# Chapter 10
# Gene Expression Profiling in Asthma

**Joanne Sordillo and Benjamin A. Raby**

**Keywords** Gene expression • Transcriptomics • Differential expression • Normalization • RNA-Seq • Microarray • Pathway analysis • Systems biology • Ribose nucleic acid

## 10.1 Introduction

Transcriptomics (gene expression profiling) refers to the quantitative and qualitative characterization of the collection of ribose nucleic acid (RNA) elements expressed in a biological system and represents one of the first truly genome-wide hypothesis-free investigative approaches in molecular biology. The advent of synthetic oligonucleotide microarray technologies has enabled large-scale application of gene expression profiling in the study of human disease, particularly malignant and hematological processes. Due to favorable characteristics of these processes, including their involvement of one cellular compartment (and often a specific, monoclonal cell type), the severity of the underlying cellular perturbation under study (malignant vs. benign cells), and the accessibility to large numbers of available banked samples obtained during clinically indicated medical procedures, the study of transcriptomics in oncology has been quite fruitful, with notable translation of these techniques to novel clinical applications with diagnostic, prognostic, and therapeutic implications. Furthermore, the discovery of large populations of noncoding RNA

J. Sordillo, Sc.D. (✉) • B.A. Raby, M.D.C.M., M.P.H.
Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA
e-mail: rejoa@channing.harvard.edu; rebar@channing.harvard.edu

A.R. Brasier (ed.), *Heterogeneity in Asthma*, Advances in Experimental Medicine and Biology 795, DOI 10.1007/978-1-4614-8603-9_10,
© Springer Science+Business Media New York 2014

elements, including microRNA and long-intergenic noncoding RNA (LINCC-RNA) has expanded the scope of transciptomic profiling beyond the protein-coding messenger RNAs (mRNA).

In this chapter, we provide a brief survey of prior applications of this approach to the study of asthma, followed by an overview of the primary technical and analytical considerations that should be addressed when conducting such studies. For more detailed review of study protocols and specific analytical platforms, readers are referred to several recent publications (Matson 2009; Yakovlev et al. 2013; Dehmer and Basak 2012; Rodriguez-Ezpelete et al. 2012).

## 10.2 Applications of Transcriptomics in Asthma Research

Asthma is a complex syndrome arising from the interplay of genetic and environmental perturbations of two multicellular organ systems (the respiratory and immune systems) operating in a developmental context. Thus, in contrast to malignancies that arise from radical molecular alterations in one population of cells, the genomic study of asthma presents a more complicated model with inherent challenges, relating to selection of disease model and tissue sampling, and analytical interpretation. Nonetheless, transcriptomic profiling has been applied widely in the study of asthma and will likely continue to play an important role in biological and translational asthma research. Broad goals of such studies include discovery of novel pathways underlying asthma pathogenesis, reclassification of asthma subtypes on the basis of distinct genetic signatures, and understanding the cellular response of asthma-relevant environmental exposures, pharmaceuticals, and other perturbations, all with the hope of identifying new therapeutic targets.

Tables 10.1 and 10.2 summarize many of the published asthma genomic studies performed in patient populations, in the peripheral immune (Table 10.1) and pulmonary (Table 10.2) compartments. Thus far, virtually all published studies have employed chip-based expression platforms in conjunction with traditional analytical methods of differential gene expression analysis, cluster analysis, or a combination of both. The majority of these studies have focused on contrasts between asthma case and unaffected control status, or across strata of asthma severity. With notable exception, sample size has been small, most studying fewer than 100 patients. Given the complexity of asthma, the heterogeneity in asthma phenotype, and wide spectrum of severity, small-sized studies are of limited value, particularly with regard to broad generalization. Complicated by differences in tissue of study, patient inclusion criteria, expression microarray employed, and analytical methods, the results of these studies are largely nonoverlapping, and few discernable common insights are apparent. Moreover, few of these studies presented rigorous evidence of either intrinsic (technical and computational) or extrinsic (replication) validation, making uncertain the generalizability of their findings. Nonetheless, several of these early efforts have provided important insights worthy of discussion here.

**Table 10.1** Asthma gene expression studies (peripheral blood/immune cells)

| Study | Experimental description | Sample size | Platform | Implicated genes/pathways |
|---|---|---|---|---|
| *WBCs* | | | | |
| Orsmark-Pietras et al. (2013) | Pediatric and adult populations<br>Primary comparisons:<br>1. Asthma severity/asthma affection status in children<br>2. Adult asthmatics- expression in fractionated leukocytes | 71 subjects | Affymetrix ST 1.0 | Major finding: Bitter taste transduction (TAS2R) differentially expressed in severe asthma, highest expression in blood lymphocytes<br>Other findings: NK cell cytotoxicity, N-glycan synthesis |
| *CD4+/CD8+ T cells* | | | | |
| Tsitsiou et al. (2012) | Gene expression levels compared in severe and mild asthmatic groups vs. controls | 24 subjects | Affymetrix U133 Plus 2.0 | Activation of CD8+ T cells in severe asthma, with upregulation IL-2 regulatory pathways; less differential expression in CD4+ T cells, with upregulation of vitamin D pathway |
| *CD4+ T cells* | | | | |
| Hunninghake et al. (2011) | Analysis of gene expression and total serum IgE in pediatric asthma population | 223 subjects | Illumina Ref8 v2 | IL17RB associated with total serum IgE; sexual dimorphism of implicated gene pathways suggest sex-specific regulation of IgE |
| Kapitein et al. (2008) | Differential gene expression in infant wheeze (transient, persistent) vs. controls | 19 subjects | Affymetrix U133A | Wheezing phenotypes show differential expression of apoptosis and T-cell proliferation genes |
| *Neutrophils* | | | | |
| Baines et al. (2010) | Neutrophilic gene expression profiles were studied in noneosinophilic asthma (vs. eosinophilic asthma or control) | 28 subjects | Illumina Ref-8 v1.1 | Noneosinophilic asthma associated with distinct neutrophil gene expression profile, implicating cell motility and apoptosis regulation |

(continued)

**Table 10.1** (continued)

| Study | Experimental description | Sample size | Platform | Implicated genes/pathways |
|---|---|---|---|---|
| *Basophils* | | | | |
| Youssef et al. (2007) | Expression profiles from "releaser" basophils (produce mediators with FcεRI cross-linking) vs. "nonreleasers" in asthmatics vs. controls | 129 subjects | Affymetrix U133A | FcεRI cross-linking induces multiple distinct genes (FcεRI α, β subunit, histamine 4 receptor) in basophil "releasers" vs. "nonreleasers" |
| *PBMCs* | | | | |
| Aoki et al. (2009) | Expression profiles in exacerbation (vs. stable asthma); compared to nonasthmatics with respiratory infection (vs. control) | 34 subjects | Illumina Human Ref8 | Many of 153 differentially expressed genes in asthma exacerbations also observed in during respiratory infection |
| Bjornsdottir et al. (2011) | Peripheral expression profiles during asthma exacerbation vs. quiescent asthma | 118 subjects | Affymetrix U133A | Distinct exacerbation-associated signatures identified in innate immunity, lymphocyte activation, and downstream adaptive immune pathways |
| Hakonarson et al. (2005) | Expression patterns in GC (Glucocorticoid)-sensitive vs. GC-resistant asthmatic subjects, at baseline, and after treatment with IL-β, TNF-α | 106 subjects | Affymetrix A arrays U95 set | Expression changes in 923 genes after IL-β, TNF-α were reversed in GC responders' cells with GC treatment (15 of these genes predicted GC response in an independent test set) |
| Subrata et al. (2009) | Expression profiling of PBMCs during asthma exacerbations vs. convalescence; qRT-PCR on specific cell populations | 67 subjects | Affymetrix U133 Plus 2.0 Gene Chips | Expression of pathways during exacerbation: arachidonic acid, leukocyte migration, innate, adaptive immunity; gene specific upregulation mainly in monocytes/DCs |
| Shin et al. (2011) | Differential gene expression in asthmatics vs. controls; of 170 differentially expressed genes, top 8 used in prediction models | 52 subjects | Illumina Human Ref8 BeadChip | Best predictive for asthma model contained *MEPE*, *MLSTD1*, and *TRIM37* (98 % sensitivity, 80 % specificity) |

**Table 10.2** Asthma gene expression studies (lung)

| Study | Experimental description | Sample size | Platform | Major genes pathways identified/ findings |
|---|---|---|---|---|
| *Endobronchial biopsies* | | | | |
| Yick et al. (2013) | RNA-Seq for differential gene expression study in asthma | 9 subjects | Ovation RNA-Seq/GS FLX+ | Differential expression of individual genes (Pendrin, Periostin, and BCL2), ten gene networks in cell morphology, movement, and development |
| Choy et al. (2011) | Quantitative description of Th2 inflammation in asthmatics; and correlation with inflammatory markers | 40 Subjects | Agilent two-color 44K array | Th2 signature correlated with CCL26, IL-13, IL-5, Wnt, TGF-β, PDGF expression, and associated with increased IgE, blood and BAL eosinophils, decreased neutrophilic and Th1 responses |
| *Airway epithelial cells* | | | | |
| Woodruff et al. (2007) | Expression profiling of asthmatics, smoking controls, non-smoking controls for epithelial dysfunction markers, effects of corticosteroids | 86 subjects | Affymetrix U133 Plus 2.0 | Expression of CLCA1, periostin, and serpinB2, upregulated in asthma vs. smoking controls; corticosteroids associated with decreased expression of CLCA1, periostin, Serpin B2, upregulation of FKBP51 in asthmatics |
| Woodruff et al. (2009) | Expression of IL-13 inducible genes used for molecular Th2 phenotype classification, validated using cytokine, inflammatory and ICS (inhaled corticosteroid) responses | 70 subjects | Affymetrix U133 Plus 2.0 | Gene expression analysis classified asthmatics into two subgroups "Th-2 High" and "Th2-low" (indistinguishable from controls) "Th2-high" showed greater peripheral and BAL eosinophilia, ICS response, serum IgE and mucin expression than "Th2-low" subjects |

(continued)

**Table 10.2** (continued)

| Study | Experimental description | Sample size | Platform | Major genes pathways identified/ findings |
|---|---|---|---|---|
| Bochkov et al. (2010) | Rhinovirus(RV)-induced gene expression compared in asthmatics vs. nonasthmatics | 18 subjects | Affymetrix U133 Plus 2.0 | After RV-infection, some immune-related genes differentially expressed in asthmatics; most of asthma-related differences in expression seen prior to RV infection |
| Freishtat et al. (2009) | Human, mouse public datasets used to study overlapping gene expression in asthma, cigarette smoke exposure | 4 datasets:2 human, 2 mouse | Affymetrix U133A MOE430A | 26 Overlapping genes for expression profiles in asthma and cigarette smoke exposure; 18 genes in lung oxidative stress pathways (Thrombospondin 1, TIMP1 central to gene network) |
| Kicic et al. (2010) | Study of airway repair mechanisms in asthmatics (compared to atopic subjects and healthy non-atopic controls) | 112 subjects | Affymetrix U133A | Slower epithelial cell repair in asthmatics (vs. atopic or healthy controls); differential expression of gene sets for repair/remodeling in asthmatics; fibronectin expression downregulated in asthma |
| *Airway smooth muscle* | | | | |
| Sutcliffe et al. (2012) | Oxidative stress burden evaluated by DNA damage, intracellular ROS; measurements then related to genome-wide expression data in asthmatics and controls | 79 subjects | Affymetrix U133A | Asthmatic subjects showed increased oxidative burden in ASM cells and also had higher NOX4 expression vs. controls; blocking NOX4 expression in siRNA knock-down demonstrated abrogated contractility in asthmatic ASM |

| | | | | |
|---|---|---|---|---|
| *Alveolar macrophages* | | | | |
| Madore et al. (2010) | Gene expression profiles in allergic asthmatics vs. control subjects | 10 subjects | AffymetrixU133A | Of the 50 differentially expressed genes, 19 in stress/immune response pathways, including 9 in the heat shock protein family |
| *Induced sputum* | | | | |
| Baines et al. (2011) | Unsupervised hierarchical clustering of expression data for molecular phenotyping of asthma in adults | 59 subjects | Illumina Humanref-8V2 | 3 Transcriptional asthma phenotypes defined by IL-1,TNF-α, NF-κB pathways; 2 corresponded to neutrophil or eosinophil inflammation<br>Phenotype clusters associated with clinical outcomes (FEV1 and FENO) |
| *Fetal lung tissue* | | | | |
| Melen et al. (2011) | Differential gene expression during fetal lung development | 38 subjects | Affymetrix U133 Plus 2.0 | No overrepresentation of asthma candidate genes during lung development; asthma GWAS genes differentially expressed during lung development (ROBO1, RORA, HLA-DQB1, IL2RB, and PDE10A) |

### *10.2.1 Immune System Genomic Profiling Studies*

The easy access of peripheral blood cells, together with the central role of the immune system, in asthma pathogenesis has motivated multiple genomic studies of cell populations collected by phlebotomy. In theory, studies of specific, homogeneous cell populations (for example, CD4+ lymphocytes, eosinophils, or neutrophils) should provide more readily reproducible biological insights regarding specific pathogenic mechanisms than studies of heterogeneous cell populations (for example, peripheral blood mononuclear cells; PBMCs). However, the latter may be more powerful for genomic classification in clinically motivated studies, for example, when the primary goal is to reclassify patients into molecularly similar subgroups. While this has largely borne true, notable exceptions are evident. For example, Hakonarson and colleagues examined the genomic profiles of 106 glucocorticoid-sensitive and resistant asthmatics using PBMC-derived RNA extracted in a resting state and following in vitro treatment with IL-1 (Hakonarson et al. 2005). A total of 11,812 genes were examined with high-density oligonucleotide microarrays in both resting PBMC (106 patients) and IL-1β/TNF-α stimulated cells treated with or without dexamethasone. More than 5,011 differential expressed genes were detected, of which 923 were reversed by dexamethasone in glucocorticoid responsive patients. A smaller subset of 15 genes classified responders from nonresponders with 84 % accuracy. Technical validation for 11 of these genes was confirmed, with one gene—NFKB—demonstrating predictive accuracy of 81.2 %. Studies in other individual peripheral blood cell types have provided other insights. Tsitsiou and colleagues compared CD4+ and CD8+ T-lymphocyte expression profiles of severe asthmatics vs. controls. They found that compared to CD8+ cells, CD4+ profiles yielded relatively few differentially expressed genes, with the exception of upregulation of the Vitamin D signaling pathway. CD8+ derived profiles showed multiple upregluated genes in severe asthmatics, with enrichment for T-cell activation and inflammation pathways (Tsitsiou et al. 2012). In a transcriptomic study of CD4+ T lymphocytes from mild-to-moderate childhood asthmatics, Hunnighake and colleagues found striking differences between boys and girls in those genes and gene pathways associated with total serum IgE levels (Hunninghake et al. 2011). For example, the gene most strongly correlated with IgE levels—the Interleukin 17 Receptor B (IL17RB)—was only correlated with IgE in boys, with no single gene demonstrating strong correlation in girls. The sets of gene pathways correlated with IgE levels in boys and girls were also nonoverlapping, suggesting distinct molecular mechanisms underlying the noted sexual dimorphism of serum IgE—a critical allergic intermediary phenotype in asthma.

While transcriptional phenotyping efforts aim to describe underlying heterogeneity in asthma cases (which is presumed to be relatively constant), other studies attempt to capture transient changes in gene expression that occur during asthma exacerbations. Asthmatic subjects' gene expression profiles from PBMCs collected during an exacerbation show increased expression in innate immune pathways (TLRs, interferon response genes), adaptive immunity (B-cell and T-cell lymphocyte activation genes), and upregulation of arachidonic acid/prostaglandin pathway

genes as compared to PBMC samples drawn during a quiescent time period (Bjornsdottir et al. 2011; Subrata et al. 2009). PBMC expression profiles during acute asthma exacerbation also show considerable overlap with expression levels from nonasthmatics experiencing upper respiratory infection (Aoki et al. 2009), suggesting that immune pathways activated in response to infection may amplify Th2-mediated responses during asthma exacerbations. An in vitro stimulation experiment of infection-related inflammation (atopic monocytes exposed to IFN-α) demonstrated upregulation of the same atopic pathway genes observed in PBMCs from asthmatic children during exacerbation (Subrata et al. 2009).

### *10.2.2 The Pulmonary Compartment*

Gene expression studies of pulmonary tissues in asthma are equally heterogeneous, ranging from studies of whole lung tissue (from surgical specimens) to studies of bronchial epithelium (derived by endobronchial brushing) to studies of alveolar macrophages from induced sputum. Expectedly, the results from these studies are similarly disparate to those observed in the peripheral compartment. They are also no less revealing. Much attention has focused on a transcriptomic profiling study of bronchial epithelium collected during a clinical trial of inhaled corticosteroid therapies, resulting in the identification of a potential pharmacogenetic biomarker—Periostin (Woodruff et al. 2007). Profiling of airway epithelial brushings obtained from nonsmoking asthmatics ($n=42$) and healthy controls ($n=28$), identified 22 differentially expressed genes, including three genes whose expression reverted to levels similar to those observed in healthy controls following treatment with inhaled corticosteroids: chloride channel, calcium-activated, family member 1 (CLCA1), periostin, and serine peptidase inhibitor, clade B (ovalbumin), and member 2 (serpinB2). In vitro studies confirmed increased expression of these three genes upon cell culture with interleukin-13, a phenomenon also reversed with corticosteroid treatment. A fourth gene, FK506-binding protein 51 (FKBP51), was markedly upregulated in vivo following inhaled corticosteroid treatment, and the expression of all four genes was predictive of clinical corticosteroid response. An independent RNA sequencing study (RNA-Seq) of endobronchial biopsies revealed differential expression of both novel and confirmative asthma-related genes and included upregulation of pendrin, periostin, and downregulation of BCL2 (Yick et al. 2013).

A common observation of many lung compartment studies (of varying cell types) is the induction of oxidative stress response genes in response to relevant exposures, such as mechanical and oxidative stress, among asthmatics compared with healthy subjects. For instance, primary airway smooth muscle cells from asthmatics show a higher burden of oxidative stress (including oxidative stress-induced DNA damage and increased production of reactive oxygen species) and also demonstrate increased expression of NADPH oxidase (NOX) subtype 4, an enzyme which may be involved in airway hypercontractility (Sutcliffe et al. 2012). A systems biology analysis based on the integration of publically available datasets (two mice and two human) revealed 18 oxidative stress genes common to both cigarette-exposed and asthmatic

lung cells, including TIMP1 (tissue inhibitor of metalloproteinase 1) and THBS1 (thrombospondin 1), which were both central to the molecular network constructed using the overlapping transcripts (Freishtat et al. 2009).

## 10.3 Methods

### *10.3.1 Tissue Sampling and RNA Isolation*

The most important determinants of the quality and reproducibility of genomic data relate to the quality of the input RNA sample, including its purity, integrity, and quantity (Table 10.3); careful consideration of these features is critical during study design and execution. Due the inherent instability of RNA and the ubiquity of RNase enzymes, minimizing RNA degradation is a priority. While many technical issues can be addressed by normalization in the later stages of study (particularly those related to RNA extraction), the choice of tissue type, methods of tissue procurement, and methods of cell isolation must be designed specifically with reference to their impact on RNA quality, as inferiorities introduced in these earliest stages are often not addressable later on. Moreover, the sensitivity of the transcriptome to changes of the cellular environment (including the induction of hypoxia or temperature-related stress responses) mandates that any technical deficiencies that

**Table 10.3** Determinants of RNA quality

| Feature | Sources of inferiority | Solutions |
|---|---|---|
| Purity | Multicellular tissue | Tissue microdissection |
| | DNA contamination | Cell sorting |
| | Organic or inorganic contamination | Cell culture |
| | Formaldehyde-fixed paraffin-embedded (FFPE) samples | Analytical considerations |
| | | DNase treatment |
| | | Rigorous protocol adherence/ repeat extraction protocol |
| | | FFPE-specific extraction procedures and platforms |
| Integrity | Sampling-related gene expression induction (hypoxia- or temperature-induced stress response, autophagy, and apoptosis) | Rapid sample preservation, including flash freezing. |
| | | Modest sample cooling |
| | RNA degradation (during extraction) | Immediate RNA extraction |
| | RNA degradation (from multiple freeze–thaw cycles) | RNA preservatives |
| | | Sample storage in multiple aliquots |
| Quantity | Low cellular yield (small sample size, low percentage of target cell) | Sample pooling |
| | | Cell culture and expansion |
| | Low intracellular RNA content (granulocytes) | RNA amplification procedures |
| | High RNase content (eosinophilic inflammation) | Low yield protocols |

arise during sampling and RNA extraction are introduced nondifferentially between study groups (i.e., between cases and controls and treated vs. untreated samples), so as to avoid technical biases that irreversibly confounds data analysis and interpretation. Thus, study design should ensure that all samples are obtained and processed uniformly (Kerr 2003).

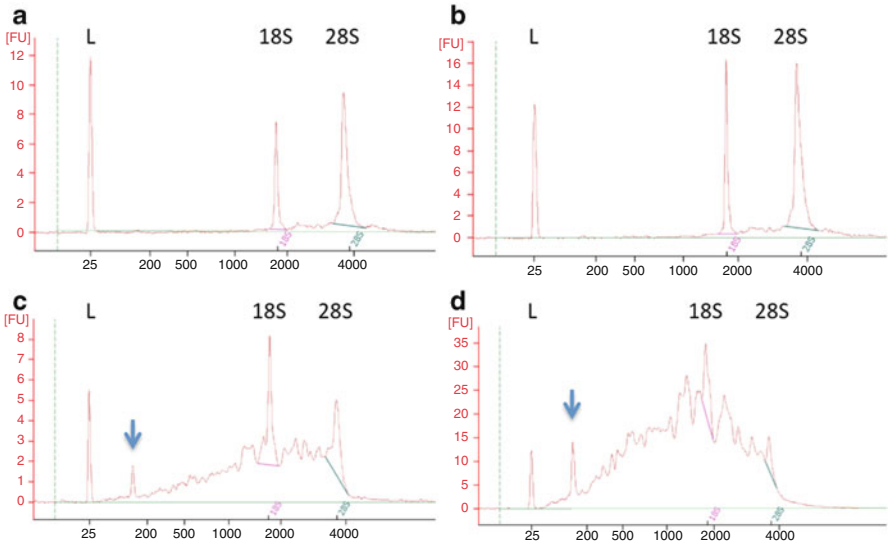
Guanidinium thiocyanate–phenol–chloroform extraction (Chomczynski and Sacchi 1987), a phase separation protocol, is the most commonly used method of RNA extraction. Several extraction kits from Invitrogen (TRIzol), Bioline (Trisure), and Tel-Test (Stat-60) are commercially available that offer high-throughput scalability for large-scale studies and small amounts of starting substrate. Recent modifications include chemistries for isolation of a wider range of RNA moieties, including shorter length microRNAs, without sacrificing yield of larger mRNA populations. Regardless of methods used, the resultant products should undergo rigorous quality assessment. RNA integrity, sizing, and concentration are estimated by either agarose gel (with ethidium bromide) or using microfluidic instrumentation (BioAnalyzer 2100, Agilent Inc.), estimating the ratios of the two ubiquitously expressed 28S and 18S ribosomal protein subunits (see Fig. 10.1). Traditional spectrophotometric analysis (optical density measurement) provides additional information regarding sample purity, with A260:A280 ratios of ~2.0 reflecting an absence of organic contamination. Concentration measurement by fluorescent dye analysis on agarose gel (with ethidium bromide) or by photometer is advised, particularly for low yield samples, though such samples can also be accommodated with the BioAnalyzer using modified (Pico) protocols. Extracted samples should be stored in liquid nitrogen in RNase-free tubing, with division of samples into multiple aliquots to avoid sample degradation from repetitive freeze–thaw cycles.

### *10.3.2 Sample Profiling (Table 10.4)*

**Platform Considerations**

Oligonucleotide Microarrays

Until very recently, most genome-wide expression profiling was performed using single-channel, hybridization-based oligonucleotide microarray technology (Schena et al. 1995). In this method, RNA samples are converted to fluorescently labeled cDNA by in vitro reverse transcription (IVT) (Rajeevan et al. 2003), generating a pool of targets. This complex pool is subsequently hybridized against a microarray surface densely studded with populations of oligonucleotide DNA sequences, 20–50 bases in length, each of which is complementary to a specific target RNA sequence. These oligonucleotide probes, situated at fixed positions on the microarray slide, bind their complementary targets. Quantitative measures of the fluorescent intensities at each fixed probe site, captured by confocal microscopic




| Sample | 28S:18S | Concentration | RIN | Interpretation |
|---|---|---|---|---|
| A | 2.1 | 22 ng/μl | 9.9 | High quality, modest yield |
| B | 1.8 | 42 ng/μl | 9.9 | High quality, high yield |
| C | 0.7 | 111 ng/μl | 6.0 | Partial degradation |
| D | 0.3 | 467 ng/μl | 3.1 | Extensively degraded |

**Fig. 10.1** RNA quality assessment: Illustrative examples of BioAnalyzer 2100 RNA assessment analysis from four total RNA samples derived from CD4+ T lymphocytes with plots of fluorescent units (FU) as a function of RNA size (in base pairs). Size standard from ladder (L) correctly situated at 25 bp. Samples A and B represent good quality samples with no degradation and adequate RNA concentrations. Samples C and D are both of poor quality, with evidence of degradation, including accumulation of short RNA products at ~100 bp (*arrows*). Neither sample C nor D is suitable for transcriptome profiling. Note that peaks between 25 and 100 bases are desired in preparations derived from small RNA extraction protocols. Small peaks at 100 bp are often observed following TRIzol or phenol extraction, denoting small ribosome proteins 5S and 5.8S, as well as tRNAs, and do not represent poor quality sample

fluorometry, correspond to the relative abundance of the target RNA in the biological sample. Improvements in chip manufacturing, probe density, and imaging resolution have facilitated development of arrays with more than one million unique features at relatively low cost (<$150 per sample), enabling simultaneous characterization of virtually all known RNA sequences, including numerous splicing isoforms, using relatively small amounts (~100 ng) of starting RNA. Manufacturers have developed a wide range of arrays that assay human and nonhuman model organismal genomes and typically offer a range of chip designs that differ with

**Table 10.4** High-throughput gene expression profiling platforms

| | Oligonucleotide microarrays | | |
|---|---|---|---|
| | Two channel | Single channel | RNA-Seq |
| Advantages | Low RNA requirement<br>Low array costs<br>More direct comparison of paired samples via competitive hybridization<br>Analytic methods well established | Low RNA requirement<br>Low array costs<br>Wide availability<br>Wide selection of array types/content<br>Analytic methods well established | Sequence-independent measurement<br>Novel isoform, sequence identification<br>Accommodates both long and short-length RNA species.<br>Detects sequence polymorphism and allelic expression<br>Read-depth measurement of transcript abundance |
| Disadvantages | More limited availability<br>Lower target content<br>Fixed content<br>Labor intensive (equimolar sample mix)<br>Sequence-dependent hybridization<br>Intensity-based relative measure of transcript abundance | Fixed content<br>Labor intensive (equimolar sample mix)<br>Sequence-dependent hybridization<br>Intensity-based indirect measure of transcript abundance | High RNA requirement<br>High array cost<br>Sensitive, labor-intensive library preparation<br>Intensive bioinformatics support for sequence alignment required<br>Analytic methods less well established<br>Considerable data storage challenges |

respect to array content (number of genes, isoforms, and RNA type), as well as the number of samples (arrays) that can be assayed per chip. As a consequence, this technology has been the most widely adopted, resulting in a well-developed understanding among the scientific community of array performance (including each platform's strengths and liabilities) and a comprehensive set of statistical approaches for image processing, sample normalization and quality control, and data analysis (see below). Using standardized approaches, implementing stringent adherence to quality assessment and uniform analysis methodologies, data reproducibility has been demonstrated to be high, both within and across laboratories (Irizarry et al. 2005; Shi et al. 2010).

The earliest oligonucleotide microarray protocols implemented a two-channel competitive hybridization approach, whereby the relative expression of two biological samples are contrasted directly by separate labeling of each with fluorophores of differing fluorescent spectra, hybridizing both to the same array in equimolar concentrations, and measuring their relative intensities by image capture of each fluorescent spectrum (i.e., two channel). There are several limitations of the two-channel approach, most notably the needs for stringent equimolar mixing of two target pools

and for performing technical replicates with dye-swap to avoid dye-dependent bias, as well as a reduced feature (probe) content per array. Though these disadvantages have led to preferential adoption of single-channel protocols for most studies, a natural application of two-channel arrays is in studies of matched, paired samples (for example, within subject comparisons of samples obtained pre- and posttreatment), where the contrast of interest can be assessed directly using individual assays, reducing the potential for cross-array technical bias.

The inherent disadvantage of oligonucleotide microarray expression measurement is the dependency on efficient, unbiased hybridization of target sequence to a predefined, fixed probe set. Due to differences in GC-content and sequence complexity across probe types, hybridization kinetics is not uniform across probe types. Limitations in microarray probe capacity restrict the number of discrete targets that can be assayed (for example, splice isoforms) and also limit the dynamic range of intensity measurement (the upper-limit of which is bound by complete probe saturation). The dependence of efficient hybridization on sequence alignment with features prespecified on the chip precludes novel transcript identification. In addition, DNA polymorphism can differentially impact hybridization between subjects, spuriously generating differences in measured gene expression, even for variants that themselves have no functional impact on RNA transcript abundance (Alberts et al. 2007). Until recently, these limitations were largely not addressable and considered recognized trade-offs for comprehensive, inexpensive genome-wide surveys of expression. However, the development of hybridization-independent, highly parallelized (so-called next-generation) sequencing platforms have largely solved these issues (Schuster 2008).

## Next-Generation Sequencing

Over the past half-decade, several platforms have been developed to enable sequencing of oligonucleotide sequence (DNA or RNA) in a highly paralleled fashion, without the need for predefined sequence-dependent hybridization. These include platforms developed by Roche, Illumina, Pacific Biosciences, and Helicos BioSciences, among others. Detailed reviews of these technologies are available elsewhere (Metzker 2010). These methods generate sequence reads of between 30 and 135 bases in length (depending on platform), sampled (fairly) randomly from the target sample. With sufficient sequencing, adequate coverage can be attained to accurately call bases from complete genomes. In addition to quantitative sequence analysis for the detection of genetic variation, sequence data can be analyzed quantitatively, as read-count—the number of times a given base is represented in a random read—correlates with the amount of target sequence starting concentration (Wold and Myers 2008). Unlikely microarrays, where transcript abundance is indirectly quantified by measuring hybridization events, sequencing-based measures represent more direct observations, namely individual transcript reads. As such, next-generation sequencing of RNA (RNA-Seq) represents a powerful tool for comprehensive characterization of transcript abundance at a genome-wide level (Wang et al. 2009), with excellent technical reproducibility (Marioni et al. 2008; Mortazavi

et al. 2008). Freed from the constraints of prespecified probes, RNA-Seq also enables complete enumeration of alternative splicing events and use of alternative transcription start sites, including novel isoform identification. These features make RNA-Seq quite attractive over microarray methods. However, limiting its introduction in most laboratories is the considerable cost, which is currently four to five times that of traditional microarray expression profiling, though it is expected that this will continue to drop to a more competitive price point in the near future. The other critical limitation of RNA-Seq is that regarding sequence preprocessing and downstream analysis, which are considerably more involved, and comparatively less well developed, compared to microarray analysis. Unlike microarrays, where gene identities of each probe are known by their coordinates on the array, the sequences generated from RNA-Seq must first be aligned to a reference genome, then annotated and assigned gene identity. This requires experienced bioinformatics analysis, also adding to the total costs of transcriptome characterization by RNA-Seq. The ability to accurately align sequence is dependent on sequence length (platform dependent), read-depth (which is contingent on the capacity of the platform, the size of the targeted genome, and the number of genomes analyzed per sequence run), and genome complexity. Moreover, the amount of data generated per sequence run (terabytes) is considerably larger than that for microarrays (megabytes), necessitating access to large-scale computing and storage capacity.

RNA-Seq, though powerful, is not without its challenges. Paramount are several recognized sampling biases, and the need for novel analytical strategies to accommodate these issues and perform statistically robust experimental analysis. With regard to sequence bias, the most well understood relate to local sequence complexity, inadequate library preparation, and sampling dependency on gene length. Though a random process, larger genes have greater likelihood of being sequenced than smaller transcripts, introducing gene-specific biases. In addition, the more complex (unique) a particular sequence, the more likely it will be accurately aligned. Similar to microarray-based assessments, target GC content induces systematic differences in read depth (Pickrell et al. 2010). As an additional complication, sequences (genomic regions) are sampled insufficiently if library preparation is inadequate. The process of library preparation demonstrates inter- and intratechnician variability. Thus, like with traditional microarray analysis, it is likely that RNA-Seq is similarly susceptible to technical batch effects, which should be accounted for during study design and analysis (Hansen et al. 2012).

### Sample Processing Considerations

Despite substantial advancement in sample processing, labeling chemistries, array synthesis and hybridization protocols, microarray profiling remains highly sensitive to technical artifact, introducing the potential for biased measurement and erroneous data interpretation (Benito et al. 2004; Fare et al. 2003). These biases can be introduced at various stages, during chip manufacturing, sample labeling, hybridization, or image capture. Technician-dependent variance is also frequently observed. Though such issues can be largely overlooked in small studies of a handful of

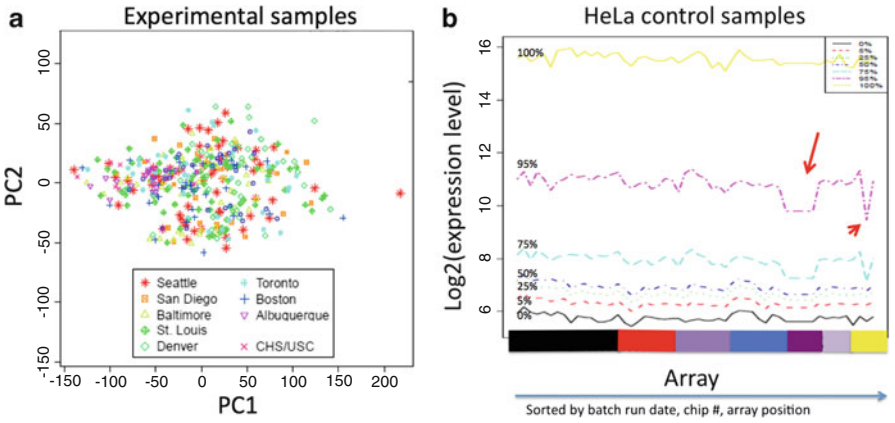



**Fig. 10.2** Sample processing strategies: (**a**) Sample randomization for the avoidance of technical bias: dot plot of first two principle components (PC) analysis of whole blood gene expression profiles of 285 samples collected from asthmatic subjects at nine centers across the USA over a 4-year periods. Samples were hybridized and imaged over 6 months on 32 Illumina HumanHT12-v3 BeadChips. The homogenous distribution of samples from disparate sites over the two-dimensional PC space supports an absence of confounding between technical batch and study site, as confirmed by formal statistical testing, illustrating successful sample randomization during processing. (**b**) Longitudinal internal quality control analysis: line plot of global gene expression intensity patterns of a replicate HeLa-cell derived RNA sample hybridized to unique array positions for 32 Illumina HumanHT12-v3 BeadChips over a 6-month period. *Lines* denote 0, 5th, 25th, 50th, 75th, 95th, and 100th percentiles of expression for each sample. Arrays processed together in batches denoted by color coding along abscissa. Note a *purple cluster* of four arrays (*long arrow*) with deviant intensities, suggesting improper processing of one batch of samples that should be considered suspect and candidates for repeat profiling. In contrast, a *yellow cluster* of three arrays processed simultaneously revealed one of three (*short arrow*) with deviant intensities. The 11 experimental samples corresponding to the same chip revealed similar deviations, resulting in removal of these data from consideration and repeat profiling of the samples

samples, processed by one technician over several days, the asthma studies typically conducted in larger patient populations are highly susceptible to these concerns.

While subtle technical variability is both expected and tolerable, it cannot be overstated that, if introduced in a nondifferential way, technical bias is not easily amenable to downstream statistical correction. For example, if samples from asthmatic cases and healthy controls are labeled and hybridized in separate batches, resulting in systematic (technical) differences in global gene expression measurements, no analytical trick can reliably disentangle the true biological expression differences from these so-called technical "batch effects" (Scherer 2009). To avoid this, it is imperative that design strategies aimed at bias prevention and detection be implemented at the outset. Most useful is adherence to a strategy of consistent, repeated sample randomization. At each major processing step (sample extraction, labeling, hybridization, and imaging), samples should be assigned to random batches, irrespective of disease status or other distinguishing clinical characteristics

(treatment group, severity, and gender), so as to minimize the possibility that any relevant covariate is confounded with latent technical artifact. This strategy is particularly useful for studies carried out over many months, where variations in reagent manufacturing, laboratory staffing, or ambient environmental conditions are almost certain (Fig. 10.2a). For such large studies, we also recommend introduction of routine surveillance for reproducibility by randomly selecting samples for repeat testing (to screen for within sample technical variability) and inclusion of standard control samples (for example, a pool of equimolar concentrations of total RNA from all available study sample) that is run with each batch of samples over the course of the study (Fig. 10.2b). Such data can identify outlier batches and potentially be used during preprocessing procedures.

### *10.3.3 Analytical Considerations*

A comprehensive discussion of the many analytical considerations surrounding traniscriptomic analysis is beyond the scope of this chapter, and many detailed, accessible references elegantly explore these issues. Here, we provide a general overview of the basic principles of microarray analysis to orient the inexperienced reader.

Following image processing and data capture, there are four main components of transcriptomic analysis (1) quality control analysis; (2) data preprocessing; (3) feature selection; and (4) experimental analysis. While the last component is of greatest scientific interest, it is entirely dependent on careful execution of the first three.

#### Quality Control Assessment

Quality control assessment is performed both study wide to screen for systematic bias and for individual samples. Individual arrays with mean intensities >1 SD from the mean should be evaluated as potential outliers and considered for removal. Cluster analysis of genes mapping to the sex chromosomes can be used to identify gender mismatches, suggesting potential sample mixup. Replicate samples should be compared to estimate technical variance.

#### Preprocessing

Multiple methods have been proposed for sample preprocessing, the most accepted of which include regression based methods, probabilistic models, and multivariate models (including principle components adjustment and surrogate variable analysis modeling). The underlying premise of many regression and probabilistic normalization methods is that, under most circumstances, the majority of genes show either no, or relatively similar, expression across samples. Batch effects can also be modeled using Bayesian (Johnson et al. 2007) or other (Luo et al. 2010) approaches.

Nonparametric procedures, like quantile normalization, scale measures across arrays uniformly, while preserving the rank order of genes (Irizarry et al. 2003). In contrast, multivariate models like PCA and SVA (Leek and Storey 2007) characterize the bulk structure of the data, defining latent variables from the expression data itself that often reflect the effects of either known (for example, differences in genetic ancestry) or unknown technical factors that impact global gene expression patterns. These variables can then be adjusted out during downstream data analysis, provided none are strongly associated with the contrast (i.e., phenotype) of interest. Though described as distinct from the experimental analysis phase, these later normalization methods are often scripted and executed together with statistical inference. Often, these methods can be applied in series, so as to reduce computation time during iterative experimental analyses steps.

## Feature Selection

Feature selection refers to the removal from consideration during analysis of subsets of probes with undesirable characteristics, with the goal of reducing the potential for spurious gene detection while preserving experimental sensitivity. Examples of feature selection include filtering of probes that show either no expression, or minimal variance in expression, across the population (there is little utility in formally testing genes whose expression is static), or probes whose sequence aligns to more than one potential gene target. Additional filtering could include those probes that target RNA sequence known to harbor common genetic polymorphism (i.e., SNPs), which would interfere with hybridization and generate spurious association, particularly in integrative genomic studies that consider both expression and genetic variation simultaneously (Murphy et al. 2010). All of these aforementioned filters are nonspecific, in that they are not imposed with reference to the biological question of interest, and thus do not bias statistical inference.

## Experimental Analysis

### Differential Gene Expression Analysis

The procedures used for experimental data analysis are dictated primarily by the biological question of interest (Table 10.5). Most analyses begin by defining the subset of genes that demonstrate statistically significant fold-differences in expression between cellular states (i.e., asthma vs. no asthma). For studies of dichotomous conditions, standardized *t*-tests, such as those implemented in the RMA procedure (Bolstad et al. 2003), are applied with significance determined using genome-wide thresholds that account for the large number of statistical tests performed. Numerous software packages, including MAS5 (Lim et al. 2007) and limma (Smyth 2005), are available for efficient implementation of these methods. Appropriate statistical models have been developed to accommodate other phenotypes, including continuous or censored phenotypes. The gene lists generated from these analyses is then

**Table 10.5** Experimental data analysis procedures

| Test procedures | Question of interest | Commonly used methods |
|---|---|---|
| Differential expression: *t*-tests, ANOVA, regression | Which genes are differentially expressed in condition X compared to condition Y? | limma, RMA |
| Gene set enrichment analysis | Are specific pathways/subsets of genes differentially expressed in condition X? | GSEA, GOstats |
| Network modeling | How are the genes in this condition related to each other? | Coexpression networks, Graphical Gaussian Models |
| Supervised machine learning | Can I differentiate two or more known cellular states (i.e., cases vs. controls; good vs. poor prognosis) based on gene expression profiling? | Support vector machine, Self-organizing maps |
| Unsupervised machine learning | Are there subgroups of my disease of which I am not aware? | K-means clustering |

examined for biological insight. Though informal gene list interpretation, based on investigators' knowledge, is invariably performed, numerous bioinformatics approaches are available to evaluate the biological significance of observed profiles in a rigorous, statistically motivated framework (Alonzi et al. 2001; Subramanian et al. 2005). These pathway, or gene set enrichment analyses, evaluate whether the observed set of differentially expressed genes are members of specific, predefined gene groups with common biology. Examples of such groupings include membership within specific metabolic pathways, chromosomal locations, similar sequence features, or having similar patterns of expression in response to cellular perturbation. For example, Pietras and colleagues demonstrated enrichment of several previously unrecognized signaling pathways, including the downregulation of N-glycan biosynthesis and the upregulation of the bitter taste transduction signaling pathways in severe asthma (Orsmark-Pietras et al. 2013).

### Network Modeling

Though powerful data mining approaches, the ultimate utility of pathway-based analytical approaches are dependent on the quality of the databases queried. Though some gene collections are in near complete form (for example, detailed physical genetic maps), others, including many poorly characterized metabolic pathways, are sparser. Network modeling represents an alternative analytic strategy that attempts to model the biological process under study by developing gene networks using the experimental data itself (Hyduke and Palsson 2010; Vidal et al. 2011). These methods are motivated by the notion that most biological states are determined by the interaction of numerous genes and by the observation that biological systems operate as scale-free networks, displaying a so-called small world property, where any two genes in a network are connected by a small number of links (Barabasi 2009). Using network-modeling approaches, one can define the interrelationship of genes

within the transcriptome and then define the subnetwork of genes that demonstrate greatest change with by experimental state (Chu et al. 2009, 2011; Schafer and Strimmer 2005). Coexpression models define network structure by identifying gene sets with similar expression patterns across disease states, experimental conditions, or temporally. Models that consider additional genomic factors that influence transcriptional regulation, including promoter sequences, chromatin modifications, regulatory genetic polymorphism, and microRNA binding, offer more complete modeling, though are reliant on external data sources. While widely applied in the study of oncology, use of such strategies in asthma to date has been largely restricted to modeling of protein–protein interaction data (Hwang et al. 2008).

### Clustering Algorithms and Machine Learning Approaches

Machine learning algorithms represent a broad class of methods that mine multivariate datasets for underlying patterns, with the goal of developing predictive functions (classifiers) that can reliably differentiate samples into specific subgroups (Inza et al. 2010; Larranaga et al. 2006). In gene expression analyses, the premise is that the predictive functions elucidate subgroups that correspond to inherent biological differences between samples. In *supervised machine learning approaches*, the predictive function learns using analyst-predefined labels (for example, case–control status), with the goal of defining gene subsets that can accurately classify samples into their respective subgroups. Similarly, supervised methods can be applied for the classification of expression patterns across samples, to define subsets of genes that follow known patterns of expression (as applied in support vector machine learning). Conversely, in *unsupervised machine learning* approaches, functions are applied with few predetermined notions regarding the underlying data structure, enabling unbiased data mining, with the goal of defining previously unknown sample subgroups with unique biological properties. Numerous clustering algorithms have been developed for these purposes, including both hierarchical and nonhierarchical methods, each with inherent advantages and disadvantages related to underlying assumptions of the data structures, the questions being addressed, and their computational burden. These issues are discussed in detail elsewhere (Kerr et al. 2008). Regardless of method employed, it must be stressed that due to the inherent $p>n$ problem inherent in genomic analysis (where the number of features being tested far exceeds the number of subjects), machine learning approaches are highlight susceptible to model over fitting, resulting in partitioning of data into biologically meaningless, yet statistically robust, subgroups. As such, it is recommended that procedures be implemented at the outset that address this concern, including a priori creation of test and validation datasets and use of both internal and external cross-validation procedures. Only those models that survive stringent validation assessments should be considered viable for further interpretation.

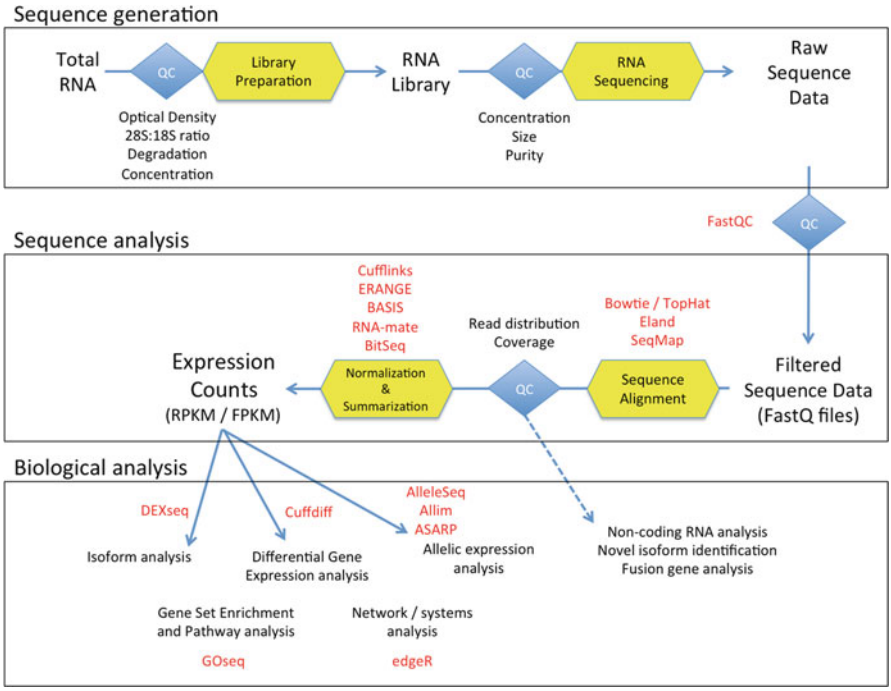



**Fig. 10.3** RNA-Seq analysis pipeline: Raw sequence reads are converted to summary measures of transcript abundance through a series of analytic and quality control steps that include filtering of poor quality reads, alignment of reads to reference sequence, gene-based read count normalization, and summarization. Transcript abundance is expressed as either reads or fragments per kilobase exon mapped per megabase sequence (RPKM or FPKM, respectively), metrics normalized by gene length. These counts serve as input for downstream biological inference and interpretation, including traditional differential gene expression analysis, exon-specific (isoform) analysis, or allelic expression analysis, where polymorphic transcript sequence variants are assessed for preferential expression of one allele over the other in heterozygous subjects. Results systems based analyses, including network building, gene set enrichment, or pathways analyses. Exemplars of available RNA-Seq software at each analytic step are indicated in *red*

## RNA-Seq Analysis

In broad terms, the analysis of RNA-Seq data is conceptually similar to that of microarray analysis, with a similar framework that includes quality control assessment and preprocessing, with screening for technical covariation, feature selection, and experimental analysis. However, due to the inherent differences in data structures, entirely distinct suites of software are required for RNA-Seq analysis (Fig. 10.3). Data preprocessing includes mapping of reads and alignment to reference genomes, followed by data normalization and summarization, where by aligned reads are translated into more biologically meaningful transcript counts. The number of reads generated for a given transcript is proportional to both the abundance of the transcript and transcript length, as larger transcripts will be

represented by a larger number of random sequence fragments. As such, to avoid systematic biases, normalization techniques must account for differences in gene length. Counts are thus expressed as reads (or fragments) per kilobase exon mapped per megabase sequence. Differential expression testing employs models that consider binomial or Poisson data distributions (in contrast to normal distributions assumed by most microarray-dedicated procedures). Analysis workflows have been packaged for several analytic environments, including the open-source Bioconductor programming environment (Gentleman et al. 2004), Galaxy (Giardine et al. 2005), and MeV (Howe et al. 2011), in addition to several commercial packages.

### Postanalysis Considerations

Like all high-throughput, hypothesis-free studies, transcriptomic studies must be considered as hypothesis generating exercises, requiring confirmation, validation, and replication through a variety of means. Individual gene findings deemed of particular relevance should be confirmed by direct technical validation using single-gene based methods, like quantitative reverse transcription PCR. As described above, expression profiles or sample subgrouping derived by machine learning should be validated using both internal and external procedures. External validation of predictive signatures in independently ascertained populations remains the gold standard. However, to date, few studies have successfully met this mark.

## References

Alberts R, Terpstra P, Li Y et al (2007) Sequence polymorphisms cause many false cis eQTLs. PLoS One 2:e622

Alonzi T, Maritano D, Gorgoni B et al (2001) Essential role of STAT3 in the control of the acute-phase response as revealed by inducible gene inactivation [correction of activation] in the liver. [erratum appears in Mol Cell Biol 2001 Apr;21(8):2967]. Mol Cell Biol 21:1621–1632

Aoki T, Matsumoto Y, Hirata K et al (2009) Expression profiling of genes related to asthma exacerbations. Clin Exp Allergy 39:213–221

Baines KJ, Simpson JL, Bowden NA et al (2010) Differential gene expression and cytokine production from neutrophils in asthma phenotypes. Eur Respir J 35:522–531

Baines KJ, Simpson JL, Wood LG et al (2011) Transcriptional phenotypes of asthma defined by gene expression profiling of induced sputum samples. J Allergy Clin Immunol 127:153–160, 60.e1-9

Barabasi AL (2009) Scale-free networks: a decade and beyond. Science 325:412–413

Benito M, Parker J, Du Q et al (2004) Adjustment of systematic microarray data biases. Bioinformatics 20:105–114

Bjornsdottir US, Holgate ST, Reddy PS et al (2011) Pathways activated during human asthma exacerbation as revealed by gene expression patterns in blood. PLoS One 6:e21902

Bochkov YA, Hanson KM, Keles S et al (2010) Rhinovirus-induced modulation of gene expression in bronchial epithelial cells from subjects with asthma. Mucosal Immunol 3:69–80

Bolstad BM, Irizarry RA, Astrand M et al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162:156–159
Choy DF, Modrek B, Abbas AR et al (2011) Gene expression patterns of Th2 inflammation and intercellular communication in asthmatic airways. J Immunol 186:1861–1869
Chu JH, Weiss ST, Carey VJ et al (2009) A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. BMC Syst Biol 3:55
Chu JH, Lazarus R, Carey VJ et al (2011) Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. BMC Syst Biol 5:89
Dehmer M, Basak SC (2012) Statistical and machine learning approaches for network analysis. Hoboken, N.J.: Wiley
Fare TL, Coffey EM, Dai H et al (2003) Effects of atmospheric ozone on microarray data quality. Anal Chem 75:4672–4675
Freishtat RJ, Benton AS, Watson AM et al (2009) Delineation of a gene network underlying the pulmonary response to oxidative stress in asthma. J Investig Med 57:756–764
Gentleman RC, Carey VJ, Bates DM et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80
Giardine B, Riemer C, Hardison RC et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451–1455
Hakonarson H, Bjornsdottir US, Halapi E et al (2005) Profiling of genes expressed in peripheral blood mononuclear cells predicts glucocorticoid sensitivity in asthma patients. Proc Natl Acad Sci USA 102:14789–14794
Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13:204–216
Howe EA, Sinha R, Schlauch D et al (2011) RNA-Seq analysis in MeV. Bioinformatics 27:3209–3210
Hunninghake GM, Chu JH, Sharma SS et al (2011) The CD4+ T-cell transcriptome and serum IgE in asthma: IL17RB and the role of sex. BMC Pulm Med 11:17
Hwang S, Son SW, Kim SC et al (2008) A protein interaction network associated with asthma. J Theor Biol 252:722–731
Hyduke DR, Palsson BO (2010) Towards genome-scale signalling network reconstructions. Nat Rev Genet 11:297–307
Inza I, Calvo B, Armananzas R et al (2010) Machine learning: an indispensable tool in bioinformatics. Methods Mol Biol 593:25–48
Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264
Irizarry RA, Warren D, Spencer F et al (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2:345–350
Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127
Kapitein B, Hoekstra MO, Nijhuis EH et al (2008) Gene expression in CD4+ T-cells reflects heterogeneity in infant wheezing phenotypes. Eur Respir J 32:1203–1212
Kerr MK (2003) Design considerations for efficient and effective microarray studies. Biometrics 59:822–828
Kerr G, Ruskin HJ, Crane M et al (2008) Techniques for clustering gene expression data. Comput Biol Med 38:283–293
Kicic A, Hallstrand TS, Sutanto EN et al (2010) Decreased fibronectin production significantly contributes to dysregulated repair of asthmatic epithelium. Am J Respir Crit Care Med 181:889–898
Larranaga P, Calvo B, Santana R et al (2006) Machine learning in bioinformatics. Brief Bioinform 7:86–112
Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3:1724–1735
Lim WK, Wang K, Lefebvre C et al (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics 23:i282–i288

Luo J, Schumacher M, Scherer A et al (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J 10:278–291
Madore AM, Perron S, Turmel V et al (2010) Alveolar macrophages in allergic asthma: an expression signature characterized by heat shock protein pathways. Hum Immunol 71:144–150
Marioni JC, Mason CE, Mane SM et al (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517
Matson RS (2009) Microarray methods and protocols. Boca Raton: CRC Press
Melen E, Kho AT, Sharma S et al (2011) Expression analysis of asthma candidate genes during human and murine lung development. Respir Res 12:86
Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46
Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
Murphy A, Chu JH, Xu M et al (2010) Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. Hum Mol Genet 19:4745–4757
Orsmark-Pietras C, James A, Konradsen JR et al (2013) Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. Eur Respir J 42:65–78
Pickrell JK, Marioni JC, Pai AA et al (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772
Rajeevan MS, Dimulescu IM, Vernon SD et al (2003) Global amplification of sense RNA: a novel method to replicate and archive mRNA for gene expression analysis. Genomics 82:491–497
Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM (2012) Bioinformatics for high throughput sequencing. New York, NY: Springer
Schafer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21:754–764
Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
Scherer A (2009) Batch effects and noise in microarray experiments: sources and solutions. Wiley, Chichester, U.K
Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18
Shi L, Campbell G, Jones WD et al (2010) The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 28:827–838
Shin SW, Oh TJ, Park SM et al (2011) Asthma-predictive genetic markers in gene expression profiling of peripheral blood mononuclear cells. Allergy Asthma Immunol Res 3:265–272
Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York, NY, pp 397–420
Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545–15550
Subrata LS, Bizzintino J, Mamessier E et al (2009) Interactions between innate antiviral and atopic immunoinflammatory pathways precipitate and sustain asthma exacerbations in children. J Immunol 183:2793–2800
Sutcliffe A, Hollins F, Gomez E et al (2012) Increased nicotinamide adenine dinucleotide phosphate oxidase 4 expression mediates intrinsic airway smooth muscle hypercontractility in asthma. Am J Respir Crit Care Med 185:267–274
Tsitsiou E, Williams AE, Moschos SA et al (2012) Transcriptome analysis shows activation of circulating CD8+ T cells in patients with severe asthma. J Allergy Clin Immunol 129:95–103
Vidal M, Cusick ME, Barabasi AL (2011) Interactome networks and human disease. Cell 144:986–998
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5:19–21

Woodruff PG, Boushey HA, Dolganov GM et al (2007) Genome-wide profiling identifies epithelial cell genes associated with asthma and with treatment response to corticosteroids. Proc Natl Acad Sci USA 104:15858–15863

Woodruff PG, Modrek B, Choy DF et al (2009) T-helper type 2-driven inflammation defines major subphenotypes of asthma. Am J Respir Crit Care Med 180:388–395

Yakovlev AY, Klebanov L, Gaile D (2013) Statistical methods for microarray data analysis: Methods and Protocols. New York, NY: Springer New York

Yick CY, Zwinderman AH, Kunst PW et al (2013) Transcriptome sequencing (RNA-Seq) of human endobronchial biopsies: asthma versus controls. Eur Respir J, in press

Youssef LA, Schuyler M, Gilmartin L et al (2007) Histamine release from the basophils of control and asthmatic subjects and a comparison of gene expression between "releaser" and "nonreleaser" basophils. J Immunol 178:4584–4594