Functional annotation of genomic signatures using array CGH and NGS analysis of phenotypic variants with Mendelian inheritance (OTKA 103983).

In the moment of the submission the concept of the application was not usual: with the new generation methods, like the next generation full genome sequencing, exome sequencing, and array CGH we wanted to study those genomic diseases, in which not only the point mutations of the coding sequences, but the alterations of some non-coding genomic sequences are associated with not normal phenotypes, occasionally with accompanying dysmorphology, using analysis of pedigrees with Mendelian inheritance. Majority of these disases belnog to the Rare-disease group. Albeit our mind is still under the influence of the dogma which automatically associates the diseases with coding genes, however, thanks to the Human Genome Project it was already known that the coding sequence is only in the 1% range of the genome, and errors and variants of the so called genomic desert, containing functional elements in the intergenic sequences, can also associate with human pathology. Soon before the submission of the application the researchers clearly revealed the impact of the analysis of the human pedigrees with unknown genetic ethiology, not speaking about that some of these fields are not or hardly can be studied with approaches of in vitro-. cellular-, or animal-models. Thus, the research also focused on complex DNA sequence and copy number variation (CNV) analysis, and the research result can reveal a molecular genetic diagnosis for the patient, either. In the study we will use the existing biobank resources, that are parts of the national and an EU based international consortium (www.biobanks.hu, www.bbmri.eu).

The basic concept could be summarized as follows. In the history of the OTKA funded research projects such approach is likely very unique might be in the moment of the submission, however, nowadays the problems subjected in the current application can be regarded still as basic or translational research but not as routine, insurance funded examinations, and certain rare-diseases cannot be investigated without the patients, since they cannot be modeled (by cellular or more complex systems), because of their genetic background is still not known. The new generation sequencing techniques already belong to the routine diagnostic methods; it was not so in the moment of the submission of this grant. Today it is clear on the one side, that next generation sequencing, and the array CHG provide often useful and straightforwarding results, but it is also clear on the other side, that lots of data, sequences, signatures, copy number variants, still remains unsolved with their information content. Therefore, these clinical translation research areas still have a kind of pioneering role, including the IT interpretation problematics. The leading medical journals already had introduced the term of the "patient recontacting", even if the data were originally obtained in basic or applied research projects. The concept of the current application came, in part, from the observation that in patients having abnormal major or minor morphogenetic clinical variants, their diseases are associated in less extent with point mutations, but mainly with genomic rearrangements. Relative huge amount of such patients can be found in the VRONY registry; these patients are not really examined by molecular methods.

Until the end of the previous decade, the biomedicine science had already revealed the exceptional impact of the research of the formerly non-favored rare-disease tasks, as the usefulness of the advantages of the next generation sequencing and the spread of the array based methods became supportive in the research of orphan diseases. It is now unequivocally known, and clearly supported by the publications appeared in the highest ranked scientific journals, that besides the analysis of the pedigrees with unknown genetic origin, even the analysis of single entries can result in successful annotation of new genome functions, verification of new functional genomic elements, that could be identified exclusively by the use of the human model, by the analysis of the rare-disease affected individuals, and this entire process just entered into the hyped period. In other words, the phenotypes with unknown genetic origin are now in the high priority focus due to their scientific impact in the new gene or genomic element - new function research axis. The scientific value of the biobanks strongly increased, and the biobanks will maximizing their use comparing the previous decades, warranting future great progress due to the next generation technical achievements. Moreover, lots of the resulting findings will have to be regarded as immediate implementation of the results, as lots of the patients likely will get diagnosis in association with these research efforts.

In summary, the major goals were based on the following consideration. During the bio-medical research of the rare-diseases there had not been such direct relationship between the research and diagnostics in that extensive form like it can be observed in the time of the submission of the current grant. That time an exponential gain could be observed for certain novel DNA examining methods, due to the Moore's-law associated extreme fall of the consumable's costs; the research entered into a novel age, when the extended use of the novel techniques became a daily reality. Diseases belonging to the rare-disease definition presented a great clinical, research, and therapeutical challenge. The new generation techniques opened new perspectives for this disease-group, leading to a period when the research had a direct merit for the patients as well. The patients could get a molecular biology based diagnosis, and chance for preconception genetic support, while the science will discover new philology and biochemistry mechanisms, including those that could have not been discovered just with the already existing tools and methods. The procedure was advantageous also for the contemporary and future society, since appropriate handling of the rare-disease research is a pan-national necessity and political, economical goal as well.

The span was originally 4 years, after the third year we asked for extension because of serious health issues in my family, that was coincided with personal changes in the department. Therefore, the total period was 5 years. This extension did not influenced the major original issues and goals of the workplan. Besides these changes (for what we requested and got NKFIH approval), we also asked some reshuffles in the budget items. Perhaps the most important was the request for purchase of high speed sophisticated laptop and benchtop personal computer, which are ready to manage the courtesy programs from Harvard and Board Institute suitable for genomic analyses.

According to the annual reports, in the first year a reasonable progress has been achieved in the collection of Mendelian pedigrees with unknown genetic origin, which is the most important achievement in the future successful implementation of the project. The collection included mainly neuromuscular diseases, including dominant ataxias and spastic paraplegias, some are already included into next generation sequencing projects. One paper was accepted for publication, in which we studied a Roma pedigree with unusual muscular phenotype, and the whole exome sequencing revealed a new mutation of an already known gene, the MYH7, but the phenotype was absolutely new, which explains why this gene did not come into the focus during the clinical examination of the subjects.

In the second year we further collected the Mendelian pedigrees (ataxias, neuromscular diseases of unknown origin, Huntington-like disease, mtDNA disease). lots of the were subjected to next generation sequencing. The array (including CNV tests) were quite successful, we described new rearrangement associated phenotypes, like a case with deletion of 4q28.3-31.23. Still related to the array based gene annotation: we identified 108 schizophrenia associated loci as a member of a large consortium; due to the large author number the authors agreed not to show any monetary support in the article (Nature,2014 Jul 24;511(7510):421). Interethnic differences were revealed for some functional SNPs when we compared Roma biobank samples with average Hungarians. Using next generation sequencing of large chromosomal segments a new Neolith European lineage was discovered in a large US EU collaboration when about 8,000-year-old hunter-gatherers from Luxembourg were analyzed; our genomewide data were used as contemporary Hungarian controls (Nature2014 Sep 18;513(7518):409).

In the third period we made further steps in the annotation of variants of population genetic importance, the statistical calculations were still on the way (we faced computer problems; these programs required special hardware). The publication activity was related to the Roma biobank associated research, including research on lipid level modifier SNPs. In collaboration we identified a new mutation in a Roma family with Huntington-like disease (still under publication). A big consortium used our data for refining LOD score. We published 4 cases of chromosomal rearrangement, and papers on significance of drug metabolizing systems in Romani population samples.

According to the last annual report, the originally planned 10 pedigrees were already completed (we had over 25 NGS, and over 150 arrays completed; over 300 array were completed for SNP data). The findings were somewhat surprising and deviated from the international experience; the reason is not known.

Taken together, the findings could be divided into four major groups.

The monogenic disease (where a single gene is affected by some kind of mutations). As it was already mentioned, an MYH7 novel mutation with absolutely new phenotype was described by us. Further, we identified a Huntington-like phenotype, also, as very rare conditions we discovered new mutations and new phenotypic variations of Limb-girdle muscular dystrophy type 1E and Limb-girdle muscular dystrophy type 1F (all the 3 cases are under publications). In this groups we had extensive external collaborations.

The genomic disease category (the condition or disease is caused by alterations of the genome, rather that mutation of a coding gene) can be the next. In this category we had 4 publications with description of new phenotypes (4q28.3-31.23; long arm of chromosome 15; 4q21, and Kleefstra syndrome) and we have 2 further papers on the way.

In the population genomics and pharmacogenomics group we investigated susceptibility (IL23R, IL10R-1087, GJB2 W24X, MLXPL, GCKR, GALNT2, CILP2, ANGPTL3, TRIB1) or pharmacogenetic, pharmacogenomic variants (CYP26B, CYP26D, GLCCII, FCER2, 6q21, SLCO1B1, SLCO1B3, CYP4F/V433M/). Total of 13 papers have been published; in these papers we mainly used our biobanks.

In the multifactorial disease/condition groups we had mainly collaboration papers, with large consortia. The most successful was the schizophrenia working group, in which collaboration we achieved publications with really high rankings.

The data were also used for population genetic analyses as well; origin of human populations (Roma people) was investigated. Total of 2 papers are under publications.

Altogether, we had 29 research papers appeared, and still 8 are submitted or are under preparation. In almost all papers we indicated the OTKA support, with except a couple of consortial papers, where the authors agreed not show any such label because of the huge number of sources of monetary support.

A total of 5 PhD students (Balázs Duga, Renáta Szalai, Dalma Várszegi, András Szabó, and Ágnes Nagy) got their PhD degree using the support of the current grant, and 2 other candidates (Katalin Sümegi, and Petra Mátyás), completed their thesis.