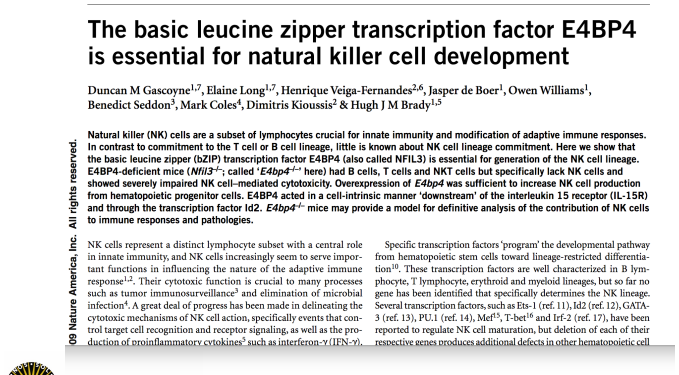
Today's computer exercise



The basic leucine zipper transcription factor E4BP4 is essential for natural killer cell development

Duncan M Gascoyne[1,7], Elaine Long[1,7], Henrique Veiga-Fernandes[2,6], Jasper de Boer[1], Owen Williams[1], Benedict Seddon[3], Mark Coles[4], Dimitris Kioussis[2] & Hugh J M Brady[1,5]

Natural killer (NK) cells are a subset of lymphocytes crucial for innate immunity and modification of adaptive immune responses. In contrast to commitment to the T cell or B cell lineage, little is known about NK cell lineage commitment. Here we show that the basic leucine zipper (bZIP) transcription factor E4BP4 (also called NFIL3) is essential for generation of the NK cell lineage. E4BP4-deficient mice (*Nfil3*$^{-/-}$; called '*E4bp4*$^{-/-}$ here) had B cells, T cells and NKT cells but specifically lack NK cells and showed severely impaired NK cell–mediated cytotoxicity. Overexpression of *E4bp4* was sufficient to increase NK cell production from hematopoietic progenitor cells. E4BP4 acted in a cell-intrinsic manner 'downstream' of the interleukin 15 receptor (IL-15R) and through the transcription factor Id2. *E4bp4*$^{-/-}$ mice may provide a model for definitive analysis of the contribution of NK cells to immune responses and pathologies.

NK cells represent a distinct lymphocyte subset with a central role in innate immunity, and NK cells increasingly seem to serve important functions in influencing the nature of the adaptive immune response[1,2]. Their cytotoxic function is crucial to many processes such as tumor immunosurveillance[3] and elimination of microbial infection[4]. A great deal of progress has been made in delineating the cytotoxic mechanisms of NK cell action, specifically events that control target cell recognition and receptor signaling, as well as the production of proinflammatory cytokines[5] such as interferon-γ (IFN-γ).

Specific transcription factors 'program' the developmental pathway from hematopoietic stem cells toward lineage-restricted differentiation[10]. These transcription factors are well characterized in B lymphocyte, T lymphocyte, erythroid and myeloid lineages, but so far no gene has been identified that specifically determines the NK lineage. Several transcription factors, such as Ets-1 (ref. 11), Id2 (ref. 12), GATA-3 (ref. 13), PU.1 (ref. 14), Mef[15], T-bet[16] and Irf-2 (ref. 17), have been reported to regulate NK cell maturation, but deletion of each of their respective genes produces additional defects in other hematopoietic cell

Universiteit Utrecht

Today's computer exercise

How old are NK cells?

How do we tackle this?

1. Blast human E4BP4 to see how many homologs we can find
2. Choose the "good" sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

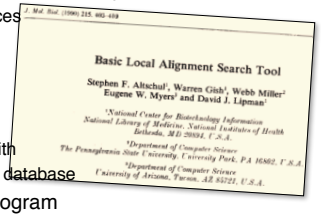
Universiteit Utrecht

1) Blast: Looking for homolog genes in other species

Universiteit Utrecht

Basic Local Alignment Search Tool (BLAST)

- Heuristic search algorithm
  - Makes shortcuts that are likely (but not guaranteed) to find the optimal hits
- BLAST finds good potential homologs at reasonable speed
  - 10-50x faster than if we align all the sequences
  - More than 100,000 queries per day on the NCBI BLAST server



Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2] Eugene W. Myers[2] and David J. Lipman[1]

[1]National Center for Biotechnology Information National Library of Medicine, National Institutes of Health Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science The Pennsylvania State University, University Park, PA 16802, U.S.A. [3]Department of Computer Science University of Arizona, Tucson, AZ 85721, U.S.A.

- Terminology:
  - Query: sequence we search the database with
  - Hit or Subject: similar sequence found in the database
- BLAST is the most used bioinformatics program
  - The BLAST article has been cited >54,000 times

Universiteit Utrecht

BLAST flavors

Query	Database	Blast
Protein	Protein	blastp
DNA	DNA	blastn
Translated DNA	Protein	blastx
Protein	Translated DNA	tblastn
Translated DNA	Translated DNA	tblastx

- Nucleotide-nucleotide searches
  - Nucleotide database, nucleotide query
  - blastn (default: W = 11 nucleotides)
    - Find homologous genes in different species
  - Megablast (default: W = 28 nucleotides)
    - Designed to efficiently find longer alignments between very similar nucleotide sequences
    - Best tool to find highly identical hits for a query sequence
    - For example: find sequences from the same species
  - Discontiguous Megablast
    - Uses discontiguous words (e.g. W = 11 nucleotides: **AT**-**GT**-**AC**-**CG**-**CG**-**T**)
    - For example, this can focus the search on codons (the third nucleotide of codons is less conserved due to the degeneracy of the genetic code → next slide)
    - Best tool to find nucleotide-nucleotide hits at larger evolutionary distances for protein-coding query sequences
- Protein-protein searches
  - Protein database, protein query sequences
  - blastp (default: W = 3 amino acids)
    - Find homologous proteins in different species

The alignment bit-score

- For a given query, we are mostly interested in finding good hits (highly similar, likely true homologs)
- We could estimate this based on a score derived only from the alignment like the bit-score or percent identity
  - … but the chance of finding a hit with a high score by random chance increases if you use a larger database
  - … so we have to correct for that

PREDICTED: n-acetyl-D-glucosamine kinase-like [Xenopus (Silurana) tropicalis]
Sequence ID: ref|XP_002938443.1| Length: 344 Number of Matches: 1

Range 1: 58 to 130 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
32.7 bits(73)	2.2	Compositional matrix adjust.	28/80(35%)	38/80(47%)	10/80(12%)

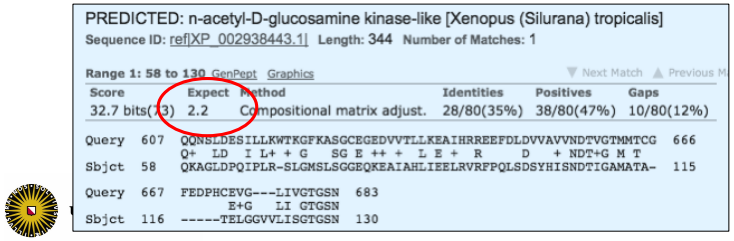
```
Query  607  QQNSLDESILLKWTKGFKASGCEGEDVVTLLKEAIHRREEFDLDVVAVVNDTVGTMMTCG  666
            Q+  LD  I L+ + G   SG E ++ +  L E +  R     D   + NDT+G M T
Sbjct  58   QKAGLDPQIPLR-SLGMSLSGGEQKEAIAHLIEELRVRFPQLSDSYHISNDTIGAMATA-  115

Query  667  FEDPHCEVG---LIVGTGSN  683
                  E+G   LI GTGSN
Sbjct  116  -----TELGGVVLISGTGSN  130
```

Expect value (E-value)

- E-value: how many times would you expect a hit this good, by random chance
  - Of course, this depends on the alignment score (*S*), the length of the query sequence (*m*), and the size of the database (*n*):

$$E = Kmne^{-\lambda S}$$

  - *K*: constant for search space scaling
  - λ: constant for substitution matrix correction



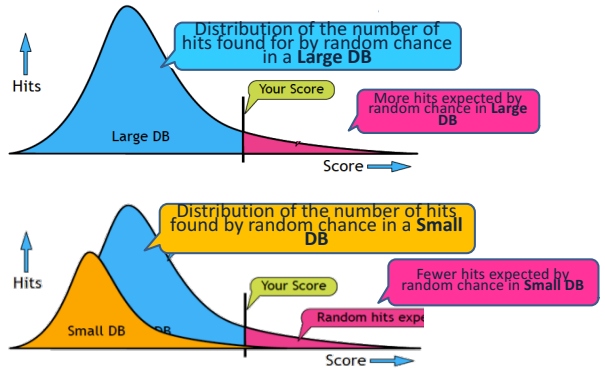
PREDICTED: n-acetyl-D-glucosamine kinase-like [Xenopus (Silurana) tropicalis]
Sequence ID: ref|XP_002938443.1| Length: 344 Number of Matches: 1

Range 1: 58 to 130 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
32.7 bits(73)	2.2	Compositional matrix adjust.	28/80(35%)	38/80(47%)	10/80(12%)

```
Query  607  QQNSLDESILLKWTKGFKASGCEGEDVVTLLKEAIHRREEFDLDVVAVVNDTVGTMMTCG  666
            Q+  LD  I L+ + G   SG E ++ +  L E +  R     D   + NDT+G M T
Sbjct  58   QKAGLDPQIPLR-SLGMSLSGGEQKEAIAHLIEELRVRFPQLSDSYHISNDTIGAMATA-  115

Query  667  FEDPHCEVG---LIVGTGSN  683
                  E+G   LI GTGSN
Sbjct  116  -----TELGGVVLISGTGSN  130
```


E-value differs for different databases (because of database size)



Today's computer exercise

How old are NK cells?

How do we tackle this?

1. Blast human E4BP4 to see how many homologs we can find
2. Choose the "good" sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

2) Which sequences should we choose? What is a good E-value?

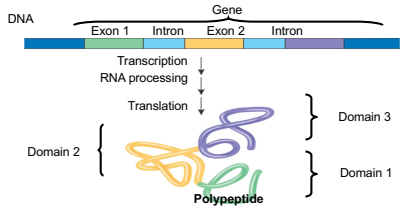
- This is very difficult to say!
- But as a rule of thumb:
  - E-value $<10^{-6}$ for nucleotide blast (blastn, megablast) are good
  - E-value $<10^{-3}$ for protein blast (blastp, blastx) are good
- If you want to be very sure that your query and hit sequences are homologs, you should only trust extremely low E-values
- … but sometimes you really have no other information about a protein, except a distant homolog with a very bad (= high, like 0.1) E-value

2) Which sequences should we choose?

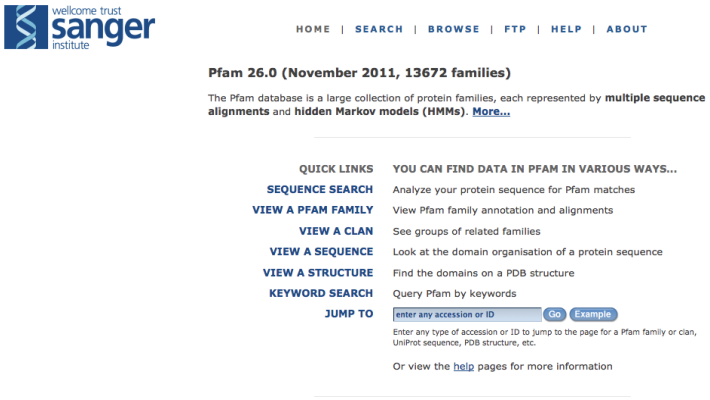
- Look for species that you did not yet select, and that you can also say something about the evolution of the NK cells. That is, instead of 10 mammals, choose few birds, reptils, fish, etc
- Check where the HSP, Blast hit, is: is it on a common domain, or on a specific domain. Hits on common domains do not indicate any homology.

Composition of the proteins: Domains

- Proteins often have a modular architecture
  - Consisting of discrete structural and functional regions called domains
- In many cases
  - Different exons code for the different domains in a protein



Campbell, Figure 17.12



Today's computer exercise

How old are NK cells?

How do we tackle this?

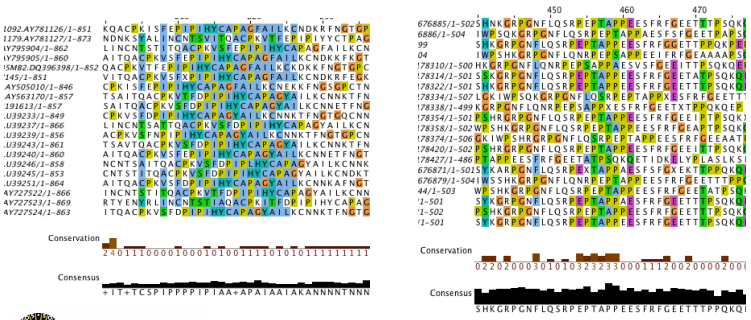
1. Blast human E4BP4 to see how many homologs we can find
2. Choose the "good" sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

Universiteit Utrecht

3) Sequence alignment: why should we align the sequences?

Which HIV-1 proteins should we use in vaccine?

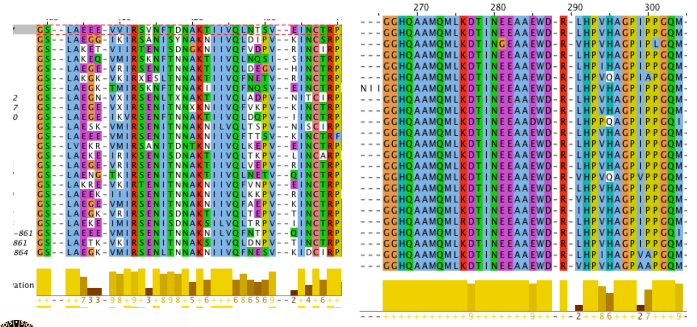
- ENV?
- Capsid??



Universiteit Utrecht

Which HIV-1 proteins should we use in vaccine?

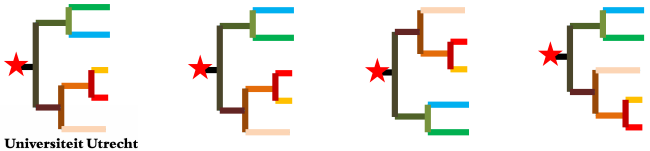
ENV? **Capsid??**



Universiteit Utrecht

4) Phylogenetic trees

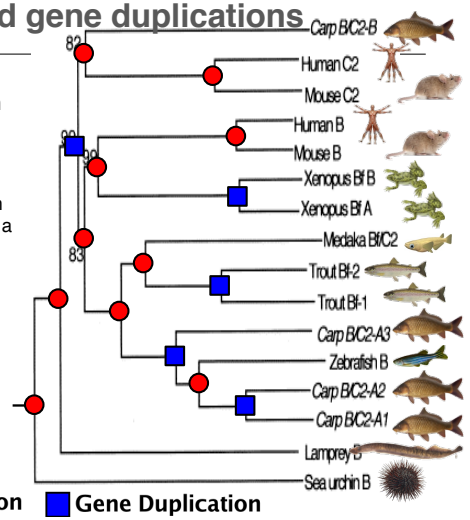
- A phylogenetic tree represents the phylogeny of species or sequences
  - Evolutionary signatures reveal the phylogenetic history
- Phylotenetic trees contain:
  - Present day sequences
  - Ancestral nodes
  - A root
- The same tree can be represented in many different ways:



Speciations and gene duplications

Phylogenetic trees of protein families contain two types of nodes

- Speciation nodes where the protein sequences in the tree diverged due to a speciation event
- Gene duplication nodes where the protein sequences in the tree diverged due to a gene duplication within one genome



Speciation Gene Duplication