Genomics 97 (2011) 200–204

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Insights into the molecular correlates modulating functional compensation between monogenic and polygenic disease gene duplicates in human

Soumita Podder, Tapash Chandra Ghosh *

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

ARTICLE INFO

Article history:
Received 29 October 2010
Accepted 16 January 2011
Available online 31 January 2011

Keywords:
Monogenic disease
Polygenic disease
Expression profile similarity
Functional buffering
Duplicates

ABSTRACT

Functional redundancy by gene duplication appears to be a common phenomenon in biological system and hence understanding its underlying mechanism deserves much attention. Here, we investigated the differences between functional compensation of monogenic and polygenic disease genes which are unexplored till date. We found that the competence of functional buffering varies in the order of non-disease genes>monogenic disease genes>polygenic disease genes. This fact has been explained by the sequence identity, expression profile similarity, shared interaction partners and cellular locations between duplicated pairs. Moreover, we observed an inverse relationship between backup capacity and the non-synonymous substitution rate of disease and non-disease genes while the opposite trend is found for their corresponding paralogs. Logistic regression analysis among sequence identity, sharing of expression profile, interaction partners and cellular locations with backup capacity between duplicated pairs demonstrated that the sharing of expression profile is the most dominant regulator of backup capacity.

1. Introduction

Gene duplication is considered as an important prerequisite for gene innovation which facilitates functional divergence through neofunctionalization or subfunctionalization [1] or may result in functionally redundant genes [2,3]. The importance of redundant duplicates in understanding the mechanisms for resilience against mutations is increasingly appreciated, as illustrated by the initiative aiming to identify the factors behind the mechanism. Since higher metazoan genomes have duplicates with several overlapping functions in the same genome, loss-of-function in one gene will thus have little phenotypic effect if there are additional genes with similar functions. Hence it seems plausible that redundancy among duplicates is one of the main mechanisms executing robustness against deleterious mutations [4]. Particularly in vertebrates many genes with similar functions are leftovers of ancient (>400 Myr) gene or genomic duplications [5,6]. Even unicellular eukaryotes retain distantly related gene pairs with similar functions [7]. Of late, the strong anti-correlation between the fitness effect of a deleted gene and the sequence similarity of its closest paralogs has also been observed in *S. cerevisiae*, demonstrating that genes with highly similar paralogs are compensated for mutations more often than genes with distant paralogs [3].

The human genome is comprised with disease and non-disease genes. Degree of duplicates as backups against gene loss varies widely within organism. In human, disease genes are less likely to have similar paralogs since they could rescue the mutant phenotype [8]. Several molecular correlates such as organismal complexity, genomic function, sequence similarity, expression profile similarity, protein interaction profile and sharing of regulatory elements have been reported to be causative agents for functional buffering by gene duplicates [9–11], though contributions of some of the factors are still debated. From an evolutionary standpoint redundant genes are relaxed to accumulate mutations since mutations in those genes would have no effect on the phenotype of the corresponding organism [12,13].

Till date, the post duplication fates of genes have been studied in the ground of non-disease and disease genes where disease genes are exemplified as inherited by Mendelian fashion only. Currently the rising incidence of polygenic diseases that do not follow the Mendelian pattern of inheritance demands an in depth study of complex or polygenic disorders. Taking advantage of the Human Gene Mutation Database (HGMD) and Genetic Association Database (GAD), the reliable archives of human monogenic and complex diseases respectively, we examined the fundamental differences that exist between functional compensation for monogenic (MD) and polygenic diseases (PD) with respect to non-disease (ND) genes. Our studies reveal that monogenic disease genes are functionally more buffered by their duplicates than polygenic disease genes, a feature that was unexplored so far. The role of expression profile similarity, interacting partner and regulatory motif sharing, and sub-localization pattern between the duplicated pairs have been elucidated as the underlying factors that escort the functional buffering variation between disease and non-disease genes.

Abbreviations: ND, non-disease genes; MD, monogenic disease genes; PD, polygenic disease genes.

* Corresponding author. Fax: +91 33 2355 3886.
E-mail address: tapash@boseinst.ernet.in (T.C. Ghosh).

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.01.004

2. Results

2.1. Variation in functional compensation between MD and PD duplicated pairs

Investigation on the molecular genetics of disease genes raises a key question, why severe mutations often do not result in a detectable abnormal phenotype. Previous report [14] illustrates the contribution to functional compensation by gene duplicates against deleterious human mutation. Therefore, it would be interesting to investigate whether any fundamental differences exist between functional compensation of monogenic and polygenic disease gene duplicates with respect to non disease duplicates. We examined functional compensation by calculating the functional buffering using GO terms in MD, PD and ND duplicated pairs respectively and observed that MD duplicates encompass significantly more functional backup than PD duplicates (Fig. 1). We also found that ND duplicates themselves have more backup capacity (35.42%) than both the classes of disease gene duplicates (Fig. 1). While separating the ND genes into housekeeping (HK) and other (OTH) genes, it was observed that housekeeping duplicates are functionally most redundant among all categories of genes, and other ND duplicated genes show intermediate functional compensation capacity between MD and PD duplicates (Fig. 1). Formerly, the significant contribution of close sequence homologs to genetic robustness against deleterious mutations has been demonstrated in the organisms like *C. elegans* as well as in human [3]. Additionally it has been proposed that close sequence paralogs are about 2–3 times less likely to harbor diseases [14]. In our dataset we observed that the sequence identity between duplicated pairs is lower in the case of disease genes as compared to non-disease genes (Disease = 41% [MD = 42%, PD = 36%], ND = 49% [HK = 55%, OTH = 39%] and each value was significant with each other at least at 0.05 level in M–W test) which is in agreement with the previous observation [15]. This result may provide a probable explanation for the observed decrease in backup capacity with decreased sequence identity.

2.2. Functional buffering in MD vs. PD duplicates: role of expression profile similarity

Intuitively, it might be expected that co-expressed paralogs would be more likely to be functionally redundant compared to those whose expression profiles differed significantly. To explain the disparity in backup capacity among MD, PD and ND duplicates, we measured the expression profile similarity between each protein and its paralogs. Unexpectedly, our results show an opposite trend of association between expression profile similarity and backup capacity since we noticed a significant gradual increase of expression profile similarity from PD to HK duplicates (PD = 0.187, MD = 0.160, ND = 0.174 [HK = 0.150, OTH = 0.179] and each value was significant with each other at least at 0.05 level in M–W test). Previous study [10] reasonably explained that in the case of remote paralogs (ds > 1), the backup capacity of a gene decreases with increased coexpression with its duplicated pairs. Then the proportion of close (ds < 1) and remote duplicates in our datasets including both disease and non-disease genes was measured and it was observed that the duplicated genes selected for our study mostly belong to the remote duplicates category (total close duplicates = 15.52% and remote duplicates = 84.48%; proportion of remote duplicates in PD, MD, HK and OTH are 85.5%, 78.72%, 62.95%, and 81.68% respectively). Thus, the expression profile similarity may be one of the aspects that can guide the varying backup capacity among PD, MD and ND duplicates. Furthermore, the earlier study [10] proposed that the regulatory motifs those are partially overlapped with paralogs have the most efficient backup activity compared to paralogs with highly similar or dissimilar sets of motifs since such an arrangement reconciles the differential expression of paralogs to provide compensation when needed. Investigation on the regulatory motifs arrangement reveals that MD and HK duplicated pairs overlap 39.55% and 47.01% with the regulatory motifs of considered disease and non-disease genes respectively whereas OTH and PD duplicated pairs have highly similar (82.1%) and dissimilar regulatory motifs (4.35%) respectively (difference between each of the above values was significant at least at 0.05 level in M–W test). This result has been interpreted as evidence that the functional compensation by disease and non-disease gene duplicates varies due to their differential sharing pattern of regulatory motifs that indeed modulates the expression profile similarity between duplicated pairs.

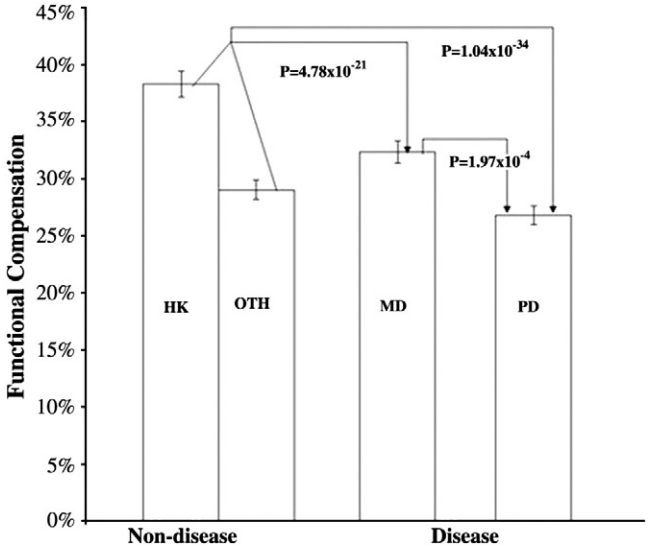

Fig. 1. Differences in functional compensation among the disease and non-disease genes with P value that indicates Mann–Whitney significance level.

2.3. Functional compensation in MD vs. PD duplicates: influence of sublocalization and protein interacting partners

Formerly, it was revealed that after duplication relocalization to new compartments (neolocalization) triggers functional diversification of duplicated proteins to endorse adaptation in their new subcellular environments [16,17]. According to this concept we intended to assess whether such subcellular reprogramming could explain the functional compensation variation among the gene classes or not. We have detected that significantly higher fraction of MD duplicates follows sublocalization compared to PD duplicates and among them paralogs of HK genes are mostly preserved HK gene localization patterns than any other subsets of duplicates which is also echoed in the previous study (Table 1) [17].

This subcellular redistribution of duplicates should often entail changes in their interactions with other proteins [17]. Moreover, correlated interaction profiles have been shown to be a strong determinant of shared functionality [11,18–20]. Hence we assessed the degree of interacting proteins shared by paralog pairs and observed that

Table 1
Comparison of sharing the subcellular localization among four classes of duplicated pairs.

Gene classes	% of duplicated proteins undergo sublocalization
MD	43.47
PD	35.10
OTH	39.51
HK	55.02

N.B. Z-score = 6.8, confidence level = 99% (MD vs. PD); Z-score = 7.47, confidence level = 97% (PD vs. OTH); Z-score = 1.56, confidence level = 94.1% (OTH vs. MD); Z-score = 1.23, confidence level = 95.6% (HK vs. MD); Z-score = 2.1, confidence level = 95.2% (HK vs. PD); Z-score = 3.91, confidence level = 98.7% (HK vs. OTH).

the propensity of sharing protein interaction partners is significantly higher in HK followed by MD, OTH and PD paralog pairs (Table 2). From this analysis it can be interpreted that the duplicate copies which are compelled to share the subcellular location of their corresponding disease and non-disease proteins are bound to interact with the other proteins specific for the aforementioned gene group to accomplish same biological processes. Therefore it could be suggested that the capacity of proteins to share the localization as well as the interaction partners of the disease and non-disease proteins plays a vital role in determining the variation of functional compensation between these groups of duplicated genes.

2.4. Functional redundancy of duplicated pairs: insight from evolutionary perspective

To investigate the relationship between backup capacity and non-synonymous substitution rate (*dn*) of disease, non-disease genes and their respective paralogs separately, we measured the *dn* of protein coding sequence by the realistic evolutionary models [21] among these groups of genes. Interestingly, we observed markedly different trends in the association between backup capacity and *dn* for disease, non-disease genes and for their respective duplicates. In the case of disease and non-disease genes, *dn* decreases with increased backup capacity of genes while the backup capacity of paralogs, measured within the paralogous pairs of gene was noticed to be increased with increasing *dn* of paralogs (Fig. 2).

2.5. Variation in functional compensation among different subsets of duplicated pairs: relative influence of the four factors

In order to investigate whether the aforementioned parameters independently influence the gene backup capacity we performed logistic regression analysis. Since not all the variables belong to the same category (some are continuous and some are binary), we grouped all the continuous variables (backup capacity, coexpression, sequence identity, and shared interaction) into two clusters to transform all the factors into binary variables as described in the Materials and Methods section. Subsequently, we computed the duplicated genes that are targeted to the corresponding disease and non-disease gene location or not as "1" and "0" respectively. Logistic regression analysis was preferred as it was observed that the independent influence of each parameter (if any) on the backup capacity and at the same time can measure the contribution of all potential predictor variables to the regression model. We observed a positive association between backup capacity and sharing of interaction with its partners in the network ($P<0.027$), protein sublocalization ($P<0.001$), and sequence identity ($P<0.004$), while protein coexpression ($P<0.007$) regulates the backup capacity inversely. From the regression coefficient value (β) of each potential parameter it can be inferred that protein coexpression level ($\beta=3.763$) is the most influential predictor of the evolutionary rate followed by the sharing interaction partners ($\beta=2.001$), protein sublocalization ($\beta=1.425$) and sequence identity in the paralog pairs ($\beta=1.029$).

Table 2
Comparisons of sharing the interaction partners among four classes of duplicated pairs.

Gene classes	% of interacting partners shared by the duplicates
MD	51.11
PD	31.63
OTH	47.06
HK	59.22

N.B. Significant level (Mann–Whitney test): $P=1.32\times10^{-22}$ (MD vs. PD); $P=2.1\times10^{-14}$ (PD vs. OTH); $P=1\times10^{-4}$ (OTH vs. MD); $P=3.1\times10^{-5}$ (HK vs. MD); $P=1.1\times10^{-19}$ (HK vs. PD); $P=4.78\times10^{-21}$ (HK vs. OTH).

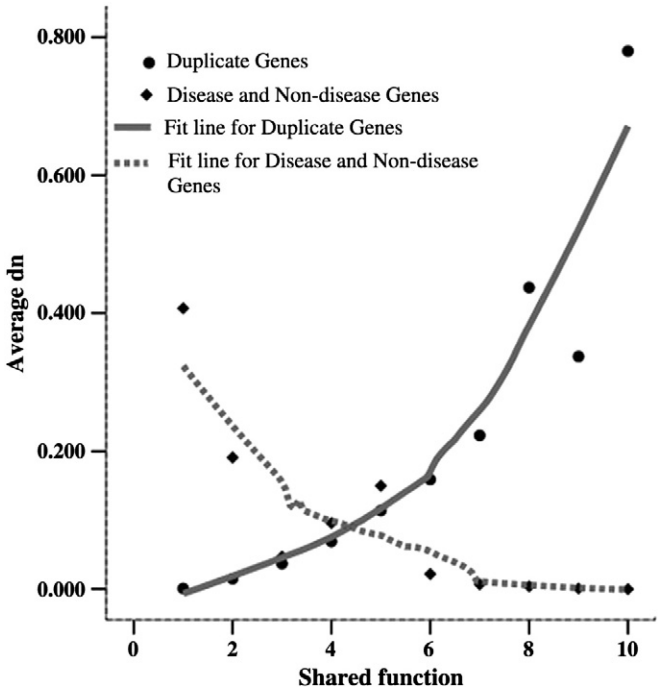

Fig. 2. Non-synonymous substitution rate (*dn*) with shared function of duplicates and their corresponding gene groups (disease and non-disease genes).

3. Discussions

Researchers have made dramatic inroads into the study of duplicate genes contributing a remarkable robustness against loss of one or more genes [14,22–24]. So far their focus has been concentrated in studying the non disease and disease gene duplicates where the disease gene sets are significantly biased towards Mendelian diseases. Our study was driven by the idea to explore the fundamental differences which exist between functional compensation for Mendelian (Monogenic) and Complex (Polygenic) diseases that may provide new information about the robustness into the realm of human genetic diseases. Indeed, the first important outcome of our study, based on the comparison of functional buffering for MD and PD duplicated pairs reveals that MD duplicates have more buffering capacity as compared to PD duplicates and that non-disease genes are themselves more buffered than both the classes of disease genes. This result is akin to the previous proposal [8] that proteins with similar paralogs should be less often involved in disease since the compromised function of such proteins when mutated could be compensated by their functional paralogs. Less backup capacity of polygenic disease duplicates may therefore increases the incidence of polygenic disease compared to monogenic disease. Although some report [25] found no simple relationship between sequence identity and functional similarity, yet we observed that buffering capacity of genes increased with the sequence identity between the corresponding gene paralog members, similar to the previous study [9].

Secondly, we have examined the genomic characteristics which have been suggested to influence the buffering capacity of genes. In yeast, the transcriptional reprogramming mechanism behind the genetic backup between paralogs elucidated that genes with low to moderate similarity of expression profiles to those of their paralogs are much more likely to be backed up [10]. Intending to verify this mechanism in human duplicates we also observed that the genes with higher buffering capacity (HK and MD genes) are rarely coexpressed with their paralogs whereas an opposite trend was found in the case of genes with lower buffering capacity (OTH and PD genes). To confirm this mechanism more evidently we also investigated the sharing of regulatory motifs between the duplicated pairs and noticed the same trend exists in human as observed in yeast. Sharing of interacting partners between paralog pairs also

remains as a strong determinant of rendering similar functionality in human. Moreover, in our study, we further enlighten a phenomenon that functionally redundant copies of disease and non-disease genes are redirected to the corresponding subcellular location of the above gene groups to maintain precise microenvironment for accomplishing similar functions.

Finally, we have emphasized the connection between backup capacity and non-synonymous substitution rate of disease, non-disease genes and their respective duplicates separately. Additionally, here we noticed that the backup capacity is negatively correlated with the non-synonymous substitution rate of disease and non-disease genes. Along with the rate of non-synonymous substitution (*dn*), non-synonymous to synonymous substitution rate (*dn*/*ds*) also varies with the backup capacity among the gene classes in a same fashion. The rate of coding sequence mutation of disease and non-disease genes decreases in the order of PD>OTH>MD>HK (*dn*/*ds* of PD = 0.199; MD = 0.143; OTH = 0.178; HK = 0.113; each value was significant with each other at least at 0.05 level in M–W test) which is consistent with our earlier study [26]. The most intuitive biological reason underpinning this observation may be the urge of important subset of genes to retain a backup copy for shielding themselves from deleterious mutation. However, the rate of non-synonymous to synonymous substitution of paralogs increases in the order of PD>OTH>MD>HK (*dn*/*ds* of PD = 0.101; MD = 0.134; OTH = 0.166; HK = 0.203; each value was significant with each other at least at 0.05 level in M–W test) which further signifies the relaxation of purifying selection on redundant duplicated genes [2,27]. Prior study [28] also recommended that if a gene's function is compensated by a redundant duplicate, mutation in that gene would have no effect on the phenotype. As a result, such mutations could not be selected against, and redundancy would be gradually lost.

To the best of our knowledge, this is the first extensive comparison of functional compensation among several classes of human genes. Although, a range of biological variables guiding the functional buffering of duplicates is reported previously, none of them quantifies the relative contribution of each factor. Assessing logistic regression analysis we conclude that the relative importance of individual factors modulating the variation in backup capacity among the aforementioned classes of genes is in the order of sharing expression level>sharing interaction partners>-protein subcellular relocalization>sequence identity in the paralogs pairs. Our results will obviously open a new paradigm in understanding the genetic robustness of human genetic diseases.

4. Materials and methods

4.1. Human disease and non-disease genes data

A list of 3959 monogenic and 2203 polygenic disease genes were obtained from the Human Gene Mutation Database (http://hgmd.org) [29] and Genetic Association Database (http://geneticassociationdb.nih.gov/) [30] respectively (Supplementary Table 1). Human protein coding genes were retrieved from Ensembl (http://www.ensembl.org/biomart/martview/) [31]. Disease genes (monogenic and polygenic) were excluded from the set of human protein coding genes and the rest of the genes were termed as non-disease genes. Finally, we obtained 9320 non-disease genes.

4.2. Identification of housekeeping genes

We used the tissue specificity index τ [32] to measure the tissue specificity of human genes. The τ of human gene i is defined by

$$\tau_H = \frac{\sum_{j=1}^{n_H}\left(1-\left[\frac{log_2 S_H(i,j)}{log_2 S_H(i,\ max)}\right]\right)}{n_H - 1},$$

where n_H is the number of human tissues examined and S_H (i, max) is the highest expression signal of gene i across the n_H tissues. The values of τ range from zero to one with higher values indicating higher variations in expression level across tissues or higher tissue specificities. If a gene is expressed in only one tissue, τ approaches to one. In contrast, if a gene is equally expressed in all tissues, $\tau = 0$. We assigned housekeeping genes by sorting our dataset according to the increase in τ values and taking out genes from the extreme 20% of population from the top end [26]. Finally, we obtained 660 housekeeping genes within 9320 non-disease genes (Supplementary Table 1).

4.3. Identification of paralogs and orthologs sets

Human paralogous genes and their sequence identity with the corresponding four classes of genes (monogenic disease, polygenic disease, housekeeping and other non-disease genes) were retrieved from Ensembl, (http://www.ensembl.org/biomart/martview/). The human–mouse orthologs with 1:1 relationship of the genes, concerned in our study and their corresponding paralogs were also downloaded from Ensembl, (http://www.ensembl.org/biomart/martview/). The corresponding Ensembl IDs from the database were used to extract the coding and orthologous sequences of the aforementioned gene sets from Ensembl. Generally remote paralogs are defined as pairs with ds >1 and close paralogs as pairs with ds <1. To avoid potential misclassification of borderline cases, we regarded remote pairs as those with ds >1.2 and close pairs as those with ds <0.8 [10].

4.4. Expression profile similarity between human genes and duplicates

The spatial expression information of human disease and non disease genes were obtained from the Gene Atlas V2 dataset (http://biogps.gnf.org/) [33]. Since our dataset was not normally distributed, we transformed the raw data into the rank in an ascending order. Because of the presence of several tied ranks in the dataset we have performed Pearson correlation coefficient (PCC), "*r*" to measure the expression profile similarity for each disease and non-disease gene [34]. Average "*r*" was calculated between the expression ranks of the considered genes with the expression ranks of each of its respective duplicated pairs [35].

$$r = \frac{\sum_{j=1}^{n}\left[S_H(i,j)S_p(i,j)\right] - \left[\sum_{j=1}^{n} S_H(i,j)\right]\left[\sum_{j=1}^{n} S_p(i,j)\right]/n}{\sqrt{\sum_{j=1}^{n}[S_H(i,j)]^2 - \left[\sum_{j=1}^{n} S_H(i,j)\right]^2/n}\sqrt{\sum_{j=1}^{n}\left[S_p(i,j)\right]^2 - \left[\sum_{j=1}^{n} S_p(i,j)\right]^2/n.}}$$

Here, n = the number of common tissues considered, H indicates human protein, and P indicates paralogs of the corresponding protein. S_H (i, j) and S_P (i, j) are the expression signal intensities of gene i in human tissue j for disease and non-disease proteins and gene i in human tissue j for their corresponding paralogs respectively.

4.5. Prediction of shared protein–protein interaction partners, promoter content and functional relationships

Human protein interaction data was retrieved from HPRD-version7 (http://hprd.org/) [36]. Promoters of the aforesaid genes were obtained from Transcriptional Regulatory Element Database (http://www.rulai.cshl.edu/cgi-bin/TRED/tred.cgi) [37] and the biological functions of the corresponding genes were downloaded from Ensembl. The percentage of shared interaction partners/promoters/function between paralogs was calculated by using Bayesian data integration method [38].

$$p_{\text{shared}(x,y)} = \frac{2 \times n_{(x,y)}}{n_x + n_y} \times 100\%,$$

where n_x and n_y represent the number of interactions/promoters/functional relationships for x and y proteins, respectively and $n_{(x,\ y)}$

represents the number of common interactions/promoters/functional relationships between duplicated pairs (x and y).

4.6. Prediction of sublocalization

Protein-coding human genes with cellular localization were extracted from Ensembl for 'cellular component' GO classification. Duplicated proteins partitioning same localization of disease and non-disease proteins were assigned as sublocalization.

4.7. Sequence analysis

Pair-wise synonymous (*ds*) and non-synonymous (*dn*) distances between the orthologous genes of human and mouse were calculated using the PAML package with default parameters [20]. The non parametric Mann–Whitney *U* test was used to evaluate the significance of all the pair-wise differences. Logistic regression analysis was done to analyze independent influence of each parameter in guiding the backup capacity of disease and non-disease genes. All the statistical tests were performed using the SPSS (13.0) package.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.01.004.

Acknowledgments

Authors are thankful to the Department of Biotechnology, Govt. of India for financial help (sanction number 102/IFD/SAN/PR-1860/2008-09).

References

[1] J.S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old ideas, Annu. Rev. Genet. 38 (2004) 615–643.
[2] M.A. Nowak, M.C. Boerlijst, J. Cooke, J.M. Smith, Evolution of genetic redundancy, Nature 388 (1997) 167–171.
[3] Z.L. Gu, L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, W.H. Li, Role of duplicate genes in genetic robustness against null mutations, Nature 421 (2003) 63–66.
[4] A. Wagner, Robustness against mutations in genetic networks of yeast, Nat. Genet. 24 (2000) 355–361.
[5] Y.K. Wang, P.N.J. Schnegelsberg, J. Dausman, R. Jaenisch, Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin, Nature 379 (1996) 823–825.
[6] W.J. Bailey, J. Kim, G.P. Wagner, F.H. Ruddle, Phylogenetic reconstruction of vertebrate Hox cluster duplications, Mol. Biol. Evol. 14 (1997) 843–853.
[7] K. Nasmyth, Control of the yeast cell cycle by the Cdc28 protein kinase, Curr. Opin. Cell Biol. 5 (1993) 166–179.
[8] N. Lopez-Bigas, C.A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease, Nucleic Acids Res. 32 (2004) 3108–3114.
[9] K. Hannay, E.M. Marcotte, C. Vogel, Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation, BMC Genomics 9 (2008).
[10] R. Kafri, A. Bar-Even, Y. Pilpel, Transcription control reprogramming in genetic backup circuits, Nat. Genet. 37 (2005) 295–299.
[11] X.W. Pan, P. Ye, D.S. Yuan, X.L. Wang, J.S. Bader, J.D. Boeke, A DNA integrity network in the yeast *Saccharomyces cerevisiae*, Cell 124 (2006) 1069–1081.
[12] D.C. Krakauer, M.A. Nowak, Evolutionary preservation of redundant duplicated genes, Semin. Cell Dev. Biol. 10 (1999) 555–559.
[13] M. Lynch, J.S. Conery, The evolutionary fate and consequences of duplicate genes, Science 290 (2000) 1151–1155.
[14] T.L. Hsiao, D. Vitkup, Role of duplicate genes in robustness against deleterious human mutations, PLoS Genet. 4 (2008).
[15] J.J. Cai, E. Borenstein, R. Chen, D.A. Petrov, Similarly strong purifying selection acts on human disease genes of all evolutionary ages, Genome Biol. Evol. (2009) 131–144.
[16] S.A. Byun-McKay, R. Geeta, Protein subcellular relocalization: a new perspective on the origin of novel genes, Trends Ecol. Evol. 22 (2007) 338–344.
[17] A.C. Marques, N. Vinckenbosh, D. Brawand, H. Kaessmann, Functional diversification of duplicate genes through subcellular adaptation of encoded proteins, Genome Biol. 9 (2008).
[18] B. Jacq, Protein function from the perspective of molecular interactions and genetic networks, Brief. Bioinform. 2 (2001) 38–50.
[19] C. Brun, A. Guenoche, B. Jacq, Approach of the functional evolution of duplicated genes in *Saccharomyces cerevisiae* using a new classification method based on protein–protein interaction data, J. Struct. Funct. Genomics 3 (2003) 213–224.
[20] A.H.Y. Tong, M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C.W.V. Hogue, H. Bussey, B. Andrews, M. Tyers, C. Boone, Systematic genetic analysis with ordered arrays of yeast deletion mutants, Science 294 (2001) 2364–2368.
[21] Z.H. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, Mol. Biol. Evol. 17 (2000) 32–43.
[22] J.H. Thomas, Thinking about genetic redundancy, Trends Genet. 9 (1993) 395–399.
[23] M. Kupiec, R. Sharan, E. Ruppin, Genetic interactions in yeast: is robustness going bust? Mol. Syst. Biol. 3 (2007).
[24] J. Ihmels, S.R. Collins, M. Schuldiner, N.J. Krogan, J.S. Weissman, Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss, Mol. Syst. Biol. 3 (2007).
[25] A. Baudot, B. Jacq, C. Brun, A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein–protein interaction network, Genome Biol. 5 (2004).
[26] S. Podder, T.C. Ghosh, Exploring the differences in evolutionary rates between monogenic and polygenic disease genes in human, Mol. Biol. Evol. 27 (2010) 934–941.
[27] J. Brookfield, Evolutionary genetics: can genes be truly redundant? Curr. Biol. 2 (1992) 553–554.
[28] M. Lynch, M. O'Hely, B. Walsh, A. Force, The probability of preservation of a newly arisen gene duplicate, Genetics 159 (2001) 1789–1804.
[29] P.D. Stenson, M. Mort, E.V. Ball, K. Howells, A.D. Phillips, N.S. Thomas, D.N. Cooper, The human gene mutation database: 2008 update, Genome Med. 1 (2009) 13.
[30] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, The genetic association database, Nat. Genet. 36 (2004) 431–432.
[31] P. Flicek, B.L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y.A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S.P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernandez-Suarez, J. Herrero, T.J.P. Hubbard, A. Parker, G. Proctor, J. Smith, S.M.J. Searle, Ensembl's 10th year, Nucleic Acids Res. 38 (2010) D557–D562.
[32] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics 21 (2005) 650–659.
[33] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl Acad. Sci. USA 101 (2004) 6062–6067.
[34] L. Louis, G. Lise, Indirect two-sided relative ranking: a robust similarity measure for gene expression data, BMC Bioinform. 11 (2010).
[35] B.Y. Liao, J.Z. Zhang, Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution, Mol. Biol. Evol. 23 (2006) 1119–1128.
[36] T.S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D.S. Somanathan, A. Sebastian, S. Rani, S. Ray, C.J. Harrys Kishore, S. Kanth, M. Ahmed, M.K. Kashyap, R. Mohmood, Y.L. Ramachandra, V. Krishna, B.A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, A. Pandey, Human protein reference database—2009 update, Nucleic Acids Res. 37 (2009) D767–D772.
[37] C. Jiang, Z. Xuan, F. Zhao, M.Q. Zhang, TRED: a transcriptional regulatory element database, new entries and other development, Nucleic Acids Res. 35 (2007) D137–D140.
[38] Y.F. Guan, M.J. Dunham, O.G. Troyanskaya, Functional analysis of gene duplications in *Saccharomyces cerevisiae*, Genetics 175 (2007) 933–943.