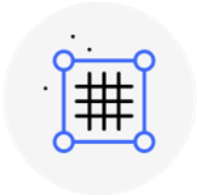
DNA Screening Basics

Kevin Flyangolts

kevin@aclid.bio

aclid



Secure

DNA/RNA and Oligo screening.

Real-time pathogen and toxin screening integrated with ordering system or embedded in a bench-top synthesizer.

Surveil

Biological threat detection.

In-field detection and identification of engineered or emerging natural threats in complex samples from austere environments.

Safe

Research and bioproduct safety.

Environmental, health, and safety screening for synthetic biology including pathogen, toxin, allergen, and antimicrobial resistance.

RICE aclid signature science LLC

Biological Risk

- RED
  - Sequences derived from Variola major or Variola minor genomes (smallpox)
  - Sequences coding functional forms of toxins or subunits
  - Sequences capable of transferring or endowing pathogenicity
- YELLOW
  - Sequences that aren't known to transfer or endow pathogenicity and are not typically regarded as housekeeping or metabolic genes
- GREEN
  - Sequences derived from organisms or viruses that do not pose a high safety or security concern
  - Sequences that are typically regarded as housekeeping or metabolic genes

Regulated Sequences

- Australia Group
  - An informal global forum of countries ensuring exports do not contribute proliferation of chemical or biological weapons
- Export Controls
  - National controls on exports of materials and goods unique to every country
- US Federal Select Agent Program
  - US control on possession and transfer of highly dangerous pathogens and toxins
- US Department of Health and Human Services Screening Framework Guidance
  - US guidance to industry on secure, safe, and responsible synthesis of genes and oligos

Scientific Review

- When a sequence is regulated or risky, further investigation is warranted
  - Identify the purpose of the sequence
  - Determine whether the sequence aligns with its intended use
  - Determine whether the sequence and its intended use are scientifically sound

- Documenting can help with future assessments and provide more standardization to the process

Similarity Search, Aligners, HMMs, and Metagenomics

- No consensus on best tool or pipeline
  - Tradeoffs on computational requirements, speed, sensitivity, and specificity
- Query coverage and percent identity
  - No standard thresholds
  - A coarse configuration for detecting homologs could be 50% for each
- E-value and bitscore
  - Significance thresholds vary by sequence lengths, database, and use case
  - Generally, hits with e-values greater than 1 are considered lower quality
- An iterative approach with a good test set leads to the best results

Effects of Sensitivity & Specificity on False Positives & Negatives

- Sensitivity and specificity are inversely related
  - **Higher specificity** may lead to lower sensitivity and **more false-negatives**
  - **Higher sensitivity** may lead to lower specificity and **more false-positives**

- **False-positives** increase **burden** by requiring a **follow-up screen / review**

- **False-negatives** increase **liability** and **risk** of misuse and non-compliance

False Positives

Near-Neighbor & Remote Similarity

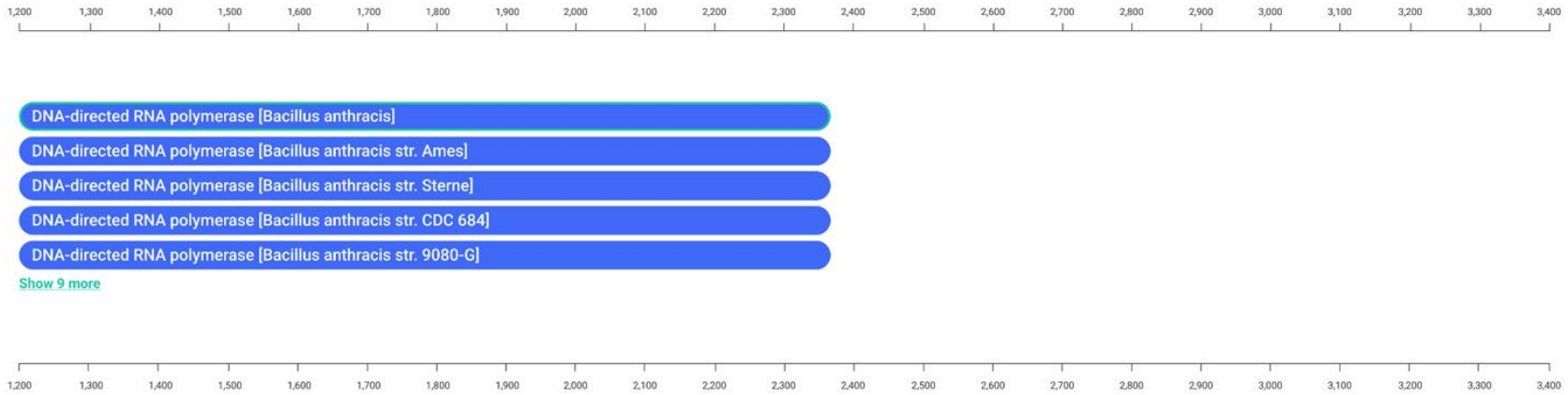
- Short segments (e.g., 20 bp)
- High sensitivity / low thresholds for significance or identity
- Repeats (e.g., tandem, interspersed)
- Must strike the right balance between sensitivity and specificity

Housekeeping Gene

- Structural, metabolic, expression, etc.
- Conserved across pathogenic and non-pathogenic organisms
- Functional annotation helps reduce false-positives

* Occasionally mislabeled reference sequences can cause false flags (e.g., eGFP fused with Ebola virus glycoprotein)

RICE aclid signature science LLC



ORGANISM: Bacillus anthracis

GENE: DNA-directed RNA polymerase

THREAT LEVEL: Not Controlled

% IDENTITY: 100.00

% COVERAGE (QUERY): 96.83

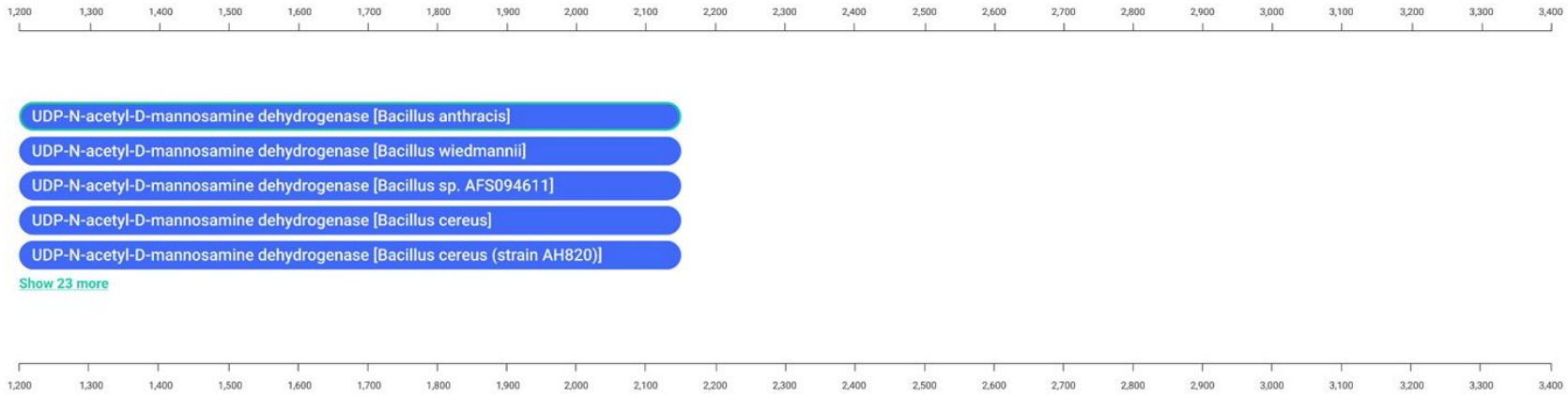
% COVERAGE (SUBJECT): 100.00

LOCATION (QUERY): 1,200 - 2,350

LOCATION (SUBJECT): 1 - 1,179

FUNCTION: DNA-dependent RNA polymerase (RNAP) catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates.

ONTOLOGY: DNA binding | DNA-directed 5'-3' RNA polymerase activity | DNA-templated transcription | ribonucleoside binding | transferase activity | nucleotidyltransferase activity | DNA-directed RNA polymerase complex



ORGANISM: Bacillus anthracis

GENE: UDP-N-acetyl-D-mannosamine dehydrogenase

THREAT LEVEL: Not Controlled

% IDENTITY: 100.00

% COVERAGE (QUERY): 93.33

% COVERAGE (SUBJECT): 67.22

LOCATION (QUERY): 1,200 - 2,150

LOCATION (SUBJECT): 1 - 973

FUNCTION: Catalyzes the four-electron oxidation of UDP-N-acetyl-D-mannosamine (UDP-ManNAc), reducing NAD+ and releasing UDP-N-acetylmannosaminuronic acid (UDP-ManNAcA).

ONTOLOGY: NAD binding | oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor | polysaccharide biosynthetic process

False Negatives

Missing or Incorrect Alignment

- Many bioinformatics tools rely on heuristics for performance
  - Large reference databases can generate spurious alignments
- New research may reveal gaps
- Periodic database updates and select-agent specific reference databases can help

Sequence Obfuscation

- Artificial sequences may look different enough from their original forms to avoid detection
- Identifying select agent signatures and splitting sequences into shorter segments can increase sensitivity

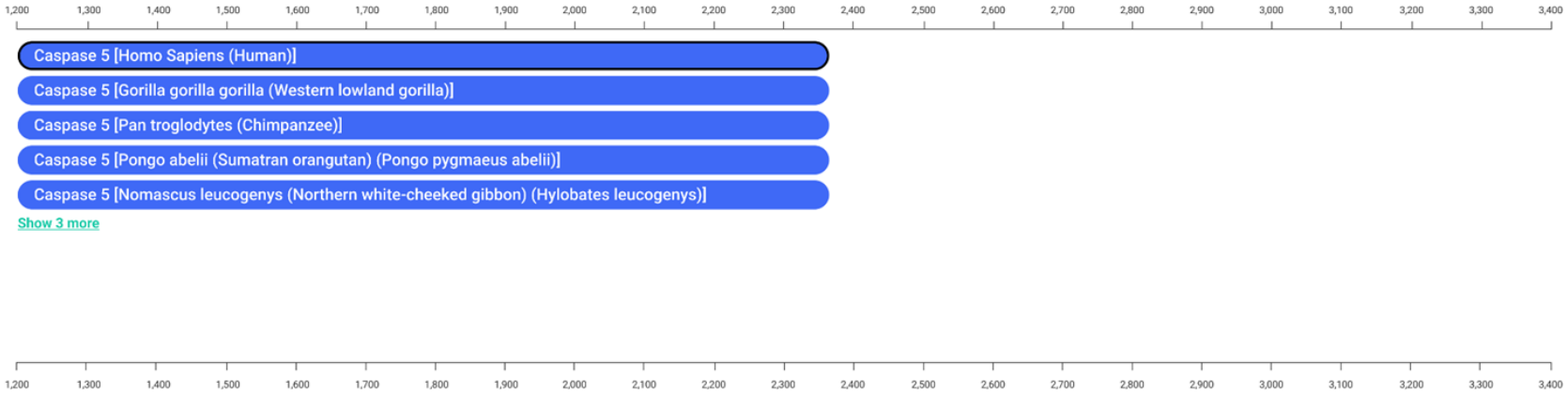
Sequence Obfuscation

Sequence modification

- Codon optimization
- Modified genetic codes
- Fusion proteins
- Low similarity homologs

Sequence recombination

- Engineered plasmids
- Sequence scrambling
- Oligo pool assembly



ORGANISM: Homo sapiens (Human)

GENE: Caspase 5

THREAT LEVEL: Not Controlled

% IDENTITY: 100.00

% COVERAGE (QUERY): 72.17

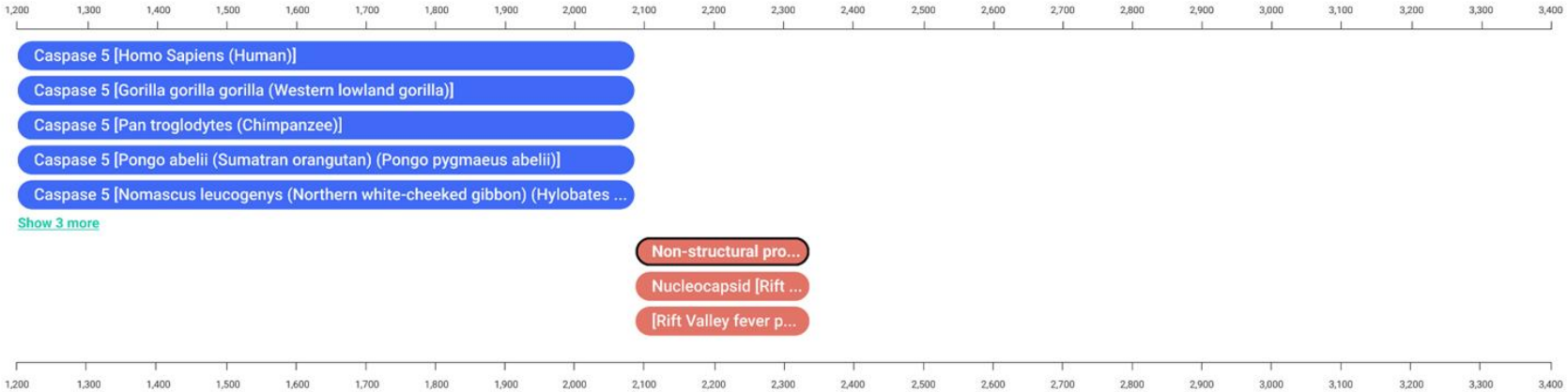
% COVERAGE (SUBJECT): 61.76

LOCATION (QUERY): 1,200 - 2,350

LOCATION (SUBJECT): 9 - 983

FUNCTION: Thiol protease that acts as a mediator of programmed cell death. Initiates pyroptosis, a programmed lytic cell death pathway through cleavage of Gasdermin-D (GSDMD): cleavage releases the N-terminal gasdermin moiety (Gasdermin-D, N-terminal) that binds to membranes and forms pores, triggering pyroptosis. During non-canonical inflammasome activation, cuts CGAS and may play a role in the regulation of antiviral innate immune activation.

ONTOLOGY: nucleus | cytoplasm | protein maturation | neurotrophin TRK receptor signaling pathway | peptidase activity | protein catabolic process | regulation of macroautophagy | cysteine-type endopeptidase activity involved in apoptotic signaling pathway | cysteine-type endopeptidase activity involved in apoptotic process



ORGANISM: Rift valley fever virus (strain ZH-548 M12) (RVFV)

GENE: Non-structural protein

THREAT LEVEL: ⊗ Controlled

% IDENTITY: 100.00

% COVERAGE (QUERY): 96.83

% COVERAGE (SUBJECT): 33.32

LOCATION (QUERY): 2,100 - 2,350

LOCATION (SUBJECT): 28 - 353

FUNCTION: Plays a role in the escape of host innate immune response by promoting the degradation of host EIF2AK2/PKR and inhibiting host transcription. Cytoplasmic NSs interacts with host FBXW11 to degrade PKR whereas nuclear pool binds to host FBXO3 to target TFIIH subunit GTF2H1 for proteasomal degradation with the help of the linker protein SKP1. Removes FBXO3 isoform 1 from the nucleus. Forms nuclear amyloid-like filaments of about 12 nm in width that may sequester NSs binding partners, causing cell cycle arrest. Also aggragates in the cytosol as short fibrils late after host cell infection. Plays a role in cell morphology and/or motility, reduction of lamellipodia, cell spreading, and dissolution of adherens junctions.

VIRULENCE FACTORS: host transcription | viral counter signaling

ONTOLOGY: host cell nucleus | suppression by virus of host transcription initiation from RNA polymerase II promoter | protein serine/threonine kinase inhibitor activity | suppression by virus of host type I interferon-mediated signaling pathway | suppression by virus of host PKR signaling | negative stranded viral RNA replication | suppression by virus of host viral-induced cytoplasmic pattern recognition receptor signaling pathway via inhibition of host RIG-I activity | suppression by virus of host type I interferon production | host cell cytoplasm