

# The human genome: an immuno-centric view of evolutionary strategies

Yin Liu and Stephen Shaw



A hallmark of modern biology is the realization of the fundamental unity of biological processes in all life forms. Consequently, the complete genome sequencing of various bacteria, yeast (*Saccharomyces cerevisiae*), fly (*Drosophila melanogaster*) and worm (*Caenorhabditis elegans*) over the past five years has already had an impact on all of biology. 'Model organisms' have contributed a great deal to immunology; for example, the Toll receptors of the fly provided the impetus for the investigation of Toll-like receptors, which proved to be fundamental elements in the mammalian innate immune system. The recent release of a draft sequence of the human genome provides the first panoramic view of the 30 000–35 000 human genes in the human genetic blueprint and provides a plethora of new details, the significance of which will take some time to appreciate. The over-riding concepts that emerge from these studies relate primarily to general evolutionary processes that are equally as relevant to immunology as they are to other disciplines of biology.

As the adaptive immune system is one of the major biological processes unique to vertebrates, analysis of the first vertebrate genome might be expected to provide revelations of particular relevance to immunologists. Although it does provide additional insights, these insights are affirmations and extensions of our previous understanding rather than fundamentally new observations. Here, we highlight conserved domain evolution as a useful perspective for beginning to appreciate and integrate this new information.

## Evolution of conserved domains

The concept of conserved evolutionary domains and their use in evolution plays a prominent role in the reports from the public consortium and the Celera project<sup>1,2</sup>. For example, the Ig domain<sup>3–5</sup> has evolved from a 'bit player' in simple multicellular organisms to a 'superstar' in

vertebrates. Although the importance of the Ig domain comes as no surprise to immunologists, it is instructive to review the emerging understanding of its participation in vertebrate evolution.

Biological evolution is a modular process in which functional units evolve and become the building blocks for more-complex functional units. We are accustomed to the fact that polypeptides fold into globular protein structures. Although one might imagine an endless variety of ways in which stretches of amino acids could assemble into a compact structure, evolution renders a different judgment. Evolution has settled on a limited number of these possibilities that have proven to be especially suitable as modules from which to build a myriad of different proteins. These modules are 'conserved domains'; typically, they are 60–120 amino acids in length. Each conserved domain has a characteristic fold consisting of  $\alpha$  helices and  $\beta$  strands assembled onto each other in a fashion that is predictable and stable in evolution. The versatility of a conserved domain lies in the fact that amino acids that are oriented towards the exterior can evolve relatively freely (especially the loops, which are most tolerant of change) giving rise to enormous functional diversity. However, key residues that stabilize the domain are sufficiently conserved during evolution to maintain the domain structure. Typically, these conserved residues include hydrophobic residues oriented to the interior of the domain. In addition, Ig domains and others 'designed' for use in the extracellular compartment often have intrachain disulfide bonds between conserved cysteines. Structural and evolutionary biologists are assembling an increasingly comprehensive understanding of conserved domains; there are scholarly resources which systematically catalog them, including PFAM (<http://www.sanger.ac.uk/Software/Pfam>)<sup>6</sup> and SMART (<http://smart.embl-heidelberg.de>)<sup>7</sup>, and sequence analysis

tools to identify conserved domains within an amino acid sequence, such as CDD (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>).

There are only a few thousand conserved domains (depending on how you enumerate them). The majority of domains evolved early in evolution, either in prokaryotes or single-cell eukaryotes (as exemplified by yeast). Only  $\approx 7\%$  of the domains are unique to vertebrates, indicating the limited recent evolution of domains. Analysis of the human genome indicates that the immune system has a disproportionate share of these recently evolved domains, particularly with reference to the cytokines<sup>2</sup>. For example, the chemokine domain evolved in vertebrates and underwent rapid gene duplication, giving rise to at least 42 genes, and likewise the vertebrate-restricted interleukin (IL)-1 domain occurs in at least seven genes. This innovation of domains appears to reflect strong selective pressures for immune system enhancement, presumably because the immune system provides protection from infectious agents.

Rather than the creation of new domains, a more typical strategy in vertebrate evolution is the innovative utilization of pre-existing domains, dramatically illustrated by the Ig domain. The Ig domain is absent in yeast and evolved in simple metazoans. It underwent a systematic expansion from being the 34th most abundant domain in the worm, to the ninth most abundant in the fly, to the most popular domain in humans, present in 765 genes<sup>1</sup>. Although spectacular, the rise in the number of genes with Ig domains is a less striking evolutionary achievement than the diversification of ways in which the Ig domain is used. The phrase 'domain architecture' of a protein refers to the combination of domain types present in that protein. Consider three different Ig-domain containing architectures: in MHC class II, the Ig domain occurs together

with a class II  $\alpha$  domain and a transmembrane region; in *flt3*, it occurs together with a transmembrane region and two kinase domains; and in the IL-6 receptor  $\alpha$  chain, it occurs together with a fibronectin type III domain and a transmembrane domain. A tally of Ig domain usage indicates that it participates in >75 different domain architectures, more than for any other domain, including the eukaryotic protein kinase domain<sup>1</sup>.

Although the Ig domain is an extreme example, it illustrates the general rule that the diversification of domain architecture (by gene duplication and the shuffling of domains between different genes) was a major strategy of innovation during the evolution of vertebrates from invertebrates. The greatest use of this innovation strategy has occurred in transmembrane proteins ( $\approx 200\%$  increase in diversity from fly to human), compared with only  $\approx 50\%$  increase in the diversity of architectures of intracellular proteins<sup>1</sup>. This finding is in keeping with our general impression that a great deal of the innovation in the immune system has occurred in transmembrane molecules.

It is worth inserting a cautionary note that an understanding of evolutionarily conserved domains can only take us so far. It is useful to know, for example, that mucosal addressin cell adhesion molecule (MadCAM)-1 is composed of multiple Ig domains<sup>8</sup>. If the function of MadCAM was unknown, the presence of Ig domains would provide useful suggestions as to potential ligands (e.g. other Ig domains or integrin molecules). However, precisely because the externally-oriented residues of domains can evolve relatively freely, different domains of the same conserved family can (and do) end up serving multiple functions. Indeed, detailed study of domain families generally identifies subfamilies, which give rise to more specific predictions of function<sup>4</sup>. In addition, there are other limitations. First, the computational algorithms for detection of conserved domains are not yet sufficiently robust, and therefore some domain assignments are missed. Second, many protein functions are not mediated by such globular protein domains. Rather, 'intrinsically disordered' regions of proteins play crucial roles<sup>9</sup>. For example, part of vascular cell adhesion molecule (VCAM)-1 constitutes a mucin-like region, containing multiple sites for O-linked

glycosylation, whose amino acid sequence conservation is subject to different kinds of constraints than the 'conserved domains' we have been discussing.

#### Evolutionary singularities of the immune system

Immunologists have come to view two aspects of the immune system as genuinely unique from an evolutionary perspective: the extraordinary polymorphism of the human histocompatibility region (HLA) and somatic cell gene rearrangement for B- and T-cell receptors. The increasingly complete view of the human genome now available does not reveal new information to challenge the uniqueness of these two phenomena. The polymorphism of the HLA region is understood to reflect balanced selection for many alleles, reflecting the very distinctive role of these gene products in regulating immune responses to environmental pathogens<sup>10</sup>. Previous analyses indicate, more generally, that immune genes have been subjected to greater evolutionary pressure than non-immune genes<sup>11</sup>. Although none rival HLA polymorphism, it will be interesting to learn which immune system genes have greater polymorphism than expected and which contribute to the range of normal variation in immune responses between individuals. Such analysis of immune system polymorphisms will be a component of the emerging whole-genome cataloging of polymorphism within the human population.

Somatic cell gene rearrangement is the second evolutionary singularity in the immune system. It has been postulated that recombinase-activating gene 1 (*RAG1*) and *RAG2* responsible for this recombination are derived from transposable elements<sup>12</sup>. The new genome analyses provide an increasing understanding of multiple classes of transposable elements and their prominent role in evolution of the human genome. Approximately half of the human genome ('junk DNA') is thought to originate from transposable elements and 47 genes have been provisionally assigned as originating from transposable elements<sup>1</sup>. However, alone among them, *RAG1* and *RAG2* gave rise to an evolutionary breakthrough in somatic cell gene rearrangement. Perhaps, 1.5 billion base pairs of junk

DNA is a small price to pay for the *RAG1* and *RAG2* innovation!

#### Future directions

The public consortium's article<sup>1</sup> closes with a fitting quote from T.S. Eliot: 'We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time.' Indeed, such exploration is continuing at a furious pace. The next major goal likely to be fulfilled is the cataloging of the entire set of human transcripts (transcriptome)<sup>13</sup>, which is predicted to be at least three times greater than the number of human genes, owing primarily to alternative splicing. Cataloging the entire collection of post-translationally modified human protein products (the proteome) is a good deal more daunting, but is now possible given the improvements in mass spectrometry and other methodologies<sup>14</sup>. How do we efficiently and systematically link these defined genetic, transcriptional and proteomic elements to immune system processes? Using the power of large-scale mouse mutagenesis screens<sup>15</sup>, investigators at the John Curtin School of Medical Research, Australian National University are screening for 'all' genes that are key regulators of the immune system<sup>16</sup>. The results of this, and other systematic approaches, promise to revolutionize our understanding of vertebrate immune system design and regulation.

#### Acknowledgements

We thank Phil Murphy and Ken Katz for suggestions on the manuscript.

#### References

- 1 The genome international sequencing consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- 2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- 3 Bork, P. *et al.* (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol.* 242, 309–320
- 4 Wang, J. and Springer, T.A. (1998) Structural specializations of immunoglobulin superfamily members for adhesion to integrins and viruses. *Immunol. Rev.* 163, 197–215
- 5 Halaby, D. *et al.* (1999) The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng.* 12, 563–571
- 6 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266
- 7 Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234

- 8 Briskin, M.J. *et al.* (1993) MAdCAM-1 has homology to immunoglobulin and mucin-like adhesion receptors and to IgA1. *Nature* 363, 461–464
- 9 Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* 293, 321–331
- 10 Dawkins, R. *et al.* (1999) Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol. Rev.* 167, 275–304
- 11 Murphy, P.M. (1993) Molecular mimicry and the generation of host defense protein diversity. *Cell* 72, 823–826
- 12 Agrawal, A. *et al.* (1998) Transposition mediated by *RAG1* and *RAG2* and its implications for the evolution of the immune system. *Nature* 394, 744–751
- 13 Dias Neto, E. *et al.* (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. U. S. A.* 97, 3491–3496
- 14 Smith, R.D. (2000) Probing proteomes – seeing the whole picture? *Nat. Biotechnol.* 18, 1041–1042
- 15 Justice, M.J. (2000) Capitalizing on large-scale mouse mutagenesis screens. *Nat. Genet.* 1, 109–115
- 16 Goodnow, C.C. *et al.* (1999) Mechanisms of self-tolerance and autoimmunity: from whole animal phenotypes to molecular pathways. *Cold Spring Harbor Symp. Quant. Biol.* 64, 313–322

Yin Liu

Stephen Shaw\*

Experimental Immunology Branch, National Cancer Institute, Bethesda, MD 20892, USA.

\*e-mail: sshaw@nih.gov

## Human mast cells and basophils in HIV-1 infection

Gianni Marone, Giovanni Florio, Angelica Petraroli, Massimo Triggiani and Amato de Paulis

Mast cells and basophils (FcεRI<sup>+</sup> cells) are classically involved in allergic disorders. HIV-1 glycoprotein gp120 acts as a viral superantigen by interacting with the heavy chain, variable 3 (V<sub>H</sub>3) region of IgE to induce cytokine release from FcεRI<sup>+</sup> cells. The chemokine receptors CCR3 and CXCR4, co-receptors for HIV-1, are expressed by FcεRI<sup>+</sup> cells. Via its interaction with CCR3, HIV-1 transactivation (Tat) protein is a potent chemoattractant for FcεRI<sup>+</sup> cells. Incubation of basophils with Tat protein upregulates the surface expression of the CCR3 receptor. There is some evidence that human FcεRI<sup>+</sup> cells could be infected *in vitro* by M-tropic HIV-1 strains.

Human mast cells and basophils are the only cells that express high-affinity receptors for IgE (FcεRI) and that also synthesize various proinflammatory mediators and cytokines<sup>1</sup>. FcεRI<sup>+</sup> cells play a fundamental role not only in the pathophysiology of allergic disorders, but also in the host's immune response to parasites and bacteria<sup>2</sup>. By contrast, the involvement of FcεRI<sup>+</sup> cells in viral infections is as yet largely unknown.

The past two decades have seen two apparently unrelated events: AIDS, which is caused by the viruses HIV-1 and HIV-2 (Ref. 3) and a pandemic of allergic disorders, which affects ≈15% of the population of industrialized countries<sup>4</sup>. There is no apparent association between these two disorders: AIDS is an acquired immunodeficiency of viral etiology<sup>3</sup> whereas atopic disorders result from the overproduction of IgE in response to allergens<sup>4</sup>. However, serum IgE levels are

increased in HIV-1-infected children<sup>5</sup> and adults<sup>6</sup> and the T helper1 (Th1) and Th2 cytokine profile is disturbed in HIV-1 infection because of a shift towards Th2-type cytokines<sup>7</sup>. Moreover, HIV-1-infected patients have an increased prevalence and severity of allergic reactions and adverse reactions to drugs<sup>8</sup>, and HIV-1 antigens induce histamine release from basophils<sup>9</sup>. Here, we review the data that implicate mast cells, basophils and their mediators in HIV-1 infection.

### Increased IgE levels in patients with HIV-1 infection

At least two cytokines [interleukin (IL)-4 and IL-13] play a key role in IgE production in humans<sup>4</sup>. IL-4 and IL-13 induce ε germline mRNA expression in B cells, which accounts for the ability of these cytokines to drive the switch to an IgE isotype. IL-4 and IL-13 are produced by Th2 cells<sup>4</sup>, basophils<sup>9</sup> and mast cells<sup>10</sup>. An additional signal for IgE synthesis is provided by the interaction of CD40 expressed by B cells with its natural ligand (CD40L) expressed by activated T cells. Activated mast cells and basophils also express CD40L (Refs 10, 11). This finding is consistent with the observation that activated mast cells and basophils, which release IL-4 and IL-13, can induce B cells to synthesize IgE.

IgE levels are elevated in HIV-1-infected children<sup>5</sup> and adults<sup>6</sup> and have been associated with disease progression<sup>12</sup>. The immune disturbance responsible for the increased IgE levels in HIV-1 infection is not fully understood. Although HIV-1-infected individuals

display polyclonal gammopathy and activated B cells<sup>13</sup>, there is no correlation between IgE levels and either IgG or IgA levels<sup>12</sup>.

### Th1–Th2 imbalance in HIV-1 infection

Clerici *et al.* provided the first suggestion that, during the early stages of HIV-1 infection, there is a shift from a Th1-type towards a Th2-type pattern of cytokine production by peripheral blood lymphocytes<sup>7</sup>. Increased Th2 cytokine production during HIV-1 infection has been confirmed<sup>14</sup>, although not all investigators have detected an overall shift in the cytokine pattern towards the Th2 subset in the lymph nodes or T-cell clones of HIV-1-infected individuals<sup>15,16</sup>. However, Maggi *et al.* found that HIV-1 replicates preferentially in Th2 as opposed to Th1 clones<sup>15</sup>.

These apparently conflicting results concerning the balance between Th1 and Th2 polarization in HIV-1-infected subjects could be caused by diverse factors: technical reasons, the production of Th2-type cytokines by cell types other than lymphocytes (e.g. FcεRI<sup>+</sup> cells), stimulation by specific HIV-1 superantigens or the effects of cytokines other than IL-4 (e.g. IL-13) on Th2-cell polarization.

### HIV-1 gp120 induces IL-4 and IL-13 release from FcεRI<sup>+</sup> cells

The HIV-1 envelope glycoprotein gp120 mediates entry into immune cells by binding to CD4, the 'primary receptor'<sup>17</sup>. CD4 binding causes conformational changes in gp120, resulting in the