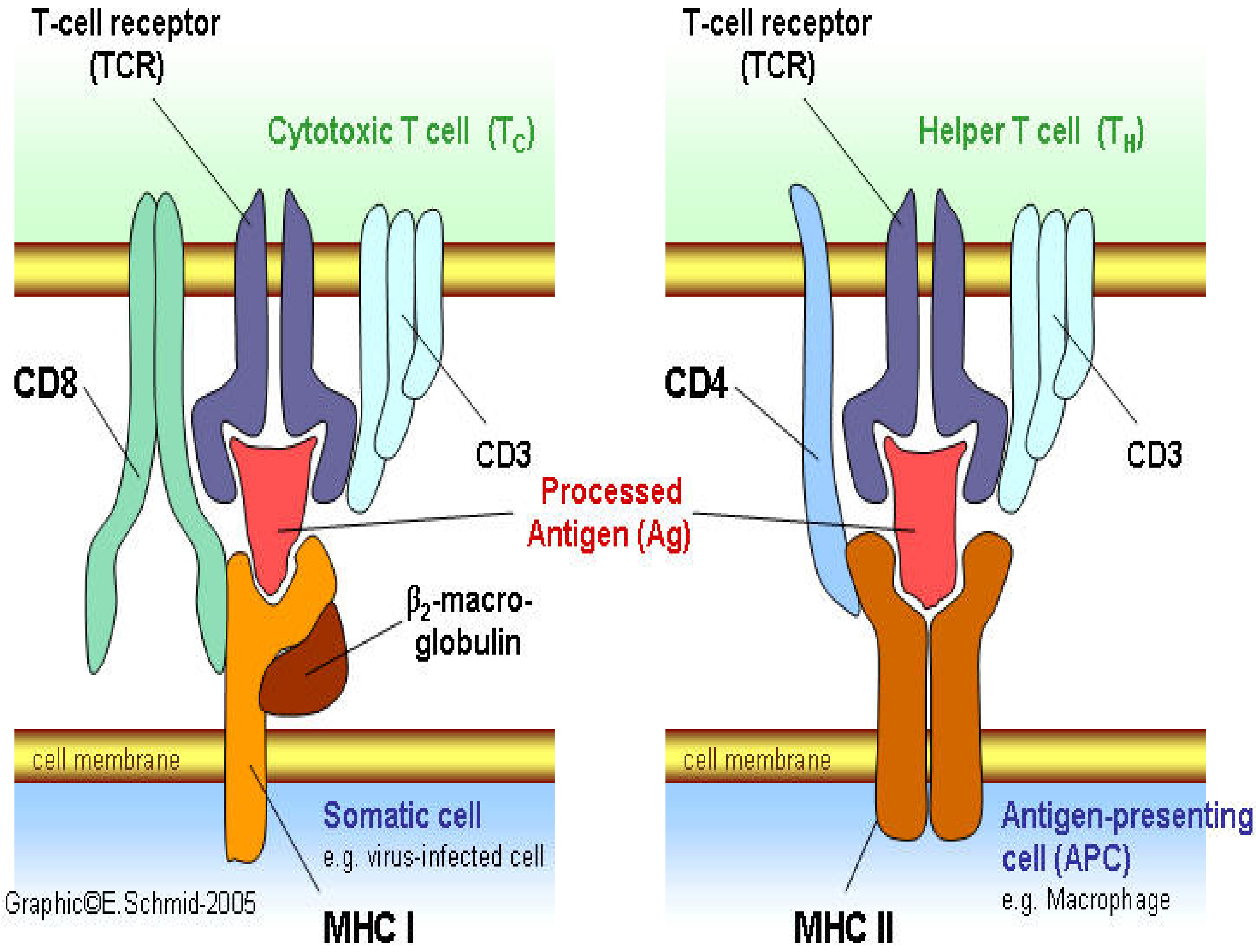
Peptide Antigen Display and Recognition: a New Fusion of Genomics and Proteomics?

T-cell antigen recognition



Motivation

- We are on the brink of learning to control critical aspects of the immune system in a very specific manner
- Accuracy will depend on vast and detailed knowledge of combinations and interactions that can only be obtained by a large-scale targeted genomics-proteomics project
- Combinatorics indicate peptide antigen display and T-cell recognition, key components of cell-mediated immunity, are amenable to a large-scale study approach
- Control of an individual’s cell-mediated immune response will have broad and profound applications in medicine, including infectious disease, cancer, autoimmunity, aging, and stem cell and transplant therapies

Scientific Challenges

- Given its genome and its RNA expression, predict the peptide antigens that will be displayed by a cell
- Predict whether a T-cell receptor will recognize a given displayed peptide antigen given the sequences of the receptor, the peptide, and the HLA molecule that displays it
- Create a master reference of human recurrent T-cell recognition events accompanied by selected live cell models
- Create a predictive computational model of cell-mediated immune response in humans

Key Data Elements

- Normal human germline and somatic whole genomes with accurate MHC sequencing
- Mutant cell panels (CRISPR + primary from patient tissues) with DNA & single cell RNA sequencing including T-cells at various states of activation
- Receptor repertoire from T-cell populations, coupled with cell state
- Large-scale experimental methods to determine peptide antigens displayed by professional antigen-presenting cells and other cells (including tumors), coupled with the T-cell receptors that recognize them (***key challenge)
- Data specific to local immune environments (e.g. lymph, gut, site of infection, autoimmune site, tumor); also microbiomes
- Time course during vaccination, infection, cancer

Possible Participants

- NHGRI, NIAID, NCI (leads)
- Other NIH institutes (maybe just make it pan-NIH)
- Cancer Immunotherapy Organizations (e.g. Alex's Lemonade Stand, …)
- Autoimmune Disease Organizations (Arthritis, Lupus, Psoriasis, Crohn's, Ankylosing Spondylitis, etc.)
- Stem cell and organ transplant organizations

Existing projects

- HudsonAlpha Repertoire 10K project http://www.r10k.org/R10K/About_R10K.html
- Human Immunology Project Consortium (NIAID) Stanford genome core (Davis)
- NIAID Immune Epitope Database http://www.iedb.org/ (~100k peptides, T and B-cell epitope prediction code, e.g. netMHC by S. Brunak et al., Tech. U. Denmark with predictions for 78 human HLA alleles, Multipred2 from Reinherz and Brusic)
- NIAID ImmPort (Contract to Northrup Grumman. This is a grab bag, certainly no Google)
- NCI TCGA (>1K full tumor genomes and more with RNA-seq)
- UCSC immunobrowser, unfunded prototype in its infancy, but includes full text literature search for immune-related DNA, RNA and protein sequences, track for IEDB, will link to https://genome.ucsc.edu/

Start with MHC Class I: more limited peptide antigen and TCR Diversities

- Typical MHC Class I displayed peptide is 9 amino acids (512 billion possibilities, small fraction of these are actually created that have enough affinity to nest in any particular one of the human ~1000 HLA alleles that display them).
- Total human T-cell TCR beta CDR3 receptor diversity in one person can approach 100 million (Robbins, 2009), theoretical diversity for TCR alpha-beta combinations for all typical human haplotypes combined is ~10^15 (Davis, 1988).
- There are many more combinations in, e.g., speech recognition, and computational methods worked there. Big data machine learning approaches will work for modeling cell-mediated immune response, if we can provide the appropriate experimentally-derived training data.
- If successful in this, we could try MHC Class II and the much more complex antibody/B-cell recognition problem.

Four key players in a recognition event

- Define an “MHC Class I recognition event” (for short just “recognition event”) as a 4-tuple:
  - Peptide of length 8 to 10
  - HLA allele that displays the peptide (represented both as a code and as an actual amino acid sequence, if the latter is available)
  - CDR3 amino acid sequence from TCR beta chain that recognizes the displayed peptide
  - Corresponding amino acid sequence from TCR alpha chain

Concrete proposal

- Collect 1 million recognition events. Realistically, for quite some time we will only be able to collect partial events in large quantities, e.g. peptide, or peptide+HLA molecule, or beta chain, or (hopefully soon) beta+alpha chain. Goal is to eventually get to 1M full 4-tuples.
- Create a stochastic “imputation” model that, given a partial event, can calculate posterior probabilities for the missing elements of the tuple.
- Train and test imputation on actual data. If we need other information in order to generate useful probabilities, add it. When imputation performance is reasonable, deploy it in research applications, and finally in clinical applications.