Anàlisi bioinformàtica de seqüències reguladores de l'expressió gènica en eucariotes

Domènec Farré Marimon

Tesi Doctoral

Barcelona, abril de 2008

Programa de Doctorat de Genètica
Departament de Genètica
Facultat de Biologia
Universitat de Barcelona
Bienni 2001-2003

Anàlisi bioinformàtica de seqüències reguladores de l'expressió gènica en eucariotes

Memòria presentada per

Domènec Farré Marimon

per optar al grau de

Doctor per la Universitat de Barcelona

Tesi Doctoral realitzada al Centre de Regulació Genòmica (CRG) sota la direcció de la Dra. **M. Mar Albà** de l'Institut Municipal d'Investigació Mèdica (IMIM) / Universitat Pompeu Fabra (UPF) / Institució Catalana per a la Recerca i Estudis Avançats (ICREA) i el Dr. **Xavier Messeguer** del Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (UPC).

La Directora,

M. Mar Albà

El Director,

Xavier Messeguer

La Tutora,

Sílvia Atrian

L'Autor,

Domènec Farré Marimon

Barcelona, abril de 2008

“El mirall s’havia trencat. Els bocins s’aguantaven en el marc però uns quants havien saltat a fora. Els anava agafant i els anava encabint en els buits on li semblava que encaixaven. Les miques de mirall, desnivellades, ¿reflectien les coses tal com eren? I de cop a cada mica de mirall veié anys de la seva vida viscuda en aquella casa.”

MERCÈ RODOREDA

Agraïments

TOT, LO BO I LO DOLENT, PER BÉ O PER MAL, ACABA ALGUNA DIA. Un doctorat té molt de bo, però també moments difícils o fins i tot durs. Per això vull agrair a tota la gent que, conscient o inconscientment, ha alleugerat els moments difícils pels que he passat aquests anys de doctorat i/o ha compartit amb mi els bons moments.

A la Mar i al Xavier per oferir-me l'oportunitat de realitzar el doctorat amb ells en aquest camp tan interessant i complex. Gràcies per les seves idees, els seus consells, el seu suport i, sobretot, la seva paciència.

A la Sílvia per la seva ajuda abans del DEA i després del DEA, sobretot aquests últims dies.

Al Roderic pel seu suport, el seu interés per la meva feina i les seves preguntes sempre tan encertades.

Al Robert, l'Eduardo i la Núria pels seus comentaris, la seva disponibilitat a ajudar en tot i la seva actitut tan positiva.

Als meus companys de patiments, els altres membres del grup *Evolutionary Genomics*: Nicolás, Loris, Macarena, Medya, Aliche. Amb ells, sobretot amb els tres primers, he compartit moments de nervis i d'estrès, per exemple abans d'un seminari (eh, Nicolás?), peró també molts bons moments, per exemple un esmorçar abans de pujar al Pedraforca (aquella montanya que acabaria amb les botes de la Macarena i amb la meva força física).

A la Meritxell, per la seva vitalitat i bon rotllo, i al Ramon i la Bet, la proveïdora d'aigua :-) que fa dies que no veig, pels *breaks* tan distrets.

Als estudiants de la UPC que han participat en el projecte de PROMO, especialment al David.

Als informàtics (sense bio): Òscar, Alfons, Judith, ..., per la seva constant ajuda.

Als que varen intervenir en versions anteriors de la plantilla de LaTeX que he adaptat per a redactar aquesta tesi: Enrique, Pep, Genís, Sergi, Robert.

En general, a tota la gent del PRBB (GRIB/CRG/IMIM/UPF/...) i de l'INB amb qui he estat en contacte, investigadors i no investigadors. Especialment, als que han hagut de soportar algun dels meus seminaris.

Al Miquel, a la Gloria, a la Rosa, al Carles i a la resta de companys de feina en aquell moment en que vaig decidir començar l'aventura del doctorat; em van animar en uns moments difícils per a mí.

A l'*Instituto Nacional de Bioinformática* (INB), a la *Fundación BBVA* i al *Plan Nacional de I+D del*

Ministerio de Educación y Ciencia pel seu suport econòmic.

A la meva família i amics, per estar sempre al meu costat quan cal. Especialment a la meva mare, als meus germans, a les meves ties, als meus cosins, a la Marta, a la Monique, al Dani i, sobretot, al Toni.

Contingut

Llista de Figures

Llista de Taules

Llista d'Abreviatures

cDNA	*Complementary DNA* (Àcid Desoxirribonucleic Complementari)
DDC	Duplicació-Degeneració-Complementació
DNA	*Deoxyribonucleic Acid* (Àcid Desoxirribonucleic)
dSM	*shared motif divergence* (mesura la divergència entre 2 seqüències)
HK	*housekeeping* (s'expressa gairebé en tots els teixits)
IUPAC	*International Union of Pure and Applied Chemistry*
Ka	*rate of non-synonymous substitutions* (taxa substitucions no sinònimes)
Kb	1000 pb
Ks	*rate of synonymous substitutions* (taxa substitucions sinònimes)
no-HK	no *housekeeping*
pb	parell(s) de bases
PWM	*Position Weight Matrix* (matriu de pesos)
RNA	*Ribonucleic Acid* (Àcid Ribonucleic)
TF	*Transcription Factor* (factor de transcripció)
TFBS	*Transcription Factor Binding Site* (lloc d'unió de factor de transcripció)
TSS	*Transcription Start Site* (lloc d'inici de la transcripció)

Part I

INTRODUCCIÓ

Resum

Aquesta introducció mostra l'àmbit científic al que correspon aquesta tesi. Resumeix els aspectes més importants del camp d'estudi: les regions de DNA que regulen la transcripció gènica. S'indica la situació dels coneixements en el moment en que s'inicià el treball, o millor dit els diferents treballs que composen la tesi, amb èmfasi a les qüestions no resoltes. Algunes d'aquestes qüestions s'han intentat contestar al llarg del treball de la tesi, al temps que s'han corroborat aspectes ja coneguts.

Elements reguladors de la transcripció

L'IMPACTE FENOTÍPIC D'UN GEN ÉS RESULTAT DE DOS COMPONENTS DIFERENTS: l'activitat biològica de la proteïna que codifica i les condicions específiques en que la proteïna s'expressa i pot exercir la seva activitat. La regulació de la transcripció és el més important punt de control de l'expressió d'un gen sota unes determinades condicions. Per tant, un gen està format per les regions codificants i les regions reguladores (principalment de la transcripció).

La dilucidació i anotació de les característiques biològicament rellevants de les seqüències genòmiques és essencial perquè aquestes siguin realment útils. L'anotació de genomes s'ha enfocat en identificar gens (les regions codificants) i predir les seves funcions, deixant sovint de banda la predicció d'elements reguladors de la transcripció en seqüències no codificants, tot i el paper essencial que tenen. Bucher (1999) afirmava que el problema de relacionar una seqüència de DNA amb una funció de regulació gènica semblava un problema tant difícil que pronosticava que els mètodes de predicció exactes no estarien disponibles quan estès completada la seqüenciació del genoma humà, al cap de 4 anys. Efectivament, actualment els mètodes de predicció de regions de regulació gènica continuen sent un repte.

Les seqüències codificants tenen una relació regular, directa, precisa i fàcil d'interpretar amb el seu fenotip immediat (bioquímic): una seqüència específica d'aminoàcids. En canvi, i aquest és el problema essencial, les seqüències reguladores tenen una relació peculiar, indirecta, no lineal i depenent del context amb el seu fenotip immediat: un pèrfil particular de transcripció, d'expressió gènica (Wray et al., 2003).

Afortunadament, tal com indicà Bucher, hi ha un paradigma ben acceptat en quan a l'organització de les regions reguladores de la transcripció (figura 1):

- Per cada gen, hi ha una o més d'una regió de control *upstream* o *downstream* del TSS (lloc d'inici de la transcripció): promotor, *enhancer*, *silencer*, . . .
- Aquestes regions presenten una organització modular. Cada regió reguladora està formada per un o més mòduls (de 30-500 pb).

- El mecanisme de la regulació transcripcional està orquestrat per factors de transcripció (TFs o *transcription factors*) que s'uneixen a segments específics del DNA. Cada TF té propietats característiques d'unió, incloent el patró i l'amplada de la seqüència de DNA a la que s'uneix, així com l'energia amb què ho fa.

- Podem considerar els llocs d'unió dels TFs (*transcription factor binding sites* o, abreujadament, TFBSs) com les unitats elementals de l'organització. Cada mòdul regulador conté una combinació de diversos TFBSs funcionals. La regulació gènica depèn de la combinació específica dels elements, així com de l'ordre i orientació en què apareixen.

- Cada TF individual interactua amb el lloc d'unió al DNA per activar, amplificar o reprimir l'expressió gènica. La regulació gènica s'efectua mitjançant complexos multiproteics de TFs (activadors o repressors) que interactuen sinèrgicament.

Per defecte, la transcripció està apagada, en *off* (Wray et al., 2003), és a dir, no pot haver transcripció efectiva d'un gen en absència de factors de transcripió específics. Per tant, tots els promotors contenen TFBSs d'activadors de la transcripció, en canvi només alguns contenen TFBSs de repressors (Davidson, 2001).

Dues parts funcionals estan sempre presents en els promotors dels gens eucariotes, però sovint són difícils de reconèixer a partir de només la informació de la seqüència. Una part és el promotor basal (o *core promoter*), on s'uneix la maquinària iniciadora de la transcripció (RNA polimerasa, TBP, TAFs, TFIIA, . . .). L'altra part funcional és tota la col·lecció de mòduls de TFBSs que assigna especificitat a la transcripció. Ara bé, la composició i organització d'aquests mòduls i TFBSs varien enormement entre diferents gens d'eucariotes (Wray et al. 2003; figura 2).

Consideracions sobre la interacció proteïna-DNA

Als anys 70, Seeman et al. (1976) indicaven que, en el complexos proteïna-DNA, les interaccions entre els aminoàcids i les bases de la doble hèlix es realitzaven mitjançant ponts d'hidrogen i interaccions hidrofòbiques. Proposaven ponts d'hidrogen dobles entre aminoàcid-parell de base. També indicaven importants diferències entre les interaccions en els solc major respecte a les interaccions en el solc menor.

Durant els anys 90, Mandel-Gutfreund et al. (1995) i Mandel-Gutfreund and Margalit (1998) realitzaren estudis sobre dades experimentals de cristal·lografia de complexos proteïna-DNA i experiments combinant seqüències de DNA i proteïnes. Com indicaren, existeixen desviacions respecte a l'esperat segons el model teòric de ponts d'hidrogen i interaccions hidrofòbiques. A partir de les dades experimentals establiren unes puntuacions de les interaccions aminoàcid-base, tot i que constataren que manquen dades experimentals per establir unes puntuacions realment vàlides.

Dels estudis d'interacció DNA-proteïna realitzats durant els anys 80 i 90 es pot concloure que, tot i que hi ha interaccions afavorides, l'especificitat per les seqüències de DNA rarament es pot explicar per una correspondència un a un entre aminoàcid i base. La unió al DNA varia substancialment entre famílies de proteïnes i no es pot establir un codi simple que adequadament descrigui el reconeixement dels llocs d'unió.

Més tard, Luscombe et al. (2001) assenyalava que apart dels ponts d'hidrogen, les forces de van der Waals jugaven un paper molt important en la interacció proteïna-DNA. Són els ponts d'hidrogen, sobretot, i les forces de van der Waals els determinants de l'especificitat de la interacció. A més, poden existir interaccions complexes: entre un aminoàcid i 2 bases de diferent parell de bases. Per tant, podem parlar d'un codi de reconeixement proteïna-DNA (Pabo and Sauer, 1984) que està degenerat en ambdues direccions: un aminoàcid pot reconèixer diverses bases i una base pot ser reconeguda per diferents aminoàcids. Aquest codi no és determinista, és probabilístic (Benos et al., 2002).

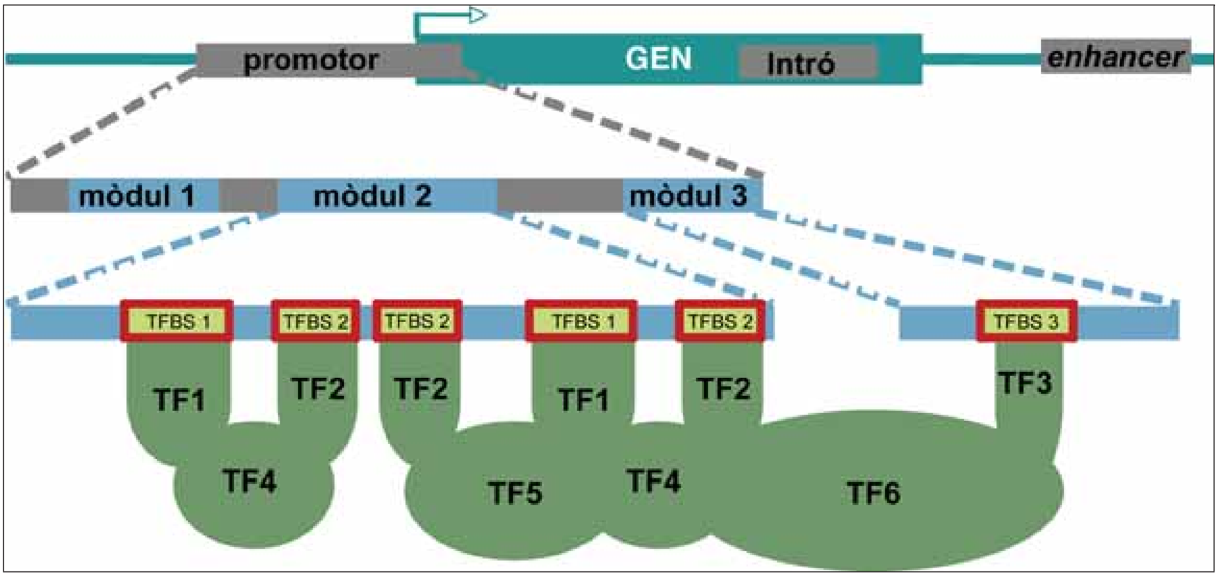


Figura 1 **Elements reguladors de la transcripció.** Els factors de transcripció (TFs) s'uneixen als motius reguladors (TFBSs) o interaccionen amb altres factors.

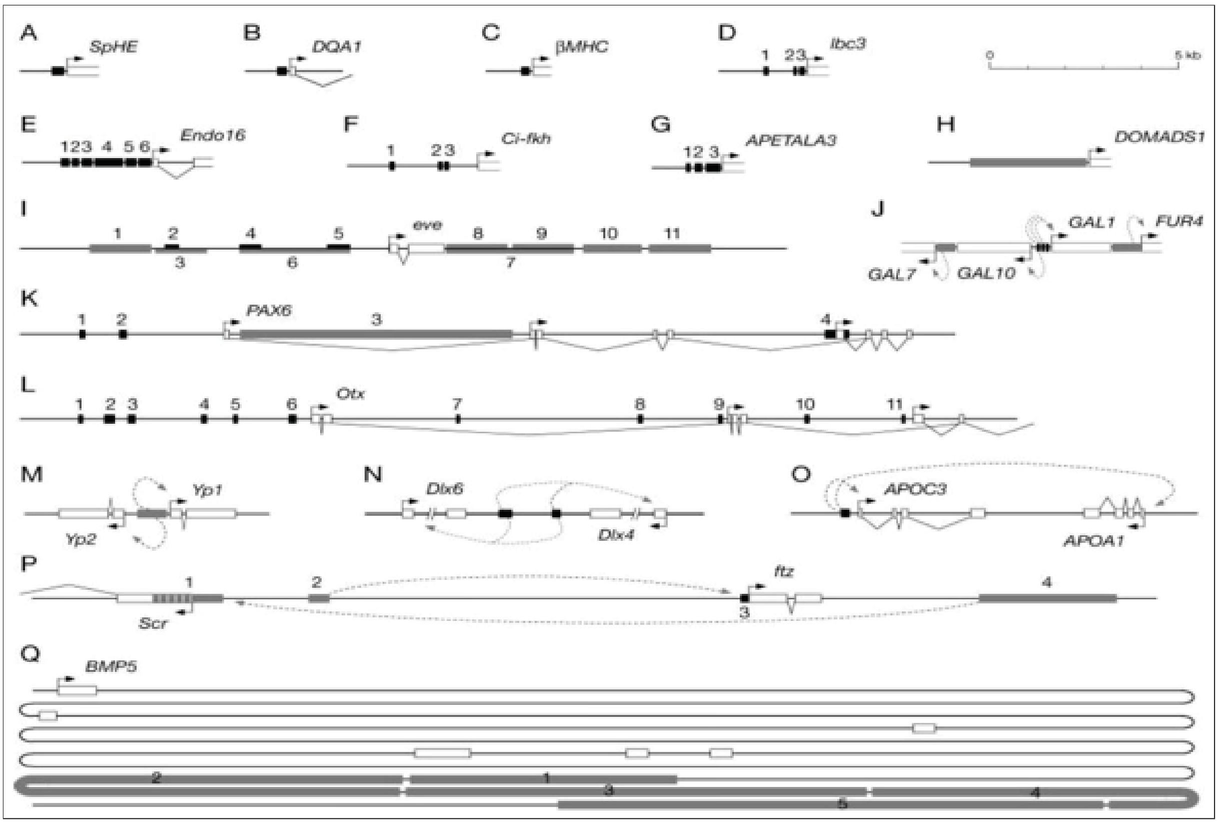


Figura 2 Variabilitat en la composició i organització de les regions reguladores. S'indiquen regions reguladores conegudes de diferents gens d'eucariotes (A-Q). Les caixes negres indiquen regions reguladores mapades de forma precisa; les caixes grises contenen seqüències reguladores que no han estat mapades de manera precisa; les caixes blanques són els exons; les fletxes amb L indiquen els llocs d'inici de la transcripció; els números distingeixen regions reguladores per les que s'ha definit experimentalment la contribució al perfil transcripcional; les fletxes amb línies discontinues indiquen interaccions entre un mòdul i més d'un *locus* o un *locus* no adjacent. Adaptat a partir de Wray et al. 2003.

Problemàtica de l'estudi dels TFBSs

En intentar fer predicció de llocs d'unió de factors de transcripció ens trobem per tant amb la següent problemàtica:

- El model de base no és determinista.
- Manquen dades d'interacció proteïna-DNA suficients per definir un model probabilístic complet.
- Els TFBSs són curts (4-20 pb) i amb alta variabilitat, fet que fa difícil separar els resultats reals dels deguts a l'atzar.

➼ Hi han dependències contextuals encara poc conegudes (Kadonaga, 1998): interaccions amb altres TFs, relacions amb factors que remodelen la cromatina, . . .

Predicció de TFBSs

En general quan parlem de predicció de TFBSs ens podem trobar en 2 situacions:

① *Pattern matching*: Tenim un conjunt de seqüències de DNA de les quals existeixen dades experimentals de la seva interacció amb un determinat TF i volem utilitzar aquesta informació per cercar TFBS d'aquest factor en altres seqüències. És fonamentalment un problema de caracterització i representació del senyal. Un cop caracteritzat i representat el TFBS només resta com fer la cerca d'aparicions en les seqüències de DNA problema (el *matching* pròpiament dit), una qüestió menys complexa.

② *Pattern discovery*: Tenim un conjunt de seqüències de DNA que sospitem tenen TFBSs en comú i volem trobar-los. Aquest cas es tracta d'un problema de detecció del senyal. Consisteix en detectar motius comuns que apareixen amb una freqüència superior a l'esperada per atzar (*overrepresented DNA motifs*).

En el cas (1), per a la caracterització i representació de TFBSs, normalment es parteix de dades experimentals directament de la literatura o que s'han compilat en alguna base de dades de regulació transcripcional (Ghosh, 2000; Higo et al., 1999; Lescot et al., 2002; Matys et al., 2003; Salgado et al., 2001; Zhu and Zhang, 1999; Kolchanov et al., 2002). Són dades que associen seqüències de DNA de zones reguladores a TFs que s'hi uneixen. Aquestes seqüències solen ser més llargues que la subseqüència en què realment es produeix la unió amb la proteïna. Per caracteritzar el TFBS de cada factor a partir d'aquestes seqüències s'han proposat diferents tipus d'algorismes: alineament local múltiple, xarxes neuronals, algorismes *greedy*, ...

Un cop caracteritzats aquests TFBSs, cal representar-los. També s'han proposat diferents tipus de representació: seqüències *consensus*, expressions regulars, matrius de pes o matrius de freqüències, ... Els programes més antics utilitzaven la representació de seqüència *consensus* de codis IUPAC, però en aquest tipus de representació es perd molta informació. La representació àmpliament més utilitzada és la de matrius de pes o de freqüències, abreviada sovint PWM (*position weight matrix*) (Stormo et al., 1982; Harr et al., 1983). En una matriu de pes assignem un pes a cada possible nucleòtid en cada posició del lloc d'unió a representar. A diferència de les seqüències *consensus*, en els programes que utilitzen matrius de pes es pot obtenir una classificació quantitativa (una puntuació) dels resultats que suggereix la possibilitat d'unió de la proteïna al lloc analitzat (Frech et al., 1997b). Les matrius de pes tenen una interpretació termodinàmica; hi ha una relació entre la probabilitat d'ocurrència de cada base en una posició determinada i la seva contribució energètica a l'energia total de la unió DNA-proteïna (Benos et al., 2002; Stormo, 1990, 2000; Stormo and Fields, 1998). Diferents programes s'han desenvolupat que utilitzen matrius de pes per fer les prediccions: Signal Scan 4.0 (Prestridge, 1996; Chen et al., 1995), MatInspector (Quandt et al., 1995), ConsInspector (Frech et al., 1997a), MATCH (Kel et al., 2003), tfscan (una de les aplicacions d'EMBOSS (Rice et al., 2000)), TESS (Schug and Overton, 1997), ConSite (Lenhard et al., 2003), . . . Utilitzen matrius pròpies o bé les definides en la base de dades de TRANSFAC (Matys et al., 2003). Però pocs d'ells permeten fer prediccions sobre vàries seqüències a la vegada i utilitzar matrius específiques de qualsevol nivell taxonòmic o espècie.

En el cas (2), detecció de TFBSs, s'han proposat també una gran diversitat de tipus d'algorisme: alineament múltiple (Hertz and Stormo, 1999), xarxes neuronals (Heumann et al., 1994; Workman and Stormo, 2000), *hidden Markov models* (Pesole et al., 2000), *suffix trees* (Pavesi et al., 2001; Brazma et al., 1998), *Gibbs sampling* (Lawrence et al., 1993), *expectation maximization* (EM) (Bailey and Elkan, 1994, 1995), *random projections* (Buhler and Tompa, 2002), ... La detecció de TFBSs es pot utilitzar per a l'anomenat *phylogenetic footprinting*, cerca de motius comuns entre promotors de gens ortòlegs (Hardison, 2000; McCue et al., 2001; Tompa, 2001; Dermitzakis and Clark, 2002; Lenhard et al., 2003). També es pot utilitzar per analitzar els promotors de gens que tenen similar perfil d'expressió.

Complexitat del promotor dels gens d'organismes multicel·lulars

En els organismes multicel·lulars en que existeixen diferents tipus cel·lulars i les cèl·lules s'organitzen en diferents teixits i òrgans, la regulació transcripcional és la principal responsable d'assegurar l'expressió dels gens en el moment, el lloc (tipus cel·lular, teixit) i al nivell adequats. Les mutacions en els motius reguladors poden modificar l'afinitat d'unió dels factors de transcripció i afectar l'expressió gènica. Aquests canvis en l'expressió dels gens poden produir importants alteracions fenotípiques.

Les tècniques de *microarrays* de DNA, molt desenvolupades en els darrers anys, permeten la caracterització de perfils d'expressió gènica (Su et al., 2004; Zhang et al., 2004). Podem així identificar els gens que s'expressen en cada teixit. Però la informació concernent a les xarxes de regulació transcripcional responsables dels patrons d'expressió observats no està continguda en les seqüències cDNA dels *arrays*, sinó en les seqüències reguladores dels gens corresponents. La seqüenciació completa dels genomes proveeix la base per identificar els promotors i l'anàlisi comparativa, intraespecífica i interespecífica, d'aquests és clau per determinar les xarxes de regulació gènica (Werner, 2001; Dermitzakis and Clark, 2002; Lenhard et al., 2003).

Una qüestió no resolta és en quin grau la comparació de genomes pot contribuir a la identificació *in silico* de regions reguladores de la transcripció. Ens preguntem si pot ser tant útil com ho ha estat per a la identificació de gens (la part codificant dels gens) i quins factors fan que els resultats de la comparació genòmica siguin més difícils d'analitzar per aquestes regions.

Com passa amb les seqüències codificants, les seqüències promotores semblen tenir taxes evolutives molt diverses (Wray et al., 2003). Hi han casos de conservació molt elevada entre les espècies, però també casos de ràpida divergència, fins i tot entre espècies molt properes. Els canvis en la seqüència promotora poden produir-se per diferents tipus de mutacions: des de mutacions locals, a petita escala, fins a insercions i delecions, a escala més àmplia (Wray et al., 2003). Cal tenir en compte que les diferències en la seqüència promotora poden o no alterar la transcripció. De fet, els efectes dels canvis de la seqüència promotora sobre la transcripció, i finalment sobre el fenotip, varien ampliament (Wray et al., 2003): poden produir-se canvis dràstics, però normalment l'efecte és neutral.

Un problema fonamental a l'hora de fer l'anàlisi comparativa de promotors és el fet que els mòduls i les regions reguladores no tenen uns límits ben definits. Prácticament tots el gens d'eucariotes tenen un promotor basal, localitzat aproximadament entre el 100 bp *upstream* del TSS i el propi TSS, on es produeix la unió del complex iniciador de la transcripció. Més *upstream* dels 100 bp, les seqüències reguladores varien àmpliament en longitud en els diferents gens (Wray et al., 2003). De totes formes, sembla ser que a partir 2 Kb *upstream* del TSS la semblança entre gens ortòlegs es redueix dràsticament, indicant que la majoria de TFBS es concentren en la regió promotora de 2 Kb (Keightley et al., 2005). Assaigs funcionals en cèl·lules en cultiu mostren que la regió que va de -500 a +50 en relació al TSS és suficient per produir transcripció de la majoria dels gens humans (Trinklein et al., 2003).

En mamífers, la conservació de les seqüències *upstream* del TSS està relacionada amb la funció del gen (Iwama and Gojobori, 2004; Lee et al., 2005). Els factors de transcripció i els gens involucrats en processos adaptatius i complexos (desenvolupament, comunicació cel·lular, funció neural, senyal) tenen un promotor més conservat, per tant amb més TFBSs. Els gens involucrats en processos bàsics, com el metabolisme i la funció ribosomal, tenen promotors poc conservats, el que indica que són més simples.

Queda per saber si, independentment de la funció, hi ha diferències en la conservació de les seqüències *upstream* en funció de l'amplitud d'expressió dels gens (entenent per amplitud d'expressió el nombre de teixits en que s'expressen). Tenen els gens *housekeeping*, els que s'expressen en gairebé tots els teixits, un promotor més simple? I els gens que s'expressen en pocs teixits?

Duplicació gènica i evolució de les seqüències reguladores de la transcripció

S'HA POSTULAT UN PAPER PREPONDERANT DE LES DUPLICACIONS GÈNIQUES en l'evolució de nous gens i nous fenotips. Segons Ohno, l'evolució dels gens i els genomes és típicament conservativa en absència de duplicacions gèniques (Ohno, 1970). La duplicació d'un gen sol ser el primer pas en la creació d'un nou gen amb noves funcionalitats. S'ha estimat que al menys un 50 % del gens procariotes i més d'un 90 % dels gens eucariotes són producte de duplicacions gèniques (Brenner et al., 1995; Teichmann et al., 1998; Gough et al., 2001).

La duplicació gènica és un procés molt important en l'evolució de famílies gèniques. Un exemple clàssic molt conegut és la família dels gens HOX. Un altre exemple és la família de les globines. Mentre que els invertebrats solen tenir un únic gen globina, en mamífers hi ha diverses globines (mioglobina, hemoglobines) que s'han generat per duplicacions gèniques en diferents moments de l'evolució (figura 3). Aquests gens solen tenir diferents perfils d'expressió. Per exemple, els cinc gens de β-hemoglobina humana, agrupats dins una regió del cromosoma 11 (figura 4), s'expressen en diferents moments de l'evolució: ϵ-hemoglobina s'expressa en el sac embrionari, γ-A-hemoglobina i γ-G-hemoglobina s'expressen al fetus, mentre que δ-hemoglobina i β-hemoglobina s'expressen a l'adult.

La duplicació génica es pot produir per diferents vies (Prince and Pickett, 2002). Un mecanisme a gran escala és la duplicació completa del genoma mitjançant poliploïdia. La poliploïdia s'ha postulat per explicar el gran nombre de duplicacions gèniques en peixos (Van de Peer et al., 2003) i plantes (Raes et al., 2003). També és la base de la hipòtesi "2R" que intenta explicar el gran nombre de duplicacions gèniques en vertebrats per dues sèries (2 *rounds*) de duplicacions completes del genoma en l'origen del llinatge dels vertebrats (Sidow, 1996; Meyer and Schartl, 1999; Wolfe, 2001).

Un altre mecanisme, a escala més local, és l'entrecreuament desigual (*unequal crossing-over*) en el procés de recombinació durant la meiosi. Un entrecreuament desigual es dona quan hi ha un alineament

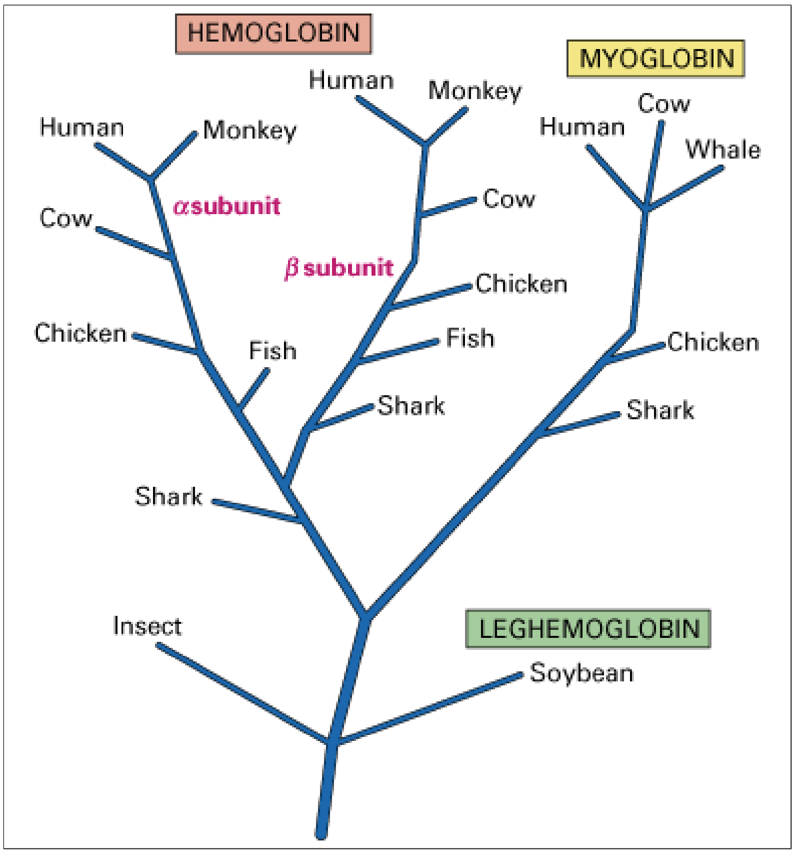


Figura 3 Arbre filogenètic de la família de les globines. Adaptat a partir de Lodish et al. 2007.

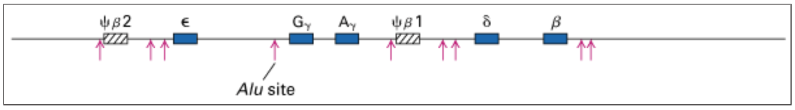


Figura 4 Gens de la família de les β-globines humanes. Els gens de la família de les β-globines humanes s'agrupen dins una regió del cromosoma 11. Són 5 gens: ϵ-hemoglobina, γ-G-hemoglobina, γ-A-hemoglobina, δ-hemoglobina i β-hemoglobina. S'hi troben també dos pseudogens Ψβ1 i Ψβ2. Adaptat a partir de Lodish et al. 2007.

incorrecte entre gens durant la meiosi i la recombinació produeix gàmetes amb diferent nombre de còpies de gens. Depenent de la mida de DNA mal alineat, un entrecreuament desigual pot involucrar més d'un gen, un sol gen o només una part d'un gen. Els alineaments incorrectes són especialment probables entre famílies de gens, conjunts de gens que codifiquen proteines amb seqüències molt similars (Lodish et al., 2007). Això és deu a que cada membre sol tenir també un alt grau de similitud de seqüència genòmica amb la resta i a més, com a resultat del mateix mecanisme d'entrecreuament desigual, aquests gens duplicats estan sovint localitzats un a prop de l'altre en una mateixa regió cromosòmica (duplicacions en tàndem). És a dir, és un procés amb retroalimentació positiva i és per això que és particularment important en l'augment de nombre de còpies de les famílies gèniques.

Un darrer mecanisme important de duplicació gènica és la retrotransposició (Soares et al., 1985). En aquest cas, la còpia es crea per transcripció inversa a partir un RNA missatger madur (sense introns), originant una còpia amb un únic exó quan el gen original era multiexònic. La majoria dels gens generats per retrotransposició perden el lligam físic amb el gen original, és a dir, es situen en un altre cromosoma o en una regió allunyada dins del mateix cromosoma.

Un altre aspecte molt important de les duplicacions gèniques és el mecanisme pel que es produeix la preservació de les còpies. La hipòtesi clàssica d'Ohno suggereix que el mecanisme de duplicació gènica produeix còpies redundants amb la capacitat d'acumular mutacions que poden conduir a la pèrdua o guany de funció, mentre una de les còpies no varia i manté la funcionalitat ancestral (Ohno, 1970). Segons aquesta hipòtesi, la retenció de les còpies redundants només es produeix per neofuncionalització, guany de funció, i requereix selecció positiva. La gran majoria de les còpies es perd per nofuncionalització (o pseudogenització), és a dir, pèrdua de funció. El gran problema d'aquesta hipòtesi és que no explica l'enorme quantitat de duplicacions gèniques funcionals que existeixen a la majoria de genomes d'eucariotes (Nadeau and Sankoff, 1997; Li et al., 2001; Postlethwait et al., 2000; Initiative, 2000).

Una hipòtesi alternativa és la proposada als anys 90 per Hughes i Lynch, basada en processos purament neutrals. Segons aquesta hipòtesi la retenció de les còpies es produeix bàsicament per subfuncionalització, pèrdua de part de la funció en cadascuna de les còpies per mutacions degeneratives, amb complementació de les funcions de les còpies per mantenir la funcionalitat ancestral (Hughes, 1994; Force et al., 1999; Lynch and Force, 2000). D'acord amb aquesta hipòtesi, Lynch i Force vàren proposar un model anomenat DDC, duplicació-degeneració-complementació (Force et al., 1999). Segons aquest model el mecanisme usual de preservació de les duplicacions gèniques seria la partició entre les còpies dels patrons d'expressió ancestral que s'iniciaria per mutacions degeneratives en elements reguladors.

Tot i que el procés DDC està basat totalment en mutacions degeneratives, hi ha almenys tres vies per les que pot jugar un paper molt important en la generació de novetat biològica (Lynch and Force, 2000). Primer, a l'estabilitzar les duplicacions gèniques dins del genoma, aquest procés estén el període de temps en que els gens estan exposats a la selecció natural, augmentant la possibilitat de mutacions beneficioses, que produeixin noves funcions. Segon, la partició de l'expressió gènica en les còpies duplicades pot reduir les restriccions de pleiotropia que operen en un únic gen, permetent que la selecció natural afini les subfuncions específiques de cada còpia. Tercer, les duplicacions gèniques amb subfuncions no resoltes quan es produeix un esdeveniment d'isolament reproductiu podrien ser un important mecanisme en el

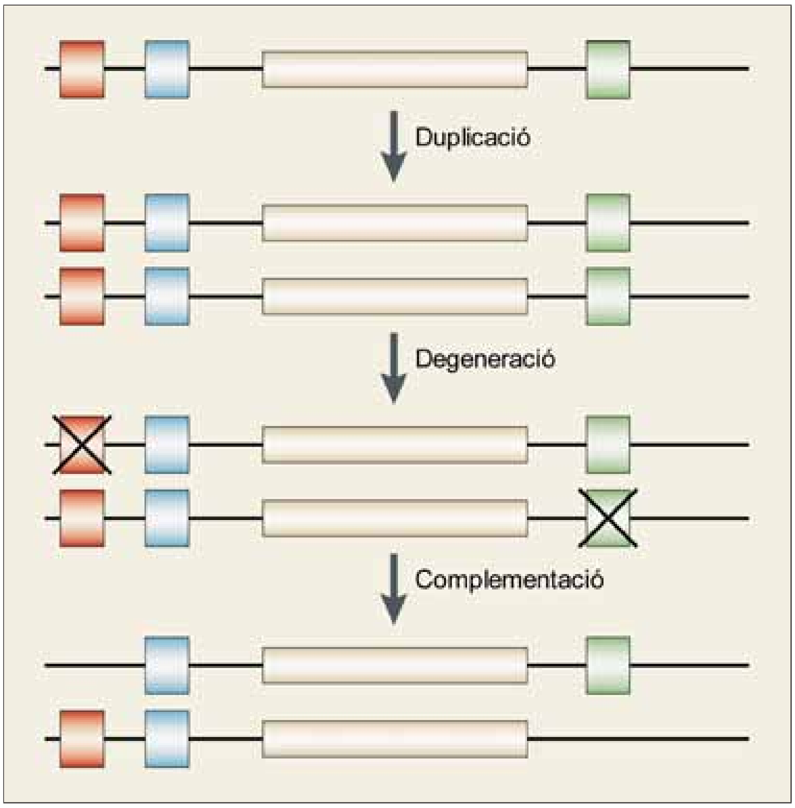


Figura 5 **Model DDC (duplicació-degeneració-complementació).** Els rectangles vermells, blaus i verds simbolitzen elements reguladors (TFBSs o mòduls de TFBSs); el rectangle gris correspon a la part codificant. Les creus indiquen mutacions degeneratives en una de les còpies, de manera que els duplicats junts mantenen les funcions original del gen ancestre. Adaptat a partir de Prince and Pickett 2002.

procés d'especiació.

Donat que les mutacions degeneratives són molt més freqüents que les mutacions benèfiques i que molts gens tenen complexes regions de regulació que controlen l'expressió en diferents teixits, la subfuncionalització seria un mecanisme de preservació dels gens duplicats més versemblant que la neofuncionalització.

En els darrers anys s'han proposat diferents hipòtesis híbrides que combinen subfuncionalització i neofuncionalització (He and Zhang, 2005; Rastogi and Liberles, 2005). De fet Lynch i Force no negaven que a més de subfuncionalització es produís també neofuncionalització, però remarcaven el paper molt més important de la subfuncionalització. Aquestes noves hipòtesis, tot i admetre el paper clau de la subfuncionalització en la retenció a curt termini dels gens duplicats, remarquen el paper fonamental a llarg termini de la neofuncionalització per a la creació de novetat biològica. És a dir, la subfuncionalització és considerada un estat transitori cap a la neofuncionalització. Com ja indicaren Lynch i Force, aquestes noves hipòtesis assenyalen també que la importància relativa de la subfuncionalització i de la neofuncionalització depèn de la mida de la població; com més petita és la mida de la població, més importància té la subfuncionalització en la preservació dels gens duplicats.

S'ha observat que els gens duplicats tendeixen a tenir proteïnes amb taxes evolutives més grans (Lynch and Conery, 2000; Van de Peer et al., 2001; Kondrashov et al., 2002; Nembaware et al., 2002; Castillo-Davis et al., 2004; Cusack and Wolfe, 2007; Mikkelsen et al., 2007). D'altra banda, les duplicacions gèniques estan associades a canvis substancials en el patrons d'expressió gènica (Huminiecki and Wolfe, 2004). A més, s'ha trobat una relació positiva entre la divergència d'expressió i la divergència de la seqüència proteica en els gens que han experimentat duplicacions (Gu et al., 2002; Makova and Li, 2003; Ganko et al., 2007).

Com s'ha indicat, el model DDC es va descriure originalment en base a mutacions degeneratives en elements reguladors (Force et al., 1999). Diferents investigadors han emfatitzat que els canvis evolutius més importants poden haver-se produït principalment a nivell de la regulació gènica més que a nivell de la funció de la proteïna (Yuh et al., 2001; Carroll, 2000). Tot i així, pocs estudis s'han fet en relació a l'evolució de les seqüències reguladores de la transcripció en les duplicacions gèniques. Un estudi força complet d'un cas concret és el de McClintock et al. que demostra subfunciolització en la duplicació gènica de Hoxb1 a *zebrafish* (McClintock et al., 2002; Prince and Pickett, 2002).

Sabem que una substitució nucleotídica pot originar el guany o la pèrdua d'un TFBS, afectant la regulació d'un gen. De fet, s'ha observat un alt moviment de motius reguladors entre gens ortòlegs d'*Homo sapiens* i *Mus musculus* (Dermitzakis and Clark, 2002; Odom et al., 2007). El truncament o la pèrdua total del promotor, per exemple quan es produeix retrotransposició, probablement ha de produir fortes asimetries entre les dues còpies (Cusack and Wolfe, 2007). De fet, s'ha publicat la tendència a una divergència de la seqüència promotora més alta en paràlegs que en ortòlegs, en *C. elegans* i *C. briggsae* (Castillo-Davis et al., 2004), però queda per investigar la relació d'aquest canvis en el promotor respecte a la divergència de l'expressió gènica.

Part II

OBJECTIUS

Objectius

El treball realitzat durant el meu doctorat ha estat dirigit a entendre millor l'organització i l'evolució de les regions de DNA que regulen la transcripció.

En concret els objectius d'aquesta tesi són:

- Millorar els mètodes *in silico* de caracterització, representació i predicció de llocs d'unió de factors de transcripció.
- Estudiar les restriccions que actuen sobre la conservació evolutiva de les seqüències reguladores de la transcripció.
- Estudiar l'efecte de la duplicació gènica sobre l'evolució de les seqüències reguladores de la transcripció i sobre l'expressió gènica.

Part III

RESULTATS

Capítol 1

Predicció de llocs d’unió de factors de transcripció

Resum

El treball presentat en aquest capítol és fruit de la col·laboració entre el Grup d’Algorísmica i Genètica, del Departament de Llenguatges i Sistemes Informàtics de la Universitat Politècnica de Catalunya, i el laboratori de Genòmica i Bioinformàtica de Virus del Wohl Virion Centre, del Departament d’Immunologia i Patologia Molecular del University College London. Un dels resultats d’aquesta col·laboració és el programa PROMO, per predir llocs d’unió de factors de transcripció (TFBSs) en una seqüència de DNA o en un conjunt de seqüències. A partir de la versió 2.0 sóc el principal autor i responsable de PROMO, però en el projecte han intervingut diferents col·laboradors al llarg dels anys (R. Escudero, O. Núñez, J. Martínez, L. Roselló, D. García) a més a més dels meus directors (X. Messeguer i M.M. Albà).

Aquest capítol inclou 2 publicacions referents a PROMO:

Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM: **PROMO: detection of known transcription regulatory elements using species-tailored searches**. Bioinformatics 2002, 18:333-334.

Farré D, Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X: **Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN**. Nucleic Acids Res 2003, 31:3651-3653.

Per una espècie determinada o un determinat nivell taxonòmic, PROMO construeix matrius de pes a partir de seqüències de TFBSs coneguts que són utilitzades per fer la predicció. La predicció simultània sobre múltiples seqüències permet descobrir elements de regulació comuns i es pot aplicar, per exemple, per a comparar regions reguladores de gens ortòlegs o de gens que s’expressen en el mateix teixit.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 18 no. 2 2002
Pages 333–334

PROMO: detection of known transcription regulatory elements using species-tailored searches

Xavier Messeguer[1], Ruth Escudero[1], Domènec Farré[1], Oscar Núñez[1], Javier Martínez[1] and M.Mar Albà[2,*]

[1]Algorithmics and Genetics Group, Software Department, Universitat Politècnica de Catalunya, C/Jordi Girona Salgado, 1–3, 08034 Barcelona, Spain and [2]Virus Genomics and Bioinformatics, Wohl Virion Centre, Department of Immunology and Molecular Pathology, University College London, 46 Cleveland St, London W1T 4JF, UK

Received on June 5, 2001; revised on August 3, 2001; accepted on September 12, 2001

ABSTRACT

Summary: We have developed a set of tools to construct positional weight matrices from known transcription factor binding sites in a species or taxon-specific manner, and to search for matches in DNA sequences.

Availability: PROMO can be accessed online at http://www.lsi.upc.es/∼alggen under the research link.

Supplementary information: An example of the graphic interface (Figure 1) can be visualized at http://www.lsi.upc.es/∼alggen/recerca/promo/figuraBioinformatics.html.

Contact: peypoch@lsi.upc.es; m.alba@ucl.ac.uk

One of the major challenges that follow the sequencing of genomes is to unravel the gene expression regulatory networks that operate in different types of cells. The transcription of a gene typically requires and is regulated by a number of cellular factors, that recognize and bind to short sequence motifs, in many cases located upstream of the gene coding sequence, in the so-called promoter and enhancer regions. Genes expressed in the same tissue or under similar conditions often share common regulatory motifs (Wasserman and Fickett, 1998), therefore the motifs found in a gene can be understood as a 'footprint' of its transcriptional regulatory mechanisms and to some extent gene function. The *in silico* prediction of potential regulatory sites is therefore a valuable tool to characterize new genes and to limit the amount of protein–DNA interactions to be tested experimentally.

Many binding sites for transcription factors have been experimentally identified and this information can be used to perform computational-based searches. A number of public databases store information on individual transcription factors and their binding sites, such as TRANSFAC (Wingender *et al.*, 2001) or RegulonDB (Salgado *et al.*, 2001). Due to the intrinsic sequence variability of the motifs recognized by particular regulatory proteins appropriate representations of the sites are IUPAC consensi or positional weight matrices (Bucher, 1990). The latter store the frequency of the different nucleotides in the different positions of the motif and are generally considered superior as they are more specific and allow rating of the matches (Frech *et al.*, 1997). The TRANSFAC database (Wingender *et al.*, 2001) contains the largest available collection of eukaryotic factor-specific weight matrices, which can be used to search for potential matches in a DNA sequence of interest, for example by the MatInspector program (Quandt *et al.*, 1995).

The TRANSFAC collection of matrices is subdivided into very broad taxonomic groups (vertebrates, fungi, plants, insects and miscellaneous). The lack of flexibility in the taxonomy levels that can be considered may lead to problems in the interpretation of the results, specially when the binding sites have only been identified in a species which is distantly related to the one under study. In addition, searching with large collections of matrices may result in an increment in the number of false positives in the predictions, a general problem when attempting to identify short and variable sequences such as transcription factor binding sites. Bearing in mind these caveats we have developed a new approach to perform searches with weight matrices, which allows the user to tailor the searches to the species or group of species of interest. By selecting a particular species instead of a general group of organisms more specificity in the searches can be achieved. If few sites are known for the species under study selecting matrices from related species may still provide valuable information. Comparing different

*To whom correspondence should be addressed.

species settings may be useful to analyze the cross-species conservation of particular known binding sites. Additional novel features of PROMO are the generation of the factor-specific matrices on the fly and the incorporation, as part of the output, of information on other genes which are known to be regulated by the subset of transcription factors that appear in the prediction.

PROMO has been written in C^{++} and includes different modules, all available through a web server. The complete collection of TRANSFAC site, factor and gene entry files is used as a source of sequences and information. The species, or group of species, of interest is selected by the user. After the species selection weight matrices are automatically derived from at least three different binding sites per transcription factor, by anchoring the alignment of the relevant sequences on the completely conserved positions or 'core' of the binding site. An automata is then constructed which contains all the different possible subsequences that score above a given similarity threshold to any of the matrices. The similarity of a sequence to a matrix is calculated according to Quandt *et al.* (1995) and the default similarity threshold used by the program is 85% (or dissimilarity 15%). Exact matches of the query sequence to the automata represent putative transcription factor binding sites in the sequence. The two steps that the user is required to perform are: (1) 'SelectSpecies', select the species or taxonomic group of interest by using a taxonomic tree derived from the organism annotations in TRANSFAC site and factor entries and; (2) 'SearchSite', input a query sequence to search for matches to the matrices in any of the two strands. Other available options are 'ViewMatrices' and 'MatrixSpecificity'. The first one allows the visualization of the matrices that have been constructed including information on genes known to contain sites represented in the matrices. The second option is a Java applet for the comparison of the specificity of matrices corresponding to pre-defined taxonomic groups (see below for a definition of specificity). After steps 1 and 2 the program typically takes a few seconds to run and the results are presented online (example in supplementary material, Figure 1). The output includes the following: matches of the sequence to the factor-specific matrices in the corresponding sequence location, including the name of the factor that binds to the motif and dissimilarity percentage; expectation values of finding the different matches by chance alone, using a model with equiprobability of the four nucleotides, or a model with nucleotide frequency as in the query sequence and; information on the location of the predicted regulatory sites, either individually or combined, in other genes. The latter feature includes a graphical representation of the different sites in the regulatory regions of the genes, following the annotations in the TRANSFAC gene entries. The information on other genes may be very useful as the observation of functional relatedness with the gene of interest may highlight particularly interesting hits.

When no species restriction is applied the number of matrices that PROMO generates is 452, derived from 4308 different sites. A variable number of matrices are created for different organism groups, for example 268 for animals, 26 for fungi, 19 for plants, 245 for vertebrates and 55 for humans. A simple way to measure and compare the specificity of different matrices is by defining the specificity for each position in the sequence as the distance between a vector representing the probability of each nucleotide and a vector where all nucleotides are equiprobable and the specificity of a matrix as a whole as the normalized average distance for all columns. Decreasing the taxonomic level under consideration leads to the expected increase in specificity when comparing matrices for the same transcription factor (same TRANSFAC entry), which will lead to a less noisy output. For example when we compare equivalent matrices from humans and vertebrates (55 different matrices), the human matrices are more specific ($p < 10^{-2}$), the average specificity being 0.8 in contrast to 0.7 for the vertebrate matrices. Future developments we envisage are the explicit modelling of combinations of factor binding sites and the use of additional regulatory site databases.

ACKNOWLEDGEMENT

This work was partially supported by project DGI BIO2001-2199.

REFERENCES

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

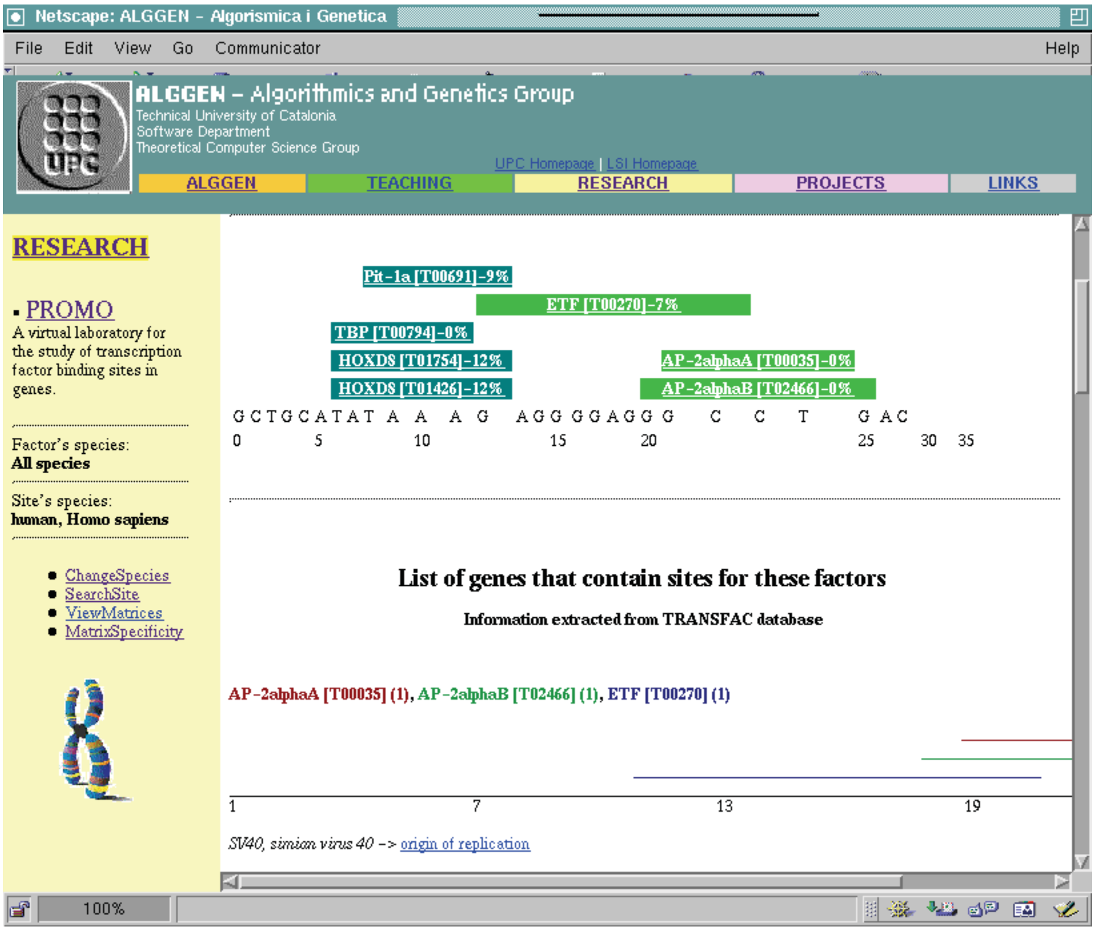
Frech,K., Quandt,K. and Werner,T. (1997) Finding protein-binding sites in DNA sequences: the next generation? *Trends Biochem. Sci.*, **22**, 103–104.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C. and Collado-Vives,J. (2001) RegulonDB (Version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.

Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.

Wingender,E., Chen,X., Fricke,R., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R., Prüss,M., Schacherer,F., Thiele,S. and Urbach,S. (2001) The TRANSFAC system on gene expresssion regulation. *Nucleic Acids Res.*, **29**, 281–283.



ALGGEN - Algorithmics and Genetics Group

Technical University of Catalonia
Software Department
Theoretical Computer Science Group

UPC Homepage | LSI Homepage

RESEARCH

- PROMO
A virtual laboratory for the study of transcription factor binding sites in genes.

Factor's species:
All species

Site's species:
human, Homo sapiens

- ChangeSpecies
- SearchSite
- ViewMatrices
- MatrixSpecificity

List of genes that contain sites for these factors

Information extracted from TRANSFAC database

AP-2alphaA [T00035] (1), AP-2alphaB [T02466] (1), ETF [T00270] (1)

SV40, simian virus 40 -> origin of replication

Supplementary material - Figure 1

Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3651–3653***
DOI: 10.1093/nar/gkg605

Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN

Domènec Farré, Romà Roset[1], Mario Huerta[2], José E. Adsuara[2], Llorenç Roselló[2], M. Mar Albà[3] and Xavier Messeguer[1,2,*]

Computing Unit, Institut de Recerca Oncològica, L'Hospitalet, Spain, [1]CEPBA-IBM Research Institute, [2]Algorithmics and Genetics Group, Software Department, Universitat Politècnica de Catalunya, Barcelona, Spain, [3]Biomedical Informatics Research Group, Health and Experimental Sciences Department, Universitat Pompeu Fabra, Barcelona, Spain

Received February 14, 2003; Revised and Accepted April 4, 2003

ABSTRACT

In this paper we present several web-based tools to identify conserved patterns in sequences. In particular we present details on the functionality of PROMO version 2.0, a program for the prediction of transcription factor binding site in a single sequence or in a group of related sequences and, of MALGEN, a tool to visualize sequence correspondences among long DNA sequences. The web tools and associated documentation can be accessed at http://www.lsi.upc.es/~alggen (RESEARCH link).

INTRODUCTION

The sequencing of a large number of genomes has greatly stimulated the development of computational methods for the identification of signals or patterns in biological sequences. Conserved patterns, preserved during evolution, may be indicative of functionality and generate testable hypotheses. In our group we have developed a number of pattern-search algorithms and web-based tools that can assist in the discovery of biological function: TRANSPO (1), to search for miniature inverted repeats transposable elements in genomic sequences; MREPATT (Roset *et al.*, manuscript submitted), to identify statistically meaningful consecutive repeated patterns in multiple genomes; MALGEN, to detect sequence motifs that are conserved among two or more very large sequences; and PROMO (2), to identify transcription factor binding sites in one or more sequences. In the ALGGEN web server one can also access clustering tools for DNA sequences using spanning trees (1) and an assembly program for sets of EST (expressed sequence tag) sequences. To facilitate the use of the programs we have developed very time-efficient algorithms so that, in most cases, the output can be provided online in a matter of seconds. The web interface of the different programs has been designed to be as user-friendly as possible.

Among the programs that can be accessed at our server, the present paper focuses on MALGEN and PROMO version 2.0. TRANSPO and MREPATT are extensively documented in recent publications. A first version of PROMO has also been published (2), but we have made significant improvements and added new features that justify its treatment in this paper.

PROMO VERSION 2.0

One of the most challenging aspects of genome biology is the understanding and modelling of the gene expression regulatory networks that operate in cells and tissues. The reliable identification of transcription factor binding sites in DNA sequences is an important step. To predict binding sites in sequences one can use the available information on known target sequences in regulatory regions of genes. Several databases contain collections of known binding sites, such as TRANSFAC (3), which contains the largest available collection of DNA binding sites in eukaryotes. Given the intrinsic variability of the protein recognition signals an appropriate representation of the binding sites are positional weight matrices (4), which store information of the relative frequency of different nucleotides in the recognition sites. In PROMO, weight matrices are constructed from known binding sites extracted from TRANSFAC and used for the identification of potential binding sites in sequences. A number of other programs exist for the prediction of transcription factor binding sites that use weight matrices (5,6) but PROMO contains a number of unique features. Among them we would like to highlight the following: (i) the possibility to select sites from any species or group of species of interest; (ii) the automatic construction of matrices that correspond to the selected taxonomic level; (iii) information in the output on other genes that may be similarly regulated; and (iv) the possibility to analyze and compare multiple sequences at the same time.

The first step when using PROMO is the selection of species or taxonomic level, both for factors and binding sites. This is aided by a Java applet that can be accessed from 'SelectSpecies' at the main menu. After the species selection, the matrices are constructed on the fly and can then be

[*]To whom correspondence should be addressed at Algorithmics and Genetics Group, Software Department, Universitat Politècnica de Catalunya, Jordi Girona 1-3, C6-117, Barcelona 08034, Spain. Tel: +34 93 4017333; Fax: +34 93 4017014; Email: messeguer@lsi.upc.es

Published by Oxford University Press 2003

inspected using the 'ViewMatrices' option. Subsequently, the user can enter the sequence at the 'SearchSites' form page. The query sequence is scanned for sites with high similarity to the matrices (6). To optimize the search time we use an automata that contains all possible subsequences in the query sequence that score above the similarity threshold to any of the matrices. The output contains a graphical representation of the predicted binding sites, expectation values to assess the significance of the matches and a list of genes that are known to be regulated by the transcription factors that appear in the predictions, either individually or in all possible combinations. The information on other genes may be very useful, as the observation of functional relatedness between these genes and the gene under study may point to particularly relevant hits. The main PROMO menu also contains an option to visualize the specificity of matrices derived from different groups of organisms, 'MatrixSpecificity' (2) and a help page.

The construction of matrices in PROMO is an automated process. The algorithm finds, given *n* factor-specific binding site sequences, the subset of at least *n*/2 sequences which results in the longest number of consecutive completely conserved positions. By doing this we maximize specificity while keeping a representative number of sequences. For example, from 34 binding sequences available for the AP-1 transcription factor, we use 20 sequences to construct the matrix, as we have determined that 20 (higher than 34/2), but no more, can be aligned with five conserved positions. We recently tested the algorithm by comparing the results obtained using the factor-specific binding sites with those obtained with random sequences of the same length and composition as the binding sites considered. Matrices with the same number of conserved sites as expected by chance are rare and discarded by the program. In the example above the random model resulted in an expectation of three conserved positions. The matrix derived from the AP-1 binding sites, with five conserved positions, is clearly significant. The number of matrices depends on the species or taxonomic level selected by the user. For example, in the current version the program generates 503 matrices when all species are considered, 313 when only sites from animals are used and 163 when only sites from human sequences are used.

The prediction of transcription factor binding sites using weight matrices derived from collections of known sites is likely to detect the occurrence of existing sites in a sequence but will also result in the prediction of many sites which are not real, that is, false positives (5). This is a consequence of the fact that binding sites tend to be short and therefore they have a high probability of occurring by chance in any sequence. Thus, although computational prediction clearly reduces the candidate number of regulatory factors to be tested, it is not sufficient to obtain a reliable map of gene expression regulatory elements in a sequence of interest. Other biological support needs to be sought. In particular, comparative analysis of functionally related genes may provide very valuable information as these genes may share regulatory elements. Functionally related genes may be those that show similar expression patterns, as determined by array-based experiments (7) or those that have a common ancestor (orthologs). A new module of PROMO, in version 2.0, 'MultiSearchSites', has been designed to identify those binding sites that are present in several, or all, out of a set of user input sequences, which may for example correspond to a cluster of similarly expressed genes. For the analysis of a group of related sequences the 'MultiSearchSites' option, instead of the 'SearchSites' option, should be selected from the main menu. Parameters that can be modified by the user are the percentage of sequences that are required to contain a match to the binding site so that the match is reported and the similarity threshold used in the predictions. The requirement that several sequences must contain the match reduces the number of total predictions while keeping those that may be more relevant. An example is shown in Figure 1, where sites above 85% similarity and present in all sequences are reported. The example corresponds to the regulatory region of the cardiac alpha-actin gene from four different vertebrate species. The prediction of the SRF (serum response factor) binding site corresponds to the experimentally verified site (8).

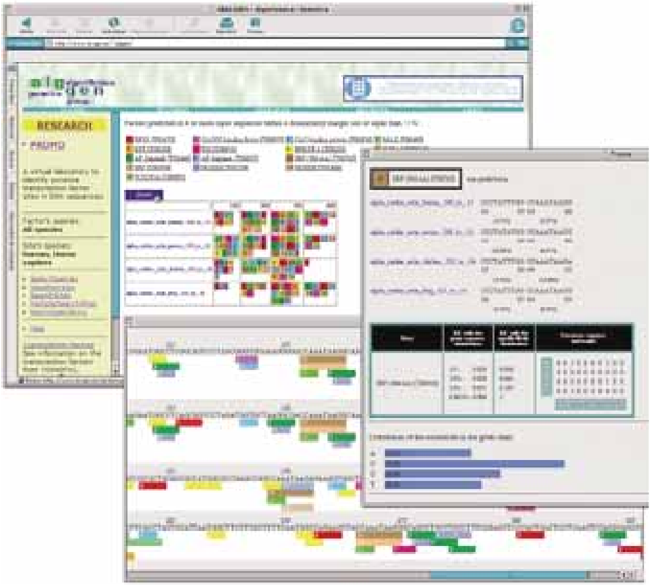


Figure 1. PROMO 'MultiSearchSites' output example. The example corresponds to the regulatory region of the cardiac alpha-actin gene from four different vertebrate species: humans, mouse, chicken and frog. Only those binding site predictions that appear in all four sequences are shown, as boxes of different colour and number. The image below, where the sequences are shown, is the result of selecting 'Zoom' in the main results page above. The image on the right is a detail of the SRF (serum response factor)—binding site predictions on the sequences. It also shows the weight matrix for the SRF recognition site and random expectation (RE) values for different levels of sequence-matrix similarity. The RE is calculated with a model that considers that all nucleotides are equally probable and also with a model that considers the nucleotide composition in the query sequence (in the picture represented by blue bars below matrix).

MALGEN

MALGEN is the acronym of Multiple ALignment of GENomes and it is a web tool to explore sequence relationships among large DNA sequences. Sequence segments of a minimum user-defined length present in two different sequences are identified and represented graphically (see examples in Fig. 2). Regions of identity, or matching

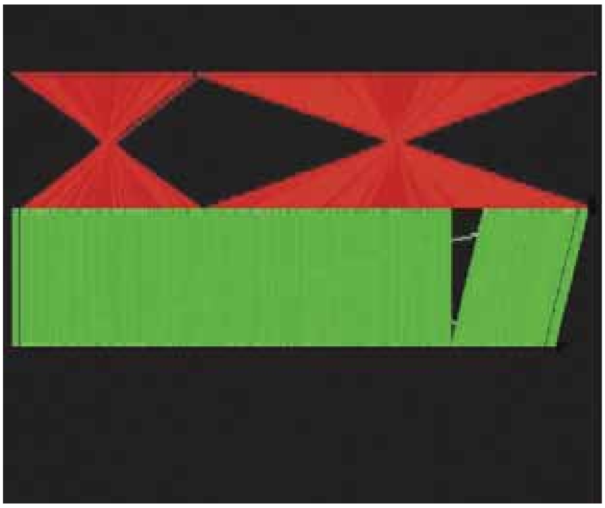


Figure 2. Comparison of three *C.pneumoniae* strain genomes: AR39 (1.247 Mb), CWL029 (1.248 Mb), J138 (1.175 Mb), in this order top to bottom. Horizontal white lines are the DNA sequences, vertical green lines are exact direct identities and vertical red line exact inverse identities.

segments, are marked with vertical lines, where green lines represent exact direct identities and red lines exact inverse identities. More than two sequences can also be represented. The identity segments are of maximum length, as they cannot be extended further, and they are unique, as they represent one-to-one correspondences, that is, matches that occur only once in each DNA sequence. For these reasons they are called Maximal Unique Matchings, for short MUMs. Note that the uniqueness property is a strong requirement but it may reinforce the biological interpretation of the matches.

The comparison of long DNA sequences is computationally expensive and not many tools for this purpose have so far been developed. Of the existing ones, MUMER v2 (9) can only compare two genomes and MGA (10) aligns multiple sequences but with a very high cost of space. We have designed an efficient space–time algorithm that allows the comparison between many genomes simultaneously. The algorithm only needs a linear space with respect to the shortest sequence. Its theoretical basis is described in detail elsewhere (11).

The web interface of MALGEN can be accessed from our web site under the Research and Align Tools links. The email of the user is required as his identifier. The user may access previously submitted jobs or, otherwise, start a new job. The process has two parts. In the first one the sequence files are provided by the user through the entry form and the server searches for the collection of MUMs, stores the list of MUMs as a new job and sends an email to the user. The list of MUMs is stored because their search is the most time-expensive process. In the second part, the user can access the existing job, containing the list of MUMs and generate a graphical representation on the fly. Different options for visualization are provided by a user-friendly interface. Options include the minimum length of the MUMs that will appear in the picture, the possibility to show direct or inverse matches, or both at the same time, the setting of the distance between consecutive sequences and the selection of the order of appearance of the sequences. The current MALGEN web tool runs the pair-wise version of the algorithm; the implementation for multiple genomes is in progress.

Depending on the nature of the sequences and the MUM minimum size, MALGEN may provide information on different kinds of signals. As the program can deal with very long sequences, it is particularly suited for the identification and visualization of chromosome, or genome, rearrangements among related species. At the same time it can provide an accurate and exhaustive mapping, in the form of MUMs, between chromosomes or genomes. Figure 2 shows a comparison between the genomes of three *Chlamidophila pneumoniae* strains (minimum MUMs size of 18 bases). In this case MUMs occupy 98% of the sequence. It can be observed that the superior one, which corresponds to the AR39 strain, contains two large inverse translocated segments in respect to the other two strains. Other examples can be visualized at the MALGEN site by submitting with the default email (malgen@lsi.upc.es). MALGEN can be used to identify different types of functional elements on genomes. For example, we are currently exploring its use in the identification of exons by comparative genomics. An additional use of MALGEN is in the alignment of very long sequences by using the MUMs to anchor the alignment (12).

ACKNOWLEDGEMENTS

ALGGEN group is supported by IST programme of the EU under contract IST-1999-14186 (ALCOM-FT).

REFERENCES

1. Santiago,N., Herraiz,C., Goñi,J.R., Messeguer,X. and Casacuberta,J.M. (2002) Genome-wide analysis of the emigrant family of mites of *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **19**, 2285–2293.
2. Messeguer,X., Escudero,R., Farré,D., Núñez,O., Martínez,J. and Alba,M.M. (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, **18**, 333–334.
3. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
4. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
5. Roulet,E., Fisch,I., Junier,T., Bucher,P. and Mermod,N. (1998) Evaluation of computer tools for the prediction of transcription factor binding site on genomic DNA. *In Silico Biol.*, **1**, 21–28.
6. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acid Res.*, **23**, 4878–4884.
7. Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.*, **9**, 681–688.
8. Sartorelli,V., Webster,K.A. and Kedes,L. (1990) Muscle-specific expression of the cardiac alpha-actin gene requires MyoD1, CArG-box binding factor, and Sp1. *Genes and Dev.*, **4**, 1811–1822.
9. Delcher,A.L., Phillippy,A., Carlton,J. and Salsberg,L. (2002) Fast algorithm for large-scale genome alignment and comparison. *Nucleic Acid Res.*, **11**, 2478–2483.
10. Höhl,M., Kurtz,S. and Ohlebush,E. (2002) Efficient multiple genome alignment. *Bioinformatics*, **18**, 1–9.
11. Huerta,M. and Messeguer,X. (2002) Efficient space and time multi-comparison of genomes. Research Report LSI-02-64-R, Dep. Llenguatge i Sistemes Informàtics, Universitat Politècnica de Catalunya.
12. Delcher,A.L., Kasif,S., Fleischmann,R.D., Peterson,J., White,O. and Salsberg,L. (1999) Alignment of whole genomes. *Nucleic Acid Res.*, **27**, 2369–2376.

Capítol 2

Expressió tissular i conservació del promotor en gens ortòlegs

Resum

El treball presentat en aquest capítol es va desenvolupar al Grup *Evolutionary Genomics*, format per membres de vàries institucions: Fundació IMIM (M.M. Albà), Universitat Pompeu Fabra (M.M. Albà, N. Bellora, L. Mularoni) i Centre de Regulació Genòmica (N. Bellora, D. Farré). També hi va intervenir el meu director, X. Messeguer, de la Universitat Politècnica de Catalunya. La major part del treball (90 %) va ser realitzada per mi amb la col·laboració d'en Loris Mularoni, del que vaig aprofitar els seus *scripts* per calcular Ka i Ks, i d'en Nicolás Bellora, que va intervenir en l'estudi de sobrerrepresentació de TFs.

Aquest capítol inclou la publicació:

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM: **Housekeeping genes tend to show reduced upstream sequence conservation**. Genome Biol 2007, 8:R140.

Entendre les restriccions que operen en les seqüències promotores dels gens de mamífers és d'importància clau per entendre l'evolució de les xarxes de regulació gènica. El nivell de conservació dels promotors varia enormement entre gens ortòlegs, indicant diferències en la intensitat de les restriccions evolutives. En aquest capítol es presenta l'estudi en que vàrem demostrar la hipòtesi de que el nombre de teixits en que un gen s'expressa està relacionat de manera significativa amb el grau de conservació de la seqüència del promotor.

Mostrem que els gens *housekeeping* de mamífers, els que s'expressen en tots o gairebé tots els teixits, tenen una menor conservació de la seqüència del promotor que els gens que s'expressen en un subconjunt de teixits. La diferència en la conservació de seqüència es fa més forta a partir de la posició -500, és a dir, més lluny del lloc d'inici de la transcripció. Això sembla suggerir una expressió gènica més simple en els gens *housekeeping*, regulada per un menor nombre de motius reguladors funcionals. De fet, vàrem identificar un subconjunt de factors de transcripció que s'uneixen a motius que estan especialment sobrerrepresentats en el promotors dels gens *housekeeping*.

Research

Open Access

Housekeeping genes tend to show reduced upstream sequence conservation

Domènec Farré[*], Nicolás Bellora[*†], Loris Mularoni[‡], Xavier Messeguer[§] and M Mar Albà[†‡¶]

Addresses: [*]Centre for Genomic Regulation, Dr Aiguader 88, Barcelona 08003, Spain. [†]Universitat Pompeu Fabra, Dr Aiguader 88, Barcelona 08003, Spain. [‡]Fundació Institut Municipal d'Investigació Mèdica, Dr Aiguader 88, Barcelona 08003, Spain. [§]Universitat Politècnica de Catalunya, Jordi Girona 1-3, Barcelona 08034, Spain. [¶]Catalan Institution for Research and Advanced Studies, Pg Lluis Companys 23, Barcelona 08010, Spain.

Correspondence: M Mar Albà. Email: malba@imim.es

Published: 13 July 2007

Genome ***Biology*** 2007, **8:**R140 (doi:10.1186/gb-2007-8-7-r140)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/7/R140

Received: 20 October 2006
Revised: 16 February 2007
Accepted: 13 July 2007

Abstract

Background: Understanding the constraints that operate in mammalian gene promoter sequences is of key importance to understand the evolution of gene regulatory networks. The level of promoter conservation varies greatly across orthologous genes, denoting differences in the strength of the evolutionary constraints. Here we test the hypothesis that the number of tissues in which a gene is expressed is related in a significant manner to the extent of promoter sequence conservation.

Results: We show that mammalian housekeeping genes, expressed in all or nearly all tissues, show significantly lower promoter sequence conservation, especially upstream of position -500 with respect to the transcription start site, than genes expressed in a subset of tissues. In addition, we evaluate the effect of gene function, CpG island content and protein evolutionary rate on promoter sequence conservation. Finally, we identify a subset of transcription factors that bind to motifs that are specifically over-represented in housekeeping gene promoters.

Conclusion: This is the first report that shows that the promoters of housekeeping genes show reduced sequence conservation with respect to genes expressed in a more tissue-restricted manner. This is likely to be related to simpler gene expression, requiring a smaller number of functional *cis*-regulatory motifs.

Background

The correct functioning of multicellular organisms depends on a complex orchestration of gene regulatory events, which ensure that genes are expressed at the right time, place and level. Much of this regulation occurs at the level of gene transcription, and is mediated by specific interactions between transcription factors and *cis*-regulatory DNA motifs. Regulatory motifs concentrate in sequences upstream of the transcription start site (TSS), the region known as the gene promoter (for a recent review, see [1]).

Changes in gene expression patterns can cause important phenotypic modifications. Mutations in *cis*-regulatory motifs can alter the binding affinity of transcription factors and affect the expression of a gene. However, the evolutionary dynamics of promoter sequences are still poorly understood. A commonly used approach to assess the existence of evolutionary constraints and identify regulatory motifs is the identification of conserved non-coding sequences across orthologues. This rationale is behind several described 'phylogenetic footprinting' methods to discover functional regulatory sequences [2-4].

Contrary to coding sequences, gene expression regulatory sequences do not have very well defined boundaries. A region spanning approximately 100 base-pairs (bp) upstream of the TSS, known as the basal promoter, plays a fundamental part in the assembly of the transcription initiation complex. Further upstream regulatory sequences are of variable length depending on the particular gene [1]. Nevertheless, a recent study has shown that, at distances longer than 2 Kb from the TSS, the similarity between orthologous promoters drastically drops, indicating that most of the functional elements concentrate in the 2 Kb promoter region [5]. In accordance, about 85% of the known mouse transcription regulatory motifs are located within 2 Kb of the gene promoter region [6] and functional assays have shown that a region spanning -500 to +50 relative to the TSS region is sufficient to drive transcription in cultured cells for most human genes [7].

Promoter sequence comparisons across different species have shed light on the different constraints exhibited by promoters of different types of genes. In particular, it has been observed that the promoters of genes encoding regulatory proteins, such as transcription factors and/or developmental proteins, tend to show remarkably strong sequence conservation [8,9], suggesting that the expression of this class of genes requires a relatively large amount of *cis*-regulatory motifs.

Another important factor that may be related to promoter sequence conservation is the number of tissues in which a gene is expressed. In the adult organism, some genes show high tissue-specificity while others show little or no tissue expression restrictions (ubiquitous expression). The effect of expression breadth on promoter conservation has not been addressed previously. Here we provide evidence that, in mammals, the simple expression patterns exhibited by housekeeping genes - expressed in all or nearly all tissues - are often associated with limited promoter sequence conservation, while tissue expression restrictions are associated with increasingly high promoter conservation. This defines a new important property of mammalian gene promoters.

Results

Divergence of orthologous human and mouse promoter sequences

The promoters of different genes exhibit varying degrees of sequence divergence [8-10]. In genes from nematodes [11] and yeast [12], the level of promoter sequence divergence is positively correlated with the evolutionary rate of the encoded protein. An interesting question is whether such a correspondence also exists in mammals. We collected human and mouse orthologous promoters (6,698 pairs, 2 Kb from the transcription start site) and applied different measures of sequence divergence. We aimed at quantifying promoter sequence divergence, evaluating the strength of selection and identifying any significant relationship between the divergence of promoter and coding sequences.

First, we calculated the fraction of the promoter sequence that failed to align between human and mouse orthologues. We used the local pairwise sequence alignment program described in Castillo-Davis *et al.* [11], which provides a score, d_{SM} (shared motif divergence), that corresponds to the fraction of non-aligned sequence. The average value was 0.701, which means that, on average, 29.9% of the 2 Kb promoter sequence was successfully aligned. On the promoter alignments we estimated the number of nucleotide substitutions per site using PAML [13]. This promoter substitution rate, which we term Kp, was, on average, 0.334 substitutions per site.

Next we estimated the synonymous (Ks) and non-synonymous (Ka) substitution rates of the corresponding gene coding sequences using PAML. In mammals, Ks can be used to account for the background mutation level. Ka, on the contrary, corresponds to changes at the amino acid level and reflects the strength of selection on the protein. In the orthologous dataset, the average Ks was 0.709 and the average Ka 0.084. The approximately two-fold difference between Kp and Ks (0.334 and 0.709, respectively) indicates stronger negative or purifying selection in the evolution of promoter sequences with respect to synonymous sites in coding regions.

We subsequently addressed the question of whether the level of promoter sequence divergence is related to the evolutionary rate in the corresponding coding sequence in mammals. Interestingly, we found a modest although significant positive correlation between the promoter divergence (d_{SM}) and the coding sequence substitution rate (d_{SM} and Ka, r = 0.20, $p < 10^{-58}$; d_{SM} and Ka/Ks, r = 0.14, $p < 10^{-29}$; d_{SM} and Ks, r = 0.18, $p < 10^{-48}$). That is, in general, proteins that showed high divergence between human and mouse (high Ka or Ka/Ks) showed a tendency to be encoded by genes with reduced promoter sequence conservation.

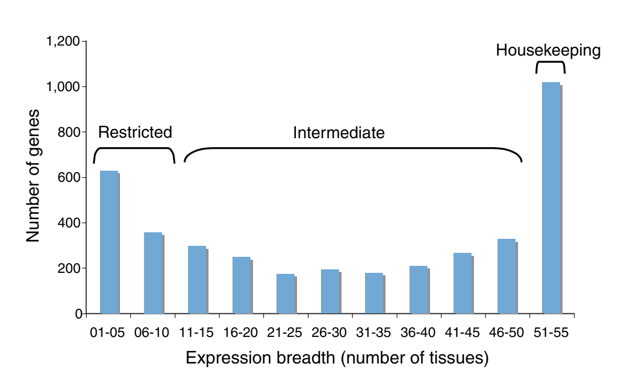


Figure 1
Mouse tissue expression distribution. We define three groups: low expression breadth (Restricted; 1-10 tissues), intermediate expression breadth (Intermediate; 11-50 tissues), high expression breadth (Housekeeping; 51-55 tissues).

Gene expression breadth

We used mouse transcriptome microarray data from Zhang *et al.* [14] to classify the previously defined genes into different groups according to their expression in 55 mouse organs and tissues (see Supplementary table S5 in Additional data file 1). The orthologous dataset with expression data contained 3,893 genes. The tissue distribution profile in five-tissue bins (Figure 1) showed a bimodal shape with a moderate excess of genes expressed in a few tissues and a more acute excess of genes expressed in a very large number of tissues. Genes with expression restricted to 1-10 tissues were classified as 'restricted' (986 genes), those with ubiquitous or nearly ubiquitous expression (51-55 tissues) as 'housekeeping' (HK; 1,018 genes), and the rest, expressed in 11-50 tissues, as 'intermediate' (1,889 genes).

We compared d_{SM}, Kp, Ka and Ks values for genes classified in the three different expression groups (Table 1). We observed that the average d_{SM} score, which corresponds to the fraction of the 2 Kb promoter that cannot be aligned, consistently increased with the expression breadth. The average d_{SM} in HK genes was 0.732 (26.8% promoter conservation), whereas in genes with 'restricted' expression it was 0.688 (31.2% promoter conservation). The d_{SM} values were significantly different between HK genes and the other non-HK groups (Wilcoxon-Mann-Whitney and Kruskal-Wallis tests, $p < 10^{-5}$). The nucleotide substitution rate within aligned regions, Kp, was, instead, not significantly different across the different datasets. Kp also showed decreased variability with respect to Ks, with about three times lower standard deviation values (Table 1). In contrast to promoter divergence, both Ka and Ka/Ks in coding sequences were significantly lower in HK genes than in the other groups (Table 1). In fact, we observed a negative correlation between expression breadth and Ka (r = -0.31, $p < 10^{-87}$), in accordance with previous results [15,16]. Therefore, while at the promoter level the constraints appeared to be weaker in HK genes than in the rest of the genes, at the level of the protein sequence the situation was reversed.

Additional support for the results was obtained using human gene expression data. We mapped the orthologous genes to the eVOC database (anatomical system and cell type) [17], based on expressed sequence tag data, and to Gene Atlas [18]. The results obtained using these datasets were in strong agreement with the results presented in Table 1 (see Supplementary tables S1, S2 and S3, respectively, in Additional data file 1). That is, the fraction of human genes with the broadest tissue expression (HK genes) always showed significantly higher promoter divergence values.

Table 1

Sequence divergence versus tissue expression breadth

No. of tissues	N (total = 3,893)	d_{SM}	Kp	Ka	Ks	Ka/Ks
01-10	986	0.688	0.337	**0.107**	**0.733**	**0.150**
		0.735	0.328	0.084	0.673	0.119
		0.221	0.110	0.093	0.299	0.122
11-50	1,889	0.701	0.333	**0.079**	0.708	0.116
		0.752	0.328	0.058	0.633	0.089
		0.216	0.093	0.073	0.307	0.103
51-55	1,018	**0.732**	0.328	**0.050**	**0.639**	**0.079**
		0.791	0.323	0.031	0.572	0.054
		0.208	0.079	0.057	0.305	0.085
	p value (K-W test)	$<10^{-5}$	0.226	$<10^{-75}$	$<10^{-18}$	$<10^{-62}$

N, number of genes; d_{SM}, promoter divergence (see text); Kp, promoter substitution rate; Ka, non-synonymous substitution rate; Ks, synonymous substitution rate. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Numbers in bold indicate significant differences at $p < 0.001$ in each expression group with respect to the rest (two-sample Wilcoxon-Mann-Whitney test). The last row shows the *p* value of Kruskal-Wallis (K-W) test that evaluates differences between the three tissue expression breadth groups.

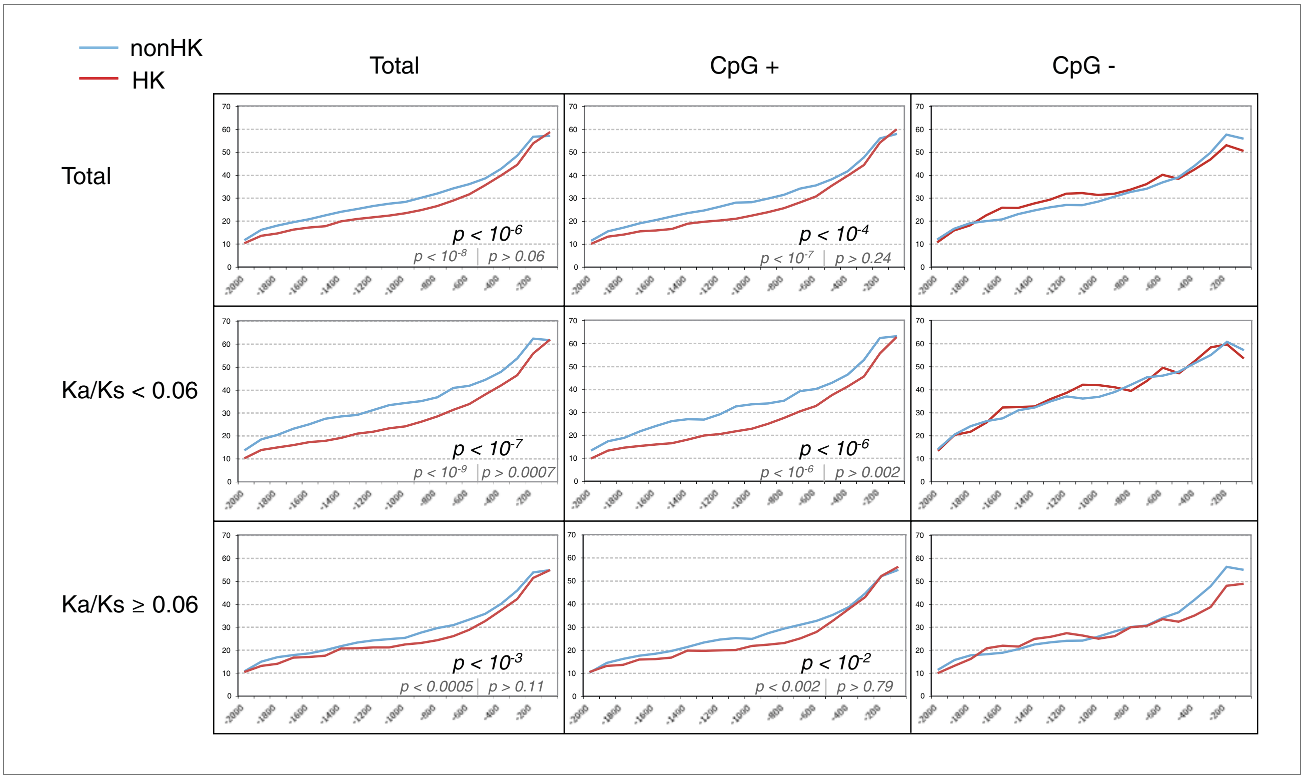


Figure 2

Promoter sequence conservation in HK and non-HK genes. The x-axis shows 100 nucleotide bins along 2 Kb upstream of the TSS. The y-axis shows percent conservation ((1 - d_{SM}) × 100). Genes were grouped according to the presence or absence of a CpG island and Ka/Ks values. Significant *p* values for 2 Kb promoter sequence divergence comparisons are indicated below the curves. Beneath these, the *p* values obtained for regions -2,000 to -500 (left), and -500 to the TSS (right), are given in smaller font size.

The next question we addressed was whether the reduced sequence conservation observed in HK genes was uniformly distributed along the 2 Kb upstream sequence or, alternatively, it could be mapped to a particular region of the promoter. Considering the complete 2 Kb sequences, d_{SM} differences between HK and non-HK datasets were significant at $p < 10^{-6}$ (Wilcoxon-Mann-Whitney test). Then, we calculated the average sequence conservation (1 - d_{SM}) in 100 nucleotide overlapping sequence windows (bins) along the 2 Kb promoter sequence in HK and non-HK genes (Figure 2, top row, left). We found that the region spanning from the TSS to position -100 showed the highest level of sequence conservation (average 1 - d_{SM} 0.576, or 57.6% promoter conservation). Further upstream, the sequence conservation gradually dropped, with a stronger decay in HK than in non-HK genes (Figure 2, top row, left). If we considered only the proximal promoter region, from the TSS to position -500, we did not detect statistically significant differences ($p = 0.0633$). However, using the region from the TSS to -600, differences became significant at $p < 0.05$ ($p = 0.0195$). On the other hand, when we considered the distal promoter region only, from -500 to -2,000, the gap between the two types of sequences regarding promoter divergence increased ($p < 10^{-8}$). Therefore, we concluded that the observed lower promoter sequence conservation of HK genes concentrated in regions upstream from position -500.

Functions of encoded gene products

Our data show that HK genes contained poorly conserved promoters, particularly in the promoter distal part (upstream from -500). Other studies reported differences in the conservation of promoter sequences in relation to the function of the protein [8,9]. As HK genes encode proteins with biased function composition [19,20], we measured the over- and under-representation of different Gene Ontology (GO) terms [21] in the group of HK genes. We also assessed whether the functional biases in HK genes could alone explain the differences observed in promoter sequence conservation.

We determined which GO classes were over- or under-represented among HK genes ($p < 0.01$, χ^2 test), using the 'molecular function', 'biological process', and 'cellular component' classification systems (Supplementary table S4 in Additional data file 1). As expected, an important fraction of the classes statistically over-represented among HK genes showed significantly high promoter sequence divergence. For example, in genes classified as 'structural constituent of ribosome', and 'mitochondrion' the average promoter sequence conservation

Table 2

Average promoter divergence values (d_{SM}) for HK and non-HK genes classified in different GO classes

		All			CpG+			CpG-		
GO term	Description	N	d_{SM} (HK)	d_{SM} (nonHK)	N	d_{SM} (HK)	d_{SM} (nonHK)	N	d_{SM} (HK)	d_{SM} (nonHK)
Molecular function										
GO:0000166	Nucleotide binding	464	0.727	0.699	**363**	**0.732**	**0.698**	101	0.684	0.700
GO:0004872	Receptor activity	259	0.734	0.675	131	0.747	0.656	128	0.655	0.692
GO:0004871	Signal transducer activity	440	0.689	0.658	246	0.692	0.656	194	0.663	0.661
GO:0003700	Transcription factor activity	183	0.673	0.602	113	0.657	0.600	70	0.766	0.605
GO:0043169	Cation binding	485	0.711	0.671	**308**	**0.732**	**0.670**	177	0.582	0.671
Biological process										
GO:0044249	Cellular biosynthesis	**256**	**0.765**	**0.735**	**183**	**0.781**	**0.729**	73	0.629	0.741
GO:0045184	Establishment of protein transport	162	0.720	0.737	138	0.723	0.731	24	0.677	0.760
GO:0007049	Cell cycle	188	0.697	0.706	152	0.703	0.724	36	0.656	0.646
GO:0019538	Protein metabolism	**700**	**0.748**	**0.703**	**523**	**0.755**	**0.698**	177	0.682	0.713
GO:0044260	Cellular macromolecule metabolism	**761**	**0.748**	**0.705**	**560**	**0.755**	**0.700**	201	0.686	0.713
GO:0050874	Organismal physiological process	**292**	**0.795**	**0.681**	**109**	**0.813**	**0.675**	183	0.756	0.685
GO:0009605	Response to external stimulus	209	0.676	0.711	85	0.758	0.699	**124**	**0.538**	**0.718**
GO:0007166	Cell surface receptor linker signal transduction	221	0.683	0.626	113	0.659	0.645	108	0.762	0.609
GO:0048513	Organ development	**214**	**0.677**	**0.566**	**103**	**0.699**	**0.528**	111	0.633	0.598
GO:0009653	Morphogenesis	**262**	**0.679**	**0.584**	**132**	**0.685**	**0.549**	130	0.664	0.615
GO:0009607	Response to biotic stimulus	166	0.761	0.723	**74**	**0.783**	**0.686**	92	0.680	0.745
GO:0007165	Signal transduction	563	0.684	0.656	342	0.687	0.668	221	0.666	0.643
Cellular component										
GO:0005739	Mitochondrion	171	0.785	0.756	148	0.780	0.770	23	0.869	0.707
GO:0005737	Cytoplasm	**773**	**0.756**	**0.719**	**579**	**0.759**	**0.727**	194	0.728	0.707
GO:0005783	Endoplasmic reticulum	**153**	**0.791**	**0.713**	**112**	**0.776**	**0.712**	**41**	**0.881**	**0.713**
GO:0005576	Extracellular region	219	0.653	0.621	77	0.718	0.591	142	0.523	0.635
GO:0005886	Plasma membrane	**373**	**0.720**	**0.661**	**189**	**0.735**	**0.656**	184	0.663	0.666

Entries in bold are those that have a significantly different d_{SM} distribution ($p < 0.05$). The number of genes (N) is indicated for each GO class. Results for CpG+ and CpG- genes are shown.

was only 23% (d_{SM} = 0.77). On the other hand, many classes under-represented among HK genes showed significantly high promoter sequence conservation (low d_{SM}). For example, genes annotated as 'transcription factor activity' or 'nervous system development' showed an average promoter conservation of 42% (d_{SM} = 0.58), and genes annotated as 'cell differentiation' showed an average promoter conservation of 43% (d_{SM} = 0.57).

Given the promoter sequence divergence differences among gene functional classes, one possibility was that the functional class bias in HK genes could fully explain the differences found between HK and non-HK genes. For this reason we tested whether there were any d_{SM} differences between HK and non-HK genes within the same GO class. For statistical robustness we considered only GO classes with a minimum of 150 genes (22 classes; Table 2). In 19 of these classes, the average d_{SM} of HK genes was higher than that of non-HK genes. For example, transcription factors with HK expression had an average d_{SM} of 0.673 (32.7% promoter conservation), while those with no HK expression had an average d_{SM} of 0.602 (39.8% promoter conservation). Of the 19 classes, 9 showed significant d_{SM} differences between HK and non-HK genes ($p < 0.05$). On the other hand, in the three classes with higher average d_{SM} scores in non-HK than in HK genes the differences were not significant ($p > 0.64$). Therefore, we concluded that the promoter sequence divergence differences between HK and non-HK genes were essentially maintained within the different GO classes.

CpG island content and coding sequence evolutionary rate

The promoters of HK genes are rich in CpG islands [22-25]. This could potentially influence the level of conservation of promoter sequences. Therefore, we divided the gene dataset into genes containing CpG islands (CpG+) and genes not containing CpG islands (CpG-), according to the presence or absence of a CpG island in the region -100 to +100 (see Materials and methods), and analyzed the two groups separately. Of the mouse genes, 65% were classified as CpG+ (91% of the human orthologs of these were also CpG+). Among the genes classified as HK, this number went up to 88%. The length of CpG islands was not significantly different in HK and non-HK genes.

Within CpG+ genes, we observed the previously described positive relationship between promoter sequence divergence (d_{SM}) and expression breadth. HK genes (expressed in 51-55 tissues) had an average d_{SM} of 0.739, whereas genes expressed in an intermediate number of tissues (11-50) and those with restricted expression (1-10 tissues) had average d_{SM} scores of 0.708 and 0.679, respectively. These scores are comparable to those obtained previously (Table 1) and the differences between HK and non-HK genes were highly significant ($p < 10^{-4}$; Figure 2, top row, middle). Similar results were obtained with other gene expression datasets (Figures S1, S2 and S3 in Additional data file 2).

In contrast, in CpG- genes the differences between HK and non-HK genes were smaller, and did not reach statistical significance in the mouse gene dataset (Figure 2, top row, right). Indeed, HK genes that did not contain CpG islands (12% of the HK genes) showed average promoter sequence divergence similar to that of non-HK genes (around 0.69). Thus, this minority of HK genes with no CpG islands appeared to have increased sequence evolutionary constraints in relation to the rest of the HK genes.

We also assessed if the presence or absence of CpG islands influenced d_{SM} differences between HK and non-HK genes within the same GO class. In CpG+ genes the differences between HK and non-HK genes were even more marked than in the complete dataset, and three additional GO functions showed statistical differences (Table 2). In CpG- genes, instead, the differences between HK and non-HK genes per GO class were, in almost all cases, not significant.

We had previously described a positive correlation between the non-synonymous substitution rate, Ka (or Ka/Ks), and promoter sequence divergence (d_{SM}). That is, many rapidly evolving coding sequences were associated with poorly conserved promoters. This seemed at first to contradict the finding that HK genes, with typically low Ka values, tended to have highly divergent promoters. To unravel the effect of coding sequence evolutionary rate and expression breadth in promoter sequence evolution, we divided the gene dataset into two groups, genes with Ka/Ks < 0.06, a fraction representing about one-third of the genes and highly enriched in HK genes, and the rest of the genes, with Ka/Ks ≥ 0.06.

The first observation was that, according to the general correlation, genes with more slowly evolving coding sequences (Ka/Ks < 0.06) showed higher promoter conservation than those with Ka/Ks ≥ 0.06 (average d_{SM} of 0.663 and 0.722, respectively). However, this was mostly due to genes that were not HK genes (Figure 2, middle row, left), which explained the apparent contradiction mentioned before. Among genes with Ka/Ks < 0.06, the average d_{SM} was 0.72 for HK genes, but 0.65 for non-HK genes. Not surprisingly, we found that the previously observed correlation between d_{SM} and Ka/Ks was more relevant in non-HK genes (r = 0.17, $p < 10^{-19}$) than in HK genes (r = 0.10, $p < 0.002$).

Cis-regulatory motif content in housekeeping gene promoters

The differences in promoter sequence divergence associated with expression tissue distribution are likely to reflect the presence of different functional regulatory motifs in genes with diverse expression patterns. Among the expression groups previously defined (restricted, intermediate and HK) only the HK gene group probably represents a rather homogeneous class from a gene expression regulatory perspective. Other groups include genes that are active in diverse tissues and that are likely to be regulated by very different factors. We thus investigated whether the promoters of HK genes were enriched in specific transcription factor binding motifs.

In the first place, we mapped all experimentally verified transcription factor binding sites (TFBSs) from TRANSFAC [26] in the human and mouse promoter sequences. We observed that approximately 75% of mapped TFBSs fell into conserved regions, which only occupy approximately 30% of the sequence analyzed. However, as only less than 2% of the genes in the dataset contained known TFBSs, we could not infer any statistically significant biases from these data. For this reason, we decided to use motifs predicted by weight matrices representing known TFBSs. We performed separate analysis with the vertebrate TFBS weight matrix collections available from TRANSFAC and PROMO [27]. We identified nine motifs that were consistently over-represented in the aligned parts of HK gene promoters using the two weight matrix datasets (χ^2 test, $p < 10^{-5}$; Table 3). The motifs were recognized by particular transcription factors or families of transcription factors, according to data in TRANSFAC and PROMO. Among them were commonly found regulators such as Sp1, or members of the ATF (activating transcription factor) family. We also analyzed HK motif over-representation separately in aligned regions located either downstream or

Table 3

Transcription factors with predicted binding motifs over-represented in HK gene promoters

Transcription factor	Description	Expression breadth
AHR and ARNT	Aryl hydrocarbon receptor; it can interact with ARNT (AHR:ARNT heterodimer)	INT
ATF family	Activating transcription factor	HK
CREB family	cAMP responsive element binding protein	INT
E2F family	E2F transcription factor	INT and HK
HIF1A	Hypoxia inducible factor 1, alpha subunit; as AHR, it can interact with ARNT	HK
MYC and MAX	Proto-oncogene protein c-myc and MYC associated factor X; they can form MYC:MAX heterodimers	INT and HK
NRF1 and NRF2	Nuclear respiratory factor 1 and 2	INT and HK
SP1	SP1 transcription factor	HK
USF	Upstream transcription factor (USF1 and USF2)	INT

HK, housekeeping; INT, intermediate.

upstream of position -500. Whereas in the region from the TSS to -500 the nine distinct motifs became even more strongly over-represented than in the 2 Kb promoter, in the more distal promoter region, upstream of -500, four of the motifs - ATF, CREB, NRF1/2 and USF - were no longer significant. We next determined the expression class of the transcription factors that could bind to the nine motif types, using the previously defined three expression groups. Importantly, all transcription factors showed HK or intermediate expression patterns (Table 3), and none showed tissue-restricted expression, which is consistent with a putative role in the regulation of HK genes. Therefore, we could define a group of factors that, mainly through interactions with HK proximal promoter regions, are likely to play important roles in the maintenance of adequate levels of expression of this type of genes.

Discussion

In this work we present the first evidence, at least to our knowledge, of a relationship between promoter sequence divergence and gene expression breadth. We have observed that the promoters of HK genes tend to be less conserved than those of non-HK genes, especially in the distal promoter region, upstream of position -500. Given the strong conservation of HK gene expression patterns across organisms [28], high promoter sequence divergence is likely to reflect weak functional constraints rather than sequence diversification driven by the acquisition of new functionalities. These observations raise the interesting possibility that HK genes have shorter functional promoters. Interestingly, other features of HK genes tend to shortness; in particular, they have been described to have shorter coding, intronic, and intergenic sequences [29-31]. As a consequence, and with the exception of plants [32], transcripts of HK genes tend to be short. One hypothesis put forward to explain this observation is selection for economy in transcription and translation [30,31]. An alternative hypothesis, called 'genome design', is that tissue-specific genes require a greater amount of non-coding DNA due to their more complex regulation [29]. Our results show that HK genes contain more divergent distal promoter sequences than non-HK genes. In line with the 'genome design' hypothesis, this may be due to their relatively simple expression patterns, requiring less regulatory sequences.

In mammals, conservation of a gene's upstream sequence is related to the function of the encoded protein [8,9]. Iwama and Gojobori [9] found that genes encoding transcription factors and developmental proteins showed high gene upstream sequence conservation. Similarly, Lee *et al.* [8] showed that genes involved in complex and adaptative processes, such as development, cell communication, neural function, and signaling, were associated with higher promoter sequence conservation despite their relative recent emergence during evolution. On the contrary, genes involved in basic processes, such as metabolism and ribosomal function, contained poorly conserved promoters. Our study is consistent with these findings, as the former genes are under-represented in HK genes, while the later are over-represented. However, by directly relating promoter conservation to mode of expression, we are able to propose a more direct explanation for the differences in promoter sequence conservation between genes that perform basic housekeeping functions, and which are simply regulated, and genes that are important for tissue- or organ-specific processes, which may require a more complex regulation. In addition, function alone cannot explain the differences across genes, as the reduced promoter sequence conservation in HK genes with respect to non-HK genes is essentially maintained within different functional (GO) classes.

The existence of a positive correlation between the speed of evolution of regulatory sequences and that of coding sequences in orthologous genes is suggestive of a link between rapid diversification of a protein and its expression pattern. We have found that in mammals there is a weak but significant correlation between these two factors, in accordance with previous observations in nematodes [11] and yeast

[12]. Interestingly, we have observed that this relationship is especially relevant for non-HK genes, while in HK genes the effect is practically negligible.

The CpG island gene classification and association with expression breadth observed here is consistent with other reports [22,24]. The majority of mammalian promoters contain CpG islands and HK genes are particularly rich in this type of sequence. Our study shows that promoters that do not contain CpG islands are more strongly conserved than those that do, and even more so if the genes encode slowly evolving proteins. Promoters with no CpG islands correspond to classical TATA-containing promoters and it has been recently shown in a large-scale analysis that they are particularly well-conserved across different mammalian species [33].

We identify nine different motifs, corresponding to known transcription factor binding sites, that are significantly over-represented in HK genes. Most of the transcription factors that bind to these sites are themselves encoded by HK genes and the rest are encoded by genes classified as of intermediate expression breadth. Five of the motifs (binding Sp1, USF, NRF1, CREB, or ATF) show high frequency peaks in the vicinity of the TSS (-200 to -1) in a large collection of human promoters, and the combination of two of them (binding Sp1 and NRF1) is over-represented in HK gene promoters [34]. Some of the motifs identified are bound by known regulators of HK genes; examples are Sp1 and USF for the APEX nuclease gene [35] or Sp1 and HIF-1 for the endoglin gene [36].

Of note, besides HK genes, we also find differences between the groups of genes with restricted expression (1-10 tissues) and intermediate expression (11-50 tissues). 'Restricted' genes tend to show higher promoter conservation than 'intermediate' genes (Table 1; Aupplementary Tables S1, S2, and S3 in Additional data file 1). These results may seem counter-intuitive, as one could argue that genes expressed in only a few tissues should have more simple regulation than genes expressed in an intermediate number of tissues. However, one possibility is that 'restricted' genes contain a larger number of negative regulatory elements. Interestingly, gene reporter assays of promoter activity in ENCODE regions (approximately 1% of the genome) have shown that negative elements appear to be present from 1,000 to 500 nucleotides upstream of the TSS in 55% of genes [37]. This indicates that motifs for inhibitory transcription factors may be present in a substantial fraction of genes. One expects that such regions will be more common in tissue-specific 'restricted' genes, which would be consistent with the observed stronger distal promoter sequence conservation.

It has been observed that metazoan-specific proteins tend to be more tissue-specific than universal eukaryotic proteins [20]. In other words, HK genes are enriched for proteins of ancient origin. Old eukaryotic proteins typically evolve more slowly and are longer than proteins of a more recent origin, probably due to increased functional constraints [38]. However, at the level of gene expression regulatory regions they may be simpler and less constrained than genes that represent innovations in multi-cellular organisms. Cross-species comparisons will be used in future studies to gain further insight into these questions.

Conclusion

We describe that genes with housekeeping expression contain more divergent promoters than genes with a more restricted tissue expression. Importantly, this property cannot be fully explained by the functional class of the encoded gene products, or by a higher prevalence of CpG islands in HK gene promoters. In addition, we have identified a number of transcription factors that are likely to play a predominant role in the control of HK gene expression. We argue that the lower promoter conservation observed in HK genes could be due to a more simple regulation of gene transcription.

Materials and methods

Sequence retrieval and alignment

We identified human and mouse orthologous genes using the Ensembl database (release 34) [39]. We considered only orthology relationships of type UBRH (unique best reciprocal hit): 17,620 records of human genes with orthologous mouse genes (human-mouse dataset) and 12,868 of mouse genes with orthologous human genes (mouse-human dataset). We extracted the promoter sequences from these genes, comprising 2 Kb upstream of the TSS, from the UCSC database (hg17 and mm6 releases) [40], excluding genes with multiple TSSs, discarding duplicates, and considering only gene pairs with human-mouse and mouse-human orthology data that were both available and congruent. The resulting dataset contained 8,972 orthologous promoter sequence pairs. We discarded repeats from alignments using RepeatMasker (release 1.1.65) [41]. We aligned the sequences with the local pairwise sequence alignment program described in Castillo-Davis *et al.* [11], using a minimum alignment length of 16 nucleotides. For each orthologous pair we obtained the promoter sequence divergence score (d_{SM}; shared motif divergence), which is the fraction of the sequence that does not align, taking the average between the human and mouse promoter sequences. The fraction of sequence aligned was then 1 - d_{SM}. We calculated the average 1 - d_{SM} in 100 nucleotide sequence windows overlapping by 20 nucleotides. Failure to align portions of the promoter may be due to very high divergence or the occurrence of insertions/deletions. To obtain an estimate of the d_{SM} random expectation we aligned, with the same program, 1,000,000 pairs of 2 Kb random sequences and calculated their d_{SM} scores. We discarded orthologous pairs with an overall average d_{SM} > 0.97 (random expectation ≥0.01), obtaining 7,330 orthologous promoter sequence pairs. Coding sequences were extracted from the Ensembl database (release 34) and aligned with ClustalW [42].

Substitution rate estimation

Synonymous (Ks) and non-synonymous (Ka) substitution rates were estimated with the codeml program in PAML [13]. From the 7,330 orthologous pairs, 6,698 remained after discarding those with Ka ≥ 0.5, Ks ≥ 2.0, or Kp ≥ 2.0 (saturated pairs). We estimated, for each gene, the number of nucleotide substitutions per site in the concatenated promoter sequence alignment, using the baseml program, with the Hasegawa, Kishino and Yano (1985) model [43], in PAML. This substitution rate was termed Kp.

Gene expression datasets

We used mouse transcriptome microarray data from Zhang *et al.* [14] to classify the previously defined genes into different groups according to their expression in 55 different mouse organs and tissues (see Supplementary table S5 in Additional data file 1). Zhang *et al.* [14] considered genes to be expressed only if their intensity exceeded the 99th percentile of intensities from the negative controls.

In addition, we used human gene expression data from Gene Atlas (GNF1H), based on transcriptome microarray data [18], and human gene expression data from the eVOC database (anatomical system and cell type ontologies, release 2.7), based on expressed sequence tag data [17]. We considered genes to be expressed in a tissue according to Gene Atlas data only if the expression level was ≥200. Gene Atlas covers 79 human organs and tissues (see Supplementary table S5 in Additional data file 1). For eVOC anatomical systems and cell types we discarded classes with a very small number of genes (<1,000) or large classes with high redundancy (>90% of genes shared with other classes). This resulted in 57 anatomical systems and 10 cell types (see Supplementary table S5 in Additional data file 1). HK, intermediate and restricted expression groups were defined following similar criteria as for the mouse transcriptome data.

Complete sequence divergence data for the different expression groups are available in Additional data file 3.

Statistical tests and correlations

Correlations were calculated with the Spearman Rank correlation method. Two-sample Wilcoxon-Mann-Whitney statistical test was used to assess differences between groups unless stated. The R statistical package was used [44].

Gene Ontology functions

GO annotations were extracted from Ensembl (release 34) [39]. We used the GO term definitions of 30 March, 2005 [21]. Over-representation and under-representation of HK genes in different GO classes were verified by chi-square test ($p < 0.01$), using expected values calculated from the percent number of HK genes in the root GO term of each ontology (GO:0003674, molecular function; GO:0008150, biological process; GO:0005575, cellular component). Only GO terms containing a number of genes between 50 and 1,000, both included, were considered. Some GO terms were discarded to reduce redundancies.

Transcription factor binding site predictions

We used weight matrices from PROMO (release 3) [27,45] and TRANSFAC (release 7.0) [26] to predict transcription factor binding sites. Motif searches were carried out with a similarity cut-off of 0.85. We selected motifs consistently predicted by both matrix collections that were over-represented in HK genes versus all the genes taken together using the chi-square test.

CpG islands

We extracted sequences -100 to +100 with respect to the TSS. We classified genes as CpG+ (CpG island-positive near TSS), when the C+G content exceeded 0.55 and the CpG score (observed CpG/expected CpG) exceeded 0.65 in the -100 to +100 region, or as CpG- (CpG island-negative near TSS), otherwise. This classification is similar to that used by Yamashita *et al.* [22], but with more stringent values for CpG+ determination, in line with the CpG island definition proposed by Takai and Jones [46]. To study differences in CpG island sequence conservation between HK and non-HK genes, we extended the CpG islands upstream, such that the G+C content exceeded 0.55 and the CpG score exceeded 0.65, calculating in this manner the 5' end point of CpG islands.

Additional data files

The following additional data are available with the online version of this manuscript. Additional data file 1 contains Supplementary tables S1-S5: Table S1 lists human gene sequence divergence values in expression groups according to Gene Atlas (GNF1H); Table S2 lists human gene sequence divergence values in expression groups according to the eVOC anatomical system classification; Table S3 lists human gene sequence divergence values in expression groups according to the eVOC cell type classification; Table S4 lists GO terms over-represented and under-represented in HK genes with their average d_{SM} values; and Table S5 lists the organs, tissues, and cell types considered in each expression dataset. Additional data file 2 contains figures plotting promoter sequence conservation along 2 Kb upstream of the TSS in HK and non-HK genes considering expression groups according to Gene Atlas GNF1H (Figure S1), the eVOC anatomical system classification (Figure S2), and the eVOC cell type classification (Figure S3). Additional data file 3 contains the complete sequence divergence and expression group data used in this manuscript. Additional data file 4 contains human 2 Kb upstream sequences (human promoters), in fasta format. Additional data file 5 contains mouse 2 Kb upstream sequences (mouse promoters), in fasta format.

Acknowledgements

The authors thank Miriam Subirats and Neus Xivillé, and members of the Computational Genomics group at GRIB/CRG, for helpful comments. We acknowledge support from Instituto Nacional de Bioinformática, Fundación Banco Bilbao Vizcaya Argentaria, Plan Nacional de I+D Ministerio de Educación y Ciencia (BIO2006-07120/BMC), European Comission Infobiomed NoE, and Fundació ICREA.

References

1. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377-1419.
2. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.** *J Mol Biol* 1988, **203:**439-455.
3. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**13.
4. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover.** *Mol Biol Evol* 2002, **19:**1114-1121.
5. Keightley PD, Lercher MJ, Eyre-Walker A: **Evidence for widespread degradation of gene control regions in hominid genomes.** *PLoS Biol* 2005, **3:**e42.
6. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
7. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome Res* 2003, **13:**308-312.
8. Lee S, Kohane I, Kasif S: **Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes.** *BMC Genomics* 2005, **6:**168.
9. Iwama H, Gojobori T: **Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network.** *Proc Natl Acad Sci USA* 2004, **101:**17156-17161.
10. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S: **Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions.** *Genome Res* 2004, **14:**1711-1718.
11. Castillo-Davis CI, Hartl DL, Achaz G: ***cis*-Regulatory and protein evolution in orthologous and duplicate genes.** *Genome Res* 2004, **14:**1530-1536.
12. Chin CS, Chuang JH, Li H: **Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.** *Genome Res* 2005, **15:**205-213.
13. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
14. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, *et al.*: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3:**21.
15. Zhang L, Li WH: **Mammalian housekeeping genes evolve more slowly than tissue-specific genes.** *Mol Biol Evol* 2004, **21:**236-239.
16. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17:**68-74.
17. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, *et al.*: **eVOC: a controlled vocabulary for unifying gene expression data.** *Genome Res* 2003, **13:**1222-1230.
18. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101:**6062-6067.
19. Lehner B, Fraser AG: **Protein domains enriched in mammalian tissue-specific or widely expressed genes.** *Trends Genet* 2004, **20:**468-472.
20. Freilich S, Massingham T, Bhattacharyya S, Ponsting H, Lyons PA, Freeman TC, Thornton JM: **Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins.** *Genome Biol* 2005, **6:**R56.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
22. Yamashita R, Suzuki Y, Sugano S, Nakai K: **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity.** *Gene* 2005, **350:**129-136.
23. Vinogradov AE: **Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth.** *Trends Genet* 2005, **21:**639-643.
24. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ Jr: **Promoter features related to tissue specificity as measured by Shannon entropy.** *Genome Biol* 2005, **6:**R33.
25. Antequera F: **Structure, function and evolution of CpG island promoters.** *Cell Mol Life Sci* 2003, **60:**1647-1658.
26. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, *et al.*: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
27. Farre D, Roset R, Huerta M, Adsuara JE, Rosello L, Alba MM, Messeguer X: **Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN.** *Nucleic Acids Res* 2003, **31:**3651-3653.
28. Yang J, Su AI, Li WH: **Gene expression evolves faster in narrowly than in broadly expressed mammalian genes.** *Mol Biol Evol* 2005, **22:**2113-2118.
29. Vinogradov AE: **"Genome design" model: evidence from conserved intronic sequence in human-mouse comparison.** *Genome Res* 2006, **16:**347-354.
30. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19:**362-365.
31. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes.** *Nat Genet* 2002, **31:**415-418.
32. Ren XY, Vorst O, Fiers MW, Stiekema WJ, Nap JP: **In plants, highly expressed genes are the least compact.** *Trends Genet* 2006, **22:**528-532.
33. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
34. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters.** *Genome Res* 2004, **14:**1562-1574.
35. Ikeda S, Ayabe H, Mori K, Seki Y, Seki S: **Identification of the functional elements in the bidirectional promoter of the mouse O-sialoglycoprotein endopeptidase and APEX nuclease genes.** *Biochem Biophys Res Commun* 2002, **296:**785-791.
36. Sanchez-Elsner T, Botella LM, Velasco B, Langa C, Bernabeu C: **Endoglin expression is regulated by transcriptional cooperation between the hypoxia and transforming growth factor-beta pathways.** *J Biol Chem* 2002, **277:**43799-43808.
37. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16:**1-10.
38. Alba MM, Castresana J: **Inverse relationship between evolutionary rate and age of mammalian genes.** *Mol Biol Evol* 2005, **22:**598-606.
39. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, *et al.*: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-561.
40. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, *et al.*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.
41. **RepeatMasker** [http://www.repeatmasker.org/]
42. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
43. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22:**160-174.
44. **R Project** [http://www.r-project.org/]
45. Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM: **PROMO: detection of known transcription regulatory elements using species-tailored searches.** *Bioinformatics* 2002, **18:**333-334.
46. Takai D, Jones PA: **Comprehensive analysis of CpG islands in human chromosomes 21 and 22.** *Proc Natl Acad Sci USA* 2002, **99:**3740-3745.

Additional data file 1

Number of tissues	N *(total=5838)*	d_{SM}	Kp	Ka	Ks	Ka/Ks
01-26	1413	0.687	0.336	**0.096**	**0.745**	**0.135**
		0.729	0.327	0.071	0.676	0.103
		0.217	0.104	0.087	0.317	0.116
27-77	3010	0.695	0.334	0.079	0.699	0.114
		0.749	0.326	0.054	0.630	0.081
		0.223	0.098	0.079	0.302	0.108
78-79	1415	**0.722**	0.329	**0.065**	**0.662**	**0.101**
		0.777	0.327	0.045	0.596	0.074
		0.208	0.088	0.067	0.307	0.100
	p-value (K-W test)	<10^{-5}	0.583	<10^{-27}	<10^{-14}	<10^{-17}

Table S1. Sequence divergence values versus Gene Atlas (GNF1H) expression breadth. N: number of genes; d_{SM}: promoter divergence (see text); Kp: promoter substitution rate; Ka: non-synonymous substitution rate; Ks: synonymous substitution rate. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Numbers in bold indicate significant differences at $p < 0.001$ in each expression group with respect to the rest (two-sample Wilcoxon-Mann-Whitney test). Last row shows the p-value of Kruskal-Wallis test that evaluates differences between the three tissue breadth expression groups.

Number of anat. systems	N *(total=6660)*	d_{SM}	Kp	Ka	Ks	Ka/Ks
01-21	1591	**0.677**	0.340	**0.118**	**0.789**	**0.158**
		0.730	0.326	0.093	0.712	0.124
		0.229	0.133	0.098	0.329	0.135
22-41	3295	**0.707**	0.333	**0.084**	**0.714**	**0.122**
		0.763	0.328	0.063	0.638	0.092
		0.215	0.091	0.077	0.307	0.106
42-57	1774	**0.711**	0.330	**0.051**	**0.625**	**0.082**
		0.766	0.326	0.033	0.565	0.059
		0.214	0.084	0.057	0.287	0.084
	p-value (K-W test)	<10^{-5}	0.658	<10^{-142}	<10^{-64}	<10^{-95}

Table S2. Sequence divergence values versus eVOC anatomical system expression breadth. N: number of genes; d_{SM}: promoter divergence (see text); Kp: promoter substitution rate; Ka: non-synonymous substitution rate; Ks: synonymous substitution rate. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Numbers in bold indicate significant differences at $p < 0.05$ in each expression group with respect to the rest (two-sample Wilcoxon-Mann-Whitney test). Last row shows the p-value of Kruskal-Wallis test that evaluates differences between the three tissue breadth expression groups.

Number of cell types	N *(total=6195)*	d_{SM}	Kp	Ka	Ks	Ka/Ks
01-02	1339	**0.676**	0.339	**0.109**	**0.765**	**0.152**
		0.731	0.326	0.083	0.691	0.112
		0.230	0.123	0.096	0.322	0.137
03-07	3637	0.705	0.331	0.080	0.700	0.117
		0.762	0.327	0.057	0.628	0.088
		0.216	0.093	0.075	0.302	0.104
08-10	1219	**0.726**	0.331	**0.054**	**0.625**	**0.085**
		0.784	0.326	0.033	0.568	0.060
		0.205	0.084	0.063	0.290	0.085
	p-value (K-W test)	$<10^{-6}$	0.570	$<10^{-77}$	$<10^{-36}$	$<10^{-50}$

Table S3. Sequence divergence values versus eVOC cell type expression breadth. N: number of genes; d_{SM}: promoter divergence (see text); Kp: promoter substitution rate; Ka: non-synonymous substitution rate; Ks: synonymous substitution rate. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Numbers in bold indicate significant differences at $p < 0.0001$ in each expression group with respect to the rest (two-sample Wilcoxon-Mann-Whitney test). Last row shows the p-value of Kruskal-Wallis test that evaluates differences between the three tissue breadth expression groups.

GO TERMS OVER-REPRESENTED IN HOUSEKEEPING GENES		N	% HK	Av. d_{SM}
MOLECULAR FUNCTION				
GO:0003735	Structural constituent of ribosome	54	66.67	**0.77**
GO:0003723	RNA binding	138	57.97	**0.74**
GO:0016874	Ligase activity	124	40.32	0.74
GO:0016817	Hydrolase activity, acting on acid anhydrides	146	37.67	0.72
GO:0000166	Nucleotide binding	464	32.97	0.71
BIOLOGICAL PROCESS				
GO:0043037	Translation	55	74.55	0.74
GO:0006412	Protein biosynthesis	141	59.57	**0.72**
GO:0009117	Nucleotide metabolism	54	57.41	0.71
GO:0006886	Intracellular protein transport	117	52.14	0.72
GO:0051186	Cofactor metabolism	51	50.98	**0.77**
GO:0044249	Cellular biosynthesis	256	49.61	**0.73**
GO:0045184	Establishment of protein localization	162	48.15	0.72
GO:0016070	RNA metabolism	104	47.12	**0.75**
GO:0006512	Ubiquitin cycle	124	44.35	0.72
GO:0006457	Protein folding	59	44.07	0.74
GO:0051243	Negative regulation of cellular physiological process	131	38.17	0.69
GO:0006091	Generation of precursor metabolites and energy	142	38.03	**0.74**
GO:0007049	Cell cycle	188	36.17	0.69
GO:0019538	Protein metabolism	700	36.00	**0.71**
GO:0044260	Cellular macromolecule metabolism	761	35.74	**0.71**
CELLULAR COMPONENT				
GO:0030529	Ribonucleoprotein complex	88	69.32	**0.74**
GO:0005829	Cytosol	78	56.41	0.73
GO:0005739	Mitochondrion	171	54.97	**0.77**
GO:0031090	Organelle membrane	117	52.14	**0.75**
GO:0012505	Endomembrane system	57	43.86	0.74
GO:0005737	Cytoplasm	773	42.95	**0.73**
GO:0005783	Endoplasmic reticulum	153	41.18	**0.73**

GO TERMS UNDER-REPRESENTED IN HOUSEKEEPING GENES		N	% HK	Av. d_{SM}
MOLECULAR FUNCTION				
GO:0004872	Receptor activity	259	8.49	0.69
GO:0015267	Channel or pore class transporter activity	73	12.33	0.66
GO:0004871	Signal transducer activity	440	14.32	**0.66**
GO:0003700	Transcription factor activity	183	18.58	**0.58**
GO:0043169	Cation binding	485	20.21	**0.68**
BIOLOGICAL PROCESS				
GO:0007267	Cell-cell signaling	96	7.29	**0.63**
GO:0030001	Metal ion transport	79	7.59	0.65
GO:0030154	Cell differentiation	78	7.69	**0.57**
GO:0007155	Cell adhesion	141	9.93	**0.64**
GO:0050874	Organismal physiological process	292	10.62	0.69
GO:0009605	Response to external stimulus	209	11.48	0.71
GO:0007166	Cell surface receptor linker signal transduction	221	11.76	**0.64**
GO:0007600	Sensory perception	66	12.12	0.69
GO:0048513	Organ development	214	13.08	**0.59**
GO:0007399	Nervous system development	90	13.33	**0.58**
GO:0009653	Morphogenesis	262	13.36	**0.60**
GO:0009607	Response to biotic stimulus	166	16.87	0.71
GO:0007165	Signal transduction	563	19.89	**0.66**
CELLULAR COMPONENT				
GO:0005576	Extracellular region	219	9.59	**0.64**
GO:0005886	Plasma membrane	373	15.55	**0.67**

Table S4. Gene Ontology (GO) terms over-represented and under-represented in housekeeping genes. N: number of human genes represented in our dataset (N>=50). % HK: percentage of housekeeping genes. d_{SM}: promoter divergence mean. Over-representation and under-representation were verified by χ^2 test ($p<0.01$). GO terms with average d_{SM} values in bold showed a significantly biased d_{SM} distribution ($p<0.01$).

Table S5. Classes (organs, tissues, cell types) of the expression datasets used to estimate expression breadth.

Zhang (organs and tissues)

- Adrenal
- Aorta
- Bladder
- Bone Marrow
- Brain
- Brown fat
- Calvaria
- Cerebellum
- Colon
- Cortex
- Digit
- E10.5 Head
- E14.5 Head
- ES
- Embryo
- Embryo 12.5
- Embryo 9.5
- Epididymus
- Eye
- Femur
- Heart
- Hindbrain
- Kidney
- Knee
- Large intestine
- Liver
- Lung
- Lymph node
- Mammary gland
- Mandible
- Midbrain
- Olfactory bulb
- Ovary
- Pancreas
- Placenta 12.5
- Placenta 9.5
- Prostate
- Salivary
- Skeletal Muscle
- Skin
- Small intestine
- Snout
- Spinal cord
- Spleen
- Stomach
- Striatum
- Teeth
- Testis
- Thymus
- Thyroid
- Tongue
- Tongue surface
- Trachea
- Trigeminus
- Uterus

Gene Atlas GNF1H

- ColorectalAdenocarcinoma
- WHOLEBLOOD
- BM-CD33+Myeloid
- PB-CD14+Monocytes
- PB-BDCA4+Dentritic_Cells
- PB-CD56+NKCells
- PB-CD4+Tcells
- PB-CD8+Tcells
- PB-CD19+Bcells
- BM-CD105+Endothelial
- BM-CD34+
- leukemialymphoblastic(molt4)
- 721_B_lymphoblasts
- lymphomaburkittsRaji
- leukemiapromyelocytic(hl60)
- lymphomaburkittsDaudi
- leukemiachronicmyelogenous(k562)
- thymus
- Tonsil
- lymphnode
- fetalliver
- BM-CD71+EarlyErythroid
- bonemarrow
- TemporalLobe
- globuspallidus
- CerebellumPeduncles
- cerebellum
- caudatenucleus
- WholeBrain
- ParietalLobe
- MedullaOblongata
- Amygdala
- PrefrontalCortex
- OccipitalLobe
- Hypothalamus
- Thalamus
- subthalamicnucleus
- CingulateCortex
- Pons
- spinalcord
- fetalbrain
- adrenalgland
- Lung
- Heart
- Liver
- kidney
- Prostate
- Uterus
- Thyroid
- fetalThyroid
- fetallung
- PLACENTA
- CardiacMyocytes
- SmoothMuscle
- bronchialepithelialcells
- ADIPOCYTE
- Pancreas
- PancreaticIslets
- testis
- TestisLeydigCell
- TestisGermCell
- TestisInterstitial
- TestisSeminiferousTubule
- salivarygland
- trachea
- AdrenalCortex
- Ovary
- Appendix
- skin
- ciliaryganglion
- TrigeminalGanglion
- atrioventricularnode
- DRG
- SuperiorCervicalGanglion
- SkeletalMuscle
- UterusCorpus
- TONGUE
- OlfactoryBulb
- Pituitary

eVOC (anatomical systems)

- bone
- cartilage
- skeletal muscle
- smooth muscle
- ganglion
- sympathetic chain
- head and neck
- heart
- liver
- bile duct
- kidney
- bladder

dermal system
brain
cerebrum
cerebral cortex
basal nuclei
hypothalamus
medulla oblongata
cerebellum
choroid
retina
optic nerve
cornea
lens
auditory apparatus
peripheral nerve

artery
larynx
lung
bone marrow
blood
tonsil
spleen
tongue
salivary gland
pancreas
pharynx
oesophagus
stomach
small intestine
large intestine

testis
prostate
ovary
breast
umbilical cord
blastocyst
uterus
placenta
endocrine pancreas
pineal gland
pituitary gland
thyroid
parathyroid
adrenal gland
thymus

<u>eVOC (cell types)</u>

squamous cell
transitional cell
lymphoblast
B-lymphocyte

T-lymphocyte
natural killer cell
macrophage
adipocyte

smooth muscle cell
skeletal muscle cell

Additional data file 2

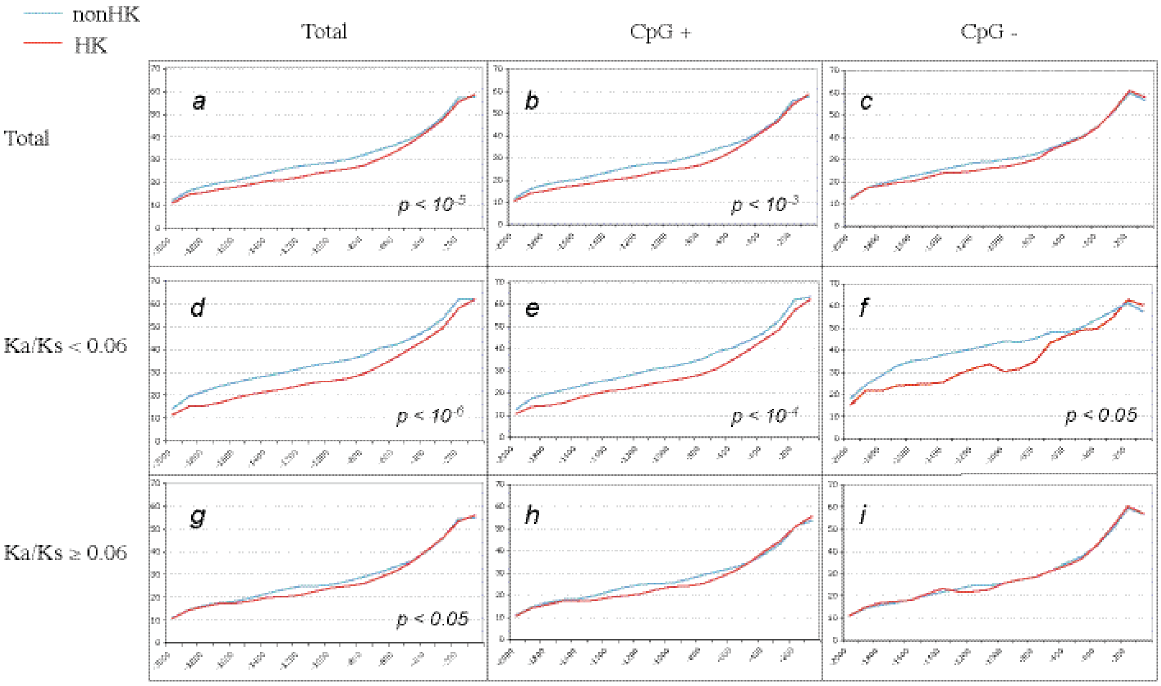


Figure S1. Promoter sequence conservation in housekeeping (HK) and non-housekeeping (non-HK) genes. Expression breadth estimated from Gene Atlas (GNF1H) data. The X-axis shows 100 nucleotide bins along 2Kb upstream of the TSS. The Y-axis shows percent conservation ((1 - d_{SM}) · 100). Genes were grouped according to the presence or absence of a CpG island and Ka/Ks values. Significant p-values (p<0.05) are indicated.

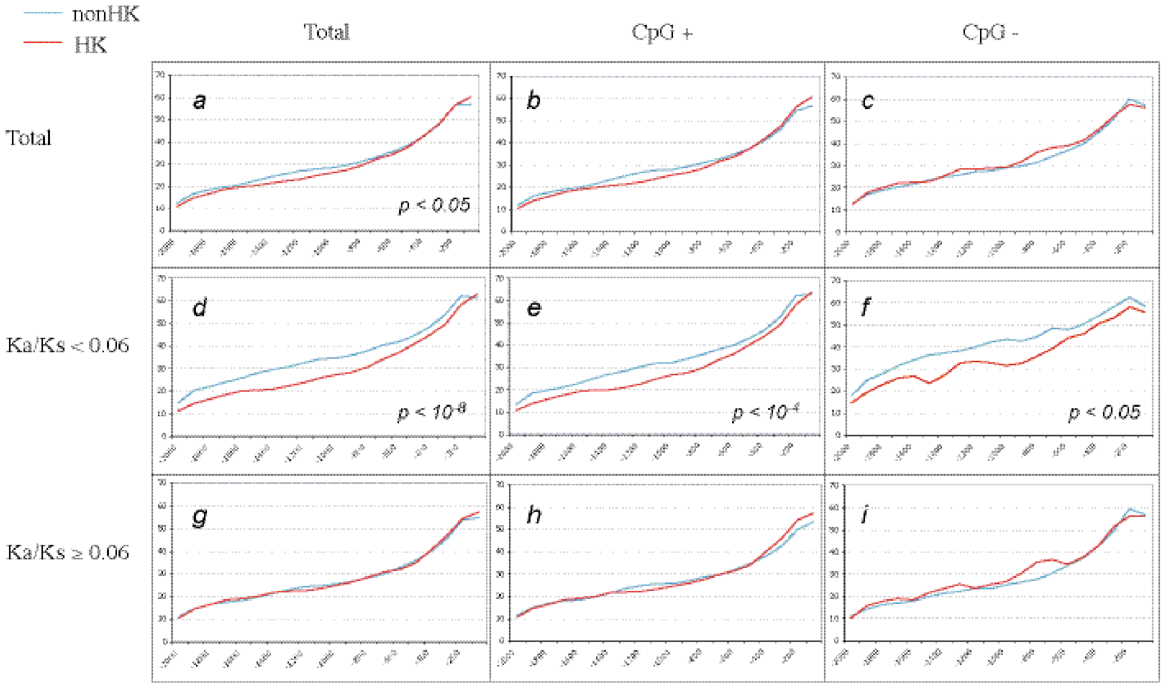


Figure S2. Promoter sequence conservation in housekeeping (HK) and non-housekeeping (non-HK) genes. Expression breadth estimated from eVOC anatomical systems data. The X-axis shows 100 nucleotide bins along 2Kb upstream of the TSS. The Y-axis shows percent conservation $((1 - d_{SM}) \cdot 100)$. Genes were grouped according to the presence or absence of a CpG island and Ka/Ks values. Significant p-values (p<0.05) are indicated.

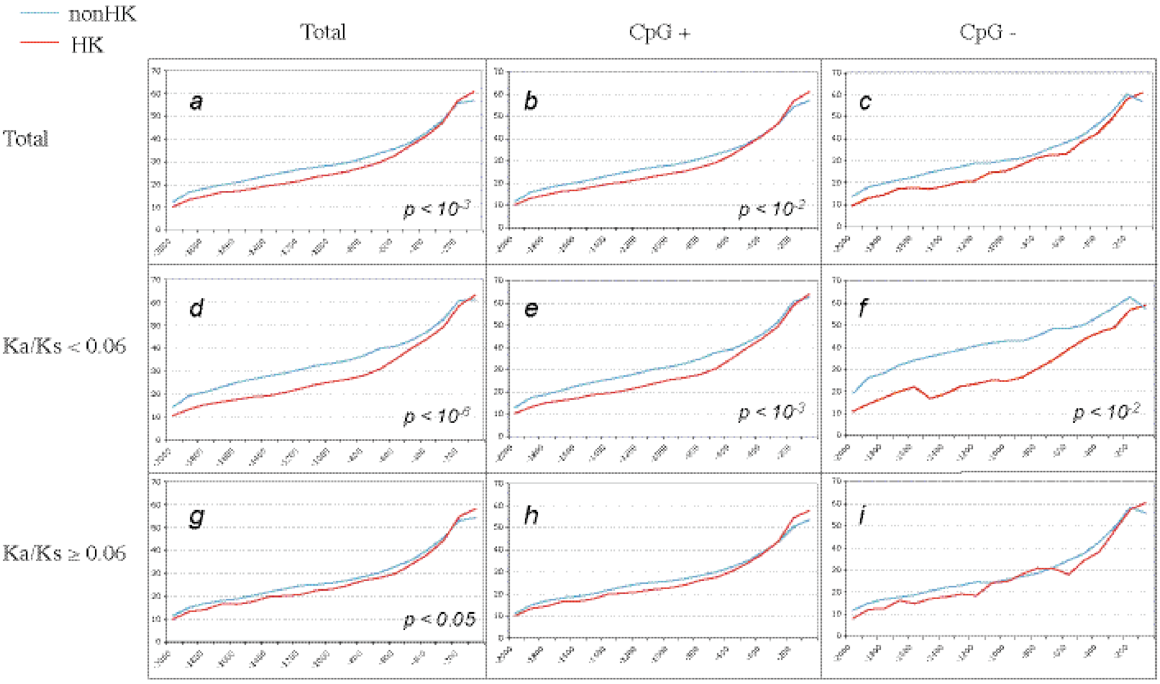


Figure S3. Promoter sequence conservation in housekeeping (HK) and non-housekeeping (non-HK) genes. Expression breadth estimated from eVOC cell types data. The X-axis shows 100 nucleotide bins along 2Kb upstream of the TSS. The Y-axis shows percent conservation ((1 - d_{SM}) · 100). Genes were grouped according to the presence or absence of a CpG island and Ka/Ks values. Significant p-values (p<0.05) are indicated.

Capítol 3

Evolució de les seqüències promotores de gens duplicats

Resum

Aquest capítol presenta un estudi sobre l'efecte de la duplicació gènica sobre l'evolució de les seqüències reguladores de la transcripció. La duplicació gènica està associada a la diversificació funcional i la innovació evolutiva dels gens. S'han fet molts estudis sobre l'efecte de la duplicació gènica sobre la divergència de les seqüències proteiques, però poc es coneix de com aquest procés afecta l'evolució de les seqüències reguladores de l'expressió gènica.

En aquest treball hem quantificat la divergència de la seqüència del promotor, així com la divergència de la seqüència codificant i la divergència de l'expressió tissular, en gens duplicats recentment en mamífers. Hem estudiat arbres de famílies gèniques que contenen gens paràlegs humans i/o de ratolí originats després de la divisió entre els primats i els rosegadors. Això ens ha permès valorar directament l'efecte del nombre de duplicacions gèniques sobre el nivell de divergència respecte als gens ortòlegs. Hem trobat que el nombre de duplicacions gèniques mostra una relació positiva amb la divergència del promotor, la taxa de substitucions no-sinònimes de la seqüència codificant i la divergència de l'expressió tissular. Per tant, la duplicació gènica no només condueix a un increment de les mutacions aminoacídiques, sinó a una acceleració de la divergència de seqüències involucrades en la regulació de l'expressió gènica. A més a més, les còpies gèniques que han experimentat un major nombre d'episodis de duplicacions gèniques tendeixen a expressar-se en un nombre menor de teixits. Aquest resultat dona suport a la pèrdua parcial de funció gènica en els gens duplicats (subfuncionalització).

Divergence of gene expression regulatory sequences after gene duplication

Domènec Farré[1], M.Mar Albà[2,3,4]

[1]Centre for Genomic Regulation, Barcelona, Spain. [2]Fundació Institut Municipal d'Investigació Mèdica, Barcelona, Spain. [3]Universitat Pompeu Fabra, Barcelona, Spain. [4]Catalan Institution for Research and Advanced Studies, Barcelona, Spain

ABSTRACT

Gene duplication is associated to gene functional diversification and evolutionary innovation. However, while many studies have addressed the effect of gene duplication on protein sequence divergence, little is known about how this process affects the evolution of gene expression regulatory sequences. Here we have quantified promoter sequence divergence, as well as coding sequence and tissue expression divergence, in recent mammalian duplicate genes. We have used gene family trees containing human and/or mouse paralogous genes originated after the Primate-Rodent split. This has allowed us to directly assess the effect of the number of gene duplications on the level of divergence with respect to the orthologous genes. We have found that the number of gene duplications shows a positive relationship with promoter divergence, coding non-synonymous substitution rate and expression tissue divergence. Therefore, gene duplication does not only lead to an increase in amino acid-altering mutations, but to an acceleration of the divergence of sequences involved in the regulation of gene expression. In addition, gene copies that have experienced a greater number of gene duplication events tend to be expressed in a smaller number of tissues. This supports partial loss of function of duplicate genes (subfunctionalization).

INTRODUCTION

Gene duplication is a major source of evolutionary innovation (Ohno, 1970). Gene duplicates can be generated by different mechanisms [1]. They may originate as a result of large-scale events that duplicate chromosomal regions or complete genomes (polyploidization). They can also originate by local events involving unequal crossover during meiosis, which generates tandem sequence duplications. Duplicate genes can also arise from retrotransposition [2]. It has been estimated that at least 50% of the prokaryotic genes, and 90% of eukaryotic genes, are the result of gene duplications [3-5].

After a gene duplication event, the resulting gene copies can have different fates. It is believed that in most cases one of the redundant copies becomes silenced by the accumulation of degenerative mutations within, at most, a few million years [6]. In other cases, however, the two copies survive. Retention of both copies may be due to the fixation of beneficial mutations in one of the copies, which acquires a novel function (neofunctionalization), while the second copy maintains the original function [7]. But beneficial mutations are rare events, and this mechanism does not seem to be sufficient to explain the high rate of gene copy survival observed after genome duplications, on the order of 20-50% [8]. A second mechanism that seems more compatible with frequent retention of duplicate genes is subfunctionalization, which is based on the acquisition of complementary degenerative mutations [8-10]. As a result, each of the two copies may specialize in a subset of functions from the ancestral protein [10]. The mutations may also affect different regulatory elements, leading to expression pattern partitioning [9].

It has been observed that genes that have experienced duplications tend to show increased protein evolutionary rates, reflecting relaxation of functional constraints and/or adaptive evolution [6,11-16]. Besides, genes that have been formed by retrotransposition evolve faster than their paralogs, which may be related to the sudden relocation in a new genomic environment and the acquisition of new regulatory sequences [13]. In general, gene duplication is associated to significant changes in gene expression patterns. Human and mouse orthologous genes that have experienced gene duplications tend to show higher tissue expression divergence than 1:1 orthologues [17]. Expression divergence between duplicate genes is rapid and, at least in its initial stages, it shows a positive relationship with protein sequence divergence [18-20]. Besides, the degree of gene expression divergence in gene duplicates depends on the function of the gene [18,20].

Expression pattern shifts in duplicate genes are likely to be associated to changes in gene expression regulatory sequences. Many of these sequences map to the gene upstream region, the region known as the promoter. Promoters are highly evolvable sequences [21].

They often exhibit a modular structure, with different regions being involved in the regulation of expression in different tissues or conditions. These regions can evolve independently, providing flexibility to the evolution of gene expression. Single nucleotide substitutions can potentially cause the gain or loss of transcription factor regulatory interactions. Indeed, a high turnover of functional regulatory motifs is observed in human and mouse orthologues [22,23]. Truncation of regulatory sequences is likely to be important in events related to gene duplication, creating asymmetries between the two copies [13]. Indeed, it has been reported that gene duplicates tend to show accelerated promoter evolution with respect to orthologues in *C.elegans* and *C.briggsae* [14]. But how these changes may relate to shifts in gene expression remains to be investigated.

In this study we wished to address several fundamental questions: 1. Is promoter sequence divergence related to expression pattern divergence in duplicate gene pairs? 2. Is there an accumulative effect in sequence or expression divergence related to the number of duplication events undergone by a gene? 3. Which types of regulatory modifications are observed after gene duplication? To be able to address these issues we chose a dataset of human and mouse genes that had duplicated after the Rodent-Primate split. Restricting the analysis to recent duplicates has the advantage that the effects of gene duplication should be easier to detect than for more distant pairs. We built gene family trees and counted the number of gene duplication events separating each gene in a lineage from the orthologue/s in the other lineage. This framework allowed us to directly investigate the effect of the number of gene duplications on sequence and expression divergence, as we considered a fixed interval of time, from the speciation event separating the two lineages to the present time. The results strongly indicate the existence of a link between promoter sequence divergence and expression pattern divergence. Besides, gene duplication is associated to a decrease in the number of tissues in which a gene is expressed, supporting partial loss of function of duplicate genes.

RESULTS

Sequence divergence of orthologous and paralogous genes

We collected pairs of human and mouse orthologues, as well as pairs of human and mouse paralogues originated after the Primate-Rodent split . Such recent duplicates were enriched in particular functions (Sup. Table 1). In particular, functionally annotated mouse paralogues contained a relatively large fraction of olfactory receptors (25%) and receptors in general (34%). These functions were also significant, although less abundant, in the human paralogues (8% and 11%, respectively). Immune response proteins and extracellular protein were also significantly over-represented among functionally annotated human paralogues (7% and 9% of genes, respectively).

We measured the pairwise divergence of promoter sequences (2 Kb from the transcription start site) as well as coding sequence synonymous and non-synonymous substitution rates (Ks and Ka, respectively), in paralogous and orthologous gene pairs. Promoter divergence was quantified using the local pairwise sequence alignment program described in Castillo-Davis et al. [14], which provides a score, d_{SM} (shared motif divergence), which corresponds to the fraction of non-aligned sequence. A local alignment approach seems the most appropriate way to measure promoter divergence given the discontinuous manner of sequence conservation of promoters [24]. In coding sequences, Ka and Ks were measured with PAML [25].

In orthologues, the d_{SM} average value was 0.752 (Table 1), which means that, on average, 24.8% of the 2 Kb promoter sequence was successfully aligned. In mouse duplicates, the average d_{SM} was 0.767, and in human duplicates 0.696, quite similar to the orthologous pair values. Average Ks in paralogous pairs was about half the average Ks in orthologous pairs (Table 1), as expected given that paralogues had originated after the split of the two lineages. Bearing this in mind, the similar d_{SM} in paralogous and orthologus pairs indicates an acceleration of promoter evolution in paralogues with respect to orthologues.

Gene pairs	N1	d_{SM}	Ka	Ks	Ka/Ks	N2	dT	dEK
Human-mouse orthologs	12,726	0.752	0.110	0.745	0.142	8,389	0.494	0.439
		0.815	0.068	0.666	0.099		0.522	0.439
		0.229	0.135	0.340	0.144		0.366	0.079
Primate duplicates	958	0.696	0.173	0.292	0.690	639	0.380	0.277
		0.785	0.132	0.223	0.579		0.185	0.347
		0.276	0.169	0.277	0.479		0.398	0.196
Rodent duplicates	1,719	0.767	0.277	0.539	0.661	683	0.656	0.363
		0.891	0.217	0.370	0.551		0.800	0.383
		0.270	0.230	0.467	0.531		0.354	0.138

Table 1. Sequence divergence of orthologues and paralogues. N1: number of gene pairs for sequence divergence analysis; N2: number of gene pairs with expression data. d_{SM}: promoter divergence; Ka: non-synonymous substitution rate; Ks: synonymous substitution rate; dT: divergence in tissues with significant expression; dEK: divergence in relative expression profiles. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Differences between orthologues and each set of paralogues are significant for each measure (p-value < 10^{-5}, two-sample Wilcoxon-Mann-Whitney test).

A strong difference was observed for non-synonymous to synonymous substitution ratio in orthologues and paralogues (Table 1). Average Ka/Ks was 0.690 for human paralogues and 0.661 for mouse paralogues, in contrast to the much lower 0.142 for orthologues. These results were highly consistent to those previously obtained by Huminiecki and Wolfe in a similar dataset [17]. As comparison of Ka/Ks ratios may be more meaningful if Ks values are similar [16], we divided the dataset in groups of similar Ks values and found that, for each group, the mean Ka/Ks of paralogous pairs was significantly higher than the mean Ka/Ks of orthologous pairs (data not shown).

We then asked whether there was any correlation between promoter divergence (d_{SM}) and protein divergence (Ka). In orthologues, we found a positive correlation (r = 0.266) between d_{SM} and Ka (Sup. Table 2), which was maintained in multiple regression analysis (coefficient = 0.379, $p < 10^{-5}$, d_{SM} vs. Ka using multiple regression formula: d_{SM} ~ Ka + Ks), discarding the possibility that the correlation between d_{SM} and Ka was a consequence of their correlation with Ks alone. In contrast, in paralogues we observed that the correlation between promoter evolution (d_{SM}) and protein evolution (Ka) was a result of their correlation with Ks alone (human genes: coefficient = -0.021, p = 0.847; mouse genes: coefficient = 0.085, p = 0.107; d_{SM} vs. Ka using multiple regression formula: d_{SM} ~ Ka + Ks). So, whereas there was a significant coupling between protein and promoter sequence evolution in orthologues, this was not the case for paralogues. These results are similar to those obtained in a study of *C.elegans* and *C. briggsae* [14].

Expression divergence of orthologous and paralogous genes

We used gene expression data from Gene Atlas (GNF1H, for human, and GNF1M, for mouse) to estimate expression divergence of orthologous and paralogous gene pairs (Table 1). We selected 29 tissues/organs that appeared in both human and mouse microarray datasets (see Methods). Two measures of expression divergence were calculated: dT, distance based on the presence or absence of tissue expression, and dEK, distance based on relative expression pattern profiles (see Methods).

We then examined the possible relationship between expression and sequence divergence in paralogues and orthologues. We found significant coefficients that correlated promoter divergence (d_{SM}) with expression divergence (dT and dEK), as well as coding sequence divergence (Ka) with expression divergence (dT and dEK), in the paralogous gene pairs (Table 2). However, there was no correlation between promoter divergence and expression divergence in the orthologous gene pairs. One possibility is that these correlations were due to the separate correlation of the different variables with Ks. However, multiple regression confirmed the significant correlation between Ka and dT, and Ka and dEK, in human paralogues (coefficients 0.476 and 0.205, respectively, $p < 0.005$). We also found significant correlation between d_{SM} and dT in human paralogues (coefficient 0.151, $p<0.003$), and between d_{SM} and dEK in mouse paralogues (coefficient 0.118, $p<10^{-5}$). The correlation between dT and Ka observed in orthologues (Table 2) was also supported by multiple regression (coefficient = 0.738, $p < 10^{-5}$). Our results indicate a coupling between promoter sequence divergence and expression divergence in paralogues but not in orthologues. They also point to a relationship between coding sequence divergence and expression divergence in both paralogues and orthologues.

		promoter sequence and expression divergence		non-syn coding seq. and expression divergence		syn. coding seq. and expression divergence	
Gene Pairs	N	r(d_{SM},dT)	r(d_{SM},dEK)	r(Ka,dT)	r(Ka,dEK)	r(Ks,dT)	r(Ks,dEK)
Human-mouse orthologs	8,389	0.044	0.080	**0.245**	0.081	0.138	0.008
Primate duplicates	639	**0.204**	0.187	**0.481**	**0.533**	**0.502**	**0.600**
Rodent duplicates	683	0.111	**0.228**	0.150	0.188	0.148	**0.223**

Table 2. Correlation between expression divergence and sequence divergence measures. N: number of gene pairs; d_{SM}: promoter divergence; Ka: non-synonymous substitution rate; Ks: synonymous substitution rate; dT: divergence in tissues with significant expression; dEK: divergence in relative expression profiles. r: Spearman rank correlation coefficient. Coefficients above 0.2 are in bold. All pairwise correlations were significant at $p < 10^{-4}$ except r(d_{SM},dT) and r(Ks,dT) in Rodent duplicates, and r(Ks,dEK) in human and mouse orthologues.

Divergence in paralogues with respect to their corresponding orthologues

Some studies reported differences in the conservation of promoter sequences between orthologous genes in relation to the function of the protein [26,27] or the expression breadth of the gene [28]. To account for such family-specific differences, we decided to analyze divergence of each paralogous pair in relation to the divergence between orthologues in the same gene family.

We built gene families that contained several gene duplicates in the Primate lineage, the Rodent lineage, or both. Next, for each gene family, we calculated the minimum d_{SM} between all possible pairs of orthologues, obtaining a measure of promoter conservation between the "closest" orthologues in each family. Then, for each paralogous pair, we calculated d_{SM}' as the d_{SM} of the paralogous pair divided by the minimum d_{SM} in the family. Additionally, for each paralogous pair, we calculated Ka', Ks', dT', and dEK' dividing its Ka, Ks, dT, and dEK values by the Ka, Ks, dT, and dEK in relation to the orthologous pair that showed the minimum d_{SM}. Subsequently, we separated paralogous gene pairs in two groups, those with $d_{SM}' \leq 1$ (CON gene pairs) and those with $d_{SM}' > 1$ (DIV gene pairs: their promoter sequences are more different to each that between orthologues).

Pairs classified as DIV made a large fraction of the pairs, 30% among human duplicates and 41% among mouse duplicates. In general, DIV gene pairs had higher Ka', Ks', dT', and dEK' values than CON gene pairs (Sup. Table 3, Sup. Figure 1). In both CON and DIV groups, average Ks' was lower than 1, denoting the more recent age of paralogous pairs with respect to orthologous pairs. We also noted that, despite the weak differences in the median dT' and dEK' between CON and DIV groups, a large number of DIV gene pairs showed a displacement towards higher dT' and dEK' values compared to the CON gene pairs (Sup.

Figure 1). This reinforced the link between promoter and expression divergence.

Effect of duplication events over sequence and expression divergence

Even when we consider recently duplicated paralogous gene pairs, as it is the case of this study, some paralogues in a family will be related to each other by more than one gene duplication event. To better dissect the effect of gene duplication on sequence and expression pattern evolution we counted the number of duplication events undergone by each gene, in each lineage, since the Primate-Rodent split (Figure 1). Then, we calculated, for each paralogous gene, sequence and expression divergence with respect to the orthologue/s. Here the divergence time considered will always be the same (~ 65 Mya x 2). For 33% of Primate gene families, and 19% of Rodent gene families, more than one orthologue existed in the other lineage. In such cases we chose the minimum value in each comparison (d_{SM}, Ka/Ks, dT, and dEK). We partitioned the duplicates in 4 groups depending on the number of duplication events undergone by the gene: 0, 1, 2, or more than 2. We found that the calculated d_{SM}, Ka/Ks, and dT increased in a significant manner with the number of gene duplications (Sup. Table 4, Figure 2). Similar results were obtained when we used the average d_{SM}, Ka/Ks, dT, and dEK values, instead of the minimum value (Sup. Table 5).

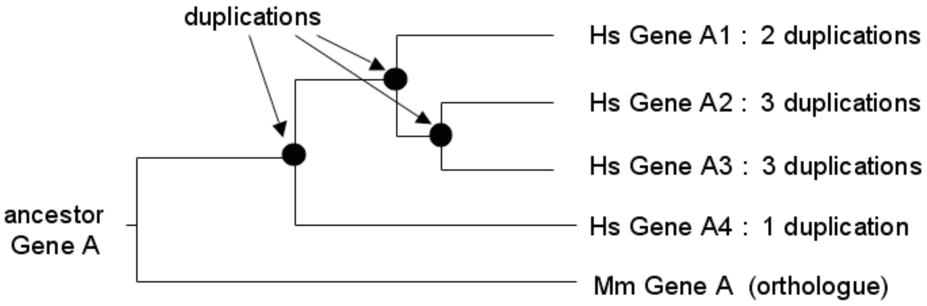

duplications
Hs Gene A1 : 2 duplications
Hs Gene A2 : 3 duplications
Hs Gene A3 : 3 duplications
Hs Gene A4 : 1 duplication
ancestor Gene A
Mm Gene A (orthologue)

Figure 1. Schematic representation of the number of duplications events undergone by Primate-specific duplicates in a gene family. Hs: Homo sapiens. Mm: Mus musculus.

These results showed that the process of gene duplication is associated with an acceleration of the divergence of both coding and promoter sequences, as well as with increased differences in the tissues in which duplicated genes are expressed above a given threshold (dT). Interestingly, dEK did not show any significant trend (Figure 2). This measure takes into account relative expression profiles and will be less sensitive to a general increase or decrease in levels of expression than dT. These results suggest that the effects of gene duplication on gene expression are often quantitative rather than qualitative. This is supported by the significant decrease in tissue expression breadth with the number of gene duplications undergone by the gene (Figure 3, Sup. Table 4).

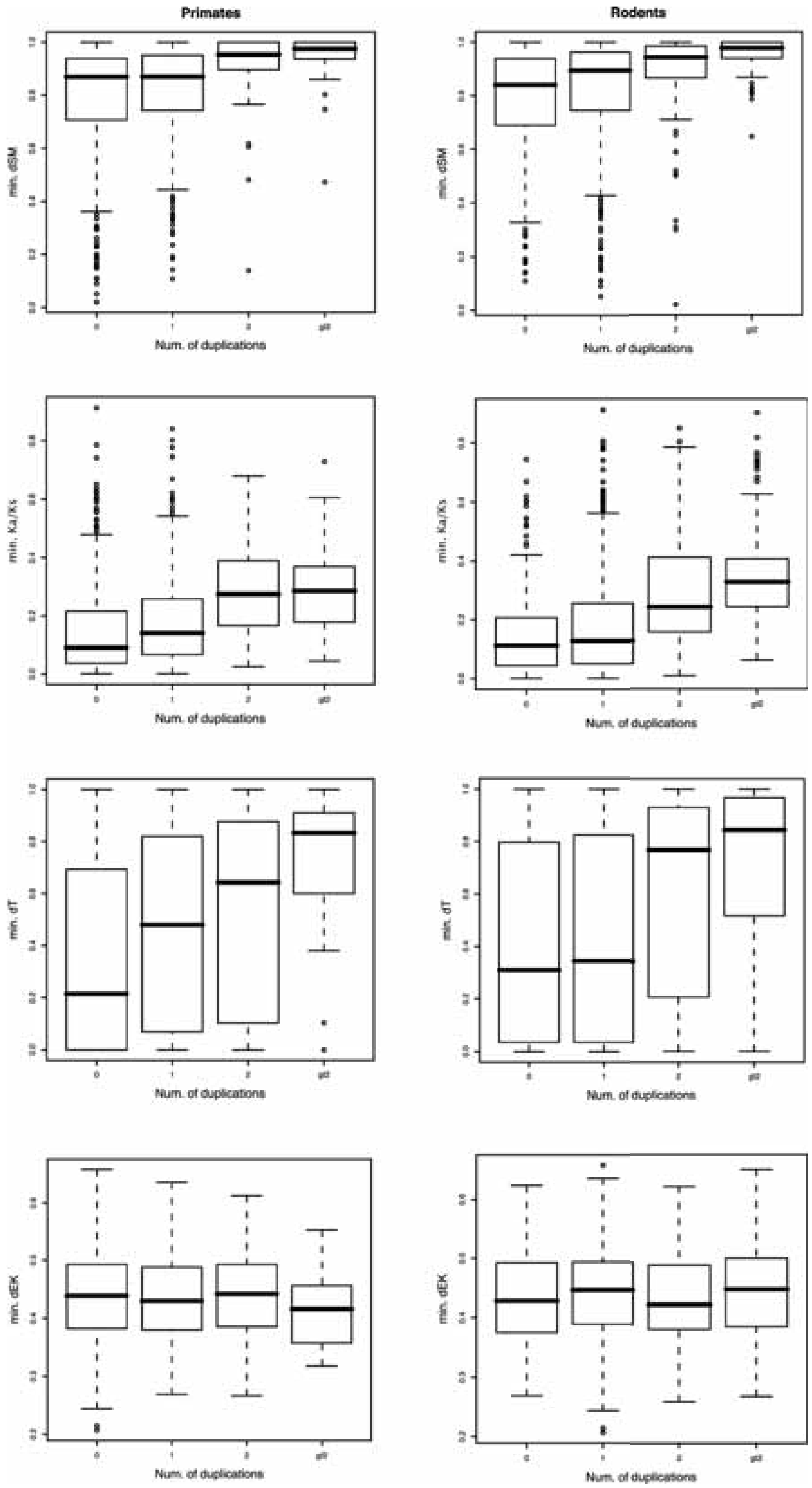


Figure 2. Effect of the number of gene duplications in several divergence measures in relation to the orthologues. d_{SM}: promoter divergence. dN/dS: non-synonymous to synonymous ratio. dT: expression divergence based on the presence or absence of expression in a tissue above a given threshold (>200), and dEK, distance based on relative expression pattern profiles. min. refers to the fact that, when several orthologues were present, the minimum value for the measure was taken. The horizontal line in the box-plot indicates the media.

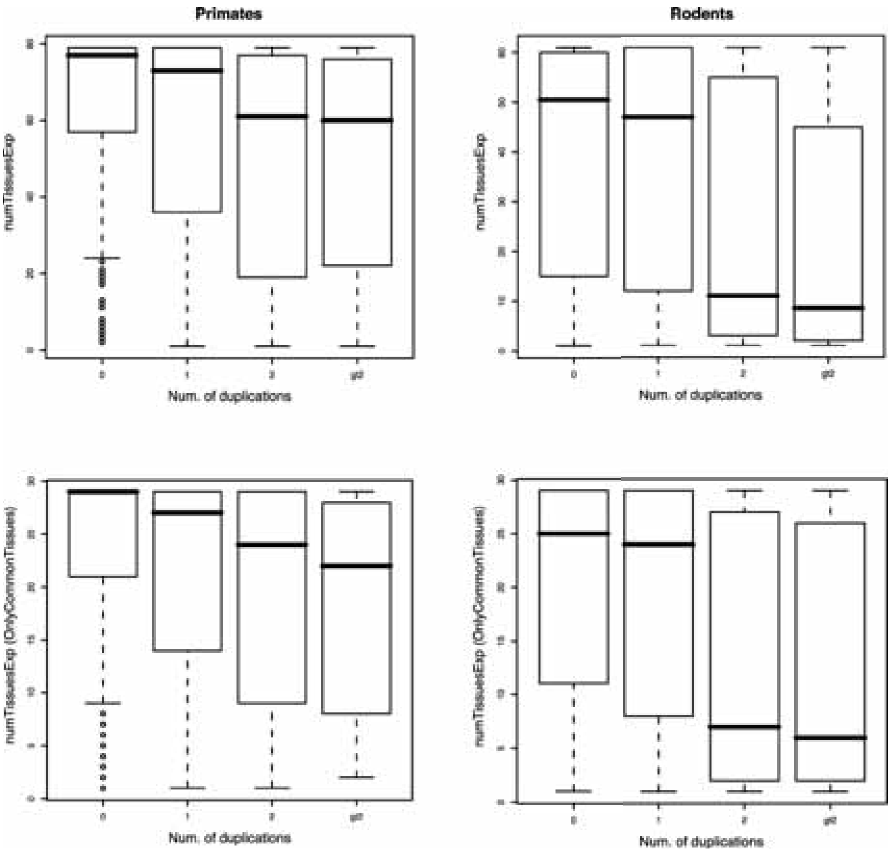


Figure 3. Relationship between number of gene duplications and tissue expression breadth. numTissueExp: number of tissues in which a gene is expressed above a given threshold (>200). Analysis was performed using the 29 common tissues with microarray data from Gene Atlas (OnlyCommonTissues) or the complete tissue collection (79 tissues for human genes and 61 for mouse genes).

DISCUSSION

Many different studies have analyzed the effects of gene duplication in protein sequence evolution and, more recently, in expression pattern divergence (see Introduction). In general, the pace of evolutionary change has been shown to increase at both levels, presumably due to relaxation of evolutionary constraints, adaptation to new functions and partial disruption of the gene by the process of gene duplication [29]. Changes in expression in gene duplicates will often be due to changes in gene expression regulatory sequences. However, little is understood about the relationship between regulatory sequence divergence and expression divergence. This is the first report that, to our knowledge, addresses this question.

The study of gene expression regulatory sequences is complicated for several reasons. First, the location of regulatory sequences is often unknown and some of them may be far away from the gene. Second, it is at present virtually impossible to estimate the effect of single mutations, as most of the functional motifs cannot be distinguished from non-functional sequences. Regarding the location of functional regulatory sequences, many of them are known to cluster near the transcription start site [30,31], including some tissue-specific regulatory motifs [30]. Additionally, at distances longer than 2 Kb from the TSS, the similarity between orthologous promoters drastically drops, indicating that most functional regions concentrate in the 2 Kb promoter region [32]. So the analysis of this region should be sufficient to capture most of the regulatory sequence information. The second issue is how to assess promoter sequence divergence. Complete alignment of long promoter regions is not always possible, as some regions are very poorly conserved even across human and mouse orthologues [24,28]. In general, proximal promoter regions are better preserved, whereas the conservation of distal sequences, upstream from –500, varies greatly from gene to gene [28]. For this reason, a local alignment approach is appropriate to study this part of the gene. It also has the advantage that the fraction of the promoter than cannot be reliably aligned can then be used as a measure of sequence divergence [14]. Indeed, no additional information is obtained by calculating nucleotide substitutions in the aligned regions as most of the variability is encapsulated in the length of such regions [28].

One main finding of this study is the significant association between promoter sequence divergence and tissue expression divergence in paralogous pairs. Interestingly, such an association is not observed in orthologous pairs. The degree of divergence of paralogous promoters is in general comparable to that of orthologous promoters, even if the divergence time of paralogous copies is, on average, about half that of orthologous copies. Generation of partial gene copies or retrocopies can have drastic effects on promoter conservation between duplicates [13,29]. Taken together, these observations seem compatible with a model in which promoters tend to diverge quickly after gene duplication, which, in some cases, may entail the gain or loss of entire functional regions. In contrast, the differences in the fraction of the promoter than can be reliably aligned across orthologous genes are more likely to reflect gradual degeneration of non-functional regions, with little or no consequences for the expression of the gene.

The effect of gene duplication on promoter sequence divergence is most clearly seen when divergence time is kept constant, by comparing orthologous human and mouse genes belonging to families with lineage-specific duplications. Such an approach has been used to investigate other aspects related to gene duplication [17,33]. Specifically, we focused on the number of gene duplications undergone by the gene (Figure 1), rather than on the size of the gene family, as the first is more directly related to the process of duplicating one gene into two daughter copies. We found a progressive increase in promoter divergence with the number of gene duplications. A similar result was obtained for protein divergence (Ka/Ks ratio), although there was no co-variation in the evolution of the two types of regions. Regarding expression, differences in the tissues where genes were expressed increased with the number of gene duplications, but we did not detect any significant differences in the relative expression profiles. One possible explanation is that for the latter we did not use any expression threshold, and any general effects on the level of expression would not have been visible. Our results were consistent with the previously reported positive relationship between the size of mammalian gene families and expression tissue divergence [17]. These authors also observed that genes from large families tended to be expressed in a low number of tissues. Similarly, we found a progressive reduction in expression breadth with the number of gene duplications. This supports the subfunctionalization model for gene duplicates and specifically, it could reflect an accumulation of degenerative mutations leading to reduced levels of expression. But several examples described by Huminiecki and Wolfe [17], and observations from our own dataset, show that duplicates sometimes show novel expression in particular tissues. These cases would fit the neofunctionalization model, and in fact both models appear to coexist in many families.

In general, the results obtained with Primate and Rodent duplicates are highly consistent, although there are some quantitative differences. First, the number of Rodent duplicates is around twice the number of Primate duplicates. It has been argued that the increased gain of gene duplicates in the mouse lineage with respect to the human lineage supports a role of positive selection in the retention of duplicate genes, as the larger population size of mouse implies more effective natural selection [33]. We have observed that the rate of lineage-specific expansion also depends on the function of the gene. For example olfactory receptors genes show much greater expansion in mouse (Sup. Table 1). Second, mouse duplicates show higher sequence divergence than human duplicates, both in the promoter region and in coding sequences (Table 1). This can be explained by the general observation that about twice as many substitutions have occurred in the mouse lineage compared to the human lineage [34]. The reason why gene expression divergence in mouse duplicates is higher than in human duplicates is not yet clear, but it may be related to the differences in the rate of promoter sequence divergence.

METHODS

Sequence retrieval, sequence alignment, and substitution rate estimation

We identified human and mouse orthologous genes using the Ensembl database release 44 [35]: 23,723 human-mouse gene pairs (17,217 human genes and 18,378 mouse genes). We also used Ensembl to identify human and mouse paralogues. For the human genes, we selected 6879 paralogous gene pairs with Ancestor labelled as “Primate”, “Catarrhini”, “Homo/Pan/Gorilla group”, or “Homo sapiens”. For the mouse genes, we selected 27,544 paralogous gene pairs with Ancestor labelled as “Rodentia”, “Sciurognathi”, “Murinae”, or “Mus musculus”. These paralogous gene pairs, derived from gene duplications occurred after the Primate and Rodent split, comprised to 2,885 human genes and 5,618 mouse genes. We found Gene Ontology (GO) annotations [36] for 1,549 of the human genes (49%) and 3,138 of the mouse genes (56%) using Ensembl Biomart [37].

We extracted promoter sequences from these genes (orthologues and paralogues), comprising 2 Kb upstream of the transcription start site (TSS), from the UCSC database hg18 and mm8 releases [38]. We discarded repeats using RepeatMasker (release 1.1.65)[39]. We aligned promoter sequences with the local pairwise sequence alignment program described in Castillo-Davis et al. [14], using a minimum alignment length of 16 nucleotides, as previously reported [28]. For each gene pair we obtained the promoter sequence divergence (d_{SM}; shared motif divergence), which is the fraction of the promoter sequence that does not align, taking the average between the human and mouse sequences.

We extracted the corresponding coding sequences from Ensembl database (release 44) and performed pairwise sequence alignments with ClustalW [40]. The number of synonymous substitutions per synonymous site (Ks), and the number of non-synonymous substitutions per non-synonymous sites (Ka) were estimated with the codeml program in the PAML package [25]. We discarded those pairs with $Ks < 10^{-4}$, $Ks \geq 2.0$, $Ka \geq 2.0$, or $Ka/Ks \geq 10.0$, containing too few or too many changes for meaningful substitution rate estimation. We successfully calculated d_{SM}, Ks, Ka, and Ka/Ks values for 12,726 orthologous gene pairs, 958 human paralogous gene pairs, and 1,719 mouse paralogous gene pairs.

Phylogenetic trees and duplication events estimation

Using the paralogy information from Ensembl, we grouped the 2,885 human genes and the 5,618 mouse genes in 935 and 1,488 gene families, respectively. For each family, we obtained the longer peptide sequence of each gene from Ensembl and performed a multiple alignment with T-Coffee [41]. If available, one orthologous gene was also included in the alignment. In 402 human gene families and 570 mouse gene families this was not possible, as no orthologues were annotated in Ensembl. In 351 human gene families and 738 mouse gene families there was a unique orthologue annotated. However, in 176 human gene families and 176 mouse gene families there was more than one orthologue. We built separate sequence alignments for each of these orthologues. Once the alignment had been built, we used SEQBOOT, PROTDIST, NEIGHBOR, and CONSENSE programs of the PHYLIP package [42] to estimate the phylogenetic tree of the family, applying neighbour-joining distance method with bootstrapping (1000 replications). In cases where more than one orthologous gene existed, we obtained a phylogenetic tree for each alignment and finally calculated a consensus tree. To estimate the number of duplication events for each gene in its phylogeny (Primates or Rodents), we counted the number of duplication nodes with bootstrap values of 700 (70%) or greater from the tree root to the gene node. Branches with lower bootstrap values were collapsed.

Expression divergence

We used human and mouse gene expression data from Gene Atlas (GNF1H and GNF1M), based on transcriptome microarray data [43]. Gene Atlas covers 79 human and 61 mouse organs and tissues. To calculate expression divergence of orthologous and paralogous gene pairs, we selected 29 tissues/organs that appear in both human and mouse datasets (adipose tissue, bone marrow, skin, skeletal muscle, lymph node, heart, trachea, lung, tongue, liver, pancreas, kidney, ovary, placenta, prostate, testis, uterus, adrenal gland, amygdala, pituitary gland, salivary gland, thyroid, thymus, hypothalamus, trigeminal, dorsal root ganglia, spinal cord, cerebellum, olfactory bulb).

We defined a first tissue expression divergence measure between gene pairs, based on the presence or absence of transcripts in different tissues, as follows:

$$dT = \frac{(ntu - nti)}{ntu}$$

where ntu is the number of tissues that show expression of any of the two genes (union), and nti is the number of tissues that show expression of both genes (intersection). We considered genes to be expressed in a tissue according to Gene Atlas data only if they show an Affymetrix average difference >= 200 [43]. Values of dT are between 0 (same expression) and 1 (completely different expression).

We also defined a second, relative measure of gene expression divergence, based on the correlation of expression profiles, as follows:

$$dEK = \frac{(1-\tau)}{2}$$

where τ is the Kendall tau rank correlation coefficient of expression levels in different tissues of the Gene Atlas dataset. In this case no expression threshold was used. Values of dEK are between 0 (τ = 1, full positive correlation) and 1 (τ = -1, full negative correlation).

Statistical tests and correlations

Correlations were calculated with the Spearman Rank correlation method (except the Kendall tau rank correlation used to calculate the expression divergence measure dEK). Two-sample Wilcoxon-Mann-Whitney statistical test was used to assess differences between groups unless stated. We also used multiple regression (formulas: dT ~ d_{SM} + Ka + Ks and dEK ~ d_{SM} + Ka + Ks) to asses if the correlation between promoter or protein sequence divergence and expression divergence could be fully attributed to the separate correlation of these variables with Ks. The R statistical package was used for all calculations[44].

REFERENCES

1. Prince VE, Pickett FB: **Splitting pairs: the diverging fates of duplicated genes**. *Nat Rev Genet* 2002, **3**:827-837.
2. Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A: **RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon**. *Mol Cell Biol* 1985, **5**:2090-2103.
3. Brenner SE, Hubbard T, Murzin A, Chothia C: **Gene duplications in H. influenzae**. *Nature* 1995, **378**:140.
4. Teichmann SA, Park J, Chothia C: **Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements**. *Proc Natl Acad Sci U S A* 1998, **95**:14658-14663.
5. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919.
6. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151-1155.
7. Ohno S: **Evolution by Gene Duplication**. New York: Springer-Verlag; 1970.
8. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459-473.
9. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**:1531-1545.
10. Hughes AL: **The evolution of functionally novel proteins after gene duplication**. *Proc Biol Sci* 1994, **256**:119-124.
11. Van de Peer Y, Taylor JS, Braasch I, Meyer A: **The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes**. *J Mol Evol* 2001, **53**:436-446.
12. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A *et al*: **Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences**. *Nature* 2007, **447**:167-177.
13. Cusack BP, Wolfe KH: **Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates**. *Mol Biol Evol* 2007, **24**:679-686.
14. Castillo-Davis CI, Hartl DL, Achaz G: **cis-Regulatory and protein evolution in orthologous and duplicate genes**. *Genome Res* 2004, **14**:1530-1536.
15. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications**. *Genome Biol* 2002, **3**:RESEARCH0008.
16. Nembaware V, Crum K, Kelso J, Seoighe C: **Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs**. *Genome Res* 2002, **12**:1370-1376.
17. Huminiecki L, Wolfe KH: **Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse**. *Genome Res* 2004, **14**:1870-1879.
18. Makova KD, Li WH: **Divergence in the spatial pattern of gene expression between human duplicate genes**. *Genome Res* 2003, **13**:1638-1645.

19. Ganko EW, Meyers BC, Vision TJ: **Divergence in expression between duplicated genes in Arabidopsis**. *Mol Biol Evol* 2007, **24**:2298-2309.
20. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data**. *Trends Genet* 2002, **18**:609-613.
21. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes**. *Mol Biol Evol* 2003, **20**:1377-1419.
22. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover**. *Mol Biol Evol* 2002, **19**:1114-1121.
23. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nat Genet* 2007, **39**:730-732.
24. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S: **Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions**. *Genome Res* 2004, **14**:1711-1718.
25. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**:1586-1591.
26. Lee S, Kohane I, Kasif S: **Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes**. *BMC Genomics* 2005, **6**:168.
27. Iwama H, Gojobori T: **Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network**. *Proc Natl Acad Sci U S A* 2004, **101**:17156-17161.
28. Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM: **Housekeeping genes tend to show reduced upstream sequence conservation**. *Genome Biol* 2007, **8**:R140.
29. Katju V, Lynch M: **On the formation of novel genes by duplication in the Caenorhabditis elegans genome**. *Mol Biol Evol* 2006, **23**:1056-1067.
30. Bellora N, Farre D, Alba MM: **Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters**. *BMC Genomics* 2007, **8**:459.
31. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters**. *Genome Res* 2004, **14**:1562-1574.
32. Keightley PD, Lercher MJ, Eyre-Walker A: **Evidence for widespread degradation of gene control regions in hominid genomes**. *PLoS Biol* 2005, **3**:e42.
33. Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH: **Role of positive selection in the retention of duplicate genes in mammalian genomes**. *Proc Natl Acad Sci U S A* 2006, **103**:2232-2236.
34. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
35. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2007**. *Nucleic Acids Res* 2007, **35**:D610-617.
36. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and**

informatics resource. *Nucleic Acids Res* 2004, **32**:D258-261.

37. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data**. *Genome Res* 2004, **14**:160-169.
38. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F *et al*: **The UCSC Genome Browser Database: update 2006**. *Nucleic Acids Res* 2006, **34**:D590-598.
39. **RepeatMasker** [http://www.repeatmasker.org/]
40. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.
41. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**:205-217.
42. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6**. In.; 2005.
43. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**:6062-6067.
44. **R Project** [http://www.r-project.org/]

SUPPLEMENTARY MATERIAL

Gene Ontology (GO)		Recent duplicates				All genes	
Id	Description	N Hs (total 1549)	N Mm (total 3138)	p obs Hs	p obs Mm	p exp Hs	p exp Mm
GO:0004872	receptor activity	180 ***	1065 ***	0.116	0.339	0.082	0.132
GO:0007186	G-protein coupled receptor protein signaling pathway	176 ***	1046 ***	0.113	0.333	0.072	0.114
GO:0001584	rhodopsin-like receptor activity	150 ***	968 ***	0.096	0.308	0.049	0.087
GO:0007165	signal transduction	205 *	865 ***	0.132	0.275	0.115	0.117
GO:0004984	olfactory receptor activity	127 ***	776 ***	0.082	0.247	0.026	0.057
GO:0050896	response to stimulus	94 ***	701 ***	0.060	0.223	0.027	0.055
GO:0007608	sensory perception of smell	86 ***	678 ***	0.055	0.216	0.018	0.050
GO:0005886	plasma membrane	31	880 ***	0.020	0.280	0.042	0.129
GO:0004977	melanocortin receptor activity	33	305 ***	0.010	0.097	0.021	0.025
GO:0005576	extracellular region	137 ***	220	0.088	0.070	0.045	0.077
GO:0006955	immune response	104 ***	86 **	0.067	0.027	0.031	0.018

Supplementary Table 1. Gene Ontology (GO) functions over-represented in mouse and/or human rencent gene duplicates. Recent duplicates have originated after the Primates-Rodent split. N: number of genes. The total number is the number of genes annotated with GO functions. Hs: Homo sapiens. Mm: Mus musculus. p obs: observed frequency. p exp: expected frequency (all genes in Ensembl). Only the top 20 associated GO terms in human, mouse, or both, were considered. Binomial probabilities wre used to test statistical significance. *** $p < 10^{-6}$; ** $p < 10^{-4}$; * $p < 0.05$. Pale grey: functions with higher relative frequency in Mm than in Hs recent duplicates. Dark grey: functions with higher relative frequency in Hs than in Mm recent duplicates.

Gene Pairs	N	r(Ks,d_{SM})	r(Ks,Ka)	r(Ka,d_{SM})
Human-mouse orthologs	10,511	0.227	0.403	0.266
Primate duplicates	787	0.318	0.858	0.258
Rodent duplicates	1,155	0.561	0.757	0.408

Supplementary Table 2. Correlation between various sequence divergence measures. N: number of gene pairs. d_{SM}: promoter divergence; Ks: synonymous substitution rate; Ka: non-synonymous substitution rate. r: Spearman rank correlation coefficient. All correlations were highly significant ($p < 10^{-12}$).

Gene pairs	d_{SM} '	N	Ka'	Ks'	dT'	dEK'
Primate duplicates	≤ 1 (CON)	269	0.415	0.228	0.854	0.721
			0.361	0.200	0.875	0.919
			0.497	0.175	1.355	0.606
	> 1 (DIV)	119	0.815	0.322	1.181	0.913
			0.464	0.340	1.055	1.108
			2.310	0.187	1.850	0.504
	p-value		$<10^{-4}$	$<10^{-5}$	<0.2	<0.02
Rodent duplicates	≤ 1 (CON)	414	0.619	0.336	5.633	0.982
			0.484	0.226	1.667	1.073
			0.565	0.385	6.469	0.439
	> 1 (DIV)	544	1.268	0.942	7.964	1.156
			0.979	0.859	1.667	1.161
			1.173	0.691	10.390	0.394
	p-value		$<10^{-48}$	$<10^{-66}$	<0.3	$<10^{-3}$

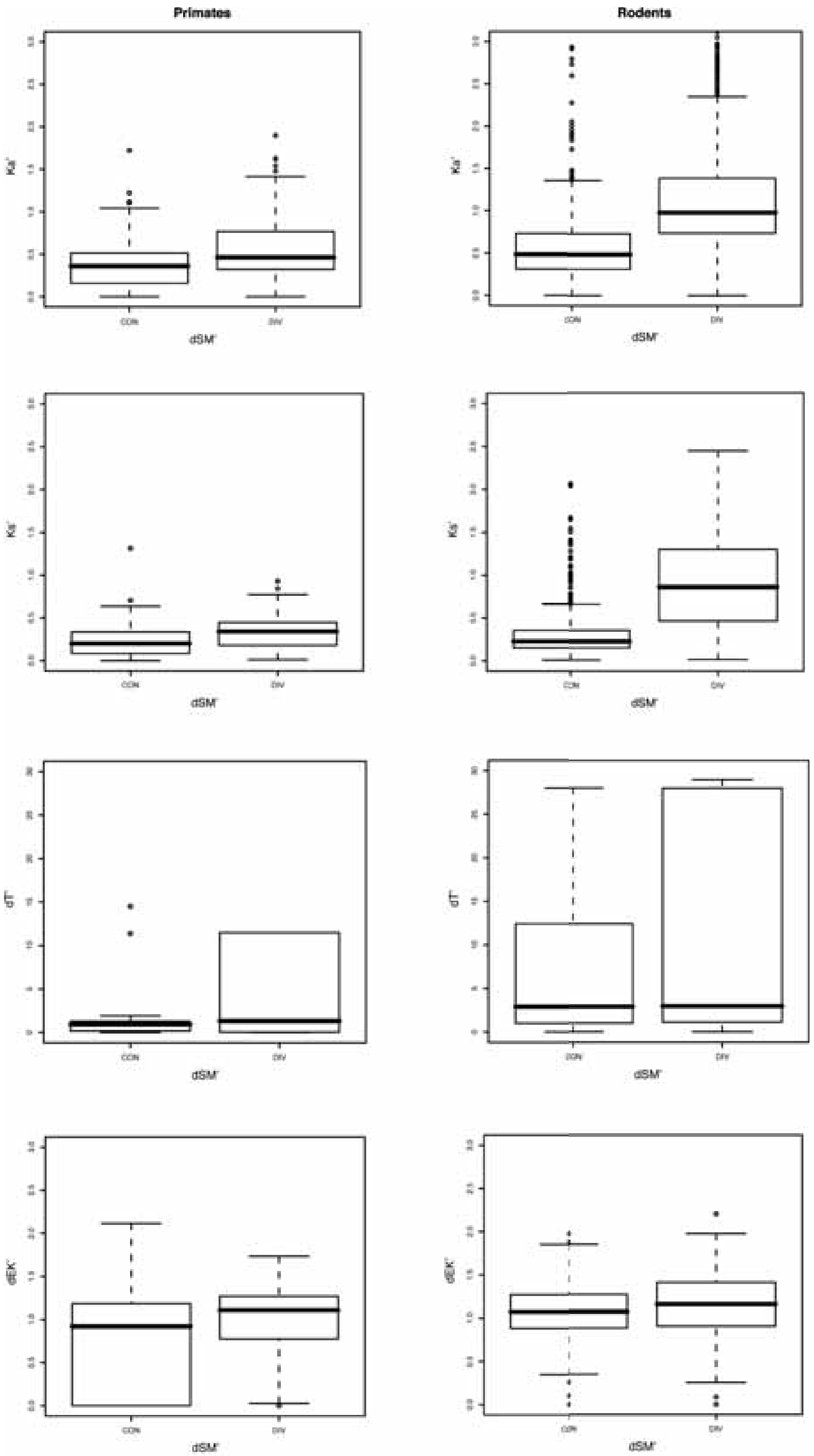
Supplementary Table 3. Sequence and expression divergence of duplicate pairs with respect to ortholoous pairs of the same gene family (with the lowest promoter divergence).. N: number of gene pairs; d_{SM}', Ks', dT', dEK': promoter sequence divergence, non-synonymous substitution rate, synonymous substitution rate, tissue expression divergence, and relative expression profile divergence with respect to the "closest" (less divergent) orthologue (see text for details). Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. Two-sample Wilcoxon-Mann-Whitney test p-values indicated.

Genes	Number of duplication events	N	d_{SM}	Ka/Ks	N (dT)	dT	N (dEK)	dEK	N (NT)	NT
Primate duplicates	0	506	0.796	0.146	349	0.355	368	0.436	451	23.557
			0.870	0.091		0.214		0.439		29.000
			0.202	0.155		0.360		0.079		8.382
	1	417	0.816	0.188	287	0.466	305	0.435	360	21.861
			0.871	0.141		0.480		0.430		27.000
			0.187	0.160		0.372		0.076		9.251
	2	86	**0.920**	**0.281**	56	0.536	62	0.444	68	18.897
			0.954	0.275		0.643		0.442		24.000
			0.125	0.143		0.393		0.070		10.805
	> 2	58	**0.943**	**0.288**	39	**0.690**	41	0.419	45	17.956
			0.975	0.285		0.833		0.416		22.000
			0.103	0.137		0.345		0.071		10.675
	p-value (K–W test)		$<10^{-15}$	$<10^{-12}$		0.006		0.149		0.003
Rodent duplicates	0	246	0.779	0.151	171	0.409	182	0.436	169	19.899
			0.841	0.112		0.310		0.429		25.000
			0.213	0.143		0.375		0.079		10.261
	1	665	0.825	**0.180**	427	0.431	454	0.442	426	19.056
			0.895	0.129		0.345		0.447		24.000
			0.190	0.184		0.378		0.075		10.704
	2	147	0.886	**0.312**	89	0.602	96	0.435	91	12.747
			0.944	0.249		0.769		0.422		7.000
			0.156	0.246		0.371		0.079		11.606
	> 2	139	**0.960**	**0.391**	86	**0.688**	95	0.443	78	**11.590**
			0.978	0.331		0.845		0.448		6.000
			0.055	0.325		0.353		0.084		11.122
	p-value (K–W test)		$<10^{-22}$	$<10^{-35}$		$<10^{-8}$		0.621		$<10^{-9}$

Supplementary Table 4. Effect of number of duplication events on sequence and expression divergence. All values were calculated for pairs of orthologues from the duplicate gene families. When several orthologues was present the minimum value for each variable was taken. For consistence, genes with 0 duplication belonged to the duplicated gene family dataset, and had duplications in the other lineage. Mean (top), median (middle), and standard deviation (bottom) are indicated for each variable. N: number of genes with available data (depends on the variable); d_{SM}:promoter divergence; Ka/Ks: non-synonymous substitution to synonymous substitution ratio; dT: tissue expression divergence; dEK: relative expression profile divergence; NT: number of tissues in which a gene is expressed out of 29 tissues. Numbers in bold indicate significant differences at p < 10-5 with the other groups (two-sample Wilcoxon-Mann-Whitney test). K-W test: Kruskal-Wallis test that evaluates differences between the different groups.

Genes	Number of duplication events	N	d_{SM}	Ka/Ks	N (dT)	dT	N (dEK)	dEK	N (NT)	NT
Primate duplicates	0	506	0.812	0.186	349	0.391	368	0.445	451	23.557
			0.881	0.134		0.276		0.451		29.000
			0.194	0.188		0.367		0.075		8.382
	1	417	0.835	0.201	287	0.513	305	0.446	360	21.861
			0.898	0.148		0.556		0.446		27.000
			0.179	0.169		0.362		0.074		9.251
	2	86	**0.937**	**0.318**	56	0.647	62	0.470	68	18.897
			0.975	0.290		0.732		0.468		24.000
			0.120	0.166		0.334		0.062		10.805
	> 2	58	**0.958**	**0.336**	39	**0.791**	41	0.448	45	17.956
			0.987	0.358		0.833		0.452		22.000
			0.090	0.168		0.213		0.063		10.675
	p-value (K–W test)		$<10^{-18}$	$<10^{-14}$		$<10^{-4}$		0.064		0.003
Rodent duplicates	0	246	0.794	0.179	171	0.436	182	0.446	169	19.899
			0.860	0.135		0.380		0.445		25.000
			0.206	0.150		0.373		0.078		10.261
	1	665	**0.839**	**0.184**	427	0.438	454	0.446	426	19.056
			0.906	0.131		0.362		0.451		24.000
			0.181	0.187		0.377		0.074		10.704
	2	147	**0.903**	**0.333**	89	0.619	96	0.445	91	12.747
			0.954	0.258		0.793		0.444		7.000
			0.150	0.253		0.376		0.077		11.606
	> 2	139	**0.969**	**0.419**	86	**0.709**	95	0.455	78	**11.590**
			0.990	0.368		0.882		0.458		6.000
			0.051	0.325		0.343		0.085		11.122
	p-value (K–W test)		$<10^{-25}$	$<10^{-39}$		$<10^{-9}$		0.434		$<10^{-9}$

Supplementary Table 5. Effect of number of duplication events on sequence and expression divergence. As in Sup. Table 4 but in this case when several orthologues are present the average value is taken. See Sup. Table 4 legend.



Supplementary Figure 1. Divergence in paralogous with respect to their corresponding orthologous.

Part IV

DISCUSSIÓ

Resum i Discussió Global

El treball realitzat durant el meu doctorat ha estat enfocat a l'estudi de les seqüències reguladores de la transcripció. L'estudi ha contemplat diferents aspectes d'aquestes seqüències: la caracterització i representació de TFBSs coneguts, la predicció de nous motius, les restriccions que intervenen en la conservació d'aquestes seqüències, la seva dinàmica evolutiva, . . . A continuació analitzo globalment el treball realitzat, els resultats obtinguts, les qüestions pendents i les perspectives d'aquest camp d'estudi.

Caracterització, representació i predicció de TFBSs

La identificació de les regions de regulació transcripcional i dels elements que les composen juga un paper fonamental a l'hora de definir les xarxes de regulació gènica. Fins recentment la descodificació d'aquestes xarxes l'han realitzat el biòlegs mitjançant anàlisis d'alteracions molt costosos (Davidson et al., 2002). La predicció *in silico* de potencials llocs reguladors és un eina útil per reduir la quantitat d'interaccions proteïna-DNA a testar experimentalment. Però la problemàtica pròpia de la predicció de TFBSs fa que els programes existents siguin sovint de poca utilitat per a l'investigador degut a la gran quantitat de falsos positius.

Amb PROMO (Messeguer et al., 2002; Farre et al., 2003) vàrem intentar millorar la predicció de TFBSs principalment per dues vies: (1) creant matrius de pes més específiques (per a una determinada espècie biològica o un determinat tàxon) i (2) fent predicció a vàries seqüències a la vegada. Proves realitzades comparant PROMO amb altres programes que també utilitzen TRANSFAC com a base de dades (per exemple, MatInspector) mostraren que PROMO millorava significativament la sensibilitat, gràcies a l'ús de matrius per nivell taxonòmic. En canvi, amb prou feines vàrem aconseguir millorar significativament l'especificitat, el gran problema de les prediccions de TFBSs. La predicció en vàries seqüències a la vegada, de motius comuns a totes les seqüències o com a mínim de una determinada fracció de elles, sí permet millorar en un cert grau l'especificitat (reduir el nombre de falsos positius), depenent sempre del nombre de seqüències estudiades i del grau de similitud entre elles. A més d'aquestes millores en la predicció, vàrem dissenyar PROMO amb una interfície d'usuari intuïtiva, amb una sortida gràfica fàcil d'interpretar.

Després de desenvolupar PROMO ens hem plantejat si és possible millorar les prediccions millorant la qualitat de les matrius. Certament crec que hi ha marge per millorar la qualitat de les matrius utilitzant algorismes més sofisticats en la seva creació i així millorar les prediccions. Però hi ha una problemàtica determinada per dues causes.

La primera causa són els límits propis del model de matrius de pes (*position-specific weight matrix* o

PWM). Aquest model assumeix que cada posició del lloc d'unió contribueix de manera independent a la interacció proteïna-DNA. Experimentalment s'ha demostrat que la hipòtesi d'independència posicional és falsa (Bulyk et al., 2002). S'han proposat models, molt més complexos, que incorporen dependències entre les posicions (Zhou and Liu, 2004; Osada et al., 2004) que han demostrat millorar de forma significativa tant la sensibilitat com l'especificitat de les prediccions. Un altre problema del model de PWM bé donat per la impossibilitat d'espaiadors de longitud variable dins la matriu, entenent per espaiador una sèrie de nucleòtids que l'únic paper que té és separar dues parts del TFBS. Certs factors de transcripció, per exemple el receptor de l'estrogen (Gruber et al., 2004), s'uneixen al DNA com homodímers o heterodímers, amb un nombre lleugerament variable de nucleòtids separant els dos llocs dels monòmers. La predicció del monòmers, per separat, és sovint molt difícil ja que són molt curts (4-5 pb) i força variables. Sovint també hi ha dependències entre les posicions entre els monòmers. Per poder fer prediccions dels dímers, caldria un model de representació més complex que el de les matrius de pes. Actualment hi ha disponible per algun d'aquests motius programes específics per a la seva predicció (Bajic et al., 2003; Favorov et al., 2005).

L'altra causa, molt més important, que limita actualment la qualitat de les prediccions, és la quantitat i la qualitat de les dades disponibles. PROMO actualment utilitza les dades de la versió 8.3 de TRANSFAC (Matys et al., 2003), la base de dades que conté una col·lecció més amplia de TFs i TFBSs d'eucariotes. La versió 8.4 (amb molt poques diferències respecte a la 8.3) té 7.796 entrades de gens (taula 1), comptant totes les espècies. Aquest nombre queda lluny del total del nombre de gens humans i molt més lluny del total de gens de totes les espècies incloses a TRANSFAC. De fet, en l'estudi fet sobre gens ortòlegs d'humà i ratolí (Farre et al., 2007) vàrem poder mapar TFBSs de TRANSFAC sobre menys del 2 % dels 17.944 gens estudiats (8.972 parells d'ortòlegs). Analitzant l'última versió de TRANSFAC (v. 11.4) veiem que s'ha produït un important increment en nombre de gens i de factors, aproximadament el doble respecte a la versió 8.4, però en canvi el nombre de llocs d'unió anotats s'ha incrementant en una proporció molt inferior (taula 1).

A més del problema de la poca quantitat de dades, hi ha un problema de qualitat. Hi ha un biaix cap a determinats gens que estan més exhaustivament anotats i hi ha un biaix cap a determinats TFs (per exemple, TBP o Sp1) que tenen molts més motius anotats que la resta de factors (Fogel et al., 2005). En canvi, molts TFs tenen pocs motius anotats, sent impossible construir matrius en aquests casos. Per molts dels motius falten dades imprescindibles (posicions, seqüència de nucleòtids, gen) per a poder validar la seva procedència. A més, la immensa majoria dels motius no tenen anotada una mínima valoració de la

Taula	**v. 6.4**	**v. 7.4**	**v. 8.4**	**v. 11.4**
FACTOR	5.072	5.401	5.919	10.622
SITE	12.760	13.302	14.782	20.925
FACTOR-SITE Links	15.719	16.479	18.748	26.951
GENE	4.065	6.692	7.796	19.688
CHIP-chip FRAGMENTS	-	-	-	16.884
MATRIX	610	695	741	834
Class	50	51	54	57
Cell	1.536	1.567	1.800	2.907
REFERENCE	10.619	11.095	11.900	16.155

Taula 1 **Estadístiques de les diferents versions de TRANSFAC.** S'indiquen el nombre d'entrades que hi ha en cada taula en les diferents versions que hem pogut analitzar (6.4, 7.4, 8.4) i en l'última versió tal com figura a la *web* de TRANSFAC (http://www.biobase-international.com).

qualitat del *binding site*, informació molt important a l'hora de decidir quins motius es descarten i quins es tenen en compte a l'hora de construir les matrius de pes. Per tant, necessitem bases de dades més completes i més ben anotades. Per sort en els darrers anys s'estan fet importants esforços en aquest sentit (Montgomery et al., 2006; Griffith et al., 2008; Blanco et al., 2006; Portales-Casamar et al., 2007).

En relació amb la quantitat de dades, cal tenir en compte que les tècniques tradicionals d'estudi de motius d'unió de TFs treballen a petita escala i no permeten un increment accelerat de la quantitat de resultats. Les noves tècniques a gran escala, com ChIP-on-chip (Kim and Ren, 2006) o ChIPSeq (Johnson et al., 2007; Robertson et al., 2007) són la solució a aquest coll d'ampolla. Aquestes noves tècniques apliquen a escala genòmica la tècnica de la immunoprecipitació de la cromatina (*chromatin immunoprecipitation* o simplement ChIP), molt utilitzada per a detectar interaccions proteïna-DNA *in vivo* (Solomon and Varshavsky, 1985; Solomon et al., 1988). La tècnica Chip-on-chip combina ChIP amb *microarrays* de DNA genòmic (figura 6). Aquesta tècnica s'ha utilitzat bastant els darrers anys i TRANSFAC ha incorporat informació sobre resultats d'experiments Chip-on-chip en les darreres versions (taula 1).

ChIP-sequencing o ChIPSeq és una tècnica molt més nova que combina ChIP amb les tècniques de seqüenciació d'última generació (d'Illumina/Solexa, per exemple). Aquesta tècnica no requereix l'ús de *microarrays* existents o de creació de nous; per això, resulta més barata, més ràpida i de més amplia

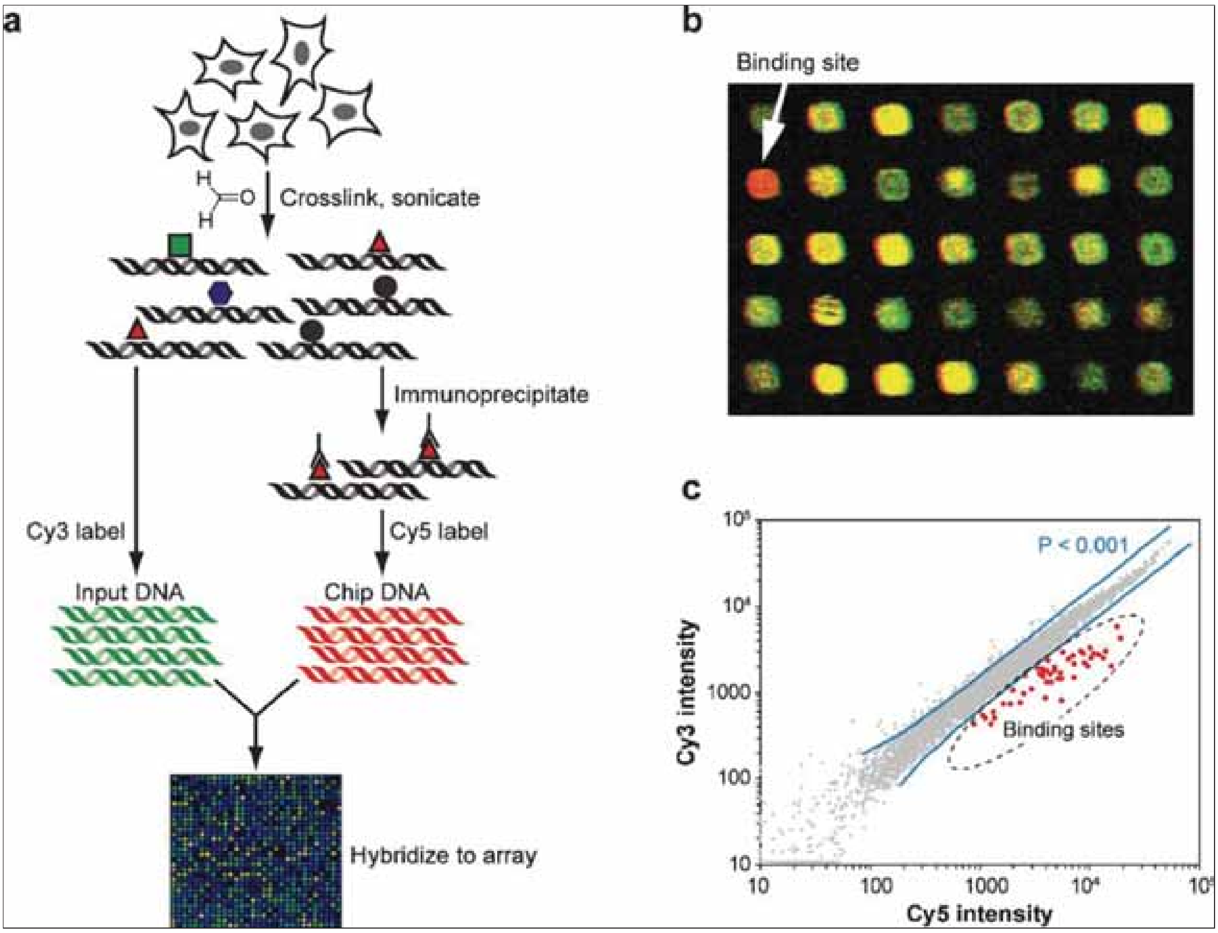


Figura 6 ChIP-on-chip. (a) Esquema del procediment experimental. (b) Una visió ampliada de les proves sobre l'*array* després de la hibridació. Els punts vermells indiquen els llocs d'unió de les proteïnes. (c) Mètode estadístic per a detectar els fragments de DNA ChIP-enriquits. Adaptat a partir de Kim and Ren 2006.

aplicació (es pot aplicar a qualsevol genoma).

En resum, per obtenir millors prediccions de TFBSs, no només necessitem un model de caracterització de TFBSs més sofisticat, sinó també una representació més sofisticada (amb dependències entre les posicions). I sobretot, necessitem moltes més dades ben anotades a partir de les que construir la representació dels TFBSs. Com he indicat, l'ús de matrius mé específiques (d'una espècie o un tàxon) millora les prediccions; però molt sovint ha estat difícil construir una matriu per a una determinada espècie o tàxon per manca de dades suficients en aquest nivell taxonòmic. Properes versions de PROMO han de tenir en compte aquests aspectes problemàtics, incorporant les solucions disponibles.

Tot i la gran quantitat d'eines computacionals que s'han desenvolupat i es desenvoluparan en aquest camp, tot i les millores que es produiran en la quantitat de dades a partir de les que construir les representacions dels TFBSs, la predicció computacional d'elements reguladors i, en particular, la predicció de TFBSs individuals continuarà sent un tasca molt difícil, sobretot degut a les dependències contextuals de la regulació transcripcional (Wray et al., 2003). Per millorar encara més els resultats de les prediccions caldrà incorporar dades de context (interaccions TF-TF, cofactors, nucleosomes, ...) en un model de prediccions molt més complex. De fet existeixen alguns programes que ja fan prediccions de combinacions de motius, que de forma explícita o implícita tenen en compte les interaccions entre factors de transcripció (Kel-Margoulis et al., 2003; Bulyk et al., 2004; Zhu et al., 2005; Waleev et al., 2006); però les dades sobre interaccions entre factors de transcripció són molt escasses (Kel-Margoulis et al., 2002).

Restriccions sobre les seqüències reguladores de la transcripció

Bona part del meu doctorat ha estat dedicada a estudiar un aspecte molt interessant, que pot ajudar a millorar els mètodes de *phylogenetic footprinting* (Tagle et al., 1988; Dermitzakis and Clark, 2002; Lenhard et al., 2003) o la detecció de motius reguladors comuns en gens amb similar perfil d'expressió: conèixer millor les restriccions que operen en la conservació de les seqüències promotores entre gens ortòlegs, entre gens paràlegs i entre gens que tenen similars patrons d'expressió.

Les regularitats que caracteritzen les seqüències codificants estan absents dels promotors. No existeix un codi genètic o una altra característica que relacioni de forma consistent la seqüència amb la funció i aquest fet té tremendes implicacions en l'estudi de l'evolució de l'estructura del promotor i de la funció.

Estudis de genòmica comparativa combinada amb experiments *in vivo* han demostrat que les seqüències no-codificants conservades són bones candidates a contenir elements reguladors de la transcripció (Aparicio et al., 1995; Loots et al., 2000; Hardison, 2000; Nobrega et al., 2003; Woolfe et al., 2005). Però no coneixem quines restriccions actuen sobre aquesta conservació.

La conservació entre promotors es dóna de manera discontínua, per blocs (Suzuki et al., 2004). Per això, per comparar les seqüències de dos promotors, l'alineament local és molt més adequat que l'alineament global. En els estudis que he realitzat, per comparar promotors he fet servir un programa d'alineament local desenvolupat per Castillo-Davis et al. (2004) que calcula un *score*, dSM (*shared motif divergence*) que correspon a la fracció de seqüència no alineada. Aquesta dSM és una estimació de la divergència entre les seqüències de dos promotors. Aquesta mesura ens ha permès tenir una manera de comparar la divergència de la seqüència promotora amb la divergència de la seqüència codificant (Ka, Ks, Ka/Ks) i amb l'expressió d'un dels gens o la divergència d'expressió entre els dos gens.

En els estudis que hem fet comparant parelles de gens ortòlegs o parelles de gens paràlegs hem utilitzat seqüències de 2000 pb (*upstream* del TSS). A distàncies superiors a 2000 pb del TSS, la semblança entre promotors ortòlegs es redueix dràsticament, indicant que la majoria d'elements reguladors es concentren dintre de la regió de 2000 pb *upstream* del TSS (Keightley et al., 2005). Per tant, l'anàlisi d'aquesta regió ha de ser suficient per capturar la majoria de les seqüències reguladores funcionals. Ara bé, aquesta aproximació clarament no té en compte elements reguladors en localitzacions menys obvies, com els introns i les regions més llunyanes *upstream* o *downstream* dels gens.

A partir d'una amplia col·lecció de gens ortòlegs humà-ratolí vàrem demostrar, per primera vegada, la relació existen entre la divergència de la seqüència promotora i l'amplitud d'expressió (*expression breadth*: el nombre de teixits, òrgans o tipus cel·lulars en que s'expressa un gen). Els gens que s'expressen en tots o gairebé tots el teixits (*housekeeping genes* o HK) tendeixen a tenir un promotor menys conservat respecte als que tenen una expressió més restringida. Les diferències es fan més fortes en posicions més distals, aproximadament a partir de -500 *upstream* del TSS. Donada l'alta conservació dels patrons d'expressió dels gens *housekeeping* entre els diferents organismes (Yang et al., 2005), aquesta alta divergència del promotor més aviat indica unes restriccions funcionals més dèbils en aquest tipus de gens i no una diversificació evolutiva.

Els nostres resultats suggereixen que els promotors de gens *housekeeping* tenen promotors més curts, més simples. Sembla ser que la tendència a una longitud menor és comú en altres parts d'aquests gens (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Vinogradov, 2006): la seqüència codificant, les seqüències intròniques, les seqüències intergèniques. També en el(s) transcript(s), però amb la curiosa excepció de les plantes (Ren et al., 2006).

Una hipòtesi que intenta explicar aquesta tendència és la que diu que hi ha una selecció per a l'economia en la transcripció i la traducció (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003). Una hipòtesi alternativa, anomenada '*genome design*', afirma que els gens específics de teixits necessiten més quantitat de DNA no codificant degut a la seva més complexa regulació (Vinogradov, 2006). Els nostres resultats estan més d'acord amb aquesta segona hipòtesi, de forma que els gens *housekeeping* tenen un promotor més simple perquè els seus patrons d'expressió són relativament més simples.

El nostre estudi és consistent amb el resultats d'altres grups que varen veure que els gens de determinats grups funcionals mostraven diferències en la conservació de la seqüència del promotor. Els factors de transcripció i els gens implicats en processos complexos i adaptatius, com el desenvolupament, la comunicació cel·lular, la funció neural o la senyalització mostren una major conservació de la seqüència del promotor (Iwama and Gojobori, 2004; Lee et al., 2005). Efectivament aquests grups funcionals tenen una infrarrepresentació de gens HK segons el nostre estudi. En canvi, els gens implicats en processos bàsics, com el metabolisme o la funció ribosomal, tenen promotors menys conservats (Lee et al., 2005) i el nostre estudi mostra, de forma congruent, que aquest grups funcionals tenen sobrerrepresentació de gens HK. A més a més, les diferències funcionals no expliquen per si soles les diferències de conservació de la seqüència promotora entre gens HK i no-HK, com hem demostrat. De fet el nostre estudi proposa una explicació més directa de les diferències de conservació del promotor entre diferents grups funcionals en base a les diferències entre els gens *housekeeping* i els gens no *housekeeping*.

Utilitzant matrius de PROMO i de TRANSFAC, vàrem identificar una sèrie de factors sobrerrepresentats en els gens HK. Quatre d'ells (Sp1, NRF1, CREB i ATF) mostren alts pics de freqüència a prop del TSS, -200 a -1 *upstream* del TSS, en promotors de ratolí (Bellora et al., 2007a) i també en promotors humans (dades no mostrades), especialment en gens HK. Alguns d'aquest factors (Sp1, USF, HIF-1) són reguladors coneguts de gens HK (Sanchez-Elsner et al., 2002; Ikeda et al., 2002). L'estudi de sobrerrepresentació el vàrem haver de fer a partir de prediccions i no a partir de TFBSs reals, ja que, com he esmentat anteriorment, quan vàrem intentar mapar tots els TFBSs de gens humans i de ratolí recollits de TRANSFAC, només un 2 % dels nostres gens tenien TFBSs. És interessant senyalar que aproximadament el 75 % dels llocs mapats queien sobre regions alineades (que correspon de mitja al 30 % de la longitud de la seqüència). Aquesta observació sembla reforçar la vàlua del *phylogenetic footprinting* per ajudar a la predicció de TFBSs; però no podem descartar que sigui conseqüència d'un altre biaix en les dades de TRANSFAC.

A part de les diferències dels gens HK també vàrem trobar diferències entre gens amb expressió restringida (1-10 teixits) i gens amb expressió intermitja (11-50 teixits). Els primers mostren una major conservació de la seqüència del promotor. Aquest resultat ens va sorprendre inicialment, ja que esperàvem que els gens que s'expressen en pocs teixits tinguessin un promotor més simple. Però aquesta major conservació d'aquest tipus de gens es pot explicar per l'existència de motius repressors en les regions més distals. Resultats del projecte ENCODE mostren que en un 55 % dels gens els elements reguladors negatius es localitzen entre -1000 i -500 pb *upstream* del TSS (Cooper et al., 2006).

La posició dels llocs d'unió de determinats factors de transcripció poden estar funcionalment restringides (Wray et al., 2003; Bellora et al., 2007a). Per exemple, els llocs d'unió del factor CBP (la capsa CCAAT) solen estar a 50-100 pb *upstream* del TSS. En general, poc es coneix de les conseqüències de la posició del lloc d'unió. Un dels projectes en que he col·laborat es centra en estudiar les restriccions posicionals del TFBSs i altres motius reguladors (Bellora et al., 2007a,b). Els resultats obtinguts indiquen que les restriccions posicionals defineixen unes arquitectures de promotor força diferents depenent de l'amplitud d'expressió i el tipus de teixit (Bellora et al., 2007a).

El nostres estudis mostren també que els gens que no tenen illes CpG tenen el promotor més conservats. Els promotors no-CpG corresponen al promotors clàssics que contenen TATA-*box*. Anàlisis recents mostren que aquests gens estan particularment ben conservats entre diferents espècies de mamífers (Carninci et al., 2006).

Respecte a les diferències entre gens HK i no-HK, un aspecte que no hem pogut estudiar en profunditat es la relació amb l'edat del gen. Els gens HK estan enriquits en proteïnes d'origen antic (Freilich et al., 2005). Les proteïnes més antigues evolucionen més lentament i són més llargues que les proteïnes d'origen més recent, probablement per tenir més restriccions funcionals (Alba and Castresana, 2005). En canvi, les regions reguladores de la transcripció dels gens més antics podrien ser més simples, amb menys restriccions, que les dels gens que representen innovacions en els organismes multicel·lulars. Resultats preliminars no confirmen aquesta hipòtesi, però tampoc la descarten totalment.

Un altre treball que queda pendent de realitzar, i que seria força interessant, és ampliar aquest estudi (HK vs. no-HK) a altres espècies, a ser possible a tots els eucariotes que tenen el genoma seqüenciat.

Un resultat interessant d'aquest estudi dels gens ortòlegs d'humà-ratolí és la correlació existent entre la velocitat d'evolució de la seqüència reguladora de la transcripció i la velocitat d'evolució de la seqüència codificant. No és molt forta però significativa en gens no-HK. Aquesta correlació ja havia estat senyalada per altres autors en nematodes (Castillo-Davis et al., 2004) i llevats (Chin et al., 2005). És interessant remarcar que aquesta correlació pràcticament es perd en gens HK. És a dir, en els gens HK la proteïna pot estar molt conservada i, en canvi, la seqüència *upstream* ser molt divergent (que en realitat vol dir que el promotor és més curt, més simple).

Aquesta correlació entre la divergència de la seqüència del promotor i la divergència de la seqüència de la proteïna suggereix un lligam entre la ràpida diversificació d'una proteïna i el seu patró d'expressió. Curiosament aquest lligam es trenca entre gens paràlegs, com hem pogut comprovar quan estudiàrem gens duplicats humans i de ratolí i com ja havien apuntat Castillo-Davis et al. (2004).

Efecte de la duplicació gènica sobre el promotor i l'expressió

S'HAN FET MOLTS ESTUDIS DELS EFECTES DE LA DUPLICACIÓ GÉNICA sobre l'evolució de les seqüències codificants (Lynch and Conery, 2000; Van de Peer et al., 2001; Nembaware et al., 2002; Kondrashov et al., 2002; Castillo-Davis et al., 2004; Cusack and Wolfe, 2007; Mikkelsen et al., 2007) i, més recentment, sobre la divergència dels patrons d'expressió (Gu et al., 2002; Makova and Li, 2003; Huminiecki and Wolfe, 2004; Ganko et al., 2007). Els canvis en l'expressió de gens duplicats ha de ser degut molt sovint a canvis en les seqüències reguladores de la transcripció. En canvi, s'ha estudiat molt poc la relació entre la divergència de la seqüència reguladora i la divergència de l'expressió. És realment estrany, tenint en compte que alguns del models teòrics que s'han proposat per explicar el mecanisme de preservació de les duplicacions gèniques com el model DDC, duplicació-degeneració-complementació (Force et al., 1999), remarca el paper fonamental dels elements reguladors. Segurament les dificultats inherents a l'estudi de les regions reguladores de la transcripció han endarrerit aquest estudi. Per tant, que sapiguem, el nostre treball sobre l'efecte de la duplicació gènica sobre la seqüència del promotor és el primer que aborda aquesta qüestió.

Un del principals resultats de l'estudi és la correlació significativa entre la divergència de la seqüència promotora i la divergència de l'expressió tissular en els gens duplicats i no en els ortòlegs. És interessant remarcar també el fet que, tot i que el temps de divergència mitjà és la meitat del temps de divergència dels ortòlegs, el grau de divergència de la seqüència del promotor és similar al dels ortòlegs. És a dir, la divergència en el promotor és més ràpida entre paràlegs que entre ortòlegs, indicant una relaxació de les restriccions en els promotors dels gens duplicats. Aquests resultats recolzen un model en que el promotor divergeix ràpidament després de la duplicació gènica o en el mateix procés de duplicació (còpies parcials o retrocòpies).

Per a entendre millor l'efecte de la duplicació gènica sobre la divergència vàrem utilitzar una aproximació ja emprada anteriorment per altres autors (Huminiecki and Wolfe, 2004; Shiu et al., 2006) que assegura un temps de divergència constant. Consisteix en comparar gens ortòlegs d'humà-ratolí pertanyents a famílies amb duplicacions específiques de cada llinatge. En la majoria de casos la duplicació només es produeix en un dels llinatges, peró hi ha casos en que hi ha duplicació en ambdós llinatges. Els nostres resultats mostren un increment progressiu de la divergència del promotor a l'augmentar el nombre de duplicacions gèniques. Una tendència similar s'observa respecte a la divergència de la proteïna; però com he dit anteriorment, a diferència dels ortòlegs, no hi ha co-variació entre les dues parts.

Respecte a l'expressió, els nostres resultats mostren també un increment de la divergència de teixits a l'augmentar el nombre de duplicacions gèniques. També s'observa una reducció de l'amplitud d'expressió amb el nombre de duplicacions, d'una manera congruent respecte a resultats obtinguts per altres autors (Huminiecki and Wolfe, 2004; Freilich et al., 2006). Aquest resultats recolzen, en general, la hipòtesi de la subfuncionalització. Però un estudi detallat mostra que hi han casos de canvis d'expressió que indiquen neofuncionalització. De fet, ambdós models (subfuncionalització i neofuncionalització) coexisteixen en moltes de les famílies. Caldria un estudi exhaustiu, família per família, per a esbrinar si primer es produeix subfuncionalització i després neofuncionalització o si ambdues poden produir-se simultàniament. Seria també interessant estudiar, al mateix temps, quins canvis en la seqüència promotora poden explicar la subfuncionalització o neofuncionalització en cada cas.

Així com la tendència d'increment de la divergència de la seqüència del promotor i de la proteïna és molt clara, la de la divergència de teixits, tot i ser significativa, mostra una variància molt gran. Una certa part d'aquesta variància es pot explicar per l'origen de les dades amb que calculem l'expressió: el *microarrays*. Ara bé, crec que la major part d'aquesta variància es deu al fet que els canvis en la seqüència (promotora o codificant) no impliquen forçosament canvis en l'expressió: poden no tenir cap efecte o poden tenir un efecte molt intens sobre l'expressió tissular.

En general, els resultats obtinguts són molt consistent entre ambdós llinatges (primats i rosegadors), però s'observen certes diferències quantitatives interessants. El nombre de duplicats de rosegadors és aproximadament el doble que el nombre de duplicats de primats. S'ha proposat que, com la mida de la població en rosegadors és en general més gran que en primats, el major nombre de gens duplicats en rosegadors respecte als primats és resultat de l'acció de la selecció positiva en la retenció dels gens duplicats, que seria més efectiva en rosegadors donada aquesta major mida de la població (Shiu et al., 2006).

A més el duplicats de ratolí tenen una divergència de seqüència (promotor i proteïna) més alta que els duplicats humans. Una explicació d'això és l'observació general de que en el llinatge de ratolí s'han produït el doble de substitucions neutrals que en el llinatge humà (Waterston et al., 2002).

Finalment, voldria insistir en que seria interessant estudiar d'una forma exhaustiva les famílies gèniques identificades en aquest estudi (sobretot les que tenen 3 o més gens) per acabar d'esbrinar el paper de la subfuncionalització i la neofuncionalització i com es relaciona amb els canvis de possibles elements reguladors en el promotor. Aquest treball podria servir per a identificar TFBSs funcionals determinants de l'expressió tissular.

Part V

CONCLUSIONS

Conclusions

1. Hem millorat les prediccions de TFBSs mitjançant matrius més específiques i *matching* sobre vàries seqüències a la vegada.

2. Millores en les prediccions de TFBSs es veuen limitades pel model de representació de matrius de pes (PWM), per la poca quantitat i qualitat de les dades de TFBSs disponibles i per les dependències contextuals del propi sistema de regulació transcripcional.

3. Els gens *housekeeping* tendeixen a tenir un promotor menys conservat (o més curt) respecte als gens que tenen una expressió més restringida.

4. Les diferències funcionals en les proteïnes no expliquen per si soles les diferències de conservació de la seqüència promotora entre gens HK i no-HK.

5. Hem identificat un petit conjunt de TFs amb llocs d'unió sobrerrepresentats en els gen HK i alguns d'ells són reguladors coneguts dels gens HK.

6. Els gens amb expressió restringida mostren una major conservació de la seqüència del promotor que els gens amb expressió intermitja.

7. Existeix correlacció entre la velocitat d'evolució de la seqüència reguladora de la transcripció i la velocitat d'evolució de la seqüència codificant en ortòlegs, però no en paràlegs.

8. Existeix correlació entre la divergència de la seqüència promotora i la divergència de l'expressió tissular en els gens duplicats i no en els ortòlegs.

9. Hi ha un increment progressiu de la divergència de la seqüència del promotor i de la proteïna a l'augmentar el nombre de duplicacions gèniques.

10. Es dona també un increment de la divergència de teixits i una reducció de l'amplitud d'expressió a l'augmentar el nombre de duplicacions gèniques.

11. Hi ha el doble de gens duplicats i amb major divergència en el llinatge de ratolí que en el d'humans.

Part VI

BIBLIOGRAFIA

Bibliografia

M. M. Alba and J. Castresana. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol*, 22(3):598–606, 2005.

S. Aparicio, A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, and S. Brenner. Detecting conserved regulatory elements with the model genome of the japanese puffer fish, fugu rubripes. *Proc Natl Acad Sci U S A*, 92(5):1684–8, 1995.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with meme. *Proc Int Conf Intell Syst Mol Biol*, 3:21–9, 1995.

V. B. Bajic, S. L. Tan, A. Chong, S. Tang, A. Strom, J. A. Gustafsson, C. Y. Lin, and E. T. Liu. Dragon ere finder version 2: A tool for accurate detection and analysis of estrogen response elements in vertebrate genomes. *Nucleic Acids Res*, 31(13):3605–7, 2003.

N. Bellora, D. Farre, and M. M. Alba. Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics*, 8:459, 2007a.

N. Bellora, D. Farre, and M. Mar Alba. Peaks: identification of regulatory motifs by their position in dna sequences. *Bioinformatics*, 23(2):243–4, 2007b.

P. V. Benos, A. S. Lapedes, and G. D. Stormo. Is there a code for protein-dna recognition? probab(ilistical)ly. *Bioessays*, 24(5):466–75, 2002.

E. Blanco, D. Farre, M. M. Alba, X. Messeguer, and R. Guigo. Abs: a database of annotated regulatory binding sites from orthologous promoters. *Nucleic Acids Res*, 34(Database issue):D63–7, 2006.

A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–15, 1998.

S. E. Brenner, T. Hubbard, A. Murzin, and C. Chothia. Gene duplications in h. influenzae. *Nature*, 378(6553):140, 1995.

P. Bucher. Regulatory elements and expression profiles. *Curr Opin Struct Biol*, 9(3):400–7, 1999.

J. Buhler and M. Tompa. Finding motifs using random projections. *J Comput Biol*, 9(2):225–42, 2002.

M. L. Bulyk, P. L. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–61, 2002.

M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Res*, 14(2):201–8, 2004.

P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–35, 2006.

S. B. Carroll. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101(6):577–80, 2000.

C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. Selection for short introns in highly expressed genes. *Nat Genet*, 31(4):415–8, 2002.

C. I. Castillo-Davis, D. L. Hartl, and G. Achaz. cis-regulatory and protein evolution in orthologous and duplicate genes. *Genome Res*, 14(8):1530–6, 2004.

Q. K. Chen, G. Z. Hertz, and G. D. Stormo. Matrix search 1.0: a computer program that scans dna sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci*, 11(5):563–6, 1995.

C. S. Chin, J. H. Chuang, and H. Li. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res*, 15(2):205–13, 2005.

S. J. Cooper, N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res*, 16(1):1–10, 2006.

B. P. Cusack and K. H. Wolfe. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol*, 24(3):679–86, 2007.

E. H. Davidson. *Genomic regulatory systems: development and evolution.* Academic Press, San Diego, 2001.

E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, M. J. Schilstra, P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–78, 2002.

E. T. Dermitzakis and A. G. Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, 19(7):1114–21, 2002.

E. Eisenberg and E. Y. Levanon. Human housekeeping genes are compact. *Trends Genet*, 19(7):362–5, 2003.

D. Farre, R. Roset, M. Huerta, J. E. Adsuara, L. Rosello, M. M. Alba, and X. Messeguer. Identification of patterns in biological sequences at the alggen server: Promo and malgen. *Nucleic Acids Res*, 31(13):3651–3, 2003.

D. Farre, N. Bellora, L. Mularoni, X. Messeguer, and M. M. Alba. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol*, 8(7):R140, 2007.

A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov, and V. J. Makeev. A gibbs sampler for identification of symmetrically structured, spaced dna motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–5, 2005.

G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, A. M. Craven, H. B. Harlow, E. W. Su, J. E. Onyia, and C. Su. A statistical analysis of the transfac database. *Biosystems*, 81(2):137–54, 2005.

A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, 1999.

K. Frech, P. Dietze, and T. Werner. Consinspector 3.0: new library and enhanced functionality. *Comput Appl Biosci*, 13(1):109–10, 1997a.

K. Frech, K. Quandt, and T. Werner. Finding protein-binding sites in dna sequences: the next generation. *Trends Biochem Sci*, 22(3):103–4, 1997b.

S. Freilich, T. Massingham, S. Bhattacharyya, H. Ponsting, P. A. Lyons, T. C. Freeman, and J. M. Thornton. Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol*, 6(7):R56, 2005.

S. Freilich, T. Massingham, E. Blanc, L. Goldovsky, and J. M. Thornton. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol*, 7(10):R89, 2006.

E. W. Ganko, B. C. Meyers, and T. J. Vision. Divergence in expression between duplicated genes in arabidopsis. *Mol Biol Evol*, 24(10):2298–309, 2007.

D. Ghosh. Object-oriented transcription factors database (ootfd). *Nucleic Acids Res*, 28(1):308–10, 2000.

J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol*, 313(4):903–19, 2001.

O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, and S. J. Jones. Oreganno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, 36(Database issue):D107–13, 2008.

C. J. Gruber, D. M. Gruber, I. M. Gruber, F. Wieser, and J. C. Huber. Anatomy of the estrogen response element. *Trends Endocrinol Metab*, 15(2):73–8, 2004.

Z. Gu, D. Nicolae, H. H. Lu, and W. H. Li. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet*, 18(12):609–13, 2002.

R. C. Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*, 16(9): 369–72, 2000.

R. Harr, M. Haggstrom, and P. Gustafsson. Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Res*, 11(9):2943–57, 1983.

X. He and J. Zhang. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 169(2):1157–64, 2005.

G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999.

J. M. Heumann, A. S. Lapedes, and G. D. Stormo. Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc Int Conf Intell Syst Mol Biol*, 2:188–94, 1994.

K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. Plant cis-acting regulatory dna elements (place) database: 1999. *Nucleic Acids Res*, 27(1):297–300, 1999.

A. L. Hughes. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*, 256(1346):119–24, 1994.

L. Huminiecki and K. H. Wolfe. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res*, 14(10A):1870–9, 2004.

S. Ikeda, H. Ayabe, K. Mori, Y. Seki, and S. Seki. Identification of the functional elements in the bidirectional promoter of the mouse o-sialoglycoprotein endopeptidase and apex nuclease genes. *Biochem Biophys Res Commun*, 296(4): 785–91, 2002.

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

H. Iwama and T. Gojobori. Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proc Natl Acad Sci U S A*, 101(49):17156–61, 2004.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–502, 2007.

J. T. Kadonaga. Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92(3):307–13, 1998.

P. D. Keightley, M. J. Lercher, and A. Eyre-Walker. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol*, 3(2):e42, 2005.

A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. Match: A tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res*, 31(13):3576–9, 2003.

O. V. Kel-Margoulis, A. E. Kel, I. Reuter, I. V. Deineko, and E. Wingender. Transcompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res*, 30(1):332–4, 2002.

O. V. Kel-Margoulis, D. Tchekmenev, A. E. Kel, E. Goessling, K. Hornischer, B. Lewicki-Potapov, and E. Wingender. Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol*, 3(1-2):145–71, 2003.

T. H. Kim and B. Ren. Genome-wide analysis of protein-dna interactions. *Annu Rev Genomics Hum Genet*, 2006.

N. A. Kolchanov, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, M. A. Pozdnyakov, N. L. Podkolodny, A. N. Naumochkin, and A. G. Romashchenko. Transcription regulatory regions database (trrd): its status in 2002. *Nucleic Acids Res*, 30(1):312–7, 2002.

F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Selection in the evolution of gene duplications. *Genome Biol*, 3(2):RESEARCH0008, 2002.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

S. Lee, I. Kohane, and S. Kasif. Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics*, 6:168, 2005.

B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

M. Lescot, P. Dehais, G. Thijs, K. Marchal, Y. Moreau, Y. Van de Peer, P. Rouze, and S. Rombauts. Plantcare, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*, 30(1):325–7, 2002.

W. H. Li, Z. Gu, H. Wang, and A. Nekrutenko. Evolutionary analyses of the human genome. *Nature*, 409(6822): 847–9, 2001.

H. Lodish, A. B. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. T. Matsudaira. *Molecular Cell Biology*. W. H. Freeman, 6th edition, 2007.

G. G. Loots, R. M. Locksley, C. M. Blankespoor, Z. E. Wang, W. Miller, E. M. Rubin, and K. A. Frazer. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463): 136–40, 2000.

N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-dna interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–74, 2001.

M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–5, 2000.

M. Lynch and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1): 459–73, 2000.

K. D. Makova and W. H. Li. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res*, 13(7):1638–45, 2003.

Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-dna binding sites. *Nucleic Acids Res*, 26(10):2306–12, 1998.

Y. Mandel-Gutfreund, O. Schueler, and H. Margalit. Comprehensive analysis of hydrogen bonds in regulatory protein dna-complexes: in search of common principles. *J Mol Biol*, 253(2):370–82, 1995.

V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31 (1):374–8, 2003.

J. M. McClintock, M. A. Kheirbek, and V. E. Prince. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. *Development*, 129(10):2339–54, 2002.

L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, 29(3):774–82, 2001.

X. Messeguer, R. Escudero, D. Farre, O. Nunez, J. Martinez, and M. M. Alba. Promo: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, 18(2):333–4, 2002.

A. Meyer and M. Schartl. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr Opin Cell Biol*, 11(6):699–704, 1999.

T. S. Mikkelsen, M. J. Wakefield, B. Aken, C. T. Amemiya, J. L. Chang, S. Duke, M. Garber, A. J. Gentles, L. Goodstadt, A. Heger, J. Jurka, M. Kamal, E. Mauceli, S. M. Searle, T. Sharpe, M. L. Baker, M. A. Batzer, P. V. Benos, K. Belov, M. Clamp, A. Cook, J. Cuff, R. Das, L. Davidow, J. E. Deakin, M. J. Fazzari, J. L. Glass, M. Grabherr, J. M. Greally, W. Gu, T. A. Hore, G. A. Huttley, M. Kleber, R. L. Jirtle, E. Koina, J. T. Lee, S. Mahony, M. A. Marra, R. D. Miller, R. D. Nicholls, M. Oda, A. T. Papenfuss, Z. E. Parra, D. D. Pollock, D. A. Ray, J. E. Schein, T. P. Speed, K. Thompson, J. L. VandeBerg, C. M. Wade, J. A. Walker, P. D. Waters, C. Webber, J. R. Weidman, X. Xie, M. C. Zody, J. A. Graves, C. P. Ponting, M. Breen, P. B. Samollow, E. S. Lander, and K. Lindblad-Toh. Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167–77, 2007.

S. B. Montgomery, O. L. Griffith, M. C. Sleumer, C. M. Bergman, M. Bilenky, E. D. Pleasance, Y. Prychyna, X. Zhang, and S. J. Jones. Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22(5):637–40, 2006.

J. H. Nadeau and D. Sankoff. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147(3):1259–66, 1997.

V. Nembaware, K. Crum, J. Kelso, and C. Seoighe. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res*, 12(9):1370–6, 2002.

M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, 2003.

D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6):730–2, 2007.

S Ohno. *Evolution by Gene Duplication*. Springer-Verlag, New York, 1970.

R. Osada, E. Zaslavsky, and M. Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–25, 2004.

C. O. Pabo and R. T. Sauer. Protein-dna recognition. *Annu Rev Biochem*, 53:293–321, 1984.

G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17 Suppl 1:S207–14, 2001.

G. Pesole, S. Liuni, and M. D'Souza. Patsearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 16(5):439–50, 2000.

E. Portales-Casamar, S. Kirov, J. Lim, S. Lithwick, M. I. Swanson, A. Ticoll, J. Snoddy, and W. W. Wasserman. Pazar: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*, 8(10):R207, 2007.

J. H. Postlethwait, I. G. Woods, P. Ngo-Hazelett, Y. L. Yan, P. D. Kelly, F. Chu, H. Huang, A. Hill-Force, and W. S. Talbot. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res*, 10(12):1890–902, 2000.

D. S. Prestridge. Signal scan 4.0: additional databases and sequence formats. *Comput Appl Biosci*, 12(2):157–60, 1996.

V. E. Prince and F. B. Pickett. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, 3(11):827–37, 2002.

K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. Matind and matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23(23):4878–84, 1995.

J. Raes, K. Vandepoele, C. Simillion, Y. Saeys, and Y. Van de Peer. Investigating ancient duplication events in the arabidopsis genome. *J Struct Funct Genomics*, 3(1-4):117–29, 2003.

S. Rastogi and D. A. Liberles. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol*, 5(1):28, 2005.

X. Y. Ren, O. Vorst, M. W. Fiers, W. J. Stiekema, and J. P. Nap. In plants, highly expressed genes are the least compact. *Trends Genet*, 22(10):528–32, 2006.

P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends Genet*, 16 (6):276–7, 2000.

G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–7, 2007.

H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides. Regulondb (version 3.2): transcriptional regulation and operon organization in escherichia coli k-12. *Nucleic Acids Res*, 29(1):72–4, 2001.

T. Sanchez-Elsner, L. M. Botella, B. Velasco, C. Langa, and C. Bernabeu. Endoglin expression is regulated by transcriptional cooperation between the hypoxia and transforming growth factor-beta pathways. *J Biol Chem*, 277 (46):43799–808, 2002.

J. Schug and G. C. Overton. Tess: Transcription element search software on the www. Technical Report Technical Report CBIL-TR-1997-1001-v0.0., Computational Biology and Informatics - LaboratorySchool of Medicine - University of Pennsylvania, 1997.

N. C. Seeman, J. M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A*, 73(3):804–8, 1976.

S. H. Shiu, J. K. Byrnes, R. Pan, P. Zhang, and W. H. Li. Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci U S A*, 103(7):2232–6, 2006.

A. Sidow. Gen(om)e duplications in the evolution of early vertebrates. *Curr Opin Genet Dev*, 6(6):715–22, 1996.

M. B. Soares, E. Schon, A. Henderson, S. K. Karathanasis, R. Cate, S. Zeitlin, J. Chirgwin, and A. Efstratiadis. Rna-mediated gene duplication: the rat preproinsulin i gene is a functional retroposon. *Mol Cell Biol*, 5(8):2090–103, 1985.

M. J. Solomon and A. Varshavsky. Formaldehyde-mediated dna-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A*, 82(19):6470–4, 1985.

M. J. Solomon, P. L. Larsen, and A. Varshavsky. Mapping protein-dna interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–47, 1988.

G. D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

G. D. Stormo. Consensus patterns in dna. *Methods Enzymol*, 183:211–21, 1990.

G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-dna interactions. *Trends Biochem Sci*, 23(3):109–13, 1998.

G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the ’perceptron’ algorithm to distinguish translational initiation sites in e. coli. *Nucleic Acids Res*, 10(9):2997–3011, 1982.

A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–7, 2004.

Y. Suzuki, R. Yamashita, M. Shirota, Y. Sakakibara, J. Chiba, J. Mizushima-Sugano, K. Nakai, and S. Sugano. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res*, 14(9):1711–8, 2004.

D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55, 1988.

S. A. Teichmann, J. Park, and C. Chothia. Structural assignments to the mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci U S A*, 95(25):14658–63, 1998.

M. Tompa. Identifying functional elements by comparative dna sequence analysis. *Genome Res*, 11(7):1143–4, 2001.

N. D. Trinklein, S. J. Aldred, A. J. Saldanha, and R. M. Myers. Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13(2):308–12, 2003.

Y. Van de Peer, J. S. Taylor, I. Braasch, and A. Meyer. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol*, 53(4-5):436–46, 2001.

Y. Van de Peer, J. S. Taylor, and A. Meyer. Are all fishes ancient polyploids? *J Struct Funct Genomics*, 3(1-4):65–73, 2003.

A. E. Vinogradov. Genome design model: evidence from conserved intronic sequence in human-mouse comparison. *Genome Res*, 16(3):347–54, 2006.

T. Waleev, D. Shtokalo, T. Konovalova, N. Voss, E. Cheremushkin, P. Stegmaier, O. Kel-Margoulis, E. Wingender, and A. Kel. Composite module analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res*, 34(Web Server issue):W541–5, 2006.

R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson,

M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

T. Werner. Target gene identification from expression array data by promoter analysis. *Biomol Eng*, 17(3):87–94, 2001.

K. H. Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet*, 2(5):333–41, 2001.

A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7, 2005.

C. T. Workman and G. D. Stormo. Ann-spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, pages 467–78, 2000.

G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–419, 2003.

J. Yang, A. I. Su, and W. H. Li. Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Mol Biol Evol*, 22(10):2113–8, 2005.

C. H. Yuh, H. Bolouri, and E. H. Davidson. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, 128(5):617–29, 2001.

W. Zhang, Q. D. Morris, R. Chang, O. Shai, M. A. Bakowski, N. Mitsakakis, N. Mohammad, M. D. Robinson, R. Zirngibl, E. Somogyi, N. Laurin, E. Eftekharpour, E. Sat, J. Grigull, Q. Pan, W. T. Peng, N. Krogan, J. Greenblatt, M. Fehlings, D. van der Kooy, J. Aubin, B. G. Bruneau, J. Rossant, B. J. Blencowe, B. J. Frey, and T. R. Hughes. The functional landscape of mouse gene expression. *J Biol*, 3(5):21, 2004.

Q. Zhou and J. S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–16, 2004.

J. Zhu and M. Q. Zhang. Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7-8): 607–11, 1999.

Z. Zhu, J. Shendure, and G. M. Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res*, 15(6):848–55, 2005.

Part VII

APÈNDIXS

Apèndix 1

ABS: una base de dades de TFBS de promotors d'ortòlegs

Resum

Aquest apèndix presenta una base de dades de TFBSs (ABS), ideada per Enrique Blanco, en la creació de la qual vaig participar recollint dades de motius reguladors, sobretot a partir de la literatura, per a la seva inclusió a ABS.

ABS intenta posar fàcilment accessibles dades de TFBSs experimentals d'alta qualitat per ajudar al disseny, l'avaluació i la millora dels sistemes computacionals per identificar TFBSs en seqüències de promotors relacionats.

Aquest treball es veu reflectit en la publicació:

Blanco E, Farré D, Albà MM, Messeguer X, Guigó R: **ABS: a database of Annotated regulatory Binding Sites from orthologous promoters**. Nucleic Acids Res 2006, 34:D63-67.

ABS és una base de dades d'accés públic via *web* que conté TFBSs identificats en promotors de gens ortòlegs de vertebrats, verificats a partir de la bibliografia. Vàrem anotar 650 TFBSs experimentals corresponents a 68 TFs i 100 gens ortòlegs humans, de ratolí, de rata o de pollastre. Es van anotar també alineaments de les seqüències dels promotors i prediccions realitzades. A més a més, es va afegir un generador de conjunts de dades artificials, a partir dels TFBSs recollits a la base de dades, i una eina d'anàlisi per a ajudar a l'entrenament i l'avaluació dels motius trobats.

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D63–D67***
doi:10.1093/nar/gkj116

ABS: a database of Annotated regulatory Binding Sites from orthologous promoters

Enrique Blanco[1,2,*], Domènec Farré[1,2], M. Mar Albà[1], Xavier Messeguer[2] and Roderic Guigó[1]

[1] Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra/Centre de Regulació Genòmica, C/Doctor Aiguader 80, 08003 Barcelona, Spain and
[2] Grup d'algorísmica i genètica, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, C/Jordi Girona 1-3, 08034 Barcelona, Spain

Received August 1, 2005; Revised September 19, 2005; Accepted October 18, 2005

ABSTRACT

Information about the genomic coordinates and the sequence of experimentally identified transcription factor binding sites is found scattered under a variety of diverse formats. The availability of standard collections of such high-quality data is important to design, evaluate and improve novel computational approaches to identify binding motifs on promoter sequences from related genes. ABS (http://genome.imim.es/datasets/abs2005/index.html) is a public database of known binding sites identified in promoters of orthologous vertebrate genes that have been manually curated from bibliography. We have annotated 650 experimental binding sites from 68 transcription factors and 100 orthologous target genes in human, mouse, rat or chicken genome sequences. Computational predictions and promoter alignment information are also provided for each entry. A simple and easy-to-use web interface facilitates data retrieval allowing different views of the information. In addition, the release 1.0 of ABS includes a customizable generator of artificial datasets based on the known sites contained in the collection and an evaluation tool to aid during the training and the assessment of motif-finding programs.

INTRODUCTION

Expression of genes is regulated at many different levels, transcription of DNA being one of the most critical stages. Specific configurations of transcription factors (TFs) that interact with gene promoter regions are recruited to activate or modulate the production of a given transcript. Many of these TFs possess the ability to recognize a small set of genomic sequence footprints called TF-binding sites (TFBSs). These motifs are typically 6–15 bp long and in some cases, they show a high degree of variability. In addition, many motifs may ambiguously be recognized by members of different TF families. Because of these flexible binding rules, computational methods for the identification of regulatory elements in a promoter sequence tend to produce an overwhelming amount of false positives. However, the identification of conserved regulatory elements present in orthologous gene promoters (also called phylogenetic footprinting) has proved to be more effective to characterize such sequences (1–3). In fact, the ever-growing availability of more genomes and the constant improvement of bioinformatics algorithms hold great promise for unveiling the overall network of gene interactions of each organism (4).

Typically, computational methods to detect regulatory elements use their own training set of experimental annotated TFBSs. These annotations are usually collected from bibliography or from general repositories of gene regulation information, such as JASPAR (5) and TRANSFAC (6). However, each program establishes different criteria and formats to retrieve and display the data that forms the final training set, which makes the comparison between different methods very difficult. The construction of a good benchmark to evaluate the accuracy of several pattern discovery methods is therefore not a trivial procedure (7).

Although important efforts are being carried out to standardize the construction of collections of promoter regions (8) or the presentation of experimental data (9), there is a clear necessity to provide stable and common datasets for future algorithmic developments. In this direction, we present here the release 1.0 of the ABS database constructed from literature annotations that have been experimentally verified in human, mouse, rat or chicken.

DATABASE CONSTRUCTION

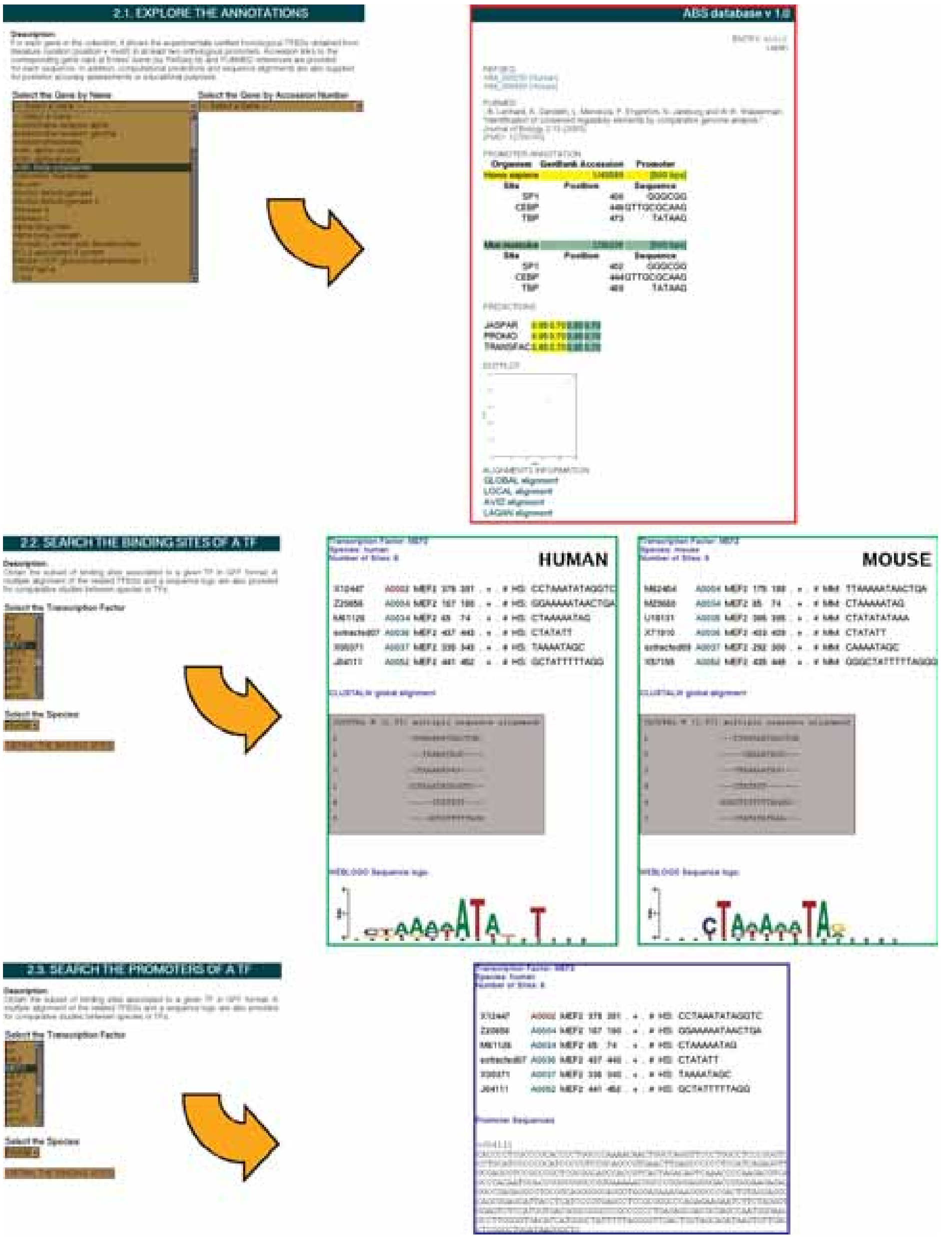
We have gathered from the literature a collection of experimentally validated binding sites that are conserved in at least

*To whom correspondence should be addressed. Tel: +34 93 2240891; Fax: +34 93 2240875; Email: eblanco@imim.es two orthologous vertebrate promoters. The sites and the promoter sequences have been manually curated to ensure data consistency. The compiled data are suitable for training both classical pattern discovery programs and new emerging comparative methods. Flat files accomplishing the GFF standard format were used to store and query the information.

The GenBank accession number of the sequences in each bibliographical reference was utilized to retrieve the promoters. Such sequences were mapped on to the corresponding RefSeq annotations to ensure we were retrieving the actual promoter. The DBTSS database (10) was finally used to refine the annotation of the TSSs. Since it is the region in which most experimental studies have been focused on, we considered the sequence 500 bp immediately upstream the annotated TSS, as the promoter region in this first release.

For each annotated promoter, we only included experimentally tested sites in this proximal region whose motifs were correctly identified in at least two species, i.e. orthologous sites. Every known binding site was mapped on to the corresponding promoter sequence by BLASTN (11). Those matches that exhibited <80% of identity between the sequence of the original site and the mapped motif in the promoter region were rejected.

We computed BLASTN (11), CLUSTALW (12), AVID (13) and LAGAN (14) alignments of the orthologous promoters



```
X12447      A0002 MEF2 378 391 . + . # HS: CCTAAATATAGGTC
Z20856      A0004 MEF2 167 180 . + . # HS: GGAAAAATAACTGA
M61126      A0034 MEF2 65  74  . + . # HS: CTAAAAATAG
extracted07 A0036 MEF2 437 443 . + . # HS: CTATATT
X00371      A0037 MEF2 339 340 . + . # HS: TAAAATAGC
J04111      A0052 MEF2 441 452 . + . # HS: GCTATTTTTAGG
```

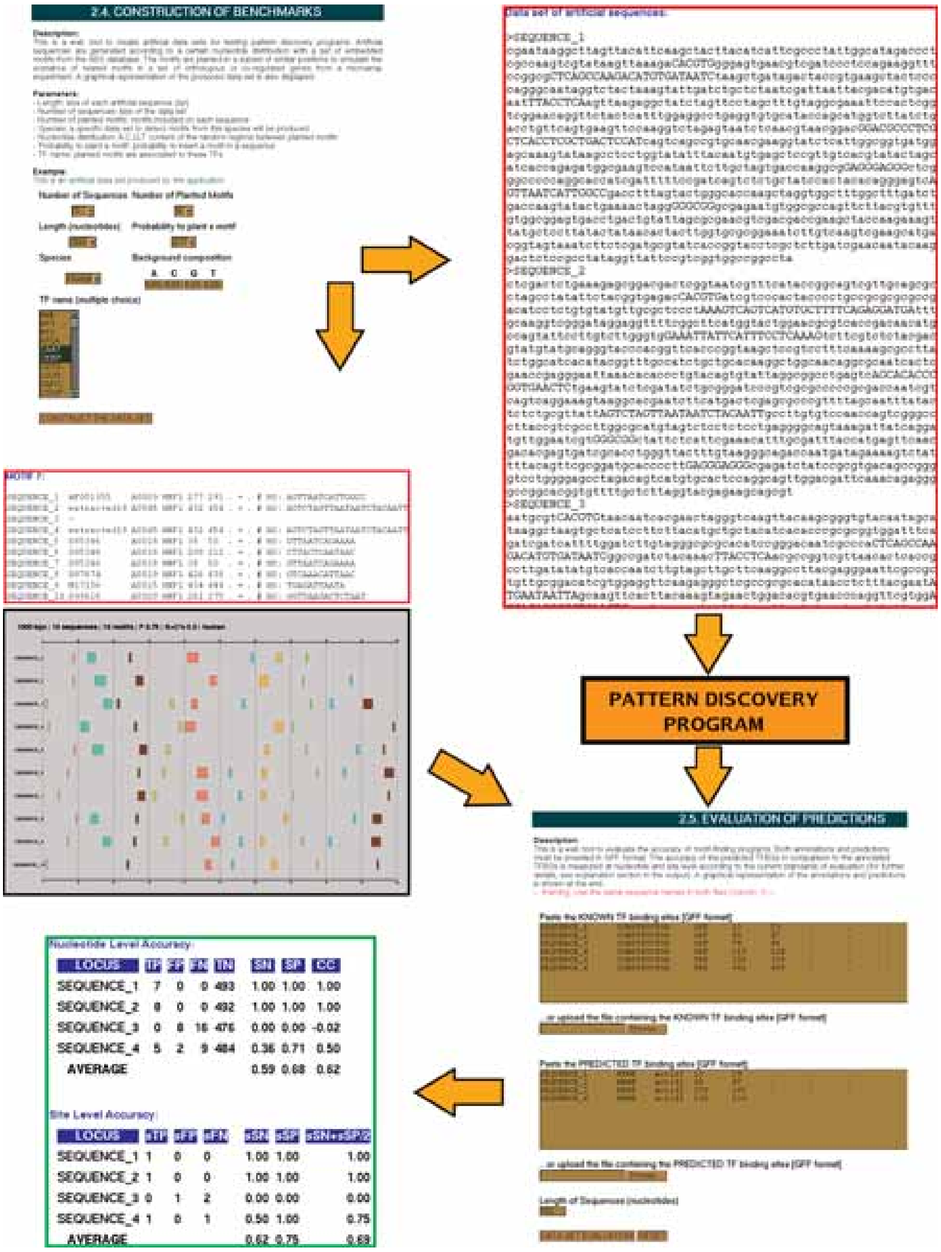
```
M62464      A0004 MEF2 175 188 . + . # MM: TTAAAAATAACTGA
M25660      A0034 MEF2 65  74  . + . # MM: CTAAAAATAG
U18131      A0035 MEF2 386 395 . + . # MM: CTATATATAAA
X71910      A0036 MEF2 423 429 . + . # MM: CTATATT
extracted09 A0037 MEF2 292 300 . + . # MM: CAAAATAGC
X57158      A0052 MEF2 435 448 . + . # MM: GGGCTATTTTTAGGG
```

Figure 1. Examples of the ABS data retrieval system showing the annotation of a gene, the set of binding motifs from a given TF in human and mouse and the extraction of the promoter sequences containing such annotations.

from each gene. Moreover, we produced a dotplot of word matches with EMBOSS (15) to visualize unusually conserved regions. For comparative assessments, computational predictions using the JASPAR (5), TRANSFAC (6) and PROMO (16) collections of position weight matrices were calculated. A very restrictive threshold of 0.85 was used to remove those predicted TFBSs whose score was below this value, creating a first group of more reliable predictions. A second group of predictions was produced using a more flexible threshold of 0.70 (see the ABS website for further information about the scoring method).

DATABASE CONTENTS

Release 1.0 of ABS database contains 100 annotated orthologous genes, each entry corresponding to two or more species. The total number of promoter sequences is 211 (105 500 nt). There is a clear predominance of human and mouse annotations: 73 entries contain at least annotations for human–mouse orthologs. A total of 650 experimental binding sites from 68 different TFs are associated with ABS entries, covering 8624 nt. In average, three TFBSs per sequence have been mapped on to the promoters with an average length of 13.3 nt per site. The majority of the TFBSs are found near


2.4. CONSTRUCTION OF BENCHMARKS
Data set of artificial sequences:
>SEQUENCE_1
>SEQUENCE_2
>SEQUENCE_3
MOTIF F:
PATTERN DISCOVERY PROGRAM
2.5. EVALUATION OF PREDICTIONS

Nucleotide Level Accuracy:

LOCUS	TP	FP	FN	TN	SN	SP	CC
SEQUENCE_1	7	0	0	493	1.00	1.00	1.00
SEQUENCE_2	8	0	0	492	1.00	1.00	1.00
SEQUENCE_3	0	8	16	476	0.00	0.00	-0.02
SEQUENCE_4	5	2	9	484	0.36	0.71	0.50
AVERAGE					0.59	0.68	0.62

Site Level Accuracy:

LOCUS	sTP	sFP	sFN	sSN	sSP	sSN+sSP/2
SEQUENCE_1	1	0	0	1.00	1.00	1.00
SEQUENCE_2	1	0	0	1.00	1.00	1.00
SEQUENCE_3	0	1	2	0.00	0.00	0.00
SEQUENCE_4	1	0	1	0.50	1.00	0.75
AVERAGE				0.62	0.75	0.69

Figure 2. Protocol to evaluate the accuracy of an external motif-finding program on a synthetic dataset generated by planting motifs from ABS in randomly generated sequences.

the TSS, as expected. The TFs that appear more frequently are TBP (14.6% of sites), SP1 (13.6%) and CEBP (5.6%). Those TFs are known to be part of the core of many eukaryotic promoters (see the ABS documentation for further details about the contents of the database).

WEB INTERFACE

Data retrieval

The ABS database can be accessed through a simple CGI/Perl-based web interface at http://genome.imim.es/datasets/abs2005/index.html. On-line documentation and tutorials are provided for each web service. The following functionalities are implemented in the current release (see Figure 1):

(i) For each gene in the collection, show the orthologous promoters and a list of experimentally verified TFBSs annotated on the corresponding sequence. Promoter sequence alignments, computational predictions, dotplots and cross-references to other well-known databases, such as GenBank, Entrez Gene and PubMed, are also provided for each annotation.

(ii) Retrieve all of the binding motifs associated with a given TF, filtering by species. Moreover, a global alignment of the motifs is provided and the corresponding sequence logo representation is displayed by using WebLogo (17). This information could be used to produce new profiles for subsequent detection of this TF in other promoters.

(iii) Retrieve all of the promoter sequences in which at least one binding site for a given TF was annotated. These sequences and the associated motifs could be used to generate datasets based on known sites to train motif-finding programs.

(iv) The gene catalogue, the promoter sequences, the collection of annotations, the sequence alignments and the computational predictions are also individually distributed in several flat files.

Benchmarking and evaluation tools

The ABS database aims to become a platform to evaluate new algorithms for the discovery of novel regulatory elements in a set of related gene promoters (e.g. orthologous promoters or co-regulated genes from microarray experiments). In addition to the data retrieval functions, two on-line applications are available to perform the benchmarking of such algorithms (see Figure 2):

(i) Constructor is a web server to produce synthetic datasets based on the ABS annotations. The design of the benchmark is highly flexible allowing to customize the number of sequences, their length, the background nucleotide distribution, the number of motifs to plant on them, the probability to plant a motif, the species and the TFs for which the associated motifs will be randomly selected from the known sites collection. The output consists of the artificial sequences with the embedded motifs, the list of motifs and a graphical representation of the occurrences in the sequences produced with the program gff2ps (18).

(ii) Evaluator is a web server to determine the accuracy of a set of predicted motifs in several sequences using a list of known binding sites as a reference set. Both sets must be provided by the user in GFF format. A complete accuracy assessment at both nucleotide and site levels is computed using the standard measures in the field (7,19).

CONCLUSIONS AND FUTURE WORK

The ABS database has been developed to fill the existing gap in the availability of consistent datasets to train and compare different pattern discovery programs. The lack of standard collections of TFBSs is specially serious in the case of phylogenetic footprinting data. The collection described here contains 650 experimental TFBSs identified in human, mouse, rat and chicken genes. Orthologous promoter sequences and their binding sites have been manually curated from bibliography. Supplementary information about the promoters is also provided for each entry. In addition, two web applications (Constructor and Evaluator) are included in this first release to facilitate the development of new motif-finding programs using the ABS annotations. In the next release, we plan to increase the number of annotations adding known sites in regulatory regions different from the proximal promoter and eventually incorporate binding motifs from other species.

ACKNOWLEDGEMENTS

E.B. is recipient of a predoctoral fellowship from Ministerio de Ciencia y Tecnología (Spain). D.F. is recipient of a predoctoral fellowship from Instituto Nacional de Bioinformática (Spain). This work has been supported by grants from Plan Nacional de I+D (BIO2000-1358-C02-02 and BIO2002-04426-C02-01), Ministerio de Ciencia y Tecnología (Spain) and from a FBBVA Bioinformatics grant. Funding to pay the Open Access publication charges for this article was provided by Ministerio de Cienciá y Tecnología (Spain).

Conflict of interest statement. None declared.

REFERENCES

1. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
2. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W., Chiaromonte,F. *et al.* (2003) Distinguishing regulatory DNA from neutral sites. *Genome Res.*, **13**, 64–72.
3. Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.*, **5**, 276–287.
5. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
6. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
7. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
8. Barta,E., Sebestyen,E., Palfy,T.B., Toth,G., Ortutay,C.P. and Patthy,L. (2005) DoOP: Databases of Orthologous Promoters, collections of

clusters of orthologous upstream sequences from chordates and plants. *Nucleic Acids Res.*, **33**, D86–D90.

9. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res.*, **33**, D103–D107.
10. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
13. Bray,N., Dubchak,I. and Patcher,L. (2003) AVID: a Global Alignment Program. *Genome Res.*, **13**, 97–102.
14. Brudno,M., Do,C., Cooper,G., Kim,M.F., Davydov,E., Green,E.D., Sidow,A. and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignments of genomic DNA. *Genome Res.*, **13**, 721–731.
15. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
16. Farré,D., Roset,R., Huerta,M., Adsuara,J.E., Rosello,L., Albà,M.M. and Messeguer,X. (2003) Identification of patterns in biological sequences at the ALGEN server: PROMO and MALGEN. *Nucleic Acids Res.*, **31**, 3651–3653.
17. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
18. Abril,J.F. and Guigó,R. (2000) gff2ps: visualizing genomic annotations. *Bioinformatics*, **16**, 743–744.
19. Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Apèndix 2

PEAKS: predicció de motius reguladors amb biaix posicional en promotors gènics

Resum

Aquest apèndix presenta PEAKS, un programa accessible via *web* per fer prediccions de motius de DNA (per exemple, TFBSs) que mostren un biaix posicional respecte a un element de referència (per exemple, el TSS). PEAKS ha estat desenvolupat al grup *Evolutionary Genomics* (GRIB-IMIM/UPF/CRG) per Nicolás Bellora i jo he participat en la preparació de les llibreries de motius coneguts que utilitza el programa i en les anàlisis realitzades utilitzant aquest programa.

Aquest treball es veu reflectit en les publicacions:

Bellora N, Farré D, Mar Albà M: **PEAKS: identification of regulatory motifs by their position in DNA sequences**. Bioinformatics 2007, 23:243-244.

Bellora N, Farré D, Albà MM: **Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters**. BMC Genomics 2007, 8:459.

S'ha utilitzat PEAKS per fer prediccions sistemàtiques de motius que mostren un biaix posicional significatiu dins del promotor en una extensa col·lecció de gens *housekeeping* i gens específics de teixit de ratolí. Els resultats mostren que els promotors dels gens *housekeeping* estan enriquits en determinats motius que tenen un fort biaix posicional, com YY1; aquest motius són en canvi poc rellevants en gens que s'expressen només en uns determinats teixits. Vàrem identificar també un ampli nombre de motius (559) amb biaix posicional en gens amb una expressió tissular altament específica, alguns força coneguts (HNF1, HNF4, RFX) i altres nous.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 2 2007, pages 243–244
doi:10.1093/bioinformatics/btl568

Genome analysis

PEAKS: identification of regulatory motifs by their position in DNA sequences

Nicolás Bellora[1], Domènec Farré[2] and M. Mar Albà[1,3,*]

[1]Research Unit on Biomedical Informatics, Universitat Pompeu Fabra, [2]Centre for Genomic Regulation and [3]Catalan Institution for Research and Advanced Studies—Municipal Institute of Medical Research, Barcelona 08003, Spain

Received on July 4, 2006; revised on October 18, 2006; accepted on November 5, 2006
Advance Access publication November 10, 2006
Associate Editor: David Rocke

ABSTRACT

Summary: Many DNA functional motifs tend to accumulate or cluster at specific gene locations. These locations can be detected, in a group of gene sequences, as high frequency 'peaks' with respect to a reference position, such as the transcription start site (TSS). We have developed a web tool for the identification of regions containing significant motif peaks. We show, by using different yeast gene datasets, that peak regions are strongly enriched in experimentally-validated motifs and contain potentially important novel motifs.

Availability: http://genomics.imim.es/peaks

Contact: malba@imim.es

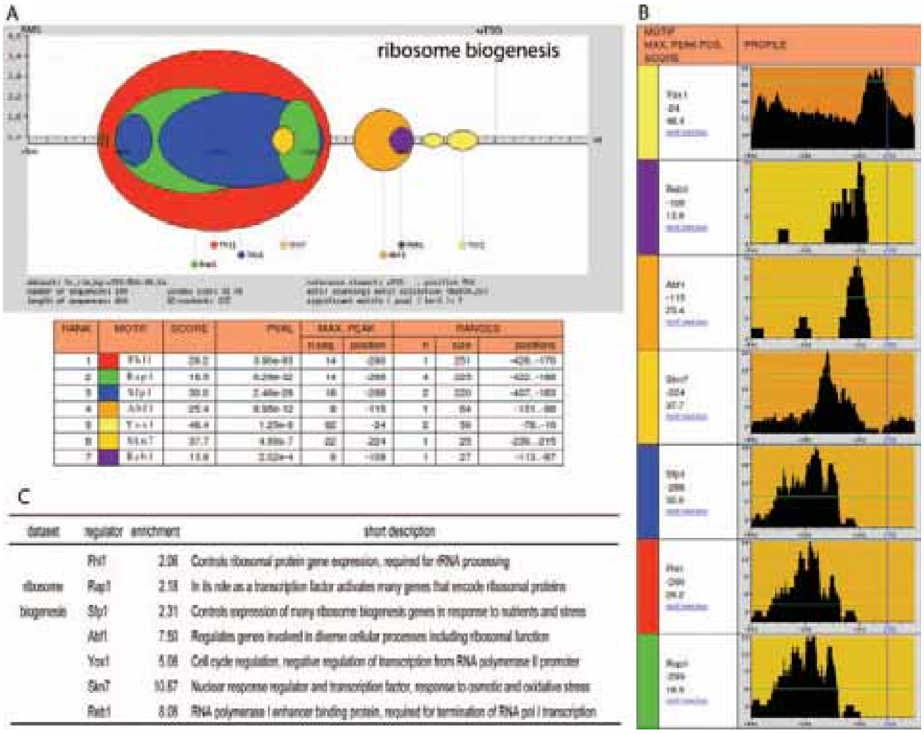
Supplementary information: Supplementary Data are available at *Bioinformatics* online.

The identification of regulatory motifs in DNA sequences is a challenging problem in bioinformatics. Computational predictions of known motifs, such as transcription factor binding sites (TFBS) often contain an unacceptable number of false positives, due to the short size and variability of the motifs. Focusing on motifs that are shared by several sequences can increase the specificity of motif predictions. For example, one can select sequences that have been conserved during evolution, a strategy known as phylogenetic footprinting (Lenhard *et al*., 2003). A different type of evolutionary constraint is related to the position of motifs along the gene sequence. There is ample evidence that many gene expression regulatory motifs show a biased location within promoter sequences (FitzGerald *et al*., 2004; Xie *et al*., 2005). That is, they are not randomly distributed but tend to accumulate or cluster in particular regions, forming high abundance 'peaks'. This presumably reflects specific requirements of motif-binding proteins that need to interact with each other to regulate transcription. The identification of significant motif peaks can be used to increase the specificity of motif predictions, provide information on the promoter structure, and help discover regulatory motifs that are specifically involved in the regulation of genes with similar expression or function. Motivated by the lack of available computational methods to detect motif clustering we have developed a novel algorithm for this purpose, which we have termed 'positional footprinting' and which is implemented in the web server PEAKS.

PEAKS can be used to analyze any group of sequences that share a known reference element, such as the transcription start site (TSS), the initiation codon, a known TFBS or any other predefined site. The scope is to detect any other motifs that show a significant clustering at a particular distance from the reference element. In the first step of the procedure the sequence positions that show matches to motifs from a user-selected library are recorded. Available motif libraries are: (1) compilations of TFBS position-specific weight matrices (PSWMs), (2) all possible DNA words of a given length or (3) pre-built consensus motif collections (Zhu and Zhang, 1999; Harbison *et al*., 2004). Several PSWM libraries can be used: TRANSFAC (Matys *et al*., 2003), Jaspar (Sandelin *et al*., 2004) and PROMO (Messeguer *et al*., 2002). Using DNA words can aid in the discovery of putative new motifs in different types of DNA sequences. In the second step, the positions of predicted motifs are used to build motif frequency profiles along the sequences. A position is considered positive for a motif is the motif occurs within a sequence window surrounding that position. Increasing the window size above the default value (31) allows the detection of motifs that do no have a very precise location at the cost of decreasing the significance of motifs located at very well defined positions (see Supplementary Table S1 for a full list of program parameters). The third step is the calculation of the positional footprinting score, S_{pf}, which measures the relative over-representation of a motif at a particular position (see PEAKS web server for a full mathematical description). The fourth step is the statistical evaluation of the maximum S_{pf} score obtained for each motif. To this end, we apply the same procedure described above to simulated random sequence datasets, which can be generated using an order 1 Markov model, to obtain an empirical p-value associated with the maximum S_{pf} score. If significant, we extract any other positions with a S_{pf} score above the p-value cut-off, which define the motif significant regions. The output includes a graphical representation of all the significant motifs and regions, a list of sequences containing significant motifs, motif profile pictures and a summary table.

Figure 1 shows the output produced by PEAKS in a dataset of 180 yeast genes involved in ribosome biogenesis (Mewes *et al*., 2002). Sequences spanned from -500 to $+100$ with respect to the most used TSS (Zhang and Dietrich, 2005). Motifs were detected using exact matches to a consensus motif collection containing 102 different TFBS (Harbison *et al*., 2004), and a sliding window of 31 nucleotides. An integrated picture (Fig. 1A) was derived from the significant regions in the profiles at p-value < 1e−3 (Fig. 1B).

[*]To whom correspondence should be addressed.



RANK	MOTIF	SCORE	PVAL	MAX PEAK		RANGES		
				n.seqs	position	n	size	positions
1	🟥 Fhl1	29.2	3.90e-83	14	-280	1	251	-426..-176
2	🟩 Rap1	16.5	6.29e-52	14	-288	4	225	-422..-198
3	🟦 Sfp1	30.6	2.46e-26	16	-288	2	220	-407..-183
4	🟧 Abf1	25.4	8.55e-12	8	-115	1	64	-131..-88
5	🟨 Yox1	48.4	1.25e-8	62	-24	2	59	-76..-15
6	🟨 Skn7	37.7	4.89e-7	22	-204	1	25	-239..-215
7	🟪 Reb1	13.6	2.02e-4	6	-108	1	27	-113..-87

dataset	regulator	enrichment	short description
ribosome biogenesis	Fhl1	2.06	Controls ribosomal protein gene expression, required for rRNA processing
	Rap1	2.18	In its role as a transcription factor activates many genes that encode ribosomal proteins
	Sfp1	2.31	Controls expression of many ribosome biogenesis genes in response to nutrients and stress
	Abf1	7.50	Regulates genes involved in diverse cellular processes including ribosomal function
	Yox1	5.08	Cell cycle regulation, negative regulation of transcription from RNA polymerase II promoter
	Skn7	10.67	Nuclear response regulator and transcription factor, response to osmotic and oxidative stress
	Reb1	8.08	RNA polymerase I enhancer binding protein, required for termination of RNA pol I transcription

Fig. 1. Results from PEAKS. Dataset comprising 180 *Saccharomyces cerevisiae* promoter sequences (−500 to +100 with respect to the most used TSS) from genes involved in ribosome biogenesis. Window size 31 nt. (**A**) Integrated representation. Significant regions were detected for Fhl1, Rap1, Sfp1, Abf1, Reb1, Yox1 and Skn7 motifs ($P < 1e{-}3$). Oval width indicates significant region boundaries. Oval height is the relative motif signal (RMS), the ratio between the number of sequences that correspond to the maximum peak and the number of sequences that contain the motif at the P-value cut-off. The table shows the score, P-value (PVAL), number of sequences and position with the maximum score (maximum peak), and the significant regions (ranges). (**B**) Significant motif profiles. The x-axis represents the sequence positions and the y-axis the number of sequences with a match to the motif. The green line represents the P-value cut-off, regions above the line are significant. Left of the profile picture is the motif name (e.g. Yox1), the position of maximum peak (−24 for Yox1) and associated S_{pf} score (48.4 for Yox1), and below a link to a list of genes containing significant motifs (motif matches). (**C**) Description and enrichment in experimentally-validated sites for significant motifs (see main text).

Five of the seven significant motifs, Fhl1, Rap1, Sfp1, Abf1 and Reb1, are known to be involved in the regulation of ribosomal-related genes (Fig. 1C). Yox1 and Skn7, have, so far, not been associated with this function, but their distribution indicates that they are strong candidates. We calculated the ratio between the observed fraction of experimentally-validated motifs falling into a significant region and the fraction of motifs expected in this region under a random motif distribution (size of the significant region divided by the total length of the sequence). The enrichment in real motifs ranged from 2.06 for Fhl1 to 10.67 for Skn7 (Fig. 1C). New putative binding sites for these transcription factors were discovered. For example, among the 24 different Skn7 motifs in the significant region (−239 to −215) only four were previously known. A second example, using a dataset of 86 yeast genes involved in amino acid metabolism, is provided in Supplementary Figure 1S.

ACKNOWLEDGEMENTS

The authors thank Loris Mularoni, Eduardo Eyras, Robert Castelo and Oscar González for useful discussions during this work. The authors acknowledge support from Fundación Banco Bilbao Vizcaya Argentaria (FBBVA), Plan Nacional de I + D MCyT (BIO2002-04426-C02-01), EC Infobiomed NoE and Fundació ICREA.

Conflict of Interest: none declared.

REFERENCES

FitzGerald,P.C. *et al*. (2004) Clustering of DNA sequences in human promoters. *Genome Res*., **14**, 1562–1574.

Harbison,C. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Lenhard,B. *et al*. (2003) Identification of conserved regulatory elements by comparative genome analysis. *J. Biol*., **2**, 13.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Messeguer,X. *et al*. (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, **18**, 333–334.

Mewes,H.W. *et al*. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*., **30**, 31–34.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.

Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3′-UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Zhang,Z. and Dietrich,F. (2005) Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res*., **33**, 2838–2851.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.

BMC Genomics

BioMed Central

Research article

Open Access

Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters

Nicolás Bellora[1,2], Domènec Farré[2] and M Mar Albà*[1,3,4]

Address: [1]Research Unit on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain, [2]Centre for Genomic Regulation, Barcelona, Spain, [3]Fundació Institut Municipal d'Investigació Mèdica, Barcelona, Spain and [4]Catalan Institution for Research and Advanced Studies, Barcelona, Spain

Email: Nicolás Bellora - nicolas.bellora@upf.edu; Domènec Farré - dfarre@imim.es; M Mar Albà* - malba@imim.es

* Corresponding author

Published: 13 December 2007

BMC Genomics 2007, **8**:459 doi:10.1186/1471-2164-8-459

Received: 27 April 2007

Accepted: 13 December 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/459

Abstract

Background: The arrangement of regulatory motifs in gene promoters, or promoter architecture, is the result of mutation and selection processes that have operated over many millions of years. In mammals, tissue-specific transcriptional regulation is related to the presence of specific protein-interacting DNA motifs in gene promoters. However, little is known about the relative location and spacing of these motifs. To fill this gap, we have performed a systematic search for motifs that show significant bias at specific promoter locations in a large collection of housekeeping and tissue-specific genes.

Results: We observe that promoters driving housekeeping gene expression are enriched in particular motifs with strong positional bias, such as YY1, which are of little relevance in promoters driving tissue-specific expression. We also identify a large number of motifs that show positional bias in genes expressed in a highly tissue-specific manner. They include well-known tissue-specific motifs, such as HNF1 and HNF4 motifs in liver, kidney and small intestine, or RFX motifs in testis, as well as many potentially novel regulatory motifs. Based on this analysis, we provide predictions for 559 tissue-specific motifs in mouse gene promoters.

Conclusion: The study shows that motif positional bias is an important feature of mammalian proximal promoters and that it affects both general and tissue-specific motifs. Motif positional constraints define very distinct promoter architectures depending on breadth of expression and type of tissue.

Background

The control of gene transcription is mediated by transcription factors, which interact in a sequence-specific manner with DNA motifs, known as transcription factor binding sites (TFBS). These motifs are abundant in gene promoter regions, upstream from the transcription start site (TSS). The promoter is often divided into the basal or core promoter, covering approximately 100 bp upstream of the TSS, and the proximal promoter, which extends up to a few hundred base pairs and typically contains multiple sites for activators [1]. Other functional regions, such as enhancers, can be found at very distant locations from the TSS. However, it appears that the region spanning from -550 to +50 with respect to the TSS is sufficient, in a large proportion of human genes, to drive transcriptional activity in cultured cells [2].

One important aspect of promoter sequences is the specific arrangement of regulatory motifs along the DNA sequence, and the existence of recurrent patterns in the relative position of motifs. It has been observed that a number of TFBS, including motifs for some of the most abundant transcription factors, show a tendency to cluster in the proximal promoter [3-7]. For example CCAAT enhancer binding protein (CEBP) motifs are basically found within an area from -100 to -50 with respect to the TSS [8]. Another example is cyclic-AMP response element (CRE), found in mammals far more frequently within 150 bp upstream from the TSS than in any other region [9]. On the other hand, it has been recently observed that a number of motifs that are likely to be important for the regulation of the expression of ribosomal protein genes are located at fixed positions within the promoter [10]. It is also well-known that TFBS can be arranged in particular combinations forming functional regulatory units, known as *cis*-regulatory modules [11,12]. Spacing between motifs can be the result of transcription factor interaction requirements in the context of particular *cis*-regulatory modules. This type of constraints can be revealed by the analysis of relative motif positions in many different genes, with the discovery of recurrent motif location patterns or 'positional footprints'. A tool that can be used to detect motif frequency profiles, using DNA words or a restricted set of known TFBS matrices, is Signal search analysis server [13]. We have recently developed another application, PEAKS [14,15], which, in addition to oligomers, uses existing TFBS matrix libraries, calculates 'positional footprinting' scores and associated p-values, and produces integrated motif views from large gene datasets. Here we use PEAKS to perform the most exhaustive to date analysis of motif positional biases in mammalian gene promoters. To explore the effect of tissue-specificity we use microarray data from 55 mouse tissues [16]. The analysis identifies distinctive features of promoters driving housekeeping or tissue-specific expression, shows that a number of well-known tissue-specific regulatory motifs are subject to strong positional constraints and predicts novel regulatory elements in different tissue expression gene datasets.

Results

Positional bias of general motifs

We collected mouse gene sequences, spanning from -600 to +100 with respect to the TSS, using the UCSC genome database [17]. This is what we will term "promoters", although it approximately corresponds to what is generally understood as the proximal promoter region. In the first place, we aimed at identifying general motifs that showed a positional bias in a significant number of promoters. We analyzed 6,372 non-redundant mouse gene promoter sequences with the previously developed program PEAKS [14,15]. A scheme of the procedure employed by PEAKS is shown in Figure 1. The first step is the identification of putative motifs on the sequences using one or more motif libraries. In this study we used four different libraries: 508 vertebrate weight matrices corresponding to known transcription factor binding sites from TRANSFAC [18]; 91 vertebrate weight matrices corresponding to known transcription factor binding sites from JASPAR, or JASPAR CORE matrices [19]; 174 weight matrices from JASPAR corresponding to putative regulatory sequences on the basis of phylogenetic conservation, or JASPAR phyloFACTS [19]; and a non-redundant set of 2080 oligomers of size 6 (6mers). The second step of the procedure is the generation of motif frequency profiles along the promoter. The profiles represent the number of sequences in which a motif is predicted at least once in a sequence window surrounding each position. In this analysis, we used a window size of 31 nucleotides, so occurrence of motifs anywhere from -15 to +15 with respect to the central position was sufficient for that position to be positive. The use of sliding windows, instead of strict positions, provides a certain degree of flexibility to accommodate functional motif and TSS position variability. The third step is the calculation of the positional footprinting score (*Spf*) of the position with the highest motif frequency (maximum peak in the profile). This score measures the tendency of the motif to be located in a particular region of the promoter, taking into account its overall abundance and distribution [14]. Using random sequences that mimic nucleotide variability along promoters, we obtain the p-value that corresponds to any particular *Spf* score. Promoter sequences contain regions with very biased GC content. To model realistic sequence datasets we first partition all mouse promoter sequences into three distinct types of regions according to their composition: 1. CpG islands; 2. GC-rich regions that are not CpG islands and; 3. The rest of regions (see Methods for an exact definition). We derive three distinct order 1 Markov chain models from sequence regions that belong to the same compositional class. Using these Markov chains, we generate random sequence datasets with the same number of sequences, and same partitioning in region types, as in the real sequences. As a result, the random sequences show similar composition to the real sequences along the promoter (Additional file 1). Throughout this work, we used a p-value <= 10^{-5} to identify motifs with significant *Spf* scores, unless stated otherwise.

In the complete mouse promoter dataset, we identified 29 significant motifs corresponding to matches to TRANSFAC matrices, 4 to JASPAR CORE matrices, 9 to JASPAR phyloFACTS, and 22 to 6mers. In many cases, the same motif was found by several of the libraries, but in other cases the information obtained was complementary. Although we considered a promoter region of length 700 nucleotides (from -600 to +100), all motifs were found in

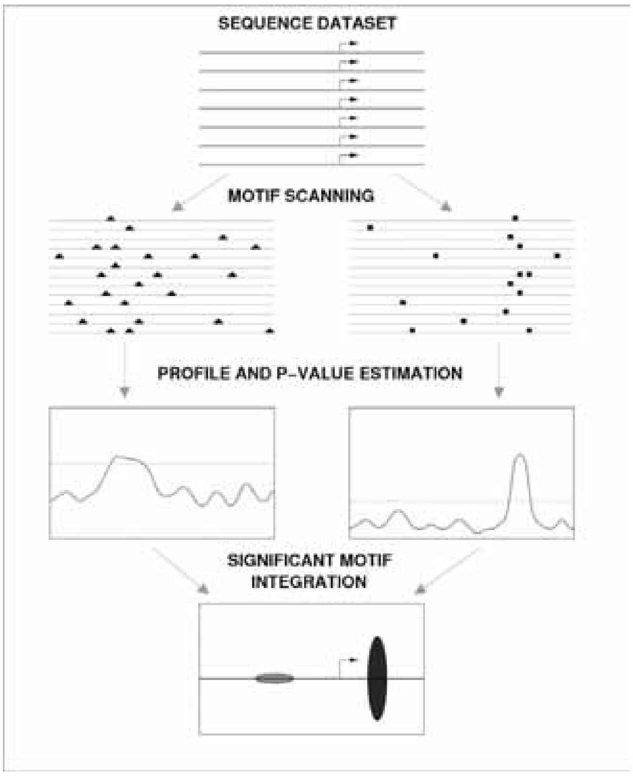


Figure 1
Schematic representation of the PEAKS method. Detection of positional bias of two hypothetical motifs in a promoter sequence dataset is shown. After motif scanning, a profile of motif frequency is obtained. The horizontal line delineates the region above a given p-value cut-off. Significant regions are plotted into a single integrated representation.

a much smaller region, between around -150 and +100. Basically, they were binding sites for general, commonly found, transcription factors. Figure 2 shows an integrated representation of the significant motifs obtained with the four libraries. To deal with motif redundancy, both within and across libraries, we clustered the motifs on the basis of the degree of overlap in all promoter sequences (see Methods and Additional file 2). We obtained nine different motif clusters, which are plotted with the same color in Figure 2.

The nine different types of motifs showed characteristic preferential positions with respect to the TSS (Figure 2, Figure 3 ALL, and Additional files 3, 4, 5, 6). The Ying and Yang (YY1/NF-E1) binding site motif was found downstream of the TSS (maximum peak at +26 with the TRANSFAC matrix V$YY1_02). In addition to several TRANSFAC matrices (V$YY1_02, V$YY1_Q6 and V$NFMUE1_Q6), this motif was detected by three different JASPAR phylo-FACTS (AAGWWRNYGGC, ACAWNRNSRCGG and GCCATNTTG) and three 6mers (GCCATG, ATGGCG, TGGCGG). Another characteristic element that we detected in this region was made of repeats of GGC or AGC. This motif was only detected with 6mers, and maximum peaks were located between +18 and +28 depending on the specific 6mer (see Figure 2 and Additional file 6). Upstream from the TSS, in a region around -20 to -40, we detected three types of motifs corresponding to known transcription factor binding sites. The first one corresponded to binding sites for the ETS-domain containing family of transcription factors: TEL, ELK and GABP (maximum peak at -31 with V$ELK_01). The second one was the TATA box (maximum peak at -36 with the JASPAR CORE TBP matrix). The third motif was the E2F binding site (maximum peak at -38 with V$E2F1_Q3_Q1). Transcription factors containing the ETS domain are involved in the regulation of transcription in a great variety of biological processes in metazoans [20]. On the other hand, E2F factors have been reported to be important for the control of the cell cycle [21]. Further upstream we found CREB-type motifs (cAMP response element-binding), which are bound by CBP/ATF/E4F transcription factors (maximum peak at -45 with V$E4F1_Q6). The region that was significant for GC-box/SP1 motifs was located further upstream (maximum peak at -62 with V$GC_01). The transcription factor SP1 is involved in the expression of many different genes, and can interact with other transcription factors, such as TBP (TATA-binding protein), Ying and Yang and E2F [22]. Other GC-rich motifs, corresponding to binding sites for factors ZF5 and ETF were part of the same motif cluster. A motif resembling the SP1 motif, but sufficiently distinct to be part of a different cluster, was identified with 6mers AGGCGG and TCCGCC (GGCGGA when reversed), around the same region. Finally, we identified CAAT-box/NF-Y motifs in a more upstream position (maximum peak at -76 with V$NFY_C).

Widely expressed versus tissue-specific genes

We next classified the mouse promoter sequences in several groups according to where the gene was expressed, using normalized microarray data from 55 different mouse organs and tissues [16]. In the first place we wanted to investigate if the arrangement and nature of the most common motifs depended on the breadth of expression. We defined a group of genes with expression limited to 1–10 tissues ('restricted', 1822 genes) and a second group of genes with expression in 51–55 different tissues ('housekeeping', 1544 genes). A comparison of the results obtained in the three different datasets – complete promoter dataset (ALL), housekeeping genes (HK) and restricted genes (RT) – is shown in Figure 3. In each motif profile, the region of the peak that is above the line is the motif significant region (represented by the width of the oval in Figure 2). Only one representative TRANSFAC or

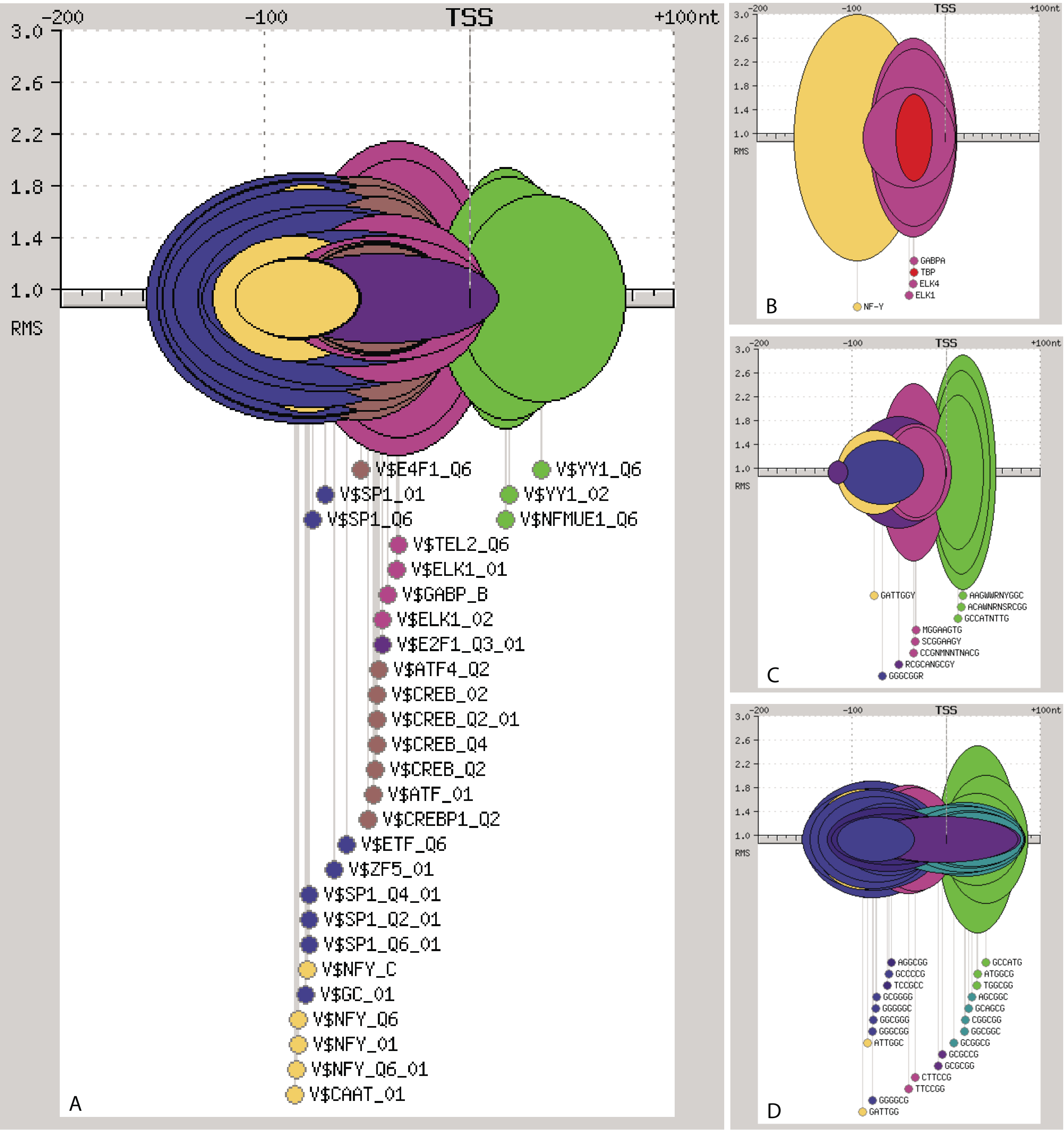
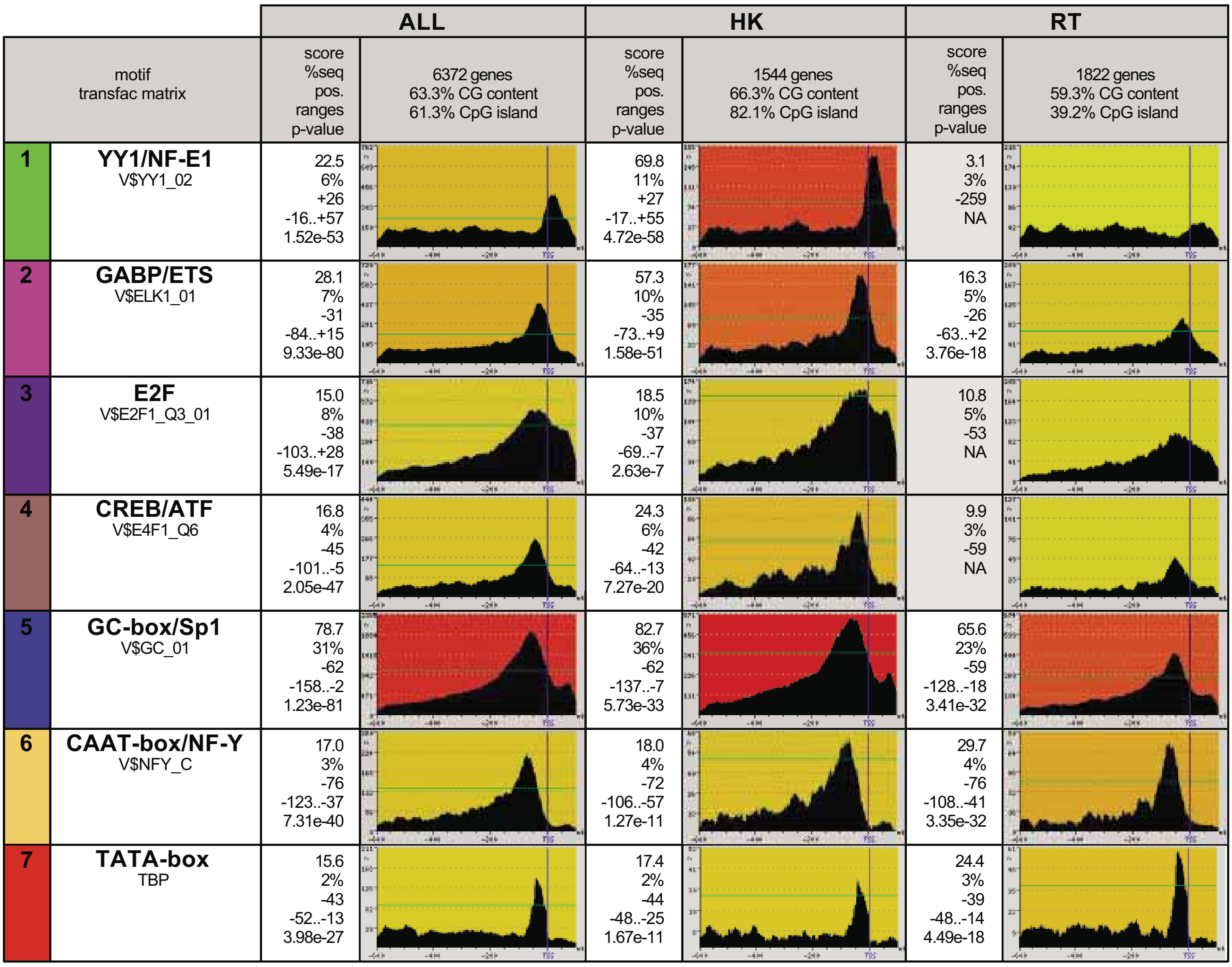


Figure 2

Integrated representation of motifs with significant positional bias in mouse promoters. The results were obtained by the program PEAKS, using different motif libaries. A. TRANSFAC PSWMs. B. JASPAR CORE PSWMs. C. JASPAR phyloFACTS. D. oligomers of size 6 (6mers). Motifs that belong to the same motif cluster are shown with the same color. A region from -200 to +100 with respect to the TSS is shown. The width of the ovals is the significant region of each motif (p-value <= 10^{-5}). The height of the ovals, the relative motif signal (RMS), is the number of sequences that contain a motif located at the position with the maximum score divided by the minimum number of sequences containing that motif that would be required to pass the p-value cut-off.



		ALL		HK		RT	
	motif transfac matrix	score %seq pos. ranges p-value	6372 genes 63.3% CG content 61.3% CpG island	score %seq pos. ranges p-value	1544 genes 66.3% CG content 82.1% CpG island	score %seq pos. ranges p-value	1822 genes 59.3% CG content 39.2% CpG island
1	**YY1/NF-E1** V$YY1_02	22.5 6% +26 -16..+57 1.52e-53		69.8 11% +27 -17..+55 4.72e-58		3.1 3% -259 NA	
2	**GABP/ETS** V$ELK1_01	28.1 7% -31 -84..+15 9.33e-80		57.3 10% -35 -73..+9 1.58e-51		16.3 5% -26 -63..+2 3.76e-18	
3	**E2F** V$E2F1_Q3_01	15.0 8% -38 -103..+28 5.49e-17		18.5 10% -37 -69..-7 2.63e-7		10.8 5% -53 NA	
4	**CREB/ATF** V$E4F1_Q6	16.8 4% -45 -101..-5 2.05e-47		24.3 6% -42 -64..-13 7.27e-20		9.9 3% -59 NA	
5	**GC-box/Sp1** V$GC_01	78.7 31% -62 -158..-2 1.23e-81		82.7 36% -62 -137..-7 5.73e-33		65.6 23% -59 -128..-18 3.41e-32	
6	**CAAT-box/NF-Y** V$NFY_C	17.0 3% -76 -123..-37 7.31e-40		18.0 4% -72 -106..-57 1.27e-11		29.7 4% -76 -108..-41 3.35e-32	
7	**TATA-box** TBP	15.6 2% -43 -52..-13 3.98e-27		17.4 2% -44 -48..-25 1.67e-11		24.4 3% -39 -48..-14 4.49e-18	

Figure 3

Promoter motif profiles in mouse genes with different expression width. ALL: complete promoter dataset; HK: housekeeping genes; RT: genes with restricted expression. Profiles were obtained with the program PEAKS using window size 31. Profiles with no significant sequence ranges (NA) did not accomplish p-value <= 10^{-5}. Left-most cells contain the TRANSFAC matrix (or JASPAR for TBP) used for motif prediction and the significant regions in the different datasets. Background color indicates score value grading, from intense red (highest) to pale yellow (lowest). 'score' is the positional footprinting score; '%seq' percentage of sequences at maximum peak; 'pos.', position of the maximum peak; 'ranges' sequence interval significant above the p-value cut-off.

JASPAR matrix per motif cluster is shown, the complete data is available in Additional files 3, 4, 5, 6.

Interestingly, there were very clear differences between HK and RT genes. The peak corresponding to motifs for Ying and Yang and nuclear factor E1 (YY1/NF-E1) was much sharper in housekeeping genes than in the general dataset (compare HK to ALL in Figure 3), and completely absent from genes with restricted expression (RT). In the HK dataset, 11% of the genes contained this motif in position +27, while this number was only 3% for RT genes, around the level of background signal for this motif. The YY1 factor is ubiquitous and involved in the control of basal transcription [23], which is consistent with our results. Besides YY1, two other motifs did not achieve statistical significance in the RT dataset. The first one was the cluster CREB/ATF/E4F, which showed a much sharper peak in the HK dataset than in the RT dataset. In particular, the percentage of HK genes containing E4F motifs at the maximum peak position was 6%, twice that of RT genes. E4F is a ubiquitously expressed protein reported to be important for mitotic progression [24]. The other motif that was not significant in RT genes was the E2F binding site, which was also about twice as frequent in HK genes than in RT genes. On the other hand, GABP/ETS and GC-box/Sp1 motifs also showed higher *Spf* scores in HK than in RT. Contrary to the motifs mentioned above, the TATA-box, as well as CAAT-box/NF-Y, were stronger in the RT dataset than in the HK dataset, although clearly significant in both.

An important outcome of the comparison between ALL, HK and RT datasets was that most of the motifs showed higher scores in the HK dataset than in the RT dataset. This is not surprising, as the latter are an amalgam of genes with very diverse expression patterns, and so they are likely to have more varied motif configurations or more distally located regulatory regions. We also observed that the average GC content, and in special the proportion of genes with CpG islands, was higher in HK than in RT gene promoters (Figure 3). Average GC content was 66.3% in HK and 59.3% in RT, while the number of CpG island-containing promoters was 82% in HK and 39.2% in RT. These differences are in agreement with previous reports [25,26].

Tissue-specific motifs

We analyzed in greater detail the genes showing strong tissue-specificity, by performing a separate analysis of groups of RT genes expressed in each of the different adult tissues (N = 47). For example the dataset 'liver' was composed of genes from the RT class (expressed in 1–10 tissues) that showed expression in liver. One can expect that some tissues will be more similar to each other, in regard to the genes that they express, than others. To learn about this, we clustered the tissues according to the number of shared expressed genes. We identified four main clusters, in which every pair of tissues shared at least 30% of the genes of one tissue. The clusters, A to D, corresponded, to a large extent, to known physiological systems (Additional file 7). Cluster A was composed by diverse tissues from the nervous system; cluster B was mainly composed by tissues related to the digestive system; cluster C by muscle and skin tissues; and cluster D by bone, lymph and bladder. We obtained non-redundant gene datasets for each cluster. These datasets were composed of RT genes for which at least 50% of the tissues in which they were expressed belonged to that cluster. Surprisingly, commonly found motifs (those shown in Figure 2) showed a very different distribution in different RT gene clusters (Additional data files 3, 4, 5, 6). For example, the GABP/ETS motif, as well as CREB/ATF, only reached significant scores in cluster D; the GC-box/SP1 was only significant in cluster A and B; and, the CAAT-box/NF-Y was only significant in cluster A.

In the analysis of RT genes expressed in particular tissues (47 datasets) we obtained 337 significant motif peaks at p-value <= 10^{-5}: 169 with TRANSFAC matrices, 18 with JASPAR CORE matrices, 48 with JASPAR phyloFACTS matrices and, 102 with 6mers (Additional data files 3, 4, 5, 6). Many of the motifs corresponded to common transcription factor binding sites, already detected in the analysis of all genes. To identify motifs that were directly related to tissue-specificity, we obtained a list of motifs that were significant in RT genes expressed in a given tissue but not in HK genes. We identified 58 different ones, found in one or a few related tissues. Of these, 14 corresponded to TRANSFAC matrices, 2 to JASPAR CORE matrices, 10 to JASPAR phyloFACTS matrices and, 32 to 6mers. Figure 4 shows a selection of such motifs. A number of them are well-known tissue-specific motifs. For example in genes expressed in liver, aside from the more general TATA and CAAT sites, there were significant peaks for HNF-1 (maximum peak at -79 with matrix V$HNF1_01), and HNF-4 (maximum peak at -92 with V$HNF4_01_B). HNF-1 and HNF-4 are members of the hepatocyte nuclear factor (HNF) family, and are well-known regulators of expression in liver and other related tissues [27]. Accordingly, the HNF-4 motif was also found in large intestine (main peak at -82 with V$DR1_Q3), and, with p-value <= 10^{-3}, in small intestine (main peak at -78 with V$HNF4_01_B) and kidney (main peak at -91 with V$HNF4_DR1_Q3). The HNF-1 motif was also significant, at p-value <= 10^{-3}, in kidney (maximum peak at -70 with V$HNF1_Q6).

Several motifs were repeatedly found in tissues from the nervous system (cluster A, Additional file 7). GC-box/SP1 and alphaCP1 motifs were particularly strong in nervous tissue genes. Among tissue-specific motifs, MZF1 was significant in cortex, hindbrain and midbrain (maximum peak between -39 to -44, p-value <= 10^{-3}); AP2 in brain, cortex, hindbrain and striatum (maximum peak between -50 and -58, p-value <= 10^{-3} in the three latter tissues); and EGR in striatum (maximum peak at -81, p-value = 6.08 × 10^{-4}). There is evidence that the factors EGR1, AP2 and SP1 are required for the neuroendocrine-specific expression of chromogranine B gene [28]. Myeloid zinc finger 1 (MZF-1) is known to play a major role in myeloid cell differentiation. The enrichment we find in neural tissue expressed genes may mean that this factor regulates neural processes as well, or that the motif resembles the consensus sequences for another, yet uncharacterized, neural factor.

In testis, the RFX1 motif was significant (max. peak at -16 with V$RFX1_02), which is consistent with the abundance of RFX factors in this tissue [29]. This motif was not found in any other tissue. Similarly, MYB and PBX1 were only found only in bone marrow (max. peak at -4 and -473, respectively, p-value <= 10^{-3}). MYB is known to be important for the regulation of hematopoiesis [30].

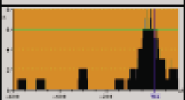
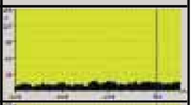
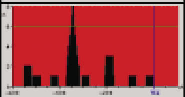
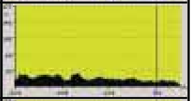
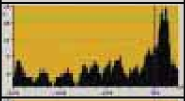
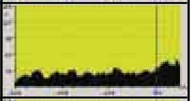
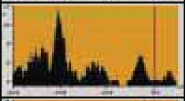
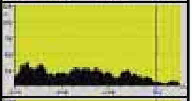
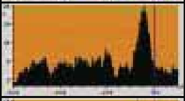
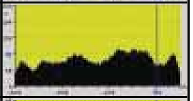
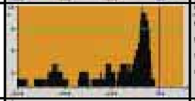
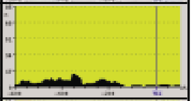
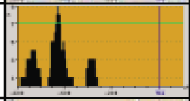
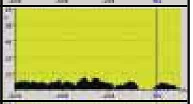
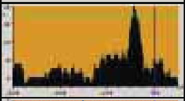
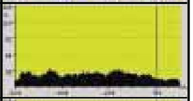
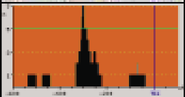
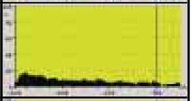
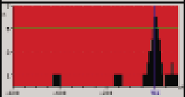
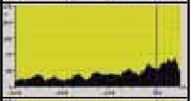
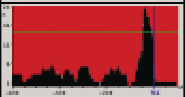
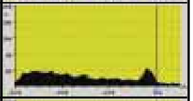
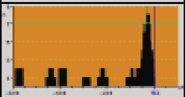
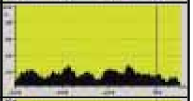
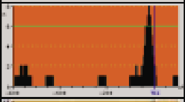
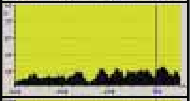
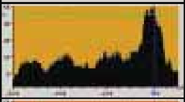
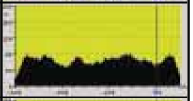
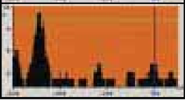
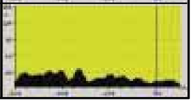
	tissue dataset motif library p-value	score %seq pos. ranges	tissue profile	HK profile
1	Bladder RGAGGAARY jaspar phylofacts 1.01e-6	41.0 10% -34 -42..-14		
2	Bone_Marrow TAGAAC 6mer 6.37e-8	95.3 11% -353 -354..-339		
3	Brain CTGCAGY jaspar phylofacts 1.00e-6	29.0 10% +31 +25..+54		
4	Hindbrain ATGAGA 6mer 4.57e-6	37.6 8% -413 -415..-407		
5	Kidney V$CACBINDINGPROTEIN_Q6 transfac 3.11e-6	45.1 15% -47 -49..-43		
6	Liver V$HNF1_01 transfac 1.33e-7	38.5 5% -79 -82..-49		
7	Liver V$MEIS1BHOXA9_01 transfac 8.82e-6	32.9 4% -433 -438..-422		
8	Liver V$HNF4_01_B transfac 9.42e-6	36.9 9% -92 -92..-79		
9	Mammary_gland YCATTAA jaspar phylofacts 1.90e-6	57.4 7% -307 -311..-300		
10	Mandible GGGTCG 6mer 4.54e-6	101.6 11% +1 -3..+10		
11	Snout V$MTATA_B transfac 2.74e-13	141.4 16% -42 -49..-12		
12	Stomach CCTAGG 6mer 3.15e-6	43.4 7% -33 -36..-22		
13	Teeth GCAACG 6mer 4.75e-7	58.6 6% -29 -32..-16		
14	Testis V$RFX1_02 transfac 1.35e-7	35.1 13% -16 -43..+1		
15	Uterus AGATTC 6mer 8.24e-6	54.9 10% -490 -500..-484		

Figure 4

Promoter motif profiles in mouse genes expressed in particular tissues. Selection of motifs that were significant in genes expressed in a particular tissue but not in the housekeeping (HK) dataset. See also Legend to Figure 3.

Interestingly, there were several tissue-specific motifs that could be detected with JASPAR phyloFACTS, or by 6mers, but not using matrices for known transcription factor binding sites. Many of these motifs are likely to correspond to yet uncharacterized transcription factor binding sites. For example phyloFACTs motif CTGCAGY showed a significant peak at +31 in brain, RGAGGAARY at -34 in bladder, and YCATTAA at -307 in mammary gland (Figure 4). Other putative tissue-specific motifs were only detected with 6mers. Examples include TAGAAC, at -353 in bone marrow, ATGAGA at -413 in hindbrain, and AGATTC at -490 in uterus.

Transcription factor target predictions

An important outcome of this work was the prediction of many novel potential transcription factor sequence targets in the regions showing significant positional bias (p-value <= 10^{-5}). It is a well-known fact that predictions of regulatory motifs suffer from the problem of false positive detection. However, given the strong position-dependency of the motifs found by PEAKS, predictions within the identified significant regions are expected to be much more reliable than predictions elsewhere in the promoter (see also next section). Using TRANSFAC matrices, predictions for commonly found binding sites (those in Figure 2A) were mapped to 5,798 different promoters (Additional file 8). This means that the vast majority of promoters (91%) contain at least one of the general regulatory motifs in the significant sequence range. Besides, we also obtained 559 predictions for motifs not significant in the ALL or HK datasets, providing annotations for putative tissue-specific transcription factor binding sites in 394 different promoters (Additional file 9). The total number of genes with one or more predicted motifs was 5942 (Additional file 10). Among tissue-specific motifs we found 86 RFX1 matches in 74 different promoters, 61 AP2 matches in 47 promoters, 40 PU1 matches in 34 promoters and, 32 HNF4 matches in 20 promoters.

Comparison with experimental data

In a previous study using yeast promoters, we showed that regions identified by PEAKS were significantly enriched in real binding sites [14]. To compare the computational results of this study with experimental data, we systematically search all the experimental binding site annotations for mouse genes in TRANSFAC, and map them onto our genes. We recovered 35 non-redundant experimentally validated sites that could be successfully mapped to genes in our dataset, for GC-box/SP1, CAAT-box/NF-Y, CREB/ATF, YY1/NF-E1, GAB/ETS and HNF-4. In general, the computational and experimental results were in very good agreement, and 25 of the 35 sites fell within significant regions (p-value <= 10^{-5}). By individual motifs, 15/18 of the GC-box/SP1 experimental sites, 5/7 of the experimental YY1/NF-E1 sites and 3/4 of the experimental CAAT-box/NF-Y fell within regions that were significant in the PEAKS analysis. For CREB/ATF, instead, only 1 out of 4 sites were located in PEAKS significant regions. For GABP/ETS and HNF-4 we only had one experimental site to compare with. The GABP/ETS site fell within the significant region. However, the HNF-4 site, in cytochrome P450 Cyp3a16, was located upstream from the region identified by PEAKS. This finding prompted us to scrutinize all other HNF-1 and HNF-4 experimental sites in TRANSFAC mouse gene entries, even if the genes were not in our dataset. These motifs were present in four additional TRANSFAC mouse gene entries: albumin 1 (HNF-1), alpha-fetoprotein (HNF-1 and HNF-4), retinol-binding protein II (HNF-4) and, transthyretin (HNF4). Of these 5 cases, 4 fell within the regions identified by PEAKS (-92 to -79), and only the HNF-4 motif in retinol-binding protein II was outside the significant region. In the work presented here, we found 31 additional putative HNF-4 sites, in different mouse promoters, which fell within the significant region. Given the positive outcome of the comparison between computational and experimental site locations, many of these sites are likely to be functional. In support of this, a region in which we predict HNF-4 sites in the hepatic lipase gene, has been recently observed to be responsible for enhanced promoter activity in liver cells, and for silencing expression in non-liver cells [31].

Discussion

Important information on the spatio-temporal expression pattern of a gene is encrypted in gene promoter sequences. Within promoters, particular arrangements of regulatory motifs facilitate specific transcription factor interactions, which result in transcription activation or repression. Transcription regulatory regions can evolve quickly, and similar motifs are often present in genes with coordinated gene expression, even if the genes are not homologues. Recurrent motif arrangements are thus presumably the result of similar evolutionary constraints in genes that are part of the same regulatory network. In the present study we have focused on motif arrangement in the proximal promoter, using the distance from the transcription start site. Until now, studies on positionally biased regulatory motifs had only been performed for general promoter motifs [6,7,32], or, at the other extreme, for motifs found in very specific datasets of functionally related genes [10,33]. Here we have investigated the impact of motifs with positional bias in the configuration of promoters driving expression in various body tissues, and used this property to uncover potentially novel tissue-specific regulators.

A number of computational studies have established that particular DNA words tend to cluster in the vicinity of the transcription start site in mammalian gene promoters [6,7,32]. Our analysis indicates that the TATA box is not a particularly common motif, the peak observed using TRANSFAC matrix V$MTATA_B corresponds to only 3,4% of the genes, although given that this refers to a region +/- 15 bp of position -41, it is likely to be an under-estimation of the real number of sites. The low frequency of this motif is in strong contrast with previous ideas on the central role of this motif in transcription, but more in line with more recent estimates based on larger datasets [34,35]. Indeed, TATA-containing promoters are more typical of tissue-specific genes than of housekeeping genes, and show a high degree of conservation across species [34]. Promoters containing GC-rich SP1 binding sites, on the contrary, appear to be very widespread, and their frequency is higher in housekeeping than in tissue-specific genes (Figure 3). Other very common motifs in mammalian promoters include binding sites for the ETS family of transcription factors, for E2F1, and CAAT-box/NF-Y motifs. None of the known basic motifs in the core promoter appears to be universal, and each one is present in only a fraction of genes. Basic motifs can combine in different ways, and it has been shown that some combinations – such as CAAT and SP1 sites – are particularly common [6]. Interactions between several of the transcription factors that assemble at the core promoter have been described, including YY1 and SP1 [36], E2F and SP1 [37] and, NF-Y (CAAT-box) and TATA binding protein (TBP)-associated factors [38]. These protein interactions are likely to impose constraints on the relative positions of the corresponding DNA motifs, which would explain why we find such strong motif positional dependencies. In support of this, it has been shown that the activity of the thymidine kinase promoter depends on the distance between E2F motifs and upstream SP1 binding sites [37].

Our results strongly indicate that housekeeping gene promoters have more fixed promoter structures than the class composed of promoters driving restricted tissue expression. This is not surprising, as distinct regulators are expected to control expression in different tissue types. On the other hand, we have shown that the Ying and Yang (YY) downstream motif is a very important constitutive element of genes with broad expression, whereas it appears to be of little relevance in genes that show tissue-specific expression. Other motifs, such as E2F and CREB/ATF/E4F, also show much stronger peaks in housekeeping genes than in tissue-restricted genes. Interestingly, in the latter the maximum peak position is displaced towards a more upstream position (-59 in RT, versus -42 in HK, for E4F, Figure 3), pointing to possible mechanistic differences in the way these factors interact with the initiation complex in the two classes of genes.

The control of tissue-specific expression is still poorly understood. We have been able to identify a number of motifs that show positional bias in tissue-restricted datasets. Previous studies on the identification of tissue-specific motifs were based on cross-species conservation and subsequent detection of tissue enrichment [5], or on the identification of *cis*-regulatory modules with high tissue-specific expression predictive value [39]. In relation to the latter study, Smith et al. [40] provided a list of tissue-specific expression important motifs: HNF-1, HNF-3, HNF-4, C/EBP and DBP in liver; MEF-2, SRF, Myogenin and SP1 in skeletal muscle and; SRY, CREM, RFX in testis (see Table III of [40]). Of these motifs, we found that HNF-1 and HNF-4 in liver, and RFX in testis, showed significant positional biases. Instead, the above-mentioned muscle-specific motifs were not identified in our analysis. This could be due to a more flexible and variable arrangement of motifs in these genes, or simply to the motifs being outside the region of the promoter considered (proximal region). In relation to this, it has been recently proposed that motifs bound by RFX factors are very abundant in conserved non-coding regions, scattered throughout the genome [41]. In another study [42], using cross-specific conservation criteria, it was found that AP-2, SP1 and EGR-1 were over-represented in neural tissues. AP-2 and EGR-1 showed positional bias in several nervous system tissues. On the other hand, SP1, while significant in the majority of tissues, achieved the largest positional footprinting scores in mammary gland, brown fat and pancreas.

Many of the motifs that show significant positional bias in our analysis are located within the first 100 bp upstream of the TSS. This is not surprising considering that the sequences are anchored at the TSS in this analysis, and position dependencies between interacting motif-binding proteins are expected to be more relevant for short distances [36,37]. More unexpected is the presence of motifs with positional bias much further upstream, in several tissue-restricted datasets. This includes MEIS1BHOXA9 in liver (maximum peak at -433), PBX1 in bone marrow (maximum peak at -473, p-value = 4.22×10^{-4}), STAT5A in eye (maximum peak at -469, p-value = 4.26×10^{-4}), and OCT1 in olfactory bulb (maximum peak at -540, p-value = 8.35×10^{-4}). One possibility to explain these cases is the existence of stronger evolutionary constraints in a longer portion of the promoter. Our own data on the weaker sequence conservation of housekeeping promoters with respect to tissue-specific distal promoters, particularly upstream from position -500, points in this direction [43]. On the other hand, from this study it can also be concluded that, contrary to what is generally assumed, the motif content of the region around the TSS can vary greatly depending on specific tissue expression. Dataset-specific motifs with positional bias have also been identified in ribosomal gene [10] or histone-coding gene promoters [33]. Therefore, both shared motif content and shared relative motif positions appear to be important for the regulation of genes with similar tissue expression patterns.

Conclusion

In this work we have shown that motifs with positional bias are abundant in mammalian promoters and can be used to define distinct promoter architectures depending on breadth or tissue of gene expression. The results offer new insights into the shaping of motif arrangement in promoter sequences by evolutionary processes. We provide predictions for a large number of motifs, including general as well as tissue-specific motifs, that show positional bias. This work provides a foundation for future studies on motif position constraints in gene regulatory sequences.

Methods

DNA sequences and tissue expression data

Gene datasets were defined from mouse transcriptome microarray data from Zhang et al. [16]. The corresponding gene promoter sequences were extracted from UCSC database (mm6) [17]. We selected genes that had a unique annotated TSS in the database as a representative set. The analysis comprised 6,372 non-redundant promoter sequences, which spanned from -600 to +100 relative to the TSS position. These sequences define the ALL dataset. Subsequently, genes were classified in 3 classes according to expression breadth: housekeeping (HK), 1,544 genes expressed in 51–55 tissues; intermediate, 3,006 genes expressed in 11–50 tissues and; restricted (RT), 1,822 genes with expression restricted to 1–10 tissues. Because many tissues can share cell types, or cell functions, we calculated the number of shared genes between tissues. We measured the overlap between all pairs of tissues and selected those pairs sharing at least 30% of genes. We selected 4 clusters that contained more than 2 adult mouse tissues. They showed a good agreement with physiological systems: 'nervous' (A), 'digestive/kidney' (B), muscular/skin' (C) and, 'skeletal/lymphatic/bladder' (D). They are shown in Additional file 7.

DNA motif prediction

For the detection of known motifs in the sequences we used three weight matrix collections of transcription factor binding sites: TRANSFAC 7 containing 508 vertebrate position specific weight matrices (PSWMs), JASPAR containing 91 vertebrate CORE PSWMs and, JASPAR 174 phyloFACTS PSWMs. Sequence hits to a matrix were defined as those that showed an overall matrix relative similarity score >= 0.90 and, for TRANSFAC matrices, an overall matrix relative similarity score >= 0.85 and core similarity score >= 0.99 [18]. To measure similarity to the TRANSFAC matrices we implemented the metrics described in [44], as used in the program MatInspector. For JASPAR matrices we used log-likelihood ratio scores. We also scanned the sequences for perfect matches to all oligomers of size 6 (6mers). Matches to both the sense and the anti-sense strand were considered. For this reason, the number of effective 6mers to be tested could be reduced from 4096 to 2080 (including 64 palindromic 6mers).

Positional footprinting (PEAKS analysis)

For those DNA motifs that showed at least one match in any promoter sequence we performed PEAKS analysis. In this analysis, all sequences were of the same length (l) and contained a common element, the transcription start site (TSS), used as the reference position. For each DNA motif we scanned the sequences with a sliding window (w, uneven size) and counted the number of sequences that contained at least one occurrence of the DNA motif (motifs were matches to PSWMs or 6mer, see above) within that window, assigning this number, $n(i)$, to the window central nucleotide, i. We used these values to build a motif profile along the sequence positions. In order to determine the positional bias of a motif we assigned a signal to noise score to each profile and estimated its p-value using random sequence datasets (see below). We then extracted the significant positions where the motif was located.

To measure the positional bias of a motif, which is basically the number of motif occurrences, at a particular position, $n(i)$ relative to the background signal level, we use the positional footprinting score *Spf* [14]. It results from three diverse scores. The first score (*Sn*) measures the number of motif occurrences at a specific position with respect to the average number along the sequence. The second score (*Sr*) penalizes signals present in only a very small percentage of sequences, by dividing the number of ocurrences at the specific location by the number of sequences used. Finally, the third score (*Sm*) is the number of occurrences at that position divided by the total number of motif predictions, used to penalize matrices that are very noisy and occur at a very high frequency, which is often due to low specificity of the matrix. As the scores account for different aspects of the signal to noise ratio, we multiply them to obtain a single final score: *Spf* = *Sn Sr Sm*. See PEAKS web documentation for a more detailed description of the *Spf* score [15].

The maximum value of the positional footprinting score *Spf*, which corresponds to the maximum peak of the motif, was defined as *Spf* _max. To assess the significance of *Spf* _max for each DNA motif tested in the dataset, we used 1000 different synthetic sequence datasets (see below). In each simulation we kept the random *Spf* _max. We then counted how many simulations showed a random *Spf* _max equal or higher than the observed *Spf* _max, and obtained the corresponding empiric p-value. The *Spf* _max values were distributed according to an extreme value distribution. We used this property to estimate the p-value that corresponded to a given score using linear interpolation. After selection of a p-value cut-off, the significant regions were defined by the concatenation of all positions that showed a score associated with a p-value below the cut-off.

Throughout this study we used $w = 31$ and p-value <= 10^{-5}, unless stated otherwise. In addition, we filtered those motifs that, even if statistically significant, showed multiple peaks or very weak peaks (less than two fold motif frequency at the maximum peak position with respect to the background or *Spf* _max < 15).

Construction of synthetic datasets

In the synthetic datasets, each random sequence had a similar composition than a real sequence in the dataset. This was achieved by using three order 1 Markov models, each of which corresponded to a compositionally different region. The three compositionally different regions were defined in the complete mouse promoter sequence dataset. The first type corresponded to CpG islands, regions of length at least 200 bp, with a minimum GC content of 55% and a minimum observed/expected GpG content ratio of 0.65 [43]. The second type corresponded to GC-rich regions that did not conform to the CpG island definition. They were at least 200 bp long and had a minimum GC content of 55%. The remaining regions made the third type. Each promoter in the study was partitioned into these three regions. Of course, different promoters varied in the number and extension of these regions. We then concatenated all the regions that were of the same type to construct three different order 1 Markov chains. Each random sequence was generated using one, two or three Markov chains, preserving the partitioning in different regions observed in the original sequence. By this approach, we obtain synthetic datasets that were remarkably similar in composition to real datasets along the promoter (Additional file 1).

Motif clustering

There was a considerable amount of redundancy in the motifs identified, both within and across motif libraries. To disentangle it, we clustered the motifs using hierarchical clustering (R package complete hierarchical clustering, [45]). Distance between motifs was based on the proportion of overlapping motif matches along all non-redundant promoter sequences. Specifically, say we have motif A and motif B (represented as matches to PSWMs or 6mers). Then the distance between A and B will the dist $(A, B) = ((N(A, B)/N(A))+(N(A, B)/N(B))/2$, where $N(A, B)$ is the number of predictions of motif A and predictions of motif B that overlap, $N(A)$ the total number of predictions of motif A and, $N(B)$ the total number of predictions of motif B. The dissimilarity cut-off used was 0.98. This approach resulted in 9 different clusters out of a total of 65 significant motifs in the complete mouse promoter dataset (Additional file 2).

Mapping of significant motifs in promoter sequences

The PEAKS analysis yielded significant regions for various motifs in each of the datasets tested. We extracted the actual predictions of the motifs in the promoter sequences, considering those motifs that fell within the significant region, and those located up to 15 nucleotides upstream or downstream of this region, as they also contributed to the peak considering that the window size employed was 31. Additional files 8 and 9 contain the predictions of general (significant in the complete collection of mouse promoters) as well as non-general motifs. Additional file 10 is a zipped file containing individual files with predictions of general and non-general motifs per each gene, in BED format. This includes 5942 genes, for which we found significant motif predictions, and a README file with instructions on how to visualize them using UCSC Genome Browser.

Global over-representation statistics

We calculated motif frequencies in complete promoter sequences using the PSWM predictions as described previously. To assess if a motif was over-represented in a particular dataset we calculated the corresponding p-value using synthetic datasets as described for positional footprinting. In addition, we compared the relative abundance of the motif in the particular dataset to that obtained in the general dataset (ALL). The values for each motif that showed significant positional bias are provided in Additional files 3, 4, 5, 6.

Authors' contributions

NB and MMA designed the study. NB and DF carried out the computations. NB, DF and MMA analyzed the data. NB and MMA wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Additional file 1 contains average nucleotide composition along the promoter for real and synthetic datasets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S1.pdf]

Additional file 2

Additional file 2 contains motif clustering for motifs in Figure 2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S2.pdf]

Additional file 3

Additional file 3 contains results of motif positional bias searches using TRANFAC matrices, at p-value <= 10^{-5}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S3.xls]

Additional file 4

Additional file 4 contains results of motif positional bias searches using JASPAR CORE matrices, at p-value <= 10^{-5}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S4.xls]

Additional file 5

Additional file 5 contains results of motif positional bias searches using JASPAR phyloFACTS matrices, at p-value <= 10^{-5}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S5.xls]

Additional file 6

Additional file 6 contains results of motif positional bias searches using 6mers, at p-value <= 10^{-5}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S6.xls]

Additional file 7

Additional file 7 contains tissue clusters in which every pair of tissues shares at least 30% of the genes of restricted (RT) expression class.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S7.eps]

Additional file 8

Additional file 8 contains predictions of general motifs in mouse promoters, in significant regions described in Additional file 3.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S8.txt]

Additional file 9

Additional file 9 contains predictions of non-general motifs in mouse promoters, in significant regions described in Additional file 3.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S9.txt]

Additional file 10

Additional file 10 contains the data in Additional files 8 and 9 in BED format, for visualization using UCSC Genome Browser.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-459-S10.zip]

Acknowledgements

We acknowledge the researchers at the Computational Genomics and Structural Biology labs (GRIB, CRG, PRBB) for many useful and enriching discussions during this work. We received financial support from Fundación Banco Bilbao Vizcaya Argentaria (FBBVA), Plan Nacional de I+D Ministerio de Educación y Ciencia (BFU2006-07120), Instituto Nacional de Bioinformática (INB), European Commission Infobiomed NoE and, Fundació ICREA.

References

1. Maston GA, Evans SK, Green MR: **Transcriptional Regulatory Elements in the Human Genome.** *Annu Rev Genomics Hum Genet* 2006, **7:**29-59.
2. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM: **Identification and functional analysis of human transcriptional promoters.** *Genome Res* 2003, **13(2):**308-312.
3. Berendzen KW, Stuber K, Harter K, Wanke D: **Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves.** *BMC Bioinformatics* 2006, **7:**522.
4. Sharov AA, Dudekula DB, Ko MS: **CisView: a browser and database of cis-regulatory modules predicted in the mouse genome.** *DNA Res* 2006, **13(3):**123-134.
5. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434(7031):**338-345.
6. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C: **Clustering of DNA sequences in human promoters.** *Genome Res* 2004, **14(8):**1562-1574.
7. Marino-Ramirez L, Spouge JL, Kanga GC, Landsman D: **Statistical analysis of over-represented words in human promoter sequences.** *Nucleic Acids Res* 2004, **32(3):**949-958.
8. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-1419.
9. Smith B, Fang H, Pan Y, Walker PR, Famili AF, Sikorska M: **Evolution of motif variants and positional bias of the cyclic-AMP response element.** *BMC Evol Biol* 2007, **7 Suppl 1:**S15.
10. Roepcke S, Zhi D, Vingron M, Arndt PF: **Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters.** *Gene* 2006, **365:**48-56.
11. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11(9):**1559-1566.
12. Howard ML, Davidson EH: **cis-Regulatory control circuits in development.** *Dev Biol* 2004, **271(1):**109-118.
13. Ambrosini G, Praz V, Jagannathan V, Bucher P: **Signal search analysis server.** *Nucleic Acids Res* 2003, **31(13):**3618-3620.
14. Bellora N, Farre D, Mar Alba M: **PEAKS: identification of regulatory motifs by their position in DNA sequences.** *Bioinformatics* 2007, **23(2):**243-244.
15. **PEAKS** [http://genomics.imim.es/peaks/]
16. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3(5):**21.
17. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31(1):**51-54.
18. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31(1):**374-378.
19. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34(Database issue):**D95-7.
20. Sharrocks AD: **The ETS-domain transcription factor family.** *Nat Rev Mol Cell Biol* 2001, **2(11):**827-837.
21. Rowland BD, Bernards R: **Re-evaluating cell-cycle regulation by E2Fs.** *Cell* 2006, **127(5):**871-874.

22. Suske G: **The Sp-family of transcription factors.** *Gene* 1999, **238(2):**291-300.
23. Smale ST, Baltimore D: **The "initiator" as a transcription control element.** *Cell* 1989, **57(1):**103-113.
24. Le Cam L, Lacroix M, Ciemerych MA, Sardet C, Sicinski P: **The E4F protein is required for mitotic progression during embryonic cell cycles.** *Mol Cell Biol* 2004, **24(14):**6467-6475.
25. Vinogradov AE: **Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth.** *Trends Genet* 2005, **21(12):**639-643.
26. Yamashita R, Suzuki Y, Sugano S, Nakai K: **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity.** *Gene* 2005, **350(2):**129-136.
27. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA: **Control of pancreas and liver gene expression by HNF transcription factors.** *Science* 2004, **303(5662):**1378-1381.
28. Mahapatra NR, Mahata M, Ghosh S, Gayen JR, O'Connor DT, Mahata SK: **Molecular basis of neuroendocrine cell type-specific expression of the chromogranin B gene: Crucial role of the transcription factors CREB, AP-2, Egr-1 and Sp1.** *J Neurochem* 2006, **99(1):**119-133.
29. Wolfe SA, van Wert J, Grimes SR: **Transcription factor RFX2 is abundant in rat testis and enriched in nuclei of primary spermatocytes where it appears to be required for transcription of the testis-specific histone H1t gene.** *J Cell Biochem* 2006, **99(3):**735-746.
30. Thomas MD, Kremer CS, Ravichandran KS, Rajewsky K, Bender TP: **c-Myb is critical for B cell development and maintenance of follicular B cells.** *Immunity* 2005, **23(3):**275-286.
31. van Deursen D, Botma GJ, Jansen H, Verhoeven AJ: **Comparative genomics and experimental promoter analysis reveal functional liver-specific elements in mammalian hepatic lipase genes.** *BMC Genomics* 2007, **8:**99.
32. Bajic VB, Choudhary V, Hock CK: **Content analysis of the core promoter region of human genes.** In *Silico Biol* 2004, **4(2):**109-125.
33. Chowdhary R, Ali RA, Albig W, Doenecke D, Bajic VB: **Promoter modeling: the case study of mammalian histone promoters.** *Bioinformatics* 2005, **21(11):**2623-2628.
34. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006.
35. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y: **Mice and men: their promoter properties.** *PLoS Genet* 2006, **2(4):**e54.
36. Lee JS, Galvin KM, Shi Y: **Evidence for physical interaction between the zinc-finger transcription factors YY1 and Sp1.** *Proc Natl Acad Sci U S A* 1993, **90(13):**6145-6149.
37. Karlseder J, Rotheneder H, Wintersberger E: **Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F.** *Mol Cell Biol* 1996, **16(4):**1659-1667.
38. Frontini M, Imbriano C, diSilvio A, Bell B, Bogni A, Romier C, Moras D, Tora L, Davidson I, Mantovani R: **NF-Y recruitment of TFIID, multiple interactions with histone fold TAF(II)s.** *J Biol Chem* 2002, **277(8):**5841-5848.
39. Smith AD, Sumazin P, Xuan Z, Zhang MQ: **DNA motifs in human and mouse proximal promoters predict tissue-specific expression.** *Proc Natl Acad Sci U S A* 2006, **103(16):**6275-6280.
40. Smith AD, Sumazin P, Zhang MQ: **Tissue-specific regulatory elements in mammalian promoters.** *Mol Syst Biol* 2007, **3:**73.
41. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.** *Proc Natl Acad Sci U S A* 2007, **104(17):**7145-7150.
42. Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P: **Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals.** *BMC Genomics* 2005, **6(1):**68.
43. Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM: **Housekeeping genes tend to show reduced upstream sequence conservation.** *Genome Biol* 2007, **8(7):**R140.
44. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23(23):**4878-4884.
45. **R Project** [http://www.r-project.org/]

Apèndix 3

Publicacions i presentacions a congressos

Publicacions de Domènec Farré

Article 1:

Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM: **PROMO: detection of known transcription regulatory elements using species-tailored searches**. Bioinformatics 2002, 18:333-334.

Factor d'impacte: 4.615 (de 2002)

Article 2:

Farré D, Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X: **Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN**. Nucleic Acids Res 2003, 31:3651-3653.

Factor d'impacte: 6.575 (de 2003)

Article 3:

Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM: **Housekeeping genes tend to show reduced upstream sequence conservation**. Genome Biol 2007, 8:R140.

Factor d'impacte: 7.172 (de 2006)

Article 4:

Blanco E, Farré D, Albà MM, Messeguer X, Guigó R: **ABS: a database of Annotated regulatory Binding Sites from orthologous promoters**. Nucleic Acids Res 2006, 34:D63-67.

Factor d'impacte: 6.317 (de 2006)

Article 5:

Bellora N, Farré D, Mar Albà M: **PEAKS: identification of regulatory motifs by their position in DNA sequences**. Bioinformatics 2007, 23:243-244.

Factor d'impacte: 4.894 (de 2006)

Article 6:

Bellora N, Farré D, Albà MM: **Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters**. BMC Genomics 2007, 8:459.
Factor d'impacte: 4.029 (de 2006)

Contribucions a congressos

2008: Annual Meeting of the Society for Mol. Biology and Evolution (SMBE2008), Barcelona (Espanya)
Títol: Effect of gene duplication on gene expression regulatory sequence evolution.
Autors: Farré D. i Albà M.M. (acceptat com a pòster)

2008: VIII Jornadas de Bioinformática, València (Espanya)
Títol: Identification of regulatory motifs by positional bias in mammalian promoters.
Autors: Bellora N., Farré D. i Albà M.M. (comunicació oral)

2007: The Biology of Genomes, Cold Spring Harbor Laboratory, New York (USA)
Títol: Distinct architecture of housekeeping gene promoters versus tissue-specific promoters.
Autors: Farré D., Bellora N., Mularoni L., Messeguer X. i Albà M.M. (pòster)

2006: Biospain-Biotec 2006, Madrid (Espanya)
Títol: Identification of functional motifs in DNA sequence.
Autors: Farré D., Bellora N., Messeguer X., Guigó R. i Albà M.M. (comunicació oral)

2005: 4th European Conference on Computational Biology 2005 (ECCB05), Madrid (Espanya)
Títol: Promoter divergence in mammals.
Autors: Farré D., Messeguer X. i Albà M.M. (pòster)

2005: ESF workshop on Transcription Networks: a Global View, Madrid (Espanya)
Títol: Variability and conservation in vertebrate promoters.
Autors: Bellora N., Farré D. i Albà M.M. (comunicació oral)

2004: V Jornades de Bioinformàtica (JBI 2004), Barcelona (Espanya)
Títol: Prediction of transcription factor binding sites with PROMO v.3: Improving the specificity of weight matrices and the searching process.
Autors: Farré D., García D., Albà M.M. i Messeguer X. (pòster)

2002: I Reunió 'Xarxa Catalana de Bioinformàtica', Les Avellanes - Balaguer (Espanya)
Títol: PROMO: Prediction of known transcription regulatory elements common to multiple sequences.
Autors: Farré D., Messeguer X. i Albà M.M. (comunicació oral)

2002: III Jornadas de Bioinformática, Salamanca (Espanya)

Títol: PROMO: Detection of known transcription regulatory elements common to multiple sequences.

Autors: Farré D., Messeguer X. i Albà M.M. (pòster)

Agraïments als Editors

Vull donar gràcies als editors de les revistes en que s'han publicat els articles inclossos en aquesta tesi. Especialment vull agraïr els drets de reproduir-los aquí.

Gràcies a *Bioinformatics* i *Nucleic Acids Research* i als seus editors, Oxford University Press, per la publicació dels següents articles:

> Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 2002, 18(2):333-334.
>
> Farré D, Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. Nucleic Acids Res 2003, 31(13):3651-3653.
>
> Blanco E, Farré D, Albà MM, Messeguer X, Guigó R. ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. Nucleic Acids Res 2006, 34(Database issue):D63-67.
>
> Bellora N, Farré D, Mar Albà M. PEAKS: identification of regulatory motifs by their position in DNA sequences. Bioinformatics 2007, 23(2):243-244.

Gràcies també a BioMed Central per les publicacions a *Genome Biology* i *BMC Genomics*, ambdues revistes d'accés públic (*open access*):

> Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol 2007, 8(7):R140.
>
> Bellora N, Farré D, Albà MM. Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Genomics 2007, 8:459.

Acknowledgements to Publishers

I acknowledge rights to reproduce the articles included in this thesis from the magazines that originally published them. I thank the magazines and their publishers.

I thank Oxford University Press for the articles published in *Bioinformatics* and *Nucleic Acids Research*:

> Messeguer X, Escudero R, Farré D, Núñez O, Martínez J, Albà MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 2002, 18(2):333-334.
>
> Farré D, Roset R, Huerta M, Adsuara JE, Roselló L, Albà MM, Messeguer X. Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. Nucleic Acids Res 2003, 31(13):3651-3653.
>
> Blanco E, Farré D, Albà MM, Messeguer X, Guigó R. ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. Nucleic Acids Res 2006, 34(Database issue):D63-67.
>
> Bellora N, Farré D, Mar Albà M. PEAKS: identification of regulatory motifs by their position in DNA sequences. Bioinformatics 2007, 23(2):243-244.

I also thank BioMed Central for the articles published in *Genome Biology* and *BMC Genomics*, both open access magazines:

> Farré D, Bellora N, Mularoni L, Messeguer X, Albà MM. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol 2007, 8(7):R140.
>
> Bellora N, Farré D, Albà MM. Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. BMC Genomics 2007, 8:459.

BioMed Central copyright and license agreement

In submitting a research article ('article') to any of the journals published by BioMed Central Ltd ('BioMed Central') I certify that:

1. I am authorized by my co-authors to enter into these arrangements.

2. I warrant, on behalf of myself and my co-authors, that:

a. the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;

b. I am/we are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other obligation. If the law requires that the article be published in the public domain, I/we will notify BioMed Central at the time of submission upon which clauses 3 through 6 inclusive do not apply;

c. the article contains nothing that is unlawful, libellous, or which would, if published, constitute a breach of contract or of confidence or of commitment given to secrecy;

d. I/we have taken due care to ensure the integrity of the article. To my/our - and currently accepted scientific - knowledge all statements contained in it purporting to be facts are true and any formula or instruction contained in the article will not, if followed accurately, cause any injury, illness or damage to the user.

And I agree to the following license agreement:

BioMed Central Open Access license agreement

Brief summary of the agreement

Anyone is free:

- to copy, distribute, and display the work;
- to make derivative works;
- to make commercial use of the work;

Under the following conditions: Attribution

- the original author must be given credit;
- for any reuse or distribution, it must be made clear to others what the license terms of this work are;
- any of these conditions can be waived if the authors gives permission.

Statutory fair use and other rights are in no way affected by the above.

Full BioMed Central Open Access license agreement

(Identical to the 'Creative Commons Attribution License')

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS BIOMED CENTRAL OPEN ACCESS LICENSE ("LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

a. **"Collective Work"** means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.

b. **"Derivative Work"** means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered a Derivative Work for the purpose of this License.

c. **"Licensor"** means the individual or entity that offers the Work under the terms of this License.

d. **"Original Author"** means the individual or entity who created the Work.

e. **"Work"** means the copyrightable work of authorship offered under the terms of this License.

f. **"You"** means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

2. Fair Use Rights
Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. License Grant
Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

a. to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b. to create and reproduce Derivative Works;

c. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works;

d. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission Derivative Works;

e. For the avoidance of doubt, where the work is a musical composition:

   i. **Performance Royalties Under Blanket Licenses.** Licensor waives the exclusive right to collect, whether individually or via a performance rights society (e.g. ASCAP, BMI, SESAC), royalties for the public performance or public digital performance (e.g. webcast) of the Work.

   ii. **Mechanical Rights and Statutory Royalties.** Licensor waives the exclusive right to collect, whether individually or via a music rights agency or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work ("cover version") and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions).

f. **Webcasting Rights and Statutory Royalties.** For the avoidance of doubt, where the Work is a sound recording, Licensor waives the exclusive right to collect, whether individually or via a performance-rights society (e.g. SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions).

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4. Restrictions
The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a. You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients' exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested. If You create a Derivative Work, upon notice from any Licensor You must, to the extent practicable, remove from the Derivative Work any reference to such Licensor or the Original Author, as requested.

b. If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Derivative Works or Collective Works, You must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied; to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and in the case of a Derivative Work, a credit identifying the use of the Work in the Derivative Work (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Derivative Work or Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability

EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Derivative Works or Collective Works from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

a. Each time You distribute or publicly digitally perform the Work or a Collective Work, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

b. Each time You distribute or publicly digitally perform a Derivative Work, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

c. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

d. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.