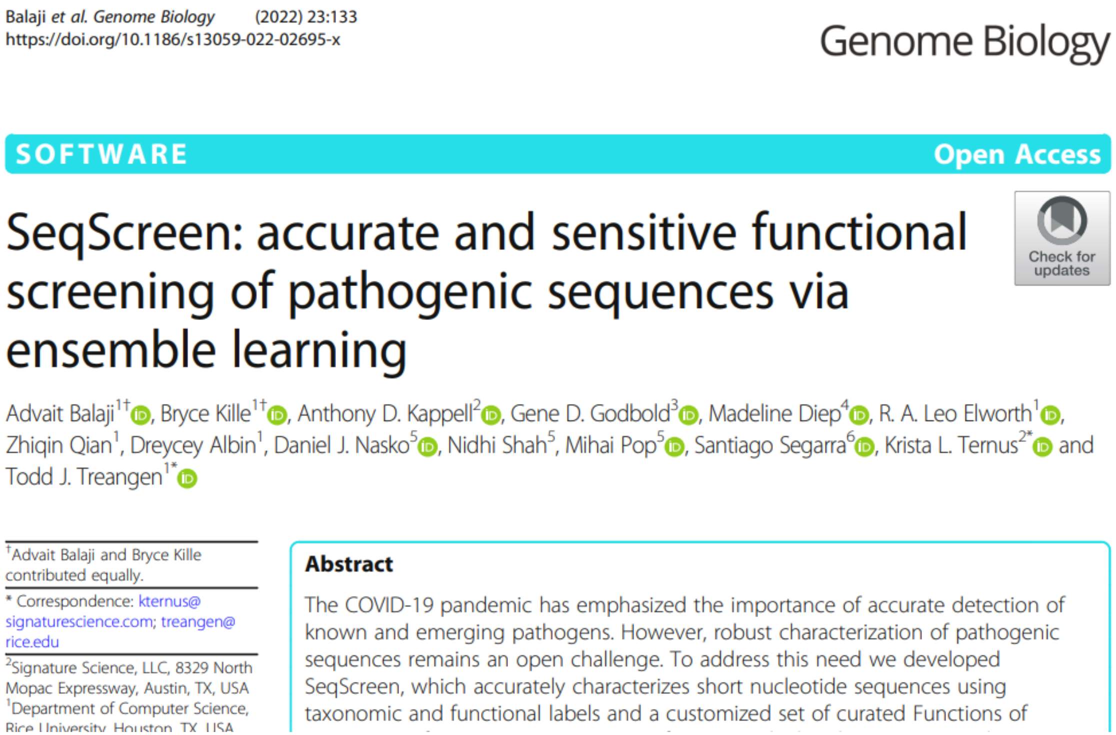
Alignment to Biothreats

Krista Ternus, PhD, PMP

Signature Science, LLC

SeqScreen Publications



Balaji *et al. Genome Biology* (2022) 23:133
https://doi.org/10.1186/s13059-022-02695-x

Genome Biology

SOFTWARE Open Access

SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning

Check for updates

Advait Balaji[1†], Bryce Kille[1†], Anthony D. Kappell[2], Gene D. Godbold[3], Madeline Diep[4], R. A. Leo Elworth[1], Zhiqin Qian[1], Dreycey Albin[1], Daniel J. Nasko[5], Nidhi Shah[5], Mihai Pop[5], Santiago Segarra[6], Krista L. Ternus[2*] and Todd J. Treangen[1*]

Abstract

The COVID-19 pandemic has emphasized the importance of accurate detection of known and emerging pathogens. However, robust characterization of pathogenic sequences remains an open challenge. To address this need we developed SeqScreen, which accurately characterizes short nucleotide sequences using taxonomic and functional labels and a customized set of curated Functions of

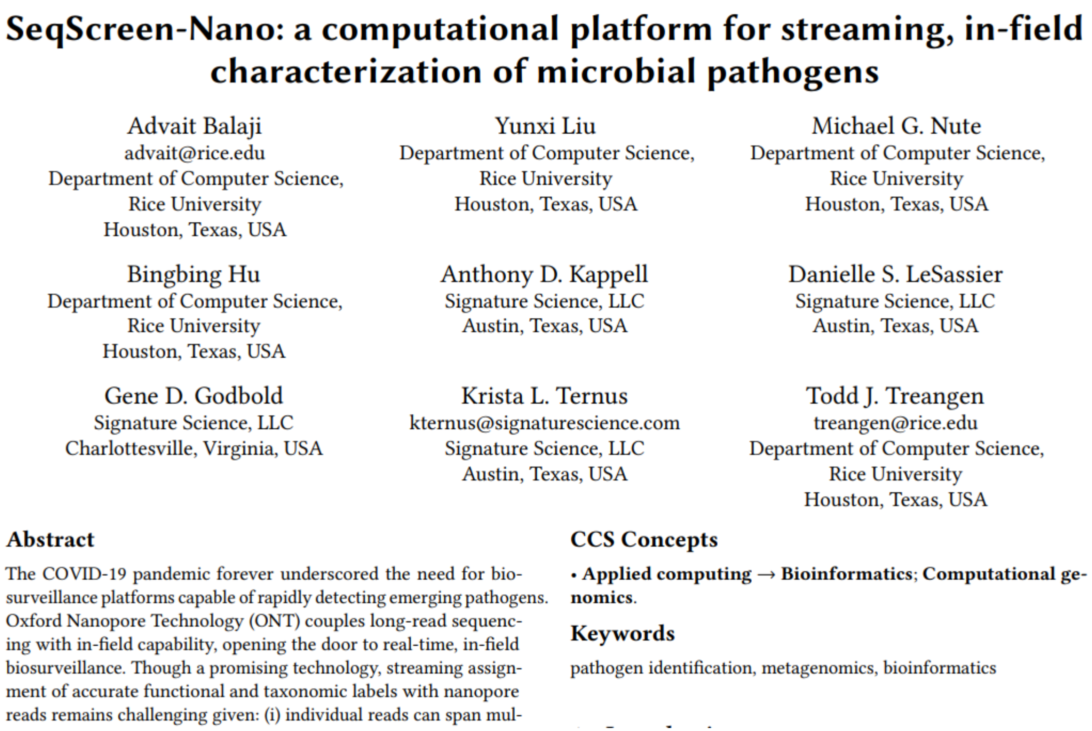
[†] Advait Balaji and Bryce Kille contributed equally.

[*] Correspondence: kternus@signaturescience.com; treangen@rice.edu

[2] Signature Science, LLC, 8329 North Mopac Expressway, Austin, TX, USA

[1] Department of Computer Science, Rice University, Houston, TX, USA

https://link.springer.com/article/10.1186/s13059-022-02695-x



SeqScreen-Nano: a computational platform for streaming, in-field characterization of microbial pathogens

Advait Balaji
advait@rice.edu
Department of Computer Science,
Rice University
Houston, Texas, USA

Yunxi Liu
Department of Computer Science,
Rice University
Houston, Texas, USA

Michael G. Nute
Department of Computer Science,
Rice University
Houston, Texas, USA

Bingbing Hu
Department of Computer Science,
Rice University
Houston, Texas, USA

Anthony D. Kappell
Signature Science, LLC
Austin, Texas, USA

Danielle S. LeSassier
Signature Science, LLC
Austin, Texas, USA

Gene D. Godbold
Signature Science, LLC
Charlottesville, Virginia, USA

Krista L. Ternus
kternus@signaturescience.com
Signature Science, LLC
Austin, Texas, USA

Todd J. Treangen
treangen@rice.edu
Department of Computer Science,
Rice University
Houston, Texas, USA

Abstract

The COVID-19 pandemic forever underscored the need for biosurveillance platforms capable of rapidly detecting emerging pathogens. Oxford Nanopore Technology (ONT) couples long-read sequencing with in-field capability, opening the door to real-time, in-field biosurveillance. Though a promising technology, streaming assignment of accurate functional and taxonomic labels with nanopore reads remains challenging given: (i) individual reads can span mul-

CCS Concepts

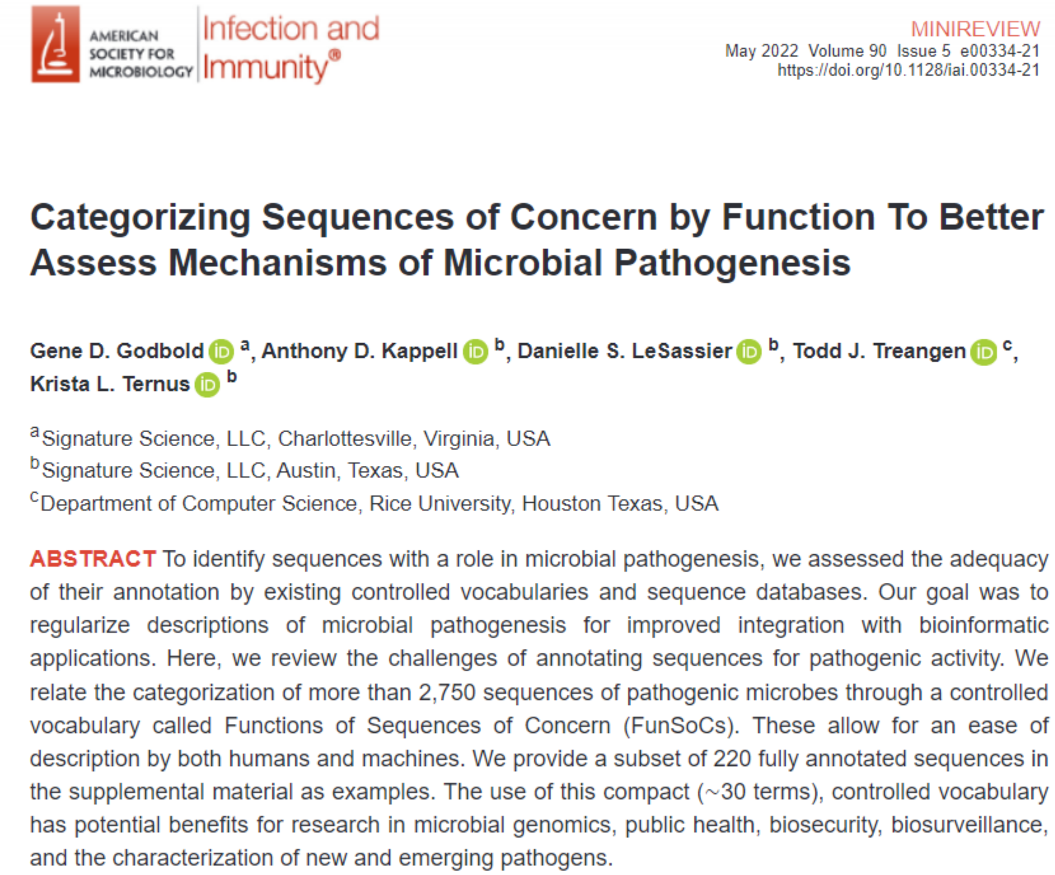
• **Applied computing** → **Bioinformatics**; **Computational genomics**.

Keywords

pathogen identification, metagenomics, bioinformatics

https://dl.acm.org/doi/pdf/10.1145/3584371.3612960

Publications about Sequences of Concern



AMERICAN SOCIETY FOR MICROBIOLOGY | Infection and Immunity®

MINIREVIEW
May 2022 Volume 90 Issue 5 e00334-21
https://doi.org/10.1128/iai.00334-21

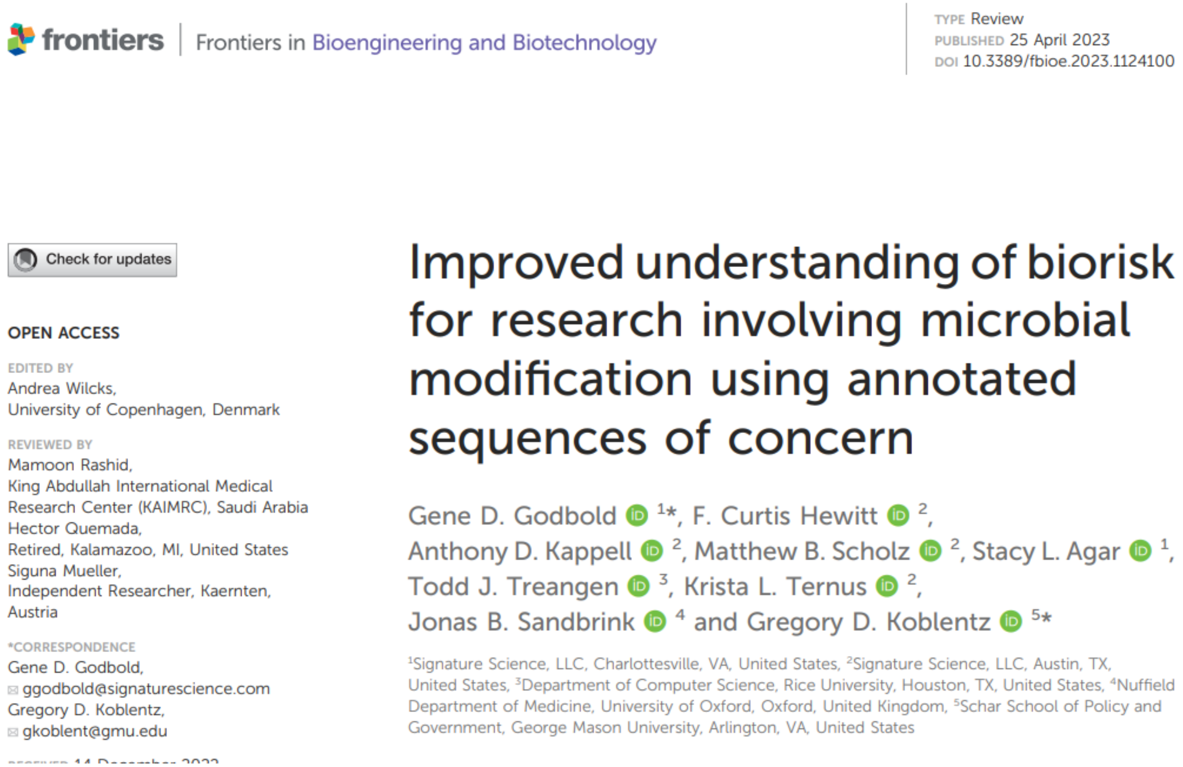
Categorizing Sequences of Concern by Function To Better Assess Mechanisms of Microbial Pathogenesis

Gene D. Godbold[a], Anthony D. Kappell[b], Danielle S. LeSassier[b], Todd J. Treangen[c], Krista L. Ternus[b]

[a]Signature Science, LLC, Charlottesville, Virginia, USA
[b]Signature Science, LLC, Austin, Texas, USA
[c]Department of Computer Science, Rice University, Houston Texas, USA

ABSTRACT To identify sequences with a role in microbial pathogenesis, we assessed the adequacy of their annotation by existing controlled vocabularies and sequence databases. Our goal was to regularize descriptions of microbial pathogenesis for improved integration with bioinformatic applications. Here, we review the challenges of annotating sequences for pathogenic activity. We relate the categorization of more than 2,750 sequences of pathogenic microbes through a controlled vocabulary called Functions of Sequences of Concern (FunSoCs). These allow for an ease of description by both humans and machines. We provide a subset of 220 fully annotated sequences in the supplemental material as examples. The use of this compact (~30 terms), controlled vocabulary has potential benefits for research in microbial genomics, public health, biosecurity, biosurveillance, and the characterization of new and emerging pathogens.

https://journals.asm.org/doi/epub/10.1128/iai.00334-21



frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Review
PUBLISHED 25 April 2023
DOI 10.3389/fbioe.2023.1124100

Check for updates

OPEN ACCESS

EDITED BY
Andrea Wilcks,
University of Copenhagen, Denmark

REVIEWED BY
Mamoon Rashid,
King Abdullah International Medical Research Center (KAIMRC), Saudi Arabia
Hector Quemada,
Retired, Kalamazoo, MI, United States
Siguna Mueller,
Independent Researcher, Kaernten, Austria

*CORRESPONDENCE
Gene D. Godbold,
✉ ggodbold@signaturescience.com
Gregory D. Koblentz,
✉ gkoblent@gmu.edu

RECEIVED 14 December 2022

Improved understanding of biorisk for research involving microbial modification using annotated sequences of concern

Gene D. Godbold[1*], F. Curtis Hewitt[2], Anthony D. Kappell[2], Matthew B. Scholz[2], Stacy L. Agar[1], Todd J. Treangen[3], Krista L. Ternus[2], Jonas B. Sandbrink[4] and Gregory D. Koblentz[5*]

[1]Signature Science, LLC, Charlottesville, VA, United States, [2]Signature Science, LLC, Austin, TX, United States, [3]Department of Computer Science, Rice University, Houston, TX, United States, [4]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, [5]Schar School of Policy and Government, George Mason University, Arlington, VA, United States

https://www.frontiersin.org/articles/10.3389/fbioe.2023.1124100/full

SeqScreen Installation

- SeqScreen documentation: https://gitlab.com/treangenlab/seqscreen/-/wikis/home
  - Runs in a Linux environment (needs up to 256 GB RAM, depending on use case)
  - We recommend installing SeqScreen via conda (or mamba)

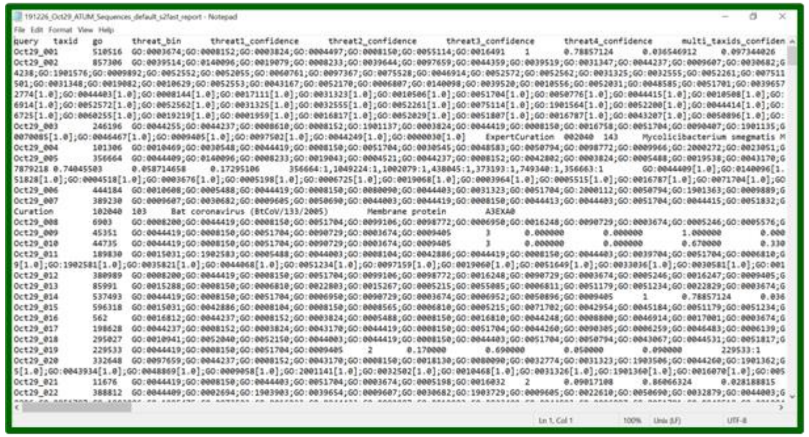
- If you are not able to use conda, you can also download SeqScreen to run via Singularity or Docker thanks to quay.io
  - See SeqScreen versions here: https://quay.io/repository/biocontainers/seqscreen?tab=tags

  -
    ```
    singularity pull docker://quay.io/biocontainers/seqscreen:4.2--hdfd78af_0
    ```

  -
    ```
    docker pull quay.io/biocontainers/seqscreen:4.2--hdfd78af_0
    ```

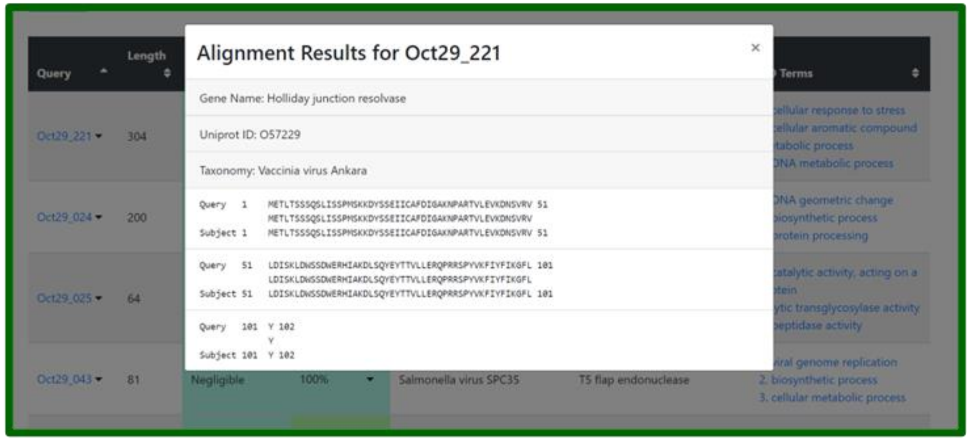
- SeqScreen database, version released in March 2023:
  - https://s3.wasabisys.com/seqscreenv4/SeqScreenDB_23.3.tar.gz
  - md5sum = `4f01938a1f8d1a61e52ef9165e737824`
  - Compressed database file is ~170 GB, uncompressed database and dependences are ~234 GB
  - Uncompress the database directory after download and leave subdirectory structure as is

SeqScreen Outputs



191226_Oct29_ATUM_Sequences_default_s2fast_report - Notepad

```
query   taxid   go      threat_bin      threat1_confidence      threat2_confidence      threat3_confidence      threat4_confidence      multi_taxids_confiden
Oct29_001       510516  GO:0003674;GO:0008152;GO:0003824;GO:0004497;GO:0008150;GO:0055114;GO:0016491    1       0.78857124      0.036546912     0.097344026
Oct29_002       857306  GO:0039514;GO:0140096;GO:0019079;GO:0008233;GO:0039644;GO:0097659;GO:0044359;GO:0039519;GO:0031347;GO:0044237;GO:0009607;GO:0030682;G
4238;GO:1901576;GO:0009892;GO:0052552;GO:0052055;GO:0060761;GO:0097367;GO:0075528;GO:0046914;GO:0052572;GO:0052562;GO:0031325;GO:0032555;GO:0052261;GO:007511
501;GO:0031348;GO:0019082;GO:0010629;GO:0052553;GO:0043167;GO:0052170;GO:0006807;GO:0140098;GO:0039520;GO:0010556;GO:0052031;GO:0048585;GO:0051701;GO:0039657
2774[1.0];GO:0044403[1.0];GO:0008144[1.0];GO:0017111[1.0];GO:0031323[1.0];GO:0010506[1.0];GO:0051704[1.0];GO:0050776[1.0];GO:0044415[1.0];GO:0010508[1.0];GO:
6914[1.0];GO:0052572[1.0];GO:0052562[1.0];GO:0031325[1.0];GO:0032555[1.0];GO:0052261[1.0];GO:0075114[1.0];GO:1901564[1.0];GO:0052200[1.0];GO:0044414[1.0];GO:
6725[1.0];GO:0060255[1.0];GO:0019219[1.0];GO:0001959[1.0];GO:0016817[1.0];GO:0052029[1.0];GO:0051807[1.0];GO:0016787[1.0];GO:0043207[1.0];GO:0050896[1.0];GO:
Oct29_003       246196  GO:0044255;GO:0044237;GO:0008610;GO:0008152;GO:1901137;GO:0003824;GO:0044419;GO:0008150;GO:0016758;GO:0051704;GO:0090407;GO:1901135;G
0070085[1.0];GO:0046467[1.0];GO:0009405[1.0];GO:0097502[1.0];GO:0044249[1.0];GO:0000030[1.0]    ExpertCuration  002040  143     Mycolicibacterium smegmatis M
Oct29_004       101306  GO:0010469;GO:0030548;GO:0044419;GO:0008150;GO:0051704;GO:0030545;GO:0048583;GO:0050794;GO:0098772;GO:0009966;GO:2000272;GO:0023051;G
Oct29_005       356664  GO:0044409;GO:0140096;GO:0008233;GO:0019043;GO:0004521;GO:0044237;GO:0008152;GO:0042802;GO:0003824;GO:0005488;GO:0019538;GO:0043170;G
7879218 0.74045503      0.058714658     0.17295106      356664:1,1049224:1,1002079:1,438045:1,373193:1,749340:1,356663:1        GO:0044409[1.0];GO:0140096[1.
51828[1.0];GO:0004518[1.0];GO:0003676[1.0];GO:0005198[1.0];GO:0006725[1.0];GO:0019068[1.0];GO:0003964[1.0];GO:0005515[1.0];GO:0016787[1.0];GO:0071704[1.0];GO
Oct29_006       444184  GO:0010608;GO:0005488;GO:0044419;GO:0008150;GO:0080090;GO:0044403;GO:0031323;GO:0051704;GO:2000112;GO:0050794;GO:1901363;GO:0009889;G
Oct29_007       389230  GO:0009607;GO:0030682;GO:0009605;GO:0050690;GO:0044003;GO:0044419;GO:0008150;GO:0044413;GO:0044403;GO:0051704;GO:0044415;GO:0051832;G
Curation        102040  103     Bat coronavirus (BtCoV/133/2005)        Membrane protein        A3EXA0
Oct29_008       6903    GO:0008200;GO:0044419;GO:0008150;GO:0051704;GO:0099106;GO:0098772;GO:0006950;GO:0016248;GO:0090729;GO:0003674;GO:0005246;GO:0005576;G
Oct29_009       45351   GO:0044419;GO:0008150;GO:0051704;GO:0090729;GO:0003674;GO:0009405       3       0.000000        0.000000        1.000000        0.000
Oct29_010       44735   GO:0044419;GO:0008150;GO:0051704;GO:0090729;GO:0003674;GO:0009405       3       0.000000        0.000000        0.670000        0.330
Oct29_011       189830  GO:0015031;GO:1902583;GO:0005488;GO:0044003;GO:0008104;GO:0042886;GO:0044419;GO:0008150;GO:0044403;GO:0039704;GO:0051704;GO:0006810;G
9[1.0];GO:1902581[1.0];GO:0035821[1.0];GO:0044068[1.0];GO:0051234[1.0];GO:0097159[1.0];GO:0019060[1.0];GO:0051649[1.0];GO:0033036[1.0];GO:0030581[1.0];GO:001
Oct29_012       380989  GO:0008200;GO:0044419;GO:0008150;GO:0051704;GO:0099106;GO:0098772;GO:0016248;GO:0090729;GO:0003674;GO:0005246;GO:0016247;GO:0009405;G
Oct29_013       85991   GO:0015288;GO:0008150;GO:0006810;GO:0022803;GO:0015267;GO:0005215;GO:0055085;GO:0006811;GO:0051179;GO:0051234;GO:0022829;GO:0003674;G
Oct29_014       537493  GO:0044419;GO:0008150;GO:0051704;GO:0006950;GO:0090729;GO:0003674;GO:0006952;GO:0050896;GO:0009405       1       0.78857124      0.036
Oct29_015       596318  GO:0015031;GO:0042886;GO:0008104;GO:0008150;GO:0008565;GO:0006810;GO:0005215;GO:0071702;GO:0042954;GO:0045184;GO:0051179;GO:0051234;G
Oct29_016       562     GO:0016812;GO:0044237;GO:0008152;GO:0003824;GO:0005488;GO:0008150;GO:0016810;GO:0044248;GO:0008800;GO:0046914;GO:0017001;GO:0003674;G
Oct29_017       198628  GO:0044237;GO:0008152;GO:0003824;GO:0043170;GO:0044419;GO:0008150;GO:0051704;GO:0044260;GO:0090305;GO:0006259;GO:0046483;GO:0006139;G
Oct29_018       295027  GO:0010941;GO:0052040;GO:0052150;GO:0044003;GO:0044419;GO:0008150;GO:0044403;GO:0051704;GO:0050794;GO:0043067;GO:0044531;GO:0051817;G
Oct29_019       229533  GO:0044419;GO:0008150;GO:0051704;GO:0009405       2       0.170000        0.690000        0.050000        0.090000        229533:1
Oct29_020       332648  GO:0097659;GO:0044237;GO:0008152;GO:0043170;GO:0008150;GO:0018130;GO:0080090;GO:0032774;GO:0031323;GO:1903506;GO:0044260;GO:1901362;G
5[1.0];GO:0043934[1.0];GO:0048869[1.0];GO:0009058[1.0];GO:2001141[1.0];GO:0032502[1.0];GO:0010468[1.0];GO:0031326[1.0];GO:1901360[1.0];GO:0016070[1.0];GO:005
Oct29_021       11676   GO:0044419;GO:0008150;GO:0044403;GO:0051704;GO:0003674;GO:0005198;GO:0016032    2       0.09017108      0.86066324      0.028188815
Oct29_022       388812  GO:0044409;GO:0002694;GO:1903903;GO:0039654;GO:0009607;GO:0030682;GO:1903729;GO:0009605;GO:0022610;GO:0050690;GO:0032879;GO:0044003;G
```



Alignment Results for Oct29_221

Gene Name: Holliday junction resolvase

Uniprot ID: O57229

Taxonomy: Vaccinia virus Ankara

```
Query   1    METLTSSSQSLISSPMSKKDYSSEIICAFDIGAKNPARTVLEVKDNSVRV 51
             METLTSSSQSLISSPMSKKDYSSEIICAFDIGAKNPARTVLEVKDNSVRV
Subject 1    METLTSSSQSLISSPMSKKDYSSEIICAFDIGAKNPARTVLEVKDNSVRV 51

Query   51   LDISKLDWSSDWERHIAKDLSQYEYTTVLLERQPRRSPYVKFIYFIKGFL 101
             LDISKLDWSSDWERHIAKDLSQYEYTTVLLERQPRRSPYVKFIYFIKGFL
Subject 51   LDISKLDWSSDWERHIAKDLSQYEYTTVLLERQPRRSPYVKFIYFIKGFL 101

Query   101  Y 102
             Y
Subject 101  Y 102
```

Oct29_221 304
cellular response to stress
cellular aromatic compound
tabolic process
DNA metabolic process

Oct29_024 200
DNA geometric change
biosynthetic process
protein processing

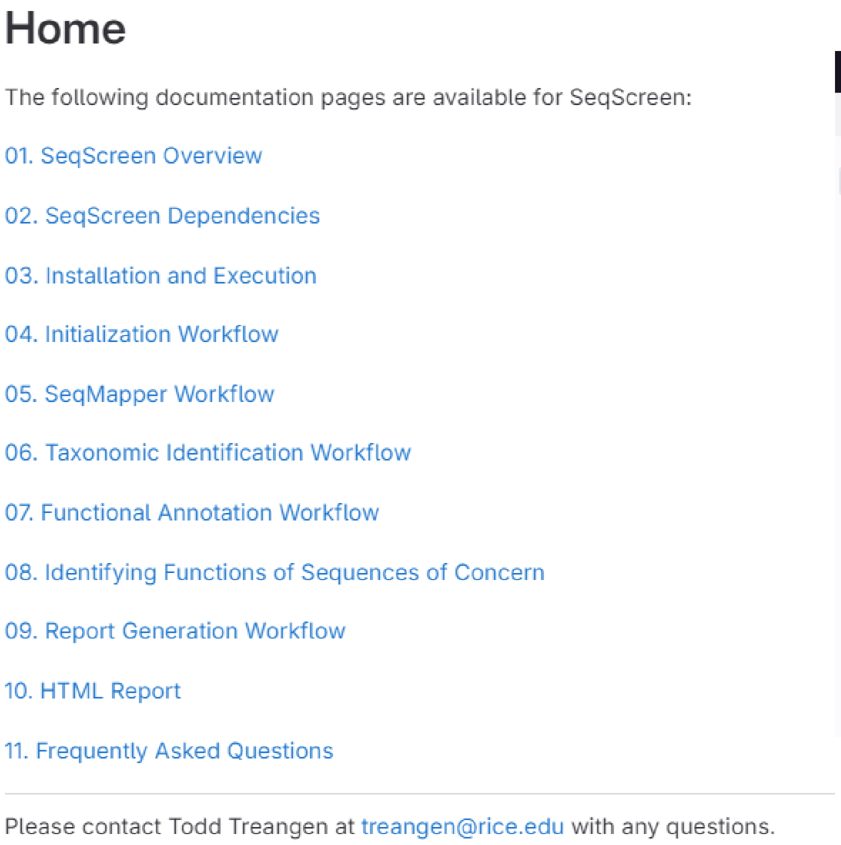
Oct29_025 64
catalytic activity, acting on a
tein
ytic transglycosylase activity
peptidase activity

Oct29_043 81 Negligible 100% Salmonella virus SPC35 T5 flap endonuclease
viral genome replication
2. biosynthetic process
3. cellular metabolic process

Computer-friendly text file format that enables easy downstream automated parsing

Human-friendly HTML report that allows end users to interactively explore the results

SeqScreen Code and Documentation on GitLab



Home

The following documentation pages are available for SeqScreen:

01. SeqScreen Overview

02. SeqScreen Dependencies

03. Installation and Execution

04. Initialization Workflow

05. SeqMapper Workflow

06. Taxonomic Identification Workflow

07. Functional Annotation Workflow

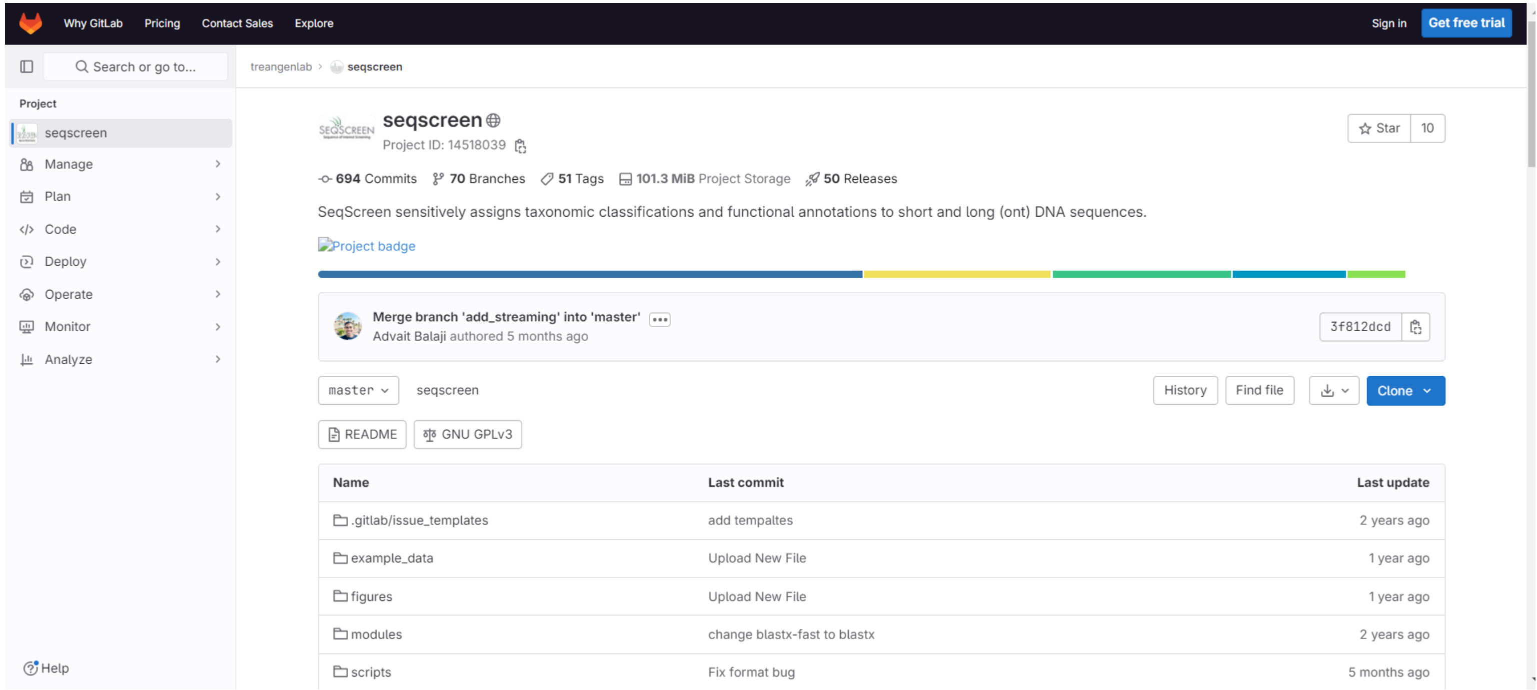
08. Identifying Functions of Sequences of Concern

09. Report Generation Workflow

10. HTML Report

11. Frequently Asked Questions

Please contact Todd Treangen at treangen@rice.edu with any questions.



seqscreen

Project ID: 14518039

694 Commits 70 Branches 51 Tags 101.3 MiB Project Storage 50 Releases

SeqScreen sensitively assigns taxonomic classifications and functional annotations to short and long (ont) DNA sequences.

Project badge

Merge branch 'add_streaming' into 'master'

Advait Balaji authored 5 months ago

3f812dcd

Name	Last commit	Last update
.gitlab/issue_templates	add tempaltes	2 years ago
example_data	Upload New File	1 year ago
figures	Upload New File	1 year ago
modules	change blastx-fast to blastx	2 years ago
scripts	Fix format bug	5 months ago

https://gitlab.com/treangenlab/seqscreen

https://gitlab.com/treangenlab/seqscreen/-/wikis/Home

Single Gene Sequence Outputs

- Input file = `single_gene_sequences/single_gene_seqs.fasta`
- Output files = `single_gene_sequences/single_gene_seqs_precomputed_results/`
  - `final_tsv_report_single_gene_seqs.xlsx`
  - `report_generation.zip`
  - `seqscreen_command.txt`
  - `single_gene_sequences_key.xlsx`

The *seqscreen_report_pathgo.tsv output file content was copied and pasted into an Excel file for ease of interpretation

```
single_gene_seqs_sensitive/report_generation/
single_gene_seqs_seqscreen_report_pathgo.tsv
```

RICE

Aclid

signature®
science LLC