Mutation Patterns in Human Cancer and Coexpressed Genes in the *Drosophila* Genome

Alan F. Rubin

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Phil Green, Chair
John A. Stamatoyannopoulos
James H. Thomas

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Mutation Patterns in Human Cancer and Coexpressed Genes in the *Drosophila* Genome

Alan F. Rubin

Chair of the Supervisory Committee:
Professor Phil Green
Genome Sciences

For my dissertation I focused on two distinct projects. In the first project, I analyzed early cancer exome datasets to identify patterns of mutation that offer insight into somatic mutational processes in humans. This project began as a critique of the statistical methodology used to identify cancer-causing ("driver") genes in these datasets in one high-profile study. We then analyzed the available data more broadly, and showed that most mutations are not under selection and are therefore a product of the underlying mutational processes in the cancer tissue. Examining these processes in detail, we showed that increased CpG mutation frequency is not a byproduct of aberrant CpG island methylation and that an A→G/T→C associated with gene expression in the germline may be associated with transcription in some cancers. In the second project, I describe a novel approach for identifying clusters of coexpressed genes in eukaryotic genomes and our results applying this model to data from *Drosophila melanogaster.* We found that two-thirds of genes in the *Drosophila* genome are coexpressed with a neighbor. The boundaries of these coexpression clusters are enriched for insulator binding sites, and are correlated with physical interaction domains, suggesting that nuclear structure may play a role in coexpression. We hypothesize that coexpression segments may represent a type of substructure within these interaction domains.

Table of Contents

Acknowledgements

I would like to thank the many current and former members of the Department of Genome Sciences who both supported and challenged me during my graduate career. Phil Green has been a wonderful mentor who embodies the ideals of rigorous scholarship and scientific integrity. The members of my thesis committee, Steve Henikoff, Larry Loeb, Jay Shendure, John Stamatoyannopoulos, and Jim Thomas, generously provided their time, ideas, and encouragement during my thesis work. Graham McVicker was a source of invaluable advice and support during my time in the Green Lab, and I will always be grateful for his guidance. I would also like to thank Sara Di Rienzi and Doug Fowler for their friendship and the many good meals we have enjoyed during my time in graduate school. Finally, and most importantly, I would like to thank my family for their patience, love, and unwavering support.

Chapter 1: Introduction

To understand the eukaryotic genome, we must try to understand the genomic sequence [1-3] and its functional properties [4-6]. Central to understanding the sequence is the knowledge of how it changes over time. Comparative genomics and evolutionary analyses are restricted to germline sequence changes, but somatic mutations are also a source of DNA sequence diversity within an organism. Patterns of somatic mutation carry a variety of information, such as cell lineage [7], but they attract the most research interest in the context of cancer [8].

Much of the research focuses on identifying new "driver" genes that may be responsible for the development and malignancy of cancer [9, 10]. Complicating these searches for "driver" mutations, many cancers exhibit a "mutator phenotype" that results in a proliferation of somatic mutations [11, 12]. This elevated mutation rate allows the cell to acquire the genetic changes it requires to become malignant, but most of the mutations are likely to be neutral. Different cancer types are known to have specific mutational signatures, such as increased mutation at TpC dinucleotides in breast cancer [13], and understanding these patterns may help identify clinically relevant similarities and differences between tumor types, as well as illuminate the mutational mechanisms.

A gene's sequence is most biologically relevant when it is transcribed into mRNA and translated into a functional protein. Transcription is a complex process, with many contributing factors [14]. In addition to *cis*-regulatory elements, modeled linearly by single-gene-centric models such as "Omes Law" [15], neighboring genes may affect gene expression. Clusters of coexpressed genes have been described in a variety of eukaryotes [16-20], and it is well known that eukaryotic gene order is nonrandom [21]. Epigenetic mechanisms such as chromatin loop domains mediated by insulators [22] or nuclear pore localization [23] present plausible models for coordination of coexpressed genes.

In addition to genetic changes, cells may acquire epigenetic changes that contribute to cancer progression [24]. For example, a tumor suppressor can be deactivated effectively by either a destabilizing mutation in the coding sequence or by epigenetic silencing that blocks gene expression. This idea shifted my interests away from somatic mutation and towards gene regulation. With the widespread availability of epigenetic data, such as from the ENCODE [4, 5] and modENCODE [6] projects, on the horizon, I sought to develop a method

for identifying coexpression clusters using only expression data, and then mining the rapidly accumulating epigenetic datasets for plausible mechanisms.

Chapter 2: Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers"

Sjöblom *et al.* [9] found 189 genes with an apparent excess of mutations in breast and colorectal tumors, and concluded that the array of mutations selected in cancer growth is much larger than previously suspected. We reanalyzed their data to correct a statistical error and a questionable background mutation rate assumption, and reach a strikingly different conclusion: mutations in only a handful of genes, not hundreds, play a significant role in the development of these cancers. We also note several interesting patterns that may illuminate mutational mechanisms.

The statistical error arose in correcting for the large number of statistical tests (one per gene). For this purpose, Sjöblom *et al.* use a standard approach, false discovery rates (FDRs) [25], but their calculations incorrectly substitute probabilities of the specific observed mutation patterns for each gene in place of P-values (which are required by FDR theory [25]). This results in substantially underestimating the FDRs and overstating the significance of many genes. Correcting this error, while retaining Sjöblom *et al.*'s assumed background mutation rate of 1.2 / megabase, sharply reduces the number of significant genes (Fig 1).

However the assumed background rate is also doubtful. It is based on two much smaller studies, one of which [26] analyzed only 3.2 Mb of sequence data in colorectal cancer samples and obtained a very broad rate confidence interval of 0.22-2.5/Mb; and another [13] of 518 kinase genes in 25 breast cancer samples, which (after eliminating two outlier samples) yielded a rate of 1.0/Mb. Since 367 of the kinase genes were also sequenced in Sjöblom *et al.*'s discovery screen, we could test consistency between studies by computing the discovery screen rates in these genes. We find a rate of 2.4/Mb, significantly higher ($P < 0.001$) than in [13]. This discrepancy may reflect protocol or tumor sample differences between studies, but in any case indicates that background rates should be determined separately for each study.

Consequently we feel a more appropriate choice of background rate in the present study is that from the discovery screen. (Better, were it available, would be a 'neutral' rate for each sample, based for example on synonymous mutations). Using this, we find that only a

handful of genes (TP53 in breast cancer, and APC, KRAS, TP53, SMAD4, and FBXW7 in colorectal cancer), all of them previously known to be mutated at high frequency in these cancers, are significant at a false discovery rate of 10% (Fig. 1). Moreover, this conclusion is relatively robust to the specific rate assumption. Using a smaller value of 2.0/Mb (which is well within the confidence interval from [26]) yields only 3 additional genes (Fig. 1).

After eliminating the above significant genes, the mutation rate in the remaining validation screen genes of [9] is still significantly higher than the discovery screen rate (5.2/Mb vs. 3.5/Mb for breast cancers, 4.5/Mb vs. 2.8/Mb for colorectal cancers). The higher rate in these genes could reflect weak positive selection for cancer growth mutations, distributed over many genes; however a more likely possibility is that there is gene-to-gene heterogeneity in the strength of purifying selection and/or in the underlying (neutral) mutation rate, such that genes with intrinsically higher 'background' rates are more likely to have been identified. The existence of both kinds of heterogeneity is well established for germline mutations [2, 27].

In examining specific patterns of mutations, we noted several interesting trends (Fig 2). CpG mutation rates are higher for CpGs outside of CpG islands, presumably reflecting the fact that CpG islands are generally not methylated [28]. Somewhat surprisingly, the significantly higher CpG rate for colorectal cancers relative to breast cancers, noted in [9], primarily reflects non-CpG island mutations, rather than (as one might have expected) differential methylation of CpG islands. We also noted (Fig. 2b) an asymmetry to the mutation patterns in breast but not colorectal cancers; such an asymmetry has been found in germline mutations in many genes and appears to be associated with transcription [29].

In summary, the data of Sjöblom *et al.* do not implicate any additional cancer genes beyond the handful known from previous studies. We cannot rule out the possibility that there are additional genes with subtle effects, but if they exist, establishing their importance will require a substantially larger set of samples, and/or alternative experimental approaches. Sjöblom *et al.*'s study is a prelude to the Cancer Genome Atlas project (http://cancergenome.nih.gov/), which seeks to identify causal mutations for a large number of cancers. Our analyses highlight the crucial importance of determining valid background rates for this study.

Figures

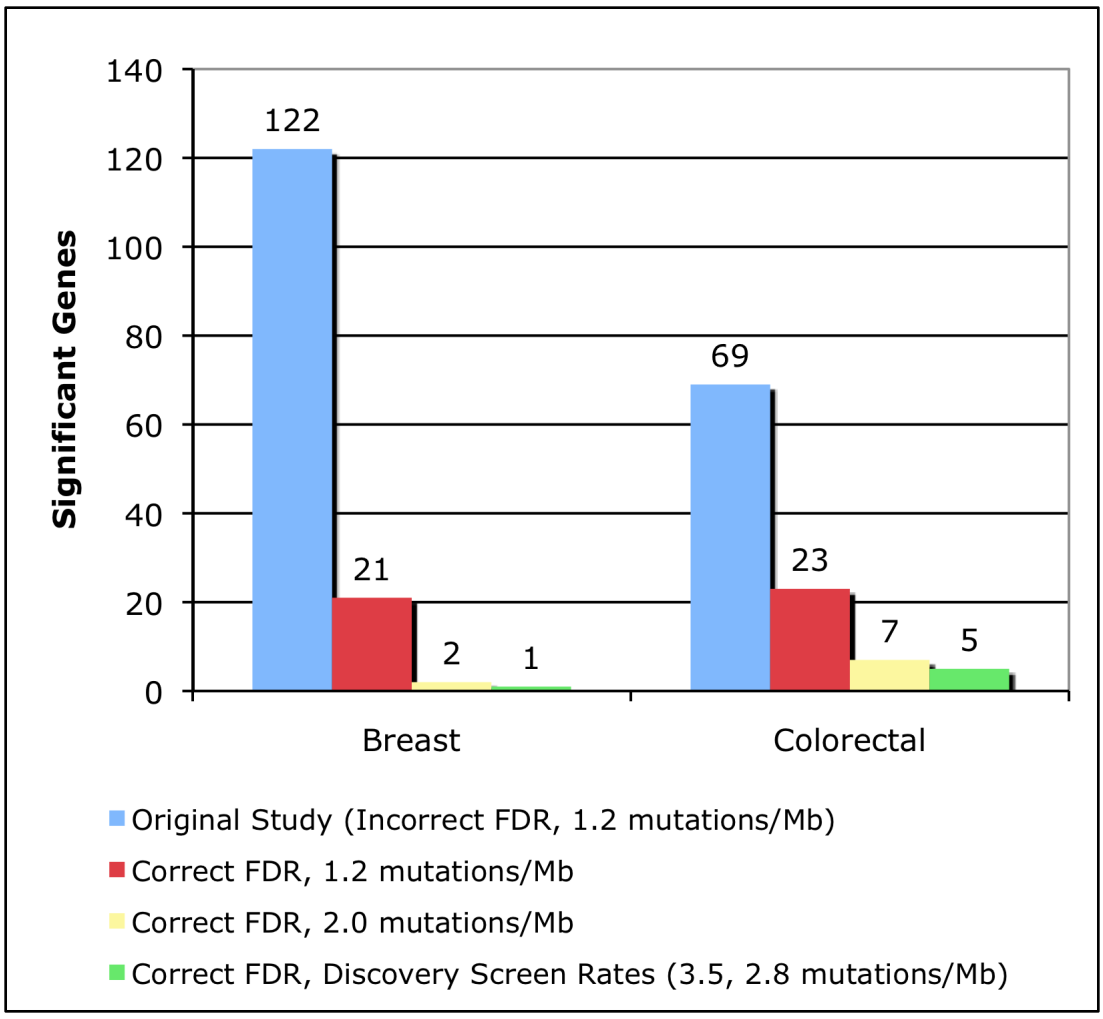


Fig. 1. Number of genes significant at a false discovery rate of 10%, as originally found by Sjöblom *et al.* [9], and upon reanalysis. We estimated relative rates for different substitution types essentially as described [9], but with the following differences: we used only discovery screen data; we estimated distinct rates for C's and for G's within CpG and TpC / GpA dinucleotides; and we estimated distinct rates for CpGs inside and outside of CpG islands. We then used these rates to estimate for each gene, based on its sequence composition, the expected total number, λ, of mutations in the discovery and validation screens combined. The P-value for a gene having a n observed mutations (in both screens combined) is then given by $\sum_{i=n}^{\infty} q_i r_i$ where $q_i = \left(\frac{\lambda^i}{i!}\right) e^{-\lambda}$ is the Poisson probability, and

$r_i \approx \left(1 - \left(\frac{24}{35}\right)^i\right)$ is the probability that at least one of the *i* mutations occurs during the discovery screen. We assumed that sequencing failures were distributed proportionately across all genes. Significant genes using the discovery screen rates (green bars) are TP53 (breast cancer) and APC, KRAS, TP53, SMAD4, and FBXW7 (colorectal cancer).

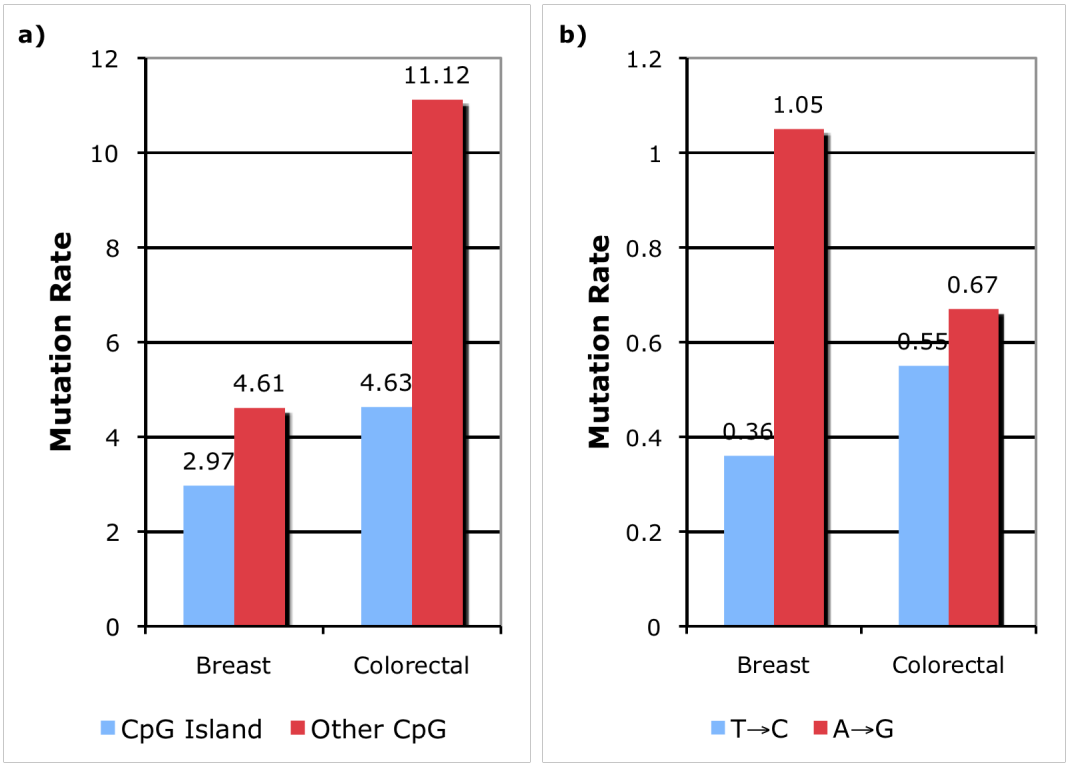


Fig. 2. Different CpG rates inside and outside CpG islands (a); and differing mutational asymmetry between cancer types (b). Rates are for the discovery screen, and are computed by dividing the observed number of nonsynonymous changes of a given type by the number of potential nonsynonymous mutations of that type (in millions). CpG island annotations are from the University of California Santa Cruz genome website [30].

Chapter 3: Mutation Patterns in Cancer Genomes

Abstract

Recent large-scale cancer sequencing studies have focused primarily on identifying cancer-associated genes, but as an important byproduct provide "passenger mutation" data that can potentially illuminate the mutational mechanisms at work in cancer cells. Here we explore patterns of nucleotide substitution in several cancer types using published data. We first show that selection (negative or positive) has affected only a small fraction of mutations, allowing us to attribute observed trends to underlying mutational processes rather than selection. We then show that the increased CpG mutation frequency observed in some cancers primarily occurs outside of CpG islands and CpG island shores, thus rejecting the hypothesis that the increase is a byproduct of island or shore methylation followed by deamination. We observe an A→G vs. T→C mutational asymmetry in some cancers similar to one that has been observed in germline mutations in transcribed regions, suggesting that the mutation process may be influenced by gene expression. We also demonstrate that the relative frequency of mutations at dinucleotide "hotspots" can be used as a tool to detect likely technical artifacts in large-scale studies.

Introduction

Large-scale sequencing of cancer genomes is beginning to have a major impact on cancer research [9, 10, 31-33]. The primary target of such studies is 'driver' mutations, i.e. those that play a role in cancer initiation or progression or provide a cell growth or survival advantage. However, the majority of mutations actually identified are presumed to be 'passenger' mutations that are not advantageous to the cancer cell [10]. From one point of view, passenger mutations are an annoying 'haystack' complicating the search for causal mutations. However, they are also a potentially rich source of information about the specific mutational mechanisms at work in somatic cells and cancer, an aspect we pursue in this paper.

Analyses of neutrally evolving genomic sequences in a variety of organisms have revealed several patterns that are presumed to reflect the nature of the mutational processes in germline cells. Transition mutations occur at a significantly higher rate than transversion mutations [34-36]. Substitution rates depend on flanking nucleotides, a notable example being cytosine in CpG dinucleotides which, in mammals, is usually methylated at the 5-carbon and undergoes hydrolytic deamination to thymine at a relatively high rate [28, 36,

37]. The germline mutation rate at CpGs is much lower within CpG islands (regions enriched in CpGs surrounding or near the transcription start sites of many genes) reflecting, at least in part, the fact that most islands are likely unmethylated in the germline [28]. Substitution rates are asymmetric in transcribed regions of the genome, with A→G transitions occurring at a higher rate than T→C transitions on the coding strand [29]. The magnitude of the effect correlates with gene expression level [38, 39]. While mechanisms for transcriptional mutagenesis have been described in model systems [40, 41], the mechanism responsible for this asymmetry is currently unknown.

Cancer sequencing studies have provided preliminary evidence for several mutational rate trends, some, but not all, of which are similar to those seen in germline mutations. Transitions are more frequent than transversions, and CpG dinucleotides are a mutation hotspot, particularly in colorectal cancer [9, 10, 31-33]. In contrast, TpC dinucleotides are a mutation hotspot in breast cancers [13], but not in the germline [42]. Analysis of mutations in *TP53* suggests that diverse tumor types are affected by different mutation processes [43]. Note that such trends could reflect mutations arising prior to carcinogenesis that are carried along by subsequent clonal expansion, in addition to mutations that occur within the cancer cells themselves.

Here we use published data from several studies to further explore patterns of nucleotide substitution in cancer cells. We first examine the overall impact of selection on the mutation spectra by comparing synonymous and nonsynonymous substitution frequencies (mutations per sites sequenced) in pancreatic cancer and glioblastoma multiforme [32, 33], and by examining the nature of amino acid changes in breast and colorectal cancers [9]. Our results suggest that most coding sequence mutations in cancer are neutral with respect to cancer growth. This finding justifies our assumption that mutation patterns in the data are more likely to reflect the mutation process than selection.

Since methylation of CpG islands and of regions within 2 kb of islands, called CpG island shores, are known features of some cancers [44-47], it has been hypothesized that the elevated CpG mutation frequency is due to increased DNA methylation of islands and shores, followed by deamination of the methylated CpGs [48]. We reject this hypothesis by showing that the elevated frequency primarily reflects increased mutation of CpGs outside of islands and shores. We also find an A→G vs. T→C mutational asymmetry similar to that previously observed in the germline [29], which suggests that the mutation pattern within a gene is influenced by its expression. Finally, we show that the fraction of mutations occurring at dinucleotide hotspots can be a useful metric for identifying technical artifacts in

cancer sequencing studies, by detecting an inconsistency between the discovery and validation screens in one study that is likely due to error-prone sample amplification.

Results and Discussion

Strength of Selection on Coding Sequence Mutations

The majority of mutations observed in cancer sequencing studies are believed to be 'passenger' mutations having little impact on the cancer cell [10]. However, it remains possible that many mutations occurring during cancer growth are deleterious to the cell and consequently eliminated by selection. If so, the set of passenger mutations would not faithfully reflect the underlying mutation process. Since selection on germline mutations in coding sequences acts mainly at the amino acid level [49], we assume that this is also true of somatic mutations and that we can therefore explore the effects of selection by comparing frequencies of nonsynonymous and synonymous substitutions. Using data from two studies where both types of substitution were catalogued [32, 33], we find the overall nonsynonymous/synonymous frequency ratios for pancreatic cancer to be 0.95, and for glioblastoma multiforme to be 1.10, neither of which is significantly different from 1 (P=0.57 and 0.43, respectively). This indicates that (in contrast to germline mutations) the set of nonsynonymous mutations in cancer is not strongly biased by selection. We also calculated separate ratios for dinucleotide contexts that are known to have elevated mutation frequencies in cancer [9, 13], and again found that the nonsynonymous/synonymous substitution frequency ratios are not significantly different from 1 (Fig. S1).

It has been claimed that hundreds of genes are subject to significant positive selection in cancer [9] although this has been disputed [50-52]. Analyzing data in [9] we find that the amino acid substitutions in known cancer genes are significantly different from substitutions in other putative (according to [9]) cancer-associated genes (P=1.61e-4), which in turn are not significantly different from genes that were not cancer-associated in [9] (P=0.133) (Fig. S2). This suggests that the mutations in the putative cancer-associated genes in [9] are primarily passenger mutations, and the set of mutations is therefore unlikely to be strongly biased by positive selection. This is not to say that there is no such selection and indeed a more sensitive test can detect evidence for positive selection in cancer [10]. However, our results suggest the proportion of selected mutations is small. In combination, these results suggest that, while selection certainly acts on some mutations, the set of mutations in cancer cells is not significantly biased by negative or positive selection, and we therefore

assume in the following that the sets of nonsynonymous nucleotide substitutions reported by cancer sequencing studies mainly reflect the underlying mutation process that generated them.

CpG Mutations

CpG dinucleotides are known to be mutation hotspots in some cancer types [9, 31-33]. 32% of the CpGs found in the coding sequence of genes we analyzed (302,780 of 940,319) are in CpG islands, and another 11% (106,691 of 940,319) are found within 2kb of CpG islands, in regions called CpG island shores [47, 53]. Since methylation of CpG islands and/or shores is known to occur in some genes in cancer [44-47], and the higher germline mutation rate at CpGs appears confined to methylated CpGs [54], it has been hypothesized that the elevated CpG mutation rate in some cancers could result primarily from mutations in methylated CpG islands or shores [48]. To test this, we classified the CpG sites in the studied genes as island, shore, or other, and calculated mutation frequencies for each class using published data [31-33]. In three of the four cancer types analyzed (colorectal, pancreatic, and glioblastoma), CpG frequencies are significantly different among classes (Fig. 1), but in all cases the island frequency is lowest and the 'other' frequency is highest. Moreover, the dramatic increase in overall CpG frequency in colorectal cancer relative to the other cancer types is mostly confined to the non-island, non-shore CpGs. The elevated overall mutation frequency at CpG dinucleotides in the cancers is therefore not due primarily to deamination of cytosines in methylated CpG islands or shores. A similar analysis using 500 bp shores (which have a higher density of CpGs and a stronger tendency to cancer-specific hypermethylation than the rest of the 2kb shore [47]) gives similar results (Fig. S3).

We also recalculated CpG mutation frequencies for protein kinase genes in various cancers sequenced in another study (Fig. S4A) [10]. Although the smaller size of this dataset prevents robust conclusions, in colorectal, lung, and ovarian cancers, the CpG mutation frequency outside of islands is again higher, suggesting this may be a general trend.

Excision of thymine in a deaminated CpG is performed by MBD4 and TDG DNA glycosylases [55]. Neither enzyme has coding sequence mutations in the samples used in our analysis, but there could be noncoding changes affecting expression or splicing, or mutations in interaction partner genes. Other possibilities are that cellular changes increasing methylation or deamination rates, or shortening cell division times could reduce the probability of repair occurring prior to DNA replication, in tumor or precursor tissue.

A→G/T→C Mutational Asymmetry

An elevated rate of germline A→G mutations compared to the complementary rate of T→C mutations on the coding strand been associated with transcribed regions in mammals [29]. We tested whether this asymmetry is detectable in cancer mutations, using data from three studies (Fig. 2) [31-33]. All cancer types appear to have some excess of A→G mutations, and the asymmetry is statistically significant in breast cancer (P=2.52e-3). To examine whether this asymmetry is associated with expression, we calculated separate mutation frequencies for genes with higher or lower than median expression, using expression data from ten breast cancer samples to classify the genes [56]. The asymmetry is more pronounced in genes with higher than median expression, suggestive of an expression-related effect, although the difference does not reach statistical significance (Fig. S5). (We found similar results using expression data from normal breast tissue [56]). The fact that the same asymmetry is associated with germline transcription [29] could point to a common (currently unknown) mechanism, for example involving in transcription-coupled DNA repair [57] as hypothesized in [29].

Dinucleotide Hotspots as a Signature

TpC (and the complementary GpA) dinucleotides are mutation hotspots for C→G transversions at the C:G base pair in breast cancer [9, 13]. In addition to breast cancers, three other cancer types sequenced in [10] have a significantly elevated mutation frequency at TpC/GpA dinucleotides: lung cancer, melanoma, and ovarian cancer (Fig. S4B). These results indicate that the TpC/GpA dinucleotide is a mutation hotspot in a subset of cancer types.

As discussed above, CpG dinucleotides are a mutation hotspot in breast cancer, colorectal cancer, pancreatic cancer, and glioblastoma [9, 31-33]. The relative fractions of mutations occurring in TpC and CpG hotspots vs. other sites may thus be viewed as a "signature" that presumably reflects the nature of the mutational mechanisms in different cancers. As such, it may provide a useful tool for monitoring data quality in large sequencing studies.

We compared the proportion of mutations at TpC and CpG sites in the discovery and validation screens for [9] in the genes that were mutated in both screens and found a statistically significant difference between the screens, for each tissue type (Fig. 3) (breast P=1.73e-4, colorectal P=9.06e-3), with the validation screen samples having a higher proportion of non-hotspot mutations. This difference suggests differing mutational mechanisms underlying the data from the two screens. It is not due to enrichment for

cancer-associated genes in the validation screen, because our calculations use the same gene set for both screens. One possible explanation is biological differences in the tissue samples: the breast cancer discovery screen used cell lines whereas the breast cancer validation screen used primary tumor tissue, and the colorectal cancer validation screen had a higher proportion of xenografts (as opposed to cell lines) than the colorectal cancer discovery screen [9]. An alternative possibility is that the difference is due to the whole genome amplification protocol, which was applied to the validation screen samples, but not the discovery screen samples. To discriminate between these possibilities, we examined a subsequent study by the same investigators [31] that analyzed additional genes in the same samples as [9] (excluding one validation screen breast tumor) using generally similar methods, but excluding mutations detected in amplified samples that could not be verified by sequencing of the unamplified sample. We analyzed hotspot proportions in the genes sequenced in [31] and found no significant difference between screens (Fig. S6). Because the same samples were used in both studies, this excludes sample differences as the primary cause for the discrepancy between screens in [9] and points instead to amplification-induced mutations.

The error rate of the ϕ-29 polymerase used in the whole genome amplification reactions in [9] is at least 12.5 mutations per Mb [58], which is significantly higher than the cancer genome mutation rates reported in recent studies [9]. Thus some strategy is needed to eliminate amplification-induced mutations. In [9], five independent amplification reactions were pooled; our results suggest this strategy was not successful, possibly because of variable amplification efficiency among replicates or reproducibility of particular amplification errors. However, combining it with sequencing of unamplified samples to validate putative mutations appears to have successfully eliminated artifacts in [31]. Note that the presence of amplification errors may contribute to the higher overall mutation rate observed in the validation screen for [9], an observation originally attributed to enrichment for cancer-associated genes [9].

Conclusions

Our results indicate that only a small subset of nonsynonymous substitutions in cancer are affected by selection, thus making it possible to interpret substitution trends as reflecting underlying mutational processes. Our analyses eliminate CpG island methylation as a significant factor in increased CpG mutation frequencies, and detect a mutational asymmetry in breast cancers that may be linked to gene expression. We also demonstrate the utility of mutation patterns for detecting technical artifacts in cancer sequencing studies.

Methods

Sources of Data

Mutation data were obtained from supplementary online material provided with [9, 31] (breast and colorectal cancers), [33] (glioblastoma multiforme), [32] (pancreatic cancer), and [10] (protein kinases). Gene sequences and CpG island annotations were downloaded from the UCSC Genome Browser database, release hg18, except for Fig. 3, which used hg17 [59]. Where necessary, we used the UCSC liftOver tool to convert mutation coordinates from human genome build 35 to build 36. For type-specific analyses, tumors were classified by tissue of origin. Normalized gene expression data was obtained from GEO [60] series accession number GSE5764 [56].

Data Filtering

For analysis of data from [31-33], we obtained the list of genes sequenced by Wood *et al.* and their RefSeq identifiers from the table of PCR primers in supplemental online material [31, 53]. We discarded genes that are no longer in RefSeq (release 35) and 68 genes containing a total of 112 nonsynonymous mutations that could not be reconciled with sequences obtained from the UCSC Genome Browser database [59]. For gene models that overlap (share a portion of a coding exon), the gene model reported to contain a mutation was selected for analysis, or if no mutation was found, the gene model with the longest coding sequence was selected. Our final gene set contained 17180 non-overlapping genes present in RefSeq release 35 (March 2009). Genomic coordinates in some FASTA headers were corrected to reflect the fact that short intergenic sequences had been included. We removed all mutations from glioblastoma multiforme tumor BR27P, which had a 17-fold excess of mutations apparently resulting from chemotherapy [33]. The majority of mutation data for breast and colorectal cancer is from [9], which did not report synonymous substitutions. For compatibility across studies, synonymous substitutions were excluded from our analyses of data from [31-33], except when analyzing synonymous versus nonsynonymous substitution frequencies. Analyses of data from [31-33] include only those genes analyzed in [31]. We also excluded mutations in intronic splice sites. Analysis of mutation data from [9] for Fig. 3 used version 1 of the consensus coding sequence (CCDS) gene set [61] instead of RefSeq.

RefSeq [53] identifiers and sequences were obtained from the UCSC genome browser (release hg18) for 498 of 518 genes analyzed in [10]. Point mutation positions reported in [10] were manually curated to resolve inconsistencies between RefSeq sequences and

sequences downloaded from the Cancer Genome Project (CGP) database [62]. For five genes, we manually corrected apparent frameshift errors in the RefSeq sequences. Genes containing mutations that were inconsistent with CGP sequences were not included. Only coding sequence point mutations identified as part of the main screen were considered. We excluded data from gliomas and acute lymphocytic leukemias (because only the kinase domains were sequenced in these samples), from MMR-deficient tumors, and from cancer types with fewer than 20 point mutations after filtering, which left breast, colorectal, gastric, lung, melanoma, ovarian, and renal cancers. We also discarded 47 mutations for which the mutated nucleotide and two flanking nucleotides on each side did not all match in the RefSeq and CGP sequences or which had inconsistent codon positions in the two sequences. Of 1007 total mutations reported for the main screen in [10], 610 passed these filters.

Analysis

Custom software was written to determine nucleotide context and CpG island membership of bases in sequenced genes. Statistical analysis was performed using the R statistical package [63].

Nonsynonymous frequencies were calculated by dividing the number of mutations at second codon positions (at which substitutions are always nonsynonymous) by the number of codons sequenced. Synonymous mutation frequencies were calculated by dividing the number of mutations at four-fold degenerate sites (at which substitutions are always synonymous) by the number of four-fold degenerate sites sequenced. We exclude sites at which both synonymous and nonsynonymous substitutions are possible in order to avoid issues relating to rate differences between substitution types. Nonsynonymous and synonymous frequencies were calculated separately for each cancer type and dinucleotide context (CpG in island, CpG in 2kb shore, other CpG, TpC/GpA, or other). Overall nonsynonymous/synonymous frequency ratios were computed by taking a weighted average of frequency ratios for each nucleotide context, using as weights the inverse of the variance of each frequency ratio, calculated as described in section 3.1 of [64]. P-values for overall frequency ratios were calculated using a two-tailed Z-test.

Mutation frequencies for data from [9, 31-33] were calculated by dividing the observed number of nonsynonymous mutations by the number of sequenced sites of that type using discovery screen data. Mutation frequencies for data from [10] were calculated by dividing the observed number of synonymous and nonsynonymous mutations by the number of

sequenced sites of that type using main screen data. For frequency comparisons, 2x2 contingency tables were constructed with entries equal to the number of mutated and unmutated sequenced bases for each of the two frequencies being compared. P-values were calculated using Fisher's exact test. Confidence intervals were calculated using the normal approximation to the binomial.

Figures and Tables

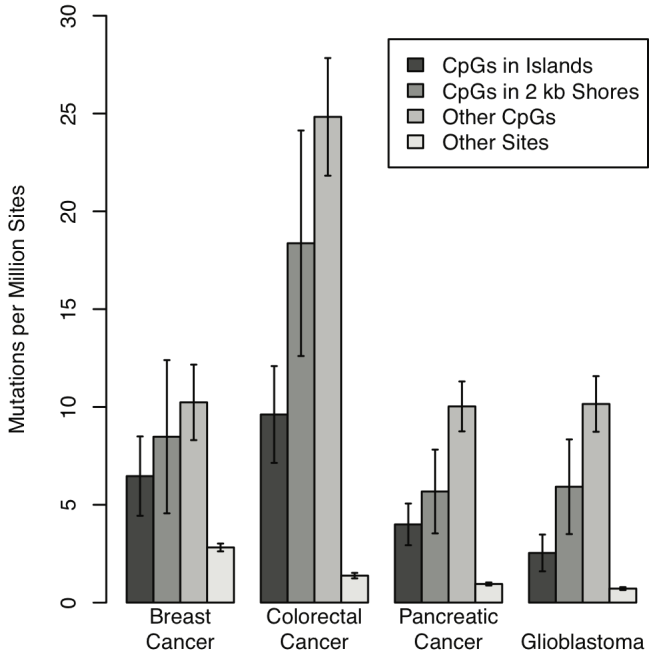


Fig. 1. CpG mutation frequencies in CpG islands, 2 kb CpG island shores, and remainder of gene. Frequencies were computed from discovery screen data in [31-33] by dividing the observed number of nonsynonymous changes at the site by the number of sequenced sites of that type. See table S3 for details. Error bars indicate 95% confidence intervals.

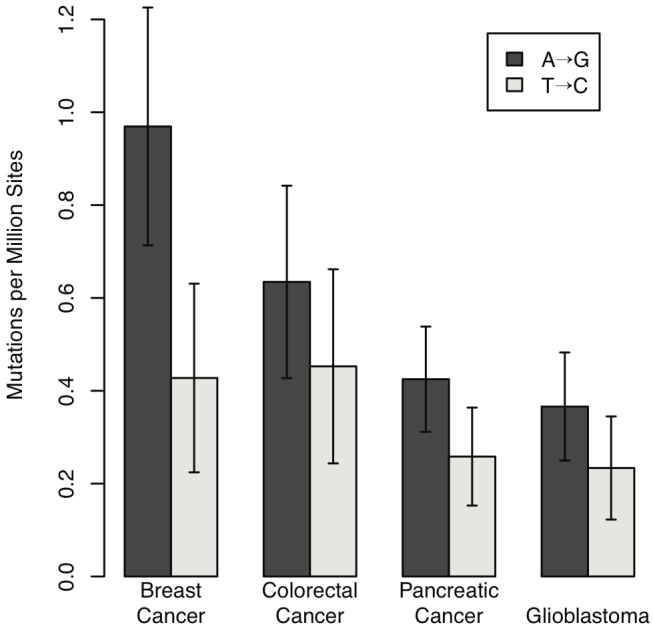


Fig. 2. A→G/T→C mutational asymmetry. Frequencies were computed as for Fig. 1. See table S5A for counts. Error bars indicate 95% confidence intervals.

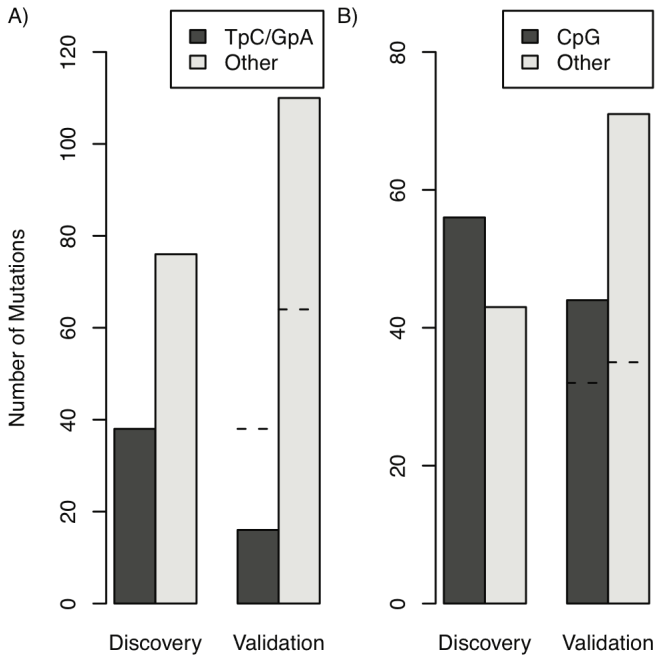


Fig 3. Screen differences in dinucleotide hotspot mutation proportion in (A) breast and (B) colorectal cancers. We used mutation data from [9], excluding known cancer genes (breast cancer: TP53; colorectal cancer: APC, KRAS, TP53, SMAD4, FBXW7), genes that were unmutated in one or both screens, and mutations in breast validation tumor BB23 because it was excluded from analysis in [31]. Dashed lines indicate expected number of mutations in the validation screen based on discovery screen frequencies, calculated by multiplying the number of bases of each dinucleotide context (CpG in island, CpG in 2kb shore, other CpG, TpC/GpA, or non-hotspot) sequenced in the validation screen by the appropriate discovery screen mutation rate. Only the C:G base pair in TpC/GpA dinucleotides is considered to be the hotspot. TpCpG and CpGpA trinucleotides were counted as CpGs. See table S6A for counts.

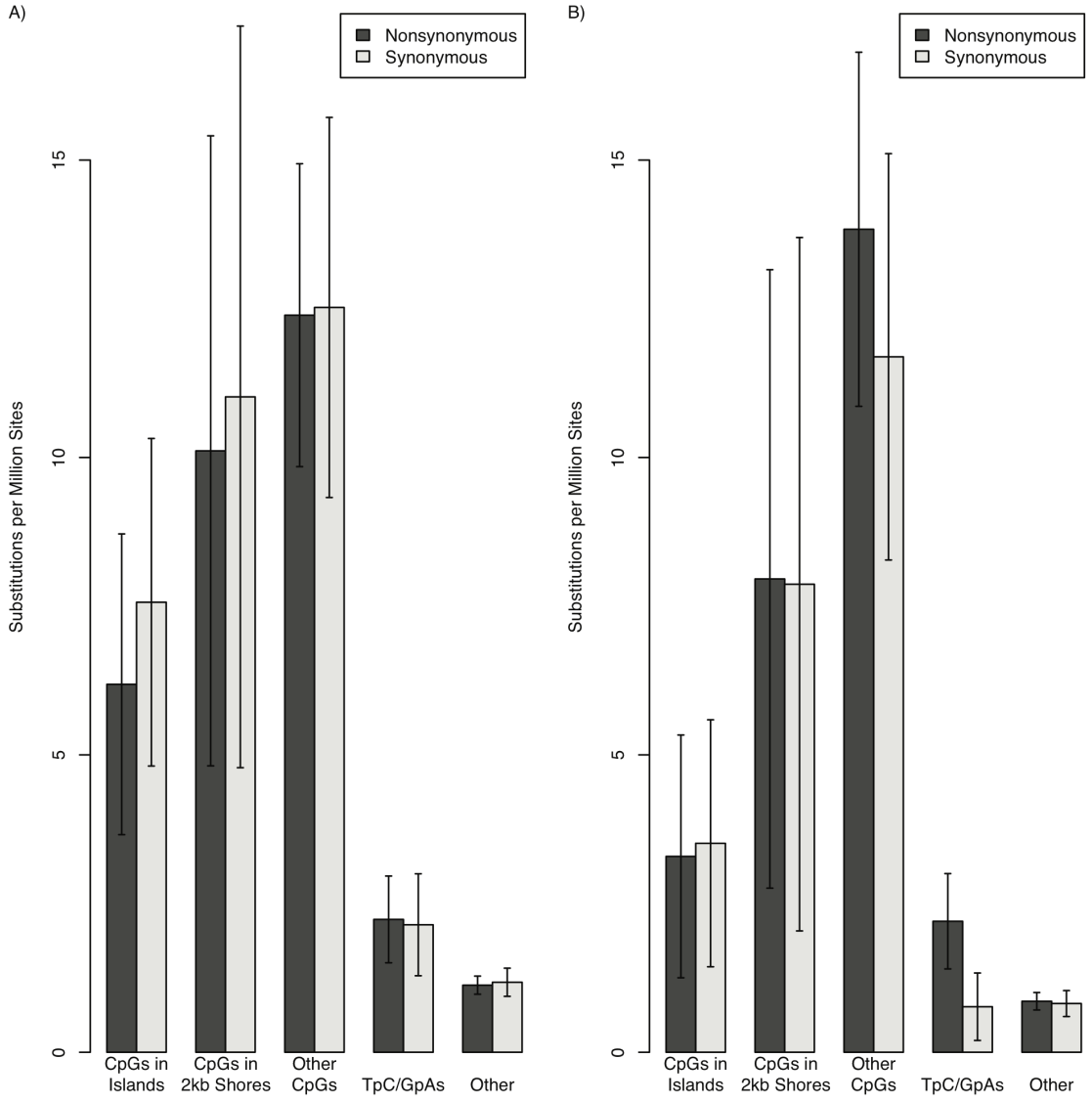


Fig. S1. Nonsynonymous and synonymous nucleotide substitution frequencies in (A) pancreatic cancer and (B) glioblastoma. Substitution data were obtained from discovery screens in [32, 33] and classified by dinucleotide context. The only significant frequency difference before correcting for multiple testing was TpC/GpA dinucleotides in glioblastoma, and this result was not significant after Bonferroni correction (uncorrected P=0.019, corrected P=0.19). See table S1 for counts. Error bars indicate 95% confidence intervals.

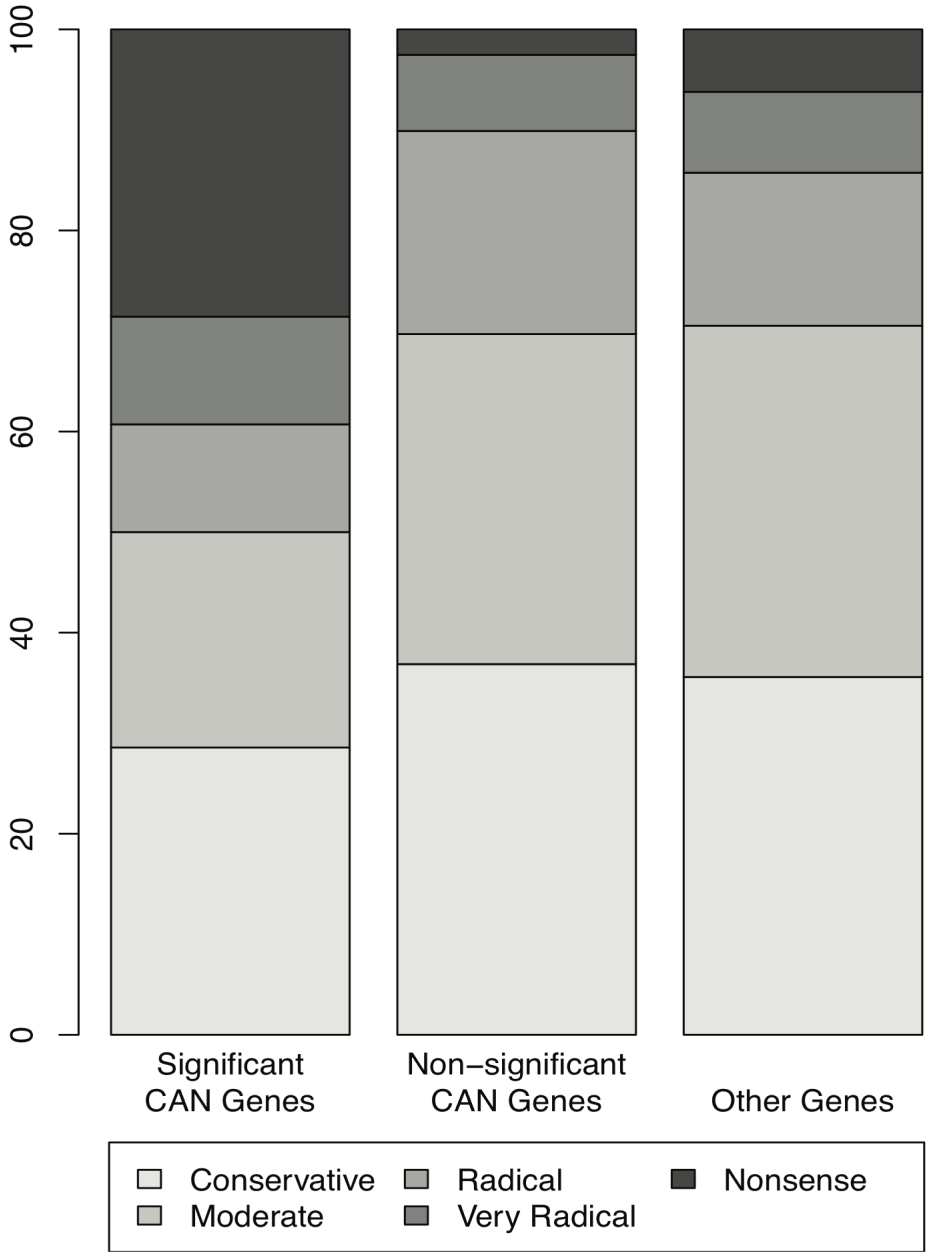


Fig. S2. Amino acid substitutions in data from discovery screen in [9]. Significant candidate cancer (CAN) genes are those that had significantly elevated mutation frequencies after reanalysis as described in [52]. Non-significant CAN genes are those originally reported as CAN genes [9] but not significant after reanalysis [52]. Shading indicates amino acid substitution severity category are as defined in [35, 65]. See table S2 for counts.

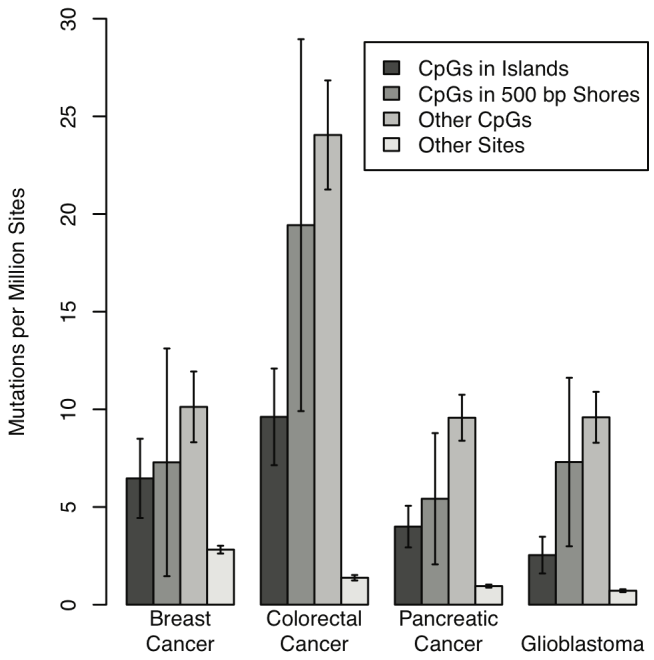


Fig. S3. CpG mutation frequencies in CpG islands, 500 bp CpG island shores, and remainder of gene. Frequencies were computed from substitution data from discovery screens in [31-33]. See table S3B for counts. Error bars indicate 95% confidence intervals.

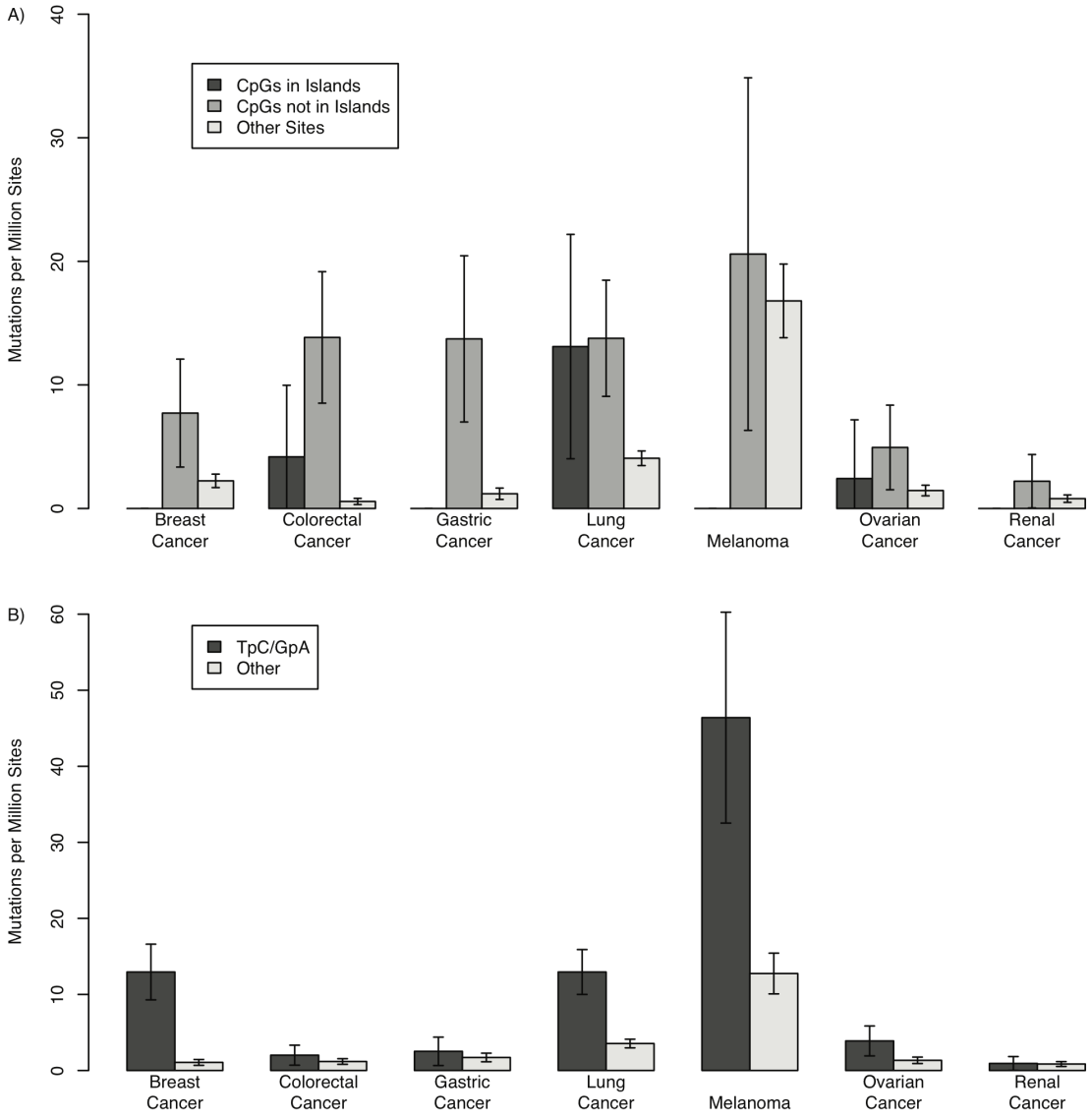


Fig S4. Mutation frequencies at (A) CpG and (B) TpC/GpA hotspots based on data from main screen in [10]. See table S4 for counts. Error bars indicate 95% confidence intervals.

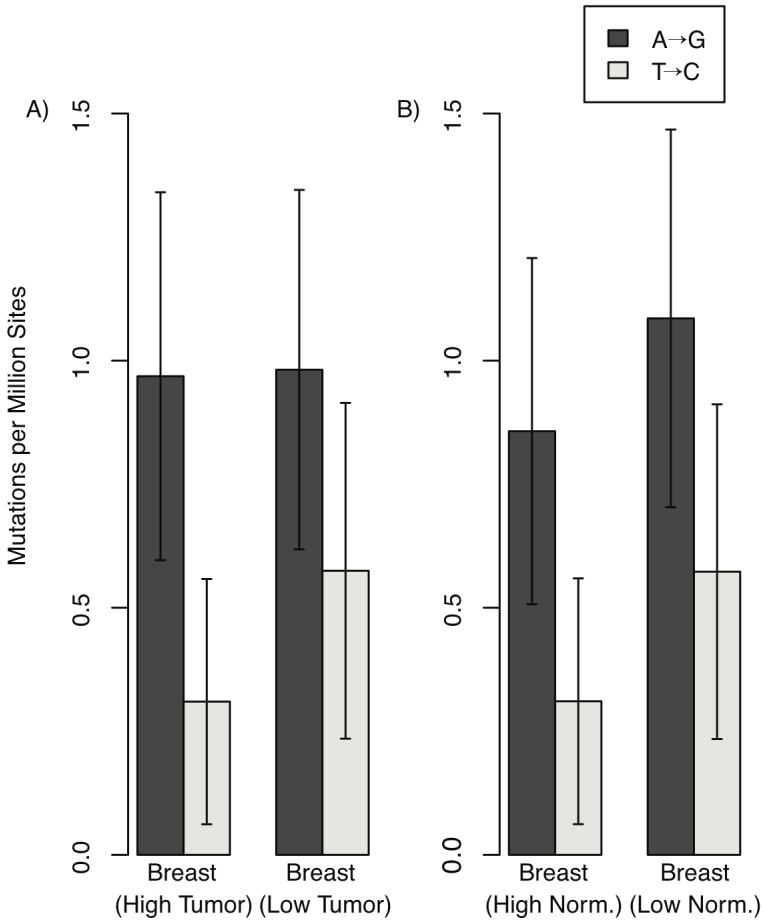


Fig. S5. A→G/T→C mutational asymmetry in breast cancer by (A) tumor and (B) normal tissue gene expression level, based on data from discovery screen in [31]. Genes were classified into higher than median and lower than median expression classes using expression data from ten breast cancer samples or twenty normal breast tissue samples [56]. See tables S5B and C for counts. Error bars indicate 95% confidence intervals.

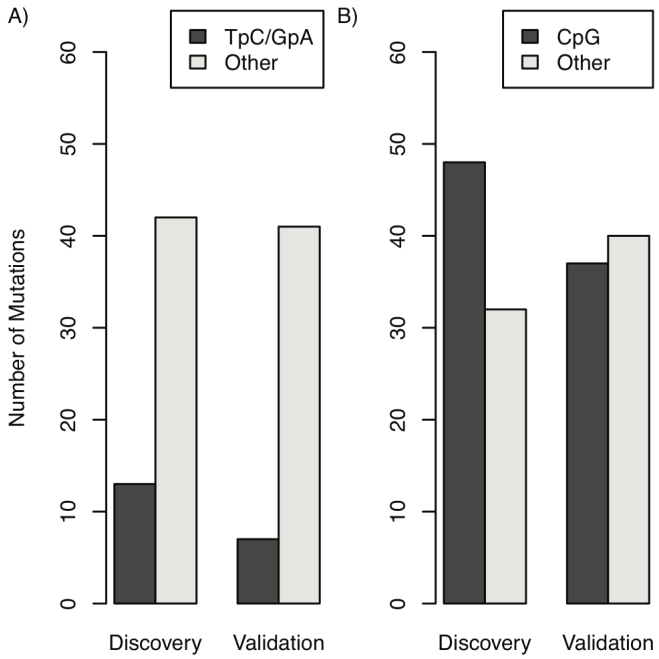


Fig S6. Dinucleotide hotspot mutation proportion by screen in (A) breast and (B) colorectal cancers. We used data from [31], excluding genes that were unmutated in one or both screens and genes that had been sequenced in [9]. Only the C:G base pair in TpC/GpA dinucleotides is counted as the hotspot. TpCpG and CpGpA trinucleotides were counted as CpGs. See table S6B for counts.

Table S1. Nonsynonymous and Synonymous Substitution Frequencies for (A) Pancreatic Cancer and (B) Glioblastoma

A)	Nonsynonymous Substitution Frequency†	Synonymous Substitution Frequency	Nonsynonymous/Synonymous Frequency Ratio
CpGs in Islands	6.187 (3.658-8.715)‡	7.566 (4.812-10.320)	0.818
CpGs in 2kb Shores	10.112 (4.815-15.408)	11.019 (4.785-17.253)	0.918
Other CpGs	12.392 (9.846-14.938)	12.522 (9.327-15.717)	0.990
TpC/GpA	2.233 (1.504-2.963)	2.143 (1.286-3.001)	1.042
Other Nucleotides	1.128 (0.976-1.280)	1.177 (0.940-1.413)	0.959

	Nonsynonymous Substitutions	Nonsynonymous Substitutions	Synonymous Non-subs.	Synonymous Non-subs.	P-value
CpGs in Islands	23	3717654	29	3832915	0.491
CpGs in 2kb Shores	14	1384535	12	1089022	0.846
Other CpGs	91	7343353	59	4711616	1.000
TpC/GpA	36	16121241	24	11197354	0.897
Other Nucleotides	211	187037633	95	80727634	0.755

B)	Nonsynonymous Substitution Frequency	Synonymous Substitution Frequency	Nonsynonymous/Synonymous Frequency Ratio
CpGs in Islands	3.293 (1.252-5.334)	3.513 (1.437-5.590)	0.937
CpGs in 2kb Shores	7.958 (2.759-13.157)	7.869 (2.040-13.698)	1.011
Other CpGs	13.837 (10.860-16.814)	11.692 (8.276-15.109)	1.183
TpC/GpA	2.202 (1.401-3.004)	0.765 (0.198-1.332)	2.878
Other Nucleotides	0.857 (0.711-1.004)	0.819 (0.600-1.037)	1.047

	Nonsynonymous Substitutions	Nonsynonymous Substitutions	Synonymous Non-subs.	Synonymous Non-subs.	P-value
CpGs in Islands	10	3036731	11	3130885	1.000
CpGs in 2kb Shores	9	1130944	7	889558	1.000
Other CpGs	83	5998323	45	3848631	0.415
TpC/GpA	29	13168448	7	9146441	0.010
Other Nucleotides	131	152779550	54	65941449	0.811

Substitution data were obtained from discovery screens in (1, 2).
†All mutation frequencies are reported as mutations per million sites.
‡95% confidence intervals for mutation frequencies are displayed in parentheses.

Table S2. Amino Acid Substitution Counts by Severity Class and P-values

	Conservative	Moderate	Radical	Very Radical	Nonsense
Significant CAN Genes†	8	6	3	3	8
Non-significant CAN Genes‡	73	65	40	15	5
Other Genes	332	326	142	75	58

	P-value
Significant CAN Genes vs. Non-significant CAN Genes	1.61×10^{-4}
Significant CAN Genes vs. Other Genes	2.96×10^{-3}
Non-significant CAN Genes vs. Other Genes	0.133

Substitution data were obtained from discovery screen in (3). Amino acid substitution severity categories are as defined in (5, 6).
†Significant candidate cancer (CAN) genes are those that had significantly elevated mutation rates after reanalysis of data from (3) as described in (4).
‡Non-significant CAN genes are those originally reported as CAN genes (3) but not significant after reanalysis.

Table S3. CpG Mutation Frequencies and Count Data for (A) 2kb and (B) 500bp CpG Island Shores

A)	CpG in Island Mutation Frequency†	CpG in Shore Mutation Frequency	Other CpG Mutation Frequency	Non-CpG Mutation Frequency
Breast Cancer	6.46 (4.44-8.49)‡	8.48 (4.56-12.39)	10.23 (8.30-12.17)	2.82 (2.62-3.02)
Colorectal Cancer	9.61 (7.14-12.09)	18.37 (12.60-24.13)	24.83 (21.82-27.84)	1.38 (1.24-1.52)
Pancreatic Cancer	4.00 (2.93-5.06)	5.68 (3.54-7.82)	10.03 (8.75-11.30)	0.95 (0.87-1.03)
Colorectal Cancer	2.54 (1.60-3.48)	5.92 (3.50-8.34)	10.15 (8.73-11.57)	0.71 (0.64-0.79)

	CpG Island Mutations	CpG Shore Mutations	CpG Island Non-mutations	CpG Shore Non-mutations	P-value
Breast Cancer	39	18	6032658	2123233	0.365
Colorectal Cancer	58	39	6032639	2123212	2.42×10^{-3}
Pancreatic Cancer	54	27	13511494	4755459	0.163
Glioblastoma	28	23	11036722	3884440	3.55×10^{-3}

	CpG Shore Mutations	Other CpG Mutations	CpG Shore Non-mutations	Other CpG Non-mutations	P-value
Breast Cancer	18	108	2123233	10551921	0.551
Colorectal Cancer	39	262	2123212	10551767	0.089
Pancreatic Cancer	27	237	4755459	23633344	3.77×10^{-3}
Glioblastoma	23	196	3884440	19304619	0.011

B)	CpG in Island Mutation Frequency	CpG in Shore Mutation Frequency	Other CpG Mutation Frequency	Non-CpG Mutation Frequency
Breast Cancer	6.46 (4.44-8.49)	7.29 (1.46-13.12)	10.13 (8.31-11.94)	2.82 (2.62-3.02)
Colorectal Cancer	9.61 (7.14-12.09)	19.43 (9.91-28.95)	24.05 (21.26-26.84)	1.38 (1.24-1.52)
Pancreatic Cancer	4.00 (2.93-5.06)	5.42 (2.06-8.78)	9.57 (8.39-10.75)	0.95 (0.87-1.03)
Colorectal Cancer	2.54 (1.60-3.48)	7.30 (2.99-11.62)	9.59 (8.29-10.90)	0.71 (0.64-0.79)

	CpG Island Mutations	CpG Shore Mutations	CpG Island Non-mutations	CpG Shore Non-mutations	P-value
Breast Cancer	39	6	6032658	823446	0.817
Colorectal Cancer	58	16	6032639	823436	0.018
Pancreatic Cancer	54	10	13511494	1844292	0.339
Glioblastoma	28	11	11036722	1506485	5.03×10^{-3}

	CpG Shore Mutations	Other CpG Mutations	CpG Shore Non-mutations	Other CpG Non-mutations	P-value
Breast Cancer	6	120	823446	11851708	0.586
Colorectal Cancer	16	285	823436	11851543	0.483
Pancreatic Cancer	10	254	1844292	26544511	0.080
Glioblastoma	11	208	1506485	21682574	0.491

Substitution data were obtained from discovery screens in (1, 2, 7).
†All mutation frequencies are reported as mutations per million sites.
‡95% confidence intervals for mutation frequencies are displayed in parentheses.

Table S4. Mutation Frequencies and Count Data for (A) CpG and (B) TpC/GpA Dinucleotide Hotspots

A)	CpG in Island Mutation Frequency†	CpG not in Island Mutation Frequency	Non-CpG Mutation Frequency
Breast Cancer	0	7.72 (3.35-12.1)‡	2.24 (1.69-2.78)
Colorectal Cancer	4.18 (0-9.97)	13.8 (8.52-19.2)	0.57 (0.32-0.82)
Gastric Cancer	0	13.7 (7.00-20.5)	1.19 (0.735-1.65)
Lung Cancer	13.1 (4.02-22.2)	13.8 (9.07-18.5)	4.06 (3.47-4.65)
Melanoma	0	20.6 (6.32-34.9)	16.8 (13.8-19.8)
Ovarian Cancer	2.42 (0-7.17)	4.94 (1.52-8.37)	1.45 (1.02-1.88)
Renal Cancer	0	2.21 (0.0442-4.37)	0.797 (0.496-1.10)

	CpG Island Mutations	CpG Island Sites	CpG Mutations	CpG Sites	Other Mutations	Other Sites
Breast Cancer	0	396048	12	1554288	65	29048616
Colorectal Cancer	2	478558	26	1878098	20	35100411
Gastric Cancer	0	297036	16	1165716	26	21786462
Lung Cancer	8	610574	33	2396194	182	44783283
Melanoma	0	99012	8	388572	122	7262154
Ovarian Cancer	1	412550	8	1619050	44	30258975
Renal Cancer	0	462056	4	1813336	27	33890052

B)	TpC/GpA Mutation Frequency	Other Mutation Frequency
Breast Cancer	12.9 (9.29-16.6)	1.06 (0.676-1.45)
Colorectal Cancer	2.01 (0.697-3.32)	1.18 (0.811-1.55)
Gastric Cancer	2.52 (0.653-4.38)	1.71 (1.14-2.28)
Lung Cancer	12.9 (10.0-15.9)	3.54 (2.97-4.11)
Melanoma	46.4 (32.5-60.3)	12.8 (10.1-15.4)
Ovarian Cancer	3.88 (1.92-5.85)	1.34 (0.912-1.76)
Renal Cancer	0.925 (0.0185-1.83)	0.848 (0.528-1.17)

	TpC/GpA Mutations	Other Mutations	TpC/GpA Non-mutations	Other Non-mutations	P-value
Breast Cancer	48	29	3706992	27291883	$<2.2\times10^{-16}$
Colorectal Cancer	9	39	4479331	32977688	0.176
Gastric Cancer	7	35	2780273	20468899	0.339
Lung Cancer	74	149	5714946	42074882	$<2.2\times10^{-16}$
Melanoma	43	87	926717	6822891	2.33×10^{-10}
Ovarian Cancer	15	38	3861485	28429037	1.05×10^{-3}
Renal Cancer	4	27	4324876	31840537	0.783

Substitution data were obtained from main screen in (8).
†All mutation frequencies are reported as mutations per million sites.
‡95% confidence intervals for mutation frequencies are displayed in parentheses.

Table S5. A→G/T→C Asymmetry Mutation Frequencies and Count Data for (A) Multiple Tissues and (B) Breast Cancer by Expression Level

A)	A→G Mutation Frequency†	T→C Mutation Frequency
Breast Cancer	0.969 (0.713-1.226)‡	0.428 (0.224-0.631)
Colorectal Cancer	0.635 (0.427-0.842)	0.453 (0.244-0.662)
Pancreatic Cancer	0.425 (0.312-0.538)	0.258 (0.153-0.364)
Glioblastoma	0.366 (0.250-0.483)	0.234 (0.123-0.345)

	A→G Mutations	T→C Mutations	A→G Non-mutations	T→C Non-mutations	P-value
Breast Cancer	55	17	56733572	39763294	2.52×10^{-3}
Colorectal Cancer	36	18	56733591	39763293	0.270
Pancreatic Cancer	54	23	127067337	89058627	0.049
Glioblastoma	38	17	103793478	72746502	0.133

B)	A→G Mutation Frequency	T→C Mutation Frequency
Breast Cancer (high expr.)	0.968 (0.596-1.341)	0.310 (0.062-0.558)
Breast Cancer (low expr.)	0.982 (0.618-1.345)	0.575 (0.235-0.914)

	A→G Mutations	T→C Mutations	A→G Non-mutations	T→C Non-mutations	P-value
Breast Cancer (High)	26	6	26846951	19350029	7.05×10^{-3}
Breast Cancer (Low)	28	11	28520284	19137703	0.143

Substitution data were obtained from discovery screen in (7). Genes were classified into high and low expression classes based on data from (9).
†All mutation frequencies are reported as mutations per million sites.
‡95% confidence intervals for mutation frequencies are displayed in parentheses.

Table S6. Dinucleotide Hotspot Count Data and P-values by Screen

A)	Breast Cancer		Colorectal Cancer	
	TpC/GpA	Other	CpG	Other
Discovery Screen	38	76	56	43
Validation Screen	16	110	44	71
Validation Screen (Expected)	38	64	32	35

	P-value	
	Breast Cancer	Colorectal Cancer
Discovery Screen vs. Validation Screen	1.73×10^{-4}	9.06×10^{-3}
Validation Screen vs. Validation Screen (Expected)	1.85×10^{-5}	0.217

B)	Breast Cancer		Colorectal Cancer	
	TpC/GpA	Other	CpG	Other
Discovery Screen	13	42	48	32
Validation Screen	7	41	37	40

	P-value	
	Breast Cancer	Colorectal Cancer
Discovery Screen vs. Validation Screen	0.320	0.151

A: Substitution data were obtained from (3), excluding known cancer genes (breast cancer: TP53; colorectal cancer: APC, KRAS, TP53, SMAD4, FBXW7), genes unmutated in one or both screens, and mutations in validation tumor BB23.
B: Data from (7), excluding genes sequenced in (3), and genes unmutated in one or both screens.

Chapter 4: Expression-based segmentation of the *Drosophila* genome

Abstract

Background

It is generally accepted that gene order in eukaryotes is nonrandom, with adjacent genes often sharing expression patterns across tissues, and that this organization may be important for gene regulation. Here we describe a novel method, based on an explicit probability model instead of correlation analysis, for identifying coordinately expressed gene clusters ('coexpression segments'), apply it to *Drosophila melanogaster*, and look for epigenetic associations using publicly available data.

Results

We find that two-thirds of *Drosophila* genes fall into multigenic coexpression segments, and that such segments are of two main types, housekeeping and tissue-restricted. Consistent with correlation-based studies, we find that adjacent genes within the same segment tend to be physically closer to each other than to the adjacent genes in different segments, and that tissue-restricted segments are enriched for testis-expressed genes. Our segmentation pattern correlates with Hi-C based physical interaction domains, but segments are generally much smaller that domains. Intersegment regions (including those which do not correspond to physical domain boundaries) are enriched for insulator binding sites.

Conclusions

We describe a novel approach for identifying coexpression clusters that does not require arbitrary cutoff values or heuristics, and find that coexpression of adjacent genes is widespread in the *Drosophila* genome. Coexpression segments appear to reflect a level of regulatory organization related to, but below that of physical interaction domains, and depending in part on insulator binding.

Background

Many factors contribute to genome organization, but one feature seen broadly across eukaryotes is that genes with similar patterns of expression often are physically clustered [21, 66]. The *S. cerevisiae* genome is enriched for pairs and triplets of coexpressed genes, which also often have shared function [17, 67, 68]. Essential genes also form clusters in

yeast, independently of coexpression clustering [69]. The ordering of coexpressed genes and essential genes in yeast is conserved over large evolutionary distances [69-71]. *Arabidopsis thaliana* also shows evidence of clustering by expression and by function [72, 73], but unlike in yeast, *Arabidopsis* clusters can be quite large, including up to 20 genes [73], and up to 10% of *Arabidopsis* genes belong to such clusters [74]. The nematode *C. elegans* has small coexpression clusters of 2-5 genes [18, 75] that are not attributable to operons [76]. Unlike other eukaryotes studied, tandem duplicates are heavily represented in *C. elegans* expression clusters [75].

Initial analyses in *Drosophila* described clusters of three or more tissue-specific genes, particularly for testis [77], and large domains of 10-30 coordinately expressed genes [19]. Subsequent statistical analyses indicate that the large domains are actually artifactual aggregates of smaller coexpression clusters, comprised of housekeeping genes and functionally coordinated genes [78], and experiments measuring the effect of chromosomal rearrangements that disrupt the large domains did not support the idea that they are important for controlling gene expression [79]. Evidence for conservation of expression clusters across *Drosophila* species is mixed. Genes within syntenic blocks are more likely to have correlated expression than expected by chance [80], and some regions show evidence of coevolution of expression [81]. However, other studies associate short intergenic distance and coexpression with higher rates of genomic rearrangement [78].

In mammals, housekeeping genes form clusters [16, 82], as do low-expression genes that are inactive in most tissues [83]. There is evidence of clustering of testis-specific genes in mouse [84]. In contrast to yeast, there is little evidence for clustering based on gene function in mammals [82]. A screen for mouse essential genes showed that they are enriched in certain chromosomal regions [85], although it is unclear if the genes in these clusters are coordinately expressed. Vertebrate coexpression clusters are thought to arise gradually over evolutionary time, and some are conserved between human and chicken [86], and human and mouse [87]. Clusters that include highly expressed genes are not more likely to be conserved than expected by chance [87], and linkage between highly expressed genes may in fact be deleterious [88].

Functionally coordinated gene clusters, which often overlap with coexpression clusters, are not conserved across eukaryotes, and the genes and functions that cluster differ widely across the species studied [89, 90].

The appreciation that genome location affects expression dates back to observations of differential expression of transgene insertions [91], but the mechanisms that maintain coexpression clusters remain unknown. Proposed mechanisms include LCR-mediated activity such as in the β-globin locus [92], sharing of proximal regulatory features [93], or regional enhancers [94]. Analyses in several species have shown that adjacent coexpressed genes tend be physically closer than the average [70, 71, 73, 95-97], but it is not known if this is required for coexpression. Insulator proteins are thought to help separate genomic regions into domains of activity or inactivity governed by long-range regulatory elements that affect many genes [98]. The insulator protein CTCF has been implicated in the creation and maintenance of chromatin loop domains [22]. Other experiments associate localization to the nuclear pore with increased expression [23], or proximity to the nuclear lamina with repression of transcription [99]. Recent advances in chromatin conformation capture and other methods for interrogating the three-dimensional structure of the nucleus allow characterization of physical contacts between genomic regions [100-104]. These studies provide evidence for interactions among neighboring genes, which may be related to gene coexpression.

Here we describe a novel method for identifying coexpression clusters and apply it to *Drosophila* expression data from a diverse set of tissues. In contrast to previous studies, we use an explicit probability model for segment-dependent gene expression that allows us to find a best-fitting partition of the genome into contiguous segments of coordinately expressed genes. Our approach avoids prior assumptions about segment size, the magnitude of coexpression effects, or other heuristics, and is based on parameters with natural mechanistic interpretations. We identify widespread small clusters of coexpressed genes and explore their properties. In particular we provide evidence for an association with physical interaction domains and insulator binding sites.

Results and Discussion

Expression Model

Previous work using correlation-based methods to identify clusters of coordinately expressed genes has had mixed success [17, 19, 20, 73]. Correlation-based results are strongly affected by the choice of arbitrary cutoffs that may over- or under-estimate coexpression and may lead to artifactual clustering [19, 78, 79]. We instead use an approach that is based on an explicit probability model for the observed expression data. The model assumes that the genome can be partitioned into contiguous groups of genes

(coexpression segments) such that the genes within a segment tend to have similar expression levels across tissues. Specifically, the (tissue dependent) expression value for a gene in a given segment is assumed to be the sum of a segment effect, which represents a regional effect on expression in a given tissue that influences all genes in the segment equally and represents shared regulation, and a gene-specific deviation, which reflects private regulation and 'noise' (stochastic or measurement). A segment may consist of one or many genes. We performed our analyses using microarray data, taking the steady-state mRNA abundance measured by these arrays as a proxy for transcriptional activity, however our method is easily adaptable to data from other technologies. Model details, and our procedure for finding an optimal segmentation of the genome, are described in Methods.

Simulations

We tested our analysis method by simulating 40 datasets each with 2000 genes and 27 tissues (comparable to the FlyAtlas [105] data for a single chromosome arm), using similar distribution parameters to those trained from the real data, and analyzing each simulated dataset. In all datasets, the parameter training converged to within 2% of the correct value, and the 'true' segmentation pattern used in the simulation was recovered exactly regardless of random starting segmentation. We also simulated data for alternative parameter sets, and found that is it robust to most parameter choices (Table S1).

Properties of *Drosophila* Expression Segments

We analyzed expression data from *Drosophila melanogaster* generated by the FlyAtlas project [105]. The FlyAtlas dataset samples 32 diverse tissues, of which we analyzed 27 after quality filtering, and 11363 genes. Optimal segmentation identification is reasonably robust (see Methods), and we chose the best scoring segmentation for followup analysis. Roughly two thirds of genes fall into multi-gene segments and thus appear to have coordinated expression with their neighbors across tissues. Multigenic segments have a mean of 3.1 genes (median 2.0) (Fig. S1A). To examine across-tissue expression patterns, we plotted the across-tissue means and standard deviations of segments with three or more genes. These segments cluster into two classes: one having low mean and high standard deviation (indicating highly variable expression across tissues), and the other having low standard deviation (suggesting "housekeeping" style expression) (Fig. 1). The 302 segments in the top quartile for standard deviation value tend to have highly tissue-restricted expression patterns, with mean expression that exceeds the dataset median in only a small number of tissues. For 119 of these, expression is restricted in this sense to a single tissue, and for 106 of the 119, the tissue is testis. This supports previous studies in

Drosophila and mouse showing that testis-expressed genes often form coexpression clusters [77, 84]. In contrast, segments expressed in non-testis tissues are often expressed in at least one other tissue (Fig. S2).

To identify segments with shared function as well as coexpression, we tested each segment for significant enrichment of GO Slim categories associated to its genes. Enriched segments are uncommon but more frequent than expected by chance (based on comparison to shuffled segmentation patterns) with 209 of 2442 multigenic segments having a significantly enriched term (P=0.00324) (Table S2).

We then looked for features that may illuminate mechanisms for the formation and maintenance of coexpression segments. Intergenic regions between segments are longer than intergenic regions within segments (P=1.39e-24by Kolmogorov-Smirnov test) (Fig. S1B and S1C). This length difference is consistent with previous work on coexpressed genes in *Drosophila* and other organisms [70, 71, 73, 95-97]. We verified that it is independent of repeats in the intergenic regions (Fig. S1D and S1E). This suggests that (perhaps not surprisingly) the mechanisms involved in establishing or maintaining coexpression may be less effective over longer distances.

We analyzed gene orientation for adjacent pairs of genes and found that two-gene segments are enriched for "head-to-head" gene pairs, which may be regulated by a bidirectional promoter [106], relative to pairs flanking intersegment regions (P=1.24e-6) or adjacent pairs within longer (three or more gene) segments (P=0.0021). 31.6% (390/1236) of all two-gene segments have head-to-head orientation, and 48.8% (1521/3118) of all head-to-head pairs lie within a segment, indicating that while head-to-head orientation may facilitate coexpression, it is neither required nor diagnostic (Table S3).

Physical interaction domains [104] represent an intriguing candidate mechanism for coexpression regulation. We find a highly significant sharing (P=2.28e-20) of segment and interaction domain endpoints, with 60.8% (571/939) of interaction domain endpoints also segment endpoints. However, segments are much smaller than interaction domains (mean sizes 1.8 genes vs. 10.3 genes) and only 49 interaction domains consist of a single segment.

Insulators may play a role in establishing interaction domain boundaries [104]. However, many insulator binding sites do not lie at interaction domain boundaries. We investigated the possibility that insulators may play a broader role in defining segments, using insulator ChIP-seq peak data generated by Nègre *et al.* [107]. We first confirmed that (consistent

with the results of Sexton *et al.*) peaks for BEAF-32, CP190, CTCF, GAF, and Mod(mdg4) are significantly enriched (per kilobase) in intersegment regions that do include a physical interaction domain boundary (by enrichment factors of 1.83, 1.54, 1.69, 1.54, and 1.42 respectively) and that this enrichment disappears after masking those peaks that overlap the 2kb windows centered on interaction domain boundaries as identified by [104] (Table S4). We then investigated intersegment regions that do not contain an interaction domain boundary. In the set of all such regions, we see no significant enrichment for insulator peaks. However, if we restrict to intersegment regions adjacent to long (three or more gene) segments, we find that the insulators BEAF-32, CP190, CTCF, and Su(Hw) are significantly enriched by factors of 1.30, 1.23, 1.19, and 1.23 (Table S4), as compared to the rest of the genome (excluding peaks overlapping the interaction domain windows). Thus it appears that insulators may play a role in defining coexpression segments, beyond their association with physical interaction domains. We also looked for insulator enrichment in intersegment regions adjacent to highly tissue-restricted segments, and found that only Su(Hw) is significantly enriched (by a factor of 1.21 (Table S4)), consistent with previous findings that Su(Hw) binds in regions where transcription is repressed in most tissues [108].

Conclusions

We developed a novel method, based on an explicit probability model, for identifying coexpression clusters that in contrast to previous approaches does not rely on arbitrary cutoffs or heuristics. We find that two thirds of *Drosophila* genes fall into multigene coexpression segments, that these segments are of two broad types, housekeeping and tissue restricted, and that clustering of genes expressed in a single tissue is largely confined to testis genes.

Adjacent genes within segments are physically closer to one another than adjacent genes in different segments. Our segmentation pattern is correlated with physical interaction domains [104], and with insulator binding, suggesting that coexpression segments may represent substructure within the interaction domains, and that they may be in part determined by insulator binding. Since coexpression segments are determined from expression data across diverse tissues, their association with physical interaction domains suggests that aspects of the domain structure may be shared between tissues. Although our analyses were confined to *Drosophila*, the observation that coexpression clusters across many eukaryotes tend to have similar properties [66], suggests that an association with insulator binding and physical interaction domains may hold more broadly.

Methods

Data Sources

Gene models were downloaded from Ensembl release 66 [109] and genomic sequence from dmel release 5 [110]. We performed an all-by-all BLATP search of annotated proteins in FlyBase 5.39 [110] to identify candidate paralogs, and found that 167 genes in the expression dataset have high identity (defined as greater than 50% amino acid identity over a 50 amino acid stretch) with another gene on the same chromosome, comprising 1.5% of all genes. No paralogs were removed from the analysis. Repeats were annotated using RepeatMasker [111].

Raw expression data were downloaded from NCBI GEO accession GSE7763 [105] in CEL format and normalized using RMA [112]. Probes were mapped to genes using the *drosophila*2.db annotation package in Bioconductor [113]. Genes with multiple probesets were assigned a single expression value by taking the median value for the probesets assigned to that gene. We computed the Pearson correlation across genes for each pair of tissue biological replicates and eliminated from further analysis any tissue for which this correlation was less than 0.98 for any biological replicate pair (Table S5). Tissue-specific expression values for each gene were taken to be the mean of the four biological replicate measurements.

Gene ordering for purposes of assigning to segments or determining gene adjacency was based on the annotated gene transcription start coordinates. The intergenic region between two adjacent genes is defined as the region between their annotated start coordinates; the intersegment region between adjacent segments is the intergenic region between the genes at the proximal segment ends.

Model Implementation

The probability model calculations were implemented as a custom C program that uses parts of the Gnu Scientific Library [114], the R Math Library [115], Argtable2 [116], LibDS [117], Bzip2 [118], and Jansson [119]. Programs for visualization and analysis of the segmentation patterns were implemented in Python and R. Software is available from A. R. by request.

Model Details

Our probability model for expression values involves, for each tissue type, a distribution $f(s)$ of segment effects, and a distribution $g(s)$ of gene-specific deviations. The expression

values we use are normalized microarray fluorescence intensities, but could in principle be derived from RNA-seq or other quantitative assays. The probability for a given segment's expression data in a single tissue is then:

$$\int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} g(x_i - s) \right) f(s) ds$$

where the x_i are the tissue-specific expression values for the n genes in the segment. We take f to be a mixture of two normal distributions, which provides a good fit to gene expression values over all tissues (Fig. S3), and g to be a normal distribution with mean 0:

$$f(s) = \phi \mathcal{N}(s; \mu_1, \sigma_1) + (1 - \phi) \mathcal{N}(s; \mu_2, \sigma_2)$$
$$g(x) = \mathcal{N}(x; 0, \sigma)$$

We assume independence of tissues and of segments, so the overall likelihood of a segmentation is a product of probabilities across tissues and segments.

The score (based on BIC) for a segmentation is the log likelihood modified by a parameter penalty that scales with the number of segments [120]:

$$\text{score} = -2 \ln(P(x|\theta)) + K \ln(n)$$

where x is the set of observed expression values, θ is the set of parameters for f and g, K is the number of estimated parameters (lengths of all segments, and distribution parameters), and n is the number of data points in the expression dataset (genes by tissues).

Model Estimation

Finding a best-fitting genome segmentation model for a given expression dataset is challenging, because it requires in principle searching the Cartesian product of the space of all possible genome partitions into segments with the space of parameters for the distributions f and g. We structure this as a search of the parameter space (carried out using the Simplex algorithm as implemented in the GNU Scientific Library [114]), in which the score associated to each set of parameter values is computed by optimizing over segmentations. Each chromosome arm is analyzed separately. For particular values of f and g parameters, we search the segmentation space as follows. First, we partition the chromosome arm into segments of random lengths (*i.e.* number of genes), drawn from a geometric distribution having (by default) a mean of 2 genes (in practice, the choice of

mean has a negligible impact on the segmentation patterns that the model converges to). We then consider three possible types of "move": split, which divides a multigenic segment into two segments; merge, which combines two adjacent segments into a single segment; and shift, which changes the boundary between two existing segments by expanding one and shrinking the other, such that at least one gene remains in each segment. Given a segmentation pattern, we evaluate each possible move and select the one that gives the greatest score improvement. The process is iterated until a segmentation is reached for which no moves improve the score. Because this search is strictly downhill, we consider multiple random initial segmentations ('replicates'), generally 1024, and carry out the above search for each of them. This yields a set of 1024 "locally optimal" segmentations; the median of their scores is then taken as the score value assigned to the specified parameter values for purposes of the parameter space search. Our analysis software also supports using the best replicate score or mean replicate score, but exploratory analyses indicated the median gave the most robust results. The best-scoring segmentation that is found with the best-scoring analysis parameters is used for subsequent analysis.

For convenience and computational speed, we made several simplifying assumptions regarding *f* and *g.* First, we assume that a single *f* and a single *g* (per chromosome arm) apply to all tissues, i.e. we do not allow tissue dependent parameter values. Second, we assume that *f* may be estimated as the mixture of normals that best fits the observed distribution of gene expression values over all tissues (Fig. S3). This *f* is found using an EM algorithm implemented in the PyMix package [121], and fixed during subsequent analysis. Thus only the parameter σ that defines the distribution of gene deviations *g* is estimated iteratively.

Simulations

Because each chromosome arm is analyzed independently in our real-data analyses, our simulated datasets each consist of a single simulated chromosome arm. To simulate a dataset with a given number of genes and tissues, we first simulate a segmentation by drawing segment lengths (*i.e.* number of genes) randomly from a geometric distribution with a specified mean until all genes have been assigned. We then simulate expression data for each tissue that conforms to the assumptions of our probability model for a specific choice of *f* and *g*, as follows. For each segment and tissue, a segment effect is drawn randomly from *f*, and for each gene in that segment, a gene-specific deviation is drawn from *g*. These are added to get the gene expression value. For simulations where *g* varies across segments and tissues, an independent σ is drawn for each draw from *f*. Our analysis of

simulated datasets used 512 replicates (starting random segmentations) per round of parameter training.

Robustness of Real Data Estimates

The spread of replicate scores for the optimal parameter values for each chromosome arm is much wider for the real data than for the simulated data (Fig. S4), and the best scoring segmentation is only found in one replicate for each chromosome arm. This suggests that the score surface for the real data is more complex than that for the simulated data (where the best scoring segmentation was found repeatedly). However, we find that the best-scoring replicates share 94.3% of their segment endpoints with the endpoints in the second-best replicates, and 91.7% of their segment endpoints with the endpoints in the worst-scoring replicates. Moreover, 84% of segments found in at least one replicate appear in more than half of the replicates. Thus, despite some variability in exact segmentation and score, the replicates are highly similar, implying that and our method is reasonably robust to the choice of starting segmentation and that for most of the genome our model finds the same local segmentation regardless of the starting pattern.

Other Analysis Procedures

For the promoter-orientation analysis, we counted the number of adjacent gene pairs in the dataset with the same orientation, "head-to-head" opposite orientation, or "tail-to-tail" opposite orientation for three classes: pairs within two-gene segments, pairs within larger multigenic segments, and intersegment pairs. P-values for comparing two classes were calculated using a 2x3 Chi-squared test.

Nègre *et al.*'s [107] insulator peaks were converted from dm3 to dm5 using FlyBase's coordinate conversion tool [110]. We removed 571 (of a total of 35365) insulator peaks (1.6%) that could not be converted to dm5 coordinates due to assembly incompatibilities. We counted the number of ChIP peaks that overlap regions of a given type using BEDTools [122] and Pybedtools [123], and converted these to peaks per kilobase by dividing by the total size of the regions. Enrichment was calculated as the ratio of the peaks per kilobase values for two specified region types. Significance for comparing two sets of regions was determined by Fisher's exact test, for a 2x2 table in which the first cell in each row gives the number of peaks overlapping regions of the given type, and the second cell gives the number of "non-peaks" of the same size as peaks, defined as and the number of bases in the regions minus the number of peaks, divided by the average peak size. For analyses of intersegment regions for a particular type of segment (e.g. multigenic segments), we

consider regions that border a segment of that type as belonging to the analyzed set. Some analyses exclude the subset of peaks that overlaps the 2kb windows identified by Sexton *et al.* as marking interaction domain boundaries.

Sexton *et al.*'s [104] physical interaction domain coordinates were converted from dm3 to dm5 using FlyBase's coordinate conversion tool [110]. We removed 12 domains that could not be converted to dm5 coordinates due to assembly incompatibilities. Genes were assigned to interaction domains based on the position of their annotated start site. We tested for significant sharing of endpoints between coexpression segments and interaction domains by performing a Fishers' exact test on the 2x2 table with cell counts giving the number of intergenic regions (*i.e.* regions between the starts of adjacent genes) that are: (i) segment endpoint and interaction domain endpoint, (ii) segment endpoint only, (iii) interaction domain endpoint only, or (iv) neither segment nor interaction domain endpoint.

Our gene Ontology analysis used generic GO Slim [124]. GO term enrichments were calculated using goatools [125]. Segments in which only one gene was annotated with the enriched term were removed from the list of significant results for that term. We compared the number of segments with one or more enriched terms to the number of segments with one or more enriched terms in 10 shuffled segmentation patterns using Fisher's exact test. Shuffled segmentations were generated by preserving chromosome gene order while randomly permuting the list of segment lengths and requiring that no segment endpoints were shared between the random segmentation and the real segmentation.

Figures and Tables

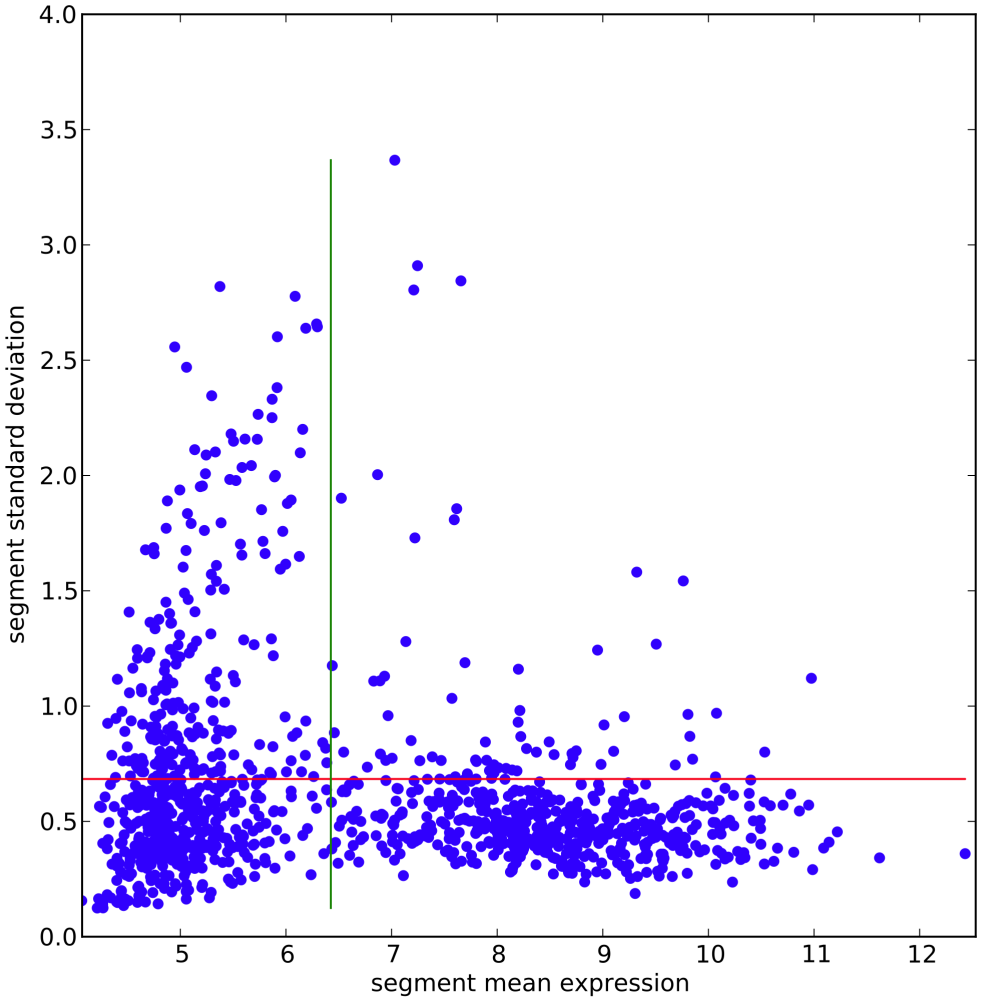


Fig 1. Scatterplot of across-tissue mean expression vs. across-tissue standard deviation reveals two classes of segments. Segments with three or more genes are plotted. The tissue-specific expression value of a segment is taken to be the average of its component genes' values; the mean and standard deviation across tissues of each segment's values are the coordinates for the plotted point. The red line denotes the cutoff for the top quartile of segments by standard deviation. The green line denotes the median gene expression value across all genes and tissues. Segments close to the X-axis have similar expression values for every tissue; those close to the Y-axis have high expression in a minority of tissues.

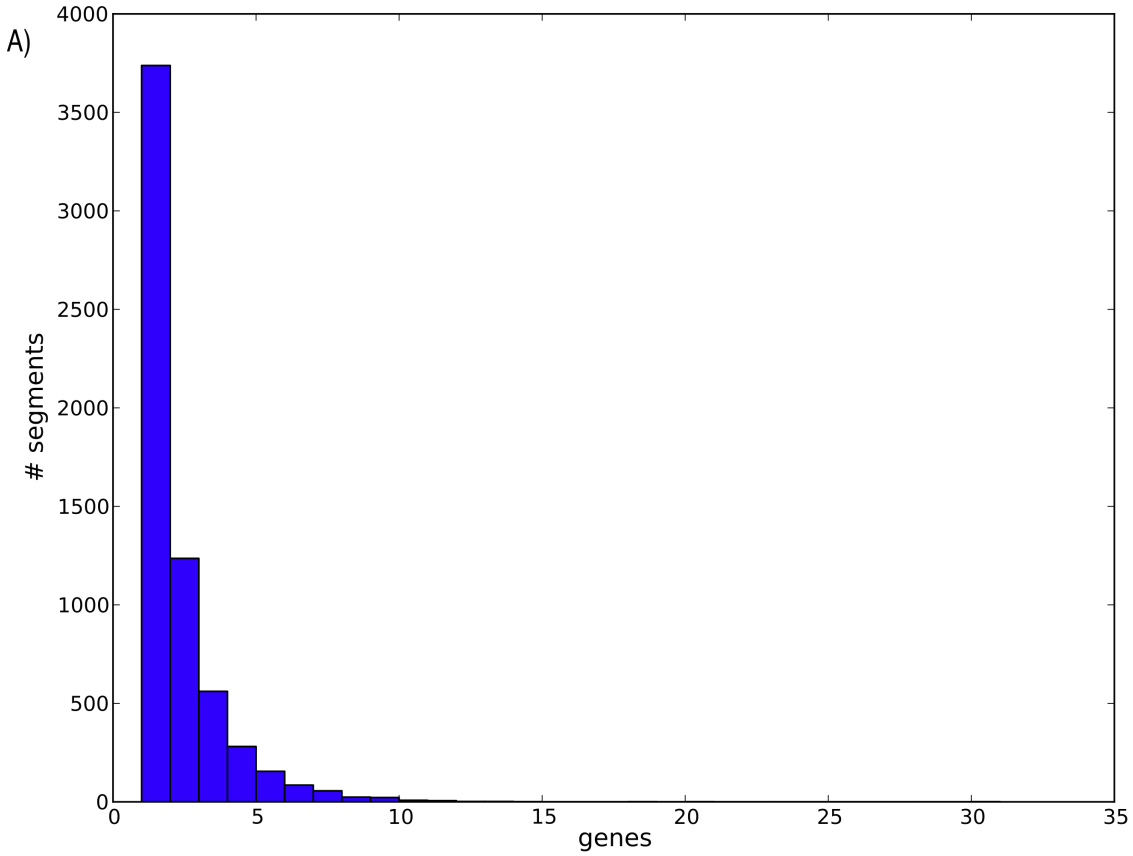
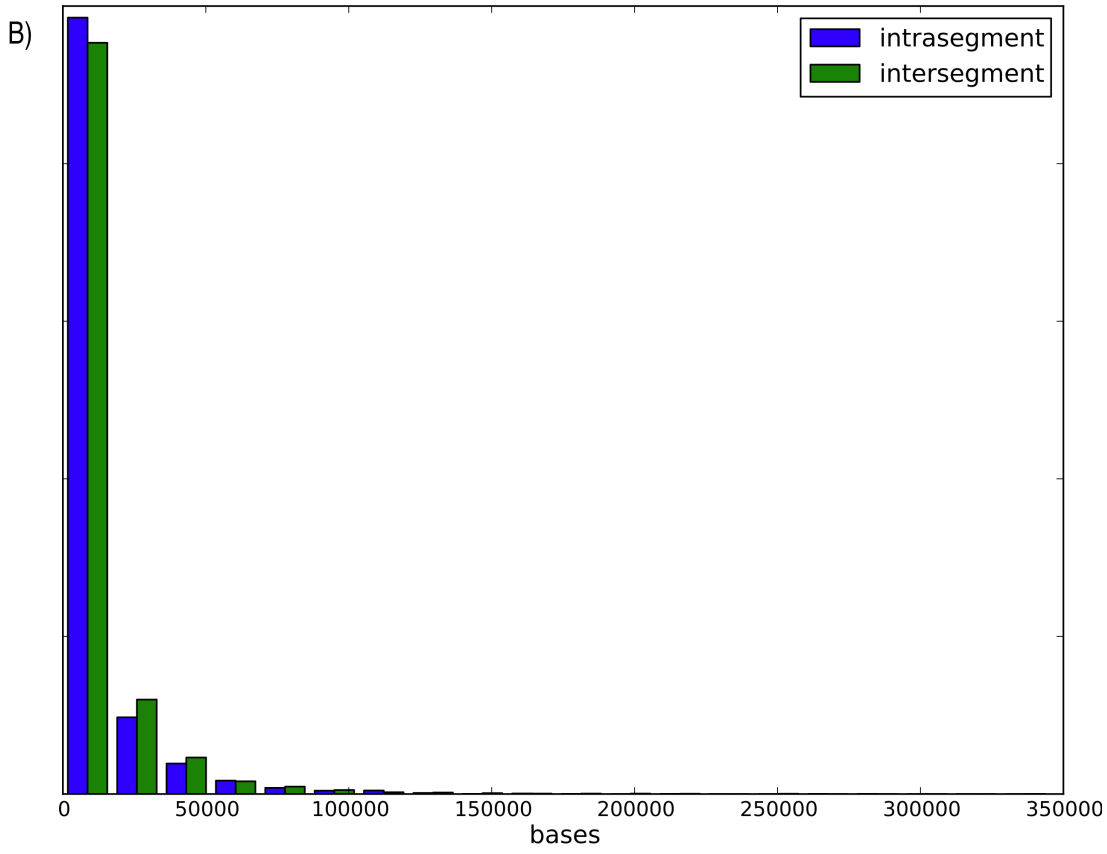
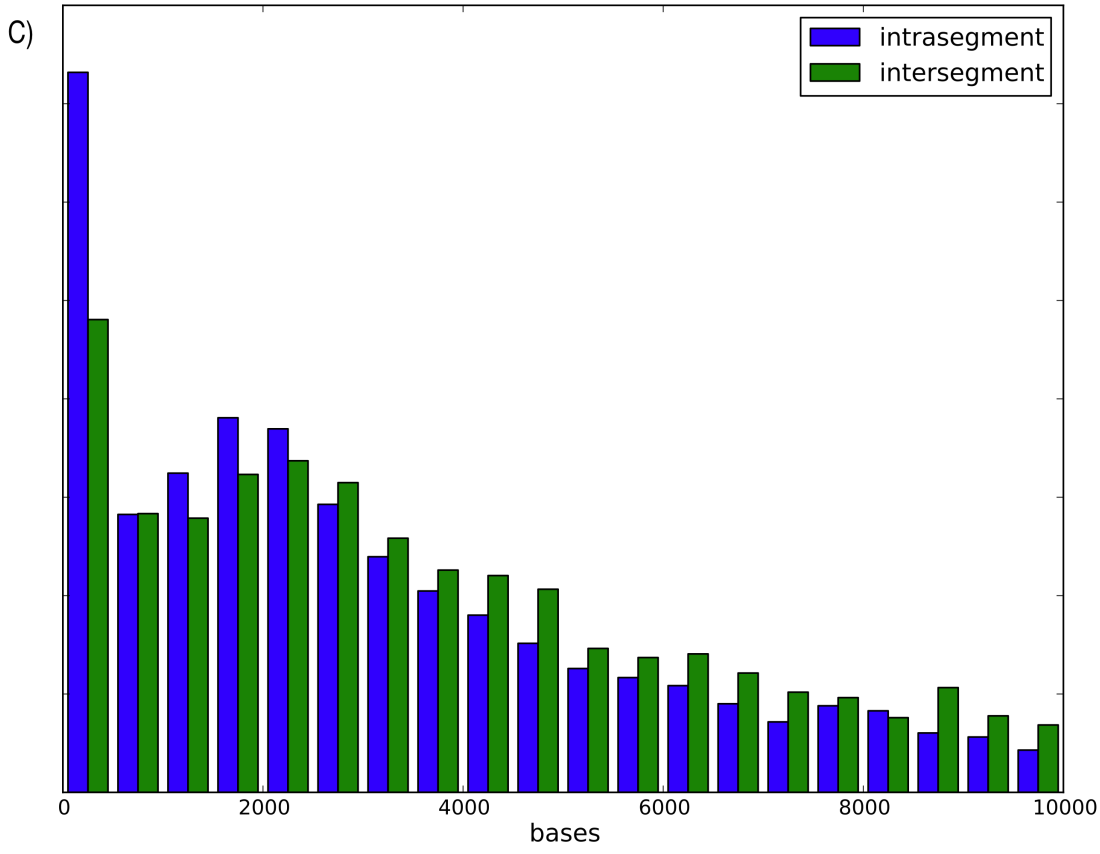
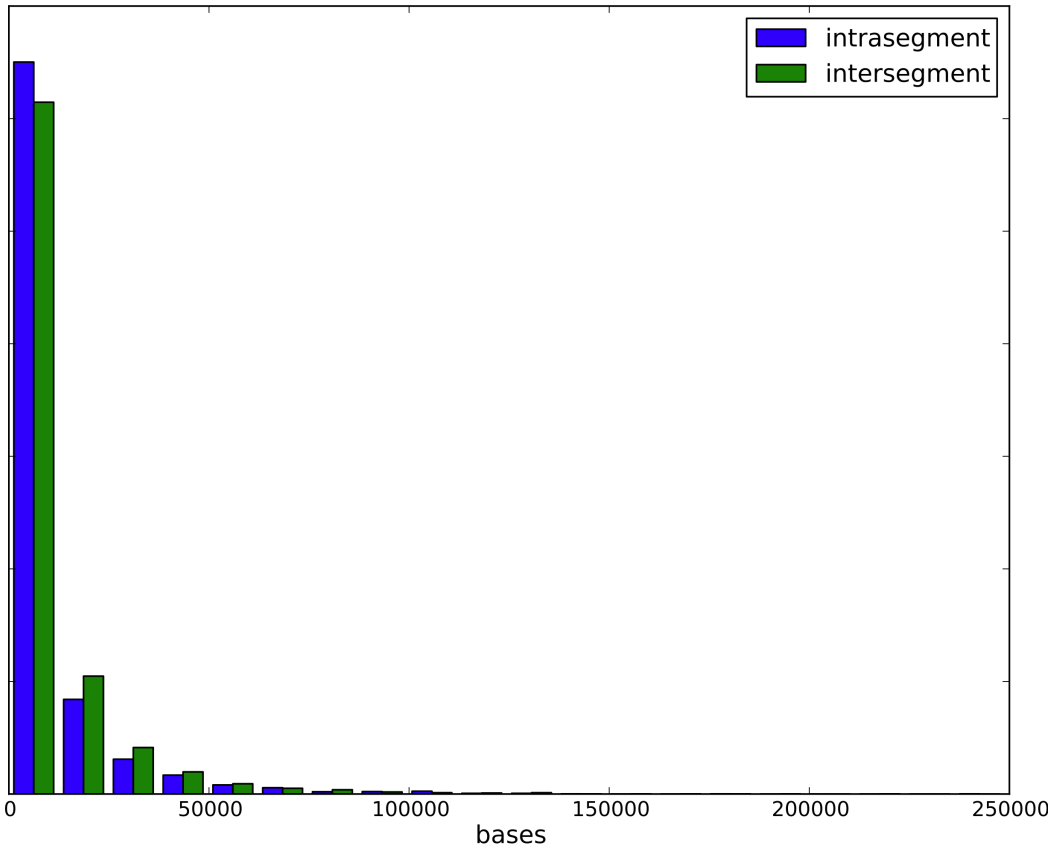


Fig. S1. Length distributions. (A) Histogram of segment lengths (# genes) for the best scoring segmentation pattern. (B and C) Normalized histogram of all intergenic lengths (in base pairs) within segments and between segments (B) and lengths below 10kb (C). Intersegment lengths are significantly longer than intergenic lengths within segments (P=1.39e-24). (D and E) Normalized histogram of all repeat-masked intergenic lengths (in base pairs) within segments and between segments (D) and lengths below 10kb (E). Repeat-masked intersegment lengths are significantly longer than repeat-masked intergenic lengths within segments (P=2.21e-24).

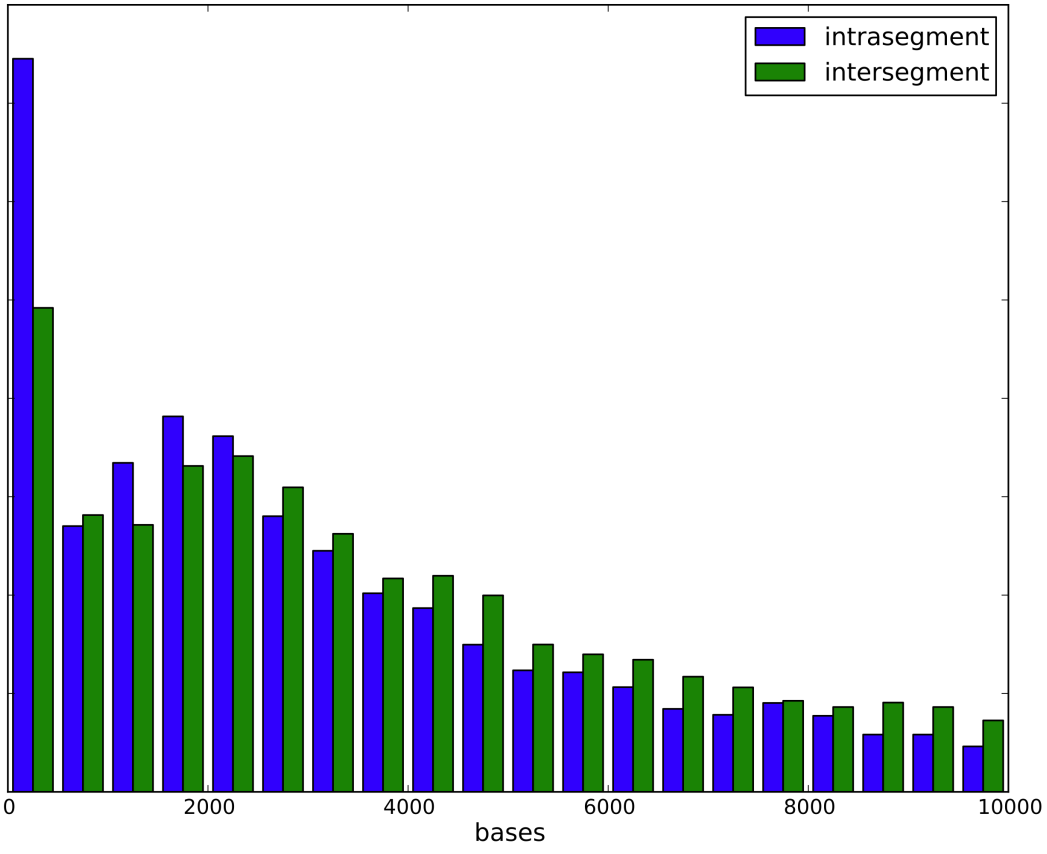

B)
intrasegment
intersegment
0
50000
100000
150000
200000
250000
300000
350000
bases


C)
intrasegment
intersegment
0
2000
4000
6000
8000
10000
bases

D)


intrasegment
intersegment
0
50000
100000
150000
200000
250000
bases

E)


intrasegment
intersegment
0
2000
4000
6000
8000
10000
bases

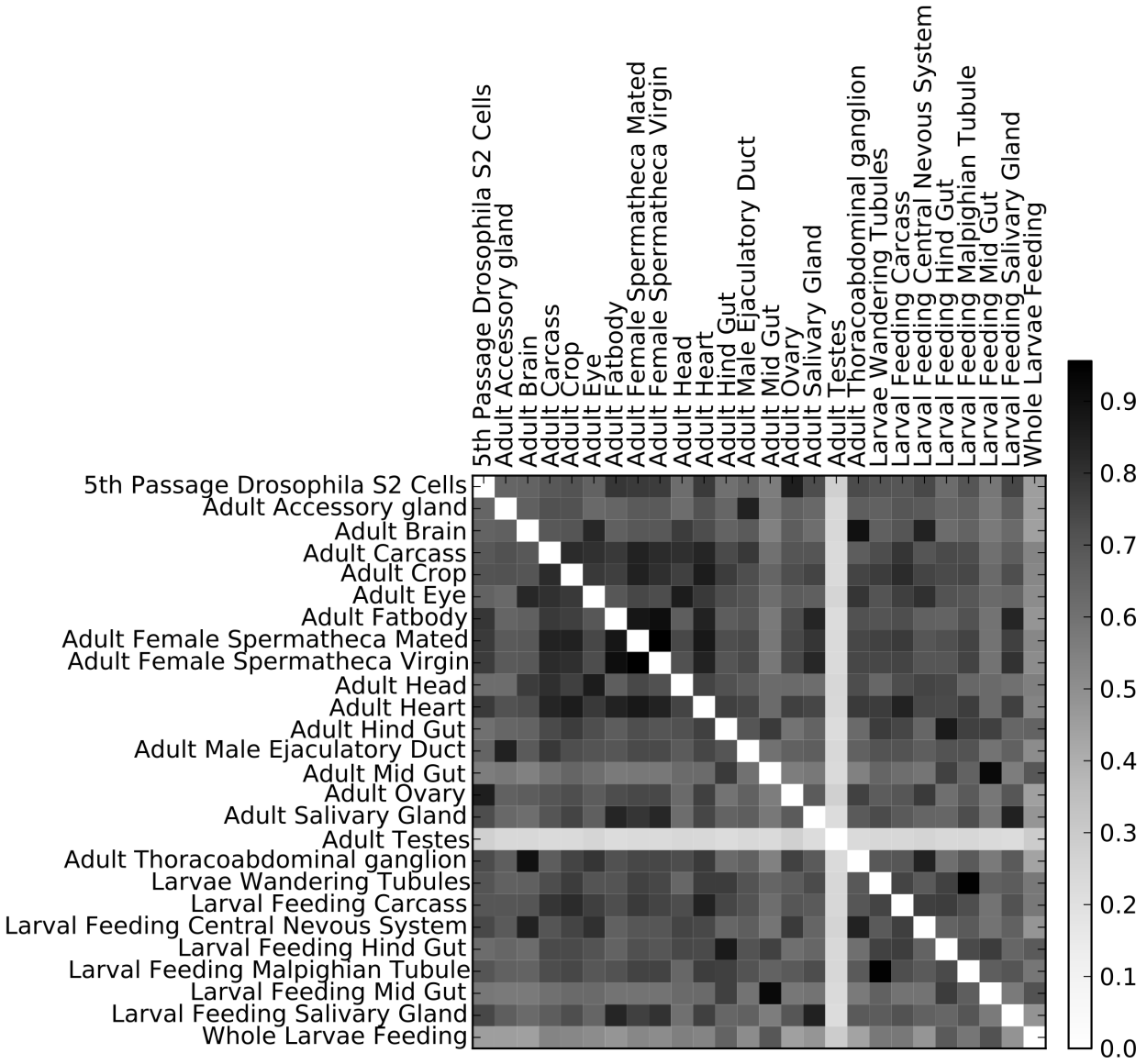
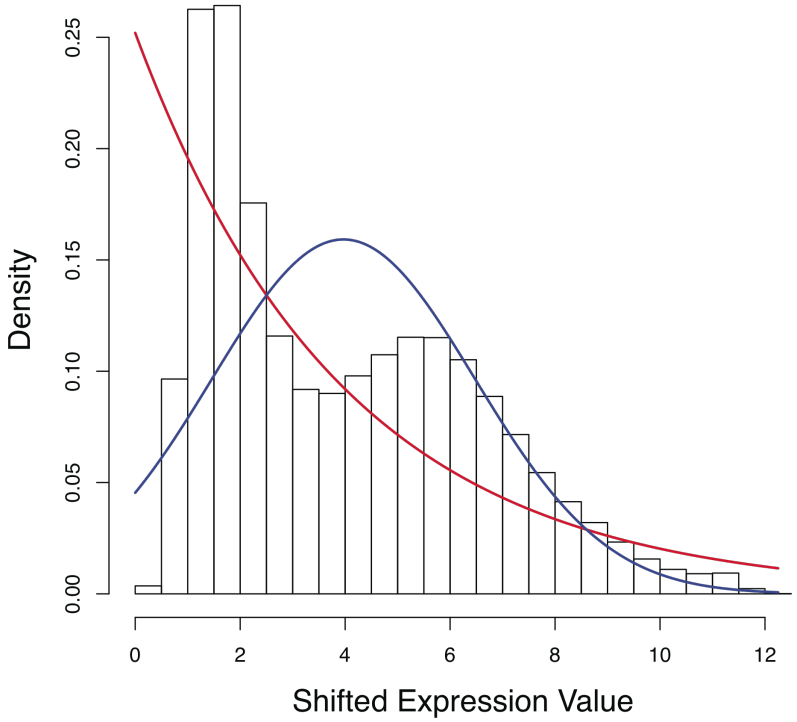
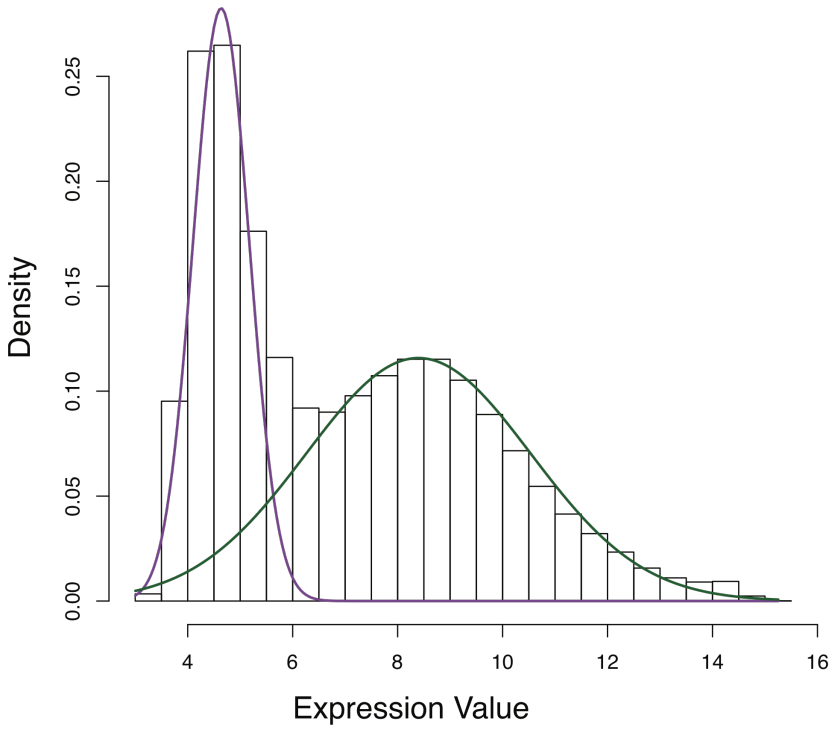


Fig. S2. Tissue coexpression. Cell color indicates ratio of the number of segments in which both tissues are expressed to the number of segments in which either tissue is expressed.

A)


Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
Shifted Expression Value

B)


Density
0.00
0.05
0.10
0.15
0.20
0.25
4
6
8
10
12
14
16
Expression Value

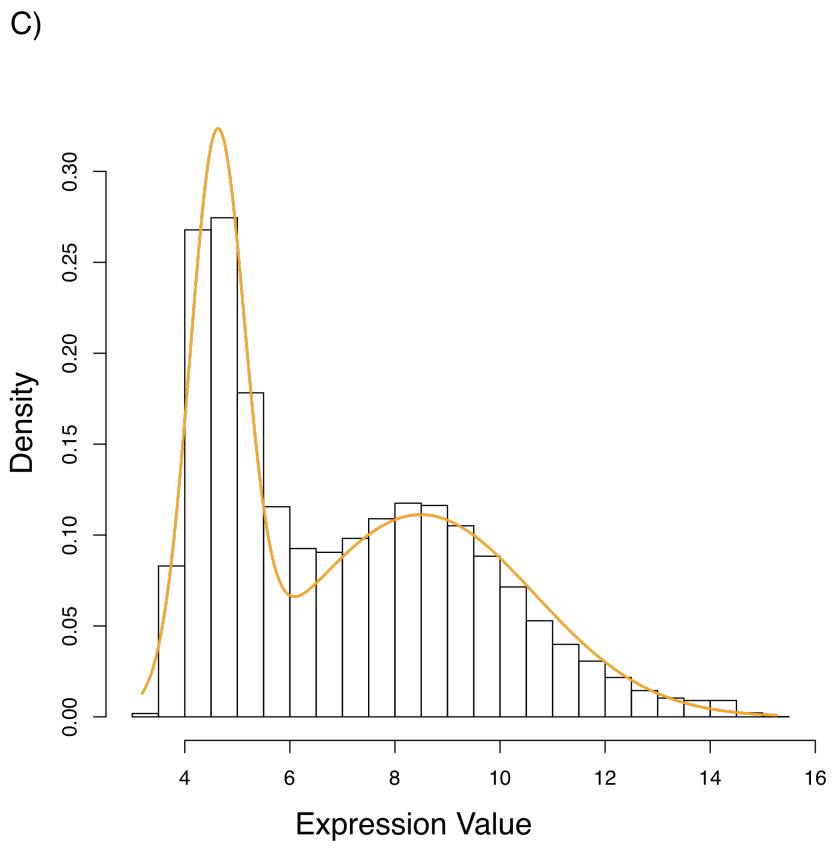


Fig. S3. FlyAtlas expression data, with best fitting distributions of various types. (A) Best fitting normal distribution (blue) and exponential distribution (red). To visualize the exponential fit, expression values are linearly shifted such that the smallest expression value is zero. (B, C) Mixture of two normal distributions (components: green and purple; sum: orange). This was taken as the segment effect distribution.

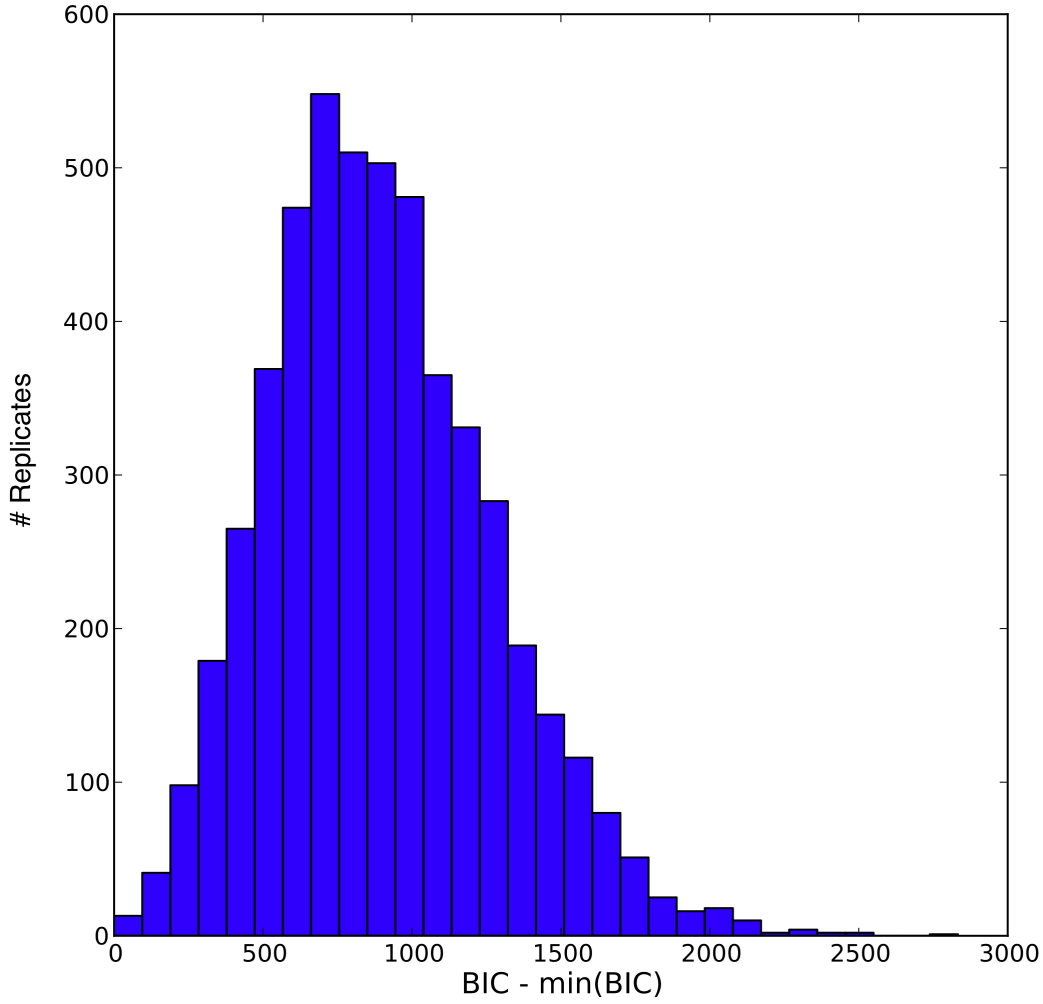


Fig. S4. Replicate score distribution. The scores of the optimal segmentations found by starting from 1024 random starting segmentations for each chromosome arm using the optimal parameter values are plotted. Horizontal axis indicates the difference between the replicate score and the best score found (lower is better).

Table S1 Simulated data robustness

Mean size (simulation)	Mean size (replicates)	Sigma	Sigma mean	Sigma std. dev.	Matching replicates per dataset
2	2	1.00			512
2	5	1.00			512
2	10	1.00			512
5	2	1.00			512
5	5	1.00			512
5	10	1.00			512
10	2	1.00			512
10	5	1.00			512
10	10	1.00			512
2,5,10	2	1.00			512
2,5,10	5	1.00			512
2,5,10	10	1.00			512
2	2		1.00	0.10	512
2	5		1.00	0.10	512
2	10		1.00	0.10	512
5	2		1.00	0.10	512
5	5		1.00	0.10	512
5	10		1.00	0.10	512
10	2		1.00	0.10	512
10	5		1.00	0.10	512
10	10		1.00	0.10	512
2,5,10	2		1.00	0.10	512
2,5,10	5		1.00	0.10	512
2,5,10	10		1.00	0.10	512
2	2		1.00	0.20	512
2	5		1.00	0.20	512
2	10		1.00	0.20	512
5	2		1.00	0.20	512
5	5		1.00	0.20	512
5	10		1.00	0.20	512
10	2		1.00	0.20	512
10	5		1.00	0.20	512
10	10		1.00	0.20	512
2,5,10	2		1.00	0.20	512
2,5,10	5		1.00	0.20	512
2,5,10	10		1.00	0.20	512
2	2		1.00	0.50	25
2	5		1.00	0.50	21
2	10		1.00	0.50	8
5	2		1.00	0.50	29
5	5		1.00	0.50	50
5	10		1.00	0.50	17
10	2		1.00	0.50	9
10	5		1.00	0.50	12

10	10	1.00	0.50	40
2,5,10	2	1.00	0.50	19
2,5,10	5	1.00	0.50	46
2,5,10	10	1.00	0.50	20

We simulated 10 datasets with the fixed sigma parameter for each mean size (40 total), as described in the text. Additional datasets were simulated to test our assumption that modeling a single value of sigma for all tissues and all segments is sufficient

Table S2 Segments with enriched GO Slim terms

Genes	GO Term	Annotation	P-value (FDR-corrected)	Segment genes with term	Segment length
FBgn0029095,FBgn0020305, FBgn0053526	GO:0043226	organelle	0.0445	3	3
FBgn0031294,FBgn0003310, FBgn0015905	GO:0005886	plasma membrane	0.0201	2	3
FBgn0024314,FBgn0051922, FBgn0031317,FBgn0031318, FBgn0031319,FBgn0031320, FBgn0031321,FBgn0031322	GO:0003729	mRNA binding	0.0249	2	8
FBgn0051668,FBgn0053124	GO:0055085	transmembrane transport	0.00155	2	2
FBgn0051668,FBgn0053124	GO:0006810	transport	0.00412	2	2
FBgn0031403,FBgn0051679, FBgn0052463,FBgn0051682	GO:0005739	mitochondrion	5.10E-05	3	4
FBgn0031403,FBgn0051679, FBgn0052463,FBgn0051682	GO:0043226	organelle	0.0299	3	4
FBgn0031407,FBgn0054049	GO:0005576	extracellular region	0.00233	2	2
FBgn0031423,FBgn0031424, FBgn0031426,FBgn0051948, FBgn0085477	GO:0006629	lipid metabolic process	0.00784	2	5
FBgn0031457,FBgn0028398, FBgn0026324	GO:0005737	cytoplasm	0.0248	2	3
FBgn0031462,FBgn0031463, FBgn0031465,FBgn0031464, FBgn0031466,FBgn0031467, FBgn0031468,FBgn0031469, FBgn0031470,FBgn0031471, FBgn0031472,FBgn0031473	GO:0016757	transferase activity, transferring glycosyl groups	2.87E-06	3	12
FBgn0004584,FBgn0031483, FBgn0031484,FBgn0031485, FBgn0250786,FBgn0025109, FBgn0002989,FBgn0025681	GO:0003677	DNA binding	0.0154	3	8
FBgn0031520,FBgn0053281, FBgn0053282,FBgn0031522, FBgn0031523	GO:0055085	transmembrane transport	2.58E-08	5	5
FBgn0031520,FBgn0053281, FBgn0053282,FBgn0031522, FBgn0031523	GO:0006810	transport	3.18E-07	5	5
FBgn0031520,FBgn0053281, FBgn0053282,FBgn0031522, FBgn0031523	GO:0022857	transmembrane transporter activity	0.000637	2	5

FBgn0031520,FBgn0053281, FBgn0053282,FBgn0031522, FBgn0031523	GO:0008150	biological_ process	0.00135	5	5
FBgn0053123,FBgn0031544	GO:0005737	cytoplasm	0.00426	2	2
FBgn0024244,FBgn0004892, FBgn0002985	GO:0005622	intracellula r	0.00892	2	3
FBgn0031746,FBgn0002856, FBgn0002855	GO:0005615	extracellula r space	9.78E-06	3	3
FBgn0031746,FBgn0002856, FBgn0002855	GO:0005576	extracellula r region	0.00517	2	3
FBgn0031756,FBgn0031757, FBgn0031758	GO:0022857	transmemb rane transporter activity	0.000192	2	3
FBgn0051642,FBgn0031782, FBgn0031784,FBgn0031785, FBgn0040950,FBgn0031786, FBgn0085409,FBgn0026755, FBgn0053531	GO:0007155	cell adhesion	0.0218	2	9
FBgn0026196,FBgn0028554, FBgn0004838	GO:0003729	mRNA binding	0.00449	2	3
FBgn0026196,FBgn0028554, FBgn0004838	GO:0003723	RNA binding	0.0131	2	3
FBgn0031874,FBgn0031875, FBgn0031876	GO:0003677	DNA binding	0.0126	2	3
FBgn0031874,FBgn0031875, FBgn0031876	GO:0003674	molecular_ function	0.0237	3	3
FBgn0031900,FBgn0051909, FBgn0010453,FBgn0004009, FBgn0031902,FBgn0031903, FBgn0002938	GO:0005576	extracellula r region	0.000125	4	7
FBgn0020618,FBgn0004177	GO:0005737	cytoplasm	0.0128	2	2
FBgn0031977,FBgn0044323	GO:0005634	nucleus	0.0452	2	2
FBgn0013531,FBgn0010287	GO:0005634	nucleus	0.0452	2	2
FBgn0032050,FBgn0032051, FBgn0032052,FBgn0032053, FBgn0032054	GO:0003735	structural constituent of ribosome	0.0194	2	5
FBgn0032050,FBgn0032051, FBgn0032052,FBgn0032053, FBgn0032054	GO:0006412	translation	0.0249	2	5
FBgn0032050,FBgn0032051, FBgn0032052,FBgn0032053, FBgn0032054	GO:0005198	structural molecule activity	0.0279	2	5
FBgn0032050,FBgn0032051, FBgn0032052,FBgn0032053, FBgn0032054	GO:0009058	biosyntheti c process	0.0295	2	5
FBgn0032066,FBgn0032067, FBgn0032068,FBgn0032069	GO:0005764	lysosome	3.53E-10	4	4

FBgn0032066,FBgn0032067, FBgn0032068,FBgn0032069	GO:0005773	vacuole	3.53E-10	4	4
FBgn0032066,FBgn0032067, FBgn0032068,FBgn0032069	GO:0043226	organelle	0.00156	4	4
FBgn0051709,FBgn0014179, FBgn0032132,FBgn0019809, FBgn0051882	GO:0008283	cell proliferation	0.00247	2	5
FBgn0051709,FBgn0014179, FBgn0032132,FBgn0019809, FBgn0051882	GO:0003674	molecular_function	0.0319	4	5
FBgn0051709,FBgn0014179, FBgn0032132,FBgn0019809, FBgn0051882	GO:0008150	biological_process	0.0416	4	5
FBgn0024285,FBgn0040064, FBgn0025700,FBgn0032160	GO:0005739	mitochondrion	0.0128	2	4
FBgn0024285,FBgn0040064, FBgn0025700,FBgn0032160	GO:0005811	lipid particle	0.0202	2	4
FBgn0032161,FBgn0032162	GO:0005739	mitochondrion	0.000817	2	2
FBgn0032236,FBgn0000413, FBgn0032237,FBgn0051716	GO:0003723	RNA binding	0.0258	2	4
FBgn0051872,FBgn0043825, FBgn0032275	GO:0006629	lipid metabolic process	7.69E-06	3	3
FBgn0051872,FBgn0043825, FBgn0032275	GO:0005615	extracellular space	0.0027	2	3
FBgn0051872,FBgn0043825, FBgn0032275	GO:0008150	biological_process	0.0291	3	3
FBgn0024740,FBgn0032289, FBgn0032291	GO:0006629	lipid metabolic process	0.00239	2	3
FBgn0028919,FBgn0032533, FBgn0001965,FBgn0000153, FBgn0004406,FBgn0015271	GO:0003677	DNA binding	0.00385	3	6
FBgn0001961,FBgn0026373, FBgn0028509	GO:0043226	organelle	0.0306	3	3
FBgn0027929,FBgn0028543, FBgn0054003,FBgn0028542, FBgn0028936	GO:0005576	extracellular region	1.42E-08	5	5
FBgn0027929,FBgn0028543, FBgn0054003,FBgn0028542, FBgn0028936	GO:0005575	cellular_component	0.00993	5	5
FBgn0032553,FBgn0032554, FBgn0013300,FBgn0013301, FBgn0028887,FBgn0259213	GO:0003677	DNA binding	0.0032	3	6
FBgn0015338,FBgn0011708, FBgn0000339	GO:0016192	vesicle-mediated transport	0.000616	2	3

FBgn0015338,FBgn0011708, FBgn0000339	GO:0006810	transport	0.0402	2	3
FBgn0020416,FBgn0020415	GO:0005576	extracellular region	0.00117	2	2
FBgn0032720,FBgn0032721	GO:0003735	structural constituent of ribosome	0.00179	2	2
FBgn0032720,FBgn0032721	GO:0006412	translation	0.00231	2	2
FBgn0032720,FBgn0032721	GO:0005198	structural molecule activity	0.00259	2	2
FBgn0032720,FBgn0032721	GO:0009058	biosynthetic process	0.00274	2	2
FBgn0032815,FBgn0032817, FBgn0032818,FBgn0032819	GO:0005622	intracellular	0.0219	2	4
FBgn0032824,FBgn0032825, FBgn0032827,FBgn0044811, FBgn0053117	GO:0006950	response to stress	0.000164	2	5
FBgn0032824,FBgn0032825, FBgn0032827,FBgn0044811, FBgn0053117	GO:0005615	extracellular space	0.0132	2	5
FBgn0032824,FBgn0032825, FBgn0032827,FBgn0044811, FBgn0053117	GO:0005576	extracellular region	0.0334	2	5
FBgn0014127,FBgn0019686, FBgn0003978,FBgn0045064	GO:0007067	mitosis	0.002	2	4
FBgn0032906,FBgn0032907	GO:0003677	DNA binding	0.00215	2	2
FBgn0032940,FBgn0000370	GO:0005634	nucleus	0.0226	2	2
FBgn0037224,FBgn0037225, FBgn0250821,FBgn0037227, FBgn0041621	GO:0005578	proteinaceous extracellular matrix	7.80E-06	3	5

FBgn0051559,FBgn0037405, FBgn0037406,FBgn0051562, FBgn0037408,FBgn0037409, FBgn0037410,FBgn0037411, FBgn0037412,FBgn0037413, FBgn0027527,FBgn0037414, FBgn0037415,FBgn0037416, FBgn0037417,FBgn0037418, FBgn0037419,FBgn0037420, FBgn0037421,FBgn0037422, FBgn0040279,FBgn0037424, FBgn0051561,FBgn0051556, FBgn0051560,FBgn0037427, FBgn0037428,FBgn0037429, FBgn0037430,FBgn0037431, FBgn0037432	GO:0003674	molecular_ function	1.59E-07	20	31
FBgn0051559,FBgn0037405, FBgn0037406,FBgn0051562, FBgn0037408,FBgn0037409, FBgn0037410,FBgn0037411, FBgn0037412,FBgn0037413, FBgn0027527,FBgn0037414, FBgn0037415,FBgn0037416, FBgn0037417,FBgn0037418, FBgn0037419,FBgn0037420, FBgn0037421,FBgn0037422, FBgn0040279,FBgn0037424, FBgn0051561,FBgn0051556, FBgn0051560,FBgn0037427, FBgn0037428,FBgn0037429, FBgn0037430,FBgn0037431, FBgn0037432	GO:0008150	biological_ process	7.04E-07	19	31
FBgn0051559,FBgn0037405, FBgn0037406,FBgn0051562, FBgn0037408,FBgn0037409, FBgn0037410,FBgn0037411, FBgn0037412,FBgn0037413, FBgn0027527,FBgn0037414, FBgn0037415,FBgn0037416, FBgn0037417,FBgn0037418, FBgn0037419,FBgn0037420, FBgn0037421,FBgn0037422, FBgn0040279,FBgn0037424, FBgn0051561,FBgn0051556, FBgn0051560,FBgn0037427, FBgn0037428,FBgn0037429, FBgn0037430,FBgn0037431, FBgn0037432	GO:0005575	cellular_co mponent	0.00236	2	31

FBgn0051481,FBgn0004054, FBgn0085326,FBgn0004053, FBgn0000166	GO:0005634	nucleus	0.00139	4	5
FBgn0051481,FBgn0004054, FBgn0085326,FBgn0004053, FBgn0000166	GO:0043226	organelle	0.00552	4	5
FBgn0000439,FBgn0003339, FBgn0001077	GO:0005634	nucleus	0.00605	3	3
FBgn0000439,FBgn0003339, FBgn0001077	GO:0043226	organelle	0.0175	3	3
FBgn0015770,FBgn0037501, FBgn0051544,FBgn0054023, FBgn0037503	GO:0005886	plasma membrane	0.0274	2	5
FBgn0042104,FBgn0042102, FBgn0042105,FBgn0042103, FBgn0037517	GO:0007165	signal transductio n	0.00705	2	5
FBgn0037617,FBgn0037618, FBgn0037619,FBgn0037620	GO:0005634	nucleus	0.0112	3	4
FBgn0037617,FBgn0037618, FBgn0037619,FBgn0037620	GO:0043226	organelle	0.0313	3	4
FBgn0014380,FBgn0037700, FBgn0003205	GO:0003924	GTPase activity	0.00326	2	3
FBgn0051390,FBgn0010421	GO:0005634	nucleus	0.0303	2	2
FBgn0037876,FBgn0037877, FBgn0037878,FBgn0020910, FBgn0037880,FBgn0037881, FBgn0037882,FBgn0037883, FBgn0037885,FBgn0037884	GO:0051082	unfolded protein binding	0.0294	2	10
FBgn0037876,FBgn0037877, FBgn0037878,FBgn0020910, FBgn0037880,FBgn0037881, FBgn0037882,FBgn0037883, FBgn0037885,FBgn0037884	GO:0006457	protein folding	0.0347	2	10
FBgn0051441,FBgn0051388	GO:0005634	nucleus	0.0303	2	2
FBgn0037892,FBgn0037893, FBgn0011774,FBgn0037894	GO:0003674	molecular_ function	0.0178	4	4
FBgn0037895,FBgn0037896	GO:0006810	transport	0.00846	2	2
FBgn0037918,FBgn0051363, FBgn0037920,FBgn0037922	GO:0005634	nucleus	0.0261	3	4
FBgn0037921,FBgn0037923, FBgn0042205,FBgn0037924	GO:0005634	nucleus	0.0149	3	4
FBgn0037921,FBgn0037923, FBgn0042205,FBgn0037924	GO:0043226	organelle	0.0417	3	4
FBgn0037921,FBgn0037923, FBgn0042205,FBgn0037924	GO:0005575	cellular_co mponent	0.0489	4	4
FBgn0042207,FBgn0038067, FBgn0038068,FBgn0038069, FBgn0038070	GO:0006629	lipid metabolic process	6.65E-11	5	5

FBgn0042207,FBgn0038067, FBgn0038068,FBgn0038069, FBgn0038070	GO:0008150	biological_ process	0.000502	5	5
FBgn0038244,FBgn0020299	GO:0005622	intracellula r	0.00143	2	2
FBgn0038260,FBgn0038261, FBgn0038262,FBgn0038266, FBgn0053555,FBgn0038267, FBgn0038268	GO:0055085	transmemb rane transport	0.00204	3	7
FBgn0038260,FBgn0038261, FBgn0038262,FBgn0038266, FBgn0053555,FBgn0038267, FBgn0038268	GO:0006810	transport	0.00491	3	7
FBgn0002783,FBgn0022787	GO:0003677	DNA binding	0.00725	2	2
FBgn0038402,FBgn0063649, FBgn0038404,FBgn0038405	GO:0055085	transmemb rane transport	0.0268	2	4
FBgn0038402,FBgn0063649, FBgn0038404,FBgn0038405	GO:0006810	transport	0.0483	2	4
FBgn0038414,FBgn0038415, FBgn0038416	GO:0022857	transmemb rane transporter activity	1.24E-06	3	3
FBgn0038414,FBgn0038415, FBgn0038416	GO:0055085	transmemb rane transport	0.000105	3	3
FBgn0038414,FBgn0038415, FBgn0038416	GO:0006810	transport	0.000261	3	3
FBgn0038414,FBgn0038415, FBgn0038416	GO:0008150	biological_ process	0.0347	3	3
FBgn0038414,FBgn0038415, FBgn0038416	GO:0003674	molecular_ function	0.0379	3	3
FBgn0038471,FBgn0038472, FBgn0020493,FBgn0038474, FBgn0038473,FBgn0038475, FBgn0038476,FBgn0086736, FBgn0038478	GO:0005730	nucleolus	0.0124	2	9
FBgn0038547,FBgn0038548, FBgn0038549,FBgn0038550	GO:0005634	nucleus	0.000301	4	4
FBgn0038547,FBgn0038548, FBgn0038549,FBgn0038550	GO:0043226	organelle	0.00125	4	4
FBgn0038547,FBgn0038548, FBgn0038549,FBgn0038550	GO:0005575	cellular_co mponent	0.0367	4	4
FBgn0004652,FBgn0038626	GO:0005622	intracellula r	0.00191	2	2
FBgn0038629,FBgn0038632, FBgn0038633	GO:0005576	extracellula r region	0.00282	2	3

FBgn0038680,FBgn0038681	GO:0005739	mitochondrion	0.000515	2	2
FBgn0038716,FBgn0038717, FBgn0038718,FBgn0038719	GO:0055085	transmembrane transport	1.70E-06	4	4
FBgn0038716,FBgn0038717, FBgn0038718,FBgn0038719	GO:0006810	transport	5.76E-06	4	4
FBgn0038716,FBgn0038717, FBgn0038718,FBgn0038719	GO:0008150	biological_process	0.00396	4	4
FBgn0038765,FBgn0038766, FBgn0038767,FBgn0038768, FBgn0038769	GO:0005634	nucleus	0.00139	4	5
FBgn0038765,FBgn0038766, FBgn0038767,FBgn0038768, FBgn0038769	GO:0043226	organelle	0.00552	4	5
FBgn0038838,FBgn0044809	GO:0006950	response to stress	4.10E-05	2	2
FBgn0038838,FBgn0044809	GO:0005615	extracellular space	0.000439	2	2
FBgn0038838,FBgn0044809	GO:0005576	extracellular region	0.00286	2	2
FBgn0038851,FBgn0038852	GO:0005634	nucleus	0.0303	2	2
FBgn0038886,FBgn0038887, FBgn0038888,FBgn0051438, FBgn0008651,FBgn0011278	GO:0008289	lipid binding	2.56E-05	2	6
FBgn0038886,FBgn0038887, FBgn0038888,FBgn0051438, FBgn0008651,FBgn0011278	GO:0005615	extracellular space	0.00753	2	6
FBgn0038886,FBgn0038887, FBgn0038888,FBgn0051438, FBgn0008651,FBgn0011278	GO:0005576	extracellular region	0.0473	2	6
FBgn0038918,FBgn0038919	GO:0005615	extracellular space	0.000293	2	2
FBgn0038922,FBgn0019960, FBgn0038923,FBgn0038924, FBgn0038925,FBgn0019957	GO:0005739	mitochondrion	0.0249	2	6
FBgn0038978,FBgn0038979	GO:0003677	DNA binding	0.00604	2	2
FBgn0039051,FBgn0039052	GO:0006520	cellular amino acid metabolic process	3.42E-05	2	2
FBgn0039051,FBgn0039052	GO:0044281	small molecule metabolic process	3.42E-05	2	2
FBgn0039051,FBgn0039052	GO:0005737	cytoplasm	0.0104	2	2

FBgn0039140,FBgn0039141, FBgn0043455,FBgn0005649, FBgn0026576,FBgn0010709, FBgn0039145,FBgn0001230, FBgn0040283	GO:0051082	unfolded protein binding	0.0344	2	9
FBgn0039140,FBgn0039141, FBgn0043455,FBgn0005649, FBgn0026576,FBgn0010709, FBgn0039145,FBgn0001230, FBgn0040283	GO:0006457	protein folding	0.0406	2	9
FBgn0039140,FBgn0039141, FBgn0043455,FBgn0005649, FBgn0026576,FBgn0010709, FBgn0039145,FBgn0001230, FBgn0040283	GO:0003674	molecular_ function	0.0441	6	9
FBgn0004897,FBgn0004898	GO:0048856	anatomical structure developme nt	5.27E-05	2	2
FBgn0004897,FBgn0004898	GO:0009790	embryo developme nt	5.27E-05	2	2
FBgn0039282,FBgn0039283, FBgn0039286	GO:0043226	organelle	0.0175	3	3
FBgn0039282,FBgn0039283, FBgn0039286	GO:0003677	DNA binding	0.0213	2	3
FBgn0039329,FBgn0028647	GO:0005634	nucleus	0.0303	2	2
FBgn0039366,FBgn0039367, FBgn0039368,FBgn0039369, FBgn0039370,FBgn0039371	GO:0005783	endoplasmi c reticulum	8.36E-09	5	6
FBgn0039366,FBgn0039367, FBgn0039368,FBgn0039369, FBgn0039370,FBgn0039371	GO:0043226	organelle	0.000936	5	6
FBgn0011670,FBgn0039378, FBgn0011668	GO:0005615	extracellula r space	0.000874	2	3
FBgn0011670,FBgn0039378, FBgn0011668	GO:0005576	extracellula r region	0.00564	2	3
FBgn0039386,FBgn0039387, FBgn0051091,FBgn0051089	GO:0006629	lipid metabolic process	0.000869	2	4
FBgn0002592,FBgn0002609, FBgn0002629,FBgn0002631, FBgn0002632,FBgn0002633	GO:0005634	nucleus	0.00767	4	6
FBgn0002592,FBgn0002609, FBgn0002629,FBgn0002631, FBgn0002632,FBgn0002633	GO:0043226	organelle	0.0294	4	6

FBgn0039434,FBgn0039435, FBgn0039436,FBgn0039438, FBgn0039439,FBgn0039441, FBgn0051080	GO:0005578	proteinaceous extracellular matrix	1.37E-12	7	7
FBgn0039434,FBgn0039435, FBgn0039436,FBgn0039438, FBgn0039439,FBgn0039441, FBgn0051080	GO:0005575	cellular_component	0.000892	7	7
FBgn0051081,FBgn0039443, FBgn0039444,FBgn0243586, FBgn0039448	GO:0005578	proteinaceous extracellular matrix	3.30E-11	5	5
FBgn0051081,FBgn0039443, FBgn0039444,FBgn0243586, FBgn0039448	GO:0005575	cellular_component	0.00812	5	5
FBgn0039470,FBgn0039471	GO:0006629	lipid metabolic process	0.000146	2	2
FBgn0039472,FBgn0039473, FBgn0039474	GO:0006629	lipid metabolic process	1.18E-06	3	3
FBgn0039472,FBgn0039473, FBgn0039474	GO:0008150	biological_process	0.0139	3	3
FBgn0004387,FBgn0039525, FBgn0027492,FBgn0040080	GO:0007165	signal transduction	0.00533	2	4
FBgn0039566,FBgn0003890	GO:0003924	GTPase activity	0.000685	2	2
FBgn0000278,FBgn0000279	GO:0005615	extracellular space	0.000293	2	2
FBgn0000278,FBgn0000279	GO:0005576	extracellular region	0.00191	2	2
FBgn0015221,FBgn0015222	GO:0005794	Golgi apparatus	0.000111	2	2
FBgn0015221,FBgn0015222	GO:0016491	oxidoreductase activity	0.000256	2	2
FBgn0026190,FBgn0039780, FBgn0051371	GO:0005783	endoplasmic reticulum	0.00178	2	3
FBgn0051014,FBgn0039782, FBgn0054041,FBgn0039783, FBgn0051524,FBgn0039784	GO:0005783	endoplasmic reticulum	2.95E-11	6	6
FBgn0051014,FBgn0039782, FBgn0054041,FBgn0039783, FBgn0051524,FBgn0039784	GO:0043226	organelle	4.18E-05	6	6
FBgn0051014,FBgn0039782, FBgn0054041,FBgn0039783, FBgn0051524,FBgn0039784	GO:0005575	cellular_component	0.00673	6	6

FBgn0051021,FBgn0051017, FBgn0051016	GO:0005783	endoplasmic reticulum	8.19E-06	3	3
FBgn0051021,FBgn0051017, FBgn0051016	GO:0043226	organelle	0.00876	3	3
FBgn0039792,FBgn0051013, FBgn0051015,FBgn0039795, FBgn0039796,FBgn0051010, FBgn0039797,FBgn0039798, FBgn0039799	GO:0005783	endoplasmic reticulum	0.0473	2	9
FBgn0051005,FBgn0039835, FBgn0039836	GO:0009058	biosynthetic process	0.00393	2	3
FBgn0039872,FBgn0017448, FBgn0039873	GO:0055085	transmembrane transport	6.31E-05	3	3
FBgn0039872,FBgn0017448, FBgn0039873	GO:0006810	transport	0.000157	3	3
FBgn0039872,FBgn0017448, FBgn0039873	GO:0008150	biological_process	0.0208	3	3
FBgn0040370,FBgn0040371, FBgn0029521,FBgn0004034, FBgn0000022,FBgn0004170, FBgn0011822,FBgn0000137	GO:0003677	DNA binding	0.0203	3	8
FBgn0040344,FBgn0040339	GO:0005634	nucleus	0.0363	2	2
FBgn0040337,FBgn0026143	GO:0005737	cytoplasm	0.0141	2	2
FBgn0000377,FBgn0023525, FBgn0029608	GO:0006810	transport	0.0141	2	3
FBgn0000377,FBgn0023525, FBgn0029608	GO:0005622	intracellular	0.0247	2	3
FBgn0000520,FBgn0250874, FBgn0024975,FBgn0000376, FBgn0024980,FBgn0040384	GO:0043226	organelle	0.0372	4	6
FBgn0028369,FBgn0003285, FBgn0029649,FBgn0004647	GO:0005886	plasma membrane	0.000608	3	4
FBgn0040393,FBgn0010294, FBgn0010295,FBgn0002933, FBgn0010296,FBgn0003086, FBgn0003374	GO:0005198	structural molecule activity	4.54E-07	5	7
FBgn0040393,FBgn0010294, FBgn0010295,FBgn0002933, FBgn0010296,FBgn0003086, FBgn0003374	GO:0005576	extracellular region	5.14E-07	5	7
FBgn0040393,FBgn0010294, FBgn0010295,FBgn0002933, FBgn0010296,FBgn0003086, FBgn0003374	GO:0003674	molecular_function	0.00148	6	7

FBgn0040393,FBgn0010294, FBgn0010295,FBgn0002933, FBgn0010296,FBgn0003086, FBgn0003374	GO:0005575	cellular_component	0.031	6	7
FBgn0052772,FBgn0029736, FBgn0029737,FBgn0029738	GO:0005622	intracellular	0.0322	2	4
FBgn0004403,FBgn0004404	GO:0005840	ribosome	0.000397	2	2
FBgn0004403,FBgn0004404	GO:0003735	structural constituent of ribosome	0.00211	2	2
FBgn0004403,FBgn0004404	GO:0006412	translation	0.00211	2	2
FBgn0004403,FBgn0004404	GO:0009058	biosynthetic process	0.00261	2	2
FBgn0004403,FBgn0004404	GO:0005198	structural molecule activity	0.00493	2	2
FBgn0030008,FBgn0030009	GO:0005622	intracellular	0.00282	2	2
FBgn0030010,FBgn0030011, FBgn0030012	GO:0005622	intracellular	0.0165	2	3
FBgn0000233,FBgn0020378	GO:0005634	nucleus	0.0436	2	2
FBgn0030163,FBgn0030164	GO:0007165	signal transduction	0.000391	2	2
FBgn0005391,FBgn0004045	GO:0006629	lipid metabolic process	0.000198	2	2
FBgn0005391,FBgn0004045	GO:0005811	lipid particle	0.00202	2	2
FBgn0005391,FBgn0004045	GO:0005198	structural molecule activity	0.00296	2	2
FBgn0030314,FBgn0030316, FBgn0002948	GO:0003677	DNA binding	0.0178	2	3
FBgn0000808,FBgn0003865	GO:0005576	extracellular region	0.00207	2	2
FBgn0030506,FBgn0030507	GO:0005634	nucleus	0.0363	2	2
FBgn0086448,FBgn0004227, FBgn0004598,FBgn0030744, FBgn0030745,FBgn0010416	GO:0003729	mRNA binding	0.0117	2	6
FBgn0086448,FBgn0004227, FBgn0004598,FBgn0030744, FBgn0030745,FBgn0010416	GO:0003723	RNA binding	0.0213	2	6
FBgn0011742,FBgn0030759, FBgn0015615,FBgn0015374, FBgn0030761,FBgn0061200, FBgn0025743	GO:0007010	cytoskeleton organization	0.00168	2	7

FBgn0052570,FBgn0052569, FBgn0052568,FBgn0052574	GO:0005578	proteinaceous extracellular matrix	1.89E-06	3	4
FBgn0052570,FBgn0052569, FBgn0052568,FBgn0052574	GO:0007165	signal transduction	0.0046	2	4
FBgn0031011,FBgn0031012, FBgn0259834	GO:0055085	transmembrane transport	6.00E-06	3	3
FBgn0031011,FBgn0031012, FBgn0259834	GO:0006810	transport	6.56E-05	3	3
FBgn0031011,FBgn0031012, FBgn0259834	GO:0008150	biological_process	0.0153	3	3
FBgn0031047,FBgn0031048, FBgn0031049	GO:0003674	molecular_function	0.0287	3	3
FBgn0031109,FBgn0031110, FBgn0031111	GO:0005576	extracellular region	0.00306	2	3
FBgn0031116,FBgn0052506	GO:0005622	intracellular	0.00282	2	2
FBgn0031118,FBgn0031119, FBgn0026323	GO:0007165	signal transduction	0.00232	2	3
FBgn0259162,FBgn0083981	GO:0003677	DNA binding	0.00606	2	2
FBgn0259162,FBgn0083981	GO:0005634	nucleus	0.0363	2	2
FBgn0033047,FBgn0033048	GO:0055085	transmembrane transport	0.00263	2	2
FBgn0033047,FBgn0033048	GO:0006810	transport	0.00491	2	2
FBgn0033083,FBgn0000054	GO:0005634	nucleus	0.0427	2	2
FBgn0033096,FBgn0033097	GO:0055085	transmembrane transport	0.00263	2	2
FBgn0033096,FBgn0033097	GO:0006810	transport	0.00491	2	2
FBgn0066293,FBgn0053349, FBgn0053558	GO:0043234	protein complex	0.000403	2	3
FBgn0029507,FBgn0029506, FBgn0033127	GO:0003674	molecular_function	0.0159	3	3
FBgn0029507,FBgn0029506, FBgn0033127	GO:0008150	biological_process	0.0163	3	3
FBgn0033128,FBgn0033129, FBgn0033130	GO:0003674	molecular_function	0.0159	3	3
FBgn0033128,FBgn0033129, FBgn0033130	GO:0008150	biological_process	0.0163	3	3
FBgn0033135,FBgn0033136, FBgn0033137	GO:0003674	molecular_function	0.0159	3	3
FBgn0033135,FBgn0033136, FBgn0033137	GO:0008150	biological_process	0.0163	3	3

FBgn0025185,FBgn0033185, FBgn0033186	GO:0005622	intracellular	2.84E-05	3	3
FBgn0033204,FBgn0033205, FBgn0033206	GO:0016491	oxidoreductase activity	0.000342	2	3
FBgn0028562,FBgn0028561, FBgn0033257	GO:0055085	transmembrane transport	0.00775	2	3
FBgn0028562,FBgn0028561, FBgn0033257	GO:0006810	transport	0.0143	2	3
FBgn0002570,FBgn0002569, FBgn0002571,FBgn0033294	GO:0005975	carbohydrate metabolic process	5.15E-09	4	4
FBgn0002570,FBgn0002569, FBgn0002571,FBgn0033294	GO:0008150	biological_process	0.00327	4	4
FBgn0050360,FBgn0033296, FBgn0033297	GO:0005975	carbohydrate metabolic process	7.93E-07	3	3
FBgn0050360,FBgn0033296, FBgn0033297	GO:0008150	biological_process	0.0163	3	3
FBgn0033322,FBgn0033323, FBgn0033324,FBgn0033326, FBgn0050357	GO:0055085	transmembrane transport	0.00146	3	5
FBgn0033322,FBgn0033323, FBgn0033324,FBgn0033326, FBgn0050357	GO:0006810	transport	0.00368	3	5
FBgn0033322,FBgn0033323, FBgn0033324,FBgn0033326, FBgn0050357	GO:0008150	biological_process	0.0412	4	5
FBgn0050343,FBgn0020621, FBgn0033400,FBgn0033401, FBgn0026326,FBgn0033402	GO:0007165	signal transduction	0.00693	2	6
FBgn0050343,FBgn0020621, FBgn0033400,FBgn0033401, FBgn0026326,FBgn0033402	GO:0005622	intracellular	0.042	2	6
FBgn0085436,FBgn0033427	GO:0005737	cytoplasm	0.00914	2	2
FBgn0033562,FBgn0010356, FBgn0033566,FBgn0050020, FBgn0017414,FBgn0033569	GO:0005634	nucleus	0.0395	3	6
FBgn0043470,FBgn0043471, FBgn0011556,FBgn0011554, FBgn0011555	GO:0005576	extracellular region	0.000305	3	5
FBgn0003863,FBgn0010425, FBgn0010357	GO:0005576	extracellular region	3.16E-05	3	3
FBgn0033663,FBgn0000556	GO:0005811	lipid particle	0.00237	2	2

FBgn0033667,FBgn0033668	GO:0005622	intracellular	0.00119	2	2
FBgn0050047,FBgn0050049, FBgn0050043	GO:0008233	peptidase activity	1.75E-07	3	3
FBgn0050047,FBgn0050049, FBgn0050043	GO:0003674	molecular_function	0.0238	3	3
FBgn0033748,FBgn0033749	GO:0005634	nucleus	0.0183	2	2
FBgn0033748,FBgn0033749	GO:0043226	organelle	0.0405	2	2
FBgn0250842,FBgn0050488, FBgn0050486	GO:0005576	extracellular region	7.90E-05	3	3
FBgn0033903,FBgn0033904, FBgn0033905	GO:0022857	transmembrane transporter activity	3.40E-08	3	3
FBgn0033903,FBgn0033904, FBgn0033905	GO:0003674	molecular_function	0.0159	3	3
FBgn0034008,FBgn0004698, FBgn0034009,FBgn0013750	GO:0005622	intracellular	0.0312	2	4
FBgn0050467,FBgn0034027, FBgn0034030	GO:0005576	extracellular region	0.00378	2	3
FBgn0014020,FBgn0017549	GO:0003924	GTPase activity	0.00092	2	2
FBgn0034152,FBgn0034153, FBgn0015584,FBgn0053530	GO:0005615	extracellular space	5.04E-05	3	4
FBgn0034152,FBgn0034153, FBgn0015584,FBgn0053530	GO:0005576	extracellular region	0.0149	2	4
FBgn0034152,FBgn0034153, FBgn0015584,FBgn0053530	GO:0005575	cellular_component	0.0427	4	4
FBgn0050456,FBgn0034194	GO:0005622	intracellular	0.00119	2	2
FBgn0034219,FBgn0028956	GO:0006950	response to stress	2.25E-06	2	2
FBgn0034329,FBgn0034330, FBgn0034331,FBgn0025583, FBgn0040735,FBgn0040734, FBgn0040733,FBgn0064237	GO:0005576	extracellular region	0.00162	3	8
FBgn0027535,FBgn0034419, FBgn0000578,FBgn0034418, FBgn0034420,FBgn0034421, FBgn0034422,FBgn0010434	GO:0007010	cytoskeleton organization	0.00651	2	8
FBgn0034440,FBgn0034441	GO:0008233	peptidase activity	3.37E-05	2	2
FBgn0022700,FBgn0034582	GO:0005615	extracellular space	0.000445	2	2
FBgn0034622,FBgn0034623, FBgn0034624	GO:0005576	extracellular region	0.00756	2	3
FBgn0050280,FBgn0050281	GO:0007165	signal transduction	0.000379	2	2

FBgn0050280,FBgn0050281	GO:0005615	extracellular space	0.000891	2	2
FBgn0034718,FBgn0013272	GO:0005886	plasma membrane	0.00142	2	2
FBgn0034766,FBgn0050259, FBgn0034768,FBgn0034769, FBgn0034770,FBgn0050275, FBgn0050268	GO:0016491	oxidoreductase activity	0.00353	2	7
FBgn0050272,FBgn0050265, FBgn0034782,FBgn0034783, FBgn0034784,FBgn0034785	GO:0055085	transmembrane transport	1.71E-09	6	6
FBgn0050272,FBgn0050265, FBgn0034782,FBgn0034783, FBgn0034784,FBgn0034785	GO:0006810	transport	1.16E-08	6	6
FBgn0050272,FBgn0050265, FBgn0034782,FBgn0034783, FBgn0034784,FBgn0034785	GO:0008150	biological_process	0.000196	6	6
FBgn0034883,FBgn0034884	GO:0055085	transmembrane transport	0.00263	2	2
FBgn0034883,FBgn0034884	GO:0006810	transport	0.00491	2	2
FBgn0041582,FBgn0021875, FBgn0034982,FBgn0019643, FBgn0015544	GO:0008283	cell proliferation	0.00355	2	5
FBgn0041582,FBgn0021875, FBgn0034982,FBgn0019643, FBgn0015544	GO:0005737	cytoplasm	0.00803	3	5
FBgn0035023,FBgn0035021, FBgn0035024,FBgn0035025, FBgn0035026,FBgn0035027, FBgn0035028	GO:0006457	protein folding	0.0212	2	7
FBgn0035063,FBgn0035064, FBgn0035065,FBgn0041205	GO:0003723	RNA binding	0.0249	2	4
FBgn0085434,FBgn0035077, FBgn0035078	GO:0055085	transmembrane transport	0.0181	2	3
FBgn0085434,FBgn0035077, FBgn0035078	GO:0006810	transport	0.0334	2	3
FBgn0011694,FBgn0004181	GO:0005615	extracellular space	0.000668	2	2
FBgn0035131,FBgn0035132	GO:0006950	response to stress	1.63E-05	2	2
FBgn0035155,FBgn0035157, FBgn0035158,FBgn0035159, FBgn0035160	GO:0003677	DNA binding	0.0171	2	5
FBgn0035283,FBgn0004636, FBgn0042712,FBgn0035285	GO:0003924	GTPase activity	0.00407	2	4
FBgn0035325,FBgn0052302	GO:0005576	extracellular region	0.00317	2	2

FBgn0053233,FBgn0035364, FBgn0053234,FBgn0035366	GO:0055085	transmembrane transport	0.0107	2	4
FBgn0053233,FBgn0035364, FBgn0053234,FBgn0035366	GO:0006810	transport	0.0293	2	4
FBgn0035563,FBgn0052238, FBgn0046793,FBgn0085295, FBgn0035567,FBgn0035568	GO:0022857	transmembrane transporter activity	0.00152	2	6
FBgn0035690,FBgn0022699, FBgn0022935,FBgn0035691	GO:0005634	nucleus	0.000137	4	4
FBgn0035690,FBgn0022699, FBgn0022935,FBgn0035691	GO:0043226	organelle	0.00074	4	4
FBgn0035690,FBgn0022699, FBgn0022935,FBgn0035691	GO:0005575	cellular_component	0.0439	4	4
FBgn0035695,FBgn0004513	GO:0055085	transmembrane transport	0.00183	2	2
FBgn0035695,FBgn0004513	GO:0006810	transport	0.00509	2	2
FBgn0010406,FBgn0027554, FBgn0035829,FBgn0035830, FBgn0035831,FBgn0011817, FBgn0026263	GO:0003723	RNA binding	0.0463	2	7
FBgn0035915,FBgn0052351	GO:0005737	cytoplasm	0.00733	2	2
FBgn0035954,FBgn0035955, FBgn0035956,FBgn0028789, FBgn0035957	GO:0005634	nucleus	0.0148	3	5
FBgn0035954,FBgn0035955, FBgn0035956,FBgn0028789, FBgn0035957	GO:0043226	organelle	0.0488	3	5
FBgn0036046,FBgn0044050	GO:0005576	extracellular region	0.00635	2	2
FBgn0040823,FBgn0036066, FBgn0052053,FBgn0052054	GO:0022857	transmembrane transporter activity	2.25E-06	3	4
FBgn0040823,FBgn0036066, FBgn0052053,FBgn0052054	GO:0055085	transmembrane transport	0.000134	3	4
FBgn0040823,FBgn0036066, FBgn0052053,FBgn0052054	GO:0006810	transport	0.000622	3	4
FBgn0036070,FBgn0036072	GO:0005576	extracellular region	0.00635	2	2
FBgn0003378,FBgn0003377	GO:0005198	structural molecule activity	0.00133	2	2
FBgn0003378,FBgn0003377	GO:0005576	extracellular region	0.00635	2	2

FBgn0036225,FBgn0036226, FBgn0036227,FBgn0036229	GO:0005576	extracellular region	4.83E-06	4	4
FBgn0036225,FBgn0036226, FBgn0036227,FBgn0036229	GO:0005575	cellular_component	0.0292	4	4
FBgn0053265,FBgn0036232	GO:0005576	extracellular region	0.00317	2	2
FBgn0036233,FBgn0036234	GO:0005576	extracellular region	0.00317	2	2
FBgn0036248,FBgn0036249	GO:0003677	DNA binding	0.00239	2	2
FBgn0036274,FBgn0052105	GO:0005634	nucleus	0.0204	2	2
FBgn0036274,FBgn0052105	GO:0043226	organelle	0.0474	2	2
FBgn0036285,FBgn0000625	GO:0005634	nucleus	0.0204	2	2
FBgn0036285,FBgn0000625	GO:0043226	organelle	0.0474	2	2
FBgn0036302,FBgn0015904, FBgn0015919,FBgn0052111, FBgn0014343,FBgn0014007	GO:0005575	cellular_component	0.0123	6	6
FBgn0036361,FBgn0036362, FBgn0036363	GO:0005576	extracellular region	0.000124	3	3
FBgn0036411,FBgn0042630	GO:0005634	nucleus	0.0341	2	2
FBgn0003459,FBgn0036423	GO:0003677	DNA binding	0.00358	2	2
FBgn0036518,FBgn0036519, FBgn0036520,FBgn0036522	GO:0005622	intracellular	0.0205	2	4
FBgn0004588,FBgn0004589, FBgn0004590,FBgn0004591	GO:0005576	extracellular region	4.83E-06	4	4
FBgn0004588,FBgn0004589, FBgn0004590,FBgn0004591	GO:0005575	cellular_component	0.0292	4	4
FBgn0004593,FBgn0004594, FBgn0014848,FBgn0014849, FBgn0014850,FBgn0014851	GO:0005576	extracellular region	7.02E-09	6	6
FBgn0004593,FBgn0004594, FBgn0014848,FBgn0014849, FBgn0014850,FBgn0014851	GO:0005575	cellular_component	0.00352	6	6
FBgn0036537,FBgn0036538	GO:0005622	intracellular	0.00142	2	2
FBgn0000212,FBgn0000115, FBgn0259824,FBgn0036556, FBgn0036557,FBgn0036558	GO:0005794	Golgi apparatus	0.00902	2	6
FBgn0036747,FBgn0036749, FBgn0036750,FBgn0052182	GO:0016491	oxidoreductase activity	0.000509	2	4
FBgn0015946,FBgn0011706, FBgn0036786	GO:0008219	cell death	0.000105	2	3
FBgn0015946,FBgn0011706, FBgn0036786	GO:0005739	mitochondrion	0.00411	2	3
FBgn0036794,FBgn0036795, FBgn0036796,FBgn0052201, FBgn0052199	GO:0005783	endoplasmic reticulum	4.95E-05	3	5

FBgn0036794,FBgn0036795, FBgn0036796,FBgn0052201, FBgn0052199	GO:0043226	organelle	0.0488	3	5
FBgn0036807,FBgn0036808, FBgn0036809,FBgn0036810	GO:0005739	mitochondrion	0.0109	2	4
FBgn0036889,FBgn0005386, FBgn0010417	GO:0003677	DNA binding	0.0106	2	3
FBgn0036889,FBgn0005386, FBgn0010417	GO:0003674	molecular_function	0.0396	3	3
FBgn0023094,FBgn0036911, FBgn0013799,FBgn0003744, FBgn0052221,FBgn0001324, FBgn0020389,FBgn0036913, FBgn0036915,FBgn0036916	GO:0007165	signal transduction	0.0203	2	10
FBgn0036949,FBgn0036950, FBgn0036951,FBgn0036952, FBgn0036953	GO:0005576	extracellular region	1.85E-07	5	5
FBgn0036949,FBgn0036950, FBgn0036951,FBgn0036952, FBgn0036953	GO:0005575	cellular_component	0.0101	5	5
FBgn0037003,FBgn0037004, FBgn0037005	GO:0055085	transmembrane transport	3.39E-05	3	3
FBgn0037003,FBgn0037004, FBgn0037005	GO:0005886	plasma membrane	0.000107	3	3
FBgn0037003,FBgn0037004, FBgn0037005	GO:0006810	transport	0.000159	3	3
FBgn0037003,FBgn0037004, FBgn0037005	GO:0008150	biological_process	0.0389	3	3
FBgn0037149,FBgn0027532, FBgn0037150,FBgn0037151	GO:0051082	unfolded protein binding	0.00105	2	4
FBgn0037149,FBgn0027532, FBgn0037150,FBgn0037151	GO:0006457	protein folding	0.00219	2	4
FBgn0037149,FBgn0027532, FBgn0037150,FBgn0037151	GO:0008150	biological_process	0.00768	4	4

Table S3 Segment gene orientation counts

	head-to-head	tail-to-tail	same strand
2 genes intrasegment	390	276	570
3+ genes intrasegment	1131	1076	1740
intersegment	1597	1766	2812

Table S4 Insulator enrichment

Peaks	Region Type	Subset	Insulator	Enrichment	P-value
masked	intersegment	no PID	BEAF	0.901	0.0011435
masked	intersegment	no PID	CP190	0.93	0.0064506
masked	intersegment	no PID	CTCF	0.9	**4.10E-06**
masked	intersegment	no PID	CTCF_C	0.874	0.0005602
masked	intersegment	no PID	CTCF_N	0.896	0.0121254
masked	intersegment	no PID	CTCF_N_KC	0.92	0.0858547
masked	intersegment	no PID	CTCF_N_S2	0.923	0.0845748
masked	intersegment	no PID	GAF	0.893	0.0008662
masked	intersegment	no PID	MDJ4	0.941	0.1212785
masked	intersegment	no PID	class1	0.91	**9.47E-09**
masked	intersegment	no PID	class2	1.011	0.6559713
masked	intersegment	no PID	suHw	1.013	0.7161517
masked	intersegment	no PID	suHw_pam	1.009	0.7881194
masked	intersegment	no PID	total	0.932	**1.94E-07**
all	intersegment	no PID	BEAF	0.731	**5.90E-27**
all	intersegment	no PID	CP190	0.774	**5.52E-25**
all	intersegment	no PID	CTCF	0.711	**2.58E-59**
all	intersegment	no PID	CTCF_C	0.698	**8.18E-24**
all	intersegment	no PID	CTCF_N	0.712	**6.95E-18**
all	intersegment	no PID	CTCF_N_KC	0.724	**2.65E-13**
all	intersegment	no PID	CTCF_N_S2	0.715	**9.49E-16**
all	intersegment	no PID	GAF	0.788	**9.88E-14**
all	intersegment	no PID	MDJ4	0.766	**1.14E-13**
all	intersegment	no PID	class1	0.735	**1.77E-88**
all	intersegment	no PID	class2	0.953	0.0495363
all	intersegment	no PID	suHw	0.954	0.1649845
all	intersegment	no PID	suHw_pam	0.953	0.1481105
all	intersegment	no PID	total	0.783	**1.38E-79**
masked	intersegment	PID	BEAF	1.163	0.000756
masked	intersegment	PID	CP190	1.006	0.8789309
masked	intersegment	PID	CTCF	1.029	0.3745567
masked	intersegment	PID	CTCF_C	1.074	0.1882005
masked	intersegment	PID	CTCF_N	1.049	0.4385505
masked	intersegment	PID	CTCF_N_KC	0.914	0.2136986
masked	intersegment	PID	CTCF_N_S2	1.048	0.4697201
masked	intersegment	PID	GAF	1.193	**0.000111**
masked	intersegment	PID	MDJ4	0.857	0.0074384
masked	intersegment	PID	class1	1.05	0.0346615
masked	intersegment	PID	class2	0.739	**6.18E-16**
masked	intersegment	PID	suHw	0.733	**3.77E-09**
masked	intersegment	PID	suHw_pam	0.744	**8.21E-09**
masked	intersegment	PID	total	0.976	0.2078031
all	intersegment	PID	BEAF	1.833	**9.02E-63**
all	intersegment	PID	CP190	1.535	**1.92E-41**
all	intersegment	PID	CTCF	1.686	**1.59E-89**
all	intersegment	PID	CTCF_C	1.694	**2.84E-33**

all	intersegment	PID	CTCF_N	1.708	**5.85E-28**
all	intersegment	PID	CTCF_N_KC	1.598	**1.87E-17**
all	intersegment	PID	CTCF_N_S2	1.731	**1.04E-26**
all	intersegment	PID	GAF	1.537	**3.75E-26**
all	intersegment	PID	MDJ4	1.417	**5.22E-14**
all	intersegment	PID	class1	1.67	**1.96E-154**
all	intersegment	PID	class2	0.882	0.0003683
all	intersegment	PID	suHw	0.872	0.0059781
all	intersegment	PID	suHw_pam	0.891	0.0165766
all	intersegment	PID	total	1.464	**6.37E-116**
masked	tissue restricted intersegment	-	BEAF	0.994	0.9721233
masked	tissue restricted intersegment	-	CP190	1	1
masked	tissue restricted intersegment	-	CTCF	1.132	0.0082345
masked	tissue restricted intersegment	-	CTCF_C	1.093	0.26485
masked	tissue restricted intersegment	-	CTCF_N	1.18	0.0620138
masked	tissue restricted intersegment	-	CTCF_N_KC	1.074	0.4660336
masked	tissue restricted intersegment	-	CTCF_N_S2	1.186	0.0672764
masked	tissue restricted intersegment	-	GAF	1.061	0.3884435
masked	tissue restricted intersegment	-	MDJ4	0.892	0.1979827
masked	tissue restricted intersegment	-	class1	1.059	0.0965906
masked	tissue restricted intersegment	-	class2	1.213	**9.85E-05**
masked	tissue restricted intersegment	-	suHw	1.22	0.0039385
masked	tissue restricted intersegment	-	suHw_pam	1.206	0.0056066
masked	tissue restricted intersegment	-	total	1.08	0.0059213
all	tissue restricted intersegment	-	BEAF	1.016	0.77835
all	tissue restricted intersegment	-	CP190	0.999	1
all	tissue restricted intersegment	-	CTCF	1.114	0.0111768
all	tissue restricted intersegment	-	CTCF_C	1.078	0.2973788
all	tissue restricted intersegment	-	CTCF_N	1.168	0.0508653
all	tissue restricted intersegment	-	CTCF_N_KC	1.06	0.5151474
all	tissue restricted intersegment	-	CTCF_N_S2	1.153	0.0876751
all	tissue restricted intersegment	-	GAF	1.057	0.3979704
all	tissue restricted intersegment	-	MDJ4	0.936	0.4258024
all	tissue restricted intersegment	-	class1	1.057	0.0845145
all	tissue restricted intersegment	-	class2	1.21	**9.09E-05**
all	tissue restricted intersegment	-	suHw	1.213	0.0043172
all	tissue restricted intersegment	-	suHw_pam	1.207	0.0044068
all	tissue restricted intersegment	-	total	1.077	0.0053649
masked	3+ gene intersegment	no PID	BEAF	1.298	**1.64E-09**
masked	3+ gene intersegment	no PID	CP190	1.231	**9.80E-09**
masked	3+ gene intersegment	no PID	CTCF	1.194	**1.44E-08**
masked	3+ gene intersegment	no PID	CTCF_C	1.135	0.0181426
masked	3+ gene intersegment	no PID	CTCF_N	1.247	**0.00017**
masked	3+ gene intersegment	no PID	CTCF_N_KC	1.205	0.004669
masked	3+ gene intersegment	no PID	CTCF_N_S2	1.207	0.0027635
masked	3+ gene intersegment	no PID	GAF	1.109	0.0277827
masked	3+ gene intersegment	no PID	MDJ4	1.127	0.025625
masked	3+ gene intersegment	no PID	class1	1.228	**3.91E-20**
masked	3+ gene intersegment	no PID	class2	1.231	**5.73E-10**

masked	3+ gene intersegment	no PID	suHw	1.239	**4.81E-06**
masked	3+ gene intersegment	no PID	suHw_pam	1.223	**9.85E-06**
masked	3+ gene intersegment	no PID	total	1.206	**1.75E-24**
all	3+ gene intersegment	no PID	BEAF	1.127	0.002818
all	3+ gene intersegment	no PID	CP190	1.083	0.0205426
all	3+ gene intersegment	no PID	CTCF	0.999	0.9883041
all	3+ gene intersegment	no PID	CTCF_C	0.961	0.4414925
all	3+ gene intersegment	no PID	CTCF_N	1.043	0.4402123
all	3+ gene intersegment	no PID	CTCF_N_KC	1.009	0.8772373
all	3+ gene intersegment	no PID	CTCF_N_S2	0.994	0.9533712
all	3+ gene intersegment	no PID	GAF	1.018	0.7017375
all	3+ gene intersegment	no PID	MDJ4	0.972	0.5960704
all	3+ gene intersegment	no PID	class1	1.054	0.014361
all	3+ gene intersegment	no PID	class2	1.188	**1.96E-07**
all	3+ gene intersegment	no PID	suHw	1.195	**0.000118**
all	3+ gene intersegment	no PID	suHw_pam	1.181	**0.000217**
all	3+ gene intersegment	no PID	total	1.069	**0.000202**
masked	3+ gene intersegment	PID	BEAF	1.159	0.0346046
masked	3+ gene intersegment	PID	CP190	1.029	0.6313606
masked	3+ gene intersegment	PID	CTCF	1.021	0.6653372
masked	3+ gene intersegment	PID	CTCF_C	1.087	0.3212066
masked	3+ gene intersegment	PID	CTCF_N	1.099	0.3052124
masked	3+ gene intersegment	PID	CTCF_N_KC	0.92	0.5125903
masked	3+ gene intersegment	PID	CTCF_N_S2	0.934	0.5679089
masked	3+ gene intersegment	PID	GAF	1.113	0.1342093
masked	3+ gene intersegment	PID	MDJ4	0.841	0.0680325
masked	3+ gene intersegment	PID	class1	1.053	0.1489509
masked	3+ gene intersegment	PID	class2	0.739	**4.57E-07**
masked	3+ gene intersegment	PID	suHw	0.738	**0.000336**
masked	3+ gene intersegment	PID	suHw_pam	0.739	**0.000265**
masked	3+ gene intersegment	PID	total	0.969	0.3099148
all	3+ gene intersegment	PID	BEAF	1.743	**1.89E-23**
all	3+ gene intersegment	PID	CP190	1.471	**5.32E-15**
all	3+ gene intersegment	PID	CTCF	1.546	**6.80E-27**
all	3+ gene intersegment	PID	CTCF_C	1.546	**1.53E-10**
all	3+ gene intersegment	PID	CTCF_N	1.605	**4.06E-10**
all	3+ gene intersegment	PID	CTCF_N_KC	1.469	**9.20E-06**
all	3+ gene intersegment	PID	CTCF_N_S2	1.55	**4.83E-08**
all	3+ gene intersegment	PID	GAF	1.403	**1.08E-07**
all	3+ gene intersegment	PID	MDJ4	1.276	0.0008429
all	3+ gene intersegment	PID	class1	1.565	**4.14E-50**
all	3+ gene intersegment	PID	class2	0.839	0.0019139
all	3+ gene intersegment	PID	suHw	0.842	0.0322006
all	3+ gene intersegment	PID	suHw_pam	0.835	0.0209903
all	3+ gene intersegment	PID	total	1.372	**6.95E-34**

Enrichment ratios greater than one indicate that the insulator is enriched in the specified region compared to the rest of the genome. P-values that are significant after Bonferroni correction are bolded (P < 5.95E-04).

Table S5
Tissue biological replicate pairwise correlation values

Tissue	Mean	Min	Max
Larval Wandering fat body	0.949	0.877	0.987
Larval Feeding Fatbody	0.958	0.924	0.991
Larval Feeding Trachea	0.979	0.971	0.992
Larval Feeding Carcass	0.989	0.980	0.993
Adult Whole Fly	0.988	0.983	0.992
Adult Carcass	0.988	0.985	0.991
Adult Male Ejaculatory Duct	0.991	0.986	0.995
Adult Salivary Gland	0.990	0.987	0.992
Adult Eye	0.990	0.988	0.993
Larval Feeding Mid Gut	0.994	0.989	0.996
Adult Fatbody	0.991	0.989	0.994
Adult Female Spermatheca Mated	0.992	0.990	0.993
Adult Heart	0.993	0.991	0.995
Adult Head	0.993	0.991	0.995
Adult Testes	0.994	0.991	0.996
Larval Feeding Hind Gut	0.994	0.992	0.997
Whole Larvae Feeding	0.995	0.992	0.998
Larval Feeding Malpighian Tubule	0.995	0.992	0.996
Adult Thoracoabdominal ganglion	0.995	0.992	0.997
Larval Feeding Salivary Gland	0.994	0.992	0.996
Adult Accessory gland	0.994	0.993	0.995
Adult Female Spermatheca Virgin	0.994	0.994	0.995
Larvae Wandering Tubules	0.995	0.994	0.996
Adult Crop	0.995	0.994	0.997
Larval Feeding Central Nevous System	0.997	0.995	0.997
Adult Brain	0.997	0.995	0.998
Adult Hind Gut	0.997	0.996	0.997
Adult Mid Gut	0.997	0.997	0.997
Adult Ovary	0.997	0.997	0.998
5th Passage Drosophila S2 Cells	0.998	0.998	0.999

Red values are below the 0.98 cutoff for pairwise correlation.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

2. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520–562.

3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de P, B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287**:2185–2195.

4. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot ….** *Nature* 2007,

5. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

6. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di S, Luisa, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, van B, Marijke, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B, Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M: **Identification of functional elements and regulatory circuits by Drosophila modENCODE.** *Science* 2010, **330**:1787–1797.

7. Salipante SJ, Kas A, McMonagle E, Horwitz MS: **Phylogenetic analysis of developmental and postnatal mouse cell lineages.** *Evol Dev* 2010, **12**:84–94.

8. Collins FS, Barker AD: **Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies.** *Sci Am* 2007, **296**:50–57.

9. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314**:268-274.

10. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew Y-E, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan M-H, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**:153-158.

11. Loeb LA: **A mutator phenotype in cancer.** *Cancer Res* 2001, **61**:3230–3239.

12. Loeb LA, Springgate CF, Battula N: **Errors in DNA replication as a basis of malignant changes.** *Cancer Res* 1974, **34**:2311-2321.

13. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, O'Meara S, Parker A, Tarpey P, Avis T, Barthorpe A, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Edwards K, Forbes S, Gorton M, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jones D, Kosmidou V, Laman R, Lugg R, Menzies A, Perry J, Petty R, Raine K, Shepherd R, Small A, Solomon H, Stephens Y, Tofts C, Varian J, Webb A, West S, Widaa S, Yates A, Brasseur F, Cooper CS, Flanagan AM, Green A, Knowles M, Leung SY, Looijenga LHJ, Malkowicz B, Pierotti MA, Teh B, Yuen ST, Nicholson AG, Lakhani S, Easton DF, Weber BL, Stratton MR, Futreal PA, Wooster R: **A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer.** *Nat Genet* 2005, **37**:590-592.

14. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**:147–151.

15. Bussemaker HJ, Foat BC, Ward LD: **Predictive modeling of genome-wide mRNA expression: from modules to molecules.** *Ann Rev Bioph Biom* 2007, **36**:329-347.

16. Caron H, van S, B, van dM, M, Baas F, Riggins G, van S, P, Hermus MC, van A, R, Boon K, Voûte PA, Heisterkamp S, van K, A, Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291**:1289–1292.

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

18. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans.** *Nature* 2002, **418**:975–979.

19. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol* 2002, **1**:5.

20. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA*2004, **101**:6062–6067.

21. Hurst LD, Pál C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5**:299–310.

22. Filippova GN: **Genetics and epigenetics of the multifunctional protein CTCF.** *Curr Top Dev Biol* 2008, **80**:337–360.

23. Akhtar A, Gasser SM: **The nuclear envelope and transcriptional control.** *Nat Rev Genet* 2007, **8**:507–517.

24. Feinberg AP, Ohlsson R, Henikoff S: **The epigenetic progenitor origin of human cancer.** *Nat Rev Genet* 2006, **7**:21–33.

25. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B* 1995,

26. Wang T-L, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW, Velculescu VE: **Digital karyotyping.** *Proc Natl Acad Sci USA*2002, **99**:16156–16161.

27. Li W-H: **Molecular Evolution.** Sinauer Associates; 1997.

28. Antequera F: **Structure, function and evolution of CpG island promoters.** *Cell Mol Life Sci* 2003, **60**:1647-1658.

29. Green P, Ewing B, Miller W, Thomas PJ, Program NISCCS, Green ED: **Transcription-associated mutational asymmetry in mammalian evolution.** *Nat Genet* 2003, **33**:514-517.

30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12**:996-1006.

31. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JKV, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PVK, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**:1108-1113.

32. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321**:1801-1806.

33. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW: **An integrated genomic analysis of human glioblastoma multiforme.** *Science* 2008, **321**:1807-1812.

34. Gojobori T, Li WH, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes.** *J Mol Evol* 1982, **18**:360-369.

35. Li WH, Wu CI, Luo CC: **Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications.** *J Mol Evol* 1984, **21**:58-71.

36. Blake RD, Hess ST, Nicholson-Tuell J: **The influence of nearest neighbors on the rate and pattern of spontaneous point mutations.** *J Mol Evol* 1992, **34**:189-200.

37. Hess ST, Blake JD, Blake RD: **Wide variations in neighbor-dependent substitution rates.** *J Mol Biol* 1994, **236**:1022–1033.

38. Majewski J: **Dependence of mutational asymmetry on gene-expression levels in the human genome.** *Am J Hum Genet* 2003, **73**:688-692.

39. Comeron JM: **Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence.** *Genetics* 2004, **167**:1293-1304.

40. Frank AC, Lobry JR: **Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms.** *Gene* 1999, **238**:65–77.

41. Francino MP, Ochman H: **Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences.** *Mol Biol Evol* 2001, **18**:1147–1150.

42. Hwang DG, Green P: **Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution.** *Proc Natl Acad Sci USA* 2004, **101**:13994-14001.

43. Strauss BS: **Role in tumorigenesis of silent mutations in the TP53 gene.** *Mutat Res* 2000, **457**:93-104.

44. Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, Wright FA, Feramisco JD, Peltomäki P, Lang JC, Schuller DE, Yu L, Bloomfield CD, Caligiuri MA, Yates A, Nishikawa R, Su Huang H, Petrelli NJ, Zhang X, O'Dorisio MS, Held WA, Cavenee WK, Plass C: **Aberrant CpG-island methylation has non-random and tumour-type-specific patterns.** *Nat Genet* 2000, **24**:132-138.

45. Esteller M: **CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future.** *Oncogene* 2002, **21**:5427-5440.

46. Shen L, Toyota M, Kondo Y, Lin E, Zhang L, Guo Y, Hernandez NS, Chen X, Ahmed S, Konishi K, Hamilton SR, Issa J-PJ: **Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer.** *Proc Natl Acad Sci USA* 2007, **104**:18654-18659.

47. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.** *Nat Genet* 2009, **41**:178-186.

48. Gonzalgo ML, Jones PA: **Mutagenic and epigenetic effects of DNA methylation.** *Mutat Res* 1997, **386**:107-118.

49. The Chimpanzee Sequencing and Analysis Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437**:69–87.

50. Forrest WF, Cavet G: **Comment on "The consensus coding sequences of human breast and colorectal cancers".** *Science* 2007, **317**:1500a.

51. Getz G, Höfling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES: **Comment on "The consensus coding sequences of human breast and colorectal cancers".** *Science* 2007, **317**:1500b.

52. Rubin AF, Green P: **Comment on "The consensus coding sequences of human breast and colorectal cancers".** *Science* 2007, **317**:1500c.

53. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**:D61-5.

54. Bird AP: **CpG-rich islands and the function of DNA methylation.** *Nature* 1986, **321**:209-213.

55. Barnes DE, Lindahl T: **Repair and genetic consequences of endogenous DNA base damage in mammalian cells.** *Annu Rev Genet* 2004, **38**:445-476.

56. Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, Klein J, Fridman E, Skarda J, Srovnal J, Hajduch M, Murray P, Kolar Z: **Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis.** *BMC Cancer* 2007, **7**:55.

57. Hanawalt PC, Spivak G: **Transcription-coupled DNA repair: two decades of progress and surprises.** *Nat Rev Mol Cell Biol* 2008, **9**:958-970.

58. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR: **Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification.** *Nucleic Acids Res* 2004, **32**:e71.

59. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Kober KM, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36**:D773-9.

60. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles--database and tools update.** *Nucleic Acids Res* 2007, **35**:D760-5.

61. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res* 2009, **19**:1316-1323.

62. **Cancer Genome Project** [http://www.sanger.ac.uk/genetics/CGP/Studies/Kinases/]

63. R DCT: **R: A Language and Environment for Statistical Computing.** 2008,

64. Gart JJ, Nam J: **Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness.** *Biometrics* 1988, **44**:323-338.

65. Grantham R: **Amino Acid Difference Formula to Help Explain Protein Evolution.** *Science* 1974,

66. Michalak P: **Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes.** *Genomics* 2008, **91**:243–248.

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

68. Kruglyak S, Tang H: **Regulation of adjacent yeast genes.** *Trends Genet* 2000, **16**:109–111.

69. Pál C, Hurst LD: **Evidence for co-evolution of gene order and recombination rate.** *Nat Genet* 2003, **33**:392–395.

70. Hurst LD, Williams EJB, Pál C: **Natural selection promotes the conservation of linkage of co-expressed genes.** *Trends Genet* 2002, **18**:604–606.

71. Poyatos JF, Hurst LD: **The determinants of gene order conservation in yeasts.** *Genome Biol* 2007, **8**:R233.

72. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN: **A gene expression map of the Arabidopsis root.** *Science* 2003, **302**:1956–1960.

73. Williams EJB, Bowles DJ: **Coexpression of neighboring genes in the genome of Arabidopsis thaliana.** *Genome Res* 2004, **14**:1060–1067.

74. Schmid M, Davison T, Henz S, Pape U, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann J: **A gene expression map of Arabidopsis thaliana development.** *Nat Genet* 2005, **37**:501–506.

75. Lercher MJ, Blumenthal T, Hurst LD: **Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes.** *Genome Res* 2003, **13**:238–243.

76. Blumenthal T, Gleason KS: **Caenorhabditis elegans operons: form and function.** *Nat Rev Genet* 2003, **4**:112–120.

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

78. Weber CC, Hurst LD: **Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation.** *Genome Biol* 2011, **12**:R23.

79. Meadows LA, Chan YS, Roote J, Russell S: **Neighbourhood continuity is not required for correct testis gene expression in Drosophila.** *PLoS Biol* 2010, **8**:e1000552.

80. Stolc V, Gauhar Z, Mason C, Halasz G, Batenburg M, Rifkin S, Hua S, Herreman T, Tongprasit W, Barbano P, Bussemaker H, White K: **A gene expression map for the euchromatic genome of Drosophila melanogaster.** *Science* 2004, **306**:655–660.

81. Mezey JG, Nuzhdin SV, Ye F, Jones CD: **Coordinated evolution of co-expressed gene clusters in the Drosophila transcriptome.** *BMC Evol Biol* 2008, **8**:2.

82. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31**:180–183.

83. Versteeg R, van S, Barbera D C, van B, Marinus F, Roos M, Monajemi R, Caron H, Bussemaker HJ, van K, Antoine H C: **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* 2003, **13**:1998–2004.

84. Li Q, Lee BTK, Zhang L: **Genome-scale analysis of positional clustering of mouse testis-specific genes.** *BMC Genomics* 2005, **6**:7.

85. Hentges KE, Pollock DD, Liu B, Justice MJ: **Regional variation in the density of essential genes in mice.** *PLoS genetics* 2007, **3**:e72.

86. Sémon M, Duret L: **Evolutionary origin and maintenance of coexpressed gene clusters in mammals.** *Molecular biology and evolution* 2006, **23**:1715–1723.

87. Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH: **Clusters of co-expressed genes in mammalian genomes are conserved by natural selection.** *Mol Biol Evol* 2005, **22**:767–775.

88. Liao B-Y, Zhang J: **Coexpression of linked genes in Mammalian genomes is generally disadvantageous.** *Mol Biol Evol* 2008, **25**:1555–1565.

89. Al-Shahrour F, Minguez P, Marquás-Bonet T, Gazave E, Navarro A, Dopazo J: **Selection upon genome architecture: conservation of functional neighborhoods with changing genes.** *PLoS Comput Biol* 2010, **6**:e1000953.

90. Lee J, Sonnhammer E: **Genomic gene clustering analysis of pathways in eukaryotes.** *Genome Res* 2003, **13**:875–882.

91. Karpen GH: **Position-effect variegation and the new biology of heterochromatin.** *Curr Opin Genet Dev* 1994, **4**:281–291.

92. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G: **Locus control regions.** *Blood* 2002, **100**:3077–3086.

93. Bonifer C: **Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required?** *Trends Genet* 2000, **16**:310–315.

94. Levine M: **Transcriptional enhancers in animal development and evolution.** *Curr Biol* 2010, **20**:R754–63.

95. Fukuoka Y, Inaoka H, Kohane IS: **Inter-species differences of co-expression of neighboring genes in eukaryotic genomes.** *BMC Genomics* 2004, **5**:4.

96. Zhan S, Horrocks J, Lukens L: **Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains.** *Plant J* 2006, **45**:347–357.

97. Chen N, Stein LD: **Conservation and functional significance of gene topology in the genome of Caenorhabditis elegans.** *Genome Res* 2006, **16**:606–617.

98. Moltó E, Fernández A, Montoliu L: **Boundaries in vertebrate genomes: different solutions to adequately insulate gene expression domains.** *Brief Funct Genomic Proteomic* 2009, **8**:283–296.

99. Reddy KL, Zullo JM, Bertolino E, Singh H: **Transcriptional repression mediated by repositioning of genes to the nuclear lamina.** *Nature* 2008, **452**:243–247.

100. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295**:1306–1311.

101. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465**:363–367.

102. Lanctôt C, Cheutin T, Cremer M, Cavalli G, Cremer T: **Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions.** *Nat Rev Genet* 2007, **8**:104–115.

103. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van B, Nynke L, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485**:381–385.

104. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148**:458–472.

105. Chintapalli VR, Wang J, Dow JAT: **Using FlyAtlas to identify better Drosophila melanogaster models of human disease.** *Nat Genet* 2007, **39**:715–720.

106. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: **An abundance of bidirectional promoters in the human genome.** *Genome Res* 2004, **14**:62–66.

107. Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RAH, Stein L, Henikoff S, Kellis M, White KP: **A comprehensive map of insulator elements for the Drosophila genome.** *PLoS Genet* 2010, **6**:e1000814.

108. Filion GJ, van B, Joke G, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de C, Inês J, Kerkhoven RM, Bussemaker HJ, van S, Bas: **Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.** *Cell* 2010, **143**:212–224.

109. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ: **Ensembl 2012.** *Nucleic Acids Res* 2012, **40**:D84–90.

110. Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ, consortium F: **FlyBase: improvements to the bibliography.** *Nucleic Acids Res* 2013, **41**:D751–7.

111. **RepeatMasker Open-3.3.0** [http://www.repeatmasker.org]

112. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249–264.

113. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**:R80.

114. Gough B: **GNU Scientific Library.** Network Theory.; 2009.

115. R CT: **R: A Language and Environment for Statistical Computing.** 2013,

116. **Argtable2** [http://argtable.sourceforge.net/doc/argtable2.html]

117. **LibDS 2.1** [http://libds.sourceforge.net/doc/index.html]

118. **bzip2 and libbzip2, version 1.0.6** [http://www.bzip.org]

119. **Jansson 2.4** [http://www.digip.org/jansson/]

120. Schwarz G: **Estimating the Dimension of a Model.** *Ann Stat* 1978, **6**:461–464.

121. Georgi B, Costa IG, Schliep A: **PyMix--the python mixture package--a tool for clustering of heterogeneous biological data.** *BMC Bioinf* 2010, **11**:9.

122. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841–842.

123. Dale RK, Pedersen BS, Quinlan AR: **Pybedtools: a flexible Python library for manipulating genomic datasets and annotations.** *Bioinformatics* 2011, **27**:3423–3424.

124. **GO Slim** [http://www.geneontology.org/GO.slims.shtml]

125. **goatools** [https://github.com/tanghaibao/goatools]

Vita

Alan Frederick Rubin was born in Boston, Massachusetts, the only child of Gerald M. Rubin and Lynn M. Rubin. The family moved to Baltimore, Maryland six months after his birth, and finally settled in Berkeley, California when he was three years old. He graduated from the Head-Royce School in Oakland, California in 1998. He attended the University of California, Los Angeles for two years and subsequently received his Bachelor of Science degree from Pacific University in Forest Grove, Oregon in 2006 with a major in Bioinformatics. In September of 2006 he entered graduate school at the University of Washington, completing his Doctor of Philosophy in June of 2013.