Spatial distribution of coding diseases-associated single-nucleotide variants in complexes of the corresponding protein with other molecules

Alexander Gress[1,2], Vasily Ramensky[3,4,5], Olga V. Kalinina[1]

1. Department for Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Building E1.4, 66123 Saarbrücken, Germany.

2. Graduate School of Computer Science, Saarland University, Campus E1.3, 66123 Saarbrücken, Germany.

3. Center for Neurobehavioral Genetics, University of California, Los Angeles, 695 Charles E. Young Drive South, Los Angeles, CA 90095, USA.

4. Moscow Institute of Physics and Technology, Dolgoprudny, Institusky 9, Moscow, Moscow Region, Russian Federation.

5. Immanuel Kant Baltic Federal University, Aleksandra Nevskogo, Kaliningrad, Kaliningrad Oblast, Russian Federation

kalinina@mpi-inf.mpg.de

Next-generation sequencing enables simultaneous analysis of human hundreds of genomes associated with a particular phenotype, e.g. a disease. These genomes naturally contain a lot of sequence variation that ranges from single nucleotide variants (SNVs) to large-scale structural rearrangements. In order to establish a functional connection between genotype and disease-associated phenotypes, one needs to distinguish disease drivers from neutral passenger mutations. Functional annotation based on experimental assays is feasible only for limited number of candidate mutations. Thus alternative computational tools are needed. A possible approach to annotating mutations functionally is to consider their spatial location with respect to functionally relevant sites in three-dimensional (3D) structures of the harboring proteins. This is impeded by the lack of available protein 3D structures. Complementing experimentally resolved structures with reliable computational models is an attractive alternative.

We developed a structure-based approach to characterizing three comprehensive sets of non-synonymous single-nucleotide variants (nsSNVs): associated with cancer, with non-cancer diseases and putatively functionally neutral. Our approach is based on a previously introduces tool, StructMAn [1], which performs structural annotation of nsSNVs with respect to their position in the three-dimensional structures of the corresponding protein relative to

interaction interfaces with other proteins, nucleic acids, or ligand-binding sites.

We have analyzed the largest to date collection of nsSNVs associated with cancer, non-cancer diseases, as well as common and benign variants from COSMIC [2] (cancer-associated nsSNVs), ClinVar [3] (cancer-associated nsSNVs, nsSNVs associated with non-cancer diseases, benign nsSNVs), UniProt annotations [4] (cancer-associated nsSNVs, nsSNVs associated with non-cancer diseases), and ExAC [5] (common variants). We searched experimentally-resolved protein three-dimensional structures from the Protein Data Bank [6] for potential homology-modeling templates for proteins harboring corresponding mutations. We found such templates for all proteins harboring cancer-associated nsSNVs or nsSNVs associated with other diseases, and 51% and 66% of proteins carrying common and clinically benign variants, respectively.

Many mutations caused by nsSNVs can be found in protein-protein, protein-nucleic acids or protein-ligand complexes. We observe a significant enrichment of cancer-associated mutations in protein-protein interaction interfaces, which is probably due to the fact that these proteins more frequently act as protein interaction hubs. Whereas cancer-associated mutations are enriched in DNA-binding proteins, they are rarely located directly in DNA-interacting interfaces. In contrast, mutations associated with non-cancer diseases are enriched in DNA-interacting interfaces. A slight, but significant enrichment is observed for disease-associated nsSNVs in ligand-binding sites, which holds even if all drugs are removed from the consideration. In contrast, common and benign variants are depleted from all interaction interfaces and enriched on protein surface not involved in interactions.

Interestingly, nsSNVs are significantly enriched in protein core, where they might affect the overall protein stability. We do not observe this trend for cancer-associated nsSNVs.

In protein complexes, in which more than one subunit is affected by diseases-associated mutations, in isolated cases we show that mutations from different subunits tend to be located on interaction interfaces in close proximity to each other, but this does not appear to be a universal trend.

1. Gress, A., Ramensky, V. E., Buech, J., Keller, A. & Kalinina, O. V. (2016) StructMAn: annotation of single-nucleotide polymorphisms in the structural context. *Nucleic Acids Res* ,

44(W1): W463-W468.

2. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H. et al. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43:** D805–D811.

3. Landrum, M. J., Lee, J. N., Bensom, M., Brown, G., Chao, C., Chitipiralla, S. et al. (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44:** D862–D868.

4. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res* **43:** D204–D212.

5. Exome Aggregation Consortium. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285-291.

6. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. et al. (2000) The Protein Data Bank. *Nucleic Acids Res* **28:** 235–42