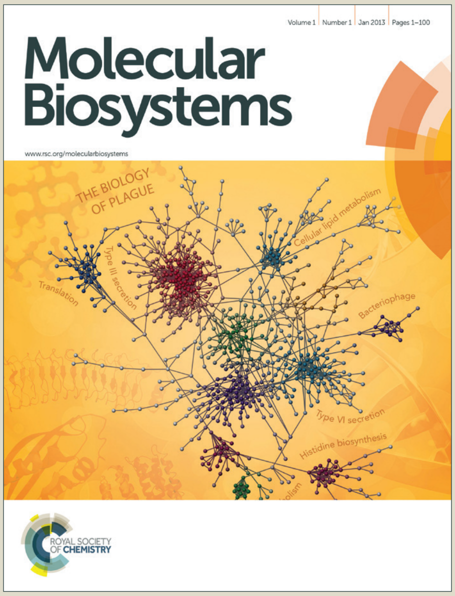
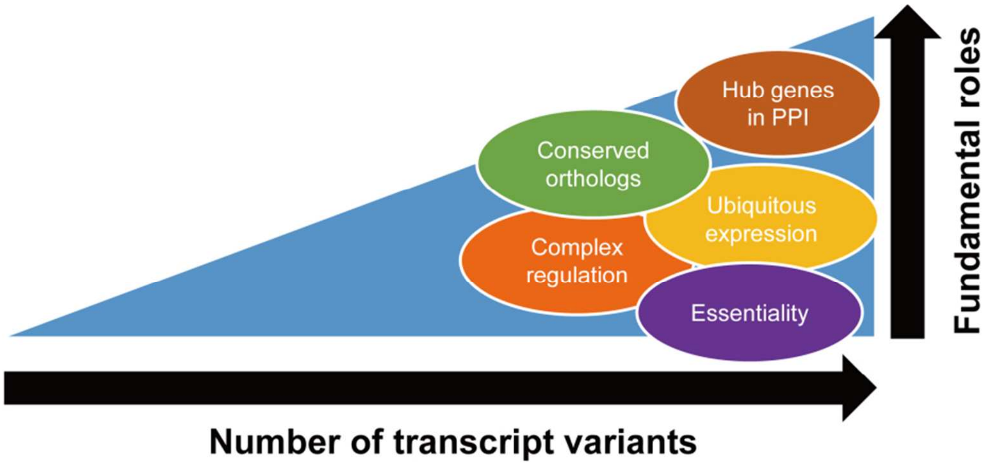




Human genes with greater transcript variants are more likely to play functionally important roles such as cellular maintenance and survival.
160x74mm (150 x 150 DPI)

Human genes with a greater number of transcript variants tend to show biological features of housekeeping and essential genes

Jae Yong Ryu,[a] Hyun Uk Kim[abd] and Sang Yup Lee*[abcd]

[a] Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Republic of Korea.

[b] BioInformatics Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[c] BioProcess Engineering Research Center, KAIST, Daejeon 305-701, Republic of Korea.

[d] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark.

† Electronic supplementary information (ESI) available: Figures S1-S4 and Tables S1-S8.

Corresponding author: Sang Yup Lee. Mailing address: Department of Chemical and Biomolecular Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305–701, Republic of Korea. Phone: 82-42-350-3930. Fax: 82-42-350-3910.

* e-mail: leesy@kaist.ac.kr

Abstract

Alternative splicing is a process observed in gene expression that results in a multi-exon gene to produce multiple mRNA variants which might have different functions and activities. Although physiologically important, many aspects of genes with different number of transcript variants (or splice variants) still remain to be characterized. In this study, we provide bioinformatic evidences that genes with a greater number of transcript variants are more likely to play functionally important roles in cells, compared with those having fewer transcript variants. Among 21,983 human genes, 3,728 genes were found to have a single transcript, and the remaining genes had 2 to 77 transcript variants. The genes with more transcript variants exhibited greater frequencies of acting as housekeeping and essential genes rather than tissue-selective and non-essential genes. They were found to be more conserved among 64 vertebrate species as orthologs, subjected to regulations by transcription factors and microRNAs, and showed hub node-like properties in the human protein-protein interaction network. These findings were also confirmed by metabolic simulations of 60 cancer metabolic models. All these results indicate that genes with a greater number of transcript variants play biologically more fundamental roles.

Introduction

During the expression of a multi-exon gene, alternative splicing results in generation of multiple transcripts.[1, 2] In human, 92-97% of the multi-exon genes undergo alternative splicing.[3] These alternatively spliced variants from a gene can have important implications in mammalian physiology and have been a source of functional diversity of many human genes by providing multiple protein products with alternative functional domains.[4] In particular, correlations between the number of splice variants of a gene and its functional role have been important topics of human genomic studies. In recent years, advent of next-generation sequencing technology such as RNA-Seq with high resolution has facilitated elucidation of functional features of splice variants of genes.[4, 5] RNA-Seq data revealed that alternative splicing events are differentially regulated in human tissues, leading to tissue-specifically coordinated splicing events.[6] Such tissue-specific alternative splicing events allow the same gene to have different combinations of exons (i.e., splice variants) across the tissues, and therefore tissue-specifically generated splice variants can have differentiated protein structures and functions.[7] Importantly, the protein isoforms from the same gene can have different degrees of disorder (i.e., lack of a well-defined three-dimensional structure) depending on the inclusion of tissue-specific exons. Such protein isoforms can rewire the overall protein-protein interaction (PPI) network by interacting with different proteins. Recent studies on generic PPI[8] and tissue-specific PPI networks of human[9] revealed that proteins encoded by genes with a greater number of splice variants tend to have more neighbor nodes and higher centralities in contrast to those encoded by genes with fewer splice variants. The number of neighbor nodes and node centralities are indicators of biologically important functions, and their greater values tend to get greater for the functionally important nodes (e.g., proteins).[10] Interestingly, tissue-specific exons, which are often observed in proteins with large values of neighbor nodes and node centralities, also appeared to be more associated with post-translational modifications and evolutionary conservations than constitutive exons.[7] More functional features of splice variants remain to be elucidated through combined experimental (i.e., RNA-Seq analysis) and theoretical studies.

In characterizing functional roles of genes, their expression patterns and essentiality are important criteria to consider. Genes can typically be categorized into housekeeping (HK) and tissue-selective (TS) genes depending on their expression patterns[11, 12]; the former being defined as genes expressed across all tissues to maintain cellular functions and the latter being expressed in only certain tissues. Essential (ES) genes are those that are critical to cell growth and survival, whereas non-essential (NE) genes are not. There was a recent study on identifying human essential genes based on the essential orthologs of mouse.[13] Evaluation of the expression patterns in different tissues and essentiality of genes based on the different number of splice variants can be useful in determining their biological importance.

In this study, we provide systemic evidences through bioinformatic analyses that genes with a greater number of transcript variants (or splice variants) have a greater chance of playing biologically important roles than those with fewer transcript variants. First, genes were grouped based on the number of their transcript variants in order to identify correlations between the number of their transcript variants and their expression patterns (as HK and TS genes) / essentiality (as ES and NE genes). For the comparative analyses of genes with the different number of transcript variants, a series of analyses were carried out to elucidate the degree of their functional conservations *via* ortholog analysis across genomes of vertebrate species, regulations by transcription factors and microRNAs, and central hub-like network properties in human PPI network. Finally, we used 60 cancer metabolic models for essentiality simulation of human metabolic genes upon their knockouts in order to further validate our findings on correlations between the number of transcript variants and gene functions. The present system-wide study provides additional evidences on the biological importance of transcript variants of human genes.

Results and Discussion

Human genes with a greater number of transcript variants play biologically more important roles

In order to examine the distribution of human genes showing different number of transcript variants, the number of transcript variants for 21,983 human genes was examined (Table S1, ESI†). These genes were downloaded from the Ensembl BioMart (release 78), and only the protein-coding genes (covering both multi- and single-exon) and their transcripts including transcript variants were considered. Meanwhile, we considered all types of transcript variants for a gene, including both that have protein IDs and do not lead to protein products. The reason is that all types of transcript variants have the chance to influence cellular physiology, for instance in the form of microRNA sponge (see Experimental for details). On average, there were 6.95 transcript variants per human gene. Among 21,983 human genes, 3,728 genes were found to have a single transcript, and the remaining genes had 2 to 77 transcript variants (Fig. 1). Overall, 83% of the human genes had 2-28 transcript variants, and the rest 0.01% (219 genes) had 29 or more transcript variants. In order to investigate correlations between the number of transcript variants and functions of human genes, human genes were categorized into a total of 77 groups according to the number of transcript variants. Among them, 60 groups had at least one or more genes, and none of the human genes had the following numbers of transcript variants: 46, 51, 54, 59, 62, 63, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, and 76. Thus, there were 60 groups that were analyzed as below, excluding those 17 groups having no genes belonging to.

First, these grouped genes having different number of transcript variants were analyzed with respect to the HK, TS, ES and NE gene categories by using gene expression pattern data covering 3,804 HK and 2,293 TS genes[12, 14] and gene essentiality data for 2,472 ES and 3,811 NE genes[13] (Table S2, ESI†). Here, HK and ES genes can be considered to be biologically important and fundamental genes, compared with their counterparts (i.e., TS and NE genes). The numbers of transcript variants for HK, TS, ES and NE genes were determined and compared (Fig. 2A). The results show that HK and ES genes tend to have a greater number of transcript variants compared with TS and NE genes, respectively. Also, the average numbers (9.12 and 9.29) of transcript variants for HK and ES genes are greater than the average number (6.95) of transcript variants for all human genes. This observation suggests that genes having important roles (HK and ES) tend to have a greater number of transcript variants compared with their counterparts (6.95 for TS and 7.64 for NE). It should be noted that the lines inside the boxplots in Figure 2A are median values, not averages. In addition, our analysis on the correlation between the number of exons in all the genes considered in this study and the number of their transcript variants revealed that they were not significantly correlated (Fig. S1, ESI†; Pearson correlation coefficient = 0.39 in Fig. S1). This observation manifests that the greater number of transcript variants for the HK and ES genes was caused by various forms of alternative splicing events, not simply by a greater number of exons in their genes. Splice variants can arise from several different mechanisms, including exon skipping, mutual exclusion of exons, alternative 5' donor site, alternative 3' acceptor site, and intron retention.[4]

The finding that the HK and ES genes overall generated a greater number of transcript variants was further supported by increasing percentages of HK, TS, ES and NE genes in each group as the number of transcript variants increased (Fig. 2B and C). The percentages of TS and NE genes showed somewhat different patterns; they did not increase as a function of the number of transcript variants. Statistical significances for the presence of HK, TS, ES and NE genes in each group with different numbers of transcript variants were calculated with Fisher's exact test, and are available in Table S3, ESI†.

In order to confirm that genes having more transcript variants are playing more important roles, we analyzed expression levels of the genes belonging to 60 groups using a recent proteomic study on 32 different human tissues by Uhlen et al.[15]; the percentages of genes in 60 groups that are expressed and appeared in proteome data were calculated (Fig. 3). Also, the percentages of HK, TS, ES, and NE genes in each tissue were calculated (Fig. 3). It was found that genes with a greater number of transcript variants were more ubiquitously expressed in all 32 different human tissues (red region in the heat map in Fig. 3), compared with those having fewer transcript variants (blue and green regions in the heat map in Fig. 3). As expected, the HK genes were ubiquitously expressed in all the 32 tissues, while 33-76% of the TS and NE genes were expressed in the 32 tissues (Fig. 3). Also, greater than 75% of the ES genes were expressed in all the tissues except for bone marrow and skeletal muscle (Fig. 3). In order to clearly show that the tissue-specific expression patterns of the examined genes were not affected by the presence of the HK, TS, ES and NE genes in each group, the gene expression patterns were re-examined by excluding all the HK, TS, ES and NE genes from each group, and the new results appeared to be consistent (Fig. S2, ESI†). These results confirm that expression of those genes having more transcript variants is more demanded in human cell compared with those having fewer transcript variants, which suggests that these genes with more transcript variants are likely play more important functional roles in the cell.

Analysis of orthologs in gene groups having different number of transcript variants

Next, we examined the number of conserved orthologs across 64 vertebrate species in each gene group (i.e., 60 groups having different number of transcript variants) to examine whether the functional conservation is correlated with the number of transcript variants (or splice variants). The number of orthologs would indicate the level of functional conservation across the examined species.[16] First, the orthologs in all 60 groups having different number of transcript variants were searched against genomes of 64 vertebrate species and counted (Fig. 4). Also, the orthologs of the human HK, TS, ES, and NE genes were searched for these genomes. In this analysis, all the protein-coding genes known to be present in the 64 vertebrates were obtained from the OrthoDB.[17] Among genes in the vertebrates, human orthologs were selected in order to examine the presence of conserved orthologs.

As a result, genes in groups having a large number of transcript variants appeared to be more conserved in 64 vertebrates compared with those having fewer transcript variants (Fig. 4). In particular, human orthologs were highly conserved in the orders such as *Carnivora*, *Cetartiodactyla*, *Glires*, and *Primates*, whereas human orthologs including HK and ES genes were not well conserved in the other orders such as *Ctenosquamata* and *Saurischia* (Fig. 4). High-level conservation of human orthologs in *Carnivora*, *Cetartiodactyla*, *Glires*, and *Primates* could be attributed to their common ancestor (the magnorder *Boreoeutheria*) according to the NCBI Taxonomy database[18]. Furthermore, a vertebrate species sharing orthologs with human to the greatest extent was olive baboon (*Papio anubis*), which appeared to have all genes from 32 groups and 92.3-99.6% genes from the remaining groups conserved in human. In contrast, sea lamprey (*Petromyzon marinus*) had the lowest number of human orthologs, having all genes from 11 groups and 0-83.3%

from the remaining groups conserved in human. Sea lamprey (*Petromyzon marinus*) was found to be phylogenetically located in the farthest distance from the rest of the vertebrate species.[19] Groups having 42, 52, 55, 56 and 64 transcript variants, which are conserved among all the 64 vertebrates, had 7 genes in total (i.e., *AKT2*, *EEF1D*, *MOK*, *MYB*, *NDRG4*, *RUNX1T1* and *SORBS2*). These genes were associated with fundamentally important functions such as protein kinases (*AKT2* and *MOK*), eukaryotic translation elongation factor (*EEF1D*), transcription factors (*MYB* and *RUNX1T1*), cell cycle progression (*NDRG4*), and adaptor protein for signaling complex (*SORBS2*). The same consistent results were obtained from the ortholog conservation analysis conducted with the same gene groups, but by excluding all the HK, TS, ES and NE genes, in order to confirm that the conservation patterns were affected by the number of transcript variants (or splice variants), not by the HK, TS, ES and NE genes present in each group (Fig. S3, ESI†). Thus, analysis of the conserved orthologs of human genes across the vertebrate species suggested another clue that genes having many transcript variants are playing functionally more important roles (e.g., conserved functions across vertebrates) due to their greater level of conservation across the examined species. On the contrary, genes with fewer transcript variants might play rather species-specific roles as shown by the lower level of conservation among the examined species for the group with a single transcript.

Regulation of genes by transcription factors and microRNAs for the genes having different number of transcript variants

We next investigated to what extent genes with different number of transcript variants (or splice variants) are subject to regulations by transcription factors and microRNAs, two important intracellular regulators. Transcription factors activate or repress their target genes by binding to their promoter regions, whereas microRNAs repress target genes by binding to their seed sites (or microRNA-binding sites) in 3' UTR. Both transcription factors and microRNAs can regulate multiple genes.[20]

The average numbers of transcription factors and microRNAs that regulate genes in the 60 groups having different number of transcript variants, and the HK, TS, ES, and NE genes were calculated. Data on target genes regulated by transcription factors and microRNAs were obtained from HTRIdb[21] and miRTarBase,[22] respectively; both databases provide information on experimentally validated target genes regulated by transcription factors and microRNAs. Information on all the microRNAs and transcription factors available in the abovementioned databases was used for this analysis in order to grasp overall relationship between the average numbers of regulators and their target genes. A full list of microRNAs, transcription factors and their target genes are available Table S4 (ESI†).

For transcription factors, genes with a single transcript appeared to be regulated by 1.49 transcription factors on average. The number of transcript variants and transcription factors regulating the corresponding genes showed a positive correlation up to the group with 31 transcript variants; the genes in the group with 31 transcript variants were found to be regulated by 3.21 transcription factors per gene on average. Correlations could not be inferred for the groups having greater than 31 transcript variants because of the lack of sufficient number of genes in these groups; less than 0.01% of human genes belong to these groups. Nonetheless, the overall pattern observed was that genes with many transcript variants tend to be subject to regulations by more transcription factors (Fig. 5A).

In a similar manner, gene regulations by microRNAs were examined. Genes with a single transcript appeared to be subject to regulations by 0.92 microRNAs on average. Positive correlations between the number of transcript variants and the number of microRNAs regulating the corresponding genes were observed for the gene groups having up to 30 transcript variants; the group with 30 transcript variants showed presence of 2.76 regulatory microRNAs per gene (Fig. 5B). Similarly to the transcription factor case, groups having greater than 30 transcript variants could not be considered for inferring correlations due to too few genes in these groups. Interestingly, three genes (*AKT2*, *MOK* and *MYB*) in a group having 42 transcript variants appeared to be regulated by 7.33 transcription factors and 7.67 microRNAs on average; this group showed the greatest number of regulators among all the gene groups having different number of transcript variants. In particular, *MYB,* known to be an essential gene crucial in hematopoiesis[23]*,* was found to be regulated by 13 transcription factors and 20 microRNAs.

Because transcription factors and microRNAs generate 9.12 and 9.29 transcript variants on average, respectively, the number of transcription factors and microRNAs regulating these gene sets were compared with genes in the group having 9 transcript variants. Interestingly, the HK (regulated by 3.02 transcription factors and 3.23 microRNAs) and ES (regulated by 3.54 transcription factors and 3.56 microRNAs) genes appeared to be more regulated than the group having 9 transcript variants regulated by 2.91 transcription factors and 2.17 microRNAs. This observation suggests that biologically more important genes such as HK and ES genes tend to be subject to more complex regulations.

Taken together, these results confirm that functionally important genes such as those with a greater number of transcript variants, and the HK and ES genes are subject to more complex regulations by more transcription factors and microRNAs. Genes with multiple transcript variants are likely to be involved in complex regulations through different promoter binding and polyadenylations by creating alternative 5' and/or 3' exons of the variant structures, and consequently help cells better adapt to environmental and/or genetic perturbations.[24]

Analysis of genes with different number of transcript variants from a network perspective

The above grouped genes (i.e., genes in 60 groups, and the HK, TS, ES, and NE genes) were then analyzed at large-scale protein level by utilizing a human PPI network from the PINA 2.0 database.[25] This network consists of a total of 17,109 nodes and 166,776 edges, each representing proteins and their interactions, respectively. In the PPI network, the degree is defined as the number of interacting proteins. We examined average degrees of proteins encoded by genes in the 60 groups having different number of transcript variants (or splice variants), and the HK, TS, ES, and NE genes. It was hypothesized that genes having more splice variants are likely to be central hubs that have a large number of connections with other proteins, and are also related to cellular essentiality.[26] Consistent with the comparative analyses presented above, central hubs were more frequently mapped to proteins encoded by genes with a greater number of transcript variants, and the HK and ES genes (Fig. 6A). For proteins encoded by the HK and ES genes, the average degrees of interactions were 36.10 and 49.07, respectively; these values are almost twice the average degree (20.29) of interactions for proteins encoded by genes having 9 transcript variants (i.e., similar to the average number of transcript variants for the HK and ES genes). Interestingly, genes with 14 transcript variants showed proteins with average degree of 53.68, which is a value substantially greater than nearby gene groups. This outlier (group with 14 transcript variants) is due to the presence of *UBC* gene encoding ubiquitin which interacts with 9,136 proteins in the PPI network for protein degradation. Network hub nodes are in general known to be essential because of a large number of their connections with other nodes and hence greater damages to the network stability upon their removal.[26] The observation that proteins encoded by the genes having a greater number of transcript variants are more likely to have central hub-like properties is not strange because multiple proteins are generated from such genes, and therefore allow more interactions with other proteins.[10] Consistently to these results, a previous study revealed that the number of degrees of protein nodes in human generic and tissue-specific PPI networks was positively correlated with the number of transcript variants for their respective genes.[8, 9]

Finally, the correlation between the average degree of the PPI network and the number of transcript variants was examined for the proteins with similar levels of disorder. Intrinsically disordered proteins are known to interact with more diverse proteins than ordered proteins because of their structural flexibility, and they also have regions enriched for alternative splicing.[27] Therefore, it was important to confirm that the observed average degrees were purely caused by the number of transcript variants (or splice variants), and not the level of protein disorder. For this analysis, disorder levels of all the proteins in the PPI network were calculated using MobiDB 2.0.[28] As a result, the proteins with 0% and 50-70% disorders all consistently showed that their average degrees and the number of their transcript variants were correlated in a positive manner (Fig. 6B and C). The results were also similar for the proteins with > 50% and >70% disorders (Fig. S4, ESI†). Taken together, we found that genes having a greater number of transcript variants indeed followed patterns of the HK and ES genes. This should be useful additional information for better characterization of the human PPI network.

Characterizing essentiality of metabolic genes having different number of transcript variants using *in silico* genome-scale metabolic models

Comprehensive human genome-scale metabolic models have proven useful in human metabolic studies including the understanding of physiological phenomena[29, 30], prediction of disease-specific biomarkers,[31] and drug targeting.[32, 33] To this end, we used recently reported 60 different NCI-60 cancer cell line-specific metabolic models[34] to further validate that metabolic prediction outcomes are consistent with the observed functional characteristics of genes having different number of transcript variants (or splice variants). Here, cancer cell metabolic models, instead of generic metabolic or normal cell type-specific models, were used in simulations. This is because the objective of cancer cell can be assumed to be biomass maximization, while that of normal cell cannot be.[35]

In order to get the number of metabolic genes in each group having different number of transcript variants, metabolic genes in the human generic model Recon 2 were searched against all the genes in 60 groups. Recon 2 is the latest version of the large-scale human metabolic model that has information on 1,789 metabolic genes, which appear to be present in human genome and correspond to 7,440 reactions and 2,626 metabolites.[36] Metabolic genes were found to have 8.78 transcript variants on average, which is a value greater than the average of all human genes (6.95 transcript variants). The percentage of metabolic genes to the total genes in each group increased as the number of transcript variants increases (Fig. 7A); this observation is reasonable because metabolism plays an important role in cellular growth through energy and biomass generation.

In order to confirm the aforementioned finding that genes with a greater number of transcript variants more frequently followed behaviors shown by the HK and ES genes, we next performed essentiality simulation for the metabolic genes using 60 cancer metabolic models; resulting growth rates of the cancer metabolic models were predicted using constraint-based flux analysis with each gene knocked out individually (see Experimental). In each model, the deleted genes were considered to be essential if the resulting predicted growth rate is lower than 5% of the maximum growth rate. As a result, none of the genes with a single transcript were predicted to be essential, while genes having 2 to 7 transcript variants were increasingly predicted to essential; however, the percentages of essential genes in the groups having 2 to 7 transcript variants were low (Fig. 7B and Table S5, ESI†). Taken together, the results from the simulation of 60 cancer cell metabolic models support our hypothesis that genes having a greater number of transcript variants are more likely associated with cellular essentiality.

Conclusions

In this study, we examined functional characteristics of genes according to the number of transcript variants (or splice variants) at genome-scale. It was found that genes having a greater number of transcript variants showed characteristics more similar to those of the HK and ES genes, suggesting that these genes play biologically more important roles. Biological importance of these genes with a greater number of transcript variants was further supported by greater conservation of orthologs across vertebrates, more complex regulations by greater number of transcription factors and microRNAs, and more hub-like properties in the human PPI network compared with genes having fewer transcript variants. Finally, we employed 60 cancer genome-scale metabolic models to further examine correlation between the essentiality of genes and the number of transcript variants. Genes having a greater number of transcript variants caused more deleterious effects on cell essentiality upon their knockouts. In summary, several different genome-wide analyses on the genes having different number of transcript variants consistently suggested that those genes having greater number of transcript variants indeed play biologically more important roles, and thus these genes and various transcript variants produced from these genes should receive much more attention in biological studies.

Experimental

Sources of data on human genes used for various comparative analyses

Data on a total of 21,983 protein-coding genes (covering both multi- and single-exon) and their transcripts including splice variants were downloaded from the Ensembl BioMart (release 78).[37] Only the protein-coding genes, not pseudogenes, were considered, but in case of their transcript variants (or splice variants), those given any category of the Transcript Support Level (TSL) were considered because they all have the chance to influence cellular physiology. When only transcripts having the TSL category of tsl1 were counted for the HK, TS, ES and NE genes, it was not possible to observe the differences in the number of their transcript variants; this contrasts with the data presented in Figure 2A. In fact, the percentage of the transcript variants with the tsl1 category was only 27.1% among all the transcript variants theoretically and/or experimentally identified in human genes. Therefore, it was considered reasonable to treat all the transcript variants to more precisely grasp hidden features of transcript variants of the human genes.

Information on 3,804 HK and 2,293 TS genes were obtained from Eisenberg et al. (2013)[14] and Chang et al. (2011),[12] respectively. Information on 2,472 ES and 3,811 NE genes was collected from Georgi et al. (2013).[13] As to the analysis of metabolic genes, those defined in the human generic metabolic model Recon 2 were considered.[36] Finally, the orthologs data in 64 vertebrates were obtained from OrthoDB, and among them, only human orthologs were selected.[17] Tissue-specific proteome expression data were obtained from Uhlén et al (2015),[15] which were used to analyze tissue-specific expressions of protein-associated genes in 60 groups, and the HK, TS, ES and NE genes. The 32 human tissues were considered in this study.

Analysis of microRNA and transcription factor regulating human genes

Information on target genes regulated by transcription factors and microRNAs was obtained from two experimentally validated databases, HTRIdb[21] and miRTarBase,[22] respectively. Ensembl gene IDs for genes obtained from the Ensembl BioMart were next converted to Entrez gene IDs used in the HTRIdb and miRTarBase using gene2ensembl available at the NCBI FTP Site (Feb. 2015) in order to map the genes onto those regulated by microRNAs and transcription factors. As a result of the gene ID conversion, 17,362 genes were considered for this analysis, and the numbers of microRNAs and transcription factors regulating them were counted. For the statistical significances, one-sided Wilcoxon rank sum tests were performed for each pair of the groups with different number of transcript variants presented in Figure 5 (Table S6 and S7, ESI†).

Analysis of protein-protein interaction network for the genes with a single transcript and multiple transcript variants

Protein interactome data were downloaded from the Protein Interaction Network Analysis (PINA) 2.0 database.[25] This network contains a total of 17,109 nodes and 166,776 edges, each representing proteins and their interactions, respectively. NetworkX version 1.8 (http://networkx.lanl.gov/) python package was used to calculate degree distributions of protein nodes. The same statistical procedure used for the target genes regulated by microRNAs and transcription factors was used for this analysis to obtain statistical significances (Fig. 6). For the analysis of correlations between the degree of protein disorder and the number of transcript variants, the degree of protein disorder was calculated using a python script available at MobiDB 2.0.[28]

In silico genome-scale metabolic simulation

Metabolic simulations are typically conducted by using an optimization technique for a metabolic model that has stoichiometric coefficients of all the metabolites in metabolic reactions that appear to be present in an organism.[38] The genome-scale metabolic models are usually underdetermined systems for which the optimization is needed, and the objective function is typically set to maximization of biomass formation for human cancer cells and microorganisms.[35] In contrast to kinetic modeling, this genome-scale metabolic modeling does not require kinetic parameters, but optionally can take omics data which can be set as optimization constraints for a human system.[39] In this study, recently reported 60 NCI-60 cancer cell line-specific metabolic models were used for the metabolic simulations.[34] Gene essentiality simulation was conducted using minimization of metabolic adjustment (MOMA).[40] Essential genes were defined as genes causing the cellular growth rate lower than 5 % of its maximum value upon their knockouts. All the metabolic simulations were conducted under the COBRApy environment[41] with Gurobi Optimizer (Gurobi Optimization, Inc., Houston, TX).

Acknowledgements

We thank Jae Ho Shin and Yun Sung Cho for critical comments on the manuscript. This work was supported by the Bio-Synergy Research Project (2012M3A9C4048759) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

Competing financial interests

The authors declare no competing financial interests.

References

1. T. Maniatis, *Science*, 1991, **251**, 33-34.
2. Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe and B. J. Frey, *Nature*, 2010, **465**, 53-59.
3. Q. Pan, O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe, *Nat. Genet.*, 2008, **40**, 1413-1415.
4. H. D. Li, R. Menon, G. S. Omenn and Y. Guan, *Trends Genet.*, 2014, **30**, 340-347.
5. Z. Wang, M. Gerstein and M. Snyder, *Nat. Rev. Genet.*, 2009, **10**, 57-63.
6. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth and C. B. Burge, *Nature*, 2008, **456**, 470-476.
7. M. Buljan, G. Chalancon, S. Eustermann, G. P. Wagner, M. Fuxreiter, A. Bateman and M. M. Babu, *Mol Cell*, 2012, **46**, 871-883.
8. A. Sinha and H. A. Nagarajaram, *J. Proteome. Res.*, 2013, **12**, 1980-1988.
9. A. Sinha and H. A. Nagarajaram, *Proteomics*, 2014, **14**, 2242-2248.
10. C. J. Tsai, B. Ma and R. Nussinov, *Trends Biochem. Sci.*, 2009, **34**, 594-600.
11. A. J. Butte, V. J. Dzau and S. B. Glueck, *Physiol. Genomics*, 2001, **7**, 95-96.
12. C. W. Chang, W. C. Cheng, C. R. Chen, W. Y. Shu, M. L. Tsai, C. L. Huang and I. C. Hsu, *PLoS One*, 2011, **6**, e22859.
13. B. Georgi, B. F. Voight and M. Bucan, *PLoS Genet.*, 2013, **9**, e1003484.
14. E. Eisenberg and E. Y. Levanon, *Trends Genet.*, 2013, **29**, 569-574.
15. M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Ponten, *Science*, 2015, **347**, 1260419.
16. R. L. Tatusov, E. V. Koonin and D. J. Lipman, *Science*, 1997, **278**, 631-637.
17. E. V. Kriventseva, N. Rahman, O. Espinosa and E. M. Zdobnov, *Nucleic Acids Res.*, 2008, **36**, D271-275.
18. S. Federhen, *Nucleic Acids Res.*, 2012, **40**, D136-143.
19. J. J. Smith, S. Kuraku, C. Holt, T. Sauka-Spengler, N. Jiang, M. S. Campbell, M. D. Yandell, T. Manousaki, A. Meyer, O. E. Bloom, J. R. Morgan, J. D. Buxbaum, R. Sachidanandam, C. Sims, A. S. Garruss, M. Cook, R. Krumlauf, L. M. Wiedemann, S. A. Sower, W. A. Decatur, J. A. Hall, C. T. Amemiya, N. R. Saha, K. M. Buckley, J. P. Rast, S. Das, M. Hirano, N. McCurley, P. Guo, N. Rohner, C. J. Tabin, P. Piccinelli, G. Elgar, M. Ruffier, B. L. Aken, S. M. Searle, M. Muffato, M. Pignatelli, J. Herrero, M. Jones, C. T. Brown, Y. W. Chung-Davidson, K. G. Nanlohy, S. V. Libants, C. Y. Yeh, D. W. McCauley, J. A. Langeland, Z. Pancer, B. Fritzsch, P. J. de Jong, B. Zhu, L. L. Fulton, B. Theising, P. Flicek, M. E. Bronner, W. C. Warren, S. W. Clifton, R. K.

Wilson and W. Li, *Nat. Genet.*, 2013, **45**, 415-421, 421e411-412.

20. M. S. Ebert and P. A. Sharp, *Cell*, 2012, **149**, 515-524.
21. L. A. Bovolenta, M. L. Acencio and N. Lemke, *BMC Genomics*, 2012, **13**, 405.
22. S. D. Hsu, Y. T. Tseng, S. Shrestha, Y. L. Lin, A. Khaleel, C. H. Chou, C. F. Chu, H. Y. Huang, C. M. Lin, S. Y. Ho, T. Y. Jian, F. M. Lin, T. H. Chang, S. L. Weng, K. W. Liao, I. E. Liao, C. C. Liu and H. D. Huang, *Nucleic Acids Res.*, 2014, **42**, D78-85.
23. M. L. Mucenski, K. McLain, A. B. Kier, S. H. Swerdlow, C. M. Schreiner, T. A. Miller, D. W. Pietryga, W. J. Scott, Jr. and S. S. Potter, *Cell*, 1991, **65**, 677-689.
24. D. D. Licatalosi and R. B. Darnell, *Nat. Rev. Genet.*, 2010, **11**, 75-87.
25. M. J. Cowley, M. Pinese, K. S. Kassahn, N. Waddell, J. V. Pearson, S. M. Grimmond, A. V. Biankin, S. Hautaniemi and J. Wu, *Nucleic Acids Res.*, 2012, **40**, D862-865.
26. H. Jeong, S. P. Mason, A. L. Barabasi and Z. N. Oltvai, *Nature*, 2001, **411**, 41-42.
27. M. Buljan, G. Chalancon, A. K. Dunker, A. Bateman, S. Balaji, M. Fuxreiter and M. M. Babu, *Curr Opin Struct Biol*, 2013, **23**, 443-450.
28. E. Potenza, T. Di Domenico, I. Walsh and S. C. Tosatto, *Nucleic Acids Res.*, 2015, **43**, D315-320.
29. T. Shlomi, T. Benyamini, E. Gottlieb, R. Sharan and E. Ruppin, *PLoS Comput. Biol.*, 2011, **7**, e1002018.
30. A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen and J. Nielsen, *Nat. Commun.*, 2014, **5**, 3083.
31. T. Shlomi, M. N. Cabili and E. Ruppin, *Mol. Syst. Biol.*, 2009, **5**, 263.
32. R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen and J. Nielsen, *Mol. Syst. Biol.*, 2014, **10**, 721.
33. K. Yizhak, O. Gabay, H. Cohen and E. Ruppin, *Nat. Commun.*, 2013, **4**, 2632.
34. K. Yizhak, S. E. Le Devedec, V. M. Rogkoti, F. Baenke, V. C. de Boer, C. Frezza, A. Schulze, B. van de Water and E. Ruppin, *Mol. Syst. Biol.*, 2014, **10**, 744.
35. O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin and T. Shlomi, *Mol. Syst. Biol.*, 2011, **7**, 501.
36. I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bolling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novere, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, Sr., M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes and B. O. Palsson, *Nat. Biotechnol.*, 2013, **31**, 419-425.
37. R. J. Kinsella, A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey and P. Flicek, *Database (Oxford)*, 2011, **2011**, bar030.
38. J. D. Orth, I. Thiele and B. O. Palsson, *Nat. Biotechnol.*, 2010, **28**, 245-248.
39. J. Y. Ryu, H. U. Kim and S. Y. Lee, *Integr. Biol.*, 2015, DOI: 10.1039/c5ib00002e.
40. D. Segre, D. Vitkup and G. M. Church, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 15112-15117.
41. A. Ebrahim, J. A. Lerman, B. O. Palsson and D. R. Hyduke, *BMC Syst. Biol.*, 2013, **7**, 74.

Figures

Fig. 1 Distribution of genes with respect to the number of their transcript variants. Among 21,983 genes obtained from Ensembl BioMart[37], 3,728 genes were found to have a single transcript, while the remaining genes had 2 to 77 transcript variants. There were 6.95 transcript variants per human gene on average. The inset shows distribution of 219 genes, each having more than 29 transcript variants, which represents 0.01% of human genes.

Fig. 2 Correlations between the number of transcript variants of human genes and the functional characteristics (i.e., HK, housekeeping; TS, tissue-selective; ES, essential; and NE, non-essential). (A) Distribution of the number of transcript variants for the HK (n = 3,804), TS (n = 2,293), ES (n = 2,472), and NE genes (n = 3,811). Boxes represent the 25th-75th percentiles, while whiskers represent the 5th-95th percentiles. The line inside the box indicates the median value of the distribution. (B) The percentages of the HK and TS genes, and (C) ES and NE genes among all the genes present in each group according to the number of transcript variants. The x-axis is the group name corresponding to the number of transcript variants and the y-axis is the percentages of HK, TS, ES, and NE genes in each group. Statistical significances for the presence of HK, TS, ES and NE genes in each group with different number of transcript variants were calculated using Fisher's exact test (Table S3, ESI†).

Fig. 3 A heat map showing percentage of the number of expressed genes in each group against each tissue. Tissue-specific expression data were obtained from proteomics studies on 32 different human tissues.[15] The percentage represents the number of expressed genes

among all the genes in each tissue. Tissue names are shown in the *x*-axis, and group names corresponding to the number of transcript variants, and the HK, TS, ES, and NE genes are indicated on the *y*-axis. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

Fig. 4 A heat map showing the percentages of orthologs in gene groups having different transcript variants in 64 vertebrate species. Data on orthologs were obtained from OrthoDB.[17] The percentage represents the number of orthologs among all the genes present in the corresponding gene group and species. The *x*-axis shows the vertebrate species, which were clustered according to their order; the order names are shown only if it has more than 3 relevant species. The *y*-axis is the group name corresponding to the number of transcript variants, and the HK, TS, ES, and NE genes. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

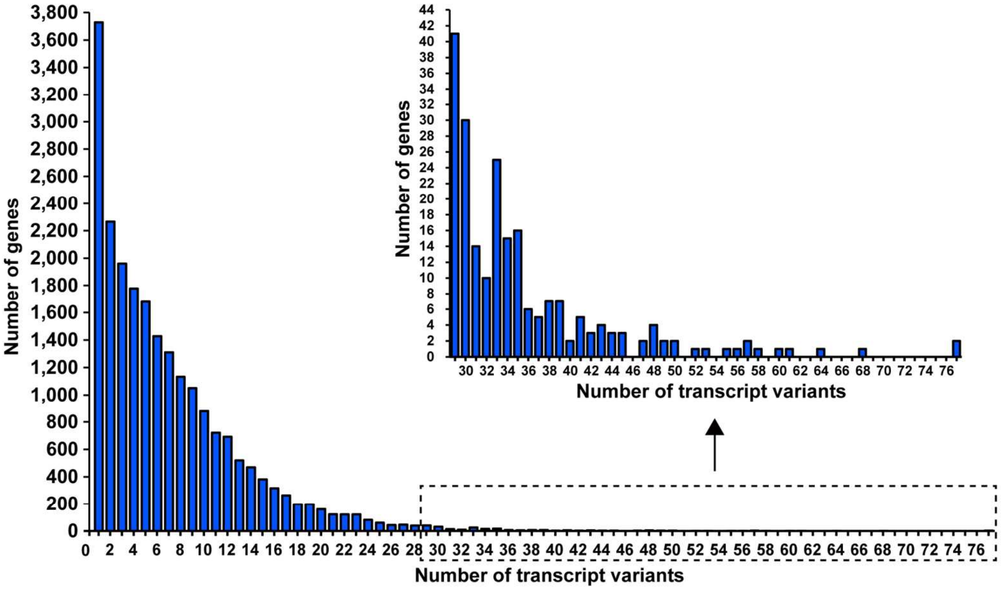
Fig. 5 Bubble plots representing average numbers of (A) transcription factors and (B) microRNAs regulating genes in each group. Information on target genes regulated by transcription factors and microRNAs was obtained from HTRIdb[21] and miRTarBase,[22] respectively. Average numbers of transcription factors and microRNAs increased for the genes having a greater number of transcript variants. Red and green bubbles represent groups having the HK and ES genes, respectively. Blue bubbles represent 60 groups classified by the number of transcript variants. The TS and NE genes (6.95 and 7.64 transcript variants on average, respectively) appeared to be regulated by 2.36 and 2.87 transcription factors, and 1.19 and 1.94 microRNAs, respectively. The bubbles for these genes are not shown because they block those of genes in 60 groups. Bubble size indicates the number of genes in each

group. Statistical significances calculated for each pair of the groups using Wilcoxon rank sum test are available in Table S6 and S7, ESI†. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

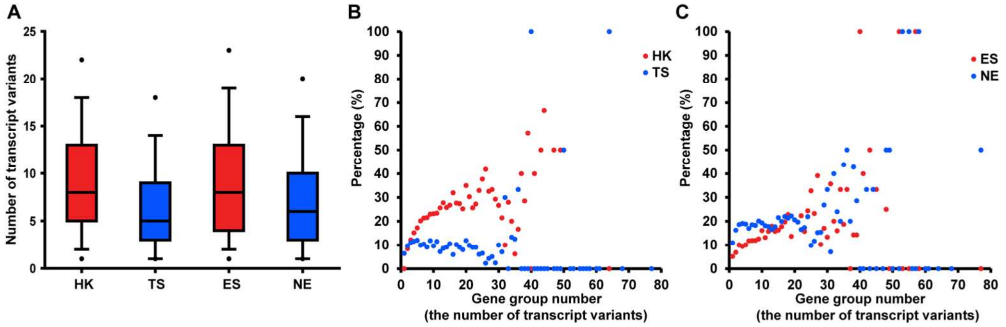
Fig. 6 Bubble plot showing the average degrees of proteins encoded by human genes in 60 groups, and the HK and ES genes in human PPI network. Protein interactome data were downloaded from PINA 2.0 database.[25] This network contains 17,109 nodes and 166,776 edges. The three bubble plots were presented for (A) all the proteins, and proteins with (B) 0% disorder only and with (C) 50-70% disorder only. Red and green bubbles represent groups having the HK and ES genes, respectively. Blue bubbles represent 60 groups classified by the number of transcript variants. Proteins encoded by the TS and NE genes (on average 6.95 and 7.64 transcript variants, respectively) had average degrees of 13.87 and 22.08, respectively. The bubbles for these genes are not shown because they block those of genes in 60 groups. Bubble size indicates the number of genes in each group. Statistical significances calculated for each pair of the groups using Wilcoxon rank sum test are available in Table S8, ESI†. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.

Fig. 7 Bubble plots showing (A) the percentage of metabolic genes to the total genes found in each group having different number of transcript variants and (B) the percentage of predicted essential genes for each group in 60 NCI-60 cancer cell line-specific metabolic models. Red and green bubbles represent groups having the HK and ES genes, respectively. Blue bubbles represent 60 groups classified by the number of transcript variants. The HK and ES genes (9.12 and 9.29 transcript variants on average, respectively) had 13.3% and 10.6% metabolic

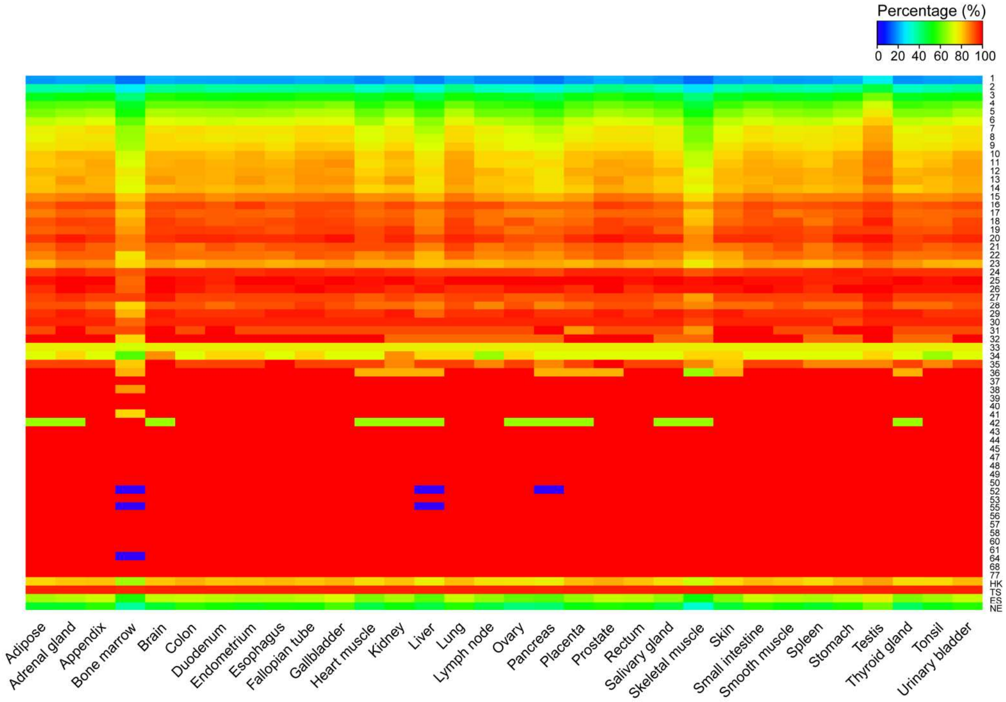
genes, and 4.5% and 4.4% essential genes, respectively. The TS and NE genes (6.95 and 7.64 transcript variants on average, respectively) had 14.8% and 10.9% metabolic genes, and 1.4% and 1.2% essential genes, respectively. The bubbles for the TS and NE genes are not shown because they block those of genes in 60 groups. Bubble size indicates the number of genes in each group. Abbreviations are: HK, housekeeping; TS, tissue-selective; ES, essential; NE, non-essential.



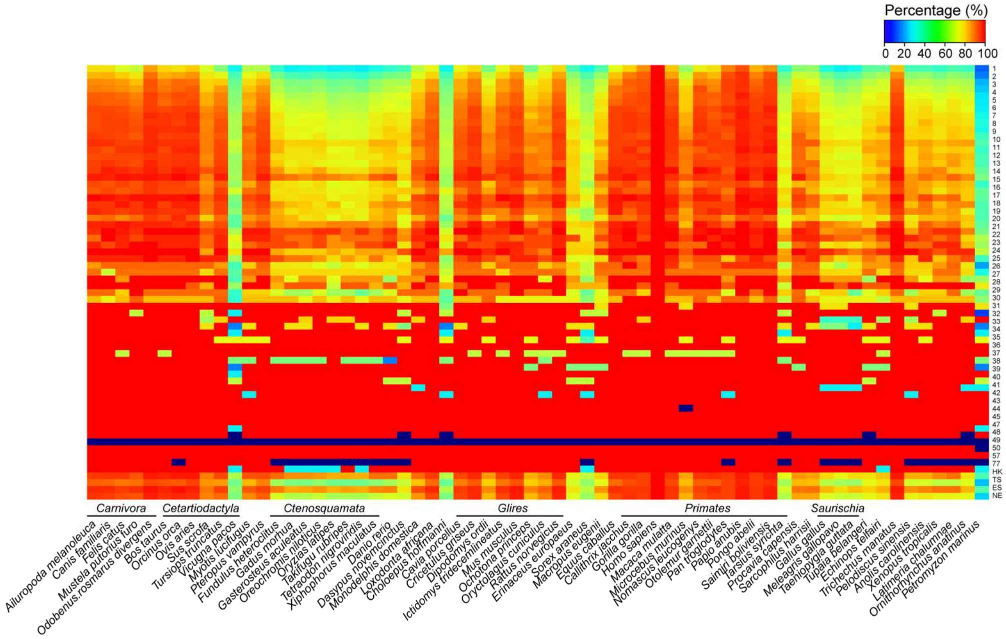
342x209mm (150 x 150 DPI)



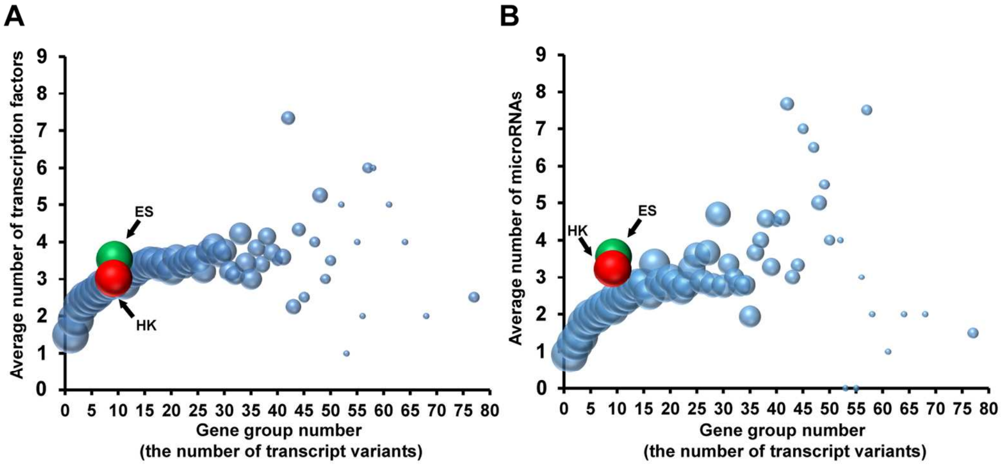
343x110mm (150 x 150 DPI)



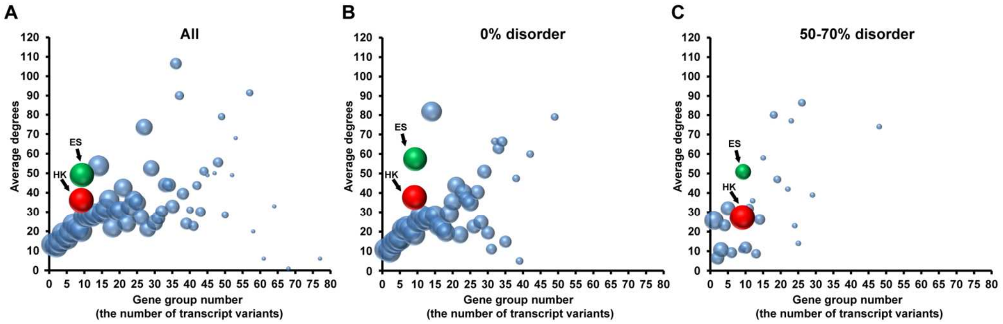
360x252mm (150 x 150 DPI)



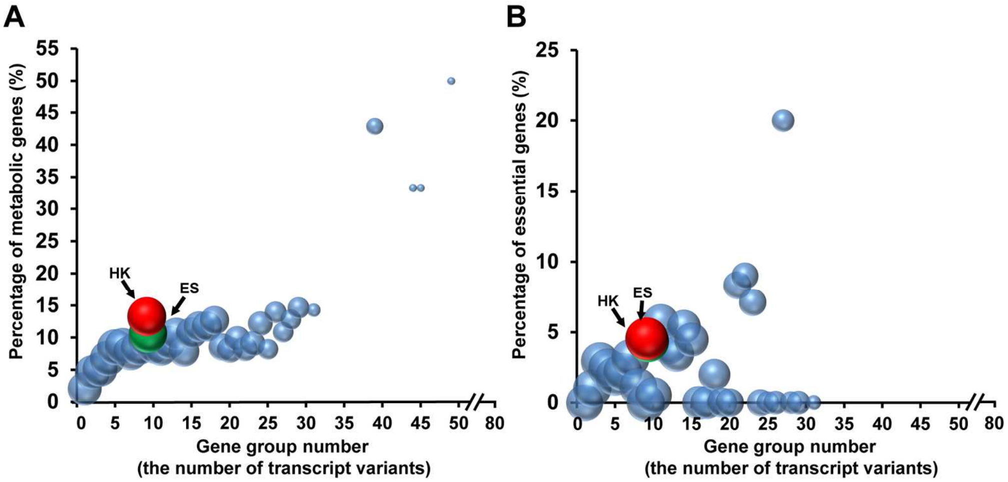
360x229mm (150 x 150 DPI)



349x161mm (150 x 150 DPI)



355x113mm (150 x 150 DPI)



360x172mm (150 x 150 DPI)