

# A genomic view of immunology

Aude M. Fahrer\*, J. Fernando Bazan†, Peter Papathanasiou\*, Keats A. Nelms\* & Christopher C. Goodnow\*

\*ACRF Genetics Laboratory and Medical Genome Centre, John Curtin School of Medical Research, Australian National University, Canberra 2601, Australia

†Department of Molecular Biology, DNAX Research Institute, 901 California Avenue, Palo Alto, California 94304-1104, USA

**The outstanding problems facing immunology are whole system issues: curing allergic and autoimmune disease and developing vaccines to stimulate stronger immune responses against pathogenic organisms and cancer. We hope that the human genome sequence will reveal the molecular checks and balances that ensure both an effective immunogenic response against pathogenic microorganisms and a suitably tolerogenic response to self antigens and innocuous environmental antigens. Three synergistic approaches—sequence homology searches, messenger RNA expression profiling on microarrays, and mutagenesis in mice—provide the best opportunities to reveal, in the genome sequence, key proteins and pathways for targeting by new immunomodulatory treatments.**

**T**he human genome sequence contains a list of all the parts controlling beneficial and harmful immune responses. To identify these parts, the immediate hurdles of gene identification and genome assembly will need to be overcome. The ultimate challenge will be to bridge the gulf between a list of sequences and whole organism biology.

## Predicting immune genes from sequence homology

The most immediate need is a list of all genes, exons and control elements in the human genome, but our experience is that the public domain data are still some way from this goal. To investigate the usefulness of searching the genome sequence on the basis of predicted protein homology alone (Fig. 1), we chose three examples of immunologically important proteins and asked how many more protein relatives could be found. Only the first of these examples could be searched with publicly available web-based databases. For details and advice about search options and databases, see Supplementary Information.

First, we looked for proteins of the tumour necrosis factor receptor (TNFR) family, members of which control proliferation or apoptosis of lymphocytes. Sequence motifs for the TNFR family have been assembled, for example in Interpro (see Supplementary Information). The search identified 21 of the 22 known TNFR family members<sup>1</sup>, the exception being TNFR superfamily member 18. Five proteins did not immediately match known TNFR family proteins. One was the recently published TNFR family member TA $\alpha$ , and two ultimately corresponded to family members OPG and DR5. The two remaining proteins represented potentially novel family members: IGI\_M1\_ctg13384\_53 was similar to OX40, and IGI\_M1\_ctg13980\_78 was similar, but only over its amino-terminal half, to CD30. The fact that we found so few new members of the TNFR family may reflect either intense mining of expressed sequence tag (EST) databases or the incomplete assembly of raw genome data at present.

Second, we looked for members of the B7 family of costimulatory proteins. CD80 and CD86 (B7.1 and B7.2) are expressed on antigen-presenting cells. They engage CD28 and CD152 (CTLA-4) receptors on T cells to transmit either costimulatory or inhibitory signals to the T cell. Two homologues of CD80 and CD86 have recently been identified<sup>2</sup>: ICOSL, which binds to the T-cell costimulatory receptor ICOS (inducible costimulator); and B7H-1, which binds to the lymphocyte inhibitory receptor PD-1 (ref. 3). The B7 costimulatory proteins are not identified by specific motifs in the major protein databases, apart from being members of the immunoglobulin superfamily. Using PSI-BLAST searches (see Supplementary Information), we identified 21 proteins with homology to the B7 proteins.

Nine of these were members of the butyrophilin family.

Butyrophilin is a cell-surface protein, also found in breast milk, that has been identified as a B7 homologue<sup>4</sup>. Three of the proteins corresponded to signal regulatory protein (SIRP)- $\alpha$ -1 and one to SIRP- $\beta$ -1. SIRP- $\alpha$ -1 is an inhibitory receptor, expressed by splenic macrophages, which binds to the CD47 self marker on erythrocytes and prevents their elimination; SIRP- $\beta$ -1 seems to be involved in the activation of myeloid and dendritic cells.

We identified four other proteins corresponding to previously cloned genes: HHLA2, a human endogenous retrovirus sequence encoding a potentially secreted protein expressed in several tissues, including lymphocytes; MCAM, a melanoma adhesion protein apparently involved in tumour progression; CXADR, a surface molecule of unknown function; and VEJAM, a vascular endothelial junction-associated molecule, which may be involved in lymphocyte homing.

We also found four novel proteins (see Supplementary Information). Extrapolating from the known members, the less characterized proteins could be involved in the costimulation or modulation of lymphocytes, macrophages or other cell types.

The third homology search was performed on the basis of three-dimensional structure rather than amino-acid homology. Cytokines are an important class of immune regulators with protein folds that accommodate considerable sequence divergence<sup>5</sup>. Family relationships have often emerged only after the resolution of prototype cytokine folds; for example, cementing the distant similarity between fibroblast growth factors and interleukin-1 (IL-1)-like molecules<sup>6</sup>, or suggesting a link between TNF proteins and an extended family of complement C1q-like cytokines<sup>7</sup>. These findings often broaden our view of the evolution of cytokines as well as their biological functions. The challenge is to detect novel molecules with weak or unapparent ties to existing cytokine groups. We can do this best with computational techniques that are being used sensitively to annotate genome-derived sequences, and to fuse knowledge of the structural templates with sensitive sequence searching and prediction routines<sup>8</sup>.

We looked at the superfamily of haemopoietic cytokines, which is distinguished by a four-helix bundle fold that engages a special class of transmembrane receptor<sup>5</sup>. Although difficult to align by sequence similarity, the helical scaffolds of these cytokines reveal faint, subfamily-distinctive motifs when superimposed. Taking the best conserved 'D' helix alignment of a diverse series of IL-6-like cytokine structures (IL-6, granulocyte colony stimulating factor, ciliary neurotrophic factor (CNTF), oncostatin-M, leukaemia inhibitory factor and a carefully arrayed set of cardiotrophin-1, IL-11 and IL-12 sequences), we constructed weighted profiles, position-specific scoring matrices and hidden Markov models and used them iteratively to search both EST and genomic databases. These profiles collected all extant IL-6-type sequences, as well as a set of novel,

predicted open reading frames (ORFs) that were then used to clone their complete gene sequences. Among these orphan cytokines is a molecule that distantly resembles CNTF, called novel neurotrophin-1 or cardiotrophin-like cytokine—not surprisingly, this molecule signals by co-opting the CNTF receptor complex<sup>9</sup>.

Another outlier sequence resembles the p35 subunit of IL-12, and competes for binding to the p40 chain (creating a cytokine labelled as IL-23); it then binds to a receptor complex that includes elements of the IL-12 signalling machinery<sup>10</sup>. Thus, in cases where no sequence similarity is detectable—but secondary structure prediction indicates, for example, a compatible register of helices and loops—fold recognition or threading techniques can tease out a reliable alignment of a novel sequence with a helical cytokine template.

It is difficult to gauge how completely these searches scan the genome, because the genome sequence is not complete, and considerable work is still needed to identify correctly all protein-coding segments and to remove intronic sequences. It must also be remembered that to use protein homology queries at least one member of a family must already be known and must have a searchable motif.

### Predicting immunological proteins from expression pattern

Sequence homology alone is a very incomplete predictor of genes in the immunological parts list, as many will not have known homologues and those that are revealed by this approach may not be expressed in the immune system. Expression of mRNAs or proteins in lymphoid cells is a stronger, complementary predictor (Fig. 1), particularly when the expression pattern changes during lymphocyte development or activation. Already, compilations of known genes and ESTs have been used to link sets of genes with differences in signalling brought about by tolerance or immunosuppression<sup>11</sup>, and to link clusters of genes with different normal and neoplastic lymphoid cell types<sup>12</sup>. Alignment of the human and mouse genome will illuminate regions of conservation, allowing oligonucleotide-based microarrays to assay all putative exons for transcription into mRNA in lymphoid cell types. This strategy holds the most

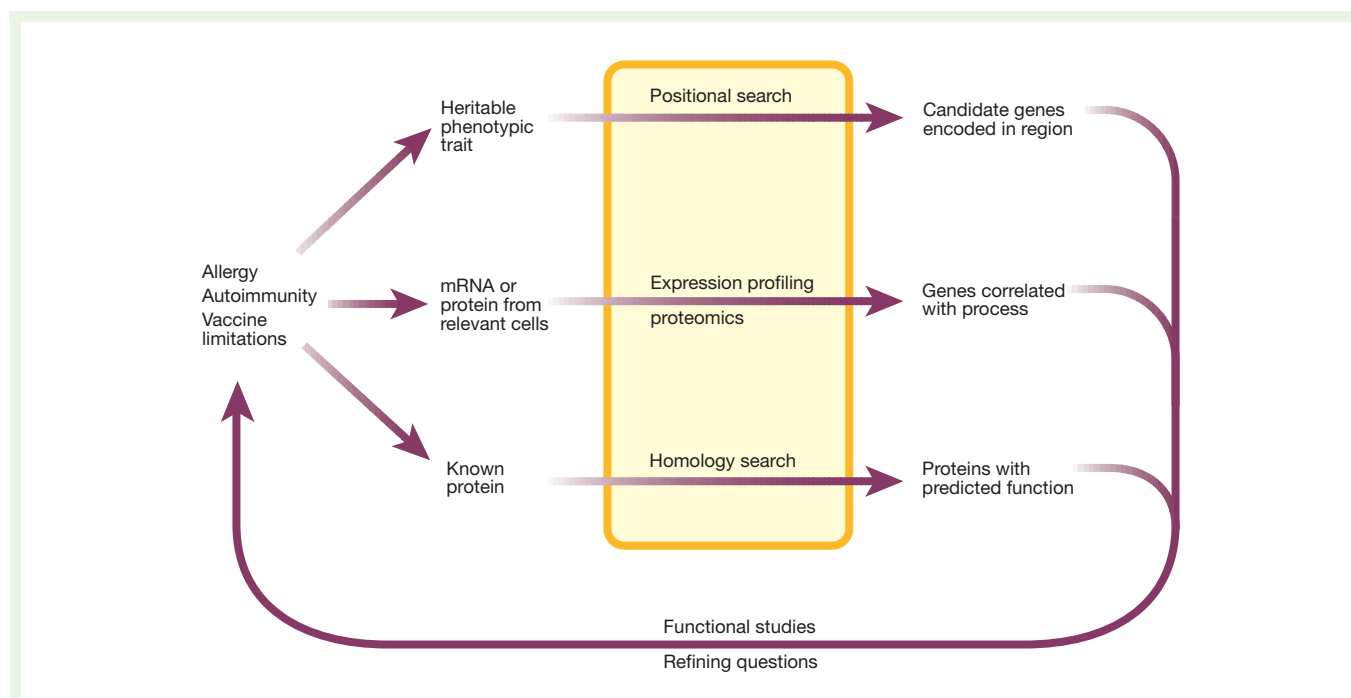
immediate promise for compiling a comprehensive parts list for the immune system, revealing differentially spliced forms as well as differences in overall mRNA abundance among immune cell subsets.

The genome sequence will also facilitate rapid identification of differentially expressed or modified proteins in immune cells or fluids by mass spectrometry and limited peptide microsequencing. This technique is limited by the need for better ways to resolve and quantify tens of thousands of proteins, which vary in abundance over many orders of magnitude.

### Mutations and polymorphisms link genes and function

Showing that a variant gene alters an immunological trait provides the most certain way to bridge the gap between genome sequence and immune responses at the organismal level (Fig. 1). As many new genes are implicated in the immune system by expression profile and sequence homology, the chief bottleneck lies in testing the immunological effects of mutant alleles of these genes. Important functional data can be obtained relatively rapidly by *in vitro* assays and *in vivo* blocking with antibodies or Fc fusion proteins. There are nevertheless many examples where the function of a gene for the immune system as a whole was inaccurately extrapolated from such assays, compared to the role that was revealed in mice with loss-of-function mutations.

It is reasonable to propose that a systematic effort be mounted to knock out every immunologically expressed gene. This could be done by homologous recombination, but at an estimated cost of US\$800 million, assuming that there are 20,000 immune genes. Lexicon Inc. have developed a large panel of embryonic stem cell lines carrying individual genes inactivated and tagged by retroviral insertional mutagenesis which can be searched via the web ([www.lexgen.com](http://www.lexgen.com)). The limitation of this method is that insertion is nonrandom. Of 67 B-cell activation genes identified by microarray profiling<sup>11</sup>, one-third had retroviral insertions. For those that were targeted, there were usually multiple separate lines, often with insertions in the same location.



**Figure 1** Impact of the human genome sequence on immunology. Three paths are shown for linking immune phenomena with individual genes; in each case, the human genome sequence has a central role.

In contrast to the 'reverse genetics' strategies above, which create mutants in predicted immunological genes using a bottom-up strategy, perhaps the greatest contribution made to immunology by the genome sequence lies in facilitating 'forward genetics' approaches. These start with an altered immunological trait and proceed from the top down to identify the causative gene. The trait may be a difference between strains of mice or a difference in human families. Research efforts can be focused on genes with key roles in immunological processes, regardless of sequence or expression pattern. A clear example is the identification of the mutated gene, AIRE, responsible for a mendelian syndrome of endocrine autoimmunity and candida infections (APECED)<sup>13,14</sup>. The genome sequence allowed identification of candidate genes within the chromosomal interval to which the APECED mutation had been mapped. Resequencing of these candidates from affected individuals revealed the mutant AIRE gene. Our experience is that public genome databases have not quite reached the goal of laying out all candidate genes between two markers. Using a test search for an interval on chromosome 21, not all known markers could be located on the chromosome map, and there were gaps in gene annotation (see Supplementary Information).

With the mouse genome sequence to be completed in 2001, phenotype-driven gene identification strategies can now be performed on a large scale by genome-wide mutagenesis in the mouse. Use of the supermutagen ethylnitrosourea (ENU) to induce point mutations in a large fraction of genes, coupled with sensitive screens for immunological traits, is allowing identification of many key regulators of the immune system<sup>15,16</sup>.

The complete list of parts provided by the human genome sequence will, on its own, not solve the major questions facing immunologists. Rather, it will form the basis of experiments designed to understand how the parts fit together to control immune responses. □

1. Screaton, G. & Xu, X. N. T cell life and death signalling via TNF-receptor family members. *Curr. Opin. Immunol.* **12**, 316–322 (2000).
2. Mueller, D. L. T cells: A proliferation of costimulatory molecules. *Curr. Biol.* **10**, R227–230 (2000).
3. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).
4. Henry, J., Miller, M. M. & Pontarotti, P. Structure and evolution of the extended B7 family. *Immunol. Today* **20**, 285–288 (1999).
5. Sprang, S. R. & Bazan, J. F. Cytokine structural taxonomy and mechanisms of receptor engagement. *Curr. Opin. Struct. Biol.* **3**, 815–827 (1993).
6. Zhang, J. D., Cousens, L. S., Barr, P. J. & Sprang, S. R. Three-dimensional structure of human basic fibroblast growth factor, a structural homolog of interleukin 1 beta. [published erratum appears in *Proc. Natl Acad. Sci. USA* **88**, 5477 (1991)]. *Proc. Natl Acad. Sci. USA* **88**, 3446–3450 (1991).
7. Shapiro, L. & Scherer, P. E. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* **8**, 335–338 (1998).
8. Fischer, D. & Eisenberg, D. Predicting structures for genome proteins. *Curr. Opin. Struct. Biol.* **9**, 208–211 (1999).
9. Elson, G. C. *et al.* CLF associates with CLC to form a functional heteromeric ligand for the CNTF receptor complex. *Nature Neurosci.* **3**, 867–872 (2000).
10. Oppmann, B., Bazan, J. F. & Kastelein, R. A. IL-23. *Immunity* (in the press).
11. Glynne, R. *et al.* How self-tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature* **403**, 672–676 (2000).
12. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
13. The Finnish-German APECED Consortium. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genet.* **17**, 399–403 (1997).
14. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17**, 393–398 (1997).
15. Goodnow, C. C. *et al.* Mechanisms of self-tolerance and autoimmunity: From whole-animal phenotypes to molecular pathways. *Cold Spring Harb. Symp. Quant. Biol.* **64**, 313–322 (1999).
16. Justice, M. J. Capitalizing on large-scale mouse mutagenesis screens. *Nature Rev. Genet.* **1**, 109–115 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

#### Acknowledgements

The authors would like to thank Jason Bock and Greg Quinn for invaluable help with computer searches, and Iain Wilson for helpful discussions.

Correspondence and requests for materials should be addressed to A.M.F. (e-mail: [aude.fahrer@anu.edu.au](mailto:aude.fahrer@anu.edu.au)).