The evolution of gene expression in primates

Tashakkori Ghanbarian, Avazeh

Award date:
2015

Awarding institution:
University of Bath

The evolution of gene expression in primates

Avazeh Tashakkori Ghanbarian

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology

May 2015

Table of contents

Acknowledgments

This thesis is dedicated to my dearest mother who so delicately ignited my passion for reading and science; your persistence on me reading books, all those curious books in your library, non-stop science magazine subscriptions even when the austerity has hit the country so hard that white papered notebooks looked like a treasure at school. You have made me a budding scientist!

This thesis is nonetheless dedicated to my dear father whose support never failed to warm my heart and strengthen my steps, and to my dearest brother, Mehrdad, who is the best big brother I could have ever wished for, to my cousins, Golzar and Hamaseh, who kept me sane and happy, love you girls.

To my supervisor, Laurence, I am so much going to miss all those wonderful discussions we had. To Jennifer for never failing to bring all tired souls together for a happy Friday. To Atahualpa for all lunchtime chats whenever I could leave the cave! To Paul Kersey for being an inspiration and the best boss I have ever had.

Abbreviations

DHS: Dnase Hypersensitive Sit
ESC: Embryonic Stem Cell line
ESE: Exonic Splice Enhancer
HERV: Human Endogenous Retroviruses
HESC: Human Embryonic Stem Cell line
LAD: Lamina Associated Domain
lincRNA: Long Intergenic Non-Coding RNA
LTR: Long Terminal Repeat
ncRNA: non-coding RNA
TE: Transposable Element
TF: Transcription Factor

Abstract

The evolution of a gene's expression profile is commonly assumed to be independent of its genomic neighborhood. This is, however, in contrast to what we know about the lack of autonomy between expression of neighboring genes in extant taxa. Indeed, in all eukaryotic genomes, genes of similar expression-profile tend to cluster, reflecting chromatin level dynamics. Does it follow that if a gene increases expression in a particular lineage then the genomic neighbors will also increase in their expression or is gene expression evolution autonomous? To address this, I consider evolution of human gene expression since the human-chimp common ancestor, allowing for both variation in estimation of current expression level and error in Bayesian estimation of the ancestral state. I find that in all tissues and both sexes, the change in gene expression of a focal gene on average predicts the change in gene expression of neighbors. The effect is highly pronounced in the immediate vicinity but extends much further. Sex-specific expression change is also genomically clustered. As genes increasing their expression in humans tend to avoid nuclear lamina domains and be enriched for the gene activator 5-hydroxymethylcytosine, chromatin level mechanisms are likely regulators of this phenomenon. Firstly established in Primates and then expanded to compacted genome of yeasts, the phenomenon of correlation in change in gene expression of the neighbouring genes I describe as "expression piggy-backing", an analog of hitchhiking. Extending the same principle to non-coding genes I find a possible role of lincRNAs in regulating expression of their neighbours, mediated by a coupling between splicing and chromatin modification. Finally I employ insertions of human endogenous retroviruses (HERVs), as a naturally occurring transgene experiment, to find out how randomly scattered sequences would affect the expression profile of their neighboring genes. I show these retroviruses to be the focus of transcription in human ES cells and define a transcription factor, LBP9, as a novel pluripotency-associated agent. Transcription results in neighbourhood modification including the generation of chimaeric transcripts. Predictions were confirmed experimentally by collaborators.

Chapter 1. Introduction

Differences between species might be the result of either changes in gene expression or changes in genes. My thesis concentrates on the former. Most of the studies on the evolution of gene expression are based on gene centric model and frequently focus on the regulatory elements affecting a given gene (Hammock and Young 2002; Carninci et al. 2006; Tirosh et al. 2006; Wray 2007; Tirosh et al. 2009; Wang and Rekaya 2009; Molineris et al. 2011; Hornung et al. 2012; Rosin et al. 2012; Wittkopp and Kalay 2012; Forrest et al. 2014; Yang et al. 2014). Such studies typically concentrate analysis on changes to a gene's promoters and enhancers. In such a perspective, changes in the promoter change the expression of the gene controlled by that promoter but nothing else. With the exception of downstream effect of the for example changes in dose of transcription factors, these studies often fail to consider any downstream gene regulatory effects. Indeed, little effort has been put into examining up or down regulation of the neighboring genes. But is it likely that changes in the expression of a given gene are isolated (or insulated) or do they propagate through a gene's physical neighbors? Put differently, are genes autonomous in their evolution in the sense that the change in expression of a focal gene has no effects on its immediate genomic neighbors?

In contrast to the prevailing autonomous view of gene expression evolution, examining profiles of gene expression across chromosomes in eukaryotes genomes has revealed evidence for genomic clustering of genes of similar expression profile (Cho et al. 1998; Cohen et al. 2000; Caron et al. 2001; Reik and Walter 2001; Blumenthal et al. 2002; Hurst et al. 2002; Roy et al. 2002; Spellman and Rubin 2002; Birnbaum et al. 2003; Lee and Sonnhammer 2003; Lercher et al. 2003; Versteeg et al. 2003; Khaitovich et al. 2004; Stolc et al. 2004; Williams and Bowles 2004; Denver et al. 2005; Liu et al. 2005; Mijalski et al. 2005; Oliver and Misteli 2005; Singer et al. 2005; Sproul et al. 2005; Lercher and Hurst 2006; Sémon and Duret 2006; Purmann et al. 2007; Ebisuya et al. 2008; Nutzmann and Osbourn 2014).

This is seen both at a fine scale and a more gross chromosomal scale (Cohen et al. 2000; Caron et al. 2001; Lercher et al. 2003; Pal and Hurst 2003; Williams and Bowles 2004; Purmann et al. 2007; Michalak 2008; Woo and Li 2011). On a fine

scale neighboring genes tend to be co-expressed more than expected by chance across multiple taxa (Blumenthal et al. 2002; Boutanaev et al. 2002; Roy et al. 2002; Lercher et al. 2003; Fukuoka et al. 2004; Williams and Bowles 2004; Purmann et al. 2007; Davila Lopez et al. 2010), the effect being most pronounced often for genes in a bidirectional orientation, in which promoters sit in close proximity to each other (Cohen et al. 2000; Williams and Bowles 2004; Davila Lopez et al. 2010; Wei et al. 2011; Uesaka et al. 2014). On a more gross scale, genes expressed in most tissues (housekeeping genes) and highly expressed genes tend to cluster in domains corresponding to tens of genes (Caron et al. 2001; Lercher et al. 2002; Versteeg et al. 2003; Weber and Hurst 2011).

These genes co-expression clusters could contain genes controlled by the same transcription factors, as shown in yeast (Képès 2003; Janga et al. 2008). Chromatin level regulation of transcription could also facilitate establishment of these clusters (Grunstein 1997; Cohen et al. 2000; Sémon and Duret 2006; Batada et al. 2007; Li et al. 2007). Interestingly, in yeast even after controlling for transcription factor similarity, neighboring genes still show striking similarity in co-expression (Batada et al. 2007). Similarly, in mammals, incorporation of transgenes into chromosomes demonstrates that these adopt the expression profile of neighbors within a broad span (Gierman et al. 2007; Symmons et al. 2014). In both yeast and mammals, the upregulation of one gene causes time lagged ripples of gene expression that correspond to changes in chromatin state (Cohen et al. 2000; Janicki et al. 2004; Ebisuya et al. 2008). In humans these ripple domains are around 100kb in size. Whether evidence for gene expression clusters could also imply selection for such clusters is unresolved. In yeast, the most highly co-expressed gene pairs tend to be more similar in functionality and more commonly conserved as a pair (Hurst et al. 2002; Poyatos and Hurst 2007). However, results in other lineages are less decisive (Lee and Sonnhammer 2003; Lee and Sonnhammer 2004; Liao and Zhang 2008; Weber and Hurst 2011). It might be that selection is weak in which case we might expect to see evidence for selection in yeast (with large population sizes) but less so in mammals.

The contrast between intra-specific and inter-specific analyses is striking and suggests an evident question: if a gene is evolutionarily up-regulated in a given lineage, are its

neighbors more likely to be evolutionarily up-regulated too? If this is so, do the neighboring genes on the same strand show similar change in their gene expression as to the ones on the opposite direction? Does this just affect expression of the closest gene or is there a relative range of operation, e.g. 100Kb as suggested in ripple effect? Is there any tissue specific pattern in evolution of neighboring genes?

These questions are the main focus of the chapter 2 in this thesis. Using gene expression data provided for six tissues across 5 primates generated by Brawand et al (Brawand et al. 2011), I ask whether genes are autonomous in their expression evolution. Reconstructing the human-chimp ancestral state of expression for homologous genes in these primates, made it possible to estimate the extent of change in expression between humans and this ancestor. Several measures were then applied to quantify this change, including a Z score and a fold change. As Z score is more robust to the variation in expression between replicates (expression or measurement noise) and uncertainty in ancestral state reconstruction, most of the analysis were done considering Z score values. In addition, by considering the residuals of the orthogonal regression of Z for a gene in a given tissue in males against the same in females we could define the degree of sex bias in expression change. This enabled us to ask in turn whether this too shows evidence of autonomy or clustering.

Challenging the gene centric perception of evolution of gene expression in Primates, I then asked if the co-expression clusters observed in Primates also exists in other organisms. Are these co-evolving gene clusters a feature of large genomes like mammals or could we find evidence in more compact genomes, like yeasts. Would short intergenic regions in yeast result in a conflicting pattern compared to the one observed in Primates with extensive intergenic DNA?

To pave the way to extend this concept to other kingdoms of life, I have also investigated if phylogenetic assumption affects the correlation observed in the change of gene expression of the neighboring genes. The near ubiquitous agreement on approximate divergence date of 5 primates considered in the Primates study made analysing the effects of prior phylogenetic tree used unnecessary. However, phylogenetic relationships are still disputed across many taxa (Rutschmann 2006) and their effect on analysing evolution of gene expression is unknown. This is discussed

in the third chapter of my thesis, when piggybacking is examined in the content of single cell compact genome of yeasts.

While these two studies in Primates and Yeasts revealed hidden relationship in evolution of gene expression of coding genes in Primates and Yeasts, they both exclude non-coding genes. As we found evidence for the co-evolving gene expression clusters are possibly owing to the chromatin regulatory mechanisms, we might ask if non-coding genes could also have a role in evolution of gene expression of their neighboring coding genes through changing the chromatin state. This is indeed the focus of the fourth chapter of this thesis. In this chapter I ask whether transcription of non-coding RNA could regulate the neighboring genes? To enable as close a resemblance to the analysis of protein coding genes (which typically do not overlap), we decided to consider a particular class of non-coding RNAs, long intergenic non-coding RNA or lincRNA. Investigating this class of non-coding RNAs would not only enable us to answer this question but would also elucidate on an intriguing pattern of purifying selection acting on exonic splice enhancer (ESE) motifs in lincRNAs as found by my collaborator on this project, Andreas Schueler.

The exons of human non-coding RNAs, ncRNAs, are known to be poorly conserved compared to protein-coding genes (Marques and Ponting 2009). On average they evolve a little slower than their flanking introns (Hurst and Smith 1999; Pang et al. 2006), suggesting weak purifying selection. The causes of this weak purifying selection are unknown (Pang et al. 2006). However, this relatively rapid evolution need not imply an absence of function as the opposite was shown in a few well-studied functional ncRNAs, like *Xist* (Engreitz et al. 2013).

Using a well-defined set of lincRNAs (Cabili et al. 2011), Andreas asked where in ncRNAs purifying selection operates and what predicts rates of evolution of ncRNAs. He found evidence for weak purifying selection especially in lincRNAs' ESE motifs and then asked if this splice related selection could be explained by lincRNAs stability, in case they were to function as a scaffold, but could not find any evidence to support this hypothesis. He also investigated whether this purifying selection could be related to nonsense mediated decay pathway to capture transcripts incorrectly processed by ribosomes, which was not the case either. So then we asked whether the

purifying selection on ESE motifs could be explained by their role in regulation of the neighboring genes through a unique class of chromatin modifiers: spliced-coupled chromatin modifiers, like CHD1 (Marfella and Imbalzano 2007; Sims et al. 2007; Persson and Ekwall 2010; Hnilicova and Stanek 2011; Zentner et al. 2013). This was indeed the case; I found that intron rich lincRNAs are also enriched in CHD1's binding sites and they also correlate with DNase hypersensity sites, DHSs, a marker for open chromatin (Thurman et al. 2012), both of which also correlate with expression of neighbors. Hence these linRNAs are more likely to be involved in regulation of their neighboring genes through changing chromatin structure.

So far I have shown evidence to link locality of genes with their evolution of expression across the coding and non-coding genes in a large genome and also a compact genome. These results would lead a curious mind to ask if one is to insert similar sequences across a genome, to what extent these randomly scattered sequences would affect the expression profile of the genes in their vicinity. Human specific endogenous retroviruses, HERVs, would provide an excellent base as a naturally occurring transgene experiment to investigate this. Members of each HERV family also exhibit high sequence similarity (although in part this is tautological). In collaboration with Prof. Zsuzsanna Izvak, at Max Delbruck centre in Berlin, we have shown that actively transcribed members of a particular class of HERVs, HERV-H, are involved in regulating their neighbors in human ES cells. This HERV we observed to be common in the transcriptome of ES cells. To determine whether this is transcriptional read-through or because of the HERVH providing functional binding sites for transcription factors, I performed a set of epigenetic analyses. These studies indicated that the epigenetic markers for transcription initiation were indeed found in the HERVH. Next I determined what the possible binding partners might be. I found that HERVH not only provides functional binding sites for a combination of naïve pluripotency transcription factors but also discovered a marker for naïve stem cells. Also the long terminal repeats, LTRs, associated with HERV-H family in particular rewire regulating mechanisms associated with pluripotency. These LTRs, I discovered, should provide a harbour for a novel transcription factor, LBP9, to initiate transcription. The role of LBP9 was confirmed by my experimental colleagues.

When we analysed the products of transcription and the relationship to the neighbours we discovered that the HERVs do not simply modulate expression levels of neighbours, but, surprisingly, HERV-H elements have a few more tricks in their arsenal and are a source of novelty in our genome. Through creating chimeric transcripts, they not only create new genes but also affect splicing of their neighboring genes. All of this is explained in more detail in chapter five, where I have also clarified my contribution to the resulting paper (published in Nature).

References

Batada NN, Urrutia, AO, Hurst, LD. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. 23:480-484.

Birnbaum K, Shasha, D, Wang, J, Jung, J, Lambert, G, Galbraith, D, Benfey, P. 2003. A gene expression map of the Arabidopsis root. Science. 302:1956-1960.

Blumenthal T, Evans, D, Link, C, Guffanti, A, Lawson, D, Thierry-Mieg, J, Thierry-Mieg, D, Chiu, W, Duke, K, Kiraly, M, et al. 2002. A global analysis of caenorhabditis elegans operons. Nature. 417:851-854.

Boutanaev AM, Kalmykova, AI, Shevelyov, YY, Nurminsky, DI. 2002. Large clusters of co-expressed genes in the Drosophila genome. Nature. 420:666-669.

Brawand D, Soumillon, M, Necsulea, A, Julien, P, Csardi, G, Harrigan, P, Weier, M, Liechti, A, Aximu-Petri, A, Kircher, M, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature. 478:343-348.

Cabili MN, Trapnell, C, Goff, L, Koziol, M, Tazon-Vega, B, Regev, A, Rinn, JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915-1927.

Carninci P, Sandelin, A, Lenhard, B, Katayama, S, Shimokawa, K, Ponjavic, J, Semple, C, Taylor, M, Engstrom, P, Frith, M, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 38:626-635.

Caron H, van Schaik, B, van der Mee, M, Baas, F, Riggins, G, van Sluis, P, Hermus, M, van Asperen, R, Boon, K, Voute, P, et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science. 291:1289.

Cho RJ, Campbell, MJ, Winzeler, EA, Steinmetz, L, Conway, A, Wodicka, L, Wolfsberg, TG, Gabrielian, AE, Landsman, D, Lockhart, DJ, et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 2:65-73.

Cohen BA, Mitra, RD, Hughes, JD, Church, GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 26:183-186.

Davila Lopez M, Martinez Guerra, J, Samuelsson, T. 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLoS One. 5:e10654.

Denver D, Morris, K, Streelman, J, Kim, S, Lynch, M, Thomas, W. 2005. The transcriptional consequences of mutation and natural selection in caenorhabditis elegans. Nat Genet. 37:544-548.

Ebisuya M, Yamamoto, T, Nakajima, M, Nishida, E. 2008. Ripples from neighboring transcription. Nat Cell Biol. 10:1106-1113.

Engreitz JM, Pandya-Jones, A, McDonel, P, Shishkin, A, Sirokman, K, Surka, C, Kadri, S, Xing, J, Goren, A, Lander, ES, et al. 2013. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science. 341:1237973.

Forrest A, Kawaji, H, Rehli, M, Baillie, J, de Hoon, M, Lassmann, T, Itoh, M, Summers, K, Suzuki, H, Daub, C, et al. 2014. A promoter-level mammalian expression atlas. Nature. 507:462-470.
Fukuoka Y, Inaoka, H, Kohane, IS. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics. 5:4.
Gierman H, Indemans, M, Koster, J, Goetze, S, Seppen, J, Geerts, D, van Driel, R, Versteeg, R. 2007. Domain-wide regulation of gene expression in the human genome. Genome Res. 17:1286-1295.
Grunstein M. 1997. Histone acetylation in chromatin structure and transcription. Nature. 389:349-352.
Hammock EA, Young, LJ. 2002. Variation in the vasopressin V1a receptor promoter and expression: implications for inter- and intraspecific variation in social behaviour. Eur J Neurosci. 16:399-402.
Hnilicova J, Stanek, D. 2011. Where splicing joins chromatin. Nucleus. 2:182-188.
Hornung G, Oren, M, Barkai, N. 2012. Nucleosome organization affects the sensitivity of gene expression to promoter mutations. Mol Cell. 46:362-368.
Hurst LD, Smith, NGC. 1999. Molecular evolutionary evidence that H19 mRNA is functional. Trends Genet. 15:134-135.
Hurst LD, Williams, EJ, Pál, C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. 18:604-606.
Janga S, Collado-Vides, J, Babu, M. 2008. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. Proc Natl Acad Sci USA. 105:15761-15766.
Janicki SM, Tsukamoto, T, Salghetti, SE, Tansey, WP, Sachidanandam, R, Prasanth, KV, Ried, T, Shav-Tal, Y, Bertrand, E, Singer, RH. 2004. From silencing to gene expression: real-time analysis in single cells. Cell. 116:683-698.
Képès F. 2003. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. J Mol Biol. 329:859-865.
Khaitovich P, Muetzel, B, She, X, Lachmann, M, Hellmann, I, Dietzsch, J, Steigele, S, Do, HH, Weiss, G, Enard, W, et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. Genome Res. 14:1462-1473.
Lee J, Sonnhammer, E. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13:875-882.
Lee JM, Sonnhammer, EL. 2004. Genomic gene clustering analysis of pathways in eukaryotes (vol 13, pg 875, 2003). Genome Res. 14:2510-2510.
Lercher MJ, Blumenthal, T, Hurst, LD. 2003. Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 13:238-243.
Lercher MJ, Hurst, LD. 2006. Co-expressed yeast genes cluster over a long range but are not regularly spaced. J Mol Biol. 359:825-831.
Lercher MJ, Urrutia, AO, Hurst, LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180-183.
Li B, Carey, M, Workman, JL. 2007. The role of chromatin during transcription. Cell. 128:707-719.
Liao BY, Zhang, J. 2008. Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Mol Biol Evol. 25:1555-1565.

Liu C, Ghosh, S, Searls, DB, Saunders, AM, Cossman, J, Roses, AD. 2005. Clusters of adjacent and similarly expressed genes across normal human tissues complicate comparative transcriptomic discovery. OMICS: J Integrative Biol. 9:351-363.
Marfella CG, Imbalzano, AN. 2007. The Chd family of chromatin remodelers. Mutat Res. 618:30-40.
Marques AC, Ponting, CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol. 10:R124.
Michalak P. 2008. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 91:243-248.
Mijalski T, Harder, A, Halder, T, Kersten, M, Horsch, M, Strom, TM, Liebscher, HV, Lottspeich, F, de Angelis, MH, Beckers, J. 2005. Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. Proc Natl Acad Sci USA. 102:8621-8626.
Molineris I, Grassi, E, Ala, U, Di Cunto, F, Provero, P. 2011. Evolution of promoter affinity for transcription factors in the human lineage. Mol Biol Evol. 28:2173-2183.
Nutzmann HW, Osbourn, A. 2014. Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 26:91-99.
Oliver B, Misteli, T. 2005. A non-random walk through the genome. Genome Biol. 6:214.
Pal C, Hurst, LD. 2003. Evidence for co-evolution of gene order and recombination rate. Nat Genet. 33:392-395.
Pang KC, Frith, MC, Mattick, JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22:1-5.
Persson J, Ekwall, K. 2010. Chd1 remodelers maintain open chromatin and regulate the epigenetics of differentiation. Exp Cell Res. 316:1316-1323.
Poyatos J, Hurst, L. 2007. The determinants of gene order conservation in yeasts. Genome Biol. 8:R233.
Purmann A, Toedling, J, Schueler, M, Carninci, P, Lehrach, H, Hayashizaki, Y, Huber, W, Sperling, S. 2007. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics. 89:580-587.
Reik W, Walter, J. 2001. Genomic imprinting: parental influence on the genome. Nat Rev Genet. 2:21-32.
Rosin D, Hornung, G, Tirosh, I, Gispan, A, Barkai, N. 2012. Promoter nucleosome organization shapes the evolution of gene expression. PLoS Genet. 8:e1002579.
Roy PJ, Stuart, JM, Lund, J, Kim, SK. 2002. Chromosomal clustering of muscle-expressed genes in caenorhabditis elegans. Nature. 418:975-979.
Rutschmann F. 2006. Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. Diversity and Distributions. 12:35-48.
Sémon M, Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 23:1715-1723.
Sims RJ, 3rd, Millhouse, S, Chen, CF, Lewis, BA, Erdjument-Bromage, H, Tempst, P, Manley, JL, Reinberg, D. 2007. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 28:665-676.

Singer GA, Lloyd, AT, Huminiecki, LB, Wolfe, KH. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol. 22:767-775.
Spellman PT, Rubin, GM. 2002. Evidence for large domains of similarly expressed genes in the drosophila genome. J Biol. 1:5.
Sproul D, Gilbert, N, Bickmore, WA. 2005. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet. 6:775-781.
Stolc V, Gauhar, Z, Mason, C, Halasz, G, van Batenburg, MF, Rifkin, SA, Hua, S, Herreman, T, Tongprasit, W, Barbano, PE, et al. 2004. A gene expression map for the euchromatic genome of drosophila melanogaster. Science. 306:655-660.
Symmons O, Uslu, VV, Tsujimura, T, Ruf, S, Nassari, S, Schwarzer, W, Ettwiller, L, Spitz, F. 2014. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 24:390-400.
Thurman RE, Rynes, E, Humbert, R, Vierstra, J, Maurano, MT, Haugen, E, Sheffield, NC, Stergachis, AB, Wang, H, Vernot, B, et al. 2012. The accessible chromatin landscape of the human genome. Nature. 489:75-82.
Tirosh I, Barkai, N, Verstrepen, KJ. 2009. Promoter architecture and the evolvability of gene expression. J Biol. 8:95.
Tirosh I, Weinberger, A, Carmi, M, Barkai, N. 2006. A genetic signature of interspecies variations in gene expression. Nat Genet. 38:830-834.
Uesaka M, Nishimura, O, Go, Y, Nakashima, K, Agata, K, Imamura, T. 2014. Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics. 15:35.
Versteeg R, van Schaik, BD, van Batenburg, MF, Roos, M, Monajemi, R, Caron, H, Bussemaker, HJ, van Kampen, AH. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. 13:1998-2004.
Wang Y, Rekaya, R. 2009. A comprehensive analysis of gene expression evolution between humans and mice. Evol Bioinform Online. 5:81.
Weber CC, Hurst, LD. 2011. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. Genome Biol. 12:R23.
Wei W, Pelechano, V, Jarvelin, AI, Steinmetz, LM. 2011. Functional consequences of bidirectional promoters. Trends Genet. 27:267-276.
Williams EJB, Bowles, DJ. 2004. Coexpression of neighboring genes in the genome of arabidopsis thaliana. Genome Res. 14:1060-1067.
Wittkopp PJ, Kalay, G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 13:59-69.
Woo YH, Li, W-H. 2011. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. Proc Natl Acad Sci USA. 108:3306-3311.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8:206-216.
Yang H, Li, D, Cheng, C. 2014. Relating gene expression evolution with CpG content changes. BMC Genomics. 15:693.

Zentner GE, Tsukiyama, T, Henikoff, S. 2013. ISWI and CHD chromatin remodelers bind promoters but act in gene bodies. PLoS Genet. 9:e1003317.

Chapter 2. Piggybacking neighbours, Part 1: Evolution of gene expression in Primates

MBE Advance Access published April 2, 2015

Neighboring Genes Show Correlated Evolution in Gene Expression

Avazeh T. Ghanbarian[1] and Laurence D. Hurst[1,*]
[1]Department of Biology and Biochemisty, University of Bath, Bath, United Kingdom
***Corresponding author:** E-mail: l.d.hurst@bath.ac.uk.
Associate editor: Gunter Wagner

Abstract

When considering the evolution of a gene's expression profile, we commonly assume that this is unaffected by its genomic neighborhood. This is, however, in contrast to what we know about the lack of autonomy between neighboring genes in gene expression profiles in extant taxa. Indeed, in all eukaryotic genomes genes of similar expression-profile tend to cluster, reflecting chromatin level dynamics. Does it follow that if a gene increases expression in a particular lineage then the genomic neighbors will also increase in their expression or is gene expression evolution autonomous? To address this here we consider evolution of human gene expression since the human-chimp common ancestor, allowing for both variation in estimation of current expression level and error in Bayesian estimation of the ancestral state. We find that in all tissues and both sexes, the change in gene expression of a focal gene on average predicts the change in gene expression of neighbors. The effect is highly pronounced in the immediate vicinity (<100 kb) but extends much further. Sex-specific expression change is also genomically clustered. As genes increasing their expression in humans tend to avoid nuclear lamina domains and be enriched for the gene activator 5-hydroxymethylcytosine, we conclude that, most probably owing to chromatin level control of gene expression, a change in gene expression of one gene likely affects the expression evolution of neighbors, what we term expression piggybacking, an analog of hitchhiking.

Key words: **gene expression evolution, gene clustering, sex-biased evolution.**

Introduction

Work on the evolution of gene expression has commonly been gene centric, concentrating on, for example, changes in the promoter elements of a given gene (Hammock and Young 2002; Carninci et al. 2006; Tirosh et al. 2006; Wray 2007; Tirosh et al. 2009; Wang and Rekaya 2009; Molineris et al. 2011; Hornung et al. 2012; Rosin et al. 2012; Wittkopp and Kalay 2012; Forrest et al. 2014; Yang et al. 2014). In such a model, changes in the promoter change the expression of the gene controlled by that promoter but nothing else (baring downstream effects of, for example, up- or downregulation of a transcription factor). But are genes autonomous in their evolution in the sense that the change in expression of a focal gene has no effects on its immediate genomic neighbors? In contrast to such an autonomous view of gene expression evolution, when examining profiles of gene expression across chromosomes, it is now evident that in eukaryotes genes of similar expression tend to cluster (Cho et al. 1998; Cohen et al. 2000; Caron et al. 2001; Reik and Walter 2001; Blumenthal et al. 2002; Hurst et al. 2002; Roy et al. 2002; Spellman and Rubin 2002; Birnbaum et al. 2003; Lee and Sonnhammer 2003; Lercher et al. 2003; Versteeg et al. 2003; Khaitovich et al. 2004; Stolc et al. 2004; Williams and Bowles 2004; Denver et al. 2005; Liu et al. 2005; Mijalski et al. 2005; Oliver and Misteli 2005; Singer et al. 2005; Sproul et al. 2005; Lercher and Hurst 2006; Sémon and Duret 2006; Purmann et al. 2007; Ebisuya et al. 2008; Nutzmann and Osbourn 2014). This is seen both at a fine scale and a more gross chromosomal scale (Cohen et al. 2000; Caron et al. 2001; Lercher et al. 2003; Pal and Hurst 2003; Williams and Bowles 2004; Purmann et al. 2007; Michalak 2008; Woo and Li 2011). On a fine scale, neighboring genes tend to be coexpressed more than expected by chance across multiple taxa (Blumenthal et al. 2002; Boutanaev et al. 2002; Roy et al. 2002; Lercher et al. 2003; Fukuoka et al. 2004; Williams and Bowles 2004; Purmann et al. 2007; Davila Lopez et al. 2010), the effect being most pronounced often for genes in a bidirectional orientation, in which promoters sit in close proximity to each other (Cohen et al. 2000; Williams and Bowles 2004; Davila Lopez et al. 2010; Wei et al. 2011; Uesaka et al. 2014). On a more gross scale, genes expressed in most tissues (housekeeping genes) and highly expressed genes tend to cluster in domains corresponding to tens of genes (Caron et al. 2001; Lercher et al. 2002; Versteeg et al. 2003; Weber and Hurst 2011).

Although genes controlled by the same transcription factors are themselves not randomly organized, at least not in yeast (Képès 2003; Janga et al. 2008), in large part broad and narrow span clustering tendencies probably reflect chromatin dynamics rather than shared transcription factors (Grunstein 1997; Cohen et al. 2000; Sémon and Duret 2006; Batada et al. 2007; Li et al. 2007). In yeast, for example, controlling for transcription factor similarity neighboring genes still show striking similarity in coexpression (Batada et al. 2007). Similarly, in mammals, incorporation of transgenes into chromosomes demonstrates that these adopt the expression

profile of neighbors within a broad span (Gierman et al. 2007; Symmons et al. 2014). In both yeast and mammals, the upregulation of one gene causes time-lagged ripples of gene expression that correspond to changes in chromatin state (Cohen et al. 2000; Janicki et al. 2004; Ebisuya et al. 2008). In humans these ripple domains are around 100 kb in size (Ebisuya et al. 2008). Whether the fact of clusters of gene expression implies selection for such clusters is unresolved. In yeast, the most highly coexpressed gene pairs tend to be more similar in functionality and more commonly conserved as a pair (Hurst et al. 2002; Poyatos and Hurst 2007). However, results in other lineages are less decisive (Lee and Sonnhammer 2003, 2004; Liao and Zhang 2008; Weber and Hurst 2011).

Here we ask whether genes are autonomous in their expression evolution. To this end we consider RNASeq data for several tissues in male and female primates. Reconstructing the human–chimp ancestral state permits us to estimate the extent of expression change between humans and this ancestor and represent this as a *Z* score that factors in both current variation in expression between replicates (expression or measurement noise) and uncertainty in ancestral state reconstruction. We then consider the extent to which neighboring genes show correlated *Z* scores. Under the null that genes are autonomous in their expression evolution the correlation in *Z* score between neighbors should be zero. In addition, by considering the residuals of the orthogonal regression of *Z* for a gene in a given tissue in males against the same in females we can define the degree of sex bias in expression change. We can thus in turn ask whether this too shows evidence of autonomy.

Results

Neighboring Genes Are Correlated in the Expression Change in All Tissues in Both Sexes

So as to gauge what the possible mechanisms might be, we considered several methods to ask whether the expression change of a focal gene (*Z*) is correlated with that of its neighbors. In the first instance we consider for each gene (regardless of which strand they reside on) the nearest neighbor downstream of the focal gene (downstream here is by reference to the published chromosomal strand not to the orientation of the gene), allowing only those instances where the intergene distance is less than 100 kb, this being the estimated size of the ripple effect (Ebisuya et al. 2008), wherein upregulation of one gene causes a time-lagged upregulation of the neighbors (the ripple). In the second instance we consider the correlation between a focal gene and its nearest pair of neighbors, one upstream one downstream, assuming both were within 100 kb (this is comparable to the first method but could be less noisy). In this instance we take the mean *Z* of the neighbors. In the third, we considered for each focal gene the mean *Z* of all neighbors within 100 kb. While the first method might be detecting immediate and local interactions between any given gene pair (e.g., mediated by bidirectional promoters), the latter most likely recovers broader scale chromatin effects. Under the premise that we must be missing the site of expression, we excluded genes with *Z* (prior to modification—see Materials and Methods) of zero owing to lack of expression in a given tissue. In the first and second cases we consider only nonoverlapping genes. For the third case, if the focal gene overlaps any of its adjacent neighbors, it is removed from the analysis; but if there are nonfocal overlapping genes in the neighborhood, they are included.

Strikingly we find that for all tissues in both sexes, all analyses report a highly significant positive correlation between *Z* of focal genes and *Z* of neighbors (fig. 1, tables 1–3). The correlation stays highly significant and in positive direction if one is to consider fold change since ancestor instead of *Z* score (supplementary table S1, Supplementary Material online). Note too that our correction of *Z* to a median of zero is here irrelevant as our statistics are based on rank ordering. These results strongly supports the hypothesis that gene evolution is nonautonomous, or at least that it occurs on a cluster-by-cluster basis. We note too that our *Z* scores accord well with the metric to define significantly changed expression employed by Brawand et al (2011) (supplementary fig. S1 and table S2, Supplementary Material online).

While the earlier results provide evidence of clustering it does not identify clusters nor does it suggest their dimension. As alternative means to test for clustering and to identify unusually large clusters, we consider the number of switches in *Z* score as one runs along a chromosome. We represent all genes as having a positive, negative, or zero Z score. Those with a zero we consider to be too indecisive to be permitted for this test so are excluded. We then consider, running down each chromosome, the number and lengths of spans with uniform *Z* sign. That is we ask about the size of runs of positive and negative *Z* scores (*Z*+ and *Z*− we then consider as states + and −). To address whether there are fewer but larger runs than expected (clustering) we ask about the number of edges of runs. A series +++−+++ for example has two edges, a + to − switch and a − to + switch. We then compare the observed genomic number of switches to the number expected under a null of random ordering. The null is derived from randomisation of character states (i.e., loci) within each chromosome, thus preserving the absolute number of + and − genes on each chromosome. For all tissues in both sexes, we observe that the observed number of clusters is lower than expected; hence, their length is greater than expected ($P < 0.0001$ in all cases). Put differently, longer runs of uniform expression change are more commonly observed than expected by chance and shorter runs are less common (fig. 2). The largest clusters even by this conservative definition (a single gene of opposite sign breaks a cluster) run to tens of genes. For illustration of some very large clusters, see supplementary figure S2*a* and *b*, Supplementary Material online. This result provides further evidence that our core result, the clustering of genes showing similar change in expression is largely immune to assumption about the precise metric of change, it being seen with *Z* metric (tables 1–3), fold change (supplementary table S1, Supplementary Material online), and digital parametrization (fig. 2).

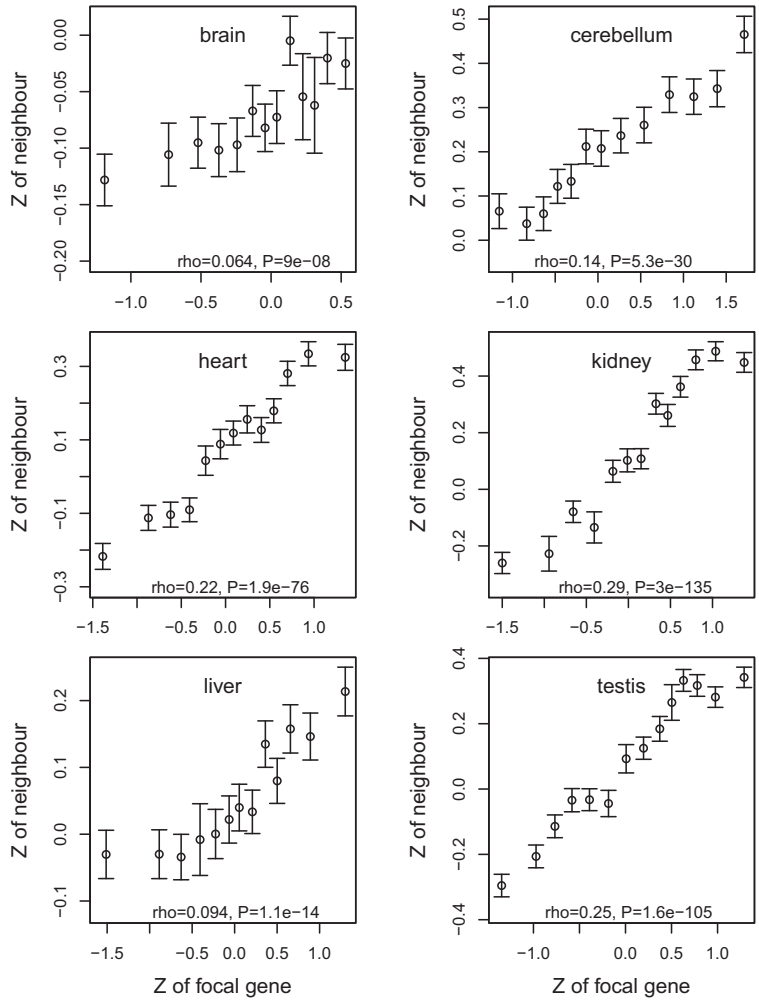


Fig. 1. Relationship between *Z* of a focal gene and *Z* of the nearest downstream neighbor for six male tissues. In this instance we consider all genes are nearest downstream neighbors if the distance between the start codons is <100 kb. This slightly contrasts with data in table 1, where the distance is defined as minimum distance between gene bodies. Trends are robust to alternative definitions. Data are split into equal sized bins (of 500 genes) defined after rank ordering with respect to *Z* score of the focal gene. The value on the *X* axis represents the mean *Z* of the genes in that bin. The value of the *Y* axis indicates the mean (±SEM) for the relevant flanking genes. The presented statistics are from Spearman correlation on raw data.

Weak Evidence Only That Gene Orientation Is Relevant to Correlated Change in Gene Expression

When considering the correlation between a focal gene and the nearest neighbour, we ignored any effects of orientation between the neighbor and the focal gene. Prior work has suggested that genes in divergent orientation may be particular in the extent of coupling in their expression (Wright et al. 1995; Cho et al. 1998; Cohen et al. 2000; Kruglyak and Tang 2000; Hurst et al. 2002; Trinklein et al. 2004; Williams and Bowles 2004; Woo and Li 2011; Wakano et al. 2012). This may be for no better reason that genes in divergent orientation will have a lower distance between their promoters (Wakano et al. 2012), all else being equal. Genes sharing bidirectional promoters are, under this model, the most highly coupled. Do we then see any effect of the correlation between *Z* scores as a function of orientation?

For every focal gene and its unique nearest downstream neighbour, we consider the two to be in one of three orientations: divergent (<– >), convergent (- > <-) and cooriented (- > - > or <-<-). For each of the three classes we calculated the Spearman's ρ value for the correlation of *Z* scores between the neighbors, this being repeated for each tissue in each sex (table 4). Very weakly suggestive of a greater coordination of genes in divergent orientation,

Table 1. Spearman Correlation between Focal Gene's *Z* Score and *Z* Score of Its Closest Nonoverlapping Downstream Neighbor.

Tissue	Male *P* Value	Male ρ	Female *P*-Value	Female ρ
Brain	8.71E−07	0.05504	2.81E−08	0.06247
Cerebellum	1.71E−19	0.10246	9.25E−21	0.10539
Kidney	3.97E−126	0.26420	3.37E−07	0.05751
Heart	4.13E−66	0.19308	7.14E−20	0.10423
Liver	5.91E−12	0.07786	NA	NA
Testis	6.92E−83	0.21132	NA	NA

NOTE.—All statistics are significant after Bonferroni testing.

Table 2. Spearman Correlation between Focal Gene's *Z* Score and Mean of Its Closest Nonoverlapping Neighbors on Both Sides.

Tissue	Male *P*-Value	Male ρ	Female *P*-Value	Female ρ
Brain	2.95E−10	0.08015	8.70E−12	0.08727
Cerebellum	1.96E−31	0.15009	1.51E−33	0.15433
Kidney	1.44E−155	0.33054	6.07E−10	0.07925
Heart	2.03E−86	0.24993	2.16E−28	0.14318
Liver	8.86E−17	0.10676	NA	NA
Testis	4.43E−118	0.28520	NA	NA

NOTE.—All statistics are significant after Bonferroni testing.

Table 3. Spearman Ranked Correlation of *Z* Score of Focal Gene with Mean *Z* Score of All Its Nonoverlapping Neighboring (within ±100 kb) Genes.

Tissue	Male *P*-value	Male ρ	Female *P*-value	Female ρ
Brain	7.75E−08	0.04780	6.93E−17	0.07465
Cerebellum	8.67E−61	0.14784	1.17E−41	0.12111
Kidney	1.32E−274	0.30926	2.81E−15	0.07078
Heart	8.82E−160	0.23968	2.07E−44	0.12681
Liver	8.51E−26	0.09458	NA	NA
Testis	6.27E−187	0.25247	NA	NA

NOTE.—All statistics are significant after Bonferroni testing.

we find that in 6 of 10 incidences the divergent orientation genes have the highest ρ value (these being male liver, brain and testis, and female kidney, heart, and cerebellum). Assuming that the divergent orientation should have the highest ρ value one-third of the time, a 6:4 split is not significant (two-tailed, binomial test $P = 0.094$; one-tailed binomial test $P = 0.076$).

To check whether the three Spearman's ρ values (for each tissue for each sex) differed from ρ score of a randomly selected subset of the same size, we performed Monte Carlo randomizations. Each simulation extracted the appropriate but randomly selected number of gene neighbors using the same underlying data (i.e., same tissue, same sex). Each simulation was repeated 10,000 times. The ρ score of each random sample was calculated and compared with that observed in the simulants to determine *P* (Materials and Methods). We find that in two incidences (male testis and female cerebellum) genes in divergent orientation have a significantly higher ($P < 0.05$) correlation in the *Z* scores than expected by chance (table 5). The effects are, however, marginal ($0.01 < P < 0.05$) and not robust to Bonferroni correction.

Prior evidence suggests that bidirectional orientation may have its most profound influence at the sub 1 kb scale (Hurst et al. 2002; Li et al. 2006; Franck et al. 2008), although another study found a marginally lower correlation among divergent genes at 1 kb distance (Takai and Jones 2004). Unfortunately there are few genes in the sample at such proximity. Nonetheless we can repeat the analyses above on this more limited subset. We observe that in five incidences (male brain, male kidney, male liver, female cerebellum, female kidney) divergent orientation records the highest ρ value, again not a significant difference (table 6). Weak significance from Monte Carlo simulations is observed in only one case (male liver), again not robust to Bonferroni correction (table 7). We conclude that we see weak, at best, evidence that gene orientation has an influence on the degree of correlated expression change.

Overlapping Genes Are the Most Strongly Positively Correlated in Expression Change

Thus far we excluded from consideration overlapping genes. A priori we might expect these to behave differently, not least because simultaneous expression of both genes might lead to transcriptional interference (Noguchi et al. 1994; Prescott and Proudfoot 2002; Osato et al. 2007). Hence upregulation of one might force downregulation of the other, if only through forcing premature transcriptional termination. Alternatively, upregulation of one might make the chromatin environment of the promoter of the neighbor even more likely to be accessible, so proving an even stronger signal of nonautonomous evolution.

While the original data set (Brawand et al. 2011) was specified as excluding all incidences in which genes overlap within their protein coding sequence, many overlap in their full-length transcript. Examining these we find that the nearest neighbors still show a strong positive correlation in *Z* scores (tables 8 and 9). Indeed, in all cases, the correlation is stronger for the overlapping genes than for the nearest nonoverlapping neighbor. Assuming each sample to be independent, the probability of such agreement is low (binomial test, $P = 0.002$). However, all samples are not independent (male and female expression change correlates—see later). Thus to evaluate whether the strength of this correlation was any different to that expected for any pair of nearest downstream neighbors, we repeatedly extracted from the larger set of nonoverlapping neighbors a random subset of the nearest downstream neighbors. The random subsets had the same number of genes as seen in the overlapping genes' set. We then asked how often we see a ρ value as great or greater than that observed for the overlapping case. Overlapping genes had consistently stronger correlation than the nonoverlapping gene sets in all tissues in both sexes (table 10). These results support the view that close proximity, possibly owing to a greater likelihood of shared chromatin environment, is a more important determinant

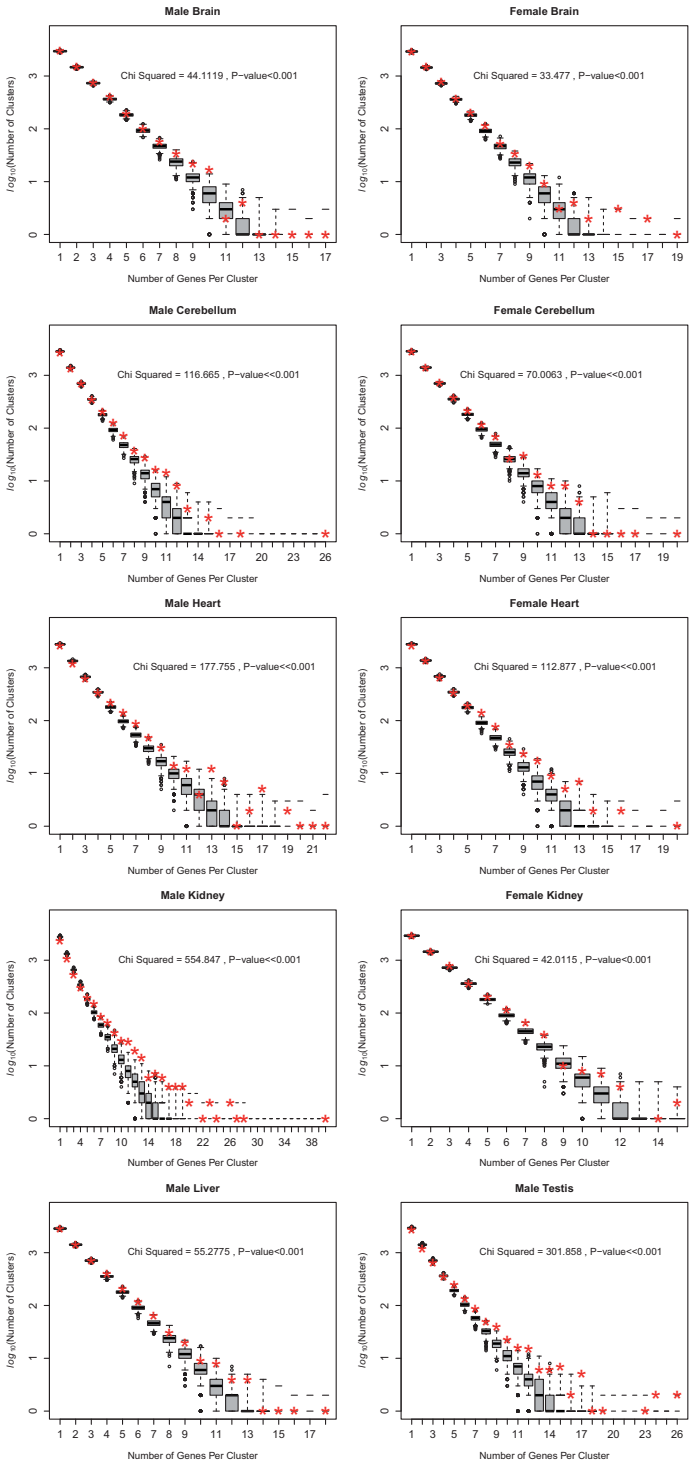


Fig. 2. Numbers of clusters of a given size compared to that expected under a random null. Observed number of clusters including certain number of genes is shown by red stars, boxplots show variation across number of clusters in 1,000 random sets.

Table 4. Spearman Correlation between Z of Divergent, Convergent, and Cooriented Closest Gene Pairs.

Tissue/Gender	Divergent *P-Value*	Divergent ρ	Convergent *P-Value*	Convergent ρ	Cooriented *P-Value*	Cooriented ρ
Brain/male	*0.000474*	0.07738	0.105396	0.03449	*0.00031*	0.05483
Cerebellum/male	*1.76E−05*	0.09616	*2.27E−07*	0.11123	*7.63E−13*	0.11086
Kidney/male	*1.76E−35*	0.27214	*9.33E−30*	0.23963	*5.83E−78*	0.27992
Heart/male	*4.52E−18*	0.19287	*3.26E−16*	0.17496	*5.52E−42*	0.20693
Liver/male	*8.23E−07*	0.11054	*0.008186*	0.05694	*1.07E−06*	0.07485
Testis/male	*3.47E−30*	0.24745	*1.24E-22*	0.20458	*3.24E-42*	0.20261
Brain/female	*0.003130*	0.06569	*0.000440*	0.07510	*0.00107*	0.05010
Cerebellum/female	*3.01E-10*	0.14002	*0.000271*	0.07796	*1.24E-10*	0.09887
Kidney/female	0.004371	0.06349	0.032372	0.04601	*0.000205*	0.05684
Heart/female	*4.75E-09*	0.13200	*3.77E-06*	0.10040	*2.52E-11*	0.10359

NOTE.—Results significant after Bonferroni testing are highlighted in italic.

Table 5. *P*-Values of Monte Carlo Simulations Comparing Spearman's Correlation ρ Score between Z Score of Focal Gene and Z Score of Its Downstream Neighbor across Divergent, Convergent, and Cooriented Subsets against ρ of a Randomly Selected Set of Genes of the Same Size as Those Subsets.

Tissue	Divergent Male *P*-Value	Convergent Male *P*-Value	Cooriented Male *P*-Value	Divergent Female *P*-Value	Convergent Female *P*-Value	Cooriented Female *P*-Value
Brain	0.12059	0.87421	0.86861	0.37086	0.20748	0.20998
Cerebellum	0.70893	0.40776	0.40526	0.03330	0.91901	0.92151
Kidney	0.41026	0.94881	0.95150	0.36286	0.70813	0.71763
Heart	0.55744	0.86571	0.86821	0.12109	0.68713	0.67243
Liver	0.05359	0.88301	0.88571	NA	NA	NA
Testis	0.03550	0.72293	0.71803	NA	NA	NA

NOTE.—If the number of genes in divergent orientation, for example, after removing zero Z scores in a specific tissue and sex is shown by ts*ND* and Spearman's correlation's ρ score between those focal genes and their divergent downstream is shown by tsρ. Then ρ score of 10,000 random sets of linked gene pairs of ts*ND* size, selected from pool of all genes in this study regardless of their orientation, is calculated and compared with tsρ in corresponding tissue/gender. If the number of random sets with their ρ great or greater than tsρ is shown by *M*, Monte Carlo *P*-values are then calculated as (*M*+1)/10,001. No observations are significant after Bonferroni testing.

Table 6. Spearman Correlation between Z Score of Focal Gene and Z Score of Its Closest Downstream Neighbor across Divergent, Convergent, and Cooriented Closest Gene Pairs Which Are Closer than 1 kb.

Tissue/Gender	Divergent *P*-value	Divergent ρ	Convergent *P*-value	Convergent ρ	Cooriented *P*-value	Cooriented ρ
Brain/male	0.10085	0.08288	0.81912	0.01280	0.95651	−0.00366
Cerebellum/male	0.01006	0.13001	0.01738	0.13288	0.02453	0.15090
Kidney/male	*7.07E−16*	0.39189	*1.30E-08*	0.31211	0.00327	0.19567
Heart/male	*7.80E-06*	0.22392	*7.79E-09*	0.31661	0.00752	0.17813
Liver/male	*0.00044*	0.17669	0.20270	0.07196	0.69872	0.02606
Testis/male	*1.02E-11*	0.33586	*1.49E-10*	0.34886	0.04807	0.13197
Brain/female	0.36058	0.04629	0.86790	−0.00929	0.43382	0.05267
Cerebellum/female	*1.32E-05*	0.21838	0.00461	0.15838	0.05900	0.12635
Kidney/female	0.12010	0.07853	0.64196	−0.02613	0.72420	−0.0237
Heart/female	0.00250	0.15248	0.00302	0.16604	0.02574	0.14933

NOTE.—Results significant after Bonferroni testing are highlighted in italic.

of coupled gene expression change than is transcriptional interference or gene orientation.

A Ripple Effect Cannot Explain the Dimensions of the Expression Change Clusters

Although the earlier more extreme correlation in changes at very small distances is potentially consistent with the ripple effect, this same effect suggests that expression clusters should be of ~100 kb in magnitude (Ebisuya et al. 2008). To estimate physical cluster size, we consider the strength of the correlation between genes in their Z score as a function of the distance between them. We consider all focal genes and the correlation between Z scores for these genes and the nearest downstream gene at a minimum of *x* base pairs away. By incrementing the minimum distance of *x*, we can then ask at what physical distance on average is ρ between the focal genes and nearest "neighbors" is less than the mean $\pm$ 1.96 SD of 1,000 randomized null sets.

For three tissues (heart, kidney, testes), the data appear to be relatively noise free, suggesting the span of local correlation to extend up to tens of megabytes (10–25 MB) (fig. 3*a*). For

Table 7. *P*-Values of Monte Carlo Simulation Comparing Spearman's Correlation ρ Score between Focal Gene and Its Downstream Neighbor across Divergent, Convergent, and Coordinated Subsets to a Randomly Selected Subset of the Same Size for Gene Pairs Closer than 1 kb.

Tissue	Divergent Male *P*-Value	Convergent Male *P*-Value	Cooriented Male *P*-Value	Divergent Female *P*-Value	Convergent Female *P*-Value	Cooriented Female *P*-Value
Brain	0.13399	0.71823	0.72053	0.33787	0.79582	0.79852
Cerebellum	0.64264	0.60364	0.59444	0.33907	0.85431	0.84622
Kidney	0.17848	0.87671	0.87581	0.07129	0.84862	0.84202
Heart	0.91831	0.15938	0.15298	0.78032	0.62664	0.63754
Liver	0.02850	0.76262	0.75932	NA	NA	NA
Testis	0.57334	0.42326	0.43336	NA	NA	NA

NOTE.—Monte Carlo simulation's steps and number of repetition are the same as explained in table 5. No observation is significant after Bonferroni testing.

Table 8. Spearman Correlation between Focal Gene's *Z* Scores and *Z* of Its Overlapping Downstream Neighbor on the Opposite Strand.

Tissue	Male *P*-value	Male ρ	Female *P*-value	Female ρ
Brain	0.00392	0.10783*	0.00368	0.10886*
Cerebellum	8.37E−14	0.27613*	8.45E−06	0.16696*
Kidney	2.75E−26	0.38295*	0.01655	0.08992*
Heart	4.90E−15	0.28986*	1.18E−06	0.18234*
Liver	0.00019	0.13979*	NA	NA
Testis	<2.2E−16	0.3942*	NA	NA

NOTE.—Those incidences marked with an asterisk have a higher correlation than seen in the comparable nonoverlapping case (shown in table 1). All observations are significant after Bonferroni testing. As the underlying data are strand-specific transcriptomics, employing overlapping sequence from opposite strands obviates problems with mismapping, causing artifactual signals of high correlation.

Table 9. Spearman Correlation between Focal Gene's *Z* Scores and Mean of Its Closest Up and Downstream Neighbors, at Least One of Which Overlaps the Focal Gene.

Tissue	Male *P*-Value	Male ρ	Female *P*-value	Female ρ
Brain	0.00013	0.11001*	0.0002	0.10724*
Cerebellum	1.18E−24	0.29169*	1.52E−11	0.19365*
Kidney	<2.2E−16	0.41596*	0.00126	0.09303*
Heart	2.93E−29	0.31778*	4.58E−13	0.20841*
Liver	7.60E−07	0.14236*	NA	NA
Testis	<2.2E−16	0.4018*	NA	NA

NOTE.—Those incidences marked with an asterisk have a higher correlation than seen in the comparable nonoverlapping case (shown in table 2). All observations are significant after Bonferroni testing.

Table 10. Monte Carlo Simulation of Overlapping Genes' Z.

Tissue	Male *P*-Value	Female *P*-Value
Brain	0.005999	0.0095
Cerebellum	0.000099	0.003
Kidney	0.000099	0.0132
Heart	0.000499	0.0004
Liver	0.007399	NA
Testis	0.000099	NA

NOTE.—Comparing Spearman correlation's ρ score of overlapping genes against randomly selected set of gene pairs of the same size over 1,000 repetitions. The number of incidents when ρ of randomly selected set is equal or higher than ρ in overlapping set was counted to calculate empirical *P*-values. All observations are significant after Bonferroni testing.

the remaining three, brain suggests a much more limited domain, while cerebellum and liver are consistent with ~10 MB span. Looking in more details at trends under 1 MB from the focal genes (fig. 3*b*), we observe that all tissues report the local correlation of *Z* to be most profound under 100 kb, with brain tissue indeed, suggesting this to be the upper limit. The discrepancy between brain and the other tissues might, we suggest be owing to heterogeneity in sampling procedures and intrinsic heterogeneity of brain tissue. A ripple effect (Ebisuya et al. 2008) that extends over ~100 kb might be able to explain the intensity of the signal at such short range (fig. 3*b*) (notice the nonlinear trends seen in 3*b* and the extent to which the left most data point in 3*a* appears as an outlier). The ripple effect appears, however, to be incompatible with the much longer-range effects as these extend in many cases well beyond the 100 kb limit of the ripple effect.

Changes in Gene Expression Accord with Lamina Domains and 5-Hydroxymethylcytosine

Do the genes changing expression accord with any chromatin signatures? Nuclear compartmentalization and lamina-associated chromatin domains (LADs) in particular have been shown to be involved in regulating genes in Metazoan (Reddy et al. 2008; Van Bortle and Corces 2013). Moreover, recent analysis of gene disregulation in Downs syndrome suggested that LADs represent a level of expression organization in the human genome (Letourneau et al. 2014). LADs have also been shown to associate with low gene expression (Guelen et al. 2008). Hence LADs would provide a good measure for investigating chromatin level regulation's involvement in evolution of gene expression. Using a high-resolution map of LADs in fibroblast (Guelen et al. 2008), we find that in all six tissues genes residing in putative lamina domains tend to have lower *Z* scores than those not in lamina domains (fig. 4 [before multitest correction, Mann–Whitney *U* test $P < 10^{-9}$ except brain $P = 4 \times 10^{-4}$]). Thus increases in expression level tend to be outside of lamina domains.

5-Hydroxymethyl cytosine (hmC) and 5 methylcytosine (mC) are also involved in chromatin level regulation of gene expression through recruiting chromatin modifiers (Mellen et al. 2012; Spruijt et al. 2013). Recent evidence also indicates that gene activity is associated with hmC on the coding strand (Wen et al. 2014). Inactive genes or noncoding

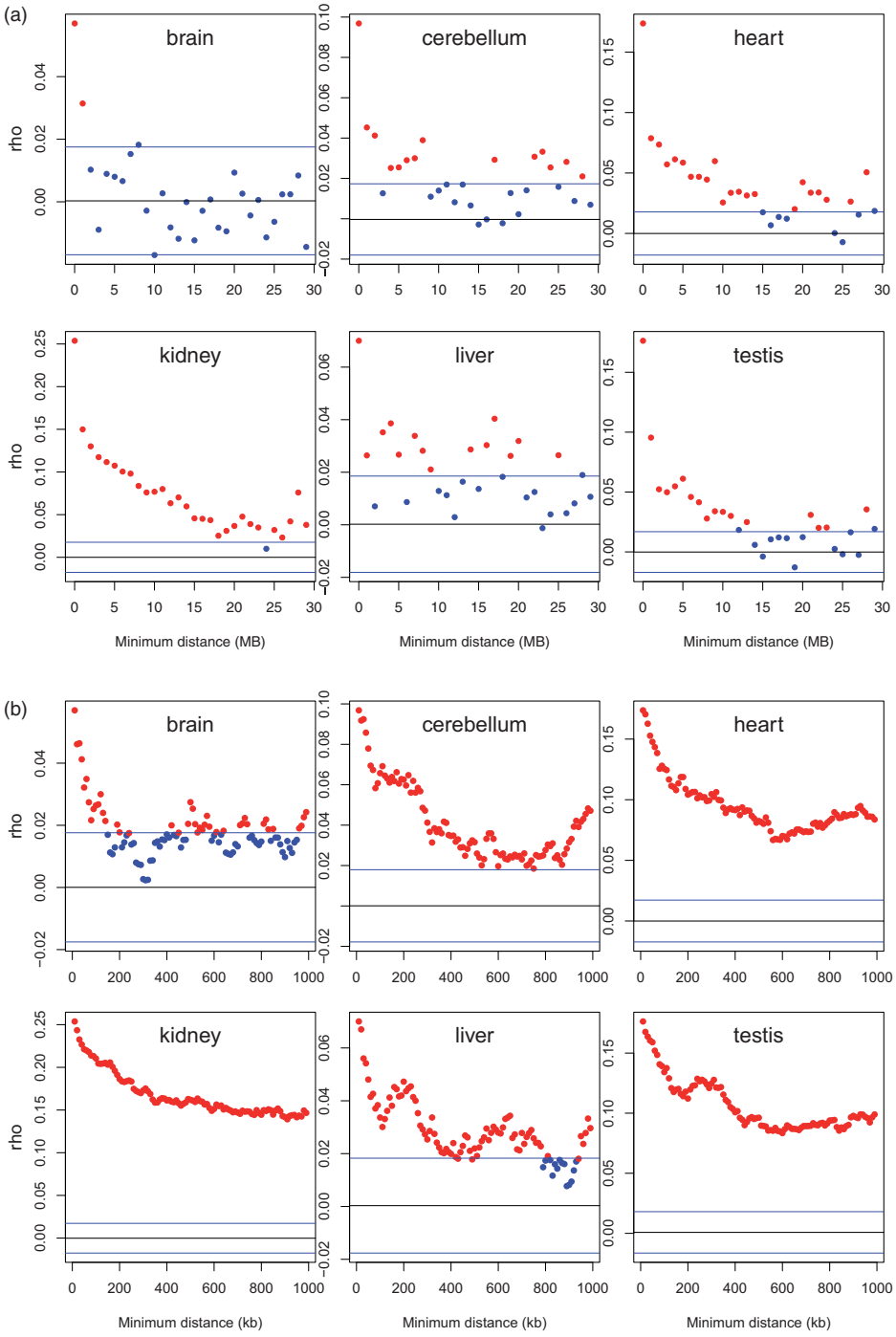


Fig. 3. Correlation between *Z* of each focal gene and *Z* of nearest downstream neighbor more than a given minimum physical distance away. (*a*) We plot data considering increments of minimum distance 1 MB at a time up to a maximum of 30 MB. (*b*) We consider 10-kb increments up to a maximum of 1 MB. For each focal gene we extract the nearest neighbor downstream that is at least the distance *x* away, *x* being the units on the *x* axis. From a list of focal and neighbor *Z* scores, we consider then the correlation between these. Correlations significant at the 0.05 level are shown in red, otherwise in blue. The blue horizontal lines indicate 1.96 SD limits determined by randomization (which should in principle correspond with the *P* from Spearman's ρ), with the black line indicating mean of null expectation from randomization (which should be around zero).

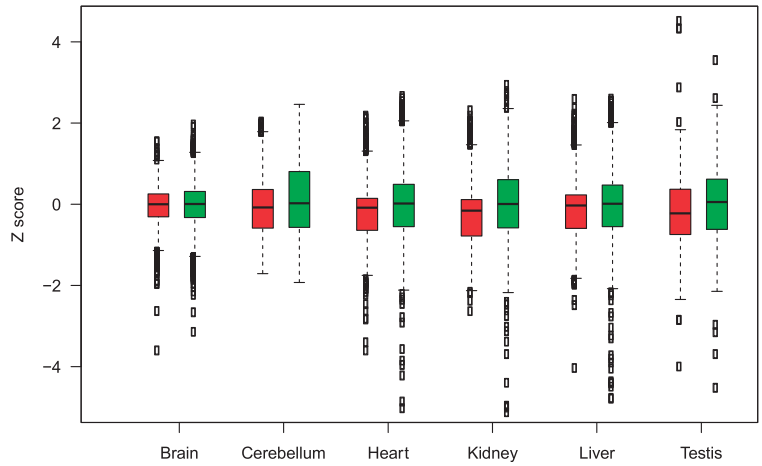


FIG. 4. Z scores of genes in and out of lamina domains across six tissues. All pairwise comparisons are highly significant (before multitest correction, Mann–Whitney U test $P < 10^{-9}$ except brain $P = 4 \times 10^{-4}$). Z score of the genes on Lamina domains are shown with boxplots in red and the rest are in green. Genes with very high or very low Z are excluded from the plot as outliers to improve presentation but have been included in Mann–Whitney U test.

Table 11. Number of Positive and Negative Z Score Genes Overlapping at Least One H3K4me3 Peak.

Tissue	Number of Genes	Number of Z+	Number of Z-	AVG (Number of Z+ with H3K4me3)	Number of Expected Z+	AVG (Number of Z− with H3K4me3)	Number of Expected Z-	χ^2 P-Value
Astrocytes-cerebellar	12,418	5,923	6,495	5,108	4,812.38	4,981.5	5,277.12	3.806E−09
Cardiac fibroblasts	12,098	5,605	6,493	4,702	4,548.21	5,115	5,268.78	0.00185
Cardiac myocytes	12,098	5,605	6,493	4,920.5	4,759.71	5,353	5,513.79	0.00146

Table 12. Number of Highly Positive and Negative Z Score Genes Overlapping at Least One H3K4me3 Peak.

Tissue	Number of Genes	Number of Z+	Number of Z-	AVG (Number of Z+ with H3K4me3)	Number of Expected Z+	AVG (Number of Z− with H3K4me3)	Number of Expected Z−	χ^2 P-Value
astrocytes-Cerebellar	6,164	3,708	2,456	3,206.5	31,32.91	2,001.5	2,075.089	0.03727
Cardiac fibroblasts	4,679	2,941	1,738	2,389	2,394.47	1,420.5	1,415.027	0.8544
Cardiac myocytes	4,679	2,941	1,738	2,520	2,516.10	1,483	1,486.902	0.8984

NOTE.—Genes with Z score higher than 1 are considered highly positive Z and the ones with Z score lower than −1 are studied as highly negative Z.

strands by contrast tend to be enriched in mC (Dahl et al. 2011). Do we see then any correspondence between hmC, mC (in cortex samples), and Z? Employing base pair resolution data (Wen et al. 2014), we indeed observe that Z (for brain) is positively correlated with hmC (Spearman correlation: $\rho = 0.17$, $P < 10^{-107}$) and negatively correlated with mC (Spearman correlation: $\rho = -0.07$, $P < 10^{-18}$).

A priori we might expect that genes associated with positive Z scores are associated with activating chromatin marks like H3K4me3 (Santos-Rosa et al. 2002; Sims et al. 2003; Martin and Zhang 2005; Greer and Shi 2012). We approach this issue using data from cardiac fibroblast, cardiac myocyte (muscle cells in heart), and astrocytes, chromatin data for which is available. Astrocytes are the most abundant cells in the brain and cerebellum (Tower and Young 1973; Chen and Swanson 2003; Tsai et al. 2012), hence would provide a defendable approximation for histone methylation profile of the whole organ. As expected Z score positive genes differ from Z score negative ones in H3K4me3 (table 11).

Given the earlier result, we might in addition expect that for genes with relatively extreme changes in Z the correspondence with H3K4me3 marks should be more pronounced. To address this we consider the subset of genes whose Z score is greater than or equal to 1 or less than or equal to −1. Unexpectedly, these genes show no significant difference in their activating histone mark methylation in two instances and only a marginal effect (astrocytes) in one (table 12).

The points mentioned earlier shows association of H3K4me3 with elevated expression in human lineage but does not elucidate whether relative gain or depletion of

Table 13. Observed Number of Concerted Genes Is Higher than Expected.

	Proportion in:						Expected Proportion	Expected Number	Observed Number	χ^2	*P*-Value
	Brain	Cerebellum	Heart	Kidney	Liver	Testis					
Z+	0.4916	0.49996	0.4999	0.4999	0.4999	0.4999	0.015356	200.0482	1216	5159	<<0.001
Z-	0.4804	0.49996	0.4999	0.4999	0.4999	0.4999	0.015006	195.4874	1165	4808	<<0.001

NOTE.—Concerted genes are either Z+ or Z− across all six tissues. So the expected number is the mean expectation of the number of concerted genes against a null of independent evolution in all tissues. The total number of genes included in this analysis is 13,027.

activating histone marks in human compared with other primates are associated with upregulation or downregulation of clusters in human lineage. To address this, we looked for evidence of H3K4me3 peaks with 1.5-fold gain or depletion in human prefrontal neuron samples compared with chimps and macaques (Shulha et al. 2012), in Z+ and Z− clusters in brain. We found that while Z+ clusters are significantly enriched in gained H3K4me3 peaks in both female and male compared with Z− clusters, Z− clusters are significantly enriched in deplete H3K4me3 peaks compared with Z+ clusters only in clusters found in female brain and not males (supplementary tables S4*a* and *b*, Supplementary Material online).

Genes with Between–Tissue Concordance in Expression Change Are Common and Clustered

Earlier, we have considered each gene's expression change in each tissue independently. Is it, however, the case that a gene upregulated in one tissue is also upregulated in other tissues or is the effect tissue specific? For those genes showing across-tissue concordance in expression change, do we find that their neighbors also tend to show across tissue concordance? That is, if a gene is up- or down-regulated in all tissues, do the neighbors also show concerted change across all tissues in the same direction as the focal gene?

To ask whether genes tend to show concerted change across all tissues, we start by analysing the six male tissues (as these have multiple replicates making the data more robust). For each gene we then convert the *Z* score into a simple classification ($Z > 0 = +1$; $Z < 0 = -1$), leaving $Z = 0$ class as is. We then consider the sum of these scores for each gene (*Z* sum). At the limit genes may be downregulated in all tissues compared with the ancestor (*Z* sum = −6) or upregulated in all (*Z* sum = +6). We compare the frequencies of *Z* sum against a null derived from randomizations in which we preserve the sum number of Z+, Z−, and $Z = 0$ seen in each tissue. We observe a great excess of incidences of concerted change, meaning an excess of more extreme scores ($\chi^2 = 12{,}409.04$, df = 12, $P << 0.01$; supplementary fig. S3, Supplementary Material online). Indeed, we find 6-fold more genes showing concerted change across all tissues than expected under a null in which the *Z* score in any given tissue is independent of that in any other tissue (table 13). We conclude that there is a strong tendency for change in expression of a given gene to be in the same direction across multiple tissues.

Those genes showing concerted evolution across all tissues belong to an eclectic mix of Gene Ontology (GO) terms including sensory perception (for positive concerted *Z* genes) and muscle development regulation (for negative concerted *Z* genes), the logic of which is not transparent to us (supplementary tables S5*a* and *b*, Supplementary Material online).

We can also ask about the expression profile of genes that show high mean *Z* scores. We consider four different metrics of expression, these being expression breadth, peak expression, mean expression level (in the tissues within which the gene is expressed), and expression skew (tau) (for definitions see Materials and Methods). We find that genes with a high mean *Z* score are more broadly expressed ($\rho = 0.14$), more highly expressed ($\rho = 0.39$), have higher maximal expression ($\rho = 0.38$), and have a low degree of skew (i.e., more evenly expressed across tissues) ($\rho = -0.13$) (in all cases $P < 10^{-14}$). In many regards, these results are to be expected as high *Z* genes are more likely to be highly expressed genes as *Z* is in part the difference between current and ancestral state and those with the highest current state are likely to be $Z > 0$. Consistent with the Z+ concerted clusters being housekeeping/highly expressed clusters, in most tissues Z+ clusters are shorter and hence denser (although the reverse is observed in clusters in brain), supplementary figure S4 and tables S3*a* and *b*, Supplementary Material online.

To ask whether genes with concerted expression evolution across tissues (all + or all−) are themselves clustered, we ask whether their neighbors are similarly concerted. To this end we identify all genes that show concerted change across all tissues either with positive *Z* or negative *Z* (absolute *Z* sum = 6). We then ask how often we find clusters of such genes (of the same sign). That is, how often do we find two concerted genes of the same sign together, how often we find triplets, *etc*. We compare these numbers to those observed in simulations in which the position of concerted genes is randomized. We find strong evidence that concerted genes clusters occur more than expected by chance (table 14). This suggests a strong principle of clustering of genes that uniformly change expression in the same direction across multiple tissues. Supplementary figure S5, Supplementary Material online, provides some examples.

Tissue-Specific Upregulation Affects Neighbors and Is Common in Cerebellum

If genes that are evolutionarily up- or downregulated across all tissues in humans cluster, do we also see that those showing tissue-specific evolutionary increase tend to sit next to genes showing evolutionary increase in the same tissue? To address this we consider those genes which, in males, show strong ($Z > 1$) increase in evolutionary change in one tissue alone,

Table 14. Monte Carlo Simulation's *P*-Value and the Number of Clusters of Concerted Genes of the Same Direction of Evolution of Expression Are Shown by Cluster Size.

Z Score Sign of the Cluster	Randomization *P*-Values Per Number of Genes in Clusters/Number of Clusters of This Size				
	2	3	4	5	6
Positive	9.999E−05/137	9.999E−05/29	9.999E−05/9	0.0059/2	0.0158/1
Negative	9.999E−05/137	9.999E−05/26	9.999E−05/8	1/0	1/0

NOTE.—Number of Z+ and Z− concerted genes are kept unchanged, but their order has been randomized, this is repeated for 1,000 iterations. Concerted gene clusters are found, and the number of occurrences of each cluster is compared with observed number of clusters of specific number of concerted genes. If the number is the same or exceeds the observed number of clusters of specific size, Monte Carlo counter is incremented. At the end of the simulation, *P*-value is calculated.

showing zero or negative *Z* in all others. This definition allows recognition of very few genes (170) but suggests the cerebellum to be a hotspot for such change (supplementary table S6*a*, Supplementary Material online). Given the low sample size, we relax the definition to include genes which are $Z > 1$ in one and only one tissue, with $Z < 1$ in all others. Henceforth, these we will refer to as tissue-specific upregulated (TSU) genes. Analysis of these provide a striking result, namely TSU genes in cerebellum alone are much common than TSU genes in other tissues (supplementary table S6*b*, Supplementary Material online), as indeed are the more strictly defined tissue-specially upregulated genes. We identified 1,230 such genes in cerebellum while only 39 genes show brain-specific upregulation. This we suggest agrees with the recent finding that the cerebellum is a focus of evolution within the primates (Barton and Venditti 2014).

Genes showing tissue-specific upregulation, in contrast to those showing coordinated change across multiple tissues, tend to be in domains of low gene density (the number of genes in ±100 kb of focal gene is low compared with coordinated ones, Mann–Whitney *U* test *P*-value = 1.26 E−43, supplementary fig. S6, Supplementary Material online). This density effect enabled us to compare the local *Z* similarity for the genes with at least one neighbor closer than 100 kb against those whose closest neighbor is further than 100 kb (of which there is an appreciable number). As shown in supplementary table S6*c*, Supplementary Material online, for the genes with a neighbor in 100 kb, the number of focal genes having a $Z > 0$ (in the focal tissue) closest neighbor is more than expected by chance ($\chi^2 = 68$, df = 5, $P << 0.001$). Indeed in all tissues the number of incidences where the nearest neighbor shows upregulation in the tissue of the focal gene is greater than expected, the deviation being significant in four of six tissues. For the genes lacking a close neighbor (supplementary table S6*d*, Supplementary Material online), the trend is mixed but the overall χ^2 statistic is weakly significant ($\chi^2 = 12.4$, df = 5, $P < 0.05$). This, however, is mostly owing to two tissues showing a strong dearth of *Z*+ genes in the vicinity of the TSU genes. That we could not detect an excess of *Z*+ genes outside of 100 kb limit suggests that many tissue-specific change genes are relatively insulated in their effects (compared with what is seen overall), possibly mediated by low gene density.

While earlier we asked merely if the neighbors have an excess of incidence of $Z > 0$ in the tissue concerned, we can also ask how many TSU genes have a TSU neighbor ($Z > 1$), with that upregulation being in the same tissue (i.e., do we see clusters of tissue-specific upregulation). While no TSU gene has any TSU neighbor in the same tissue in brain and testis, in cerebellum there are 128 genes whose closest downstream neighbor also exhibits cerebellum tissue-specific upregulation. This is not more than expected by chance (one-tailed Monte Carlo simulation keeping the same number of TSU genes in each tissue and randomizing gene order, $P > 0.05$; supplementary table S6*e*, Supplementary Material online). More generally, we see no evidence that TSU genes cluster in any tissue (supplementary table S6*e*, Supplementary Material online) and, through combining individual *P*-values across tissues with Fisher method, we find no overall support for the hypothesis of TSU clustering ($\chi^2 = 15.84$, df = 12, *P*-value > 0.1).

No Evidence for Unusual Expression Change in the Vicinity of the Human Chromosome 2 Fusion Event

Earlier, we have considered trends en masse. Close scrutiny of some forms of gross chromosomal change suggest that genes neighboring chromosomal disruption sites tend to have altered gene expression (Milot et al. 1996; Dillon et al. 1997; Kleinjan and van Heyningen 1998; Kleinjan and van Heyningen 2005; Harewood and Fraser 2014). Do we see any evidence of this on the broader evolutionary scale? To address this we consider the genes in the vicinity of the human chromosome 2 fusion event.

Human chromosome 2 is fusion of two chromosomes present in the great apes, chimp included (Miller and Reis 1982). The fusion zone is reported to be in the vicinity of 2q13-2q14.1 (Fan et al. 2002). Via the Ensembl web browser (Flicek et al. 2014) under comparative genomic mode, we determined that human gene ENSG00000146556 was in the vicinity of the fusion boundary, its neighbors in chimp being ENSPTRG00000014555 on chromosome 2b in one direction and ENSPTRG00000012388 and ENSPTRG00000012383 on chromosome 2a in the other direction. We then asked whether the mean *Z* for genes in proximity to this site were in any manner unusual. To this end we considered a 1 MB window upstream and downstream of the fusion sites and considered *Z* for all genes within this domain. As expected, in one direction there are relatively few genes, this corresponding to the ancient telomeric end of one of the fusion chromosomes. The mean *Z* score for genes in this window is no different to zero (mean *Z* = 0.002, SD = 0.396), suggesting that this is not a zone associated with either up- or downregulation (supplementary fig. S7, Supplementary Material online).

Sex-Biased Gene Expression Change Is Clustered

As we have, for several tissues, change in expression data in both males and females, we can ask, for any given gene, whether the change in expression in one sex correlates with that in the other sex. Under a null of no change in the degree

Table 15. Spearman Correlation between Female and Mean of Male Z Scores Per Tissue.

Tissue	ρ	P-Value
Brain	0.52967	<<0.0001
Cerebellum	0.32532	<<0.0001
Heart	0.45401	<<0.0001
Kidney	0.43073	<<0.0001

Table 16. Spearman Correlation between Sex Bias Standard Residual of Standard Major Axis Estimation between Z of Male and Female for a Focal Gene and Standard Residual of Its Nearest Downstream Neighbor.

Tissue	Nonoverlapping P-Value	Nonoverlapping ρ	Overlapping P-Value	Overlapping ρ
Brain	*0.00018*	0.03995	*0.00325*	0.10407
Cerebellum	0.03109	0.02304	*9.10E−06*	0.15636
Heart	*1.42E−05*	0.04638	*8.04E−05*	0.13913
Kidney	*6.95E−19*	0.09465	0.01206	0.08883

NOTE.—Incidences significant after Bonferroni testing are shown in italic.

of sex bias in expression, such a check also provides an internal consistency check for our mode of analysis and the data. Indeed, as for female tissues we have only one sample, and it might be that data from females are too noisy to be dependable. We find a strong correlation, on a gene-by-gene basis for Z in males in given tissue and Z in females for the same tissue (table 15). The correlation stays significant when zero Z score (after correction) genes are left in (supplementary table S7, Supplementary Material online). This provides support for the hypothesis that the dominant trend in change in gene expression is not sex biased.

By considering the standardized residuals from orthogonal regression between the male and female Z scores, we can also obtain information on the extent of sex bias in the evolution of gene expression. Note this is not the same as the degree of sex bias, but rather the degree of change in sex bias. We can then ask whether the degree of change in sex bias is also nonautonomous. To this end, we consider the correlations as mentioned earlier. For each focal gene, we consider the correlation between residuals for a focal gene and its nearest downstream neighbor, between the focal gene and its two nearest neighbors (one upstream one down) and between the focal gene and the mean of all neighbors within 100 kb of the focal gene. In all examples we find a significant and positive correlation indicating the sex-biased expression change also occurs in a clustered mode (tables 16–18). In 6 of 8 nearest neighbor comparisons, the effect is more pronounced for overlapping genes. The genomic sizes of the clusters of genes with correlated residuals is varied across tissues, starting with cerebellum and heart clusters below 50 kb, going up to 100 kb in brain and exceeding 200 kb in kidney (fig. 5).

These results support the hypothesis that the extent of change in sex bias is also genomically regionalized. This is further supported by the finding that when we score residuals as positive or negative states, we again find fewer switches in

Table 17. Spearman Correlation between Standard Residual of Standard Major Axis Estimation between Z of Male and Female for a Focal Gene and Mean Standard Residual of Its Two Nearest Neighbors.

Tissue	Nonoverlapping P-Value	Nonoverlapping ρ	Overlapping P-Value	Overlapping ρ
Brain	*1.46E−05*	0.05452	*0.00281*	0.07649
Cerebellum	0.01433	0.03082	*6.07E−07*	0.12738
Heart	*4.50E−07*	0.06346	*3.05E−08*	0.14127
Kidney	*7.02E−23*	0.12348	*4.32E−06*	0.11740

NOTE.—Incidences significant after Bonferroni testing are shown in italic.

Table 18. Spearman Correlation between Standard Residual of Standard Major Axis Estimation between Z of Male and Female of the Focal Gene and the Mean of Standard Residual of All Its Neighbors within 100 kb of the Focal Gene.

Tissue	Spearman P-Value	Spearman ρ
Brain	4.00E−08	0.04817
Cerebellum	0.00848	0.02310
Kidney	1.71E−39	0.11504
Heart	1.87E−05	0.03755

NOTE.—All incidences are significant after Bonferroni testing.

state than expected by chance, implying clustering (P from randomisation, brain $P = 0.0009$; cerebellum $P = 0.01$; heart $P = 0.007$; kidney $P = 0.001$).

The earlier analysis ignores those instances where Z is zero (before median correction) for a gene in either sex. This may be biasing results as the genes with $Z = 0$ in one sex, but not the other, are sex biased in their change of expression. This makes little difference to results (supplementary tables S8*a–c*, Supplementary Material online).

No Evidence That the X Chromosome Is Enriched for Genes Changing Sex Bias

With the same data we can also ask whether another form of clustering is seen, i.e., chromosomal scale clustering. According to Rice's hypothesis (Rice 1984) the X chromosome should be a hotspot for sex-biased gene expression change. He postulates that genes with sexually antagonistic fitness effects can be more likely to spread if on a sex chromosome. The spread of such alleles creates the context for the spread of modifiers that limit the expression of the deleterious allele in the sex in which the effect is deleterious, i.e., modifiers of sex-specific change in expression. Hence sex biased gene expression change is expected to be more pronounced on the X chromosome than on autosomes. This can mean both the evolution toward male-biased and female-biased gene expression.

Given that we have no strong prior on the direction of sex-biased change on the X, we consider for all genes the modulus of the degree of sex-biased change. We then ask whether these values are different for X than for autosomes. We find no evidence for a difference (Mann–Whitney *U* test, brain *P*-value = 0.4906; cerebellum *P*-value = 0.8944; heart *P*-value = 0.9374; kidney *P*-value = 0.7523). In addition we

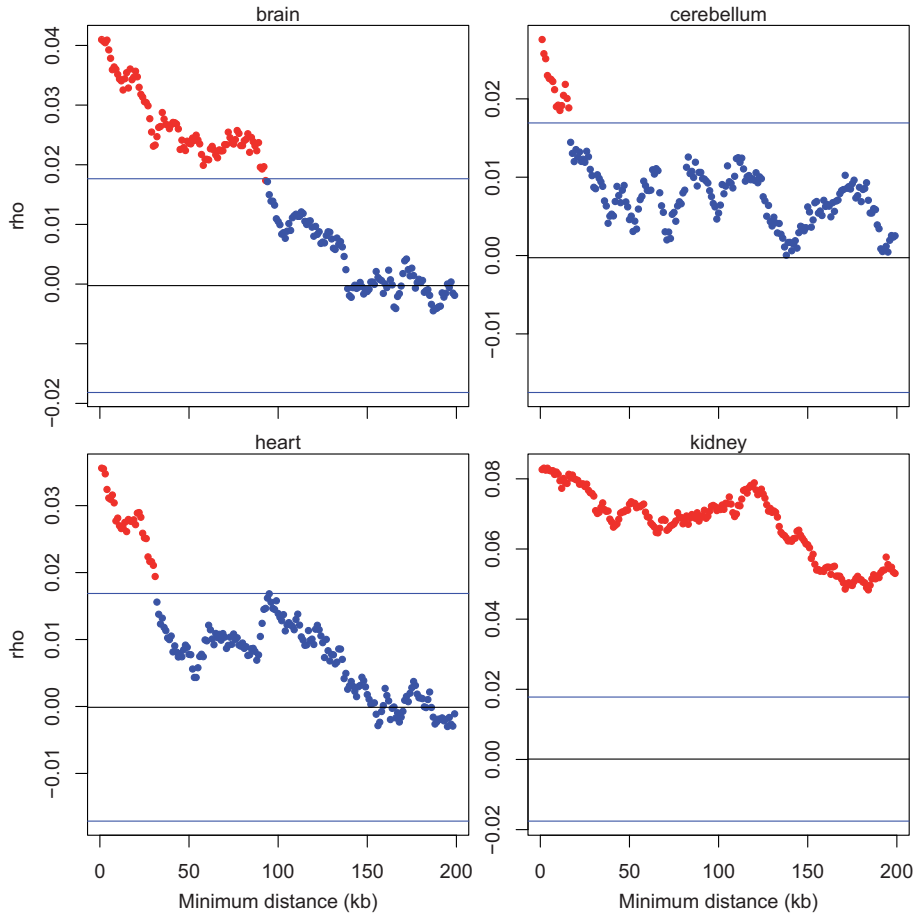


Fig. 5. The extent of local correlation in sex-biased expression change for four tissues. Method is the same as that for figure 3, excepting that here we employ standardized residuals of the orthologous regression on *Z* between sexes (rather than *Z*). We consider all focal genes and the correlation between residuals of *Z* scores for these genes and the nearest downstream gene on the same chromosome a minimum of *x* base pairs away. Correlations significant at the 0.05 level are shown in red, otherwise in blue. The blue horizontal lines indicate 1.96 SD limits determined by randomization, with the black line indicating mean of null expectation (which should be around zero).

can ask about the 5% of genes with the most extreme change in sex bias (the 5% with the highest modulus of residual score). Are these more commonly found on the X chromosome? We find no evidence to support this proposition either (supplementary table S9, Supplementary Material online). We conclude that we see no evidence that the X chromosome is a hotspot for sex-biased gene expression change. However, if instead we consider the change in expression of genes in the testis, we do find that X-linked genes show a different median *Z* compared with autosomal genes. Considering only those genes with expression > 0 in the ancestor, the median *Z* for X-linked genes is 0.15, while for autosomes it is -0.012 (Mann–Whiney *U* test, $P = 0.00023$). In no other tissue is the median *Z* on the X greater than the median *Z* on the autosomes.

Discussion

Here we have presented evidence that gene expression change, at least in humans, occurs on a cluster-by-cluster basis, such that the expression change of any given focal gene predicts the expression change of genes in its vicinity in any given tissue. The result is insensitive to the metric of expression change. Moreover, many genes show coordinated changes in expression across multiple tissues and in the same tissue in different sexes. Genes that show coordinated expression changes across multiple tissues tend to sit next to other genes showing similar coordination. This suggests that a dominant mode of expression change evolution may be nothing more than a switch of a chromosomal block to a state of permanently open (or predominantly closed) chromatin in multiple tissues (or open/closed longer in multiple tissues),

thereby causing increases or decreases in expression of spans of genes in all circumstances.

Gene density effects we suggest might in addition also be relevant. If much of the expression change is owing to local chromatin modification, we might expect that domains of high gene density are more coordinated in their expression change, simply because the chances that a local change to one gene might affect another would be greater. Such a model is consistent with our finding that genes showing tissue-specific upregulation and that have no gene neighbor within 100 kb do not affect expression of their nearest (over 100 kb) neighbor, while other genes in high density domains do. If upregulation of one gene in a zone of high gene density affects the neighbors whose upregulation affects the neighbors on and so forth, this might in turn generate self-propagating domains of expression change. It is notable then that genes showing increased expression across multiple tissues tend to be in domains of high gene density.

Why gene expression for the focal gene changes is unclear, although we found no evidence for a coupling with chromosomal alternations (i.e., in the chromosome 2 fusion event). While the precise mechanisms of nonautonomous evolution are unclear, the form of the curves relating genomic distance to correlation in *Z* score, suggest much more profound effects in immediate vicinity, a conclusion supported by the stronger correlations seen for overlapping genes. We suggest that there may thus be more than one mechanism at play. Perhaps in the immediate vicinity of a gene, expression of one gene directly impacts the expression of its neighbors (cf. the ripple effect [Ebisuya et al. 2008]), while over broader spans ($>$ 100 kb), a more generic chromatin opening/closing and self-propagation mechanism (Batada et al. 2007; Gierman et al. 2007) may be more relevant. Either way, our results suggest that a promoter-focused concentration on the causes of expression change (Tirosh et al. 2009; Rosin et al. 2012; Wittkopp and Kalay 2012; Yang et al. 2014) is likely to provide too restricted a view of expression change viewed more globally, at least within primates.

While we detected expression change clusters defined on an intrachromosomal scale, which for the most part is not predicted by population genetical theory, we did not observe a form of clustering that we had expected from such theory. Rice's theory (Rice 1984) would suggest that X-linked genes should be prone to changes in sex-biased gene expression; however, we did not detect this for expression in tissues present in both sexes. One possible explanation for this might be that the tissues examined may not be those most likely to be subject to the strongest sex-biased gene expression. Indeed, testes show a large increase in *Z* for X-linked genes compared with autosomal genes, potentially compatible with Rice's model (note this is not change in degree of sex bias as there is no female testicular expression to compare it with). The data thus accord with a model in which for nonsex-specific tissues the degree of sex-biased change in gene expression is a largely neutral process and thus outside of the domain of Rice's hypothesis.

More generally, given the extent to which one gene's expression change affects that of the neighbors, it is simplest to suppose as a null model that much of the expression change we observe is neutral and what might be called expression "piggybacking." That is to say, the upregulation of one gene may be selectively favored but, because its upregulation increases the chances that the neighbors are upregulated, the spread through the population of the focal heritable expression change causes expression divergence (from the ancestral state) of near neighbors of that focal gene. The expression change of the neighbours need not be the focus of selection but rather a necessary consequence of the change to the focal gene.

Expression piggybacking may be considered an analog of genetic hitchhiking, in so much as it suggests correlated changes at genomically neighboring sites. Piggybacking is different, however, in so much as it does not require linkage disequilibrium between alleles at closely linked sites. Indeed, in piggybacking there need only be one allele affecting the expression of the focal gene while the neighboring genes can, in principle, be genetically uniform across the population. Nonetheless, the flanking genes will change, over evolutionary time, their expression profile, piggybacking on the heritable expression change at the focal allele. Alternatively put, estimation of the net selective impact, if any, of any mutation affecting the expression of any given gene, needs also to factor in the effects this focal expression change has on the expression of neighbors as well. Our data are broadly consistent with expression piggybacking, possibly largely selectively neutral, being a fundamental cause of expression divergence in primates.

Materials and Methods

Estimation of Z Scores

Gene expression data were obtained from Brawand et al (2011). We used expression values reported in NormalizedRPKM_ConstitutiveAlignedExons_Primate1to1-Orthologues.txt and extracted loci and strand information from Human_Ensembl57_TopHat_UniqueReads.txt also provided in the supplementary materials of the relevant paper. This provides RPKM figures for 13,027 genes in six tissues across five primate species. To determine the change in gene expression between current levels in humans and that seen in the human–chimp common ancestor we employed BayesTraits (Pagel et al. 2004). The assumed phylogeny and branch lengths are the same as those employed by Brawand et al. (2011).

BayesTraits was run in the following manner. Normalized RPKM, as provided by Brawand et al. (2011), were passed to BayesTraits as measures of gene expression. For each gene, mean of normalized RPKM values across different individuals in Human was calculated separately for male and female samples. Also if more than one male or female sample is available in any of the tissues in chimpanzee or any of the outgroups, their mean is computed and passed to BayesTraits, otherwise a single expression value was used. To find the estimated gene expression level in the ancestor of human and chimpanzee, for each gene in each tissue, BayesTraits program was run twice, first to build the

estimated gene expression tree for males and second for female samples. Each time, the primate phylogenic tree and means of normalized RPKM of the gene in human and also its orthologous genes in chimpanzee and three primate outgroups (gorilla, orangutan, and macaque), in corresponding gender, are passed to BayesTraits, to build the estimated gene expression model. BayesTraits employs Markov chain Monte Carlo and maximum likelihood to find the posterior distribution of this model and estimate the level of expression in this tree's middle nodes (Pagel et al. 2004). Through examination of the convergence trends of the BayesTraits output, we considered that the final 10% of BayesTraits estimates would be robust. From this sample we estimate both the mean (E_a) and variance (V_a) in the estimation of the human–chimp ancestral state. Relaxation of the 10% cutoff makes no important difference to results (data not shown).

These simulations were run independently for each gene, for each tissue in each sex. If the mean expression of given gene, in given tissue in a given sex is $E_{current}$, or E_c in abbreviated form, and its variance is V_c, if estimable, while that for the ancestral condition is E_a and V_a, then we can define the degree of expression divergence in human lineage from human–chimp ancestor as a Z score:

$$Z = \frac{E_c - E_a}{\sqrt{V_c + V_a}}$$

This metric compares the extent of difference between mean current expression level and ancestral level, scaled by the degree of variation both in current estimates (expression noise or measurement error) and the degree of uncertainty in the ancestral state's estimation. A positive Z implies an increase in gene expression since the ancestor. In part the defense for our metric is the same as the defense for any application of a Z score, namely it measures difference in standard deviation units. That is, a gene with largely variable expression across individuals or high fluctuation and uncertainty in estimation of expression in ancestor would have a lower Z score compared to a gene with similar but steadier level of current expression and/or one with similar but more stable estimation of ancestral level of expression. However, another part of the defense is that in our model, inspired by the ripple hypothesis, increased opening of chromatin can lead to increased spurious expression. Our supposition is that this might cause an approximately constant absolute increase in the amount of transcription in all neighbors a given distance away, not an increase proportional to the current level (as measured by fold change). Nonetheless to examine the possibility that results might be contingent on metric we also consider 1) a digital representation (increase or decrease since ancestor) and 2) fold change. Note too that we are not concerned with whether our metric calls significance in gene expression change as most of the gene expression in our model is neutral drift owing to ripple effects. Rather, we wish to present a quantitative variable that captures the absolute amount of expression change factored in standard deviation units.

For each tissue in each sex we assume that the median expression change must be zero. This is equivalent to assuming an absence of net increase or decrease in overall expression levels. This required a minor adjustment of Z scores for all genes in all tissues. If the median Z in any given tissue in a given sex is M, then we defined modified Z as $Z_{mod} = Z - M$. This forces all tissues to have a modified median of zero and as many genes increasing expression as decreasing (this being approximately equivalent to an assumption that the net transciptome size is no different; hence, for every gene increasing expression there should be one decreasing expression). All analyses were performed on Z_{mod}. Henceforth, we shall refer to Z, for convenience, where Z_{mod} is what we are employing. In practice the correction makes little or no difference as 1) the correction is usually very small and 2) many of our statistics are rank order based and so unaffected by the modification. We note that our method has the advantage that it largely eliminates any RNAseq amplication biases (e.g., owing to GC content) from affecting our metric of expression change. This is because nucleotide content is almost unchanged between human and chimp, and hence any bias in amplification of a given transcript is likely to affect human and chimp equally. By considering only the change from the ancestor we thus exclude amplification biases from derivation of Z. As evidence for this, the mean correlation, across all tissues, between Z and the change in GC between human and chimp is indistinguishable from zero.

Chromatin Data

For a few human cell lines, ChIP-seq histone methylation data produced by University of Washington is available through ENCODE's portal (Bernstein et al. 2012; Gerstein et al. 2012; Rosenbloom et al. 2012). We could approximate whole tissue histone methylations profile by matching the most abundant cell lines in heart and cerebellum to three of the cell lines available in ENCODE. Among many cell types composing heart, Cardiac fibroblast and cardiac myocyte (muscle cells in heart) are consequently mostly abundant ones. Furthermore, astrocytes are the most numerous cell type in the central nervous system (Chen and Swanson 2003; Tsai et al. 2012). Hence, HAc, an astrocytes-cerebellar cell line, was used to approximate histone methylation profile in cerebellum (Tower and Young 1973).

To do the histone methylation analysis, H3K4me3 peak data were downloaded as an activating histone mark (Santos-Rosa et al. 2002; Sims et al. 2003; Martin and Zhang 2005; Greer and Shi 2012) for above cell lines. Then Z score positive and negative genes overlapping one or more H3K4me3 peak(s) were found using Bedtools (Quinlan and Hall 2010). Due to the histone mark protocol used in ENCODE, each experiment was repeated twice and peak data are reported separately for each repetition. So here we report the average number of Z score positive or negative genes overlapping one or more peaks across these two repeated peak data sets.

We also compared Z score positive and negative clusters with regard to gain and depletion of H3K4me3 peaks in

humans compared with chimps and macaques. To do this, we took 885 H3K4me3 peaks which were shown to have 1.5-fold higher human-specific gain in human samples compared with macaque and chimpanzee samples as shown by Shulha (2012). Intersect command from bedtools was then used to find the clusters overlapping a gained H3K4me3 peak. An ad hoc script was used to count the number of $Z+$ and $Z-$ clusters with at least one gained peak. Similarly we also compared the number of $Z+$ and $Z-$ clusters with evidence for at least one H3K4me3 depleted peak using 177 H3K4me3 peaks with human-specific depletion which had at least 1.5-fold lower tag density in human samples compared with chimps and macaques as shown by Shulha (2012).

GO Analysis

Is there a functional link between the genes that show the same sign of expression change across all tissues (concerted genes)? Is there a functional clue to link the genes with elevated changed expression across all tissues? To determine this, the concerted genes (same profile of change across all tissues) are divided into two sets:first the ones with elevated expression in human lineage compared with human–chimp ancestor across all tissues in male samples and second the ones with reduced expression than the estimated expression in the ancestor. We just used male sample tissues for this analysis as there are more repeats available for these, also as shown, their expression is more stable and less noisy. Doing this, we found 1,244 concerted Z score positive genes and 1,053 concerted negative ones. Then GO term enrichment analysis was performed on these two sets, using GOrilla (Eden et al. 2009), to find the enriched GO functions and processes.

Expression Measures

To address the correlates of Z we also ask about a series of expression measures, these being breadth, mean rate, peak rate, and tau. For a gene to be considered as being expressed in a given tissue in a given species we required that the mean across replicates for that tissue to be more than at least 2 RPKM. If it was less than 2, it was set to zero for that tissue. Breadth is defined as the proportion of tissues within which a gene is expressed. To prevent nonindependence between rate and breadth, we defined rate as the mean rate of expression of that gene across all tissue within which it is expressed (i.e., at rate $>$ 2). Peak rate is the maximum expression level considered across all tissues. Tau is a measure of skew in expression and is defined as:

$$\tau = \frac{\sum_{j=1}^{n}\left(1 - \frac{\log(e_j)}{\log(e_{\max})}\right)}{n-1}$$

where there are n tissues, the expression in any one being e_j and the maximal for that gene across all tissues is $e_{\max}$. A gene with very highly skewed expression (very high in only one tissue) take a high value of tau (limit approaching 1) while those expressed uniformly take a low value (limit zero).

hmC and mC Assays

Base resolution map of hydromethylome in prefrontal cortex has been produced by Wen et al. (2014). First shown in Bacteriophage, hmC is able to turn genes on or off (Wyatt and Cohen 1952; Dahl et al. 2011). Wen et al. (2014) has recently shown 10-fold increase in hmC in adult prefrontal cortex compared with fetal. Also, hmC correlates positively with gene expression while mC correlates negatively with gene expression (Colquitt et al. 2013; Wen et al. 2014). Furthermore, there is disparity between hmC and mC enrichment on sense and antisense strands, hmC being enriched on sense and mC on antisense strands (Peric-Hupkes et al. 2010). To find out if they correspond with change in gene expression, we took hmC and mC percentages as reported by Wen et al (2014) and calculated how they correlated with Z scores of genes in brain.

Lamina Domain Assignment

LADs originally produced by Guelen et al (2008) using Lung fibroblast cell line, are available through UCSC's table browser for hg19. Intersect command from bedtools (Quinlan and Hall 2010) was used to find the genes overlapping these domains. For this analysis, genes with zero Z scores (prior to modification) are not removed due to expectation of the genes on LAD domains to be very lowly, if at all, expressed. Then Z of genes on and off LAD domains were compared using Mann–Whitney U test and also Brunner Munzel test, to correct for robustness to the form of distributions.

Statistics

Where appropriate statistics were performed in R, many analyses were performed using Monte Carlo simulations. In these incidences, if N is the number of observations as extreme or more extreme as observed and M is the number of simulants, then the unbiased estimator of the type I error rate (what may be regarded as an empirical P) is:

$$P = \frac{N+1}{M+1}.$$

Supplementary Material

Supplementary figures S1–S7 and tables S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The work was funded by a studentship award from University of Bath to A.T.G. and Medical Research Grant MR/L007215/1.

References

Barton RA, Venditti C. 2014. Rapid evolution of the cerebellum in humans and other great apes. *Curr Biol.* 24:2440–2444.

Batada NN, Urrutia AO, Hurst LD. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet.* 23:480–484.

Bernstein BE, Birney E, Dunham I, Green E, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Birnbaum K, Shasha D, Wang J, Jung J, Lambert G, Galbraith D, Benfey P. 2003. A gene expression map of the *Arabidopsis* root. *Science* 302: 1956–1960.
Blumenthal T, Evans D, Link C, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu W, Duke K, Kiraly M, et al. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* 417:851–854.
Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* 420:666–669.
Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343–348.
Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C, Taylor M, Engstrom P, Frith M, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet.* 38:626–635.
Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus M, van Asperen R, Boon K, Voute P, et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* 291:1289.
Chen Y, Swanson RA. 2003. Astrocytes and brain injury. *J Cereb Blood Flow Metab.* 23:137–149.
Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell.* 2:65–73.
Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet.* 26:183–186.
Colquitt BM, Allen WE, Barnea G, Lomvardas S. 2013. Alteration of genic 5-hydroxymethylcytosine patterning in olfactory neurons correlates with changes in gene expression and cell identity. *Proc Natl Acad Sci U S A.* 110:14682–14687.
Dahl C, Gronbaek K, Guldberg P. 2011. Advances in DNA methylation: 5-hydroxymethylcytosine revisited. *Clinica Chim Acta.* 412:831–836.
Davila Lopez M, Martinez Guerra J, Samuelsson T. 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. *PLoS One* 5:e10654.
Denver D, Morris K, Streelman J, Kim S, Lynch M, Thomas W. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet.* 37:544–548.
Dillon N, Trimborn T, Strouboulis J, Fraser P, Grosveld F. 1997. The effect of distance on long-range chromatin interactions. *Mol Cell.* 1:131–139.
Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol.* 10:1106–1113.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
Fan Y, Linardopoulou E, Friedman C, Williams E, Trask B. 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* 12:1651–1662.
Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42:D749–D755.
Forrest A, Kawaji H, Rehli M, Baillie J, de Hoon M, Lassmann T, Itoh M, Summers K, Suzuki H, Daub C, et al. 2014. A promoter-level mammalian expression atlas. *Nature* 507:462–470.
Franck E, Hulsen T, Huynen M, de Jong W, Lubsen N, Madsen O. 2008. Evolution of closely linked gene pairs in vertebrate genomes. *Mol Biol Evol.* 25:1909–1921.
Fukuoka Y, Inaoka H, Kohane IS. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* 5:4.
Gerstein M, Kundaje A, Hariharan M, Landt S, Yan K, Cheng C, Mu X, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489: 91–100.
Gierman H, Indemans M, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R. 2007. Domain-wide regulation of gene expression in the human genome. *Genome Res.* 17:1286–1295.
Greer E, Shi Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Gen.* 13:343–357.
Grunstein M. 1997. Histone acetylation in chromatin structure and transcription. *Nature* 389:349–352.
Guelen L, Pagie L, Brasset E, Meuleman W, Faza M, Talhout W, Eussen B, de Klein A, Wessels L, de Laat W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453:948–951.
Hammock EA, Young LJ. 2002. Variation in the vasopressin V1a receptor promoter and expression: implications for inter- and intraspecific variation in social behaviour. *Eur J Neurosci.* 16:399–402.
Harewood L, Fraser P. 2014. The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet.* 23: R76–R82.
Hornung G, Oren M, Barkai N. 2012. Nucleosome organization affects the sensitivity of gene expression to promoter mutations. *Mol Cell.* 46:362–368.
Hurst LD, Williams EJ, Pál C. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* 18: 604–606.
Janga S, Collado-Vides J, Babu M. 2008. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc Natl Acad Sci U S A.* 105:15761–15766.
Janicki SM, Tsukamoto T, Salghetti SE, Tansey WP, Sachidanandam R, Prasanth KV, Ried T, Shav-Tal Y, Bertrand E, Singer RH. 2004. From silencing to gene expression: real-time analysis in single cells. *Cell* 116: 683–698.
Képès F. 2003. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J Mol Biol.* 329:859–865.
Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, et al. 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.* 14: 1462–1473.
Kleinjan D-J, van Heyningen V. 1998. Position effect in human genetic disease. *Hum Mol Genet.* 7:1611–1618.
Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet.* 76:8–32.
Kruglyak S, Tang H. 2000. Regulation of adjacent yeast genes. *Trends Genet.* 16:109–111.
Lee J, Sonnhammer E. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13:875–882.
Lee JM, Sonnhammer EL. 2004. Genomic gene clustering analysis of pathways in eukaryotes (vol 13, pg 875, 2003). *Genome Res.* 14:2510.
Lercher MJ, Blumenthal T, Hurst LD. 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res.* 13:238–243.
Lercher MJ, Hurst LD. 2006. Co-expressed yeast genes cluster over a long range but are not regularly spaced. *J Mol Biol.* 359:825–831.
Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet.* 31:180–183.
Letourneau A, Santoni FA, Bonilla X, Sailani MR, Gonzalez D, Kind J, Chevalier C, Thurman R, Sandstrom RS, Hibaoui Y, et al. 2014. Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature* 508:345–350.
Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* 128:707–719.
Li Y-Y, Yu H, Guo Z-M, Guo T-Q, Tu K, Li Y-X. 2006. Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comp Biol.* 2:e74.
Liao BY, Zhang J. 2008. Coexpression of linked genes in mammalian genomes is generally disadvantageous. *Mol Biol Evol.* 25:1555–1565.
Liu C, Ghosh S, Searls DB, Saunders AM, Cossman J, Roses AD. 2005. Clusters of adjacent and similarly expressed genes across normal

human tissues complicate comparative transcriptomic discovery. *OMICS: J Integr Biol.* 9:351–363.
Martin C, Zhang Y. 2005. The diverse functions of histone lysine methylation. *Nat Rev Mol Cell Biol.* 6:838–849.
Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N. 2012. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 151:1417–1430.
Michalak P. 2008. Coexpression coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* 91:243–248.
Mijalski T, Harder A, Halder T, Kersten M, Horsch M, Strom TM, Liebscher HV, Lottspeich F, de Angelis MH, Beckers J. 2005. Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc Natl Acad Sci U S A.* 102:8621–8626.
Miller R, Reis D. 1982. The origin of man: a chromosomal pictorial legacy. *Science* 215:1526.
Milot E, Strouboulis J, Trimborn T, Wijgerde M, de Boer E, Langeveld A, Tan-Un K, Vergeer W, Yannoutsos N, Grosveld F, et al. 1996. Heterochromatin effects on the frequency and duration of LCR-mediated gene transcription. *Cell* 87:105–114.
Molineris I, Grassi E, Ala U, Di Cunto F, Provero P. 2011. Evolution of promoter affinity for transcription factors in the human lineage. *Mol Biol Evol.* 28:2173–2183.
Noguchi M, Miyamoto S, Silverman TA, Safer B. 1994. Characterization of an antisense Inr element in the eIF-2 alpha gene. *J Biol Chem.* 269: 29161–29167.
Nutzmann HW, Osbourn A. 2014. Gene clustering in plant specialized metabolism. *Curr Opin Biotechnol.* 26:91–99.
Oliver B, Misteli T. 2005. A non-random walk through the genome. *Genome Biol.* 6:214.
Osato N, Suzuki Y, Ikeo K, Gojobori T. 2007. Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* 176:1299–1306.
Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol.* 53:673–684.
Pal C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet.* 33:392–395.
Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Graf S, Flicek P, Kerkhoven RM, Van Lohuizen M, et al. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell.* 38:603–613.
Poyatos J, Hurst L. 2007. The determinants of gene order conservation in yeasts. *Genome Biol.* 8:R233.
Prescott EM, Proudfoot NJ. 2002. Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci U S A.* 99: 8796–8801.
Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W, Sperling S. 2007. Genomic organization of transcriptomes in mammals: coregulation and cofunctionality. *Genomics* 89: 580–587.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.
Reddy KL, Zullo JM, Bertolino E, Singh H. 2008. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* 452:243–247.
Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. *Nat Rev Genet.* 2:21–32.
Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.
Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, et al. 2012. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* 40:D912–917.
Rosin D, Hornung G, Tirosh I, Gispan A, Barkai N. 2012. Promoter nucleosome organization shapes the evolution of gene expression. *PLoS Genet.* 8:e1002579.
Roy PJ, Stuart JM, Lund J, Kim SK. 2002. Chromosomal clustering of muscle-expressed genes in caenorhabditis elegans. *Nature* 418: 975–979.
Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NCT, Schreiber SL, Mellor J, Kouzarides T. 2002. Active genes are trimethylated at K4 of histone H3. *Nature* 419:407–411.
Sémon M, Duret L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol Biol Evol.* 23:1715–1723.
Shulha H, Crisci J, Reshetov D, Tushir JS, Cheung I, Bharadwaj R, Chou HJ, Houston IB, Peter CJ, Mitchell AC, et al. 2012. Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. *PLoS Biol.* 10:e1001427.
Sims R Jr, Nishioka K, Reinberg D. 2003. Histone lysine methylation: a signature for chromatin function. *Trends Genet.* 19:629–639.
Singer GA, Lloyd AT, Huminiecki LB, Wolfe KH. 2005. Clusters of coexpressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol.* 22:767–775.
Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol.* 1:5.
Sproul D, Gilbert N, Bickmore WA. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet.* 6: 775–781.
Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PW, Bauer C, Munzel M, Wagner M, Muller M, Khan F, et al. 2013. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* 152:1146–1159.
Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 306:655–660.
Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F. 2014. Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 24:390–400.
Takai D, Jones PA. 2004. Origins of bidirectional promoters: computational analyses of intergenic distance in the human genome. *Mol Biol Evol.* 21:463–467.
Tirosh I, Barkai N, Verstrepen KJ. 2009. Promoter architecture and the evolvability of gene expression. *J Biol.* 8:95.
Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet.* 38:830–834.
Tower DB, Young OM. 1973. The activities of butyrylcholinesterase and carbonic anhydrase, the rate of anaerobic glycolysts, and the question of a constant density of glial cells in cerebral cortices of various mammalian species from mouse to whale. *J Neurochem.* 20:269–278.
Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res.* 14:62–66.
Tsai HH, Li H, Fuentealba LC, Molofsky AV, Taveira-Marques R, Zhuang H, Tenney A, Murnen AT, Fancy SP, Merkle F, et al. 2012. Regional astrocyte allocation regulates CNS synaptogenesis and repair. *Science* 337:358–362.
Uesaka M, Nishimura O, Go Y, Nakashima K, Agata K, Imamura T. 2014. Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genomics* 15:35.
Van Bortle K, Corces VG. 2013. Spinning the web of cell fate. *Cell* 152: 1213–1217.
Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13:1998–2004.
Wakano C, Byun JS, Di LJ, Gardner K. 2012. The dual lives of bidirectional promoters. *BBA-Gene Regul Mech.* 1819:688–693.
Wang Y, Rekaya R. 2009. A comprehensive analysis of gene expression evolution between humans and mice. *Evol Bioinform Online.* 5:81.
Weber CC, Hurst LD. 2011. Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation. *Genome Biol.* 12:R23.
Wei W, Pelechano V, Jarvelin AI, Steinmetz LM. 2011. Functional consequences of bidirectional promoters. *Trends Genet.* 27:267–276.
Wen L, Li X, Yan L, Tan Y, Li R, Zhao Y, Wang Y, Xie J, Zhang Y, Song C, et al. 2014. Whole-genome analysis of 5-hydroxymethylcytosine and

5-methylcytosine at base resolution in the human brain. *Genome Biol.* 15:R49.
Williams EJB, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.* 14: 1060–1067.
Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 13:59–69.
Woo YH, Li W-H. 2011. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proc Natl Acad Sci U S A.* 108:3306–3311.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 8:206–216.
Wright KL, White LC, Kelly A, Beck S, Trowsdale J, Ting JP. 1995. Coordinate regulation of the human TAP1 and LMP2 genes from a shared bidirectional promoter. *J Exp Med.* 181: 1459–1471.
Wyatt GR, Cohen SS. 1952. A new pyrimidine base from bacteriophage nucleic acids. *Nature* 170:1072–1073.
Yang H, Li D, Cheng C. 2014. Relating gene expression evolution with CpG content changes. *BMC Genomics* 15:693.

Table S1a. Spearman correlation between focal gene's fold change and fold change of its closest non-overlapping downstream neighbor. Fold change is defined as mean of current level of expression in human divided by mean of estimated expression in ancestor All statistics are significant after Bonferroni testing.

Tissue	Male *P-value*	Male *Rho*	Female *P-value*	Female *Rho*
Brain	9.61E-10	0.06680	1.71E-10	0.07036
Cerebellum	5.15E-22	0.10794	7.50E-31	0.12778
Kidney	1.29E-107	0.23983	1.57E-08	0.06260
Heart	2.29E-43	0.15370	3.10E-22	0.11004
Liver	3.41E-16	0.09092	NA	NA
Testis	7.16E-92	0.21573	NA	NA

Table S1b. Spearman correlation between focal gene's fold change and mean fold change of its closest non-overlapping neighbors on both sides. All statistics are significant after Bonferroni testing.

Tissue	Male *P-value*	Male *Rho*	Female *P-value*	Female *Rho*
Brain	5.12E-18	0.09341	2.49E-19	0.09756
Cerebellum	3.88E-38	0.14116	6.95E-57	0.17230
Kidney	6.99E-149	0.27710	3.74E-17	0.09157
Heart	2.48E-59	0.17661	1.72E-35	0.13684
Liver	1.24E-23	0.10929	NA	NA
Testis	5.20E-145	0.26955	NA	NA

Table S1c. Spearman ranked correlation of fold change of focal gene with mean of fold change of all its non-overlapping neighboring (within ±100Kb) genes. All statistics are significant after Bonferroni testing.

Tissue	Male *P-value*	Male *Rho*	Female *P-value*	Female *Rho*
Brain	3.51E-09	0.05254	3.43E-10	0.05617
Cerebellum	4.81E-55	0.14054	2.47E-59	0.14527
Kidney	9.39E-251	0.29607	9.68E-26	0.09391
Heart	4.99E-109	0.19835	4.44E-42	0.12333
Liver	2.30E-53	0.13815	NA	NA
Testis	5.32E-178	0.24651	NA	NA

Table S2. Concordance between genes called as of changed expression by Z score method and the method of Brawand et al. Mann-Whitney test comparing Z scores of the genes shown by Brawand et al to have significantly shifted their expression compare to the rest of the genes shows significant difference between Z score of the two groups. The number of genes reported as significant in Brawand et al. is very low in some tissues.

Tissue	Male *P-value*	Female *P-value*	Number of genes in Brawand's significant genes list
Brain	0.00068	0.00177	4
Cerebellum	9.87E-88	4.71E-94	268
Kidney	1.42E-42	8.04E-75	207
Heart	6.01E-07	1.16E-06	10
Liver	3.42E-36	NA	80
Testis	4.83E-154	NA	567

Table S3a. GO processes enriched in concerted Z score positive genes across tissues in male samples

GO term	Description	*P-value*
GO:0050909	Sensory perception of taste	1.97E-5
GO:0007606	Sensory perception of chemical stimulus	1.41E-4
GO:0009968	Negative regulation of signal transduction	1.48E-4
GO:0048585	Negative regulation of response to stimulus	7.82E-4
GO:0060041	Retina development in camera-type eye	8.78E-4
GO:0017038	Protein import	9.03E-4

Table S3b. GO processes enriched in concerted Z score negative genes across tissues in male samples

GO term	Description	*P-value*
GO:0016202	Regulation of striated muscle tissue development	4.99E-4
GO:1901861	Regulation of muscle tissue development	4.99E-4
GO:0048634	Regulation of muscle organ development	4.99E-4

Table S4a. In Brain Z+ clusters, in comparison to Z- clusters, are significantly enriched in gained H3K4me3 peaks in human compared to chimps and macaques in both males and females.

Tissue	#Clusters	#Positives Cluster	#Negatives Negative	#Positives cluster gained at least one H3K4me3	#Expected positives clusters gained	#Negatives cluster gained H3K4me3	#Expected Negatives clusters at least one gained	Chi Square *P-value*
Female	5975	2987	2988	256	220.46	185	220.54	11.457 P-value <0.001
Male	6053	3023	3030	275	226.24	178	226.76	20.994 P-value <<0.001

Table S4b. In brain, Z- clusters, in comparison to Z+ clusters, are significantly enriched in depleted H3K4me3 peaks in human compared to chimps and macaques only in females.

Tissue	#Clusters	#Positives Cluster	#Negatives Negative	#Positives cluster depleted at least one H3K4me3	#Expected positives gained	#Negatives cluster depleted at least one H3K4me3	#Expected Negatives gained	Chi Square *P-value*
Female	5975	2987	2988	39	71.49	104	71.51	29.527 P-value <<0.001
Male	6053	3023	3030	62	66.92	72	67.08	0.722 p-value ~0.4

Table S5a. *P-values* of Mann-Whitney test comparing physical dimensions of Z+ and Z- clusters

Tissue	Median length of Z+ clusters (Male)	Median length of Z- clusters (Male)	Male *P-value*	Median length of Z+ clusters (Female)	Median length of Z- clusters (Female)	Female *P-value*
Brain	100813	81953	1.74E-06	90674	91422	0.86969
Cerebellum	90838	96469	0.15828	92790	102161	0.05200
Kidney	98079.5	115206	2.33E-07	83051.5	99106	0.00382
Heart	87304	115284	7.89E-07	91020	112959	2.91E-05
Liver	73028	106214	4.96E-17	NA	NA	NA
Testis	97548	106808	0.00039	NA	NA	NA

Table S5b. *P-values* of one-tailed Monte Carlo simulation to determine whether the difference observed between density of Z+ and Z- clusters could have happened by chance, as a function of the number of genes in a cluster. For this simulation, the number of Z+ and Z- genes are kept the same as observed and gene order is randomized. In each iteration, the number of occurrence of clusters of specified size is counted, if it is great or greater than the observed number of clusters of that size, Monte Carlo counter is incremented. Empirical P is then calculated after 1000 iterations.

Tissue/ Gender	*P-value* per Gene Cluster containing this number of genes											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Brain/M	0.9980	1.0000	0.4376	0.9940	0.5554	0.6943	0.9301	0.8262	0.9031	0.6913	NA	NA
Cere./M	1.0000	0.9950	0.0270	0.0130	0.0959	0.0090	0.0350	0.0230	0.0290	0.5504	0.0230	0.0819
Kidney/M	0.9990	0.2527	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0160	0.0020	0.0010
Heart/M	1.0000	0.3127	0.0010	0.0010	0.0010	0.0010	0.0010	0.0200	0.0010	0.0420	0.0010	0.0010
Liver/M	0.1968	0.0010	0.0010	0.0010	0.0010	0.0030	0.0789	0.5724	0.0440	0.7592	0.2308	NA
Testis/M	0.8511	0.5015	0.0310	0.0020	0.0010	0.0010	0.0010	0.0010	0.0010	0.0020	0.0020	0.0010
Brain/F	0.8971	0.4525	0.4236	0.9530	0.8591	1.0000	0.3137	1.0000	0.4236	1.0000	NA	NA
Cere./F	0.9980	0.9880	0.7493	0.0010	0.0999	0.0010	0.0420	0.2238	0.0310	0.1039	0.3267	NA
Kidney/F	0.9970	0.2907	0.0010	0.3586	0.3117	0.5544	0.2807	0.1888	0.0559	0.7732	NA	NA
Heart/F	1.0000	0.8561	0.1269	0.0060	0.0010	0.0100	0.0080	0.0020	0.0050	0.5954	0.2817	NA

Cere. = Cerebellum
M = Male
F = Female

Table S6a. There are more strictly tissue-specifically up-regulated genes in Cerebellum and Liver than expected by chance. If we define strictly tissue specific up-regulated genes as the genes with Z score of more than one in one tissue and zscore of negative or zero in all other tissues and then estimate their number in each tissue based on the ratio of total number of strictly tissue specific up-regulated genes to the total number of expressed genes (across all tissues), and define expected number of strictly tissue specific genes by number of genes which are expressed in that tissue multiplied to this ratio, the observed number of tissue-specific genes in Cerebellum and Liver are significantly higher than the number expected by chance. Number of strictly tissue-specific genes is significantly below the expected number in Brain and Heart.

Tissue	#Genes Expressed (RPKM>2)	#Strictly Tissue Specific Genes	Expected	Chi-Squared	*P-value*
Brain	10020	7	29.3038	16.9759	<<0.01
Cerebellum	9400	45	27.4906	11.1522	<0.01
Kidney	9619	15	28.1311	6.1293	~0.02
Heart	9328	13	27.2800	7.4750	<0.01
Liver	8917	51	26.0780	23.8171	<<0.01
Testis	10845	39	31.7165	1.6726	~0.2

Table S6b. There are more tissue-specifically up-regulated genes in Cerebellum than expected by chance. By relaxing the definition of tissue-specific up-regulated genes to genes with Z>1 in one and only one tissue, more tissue-specifically up-regulated genes are detected. If we estimate number of tissue specifically up-regulated genes in each tissue based on the ratio of total number of tissue specifically up-regulated genes to the total number of expressed genes (across all tissues), and define expected number of tissue specific genes by number of genes which are expressed in that tissue multiplied to this ratio, the observed number of tissue-specific genes in Cerebellum is significantly higher than the number expected by chance. Number of tissue specific genes is significantly below the expected number in Brain and Heart.

Tissue	#Genes Expressed (RPKM>2)	#Tissue Specific Genes	Expected	Chi-Squared	*P-value*
Brain	10020	39	557.9786	482.7045	<<0.01
Cerebellum	9400	1230	523.4530	953.6838	<<0.01
Kidney	9619	524	535.6484	0.2533	~0.6
Heart	9328	365	519.4436	45.9199	<<0.01
Liver	8917	457	496.5564	3.1511	~0.07
Testis	10845	622	603.9200	0.5413	~0.55

Table S6c. If a focal tissue-specific up-regulated gene has a neighbor closer than 100Kb, overall this neighbor is more likely to be up-regulated. Binomial test between the number of tissue-specific up-regulated genes with at least a neighbor in ±100Kb to the number of these genes whose closest neighbor in 100Kb is also a Z+ gene, Chi-Squared = 68.03, p-value<<0.01. Overall 59.45% of these genes have a Z+ gene as their closest neighbor. Expected is the expected number of tissue specifically up-regulated genes with a neighbour up-regulated in the same tissue. Tissue-specific up-regulated genes are defined as explained in table S6b.

Tissue	#TSU* with a neighbor in ±100Kb	#TSU* with Z+ neighbor in ±100Kb	%with Z+ neighbor in ±100Kb	Expected Value	Binomial *P-value*
Brain	32	19	59.375	15.731	0.2902
Cerebellum	1076	606	56.3197	537.9587	3.33E-05
Kidney	501	348	69.4611	250.4808	1.16E-18
Heart	336	196	58.3333	167.9871	0.0022
Liver	413	220	53.2688	206.4841	0.1844
Testis	547	338	61.7916	273.479	3.09E-08

* TSU: Tissue-Specific Up-regulated genes

Table S6d. If a focal tissue-specific up-regulated gene lacks a neighbor closer than 100Kb, overall its closest neighbor is less likely to be a Z+ gene. Binomial test between the number of tissue-specific up-regulated genes without any neighbor in ±100Kb to the number of these genes whose closest neighbor is a Z+ gene. Expected value for the number of tissue-specific up-regulated genes having a Z+ as their closest neighbor is calculated by multiplying number of tissue-specific up-regulated genes to probability of a gene being Z+ in the corresponding tissue, Chi-Squared = 12.43, p-value<0.04. Overall 39.46% of these genes were observed to have a Z+ gene as their closest neighbor. Tissue-specific up-regulated genes are defined as explained in table S6b.

Tissue	#TSU* without a neighbor in 100Kb	#TSU* with Z+ as its closest neighbor	%with Z+ closest neighbor	Expected Value	Binomial *P-value*
Brain	7	4	57.1429	3.4412	0.7222
Cerebellum	154	57	37.013	76.9941	0.0016
Kidney	23	14	60.8696	11.4991	0.3074
Heart	29	13	44.8276	14.4989	0.7111
Liver	44	21	47.7273	21.9983	0.8804
Testis	75	22	29.3333	37.4971	0.0004

* TSU: Tissue-Specific Up-regulated genes

Table S6e. Genes with Tissue-specific upregulation are not clustered. The number of tissue-specific up-regulated (TSU) genes with a TSU downstream neighbor in the same tissue is shown. *P-value* of one-tailed Monte Carlo simulation is also reported. For this simulation the number of TSU genes are kept the same as observed in corresponding tissue but their gene order is randomized by shuffling all the genes including TSU ones in each tissue to ask if similar number of TSU downstream neighbors could have happened just by chance under the null of random gene order. This is repeated for 10,000 times and the number of TSU genes with a TSU downstream neighbor, regardless of their distance to the focal gene, is counted; If equal or greater than the observed number of genes with a TSU downstream neighbor, Monte Carlo counter is incremented and overall P-value calculated. We also show, for information, the proportion of TSU pairs where they are within 100kb of each other. Tissue-specific up-regulated genes are defined as explained in table S6b.

Tissue	#TSU* genes	#Downstream neighbour is also TSU in the same tissue	%Within 100Kb	Monte Carlo Mean (SD)	Monte Carlo P-value
Brain	39	0	0	0.1135 (0.3347)	1
Cerebellum	1230	128	77.343	116.0135 (9.661)	0.11979
Kidney	524	28	92.857	21.056 (4.407)	0.07629
Heart	365	13	92.307	10.1909 (3.082)	0.22058
Liver	457	20	85	16.0068 (3.865)	0.17988
Testis	622	0	0	29.5799 (5.112)	1

* TSU: Tissue-Specific Up-regulated

Table S7. Spearman correlation between female and mean of male Z scores per tissues, without removing zero Z scores

Tissue	*Rho*	*P-value*
Brain	0.52967	<<0.0001
Cerebellum	0.32532	9.01e-319
Heart	0.45401	<2.2e-16
Kidney	0.43073	<2.2e-16

Tables S8a. Spearman correlation between standard residual of standard major axis estimation between Z of male and female for a focal gene and residual of its nearest downstream neighbor, without removing zero Z scores

Tissue	Non-overlapping *P-value*	Non-overlapping *Rho*	Overlapping *P-value*	Overlapping *Rho*
Brain	0.00018	0.03994	0.00325	0.10408
Cerebellum	0.03109	0.02304	9.10E-06	0.15636
Heart	1.42E-05	0.04638	8.04E-05	0.13913
Kidney	6.95E-19	0.09465	0.01206	0.08883

Table S8b. Spearman correlation between standard residual of standard major axis estimation between Z of male and female for a focal gene and mean of residuals of its two nearest neighbors, without removing zero Z scores

Tissue	Non-overlapping *P-value*	Non-overlapping *Rho*	Overlapping *P-value*	Overlapping *Rho*
Brain	5.71E-06	0.05461	0.00613	0.09694
Cerebellum	0.002448527	0.03648	0.00076	0.11884
Heart	5.86E-09	0.07003	0.00019	0.13129
Kidney	3.24E-23	0.11913	0.00018	0.13190

Table S8c. Spearman correlation between standard residual of focal gene and the mean of standard of residual of all neighbors within 100kb of the focal gene, without removing zero Z scores

Tissue	Spearman *P-value*	Spearman *Rho*
Brain	4.00E-08	0.04816
Cerebellum	0.00847	0.02310
Kidney	1.71E-39	0.11504
Heart	1.87E-05	0.03755

Table S9. No evidence for X chromosome enrichment in sex biased genes. If top 5% of the genes are selected based on their standard residual to standard major axis estimation between Z of male and female, no enrichment of x linked genes is observed compared to autosomal genes. Brawand's dataset includes 466 genes on chromosome X and 12561 genes on autosomes. Note, genes with Z=0 are not excluded from this analysis so expected number of genes are the same across different tissues.

	Observed X	Expected X	Observed Autosome	Expected Autosome	Chi Squared	*P-value*
Brain	21	23.3	628	628.05	0.2270	>0.6
Cerebellum	17	23.3	632	628.05	1.7283	>0.15
Heart	26	23.3	623	628.05	0.3535	>0.5
Kidney	21	23.3	628	628.05	0.2270	>0.6

Fig S1. The genes Brawand et al found to have changed their expression significantly in human also have high Z score. Genes listed with significant lineage-specific expression switches in human in Brawand et al. are divided in two groups: Brawand's significant genes which are Z score positive (in blue) and the ones with negative Z scores (in yellow). These two subsets are then plotted against the rest of the genes, shown in green, across different tissues. There was no Z score negative gene in 4 genes Brawand found to have significantly changed their expression in human brain. Please also note that genes found to have extremely high or low Z score have been removed from this plot to improve clarity of comparison between our method and Brawand et al's method.

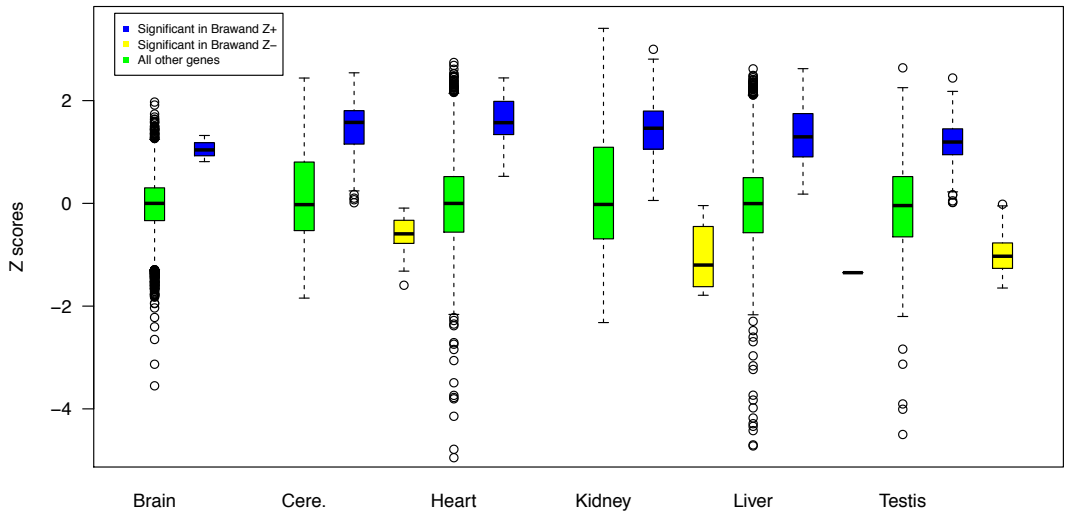


Fig S2a. Examples of large positive Z clusters

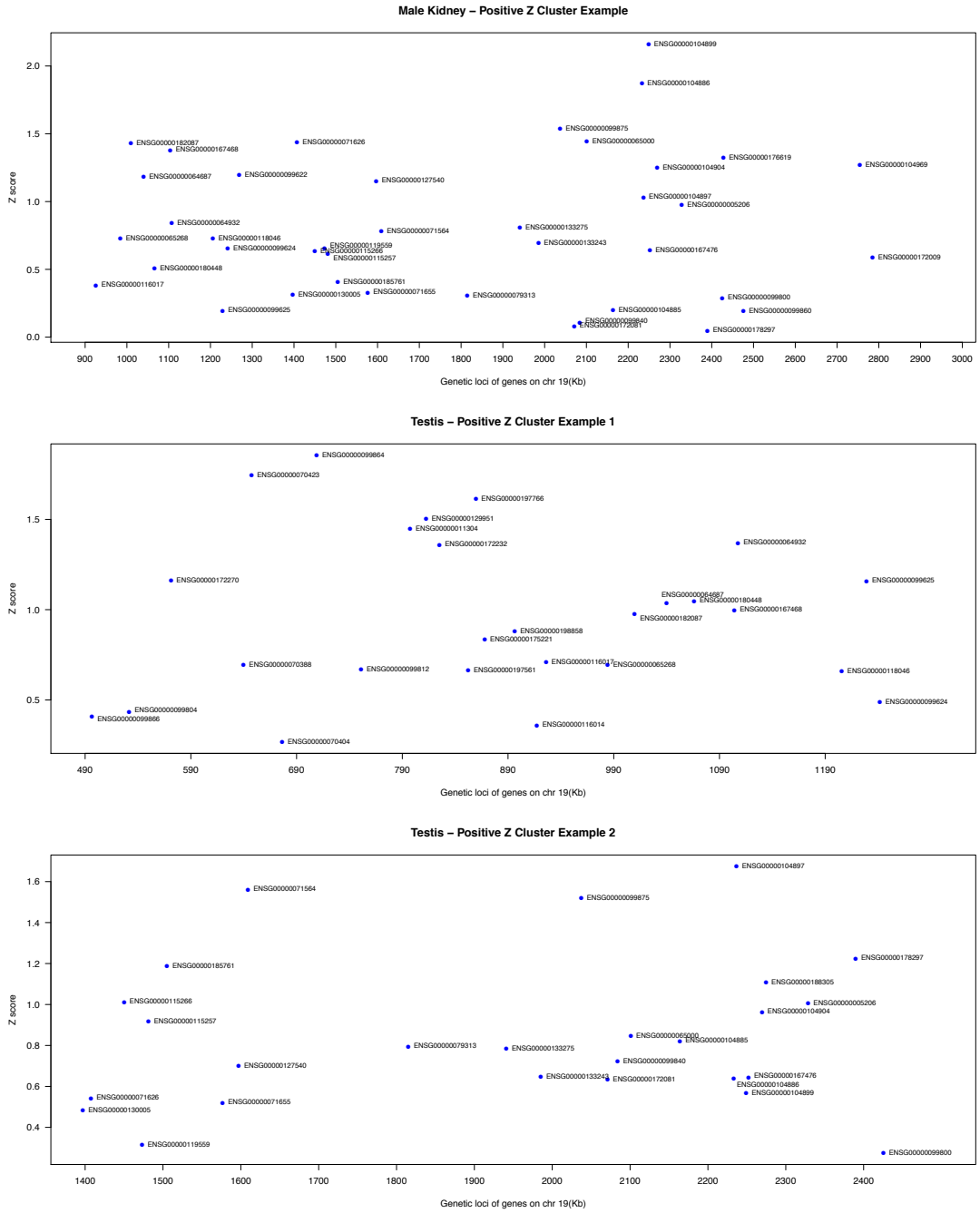


Fig S2b. Examples of large negative Z clusters

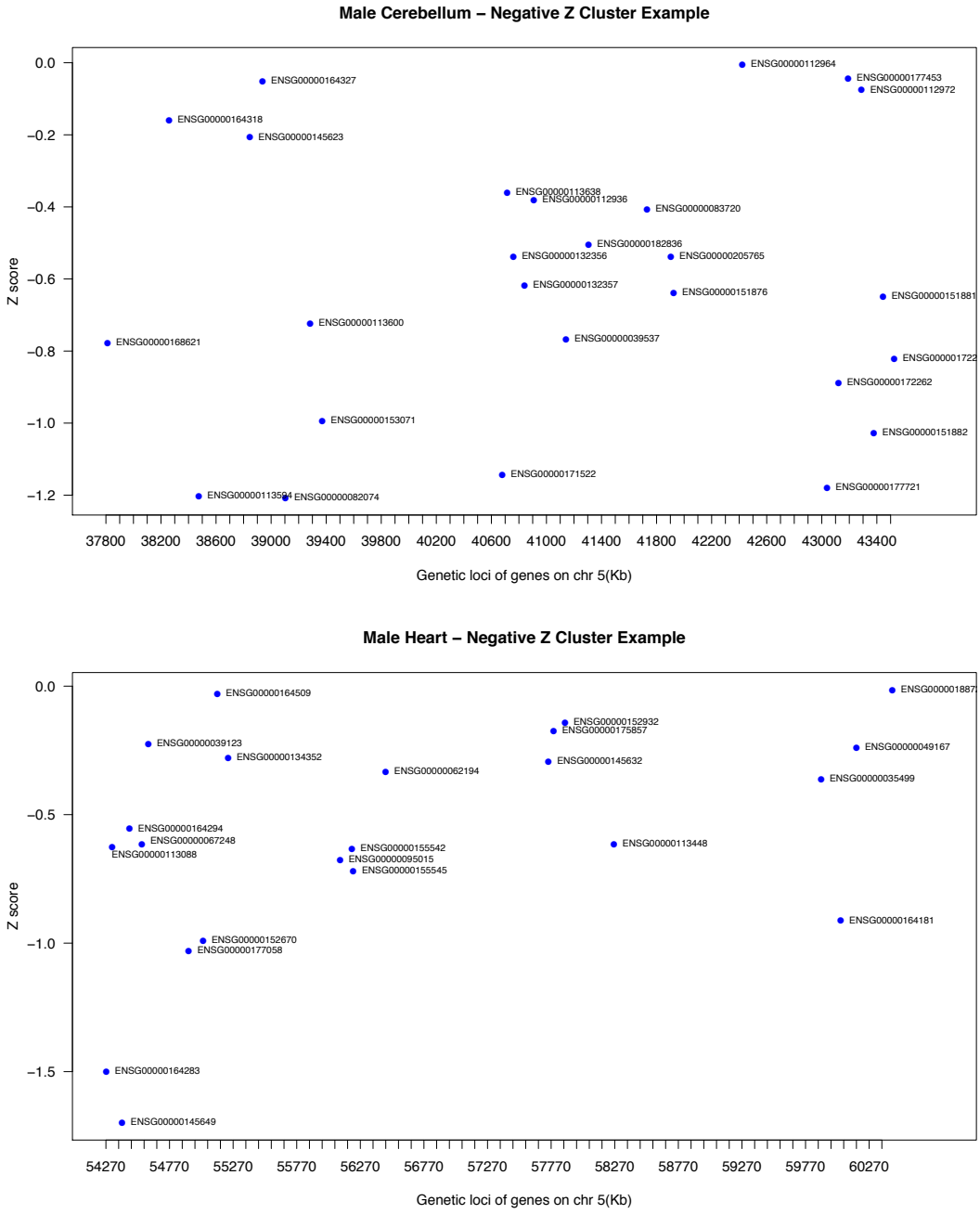


Fig S3. Genes showing increase in expression across all tissues or decrease in expression across all tissues are more common than expected by chance. The sign of Z score indicates the direction of change in gene expression across time (from common ancestor to human). One may represent Z positive genes with +1 and Z negative with -1, keeping zero Z scores unchanged, and calculate for each gene their sum across 6 male tissues. Genes with sums equal to +6 (or -6) have gone up (or down) across all 6 tissues. The number of concerted up-regulated genes is shown by the right-most bar in histogram and the number of concerted down-regulated genes is shown by the left-most bar. Null expectations are shown in blue. Null is derived from randomizations in which the number of Z+, Z- and Z=0 genes are kept the same as observed, but the gene order is randomised across tissues. Errors bars are +/-SD from 1000 simulations.

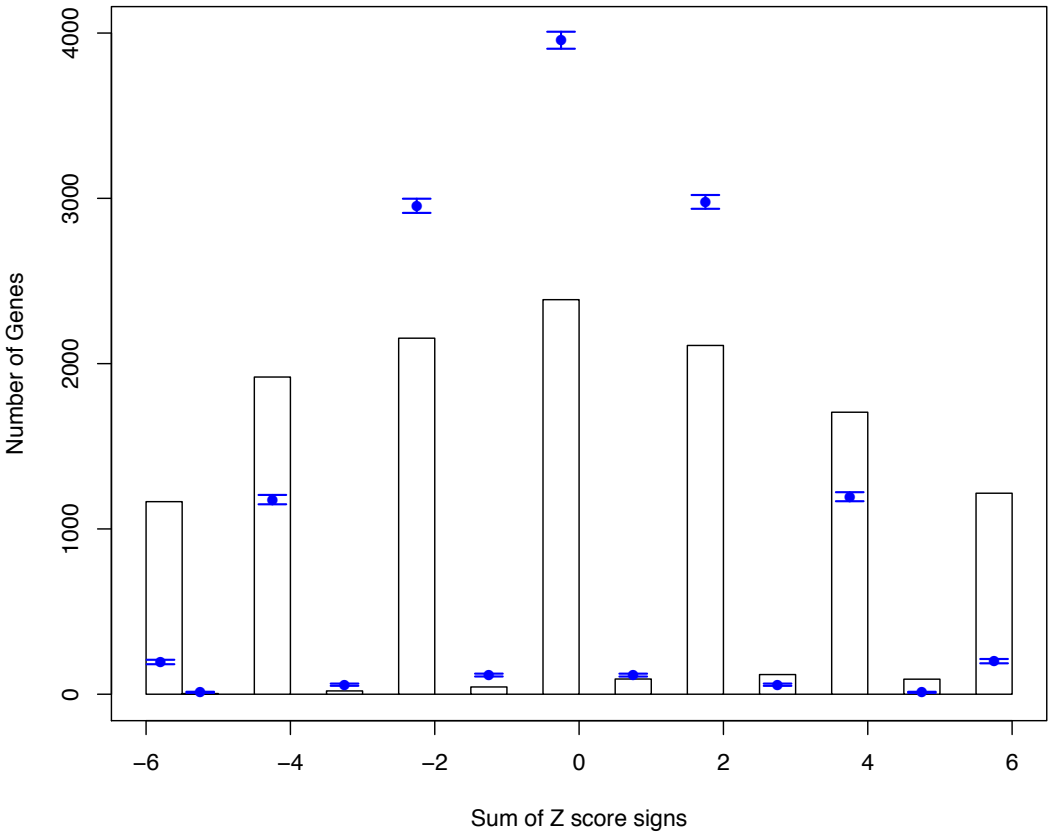


Fig S4. Positive Z score clusters are denser in most tissues except in brain. Positive Z score clusters are shown in green and negative ones in red. The effect is most pronounced in male kidney.

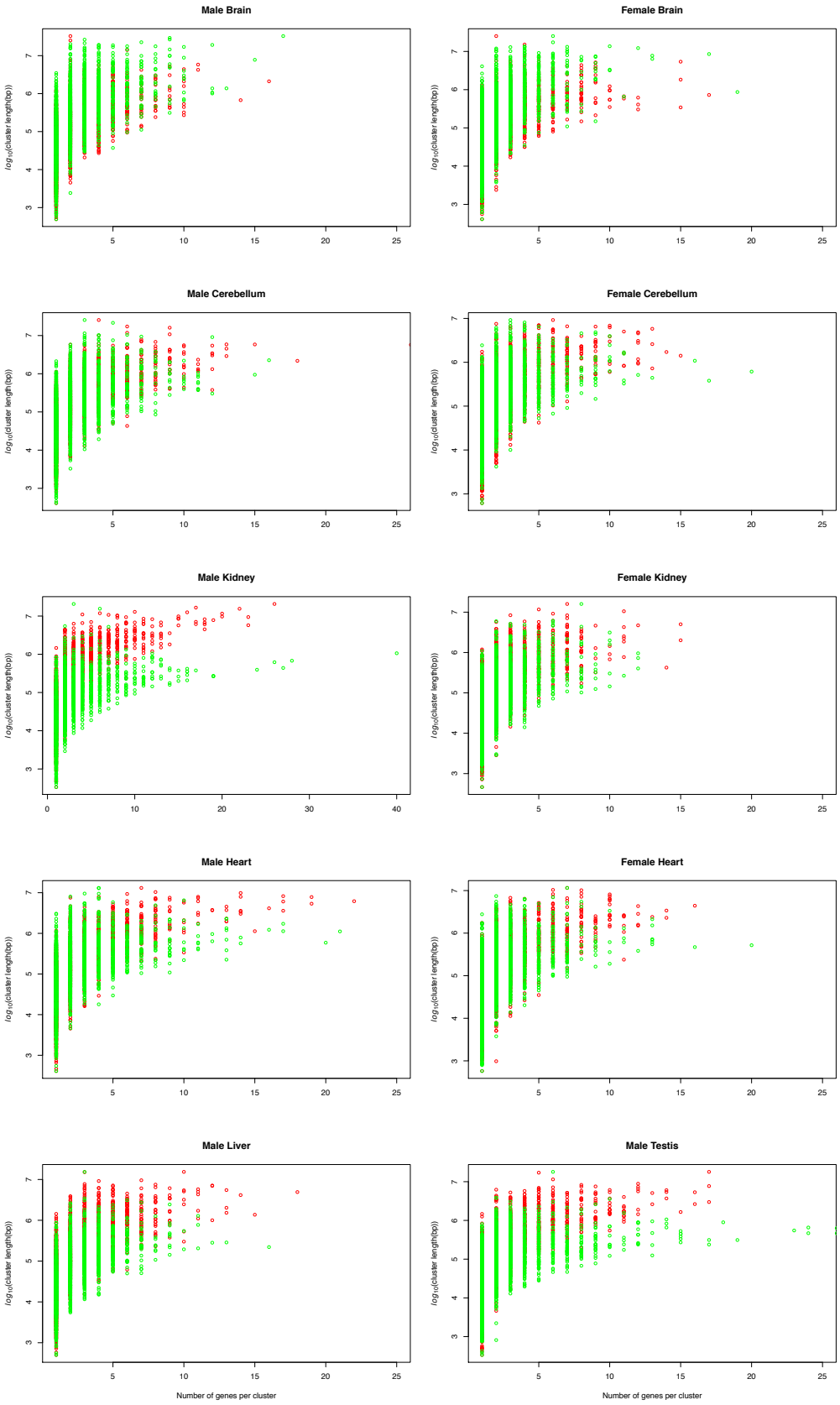


Fig S5. The longest concerted Z score positive cluster based on expression profile in males

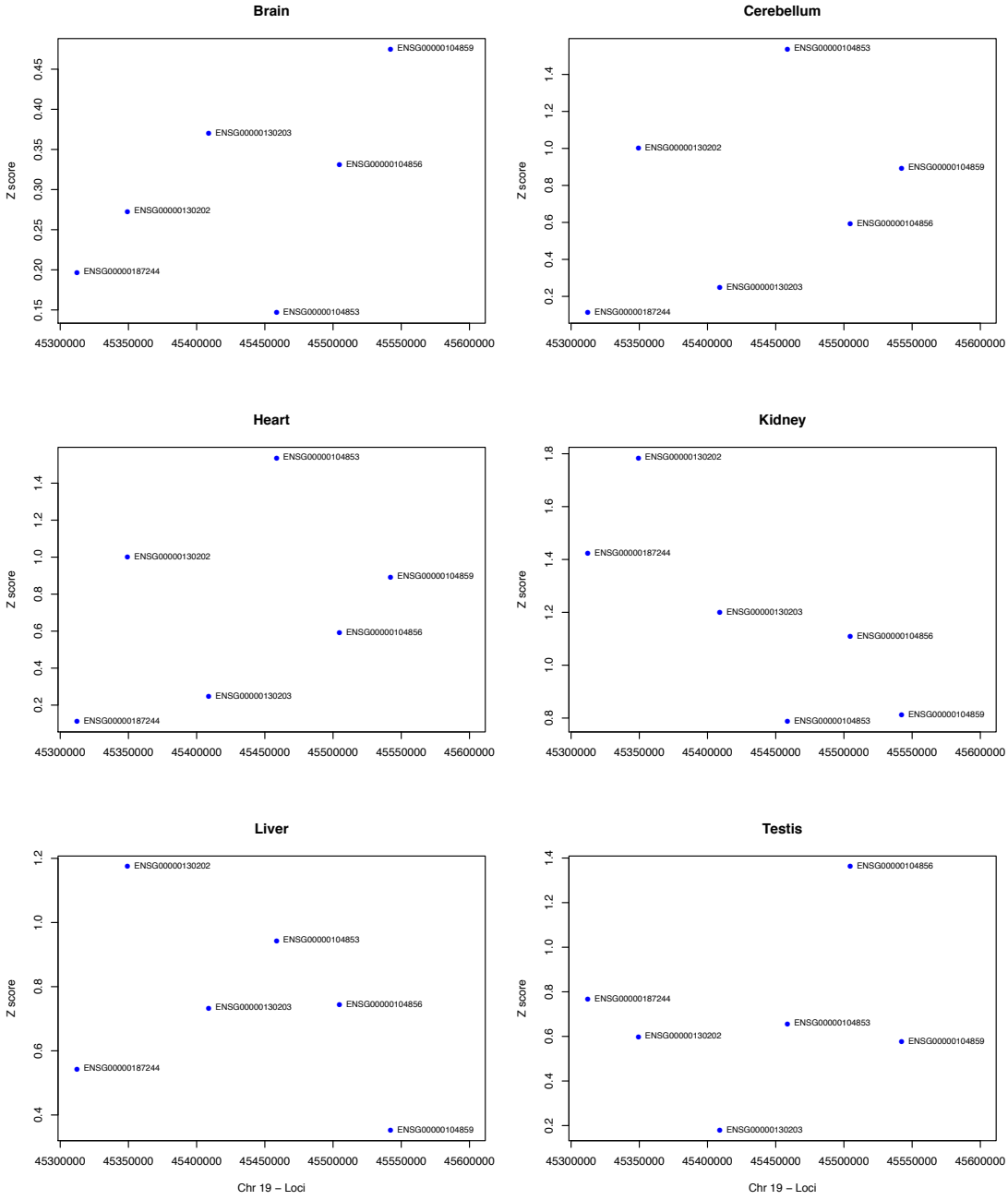


Fig S6. Tissue-specific up-regulated genes tend to be in domains of low gene density compared to genes with concerted change in expression across all 6 tissues. Number of genes in ±100Kb window around the tissue-specific up-regulated and concerted genes are shown below. There is a significant difference between number of genes in ±100Kb of focal genes is tissue-specific genes compared to concerted ones, Mann-Whitney U test *P-value* = 1.26e-43.

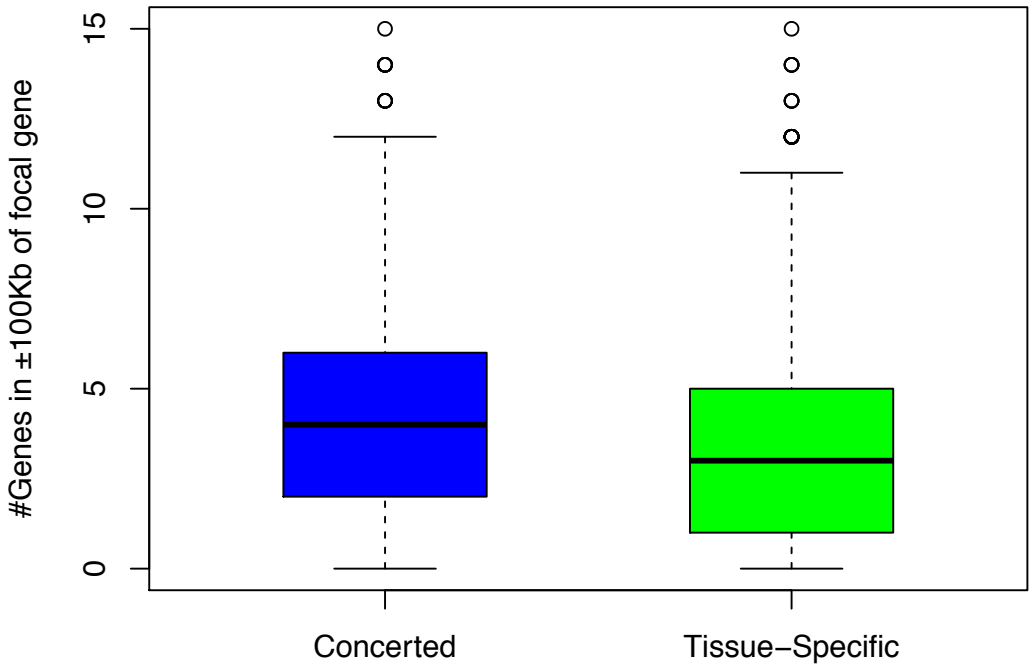
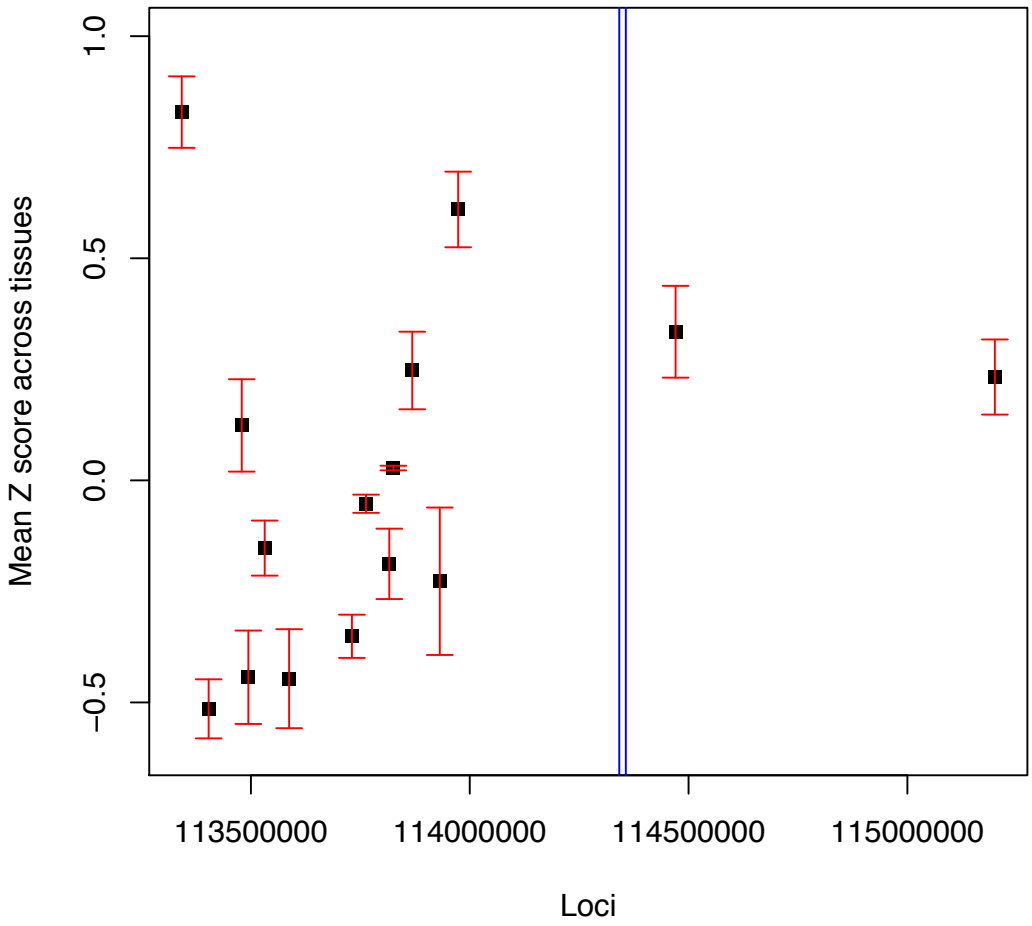


Fig S7. Mean of Z scores of genes across 6 tissues in the vicinity of fusion point on human chromosome 2. Means are averages across all tissues. Error bars indicate SEM. The vertical line indicates the approximate location of the fusion zone.



Chapter 3. Piggybacking neighbors, Part 2:

Evolution of gene expression in Yeasts

The following section is to be submitted to Journal of Molecular Evolution as a single author paper.

Evidence for small scale expression piggybacking in a compact genome

Avazeh T. Ghanbarian[1,2*]

1. Department of Biology and Biochemisty, University of Bath, Bath, BA2 7AY
2. Wellcome Trust - Medical Research Council Stem Cell Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QR

Abstract As short term upregulation of a gene causes, in humans and yeast, a time-lagged upregulation of genes in the near vicinity (this being a so called ripple effect), it was conjectured that, over evolutionary time the change in expression of one gene might also correlate with that of the neighbors. Indeed this was recently shown to be true in primates. The notion that evolutionary change in expression of one gene might be correlated with that of the neighbors was dubbed expression piggybacking. In primates the ripple effect extends to circa 100kb and the piggybacking is most acute within this span but extends further. In yeast the ripple span is more limited, extending to only a few kb (~3kb). Do we then see correlated changes in expression between neighboring genes in yeast and is the span much more limited than seen in primates? Here I examine these hypotheses, testing results for resilience to phylogenetic assumptions. I find that gene expression evolution within the yeasts is indeed correlated on a local scale and that the scale (< 3kb) is very much more limited than seen in primates. These results indicate that expression piggybacking is seen in highly compact genomes but that compaction is also associated with stronger insulation of gene expression change.

Keywords: Evolution of gene expression, Yeast, Ripple effect

Introduction

Classically when we think about the evolution of gene expression, we tend to think in gene centric manner. That is to say, we look, for example, for changes in the promoter of a given gene to explain the changes in the expression of this gene. Recently, this view has been challenged. Starting from the observation that a short term increase in a gene's expression leads to a commensurate increase in expression of the neighbors, the so-called ripple effect (Ebisuya et al. 2008), it was asked whether, on an evolutionary time scale, the change in expression of a gene correlates with that of its neighbors. Such changes could explain, for example, the regularly observed similarity of gene expression of neighboring genes (over small and broad scale) seen in eukaryotes (Caron et al. 2001; Cohen et al. 2000; Lercher et al. 2003; Michalak 2008; Pal and Hurst 2003; Purmann et al. 2007; Williams and Bowles 2004; Woo and Li 2011). Through analysis of expression change in humans since the human chimp common ancestor, it was found that across both sexes and all tissues

* Corresponding author, atg20@bath.ac.uk

examined, change in expression of a given gene was indeed correlated with that of the neighbors. This coupled alteration in gene expression of physically clustered genes was termed expression piggybacking (Ghanbarian and Hurst 2015).

In the case of primates, the domain of influence is most acute in the immediate vicinity of the focal gene. However, while the ripple effect extends only to circa 100kb in primates, the correlation in expression change could be detected at much greater distances. Might expression piggybacking be a peculiarity of primates with their large genomes with abundant intergene DNA and possibly slight selection for tight insulation of changes in gene expression? Yeast has been also demonstrated to have a ripple effect (Ebisuya et al. 2008). Importantly in this instance the ripple is much more localised (circa 3kb) than seen in primates. This might be explained by a greater need to insulate gene expression in genomes with shorter sequences interspersed between genes. Here then I ask whether yeast also shows evidence of piggybacking (i.e. correlated expression evolution between physically linked genes) and whether the dimensions of such clusters are more restricted than seen in primates. Availability of well-annotated genomes and comparative transcriptome data in 4 Yeasts (Busby et al. 2011), makes yeast species a suitable model to investigate these questions.

Materials and Methods

Gene expression data

Gene expression data was obtained from Busby et al (2011). For the analysis reported in this paper, cross-species expression values reported in 1471-2164-12-635-s2.xls was complemented with annotations from the Saccharomyces Genome Database, http://downloads. yeastgenome.org accessed on 26'th, Feb, 2015, (Cherry et al. 2012). An ad-hoc script is used to find 2963 homologous genes across all 4 Yeasts included in this study, *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*. Read counts for these homologous genes are reported for two replicates in Busby et al (2011). Level of expression in *S. cerevisiae* and *S. paradoxus*'s ancestor is then estimated by employing BayesTraits (Pagel et al. 2004).

Estimating ancestral level of gene expression

The changes in expression of a gene over evolutionary time can be estimated by standard Z score and/or fold change. To calculate either, we first need to estimate the level of gene expression in the ancestor of the species of interest. For this study, I calculate the change in expression of *Saccharomyces cerevisiae*'s genes to their estimated expression in *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*'s common ancestor. While inclusion of expression values in at least one outgroup species is necessary in this estimation, the accuracy can be improved by including more outgroups. Hence expression values in *S. mikatae* and *S. bayanus* were added as two outgroups. To this end, a Bayesian model of gene expression across the four yeasts was built. Through applying a Bayesian method, BayesTraits, to the read counts per gene across homologous genes in these 4 yeast species, the level of expression in the common ancestor of *S. cerevisiae* and *S. paradoxus* can be estimated. Subsequently the difference between current level of expression in *S.*

cerevisiae to this ancestor is calculated as a Z score and fold change, the formulas are explained in detail below. The method applied here is similar to the one we employed previously to study evolution of gene expression in Primates (Ghanbarian and Hurst 2015), except for a couple of necessary changes. First to account for yeasts' sensitivity to minute changes in the experimental conditions and second to examine the effect of chosen phylogenetic tree on the estimated change in gene expression in yeast as explained below.

Addressing variation across replicates

Z score and fold change are first calculated based on the mean of read counts per gene across two replicates. The analysis based on these Z score or fold change values are tagged as "Mean" across the results reported here. Nonetheless, Busby et al attributed more than 60% of observed variance between species to their environmental response or measurement imprecision (Busby et al. 2011). Hence to account for yeasts' sensitivity to environmental changes and experimental conditions, the level of gene expression in ancestor is also estimated in each replicate separately. Z score and fold change calculated based on these estimates are tagged as "Rep1" or "Rep2" across all analysis reported here. Having these three separate measures available per homologous gene, helps to tease apart the variation triggered by environmental response or measurement imprecision from the variation implied by the changes in gene expression during evolutionary time. Variation caused by environmental response are assumed to be due to unaccounted experimental conditions, since Busby et al. claimed the growth conditions and other experimental conditions to be the same across both replicates.

Addressing the difference across prior phylogenetic trees

As mentioned above, the Bayesian method, BayesTraits, requires a prior phylogenetic tree to be able to build an accurate model of gene expression across 4 yeasts in this study. Phylogenetic relationships are still disputed across many taxa (Rutschmann 2006). While the ubiquitous agreement on approximate divergence date of 5 primates considered in the Primates study (Ghanbarian and Hurst 2015), made analysing the effects of prior phylogenetic tree used unnecessary, this could make a difference in expanding this analysis to other species. In species lacking general agreement on their phylogenetic relationship, especially their divergence date, we might ask how the choice of phylogenetic relationship in estimation of the ancestral state might affect the correlation between evolutionary changes in expression of the neighboring genes. In particular, it would be relevant to compare the difference in using two specific phylogenetic trees: one inferred based on the substitution rate in genic regions and the second based on substitution rate in intergenic regions. This would be especially important to test hypotheses regarding not only the resilience to phylogenetic assumptions but also the regulatory mechanisms involved in controlling the correlated change in expression. This would be particularly interesting since intergenic regions are enriched in such regulatory elements, like promoters and enhancers, compared to the gene bodies. By considering these two phylogenetic trees, one may ask if the magnitude or domain of correlation in change in gene expression is substantially and consistently different when genic or intergenic phylogenetic trees is applied. For this comparison, I used the genic and intergenic phylogenetic trees produced by Kellis et al (2003) based on the substitution rate in genic and intergenic sequences across the same 4 yeasts included in the study reported here

Measure of change in gene expression

If one is to use mean of replicates, Z score is calculated as follows. If the mean read count of a given gene is $E_{current}$, or E_c in abbreviated form, and its variance is V_c, while that for the ancestral condition is E_a and V_a, then we can define the degree of expression divergence in *S. cerevisiae* lineage from cerevisiae-paradoxus ancestor as a Z score calculated as below:

$$Z = \frac{E_c - E_a}{\sqrt{V_c + V_a}}$$

If one is to consider each replicate separately, E_c would represent the read count in corresponding replicates and V_c would be equal to zero, as there is no variance in a single read count for each gene in each replicate. Hence Z score formula would be reduce to the one below:

$$Z = \frac{E_c - E_a}{\sqrt{V_a}}$$

Nevertheless E_a and V_a could be calculated based on different percentage of BayesTraits estimates, as BayesTraits provides us with a list of estimated values rather than just one value. So as to gauge what percentage of estimates should be used as a robust estimation of expression in cerevisiae-paradoxus ancestor, a benchmark was applied. For each gene, Z scores were calculated based on all or last 50%, 20% and 10% of BayesTraits estimate. Distribution of Z scores including their median, maximum and minimum is shown in figs S1 and S2, in supplementary materials. Under the assumption that size of transcriptome has not changed since this ancestor, median of Z score across all genes should be close to zero. This is because if as many genes increase their expression as there are genes decreasing their expression, median of Z score would be near zero. Overall, using last 10% of BayesTraits' estimates produces medians closer to zero, across different phylogenetic trees and strategies to treat replicate read counts. So all Z scores and fold changes reported in the body of this paper are calculated based on using last 10% of BayesTraits estimate as the measure of ancestral level of gene expression.

Anyway, Z score metric compares the extent of difference between current level of expression and ancestral level, scaled by the degree of variation both in current estimates (expression noise or measurement error) and the degree of uncertainty in the ancestral state's estimation. In other words, Z score measures the difference in standard deviation units. That is, a gene with largely variable expression across individuals or high fluctuation and uncertainty in estimation of expression in ancestor would have a lower Z score compared to a gene with similar but steadier level of current expression and/or one with similar but more stable estimation of ancestral level of expression. Also a positive Z implies an increase in gene expression since the ancestor and a negative Z shows a gene which is down-regulated since the ancestor.

We also calculated a second measure of change in gene expression: fold change. We define fold change simply by dividing current level of expression, or its mean if

applied, to the mean of estimated level of expression in ancestor in the last 10% of BayesTraits estimates.

Median Correction

Assuming an absence of net increase or decrease in overall expression levels requires a minor adjustment of the Z scores for all genes in all tissues. If the median Z in any given tissue in a given sex is M, then we define modified Z as Zmod = Z – M. This forces all datasets to have a modified median of zero and as many genes increasing in expression as decreasing, as the current net transcriptome size is assumed not to be any different from that of the ancestor. All analyses were performed on Zmod. Henceforth we shall refer to Z, for convenience, where Zmod is what we are employing. In practice the correction makes little or no difference as a) the correction is usually very small and b) many of our statistics are rank order based and so unaffected by the modification.

Results

Neighboring genes correlate significantly in their evolution of gene expression regardless of phylogenetic tree applied

Regardless of the approach to treat replicates separately or use their average, or even irrespective of the phylogenetic tree used in estimating the gene expression model, there is a significant correlation between the change in gene expression of the focal gene and that of its neighboring genes (Fig 1, Tables 1-3). The correlation is also consistently significant and in positive direction irrespective of the measure of change in gene expression applied, whether it is Z score (shown in tables 1-3) or fold change (shown in tables S1-3 in supplementary material). Moreover, the correlation is always in a positive direction, implying that the change in gene expression of the focal gene predicts the change in gene expression of its neighbors. Across the correlations reported above, only the neighbors within 3Kb boundary of the focal gene are considered, 3Kb being the boundary of expression correlation in yeast as suggested by the ripple effect (Ebisuya et al. 2008).

While this indicates that the prior result from Primates are by no means exceptional, the degree of correlation is far less than what was seen in Primates (Ghanbarian and Hurst 2015). Spearman rho scores in yeasts are about half or even one third of that observed in the Primate study. This might be due to relative compaction of yeast's genome in comparison to Primates. As the intergenic regions are far smaller in Yeasts in comparison to Primates, this could potentially increase the chance of interference in expression leading to a more localised correlation in evolution of gene expression (Kristiansson et al. 2009; Zeitlinger and Stark 2010).

No evidence for genic or intergenic prior phylogenetic trees to have a consistent effect on the correlation in change in gene expression in the neighboring genes

Intergenic regions are enriched in the regulatory elements which could influence up-regulation or down-regulation of the genes. So if the ancestral state of expression is estimated based on intergenic phylogenetic relationship between species, do neighboring gene correlate the change in their expression better? To this end, I used genic and intergenic phylogenic trees and calculated Z score as the measure of change

in gene expression across three sets of neighboring genes and also different strategies to treat replicates (Tables 1-3). Neither a consistent nor a drastic change was found in the correlation of Z scores of the neighboring genes; only a minor disparity was observed in the score and significance of the correlations across the two phylogenetic trees. This is irrespective of the strategy applied in treating the replicates separately or considering their average. In short, while genic tree leads to small increase in correlation of Z score of the neighboring genes in some datasets, intergenic tree improves this correlation slightly in other datasets. For example in the correlation between Z scores of the focal gene and Z score of its closest downstream neighbour closer than 3Kb, shown in Table 1, the use of intergenic phylogenetic tree results in slightly more significant and stronger correlation between the change in gene expression of the focal gene and its closest downstream neighbour when mean of replicates is considered; however, the genic phylogenetic tree yields slightly more significant and stronger correlation in replicate 1. Intriguingly, both correlations are insignificant after Bonferroni correlation regardless of the phylogenetic tree used in replicate 2. Next, when the correlation between Z score of the focal gene to the Z score of mean of closest upstream and downstream neighbors is considered, mean and replicate 2 show an small increase in correlation when intergenic tree is used; while correlation is slightly higher for genic tree in replicate 1. The results stay significant after Bonferroni correction (Table 2). A similar pattern is observed when the correlation between Z score of the focal gene and mean Z score of all its neighbors closer than 3Kb is calculated; again neither of the two phylogenetic trees outperform the other consistently across all mean and replicates sets. Hence, the correlation between the change in expression of the neighboring genes is resilient to phylogenetic assumptions, at least when the trees only moderately differ in their branch length rather than their structure.

Genic phylogenetic tree improves the accuracy in estimation of ancestral state of expression and complies with assumption of constant transcriptome size in most of the datasets

If the transcriptome size has been constant since the ancestor, one would expect to see as many genes increasing their expression as there are genes decreasing their expression. Under such assumption, median of all genes' Z scores is expected to be near zero. Using genic phylogenetic tree to estimate the ancestral level of expression results in lower Z score median in mean and replicate 1 but not replicate 2 (supplementary figs S1-2). In other words, in mean and replicate 1, using genic phylogenetic tree yields Z scores agreeable to the general assumption that the size of transcriptome has not changed drastically since the ancestor. On the other hand, in replicate 2 use of intergenic tree produces Z scores compliant with constant transcriptome assumption.

However, the genic phylogenetic tree consistently produces more extreme Z scores in comparison to Z scores generated by intergenic trees across all mean and two replicates. This could be due to improvement in estimating level of expression in the ancestor for highly up-regulated or down-regulated genes. More accurate estimation of level of expression of such genes in the ancestor would reduce the variation in the estimated values in BayesTraits output. This would finally push Z score of these highly up-regulated or down-regulated genes to more extreme values.

Domain of correlation in change in gene expression is smaller in Yeast compared to Human

So far we have shown that there is a significant correlation between the change in expression of the focal gene and that of its immediate neighbors. This is regardless of the choice of phylogenetic tree or the strategy to treat the replicate information. The results represented so far only include the neighboring genes located in the boundaries suggested by the ripple effect empirical data for yeast, 3Kb. But how far does the correlation between the change in gene expression of the neighboring genes extend? Ebisuya et al have shown the ripple effect in Mammals would cause the neighboring genes to correlate their expression in a time-lagged manner if they are closer than 100Kb to each other. They have also characterised the boundaries of this ripple effect in Yeasts, which only expands to 3Kb of the focal loci of expression (Ebisuya et al. 2008). In our previous analyse of evolution of gene expression in Primates, we found evidence for the span of local correlation to extend an order of magnitude further than ripple effect's suggested boundary (Ghanbarian and Hurst 2015). In some tissues a significant correlation was observed even as far as tens of mega base pairs away from the focal gene. So does the dimension of correlation similarly exceed the boundary suggested by ripple effect in yeasts?

The correlation between Z score of the neighboring genes sitting x-min kilobase pairs away was calculated for the gene pairs at 1Kb, 2Kb and 3Kb (Table 5). While the correlation in change in gene expression is significant for the genes closer than 1Kbs away, it fell into insignificance for the neighboring genes in the next bit. Extending the analysis to 3Kb, as ripple's boundary in yeasts, restores the significant correlation between the neighboring genes across all but one of the datasets, rep1 dataset when intergenic phylogenetic tree is used. So the boundary of significant correlation in evolution of gene expression of the neighboring genes in yeast is perhaps smaller than the ripple effect's boundary. The limited dimensions of the zone of piggy-backing could be due to smaller intergenic regions in yeasts compared to larger intergenic regions in Primates. Smaller intergenic regions might increase the chance of transcriptional interference leading to a more localised correlation in gene expression.

It needs to be mentioned that all overlapping genes have been excluded from the results reported so far. However, when analysing the evolution of gene expression in Primates, we found that very close proximity in overlapping genes leads to increase in correlation in change in gene expression (Ghanbarian and Hurst 2015). This is possibly because overlapping genes are more likely to be regulated by the same chromatin level regulatory elements. These regulatory elements would facilitate high correlation in evolution of their gene expression profiles. However, Busby et al (2011) included very few overlapping genes, less than 20. So our effort to investigate whether overlapping genes in more compact genomes, like yeasts, would also demonstrate higher correlation in their evolution of gene expression in comparison to non-overlapping neighbors was ineffective. The correlations were not statistically significant across any of the sets but given the limited sample size we prefer not to make a strong conclusion (Table 4 and table S4 in supplementary material).

Discussion

The phenomena of correlated change in expression of the neighboring genes was dubbed “piggy-backing” when it was first established in Primates. Here I have shown evidence for the piggy-backing in yeast, an organism with a compact genome. The correlation between the change in gene expression of the neighboring genes is shown to be significant and in positive direction regardless of the phylogeny assumption; whether genic or intergenic substitution rate is used to infer phylogenetic relationship between yeast species in this study. Moreover, the strategy applied to treat the replicate information and yeast sensitivity to the environmental changes makes no qualitative difference in the direction or significance of this correlation. Only a minor disparity was observed in p-values across the datasets generated by processing each replicate separately or employing their average read count in our study.

While there were not enough overlapping genes in the homologous gene set to evaluate the strength of correlation at very close proximity, it was still possible to investigate the boundary of this correlation. Calculating Spearman correlation of Z score of two genes at x-min distance away from each other revealed the genes significantly correlate the change in their expression at 1Kb, however, the correlation falls into insignificant p-values very quickly at only 2Kb. Henceforth, the boundary of co-evolving gene clusters in not only slightly smaller than suggested by ripple effect in yeast but more importantly it is vastly smaller than the boundary observed in Primates. This could be due to shorter intergenic regions in yeast in comparison to large ones in Primates. Short intergenic regions increase the possibility of transcriptional interference and also decrease the number of regulatory elements necessary to create a more comprehensive expression profile similar to the one observed in Primates and multicellular organisms. Investigating the boundary of co-evolving gene expression clusters in other organisms with an intermediate average length of intergenic DNA would allow one to investigate if boundary of piggybacking can be predicted solely by the length of intergenic DNA.

Although the choice of genic or intergenic substitution rates to infer the phylogenetic relationship does not make a quantitative difference in the correlation of change in gene expression in neighboring genes, it makes a difference in the median of Z score. Using genic phylogenetic tree in two third of the datasets resulted in Z scores with median closer to zero. In other words, use of genic tree leads to Z scores more agreeable to the general assumption of constant transcriptome size since the common ancestor. However, more comparative transcriptome replicates should be employed to make a statistically valid conclusion. It needs to be mentioned that the evolutionary distance between the yeast species included in this study is relatively short. This might not be consistent if the change in expression of genes were to be estimated from a more distant ancestor.

References

Busby MA, Gray JM, Costa AM, Stewart C, Stromberg MP, Barnett D, Chuang JH, Springer M, Marth GT (2011) Expression divergence measured by transcriptome sequencing of four yeast species. BMC Genomics 12:635

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus M, van Asperen R, Boon K, Voute P, Heisterkamp S, van Kampen A, Versteeg R (2001) The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291:1289

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40:D700

Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genetics 26:183

Ebisuya M, Yamamoto T, Nakajima M, Nishida E (2008) Ripples from neighboring transcription. Nature Cell Biology 10:1106

Ghanbarian AT, Hurst LD (2015) Neighboring Genes Show Correlated Evolution in Gene Expression. Mol Biol Evol

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241

Kristiansson E, Thorsen M, Tamas MJ, Nerman O (2009) Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements. Mol Biol Evol 26:1299

Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Research 13:238

Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91:243

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Systematic biology 53:673

Pal C, Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nature Genetics 33:392

Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W, Sperling S (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics 89:580

Rutschmann F (2006) Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. Diversity and Distributions 12:35

Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of arabidopsis thaliana. Genome Research 14:1060

Woo YH, Li W-H (2011) Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. Proceedings of the National Academy of Sciences 108:3306

Zeitlinger J, Stark A (2010) Developmental gene regulation in the era of genomics. Dev Biol 339:230

Table 1. Spearman correlation between Z score of the focal gene and its closest non-overlapping downstream neighborwhich is closer than 3Kb away. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2214	**0.00113**	0.06915	**2.51E-05**	0.08943
Rep1	2214	**0.00016**	0.08015	**0.00062**	0.07268
Rep2	2214	0.03589	0.04459	0.02307	0.04829

Table 2. Spearman correlation between Z score of the focal gene and mean Z score of its up and downstream neighbors, when at least one neighborin closer than 3Kb. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2751	**7.70E-07**	0.09443	**5.13E-10**	0.11859
Rep1	2750	**1.97E-08**	0.10722	**1.30E-07**	0.10084
Rep2	2750	**0.00037**	0.06814	**0.00018**	0.07162

Table 3. Spearman correlation between Z score of the focal gene and mean Z score of all genes in ±3Kb neighborhood, when the focal gene has at least one neighborin ±3Kb. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2751	**0.00017**	0.07168	**2.66E-06**	0.08938
Rep1	2750	**0.00014**	0.07271	**0.00354**	0.05561
Rep2	2750	**0.00026**	0.06951	**0.00022**	0.07039

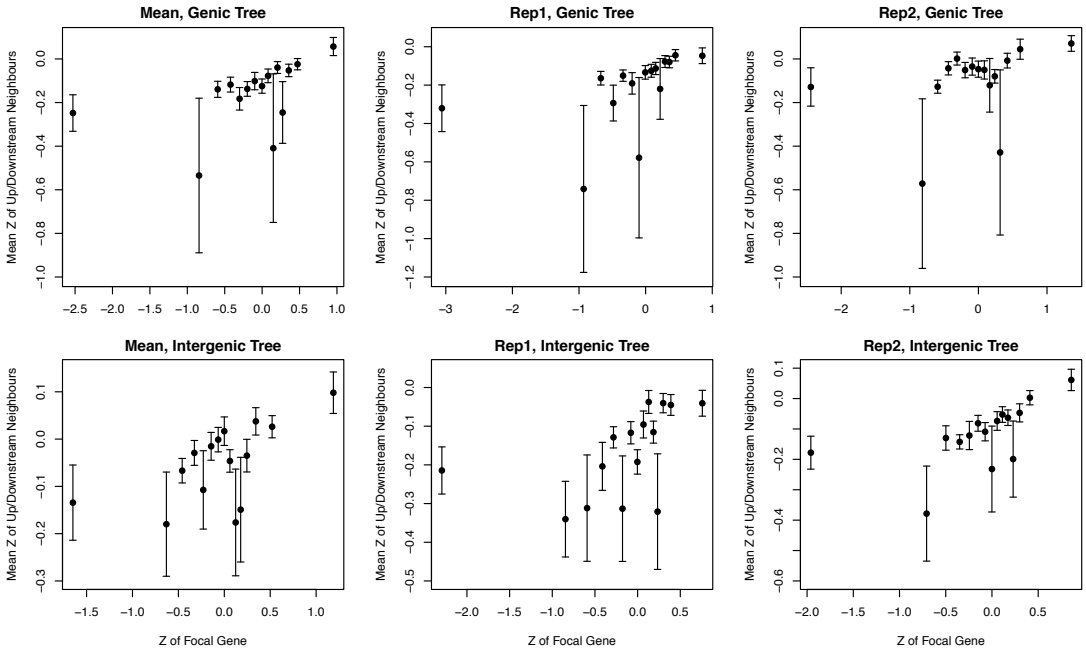
Table 4. Spearman correlation between Z score of the focal gene and Z score of its overlapping downstream neighbor. None of the correlations is statistically significant due to low number of overlapping genes in our study.

Mean/Rep	#overlapping Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	19	0.30940	0.24561	0.61043	0.12456
Rep1	18	0.49789	0.17028	0.38844	0.21569
Rep2	19	0.57028	0.13860	0.52181	0.15614

Table 5. Correlation in change in gene expression only extends to 1Kb. The correlation between Z score of the focal gene and its nearest downstream gene on the same chromosome a minimum of 1K, 2K and 3K base pairs away is shown across the datasets generated by using genic and intergenic phylogenetic trees and different strategies to treat replicates. Results significant after Bonferroni testing are highlighted in bold.

Mean-Rep/Tree	Correlation to 1Kb		Correlation to 2Kb		Correlation to 3Kb	
	P-value	*Rho*	*P-value*	*Rho*	*P-value*	*Rho*
Mean/Genic	**0.00033**	0.07372	0.32129	0.02701	0.02448	0.05953
Mean/Intergenic	**2.20E-05**	0.08706	0.21298	0.03392	0.02496	0.05933
Rep1/Genic	**0.00112**	0.06690	0.16517	0.03780	0.05061	0.05176
Rep1/Intergenic	**0.00399**	0.05914	0.44299	0.02090	0.46728	0.01926
Rep2/Genic	**0.00206**	0.06328	0.22554	0.03302	0.01302	0.06570
Rep2/Intergenic	**0.00203**	0.06336	0.60948	0.01392	0.01356	0.06532

Fig 1. Relationship between Z of a focal gene and mean Z of the its nearest up/downstream neighbors across two phylogenetic trees and mean or separate replicates read counts. In this instance we consider all genes are nearest neighbors if the distance between the gene bodies of the focal and at least one of its immediate neighboring genes is less than 3kb. Data is split into 15 equal sized bins defined after rank ordering with respect to Z score of the focal gene. The value on the X axis represents the mean Z of the genes in that bin. The value of the Y axis indicates the mean (+/-sem) for the relevant flanking genes.



Supplementary material

Supplementary Tables

Table S1. Spearman correlation between fold change of the focal gene and fold change of its closest non-overlapping downstream neighborwhich is closer than 3Kb away. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2214	**0.00567**	0.05877	**0.00014**	0.08071
Rep1	2214	**0.00082**	0.07107	**0.00156**	0.06719
Rep2	2214	0.05760	0.04036	0.03768	0.04417

Table S2. Spearman correlation between fold change of the focal gene and mean fold change of its up and downstream neighbors, when at least one neighborin closer than 3Kb. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2751	**2.06E-05**	0.08140	**5.81E-08**	0.10359
Rep1	2750	**1.07E-05**	0.08417	**1.55E-05**	0.08262
Rep2	2750	**0.00071**	0.06478	**0.00042**	0.06742

Table S3. Spearman correlation between fold change of the focal gene and mean fold change of all genes in ±3Kb neighborhood, when the focal gene has at least one neighborin ±3Kb. Results significant after Bonferroni testing are highlighted in bold.

Mean/Rep	#Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	2751	**0.00336**	0.05590	**0.00037**	0.06782
Rep1	2750	**0.00249**	0.05765	**0.00591**	0.05248
Rep2	2750	**0.00069**	0.06468	**0.00153**	0.06041

Table S4. Spearman correlation between fold change of the focal gene and fold change of its overlapping downstream neighbor

Mean/Rep	#overlapping Genes	Genic Tree *P-value*	Genic Tree *Rho*	Intergenic Tree *P-value*	Intergenic Tree *Rho*
Mean	19	0.27518	0.26316	0.30587	0.24737
Rep1	18	0.17642	0.33333	0.14988	0.35397
Rep2	19	0.61553	0.12281	0.55553	0.14386

Supplementary Figures

Figure S1. Distribution of Z scores calculated based on mean of read counts across two replicates for two phylogenetic trees shown by the percentage of BayesTraits' estimates used in calculation of Z score, all or last 50%, 20% and 10%. Median, maximum and minimum Z score are also shown. First row shows Z scores calculated by using genic phylogenetic tree accompanied by the corresponding intergenic one on the second row.

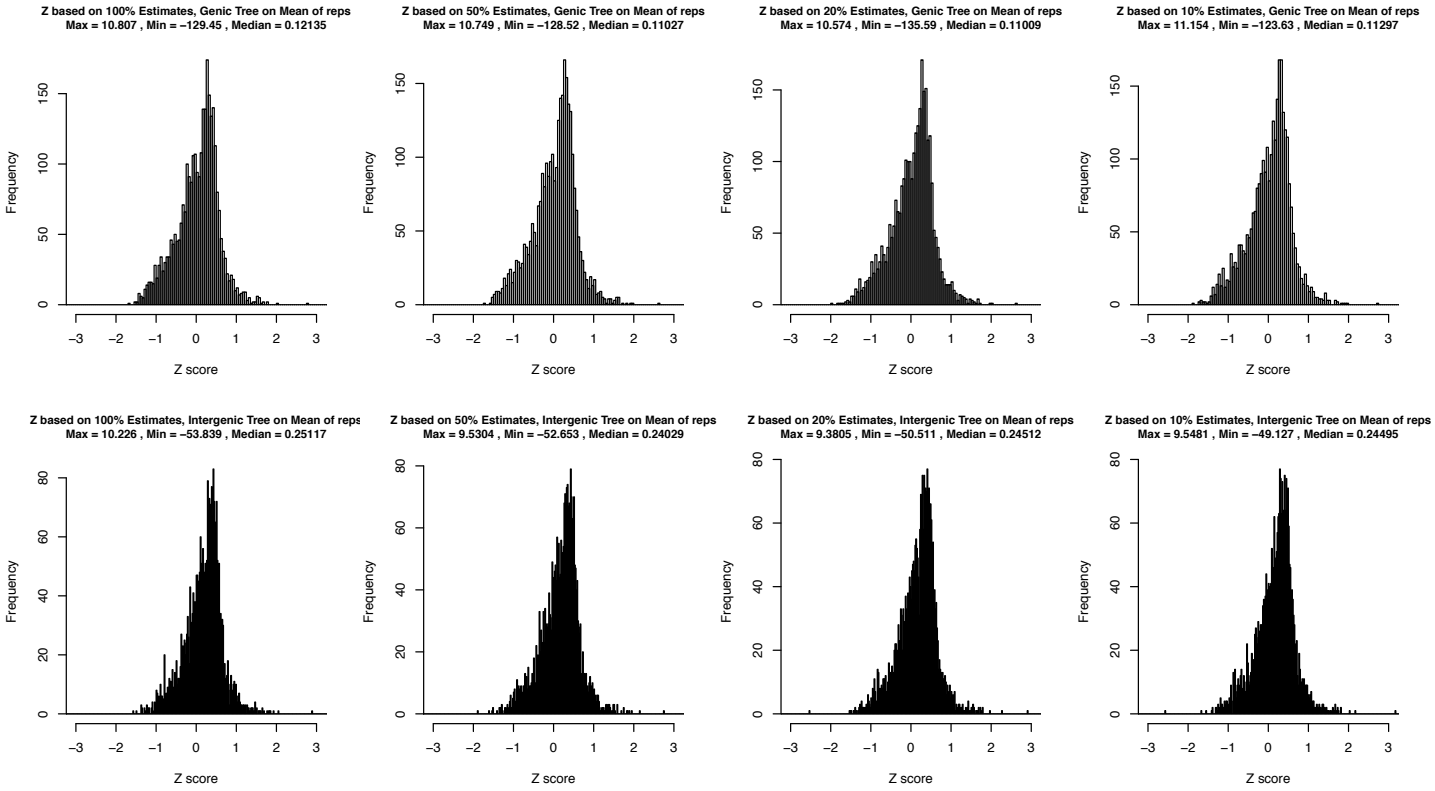
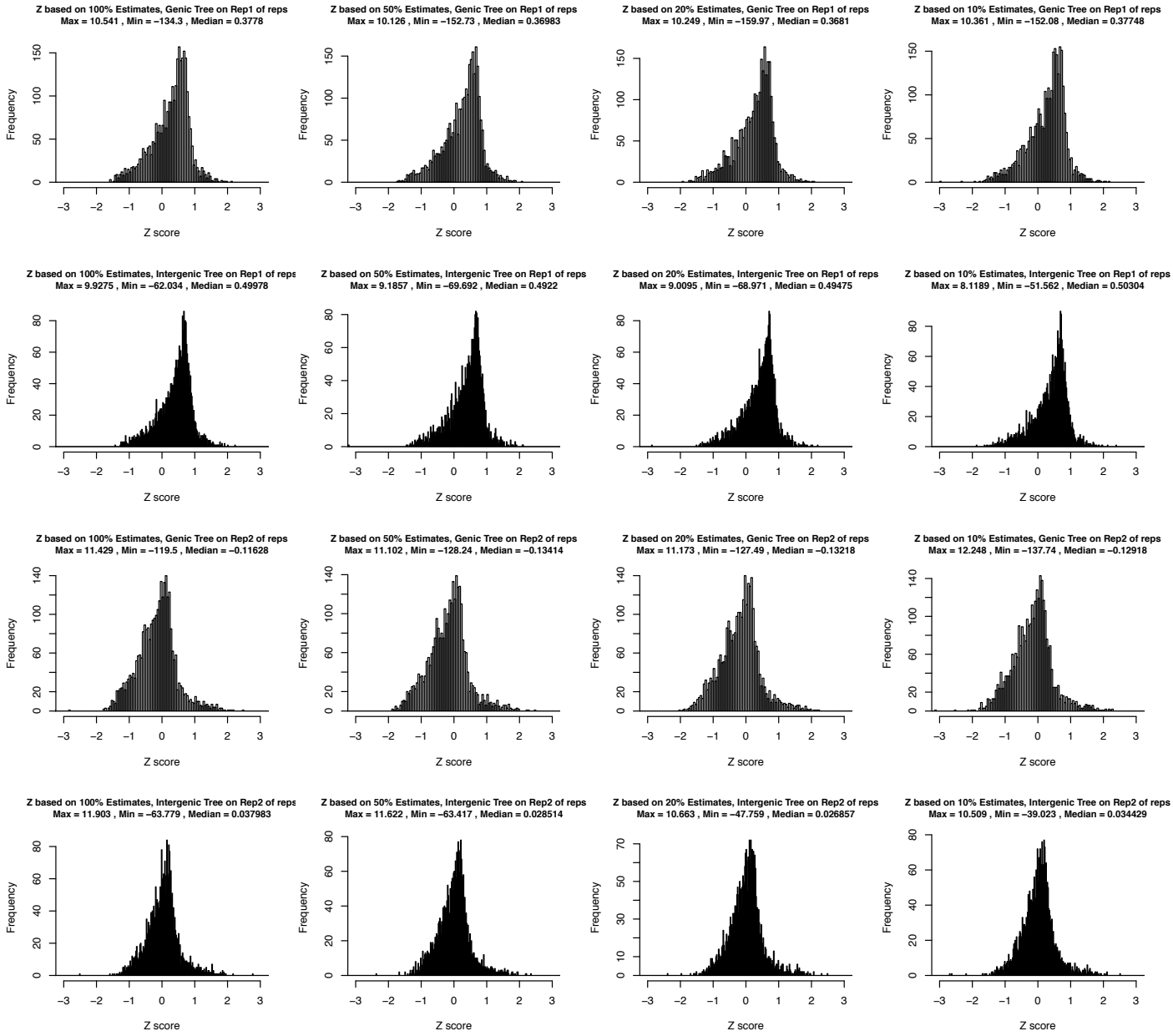


Figure S2. Distribution of Z score calculated based on separate read counts per two replicates for two phylogenetic trees shown by the percentage of BayesTraits' estimates used in calculation of Z score, all or last 50%, 20% and 10%. Median, maximum and minimum Z score are also shown. First and second rows show Rep1 and third and forth rows show Rep2. On even rows Z scores are calculated by using genic phylogenetic tree and corresponding intergenic ones are shown on odd rows below.



Chapter 4. Double agents: How lincRNAs regulate expression of their neighbors

Contribution

Having shown evolution of gene expression in the coding genes to be linked to their neighbours in Primates and Yeasts in previous two chapters, in this chapter I delve into the world of non-coding genes to discuss evolution of lincRNAs. The paper, for which I was co-first author with Andreas Schuler, is published in Molecular Biology and Evolution. This paper is presented as chapter 4. Here I clarify my contributions in this work, not the least to fulfil the requirement of the thesis specifications required by the University of Bath.

Andreas has done all the analyses regarding to the ESE usage and within gene variations in rates of evolution and RNA stability. Laurence Hurst did the nonsense-mediated decay hypothesis. I have done all the analyses to find evidence for the process hypothesis. These include analysing the chromatin remodellers when I found evidence for intron rich active lincRNAs being enriched in CHD1. I have also proceeded with the chromatin state analysis where I found intron density of lincRNA not only correlates with local DHS density but also possibly regulates the expression of neighbours. A few of the epigenetics analyses I have conducted helped us to formulate the process scenario better but were removed from the publication to make the paper as concise as possible. Hence my contributions were instrumental in explaining weak purifying selection on ESE motifs in lincRNA through the process hypothesis versus the product hypothesis.

As you will read in more detail in the paper below, the process hypothesis is an ingenious idea to draw attention to the possibility of the process of splicing to recruit splice-coupled chromatin modifier, CHD1, which would in turn increase the possibility of the neighbouring genes to be transcribed. If one is to see lincRNAs through the glasses of the process hypothesis, one might appreciate how cleverly they function as a double agent; appearing to be only weakly conserved on their ESE motifs but in reality regulating the neighbouring genes through engaging with an splice-coupled chromatin modifier, CHD1.

Purifying Selection on Splice-Related Motifs, Not Expression Level nor RNA Folding, Explains Nearly All Constraint on Human lincRNAs

Andreas Schüler,[†,1] Avazeh T. Ghanbarian,[†,2] and Laurence D. Hurst[*,2]
[1]Institute for Evolution and Biodiversity, University of Münster, Münster, Germany
[2]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom
[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: l.d.hurst@bath.ac.uk.
Associate editor: Eduardo Rocha

Abstract

There are two strong and equally important predictors of rates of human protein evolution: The amount the gene is expressed and the proportion of exonic sequence devoted to control splicing, mediated largely by selection on exonic splice enhancer (ESE) motifs. Is the same true for noncoding RNAs, known to be under very weak purifying selection? Prior evidence suggests that selection at splice sites in long intergenic noncoding RNAs (lincRNAs) is important. We now report multiple lines of evidence indicating that the great majority of purifying selection operating on lincRNAs in humans is splice related. Splice-related parameters explain much of the between-gene variation in evolutionary rate in humans. Expression rate is not a relevant predictor, although expression breadth is weakly so. In contrast to protein-coding RNAs, we observe no relationship between evolutionary rate and lincRNA stability. As in protein-coding genes, ESEs are especially abundant near splice junctions and evolve slower than non-ESE sequence equidistant from boundaries. Nearly all constraint in lincRNAs is at exon ends (N.B. the same is not witnessed in *Drosophila*). Although we cannot definitely answer the question as to why splice-related selection is so important, we find no evidence that splicing might enable the nonsense-mediated decay pathway to capture transcripts incorrectly processed by ribosomes. We find evidence consistent with the notion that splicing modifies the underlying chromatin through recruitment of splice-coupled chromatin modifiers, such as CHD1, which in turn might modulate neighbor gene activity. We conclude that most selection on human lincRNAs is splice mediated and suggest that the possibility of splice–chromatin coupling is worthy of further scrutiny.

Key words: **ncRNA, rate of evolution, splicing.**

Article

Introduction

Understanding how genes evolve and where purifying selection is acting to maintain the status quo can, in principle, be highly informative of the function of a gene and the reasons that mutations might be deleterious and potentially causative of disease. In the simplest instance, for example, selection to preserve functional protein motifs is commonly taken to imply a function for that motif and possible pathogenic consequences for mutations that disrupt the motif. On a broad scale, we can approach these issues by asking where in genes we see purifying selection and what determines the variation between genes in their rate of evolution. Although the determinants of the rate of protein evolution are much studied (Pal et al. 2006; Zeldovich and Shakhnovich 2008), much less well understood are the determinants of the evolutionary rate of noncoding RNAs (ncRNAs). The exons of human ncRNAs are typically poorly conserved compared with protein-coding genes (Marques and Ponting 2009) and on average evolve a little slower than their flanking introns (Hurst and Smith 1999; Pang et al. 2006), suggesting weak purifying selection. The causes of this are unclear (Pang et al. 2006). The relatively rapid evolution need not imply an absence of function, as even highly functional ncRNAs, such as *Xist*, contain only a few conserved stretches (Pang et al. 2006). The determinants of between-gene variation in the rate of evolution of ncRNAs are only beginning to be explored (Managadze et al. 2011). Here then we ask about where in ncRNAs purifying selection operates and what predicts rates of evolution of ncRNAs.

Understanding the evolution of ncRNA can, conversely, potentially shed important light on the mode of selection on protein-coding genes. For example, it has recently been suggested that selection on RNA stability is an important determinant of rate of protein evolution (Park et al. 2013). It is, however, unknown whether this selection is particular to RNAs that are translated or to all RNA species. In principle, one can imagine models for either possibility. For example, RNA stability selection may be important in altering translational dynamics if RNA structure modulates ribosomal speed. Conversely, the selection may simply be to enable RNA to persist in a stable configuration, in which case ncRNA might be under similar selection. For proteins there is at least one universal predictor of between-gene variation in rate of

evolution, namely the more a gene is expressed the lower its evolutionary rate (Pal et al. 2001; Drummond et al. 2006). One hypothesis to explain this concerns selection on protein-folding accuracy (Drummond and Wilke 2008; Yang et al. 2010). If the correlation between protein rate of expression and rate of evolution is mediated exclusively by selection on protein folding, then we expect no such correlation in ncRNAs. One prior claim (Managadze et al. 2011) identified slower evolution in more highly expressed ncRNA and found a coupling between long intergenic ncRNA (lincRNA) stability and evolutionary rate. They concluded there to be a universal (all transcript types) correlation between expression level and evolutionary rate. This they took to suggest a possible universal selection on folding, be it RNA or protein level. Given the importance of such a result, we now return to this issue.

In mammals, splice-related constraints are of an approximately equal magnitude to the expression-related parameters as a predictor of rate of protein evolution (Parmley et al. 2007). Although splice sites are necessary for exon–intron junction recognition, they carry only some of the information required for accurate splicing of protein-coding genes (Lim and Burge 2001). Exonic splice enhancers (ESEs) are also necessary to maintain proper splicing. ESE motifs are purine-rich hexamers that bind serine arginine-rich (SR) proteins to aid exonic splice site recognition (Blencowe 2000; Cartegni et al. 2002). They mostly operate close to (within 70 bp) exon–intron junctions (Fairbrother, Holste, et al. 2004) in a quantitative fashion, such that the higher the density of ESEs the higher the splice rate (Graveley 2000; Fairbrother et al. 2002; Fairbrother, Holste, et al. 2004; Fairbrother, Yeo, et al. 2004; Ke et al. 2011). On average 30–40% of bases at the flanks of protein-coding exons feature in at least one experimentally confirmed motif, this proportion being higher for exons flanked by larger introns (Dewey et al. 2006) where exon definition is especially difficult.

Owing to their abundance, importance, and skewed nucleotide content, ESEs leave strong and easily identified footprints in the molecular evolution of mammalian protein-coding genes (Cáceres and Hurst 2014). ESE motifs evolve at considerably lower rates than non-ESE sites, at both the synonymous (Carlini and Genut 2006; Parmley et al. 2006) and nonsynonymous levels (Parmley et al. 2007). The abundance of ESEs near exon junctions skews amino acid content and codon usage patterns (Parmley and Hurst 2007; Parmley et al. 2007), with the majority of amino acids and codons showing avoidance or preference near boundaries, these trends being well predicted by ESE nucleotide content. As "boundary" regions are large with respect to the average size of an exon, the biology of ESEs is one of the major influences on human protein-coding genes.

ncRNAs frequently contain conserved promoter regions and splice sites and also show a reduced rate of insertions and deletions (Ponjavic et al. 2007), indicative of selection for splicing and transcription. Indeed, conserved splice sites have been employed to identify noncoding transcripts (Rose et al. 2011) and splice sites in ncRNA often show considerable degrees of conservation (Nitsche et al. 2014; Washietl et al. 2014). It is, however, unknown whether ESEs are involved in splice regulation and, assuming that they are, whether they contribute to purifying selection operating on sequence. Here then we employ a robust and appropriate high-quality data set of human ncRNAs (Cabili et al. 2011), wherein we can both have a good measure of confidence that the ncRNAs are not protein coding, that the ncRNA are real (by them being identified more than once), and that, being intergenic, there are minimal issues with overlapping transcripts. Of these data, we ask 1) whether ncRNAs show evidence of splice-related constraint with reduced rates of evolution at exonic ends, especially in residues associated with exonic splice enhancer motifs; 2) if so, what proportion of the reduced rate of evolution of ncRNA exons, when compared with flanking introns, can be explained as owing to splice-related selection; and 3) how important is splice-related selection in explaining between-gene variation in rate of evolution of ncRNAs compared with other possible predictors. We report that the great majority of selection on ncRNAs is splice related, purifying selection being dominantly on exon ends with ESE motifs especially slow evolving. We consider a series of models to explain this unexpected result.

Results

Do lincRNAs employ ESEs? If they do, can we find evidence for splice-related constraints within the exons on lincRNAs? If we can, how important are splice-related constraints, both in explaining any purifying selection operating on lincRNAs exons when compared with their introns and in explaining between-gene variation in rates of exonic evolution? To address the former issues, we start by considering whether exons of lincRNAs use ESEs in the same manner as protein-coding genes and in turn whether they impose comparable degrees of constraint.

ESE Usage and within-Gene Variation in Rate of Evolution

ESE Usage at lincRNA Exonic Flanks Resembles That in Protein-Coding Exons

ESEs are most efficient close to the splice junction (Fairbrother et al. 2002; Fairbrother, Holste, et al. 2004; Fairbrother, Yeo, et al. 2004) and if ESEs are involved in splicing regulation for lincRNAs, putative ESE motifs should be enriched close to splice junctions. To test this hypothesis, we annotated putative ESE motifs in the lincRNA and protein-coding alignments by using the set of experimentally confirmed human ESE-hexamers employed in a previous study (Parmley et al. 2006) as defined by Fairbrother, Yeo, et al. (2004). We temporarily removed gaps from the alignments to scan for matches to the set of known ESE-hexamers. Matching hexamers were masked and gaps were reinserted after the scan. As expected, the density of putative ESE motifs is highest in direct proximity to the splice sites and decreases with distance from the splice site. This trend is observed in both lincRNA and protein-coding exons (fig. 1*a* and *b*).

It has been shown that large introns are correlated with a high density of ESEs in the flanking exons (Dewey et al. 2006;

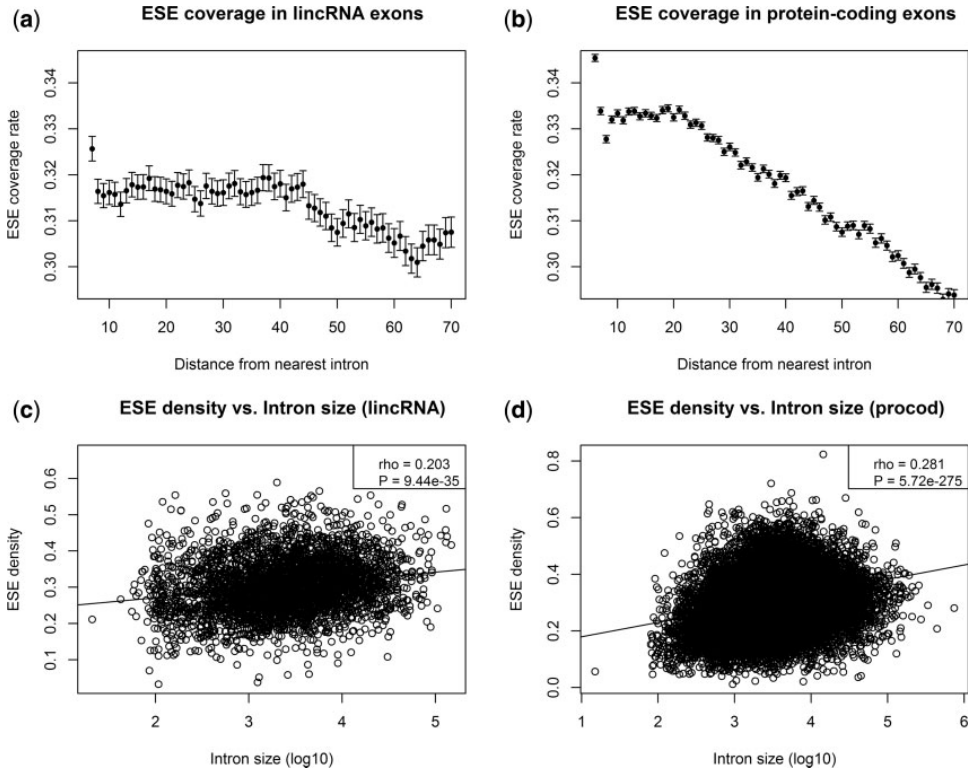


Fig. 1. Relative frequencies of bases ± SEM predicted to be part of an ESE motif as a function of the distance to the nearest intron, starting at a distance of 6 [(*a*) and (*b*)]. The decadic logarithm of the average intron length for lincRNA and protein-coding genes versus the density of ESE motifs on the exon sequences of this gene is shown in (*c*) and (*d*). For (*c*) and (*d*), "density" has been measured as the number of nucleotides that belong to a putative ESE motif divided by the summed length of exons for the respective gene. This figure includes only the conservative lincRNAs. For the complete set, see supplementary figure S1, Supplementary Material online. Error bars for (*a*) and (*b*) = ± SEM.

Cáceres and Hurst 2014). We can reproduce this observation for both the protein-coding genes and the lincRNAs in our data set (fig. 1*c* and *d*). The trend is weaker in lincRNAs compared with protein-coding genes (rho = 0.2 and 0.28, respectively) but both correlations are highly significant (from Spearman: $P < 10^{-16}$). Using all lincRNAs instead of only the conservative subset does not qualitatively change these results (supplementary fig. S1, Supplementary Material online).

Exonic Splice Enhancers Evolve Considerably Slower Than Nonenhancers in lincRNA Evolution

In protein-coding genes, exon residues that specify ESEs evolve slower than non-ESE sequence (Parmley et al. 2006, 2007). Our data replicate this result (median K [ESE] = 0.021, median K [non-ESE] = 0.028, Wilcoxon test: $P < 10^{-16}$). More importantly, we find that substitution rates in ESEs are significantly lower than the ones in non-ESE sites in lincRNAs (median K [ESE] = 0.055, median K [non-ESE] = 0.066, Wilcoxon test: $P < 10^{-16}$).

Given that ESEs function close to exon boundaries, it might in turn be helpful to control for distance from an exon boundary. We thus compared the evolutionary rates between ESE and non-ESE sites as a function of the distance from the nearest splice junction. Conceptually, every exon was split in half and for each alignment site we assigned the base pair-distance to the 5′-splice junction for the first exon-half or to 3′-junction for the second exon-half. We calculated the substitution rates for sites up to 70 bp away from the nearest splice junction and distinguished between ESE and non-ESE for both protein coding RNAs (fig. 2*a*) and lincRNA (fig .2*b*). We again observe a trend of ESE sites evolving slower than non-ESE sites, and also a positional effect with the average substitution rates increasing with the distance from the nearest splice junction. This positional effect is observed both in lincRNA and in protein-coding genes (fig. 2). Using all lincRNAs instead of the conservative subset again does not qualitatively affect the results (supplementary fig. S2, Supplementary Material online).

ESEs are purine-rich and, as ESE density is also decreasing with distance from the nearest splice junction, a biased nucleotide composition might be responsible for the overall increase in evolutionary rates with increasing distance from the splice site. To test this, we concatenated the alignments of all exons and for each distance value, we extracted a random sample from this concatenated alignment with the same sample size and nucleotide composition as the alignment sites for the respective distance from the splice junction. The overall evolutionary rates in the randomized samples are higher compared with the putative ESE motifs and

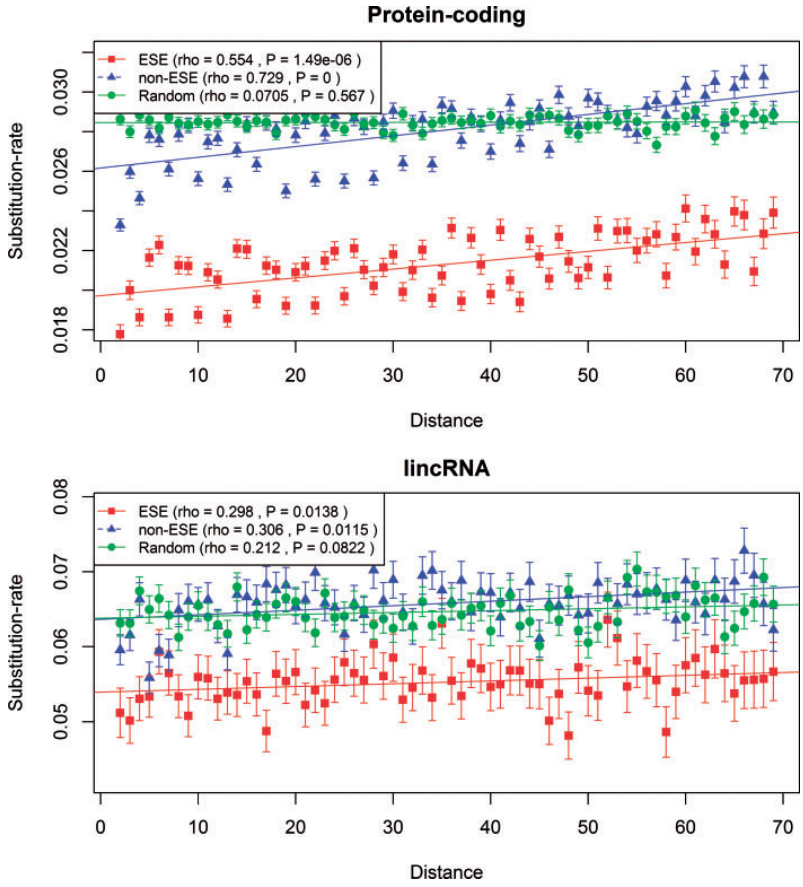


Fig. 2. ESE motifs evolve slower than non-ESE sites. The substitution rates (number of substitutions divided by number of sites) in ESEs and non-ESEs are shown as a function of the distance in base pairs from the nearest splice-junction, for lincRNA (bottom) and protein-coding (top) genes. This figure includes only the conservative lincRNAs. For the complete set, see supplementary figure S2, Supplementary Material online. Bars indicate ± SEM.

comparable to the non-ESE sites, demonstrating that biased nucleotide composition cannot explain the lower evolutionary rates in putative ESE motifs. The magnitude of the difference between the ESE and non-ESE sequences in their rate of evolution in lincRNAs, with ESE evolving around 15% slower than nucleotide controlled null sequence, is the same as witnessed at 4-fold degenerate synonymous sites in protein-coding genes (Cáceres and Hurst 2014).

Although we employ experimentally confirmed ESEs, these are unlikely to correspond to all biologically meaningful ESEs. Nonetheless, selective pressure to maintain the experimentally defined set of ESE-motifs does not seem to be the only cause for this trend because non-ESE sequences show the same trend of increasing substitution rates with increasing distance (rho [ESE] = 0.554 and rho [non-ESE] = 0.729 for protein-coding genes; fig. 2, and for lincRNAs rho [ESE] = 0.298, rho [non-ESE] = 0.306; fig. 2). To consider more generally the role of splice-related constraint, we therefore also compare exon flanks with exon cores and with intronic cores. We presume any differences in rates to be owing to splice-related features.

Weak Constraint on lincRNAs Is Dominantly Owing to Selection in Exonic Flanks in Humans

We employed the ratio of the substitution rate in exons (*Ke*) over the substitution rate in introns (*Ki*) to scan the lincRNA alignments for signatures of purifying selection. For the interpretation of this ratio, introns are used as a proxy for background, possibly neutral, rate (Hoffman and Birney 2007; Resch et al. 2007). A *Ke*/*Ki* ratio < 1 (or < 0 after log-transformation) would thus be indicative of purifying selection. For protein-coding genes, there is evidence of higher selective constraints near exon–intron boundaries, both in the exonic and in the intronic regions flanking the splice junction (Chamary and Hurst 2004; Warnecke et al. 2008) and we therefore analyzed the regions in exon and intron cores and those flanking the splice junction separately (we do not analyze intron flanks). For each aligned gene in the protein-coding and the lincRNA data set, we concatenated 70 bp of exonic sequences flanking the splice junctions and calculated the number of substitutions in the concatenated exon flanks (*Kef*). We defined exon cores as the sequences enclosed by two exon flanks, concatenated them as well, and calculated

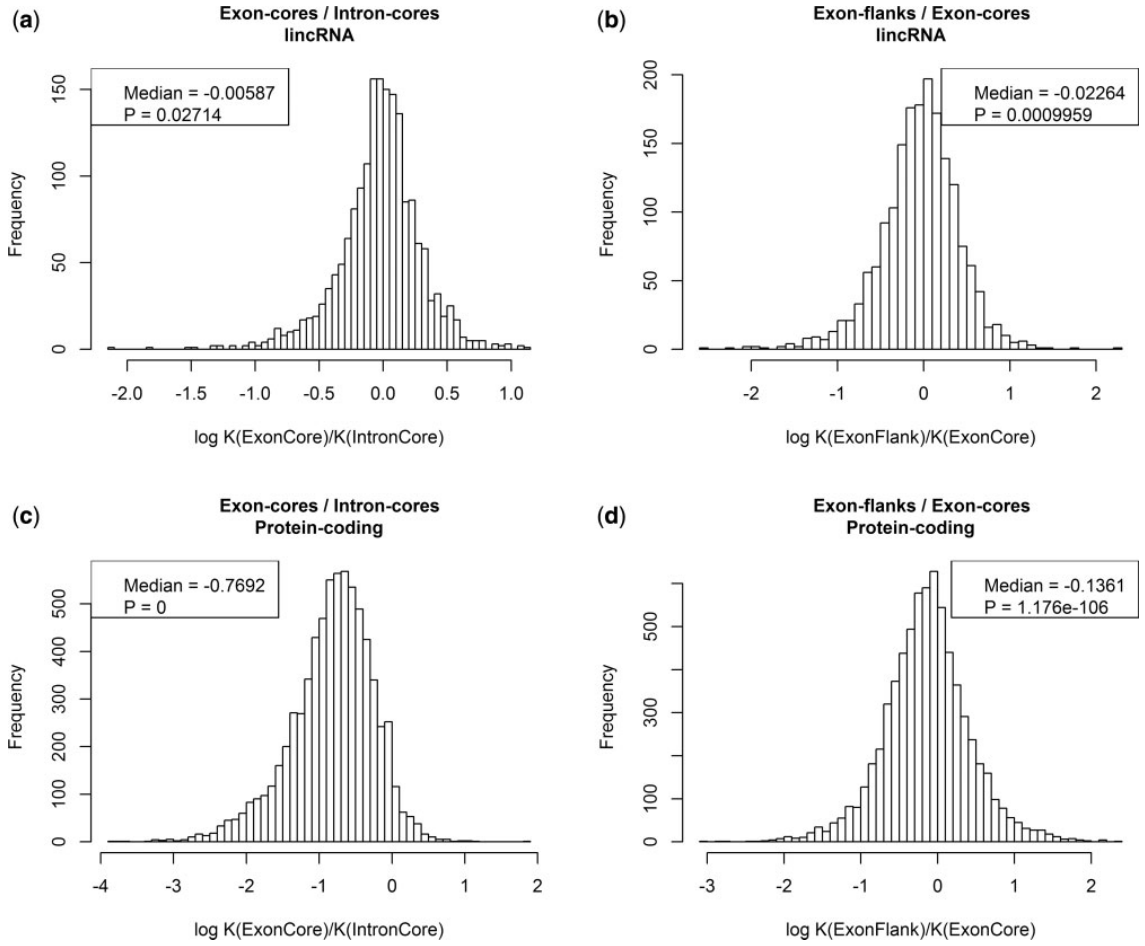


Fig. 3. Exon cores and flanks evolve at different rates. The distributions of *Kec/Kic* and *Kef/Kec* values are shown for protein-coding genes and lincRNAs.

their substitution rates (*Kec*). We compared those rates with the substitution rates in concatenated intron cores (*Kic*), defined as the intronic sequences without the 20 bp flanking the splice junction, as done previously (Warnecke et al. 2008). For this analysis, to reduce the impact of noisy short sequences, we excluded all genes for which the concatenated exon- and intron flanks and intron cores were shorter than 100 bp, which leaves us with 1,810 (53%) lincRNA genes.

The logged (natural logarithm) distributions of *Ke/Ki* ratios for the alignments of lincRNAs and protein-coding genes are shown in figure 3*a* and *c*. The *Kec/Kic* distribution in protein-coding genes is, as expected, consistent with the majority of genes evolving under strong purifying selection (median log *Kec/Kic* = −0.769; fig. 3*c*). For the core exon and core intron regions of lincRNAs, we observe a similar but much weaker trend (median log *Ke/Ki* = −0.005, Wilcoxon test: $P = 0.027$; fig. 3*a*). This trend is still significant at the 0.05 level, but very weak compared with the same effect seen in protein-coding genes, which is consistent with earlier studies that found little evidence for purifying selection acting on lincRNA exons. The *Ke/Ki* ratio is slightly more pronounced when the entire set of lincRNAs instead of only the conservative subset is used (supplementary fig. S3, Supplementary Material online) which is not unexpected because the nonconservative set might contain some genes with protein-coding potential (see Materials and Methods for filtering with respect to coding potential).

We notice that with ESE sequence close to exon flanks evolving 15% slower than neighboring sequence, that this effect alone might explain all or nearly all constraint on ncRNAs. With about 27% of sequence near exon flanks (in the relevant sample), a density of ESE around 30% and 15% slower rate of evolution, assuming exonic non-ESE sequence evolves at about the same rate as introns, we predict that log *Ke/Ki* should be approximately $\log(1 - [0.27 \times 0.3 \times 0.15]) = -0.01$. This is, if anything, greater than the proportional difference that is actually observed, suggesting that selection on ESEs may account for all of the reduced rate of evolution of exons compared with introns.

Assuming this to be the case, we would also expect a low rate of evolution at exon flanks to account for most of the difference between exon and intron. The exon flanks of protein-coding genes have previously been shown to evolve

slower than the exon-core regions owing to the fact that splice-related motifs tend to be in the flanks (Parmley et al. 2007; Warnecke et al. 2008). As expected, we also observe this in our set of protein-coding genes (median log *Kef*/*Kec* = −0.136; fig. 3*d*). Importantly, a similar pattern can be observed for the lincRNA data set, where the exon flanks also evolve slower on average than the exon-core regions (fig. 3*b*). The difference between exon flanks and cores in lincRNAs is again not as pronounced as in the protein-coding data set but still highly significant (median log *Kef*/*Kec* = −0.022, Wilcoxon test: $P = 0.0009$). Overall then, these results support the view that the majority of selective constraint on lincRNAs is at exonic ends with little if any at exon cores. Put differently, if there are conserved motifs in exon cores (as seen if *Xist*), then they are rare. As we excluded genes where the concatenated exon flanks and cores or the concatenated intron cores were shorter than 100 nucleotides, these results do not necessarily reflect trends in short lincRNA genes.

These results replicate in part those of Chodroff et al. (2010) who noted a tendency for exon cores of lincRNAs to evolve faster than flanks. They suggest that this might not be owing to splice-related constraint but instead reflect transposable element insertion in cores. That in the flanks ESEs and non-ESE evolve at different rates (fig. 2) and that the net exonic rate can be predicted from knowing this constraint alone strongly argue in favor of splice-related constraint particular to the exon flanks.

Constraint on lincRNAs in Drosophila Is Stronger Than in Humans but Is Dominantly Not Owing to Selection in Exonic Flanks

We can ask whether the above result might be general. Recently, it has been reported that in *Drosophila* selection on ncRNAs is more intense than seen in humans (Young et al. 2012; Haerty and Ponting 2013). Does this mean that the difference in evolutionary rate between flanks and cores is all the more profound? To address this, we considered rates of evolution of ncRNAs as previously annotated comparing *D. melanogaster* and *D. yakuba*. Confirming the strong constraint in ncRNA exons we find that the log of the ratio of rate of evolution between exon core and intron core is −0.6 (Wilcoxon test: $P = 4.8 \times 10^{-18}$). Unexpectedly, we find that flanks evolve if anything faster than the cores (median ratio of log [flank/core] = 0.25, Wilcoxon tests: $P = 1 \times 10^{-8}$). We conclude that the stronger selection on flanks of ncRNAs is not universal.

Causes of between-Gene Variation in Rates of Evolution

The above analyses indicate that ncRNAs use ESEs much as protein-coding genes do and that splice-related constraints explain the great majority of within-gene purifying selection. These results suggest a further issue. If splicing is so important in explaining intragene variation in rates of evolution, is it also the most important predictor of between-gene variation in rates of evolution? It is not trivially the case that this need be so. For protein-coding genes, a universal and highly significant negative correlation between gene expression and rate of protein evolution has repeatedly been observed (Pal et al. 2001; Drummond and Wilke 2008; Wolf et al. 2010). As this is effectively controlled for by considering intragene analyses, it could be that the causes of intragene variation are dwarfed by a feature, such as expression level, which only becomes important when considering intergene comparisons.

Partial Correlation Analysis Suggests ESE Density Is the Best Predictor of lincRNA Rate of Evolution

The above analysis suggests that ESEs and exon flanks impose major constraint on sequence evolution of lincRNAs. Indeed, the difference in the extent of constraint between the exon core and exon flank suggests that most constraint is splice related. This analysis, while controlled at a pairwise level, does not address the issue of how well splice-related constraints explain between-gene variations in evolutionary rate. How then do splice-related constraints compare with other putative predictors of evolutionary rate and how relatively important is each predictor when allowing for covariance with the others?

To this end, we carried out a partial correlation analysis using the *pcor* R script (Kim and Yi 2006). We considered three expression parameters (maximum expression rate, median expression rate and expression breadth, breadth being the proportion of tissues within which a gene is expressed), two splicing-related parameters (fraction of exon sequence in 70-bp windows flanking splice junctions [frac70] and the fraction of exonic sequence that matches known ESE motifs [ESE density]), folding stability and GC content. Normal and partial correlations are shown in table 1 (see also supplementary table S1, Supplementary Material online). In addition, for comparison, we consider the same parameters in their ability to predict rates of evolution of protein-coding genes.

As regards the rate of evolution of lincRNAs, one parameter stands out. Out of all parameters we considered, the density of ESE motifs in exon sequences is the best predictor for evolutionary rates in lincRNAs, both in normal and in full partial correlation analyses. The other splicing-related parameter, the fraction of sequence within 70 bp of an exon junction (frac70), is however not significantly correlated with evolutionary rates of the lincRNAs.

For the protein-coding genes, the situation is somewhat different. Both the fractions of sequence within 70 bp of an exon boundary and ESE density are correlated with evolutionary rate in the normal correlation analyses, whereas ESE density is no longer significantly correlated in the partial correlation analyses. This seems to be an interaction effect with GC content because ESE density shows a significant partial correlation, comparable to the normal correlation, when GC content is removed from the set of controlled variables. The overall GC content of lincRNA exons is very low compared with the exons of protein-coding genes (median GC content = 0.309 and 0.515, respectively) which might explain why GC content masks the effect of ESE density on evolutionary rate in protein-coding genes but not in lincRNAs.

Table 1. Normal and Partial Correlations with Evolutionary Rate (measured as Tamura–Kumar distance, see Materials and Methods) Using Spearman Correlation (for Pearson correlation, see supplementary table S1, Supplementary Material online).

	Normal	Partial
lincRNA		
Max. expression rate	−0.005	0.032
Med. expression rate	−0.025	−0.035
Exp. breadth	−0.038	*−0.091***
RNA stability	0.048†	0.009
Frac70	*−0.051*†	−0.011
ESE density	*−0.182****	*−0.194****
GC	*−0.058**	*−0.102****
Protein coding		
Max. expression	*−0.203****	−0.019
Med. expression	*−0.339****	*−0.063****
Exp. breadth	*−0.369****	*−0.189****
RNA stability	*0.154****	*0.028**
Frac70	*−0.222****	*−0.101****
ESE density	*−0.29****	0.008
GC	*0.313****	*0.168****

NOTE.—Numbers highlighted in italic are significant after Bonferroni correction (at 5% level, raw $P < 0.00357$ with $N = 14$). Significance codes for P values prior to Bonferroni correction: †$P < 0.01$; *$P < 10^{-3}$; **$P < 10^{-6}$; *** $P < 10^{-9}$.

Expression Level Does Not Predict Evolutionary Rates of lincRNAs

A universal and highly significant negative correlation between gene expression and rate of protein evolution has repeatedly been observed (Pal et al. 2001; Drummond and Wilke 2008; Wolf et al. 2010). It has been proposed that the dominant underlying cause for this correlation is the cost imposed by protein misfolding, which is higher for highly expressed genes (Drummond and Wilke 2008; Yang et al. 2010). It has also been demonstrated however that the cost imposed by protein misfolding cannot be the only cause underlying the observed negative correlation, other possible mechanisms that underlie this correlation include the avoidance of protein misinteractions (Yang et al. 2012) and differential requirements for mRNA folding (Park et al. 2013).

If selective pressure to avoid protein misfolding (or indeed any protein related feature) is the cause of this correlation, a similar trend for lincRNAs should be absent, assuming that they are never translated into a protein product. However, Managadze et al. (2011) have shown that a weak but significant negative correlation between expression level and evolutionary rate is indeed observable in human and mouse lincRNA data sets. We tested the lincRNA data set produced by Cabili et al. to see whether we can reproduce these findings.

For each lincRNA sequence, we plotted the evolutionary distance against the maximal and median expressions and the expression breadth of the respective lincRNA (fig. 4*a*–*c*). Surprisingly, given prior evidence (Managadze et al. 2011), we observe no significant correlation for either maximal expression (rho = −0.005, $P = 0.76$; fig. 4*a*) or median expression (rho = −0.025, $P = 0.12$; fig. 4*b*). This remains true after partial correlation. For expression breadth, we observe a weak negative correlation that is significant at the 0.05 level (rho = −0.038, $P = 0.02$; fig. 4*c*). This result is a little more robust on partial correlation analysis (table 1). Thus lincRNAs that are highly tissue-specific are, on average, less conserved between humans and macaques than those with a larger expression breadth.

The weak correlations between expression parameters and rate of evolution of lincRNAs contrast strikingly with what we find for protein-coding genes. For the protein-coding data set, maximum expression (rho = −0.204, $P < 2.2 \times 10^{-16}$; fig. 4*d*), median expression (rho = −0.34, $P = 2.2 \times 10^{-16}$; fig. 4*e*), and expression breadth (rho = −0.369, $P = 2.2 \times 10^{-16}$; fig. 4*f*) all show a highly significant negative correlation with expression, as expected. These results are diminished on partial correlation analysis but expression level and breadth remain predictors.

lincRNA Folding Stability Does Not Explain Evolutionary Rates

The possibility that RNA structure might be a determinant of protein rate of evolution has recently been proposed (Park et al. 2013). Given this it is relevant to ask whether the same may be perhaps an even more profound predictor for sequences where the RNA alone may be functionally relevant, that is, lincRNAs. We find that folding stability shows a very weak Spearman correlation with evolutionary rates of lincRNAs, but this effect vanishes when the other parameters are controlled for. We conclude that selection on folding strength does not explain the rate of evolution of long ncRNAs (lncRNAs).

Although one may question the ability of any method to correctly infer RNA stability (not least because they fail to acknowledge the presence of the exon-junction complex (EJC) on mature RNA), it is notable that this result contrasts with what is seen for protein-coding genes. In this instance, as previously reported (Park et al. 2013), folding stability shows a strong positive Spearman correlation with evolutionary rate (rho = 0.154, $P < 2.2 \times 10^{-16}$). We note an important word of caution, however, as this correlation is substantially reduced in the partial correlation analysis (partial rho = 0.028, $P < 10^{-3}$). Moreover in a partial Pearson product–moment correlation the sign of the correlation shifts to being negative (supplementary table S1, Supplementary Material online). The strong correlation in the normal Spearman analysis (and that recently reported; Park et al. 2013) seems to be caused by an interaction effect with GC content and ESE density and the removal of those two parameters from the partial correlation analysis yields a partial correlation of comparable magnitude to the normal correlation (partial rho = 0.158, $P < 2.2 \times 10^{-16}$). GC content and folding stability are positively correlated and the negative correlation of folding stability and evolutionary rate thus seems to be caused by stable protein-coding RNAs having a higher GC content than average. Whether the GC content is high to ensure strong folding or whether strong folding is an incidental side consequence of GC content remains to be discovered.

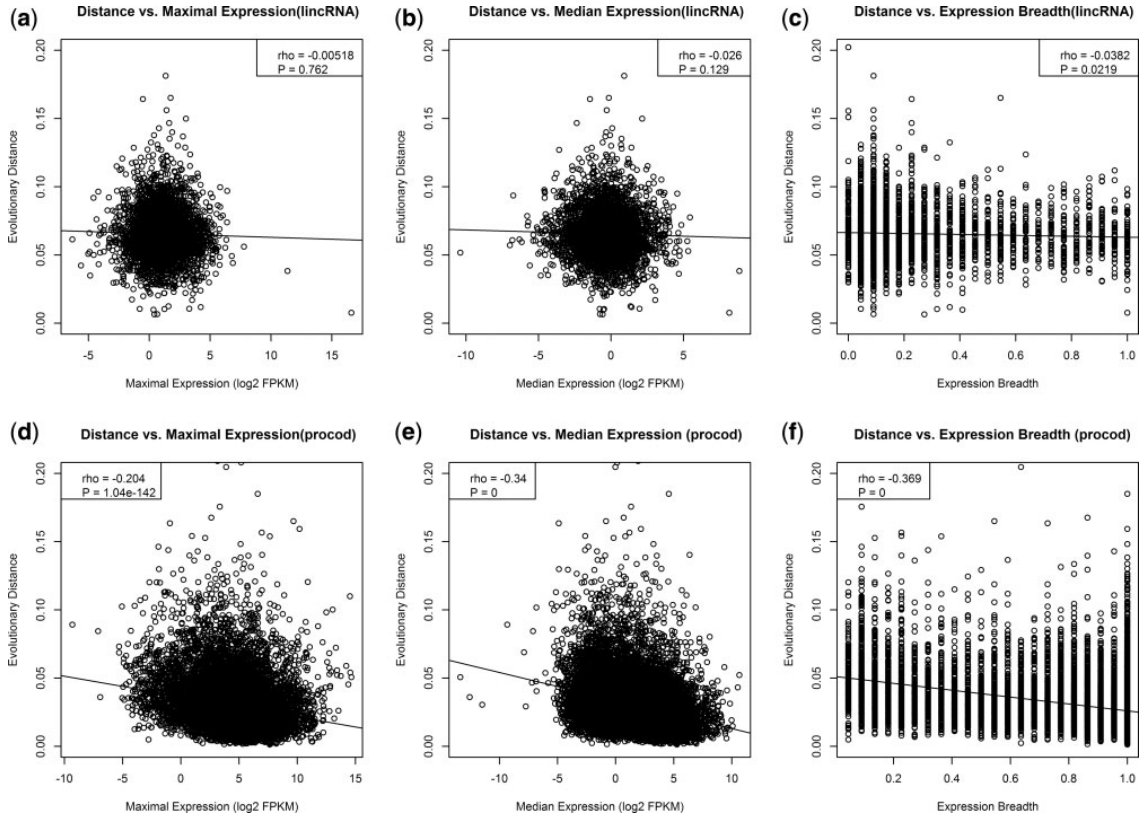


Fig. 4. Correlation of expression parameters with evolutionary rates of lincRNAs and protein-coding genes. The evolutionary distance to the macaque homologue was plotted versus the values of maximum expression (*a*), median expression (*b*), expression breadth (*c*) for each lincRNA, and for protein coding genes (*d-f*).

Complex Correlations between lincRNA Folding Stability and Expression Parameters but No Relation with Evolutionary Rates

Although RNA stability does not predict evolutionary rates of evolution, it remains valid to ask whether stability might correlate with expression parameters. It has been proposed that many, if not most, lncRNA transcripts are highly unstable (Houseley and Tollervey 2009). However, genome-wide studies on lncRNA stability have revealed that lncRNA transcripts are not generally unstable, but rather show a wide range of stabilities that is on average lower, but still comparable to that of protein-coding mRNAs (Clark et al. 2012). As the stability of protein-coding RNAs is correlated with expression (Liebhaber 1997; Shabalina et al. 2006), we tested the lincRNA data set for the presence of a similar pattern.

We detected not only a significant positive correlation between folding stability and maximal expression level (rho = 0.105, $P \sim 10^{-9}$; fig. 5*a*) but also a highly significant negative correlation between folding stability and median expression (rho = −0.19, $P < 10^{-16}$; fig. 5*b*) and a positive correlation with expression breadth (rho = 0.309, $P < 10^{-16}$; fig. 5*c*). To see whether these trends are statistically independent from each other, we constructed a linear regression model to predict RNA stability based on all three expression parameters and conducted an analysis of variance. There is a significant three-way interaction between maximum expression, median expression, and expression breadth (*F*-test: $P \sim 10^{-5}$). These trends suggest that stable lincRNAs are associated with a high maximum expression and are expressed in several tissues, but are highly expressed in few or only one of these tissues and thus also have a low median expression.

For the protein-coding genes in our data set, we observe negative correlations between folding stability and both maximal expression (rho = −0.05, $P < 10^{-9}$; fig. 5*d*) and median expression (rho = −0.07, $P < 10^{-16}$; fig. 5*e*) but no significant correlation with expression breadth (fig. 5*f*). This is perhaps surprising as expression breadth is the strongest predictor of protein evolutionary rates.

Differential Sampling with Respect to Expression Level Explains Differences between Analyses

The above analyses have thrown up two possibly surprising results: Splice-related features are centrally important for predicting between-gene variation in rates of evolution and expression level appears not to be an important predictor. The latter result is doubly surprising given how important expression level is for predicting protein rates of evolution and

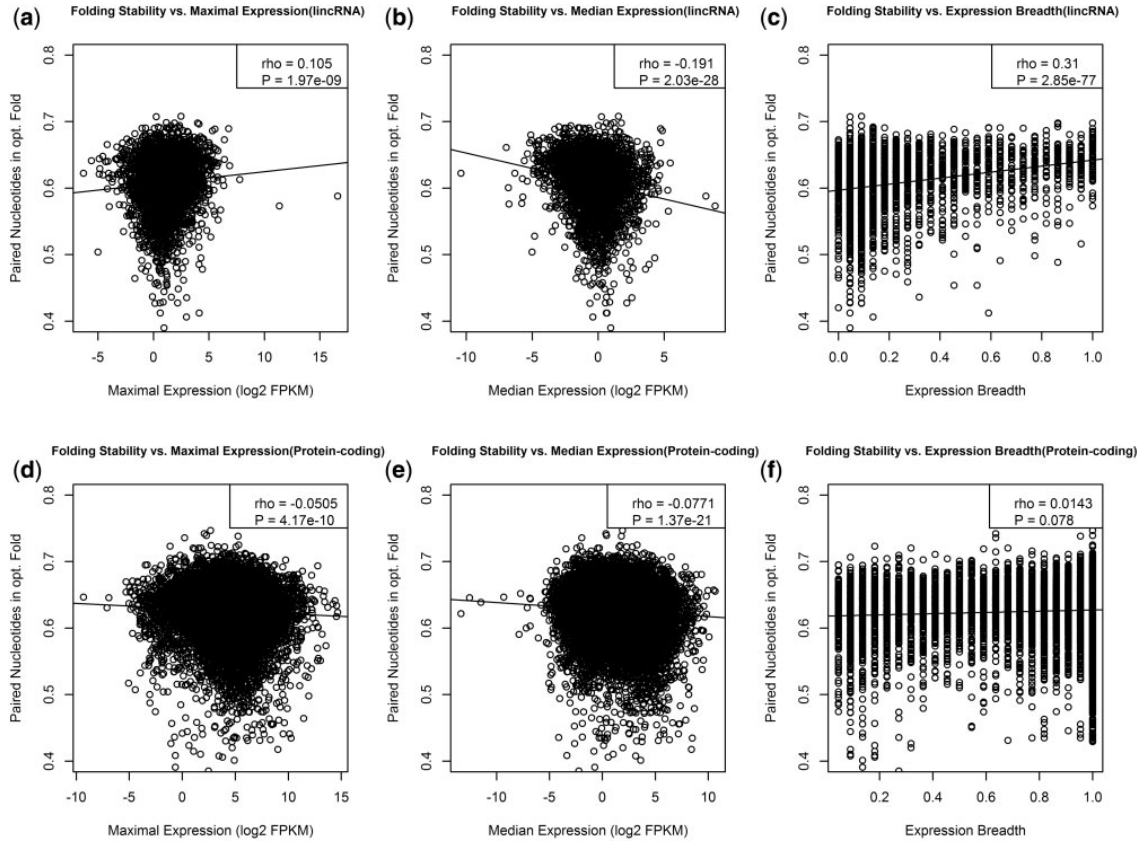


Fig. 5. Folding stability and expression of lincRNAs. The folding stability, assessed as the fraction of paired nucleotides in the minimum energy fold, is plotted against maximum (*a*) and median expression (*b*) and expression breadth (*c*) comparable plots for protein coding genes are shown in *d*, *e*, and *f*.

because Managadze et al. (2011) had previously reported a coupling between evolutionary rate and expression rate. There may be several reasons why this correlation is not apparent in the lincRNA data set produced by Cabili et al. The data set used in the study by Managadze et al. was based on lincRNA data from the NRED database (Dinger et al. 2009) and has a smaller sample size compared with the data set we used in this study. Using a gap threshold of 15% (to match that of Managadze et al., see Materials and Methods) we are left with 3,592 lincRNAs compared with 519 human and 2,013 mouse lincRNAs in the study by Managadze et al. This may itself have some influence, as if we reduce our sample size to 1,500 transcripts we recover a negative correlation at least as extreme as that seen in Managadze et al.'s study 18% of the time.

However, the more important reason for the discrepancy appears to be that the data produced by Cabili et al. is based on deeper transcriptome sequencing that is less biased toward highly expressed lincRNA genes. We tested this hypothesis by analyzing the correlation between expression level and evolutionary rate using only the 50% of lincRNAs that show the highest maximum expression level. For this data set, we do observe a weak but significant negative correlation (supplementary fig. S4*A*–*C*, Supplementary Material online, *P* values for all correlations < 0.05). Indeed, when we repeatedly subsample from only the more highly expressed gene set we recover a negative correlation at least as extreme as that seen in Managadze et al.'s (2011) study 98% of the time. We obtained the lincRNA data set used in the study by Managadze et al. (2011) and can reproduce their results using this data set.

These results suggest that the correlation observed before between expression level and rate of evolution of lincRNAs (Managadze et al. 2011) is dependent on limited sampling. It might be that the more in-depth analysis of Cabili et al. (2011) is more noisy, especially for lowly expressed transcripts, and that this extra noise led to the removal of the correlation. Alternatively, there may not be a monotonic relationship between expression level and rate, in which case sampling only the more highly expressed transcripts could enable detection of a strong trend unique to the highly expressed genes. Alternatively, the prior result may simply be an artifact of limited sampling. As we cannot discriminate between these alternatives, we suggest that the evidence against the misfolding hypothesis on the basis of correlation between expression level on ncRNA evolution (Managadze et al. 2011) be considered provisional. The possibly stronger evidence (although this is relative) appears to derive from a weak correlation between expression breadth and rate of evolution.

Discussion and Further Results

Human lincRNAs have been shown to be almost as poorly conserved as other intergenic sequences and even highly functional lincRNAs such as *Xist* only contain few short stretches that are well conserved (Pang et al. 2006). However, many lincRNA loci have conserved promoter sequences (Carninci et al. 2005) and conserved splice sites (Ponjavic et al. 2007). The 5′- and 3′-splice sites alone are usually not sufficient to maintain proper intron excision (Lim and Burge 2001) without the additional presence of ESE motifs near the splice junction (Wang et al. 2005). In protein-coding genes, purifying selection is acting to maintain these motifs and we thus hypothesized that, given the presence of conserved splice sites, a similar trend might be observable for lincRNAs. Consistent with this hypothesis, we do observe that the exon-flank regions of lincRNAs evolve slower than their exon-core regions. This trend is weaker than the one observed in protein-coding genes, but qualitatively similar and highly statistically significant. We further showed that the putative ESE motifs within exon flanks evolve significantly slower than the sites which do not correspond to known ESEs, indicating that purifying selection to maintain ESE motifs is at least partly responsible for the slower evolution in exon flanks compared with core regions.

One possible explanation for the difference in evolutionary rate between ESE and non-ESE at exon flanks (fig. 2) is that our data set of lincRNAs is contaminated with protein-coding genes and these and these alone employ ESEs (or employ them much more often). Note that if both protein-coding genes and noncoding genes both employ ESEs at similar densities (which we show they do—fig. 1*a* and *b*), then the difference in rate between ESE and non-ESE cannot be explained as a contamination artifact. On a priori grounds, it is indeed hard to see how an SR protein might distinguish an ESE in a protein-coding immature transcript from the same ESE in a lincRNA immature transcript. The artifact explanation we suggest is unparsimonious for numerous reasons. First, all of our results strongly argue against a large contamination issue: The rate of evolution is extremely high on average in our lincRNAs and we do not recover the strongest protein-related correlations, such as with expression level and RNA stability. Perhaps more directly, if we split our internal exons into those with at least one stop codon in every frame (very unlikely to be protein coding) and all others and repeat the analysis of ESE and non-ESE rates, we observe that the two partitions of the data are nearly identical in absolute rates and the difference between ESE and non-ESE (supplementary fig. S5, Supplementary Material online). Indeed, the difference between ESE and non-ESE in the set with stops in all frames is approximately 15% as it is for the data set en mass and for 4-fold degenerate sites in protein-coding exons (Cáceres and Hurst 2014). We conclude that our results are not affected by contamination from protein-coding sequence.

Given the evidence for purifying selection on ESEs we might expect to see some genetic diseases associated with splicing defects in lincRNAs owing to single nucleotide polymorphisms close to but not at the splice junctions. The evidence for selection on ESE refutes the hypothesis that lincRNAs are the all the product of junk transcription. That the majority of the selection on lincRNAs is on splicing presents a new paradox, why it is that selection acts on the splicing process. In principle there might be at least three classes of explanation, which we term the product hypothesis, the error-proofing hypothesis, and the process hypothesis.

Why Is Most Selection on Human lincRNAs Splice Related?

Absence of Constraint in Exon Cores Does Not Refute the Product Hypothesis

The product hypothesis proposes that the product of transcription and splicing is important and the precise exonic structure of the mature ncRNA relevant to this function (as with most protein-coding genes). Our finding that exon cores evolve at rates very similar to those of flanking introns provides little or no support for the idea that functionality of the ncRNA product impacts evidently on sequence conservation. This does not, however, refute the product hypothesis for several reasons. First, some well-described lincRNAs have known functions (e.g., Xist) but apparently little or no sequence conservation (Pang et al. 2006). Further, were our lincRNAs to contain very small subset of sites under strong purifying selection owing to selection on the operation of the RNA, we would almost certainly be unable to detect it with our metrics, the sites being too rare and hence diluted. In addition, conservation of function may be reflected not in conservation of nucleotide sequence but in tolerated indel events.

One way to rationalize the apparent rare selection on nucleotide sequence is that ncRNA might be under selection to enable strong structure, which imposes only weak selection on the primary sequence. This hypothesis would also be consistent with the observations that many lincRNAs with distinct sequences are able to bind the same protein complex (Guttman et al. 2011; Khalil and Rinn 2011) and that the rates of insertions and deletions, which would be much more disruptive to the secondary structure than point mutations, are reduced in lincRNAs (Ponjavic et al. 2007). Were this the case, however, we might expect that lincRNAs that are more stable might evolve slower to preserve that structure. Our data, however, find no support for the view that lincRNA structure is under selection, or at least that any selection on structure is operating uniformly in the same direction (e.g., to always increase stability). This contrasts with the picture for protein-coding genes and with prior claims for ncRNA (Managadze et al. 2011). Selection against indels may well also disrupt splicing, potentially explaining this prior result. In sum, given evidence of function in the absence of sequence conservation, we cannot eliminate the hypothesis that lincRNAs have a direct function, we just find little or no evidence to support it from the mode of sequence evolution in humans, although the fly data are compatible with such a model.

No Strong Evidence for the Nonsense-Mediated Decay Error-Proofing Hypothesis

Another possibility is that the selection on splicing may be part of an error-control mechanism. Splicing commonly results in the mature transcript being bound with the EJC in proximity to the exon–exon junction (Le Hir et al. 2001). The complex is known to mediate the effect of splicing on mRNA expression levels (Wiegand et al. 2003) and so might be directly beneficial were the lincRNA functional (i.e., the product hypothesis above). However, many of the effects of the EJC may well be undesirable for ncRNAs: The EJC acts to promote export from the nucleus, enable polyadenylation, and enhance translation (Le Hir et al. 2001; Wiegand et al. 2003). In addition, however, in mammals the EJC is also necessary for the initiation of NMD (Le Hir et al. 2001; Isken et al. 2008). Might the selection on splicing be to enable the EJC to be attached to initiate NMD should a ribosome inappropriately bind an ncRNA? Binding of ribosomes to ncRNA has been described (Wilson and Masel 2011), but whether this reflects improperly annotated coding genes or accidental ribosome initiation in unclear.

At first sight this is a possibly attractive explanation, not least because it is consistent with the apparent-reduced constraint in exon flanks compared with cores in *Drosophila*, as flies do not employ the EJC to initiate NMD (Brogna and Wen 2009). The hypothesis also fits with the notion that many otherwise paradoxical features of gene and genome evolution are error-correcting or error-proofing (Warnecke and Hurst 2011). However, the hypothesis has at least one major problem. Although a polyA tail is required for NMD activation (Brogna and Wen 2009), many ncRNAs are possibly not polyadenylated. Indeed, the great majority of the transcripts that can be detected uniquely in protocols that do not require polyA tail tagging, compared with methods that require such tagging, are lncRNAs (Cui et al. 2010).

We can in addition ask whether we can find a trace of selection for NMD triggering on the ncRNA sequences. Stop codons less than about 50 bp upstream of the terminal exon–intron junction are thought to be invisible to the activity of NMD (Zhang et al. 1998), what we term the NMD shadow. This provides grounds for potentially instructive tests. If introns are there to trigger NMD, then sequence prior to this 50-bp window might be expected to have a higher frequency of stops (in any frame). To address this, then we considered instances of lncRNAs where the last but one exon was more than 100 bp. We then considered the 50 bp at the 3′-end of this exon and the 50 bp at the 5′-end of the same exon. We then compare stop codon frequency in the 5′- and 3′-end in a paired fashion. This method allows us to control for the amount of sequence analyzed per exon, the proximity to an exon junction (given that these are expected to be purine loaded owing to the presence of ESEs), isochore level nucleotide content, and the possibility that the last but one exon may actually be protein coding. A small and nonsignificant minority (47.5%) of last but one exons have more stops at the 5′-end than the 3′-end (binomial test: $P = 0.31$). On average, each 50-bp exon end has about 0.6–0.7 stop codons in each reading frame. We thus see no evidence that stops are enriched outside of the NMD 50-bp shadow.

One might object that this test fails to recognize the possibility that a stop may have occurred prior to the last but one exon and only one stop is required (per possible frame). To consider this, then we consider the class of ncRNAs with just two exons and consider the sequence −100 to −51 prior to the single exon junction in the first exon and compare this to sequence −50 to the 3′-end of exon (the NMD shadow). As before the stop codon frequency is no different in the two (in 50.03% of cases the first 50 bp has the higher stop codon frequency). Of all two exon ncRNAs 20% have no stop codon outside of the final 50 nucleotides and 55% have fewer than three, meaning that at least one prospective reading frame is NMD unprotected. Were there selection for stops outside of the NMD shadow this should be most apparent in those first exons longer than 50 bp but still relatively short. Of those first exons that have more than 50 bp of sequence but less than 101 bp, 45% have no stop codons outside of the NMD shadow. Randomizing the same 5′-sequence we predict that around 41% would lack a stop codon in any frame by chance alone suggesting, if anything, that the real sequence is slightly diminished for stops. Ninety-one percent of the 5′-sequences have fewer than three stop codons meaning that at least one frame of reading is NMD unprotected. Eighty-six percent of random sequence is expected to have fewer than three, again suggesting no enrichment of stop codons to initiate NMD. We can more generally ask about stop codon density in the 5′-exon of two exon genes. If stops are there to trap ribosomes, then we would expect a higher density in small first exons as these would be under particular pressure to encode them, longer first exons likely having a stop codon by chance. However, stop codon density is unrelated to exon length, with no hint of the expected negative correlation (rho = 0.14, $P = 0.51$). In sum, we find no good evidence that selection is enriching these exons for stop codons to trigger NMD.

The Process Hypothesis: Intron Density Is Associated with Chromatin and Gene Activity

The final possibility is that it is the process of splicing that is important. Implicit in the process argument, and contrary to the product hypothesis, is the notion that after the splicing event the RNA could be destroyed instantaneously with no negative consequence. Although it is unclear why the process might be relevant, we note that recent evidence suggests that the splicing process is somehow coupled with epigenetic marks on the DNA (Adam-Hall and Georgel 2011; Luco et al. 2011). This can mean both that the epigenetic status of the DNA can affect the process of splicing and, more importantly in this context, that the splicing process can modify the underlying DNA (Hnilicova and Stanek 2011). Evidence exists for both directions of interaction (Hnilicova and Stanek 2011). Mechanistically it is unclear how this operates but four chromatin adaptor proteins (including the chromatin remodeler CHD1 [Sims et al. 2007]) are recognized that permit coupling between splicing factors and histone posttranslational modifications (Adam-Hall and Georgel

2011). Similarly, the SWI/SNF chromatin remodeling factors are known to interact with many components of the spliceosome (Adam-Hall and Georgel 2011). We note that a potential role for introns in modulating the chromatin of the underlying gene has relevance for explaining why mammalian transgenes typically require introns for efficient expression. As this effect is mediated, at least in part, by the recruitment of the EJC, rather than the presence of an intronic sequence per se (Wiegand et al. 2003), interaction between the EJC and any of the above splice/chromatin modifiers would provide a mechanistic rationale.

Given that ncRNAs are thought to play a role in chromatin modulation (Mercer and Mattick 2013) (although splicing is not always necessary [Beckedorff et al. 2013]) this suggests a hypothesis that is, to the best of our knowledge, novel. Might the expression and splicing on lincRNAs be a mechanism to alter the epigenetic landscape of the underlying DNA? If it does, then might this simply be a mechanism to control expression of the lincRNA or might it have knock on consequences for flanking genes? In yeast and mammals, for example, the expression of one gene causes a time-lagged ripple of gene activation of neighbors associated with spreading altered chromatin (Ebisuya et al. 2008). NcRNAs are well known to have *cis*-effects on genes in their vicinity (Pauler et al. 2012), so this possibility is not without precedent. Here then we ask two questions. First, is there evidence consistent with the possibility that lincRNAs affect, through splicing, underlying chromatin? Second, is there evidence consistent with the possibility that the activity and splicing of lincRNAs impact on the chromatin and expression of neighbors? Note that expression alone might have effects on chromatin even in the absence of splicing (as in yeast), such a model can also apply to expression of lincRNAs without introns. Our hypothesis is that splicing can bolster such an effect.

This hypothesis, like the NMD hypothesis, can potentially explain why constraint is not so evident in exonic flanks in *Drosophila*. Although chromatin modifiers can also be splice modifiers in *Drosophila* (Hnilicova and Stanek 2011), in flies ESE density is thought to be relatively low as introns are short and exons typically have strong splice sites (Warnecke et al. 2008). Humans, in contrast, have much longer introns and quite often weaker splice sites, both of which predict higher ESE density (Dewey et al. 2006). Thus information in the flanks is thought to be of lesser importance in *Drosophila* than it is in humans for the specification of splice location and, while detectable, the impact on codon usage and rates of evolution at exonic flanks of selection for ESEs is marginal in protein-coding genes (Warnecke and Hurst 2007).

Intron-Rich Active lincRNAs Are Enriched in CHD1. To test the first hypothesis, we assessed 1) whether actively transcribed lincRNAs are enriched in CHD1-binding sites compared with inactive lincRNAs and 2) whether the density of CHD1 is correlated with the intron density (and hence the amount of splicing per base pair of a gene). This analysis is limited to the cell lines H1-HESC, from human embryonic stem cells, and K562, a leukemia cell line, as those are the only cell lines for which CHD1 modifications are available in the ENCODE data set. lincRNA expression status and expression of neighbors we derive from the same two cell types. Note that the genes considered active or inactive in the two cells in these analyses are specific to each cell and the CHD1 measure is similarly specific to each cell type. Thus, the two cell types are independent tests of the same hypothesis.

We find that active lincRNAs are more dense in CHD1 on the DNA containing the gene compared with the transcriptionally inactive lincRNAs (Mann–Whitney *U* test, two tailed, $P = 3.9 \times 10^{-9}$ in H1 and $P = 4.8 \times 10^{-9}$ in K562). Concerned that this statistic may be misled by large number of sequences with no CHD1 binding, we repeated the analysis using a Monte Carlo simulation (see Materials and Methods) which may be more robust to the data structure. The results remain robust (from simulation: $P << 10^{-4}$).

In addition, we can ask whether the CHD1 density on a ncRNA is predicted by the density of introns. As can be seen (fig. 6), active genes have higher CHD1 density the more introns they have (H1; rho = 0.23, $P < 2.2 \times 10^{-16}$, for K562 rho = 0.16, $P < 2.2 \times 10^{-16}$). For the inactives, the inverse is seen, the effect being greatly owing to the great number of intron-rich genes without any CHD1 (H1 inactive, rho = −0.11, $P < 5.2 \times 10^{-15}$, for K562 rho = −0.19, $P < 2.2 \times 10^{-16}$). This might suggest active purging of CHD1 from inactive genes. Considering CHD1 coverage (i.e., proportion of gene covered by at least one CHD1 span) does not affect conclusions: H1 active, rho = 0.1, $P < 10^{-12}$, K562 active rho = 0.08, $P < 10^{-8}$, inactives: H1 rho = −0.15, $P < 2.2 \times 10^{-16}$, K562 rho = −0.23, $P < 2.2 \times 10^{-16}$. These tests are robust to application of Goodman–Kruskall gamma test, a test more robust to tied values (see fig. 6). In turn we can ask whether CHD1 occupancy correlates with the extent of open chromatin within the genes in question, as assayed by the density of DNAase Hypersensity Sites (DHS). As expected, active genes have higher DHS than inactive ones and the extent of DHS correlates positively with intron density (fig. 7). There is a strong positive correlation between CHD1 occupancy and DHS occupancy in both cell types (table 2).

These findings are consistent with the chromatin modification/splicing hypothesis, in which splicing recruits CHD1 to the underlying sequence which in turn acts to maintain or force opening of chromatin. Moreover, consistent with the notion that splicing enables the focal gene to remain open and active we find that the intron density in both cell lines is higher for active genes than for inactive ones (Mann–Whitney *U* test: H1, $P = 0.0003$, K562 $P = 0.04$).

Intron Density of lincRNA Predicts Local DHS Density and Expression of Neighbors. Although the above evidence is consistent with the hypothesis that splicing of lincRNAs mediates recruitment of CHD1 to the underlying DNA, it provides no evidence that this has consequences for the neighboring genes. It may simply be the case that CHD1 recruitment aids the maintained expression of the focal lincRNA (for whatever reason) or indeed, that CHD1 recruitment is an incidental occurrence, a necessary consequence of splicing. We can then also ask whether active lincRNAs define

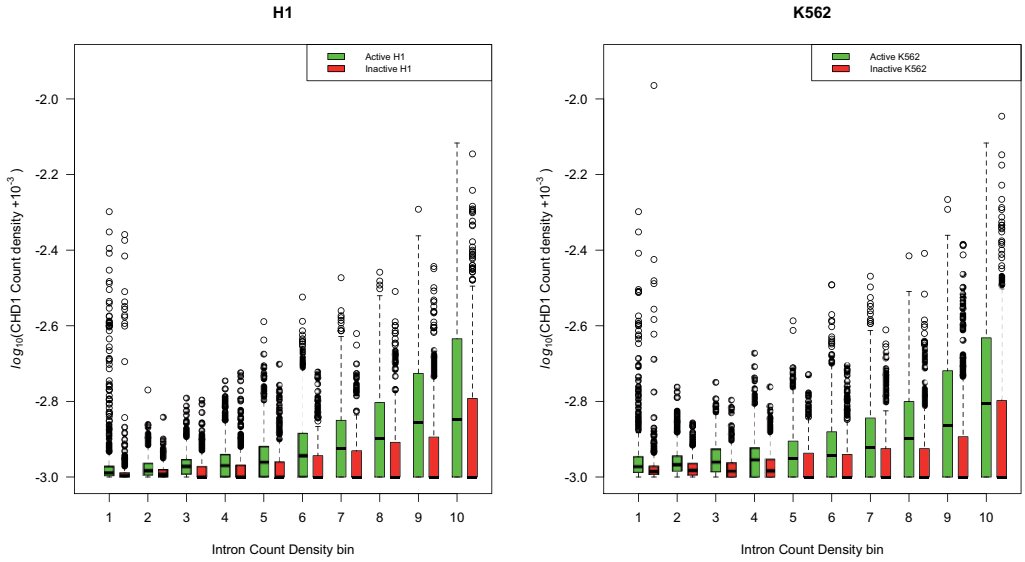


FIG. 6. CHD1 density within lincRNAs is higher in active intron rich genes. Here, for each gene, we consider the number of CHD1 peaks (as specified by ENCODE) per unit base pair of each gene and compare this with the number of introns per unit base pair of gene length (in both cases we employ the length of the unspliced gene). We consider those lincRNAs that are transcriptionally active or inactive in each cell type separately. As can be seen, active genes have higher CHD1 density the more introns they have. For H1 active, rho = 0.23, $P < 2.2 \times 10^{-16}$, for K562 rho = 0.16, $P < 2.2 \times 10^{-16}$. For the inactives, the inverse is seen the effect being greatly owing to the great number of intron rich genes without any CHD1: For H1 inactive, rho = −0.11, $P < 5.2 \times 10^{-15}$, for K562 rho = −0.19, $P < 2.2 \times 10^{-16}$. Concerned that there were many tied values we examined the latter result using the Goodmans Kruskall gamma test, this being more robust to tied values. Results are unaffected (for H1 active, gamma = 0.2048, H1 inactive gamma = −0.0863, K562 active gamma = 0.1353, and K562 inactive gamma = −0.1382; all P's < 0.001 from 1,000 simulations). Note that the genes considered active or inactive in the two cells are specific to each cell and the CHD1 measure is similarly specific to each cell type. Thus, the two cell types are independent tests of the same hypothesis. Considering CHD1 coverage (i.e., proportion of gene covered by at least one CHD1 span) does not affect conclusions: H1 active, rho = 0.1, $P < 10^{-12}$, K562 active rho = 0.08, $P < 10^{-8}$, inactives: H1 rho = −0.15, $P < 2.2 \times 10^{-16}$, K562 rho = −0.23, $P < 2.2 \times 10^{-16}$. Results are again robust to application of Goodmans Kruskal gamma (H1 active gamma = 0.0865, K562 active gamma = 0.0654 and H1 inactive gamma = −0.142 and K562 inactive gamma = −0.1834 and all $P < 0.001$, from 1,000 simulations).

a broader domain of open chromatin and a domain of increased gene activation, as supposed by the ripple hypothesis (Ebisuya et al. 2008). To this end we ask about activity in the domains flanking the focal genes, both in terms of chromatin and gene activity. As in humans the ripple effect is thought to extend approximately 100 kb (Ebisuya et al. 2008), this defines the span that we examine.

To analyze the chromatin state, we examine the density of DHS in spans around the focal active or inactive lincRNAs. We calculated the DHS density in independent 10-kb windows either side of active and inactive lincRNA and then compare the density, at a given distance between the active and inactive ones. As can be seen the DHS density is highest in the immediate vicinity of active loci in both cell types (fig. 8 and supplementary fig. S6, Supplementary Material online). It is striking that active lincRNAs appear to be at the position of maximal chromatin opening. This is what would be expected were activity of the lincRNA causing a rippling/spreading opening of chromatin.

If the chromatin splice model is correct and enables spreading of open chromatin, then we might expect that the local DHS density is correlated both with intronic density and with activity of the focal gene. To examine this, we consider the 50-kb blocks either side of the focal gene and take the average DHS density. We then ask whether that density is correlated with the intron density of the focal gene. We find that it is both for active and inactive genes (table 3; supplementary fig. S7, Supplementary Material online). Similarly, the CHD1 density in the focal gene predicts DHS coverage in the flanking sequence (table 4), this effect being either about the same magnitude as in the inactives or much more profound when the focal gene is active, depending on the data set. To ask then whether the inactives and actives differ in the local DHS density controlling for gene intron content, we perform a loess regression and compare the residuals for the actives and inactives. We consider numerous alternative kernels for the loess to consider the consequence of different smoothing parameters. In all cases, the actives have a higher DHS density in their vicinity than the inactives (supplementary table S2, Supplementary Material online). We conclude that high intron density, high CHD1 occupancy, and gene activity of the focal lincRNAs all predict higher DHS levels in the neighborhood of the active gene, consistent with a spreading chromatin model.

We can in addition ask whether this open chromatin has any functional correlates. We might, for example, imagine that upregulation of a lincRNA with a high intron density modifies local chromatin and enables genes in the vicinity

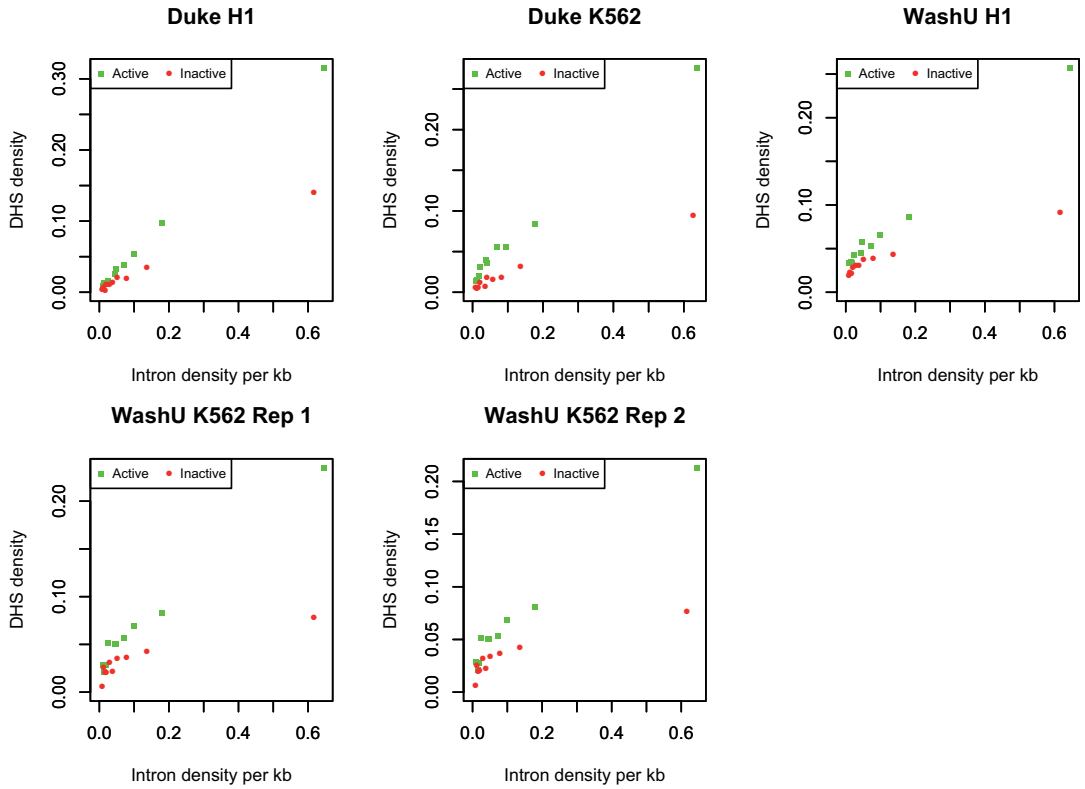


FIG. 7. DHS density as a function of intron density for active and inactive genes. As within the WashU data set DHS density is rather low (such that most short genes have no DHS peak within the gene), we here analyze the data in manner designed to avoid the inherent stochasticity this induces. First, we rank all genes by total gene length (including introns). We then divide the data into bins of equal total gene size. With ten bins, the first bin contains the longest genes whose total length in approximately 1/10 the total gene length. Thus, each bin has different numbers of genes but an equal amount of total sampled DNA. We then calculate for each bin the total number of introns to derive the number of introns per kilobase of sequence. We also consider the total number of DHS peaks and calculate the number of these per kb. All correlations are significant at $P < 0.0002$ (Spearman). In all incidences, the mean DHS density is higher in the actives than the inactives (paired *t*-test, $P < 0.05$).

Table 2. Correlation between Intragenic DHS Density and CHD1 Coverage Density Occupancy within Active Genes.

Source of DHS Data	H1	K562
Duke	$P < 2.2$ E-16 rho = 0.43	$P < 2.2$ E-16 rho = 0.51
WashU	$P < 2.2$ E-16 rho = 0.44	Rep1: $P < 2.2$ E-16 rho = 0.30 Rep2: $P < 2.2$ E-16 rho = 0.31

NOTE.—lincRNA data from Derrien et al. (2012). WashU DHS data provide two replicates for K562. We analyze both separately.

to be expressed by spreading of chromatin. Consistent with this, the neighbors of active lincRNAs are themselves especially active. Just as active genes sit at local DHS peaks, so too active lincRNAs sit at the centre of peaks of expression (fig. 9 and supplementary fig. S8, Supplementary Material online). The DHS and the expression modulation peaks extend approximately the same distance. As expected given this model, intron-rich and intron-poor lincRNAs have differing gene activity in their vicinity (correlation between intron density and percentage of genes expressed, for H1 and K562 active rho = 0.08, $P < 1 \times 10^{-8}$) (supplementary fig. S9, Supplementary Material online). This correlation is slightly weaker when the focal gene is inactive (for both rho = 0.07, $P < 10^{-6}$). Similarly the focal gene's CHD density positively correlates with the expression of neighbors (H1 actives, rho = 0.19, $P = 1.2 \times 10^{-43}$; K562 actives rho = 0.18, $P = 6.7 \times 10^{-35}$). Again controlling for intron density, using the loess method, we find that active lincRNAs have higher gene expression in their vicinity than inactive ones (supplementary table S3, Supplementary Material online).

The above evidence is consistent with the model that transcription and splicing of lincRNAs modulate chromatin of the underlying gene body which can in turn have a spreading effect, modulating expression of neighbors.

Alternative Models and Interpretations

Above we discussed three possibilities but this catalogue of possible explanations is by no means exhaustive. It has, for

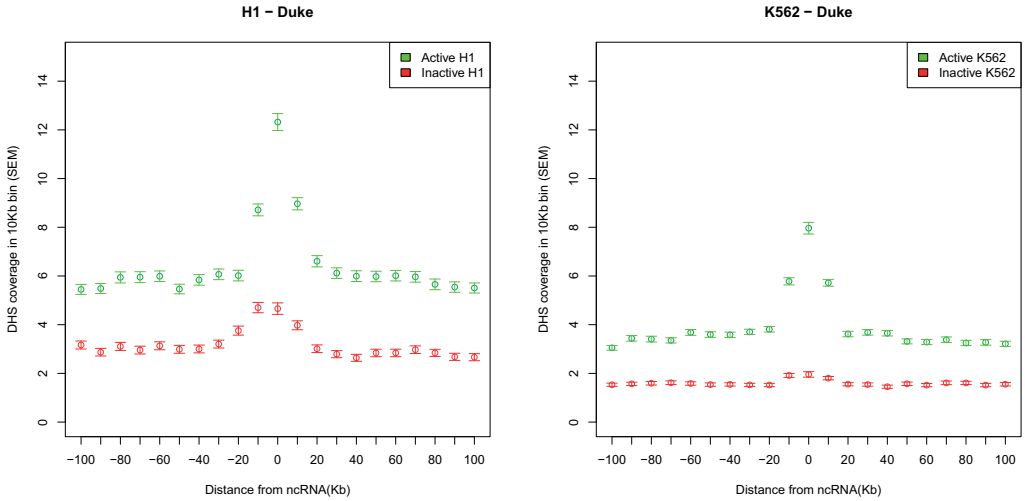


Fig. 8. The DHS density in sites flanking active and inactive lncRNAs.

Table 3. Spearman Correlation between the Focal lincRNA Gene's Intron Count Density per Kilobase and DHS Coverage per Kilobase in ± 50 kb Flanks.

Flanking Data (±50 kb)	Active Rho	Active *P*	Inactive Rho	Inactive *P*
Duke H1	0.264	2.34E-78	0.219	1.82E-50
Duke K562	0.239	1.11E-63	0.223	3.38E-52
WashU H1	0.183	5.49E-38	0.132	6.62E-19
WashU K562 Rep1	0.190	1.49E-40	0.191	1.31E-40
WashU K562 Rep2	0.146	4.12E-23	0.142	7.16E-22

Table 4. Spearman Correlation between the Focal lincRNA Gene's CHD1 Density per Kilobase and DHS Coverage per Kilobase in ±50 kb Flanks.

Flanking Data (±50 kb)	Active Rho	Active *P*	Inactive Rho	Inactive *P*
Duke H1	0.334	1.72E-127	0.303	3.84E-97
Duke K562	0.465	1.02E-258	0.264	5.14E-74
WashU H1	0.455	2.09E-247	0.369	6.16E-147
WashU K562 Rep1	0.528	0	0.529	0
WashU K562 Rep2	0.286	9.65E-87	0.286	5.69E-87

example, been demonstrated that ncRNAs can act as long-range *cis*-silencers by transcriptional interference, and could thus be regulatory active without the mature RNA being involved in the process (Pauler et al. 2012). This would be a further manifestation of the process argument, although why this process requires splicing is unclear. There might in turn be selection for the correct placement of the EJC, for reasons other than the initiation of NMD. EJCs are, for example, thought to regulate RNA localization (Giorgi and Moore 2007). Given that EJC placement is not now thought to be constitutive (Sauliere et al. 2010), it will be informative to know whether lincRNAs are unusual in their ability to attract these complexes.

Our chromatin model results are consistent with a model in which splicing of lincRNAs recruits CHD1 (and related splice associated chromatin modifiers) to transcriptionally active DNA, and this in turn enables both the chromatin within the focal gene to remain open and for there to be some spreading away from the focal active gene which is permissive for expression of neighbors. The same model, we note, also suggests a novel hypothesis for the positive correlation between intron density and expression breadth of protein-coding genes (Parmley et al. 2007) on the one hand, and the tendency for house-keeping genes to genomically cluster (Lercher et al. 2002). Broadly expressed (housekeeping) genes may be selectively favored to have absolutely more introns to enable a self-reinforcing open chromatin (N.B. intron density is higher in active genes). This would be mediated by splicing increasing the chances of recruiting CHD1 (and similar splice/chromatin modifiers) to the local DNA, which increases the chances of keeping chromatin open, enabling a higher likelihood of further transcription of the neighboring broadly expressed genes.

However while consistent with the model, our results are also consistent with alternative models. Notably, if for some other reason intron-rich genes tend to reside in domains of high gene activity then it is possible that the lincRNA has expression passively dependent on the local DHS/expression environment, much as transgenes adopt the expression profile of neighbors (Gierman et al. 2007). Consistent with this model highly and broadly expressed genes cluster in mammalian genomes (Caron et al. 2001; Lercher et al. 2002). Similarly, GC rich isochores tend to be domains of small introns and hence a higher intron density measured as introns per base pair of full gene. Note, however, in this model, given the evidence of local transgene adoption of expression profiles (Gierman et al. 2007), there is still a need to evoke the notion that local gene expression influences genes in the

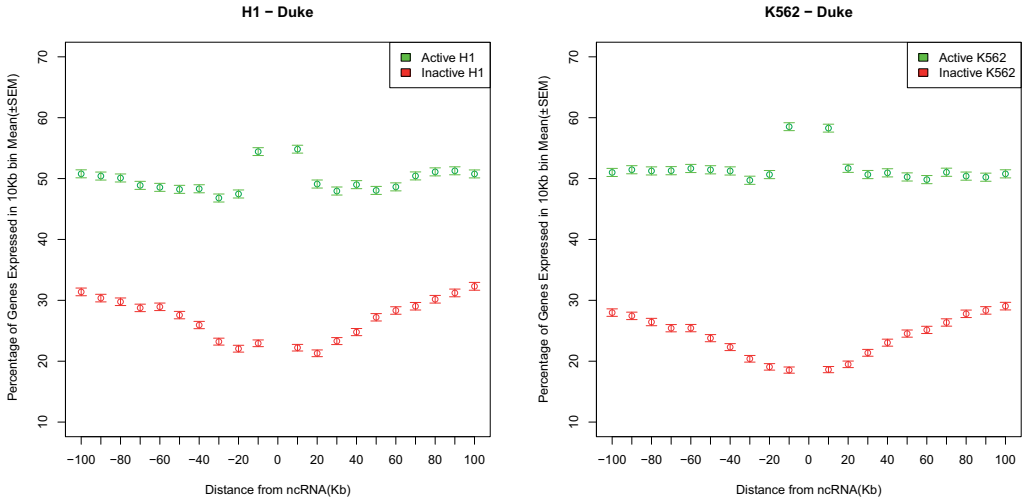


FIG. 9. Gene expression in the vicinity of active and inactive lincRNAs.

vicinity. Indeed, to disallow the possibility that lincRNAs do not affect their neighborhood, one would have to make a special case as to why one class of gene (protein coding) might affect neighbors but another (lincRNAs) do not. Why in this passive expression model intron dimensions of the focal lincRNA covary with local expression level is unclear, but could reflect a local mutational bias toward deletions. Indeed, intron size and intergene distance tend to covary (Urrutia and Hurst 2003).

The required experiment to distinguish these two explanations would be the introduction of a lincRNA with and without introns and ask whether the intron containing one affects the local DHS and expression of neighbors more than the same insertion when lacking the intron. It would also be helpful to know whether transcription rate of a lincRNA, with or without introns, might predict the extent of any upregulation of neighbors. Should it prove to be the case that intron-bearing genes modulate the expression of their neighbors, this would have consequences for the assessment of the safety of transgene inserts, as, for example, in gene therapy.

Materials and Methods

Sequences, Alignments, and Evolutionary Distances

ncRNAs are commonly classified by their length into small (18–31 nt), medium (32–200 nt) and long (from 200 nt up to several hundred kilobases) ncRNAs (Wilusz et al. 2009; Nagano and Fraser 2011). The lncRNAs are the most mysterious group among those three. Few of them have been experimentally characterized and many are poorly conserved on the sequence level (Amaral et al. 2011; Lee 2012). The group of lncRNAs can be further divided into those transcripts that overlap protein-coding genes and lincRNAs (Ponting et al. 2009). The lncRNAs that overlap protein-coding genes are most likely involved in sense–antisense regulation (Chen et al. 2005). Their evolution is likely to be constrained by the evolution of the antisense target and hence is not optimal to ask about selection on ncRNAs more generally. Here then we solely examine lincRNAs. So far, few lincRNAs have been experimentally characterized, but functional lincRNAs seem to be involved in protein-coding gene regulation by means of chromatin remodeling, transcriptional control, and posttranscriptional processing (Mercer and Mattick 2013).

The data set of putative human lincRNAs identified by Cabili et al. (2011) was downloaded as BED (Browser Extensible Data, including genomic coordinates) formatted data (supplementary material in Cabili et al. 2011). These putative lincRNAs were inferred based on the reconstruction of transcripts based on greater than 4 billion RNA-seq reads collected from 24 human tissues. In total, 10,500 putative lincRNAs have been identified by the authors. This set of candidate lincRNAs was filtered to remove transcripts where evidence for protein-coding potential could be detected (as specified by the original authors), which leads to a subset of 8,195 lincRNAs. This subset was further been filtered to remove lincRNA genes that could not be reconstructed in at least two different tissues, or reconstructed by two different assemblers in the same tissue, leaving a stringent subset of 4,662 lincRNAs (Kapranov et al. 2007). Unless otherwise noted, this stringent lincRNA subset was used for analyses in this study.

The intron and exon sequences (based on the hg19 assembly) corresponding to the lincRNA BED data were downloaded from the Galaxy server (Blankenberg et al. 2011). The galaxy server was also used to extract alignments of these regions to the rhesus macaque genome (rheMac2 assembly), based on the UCSC 46-way whole-genome multiZ alignment (Kent et al. 2002). The intron and exon alignment blocks were concatenated with the "stitch gene blocks" function provided by the Galaxy server to produce alignments of concatenated exons and concatenated introns for each lincRNA gene. The fraction of alignment positions that correspond to insertions/deletions (indels) was calculated with a

custom script and alignments with a fraction of indels higher than a given threshold were discarded. Unless otherwise noted, this threshold was set to 15% (this threshold, while arbitrary, enables comparison to other analyses).

To compare the properties of lincRNAs with those of protein-coding genes, we gathered BED12 data for the 17,132 reconstructed protein-coding transcripts from the data set and constructed alignments to the homologous regions in the macaque genome with the same approach as described for the lincRNAs. Note that we employ intronic sequence away from exon ends as a comparator not because all the sequence is necessarily neutrally evolving but because it 1) controls for local variation in the mutation rate (Matassi et al. 1999; Lercher et al. 2001), 2) conforms with numerous prior analyses (Hurst and Smith 1999; Pang et al. 2006), and 3) controls for transcription-coupled mutational/repair processes (Hanawalt and Spivak 2008). Importantly, comparison with flanking nontranscribed sequence, even if GC matched, does not control for this. If transcription-coupled repair is prevalent even on neutrally evolving sequence, in comparing exonic rates of evolution to flanking but untranscribed and hence unrepaired sequence, one could potentially misinfer purifying selection on the exon. In contrast, as introns may contain hidden residues under constraint, the comparison of exonic to intronic rates to infer purifying selection on the exons is most probably conservative. We note in addition that with biased gene conversion prevalent in the human genome (Duret and Galtier 2009) no sequence can be guaranteed to provide a perfect neutral proxy.

Evolutionary distances between human and macaque sequences were calculated with a custom implementation of the method proposed by Tamura and Kumar (2002). This method relaxes the assumption of substitution pattern homogeneity among lineages and thus allows for a more accurate distance estimation. Note that to enable fair comparison between protein-coding genes and lincRNAs we use the same metric for both. This is also meaningful as the dominant constraints that we are examining, splice-related selection and RNA folding, operate at the RNA rather than the protein level.

Expression Data

We used the expression patterns of lincRNAs and protein-coding transcripts based on the supplementary tables S2 and S6 of Cabili et al. (2011). For each lincRNA, the FPKM (fragments per kilobase of exon per million fragments mapped) value for each of the 24 studied tissues was extracted. We log-normalized the FPKM values and calculated the maximum and median FPKM for each lincRNA. The expression breadth was assessed by calculating the fraction of tissues where the respective lincRNA was detectably expressed (FPKM > 0).

For analysis of the expression of genes neighboring focal ncRNA genes, we used the profiles available for H1 and K562 cell lines on Encode portal, generated with Gencode V7 annotation (2012). We considered gene expression in bins flanking focal ncRNA genes. Average gene expression per bin is calculated as below:

Note here we simply consider whether a gene is expressed or not, not its absolute level.

ESE Hexamers

We annotated putative ESE motifs in the lincRNA and protein-coding alignments by using the set of experimentally confirmed human ESE-hexamers employed in a previous study (Parmley et al. 2006) as defined by Fairbrother, Yeo, et al. (2004). These are presented in supplementary table S4, Supplementary Material online.

RNA Folding Simulation

We used the UNAfold (Markham and Zuker 2008) software package to computationally predict the minimum energy folding of each lincRNA sequence. The "hybrid-ss-min" tool from the UNAfold package was run on each sequence with default parameters and we subsequently inferred the number of paired nucleotides from the output file. The proportion of folded nucleotides in the minimum energy RNA structure was used as a proxy for RNA-folding stability.

Assessing CHD1-Binding Sites in Active lincRNAs

We used ENCODE Project Consortium (2012) data in the latest release to find the CHD1-binding sites for the lincRNAs that both correspond to our stringent subset of the Cabili et al.'s data and are also found in the Macaque genome. Specifically, we downloaded the broadPeak data sets for the only human cell lines for which CHD1 modifications are available—k562 and h1-hesc (available on http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone, last accessed September 1, 2014).

To calculate the density for each lincRNA, the number of CHD1's peaks which are overlapping this lincRNA is divided by the lincRNA-length. In addition, we consider the sum breadth of CHD1 spans (as specified by ENCODE) and consider the proportion of this span to the gene length. Unless specified otherwise, analysis is on the number of CHD1 peaks per base pair. As the K562 and H1-hesc cell lines have not been considered in the data set of Cabili et al., we assessed whether lincRNAs were expressed in these cell lines based on Caltech and CSHL RNA seq data sets available from ENCODE (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCaltechRnaSeq and http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCshlLongRnaSeq [last accessed September 1, 2014] respectively). Based on these data sets, we find 346 lincRNAs to be actively expressed in the H1-hesc cell line and 338 in the K562 cell line. For analysis comparing various features against intron density in lincRNAs, we employ the larger lincRNA data set of Derrien et al. (2012).

As there are multiple sequences with no CHD1 binding, we were concerned that the Mann–Whitney U test might be misleading. To explore this, we used a Monte Carlo simulation to test whether the enrichment of CHD1-binding sites in active sequences could be explained by chance. To do this

for each cell type, we combined the active and inactive sets and randomly selected two sets: A hypothetical active set and a hypothetical inactive set, including the same number of sequences as observed in each cell line by querying ENCODE data. This was iterated 10,000 times, each time the difference between medians of CHD1 density in two sets was calculated and compared with the difference in medians observed in the real data. The number of times the median difference in hypothetical and randomly generated sets was as high or higher than the median difference was observed. In this test the unbiased estimation of the P of this Monte Carlo simulation is $P = (n + 1)/(m + 1)$, where n is the number of randomization as extreme or more extreme in the difference between the two classes as seen in the real data and m is the number of randomizations.

DHS-Binding Profile

DNase hypersensitive sites (DHSs) point at open chromatin segments on chromosomes. Different tissues diverge in locations of DHSs, encouraging tissue-specific gene expression patterns. The DHSs data available through ENCODE portal are generated in two production centers, University of Washington and Duke University, through a similar procedure. WashU provides two sets for K562 accessible through: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwDnase, last accessed September 1, 2014. Duke's data sets are accessible from: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeOpenChromDnase, last accessed September 1, 2014.

Supplementary Material

Supplementary figures S1–S9 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We should like to thank Chris Ponting, Ji-Long Liu and Rob Young for providing the Drosophila ncRNA annotations. This study was funded by the Volkswagen foundation, grant I/84 830 to A.S. This study was supported by a studentship award from University of Bath to A.T.G and by Medical Research Grant MR/L007215/1. The funders had no influence on the work. A.S., A.T.G., and L.D.H. performed the analyses and wrote the manuscript. L.D.H. devised the research. All authors read and approved the final manuscript. The authors declare that they have no competing interests.

References

Adam-Hall J, Georgel PT. 2011. The worlds of splicing and chromatin collide. In: Grabowski P, editor. RNA processing. InTech. Available from: http://www.intechopen.com/books/rna-processing/the-worlds-of-splicing-and-chromatin-collide.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* 39:D146–D151.

Beckedorff FC, Ayupe AC, Crocci-Souza R, Amaral MS, Nakaya HI, Soltys DT, Menck CF, Reis EM, Verjovski-Almeida S. 2013. The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation. *PLoS Genet.* 9:e1003705.

Blankenberg D, Taylor J, Nekrutenko A, Galaxy T. 2011. Making whole genome multiple alignments usable for biologists. *Bioinformatics* 27: 2426–2428.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* 25:106–110.

Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol.* 16:107–113.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25: 1915–1927.

Cáceres EF, Hurst LD. 2014. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* 14:R143.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol.* 62:89–98.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309: 1559–1563.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voûte PA, et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291:1289–1292.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 3:285–298.

Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol.* 21: 1014–1023.

Chen JJ, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* 21:203–207.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnar Z, Ponting CP. 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.* 11:R72.

Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res.* 22:885–898.

Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, Geng J, Zhang B, Yu X, Yang J, et al. 2010. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* 96: 259–265.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22: 1775–1789.

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.

Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. 2009. NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.* 37:D122–D126.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol.* 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10:285–311.

Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol.* 10:1106–1113.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268.
Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007–1013.
Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* 32: W187–W190.
Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R. 2007. Domain-wide regulation of gene expression in the human genome. *Genome Res.* 17:1286–1295.
Giorgi C, Moore MJ. 2007. The nuclear nurture and cytoplasmic nature of localized mRNPs. *Semin Cell Dev Biol.* 18:186–193.
Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* 6:1197–1211.
Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477: 295–300.
Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 14:R49.
Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol.* 9: 958–970.
Hnilicova J, Stanek D. 2011. Where splicing joins chromatin. *Nucleus* 2: 182–188.
Hoffman MM, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. *Mol Biol Evol.* 24:522–531.
Houseley J, Tollervey D. 2009. The many pathways of RNA degradation. *Cell* 136:763–776.
Hurst LD, Smith NGC. 1999. Molecular evolutionary evidence that H19 mRNA is functional. *Trends Genet.* 15:134–135.
Isken O, Kim YK, Hosoda N, Mayeur GL, Hershey JW, Maquat LE. 2008. Upf1 phosphorylation triggers translational repression during nonsense-mediated mRNA decay. *Cell* 133:314–327.
Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484–1488.
Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360–1374.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.
Khalil AM, Rinn JL. 2011. RNA-protein interactions in human health and disease. *Semin Cell Dev Biol.* 22:359–365.
Kim SH, Yi SV. 2006. Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. *Mol Biol Evol.* 23:1068–1075.
Le Hir H, Gatfield D, Izaurralde E, Moore MJ. 2001. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* 20: 4987–4997.
Lee JT. 2012. Epigenetic regulation by long noncoding RNAs. *Science* 338: 1435–1439.
Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet.* 31:180–183.
Lercher MJ, Williams EJB, Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol.* 18: 2032–2039.
Liebhaber SA. 1997. mRNA stability and the control of gene expression. *Nucleic Acids Symp Ser.* 36:29–32.
Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A.* 98: 11193–11198.
Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. 2011. Epigenetics in alternative pre-mRNA splicing. *Cell* 144:16–26.
Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. 2011. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol.* 3:1390–1404.
Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol.* 453:3–31.
Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10:R124.
Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol.* 9:786–791.
Mercer TR, Mattick JS. 2013. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol.* 20:300–307.
Nagano T, Fraser P. 2011. No-nonsense functions for long noncoding RNAs. *Cell* 145:178–181.
Nitsche A, Doose G, Tafer H, Robinson M, Saha NR, Gerdol M, Canapa A, Hoffmann S, Amemiya CT, Stadler PF. 2014. Atypical RNAs in the coelacanth transcriptome. *J Exp Zool B Mol Dev Evol.* 322: 342–351.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet.* 7:337–348.
Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* 22:1–5.
Park C, Chen XS, Yang JR, Zhang JZ. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 110:E678–E686.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.
Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol.* 24:1600–1603.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5: 343–353.
Pauler FM, Barlow DP, Hudson QJ. 2012. Mechanisms of long range silencing by imprinted macro non-coding RNAs. *Curr Opin Genet Dev.* 22:283–289.
Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17:556–565.
Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* 136:629–641.
Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol.* 24:1821–1831.
Rose D, Hiller M, Schutt K, Hackermuller J, Backofen R, Stadler PF. 2011. Computational discovery of human coding and non-coding transcripts with conserved splice sites. *Bioinformatics* 27:1894–1900.
Sauliere J, Haque N, Harms S, Barbosa I, Blanchette M, Le Hir H. 2010. The exon junction complex differentially marks spliced junctions. *Nat Struct Mol Biol.* 17:1269–1271.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* 34:2428–2437.
Sims RJ, Millhouse S, Chen C-F, Lewis BA, Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D. 2007. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol Cell.* 28:665–676.

Tamura K, Kumar S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol.* 19: 1727–1736.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* 13:2260–2264.
Wang J, Smith PJ, Krainer AR, Zhang MQ. 2005. Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.* 33:5053–5062.
Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol.* 24: 2755–2762.
Warnecke T, Hurst LD. 2011. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet.* 12: 875–881.
Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9:r29.
Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24:616–628.
Wiegand HL, Lu SH, Cullen BR. 2003. Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. *Proc Natl Acad Sci U S A.* 100:11327–11332.
Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol.* 3:1245–1252.
Wilusz JE, Sunwoo H, Spector DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23:1494–1504.
Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. *Genome Biol Evol.* 2: 190–199.
Yang JR, Liao BY, Zhuang SM, Zhang JZ. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A.* 109:E831–E840.
Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol.* 6:13.
Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP. 2012. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol Evol.* 4: 427–442.
Zeldovich KB, Shakhnovich EI. 2008. Understanding protein evolution: from protein physics to Darwinian selection. *Annu Rev Phys Chem.* 59:105–127.
Zhang J, Sun XL, Qian YM, LaDuca JP, Maquat LE. 1998. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol.* 18:5272–5283.

SUPPLEMENTARY DATA

Supplementary figure 1. Relative frequencies of bases predicted to be part of an ESE motif as a function of the distance to the nearest intron, starting at a distance of 6 (a+b). Shown in c+d is the decadic logarithm of the average intron length for lincRNA and protein-coding genes vs. the density of ESE motifs on the exon sequences of this gene. For c+d, "density" has been measured as the number of nucleotides that belong to a putative ESE motif divided by the summed length of exons for the respective gene. This figure includes all lincRNAs instead of only the conservative subset (see methods).

Supplementary Figure 2. ESE motifs evolve slower than non-ESE sites. The substitution rates in ESEs and non-ESEs are shown as a function of the distance from the nearest splice-junction. This figure includes all lincRNAs instead of only the conservative subset (see methods).

Supplementary Figure 3. Exon cores and flanks evolve at different rates. The distributions of *Kec*/*Kic* and *Kef*/*Kec* values are shown for protein-coding genes and lincRNAs. This figure includes all lincRNAs instead of only the conservative subset (see methods).

Supplementary Figure 4. Evolutionary rates and lincRNA expression. The evolutionary distance to the macaque homologue was plotted vs the values of maximum expression (a), median expression (b) and expression breadth (c) for each lincRNA. This figure includes all lincRNAs instead of only the conservative subset (see methods).

Supplementary Figure 5. DHS density in the vicinity of active and inactive lincRNAs from the WashU data.

Supplementary Figure 6. The relationship between local DHS density (+/-50kb either side of a focal gene) and the intron density of that gene.

Supplementary Figure 7. Proportion of gene expressed in the vicinity of active and inactive lincRNAs from the WashU data.

Supplementary Figure 8. The relationship between the local expression (proportion of genes expressed in +/- 50kb window) and the intron density of the focal gene.

Table S1. Normal and partial correlations with evolutionary rate (measured as Tamura-Kumar distance, see methods) using Pearson product moment correlation and Spearmans. Numbers highlighted in bold are significant after Bonferonni correction with N=14.

Table S2. Analysis of residuals of loess plot of intron density versus local DHS density

Table S3. Analysis of residuals of loess plot of intron density versus local expression

Table 1. Normal and partial correlations with evolutionary rate (measured as Tamura-Kumar distance, see methods) using Spearman's correlation (for Pearson correlation see supplementary table 1). Numbers highlighted in bold are significant after Bonferonni correction with N=14.

LincRNA	Normal	Partial
Max. Expression rate	-0.005	0.032
Med. Expression rate	-0.025	-0.035
Exp. breadth	-0.038	**-0.091****
RNA stability	0.048$^{\#}$	0.009
Frac70	**-0.051**$^{\#}$	-0.011
ESE density	**-0.182*****	**-0.194*****
GC	**-0.058***	**-0.102*****
Protein coding	Normal	Partial
Max. expression	**-0.203*****	-0.019
Med. Expression	**-0.339*****	**-0.063*****
Exp. Breadth	**-0.369*****	**-0.189*****
RNA stability	**0.154*****	**0.028***
Frac70	**-0.222*****	**-0.101*****
ESE density	**-0.29*****	0.008
GC	**0.313*****	**0.168*****

Significance codes for p-values: $^{\#}$ $P < 0.01$;* $P < 10^{-3}$; ** $P < 10^{-6}$;*** $P < 10^{-9}$

Table 2. Correlation between intragenic DHS density and CHD1 coverage density occupancy within active genes. LincRNA data from Derrien et al.(Derrien, et al. 2012). WashU DHS data provides two replicates for K562. We analyse both separately.

Source of DHS data	H1	K562
Duke	p-value < 2.2e-16 rho = 0.429496	p-value < 2.2e-16 rho = 0.5130903
WashU	p-value < 2.2e-16 rho = 0.4447778	Rep1: p-value < 2.2e-16 rho = 0.3011269 Rep2: p-value < 2.2e-16 rho = 0.3076285

Table 3. Spearman correlation between the focal lincRNA gene's Intron count density per Kb and DHS coverage per Kb in ±50Kb flanks

Flanking data (+/- 50kb)	Active rho	Active P	Inactive rho	Inactive P
Duke H1	0.263832376	2.34E-78	0.218965279	1.82E-50
Duke K562	0.238683917	1.11E-63	0.22204091	3.38E-52
WashU H1	0.183116797	5.49E-38	0.131202149	6.62E-19
WashU K562 Rep1	0.189931885	1.49E-40	0.190064325	1.31E-40
WashU K562 Rep2	0.145656374	4.12E-23	0.141437648	7.16E-22

Table 4 Spearman correlation between the focal lincRNA gene's CHD1 density per Kb and DHS coverage per Kb in ±50Kb flanks

Flanking data (+/- 50kb)	Active rho	Active P	Inactive rho	Inactive P
Duke H1	0.334331422	1.72E-127	0.303009846	3.84E-97
Duke K562	0.465303297	1.02E-258	0.264378638	5.14E-74
WashU H1	0.454898251	2.09E-247	0.36938285	6.16E-147
WashU K562 Rep1	0.528028699	0	0.529417285	0
WashU K562 Rep2	0.285852821	9.65E-87	0.286223385	5.69E-87

Figure 1. Relative frequencies of bases predicted to be part of an ESE motif as a function of the distance to the nearest intron, starting at a distance of 6 (a+b). Shown in c+d is the decadic logarithm of the average intron length for lincRNA and protein-coding genes vs. the density of ESE motifs on the exon sequences of this gene. For c+d, "density" has been measured as the number of nucleotides that belong to a putative ESE motif divided by the summed length of exons for the respective gene. This figure includes only the conservative lincRNAs. For the complete set see Supplementary Figure 1.

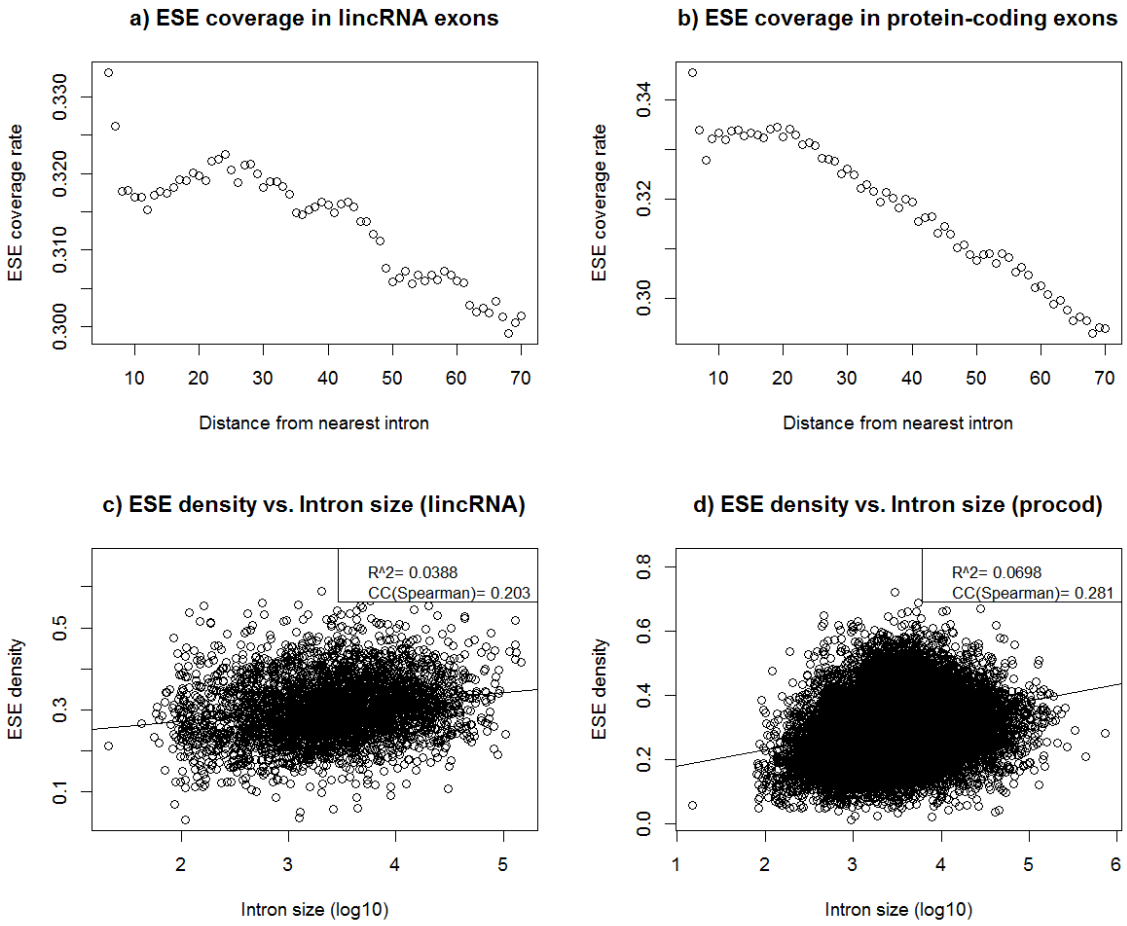


Figure 2. ESE motifs evolve slower than non-ESE sites. The substitution rates in ESEs and non-ESEs are shown as a function of the distance from the nearest splice-junction. This figure includes only the conservative lincRNAs. For the complete set see Supplementary Figure 2.

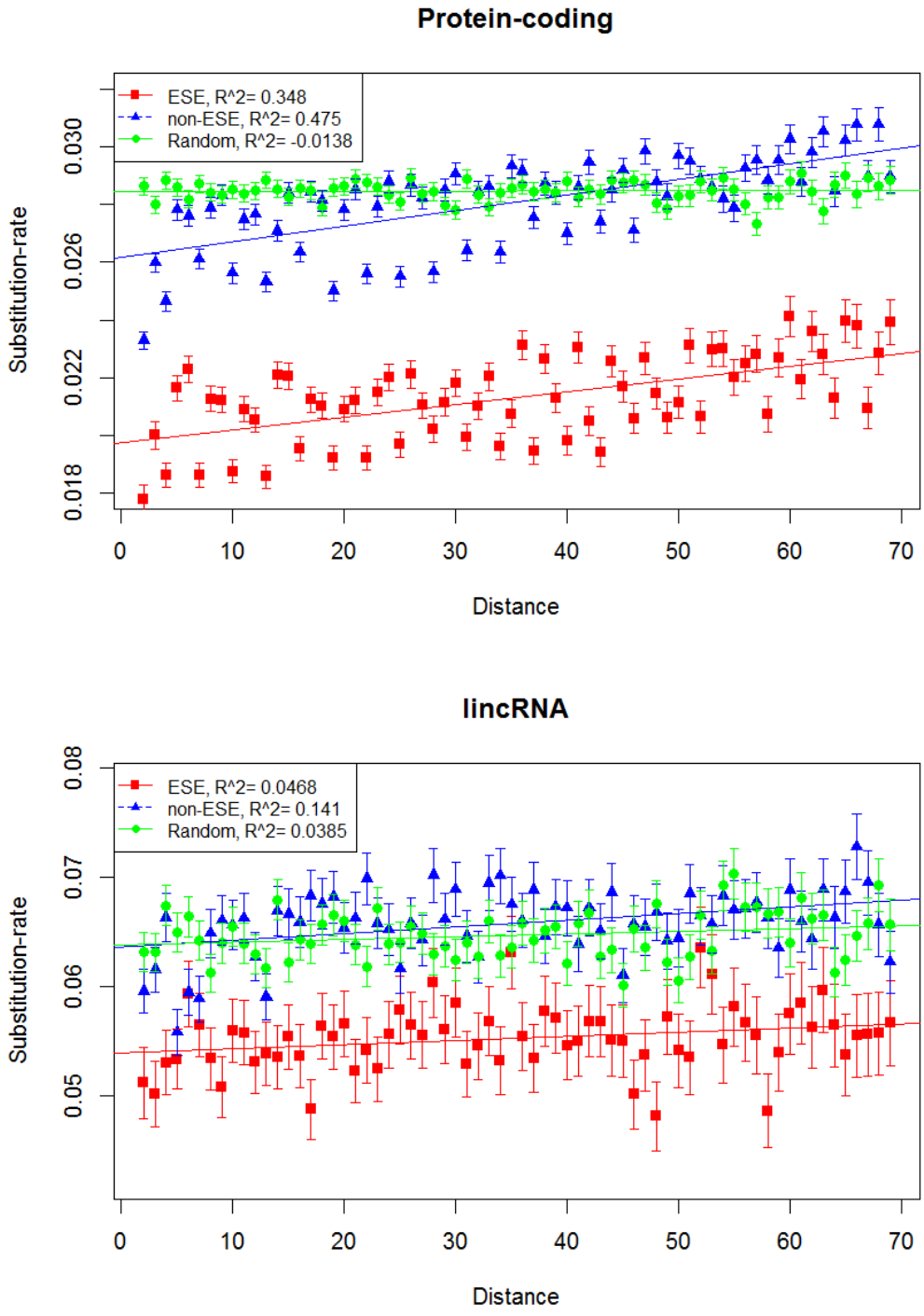


Fig. 3 Exon cores and flanks evolve at different rates. The distributions of *Kec/Kic* and *Kef/Kec* values are shown for protein-coding genes and lincRNAs.

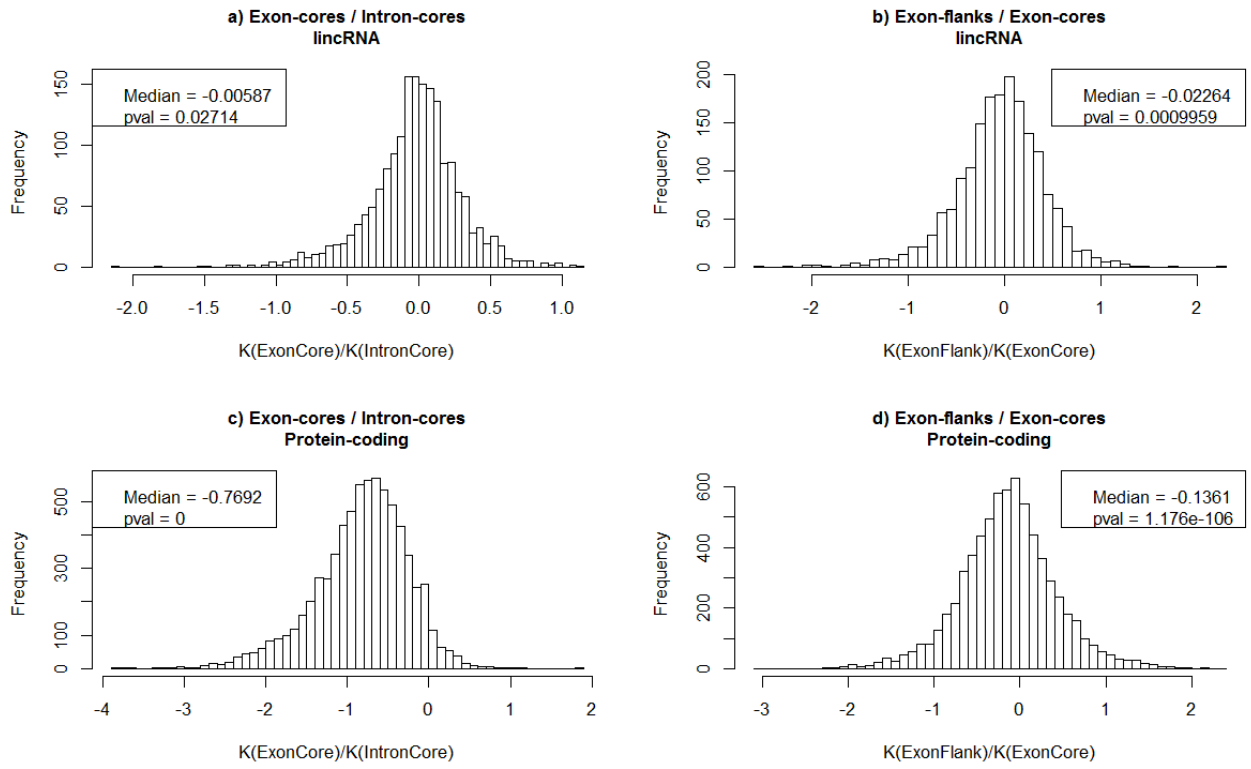


Figure 4 Correlation of expression parameters with evolutionary rates of lincRNAs and protein-coding genes. The evolutionary distance to the macaque homologue was plotted vs the values of maximum expression (a), median expression (b) and expression breadth (c) for each lincRNA.

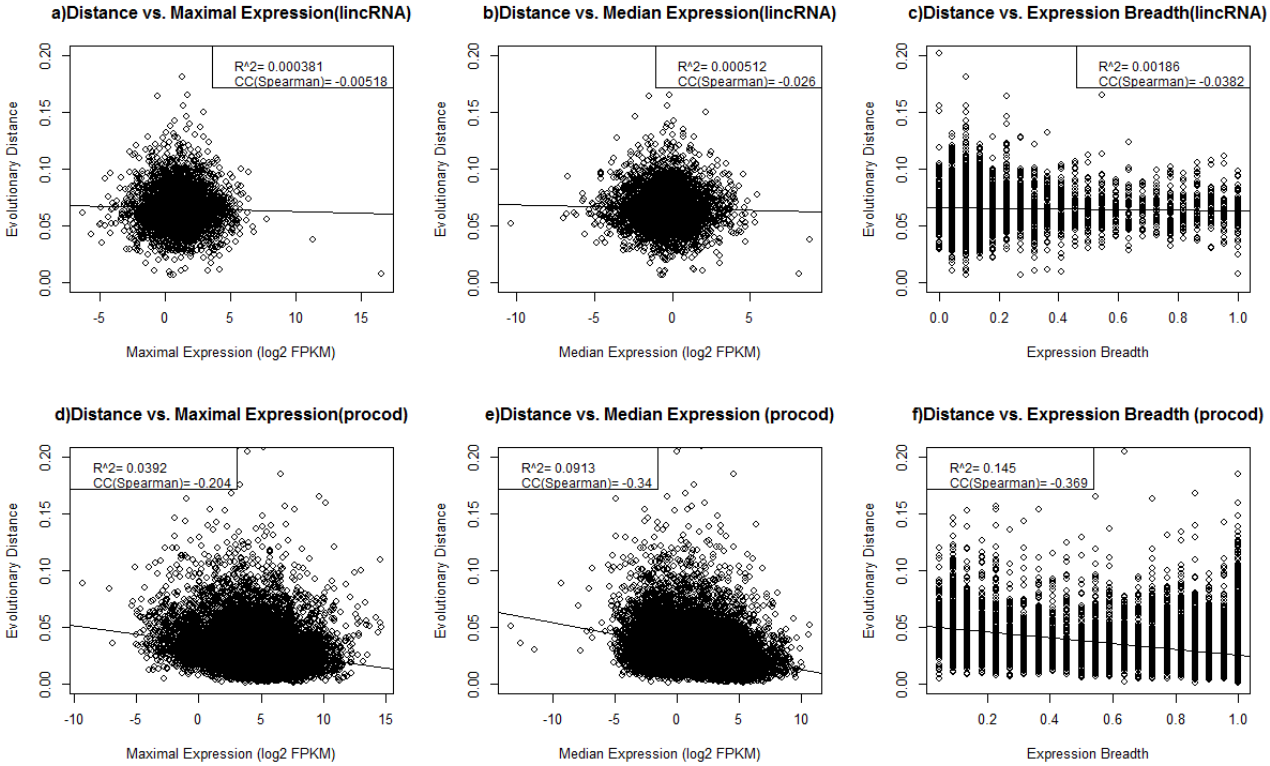


Figure 5 Folding stability and expression of lincRNAs. The folding stability, assessed as the fraction of paired nucleotides in the minimum energy fold, is plotted against maximum (a) and median expression (b), expression breadth (c) and evolutionary distance (d).

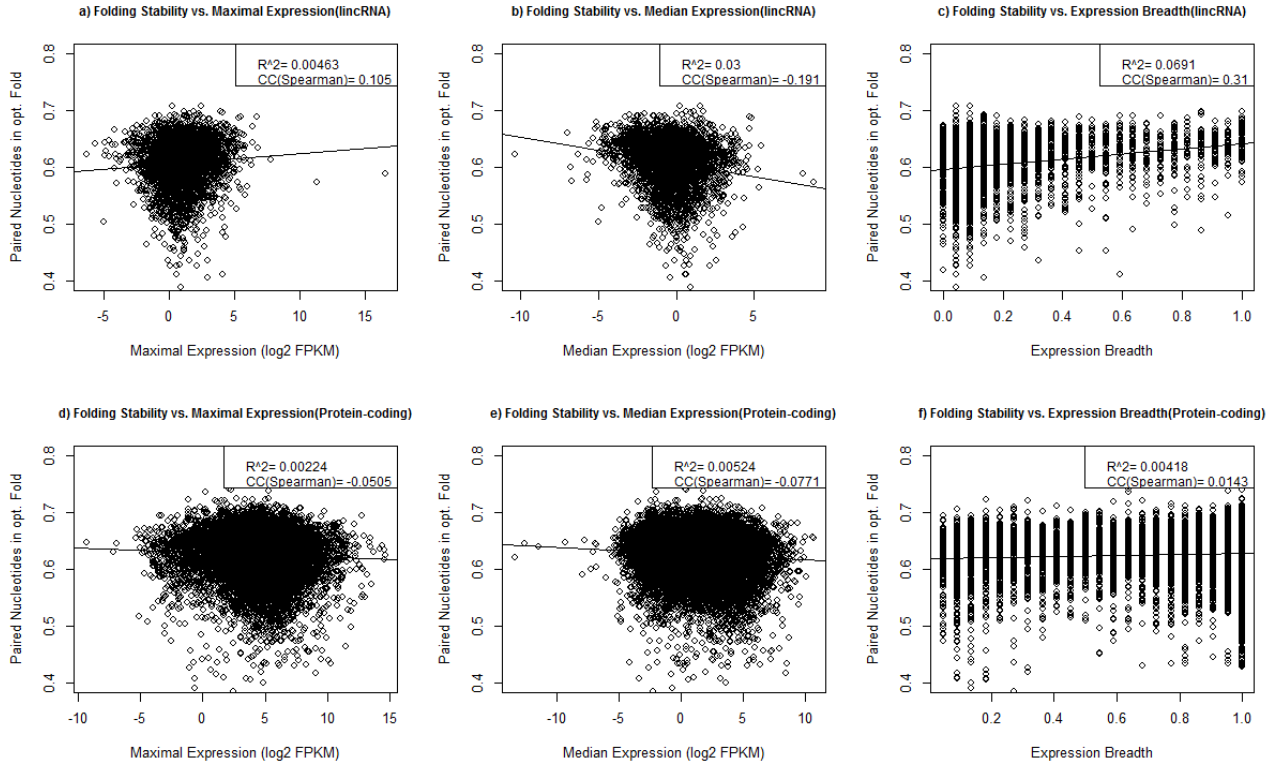


Figure 6. CHD1 density within lincRNAs is higher in active intron rich genes. Here, for each gene, we consider the number of CHD1 peaks (as specified by ENCODE) per unit base pair of each gene and compare this with the number of introns per unit base pair of gene length (in both cases we employ the length of the unspliced gene). We consider those lincRNAs that are transcriptionally active or inactive in each cell type separately. As can be seen, active genes have higher CHD1 density the more introns they have. For H1 active, rho=0.23, $P<2.2 \times 10^{-16}$, for K562 rho=0.16, $P<2.2 \times 10^{-16}$. For the inactives, the inverse is seen the effect being greatly owing to the great number of intron rich genes without any CHD1: For H1 inactive, rho=-0.11, $P<5.2 \times 10^{-15}$, for K562 rho=-0.19, $P<2.2 \times 10^{-16}$. Note that the genes considered active or inactive in the two cells are specific to each cell and the CHD1 measure is similarly specific to each cell type. Thus the two cell types are independent tests of the same hypothesis. Considering CHD1 coverage (i.e. proportion of gene covered by at least one CHD1 span) doesn't affect conclusions: H1 active, rho=0.1, $P<10^{-12}$, K562 active rho=0.08, $P < 10^{-8}$, inactives: H1 rho=-0.15, $P<2.2 \times 10^{-16}$, K562 rho=-0.23, $P<2.2 \times 10^{-16}$.

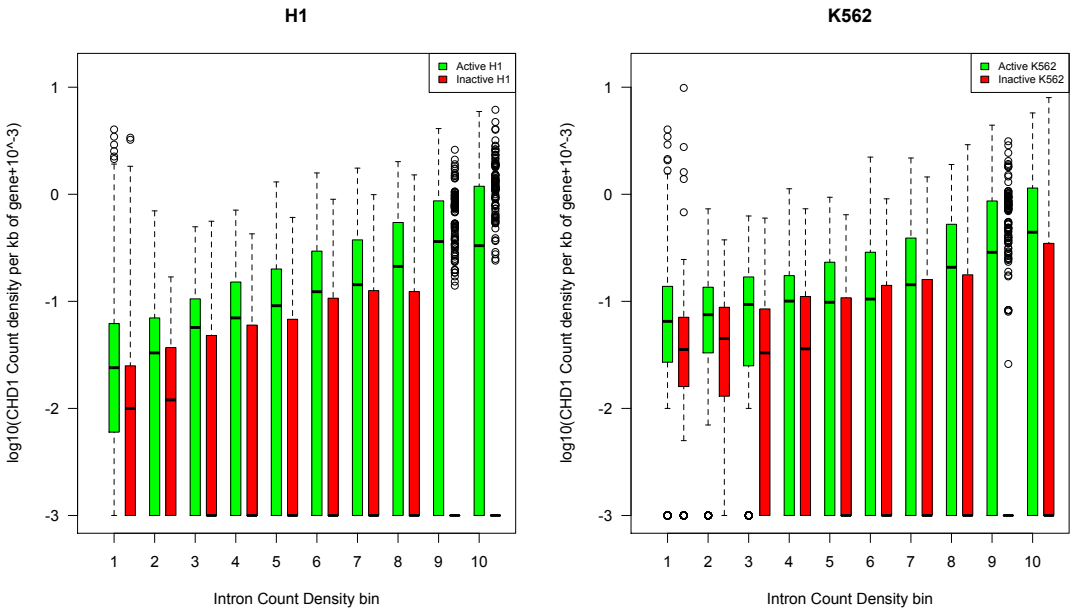


Figure 7 DHS density as a function of intron density for active and inactive genes. As within the WashU dataset DHS density is rather low (such that most short genes have no DHS peak within the gene), we here analyse the data in manner designed to avoid the inherent stochasticity this induces. First we rank all genes by total gene length (including introns). We then divide the data into bins of equal total gene size. With ten bins, the first bin contains the longest genes whose total length in approximately 1/10 the total gene length. Thus each bin has different numbers of genes but an equal amount of total sampled DNA. We then calculate for each bin the total number of introns to derive the number of introns per kb of sequence. We also consider the total number of DHS peaks and calculate the number of these per kb. All correlations are significant at P <0.0002 (Spearmans). In all incidences the mean DHS density is higher in the actives than the inactives (paired t. test, P<0.05)

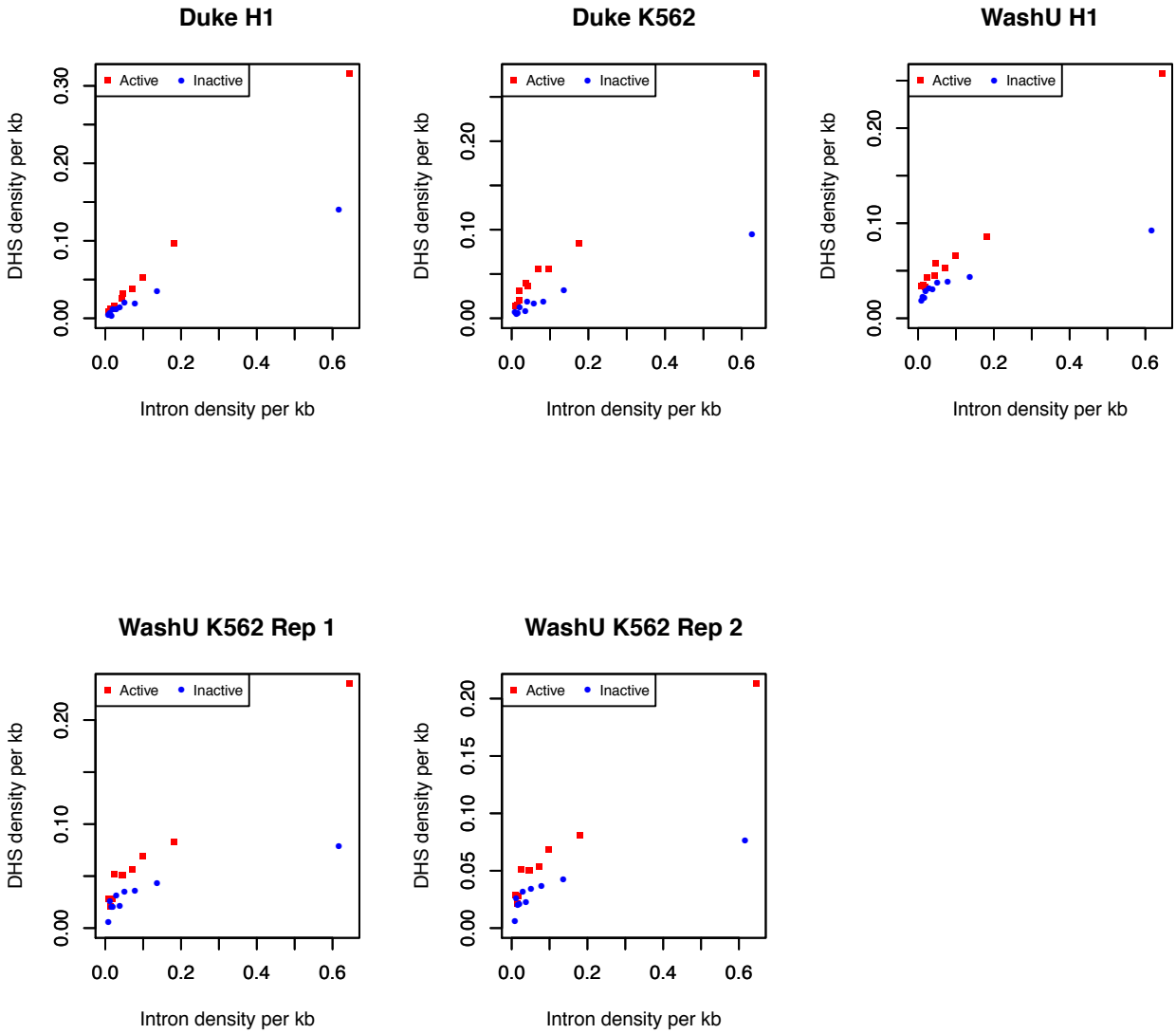


Figure 8 The DHS density in sites flanking active and inactive lncRNAs.

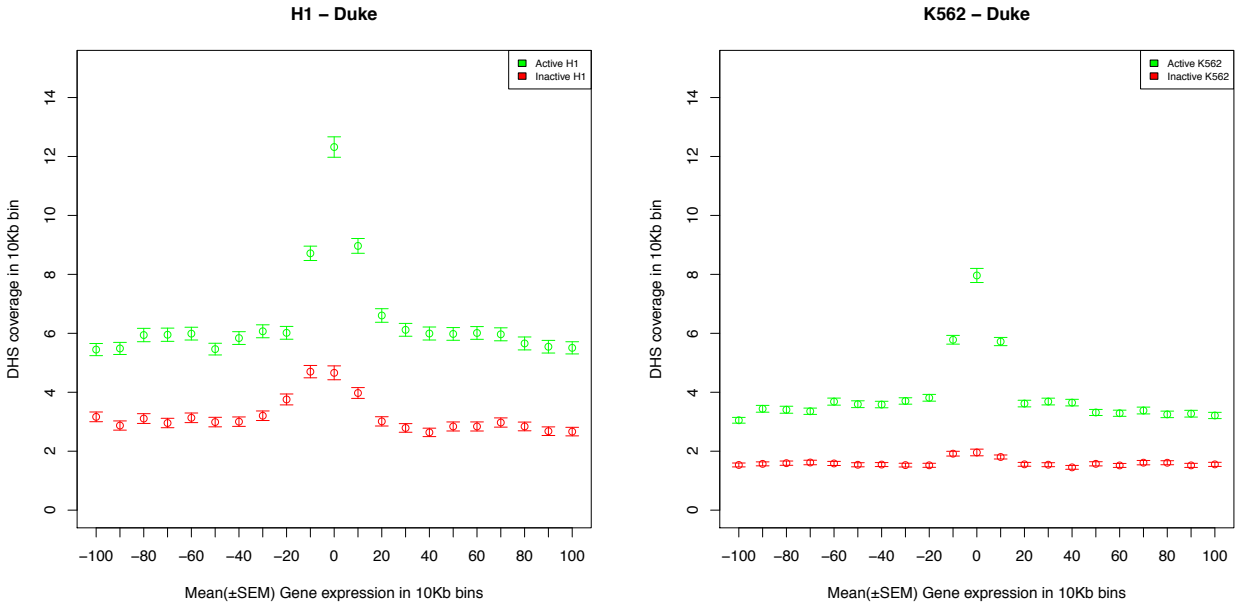
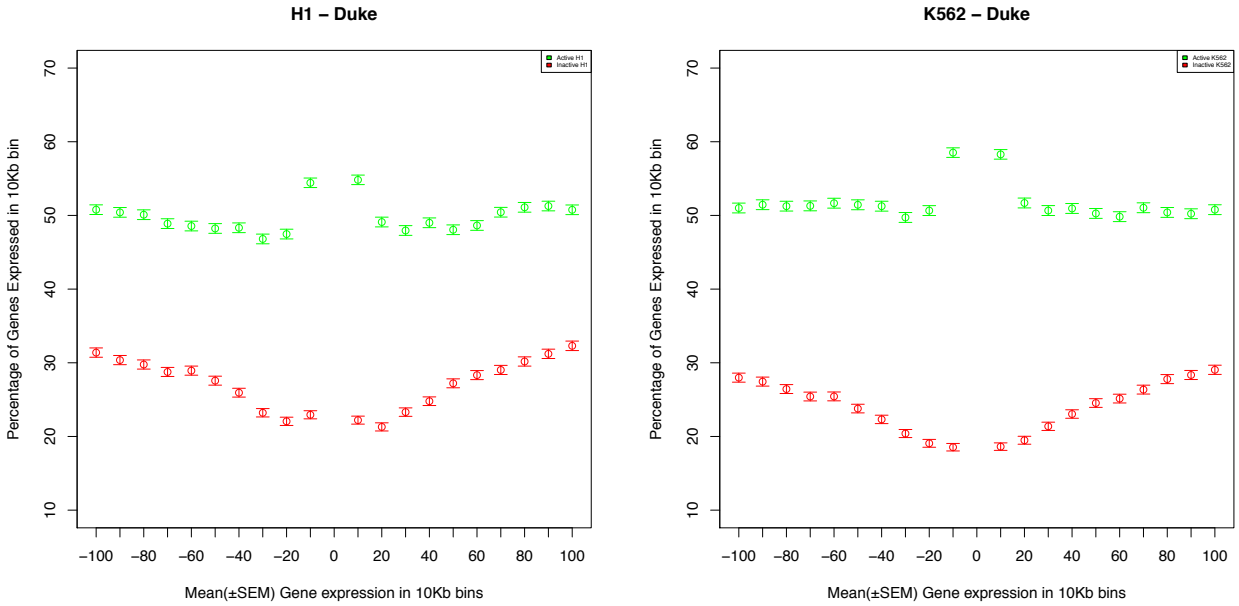


Figure 9 Gene expression in the vicinity of active and inactive lincRNAs



Chapter 5. Highly prized immigrants: How endogenous retrovirus elements rewired our pluripotency network

Introduction

In previous chapters, I have shown that neighboring coding genes correlate in their evolution of expression across Primates and Yeasts. I have also presented evidence on how weak purifying selection found splice enhancer motifs points at non-coding genes being involved in regulating their neighboring coding genes. In this chapter, I will take one step further and show how study of a primate specific endogenous retrovirus family in the context of stem cells has revealed the critical role these foreign element play in rewiring our pluripotency network. Human specific endogenous retroviruses, HERVs, provide an excellent opportunity to investigate the effect of randomly scattered almost-identical sequences on evolution of expression profile of their neighboring genes. In this sense, HERVs can be seen as a naturally occurring transgene experiment.

It needs to be clarified that the data and analysis provided in this chapter are result of a collaboration with the lab of Prof. Zsuzsanna Izvak, at Max Delbruck center in Berlin. Result of this collaboration has been published in *Nature*. Some of the analyses I have conducted were removed from this publication due to word count restrictions, accepted number of figures and other limitations enforced by the journal. Also due to replication of many of my *in silico* analyses *in vivo*, we have decided to drop some of the Bioinformatics analyses out of the final paper. However, as these analyses were instrumental in the discovery and final conclusion stated in the publication, here I explain them and also clarify my contribution.

Prof. Izvak and her team have discovered a particular family of HERVs, HERV-H, to be highly expressed in human embryonic stem cells and were interested to find out why HERV-Hs, among all other families of repeat elements, exhibit such a curious expression profile. They provided us with a list of transcriptionally active HERV-H and I started analyzing them *in silico*. I first compared their expression profile in stem cells in comparison to other tissues. As their expression pattern proved to be distinctive, I then analysed the chromatin state and epigenetics marks in their vicinity. This was followed by several transcription factor analyses and comparison of binding sites for several chromatin remodelers.

These Bioinformatics analyses have shown actively transcribed members of HERV-H are involved in regulating their neighbours. They not only provide functional binding sites for a combination of naïve pluripotency transcription factors but also, I discovered, a novel transcription factor, LBP9, which interestingly was found to be a marker for naïve stem cells by several *in vitro* experiments done in Prof. Izvak's lab and is

currently being patented. Hence, through providing a harbor for LBP9, the long terminal repeats, LTRs, associated with HERV-H family rewire regulating mechanisms associated with pluripotency and also initiate transcription of the neighboring genes. The full *Nature* publication resulting from our collaboration has been included at the end of this chapter, but first I will explain my contribution *in silico* analyses in more detail.

A clarification on the definition of "active sequences" is needed before any explanation on the analyses themselves could be initiated. Prof. Izvak and her team has first provided us with a list of manually assembled HERV sequences. These are composed of long sequences each consisting on one or more HERV and their terminal repeats. Those with unique RNA seq reads attached to them were consider to be active. This definition of active sequences suits the type of analyses addressing patterns in larger segments of genome, like open chromatin analysis. Prof. Izvak and her team have also specified the list of individual HERV-Hs and their LTRs plus a shorted list of HERV-Ks and their LTRs which were all found to be actively transcribed in HESC cell line under study. All of the analyses stated in this chapter are done on the later list of active genes if not otherwise stated.

It is also important to clarify the definition of "extended active" and "extended inactive" sequences. In the primitive analyses conducted, I found an interesting signature in active HERV-Hs and their LTRs, LTR7s. As explained below, we found these active sequences to be in open stretches on chromosomes or be very close to open chromatin. This inspired us to not only include the bodies of these active sequences in our study but also include a short stretch of DNA up and downstream to them. After a consultation with our wet lab collaborators and running a benchmark, it was decided to extend active sequences 1.5Kb on either sides.

Active HERV-Hs are more often in vicinity of open chromatin in stem cells rather than any other tissues studied

DNase I hypersensitive sites, DHSs, are used as indicators for segments of open chromatin across chromosomes (Thurman et al. 2012). DHS data is available for a few cell lines through ENCODE portal (Bernstein et al. 2012). Using these publically available datasets, I asked whether active manually assembled HERV-Hs represent a tissue specific pattern. In other words, could we find evidence for open chromatin in one specific cell line compared to the others available. And the answer is yes, active HERV-Hs are enriched in DHS peaks in human embryonic stem cell line, HESC, compared to any other cell lines, as shown in table 1. The stark difference between HESC and induced pluripotent stem cell, iPS is especially interesting and at the time was instrumental in follow up *in vitro* analysis.

Table 1. Active HERV-Hs are enriched in DHS peak in HESC cell lines compared to induced pluripotent stem cell

	HESC	iPS	Neuron	Fibroblast	Trophoblast
#active HERV-Hs overlapping at least one DHS(s)	214	16	18	15	0
#active HERV-Hs not overlapping any DHS(s)	183	381	379	382	397

I have also conducted a Monte Carlo simulation to investigate statistical significance of having the same or more sequences overlapping one or more DHSs just by chance. To do this for 10,000 iterations, I have computationally generated 397 random sequences and then counted the number of sequences including one or more DHSs, counted the number of times the same or more sequences were found to overlap DHSs. From this count one could calculate a Monte Carlo derived p-value to find out if the number observed could just have happened by chance. However, the p-value of observing this number of overlaps by chance is 1×10^{-6}. This means that HERV-Hs are highly significantly enriched in open chromatin marks,

I have also expanded on this analysis by investigating HERV-Hs for which there was no evidence for overlapping a DHS peak, as a mark for open chromatin. I asked if no open chromatin evidence is found for these active HERV-Hs, how far are they from the closest open chromatin. So the distance between these active HERV-Hs to their closest DHS peak were calculated. As shown on figure 1, we found that while active HERV-Hs not overlapping a DHS peak themselves are very close to a DHS peak in HESC cell lines, they are mostly mega base pairs away from an open chromatin mark in other cell lines. Some members of HERV-K are also active in HESC cell line but they don't exhibit a similar patter in overlapping DHSs.

Active HERV-H and their long terminal repeats are highly enriched in open chromatin DHS peaks

Now that longer manually assembled active sequences were found to be overlapping a DHS or be very close to one, do the active individual HERV-H and their long terminal repeats, LTR7s, show the same pattern? Using ENCODE data for HESC cell line, the number of active and inactive HERV-Hs and LTR7s were counted. A simple chi test revealed they are both enriched in open chromatin DHS peaks, table 2. Expected value shown in table 2 is calculated as follows. If the number of active sequences were shown by N_a, the number of inactive ones by N_i, the number of in active sequence overlapping at least one DHS peak by H_a and the number of inactive sequences overlapping at least one DHS peak by H_i, then the number of active sequences expected to overlap at least one DHS peak when there was no statistical difference between the active and inactive sequences could be calculated by:

$$\#Expected\ H_a = \frac{(H_a + H_i)N_a}{N_a + N_i}$$

Some members of HERV-K are also active in HESC cell line but they are not significantly enriched in DHSs as chi-squared test showed.

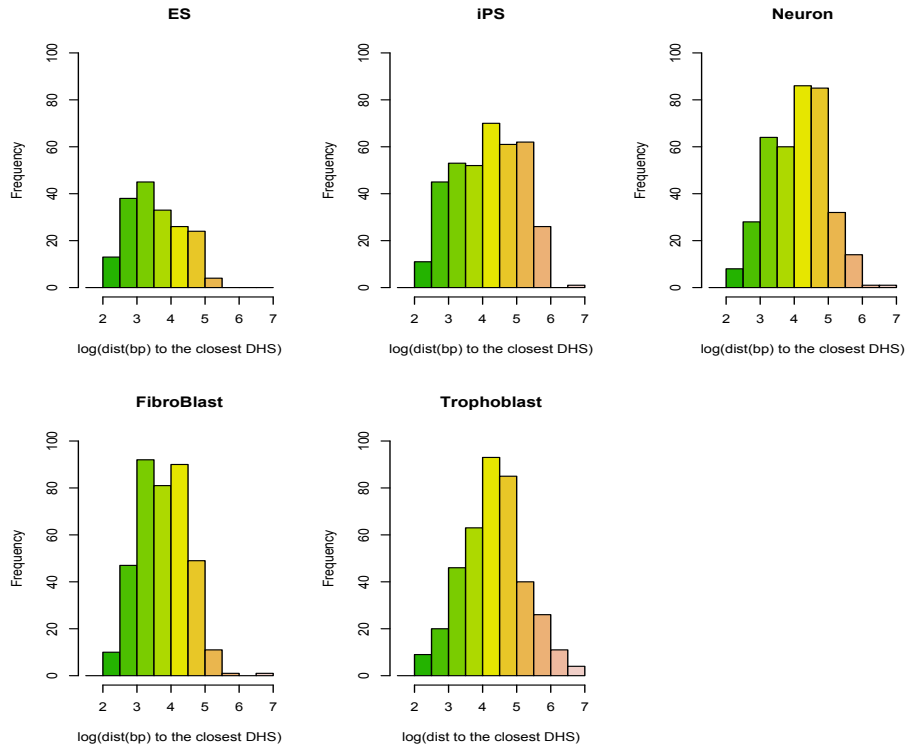


Figure 1. Cross-tissue comparison of the distance of the closest DHS to the active sequences not overlapping any DHS peaks. The distances are presented in log ratio.

HERV-H is associated with activating but not repressive histone marks

Now that active HERV-Hs and their LTRs are enriched in open chromatin DHS peaks, are they also enriched in activating or repressive epigenetic marks? To investigate this, I used two of the best studied epigenetics marks, H3K4me3 and H3K27me3, both of which were available through ENCODE for HESC cell line. H3K4me3 is associated with transcriptional activation while H3K27me3 is shown to be repressive (Sims et al. 2007; Cedar and Bergman 2009; Bock 2012; Greer and Shi 2012). Comparison of H3K4me3 and H3K27me3, across the active and inactive HERV-H and also LTR7 have shown the active HERV-H and LTR7s to be highly enriched in activating histone mark but not the repressive histone mark (Table 2). Chi squared P-values are less than 0.001

in H3K4me3 analysis, indicating the difference between activating histone mark enrichment in active and inactive HERV-Hs and LTR7s to be statistically significant.

No evidence for bivalent chromatin mark in active or inactive HERV-H and LTR7

Highly conserved non-coding sequences in ESC also manifest a distinctive histone mark signature. It is thought that these regions are poised into a dynamic state marked by large segments enriched in H3K4me3 containing short and sporadic intervals of H3K27me3 marks (Bernstein et al. 2006). This bivalent mark is also known to be reduced in cancer stem cells (Soejima 2010). I have examined this bivalent histone mark but did not find any evidence for it in active or inactive sequences of HERV-H or LTR7.

Active HERV-Hs and their long terminal repeats are enriched in chromatin modifier binding sites compared to inactive HERV-Hs

CHD1 is a member of the chromodomain helicase DNA binding, CHD, family of proteins that interacts with nucleosomes and plays a role in chromatin remodeling to modulate transcription. The members of the CHD family of proteins possess 3 common structural and functional domains: a chromodomain, to act as chromatin organization modifier, an SNF2-like helicase/ATPase domain and a C-terminal DNA-binding domain. CHD1 is the most studied member of this family and has been shown to interact with the transcriptional co-repressor NCoR and histone deacetylase 1 indicating a role in transcriptional regulation. (Marfella and Imbalzano 2007; Sims et al. 2007; Zentner et al. 2013). CHD1 has also been shown to interact with the Paf1 complex and Rtf1 implicating an additional role in transcriptional elongation (Warner et al. 2007). Importantly, recently CHD1 was shown to be involved regulation of pluripotency in ESCs (Piatti et al. 2015; Siggens et al. 2015)

The previous analysis in lincRNAs, explained in previous chapter, I found evidence for CHD1's role in regulating the neighboring genes. So do we see a similar pattern emerging in repeat elements too? To find the answer, I have compared CHD1's binding sites in active and non-active extended HERV-Hs and LTR7s and found active HERV-Hs and LTR7s to be highly enriched in CHD1's binding sites compared to non-active ones. This is shown in table 2 below and the source of CHD1 binding sites is the dataset provided in HESC on ENCODE (Bernstein et al. 2012). I also compared binding sites of three other chromatin remodelers, MYC, MAX and CHD2, for which experimental data is provided through the ENCODE's portal. A very similar pattern emerged, active HERV-Hs and LTR7s were found to be highly enriched in all three of these chromatin remodelers.

I have also asked if this is a signature of HERV-H family of endogenous retroviruses. Since our wet lab collaborators have found a few members of HERV-Ks to be also transcribed in embryonic stem cells in human, I could inspect them and their long

terminal repeats to find the answer. Active HERV-Ks and their long terminal repeats, LTR5s, were not found to be enriched in any of these four chromatin remodelers. So we concluded this to be a signature of HERV-H family and our wet lab collaborators then have conducted experiments to show this *in vivo*. More detail is included in the paper attached below.

Table 2. Chi-squared comparison between active and inactive HERV-H, their long terminal repeats, LTR7 across open chromatin DHS peaks, activating and repressive histone marks and also four chromatin modifiers are represented. Active HERV-H and LTR7 are highly enriched in open chromatin DHS peaks and also are highly enriched in activating histone mark, H3K4me3. Nonetheless, neither is enriched in repressive histone mark, H3K27me3. More interestingly, active HERV-H and LTR7s are highly enriched in four chromatin modifiers, CHD1, CHD2, c-MYC and MAX. Neither of these statements is true for HERV-K and their long terminal repeats, LTR5.

		Total	Active	Inactive	In extended active	Expected active	In extended inactive	Expected inactive	Chi Squared (P)
DHS	HERVH	6030	874	5156	473	159.7	629	942.3	718.8 (P<<0.001)
	LTR7	3218	212	3006	293	112.06	1408	1588.94	312.75 (P<<0.001)
	HERVK	247	13	234	6	5.1	91	91.9	0.168 (P≅0.7)
	LTR5	615	21	594	5	5.6	159	153.6	0.254 (P≅0.6)
H3K4me3	HERVH	6030	874	5156	2527.5	644.4	1918.5	3801.6	6435.6 (P<<0.001)
	LTR7	3218	212	3006	473	91.24	912	1295.05	1709.9(P<<0.001)
	HERVK	247	13	234	10	2.2	31.5	39.3	29.2 (P<0.001)
	LTR5	615	21	594	1.5	3.4	97	95.1	1.09 (P≅0.95)
H3K27me3	HERVH	6030	874	5156	11.5	14.7	90	86.8	0.86 (P>0.05)
	LTR7	3218	212	3006	4.5	7.15	104	101.35	1.05 (P>0.05)
	HERVK	247	13	234	1.5	1.6	29	28.9	0.0066 (P≅0.95)
	LTR5	615	21	594	7	2.7	72	76.3	7.09 (p<0.01)
CHD1	HERV-H	6030	874	5156	435	83.5	141	492.5	1730.6(P<<0.001)
	LTR7	18079	212	17867	88	4.43	290	373.57	1594.2(P<<0.001)
	HERVK	247	13	234	7	7.6	137	136.4	0.05(P>0.05)
	LTR5_Hs	615	21	594	13	11.7	331	332.3	0.15(P>0.05)
c-MYC	LTR7	3218	212	3006	19	10.21	136	144.79	8.09 (P=0.0044)
	HERVK	247	13	234	1	1.11	20	19.89	0.01 (P>0.05)
	LTR5	615	21	594	4	1.98	54	56.02	2.13 (P>0.05)
MAX	HERVH	6030	874	5156	99	47.68	230	281.31	64.57 (P<<0.001)
	LTR7	3218	212	3006	51	23.85	311	338.15	33.09 (P<<0.001)
	HERVK	247	13	234	2	1.84	33	33.15	0.014 (P>0.05)
	LTR5	615	21	594	5	4.54	128	128.46	0.047 (P>0.05)
CHD2	HERVH	6030	874	5156	174	50	171	294.99	359.58 (P<<0.001)
	LTR7	3218	212	3006	76	18.38	203	260.62	193.36 (P<<0.001)
	HERVK	247	13	234	0	0.79	15	14.21	0.833 (P>0.05)
	LTR5	615	21	594	4	1.50	40	42.49	4.29 (P= 0.038)

LTR7s provide binding sites for a novel transcription factor rewiring pluripotency network

So far I have shown active HERV-Hs and LTR7s are associated with open chromatin, are enriched in activating histone marks and four chromatin remodelers. So they are changing chromatin state from close to open which would allow for transcription of their neighboring genes. In this content, one might ask if they also provide binding site for specific transcription factor(s), TF(s). In other word, is the expression of active HERV-Hs or LTR7s regulated by a particular transcription factor or combination of transcription factors? While in vivo experiments are essential in verifying TFs, they are

largely ineffective in finding the transcription factors candidates in the first instance. There are thousands of TFs discovered and the list is still growing. In vivo TF analyses are also time consuming and expensive. So the best approach is to use computational methods to narrow down the list of TFs and/or find a novel TF. However, it is necessary for these TFs to be then verified as the computational methods have high false positive rates.

Clover and Rover were used to implement a comparative solution to find TFs which are significantly enriched in HERVHs and/or LTR7s (Haverty et al. 2004), however, both packages depend on a list of TF motifs to be able to conduct TF analysis on the sequences of interest. For this purpose I have used Jaspar database which is a relatively comprehensive database and one of the few openly accessible TF databases (Bryne et al. 2008). But Jaspar does not include all the TFs discovered so far. To cover for this shortcoming, I later complemented my TF analyses by probing other resources, including ENCODE and hmChIP databases explained in details below.

Clover and Rover Analysis

To find the motifs found in the active HERVHs, Clover was used to compare Jaspar motifs enriched in the active HERV-Hs against GC matched background sequences. This allows us to ask which TFs, within the Jaspar dataset, have more TF binding sites than expected by chance in GC matched control sequences. In addition, we ask using Rover which motifs might be significantly enriched in the active HERV-Hs compared with those with LTR7C/Y, which are slightly less active members of LTR7. I have also used Rover to find which TFs are enriched in active LTR7 sequences in comparison to inactive HERV-Hs. The analyses where then compiled together to find the set of TFs found significant across all, table 3. The only TFs found to be significant across all analyses is MA0145.1 Tcfcp2l1, also known as LBP9.

ENCODE and hmChIP

ENCODE and hmChIP databases include chip-seq datasets with much lower false positive rate compared to the computation approach presented above. However, they include far fewer TFs, so the list provided by them is by no means comprehensive.

For ENCODE analysis, second release of Txn Factor ChIP was downloaded through ENCODE's portal (http://genome.ucsc.edu/ENCODE/downloads.html). This data set is the result of collaboration between Myers Lab at the HudsonAlpha Institute for Biotechnology and the labs of Michael Snyder, Mark Gerstein and Sherman Weissman at Yale University; Peggy Farnham at UC Davis; and Kevin Struhl at Harvard. Kevin White at The University of Chicago. Vishy Iyer at The University of Texas Austin (Bernstein et al. 2012). I then used bedtools to process these files to find which transcription factors are enriched in which sequences. These three sequence sets were processed in this fashion: 1- manually assembled longer HERV-H sequence, 2- extended

manually assembled longer HERV-H sequence and 3- 1090 highly expressed HERV-Hs. The number of transcription factor binding sites found in each set is shown in table 4.

Table 3. Motifs found significantly enriched in active sequences of interest by Clover and Rover analyses of transcription factors are listed.

Jaspar ID / Motif	Clover HERV-H v GC matched (Raw score)	Clover HERV-H v GC matched (P-value)	Rover HERV-h v LTR7C/Y (P-value)	Rover LTR7 v inactive HERV-H (P-value)
MA0145.1_Tcfcp2l1 (LBP9)	111	0.001	8.30E-36	0.000576175
MA0035.2_Gata1	96.4	0	2.40E-27	
MA0109.1_Hltf	86.2	0	1.17E-18	
MA0063.1_Nkx2-5	-3.97	0.002	9.80E-15	0.000265626
MA0259.1_HIF1A::ARNT	-3.68	0	1.47E-14	
MA0047.2_Foxa2	365	0	1.43E-08	
MA0148.1_FOXA1	416	0	8.38E-07	
MA0101.1_REL	184	0	0.000411012	
MA0143.1_Sox2	64.9	0.001	0.00150613	
MA0055.1_Myf	285	0	0.00229056	
MA0140.1_Tal1::Gata1	175	0	0.00700998	
MA0144.1_Stat3	-4.94	1	5.80E-83	0.000910078
MA0066.1_PPARG	-6.61	1	9.08E-75	9.99431e-06
MA0007.1_Ar	-5.98	0.989	3.37E-45	0.000755461
MA0158.1_HOXA5	-4.36	1	2.34E-08	
MA0135.1_Lhx3	6.79	0.999	1.17E-07	
MA0137.2_STAT1	11.3	0.994	2.62E-06	
MA0113.1_NR3C1	-4.76	1	6.50E-06	
MA0048.1_NHLH1	-4.03	1	7.14E-06	
MA0163.1_PLAG1	-6.04	1	4.51E-05	
MA0099.2_AP1	50.9	0.982	6.82E-05	
MA0025.1_NFIL3	-7.09	1	0.000107307	
MA0078.1_Sox17	-4.43	1	0.000281809	
MA0030.1_FOXF2	-5.5	1	0.000616575	
MA0139.1_CTCF	-7.09	1	0.00160586	
MA0040.1_Foxq1	-4.45	1	0.00233706	
MA0014.1_Pax5	-3.79	1	0.00235216	
MA0102.2_CEBPA	-3.99	0.958	0.00535611	

I have also proceeded with analyzing the manually assembled active sequences' transcription factors profile in hmChIP database (Chen et al. 2011). As mentioned above, hmChIP is a database of genome-wide chromatin immunoprecipitation (ChIP) data in human and mouse, including 2016 samples from 492 ChIP-seq and ChIP-chip experiments. hmChIP also lists some of the older data releases from the labs involved in ENCODE project. The latest version of hmChIP is based on used hg18, so liftover was used to convert our sequences, which were based on hg19 annotation, to hg18.

Conversion was successful except for one sequence which could not be done due to a split introduced in hg19. The top hits for our manually assembled sequences in hmChIP database included P300, Pol2-4H8, Pol2 and TAF1 which were all found in ENCODE analyses too. However, at the end we decided to discard results from hmChIP's TF analyses due to a number of caveats concerning a large number of these hits being based on old datasets, whose newer releases were accessible through ENCODE portal. These new releases were anyway included in ENCODE analysis reported above.

Table 4. The number of transcription factor binding sites of the transcription factor of interest in ENCODE is shown below.

TF	Count in assembled HERV-H seqs	Count in extended assembled HERV-H seqs	Counts in highly expressed seqs
Pol2-4H8	118	140	199
Pol2	104	128	245
CTCF	41	45	78
TAF7_(SQ-8)	40	53	56
TAF1	35	39	46
NANOG_(SC-33759)	27	33	41
CEBPB	25	33	50
MafK_(ab50322)	22	28	43

Conclusion

In previous chapters, I have shown evidence regarding the effects of evolution of coding and non-coding sequences on the expression of their neighbours. In this chapter, I used Human specific endogenous retroviruses, HERVs, to provide an excellent base as a naturally occurring transgene experiment to examine how these randomly scattered sequences might affect the expression profile of their neighboring genes, in the context of stem cells. In collaboration with Prof. Zsuzsanna Izvak, I have shown actively transcribed members of a special class of HERVs, HERV-H, are involved in regulating their neighbours. Above I have presented a summary on evidence of open chromatin DHS peaks and activating histone marks on and in vicinity of active HERV-Hs and their long terminal repeats, LTR7. TF analyses were especially instrumental in guiding wet lab team to characterize role of a novel transcription factor, LBP9, in rewiring pluripotency network in stem cells. I also helped with analyzing chimeric transcripts. HERV-Hs not only create new genes but also affect splicing of the genes. All of this is explained in more detail in the resulting paper published in Nature, which is attached below.

References

Bernstein BE, Birney, E, Dunham, I, Green, ED, Gunter, C, Snyder, M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489:57-74.

Bernstein BE, Mikkelsen, TS, Xie, X, Kamal, M, Huebert, DJ, Cuff, J, Fry, B, Meissner, A, Wernig, M, Plath, K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 125:315-326.

Bock C. 2012. Analysing and interpreting DNA methylation data. Nat Rev Genet. 13:705-719.

Bryne JC, Valen, E, Tang, MH, Marstrand, T, Winther, O, da Piedade, I, Krogh, A, Lenhard, B, Sandelin, A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36:D102-106.

Cedar H, Bergman, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet. 10:295-304.

Chen L, Wu, G, Ji, HK. 2011. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. Bioinformatics. 27:1447-1448.

Greer EL, Shi, Y. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. Nat Rev Genet. 13:343-357.

Haverty PM, Hansen, U, Weng, Z. 2004. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res. 32:179-188.

Marfella CG, Imbalzano, AN. 2007. The Chd family of chromatin remodelers. Mutat Res. 618:30-40.

Piatti P, Lim, CY, Nat, R, Villunger, A, Geley, S, Shue, YT, Soratroi, C, Moser, M, Lusser, A. 2015. Embryonic stem cell differentiation requires full length Chd1. Sci Rep. 5:8007.

Siggens L, Cordeddu, L, Ronnerblad, M, Lennartsson, A, Ekwall, K. 2015. Transcription-coupled recruitment of human CHD1 and CHD2 influences chromatin accessibility and histone H3 and H3.3 occupancy at active chromatin regions. Epigenetics Chromatin. 8:4.

Sims RJ, 3rd, Millhouse, S, Chen, CF, Lewis, BA, Erdjument-Bromage, H, Tempst, P, Manley, JL, Reinberg, D. 2007. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 28:665-676.

Soejima. 2010. Distinct epigenetic regulation of tumor suppressor genes in putative cancer stem cells of solid tumors. International Journal of Oncology. 37.

Thurman RE, Rynes, E, Humbert, R, Vierstra, J, Maurano, MT, Haugen, E, Sheffield, NC, Stergachis, AB, Wang, H, Vernot, B, et al. 2012. The accessible chromatin landscape of the human genome. Nature. 489:75-82.

Warner MH, Roinick, KL, Arndt, KM. 2007. Rtf1 is a multifunctional component of the Paf1 complex that regulates gene expression by directing cotranscriptional histone modification. Mol Cell Biol. 27:6103-6115.

Zentner GE, Tsukiyama, T, Henikoff, S. 2013. ISWI and CHD chromatin remodelers bind promoters but act in gene bodies. PLoS Genet. 9:e1003317.

This paper has been removed due to copyright concerns.

The paper in question is:

Wang, J, Xie, G, Singh, M, Ghanbarian, AT, Raskó, T, Szvetnik, A, Cai, H, Besser, D, Prigione, A, Fuchs, NV, Schumann, GG, Chen, W, Lorincz, MC, Ivics, Z, Hurst, LD & Izsvák, Z 2014, 'Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells' *Nature*, vol 516, no. 7531, pp. 405-409

It can be found here: http://dx.doi.org/10.1038/nature13804

Chapter 6. Discussion

In the results chapters presented so far, I have shown that the change in gene expression of a focal gene on average predicts the change in gene expression of neighbors in six human tissues and both sexes. The effect is highly pronounced in the immediate vicinity but extends much further. Sex-specific expression change is also genomically clustered. As genes increasing their expression in humans tend to avoid nuclear lamina domains and be enriched for the gene activator 5-hydroxymethylcytosine, chromatin level mechanisms are probably involved in regulating this phenomenon. The phenomenon of correlation in change in gene expression of the neighbouring genes is termed expression piggy-backing, an analog of hitchhiking. I have also shown evidence of piggybacking in the compact genome of yeast. However, the size of co-evolving clusters found in yeast was much smaller than that observed in human.

I have also investigated whether non-coding genes might initiate a similar regulation of neighbours. To this end I conducted a follow-up study in lincRNAs. Finding most selection on lincRNA to be splice related, I have shown intron-rich lincRNAs to be enriched in CHD1's, an splice-related chromatin remodeller binding sites. I also found intron rich lincRNAs to be enriched in DHS peaks, a marker for open chromatin. And both CHD1 and DHS were found to correlate with expression of the neighbours, the effect possibly modulated through intron density effects. In other words, lincRNAs were found to be regulating expression of their neighbouring genes through providing binding site for a splice-related chromatin remodeller, CHD1. I then studied HERVs, as a naturally occurring transgene experiment, to investigate how randomly scattered similar sequences might affect the expression profile of their neighboring genes.

The publications resulting from my PhD are all reported in this thesis. They include 2 papers in *Molecular Biology and Evolution*, a paper in *Nature* and another recently submitted to *Journal of Molecular Evolution*. Each paper includes an in depth conclusion and discussion. As a consequence, here I will just point out the future work. These include a few interesting questions which the limited time and resources

did not allow me to investigate during my PhD. Nevertheless, they are important questions in context of evolution of gene expression.

In my MBE paper "Neighboring Genes Show Correlated Evolution in Gene Expression", I have presented evidence for co-evolving gene expression clusters across 6 tissues, analysed clusters of genes exhibiting between–tissue concordances and also established sex-biased gene expression clusters. However, I have just scratched the surface on characterising co-evolving gene expression clusters. Knowing evolution of gene expression happens in clusters expands the horizon to study variations in gene expression in both healthy and affected individuals. One could ask how disease-associated variants relate to these clusters and whether these co-evolving gene expression clusters could explain why some diseases prevailed throughout populations even after long periods of purifying selection. These clusters might also explain why many gene therapy efforts had unexpected outcomes and many have failed (Schneider et al. 2010; Kay 2011).

In the first couple of results chapter, I have shown evidence for correlation in gene expression in Primates and Yeasts. Although we found evidence for piggybacking in both genomes, the boundary of correlation was drastically different. While numerous co-evolving gene expression clusters were found to expand over large segments of the genome in human, the size of clusters are far smaller in yeast and usually include only very close neighbours. In comparison to the large genome of human, yeast has a compact genome with short intergenic regions that might necessitate greater insultation. A compact genome might also increase the chance of transcriptional interference, hence would bar expansion of correlated gene expression domains. In turn does this render the yeast genome more evolutionary constrained in exploring possible expression space, as change in one gene affects relatively few others. A compact genome might also be limiting in that longer intergenic regions in theory could increase diversity of regulatory elements which are necessary to evolve a more complex gene expression profile.

But could we conclude that the size of co-evolving gene expression clusters are dictated by the average size of intergenic regions in any particular organism? The results presented here in two genomes are too limited to answer this question. To find the answer, one shall study piggybacking in several other genomes across varied

average intergenic regions size to investigate average size of co-evolving gene expression clusters as a function of average size of intergenic regions.

On the other hand, while piggybacking might facilitate evolution of complex gene expression profiles, at least in Primates and yeasts, not much is known about how this phenomena itself came to existence. In other words, how has piggybacking expression itself evolved? Was there a benefit to the correlation in gene expression when it first evolved or is it best seen as a deleterious trait that needs to be neutralized as much as feasible? It need not be that one or the other need be true. It could be that modulated change in expression is beneficial for some gene clusters, disadvantageous for others.

Mechanistically, I presumed that piggybacking emerged with the same mechanism through all phylogenies. This need not be true and it might have developed several times in parallel? These questions and many others regarding the evolution of piggybacking phenomena itself cannot be answered by limited results provided here. A through examination of piggybacking in several other organisms across different kingdom of life is necessary to shed light on how this phenomena itself was evolved.

One could also ask how evolution of novel genes correlate with evolution of its neighboring genes' expression? Do novel genes more often emerge in a particular type of co-evolving gene expression clusters? Do they popup in shorter clusters or longer ones? Above I have shown up-regulated clusters are denser compared to the down-regulated ones. So are up-regulated gene clusters functioning as hotspots for novel genes to emerge? In our *Nature* paper, we have shown how HERV-Hs provides a source of novelty in our genome through creating chimeric transcripts. We have shown HERV-Hs not only create new genes but also affect splicing of their neighboring genes. So if one is to make a list of novel genes created through retrovirus insertions, through mutations creating new ORFs in intergenic regions or through other means established in literature (Long et al. 2003; Hoekstra and Coyne 2007; Innan and Kondrashov 2010), do we see any preference for novel genes to emerge in any specific type of clusters? Also how does expression profile of these novel genes differ from that of their neighbours? Do they always adopt a similar expression profile or may they function as an insulator and break clusters? The answer to these questions could have direct implications in gene therapy.

This novel gene analysis could be conducted differently by studying age of genes in relationship to the characteristics of co-evolving gene expression cluster they belong to. In this context, the first question to ask would be: are clusters generally homogeneous in terms of the age of the genes they encompass? Do younger genes often appear on cluster edges? Are down-regulated clusters populated with older genes compared to up-regulated clusters? This analysis might also help to understand how larger clusters came to existence.

Here, I have not presented results of my preliminary study of insulators. My first analysis of CTCF binding sites across up-regulated and down-regulated clusters did not find any significant difference between the two types of cluster. However, there are many other insulators one could study to probe the differences in insulation mechanisms across up-regulated and down-regulated clusters (Guelen et al. 2008; Symmons et al. 2014). One could also investigate the difference in pattern of insulation across small and large clusters. Not all insulation mechanisms and components involved are yet known, nonetheless, the topic is attracting more attention and novel mechanisms and components are being discovered. Hence, it would be interesting to revisit the role of insulators in defining boundaries of co-evolving gene expression clusters.

I have shown evidence for chromatin level regulation of evolution of gene expression. Although, these mechanisms can explain correlation in scales of hundred kilo base pairs, as ripple effect suggests, they overreach their limits to explain the observed cluster sizes. It was shown in chapter 2 that the ripple effect alone cannot explain the domain of influence observed in Primates. Nevertheless, astounding clusters expanding a few mega base pairs calls for novel mechanisms perhaps yet to be discovered or ground breaking clarification on actual boundaries of currently known mechanisms. Could newly established chromosomal looping concept be able to explain how large clusters of co-evolving gene are maintained (Rao et al. 2014)? Could 3D structure of chromosomes be able to shed light on obscurities in regulation of these large clusters? It needs to be mentioned that the current methods applied to develop 3D models of chromosomal structure still suffer from high false positives and quality and accuracy of the resulting models have been frequently questioned (Jin et al. 2013). Perhaps in near future a novel method would lift this limit and then we

could investigate large co-evolving gene clusters encompassing 3D dynamics to build a more accurate model of evolution of gene expression.

Last but not least, in chapter 4, I have shown evidence for lincRNAs being involved in regulating expression of their neighbouring genes through providing binding site for an splice-related chromatin remodeller, CHD1. But I have not checked if this only happens in lincRNAs or whether a similar pattern is observed in other classes of non-coding RNAs. In other words, do other classes of non-coding RNAs adopt a similar mechanism and are involved in regulating their neighboring coding genes? By the method presented here, one could investigate this further.

So with this I would like to conclude this thesis and thank you for your attention.

References

Guelen L, Pagie, L, Brasset, E, Meuleman, W, Faza, M, Talhout, W, Eussen, B, de Klein, A, Wessels, L, de Laat, W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 453:948-951.

Hoekstra HE, Coyne, JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 61:995-1016.

Innan H, Kondrashov, F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:97-108.

Jin F, Li, Y, Dixon, JR, Selvaraj, S, Ye, Z, Lee, AY, Yen, CA, Schmitt, AD, Espinoza, CA, Ren, B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 503:290-294.

Kay MA. 2011. State-of-the-art gene-based therapies: the road ahead. Nat Rev Genet. 12:316-328.

Long M, Betran, E, Thornton, K, Wang, W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865-875.

Rao SS, Huntley, MH, Durand, NC, Stamenova, EK, Bochkov, ID, Robinson, JT, Sanborn, AL, Machol, I, Omer, AD, Lander, ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 159:1665-1680.

Schneider CK, Salmikangas, P, Jilma, B, Flamion, B, Todorova, LR, Paphitou, A, Haunerova, I, Maimets, T, Trouvin, JH, Flory, E, et al. 2010. Challenges with advanced therapy medicinal products and how to meet them. Nat Rev Drug Discov. 9:195-201.

Symmons O, Uslu, VV, Tsujimura, T, Ruf, S, Nassari, S, Schwarzer, W, Ettwiller, L, Spitz, F. 2014. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 24:390-400.