Computational Molecular Bioscience, 2020, 10, 12-44
https://www.scirp.org/journal/cmb
ISSN Online: 2165-3453
ISSN Print: 2165-3445

Conserved Immunoglobulin Domain Similarities of Higher Plant Proteins

Jaroslav Kubrycht[1*], Karel Sigler[2]

[1]Department of Physiology, Second Faculty of Medicine, Charles University, Prague, Czech Republic
[2]Laboratory of Cellular Biology, Institute of Microbiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
Email: *jkub@post.cz, sigler@biomed.cas.cz

How to cite this paper: Kubrycht, J. and Sigler, K. (2020) Conserved Immunoglobulin Domain Similarities of Higher Plant Proteins. *Computational Molecular Bioscience*, **10**, 12-44.
https://doi.org/10.4236/cmb.2020.101002

Received: January 14, 2020
Accepted: March 28, 2020
Published: March 31, 2020

Abstract

The traces of immunoglobulin domain similarities were searched in sequences of higher plants using bioinformatic tools to look for possible early phylogenic structural relationships. 280 thousand sequence IDs, obtained by sixteen types of primary BLAST searches, were differently processed by seventeen selection procedures and an anti-redundant sequence-related approach using JavaScript, PHP, Windows programs and conserved domain searches by means CDD. The resulting seventeen sets of records describing conserved domain similarities of 1323 different sequence IDs yielded a set of next generation (final set) comprising forty-nine records containing superior ("non-refutable") conserved immunoglobulin domain similarities. The selected sets and their subsets were mapped and subsequently statistically compared with respect to immunoglobulin-related as well as other reciprocal domain linkages. The list of frequently occurring conserved domain similarities concerned first of all domains important for plant and metazoan immunity, e.g. tyrosine kinases accompanying variable immunoglobulin domains in early *Metazoa*, toll-like receptors, lectin and leucine-rich repeat domains. Detailed description of immunoglobulin domain similarities occurring in the final set was completed by fold analysis of the restricted segments. The data were then discussed with respect to i) immunoglobulin fold evolution, ii) possible structural importance of domains cd14066 (IRAK) and PLN00113 (LRR-associated kinase) for deep evolution of catalytic serine/threonine/tyrosine kinase domains, iii) interatomic, structural and specificity standpoints and iv) traces of antibody-like phosphorylation sites described in our previous paper.

Keywords

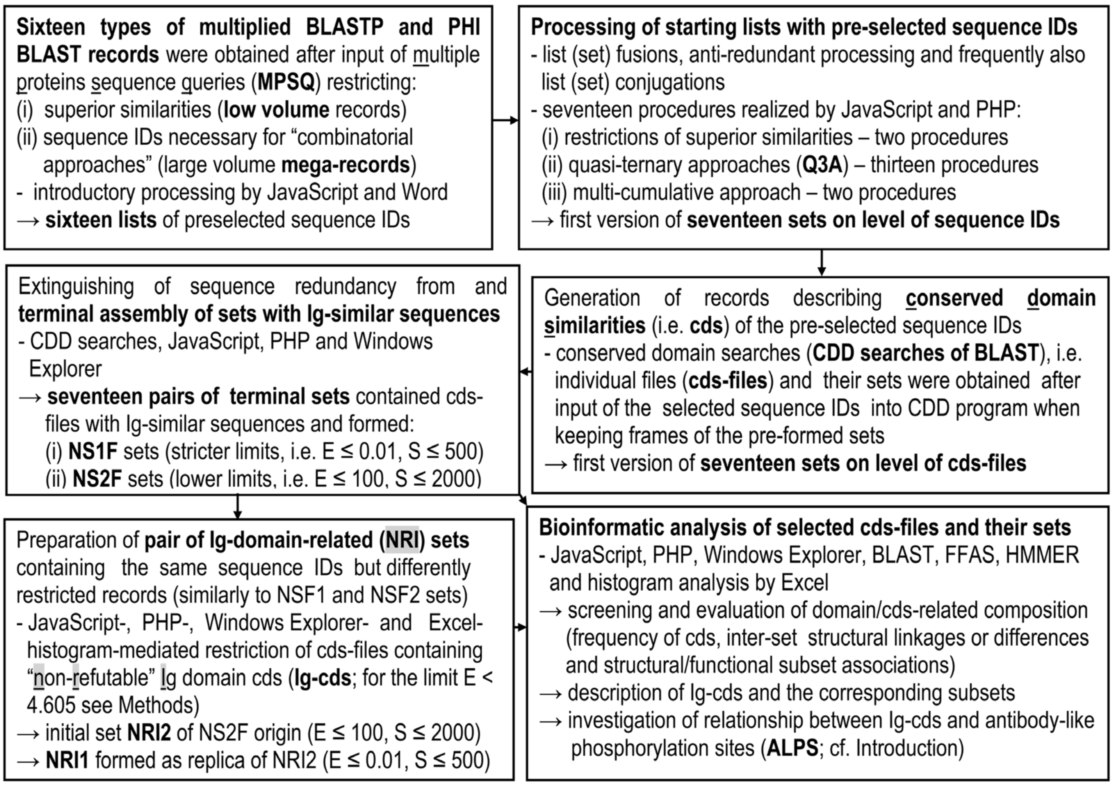
Conserved Domain, Domain Shuffling, Deep Evolution, Evolution, Fold, Immunoglobulin, Kinase, Plant, Immunity, Resistance

1. Introduction

In accordance with recent opinions, we distinguish two main layers of plant immunity, *i.e.* pathogenic pattern-triggered immunity (**PTI**) and effector-triggered immunity (**ETI**). These layers are frequently accompanied by activation of kinase cascades, Ca^{2+} influxes, generation of reactive oxygen species, transcriptional reprogramming, phytohormone signalling, proteasome degradation pathways and more specifically (*i.e.* in special cases of virus infections) also RNA silencing machinery [1] [2]. PTI is initiated by interactions of pathogenic motifs (patterns) with cell surface signaling molecules exposed to the cell wall mostly belonging to the superfamily of receptor like kinases (**RLK**; [3] [4] [5] [6]). Among familiar groups of RLK, serine/threonine kinases receptor-like kinases (**STRK**) are frequent. In addition to their specific catalytic domains, these STRK can contain binding sites formed by e.g. lectin or leucine-rich repeat (**LRR**) domains [7] [8] [9] [10]. ETI is mostly mediated by intracellular receptors called nucleotide binding sites/LRR proteins (**NLR**). Molecules of NLR are composed of i) nucleotide-binding LRR domain and ii) alternatively of either coiled-coil (**CC**) or Toll-like/interleukin 1 receptor (**TIR)** domain [11] [12] [13].

The domains mentioned above, *i.e.* catalytic domains or STRK and binding TIR or LRR domains, also take part in mechanisms of metazoan immunity [14] [15]. In accordance with this parallel occurrence, we can pose a question of whether at least some ancestor-like traces of typical animal superfamilies can be found in higher plants (*Embryophyta*), as consequences of horizontal transfer or co-evolution of superfamiliar ancestors. This concerns first of all traces of 800 million-years-old immunoglobulin (**Ig**) superfamily representing a very important group of immune proteins in most of metazoans including spongi and vertebrates [16] [17] [18] [19].

In our four preceding papers [20] [21] [22] [23], we hypothesized step-by-step a possible role of antibody-like phosphorylation sites (**ALPS**) or the corresponding nucleotide repeats in the evolution of antigen receptors and their ancestors. Consequently, ALPS sequences were used together with specifically restricted segments of conserved Ig domain sequences (**Ig-cd**) to compose four multiple protein sequence queries (**MPSQ**) inputted in our **starting procedural step** including BLAST searches. More precisely, the corresponding five-step selection procedures and the following data analysis were performed as described in **Figure 1**. Concerning superior ("non-refutable") conserved domain similarities (**cds**) of Ig-cd (*i.e.* **Ig-cds**) present in the set NRI2 (cf. **Figure 1**), we found cds-recording files (**cds-files**) representing protein sequences attaining i) **significant Ig-cds** ($p < 0.01$), ii) quasi-significant Ig-cds ($0.01 \le p < 0.1$) and iii) less significant Ig-cds supported by FFAS-fold searches or literature contexts. Two types of Ig-cds were distinguished in NRI set. **Dominant Ig-cds** attained superior regional evaluation of cds. **Recessive Ig-cds** co-located with cds of non-Ig domains achieving higher bit score but sometimes an interesting context. Two significant dominant Ig-cds included **bacterial Ig** domains. In addition to Ig-cds evaluation, we indicated multiple associations of robust cds of catalytic-kinase-, lectin-, LRR- and TIR-related-domains, broadly occurring in the selected cds-files.



Sixteen types of multiplied BLASTP and PHI BLAST records were obtained after input of multiple proteins sequence queries (**MPSQ**) restricting:
(i) superior similarities (**low volume** records)
(ii) sequence IDs necessary for "combinatorial approaches" (large volume **mega-records**)
- introductory processing by JavaScript and Word
→ **sixteen lists** of preselected sequence IDs

Processing of starting lists with pre-selected sequence IDs
- list (set) fusions, anti-redundant processing and frequently also list (set) conjugations
- seventeen procedures realized by JavaScript and PHP:
(i) restrictions of superior similarities – two procedures
(ii) quasi-ternary approaches (**Q3A**) – thirteen procedures
(iii) multi-cumulative approach – two procedures
→ first version of **seventeen sets on level of sequence IDs**

Generation of records describing **conserved domain similarities** (i.e. **cds**) of the pre-selected sequence IDs
- conserved domain searches (**CDD searches of BLAST**), i.e. individual files (**cds-files**) and their sets were obtained after input of the selected sequence IDs into CDD program when keeping frames of the pre-formed sets
→ first version of **seventeen sets on level of cds-files**

Extinguishing of sequence redundancy from and **terminal assembly of sets with Ig-similar sequences**
- CDD searches, JavaScript, PHP and Windows Explorer
→ **seventeen pairs of terminal sets** contained cds-files with Ig-similar sequences and formed:
(i) **NS1F** sets (stricter limits, i.e. E ≤ 0.01, S ≤ 500)
(ii) **NS2F** sets (lower limits, i.e. E ≤ 100, S ≤ 2000)

Preparation of **pair of Ig-domain-related (NRI) sets** containing the same sequence IDs but differently restricted records (similarly to NSF1 and NSF2 sets)
- JavaScript-, PHP-, Windows Explorer- and Excel-histogram-mediated restriction of cds-files containing "non-refutable" Ig domain cds (**Ig-cds**; for the limit E < 4.605 see Methods)
→ initial set **NRI2** of NS2F origin (E ≤ 100, S ≤ 2000)
→ **NRI1** formed as replica of NRI2 (E ≤ 0.01, S ≤ 500)

Bioinformatic analysis of selected cds-files and their sets
- JavaScript, PHP, Windows Explorer, BLAST, FFAS, HMMER and histogram analysis by Excel
→ screening and evaluation of domain/cds-related composition (frequency of cds, inter-set structural linkages or differences and structural/functional subset associations)
→ description of Ig-cds and the corresponding subsets
→ investigation of relationship between Ig-cds and antibody-like phosphorylation sites (**ALPS**; cf. Introduction)

Figure 1. Five-step selection and the following analysis of Ig-similar and Ig-domain-related sequences. A concise methodological overview related to the content of our paper can be seen here. E—Expects, S—sample sizes.

2. Methods

2.1. Overall Description of Softwares and Procedures

For an overall scheme of procedures and approaches used in this paper see Figure 1. Selection of sequence IDs and cds-files was performed with the assistance of online accessible programs BLASTP, PHI-BLAST [24] [25] and conserved domain searches of BLAST (**CDD searches** [26] [27] [28] [29]) based on **domain structures** selected with the contribution of **three-dimensional analysis** [30]. Further analysis of the data required on-line bioinformatic programs such as FFAS03 and HMMER [31] [32] [33] and a downloadable active web page with Fisher's exact test for 2 × 2 tables [34]. Our JavaScript and PHP codes of active web pages were written by freely downloaded PSPad editor. Easy PHP 12.1 then enabled runs of PHP codes. For some purposes, Word, Window Explorer or conditioned formatting and histograms both generated by Excel were necessary. For details see Sections of Chapter WP1.

2.2. Multiple Protein Sequence Queries (MPSQ)

Conserved variable Ig domains (**IgV-cd**) were used to form two MPSQ (**MQI**), *i.e.* MQI1 and MQI2. Segments forming our MQI composed the sequence block presented in Figure 2 of our previous paper [23]. Besides local high sequence similarities in the selected sequence sub-blocks, additional phylogenic parameters decided about the final restriction of MQI segments. MQI1 included segments of sub-block (pm3) restricted based on fold relationships (block positions 15 - 42), the corresponding consensus segment and accompanying pattern WXXQXP. MQI2 comprised pattern DX(3)YXC and the segments containing common amino acids L,D,Y,C (block positions 81 - 97) restricting also the corresponding conserved sub-block of IgV-cd-related invertebrate sequences [21]. Consensus was not used to form MQI2 due to its identity with the participating segment of cd00099. Two ALPS-related MPSQ (**MQP**), *i.e.* **MQP1** and **MQP2,** were composed of the two different groups of ALPS achieving superior evaluation in database confirmation or prediction of ALPS. For the ALPS groups and protocol describing MPSQ formation from segment sequences and spacers see [23].

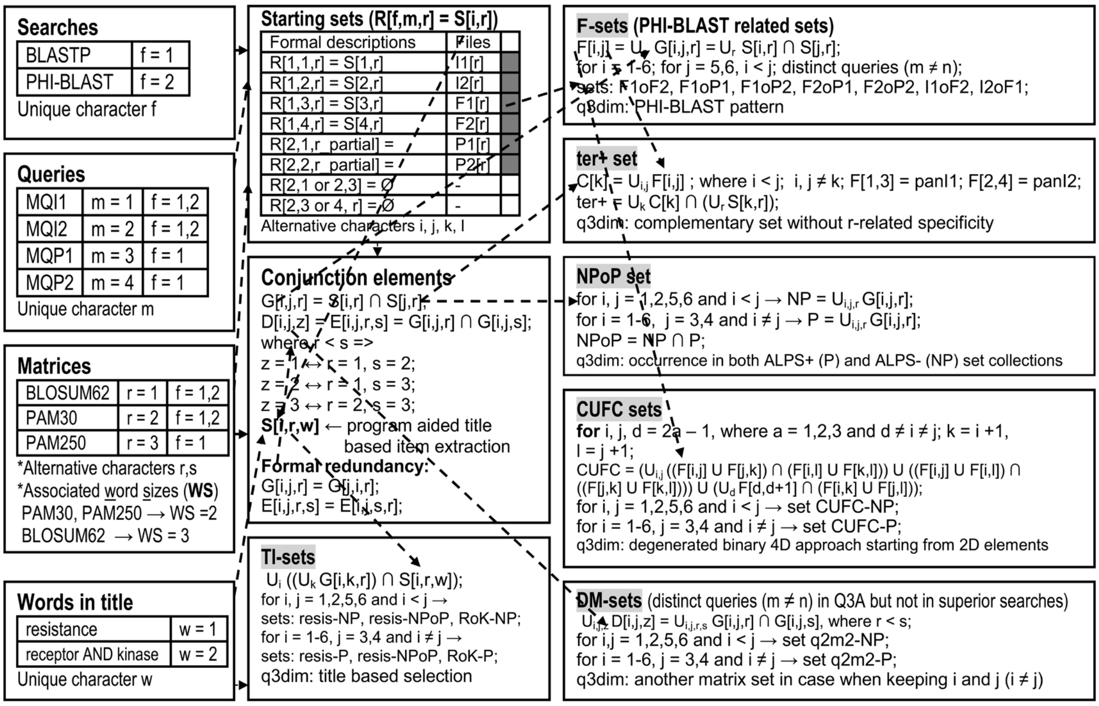

Searches
BLASTP | f = 1
PHI-BLAST | f = 2
Unique character f
Queries
MQI1 | m = 1 | f = 1,2
MQI2 | m = 2 | f = 1,2
MQP1 | m = 3 | f = 1
MQP2 | m = 4 | f = 1
Unique character m
Matrices
BLOSUM62 | r = 1 | f = 1,2
PAM30 | r = 2 | f = 1,2
PAM250 | r = 3 | f = 1
*Alternative characters r,s
*Associated word sizes (WS)
PAM30, PAM250 → WS =2
BLOSUM62 → WS = 3
Words in title
resistance | w = 1
receptor AND kinase | w = 2
Unique character w
Starting sets (R[f,m,r] = S[i,r])
Formal descriptions | Files
R[1,1,r] = S[1,r] | I1[r]
R[1,2,r] = S[2,r] | I2[r]
R[1,3,r] = S[3,r] | F1[r]
R[1,4,r] = S[4,r] | F2[r]
R[2,1,r_partial] = | P1[r]
R[2,2,r_partial] = | P2[r]
R[2,1 or 2,3] = Ø | -
R[2,3 or 4, r] = Ø | -
Alternative characters i, j, k, l
Conjunction elements
G[i,j,r] = S[i,r] ∩ S[j,r];
D[i,j,z] = E[i,j,r,s] = G[i,j,r] ∩ G[i,j,s];
where r < s =>
z = 1 ↔ r = 1, s = 2;
z = 2 ↔ r = 1, s = 3;
z = 3 ↔ r = 2, s = 3;
S[i,r,w] ← program aided title based item extraction
Formal redundancy:
G[i,j,r] = G[j,i,r];
E[i,j,r,s] = E[i,j,s,r];
TI-sets
Ui ((Uk G[i,k,r]) ∩ S[i,r,w]);
for i, j = 1,2,5,6 and i < j →
sets: resis-NP, resis-NPoP, RoK-NP;
for i = 1-6, j = 3,4 and i ≠ j →
sets: resis-P, resis-NPoP, RoK-P;
q3dim: title based selection
F-sets (PHI-BLAST related sets)
F[i,j] = Ur G[i,j,r] = Ur S[i,r] ∩ S[j,r];
for i = 1-6; for j = 5,6, i < j; distinct queries (m ≠ n);
sets: F1oF2, F1oP1, F1oP2, F2oP1, F2oP2, I1oF2, I2oF1;
q3dim: PHI-BLAST pattern
ter+ set
C[k] = Ui,j F[i,j] ; where i < j; i, j ≠ k; F[1,3] = panI1; F[2,4] = panI2;
ter+ = Uk C[k] ∩ (Ur S[k,r]);
q3dim: complementary set without r-related specificity
NPoP set
for i, j = 1,2,5,6 and i < j → NP = Ui,j,r G[i,j,r];
for i = 1-6, j = 3,4 and i ≠ j → P = Ui,j,r G[i,j,r];
NPoP = NP ∩ P;
q3dim: occurrence in both ALPS+ (P) and ALPS- (NP) set collections
CUFC sets
for i, j, d = 2a – 1, where a = 1,2,3 and d ≠ i ≠ j; k = i +1, l = j +1;
CUFC = (Ui,j ((F[i,j] U F[j,k]) ∩ (F[i,l] U F[k,l])) U ((F[i,j] U F[i,l]) ∩ ((F[j,k] U F[k,l]))) U (Ud F[d,d+1] ∩ (F[i,k] U F[j,l]));
for i, j = 1,2,5,6 and i < j → set CUFC-NP;
for i = 1-6, j = 3,4 and i ≠ j → set CUFC-P;
q3dim: degenerated binary 4D approach starting from 2D elements
DM-sets (distinct queries (m ≠ n) in Q3A but not in superior searches)
Ui,j,z D[i,j,z] = Ui,j,r,s G[i,j,r] ∩ G[i,j,s], where r < s;
for i,j = 1,2,5,6 and i < j → set q2m2-NP;
for i = 1-6, j = 3,4 and i ≠ j → set q2m2-P;
q3dim: another matrix set in case when keeping i and j (i ≠ j)

Figure 2. Quasi-ternary approaches (Q3A). In accordance with Methods, Q3A added a weak condition (quasi-third dimension) to the selection of common sequence IDs coming from certain pairs of specific BLAST mega-records obtained always with two different MPSQ. This weak condition was realized via four alternative ways, when requiring: i) additional participation of PHI-BLAST-associated sequence pattern in the selection process (this condition restricts the collection of **F-sets**), ii) occurrence of selected sequence ID in two mega-records of searches differing only in adjusted matrix (double matrix sets, *i.e.* **DM-sets**), iii) the presence of one or two words or abbreviations in molecular titles associated with the pre-selected sequence IDs (**TI-sets**) and iv) additional combinations of set fusion and conjugation in cases considered as spreading or analytical for our data processing (**sets CUFC**, **NPoP** and **ter+**). Ø—BLAST mega-records are formally assumed but in fact do not exist (this status followed from the absence of sequence patterns associated with MQP1 or MQP2 and impossibility to perform PHI-BLAST with the matrix PAM250); panI1, panI2—sets prepared by conjugation of five sets (cf. Section 2.6); q3dim—condition representing "quasi-third" dimension; oP—sets obtained with ALPS-related MPSQ (*i.e.* MQP1 or MQP2); -P, -NP, -NPoP—disjunctive subsets obtained only with, only without or simultaneously with and without ALPS-related MPSQ, respectively (similarly to the unique subset—NPoP, the set NPoP exists); r_partial—only incomplete list of r-values was used due to incompatibilities described in explanation of abbreviation Ø.

2.3. Two Different Restrictions of Sequence IDs Coming from Top Sequence Items of BLAST Records

Five parallel "sample records" differing in limiting Expect values or the download date but belonging to the same type of BLAST records were obtained in each of sixteen possible cases (cf. Figure 2). "Sample records" differed in the numbers of extracted items when forming top10 and t100 samples composed of upper 10 or 100 sequence items, respectively. These records were specifically non-redundantly fused forming *S*[*i*, *r*]-related sets **top10[*i*, *r*]** and ***t*100[*i*, *r*]** keeping sixteen lists corresponding to different types of multiplied BLAST records (cf. Figure 2). Terminal top10-NP subset (related to MQI1 or MQI2 inputs) and top10-P subset (related to MQP1 or MQP2 inputs) were prepared by selective non-redundant fusion of the corresponding top10[*i*, *r*]. Subsets of *t*100[*i*, *r*] were processed similarly to quasi-ternary DM-sets *D*[*i*, *j*, *z*] (described in Figure 2) using the same PHP program for the case of a less complex formula comprising two dimensional *D*[*k*, *z*] (cf. also WP1.1):

$$\bigcup\nolimits_k C[k] = \bigcup\nolimits_k \bigcup\nolimits_{z=1\sim3} D[k,z] = \bigcup\nolimits_k \bigcup\nolimits_{r<s} t100[k,r] \cap t100[k,s] \tag{1}$$

This means that non-redundant fusion of ALPS-unrelated (*k* = 1 - 10) and ALPS-related (*k* = 11 - 16) *C*[*k*] formed immediately disjunctive subsets **t100dm-NP** and **t100dm-P,** respectively.

2.4. BLAST-Derived Mega-Records Necessary for Quasi-Ternary and Multi-Cumulative Approaches

The aim of the searches was to obtain as many as possible long BLAST records of sequence items. Consequently, the productive Expect limits achieved the value range 2 × 10^5 - 2 × 10^7. This determined the 20000 sequence IDs for each BLASTP record, whereas PHI-BLAST searches yielded lower maximum numbers of sequence IDs, *i.e.* 6899-15.300. Each of the sixteen types of mega-records (cf. Figure 2) was downloaded in at least three versions differing in the date of download or limiting Expects.

2.5. Quasi-Ternary Approaches (Q3A) in Selection of Sequence IDs

Q3A represented compromising solution between selections of too extended sets determined by certain conjugations of two starting sets and too poor sets restricted by the conjugation of three sets. In accordance with Figure 2, Q3A comprised four manners how to perform weakened third dimension (quasi-third dimension) of SW architecture. For additional details see Figure 2.

2.6. Selection of Sequence IDs Based on Multi-Cumulative Approach

We selected sequence IDs simultaneously found by the same MPSQ in five required BLAST mega-records (cf. Figure 2), *i.e.* in three BLASTP and two PHI-BLAST searches. Due to the absence of patterns associated with MQP1 and MQP2, *i.e.*

disability of these MPSQ to perform PHI-BLAST searches, only the lists related to MQI1 or MQI2 determined the sets of cds-files entitled **panI1** and **panI2**, respectively.

2.7. Regularity of cds-Related Subsets

To avoid a false selection of cds-files containing only domain references but not searched cds, three keyword candidates (*i.e.* Pssm-ID, accession and name) were tested for numbers of selected cds-files. At least two of these numbers had to achieve minimum values to confirm regularity of the keywords associated with the minima (cf. WP2.1).

2.8. Two Set Families Were Derived When Removing Sequence Redundancy of cds-Files

Current conserved domain searches (see above) determining cds restricted by p < 0.01 and sample size 500 were performed with all items pre-selected by the preceding procedures (Figure 2). This resulted in a class of sets with **starting sets of cds-files**. The subsets of cds-files with positive domain occurrence of valid cds were reduced when removing sequence redundancy (see above and WP1.2) determining **NS1a family** of non-redundant sets of cds-files (*i.e.* **NS1aF sets**). Sequence redundancy was also checked in the cds-files without valid cds under modified conditions represented by limit E = 100 and sample size 2000. This restricted the NS2b family of cds-file subsets (**NS2bF**) and the corresponding non-redundant subsets of empty files of family NS1b (**NS1bF**). The following fusion of the corresponding sets of NS1aF and NS1bF then yielded sets of the reference family of NS1 sets (**NS1F**). Similarly to the formation of NS2bF and NS1F, we determined sets of NS2a family (**NS2aF**) including the same sequence IDs like NS1aF sets and then completed the assembly of sets of NS2 family (**NS2F**).

2.9. Three-Step Selection of Model Tyrosine Kinase Domains Forming Robust cds with Sequences of Plant Proteins and the Proteins of Early *Metazoans* Containing Variable Ig Domains in Addition

In the first step, we searched for NS1F-related files including receptor terms: “Ig) domain”, “Ig domain”, “Ig)-like”, “Ig-like”, “Immunoglobulin”, “B-cell receptor” and “T-cell receptor”. **In the second step**, the records of conserved domains extracted from the files were reduced to those including required cell type or cell-type-associated process, *i.e.* looking for the terms: “B-cell”, “T-cell”, “lymphocyte”, “lymphoma”, “amutoimmune”, “leukemia”, “macrophage”, “phagocyte”, “immune system”, “immunity”, “hematopoietic”. **In the third step**, we kept only such regular records of conserved domains whose names compose cds-files of three well defined signaling molecules containing both IgV-cd and tyrosine-kinase activity (RTK, SRTK, GCTK2) endogenously expressed in the immunologically important model living fossil *Geodia cydonium* [17] [18]. Since cd05034 denotes the family of Src kinase-like protein tyrosine kinases including four kinase subfamilies important for specific immunity (Lck, Blk, Lyn and Fyn), we substituted the preceding selection procedures with the knowledge about the linkage of this domain to specific immunity and fossil IgV-cd mentioned above and complemented then the former set of nine selected cdigvtk-related accessions with cd05034 (for the list of ten selected cdigvtk see Results). Four types of strategies comprising searches with ten selected regular domain keywords (see above) of cdigvtk (**dci**) were employed, *i.e.* i) enumeration of the most frequently selected subsets of the same cds-files (**mfs**) when using dci, ii) identification of item number determined by any of dci (**cdigvtk(max)**) and iii) enumeration of cds-files containing all dci, *i.e.* total coincidence of dci (**cdigvtk(tc)**).

2.10. Statistical Evaluations

For multiple notes to statistical processing including enumeration of odds ratio values (OR and OR* for 2 × 2 tables including zero values) and t-test see our preceding paper [23], its important associated sources [35] [36] and certain improvements mentioned in the webpage (WP2.2-4). Maximum Expect limit (**E[max]**) for significantly "non-refutable" Ig-domain related (**NRI**) cds was derived based on: i) significance limit related to inverse-phenomenon (for p < w max = 1 − w, *i.e.* all E achieving $E \geq E[\max]$ can be seen as significantly refutable), ii) current BLAST formula determining relationships between Expect and p-values [37] and iii) the value $w = 0.01$ limiting significance in CDD searches:

$$E[\max] = -\ln(1-\max) = -\ln(w) = -\ln(0.01) = 4.605. \tag{2}$$

E[max] enabled us to assemble of Ig-cds-related set selected from all sets of NSI2F. This set was called **NRI2** set (cf. Figure 1). For evaluation of Expect-related specificity and overall hierarchy of levels classifying cds in our paper see WP2.5 and WP2.6, respectively.

2.11. Terms, Acronyms, Abbreviations, Texts Denoted by WP-Associated References and the Reasons for Color Grading in Map-Like Pictures or Absence of Phylogenic Trees

The corresponding information composes our web page supplement, *i.e.* a pdf-file accessible in the corresponding section of our web page http://www.papersatellitesjk.com or via e-mail correspondence.

3. Results

3.1. Selected Records, Sets and Subsets

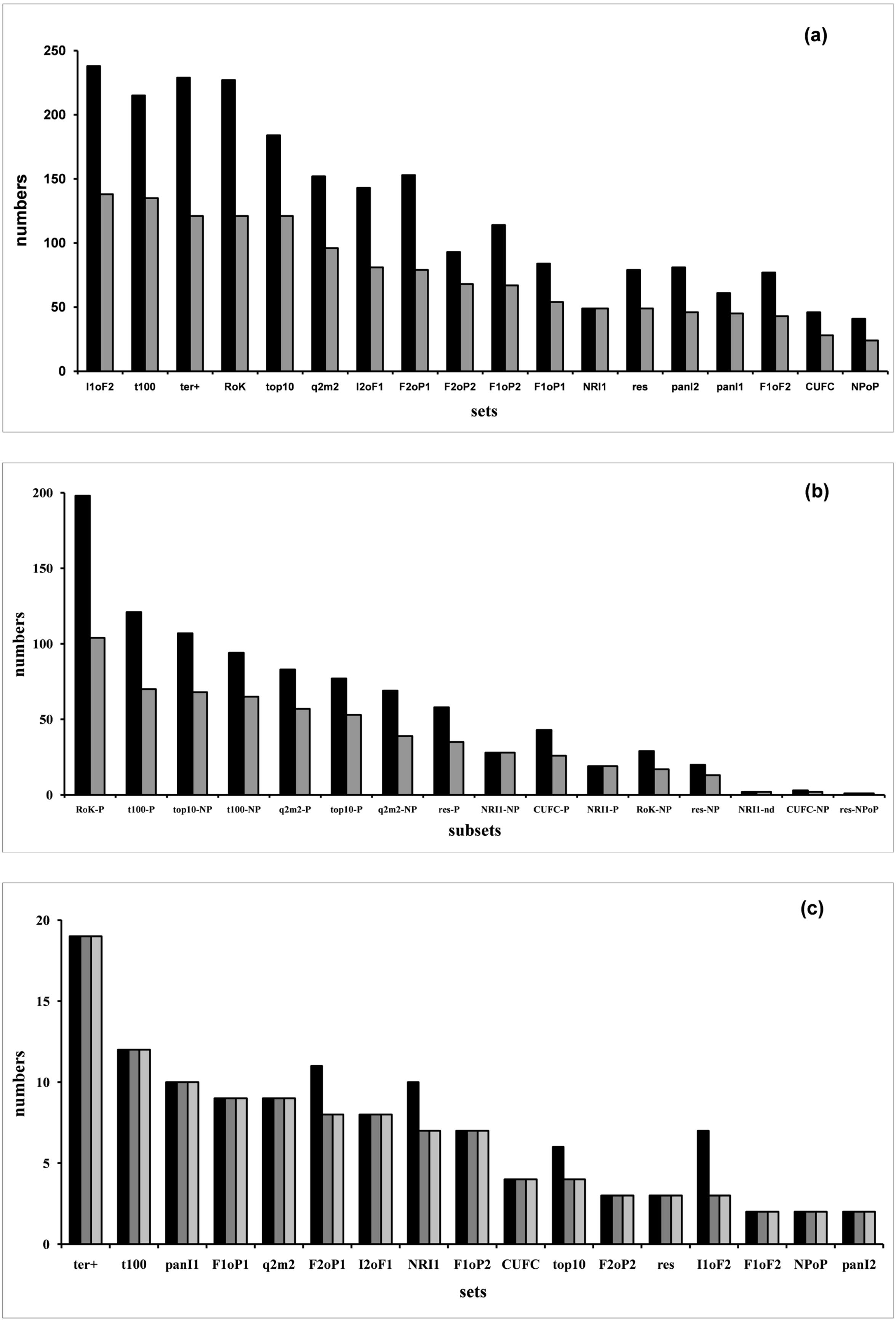
The starting set included sixteen types of BLAST search records containing 280.423 sequence IDs (for details see WP3.1-2; cf. Section 2.11). The selection procedures and the following removal of sequence redundancy resulted in 1323 cds-files present in 17 sets (composing the family of NS1 sets, *i.e.* **NS1F**). The same sequence IDs and sets then similarly composed the family of NS2 sets (**NS2F**) when using other limits for Expects (E ≤ 100) and sample sizes (s ≤ 2000; cf. Section 2.8). In accordance with the usage of ALPS-related MQP in selection of cds-files, we distinguished between **ALPS-related**, **ALPS-unrelated**, **ALPS-diversified** and **undiversified** intertwined sets (ter+, NPoP). In fact, six ALPS-diversified sets were derived here. These sets contained pairs of disjunctive ALPS-related and ALPS-unrelated subsets (denoted -P and -NP, respectively), whereas only unique additional undiversified subset was found in the case of resis set (subset resis-NPoP including unique cds-file). In addition to these primarily derived sets and subsets, 238 cds-files in different NS2F sets included conserved Ig-domain similarities (Ig-cds) in CDD records. 223 of these cds-files were non-redundant with respect to the recorded sequences. Forty-nine of the pre-selected 223 cds-files of NS2F belonged to those containing "non-refutable" Ig-domain-related cds (for the corresponding Ig-cds restriction see Section 2.10), *i.e.* they formed **NRI2** set. The same sequence IDs like those present in cds-files of NRI2 enabled then to assemble NRI1 set from sets of NS1F origin. Both NRI sets then consisted of three subsets, *i.e.* NRI-NP, NRI-NP and NRI-nd, where the latest name denotes cds-files coming only from undiversified intertwined sets of NSF2. For the numbers of cds-files in all sets and subsets described here see Figure 3. For names of sets and subsets see Figure 2 and WP5.

3.2. Cds-Files, Whose Descriptions of Individual Cds Contain the Term Immunoglobulin

The robust cds with catalytic tyrosine kinases domains frequently included term "immunoglobulin" in their descriptions. This concerned also most of descriptions of cds with evolutionary important model IgV-domain-associated catalytic tyrosine kinases domains from *Geodia cydonium* (**cdigvtk**; preselected here according to Methods). The corresponding coincidence with the term immunoglobulin was demonstrated using cds of domain **cd05034** (representing Src-kinase-like family) achieving the highest score among cdigvtk-derived cds (Figure 3 and Figure 4). In spite of these favorable results and the robust score values evaluating cds with cdigvtk (55 - 140 bits), the other kinases (serine/threonine kinases without immunoglobulin-related contexts) achieved considerably higher scores in their cds with the same sequence regions (Figure 4). In addition, the **hierarchy of cds** demonstrated in Figure 4 appeared to be consistent with the hierarchy of the cds-file-related subsets composing the receptor-kinase-domain-rich set RoK representing the set attaining the most statistically deviated occurrences of cdigvtk (cf. Table 1) and ALPS (104 cds-files; p < 0.05 in t-test):

$$\begin{aligned}&\text{cdigvtk}(\text{tc})[102]=\text{cd05102}[102]\subset\text{cdigvtk}(\max)[112]=\text{cd05034}[112]\\&\subset\text{cd14066}[113]=\text{I3R}[113]=\text{receptor-like kinase}[113]=\text{Ig}[113]\\&\subset\text{PLN00113}[118]\subset\text{RoK}[121]\end{aligned}\tag{3}$$

where square brackets include numbers of cds-files (**NCF**) **specifically double-**


(a)
250
200
150
100
50
0
numbers
I1oF2
t100
ter+
RoK
top10
q2m2
I2oF1
F2oP1
F2oP2
F1oP2
F1oP1
NRI1
res
panI2
panI1
F1oF2
CUFC
NPoP
sets
(b)
200
150
100
50
0
numbers
RoK-P
t100-P
top10-NP
t100-NP
q2m2-P
top10-P
q2m2-NP
res-P
NRI1-NP
CUFC-P
NRI1-P
RoK-NP
res-NP
NRI1-nd
CUFC-NP
res-NPoP
subsets
(c)
20
15
10
5
0
numbers
ter+
t100
panI1
F1oP1
q2m2
F2oP1
I2oF1
NRI1
F1oP2
CUFC
top10
F2oP2
res
I1oF2
F1oF2
NPoP
panI2
sets

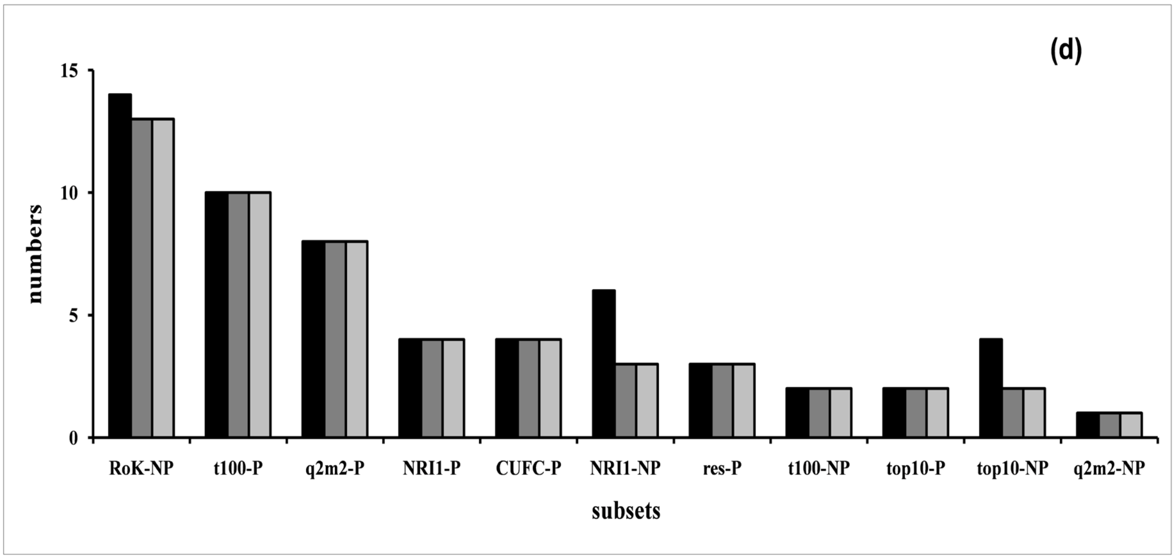
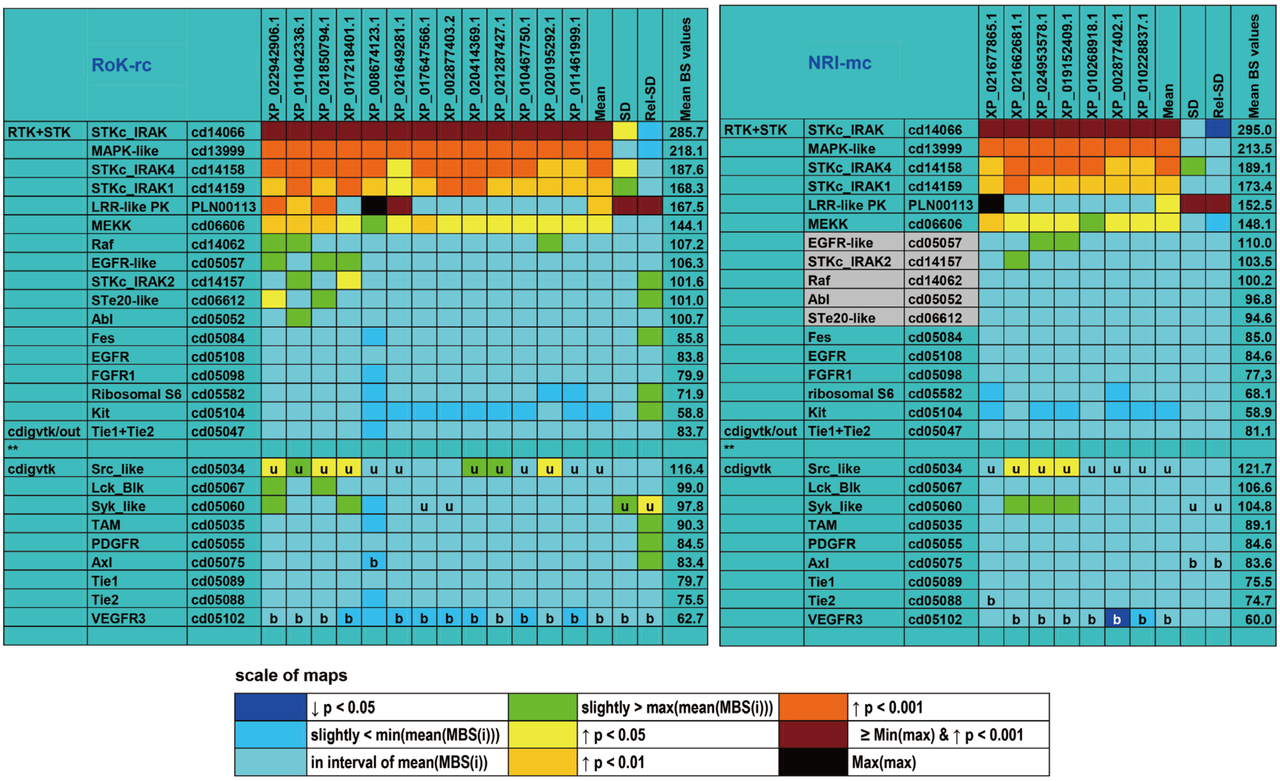


Figure 3. Introductory evaluation of the selected Ig-similar and Ig-domain-related sets and subsets of cds-files. The heights of columns shown here illustrate the numbers of sequence IDs and the corresponding cds-files (the files recording sequence-ID-derived conserved domain similarities, *i.e.* cds) in the selected sets and subsets. Set/subset-related clusters of columns are arranged descendently according to the heights of their second columns. Terms res and t100 substitute here set names resis and t100dm, respectively. For explanation of full set and subset names see Figure 2 or WP5. **Sections (a) and (b):** Sizes of sets and subsets. Colors: black—primarily derived sets and subsets; gray—sets and subsets obtained after a two-step procedure abolishing sequence redundancy (cf. Section 2.8, Figure 2 and WP1.2). **Sections (c) and (d):** Coincidences of term "immunoglobulin" and presence of cds with cd05034 in the selected sets and subsets. Only positive subset occurrences were displayed here. Extremely high and almost the same NCF (*i.e.* 113 for term "immunoglobulin" and 112 for other two cases found in RoK set and 99 resulting in all cases of RoK-P subset) were not comprised into the graph due to worsened resolution of complete presentation. Colors: black—searched term immunoglobulin; dark gray—cds with cd05034 (cf. Sections 2.9 and 3.2); light gray—coincidence of both indicated items in each enumerated cds-file.

restricted in cases of domain keywords (cf. Sections 2.7 and WP2.1); cdigvtk(max), cdigvtk(tc) denote subsets including cds with any of cdigvtk or all cdigvtk, respectively; cd05034, cd05102, cd14066, PLN00113 represent subsets including cds with catalytic domains of kinases, *i.e.* **Src-kinase-like** family, tyrosine **kinase of VEGF receptor 3**, **STKc-IRAK** (*i.e.* interleukin-1 receptor associated serine/threonine kinases; mostly dominant cds co-locating with cdigvtk) or **LRR-associated kinases** (achieving global maxima in Figure 4), respectively; I3R constitutes the search for two types of **global keywords** of plant immunity, *i.e.* conjugation of strategy immun3 (cf. Figure 5) and the term resistance (**I3R** is significantly associated with RoK as follows from specifically increased NCF and NCF-derived fractions of cds files, *i.e.* **FCF**; for both NCF and FCF holds $p < 0.001$; for details see WP3.4); Ig denotes a subset of cds-files containing the term **immunoglobulin**.

The only difference in a single cds-file can be seen in Formula (3), when comparing large RoK-derived subsets **Ig[113]** with the **cdigvtk(tc)[112]**. The presented subset hierarchy in RoK set and data in Figure 4 moreover suggest phylogenic importance of two kinases, *i.e.* PLN00113 and cd14066 (see Section 4.3).



RoK-rc			XP_022942906.1	XP_011042336.1	XP_021850794.1	XP_017218401.1	XP_008674123.1	XP_021649281.1	XP_017647566.1	XP_002877403.2	XP_020414369.1	XP_021287427.1	XP_010467750.1	XP_020195292.1	XP_011461999.1	Mean	SD	Rel-SD	Mean BS values
RTK+STK	STKc_IRAK	cd14066																	285.7
	MAPK-like	cd13999																	218.1
	STKc_IRAK4	cd14158																	187.6
	STKc_IRAK1	cd14159																	168.3
	LRR-like PK	PLN00113																	167.5
	MEKK	cd06606																	144.1
	Raf	cd14062																	107.2
	EGFR-like	cd05057																	106.3
	STKc_IRAK2	cd14157																	101.6
	STe20-like	cd06612																	101.0
	Abl	cd05052																	100.7
	Fes	cd05084																	85.8
	EGFR	cd05108																	83.8
	FGFR1	cd05098																	79.9
	Ribosomal S6	cd05582																	71.9
	Kit	cd05104																	58.8
cdigvtk/out	Tie1+Tie2	cd05047																	83.7
**																			
cdigvtk	Src_like	cd05034	u	u	u	u	u	u			u	u	u	u	u	u			116.4
	Lck_Blk	cd05067																	99.0
	Syk_like	cd05060							u	u							u	u	97.8
	TAM	cd05035																	90.3
	PDGFR	cd05055																	84.5
	Axl	cd05075					b												83.4
	Tie1	cd05089																	79.7
	Tie2	cd05088																	75.5
	VEGFR3	cd05102	b	b	b	b		b	b	b	b	b	b	b	b	b	b	b	62.7

NRI-mc			XP_021677865.1	XP_021662681.1	XP_024953578.1	XP_019152409.1	XP_010268918.1	XP_002877402.1	XP_010228837.1	Mean	SD	Rel-SD	Mean BS values
RTK+STK	STKc_IRAK	cd14066											295.0
	MAPK-like	cd13999											213.5
	STKc_IRAK4	cd14158											189.1
	STKc_IRAK1	cd14159											173.4
	LRR-like PK	PLN00113											152.5
	MEKK	cd06606											148.1
	EGFR-like	cd05057											110.0
	STKc_IRAK2	cd14157											103.5
	Raf	cd14062											100.2
	Abl	cd05052											96.8
	STe20-like	cd06612											94.6
	Fes	cd05084											85.0
	EGFR	cd05108											84.6
	FGFR1	cd05098											77,3
	ribosomal S6	cd05582											68.1
	Kit	cd05104											58.9
cdigvtk/out	Tie1+Tie2	cd05047											81.1
**													
cdigvtk	Src_like	cd05034	u	u	u	u	u	u	u	u			121.7
	Lck_Blk	cd05067											106.6
	Syk_like	cd05060									u	u	104.8
	TAM	cd05035											89.1
	PDGFR	cd05055											84.6
	Axl	cd05075									b	b	83.6
	Tie1	cd05089											75.5
	Tie2	cd05088	b										74.7
	VEGFR3	cd05102		b	b	b	b	b	b	b			60.0

scale of maps

🟦	↓ p < 0.05	🟩	slightly > max(mean(MBS(i)))	🟧	↑ p < 0.001
🟦	slightly < min(mean(MBS(i)))	🟨	↑ p < 0.05	🟥	≥ Min(max) & ↑ p < 0.001
🟦	in interval of mean(MBS(i))	🟨	↑ p < 0.01	⬛	Max(max)

Figure 4. Statistically based comparison of co-locating cds indicating RTK- or STRK-related catalytic domains. The extents of reciprocally co-locating robust cds with catalytic kinase domains were evaluated in two preformed model subsets of cds-files limited by E ≤ 0.01 and content of the term immunoglobulin (*i.e.* RoK-rc and NRI-mc; for details see below). Bit score of cds with cdigvtk (*i.e.* catalytic tyrosine kinase domains accompanying IgV domains in molecules of *Metazoa*; cf. Section 2.9) were used as compared model bit scores (MBS(i)) to perform united t-test statistics concerning subset-related sample sizes 117 and 63 (see WP2.2; cf. also [35]). This enabled us to made two color maps of statistical relationships. As follows from these maps and cds-records, PLN00113 achieved extreme variability of bit score values in its cds (maximum SD and Rel-SD) accompanied by considerable differences in cds lengths (data not shown) and maximum bit score values in both mentioned subsets (cf. black elements). In contrast, minimum Rel-SD was found in rows of cd14066 mostly superior in individual cases of evaluated sequence IDs. Regular records Ras (cd00882), IRAK3 (pseudogene), Lyk3 and Sek (neither found among CDD sequence IDs) were not found in our cds-files. Abbreviations: b, u—bottom and upper bit score values within cdigvtk set in evaluated cds-file; BS—bit score; cdigvtk/out—domain cds excluded from the compared model set due to repeated Tie specificity; gray background—slight differences in mean-score-derived order were observed in RoK-rc and NRI-mc subsets; in interval of mean(MBS(i))—the evaluated value occurs within minimum interval containing all mean values of MBS(i), *i.e.* Mean BS values occurring in cdigvtk-related bottom parts of both maps; Max(max), Min(max)—global picture-related maximum or minimum of column-related maxima, respectively; Mean—mean score value in color representation; NRI-mc—multi-connective files (cf. the Section 3.3) representing independently also all cds-files containing the observed kinase domains present in NRI1; Rel-SD—relative standard deviation (Rel-SD = SD/Mean); RoK-rc—cds-files randomly selected from RoK subset of cds-files including the term immunoglobulin but not any of the five files composing NRI-mc subset (number of selected sequences corresponded to the formula n = ceil(N/10), where N = 121 denoted number of cds-files in RoK set); STK/STKc—catalytic domains specifically phosphorylating mainly serines or threonines (cf. Section 4.3); SD—standard deviation of selective mean value; slightly > max(mean(MBS(i))), slightly < min(mean(MBS(i)))—non-significantly different values slightly higher or lower than maximum or minimum mean(MBS(i)) values (see above), respectively; ↓, ↑—bottom and upper number limits following from the adequate two-tailed t-test statistics were used, respectively. For additional abbreviations see WP4 or WP5.

3.3. Monitoring of Conserved Domain Similarities and Terms in NS1F-Related Sets of cds-Files

In accordance with the results in Table 1, we introduced here five selected color

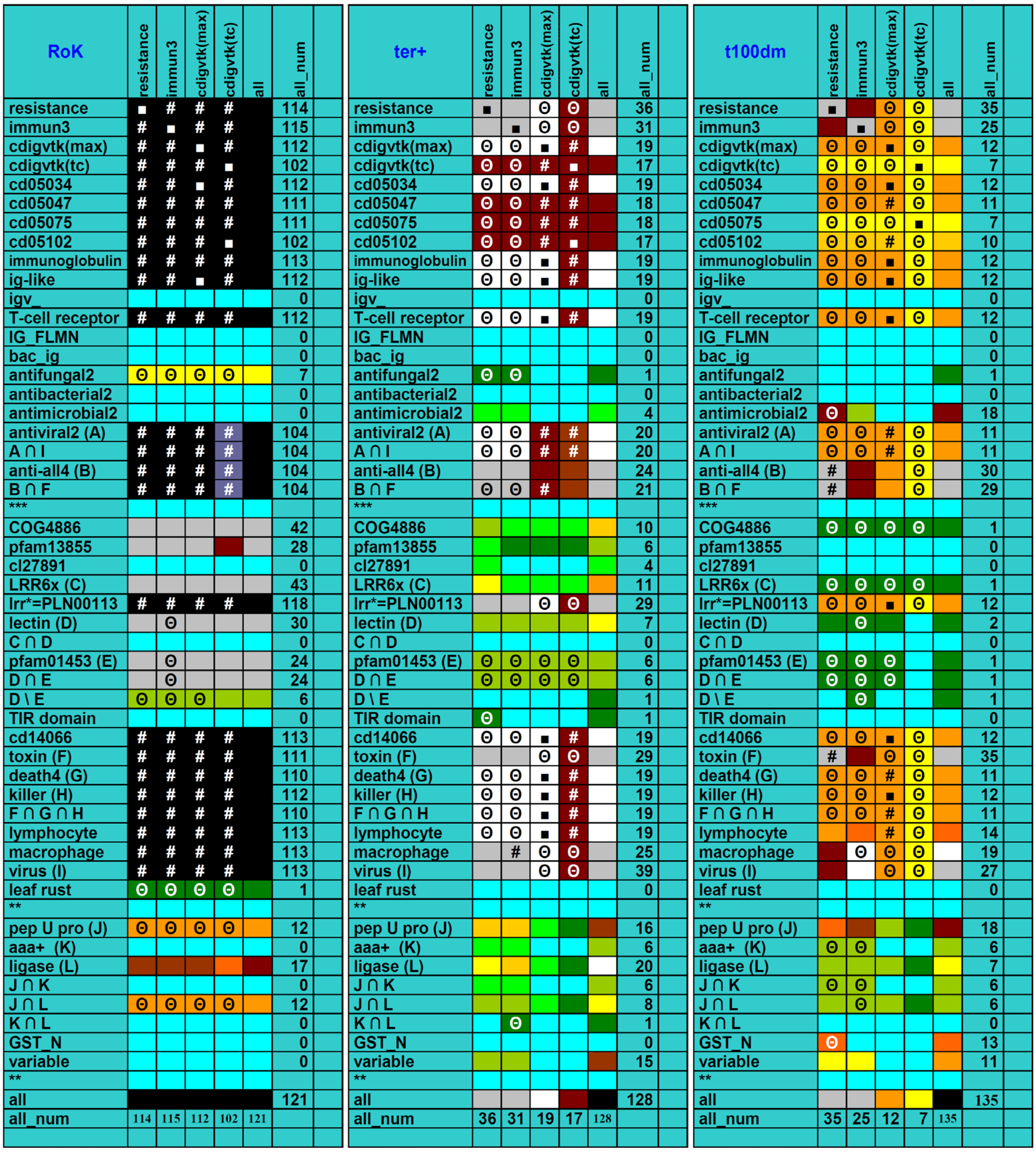
Table 1. Occurrences of cdigvtk-derived cds in focus of simple comparative approach.

sets and subsets	mc	mc-	mc+	mc+/-	all_num	cdigvtk(max)	cdigvtk(tc)	fractions of cdigvtk(max)	fractions of cdigvtk(tc)
CUFC-NP	*10*	*0*	*0*	*0*	*2*	0	0	0	0
CUFC-P	*10*	*0*	*0*	*0*	*26*	4	4	0.1538	**0.1538**
CUFC	*10*	*0*	*0*	*0*	*28*	4	4	0.1429	0.1429
F1oF2	10	0	0	0	**43**	2	2	0.0465	0.0465
F1oP1	9	1	0	0	**54**	9	**8**	**0.1667**	0.1481
F1oP2	6	0	4	0	**67**	7	6	0.1045	0.0896
F2oP1	7	0	3	0	**79**	8	7	0.1013	0.0886
F2oP2	10	0	0	0	**68**	3	3	0.0441	0.0441
I1oF2	9	0	1	0	**138**	4	3	0.029	0.0217
I2oF1	2x4	2	0	0	**81**	8	2	0.0988	0.0247
NPoP	*6*	*0*	*4*	*0*	*24*	2	1	0.0833	0.0417
NRI1	9	1	0	0	**49**	7	6	01429	0.1224
NRI1-NP	*10*	*0*	*0*	*0*	*28*	3	3	0.1071	0.1071
NRI1-P	*9*	*1*	*0*	*0*	*19*	4	3	**0.2105**	**0.1579**
NRI1-nd	*10*	*0*	*0*	*0*	*2*	0	0	**0**	**0**
panI1	10	0	0	0	**45**	10	10	**0.2222**	**0.2222**
panI2	10	0	0	0	**46**	2	2	0.0435	0.0435
q2m2-NP	*10*	*0*	*0*	*0*	*39*	1	1	0.0256	0.0256
q2m2-P	5	0	4	1	**57**	8	5	0.1404	0.0877
q2m2	5	0	4	1	**96**	9	6	0.0938	0.0625
resis	10	0	0	0	**49**	3	3	0.0612	0.0612
RoK	3	4	1	2	**121**	**112**	**102**	**0.9256**	**0.8347**
RoK-NP	*7*	*2*	*0*	*1*	*17*	**13**	**12**	**0.7647**	**0.7059**
RoK-P	4	2	3	1	**104**	**99**	**90**	**0.9519**	**0.8654**
t100dm-NP	10	0	0	0	**65**	2	2	0.0308	0.0308
t100dm-P	3	7	0	0	**70**	10	5	0.1429	0.0714
t100dm	2x3	4	0	0	**135**	**12**	7	0.0889	0.0519
ter+	5	1	4	0	**128**	**19**	**17**	0.1484	0.1328
top10-NP	10	0	0	0	**68**	2	2	0.0294	0.0294
top10-P	8	2	0	0	**53**	2	1	0.0377	0.0189
top10	8	2	0	0	**121**	4	3	0.0331	0.0248
Selected limits for p < 0.05									
One-tailed						-	**9.300**	**0.1776**	**0.1544**
Two-tailed						**12.696**	-	-	-

Current t-test evaluation was used to indicate remote (statistically deviated) occurrences of cds-files with cds including cdigvtk, when evaluating numbers or fractions of such cds-files. Collection of PHI-BLAST derived F-sets (cf. Figure 2) represented here a model reference entity due to lucid and simple definition of its non-tendentiously selected sets. **Abbreviations**: all_num—number of cds-files (**NCF**) in an evaluated set; cdigvtk—tyrosine kinases associated with IgV domains broadly occurring in *Metazoa* (cf. Section 2.9); cdigvtk(max), cdigvtk(tc)—NCF containing at least one or all cdigvtk-derived cds, respectively (cf. WP2.1); mc—number of most frequently selected subsets (**mfs**) with the same file content, when using ten different cdigvtk (the expression 2x denotes the existence of two different maximum subsets); mc-, mc+—frequency of subsets which can be exclusively derived when either diminishing or extending the mfs, respectively; mc+/-—frequency of subsets which can be derived only when simultaneously diminishing and extending the mfs; bold in the column all_num—sufficient sample sizes. **Colors in backround**: gray—compared F-sets; black—significantly different values. For abbreviations see also Figure 2, WP4 or WP5.

maps, *i.e.* four maps of NS1F-sets achieving superior values in their upper sections and Ig-domain associated set NRI1. These maps demonstrated also certain relationships between the subsets within the observed sets (**Figure 5**; for color grading see Sections 2.11 and WP1.3).

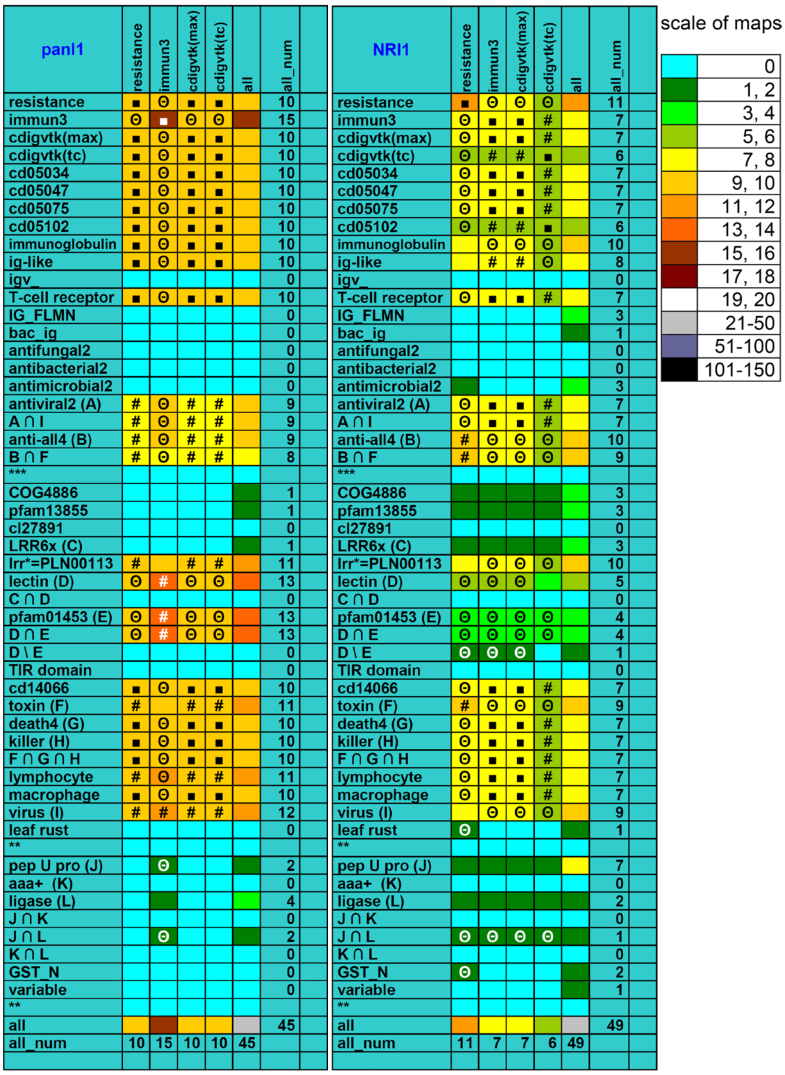
Four alternative search strategies each looking for the occurrence of the same term written by two alternative entries (**S2E**) were described in **Figure 5**. The numbers of cds-files obtained with S2E antiviral2 in the RoK set (104) significantly differed from the corresponding numbers achieved in other sets of NS1F ($p <$ 0.001 in t-test). 97 of these 104 cds-files contained cds of all cdigvtk. The records containing cdigvtk and terms of S2E antiviral2 were also present in all seven cds-files composing the RoK subset restricted with S2E antifungal2. In contrast,



RoK	resistance	immun3	cdigvtk(max)	cdigvtk(tc)	all	all_num	
resistance	■	#	#	#		114	
immun3	#	■	#	#		115	
cdigvtk(max)	#	#	■	#		112	
cdigvtk(tc)	#	#	#	■		102	
cd05034	#	#	■	#		112	
cd05047	#	#	#	#		111	
cd05075	#	#	#	#		111	
cd05102	#	#	#	■		102	
immunoglobulin	#	#	#	#		113	
ig-like	#	#	■	#		112	
igv_						0	
T-cell receptor	#	#	#	#		112	
IG_FLMN						0	
bac_ig						0	
antifungal2	Θ	Θ	Θ	Θ		7	
antibacterial2						0	
antimicrobial2						0	
antiviral2 (A)	#	#	#	#		104	
A ∩ I	#	#	#	#		104	
anti-all4 (B)	#	#	#	#		104	
B ∩ F	#	#	#	#		104	
***							
COG4886						42	
pfam13855						28	
cl27891						0	
LRR6x (C)						43	
lrr*=PLN00113	#	#	#	#		118	
lectin (D)		Θ				30	
C ∩ D						0	
pfam01453 (E)		Θ				24	
D ∩ E		Θ				24	
D \ E	Θ	Θ	Θ			6	
TIR domain						0	
cd14066	#	#	#	#		113	
toxin (F)	#	#	#	#		111	
death4 (G)	#	#	#	#		110	
killer (H)	#	#	#	#		112	
F ∩ G ∩ H	#	#	#	#		110	
lymphocyte	#	#	#	#		113	
macrophage	#	#	#	#		113	
virus (I)	#	#	#	#		113	
leaf rust	Θ	Θ	Θ	Θ		1	
**							
pep U pro (J)	Θ	Θ	Θ	Θ		12	
aaa+ (K)						0	
ligase (L)						17	
J ∩ K						0	
J ∩ L	Θ	Θ	Θ	Θ		12	
K ∩ L						0	
GST_N						0	
variable						0	
**							
all						121	
all_num	114	115	112	102	121		

ter+	resistance	immun3	cdigvtk(max)	cdigvtk(tc)	all	all_num	
resistance	■		Θ	Θ		36	
immun3		■	Θ	Θ		31	
cdigvtk(max)	Θ	Θ	■	#		19	
cdigvtk(tc)	Θ	Θ	#	■		17	
cd05034	Θ	Θ	■	#		19	
cd05047	Θ	Θ	#	#		18	
cd05075	Θ	Θ	#	#		18	
cd05102	Θ	Θ	#	■		17	
immunoglobulin	Θ	Θ	■	#		19	
ig-like	Θ	Θ	■	#		19	
igv_						0	
T-cell receptor	Θ	Θ	■	#		19	
IG_FLMN						0	
bac_ig						0	
antifungal2	Θ	Θ				1	
antibacterial2						0	
antimicrobial2						4	
antiviral2 (A)	Θ	Θ	#	#		20	
A ∩ I	Θ	Θ	#	#		20	
anti-all4 (B)						24	
B ∩ F	Θ	Θ	#			21	
***							
COG4886						10	
pfam13855						6	
cl27891						4	
LRR6x (C)						11	
lrr*=PLN00113			Θ	Θ		29	
lectin (D)						7	
C ∩ D						0	
pfam01453 (E)	Θ	Θ	Θ	Θ		6	
D ∩ E	Θ	Θ	Θ	Θ		6	
D \ E						1	
TIR domain	Θ					1	
cd14066	Θ	Θ	■	#		19	
toxin (F)			Θ	Θ		29	
death4 (G)	Θ	Θ	■	#		19	
killer (H)	Θ	Θ	■	#		19	
F ∩ G ∩ H	Θ	Θ	■	#		19	
lymphocyte	Θ	Θ	■	#		19	
macrophage		#	Θ	Θ		25	
virus (I)			Θ	Θ		39	
leaf rust						0	
**							
pep U pro (J)						16	
aaa+ (K)						6	
ligase (L)						20	
J ∩ K						6	
J ∩ L						8	
K ∩ L		Θ				1	
GST_N						0	
variable						15	
**							
all						128	
all_num	36	31	19	17	128		

t100dm	resistance	immun3	cdigvtk(max)	cdigvtk(tc)	all	all_num	
resistance	■		Θ	Θ		35	
immun3		■	Θ	Θ		25	
cdigvtk(max)	Θ	Θ	■	Θ		12	
cdigvtk(tc)	Θ	Θ	Θ	■		7	
cd05034	Θ	Θ	■	Θ		12	
cd05047	Θ	Θ	#	Θ		11	
cd05075	Θ	Θ	Θ	■		7	
cd05102	Θ	Θ	#	Θ		10	
immunoglobulin	Θ	Θ	■	Θ		12	
ig-like	Θ	Θ	■	Θ		12	
igv_						0	
T-cell receptor	Θ	Θ	■	Θ		12	
IG_FLMN						0	
bac_ig						0	
antifungal2						1	
antibacterial2						0	
antimicrobial2	Θ					18	
antiviral2 (A)	Θ	Θ	#	Θ		11	
A ∩ I	Θ	Θ	#	Θ		11	
anti-all4 (B)	#			Θ		30	
B ∩ F	#			Θ		29	
***							
COG4886	Θ	Θ	Θ	Θ		1	
pfam13855						0	
cl27891						0	
LRR6x (C)	Θ	Θ	Θ	Θ		1	
lrr*=PLN00113	Θ	Θ	■	Θ		12	
lectin (D)		Θ				2	
C ∩ D						0	
pfam01453 (E)	Θ	Θ	Θ			1	
D ∩ E	Θ	Θ	Θ			1	
D \ E		Θ				1	
TIR domain						0	
cd14066	Θ	Θ	■	Θ		12	
toxin (F)	#		Θ	Θ		35	
death4 (G)	Θ	Θ	#	Θ		11	
killer (H)	Θ	Θ	■	Θ		12	
F ∩ G ∩ H	Θ	Θ	#	Θ		11	
lymphocyte			#	Θ		14	
macrophage		Θ	Θ	Θ		19	
virus (I)			Θ	Θ		27	
leaf rust						0	
**							
pep U pro (J)						18	
aaa+ (K)	Θ	Θ				6	
ligase (L)						7	
J ∩ K	Θ	Θ				6	
J ∩ L		Θ				6	
K ∩ L						0	
GST_N	Θ					13	
variable						11	
**							
all						135	
all_num	35	25	12	7	135		



panI1	resistance	immun3	cdigvtk(max)	cdigvtk(tc)	all	all_num
resistance	■	Θ	■	■		10
immun3	Θ	■	Θ	Θ		15
cdigvtk(max)	■	Θ	■	■		10
cdigvtk(tc)	■	Θ	■	■		10
cd05034	■	Θ	■	■		10
cd05047	■	Θ	■	■		10
cd05075	■	Θ	■	■		10
cd05102	■	Θ	■	■		10
immunoglobulin	■	Θ	■	■		10
ig-like	■	Θ	■	■		10
igv_						0
T-cell receptor	■	Θ	■	■		10
IG_FLMN						0
bac_ig						0
antifungal2						0
antibacterial2						0
antimicrobial2						0
antiviral2 (A)	#	Θ	#	#		9
A ∩ I	#	Θ	#	#		9
anti-all4 (B)	#	Θ	#	#		9
B ∩ F	#	Θ	#	#		8
***						
COG4886						1
pfam13855						1
cl27891						0
LRR6x (C)						1
lrr*=PLN00113	#		#	#		11
lectin (D)	Θ	#	Θ	Θ		13
C ∩ D						0
pfam01453 (E)	Θ	#	Θ	Θ		13
D ∩ E	Θ	#	Θ	Θ		13
D \ E						0
TIR domain						0
cd14066	■	Θ	■	■		10
toxin (F)	#		#	#		11
death4 (G)	■	Θ	■	■		10
killer (H)	■	Θ	■	■		10
F ∩ G ∩ H	■	Θ	■	■		10
lymphocyte	#	Θ	#	#		11
macrophage	■	Θ	■	■		10
virus (I)	#	#	#	#		12
leaf rust						0
**						
pep U pro (J)		Θ				2
aaa+ (K)						0
ligase (L)						4
J ∩ K						0
J ∩ L		Θ				2
K ∩ L						0
GST_N						0
variable						0
**						
all						45
all_num	10	15	10	10	45	

NRI1	resistance	immun3	cdigvtk(max)	cdigvtk(tc)	all	all_num
resistance	■	Θ	Θ	Θ		11
immun3	Θ	■	■	#		7
cdigvtk(max)	Θ	■	■	#		7
cdigvtk(tc)	Θ	#	#	■		6
cd05034	Θ	■	■	#		7
cd05047	Θ	■	■	#		7
cd05075	Θ	■	■	#		7
cd05102	Θ	#	#	■		6
immunoglobulin		Θ	Θ	Θ		10
ig-like		#	#	Θ		8
igv_						0
T-cell receptor	Θ	■	■	#		7
IG_FLMN						3
bac_ig						1
antifungal2						0
antibacterial2						0
antimicrobial2						3
antiviral2 (A)	Θ	■	■	#		7
A ∩ I	Θ	■	■	#		7
anti-all4 (B)	#	Θ	Θ	Θ		10
B ∩ F	#	Θ	Θ	Θ		9
***						
COG4886						3
pfam13855						3
cl27891						0
LRR6x (C)						3
lrr*=PLN00113		Θ	Θ	Θ		10
lectin (D)	Θ	Θ	Θ			5
C ∩ D						0
pfam01453 (E)	Θ	Θ	Θ	Θ		4
D ∩ E	Θ	Θ	Θ	Θ		4
D \ E	Θ	Θ	Θ			1
TIR domain						0
cd14066	Θ	■	■	#		7
toxin (F)	#	Θ	Θ	Θ		9
death4 (G)	Θ	■	■	#		7
killer (H)	Θ	■	■	#		7
F ∩ G ∩ H	Θ	■	■	#		7
lymphocyte	Θ	■	■	#		7
macrophage	Θ	■	■	#		7
virus (I)		Θ	Θ	Θ		9
leaf rust	Θ					1
**						
pep U pro (J)						7
aaa+ (K)						0
ligase (L)						2
J ∩ K						0
J ∩ L	Θ	Θ	Θ	Θ		1
K ∩ L						0
GST_N	Θ					2
variable						1
**						
all						49
all_num	11	7	7	6	49	

scale of maps
0
1, 2
3, 4
5, 6
7, 8
9, 10
11, 12
13, 14
15, 16
17, 18
19, 20
21-50
51-100
101-150

Figure 5. Monitoring of cds- and term-related subsets of cds-files in four selected NS1F-related sets and NRI1. Three lengthwise sections are separated by rows with "**" in each of five set-related pictures forming this figure. The first sections describe the occurrence (number) of Ig-related terms (terms and cds-associated domain identifiers) in cds-files of the evaluated sets. The second sections concern immunologically important terms, whereas the third sections comprises interesting or somewhere frequent terms. For our solution of problems with dissociation of conserved domain keywords/identifiers from cds records see Section 2.7 and WP2.1. Icons defining the extent of conjugations between subset pairs defined by rows and columns: ▪—subset identities; #—at least 80% of cds-files are present in both subsets; Θ—one subset is fully included in the counterpart subset (cf. also all_num values). Abbreviations: all, al_num—color and numbers related to all items in the corresponding row or column, respectively; antifungal2, antibacterial2, antimicrobial2, antiviral2—all terms ahead of number 2 were scanned two times (with or without dash after term anti) to get a fusion subset; anti-all4—all four preceding sets were fused to a unique set; bac_ig—bacterial Ig-like domains whose accessions begin with big_ or BID_; cdigvtk(max), cdigvtk(tc)—the presence of any or all cds recording cdigvtk domains was required, respectively (cf. 2.9); cl27891—domain of the third LRR superfamily; death4—the subset of cds-files comprises all the following four terms: programmed cell death, apoptosis, necrosis, caspase; GST_N—glutathione S-transferase family; immun3—fusion of three subsets containing terms: immune, immunity, immunol; IG_FLMN, ig_v—conventional abbreviations denoting filamin-type Ig domains or multiple different variable Ig domains, respectively; LRR6x—only six leucine-rich repeat domains (LRR) formed cds-records in all our sets yielding this fusion subset, *i.e.* COG4886, pfam13855, cl27891, pfam12799, pfam14580 and pfam07725 (cf. WP1.4 and WP3.3); lrr* = PLN00113—the term "leucine-rich repeat" in fact restricts the same cds-files like cds with PLN00113 but much less frequently cds with LRR domains; pep U pro—a set with the terms peptidase or protease; pfam01453—domain of mannose specific lectins; TIR domain—toll-like/interleukin-1 receptor domain. For additional abbreviations see WP4 and WP5.

a unique set-derived subset of significantly increased size (NCF = 17; p < 0.01 in t-test) was selected with S2E antifungal2, *i.e.* a subset of F2oP1, where sixteen cds-files included robust cds with the domain of antifungal thaumatin-like proteins. This subset comprised a unique cds-file independently containing all cdigvtk-related accessions and terms of S2E antiviral2 but not the terms of S2E antimicrobial2. The cds-file of the same content like this file composed also the RoK set. The terms of S2E antimicrobial2 were not observed in RoK, but it still exhibited significantly increased occurrence in the set t100dm set (NCF = 18; p < 0.01 in t-test). The corresponding subset of exclusive term occurrence, *i.e.* t100dm-NP, did not include cds-files with cdigvtk-related accessions and the other compared terms of S2E antiviral2 or antifungal2. However, only six of the selected cds-files realized in fact the aim of this S2E indicating robust cds with ABC-type antimicrobial peptide transport system.

Twenty cds-files with TIR domains (item **tir dom** in Figure 5) were found in the **resis** set composed of cds-records generated by sequence IDs which contained the term resistance in the molecular title (cf. Figure 2). Other sets significantly differed from such extreme TIR domain occurrence in the number and density values (at least p < 0.001 in t-test; cf. Figure 5; the corresponding NCF achieved only values in interval 0 - 3 in other sets). Significantly increased values of densities concerned occurrences of LRR6x- and NB-ARC-related cds-files in TIR domain-derived subset of resis set (both 18 of 19 cds-files) and NB-ARC domains (35 of 38 cds-files) in the subset restricted with LRR6x (including all found LRR domains; Table 2). Simultaneous occurrence of NB-ARC, LRR6x and TIR domains in cds-files indicated the presence of 18 NLR receptors in the resis set. Additional distinct six sequences (cds-files) alternatively comprised NB-ARC, LRR6x and coiled-coil domains in accordance with alternative NLR composition described in Introduction.

Most of cds-files containing the term "-lectin" comprised the domain with accession pfam01453 (*i.e.* mannose specific lectin). Term lrr* (leucine-rich repeats) revealed mostly kinase domains associated with LRR but much less frequently LRR domains (for details concerning distribution of LRR domains see WP3.3 and Figure 5).

Seven multi-connective cds-files (five of RoK origin) were independently found in NRI1-set when: i) individually searching for the terms antiviral, killer, lymphocyte, macrophage, or the term programmed cell death; ii) combining the term resistance and strategy immun3 (cf. I3R mentioned above) or using conjugation strategy death4; and iii) looking for the occurrence of most frequent cdigvtk cd05034 or co-locating cds with cd14066 (Figure 5). The cds-records with TIR domains and aaa+ (superfamily of ATP-ases) were found neither in NRI1 nor in NRI2 set (Figure 5 and WP2.1). Apparent contradictions between NRI-related data present in Figure 5 and last table in fact followed only from the difference in sample sizes restricting content of cds-files composing NRI1 and NRI2 (see Figure 1).

Table 2. Statistical analysis of domain- (cds-) and term-related associations using 2 × 2 tables.

Set/subset A[a]	Set/subset B[a]	Q[a]	AQ+	AQ-	A[a]	BQ+	BQ-	B[a]	OR[b]	One-t[b]	Two-t[b]	SQ-eval[c]
i) Dominance of RoK in content of cdigvtk (extension of studies in Table 1)												
RoK	ter+	C(m)	112	9	121	19	109	128	71.4	7.72E−39	1.17E−38	**ssl_C**
RoK	ter+	C(t)	102	19	121	17	111	128	35.1	1.22E−31	1.30E−31	**ssl_C**
RoK	panI1	C(m)	112	9	121	10	35	45	43.6	1.07E−18	1.07E−18	**ssl_C**
RoK	panI1	C(t)	102	19	121	10	35	45	18.8	1.18E−13	1.18E−13	**ssl_C**
RoK	NRI1	C(m)	112	9	121	7	42	49	74.7	1.17E−23	1.17E−23	**ssl_C**
RoK	NRI1	C(t)	102	19	121	6	43	49	38.5	5.70E−19	5.70E−19	**ssl_C**
RoK	t100dm	C(m)	112	9	121	12	123	135	**128**	**1.03E−46**	**1.12E−46**	**ssl_C**
RoK	t100dm	C(t)	102	19	121	7	128	135	98.2	**2.81E−42**	**2.82E−42**	**ssl_C**
RoK	resis	C(m)	112	9	121	3	46	49	191	1.07E−28	1.07E−28	**ssl_C**
RoK	resis	C(t)	102	19	121	3	46	49	82.3	1.59E−22	1.59E−22	**ssl_C**
ii) Linkages between cdigvtk and ALPS-related subsets												
RoK-P	resub	C(m)	99	5	104	13	4	17	6.09	2.21E−2	2.21E−2	si_QC
RoK-P	resub	C(t)	90	14	104	12	5	17	2.68	9.87E−2	0.14214	q/poor_S
t100dm-P	resub	C(m)	10	60	70	2	63	65	5.25	2.11E−2	3.19E−2	si_QC
t100dm-P	resub	C(t)	5	65	70	2	63	65	2.42	0.25208	0.44297	poor_S
q2m2-P	resub	C(m)	8	49	57	1	38	39	6.20	5.66E−2	7.84E−2	q_QC
q2m2-P	resub	C(t)	5	52	57	1	38	39	3.65	0.21529	0.39581	poor_S
F1oP1	I2oF1	C(m)	9	45	54	8	73	81	1.83	0.18348	0.29343	poor_W
F1oP1	I2oF1	C(t)	8	46	54	2	79	81	6.87	9.71E−3	1.45E−2	si_QC
F1oP1	panI1	C*	9	45	54	10	35	45	0.7	0.32800	0.60969	poor_inv
iii) Associations of lectin domain pfam01453 (**P1453**)												
pan I1	F1oP1	P1453	13	32	45	3	51	54	6.91	1.83E−3	2.19E−3	si2_QC
RoK-NP	resub	P1453	9	8	17	15	89	104	6.68	9.81E−4	9.81E−4	si2_QC
RoK_nLRR6x	resub	P1453	24	54	78	0	43	43	20.0	5.83E−6	1.87E−5	si2_C
iv) Combined analysis and linkages of functionally related terms in different RoK subsets												
RoK-P	resub	antifu2	7	97	104	0	17	17	1.47	0.33645	0.59146	poor_W
RoK_nLRR6x	resub	antifu2	7	71	78	0	43	43	4.89	4.18E−2	9.57E−2	si/q_QC
RoK-P_nLRR6x	resub	antifu2	7	57	64	0	57	57	8.00	9.84E−3	1.40E−2	si_QC
RoK-P_nLRR6x_C(m)	resub	antifu2	7	56	63	0	49	49	7.02	1.53E−2	1.76E−2	si_QC
RoK-P_nLRR6x_C(t)	resub	antifu2	7	54	61	0	41	41	6.11	2.36E−2	3.97E−2	si_QC
RoK-P	resub	antivir2	93	11	104	11	6	17	4.61	1.51E−2	1.51E−2	si_QC
RoK_nLRR6x	resub	antivir2	71	7	78	33	10	43	3.07	3.13E−2	5.28E−2	si/q_S
RoK-P_nLRR6x	resub	antivir2	62	2	64	42	15	57	11.1	2.19E−4	3.51E−4	si2_C
RoK-P_nLRR6x_C(m)	resub	antivir2	62	1	63	42	7	49	10.3	1.23E−2	2.05E−2	si_C
RoK-P_nLRR6x_C(t)	resub	antivir2	60	1	61	37	4	41	6.49	0.08316	0.15459	q/poor_QC

Continued

v) Other selected linkages												
resis	RoK	cl27891	34	15	49	0	121	121	**267**	2.35E−24	2.35E−24	**ssl_C**
resis	NRI1/2	tir dom	19	30	49	0	49	49	32.3	2.18E−7	4.35E−7	**ssl_C**
resis	RoK	tir dom	19	30	49	0	121	121	78.7	2.73E−12	2.73E−12	**ssl_C**
resis_LRR6x	resub	tir dom	18	20	38	1	10	11	9	2.15E−2	3.35E−2	si_QC
resis_LRR6x	resub	NB-ARC	35	3	38	3	8	11	31.1	4.91E−5	4.91E−5	si2_C
resis_cl27891	resub	tir dom	18	16	34	1	14	15	15.8	1.85E−3	3.30E−3	si2_C
RoK_cd14066	resub	C(m)	112	1	113	0	8	8	**509**	1.00E−11	1.00E−11	**ssl_C**
RoK_cd14066	resub	C(t)	102	11	113	0	8	8	77.3	8.40E−8	8.40E−8	**ssl_C**
RoK_PLN00113	resub	C(m)	112	6	118	0	3	3	64.6	2.92E−4	2.92E−4	si2_C
RoK_PLN00113	resub	C(t)	102	16	118	0	3	3	24.2	3.36E−3	3.36E−3	si2_C
RoK_LRR6x	resub	ligase	11	32	43	6	72	78	4.13	8.42E−3	1.20E−2	si_S
t100-NP_cd04982#	resub	cd03053	12	6	18	1	46	47	92.0	5.37E−8	5.37E−8	**ssl_C**
vi) ALPS associations in the light of model evaluation using odds ratios[d]												
CUFC	GM	ALPS	26	2	28	ξ	η	91.5	5.14	1.77E−2	3.65E−2	si_QC
RoK	GM	ALPS	104	17	121	ξ	η	91.5	2.29	1.24E−2	2.29E−2	si_S

[a]Sets or pairs of complementary subsets were compared with respect to the differences in occurrence of a certain query terms (see **Q** below). _—conjunction; _n—conjunction with a complementary subset; AQ+, BQ+—subsets of sets/subsets A or B positive for presence of Q or associated cds (if Q was a domain identifier), respectively; cd03053—Phi subfamily of glutathione S-transferases; cd04982—IgV domain of the gamma chain of gamma/delta TCR; NB-ARC—domain pfam00931 forming NLR proteins (cf. Introduction); Q—terms used as search queries, *i.e.* as: i) regular conserved-domain-derived keywords associated with cds or domain-related terms (cf. Sections 2.7, WP1.5 and WP2.1) or ii) functionally related terms; resub—residual complementary subset of compared set. **Specifically shortened versions of abbreviations** otherwise used in this paper: antifu2—antifungal/anti-fungal; antivir2—antiviral/anti-viral; C(m), C(t)—cdigvtk(max), cdigvtk(tc) described in Table 1, respectively; C*—case, if C(m) = C(t). For additional abbreviations see Figure 2 or Figure 5 and WP4 or WP5. [b]Odds ratio values were enumerated conventionally or as zero-associated Bayesian OR* (cf. Section 2.10). Statistical evaluation of the corresponding 2 × 2 tables was performed by Fisher's exact test. **Two tailed exact Fisher's test** can be always assumed as valid, whereas one-tailed test consists in the usage of additional orienting conditions (sometimes depends on interpretation or context). #—for the details see the Section 3.6; a**E-n** = a × **10^{-n}**; One-t, Two-t—values following from one-tailed and two-tailed Fisher's test, respectively. [c]**Semi-quantitative odds-ratio-related evaluation** represents lucid insight into the presented data. Different comments to one and two tailed test are separated by "/", if such difference exists. **Classification of significance levels**: poor—insufficiently low or none validity ($p \geq 0.1$); q—quasi-significant (in this case $0.05 \leq p < 0.1$; cf. WP2.6); si—minimally significant ($0.005 \leq p < 0.05$); si2—of improved significance ($10^{-6} \leq p < 0.005$; cf. PSI BLAST limit); ssl—of superior significance ($p < 10^{-6}$). **Classification of linkages according to OR values**: C—causal linkage (OR ≥ 10); inv—inverse linkage (OR < 1/1.4); QC—quasi-causal linkage (4.5 ≤ OR < 10; interval added here by authors to the current classification; cf. WP2.3); S—strong linkage (2 ≤ OR < 4.5); W = weak (1.4 ≤ OR < 2). [d]The values ξ and η characterizing model mean set were enumerated here with the help of geometrical mean value (**GM**) of ratios derived from numbers of positivities and negativities typical for each topically compared/comparable set, respectively (cf. WP2.4; $\xi = 65.58424789$; $\eta = 25.91575211$).

3.4. Statistical Analysis of Linkages between Sets and Subsets of NS1F Based on Evaluation of 2 × 2 Tables

Three types of **strong significant statistical associations** following from robust cds can be seen in the **different segments of** Table 2: **i)** association of cdigvtk with RoK set or their related linkage to dominant STRK domain 14066 (segments i) and v)), **ii)** total disjunction of lectin (pfam01453) and LRR domains in the RoK set (segment iii)); for similar situation in the other sets see Figure 5), and **iii)** an extreme difference between the absence of cds with cd27891 and TIR domains in the RoK set (contrasting in case of LRR domain cl27891 with the presence of forty-three cds formed by distinct LRR in RoK) and their abundant co-occurrences in individual cds-files of resis set (segment v) and Figure 5). Though the linkages between cdigvtk and ALPS-related segments were mostly strong, they achieved **only boundary line** of significance (segment ii)).

3.5. Statistics of ALPS Occurrence

The occurrence of ALPS-selected cds-files (**ASC**; ASC represents molecules or sequence IDs selected with the assistance of sequences of ALPS forming two sequence queries MQP; cf. Section 2.2) was evaluated with respect to the number of selected cds-files (four sets with oP in the name and six subsets with -P in the name) and their relative occurrence (six pairs of subsets). Two values were significantly higher in t-test than the values in the compared collection of the residual sets (cf. WP2.2), *i.e.* i) maximum ratio (13) between ALPS-positive and such negative samples in CUFC set ($p < 0.01$) and ii) the number (104) of ASC in RoK set ($p < 0.05$). Significantly deviated maximum odds ratio (**OR**) and the best significance level of OR evaluation, selected CUFC and RoK sets as significantly ALPS-associated sets, respectively. For the data and methodology concerning model odds ratios see the sixth section of Table 2 and WP2.4, respectively.

3.6. Conserved Ig Domain Similarities in the Maximum Expect Extension of NSF2 and NRI2 Sets

Figure 6 represented not only a map of similarities in broad range of Expects but also starting point for screening of the corresponding Ig-cds (see also WP3.5). Among others the domain cd04982 (IgV_TCR_gamma) achieved significantly increased NCF ($p < 0.05$) in the t100dm-NP subset of NS2F (NCF = 18 holds also for the set t100dm). Twelve of the corresponding eighteen cd04982-related sequence IDs restricted cds-files specifically including cds with cd03053 (glutathione S-transferase GST_N_Phi frequent in *Arabidopsis* and *Oryza genomes*) in NS1F-related t100dm-NP set. This indicated a significant molecular association of short TCR-gamma-related peptides and cds with cd03053 (Table 2). These immunologically interesting peptides were found among others in certain species of vegetable and tobacco (for further comments see WP3.6).

3.7. More Detailed Description of the Selected "Non-Refutable" Ig-cds Forming NRI2 Set

Only **seven sequences** included "non-refutable" Ig-cds, which were called here as **dominant Ig-cds**, *i.e.* as Ig-cds that achieved the highest score and minimum Expect among cds co-locating with the same segments of individually evaluated sequences (cf. Table 3). Three and four of the selected sequences contained bacterial and metazoan dominant Ig-cds, respectively. Two significant and one quasi-significant **dominant Ig-cds including bacterial Ig domains** occurred in different sequence regions of nuclear pore complex protein GP21 (sequence ID XP_010248630.1; Table 3). All these three segments restricted by Ig-cds were approved by our Ig-fold evaluation. Another protein LOC105056499 (XP_010937019.2) was unique one selected by three different procedures. Its Ig-cds was composed

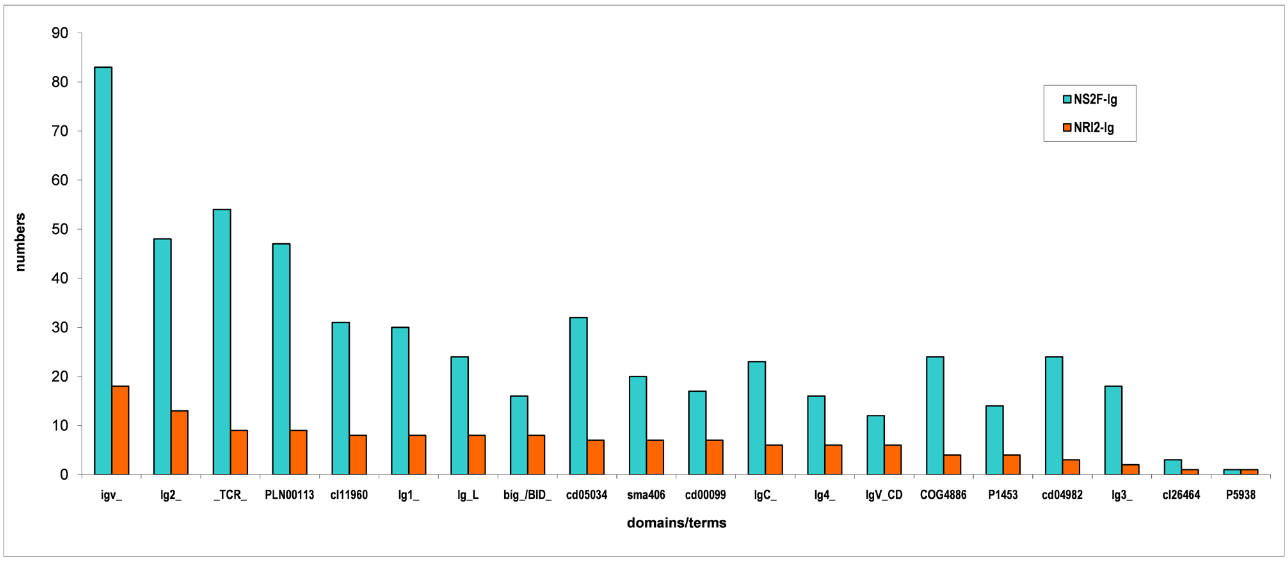


Figure 6. Occurrence of Ig-cds, Ig-cds-overlapping cds and Ig-domain-related terms in cds-files forming "expanded" NS2F and NRI2 sets. The cds limited by Expect values E ≤ 100, *i.e.* cds at least comparable with dense similarities between segments achieving lengths of secondary structures (cf. WP2.6), are described here. For distribution of re-selected "non-refutable" Ig-cds see Table 3. Abbreviations: big_/BID_—terms denoting major part of bacterial domains; cl26464—atrophin-1 family; igv_—cluster of IgV-related terms; NS2F-Ig, NRI2-Ig—content of Ig-cds and related terms or term clusters within non-redundant set NS2F-Ig (comprising 223 cds-files extracted from all sets of NS2F as subset of cds-files including Ig-cds records) and NRI2 described in Section 3.1, respectively; P1453—pfam01453 (domains of mannose specific lectins); P5938—pfam05938, *i.e.* domain of plant self-incompatibility protein S1; sma406—smart00406, *i.e.* IgV domain frequent in non-chordate *Metazoa*. For additional information and abbreviations see Figure 5 or Sections 2.11, WP1.5, and WP4.

of **typical metazoan and dominant Ig domain** Ig-cd Ig1_ IL1R_like. The segment restricted by Ig-cds achieved maximum fold score 6.81 in case of Ig light chain, but not prevailing number of required Ig folds (rule Q1 in Table 3). Nevertheless, the Ig-cds-related specificity of 96.3% was enumerated based on Expect values of all cds records overlapping the sequence segment in the cds-file of NRI2 origin (limited by E ≤ 100; for details see WP2.5). Other three dominant typical metazoan Ig-cds with sequences XP_021903093.1, XP_ 017226294.1, XP_022544643.1 achieved higher Expects than XP_010937019.2, but were fully approved by our folding analysis with FFAS03 (Table 3). Lower fraction (23.8%) of FFAS03-approved sequences was selected among forty-two "non-refutable" Ig-cds called here as **recessive Ig-cds** and representing Ig-cds co-locating with more valid non-Ig cds. Three types of interesting phenomena were observed in the case of recessive "non-refutable" Ig-cds: i) FFAS- and CDD-confirmed chimera of filamin domains and Ig-cds (see Table 3 and Section 4.1), ii) the existence of a plant self-incompatibility protein looking like a functional and structural analogue of plant T-cell receptor (E = 5.0 for co-locating weak cds with IgV-TCR_gamma) and iii) extremely frequent co-locations of **weak Ig-cds** (4.605 ≤ E ≤ 100) with "non-refutable" Ig-cds. The last phenomenon concerned mainly thirteen and fourteen weak Ig-cds co-locating with five and two also co-locating "non-refutable" Ig-cds found in cds-records of sequences XP_021662681.1 (ALE2 STRK) and XP_009381707.1 (potassium transporter) by CDD searches, respectively (cf. Table 3).

Table 3. Three groups of re-selected plant protein sequences forming "non-refutable" Ig-cds.

Title[a]	Sequence ID[a]	Dom[b]	cd-specification[b]	cd access[b]	Bit	Expect[b]	Peptide	CA[c]	CA	FM[d]	FR[d]	HME[e]	HMG[e]	HMT[e]
Species[a]	Search strategy[a]				score[b]		position[b]		position[c]					
i) combined cds- and fold-based selection of Ig-cds[f]														
nuclear pore complex protein	XP_010248630.1	max	**Big_2**	pfam02368	50.86	**1.04E−07**	1154 - 1224	2	1149 - 1224	−10.1	R1	3.6E−13	*Zobellia*	*Bacteroidetes*
GP210	top10-NP	mig	**BID_2**	smart00635	41.23	**2.35E−04**	1149 - 1224			−10.2	R1	6.9E−13	*Zobellia*	*Bacteroidetes*
Nelumbo nucifera		2ig	**BID_2**	smart00635	34.30	**0.06**	1561 - 1646	1	-	−10.4	R1	4.5E−16	*Flavobacterium*	*Bacteroidetes*
		3ig	**Big_2**	pfam02368	33.91	**0.08**	496 - 533	2	495 - 533	−10.5	R1	8.9E−17	*Neisseriaceae*	*Proteobacteria*
gamete expressed 2 i. X3	XP_019237668.1	max	Filamin	pfam00630	42.30	6.01E−05	92 - 222							
Nicotiana attenuate	I1oF2	2ig	**IG_FLMN**	smart00557	32.19	**0.20**	**92 - 228**	1	92 - 228	-	0	6.9E−13	*Natronorubrum*	*Archaea*
		mig	**IG_FLMN**	smart00557	37.58	**2.89E−03**	**289 - 329**	3	252 - 329	−6.19	0	**4.2E−27**	*Acyrthosiphon*	*Insecta*
		dii	Filamin	pfam00630	37.68	2.61E−03	224 - 324							
		-	**BID_1**	smart00634	31.14	**0.45**	**252 - 316**			−11.9	R1	**1.8E−35**	*Enterobacterales*	*Proteobacteria*
STRK ALE2 i. X1	XP_021662681.1	max	STKc_IRAK	cd14066	312.67	3.38E−98	765 - 1033							
Hevea brasiliensis	RoK-P	mig	**Ig1_Neogenin**	cd05722	36.69	**9.26E−03**	**288 - 335**	5	288 - 335	−8.22	R2	**1.5E−27**	*Pocillopora*	*Cnidaria*
		dii	Atrophin-1 SF	cl26464	94.62	2.49E−19	33 - 490							
		sit	PTKc_Src_like	cd05034	137.03	4.57E-36	763 - 961							
		iit	PTKc_VEGFR3	cd05102	65.39	5.05E-11	763 - 961							
u.p. LOC109158029	XP_019161394.1	max	Retrotran_ gag_3	cl28789	32.71	0.01	19 - 55							
Ipomoea nil	t100dm-NP	mig	**IgV**	cd00099	30.44	**0.26**	**50 - 77**	1	-	−7.96	R2	1.8E−05	*Tetranychus*	*Arachnida*
u.p. LOC108851022	XP_018479959.1	max	C1_2	pfam03107	43.92	4.18E-06	532 - 578							
Raphanus sativus	panI2	dii	C1_2	pfam03107	38.92	2.40E-04	117 - 166							
		mig	**IgV_CD8_beta**	cd07700	31.64	**0.35**	**122 - 145**	1	-	−7.31	R3	1.2E−11	*Ornithorhynchus*	*Mammalia*
u.p. LOC25498589	XP_013448976.1	max	DUF674	cl04913	404.72	3.55E−138	7 - 445							
Medicago truncatula	panI2	mig	**IgV_H**	d04981	31.15	**0.45**	**299 - 353**	3	299 - 353	−8.76	R2	0.096	*Mizuhopecten*	*Mollusca*
potassium transporter 19	XP_010027070.1	max	K_trans_SF	cl15781	825.53	0E+00	5 - 750							
Eucalyptus grandis	q2m2-P	mig	IG_like	smart00410	30.17	0.81	**610 - 646**	1	-	−6.69	R3	1.8E−20	*Crassostrea*	*Mollusca*

Continued

potassium transporter 5-like	XP_018716011.1	max	K_trans super family	cl15781	797.41	0E+00	2 - 758							
Eucalyptus grandis	q2m2-P	mig	**IG_like**	smart00410	30.17	**0.95**	**606 - 642**	2	599 - 642	−7.42	R3	**1.9E−34**	*Crassostrea*	*Mollusca*
kinase-like protein	NP_190213.2	max	Malectin_like	pfam12819	259.15	5.16E−83	27 - 347							
Arabidopsis thaliana	q2m2-P	dii	n.d.	PLN03150	70.23	9.16E−13	6-348							
		mig	**IgV_L_lambda**	cd04984	29.33	**1.5**	**17 - 47**	2	13 - 47	−8.22	Q2	1.2E−07	*Acinetobacter*	*Proteobacteria*
		-	**IgV_L_kappa**	cd04980	28.22	**3.6**	**13 - 43**			−7.85	R2	1.2E−09	*Acinetobacter*	*Proteobacteria*
potassium transporter 5	XP_009381707.1	max	K_trans SF	cl15781	908.74	0E+00	2 - 752							
Musa acuminata	q2m2-P, ter+	mig	**Ig2_PTK7**	cd05760	29.11	**1.6**	**615 - 640**	2	613 - 652	−7.75	R2	7.9E−21	*Stylophora*	*Cnidaria*
phospholipase SGR2	XP_021903093.1	max	DDHD	pfam02862	157.28	**3.20E−43**	794 - 989							
Carica papaya	F2oP2	mig	**Ig2_LILR_KL**	cd05711	30.10	**1.7**	**295 - 316**	1	-	−10.0	R1	1.3E−06	*Vulpes*	*Mammalia*
potassium channel i.X1	XP_018679675.1	max	Ank_2 SF	cl26073	398.86	1.45E−127	65 - 697							
KOR2-like	ter+	mig	**Ig1_PVR_like**	cd05718	28.93	**2.7**	**633 - 650**	2	604 - 655	−6.07	R3	6.8E−16	*Xenopus*	*Amphibia*
Musa acuminata		-	**ig**	pfam00047	28.31	**4.2**	**604 - 655**			−6.92	0	**7.9E−51**	*Metarhizium*	*Fungi*
microtubule-assoc. p. AIR9	XP_017226294.1	max	LRR	COG4886	65.37	1.10E−10	274 - 430							
Daucus carota	ter+	mig	Ig_3	pfam13927	29.39	**3.6**	**799 - 825**	1	-	−9.34	R2	5.1E−11	*Chelonia*	*Testudines*
u.p. **LOC106355531**	XP_022544643.1	max	RGS12_usC SF	cl24986	29.50	1.1	65 - 129							
Brassica napus	panI1	mig	**V-set**	pfam07686	27.44	**3.6**	120 - 162	1	-	−7.35	R3	0.17	*Flavobacterium*	*Bacteroidetes*

ii) residual significant (here p ≈ E ≤ 0.01 and p < E) and quasi-significant (0.01 ≤ p < 0.1) Ig-cds of weak (Q-related) fold relationship but sufficient domain specificity

u.p. **LOC105056499** i.X1	XP_010937019.2	max	**Ig1_IL1R_like**	cd05756	32.44	**0.09**	**228 - 267**	1	-	−6.81	**Q1**	8.8E−13	*Callipepla*	*Aves*
Elaeis guineensis	CUFC-P, ter+, q2m2-P	+mig												

iii) Ig-cds added by authors[8]

alpha-amylase/ subtilisin	XP_009376739.1	max	Kunitz_legume	pfam00197	159.38	7.23E−50	27 - 203							
	top10-NP	mig	**IgV_L_kappa**	cd04980	29.76	**0.35**	**100 - 158**	2	100-158	none	0	none	-	-
Pyrus x bretschneideri			**IgV_CD8_beta**	cd07700	27.41	**2.7**	**121 - 158**			−5.98	**Q3**	0.04	*Nematostella*	*Cnidaria*
u.p. **LOC109149200**	XP_019152409.1	max	STKc_IRAK	cd14066	281.08	1.13E−85	505 - 770							

Continued

Ipomoea nil	F1oP2, ter+	mig	**IGv**	smart00406	31.97	**0.40**	**53 - 106**	1	-	−7.73	0	2.7E−03	*Okeania*	*Cyanobacteria*
		mle	B_lectin	cd00028	137.06	1.40E−37	29 - 151							
		sit	PTKc_Src_like	cd05034	141.27	2.18E−37	504 - 698							
		iit	PTKc_VEGFR3	cd05102	65.39	8.00E−11	1416 - 1589							
u.p. **LOC108207541**	XP_017233469.1	max	Self-incomp_S1	pfam05938	104.22	5.96E−30	33 - 137							
Daucus carota	panI2	mig	**Ig5_Contactin -1**	cd05852	26.52	**2.2**	**67 - 109**	1	-	−7.17	0	1.7E−11	*Elysia*	*Mollusca*
LRR STRK RCH1	XP_021677865.1	max	PLN00113 SF	cl26793	497.45	7.31E−159	22 - 1075							
Hevea brasiliensis	RoK-P	mig	**Ig3_L1-CAML**	cd05731	28.94	**2.9**	**429 - 457**	1	-	none	0	**8.9E−31**	*Piromyces*	*Fungi*
		sit	PTKc_Src_like	cd05034	115.46	1.18E−28	798 - 1066							
		iit	PTKc_VEGFR3	cd05102	68.08	7.04E−12	797 - 1081							
G-type lectin S-STRK	XP_010228837.1	max	STKc_IRAK	cd14066	282.62	1.77E−89	504 - 770							
At2g19130	panI1, RoK-NP	mig	**IGv**	smart00406	28.12	**4.1**	**58 - 96**	1	-	−6.19	0	1.6E−07	*C. Magneobact.*	*Nitrospira*
Brachypodium distachyon		mle	B_lectin	pfam01453	113.96	5.09E−30	76 - 193							
		sit	PTKc_Src_like	cd05034	115.07	1.06E−28	503 - 696							
		iit	PTKc_VEGFR3	cd05102	58.07	7.91E−09	504 - 694							

[a] **Sequences and their cds files origin.** Search strategies are identified here by names of the selected sets (for procedures see Figure 2). Selected sequences are arranged ascendantly according to minimum or unique Expects concerning their Ig-cds. For additional comments to the selected molecules see Results and Discussion. **Abbreviations in names of proteins:** i.—isoform number; p.—protein; RK—receptor kinase; rest obs—restored obsolete file; STRK—serine/threonine receptor-like protein kinase; u.p.—uncharacterized protein. [b] **The displayed conserved domain similarities (cds) determined by CDD searches.** Framed pairs or triplets of intervals—co-locating recessive Ig-cds and dominant non-Ig cds (cf. Section 3.7); big, BID—cds of Ig-cd found in bacteria; _CAML—_CAM-like; E-n in Expect values—the order related values 10^{-n}; 2ig, 3ig—sequence segments with Ig domain-related cds achieving local chain maxima of the second or third site-related order, respectively; dii—non-Ig, non-lectin and non-max domain including less similar (recessive) Ig-cds; Ig2_LILR_KL—Ig2_LILR_KIR_like; IG_FLMN—filamin-type Ig-cd; iit, sit—records of associated domains with robust mostly inferior or superior cdigvtk similarities, *i.e.* cd05102 and cd05034, respectively; max, mig, mle—maximum cds, Ig-cds or cds of lectin domain found in cds-file of the referred sequence, respectively; SF—super family/superfamily. For more detailed information about conserved domains see special option in menu on NCBI web page. For additional abbreviations see WP4 or WP5. [c] **Parameters of existing co-locating "non-refutable" Ig-related cds.** CA—numbers of all co-locating "non-refutable" or solely found Ig-cds; CA positions—sequence positions of the minimum segments where all co-locating "non-refutable" Ig-cds occur. [d] **FFAS evaluation.** Symmetrical sequence extension or fusion of substantially overlapping segments located in very close positions were specifically used in the cases of too short Ig-cds-derived sequence segments (length < 25 amino acids) for FFAS evaluation. **Except for** the rule Q3, other five selection rules (see below) included **common requirement** for **Ig-related fold FFAS03 scores S_ig**, *i.e.* occurrence of S_ig ≥ 5 in at least two separate database outputs. FM—FFAS derived maximum S_ig; FR—FFAS-associated selection rules, *i.e.* rules **R** necessary for sequence occurrences in the first section of the table and twilight-zone-related quasi-rules **Q**; Q1—7.5 > S_Ig ≥ 5.0 and S_ig achieves maximum value; Q2—10.0 > S_ig ≥ 5.0, S_ig achieves the second or the third superior values and prevailing number of S_ig is observed among all scores ≥ 5; Q3—like in R1, R2 or R3 but S_ig ≥ 5 occur only in one database record; R1—S_ig ≥ 10; R2—10 > S_ig ≥ 7.5 and S_ig achieves maximum value; R3—7.5 > S_ig ≥ 5, S_ig achieves maximum and prevailing number of S_ig is observed among all scores ≥ 5. [e] **Maximal HMMER-derived similarities between the displayed segments restricted by Ig-cds and non-plant sequences**. HMMER evaluation achieved lower Expect values than the preceding BLAST and CDD. This let us to observe the distribution of minimum Expect values (E) enumerated by HMMER (characterizing similarities of each "non-refutable" Ig-cds-related segment selected by CDD) by means of fifteen differently restricted histograms and to estimate thus intervals important for our evaluation (see WP3.9 and WP3.10). This effort yielded two intervals indicating grading positive evaluation, *i.e.* i) $10^{-21} > E \geq 10^{-26}$ and ii) $E < 10^{-26}$, gradually restricting **suspicious** and **hot candidate segments** for recent horizontal transfer, respectively. HME—HMMER-derived Expects; HMG—generum origin of the sequence achieving maximum HMMER-derived similarity; HMT—broader taxonomic group including HMG, bold numbers—**hot candidate segments** for recent horizontal transfer. **Abbreviated expression in the column HMG:** *C. Magneobact.—Candidatus Magnetobacterium.* [f] **FFAS-mediated selection.** Three selection rules (R1, R2, R3) described in footnote d were alternatively required to display the corresponding data in this table section. [g] **Ig-cds added by authors.** These items forming also NRI sets appear to be interesting with respect to their literature or structural contexts (cf. Sections 4.1 and 4.3).

In addition to FFAS and CDD studies, the searches for maximum sequence similarities of non-plant sequences to the Ig-cds-derived plant segments dealt with here were performed using HMMER. Six non-plant sequences, including the two ones of fungal origin, were classified here as hot candidates for recent horizontal transfer exchanges (Table 3).

4. Discussion

4.1. Selected Ig-Domain-Related Sequence Segments

In fact, at least three main groups of Ig-cds can be considered with respect to out data as ensues from the three following paragraphs.

The presence of **two dominant cds recording significant bacterial Ig domains** was indicated in Results and Table 3 in case of protein GP210 (XP_010248630.1; [38] [39]). Bacterial Ig domains perhaps evolved from eukaryotic Ig domains [40]. In accordance with Table 3, bacterial Ig-cds selected here were not classified as positive with respect to recent horizontal transfer though such earlier transfer probably occurred via bacteria as vectors. These bacterial vectors are usually delivered to plants by means of insects and *Nematoda* worms [41] [42].

According to our data, the sequence XP_010937019.2 encoding protein found in oil palm (*Elaeis guineensis*; [43] [44]) formed Ig-cds with domain Ig1_IL1R_like. This Ig-cds represented maximum **dominant Ig-cds with metazoan domains** in Table 3. The sequence could be interesting from the point of view of **vanished structures** resulted via long-lasting or intensive mutation changes. Hence all cds of the sequence achieved only very low scores. Since Ig domains or Ig folds are usually resistant to mutations, the indication of superior cds with Ig-domain within the sequence appeared to be consistent with both Ig-domain relationship and vanished structure classification. Another interesting context of the discussed superior Ig-cds consists in its functional relationship to interleukin-1 receptor (**IL1R**). Hence two additional and even robust IL1R-related cds were also observed in the selected cds-files, *i.e.* cds with Toll/IL1R (TIR) and cd14066 (IL1R associated kinase) domains. Do these IL1R-related cds represent **traces of very old ancestor** IL1R-mediated regulation or only different phylogenic integrations of effective structures?

The occurrence of **recessive Ig-cds** (*i.e.* Ig-cds co-locating within more valid non-Ig domains) may have four reasons: common ancestor origin, transfer of transposons and transposon-like elements, other recombination events (cf. [25]), or a long-lasting convergent competition involving recognition of the same ligand/ligands or related set of ligands. For instance, cds of the multi-interactive **filamin** domain includes here significant recessive Ig-cds recoding Ig-FLMN co-locating with "non-refutable" bacterial Ig-cds with BID_1 restricting the segments of Ig-fold-related score of maximal value in Table 3 (see XP_019237668.1 in Table 3; for associated sequencing project see [45] [46] [47] [48]). As known, filamin domain is composed of Ig-like beta-sandwich fold [49]. Consequently, the presented facts could suggest a common molecular ancestor origin of filamin and Ig domains. Filamin domains are among others present in protozoans [49]. In plants, they even contribute to evolution of cell surface proteins in lineages from a common ancestor to glaucophytes, rhodophytes and viridiplants [50]. Since mannose-specific lectin domains prevail among the selected **lectin** domains, the possible structural relationship between Ig-like beta-sandwich fold and beta-Prism I fold found in plant mannose- and galactose-specific lectins [51] could explain the occurrence of Ig-like structures within lectin domains. The consistent presence of Ig domains in **channels** (cf. **Table 3**) has as yet been described only in mammals [52] [53]. Common molecular origin of Ig superfamily and **self-incompatibility protein** (**SIP**) recorded in **Table 3** represents an important and interesting clue for authors working with these proteins because structural models were developed [54] [55]. Similarly to compared TCR, SIP is involved in cell death events [56]. The inclusions of Ig-cds within several **catalytic** or **transporter** domains described also in **Table 3** appear to be interesting from the point of view of conformational flexibility of Ig domains (cf. [57]), though the relationship to Ig domains remains unclear. The segment restricted by cds with PLN00113 in **LRR STRK** on the other hand surprisingly forms cds with cdigvtk domains and Ig-cds located in its different subsegments (**Table 3**). For comments to important significant Ig-cds of **ALE2** see next section.

In contrast to high occurrence of cds-files recording cds with cdigvtk and STRK-related catalytic domains in RoK set, only low file fraction of such cds forms the NRI sets including top Ig-cds (**Table 3**). Nevertheless relatively increased density but low number of cds with cdigvtk was observed in cds-files of ALPS-related fraction of NRI (**Table 1**; cf. also Sections 4.2 and 5). This indicates considerable independence between occurrences of top Ig-cds and cds of the discussed kinase domains in sequences of higher plant protein origin. For comments to incomplete Ig-like domain folds composing also "non-refutable" Ig-cds-derived segments desribed here see the paper of Berisio [58] (cf. also WP3.7).

4.2. Sequence Similarities of Antibody-Like Phosphorylation Sites

As is well known, the major part of phosphorylation sites with specificity corresponding to ALPS exists in plants [59] [60] [61] [62]. Significantly increased number and relative occurrence value concerned A̲LPS-s̲elected c̲ds-files (ASC; cf. Results) forming in RLK-related RoK set. Another, functionally related resis set (including mostly NLR proteins) then contained 71.4% of ASC whereas NRI set only 38.8%. However, it is a question whether indeed functional plant phosphorylation sites form sequence similarities with mouse and human ALPS. Though eukaryotic phosphosites are generally more conserved than their non-phosphorylated counterparts with similar structural constrains [63], low sequence differences can be even observed between mouse and humans orthologues [23]. A possible alternative mechanism maintaining at least some ALPS-related structures consists in independent participation of some structural segments of ALPS in recognition (**antibody-like property of ALPS**). This would be in accordance with: i) assumed related effects of similarly long LRR [9] [64] [65] and ii) prevalence of ALPS-derived sequence IDs to molecules involved in plant immunity, *i.e.* mainly in pathogen recognition. The latter possibility would follow from:: i) significantly increased occurrence of ALPS-related cds-files in RoK set (containing PTI-related STRK involved in pathogen recognition) accompanied moreover by significant association of these files with the term antiviral (Results and **Table 2**) and ii) considerable occurrence of ALPS-derived sequences in the resis set containing otherwise distinct ETI-related sequences of NLR-proteins involved also in plant resistance.

Robust cds between ALPS-selected plant sequence of ALE2 (described in [66]) and domain of Atrophin-1 superfamily (cl26464) includes significant recessive Ig-cds with domain cd05722 (*i.e.* Ig1_Neogenin; **Table 3**) supported by a maximum number of co-locating "non-refutable" and weak Ig-cds among the cds-files of NRI2 (cf. Section 3.7). In accordance with BLASTP comparison, the corresponding Ig-cds-related segment of ALE2 contains moreover four **traces** and overlaps one trace of ALPS-related segments (cf. WP3.8). Two segments even look like as two repeats occurring at amino acid positions 305 - 328 and 329 - 359 of ALE2 (cf. WP3.8). Possible role of ALPS in ageing was previously discussed [23]. Similarly members of Atrophin-1 superfamily (domain co-located here) participate in a progressive neurodegenerative disorder in vertebrates [67]. This possible agreement raises the question whether the described coincidence of the three types of sequence similarities mentioned here can be interesting for the phylogeny of ageing.

4.3. RoK Set as a Set of Phylogenic Interest

The ALPS-associated RoK set is composed of molecules selected by combined procedures and reselected according to the content of the terms receptor and kinase in molecular title (cf. **Figure 2**). In accordance with this selection strategy and Results, this set contained mainly serine/threonine receptor/receptor-like kinases (**STRK**). Many plant STRK exhibit additional tyrosine specificity forming a group of **STY** (*i.e.* Ser/Thr/Tyr) **kinases** and also the corresponding sequence chimerism [68]. We can even think about deep evolution of STY kinases leading to the last universal common ancestor (**LUCA**) of *Archaea*, *Bacteria* and *Eukarya* [69] [70]. In accordance with this opinion, STY kinases achieve structural relationship to animal non-receptor tyrosine kinases, Src, Abl, Lyn, Fes, Sek, Kin and Ras as well as receptor-like kinase Lyk3 suggesting thus common superfamily origin of the compared kinases [71] [72]. In comparison with these data, our subsets of cds-files specifically compared in **Figure 4** did not contain the catalytic domains of Ras and Lyk3 kinases.

Occurrence of global score maxima of PLN00113 (**Figure 4**), superiority of its subset in RoK set (Formula (4)), its lower cds densities and lower frequencies of individual superior cds than in case of cd14066 (**Figure 4**) appear to be in agreement with a broader and more distant, *i.e.* ancient ancestor-like structural context of PLN00113 (LRR-associated catalytic kinase domain) common with respect to the evaluated kinase domains. On the other hand cd14066 (interleukin-1 receptor associated catalytic serine/threonine kinase domain STRK; *i.e.* STKc-IRAK), achieves the densest cds and most frequently superior kinase-related scores in individual cds-files. These scores represent sometimes even score maxima among all cds in evaluated cds-file and determine extremely low relative standard deviations (**Figure 4**). This extreme data can indicate closer but later structural relationship of cd14066 to ancestor structure than in case of domain PLN00113. The described cds-derived domain chimerism appears to be interesting in two aspects, *i.e.* i) possible structural divergence between protozoan or early metazoan RTK (including also IgV-associated cdigvtk evaluated here) and plant STY kinases [18] [73] [74] [75] and ii) inclusion of cds recording NRI-restricted domain Ig3_L1-CAM_like (28.94 bits) and cdigvtk (co-locating with cd14066) in different sites of longer and extremely robust cds (497.45 bits) of domain PLN00113 (**Table 3**).

4.4. Structures in the Light of Statistics

Removal of sequence redundancy diminished the number of sequence IDs from 2217 to 1323. In fact, this processing reflected the presence of sequences encoding duplicated isoforms and alternative splicing products. In agreement with the monitored associations between variously selected sets and subsets, many significant domain linkages and disjunctions also exist (cf. **Table 2** and the corresponding section of Results) as possible consequences of diversified selective processes of domain shuffling (cf. [15]). In accordance with statistics in **Table 2** and the relationships described in Results, we can among others pose question, whether the third superfamily of LRR domains represented by cl27891 is specifically involved in recognition by ETI-related NLR composed also of TIR domains (cf. [12]). Several significantly increased occurrences of terms in the sets were partially explained here. This comprised the occurrences of: i) term immunoglobulin, ii) terms linked in strategy I3R (related to plant immunity) in RoK set and iii) the terms antifungal2 in the set F2oP1 (see Section 3.2). Statistically predicted hot-candidate partners for recent horizontal transfer among genes of fungal origin displayed in **Table 3** represent meanwhile a question for further analysis and critical comments exceeding extent of this paper.

5. Conclusions

The described procedures selected many non-Ig molecules important for plant immunity. To explain this relationship for cases of Ig-similar proteins and proteins with weak Ig-cds, we considered mainly similarities following from motif-motif, motif-secondary structure or ligand-secondary structure interactions often generated via convergent mutation changes, repeat effects and recombination changes on DNA level [15] [22] [76] [77] [78] [79].

Due to phylogenic distance of higher plants from animals including most of

Ig-domains, we enlarged the limiting Expect values to those approximating i) significant refusing of domain similarity (Section 2.10) and among others also ii) dense similarities of short fold lengths (cf. WP2.6). Since uncertainty of such weakened restriction increased, complementary evaluation of cds specificity (uniquely proved here in case of quasi-significant Ig-cds of XP_010937019.2; cf. Sections 3.7 and WP2.5) or usage of alternative online accessible methods (cf. FFAS-fold-derived analysis in Table 3) became to be important. The function of the selected Ig-cds-related segments displayed in Table 3 remains unclear. We can only remind the discussed possible relationships to interleukin-1-receptor-related structures and structures involved in ageing or self-recognition interesting for further structural investigation.

Several extremely statistically strong and significant domain linkages or disjunctions in multi-domain molecules follow from our data. Nevertheless, it remains to determine whether all these linkages are independent of our Ig-related selection machinery. This could be decided in future revision of the described associations on larger sets keeping only domain relationships. Similarly, more detailed phylogenic evaluation of relationships between cds including cd14066, PLN00113 and cdigvtk segments could be important for the studies concerning deep evolution of SRTK, RTK and perhaps also Ig-domains (see Figure 4 and below).

In accordance with data presented in Table 2 and Table 3, an alternative role of ALPS-related segments in recognition was considered in Section 4.2. This raises the question, whether phylogenic (inter-genic or intragenic) interactions between at least certain genes encoding i) Ig domain ancestors or Ig domains and ii) STRK could include transfer of ALPS-related structures (or the corresponding repeats; cf. Section 4.2) from STRK improving sometimes recognition mediated pre-Ig or Ig domains.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Tsuda, K. and Somssich, *I.E.* (2015) Transcriptional Networks in Plant Immunity. *New Phytologist*, **206**, 932-947. https://doi.org/10.1111/nph.13286

[2] Calil, I.P. and Fontes, E.P.B. (2017) Plant Immunity against Viruses: Antiviral Immune Receptors in Focus. *Annals of Botany*, **119**, 711-723. https://doi.org/10.1093/aob/mcw200

[3] Brustolini, O.J., Silva, J.C., Sakamoto, T. and Fontes, E.P. (2017) Bioinformatics Analysis of the Receptor-Like Kinase (RLK) Superfamily. *Methods in Molecular Biology*, **1578**, 123-132. https://doi.org/10.1007/978-1-4939-6859-6_9

[4] Dezhsetan, S. (2017) Genome Scanning for Identification and Mapping of Receptor-Like Kinase (RLK) Gene Superfamily in *Solanum tuberosum*. *Physiology and Molecular Biology of Plants: An International Journal of Functional Plant Biology*, **23**, 755-765. https://doi.org/10.1007/s12298-017-0471-6

[5] Shchennikova, A.V., Kochieva, E.Z., Beletsky, A.V., Filyushin, M.A., Shulga, O.A., Ravin, N.V. and Skryabin, K.G. (2017) Identification and Expression Analysis of Receptor-Like Kinase Gene ERECTA in Mycoheterotrophic Plant *Monotropa hypopitys*. *Molecular Biology* (*Moscow*), **51**, 780-786. https://doi.org/10.1134/S002689331705017X

[6] Bacete, L., Mélida, H., Miedes, E. and Molina, A. (2018) Plant Cell Wall-Mediated Immunity: Cell Wall Changes Trigger Disease Resistance Responses. *The Plant Journal*, **93**, 614-636. https://doi.org/10.1111/tpj.13807

[7] Hervé, C., Dabos, P., Galaud, J.P., Rougé, P. and Lescure, B. (1996) Characterization of an *Arabidopsis thaliana* Gene That Defines a New Class of Putative Plant Receptor Kinases with an Extracellular Lectin-Like Domain. *Journal of Molecular Biology*, **258**, 778-788. https://doi.org/10.1006/jmbi.1996.0286

[8] Afzal, A.J., Wood, A.J. and Lightfoot, D.A. (2008) Plant Receptor-Like Serine Threonine Kinases: Roles in Signaling and Plant Defense. *Molecular Plant-Microbe Interactions*, **21**, 507-517. https://doi.org/10.1094/MPMI-21-5-0507

[9] Liu, P.L., Du, L., Huang, Y., Gao, S.M. and Yu, M. (2017) Origin and Diversification of Leucine-Rich Repeat Receptor-Like Protein Kinase (LRR-RLK) Genes in Plants. *BMC Evolutionary Biology*, **17**, 47. https://doi.org/10.1186/s12862-017-0891-5

[10] Ma, N., Liu, C., Li, H., Wang, J., Zhang, B., Lin, J. and Chang, Y. (2018) Genome-Wide Identification of Lectin Receptor Kinases in Pear: Functional Characterization of the L-Type LecRLK Gene PbLRK138. *Gene*, **661**, 11-21. https://doi.org/10.1016/j.gene.2018.03.077

[11] Di Gaspero, G. and Cipriani, G. (2003) Nucleotide Binding Site/Leucine-Rich Repeats, Pto-Like and Receptor-Like Kinases Related to Disease Resistance in Grapevine. *Molecular Genetics and Genomics*, **269**, 612-623. https://doi.org/10.1007/s00438-003-0884-5

[12] Bonardi, V., Cherkis, K., Nishimura, M.T. and Dangl, J.L. (2012) A New Eye on NLR Proteins: Focused on Clarity or Diffused by Complexity? *Current Opinion in Immunology*, **24**, 41-50. https://doi.org/10.1016/j.coi.2011.12.006

[13] Burdett, H., Bentham, A.R., Williams, S.J., Dodds, P.N., Anderson, P.A., Banfield, M.J. and Kobe, B. (2019) The Plant “Resistosome”: Structural Insights into Immune Signaling. *Cell Host and Microbe*, **26**, 193-201. https://doi.org/10.1016/j.chom.2019.07.020

[14] Johnson, G.B., Brunn, G.J., Tang, A.H. and Platt, J.L. (2003) Evolutionary Clues to the Functions of the Toll-Like Family as Surveillance Receptors. *Trends in Immunology*, **24,** 19-24. https://doi.org/10.1016/S1471-4906(02)00014-5

[15] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2008) Molecular Biology of the Cell. 5th Edition, Garland Science, New York. https://doi.org/10.1201/9780203833445

[16] Schäcke, H., Rinkevich, B., Gamulin, V., Müller, I.M. and Müller, W.E. (1994) Immunoglobulin-Like Domain Is Present in the Extracellular Part of the Receptor Tyrosine Kinase from the Marine Sponge *Geodia cydonium*. *Journal of Molecular Recognition*, **7**, 273-276. https://doi.org/10.1002/jmr.300070406

[17] Blumbach, B., Diehl-Seifert, B., Seack, J., Steffen, R., Müller, I.M. and Müller, W.E. (1999) Cloning and Expression of New Receptors Belonging to the Immunoglobulin Superfamily from the Marine Sponge *Geodia cydonium*. *Immunogenetics*, **49**, 751-763. https://doi.org/10.1007/s002510050549

[18] Müller, W.E.G. (2001) Review: How Was Metazoan Threshold Crossed? The Hy-

pothetical *Urmetazoa*. *Comparative Biochemistry and Physiology Part A*, **129**, 433-460. https://doi.org/10.1016/S1095-6433(00)00360-3

[19] Grassot, J., Gouy, M., Perrière, G. and Mouchiroud, G. (2006) Origin and Molecular Evolution of Receptor Tyrosine Kinases with Immunoglobulin-Like Domains. *Molecular Biology and Evolution*, **23,** 1232-124. https://doi.org/10.1093/molbev/msk007

[20] Kubrycht, J., Borecký, J. and Sigler, K. (2002) Sequence Similarities of Protein Kinase Peptide Substrates and Inhibitors: Comparison of Their Primary Structures with Immunoglobulin Repeats. *Folia Microbiologica*, **47**, 319-358. https://doi.org/10.1007/BF02818689

[21] Kubrycht, J., Sigler, K., Růžička, M., Souček, P., Borecký, J. and Ježek, P. (2006) Ancient Phylogenetic Beginnings of Immunoglobulin Hypermutation. *Journal of Molecular Evolution*, **63**, 691-706. https://doi.org/10.1007/s00239-006-0051-9

[22] Kubrycht, J., Sigler, K., Souček, P. and Hudeček, J. (2013) Structures Composing Protein Domains. *Biochimie*, **95**, 1511-1524. https://doi.org/10.1016/j.biochi.2013.04.001

[23] Kubrycht, J., Sigler, K., Souček, P. and Hudeček, J. (2016) Antibody-Like Phosphorylation Sites in Focus of Statistically Based Bilingual Approach. *Computational Molecular Bioscience*, **6**, 1-22. https://doi.org/10.4236/cmb.2016.61001

[24] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, **25**, 3389-3402. https://doi.org/10.1093/nar/25.17.3389

[25] Zhang, Z., Schäffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V. and Altschul, S.F. (1998) Protein Sequence Similarity Searches Using Patterns as Seeds. *Nucleic Acids Research*, **26**, 3986-3990. https://doi.org/10.1093/nar/26.17.3986

[26] Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. *Nucleic Acids Research*, **30**, 281-283. https://doi.org/10.1093/nar/30.1.281

[27] Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., *et al.* (2003) CDD: A Curated Entrez Database of Conserved domain Alignments. *Nucleic Acids Research*, **31**, 383-387. https://doi.org/10.1093/nar/gkg087

[28] Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., *et al.* (2011) CDD: A Conserved Domain Database for the Functional Annotation of Proteins. *Nucleic Acids Research*, **39**, D225-D229. https://doi.org/10.1093/nar/gkq1189

[29] Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., *et al.* (2015) CDD: NCBI's Conserved Domain Database. *Nucleic Acids Research*, **43**, D222-D226. https://doi.org/10.1093/nar/gku1221

[30] Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., *et al.* (2013) CDD: Conserved Domains and Protein Three-Dimensional Structure. *Nucleic Acids Research*, **41**, D348-D352. https://doi.org/10.1093/nar/gks1243

[31] Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. and Godzik, A. (2005) FFAS03: A Server for Profile-Profile Sequence Alignments. *Nucleic Acids Research*, **33**, W284-W288. https://doi.org/10.1093/nar/gki418

[32] Jaroszewski, L., Li, Z., Cai, X.H., Weber, C. and Godzik, A. (2011) FFAS Server: Novel Features and Applications. *Nucleic Acids Research*, **39**, W38-W44. https://doi.org/10.1093/nar/gkr441

[33] Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R. and Finn, R.D. (2018) HMMER Web Server: 2018 Update. *Nucleic Acids Research*, **46**, W200-W204. https://doi.org/10.1093/nar/gky448

[34] Lowry, R. (2001) 2x2 Contingency Table. http://vassarstats.net/tab2x2.html

[35] Zvárová, J. (2001) Biomedical Statistics. I. The Fundamentals of Statistics for Biomedical Fields. Karolinum, Prague.

[36] Mefford, J. (2012) Bayesian Statistics and Bias Analysis. http://slideplayer.com/slide/8525698

[37] Dembo, A., Karlin, S. and Zeitouni, O. (1994) Limit Distribution of Maximal Non-Aligned Two-Sequence Segmental Score. *The Annals of Probability*, **22**, 2022-2039. https://doi.org/10.1214/aop/1176988493

[38] Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.T., Zhang, Q., Kim, M.J., Schatz, M.C., Campbell, M., *et al.* (2013) Genome of the Long-Living Sacred Lotus (*Nelumbo nucifera Gaertn.*). *Genome Biology*, **14**, R41. http://genomebiology.com/2013/14/5/R41 https://doi.org/10.1186/gb-2013-14-5-r41

[39] Vimolmangkang, S., Deng, X., Owiti, A., Meelaph, T., Ogutu, C. and Han, Y. (2016) Evolutionary Origin of the NCSI Gene Subfamily Encoding Norcoclaurine Synthase Is Associated with the Biosynthesis of Benzylisoquinoline Alkaloids in Plants. *Scientific Reports*, **6**, Article No. 26323. https://doi.org/10.1038/srep26323

[40] Bateman, A., Eddy, S.R. and Chothia, C. (1996) Members of the Immunoglobulin Superfamily in Bacteria. *Protein Science*, **5**, 1939-1941. https://doi.org/10.1002/pro.5560050923

[41] Frago, E., Dicke, M. and Godfray, H.C. (2012) Insect Symbionts as Hidden Players in Insect-Plant Interactions. *Trends in Ecology and Evolution*, **27**, 705-711. https://doi.org/10.1016/j.tree.2012.08.013

[42] Zhu, S. and Gao, B. (2014) Nematode-Derived Drosomycin-Type Antifungal Peptides Provide Evidence for Plant-to-Ecdysozoan Horizontal Transfer of a Disease Resistance Gene. *Nature Communications*, **5**, 3154. https://doi.org/10.1038/ncomms4154

[43] Low, E.T., Alias, H., Boon, S.H., Shariff, E.M., Tan, C.Y., Ooi, L.C., Cheah, S.C., Raha, A.R., Wan, K.L. and Singh, R. (2008) Oil Palm (*Elaeis guineensis Jacq.*) Tissue Culture ESTs: Identifying Genes Associated with Callogenesis and Embryogenesis. *BMC Plant Biology*, **8**, Article No. 62. https://doi.org/10.1186/1471-2229-8-62 http://www.biomedcentral.com/1471-2229/8/62

[44] Uthaipaisanwong, P., Chanprasert, J., Shearman, J.R., Sangsrakru, D., Yoocha, T., Jomchai, N., Jantasuriyarat, C., Tragoonrung, S. and Tangphatsornruang, S. (2012) Characterization of the Chloroplast Genome Sequence of Oil Palm (*Elaeis guineensis Jacq.*). *Gene*, **500**, 172-180. https://doi.org/10.1016/j.gene.2012.03.061

[45] Jang, S., Hong, M.Y., Chung, Y.Y. and An, G. (1999) Ectopic Expression of Tobacco MADS Genes Modulates Flowering Time and Plant Architecture. *Molecules and Cells*, **9**, 576-586.

[46] Bohlmann, J., Stauber, E.J., Krock, B., Oldham, N.J., Gershenzon, J. and Baldwin, I.T. (2002) Gene Expression of 5-Epi-Aristolochene Synthase and Formation of Capsidiol in Roots of *Nicotiana attenuata* and *N. sylvestris*. *Phytochemistry*, **60**, 109-116. https://doi.org/10.1016/S0031-9422(02)00080-8

[47] Yukawa, M., Tsudzuki, T. and Sugiura, M. (2006) The Chloroplast Genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: Complete Sequencing Confirms that the *Nicotiana sylvestris* Progenitor is the Maternal Genome Donor of *Nicotiana tabacum*. *Molecular Genetics and Genomics*, **275**, 367-373. https://doi.org/10.1007/s00438-005-0092-6

[48] Sierro, N., Battey, J.N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C. and Ivanov, N.V. (2013) Reference Genomes and Transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biology*, **14**, R60. http://genomebiology.com/2013/14/6/R60 https://doi.org/10.1186/gb-2013-14-6-r60

[49] Light, S., Sagit, R., Ithychanda, S.S., Qin, J. and Elofsson, A. (2012) The Evolution of Filamin-a Protein Domain Repeat Perspective. *Journal of Structural Biology*, **179**, 289-298. https://doi.org/10.1016/j.jsb.2012.02.010

[50] Becker, B., Doan, J.M., Wustman, B., Carpenter, E.J., Chen, L., Zhang, Y., Wong, G.K. and Melkonian, M. (2015) The Origin and Evolution of the Plant Cell Surface: Algal Integrin-Associated Proteins and a New Family of Integrin-Like Cytoskeleton-ECM Linker Proteins. *Genome Biology and Evolution*, **7**, 1580-1589. https://doi.org/10.1093/gbe/evv089

[51] Chandran, T., Sharma, A. and Vijayan, M. (2013) Generation of Ligand Specificity and Modes of Oligomerization in *β*-Prism I Fold Lectins. *Advances in Protein Chemistry and Structural Biology*. **92**, 135-178. https://doi.org/10.1016/B978-0-12-411636-8.00004-3

[52] Fallen, K., Banerjee, S., Sheehan, J., Addison, D., Lewis, L.M., Meiler, J. and Denton, J.S. (2009) The Kir Channel Immunoglobulin Domain Is Essential for Kir1.1 (ROMK) Thermodynamic Stability, Trafficking and Gating. *Channels* (*Austin*), **3**, 57-68. https://doi.org/10.4161/chan.3.1.7817

[53] Yereddi, N.R., Cusdin, F.S., Namadurai, S., Packman, L.C., Monie, T.P., Slavny, P., Clare, J.J., Powell, A.J. and Jackson, A.P. (2013) The Immunoglobulin Domain of the Sodium Channel β3 Subunit Contains a Surface-Localized Disulfide Bond that Is Required for Homophilic Binding. *FASEB Journal*, **27**, 568-580. https://doi.org/10.1096/fj.12-209445

[54] Ashkani, J. and Rees, D.J. (2016) A Comprehensive Study of Molecular Evolution at the Self-Incompatibility Locus of *Rosaceae*. *Journal of Molecular Evolution*, **82**, 128-145. https://doi.org/10.1007/s00239-015-9726-4

[55] Nasrallah, J.B. (2017) Plant Mating Systems: Self-Incompatibility and Evolutionary Transitions to Self-Fertility in the Mustard Family. *Current Opinion in Genetics and Development*, **47**, 54-60. https://doi.org/10.1016/j.gde.2017.08.005

[56] Wilkins, K.A., Poulter, N.S. and Franklin-Tong, V.E. (2014) Taking One for the Team: Self-Recognition and Cell Suicide in Pollen. *Journal of Experimental Botany*, **65**, 1331-1342. https://doi.org/10.1093/jxb/ert468

[57] James, L.C., Roversi, P. and Tawfik, D.S. (2003) Antibody Multispecificity Mediated by Conformational Diversity. *Science*, **299**, 1362-1367. https://doi.org/10.1126/science.1079731

[58] Berisio, R., Ciccarelli, L., Squeglia, F., De Simone, A. and Vitagliano, L. (2012) Structural and Dynamic Properties of Incomplete Immunoglobulin-Like Fold Domains. *Protein and Peptide Letters*, **19**, 1045-1053. https://doi.org/10.2174/092986612802762732

[59] Lee, T.Y., Bretaña, N.A. and Lu, C.T. (2011) PlantPhos: Using Maximal Dependence Decomposition to Identify Plant Phosphorylation Sites with Substrate Site Specific-

ity. *BMC Bioinformatics*, **12**, Article No. 261.
http://www.biomedcentral.com/1471-2105/12/261
https://doi.org/10.1186/1471-2105-12-261

[60] Yoshiyama, K.O., Kobayashi, J., Ogita, N., Ueda, M., Kimura, S., Maki, H. and Umeda, M. (2013) ATM-Mediated Phosphorylation of SOG1 Is Essential for the DNA Damage Response in *Arabidopsis*. *EMBO Reports*, **14**, 817-822.
https://doi.org/10.1038/embor.2013.112

[61] Zulawski, M., Braginets, R. and Schulze, W.X. (2013) PhosPhAt Goes Kinases Searchable Protein Kinase Target Information in the Plant Phosphorylation Site Database PhosPhAt. *Nucleic Acids Research*, **41**, D1176-D1184.
https://doi.org/10.1093/nar/gks1081

[62] Takagi, M., Sakamoto, T., Suzuki, R., Nemoto, K., Obayashi, T., Hirakawa, T., Matsunaga, T.M., Kurihara, D., Nariai, Y., Urano, T., Sawasaki, T. and Matsunaga, S. (2016) Plant Aurora Kinases Interact with and Phosphorylate Transcription Factors. *Journal of Plant Research*, **129**, 1165-1178.
https://doi.org/10.1007/s10265-016-0860-x

[63] Gnad, F., Forner, F., Zielinska, D.F., Birney, E., Gunawardena, J. and Mann, M. (2010) Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes, and Mitochondria. *Molecular and Cellular Proteomics*, **9**, 2642-2653.
https://doi.org/10.1074/mcp.M110.001594

[64] Walker, J.C. (1993) Receptor-Like Protein Kinase Genes of *Arabidopsis thaliana*. *The Plant Journal*, **3**, 451-456. https://doi.org/10.1111/j.1365-313X.1993.tb00164.x

[65] Hirano, M., Guo, P., McCurley, N., Schorpp, M., Das, S., Boehm, T. and Cooper, M.D. (2013) Evolutionary Implications of a Third Lymphocyte Lineage in Lampreys. *Nature*, **501**, 435-438. https://doi.org/10.1038/nature12467

[66] Li, D., Deng, Z., Qin, B., Liu, X. and Men, Z. (2012) De Novo Assembly and Characterization of Bark Transcriptome Using Illumina Sequencing and Development of EST-SSR Markers in Rubber Tree (*Hevea brasiliensis Muell. Arg.*). *BMC Genomics*, **13**, Article No. 192. http://www.biomedcentral.com/1471-2164/13/192
https://doi.org/10.1186/1471-2164-13-192

[67] Vázquez, N., Rocha, S., López-Fernández, H., Torres, A., Camacho, R., Fdez-Riverola, F., Vieira, J., Vieira, C.P. and Reboiro-Jato, M. (2019) EvoPPI 1.0: A Web Platform for Within- and Between-Species Multiple Interactome Comparisons and Application to Nine PolyQ Proteins Determining Neurodegenerative Diseases. *Interdisciplinary Sciences, Computational Life Sciences*, **11**, 45-56.
https://doi.org/10.1007/s12539-019-00317-y

[68] Hirayama, T. and Oka, A. (1992) Novel Protein Kinase of *Arabidopsis thaliana* (APK1) That Phosphorylates Tyrosine, Serine and Threonine. *Plant Molecular Biology*, **20**, 653-662. https://doi.org/10.1007/BF00046450

[69] Hanks, S.K., Quinn, A.M. and Hunter, T. (1988) The Protein Kinase Family: Conserved Features and Deduced Phylogeny of the Catalytic Domains. *Science*, **241**, 42-52. https://doi.org/10.1126/science.3291115

[70] Stancik, I.A., Šestak, M.S., Ji, B., Axelson-Fisk, M., Franjevic, D., Jers, C., Domazet-Lošo, T. and Mijakovic, I. (2018) Serine/Threonine Protein Kinases from *Bacteria*, *archaea* and *Eukarya* Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life. *Journal of Molecular Biology*, **430**, 27-32.
https://doi.org/10.1016/j.jmb.2017.11.004

[71] Rudrabhatla, P., Reddy, M.M. and Rajasekharan, R. (2006) Genome-Wide Analysis and Experimentation of Plant Serine/Threonine/Tyrosine-Specific Protein Kinases. *Plant Molecular Biology*, **60**, 293-319. https://doi.org/10.1007/s11103-005-4109-7

[72] Klaus-Heisen, D., Nurisso, A., Pietraszewska-Bogiel, A., Mbengue, M., Camut, S., Timmers, T., Pichereaux, C., Rossignol, M., Gadella, T.W., Imberty, A., Lefebvre, B. and Cullimore, J.V. (2011) Structure-Function Similarities between a Plant Receptor-Like Kinase and the Human Interleukin-1 Receptor-Associated Kinase-4. *Journal of Biological Chemistry*, **286**, 11202-11210. https://doi.org/10.1074/jbc.M110.186171

[73] King, N. and Carroll, S.B. (2001) A Receptor Tyrosine Kinase from *Choanoflagellates*: Molecular Insights into Early Animal Evolution. *Proceedings of National Academy of Sciences of the United States of America*, **98**, 15032-15037. https://doi.org/10.1073/pnas.261477698

[74] Schultheiss, K.P., Craddock, B.P., Tong, M., Seeliger, M. and Miller, W.T. (2013) Metazoan-Like Signaling in a Unicellular Receptor Tyrosine Kinase. *BMC Biochemistry*, **14**, Article No. 4. http://www.biomedcentral.com/1471-2091/14/4 https://doi.org/10.1186/1471-2091-14-4

[75] Mohanty, S., Oruganty, K., Kwon, A., Byrne, D.P., Ferries, S., Ruan, Z., Hanold, L.E., Katiyar, S., Kennedy, E.J., Eyers, P.A. and Kannan, N. (2016) Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genetics*, **12**, e1005885. https://doi.org/10.1371/journal.pgen.1005885

[76] Hakim, I., Amariglio, N., Grossman, Z., Simoni-Brok, F., Ohno, S. and Rechavi, G. (1994) The Genome of the I Human Transposable Repetitive Elements Is Composed of a Basic Motif Homologous to an Ancestral Immunoglobulin Gene Sequence. *Proceedings of National Academy of Sciences of the United States of America*, **91**, 7967-7969. https://doi.org/10.1073/pnas.91.17.7967

[77] Pang, E. and Lin, K. (2010) Yeast Protein-Protein Interaction Binding Sites: Prediction from the Motif-Motif, Motif-Domain and Domain-Domain Levels. *Molecular Biosystems*, **6**, 2164-2173. https://doi.org/10.1039/c0mb00038h

[78] Wang, F., Liu, M., Song, B., Li, D., Pei, H., Guo, Y., Huang, J. and Zhang, D. (2012) Prediction and Characterization of Protein-Protein Interaction Networks in Swine. *Proteome Science*, **10**, Article No. 2. http://www.proteomesci.com/content/10/1/2 https://doi.org/10.1186/1477-5956-10-2

[79] Kubrycht, J., Sigler, K. and Souček, P. (2012) Virtual Interactomics of Proteins from Biochemical Standpoint. *Molecular Biology International*, **2012**, Article ID: 976385. https://doi.org/10.1155/2012/976385