

## RESEARCH ARTICLE

# Identification of candidate disease genes in patients with common variable immunodeficiency

Guojun Liu<sup>1,\*</sup>, Mikhail A. Bolkov<sup>2,3</sup>, Irina A. Tuzankina<sup>2,3</sup>, Irina G. Danilova<sup>1,2</sup>

<sup>1</sup> Department of Medical Biochemistry and Biophysics, Institute of Natural Sciences and Mathematics, Ural Federal University, Ekaterinburg 620000, Russia

<sup>2</sup> Institute of Immunology and Physiology, Ural Branch of the Russian Academy of Sciences, Ekaterinburg 620000, Russia

<sup>3</sup> Department of immunochemistry, Institute of Chemical Engineering, Ural Federal University, Ekaterinburg 620000, Russia

\* Correspondence: gjliu0325@gmail.com

Received November 23, 2018; Revised December 29, 2018; Accepted January 29, 2019

**Background:** Common variable immunodeficiency (CVID), the most prevalent form of primary immunodeficiency (PID), is characterized by hypogammaglobulinemia and recurrent infections. Understanding protein-protein interaction (PPI) networks of CVID genes and identifying candidate CVID genes are critical steps in facilitating the early diagnosis of CVID. Here, the aim was to investigate PPI networks of CVID genes and identify candidate CVID genes using computation techniques.

**Methods:** Network density and biological distance were used to study PPI data for CVID and PID genes obtained from the STRING database. Gene expression data of patients with CVID were obtained from the Gene Expression Omnibus, and then Pearson's correlation coefficient, a PPI database, and Kyoto Encyclopedia of Genes and Genomes were used to identify candidate CVID genes. We then evaluated our predictions and identified differentially expressed CVID genes.

**Results:** The majority of CVID genes are characterized by a high network density and small biological distance, whereas most PID genes are characterized by a low network density and large biological distance, indicating that CVID genes are more functionally similar to each other and closely interact with one other compared with PID genes. Subsequently, we identified 172 CVID candidate genes that have similar biological functions to known CVID genes, and eight genes were recently reported as CVID-related genes. MYC, a candidate gene, was down-regulated in CVID duodenal biopsies, but up-regulated in blood samples compared with levels in healthy controls.

**Conclusion:** Our findings will aid in a better understanding of the complex of CVID genes, possibly further facilitating the early diagnosis of CVID.

**Keywords:** common variable immunodeficiency; primary immunodeficiency; candidate CVID genes; protein-protein interactions; network density; biological distance

**Author summary:** Like many human diseases, common variable immunodeficiency (CVID) is multigenic, resulting from mutations in multiple genes that affect the same or diverse phenotypes. In fact, multigenic causes of CVID have been found in approximately 90% of cases. However, the genetic causing of CVID is still largely unclear. The multigenic traits of CVID are possibly the results of complex interactions between genes. The protein-protein interaction network-based view can, therefore, provide a deeper insight into CVID.

## INTRODUCTION

Common variable immunodeficiency (CVID), the most prevalent form of primary immunodeficiency (PID), is characterized by low serum levels of IgG, IgA, and IgM; deficient specific antibody responses to infection or vaccination; and exclusion of other causes of hypogammaglobulinemia; it has an estimated prevalence of 1:50,000 to 1:25,000 [1]. CVID shows substantial phenotype and genotype heterogeneity, and the majority of CVID cases have an unknown genetic cause [2]. The monogenic defects that have been implicated in CVID include the following: (i) recessively inherited mutations in *IL21*, *IL21R*, *LRBA*, *ICOS*, *PRKCD*, *CD19*, *CD20*, *CD21*, *CD27*, *CD81*, and *RAC2*; (ii) dominantly inherited mutations in *CTLA4*, *TNFSF12*, *NFKB1*, *PLCG2*, *NFKB2*, *PIK3CD*, *PIK3R1*, *VAV1*, *BLK*, *IKZF1*, and *IRF2BP2*; and (iii) monoallelic or biallelic mutations in *TNFRSF13B* and *TNFRSF13C* [3].

For example, the B cell co-receptor complex is composed of CD19, CD21, CD81, and CD225, which together lower the threshold for B cell activation following antigen binding to the B cell receptor, and it has been recently reported that *CD19*, *CD81*, and *CD21* deficiencies occur in autosomal recessive forms of CVID [4]. *BLK* plays an important role in BCR signaling and the recruitment of T cell help, and a heterozygous loss-of-function mutation in *BLK* was previously detected in two related patients with CVID [5]. *NFKB1* encodes the mature p52 subunit and its precursor p105 subunit, and it belongs to the NF- $\kappa$ B transcription factor family and has been associated with CVID in multiple consanguine families or sporadic cases [6,7]. Loss-of-function variants in *TNFRSF13B* might aggravate the effect of already impaired Toll-like receptor (TLR) signaling or impose TLR signaling defects [8], and they have also been detected in many patients with CVID [3]. However, not all CVID causes involve monogenic defects, as monogenic causes of CVID have been found in only approximately 10% of cases. Conversely, there are numerous examples of multigenic (or polygenic) causes of CVID, where variants in multiple genes can contribute to the same or diverse phenotypes [9–11]. Although CVID is thought to result from genetic defects, the exact cause of the disorder is unknown in the large majority of cases.

Great progress has been made in approaches used to identify PID disease genes. In the early years, Keerthikumar *et al.* used a support vector machine to classify all human genes as PID genes or non-PID genes. The underlying principle of this classification was to calculate the confidence score for each candidate PID gene based on the 69 features observed in the 148 known PID genes [12]. Recently, a research team identified PID candidate

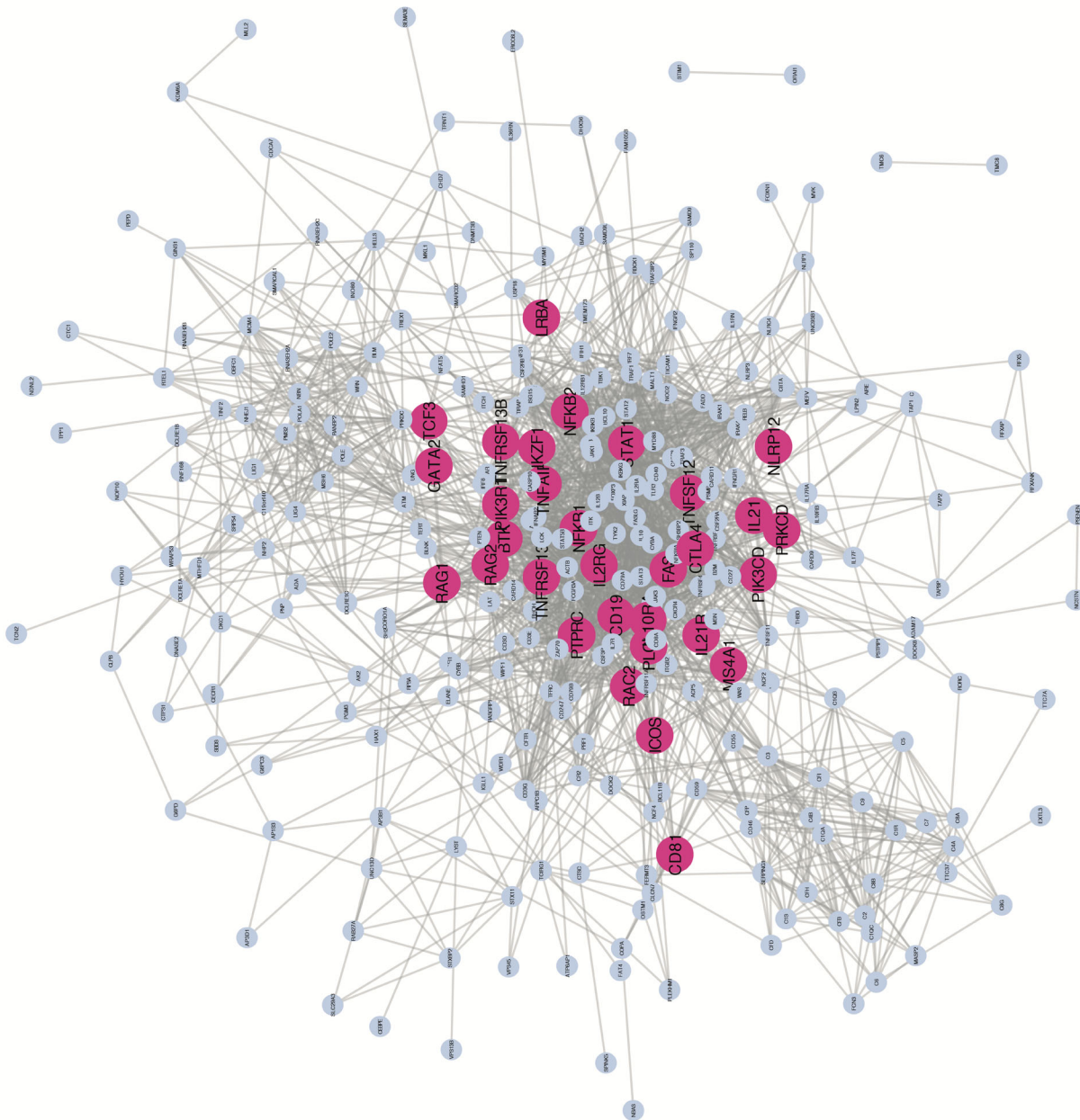
genes by analyzing the properties of protein-protein interaction (PPI) networks of all essential human immune tissue genes and their gene ontology (GO) terminology [13]. Other studies predicted 3,110 candidate PID genes using an *in silico* computational approach based on biological distance [14,15]. Importantly, they found that PID genes, compared with other human genes, tend to be in the central hub of the human genome network and to more closely interact with each other. Moreover, PID genes form several tightly intra-related sub-clusters, and most of them have at least one other PID gene as a close functional neighbor among a wide array of biological mechanisms [14]. Understanding this mechanism might provide additional insight into the diversity of genetic pathways underlying PID, which in turn might facilitate the development of new drugs and therapeutic approaches. Although the mechanism of PID has been adequately studied and several *in silico* methods have been developed to identify candidate PID genes, to our knowledge, there have been no reports on either identifying CVID candidate genes or revealing the biological network features of CVID genes. In particular, the underlying mechanisms of the PPI networks of CVID genes remain unclear and lack a systematic level of interpretation. With the recent accumulation of novel CVID genes [3,16], reliable molecular interactions [17], and the state-of-the-art computational techniques [18–20], there is an urgent need to predict candidate CVID-specific genes using systems biology and bioinformatics methods to help the rapid and accurate discovery of new CVID-related genes.

In this study, by calculating the network density and biological distance of CVID genes and PID genes, we found that CVID genes are more similar in function and interact more closely compared with most PID genes. We also present a comprehensive approach for predicting candidate CVID genes. Finally, we identified differentially expressed genes in the duodenal biopsy and blood samples from patients with CVID.

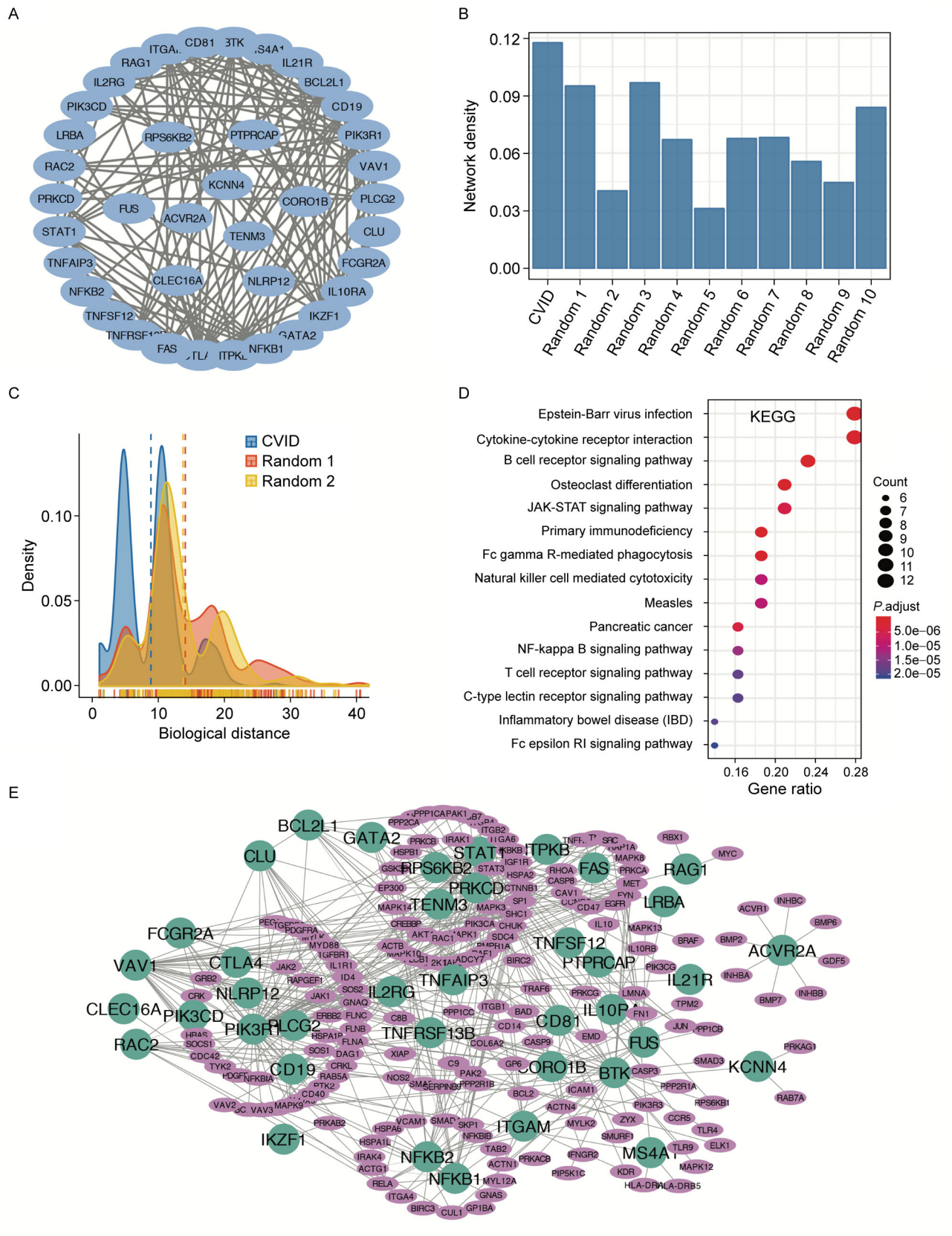
## RESULTS

### Exploration of the PPIs of CVID genes

The PPI data for PID genes were retrieved from the STRING database and visualized using Cytoscape software (Figure 1), and the original PPI network generated using the STRING database is provided in Supplementary Figure S1. We found that CVID genes have a central tendency in the network, suggesting that CVID genes might interact more frequently than PID genes. To further study the phenomenon of the complex interactions of CVID genes, the PPI data of a CVID group composed of 39 CVID genes and that of ten random groups (each



**Figure 1. Protein-protein interaction (PPI) network of the PID genes.** The PPI data for PID genes obtained from the STRING database were visualized using Cytoscape software. Red nodes represent CVID genes, and light steel blue nodes represent other PID genes.



**Figure 2. PPI network of CVID genes, network densities and biological distances of the CVID group and random groups, KEGG enrichments of CVID genes, and PPI network of candidate CVID genes and known CVID genes.** (A) PPI network of CVID genes. (B) Barplot of the network densities of a CVID group and ten random groups. (C) Density plot of the biological distances of a CVID group and two random groups. (D) Bubble plot of the KEGG enrichment results of CVID genes. The size of the displayed bubble corresponds to the count of the gene, and the color of the bubble corresponds to the adjusted *P*-value. (E) The PPI network of candidate CVID genes (orchid) and known CVID genes (dark slate gray).

group consists of 39 PID genes) were obtained from STRING. The PPI network of the CVID group is shown in Figure 2A and those of the random groups are shown in Supplementary Figure S2. Then, network density ( $D_{\text{network}}$ ) was used to measure and compare the cohesion and tightness of the PPI networks of each group (the greater the network density is for a group, the closer the interaction of the genes within the group). The results revealed that the CVID group possessed a higher network density than the ten random groups, indicating that CVID genes more closely interact with each other (Figure 2B). In addition, we utilized HGC to calculate the biological distance of a CVID group and two random groups (each group consists of 39 PID genes) and compared the density distribution of the biological distance. The smaller the biological distance is for a group, the closer the functional relevance between the genes in the group. The results revealed that the density distribution of the CVID group had a median value of 8.8, while the median values of the random 1 and random 2 groups were around 14, indicating a tighter functional interrelatedness between the CVID genes (Figure 2C). Altogether, these observations lead to the generalization that CVID genes more closely interact with each other and have a closer biological interrelatedness with one another than PID genes.

### Identification of candidate CVID genes

The first and second filtration steps mentioned in “Materials And Methods” yielded 2,751 CVID-specific interactions, including 1,716 candidate genes. We performed a KEGG analysis of known CVID genes and found that a total of 15 KEGG pathways (e.g., Epstein-Barr virus infection, cytokine-cytokine receptor interaction, and B cell receptor signaling pathway) were statistically significant ( $P < 0.05$ , Figure 2D). To identify candidate genes that are functionally similar to known CVID genes, we further screened for certain candidate genes enriched in at least one of the above-mentioned 15 KEGG pathways, resulting in 414 CVID-specific interactions comprising 172 novel CVID candidate genes (Supplementary File 1 and Supplementary materials S1). The resulting PPI network for known CVID genes and CVID candidate genes is shown in Figure 2E.

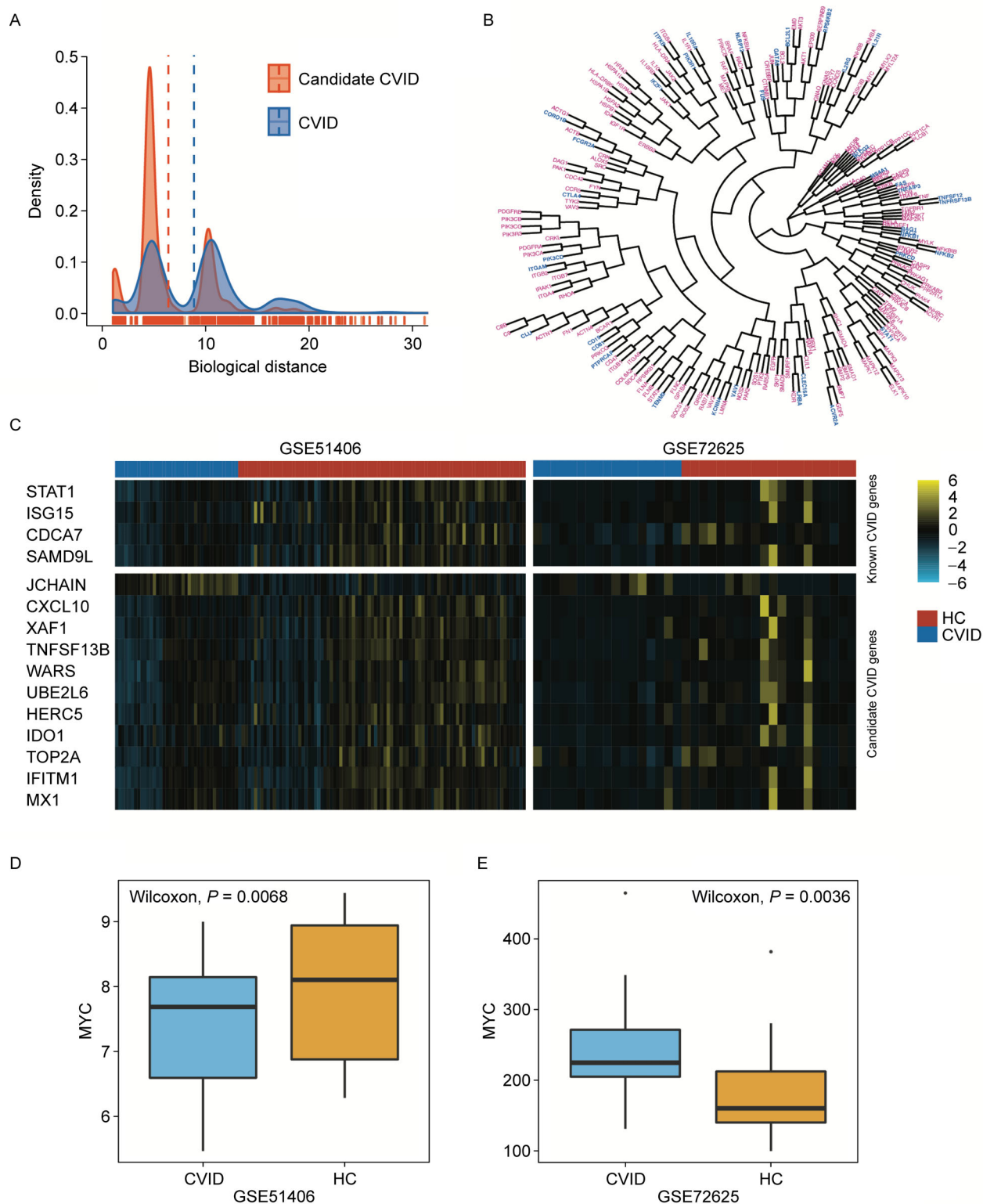
### Evaluation of candidate CVID genes

To evaluate our predictions, we calculated the biological distance of 172 CVID candidate genes and compared it to that of 39 known CVID genes. As a result, the median biological distance of candidate CVID genes was 6.01, which was smaller than (or similar to) that of known CVID genes, indicating a strong biological association

between CVID candidate genes similar to that between CVID genes (Figure 3A). Then, the CVID candidate genes were mixed up with CVID genes, and the biological distance was assessed again for the mixed genes; the results were shown in Supplementary File 2. FGA phylogenetic analysis was subsequently performed on the mixed genes to assess the biological correlation between CVID genes and CVID candidate genes. The results showed that candidate CVID genes were evenly distributed over the entire range of known CVID genes, suggesting that these CVID candidate genes were closely related to known CVID genes (Figure 4C). We also evaluated our predictions by reviewing some studies. Notably, eight genes (*AKT1*, *AKT3*, *RELA*, *SOC31*, *STAT3*, *XIAP*, *CD40*, and *CASP8*) not included in our original list of CVID genes were identified as CVID candidate genes and have been shown to cause or affect CVID in experiments, demonstrating the importance of the CVID candidate genes (Table 1).

### Identification of differentially expressed genes (DEGs) between patients with CVID and healthy controls (HCs)

In the GSE72625 dataset, compared with the levels in HCs, 464 genes were significantly differentially expressed in the duodenal biopsy of patients with CVID, with 275 genes up-regulated and 189 genes down-regulated. In the GSE51406 dataset, 489 DEGs were differentially expressed in the blood of patients with CVID compared with the levels in HCs, with 348 genes up-regulated and 141 genes down-regulated (Supplementary Figure S3A and S3B). Moreover, 31 genes were observed to be up-regulated both in the blood and duodenal biopsy of patients with CVID compared with the levels in HCs, whereas seven overlapping genes were observed to be down-regulated both in the blood and duodenal biopsy of patients with CVID compared with the levels in HCs (Supplementary Figure S3C and S3D, and Supplementary materials S2). Importantly, several well-known PID genes (e.g., *STAT1*, *ISG15*, *CDCA7*, and *SAMD9L*) and previously reported candidate PID genes (e.g., *JCHAIN*, *CXCL10*, *XAF1*, *TNFSF13B*, *WARS*, *UBE2L6*, *HERC5*, *IDO1*, *TOP2A*, *IFITM1*, *MX1*) were significantly differentially expressed [7,9]. With the exception of *JCHAIN*, all of the PID-related genes mentioned above were up-regulated in both the blood and the duodenal biopsy of patients with CVID compared with the levels in HCs (Figure 3C). *MYC*, a CVID candidate gene found both in this study and a previous study, was down-regulated in the duodenal biopsy of patients with CVID compared with the level in HCs, whereas it was up-regulated in the blood of patients with CVID compared with the level in HCs (Figure 3D).



**Figure 3. Biological distances and functional genomic alignment (FGA) of candidate CVID genes and CVID genes, heatmap plot of differentially expressed known PID genes and candidate PID genes, and boxplot of MYC in the GSE51406 and GSE72625 datasets.** (A) Density plot of biological distances of known CVID genes and predicted CVID candidate genes. (B) Phylogenetic tree of biological distances generated by FGA, showing the hierarchical clustering of all known CVID genes (blue) and predicted CVID genes (violet-red). The length of a branch indicates the strength of separation between individuals. (C) Heatmap plot of differentially expressed known PID genes and candidate PID genes. Yellow indicates high expression, and turquoise indicates low expression. (D) Boxplot of the comparison of the expression level of MYC between patients with CVID and HCs in the GSE51406 dataset. (E) Boxplot of the comparison of the expression level of MYC between patients with CVID and HCs in the GSE72625 dataset.

**Table 1** List of CVID candidate genes with a recently reported association with CVID

Gene symbol	Description	Aliases	Ref.
<i>AKT1</i>	AKT serine/threonine kinase 1	<i>AKT, CWS6, PKB, PKB-ALPHA, PRKBA, RAC, RAC-ALPHA</i>	[21], PMID: 27664934
<i>AKT3</i>	AKT serine/threonine kinase 3	<i>MPPH, MPPH2, PKB-GAMMA, PKBG, PRKBG, RAC-PK-gamma, RAC-gamma, STK-2</i>	[22], PMID: 26081581
<i>RELA</i>	RELA proto-oncogene, NF- $\kappa$ B subunit	<i>NFKB3, P65</i>	[23], PMID: 27461466
<i>SOCS1</i>	Suppressor of cytokine signaling 1	<i>CIS1, CISH1, JAB, SOCS-1, SSI-1, SS1, TIP-3, TIP3</i>	[24], PMID: 29618830
<i>STAT3</i>	Signal transducer and activator of transcription 3	<i>ADMIO, ADMIO1, APRF, HIES</i>	[53], PMID: 29180260 [25], PMID: 26360251 [26], PMID: 27379089
<i>XIAP</i>	X-linked inhibitor of apoptosis	<i>API3, BIRC4, IAP-3, ILP1, MIHA, XLP2, hIAP-3, hIAP3</i>	[27], PMID: 27492372
<i>CD40</i>	CD40 molecule	<i>Bp50, CDW40, TNFRSF5, P50</i>	[28], PMID: 23305827 [29], PMID: 30464201 [30], PMID: 28756897
<i>CASP8</i>	Caspase 8	<i>ALPS2B, CAP4, CASP-8, FLICE, MACH, MCH5</i>	[30], PMID: 28756897

## DISCUSSION

CVID, a type of PID, presents a profound heterogeneity in both phenotype and genotype, with monogenic or complex causes. Our research sheds light on the underlying molecular mechanism of the complicated CVID genes, that is, CVID genes have a closer biological interrelatedness and tighter interactions with each other than most PID genes. In addition, we identified 172 new CVID candidate genes that interact with known CVID genes in the same biological pathway and show a high biological correlation with known CVID genes.

Many experimental studies have confirmed that the eight genes in our obtained CVID gene list are true CVID-causing genes or CVID-related genes (Table 1). Akt mediates biological functions such as differentiation, survival, cell cycle, metabolism, cytokine production, growth, and activation of B cells through direct downstream molecules such as FoxO, GSK-3, Tpl-2, and TSC2 [21]. Our approach identified two isoforms of *Akt* (AKT1 and AKT3) as CVID candidate genes. AKT1 is a key signaling protein in the cellular pathways involved in skeletal muscle hypertrophy and general tissue growth. A previous study found impaired phosphorylated AKT1 expression that was significantly correlated with antibody response to a vaccine and worse clinical complications in B cells of patients with CVID [21]. AKT3, the major effector downstream of the PI3K signaling pathway, plays a key role in peripheral B cell maturation and survival, and it was found to be hypermethylated in B cells when comparing a twin with CVID with his healthy sibling [22]. RELA is initially located in the cytoplasm, and it can translocate to the nucleus after anti-IgM stimulation in

CD21<sup>+</sup> B cells of HCs, whereas nuclear translocation of RELA is strongly induced in CD21<sup>+</sup> and CD21<sup>low</sup> B cells of patients with CVID. In addition, the phosphorylation of RELA was, surprisingly, found to be significantly lower in CD21<sup>+</sup> and CD21<sup>low</sup> B cells of patients with CVID compared with that of HCs [23]. An analogous study observed that the transcriptional level of SOCS1 was up-regulated after activation of TLR4 and TLR9 in patients with CVID and HCs [24]. Several other studies have suggested that mutations in the coiled-coil domain of STAT3 can promote the transcriptional activity of STAT3, which is pathogenic for CVID [25,26]. Likewise, others have reported that increased tyrosine phosphorylation of STAT3 was detected in the memory B cell population of patients with CVID, which was associated with elevated apoptotic rates in these cells [25]. Granulomatous lymphocytic interstitial lung disease (GLILD), as a discrete histopathological entity, was first described as a manifestation of XIAP deficiency. Moreover, one study speculated the possibility of XIAP deficiency in severe CVID because GLILD has been most frequently described as a complication of CVID [27]. CASP8 mutations not only impair adequate TCR and TLR signaling, but also skew the immune responses towards a pro-inflammatory pattern [31]. CASP8 has also been found to be mutated in many patients with autoimmune lymphoproliferative syndrome (ALPS), which may be associated with CVID because their similar clinical and immunological diagnoses [32].

Some genes exhibit drastic differences in expression between HCs and patients with CVID. For example, the expression level of CD40 and ICOS was significantly lower in the B cells and T cells of patients with CVID than

in those of HCs [28], and the expression level of CASP8 was significantly lower in CD11c<sup>+</sup> B cells of patients with CVID compared with that in cells of HCs [30]. As observed in previous research, our results also demonstrated that 38 genes are significantly differentially expressed, including four well-known PID genes and 11 previously proposed candidate PID genes [12,14]. It is important to stress again that MYC, a CVID candidate gene identified in this study and previous studies, is down-regulated in the duodenal biopsy of patients with CVID compared with the level in HCs, whereas it is up-regulated in the blood sample of patients with CVID compared with the level in HCs [12,14]. Therefore, whether the mutation of MYC leads to the alteration of its expression in patients with CVID needs to be further investigated.

The bioinformatic analysis performed in this study produced reliable results regarding the investigation of PPI networks of CVID genes and identification of CVID candidate genes. It is worth noting that we can enhance our research in following aspects in the future: (i) One goal of this study was to identify candidate CVID genes that are significantly co-expressed, functionally similar, and interact with each other. However, these features only occur in some CVID genes, and it is therefore of utmost importance to design a new algorithm or approach that can predict CVID genes that are both related and unrelated to one other. (ii) Our hypothesis of close interactions between CVID genes is based only on algorithmic methods (e.g., network density and biological distance). The recent rapid advances in sequencing technology provide an opportunity to test this hypothesis using transcriptional big data. (iii) Data cleaning and processing also result in the possibility of errors. Thus, our results need to be further validated in a large cohort of CVID samples.

## CONCLUSIONS

In conclusion, our study raises the hypothesis that CVID genes are more biologically interrelated and interact closer with each other than most PID genes, which may help physicians and researchers gain a deeper understanding of the pathophysiology of CVID. In addition, we provided a list of CVID candidate genes that represent attractive targets for testing in patients whose etiology cannot be ascribed to any known CVID gene.

## MATERIALS AND METHODS

### Data acquisition and pre-processing

The transcriptomic profiles of patients with CVID were

downloaded from the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). The datasets taken into consideration included GSE72625 (20 patients with CVID and 17 HCs, duodenal biopsy) and GSE51406 (91 patients with CVID and 39 HCs, whole blood). The pre-processing procedure for the raw microarray data consisted of deleting rows containing more than 70%–75% of missing values and log<sub>2</sub>-transforming the gene expression values. The list of 351 well-known PID genes was obtained from the European Society for Immunodeficiencies (ESID), and it is shown in Supplementary materials S3. The 39 literature-based CVID genes are shown in Supplementary materials S4.

### Evaluation of network density and biological distance for CVID and PID genes

The protein-protein interaction (PPI) data for PID and CVID genes was derived from the STRING database, including known and predicted interactions from genomic contexts, high-throughput experiments, co-expression, and previous knowledge [27]. We chose *Homo sapiens* and set the minimum required interaction score to 0.4. The network of PPI data was visualized using Cytoscape software (version 3.2) [33]. Network density ( $D_{\text{network}}$ ) is the most widely used concept in gene regulatory network and PPI network studies, and it can be used to determine whether a network is tight or cohesive. The  $D_{\text{network}}$  can be defined as [34]

$$D_{\text{network}} = \frac{\sum_{i=1}^n \sum_{j \neq i} a_{ij}}{n(n-1)} \quad (1)$$

where  $a_{ij}$  is pairwise adjacency,  $\sum_{j \neq i} a_{ij}$  represents the connectivity (the unweighted network connectivity equals the number of genes that are directly linked to gene  $i$ ) of the  $i$ -th gene, and  $n$  is the number of genes in the network. Note that  $a_{ij} = 1$  if the interaction of gene  $i$  and gene  $j$  occurs in the STRING database, whereas  $a_{ij} = 0$  otherwise. The PPI data of a CVID group (39 literature-based CVID genes) and ten random groups (each group consists of 39 PID genes) were respectively converted into the symmetric adjacency matrix ( $a_{ij}$ ,  $i, j = 1, \dots, n$ ) using the “igraph” R package [35]. Network density was used to compare their network cohesion or tightness. The greater the network density is of a group, the tighter the interaction of the genes in the group.

Biological distance ( $B_{ij}$ ) was first introduced by Itan *et al.*, and it can be used to calculate the shortest distances for all possible human gene pairs [36]. Using biological distance, a previous study found that PID genes tend to be centrally located based on the human genome network and form several tightly intra-related sub-groups across diverse biological mechanisms [14].  $B_{ij}$  is defined by

$$B_{i,j} = \begin{cases} \frac{C}{S_{i,j}} & \text{if } C = 1 \\ \frac{C}{S_{i,1} + S_{1,2} + S_{2,3} + \cdots + S_{C-2,C-1} + S_{C-1,j}} & \text{if } C > 1 \end{cases}, \quad (2)$$

where  $S_{i,j}$  is the combined score between gene  $i$  and gene  $j$  provided by the STRING database, and  $C$  is the number of direct connections between gene  $i$  and gene  $j$ . The smaller the biological distance is of a group, the closer the biological interrelatedness between genes in the group. We calculated the biological distance of a CVID group (39 CVID genes) and two random groups (each group consists of 39 PID genes) with the human gene connectome (HGC) Python package provided by Itan *et al.* [36].

### Prediction of candidate CVID genes

The three following steps were taken to predict the CVID candidate genes: (i) Pairwise Pearson's correlation analysis was performed on the expression values of 39 CVID genes and each protein-coding gene (or, that is, candidate gene) based on the GSE51406 and GSE72625 datasets. Using  $|r| > 0.9$  and  $P < 0.05$  as cut-off values, the candidate genes were acquired from each of the two datasets. The overlapping candidate genes obtained from the two datasets were used for the subsequent analysis. (ii) The PPI data for all human protein-coding genes were obtained from Cheng *et al.* [37], including 217,160 interactions provided by eleven databases (*e.g.*, BioGRID [38], HI-II-14\_Net [39], HPRD [40], Instruct [41], InnateDB [42], IntAct [43], MINT [44], PINA [45], SignalLink2.0 [46], KinomeNetworkX [47], and PhosphositePlus [48]). A candidate gene was then retained if the interaction between the CVID gene and candidate gene obtained in the previous step occurred in the PPI data. (iii) Kyoto Gene and Genomic Encyclopedia (KEGG) analysis was performed using the R package “clusterProfiler” for CVID genes to estimate their biological function enrichment [49]. KEGG analysis was then performed on the remaining candidate genes, and a gene was defined as a true CVID candidate gene if the candidate gene was enriched in the same pathway as the CVID gene.

### Estimation of novel CVID candidate genes

To determine whether our method was appropriate for predicting CVID candidate genes, we calculated the biological distances of the predicted CVID candidate genes and compared them to those of 39 known CVID genes. Subsequently, functional genomics alignment (FGA), a phylogenetic clustering analysis, was performed using the “APE” package available in R to assess the biological correlation between candidate CVID genes and

CVID genes [36,50]. Specifically, we first created a biological distance matrix between CVID genes and CVID candidate genes and then applied the neighbor-joining algorithm (*nj* function) to generate a phylogenetic fan-shaped tree showing the hierarchical clustering of CVID candidate genes and CVID genes. The R package “limma” was applied to screen for differentially expressed genes between patients with CVID and HCs, with  $|\log_2$  fold change  $> 0.4$  and  $P$ -value  $< 0.05$  as the cut-off values [51]. The overlap between the DEGs obtained from GSE72625 and the DEGs obtained from GSE51406 was determined using the R package “VennDiagram” [52].

### SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at <https://doi.org/10.1007/s40484-019-0174-9>.

### AUTHOR CONTRIBUTIONS

Guojun Liu performed the computations and analysis; Mikhail A. Bolkov contributed to the preparation and interpretation of the data; Irina A. Tuzankina and Irina G. Danilova conceptualized the study. All authors contributed to the writing of the manuscript and approved the final version of the manuscript.

### ACKNOWLEDGEMENTS

This study was funded by the Act 211 Government of the Russian Federation (No. 02.A03.21.0006) and the IIP UB RAS project (No. AAAA-A18-118020590108-7).

### COMPLIANCE WITH ETHICS GUIDELINES

The authors Guojun Liu, Mikhail A. Bolkov, Irina A. Tuzankina and Irina G. Danilova declare that they have no conflict of interests.

All procedures performed in studies were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

### REFERENCES

- Gathmann, B., Mahlaoui, N., Gérard, L., Oksenhendler, E., Warnatz, K., Schulze, I., Kindle, G., Kuijpers, T. W., van Beem, R. T., Guzman, D., *et al.* (2014) Clinical picture and treatment of 2212 patients with common variable immunodeficiency. *J. Allergy Clin. Immunol.*, 134, 116–126
- Bonilla, F. A., Barlan, I., Chapel, H., Costa-Carvalho, B. T., Cunningham-Rundles, C., de la Morena, M. T., Espinosa-Rosales, F. J., Hammarström, L., Nonoyama, S., Quinti, I., *et al.* (2016)

- International Consensus Document (ICON): common variable immunodeficiency disorders. *J. Allergy Clin. Immunol. Pract.*, 4, 38–59
3. Bogaert, D. J., Dullaers, M., Lambrecht, B. N., Vermaelen, K. Y., De Baere, E. and Haerynck, F. (2016) Genes associated with common variable immunodeficiency: one diagnosis to rule them all? *J. Med. Genet.*, 53, 575–590
4. van Zelm, M. C., Reisli, I., van der Burg, M., Castaño, D., van Noesel, C. J., van Tol, M. J., Woellner, C., Grimbacher, B., Patiño, P. J., van Dongen, J. J., *et al.* (2006) An antibody-deficiency syndrome due to mutations in the *CD19* gene. *N. Engl. J. Med.*, 354, 1901–1912
5. Compeer, E. B., Janssen, W., van Royen-Kerkhof, A., van Gijn, M., van Montfrans, J. M. and Boes, M. (2015) Dysfunctional BLK in common variable immunodeficiency perturbs B-cell proliferation and ability to elicit antigen-specific CD4<sup>+</sup> T-cell help. *Oncotarget*, 6, 10759–10771
6. Lo, B., Zhang, K., Lu, W., Zheng, L., Zhang, Q., Kanellopoulou, C., Zhang, Y., Liu, Z., Fritz, J. M., Marsh, R., *et al.* (2015) Patients with LRBA deficiency show CTLA4 loss and immune dysregulation responsive to abatacept therapy. *Science*, 349, 436–440
7. Fliegauf, M., Bryant, V. L., Frede, N., Slade, C., Woon, S. T., Lehnert, K., Winzer, S., Bulashevskaya, A., Scerri, T., Leung, E., *et al.* (2015) Haploinsufficiency of the NF- $\kappa$ B1 subunit p50 in common variable immunodeficiency. *Am. J. Hum. Genet.*, 97, 389–403
8. Almejun, M. B., Cols, M., Zelazko, M., Oleastro, M., Cerutti, A., Oppezzo, P., Cunningham-Rundles, C. and Danielian, S. (2013) Naturally occurring mutation affecting the MyD88-binding site of TNFRSF13B impairs triggering of class switch recombination. *Eur. J. Immunol.*, 43, 805–814
9. Kienzler, A. K., Hargreaves, C. E. and Patel, S. Y. (2017) The role of genomics in common variable immunodeficiency disorders. *Clin. Exp. Immunol.*, 188, 326–332
10. van Schouwenburg, P. A., Davenport, E. E., Kienzler, A. K., Marwah, I., Wright, B., Lucas, M., Malinauskas, T., Martin, H. C., Lockstone, H. E., Cazier, J. B., *et al.* (2015) Application of whole genome and RNA sequencing to investigate the genomic landscape of common variable immunodeficiency disorders. *Clin. Immunol.*, 160, 301–314
11. Kelsen, J. R., Dawany, N., Moran, C. J., Petersen, B. S., Sarmady, M., Sasson, A., Pauly-Hubbard, H., Martinez, A., Maurer, K., Soong, J., *et al.* (2015) Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease. *Gastroenterology*, 149, 1415–1424
12. Keerthikumar, S., Bhadra, S., Kandasamy, K., Raju, R., Ramachandra, Y. L., Bhattacharyya, C., Imai, K., Ohara, O., Mohan, S. and Pandey, A. (2009) Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA Res.*, 16, 345–351
13. Ortutay, C. and Vihinen, M. (2009) Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res.*, 37, 622–628
14. Itan, Y. and Casanova, J. L. (2015) Novel primary immunodeficiency candidate genes predicted by the human gene connectome. *Front. Immunol.*, 6, 142
15. Requena, D., Maffucci, P., Bigio, B., Shang, L., Abhyankar, A., Boisson, B., Stenson, P. D., Cooper, D. N., Cunningham-Rundles, C., Casanova, J. L., *et al.* (2018) CDG: an online server for detecting biologically closest disease-causing genes and its application to primary immunodeficiency. *Front. Immunol.*, 9, 1340
16. van Schouwenburg, P. A., Davenport, E. E., Kienzler, A. K., Marwah, I., Wright, B., Lucas, M., Malinauskas, T., Martin, H. C., Lockstone, H. E., Cazier, J. B., *et al.* (2015) Application of whole genome and RNA sequencing to investigate the genomic landscape of common variable immunodeficiency disorders. *Clin. Immunol.*, 160, 301–314
17. Yang, Y., Wang, W., Lou, Y., Yin, J. and Gong, X. (2018) Geometric and amino acid type determinants for protein-protein interaction interfaces. *Quant. Biol.*, 6, 163–174
18. Lee, H. C., Lai, K., Lorenc, M. T., Imelfort, M., Duran, C. and Edwards, D. (2012) Bioinformatics tools and databases for analysis of next-generation sequence data. *Brief. Funct. Genomics*, 11, 12–24
19. Charoentong, P., Angelova, M., Efremova, M., Gallasch, R., Hackl, H., Galon, J. and Trajanoski, Z. (2012) Bioinformatics for cancer immunology and immunotherapy. *Cancer Immunol. Immunother.*, 61, 1885–1903
20. Vasudevaraja, V., Renbarger, J., Shah, R. G., Kinnebrew, G., Korc, M., Wang, L., Huo, Y., Liu, E., Li, L. and Cheng, L. (2017) PMTDS: a computational method based on genetic interaction networks for precision medicine target-drug selection in cancer. *Quant. Biol.*, 5, 380–394
21. Yazdani, R., Ganjalikhani-Hakemi, M., Esmaeili, M., Abolhassani, H., Vaeli, S., Rezaei, A., Sharifi, Z., Azizi, G., Rezaei, N. and Aghamohammadi, A. (2017) Impaired Akt phosphorylation in B-cells of patients with common variable immunodeficiency. *Clin. Immunol.*, 175, 124–132
22. Rodríguez-Cortez, V. C., Del Pino-Molina, L., Rodríguez-Ubreva, J., Ciudad, L., Gómez-Cabrero, D., Company, C., Urquiza, J. M., Tegnér, J., Rodríguez-Gallego, C., López-Granados, E., *et al.* (2015) Monozygotic twins discordant for common variable immunodeficiency reveal impaired DNA demethylation during naïve-to-memory B-cell transition. *Nat. Commun.*, 6, 7335
23. Keller, B., Cseresnyes, Z., Stumpf, I., Wehr, C., Fliegauf, M., Bulashevskaya, A., Usadel, S., Grimbacher, B., Rizzi, M., Eibel, H., *et al.* (2017) Disturbed canonical nuclear factor of  $\kappa$  light chain signaling in B cells of patients with common variable immunodeficiency. *J. Allergy Clin. Immunol.*, 139, 220–231.e8
24. Sanaei, R., Rezaei, N., Aghamohammadi, A., Delbandi, A. A., Teimourian, S., Yazdani, R., Tavasolian, P., Kiaee, F. and Tajik, N. (2018) Evaluation of the TLR negative regulatory network in CVID patients. *Genes Immun.*, 20, 198–206
25. Clemente, A., Pons, J., Lanio, N., Cunill, V., Frontera, G., Crespi, C., Matamoros, N. and Ferrer, J. M. (2015) Increased STAT3

- phosphorylation on CD27<sup>+</sup> B-cells from common variable immunodeficiency disease patients. *Clin. Immunol.*, 161, 77–88
26. Maffucci, P., Filion, C. A., Boisson, B., Itan, Y., Shang, L., Casanova, J. L. and Cunningham-Rundles, C. (2016) Genetic diagnosis using whole exome sequencing in common variable immunodeficiency. *Front. Immunol.*, 7, 220
  27. Steele, C. L., Doré, M., Ammann, S., Loughrey, M., Montero, A., Burns, S. O., Morris, E. C., Gaspar, B., Gilmour, K., Bibi, S., *et al.* (2016) X-linked inhibitor of apoptosis complicated by granulomatous lymphocytic interstitial lung disease (GLILD) and granulomatous hepatitis. *J. Clin. Immunol.*, 36, 733–738
  28. Berrón-Ruiz, L., López-Herrera, G., Vargas-Hernández, A., Mogica-Martínez, D., García-Latorre, E., Blancas-Galicia, L., Espinosa-Rosales, F. J. and Santos-Argumedo, L. (2014) Lymphocytes and B-cell abnormalities in patients with common variable immunodeficiency (CVID). *Allergol. Immunopathol. (Madr.)*, 42, 35–43
  29. López-Gómez, A., Clemente, A., Cunill, V., Pons, J. and Ferrer, J. M. (2018) IL-21 and anti-CD40 restore Bcl-2 family protein imbalance *in vitro* in low-survival CD27<sup>+</sup> B cells from CVID patients. *Cell Death Dis.*, 9, 1156
  30. Karnell, J. L., Kumar, V., Wang, J., Wang, S., Voynova, E. and Ettinger, R. (2017) Role of CD11c<sup>+</sup> T-bet<sup>+</sup> B cells in human health and disease. *Cell. Immunol.*, 321, 40–45
  31. Niemela, J., Kuehn, H. S., Kelly, C., Zhang, M., Davies, J., Melendez, J., Dreiling, J., Kleiner, D., Calvo, K., Oliveira, J. B., *et al.* (2015) Caspase-8 deficiency presenting as late-onset multi-organ lymphocytic infiltration with granulomas in two adult siblings. *J. Clin. Immunol.*, 35, 348–355
  32. Rensing-Ehl, A., Warnatz, K., Fuchs, S., Schlesier, M., Salzer, U., Draeger, R., Bondzio, I., Joos, Y., Janda, A., Gomes, M., *et al.* (2010) Clinical and immunological overlap between autoimmune lymphoproliferative syndrome and common variable immunodeficiency. *Clin. Immunol.*, 137, 357–365
  33. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13, 2498–2504
  34. Horvath, S. and Dong, J. (2008) Geometric interpretation of gene coexpression network analysis. *PLOS Comput. Biol.*, 4, e1000117
  35. Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal. Complex Syst.*, 1695, 1–9
  36. Itan, Y., Zhang, S. Y., Vogt, G., Abhyankar, A., Herman, M., Nitschke, P., Fried, D., Quintana-Murci, L., Abel, L. and Casanova, J. L. (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *Proc. Natl. Acad. Sci. USA*, 110, 5558–5563
  37. Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A. L. and Loscalzo, J. (2018) Network-based approach to prediction and population-based validation of *in silico* drug repurposing. *Nat. Commun.*, 9, 2691
  38. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, 34, D535–D539
  39. Rolland, T., Taşan, M., Charleaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, 159, 1212–1226
  40. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, 37, D767–D772
  41. Meyer, M. J., Das, J., Wang, X. and Yu, H. (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics*, 29, 1577–1579
  42. Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E., Brinkman, F. S. and Lynn, D. J. (2013) InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.*, 41, D1228–D1233
  43. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32, D452–D455
  44. Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, 35, D572–D574
  45. Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., Biankin, A. V., Hautaniemi, S. and Wu, J. (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res.*, 40, D862–D865
  46. Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfi, M., Dül, Z., Zsáki, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., *et al.* (2013) Signalink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.*, 7, 7
  47. Cheng, F., Jia, P., Wang, Q. and Zhao, Z. (2014) Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget*, 5, 3697–3710
  48. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V. and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, 40, D261–D270
  49. Yu, G., Wang, L. G., Han, Y. and He, Q. Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16, 284–287
  50. Paradis, E., Claude, J. and Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289–290
  51. Diboun, I., Wernisch, L., Orengo, C. A. and Koltzenburg, M. (2006) Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics*, 7, 252
  52. Chen, H. and Boutros, P. C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in

- R. BMC Bioinformatics, 12, 35
53. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res., 37, D412–D416
54. Russell, M. A., Pigors, M., Houssen, M. E., Manson, A., Kelsell, D., Longhurst, H. and Morgan, N. G. (2018) A novel *de novo* activating mutation in STAT3 identified in a patient with common variable immunodeficiency (CVID). Clin. Immunol., 187, 132–136