



# STATdb: A Specialised Resource for the STATome

C. Pawan K. Patro<sup>1</sup>, Asif M. Khan<sup>2,3</sup>, Tin Wee Tan<sup>1\*</sup>, Xin-Yuan Fu<sup>1,4,5\*</sup>

**1** Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **2** Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, **3** Perdana University Graduate School of Medicine, Serdang, Selangor Darul Ehsan, Malaysia, **4** Cancer Science Institute of Singapore (CSI), National University of Singapore, Singapore, Singapore, **5** Department of Microbiology and Immunology, Indiana University School of Medicine, Indianapolis, Indiana, United States of America

## Abstract

Signal transducers and activators of transcription (STAT) proteins are key signalling molecules in metazoans, implicated in various cellular processes. Increased research in the field has resulted in the accumulation of STAT sequence and structure data, which are scattered across various public databases, missing extensive functional annotations, and prone to effort redundancy because of the dearth of community sharing. Therefore, there is a need to integrate the existing sequence, structure and functional data into a central repository, one that is enriched with annotations and provides a platform for community contributions. Herein, we present STATdb (publicly available at <http://statdb.bic.nus.edu.sg/>), the first integrated resource for STAT sequences comprising 1540 records representing the known STATome, enriched with existing structural and functional information from various databases and literature and including manual annotations. STATdb provides advanced features for data visualization, analysis and prediction, and community contributions. A key feature is a meta-predictor to characterise STAT sequences based on a novel classification that integrates STAT domain architecture, lineage and function. A curation policy workflow has been devised for regulated and structured community contributions, with an update policy for the seamless integration of new data and annotations.

**Citation:** Patro CPK, Khan AM, Tan TW, Fu X-Y (2014) STATdb: A Specialised Resource for the STATome. PLoS ONE 9(8): e104597. doi:10.1371/journal.pone.0104597

**Editor:** Michael Nevels, University of Regensburg, Germany

**Received:** November 20, 2013; **Accepted:** July 15, 2014; **Published:** August 26, 2014

**Copyright:** © 2014 Patro et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** XYF was supported by start-up funds from the Office of the Deputy President (Research and Technology) and the Yong Loo Lin School of Medicine of the National University of Singapore (NUS), and grants from Ministry of Education of Singapore (MOE2010-T2-1-084). CPKP was funded on an NUS Research Scholarship for part of this research work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* Email: tinwee@bic.nus.edu.sg (TWT); bchfy@nus.edu.sg (XYF)

## Introduction

Signal transducers and activators of transcription (STAT) proteins are one of the most important signalling molecules in metazoans [1,2,3], playing dual roles as cytoplasmic signalling proteins and nuclear transcription factors in the cell. STATs are key components of the Janus Kinase (JAK)/STAT signalling pathway [4], an evolutionarily conserved cascade that facilitates a wide range of inter- and intra-cellular signalling roles vital for cellular differentiation, growth and survival [5,6,7]. STATs get activated via phosphorylation by kinases, such as JAKs and Src kinases, and growth factor receptors, among other activating proteins, responding to extracellular-signalling proteins [8,9]. STAT proteins, upon activation, translocate to the nucleus to regulate a diverse set of target genes [1], however several deviations to this canonical pathway have been described to date [10]. Numerous studies have shown that dysregulation of the JAK/STAT pathway is associated with chronic inflammation, neurodegenerative diseases and cancer, among other disease states [11].

The STAT protein family in mammals comprises seven members—STAT1-4, STAT5A and 5B, and STAT6—with diverse functions [1,11,12]. Knockout of either STAT1 or STAT2 results in an impaired response to interferons [1]. Furthermore, the absence of STAT1 results in impaired growth control [13] whereas STAT2 knockout mice show numerous defects in their

immune response [14]. Early embryonic lethality has been associated with STAT3 knockout mice [1,13], and additional complications, such as multiple defects in adult tissues and an impaired response to pathogens, are also linked to the absence of STAT3. STAT4 deletion affects T helper 1 (TH1) cell function, opposing STAT6 function, which impairs TH2 differentiation [1,13]. Both STAT5A and STAT5B are important for breast development/lactation: STAT5A is required for prolactin responsiveness, whereas STAT5B is required for growth hormone responsiveness [1,13]. STAT5 refers to the gene that duplicated to give rise to STAT5A and STAT5B in species ancestral to mammals [15]. Both STAT5.1 and STAT5.2 are STAT5 homologs in fishes [15].

STAT family of proteins has thus been studied intensively [1,11], which has led to the accumulation of sequence and structure data scattered across various public databases. For example, the primary NCBI sequence databases (GenBank and GenPept) are comprehensive but lack extensive functional annotations, such as status of experimental validation, STAT domains, interacting proteins, and gene and structural information, which are found in other databases, such as UniProt, RefSeq, PDB, Gene, CDD, and within the literature. Public databases, however, are prone to errors [16], and consequently an extensive analysis is required to ascertain the reliability of data in public domains by cross-checking with other databases and with what is cited in the literature. This difficult task, along with the substantial

lack of sharing amongst the scientific community, has thus led to redundant efforts in the laboratory. Therefore, there is an urgent need to assemble, organize, remove duplicates and integrate existing sequence, structure and functional data into a central repository that is enriched with annotations and provides a platform for community contribution to allow for systematic, integrated analyses of STATs.

Herein, we present STATdb, a specialised repository of STAT sequences, representing the known STATome, integrating existing sequence, structure and functional information from various databases, and the literature, and including manual annotations. This, to our knowledge, is the first reported specialised Web resource for STAT sequences. STATdb, besides the basic functionalities such as database query using keyword search and data download, provides advanced features for data visualization, analysis and prediction, and community contribution. Users can dynamically browse the STATome—the complete dataset of reported STAT sequence records in STATdb—and interactively view available 3D structures. STATdb is integrated with sequence analysis tools, such as the Basic Local Alignment Search Tool (BLAST) for sequence similarity searches and ClustalW for multiple sequence alignments on the fly. A key feature of the database is STATdbPredict, which is used to characterize STAT sequences based on a novel classification scheme that incorporates domain architecture, lineage and function. Sequence records are manually annotated with STATdb classification notation, experimental status validation, and individual domain sequences, among others. A submission/curation policy workflow has been devised for regulated and structured contribution of new records and for enrichment/correction of functional annotations of existing records by the STAT research community (curator) through an easy-to-use interface. Community contribution, based on existing data and literature, is important in biological data-warehousing [17] and the approach has been highly successful, as exemplified by numerous Wiki-based projects: PDBwiki [18], WikiProteins [19], Gene Wiki [20], RNA Wikiproject [21], EcoliWiki [22] and WikiPathways [23]. Additionally, an update policy has been devised for the regular integration of new records and annotations from public databases and/or the community.

## Materials and Methods

### Sequence Data Collection

Protein and nucleotide sequences of STAT were first collected through keyword searches using the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) [24,25] database, followed by sequence similarity searches against all reported sequences in the NCBI non-redundant (NR) database [25]. Keyword hits were manually checked and verified as STAT according to the literature. Selected verified sequences were used as query for Position-Specific Iterated (PSI)-BLAST search [26] in order to perform a comprehensive survey of STAT sequences. Significant blast hits were selected and sequence duplicates were removed using CD-HIT [27]. The remaining non-redundant sequences were used to populate STATdb.

### Database Record Annotation

Existing STAT record annotations in various public databases were studied to identify relevant fields for STATdb. The list of fields defined for STATdb records are provided in Table 1. Fields that provide information selected from the source record (NCBI Entrez Protein Database) are marked as “Source”, such as gene name, protein name, type of STAT, database cross-references, literature, species, location of the gene on the chromosome, length

of the protein sequence, and amino acid sequence for the full-length protein, as well as the list of individual STAT domains. “Assigned” fields are those not found in the source record, but were included to provide information obtained from database cross-references and/or analysis of the sequence data, existing annotations or the literature.

### STATdb Classification

STATdb-enriched annotations enabled the construction of a novel classification scheme for the characterization of STAT sequences and for the prediction of novel family members. This classification is based on a three-tier system: “Domain Architecture – Lineage – Function”.

“Domain Architecture”, or “DA”, is used to describe the observed order/arrangement of STAT domains within the protein. STAT proteins comprise five major domains: protein interaction domain (STAT\_int), all-alpha domain (STAT\_alpha), DNA-binding domain (STAT\_bind), SH2 domain (SH2) and the transactivation domain (TAZ2). The five unique domain architectures are referred to as DA I – DA V, observed to date for STATdb sequences with DA U representing uncommon combinations and artificial sequences.

“Lineage” is defined as a sub-classification of “Domain Architecture”, and is based on the taxonomy of the species from which the STAT sequence was isolated. All STATdb sequences of each domain architecture were analysed for their species lineage by interrogating the NCBI Taxonomy Database and the sequences were then grouped according to the furthest common differentiation level from the root (*i.e.*, cellular organism). As STATs are a family of paralogous loci (*e.g.*, in vertebrates), the classification does not aim to coincide the species and gene trees in instances where it is not possible.

“Function” is defined as a sub-classification of “Lineage”, and is based on the role of the STAT family members. Although there are seven mammalian STATs (STAT1-4, STAT5A, STAT5B and STAT6), numerous other STATs are commonly found in fishes or invertebrates, and the functions of these other STATs are also incorporated in this sub-classification tier.

### STATdbPredict

STATdbPredict is a meta-prediction system designed to characterize protein sequences based on STATdb classification. Users can submit one or more sequences in FASTA format to obtain a prediction of DA, lineage and/or function. The prediction process involves querying for the presence of statistically reliable STAT domains using Reversed Position Specific (RPS)-BLAST [26] and classifying them based on “DA”; this is followed by identifying the highest scoring pair (HSP) for the prediction of “Lineage” (tier two) and “Function” (tier three) using BLASTp [26] (see “STATdb Home > Help > Tools: Predict” for the prediction algorithm of STATdbPredict). STAT domain Position Specific Scoring Matrices (PSSMs) were downloaded from the NCBI conserved domain database (CDD) and used to create a local in-house RPS-BLAST-searchable database. This in-house searchable database is of much smaller size than the original CDD; thus, the values of the search output parameters (E-value, percentage identity, alignment length and bit score) will not be the same between the in-house database and original CDD. Since the E-value appears to be inversely proportional to database size [26], its values are larger for the in-house database and, thus, are deemed not appropriate as a parameter for the selection of significant domain hits. Therefore, the experimentally verified STAT records were used to determine the acceptable value range for the remaining other vital parameters (percentage identity,

**Table 1.** List of all fields defined for STATdb records.

Field Name	Description	Source#/Assigned
STATdb Id	STATdb Unique Identifier/Accession Number	<b>Assigned</b> (by STATdb authors)
gName	Gene Name	Source & <b>Assigned</b> (via NCBI Entrez Gene database)
pName	Protein Name	Source
STAT type	STAT family sub-group based on function	Source & <b>Assigned</b> (literature <sup>§</sup> )
STATdb Classification	Classification based on three-tier system	<b>Assigned</b> (by STATdb authors)
	Domain Architecture - Lineage - Function	
DBXRef	Database Cross References	Source & <b>Assigned</b> (pathway information obtained via KEGG database and other cross references are from source)
Literature	Literature (PubMed Reference Id)	Source
Species (Source Organism)	Species containing STAT	Source
Expt. Status	Experimental Status	<b>Assigned</b> (by STATdb authors)
	E - Experimentally Verified	
	P - Predicted/Hypothetical	
	U - Unknown	
Expt. Status Evidence	Experimental Status Evidence	<b>Assigned</b> (literature <sup>§</sup> )
ChromLoc	Chromosome Location	Source
IntPartners	Interacting Proteins	<b>Assigned</b> (via NCBI Entrez Gene database)
SeqLen	Sequence Length (Protein)	Source
Completeness	Completeness of the protein sequence	<b>Assigned</b> (by STATdb authors)
	Complete/Incomplete	
STAT Dom	STAT domains	Source
DomArchitecture	Domain Architecture	<b>Assigned</b> (via SMART database)
STAT DomSeq	Nucleotide & Protein Sequence of STAT domains	<b>Assigned</b> (derived from source)
STAT_int	Nucleotide & Protein Sequence for protein interaction domain	<b>Assigned</b> (derived from source)
STAT_alpha	Nucleotide & Protein Sequence for all alpha domain	<b>Assigned</b> (derived from source)
STAT_bind	Nucleotide & Protein Sequence for DNA binding domain	<b>Assigned</b> (derived from source)
STAT_sh2	Nucleotide & Protein Sequence for SH2 domain	<b>Assigned</b> (derived from source)
STAT_taz2	Nucleotide & Protein Sequence for TAZ2 domain	<b>Assigned</b> (derived from source)
BindingMotif	DNA Binding Motif	<b>Assigned</b> (via JASPAR database)
NucSeq	Nucleotide Sequence	<b>Assigned</b> (via NCBI Entrez Nucleotide database)
ProtSeq	Protein Sequence	Source
Comment	STATdb Curation Comments	<b>Assignable</b>

Fields that provide information selected from the source record (NCBI Entrez Protein database) are marked as "Source". "Assigned" fields are those not found in the source record, but were included to provide information obtained from analysis of the sequence data, existing annotations or the literature.

<sup>#</sup>NCBI Entrez protein database.

<sup>§</sup>The respective literature are indicated in the relevant records.

doi:10.1371/journal.pone.0104597.t001

alignment length and bit score) for each domain. The minimum range values of the three parameters (percentage identity, alignment length, and bit score) are used as a cut-off for statistical reliability of a domain hit (see "STATdb Home > Help > Tools: Predict" for the range values). The HSP is used to ascribe "Lineage" and "Function", and is defined as the best match to the query, with a percentage sequence identity of  $\geq 90$  and a length difference of  $\leq 10$ ; predictions based on HSPs that do not meet these criteria are indicated as hits of low confidence.

The accuracy of the prediction system was tested using a test dataset comprising new STAT sequences (as at June 2013) not found in STATdb (as at April 2013, STATdb comprised 1,424 records). These new sequences were obtained using the PSI-BLAST search against the NCBI NR database. The search resulted in 116 new STAT sequences, of which 20 were assigned a

"DA U". The remaining 96 classifiable sequences were non-redundant and used as positive samples for the test dataset, with the top 96 non-redundant, non-STAT hits from the PSI-BLAST used as negative samples. After this analysis was complete, the 116 new STAT sequences were added to STATdb.

### STATdb Construction

STATdb was created using MySQL (www.mysql.com) and the user interface was developed through the use of PHP, HTML and jQuery. MySQL is used for data storage, processing and retrieval of specific information. PHP (www.php.net) pages are used to process the forms and browse through the different sections of the database. HTML was utilised for the website design, with dynamic record browsing according to different groupings facilitated by jQuery (www.jquery.com), which is used to manage all the Java

Scripts and AJAX. Analysis tools supported by BioSLAX ([www.bioslax.com](http://www.bioslax.com)), such as BLAST similarity search and ClustalW [28] for multiple sequence alignments, are included in the database.

## Results

### Features of STATdb

Each record in STATdb is given a unique Id in the form of “STAT\_XXXXX”, where “XXXXX” represents five numerical digits. A sample record is provided in Figure 1. The records comprise standard data fields from the source databases (NCBI Entrez Protein database) and “Assigned” fields, which are defined by the authors for enriched manual annotations:

- STATdb Id – provides an Id for each individual STATdb record.
- Gene name – provides the gene name obtained from the NCBI Entrez Gene database.
- STAT type – STAT family sub-group.
- STATdb classification – provides a notation that earmarks the characteristic features of the sequence in terms of “DA”, “Lineage” and “Function” (see “Classification” section below for details).
- DBXRef – provides database cross-references that are mostly obtained from the source record; however, pathway information is obtained from the KEGG database.
- Experimental status validation – provides information from the literature and/or cross-referenced databases on the reliability of the STAT sequence, as either experimentally verified (E), or hypothetical/predicted (P) or unknown (U).
- IntPartners – lists the interacting protein partners of STAT, which were obtained from NCBI Entrez Gene database.
- Completeness of the protein sequence – the sequence is considered “complete” if all of the domains for the corresponding architecture are present.
- STAT Domain architecture – describes the order of the STAT domains in the sequence (via SMART [29,30] database).
- STAT Domain sequences – lists the amino acids and the corresponding nucleotide sequences (obtained by use of TBLASTN) of the individual STAT domains.
- Binding Motif – provides the STAT binding motif and the predicted target gene information obtained from JASPAR database.
- NucSeq – provides the nucleotide sequence obtained from the NCBI Entrez Nucleotide database for the corresponding protein.
- STATdb curation comments – this provides a platform for annotations and/or corrections by the STATdb community.

The key features of STATdb can be divided into basic and advanced, as described below:

#### A. Basic:

##### i. Keyword and Sequence Search

Keyword queries of the database include STATdb\_Id, gene name, protein name, STAT type, species, STAT domain, interacting proteins or other database cross-references. A sequence search is performed using BLAST against databases of (i) experimentally verified sequences, (ii) predicted, (iii) all STAT sequences (protein and nucleotides) and (iv) interacting partners (JAK, EGFR, and Src Kinase).

##### ii. Downloads

STATdb sequences categorized as “all sequences”, “experimentally verified”, “predicted” and sequences of interacting partners are available for download in FASTA format from the download page.

#### B. Advanced:

##### i. Browser

The Browser allows for dynamic browsing of the STATome according to all records, types of STAT, DNA or protein sequences, interacting proteins, status of experimental validation, and STAT DA (Figure 2A). Records can be selected to retrieve the full data or only the sequences in FASTA format, or they can be submitted for multiple sequence alignment on the fly using ClustalW.

##### ii. View 3D

The Jmol viewer allows for the manipulation of available 3D structures of STAT obtained from PDB. Currently, there are only 11 reported solved 3D structures for human (2), mouse (8) and the social amoeba *Dictyostelium discoideum* (1). Users can analyse the structures using the different options provided and also download primary sequences (FASTA format) and 3D structure coordinates (PDB format).

##### iii. Contribute


“Contribute” offers a platform for the STATdb community to curate annotations or submit new STAT sequences (Figure 2B). The submission of new STAT sequences will be checked and verified using the “Submission Policy” (see [http://statdb.bic.nus.edu.sg/downloads/submission\\_policy.pdf](http://statdb.bic.nus.edu.sg/downloads/submission_policy.pdf)). This would result in a database rich with annotations by expert curators in the field.

##### iv. Classification

STATs are complex proteins, but have been originally classified based simply on function and named according to their order of discovery (STAT types) [1,12,13]. The mammalian STAT family comprises seven different known members (STAT1-4, STAT5A, STAT5B and STAT6), which correspond to a determined function (see Table W4 at “STATdb Home > Classification”), and other types commonly found in fishes or invertebrates. The “STAT (s)” annotation is used to refer to the family or species-specific STATs, and the “(s)” represents the literature name of the STAT in the particular species. This includes STAT (dstA to D) of *Dictyostelium discoideum* and *Polysphondylium pallidum* PN500, STAT (D-STAT) of *Drosophila melanogaster*, and STAT (STA-1) and STAT (STATB) of *C. elegans*. Unknown, predicted or hypothetical STATs are denoted as STAT(u).

Although sequences of a STAT type are described to share the same function, our analysis shows that they possess differences in their domain architecture and, in some cases, appear to be lineage-specific [10,31,32,33,34]. As such, there might be subtle but distinct differences in the mode of function between family members of a STAT type, which merits further investigation. The rationale behind our classification system was to further stratify the original classification in a way that would allow for the quick delineation of possible structure, function and lineage of novel STATs.

The analysis of STAT type by structure revealed five distinct domain architectures (see Table W2 at “STATdb Home > Classification”). Domain architecture I (DA I) contains all of the five domains in the order of STAT\_int, STAT\_alpha, STAT\_bind, SH2 and TAZ2 from N- to C-terminus (see Materials and Methods for their descriptions). DA II lacks the TAZ2 domain, whereas DA III lacks both the TAZ2 and the STAT\_int. DA IV contains only the STAT\_bind and SH2 domains, and DA V comprises the coiled-coil domain (Dict\_STAT\_coil) and the SH2 domain. All other sequences that cannot be classified in this way—but contain or show similarity to at least one of the five major

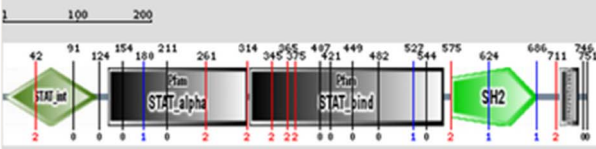
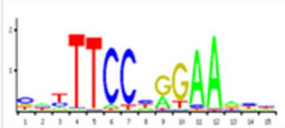
STATdb  
*A specialised resource for the STATome*


[Home](#)    [STATome](#)    [Classification](#)    [Tools](#)    [Contribute](#)    [Help](#)

[Home](#) > [Search](#) > [Keyword](#) > [Results](#) > [Sequence Record](#)

**Results**

Click [here](#) to comment or give feedback on the record.

STATdb Id	STAT_00001		
gName	ISGF-3; STAT91; DKFZp686B04100; STAT1	pName	Signal transducer and activator of transcription 1-alpha/beta isoform alpha (STAT1)
STAT type	STAT1	STATdb Classification	DA 1 : A : STAT1 ( <a href="#">STATdb classification details</a> )
DBXRef (Sequence)	<a href="#">GI (Protein) : 6274552</a> <a href="#">GenPept Acc. No. : NP_009330.1</a> <a href="#">GenBank Acc. No. : NM_007315.3</a>	DBXRef (Gene & Pathway)	<a href="#">Gene Id : 6772</a> <a href="#">KEGG Pathway Id : hsa04630</a>
Literature	<a href="#">20331378</a> <a href="#">20347693</a> <a href="#">15322115</a> <a href="#">14963018</a> <a href="#">14600148</a> <a href="#">12817007</a> <a href="#">12637327</a> <a href="#">12270932</a> <a href="#">12171910</a> <a href="#">12138178</a> <a href="#">11972023</a> <a href="#">11839738</a> <a href="#">11294897</a> <a href="#">11152457</a> <a href="#">11257227</a> <a href="#">10982844</a> <a href="#">10918587</a> <a href="#">1085106</a> <a href="#">10851046</a> <a href="#">9925928</a> <a href="#">9355737</a> <a href="#">7657660</a> <a href="#">7543024</a> <a href="#">754302</a> <a href="#">7885841</a> <a href="#">7514165</a> <a href="#">7690989</a> <a href="#">1502203</a> <a href="#">1496401</a>		
Species	<a href="#">Homo sapiens</a> ( <a href="#">List of species in this database can be found here</a> )	ChromLoc	2q32.2
Expt. Status	E	Expt. Status Evidence	<a href="#">Pubmed</a>
SeqLen	750	Completeness	Complete
IntPartners	CREBBP; GNB2L1; KIT; NMI; STAT1; SYK; CREBBP; EP300; ADRA1B; AKT1; BIMX; BRCA1; CAMK2D; CAMK2G; CCR1; CCR5; CSE1L; CSF2RB; CXCR4; DUSP3; EGFR; EIF1AD; EIF2AK2; ELP2; FADD; FANCC; FGFR3; FGFR4; FOS; FYN; GTF2I; HSF1; IFNAR2; IFNGR1; IL27RA; IL2RB; IL2RG; IRF1; IRF2; IRF9; JAK1; JAK2; JUN; KDR; KPNA1; KPNA6; LCK; LMO2; MAPK14; MCM3; MCM5; MDK; PDGFRA; PDGFRB; PIAS1; PRKCD; PRMT1; PTK2; PTPN11; PTPN2; RAC1; RELA; RPS6KA5; SRC; STAT2; STAT3; STAT5A; STAT5B; SUMO4; TNFRSF1A; TNFRSF1B; TRADD; TYK2; UBE2I; VDR; XPO1; ZNF467; ACTN4; BCL3; C20orf185; CTNBL1; DCTN1; HSP90AB1; KPNA2; LZTR1; MBD1; MTOR; RNF11; SHANK1; SMARCA4; SPTAN1; SPTBN1; TRIM28		
STAT Dom	SH2 Domain DNA-Binding Domain All Alpha Domain Protein Interaction Domain Transactivation Domain		
DomArchitecture	 <p>Domain architecture details by SMART</p>		
STAT DomSeq	<a href="#">Show Sequences (+)</a>		
BindingMotif	 <p><a href="#">Click here for details from the JASPAR core database</a></p> <p>Predicted Target gene: PAPP4</p>		
NucSeq	GCTGAGCGG ... <a href="#">See more</a>		
ProtSeq	MQQNYELQQE ... <a href="#">See more</a>		
Comment	N.A		

© Copyright 2012-2013 STATdb. All Rights Reserved.  
 updated on: 02/17/2014 23:08:45 by Chinari Pawan Kumar Patro

Browser best viewed on: Firefox 19.0 and above  
 Resolution best viewed on: (1280 x 720) or above

**Figure 1. A sample STATdb record.**  
 doi:10.1371/journal.pone.0104597.g001





**Figure 2. Snapshots of selected STATdb key features.** A) STATome Browser – allows for the dynamic browsing of the STATome, a complete set of reported STAT records in STATdb. B) Contribute – provides a platform for the STATdb community to curate annotations or submit new STAT sequences. C) Classification - provides a notation that describes the grouping of a sequence based on our three-tier classification system: “Domain Architecture – Lineage – Function” and D) Predict - characterizes protein sequences using STATdb classification. doi:10.1371/journal.pone.0104597.g002

domains—are labeled as DA U. All artificial sequences, even if they share the observed orders, are still classified as DA U.

The five domain architectures can be further differentiated into 14 unique lineages, notated as “A” to “M”, with “Z” for artificial sequences (see Table W3 at “STATdb Home > Classification”). Two lineages were observed for each bilateria (Deuterostomia and Protostomia), cnidaria (Anthozoa and Hydrozoa), choanoflagellida (Monosiga and Salpingoeca), and dictyosteliida (Dictyostelium and Polysphondylium), whereas one lineage was observed for each placozoa (trichoplax), porifera (demospongiae), ichthyosporea (Capsaspora), acanthamoeba and tracheophyta.

The stratification of STAT types into “DA” and “Lineage” resulted in a three-tier classification system, with notations, such as “DA I : A : STAT1”, which describes “DA” (tier one), “Lineage” (tier two), and “Function” (tier three), respectively (Figure 2C). Currently, the website comprises 96 notations that involve three tiers (see Table W1 at “STATdb Home > Classification”). Searches can thus be performed according to these collective notations or to each individual tier; this information is provided under the “Classification” section (“STATdb Home > Classification”) or can be found using the “Search” page (“STATdb Home > Search”). The classification will be updated regularly to

Processing Blast Search...

Searching for domains...

**RESULTS**

```
# Query: STAT_00001|STAT1|Homo_sapiens
# Processing...
# Contains

STAT domain, subject id, % identity, alignment length, q. start, q. end, s. start, s. end, evalue, bit score
STAT_bind domain      gnl|CDD|145817  68.50  254  317  567  1  254  3e-147  424
STAT_alpha domain     gnl|CDD|144561  57.14  182  136  315  1  182  2e-70  220
STAT_interaction domain gnl|CDD|198032  63.33  120  2  121  1  120  2e-59  187
SH2_STAT family domain gnl|CDD|198175  49.21  126  557  682  1  114  4e-51  164
STAT_TAZ2bind domain  gnl|CDD|152597  68.00  25  715  739  1  23  6e-08  40.5
SH2_STAT1 domain     gnl|CDD|198235  99.34  151  557  707  1  151  4e-108  319
```

**PREDICTION REPORT**

```
Query,Subject id,STATdb Classification,% identity,aln len,evalue,bit score
STAT_00001|STAT1      STAT_01046      DA I : A : STAT1      100.00  750  0.0  1552
```

See below for the alignment with the best hit

```
Query: 1
> STAT_01046|397509866:XP_003825332.1|DA_I:A:STAT1|Pan_paniscus
Length=750

Score = 1552 bits (4018), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 750/750 (100%), Positives = 750/750 (100%), Gaps = 0/750 (0%)

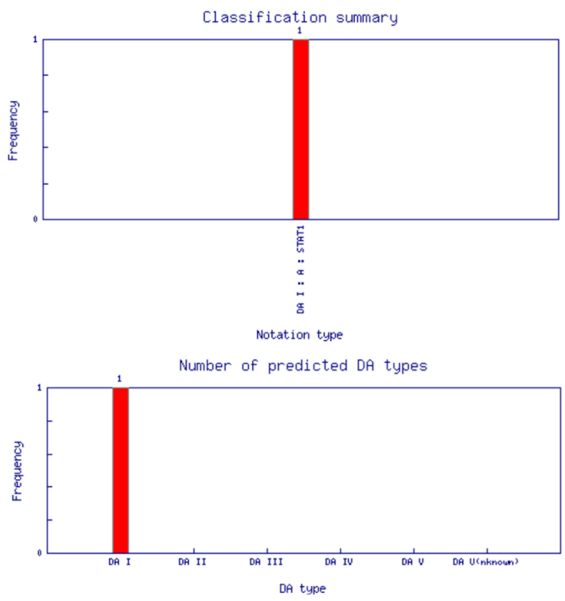
Query 1      MSQWYELQQLDKSFLEQVHQLYDDSPFMEIRQYLAQWLEKQDWEHAANDVSFATIRFHDL  60
Sbjct 1      MSQWYELQQLDKSFLEQVHQLYDDSPFMEIRQYLAQWLEKQDWEHAANDVSFATIRFHDL  60

Query 61     LSQLDDQYSRFSLENNFLLQHNIRKSKRNLQDNFQEDFIQMSMIYSCLKEERKILENAQ  120
Sbjct 61     LSQLDDQYSRFSLENNFLLQHNIRKSKRNLQDNFQEDFIQMSMIYSCLKEERKILENAQ  120
                * * * * *
                * * * * *
                * * * * *

Query 601    LRFESSREGAIFTWVERSQNGGEPDFHAVEPYTKKELSAVTFPDIIRNYKVMAAENIP  660
Sbjct 601    LRFESSREGAIFTWVERSQNGGEPDFHAVEPYTKKELSAVTFPDIIRNYKVMAAENIP  660

Query 661    ENPLKYLYPNIDKDHAFGKYSRPKEAPEPMELDGPKGTGYIKTELISVSEVHPSRLQTT  720
Sbjct 661    ENPLKYLYPNIDKDHAFGKYSRPKEAPEPMELDGPKGTGYIKTELISVSEVHPSRLQTT  720

Query 721    DNLLFMSPEEFDEVSRIVGSVEFDSMMNTV  750
Sbjct 721    DNLLFMSPEEFDEVSRIVGSVEFDSMMNTV  750
```



Prediction report in text can be downloaded [here](#)  
 Click [here](#) for STATdb Classification details

**Figure 3. STATdbPredict output report page for STAT\_00001.** The alignment is cropped to save space.  
 doi:10.1371/journal.pone.0104597.g003

**Table 2.** Performance measures of STATdbPredict (RPS-BLAST and BLASTp) versus standalone BLASTp search.

A. STATdbPredict search							
	TP	FN	TN	FP	Accuracy (%)	Sensitivity (%)	Specificity (%)
<b>Domain Architecture (DA)</b>	85	11	96	0	94.27	88.54	100
<b>Lineage</b>	95	1	96	0	99.48	98.96	100
<b>STAT type</b>	90	6	96	0	96.88	93.75	100
B. Standalone BLASTp search							
	TP	FN	TN	FP	Accuracy (%)	Sensitivity (%)	Specificity (%)
<b>Domain Architecture (DA)</b>	79	17	96	0	91.15	82.29	100
<b>Lineage</b>	95	1	96	0	99.48	98.96	100
<b>STAT type</b>	90	6	96	0	96.88	93.75	100

doi:10.1371/journal.pone.0104597.t002

provide a true representation of the STATome as the database grows.

#### v. Predict

STATdbPredict characterizes protein sequences through the STATdb classification system (Figure 2D). This prediction system reports the STATdb classification notation of the query sequence(s) along with any additional information, such as individual domain hits, the HSP, and the frequency of the different notations (Figure 3). This provides information on the potential structure, function and lineage of novel STATs, which can help in planning experiments for validation. STATdbPredict is essentially a combination of two BLAST programs: RPS-BLAST (against an in-house database of PSSM matrices downloaded from CDD) and standard BLASTp (against STATdb version without the test dataset) with optimized parameters that are applied in the context of the 3-tier classification. Outputs of STATdbPredict are annotated according to the classification, whereas a standalone BLAST search against STATdb sequences also annotated according to the 3-tier annotation would provide a similar result but of lower overall accuracy (~91% versus ~94%) and sensitivity (~82% versus ~89%) than STATdbPredict (Table 2). This is because RPS-BLAST, through the use of PSSM matrices, captures the diversity of the domains, which cannot be represented by a single HSP of a BLAST search. Even though the percentage differences in accuracy between Predict and standalone BLAST seem minor, the absolute number of records affected is significant and will be more so for a larger data size; for example, 17 were incorrectly identified by standalone BLAST for a test dataset of 96 positive and 96 negative samples. Nonetheless, both methods have a high overall accuracy because of the granular stratification of STAT sequences into the 3-tier classification system. The prediction system will be updated regularly for improved reliability as the size of the database grows. STATdb represents a platform for the future development of more sophisticated meta-predictors, with an increased number of record and corresponding annotations for scanning the tree of life genome/proteome for novel STATs in practical applications.

#### Application of STATdbPredict: Defining the STATome

The STATome represents all reported STAT sequences in nature. The sequences used to populate STATdb were obtained via two approaches: (i) a standard search of NCBI NR (see Materials and Methods for “Sequence Data Collection”) and (ii) STATdbPredict to scan UniProt UniRef100 [35] and NCBI NR datasets. At the time of collection, the UniRef100 dataset contained 20,002,214 sequences, whereas the NR dataset contained 23,075,327 sequences. The standard search returned 1,126 STAT sequences, whereas STATdbPredict identified an additional 65 unique sequences from NR and 233 from UniRef100. In addition, the 116 sequences identified during the accuracy analysis of STATdbPredict, which were obtained more recently using a standard NCBI NR search, were eventually included in STATdb, resulting in a total of 1,540 distinct sequence records.

STATdb is currently the only specialised repository of the STATome. Of the 1,540 records (as at June 2013), 186 are experimentally (“E”) verified STAT sequences, whereas 1,354 are predicted (“P”) (see submission/curation policy for grouping procedure). A total of 93 records have annotations of the interacting partners, which broadly fall under four groups: inhibitors, such as protein inhibitor of activated STAT (PIAS), and suppressors of cytokine signaling (SOCS); activators, such as JAK, Src kinase and EGFR; cytokines, such as interferons and interleukins, which comprise the majority; and unclassified, such



as JUN, BCL3, Gfap, EP300, among others. STAT is currently reported to be present in 235 species from diverse lineages, including bilateria, cnidaria, choanoflagellida, dictyosteliida, placozoa, porifera, ichthyosporea, acanthamoeba and tracheophyta. The STAT types—STAT1, STAT2, STAT4, STAT5A, STAT5B and STAT6—are represented by more than 100 records each, whereas STAT3 and STAT(u) comprise over 200 and 300 records, respectively (navigate to “STATdb Home > Help > Statistics” for the list of species and STAT types).

### Maintenance, stability and growth of STATdb

We have devised an update policy (see [http://statdb.bic.nus.edu.sg/downloads/update\\_policy.pdf](http://statdb.bic.nus.edu.sg/downloads/update_policy.pdf)) for the regular growth of the database. The stability of the database will be monitored regularly, and feedback from users will be key in addressing any bugs or issues within the system. Additionally, regression testing will be performed before major updates to ensure full functionality and stability. Plans for longevity of the database beyond the current team include a proposal for the long-term maintenance of the database by a group of volunteers selected from the list of top contributors. These users will be given the authority to make changes to the database in accordance with the standard system administrator acceptable use policy, and will also be responsible for maintaining the various policies of the database, such as new sequence submissions and update policies. Other plans include depositing the latest copy to Asia-Pacific Bioinformatics Network's (APBioNet's) cloud re-instantiation Web-accessible system (<http://biodb100.apbionet.org>; [36]) for archival and future on-demand re-instantiation by users where the original database site is not accessible. This is in line with the Minimum Information about a Bioinformatics Investigation (MIABI) standards [37], harmonised

with the BioDBcore standards of the International Society for Biocuration (ISB) and BioSharing [38].

### Discussion

STATdb is a unique Web resource that provides a comprehensive collection of STAT protein and nucleotide sequences, enriched with functional and structural annotations for data mining and analyses. The significant attributes of STATdb include: (a) integration of STAT data from different databases, creating the only unified STATome reported to date; (b) a novel classification system (comprising characteristic features of STAT protein sequences), which is used as a basis for STATdbPredict, a high accuracy (>90%) meta-predictor; (c) tools to analyse the functional and structural properties of STAT (BLAST, alignment, STATdbPredict); and (d) a platform for community contribution, which is guided by submission curation and an update policy. We envisage that this database will serve as a template for the development of a knowledgebase for signaling proteins.

### Acknowledgments

The authors thank Mr. Mark De Silva, Mr. Lim Kuan Siong and Mr. Mohammad Aslam Khan for their help and valuable suggestions.

### Author Contributions

Conceived and designed the experiments: CPKP AMK TWT. Performed the experiments: CPKP. Analyzed the data: CPKP. Wrote the paper: CPKP AMK TWT XYF. Supervised the research: AMK TWT XYF. Designed the web resource and developed the database: CPKP.

### References

- Levy DE, Darnell JE (2002) Stats: transcriptional control and biological impact. *Nat Rev Mol Cell Biol* 3: 651–662.
- Fu XY, Schindler C, Improta T, Aebersold R, Darnell JE Jr (1992) The proteins of ISGF-3, the interferon alpha-induced transcriptional activator, define a gene family involved in signal transduction. *Proc Natl Acad Sci U S A* 89: 7840–7843.
- Darnell JE Jr, Kerr IM, Stark GR (1994) Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* 264: 1415–1421.
- Stark GR, Darnell JE (2012) The JAK-STAT pathway at twenty. *Immunity* 36: 503–514.
- Leonard WJ, O'Shea JJ (1998) Jaks and STATs: biological implications. *Annu Rev Immunol* 16: 293–322.
- Ihle JN, Kerr IM (1995) Jaks and Stats in signaling by the cytokine receptor superfamily. *Trends Genet* 11: 69–74.
- Ihle JN (1996) STATs: signal transducers and activators of transcription. *Cell* 84: 331–334.
- Fu XY (1992) A transcription factor with SH2 and SH3 domains is directly activated by an interferon alpha-induced cytoplasmic protein tyrosine kinase(s). *Cell* 70: 323–335.
- Schindler C, Shuai K, Prezioso VR, Darnell JE Jr (1992) Interferon-dependent tyrosine phosphorylation of a latent cytoplasmic transcription factor. *Science* 257: 809–813.
- Decker T, Müller M (2012) *Jak-Stat Signaling: From Basics to Disease*: Springer. 448 p.
- Akira S (1999) Functional roles of STAT family proteins: lessons from knockout mice. *Stem Cells* 17: 138–146.
- Copeland NG, Gilbert DJ, Schindler C, Zhong Z, Wen Z, et al. (1995) Distribution of the mammalian Stat gene family in mouse chromosomes. *Genomics* 29: 225–228.
- Darnell JE Jr (1997) STATs and gene regulation. *Science* 277: 1630–1635.
- Park C, Li S, Cha E, Schindler C (2000) Immune response in Stat2 knockout mice. *Immunity* 13: 795–804.
- Lewis RS, Ward AC (2004) Conservation, duplication and divergence of the zebrafish stat5 genes. *Gene* 338: 65–74.
- Veeramani A, Gopalakrishnan K, Brusica V, Koh JL (2006) BioDART - Catalogue of biological data artifact examples. International Conference on Biomedical and Pharmaceutical Engineering, 2006. pp. 324–329.
- Mazumder R, Natale DA, Julio JA, Yeh LS, Wu CH (2010) Community annotation in biology. *Biol Direct* 5: 12.
- Stehr H, Duarte JM, Lappe M, Bhak J, Bolser DM (2010) PDBWiki: added value through community annotation of the Protein Data Bank. *Database* 2010.
- Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, et al. (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol* 9: R89.
- Huss JW III, Orozco C, Goodale J, Wu C, Batalov S, et al. (2008) A gene wiki for community annotation of gene function. *PLoS Biol* 6: e175.
- Daub J, Gardner PP, Tate J, Ramskold D, Manske M, et al. (2008) The RNA WikiProject: community annotation of RNA families. *RNA* 14: 2462–2464.
- McIntosh BK, Renfro DP, Knapp GS, Lairikyengbam CR, Liles NM, et al. (2012) EcoliWiki: a wiki-based community resource for *Escherichia coli*. *Nucleic Acids Res* 40: D1270–1277.
- Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, et al. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res* 40: D1301–1307.
- Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61–65.
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 38: D5–16.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
- Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680.
- Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* 95: 5857–5864.
- Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 40: D302–305.
- Wang Y, Levy DE (2006) *C. elegans* STAT: evolution of a regulatory switch. *FASEB J* 20: 1641–1652.
- Kay RR (1997) Dictyostelium development: lower STATs. *Curr Biol* 7: R723–725.

33. Hombria JC, Brown S (2002) The fertile field of *Drosophila* Jak/STAT signalling. *Curr Biol* 12: R569–575.
34. Zeidler MP, Bach EA, Perrimon N (2000) The roles of the *Drosophila* JAK/STAT pathway. *Oncogene* 19: 2598–2606.
35. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282–1288.
36. Tan TW, Xie C, De Silva M, Lim KS, Patro CPK, et al. (2013) Simple re-instantiation of small databases using cloud computing. *BMC Genomics* 14 Suppl 5: S13.
37. Tan TW, Tong JC, Khan AM, de Silva M, Lim KS, et al. (2010) Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi). *BMC Genomics* 11 Suppl 4: S27.
38. Gaudet P, Bairoch A, Field D, Sansone S-A, Taylor C, et al. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Database (Oxford)* 2011: baq027.