Published in final edited form as:

Evol Dev. 2009 ; 11(4): 382–390. doi:10.1111/j.1525-142X.2009.00345.x.

A phylogenetic approach to gene expression data: evidence for the evolutionary origin of mammalian leukocyte phenotypes

Austin L. Hughes^{*} and Robert Friedman

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

SUMMARY

The evolution of multicellular organisms involved the evolution of specialized cell types performing distinct functions; and specialized cell types presumably arose from more generalized ancestral cell types as a result of mutational event, such as gene duplication and changes in gene expression. We used characters based on gene expression data to reconstruct evolutionary relationships among 11 types of lymphocytes by the maximum parsimony method. The resulting phylogenetic tree showed expected patterns including separation of the lymphoid and myeloid lineages; clustering together of granulocyte types; and pairing of phenotypically similar cell types such as T-helper cells type 1 and T-helper cells type 2 (Th1 and Th2). We used phylogenetic analyses of sequence data to determine the time of origin of genes showing significant expression difference between Th1 and Th2 cells. Many such genes, particularly those involved in the regulation of gene expression or activation of proteins, were of ancient origin, having arisen by gene duplication before the most recent common ancestor (MRCA) of tetrapods and teleosts. However, certain other genes with significant expression difference between Th1 and Th2 arose after the tetrapod--teleost MRCA, and some of the latter were specific to eutherian (placental) mammals. This evolutionary pattern is consistent with previous evidence that, while bony fishes possess Th1 and Th2 cells, the latter differ phenotypically in important respects from the corresponding cells of mammals. Our results support a gradualistic model of the evolution of distinctive cellular phenotypes whereby the unique characteristics of a given cell type arise as a result of numerous independent mutational changes over hundreds of millions of years.

INTRODUCTION

The evolution of multicellular organisms involved the evolution of specialized cell types performing distinct functions. One type of evidence that has long been used to reconstruct the evolutionary events involved in the origin of distinct cell and tissue types is developmental evidence. For example, a common evolutionary origin has been hypothesized for tissues that arise developmentally from the neural crest cells in vertebrates (Holland et al. 2001). A hypothesis for a common evolutionary origin for two distinct cell or tissue types implies that one or more mutational events occurred in the evolutionary past, giving rise to two cell types with distinct gene expression patterns. Such mutational events might include changes in promoter regions causing changes in gene expression pattern, as well as gene duplication events giving rise to specialized protein functions (Hughes 1994).

Recent advances in molecular biology provide new sources of data, including genomic sequences and gene expression data, which can provide evidence regarding the ancient mutational events that have given rise to distinct cell types. Here we illustrate the application of such data to the question of the evolutionary origin of mammalian leukocyte types. The

vertebrate immune system depends on the actions of several distinctive cell types, classed together as leukocytes, which play a variety of roles as regulators and effectors of host defense mechanisms (Klein and Hořejší 1997). Developmentally these cells are, along with erythrocytes, derived from a single stem cell type, the hematopoietic stem cell (Iwasaki and Akashi 2007; Lai and Kondo 2007; Orkin and Zon 2008). Customarily, mature mammalian blood cells are classified as belonging to one of two separate lineages: (1) the lymphoid lineage, comprising B-cells (BC), T-cells (including cytotoxic T-cells, both T-helper cells type 1 and 2 [Th1 and Th2], and T-memory cells), and natural killer cells (NK); and (2) the myeloid lineage, comprising granulocytes (basophils [BAS], eosinophils [EOS], and neutrophils [NEUT]), monocytes, macrophages (MAC), and mast cells. Traditional hierarchical pictures of the developmental origin of these cell types are now viewed as oversimplified, because there appears to be considerable flexibility in developmental pathways (Orkin and Zon 2008).

Here we use gene expression data to reconstruct evolutionary relationships among 11 types of lymphocytes. In particular we focus on the Th1 and Th2 cells, which in mammals play key roles in coordinating immune responses. These cell populations are distinguished operationally by the cytokines they produce: interleukin-2 (IL-2), lymphotoxin α (TNF- β), and interferon- γ in the case of Th1; and IL-4, IL-5, IL-6, and IL-10 in the case of Th2 (Romagnani 1991; Del Prete 1998). There is evidence that Th1 and Th2 cells are also found in birds (Göbel et al. 2003; Degen et al. 2005); but there appears to be no evidence regarding their presence or absence in any tetrapod outside of birds and mammals. On the other hand, a few recent studies have suggested that Th1 and Th2 populations may exist at least in some form in bony fishes (teleosts). In mammals, the cytokine IL-12 is important in stimulating the development of the Th1 phenotype (Watford et al. 2003). Mammalian IL-12 is a heterodimer, consisting of 2U designated p35 and p40; teleosts have been found to possess a single orthologue of p35 but multiple p40 genes (Huising et al. 2006a). Furthermore, the discovery of an IL-4 gene in the green pufferfish *Tetraodon nigroviridis* supports the hypothesis that teleosts possess at least some degree of Th2 function (Li et al. 2007).

Here we analyze human gene expression data in order to identify genes with significantly different patterns of expression in the Th1 and Th2 cells. We focus on expression data from unstimulated cells, rather than stimulated cells, because we are interested in discovering candidate genes for major developmental switches in the differentiation between the Th1 and Th2 pathways. In the terminology of Davidson and Erwin (2006), we are searching for genes likely to be components of the “kernels” of gene regulatory networks responsible for the differentiation between these two cellular phenotypes. We then use phylogenetic analysis of such genes in order to obtain evidence regarding the evolutionary origin of the distinctive characteristics of these two T-cell populations. By timing gene duplication events relative to events of organismal divergence (cladogenesis), we place the evolution of the Th1 and Th2 phenotypes in the context of vertebrate history, particularly with regard to the divergence of the tetrapod and teleost lineages.

METHODS

Gene expression data

Gene expression data for 22,215 human transcripts were taken from Gene Expression omnibus series GSE3982, based on the Affymetrix Genome U133 Array Set HG-U133A (Jeffrey et al. 2006). Jeffrey et al. (2006) isolated pure leukocyte populations (leukocyte types) and obtained data on gene expression before and after stimulation with appropriate stimuli to induce activation. In the present analysis, only the data from unstimulated cells were used. Cells were classified as belonging to the following 11 types (with abbreviations used below): B-cells (BC); natural killer cells (NK); central memory cells (CMEM); effector memory cells (EFM); Th1 T-helper cells (Th1); Th2 T-helper cells (Th2); macrophages (MAC); cord-blood mast cells

(CMAS); basophils (BAS); eosinophils (EOS); and neutrophils (NEUT). Each cell type was represented by two replicates.

One-way analysis of variance (ANOVA) was applied to each of 22,215 transcripts to test for an overall difference in mean expression scores among the 11 cell types, corrected for multiple testing by the Bonferroni method. In the case of the transcripts for which the overall F -test was significant at the 5% level (Bonferroni-corrected), planned comparisons (Lindman 1974) were conducted between Th1 and Th2.

A standardized score for each cell type and a given transcript was computed as follows: Let x_{ij} represent the mean score of the i th cell type for the j th transcript; let $x_{..j}$ represent the mean of the x_{ij} values for the j th transcript; and let s_j represent the standard deviation (SD) of the 11 x_{ij} values for the j th transcript. Then the standardized score for cell type i and transcript j is given by

$$Y_{ij} = (x_{ij} - x_{..j}) / s_j \quad (1)$$

Cluster analysis was applied to the standardized scores using a single-link clustering algorithm implemented in the Minitab statistical analysis software package (<http://www.minitab.com>). For purposes of clustering, the distance between cell types i and k was computed by the formula

$$d_{ik} = 1 - r_{ik} \quad (2)$$

where r_{ik} is the Pearson correlation coefficient, computed across all transcripts, between the standardized scores of cell types i and k .

In order to conduct phylogenetic analysis of cell types based on gene expression, each of the 22,215 transcripts was treated as a character with a categorical value ("expression level category") based on Y_{ij} . For the j th transcript, Y_{ij} was placed into one of three expression level categories: (1) ≤ -1 ; (2) > -1 but < 1 ; or (3) ≥ 1 . These categories correspond, respectively, to expression levels 1 SD or more below the mean for a given transcript; within 1 SD of the mean for the transcript; and 1 SD above the mean for the transcript. Maximum parsimony (MP) analysis, using heuristic search, was applied to these characters using the PAUP* program (Swofford 1999). Bootstrapping was used to test for reliability of clustering patterns in the phylogenetic analysis (Felsenstein 1985); 1000 bootstrap samples were used. The consistency index (CI) for characters was computed as the maximum possible number of changes at a given character divided by the number of changes inferred by the MP tree. A CI of 1.0 thus indicates a character (transcript) with no homoplasy (forward/backward change or parallel change).

Sequence data

The amino acid sequences corresponding to transcripts of interest were used in BLAST homology searches of the NCBI nonredundant protein database to find homologous genes from model organisms. Sequences were gathered from the following vertebrate taxa: Mammalia: Eutheria: human *Homo sapiens*; dog *Canis lupus familiaris*; and mouse *Mus musculus*; Mammalia: Metatheria: opossum *Monodelphis domestica*; Mammalia: Monotremata: platypus *Ornithorhynchus anatinus*; Aves: chicken *Gallus gallus*; Amphibia: *Xenopus laevis* and *Xenopus tropicalis*; Actinopterygii: zebrafish *Danio rerio* and green pufferfish *T. nigroviridis*. When sequences from animals other than vertebrates were available, we used these sequences as outgroups to root phylogenetic trees. The following species were included: the urochordate *Ciona intestinalis*; the sea urchin *Strongylocentrotus purpuratus*; the insects *Drosophila melanogaster*, mosquito *Aedes aegypti*, red flour beetle *Tribolium castaneum*,

jewel wasp *Nasonia vitripennis*, and honeybee *Apis mellifera*; the nematode worm *Caenorhabditis elegans*; and the sea anemone *Nematostella vectensis*.

Amino acid sequences were aligned using the CLUSTAL X program (Thompson et al. 1997). Phylogenetic trees were also constructed from amino acid sequences using the following methods: (1) MP (Swofford 1999); (2) the quartet maximum likelihood (Strimmer and von Haeseler 1996); and (3) the neighbor-joining (NJ) method (Saitou and Nei 1987). NJ trees of amino acid sequences were constructed on the basis of the Poisson, equal-input, JTT models (Tamura et al. 2007). Because all methods yielded very similar results only NJ trees based on the JTT model are shown below. Bootstrapping was used to test for reliability of clustering patterns in NJ trees; 1000 bootstrap samples were used.

Branching patterns in phylogenetic trees were used to establish the timing of gene duplication events relative to events of cladogenesis in vertebrate evolutionary history. In particular, the phylogenetic analyses were used to time duplication events relative to the most recent common ancestor (MRCA) of eutherian (placental) mammals; the MRCA of all mammals; the MRCA of amniotes and amphibians; and the MRCA of tetrapods and teleosts.

Related gene pairs

All genes for which expression data were available were assigned to gene families using the BLASTCLUST program (Altschul et al. 1997), using default values except for a 30/50 criterion (Hughes et al. 2005); where 30% is the minimum amino acid similarity between any two matching sequences across at least 50% of the aligned lengths. Once all pairs of similar sequences were identified, BLASTCLUST then clustered them into sets of orthologues by a single-linkage method. A set of 631 pairs of related genes (families having two members in the expression data set) were identified. Expression levels in each of these families were compared in Th1 and Th2 cells by factorial ANOVA, with the gene and the cell type as the two factors. We looked for significant gene by cell type interactions, which indicated pairs of related genes that differed in expression pattern in the two cell types.

RESULTS

Patterns of gene expression

Gene expression data on 22,215 transcripts were tested by ANOVA for significant differences among 11 human leukocyte types. Nine hundred and forty-four transcripts showed a significant (Bonferroni-corrected) overall *F*-test. Of these transcripts, 17 corresponded to genes listed as regulated by IL-4 and/or IL-12 during differentiation of Th1 and Th2 cells by Lund et al. (2007): *AKAP12*, *AKAP2*, *CEBPB*, *EPLIN*, *HOP*, *HPGD*, *KIAA0992*, *LIMS1*, *LOC93081*, *LY9*, *PASK*, *PSCD1*, *RAB11FIP1*, *RAB27B*, *RHOQ*, *TMEM49*, and *TNSF10*. There was a significant (Bonferroni-corrected) difference in planned comparisons between the cell types of the lymphoid lineage (BC, NK, CMEM, EFM, Th1, and Th2) and those of the myeloid lineage (CMAS, MAC, BAS, EOS, and NEUT) in the case of 551 transcripts (58.4%).

Cluster analysis based on the correlation matrix of mean expression scores among the 11 cell types did not show separate clustering of the lymphoid and myeloid lineages (Fig. 1A). However, the three types of granulocytes (BAS, EOS, and NEUT) did cluster together (Fig. 1A). The most similar pairs of cell types were the two types of helper T-cells (Th1 and Th2) and the two types of memory cells (CMEM and EFM; Fig. 1A).

MP analysis applied to characters based on expression level category yielded a phylogenetic tree topologically different from that produced by cluster analysis (Fig. 1, A and B). The MP tree was based on 22,215 characters (corresponding to transcripts), of which 16,389 were parsimony informative. The overall CI for the tree was 0.678; CI computed for informative

characters only was 0.624. In the MP tree, the lymphoid and myeloid lineages were separated, although bootstrap support for this pattern (59%) was not very strong (Fig. 1B). Within the myeloid lineage, the three granulocyte cell types clustered together with very strong (99%) bootstrap support (Fig. 1B). Th1 and Th2 cells clustered together with 100% bootstrap support, as did the memory cell types CMEM and EFM (Fig. 1B).

Th1 versus Th2

There were 435 transcripts that both had a CI = 1.0 and showed a change in expression level category on the branch leading to Th1 from the common ancestor of Th1 and Th2. A nonoverlapping set of 485 characters both had a CI = 1.0 and showed a change in expression level category on the branch leading to Th2 from the common ancestor of Th1 and Th2. From among these transcripts, transcripts with a significant overall *F*-test and a significant *F*-test for the comparison between Th1 and Th2 (both Bonferroni-corrected) were selected as candidate genes likely to play key roles in the differentiation between these two cell types. This very conservative procedure was used to identify a set of genes highly likely to play a role in the differentiation of Th1 and Th2. There were 22 such genes, 16 of which showed increased expression in Th1 relative to Th2 and six of which showed increased expression in Th2 relative to Th1 (Table 1).

The IL-12 receptor β 2-subunit (IL-12R β 2), encoded by *IL12RB2*, was previously shown to be induced during differentiation of human Th1 cells but not Th2 cells, consistent with the present results (Table 1). Similarly, lymphotoxin α (also known as TNF- β , encoded by *LTA*; Table 1) is known to be expressed by Th1 cells (Romagnani 1991). *DDP4*, *DUSP5*, *DOK5*, and *ZBTB32* (Table 1) have all previously been reported to play roles in T-cell development, though not in the differentiation of Th1 and Th2 cells *per se* (Tanaka et al. 1992; Bilic and Ellmeier 2007; Kovanen et al. 2008).

Phylogenetic analyses were used to determine the time of origin of the genes showing significant differences between Th1 and Th2 relative to major cladogenic events in vertebrate evolutionary history. In 16 of the 22 genes (72.3%), the topology of the phylogenetic tree supported the hypothesis that the gene arose by gene duplication before the MRCA of tetrapods and teleosts with at least 67% bootstrap support (Table 1). In 12 of the genes (54.5%), the bootstrap support for gene duplication before the MRCA of tetrapods and teleosts was 97% or better (Table 1). *P4HA2* (encoding procollagen-proline 2-oxaloglutarate 4-dioxygenase, α) provides an example of a topology of this form (Fig. 2). The phylogenetic tree of vertebrate members of the P4HA gene family was rooted with insect homologs (Fig. 2). The vertebrate genes formed two clusters, *P4HA1* and *P4HA2*, each of which received very strong (99%) bootstrap support (Fig. 2). Each of the clusters included both teleost and tetrapod sequences (Fig. 2), supporting the hypothesis that the gene duplication that gave rise to *P4HA1* and *P4HA2* occurred before the MRCA of tetrapods and teleosts. Note that, although our database homology search uncovered no teleost orthologue of mammalian *DOK5*, orthologues of the related *DOK4* and *DOK6* genes were found in zebrafish; and the phylogeny supported the hypothesis that *DOK5* originated before the MRCA of tetrapods and teleosts.

Among the 22 genes with significant difference between Th1 and Th2, there was 1 (*FUT8*) that showed no evidence of duplication; thus, there was a single orthologous gene in insects and in vertebrates (Table 1). On the other hand, four genes were specific to mammals alone or to birds and mammals (Table 1). One of these (*CFH*) was specific to primates (Table 1). For example, *ZNF215* (encoding zinc finger protein 215) formed part of a group of zinc finger protein genes that were shown by phylogenetic analysis to be mammal specific (Fig. 3). In the phylogenetic tree, there was a cluster including *ZNF215* and certain other mammalian sequences but no sequences from chicken, *Xenopus*, or bony fishes; and this cluster received 99% bootstrap support (Fig. 3).

Related gene pairs

As a complementary approach for studying expression differences within gene families, we used factorial ANOVA to compare expression patterns of related gene pairs in the expression data set. When this approach was applied to expression in Th1 and Th2 cells, three gene pairs showed significant (Bonferroni-corrected) gene by cell-type interactions (Fig. 4). One of these pairs involved *ARID3A* and *ARID3B* (Fig. 4A), the latter of which was previously identified (Table 1). In the comparison between *ARID3A* and *ARID3B*, there was a significant interaction because, although *ARID3B* showed higher mean expression scores than *ARID3A* in both Th1 and Th2, the difference was much greater in the case of Th1 (Fig. 4A). Similarly, *GFPT1* showed higher mean expression scores than *GFPT2* in both Th1 and Th2, but the difference was much more pronounced in Th1 (Fig. 4B). Conversely, *RBPJ* showed higher mean expression scores than *RBPJL* in both Th1 and Th2, but the difference was much more pronounced in Th2 (Fig. 4C).

Phylogenetic analysis supported the hypothesis that *ARID3A* and *ARID3B* arose by gene duplication before the MRCA of tetrapods and teleosts (Fig. 5). Phylogenetic analysis likewise supported (with 100% bootstrap support) the hypothesis that the *GFPT1* and *GFPT2* genes arose by gene duplication before the tetrapod–teleost MRCA (not shown). In the case of *RBPJ* and *RBPJL*, the phylogeny supported the hypothesis that these genes duplicated before the tetrapod–teleost MRCA (with 100% bootstrap support; Fig. 6). In addition, the phylogeny supported the hypothesis that *RBPJ* and *RBPJL* arose by duplication before the MRCA of chordates and urochordates (with 99% bootstrap support; Fig. 6).

DISCUSSION

Gene expression data in cells belonging to 11 distinct human leukocyte types were used to reconstruct the evolutionary history of the cell types. A MP analysis showed a clear separation of the lymphoid and myeloid lineages and strongly supported a monophyletic cluster of the granulocytes within the myeloid lineage. Within the lymphoid lineage, there was strong support for the expected close relationships between Th1 and Th2 cells and between CMEM and EFM cells.

Two complementary statistical approaches, based on ANOVA, identified 24 genes with significant differences in expression pattern between Th1 and Th2 cells (Table 1 and Fig. 4). Six of these genes encode proteins known to be directly involved in the regulation of gene expression: *NFE2L3* (Derjuga et al. 2004); *APBB2* (Bruni et al. 2002); *ARID3B* (Wilsker et al. 2005); *ZBTB32* (Bilic and Ellmeier 2007); *ZNF215* (Alders et al. 2000); and *RBPJ* (Minoguchi et al. 1997). Also included at least five genes whose protein products play a role in the modification or activation of other proteins: *P4HA2* (Koivunen et al. 2006); *DPP4* (Tanaka et al. 1992); *FUT8* (Ihara et al. 2006); *DUSP5* (Kovanen et al. 2008); and *DOK5* (Shi et al. 2006). These 11 genes might be considered as candidates for key developmental switch genes; that is, components of the kernels (Davidson and Erwin 2006) of the gene regulatory networks responsible for the differentiation between Th1 and Th2 cells.

Phylogenetic analyses supported the hypothesis that 8 of the 11 candidate switch genes arose by gene duplication before the MRCA of tetrapods and teleosts, while 1 (*FUT8*) was unduplicated because the MRCA of tetrapods and teleosts (Table 1 and Fig. 6). In the case of *DPP4*, the phylogenetic analysis provided only modest support (72% bootstrap support) for the hypothesis that gene duplication occurred after the tetrapod–teleost MRCA (Table 1), while in the case of *ZNF215* there was strong evidence that this gene arose within the mammals (Table 1 and Fig. 3). Thus, most candidate developmental switch genes were of ancient origin and were present in the common ancestor of tetrapods and teleosts, consistent with the hypothesis that the basic Th1 and Th2 phenotypes arose before the tetrapod–teleost MRCA

(Li et al. 2007). However, a few of these genes arose later, suggesting that the Th1 and Th2 phenotypes have evolved further specific features over the evolutionary history of tetrapods.

The 24 genes with significant expression differences between Th1 and Th2 included a number of genes encoding effector molecules of these cells: genes encoding cytokines (*LIF*, *LTA*, and *IL26*); the gene (*IL12RB2*) encoding the signaling component of the IL-12; and the gene (*GNLY*) encoding the cytotoxic peptide granulysin (Callard and Gearing 1994; Rogge et al. 1997; Hör et al. 2004; Huising et al. 2006b; Latinovic-Golic et al. 2007). Of these five genes, three (*LIF*, *IL26*, and *GNLY*) were specific to mammals or to birds and mammals (Table 1). Overall, of five genes that originated within mammals or at least within amniotes (*LIF*, *CFH*, *ZNF215*, *IL26*, and *GNLY*), two showed enhanced expression in Th1, while three showed enhanced expression in Th2 (Table 1). Thus, both Th1 and Th2 phenotypes have continued to evolve significant new aspects within the amniotes lineage, although their origin is likely to predate the tetrapod–teleost MRCA.

When a new cell type evolves, it evidently can co-opt a previously existing gene through changes in expression pattern. For example, the *FUT8* gene is an ancient unduplicated gene with orthologues in insects and vertebrates that showed significantly enhanced expression in Th1 cells (Table 1). Clearly this gene was present before the evolution of any of the cell types unique to the vertebrate immune system; presumably changes in its expression pattern allowed it to play a role in the development of the Th1 phenotype. Similarly, phylogenetic analysis supported the hypothesis that *RBPJ* arose by gene duplication before the origin of the vertebrates (Fig. 6). Moreover, it is worth noting that, in the case of any of the genes that arose through gene duplication before the tetrapod–teleost MRCA, their elevated expression in Th1 or Th2 may have evolved long after their first appearance. Nonetheless, genes that have arisen after the tetrapod–teleost MRCA and have elevated expression in mammalian Th1 or Th2 must represent refinements of the Th1 or Th2 phenotype that were not present in the common ancestors of tetrapods and teleosts.

Our results thus support a gradualistic model of the evolution of distinctive cellular phenotypes whereby the unique characteristics of a given cell type arise as a result of numerous independent mutational changes over hundreds of millions of years. These mutational changes include both the origin of new genes by gene duplication as well as changes in expression pattern of already existing genes. This study suggests that the analysis of gene expression data from a phylogenetic perspective can provide insights into the evolutionary origin of phenotypically distinct cell and tissue types, and that such analyses can shed light on the evolutionary origin of cellular differentiation throughout the history of life.

Acknowledgments

This research was supported by grant GM43940 from the National Institutes of Health to A. L. H.

REFERENCES

- Alders M, et al. Disruption of a novel imprinted zinc-finger gene, ZNF215, in Beckwith-Wiedemann Syndrome. *Am. J. Hum. Genet* 2000;66:1473–1484. [PubMed: 10762538]
- Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
- Bilic I, Ellmeier W. The role of BTB domain-containing zinc finger proteins in T cell development and function. *Immunol. Lett* 2007;108:1–9. [PubMed: 17084908]
- Bruni P, et al. Fe65, a ligand of the Alzheimer's β -amyloid precursor protein, blocks cell cycle progression by down-regulating thymidate synthase expression. *J. Biol. Chem* 2002;277:35481–35488. [PubMed: 12089154]
- Callard, RE.; Gearing, AJ. *The Cytokine Factsbook*. Academic Press; London: 1994.

- Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science* 2006;311:796–800. [PubMed: 16469913]
- Degen WG, van Daal N, Rothwell L, Kaiser P, Schijns V. Th1/Th2 polarization by viral and helminth infection in birds. *Vet. Microbiol* 2005;105:163–167. [PubMed: 15708812]
- Del Prete G. The concept of type-1 and type-2 helper T cells and their cytokines in humans. *Int. Rev. Immunol* 1998;16:427–455. [PubMed: 9505198]
- Derjuga A, et al. Complexity of CNC transcription factors as revealed by gene targeting of the *Nrf3* locus. *Mol. Cell. Biol* 2004;24:3286–3294. [PubMed: 15060151]
- Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;39:783–791.
- Göbel TW, et al. IL-18 stimulates the proliferation and IFN- γ release of CD4+T cells in the chicken: conservation of a Th1-like system in a nonmammalian species. *J. Immunol* 2003;171:1809–1815. [PubMed: 12902481]
- Holland, PW.; Wada, H.; Manzanares, M.; Krumlauf, R.; Shimeld, SM. The origin of the neural crest.. In: Ahlberg, PE., editor. *Major Events in Early Vertebrate Evolution*. Taylor and Francis; London: 2001. p. 33-39.
- Hör S, et al. The T-cell lymphokine interleukin-26 targets epithelial cells through the interleukin-20 receptor 1 and interleukin-10 receptor 2 chains. *J. Biol. Chem* 2004;279:33343–33351. [PubMed: 15178681]
- Hughes AL. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B* 1994;256:119–124.
- Hughes AL, Ekollu V, Friedman R, Rose JR. Gene family content-based phylogeny of prokaryotes: the effect of search criteria. *Syst. Biol* 2005;54:268–276. [PubMed: 16012097]
- Huising MO, Kruiswijk CP, Flik G. Phylogeny and evolution of class-I helical cytokines. *J. Endocrinol* 2006b;189:1–25. [PubMed: 16614377]
- Huising MO, et al. The presence of multiple and differentially regulated interleukin-12p40 genes in bony fishes signifies an expansion of the vertebrate heterodimeric cytokine family. *Mol. Immunol* 2006a; 43:1519–1533. [PubMed: 16460805]
- Ihara H, Ikeda Y, Taniguchi N. Reaction mechanism and substrate specificity for nucleotide sugar of mammalian α 1,6-fucosyltransferase—a large-scale preparation and characterization of recombinant human FUT8. *Glycobiology* 2006;16:333–342.
- Iwasaki H, Akashi K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* 2007;26:726–740. [PubMed: 17582345]
- Jeffrey KL, et al. Positive regulation of immune cell function and inflammatory responses by phosphate PAC-1. *Nature Immunol* 2006;7:274–283. [PubMed: 16474395]
- Klein, J.; Hořejší, V. *Immunology*. 2nd Ed.. Blackwell; Oxford: 1997.
- Koivunen P, Hirsilä M, Kivirikko KI, Myllyharju J. The length of peptide substrates has a marked effect on hydroxylation by the hypoxia-inducible factor prolyl 4-hydroxylases. *J. Biol. Chem* 2006;281:28712–28720. [PubMed: 16885164]
- Kovanen PE, et al. T-cell development and function are modulated by dual specificity phosphatase DUSP5. *J. Biol. Chem* 2008;283:17362–17369. [PubMed: 18430737]
- Lai AB, Kondo M. T and B cell differentiation from hematopoietic stem cell. *Semin. Immunol* 2007;20:207–212. [PubMed: 18583148]
- Latinovic-Golic S, Walch M, Sundstrom H, Dumrese C, Groscurth P, Ziegler U. Expression, processing and transcriptional regulation of granulysin in short-term activated human lymphocytes. *BMC Immunol* 2007;8:9. [PubMed: 17596262]
- Li J-H, Shao J-Z, Xiang L-X, Wen Y. Cloning characterization and expression analysis of pufferfish interleukin-4 cDNA: the first evidence of Th2-type cytokine in fish. *Mol. Immunol* 2007;44:2078–2086. [PubMed: 17084456]
- Lindman, HR. *Analysis of Variance in Complex Experimental Designs*. W.H. Freeman; San Francisco: 1974.
- Lund RJ, et al. Genome-wide identification of novel genes involved in Th1 and Th2 cell differentiation. *J. Immunol* 2007;178:3648–3660. [PubMed: 17339462]

- Minoguchi S, et al. RBP-L, a transcription factor related to RBP-Jκ. *Mol. Cell. Biol* 1997;17:2679–2687. [PubMed: 9111338]
- Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* 2008;132:631–644. [PubMed: 18295580]
- Rogge L, et al. Selective expression of an interleukin-12 receptor component by human T helper 1 cells. *J. Exp. Med* 1997;185:825–831. [PubMed: 9120388]
- Romagnani S. Type 1 T helper and type 2 T helper cells: functions, regulation and role in protection and disease. *Int. J. Clin. Lab. Res* 1991;21:152–158. [PubMed: 1687725]
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol* 1987;4:406–425. [PubMed: 3447015]
- Shi L, et al. Dok5 is substrate of TrkB and TrkC receptors and involved in neurotrophin induced MAPK activation. *Cell. Signal* 2006;18:1995–2003. [PubMed: 16647839]
- Strimmer K, von Haeseler A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol* 1996;13:964–969.
- Swofford, DL. PAUP*: Phylogenetic Analysis using Parsimony (*and Other Methods). Sinauer; Sunderland, MA: 1999.
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol* 2007;24:1596–1599. [PubMed: 17488738]
- Tanaka T, et al. Cloning and functional expression of the T cell activation antigen CD26. *J. Immunol* 1992;149:481–486. [PubMed: 1352530]
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Diggins DG. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–4882. [PubMed: 9396791]
- Watford WT, Moriguchi M, Morinobu A, O'Shea JJ. The biology of IL-12: coordinating innate and adaptive immune responses. *Cytokine Growth Factor Rev* 2003;14:361–368. [PubMed: 12948519]
- Wilsker D, Probst L, Wain HM, Maltais L, Tucker PW, Moran E. Nomenclature of the ARID family of DNA-binding proteins. *Genomics* 2005;86:242–251. [PubMed: 15922553]

Fig. 1.

(A) Cluster analysis of gene expression data in 11 human leukocyte types. The distance (Y-axis) is 1 minus the correlation coefficient. (B) Maximum parsimony tree based on 22,215 characters derived from gene expression data (tree length = 54,501 changes; overall consistency index = 0.678). The numbers on the branches represent the percentage of 1000 bootstrap samples supporting that branch. BC, B-cells; NK, natural killer cells; CMEM, central memory cells; EFM, effector memory cells; Th1, T-helper cells type 1; Th2, T-helper cells type 2; MAC, macrophages; CMAS, cord-blood mast cells; BAS, basophils.

Fig. 2. Neighbor-joining tree of vertebrate proline 4-hydroxylases, rooted with invertebrate homologs, based on the JTT amino acid distance at 380 aligned amino acid positions. Sequences are identified by species and by accession number. The numbers on the branches represent the percentage of 1000 bootstrap samples supporting that branch.

Fig. 3. Neighbor-joining tree of zinc finger protein 215 (ZNF215) and related zinc finger proteins, based on the JTT amino acid distance at 111 aligned amino acid positions. Sequences are identified by species and by accession number. The numbers on the branches represent the percentage of 1000 bootstrap samples supporting that branch; only values $\geq 50\%$ are shown.

Fig. 4. Interaction graphs, for significant gene by cell type interactions, illustrating mean expression scores in Th1 and Th2 cells of three pairs of related genes: (A) *ARID3A* and *ARID3B*; (B) *GFPT1* and *GFPT2*; and (C) *RBPJ* and *RBPJL*.

Fig. 5. Neighbor-joining tree of ARID family of vertebrates, rooted with invertebrate sequences, based on the JTT amino acid distance at 248 aligned amino acid positions. Sequences are identified by species and by accession number. The numbers on the branches represent the percentage of 1000 bootstrap samples supporting that branch.

Fig. 6. Neighbor-joining tree of *RBPJ* and *RBPJL* of vertebrates and invertebrates, based on the JTT amino acid distance at 399 aligned amino acid positions. Sequences are identified by species and by accession number. The numbers on the branches represent the percentage of 1000 bootstrap samples supporting that branch; only values $\geq 50\%$ are shown.

Table 1

Genes with significant difference in mean expression levels between T-helper cells type 1 and T-helper cells type 2 (Th1 and Th2) cells

Increased expression	Gene	Protein function	Orthologues found ¹						Gene duplication time ²
			Euth. Mam.	Non-Euth. Mam.	Bird	Amph.	Fish	Insect	
Th1	COL1A2	Collagen, type I, α 2	+	-	+	+	+	-	Before tetrapod-teleost MRCA (100)
	P4HA2	Procollagen-proline 2-oxaloglutarate 4-dioxygenase, α	+	+	+	+	+	-	Before tetrapod-teleost MRCA (99)
	DPP4	Dipeptidyl peptidase 4	+	+	+	+	-	-	After tetrapod-teleost MRCA (72)
	FUT8	α 1,6-fucosyltransferase	+	+	+	+	+	+	[None; ancient unduplicated]
	NFE2L3	Nuclear factor (erythroid-derived 2)-like 3	+	+	-	-	+	-	Before tetrapod-teleost MRCA (97)
	STEAP1	Six-transmembrane epithelial antigen of the prostate 1	+	+	+	+	+	-	Before tetrapod-teleost MRCA (93)
	LTA	Lymphotoxin α	+	+	-	+	+	-	Before amniote-amphibian MRCA (87)
	IL12RB2	Interleukin 12 receptor, β 2	+	+	+	-	+	-	Before tetrapod-teleost MRCA (67)
	DUSP5	Dual specificity phosphatase	+	+	+	+	+	-	Before tetrapod-teleost MRCA (100)
	APBB2	Amyloid β (A4) precursor protein-binding, family B, member 2	+	+	+	+	+	-	Before tetrapod-teleost MRCA (100)
	DOK5	Docking protein 5	+	+	+	-	-	-	Before tetrapod-teleost MRCA (99)
	CFH	Complement factor H	-	-	-	-	-	-	After primate MRCA (99)
	ARID3B	AT-rich interactive domain 3B	+	+	-	-	+	-	Before tetrapod-teleost MRCA (99)
	SHCBP1	SHC SH2-domain binding protein 1	+	+	+	+	+	-	Before tetrapod-teleost MRCA (100)
Th2	ZBTB32	Zinc finger and BTB domain containing 32	+	-	-	-	+	-	Before tetrapod-teleost MRCA (100)
	GNLY	Granulysin	+	+	-	-	-	-	[None; bird- and mammal-specific]
	LIF	Leukemia inhibitory factor	+	-	-	-	-	-	[None; eutherian mammal-specific]
	ANK1	Ankyrin 1, erythrocytic	+	-	-	-	+	-	Before tetrapod-teleost MRCA (100)
	PTGIS	Prostaglandin I2 synthase	+	+	-	-	+	-	Before tetrapod-teleost MRCA (100)
	LIMA1	LIM domain and actin binding 1	+	+	+	+	+	-	Before tetrapod-teleost MRCA (99)
	ZNF215	Zinc finger protein 215	+	-	-	-	-	-	After mammal MRCA (99)

Increased expression	Gene	Protein function	Orthologues found ¹						Gene duplication time ²
			Euth. Mam.	Non-Euth. Mam.	Bird	Amph.	Fish	Insect	
	IL26	Interleukin 26	+	+	-	-	-	-	After mammal MRCA (84)

Euth. Mam, Eutherian mammals; Non-Euth. Mam, noneutherian mammals (Metatheria or Monotremata); Amph, Amphibian.

¹ Orthologues identified by homology search and phylogenetic analysis.

² Numbers in parentheses represent bootstrap support for the branch in the phylogenetic tree establishing the gene duplication time. Only branches with at least 67% support are included.