Dosage-sensitivity of human transcription factor genes

Zhihua Ni[1,2], Xiao-Yu Zhou[1], Sidra Aslam[1] and Deng-Ke Niu[1*]

[1]MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing, China

[2]College of Life Sciences, Hebei University, Baoding, China

*Correspondence:

Corresponding Author

dkniu@bnu.edu.cn, dengkeniu@hotmail.com

Running title: Dosage-sensitivity of transcription factor genes

Keywords: Transcription factor, dosage sensitive, ohnolog, conserved copy number, haploinsufficiency, disease, nuclear receptor, C2H2-ZF.

Abstract

Changes in the copy number of protein-coding genes would lead to detrimental effects if the consequent changes in protein concentration disrupt essential cellular functions. Large-scale genomic studies have identified thousands of dosage-sensitive genes in human genome. We are interested in the dosage-sensitivity of transcription factor (TF) genes whose products are essential for the growth, division and differentiation of cells by regulating the expression of

the genetic information encoded in the genome. We first surveyed the enrichment of human TF genes in four recently curated datasets of dosage-sensitive genes, including the haploinsufficient genes identified by a large-scale genomic study, the haploinsufficient genes predicted by a machine learning approach, the genes with conserved copy number across mammals, and the ohnologs. Then we selected the dosage-sensitive genes that are present in all the four dataset and regarded them as the most reliable dosage-sensitive genes, and the genes that are absent from any one of the four datasets as the most reliable dosage-insensitive genes, and surveyed the enrichments of TFs genes in these two datasets. A large number of TF genes were found to be dosage-insensitive, which is beyond the expectation based on the role of TFs. In spite of this, the likeness of TF genes to be dosage-sensitive were supported by five datasets, with the conserved-copy-number genes as the exception. The nuclear receptors are the only one family of TFs whose dosage-sensitivity was consistently supported by all the six datasets. In addition, we found that TF families with very few members are also more likely to be dosage-sensitive while the largest TF family, C2H2-ZF, are most likely dosage-insensitive. The most extensively studied TFs, p53, are not special in dosage-sensitivity. They are significantly enriched in only three datasets. We also confirmed that dosage-sensitive genes generally have long coding sequences, high expression levels and experienced stronger selective pressure. Our results indicate some TFs function in a dose-dependent manner while some other not. Gene dosage changes in some TF families like nuclear receptor would result in disease phenotypes while the effects of such changes in some TFs like C2H2-ZF would be mild.

INTRODUCTION

Transcription factors (TFs) are proteins that control the rate of transcription by binding to a specific DNA sequence (Lambert et al., 2018). Mutations affecting TFs would disrupt the accurate control of gene expression and would cause diseases. Complete homozygous loss of TF genes is rarely seen in humans while in mice the homozygote phenotype is severe and sometimes lethal (Engelkamp and vanHeyningen, 1996). Previous studies showed that some TFs work in a dose-dependent manner, heterozygous loss-of-function mutations are also associated with severe phenotypic disorders (Engelkamp and vanHeyningen, 1996; Seidman and Seidman, 2002). Meanwhile, there are also some cases where individuals with heterozygous deletion of particular TFs (e.g. THRB) did not exhibit abnormal phenotypes (Engelkamp and vanHeyningen, 1996; Seidman and Seidman, 2002). Changes in gene copy number are common both in the evolution and at the population level of animals and plants. Dosage-insensitive genes has a strong majority in the human genome (Makino et al., 2013; Lek et al., 2016; Rice and McLysaght, 2017a; Rice and McLysaght, 2017b; Shihab et al., 2017). However, considering that the functions of TFs are to insure the target genes to be expressed in the right cell at the right time and in the right amount in response to intracellular, intercellular or environmental signals, it seems odd for a TF whose dose reduction by half does not cause any phenotypic changes. Are the dosage-insensitive TF genes previously identified (Engelkamp and vanHeyningen, 1996; Seidman and Seidman, 2002) just special exceptions or is dosage-sensitivity limited within some special families of TF genes? In 2002, Seidman and Seidman surveyed the molecular causes of human syndromes (Seidman and Seidman, 2002). Among the 491 TF genes, only 27 TFs were confirmed to exhibit phenotypic haploinsufficiency by syndromic records available at that time. Members of several TF gene families appears more likely to be sensitive to copy number changes. In the past 17 years, both the annotation of TF genes in human genome and the identification of dosage-sensitive

genes have been greatly advanced. Recently, Lambert et al. curated a comprehensive list of human TFs by manually examining the putative TF lists from several sources including previous manual curations, domain searches, Gene Ontology, and crystal and NMR structures of proteins in complex with DNA taken from the Protein Data Bank (Lambert et al., 2018). This most updated list includes 1639 TFs categorized into 65 families. Another study in the same year curated 1665 TFs, a very similar number (Hu et al., 2018). On the other hand, large volumes of data have accumulated for the traditional measure of dosage-sensitive genes, haploinsufficiency. Here, we briefly describe two sources of data used in this study. To evaluate the intolerance to heterozygous protein-truncating variations, Lek et al. (2016) designed a metric, pLI, the probability of being loss-of-function intolerant. The metrics were obtained by comparing the observed rates of the truncating variations with the rates expected from variations completely tolerated by natural selection. With a cutoff of pLI $\geq$ 0.9, they identified 3230 haploinsufficient genes by analyzing 60706 human exomes. Besides the previously identified haploinsufficient genes that are associated with diseases, their dataset also embrace some genes that have not yet been assigned any human diseases. The low appearance of these genes in human populations indicates that heterozygous loss of function confers some survival or reproductive disadvantage. More recently, Shihab et al. (2017) integrated genomic and evolutionary information from several large database and predicted 7841 haploinsufficient genes in human genome using a machine learning approach called HIPred. In addition, new proxies for dosage-sensitivity have been established (Rice and McLysaght, 2017a). As the dosage-sensitivity of a gene resulting from the effects of the ratio of its product to other components of the cell, small-scale duplication containing dosage-sensitive genes would be deleterious. Therefore, the duplicability of a gene in evolution relative to all other genes in the same genome could also be used as a measure of dosage-sensitivity (Rice and McLysaght, 2017a). The pairs of genes in the same genome originating

from small-scale duplication are termed as paralogs while those originating from whole genome duplications are termed as ohnologs (Glover et al., 2016). If a gene has only ohnologs but no paralogs, its copy number changed in evolution is very likely under dosage constraint. Makino et al. (2013) compiled a list of 7294 ohnologs in human genome and confirmed that they are in deserts for copy number variations. Rice and McLysaght (2017b) found that genes in pathogenic copy-number-variation regions of human genome are more conserved in copy number across 13 mammal genomes than genes in benign regions. Therefore, the 7014 genes with the most conserved-copy-number across the 13 mammal genomes they identified could be regarded as dosage-sensitive genes. Taking advantage of the unprecedented wealth of data, we re-evaluated the dosage-sensitivity of human transcription factor genes.

Previous analyses showed that dosage-sensitive genes are generally expressed at high levels and encode large proteins, possibly because extra copies of these genes are costly in terms of cellular resources (Gout et al., 2010; Rice and McLysaght, 2017a). In their analyses, mRNA abundance was used as the measure of expression level. Although the mRNA abundance is significantly correlated with protein abundance, it is far from a perfect parameter to represent or predict protein expression levels (Greenbaum et al., 2003; Schmidt et al., 2007; Guo et al., 2008; Gry et al., 2009; Maier et al., 2009). In addition, most cellular resources consumed during the expression of a gene is the production of proteins rather than the production of mRNAs. For a potential determining factor of dosage-sensitivity resulting from the cellular resources they consumed, protein abundance is a more accurate measure than mRNA abundance. Therefore, we re-examined the relationship between dosage-sensitivity and gene expression level using protein abundance. Two additional parameters, protein size and selective force, were also used to characterize the dosage-sensitivity of TF genes.

MATERIALS AND METHODS

Two datasets of haploinsufficient genes in human genome were obtained from references (Lek et al., 2016; Shihab et al., 2017). A list of 7294 ohnologs in human genome was obtained from reference (Makino et al., 2013) and a list of 7014 copy-number conserved genes was obtained from reference (Rice and McLysaght, 2017b). The orthologous relationships between human and mouse genes and their dN and dS values were retrieved from BioMart (Ensembl version GRCh37, http://grch37.ensembl.org/biomart/martview). The data on protein abundance were downloaded from the Human Protein Atlas (Version 18.1, https://www.proteinatlas.org/) (Uhlén et al., 2015) and the length of each protein-coding sequence was retrieved from Ensembl (version GRCh37, ftp://ftp.ensembl.org/). For the multiple alternative splicing isoforms, we only retained the longest transcripts. The data including the 1639 TFs and their classifications into different families were obtained from reference (Lambert et al., 2018).

RESULTS

According to reference (Lambert et al., 2018), 1639 TF genes have been annotated in human genome. Among them, 1570 TF genes could be categorized into 64 families according to the DNA-binding domains they encoded. And 69 TFs lacking recognizable DNA-binding domains were collectively categorized into a family termed “Unknown”. The largest family is C2H2-ZF with 747 members, followed by homeodomain (196 members), bHLH (108 members), bZIP (54 members), forkhead (49 members), nuclear receptor (46 members), HMG/Sox (30 members), ETS (27 members), T-box (17 members), AT-hook (16 members), Homeodomain+POU (16 members), Myb/SANT (14 members), THAP finger (12 members), CENPB (11 members), E2F (11 members), BED ZF (10 members), GATA (10 members), Rel (10 members), etc. Chi-square test (expected value > 5) and Fisher’s exact test (expected

value ≤ 5) were used to test whether each family of TFs are more likely to be dosage-sensitive than other human genes. In addition, 46 small families, each with members ≤ 9 were also grouped together in statistical analyses and termed as small-family TFs.

The haploinsufficiency of human transcription factor genes

In the version GRCh37.p13 of human genome, 22810 protein-coding genes have been annotated. Among them, 7824 genes were predicted to be haploinsufficient by Shihab et al. (Shihab et al., 2017). That is, 34.3% of human genes are haploinsufficient. More than half (855) of the 1639 TF genes are absent from this dataset. Although the haploinsufficient TF genes (784) are in the minority, they are significantly overrepresented in this haploinsufficient gene dataset compared with other human genes (Chi-square = 19.7, df = 1, $p = 1.49\times 10^{-9}$). Among the 65 TF gene families, the gene numbers of 57 families observed in the haploinsufficient gene dataset are larger than expected by chance and the gene numbers of eight families are smaller than expected by chance. In 22 families including by homeodomain, bHLH, bZIP, forkhead, nuclear receptor and the Unknown family, the overrepresentation of TF genes in the haploinsufficient gene dataset are statistically significant ($p < 0.05$ for all cases. Table 1). In addition, the small-family TF genes are overrepresented in the haploinsufficient gene dataset (Chi-square = 24.3, df = 1, $p = 8.1\times10^{-7}$). Only one TF family are significantly underrepresented in the haploinsufficient gene dataset, it is the largest TF family C2H2-ZF (Chi-square = 5.9, df = 1, $p = 0.015$).

Among the 3230 haploinsufficient genes identified by Lek et al. (2016), there are 384 TF genes. Although TF genes are statistically overrepresented in this haploinsufficient genes (Chi-square = 37.5, df = 1, $p = 9.35 \times 10^{-10}$), the haploinsufficient TF genes take only 23% of all human TF genes. In 25 families, with gene numbers ranging from two to 108, the TF genes are significantly overrepresented in the haploinsufficient gene dataset ($p < 0.05$ for all cases.

Table 2). The other families are not significantly different from other human genes in their frequency to be haploinsufficient ($p > 0.05$ for all cases). The small-family TF genes are also overrepresented in the haploinsufficient gene dataset of Lek et al. (Chi-square = 34.0, df =1, $p = 5.3 \times 10^{-9}$).

Enrichment of transcription factor genes in the ohnolog dataset

Ohnologs are duplicates originating only from whole genome duplication. Their dosages are tightly bound to the dosage of the genome. Makino et al. (2013) identified 7294 ohnologs, but 220 of them are not present in the current version of human genome, GRCh37.p13. The remaining 7074 ohnologs take 31% of the human genome. In this dataset, we found 818 TF genes, 49.9% of all human TF genes. Chi-square test showed that TF genes are significantly enriched in the dataset of ohnologs (Chi-square = 72.3, df = 1, $p < 2 \times 10^{-16}$). The genes in 19 families of TFs (ARID/BRIGHT;RFX, Brinker, DM, CENPB, CBF/NF-Y, etc) do not present in the dataset of ohnologs genes. The underrepresentation of CENPB is statistically significant (Fisher's exact test, $p < 0.05$. Table 3). Many families (including C2H2-ZF, bHLH, nuclear receptor, bZIP, Homeodomain+POU, etc.) exhibit significant overrepresentation in the ohnolog dataset ($p < 0.05$ for all cases. Table 3). In some families (C2H2-ZF-Homeodomain, Runt, p53, EBF1, Paired box, TEA, AP-2, CSD, homeodomain-paired_box, grainyhead, STAT, and IRF), all the TF genes are ohnologs. The small-family TF genes are also overrepresented in the ohnolog dataset (Chi-square = 15.8, df =1, $p = 7.1 \times 10^{-5}$). It should be specially noted that, only in this dataset, C2H2-ZF genes are statistically overrepresented in dosage-sensitive genes.

Copy-number conservation of transcription factor genes

In total, 7014 genes (33.7%) were found to have conserved copy-numbers across mammalian genomes (Rice and McLysaght, 2017b). Among the 1639 TF genes, there are 493 TF genes whose copy-numbers were conserved. Chi-square test showed that these 493 TF genes are not significantly overrepresented or underrepresented in the dataset of conserved copy-number genes (Chi-square = 0.12, df = 1, $p = 0.73$). However, we noticed that C2H2-ZF genes and CENPB genes are significantly underrepresented in this dataset ($p < 0.05$ for both cases. Table 4). Nuclear receptor genes, IRF genes and E2F genes are significantly overrepresented in this dataset ($p < 0.05$ for all the three cases. Table 4). In addition, the small-family TF genes also significantly enriched in this dataset (Chi-square = 6.16, df = 1, $p = 0.013$).

Most reliable dosage-sensitive (MRDS) transcription factor genes

We noticed that the datasets of dosage-sensitive genes obtained from different measures vary substantially (Figure 1). The dosage-sensitivity of 853 genes was consistently supported by four independent studies, so we could regard them as the MRDS genes. In addition, there are 9459 genes that were not detected to be dosage-sensitive in any of the four studies (Supplementary Data S1). These 9459 genes could be regarded as the most reliable dosage-insensitive (MRDIS) genes (Supplementary Data S1). With the most reliable datasets, we evaluated the dosage-sensitivity of TF genes further.

Among the 853 MRDS genes, there are 125 TF genes. Chi-square test showed that TF genes are significantly enriched in the dataset of MRDS genes (Chi-square = 21.8, df = 1, $p = 3 \times 10^{-6}$). In 30 TF families, the gene numbers observed in this dataset are larger than expected by chance and in 12 TF families including bHLH, nuclear receptor, T-box, Rel, etc., the differences are statistically significant ($p < 0.05$ for all the 12 cases. Table 5). No gene families are significantly underrepresented in the dataset of MRDS genes ($p > 0.05$ for all

cases). The small-family TF genes are significantly overrepresented in the MRDS gene dataset (Chi-square = 15.2, df = 1, $p = 10^{-4}$).

Among the 9459 MRDIS genes, there are 454 TF genes, significantly fewer than the number (680) expected by chance (Chi-square = 44.9, df = 1, $p = 2 \times 10^{-11}$). In 56 TF gene families, the observed gene numbers in this dataset are smaller than expected by chance and statistically significant differences were obtained in 18 TF gene families including Homeodomain, nuclear receptor, bZIP, E2F, bHLH, Rel, etc. ($p < 0.05$ for all the 18 cases. Table 6). The CENPB TF gene family is significantly overrepresented in the dataset of MRDIS genes (Fisher exact test, $p = 0.011$). The small-family TF genes are significantly underrepresented in the MRDIS gene dataset (Chi-square = 26.6, df = 1, $p = 2.5\times 10^{-7}$).

Characterizations of dosage-sensitive transcription factor genes

Previous analyses have characterized the dosage-sensitive genes as encoding large proteins, under strong selective force and being highly expressed (Gout et al., 2010; Schuster-Böckler et al., 2010; Rice and McLysaght, 2017a). In this study, we first verified these characterizations using the datasets of MRDS genes and MRDIS genes then examined whether the dosage-sensitive TF genes exhibit the same characteristics.

The coding sequences (CDS) of the 853 MRDS genes are significantly larger than those of the 9280 MRDIS genes (Mann-Whitney *U* test, $p < 10^{-6}$), with their median values differ more than 2 folds (2247 *vs*. 927). The MRDS TF genes and the MRDIS TF genes also differ significantly in their CDS length (Mann-Whitney U test, $p < 10^{-6}$), but with a smaller difference in their median values (1770 *vs*. 1490).

The selective pressure of each gene was measured by the dN/dS value for the human gene relative to its mouse ortholog. Valid dN/dS values were obtained for 4654 MRDIS genes and all the 853 MRDS genes. Mann-Whitney *U* test showed that the MRDS genes have

experienced significantly stronger selective pressure than the MRDIS genes ($p < 10^{-6}$, median values 0.045 *vs*. 0.183). The TF genes exhibit the same pattern (Mann-Whitney U test, $p < 10^{-6}$, median values: 0.051 *vs*. 0.201).

In the Human Protein Atlas database, the expression level of a protein, if has ever been studied, was defined as one of the following four categories in each studied cell sample: high, medium, low, and not detected. For a comparative study, we retained 31 cell samples in which all the MRDS and MRDIS TF genes have been studied. In any of these 31 samples, 659 MRDS genes including 95 MRDS TF genes and 3192 MRDIS genes including 186 MRDIS TF genes have been studied. For each gene, we surveyed its expression level in the 31 cell samples. For example, the expression level of gene *ENSG00000095951*, one member of C2H2-ZF gene family, is high in 9 samples, medium in 12 samples, low in 8 samples, and not detected in 2 samples. Then we compared the MRDS genes and MRDIS genes in their presences in each of the four categories. On average, each MRDS gene has high expression levels in two samples while each MRDIS has high expression levels in one sample. Mann-Whitney *U* test showed that the difference is statistically significant ($p < 10^{-6}$; Figure. 2). Similarly, at medium and low level, MRDS genes were detected in significantly more samples than MRDIS genes (Mann-Whitney *U* test, $p < 10^{-6}$ for both cases; Figure. 2). By contrast, in the “not detected” category, MRDS genes were assigned to significantly fewer samples than MRDIS genes (Mann-Whitney *U* test, $p < 10^{-6}$ for both cases; Figure. 2). Unexpectedly, no significant differences were detected between MRDS TF genes and MRDIS TF genes in any of the four categories (Mann-Whitney *U* test, $p > 0.05$ for all cases). That is, we confirmed the previous study that dosage-sensitive genes have higher expression levels than dosage-insensitive genes. However, our results suggested that dosage-sensitive TF genes do not have higher expression levels than dosage-insensitive TF genes. Furthermore, we

found that TF genes differ from other human genes significantly at medium expression level and in the "not detected" category (Mann-Whitney U test, $p < 10^{-6}$ for both cases; Figure. 3).

DISCUSSION

Including the MRDS and MRDIS gene datasets, we have surveyed the dosage-sensitivity of TF genes among six datasets. Although the results obtained from different datasets vary substantially, some consensuses could be observed (Table 7). First, more than half of the TF genes are dosage-insensitive. The large number of dosage-insensitive TF genes is beyond the expectation from their role to ensure accurate expression of their target genes. Second, nuclear receptor is a very special TF gene family. It is the only family whose enrichment of dosage-sensitive genes were consistently supported by all the six datasets. Third, TF gene families with fewer numbers are more likely to enrich dosage-sensitive genes. This was supported by the consistent observations of overrepresentation of the small-family TF genes in the five dosage-sensitive gene dataset and underrepresentation of them in the MRDIS gene dataset. If adopting a less conservative criterion (i.e., supported by five datasets), all TF genes as a whole, bHLH, E2F, Rel, and SMAD are enriched in dosage-sensitive genes. The most extensively studied TF family, p53 genes, are significantly enriched in only three datasets, the haploinsufficient genes detected by Lek et al. (2016), the haploinsufficient genes predicted by Shihab et al. (2017) and the ohnologs identified by Makino et al. (2013). No TF gene families were consistently underrepresented in six or five datasets. The CENPB genes are significantly underrepresented in the ohnolog gene dataset and the dataset of genes with conserved-copy-number, and significantly overrepresented in the MRDIS gene dataset. In addition, members of the largest TF gene family, C2H2-ZF, are significantly underrepresented in two dataset, the haploinsufficient genes predicted by Shihab et al. (2017) and the genes with conserved-copy-number (2017b).

In *Saccharomyces cerevisiae*, one of the two genomic copies of some genes have been experimental eliminated (Springer et al., 2010; Torres et al., 2016). Most of these deletions did not significantly affect the growth rate of the mutant yeasts. It has been proposed that the levels of many eukaryotic proteins can be neutral for fitness over a broad range (Springer et al., 2010). If the dramatic changes in protein levels have little or no evolutionary consequences in yeast, organisms with smaller effective population size and more effective genetic drift would have more genes whose dosage changes are neutral. In this context, it is reasonable to see that most human genes are dosage-insensitive. Considering the role of TF genes in regulation of gene expression, the existence of large number of dosage-insensitive TF genes is beyond intuition. One possible explanation is that some TFs do not directly regulate the expression level of other genes, but simply contribute, more or less, to the maintenance of a nuclear environment compatible to the expression of required genes. Nonfunctional interactions within the accessible portions of the genome have been observed for many TFs (Fisher et al., 2012; Slattery et al., 2014). Similar to this idea, some TFs, like KRAB zinc finger proteins, are involved in the repression of transposable elements (Thomas and Schneider, 2011; Imbeault et al., 2017; Yang et al., 2017). The KRAB zinc finger proteins is a subfamily of the TF family C2H2-ZF, which are one of the most dosage-insensitive TF families. As there is no purifying selection against the accumulation of degenerative mutations in the repressed transposable elements, most transposable elements would quickly lose their transposition potential in evolution. Then, the dosage changes of the TFs that target the transposable elements with low or no transposition potential would not have severe effects. Only a few members of KRAB zinc finger proteins whose targets have high transposition potential would be dosage-sensitive. In addition, if a large proportion of TF binding is nonfunctional, the effects of dosage changes on the expression of target genes would be buffered (Todeschini et al., 2014). However, there is also evidence for the dosage-

sensitivity of the TFs that bind to conserved noncoding sequences or critically determine the chromatin structure of target genes (Lundgren et al., 2000; Schnable et al., 2011).

One consensus of our analyses of the six datasets is that small-family TF genes are more likely to be dosage-sensitive. We could not see any special common characteristics among these small-family TFs. Instead, we think that the cause and effect should be considered in opposite direction. It is the strong dosage-sensitivity of these TF genes that act as a selective force against expansion of these families and maintain the small numbers of members in evolution.

The most noticeable dosage-sensitive TF gene family is nuclear receptor. Their dosage-sensitivity were consistently supported by all the six datasets. Nuclear receptors could activate target genes after binding the nonpolar ligands like estrogen, progesterone, retinoic acid, oxysterols, and thyroid hormone. In contrast to other TFs, nuclear receptors respond directly to extracellular changes by binding the ligands diffusible across the plasma membrane (Sladek, 2011). A distinguishing characteristic of steroid hormones is the dose-response curve manifested in their regulation of gene expression (Szapary et al., 1999; Simons, 2006). Fixed dosages of the nuclear receptors for steroid hormones might be the premise of the dose-response curve.

In plants, the TF WUS provides an insightful example for the mechanism of dosage-sensitivity of TFs (Hofhuis and Heidstra, 2018). With the binding of different co-factors, WUS can activate its target gene CLV3 when its protein is expressed at low level while it will repress the target gene when its protein is at high level. For the WUS gene, both increase and decrease of the dosage would disturb its function. The dosage of WUS gene has been fixed by its dual functions as an activator and a repressor. We suppose that the most dosage-sensitivity TFs would be those whose function could be modulated towards activator and repressor properties. However, the literatures on dosage-sensitivity, we could not find any measure or

data on the degree of dosage-sensitivity. The MRDS genes we defined have just high possibility to be dosage-sensitive, rather than the most sensitive genes. However, considering that the most sensitive genes are most likely to be detected in the studies of dosage-sensitivity, the MRDS genes are likely the appropriate candidates for future studies on the mechanism of dosage sensitivity. In this sense, the nuclear receptor genes as well as genes encoding bHLH, E2F, Rel, and SMAD are likely the most dosage-sensitive TF genes. Besides the activator property, some nuclear receptors also bind to other transcription factors and deactivate the second transcription factor and its target genes in a process known as transrepression (Pascual and Glass, 2006). A TF gene with two or more functions are more likely to be sensitive to dosage changes because dosage constraint in any function would cause the gene to be dosage-sensitive.

Recently, Ishikawa et al. suggested that degradation of excessive subunits of protein complex might be the mechanism of dosage compensation that buffers gene copy number changes (Ishikawa et al., 2017). If so, a TF gene whose protein product form heteromers with the products of other genes would be more likely dosage-insensitive. Therefore, we surveyed the binding modes of the TF genes. Unexpectedly, we found only three MRDS TF genes and five MRDIS TF genes whose products form obligate heteromers with other proteins. The majority of TFs (97.6%) are monomers or homomultimers. Because more than half of TF genes are dosage-insensitive and their proteins present as monomer or homomultimer, degradation of excessive subunits is unlikely the primary mechanism for the dosage compensation of dosage-insensitive TF genes.

In line with the high expression level and long CDS of dosage sensitive genes, the expression burden was suggested to be a possible explanation for the dosage-sensitivity (Rice and McLysaght, 2017a). We confirmed that dosage-sensitive genes are generally expressed at high level. But we did not find the same pattern in TF genes. At least in TF genes, the dosage-

sensitivity could not be explained by expression burden. The small dN/dS values and so the high selective pressure experienced by the dosage-sensitive genes indicate that the functional constraints might be tightly associated with a dosage constraint. For an extreme example, if the loss and gain of a dispensable gene have not any adaptive consequence, deletion or insertion of one allele unlikely have any phenotypic effects.

CONCLUSIONS

In diploid organisms, copy number variations of protein-coding genes are common either in evolution or at population level. Some of them have been found to be associated with serious phenotypic changes, which are various diseases in humans. The sensitivity to dosage changes could be measured either from the frequency of the variations in population or the constraints of copy number changes in evolution (Rice and McLysaght, 2017a). Perhaps because of the numerous cases of the association between copy-number changes and diseases, the dosage-sensitivity of human genes are studied more extensively than other model organisms (Springer et al., 2010; Thomas and Schneider, 2011; Cook et al., 2012; Makino et al., 2013; Meinke, 2013; Lek et al., 2016; Torres et al., 2016; Rice and McLysaght, 2017a; Rice and McLysaght, 2017b; Shihab et al., 2017). TFs are the special group of proteins that regulate the expression of other genes to ensure the normal growth and division of cells. Their copy number changes are expected to affect the phenotype more seriously. 17 years ago, Seidman and Seidman surveyed the dosage-sensitivity of human TF genes using the data available at that time. They found that the dosage-sensitivity vary greatly among TF gene families (Seidman and Seidman, 2002). Taking the advantage of the rapid progresses in the identification of both TF genes and the dosage sensitive genes in the past decade, we selected four large-scale datasets and extensively surveyed the dosage-sensitivity of human TF genes. A large number of TF genes were found to be dosage-insensitive. In spite of this, TF genes

are statistically more likely to be dosage-sensitive than other genes in human genome in most datasets we used (5/6). Only in one TF family, the nuclear receptor, the dosage-sensitivity was consistently supported by all the six datasets we used. In addition, we found that TF families with very few members are also more likely to be dosage-sensitive. The largest TF family, C2H2-ZF, are most likely dosage-insensitive TFs. They were either significantly underrepresented in dosage-sensitive genes, or not significantly different from other human genes, depending on the datasets surveyed. Our results indicate that, not all TFs function in a dose-dependent manner. Gene dosage changes and consequently protein concentration changes in some TFs like nuclear receptor would affect the phenotypes seriously while such changes in some other TFs like C2H2-ZF would be mild to the host.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

DKN conceived the study and wrote the manuscript. ZN retrieved the data from online databases and performed all the statistical analyses. XYZ and SA each repeated some of the analyses. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31671321).

ACKNOWLEDGMENTS

We appreciate the kind helps from Er-Li Pang and the High Performance Computing Center of Hebei University.

Abbreviations

TF: transcription factor; MRDS: most reliable dosage-sensitive; MRDIS: most reliable dosage-insensitive.

REFERENCES

Cook, R.K., Christensen, S.J., Deal, J.A., Coburn, R.A., Deal, M.E., Gresens, J.M., et al. (2012). The generation of chromosomal deletions to provide extensive coverage and subdivision of the *Drosophila melanogaster* genome. *Genome Biology* 13(3)**,** R21. doi: 10.1186/gb-2012-13-3-r21.

Engelkamp, D., and vanHeyningen, V. (1996). Transcription factors in disease. *Current Opinion in Genetics & Development* 6(3)**,** 334-342. doi: 10.1016/s0959-437x(96)80011-6.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 109(52)**,** 21330-21335. doi: 10.1073/pnas.1209589110.

Glover, N.M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What are they and how do we infer them? *Trends in Plant Science* 21(7)**,** 609-621. doi: 10.1016/j.tplants.2016.02.005.

Gout, J.-F., Kahn, D., Duret, L., and Paramecium Post-Genomics Consortium (2010). The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLOS Genetics* 6(5)**,** e1000944. doi: 10.1371/journal.pgen.1000944.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology* 4(9)**,** 117-226. doi: 10.1186/gb-2003-4-9-117.

Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., et al. (2009). Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* 10(1)**,** 365. doi: 10.1186/1471-2164-10-365.

Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G.G., Liu, Y., et al. (2008). How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochimica et Biophysica Sinica* 40(5)**,** 426-436. doi: 10.1111/j.1745-7270.2008.00418.x.

Hofhuis, H.F., and Heidstra, R. (2018). Transcription factor dosage: more or less sufficient for growth. *Current Opinion in Plant Biology* 45**,** 50-58. doi: 10.1016/j.pbi.2018.05.008.

Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2018). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research***,** gky822-gky822. doi: 10.1093/nar/gky822.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543**,** 550-554. doi: 10.1038/nature21683.

Ishikawa, K., Makanae, K., Iwasaki, S., Ingolia, N.T., and Moriya, H. (2017). Post-translational dosage compensation buffers genetic perturbations to stoichiometry of protein complexes. *PLOS Genetics* 13(1)**,** e1006554. doi: 10.1371/journal.pgen.1006554.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y.M., Albu, M., et al. (2018). The human transcription factors. *Cell* 172(4)**,** 650-665. doi: 10.1016/j.cell.2018.01.029.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536**,** 285-291. doi: 10.1038/nature19057.

Lundgren, M., Chow, C.-M., Sabbattini, P., Georgiou, A., Minaee, S., and Dillon, N. (2000). Transcription factor dosage affects changes in higher order chromatin structure associated with activation of a heterochromatic gene. *Cell* 103(5)**,** 733-743. doi: 10.1016/S0092-8674(00)00177-X.

Maier, T., Guell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Letters* 583(24)**,** 3966-3973. doi: 10.1016/j.febslet.2009.10.036.

Makino, T., McLysaght, A., and Kawata, M. (2013). Genome-wide deserts for copy number variation in vertebrates. *Nature Communications* 4**,** 2283. doi: 10.1038/ncomms3283.

Meinke, D.W. (2013). A survey of dominant mutations in *Arabidopsis thaliana*. *Trends in Plant Science* 18(2)**,** 84-91. doi: 10.1016/j.tplants.2012.08.006.

Pascual, G., and Glass, C.K. (2006). Nuclear receptors versus inflammation: mechanisms of transrepression. *Trends in Endocrinology & Metabolism* 17(8)**,** 321-327. doi: 10.1016/j.tem.2006.08.005.

Rice, A.M., and McLysaght, A. (2017a). Dosage-sensitive genes in evolution and disease. *BMC Biology* 15**,** 10. doi: 10.1186/s12915-017-0418-y.

Rice, A.M., and McLysaght, A. (2017b). Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nature Communications* 8**,** 14366. doi: 10.1038/ncomms14366.

Schmidt, M.W., Houseman, A., Ivanov, A.R., and Wolf, D.A. (2007). Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*. *Molecular Systems Biology* 3**,** 79. doi: 10.1038/msb4100117.

Schnable, J., Pedersen, B., Subramaniam, S., and Freeling, M. (2011). Dose–sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. *Frontiers in Plant Science* 2(2). doi: 10.3389/fpls.2011.00002.

Schuster-Böckler, B., Conrad, D., and Bateman, A. (2010). Dosage sensitivity shapes the evolution of copy-number varied regions. *PLOS ONE* 5(3)**,** e9474. doi: 10.1371/journal.pone.0009474.

Seidman, J.G., and Seidman, C. (2002). Transcription factor haploinsufficiency: when half a loaf is not enough. *Journal of Clinical Investigation* 109(4)**,** 451-455. doi: 10.1172/JCI15043.

Shihab, H.A., Rogers, M.F., Campbell, C., and Gaunt, T.R. (2017). HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics* 33(12)**,** 1751-1757. doi: 10.1093/bioinformatics/btx028.

Simons, S.S. (2006). How much is enough? Modulation of dose-response curve for steroid receptor-regrulated gene expression by changing concentrations of transcription factor. *Current Topics in Medicinal Chemistry* 6(3)**,** 271-285. doi: 10.2174/156802606776173465.

Sladek, F.M. (2011). What are nuclear receptor ligands? *Molecular and Cellular Endocrinology* 334(1)**,** 3-13. doi: 10.1016/j.mce.2010.06.018.

Slattery, M., Zhou, T.Y., Yang, L., Machado, A.C.D., Gordan, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* 39(9)**,** 381-399. doi: 10.1016/j.tibs.2014.07.002.

Springer, M., Weissman, J.S., and Kirschner, M.W. (2010). A general lack of compensation for gene dosage in yeast. *Molecular Systems Biology* 6**,** 368. doi: 10.1038/msb.2010.19.

Szapary, D., Huang, Y., and Simons, J.S.S. (1999). Opposing effects of corepressor and coactivators in determining the dose-response curve of agonists, and residual agonist activity of antagonists, for glucocorticoid receptor-regulated gene expression. *Molecular Endocrinology* 13(12)**,** 2108-2121. doi: 10.1210/mend.13.12.0384.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. *Genome Research* 21(11)**,** 1800-1812. doi: 10.1101/gr.121749.111.

Todeschini, A.-L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends in Genetics* 30(6)**,** 211-219. doi: 10.1016/j.tig.2014.04.002.

Torres, E.M., Springer, M., and Amon, A. (2016). No current evidence for widespread dosage compensation in *S. cerevisiae*. *Elife* 5**,** e10996. doi: 10.7554/eLife.10996.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science* 347(6220)**,** 1260419. doi: 10.1126/science.1260419.

Yang, P., Wang, Y., and Macfarlan, T.S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends in Genetics* 33(11)**,** 871-881. doi: 10.1016/j.tig.2017.08.006.

Figure Legends

Figure 1. Venn diagram displaying the most reliable dosage-sensitive (MRDS) genes and the most reliable dosage-insensitive (MRDIS) genes. The 853 dosage-sensitive genes shared by the four datasets, Shihab et al. (2017), Lek et al. (2016), Makino et al. (2013) and Rice et al. (2017b), are regarded as MRDS genes and the 9459 genes that are absent from any of the four datasets of dosage-sensitive genes are regarded as MRDIS genes.

Figure 2. Comparison of the expression level between the most reliable dosage-sensitive (MRDS) genes and the most reliable dosage-insensitive (MRDIS) genes. The definition of MRDS genes and MRDIS genes can be seen in Figure 1. The *p* values were calculated using Mann-Whitney *U* test.

Figure 3. Comparison of the expression level between transcription factor genes and other genes in human genome. The *p* values were calculated using Mann-Whitney *U* test.

Table 1. Appearance of TF genes in the dataset of haploinsufficient genes predicted by Shihab et al. (Shihab et al., 2017)

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
bHLH	108	68	37.0	0.003	9.1
bZIP	54	39	18.5	0.007	7.3
C2H2_ZF	747	204	256.2	0.015	5.9
C2H2_ZF;Homeodomain	4	4	1.4	0.014	
CSD	5	5	1.7	0.005	
CUT;Homeodomain	7	6	2.4	0.008	
CxxC	9	7	3.1	0.010	
E2F	11	9	3.8	0.002	
Ets	27	20	9.3	0.047	3.9
GATA	10	8	3.4	0.004	
Grainyhead	6	6	2.1	0.002	
Homeodomain	196	124	67.2	4×10^{-5}	16.9
IRF	9	9	3.1	7×10^{-5}	
Myb/SANT	14	9	4.8	0.023	
Nuclear_receptor	46	41	15.8	8×10^{-4}	11.2
p53	3	3	1.0	0.040	
Paired_box	4	4	1.4	0.014	
Rel	10	10	3.4	2×10^{-5}	
RFX	8	6	2.7	0.023	
Runt	3	3	1.0	0.040	
SMAD	9	9	3.1	7×10^{-5}	
STAT	7	7	2.4	6×10^{-4}	
Unknown	69	40	23.7	0.041	4.2
All TFs in human genome	1639	784	562.2	10^{-9}	36.5
Small-family TFs	186	133	63.8	8×10^{-7}	24.3

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members.

Table 2. Appearance of TF genes in the dataset of haploinsufficient genes detected by Lek et al. (Lek et al., 2016)

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
AP-2	5	4	0.7	0.002	
ARID/BRIGHT	6	3	0.8	0.041	
AT-hook	16	11	2.3	10^{-6}	
bHLH	108	31	15.3	0.021	5.3
C2H2-ZF;Homeodomain	4	4	0.6	4×10^{-4}	
CG-1	2	2	0.3	0.020	
CUT;Homeodomain	7	4	1.0	0.010	
CxxC	9	7	1.3	3×10^{-5}	
CxxC;AT-hook	2	2	0.3	0.020	
E2F	11	5	1.6	0.012	
EBF1	4	3	0.6	0.010	
Ets	27	8	3.8	0.045	
GATA	10	4	1.4	0.041	
Grainyhead	6	5	0.8	3×10^{-4}	
HMG/Sox	30	11	4.2	0.002	
MBD	8	5	1.1	0.002	
Myb/SANT	14	6	2.0	0.009	
Nuclear receptor	46	26	6.5	6×10^{-4}	11.7
p53	3	3	0.4	0.003	
Pipsqueak	2	2	0.3	0.020	
Rel	10	10	1.4	3×10^{-9}	
RFX	8	5	1.1	0.002	
SMAD	9	5	1.3	0.004	
STAT	7	5	1.0	9×10^{-4}	
T-box	17	9	2.4	2×10^{-4}	
All TFs in human genome	1639	384	232.1	9×10^{-10}	37.5
Small-family TFs	186	89	26.3	5×10^{-9}	34.0

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members.

Table 3. Appearance of TF genes in the dataset of ohnologs identified by Makino et al. (Makino et al., 2013)

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
AP-2	5	5	1.6	0.003	
bHLH	108	63	33.5	0.003	9.0
bZIP	54	40	16.7	0.002	9.5
C2H2_ZF	747	276	231.7	0.049	3.9
C2H2_ZF;Homeodomain	4	4	1.2	0.009	
CENPB	11	0	3.4	0.022	
CSD	5	5	1.6	0.003	
CUT;Homeodomain	7	6	2.2	0.005	
CxxC	9	6	2.8	0.030	
E2F	11	11	3.4	3×10^{-6}	
EBF1	4	4	1.2	0.009	
GATA	10	7	3.1	0.013	
Grainyhead	6	6	1.9	8×10^{-4}	
HMG/Sox	30	20	9.3	0.048	3.9
Homeodomain	196	133	60.8	2×10^{-7}	26.9
Homeodomain;Paired_box	5	5	1.6	0.003	
Homeodomain;POU	16	15	5.0	3×10^{-7}	
IRF	9	9	2.8	3×10^{-5}	
Nuclear_receptor	46	43	14.3	10^{-4}	14.4
p53	3	3	0.9	0.030	
Paired_box	4	4	1.2	0.009	
Rel	10	9	3.1	2×10^{-4}	
Runt	3	3	0.9	0.030	
SMAD	9	8	2.8	6×10^{-4}	
STAT	7	7	2.2	3×10^{-4}	
TEA	4	4	1.2	0.009	
All TFs in human genome	1639	818	508.3	2×10^{-16}	72.3
Small-family TFs	186	109	57.7	7×10^{-5}	15.8

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members.

Table 4. Appearance of TF genes in the dataset of genes with conserved-copy-number across mammals identified by Rice et al. (Rice and McLysaght, 2017b)

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
C2H2-ZF	747	142	229.7	5×10^{-6}	20.7
CENPB	11	0	3.4	0.023	
E2F	11	7	3.4	0.042	
IRF	9	6	2.8	0.029	
Nuclear receptor	46	31	14.1	0.012	6.3
Small-family TFs	186	87	57.2	0.013	6.2

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members.

Table 5. Appearance of TF genes in the dataset of most reliable dosage-sensitive genes.

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
AP-2	5	3	0.2	5×10^{-4}	
bHLH	108	13	4.0	2×10^{-4}	
C2H2-ZF;Homeodomain	4	3	0.1	2×10^{-4}	
CUT;Homeodomain	7	2	0.3	0.026	
Grainyhead	6	4	0.2	3×10^{-5}	
Paired-box	4	2	0.1	0.008	
Rel	10	3	0.4	0.005	
RFX	8	3	0.3	0.003	
SMAD	9	2	0.3	0.042	
T-box	17	5	0.6	3×10^{-4}	
TCR/CxC	1	1	0.0	0.0374	
All TFs in human genome	1639	125	61.3	3×10^{-6}	21.8
Small-family TFs	186	31	7.0	10^{-4}	15.2

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members. The most reliable dosage-sensitive genes were defined as the common dosage-sensitive genes of the four datasets used in Table 1-4.

Table 6. Appearance of TF genes in the dataset of most reliable dosage-insensitive genes.

TF family	Total number in human genome	Observed number in this dataset	Expected number in this dataset	*p* value	Chi square
ARID/BRIGHT	6	0	2.5	0.045	
AT-hook	16	1	6.6	0.041	4.2
bHLH	108	22	44.8	0.005	7.8
bZIP	54	5	22.4	0.001	11.0
CENPB	11	9	4.6	0.011	
CUT;Homeodomain	7	0	2.9	0.046	
CxxC	9	0	3.7	0.013	
E2F	11	0	4.6	0.004	
Ets	27	3	11.2	0.030	4.7
Grainyhead	6	0	2.5	0.045	
HMG/Sox	30	3	12.4	0.016	5.8
Homeodomain	196	30	81.3	10^{-6}	23.6
Homeodomain;POU	16	1	6.6	0.041	4.2
IRF	9	0	3.7	0.013	
Nuclear receptor	46	0	19.1	10^{-5}	19.1
Rel	10	0	4.1	0.007	
SMAD	9	0	3.7	0.013	
STAT	7	0	2.9	0.046	
Unknown	69	7	28.6	3×10^{-4}	13.1
All TFs in human genome	1639	454	679.7	2×10^{-11}	44.9
Small-family TFs	186	25	77.1	2×10^{-7}	26.6

Chi-square test (expected value > 5) and Fisher's exact test (expected value ≤ 5) were used to test the overrepresentation or underrepresentation of the TF genes in this dataset. The small-family TFs include 46 TF families, each with ≤ 9 members. The most reliable dosage-insensitive genes were defined as the common dosage-insensitive genes of the four datasets used in Table 1-4.

Table 7 Supports of the enrichments of transcription factor genes among dosage-(in)sensitive genes

	Shihab et al. 2017	Lek et al. 2016	Makino et al. 2013	Rice et al. 2017	MRDS	MRDIS
Nuclear receptor	DS***	DS***	DS***	DS*	DS***	DS***
All TF families <10 members together	DS***	DS***	DS***	DS*	DS***	DS***
All TF genes as a whole	DS***	DS***	DS***	NS	DS***	DS***
bHLH	DS**	DS*	DS**	NS	DS***	DS**
E2F	DS**	DS*	DS***	DS*	NS	DS**
Rel	DS***	DS***	DS***	NS	DS**	DS**
SMAD	DS***	DS**	DS***	NS	DS*	DS*
CENPB	NS	NS	DIS*	DIS*	NS	DS*
C2H2-ZF	DIS*	NS	DS*	DIS***	NS	NS

Sources of the dataset: haploinsufficient genes detected by Lek et al. (Lek et al., 2016) and predicted by Shihab et al. (Shihab et al., 2017), ohnologs identified by Makino et al. (Makino et al., 2013), genes with conserved-copy-number in mammal evolution identified by Rice et al. (Rice and McLysaght, 2017b), the most reliable dosage-sensitive (MRDS) genes (i.e. the common dosage-sensitive genes of the above four datasets), and the most reliable dosage-insensitive (MRDIS) genes (i.e. the common dosage-insensitive genes of the above four datasets). DS: dosage sensitive; DI: dosage insensitive; NS: not significant, $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Figure 1

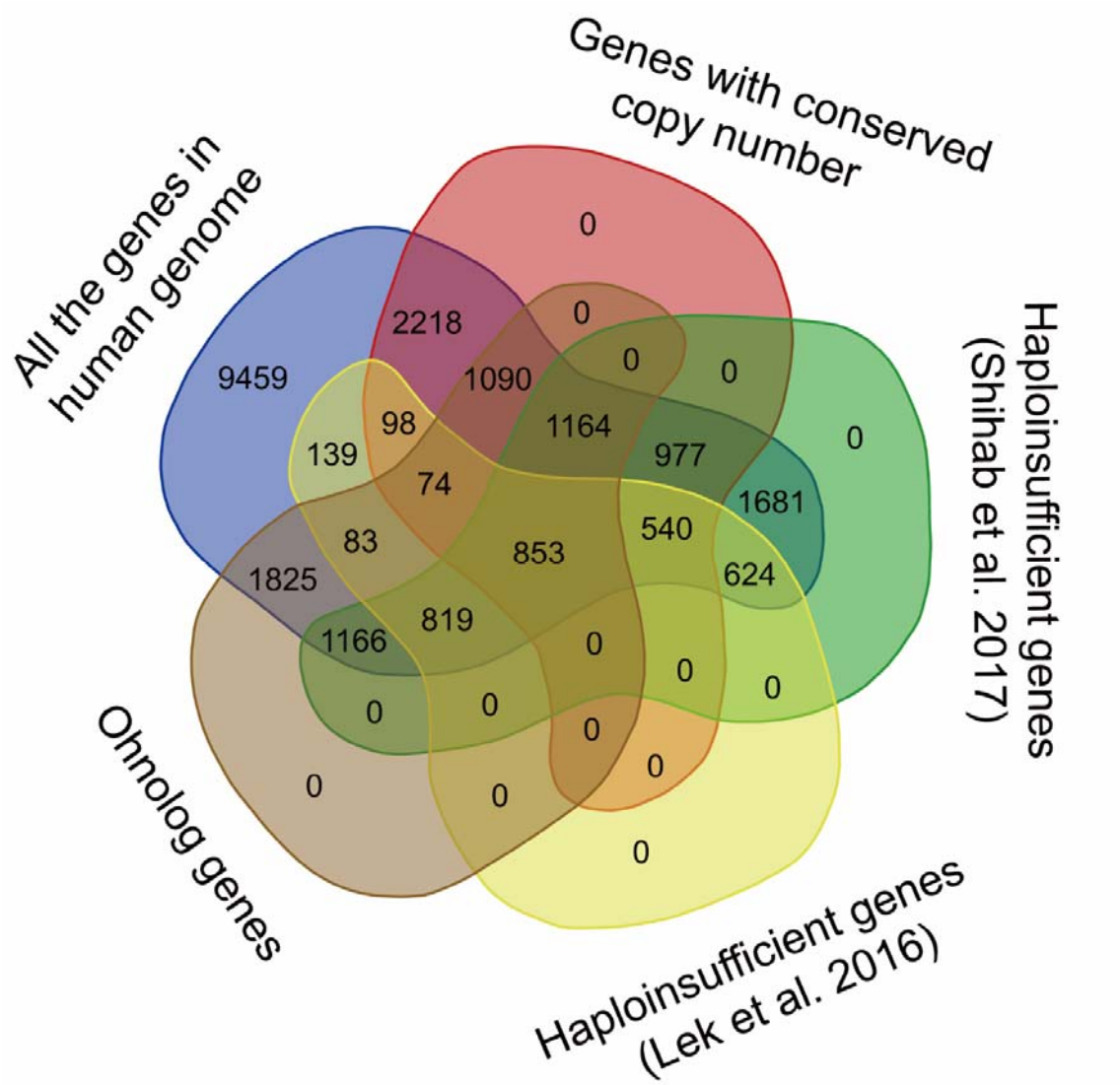

Genes with conserved copy number
All the genes in human genome
Haploinsufficient genes (Shihab et al. 2017)
Ohnolog genes
Haploinsufficient genes (Lek et al. 2016)
0
2218
0
0
0
9459
1090
98
1164
977
0
139
74
1681
83
540
853
624
1825
819
1166
0
0
0
0
0
0
0
0
0
0

Figure 2

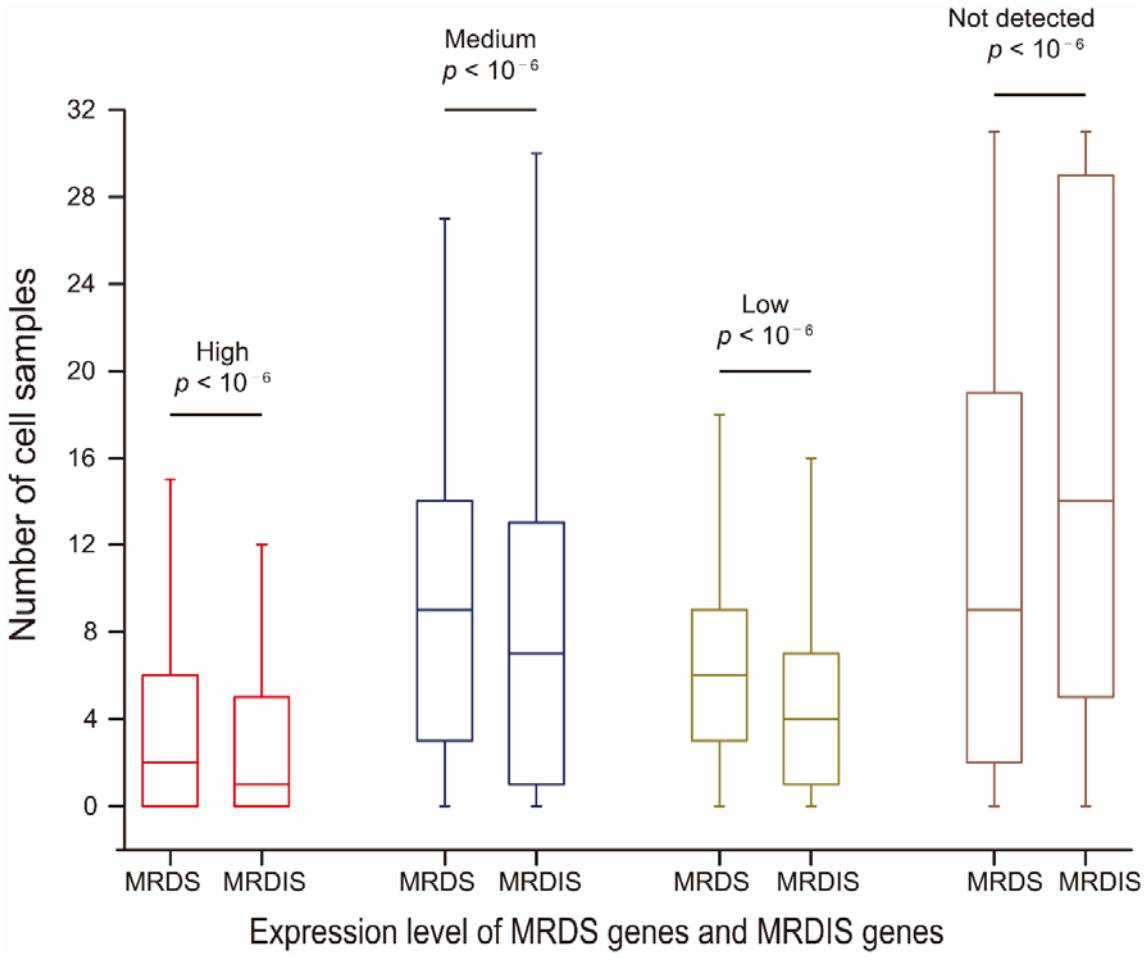
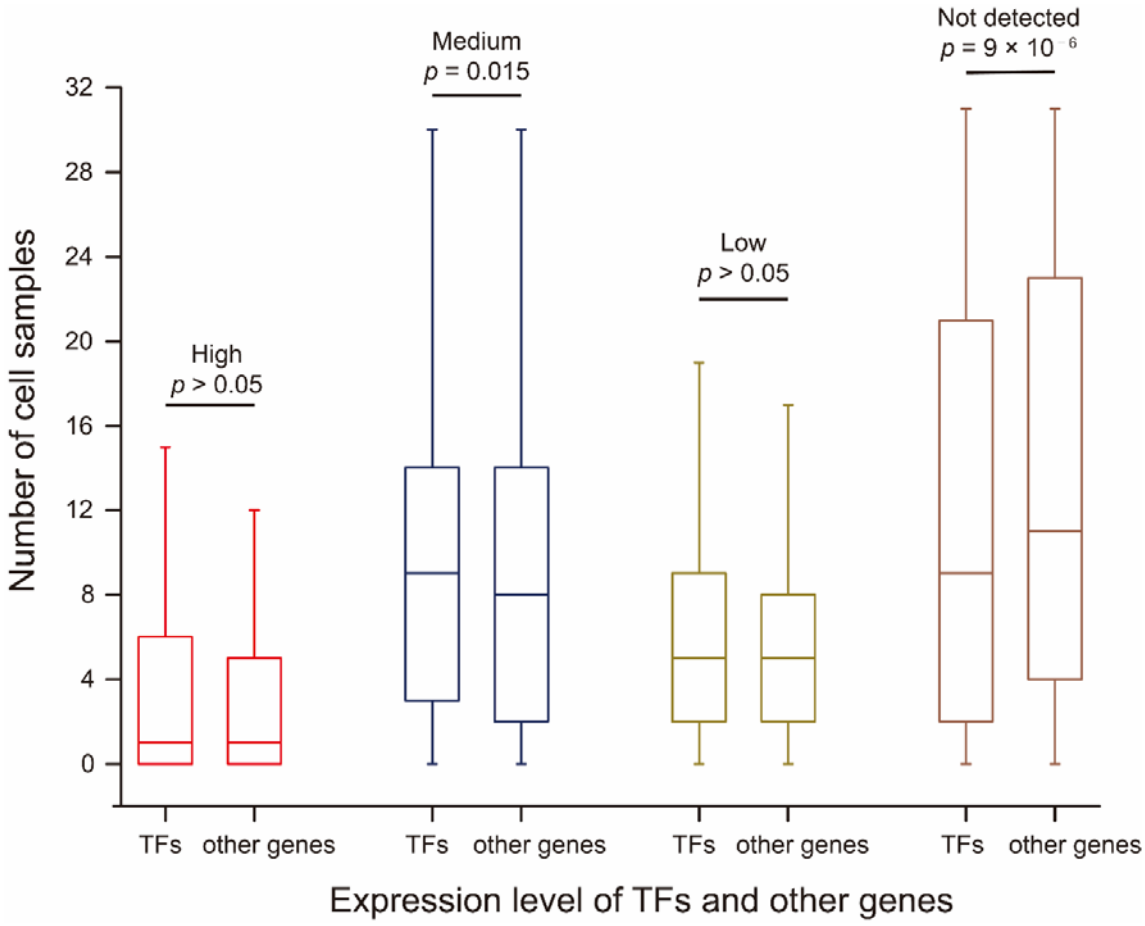

Medium
$p < 10^{-6}$
Not detected
$p < 10^{-6}$
High
$p < 10^{-6}$
Low
$p < 10^{-6}$
Number of cell samples
32
28
24
20
16
12
8
4
0
MRDS
MRDIS
MRDS
MRDIS
MRDS
MRDIS
MRDS
MRDIS
Expression level of MRDS genes and MRDIS genes

Figure 3


Medium
p = 0.015
Not detected
p = 9 × 10⁻⁶
Low
p > 0.05
High
p > 0.05
Number of cell samples
32
28
24
20
16
12
8
4
0
TFs
other genes
TFs
other genes
TFs
other genes
TFs
other genes
Expression level of TFs and other genes