Multiplex meta-analysis of RNA expression to identify genes with variants associated with immune dysfunction

Alexander A Morgan,[1] Vasilios J Pyrgos,[2] Kari C Nadeau,[3] Peter R Williamson,[2] Atul Janardhan Butte[1]

[1]Biomedical Informatics Graduate Training Program and Division of Systems Medicine, Department of Pediatrics, Stanford University, Stanford, California, USA
[2]Laboratory of Clinical Infectious Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA
[3]Division of Immunology, Department of Pediatrics, Stanford University, Stanford, California, USA

Correspondence to
Dr Atul Janardhan Butte, Division of Systems Medicine, Department of Pediatrics, Stanford University, 1265 Welch Road X-163 MS-5415, CA 94305-5415, USA; abutte@stanford.edu

Received 22 October 2011
Accepted 29 December 2011

ABSTRACT

Objective We demonstrate a genome-wide method for the integration of many studies of gene expression of phenotypically similar disease processes, a method of multiplex meta-analysis. We use immune dysfunction as an example disease process.

Design We use a heterogeneous collection of datasets across human and mice samples from a range of tissues and different forms of immunodeficiency. We developed a method integrating Tibshirani's modified t-test (SAM) is used to interrogate differential expression within a study and Fisher's method for omnibus meta-analysis to identify differentially expressed genes across studies. The ability of this overall gene expression profile to prioritize disease associated genes is evaluated by comparing against the results of a recent genome wide association study for common variable immunodeficiency (CVID).

Results Our approach is able to prioritize genes associated with immunodeficiency in general (area under the ROC curve = 0.713) and CVID in particular (area under the ROC curve = 0.643).

Conclusions This approach may be used to investigate a larger range of failures of the immune system. Our method may be extended to other disease processes, using RNA levels to prioritize genes likely to contain disease associated DNA variants.

INTRODUCTION

One of the major goals of translational bioinformatics is to equip clinical medicine with the ability to use information about a patient's genome for diagnosis and decision-making. Several commercial companies provide disease risk information according to individual genotype,[1] and other approaches have been developed and used to analyze and interpret high-depth patient sequence data to guide treatment[2 3]; these are all instances of personalized medicine approaches that integrate genomics. Genotyping using arrays is a well-established commercial service, and the technology exists to provide high-depth sequencing and coverage at a cost comparable to many commonly used diagnostic tests[4]; basic techniques exist for using these data in a medically relevant fashion.[5 6] However, genomic medicine relies on knowledge of the genetic basis of disease, and although methods such as genome-wide association studies using genotyping arrays have become the gold standard for discovering and exploring genetic variations, they have had a relatively poor success rate in explaining the chief genetic contributions to the heritability of many major, common diseases; this is known as the 'missing heritability' problem.[7–9] We need additional tools to accelerate the process of uncovering the causal variations giving rise to pathology.[10] Unfortunately, targeted candidate gene association studies have had a notoriously poor rate of replication in contrast to the much less biased genome-wide approaches.[11 12] An approach that can prioritize targeted portions of the genome implicated in disease association in an unbiased, genome-wide manner, based on data-driven, functional properties would provide a powerful tool for future medical genomics. In this paper, we describe such a method and demonstrate its ability to prioritize the genes with variants implicated in a specific form of immunodeficiency, 'common variable immunodeficiency' (CVID), characterized by patient inability to produce sufficient antibodies.

Many common, multifactorial diseases include an infectious, autoimmune or inflammatory component; the mammalian immune response is a very finely tuned, highly complex system with hundreds of signaling molecules, dozens of different cell types, and the involvement of multiple tissue types and organs.[13] It features all the motifs and elements of the most complex biological circuits and control systems, with many interacting feedback and feed-forward elements.[14] Dysfunction can arise from variation in many different components of this complex, highly inter-connected system, and the phenotypic changes in the immune system to the range of genomic variations possible is only beginning to be understood. However, characterization of the variations that lead to serious immune failure can help in the treatment of affected patients and also, hopefully, provide insight into the range of human immune response as influenced by genetics.

Although large knowledge bases on immune function have been constructed,[15] we also have access to genome-wide functional data in the form of gene expression information. Large repositories like the NIH NCBI Gene Expression Omnibus[16] provide access to tens of thousands of different highly parallelized gene expression measurements. We and others have previously described methods that look across multiple studies which each provide many gene expression measurements (ie, multiplex) in what we call 'multiplex meta-analysis'[17–20] to obtain an overall picture of gene expression across studies. Taking advantage of the central dogma that DNA codes for RNA, which codes for protein, studying RNA at the gene expression level in specific phenotypes has provided potential insight into

functional processes in the phenotype and suggests genes (DNA) that may contain variants associated with that disease.[21 22] In this study, we extend this idea and describe a method for integrating gene expression data across multiple gene expression studies for clinically/phenotypically similar presentations of a range of immune deficiencies across species and tissue types to create an integrated multiplex (parallelized) meta-profile of gene expression. This meta-expression profile allows powerful prioritization of the genes involved in general immune deficiency but also shows predictive power over a recent genome-wide association study of CVID (figure 1). We suggest that this method could be used to investigate the genetic and molecular pathology of other forms of immune dysfunction.

METHODS

We collected data from the NIH NCBI Gene Expression Omnibus for 16 different studies of immunodeficiency (table 1). Importantly, these gene expression measurements are from many different forms of immune dysfunction and span multiple species and different tissue types. The gene expression samples in each study were hand annotated and further divided into 37 different experimental comparison subgroups (such as different mouse backgrounds used in the same study). Gene expression levels were compared between immune deficient samples and controls (normal immune function samples) in each subgroup using the modified t test proposed by Tusher *et al*[37] and incorporated into the significance analysis of microarrays.[38] The annotations of all samples are available upon request.

We calculate the mean log fold change[38] of each oligonucleotide in each array (m) in each experimental class (k, either i for immunodeficiency or c for control) and each subgroup (g), with $n_{k,g}$ indicating the total number of arrays of class k in subgroup g.

$$\bar{x}_{k,g} = \frac{\sum_{m=1}^{n_{k,g}} \log(x_{m,k,g})}{n_{k,g}} \quad (1)$$

These mean log fold changes are then used in the modified t test (equation 2) introduced by Tusher *et al*[37] and explained in detail in Witten and Tibshirani,[38] using the standard deviation (σ_g) and a value $s_{o,g}$ (scaling factor) selected to minimize the coefficient of variation of T_g across all oligonucleotides.

$$T_g = \frac{\bar{x}_{i,g} - \bar{x}_{c,g}}{\sigma_g + s_{o,g}} \quad (2)$$

The modified t statistic T_g is then bootstrapped to calculate a p value (p_g) for each gene. The oligonucleotides on the array are mapped to genes using AILUN[39] and across species using HomoloGene groups.[40] Using a method proposed by Fisher[41] and based on the fact that p values selected at random should be uniformly distributed, we can look for deviations by the χ^2 test (equation 3), with p_g the p value for that gene in the subgroup g and n_g the number of subgroups measuring that gene.

$$\chi^2_{2n_g} = -2\sum_{g=1}^{n_g} \log(p_g) \quad (3)$$

Importantly, this method will freely mix different directions of variation across studies, up or down, using only the significance of the test. The final meta p values may then be used to rank the significance of expression differences between immune deficient and normal controls across studies. We predict that genes that are significantly differentially expressed across studies identified through our method are more likely to be involved in immunodeficiency; the lower the p value, the greater the priority given.

To establish a baseline and identify genes involved with the immune system in general, we also took lists of genes annotated for involvement in acquired immunity, innate immunity, and inflammation taken from the Molecular Signatures Database.[42] The genes with an annotation in each respective category were taken as predictions for having variants associated with disease risk, and those genes lacking annotations were taken to be predictions for those genes not being closely coupled to variants associated with increased disease risk. For example, one strategy to prioritize CVID related genes would be to first investigate all those genes annotated for being involved in the biological process of 'acquired immunity', using the presence of that

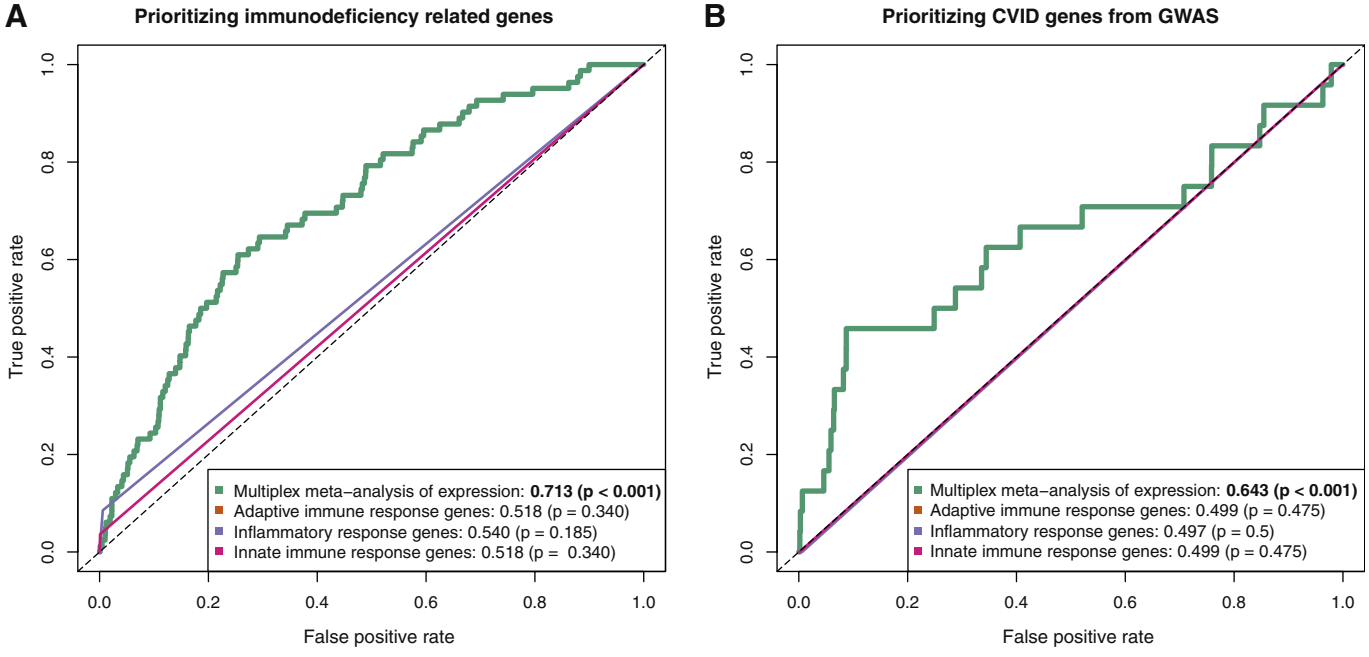

Figure 1 Receiver Operating Characteristic (ROC) curves evaluating the prioritization of disease associated genes.

Table 1 Publicly-available gene expression experimental data used in multiplex meta-analysis

GSE: GEO series accession number	Control samples	Immune deficient samples	Title	Reference
GSE10817	2	2	Mll5 is required for hematopoietic stem cell fitness and homeostasis	23
GSE11005	34	20	Immune responses to pneumocystis infection are robust in immunocompetent mice but absent in CD40 ligand deficient mice	24
GSE12464	13	6	Transcriptional signatures of Itk-deficiency using CD3+ T cells	25
GSE12465	4	5	Transcriptional signatures of Itk-deficiency using CD3+, CD4+, and CD8+ T cells	25
GSE15324	4	4	Control of CD8+ T cell proliferation by the transcription factor ELF4	26
GSE15750	8	8	Enhancing CD8 T cell memory by modulating fatty acid metabolism	27
GSE2585	4	4	Promiscuous gene expression in the mouse thymus	28
GSE3414	18	18	Immune response to *Nippostrongylus brasiliensis* in the mouse lung	29
GSE3676	2	2	Expression profile of the testis from Tslc1−/− mice.	30
GSE5654	3	3	Essential role of Jun family transcription factors in PU.1-induced leukemic stem cells	31
GSE85	3	3	Wild type and Aire −/− murine medullary thymic epithelial cells	32
GSE8507	35	35	Neutrophil and PBMC gene expression data from Job's syndrome	33
GSE8564	10	10	Analysis of Aire effects on individual mice of different genetic backgrounds	34
GSE8726	4	3	Expression data from Sod2−/− and Sod2+/+ mouse erythroblasts	Unpublished from FM Martin *et al* at The Scripps Research Institute
GSE935	8	12	NIH/NIAID chronic granulomatous disease neutrophils	35
GSE9499	30	15	DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome: gene expression analysis	36

GEO, NCBI Gene Expression Omnibus.

annotation on a gene as the predictor for involvement with CVID.

To evaluate our ability to prioritize genes with variants associated with immunodeficiency, we compiled a manually expertly curated list of 131 genes known to have variations associated with immunodeficiency in general. However, as that list might be biased toward genes discovered through differential expression, we also used the results of a recent genome-wide association study of 363 patients with CVID and 3031 healthy controls. We took the results of this study and identified 31 genes linked to variations associated with CVID,[43] prominent among them the genes of the major histocompatibility complex (MHC).

We evaluated the ability to prioritize the genes with variants implicated in immunodeficiency using the receiver operating characteristic (ROC) curve. The area under the ROC curve summarizes the power of a prediction method into a single numerical value, in this case predicting the association of genes likely to contain disease-associated variants. An area of 0.5, or a curve along the diagonal indicates no predictive power. A larger area under the curve (AUC) implies better discrimination ability. We obtain a p value by comparing the observed AUC against 1000 randomizations.

RESULTS

In figure 1, we show the ability of our multiplex meta-analysis derived gene expression profile for immunodeficiency to identify and prioritize genes with disease-associated variants. In panel A of figure 1, we see that our approach provides an AUC of 0.713, much better than a random sampling of the genome (diagonal) or selecting and prioritizing genes annotated for immune system involvement (other solid lines).

In figure 1B, we can see very strong prioritization of genes identified through an unbiased, genome-wide investigation of the genetic causes of CVID. Even though CVID represents a collection of different, related syndromes of dysfunction in antibody production, our gene expression analysis drawn from an even more heterogeneous collection of diseases provides predictive power with an AUC of 0.643.

The top 20 genes from our gene expression meta-analysis for immunodeficiency are shown in table 2. Our approach suggests that these genes should be further investigated in relation to their role in immune dysfunction. Although many of these genes have been poorly characterized for function, some are already known to have an important role in immunity, suggesting that although looking just at genes annotated for immunity does not really help selection, some key immune genes have nonetheless been found using microarray meta-analysis, such as *PECAM1* which codes for CD31 (a protein involved in leukocyte migration) and *STAT1* which is a very important immune regulator involved in response to signaling by interferons.

CONCLUSION

Our hypothesis in this study was that by integrating RNA repeatedly implicated in gene expression microarray experiments related to immunology, we could identify genes (DNA) recognized to contain variants or mutations well known to be associated with immune dysfunction. To test this, we compared sets of genes found through meta-analysis with genes selected using prior knowledge of immunology pathways, comparing against

Table 2 Top differentially expressed genes across immunodeficiency

Rank	Symbol	Multiplex meta-analysis p value	Rank	Symbol	Multiplex meta-analysis p value
1	FNIP2	7.87E-72	11	MS4A1	1.31E-44
2	PDE4DIP	7.77E-70	12	TFRC	9.33E-43
3	KLK2	8.40E-69	13	ZFYVE16	4.44E-41
4	SFI1	4.18E-64	14	MAP4K4	9.90E-41
5	SETDB2	2.36E-63	15	GAS7	7.97E-40
6	KLHL6	5.76E-49	16	GADD45B	9.01E-40
7	ZADH2	2.16E-47	17	PECAM1	8.76E-39
8	SLAMF7	4.26E-47	18	ST8SIA4	2.39E-38
9	PPAP2B	1.72E-46	19	LGALS3	8.10E-38
10	RSAD2	7.28E-46	20	STAT1	4.21E-37

a manually curated gold standard list of genes known to have variants associated with immunodeficiency.

The important result here is that prior annotations of immune function actually provide no particular insight, as the resulting curves diverge very little from the diagonal ($\sim$0.5 in both panels of figure 1). Our study of genes by their annotation for involvement in immune processes is analogous to previously commonly used candidate gene approaches; genes which were believed to be involved in a pathway or process involved in the phenotype of interest (such as the disease) were examined for genetic variation enriched in affected individuals relative to healthy controls. Although such knowledge-driven approaches have suggested many genes and variants that might influence disease risk, the variants found by such approaches have typically failed to be replicated in larger or subsequent studies.[12 44] Our findings lead to a similar result. The previous knowledge-based approach provides little value in prioritization.

There are many potential biases in the identification of genes with disease-associated variation, and it has been suggested that the best ways to address this are to interrogate the genome as widely as possible and to combine studies whenever possible to demonstrate reproducibility and consistency of effect.[45–48] In statistics, previous beliefs that lead to bias may be cast in a more formal way, for example as a Bayesian prior, which is a type of statistical bias. Other types of biases may be the result of sample collection having a particular structure that skews the results, such as patients and controls in a clinical trial selected in a non-random way. Overall, bias is not meant here as a pejorative term, merely descriptive. Selecting genes using previous knowledge would represent a form of bias. We suggest that our meta-analysis approach is relatively unbiased compared to using annotations of individual genes, in that it relies on the single assumption that the datasets synthesized relate to the phenotype under study, and is not biased on particular genes or pathways having been previously studied to a greater extent than others.

In our results, investigating this type of immune disease using biased, prior knowledge of the molecular genetics of immune function provides little advantage over a random selection of genes. A study looking for genetic association only in the genes annotated for immune function, would apparently not do much better than random. However, our data-driven approach offers a much better way to prioritize genes for genetic investigation. The meta-analysis derived expression profiles preferentially select genes with disease-associated variants without making any assumptions of function other than differential expression in a related disease.

The vastness of the human genome provides a huge challenge as we seek to investigate the genetic underpinnings of disease; however, we want to interrogate not only a huge range of variants at different loci but also systems with potentially many interacting components. Tackling such a task will require integration over many different experimental modalities. We have demonstrated a method for using raw functional data in the form of multiplexed gene expression to propose genes associated with disease. Importantly, our results shown in figure 1 suggest that this unbiased, data-driven approach is superior to using highly biased, functional annotations.

Although our method may be of particular use in further investigations of immunodeficiency and CVID in particular, it does not need to be confined to immunology. Our approach may work in other disease domains. Immune dysfunction is just one example of many disease phenotypes that derive from the interaction of many genes, proteins, and environmental factors. Future extensions may include information on sequence conservation[10] or direct interactions between genes and proteins.[49]

Acknowledgments Special thanks to Rong Chen and Alex Skrenchuk for the development of database resources.

Funding This work was supported in part by the March of Dimes, the Lucile Packard Foundation for Children's Health, the Hewlett Packard Foundation, and the National Library of Medicine through direct research funding (R01 LM009719) and a Biomedical Informatics training grant (T15 LM007033). This research was also supported, in part, by the Intramural Research Program of the NIH, NIAID.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Ng PC,** Murray SS, Levy S, *et al*. An agenda for personalized medicine. *Nature* 2009;**461**:724–6.
2. **Ashley EA,** Butte AJ, Wheeler MT, *et al*. Clinical assessment incorporating a personal genome. *Lancet* 2010;**375**:1525–35.
3. **Worthey EA,** Mayer AN, Syverson GD, *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;**13**:255–62.
4. **Pushkarev D,** Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;**27**:847–50.
5. **Morgan AA,** Chen R, Butte AJ. Likelihood ratios for genome medicine. *Genome Med* 2010;**2**:30.
6. **Stern S,** Cifu A, Altkorn D. *Symptom to Diagnosis: An Evidence-Based Guide*. 2nd edn. San Francisco: Lange Medical, 2010.
7. **Hemminki K,** Forsti A, Houlston R, *et al*. Searching for the missing heritability of complex diseases. *Hum Mutat* 2011;**32**:259–62.
8. **Maher B.** Personal genomes: the case of the missing heritability. *Nature* 2008;**456**:18–21.
9. **Manolio TA,** Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
10. **Cooper GM,** Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**:628–40.
11. **Chanock SJ,** Manolio T, Boehnke M, *et al*. Replicating genotype-phenotype associations. *Nature* 2007;**447**:655–60.
12. **Ioannidis JP,** Ntzani EE, Trikalinos TA, *et al*. Replication validity of genetic association studies. *Nat Genet* 2001;**29**:306–9.
13. **Murphy KM,** Travers P, Walport M. *Janeway's Immunobiology*. 7th edn. London, UK: Garland Science, 2007.
14. **Alon U.** *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca Raton, FL: Chapman and Hall/CRC, 2006.
15. **Shen-Orr SS,** Goldberger O, Garten Y, *et al*. Towards a cytokine-cell interaction knowledgebase of the adaptive immune system. *Pac Symp Biocomput* 2009:439–50.
16. **Barrett T,** Troup DB, Wilhite SE, *et al*. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acid Res* 2007;**35**(Database issue):D760–5.
17. **Campain A,** Yang YH. Comparison study of microarray meta-analysis methods. *BMC Bioinformatics* 2010;**11**:408.
18. **Chen R,** Sigdel TK, Li L, *et al*. Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions. *PLoS Comput Biol* 2010;**6**.pii:e1000940.
19. **Morgan AA,** Khatri P, Jones RH, *et al*. Comparison of Multiplex Meta-analysis Techniques for Understanding the Acute Rejection of Solid Organ Transplants. *BMC Bioinformatics* 2010;**11**(Suppl 9):S6.
20. **Ramasamy A,** Mondry A, Holmes CC, *et al*. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 2008;**5**:e184.

21. **English SB,** Butte AJ. Evaluation and integration of 49 genome-wide experiments and the prediction of previously Unknown Obesity-related Genes. *Bioinformatics* 2007;**23**:2910—17.
22. **Mootha VK,** Lepage P, Miller K, *et al*. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* 2003;**100**:605—10.
23. **Zhang Y,** Wong J, Klinger M, *et al*. MLL5 contributes to hematopoietic stem cell fitness and homeostasis. *Blood* 2009;**113**:1455—63.
24. **Hernandez-Novoa B,** Bishop L, Logun C, *et al*. Immune responses to Pneumocystis murina are robust in healthy mice but largely absent in CD40 ligand-deficient mice. *J Leukoc Biol* 2008;**84**:420—30.
25. **Blomberg KE,** Boucheron N, Lindvall JM, *et al*. Transcriptional signatures of Itk-deficient CD3+, CD4+ and CD8+ T-cells. *BMC Genomics* 2009;**10**:233.
26. **Yamada T,** Park CS, Mamonkin M, *et al*. Transcription factor ELF4 controls the proliferation and homing of CD8+ T cells via the Krüppel-like factors KLF4 and KLF2. *Nat Immunol* 2009;**10**:618—26.
27. **Pearce EL,** Walsh MC, Cejas PJ, *et al*. Enhancing CD8 T-cell memory by modulating fatty acid metabolism. *Nature* 2009;**460**:103—7.
28. **Derbinski J,** Gäbler J, Brors B, *et al*. Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J Exp Med* 2005;**202**:33—45.
29. **Reece JJ,** Siracusa MC, Scott AL. Innate immune responses to lung-stage helminth infection induce alternatively activated alveolar macrophages. *Infect Immun* 2006;**74**:4970—81.
30. **Yamada D,** Yoshida M, Williams YN, *et al*. Disruption of spermatogenic cell adhesion and male infertility in mice lacking TSLC1/IGSF4, an immunoglobulin superfamily cell adhesion molecule. *Mol Cell Biol* 2006;**26**:3610—24.
31. **Steidl U,** Rosenbauer F, Verhaak RG, *et al*. Essential role of Jun family transcription factors in PU.1 knockdown-induced leukemic stem cells. *Nat Genet* 2006;**38**:1269—77.
32. **Anderson MS,** Venanzi ES, Klein L, *et al*. Projection of an immunological self shadow within the thymus by the aire protein. *Science* 2002;**298**:1395—401.
33. **Holland SM,** DeLeo FR, Elloumi HZ, *et al*. STAT3 mutations in the hyper-IgE syndrome. *N Engl J Med* 2007;**357**:1608—19.
34. **Venanzi ES,** Melamed R, Mathis D, *et al*. The variable immunological self: genetic variation and nongenetic noise in Aire-regulated transcription. *Proc Natl Acad Sci U S A* 2008;**105**:15860—5.
35. **Kobayashi SD,** Voyich JM, Braughton KR, *et al*. Gene expression profiling provides insight into the pathophysiology of chronic granulomatous disease. *J Immunol* 2004;**172**:636—43.
36. **Jin B,** Tao Q, Peng J, *et al*. DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Hum Mol Genet* 2008;**17**:690—709.
37. **Tusher VG,** Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;**98**:5116—21.
38. **Witten D,** Tibshirani R. *A Comparison of Fold-change and the T-statistic for Microarray Data Analysis*. Stanford, CA: Stanford University, 2007.
39. **Chen R,** Li L, Butte AJ. AILUN: reannotating gene expression data automatically. *Nat Methods* 2007;**4**:879.
40. **Wheeler DL,** Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology information. *Nucleic acids Res* 2007;**35**(Database issue):D5—12.
41. **Hedges L,** Olkin I. *Stat Methods for Meta-analysis*. Academic Press, New York, 1985.
42. **Subramanian A,** Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545—50.
43. **Orange JS,** Glessner JT, Resnick E, *et al*. Genome-wide association identifies diverse causes of common variable immunodeficiency. *J Allergy Clin Immunol* 2011;**127**:1360—7.
44. **Ioannidis JP.** Why most published research findings are false. *PLoS Med* 2005;**2**: e124.
45. **Ioannidis JP.** Why most discovered true associations are inflated. *Epidemiology* 2008;**19**:640—8.
46. **Xiao R,** Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. *Genet Epidemiol* 2009;**33**:453—62.
47. **Zhong H,** Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* 2008;**9**:621—34.
48. **Zollner S,** Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 2007;**80**:605—15.
49. **Chen J,** Aronow B, Jegga A. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 2009;**10**:73.