Indian Journal of Biochemistry & Biophysics
Vol. 48, August 2011, pp. 215-225

Review

Recent Advances in i-Gene Tools and Analysis: Microarrays, Next Generation Sequencing and Mass Spectrometry

Michael J Moorhouse[1*] and Hari S Sharma[2]

[1]Department of Blood Cell Research, Stichting Sanquin Bloedvoorziening, Postbus 9190, 1006 AD Amsterdam
[2]Department of Pathology, VUmc, University Medical Centre, Amsterdam, The Netherlands

Received 24 March 2011; revised 25 July 2011

Recent advances in technology and associated methodology have made the current period one of the most exciting in molecular biology and medicine. Underlying these is an appreciation that modern research is driven by increasing large amounts of data being interpreted by interdisciplinary collaborative teams which are often geographically dispersed. The availability of cheap computing power, high speed informatics networks and high quality analysis software has been essential to this as has the application of modern quality assurance methodologies. In this review, we discuss the application of modern 'High-Throughput' molecular biological technologies such as 'Microarrays' and 'Next Generation Sequencing' to scientific and biomedical research as we have observed. Furthermore in this review, we also offer some guidance that enables the reader as to understand certain features of these as well as new strategies and help them to apply these i-Gene tools in their endeavours successfully. Collectively, we term this 'i-Gene Analysis'. We also offer predictions as to the developments that are anticipated in the near and more distant future.

Keywords: i-Gene analysis, Microarrays, Next Generation Sequencing, Mass spectrometry

Introduction

Scientific and medical research has always required the collection of data as part of the classical 'scientific method' paradigm to which most studies conform. What is unprecedented is the ability to generate data in such large quantities in an automated robust manner that derives an inquisitive research community to do so for the benefit of mankind. While it is difficult to separate cause and effect as to explain the current situation, a description would undoubtedly include coverage of:

- The intelligent use of basic biological chemistry concepts and biological occurring albeit in a modified, 'ruggedised' form – enzymes and molecules combined to design new protocols[1].
- The increase in cheap computer hardware to both drive automation in sample preparation stages and assist in the storage, processing and transmission of the acquired data and knowledge[2,3].
- Physical components that are mass produced, reliable and cheap which are available for instrument construction.
- The increase in the use of commercial integrated workflow systems in which one manufacturer supplies an entire processing pipeline from basic reagents, though branded/approved equipment, defined work protocols, training and customer support to primary data analysis algorithms and software. This has provided the necessary quality assurance to make many such studies robust enough to be used practical.
- A change in the mind-set of investigators to use such pre-prepared workflows which often involve committing to use the one almost exclusively compared to recruitment of a large workforce of semi/highly skilled staff.
- A willingness to accept that such commercial workflows are often far from perfect being often little more that advanced prototypes and hence will be adapted and updated while a study is in progress.

Hypothesis and experimental design

The design of scientific and medical experiments is a combination of the human imagination and desire underwritten by the technology available to implement the plan. The 'scientific method' demands the review of previous experience, the proposal of hypotheses, the design of tests (in the experiment

*Author for correspondence:
Tel: +31-20-5123168
Fax: +31-20-5123252
E-mail: m.moorhouse@sanquin.nl

sciences) with the collection of pertinent data, followed by critical assessment of the results of these test and whether they support the hypotheses. Previously, hypotheses were often confined to the assessment of single genes as this was all that was practical to study i.e. "Is the VEGF gene expressed in [tissue]?" due to the costs involved both financial and labour. With the ability to assess potentially all genes (or even spliced transcripts) – see the next section for a review of the technology – with i-Gene analysis this becomes: "What genes are expressed in [tissue]? Is it VEGF as the literature would suggest?" This has lead to the accusation, especially informal during aural presentations that modern genomics studies are 'not hypothesis driven.' We assert that this was incorrect as the hypothesis was merely more general and therefore could not recognised as such. It is possible to propose, even if it often undesirable, hypotheses that are as general as "We assume that we can use an expression array (or alternative technology – see next section) to reveal interesting genes for further study." An extreme example of this is the genetic profiling of longitudinal/prospective cohort studies. Here it is assumed that the genotype profiles will be of use when combined with phenotypical parameters long after the initial studies have been completed. Here the hypothesis could be wistfully stated as: "We believe that this resource -that is expensive to create - will be of general benefit in the future." In these terms, it is understandable why traditional biologists and clinicians are sceptical about it, yet it is common in other fields such as astronomy and space exploration where the high costs and long timescales for data collection often limit the precise definition of experimental aims at the time of commitment.

Hypotheses generally fall into three general categories, though elements of each can often be found in be found in many studies. For each, we present an example hypothesis based the same topic of interest.

Fundamental science studies

These are the most classical of investigations, where the scientific question is well defined and the application of genomics technologies is an aid to investigating a well defined hypothesis usually in a well studied topic area[4]. This is application of modern technology to a classical study. An example hypothesis is 'the VEGF gene is involved in growth of new blood vessels in tissue X'.

Lead generation

In this case, the hypothesis is far broader than 1) used when there is no, little or poor previous research in the field to drive the investigation and the hypothesis. An alternative scenario is that researchers are unaware of new directions. The aim of such a study is to generate interesting 'leads' to investigate further by other methods. An example hypothesis might be: 'Genes are involved in the recruitment of blood vessels, and we wish to identify these for further study.

Profiling studies for the purpose of classification

In this mode, genomics tools are used without a desire to extract at least in the primary instance-scientific meaning from the results obtained but rather the motivation is to obtain a numeric classification of the samples into groups. The definition of groups *a-priori* is not a fundamental requirement as this can be derived and reassigned to fit alternative experimental designs. If generality of the classifier obtained is to be claimed then care must be taken that the classifier developed is not 'over optimised' (often called 'over fitted') on the data contained in the current experiment. Space prevents an extensive discussion of this topic here, but it is a trap that such investigations should guard against if they are to be taken seriously. An example hypothesis might be: 'Sub-groups of patients exist that respond differently to treatments of myocardial infarcts. A numeric classifier to allocate future patients to different treatment programs will be derived from the current data'.

Fundamental concepts

In this review, we describe in detail three of the main data collection genomics technologies that have been developed recently and used to gather data for i-Gene analysis focusing on those we have direct experience of. We then cover some of the analysis techniques we have found to be useful for interpreting the resulting data. Prior to this, we review the 'DNA makes RNA makes Protein' paradigm and its use as a information flow model for i-Gene analysis and then data processing and analysis operations.

DNA-RNA-Protein axis in the context of i-Gene analysis

The 'DNA-makes-RNA-makes-Protein' is a classical paradigm of biology. The description here is deliberately general and omits a multitude of nuances

that are both interesting scientifically and medically. We prevent it here in sufficient details the implications in the context of information flow with a cell as this will be useful for the following discussions on i-Gene analysis.

DNA generally exists in low copy number (as low as 2 molecules per cell) and it is the master copy of the information available to the cell. From it, many copies are made in RNA that leaves the DNA unchanged. The RNA ultimately drives the synthesis of proteins which have diverse functional roles (catalytic, structural, signalling and regulatory) within the cell. As proteins and RNA are chemically different to DNA, they can be processed and degraded without damaging the reference DNA. This turnover allows for regulatory processes to act and hence the cell to respond to the environment. The 'Reserve Transcriptases' classes of enzymes that convert RNA into DNA are extremely useful as they allow the technologies that operate on DNA to be applied to RNA with ease[5].

Data storage, information processing and knowledge interpretation

The results from the technologies described below are ultimately a list of IDs in a table or a database, possibly with metrics on amount, confidence and ratio relative to the other samples studied included alongside. This data must be stored initially, processed by some combination of computer encoded algorithm and professional judgement, followed by interpretation to meaningful knowledge. We discuss this in more details in a later section.

Genomics technologies for i-Gene analysis

Presented here are three of the most interesting technologies that are most amenable to large-scale studies and gaining in popularity in research. We give an overview first and then a more detailed description of their nuances.

Microarrays

These exploit the binding of DNA labelled with a fluorescent dye prepared from a sample to DNA of complementary sequence commonly termed 'probes' attached to an inert substrate[6]. The amount of DNA bound after the non-complementary labelled DNA has rinsed off is assessed by exciting the dye using a laser scanned across the array surface. The resultant fluorescence intensity is transformed using mathematical and statistical methods to produce a quantitative value of the amount of DNA bound and hence present in the sample. The density of modern microarrays allows approximately many millions of probes/features to be included per array with content beyond this being split across 'array sets' (for example, the Affymetrix Human Tiling 1.0R Array Set that contains 14 arrays in total). Physically, microarrays are around the size of a classical microscope slide, typically 3 inch by 1 inch or 75 mm by 25 mm, the exact dimensions being manufacturer-specific. On some array platforms, gaskets can be used to isolate different sections of the surface from each other so multiple samples can be run on the same array albeit with reduced coverage for each to increase sample throughput (see Fig. 1).

All the types of microarray described in this review use DNA that has been chemical synthesised, this being the current technology of choice for most applications. An alternative is to 'spot' DNA from biological sources, a clonal BAC (Bacterial Artificial Chromosome) cDNA library for example as probes to form 'spotted arrays'[7]. While this allows the use of larger probes which reduces the number of sample points (a BAC library with 32 000 clones with insert

Fig. 1—Example of microarray (Illumina Bead Array ™, Quad 610 shown) the four regions on the surface are separated from each other using a gasket to increase the number of samples that can be processed on one microscopic slide as one array [Reproduced with copy right permission from Illumina Inc, San Diego]

sizes of ~150 kb can tile across the human genome of ~3 Gb for example) poorly characterised DNA fragments can cause problems for data interpretation and the hypotheses that can be addressed. One application where spotted arrays based on BACs have been used until quite recently is the field of human molecular diagnostic, where a comparison of change healthy: normal is sufficient and the knowledge of the location can then translated into a validated protocol for medical use. One drawback of spotted arrays is the need to prepare and most likely store the material for each spot separately prior until immediately prior to addition on to the array surface. While possible the lab management issues involve processing this should not be underestimated.

The density of modern arrays is high enough that many millions of probes can be included on one array. This allows a high number of biological features to be studied with multiple fold redundancy and also the sequence of the probes to be varied which are then combined numerically to either improve reliability or to study different aspects of the same feature – a gene, transcript or exon for example. The results from the individual probes can then be combined together in the analysis software.

This basic technology can be applied to a wide variety of studies, the general criterion being a predictable DNA sequence as an output that requires quantification. This includes genotyping and genome structure studies as well to transcriptome analysis and transcription factor pull down studies.

While a powerful technology, it has three main limitations:

- The sequences of the 'probes' must be selected during array manufacture. Thus, the array can be designed using only existing knowledge of the system it likely to assess which limits its ability to discover truly novel features.
- The hybridisation of labelled DNA is sequence and melting temperature dependent, further complicating the selection of probe sequences.
- Due to the high concentration of DNA present, there is a high likelihood of cross-hybridisation of labelled sample to probes with which it is not complimentary, leading to increased non-specific background luminescence.

The major manufactures of array systems are: of Affymetrix Inc.; Illumina Inc., Agilent Technologies Inc. and Roche NimbleGen, Inc.

High throughput genome sequencing (HTGS) or next generation sequencing (NGS)

This is a new set of technologies that produces many hundreds of millions of DNA sequences per sample at a low cost. In contrast to classical chain termination (aka. 'Sanger Sequencing') sequencing which produces reads of 600 to 900 bp from a single DNA template, NGS techniques result in many millions of short DNA (70 to 800 bp depending on the technology used) reads per sample. There are a variety of systems that are undergoing rapid development and vying for technical and commercial dominance. Some of these are more robust who's characteristics are becoming known while others the so-called 'next-next generation' exist as prototypes or have entered trials with early adopters.

In contrast to microarray sample preparation, the DNA to be sequenced via NGS is diluted to low concentration and then covalently attached to an inert substrate, such as a glass slide or beads, so that one molecule is physically separate and distinguishable from neighbouring ones. An amplification step is used to increase the number of identical copies in that area which ultimately boosts the resultant signal during sequencing increasing the signal to noise ratio remarkably.

The use of 'pair end' strategies, where both ends of the DNA fragments immobilised on the inert surface are sequenced is an extremely useful enhancement to NGS[8]. In the simplest case, it provides an easy method to double the sequence output with little extra cost in terms of sample preparation though the real use is in structural genome studies. The physical linkage of the two reads in the same piece of DNA a defined distance apart can be used to build a scaffold during a sequence assembly project, which is especially difficult in repetitive genomic regions. Another important medical application is the identification of structural rearrangements (inversions, translocations, duplications and deletions) by reference to a pre-assembled genome to identify aberrant regions.

As depicted in Fig. 2, there are three basic technological strategies (Illumina-Solexa, 454 Pyro sequencing and ABI SoLID) are being used in robust NGS systems currently:

Sequencing by synthesis

The complementary strand to those attached to the surface is synthesised one base at a time by the addition of a series of fluorescent nucleotides which is

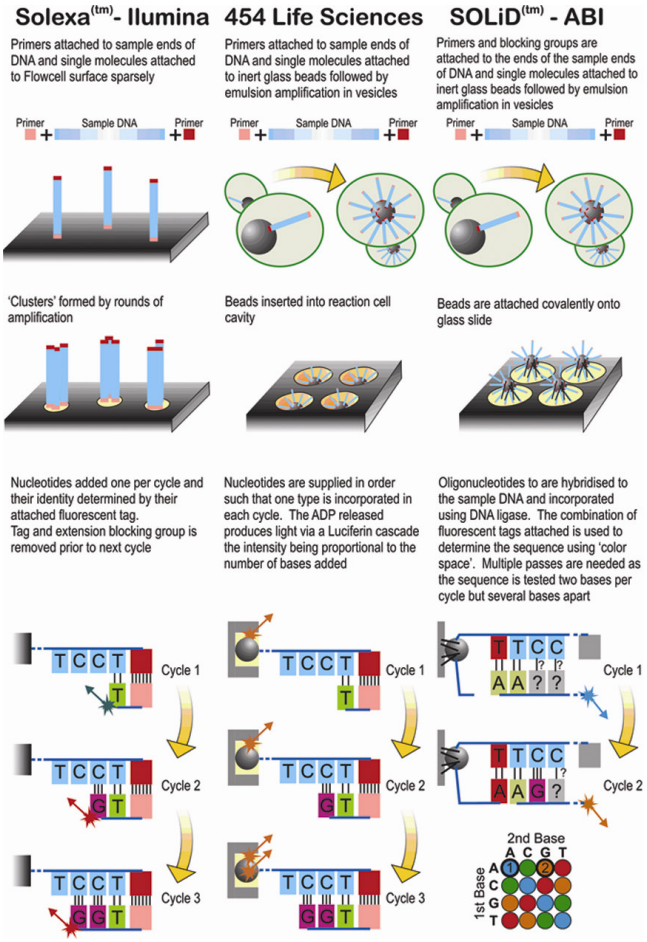

Fig. 2—Technological explanation and comparison of three most commonly used New generation of Sequencing (NGS) systems [Illumina-Solexa, 454 Pyro sequencing and ABI SoLID™]

then read with a laser and camera and lens system. The fluorescence is destroyed (or removed depending on the system) after each step to prevent confusion between the signals reported. As the base incorporation step is imperfect, in rare circumstances two bases may added or none instead of one causing 'phasing errors' in the resultant sequence. Sequencing by synthesis is one of the 'short read' technologies with a maximum read of around 150 bp. The Solexa (tm) (Illumina Inc.) and tSMS (tm) (True Single Molecule Sequencing) (Helicos Bioscience Inc.) systems make use of this methodology.

Pyrophosphate sequencing

The complementary stand to those attached to the surface is elongated by a polymerase enzyme. As each of the four nucleotides is supplied in a known order via the reaction mixture flowing across the reaction cell, the extension stops at the end of consecutive runs of the same nucleotide (termed 'homopolymer runs'). For each base incorporated, one pyrophosphate molecule is released which in turn is used by luciferin enzyme to produce light and the amount of light emitted being proportional to the number of incorporation events. An exact read length in bases/nucleotides cannot be known *a-priori* as this dependent on the bias in sequence composition of the sample being sequenced, but 800 bp reads are possible with modern systems such as those from the Genome Sequencer FLX System of 454 Life Sciences which uses this methodology.

Sequencing by ligation

This uses short oligonucleotides probes of 10 bases or less that hybridise specifically to the sample DNA bound to the surface in a manner similar to in microarrays. While the identity of the first bases usually position one and two at the 3' end of the probes are known, the rest of the sequence is degenerate (i.e. a random combination of bases). Attached to each 5' end of each probe is a fluorescent corresponding known base(s) at the 3' end which is read using an excitation/detection step and then cleaved to prepare the 5' end for the next hybridisation. As one or two bases are determined out of the total length of the probe, multiple passes are required to sequence the entire molecule. This is commonly accomplished using a process called 'primer reset', where the elongated strand removed and the hybridisation commenced from a start point such that different bases are sequenced. The ABI, Inc. SoLID™ system is the current market leader in the application of this technology. This uses a 2-base encoding method and the concept of 'color space' to improve accuracy and can differentiate sequencing errors from real genomic polymorphisms, as only certain color space transitions are valid when compared to a known reference sequence. Currently, the SoLID™ supports a 75 bp read length and up to 20 million reads per sample.

An important point is that NGS methods process the material they are given with little bias compared to other methods, for example microarrays which will report only the report intensities bound to pre-defined probes. This alters the hypotheses from "How much of sequence X, Y and Z is present in my sample?" to "What sequences are present in my sample and at what quantity?" While this is highly desirable in many applications, pre-selection often termed 'targeted sequencing' is more desirable in

others. Examples of this include gene expression studies that need to separate messenger RNA from other types or gene re-sequencing studies. Here, the same approaches developed for Sanger sequencing can be used: SAGE methodologies for gene expression and PCR amplification/cloning strategies of selected regions.

One of the perhaps surprising applications of microarrays is their use as enrichment tools for DNA samples prior to NGS. While suffering somewhat from the same problems as microarrays in general – the need to select the sequences to enrich for – which restricts the ability to detect totally new sequence features, the hybridisation conditions and probes can be adjusted such that small mismatched (SNPs or possibly insertions-deletions (indels) will still be captured and present in the enriched mixture. Roche NimbleGen Inc. for example has an integrated pipeline for the supply of capture/enrichment arrays from the design of oligonucleotide probes based on genomic regions of interest supplied in a reference genome through array delivery and protocols to individual researchers to use.

Mass spectrometry

The study of the proteins in a cell must rely on fundamentally different technique to those described previously, as there is no practical polypeptide to polynucleotide conversion process. While protein arrays similar in concept to DNA microarrays, but with proteins or antibodies spotted on to the surface are available, they suffer from the same problems of the older cDNA microarrays: low density, the requirement to prepare the material for each 'spot' separately, as well as the problem of deciding upon the content of the array. The three dimensional structure of a particular protein may be extremely useful scientifically for allowing insight into the underlying biology, but determining it is still a time-consuming, laborious process and far from high throughput.

For i-Gene analysis, mass spectrometry is a technology of choice for reasons of speed and general applicability. The basic principle is that the proteins are extracted from the sample and a mass spectrometer used to determine their mass which is then compared to protein databases to determine their identity. The sample must be volatised and ionised prior to analysis with the two most common methods used to do this being Electrospray Ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In an Electrospray System, the liquid sample containing the proteins is passed through a very small electrically charged nozzle such that a fine aerosol results; with MALDI, the sample is crystallised on to an inert substrate along with the 'matrix' that assists with ionisation and protects the sample when it is subjected to a laser beam. Ions are needed, so that the sample can be manipulated using electric fields inside the instrument.

While conceptually simple, there are many complicating technical problems at the level of data collection and subsequent analysis. For example, the instruments themselves have excellent mass resolving capabilities, but a fundamental limitation is that they detect a single mass at one time albeit very quickly, so a scan across a range of masses can be compiled into a single spectrum and many can be taken per second. These spectra can be extremely complicated as they are conceivably a composite of all the proteins in the sample. Further, the primary data are spectra of mass/charge ratio which must be post-processed, involving filtering, peak finding, the masses lookup in a database containing those known before a final list of identified proteins is compiled. Another complicating factor is the wide range of concentrations of proteins in samples that can be 10 orders of magnitude and can result in those present at low copy being missed. To help alleviate these problems, fractionation/separation methods, such as the use of 2D Gel Electrophoresis and HPLC (High Performance Liquid Chromatography) are often applied prior to analysis of the sample in the instrument.

There is a plethora of competing mass determination technologies in use and under active development, for details see[9,10]. A key technique is Tandem Mass Spectrometry (often called MS/MS) that involves the coupling of two mass selection stages together in the same instrument together by a fragmentation stage. This allows the selection of a particular protein or peptide at the first MS stage which is then fragmented at random points along its peptide backbone by collision with a gas at low pressure. The masses of the resulting fragments can be used to determine snippets of the amino acid sequence by reference to the masses of the amino acids expected thus aiding in identification.

Overall there are three general approaches: 'top-down', 'bottom-up' and 'shotgun'.

In 'top-down', whole proteins are supplied to the instrument with a minimum of pre-processing thus giving a more complete impression of the protein complement of the cell and the post-transitional modifications. This results in complex spectra and large data sets for which high mass accuracy and the ability to fragment the proteins inside the instrument is highly desirable.

In the 'bottom-up' approach, the proteins are separated at least partially on 1D or 2D gels and converted into peptides using either chemical or enzymatic means. Due to the simplicity of the spectra obtained, the resulting mass fragments are often sufficient to identifying the original proteins.

For the 'shotgun' approach, the proteins in the sample are cleaved into peptides followed and identified using MS/MS technology. Often an extra HPLC is used to help separate the peptide as they are processed by the instrument. Typically, manufacturers supply analysis software for filtering of spectra and possibly peak identification. The conversion of these into real protein identifiers (Ids) is the preserve of more specialist algorithms of which many examples exist including the popular Mascot (of Matrix Science Inc.) and SEQUEST[11] There are many manufactures of mass spectrometer instruments, the principle ones are: Waters Inc., Bruker Daltronics Inc. and Thermo Scientific Inc.

Data storage and analysis

Given the data volumes involved and the complexity of analysis in i-Gene analysis, the use of computers and largely automated processing is inevitable. This may be discouraging for those researchers who are highly skilled in the laboratory, but are uncomfortable with the use of computers. However, great improvements have been made recently in the usability of software due to the increased maturity of the field coupled with the application of best practices of commercial software development. We now describe the data storage of i-Gene data both raw and processed and some example methods used.

Data storage and computational requirements

The volume of data resulting from i-Gene analysis can be large, in the Gb (Gigabyte) to Tb (Tetrabyte) range and thus may require specialist handling techniques. Fortunately, the solutions developed in the fields of ICT (Information Communications Technology) to solve similar problems can be applied easily. For example, commodity hardware – computer workstations and servers, operating system such as Microsoft's Windows (tm), Linux and Apple's OS X (tm), networking switches and disk/disk arrays can be bought very cheaply and used almost without modification while being easy to setup.

The dataset sizes resulting from i-Gene style analysis can be large. For example, the 'CEL' file containing the intensities of the probes of an Affymetrix style GeneChip (tm) can be >50 Mb, while the raw image scans can be >250 Mb. While each is relatively small, the effort of storing and indexing even 100 to 1000 which may result from a typical experiment becomes non-trivial.

With NGS, these same problems are even more acute: a single Flowcell containing eight samples run on a Solexa Illumina Genome Analyzer II can generate 1 Tb of raw images and a further 300 Gb of primary analysis over a period of three days. Also, this is split across hundreds of directories containing thousands of files which can be highly sub-optimal for some disk systems and network transmission.

A compounding factor is often that intermediate results created with slightly different parameters can need storing which can lead to the same data appearing in many slightly different forms. This creates a problem for traceability as which version of the partially analysed data often become hard to determine unless proper records are kept.

A point that is often overlooked is the need to store data in an accessible form after the experiment has been completed and the results published. While journals and increasingly funding agencies have their own specific policies, all encourage as open access to data as quickly as possible for the reasons of scientific scrutiny to confirm the finding and addition to the corpus of scientific knowledge for general reference or re-use in meta-studies. Notable exceptions to this are where patient/participant confidentially may be compromised (an issue where comprehensive genetic profiles are used for example) or commercialisation/patenting of the results is being considered. Whatever the caveats, a period of five years after publication and longer for commercial use are not unusual. During this time, equipment must be kept maintained and housed or yearly storage contracts paid for which can severely impact on future research budgets.

For the above some solutions are:

- As mentioned, the commodity computer hardware is suitable and easy to use while cheap enough for researchers to afford. An option gaining popularity is the use of distributed computing in which the analysis is directed from local computers but done off-site at a centralised, possibly commercial, computing facility. An example of this is the Taverna workflow system coupled to myGrid[12].
- Good record keeping should be commonplace in research and data storage is merely an extension of this. The analysis packages are often highly automated and a few key parameters stored with software/packages versions are sufficient. This problem has been anticipated and solved in many software packages. In other more open systems, it is possible to write very terse 'programs' called 'scripts' to direct and hence form a of records of record - the operations done.
- Many public on-line databases/data warehouses are funded to store the majority of the information that must be made available to wider community. For example, Array Express[13] stores Gene Expression and the NCBI Trace Archive[14] stores sequence data. Many journals now require deposition of experimental data into such databases which may be an unexpected burden immediately prior to manuscript acceptance. Also application of formal ILM (Information Life Cycle) management policies as developed in the field of ICT fields can be of great benefit in alleviating data storage problems.

The computational requirements of i-Gene analysis are not high at least from the perspective of the individual researcher for their basic analysis. Little advice can be offered here beyond a medium power workstation as to the exact specification needed as this depends on the analysis done.

Data analysis

For convenience of the discussion, we divide the analysis process into four sections: raw data acquisition and processing, linking/data consolidation using annotation, determination of the interesting data points and conceptual/scientific interpretation. Note that one of more of these may be integrated into the same software package.

Raw data acquisition and processing

The data acquisition is equipment and manufacturer specific, as it involves close interactions with the physical world (spatial locations, laser powers, liquid/vacuum pump activity etc.). The analysis needed is to capture these parameters and store them for further use. The results of this are often a series of technology and manufacturer-specific data files that represent the raw data. For example, a typical output from the scanning of a microarray might be a set of images that resemble those from a microscope, a file containing a single intensity of each probe and a 'meta-data' file listing details such as the date and time of the scan, the identifier of the chip and scanner settings.

Linking of the raw data together into more biologically meaningful units and quality filtering

Often the same basic unit such as a gene will be assessed at multiple points and the results of each can be combined in different ways, depending on the level of granularity required, for example: gene, transcript, exon, SNP. Also, these groupings may be updated based on new information and insight which can lead to improved performance, for example, the re-mapping of probes to transcripts with Affymetrix GeneChip (tm) arrays[15]. Whether it is worthwhile to use these or to remain faithful to the original standard, which allows for direct comparison of results to previous experiments, should be decided upon.

For NGS, the output is a large number of short sequences. These must be aligned either to each other or to known biological sequences a completed DNA genome that is preferably well annotated with gene boundaries, a set of RNA transcripts or ESTs (Expressed Sequence Tags) for further analysis to be done. There are good algorithms to do this for example a variant of SSAHA is used in the Illumina Genome Analyzer Pipeline[16] to align reads to reference genome or VELVET to assemble the sequence reads into larger overlapping regions (called *de novo* assembly).

The raw data should be assessed for quality with bad samples excluded and aberrant ones that do not conform to expectations investigated further, the exact methodology used being dependent on the data types. Often quite crude metrics, such as the overall intensity for microarrays, number of peaks called or read sequences that map on to the reference genome are surprisingly indicative. In many cases, there will be a logical explanation for the problems encountered such

as a failed sample extraction, a bad reagent or mislabelling causing a sample swap error. One of the advantages to using commercial systems is that these have extended diagnostic controls built into the protocols to help identify problems with extensive technical support to describe how to uses these. From the manufacturer's perspective, this offers an enticement to use their system and to demonstrate that technology is operating correctly when a customer claims it is at fault.

Identifying the interesting data points

Due to the amount of data produced, which may be many millions of data points per sample, filtering to identify interesting features of the data is essential. To accomplish this, the data from one sample must be directly comparable to others which is done through a processed called 'normalisation'. The type of output required has an impact on the difficulty of this. SNP data for example is relatively easy to normalise, if the output as the allele call are essentially one of four states: AA, AB, BB or No-Call/not detected as the output itself is allows direct comparison. Further analysis of the genotype calls can then be done using program such as PLINK[17].

In gene expression, where quantitative measures of gene expression are required on a continuous scale then more complicated models must be used. A detailed discussion of the normalisation and analysis techniques used with each technology is beyond the scope of this review, but interested readers can consult others such as those used for expression microarrays[18] and for copy-number studies using SNP arrays[19]. These techniques transform the raw data using a particular set of assumptions to make them comparable such that probes or grouped set of probes that differ significantly can be identified. A classical approach with expression studies is to use a combination of the p-value and fold change to rank the results in order of interest. The p-value gives a measure of the statistical difference between two or more groups, whereas the fold change indicates whether the changes are large enough to be biologically meaningful[20].

The open source Bioconductor framework written in the 'R' programming language offers a comprehensive very suite of processing, analysis and visualisation functions for many biological data types[21]. As R and hence Bioconductor is text-based that can be intimidating to researchers and so graphical interfaces have been created that synthesise functions together into a guided workflow. For example, the 'affylmGUI' package that allows the normalisation and analysis of data produced on Affymetrix GeneChip™ arrays by linear modelling[22].

Increasing the group size i.e. running replicate samples is very helpful for statistical methods that use intra-group variability to exclude aberrant measurements that would bias the inter-group comparison. The number of samples needed per group is dependent on the variability between samples of the same type expected, so it is impossible to offer advice on absolute numbers except that less than three replicates becomes problematic. At this stage, the data is been transformed such that it is comparable to each other and filtered by criterion that define the list as 'interesting' either scientifically or to useful to numerically classify/partition samples.

Scientific interpretation

While a simple list of differentially expressed genes, proteins or SNPs associated with a particular trait is useful, explanation of the results in a biological or medical context is often necessary for assessment of the hypotheses. Described below are the more common interpretation methods to accomplish this. All these draw on existing scientific knowledge present in structured databases and published literature. Many tools exist that use pattern identification and matching techniques to give insight into the data presented some of which are covered below. An important point is that that these tools are excellent at the recall of information, but are far less deductive compared to a trained scientist. While the output of these tools is useful to get an impression of what is already known extension of this must come from the insight of a human researcher which requires real thought. It is a commonly held truism that collecting samples and processing them through commercial genomics systems is easy and the hard part is the data interpretation that can take up to 85% of the length of the entire experiment despite there being excellent tools. Sadly, the complexity of the output reflects the complexity of the input , which is the biological world.

Note that many of the tools mentioned are 'gene centric' due to the underlying corpus of biological knowledge being structured around genes. This may be a problem for the results of certain investigations, for example SNP associate studies that can ignore completely gene structure and return statistically significant results that are not close to any genes.

The classification of the genes on to knowledge frameworks can often be very insightful, the two most commonly used of which are GO (The Gene Ontology)[23] and KEGG (Kyoto Encyclopedia of Genes and Genomes)[24]. GO is functional classification scheme organised in a semi-hierarchical manner such that nodes (called 'terms') 'leaf-wards' away from the most general 'root' node represent more specific versions of the same concept. Genes and their resulting protein products are mapped to nodes by either a process of manual curation or automatic annotation. KEGG is a collection of biological pathways showing the interaction of proteins (and by proxy their corresponding genes) with each other and small molecules. KEGG pathways tend to be comprehensively documented because the underlying proteins must have been well studied for the interactions between them known. Due to the greater difficultly of obtaining high quality information about protein interactions than tentative function classification for GO, KEGG covers fewer biological entities than GO.

As the underlying data for both GO and KEGG are available in a machine readable form and are free for academic use, very good tools for analysis have been created. Two of the most comprehensive and widely used are 'BABELOMICS'[25] and 'DAVID'[26]. These identify GO terms and pathways that are significantly biased representation in the original list supplied, in addition to linking to other databases containing extended information such as protein domain structures and species homologs.

In addition to academic tools, there are a number of commercial packages that contain a combination of extra analysis functionality, proprietary content with formal support and training. Two examples are MetaCore from GeneGo Inc. and IPA from Ingenuity Systems Inc. These are very easy to use in the first instance but extremely powerful, if their nuances are mastered. A large component of the business model of such companies is to paid human researchers to read scientific literature to expand the database of interactions that their tools use. It is unsurprising given the costs involved that such content is closely guarded from automatic download.

An alternative approach is to use a database such as GeneCards to find the research papers describing the function of genes in the list[27]. While this can be very time-consuming, it may be the only viable option with high value experiments in fields that are poorly served by pathway databases. Another benefit is that because the reading done is tailored to a specific gene list by a human scientist, there is a high chance that real insight will be gained into the underlying mechanisms that software alone cannot deduce.

Conclusion

i-Gene Analysis is the application of modern genomics technology to study many data points in many samples in a single experiment. It uses commercial systems such as microarrays, next generation sequencing and mass spectrometry to study the molecular basis of key biological and medical hypotheses. Due to the large amount of data generated this must be extensively filtered to extract the important features of this data, so that they can be properly interpreted to derive conclusion(s).

References

1 Gunderson KL, Kruglyak S, Graige M S, Garcia F, Kermani B G, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan J B, Barnard S & Chee M S (2004) *Genome Res* 14, 870-877

2 Tsai T Y, Chang K W & Chen CY (2011) *J Comput Aided Mol Des* 25, 525-531

3 Pettersson E, Lundeberg J & Ahmadian A (2009) *Genomics* 93, 105-111

4 Schumacher B, van der Pluijm I, Moorhouse M J, Kosteas T, Robinson AR, Suh Y, Breit T M, van Steeg H, Niedernhofer L J, van Ijcken W, Bartke A, Spindler S R, Hoeijmakers J H, van der Horst G T & Garinis G A (2008) *PLoS Genet* 4(8), e1000161

5 Gräslund S, Larsson M, Sterky F, Uhlén M, Lundeberg J, Höög C & Ståhl S (1999) *Biotechniques* 27, 488-498

6 Forster T, Roy D & Ghazal (2003) *J Endocrinol* 178, 195-204

7 Krzywinski M, Bosdet I, Smailus D, Chiu R, Mathewson C, Wye N, Barber S, Brown-John M, Chan S, Chand S, Cloutier A, Girn N, Lee D, Masson A, Mayo M, Olson T, Pandoh P, Prabhu AL, Schoenmakers E, Tsai M, Albertson D, Lam W, Choy CO, Osoegawa K, Zhao S, de Jong P J, Schein J, Jones S & Marra M A (2004) *Nucleic Acids Res* 32, 3651-3660

8 Tuzun E, Sharp A J, Bailey J A, Kaul R, Morrison V A, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson M V & Eichler E E (2005) *Nat Genet* 37, 727-732

9 Han J, Danell R M, Patel J R, Gumerov D R, Scarlett C O, Speir J P, Parker C E, Rusyn I, Zeisel S & Borchers C H (2008) *Metabolomics* 4, 128-140

10 Aebersold R & Mann M (2003) *Nature* 422 (6928), 198-207

11 MacCoss M J, Wu C C & Yates J R 3rd (2002) *Anal Chem* 74 (21), 5593-5599

12 Hull D, Wolstencroft K, Stevens R, Goble C, Pocock M R, Li P & Oinn T Taverna (2006) *Nucleic Acids Res* 34 (Web Server Issue), W729-732

13 Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U & Brazma A (2009) *Nucleic Acids Res* 37 (Database Issue), D868-872

14 Cochrane G, Akhtar R, Aldebert P, Althorpe N, Baldwin A, Bates K, Bhattacharyya S, Bonfield J, Bower L, Browne P, Castro M, Cox T, Demiralp F, Eberhardt R, Faruque N, Hoad G, Jang M, Kulikova T, Labarga A, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Plaister S, Robinson S, Sobhany S, Vaughan R, Wu D, Zhu W, Apweiler R, Hubbard T & Birney E (2008) *Nucleic Acids Res* 36 (Database issue), D5-12

15 Sandberg R & Larsson O (2007) *BMC Bioinformatics* 8, 48

16 Ning Z, Cox A J & Mullikin J C (2001) *Genome Res* 11, 1725-1729

17 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A, Bender D, Maller J, Sklar P, de Bakker P I, Daly M J & Sham P C (2007) *Am J Hum Genet* 81, 559-575

18 Cui X & Churchill G A (2003) *Genome Biol* 4, 210

19 Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Höglund M, Borg A & Ringnér M (2008) *BMC Bioinformatics* 9, 409

20 McCarthy D J & Smyth G K (2009) *Bioinformatics* 25, 765-771

21 Reimers M & Carey V J (2006) *Methods Enzymol* 411, 119-134

22 Wettenhall J M, Simpson K M, Satterley K & Smyth G K (2006) *Bioinformatics* 22, 897-899

23 Ashburner M, Ball C A, Blake J A, Botstein D, Butler H, Cherry J M, Davis A P, Dolinski K, Dwight S S, Eppig J T, Harris M A, Hill D P, Issel-Tarver L, Kasarskis A, Lewis S, Matese J C, Richardson J E, Ringwald M, Rubin G M & Sherlock G (2000) *Nat Genet* 25, 25-29

24 Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) *Nucleic Acids Res* 36 (Database Issue), D480-484

25 Al-Shahrour F, Carbonell J, Minguez P, Goetz S, Conesa A, Tárraga J, Medina I, Alloza E, Montaner D & Dopazo J (2008) *Nucleic Acids Res* 36 (Web Server issue), W341-346

26 Huang da W, Sherman B T & Lempicki R A (2009) *Nat Protoc* 4, 44-57

27 Rebhan M, Chalifa-Caspi V, Prilusky J & Lancet D (1998) *Bioinformatics* 14, 656-664