Available online at www.sciencedirect.com

SCIENCE DIRECT

Genomics 85 (2005) 503–511

GENOMICS

www.elsevier.com/locate/ygeno

IRIS: A database surveying known human immune system genes ☆

James Kelley [a], Bernard de Bono [b], John Trowsdale [a,*]

[a] Department of Pathology, Immunology Division, University of Cambridge, Cambridge CB2 1QP, UK
[b] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

Received 11 December 2004; accepted 20 January 2005

Abstract

We have compiled an online database of known human defense genes: the Immunogenetic Related Information Source (IRIS). As of October 1, 2004, there are 1562 immune genes recorded in IRIS, representing 7% of the human genome. This resource contains searchable information including chromosomal location, sequence data, and a curated functional annotation for each entry. We used IRIS as a basis for analyzing the composition and characteristics of the immune genome, such as gene clustering, polymorphism, and relationship to disease. High protein sequence similarity correlated inversely with distance between immune genes, consistent with clustering of duplicated loci. We also found that, even though some immune genes exhibit high levels of polymorphism, such as MHC class I, the range of levels of polymorphism in immune genes is similar to that of nonimmune genes. Approximately 20% of immune genes have a known disease association. IRIS is available online at http://www.immunegene.org.

Keywords: Immune system; Human genome; Genetic information databases; Genetic polymorphisms; Gene clusters; Disease

Where comparisons have been made, such as between mouse and human [1] and *Anopheles* and *Drosophila* [2,3], the number of immune system genes differs markedly in related species. It has been suggested that other features, such as increased levels of polymorphism and clustering of immune system genes, may also reflect the strong selection required for resistance to infection [4]. A database of all known immune genes would provide an organized subset of functionally related human genes to compare immune system genes, as a set, with the rest of the genome. It could also be used in genetic studies of function and disease, in DNA arrays, and in computer prediction programs of potential therapeutic targets during drug development.

While several databases address the function of the immune system [5–7], none include an exclusive survey of the entire immune genome. To create such a database, we developed the Immunogenetic Related Information Source (IRIS). It currently includes information on chromosomal locations, protein and nucleotide sequences, and manual curations of the proposed function of each gene in immunity. This online resource is available at http://www.immunegene.org.

Using IRIS we set out to explore whether, as a group, immune system genes display any distinctive characteristics that reflect selective pressure for resistance to infection, including clustering, level of polymorphism, and genetic association with disease. This formal organization of immune genes provides a basis for functional interaction mapping, serves as a documented subset of functionally related genes, and facilitates future observations on genomic features of the immune system. Our initial analysis using IRIS suggests that immune defense genes, as a functional subset, exhibit marked gene duplication and association with disease, although these features are by no means restricted to immune defense.

☆ Supplementary data for this article may be found on ScienceDirect.

* Corresponding author. Fax: +44 1223 339768.
E-mail address: jt233@cam.ac.uk (J. Trowsdale).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.01.009

Results and discussion

Definition of an immune gene

For the purposes of this database, an immune gene was operationally defined as a complete gene that produces a functional transcript and demonstrates at least one defense characteristic listed in Table 1. Gene segments, such as those encoding antigen receptors, immunoglobulins, and T cell receptors [8,9], were not included. This exclusion prevented an overrepresentation of data when counting the number of immune genes. These immune gene segments are extensively analyzed in an online database, the International ImMunoGeneTics Information System (IMGT) (http://imgt.cines.fr:8104), maintained by LIGM at Montpellier, France [10,11]. (For more details on the definition of an immune gene, please refer to a recent review from our laboratory [4].)

Even though some genes not included under this definition may be involved in defense mechanisms in a broader sense, we were deliberately conservative for the initial compilation. We have not included genes whose products correct or prevent physical disturbances such as tissue repair from wounds or DNA repair from radiation (for example, melanin); genes whose products provide a physical barrier such as epithelia, unless there is a specific interaction with a pathogen (for example, keratin); genes whose role in immunity is used in many systems of the body (for example, actin); genes that share sequence similarity with known immune genes but have no putative defense function (for example, some immunoglobulin superfamily members); and genes whose products interact with non-pathogenic antigens (for example, some components in digestion or reproduction).

Composition and functional survey of the immune genome

As of October 1, 2004, IRIS contained 1562 immune genes, indicating that the immune genome comprises over 7% of the total 21,432 human genes. This was the number of complete human genes recorded in LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink).

In this analysis, the innate and adaptive immune systems occupied 41 and 27%, respectively, of the immune genome. The larger number of innate immune genes compared to adaptive immune system genes was affected by the absence in IRIS of immunoglobulin [8] and T cell receptor [9] gene segment repertoires, which in base pairs comprise a large portion of the adaptive immune genome. Genes dedicated to the development of immune tissues, for example in immune cell hematopoiesis, represented 8%. Approximately 24% of the immune genes were not easily placed into any of the previously mentioned categories. This "other" category included genes that can be induced by an immunomodulator, can participate in a pathway leading to the expression of an immune molecule, and are expressed primarily in immune tissues (see Fig. 1).

Table 1
Defense characteristics of the immune system

Known or putative function in innate or adaptive immunity
Participates in the development or maturation of immune system components
Induced by immunomodulators[a]
Encodes a protein expressed primarily in immune tissues[b]
Participates in an immune pathway that results in the expression of defense molecules[c]
Produces a protein that interacts directly with pathogens or their products

Any complete, human gene that produces a functional transcript and matches one or more of these criteria is included in IRIS.

[a] Immunomodulators, such as IFN-γ, regulate expression of immune components. Proteins that are direct effectors of immunomodulators are thought to have an immune function.

[b] We assume that if a gene is commonly expressed in immune tissues then it has a role in immune functioning.

[c] NF-κB is an example.

Distribution of immune genes

IRIS was used to ask whether immune system genes are randomly distributed in the genome. They were found to be fairly evenly distributed, as expected of an ancient, complex system. However, mapping immune gene distribution, which is shown as Fig. 2, revealed some variation in immune gene densities. Chromosomes 19, 17, 6, and 11 had higher, and chromosomes 18, 13, X, and 14 had lower, immune gene densities, while the Y chromosome was completely lacking in immune genes. There were some pockets of immune genes at very high density. In some cases this was accounted for by adjacent duplicated genes, such as in the leukocyte receptor complex (LRC) of chromosome 19, the natural killer complex of chromosome 12, the MHC on chromosome 6, and the cluster of interleukin genes on chromosome 1. This increased density might be due to relatively recent gene duplications, as in the LRC, but there was also evidence for examples, such as the MHC, that do not solely consist of related duplicates.

Clustering of immune genes

There are essentially two types of gene clusters. The most common are clusters of related duplicated genes whose functions diverge, such as the *KIRs* and *ULBPs*. These presumably result from tandem segmental duplications [12], caused by slipped-strand mispairing, gene conversion, or unequal crossover [13]. Tandem duplication explains the clustered arrangement of the NBS–LRR disease resistance genes in plants, for example [14].

A second type of gene cluster is of genes with related functions but with unrelated sequences. Clustering of this type is found in prokaryotic operons that encode interacting proteins [15]. With the exception of gene clustering in operons of *Caenorhabditis elegans* [16], it is generally