Supplemental Material for Genomic signatures of germline gene expression

SUPPLEMENTAL FIGURES

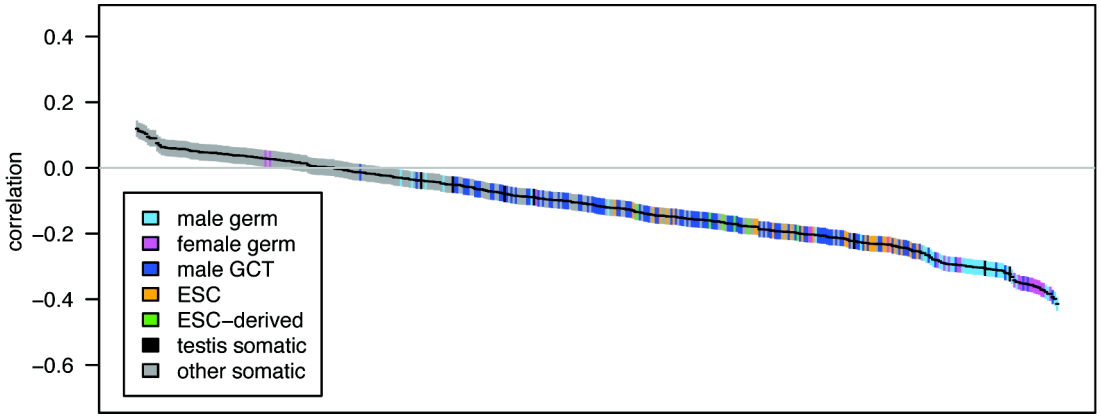

Figure S1. Pairwise correlations between gene expression and crossover rate. Each of the 409 tissue samples is represented by a single bar, colored by tissue type as defined in the key (ESC = embryonic stem cells, GCT = germ cell tumors). Bars are ordered from left to right by the correlation coefficient, with the vertical extent of the bar indicating the 95% confidence interval. A total of 8420 autosomal genes that met filtering criteria were used for this analysis.

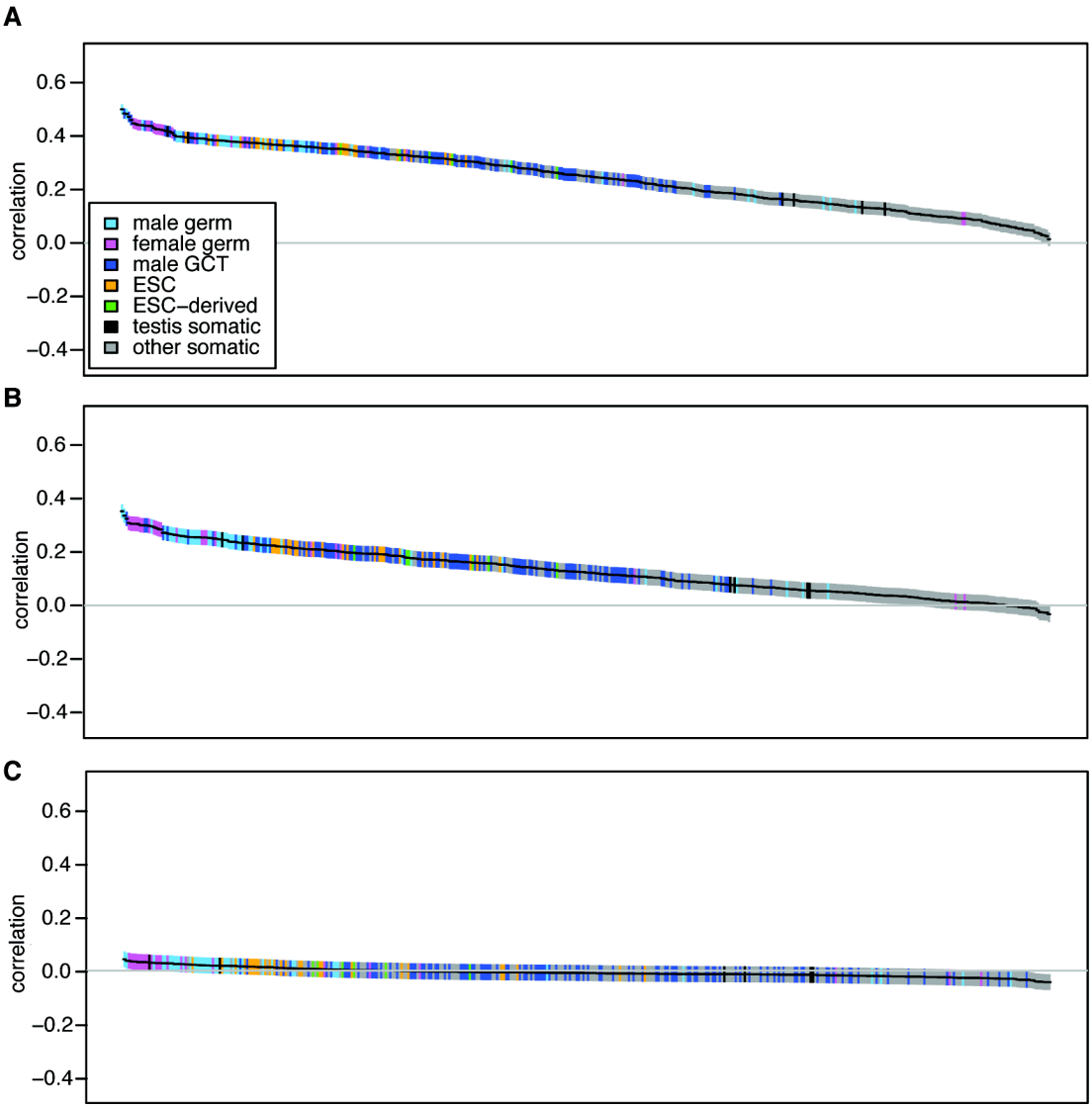

Figure S2. Pairwise correlations between gene expression and (A) G+T content, (B) A→G / T→C substitution asymmetry, (C) G→A / C→T substitution asymmetry. Figure layout is as described for Figure S1.

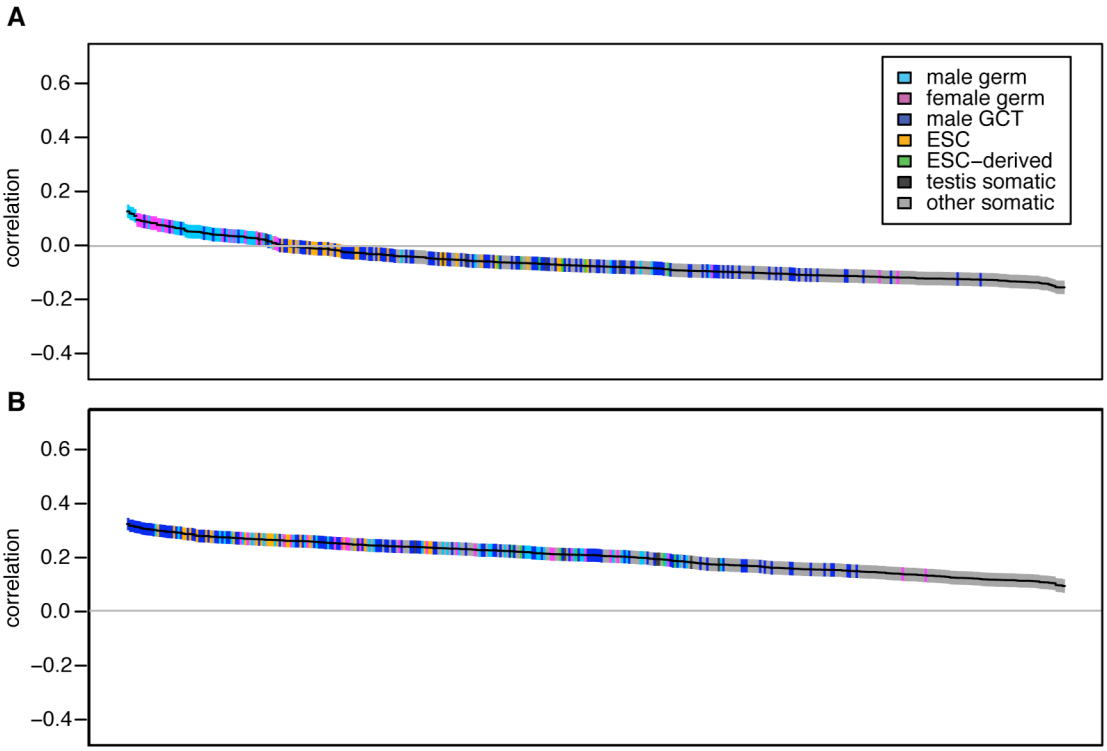

Figure S3. Pairwise correlations between gene expression and density of (A) L1 elements and (B) Alu elements. Figure layout is as described for Figure S1.

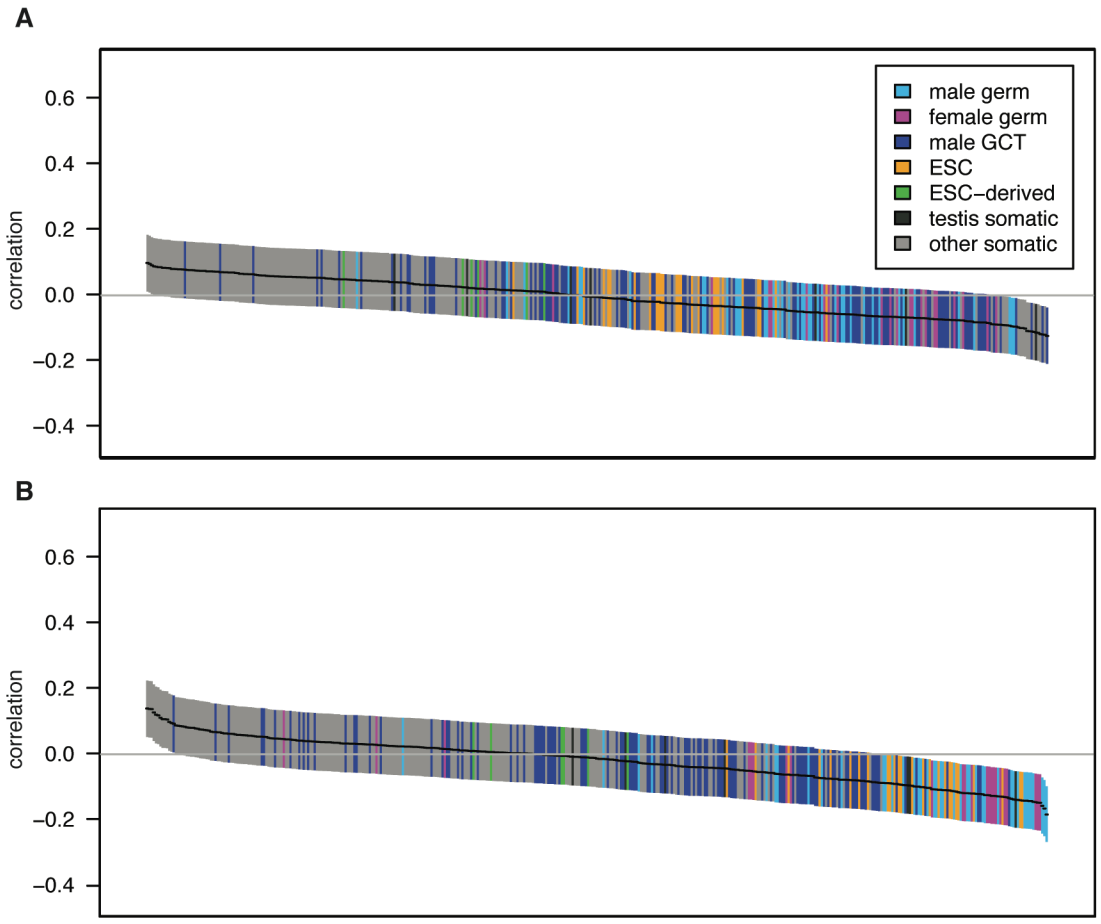

Figure S4. Pairwise correlations between gene expression and orientation bias of transposable elements for high tissue differentiation genes. The figure layout is as described in Figure S1. Correlations are between gene expression and (A) L1 orientation bias or (B) Alu orientation bias (n = 507).

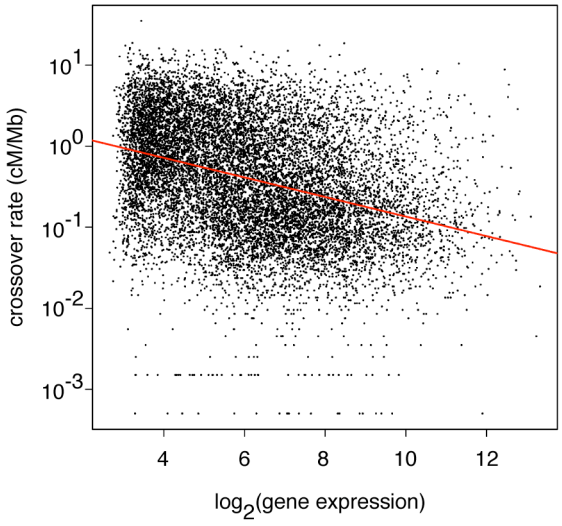

Figure S5. Scatterplot of fetal ovary gene expression versus crossover rate. A total of 12,396 autosomal genes are plotted for which expression data and at least 10 kb of filtered sequence was available. The red line is the best linear fit for $\log_2$ crossover rate versus $\log_2$ gene expression ($r^2 = 0.12$, $P < 10^{-300}$). Gene expression was estimated by averaging all fetal ovary samples from 12-18 weeks gestation.

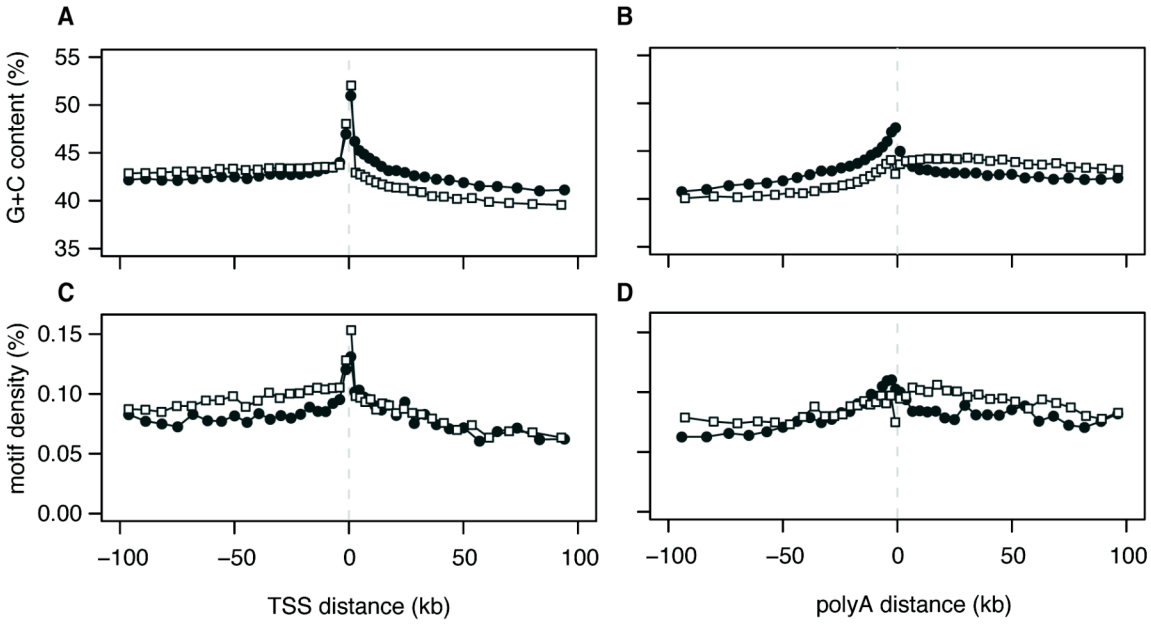

Figure S6. G+C content and density of a recombination hotspot motif as a function of distance from the transcription start site (TSS) or polyadenylation (polyA) site. Sites were binned as described in Figure 3 of the main text. (A) G+C content as a function of distance from the TSS for low (black circles) and high (open squares) expression genes. (B) G+C content as a function of distance from the polyA site. (C) Density of the recombination hotspot motif CCNCCNTNNCCNC (or its reverse complement) as a function of distance from the TSS. Density was calculated by dividing the number of sites within identified motifs by the total number of sites. (D) Density of the recombination hotspot motif as a function of distance from the polyA site.

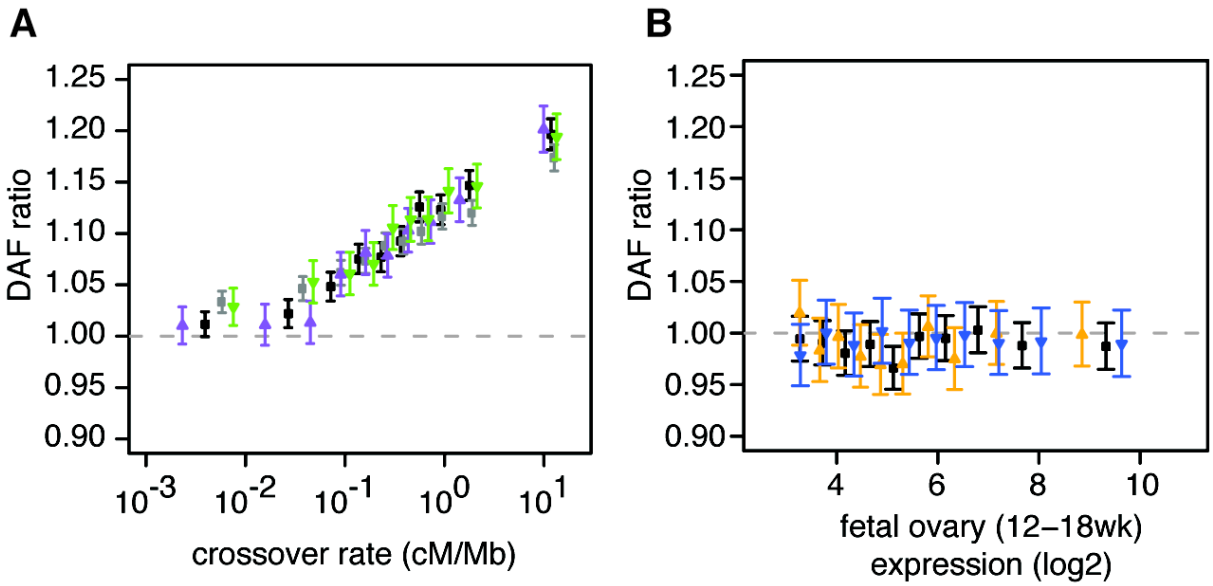

Figure S7. Derived allele frequency (DAF) ratios for filtered HapMap phase II single nucleotide polymorphisms (SNPs) binned by fetal ovary (12-18 weeks gestation) expression or finescale crossover rate. The mean DAF of each bin was estimated using the Yoruban genotypes. Similar results were obtained using other SNP datasets (see Supplementary Table S5). **(A)** Ratios of mean W (A or T) → S (G or C) and S→W DAFs as a function of local crossover rate in intergenic regions (open grey circles); introns of all genes (black squares); introns of high-expression genes (purple triangles); or introns of low-expression genes (inverted green triangles). **(B)** Ratios of A→G and T→C DAFs (where alleles indicate coding-strand nucleotide) as a function of gene expression for all intronic SNPs (black squares); intronic SNPs with high crossover rates (orange triangles); and intronic SNPs with low crossover rates (inverted blue triangles). "High" = above median, "low" = below median. Error bars are 95% confidence intervals of the mean ratios calculated using Fieller's theorem (Fieller 1954) (as implemented in R's mratios package (Dilba Djira et al. 2008)).

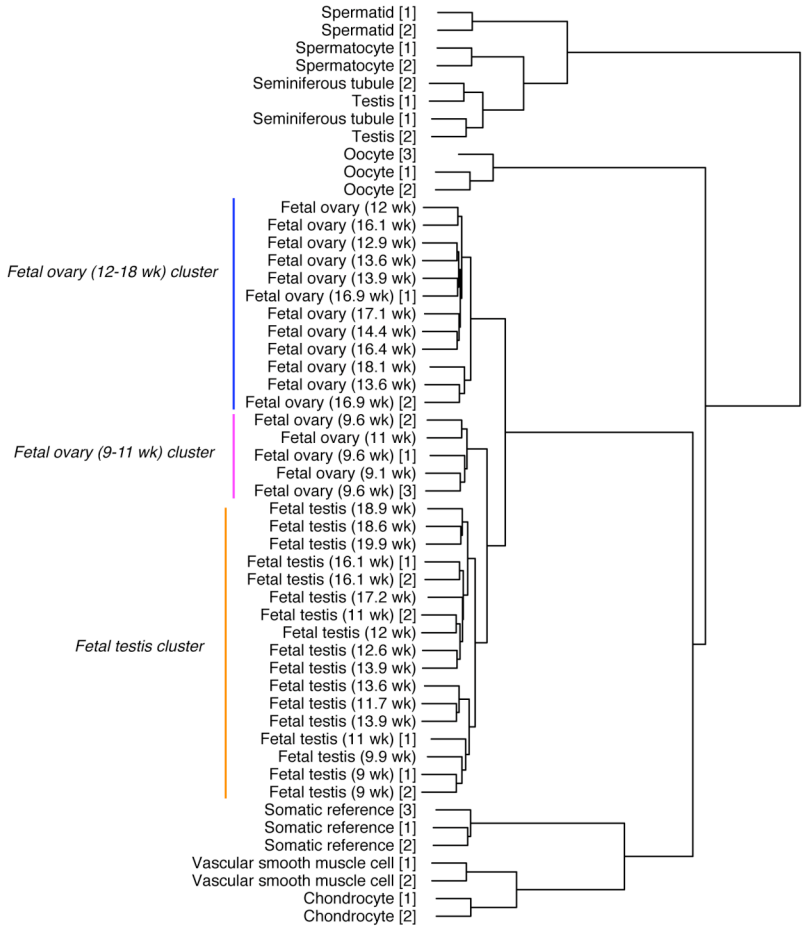

Figure S8. Clustering of expression data. We performed hierarchical clustering of gene expression samples from (Kocabas et al. 2006; Chalmel et al. 2007; Houmard et al. 2009) using R's hclust function with the "complete" method, and the distance between two expression samples defined as 1-*r*, where *r* is the Pearson correlation of their gene expression values.

SUPPLEMENTAL TABLES

Supplemental Tables S1, S2 and S3 are provided as separate files

Table S1. Summary of gene expression samples used in this study.

Table S2. Summary of pairwise correlations with gene expression.

Table S3. Summary of pairwise correlations with gene expression in high tissue differentiation genes

region	n	mean	SD	min	Q1	med	Q3	max	*P*
G+T content									
autosomes	9577	0.520	0.016	0.431	0.509	0.521	0.532	0.597	$< 10^{-300}$
chrX	348	0.518	0.014	0.479	0.508	0.518	0.528	0.574	1.5×10^{-75}
chrY	29	0.516	0.021	0.477	0.498	0.516	0.534	0.553	1.9×10^{-4}
PAR	11	0.523	0.031	0.437	0.521	0.534	0.538	0.548	3.3×10^{-2}
log2(A→G / T→C)									
autosomes	6561	0.54	0.55	-2.56	0.21	0.56	0.88	3.87	$< 10^{-300}$
chrX	222	0.54	0.52	-1.03	0.23	0.52	0.90	2.00	2.6×10^{-37}
log2(G→A / C→T)									
autosomes	6561	0.09	0.42	-2.15	-0.14	0.10	0.32	2.59	9.9×10^{-68}
chrX	222	0.14	0.56	-2.48	-0.16	0.14	0.43	1.99	2.7×10^{-4}

Table S4. G+T content and substitution asymmetry distribution summary measures, by gene chromosomal origin. Only genes with at least 10kb of sequence were included in calculations. *P* values are from two-sided t-tests for the null hypotheses that the mean G+T content is 0.5 and the mean log substitution rate ratio is 0. PAR, pseudoautosomal region.

region	expr	crossover	A→G/T→C				G→A/C→T			
			P	ratio	CI		*P*	ratio	CI	
HapMap										
intron			**4.7E-04**	0.99	0.98	0.99	3.8E-01	1.00	0.99	1.00
intergenic			7.3E-01	1.00	0.99	1.00	5.4E-01	1.00	1.00	1.01
intron		low	5.9E-02	0.99	0.98	1.00	4.9E-01	1.00	0.99	1.01
intergenic		low	5.3E-01	1.00	0.99	1.01	2.2E-01	1.01	1.00	1.01
intron		high	**1.8E-02**	0.99	0.98	1.00	5.6E-01	1.00	0.99	1.01
intergenic		high	2.5E-01	1.00	0.99	1.00	7.6E-01	1.00	0.99	1.01
intron	low		**3.4E-03**	0.99	0.98	1.00	1.4E-01	0.99	0.98	1.00
intron	high		2.3E-01	0.99	0.98	1.00	4.7E-01	1.00	0.99	1.01
intron	low	low	1.0E-01	0.99	0.97	1.00	6.5E-02	0.99	0.97	1.00
intron	low	high	**2.2E-02**	0.98	0.97	1.00	7.2E-01	1.00	0.98	1.01
intron	high	low	3.8E-01	0.99	0.98	1.01	3.7E-01	1.01	0.99	1.02
intron	high	high	8.5E-01	1.00	0.98	1.01	9.4E-01	1.00	0.99	1.02
Keinan										
intron			**2.5E-02**	0.97	0.95	1.00	2.6E-01	0.99	0.96	1.01
intergenic			7.8E-01	1.00	0.98	1.02	8.2E-01	1.00	0.98	1.02
intron		low	8.1E-01	1.00	0.96	1.03	9.2E-01	1.00	0.96	1.03
intergenic		low	2.8E-01	1.02	0.99	1.05	8.5E-01	1.00	0.97	1.03
intron		high	**5.1E-03**	0.95	0.92	0.99	1.1E-01	0.97	0.93	1.01
intergenic		high	4.6E-01	0.99	0.96	1.02	5.9E-01	1.01	0.98	1.04
intron	low		8.4E-02	0.97	0.93	1.00	**1.9E-02**	0.96	0.92	0.99
intron	high		2.6E-01	0.98	0.95	1.02	5.5E-01	1.01	0.97	1.05
intron	low	low	7.5E-01	0.99	0.94	1.05	2.1E-01	0.97	0.92	1.02
intron	low	high	**4.6E-02**	0.95	0.91	1.00	**3.8E-02**	0.94	0.89	1.00
intron	high	low	8.8E-01	1.00	0.96	1.05	2.6E-01	1.03	0.98	1.08
intron	high	high	7.1E-02	0.95	0.90	1.00	6.4E-01	0.99	0.93	1.05
EGP/PGA										
intron			6.8E-01	0.99	0.92	1.05	5.7E-01	1.02	0.96	1.08
intergenic			7.0E-02	1.19	0.99	1.44	8.4E-01	1.02	0.87	1.19
intron		low	6.1E-01	1.02	0.93	1.13	3.6E-01	1.04	0.96	1.13
intergenic		low	5.9E-01	1.08	0.82	1.41	5.8E-01	1.07	0.84	1.35
intron		high	3.4E-01	0.96	0.87	1.05	9.3E-01	1.00	0.92	1.08
intergenic		high	5.3E-02	1.30	1.00	1.69	7.9E-01	0.97	0.78	1.21
intron	low		5.7E-01	1.03	0.93	1.13	9.2E-01	1.00	0.93	1.09
intron	high		4.2E-01	0.96	0.88	1.06	4.5E-01	1.03	0.95	1.12
intron	low	low	1.3E-01	1.13	0.97	1.32	9.5E-01	1.00	0.88	1.14
intron	low	high	5.6E-01	0.96	0.85	1.09	9.4E-01	1.00	0.91	1.11
intron	high	low	8.3E-01	0.99	0.88	1.12	2.5E-01	1.07	0.95	1.19
intron	high	high	4.1E-01	0.94	0.81	1.09	8.5E-01	0.99	0.86	1.13

Table S5. Ratios of derived allele frequencies (DAFs) for three single nucleotide polymorphism (SNP) datasets. SNPs were assigned recombination rates from the finescale recombination map and expression values from fetal ovary (12-18wk) if they overlapped a gene. SNPs were then classified as intronic or intergenic and further

subdivided by expression and crossover rate. SNPs were defined as having “high” or “low” crossover and expression depending on whether their values fell above or below the median of all HapMap SNPs. Mean DAFs were calculated for SNPs with a particular ancestral state and their ratios are summarized in the table. The notation A→G indicates, for example, that the mean DAF was calculated from G alleles with a putative ancestral state of A. *P*-values for the ratio of mean DAFs being different from 1.0 were calculated using Fieller's theorem (Fieller 1954; Dilba Djira et al. 2008). *P*-values less than 0.05 are highlighted in bold; only the A→G / T→C ratio for HapMap SNPs in introns remains significantly different from 0 after correcting for the number of tests performed. In this case the mean DAF ratio is slightly below 1.0 (in the opposite direction of the A→G / T→C substitution asymmetry).

	N_G	N_S	Crossover rate			G+T content			A→G / T→C			G→A / C→T			L1 density			Alu density		
			r_G	r_S	P	r_G	r_S	P	r_G	r_S	P	r_G	r_S	P	r_G	r_S	P	r_G	r_S	P
Study																				
Barberi	3	0	-0.17			0.34			0.18			0.00			-0.06			0.27		
Chalmel	8	4	-0.31	-0.20	**8.8E-03**	0.34	0.30	2.0E-01	0.22	0.17	6.7E-02	0.02	0.00	1.1E-01	0.07	-0.04	**2.5E-05**	0.26	0.18	**3.5E-03**
Ge	2	34	-0.16	-0.06	3.7E-01	0.26	0.19	2.1E-01	0.14	0.09	3.4E-01	-0.01	-0.01	8.4E-01	-0.04	-0.09	5.1E-01	0.21	0.16	2.3E-01
Houmard	34	0	-0.32			0.40			0.27			0.03			0.06			0.22		
Kocabas	3	3	-0.21	-0.06	**3.9E-04**	0.34	0.14	**1.6E-04**	0.21	0.06	**3.4E-05**	0.00	0.00	8.0E-01	0.02	-0.10	**6.9E-04**	0.21	0.14	**1.3E-02**
Korkola	107	0	-0.16			0.30			0.16			-0.01			-0.06			0.23		
Looijenga	12	0	-0.29			0.42			0.26			0.00			0.03			0.30		
Perez-Iratxeta	6	0	-0.14			0.33			0.17			0.00			-0.06			0.24		
Sato	3	0	-0.24			0.38			0.22			0.01			0.00			0.29		
Skottman	14	0	-0.22			0.37			0.21			0.01			-0.01			0.26		
Su	8	138	-0.04	-0.01	1.3E-01	0.15	0.14	6.5E-01	0.06	0.05	3.8E-01	-0.02	-0.02	2.6E-01	-0.08	-0.11	**3.7E-02**	0.18	0.15	6.7E-02
Wu	3	3	-0.38	-0.28	6.7E-02	0.47	0.39	**3.0E-02**	0.33	0.22	**2.1E-02**	0.03	0.01	1.9E-01	0.11	0.01	**6.9E-03**	0.24	0.22	1.5E-01
Microarray												0.00								
hgu133A	143	172	-0.16	-0.02	**1.3E-48**	0.30	0.15	**7.2E-56**	0.16	0.05	**1.0E-54**	-0.01	-0.02	**6.5E-13**	-0.05	-0.10	**6.8E-28**	0.23	0.15	**6.4E-42**
hgu133plus2	60	10	-0.31	-0.18	**3.4E-03**	0.40	0.28	**6.3E-03**	0.26	0.15	**8.5E-04**	0.02	0.01	**8.4E-03**	0.05	-0.04	**9.7E-05**	0.24	0.18	**3.0E-04**

Table S6. Mean gene expression correlations for germline-like and somatic tissues. Microarray experiments were grouped either by study or microarray platform and pairwise correlations between gene expression and crossover rate, G+T content, *etc.* were calculated for each experiment. Each experiment was further classified as "germline-like" if it was from tissues containing germline cells (*e.g.* whole testis), embryonic stem cells, or germline cell tumors, and otherwise as somatic. Non-germline-like immortalized cell lines (*e.g.* HeLa cells) were excluded. n_G, the number of germ-like experiments; n_S, the number of somatic experiments; r_G, the mean germ-like correlation; r_S, the mean somatic correlation; P, the P-value for the difference in means from a two-sided Welch's t-test. P-values which are significant at the 0.05 level are highlighted in bold.

SUPPLEMENTAL NOTES

Supplemental Note S1—Microarray batch effects.

We compare expression data from numerous studies and two microarray platforms, so batch effects are a potential concern. To test whether batch effects affect our conclusions we compared expression correlations separately for each microarray platform and for each study (Supplementary Table S6). Germline-like tissues have significantly greater mean correlations than somatic tissues for both microarray types, and the trend is in the same direction for all five studies with both tissue types (significantly so for 4/5 studies for L1 density; 2/5 studies for G+T content, A→G / T→C substitution asymmetry, crossover rate, and Alu density). Thus, the correlations with gene expression are stronger for germ tissues than somatic tissues even when each study or microarray platform is considered separately.

Supplemental Note S2—Transposable element orientation bias.

We examined the orientation of intronic transposable elements with respect to the direction of transcription of the genes that they reside in. We assigned each gene a bias, *b,* calculated as $b = \log_2\left((n_f + n_p)/(n_r + n_p)\right)$ where n_f and n_r are the number of bases in forward and reverse orientation elements, and n_p = 10 is a small pseudocount to avoid division by 0. For this analysis we only considered intronic sites at least 100 bp from exons, and examined the set of 507 high tissue differentiation genes that have at least 10 kb of intronic sequence. As has been previously observed (Medstrand et al. 2002;

Glusman et al. 2006), both L1 and Alu elements have significant orientation biases, with less elements in the forward than reverse orientation (mean b_{Alu} = -0.22; $P = 4.1\times10^{-4}$; mean b_{L1} = -1.56; $P < 2.2\times10^{-16}$; by one-sample, two-sided t-tests).

We next examined if the orientation bias of transposable elements is correlated with gene expression in germline and somatic cells (Figure S4). The L1 orientation bias shows no evidence for a correlation with gene expression (the strongest correlation is r = -0.12; $P = 5.3\times10^{-3}$, by t-test, not significant after correction for multiple tests). The Alu orientation bias shows a slight negative correlation with expression in germ cells and the strongest correlation is with spermatogonial stem cells (r = -0.18; $P = 5.3\times10^{-5}$, by t-test; following Bonferroni correction for 409 tests, $P = 1.4 \times 10^{-2}$). The mean correlation of germline tissues ($\bar{r} = -0.098$) is stronger than that of somatic cells ($\bar{r} = 0.025$). This correlation is weak compared to that of repeat density alone.

As there is no correlation between gene expression and L1 orientation bias, and the correlation with Alu orientation bias is weak compared to that of Alu density, the orientation bias may result from selection against the introduction of new polyadenylation sites, rather than strand-biased insertion (Smit 1999; Glusman et al. 2006).

SUPPLEMENTAL METHODS

Allele Frequencies

We estimated derived allele frequencies using three polymorphism datasets: SNPs identified from complete resequencing of targeted gene regions in the SeattleSNPs NHLBI Program for Genomic Applications and the NIEHS Environmental Genome Project (NIEHS SNPs 2009; SeattleSNPs 2009) (downloaded April 27, 2009) ("EGP/PGA"), HapMap SNPs which were extensively filtered in order to be "cleanly ascertained" (Keinan et al. 2007) ("Keinan"), and the complete set of non-redundant HapMap phase II SNPs (October 2008 update, downloaded February 5, 2009) ("HapMap").

To account for uneven genotyping depth, we resampled down to 40 chromosomes for the EGP/PGA dataset and down to 100 chromosomes for the Keinan and HapMap datasets, discarding SNPs that had fewer genotypes than this threshold and SNPs that were monomorphic after resampling. Genotypes from children within trios were not used. For the EGP/PGA and HapMap datasets we inferred the ancestral alleles using the chimpanzee sequence, and required that one of two alleles match the human reference sequence. We omitted SNPs that were not flanked by conserved nucleotides in the human/chimp/macaque alignment and SNPs that may have arisen from deamination of 5-methyl-cytosine (A/G SNPs following a C, and T/C SNPs preceding a G). (For the Keinan dataset, the same CpG filter had already been applied, and we used their ancestral allele assignment, which was determined using chimp and orangutan).

REFERENCES

Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SS, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jégou B et al. 2007. The conserved transcriptome in human and rodent male gametogenesis. *Proc Natl Acad Sci USA* **104**: 8346-8351.

Dilba Djira G, Hasler M, Gerhard D, Schaarschmidt F. 2008. mratios: Inferences for ratios of coefficients in the general linear model.

Fieller EC. 1954. Some Problems in Interval Estimation. *Journal of the Royal Statistical Society Series B (Methodological)* **16**: 175-185.

Glusman G, Qin S, El-Gewely MR, Siegel AF, Roach JC, Hood L, Smit AF. 2006. A third approach to gene prediction suggests thousands of additional human transcribed regions. *PLoS Comput Biol* **2**: e18.

Houmard B, Small C, Yang L, Naluai-Cecchini T, Cheng E, Hassold T, Griswold M. 2009. Global gene expression in the human fetal testis and ovary. *Biol Reprod* **81**: 438-443.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**: 1251-1255.

Kocabas AM, Crosby J, Ross PJ, Otu HH, Beyhan Z, Can H, Tam WL, Rosa GJ, Halgren RG, Lim B et al. 2006. The transcriptome of human oocytes. *Proc Natl Acad Sci USA* **103**: 14027-14032.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**: 1483-1495.

NIEHS SNPs. 2009. NIEHS Environmental Genome Project.

SeattleSNPs. 2009. NHLBI Program for Genomic Applications, SeattleSNPs.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**: 657-663.