ACE2 Homo-dimerization, Human Genomic variants and Interaction of Host Proteins Explain High Population Specific Differences in Outcomes of COVID19

Swarkar Sharma[1*], Inderpal Singh[2†], Shazia Haider[3†], Md. Zubbair Malik[4†], Kalaiarasan Ponnusamy[5†], Ekta Rai[1]

1. Human Genetics Research Group, School of Biotechnology, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India.
2. Bioinfores Pvt. Ltd., R. S. Pura, Jammu, Jammu and Kashmir, India.
3. Jaypee Institute of Information Technology, Noida, sector-62, Uttar Pradesh, India
4. School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India.
5. School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.

† - All authors contributed Equally

Corresponding Author:
Dr. Swarkar Sharma
Coordinator, Human Genetics Research Group,
School of Biotechnology, Shri Mata Vaishno Devi University, Katra, J&K, India
Email: swarkar.sharma@smvdu.ac.in
Mobile: +91-9419955636; Ph: +91-1991-285535//285525 Ext. 2385

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a positive single stranded RNA virus that causes a highly contagious Corona Virus Disease (COVID19). Entry of SARS-CoV-2 in human cells depends on binding of the viral spike (S) proteins to cellular receptor Angiotensin-converting enzyme 2 (ACE2) and on S protein priming by host cell serine protease TMPRSS2. Recently COVID19 has been declared pandemic by World Health Organization yet high differences in disease outcomes across countries have been seen. We provide evidences based on analyses of existing public datasets and by using various *in-silico* approaches to explain some of these as factors that may explain population level differences. One of the key factors might be entry of virus in host cells due to differential interaction of viral proteins with host cell proteins due to different genetic backgrounds. Based on our findings, we conclude that higher expression of *ACE2* facilitated by natural variations, acting as Expression quantitative trait loci (eQTLs) and with different frequencies in different populations, results in ACE2 homo-dimerization which is disadvantageous for TMPRSS2 mediated cleavage of ACE2 and becomes more difficult in presence of broad neutral amino acid transporter, B0AT1 (coded by *SLC6A19*), that usually does not express in Lungs. We also propose that the monomeric ACE2 has higher preferential binding with SARS-CoV-2 S-Protein vis-a-vis its dimerized counterpart. Further, eQTLs in *TMPRSS2* and natural structural variations in the gene may also result in differential outcomes towards priming of viral S-protein, a critical step for entry of Virus in host cells. In addition, we suggest some other potential key host genes like *ADAM17, RPS6, HNRNPA1, SUMO1, NACA, BTF3* and some other proteases as Cathepsins, that might have a critical role. Understanding these population specific differences may help in developing appropriate management strategies.

Introduction

The recent emergence of corona virus disease (COVID19) caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2) has resulted in >25 Million infections and >170 thousands deaths worldwide so far, and the numbers are increasing exponentially [https://covid19.who.int/]. SARS-CoV-2 is reported to be originated in bats [1] and transmitted to humans via unknown intermediate host. However, its origin is still being questioned time and again. With the declaration of SARS-CoV-2 as pandemic by World Health Organisation (WHO), extensive research worldwide has been carried out. It has been established that Human ACE2 mediates SARS-CoV-2 entry into cells through its Spike (S) Protein an primarily makes entry to host body through respiratory tract with recent reporting of nasal epithelial cells as potential initial infection site. [2] Angiotensin-converting enzyme 2 (ACE2) is functional receptor on human cells for this newly originated coronavirus [3] with higher affinity then severe acute respiratory syndrome coronavirus (SARS-CoV) originated in 2002. [4]. However, no substantial evidence exists that ACE2 higher expression primarily is associated with degree of infection and in addition COVID19 lethality is mostly driven by the extent of underlying lung injury [5] whereas a negative correlation has been reported between ACE2 expression and lung injury [6], making it an interesting question to explore.

High differences in clinical outcomes and across countries have been seen [https://covid19.who.int/] and noted that neither all people who are exposed to SARS-CoV-2 develop infection nor all infected patients end up in severe respiratory illness [7] which cannot be explained by immunity alone [8] . This results in hypothesis of differential genetic susceptibility to COVID19 and virulence of SARS-CoV-2 in different populations [9]. Efforts are being made worldwide to understand it better and it becomes complex with reports that BCG vaccination reduces morbidity and mortality for COVID19 in human population [10]. Due to extensive research going on in parallel, evidences in favour [11] as well as against [12] existence of SARS-CoV-2 S-protein binding-resistant ACE2 natural variants, in different populations, have been generated recently. In addition, studies highlighting role of eQTLs in *ACE2* expression and resulting in potential differential COVID19 fatality [12, 13] are pouring in. Where most of the such studies are targeting only natural variations in ACE2 gene as SARS-CoV-2 differential susceptibility factor, recent evidences suggested that additional host proteins like cellular serine protease TMPRSS2 act as co-factors and are critical for efficient cellular infection by SARS-CoV-2 [14]. Further, it is equally important to consider that rare functional variants, though may have consequences, yet may not explain large scale population level differential clinical outcomes.

This highlights the importance of identification of such other potential co-factors and underlying mechanisms these genes can play. At the same time, understanding correlations between these host proteins and their interactions with SARS-CoV-2 and ACE2 may explain many of the unanswered questions. Further evaluating these host genes, exploring natural occurring variants in these, their expression patterns may also

shed some light on better understanding of differential susceptibility to COVID19 and virulence of SARS-CoV-2. In the present study, we have tried to exploit existing literature to understand mechanisms of correlations, of host genes, specifically associated with interactions related to ACE2 and some prominent Viral proteins, using *in-silico* approaches and also tried to understand these as factors explaining population level differences.

Methodology

As SARS-CoV-2 is primarily a respiratory pathogen, the present study was to understand mechanisms of its entry to cells in lungs and believe it can be extrapolated to other respiratory tract tissues to explore potential factors that may result in to differential outcomes in respiratory illness. To being with, we started with the most studied host protein ACE2 and tried to understand its interaction with SARS-CoV-2 S-Protein through Molecular Dynamics simulations, followed by adding other interacting proteins. We further added layers of other methods to understand better the potential causes and outcomes.

Molecular Dynamics (MD) Simulations

Recently submitted experimental structure 6M17.pdb of ACE2 dimer bound to two B0AT1 (SL6A19) and two receptor-binding domains (RBD) of SARS-CoV-2 with overall resolution of 2.90 Angstrom was downloaded from the Protein Data Bank for structural, visualization and dynamics analysis [15]. This structure was corrected for missing amino acids, side chains, missing hydrogen atoms, correction of disulphide bonds etc., optimized for pKa corresponding to physiological pH 7.2 and energy minimized to correct steric clashes using the Protein Preparation Wizard of Schrodinger Maestro [16]. Predefined solvation model TIP3P was used and overall neutrality of the system was maintained by addition of Na+ and Cl- counter ions[17]. Physiological salt concentration of 0.15 Molar was generated through addition of NaCl. Periodic boundary condition of 10 Angstrom was set using the System Builder Tool of Desmond software[18]. Total two MD simulations, each 10 nanosecond long, were conducted. In one of the simulation, six chains i.e. ACE2 dimer with each monomer bound to a B0AT1 (SL6A19) and RBD domain was simulated and the resulting solvated system consisted of 424,847 atoms, while for other simulation monomeric ACE2 bound to viral RBD consisted of 174,900 atoms. Root mean square fluctuation (RMSF) and principal component analysis were done through R based BIO3D module [19]. MMGBSA free energy of binding between proteins was calculated through Prime Software of Schrodinger Suite [20]. Structural visualizations and images were traced using pyMOL and VM D [20] [21].

Gene Expression in Alternate Transcripts and Regulation Analyses

Genotype-Tissue Expression (GTEx) portal at https://gtexportal.org/ was used to explore various gene expression. By Tissue, Multigene Query as well as transcript browser was used to understand expression of splice variants and exons in Lungs. eQTLs were viewed by GTEx IGV Browser as well as GTEx Locus

Browser was used to plot gene specific eQTLs. eQTLs and other gene regulatory information, splice variants and ESTs were also explored through UCSC genome browser https://genome.ucsc.edu/ with build GRCh37/hg19. Table browser was used to interact various datasets to look for overlaps and filter outcomes. Retrieved information was saved in files as well as plotted on UCSC browser as custom tracks. Comparison of RNA expression and Protein expression, mainly in lungs, was done at The Human Protein Atlas [22] at www.proteinatlas.org. HaploReg v4.1 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) was used to annotate effect of noncoding genome variants of regulation motifs.

Analyses of Allele frequencies distribution of Variants in different population Groups

Genetic data for global population groups was explored through the GnomAD portal at https://gnomad.broadinstitute.org/, as well as dbSNP database of NCBI at https://www.ncbi.nlm.nih.gov/snp. Gene specific SNPs data was also retrieved from 1000 genomes (1000G) Phase 3 data set available through https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes as per Homo sapiens:GRCh37.p13 (GCF_000001405.25). Data for the population groups belonging to five super population groups [African (AFR), Ad mixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS)] was analysed. The folowing sub population groups were studied. AFR: African Caribbeans in Barbados (ACB), Americans of African Ancestry in SW USA (ASW), Esan in Nigeria (ESN), Gambian in Western Divisions in the Gambia (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), Yoruba in Ibadan, Nigeria (YRI). AMR: Colombians from Medellin, Colombia (CLM), Mexican Ancestry from Los Angeles USA (MXL), Peruvians from Lima, Peru (PEL), Puerto Ricans from Puerto Rico (PUR). EAS: Chinese Dai in Xishuangbanna, China (CDX), Han Chinese in Beijing, China (CHB), Southern Han Chinese (CHS), Japanese in Tokyo, Japan (JPT), Kinh in Ho Chi Minh City, Vietnam (KHV). EUR: Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Finnish in Finland (FIN), British in England and Scotland (GBR), Iberian Population in Spain (IBS), Toscani in Italia (TSI). SAS: Bengali from Bangladesh (BEB), Gujarati Indian from Houston, Texas (GIH), Indian Telugu from the UK (ITU), Punjabi from Lahore, Pakistan (PJL) and Sri Lankan Tamil from the UK (STU).

LDproxy tools of LDlink 4.0.3 web based suite (https://ldlink.nci.nih.gov/) was used to map proxy variants in strong LD and with putatively functional role. LDpop tool of the LDlink 4.0.3 was used for geographically annotating allele frequencies in 1000G populations, alternatively also done by a web based tool Datawrapper (https://www.datawrapper.de/) in some of the figures.

Computational structural analysis on SARS CoV-2 S-protein and TMPRSS2

TMPRSS2 has been recently characterized as critical component for cell entry by SARS-CoV-2 [14]. To understand, the interactions, the protein sequence of surface glycoprotein (YP_009724390.1) of SARS CoV-2 and two transcripts of TMPRSS2 (NP_005647.3 and NP_001128571.1) protein of homo sapiens, were retrieved from the NCBI-Protein database. Pairwise sequence alignment of TMPRSS2 isoforms was carried

out by Clustal Omega tool [23]. The protein domain information and transcript variation were retrieved from Uniport [24], Prosite [25], Pfam [26] and ENSEMBL [27], respectively. The homology model of the S protein and TMRPSS2 constructed using Swiss model [28], whereas 3D structure of the ACE2 retrieved from the PDB database (PDB ID: 6M17). These structures were energy minimized by the Chiron energy minimization server [29]. The binding site residues of the proteins retrieved from Uniport and literature. The variant structure of the TMRPSS2 protein was generated using WHATIF server [30] and energy minimized. The effect of the variant was analyzed using HOPE [31] and I-mutant [32]. The I-mutant method allows us to predict stability of the protein due to mutation. The docking studies for wild and mutant TMRPSS2 with S protein and ACE2 carried out using HADDOCK [33].

Network analysis on SARS-CoV-2 with Human Proteins

We downloaded the SARS-CoV-2 genome (Accession number: MT121215) from the NCBI database. In order to find the Host-Pathogen Interactions (HPIs), the SARS-CoV-2 protein sequences were subjected to Host-Pathogen Interaction Database (HPIDB 3.0) [34, 35]. In addition to the host proteins, we added two more proteins (*TMPRSS2* and *SLC6A19*) which found to have an important role in the mechanism of viral entry [14]. The protein-protein interaction (PPI) and transcription factor regulation of human proteins were retrieved from GeneMANIA [36] and literature [37-42], respectively. The Host Pathogen Interaction Network (HPIN) was visualized using Cytoscape [43] which includes the information collected from HPIDB, PPIs and TFs. Modules are defined as the set of statistics and functionally significant interacting genes [44] which is constructed by using MCODE [45]. Further, Hubs were identified using Network analyzer [46], the plugin of Cytoscape v3.2.1. We also studied the hub proteins association with diseases using DisGeNET [47].

Perturbation by leading hub removal analysis

The topological properties (degree distribution, average clustering coefficient and average neighbourhood connectivity) of the network, which characterizes the structure and behaviour of the HPIN.

Degree Distribution (P(k)): Degree k is the number of interaction a node in the HPIN. *P(k)* is the probability of randomly chosen node to have k interaction with the neighbour. The probability of degree distribution *(P(k))* of the network is calculated by:

$$P(k) = \frac{n_k}{N}$$

where, n_k is equal to the number of nodes with degree k and N is equal to the size of the network [48, 49].

Clustering Coefficient (C(k)): Clustering coefficient defines how strongly the nodes in a network tend to cluster together. Clustering coefficient of the i^{th} node in undirected network can be obtained by:

$$C(k_i) = \frac{2e_i}{k_i(k_i - 1)}$$

where, e_i is the number of connected pairs of the nearest-neighbour of the *i-th* node and k_i denotes the degree of the respective node [50, 51].

Neighborhood Connectivity ($C_N(k)$): Neighbourhood connectivity of a node is defined as the connectivity of all the neighbours of the node [52]. The average connectivity of the nearest neighbours of a node is given by:

$$C_N(k) = \sum_q qP(q|k)$$

where, $P(q|k)$ denotes the conditional probability that a link belonging to a node with connectivity k points to a node with connectivity q.

For scale free network, $C_N(k)$ is constant while it follows power law $C_N(k) = c^{-\beta}$ for hierarchical network with β approximately equal to 0.5 [53]. The positive and negative signs in β indicates the assortivity or disassortivity of network respectively [54].

Removal of high degree nodes (hubs) in biological networks may cause lethality in various complex biological networks. This idea of understanding may enable us to determine the topological features of the network architecture, important functionally related modules and hubs [55, 56]. The topological properties were calculated after each subsequent removal using the Network Analyzer plugin in Cytoscape version 3.7.1. The calculated topological properties of knock-out experiment were compared with those of the main network and the change in the properties helped in understanding the role of leading hubs in the network.

Results and Discussion

SARS-CoV-2 uses ACE2 for entry and its S-protein priming by the serine protease TMPRSS2 is key factor. Other proteases like cathepsin B and L may be available in host cells yet only TMPRSS2 activity is essential for pathogenesis and spread [14].

Dynamics and binding analysis of ACE2 with SARS CoV-2 RBD in dimeric and monomeric state

Mutagenesis experiments have reported the importance of Arginine and Lysine residues in ACE2 positioned between 697 to 716 as an important recognition site for TMPRSS2 mediated cleavage **(Figure 1A)** of ACE2. Mapping these residues on the recently resolved structure (61M7.pdb) suggests that this region lies in symmetric dimerization interface between ACE2 homodimer and is also masked from outside by two BoAT1(SL6A19) in an overall 1:2:1 hetero-tetramer of B0AT1 (SL6A19) and ACE2 **(Figure 1A)** [15] . In a recently reported study, the overexpression of ACE2 has been reported to have an overall protective role in viral infection among patients [6]. Increased density of ACE2 on the cell surface has been observed to protect from lung injury [6] also been observed at many other instances. We propose increased expression of ACE2 due to various factors, including natural occurring variations influencing expression of the gene, may increase collisions and hence result in homo-dimerization of this protein. Binding of B0AT1 (SL6A19) to ACE2 is not dependent on ACE2 homodimerization, but its presence along with dimerised ACE2

effectively shields the later from interaction with TMPRSS2 and may prevent it or reduce its efficiency to cleave ACE2, a hypothesis based on visualization of the available experimentally resolved structure 61M7.PDB [15]. This observation is significant since the cleaved ACE2 interacts more efficiently with the SARS CoV-2 RBD domain. To understand the mechanistic effect of dimerization, we conceived and simulated two situations: one, where ACE2 dimer is bound to two B0AT1 (SL6A19) and two, RBD domains of SARS CoV-2 and second, where only monomeric ACE2 is bound to one viral RBD. From the RMSF analysis of the simulated trajectories, we observed higher order fluctuation in monomeric ACE2's homo-dimerization interface while the rest of the ACE2 displayed a similar fluctuation profile including residues interacting with RBD **(Figure 2A)**. PCA analysis and amino acid residue loadings on the PCs (PC1 in this case) reported slightly higher away ward conformational dynamics w.r.t RBD in dimeric ACE2 **(Figure 2B and C)**. This contrasting essential dynamics in dimeric ACE2 is explained well through the MM-GBSA based free energy of binding calculation between the 10th nanosecond final confirmation of ACE2-RBD complex. From both simulations through MM-GBSA method, 1.5 fold strong binding of monomeric ACE2-RBD complex compared to dimeric ACE2-RBD was observed (*i.e.* ΔG=-75.58 for Dimeric ACE2-RBD vs ΔG=-116.98 for Monomeric ACE2-RBD Complex). This observation further explains the protective role of overexpressed ACE2, resulting in its dimerization and reduced affinity for the RBD domain of SARS CoV-2 S1 protein. As our current computational resources could not support longer simulations therefore, we propose longer simulations for robust statistical inference to understand this phenomenon in more details. Further, we are also not clear whether the presence of B0AT1(SL6A19) has any allosteric effect on this observation, in addition to already observed masking of ACE2 from TMPRSS2 mediated cleavage. Additionally, another metalloprotease ADAM17 has been reported to compete with the TMPRSS2 and cleaves ACE2 in a way that only cleavage by TMPRSS2 was reported to drive the SARS CoV entry inside the cell [57] which is yet to be explored in relation to SARS CoV-2. Recent literature also indicated the potential role of other proteases like cathepsin B/L that can functionally replace TMPRSS2, need to be evaluated extensively. [2] All these hypotheses warrant further *in-silico* analyses with better computational resources, as well as experimental study designs and for these we are open to seek collaborations.

However, the observations made are of high importance and emphasise on evaluation of expression of the ACE2 and BoAT1 (SL6A19) along with TMPRSS2 among patients with varied clinical response to SARS CoV-2 infection, differential outcomes and also correlation with observed mortality. Therefore differential expression of these genes among patients with displaying varied responses from asymptomatic to acute symptoms is worth exploration which may act as biomarkers if proven, to predict severity and susceptibility.

Differential Genomics backgrounds derived expression of ACE2 and TMPRSS2

With the observations, from our MD analyses **(Figures 1 and 2)** that higher expression of ACE2 gene may promote ACE2 homo-dimerization rendering less binding affinity to SARS CoV-2 RBD as well as masking

TMPRSS2 cleavage site **(Figure 3)**, evidences on eQTLs and ACE2 higher expression in East Asians [13] and reported protective role of ACE2 in lung injury [6], we explored more about ACE2 expression as well as differential genomic backgrounds in different population groups.

ACE2 Expression and functional SNPs related to its expression

Tissue Specific evaluation of ACE2 gene in GTEx portal indicated low level expression of the ACE2 in Lungs (**Figure 4a and Figure 5**). Further, it was observed that not all the transcripts and exons of the gene express in Lungs (**Figure 4b and 4c**). As expression of *ACE2* is less in Lungs thus, probably GTEx portal did not return cis-eQTLs for ACE2 gene in lungs. However, eQTLs for the gene were found in other tissues **(Figure 6a and b)** in GTEx portal. These eQTLs remained the same across tissue sets and showed similar effect patterns in expression (**Figure 6b)**. Yet to corelate the expression of the gene with genomic variations, regions of the gene were explored for potential regulatory elements and overlapped it with common variations through table browser tool of UCSC genome browser with assumption that the variants having population level effect should be common variants **(Figure 4d and 7).** Variations data was retrieved from 1000 Genomes Phase 3 dataset and filtered for variants with at least 10% frequency of the alternate allele in global population. Overlapping the variants from both the exercises resulted in shortlisting of 2 SNPs rs1978124 and rs2106809 which were observed with differential frequency distribution in different populations of the world **(Figure 8)** at both super population group and sub population group level. Interestingly, EAS and EUR sub population groups were observed to show relatively uniform frequency distribution within group. However, in AFR, AMR and SAS sub population groups intra population group differences were observed to be higher, indicating diversity in the gene pool of population groups. HaploReg annotations indicated that these SNPs are in region with Enhancer histone marks and DNAase activity as well as the same was observed through multiple regulatory elements tracks in UCSC Genome browser **(Figure 4d and 7)** including GH0XJ015596 enhancer marked by GeneHancer database. [58] Linkage Disequilibrium (LD) values as (r^2) with other putative proxy functional variants was also explored and mapped with UCSC genome browser. Interestingly, it was observed that SNPs rs1978124 and rs2106809 have a strong LD block of >100kb with various SNPs in absolute LD but upstream of ACE2 gene across population groups **(Figure 9)**. This indicates a strong enhancer activity from the region that may affect higher expression of ACE2 gene in lungs which may facilitate ACE2 homo-dimerization (**Figure 3**). However, it also requires experimental validation.

TMPRSS2 Expression and functional SNPs in the gene with potential effect

GTEx portal indicated relatively high levels of TMPRSS2 expression in Lungs (**Figure 4a and Figure 10c**), differential transcription (**Figure 10c**) but no protein expression in lungs (**Figure 10a**). Evaluation of the Protein atlas portal, source of Figure 10a, indicated that both antibodies used for detection TMPRSS2 (not shown in Ms) were restricted to target near N terminal of the protein (either cytoplasmic domain or proximal

extracellular domain near membrane). Evaluation of the GTEx portal for cis-eQTLs for TMPRSS2 gene in lungs, returned a large number of eQTLs but with a peculiar feature (**Figure 11a and b**). The eQTLs in lungs were different as compared to other tissues and found to be concentrated in region of the gene with potential alternate transcripts (**Figure 12a,b and Figure 13**), towards end of the gene in relatively high expressing exons (**Figure 12b)** and coding for the amino acid sequence that has putative functional role in protein as serine protease domain **(Figure 12a)** critical for ACE2 cleavage and SARS CoV-2 S-Protein priming. Variations data was retrieved from 1000 Genomes Phase 3 dataset and filtered for variants with at least 10% frequency of the alternate allele in global population and overlapped it with cis-eQTLs data resulting in 10 SNPs (rs463727, rs55964536, rs4818239, rs734056, rs4290734, rs2276205, rs34783969, rs11702475, rs62217531 and rs383510). Annotation of the SNPs indicated all the SNPs cluster together and are in strong LD block **(Figure 13** and **Figure 14a)**. HaploReg annotations indicated Enhancer histone marks and DNAase activity in the regions overlapping these SNPs amongst these rs4818239 showed a prominent putative functional role (**Figure 14b**) which further requires experimental validation. However, the frequency distribution of the variant rs4818239 in different populations groups of 1000G showed interesting differential pattern (**Figure 14c**). We also explored if any alternate functional variations, yet common in populations exist by screening TMPRSS2 gene through genomAD browser and filtered for missense only variations. The search returned rs75603675 (NP_001128571.1:p.Gly8Val or G8V) and rs12329760 (NP_001128571.1:p.Val197Met or V197M), also observed with differential frequency distribution in 1000G populations (**Figure 14d and 14e**). The observations indicate TMPRSS2 variants may influence interaction with ACE2 as well as SARS CoV-2 (**Figure 3**) resulting in population specific differential outcomes.

The variation in TMPRSS2 could inhibit the ingress of SARS CoV-2

We further opted to explore more the structure and function of TMPRSS2 protein. Pairwise sequence alignment of two isoforms of TMPRSS2 suggests that Isoform-2 lack 37 residues at N-terminal compare to Isoform-1, which was the longest transcript and codes for 529 amino acids (**Figure 15**). Since the human TMPRSS2 protein structure was not available in the PDB database, We have generated the computational protein model. The model built using Serine protease hepsin (PDB ID: 5CE1) of homo sapiens. The protein 3D structure modelled from 146-491 residues of TMPRSS2 with sequence identity, GMQE and QMEAN of 33.82%, 0.53 and -1.43 values, respectively. It showed that the model was constructed with high confidence and best quality. In the similar manner 3D structure of S Protein of SARS CoV-2 was built with sequence identity, GMQE and QMEAN of 99.26%, 0.72 and -2.81 values, respectively. The protein stability analysis showed that rs12329760 (p.Val197Met or V197M) (Isoform-1) variation could decrease the stability of TMPRSS2 protein with $\Delta\Delta G$ value of -1.51 kcal/mol. The HOPE results suggest, V197M variation is located within a domain, SRCR (GO Term: Scavenger Receptor Activity as annotated in UniProt) and introduces an amino acid with different properties, which could disturb this domain and can abolish its function. Analyses of another variation rs75603675 (p.Gly8Val or G8V) showed that wild-type residue, glycine flexibility might

be necessary for the protein function and alteration in this position can abolish this function as the observed torsion angles for this residue were unusual. It could be speculated that only glycine was flexible enough to make these torsion angles, change at the location into another residue would force the local backbone into an incorrect conformation and may disturb the local structure. The variant residue was also observed to be more hydrophobic than the wild-type residue. Since sequence similarity search did not find any significant template at the N-terminal of this protein, we could not generate a quality model for isoform-2, which contains G8V variation; hence we carried out sequence-based stability analysis. The results suggest G8V variation could increase the stability of the TMPRSS2 protein with ΔΔG value of -0.10 kcal/mol. Further, it is known that Arginine and lysine residues within amino acids 697 to 716 are essential for efficient ACE2 cleavage by TMPRSS2 [57] and recent studies have shown that SARS CoV-2 uses the ACE2 for entry and the serine protease TMPRSS2 for S protein priming. Based on information from previous studies, we docked the TMPRSS2 p.Val197Met wild-type and variant protein with ACE2 and S protein of SARS CoV-2. The docking results suggest that the variant 197M protein could promote the binding to S-protein and inhibit the binding with ACE2 (**Figure 16**). However, these observations need critical revaluation as well as experimental work to understand these interactions better.

SLC6A19 (B0AT1) if expresses naturally in Lungs and other respiratory tract cells, may provide protection.

One of the interesting gene is *SLC6A19* that codes for protein B0AT1 that expresses in very limited number of tissues and reported absent in Lungs (**Figure 4a and Figure 17a,b,c**). Our findings indicated its protective role with competitive hinderance in binding to TMPRSS2. As its expression was observed to be very low in Lungs we resolved to look for indirect signatures that may have putative role in providing differential susceptibility. We noticed several variations clustering together in a potential enhancer region (**Figure 17d and e**) and with differential frequencies in different populations. eQTL analyses at GTEx portal also showed huge list of potential SNPs upregulating or down regulating SLC6A19 expression in Pancreas, leaver and whole blood cells as well as with overlap with potential Transcription factor binding sites (**Figure 17f)**. We hypothesize a potential chance of some natural occurring variation/s that can induce expression of BoAT1 in respiratory tract thus providing protective role in this scenario (**Figure 3**). However, this requires more extensive computational exploration as well as experiential validations.

Host-Pathogen Interaction Modelling

We believed additional genes and factors might be playing role thus, we further carried out a systems biology study to identify these novel key regulators that may influence this Human and SARS CoV-2 interaction. A detailed Host Pathogen Interaction Network (HPIN) was created. The constructed HPIN contained 163 interactions, involving 31 nodes, which includes 4 viral proteins, 27 human proteins and 8 Transcription Factors (TF) (**Figure 18A**). From the HPIN, we identified hubs, namely *RPS6*, *NACA*, *HNRNPA1*, *BTF3* and

SUMO1 with 19, 18, 17, 16 & 12 degrees, respectively in the network (**Figure 18A**). This indicates the affinity to attract a large number of low degree nodes towards each hub, which is a strong evidence of controlling the topological properties of the network by these few hubs [59]. Interestingly, out of five significant hubs, four hubs (*RPS6*, *NACA*, *HNRNPA1* and *BTF3*) present in module (Nodes: 16, Edges: 118, Score: 15.73) and one hub *SUMO1* is present at motif level which considered one of the most important regulating motif of biological network at a fundamental level (**Figure 18B**). From our prediction; we found, four viral proteins (*S*, *N*, *ORF1a* & *ORF1ab*) target five host protein groups (*ACE2*, *SUMO1*, *HNRNPA1*, *RPS6* & *ATP6V1G1*). *SUMO1* & *HNRNPA1* are target by same viral proteins (N). The hub protein, *RPS6* (highest degree) directly interacts with one of the important protein *TMPRSS2* which further propagate the signal to *ACE2 & SLCA19*. The Transcription factor HIF1A and BCL6, STAT5, YBX1 inhibits *ACE2* and *SUMO1* whereas MYC, AR and HNF4a, HNF1a activates *HNRNPA1*, *TMPRSS2* and *SLCA19* respectively. From the gene diseases association study few diseases are highly associated with these important hub proteins (**Figure 19A**); *HNRNPA1* (36%) followed by *RPS6* (25%), *SUMO1* (21%), *BTF3* (9%) and *NACA* (9%) (**Figure 19B**). Clinically, patients with COVID-19 present with respiratory symptoms, Anoxia, fatigue, heart failure etc. are associated with these five hub proteins mainly *HNRNPA1& SUMO1*.

By hub removal methodology; we tried to understand the effect of hub removal and calculated the topological properties of the HPIN as a control. The probability of degree distribution P(k) showed that the HPIN followed a power law scaling behavior. The power law behavior was also checked and confirmed by using statistical test for power law fitting (p-value≥0.1)[60] in the hub-removal process. The hub removal analysis showed an important magnitude of changes in network metrics such as Degree distribution (Indegree & outdegree), Clustering coefficient and Neighborhood connectivity (**Figure 20A**). This shows that removing *RPS6*, *HNRNPA1*, *SUMO1*, *NACA* & *BTF3* leading hubs from the network allows more responsibilities of the existing leading hubs in the resulting network in order to reorganize and save the network properties from breakdown. The Indegree and Outdegree distribution of HPIN showed the removal of *NACA, RPS6* & *SUMO1* differ from the complete (**Figure 20B**). In case of clustering coefficient, the increase in γ indicates the increase in compactness in the hubs removed network in order to save the network from breakdown. Perturbation of *BFT3* shows minimum γ, indicating removal of *BFT3* the network is less compactness, which may lead to the delay in flow of signal. In Neighbourhood connectivity with exponent μ which is maximum in removal of *SUMO1* hub, which reveals that the regulating roles of remaining hubs become more important [61]. The increase in μ indicates that the information processing in the network becomes faster when *SUMO1* hubs are removed which means that local perturbations due to hubs removal are strong enough to cause significant change in global scenario [62] (**Figure 20B**). The knock out hubs experiment we performed, shows no drastic change in the topological properties of the perturbed network, but enhance the regulating capabilities of the modules than the hubs, which is the consequence of strong inter-links in the network.

Overall network analysis showed *ACE2* is not only the key molecule for entry and survival of SARS-CoV-2 virus, the hub proteins like *RPS6*, *HNRNPA1*, *SUMO1*, *NACA* & *BTF3* might also play a vital role. Analysing

these interactions could provide further important understanding for the underlying biological mechanism of SARS-CoV-2 virus infection and identifying putative drug targets.

To conclude (**Figure 3**), higher expression of *ACE2* is facilitated by natural variations with different frequencies in different populations and functionally associated with expression of the gene. The higher expression of ACE2 facilitates homo-dimerization resulting in hindrance to TMPRSS2 mediated cleavage of ACE2. It becomes more difficult in presence of B0AT1 that usually does not express in Lungs. We also propose that the monomeric ACE2 has higher preferential binding with SARS-CoV-2 S-Protein vis-a-vis its dimerized counterpart. Further, natural variations in *TMPRSS2*, with potential functional role, and their differential frequencies may also result in differential outcomes towards interaction with ACE2 or priming of viral S-protein, a critical step for entry of Virus in host cells. In addition, we have identified some other potential key host genes like *ADAM17, RPS6, HNRNPA1, SUMO1, NACA* and *BTF3,* that might have a critical role. With all this background, it is anticipated that in populations like Indian populations, with highly diverse gene pool, a great variation in clinical outcomes is expected and that could be population/region specific, but primarily due to gene pool structure of the region, despite similar exposure levels to SARS CoV-2 and resources. Understanding these population specific differences may help in developing appropriate management strategies.

Acknowledgment
All the authors acknowledge Prof. RNK Bamezai (Padamshri), former Vice Chancellor Shri Mata Vaishno Devi University, Katra and former Professor and Coordinator, National Centre for Applied Human Genetics, School of Life Sciences, Jawaharlal Nehru University, New Delhi, for his critical suggestions through various virtual discussion rounds, resulting in present study design. Authors also acknowledge Prof. Gyaneshwar Chaubey, Banaras Hindu University, UP, India for inputs, suggestions and assistance with geographical mapping of the allele frequencies from 1000G dataset.

Author Contributions
SS conceived the concept. IP carried out molecular modelling simulations of ACE2 and related analyses. KP carried out TMPRSS2 modelling and related analyses. SS and ER carried out Human Population data screening and related work. SH and MZM carried out network analyses to identify additional key genes. SS, IP, SH, MZM designed, analysed, executed and wrote their part of manuscript. SS interpreted the results together, ER assisted SS in compilation of figures and SS compiled the overall MS.

Competing Interests
The authors declare no competing interests associated with MS. For declaration purposes, SS is founder, chief scientific advisor of a startup “Biodroid Innovations Pvt Ltd” and IP is director of “Bioinfores Pvt. Ltd.”.

LIST OF FIGURES

Figure 1 Structural representation and essential dynamics of ACE2

A multi chain complex of dimeric ACE2 is shown (colored as purple and forest green surface representation) bound to B0AT1 (colored as chocolate and orange surface) through its C terminal amino acids and SARS-CoV-2 RBD domain through its N terminal amino acids (shown as black and blue surface), the recognition site of TMPRSS2 is also highlighted as surface against the cartoon background [A]. The cartoon representation along with CPK spheres blue-white-red showing the direction of essential dynamics calculated through PCAs PC1 of ACE2-RBD region from both simulations is shown [B and C].

Figure 2 Root mean square fluctuation and contribution of ACE2 residues to PC1 and 2 along with proportion of variance

Root mean square fluctuation (RMSF) plots of ACE2 protein from monomeric ACE2-RBD viral complex and dimeric ACE2(single chain)-RBD complex are shown. The localized fluctuation profile of the amino acids across the protein were similar for both simulation conditions except higher fluctuation in the region that constitutes the ACE2 dimerization interface and following C terminal helix which interacts with B0AT1 (B0AT1 was deleted in the monomeric ACE2-RBD simulation [A]. The contribution of the amino acids to the principal component (PC) 1 and 2 (black and red respectively) along with the scree plot reporting the proportion of variance by each PC are shown [B and C]. It can be observed that the N terminal amino acids consisting of the ACE2-RBD binding site contributed more to the overall essential dynamics in dimeric compared monomeric ACE2 simulated in the present study.

Figure 3. The overall interaction of host protein and SARS CoV-2 entry to cell

The figure summarises various factors involved that can influence the entry of Virus in host cell. It depicts brief mechanisms that may arise due to natural occurring variations influencing expression of host genes or structural changes affecting interactions within host proteins or viral proteins resulting in effect on efficiency of virus entry in host cells which could be key factor is providing differential clinical outcomes in different population groups.

Figure 4. Expression of host genes in Lungs, alternative transcripts of *ACE2* and expression data depiction and regulatory elements and other annotations related to ACE2

(a) Expression of various human genes in Human Lungs from GTEx portal. Normalized Expression Values as Transcripts per million (TPM) are plotted as shades from yellow to deep blue in low to high order. (b) Expression levels as median read count per base as shades of blue from low to high values indicating exons expression in different tissues in GTEx portal. Gene models of alternate transcripts are also depicted. (c) Expression levels as median read count per base as shades of red from low to high values indicating exons

junctions expression in different tissues in GTEx portal. (d) UCSC genome browser showing various tracks at 5' Untranslated region of ACE2 gene, common dbSNPs 153 intersecting with conserved transcription factors and conserved transcription factor binding sites are also depicted.

Figure 5. The screenshot from Human Protein Atlas depicting Human ACE2 expression

RNA and Protein expression in different tissues is depicted. In lungs, only RNA expression of ACE2 gene is shown with no ACE2 protein expression.

Figure 6. The screenshot from GTEx portal depicting eQTLs of ACE2 Gene

eQTLs in different tissues were plotted through (a) GTEx IGV Browser as well as (b) GTEx Locus Browser. (a) Red dots indicate eQTLs that showed up in the region against the query. Size of the dot in (b) indicate level of significance (as negative p values) whereas colour depicts positive or negative correlation with Normalized effect size (NES) of the eQTL from -1 to 0 to 1. Red color shades represent upregulation whereas blue color shades show downregulation. (b) also depicts Linkage disequilibrium (LD) with value range from 0 to 1 as white to black shades with 1(dark) as absolute LD.

Figure 7. Annotation of *ACE2* depicting various regulatory elements in UCSC Genome browser

UCSC genome browser showing SNPs rs1978124 and rs2106809 in intron 1 of ACE2 gene. Various tracks covering first 2 exons and 5' Untranslated region of ACE2 gene are shown. The ACE2 gene is located on negative strand. In the region GH0XJ015596 enhancer (in red colour) can be located in the track Enhancer and promoter from GeneHancer database.

Figure 8. Frequency distribution of SNPs rs1978124 and rs2106809 in 1000G population groups

Frequency distribution of SNPs, rs1978124 and rs2106809 in 1000G populations dataset, also depicted on world map by a web based tool Datawrapper (https://www.datawrapper.de/). Derived allele frequencies in both the SNPs are also plotted for comparison, indicating differential frequency distribution in different population groups. SNPs showed relatively uniform frequency distribution trend in sub populations belonging to same super population groups EUR and EAS. However, differences can be seen amongst the AFR, AMR and SAS sub population groups.

Figure 9. UCSC Genome browser screenshot depicting proxy SNPs and LD structure in the *ACE2* region

UCSC genome browser showing r^2 plot along with other annotations for SNPs rs1978124 and rs2106809 from the intron 1 of ACE2 gene. GeneHancer regulator elements and interaction between GeneHancer regulatory elements as curved lines can be seen. In addition, it can be seen that all the proxy SNPs (in Brown)

are clustered together within potential functional regulatory enhancer element and interaction region (in brown color).

Figure 10. Expression of TMPRSS2 in different tissues

(a) Expression of TMPRSS2 in different tissues as reterived from GTEx portal. Transcripts per million (TPM) are plotted on Y axis and different tissues on X axis. (b) RNA and Protein expression in different tissues is depicted in Human Protein Atlas online portal. In lungs, only RNA expression of TMPRSS2 gene is noted with no protein expression depiction. (c) Expression levels as read count in shades of purple from low to high values indicating expression of different transcripts in different tissues in GTEx portal. Gene models of alternate transcripts are also depicted.

Figure 11. The screenshot from GTEx portal depicting eQTLs of TMPRSS2 gene

eQTLs in different tissues were plotted through (a) GTEx IGV Browser as well as (b) GTEx Locus Browser. (a) Red dots indicate eQTLs that showed up in the region against the query. Size of the dot in (b) indicate level of significance (as negative p values) whereas colour depicts positive or negative correlation with Normalized effect size (NES) of the eQTL from -1 to 0 to 1. Red color shades represent upregulation whereas blue color shades show downregulation. (b) also depicts Linkage disequilibrium (LD) with value range from 0 to 1 as white to black shades with 1(dark) as absolute LD.

It was noted [also indicated by blue arrow in (a)] that where in other tissues eQTLs are mainly towards 5' UTR of the gene, in Lungs the eQTLs are towards end of the gene extending towards 3'UTR.

Figure 12. Structure of TMPRSS2 gene, its alternative transcripts and functional domains

(A) TMPRSS2 canonical transcript is constituted of 14 exons and alternate transcripts have been seen. The coded protein is a transmembrane protein with 1-85 amino acids (aa) forming Cytoplasmic Domain, and 112-492 aa constituting Extracellular domain. (B) Expression levels as median read count per base as shades of blue from low to high values indicating exons expression in different tissues in GTEx portal. Gene models of alternate transcripts are also depicted. Interesting to note exons coding for the extracellular domain have higher expression in Lungs (row marked by blue arrow) then other exons.

Figure 13. UCSC Genome browser screenshot depicting Lung eQTLs along with SNPs and LD structure in the *TMPRSS2* region

Lung eQTLs that overlapped with the common dbSNPs are shown and along r^2 values depicting LD in the region is plotted. All the SNPs appeared to cluster towards 3' end of the gene. Other annotations include GeneHancer regulator elements and interaction between GeneHancer regulatory elements as curved lines can be seen. (dark blue color for TMPRSS2 gene). Alternate splicing graph is also indicated.

Figure 14. Elucidation of key functional SNPs of *TMPRSS2* and their Frequency distribution in 1000G population groups

(a) TMPRSS2 gene eQTLs in lungs appear to cluster in the gene towards 3' end and potentially are associated with expression of alternative transcripts in lungs. (b) Amongst these common eQTLs, HaploReg annotations indicated rs4818239 as an important SNP. (c) Frequency distribution on world map of rs4818239
(d) rs75603675 (p.Gly8Val) (e) rs12329760 (p.Val197Met)
The frequencies of SNPs are based on 1000G populations dataset and depicted on world map by LDpop tool of web based LDlink 4.0.3 suite from National Cancer Institute, USA. Derived allele frequencies from low to high are plotted as shades of white to blue.

Figure 15. Sequence alignment and structure of TMPRSS2

(A) The pairwise sequence alignment shows the N-terminal difference in two isoforms and highlights two mutation regions. (B) The predicted 3D structure of TMPRSS2 protein. Two domains of TMPRSS2 and mutation residue are highlighted in the carton model with different colors.

Figure 16. *In silico* protein-protein docking analysis of TMPRSS2.

(A1) Surface Model (A2) Carton Model of TMPRSS2 (Pink) human protein interaction with ACE2 (Green) of SARS CoV-2. (B1) Surface Model (B2) Carton Model of Docked complex of SARS CoV-2 S-protein protein (Blue) with TMPRSS2 (Pink). (C) Comparison of docking score between wild-type and mutant TMPRSS2 with ACE2 and SARS CoV-2 S-protein.

Figure 17. Summarised information about SLC6A19 gene and its variants

(a) Expression of TMPRSS2 in different tissues as retrieved from GTEx portal indicates very restricted expression of the gene. Expression levels as read count in shades of purple from low to high values. Gene models of alternate transcripts are also depicted.(b) Expression levels as median read count per base as shades of blue from low to high values indicating exons expression in different tissues in GTEx portal. Gene models of alternate transcripts are also depicted. (c) SLC6A19 RNA and Protein expression in different tissues is depicted in Human Protein Atlas online portal indicating SLC6A19 has very restricted expression in tissues and in lungs neither its RNA nor Protein expresses. (d) UCSC genome browser screenshot highlights a prominent Enhancer element GH05J001199 at 5'UTR of the gene overlapping exon 1 and extending in intron 1 of the gene. (e) Screenshot from UCSC genome browser depicting SNPs cluster overlapping and intersecting enhancer region GH05J001199. It also depicts common variants (with applied filter on track for SNPs with frequency greater than 20 percent in global populations) in dbSNP version 151. (f) eQTLs in different tissues for *SLC6A19* were plotted through GTEx Locus Browser. Size of the dot indicate level of significance (as negative p values) whereas colour depicts positive or negative correlation with Normalized effect size (NES) of the eQTL from -1 to 0 to 1. Red color shades represent upregulation whereas blue color

shades show downregulation. It also depicts Linkage disequilibrium (LD) in region with value range from 0 to 1 as white to black shades with 1(dark) as absolute LD. Significant upregulation of the gene by eQTLs was observed in Pancreas whereas downregulation in Liver and no data on Lungs as it is reported not to be expressing in Lungs at GTEx portal.

Figure 18. Host-Pathogen Interaction Network and their significant hubs.

(A) The network view of HPIN imported from Cytoscape. The Viral proteins, Human proteins and TFs are represented as nodes and edges denote the physical interaction. All the nodes of viral proteins (red), human proteins (blue) and TF (green) are filled triangles, circles and V-shaped respectively. The edges between Virus-human proteins are shown in orange-headed arrows, PPIs in grey lines and TF-human protein in pink arrow-headed (activators) and flat-headed (Inhibitors). The significant existence of sparsely distributed few main hub proteins, namely RPS6, NACA, HNPRNPA1, BTF3 and SUMO1 are colored as dark blue in the network were represented in the order of four enlarged sized circles.

(B) The Module and Motif constructed and analyzed using MCODE. All the nodes and edges of module & motif are in blue and gray color, respectively. The significant hubs, present in module & motif highlighted in dark blue color.

Figure 19. Number of diseases associated with hub proteins.

(A) The network view where nodes represent diseases and hub proteins and edges the association between them. (All the nodes of diseases rectangle (red), diseases highly associated with hub proteins rectangle (light green) and hub proteins circle (cyan) are filled color and edges in lines (grey).

(B) The pie chart graph showing the percentage distribution of each hub proteins associated with diseases.

Figure 20. Topological characteristics of the HPIN and hub removal.

(A) The figure illustrates the network properties, such as in-degree, out-degree, clustering coefficient, neighbourhood connectivity in the HPIN. The In-degree ($P(k)_{In}$) and out-degree ($P(k)_{out}$) distribution, Clustering Co-efficient C(k), Neighbourhood connectivity $C_N(k)$ is fitted to the power law distribution with exponent values (α, β, γ & μ).

(B) Power law distribution with exponent values α, β, γ & μ of HPIN (complete), hub removed proteins network (*RPS6*, *NACA*, *HNRNPA1* and *BTF3)*.

REFERENCES

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al: **A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature* 2020, **579:**270-273.
2. Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, Talavera-López C, Maatz H, Reichart D, Sampaziotis F, et al: **SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes.** *Nature Medicine* 2020.
3. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D: **Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.** *Cell* 2020.
4. Wan Y, Shang J, Graham R, Baric RS, Li F: **Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus.** *Journal of Virology* 2020, **94:**e00127-00120.
5. Kuster GM, Pfister O, Burkard T, Zhou Q, Twerenbold R, Haaf P, Widmer AF, Osswald S: **SARS-CoV2: should inhibitors of the renin-angiotensin system be withdrawn in patients with COVID-19?** *Eur Heart J* 2020.
6. Imai Y, Kuba K, Rao S, Huan Y, Guo F, Guan B, Yang P, Sarao R, Wada T, Leong-Poi H, et al: **Angiotensin-converting enzyme 2 protects from severe acute lung failure.** *Nature* 2005, **436:**112-116.
7. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, et al: **Clinical Characteristics of Coronavirus Disease 2019 in China.** *N Engl J Med* 2020.
8. Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, Bucci E, Piacentini M, Ippolito G, Melino G: **COVID-19 infection: the perspectives on immune responses.** *Cell Death Differ* 2020.
9. Kaiser J: **How sick will the coronavirus make you? The answer may be in your genes.** *Health, Coronavirus, Sciencemagorg* 2020.
10. Miller A, Reandelar MJ, Fasciglione K, Roumenova V, Li Y, Otazu GH: **Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study.** *medRxiv* 2020**:**2020.2003.2024.20042937.
11. Stawiski EW, Diwanji D, Suryamohan K, Gupta R, Fellouse FA, Sathirapongsasuti JF, Liu J, Jiang Y-P, Ratan A, Mis M, et al: **Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility.** *bioRxiv* 2020**:**2020.2004.2007.024752.
12. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, Wen F, Huang X, Ning G, Wang W: **Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations.** *Cell Discov* 2020, **6:**11.
13. Chen J, Jiang Q, Xia X, Liu K, Yu Z, Tao W, Gong W, Han JJ: **Individual Variation of the SARS-CoV2 Receptor ACE2 Gene Expression and Regulation.** *Pre prints* 2020**:**preprints.org > 202003.200191.v202001.
14. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu NH, Nitsche A, et al: **SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.** *Cell* 2020.
15. Yan R, Zhang Y: **Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2.** 2020, **367:**1444-1448.
16. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W: **Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments.** *Journal of Computer-Aided Molecular Design* 2013, **27:**221-234.
17. Mark P, Nilsson L: **Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K.** *The Journal of Physical Chemistry A* 2001, **105:**9954-9960.
18. **Proceedings of the 2006 ACM/IEEE conference on Supercomputing.** In*; Tampa, Florida*. Association for Computing Machinery; 2006
19. Grant BJ, Rodrigues AP, ElSawy KM, McCammon JA, Caves LS: **Bio3d: an R package for the comparative analysis of protein structures.** *Bioinformatics* 2006, **22:**2695-2696.

20. Jacobson MP, Friesner RA, Xiang Z, Honig B: **On the Role of the Crystal Environment in Determining Protein Side-chain Conformations.** *Journal of Molecular Biology* 2002, **320:**597-608.
21. Humphrey W, Dalke A, Schulten K: **VMD: Visual molecular dynamics.** *Journal of Molecular Graphics* 1996, **14:**33-38.
22. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al: **Proteomics. Tissue-based map of the human proteome.** *Science* 2015, **347:**1260419.
23. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.
24. UniProt C: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res* 2019, **47:**D506-D515.
25. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I: **New and continuing developments at PROSITE.** *Nucleic Acids Res* 2013, **41:**D344-347.
26. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al: **The Pfam protein families database in 2019.** *Nucleic Acids Res* 2019, **47:**D427-D432.
27. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM, Brent S, Kulesha E, Marin-Garcia P, et al: **Ensembl variation resources.** *BMC Genomics* 2010, **11:**293.
28. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al: **SWISS-MODEL: homology modelling of protein structures and complexes.** *Nucleic Acids Res* 2018, **46:**W296-W303.
29. Ramachandran S, Kota P, Ding F, Dokholyan NV: **Automated minimization of steric clashes in protein structures.** *Proteins* 2011, **79:**261-270.
30. Chinea G, Padron G, Hooft RW, Sander C, Vriend G: **The use of position-specific rotamers in model building by homology.** *Proteins* 1995, **23:**415-421.
31. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G: **Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces.** *BMC Bioinformatics* 2010, **11:**548.
32. Capriotti E, Calabrese R, Casadio R: **Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.** *Bioinformatics* 2006, **22:**2729-2734.
33. Dominguez C, Boelens R, Bonvin AM: **HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.** *J Am Chem Soc* 2003, **125:**1731-1737.
34. Ammari MG, Gresham CR, McCarthy FM, Nanduri B: **HPIDB 2.0: a curated database for host-pathogen interactions.** *Database (Oxford)* 2016, **2016**.
35. Kumar R, Nanduri B: **HPIDB--a unified resource for host-pathogen interactions.** *BMC Bioinformatics* 2010, **11 Suppl 6:**S16.
36. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res* 2010, **38:**W214-220.
37. Maitland NJ, Frame FM, Polson ES, Lewis JL, Collins AT: **Prostate cancer stem cells: do they have a basal or luminal phenotype?** *Horm Cancer* 2011, **2:**47-61.
38. David CJ, Chen M, Assanah M, Canoll P, Manley JL: **HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer.** *Nature* 2010, **463:**364-368.
39. Tumer E, Broer A, Balkrishna S, Julich T, Broer S: **Enterocyte-specific regulation of the apical nutrient transporter SLC6A19 (B(0)AT1) by transcriptional and epigenetic networks.** *J Biol Chem* 2013, **288:**33813-33823.
40. Zhang R, Wu Y, Zhao M, Liu C, Zhou L, Shen S, Liao S, Yang K, Li Q, Wan H: **Role of HIF-1alpha in the regulation ACE and ACE2 expression in hypoxic human pulmonary artery smooth muscle cells.** *Am J Physiol Lung Cell Mol Physiol* 2009, **297:**L631-640.

41. Barros P, Lam EW, Jordan P, Matos P: **Rac1 signalling modulates a STAT5/BCL-6 transcriptional switch on cell-cycle-associated target gene promoters.** *Nucleic Acids Res* 2012, **40:**7776-7787.
42. Yu YN, Yip GW, Tan PH, Thike AA, Matsumoto K, Tsujimoto M, Bay BH: **Y-box binding protein 1 is up-regulated in proliferative breast cancer and its inhibition deregulates the cell cycle.** *Int J Oncol* 2010, **37:**483-492.
43. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
44. Reichardt J, Bornholdt S: **Statistical mechanics of community detection.** *Physical Review E* 2006, **74:**016110.
45. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4:**2.
46. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M: **Computing topological parameters of biological networks.** *Bioinformatics* 2008, **24:**282-284.
47. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabasi AL: **Disease networks. Uncovering disease-disease relationships through the incomplete interactome.** *Science* 2015, **347:**1257601.
48. Albert R, Barabási A-L: **Statistical mechanics of complex networks.** *Reviews of Modern Physics* 2002, **74:**47-97.
49. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286:**509-512.
50. Ravasz E, Barabasi AL: **Hierarchical organization in complex networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67:**026112.
51. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
52. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296:**910-913.
53. Pastor-Satorras R, Vazquez A, Vespignani A: **Dynamical and correlation properties of the internet.** *Phys Rev Lett* 2001, **87:**258701.
54. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A: **The architecture of complex weighted networks.** *Proc Natl Acad Sci U S A* 2004, **101:**3747-3752.
55. Malik MZ, Chirom K, Ali S, Ishrat R, Somvanshi P, Singh RKB: **Methodology of predicting novel key regulators in ovarian cancer network: a network theoretical approach.** *BMC Cancer* 2019, **19:**1129.
56. Nafis S, Ponnusamy K, Husain M, Singh RK, Bamezai RN: **Identification of key regulators and their controlling mechanism in a combinatorial apoptosis network: a systems biology approach.** *Mol Biosyst* 2016, **12:**3357-3369.
57. Heurich A, Hofmann-Winkler H, Gierer S, Liepold T, Jahn O, Pohlmann S: **TMPRSS2 and ADAM17 cleave ACE2 differentially and only proteolysis by TMPRSS2 augments entry driven by the severe acute respiratory syndrome coronavirus spike protein.** *J Virol* 2014, **88:**1293-1307.
58. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al: **GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.** *Database (Oxford)* 2017, **2017**.
59. Good MC, Zalatan JG, Lim WA: **Scaffold proteins: hubs for controlling the flow of cellular information.** *Science* 2011, **332:**680-686.
60. Clauset A, Shalizi CR, Newman MEJ: **Power-Law Distributions in Empirical Data.** *SIAM Review* 2009, **51:**661-703.
61. Borgatti S, Carley K, Krackhardt D: **On the Robustness of Centrality Measures Under Conditions of Imperfect Data.** *Social Networks* 2006, **28:**124-136.
62. Canright G, Engø-Monsen K: **Roles in networks.** *Science of Computer Programming* 2004, **53:**195-214.

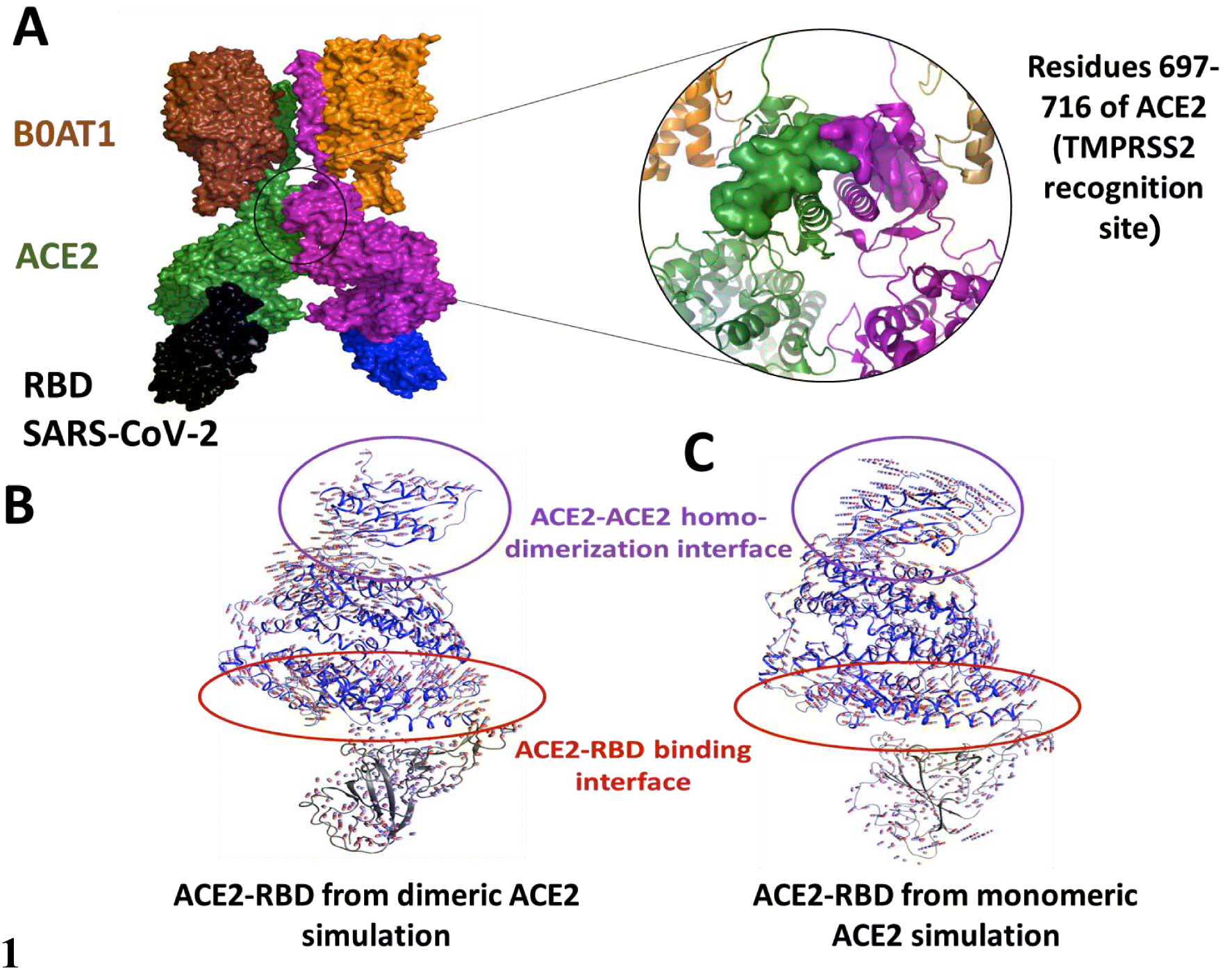


Figure 1

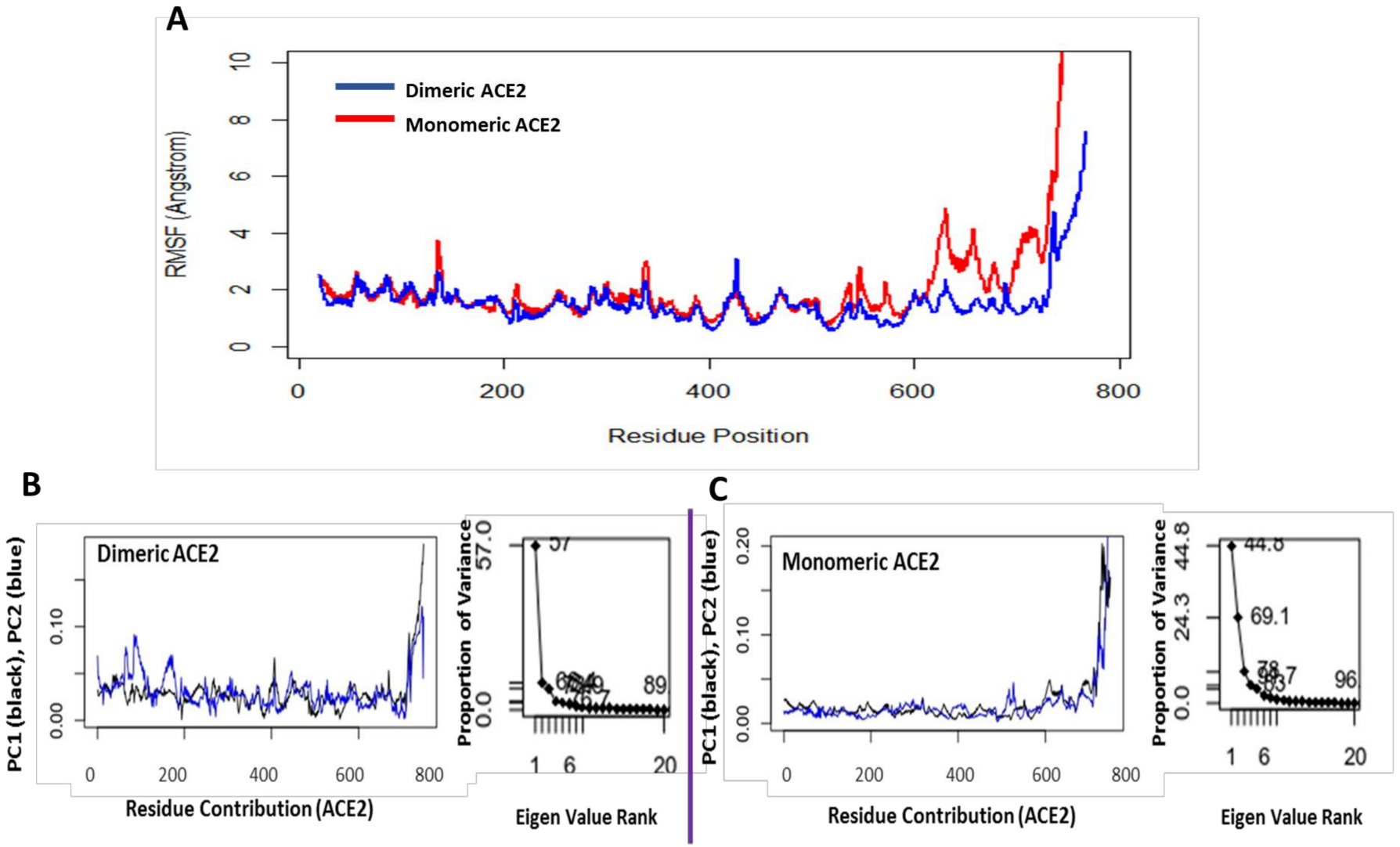


Figure 2

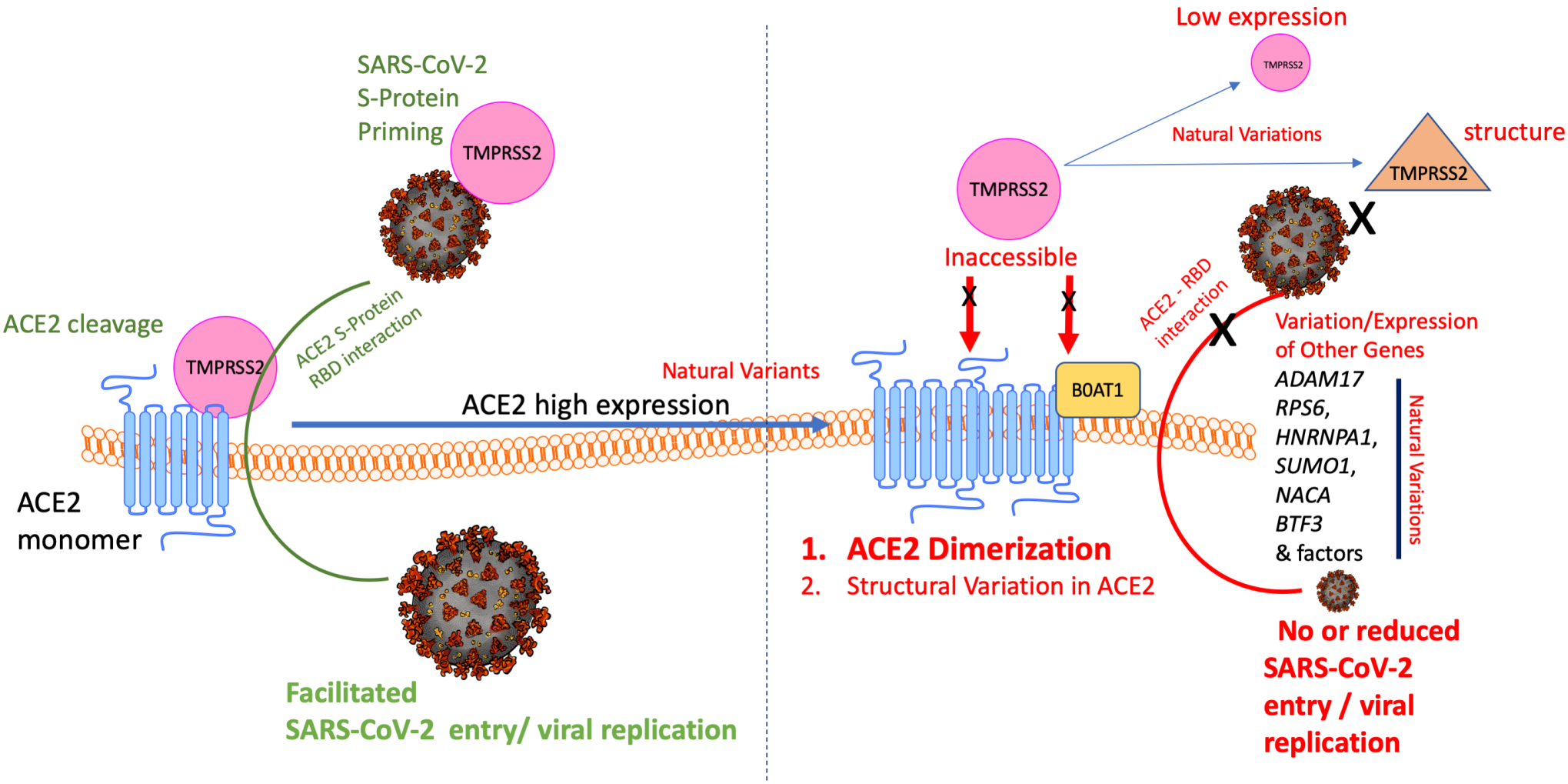


Figure 3

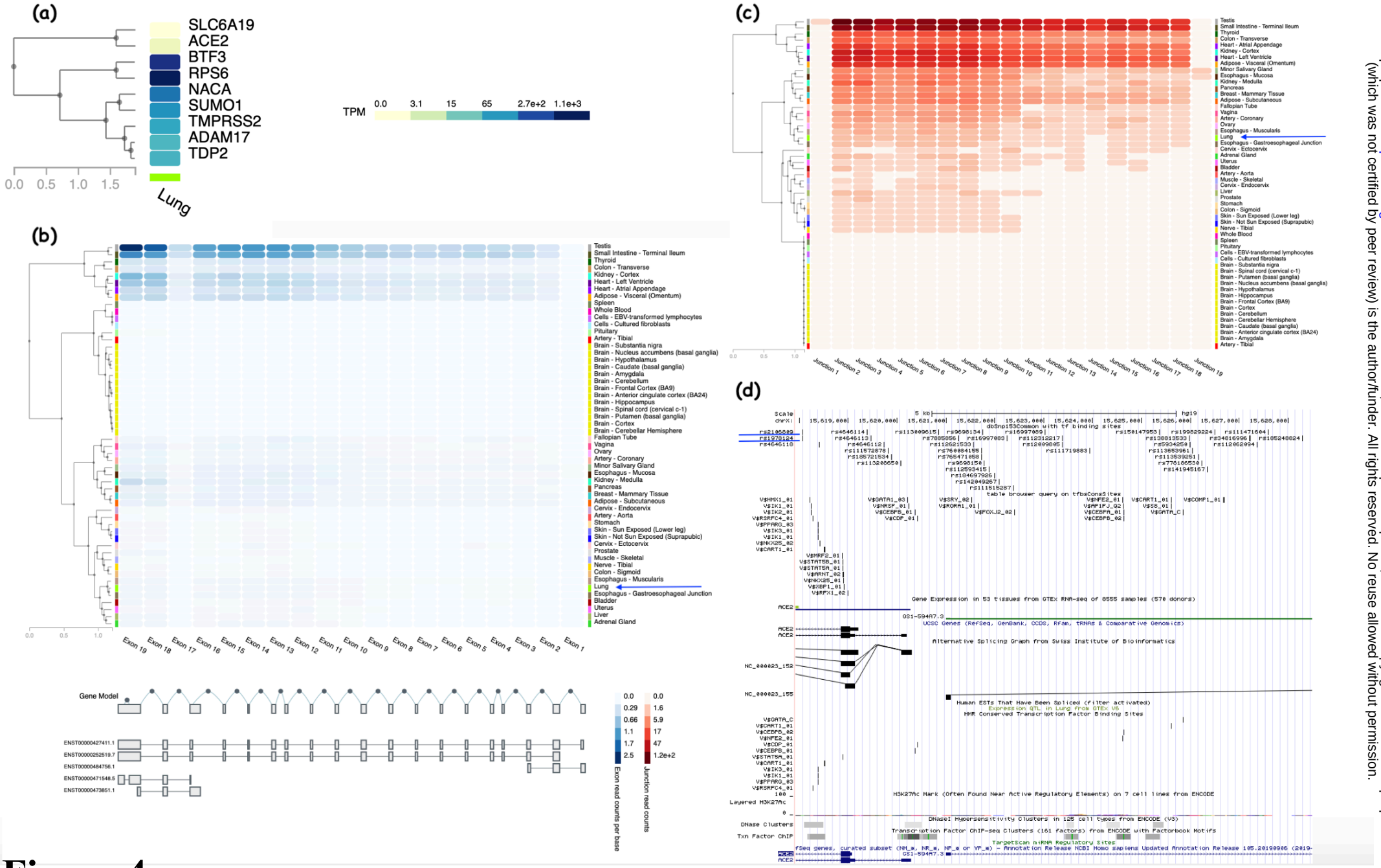


Figure 4

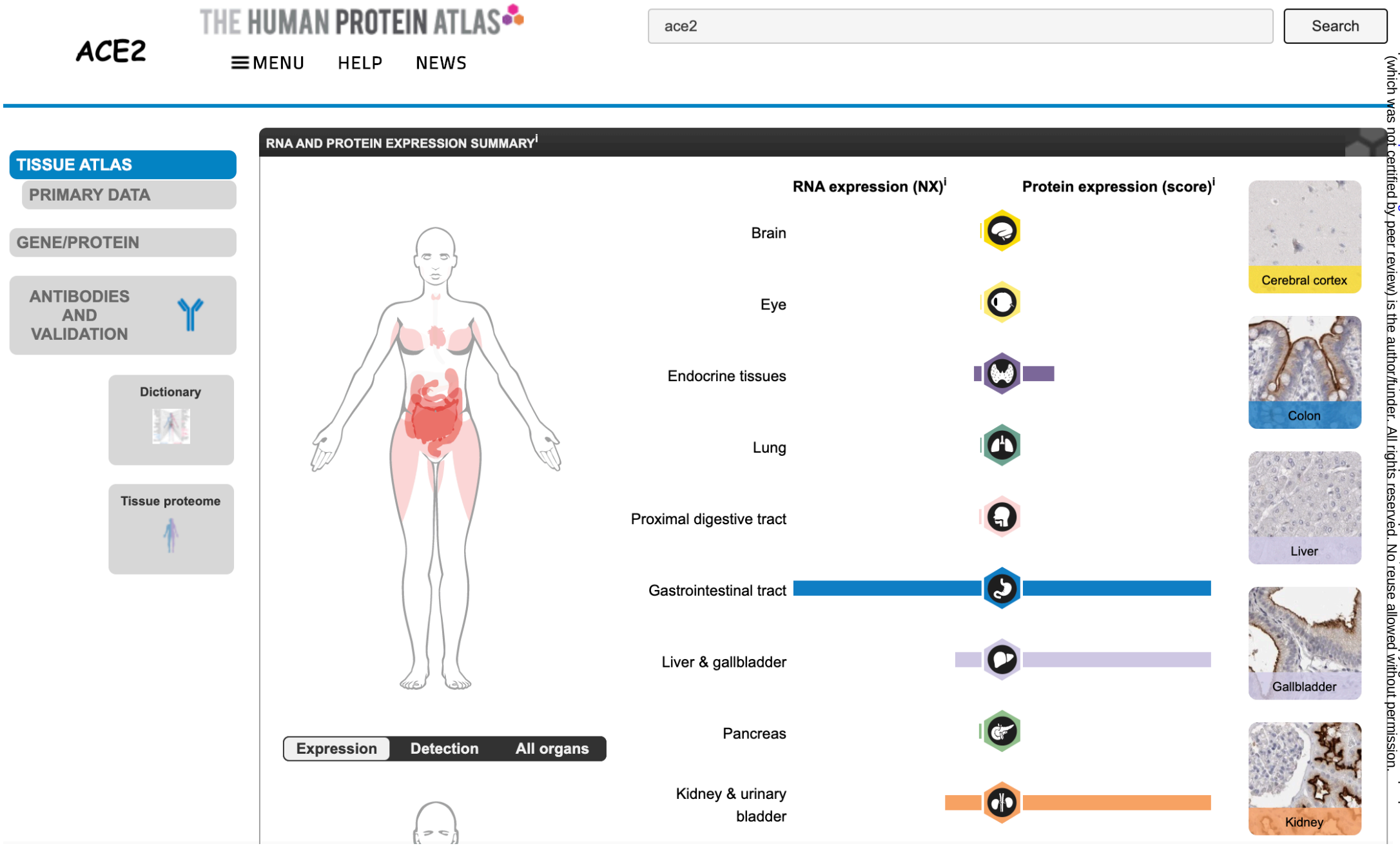


Figure 5

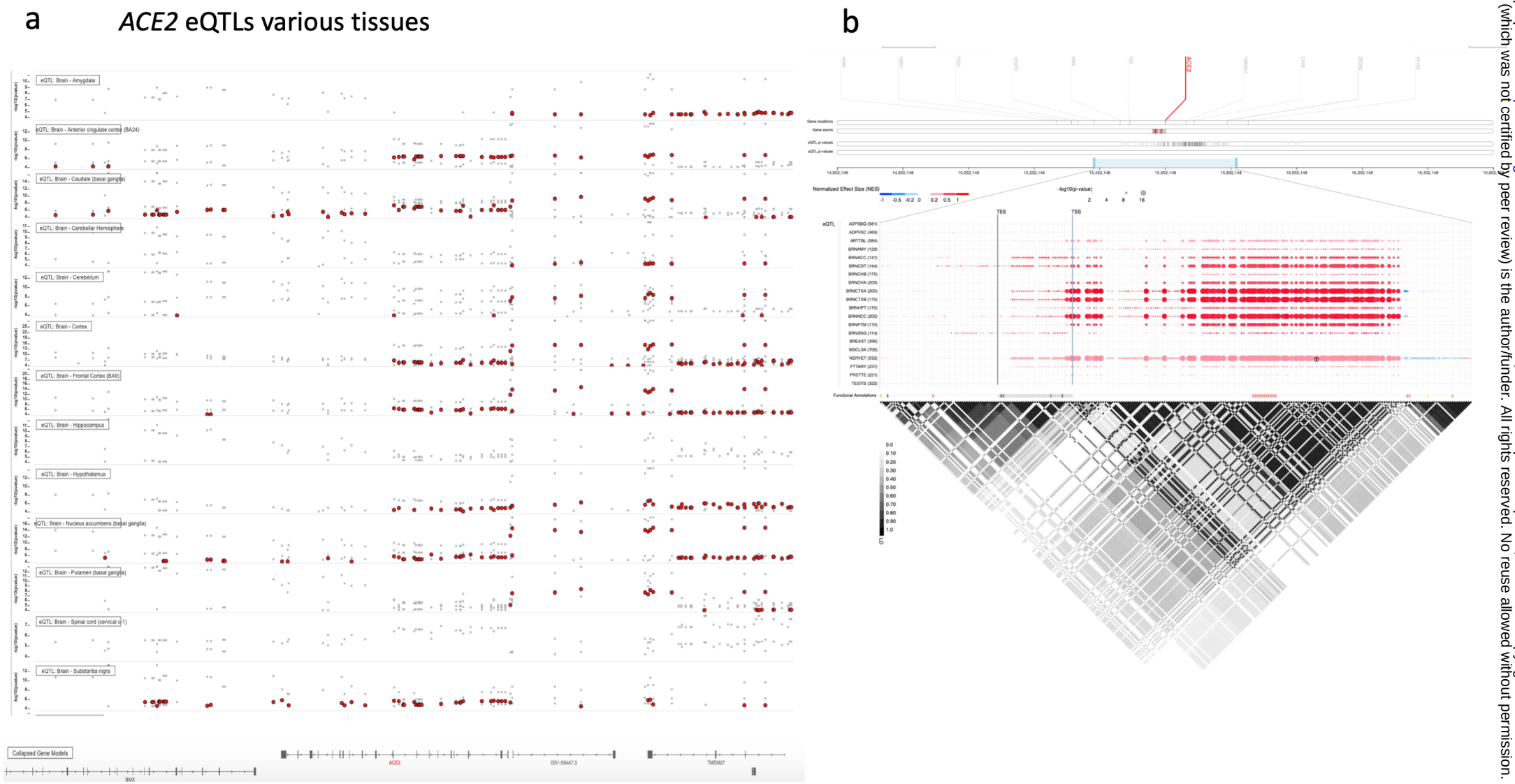


Figure 6

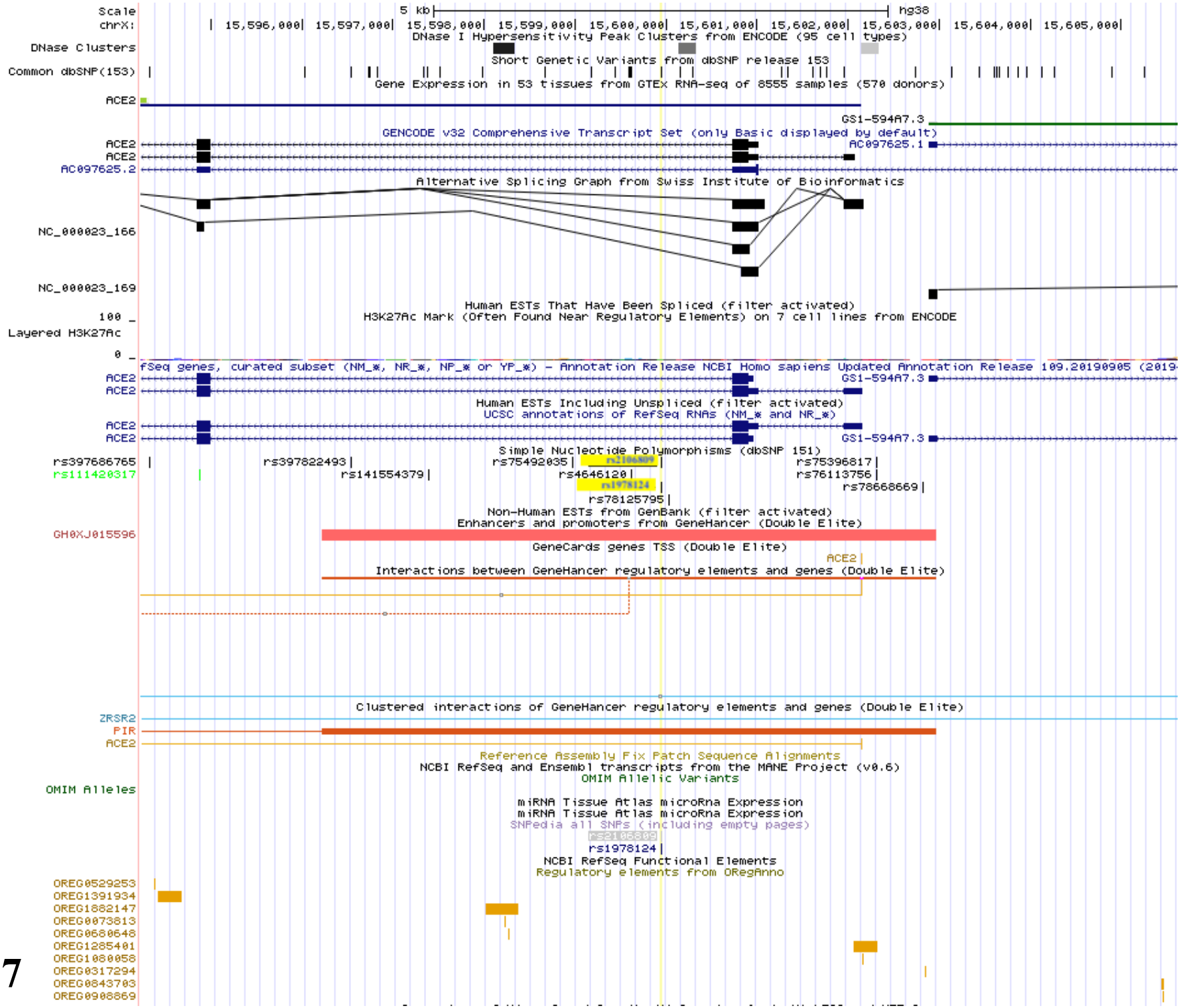


Figure 7

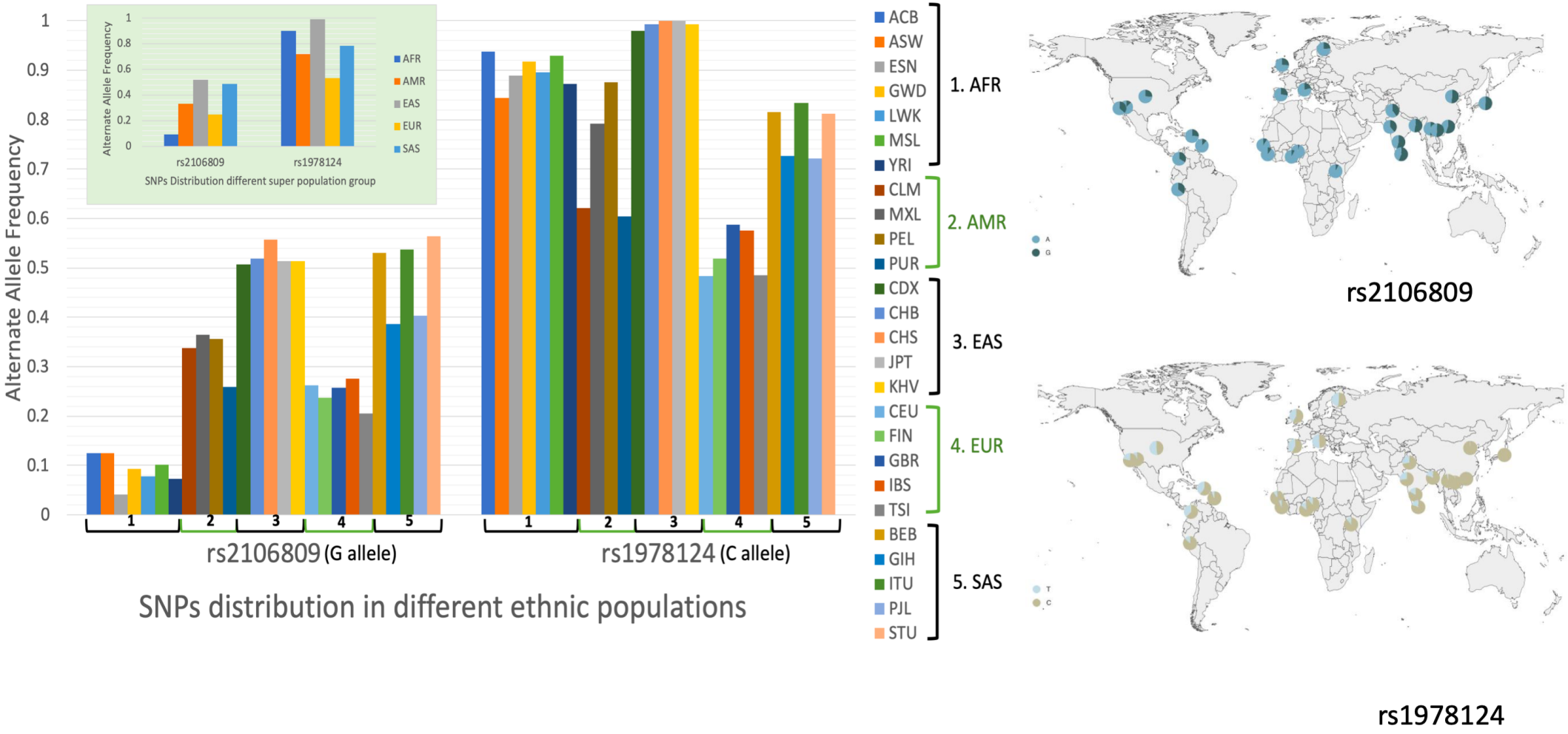


Figure 8

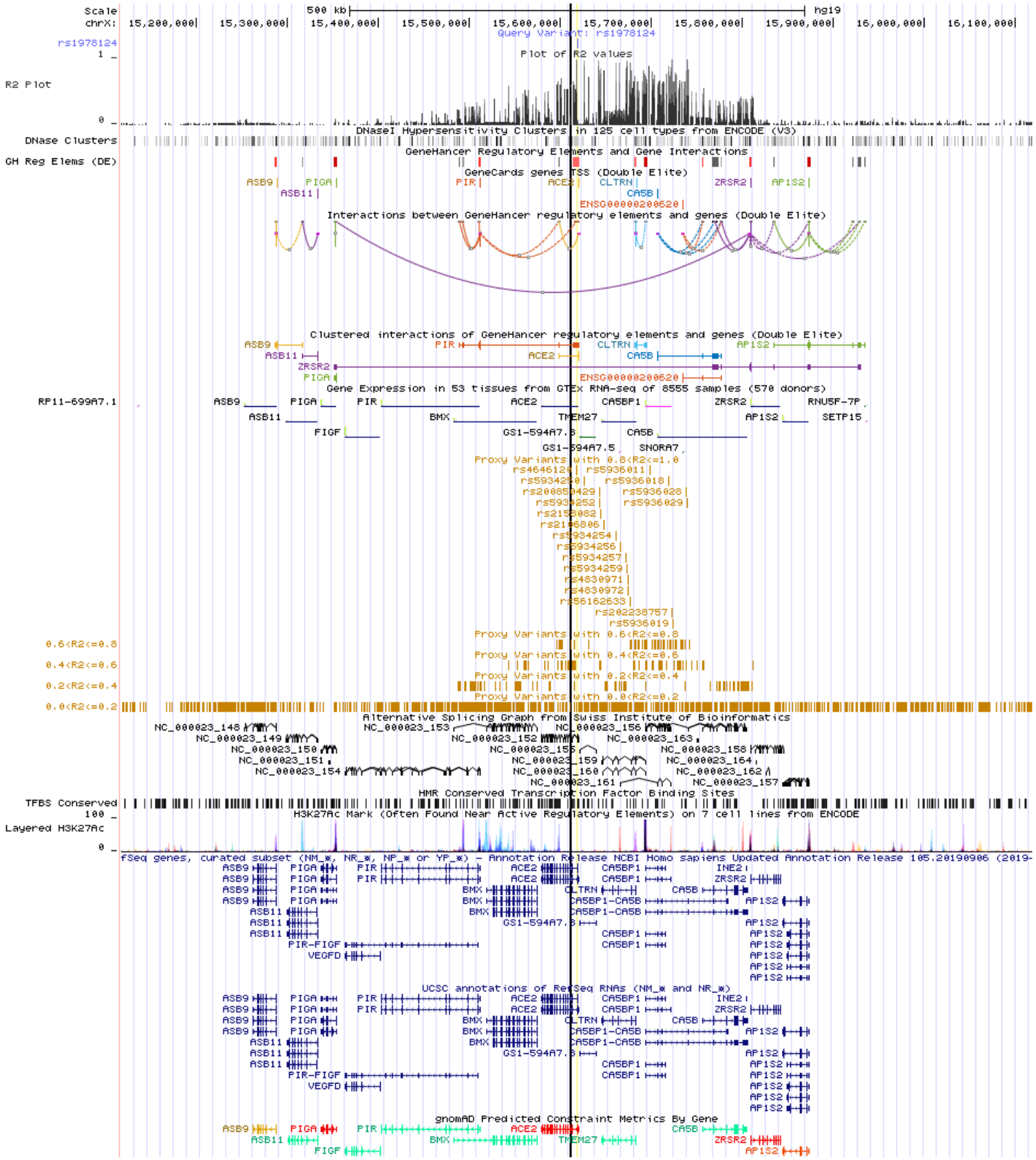


Figure 9

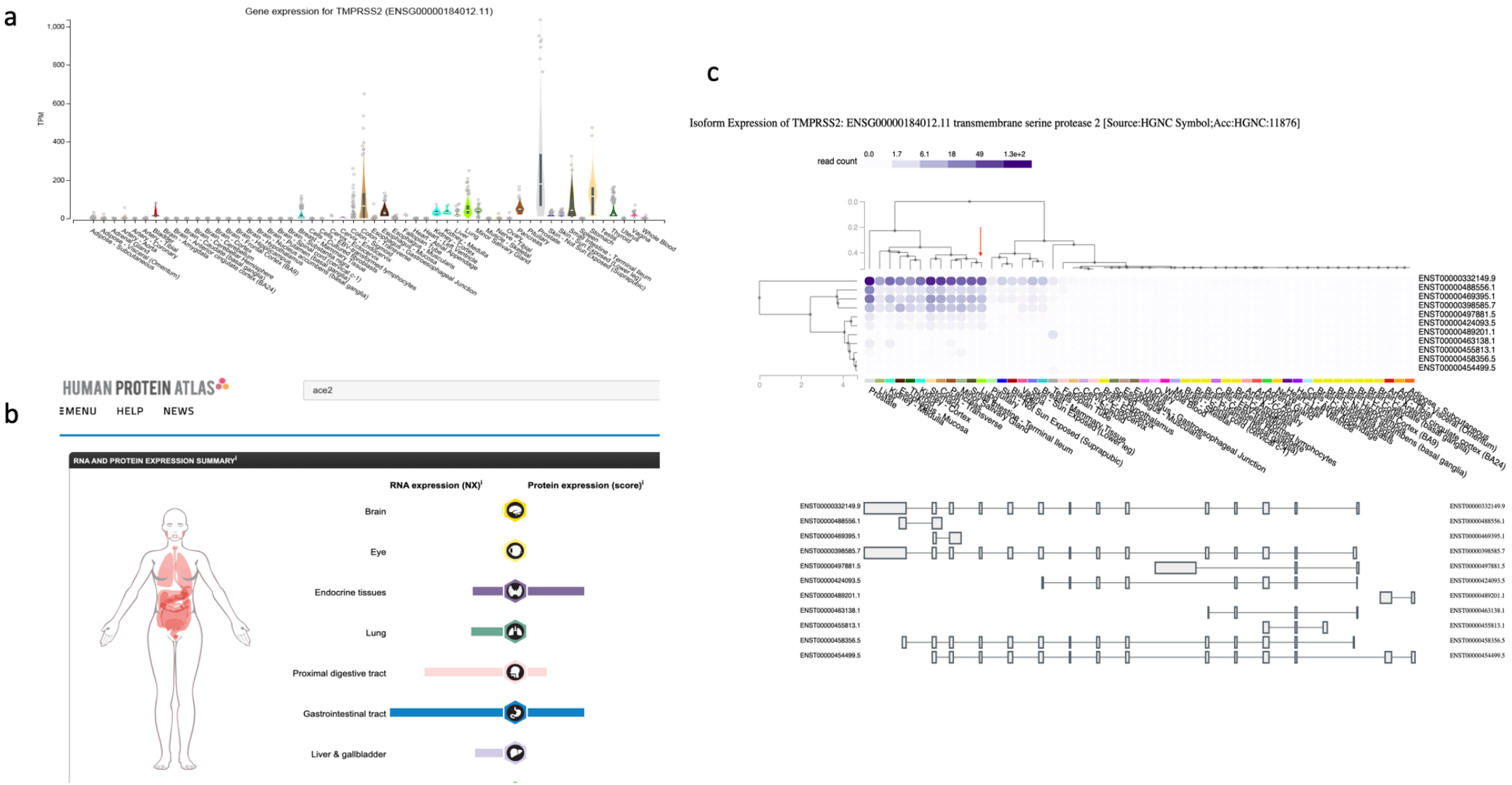


Figure 10

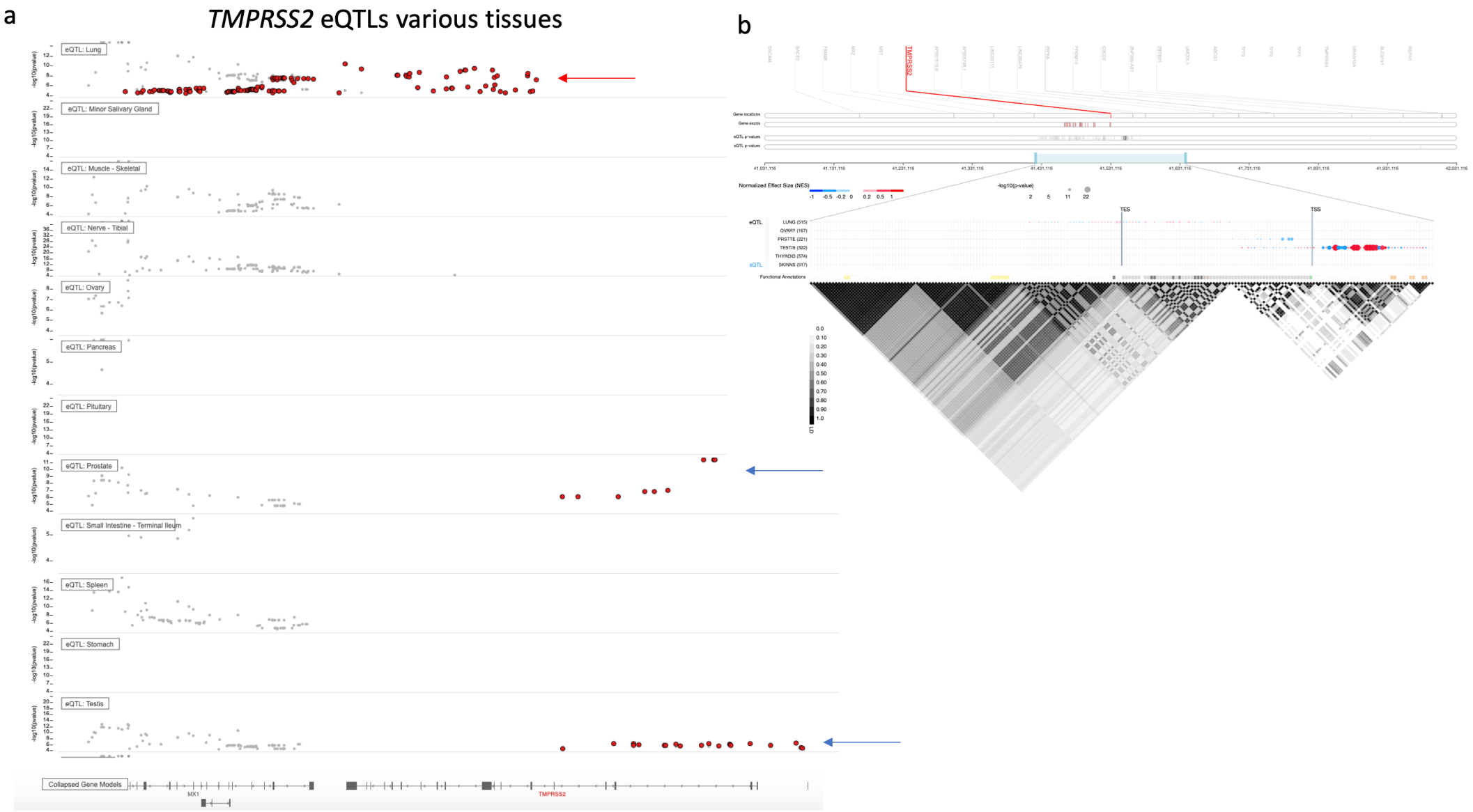
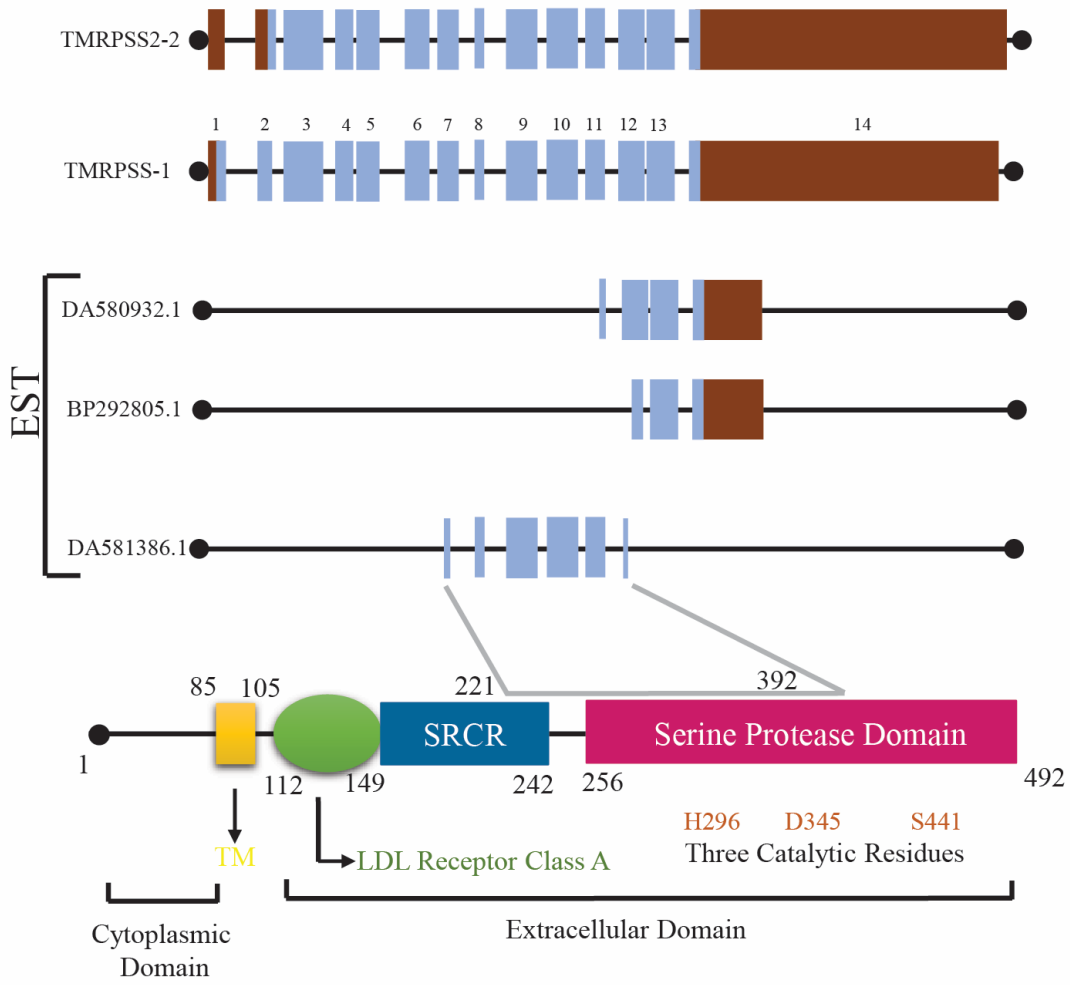


Figure 11

A.



B.

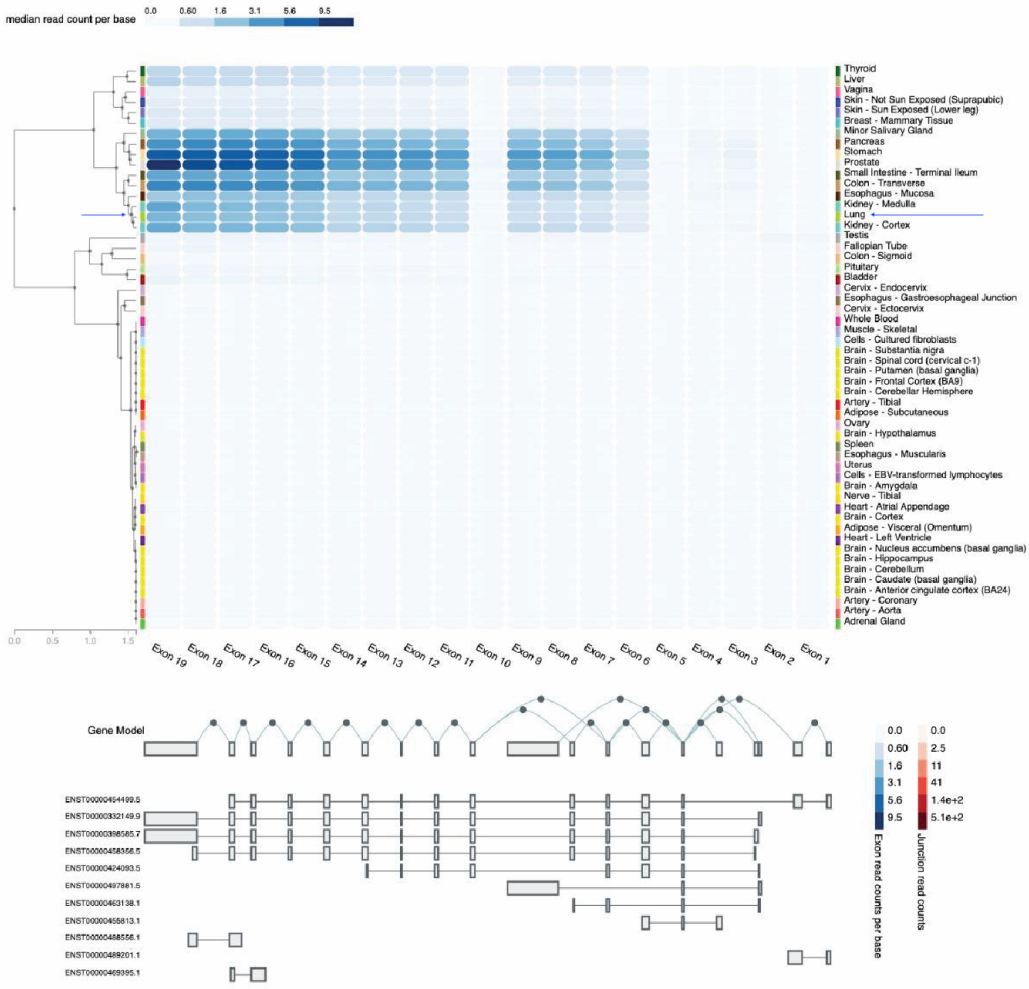


Figure 12

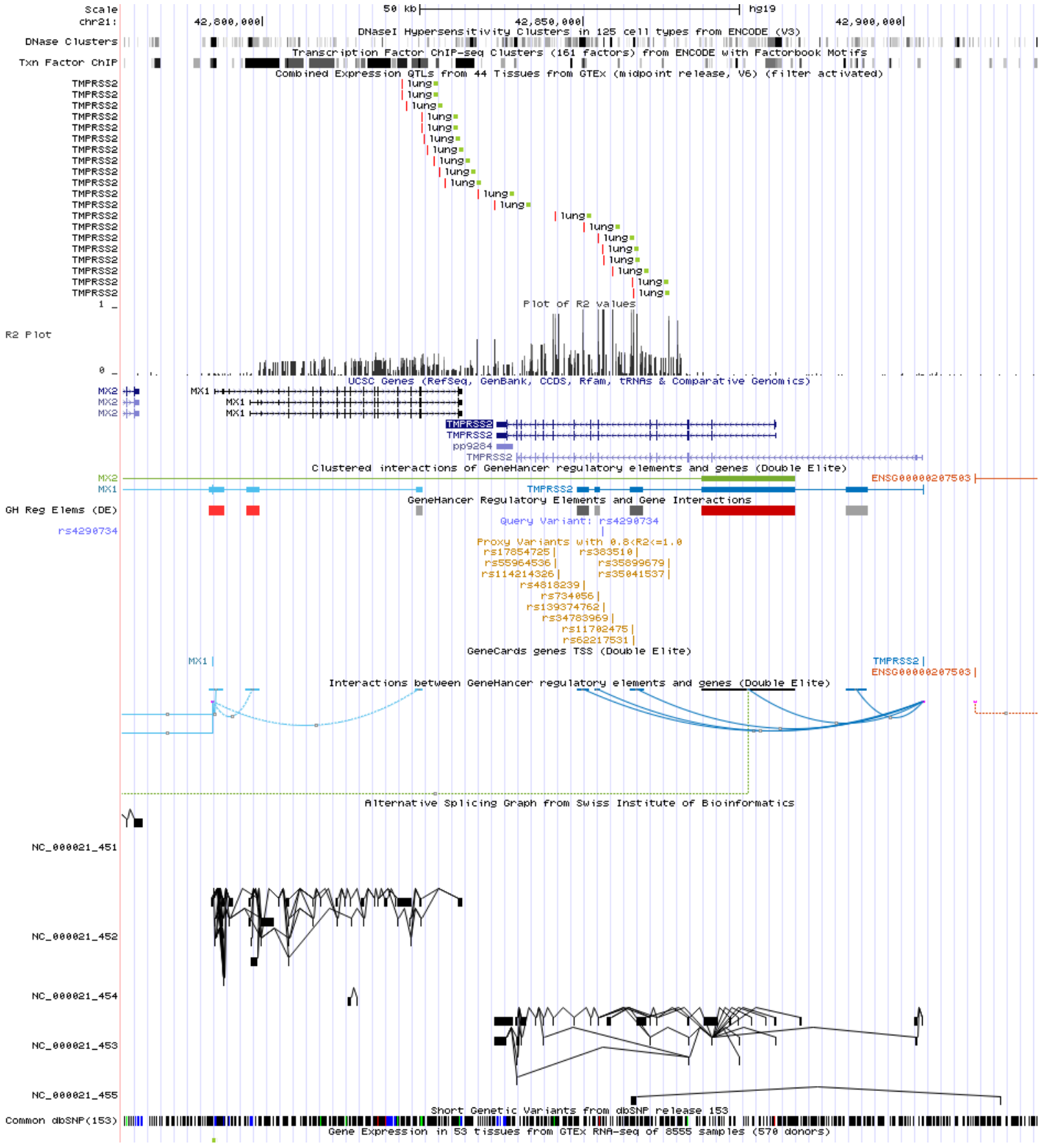
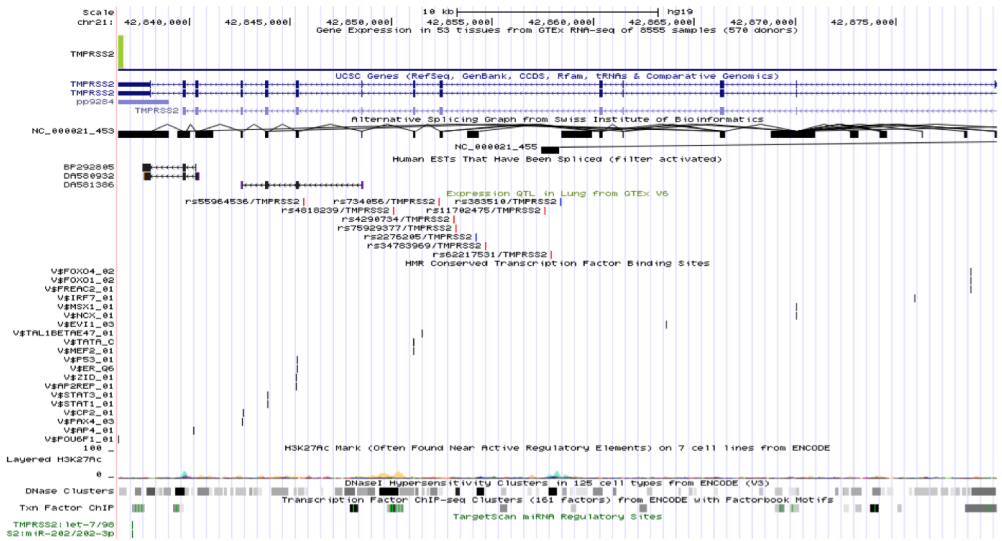


Figure13

a

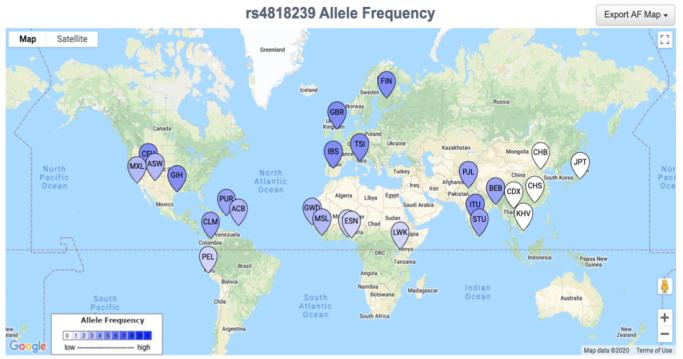


b

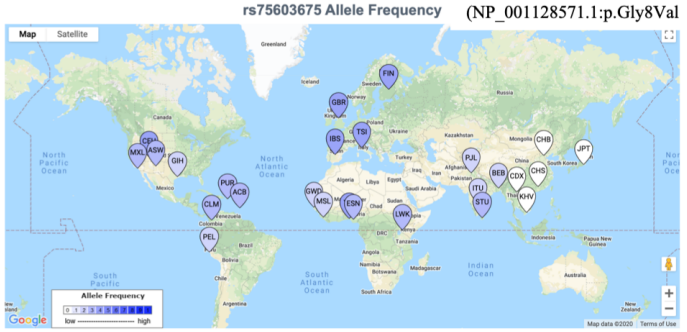
Query SNP: rs4818239 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
21	41478148	1	1	rs4818239	T	C	0.25	0.35	0.01	0.48		BRST, PLCNT	21 tissues	18 tissues	9 bound proteins	4 altered motifs			1 hit	TMPRSS2	intronic
21	41480393	0.93	0.97	rs734056	C	A	0.22	0.33	0.00	0.48			8 tissues	PANC,GI		Gfi1			3 hits	TMPRSS2	intronic
21	41481156	0.93	0.97	rs4290734	A	G	0.09	0.32	0.00	0.48			4 tissues	ESDR,BRN		5 altered motifs			3 hits	TMPRSS2	intronic
21	41481267	0.93	0.97	rs139374762	7-mer	T	0.13	0.32	0.00	0.48			4 tissues	5 tissues						TMPRSS2	intronic
21	41482711	0.92	0.97	rs34783969	A	T	0.14	0.33	0.00	0.48			8 tissues	6 tissues		5 altered motifs			1 hit	TMPRSS2	intronic
21	41485974	0.8	0.93	rs62217531	C	T	0.31	0.33	0.00	0.46			19 tissues	8 tissues		MZF1::1-4,THAP1,YY1			1 hit	TMPRSS2	intronic

c



d



e

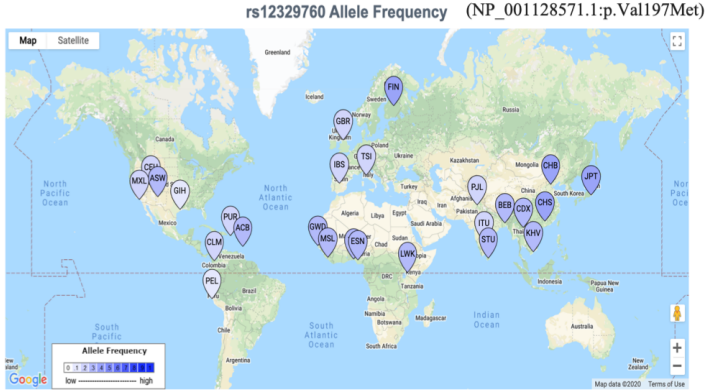
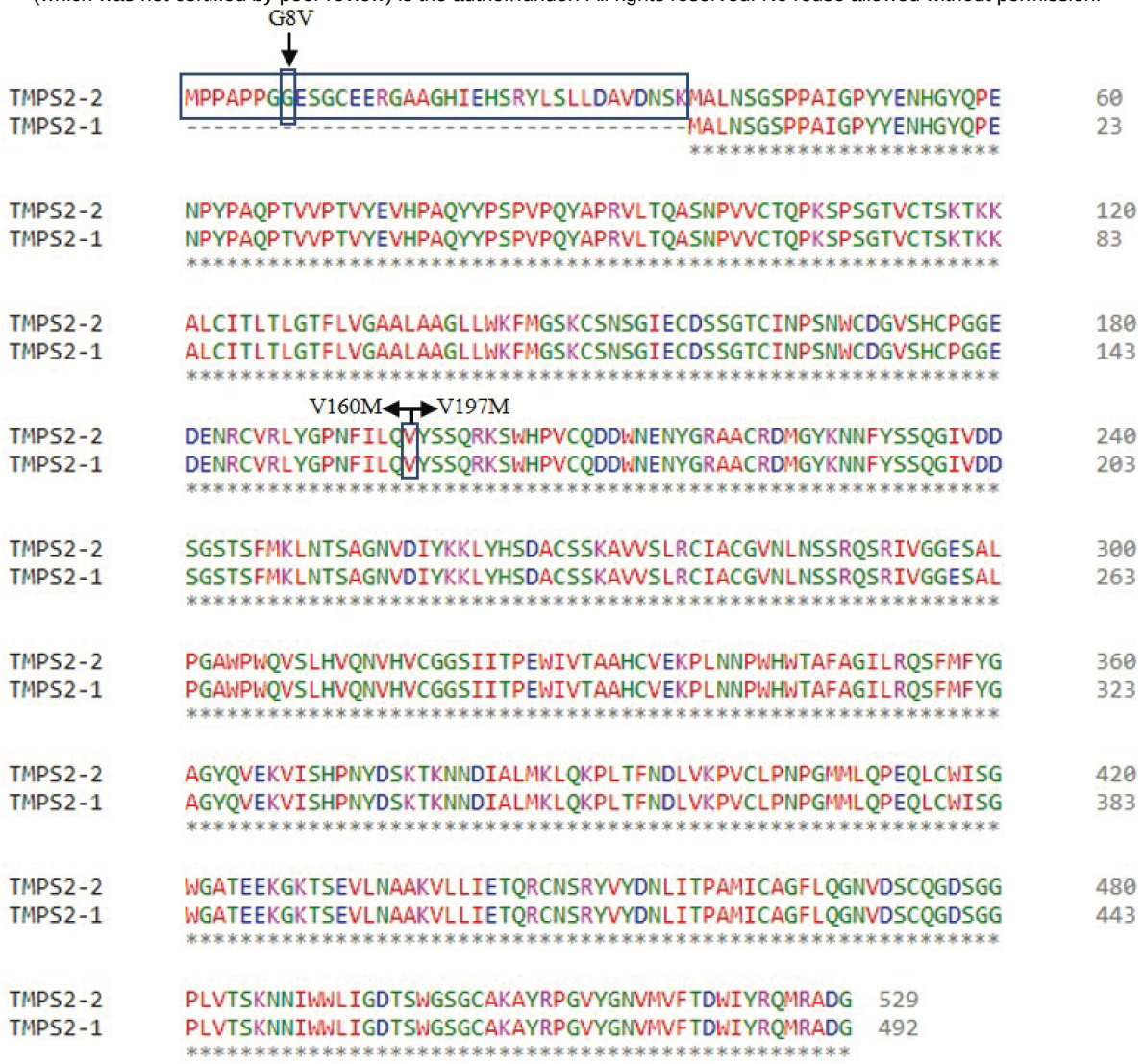


Figure 14

A.



B.

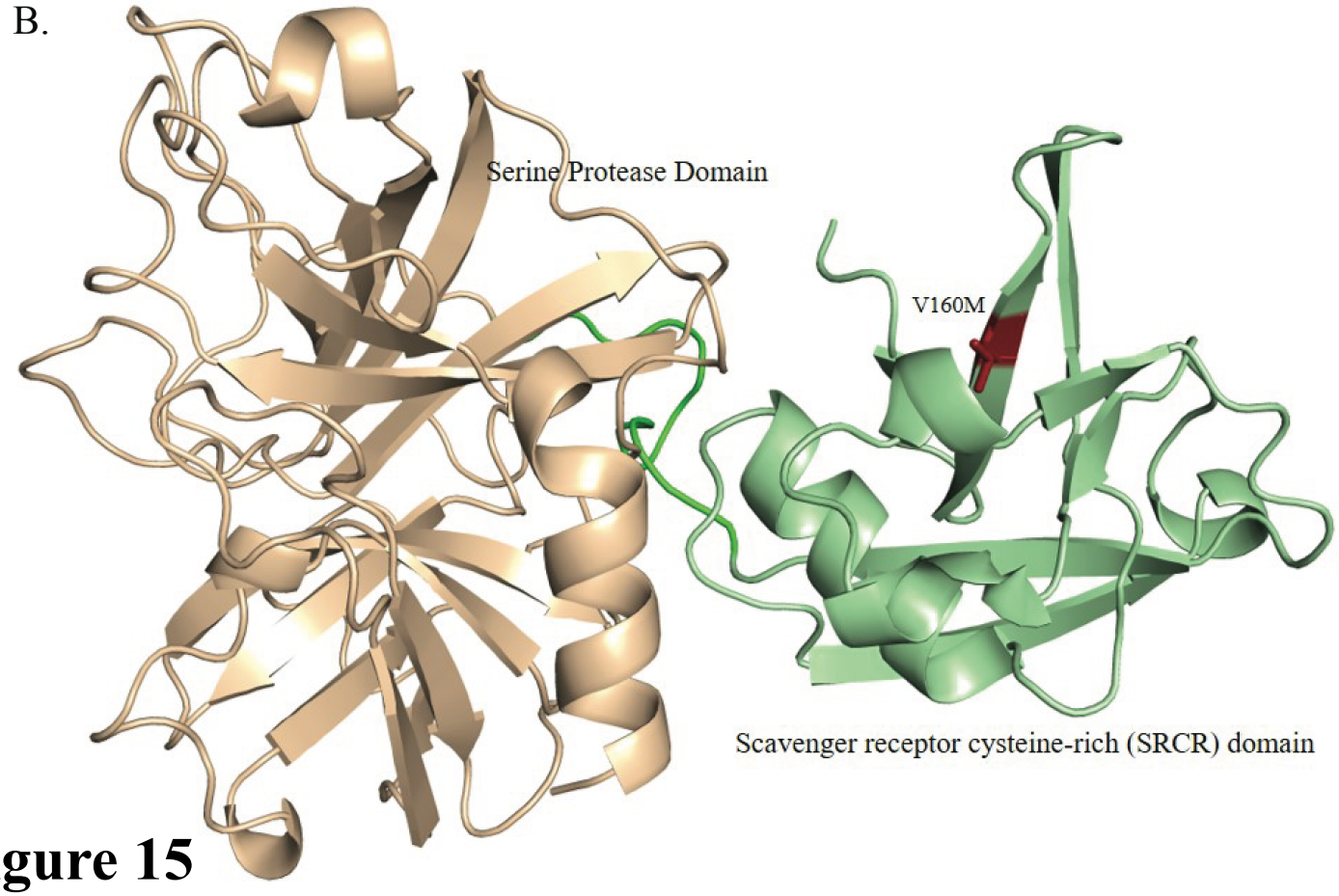


Figure 15

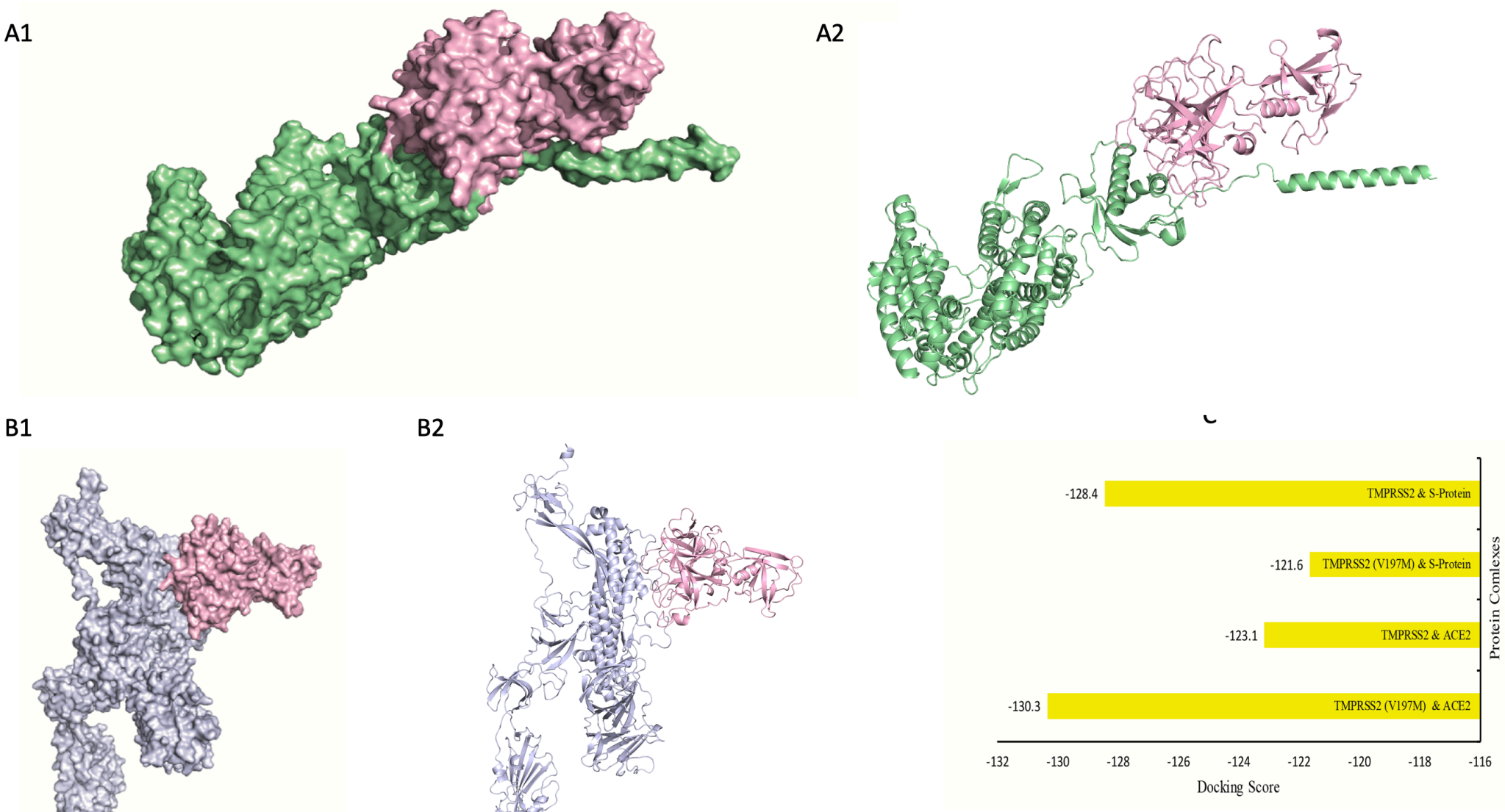


Figure 16

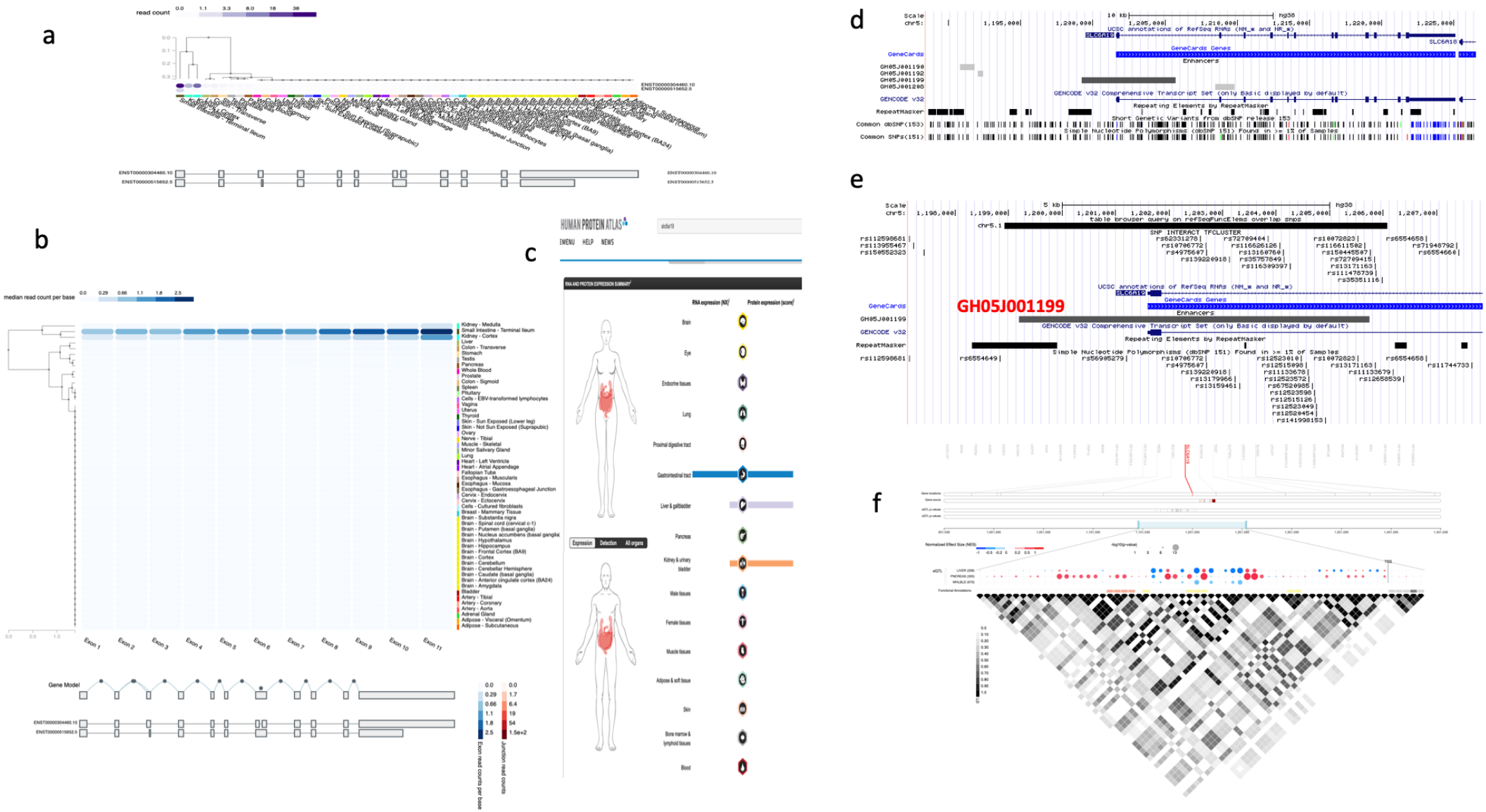


Figure 17

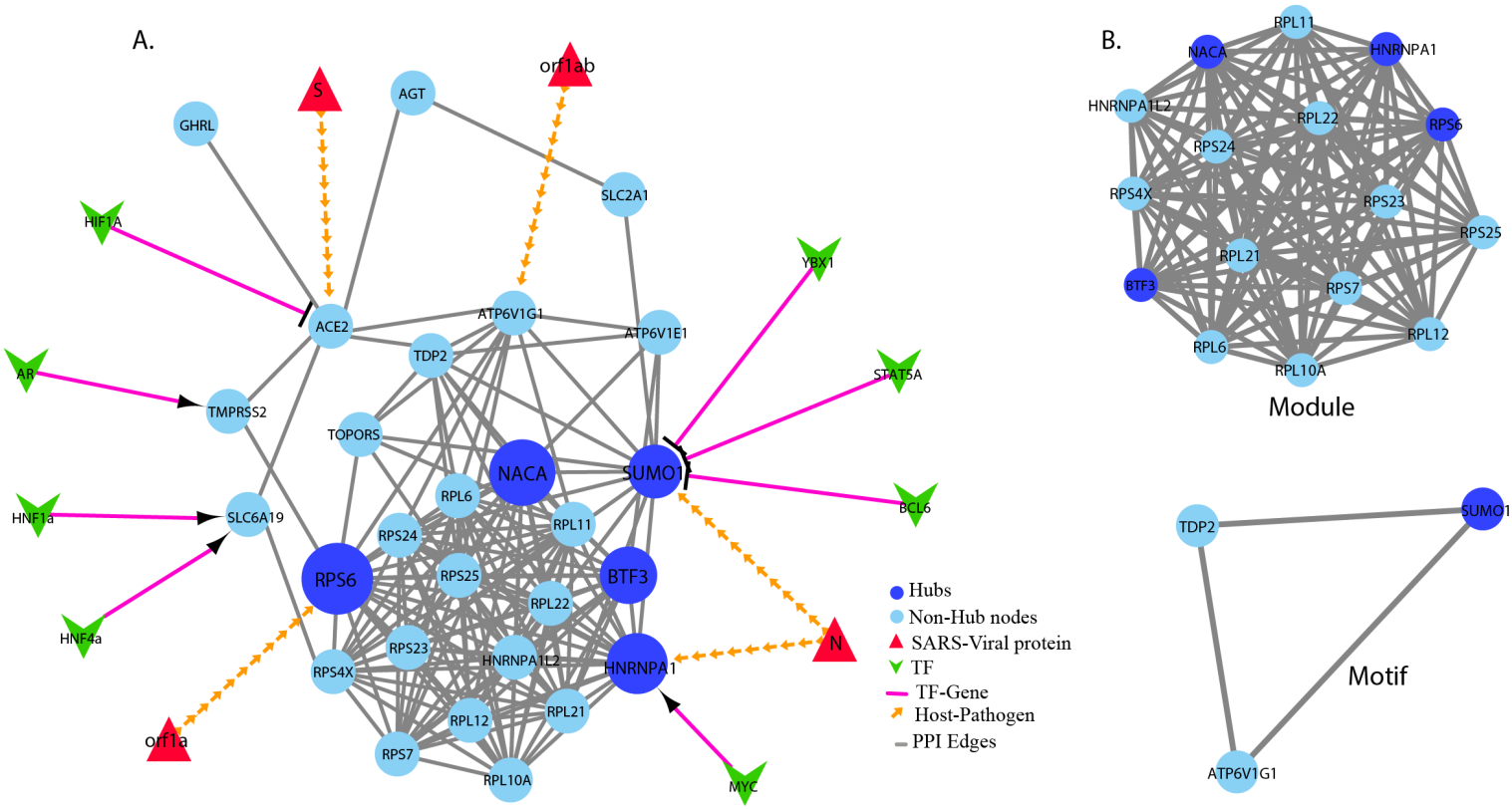


Figure 18

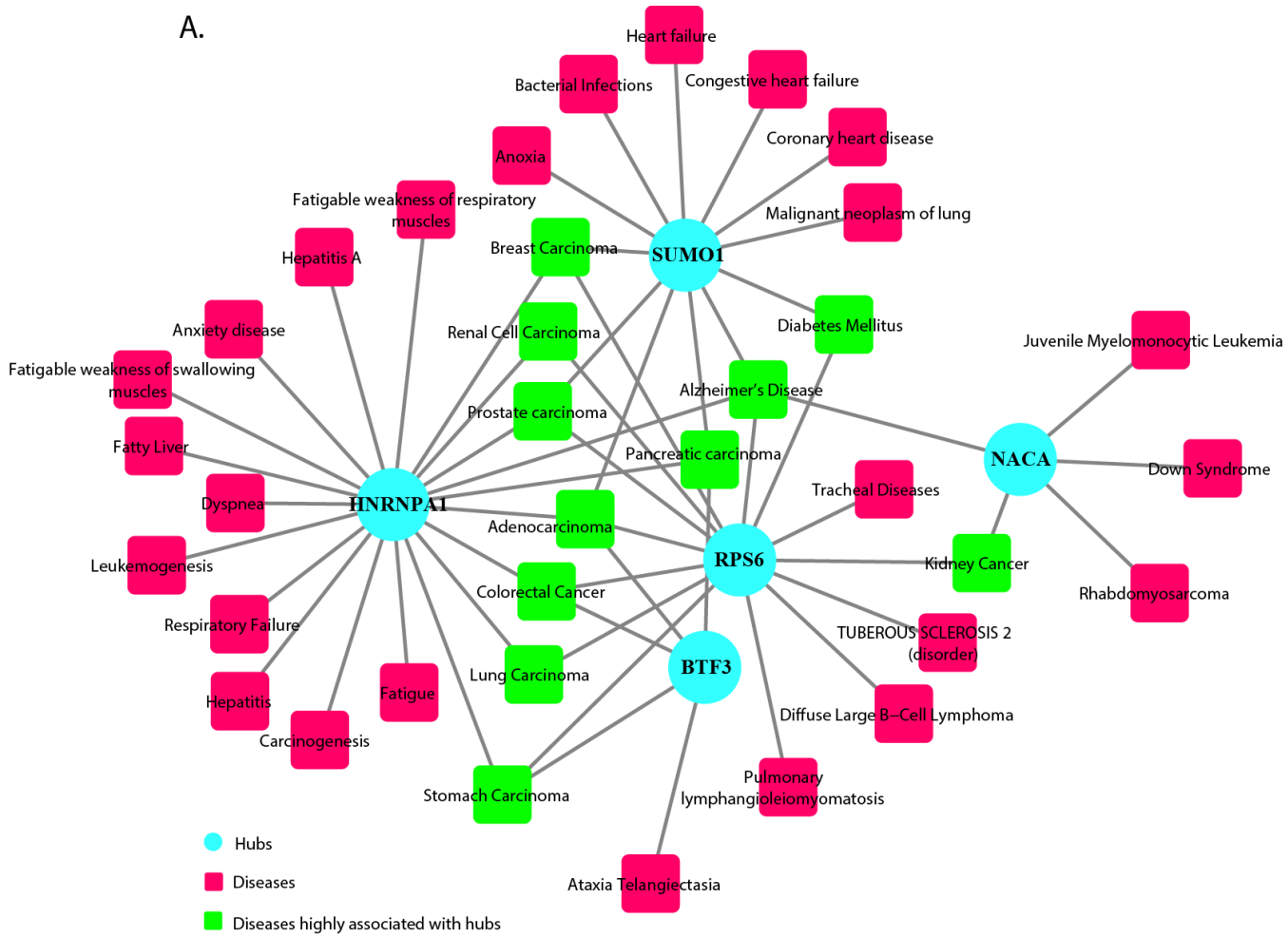


Figure 19

Figure 20