Supplementary Material

for

The impact of genomic neighbourhood on human and chimpanzee transcriptome evolution

[Supplemental material is available online at http://www.genome.org and at http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/]

SI-1: Consistency in CGN score calculation

Correlation between CGN score for human genes using different approaches

CGN score for human genes (with chimpanzee as reference species) was calculated using several different but related approaches: (a) those that fixed the window size at various values, i.e. *w* = 1Mb and 3Mb and (b) those that fixed the number of neighbouring genes around the gene of interest at specific values, i.e. N_A^{HS} = 30 genes and 40 genes (see **Figure 2a** in the main text).

1a. CGN calculated using a window size *w* of 1 Mb and 3 Mb instead of 2Mb, centering on the gene of interest. The calculated CGN score ($CGN_{1Mb\ and}$ CGN_{3Mb}) was compared against that calculated using a 2Mb window (CGN_{2Mb}) as described in the manuscript. A high correlation (Corr-coeff > 0.8) was observed in both cases showing that the CGN scores are comparable.

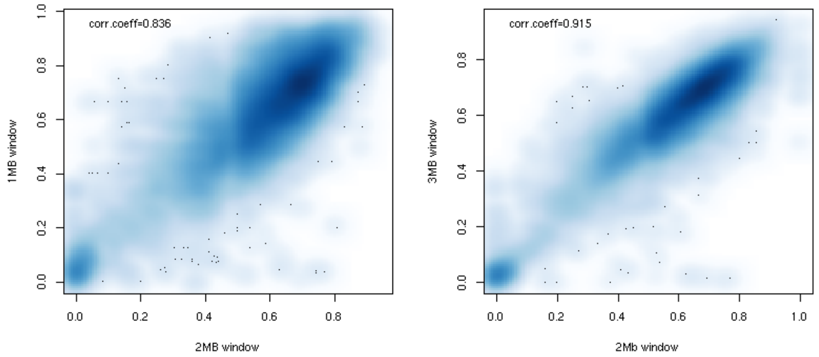


Figure SF1: Correlation between CGN_{2Mb} and CGN_{1MB} (left) and CGN_{3MB} (right). Intensity of blue colour reflects the density of data-points. Correlation coefficient values are 0.836 (for CGN_{1Mb}) and 0.915 (CGN_{3Mb}) showing that the CGN scores are comparable and generally robust to the choice of our definition used.

1b. CGN score calculated after fixing the number of neighbours N_A^{HS} at 30 genes and 40 genes, centering on the gene of interest. The calculated CGN scores were compared against that calculated using a 2Mb window (CGN_{2Mb}) as described in the manuscript. A high correlation (Corr-coeff > 0.78) was observed in both cases showing that the CGN scores are comparable.

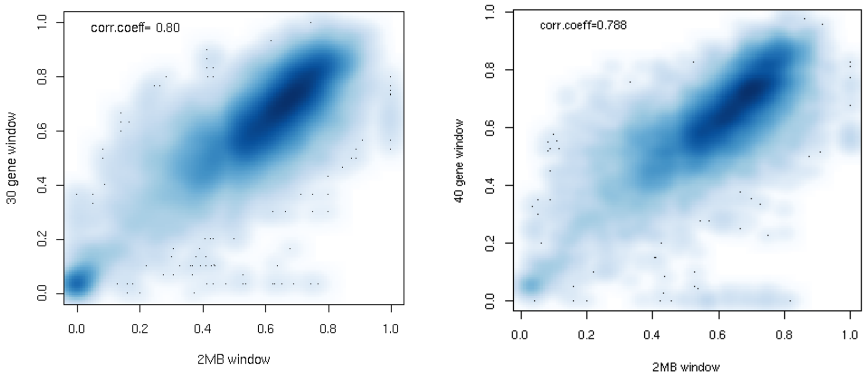
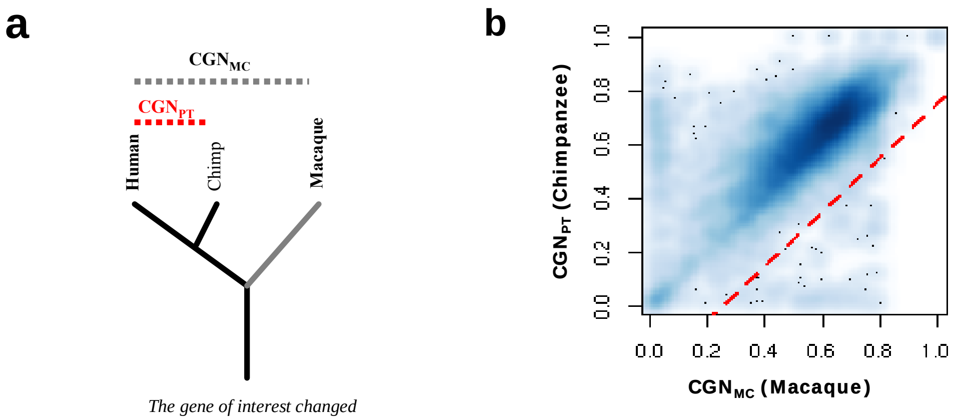


Figure SF2: Correlation between CGN_{2Mb} v/s $CGN_{30\ genes}$ (left) and CGN_{2Mb} v/s $CGN_{40\ genes}$ (right) is shown. Intensity of blue colour reflects the density of data-points. Correlation coefficient values are 0.80 ($CGN_{30\ genes}$) and 0.79 ($CGN_{40\ genes}$) suggesting that the CGN scores are comparable and generally robust to the choice of our definition used.

SI-2: Quality control in CGN score calculation

Complications in orthology-identification or genome assembly can introduce a small number of false positives in determining correct ortholog pairs and their genomic locations. To address this problem, we compared the CGN score of a human gene by using chimpanzee (denoted as CGN_{PT}) and macaque (denoted as CGN_{MC}) as reference species. Usually, if the human gene shows a change in genomic neighbourhood with respect to a reference species (e.g. chimpanzee) due to an evolutionary event after the split with the common ancestor, or vice versa (**Figure SF3a**), CGN score (*e.g.* CGN_{PT}) is expected to be small. Since the evolutionary distance between human and macaque is longer than that between human and chimpanzees, it is expected that there will be more alteration in genomic neighbourhood between human and macaque, as compared to that between human and chimpanzee, i.e. usually $CGN_{PT} > CGN_{MC.}$

However, there can be small variations in CGN score when the number of orthologous gene-pairs around a gene of interest between human and chimpanzee is different (usually greater) than that between, human and macaque (*i.e.*, $CGN_{PT} + \delta = CGN_{MC}$, where δ is small). However, if CGN_{MC} is considerably greater than CGN_{PT}, that can indicate (i) a possible problem with mis-prediction of ortholog or its chromosomal location or (ii) a chimpanzee-specific change in genomic neighbourhood. Since it is difficult to differentiate between the two possibilities with confidence, we decided to exclude all such cases. To identify such cases, we plotted the correlation between CGN_{PT} and CGN_{MC} for all human genes (**Figure SF3**). After extensive manual inspection, we accepted the CGN score for the genes that satisfy the following condition: $CGN_{PT} > CGN_{MC} - 0.25$, *i.e.*, $\delta = 0.25$. This resulted in the identification of <1% of the orthologous human genes and were excluded from further analysis. Although the 1% of the genes which we reported as either problematic or chimp-specific is true, the 99% of the cases which we analyze should not be treated as human specific changes only. Because of the way in which CGN score is calculated, the identification of genes with low CGN would include both the set of genes that have changed their neighbourhood either in the human lineage or the chimpanzee lineage after the split from their common ancestor. By using outgroup information, we identifed the number of genes that have been altered exclusively in one of the two lineages and find that roughly equal number of genes have altered their neighbourhood in a lineage specific manner (data not shown).



The gene of interest changed its neighbourhood either in human or in chimpanzee

For some genes, CGN_{MC} was much higher that $CGN_{PT,}$ pointing to a potential problem in genome assembly. Such genes (< 1% of all genes) were removed from subsequent analysis

Figure SF3: CGN_{PT} is expected to be greater or equal to CGN_{MC} (**a**). A comparison of the values shows that this is indeed the case for a majority of the genes (**b**). Intensity of blue colour reflects the density of data-points. The genes on the right side of the dotted red line with CGN_{MC} values much greater than CGN_{PT} (*i.e.*, $CGN_{MC} \geq CGN_{PT} + 0.25$) were identified and excluded from further analysis to avoid potentially spurious cases of genes with altered neighbourhood due to genome assembly errors or problems in orthology detection.

SI-3: Data on genes, chromosomal position and their CGN scores

Of the predicted 31,986 genes in humans and 25,466 genes in chimpanzee, we obtained unambiguous ortholog mapping from Ensembl-Compara for 19,256 genes. For each of the 19,256 human genes, we calculated the CGN score using chimpanzee as a reference species. To ensure that errors in genome assembly do no affect the calculation of CGN score, we used the macaque genome sequence as an out-group reference species to remove all spurious instances of genes with altered neighbourhood (*see* **SI-2** *above*). We further removed all genes that are located in incompletely assembled genomic regions and chromosomal bands nearest to telomeres and centromeres, which often experience sequencing errors and structural variations. We also decided to exclude one-to-many, many-to-many and many-to-one orthologs (~4.5%) and focussed only on one-to-one orthologs. This resulted in 16,868 genes with CGN scores. The entire list of genes and the CGN scores are provided as an excel file and a tab-delimited text file from the supplementary URL:
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS1.xls
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS1.txt

Table ST1: The description of the columns in the file is given below

The first column provides the Ensembl gene identifier for human genes (Ensembl v.48)
The second column provides the HGNC gene name
The third column provides the chromosome number and chromosomal position of the human gene
The fourth column provides the Ensembl gene identifier for the chimp ortholog (Ensembl v.48)
The fifth column provides the chromosome number and chromosomal position of the chimp gene
The sixth column provides the no. of neighbouring genes within a 2Mb window around the human gene (N_A^{HS})
The seventh column provides the no. of neighbouring genes within a 2Mb window around the chimp ortholog (N_A^{PT})
The eighth column shows the number of orthologs (N_A) that are common between N_A^{HS} and N_A^{PT} for that gene
The ninth column provides CGN score of the human genes calculated using chimpanzee as reference species (CGN_{2Mb})

SI-4: Chromosomal distribution analysis

We analysed the distribution of genes with low CGN on the different human chromosomes by comparing the distribution of CGN scores of genes from every chromosome to the genomic distribution (median = 0.637) using Mann-Whitney test (two-tailed; *see* **Table ST2**, **Figure SF4**). We note that Chr-19, Chr-17 and Chr-11 were the ones which were enriched in genes with high CGN score. It is known that a significant part of the human chromosome 17 maps to a single large segment on Chromosome 11 in mouse, suggesting high conservation of genomic neighbourhood of genes encoded in this chromosome during mammalian evolution. On the other hand, apart from the sex chromosomes, we find that chromosomes Chr-18, Chr-7 and Chr-13 show enrichment for genes with low CGN.

Chromosome	Number of genes with CGN score	Median CGN score	P value
Chr19	1016	0.714	<1.0E-10
Chr11	1037	0.68	<1.0E-10
Chr17	943	0.672	<1.0E-10
Chr12	917	0.659	<1.0E-10
Chr16	566	0.659	-
Chr9	720	0.647	-
Chr3	1003	0.645	-
Chr21	226	0.6395	-
Chr6	932	0.636	-
Chr8	580	0.636	-
Chr1	1759	0.636	-
Chr20	464	0.636	-
Chr5	771	0.623	-
Chr22	419	0.62	-
Chr2	1136	0.615	-
Chr10	674	0.613	-
Chr4	673	0.607	-
Chr15	528	0.603	-
Chr14	571	0.6	-
ChrX	607	0.571	<1.0E-10
Chr13	270	0.565	<1.0E-10
Chr7	780	0.561	<1.0E-10
Chr18	248	0.555	<1.0E-6
ChrY	28	0.277	<1.0E-6

Table ST2: Chromosomal enrichment for high and low CGN scoress

Even though we find genes with low CGN to be located near centromeres and telomeres, we did not attempt to quantify clustering patterns for several reasons. (i) We wish to note that the genes near telomeres and centromeres may prove to be interesting as such regions are unusual in terms of their evolutionary history or could be prone to assembly errors. (ii) It is difficult to isolate cases affected by assembly errors with confidence and hope that the availability of better quality data will shed more light on the biological implication of the real cases. To remove potentially spurious cases that may affect our analysis, we have removed all the genes that are located in the band nearest to the telomeres and centromeres.

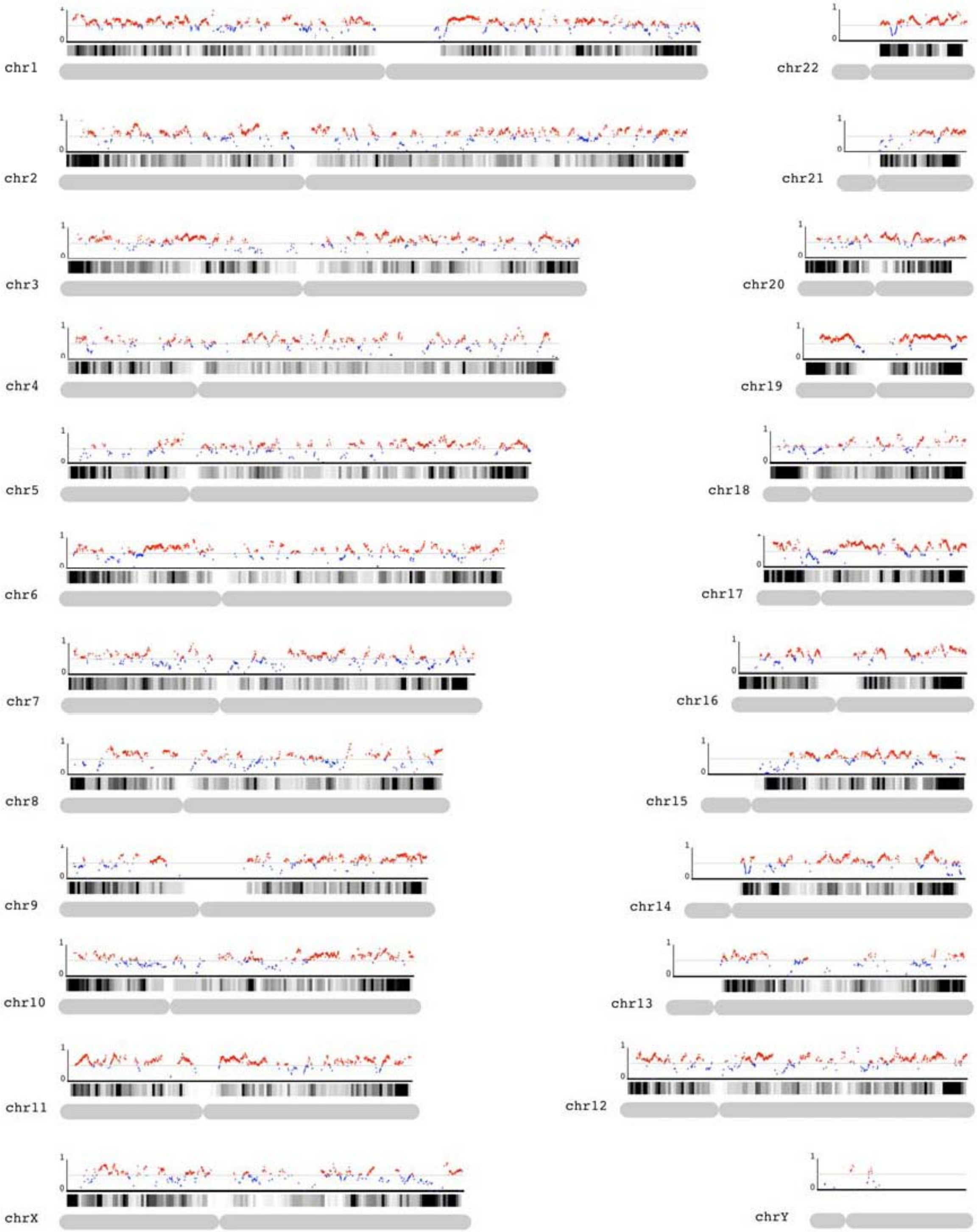
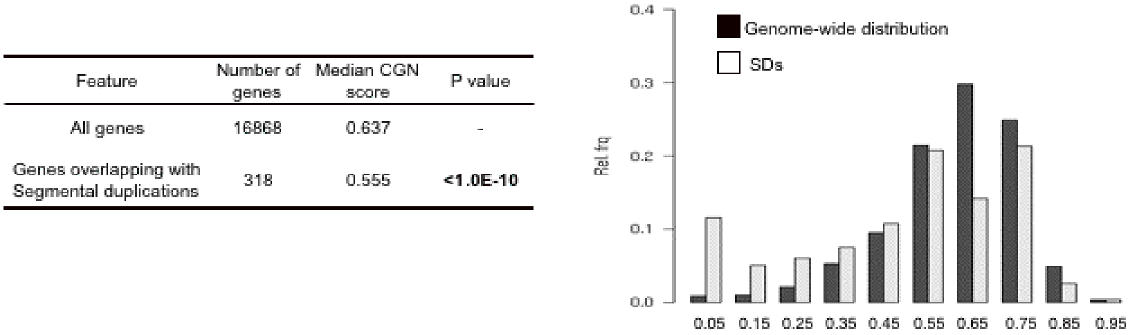


Figure SF4: Chromosomal distribution of CGN scores for human genes. Human chromosomes are shown in gray at the bottom. For each chromosome, distribution of local recombination rate is shown in the middle so that intensity of black colour of the bands indicates magnitude of local recombination rate. At the top, CGN scores of genes are shown in red (CGN score>0.5) and blue (CGN score ≤0.5).

SI-5: Structural variation analysis

Several of the structural variations are yet to be fixed in human and chimpanzee populations. Therefore, we wanted to assess whether our findings could be influenced depending on whether such variations have been fixed or not.

Overlap with Segmental duplications: We obtained the positions of human Segmental Duplications from the Segmental Duplication Database (http://humanparalogy.gs.washington.edu/). SDs were detected using Whole Genome Shotgun Sequence Detection (WSSD). We identified the 318 genes that are located in SD regions and also had CGN scores calculated. As expected, genes that reside on SDs are more likely to have undergone a change in genomic neighbourhood, and have significantly low CGN score distribution as compared to that of all genes in the human genome (p value <1.0E-10). The results are summarised in **Figure SF5**. We did not detect any correlation (correlation coefficient: -0.078) between the CGN score of a gene and the amount of genetic material introduced in its 2Mb neighbourhood due to segmental duplication. This is because majority of the duplicated regions are small (median = 3882 bp), and it is likely that the functional implication of insertion or deletion of genetic material around a gene is context dependent.



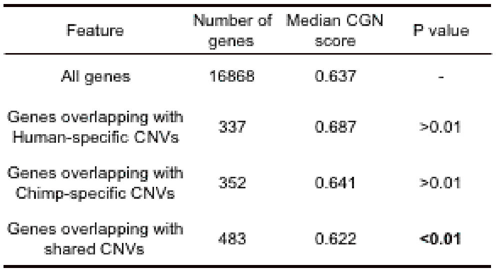
Feature	Number of genes	Median CGN score	P value
All genes	16868	0.637	-
Genes overlapping with Segmental duplications	318	0.555	**<1.0E-10**

Figure SF5: (A) Table showing the extent of overlap of segmental duplications with genes that have CGN score calculated. Distribution of CGN scores for the genes that overlap with SDs is compared against distribution of CGN scores of all the genes in the human genome using Mann Whitney test (Alternative hypothesis: $CGN_{genes\ in\ SD} < CGN_{all\ genes}$) and p value is reported. (B) Histogram showing distribution of CGN scores for all genes (black) and those overlap with SDs (white).

Overlap with Insertions and deletions: Positions for insertions and deletions (InDels) in the human populations were obtained from the Database of Genomic Variants (http://projects.tcag.ca/variation/). In this database, InDels were culled from 8 published studies (Iafrate et al. Nat Gen. 2004), and contained 9659 InDels several of which overlap between studies. The vast majority of the insertions and deletions in that dataset were small (median = 162 bp), and did not encompass entire protein coding genes. Therefore, we focussed on the extent of insertion or deletion of genetic material in the 2Mb neighbourhood region of a gene of interest. On an average, ~1% genetic material (median = 1230 bp, maximum = 18067 bp) was changed in the 2Mb neighbourhood of a gene due to insertion or deletion. We did not detect any correlation (correlation coefficient: 0.009) between the CGN score of a gene and the amount of genetic material changed in its 2Mb neighbourhood due to insertion or deletion.

Overlap with CNVs in human and chimpanzees: Copy number variations specific to human and chimpanzee lineages and those shared between the two species are obtained from Perry et al. (Perry et al. 2008). The authors detected 791 CNVs in human and chimpanzee using CGH on a single human microarray platform and identified those specific to human, specific to chimpanzee and those shared between the species.

We observed that 337, 352 and 483 genes that overlap with human-specific CNVs, Chimpanzee specific CNVs and shared CNVs and also have their CGN score calculated. We compared the CGN score of genes in each group against the distribution of CGN score for all genes in the human genome using Mann-Whitney test. The results are summarised in **Figure SF6**. We found that the genes, which overlap with CNVs shared between human and chimpanzee populations have significantly low CGN score as compared to the genome-wide background. But the number of genes that have altered their neighbourhood (CGN score ≤0.5) and also overlap with shared CNVs is small (151 genes), and our main conclusions in this paper remain unchanged even after excluding those cases (data not shown).



Feature	Number of genes	Median CGN score	P value
All genes	16868	0.637	-
Genes overlapping with Human-specific CNVs	337	0.687	>0.01
Genes overlapping with Chimp-specific CNVs	352	0.641	>0.01
Genes overlapping with shared CNVs	483	0.622	**<0.01**

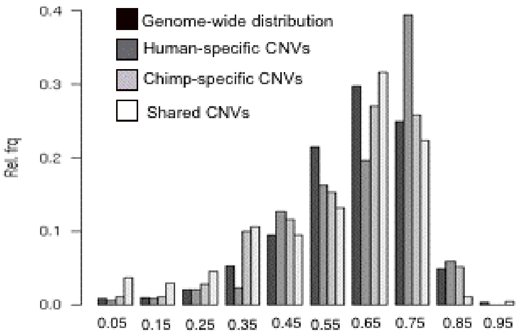
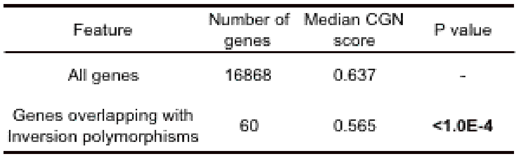


Figure SF6: (A) Table showing the extent of overlap of human-specific, chimpanzee specific and shared CNVs with genes that have CGN score calculated. Distribution of CGN scores for the genes that overlap with CNVs is compared against distribution of CGN scores of all the genes in the human genome using the Mann-Whitney test (Alternative hypothesis: $CGN_{genes\ with\ CNVs} < CGN_{all\ genes}$) and p-value is reported. (B) Histogram showing distribution of CGN scores are shown for the genes that overlap with human-specific (dark grey), chimpanzee specific (light grey), shared (white) CNVs and for for all genes (black).

Overlap with Inversion polymorphisms in human and chimpanzees: Inversion polymorphisms in the human HapMap populations are obtained from Bansal et al. (Bansal et al. 2007). The authors detected inversion polymorphism associated with 171 Refseq and/or Ensembl (Genbank) genes using LD analysis of HapMap data. From that list, 105 genes also had their CGN scores available. We compared the CGN score of these genes against the distribution of CGN score for all genes in the human genome using the Mann-Whitney test. The results are summarised in **Figure SF7**. We found that the genes, which overlap with inversion polymorphisms in human HapMap populations have significantly low CGN score as compared to the genome-wide background. But the number of genes that have altered their neighbourhood (CGN score ≤0.5) and also overlap with inversion polymorphism is small (33 genes), and our main conclusions in this paper remain unchanged even after excluding those cases (data not shown).



Feature	Number of genes	Median CGN score	P value
All genes	16868	0.637	-
Genes overlapping with Inversion polymorphisms	60	0.565	**<1.0E-4**

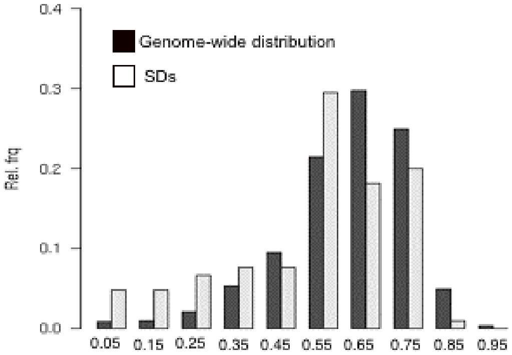


Figure SF7: (A) Table showing the extent of overlap of inversion polymorphisms with genes that have CGN score calculated. Distribution of CGN scores for the genes that overlap with inversions is compared against distribution of CGN scores of all the genes in the human genome using the Mann-Whitney test (Alternative hypothesis: $CGN_{genes\ in\ inversion} < CGN_{all\ genes}$) and p-value is reported. (B) Histogram showing distribution of CGN scores for all genes (black) and those overlap with regions containing inversions (white).

Taken together, the results from the above three analyses suggest that while several structural rearrangements gave rise to alteration of genomic neighbourhood, they are probably yet to be fixed in human (and/or in chimpanzee) populations. Although our preliminary analysis shows that effect of structural polymorphisms is small, high resolution structural variation study involving large sample size will be necessary to understand the proportion of alteration of genomic neighbourhood.

SI-6: Genomic neighborhood and expression divergence

Data on gene expression divergence was obtained from Khaitovich *et al.* [30]. Expression divergence values were available for 9,248 orthologous genes in at least one of the five tissues (brain, heart, kidney, liver and testis) and were defined by the authors as an average squared difference in normalized gene expression across all probes with detectable gene expression between species. A gene was defined as showing high expression divergence if its expression divergence value was greater than the median value of all genes for that tissue.

Table ST3: Conditional probability values for genes with conserved neighbourhood (H-CGN).

High Expression Divergence (HED)

	n(H-CGN)	n(HED)	n(HED ∩ H-CGN)	p(HED|H-CGN)	p(H-CGN|HED)
Brain	4474	2807	2185	0.488	0.778
Heart	4264	2680	2099	0.492	0.783
Kidney	4792	3004	2393	0.499	0.797
Liver	4327	2696	2127	0.492	0.789
Testis	5520	3420	2721	0.493	0.796
mean				**0.493**	**0.789**

Low Expression Divergence (LED)

	n(H-CGN)	n(LED)	n(LED ∩ H-CGN)	p(LED|H-CGN)	p(H-CGN|LED)
Brain	4474	2800	2289	0.512	0.818
Heart	4264	2643	2165	0.508	0.819
Kidney	4792	2946	2399	0.501	0.814
Liver	4327	2643	2200	0.508	0.832
Testis	5520	3411	2799	0.507	0.821
mean				**0.507**	**0.821**

Table ST4: Conditional probability values for genes with altered neighbourhood (L-CGN).

High Expression Divergence (HED)

	n(L-CGN)	n(HED)	n(HED ∩ L-CGN)	p(HED|L-CGN)	p(L-CGN|HED)
Brain	1133	2807	622	0.549	0.222
Heart	1059	2680	581	0.549	0.217
Kidney	1158	3004	611	0.528	0.203
Liver	1012	2696	569	0.562	0.211
Testis	1311	3420	699	0.533	0.204
mean				**0.544**	**0.211**

Low Expression Divergence (LED)

	n(L-CGN)	n(LED)	n(LED ∩ L-CGN)	p(LED|L-CGN)	p(L-CGN|LED)
Brain	1133	2800	511	0.451	0.183
Heart	1059	2643	478	0.451	0.181
Kidney	1158	2946	547	0.472	0.186
Liver	1012	2643	443	0.438	0.168
Testis	1311	3411	612	0.467	0.179
mean				**0.456**	**0.179**

n(HED): No genes with high expression divergence (above median); **n(LED)**: No genes with low expression divergence (below median); **n(H-CGN)**: No genes with high CGN score (>0.5); **n(L-CGN)**: No genes with low CGN score (<=0.5).

Table ST5: Comparison of conditional probability values and their interpretation

	p(HED|L-CGN)	p(LED|L-CGN)	p(HED|H-CGN)	p(LED|H-CGN)
Brain	0.549	0.451	0.488	0.512
Heart	0.549	0.451	0.492	0.508
Kidney	0.528 ▶	0.472	0.499 ◀	0.501
Liver	0.562	0.438	0.492	0.508
Testis	0.533	0.467	0.493	0.507
Significance	p-value (t-test) = $3.23x10^{-04}$		p-value (t-test) = $4.72x10^{-03}$	
Result	The probability that a gene has a high expression divergence given that it has a low CGN is systematically higher than the probability of finding genes with low expression divergence.		The probability that a gene has a high expression divergence given that it has a high CGN is systematically lower than the probability of finding genes with low expression divergence.	
Interpretation	This means that if the gene neighborhood is altered, the gene is more likely to undergo high expression divergence		This means that when gene neighbourhood is conserved, the gene is likely to experience low expression divergence	

	p(H-CGN|HED)[+]	p(L-CGN|HED)	p(H-CGN|LED)	p(L-CGN|LED)
Brain	0.778	0.222	0.818	0.183
Heart	0.783	0.217	0.819	0.181
Kidney	0.797 ▶	0.203	0.814 ▶	0.186
Liver	0.789	0.211	0.832	0.168
Testis	0.796	0.204	0.821	0.179
Significance	p-value (t-test) = $1.82x10^{-09}$		p-value (t-test) = $5.62x10^{-10}$	
Result	The probability that a gene has a high CGN given that it has a high expression divergence is greater than the probability of having a low CGN.		The probability that a gene has a high CGN given that it has a low expression divergence is greater that the probability of having a low CGN.	
Interpretation	This means that when the expression divergence is high, this could also be caused by factors (*e.g.*, evolution of trans- and cis-regulatory elements) other than position effect[+,$]		This means that when the expression divergence is low, then the genes are likely to have their genomic neighbourhood conserved[$,+]	

[+]We wish to emphasize that the data on expression divergence was available for only five tissues and patterns of expression divergence in other tissues remain unknown. Therefore, the findings reported here should be treated as a conservative estimate of the impact of alteration of genomic neighbourhood on transcriptome evolution.

[$]It is highly likely that change in genomic neighbourhood alone or in combination with other mechanisms, such as evolution of *trans*- and *cis*-regulatory elements, contribute to varying extent to expression divergence.

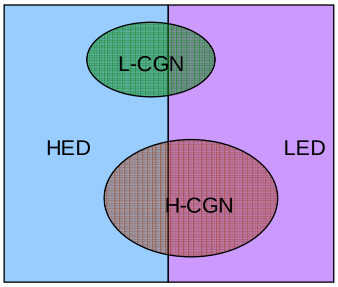


Figure SF8: Schematic representation of the results from the calculations mentioned above.

SI-7: Control calculation - choice of CGN cut-off and expression divergence in five tissues

We observed that genes with altered neighbourhood (CGN ≤ 0.5) tend to show high expression divergence (**Fig 3** in the main text). To investigate if our results are sensitive to the choice of the CGN score cut-off used to identify genes with altered and conserved neighbourhood, we performed two control calculations as described below.

First, among all the genes that have both CGN score and expression divergence between human and chimpanzee, we ranked all genes based on their CGN score. Those with CGN score above median (CGN score >0.637) were placed in the bin of 'Conserved neighbourhood' and those below median were placed in the bin of 'Altered neighbourhood'. Next, in each of the five tissues, we compared the expression divergence values of the genes in the two bins. We observed that in all of the five tissues, the genes with altered neighbourhood consistently showed a significantly higher expression divergence when compared to those genes with conserved neighbourhood (**Table ST6**). These results suggest that the results are robust to the choice of the CGN threshold used to identify genes with conserved or altered neighbourhood.

Next, we repeated the analysis by placing the genes with CGN score in the top 10 percentile (CGN score > 0.775) in the bin of 'Conserved neighbourhood' and bottom 10 percentile (CGN score < 0.416) in the bin of 'Altered neighbourhood'. The results were consistent across all five tissue-types although the p-values were slightly lower due to fewer data-points (**Table ST6)**. Taken together this shows that the results are independent of the choice of CGN score cut-off used to identify genes with conserved or altered neighbourhood.

eeeTissue	**Median Expression divergence (number of genes)**		**P value**		**Median Expression divergence (number of genes)**		**P value**
	LCGN (CGN≤0.637)	**HCGN (CGN>0.637)**			**LCGN (CGN≤0.413)**	**HCGN (CGN>0.791)**	
Brain	0.113 (3013)	0.098 (2594)	**1.284x10^{-3}**		0.116 (587)	0.097 (453)	**1.062x10^{-4}**
Heart	0.266 (2866)	0.245 (2457)	**4.651x10^{-3}**		0.289 (530)	0.244 (442)	**3.386x10^{-3}**
Kidney	0.201 (3138)	0.167 (2812)	**5.833x10^{-6}**		0.206 (590)	0.157 (509)	**1.481x10^{-3}**
Liver	0.317 (2808)	0.269 (2531)	**5.547x10^{-5}**		0.354 (504)	0.290 (458)	>0.05
Testis	0.239 (3559)	0.223 (3272)	**1.390x10^{-2}**		0.281 (653)	0.240 (607)	**4.162x10^{-2}**

Table ST6: Assessing the link between alteration in gene neighborhood and gene expression divergence for five tissue-types. Median cut-off (left panel) and 10 percentile (left panel) cutoff for the CGN score were used to identify genes with conserved or altered neighbourhood. Distributions of the expression divergence value for genes with conserved (HCGN) and altered (LCGN) neighborhood were compared using Mann-Whitney test for five tissue-types: Brain, Heart, Kidney, Liver and Testis. For visual clarity, median expression and sample size (*i.e.*, number of genes) for genes with conserved and altered neighborhood categories are listed for each tissue.

SI-8: Control calculation - choice of CGN cut-off and expression divergence in six brain parts

We observed that genes with low CGN (CGN ≤ 0.5) tend to show high expression divergence compared to those with high CGN (CGN > 0.5) in all six brain parts (**Fig 4** in the main text). To investigate if our results are sensitive to the choice of the CGN score cut-off used to identify genes with altered and conserved neighbourhood, we repeated the analysis by placing genes with CGN score above median (CGN score >0.637) in the bin of 'Conserved neighbourhood' and those below median in the bin of 'Altered neighbourhood'. We observed that in all of the six brain regions, the genes with altered neighbourhood consistently showed a significantly higher expression divergence when compared to those genes with conserved neighbourhood (**Table ST7**). These results suggest that the results are robust to the choice of the CGN threshold used to identify genes with conserved or altered neighbourhood.

Tissue	**Median Expression divergence (number of genes)**		**P value**
	LCGN (CGN≤0.637)	**HCGN (CGN>0.637)**	
Anterior cingulate cortex	0.32 (2528)	0.30 (2255)	**3.373×10^{-2}**
Broca's area	0.325 (2528)	0.31 (2255)	**7.969×10^{-3}**
Caudate nucleus	0.34 (2528)	0.32 (2255)	**7.38×10^{-3}**
Cerebellum	0.33 (2528)	0.30 (2255)	**9.55×10^{-6}**
Prefrontal cortex	0.28 (1211)	0.28 (1062)	**>0.05**
Primary visual cortex	0.31 (1211)	0.27 (1062)	**2.162×10^{-4}**

Table ST7: Assessing the link between alteration in gene neighborhood and gene expression divergence for six brain regions. Median cut-off for the CGN score were used to identify genes with conserved or altered neighbourhood. Distributions of the expression divergence value for genes with conserved (HCGN) and altered (LCGN) neighborhood were compared using Mann-Whitney test for six brain regions: Anterior cingulated cortex, Broca's area, Caudate Nucleus, Cerebellum, Prefrontal cortex and Primary visual cortex,. For visual clarity, median expression and sample size (*i.e.*, number of genes) for genes with conserved and altered neighborhood categories are listed for each tissue.

SI-9: Genomic neighborhood and expression level in different tissues

There were 33,669 probes present in GNF1Hdata.txt which was obtained from GNF Symatlas database (symatlas.gnf.org)[18]. We obtained 51,900 probe-Ensembl protein identifer mapping from Ensemble (v.48) and an additional set of 11,500 mapping of probes to Refseq and Unigene identifiers, which we mapped back to Ensembl peptide identifiers. The two sets were not mutually exclusive, and the combined non-redundant set had 57,956 probes to Ensembl protein identifiers mapped. We excluded probes mapping to multiple genes, or those mapping to no genes. If a peptide had multiple probes mapping to it, we averaged the intensity of all probes on that transcript in a given tissue-type, and reported that as the expression intensity of that transcript in that tissue-type. Finally, we removed the following tissue-types from subsequent analysis: Colorectal Adenocarcinoma, Leukemia lymphoblast (molt4), Lymphoma burkittsRaji, Leukemia Promyelocytc (hl60), Lymphoma burkitts Daudi and Leukemia chronic melogenous.

The final dataset had expression data for 33,495 transcripts from 17,185 genes in 72 tissue-types. Since some genes can have more than one transcript, the CGN score of the gene was assigned to all its transcripts. CGN scores were available for 13,963 genes of which 2,574 genes had their genomic neighbourhood altered. A transcript was considered tissue specific if it was expressed in at least one tissue and had above-median expression (>321 a.u.) in less than 10% of the tissue-types considered (*i.e.*, expression breadth < 0.1; 7 or fewer tissue-types). This resulted in the identification of 4,009 trancripts that showed a strong tissue-specific expression pattern of which 683 also had alterations in their genomic neighbourhood (*i.e.*, low CGN). The list of the 683 transcripts and the CGN scores are provided as an excel file and as a tab-delimited text file from the supplementary URL:
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS5.xls
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS5.txt

Table ST8: The description of the columns in the file is given below

The first column provides the Ensembl gene identifier for human genes (Ensembl v.48)
The second column provides the HGNC gene name
The third column provides the ENSEMBL peptide identifier for human genes (Ensembl v.48)
The fourth column chromosome number and chromosomal position of the human gene
The fifth column provides CGN score of the human genes calculated using chimpanzee as reference species (CGN_{2Mb})
The sixth column provides the tissues in which the gene is tissue-specifically expressed (expression breadth < 0.1)

Next, we analyzed if any of the 72 human tissues expressed a particularly high proportion of genes with altered neighbourhood in a tissue-specific manner (expression breadth < 0.1). First, for each human tissue-type, the number of tissue-specific transcripts (X_{obs}) and the number of transcripts of genes with low CGN that are expressed in a tissue-specific manner (x_{obs}) was calculated. In additon, we also calculated the proportion ($R_{obs} = x_{obs}/X_{obs}$) of such genes for each tissue. To assess if this proportion is what one would expect by chance, we then carried out 10^6 random simulations where each time and for each tissue, a set of X_{sim} transcripts was drawn from the set of 4,009 tissue-specifically expressed transcripts with replacement. *i.e.*, the same transcript can be assigned to two different tissues as tissue-specifically expressed. Each time, the number of transcripts with altered neighbourhood (x_{sim}) was estimated and the proportion ($R_{\mathrm{sim}} = x_{sim}/X_{sim}$) of such transcripts was calculated for each tissue. By comparing the value seen in the real data with what is expected from the simulation, a Z score was calculated for each tissue type: $Z = (R_{obs} - \langle R_{sim} \rangle)/\sigma_{sim}$ where $\langle R_{sim} \rangle$ and σ_{Rsim} are the mean standard deviation of distribution of the the proportion over 10^6 simulations, respectively. A highly positive Z-score indicates that the tissue is statistically significantly enriched to express genes with altered neighbourhood (low CGN) in a tissue-specific manner, while a highly negative score indicates under-representation. When Z score is greater (or less) than zero, we computed the p-value of over-representation (or under-representation) associated with the tissue as the proportion of times during the 10^6 simulations when $R_{\mathrm{sim}} > R_{\mathrm{obs}}$. We repeated the process for all the 72 tissues and corrected the p-value for multiple testing using the FDR method in the R package. The results obtained were similar when a more stringent threshold was used to identify genes (expression breadth <0.075; 5 or fewer tissue-types) and both results are available as tables (**Table ST9** and **ST10**) and in **Figure SF9**.

Table ST9: The table of p-values and Z-scores when expression breadth < 0.1 can be obtained from:
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS6.xls
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS6.txt

Table ST10: The table of p-values and Z-scores when expression breadth < 0.075 was used can be obtained from:
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS7.xls
http://www.mrc-lmb.cam.ac.uk/genomes/sde/CGN/tableS7.txt

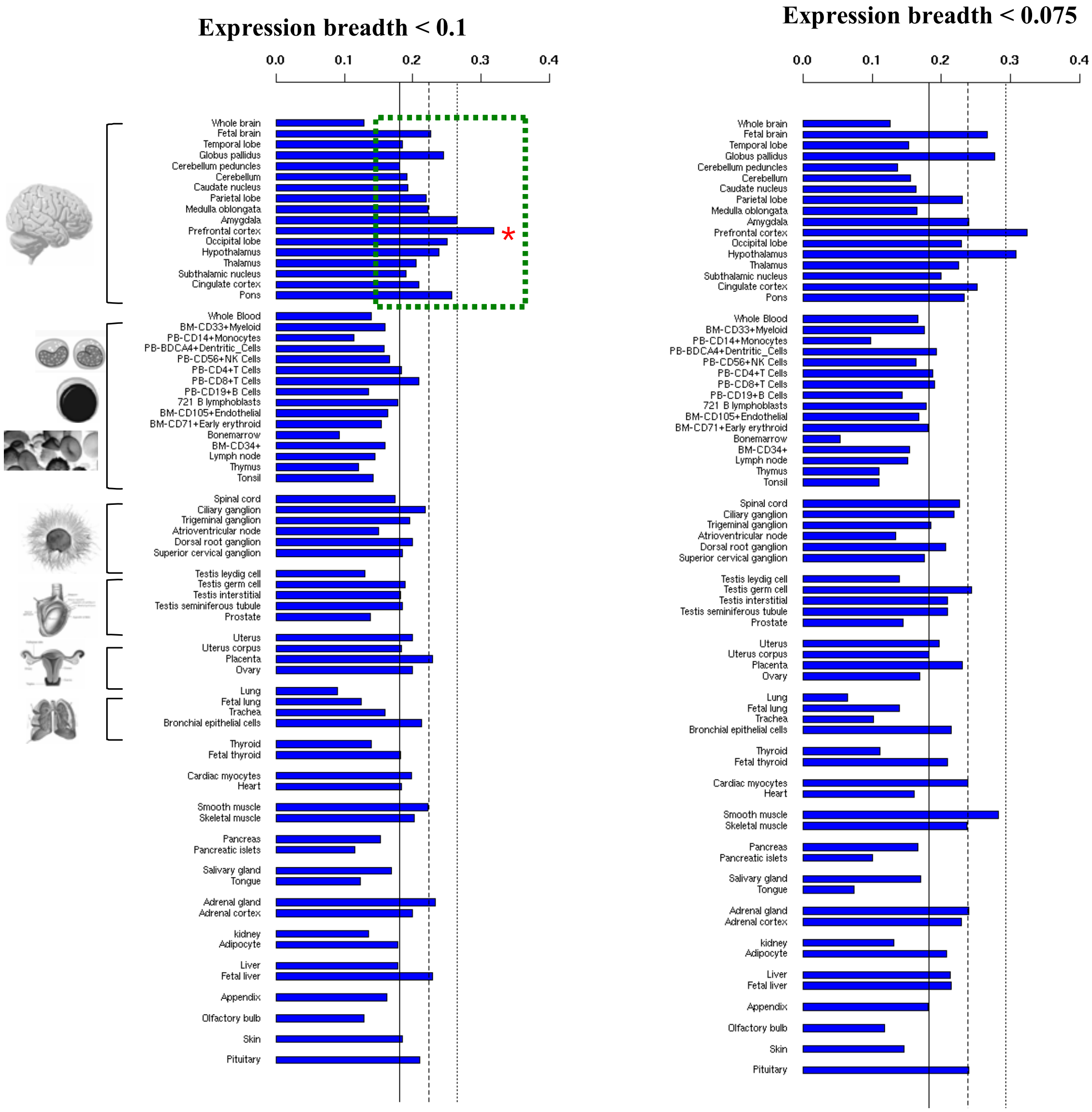


Figure SF9: Fraction of tissue specifically expressed transcripts of genes with altered neighborhood. Tissue specificity was defined by expression breadth < 0.1 (left) and < 0.075 (right). The results are comparable between the two panels, indicating that the findings are robust to the choice of parameters used to define tissue-specific genes. Mean, Mean + 1 St.Dev and Mean + 2 St.Dev are shown as full, dashed and dotted horizontal lines.

We repeated the analyses using the most stringent criteria: considering only those genes which are expressed in one tissue as tissue-specific. The result is shown in **Figure SF10**. Please note that the number of tissue-specifically expressed genes was very low for several of the tissues to get a reliable estimate.

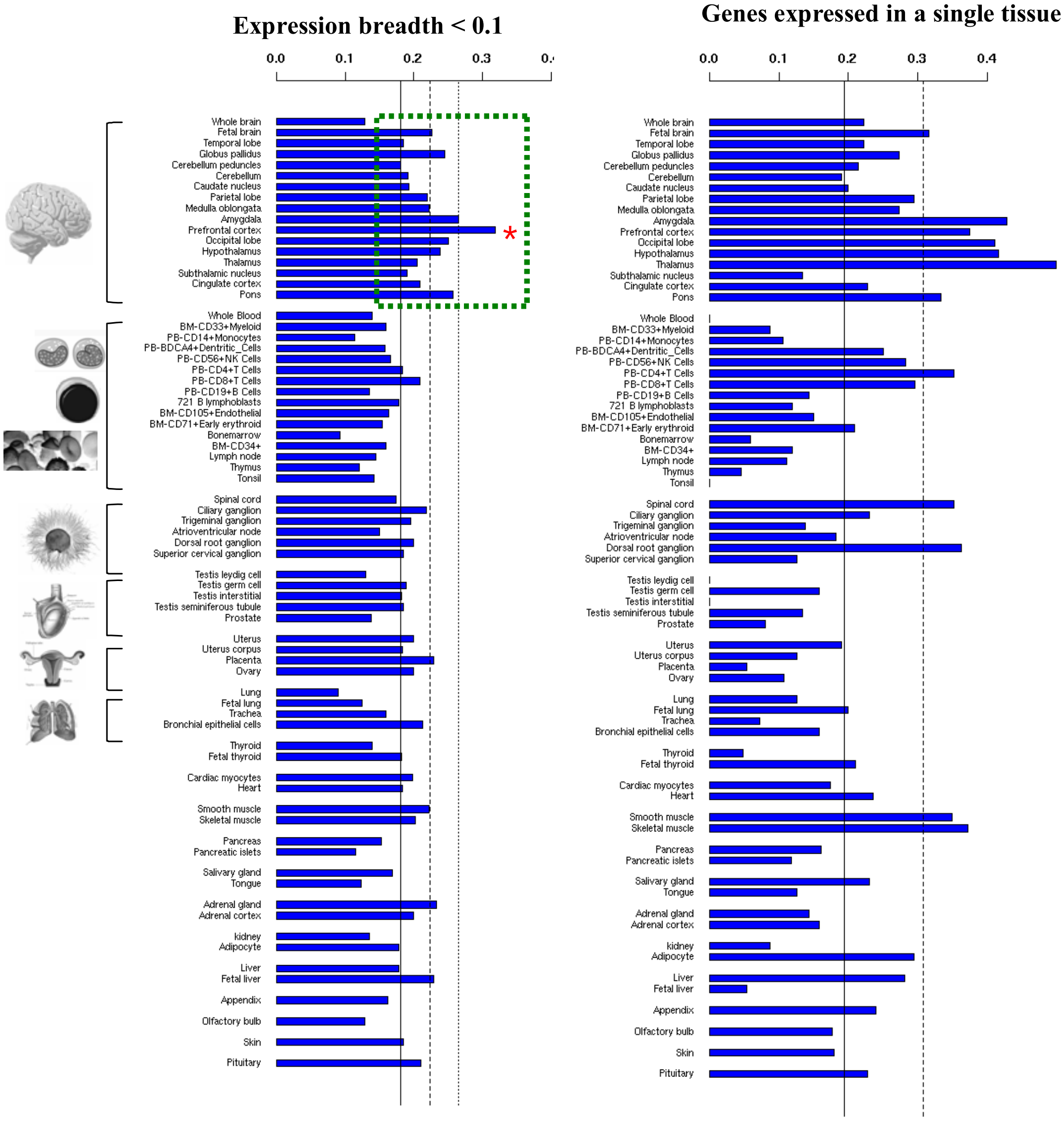
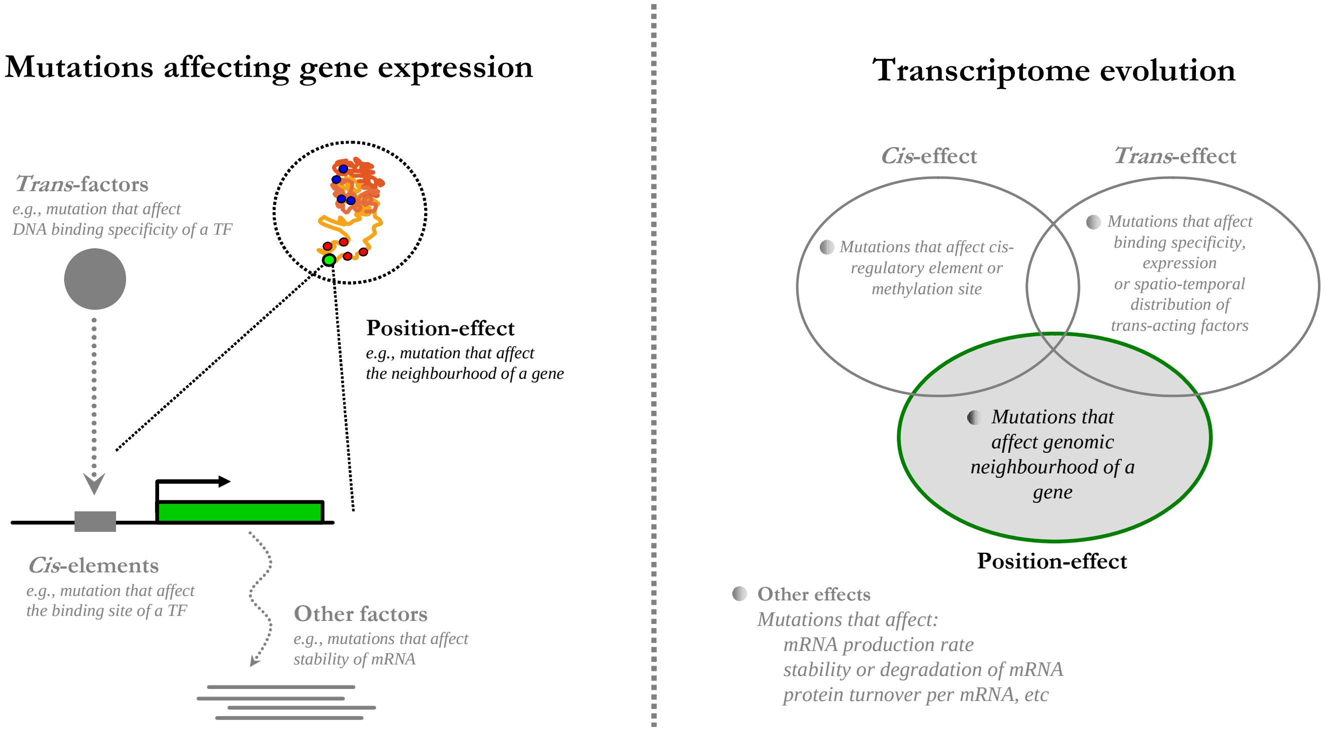


Figure SF10: Fraction of tissue specifically expressed transcripts of genes with altered neighborhood. Tissue specificity was defined by expression breadth < 0.1 (left) and expression in a single tissue (right). Mean, Mean + 1 St.Dev and Mean + 2 St.Dev are shown as full, dashed and dotted horizontal lines.

SI-10: Factors contributing to transcriptome evolution



Mutations affecting gene expression

***Trans*-factors**
e.g., mutation that affect DNA binding specificity of a TF

Position-effect
e.g., mutation that affect the neighbourhood of a gene

***Cis*-elements**
e.g., mutation that affect the binding site of a TF

Other factors
e.g., mutations that affect stability of mRNA

Transcriptome evolution

***Cis*-effect**

Mutations that affect cis-regulatory element or methylation site

***Trans*-effect**

Mutations that affect binding specificity, expression or spatio-temporal distribution of trans-acting factors

Mutations that affect genomic neighbourhood of a gene

Position-effect

Other effects
Mutations that affect:
mRNA production rate
stability or degradation of mRNA
protein turnover per mRNA, etc

In addition to other molecular mechanisms such as *trans*- and *cis*-regulatory changes, alterations in genomic neighbourhood is an important factor that drives transcriptome evolution