Bio-TDS: Bioscience Query Tool Discovery System

Etienne Z. Gnimpieba[1,2*], Menno S. VanDiermen[1], Shayla M. Gustafson[1], Bill Conn[1], Carol M. Lushbough[1,2]

[1]Biomedical Engineering Department, University of South Dakota, 4800 North Career Ave, Sioux Falls, SD 57107,
[2]BioSNTR, Brookings, SD 57006, USA

To whom correspondence should be addressed: +1 605 2749578; Etienne.Gnimpieba@usd.edu

SUPPORTING MATERIALS

The **Bio-TDS** (**Bio**Query **T**ools **D**iscovery **S**ystems) has been developed to assist researchers in retrieving the most applicable analytic tools by allowing them to formulate their questions as free text. The Bio-TDS is a flexible retrieval system that affords users from multiple bioscience domains (e.g. genomic, proteomic, bio-imaging) the ability to query over 15,000 analytic tool descriptions integrated from well-established, community repositories. One of the primary components of the Bio-TDS system is the ontology and natural language processing workflow for annotation, curation, query processing, and evaluation. The Bio-TDS's scientific impact was evaluated using sample questions posed by researchers retrieved from Biostars, a site focusing on biological data analysis. The Bio-TDS was compared to five similar bioscience analytic tool retrieval systems with the Bio-TDS outperforming the others in terms of relevance and completeness. The Bio-TDS offers researchers the capacity to associate their bioscience question with the most relevant computational toolsets required for the data analysis in their knowledge discovery process.

Table 2: Support Material Overview (also available at http://biotds.org/help/supporting.xhtml)

S1	BETS specification description and manipulation
S2	Resources extraction and semi-automatics curation
S3	TONER: Tools ontology-based annotation
S4	Bio-TDS Query processing workflow and programmatic access
S5	Bio-TDS evaluation and comparison

S1 - BETS Specification description and manipulation

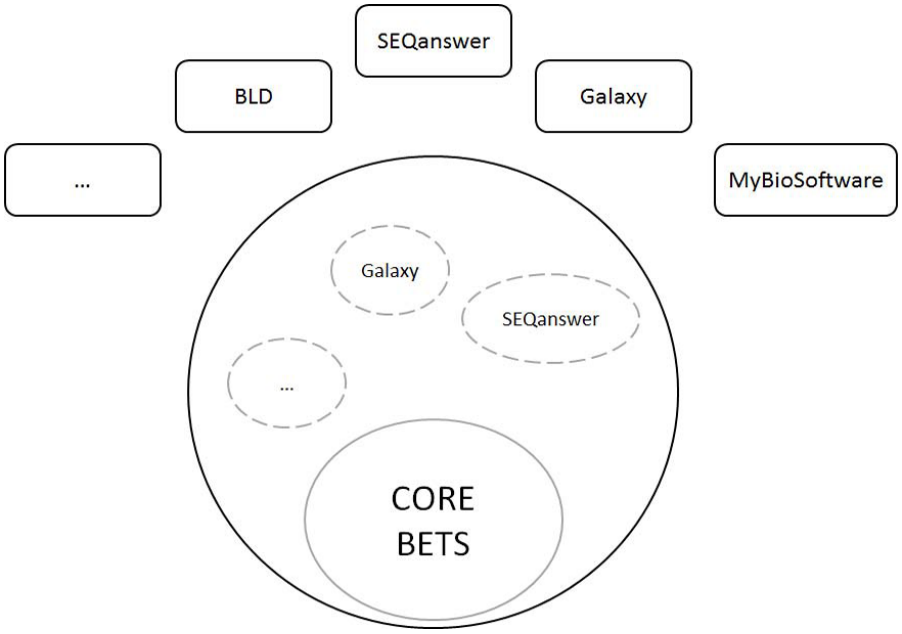


Figure S1a: BETS design Overview

Bioinformatics **E**laborated **T**ools **S**pecifications (**BETS**) provides a standard for analytic tool descriptions. The analytic tool descriptions (i.e. metadata) gathered from community tool repositories integrated into the Bio-TDS are stored in JSON format using the BETS standard. This standard consists of core BETS attributes and domains/repositories specifics attributes (Figure S1a). The core BETS attributes are manually mapped to the repository attribute.

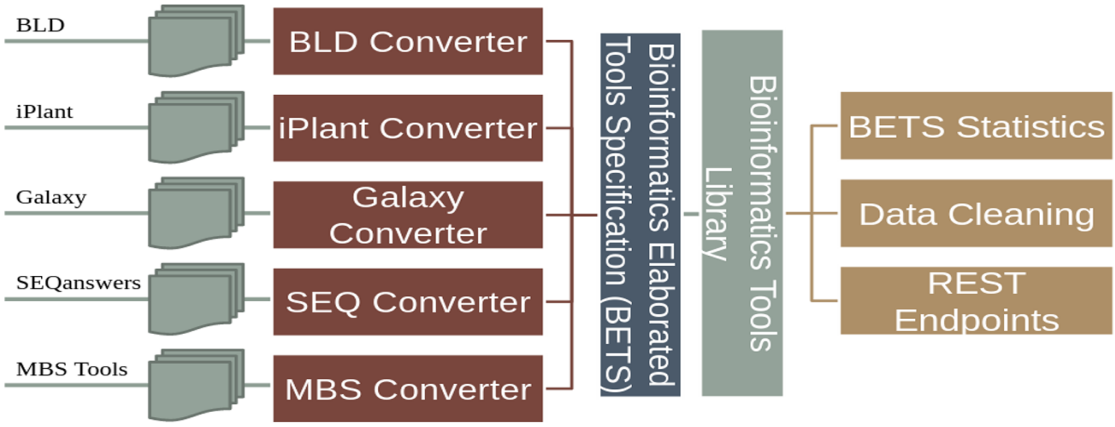


Figure S1b: BETS Converter workflow

S2 - Resources extraction and semi-automatics curation

The **Bio-TDS** combines bioinformatics tools from five other repositories and stores them in one central location, following **BETS** (**B**ioinformatics **E**laborated **T**ool **S**pecification). There are six main modules that convert the data from each of the five repositories into BETS tools and store the new tools into the Bio-TDS database.

The **BETS Checker** is a Java application that tests the compatibility of a tool with the BETS specification. A tool is considered "compatible" if it is in the format specified by the specific BETS converter. For example, the system contains a mapper called **Galaxy Converter**. A tool from the Galaxy Tool Shed can only be "compatible" if it matches the predefined Galaxy format (Figure S2a0.

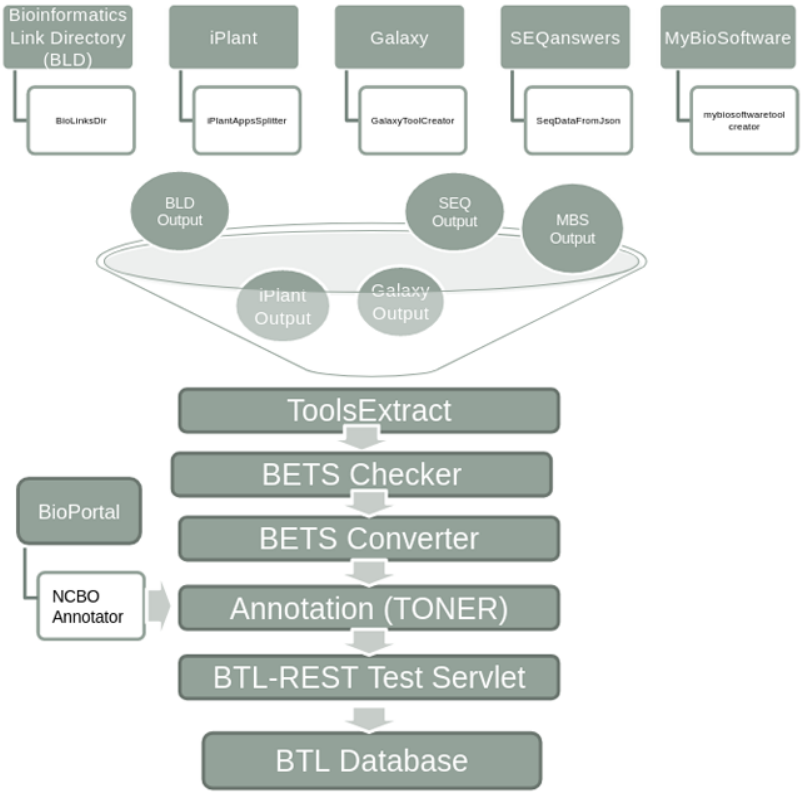


Figure S2a: Bio-TDS extraction workflow

A rule-based ***(or predicate)*** semi-automatic curation process has been developed for the Bio-TDS by combining human inspection and data mining methods. Rules are generated manually and are applied in a very specific order (Figure S2b). For example, some rules such as ***[if tool type is 'not tool', then remove the tool from the repository]*** are applied during the extraction process. If the scope of this rule is limited to one attribute, its consistency among repositories and impact on the curation process is relevant and effective. Following this logic, several categories were identified and used to lead the development of the rules: community/domain related rules, design related rules, and human intuitive rules (Table S2). After integration and curation, we have a very good improvement in our repository content.

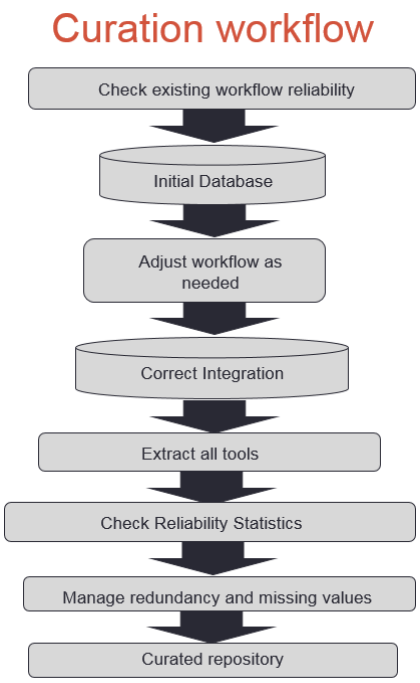


Figure S2b: Curation workflow

Community or domain related rules--are governed by information representation and information management in each community. For scientific communities such as bioinformatics, EDAM ontology is an important semantic and lexical information representation. These considerations led to the development of rules such as [if "input format" value is empty, and there is a mention in the "Description" in "list of input format from EDAM", then fill the input format field with the mention]. Ontologies also allow us to check relationships between terms. Therefore, a relationship such as "synonym" from the ontology can help to populate some missing value or remove redundant data.

Design based rules include constraints related to data integration issues. The key goal is to keep the integrated database lossless when applying operations such as merging duplicates. For example, considering the following rule [if "Name" and "version" are the same for two tools then, merge them]. This definition of tool similarity allows for a simple string comparison of two attributes (name, version) using simple edit distance. During and after the integration of analytic tool definitions, missing data management rules are used to minimize the empty values in the Bio-TDS repository.

Human inspection remains the key and most accurate action during the curation process. This include rules for consistency checking (i.e. no contradiction, complete and close rules), data quality overview, and new intuitive rules generation (i.e. looking at data and statistics, some rules can be inferred, and human walkthrough suggestions. A simple Web UI was developed to allow contributors to add curation suggestions.

At the current development stage, the Bio-TDS team has already identified 10 key rules (Table S2).The application of that rule set has helped to improve the repository accuracy by removing duplicates and invalid tools (from 19,000 tools to 15,000), removing inaccurate attributes values, and filling in missing information. This has achieved an overall improvement rate of ~30%.

Table S2: Selected curation rules

#	Type**	Condition (if…)	Consequence (then…)	Notes /Scope
1	HR	Tool type not "Tool"	Remove the tool from the repository	Extraction, Galaxy
2	DR	"Version" contain string	Remove string and keep numeric	BioQueryTool repository
3	DR	"Name" and "version" are the same for 2 tools	Merge them	BioQueryTool repository
4	CR	"input format" value is empty, and there is a mention in the "Description" in "list of input format from EDAM"	Fill the input format field with the mention	BioQueryTool repository
5	DR	The tool source specification is compatible with BETS	Send to the SUCCESS folder for BETS converter and add to the repository. Otherwise, send to ERROR folder for manual integration.	Extraction (ToolChecker), all integrated repositories
6	HR	In many tools a given attribute value is empty	Manually check and locate the most relevant related field to populate them.	Extraction, BioQueryTool repository (e.g. when platform is "R Bioconductor", "Operating System" attribute is filled)
7	DR	A tool attribute contain 2 or more duplicate	Remove duplicate	Extraction, all repositories
8	CR	The tools annotation is weak (<5 annotation terms)	Use the ontology context (hierarchy parent-child, synonyms, preferred name) to enrich the a notation	TONER, BioQueryTool repository
9	HR	A given attribute value is empty	Use homolog attribute to populate the attribute value	BioQueryTool repository e.g. if there is no "link" or "author" for a given tool, Use "reference URL" or "contact" to populate the field
10	HR	A tool doesn't have a quality attribute	Check if there is similar(name, reference, …) tool from SEQAnswer or GALAXY, and use the related quality attribute (citation count, …).	BioQueryTool repository

Rules Type: CR= Community Rules; DR= Design Rules ; HR= Human intuitive Rules

S3 - TONER: Tools ontology-based annotation

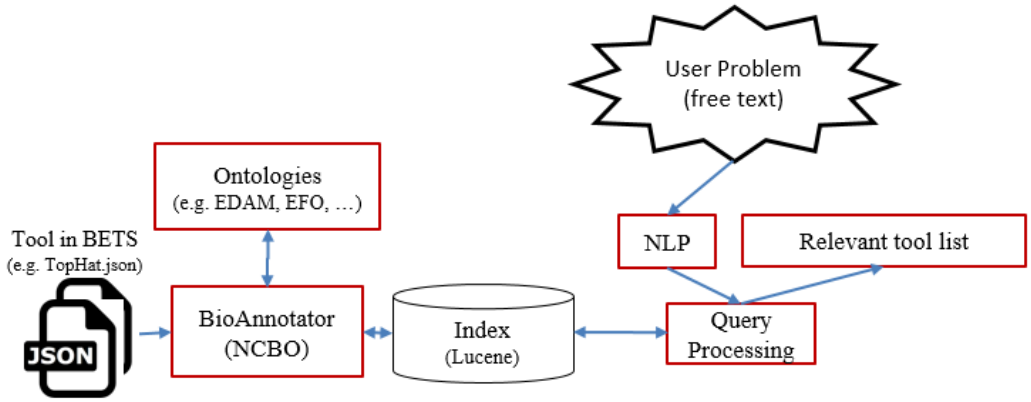


Figure S3: TONER workflow

S4- Bio-TDS Query processing workflow and programmatic access

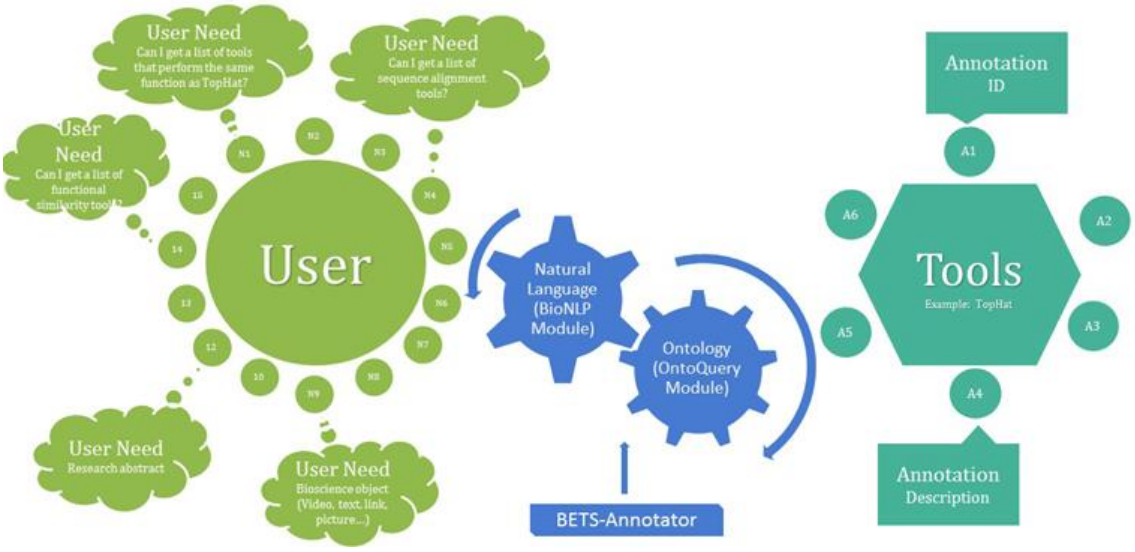


Figure S4a: Bio-TDS System Workflow

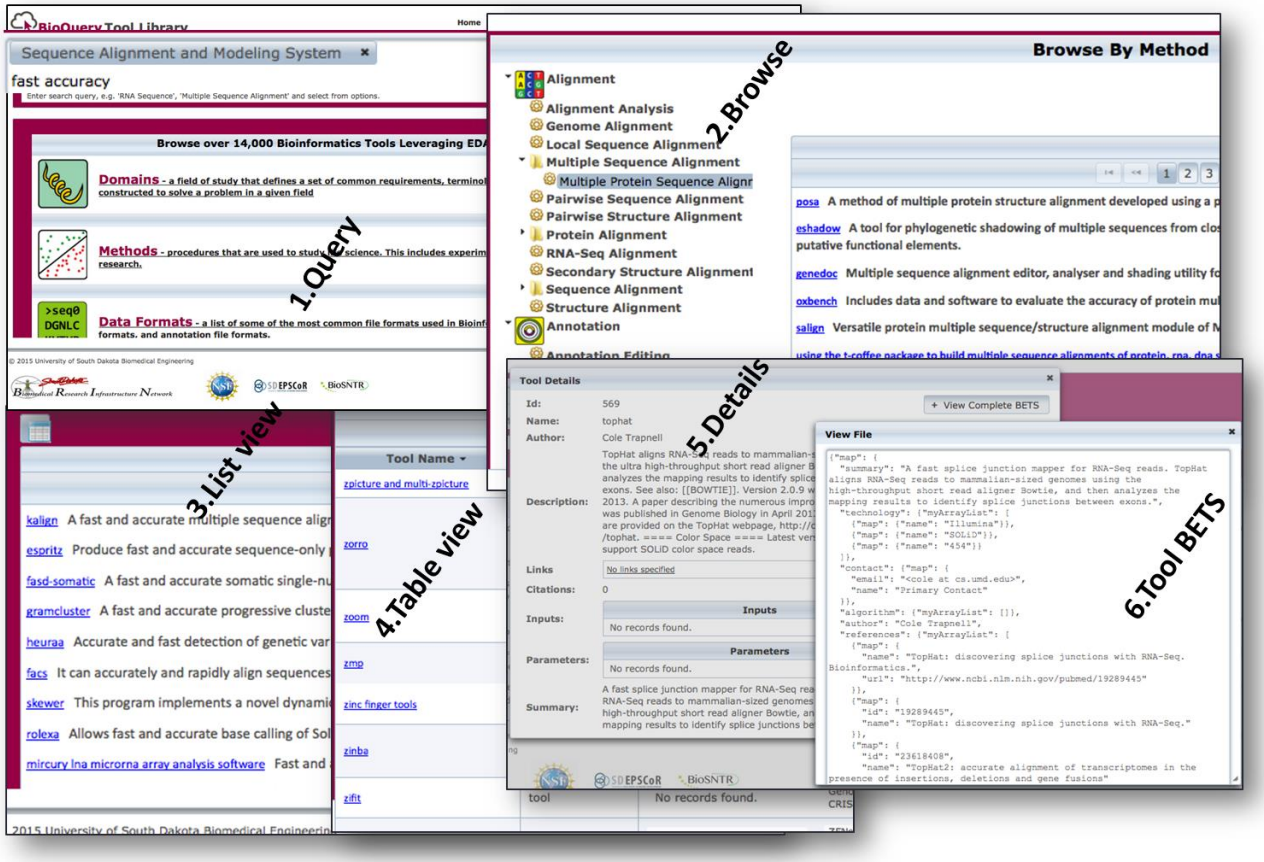


Figure S4b: Bio-TDS Display View Options

Table S4: REST Endpoints (Base URL http://jacksons.usd.edu/BTL-REST)

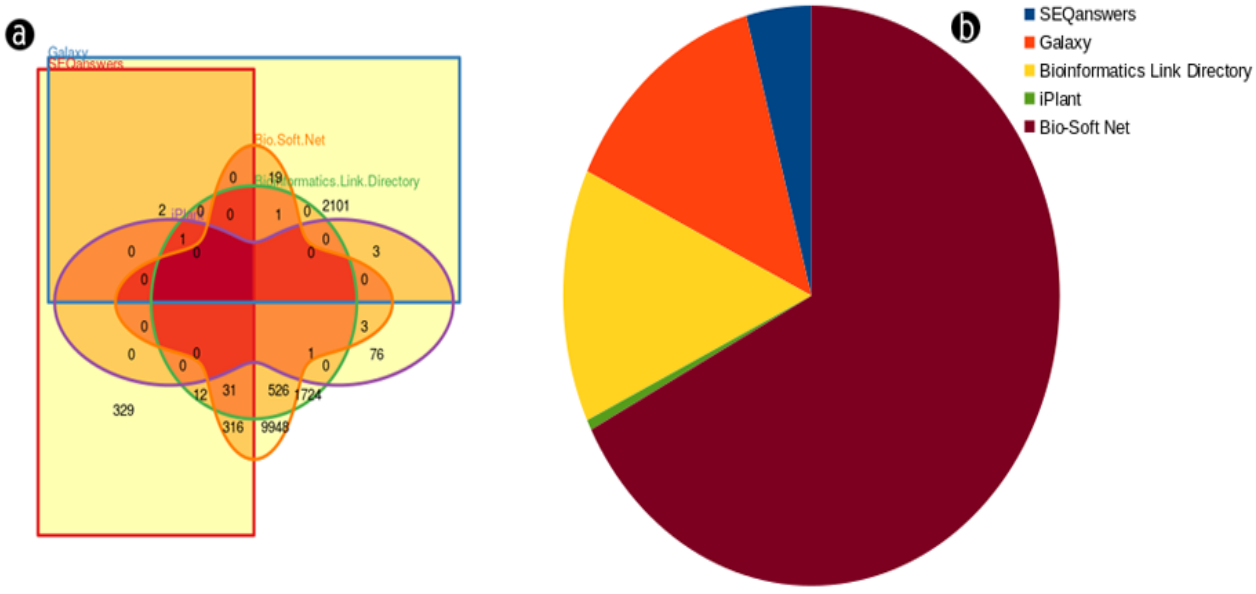
URL	HTTP VERB	RETURNS
Tools		
/resources/tools/	GET	Disabled - Get ranges instead - List of All Tools
/resources/tools/{id}	GET	Tool by ID
/resources/tools/{from}/{to}	GET	Tools within a range
/resources/tools/count	GET	Count of All Tools
/resources/tools/names	GET	All Tool Names
/resources/tools/summaries	GET	All Tool Summaries
BETS		
/resources/bets/	GET	Disabled - Get individual bets - All BETS data for each tools
/resources/bets/{id}	GET	BETS for tool by ID
/resources/bets/inputs/{id}	GET	Inputs for Tool by ID
Bridges		
/resources/bridges/	GET	All Bridges data
/resources/bridges/{id}	GET	Bridge by ID
/resources/bets/{from}/{to}	GET	Bridges within a range
/resources/bets/count	GET	Count of the Bridges
Repositories		
/resources/repos/	GET	List of the Repositories our tools came from
/resources/repos/tools	GET	List of Every Tool ID and its Repo ID
/resources/repos/tools/{id}	GET	ResourceTool by ID
/resources/repos/getbytoolids	GET	Get example JSON for posting
/resources/repos/getbytoolids	POST	Post an array of tool ids to get their repository names
Storing Bridges or Tools (these are normally disabled)		
/resources/store-bridges	GET	An example Bridge JSON to POST
/resources/store-bridges	POST	Returns the URI of the created bridge
/resources/store-bridges/list	POST	Disabled
/resources/store-bridges/{id}	POST	Accepted if it updates the bridge id specified
/resources/store-tools	GET	Returns an example Tool JSON to Post
/resources/store-tools	POST	Disabled
/resources/store-tools/list	POST	Disabled
/resources/store-tools/{id}	POST	Accepted if it updates the tool id specified

*Subject to change

*Query parameter 'PrettyPrint=true' will return prettified JSON

S5 - Bio-TDS Evaluation and comparison

Note: The data used for the current tests have been collected from the associated repositories in December 2015. Because each repository may have been updated, the reproducibility should consider repository versions for accuracy.



c

Table 1: Bio-TDS evaluation and comparison overview

Criteria*	Bio-TDS	BLD	ELIXIR	GALAXY	SeqAnswer
MRR+	**1**	0.0131	0.0087	0.0043	0.1484
MAP+	**0.0004**	NS	NS	NS	NS
MAR+	**0.8755**	0	0	0.0036	0.0339
MAF+	**0.0008**	NS	NS	NS	NS
MRR++	**0.7598**	NS	0.1441	0.131	0.6899
MAP++	0.0427	NS	NS	NS	**0.0572**
MAR++	**0.3474**	0.02	0.0696	0.0518	0.2327
MAF++	0.0801	NS	NS	NS	**0.1383**

*Evaluation Criteria: **MRR** = Mean Retrieval Rate; **MAP** = Mean Average Precision; **MAR** = recall; **MAF** = mean Average F-measure. **NS** = not significant result. **User Query Type**: +Free text Query; ++Keyword Query

Figure S5a: Bio-TDS Evaluation and Comparison Overview

Figure S5a shows the existing repository statistic (b), as of May 2016, the tools count overlap Venn diagram (a), and repositories comparison overview (c).

Note: "NS" value in a given evaluation criteria (Precision, Recall,...) indicates limited data point (missing >40% data points compare the variable dataset size) to compute an accurate meaningful criteria value. This is due to a low retrieval rate in the related repository (e.g. no result return for the query).

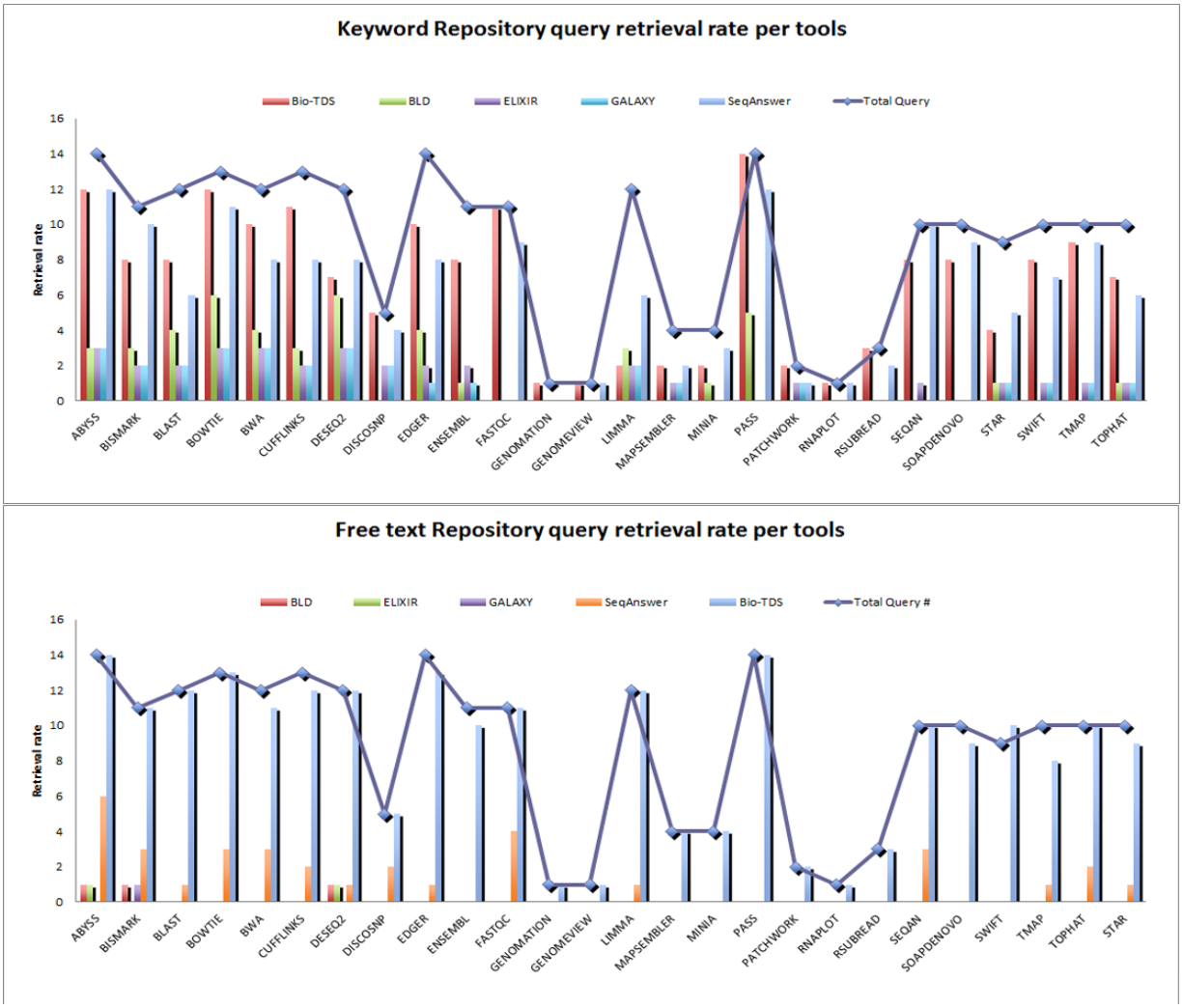
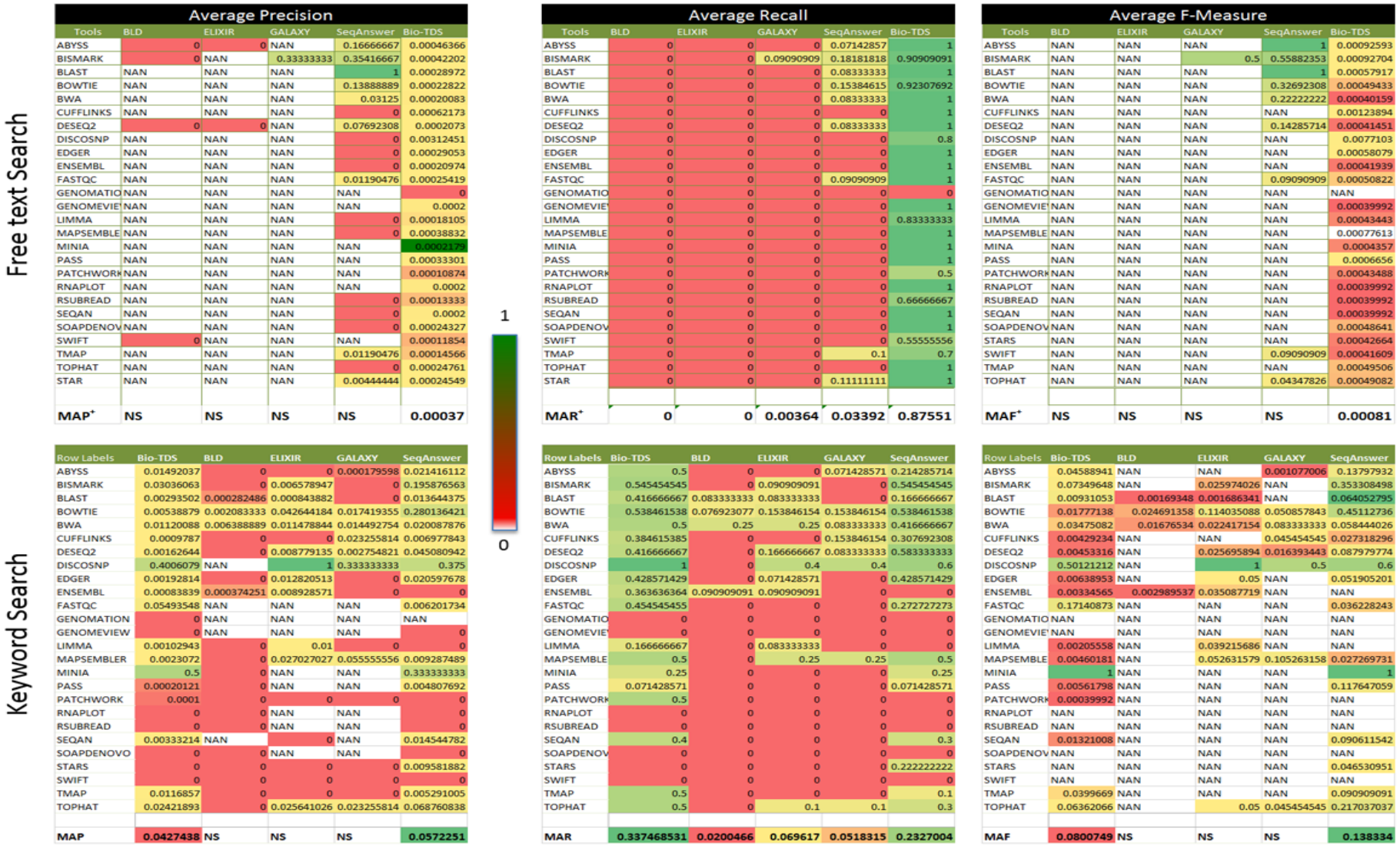


Figure S5b: Bio-TDS Retrieval Rate Comparison



Average Precision

Tools	BLD	ELIXIR	GALAXY	SeqAnswer	Bio-TDS
ABYSS	0	0	NAN	0.16666667	0.00046366
BISMARK	0	NAN	0.33333333	0.35416667	0.00042202
BLAST	NAN	NAN	NAN	1	0.00028972
BOWTIE	NAN	NAN	NAN	0.13888889	0.00022822
BWA	NAN	NAN	NAN	0.03125	0.00020083
CUFFLINKS	NAN	NAN	NAN	0	0.00062173
DESEQ2	0	0	NAN	0.07692308	0.0002073
DISCOSNP	NAN	NAN	NAN	0	0.00312451
EDGER	NAN	NAN	NAN	0	0.00029053
ENSEMBL	NAN	NAN	NAN	0	0.00020974
FASTQC	NAN	NAN	NAN	0.01190476	0.00025419
GENOMATIO	NAN	NAN	NAN	NAN	0
GENOMEVIE	NAN	NAN	NAN	NAN	0.0002
LIMMA	NAN	NAN	NAN	0	0.00018105
MAPSEMBLE	NAN	NAN	NAN	0	0.00038832
MINIA	NAN	NAN	NAN	NAN	0.0002179
PASS	NAN	NAN	NAN	NAN	0.00033301
PATCHWORK	NAN	NAN	NAN	NAN	0.00010874
RNAPLOT	NAN	NAN	NAN	NAN	0.0002
RSUBREAD	NAN	NAN	NAN	0	0.00013333
SEQAN	NAN	NAN	NAN	0	0.0002
SOAPDENOV	NAN	NAN	NAN	0	0.00024327
SWIFT	0	NAN	NAN	NAN	0.00011854
TMAP	NAN	NAN	NAN	0.01190476	0.00014566
TOPHAT	NAN	NAN	NAN	0	0.00024761
STAR	NAN	NAN	NAN	0.00444444	0.00024549
MAP⁺	**NS**	**NS**	**NS**	**NS**	**0.00037**

Average Recall

Tools	BLD	ELIXIR	GALAXY	SeqAnswer	Bio-TDS
ABYSS	0	0	0	0.07142857	1
BISMARK	0	0	0.09090909	0.18181818	0.90909091
BLAST	0	0	0	0.08333333	1
BOWTIE	0	0	0	0.15384615	0.92307692
BWA	0	0	0	0.08333333	1
CUFFLINKS	0	0	0	0	1
DESEQ2	0	0	0	0.08333333	1
DISCOSNP	0	0	0	0	0.8
EDGER	0	0	0	0	1
ENSEMBL	0	0	0	0	1
FASTQC	0	0	0	0.09090909	1
GENOMATIO	0	0	0	0	0
GENOMEVIE	0	0	0	0	1
LIMMA	0	0	0	0	0.83333333
MAPSEMBLE	0	0	0	0	1
MINIA	0	0	0	0	1
PASS	0	0	0	0	1
PATCHWORK	0	0	0	0	0.5
RNAPLOT	0	0	0	0	1
RSUBREAD	0	0	0	0	0.66666667
SEQAN	0	0	0	0	1
SOAPDENOV	0	0	0	0	1
SWIFT	0	0	0	0	0.55555556
TMAP	0	0	0	0.1	0.7
TOPHAT	0	0	0	0	1
STAR	0	0	0	0.11111111	1
MAR⁺	**0**	**0**	**0.00364**	**0.03392**	**0.87551**

Average F-Measure

Tools	BLD	ELIXIR	GALAXY	SeqAnswer	Bio-TDS
ABYSS	NAN	NAN	NAN	1	0.00092593
BISMARK	NAN	NAN	0.5	0.55882353	0.00092704
BLAST	NAN	NAN	NAN	1	0.00057917
BOWTIE	NAN	NAN	NAN	0.32692308	0.00049433
BWA	NAN	NAN	NAN	0.22222222	0.00040159
CUFFLINKS	NAN	NAN	NAN	NAN	0.00123894
DESEQ2	NAN	NAN	NAN	0.14285714	0.00041451
DISCOSNP	NAN	NAN	NAN	NAN	0.0077103
EDGER	NAN	NAN	NAN	NAN	0.00058079
ENSEMBL	NAN	NAN	NAN	NAN	0.00041939
FASTQC	NAN	NAN	NAN	0.09090909	0.00050822
GENOMATIO	NAN	NAN	NAN	NAN	NAN
GENOMEVIE	NAN	NAN	NAN	NAN	0.0003992
LIMMA	NAN	NAN	NAN	NAN	0.0004344
MAPSEMBLE	NAN	NAN	NAN	NAN	0.00077613
MINIA	NAN	NAN	NAN	NAN	0.0004357
PASS	NAN	NAN	NAN	NAN	0.0006656
PATCHWORK	NAN	NAN	NAN	NAN	0.0004348
RNAPLOT	NAN	NAN	NAN	NAN	0.0003992
RSUBREAD	NAN	NAN	NAN	NAN	0.0003992
SEQAN	NAN	NAN	NAN	NAN	0.0003992
SOAPDENOV	NAN	NAN	NAN	NAN	0.0004864
STARS	NAN	NAN	NAN	NAN	0.0004266
SWIFT	NAN	NAN	NAN	0.09090909	0.0004160
TMAP	NAN	NAN	NAN	NAN	0.0004950
TOPHAT	NAN	NAN	NAN	0.04347826	0.0004908
MAF⁺	**NS**	**NS**	**NS**	**NS**	**0.00081**

Row Labels	Bio-TDS	BLD	ELIXIR	GALAXY	SeqAnswer
ABYSS	0.01492037	0	0	0.000179598	0.021416112
BISMARK	0.03036063	0	0.006578947	0	0.195876563
BLAST	0.00293502	0.000282486	0.000843882	0	0.013644375
BOWTIE	0.00538879	0.002083333	0.042644184	0.017419355	0.280136421
BWA	0.01120088	0.006388889	0.011478844	0.014492754	0.020087876
CUFFLINKS	0.0009787	0	0	0.023255814	0.006977843
DESEQ2	0.00162644	0	0.008779135	0.002754821	0.045080942
DISCOSNP	0.4006079	NAN	1	0.333333333	0.375
EDGER	0.00192814	0	0.012820513	0	0.020597678
ENSEMBL	0.00083839	0.000374251	0.008928571	0	0
FASTQC	0.05493548	NAN	NAN	NAN	0.006201734
GENOMATION	0	NAN	NAN	NAN	NAN
GENOMEVIEW	0	NAN	NAN	NAN	0
LIMMA	0.00102943	0	0.01	0	0
MAPSEMBLER	0.0023072	0	0.027027027	0.055555556	0.009287489
MINIA	0.5	0	NAN	NAN	0.333333333
PASS	0.00020121	0	NAN	NAN	0.004807692
PATCHWORK	0.0001	0	0	0	0
RNAPLOT	0	0	NAN	NAN	0
RSUBREAD	0	0	NAN	NAN	0
SEQAN	0.00333214	NAN	0	NAN	0.014544782
SOAPDENOVO	0	0	NAN	NAN	0
STARS	0	0	0	0	0.009581882
SWIFT	0	0	0	0	0
TMAP	0.0116857	0	0	0	0.005291005
TOPHAT	0.02421893	0	0.025641026	0.023255814	0.068760838
MAP	**0.0427438**	**NS**	**NS**	**NS**	**0.0572251**

Row Labels	Bio-TDS	BLD	ELIXIR	GALAXY	SeqAnswer
ABYSS	0.5	0	0	0.071428571	0.214285714
BISMARK	0.545454545	0	0.090909091	0	0.545454545
BLAST	0.416666667	0.083333333	0.083333333	0	0.166666667
BOWTIE	0.538461538	0.076923077	0.153846154	0.153846154	0.538461538
BWA	0.5	0.25	0.25	0.083333333	0.416666667
CUFFLINKS	0.384615385	0	0	0.153846154	0.307692308
DESEQ2	0.416666667	0	0.166666667	0.083333333	0.583333333
DISCOSNP	1	0	0.4	0.4	0.6
EDGER	0.428571429	0	0.071428571	0	0.428571429
ENSEMBL	0.363636364	0.090909091	0.090909091	0	0
FASTQC	0.454545455	0	0	0	0.272727273
GENOMATIO	0	0	0	0	0
GENOMEVIE	0	0	0	0	0
LIMMA	0.166666667	0	0.083333333	0	0
MAPSEMBLE	0.5	0	0.25	0.25	0.5
MINIA	0.25	0	0	0	0.25
PASS	0.071428571	0	0	0	0.071428571
PATCHWORK	0.5	0	0	0	0
RNAPLOT	0	0	0	0	0
RSUBREAD	0	0	0	0	0
SEQAN	0.4	0	0	0	0.3
SOAPDENOV	0	0	0	0	0
STARS	0	0	0	0	0.222222222
SWIFT	0	0	0	0	0
TMAP	0.5	0	0	0	0.1
TOPHAT	0.5	0	0.1	0.1	0.3
MAR	**0.337468531**	**0.0200466**	**0.069617**	**0.0518315**	**0.2327004**

Row Labels	Bio-TDS	BLD	ELIXIR	GALAXY	SeqAnswer
ABYSS	0.04588941	NAN	NAN	0.001077006	0.13797932
BISMARK	0.07349648	NAN	0.025974026	NAN	0.353308498
BLAST	0.00931053	0.00169348	0.001686341	NAN	0.064052795
BOWTIE	0.01777138	0.024691358	0.114035088	0.050857843	0.45112736
BWA	0.03475082	0.01676534	0.022417154	0.083333333	0.058444026
CUFFLINKS	0.00429234	NAN	NAN	0.045454545	0.027318296
DESEQ2	0.00453316	NAN	0.025695894	0.016393443	0.087979774
DISCOSNP	0.50121212	NAN	1	0.5	0.6
EDGER	0.00638953	NAN	0.05	NAN	0.051905201
ENSEMBL	0.00334565	0.002989537	0.035087719	NAN	NAN
FASTQC	0.17140873	NAN	NAN	NAN	0.036228243
GENOMATIO	NAN	NAN	NAN	NAN	NAN
GENOMEVIE	NAN	NAN	NAN	NAN	NAN
LIMMA	0.00205558	NAN	0.039215686	NAN	NAN
MAPSEMBLE	0.00460181	NAN	0.052631579	0.105263158	0.027269731
MINIA	1	NAN	NAN	NAN	1
PASS	0.00561798	NAN	NAN	NAN	0.117647059
PATCHWORK	0.00039992	NAN	NAN	NAN	NAN
RNAPLOT	NAN	NAN	NAN	NAN	NAN
RSUBREAD	NAN	NAN	NAN	NAN	NAN
SEQAN	0.01321008	NAN	NAN	NAN	0.090611542
SOAPDENOV	NAN	NAN	NAN	NAN	NAN
STARS	NAN	NAN	NAN	NAN	0.046530951
SWIFT	NAN	NAN	NAN	NAN	NAN
TMAP	0.0399669	NAN	NAN	NAN	0.090909091
TOPHAT	0.06362066	NAN	0.05	0.045454545	0.217037037
MAF	**0.0800749**	**NS**	**NS**	**NS**	**0.138334**

Figure S5c: Bio-TDS exactness (Precision) and completeness (Recall) Comparison.