Supplementary File 2. **Dictionary of abbreviations** (accessible on www.papersatellitesjk.com)

to the paper: Antibody-like phosphorylation sites in focus of statistically based bilingual approach

Contents

Abbreviations frequently used in different sites of the paper (global abbreviations in the main text)

aa - amino acid residue(s)

AID – activation-induced cytidine demainase

AR – antigen receptor(s)

ATM - ataxia-telangiectasia mutated (kinase)

BS, BS$^{\#}$, BS[MUSAS], BS$_m$ – various forms of bit scores (see below)

CDR1 - complementarity determining region 1 of IgV (i.e. the first hypervariable region of antibodies)

CDR1a, CDR1all, CDR1L, CDR1light and CDR1sum – segments composing or overlapping CDR1 (see below)

CSB - conserved segment of sequence block(s) in general

CSB1, CSB2 - specifically restricted CSB in initial MSA (cf. Fig. 1 and 2)

F, F-value - degree of fuzzy-related intervals of LE (cf. section 2.2.2, Fig. 2 and 3)

HM - hypermutation motif(s) (cf. section 3.4 and Fig. 5)

HM* - HM critically located with respect to aa alteration

Ig - immunoglobulin(s)

IgV – variable Ig domain(s) of antigen receptors and related molecules (not only of immunoglobulins)

LE - length equivalent(s) (cf. sections WP2.1.1 and WP5.2)

MEP - MPL-encoded peptide(s)

MNSQ - multi-nucleotide-sequence query/queries composed of the selected segments of antigen receptors similar to the accessible defined segments of conserved domains (cf. chapter WP2.2, sections WP5.4 and WP5.5)

MNSQ1, MNSQ2 - MNSQ number one and two differing in initial steps of their generation (cf. section 2.3)

MPL - MRNS, which encode protein regions containing predicted as phosphorylation sites of the corresponding PPSIg- and critical HM*-related locations

MRNS - MNSQ-related nucleotide segment(s) present in molecules different from AR and determined in the BLAST searches with MNSQ1 and MNSQ2 queries (cf. sections 2.4, WP5.6, WP5.7 and chapter WP2.3)

MSA record – record of multiple sequence alignment(s)

NS – nucleotide sequence(s) of the corresponding both strands of cDNA

OR, OR*[0] – precise and regularly approximated odds ratio values, respectively (see section 2.1.3 or below)

PPS – phosphorylated protein segments

PPSIg – sequences related to PPS occurring in N-terminal IgV segments

SBC – sequence block column(s)

SF1, SF2, SF3 - supplementary files accessible on www.papersatellitesjk.com or via email jkub@post.cz.

TCA - three-step ternary combined (see below or sections 2.4, WP2.3.2-4 and WP5.7)

TCR – T-cell receptor(s)

WPk.m, WPk.m.n – parts of the text present in supplementary files and forming k-th chapter, m-th section or chapter (if n exists) and in case of n occurrence also n-th section

W-pairs - pairs of WRCH or WRCY co-localized in certain distances (cf. sections 3.4 and 4.4)

Overall list of local and global abbreviations used in paper and supplementary files

aa - amino acid residue(s)

AID - activation-induced cytidine demainase

AR - antigen receptor(s)

ATM - ataxia-telangiectasia mutated (kinase)

BNsup35 – global search for superior BLASTN similarities restricted by a minimum score of 35 bit

BS – bit score

$BS^{\#}$ - double sequence similarity equivalents of bit score (cf. section 2.2.3)

BS[MUSAS] – bit score of MUSAS

BS_m - mean bit score between consensus and chains forming conserved sequence blocks

CDR1- complementarity determining region 1 of IgV (i.e. hypervariable region 1 of antibodies)

CDR1a – segments achieving at least 50% overlaps with CDR1all

CDR1all – common part of CDR1 occurring in both light and heavy chains

CDRL - segments achieving at least 50% overlaps with CDR1light

CDR1light – part of CDR1 occurring only in CDR1 light chains

CDR1s - a segment overlapping both CDR1all and CDR1L, but less markedly CDRall (less than 50% of overlap)

CDS - conserved domain sequences (s)

cls - consensus like sequence (nucleotide and protein sequences of clonal names AF273898.1 and AAK20241.1, respectively; *Danio rerio* origin; ISS2), an immunoglobulin sequence segment selected in the initial TBLASTN searches necessary for MNSQ2 construction as the sequence of highest similarity with the selected three groups of CSB1-related QS (cf. section WP2.2.3); cls was used in translated form as QS participating also in initial TBLASTN searches, i.e. ISS2

CSB - conserved segment of sequence block(s) in general

CSB1, CSB2 - specifically restricted CSB in initial MSA (cf. Fig. 1 and 5)

CTAE - collisions of transcription apparatus (synthesizing mRNA copy of transcribed strand of DNA) with APOBEC enzymes interacting with HM present in the same strand

DX, DXS - double combined TBLASTX searches for co-localized sequence similarities, restricted by a minimum score of 30 and 35 bits, respectively

ELEMS - Evaluation of sequence similarities using Length Equivalent Measures as a System

ER - Entrez restriction(s) of BLAST searches described in Fig. 1 and section WP2.3.1

ER1 - ER related to molecules involved in phosphorylation (limited by 25 bits (TCA) or 30 bits (other searches); cf. WF22)

ER2 - ER concerning molecules including multiple cancer names in their title (limited by 25 bits (TCA) or 30 bits (other searches); cf. WF23)

ER3 - ER substantially diminishing occurrence of antigen receptor sequences in “global searches” for superior similarities (limited by 35 bits; cf. WF21)

ER4, ER5 - ER with molecules restricted by FFAS-scan (limited according to employed approach by 25 bits (TCA) or 30 bits (other searches); cf. WF19 and WF20)

ER6, ER7 - ER concerning special and generalized NITR searches limited by 30 (cf. WF24) and 35 bits (cf. WF25), respectively

ERα – list of Entrez restriction defined with formulas WF19, WF20, WF22, WF23 and WF24 present in section WP2.3.1 of SF1 (cf. Fig. 1)

F, F-value - degree of fuzzy-related intervals of LE

freq[MUSAS] – frequency of MUSAS

FRL - fuzzy related limit(s)

GLC - germ-line cells

hci - highly conserved island(s)

HM - hypermutation motif(s)

HM* - hypermutation motif(s) located at nucleotide sequence position critical with respect to possible aa alteration

HQS - hybrid query sequences

Ig - immunoglobulin(s)

IgV – variable Ig domain(s) of antigen receptors and related molecules (not only of immunoglobulins) (cf. conserved IgV-related domain constructs present in Fig. 5)

ISS1, ISS2 - initial sequence searches starting selection of MNSQ1 and MNSQ2 units, respectively

LE - length equivalent(s) (cf. sections WP2.1.1 and WP5.2)

mean(X_i) – mean value determined by set of X_i values

MEP - MPL-encoded peptide(s)

MEPS - MRNS encoded protein segment(s)

MNSQ - multi-nucleotide-sequence query/queries composed of the selected segments of antigen receptors similar to the accessible defined segments of conserved domains (cf. chapter WP2.2, sections WP5.4 and WP5.5)

MNSQ1, MNSQ2 - MNSQ number one and two differing in initial steps of their generation (see chapter WP2.2, sections WP5.4 and W5.5)

MPL - MRNS, which encode protein regions containing predicted or confirmed as phosphorylation sites of the corresponding PPSIg- and critical HM*-related locations

MPQnα - multi-protein-sequence queries used in revising searches restricting set of sequences finally composing MNSQ1 and MNSQ2 (n = 1, 2 and 3 corresponding to included chains of CSB1, CSB2 or MSEP, respectively; α = a or b denoting presence of LE- or σ-consensus, respectively; cf. sections 2.3, WP5.4 and WP5.5)

MPQ3ab - MPQ containing sequences of MSEP segments of initial MSA and both LE- and σ-related consensi

MRNS - MNSQ-related nucleotide segment(s) determined in the corresponding BLAST searches with MNSQ1 and MNSQ2 queries (cf. section 2.4 and chapter WP2.3, sections WP5.6 and WP5.7)

MSA - multiple sequence alignment(s)

MSA record – resultant sequence record after MSA

MSEP - MSA segment enveloping chains corresponding to PPSIg and restricting MNSQ (section 2.3)

MSDV - minimum significantly different values (cf. section WP2.3.4)

MUSAS - MNSQ-unit derived similarities with almost the same segment of subject sequence

NS – nucleotide sequence(s) of the corresponding both strands of cDNA

OR - odds ratio(s)

OR*[0] - an approximated value of odds ratio, if 2x2 table contains zero in a unique element (cf. section 2.1.3)

QS - query sequence(s)

PK - protein kinase(s)

PS - pattern-related sequence similarities

pm3 - a special CSB1-related consensus derived with the help of FFAS (cf. section 3.1 and WP5.1)

PPSIg - sequences related to PPS occurring in certain N-terminal part of IgV segments

PPS - phosphorylated protein segments (specifically recognized by the protein kinases)

SBC - sequence block (i.e. MSA) column(s)

SEM - sequences enveloping MEPS

SF – general abbreviation of supplementary files

SF1, SF2, SF3 - supplementary files accessible on www.papersatellitesjk.com or via email jkub@post.cz

SMFS – substitution-matrix-derived fuzzy-related system

SNEM - short nearly exact matches

STS-query - segment of typical sequences forming here CSB1, which was selected as source for (i) QS and MPQ1 determining ISS1 or (ii) QS, HQS and MPQ1 necessary for ISS2 (see section WP2.2.3, WP5.4)

TCA - three-step ternary combined approach including two types of searches for repeatedly co-localized similarities (see sections 2.4 , WP2.3.2-4 and WP5.7)

TCR - T-cell receptor(s)

W11 - word size 11 in BLASTN searches

W2P30, W3B62 – two current WMR based on word sizes two and three and substitution matrices PAM 30 and BLOSUM 62, respectively

#W2P30∩W3B62 - simultaneous occurrence of co-localized similarities in the searches with WMR W2P30 and W3B62 was required (see Fig. 1)

WFn - n-th formula in the main web page supplement

WMR - restriction of sequence searches based on determination of word size and type of substitution matrix (cf. W2P30 and W3B62)

W-pairs - pairs of WRCH or WRCY co-localized in certain distances (cf. sections 3.4 and 4.4)

WPk.m, WPk.m.n – parts of the text present in supplementary files and forming k-th chapter, m-th section or chapter (if n exists) and in case of n occurrence also n-th section

WPTn - n-th table of supplementary files

x-intervals - the intervals of score products with extent of 0.01 (i) quantifying axis x composing histograms in Fig. 6 (cf. sections WP3.2 and WP3.3) and (ii) associated with the localizing values denoted here as x-values

σ-consensus - a consensus containing aa achieving maximum values of column scores (cf. section 2.2)