Pharmacogenetics—Statistical Considerations

Aiden Flynn, Craig Ledgerwood and Caroline O'Hare

Abstract The growth of Pharmacogenetics (PGx), using biomarkers to diagnose, prognose and identify patient subgroups most responsive to clinical intervention, heralds the possibility of more effectively targeted therapies and personalised medicine. Whilst demonstrating clinical significance in a number of studies, greater use of PGx has been limited by the need for further technological/methodological advancement together with a more integrated approach in study design and data analysis at the outset of clinical studies. Consideration of the statistical factors to be examined over the course of biomarker studies at the planning stage, instead of the current trend for retrospective analysis, will ensure that studies will be suitably powered to address specific questions and that subsequent data analysis will account appropriately for sources of variability. This will improve confidence levels in the conclusions drawn and the overall utility of PGx research. Greater use of PGx in the development of personalised medicine will require more guidance by statisticians and quantitative biologists in the handling and extraction of information derived from the data produced from large studies within the multidisciplinary network of researchers involved. This chapter highlights the key limiting statistical factors to be considered when embarking upon investigations using PGx, affecting the quality of information obtained from clinical data generated in personalised medicine research.

Keywords Pharmacogenetics (PGx) · Biomarkers · Data analysis · Statistics · Study design optimisation · Simulation · Modelling · Personalised medicine · Bioinformatics

A. Flynn (✉) · C. Ledgerwood · C. O'Hare
Exploristics Ltd, 55-59 Adelaide Street, BT2 8FE Belfast, UK
e-mail: aiden.flynn@exploristics.com

C. Ledgerwood
e-mail: craig.ledgerwood@exploristics.com

C. O'Hare
e-mail: caroline.ohare@exploristics.com

G. Grech, I. Grossman (eds.), *Preventive and Predictive Genetics: Towards Personalised Medicine,* Advances in Predictive, Preventive and Personalised Medicine 9,
DOI 10.1007/978-3-319-15344-5_4

1 Introduction

The development of Pharmacogenetics (PGx) using biological markers (biomarkers) to identify patient groups responsive to treatment during clinical trials promises a new era in personalised medicine. Its application within recent clinical development programmes has grown considerably as both healthcare providers and drug developers have recognised its importance in directing treatments to those most likely to benefit. Where PGx has been implemented it can be used to guide decision making in clinical studies. It offers additional options over the course of drug development by helping explain unexpected variability in safety and efficacy outcomes in clinical interventions. Previously, such variability would have resulted in the termination of costly research programmes. However, as PGx can identify patient subgroups which are most responsive to treatment it can be used to focus further studies within these subsets.

Despite the potential value of PGx in improving the benefits and reducing the risks of some drugs in development by targeting treatments more effectively, so far its successes within research have been limited. This may be due in part to its predominant use, at present, as a tool to re-evaluate development plans when study outcomes are negative or ambiguous rather than being integrated at the outset in a personalised medicine approach. However, although PGx is still an evolving strategy requiring further technological and methodological development to optimise its use in data analysis and study design, its uptake in research programmes at the prospective planning stage, aiding study design and data analysis, is likely to improve its utility. To this end, this chapter identifies and quantifies the key limiting statistical factors commonly encountered when using PGx within a research study, which affect the quality of the information derived from personalised medicine research.

2 Types of Biomarkers

The aim of PGx analysis is to identify and characterise clinical responses occurring in patients subject to a given clinical intervention. These effects can be traced through data sets acquired from a variety of biomarkers. The use of biomarkers to track disease and its treatment offers the future possibility of individualised therapies providing personalised medicine for each patient. The biomarkers observed are biological characteristics that may be detected and measured objectively and used as an indicator of normal biological, pathogenic or pharmacologic processes in response to therapeutic intervention. Identification of individual biomarkers in the form of chemical, physical or biological parameters can be used either to measure progress of a disease or the efficacy of its treatment. As a result, biomarkers may be used to diagnose or predict treatment or disease outcome.

There are different types of biomarkers, with each type requiring the application of distinct statistical methods depending on their relationship to the observed

Table 1 Examples of predictive and prognostic biomarkers in current use

Biomarker	Type	Associated biological process/function	Indication
EGFR (ErbB-1)	Predictive	Signal transduction, cell proliferation, regulation of DNA replication/repair, stress response, cell adhesion, cell migration	Advanced non-small cell lung cancer, anal cancer glioblastoma multiforme
HER2/neu (ErbB-2)	Predictive	Transcriptional regulation, signal transduction, cell proliferation	Breast cancer
BluePrint ®	Predictive	80 gene panel for assessing molecular subtype of breast cancer	Breast cancer
MammaPrint®	Predictive	70 gene panel to categorise lymph node negative breast cancer	Breast cancer
OncoTypDX®	Predictive/prognostic	21 gene panel for assessing response to chemotherapy of estrogen receptor (ER) positive tumours	Breast/colon cancer
HLA-B*5701	Predictive	Immune regulation	Hypersensitivity reaction to Abacavir
K-RAS	Predictive/prognostic	Ras protein signal transduction, cell proliferation, gene expression regulation	Colorectal cancer
AB1-42	Prognostic	Protein component isoform of amyloid deposits associated with Alzheimer's Disease (AD)	Alzheimer's disease

treatment response. Therefore, the objective of a biomarker's use and its characteristics should be clear at the outset of analysis to ensure that the correct statistical approach is applied to the data. Some examples of biomarker types and their uses are given in Table 1. In this chapter, two types of biomarkers are considered, prognostic and predictive markers. For statistical purposes there is an important difference between these two marker types. Prognostic biomarkers, such as AB1-42, are linked to the prognosis or likely disease outcome in a defined patient group independent to the treatment given. As a consequence, they are usually identified with models where the biomarker is fixed as the main effect. In contrast, predictive biomarkers, including HER2, are able to help identify patients likely to respond to a given treatment but not to a comparator where response may be measured as efficacy or safety. Their identification requires the application of a statistical model which allows interaction between biomarker and treatment. In some instances, however, biomarkers may be both prognostic and predictive. An example of this is mutant K-ras which expressed in non-small cell lung tumours and can be used to predict responsiveness to EGFR Tyrosine Kinase Inhibitors.

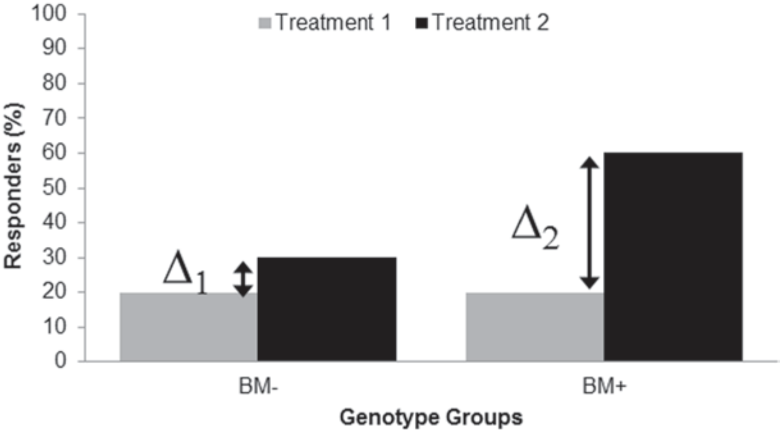

Fig. 1 An example of a predictive biomarker that is able to distinguish between groups of patients. In this case, the difference in the response rate in the BM+ group between treatment 2 and treatment 1 is greater than the equivalent difference in the BM− group. In other words, Δ_2 is greater than Δ_1

Identification of prognostic and predictive marker types is proving extremely useful in the development of personalised medicine. Prognostic markers can be used to segment populations by setting inclusion criteria at the start of a clinical study. This results in a reduction in the overall variability in the measure of response. In contrast, predictive markers are used to target treatments to patients more likely to derive benefit and are frequently further investigated as diagnostics for identifying responsive patient groups. An example of the use of a predictive biomarker to distinguish patient subgroups is shown in Fig. 1.

2.1 Biomarker Platforms

There are many different methods or platforms employed to measure biomarkers. These include a wide variety of technologies that can be used to produce biomarker data ranging from imaging modalities to the measurement of molecular biomarkers indicating gene expression, RNA expression, protein concentrations, single nucleotide polymorphisms (SNPs) or metabolites. Biomarker data can also take the form of continuous measurements, categories and ordinal scores. Different platforms may measure single markers or many thousands of markers simultaneously, generating data with specific attributes which will need to be accounted for in any statistical analysis. Data generated from some platforms may also require pre-processing steps such as scaling or normalisation [1] which must be considered prior to analysis. Consequently, it is important that the properties of the data obtained from each type of platform are factored into any statistical analysis.

2.2 Variability and Data Quality

Despite the accuracy of many of the biomarker platforms used, biomarker data can be prone to variability and bias. This can result in any subsequent analysis being subject to greater levels of statistical uncertainty leading to increased study failure rates. There are many factors which cause the observed variability and bias. These include the handling methods used for the tissue sample, when the sample was taken, patient factors such as drop-out rates, as well as inter-laboratory variation. If such factors are not addressed, inconsistent results are produced for the same biomarker across different biomarker studies [2, 3]. Therefore, it is critical that possible sources of variability should be evaluated during the development of a biomarker and suitable strategies for handling these sources and minimising their effect implemented.

2.3 Sources of Missing Data

Biomarker data often has a higher proportion of missing values than clinical data. These can arise as a result of numerous factors such as low consent rates for optional samples, patient drop-outs due to non-response or toxicity and measurements below the limits of detection of the biomarker assay. One key problem with these missing data is that the data are not missing at random. Indeed, the patients with missing values can be more likely to differ in their response to treatment compared to those with non-missing values. Therefore, it is important that missing biomarker data is not ignored as they are often informative. During any analysis involving missing biomarker data, it is important to compare key variables (e.g. those likely to impact response) between patients with and without biomarker data in order to understand differences between the missing group and the remainder of the study population. In addition, it is useful to understand reasons for missing data and take appropriate action. Where the pattern of missing data is understood, the implementation of models or imputation methods can help to recover the true underlying population statistics in the presence of missing data. On the latter point, information relating to the reason for missing data (e.g. below the limit of quantification) is often not recorded within the data set. This illustrates the need to improve on data standards and management practices relating to biomarker data.

2.4 Dimensionality and False Positives

Biomarker studies often involve the evaluation of numerous biomarkers in order to generate new hypotheses relating to the association between biomarker and response to treatment. This type of repetitive analysis results in a high number of false positive associations if the appropriate methods for controlling for the false positive rate are not used. Methods for adjusting for multiple testing have been reviewed elsewhere [4]. However, it should be noted the strategy for controlling false positives should be consistent with the aims of the experiment and the proposed use of the results. In hypothesis generating studies, it does not make sense to apply an overly conservative strategy that limits the likelihood of identifying plausible markers. Furthermore, exploratory studies do not end when a statistically significant p-value is generated. Indeed, there are often further steps in the evaluation process that will remove further spurious associations leaving those markers which are biologically plausible and have a clinically meaningful application.

3 Study Design Options

Good study design improves significantly the probability of meeting research objectives whilst minimising known sources of variability and bias. In PGx, the study design options depend on how and when PGx is being applied. At present, the early stages of PGx research is usually exploratory whereby many biomarkers are investigated, often using data collected as part of a study designed for another purpose. This is usually followed by confirmatory research where PGx becomes the primary objective in a prospectively designed study. To date, most methodological research into study designs for PGx has focused on the prospective, confirmatory applications.

3.1 Confirmatory Studies

Confirmatory studies are designed primarily to test a hypothesis based on observable pre-specified biological effects. Such studies are designed prospectively and measure markers of relatively known function which have previously been shown to explain variability in patient response. Several study designs that use prognostic and predictive markers to stratify the study population have been suggested and evaluated [5–8]. Three common designs used in confirmatory studies for predictive markers are an enriched design, a stratification design and an adaptive design, as shown in Fig. 2. The merits of each of these are discussed in the following sections.

In the targeted or enriched design (Fig. 2a) patients are selected for the study based on their biomarker status in a pre-screening step. This allows patients with the negative status to be excluded from the study. Positive status patients are then randomised to one of the treatment groups. The main advantage of this design is that a treatment effect can be observed within smaller studies. The disadvantage of this approach is that it does not provide information on the effect of treatment in the excluded population. As a result, it can only be used when there is already prior knowledge of the impact of a single biomarker.

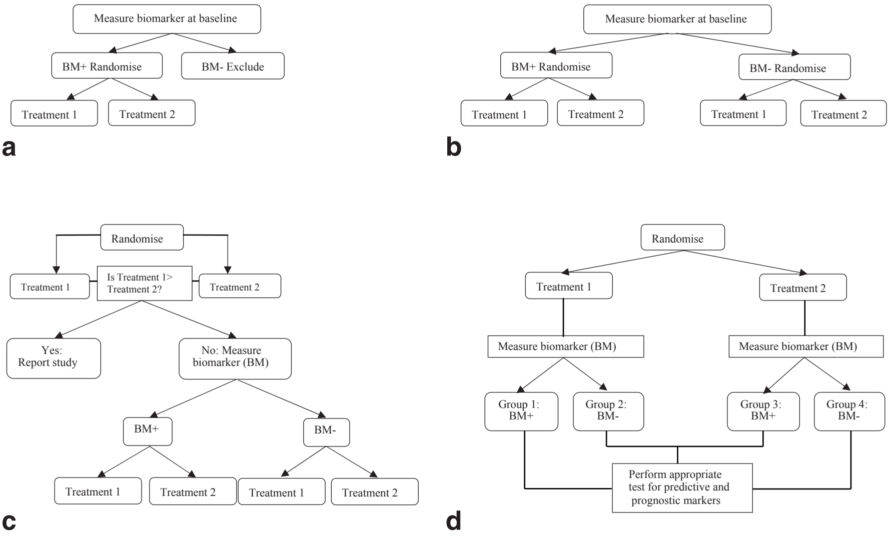

Fig. 2 Examples of study designs for Pharmacogenetics: Enriched design (**a**), Stratification design (**b**), Adaptive design (**c**), Retrospective design (**d**)

In contrast, the stratification design is less restrictive in its remit (Fig. 2b). It also has a pre-screening step whereby all the study subjects are stratified according to biomarker status and then randomised to treatment. The advantage of this design is that information can be collected on the treatment effect in the negative biomarker status group. Moreover, the performance characteristics of a diagnostic test, for example its sensitivity and specificity, can be estimated. However, as with enriched design, considerable prior knowledge about the biomarker is also required.

Both the enriched and the stratification designs are useful when studies are designed to test a single hypothesis relating to a given biomarker. However, more often studies have multiple objectives and involve evaluating a treatment effect in the entire study population as well as within sub-populations. In this instance an adaptive design is useful (Fig. 2c) [9]. With this design, patients are randomised to treatment groups and the treatments are compared. If there is no difference in the treatments, patients are stratified by biomarker status and a comparison of treatments is performed within these strata. This approach leads to a higher false-positive error rate, as multiple statistical tests are performed. Controlling for false-positives will result in larger studies. However, this design is more flexible than the targeted or stratified design as it allows the testing of multiple objectives and can be modified to include the evaluation of multiple biomarkers.

3.2 *Exploratory Studies*

PGx is used in exploratory studies for identifying useful biomarkers and to generate hypotheses for testing in further studies. Exploratory studies can range from evaluating small groups of candidate markers to large scale biomarker arrays, depending on prior knowledge of their biological function and relationships to treatment response.

Currently, clinical studies often collect blood samples with a view to using them for exploratory PGx research [10]. If a PGx study is initiated, the study population is stratified retrospectively and iteratively using the biomarkers under investigation (Fig. 2d). The advantage of this approach is its flexibility, as it does not compete with the original study objective and many biomarkers can be evaluated retrospectively. However, it has a limited ability to detect biomarker effects due to the restricted sample size and the need to control for the high false-positive rate. Moreover, bias and imbalance are introduced into the strata as patients within them have not been randomised to treatment [11]. Consequently, biomarkers identified using retrospective analysis may require further support from data derived in prospectively designed studies [12].

As a result of the limitations of the retrospective approach in exploratory studies, PGx has not provided the breakthroughs anticipated. Notable exceptions to this have been studies with drugs associated with large genetic effects, such as Abacavir [13] and Panitumumab [14]. This has highlighted the key challenge in PGx research. Where insufficient patient data has been available, it has been hard to detect more moderately sized effects and thus identify biomarkers with clinical utility. Nevertheless, retrospective PGx approaches will continue to play an important exploratory role. However, to improve the likelihood of successful exploratory PGx studies, a more integrated approach is required in research programmes at the outset of clinical study design and data analysis. Indeed, recent research using computer simulation to design studies addressing multiple objectives, including PGx investigations [15], showed that prospective planning is vital. This is particularly important when studies are designed for another purpose, so that useful PGx data can be generated without impacting the primary objectives of the study.

3.3 Data Analysis Methods

The main objective for PGx analysis is to identify and/or characterise genetic effects. Whilst there are too many methods to review adequately in this article, there are some general principles that are broadly followed in basic analyses, as shown in Fig. 3. Current approaches to biomarker or feature discovery involve a multistep process whereby biomarkers are selected for further investigation based on the strength of association with an outcome; typically by setting an arbitrary limit on the likelihood of detecting false positives (e.g. p value <0.05). Evaluation of biomarkers involves the application of a statistical model comprising the factors that are thought to contribute to the observed variability in response. These models can include two types of effects: main effects where factors make a sole contribution to

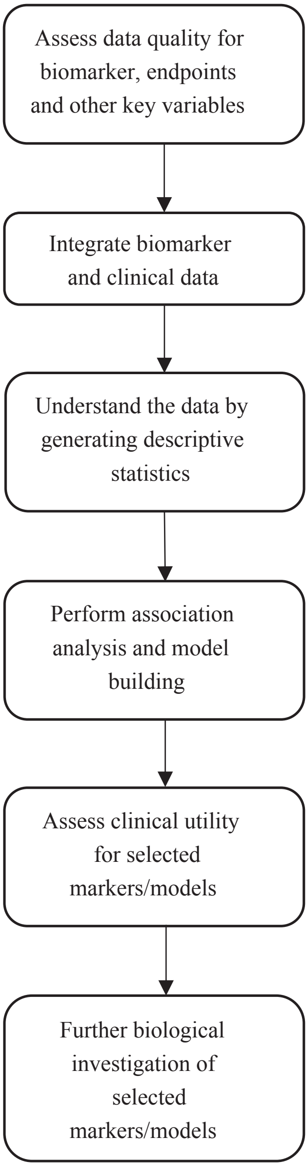
Assess data quality for biomarker, endpoints and other key variables
Integrate biomarker and clinical data
Understand the data by generating descriptive statistics
Perform association analysis and model building
Assess clinical utility for selected markers/models
Further biological investigation of selected markers/models

Fig. 3 An example of a statistical analysis strategy for a personalised medicine study. The analysis typically involves multiple steps that integrate different sources of information

the observed variability; and interaction effects where two or more factors make a combined contribution. This model can be written in the form of

$$R = B + T + B \times T$$

where R is the response to treatment, B is the biomarker measurement and T is the treatment group.

The choice of model is important as it will determine the utility and application of selected markers. Models with genotype as a main effect are useful for identifying markers that are associated with response, regardless of treatment (prognostic markers) whilst markers that are associated with response in the presence of treatment (predictive markers) can be identified using models with an interaction between genotype and treatment. Genetic markers are then selected on the basis of their prognostic or predictive utility.

3.4 Alternative Analysis Methods

The analysis methods described above relate in general to linear regression models. However, these can be limited in terms of their ability to incorporate complex relationships between different predictive and prognostic factors. In addition, linear modelling tends to require the pre-specification of the structure of the model. Clearly, this can lead to an over-simplification of the form of the relationships amongst predictive factors and outcomes. There are numerous other approaches that do not make the same assumptions and are more flexible in terms of enabling complex relationships to be modelled. There are too many to cover in any detail but neural networks [16], support vector machines [17] and random forests [18, 19] are regularly used to develop predictive models with some success. All modelling approaches need to take account of the study design and the biomarker utility and type.

High dimensional biomarker data sets are often sparse, in the sense that the model fitting process may have a limited number of observations that can be used to estimate the model parameters. There are a few useful methods that can be used to handle low density data including exact methods, lasso, elastic nets and others [20].

4 Model Building and Validation

The development of predictive and prognostic models generally involves the evaluation of biomarkers in the context of many other factors, such as demographics, baseline measures and environmental factors. As a result, these models include a combination of many factors that are additive in terms of their association with outcome. The development of these models is a multi-step process comprising variable selection and model evaluation followed by model validation [20].

Approaches to variable selection and model evaluation are generally well established. Typically, variable selection and model evaluation is an iterative process whereby variables are added or removed from a model following an evaluation of the contributions of those variables to the performance of the model. Following the model building process, the final model is the one that is considered to be the best performer.

One major problem with using high-dimensional data to build a predictive model is over-fitting of the data. In this instance, many variables are shown to have strong relationships with the outcome as a result of random selection. Consequently, any model that is based on these random relationships will not generalise to unseen data or an independent data set, highlighting the importance of model validation.

There are numerous ways to perform model validation [20]. A common approach is to train the model on data from one study and then use an independent dataset to validate the model by assessing its performance in the second dataset. One problem with this method is the lack of availability of a relevant independent dataset. An alternative approach is to split the data from one study into a training and validation set. The robustness of the model may be evaluated using an iterative procedure for selecting the test and validation set. This, however, relies on the availability of enough observations (patients) as splitting the data will reduce the power to identify useful markers. Where sample size is limited, another useful strategy is to use leave-one out cross-validation. In this case, the model is trained on all but one observation and the ability of the model to predict the outstanding observation is evaluated. This evaluation is performed repeatedly by randomly selecting the observation that is left out of the model building step.

Recent work has shown that the best approach to model building is to integrate the variable selection and the validation steps into one large iterative process [21]. The benefit of this method is that the performance of many models can be assessed at once whilst controlling for false positives. For all the cross-validation approaches described above, the model building and performance characterisation is performed in the same dataset using data that were collated under study-specific conditions. The most robust form of validation involves the use of completely independent data (external validation) to assess the performance of a model.

5 Diagnostic Development

The use of statistics and modelling is vital in demonstrating the utility of companion diagnostics, prior to regulatory approval. The sensitivity and specificity of a diagnostic in its target population, as well as its positive and negative predictive value need to be identified under the original conditions in which it has been evaluated and developed. There are also a range of criteria that need to be set [22], such as defining the optimal threshold for biomarkers on a continuous scale and evaluating the repeatability and reproducibility of the biomarker assay. In addition, the diagnostic development process can be validated by understanding and quantifying the factors that may impact its performance. When a diagnostic is being co-developed with a drug for regulatory approval, good coordination between these processes is critical. Diagnostics development may often fall behind that of its associated drug, due to identification of biomarkers over the course of a research programme. This can cause delays in drug approval unless both development programmes are well synchronised.

6 Visualisation and Presentation

A key component of any analysis in a personalised medicine study is a clear and simple visualisation of the results. The use of good graphical outputs able to display relevant information simply, help to place the results in a suitable context facilitating the interpretation of large data sets. Well-designed graphical displays can integrate information on the clinical utility of biomarkers along with biological information, such as the functional annotation of the gene region, by overlaying both sources of information on the same plot. Another important aspect of visualisation is the presentation of high dimensional data. In this case, the use of multi-panel plots, heat maps, contour and surface plots are extremely useful. In addition, it is common to reduce the dimensions of data using methods such as multi-dimensional scaling, principal components analysis and clustering. This enables the data to be displayed on standard plots in two or three dimensions and can also uncover hidden structures in the data.

The presentation of simple summary statistics can often mask effects and responses that are notably different to those of the broader population. Therefore, it is important to be able to distinguish those observations that differ in order to understand variability in the data and identify patients that derive benefit. Consequently, any analysis of personalised medicine research should include graphical displays that enable the visualisation of individual data points. It also presents an opportunity for the observations obtained from a biomarker study to be visualised alongside information derived from other sources, placing it within a wider biological context.

7 Bioinformatics and Biological Interpretation

Since the completion of the Human Genome Project (HGP) and the arrival of next generation sequencing (NGS), technological advances in genomic sequencing have increased the speed at which entire genomes can now be sequenced. In addition, the use of microarray gene chip technology to screen patient tissue samples for the presence of genetic biomarkers associated with some disease processes has become increasingly commonplace. These advances in the area of medical genetics have resulted in the generation of unprecedented volumes of raw biological data. The need to analyse this data in order to understand it and how it might be used for clinical applications has required the capabilities provided by the expanding field of bioinformatics. Bioinformatics combines the mathematics, computer sciences and statistics required for the collection, banking, deciphering, analysing and modelling that is necessary to analyse large amounts of biological information. Indeed, bioinformaticians continue to seek to address the pressing need for data analysis through the development of analytical tools that can be utilised on desktop systems to analyse and interpret the data collected.

The recent era of next generation sequencing and multiplex microarray platforms has allowed a vast expansion in the number of sequences able to be analysed in each experiment. Prior to the emergence of these technologies, the focus of molecular biology was on known sequences previously identified and attributed to a given protein and/or function. Complementary probes were used to identify the presence and abundance of those target sequences and determine differences between treated/non-treated or resistant/responsive groups. Performed initially in singleplex assays (PCR), this quickly progressed to multiplex microarrays which could simultaneously measure thousands of targets (genes, single nucleotide polymorphisms (SNP) or messenger RNA transcripts (mRNA)) thanks to the technologies developed by Affymetrix, Agilent and Illumina. However, whilst this has increased the number of sequences that can be analysed it has also raised problems in their analysis due to the high number of dimensions in the data produced and the relatively low number of observations in studies.

Bioinformaticians have played a key role in implementing these technologies and addressing the difficulty in dealing with high dimensional data. The pre-processing of data has become critical to the utility of high-throughput systems, with several normalisation techniques, such as Robust Multi-array Average (RMA) and the current Affymetrix algorithm MAS5, being developed and used routinely in both the proprietary software provided by the instrument manufacturers and in open source packages, such as that available on the Bioconductor software repository (http://www.bioconductor.org/). Following normalisation, the next problem is dealing with the high dimensional data and correcting for the false discovery rate in hypothesis testing. Both commercial and open source packages use standard statistical methods to make comparisons between groups. P-values are typically adjusted to correct for the number of tests being performed through methods such as Benjamini–Hochberg [23].

Whilst many of the current tools for analysing and interpreting microarray and next generation sequencing data are useful, the huge quantities of data they produce nevertheless continue to create new challenges in data analysis. The technical process of sequencing an entire genome may have become routine, however analysis of the data it generates remains problematic as it is very computationally intensive with over 3 billion base pairs and 50 million variations to consider. Although sequencing and microarray platforms have been around for some time, the ability to process the volume of data produced in a routine setting at an affordable cost has only become a relatively recent possibility due to the ability to store and process the terabytes of data produced.

Currently, a number of software tools exist which facilitate the process of interpreting sequencing and microarray data. Alignment tools, such as BLAST, have been used for years to identify proteins or genes from short amino acid or nucleotide sequences or to compare the similarity between two or more sequences. The concept of the algorithms used by FASTA/BLAST and software for NGS sequence alignment are similar, however, the alignment of hundreds of millions of short sequences (FASTQ) from the entire genome takes a lot longer to perform even with accelerated algorithms. The Bowtie sequence alignment tool is one of the most widely used aligners, due to its speed and the fact it is freely available to use. Even this can take several hours per sample to process and many researchers choose to run these on cloud-based platforms such as Amazon's Web Services. Other tools for analysing these alignments to identify variations, splice variants and differential expression of genes and isoforms are also freely available, such as the web-based application, Galaxy. However, these are still being actively developed and there are no clearly defined procedures as yet for the analysis and interpretation of alignments. Indeed, there are a plethora of tools both commercial and open-source for visualising, analysing and interpreting the large volume of information produced by each experiment.

Although the development of sequencing, storage and processing techniques has evolved concurrently, the ability to interpret the biological relevance of the information generated is still lagging. Some understanding of the underlying biological processes may be obtained through the mining of large gene sequence databases. These repositories of information encompass the knowledge gained to date regarding the biological relevance of genes and variations in sequences. Further biological context for biomarker studies might also be obtained through the use of data banked in public databases, such as ArrayExpress a functional genomics database containing data from both microarray and high-throughput sequencing studies. These databases now play a fundamental role in biological research and development, acting as a warehouse for storing, organising and providing large data sets relating to the occurrence and consequences of many biological processes, including gene variation, drug transport, drug targets, and other proteins of importance for drug response or toxicity [24–27]. Amongst the large number of databases generally available, some have become important bioinformatics tools within pharmacogenetics, such as the Human Genome Project (HGP) [28], Ensembl [29], the SNP databases dbSNP and JSNP [30], and HapMap [31]. These databases are rapidly expanding as they are continually updated with new submissions of genetic information, especially regarding the variation across the population.

The HGP demonstrated that the 20,000-plus genes expressed in humans only accounts for 1.5 % of the genome, with very little known about the function of the remaining 98.5 %. This is now being addressed through global collaborative projects such as ENCODE, which aims to completely annotate the non-protein encoding regions of the genome. Moreover, large sequencing projects like HapMap phase 3 and the 1000 Genomes Project, aim to give a clearer picture regarding the intrinsic genetic variation present within the human population. These projects in particular will provide a valuable resource for bioinformaticians, giving an important insight into the range of variation inherent in the human genome in general across multiple ethnicities. This will no doubt raise more considerations for analysis and interpretation of genomic data.

Possibly the most influential sequence database to date has been GenBank, an open access database storing known gene sequences from over 100,000 distinct species, along with their protein translations. It is run and maintained by the National Center for Biotechnology Information (NCBI) which plays an active and collaborative role in the development of computational biology. Their bioinformatics resources can be used to annotate and analyse an abundance of disparate data. Access to this important and expanding resource allows researchers to derive possible connections between the different aspects of biomarker data and thereby shape a more biologically meaningful view of it [32, 33].

The Ensembl database also provides genomic information with a rich source of gene variant data from humans and other species, including single nucleotide polymorphisms (SNPs). Alongside the HGP and Ensembl, dbSNP, Japan's JSNP and HapMap are another three of the more widely accessible and utilised bioinformatics resources. The Single Nucleotide Polymorphism Database, (dbSNP), created to supplement GenBank, is a public access archive for genetic variation within and between organisms developed by the NCBI. It comprises information on over 64 million distinct SNP variants in 55 species, including *Homo sapiens* [34]. Meanwhile, the HapMap project provides an alternative platform of information designed to enable researchers to carry out large scale studies to link genetic variants to the risk of specific diseases [31].

Other useful databases storing genetic data are the Gene Expression Omnibus (GEO), the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [35] and PharmGKB [36]. The analysis of the data contained in these databases is now an integral element of the bioinformatics process. Many of the data storage facilities have analytical applications bolted on as add-ons, whilst others are standalone warehouses that store the information that is used in other analytical applications. As with genetic databases, there are many bioinformatics tools available for dissecting, analysing and visualising the data, a selection of these have been listed in Table 2. This is also an evolving field within bioinformatics with new software tools under development.

Another recent tool useful in understanding the biological complexity of differentially expressed genes and proteins and identifying statistically significant sets of genes, is Gene Set Enrichment Analysis (GSEA) [37]. This is a powerful analytical method for interpreting gene expression data, deriving its power by focusing on groups of genes or gene sets that share common biological function, chromosomal location or regulation. Whilst single-gene analysis is useful it can miss important effects on biological pathways which are distributed across large networks of genes and are hard to detect at individual gene level. In contrast, GSEA which examines sets of related genes can identify many common biological pathways as cellular processes often affect networks of interacting genes. The advantage of GSEA is that it facilitates interpretation of genome-wide expression data as it focuses on gene sets which give more reproducible and therefore interpretable data.

Undoubtedly, constructing biological meaning from lists of statistically significant markers remains a challenge to bioinformaticians. However the development of complex data warehouses and other resources, that detail the structures and processes and interactions where individual genes/proteins exist, have helped to overcome such difficulties through the grouping of long lists into smaller sets of related genes or proteins that share a similar physiological/pathological function, cellular localisation, position in the genome or can be defined by similar gene ontology terms. There is a plethora of gene set databases that use examples found in the

Table 2 A non-exhaustive list of bioinformatics tools available with a brief description of the tool and an external link

Tool	Description	Link
ArrayExpress	Functional genomics database containing data from microarray and high-throughput sequencing studies	http://www.ebi.ac.uk/arrayexpress/
BioMart	Search engine allowing generation of tables of terms linked to genes and SNPs	http://www.ensembl.org/biomart
BLAST	Database allowing searching and alignment of sequences to the RefSeq genome	http://www.blast.ncbi.nlm.nih.gov/Blast.cgi
GeneMania	Search engine to find related genes through linkage of association information	http://www.genemania.org/
GSEA	Tool to perform enrichment analysis on gene sets provided in MSigDB database	http://www.broadinstitute.org.gsea/index.jsp
Haploview	Range of tools for analysing linkage disequilibrium and haplotype patterns	http://www.broad.mit.edu/mpg/haploview/
KEGG mapper	Collection for mapping gene sets to the KEGG pathways	http://www.kegg.jp/kegg/kegg1b.html
KEGG pathway	R package for analysis and visualisation of expression data within KEGG pathway	http://www.bioconductor.org/packages/release/bioc/html/KEGGprofile.html
PathNet	Tool that performs pathway analysis using topological information from pathways	http://www.bioconductor.org/packages/release/bioc/html/PathNet.html
Pathway browser	Tool for visualising pathways	http://www.reactome.org/PathwayBrowser/
SNAP	SNP annotation/proxy search tool using linkage disequilibrium and physical distance	http://www.broadinstitute.org/mpg/snap/
Stitch	Tool for exploration of known/predicted molecular interactions	http://www.stitch.embl.de/
Sweep	Tool for large scale haplotype analysis	http://www.broadinstitute.org/mpg/sweep/index.html
topGO	Compares GO term representation in an expression set accounting for GO topology	http://www.bioconductor.org/packages/2.12/bioc/html/topGO.html
UCSC genome browser	Tool providing interactive graphical interface to visualise genome annotation and chromosomal position	http://www.genome.ucsc.edu/

literature and gene sets that have been computationally derived. Databases such as, MSigDB and ConsensusPathDB, have brought together a large collection of gene sets comprising gene regulation, protein interactions, genetic interactions, biochemical reactions, drug-target interactions, pathways, gene ontology, disease regulation and many more. As differences in database structures and terminology exist, moving towards large warehouses of this information and development of advanced tools for mining the information are crucial to their implementation. Tools have been developed as stand-alone applications or as add-ons in other packages such as R (PGSEA). The principles of these tools are largely similar, in that they comprise an annotation database, a process that can assign those annotations to a given gene list, a further process that performs a statistical test to identify annotations that are significantly represented in the gene list and a method to interpret this graphically.

Examples of statistical approaches used in gene enrichment analysis are over-representation analysis (ORA), functional class scoring and pathway topology. Using a statistical test, for example the hypergeometric or binomial, ORA evaluates whether a specified functionally defined group of genes/proteins is represented within a gene list, or if it occurs merely by chance [38]. One drawback, however, is that ORA treats each gene equally, losing any possible correlation and interaction between genes. Functional class scoring overcomes some of the limitations of ORA as it treats the genes differently depending on the strength of the individual raw microarray values [38]. The pathway topology approach has advantages over both of these methods, in that it does not only consider the number of genes in a pathway to identify significant pathways, but also utilises information about inter-pathway connectivity. GSEA utilises its own novel method for performing the analysis, this calculates an enrichment score using a weight Kolmogorov-Smirnov-like test. The enrichment score generally reflects the degree of over representation at the top or bottom of a ranked gene list [37].

Whilst it remains challenging, bioinformatics has made some progress in understanding and interpreting large data sets by exploiting alternative data resources. The effort to understand in more depth the implications of the data collected and analysed within an experiment has required input from other data sources enabling a fuller understanding of its meaning. It is becoming clear that results from primary biomarker analyses might need augmentation with additional information from other studies to provide a greater biological context for the role of the biomarker in question. Further biomarker context could be given by its characterisation as well as by pathway analysis. Additional data in this form would help authenticate the patient subgroups identified from biomarker analysis by providing further supporting evidence. Alternatively, longitudinal data would allow further characterisation of subgroups using variables that change over time. This would require the pre-selection of those variables, with the selection process used described.

8 Future Opportunities

There are a number of opportunities within statistics and modelling to improve the application and implementation of pharmacogenetics in future studies. Four key areas to be considered are improving study design, integration of analysis methods, use of disparate data sources to provide biological context and better multidisciplinary collaboration involving quantitative scientists. Addressing these factors will develop PGx by increasing the success rate of exploratory studies to identify new biomarkers. Additional use of computer simulation will enable the application of smarter clinical trials that optimise the likelihood of success of a study without prohibitively increasing its size.

At present integrative analysis methods, such as Bayesian methods, are based on the idea of obtaining a consensus by combining prior information and current opinion, thereby providing a statistical framework that enables the quantitative (probabilistic) integration of information across multiple analysis steps [39]. This approach can limit biomarker discovery as such studies may be underpowered to detect small to moderate (but biologically important) effects; it filters potentially useful information in variables that fail to reach significance and it ignores the additional control of false positives that naturally occurs in the subsequent analysis steps. To progress, there is a need to develop methods that do not filter out useful information and that enable the quantitative integration of information from additional analysis steps, such as clinical and biological pathway analysis and comparisons with literature. Furthermore, there is enormous scope for developing and applying statistical models that more closely reflect the underlying biology and patterns of response; using models that better describe the data will increase the power to detect genetic markers. The development of these capabilities will require extensive methodological research and development for the integration and application of disparate data sources.

It is clear that statisticians and quantitative biologists are of increasing importance in the multidisciplinary network of researchers involved in the development of personalised medicine. Bioinformatics has helped develop new algorithms and software to facilitate the analysis of complex data sets. The development of new computational data and analytical solutions are crucial to handling and extracting the information derived from large clinical studies, improving the understanding of disease progression and treatment. Nevertheless, a knowledge gap still exists between the exploratory world of bioinformatics and the rigour and regulation of clinical statistics. Closer collaboration between quantitative scientists will break down the barriers in communication that exist between the disciplines and will enable scientists to gain experience, knowledge and an appreciation of the skills and capabilities that exist in other fields. Deeper understanding of other capabilities and technologies will lead to new innovations that make use of the extensive information available and improve the application of Pharmacogenetics in the quest for personalised medicine.

References

1. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD Jr, Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Chou J, Da-

vison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR, Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano JG, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li M, Li QZ, Li S, Liu J, Liu Y, Liu Z, Meng L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page GP, Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan PY, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, von Frese J, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD, Consortium M (2010) The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28(8):827–838. doi:10.1038/nbt.1665

2. Mattsson N, Blennow K, Zetterberg H (2010) Inter-laboratory variation in cerebrospinal fluid biomarkers for Alzheimer's disease: united we stand, divided we fall. Clin Chem Lab Med 48(5):603–607. doi:10.1515/CCLM.2010.131
3. Fenech M, Bonassi S, Turner J, Lando C, Ceppi M, Chang WP, Holland N, Kirsch-Volders M, Zeiger E, Bigatti MP, Bolognesi C, Cao J, De Luca G, Di Giorgio M, Ferguson LR, Fucic A, Lima OG, Hadjidekova VV, Hrelia P, Jaworska A, Joksic G, Krishnaja AP, Lee TK, Martelli A, McKay MJ, Migliore L, Mirkova E, Muller WU, Odagiri Y, Orsiere T, Scarfi MR, Silva MJ, Sofuni T, Surralles J, Trenta G, Vorobtsova I, Vral A, Zijno A, project HUM (2003) Intra- and inter-laboratory variation in the scoring of micronuclei and nucleoplasmic bridges in binucleated human lymphocytes. Results of an international slide-scoring exercise by the HUMN project. Mutat Res 534(1–2):45–64. doi:10.1016/S1383-5718(02)00248-6
4. Hsu JC (2010) Multiplicity adjustment big and small in clinical studies. Clin Pharmacol Ther 88(2):251–254. doi:10.1038/clpt.2010.122
5. Bromley CM, Close S, Cohen N, Favis R, Fijal B, Gheyas F, Liu W, Lopez-Correa C, Prokop A, Singer JB, Snapir A, Tchelet A, Wang D, Goldstaub D, Industry Pharmacogenomics Working G (2009) Designing pharmacogenetic projects in industry: practical design perspectives from the Industry Pharmacogenomics Working Group. Pharmacogenomics J 9(1):14–22. doi:10.1038/tpj.2008.11
6. Mandrekar SJ, Sargent DJ (2009) Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat 19(3):530–542. doi:10.1080/10543400902802458
7. Simon R (2010) Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Per Med 7(1):33–47. doi:10.2217/pme.09.49
8. Freidlin B, McShane LM, Korn EL (2010) Randomized clinical trials with biomarkers: design issues. J Natl Cancer Inst 102(3):152–160. doi:10.1093/jnci/djp477
9. Jiang W, Freidlin B, Simon R (2007) Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Cancer Inst 99(13):1036–1043. doi:10.1093/jnci/djm022
10. Goodsaid FM, Amur S, Aubrecht J, Burczynski ME, Carl K, Catalano J, Charlab R, Close S, Cornu-Artis C, Essioux L, Fornace AJ Jr, Hinman L, Hong H, Hunt I, Jacobson-Kram D, Jawaid A, Laurie D, Lesko L, Li HH, Lindpaintner K, Mayne J, Morrow P, Papaluca-Amati M, Robison TW, Roth J, Schuppe-Koistinen I, Shi L, Spleiss O, Tong W, Truter SL, Vonderscher J, Westelinck A, Zhang L, Zineh I (2010) Voluntary exploratory data submissions to the US FDA and the EMA: experience and impact. Nat Rev Drug Discov 9(6):435–445. doi:10.1038/nrd3116

11. Wang SJ, O'Neill RT, Hung HJ (2010) Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. Clin Trials 7(5):525–536. doi:10.1177/1740774510375455
12. Burns DK, Hughes AR, Power A, Wang SJ, Patterson SD (2010) Designing pharmacogenomic studies to be fit for purpose. Pharmacogenomics 11(12):1657–1667. doi:10.2217/pgs.10.140
13. Hughes AR, Brothers CH, Mosteller M, Spreen WR, Burns DK (2009) Genetic association studies to detect adverse drug reactions: abacavir hypersensitivity as an example. Pharmacogenomics 10(2):225–233. doi:10.2217/14622416.10.2.225
14. Weber J, McCormack PL (2008) Panitumumab: in metastatic colorectal cancer with wildtype KRAS. BioDrugs 22(6):403–411. doi:10.2165/0063030-200822060-00006
15. Flynn AA (2011) Pharmacogenetics: practices and opportunities for study design and data analysis. Drug Discov Today 16(19–20):862–866. doi:10.1016/j.drudis.2011.08.008
16. Wei JS, Greer BT, Westermann F, Steinberg SM, Son CG, Chen QR, Whiteford CC, Bilke S, Krasnoselsky AL, Cenacchi N, Catchpoole D, Berthold F, Schwab M, Khan J (2004) Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. Cancer Res 64(19):6883–6891
17. Lee HS, Cho SB, Lee HE, Kim MA, Kim JH, Park do J, Yang HK, Lee BL, Kim WH (2007) Protein expression profiling and molecular classification of gastric cancer by the tissue array method. Clin Cancer Res 13(14):4154–4163
18. Patterson SD, Cohen N, Karnoub M, Truter SL, Emison E, Khambata-Ford S, Spear B, Ibia E, Sproule R, Barnes D, Bhathena A, Bristow MR, Russell C, Wang D, Warner A, Westelinck A, Brian W, Snapir A, Franc MA, Wong P, Shaw PM (2011) Prospective-retrospective biomarker analysis for regulatory consideration: white paper from the industry pharmacogenomics working group. Pharmacogenomics 12(7):939–951. doi:10.2217/pgs.11.52
19. Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Sole X, Crous-Bou M, Guino E, Sanjuan X, Biondo S, Soriano A, Jurman G, Capella G, Furlanello C, Moreno V (2012) Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. PLoS One 7(11):e48877. doi:10.1371/journal.pone.0048877
20. Taylor JM, Ankerst DP, Andridge RR (2008) Validation of biomarker-based risk prediction models. Clin Cancer Res 14(19):5977–5983. doi:10.1158/1078-0432.CCR-07-4534
21. Freidlin B, Jiang W, Simon R (2010) The cross-validated adaptive signature design. Clin Cancer Res 16(2):691–698. doi:10.1158/1078-0432.CCR-09-1357
22. FDA (2005) Drug diagnostics co-development concept paper. http://www.fda.gov/downloads/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/UCM116689.pdf 2013. Accessed 29 May 2014
23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1):289–300
24. Chen YP, Chen F (2008) Identifying targets for drug discovery using bioinformatics. Expert Opin Ther Targets 12(4):383–389. doi:10.1517/14728222.12.4.383
25. Ginsburg GS, Willard HF (2009) Genomic and personalized medicine: foundations and applications. Transl Res 154(6):277–287. doi:10.1016/j.trsl.2009.09.005
26. Roos DS (2001) Computational biology. Bioinformatics–trying to swim in a sea of data. Science 291(5507):1260–1261
27. Sim SC, Altman RB, Ingelman-Sundberg M (2011) Databases in the area of pharmacogenetics. Hum Mutat 32(5):526–531. doi:10.1002/humu.21454
28. ORNL The Human Genome Management Information System (HGMIS) (2014) www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml. Accessed 29 May 2014
29. Ensembl Ensembl Genome Browser (2014) http://hapmap.ncbi.nlm.nih.gov/. Accessed 29 May 2014

30. NCBI SNP—Short Genetic Variations (2014) http://www.ncbi.nlm.nih.gov/SNP. Accessed 29 May 2014
31. NCBI International HapMap Project (2014) http://hapmap.ncbi.nlm.nih.gov/. Accessed 29 May 2014
32. NCBI Human Genome Resources (2015) http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml. Accessed 23 Mar 2015
33. Roses AD (2000) Pharmacogenetics and the practice of medicine. Nature 405(6788):857–865. doi:10.1038/35015728
34. Johnson AD (2009) Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources. Circ Cardiovasc Genet 2(5):530–536. doi:10.1161/CIRCGENETICS.109.872010
35. KEGG KEGG: Kyoto Encyclopedia of Genes and Genomes (2014) www.genome.jp/kegg/. Accessed 29 May 2014
36. PharmGKB PharmaGKB (2014) http://www.pharmagkb.org. Accessed 29 May 2014
37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43):15545–15550. doi:10.1073/pnas.0506580102
38. Leong HS, Kipling D (2009) Text-based over-representation analysis of microarray gene lists with annotation bias. Nucleic Acids Res 37(11):e79. doi:10.1093/nar/gkp310
39. Heron EA, O'Dushlaine C, Segurado R, Gallagher L, Gill M (2011) Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. Biostatistics 12(3):445–461. doi:10.1093/biostatistics/kxq072