Harvesting the Human Genome: the Israeli Perspective

Edna Ben-Asher PhD[1,2], Vered Chalifa-Caspi PhD[2,3], Shirley Horn-Saban PhD[2,3], Nili Avidan PhD[1,2] Zviya Olender PhD[1,2], Avital Adato PhD[1,2], Gustavo Glusman[2], Marilyn Safran[2,3], Menachem Rubinstein PhD[1,3] and Doron Lancet PhD[1,2]

[1]Department of Molecular Genetics, [2]Crown Genome Center and [3]Department of Biological Services, Weizmann Institute of Science, Rehovot, Israel

Key words: DNA sequencing, DNA arrays, bioinformatics, computational genomics, disease genes, mutation detection

IMAJ 2000;2:657–664

For Editorial see page 690

The genome revolution

Since its inception in 1990, the two principal goals of the Human Genome Project have been the mapping and sequencing of the entire gamut of three billion DNA bases that code for human beings. In the USA, the project is led by two government agencies: the Department of Energy and the National Institutes of Health. Other major participants in the HGP are England, Germany and Japan. As we write this review, the project is near completion: mapping has been fully attained, and the establishment of a draft DNA sequence of the human genome has been formally announced [Figure 1]. A key player in the change of schedule from the projected completion date of the year 2005 has been the company Celera Genomics, which performed a parallel genome-wide sequencing effort. It utilized a revolutionary approach of whole-genome shotgun sequencing. In the public domain, more than 20% of the human sequence is in finished form (i.e., entire genomic clones sequenced top to bottom with >99.9% accuracy), and over 70% is available in draft form (i.e., the sequence of each clone is available in segments, and with lower accuracy) [See www.ncbi.nlm.nih.gov/genome/seq]. Sequencing of chromosomes 21 [1] and 22 [2] has been completed.

A scientist with access to the cumulative assembled results of both the government project and the corporate data (currently available only as institutional paid subscriptions) can basically view the entire human genome, where >99% of the bases are represented somewhere in the sequencing results. However, the draft nature of the sequences requires considerable dexterity and knowledge in bioinformatics, crucial for harvesting the data. The world is now entering a new phase of this project – the annotation and interpretation of the results, the functional analysis of the genome sequences, and the study of genetic variation among human individuals [3].

HGP = Human Genome Project

The Israeli angle

The need for an Israeli involvement in the HGP was initially debated. Can a small country like Israel compete with the big powers? Luckily, there was an emerging recognition that by the time the project is done, Israel would be facing a problem of survival in a world whose rules had dramatically changed. It became apparent that by having even a minor involvement and capacity building throughout the project, the State of Israel will be significantly better positioned for the "post-genome era," when it arrives. Along these lines, the Israel Academy of Sciences, and later the Ministry of Science, Culture and Sports, have been active in providing support for the necessary infrastructure. The Weizmann Institute Genome Center, to be described in some detail

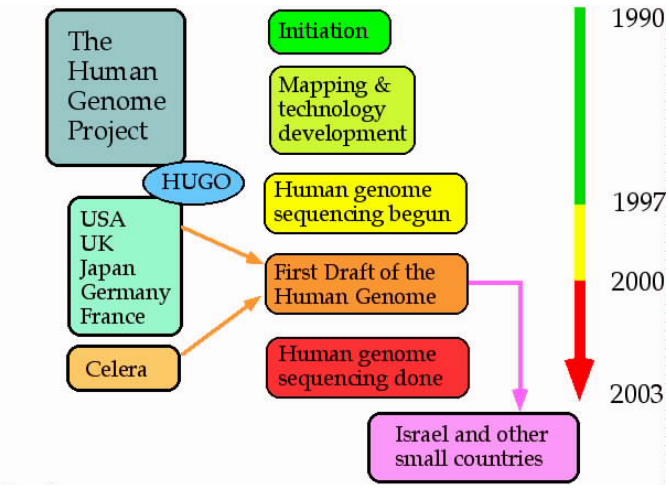


Figure 1. Steps involved in the progression of the worldwide genome project. The genome project was initiated by the Human Genome Organization and by a worldwide consortium in 1990. The first target, whole genome mapping, was accomplished by 1997. Thereafter, genome sequencing took over and is now very near to completion. Israel joined the project in 1993.

below, is part of this effort, in conjunction with a center at the Hebrew University of Jerusalem, headed by Prof. Bat-Sheva Kerem. Other national activities have been established at Tel Aviv University (Prof. Batsheva Bonne-Tamir), and several other institutes of higher education have joined in establishing genome and bioinformatics activities on their campuses. In parallel, important progress has been attained in the Israeli biomedicine and biotechnology scenes, and this trend of coping with the genome revolution in the health and corporate spheres will likely intensify considerably in the coming decade. Hopefully, a fruitful cooperation will be established among the different entities involved in the realms of the new medicine and pharmacology that stem from the genome revolution.

In broad outline, a well-established genome-related effort typically consists of three major activities: Genomic DNA sequencing, Bioinformatics, and DNA arrays. Here we describe such activities as typified by the ongoing research and development at the Weizmann Institute of Science.

Genomic DNA sequencing

This technology lies at the heart of all current efforts in the world Genome Project. The core instrumentation is the automated fluorescence-based electrophoresis DNA sequencer. In its different variations, this has been the only means for generating megabase sequences of entire human chromosomes, as well as those of many model organisms and plants. Following the first-generation instruments, namely the PE-Applied Biosystems models 373 and 377, our center has now installed the new generation automated 96 capillary sequencer, i.e., the Perkin Elmer-Applied Biosystems model 3700. This new instrument typically expands the sequencing capacity about tenfold and has been single-handedly responsible for the huge acceleration in the worldwide sequencing speed over the last 2 years.

The technology of large-scale genomic sequencing has to be complemented with a series of additional genome methodologies, which are defined under the umbrella of "shotgun sequencing." This includes DNA shearing by specific sonication methods, sub-cloning into a bacterial vector, the preparation of hundreds of DNA samples from the sub-clones and, finally, sequencing of these clones. Many of these steps are accomplished with the aid of special robotics, also present at the WIS Genome Center. Shotgun DNA sequencing introduces computational challenges – the computer-based assembly of the sequenced fragments. For this, computer programs such as PHRED/PHRAP [4], the Staden package [5] and Sequencher [6] are used, which take advantage of overlapping reads. Using these programs enables eventual reconstruction of the genomic clone full sequence.

As part of our projects at the WIS Genome Center, we perform large-scale DNA sequencing of different types of genomic clones, such as cosmids, PACs and BACs (50–150 kb long). In collaboration with other scientists in Israel and abroad, we analyzed genomic regions carrying genes of interest, thus carving our own niche in the world of genome sequencing.

As an example of the advantages of large-scale sequencing, we have sequenced 150 kb of DNA on chromosome 21 in a region that includes the very large gene for acute myeloid leukemia 1 (*AML1* or *RUNX1*), a project led by Prof. Yoram Groner of the WIS. The complete genomic structure of AML1 was thus determined and a detailed analysis of the whole gene region performed [submitted for publication]. This has contributed to the understanding of chromosome translocations associated with acute myeloid leukemia. It also earned Israel a unique position as member of the international consortium that sequenced chromosome 21.

Another large sequencing project was that of the entire 450 kb cluster of olfactory receptor genes on human chromosome 17p13.3 [7]. Analysis of this region led to identification of new OR genes and pseudogenes, and enabled deciphering of the genomic structure of OR genes, the identification of their regulatory sequences and the study of their evolution.

Gene identification following linkage analysis

The genome's first draft is most advantageous to researchers in Israel. This is because of the unique genetic consistency of the Israeli population, Jewish as well as non-Jewish. This population includes various defined ethnic groups and genetic isolates that are highly informative for genetic studies. Linkage analysis of a particular disease among members of affected families, originating from a single ethnic group, usually enables defining a rather narrow genomic interval for the disease-causing mutation. Once a particular monogenic disease has been mapped to a chromosomal region, the study is ripe for an attempt to identify the specific gene responsible for that disease. Our center has developed an expertise in this important aspect of "genome harvesting." We provide know-how to Israeli geneticists and physicians by carefully browsing and interpreting the accumulating data produced at the various genome centers. The initial steps in this transit from linkage mapping to gene identification involve database searches in order to identify the relevant clones that have been partially or fully sequenced. Computer programs are then used to bring together all the sequenced segments from different databases, including those at the National Center for Biotechnology Information and the individual genome centers worldwide. Then the sequences are analyzed in detail, and the candidate genes encoded within them are identified. Insight related to presumed function is employed

PE = Perkin Elmer
WIS = Weizmann Institute of Science

AML1 = acute myeloid leukemia 1
OR = olfactory receptor

to examine the feasibility of different candidate genes causing the disease. Annotation databases such as GeneCards [8] and Unified Database as well as annotation programs such as Rummage [9], Genotator [10] and GESTALT [11] are employed.

The next step is to determine the genomic structure (exon and intron boundaries) of these genes. This is done both by computer algorithms and by a comparison of the genomic sequences to expressed sequence tags and cDNAs using the appropriate databases and tools; e.g., NCBI human ESTs, TIGR (the Institute for Genomic Research) tentative human consensus sequences – (http://www.ncbi.nlm.nih.gov/BLAST/thcblast.html) and Compugen-LabOnWeb (http://www.labonweb.com/). Once exons are identified for the candidate gene, a re-sequencing operation is initiated. Genomic DNA samples taken from affected and control individuals are subjected to DNA sequencing. Programs such as the Staden Package are used to detect sequence variations between those affected and the controls. This is often a laborious process, since bona fide mutations must be distinguished from common polymorphisms and from occasional sequencing errors. A combination of sequence analysis dexterity and careful genetic analysis is thus employed to eliminate ambiguities. This arduous process then culminates in the identification of the mutations responsible for the disease. Identification of exclusive mutations in affected individuals is the final proof that the candidate gene is indeed involved in generation of the disease. It is important that affected individuals bearing different haplotypes be analyzed so that the identification is corroborated through the discovery of at least two different mutations.

Several specific projects aimed at identifying genes involved in monogenic diseases are currently underway at our center [Figure 2].

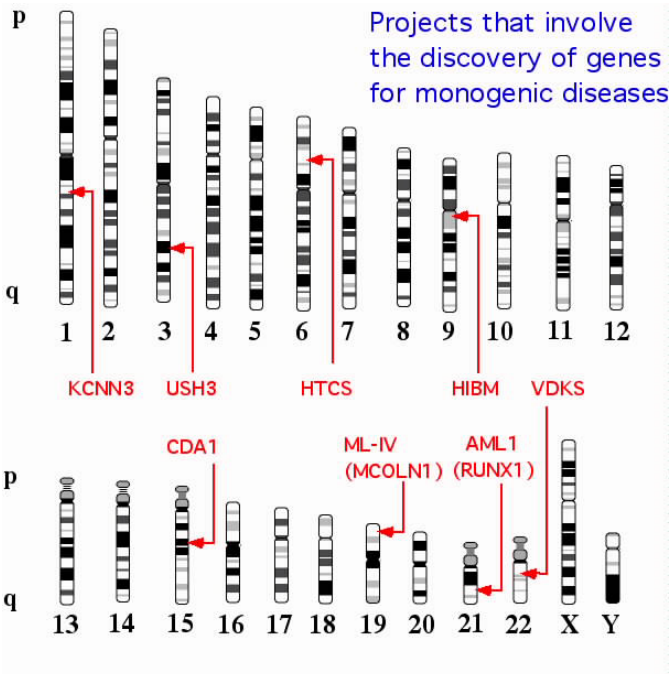


Figure 2. Projects that involve the discovery of genes for monogenic diseases carried out at the Weizmann Institute Genome Center. The chromosomal location of the genes is indicated. KCNN3 = potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3, This project is headed by Prof. Ruth Navon from Tel Aviv University. USH3 = Usher syndrome type III, in collaboration with Prof. Batsheva Bonne-Tamir, Tel Aviv University. HTCS = hypotrichosis simplex, HIBM = hereditary inclusion body myopathy, headed by Prof. Stella Mitrani-Rosenbaum, Hebrew University–Hadassah Medical School, Jerusalem. CDA1 = congenital dyserythropoietic anemia type 1, headed by Prof. Hanna Tamary, Schneider Children's Hospital, Petah Tiqva. ML-IV = mucolipidosis type IV, in collaboration with Prof. Gideon Bach, Hebrew University–Hadassah Medical School, Jerusalem. AML1 = acute myeloid leukemia type 1, headed by Prof. Yoram Groner, Weizmann Institute of Science. VDKS = Van-Der Knaap syndrome, headed by Prof. Elon Pras, Sheba Medical Center, Tel-Hashomer. These projects are carried out on a collaborative basis between the above mentioned researchers and the WIS Genome Center.

Polymorphisms and multigenic diseases

One of the most important future implications of the post-genome era is related to common polymorphisms. It is believed by many that practically all the common diseases, including elevated blood pressure, heart failure, asthma, osteoporosis, Alzheimer's disease, diabetes, mental diseases such as schizophrenia, and many more, are multigenic, i.e., caused by contributions of multiple genes. Often, they may be manifested in an individual possessing a specific combination of single nucleotide polymorphisms. SNPs are rather common in human populations, occurring on average once every 1,000 bases. Thus, the human genome may harbor up to three million SNPs defining a binary "string," which is three million bits long, where the value 1 represents one possible allele (e.g., A) and the value 0 the second possible allele, (e.g., C) [Figure 3]. Most SNPs have no functional consequences, while others appear to be important, especially those that occur inside exons (coding SNPs), at exon-intron boundaries and within promoter regions. A worldwide race is now underway to attain access to large numbers of cSNPs and to correlate them with diseases. Linkage analyses in families, or association studies in affected individuals versus controls, may be utilized. Again, a specific technological infrastructure is required (e.g., SNP chips or mass spectrometry-based SNP detectors). This should be combined with a broad theoretical knowledge of population genetics and evolutionary history. Our center is getting ready to provide for such needs, as are other genome centers in Israel.

NCBI = National Center for Biotechnology Information
EST = expressed sequence tags
SNPs = single nucleotide polymorphisms

cSNPs = coding SNPs

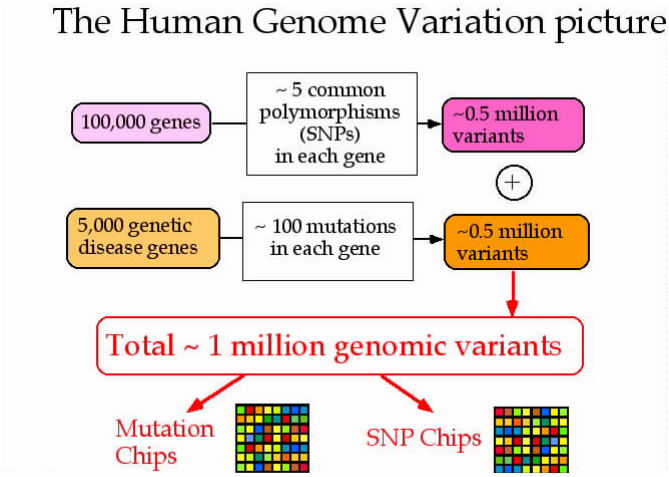


Figure 3. A schematic representation of the vast genomic variation within the human population. As can be seen in the scheme there are 0.5 million estimated cSNPs and 0.5 million estimated disease-causing mutations. This variation, which is essential for drug design, will be analyzed on the whole genome of various populations, mainly by the use of DNA chips technology.

Computational genomics

The accumulation of human genome sequences and annotation information is accompanied by the development of numerous Web-accessible databases, as well as tools for data search and analysis. These encompass the various aspects of human genome research such as genomic and mRNA sequence information, protein sequence, structure and function, biological networks, variation studies and comparative genomics. Our established computational genomics capacity helps researchers find information and Web-available tools relevant to their field of research. In addition, it helps develop new resources that enable retrieving and integrating human genome information from existing databases and presents them in a summarized comprehensive way.

Israel faces a major challenge in making use of its excellent capability in the fields of mathematics and computing on one hand, and of life sciences and biomedicine on the other. A unique opportunity exists in joining the two fields together under the umbrella of computational genomics and bioinformatics. Centralized, nationally endowed know-how and computer hardware and software are essential for achieving this goal. It is imperative that three parallel capacities be developed:

- Implementing computational genomics tools for mastering in detail all the sources of genomic information in the public domain Web resources and in purchasable corporate resources.
- Developing in-house algorithms, software packages and databases that will allow the Israeli community to communicate freely and at the same level with colleagues abroad, and gain the best access to computational genomics and bioinformatics resources.
- Implementing technologies such as DNA arrays and SNP analyses, that also require specific talents in aspects of computational biology, e.g., cluster analysis or population genetics.

Locally developed genome databases and analysis tools

As examples for the second point, our center has developed the genomic tools as described below:

• GeneCards, a database of human genes, their products and their involvement in diseases [9] (http://bioinfo.weizmann.a-c.il/cards). It is based on information retrieved from public, Web-accessible databases such as SWISS-PROT, OMIM, UniGene, GDB and others. By integrating and organizing the retrieved data in gene-related "cards," GeneCards provides a quick-glance starting point for each gene; it also enables "lateral" searches of the human genome for genes involved in specific physiological and pathological conditions. GeneCards currently includes all human genes that have a HUGO approved symbol, as well as selected others. Each card presents information about the gene symbol, name and synonyms, chromosomal location, encoded protein(s), nucleic acid sequence(s), homologous mouse gene, related disorders, research articles from PubMed, as well as links to additional sources of information about mutations, clinical information, linkage data and clone collections. A powerful free-text search engine enables retrieval of all gene cards whose text contains the entered keyword (e.g., the name of a gene, enzyme, hormone, disease condition, tissue, chromosomal location, sequence, etc.). A query-reformulation support mechanism reacts to unsuccessful user queries by suggesting further search options and links to external resources [Figure 4]. GeneCards is mirrored by academic institutions throughout the world, and has attracted wide academic and commercial interest.

• A second locally developed genomic tool is the Unified Database (http://bioinformatics.weizmann.ac.il/udb/). This resource organizes genomic information by locating genes, EST clusters, polymorphic markers and sequence tagged sites along the human chromosomes. Relevant information for UDB is automatically extracted from public, Web-accessible resources that contain human genome mapping information (such as Genome Database, Whitehead Institute/MIT Center for Genome Research, Stanford Human Genome Center, Genethon, UniGene, and GeneMap'99). Integrated map locations for each genomic object are calculated from the separate method-specific maps (genetic linkage, radiation hybrid, and content-contig maps) and presented on a megabase-scale integrated map. UDB thus provides straightforward answers to questions such as: "Which polymorphic markers are located in the vicinity to a given gene?" and "What is the approximate distance between two markers, each mapped by a different method?" Two

HUGO = Human Genome Organization
UDB = Unified Database

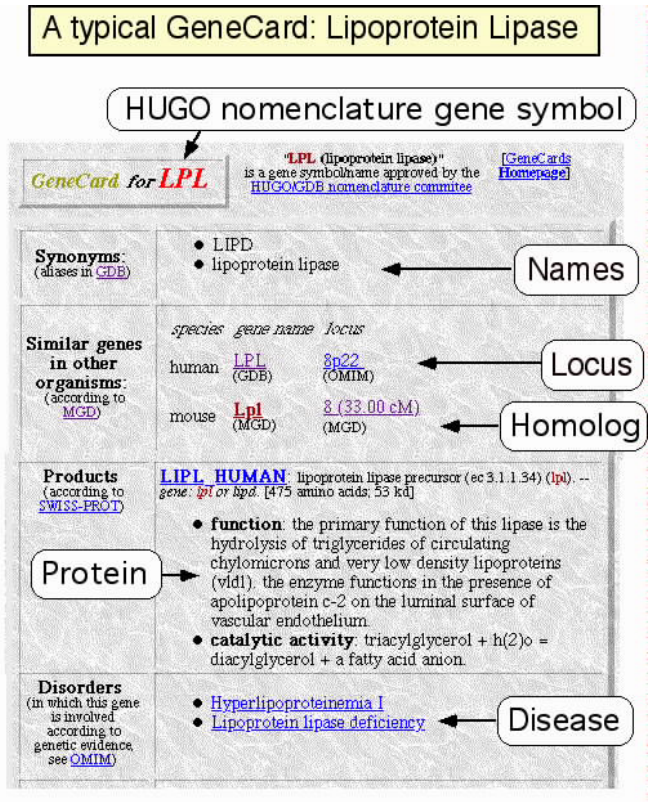


Figure 4. A typical GeneCard – example of the information that is available via the GeneCards database (see text)

modes are available for searching UDB. The first is by specifying the region of interest in the chromosome, in terms of megabase range or cytogenetic band; the second is by indicating object name or alias, thus retrieving a map segment in their neighborhood. UDB can serve as an important tool for fine mapping and positional cloning of disease genes, as well as for development of genetic diagnostic tools. An example of harnessing the power of these engines has been the development of the specialized Human Chromosome 21 Database at the Weizmann Institute of Science namely CroW21 (http://bioinformatics.weizmann.ac.il/chr21/).

- A third genome-wide analyzer is GESTALT (Genomic Sequence Total Analysis and Lookup Tool) [12] (http://bioinfo.weizmann.ac.il/GESTALT). GESTALT is a software tool that integrates several existing sequence analysis algorithms to create a single graphical display for visualizing a given genomic sequence. Thus, it enables gene finding and automatic annotation of large novel genomic sequences. It has an important role in our efforts to identify genes for genetic diseases [Figure 5].

Guides and support for database searches

Web sites and software tools for genomic and molecular biology data are available from the "Bioinformatics and Biological Computing" homepage (http://bioinfo.weizmann.ac.il). Major databases and software packages are locally mirrored on the Bioinformatics and the DNA and Protein Sequence Analysis Server (DAPSAS) computers for swift access by Israeli researchers. Training courses, email and personal support for using these resources are available by the Genome Infrastructure staff (http://bioinformatics.weizmann.ac.il/genome_center/help_desk.html). Many of these activities are conducted in collaboration with the Israel National Node (INN) of EMBNet.

A recent web site focusing on human genomic sequences, clones and maps, "Human Genome – The Third Millennium"

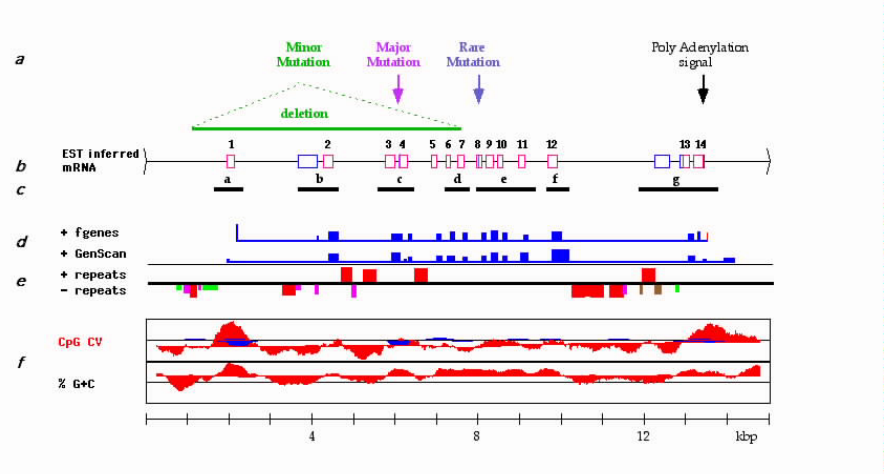


Figure 5. Gene and mutation analysis using the GESTALT program. The DNA sequence of the linkage region for gene mucolipin1 (mucolipidosis type IV) has been subjected to two major analyses: gene prediction programs [**d–f**] and homology search (BLAST) against the EST databases [**b**]. The newly discovered gene structure includes [**a**] genetic landmarks, and [**b**] exons (red boxes) inferred by EST homology searches. EST sequences not included in the putative mRNA are indicated by blue boxes; [**c**] Genomic segments amplified by PCR for mutation analysis; [**d**] Gene prediction results (fgenes and GenScan). Predicted exons are displayed in blue, with box height indicating exon quality (the scaling is arbitrary but consistent for each prediction program); complete gene structure is underlined in blue; polyA signal is indicated in red; [**e**] Repetitive sequences. *Alu* sequences are denoted in red, MIR sequences in purple, LINEs in green, other interspersed repeats in brown; [**f**] Compositional analyses. CpG contrast values and %G+C are displayed as deviations from the regional average (GESTALT analysis). Three mutations have been identified, as indicated on top, a deletion of 7 exons and two point mutations, one of which was more frequent among the affected individuals.

PCR = polymerase chain reaction
MIR = medium interspersed repeats
LINEs = long interspersed nuclear elements

(HG3M) (http://dapsas.weizmann.ac.il/hg3m) strives to guide researchers among databases containing information accumulated during the worldwide preparation of the first draft of the human genome sequence. The site provides links that are relevant to human genomic information, presents search strategies to help users with little or no experience, and contains practical examples with tips for searching and experimental follow-up.

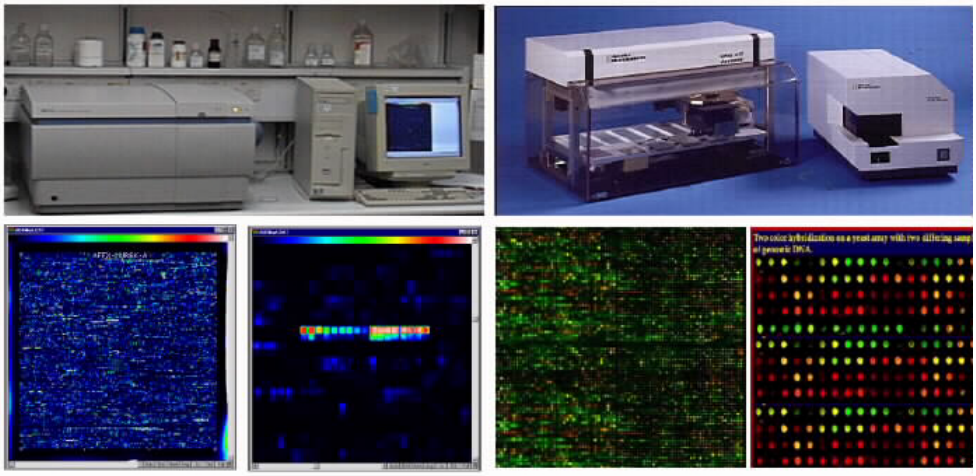


Figure 6. The machinery used for the two major DNA array technologies. **Left:** The Affymetrix GeneChip photolithography system (top) and its high density DNA chip at two different magnifications (bottom). The actual size of the chip is 1.5x1.5 cm. The false color scale represents high hybridization (red) to low hybridization (blue). The high magnification figure depicts an active gene composed of an array of bright oligonucleotide squares, with the lower row being the mutated controls. **Right:** A typical 'spotting' machinery on top (produced by Genome MycroSystems), and its microgrid of spots on the bottom, at two magnifications. The actual size of this grid is that of a microscope slide (2.5x2.5 cm). The red and green colors represent two different probes, with yellow indicating reaction with both.

DNA chips

A DNA chip or array (also known as DNA microchip or microarray) consists of between thousands and hundreds of thousands of DNA segments, immobilized on a solid surface, e.g., chemically pretreated glass. The attached DNA varies from oligonucleotides and PCR products to cDNA or genomic clones. DNA array systems are basically an extension of traditional hybridization methods that have been used in molecular biology for decades. Advanced new technologies enable miniaturization of the components, leading to construction of high density arrays. The technique enables genetic analysis on a massively parallelized scale. The DNA array unit at our center implements the relevant technologies for the benefit of the entire Israeli academic and medical community.

Different DNA chip technologies

Three main technologies are commercially available at present. The two most prominent ones have been integrated at our center [Figure 6], while the third has only recently been released and is slated for implementation at a later time. The different technologies are:

• ***Photolithography-based oligonucleotide arrays.*** The GeneChip technology, pioneered by Affymetrix, is a full "turn-key" solution for DNA array analysis. Both the hardware and the arrays are developed and manufactured as shelf products. For expression chips (see below) the arrays comprise hundreds of thousands of 25-mer oligonucleotides, representing up to 12,000 predefined genes. The oligonucleotides are photochemically synthesized, base by base, on the glass surface using masking technology (photolithography) similar to that employed for computer chip manufacturing. The arrays that are available for expression analyses encompass multiple species, including human, mouse, rat, yeast and *Escherichia coli*. Arrays of additional organisms will be available in the near future.

Other classes of photolithography arrays are re-sequencing or mutation-detection chips, e.g., for the p53 oncogene and for the cytochrome P450 biotransformation enzymes that play a key role in pharmacokinetics. There are also arrays for pathogen identification (HIV) and for scoring SNPs (see below).

The Affymetrix GeneChip system, including a fluidic station and a dedicated scanner, was installed at our center in November 1998 and is being used routinely by numerous researchers throughout Israel. It serves as a nucleus of collaboration with the medical community, notably at the Sheba Medical Center (Prof. Gideon Rechavi).

• ***"Spotted arrays."*** This technology involves deposition ("printing") of DNA fragments of choice, thus allowing the generation of customized chips ("do it yourself"). Researchers can design their own chips by controlling the type of genes as well as the pattern and number of spots (array density). DNA is typically deposited at 100 μm/spot, using a delicate robot with adequate pins (solid or split). The robot transfers the samples from microtiter plates (96 or 384 wells) to the glass surface (a chemically pretreated standard microscope slide). The deposited DNA is then immobilized on the surface by a range of chemical reactions. The resulting arrays serve as high density targets for hybridization with the queried RNA or DNA.

The main advantage of the technology is its flexibility, allowing one to work with the species of choice and an individualized selection of genes. Moreover, it is very cost effective when a large number of identical chips are being prepared. The disadvantages are the spot-to-spot and chip-to-chip variation, as well as the cost and labor of purchasing and maintaining large libraries of DNA segments.

Our center has recently implemented a configuration of a BioRobotics spotter (http://www.BioRobotics.com) and a GSI Lumonics scanner (http://www.gsilumonics.com). The spotter (or "arrayer") is capable of producing microchips with up to 32,000 spots on a single microscope slide. It can process up to 108 microscope slides and 24 source plates at the same time. In addition, the "Total Array System" configuration that is installed in the center is capable of performing macroarraying on nylon membranes, as well as re-arraying of single clones. The ScanArray 4000 scanner (GSI Lumonics) has three user-selectable confocal lasers with an excitation range from 488 to 633 nm. It allows analysis of microchips that were hybridized to complex probes with up to three different fluorescent dyes. Accompanying software quantifies and compares the hybridized spots.

• ***Inkjet chips.*** This technology also involves deposition of DNA. However, in contrast to physical contact, the DNA is injected onto the slide in a similar fashion to that of inkjet color printers, resulting in the building up of DNA fragments on the chip base by base, using regular oligosynthesis chemistry. Thus, it shares attributes with both the photolithography and spotter methods.

Applications of DNA arrays

All the applications described below may, in principle, be implemented using any of the three technologies listed in the previous section.

• ***Expression profiling.*** This allows a glimpse into the inside of the cell, monitoring the RNA at a snapshot, thus determining which genes are expressed and to what extent. Expression-profiling experiments are typically comparative, examining up- or down-regulation of genes compared to a given baseline.

The immobilized DNA is hybridized to RNA purified from the tested cells or tissues. The biotinylated RNA fragments bind to the complementary DNA fragments on the microarray and are then fluorescently labeled and detected by a high resolution laser scanner. Dedicated software then interprets the signal intensity and generates a table indicating expression levels, as well as exact comparative expression ratios for each gene.

• ***Mutation detection (re-sequencing).*** Here, all fragments represent a single gene, and each spot is used to query a different base along the gene. Hybridization of the array with DNA extracted from a tissue/blood sample of a single individual (a candidate carrier of a mutation) will result in identification of the DNA bases differing from the wild-type sequence. It is most likely that in the foreseeable future this application will become routine in hospitals and genetic counseling centers.

• ***SNP chips.*** This application allows one to score bi-allelic single nucleotide polymorphisms in individuals within a population. The currently available Affymetrix SNP-chip has 1,500 exclusively defined SNPs. These chips resemble the re-sequencing chips, in that genomic DNA of an individual is hybridized to the chip to determine the allele in each of the SNP sites. These chips represent a novel technique for gene identification by linkage analysis, parallel to microsatellite (or short tandem repeat) screening. SNP chips will revolutionize personal identification in forensic medicine, and will serve as an utterly powerful tool for genotyping of multigenic diseases.

Significance and prospects of DNA chips

Although the field of DNA arrays is still in its validation phase, the power of the technology is appreciated worldwide. There is no doubt that the post-genome era will depend largely on this technology, which is the immediate tool for reaping the benefits from the vast amount of information accumulated in the various genome projects. DNA chips allow a highly parallelized, high throughput analysis of the functional significance of pre-discovered sequences. Pharmacological industries use chip technologies for both identifying target genes for the development of drugs (thus identifying potential side effects), as well as in the attempt to design personally tailored drugs for individuals, depending on their genetic content (SNP haplotype) to increase drug efficacy.

In the post-genome era, when all genes are sequenced and annotated, it is likely that DNA chips will carry the entire genome of an organism for functional assays. In fact, a whole-genome human expression chip is likely to be issued soon by Affymetrix. In addition, restricted groups of genes affecting a single phenotype (e.g., all oncogenes) can be placed on a single chip to allow the detection of the exact mutation causing the disease. Similarly, genes contributing to complex diseases (such as asthma or hypertension) can be monitored. Last but not least, chip technologies are likely to contribute much to prenatal diagnosis of genetic diseases.

Summary

The post-genome era is at our door, and soon the complete human genome sequence will be available for the next set of goals. Israel is well equipped and skilled to join the worldwide harvest of the human genome, but additional massive government investment is required. This will affect various domains of activity, including the fields of diagnostics and therapeutics. The technologies and know-how described above constitute the basis for future human genome applications in Israel.

Acknowledgments: This work was funded by the Crown Human Genome Center, by the Ministry of Science to the National Laboratory for Genome Infrastructure, by the Krupp foundation, the German-Israel Foundation for scientific research and development, and the Weizmann Institute's Glasberg, Levy, Nathan Brunschwig and Levine funds. Doron Lancet holds the Ralph and Lois Chair in Human Genomics.

References

1. Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park H-S, Toyoda A, Ishii K, Totoki Y, Choi DK, et al. The DNA sequence of human chromosome 21. *Nature* 2000;405(6784):311–19.

2. Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, Ainscough R, Almeida JP, Babbage A, et al. The DNA sequence of human chromosome 22. *Nature* 1999;402(6761:489–95.

3. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, and the members of the DOE and NIH planning groups. New goals for the U.S. Human Genome Project: 1998–2003. *Science* 1998;282:682–9.

4. Ewing B, LaDeana H, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred. I: Accuracy assessment. *Genome Res* 1998;8:175–85.

5. Staden R, Beal KF, Bonfield JK. The Staden package, 1978. *Methods Mol Biol* 2000;132:115–30.

6. Miller MJ, Powell JI. A quantitative comparison of DNA sequence assembly programs. *J Comput Biol* 1994;1(4):257–69.

7. Glusman G, Sosinsky A, Ben-Asher E, Avidan N, Sonkin D, Bahar A, Rosenthal A, Clifton S, Roe B, Ferraz C, Demaille J, Lancet D. Sequence, structure, and evolution of a complete human olfactory receptor gene cluster. *Genomics* 2000;63(2):227–45.

8. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 1998;14(8):656–64.

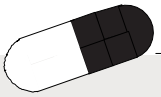
9. Riboldi Tunnicliffe G, Gloeckner G, Elgar GS, Brenner S, Rosenthal A. Comparative analysis of the PCOLCE region in Fugu rubripes using a new automated annotation tool. *Mamm Genome* 2000;11(3):213–19.

10. Harris N. Genotator: a workbench for sequence annotation. *Genome Res* 1997;7(7):754–62.

11. Glusman G, Lancet D. GESTALT: a workbench for automatic integration and visualization of large-scale genomic sequence analyses. *Bioinformatics* 2000;16(5):482–3.

Correspondence: Drs. E. Ben-Asher and D. Lancet, Dept. of Molecular Genetics and Crown Genome Center, Weizmann Institute of Science, Rehovot 76100, Israel. Tel: (972-8) 934-3099, 934-3683; Fax: (972-8) 934-4112; email: edna.ben-asher@weizmann.ac.il, doron.lancet@weizmann.ac.il

Capsule

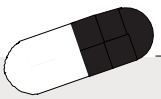


Robust survival strategy?

The eubacterium *Deinococcus radiodurans* has an extraordinary ability to survive radiation-induced DNA damage. A clue to its survival strategy has emerged from a study of the Rsr protein, which is an ortholog of the 60 kD eukaryotic protein Ro and was identified in the recently sequenced *D. radiodurans* genome. Although the function of Ro in eukaryotic cells is unclear, it associates with small cytoplasmic RNAs (called Y RNAs), as well as 55 ribosomal RNA, and is a major target in some human autoimmune diseases. Chen et al. have found that a *D. radiodurans* strain deficient in Rsr is significantly more sensitive to ultraviolet radiation than the wild type, and that UV irradiation causes accumulation of a Y-like RNA that forms a complex with Rsr. Thus, Ro-like ribonucleoproteins play a role in the recovery of UV-irradiated *D. radiodurans* and, in principle, could do likewise in eukaryotic cells.

Genes Dev 2000;14:777

Capsule



Mating in *Candida*

Candida albicans is a ubiquitous inhabitant of human beings, often as a nuisance, but increasingly as a serious pathogen. Its identity has been known for 80 years, but only very recently has it been suspected of having sex, despite the sexual nature of similar yeast-like organisms. Now two groups by two different routes have discovered mating in *C. albicans*. Hull and Johnson consolidated their recent observation of a mating locus by engineering a series of strains with deletions in components of the mating type-like locus (*MTL*) and identifying the progeny of pairs of various combinations of engineered strains that have mated during infection in mice. They only found progeny, with increased DNA content, from pairs of strains with appropriate genotypes – in this case, *MTLa* and *MTLa*. In contrast, Magee and Magee generated hemizygous strains by metabolic selection (something that could easily happen in nature) on sorbose-agar plates. This group also achieved mating between the resulting *MTLa* and *MTLx* strains and confirmed that these progeny were tetraploid. What both groups have yet to obtain are spores formed after meiosis.

Science 2000;289:307,310