Genetic mechanisms for tissue-specific essential genes

Elad Dvir[1], Shahar Shohat[1] and Sagiv Shifman[1,*]

[1] Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

*Correspondence: Sagiv Shifman, Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 91904, Israel. Email: sagiv.shifman@mail.huji.ac.il, Phone: +972-2-6585396, Fax: +972-2-658697

Abstract

Genetic diseases often manifest in a specific tissue despite the genetic risk variant being present in all the cells. The most commonly assumed mechanism for this tissue-specificity is the selective expression of the causal gene in the pathogenic tissue, but other possible mechanisms are less explored. Here, using genome-scale CRISPR screens and expression data from approximately 800 cancer cell lines, we identified 1274 tissue-specific essential genes (TSEGs) and show that only a minority of these genes can be explained by preferential expression. To elucidate the mechanisms for the TSEGs we analyzed their association with gene expression. The expression of 115 TSEGs was consistent with preferential expression in the vulnerable tissues. Counterintuitively, for 509 TSEGs we observe a decrease in expression in the vulnerable tissues, which is mainly explained by the increase in gene copy number. Analyzing the expression level of all other genes in the genome identified three other suggested mechanisms for TSEGs. First, based on the expression of paralogs, we observed reduced functional redundancy for 273 TSEGs out of the 741 with known paralogs. Second, in another 118 TSEGs, we found a surprisingly significantly higher expression of their paralogs in the vulnerable tissues, implying functional codependency. Finally, using gene set enrichment analyses, we provide evidence for the remaining 466 TSEGs that the transfer of small molecules to mutant cells by non-mutant cells could explain blood-specific essentiality. Our findings indicate that reliance only on the preferential expression of disease genes may bias views of the disease-relevant tissues.

Introduction

Hereditary diseases, including neurodegenerative disorders, cardiovascular diseases, and autoimmune diseases often affect a specific tissue although the genetic risk variants are present across the human body [1–3]. One explanation for the tissue specificity of diseases is the preferential expression, or even exclusive expression, of the causal genes in the specific tissue [1,4,5]. For example, the caveolin 3 (CAV3) gene is selectively expressed in cardiac myocytes, skeletal muscle, and smooth muscle cells, and its mutation causes cardiomyopathies and skeletal muscle disorders [6]. Indeed, a large-scale analysis that systematically explored the correlations between pathological manifestations of diseases, the expression patterns of more than 2000 implicated genes, and 1500 protein complexes found that disease genes and complexes tend to be overexpressed in tissues where defects cause pathology [4].

However, many other casual genes are expressed widely throughout the body and do not show tissue-specific expression. Alternative explanations for tissue-specificity of casual genes could be interactions with other genes [1,2,7,8]. In a recent large-scale analysis that used gene and protein expression data to map the interactomes of 16 human tissues and their relation to hereditary diseases it was found that although disease causal genes are expressed in many tissues, they tend to have an increasing number of tissue-specific protein interactions in the affected tissue [2]. Similarly, It has also been proposed that tissue specificity of genes could be explained by the expression of a functional subnetwork of genes in the relevant tissue [9].

Tissue-specificity of diseases can be explained by genetic interactions between the disease genes and other genes that are expressed in specific tissues, even if the two proteins do not physically interact. One such genetic interaction is when the effect of a pathogenic genetic variant can be modified by the expression of another gene to result in either the suppression or aggravation of the effect. Suppression in specific tissues can be caused by functional redundancy between genes that share similar functions [1,7,8,10]. In such cases, a gene with a similar function to the casual gene might be lowly expressed in a specific tissue which makes the tissue more vulnerable to mutations in the causal gene. Paralogs could provide such redundancy due to their functional overlap and ability to compensate for mutations [7,10,11]. A study of 112 hereditable diseases found

that in 20% there was a lower expression of paralogs in the disease-related tissue [7]. However, it was demonstrated in yeast that in some cases the expression of paralogs increases the deleterious effect of a mutation due to functional codependency of the paralogs [12].

Previous studies in humans focused mainly on tissue specificity in the context of hereditable monogenic diseases [2–5,7]. A systematic analysis of the effect of mutations across the genome on different tissues was not previously performed due to a lack of data. The development of CRISPR-Cas9 based approaches for genome-wide loss-of-function (LoF) screening made it possible to identify essential genes in large panels of cell lines [13,14]. Project Achilles used a large panel of human cancer cell lines that represent multiple cancer tissues to create a catalog of essential genes [15]. The fitness consequences of disrupting a specific gene are estimated by comparing the abundances of single-guide RNAs (sgRNAs) targeting the gene at the start of the experiment to the abundances of the same sgRNAs after a few weeks of growth. A relative decrease in the abundance of sgRNAs suggests that disruption of the target gene causes a viability defect. The magnitude of the impact on fitness can be inferred from the magnitude of change in abundance and is represented by a dependency score. This resource makes it possible not only to systematically investigate which genes are essential for cancer cells, but also enables the discovery of the factors that contribute to the variation in essentiality. Using these resources, we can identify tissue-specific essential genes (TSEGs) by grouping the cell lines based on the lineage of the cells.

In the current study, we utilized the large-scale CRISPR screens to identify TSEGs and study the mechanisms of tissue specificity. We identified in total 1701 TSEGs in 23 tissues and showed their relevance to human phenotypic abnormalities. Some of the TSEGs were specific to more than one tissue, therefore resulting in 1274 unique genes. Using gene expression from the same cells we identified which TSEGs are best explained by preferential expression, functional redundancy, or functional codependency. Additionally, we found two other mechanisms for TSEGs that may also be influenced by the methods used in the screen. For some genes, an increase in copy number is associated with both an increase in gene expression and a decrease in essentiality, probably due to insufficient disruption of all the copies of the target gene. For the genes essential specifically in blood cells we suggest based on gene set enrichment analyses that intracellular

communication with an exchange of small metabolites between cells (a mechanism that is absent in blood cells) can explain the large number of TSEGs identified for blood-related cells.

Results

Inferring tissue-specific essential genes from large-scale CRISPR screens

To comprehensively catalog genes that are essential for specific tissues we utilized data from large-scale CRISPR screens in cancer cell lines for the identification of genes required for growth or viability (789 cancer cell lines originated from 27 different tissues). For each gene in each cell line, the dataset contains a dependency score, such that a very negative dependency score implies that the cell line is highly dependent on that gene (scores are adjusted such as a score of 0 is equivalent to a gene that is under no selection and the median score of the highly essential genes is a score of -1). Tissue-specific essential genes (TSEGs) were defined as genes with significantly lower levels of dependency scores in a specific tissue compared to all other tissues (FDR < 0.05) and with an average dependency score < -0.5. Using this method, we were able to identify genes that show tissue-specificity even if it is in more than one tissue.

We identified in total 1701 TSEGs in 23 tissues which included 1274 unique genes (Table S1). The number of genes per tissue varied, with the highest numbers for plasma and blood (Figure 1A). The number of cell lines per tissue is unlikely to explain this variation, as there was no significant correlation between sample size and the number of genes identified per tissue ($r = -0.048$, $P = 0.83$) (Figure 1B). It is important to note that the dependency scores of TSEG were continuous across tissues for some genes (Figure 1C), while for other genes there was a clear distinction between tissues with low and high dependency scores (Figure 1D). Since some of the TSEGs were shared between tissues, we quantified the percentage of overlap between tissues (Figure 1E). The highest overlap was found between related tissues, such as blood-plasma, plasma-lymphocytes, blood-lymphocytes, ovary-uterus and pancreas-bile duct (Jaccard Index = 0.186, 0.095, 0.093, 0.091, 0.066 respectively).

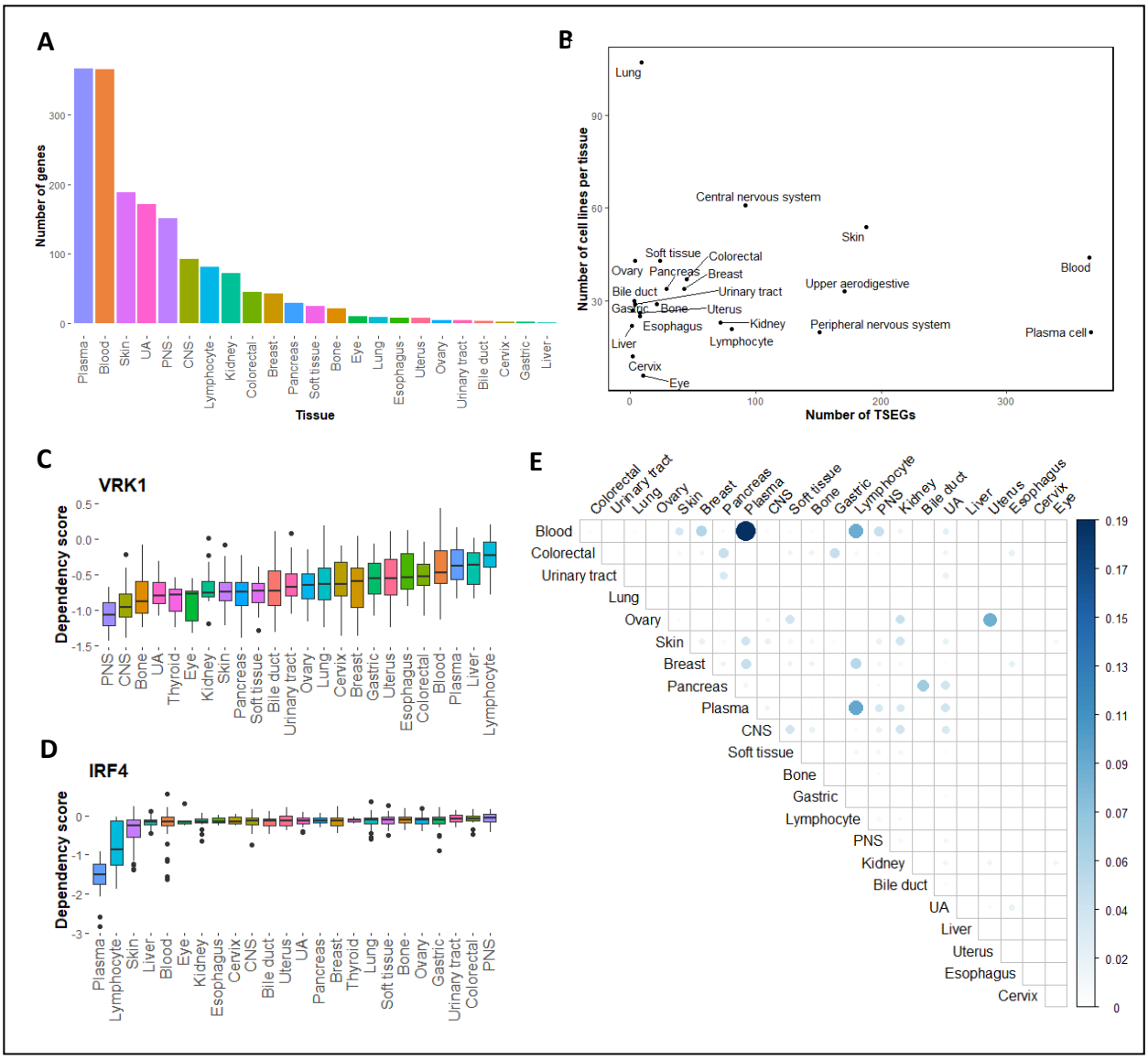

Figure 1. Identification and characterization of TSEGs. (A) The number TSEGs per tissue. (B) Number of TSEGs per tissue as a function of sample size (number per tissue). (C-D) Examples of distribution of dependency scores across tissues for two TSEGs. (C) VRK1, essential in the central nervous system and the peripheral nervous system, and (D) IRF4, essential in the plasma and the lymphocyte tissues. (E) The overlap of TSEGs between tissues. The overlap was quantified by a metric of intersection over union, also known as the Jaccard index. Abbreviations: CNS – *central nervous system*, PNS – *peripheral nervous system, UA – upper aerodigestive.*

Association of tissue-specific essential genes with relevant phenotypic abnormalities

The TSEGs were identified based on cancer cell lines, but we hypothesized that in many cases the results could also be relevant to normal tissues and explain tissue specificity of human diseases.

We performed an enrichment analysis of human phenotypes (Table S2) and found that TSEGs as a group are most enriched with phenotypes of neurodevelopmental disorders, such as decreased head circumference (*FDR* = 2.230e-13), microcephaly (*FDR* = 2.230e-13), aplasia/hypoplasia involving the central nervous system (*FDR* = 9.369e-11), and in neoplasms of the pancreas (*FDR* = 4.290e-09) and the large intestine (*FDR* = 1.977e-06). These enrichments are consistent with the role of cell proliferation in both cancer and neurodevelopmental disorders [16,17].

Next, we explored whether the TSEGs in each separate tissue are associated with phenotypic abnormalities in matching human body systems from the Human Phenotype Ontology database (Table S3) using gene-set enrichment analysis (GSEA) (Figure 2). For the majority of tissues (13/23), the TSEGs were not enriched in any of the terms. For the remaining tissues, the significant terms commonly included related body systems. For example, TSEGs of blood were most significantly enriched with abnormalities of the blood and blood-forming tissues, and TSEGs of lymphocytes were most significantly enriched with abnormalities of the immune system. For some tissues, the related term was not necessarily the most significant one. For example, abnormalities in the nervous system ranked second in association with TSEGs of the central nervous system, but significant enrichment was also found for many other non-related body systems.

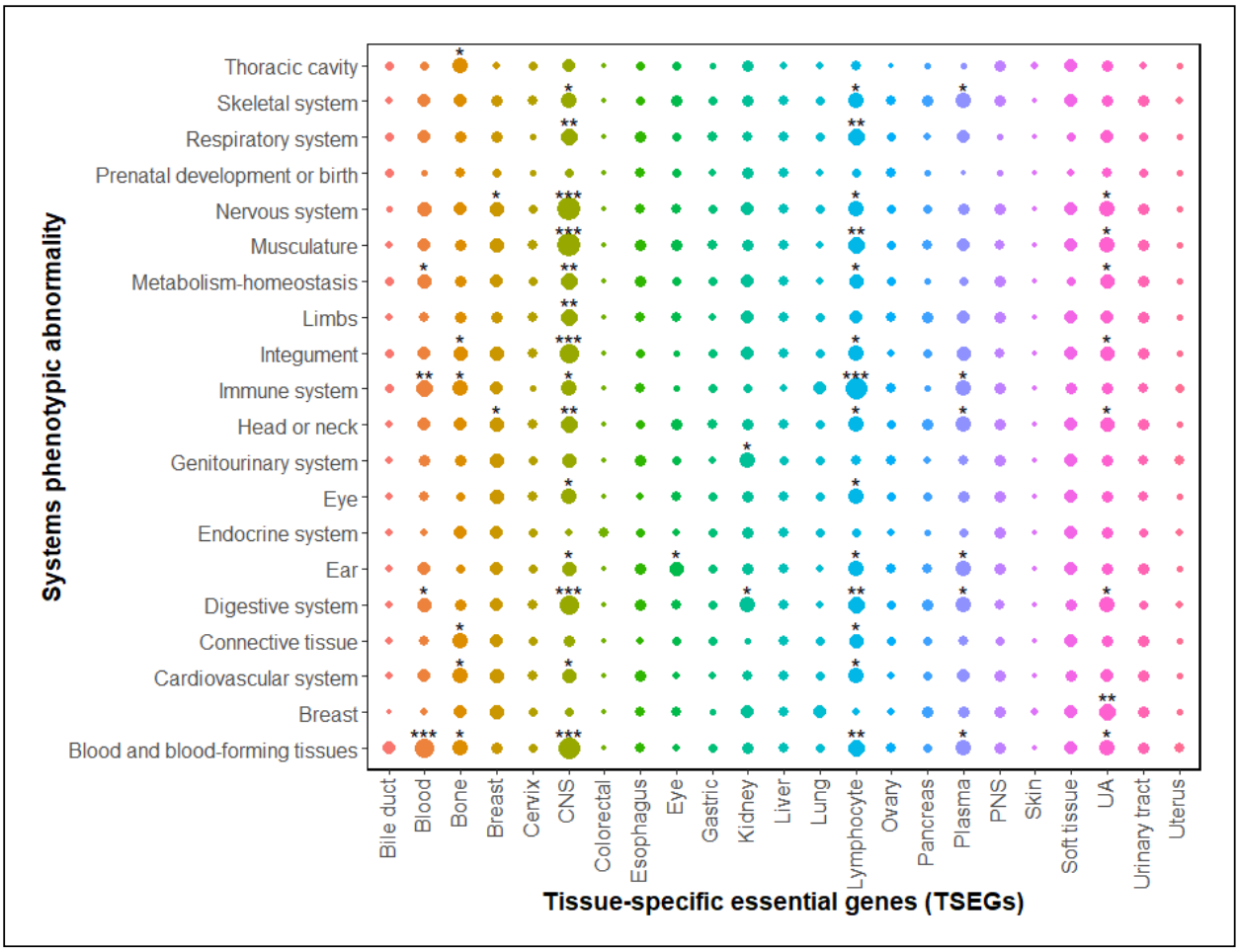

Figure 2. Enrichment of TSEGs with relevant phenotypic abnormalities. The enrichment was performed using GSEA with 20 phenotypic abnormalities from the Human Phenotype Ontology database. * *FDR* < 0.05; ** *FDR* < 0.005; *** *FDR* < 0.001. Circle sizes represent -log of the FDR q value. Abbreviations: CNS – *central nervous system*, PNS – *peripheral nervous system, UA – upper aerodigestive.*

Tissue-specific processes, including preferential expression and copy number variations, can explain TSEGs

The simplest explanation of TSEGs is a preferential expression (or even exclusive expression) of the essential genes in the susceptible tissue. To explore this possibility, we used the expression data available for the same cell lines to test for differential expression of 19,144 genes between the tissues with low dependency scores (hereafter the vulnerable tissues) and all other tissues (hereafter the non-vulnerable tissues). While in many cases the average expression of TSEGs in the vulnerable tissues was considerably higher compared to the non-vulnerable tissues,

surprisingly the expression of many other TSEGs was lower in the vulnerable tissues (Figure 3A). The tendency for more extreme expression values (lower or higher) in the vulnerable tissues was significantly different than expected by chance based on a randomization test (P = 9.9X10^{-5}) (Figure 3B).

To estimate how many of the TSEGs are likely to be explained by preferential expression, we tested the correlation between the dependency scores of each of the TSEGs and the expression levels of all available genes. For the genes that are explained by a preferential expression, we expected that the dependency scores of the TSEG will be negatively correlated with the self-expression of the TSEG (higher expression of the TSEG in cell lines with a lower score that indicates its higher essentiality). Out of the 1274 TSEGs, 624 (49%) had a significant correlation between the dependency and their expression (FDR < 0.05), however, only 115 of them were in the direction consistent with preferential expression (Spearman's ρ < 0, FDR < 0.05) (see examples in Figure 3C-D).

The remaining 509 TSEGs with a positive correlation (Spearman's ρ > 0, FDR < 0.05) are genes that are less essential in cell lines with higher expression (see example in Figure 3E). One possible explanation for this unexpected finding is the presence of copy number variations (CNVs) that may both increase the gene expression and also decrease the likelihood of a complete knockout of the gene by CRISPR. Indeed, we found that that for 469 genes out of the 509 TSEGs (92.1%), there was a significant (FDR < 0.05) positive correlation between the copy number and the dependency score (see examples in Figure 3E-G). This was not evident for TSEGs with a negative correlation, as only 5 out of 115 (4.3%) showed a significant correlation between the dependency and the copy number. Overall, the correlation direction (negative vs. positive) was highly significantly associated with CNVs as predictors (odds ratio (OR) = 250.7, P = 2.7×10^{-80}) (Figure 3H).

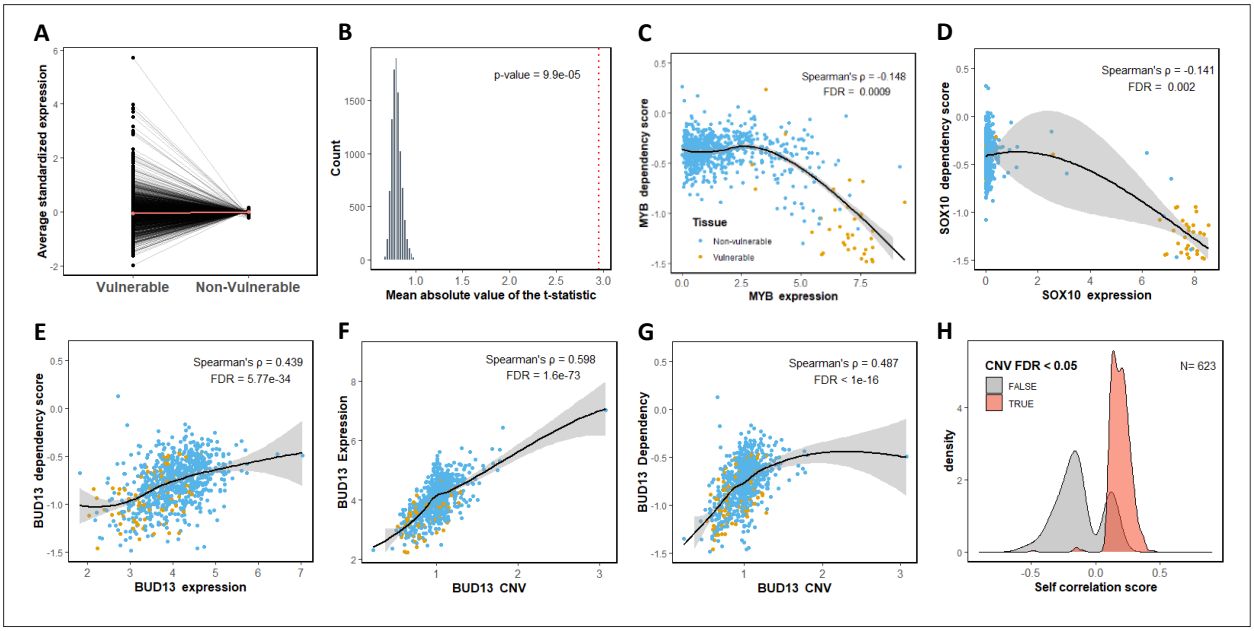

Figure 3. The increase and decrease expression of the TSEGs in vulnerable tissues is consistent with a preferential expression mechanism and the effect of CNVs. (A) The average standardized expression of TSEGs in vulnerable and non-vulnerable tissues. (B) The observed mean absolute t-statistic (dashed red line) was calculated for differences in TSEGs expression between the vulnerable and non-vulnerable tissues. The histogram shows the distribution of the expected values under the null based on 10,000 permutations. (C-D) Demonstration of preferential expression. (C) *MYB*, a TSEG in blood and lymphocytes, is more essential (lower dependency) in cells with higher expression of *MYB.* (D) *SOX10*, a TSEG in the skin, is more essential in cells with higher expression of *SOX10*. (E-G) Demonstration of TSEG that is explained by the gene copy number. (E) *BUD13*, a TSEG in the skin and the upper aerodigestive tissue is less essential in cells with higher expression of *BUD13*. (F) *BUD13* expression is associated with the copy number of *BUD13*. (G) *BUD13* dependency scores are associated with the copy number of *BUD13*. (H) Density plots of Spearman's correlation coefficients between the dependency scores and TSEGs expression. The plots are shown for 623 TSEGs with CNV information, divided into genes with significant correlation with the gene copy number (light red) or non-significant (gray). All *P*-values presented in the figures are adjusted for multiple tests by the Benjamini and Hochberg false discovery rate (FDR) procedure. The yellow dots (C-G plots) represent cell lines that belong to the vulnerable tissues and blue dots are for cell lines in the non-vulnerable tissues. In C-G, locally estimated scatterplot smoothing (LOESS) curves and 95% confidence intervals are shown alongside Spearman's ρ and the FDR value.

Tissue-specific genetic interaction of paralogs may explain a large proportion of TSEGs

Another possible mechanism for TSEGs is tissue-specific functional redundancy. For example, lower expression in the vulnerable tissues of paralogs that can compensate for the mutations may explain TSEGs. To study the expression of paralogs as a potential mechanism, we identified TSEGs with paralogs (741 genes, Table S4) and compared the expression of the paralogs in the

vulnerable vs. non-vulnerable tissues. Similar to the self-expression analysis above, we observed that the average expression of the paralogs tends to be either higher or lower in the vulnerable tissues relative to the non-vulnerable tissues (Figure 4A). The absolute differences in the expression of paralogs between the vulnerable and non-vulnerable tissues were significantly different than expected by chance based on a randomization test (P = 9.9X10^{-5}) (Figure 4B). We tested the correlation between the dependency scores and the expression of the paralogs and found that out of the 655 genes with paralogs, 273 had a significant positive correlation in the predicted direction of reduced functional redundancy (Spearman's $\rho > 0$, FDR < 0.05) (see example in Figure 4C-F). This suggests that for 273 TSEGs, the decreased expression of the paralog in the vulnerable tissues is a likely mechanistic explanation. We then tested how many of the 273 TSEGs and their paralogs show also the opposite effect of a significant correlation between the dependency of the paralog and the expression of the TSEGs (symmetric compensation). Out of 261 paralogs with available dependency scores, only 35 showed symmetric compensation (Table S4). An additional 118 TSEGs had a significant negative correlation between the dependency scores and the expression of their paralogs (Spearman's $\rho < 0$, FDR < 0.05) (see example in Figure 4G), suggesting that a higher expression of the paralogs, instead of compensating for the mutations, makes the cells more vulnerable to mutations (functional codependency).

For some TSEGs, we had data regarding the expression of multiple paralogs. We observed that when the most significant correlation with one paralog was negative (functional codependency) significant correlations with the other paralogs were in the same direction. However, when the most significant correlation was positive (functional redundancy), other paralogs could have both negative and positive significant correlations, although in most cases (204 out of 273) the correlations were in the same direction (Table S4). Functional enrichment analysis showed that TSEGs implicated in functional codependency were enriched with transcription factors while TSEGs implicated in functional redundancy were enriched with general essential functions (Table S5).

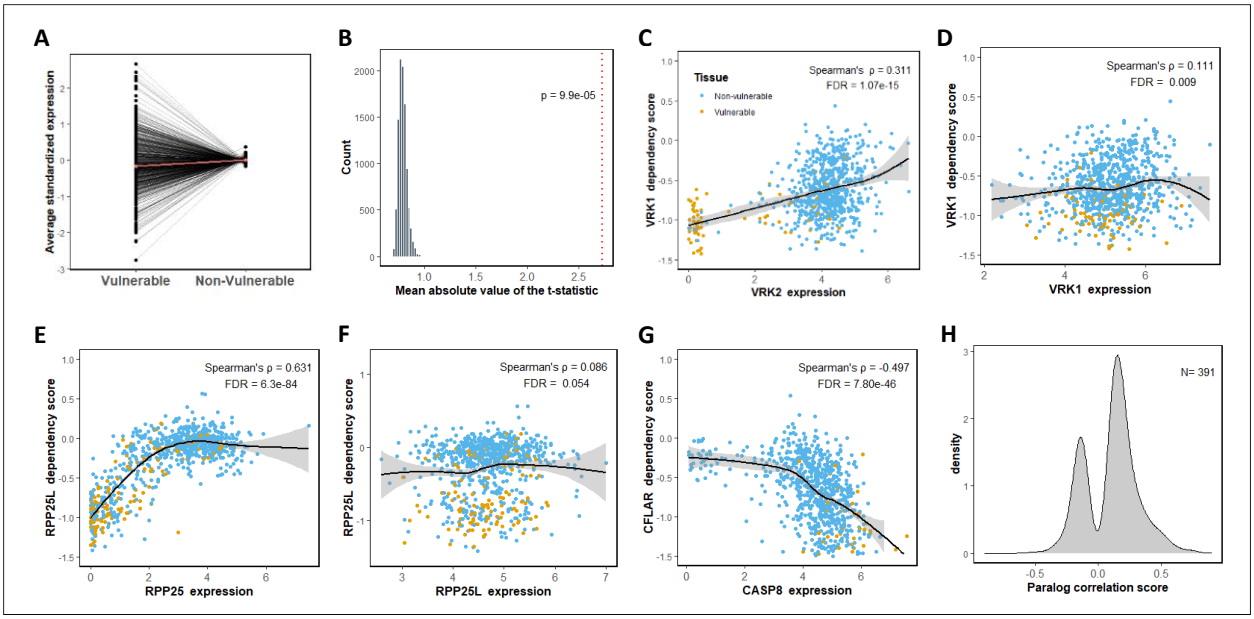

Figure 4. Analysis of paralogs expression suggests redundancy and codependency mechanisms for TSEGs. (A) The average standardized expression of TSEGs paralogs in vulnerable and non-vulnerable tissues. (B) The observed mean absolute t-statistic (dashed red line) was calculated for differences in paralog expression between the vulnerable and non-vulnerable tissues. The histogram shows the distribution of the expected values under the null based on 10,000 permutations. (C-F) Demonstration of functional redundancy. (C-D) The dependency of *VRK1*, a TSEG in the central and peripheral nervous systems, is correlated mainly with (C) the expression of its paralog, *VRK2,* (D) relative to the expression of *VRK1* itself. (E-F) The dependency of RPP25L, a TSEG in the central nervous systems, soft tissues, and lymphocytes, is significantly correlated (E) with the expression of its paralog, *RPP26*, but not (F) with the expression of RPP25L itself. (G) Demonstration of functional codependency. *CFLAR*, a TSEG in plasma and kidney, is more essential in cells with high expression of its paralog, *CASP8*. (H) Density plots of Spearman's correlation coefficients between the TSEGs dependency scores and the paralog expression. For TSEGs with more than one paralog, the graph presents the correlation coefficient for the most significant one. All *P*-values presented in the figures are adjusted for multiple tests by the Benjamini and Hochberg false discovery rate (FDR) procedure. The yellow dots (C-G plots) represent cell lines that belong to the vulnerable tissues and blue dots are for cell lines in the non-vulnerable tissues. In C-G, locally estimated scatterplot smoothing (LOESS) curves and 95% confidence intervals are shown alongside Spearman's ρ and the FDR value.

Intercellular communication may explain TSEGs particularly in blood cells

Our analysis shows that preferential expression, CNVs, functional redundancy, or functional dependence may explain 63.42% of the TSEGs (n = 808). We next turned to find alternative explanations for the remaining TSEGs (n = 466). To do so, we calculated the correlation between the dependency scores of the 466 TSEGs and gene expression of all other genes. We selected the five most significant genes with a positive correlation to the dependency scores of each of the

466 remaining TSEGs. These 2330 genes were enriched for gene ontology (GO) terms related to intercellular communication, such as cell adhesion (FDR = 3.03×10^{-12}), membrane raft (FDR = 3.9×10^{-6}), and negative regulation of cell communication (FDR = 1.8×10^{-6}) (Table S6). We noted that some of the genes were significantly correlated with multiple TSEGs (Figure 5A). We performed GSEA with the genes ranked by the number of times their expression was correlated with the dependency scores of TSEGs and found similar enrichment for GO terms related to the interaction between cells, such as cell junction assembly (FDR = 0.008), plasma membrane region (FDR = 0.01) and divalent inorganic cation transport (FDR = 0.01) (Figure 5B-D). This enrichment analysis suggests that the top-ranked genes have a direct role in intercellular communication. At the top of the list was *GJA1* (also known as connexin 43), which was correlated with 27 TSEGs. *GJA1* is a component of gap-junctions that allow the exchange of low molecular weight molecules between adjacent cells [18]. The second-ranked gene (correlated with 12 TSEGs), *CALD1,* is involved in various actin cytoskeleton-related cellular processes including endocytosis and exocytosis [19]. These results could be interpreted as follows: mutant cells that express proteins that allow uptake through gap-junctions or endocytosis from surrounding cells will be less dependent on the ability to generate small essential molecules. Consistent with this hypothesis, enrichment analysis for the 27 genes correlated with *GJA1* showed that these genes are involved in the metabolism of small molecules, including ribonucleoside monophosphate biosynthetic process (FDR = 1.3×10^{-19}), nucleoside monophosphate biosynthetic process (FDR = 1.8×10^{-18}), carbohydrate derivative metabolic process (FDR = 3.1×10^{-18}) and nucleobase-containing small molecule metabolism (FDR = 1.4×10^{-17}) (Table S6). Since in circulating blood cells gap-junctions are absent, we expected that the TSEGs that correlated with *GJA1* to be dominated by blood cells. Indeed, we found that the majority of genes (22 out 27 genes) that are significantly positively correlated with GJA1 were specifically essential in blood-related tissues (blood, lymphocytes, and plasma cells), which was significantly more than expected (OR = 4.61, P = 0.0011). We also suspected that such a mechanism might explain why we had a higher amount of TSEGs in tissues that lack interactions and junctions between cells, such as plasma or blood cells. To test this, we removed all TSEGs with a significant positive correlation with any of the 23 genes belonging to the cell junction assembly group and examined the change in the proportion of TSEGs per tissue.

The largest negative decline in representation was for blood tissue (-10%), lymphocytes (-2.4%), and plasma cells (-1.6%) (Figure 5E).

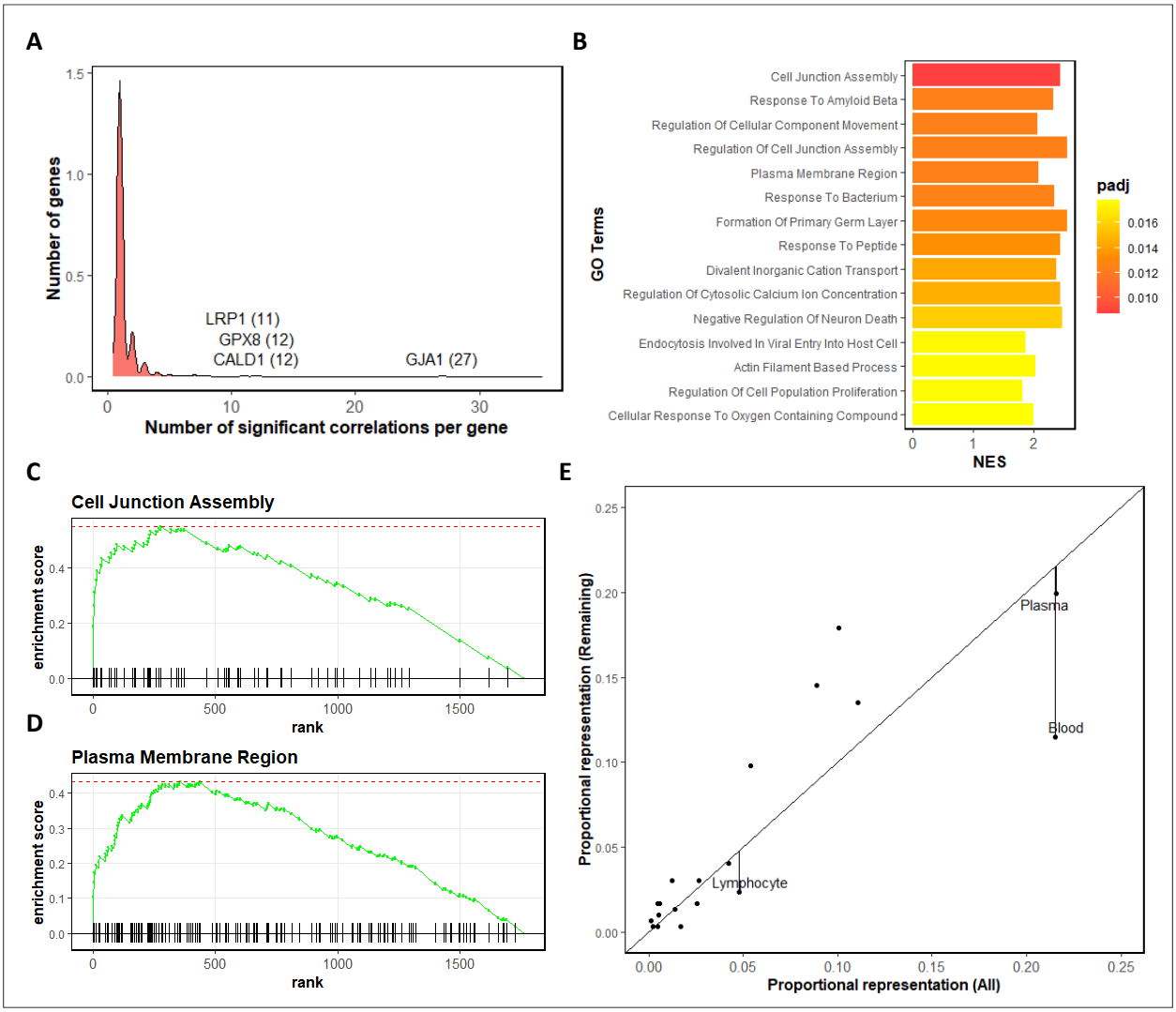

Figure 5. Cell to cell communication as a possible mechanism for TSEGs. (A) Distribution of the number of TSEGs that were significantly and positively correlated with the expression of a specific gene. The analysis was performed for 466 TSEGs not explained by other processes. The names of four highly correlated genes are shown. (B) Gene ontology (GO) terms identified by gene-set enrichment analysis (GSEA) for association with the genes correlated with the unexplained TSEGs. The dependency of the 466 unexplained TSEGs was tested for correlation with gene expression across the genome. For each gene in the genome with expression data, the number of significant positive correlations with TSEGs was recorded and was used to rank the genes for the GSEA. The colors correspond to the FDR corrected *P*-values. NES-normalized enrichment score. (C-D) Enrichment plots for the two most significant terms: (C) cell junction assembly and (D) plasma membrane region. (E) The proportional representation of TSEGs in each tissue is plotted as a function of the proportion after removing 248 TSEGs significantly correlated with the expression of any of the 23 genes that are involved in cell junction organization. The three tissues with the largest drop in representation are blood cells, plasma cells, and lymphocytes.

Discussion

We utilized data from whole-genome CRISPR screens in cancer cell lines to identify genes that are significantly more essential in specific tissues. We identified different tissue-specific mechanisms that could explain the dependency pattern of those genes. Only in 9% of cases, the dependency was explained by preferential expression in vulnerable tissues. These results are consistent with previous studies that showed that although some disease genes tend to have elevated expression in the relevant tissue, this mechanism alone is not sufficient to explain the tissue specificity of diseases [1,4,5].

In three of the proposed mechanisms, tissue-specific compensation is implicated. The first mechanism of compensation involves the gene itself (intra-allelic interaction). The compensation originates from an increased copy number of the gene, which is more likely to occur in specific tissues. We suggest that this lowers the probability that all copies will be disrupted by mutations. In our study, gene duplication explained a big proportion of TSEGs (37%). This type of mechanism may be relevant particularly to duplications that frequently occur in specific types of cancer but is unlikely to explain the tissue specificity of hereditary diseases. The second mechanism involves compensation by another gene (non-allelic interactions) that results in reduced functional redundancy in specific tissues. This was observed in 21% of the TSEGs (42% of genes with paralogs), but since we studied only genes with known paralogs, it is likely to be more widespread. The third mechanism is compensation between cells. Our analysis suggests that some TSEGs may arise from variation in intercellular communication that allows the transfer of small molecules to mutant cells only in specific tissues. Since in most cases the genetic risk variants are present in all cells, this mechanism is relevant only in the context of genetic mosaicism when there are genetically distinct cell populations within the same individual (e.g. X-inactivation) [20–22].

In addition, we found that 9% of TSEGs (18% of genes with paralogs) have an opposite correlation between their dependency scores and paralog expression, meaning that these genes are essential in tissues where the paralog has elevated expression. Thus, the suggested mechanism is functional codependency. Little is known regarding this mechanism, its extent in humans, and its relevance to heritable diseases. Previous works suggested that functional codependency could

stem from a direct interaction between proteins that form heterodimers, and that cells in which the paralogs are expressed are more dependent on the complex and therefore more vulnerable to the mutation [12,23]. We found that the TSEGs implicated in functional codependency are mainly involved in transcription regulation. The codependency of those transcription factors may stem from being in the same complex or having common target genes.

Our findings have important implications for studies of genotype-phenotype relationships. Studies of both monogenic and complex diseases are frequently based on the assumption that the susceptibility genes are preferentially expressed in the affected tissue. This type of analysis is done to connect newly identified variants to a disease or as a way to identify the tissues and developmental stages that are associated with the disease. For example, in autism spectrum disorders, where the underlying brain regions and circuits are not fully understood, the analysis of genes disrupted by *de novo* mutations showed preferential expression during development in several specific brain regions[24–26]. However, our findings indicate the need for ways to integrate both the expression of the causal genes and other interacting genes to determine which tissues or cell types are likely to be involved in the disease. Since our work is based on cancer cell lines it may have specific implications for understanding the origin of tissue specificity in response to oncogenic driver mutations. However, due to the same reason, these findings do not necessarily indicate mechanisms for diseases in adult tissues and are more likely to reflect mechanisms for developmental diseases in which cells tend to highly proliferate, in similarity to cancer cells. Indeed, the association of TSEGs also with developmental disorders suggests that the findings are relevant to a large set of developmental disorders. Additionally, studies have demonstrated that for some genes, the RNA expression is not correlated with protein expression [27–29]. Such a difference between RNA and protein expression could pose another limitation to our work since we based the search for mechanisms on RNA expression results.

An additional and more general output of our study is that tissue specificity, like other forms of biological variation, can offer important insights into genetic mechanisms of genetic interactions. The same mechanisms that can explain differences in essentiality between tissues can also explain the differences between individuals through modifier genes and can be the basis for incomplete penetrance and variable expressivity.

Methods

Inferring tissue-specific essential genes (TSEGs)

Differential dependency analysis was performed using the Limma package in R [30]. The dependency scores were obtained from the DepMap portal (https://depmap.org/portal/) (DepMap Public 20Q3, dataset doi:10.6084/m9.figshare.12931238.v1). The dataset contains scores for 18,119 genes in 789 cancer cell lines originating from 27 different lineages. Each gene from the dataset was fitted to 27 linear models, corresponding to the 27 different tissues. Each model tested for the tendency of a gene to have higher or lower dependency scores in a specific type of tissue. Moderated t-statistics, moderated F-statistic, log-odds of the differential dependency, and p-values were computed using the eBayes function from this package. The p-values were adjusted for multiple tests by a false discovery rate (FDR) Benjamini-Hochberg procedure. TSEGs were defined as genes with significantly lower levels of dependency scores in a specific tissue compared to all other tissues ($t < 0$, FDR < 0.05) and with an average dependency score < -0.5. TSEGs in tissues with a sample size < 5 cell lines were excluded from the analysis. Pearson correlation test was used to test the association between the number of cell lines and the number of TSEGs per tissue. The percentage of overlap between tissues was quantified by the Jaccard Index (metric of intersection over union).

Average expression estimation in vulnerable and non-vulnerable tissues

To test whether the expression of TSEGs in the vulnerable tissues tended to have more extreme values both in the positive and negative directions, we performed a randomized permutation test. We calculated for all TSEGs the mean of the absolute values of t-statistics between the expression in cell lines that belongs to the vulnerable tissues and other cell lines. This value served as the test statistic and was compared to 10,000 simulations (n) in which the tissue identity of TSEGs was randomized. We calculated the empirical P-value as (r+1)/(n+1), where r is the number of times the statistic in the simulations exceeded the true statistic[31]. Similarly, to check whether the expression of the paralogs tends to have more extreme values (both in the positive and negative directions), we calculated for all TSEGs the mean of the absolute values of

the t-statistics between the expression of the paralog in cell lines that belongs to the vulnerable tissues and other cell lines. This value served as a test statistic and was compared to 10,000 simulations in which the tissue identity of TSEGs was randomized. The empirical P-value was calculated as before. Human paralogs were identified using the BioMart web interface at the Ensembl genome browser (https://www.ensembl.org) [32].

Correlation tests between dependency scores, gene expression, and copy number

Spearman's rank correlation tests were used for studying the relationship between the dependency scores of the TSEGs and the expression of 19,144 genes. Gene expression available for 783 cell lines was obtained from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) expression dataset. The FDR procedure was used to adjust the *P*-values for each TSEG and across the 19,144 tests. The relationship between the copy number of each gene (obtained from the DepMap portal) and the gene's dependency scores was tested using Spearman's rank correlation test with FDR adjustment. Fisher's exact test was used to test for the association between the number of TSEGs significantly associated with copy number and the direction (positive or negative) of the correlation between the expression of the TSEG and the dependency scores. Functional redundancy and functional dependency with paralogs were identified based on a significant (FDR < 0.05) positive or negative Spearman's rank correlation between the dependency scores of the TSEGs and the expression of existing paralogs.

Gene set enrichment analyses (GSEA)

Enrichment analysis of human phenotypes associated with all TSEGs was performed using ToppGene [33]. Gene Set Enrichment Analysis (GSEA) [34] with the fgsea package in R [35] was used to test for the enrichment of human phenotypic abnormalities with TSEGs separately for each tissue. For this purpose, 20 gene sets of human phenotypic abnormalities were composed using data obtained from the Human Phenotype Ontology (HPO) database (Table S3). Genes that overlapped with more than 10 phenotypic abnormalities were excluded from the analysis. The

ranked gene list for each tissue was based on -log(P) of all 18,119 genes in each tissue as was calculated using the differential dependency pattern analysis. Enrichment analysis of biological process and molecular functions for TSEGs explained by redundancy and separately by codependency was performed using ToppGene. To identify possible common explanatory variables of TSEGs not explained by other suggested mechanisms (preferential expression, CNVs, redundancy, codependency) we selected the five genes with expression most significantly positively correlated with the dependency of the TSEGs, and then ranked those correlated genes based on the number of appearances. Next, we performed GSEA on this list with GO terms obtained from the Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org). Additional enrichment analysis for biological process and molecular functions was done on the unranked list of those genes using ToppGene. A similar analysis was also performed for TSEGs that were significantly correlated with the most common possible explanatory, the GJA1 gene.

References

1. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nature Reviews Genetics* **21**, 137–150 (2020).

2. Barshir, R., Shwartz, O., Smoly, I. Y. & Yeger-Lotem, E. Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. *PLoS Comput. Biol.* **10**, (2014).

3. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).

4. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20870–20875 (2008).

5. Reverter, A., Ingham, A. & Dalrymple, B. P. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Min.* **1**, (2008).

6. Song, K. S. *et al.* Expression of caveolin-3 in skeletal, cardiac, and smooth muscle cells: Caveolin-3 is a component of the sarcolemma and co-fractionates with dystrophin and dystrophin-associated glycoproteins. *J. Biol. Chem.* **271**, 15160–15165 (1996).

7. Barshir, R. *et al.* Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. *PLoS Genet.* **14**, (2018).

8. Papp, B., Pál, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661–664 (2004).

9. Kitsak, M. *et al.* Tissue Specificity of Human Disease Module. *Sci. Rep.* **6**, (2016).

10. Diss, G., Ascencio, D., Deluna, A. & Landry, C. R. Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **322**, 488–499 (2014).

11. Kafri, R., Levy, M. & Pilpel, Y. The regulatory utilization of genetics redundancy through responsive backup circuits. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11653–11658 (2006).

12. Diss, G. *et al.* Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science (80-.).* **355**, 630–634 (2017).

13. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* **339**, 819–823 (2013).

14. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science (80-.).* **350**, 1096–1101 (2015).

15. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

16. Shohat, S. & Shifman, S. Genes essential for embryonic stem cells are associated with neurodevelopmental disorders. *Genome Res.* **29**, 1910–1918 (2019).

17. Ernst, C. Proliferation and Differentiation Deficits are a Major Convergence Point for Neurodevelopmental Disorders. *Trends in Neurosciences* **39**, 290–299 (2016).

18. Goodenough, D. A., Goliger, J. A. & Paul, D. L. Connexins, connexons, and intercellular communication. *Annual Review of Biochemistry* **65**, 475–502 (1996).

19. Sobue, K., Kanda, K., Tanaka, T. & Ueki, N. Caldesmon: A common actin-linked regulatory protein in the smooth muscle and nonmuscle contractile system. *Journal of Cellular Biochemistry* **37**, 317–325 (1988).

20. Migeon, B. R. X-linked diseases: susceptible females. *Genetics in Medicine* **22**, 1156–1174 (2020).

21. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease-clones picking up speed. *Nature Reviews Genetics* **18**, 128–142 (2017).

22. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nature Reviews Genetics* **14**, 307–320 (2013).

23. Dandage, R. & Landry, C. R. Paralog dependency indirectly affects the robustness of

human cells. *bioRxiv* (2019). doi:10.1101/552208

24. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–21 (2013).

25. Shohat, S., Ben-David, E. & Shifman, S. Varying Intolerance of Gene Pathways to Mutational Classes Explain Genetic Convergence across Neuropsychiatric Disorders. *Cell Rep.* **18**, 2217–2227 (2017).

26. Ben-David, E. & Shifman, S. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Mol. Psychiatry* **18**, 1054–6 (2013).

27. Shankavaram, U. T. *et al.* Transcript and protein expression profiles of the NCI-60 cancer cell panel: An integromic microarray study. *Mol. Cancer Ther.* (2007). doi:10.1158/1535-7163.MCT-06-0650

28. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Mol. Cell. Biol.* (1999). doi:10.1128/mcb.19.3.1720

29. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* (2009). doi:10.1186/1471-2164-10-365

30. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (2005). doi:10.1007/0-387-29362-0_23

31. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical P values from Monte Carlo procedures [1]. *American Journal of Human Genetics* **71**, 439–441 (2002).

32. Smedley, D. *et al.* BioMart - Biological queries made easy. *BMC Genomics* **10**, (2009).

33. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305-11 (2009).

34. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

35. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* 060012 (2016).