Practical uses of quality assessment for high-dimensional gene expression data

Julia Brettschneider

1 Introduction

As J.W. Tukey famously praised our profession, "the best thing about being a statistician is that you get to play in everyone's backyard." Indeed, starting a collaboration with a scientist has a lot in common with a playdate at a new friend's house. It leads the statistician into an unfamiliar world of knowledge with unusual types of challenges. Unfamiliar rules. New toys. The goals and priorities, however important, are vaguely defined, and have to be determined and realised in collaboration with the new friend. New games and tricks arise and spread from the neighbour's backyard to their neighbours' backyards, eventually becoming widely practiced in the community. As statisticians, our control over the development, the dissemination and the ownership of methods is limited. Just at the best part of the game, getting the data into shape and generalising the methods, lab superiors may announce that priorities have shifted and there is no time for fine tuning the quantitative methodology. Later, not all members of the community will remember in which backyard the original idea was established and some believe it happened in their own.

A typical situation to call in a statistician is to find answers in data with unfamiliar characteristics generated by new technologies. In the last two decades, novel high- throughput gene expression measurement technologies such as microarrays and RNA Sequencing have created a strong connection between functional genomics and statistics. When the first microarray platforms were introduced, the most intriguing fact about them was the sheer number of genes that could be assayed simultaneously, enabling biologists to adopt new strategies in their quest for understanding which genes play which roles in a given biological process. Instead of verifying the role of a specific gene, they could explore *which* genes are involved and how they interact. Biologists enthusiastically set up experiments studying ev-

Julia Brettschneider
University of Warwick, Coventry, CV4 7AL, UK, e-mail: julia.brettschneider@warwick.ac.uk

erything on the genomic level: comparing mutants and wild types, unrevealing cell division, circadian clock, embryonic development, ageing and many more. Biomedical research looked into the molecular aetiology of complex diseases such as cancer, Alzheimer's, schizophrenia and cardio vascular diseases.

In the early days of the new technology, lab practice was dominated by mantras like microarray pioneer D. Botstein's: "If I had to replicate my experiments, I could only do half as much." Major manufacturers of the new technology would back up this attitude by conveying the impression that the technology produces high quality data with occasional outliers, perhaps, but so obviously no effort for statistical quality monitoring was even needed. Sadly, many microarray based studies turned out to be inconclusive or irreproducible. Concerns grew, especially in view of clinical use in diagnosis for treatment individualisation. D. Allison [2] reviews the epistemological issues in microarray based research. *Nature* and related journals published a series of articles about reproducibility and editorial steps to ensure transparency and robustness in published work were taken `nature.com/nature/focus/reproducibility`. Bayer, one of the worlds largest chemical companies, halted nearly two-thirds of its target-validation projects because in-house experimental findings failed to match published literature claims [3].

The biggest impact statisticians had on the field was a change of attitude in the users of the new technologies. Better experimental design, improved data preprocessing, systematic data quality control and awareness of the pitfalls of multiple testing are becoming more frequent in the genomics community at all levels, including academic labs, research institutes and industry.

2 Genomics and massively parallel measurement technology

DNA, the chemical structure of genes, is the blueprint of a biological organism. It is passed on from mother cell to daughter cells by *replication*. But why are your brain cells different from your liver cells despite having the same DNA? The molecular explanation is that through biochemical processes the information encoded in DNA contained in each cell's nucleus is *transcribed* into RNA and further *translated* into proteins (Fig. 1), the main building block of biological organisms. The amount of RNA and protein produced by a gene is *variable.* Depending on circumstances such as organism, tissue type, time point, developmental stage, disease state and environmental conditions. The abundance of RNA produced by a gene is called its *expression* and can be measured through blotting technologies. Massive parallelisation of the measurement process came with the introduction of *microarrays*, glass surfaces with large numbers of distinct fragments of DNA called probes attached to it at fixed positions. A fluorescently labelled sample containing a mixture of unknown quantities of DNA molecules is applied to the microarray. Under the right chemical conditions, single-stranded fragments of target DNA will base pair with the probes which are their complements, with great specificity, a reaction is called *hybridisation*. The informal industrial standard for microarrays are *short oligonu-*

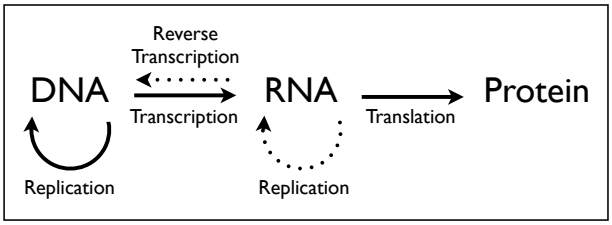

Fig. 1 Central Dogma of Molecular biology. Genetic information governs the organisms through biochemical processes including transcription, translation and replication (cell division).

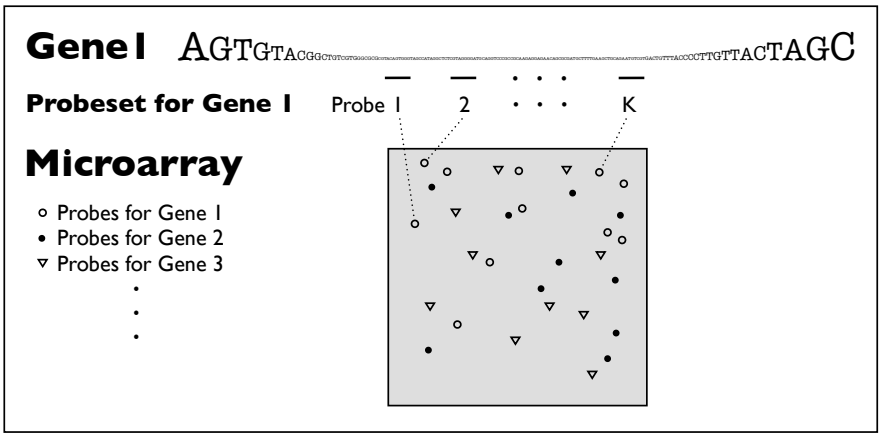

Fig. 2 Short oligonucleotide gene expression arrays. Each gene is represented by 11-20 probes scattered across the microarray. The probes are synthesised on the array and can provide expression measurements for tens of thousands of genes in one biochemical assay.

cleotide microarrays shown in Fig. 2. The intensities measured on the array will be statistically combined into an expression value estimate for the gene. Another decade later, the parallelisation of sequencing further progressed gene expression measurement. *RNA sequencing* technology is now a more precise (and more costly) alternative to microarray platforms. In terms of biomedical research, these high-throughput approaches have opened up entirely new avenues. Rather than experimentally confirming the hypothesised role of a certain candidate gene in a certain cellular process, they can use genome-wide comparisons to screen for all genes which might be involved in that process.

3 A quality assessment framework and toolbox

From a statistical point of view, high-throughput gene expression measurement technologies have created data with a particular profile of challenges: The measurement is a multi-step biochemical procedure with each step contributing to technical variation. There also is biological variation between RNA, which can be difficult to

distinguish from the variation between different species (or different parts of an organism, or different states). Huge numbers of measurements of molecular species are being taken in parallel, no gold-standards for a representative number of these species are available, their correlation structure is unknown and they are affected non-uniformly by the numerous sources of variation.

In a seminal paper, Brettschneider *et al.* [1] provide a conceptual framework for quality assessment (QA) for data obtained by these technologies and offer a toolbox with a number of concrete methods. The explicit QA goals are manifold, depending on resources, time and kind of user. Typical phenomena to look for are outliers, trends or patterns over time, effects of experimental conditions or sample characteristics, changes between batches, sample cohorts or lab sites, because all of these sources of variation may potentially interfere with the reproducibility of the study.

The QA toolbox relies on analysing the collective behaviour of the data after statistical preprocessing. It provides both numerical and spatial quality assessment. Some of the measures are tailored to short oligonucleotide microarrays, others can also be used for data from platforms or RNA sequencing. An extensive discussion of the application of the QA toolbox to experimental data sets can be found in [1]. Here we illustrate the main ideas using raw data from a fruit fly experiment by our collaborator T. Magelhaes (at the time at Corey Lab, UC Berkeley). The data set includes 89 short oligonucleotide arrays of 19 mutants and wild type with 4 to 5 replicates each and is available at the National Center for Biotechnology Informations Gene Expression Omnibus (GSE6515 at `ncbi.nlm.nih.gov/geo`).

Raw intensities. The most primitive assessment is to consider the distributions of the raw intensities. We do not consider this a full QA measure, but use them to study brightness, dimness or saturation, or in combination with more complex quality measures. For short oligonucleotide arrays raw intensities refer to the *PM values* (i.e. intensities obtained by perfect sequence matches on the array), while for printed microarrays spot intensities could be used.

Relative Log Expression (RLE). This assessment captures the amount of similarity between the overall distribution of the gene expression values of one sample and the corresponding distributions of other samples in the same data set. It can be computed from data obtained with any microarray platform as well as with RNA sequencing technology. First, the data of all samples from an experiment (or batch) is preprocessed by a suitable algorithm providing one expression value estimate for each gene in each sample. Then, a *median array*[1] is constructed by calculating, gene by gene, the median expression value over all samples measured in the experiment. Finally, again gene by gene, the RLE is defined as the difference of the gene's log expression in the sample in question to its log expression in the median array. The result is an *RLE distribution* for each sample. Their interpretation makes use of two assumptions that are justified in many experiments. Compared across biological conditions, about the same amount of genes are unregulated as down regulated, and

[1] This terminology stems from microarray technology. If another technology is used this collection of reference values can be computed the same way, though technically is not an array.

most genes are not differentially expressed. So, in a good quality array the median of the RLE (Med(RLE)) is close to 0 and its interquartile range (IQR(RLE)) is small.

The remaining assessment tools are specific to short oligonucleotide arrays and use probe-level quantities obtained as by-products of the robust multichip analysis (RMA) algorithm [4]. For a fixed probeset, RMA models the background corrected normalised intensity y_{ij} of probe j on array i as $\log y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}$, with α_j a *probe affinity effect*, μ_i the log scale expression level for array i, and ε_{ij} an i.i.d. centered error with standard deviation σ, with a zero-sum constraint on the α_j's. The model can be fitted robustly by iteratively weighted least squares delivering a probeset expression index $\hat{\mu}_i$ for each array i and residuals r_{ij} and weights w_{ij} attached to probe j. Discordant probe intensities get downweighted.

Normalized Unscaled Standard Error (NUSE): This assessment is calculated for each probeset resulting in the *NUSE distribution*. Let $\hat{\sigma}$ be the estimated residual standard deviation in the RMA model and $W_i = \sum_j w_{ij}$ the *total probe weight* (of the fixed probeset) in array i. Its expression value estimate is $\hat{\mu}_i = \sum_j y_{ij} \cdot w_{ij}/W_i$ with $SE(\hat{\mu}_i) = \hat{\sigma}\sqrt{\sum_j w_{ij}^2}\Big/W_i$. The residual standard deviations vary across the probesets within an array providing an assessment of overall goodness of fit, but no information on relative precision of estimated expressions across arrays, so we replace $\hat{\sigma}$ by 1. Other sources of heterogeneity are the probeset-dependent number of "effective" probes (in the sense of being given substantial weight by RMA) and dysfunctional probes (i.e. having high variabiliy, low affinity, or a tendency to cross hybridise). To compensate, we divide by its median over all arrays obtaining the *Normalised Unscaled Standard Error (NUSE):* $NUSE(\hat{\mu}_i) = \sqrt{\sum_j w_{ij}^2}/W_i \Big/ \text{Median}_t\left\{\sqrt{\sum_j w_{tj}^2}/W_t\right\}$. It can be thought of as the square root of the sum of the squares of the normalised *relative effectiveness of the probes* contributing to the probeset summary (see [1]). Deviations of Med(NUSE) from 1 or high IQR(NUSE) indicate low quality.

Quality landscapes: Shading the positions in a rectangular grid according to the magnitude of probe-level quantities (i.e. raw probe intensities, weights or residuals) creates images of the array (e.g. Fig. 3). The collective spatial behaviour of these quantities can reveal local damage caused e.g. by dust particles, handling, air bubbles, and spatial inhomogeneity due to insufficient mixing or drying out. The weights are in a sense the reciprocals of the absolute residuals. The two centre images show that the sign of the residuals can give additional insight.

Figure 4 shows the QA results for 14 arrays of the jointly analysed full set of arrays using a series of coloured boxplots. The most obvious fact in this selection of arrays is that array 1 of the second mutant (violet) has very low PM values. However, RLE and NUSE testify this array is of average quality. We later found out that this particular array was hybridised on a differently calibrated machine. What looked like an outlier according to PM was fixed turned into normal quality through preprocessing. This is confirmed by the weight landscapes in Fig. 3 showing only local defects near the edges for array 1, but overall low quality for array 2. In contrast, arrays 2 and 3 of the same series do not look suspicious in terms of PM, but RLE and NUSE rate them as lowest quality arrays of the data set of 89 arrays (not

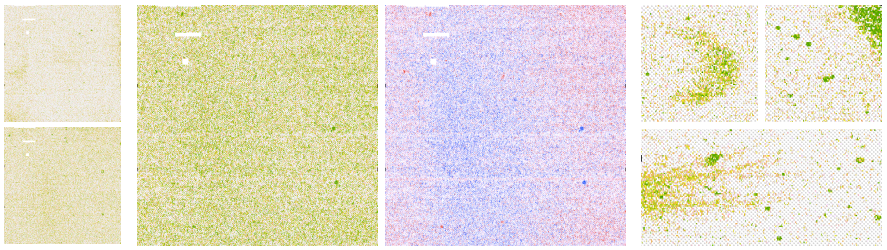

Fig. 3 Quality landscapes. The small weight landscapes on the left correspond to array 1 and array 2 of the second mutant (violet in Fig. 4) in the fruit fly dataset. The remaining images are from preliminary experiments conducted in the same lab. The weight and residual landscapes in the centre are from the same array. The distribution of the signs of the residuals (visualised as red versus blue) reveals spatial inhomogeneity. The magnified details (right) of weight landscapes show typical local defects.

all pictured). The other arrays are of comparable data quality, though noticeably the last replicate array in both the first (green) and the third (turquoise) mutant seem of lower quality than the others, again not obvious from just studying the PM boxplots.

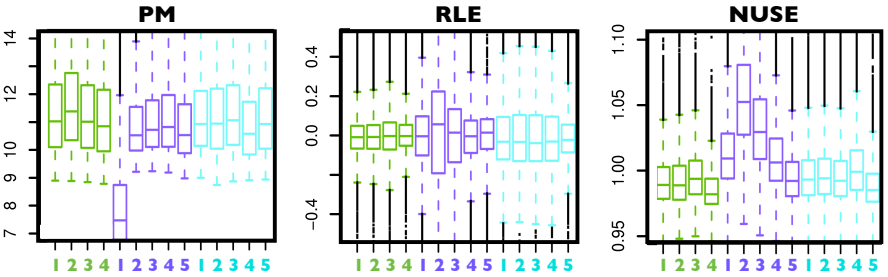

Fig. 4 Boxplots of quantities used in QA. Distribution of PM, RLE and NUSE of three different fruit fly mutants (coloured green, violet and turquoise) in 4 to 5 replicates each.

4 How impact happens

The methodology was disseminated through the journal article [1], preliminary works quoted there and conference presentations. Its use in practice was accelerated by several factors. Firstly, the methods were fleshed out by many explicit case studies of typical lab experiments. The journal publication itself discusses 5 data collections covering different designs, lab sizes and biological organisms, and related publications offer numerous other experimental data examples. In particularly, there is B. Bolstad's extraordinary initiative `plmimagegallery.bmbolstad.com`, a col-

lection of case studies featuring each array's numerical and spatial QA as well as discussions about specific technical causes of poor quality. Secondly, the methods were implemented in freely available software, mostly the open source R-packages *affyPLM* and *arrayQualityMetrics* from `bioconductor.org`, but also *Chipster, RobiNA* and other genomic data analysis software. Thirdly, the authors built strong links to users of the technology in academia, research institutes and industry through collaboration and advice and through their presence in online forums. We now discuss the roles the QA toolbox provided by [1] took on for different layers of the scientific community. Further details, references and more quotes can be found in [7].

Small labs: Academic labs and smaller research institutes run high-throughput gene expression microarray based studies of up to 100 arrays, sometimes even a few hundred arrays. Their main purpose of using the QA toolbox is the identification of outliers, of technical artefacts and of systematic quality differences between experimental conditions. This can lead to excluding part of a data set or replication of poor quality hybridisation. In the worst case, it means replication of the whole experiment with improved technology or different experimental design. The easily interpretable quality landscape are particularly popular in small labs, because they give very concrete feedback about the hybridisations.

Core facilities: Larger genomics facilities in research institutes, hospitals or companies run industrial style high-throughput measurement operations. In addition to the QA goals sketched above for small labs, they are interested in process optimisation and control. In W. Shewhart's terms, they use the QA toolbox to detect *special causes* of poor quality through the identification of artefacts and biases and modify their facility and experimental designs accordingly. The scores based on raw intensities, RLE and NUSE can be used within established multivariate statistical process control frameworks. A. Scherer (CEO of Spheromics, formerly at Novartis and the Australian Genome Research Facility) emphasises the importance of NUSE distributions to detect batch effects.

Quality benchmarking: The most prominent initiative for benchmarking high-throughput gene expression measurement quality is the *Microarray Quality Control (MAQC)* project led by the US Food and Drug Administration (FDA). It aims at establishing standards to ensure successful and reliable use in clinical practice and regulatory decision-making. The QA toolbox has contributed to Phase II of the development of MAQC, which aimed to assess and establish *best practices* for development and validation of predictive models for personalised medicine.

Medical diagnosis and treatment decision: Biotech companies have been developing test based on multivariate gene expression profiles obtained in individual patients. For example, a test returning a patient's individual recurrence estimate helps decide whether or not for this patient the protection provided by adjuvant chemotherapy outweighs its risks. The QA toolbox has been used by data analysts involved in the development of such tests. For example, the test *Afirma*, developed and validated by the molecular diagnostics company Veracyte, is expected to reduce the number of surgeries with their attendant morbidity (life-long follow-up treatments) in initially suspected thyroid cancer [5]. The traditional diagnosis produces up to 30%, inconclusive cases typically resulting in surgery, of which 70%-80% of

patients turn out to have benign tumours. *Afirma* succeeds in avoiding the need for half of these surgery cases, resulting in expected health care cost savings of $3000 per patient as well as improving patient health outcomes. An economic impact study concluded that routine use of *Afirma* in the USA would result in 74%, fewer surgeries in patients with benign tumours, corresponding to about $122 million medical savings [6]. Crucial steps for commercial success were FDA software validation and convincing clinicians by achieving a negative predictive rate above 94%. According to Veracyte, a key step was data QA based on RLE distributions from [1]. They shed light on the sources of variation in their custom-made gene expression microarrays, detected outliers and guided the removal of artefacts and batch effects arising from inconsistencies in operator, protocol or sample conditions.

The nature of impact of mathematical and statistical research is typically indirect and unforeseeable. We rephrase questions arising in interdisciplinary collaborations to construct methods applicable to general classes of similar problems, thereby planting seeds for fundamental long-lasting changes to industrial processes or clinical practice. Whether, when and how one particular seed will come to fruition and how enthusiastically its fruits will be picked by the scientific community during a particular time period is a process we can only partially influence. While requiring our effort at all stages, success of this sort is largely subject to external factors and chance. This section told the story of a lucky seed, one that grew into a QA toolbox widely used in the genomics community.

Acknowledgements I thank F. Collin (Genomic Health), B. Bolstad (Affymetrix) and T. Speed (UC Berkeley and WEHI Melbourne) for our longstanding collaboration. I am also grateful to G. Kennedy (Veracyte), D. Brewer (ICR) and A. Scherer (Spheromics) for support with demonstrating impact, and D. Firth (University of Warwick) for feedback on drafts of my REF 2014 impact case.

References

1. Brettschneider J., Collin F., Bolstad B.M., and Speed T.P.: Quality assessment for short oligonucleotide arrays, with 5 commentaries and rejoinder, Technometrics **50**, 241–264 (2008)
2. Allison, D., Cui, X., Page, G., and Sabripour, M.: Microarray data analysis: from disarray to consolidation and consensus, Nature Review Genetics **7**, 55–65 (2006)
3. Mullard, A.: Reliability of 'new drug target' claims called into question, News and analysis, Nature Reviews Drug Discovery 10, 643-644 (2011)
4. Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., and Speed, T.: Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Research **31**, e15 (2003)
5. Alexander, E.K. et al: Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology, New England Journal of Medicine, Aug 23 (2012)
6. Li H., Robinson, K., Anton, B., Saldanha, I., and Ladenson, P.: Cost-Effectiveness of a Novel Molecular Test of Cytologically Indeterminate Thyroid Nodules, JCEM **96(11)**, E1719 (2011)
7. Brettschneider J.: Quality assessment for high-throughput genomic data in research and clinical practice, Impact Statement REF 2014, Unit of assessment: Mathematical Sciences, Summary impact type: Technological