adequate power (80%) a simple 'rule of thumb' would be that a ~10% decrease in phenotypic accuracy would lead to a ~20% reduction of statistical power (Figure 1a). In a practical sense, it could be helpful to think of the relationship in reverse, and ask the question: how many additional cases are required to counteract the consequences of the phenotypic error? Based on calculations using the PAWE program [3] we show that to maintain adequate power (>80%) sample size would need to increase by 20% in order to offset a diagnostic inaccuracy of 10% (Figure 1b). Whether it is better to increase the sample size or concentrate on improving the phenotypic accuracy will depend on the relative cost of improving the phenotypic accuracy of the existing cohort versus the costs of further genetic analysis on a less well-phenotyped group of cases.

References

1 Lee, K. and Sawcer, S.J. (2010) Does phenotype accuracy really limit the horizon when detecting new disease genes? *Trends Genet.* This issue
2 Samuels, D.C. *et al.* (2009) Detecting new neurodegenerative disease genes: does phenotype accuracy limit the horizon? *Trends Genet.* 25, 486–488
3 Edwards, B.J. *et al.* (2005) Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet.* 6, 18

0168-9525/$ – see front matter
doi:10.1016/j.tig.2010.03.001 Trends in Genetics 26 (2010) 242–243

Genome Analysis

Mutational bias shaping fly copy number variation: implications for genome evolution

Margarida M. Cardoso-Moreira[1,2,3*] and Manyuan Long[1]

[1] Department of Ecology and Evolution, University of Chicago, 1101 E 57th St. 60637 Chicago, IL, USA
[2] Graduate Program in Areas of Basic and Applied Biology, Universidade do Porto, Porto, Portugal
[3] Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Copy number variants (CNVs) underlie several genomic disorders and are a major source of genetic innovation. Consequently, any bias affecting their placement in the genome will impact our understanding of human disease and genome evolution. Here we report a mutational bias affecting CNVs that generates different probabilities of duplication and deletion across the genome in association with DNA replication time. We show that this mutational bias has important consequences for genome evolution by leading to different probabilities of gene duplication for different classes of genes and by linking the probability of gene duplication with the transcriptional activity of genes.

An expanded view of genetic variation

In the last five years it was discovered that a large proportion of the genetic variation found within species lies in differences in the number of copies of DNA segments – CNVs (i.e. polymorphic duplications and deletions) [1–3]. The pervasiveness of CNVs propelled their study to the forefront of medical and genetic research because their two-fold potential to underlie disease and to be a major source of genetic innovation was immediately recognized [4,5]. Understanding the mechanisms governing the placement of CNVs along the genome is a crucial goal of CNV research because identifying these mechanisms will impact upon our understanding of genome evolution and potentially aid in the development of more precise medical diagnostic tools.

Duplications and deletions are differentially distributed across the genome

A high-resolution map of CNVs was recently generated for the genome of the fruit fly *Drosophila melanogaster* [3], and this provided the first appropriate dataset with which to investigate the mechanisms governing the genomic placement of CNVs in this species. The map comprises 2211 duplications and 1428 deletions identified in a survey of 15 natural populations of *D. melanogaster* using the published genomic sequence as a reference (see supplementary material online).

Using this dataset we independently calculated the density of duplications and deletions across each chromosome arm (supplementary material online). Surprisingly, we found the two to be differentially distributed. The highest density regions (HDRs, correspond to the regions exhibiting the 10% most extreme density values) of duplications and deletions do not overlap outside the pericentromeric region (Figure 1a). In total, we identifed 47 HDRs (25 deletion-HDRs and 22 duplication-HDRs) that encompass 25% of the analyzed genome (HDR coordinates are given in Table S1 in the supplementary material online). In fact, we found duplication and deletion densities to be also negatively correlated outside these regions [Pearson correlation coefficient: -0.21, 95% confidence interval (CI): -0.23 to -0.20, $P < 2e^{-16}$]. These data show that regions of the genome enriched with duplications (duplication-HDRs; shown in red in Figure 1a) are depleted of deletions whereas regions enriched with deletions (deletion-HDRs; shown in grey in Figure 1a) are depleted of duplications.

Corresponding author: Cardoso-Moreira, M.M. (mmc256@cornell.edu); Long, M. (mlong@uchicago.edu).
* Present address: Department of Molecular Biology and Genetics, Cornell University, 107 Biotechnology Building, 14853-2703 Ithaca, NY, USA.

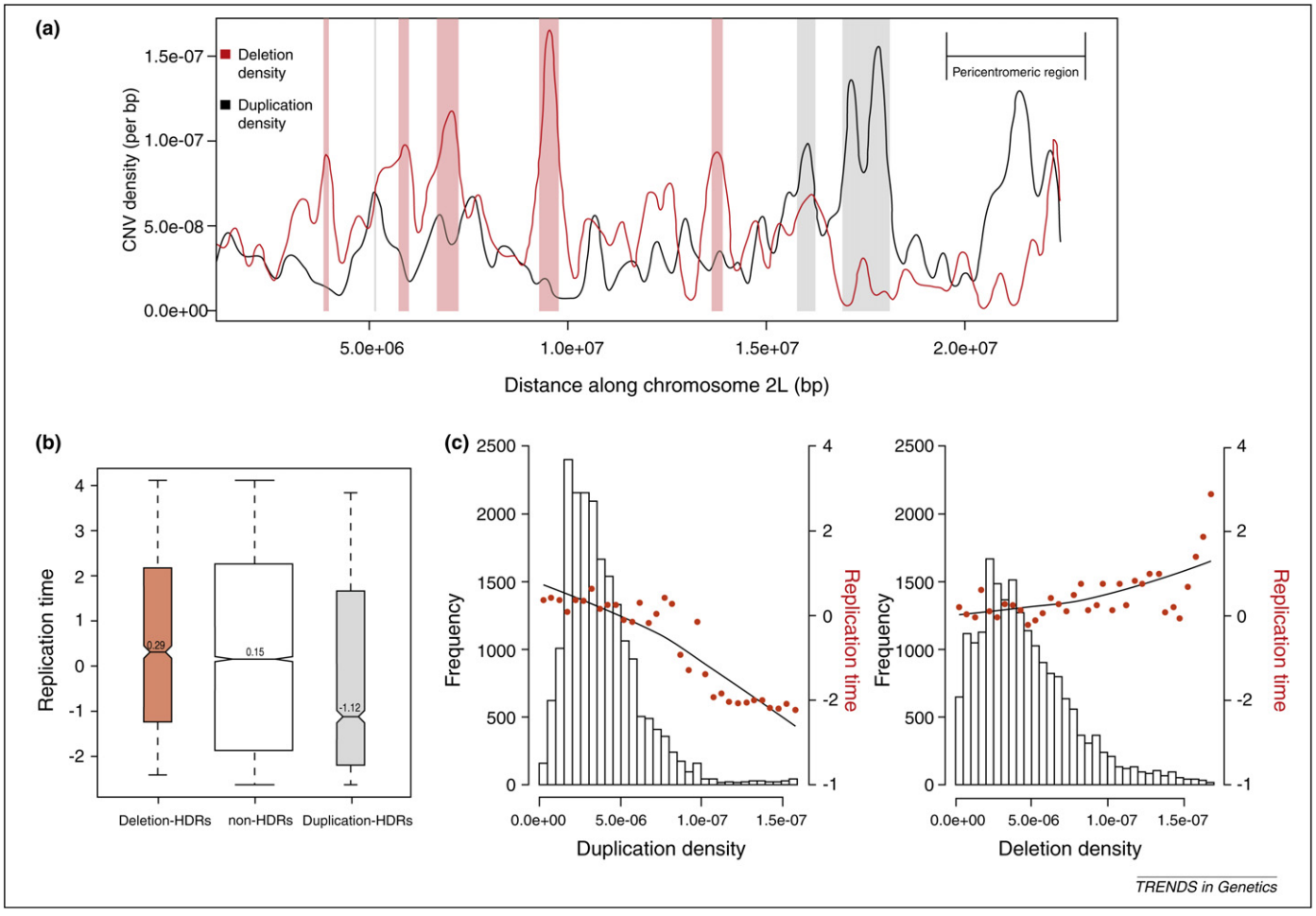

Figure 1. Relationship between the differential distribution of duplication and deletion densities and replication time. **(a)** Density of duplications and deletions along *D. melanogaster* chromosome 2L. CNV data were produced by Emerson and colleagues [3]. The red blocks correspond to deletion-HDRs and the grey blocks to duplication-HDRs. HDRs correspond to the chromosome regions with the 10% most extreme CNV density values (after excluding pericentromeric regions). **(b)** Replication time profiles of duplication-HDRs and deletion-HDRs. The boxplots represent the distribution of replication time values for duplication-HDRs (in grey), non-HDRs (in white) and deletion-HDRs (in red). The width of the boxplot is proportional to the number of observations in each group. **(c)** Association between replication time and duplication density and deletion density. For each interval of duplication and deletion density (as defined by the histogram in grey) we calculated the mean observed replication time (red dots). The curve was produced using the R function 'scatter.smooth' [23]. In **(b)** and **(c)** Replication time data are from Schwaiger and colleagues [12].

This observation is surprising because one would expect to find regions rich in CNVs to be enriched in both duplications and deletions. However, there could be a simple explanation for this observation. Because purifying selection acts more strongly on deletions involving genes than on duplications [3], we could be observing the consequences of deletions being removed from gene regions more often than duplications. To the contrary, we found that deletion-HDRs have a significantly higher gene density than duplication-HDRs (median gene density: $1.2\text{x}10^{-4}$ versus $9.3\text{x}10^{-5}$ genes/bp, respectively; Wilcoxon rank sum test, $P = 0.02$), thereby ruling out differences in selective pressures as the explanation for the different distribution of duplications and deletions across the genome. We note that most deletions fell within introns, thus avoiding coding sequences [3].

We also addressed the possibility that the differential distribution of duplications and deletions could be a consequence of the genomic distribution of segmental duplications (SDs) and transposable elements (TEs) (see supplementary material online). SDs and TEs contribute to CNV formation in mammals [1,2,6,7] by facilitating the occurrence of non-allelic homologous recombination (NAHR) and, in the case of TEs, the occurrence of non-homologous end-joining (NHEJ) [7]. Both pathways can generate CNVs but, whereas NAHR is predicted to generate duplications and deletions, NHEJ is predicted to generate predominantly deletions [6–8]. Hence, the distribution of SDs and TEs across the fly genome could underlie the differential distribution of duplications and deletions. However, because SDs and TEs are much more abundant in mammals than in flies, where they are mostly restricted to regions with very low rates of crossing over (pericentromeric regions and the 4$^{\text{th}}$ chromosome [9,10] that were excluded from this study), this explanation is unlikely to fully account for the observed pattern. Accordingly, TEs and SDs were found to be associated with only a minority of CNVs outside pericentromeric regions (M.C.M., R. Arguello and M.L. unpublished) and when we excluded these CNVs from our dataset our observations remained unchanged (data not shown). Consequently, we also ruled out the genomic distribution of SDs and TEs as the explanation for the differential distribution of duplications and deletions across the genome.

Duplication and deletion densities are associated with replication time

The association between gene density and the distribution of duplications and deletions made us suspect a possible link with DNA replication time because DNA replication time is associated with gene density. DNA is replicated following a tightly regulated time program that appears to be conserved between cell types for a large portion of the genome [11,12]. Early-replicating regions tend to be gene-rich whereas late-replicating regions tend to be gene-poor. Importantly, a link has already been established between replication time and human single base pair mutation rates [13].

We tested for an association between the distribution of duplications and deletions and replication time using a high-resolution genome-wide replication time profile recently generated for *D. melanogaster* [12]. The results reported below refer to the replication time profile of Kc cells but all results remained the same when we used the replication time profile of Cl8 cells instead or when we restricted our analyses to those regions in the genome with similar replication time profiles between these two cell lines (see supplementary material online). Low replication time values (minimum is -4) indicate late-replicating regions whereas high replication time values indicate early-replicating regions (maximum is +4).

The replication time profile of duplication-HDRs and deletion-HDRs is significantly different: duplication-HDRs tend to be associated with later-replicating times and deletion-HDRs with earlier-replicating times (median replication time: −1.12 versus +0.29, respectively; Wilcoxon rank sum test, $P = 2e^{-16}$) (Figure 1b). Whereas deletion-HDRs tend to be replicated significantly earlier than the rest of the genome (median replication time: +0.29 versus +0.15, respectively; $P = 3e^{-14}$), duplication-HDRs tend to be replicated significantly later (median replication time: -1.12 versus +0.15, respectively; $P = 2e^{-16}$) (Figure 1b). Accordingly, within duplication-HDRs 51% of the sequences are classified as late-replicating, in contrast to 36% in non-HDRs and 29% in deletion-HDRs [duplication-HDRs versus non-HDRs: $\chi^2 = 164$, 2 degrees of freedom (df), $P < 2e^{-16}$; duplication-HDRs versus deletion-HDRs: $\chi^2 = 204$, 2 df, $P < 2e^{-16}$] (Figure S1 in the supplementary material online).

The association between duplication and deletion density and replication time is not restricted to duplication- and deletion-HDRs. Figure 1c shows the relationship between replication time and duplication and deletion density for the whole genome. We binned duplication and deletion density in equally sized intervals (histogram) and for each density interval we calculated the mean replication time (red dots). Using these data we found a strong negative correlation between duplication density and replication time (Pearson correlation coefficient: −0.9; CI, −0.96 to −0.82; $P = 6e^{-13}$) and a positive correlation between deletion density and replication time (Pearson correlation coefficient: +0.6; CI, 0.34 to 0.78; $P = 0.0001$). It is important to note that if we instead perform a correlation between duplication and deletion density and replication time across all data points the correlation coefficients are much weaker. This is because the strong association between CNV density and replication time is only observed in genomic regions that have different densities of duplications and deletions. For large stretches of the genome, duplication and deletion densities are similar and in these regions the association reported above is very weak (Figures S2 and S3 in the supplementary material online).

These data point towards a biased mutational process underlying CNV formation that is linked with replication time through a mechanism that is so far unknown. This mutational bias results in higher probabilities of duplication towards late-replicating regions and higher probabilities of deletion towards early-replicating regions.

Different rates of gene duplication for different classes of genes

Because replication time and gene distribution are correlated in the genome, this CNV mutational bias is expected to impact upon genome evolution. An example of this impact is illustrated by genes with sexually dimorphic expression, usually referred to as sex-biased genes (i.e. male-, female- or unbiased genes) [14] that we found not to be distributed randomly along the autosomes with regard to replication time (supplementary material online). The classification of genes as female-, male- and un-biased was retrieved from the Sebida database [15]. Male-biased genes tend to be replicated later than female-biased and unbiased genes (median replication time: +0.2 versus +1.4 and +0.2 versus +0.4, respectively; Wilcoxon rank sum test, $P < 2e^{-16}$ and $P = 9e^{-9}$), whereas female-biased genes tend to replicate earlier than unbiased genes ($P < 2e^{-16}$) (Figure S4 in the supplementary material online). Accordingly, 51% of female-biased genes are early-replicating versus 37% of male-biased genes, whereas 32% of male-biased genes are late-replicating versus only 16% of female-biased genes ($\chi^2 = 155$, 2 df, $P < 2e^{-16}$) (Figure S5 in the supplementary material online). As predicted by their replication time profile, male-biased genes are over-represented in duplication-HDRs (8% increase, Fisher's exact test, $P = 0.001$).

The increase in the probability of gene duplication for genes located in later-replicating regions, here exemplified by male-biased genes, is predicted to have two consequences: (i) these genes will have an increased number of fixed paralogs, and (ii) some of these will show increased intraspecific variation of gene expression due to the presence of duplications and the resulting dosage variation [16]. In agreement with these two predictions: (i) male-biased genes have been shown to have a disproportionately higher number of paralogs in the genome than female-biased or un-biased genes [14,15], and (ii) male-biased genes have been shown to have higher intraspecific variation in gene expression levels [17]. Although positive selection was suggested to contribute to these two phenomena [14], we suggest that the higher duplication rates experienced by male-biased genes might also play a role.

A link between probability of gene duplication and gene expression

The CNV mutational bias is expected to further impact upon genome evolution because of the relationship between replication and transcription mechanisms

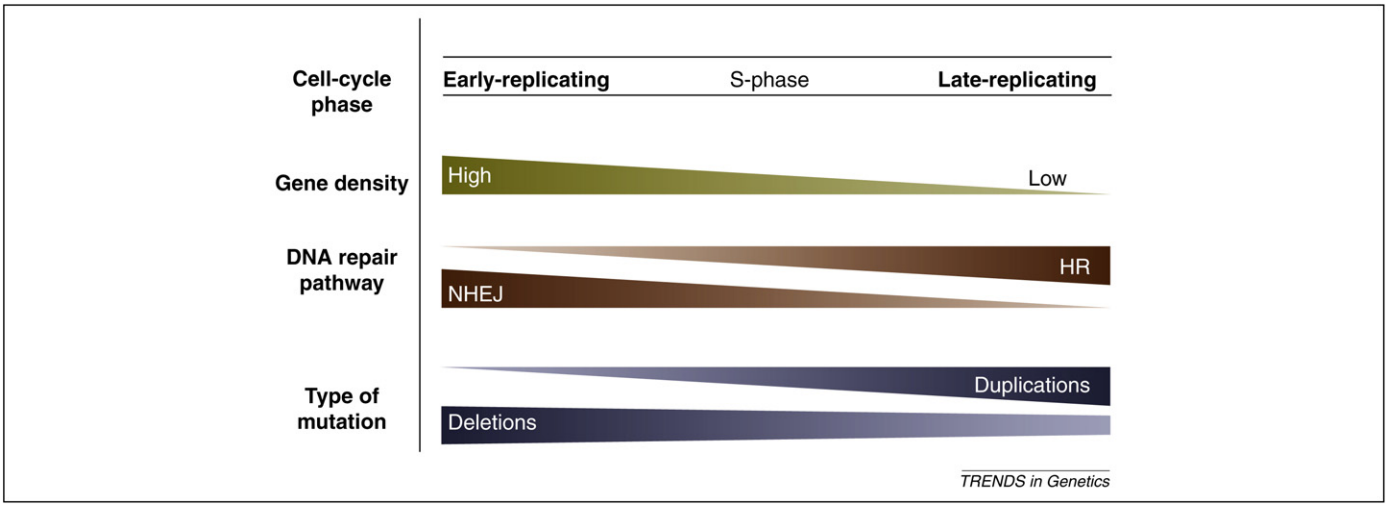

Figure 2. A mechanistic model based on the interplay between DNA replication and DNA repair. Schematic representation of how gene density, DNA repair pathways and types of CNVs are expected to vary along the S-phase of the cell cycle (i.e. during DNA replication). Although late-replicating regions are indicated by low numbers we chose to depict the cell cycle as progressing from left to right (with late-replication on the right).

[11,12,18]. Although it is not clear why replication time is associated with transcriptional activity, compelling evidence in humans and flies suggests the two are connected [11,18]. Whereas early-replicating regions are associated with regions of higher transcriptional permissiveness, and are therefore enriched for genes with broader expression patterns, late-replicating regions are enriched with genes with restricted transcriptional activity [11,18]. Hence, the CNV mutational bias is predicted to lead to higher probabilities of gene duplication for genes with restricted expression patterns than for genes with wider transcriptional activity.

Accordingly, we found using data from FlyAtlas [19] that genes located in duplication-HDRs are expressed in a significantly lower number of tissues when compared to the whole-genome (median number of tissues: 9 versus 15, respectively; Wilcoxon rank sum test, $P = 5e^{-6}$) (Figure S6 in the supplementary material online).

A mechanistic model based on the interplay between DNA replication and DNA repair

CNVs result from the formation of DNA double strand breaks (DSBs) [6,20]. Several processes create DSBs in the germ line, most notably DNA replication [6]. Because broken DNA affects cell viability and genomic stability such lesions are promptly repaired. The two main mechanisms to repair DSBs are homologous recombination (HR) that requires extensive sequence similarity to perform the repair, and non-homologous end-joining (NHEJ) that requires little or no sequence similarity [6,20,21]. Both HR and NHEJ can generate CNVs as a by-product of fixing DSBs [6–8].

Why would there be a link between the distribution of duplications and deletions across the genome and replication time? We hypothesize that this might be at least partly due to two characteristics of these two DNA repair mechanisms. First, there is evidence suggesting that the prevalence of HR and NHEJ might change throughout the cell cycle, namely during the S-phase (when DNA is replicated) [6,20,21]. Because the efficiency of HR is influenced by DNA template accessibility, and because the sister chromatid is the preferred template, HR is thought to be the dominant repair pathway during late-S and G2 phases of the cell cycle [6,20,21]. During G1 and early-S phases the dominant pathway is thought to be NHEJ [6,20,21] (Figure 2). Second, HR and NHEJ generate different types of CNVs. Whereas NHEJ predominantly creates deletions, HR generates both types of CNVs [6–8] (Figure 2). If these two phenomena apply to *Drosophila*, one predicts an enrichment of deletions in early-replicating regions where NHEJ is the dominant repair pathway, and the presence of both duplications and deletions in late-replicating regions where HR dominates. Furthermore, because purifying selection acts more strongly on deletions than on duplications [3], late-replicating regions would be expected to show an excess of duplications (Figure 2).

Although this mechanistic model qualitatively fits the observed data, it remains speculative. The data suggesting the differential use of HR and NHEJ during DNA replication were not obtained in *Drosophila*, and molecular data are also lacking for potential differences between mitotic and meiotic cells and for the use of different repair pathways in response to different types of DSBs. However, the value of this model lies in the fact that it is based on explicit assumptions that can easily be tested as more molecular data become available.

Concluding remarks

We have shown that the probabilities of duplication and deletion vary considerably across the *Drosophila* genome, that they are negatively correlated, and that they show an association with replication time. This association has important consequences for genome evolution because it predicts that some classes of genes will experience different rates of duplication and that genes with different transcriptional profiles will also mutate at different rates. The implications for genome evolution might be even more far-reaching if a link between replication time and epigenetic remodeling is also established [22].

The small and compact fly genome has allowed for an unprecedented high-resolution description of replication time and CNV density across the genome. As high-resolution data are gathered for other eukaryotic genomes (namely mammalian genomes), it will be possible to evaluate the generality of this CNV mutational bias and further study its consequences for genome evolution and human health.

Acknowledgments

We thank Roman Arguello, Hedibert Lopes, Maria Vibranovski and Beatriz Viçoso for critical discussion and reading of the manuscript and anonymous reviewers for improving the quality of the work. M.C.M. was funded by the Portuguese Foundation for Science and Technology (POCI 2010, FSE) and M.L. by the Packard Fellowship for Science and Engineering and the National Institutes of Health (R0IGM065429-0IA1 and R0IGM078070-0IAI).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2010.03.002.

References

1 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature* 444, 444–454
2 She, X. *et al.* (2008) Mouse segmental duplication and copy number variation. *Nat. Genet.* 40, 909–914
3 Emerson, J.J. *et al.* (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* 320, 1629–1631
4 Beckmann, J.S. *et al.* (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.* 8, 639–646
5 Perry, G.H. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256–1260
6 Sankaranarayanan, K. and Wassom, J.S. (2005) Ionizing radiation and genetic risks XIV. Potential research directions in the post-genome era based on knowledge of repair of radiation-induced DNA double-strand breaks in mammalian somatic cells and the origin of deletions associated with human genomic disorders. *Mutat. Res.* 578, 333–370
7 Shaw, C.J. and Lupski, J.R. (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* 13, R57–R64
8 Aguilera, A. and Gómez-González, B. (2008) Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.* 9, 204–217
9 Fiston-Lavier, A.S. *et al.* (2007) A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res.* 17, 1458–1470
10 Bergman, C.M. *et al.* (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* 7, R112
11 Donaldson, A.D. (2005) Shaping time: chromatin structure and the DNA replication programme. *Trends Genet.* 21, 444–449
12 Schwaiger, M. *et al.* (2009) Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. *Genes Dev.* 23, 589–601
13 Stamatoyannopoulos, J.A. *et al.* (2009) Human mutation rate associated with DNA replication timing. *Nat. Genet.* 41, 393–395
14 Ellegren, H. and Parsch, J. (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* 8, 689–698
15 Gnad, F. and Parsch, J. (2006) Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* 22, 2577–2579
16 Stranger, B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848–853
17 Meiklejohn, C.D. *et al.* (2003) Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9894–9899
18 Schwaiger, M. and Schübeler, D. (2006) A question of timing: emerging links between transcription and replication. *Curr. Opin. Genet. Dev.* 16, 177–183
19 Chintapalli, V.R. *et al.* (2007) Using FlyAtlas to identify better *Drosophila* models of human disease. *Nat. Genet.* 39, 715–720
20 Sonoda, E. *et al.* (2006) Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair (Amst)* 5, 1021–1029
21 Branzei, D. and Foiani, M. (2008) Regulation of DNA repair throughout the cell cycle. *Nat. Rev. Mol. Cell Biol.* 9, 297–308
22 Göndör, A. and Ohlsson, R. (2009) Replication timing and epigenetic reprogramming of gene expression: a two-way relationship? *Nat. Rev. Genet.* 10, 269–276
23 R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. (http://www.R-project.org/)

doi:10.1016/j.tig.2010.03.002 Trends in Genetics 26 (2010) 243–247