

Chapter 8: General Discussion

An understanding of any complex system requires knowledge not only of the identities of all the components involved, but how they are inter-connected and realize their functions. This thesis has undertaken a two-pronged approach to extending our current understanding of the process of leukocyte cell-cell communication via the interactions of cell surface proteins. It addresses both the depth of our knowledge of the nature of these interactions and the breadth of our understanding of the number and identity of the molecules involved. The importance and inter-relatedness of these two features of biological systems has recently been highlighted by studies of T cell costimulation. Our understanding of this system has been transformed both by the discovery of novel costimulatory receptors (e.g. ICOS (Hutloff et al. 1999) and its ligand LICOS (Brodie et al. 2000)), and by careful analysis of the structures (Ikemizu et al. 2000; Schwartz et al. 2001; Stamper et al. 2001) and interactions of these molecules (Collins et al. 2002). The course of the experiments described in this thesis was heavily influenced by the completion of the human genome sequencing projects (Lander et al. 2001; Venter et al. 2001). As in many other laboratories, an attempt has been made to begin applying systematic approaches to identifying all the molecules involved in key immune processes or that are found in a given cell or protein family, using functional genomics-based approaches.

First considered was the nature of the interactions of cell surface molecules. The importance of the weak affinities of these interactions (compared to the affinities of interactions involving soluble proteins) is well established (van der Merwe and Barclay 1994). The affinity measured for CD150 self-association demonstrates that earlier reports of an extremely high affinity ($K_d \sim 0.1\text{nM}$ (Punnonen et al. 1997)), which challenged this understanding of the importance of weak interactions, did not reflect the monovalent association of two molecules of CD150 in solution. The importance of specificity in interactions at the cell surface is intuitively obvious, otherwise these interactions would be of no use in communicating to a

cell the state of its environment. Crystallographic studies of the rat CD48 ligand binding domain tested previous understanding of the structural basis for these properties, particularly an early paradigm established from studies of the CD2-CD58 interaction.

The structure of the ligand binding domain was determined and mutations of CD48 binding-site residues were characterised. This allowed the interaction of CD48 with CD2 to be modelled using the known structure of the human CD2-CD58 complex (Wang et al. 1999) and the extensive mutagenesis data available for the rat system. These studies confirmed the importance of low surface complementarity and mainly electrostatic contacts for weak but specific interactions at the cell surface, first seen for CD2-CD58. However, differences between the human and rat CD2-ligand interactions fit an emerging view of heterogeneity in the structural features of cell surface interactions that combine to generate the specific binding properties required for each interaction. In particular, the flatter binding face and the presence of only a few putative electrostatic contacts between the two rat proteins may accommodate flexibility in the orientation, shape and charge distribution of the binding surface of its ligands. This may explain the reduced specificity of rat CD48, which was shown to bind two rat CD244-homologues that differ radically in their binding faces, in addition to its established interaction with CD2 (van der Merwe et al. 1993). The ability of CD48 to bind all three of these ligands with similar affinities and kinetics demonstrates the flexibility of the interactions at the binding face. The structural heterogeneity of mechanisms ensuring weak binding at the cell surface seen here is consistent with studies of the TCR-peptide-MHC and CD80-CD152 interactions, wherein entropic effects on the one hand, and very high degrees of surface complementarity on the other, remain compatible with weak, specific recognition. In the first case, the flexibility of the binding face responsible for the entropic effects seen in TCR interactions is thought to increase the repertoire of potential antigen binding surfaces (Willcox et al. 1999). The high degree of surface complementarity in the CD80-CD152

interaction explains why it is one of the strongest interactions at the cell surface, a property likely to be linked to its inhibitory function (Stamper, 2001 #543}.

In addition to shedding light on the structural variability of cell surface interactions, the cloning of two different rat CD244 homologues (2B4.1 and 2B4.2, confirmed by Kumaresan et al. (Kumaresan et al. 2000; Kumaresan et al. 2000)), and the evidence for positive selection acting on their binding faces, suggest the possibility of the existence of a second ligand for CD244. Whether or not this is the case, these genes represent an example of a process by which the generation of proteins with new specificities could proceed very quickly after gene duplication. The eventual outcome could be similar to that of the human CD48 and CD58 genes, which bind two separate ligands (CD244 and CD2) although these are both bound by CD48 in rats and mice. However, in humans, CD48 and CD58 are extremely divergent and their DNA sequences do not support the suggestion that they are more related to each other than to other members of the CD2 subset. It is more likely that, over time, the two rat 2B4 proteins will continue to bind the same ligand(s) but with different properties, as is the case for CD80 and CD86 where such differences appear to be functionally significant (Collins et al. 2002).

Additional insights into the evolution of protein families were obtained by the identification of what we are now confident is the complete set of CD2 subset genes and pseudogenes. One completely novel gene (CD2F-10), two further genes whose cDNAs were already in the GenBank database (19A24 and BCM-like membrane protein), and at least four pseudogenes (CD2F-Ψ1-4) belonging to this subset, were identified. Several other groups have also attempted to identify new members of the CD2 subset using diverse methods, but have not identified any genes missed by our genomic screening, supporting the suggestion that we have identified the complete family. Of the sequences published, NTBA (Bottino et al. 2001) and

SF2000 (Fraser et al. 2002) are the same and represent the human orthologue of Ly108; BLAME (Kingsbury et al. 2001) is identical to the BCM-like membrane protein; CS1 (Boles and Mathew 2001) and CRACC (Bouchon et al. 2001) are the same as 19A24 (also published by the original sequencers (Murphy et al. 2002)) and novel Ly9 (Tovar et al. 2002) is its murine orthologue; and finally, CD84-H1 (Zhang et al. 2001) and SF2001 (Fraser et al. 2002) are the same as CD2F-10. It should shortly be possible to complete the analysis of receptor-ligand relationships within this important group of proteins using surface plasmon resonance techniques and this should provide additional evidence about the completeness of the subset. The absence of ‘orphan’ receptors, which are presently a feature of the expanded family, would strongly support the hypothesis that the subset is now complete.

The two rat 2B4-like genes and the pseudogene CD2F- Ψ 1 demonstrate two alternative fates of duplicated genes: evolutionary ‘birth’ and ‘death’ respectively. Both are recent events, but in one case the introduction of a mutation encoding a stop codon appears to have resulted in the gene becoming a non-functional pseudogene, whilst in the other positive selection has acted on the binding face of the encoded protein, possibly allowing the generation of new specificities. Gene duplication has long been considered a necessary source of material for the origin of evolutionary novelty (e.g. (Ohno et al. 1968)). Recent analyses of the human genome have suggested that this is the case, and that the majority of duplicated genes are rapidly silenced with only a small proportion surviving and evolving new functions (Lynch and Conery 2000). The CD2 subset genes represent a graphic example of these processes in recently duplicated sequences. In this light, it will be interesting to determine what the specificity of CD2F- Ψ 1 was before it gained a nonsense mutation, as this might explain why it was lost. Since its ligand binding domain is highly similar to NTBA (human Ly108) it may have bound the same ligand.

It is now possible to speculate about the overall process by which a family of proteins, such as the CD2 subset, can evolve. Assuming that the family began with a single homophilic ancestor, as is likely, some of the earliest duplications probably resulted in three loci: one containing the ancestor of CD2 and CD58, a second containing the ancestor of the pseudogenes containing exons CD2F-CX6 to 10 and the last containing the ancestor(s) of the remaining family members. These three sets of genes are clearly the most divergent and are encoded at separate loci on chromosome 1 in humans. The origins of CD58 present an interesting dilemma as this gene appears to be absent from rats and mice and yet must have been duplicated some time ago as it has low homology to the rest of the family (most similar to CD2). However, BLAST searches of the Celera mouse genome with human CD58 did not identify a pseudogene with homology to CD58, as might have been expected if this gene was lost after rats and mice diverged from primates.

It has been proposed recently that genomic location is the major factor in determining levels of gene expression, although transcriptional regulation determines which genes are actually expressed (Caron et al. 2001). Thus, gene duplication where the duplicate is encoded some distance from the original is less likely to result in a productively expressed gene that may then acquire a novel function, because such distant sites are less likely to be in expressed regions of the genome. This may explain why the duplications of genes *within* the main CD2 family locus (1q21.3-22) have resulted in many proteins with different specificities and only one pseudogene (CD2F-Ψ2), whereas duplications to adjacent but distinct loci (CD2F-Ψ1 at 1q23.2; CD2F- Ψ3 and Ψ4 and possibly others at 1q21.2-21.3) have resulted only in pseudogenes.

The large differences between the cytoplasmic domains of CD2 subset members may also have an evolutionary explanation. Several receptors from this subset transmit signals to lymphocytes via their cytoplasmic domains (Tangye et al. 1999; Latour et al. 2001; Sayos et al. 2001). Their ligands meanwhile (only known in two cases - CD58 for CD2 (Selvaraj et al. 1987) and CD48 for CD244 (Brown, M. H. et al. 1998)) are expressed on antigen presenting cells and are unlikely to transmit signals to these cells, as they are GPI-linked or have very short cytoplasmic tails. If, on binding its ligand, the ancestor of this subset transmitted signals into the cell on which it was expressed; evolution of the cytoplasmic domains of gene duplicates could make it possible for two different signals to be transmitted. If this occurred simultaneously with evolution of the extracellular domains so that the interaction was no longer homophilic, two cells expressing one duplicate each would be able to communicate in a unidirectional manner, which could be advantageous. In the case of leukocyte-activating receptors, the circulating leukocyte may need to respond when it encounters a particular ligand on another cell but that cell may not need to receive such a signal or be able to act on it. Thus, while purifying selection would maintain the signalling domain of the leukocyte receptor, that of the ligand would accumulate mutations and may eventually be lost by differential splicing or the introduction of nonsense mutations. Later duplication of such ancestral genes could result in a family of genes with similar extracellular domains wherein half were receptors capable of signalling and the others were non-signalling ligands. That this is the case for the CD2 subset is supported by the sequences and expression of the new members identified. BLAME (or BCM-like membrane protein, (Kingsbury et al. 2001)) and CD2F-10 (Fennelly et al. 2001) are expressed in antigen presenting cells and have very short cytoplasmic tails, while Ly108/NTBA (Peck and Ruley 2000; Bottino et al. 2001) and 19A24/CS1/CRACC (Boles and Mathew 2001; Bouchon et al. 2001) are expressed in lymphocytes and have longer cytoplasmic domains including possible tyrosine-containing signalling motifs.

Structural studies of shCD150 and c2B4.1 would have provided insights into the nature of the putative homophilic precursor of all these molecules and the mechanisms by which multiple ligands weakly and specifically bind single receptors, if only both proteins had produced well-diffracting crystals. The unstable nature of the chimeras of their ligand binding domains (particularly that of CD150) with rat CD2 domain 2 suggests an important *caveat* in the use of domain exchanges to investigate molecular functions. The inter-domain linker region of CD150 is no more divergent from the rat CD2 sequence than those of other members of the subset, so its replacement is unlikely to be responsible for this lack of stable folding. Rather, it seems likely that the stable folding of CD150 requires some specific contacts between the two domains (as seen in the structures of srCD2, cCD58 and cCD48) and that these are disrupted in the chimera. Thus although domain exchange experiments are widely used (e.g. (Wong, H. et al. 1991; George et al. 1993; Kassner et al. 1995; Ostrowski et al. 1995)) and can be informative (as in the case of cCD48 and cCD58), they should be interpreted cautiously.

While these studies of the CD2 subset were undertaken in order to obtain insights into the mechanisms of binding and evolution of cell surface molecules, the identification of new members of this well-studied family raised the question: how well is the leukocyte surface characterised? The transcriptome of a CTL clone was investigated using serial analysis of gene expression (SAGE). By analysing a single CTL clone, we have effectively described the complete transcriptome of a single cell (albeit to a certain depth), to our knowledge for the first time. Given the size of the library and assuming a binomial distribution of sampling at the 95% confidence level, every transcript constituting 0.008% of the CTL transcriptome (~22 copies per cell) is likely to have been detected (by sequencing its tag twice, confirming that it does not contain a sequencing error).

It was demonstrated that the levels of quantitative changes in the expression of a transcript are related to the function of the encoded protein and comparisons of lymphocyte transcriptomes on such a quantitative basis unexpectedly showed that CTL are more similar to NK cells than activated CD4⁺ helper T cells. In fact, the level of differences between the CD8⁺ and CD4⁺ T cell libraries is closer in magnitude to the differences seen between unrelated tissues (e.g. colon and ovary or CTL and colon). Thus, the differentiation of T lymphocytes into subsets, or the activation of these cells, or both, induces dramatic changes in their overall gene expression profile. Microarray data supports the possibility that the majority of these changes occur on T cell activation.

A complete analysis of the transcripts associated with every tag in the library is not feasible with current techniques. Therefore, a methodology was developed for processing SAGE data and identifying tags potentially of interest to immunologists i.e. those important for T cell immune functions. This method involves statistical and quantitative comparisons of the CTL SAGE library with those derived from unrelated tissues. In this way, 316 tags likely to be linked to transcripts directly involved in immune function were identified within a larger list of 899 tags that were significantly more abundant in the CTL library compared to a colon epithelium SAGE library. The method excluded few transcripts with known immune function but did exclude tags related to the increased proliferation and protein production of CTL compared to other tissues, which although important are not of interest in investigating mechanisms of immune action. The method therefore seems well-suited to targeting transcripts of interest. It needs to be born in mind, however, that the statistical tests used in this method mean that tags absent from non-immune libraries will only be detected as CTL-specific with 95% confidence if they constitute at least 0.014% of the CTL transcriptome (~42 copies per cell). Thus some poorly expressed transcripts, as well as transcripts not restricted

to the immune system, will be excluded despite being important to the immune response, in a library of this size.

Analysis of the tags specific to CTL compared to non-immune SAGE libraries shows that a large proportion (~40%) of the transcripts involved in CTL immune-function remain to be characterised, although a smaller proportion (15-25%) have not been fully sequenced. Many of these transcripts were characterised using readily available, web-based bioinformatic tools, identifying 98 that may have molecular functions relevant to the immune response of this CTL (e.g. cell surface molecules, signalling proteins or transcription factors). As this analysis only considered transcripts of moderate to high abundance, the proportion of important T cell molecules that are uncharacterised is likely to become significantly larger when lower-abundance genes are sampled.

The SAGE experiments described in this thesis should provide, along with the work of others, an expression database for human lymphocytes. This database will complement existing resources such as BodyMap and, because of the digital nature of SAGE data, allow detailed comparisons of future SAGE libraries. Thus, the SAGE expression database will continue to be enriched, providing a comprehensive overview of gene expression changes in leukocytes. In addition, because SAGE measures the abundance of all transcripts in the cell and is not biased towards known genes, as is presently the case for microarray experiments, the expression of newly described molecules (such as the new CD2 subset members described here) can be quickly assessed. The large amounts of structural and binding data produced for leukocyte cell surface molecules over recent years means that we are now close to having the necessary intellectual framework for formulating quantitative models of leukocyte function, particularly now that the techniques are at hand for identifying the complete sets of such molecules.