ELSEVIER

From genetic discovery to future personalized health research

Aarno Palotie[1,2,3], Elisabeth Widén[2] and Samuli Ripatti[1,2]

[1] The Wellcome Trust Sanger Institute, Cambridge, UK
[2] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[3] The Broad Institute of MIT and Harvard, Cambridge, MA, United States

During the past ten years the field of human disease genetics has made major leaps, including the completion of the Human Genome Project, the HapMap Project, the development of the genome-wide association (GWA) studies to identify common disease-predisposing variants and the introduction of large-scale whole-genome and whole-exome sequencing studies. The introduction of new technologies has enabled researchers to utilize novel study designs to tackle previously unexplored research questions in human genomics. These new types of studies typically need large sample sizes to overcome the multiple testing challenges caused by the huge number of interrogated genetic variants. As a consequence, large consortia-studies are at present the default in disease genetics research. The systematic planning of the GWA-studies was a key element in the success of the approach. Similar planning and rigor in statistical inferences will probably be beneficial also to future sequencing studies. Already today, the next-generation exome sequencing has led to the identification of several genes underlying Mendelian diseases. In spite of the clear benefits, the method has proven to be more challenging than anticipated. In the case of complex diseases, next-generation sequencing aims to identify disease-associated low-frequency alleles. However, their robust detection will require very large study samples, even larger than in the case of the GWA-studies. This has stimulated study designs that capitalize on enriching sets of low-frequency alleles, for example, studies focusing on population isolates such as Finland or Iceland. One example is the collaborative SISu Project (Sequencing Initiative Suomi) that aims to provide near complete genome variation information from Finnish study samples and pave the way for large, nationwide genome health initiative studies.

During the ten years that the ESF Functional Genomics program has existed, our thinking and tools to improve the understanding of the genetic background of diseases has transformed substantially. In 2002, when the program started, the first draft of the human genome had just been launched. Trying to understand the genetic background of common diseases, our thinking at that time was heavily influenced by our previous success to identify gene mutations causing Mendelian (monogenic) diseases. The thought was that similar family-based positional cloning strategies used to pinpoint genes for Mendelian diseases, would also be powerful in case of complex traits. However, the underlying genetic architecture of complex diseases turned out to be fundamentally different from those in classical Mendelian diseases thereby needing different study designs and tools. The vision for complex disease genetics was highlighted already in 1996 when Risch and Merikangas [1] pointed out that association studies are far more powerful than linkage studies in detecting multiple variants with small effect sizes. However, at that time, our tools were not yet sufficiently developed to implement these visions in practice. During the past ten years, that is, during the lifetime of the ESF program, the tools

Corresponding author: Palotie, A. (aarno.palotie@helsinki.fi), (aarno@sanger.ac.uk)

1871-6784/$ - see front matter http://dx.doi.org/10.1016/j.nbt.2012.11.013

and opportunities have developed dramatically; the most important contributions being the drop in the cost of genotyping and sequencing.

As family-based linkage studies had modest success in identifying variants associated with the pathogenesis of common diseases, the contrary is true for the genome-wide association (GWA) studies. The work of the International HapMap consortium created a map of sequence variants common in the population (http://hapmap.ncbi.nlm.nih.gov/) that the industry used to develop standardized, cost-effective genotyping platforms. This enabled the genotyping of hundreds of thousands of variants in each individual study subject. Already the first GWA-studies clearly showed that the GWA-strategy is a successful method to detect associations between common polymorphisms and complex diseases and traits. As of July 2012, the GWA-study catalogue (http://www.genome.gov/gwastudies/) lists 1324 publications reporting robust associations between 6735 SNP markers and hundreds of diseases and traits [2].

By contrast, it also quickly became clear that most of the associated variants have very modest effect sizes (Odds ratios between 1.1 and 1.4). This inevitably means that GWA-studies need large sample sizes; samples from the low thousands up to hundreds of thousands are needed to pinpoint true associations among hundreds of thousands of markers with high statistical power. The need for large sample sizes beyond what individual study sites are able to collect quickly changed the mode of how genetic research is carried out. This new mode follows the global, collaborative pattern initially set by the Human Genome Project. No clinical center can work by them selves anymore. The groups had, and still have, to combine their resources to achieve meaningful statistical power. This was a major paradigm change in how academic research in this field is carried out. The necessity of large collaborative studies also had its impact on the funding structure. Large studies are expensive, beyond the reach of traditional academic funding modules, for example, the RO1 mechanism in US. Also in Europe, national and EU frameworks had a hard time (and still have) to respond to this quick change.

Lessons from GWA studies to guide us to design sequencing studies

One important lesson from the success of the GWA-era was that it is crucial to carefully and systematically build large collaborative networks. For example, the early collaboration between the academic HapMap scientists and the genotype reagent industry was a key-element that drove the development of standardized genotyping platforms that the research field needed. These commercial genotyping chips were subsequently used in laboratories around the world, which enabled straightforward combination of data for subsequent meta-analyses.

Another lesson was that a solid statistical framework for study design and data analysis should be developed early on. In the GWA-studies, this consisted of strategies to control for genotype measurement errors, standardized approaches to control for confounding due to differential allele frequencies across populations and the multiple testing burden. It also included strategies for replication and validation of associations. As a result, the GWA-studies have implicated an avalanche of novel candidate genes and regions for functional studies. The well-defined statistical framework saved the field from a lot of unnecessary confusion. Now, when we have moved towards sequence-based low-frequency variant association studies, the benefits for setting rigorous statistical framework early on should be kept in mind.

From the GWA locus to function

GWA-studies have been criticized for that they only provide information on genetic loci but no functional insight. Although we think that the critique in part is unjustified, it nonetheless highlights an important next step of the research. However, at present the field of functional biology is lacking efficient tools to shed more light on the function of the associated variants. One challenge is that the effect size of each individual variant is small and thus it is not trivial to develop readouts for cell- or model organism-based assays that reliably monitor small changes, e.g. in gene expression or phenotypes. Knock out assays may be a practical approach to elucidate gene function. Yet it is not an optimal model to decipher the in vivo consequence of genome variation with modest functional consequences. Secondly, the associated variants do not work in isolation. In some cases there is evidence that some cumulative effects of variants are additive [3], but there is also evidence that some variants can compensate for the effect of others [4]. Thirdly, most of the associated loci are located outside the coding regions of genes. Our understanding of the function and relevance of these intergenic, probably regulatory, regions is limited. This limited knowledge certainly limits the design of follow-up experiments. Large collaborative initiatives such as the ENCODE (Encyclopedia of DNA Elements) project aim to improve our understanding of the function of the genome. It provides new genome-wide tools to link disease-associated gene variants to gene function. One example is the open chromatin assays, including mapping of DNAase I hypersensitive sites in a large panel of cell lines (http://www.genome.gov/10005107).

Towards a more complete understanding of the human genome variation

The success of the GWA-studies stimulated the field to proceed to provide an even more comprehensive understanding of the association of disease risk with genome variations. The introduction of new sequencing technologies and the dramatic drop of sequencing costs made it possible to explore the human variation in more detail. The first international effort was the 1000 genome project that aims to establish an extensive catalogue of human variations from 25 populations (www.1000genomes.org). The 1000 genomes project consists of anonymous participants without any phenotype information. Thus its main aim is to provide an international open access resource that serves as a basis for subsequent phenotype related studies. The 1000 genome project has been followed by large disease-specific sequencing studies, for example, the UK10K (www.UK10K.org) and GoT2D (Genetics of Type 2 Diabetes).

Medical genome- and exome sequencing

The excitement of low cost sequencing has stimulated a boom of whole-genome and whole-exome sequencing studies. There are several successful examples of the identification of Mendelian gene mutations using new sequencing techniques (for references see [5]). At the same time, the challenges of this approach have also