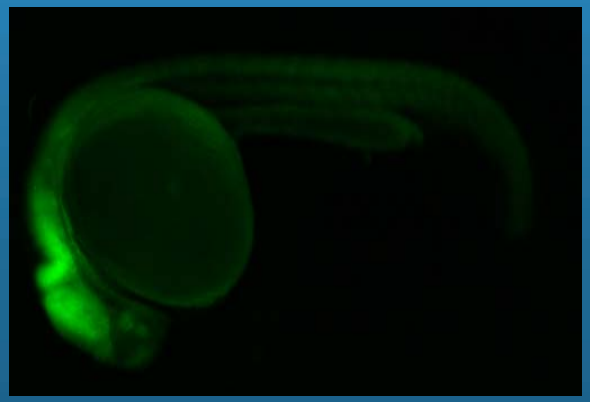


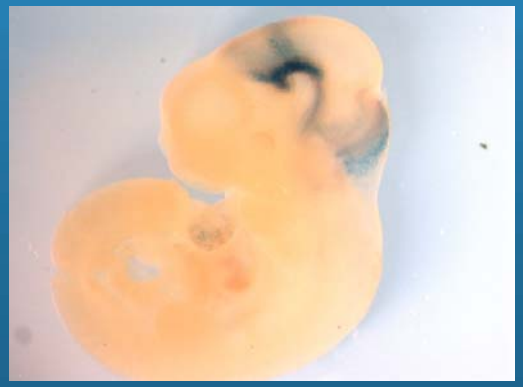
RNAC-pz*Irx3-GFP*



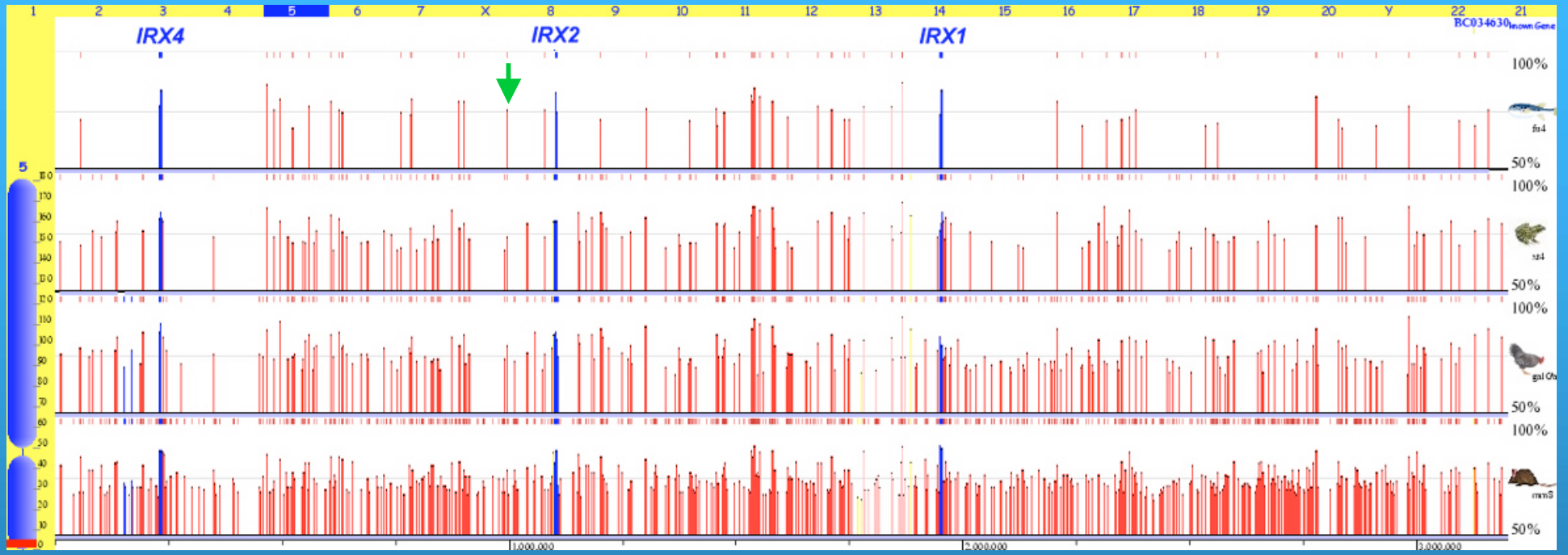
IrxA



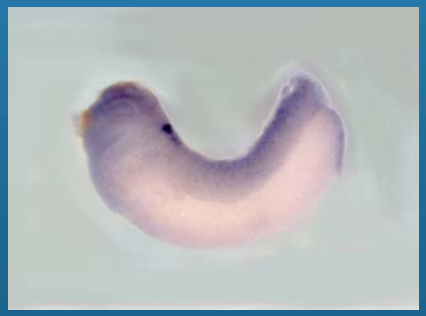




IrxA

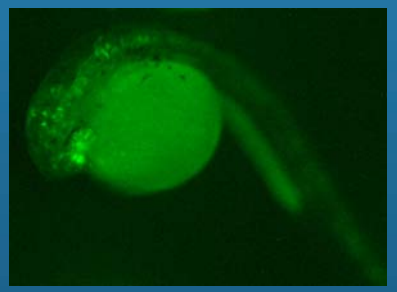


IrxA

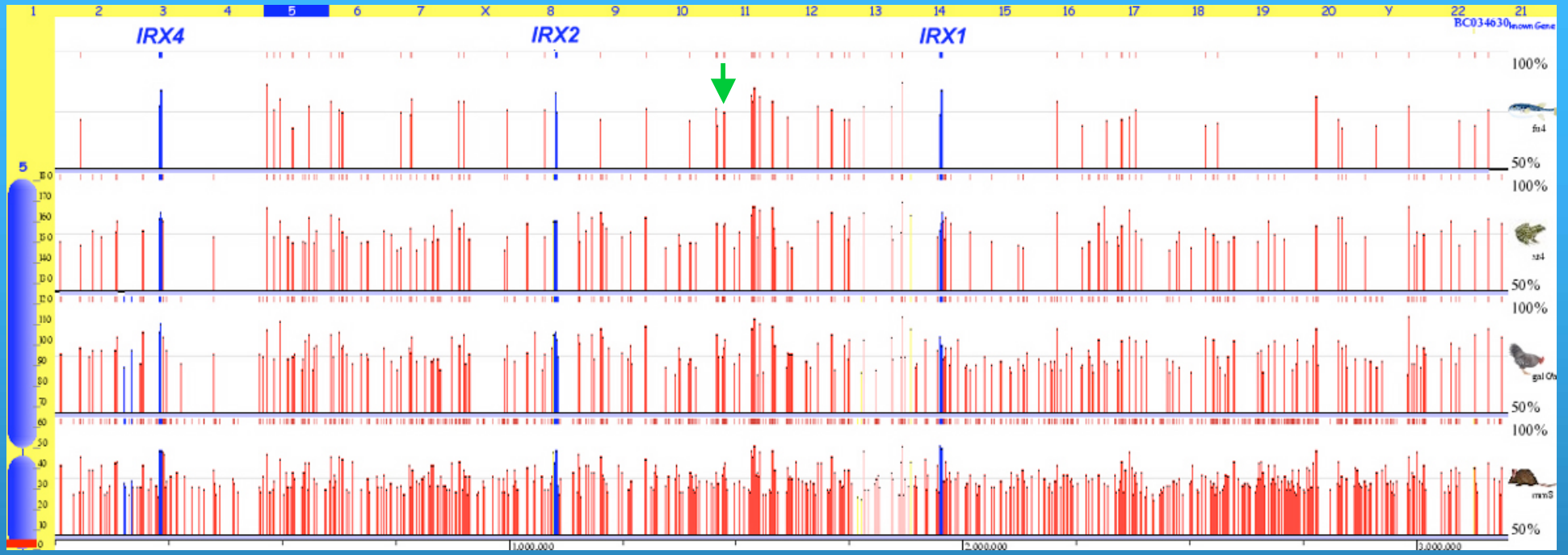


IrxB

IrxA



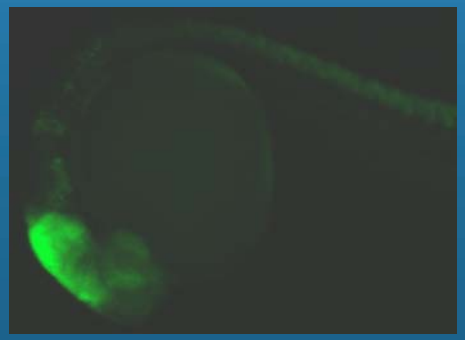
IrxA



IrxA

IrxB

IrxB



752 *IrxA*

```
Chicken   AAAATCTCACAACTTTACAGGTGGAGCCACATTGCAGCTAGTGACACAGA-CAGATGACA 106
Xenopus   AAACTCCTGCATCTTTACAGGTAGGGCCTGTTTGCAGAAAGTGACA-AGA-CTAATGACA 238
Human     AGAATCCCACCACGTTGCAGGTGGGGCC-GGCTGCTGCCAGTGGCAGAGAACGAATGACA 107
          * * **   *  * ** ***** * ***    *** *  **** ** *** *  ******

Chicken   TATTTATTAGCTTGTCAAAGCTTCCTCTGTTTATGGAAGACTCACCCAGCTAAACAAAAC 166
Xenopus   TGCTTATTAGCTTGTCAAACCCTCCCCAGTTTATAGAAGGCTCACCCAGCTAAACAAAAC 298
Human     TATTTATTAGCTTGTCAAAGCCTCCCCTGTTTATGGAAGGCTCACCCAGCTAAACAAAAC 167
          *  ***************** * *** * ****** **** ********************

Chicken   AGTCTGTTTCCCATTGGCTAATTGAGCTGGCCCCTGTTGAGAGTCTCAGCAAAGTGCTTG 226
Xenopus   AGTCTGTTTTCTATTGGCTTATTGGGTTTGTTCCTGTTGGGAATCTCAGTACAGTGCTTG 358
Human     AGTCTGCTTCCCATTGGCTCAGCCCAG-GGCCCCTGTTGAGAGCTTCAGCAAAGTGCTTG 226
          ****** ** * ******* *         *  ******* **   **** * ********

Chicken   ATGAGGCATAAAATGATGATTAATTAATTTACTGCTCAGTTGATCATCAAACTGAGAGGT 286
Xenopus   ATGAGGCATAAAATGATGATTAATTAATTTACTGCACAGTTGATCATCAAACAGAGAGGT 418
Human     ATGAGGCATAAAATGATGATTAATTAATTTACTGCTCAGCTGATCATCACACTGCGAGGT 286
          ********************************** *** ********* ** * *****

Chicken   TTATGATTCTCACAGCGGTCAACACTCTCTCTTCAGTGTGCAGCAGCATGAGCAGCCTTC 346
Xenopus   TTATGATTCACACAGCGGTCAACACTCCCACTTCAATGTGCCACAGCATGATTAGTTTTA 478
Human     TTATGGTTCTCCCAGCGGTCAGCCCGCGCTCTGCGGTGCGTGGCCGC---GGCTGCCTTT 343
          ***** *** * ********* * * * * ** *  ** *   * **       *  **
```

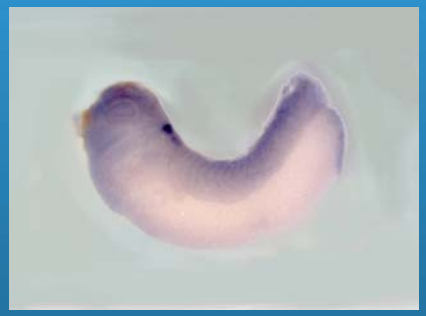
752 *IrxB*

```
Chicken   ----------------------GTTCGCACTTGCTGCACTTCTGATTCAGAAAGGGCTG 37
Xenopus   CTTAAAACGATGTGCACTTTATTGTTGGCACTTGCTGAGCTTCAGAGACAGAAAGGGCTG 60
Human     -------------------GCAGGGCGTCCCGGCCACGCCCC--GTCCCCAACGCGCTG 38
                                 *   *  *  **    *  *     *  ** * ****

Chicken   ATTTATTAGCTTGTCAAGGGAAGCATTGTTCAGGAAGGCAATCAAAAAGCACTTACCTAG 97
Xenopus   ATTTATTAGCTTGTCAAGGGAAGCATTGTTGAAGAAGGCAATCAAAAAGCATTTACCTAG 120
Human     ATGGATCGGCTTGTCAGGAGA-GCTCTGTGCGCGGAGCCCATCAGAAAGCACTTACCTCG 97
          **  **  ******** * ** **  ***    * ** * **** ****** ****** *

Chicken   CTAAACTAAAATTACTGCCTTTTCAGTGGCTAATTTGTATAACTCCCATGGAGGAACTTG 157
Xenopus   CTAAACTAAAATTACTGCCTTTTCAGTGGCTAATTTGTATAACTCCCATGAAGGAACTTG 180
Human     CTAAACTGAAATTACTGTCTTTTCAGGGGCTAATTTGTGTAACTCCCATGGAGGGGCTTG 157
          ******* ********* ******** ************ ************ ***  ****

Chicken   GGAAGTGTTTGATGAGGCATAAAATGATGATTAATGAATTTATTGGCCTGCTGCTCATCA 217
Xenopus   GAAAGTGTTTGATGAGGCATAAAATGATGATTAATGAATTTATTGCCTTGCTTCTCATAA 240
Human     GAGGGGGTTTGATGAGGCATAAAATGATGATTAATGAATTTATTGGCC-GCTGCTCATCA 216
          *   * ***************************************** *  *** ***** *

Chicken   AATA-GGAGTAGT-ATATCTT-TCCTCTATTGGGCGTTGGGGCTAATGCTACATGGCTGT 274
Xenopus   AACACAAAGTGGT-ATCTCAC-TCTCCCTTTGTAT-TTCTCTCTCTCTTTCTTTATCT 297
Human     AACAAGGACTGGAGAGTTCGCGTCTGCACCCGCGGCCTGGCGGCGGTGCTCAACACAACT 276
          ** *   * * *  *  **   **  *   *     *          **        *
```

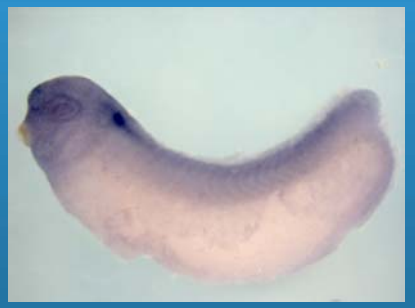
752 *IrxA vs IrxB*

```
Chicken_IrxB   TGTCAAGGGAAGCATTGTTCAGGAAGGCAATCAAAAAGCACTTACCTAGCTAAACTAAAA 108
Xenopus_IrxB   TGTCAAGGGAAGCATTGTTGAAGAAGGCAATCAAAAAGCATTTACCTAGCTAAACTAAAA 131
Human_IrxB     TGTCAGGAGA-GCTCTGTGCGCGGAGCCCATCAGAAAGCACTTACCTCGCTAAACTGAAA 108
Chicken_IrxA   TATTTATTAGCTTGTCAAAGCTTCCTCTGTTTATGGAAGACTCACCCAGCTAAACAAAAC 166
Xenopus_IrxA   TGCTTATTAGCTTGTCAAACCCTCCCCAGTTTATAGAAGGCTCACCCAGCTAAACAAAAC 298
Human_IrxA     TATTTATTAGCTTGTCAAAGCCTCCCCTGTTTATGGAAGGCTCACCCAGCTAAACAAAAC 167
               *                              * *   *    * ***  *******  **

Chicken_IrxB   TTACTGCCTTTTCAGTGGCTAATTTGTATAACTCCCATGGAGGAACTTGG-GAAGTGTTT 167
Xenopus_IrxB   TTACTGCCTTTTCAGTGGCTAATTTGTATAACTCCCATGAAGGAACTTGG-AAAGTGTTT 190
Human_IrxB     TTACTGTCTTTTCAGGGGCTAATTTGTGTAACTCCCATGGAGGGGCTTGG-AGGGGGTTT 167
Chicken_IrxA   AGTCTGT-TTCCCATTGGCTAATTGAGCTGGCCCCTGTTGAGAGTCTCAGCAAAGTGCTT 225
Xenopus_IrxA   AGTCTGT-TTTCTATTGGCTTATTGGGTTTGTTCCTGTTGGGAATCTCAGTACAGTGCTT 357
Human_IrxA     AGTCTGC-TTCCCATTGGCTCAGCCCAG-GGCCCCTGTTGAGAGCTTCAGCAAAGTGCTT 225
                  ***  **   *  **** *           **  *   *    *  *    * * **

Chicken_IrxB   GATGAGGCATAAAATGATGATTAATGAATTTATTGGCCTGCTGCTCATCAAATA--GGAG 225
Xenopus_IrxB   GATGAGGCATAAAATGATGATTAATGAATTTATTGCCTTGCTTCTCATAAAACAC-AAAG 249
Human_IrxB     GATGAGGCATAAAATGATGATTAATGAATTTATTGGCC-GCTGCTCATCAAACAA-GGAC 225
Chicken_IrxA   GATGAGGCATAAAATGATGATTAATTAATTTACTGCTCAGTTGATCATCAAACTGAGAGG 285
Xenopus_IrxA   GATGAGGCATAAAATGATGATTAATTAATTTACTGCACAGTTGATCATCAAACAGAGAGG 417
Human_IrxA     GATGAGGCATAAAATGATGATTAATTAATTTACTGCTCAGCTGATCATCACACTGCGAGG 285
               ************************ ****** **    * *  **** * *
```

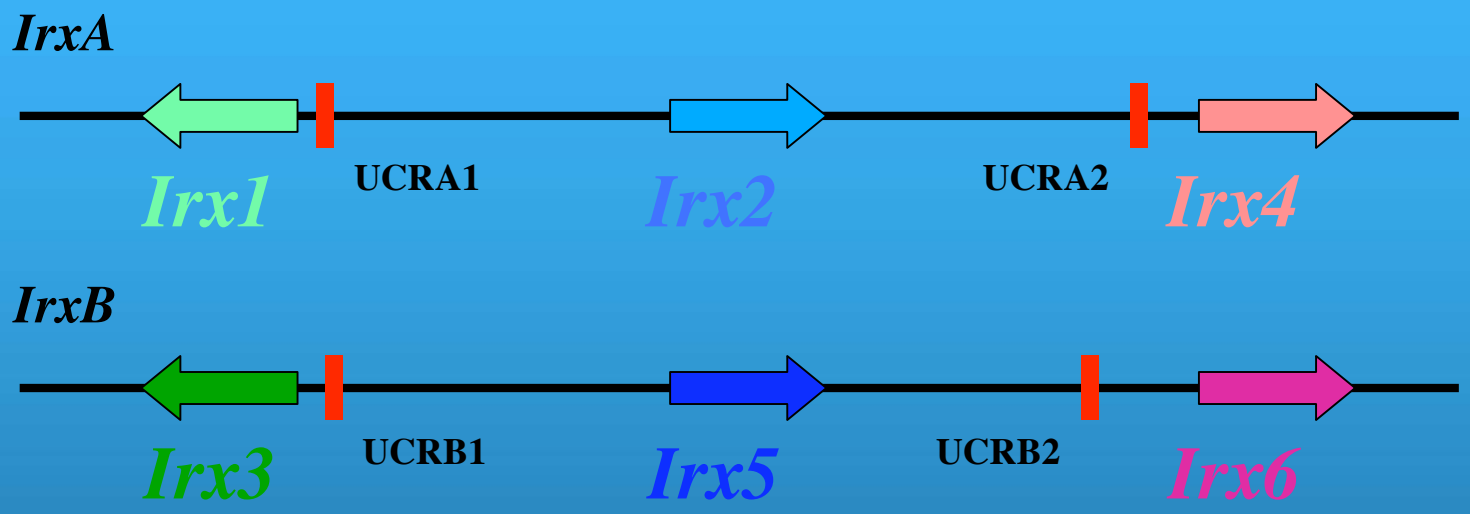
IrxA



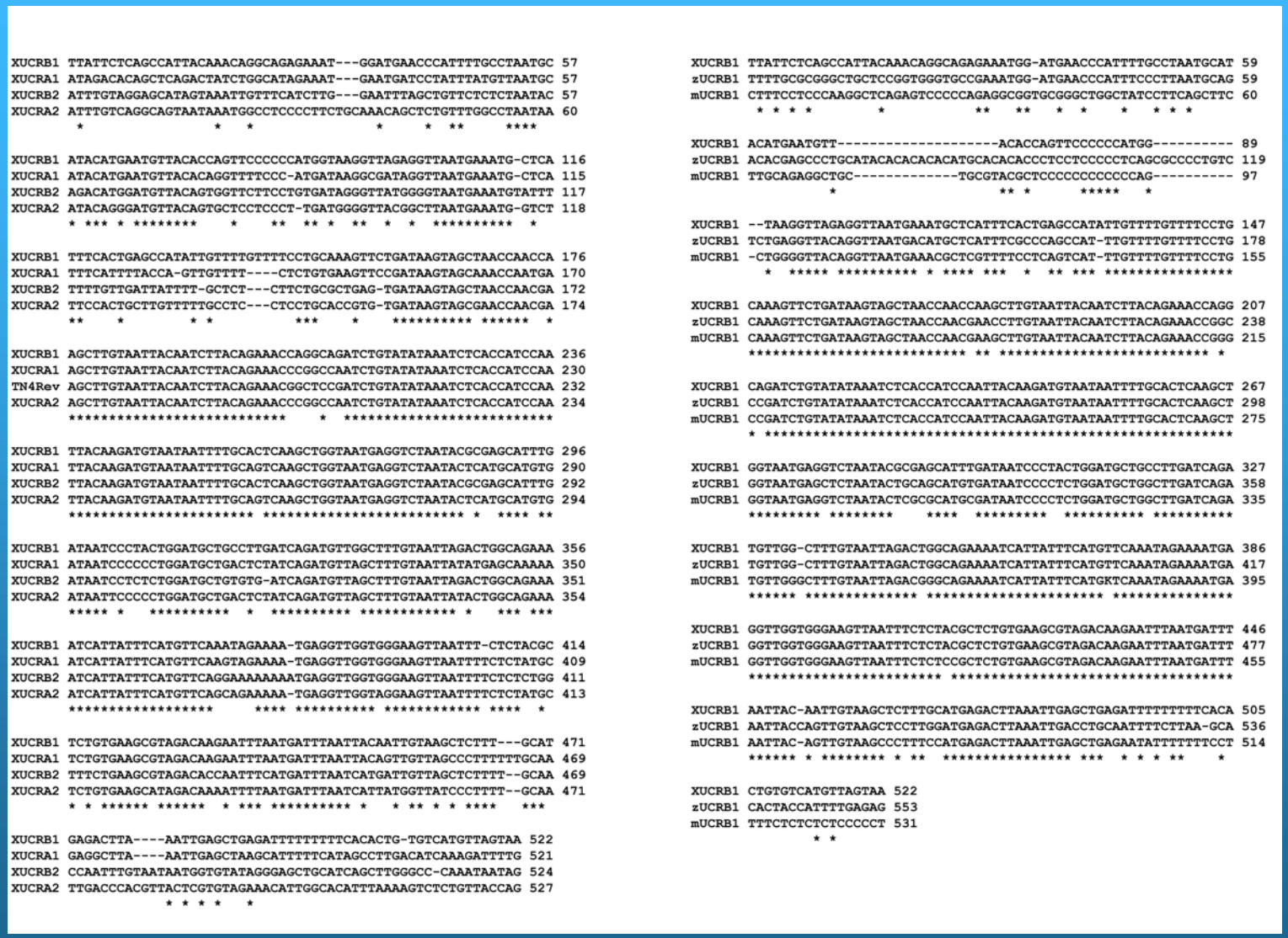
IrxB



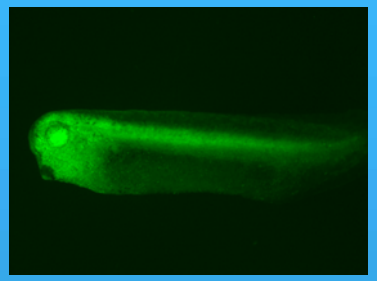
Regiones ultraconservadas en los complejos *Irx*



Regiones ultraconservadas de los complejos *Irx*



Regiones ultraconservadas de los complejos *Irx*

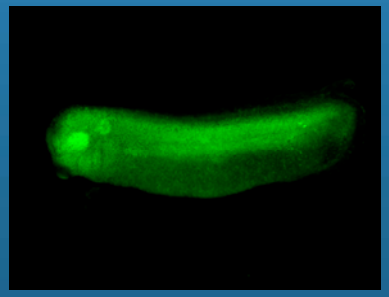


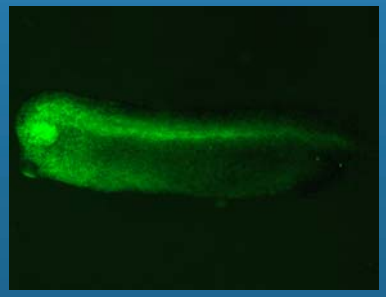
***IrxA* cluster**

Irx1 UCRA1 ***Irx2*** UCRA2 ***Irx4***

***IrxB* cluster**

Irx3 UCRB1 ***Irx5*** UCRB2 ***Irx6***





El contexto genómico modula la actividad de las regiones UCs

***IrxA* cluster**

Irx1 UCRA1 ***Irx2*** UCRA2 ***Irx4***

***IrxB* cluster**

Irx3 UCRB1 ***Irx5*** UCRB2 ***Irx6***

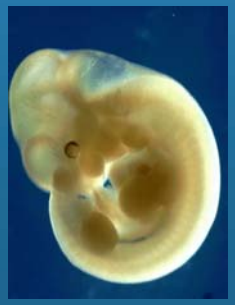
El contexto genómico modula la actividad de las regiones UCs

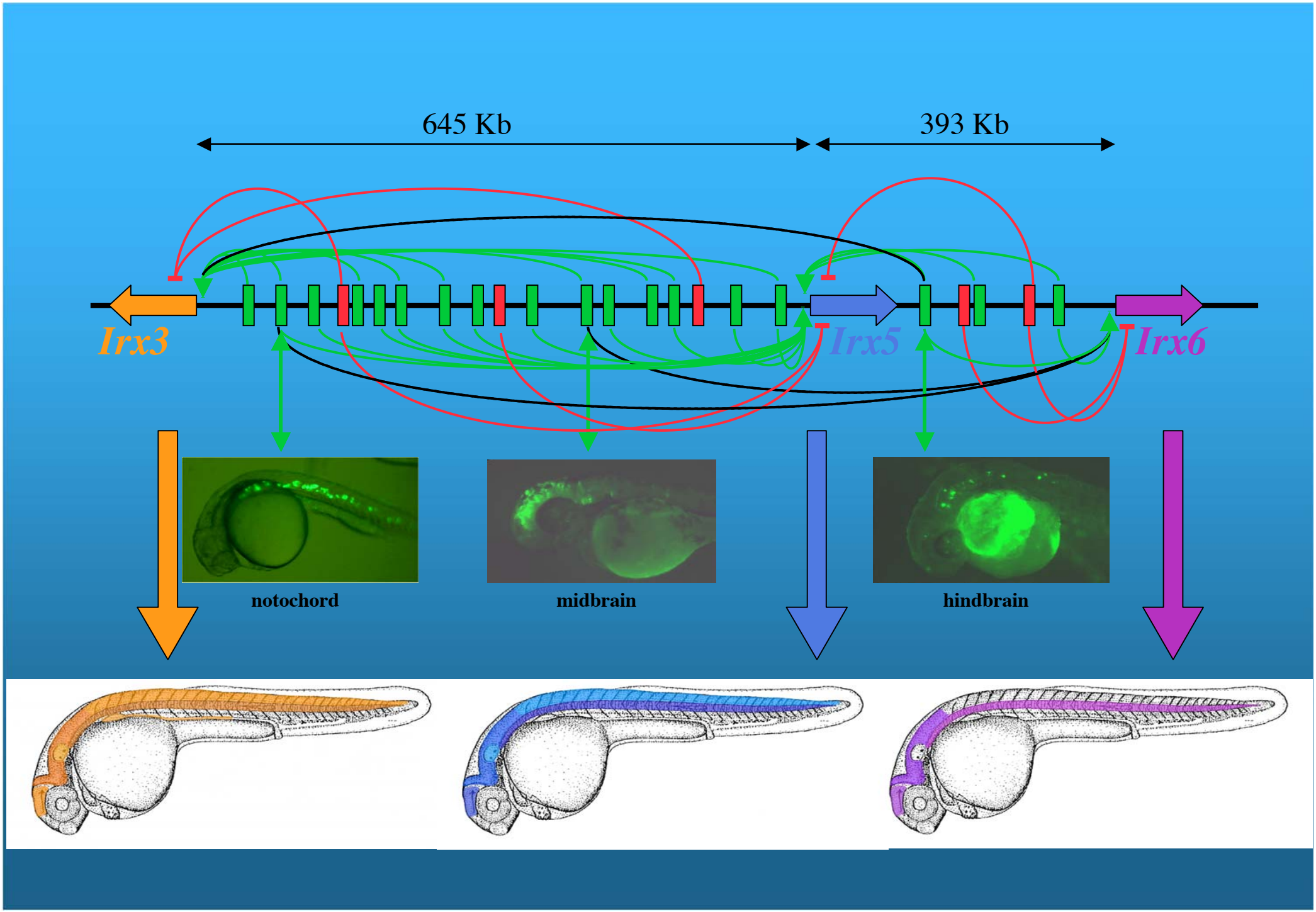
***IrxA* cluster**

Irx1 UCRA1 ***Irx2*** UCRA2 ***Irx4***

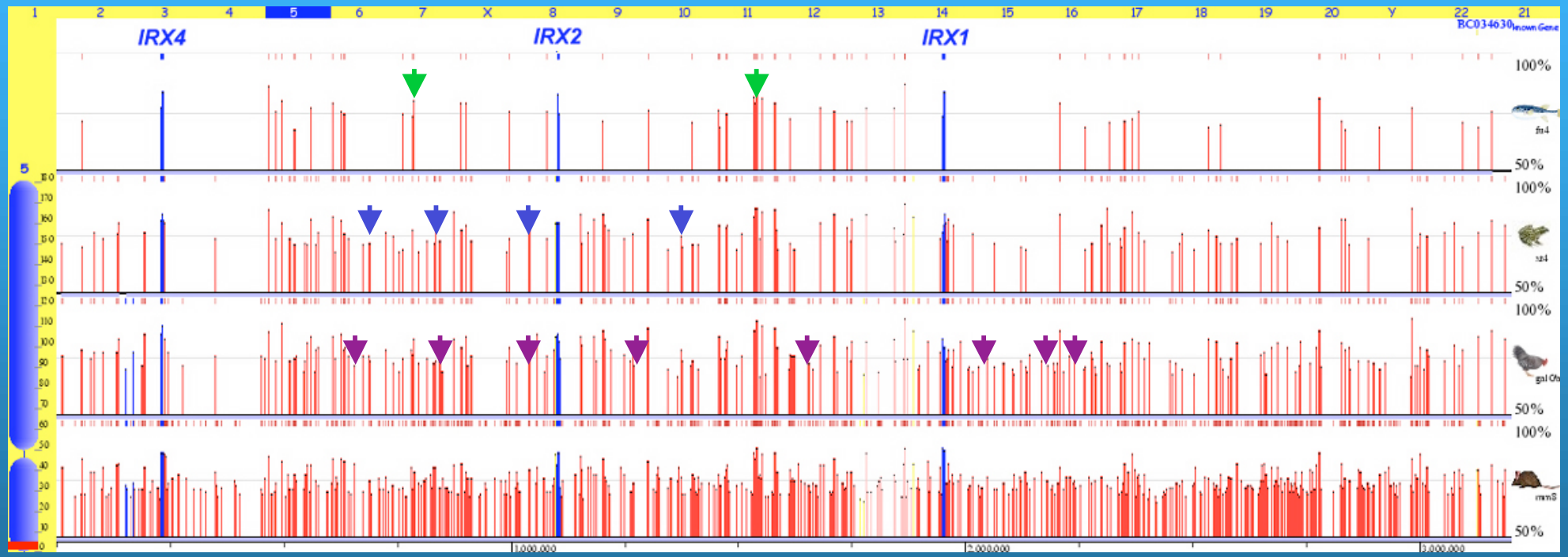
***IrxB* cluster**

Irx3 UCRB1 ***Irx5*** UCRB2 ***Irx6***




645 Kb
393 Kb
Irx3
Irx5
Irx6
notochord
midbrain
hindbrain

Nuevas RNACs aparecidas en la evolución pueden contribuir a innovaciones corporales a través de nuevos patrones de expresión



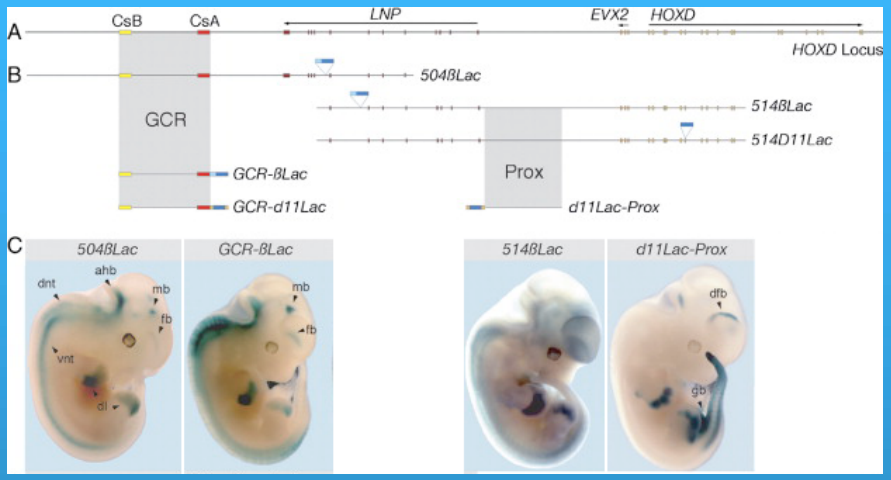
Conclusiones

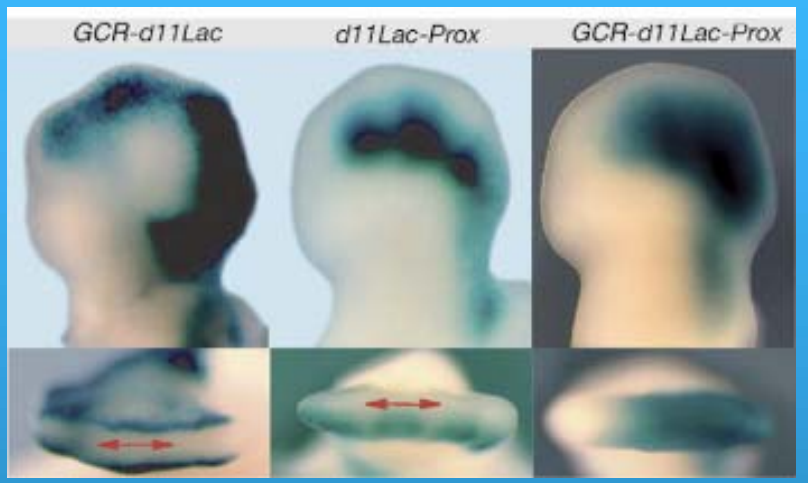
- **La mayoría de las RNACs de los desiertos genómicos del complejo *IrxB* contienen "enhancers" que activan la expresión en subdominios de los genes *IrxB***
- **Algunos de los "enhancers" son redundantes**
- **El patrón de expresión de los genes *Irx* se construye por la acción combinada de múltiples elementos cis-reguladores**
- **Algunas de las HCNRs pueden contener elementos silenciadores**
- **La existencia de regiones reguladoras compartidas podría ser responsable de mantener a los genes *Irx* organizados en complejos**
- **El contexto genómico puede modular la actividad enhancer de las RNAC**
- **El incremento del número de RNACs en vertebrados más modernos puede corresponder a nuevos "enhancers" para nuevos dominios de expresión**
- **Un mecanismo evolutivo para generar nuevas formas es añadir nuevos modulos reguladores a genes importantes en procesos de desarrollo para que estos adquieran nuevas funciones**

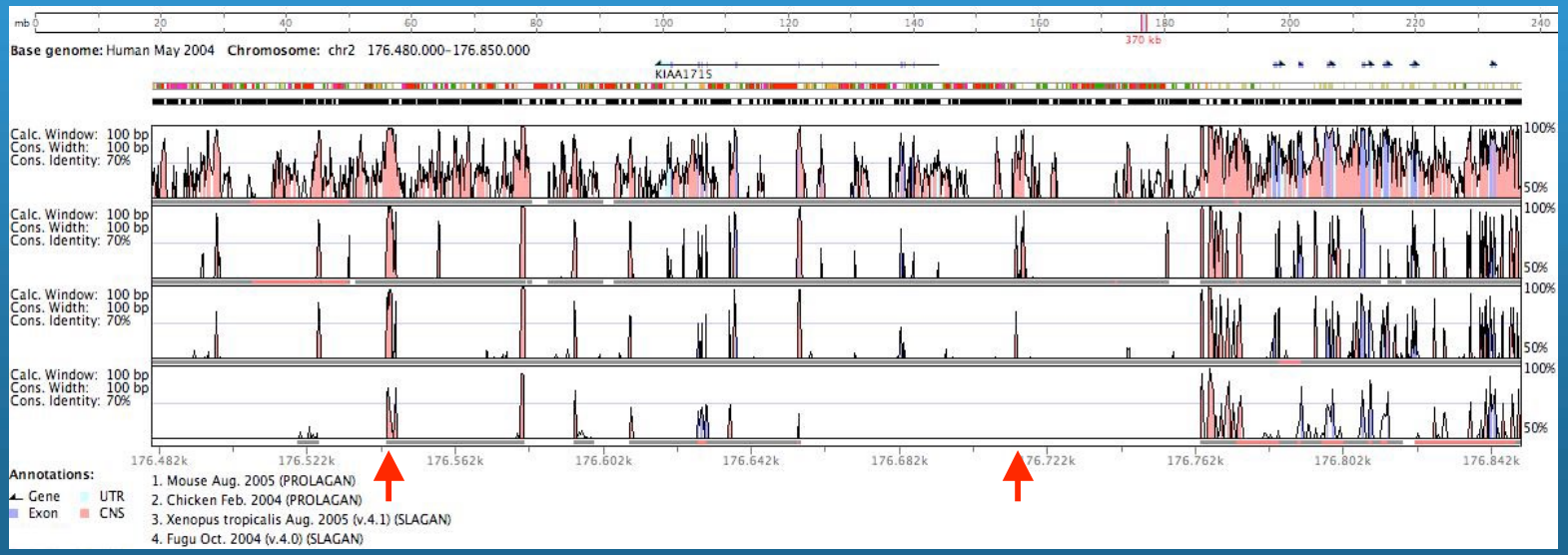
Conclusiones genómicas

- **Existen 3000-4000 RNACs comunes a todos los vertebrados**
- **Estas regiones se encuentran en regiones sinténicas alrededor de genes del desarrollo formando bloques de regulación genómica**
- **Existen conjuntos independientes de RNACs en otros linages (mosca, gusanos)**
- **La conservación implica función. Muchas RNACs contienen elementos reguladores**
- **Cada conjunto de RNACs podría contener los elementos reguladores necesarios para los diferentes planes corporales**
- **La aparición de nuevos RNACs en la evolución está asociado a innovaciones en los planes corporales**
- **Puede existir relación entre RNACs y enfermedades genéticas humanas**

Aparición de los dedos en tetrápodos durante la evolución

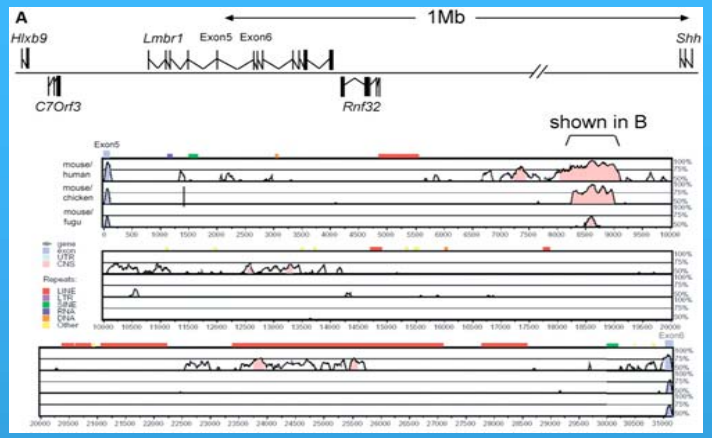


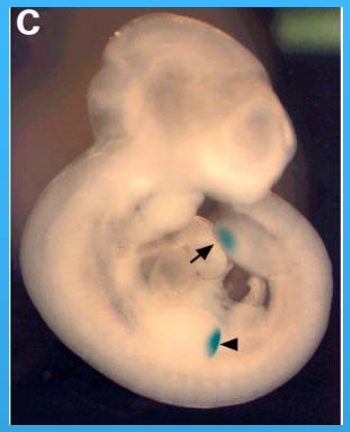


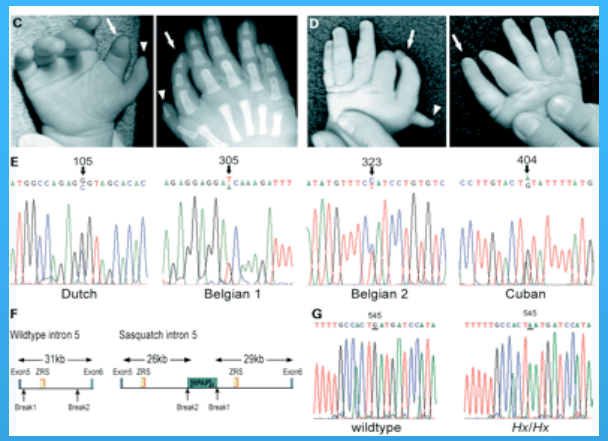


Transgenic analysis of the *Hoxd* gene regulation during digit development. González et al. (2007). Dev. Biol. 306, 847-859

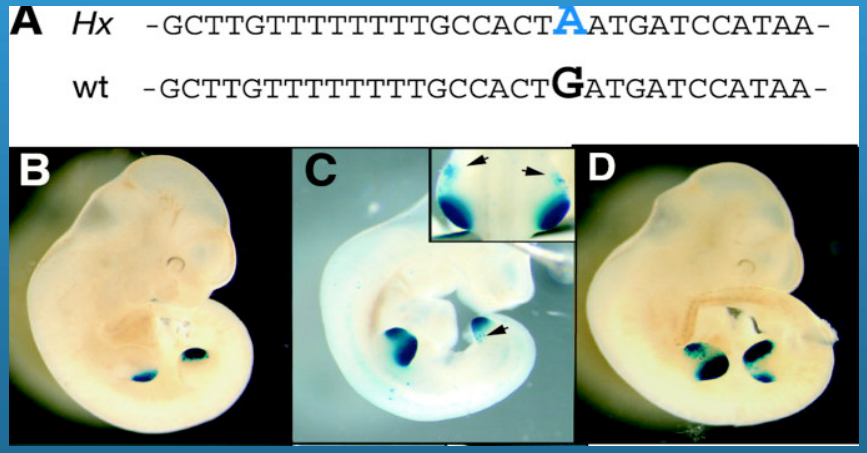
A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Lettice et al. (2003). Hum. Mol. Genet. 12, 1725-1735.







Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. Maas and Fallon (2005). Dev. Dyn. 232, 345-348



Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Sagai et al. (2005). Development 132, 797-803

PROJECT: DeVelome

Deciphering the vertebrate regulatory genome: diagnostics for human genetic disease

Consortio Preliminar

Thomas Becker, Sars Centre Norway
José Luis Gómez-Skarmeta, CSIC Spain
Boris Lenhard, Sars Centre Norway
Fernando Casares, CSIC Spain
Miguel Manzanares, CSIC Spain
Greg Elgar, University of London, UK
Denis Doubule, University of Geneva, Switzerland
Patrick Charnay, Ecole Normale Supérieure, France
Veronica van Heyningen, MRC Human Genetics Unit, UK
Antonio Jacinto, Instituto Gulbenkian, Portugal
Uwe Strähle, Karlsruhe, Germany

PRINCIPIO:

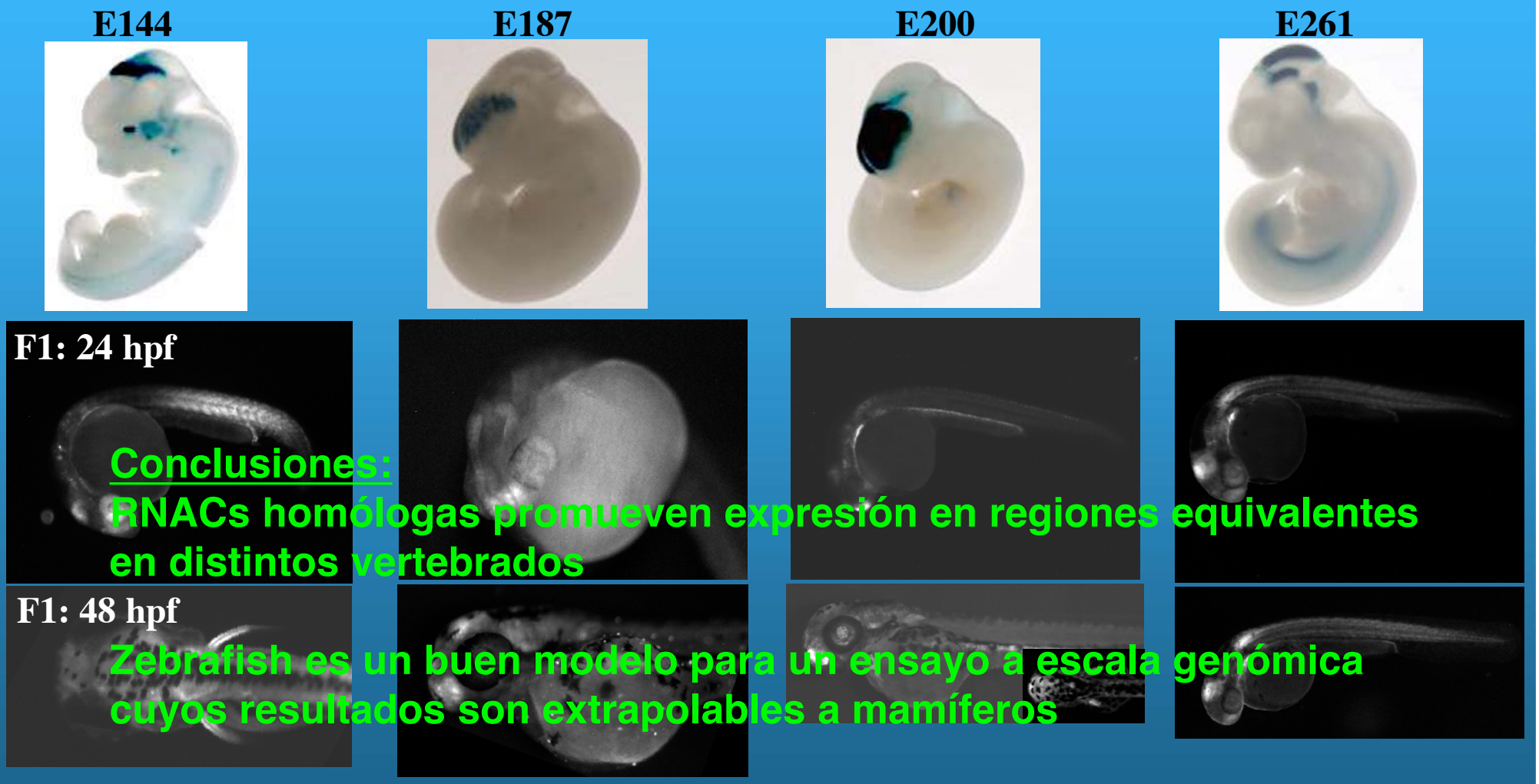
La colección de RNACs contiene el kit de elementos reguladores para generar el plan corporal de vertebrados

OBJECTIVOS:

1. **Ensayar la actividad enhancer de las 3.000 RNACs presentes en todos los vertebrados. Esta colección representará el kit de elementos reguladores necesarios para construir un vertebrado**
2. **Generar *líneas transgénicas estables* para todas las regiones con actividad enhancer**
3. **Estudiar la asociación de enfermedades genéticas humanas a las RNACs**
4. **Intentar descifrar el lenguaje de las regiones reguladoras**

Resultados Piloto:

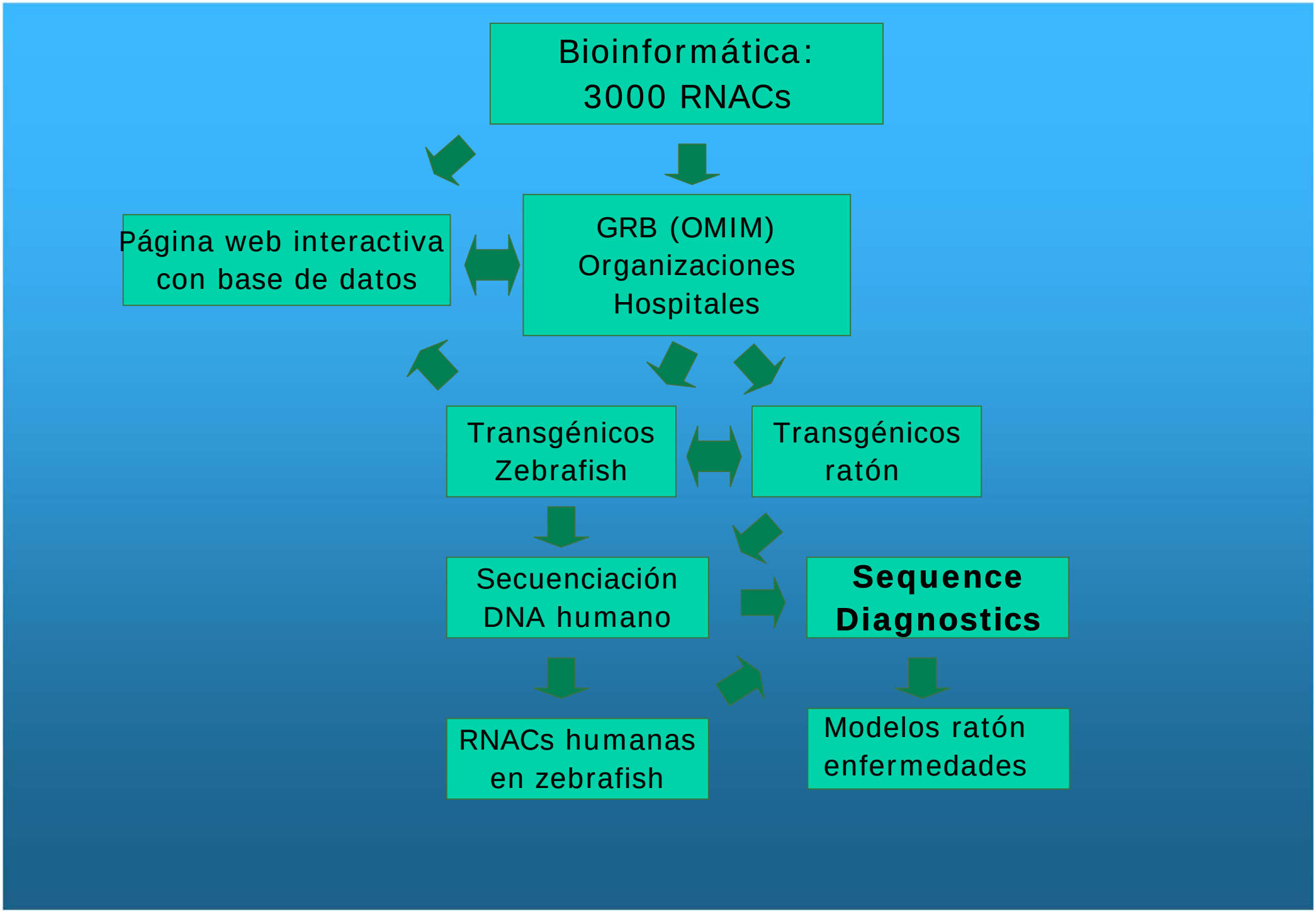
10 secuencias con actividad enhancer probada en ratón con 4 promotores: 2 X TATA and 2 X repeticiones CG. Total 40 ensayos.

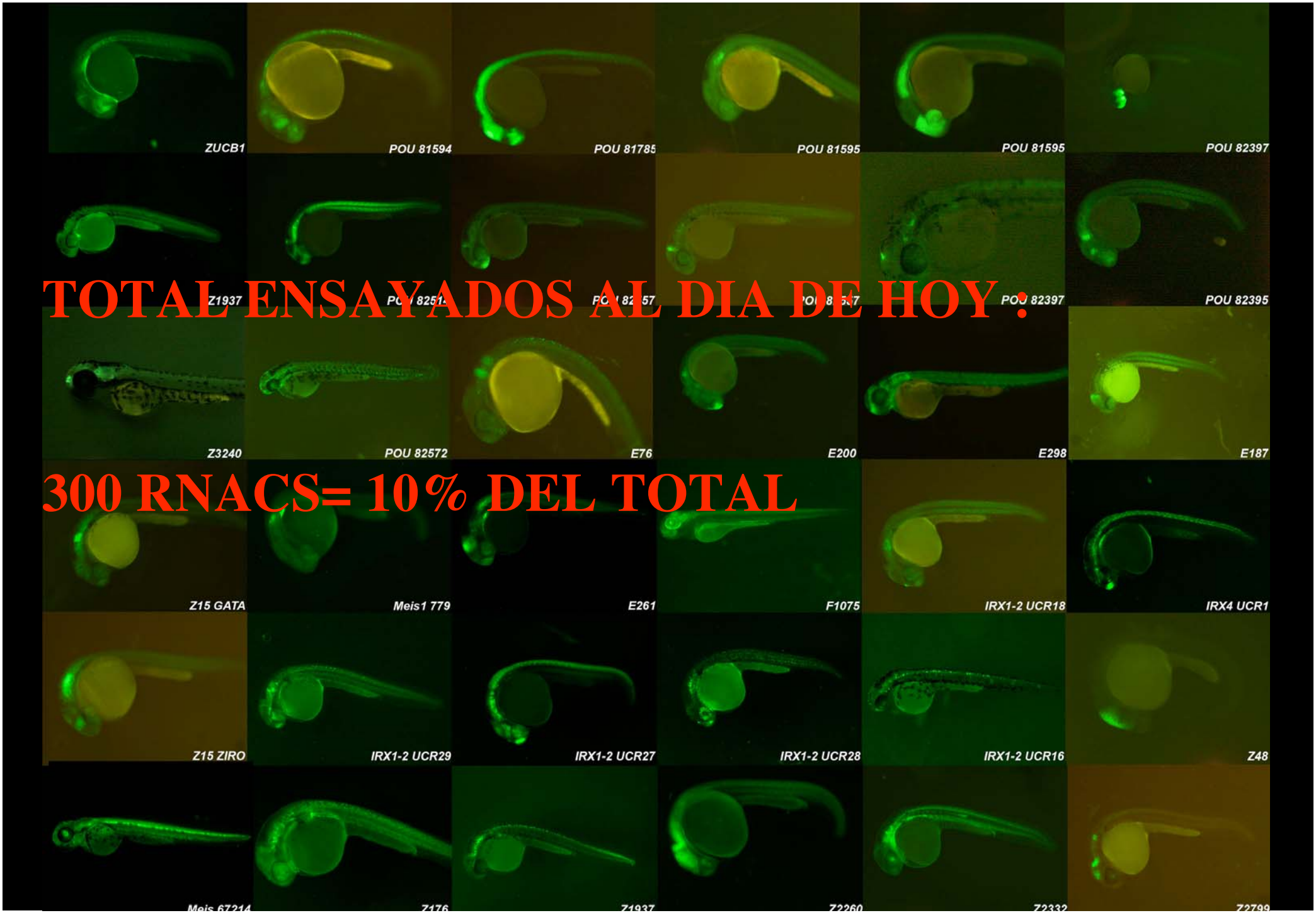


Conclusiones:

RNACs homólogas promueven expresión en regiones equivalentes en distintos vertebrados

Zebrafish es un buen modelo para un ensayo a escala genómica cuyos resultados son extrapolables a mamíferos


Bioinformática:
3000 RNACs
Página web interactiva
con base de datos
GRB (OMIM)
Organizaciones
Hospitales
Transgénicos
Zebrafish
Transgénicos
ratón
Secuenciación
DNA humano
Sequence
Diagnostics
RNACs humanas
en zebrafish
Modelos ratón
enfermedades


ZUCB1
POU 81594
POU 81785
POU 81595
POU 81595
POU 82397
Z1937
POU 82397
POU 82395
Z3240
POU 82572
E76
E200
E298
E187
Z15 GATA
Meis1 779
E261
F1075
IRX1-2 UCR18
IRX4 UCR1
Z15 ZIRO
IRX1-2 UCR29
IRX1-2 UCR27
IRX1-2 UCR28
IRX1-2 UCR16
Z48
Z176
Z1937
Z2260
Z2332
Z2799
TOTAL ENSAYADOS AL DIA DE HOY :
300 RNACS= 10% DEL TOTAL

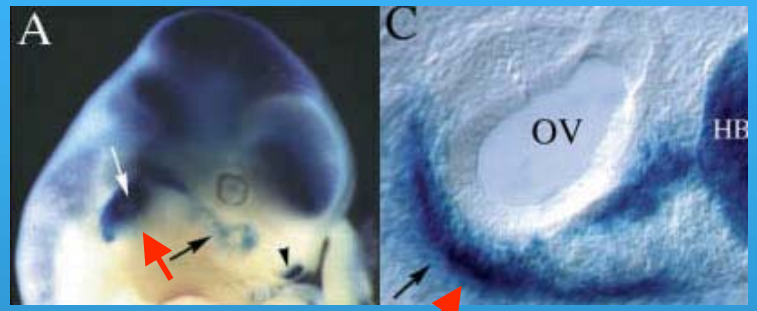
OBJETIVO:

Asociación de enfermedade genéticas humanas a mutaciones en RNACs

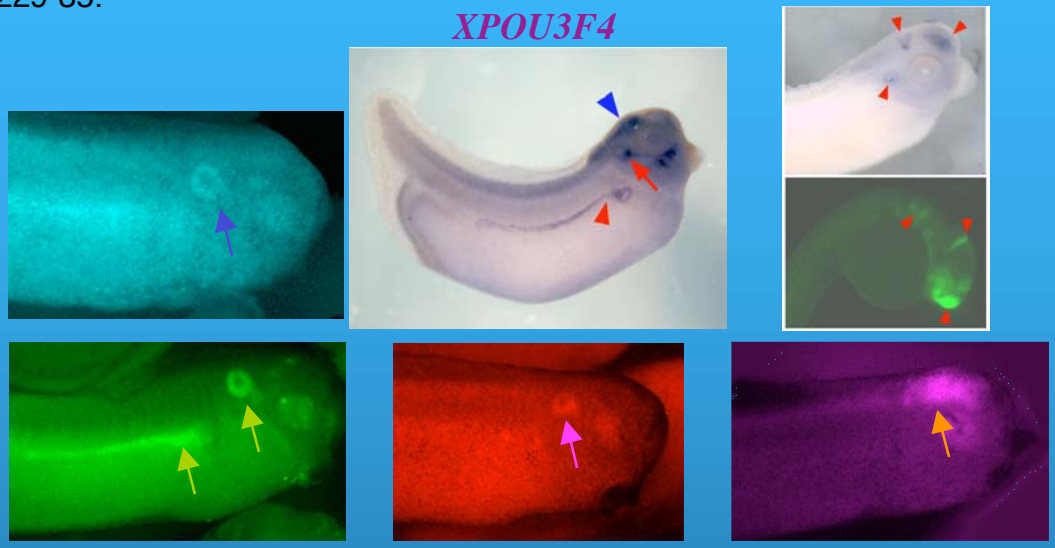
Identificación de RNACs implicadas en sordera

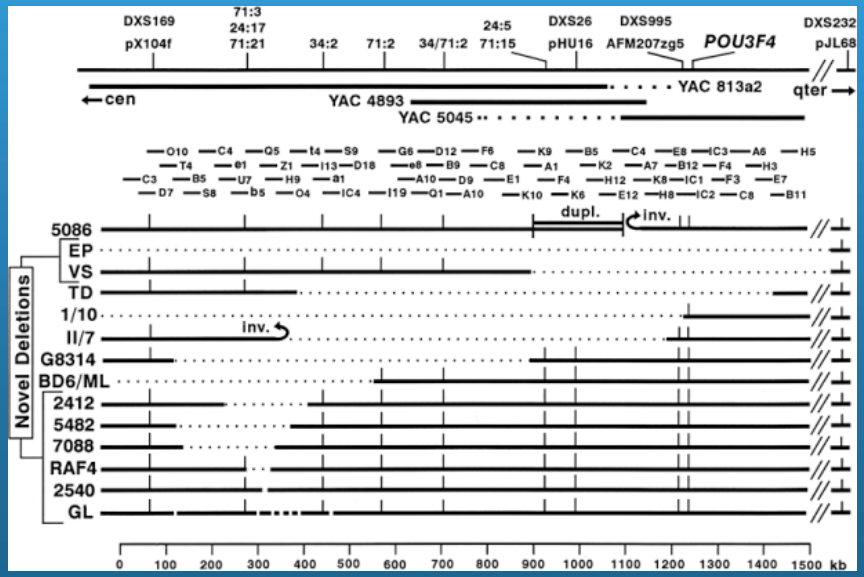
de Kok YJ et al. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene ***POU3F4***. Hum Mol Genet. 1996 Sep;5(9):1229-35.

mPOU3F4

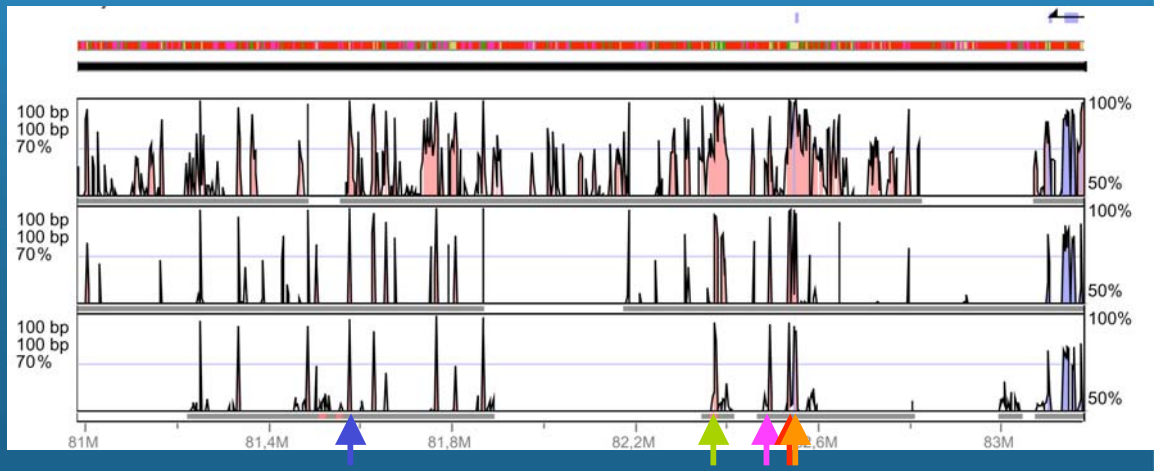


XPOU3F4





POU3F4



En colaboración:
Felipe Moreno (España), Richard Smith (EEUU) y
Frans Cremers (Holanda)

Estudios de asociación de enfermedades a SNPs

nature genetics

A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*

David A van Heel[1], Lude Franke[2,17], Karen A Hunt[1,17], Rhian Gwilliam[3,17], Alexandra Zhernakova[2], Mike Inouye[3], Martin C Wapenaar[4], Martin C N M Barnardo[5], Graeme Bethel[3], Geoffrey K T Holmes[6], Con Feighery[7], Derek Jewell[8], Dermot Kelleher[9], Parveen Kumar[1], Simon Travis[9], Julian RF Walters[10], David S Sanders[11], Peter Howdle[12], Jill Swift[13], Raymond J Playford[1], William M McLaren[3], M Luisa Mearin[14,15], Chris J Mulder[16], Ross McManus[17], Ralph McGinnis[3], Lon R Cardon[4], Panos Deloukas[3] & Cisca Wijmenga[2,4]

We tested 310,605 SNPs for association in 778 individuals with celiac disease and 1,422 controls. Outside the HLA region, the most significant finding (rs13119723; $P = 2.0 \times 10^{-5}$) was in the *KIAA1109-TENR-IL2-IL21* linkage disequilibrium block. We independently confirmed association in two further collections (strongest association at rs6822844, 24 kb 5′ of *IL21*; meta-analysis $P = 1.3 \times 10^{-14}$, odds ratio = 0.63), suggesting that genetic variation in this region predisposes to celiac disease.

Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility

Miles Parkes[1,13], Jeffrey C Barrett[2,13], Natalie J Prescott[3,13], Mark Tremelling[1], Carl A Anderson[2], Sheila A Fisher[1], Roland G Roberts[3], Elaine R Nimmo[4], Fraser R Cummings[5], Dianne Soars[3], Hazel Drummond[4], Charlie W Lees[4], Saud A Khawaja[1], Richard Bagnall[1], Denis A Burke[6], Catherine E Todhunter[6], Tariq Ahmad[3], Clive M Onnie[3], Wendy McArdle[7], David Strachan[8], Graeme Bethel[9], Claire Bryan[9], Cathryn M Lewis[3], Panos Deloukas[9], Alastair Forbes[10], Jeremy Sanderson[11], Derek P Jewell[5], Jack Satsangi[4], John C Mansfield[6], the Wellcome Trust Case Control Consortium[12], Lon Cardon[2] & Christopher G Mathew[3]

A genome-wide association scan in individuals with Crohn's disease by the Wellcome Trust Case Control Consortium detected strong association at four novel loci. We tested 37 SNPs from these and other loci for association in an independent case-control sample. We obtained replication for the autophagy-inducing *IRGM* gene on chromosome 5q33.1 (replication $P = 6.6 \times 10^{-4}$, combined $P = 2.1 \times 10^{-10}$) and for nine other loci, including *NKX2-3*, *PTPN2* and gene deserts on chromosomes 1q and 5p13.

A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer

David J Hunter[1–4], Peter Kraft[2], Kevin B Jacobs[5], David G Cox[1,2], Meredith Yeager[4,6], Susan E Hankinson[1], Sholom Wacholder[4], Zhaoming Wang[4,6], Robert Welch[4,6], Amy Hutchinson[4,6], Junwen Wang[4,6], Kai Yu[4], Nilanjan Chatterjee[4], Nick Orr[2], Walter C Willett[1,3], Graham A Colditz[3], Regina G Ziegler[4], Christine D Berg[10], Saundra S Buys[11], Catherine A McCarty[12], Heather Spencer Feigelson[13], Eugenia E Calle[13], Michael J Thun[13], Richard B Hayes[4], Margaret Tucker[4], Daniela S Gerhard[14], Joseph F Fraumeni, Jr[4], Robert N Hoover[4], Gilles Thomas[4] & Stephen J Chanock[4,7]

We conducted a genome-wide association study (GWAS) of breast cancer by genotyping 528,173 SNPs in 1,145 postmenopausal women of European ancestry with invasive breast cancer and 1,142 controls. We identified four SNPs in intron 2 of FGFR2 (which encodes a receptor tyrosine kinase and is amplified or overexpressed in some breast cancers) that were highly associated with breast cancer and confirmed this association in 1,776 affected individuals and 2,072 controls from three additional studies. Across the four studies, the association with all four SNPs was highly statistically significant (P_{trend} for the most strongly associated SNP (rs1219648) = 1.1×10^{-10}; population attributable risk = 16%). Four SNPs at other loci most strongly associated with breast cancer in the initial GWAS were not associated in the replication studies. Our summary results from the GWAS are available online in a form that should speed the identification of additional risk loci.

cases. In the past, family-based studies have been the primary focus of study in the search for genetic determinants, but with new technologies that enable analysis of hundreds of thousands of SNPs, together with new insights into the structure of variation in the human genome, it is now possible to scan the genome in an agnostic manner in studies of unrelated cases and controls in search of common genetic variants associated with disease risk[5]. One strategy for conducting a GWAS is to analyze early-onset cases, often enriched with cases with a positive family history of the disease, in order to maximize the opportunity to detect inherited causal variants. Already, two such studies, one of diabetes and one of breast cancer, have successfully identified and replicated associations of common genetic variants with these diseases[5,6]. Alternatively, GWA analysis of older subjects may identify common genetic variants associated with sporadic disease, as has been successfully demonstrated for prostate cancer[7,8].

We initially genotyped 1,183 women with postmenopausal invasive breast cancer and 1,185 individually matched controls from the Nurses'

Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer

Simon N Stacey[1], Andrei Manolescu[1], Patrick Sulem[1], Thorunn Rafnar[1], Julius Gudmundsson[1], Sigurjon A Gudjonsson[1], Gisli Masson[1], Margret Jakobsdottir[1], Steinunn Thorlacius[1], Agnar Helgason[1], Katja K Aben[2,3], Luc J Strobbe[4], Marjo T Albers-Akkers[5], Dorine W Swinkels[3], Brian E Henderson[6], Laurence N Kolonel[7], Loic Le Marchand[7], Esther Millastre[8], Raquel Andres[8], Javier Godino[9], Maria Dolores Garcia-Prats[10], Eduardo Polo[11], Alejandro Tres[8], Magali Mouy[1], Jona Saemundsdottir[1], Valgerdur M Backman[1], Larus Gudmundsson[1], Kristleifur Kristjansson[1], Jon T Bergthorsson[1], Jelena Kostic[1], Michael L Frigge[1], Frank Geller[1], Daniel Gudbjartsson[1], Helgi Sigurdsson[12], Thora Jonsdottir[12], Jon Hrafnkelsson[12], Jakob Johannsson[12], Thorarinn Sveinsson[12], Gardar Myrdal[12], Hlynur Niels Grimsson[12], Thorvaldur Jonsson[12], Susanna von Holst[13], Barbro Werelius[13], Sara Margolin[14], Annika Lindblom[13], Jose I Mayordomo[8], Christopher A Haiman[6], Lambertus A Kiemeney[3], Oskar Th Johannsson[12], Jeffrey R Gulcher[1], Unnur Thorsteinsdottir[1], Augustine Kong[1] & Kari Stefansson[1]

Familial clustering studies indicate that breast cancer risk has a substantial genetic component[1–3]. To identify new breast cancer risk variants, we genotyped approximately 300,000 SNPs in 1,600 Icelandic individuals with breast cancer and 11,563 controls using the Illumina Hap300 platform. We then tested selected SNPs in five replication sample sets. Overall, we studied 4,554 affected individuals and 17,577 controls. Two SNPs consistently associated with breast cancer: ~25% of individuals of European descent are homozygous for allele A of rs13387042 on chromosome 2q35 and have an estimated 1.44-fold greater risk than noncarriers, and for allele T of rs3803662 on 16q12, about 7% are homozygous and have a 1.64-fold greater risk. Risk from both alleles was confined to estrogen receptor–positive tumors. At present, no genes have been identified in the linkage disequilibrium block containing rs13387042. rs3803662 is near the 5′ end of *TNRC9*, a high mobility group chromatin–associated protein whose expression is implicated in breast cancer metastasis to bone[4].

Mutations in breast cancer susceptibility genes *BRCA1* and *BRCA2* account for 15%–25% of the familial component of breast cancer risk[3,6]. Much of the genetic component of risk of breast cancer remains uncharacterized and is thought to arise from combinations of less penetrant variants that, individually, may be quite common[7]. Many searches for less penetrant breast cancer risk variants have been carried out using a candidate gene, case-control association approach. Findings from these studies have often proven difficult to replicate[8]. Recently, common missense variants in two genes, *CASP8* and *TGFB1* have been shown to be associated with breast cancer risk through well-powered, multicenter analyses[9]. These reports emphasize the importance of large-scale studies with adequate replication when the goal is to identify common variants conferring modest increases in the risk of breast cancer.

In order to search widely for alleles of common SNPs associating with breast cancer susceptibility, we carried out a genome-wide SNP association study using Illumina HumanHap300 microarray technology. We genotyped 1,600 Icelandic individuals with breast cancer and 11,563 controls; we designated this discovery sample set 'Iceland 1'

Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes

John A Todd[1], Neil M Walker[1,9], Jason D Cooper[1,9], Deborah J Smyth[1,9], Kate Downes[1], Vincent Plagnol[1], Rebecca Bailey[1], Sergey Nejentsev[1], Sarah F Field[1], Felicity Payne[1], Christopher E Lowe[1], Jeffrey S Szeszko[1], Jason P Hafler[1], Lauren Zeitels[1], Jennie H M Yang[1], Adrian Vella[1,8], Sarah Nutland[1], Helen E Stevens[1], Helen Schuilenburg[1], Gillian Coleman[1], Meeta Maisuria[1], William Meadows[1], Luc J Smink[1], Barry Healy[1], Oliver S Burren[1], Alex A C Lam[1], Nigel R Ovington[1], James Allen[1], Ellen Adlem[1], Hin-Tak Leung[1], Chris Wallace[2], Joanna M M Howson[1], Cristian Guja[3], Constantin Ionescu-Tirgoviște[3], Genetics of Type 1 Diabetes in Finland[4], Matthew J Simmonds[5], Joanne M Heward[5], Stephen C L Gough[5], The Wellcome Trust Case Control Consortium[6], David B Dunger[7], Linda S Wicker[1] & David G Clayton[1]

The Wellcome Trust Case Control Consortium (WTCCC) primary genome-wide association (GWA) scan[1] on seven diseases, including the multifactorial autoimmune disease type 1 diabetes (T1D), shows associations at $P < 5 \times 10^{-7}$ between T1D and six chromosome regions: 12q24, 12q13, 16p13, 18p11, 12p13 and 4q27. Here, we attempted to validate these and six other top findings in 4,000 individuals with T1D, 5,000 controls and 2,997 family trios independent of the WTCCC study. We confirmed unequivocally the associations of 12q24, 12q13, 16p13 and 18p11 ($P_{follow\text{-}up} \leq 1.35 \times 10^{-9}$; $P_{overall} \leq 1.15 \times 10^{-14}$), leaving eight regions with small effects or false-positive associations. We also obtained evidence for chromosome 18q22 ($P_{overall} = 1.38 \times 10^{-6}$) from a GWA study of nonsynonymous SNPs. Several regions, including 18q22 and 18p11, showed association with autoimmune thyroid disease. This study increases the number of T1D loci with compelling evidence from six to at least ten.

clustering of T1D (**Supplementary Table 1** online). We have assumed for T1D[3] the classical model of a small number of genes with large effects and a large number of genes with small effects[4,5]. If this genetic model is correct, notwithstanding a major role for (unknown) environmental factors[6,7], there should be many more new genes (and pathways) to be discovered, provided sample sizes, study design and genotyping technology suffice[2,8–13].

Here, we followed up on the most statistically significant results from two GWA studies: a nonsynonymous SNP (nsSNP) case-control study of 13,378 SNPs in 3,400 affected individuals and 3,300 controls[2,8] and the WTCCC study using an Affymetrix 500K Mapping Array GWA GeneChip on 2,000 cases and 3,000 controls[1]. There was a substantial overlap of samples (1,834 cases and 1,134 controls) between these studies, but we still had independent samples available for follow-up (up to 4,000 affected individuals and 5,000 controls available from the same DNA collections and 2,997 parent-child trios from 2,839 families).

Vol 447|28 June 2007|doi:10.1038/nature05887 nature

ARTICLES

Genome-wide association study identifies novel breast cancer susceptibility loci

Douglas F. Easton[1], Karen A. Pooley[2], Alison M. Dunning[2], Paul D. P. Pharoah[2], Deborah Thompson[1], Dennis G. Ballinger[3], Jeffery P. Struewing[4], Jonathan Morrison[5], Helen Field[2], Robert Luben[5], Nicholas Wareham[5], Shahana Ahmed[2], Catherine S. Healey[2], Richard Bowman[6], the SEARCH collaborators[2*], Kerstin B. Meyer[2], Christopher A. Haiman[6], Laurence K. Kolonel[6], Brian E. Henderson[6], Loic Le Marchand[6], Paul Brennan[10], Suleeporn Sangrajrang[11], Valerie Gaborieau[10], Fabrice Odefrey[10], Chen-Yang Shen[12], Pei-Ei Wu[12], Hui-Chun Wang[12], Diana Eccles[13], D. Gareth Evans[14], Julian Peto[15], Olivia Fletcher[16], Nichola Johnson[16], Sheila Seal[17], Michael R. Stratton[17,18], Nazneen Rahman[17], Georgia Chenevix-Trench[19], Stig E. Bojesen[20], Børge G. Nordestgaard[20], Christen K. Axelsson[21], Montserrat Garcia-Closas[22], Louise Brinton[22], Stephen Chanock[23], Jolanta Lissowska[24], Beata Peplonska[25], Heli Nevanlinna[26], Rainer Fagerholm[26], Hannaleena Eerola[26,27], Daehee Kang[28], Keun-Young Yoo[28,29], Dong-Young Noh[28], Sei-Hyun Ahn[30], David J. Hunter[31,32], Susan E. Hankinson[32], David G. Cox[31], Per Hall[33], Sara Wedren[33], Jianjun Liu[34], Yen-Ling Low[34], Natalia Bogdanova[35,36], Peter Schürmann[36], Thilo Dörk[36], Rob A. E. M. Tollenaar[37], Catharina E. Jacobi[38], Peter Devilee[39], Jan G. M. Klijn[40], Alice J. Sigurdson[41], Michele M. Doody[41], Bruce H. Alexander[42], Jinghui Zhang[4], Angela Cox[43], Ian W. Brock[43], Gordon MacPherson[43], Malcolm W. R. Reed[44], Fergus J. Couch[45], Ellen L. Goode[45], Janet E. Olson[45], Hanne Meijers-Heijboer[46,47], Ans van den Ouweland[47], André Uitterlinden[48], Fernando Rivadeneira[48], Roger L. Milne[49], Gloria Ribas[49], Anna Gonzalez-Neira[49], Javier Benitez[49], John L. Hopper[50], Margaret McCredie[51], Melissa Southey[50], Graham G. Giles[52], Chris Schroen[53], Christina Justenhoven[54], Hiltrud Brauch[54], Ute Hamann[55], Yon-Dschun Ko[56], Amanda B. Spurdle[19], Jonathan Beesley[19], Xiaoqing Chen[19], kConFab[57*], AOCS Management Group[19,57*], Arto Mannermaa[58,59], Veli-Matti Kosma[58,59], Vesa Kataja[58,60], Jaana Hartikainen[58,59], Nicholas E. Day[5], David R. Cox[3] & Bruce A. J. Ponder[2,7]

Breast cancer exhibits familial aggregation, consistent with variation in genetic susceptibility to the disease. Known susceptibility genes account for less than 25% of the familial risk of breast cancer, and the residual genetic variance is likely to be due to variants conferring more moderate risks. To identify further susceptibility alleles, we conducted a two-stage genome-wide association study in 4,398 breast cancer cases and 4,316 controls, followed by a third stage in which 30 single nucleotide polymorphisms (SNPs) were tested for confirmation in 21,860 cases and 22,578 controls from 22 studies. We used 227,876 SNPs that were estimated to correlate with 77% of known common SNPs in Europeans at $r^2 > 0.5$. SNPs in five novel independent loci exhibited strong and consistent evidence of association with breast cancer ($P < 10^{-7}$). Four of these contain plausible causative genes (*FGFR2*, *TNRC9*, *MAP3K1* and *LSP1*). At the second stage, 1,792 SNPs were significant at the $P < 0.05$ level compared with an estimated 1,343 that would be expected by chance, indicating that many additional common susceptibility alleles may be identifiable by this approach.

Identificación de RNACs implicadas en diabetes

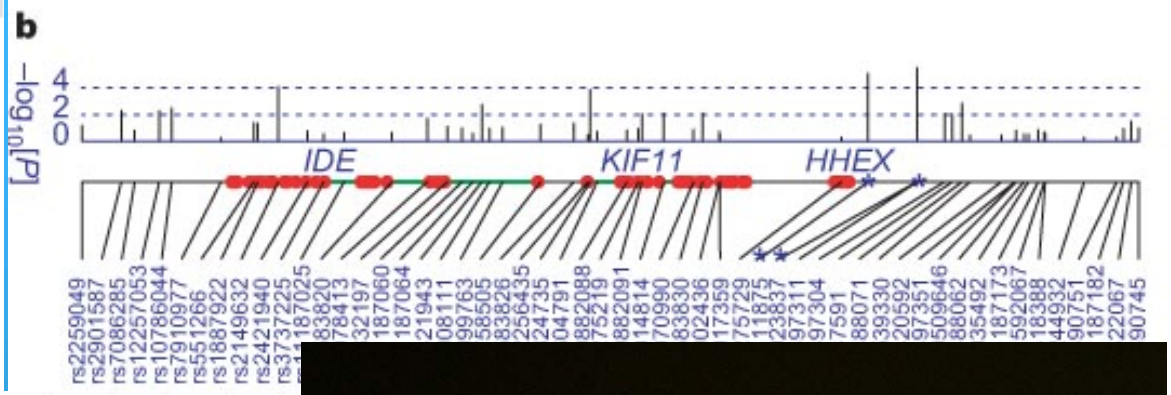
Vol 445|22 February 2007|doi:10.1038/nature05616 nature

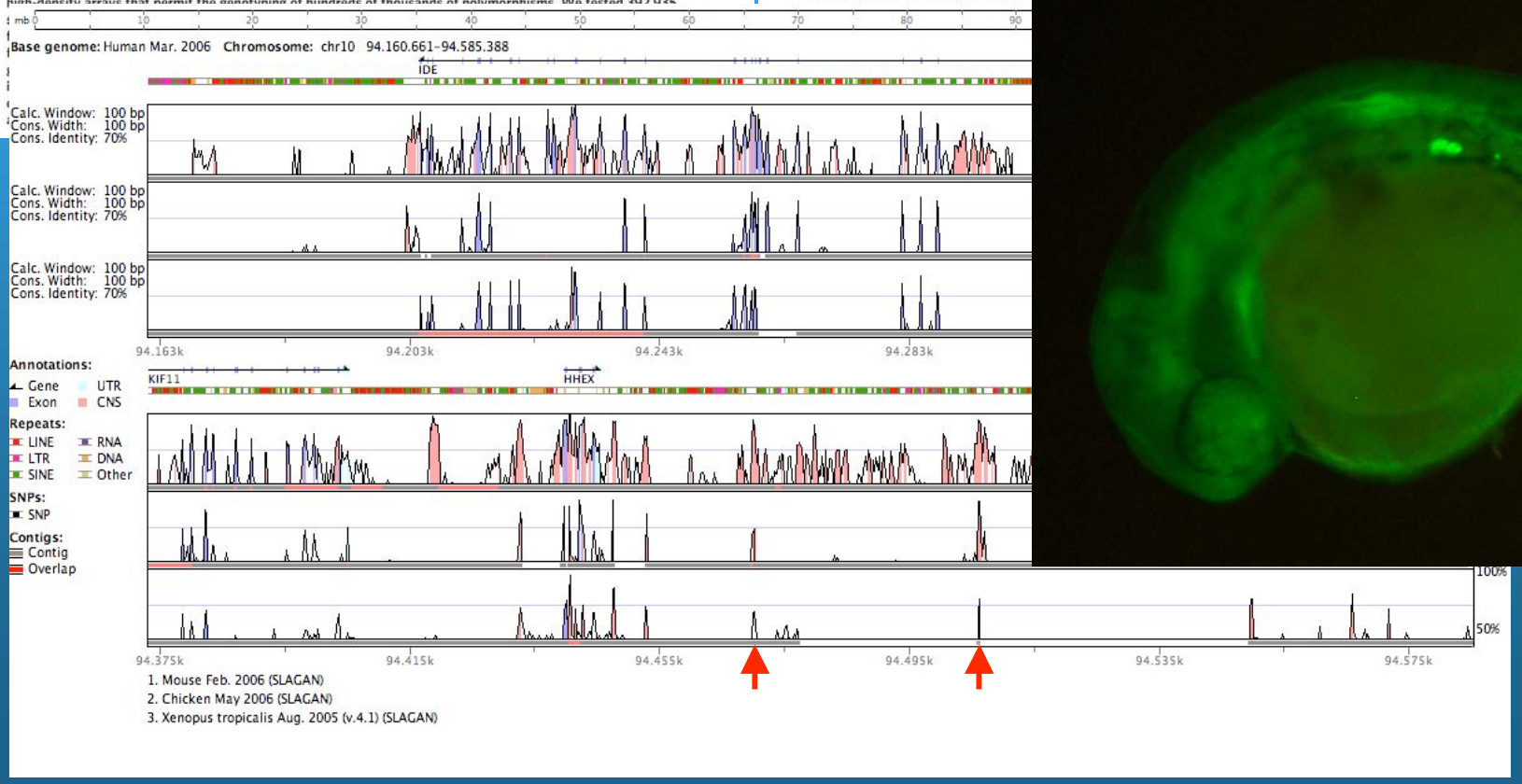
ARTICLES

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek[1,2,4], Ghislain Rocheleau[1]*, Johan Rung[4]*, Christian Dina[5]*, Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]

Type 2 diabetes mellitus results from the interaction of environmental factors with a combination of genetic variants, most of which were hitherto unknown. A systematic search for these variants was recently made possible by the development of high-density arrays that permit the genotyping of hundreds of thousands of polymorphisms. We tested 392,935





AGRADECIMIENTOS

Función de genes *Irx*

Elisa del Pilar Rodríguez
Ana Fernández
Pilar Alarcón

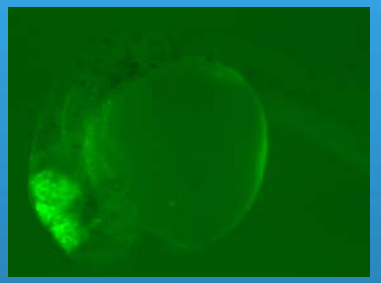
Colaboradores

Miguel Manzanares

Enhancers Vertebrados

Elisa de la Calle-Mustienes
Juan Tena
Eva Alonso
Ana Miñan
Jose Bessa
Silvia Naranjo

Cármen Gloria Feijoo
Miguel Allende



Proyecto Develome

Fernando Casares
Tom Becker
Boris Lenhard

Where do genes start, and where do they end?

Cell *129*, 823–837, May 18, 2007 ©2007 Elsevier Inc. 823

Resource

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]

[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

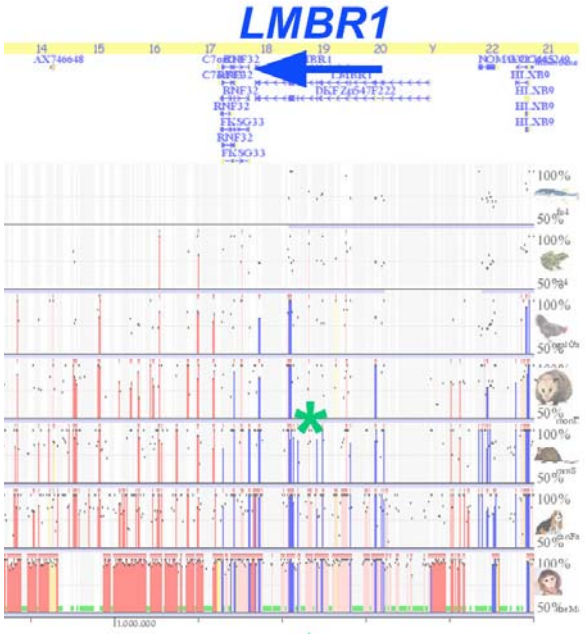
[3] These authors contributed equally to this work and are listed alphabetically.

*Correspondence: zhaok@nhlbi.nih.gov

DOI 10.1016/j.cell.2007.05.009

Histone modifications are implicated in influencing gene expression. We have generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF across the human genome using the Solexa 1G sequencing technology. Typical patterns of histone methylations exhibited at promoters, insulators, enhancers, and transcribed regions are identified. The monomethylations of H3K27, H3K9, H4K20, H3K79, and H2BK5 are all linked to gene activation, whereas trimethylations of H3K27, H3K9, and H3K79 are linked to repression. H2A.Z associates with functional regulatory elements, and CTCF marks boundaries of histone methylation domains. Chromosome banding patterns are correlated with unique patterns of histone modifications. Chromosome breakpoints detected in T cell cancers frequently reside in chromatin regions associated with H3K4 methylations. Our data provide new insights into the function of histone methylation and chromatin organization in genome function.

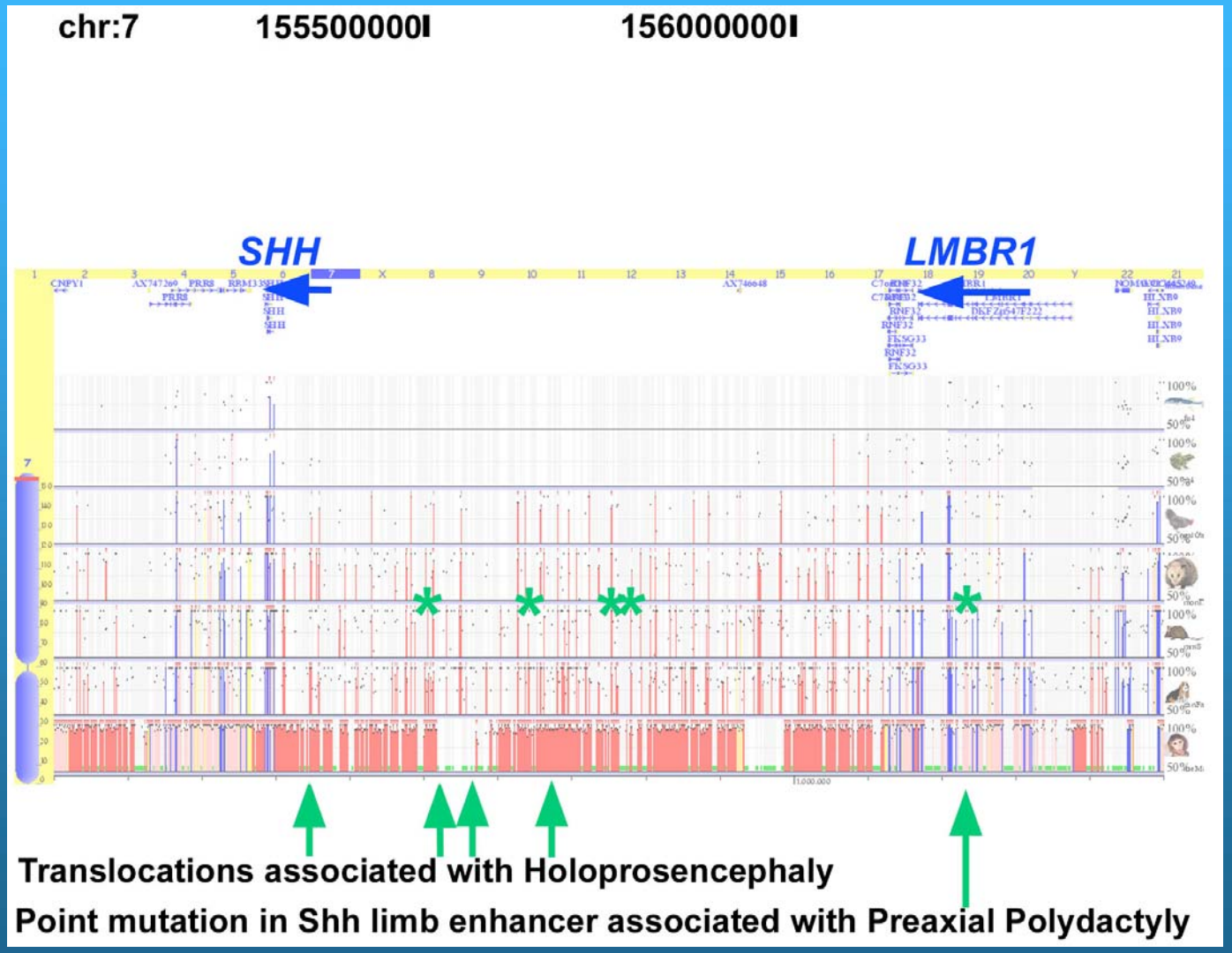
Cell



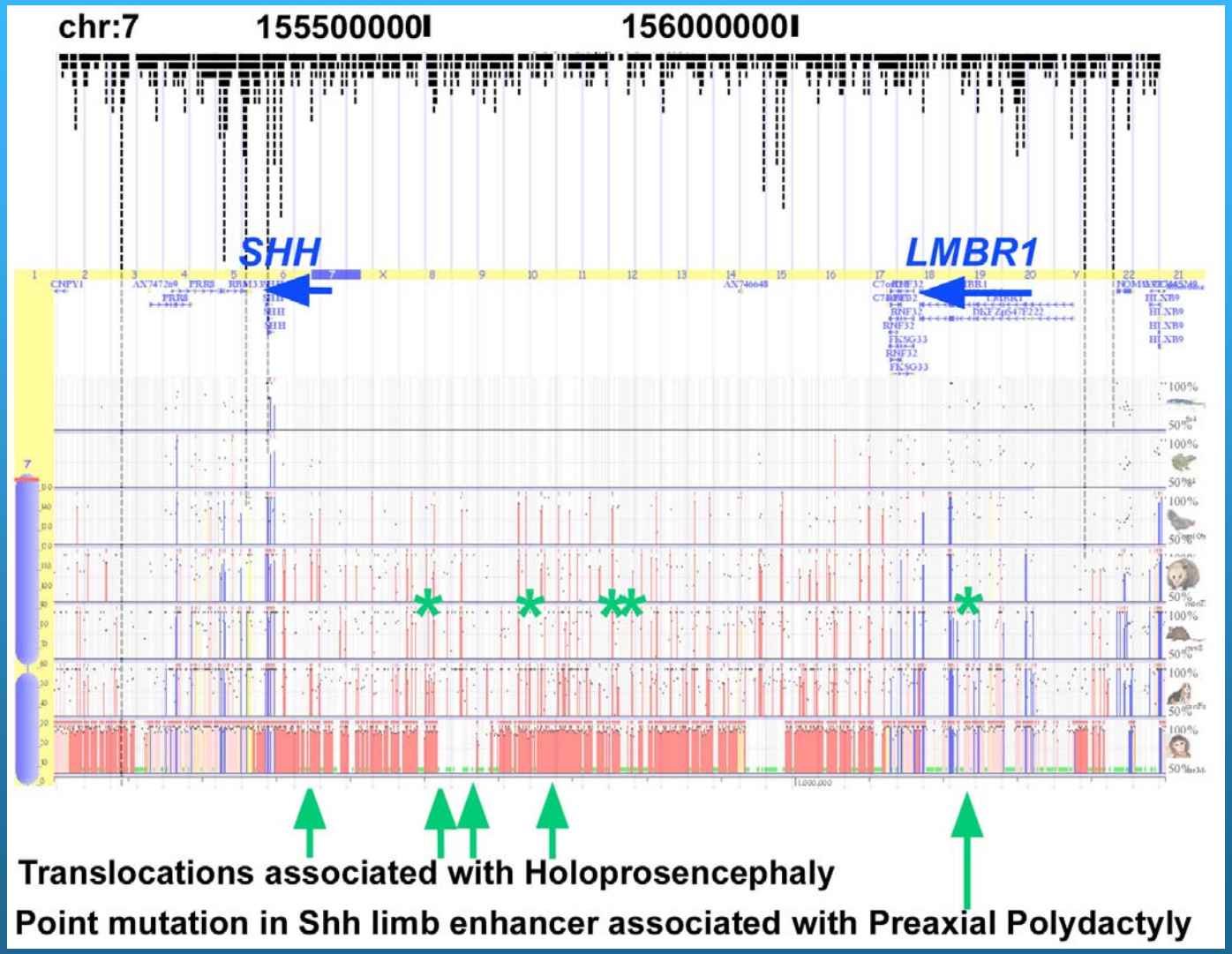
ncephaly

iated with Preaxial Polydactyly

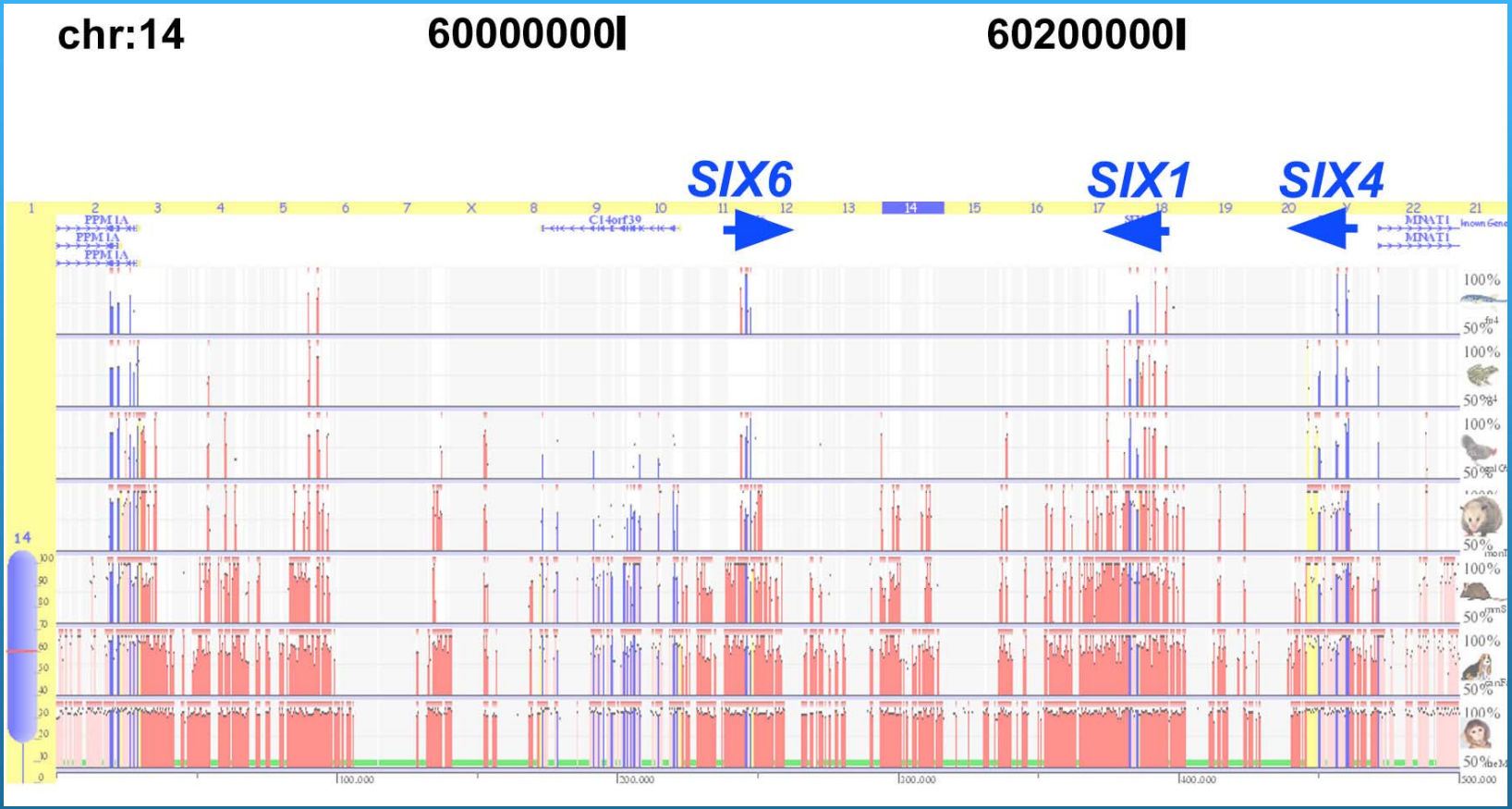
Where do genes start, and where do they end?



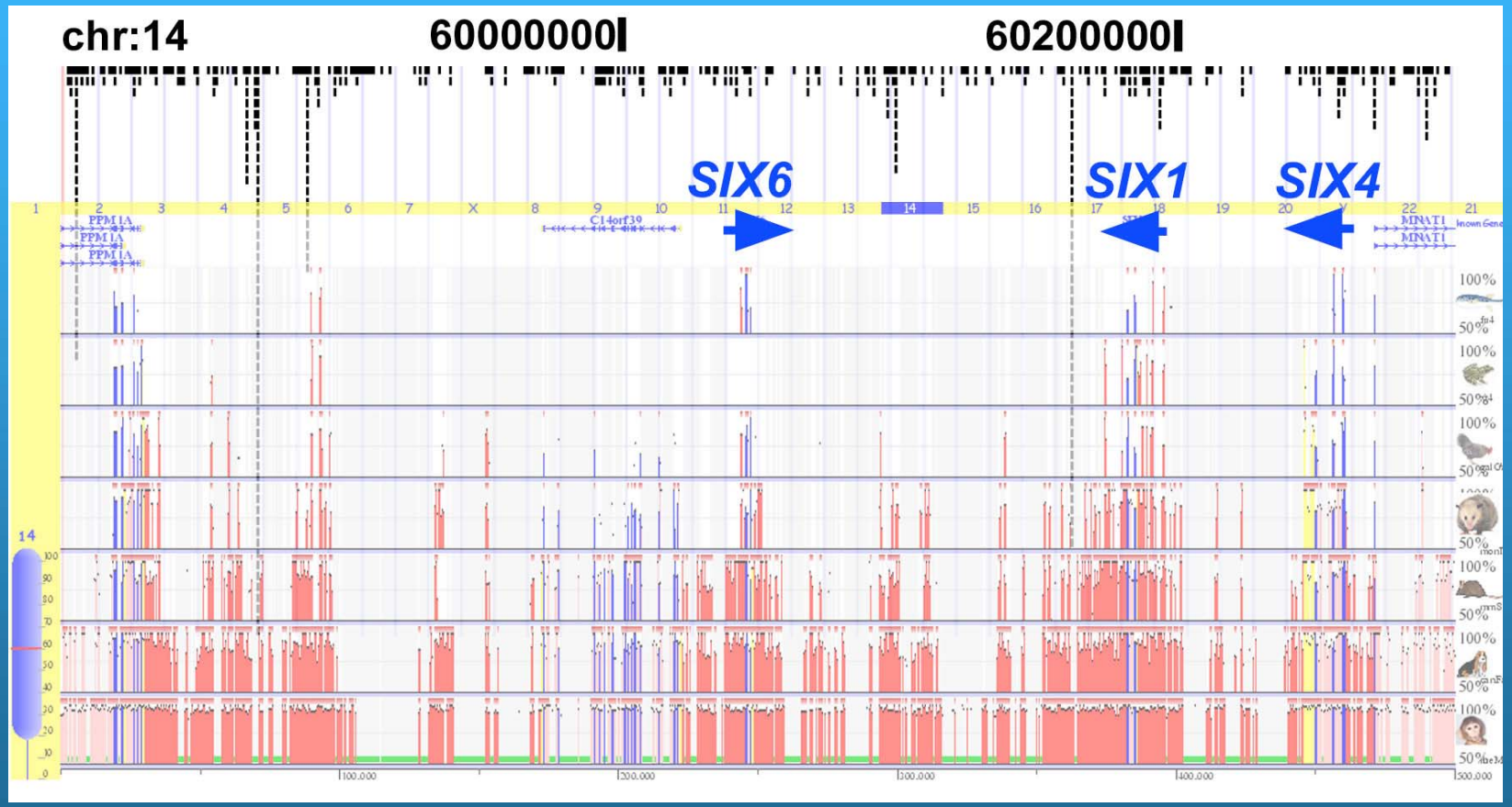
Where do genes start, and where do they end?



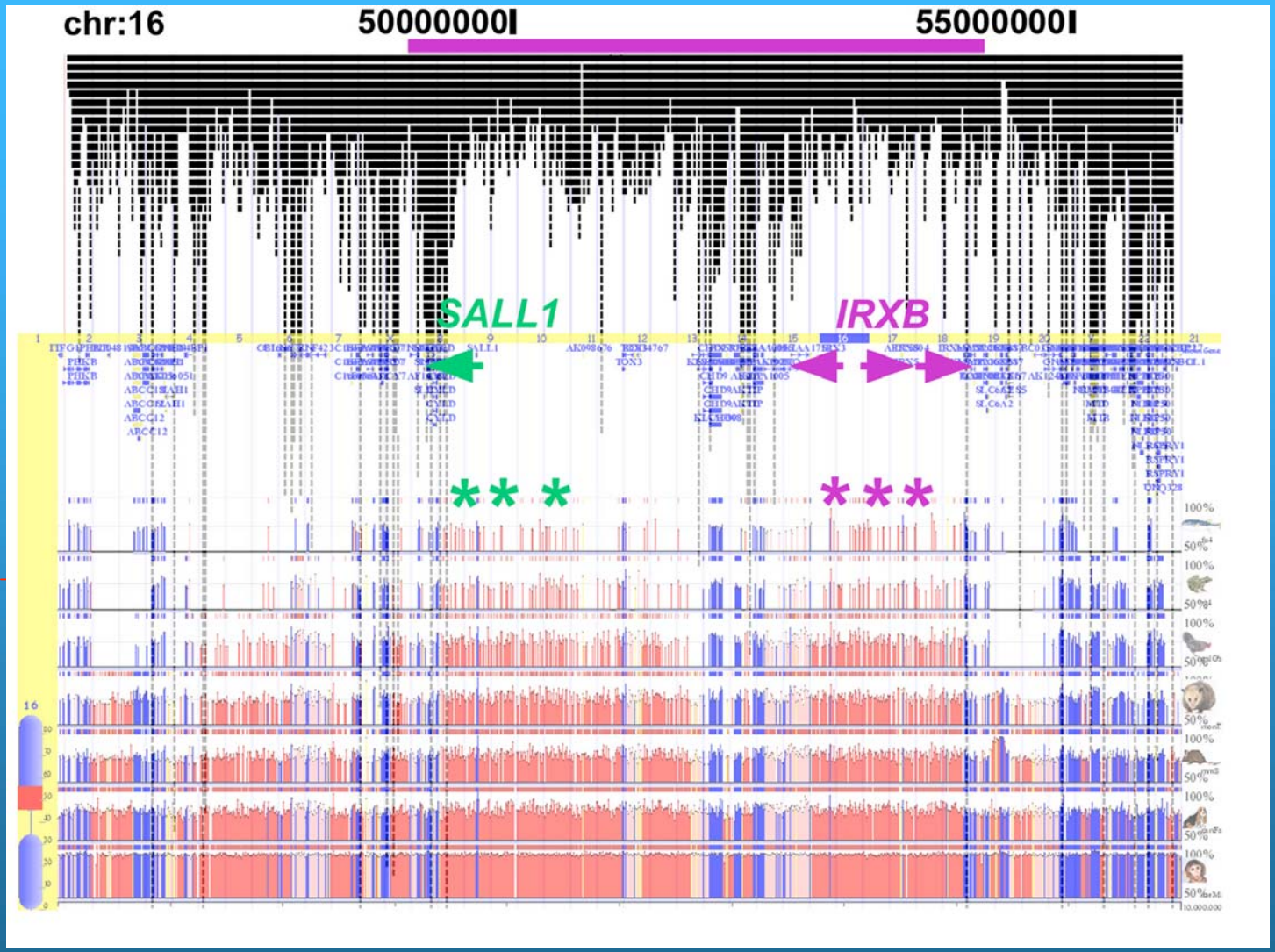
Where do genes start, and where do they end?



Where do genes start, and where do they end?



Where do genes start, and where do they end?



This simple co-representation of whole genome comparisons and CTCF distribution should aid to determine the extent of DNA around a particular transcription unit that is likely to contain most, if not all, of its cis-regulatory elements.