

Storing and using biobanks for research

Lena Jonsson

Amersham Pharmacia Biotech, Uppsala

&

Ulf Landegren

Rudbeck Laboratory, Department of Genetics and Pathology, Uppsala University

Introduction

The successful application of large-scale biological analyses to determine total genome sequences inspires similar grand attempts to understand the function of the encoded molecules in health and disease. Collections of patient samples – biobanks – will play a central role in this work by revealing genetic, other molecular and environmental factors of importance for the development of disease, as well as for properties like personality traits or longevity. It is therefore appropriate to consider some of the technological prerequisites for the use of biobanks. We will briefly discuss the accumulating knowledge base arising from genomic research, the identification of medically relevant molecular indicators, and the development of technologies and procedures to investigate these factors using banks of human tissue samples.

Genomic Progress

The Human Genome Project

The fundamental strategy in genomic research is to establish and utilise effective methods to accumulate extensive, ultimately total information about aspects of genomes, without focussing on specific biological or medical problems¹. After a recent acceleration of data gathering, information on the nucleotide sequence of the human genome is now all but complete. An understanding of the complete complement of human genes is not yet at hand, but should become available over the next few years, probably listing considerably less than 50,000 genes.

Information is also accumulating about the normal variation among human genomes. Most commonly, sequence differences between individuals involve differences in single nucleotide positions. DNA sequence variants with a reasonably high population frequency are observed approximately every 1,000 nucleotides across the genome, and are referred to as single nucleotide polymorphisms (SNPs). Information about more than one million variable sequences is already publicly available via the Internet², and more such markers are available from commercial providers of genetic information. There is a strong and widespread expectation that information about genetic variation, along with improved methods to study this in suitable biobanks, will provide invaluable insights into the basis of all major diseases.

It is already becoming apparent that this will prove a challenging proposition, however, and that further improvements in strategy and technology may be necessary to realise this potential. In parallel with sequencing of the human genome, total genome sequences of important model organisms like yeast, *C. elegans*, the fruit fly, the mouse and the rat, as well as of most important infectious agents, have become or are rapidly becoming available. This provides important opportunities for illuminating comparisons between different organisms, and for identification of the sources of infections in patient samples.

Functional Genomics

The focus of research interest is now shifting to the accumulation of global information about how genomes encode the properties of proteins, cells, and organism, through functional genomic research³. Over the next ten-year period these efforts should provide invaluable information about the structure of most or all human proteins, their interaction partners, and the connectivity of all metabolic and signal transmission chains, thereby defining the modules that perform the various cellular functions⁴. Analytical tools will be developed for efficient studies of all these properties; tools that also become available for analyses of biobanked material.

The search for molecular lesions

Current medical research yields an abundance of molecular factors of value as targets for diagnostic investigations. The various types of analyses commonly used to investigate such factors in biobanked samples may be grouped as follows:

Resequencing

Single or small sets of genes are being identified that are typically found to be altered by mutations in particular diseases, and that may therefore need to be scrutinised in individual patients for the cellular equivalents of software bugs. This requires resequencing the relevant gene in individual patients to reveal any deviations from one or a few known normal variants of the sequence. For many genes, information about hundreds or more pathologic alterations that have been observed are already archived in mutation databases available via the Internet.

Genotyping

Many known human DNA sequence variants are known to be associated with particular diseases⁵ or they may influence a patient's response to a particular drug⁶. Genotyping of such markers may therefore be of value to characterise patient populations. DNA sequence variants with no known functional consequences can also be useful in association and linkage analyses.

In association studies, the frequency of variants of individual genetic markers are compared between healthy persons and patient populations, with the hope that an observed difference in frequency may be the consequence of a direct effect by the sequence difference, or due to co-inheritance with nearby, unknown genetic variants having such an effect. Associated markers with no direct effect on the disease are said to be in linkage disequilibrium with the disease-related changes, and they may therefore guide the investigator to the gene, which is directly involved in the disease. If the DNA samples are derived from individuals in families where particular diseases are known to segregate, then the location of

disease-associated genetic changes among the chromosomes may be pinpointed by genetic linkage analysis using the same types of genetic markers. This approach has proven extremely valuable for defining the nature of conditions primarily influenced by single or a limited number of genes⁷.

However, when many genes are involved in causing disease, along with prominent influence by environmental factors, then both association studies and linkage analyses have met with far greater difficulty. This type of more complex disease causation is probably characteristic of most clinically important human diseases.

Transcript Profiling

By measuring how individual or large sets of genes are expressed in the cells of a sampled tissue, a valuable view is obtained of the state of differentiation and activity of the sampled cells. Suitable comparison between healthy and affected tissues can reveal genes whose expression is affected by disease. The analysis provides an important basis for diagnosis in e.g. malignancy⁸, and allows monitoring of the progress of disease or response to therapy. After a phase of analyses of transcription of thousands or tens of thousands of genes, it will probably be feasible to define limited sets of genes likely to be informative for disease diagnosis and follow-up.

Gene quantification

Copy number changes of specific gene regions in malignancy can reflect loss of genetic factors that protect against excessive proliferation, or conversely, gain of copies of genes that can support such proliferation, evident as altered copy numbers of certain chromosomal regions. Regions of the genome present in increased or decreased copy numbers can be demonstrated by comparing genomic DNA samples from normal and malignant tissues through a process called comparative genome hybridisation⁹, allowing genes critically important for the malignant phenotype to be demarcated.

Proteomics

Following in the footsteps of the characterisation of the human genome comes the desire to associate the gene products to functions and roles in biological networks. Analyses of the expression of large sets of proteins, or of rearrangements or secondary modifications of proteins are coming more and more to be seen as promising approaches for monitoring biological processes in patient samples. Such proteomic analyses reflect the biological state of the body fluid or tissue at the time the sample was collected, sensitively influenced by both genetic and environmental factors¹⁰. Samples of any types of tissues or body fluids can be utilised for the analyses, and the generated information adds valuable pieces to the biological puzzle. Identifying protein patterns that are specific for a certain disease can enable the identification of disease mechanisms and the development of new diagnostic and therapeutic tools, and perhaps ways of preventing the development of disease.

And much more

Other classes of molecular analyses are known to be informative about predisposition for, or about the state and progress of disease. For example, gross structural changes of chromosomes can be of diagnostic value in congenital disease and in malignancy.

Cytogenetic analyses of this kind can be combined with in situ hybridisation using specific DNA probes. In histopathology too, analysis of the microscopic location of specific proteins and RNA and DNA sequences provides more specific information. The ability to expand DNA sequence elements at the ends of chromosomes, telomers, is characteristic of a limited number of stem cells but also of many malignant cells, and this activity can therefore betray the presence of malignant cells in tissue samples¹¹. Comprehensive measurements of metabolites, sometimes referred to as metabolomics, can be highly informative in analyses of patient samples. Currently, a rapidly expanding number of analytes are being identified that can be of interest to monitor in biobanked material.

Methods for molecular analyses

One of the factors sparking increased interest in collections of patient samples is the greatly improved techniques that are becoming available for such analyses, enabling far greater numbers of increasingly precise analyses of progressively smaller tissue samples, as well as entirely new forms of examinations.

Analytical technologies

Routine analyses of human genes became feasible in the mid-1970s with the advent of the Southern blot, and the Sanger method to deduce nucleotide sequences followed a few years later¹². The preconditions further improved with the development of the polymerase chain reaction in the mid-1980s, by providing the specificity and sensitivity required to conveniently detect individual DNA sequences among the thirteen billion nucleotides constituting the genome in each human cell nucleus¹³.

Since the mid-1990s an increasing number of genetic analyses have been performed using microprocessor-like chips where large numbers of different probes have been deposited¹⁴. These DNA microarrays permit simultaneous analysis of large numbers of gene sequences in analyses such as resequencing, genotyping, gene expression measurement and so on.

The toolkit for proteomic analysis has also expanded considerably in recent years. Techniques like electrophoresis, chromatography, and mass spectrometry have been scaled to address the need for comprehensive analyses of protein complements. In another important line of development in proteomics, binding by antibodies and other affinity reagents can reveal the presence and concentrations of large sets of specific proteins.

Technological developments at home and abroad

Sweden has a strong tradition of developing methods for analysis of biomolecules. The Svedberg and Arne Tiselius both received the Nobel prize for their work on ultracentrifugation and on electrophoresis, respectively. These methods of protein characterisation were later followed by others like chromatography, isoelectric focussing, and antibody-assisted immune precipitation and ELISA techniques. More recently, a number of techniques for molecular analyses have been established by scientists in Sweden¹⁵. Clearly, the availability of unique detection technologies has played an important role in Swedish biomedical research at universities and in the pharmaceutical industry, and it has provided a basis for a viable biotech industry. Nonetheless, detection techniques continue to be an important limiting factor in the use of biobanks and further dramatic improvements are both required and expected.

Future technologies that may significantly influence the usefulness of biobanks include methods to prepare and handle large numbers of samples in parallel. One existing example of such technologies – tissue arrays – allows multiple representations of a thousand or more tissues to be prepared and accessed for parallel microscopic analysis. It is also foreseeable that far larger numbers of target molecules can be studied at one time. This is of importance both from the point of view of throughput and because it promises to limit consumption of valuable collected samples if one test can answer a thousand questions. Investigations that can report the relative location or fine structure of proteins or precise quantity of sets of proteins are likely to provide important insights into the state of tissue samples. Finally, an increasing range of tests that directly analyse the function of important cellular components will provide valuable information about biobanked samples.

Biobanks

Excellent opportunities thus exist to combine the rapidly growing knowledge about the constituents of our cells and the improved means of analysing them in the course of studies of central biological questions. Such investigations are frequently performed in model organisms that present opportunities for genetic crosses, gene modifications, or other interventions. However, many problems peculiar to human health or specific human properties will require studies of collections of tissue samples from human donors. It is therefore important to make plans for effective biobanks where samples of body fluids and tissues can be received, stored, and made available for research.

By collecting and storing well-characterised sets of patient samples, the same samples can be used in many different investigations, reducing inconvenience to the donors and allowing more effective use of research funds. Studies of banked samples can be continually extended as technology improves or more information becomes available about factors of particular interest in the investigation of specific diseases. In this way comprehensive information will accumulate about collected samples, increasing the value of any subsequent investigation.

Deposits and withdrawals in biobanks

A fundamental problem with biobanks concerns the need to collect samples today, or better still yesterday, for research needs that may arise tomorrow. Accordingly, there is a significant problem in determining which samples should be collected. The use of biobanks presents a related problem, made in that it has to be decided whether the application of a limited sample resource is justified for the study in question. Clearly, more focussed studies could be performed at some later time when more is known about the problem in question. Moreover, at a later time more analyses will be possible with the same amount of tissue, or entirely novel forms of analysis may become possible.

For molecular genetic analyses, the collected samples must include nucleated cells in order to provide a store of genomic DNA. This is commonly obtained by sampling blood or perhaps cells obtained from the lining of the mouth cavity, but a wide variety of tissues can be used. Studies of somatic genetic changes that may have arisen during life, for example in tumour tissues, naturally require nucleated cells from the relevant tissues to be sampled.

What genes are expressed in sampled tissues, and at what levels, can be investigated by mRNA analyses, again necessitating that nucleated cells to be present in the sampled tissues. Due to the proneness of mRNA to degradation, special care must be taken to preserve

integrity, e.g. by storage at very low temperatures or in the presence of denaturants that prevent enzymatic degradation.

Proteins can be analysed in a wide variety of samples of tissues and body fluids and do not require that the samples containing nucleated cells. Like samples collected for mRNA analyses, they also require careful methods for treatment and storage of the samples.

Analyses of low molecular weight substances such as hormones, nerve transmitters and metabolites also provide valuable views on the state of sampled tissues, and on factors directly or indirectly involved in ongoing disease processes. Such analyses are possible using a wide range of tissues.

Some donated samples that include live cells may be preserved by adapting the cells for tissue culture, providing a permanent, renewable record of the genotype and permitting functional assays at the level of the cell. As an alternative, tissues can be stored at low temperature and under conditions where viability is preserved for possible later adaptation for tissue culture. Further progress in the understanding of cellular differentiation and the culture of stem cells may allow widely different tissue types to be derived from stored samples of patient tissue.

The large numbers of serum samples that exist at many clinical departments can be used for proteomic studies and may permit detection of infectious agents or immune responses to such agents, but genetic studies are difficult because of the minimal and variable presence of nucleated cells. Formalin-fixed and paraffin-embedded samples are routinely preserved after pathological analyses and very large numbers of such samples are stored at departments of pathology around the country. Future analytic technologies may allow an increasing number of biomolecules to be evaluated in such samples.

The Swedish situation

Sweden has excellent, in some respects unique, advantages for the collection and application of biobanks in large epidemiological studies, as is already reflected in the scientific literature. Several aspects have been of key importance for this work:

- For genetic studies it can be of advantage that parts of Sweden were colonised by a limited number of individuals, since this increases the probability of two individuals who both have a given disease also sharing genetic risk factors.
- An advanced health care system is of importance in ensuring that sampled individuals have been correctly and promptly diagnosed and characterised with respect to relevant clinical parameters.
- Church records and other registers can provide information on the genealogy of sampled individuals, establishing relatedness as required in genetic linkage studies.
- The availability of health care records detailing diagnoses at discharge from hospitals or at time of death or listing twins, provides a potential for rapidly identifying suitable populations to sample.
- The use of personal identifiers (unique national registration numbers) can permit cross-referencing between the large number of registers, subject to permission being granted for such studies.
- Large banks of collected samples are already at hand, for instance serum samples, pathological specimens, cervical smears collected in check-ups, or the PKU-archive containing blood from new-born children, and banks of collected blood samples such as that maintained by the Medical Biobank in Umeå.
- Advanced epidemiological expertise is available, but may nonetheless remain one of several limiting factors for the exploitation of biobanks.

- It is of some importance that the Swedish population is larger than, for example, that of Iceland, where rare diseases will have a more limited representation.
- Finally, the population appears to have a generally positive attitude towards donating samples and providing background information in the interest of furthering scientific understanding of disease. The relationship between patients and scientists is based on a social contract, necessitating that the highest standards are maintained in protecting the integrity of the donors, and that collected resources are optimally stored and applied.

Conclusions

Excellent opportunities thus exist for combining the rapidly growing knowledge about the constituents of our cells and the improved means of analysis of them in the course of studies of central biological questions. Investigations of this kind are frequently performed in model organisms which present opportunities for genetic crosses, gene modifications, or other interventions. However, many problems peculiar to human health or specific human properties will require studies of collections of tissue samples from human donors. It is therefore important to make plans for effective biobanks where samples of body fluids and tissues can be received, stored, and made available for research.

References:

-
- ¹ *Nature*. 2001, vol. 409 (6822), pp. 745 ff; *Science*. 2001, vol. 291 (5507), pp 1153 ff.
- ² Kwok, P. Y., and Gu, Z. Single Nucleotide polymorphism libraries: why and how are we building them? *Mol Med Today*. 1999. vol. 5, pp. 538-453.
- ³ Vukmirovic, O. G., and Tilghman, S. M. Exploring genome space. *Nature*. 2000. vol. 405, pp.820-822.
- ⁴ Hartwell, L. H. et al. From molecular to modular cell biology. *Nature*. 1999. vol. 402 (suppl), pp. C47-C52.
- ⁵ Clain, J. et al. Two mild CF-associated mutations result in severe cystic fibrosis when combined in CIS and reveal a residue important for CFTR processing and function. *J Biol Chem*. 2000. Dec 15; Pauling, L. et al. Sickle cell anemia: A molecular disease. *Science*. 1949. vol. 110, p. 543.
- ⁶ Drysdale, C. M. et al. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci*. 2000. sep 12, vol. 97 (19), pp. 10483-10484.
- ⁷ Chapman, N. H., and Thompson, E. A. Linkage disequilibrium mapping: the role of population history, size and structure. *Adv Genet*. 2001. vol. 42, pp. 413-437.
- ⁸ Alizadeh, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000. vol. 403, pp. 503-511.
- ⁹ Forozan, F. et al. Genome screening by comparative genomic hybridisation. *Trends Genet*. 1997. vol. 13, pp. 405-409.
- ¹⁰ Galvani, M. et al. Two-dimensional gel electrophoresis/matrix-assisted laser desorption/ionisation mass spectrometry of commercial bovine milk. *Rapid Commun Mass Spectrom*. 2001. Feb 28. vol.15 (4), pp. 258-264.
- ¹¹ Buys, C. H. Telomeres, telomerase, and cancer. *New Engl J Med*. 2000. vol. 342, pp. 1282-1283.
- ¹² Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*. 1975. vol. 98, pp. 503-517.
- ¹³ Saiki, R. K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*. 1985. vol. 230, pp. 1350-1354.
- ¹⁴ Lander, E. S. Array of hope. *Nat Genet*. 1999. vol. 21 (Suppl), pp. 3-4.
- ¹⁵ Ronaghi, M. et al. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*. 1996. Nov 1, vol. 242(1), pp.84-89; Nilsson, M. et al. Padlock probes: Circularizing oligonucleotides for localized DNA detection. *Science*. 1994. vol. 265, pp. 2085-2088; Syvänen, A. C. et al. A primer-guided nucleotide incorporation assay in the genotyping of Apolipoprotein E. 1990. *Genomics*. vol.8, pp. 684-692; Howell, W. M. et al. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms *Nat Biotechnol* 1999, vol. 17, pp. 87-88. Svanvik N. et al. Detection of PCR products in real time using light-up probes. *Anal Biochem* 2000 Dec 1; vol 287(1), pp. 179-82.