Satellite information to the article "Ancient Phylogenetic Beginnings of Immunoglobulin Hypermutation"

Jaroslav Kubrycht, Karel Sigler, Michal Růžička, Pavel Souček, Jiří Borecky, Petr Ježek

Contents

WP1 Sequence alignments

WP1.1 Programs used in our paper

The programs were accompanied by manuals of an extent sufficient for use. In addition, several papers were important for our work (Altschul et al. 1997; Zhang et al. 1998; Schaffer

et al. 2001; Marchler-Bauer et al. 2002; Notredame et al. 2003; Bray and Pachter 2004; Simossis et al. 2005).

BLAST programs - www.ncbi.nlm.nih.gov/BLAST/

CLUSTAL W 1.82 - www.ebi.ac.uk/clustalw/

MUSCLE 2.01 - baboon.math.berkley.edu/mavid/

WP1.2 BLAST limits

Usual limits of Standard BLAST searches determined by word sizes 3 or 11 and Expect value 10 restricted all double-sequence comparison related protein and nucleotide searches, respectively. In addition, all our corresponding protein sequence searches (BLASTP, BLASTX, TBLASTN) were limited by the bit score value 22.3 which represents a minimal value of any BLASTN search, independently of the search strategy used. More special PSI-BLAST iterations were limited by Expect value 0.005 in accordance with established rules for this procedure. PHI BLAST search was limited only by word size 3 and Expect value 10 (i. e. by current limits for Standard Protein BLAST), without an additional bit score limit. The reason is in substantially lower PHI BLAST-derived bit scores with respect to the same Expect value than in other searches (due to ignored score of aa identities between PHI BLAST pattern sequence and both compared sequences). Therefore, we used only usual standard restriction by BLASTP Expect limit in the searches with short hybrid template derived query sequences.

WP2 Sequences and templates

WP2.1 CDR1-like segments of GCSAMS

In our previous paper we determined GCSAMS positions of CDR1-like segments related to heavy and light chains of Igs by using two independent procedures (Kubrycht et al. 2004). First of all we found positional agreement between segments similar to protein-kinase substrates and inhibitors located in N-terminal regions of Igs and GCSAMS, and overlapping with CDR1 of Ig light chains (Kubrycht and Sigler 1997; Kubrycht et al. 2002; Kubrycht et al. 2004). Secondly, we located corresponding positions when using conserved domain similarities and focused BLASTP similarities (generating among others also the important B1 template related GCSAMS segment denoted here as **GF** and restricted by GP:N132-203; Kubrycht et al. 2004). Both these procedures, together with the positions of conserved aa restricting CDR1 of Igs (Kubrycht et al. 2004; cf. also Potter et al. 1976 and Kabat et al.

1991), determined the searched positions. This means that we restricted two CDR1-like segments of GCSAMS, i. e. **GCSAMS(cdr1L.1)** related to Ig light chains (GP:aa22-34, N150-188) and **GCSAMS(cdr1.1.com)** (GP:aa30-34, N174-188) related to Ig heavy chains, and simultaneously included in both CDR1-like segments.

WP2.2 Templates derived from multiple sequence alignments

Templates derived from multiple sequence alignments (**MSA**-derived templates) contain only aa or alternative aa that occur in the same columns of evaluated sequence block more frequently than the same aa in the protein sequence encoded by random DNA (for corresponding approach see Kubrycht et al. 2002). In addition to this simple restriction, the hierarchy of template aa has been established to grade their reliability (three types of aa are considered in WP3.2 and WP3.3). For instance aa of all levels displayed in WP3.2 form a unique multi-variant MSA-derived template.

WP2.3 Literature context of selected reference sequences (RfS)

TCRL (T-cell receptor like protein) and **VpBLP** (V-pre B like protein) represent the proteins of the vertebrate jawless fish sea lamprey (Petromyzon marinus) origin, which does not contain true AR. In agreement with literature data (Pancer et al. 2004), unduplicated TCRL appears to be a diversification derivative of PPCAR gene and exhibits meanwhile the closest linkage to AR genes. According to the published ClustalW alignments, the sequence of lamprey VpBLP contains an Ig variable region with canonical V-frame and is similar to mammalian Vpre-B (Cannon et al. 2005). VpBLP is thus the first molecule with proposed close structural relationship to the B cell receptor complex in jawless vertebrates.

Though immunoglobulin W (**IgW**) and IgM VH gene families are both 460 million year old (Rumfelt et al. 2004), comparison of the shark sequences of IgW variable and constant domains shows homology greater than that found among the corresponding genes encoding IgM (Bernstein et al. 1996b). This increased homology suggests a closer relationship of IgW, detected exclusively in primitive Gnathostomata (sharks and lungfish) to primordial Ig genes (Schluter et al. 1997; Rumfelt et al. 2004). These facts predetermined also our decision about IgW not only as a source of representative RfS and but also as query sequences for additional selection of RfS (see Material and Methods).

The molecules of **SIRP** (signal-regulatory protein(s)) and **NITR** (novel imune-type receptor(s)) families represent homologues of AR. These molecules were possibly derived from late PPCAR gene close to AR (van den Berg et al. 2004).

WP3 Additional relationships concerning Table 4

WP3.1 Complete result of multiple sequence alignments partially displayed in Table 4

```
CLUSTAL W (1.82) multiple sequence alignment
ror             --NVTKYR---GQAVRIR-CEITGN--PIPNYS----WYKD--DVIINNDPSDRRMGHKP 46
apCAM           PPTIHPDNPKVGDEVKIT-CQATGV--PPPTYQ----FKKG--DVMVTDEMVNN------ 45
tractin         --PVSPMKLTEGKNTVVQ-CSVFGA--PKPLVT----CLRN--DTVISGDRFKVD----- 44
leechCAM        -------------------CKVDGL--PKPEVS----WRYK--DRKLDSERYTKV----- 28
RTK             ----PGLVVREGSEVIVLTCEVYGY--PRDSSPP--MWSSP--GRNLESGRFITTPRYTN 50
GCSAMSd2        ----SGLVVREGSEVIVLTCEVYGY--PRDSSPP--MWSSP--GRNLESGRFNITPRYTG 50
CG14162d2       ----PDLHVDKGSTINLT-CTVKFS--PEPPAYI--FWYHH--EEVINYDSSRGGVSVIT 49
CG14469-PA      -----ELHVDMGSTINLV-CIIEKS--PTPPQYV--YWQKN--DRLINYVDSRRDITIET 48
CG6867-PA       ------------SSATLE-CLVEAF--PEAIRY----WERAYDGKILDPSDKYGIESY-- 39
MDM             ------------SAVTLE-CVFSGYD-PKAPHFPKITWRTADGKVIINDSKYTLS----- 41
IgW             ---PESVVKKPGETVRLS-CGVTGFD-IDTHYIT---WVKQVPGKGLEWLLYHDSRPQEF 52
VCBP5           ----------------------AGN--PEPENIT----ISP----SFDGRVSLDAD---- 24
VDB             ------LALQPGDNANLQ-CNYTTTQSPSTGSLS---WSFNNGSGDVTFYQRLGSTEIPS 50
GCSAMSd1        ---PVSPDLSQPHSVTLT-CSAASP--PARGYQYQWQWRRN--GTLLSNTHTRFSITP-- 50
NITR2           --------VQPGDSVTLN-CTIHTKTCSGDHSVY---WFRHRSGESHPGIIYTHGDRSDQ 48


ror             T----AWGS------------RLKINDVRPSDSAVYTCKAENDFGNEETSGSLTVL- 86
apCAM           -------G-------------VLTINPLKTTDQATYTCIATNKGGFAESSNTLDV-- 80
tractin         -----THGN-------------LLVSNLQLSDSGNYICFASNKFGNDSVGANLIV-- 81
leechCAM        -----EDG--------------LLIKNITTEDDGIYQCSAN--VEND---------- 54
RTK             TL---SNGSVSSS--EKVALSQLTIFNVTAADEGEYTCSVD---GESASF------- 92
GCSAMSd2        TL---SNGSVSSS--DKVALSQLTIFNVTVADEGEYTCSVD---GESASF------- 92
CG14162d2       -----EKGDV--------TTSFLLIQNADLADSGKYSCAPS---NADVASVRVHVLN 90
CG14469-PA      -----TPGPR--------TQSRLIIREPQVTDSGNYTCSAS---NTEPAS------- 82
CG6867-PA       -----PEGFK--------TTMRLTISNLRKDDFGYYHCVAR----NE---------- 69
MDM             -----SDGR------------ALTIRSVTGSDQKKYYCSASNSAGFAGPHAVFLNV- 80
IgW             APG--IEGRFTPS--VVSNTAYLEITSLSVTDTAIYYCA------------------ 87
VCBP5           -----GSGSFT---------PTLTITDIRPSDSGRYWCAPDISEDYSNLG------- 60
VDB             AG---YQGRVTFIGDLSTGVANIRLSNMQTEDSGSYTCSVTVFGDGQDSQSITVTV- 103
GCSAMSd1        -----STNTHS---------SSLVISGLRYSDAGDYMCTVE---------------- 77
NITR2           CEKSPEAGSPTQS-----CVYNLPKRNLTLSDAGTYYCAVASCGEILFGNRTKLDV- 99
                       .              :        *   * *

MUSCLE (2.01) multiple sequence alignment
ror             -----NVTKYRGQA-VRIRCEITGNPIP--N---YSWYKDDVIINNDPSD---RRMGHKP 46
CG14162d2       ----PDLHVDKGST-INLTCTVKFSPEPPAY---IFWYHHEEVINYDSSR---GGVSVIT 49
CG6867-PA       ------------SS-ATLECLVEAFPEA--I---RYWERAYDGKILDPSD---KYGIESY 39
tractin         --PVSPMKLTEGKN-TVVQCSVFGAPKP--L---VTCLRNDTVISGDRFK---VDT---- 45
leechCAM        -------------------CKVDGLPKP--E---VSWR--YKDRKLD-SE---RYT---- 26
apCAM           PPTIHPDNPKVGDE-VKITCQATGVPPP--T---YQFKKGDVMVTDEMVN---NGV---- 47
CG14469-PA      -----ELHVDMGST-INLVCIIEKSPTPPQY---VYWQKNDRLINYVDSR---RDITIET 48
MDM             -------------SAVTLECVFSGYDPKAPHFPKITWRTADGKVIINDSK---------- 37
RTK             ----PGLVVREGSEVIVLTCEVYGYPRDSSP---PMWSSPGRNLESGRFITTPRYTNTLS 53
GCSAMSd1        ---PVSPDLSQPHS-VTLTCSAASPPARGYQ---YQWQWRRNGTLLSNTH---TRFSITP 50
GCSAMSd2        ----SGLVVREGSEVIVLTCEVYGYPRDSSP---PMWSSPGRNLESGRFNITPRYTGTLS 53
VDB             ------LALQPGDN-ANLQCNYTTTQSPSTG---SLSWSFNNGSGDVTFY---QRLGSTE 47
IgW             ----PESVVKKPGETVRLSCGVTGFDIDTHY---ITWVKQVPGKGLEWLL---YHDSRPQ 50
NITR2           --------VQPGDS-VTLNCTIHTKTCSGDH--SVYWFRHRSGESHPGIIYTHGDRSDQC 49
VCBP5           ----------------------AGNPEP------------ENITISPSFD---GRVSLDA 23


ror             TAWGS---RLK-------------INDVRPSDSAVYTCKAE--NDFGNEETSGSLTVL 86
CG14162d2       EKGDVTTSFLL-------------IQNADLADSGKYSCAPSNADVASVRVHVLN---- 90
CG6867-PA       PEGFKTTMRLT-------------ISNLRKDDFGYYHCVARNE--------------- 69
tractin         ------HGNLL-------------VSNLQLSDSGNYICFASNKFGNDSVGANLIV--- 81
leechCAM        ----KVEDGLL-------------IKNITTEDDGIYQCSANVEND------------- 54
apCAM           ---------LT-------------INPLKTTDQATYTCIATNKGGFAESSNTLDV--- 80
CG14469-PA      TPGPRTQSRLI-------------IREPQVTDSGNYTCSASNTEPAS----------- 82
MDM             -YTLSSDGRAL------------TIRSVTGSDQKKYYCSASNSAGFAGPHAVFLNV-- 80
RTK             NGSVSSSEKVA--------LSQLTIFNVTAADEGEYTCSVDGESASF----------- 92
GCSAMSd1        STNTHSSSLV--------------ISGLRYSDAGDYMCTVE----------------- 77
GCSAMSd2        NGSVSSSDKVA--------LSQLTIFNVTVADEGEYTCSVDGESASF----------- 92
VDB             IPSAGYQGRVTFIGDLSTGVANIRLSNMQTEDSGSYTCSVTVFGDGQDSQSITVTV-- 103
IgW             EFAPGIEGRFT--PSVVSNTAYLEITSLSVTDTAIYYCA------------------- 87
NITR2           EKSPEAGSPTQ------SCVYNLPKRNLTLSDAGTYYCAVASCGEILFGNRTKLDV-- 99
VCBP5           DGSGSFTPTLT-------------ITDIRPSDSGRYWCAPDISEDYSNLG-------- 60
                                         *   * *
```

WP3.2 Table 4 in more detail

Table 4. Conserved regions in the selected non-vertebrate IGv-related segments [a,b]

```
CLUSTAL W (1.82) multiple sequence alignment
ror             T----AWGS-----------RLKINDVRPSDSAVYTCKAENDFGNEETSGSLTVL- 86*177
apCAM           -------G------------VLTINPLKTTDQATYTCIATNKGGFAESSNTLDV-- 80*300
tractin         -----THGN-------------LLVSNLQLSDSGNYICFASNKFGNDSVGANLIV-- 81*516
leechCAM        -----EDG--------------LLIKNITTEDDGIYQCSAN--VEND---------- 54*200
RTK             TL---SNGSVSSS--EKVALSQLTIFNVTAADEGEYTCSVD---GESASF------- 92*129
GCSAMSd2        TL---SNGSVSSS--DKVALSQLTIFNVTVADEGEYTCSVD---GESASF------- 92*206
CG14162d2       -----EKGDV--------TTSFLLIQNADLADSGKYSCAPS---NADVASVRVHVLN 90*273
CG14469-PA      -----TPGPR--------TQSRLIIREPQVTDSGNYTCSAS---NTEPAS------- 82*140
CG6867-PA       -----PEGFK--------TTMRLTISNLRKDDFGYYHCVAR----NE---------- 69*612
MDM             -----SDGR-----------ALTIRSVTGSDQKKYYCSASNSAGFAGPHAVFLNV- 80*229
IgW             APG--IEGRFTPS--VVSNTAYLEITSLSVTDTAIYYCA------------------ 87*112
VCBP5           -----GSGSFT---------PTLTITDIRPSDSGRYWCAPDISEDYSNLG------- 60*281
VDB             AG---YQGRVTFIGDLSTGVANIRLSNMQTEDSGSYTCSVTVFGDGQDSQSITVTV- 103*140
GCSAMSd1        -----STNTHS---------SSLVISGLRYSDAGDYMCTVE---------------- 77*82
NITR2           CEKSPEAGSPTQS-----CVYNLPKRNLTLSDAGTYYCAVASCGEILFGNRTKLDV- 99*262
common aa(.,:,*)       .              :         *   * *
hoaa                                   L-I-N----D-G-YTC-A--
mhoaa                                  -----------------A--
                                       -----------------V--
CRCL(*)                                *****************
IgW/CRCL                               LEITSLSVTDTAIYYCAR
PBB+                                   ARFSSLTGYDLEWTYCAR

MUSCLE (2.01) multiple sequence alignment
ror             TAWGS---RLK-------------INDVRPSDSAVYTCKAE--NDFGNEETSGSLTVL 86*177
CG14162d2       EKGDVTTSFLL-------------IQNADLADSGKYSCAPSNADVASVRVHVLN---- 90*273
CG6867-PA       PEGFKTTMRLT-------------ISNLRKDDFGYYHCVARNE--------------- 69*612
tractin         ------HGNLL-------------VSNLQLSDSGNYICFASNKFGNDSVGANLIV--- 81*516
leechCAM        ----KVEDGLL-------------IKNITTEDDGIYQCSANVEND------------- 54*200
apCAM           ---------LT-------------INPLKTTDQATYTCIATNKGGFAESSNTLDV--- 80*300
CG14469-PA      TPGPRTQSRLI-------------IREPQVTDSGNYTCSASNTEPAS----------- 82*140
MDM             -YTLSSDGRAL------------TIRSVTGSDQKKYYCSASNSAGFAGPHAVFLNV-- 80*229
RTK             NGSVSSSEKVA--------LSQLTIFNVTAADEGEYTCSVDGESASF----------- 92*129
GCSAMSd1        STNTHSSSLV--------------ISGLRYSDAGDYMCTVE----------------- 77*82
GCSAMSd2        NGSVSSSDKVA--------LSQLTIFNVTVADEGEYTCSVDGESASF----------- 92*206
VDB             IPSAGYQGRVTFIGDLSTGVANIRLSNMQTEDSGSYTCSVTVFGDGQDSQSITVTV-- 103*140
IgW             EFAPGIEGRFT--PSVVSNTAYLEITSLSVTDTAIYYCA------------------- 87*112
NITR2           EKSPEAGSPTQ------SCVYNLPKRNLTLSDAGTYYCAVASCGEILFGNRTKLDV-- 99*262
VCBP5           DGSGSFTPTLT-------------ITDIRPSDSGRYWCAPDISEDYSNLG-------- 60*281
common aa = P1(*)                              *   * *
P2(*)                                          *   * * *
hoaa                                   --I-N----D-G-YTC-A-N
rhoaa                                  L-I-N----D-G-YTC-A-N
mhoaa                                  -----------------A--
                                       -----------------V--
tmaa                                   --I-NLT-SDSG-YTCSA-N
                                       -----V----------AV--
mtmaa                                  ------R-A--------A--
                                       ------Q-T--------V--
qaa                                    LT-S---T---A--Y---S-
                                       -------V----------D-
CR(*)                                    ***************
HTS1                                   LTISNLBVSDSGXYTCSAZN
IgW/PBB                                  ITSLSVTDTAIYYCAR
PBB/IgW                                  FSSLTGYDLEWTYCAR
PBB/tmaa                                 FSSLTGYDLEWTYCAR
```

[a]Displayed C-terminal parts of selected PSI-BLAST-derived segments contain the conserved regions (**CR**; underlined) of high similarity, without gaps and of equal overlapping in both different multiple sequence alignments. These regions possibly correspond to the primordial building block of variable immunoglobulin heavy chains described by Ohno et al. (1982). For additional important relationships see last section of Results. For more detailed description of the hybrid template formation see WP2.2, WP3.3 and last subsection of Material and Methods.

[b]**hoaa** (see below) of both alignments restrict the regions with compared aa. The first and the second numbers after the sequences denote C-terminal positions of displayed peptide chains in PSI-BLAST derived segments and whole molecules, respectively. CRCL - Clustal W derived conserved regions; qaa - questionable aa; **hoaa** - high-occurrence aa limited by a length equivalent value of three; IgW/PBB, IgW/CRCL - completed IgW segments similar to PBB (C-terminal arginin was not present in our PSI-BLAST searches) or of CRCL extent, respectively; **HTS1** - a hybrid template sequence constructed here according to last subsection of Material and Methods (**B** = Q,R,T and **Z** = D,S); mhoaa, mtmaa - pairs of frequent single mutants achieving together the required limits for tmaa or hoaa, respectively, including sometimes additional improved evaluation of pair occurrences; **P1**, **P2** - aa of PHI BLAST pattern denoted by * (last subsection of Results); **PBB** - Ohno´s modern intact primordial building block; **PBB+** - a sequence of CRCL extent proposed based on current primary structure of tandem repeats, i. e. chain composed of C-terminal PBB dipeptide followed by whole PBB repeat; **rhoaa** - hoaa reevaluated in accordance with PHI BLAST search (see also last subsection of Results); tmaa - template motif aa (see also the following section). For abbreviations of molecular names see WP5.1.

WP3.3 Hierarchy of coincident aa used in Table 4[a]

	length equivalent	level of aa coincidence
questionable aa	1.0 -1.5	aa occur slightly more than randomly
template motif aa	> 1.5	aa are more likely similar than dissimilar
high occurrence aa	> 3.0	corresponding abstract non-integer height of model twin column (**MTC**) containing exclusively identical/similar aa is larger than three, e.g. a double sequence identity/similarity have to be confirmed by more than one additional adequate random event in each site of aa positivity

[a]The presented approach possibly represents a reasonable cut-off limit alternative of entropy-related sequence logos (Schneider and Stephens 1990; Kubrycht and Sigler, manuscript in preparation). For additional comments to our approach see Kubrycht et al. 2002. For comments to motif-related length equivalent limit see Kubrycht et al. 2004.

WP4 Molecules of closest phylogram relationships

WP4.1 Principles

The short lengths of selected conserved regions, and the number of gaps together with enhanced variability of the longer IGv-related segments (derived by PSI-BLAST iteration) represent problematic properties complicating the use of usual phylogenetic analysis. Hence, if obtained by different alignment methods, these properties cause more variable multiple sequence alignment (**MSA**) records than is usual. Our approximate solution of this diminished phylogram stability is based on simplified frequency analysis of the closest linkages in original and differentially extended phylograms (i. e. phylograms defined on modified sets differentially extended by the addition of one important reference sequence; for details see below). Since the corresponding diminished phylogram stability indeed exists (WPT1 in WP4.3), we cannot expect a usual power of phylogenetic analysis when comparing sequences related to variable Ig domains. Therefore, our approach represents more likely one parameter of a multi-parameter evaluation than a decisive or main procedure.

WP4.2 Procedures necessary for generation of phylograms

WP4.2.1 Sequence sources of phylograms. We came from three sequence sets including PBB related regions (**PRR**), CRCL (for both types of segments see Table 4 and WP3.2) and PSI-BLAST derived segment(s) (**PBDS**; WP3.1). PBB or PBB+ segments (WP3.2) were added to sets of PRR or CRCL, to derive modified (alternative) sets, respectively. Similarly, we used separately each of two IGv domains (cd00099 and smart00406) to gain two alternative sets of PBDS.

WP4.2.2 Multiple sequence alignments. CLUSTAL W 1.82 and MUSCLE 2.01, on line available via pre-adjusted web page (for program addresses see WP1.1), were uploaded by FASTA formatted sequence sets present in *.txt files. The results were obtained directly or via e-mail.

WP4.2.3 Construction of phylograms. All phylograms were “phylip“ trees determined with the help of CLUSTAL W. Usual CLUSTAL W-derived phylograms were constructed from CLUSTAL W-derived MSA-record immediately after sequence comparison. On the other hand, MUSCLE-derived PHYLIP trees were derived from MSA-record using pre-reset CLUSTAL W. All four accessible variants of “phylip“ trees, reflecting tolerated/untolerated gaps and the presence/absence of Kimura´s distance correction were generated with all ten sequence sets.

WP4.3. Phylogram records

WP4.3.1 Some relationships between phylogram groups. Though CRCL and PRR resulted primarily from sequence similarities of CLUSTAL W MSA and both MSA records, respectively (Table 4; WP3.2), both MUSCLE- and CLUSTAL W-derived phylograms of conserved PRR and CRCL were identical. The effect of one reference sequence addition was lower in phylograms related to conserved regions (**CR**, i. e. CRCL or PRR) than in PBDS-related phylograms. The closest linkages of CR-related phylograms and MUSCLE-derived ones on sets including PBDS (**SPBDS**; WP3.1) were similar in contrast to markedly different CLUSTAL W-derived phylograms on SPBDS.

WP4.3.2 Approximate evaluation of phylogram similarities. Only the items present in the three closest branches (but not more) were included in our evaluation of the forty resulting phylograms. This means that only the phylogram linkages of orders 1, 1-2, 2, 2-3 and 3 (the orders 1-3 were not observed) were included in the enumeration of frequencies. The results of our approximate frequency-based evaluation of phylograms are demonstrated on associated Table WPT1.

WPT1. Sequences of close phylogram relationship to reference IgW segments

WPT1A. Map of close phylogram linkages

seq set	CU					CT					NU					NT				
	1	1-2	2	2-3	3	1	1-2	2	2-3	3	1	1-2	2	2-3	3	1	1-2	2	2-3	3
conserved regions																				
PBBS	apCAM	–	MDM	–	CG14469	apCAM	–	MDM	–	CG14469	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–
PBBS+P	MDM	–	NITR2	–	–	MDM	–	NITR2	–	–	MDM	–	apCAM	–	–	MDM	–	apCAM	–	–
CRCLS	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–
CRCLS+P+	MDM	–	NITR2	–	GCSd1	MDM	–	NITR2	–	GCSd1	MDM	–	apCAM	–	–	MDM	–	apCAM	–	–
segments pre-aligned by PSI-BLAST iteration																				
- phylograms derived based on CLUSTAL W MSA																				
SPBDS	VCBP5	–	–	apCAM ror	–	apCAM	–	ror	–	VCBP5	VCBP5	–	–	apCAM ror	–	apCAM	–	–	ror VCBP5	–
SPBDS+IGv4	CG6867	–	GCSd1	–	NITR2	CG6867	–	CG14162	–	CG14469	CG6867	–	–	GCSd1 NITR2	–	CG6867	–	–	–	–
SPBDS+IGv9	CG6867	–	–	–	–	–	–	–	–	–	CG6867	–	–	–	–	CG6867	–	–	–	–
- phylograms derived based on MUSCLE MSA																				
SPBDS	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–	apCAM	–	MDM	–	–
SPBDS+IGv4	–	NITR2 VDB	–	–	GCSd1	apCAM	–	–	ror VCBP5	–	–	NITR2 VDB	–	–	–	NITR2	–	–	GCSd1 CG6867	–
SPBDS+IGv9	–	apCAM ror	–	–	–	MDM	–	–	apCAM ror	–	–	GCSd1 NITR2	–	–	–	MDM	–	–	–	

WPT1B. Comparison of selected segments

	PBBS or PBBS+P						CRCL or PBSS+P+						SPBDS/CLUSTAL W MSA						PBDS/MUSCLE MSA						all					
	1	1-2	2	2-3	3	s1	1	1-2	2	2-3	3	s2	1	1-2	2	2-3	3	s3	1	1-2	2	2-3	3	s4	1	1-2	2	2-3	3	sum
1. apCAM	4	–	2	–	–	6	4	–	2	–	–	6	2	–	–	2	–	4	5	1	–	1	–	7	15	1	4	2	–	23
2. MDM	4		4	–	–	8	4	–	4	–	–	8	–	–	–	–	–	–	2	–	4	–	–	6	10	–	12	–	–	22
3. NITR2	–	–	2	–	–	2	–	–	2	–	–	2	–	–	–	1	1	2	1	3	–	–	–	4	1	3	4	1	1	10
4. CG6867	–	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	7	–	–	–	1	–	1	7	–	–	1	–	8
5. ror	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	3	–	4	–	1	–	2	–	3	–	1	1	5	–	7
6. GCSd1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	1	–	2	–	1	–	1	1	3	–	1	1	2	3	7
7. VCBP5	–	–	–	–	–	–	–	–	–	–	–	–	2	–	–	1	1	4	–	–	–	1	–	1	2	–	–	2	1	5
8. CG14469	–	–	–	–	2	2	–	–	–	–	–	–	–	–	–	–	1	1	–	–	–	–	–	–	–	–	–	–	3	3
9. VDB	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–	–	–	2	–	2	–	–	–	2
10. CG14162	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	1	–	–	–	–	–	–	–	–	1	–	–	1

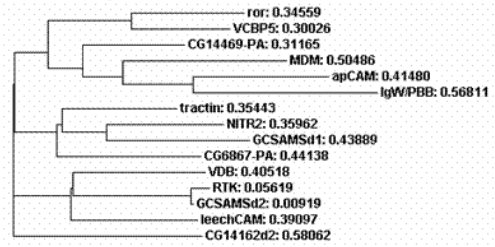
[a]apCAM, MDM and CG6867 can be frequently found in the closest branch to reference IgW segments. Only apCAM and NITR2 occur in all four displayed series. For additional information see also last section of Results and WP4.3.1.

[b]**Sequence sets**: PBBS, CRCLS, SPBDS - sets of PBB-related regions (also PRR), CRCL and PBDS, respectively. **Segment origin:** CG14162 - CG14162d2; CRCL - conserved region(s) derived by CLUSTAL W MSA (Table 4; WP3.2); PBDS - PSI-BLAST derived segment(s) (WP3.1); GCSd1 - GCSAMSd1 (for explanation of this and other abbreviations see Tables 2 and 3 or WP5.1). **Reference sequences used for generation of alternative sequence sets**: IGv4, IGv9 - variable Ig domains smart00406 and cd00099, respectively; P - PBB sequence (PBB is Ohno's primordial building block of heavy variable Ig genes; Ohno et al. 1982; WP3.2); P+ - PBB+ sequence (see WP3.2). **Other abbreviations:** 1, 1-2, 2, 1-3, 3 - for explanation see WP4.3.2; components of expressions CT, CU, NT, NU: C indicates participation of Kimura's correction of distances, N denotes absence of such correction, T and U mean evaluation tolerating and not tolerating gaps, respectively; MSA - multiple sequence alignment, s1, s2, s3, s4 - special sums in the corresponding series.

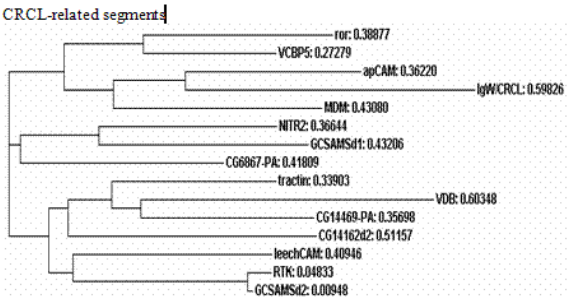
WP4.3.3 Examples of our phylograms. We show here all ten phylograms determined using one of four variants of phylogram evaluation (the variant with Kimura´s correction of distances without gap tolerance) to demonstrate various differences between phylograms determined by original and modified sequence sets.

Identical phylograms of short conserved regions were found after both ClustalW and MUSCLE derived MSA

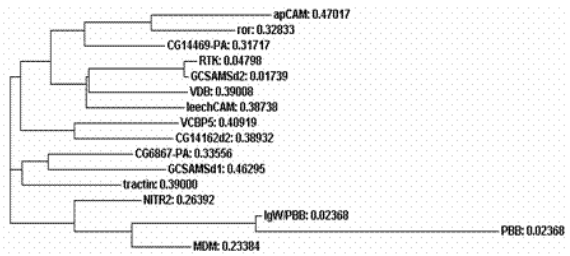
PBB-related segments



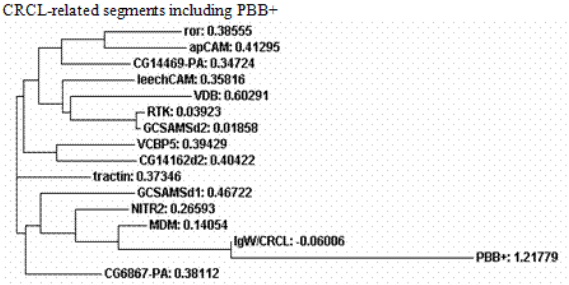
CRCL-related segments



PBB-related segments including PBB

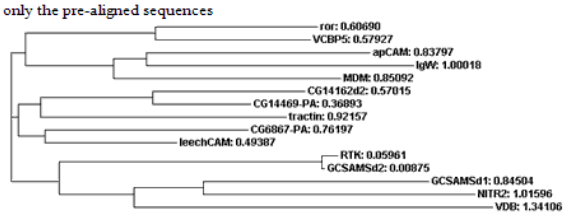


CRCL-related segments including PBB+

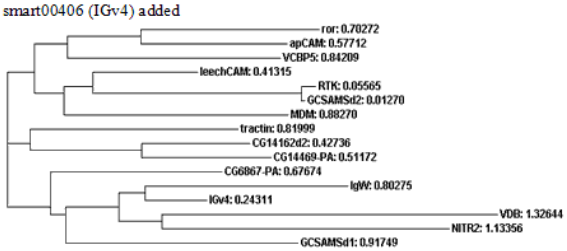


MUSCLE MSA with the whole pre-aligned segments

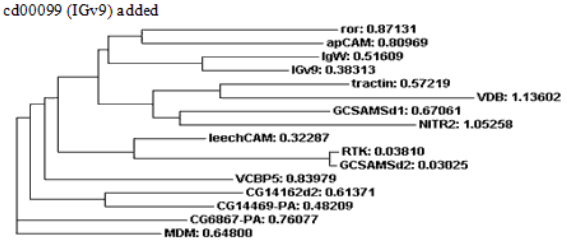
only the pre-aligned sequences



smart00406 (IGv4) added

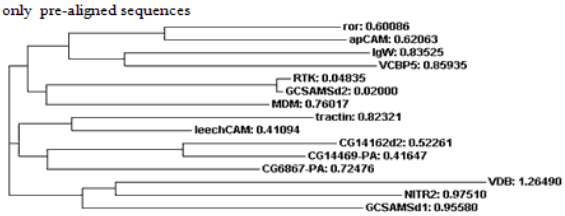


cd00099 (IGv9) added

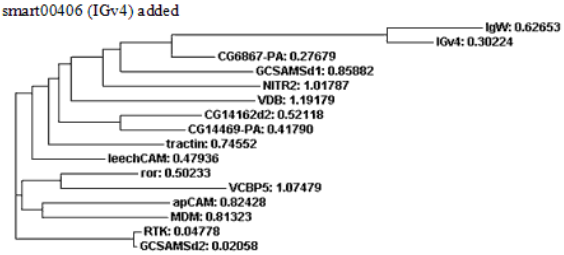


CLUSTAL W MSA with the whole pre-aligned segments

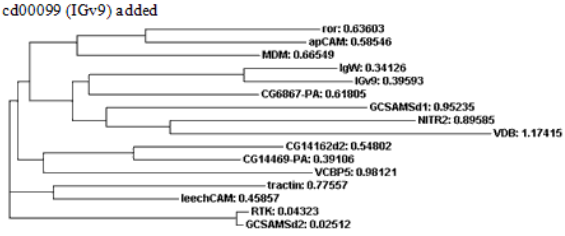
only pre-aligned sequences



smart00406 (IGv4) added



cd00099 (IGv9) added



WP5 Abbreviations

WP5.1 Names of molecules: AID - activation-induced cytidine deaminase; apCAM - adhesion molecule, identical with co-selected NCAM-related molecule of the same origin; AR - antigen receptor(s); CAM - a group of cell adhesion molecules; COG3210 - large exoproteins involved in heme utilization and adhesion; CP-933P - putative tail component of a cryptic prophage; FGFR - fibroblast growth factor receptor; GCSAM - sponge cell adhesion molecules cloned from the marine sponge *G. cydonium*; GCSAML, GCSAMS - long and short cell adhesion/recognition molecules from *G. cydonium*, respectively; HSPG2 - heparan sulphate proteoglycan 2; Ig, Igs - immunoglobulin(s); IgW - immunoglobulin(s) W; KGFR2 - keratinocyte growth factor receptor 2; leechCAM - cell adhesion molecules from medicinal leech; MDM - molluscan defence molecule; NCAM - a group of neural cell adhesion molecule; NITR - novel immune-type receptor(s); PPCAR - ("pre-historical") phylogenetic precursors of hypothetical common ancestor of antigen receptors encoded possibly by rearranging gene; RTK - receptor tyrosine kinase; RTPh - receptor tyrosine phosphatase; SIRP - signal-regulatory protein(s); SRTK - sponge RTK; VCBP5 - variable region-containing chitin-binding protein 5; VDB - a molecule bearing an Ig-like variable region of the CTX; TCRL - T-cell receptor-like molecule from *Petromyzon marinus*; TCRalpha - T- cell receptor alpha; VpBLP - lamprey V-preB-like protein.

WP5.2 Conserved domains: BID_1, BID_2, big_1, big_2 - typical bacterial Ig-like domains; CBM - carbohydrate binding domain (CBM_4_9); CelD - N-terminal ig-like domain of cellulase; CBM_14, ChtBD2 - different chitin binding domains; COG3291 - domain overlapping PKD (domain) within MM1983; COG5492 - more distinct Ig-like domains; collagen - collagen triple helix repeat (20 copies); CW_bin2 - putative cell-wall binding repeat 2; FN3 - fibronectin type III domain; fn3 - FN3 of mouse origin; Fz - Fz domain; GH_9 - glycosyl hydrolase family 9; IG4 - IG domain smart00409 present also in Igs; IG9 - cd00096 a predominant feature of most Ig domains is a disulfide bridge connecting two beta sheets with a tryptophan packing against the disulfide bond (molecules such as T-cell receptors, CD2, CD4 and CD8); ig - examples include antibodies, giant muscle titin and receptor tyrosine kinases; IGc2 - relationship to the second constant domains of Igs; IGcam - Ig domains of cell adhesion molecules, similar to: NCAML1, fascicilin II and insect immune protein hemolin; IGL - immungobulin-like domains that cannot be classified into variable Ig domains, IGc1, IGc2 or Ig; IGv6, IGv9 - variable Ig domains (**IGv**) smart406 and cd00099 (more related to

Igs and T cell receptors), respectively; KR - Kringle domain; LytB - putative cell wall domain; PKD - various polycystic kidney disease domains; OLF - olfactomedin-like domain; Pkinase - protein kinase domain; SPS1 - serine/threonine protein kinase domain of general function; STKc - serine/threonine protein kinase domain of phosphotransferases; TyrKc - tyrosine kinase, catalytic domain.

WP5.3 Names of genera: *A.* - *Archaeglobus* (Table 1), *A.* - *Anopheles* (Table 2), *A.* - *Aplysia* (Table 3), *B.* - *Branchiostoma, C.* - *Cellulomonas* (Table 1), *Cy.* - *Cytophaga, C.* - *Caenorhabditis* (Table 2), *C.* - *Ciona* (Table 3), *D.* - *Dechloromonas* (Table 1), *D.* - *Drosophila* (Table 2), *De.* - *Desulfitobacterium* (Table 1), *E.* - *Escherichia* (Table 1), *E.* - *Eptatretus* (a hagfish generum in Table 2), *G.*- *Geodia, H.* - *Hirudo, Ha* - *Halocynthia, L.* - *Leptospira* (Table 1), *L.* - *Lymnaea* (corresponds to *L. stagnalis* = great pond snail; Table 3), *M.* - *Methanosarcina, P.* - *Petromyzon* (corresponds to *L. marinus* = marine lamprey), *T.* - *Treponema, V.* - *Vibrio, Y.* - *Yersinia.*

WP5.4 Local abbreviations in the text, figure and tables: CICIDS - chimerical IGv-related conserved Ig domain similarities (Table 2); CR - conserved regions located in the C-terminal part of both IGv-related segments derived by both multiple sequence alignments (Table 4; WP3); CRCL - two aa longer regions than CR were derived by CLUSTAL W 1.82 (Table 4; WP3.2); DIF1 - a segment of frequent oligonucleotide dissimilarities (Fig.1; the second section of Results); CRIgW - the CR segment of model (reference sequence) IgW (Table 4); ER - envelope region(s), i. e. a pair or unique region(s) including enveloped segment and additional (enveloping) limited overlaps, whose extents are at most half of both lengths of oligonucleotides used in our scanning of regional sequence identities (for details see Fig.1); GF - an important segment (GP:N132-203) of the first Ig domain of GCSAMS (Fig.1; WP2.1) restricted by comparison of GCSAMS with Igs; GCSAMS(cdr1.1.com) - related to Ig heavy chains, and simultaneously included in both CDR1-like segments of GCSAMS (Fig.1; WP2.1); GCSAMS(cdr1L.1) - CDR1-like segment of GCSAMS related to Ig light chains (GP:aa22-34, N150-188) (WP2.1; Kubrycht et al. 2004); HDR - hypermutating DNA regions of about one kilobase length (chapter 1 and section 4.3); hoaa - high occurrence aa (Table 4; WP3.2 and WP3.3); HT - hypermutation tetranucleotides (frequent targets for activation-induced cytidine deaminase in HDR) (the second section of Results); HTS1 - hybrid template sequence 1 constructed in this paper according to last subsection of Material and Methods and

multiple sequence alignment displayed in Table 4 (B = Q,R,T and Z = D,S); Ig/T - these values represent maximal score values of Ig/T-cell receptor sequence similarities (Table 3); IgW/PBB, IgW/CRCL - complemented IgW segments similar to PBB, or of CRCL extent, respectively (Table 4; WP3.2); IROIL - sequence identities with human and mouse mRNA oligonucleotides of inferior possible lengths (Fig.1); MSA - multiple sequence alignment(s) (Table 4; WP3); MTC - model twin column, i. e. a sequence block column of non-integer height value defined by total X identity or similarity (meanwhile in cases of mutation convertibilities) corresponding to the same column probability of X in the observed (usual) column of integer height (WP3.3; Kubrycht et al. 2002); N MYA - N million years ago (chapter 1); ORD - oligonucleotide mRNA dissimilarities (Fig.1); PBDS - PSI-BLAST derived segments (WP3); PBS - peptide chain encoded by Ohno´s 48-base-long "modern intact primordial building block" derived from sequences of mouse variable Ig chains; PK - protein kinases (section Questions and Possibilities); PKSI - peptide protein kinase substrates and inhibitors; RA aa - residual alternative aa (see Methods); rhoaa - reevaluated high occurrence aa (reevaluation of hoaa was based on PHI BLAST procedure described in last subsection of Results, Table 4 and WP3.2); RfS - reference sequences i. e. important sequences or representatives of important sequence groups (last section of Material and Methods); S1, S2 - bacterial protein segments similar to HTS1 (last subsection of Results); SM - single mutants (mutation derivatives) of HT (see above); SO aa - aa of superior occurence (Material and Methods); SPSE - specific segments similar to HTS1 (in HTS1-derived PHI BLAST searches) and including selected conserved regions (last subsection of Results); T10L - top ten layer of PHI BLAST searches (together with SPSE); TIGV, TIGCAM - short IGv- and full length IGcam- related segments of TCRL located at aa positions 84-119 and 158-228, respectively (the third section of Results).

WP5.5 Broadly occurring abbreviations: For these abbreviations see the second footnotes of the paper.