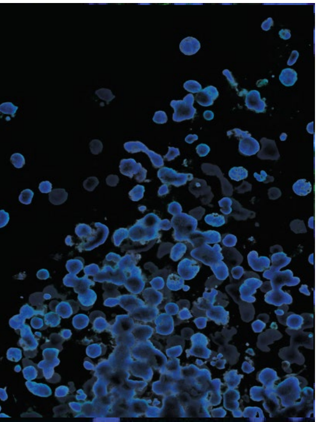


Genomic Data Quality

Connecting the Dots Between Bioinformatics and Physical Materials

Jonathan Jacobs, PhD
Senior Director, Bioinformatics
Sequencing & Bioinformatics Center
ATCC

Credible Leads to Incredible™

About ATCC

- Founded in 1925
- 501(c)(3) not-for-profit organization
- World’s largest, most diverse biorepository
- Quality Accreditation by multiple industry standards
  - ISO 9001 Certified
  - ISO 13485 Certified
  - ISO/IEC 17025 Accredited
  - ISO 17034 Accredited
- Standards development partner with multiple industry working groups
  - ANSI Standards Working Groups
  - AOAC International Working Group
  - IMMSA/NIST Microbiome Standards

- Global supplier of authenticated cell lines, microorganisms, and molecular standards
- Sales and Distribution to 150+ countries
- Full Talented team of 500+ employees

Thousands of authenticated biomaterials

- 5,000+ cell lines & primary tissue
- 2,500+ viruses
- 9,500+ bacteria
- 38,000+ fungi and protists

Visit **atcc.org**

Genomics data quality

Connecting the dots between bioinformatics and physical materials

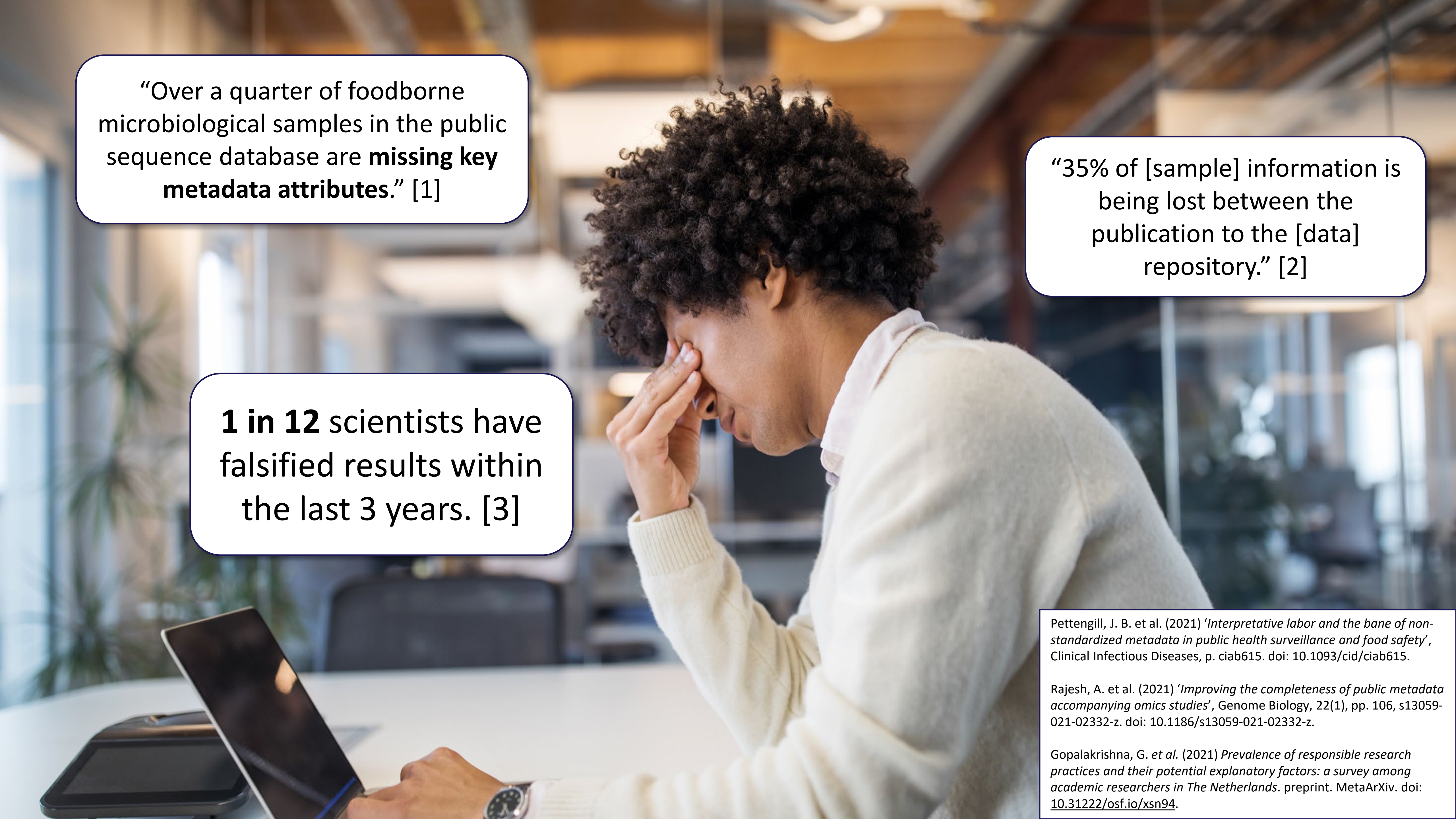
- Review challenges associated with genomics data quality and authenticity
- Discuss ***why*** ATCC is committed to **providing reference-quality genomes** for our materials
- Discuss our current efforts to produce standardized genomics reference data
- Explore the ATCC Genome Portal
- Explore the ATCC Cell Line Land

Challenges stemming from poor data quality…

“***Finding the right cell lines*** for my research is a challenge.”

“Many cell types are ***not good models*** for the disease I’m studying.”

“Pre-existing results are difficult to reproduce and often ***not reproducible***.”



“Over a quarter of foodborne microbiological samples in the public sequence database are **missing key metadata attributes**.” [1]

“35% of [sample] information is being lost between the publication to the [data] repository.” [2]

1 in 12 scientists have falsified results within the last 3 years. [3]

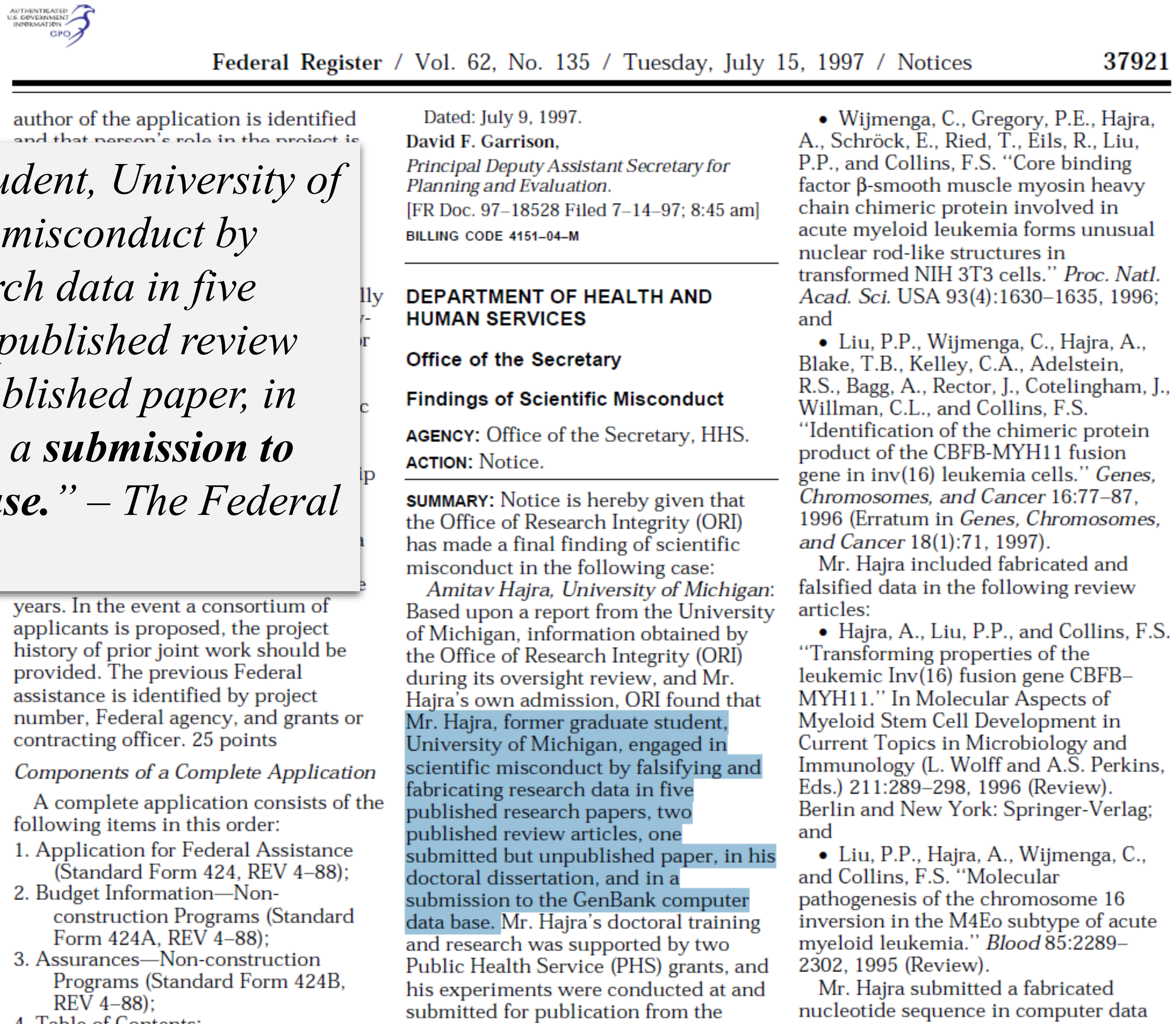
Pettengill, J. B. et al. (2021) ‘*Interpretative labor and the bane of non-standardized metadata in public health surveillance and food safety*’, Clinical Infectious Diseases, p. ciab615. doi: 10.1093/cid/ciab615.

Rajesh, A. et al. (2021) ‘*Improving the completeness of public metadata accompanying omics studies*’, Genome Biology, 22(1), pp. 106, s13059-021-02332-z. doi: 10.1186/s13059-021-02332-z.

Gopalakrishna, G. *et al.* (2021) *Prevalence of responsible research practices and their potential explanatory factors: a survey among academic researchers in The Netherlands*. preprint. MetaArXiv. doi: 10.31222/osf.io/xsn94.

Fake data was first discovered in GenBank in 1997

> *"Mr. Hajra, former graduate student, University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a* ***submission to the GenBank computer data base.****" – The Federal Register, v62, n135 (****1997****)*



AUTHENTICATED U.S. GOVERNMENT INFORMATION GPO

Federal Register / Vol. 62, No. 135 / Tuesday, July 15, 1997 / Notices 37921

author of the application is identified and that person's role in the project is

years. In the event a consortium of applicants is proposed, the project history of prior joint work should be provided. The previous Federal assistance is identified by project number, Federal agency, and grants or contracting officer. 25 points

Components of a Complete Application

A complete application consists of the following items in this order:

1. Application for Federal Assistance (Standard Form 424, REV 4–88);
2. Budget Information—Non-construction Programs (Standard Form 424A, REV 4–88);
3. Assurances—Non-construction Programs (Standard Form 424B, REV 4–88);
4. Table of Contents;

Dated: July 9, 1997.

David F. Garrison,

Principal Deputy Assistant Secretary for Planning and Evaluation.

[FR Doc. 97–18528 Filed 7–14–97; 8:45 am]

BILLING CODE 4151–04–M

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Office of the Secretary

Findings of Scientific Misconduct

AGENCY: Office of the Secretary, HHS.

ACTION: Notice.

SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has made a final finding of scientific misconduct in the following case:

Amitav Hajra, University of Michigan: Based upon a report from the University of Michigan, information obtained by the Office of Research Integrity (ORI) during its oversight review, and Mr. Hajra's own admission, ORI found that Mr. Hajra, former graduate student, University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a submission to the GenBank computer data base. Mr. Hajra's doctoral training and research was supported by two Public Health Service (PHS) grants, and his experiments were conducted at and submitted for publication from the

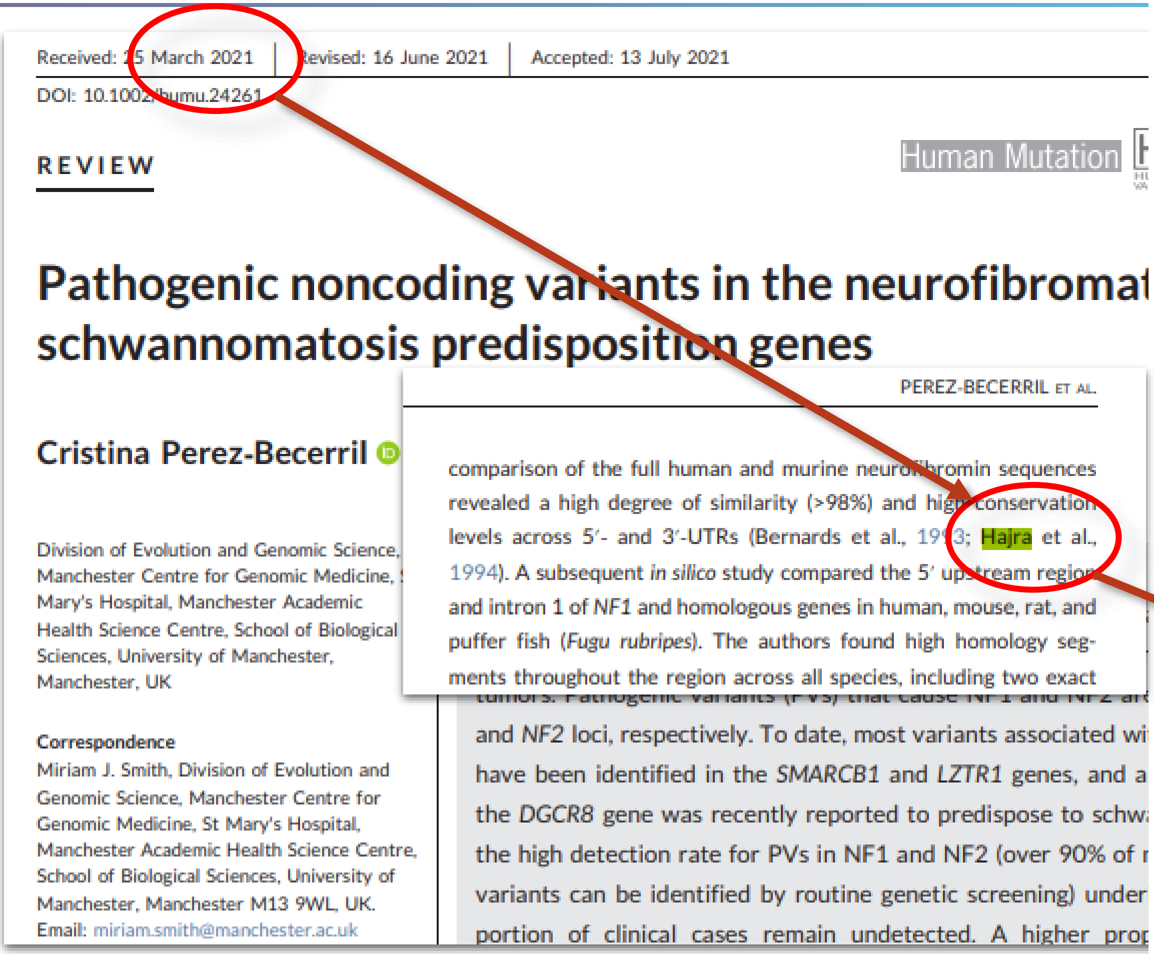
- Wijmenga, C., Gregory, P.E., Hajra, A., Schröck, E., Ried, T., Eils, R., Liu, P.P., and Collins, F.S. "Core binding factor β-smooth muscle myosin heavy chain chimeric protein involved in acute myeloid leukemia forms unusual nuclear rod-like structures in transformed NIH 3T3 cells." *Proc. Natl. Acad. Sci.* USA 93(4):1630–1635, 1996; and
- Liu, P.P., Wijmenga, C., Hajra, A., Blake, T.B., Kelley, C.A., Adelstein, R.S., Bagg, A., Rector, J., Cotelingham, J., Willman, C.L., and Collins, F.S. "Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells." *Genes, Chromosomes, and Cancer* 16:77–87, 1996 (Erratum in *Genes, Chromosomes, and Cancer* 18(1):71, 1997).

Mr. Hajra included fabricated and falsified data in the following review articles:

- Hajra, A., Liu, P.P., and Collins, F.S. "Transforming properties of the leukemic Inv(16) fusion gene CBFB–MYH11." In Molecular Aspects of Myeloid Stem Cell Development in Current Topics in Microbiology and Immunology (L. Wolff and A.S. Perkins, Eds.) 211:289–298, 1996 (Review). Berlin and New York: Springer-Verlag; and
- Liu, P.P., Hajra, A., Wijmenga, C., and Collins, F.S. "Molecular pathogenesis of the chromosome 16 inversion in the M4Eo subtype of acute myeloid leukemia." *Blood* 85:2289–2302, 1995 (Review).

Mr. Hajra submitted a fabricated nucleotide sequence in computer data

24 years later, this falsified data still being cited…



Received: 15 March 2021 | Revised: 16 June 2021 | Accepted: 13 July 2021

DOI: 10.1002/humu.24261

REVIEW

Human Mutation

Pathogenic noncoding variants in the neurofibromat schwannomatosis predisposition genes

Cristina Perez-Becerril

Division of Evolution and Genomic Science, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Academic Health Science Centre, School of Biological Sciences, University of Manchester, Manchester, UK

Correspondence
Miriam J. Smith, Division of Evolution and Genomic Science, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester Academic Health Science Centre, School of Biological Sciences, University of Manchester, Manchester, Manchester M13 9WL, UK.
Email: miriam.smith@manchester.ac.uk

tumors. Pathogenic variants (PVs) that cause NF1 and NF2 are

and NF2 loci, respectively. To date, most variants associated wi

have been identified in the *SMARCB1* and *LZTR1* genes, and a

the *DGCR8* gene was recently reported to predispose to schw

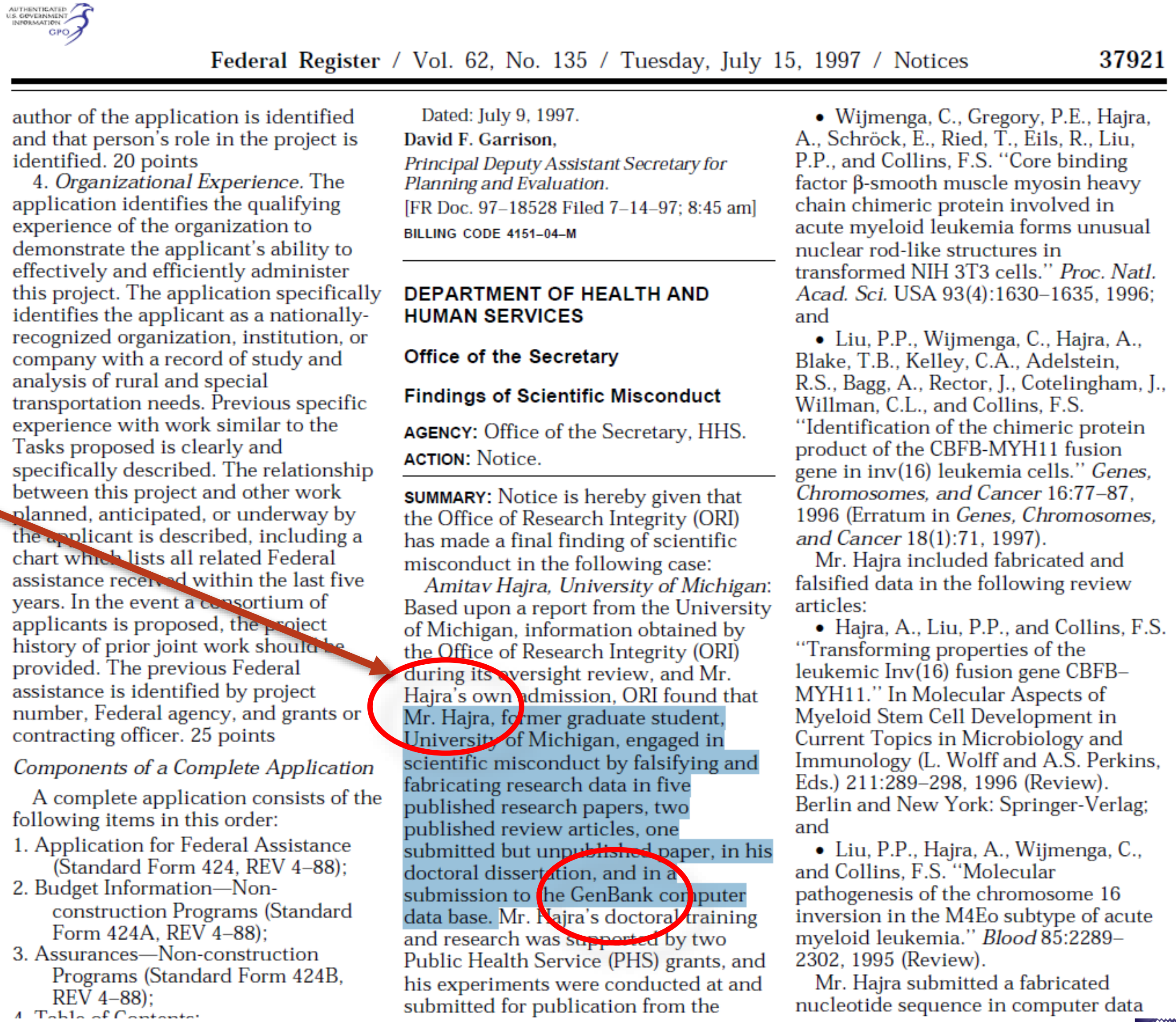
the high detection rate for PVs in NF1 and NF2 (over 90% of

variants can be identified by routine genetic screening) under

portion of clinical cases remain undetected. A higher prop

PEREZ-BECERRIL ET AL.

comparison of the full human and murine neurofibromin sequences revealed a high degree of similarity (>98%) and high conservation levels across 5′- and 3′-UTRs (Bernards et al., 1993; Hajra et al., 1994). A subsequent *in silico* study compared the 5′ upstream region and intron 1 of *NF1* and homologous genes in human, mouse, rat, and puffer fish (*Fugu rubripes*). The authors found high homology segments throughout the region across all species, including two exact



AUTHENTICATED U.S. GOVERNMENT INFORMATION GPO

author of the application is identified and that person's role in the project is identified. 20 points

4. *Organizational Experience.* The application identifies the qualifying experience of the organization to demonstrate the applicant's ability to effectively and efficiently administer this project. The application specifically identifies the applicant as a nationally-recognized organization, institution, or company with a record of study and analysis of rural and special transportation needs. Previous specific experience with work similar to the Tasks proposed is clearly and specifically described. The relationship between this project and other work planned, anticipated, or underway by the applicant is described, including a chart which lists all related Federal assistance received within the last five years. In the event a consortium of applicants is proposed, the project history of prior joint work should be provided. The previous Federal assistance is identified by project number, Federal agency, and grants or contracting officer. 25 points

Components of a Complete Application

A complete application consists of the following items in this order:

1. Application for Federal Assistance (Standard Form 424, REV 4–88);
2. Budget Information—Non-construction Programs (Standard Form 424A, REV 4–88);
3. Assurances—Non-construction Programs (Standard Form 424B, REV 4–88);
4. Table of Contents;

Dated: July 9, 1997.

David F. Garrison,

Principal Deputy Assistant Secretary for Planning and Evaluation.

[FR Doc. 97–18528 Filed 7–14–97; 8:45 am]

BILLING CODE 4151–04–M

DEPARTMENT OF HEALTH AND HUMAN SERVICES

Office of the Secretary

Findings of Scientific Misconduct

AGENCY: Office of the Secretary, HHS.

ACTION: Notice.

SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has made a final finding of scientific misconduct in the following case:

Amitav Hajra, University of Michigan: Based upon a report from the University of Michigan, information obtained by the Office of Research Integrity (ORI) during its oversight review, and Mr. Hajra's own admission, ORI found that Mr. Hajra, former graduate student, University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a submission to the GenBank computer data base. Mr. Hajra's doctoral training and research was supported by two Public Health Service (PHS) grants, and his experiments were conducted at and submitted for publication from the

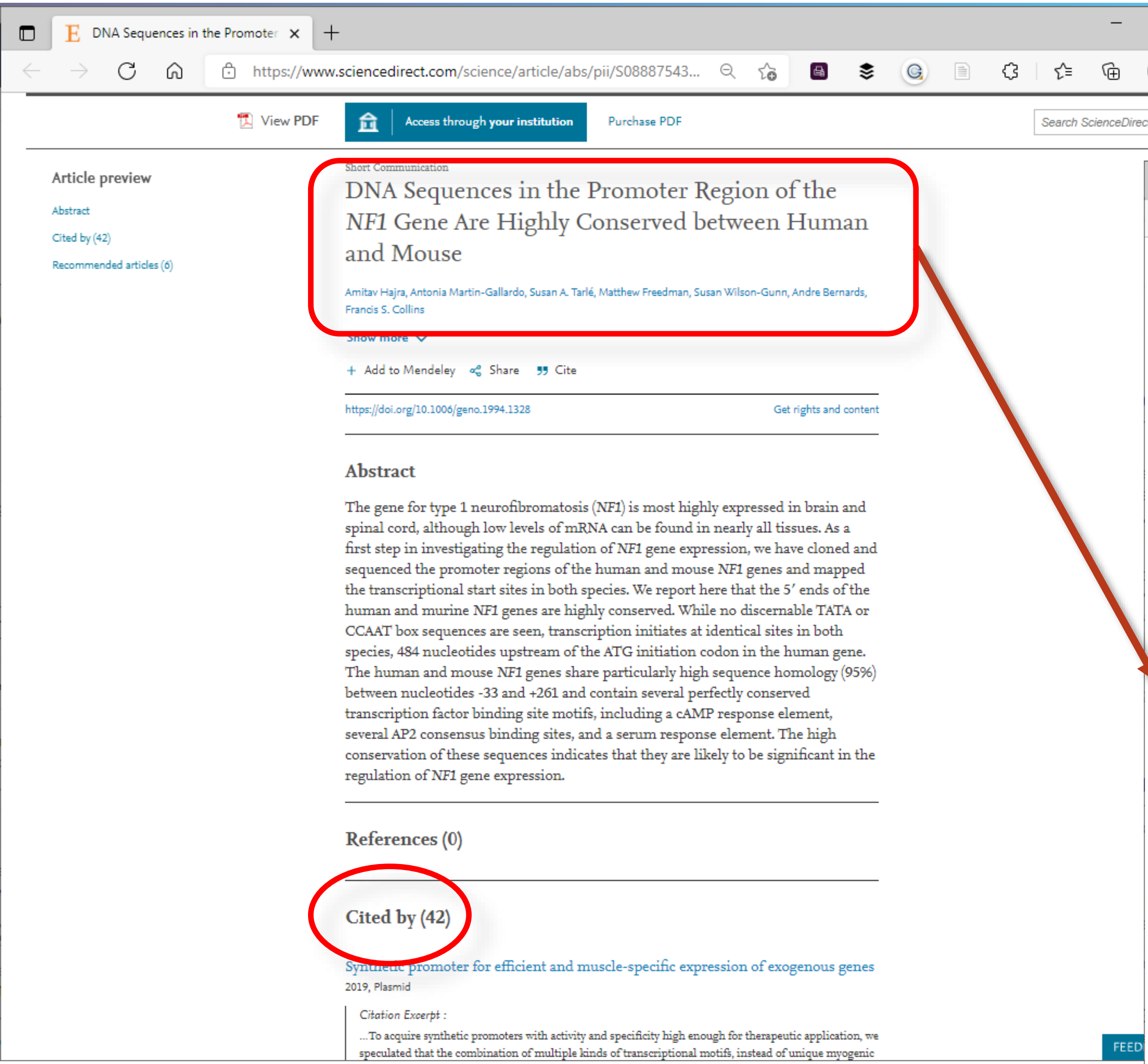
- Wijmenga, C., Gregory, P.E., Hajra, A., Schröck, E., Ried, T., Eils, R., Liu, P.P., and Collins, F.S. "Core binding factor β-smooth muscle myosin heavy chain chimeric protein involved in acute myeloid leukemia forms unusual nuclear rod-like structures in transformed NIH 3T3 cells." *Proc. Natl. Acad. Sci.* USA 93(4):1630–1635, 1996; and
- Liu, P.P., Wijmenga, C., Hajra, A., Blake, T.B., Kelley, C.A., Adelstein, R.S., Bagg, A., Rector, J., Cotelingham, J., Willman, C.L., and Collins, F.S. "Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells." *Genes, Chromosomes, and Cancer* 16:77–87, 1996 (Erratum in *Genes, Chromosomes, and Cancer* 18(1):71, 1997).

Mr. Hajra included fabricated and falsified data in the following review articles:

- Hajra, A., Liu, P.P., and Collins, F.S. "Transforming properties of the leukemic Inv(16) fusion gene CBFB–MYH11." In Molecular Aspects of Myeloid Stem Cell Development in Current Topics in Microbiology and Immunology (L. Wolff and A.S. Perkins, Eds.) 211:289–298, 1996 (Review). Berlin and New York: Springer-Verlag; and
- Liu, P.P., Hajra, A., Wijmenga, C., and Collins, F.S. "Molecular pathogenesis of the chromosome 16 inversion in the M4Eo subtype of acute myeloid leukemia." *Blood* 85:2289–2302, 1995 (Review).

Mr. Hajra submitted a fabricated nucleotide sequence in computer data

After 42 citations… the data is still in GenBank…



Short Communication

DNA Sequences in the Promoter Region of the *NF1* Gene Are Highly Conserved between Human and Mouse

Amitav Hajra, Antonia Martin-Gallardo, Susan A. Tarlé, Matthew Freedman, Susan Wilson-Gunn, Andre Bernards, Francis S. Collins

https://doi.org/10.1006/geno.1994.1328

Abstract

The gene for type 1 neurofibromatosis (*NF1*) is most highly expressed in brain and spinal cord, although low levels of mRNA can be found in nearly all tissues. As a first step in investigating the regulation of *NF1* gene expression, we have cloned and sequenced the promoter regions of the human and mouse *NF1* genes and mapped the transcriptional start sites in both species. We report here that the 5′ ends of the human and murine *NF1* genes are highly conserved. While no discernable TATA or CCAAT box sequences are seen, transcription initiates at identical sites in both species, 484 nucleotides upstream of the ATG initiation codon in the human gene. The human and mouse *NF1* genes share particularly high sequence homology (95%) between nucleotides -33 and +261 and contain several perfectly conserved transcription factor binding site motifs, including a cAMP response element, several AP2 consensus binding sites, and a serum response element. The high conservation of these sequences indicates that they are likely to be significant in the regulation of *NF1* gene expression.

References (0)

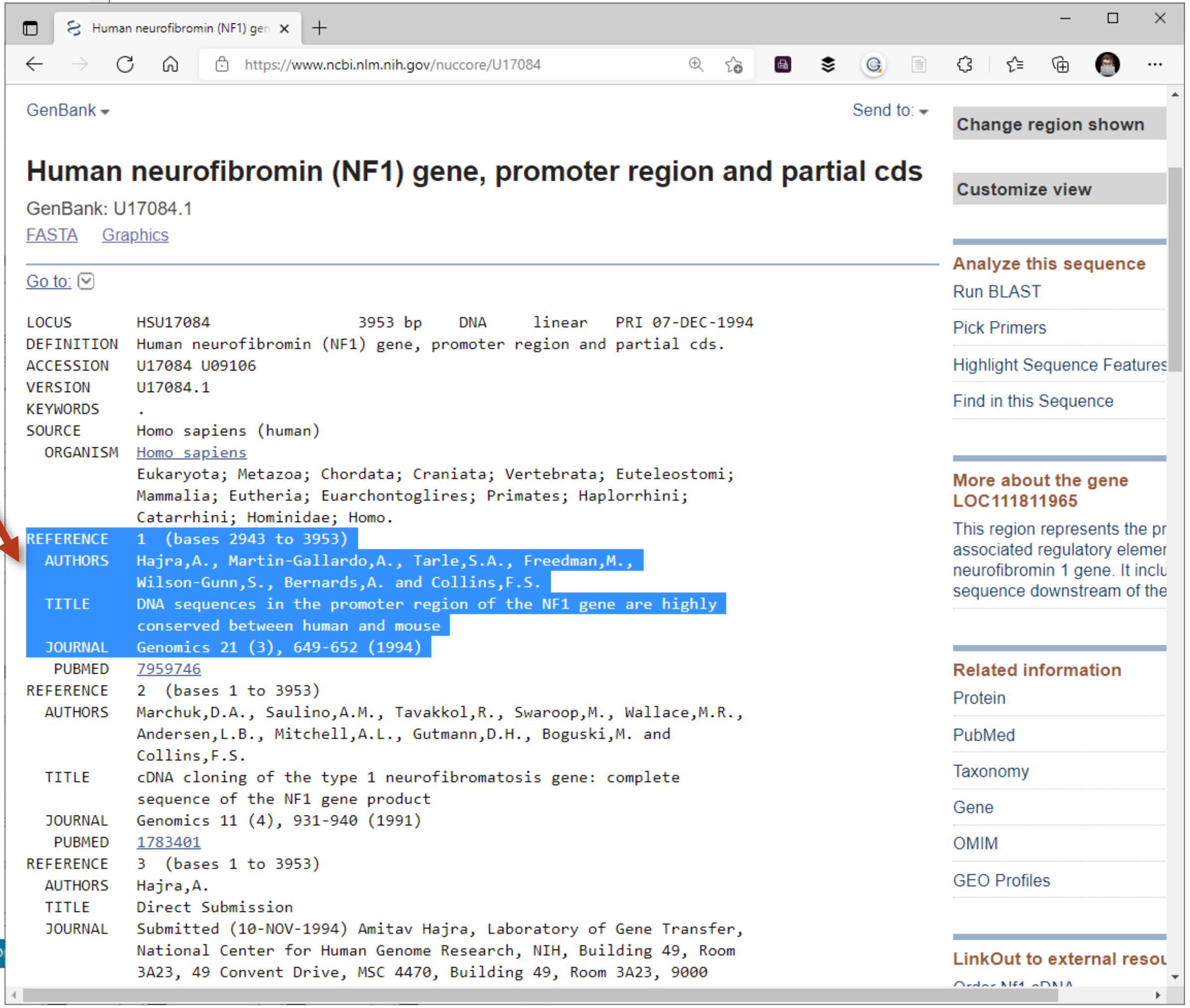
Cited by (42)

Synthetic promoter for efficient and muscle-specific expression of exogenous genes

2019, Plasmid

Citation Excerpt :

…To acquire synthetic promoters with activity and specificity high enough for therapeutic application, we speculated that the combination of multiple kinds of transcriptional motifs, instead of unique myogenic



Human neurofibromin (NF1) gen
https://www.ncbi.nlm.nih.gov/nuccore/U17084
GenBank
Send to:
Change region shown
Customize view
FASTA
Graphics
Go to:

Human neurofibromin (NF1) gene, promoter region and partial cds

GenBank: U17084.1

```
LOCUS       HSU17084                3953 bp    DNA     linear   PRI 07-DEC-1994
DEFINITION  Human neurofibromin (NF1) gene, promoter region and partial cds.
ACCESSION   U17084 U09106
VERSION     U17084.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 2943 to 3953)
  AUTHORS   Hajra,A., Martin-Gallardo,A., Tarle,S.A., Freedman,M.,
            Wilson-Gunn,S., Bernards,A. and Collins,F.S.
  TITLE     DNA sequences in the promoter region of the NF1 gene are highly
            conserved between human and mouse
  JOURNAL   Genomics 21 (3), 649-652 (1994)
   PUBMED   7959746
REFERENCE   2  (bases 1 to 3953)
  AUTHORS   Marchuk,D.A., Saulino,A.M., Tavakkol,R., Swaroop,M., Wallace,M.R.,
            Andersen,L.B., Mitchell,A.L., Gutmann,D.H., Boguski,M. and
            Collins,F.S.
  TITLE     cDNA cloning of the type 1 neurofibromatosis gene: complete
            sequence of the NF1 gene product
  JOURNAL   Genomics 11 (4), 931-940 (1991)
   PUBMED   1783401
REFERENCE   3  (bases 1 to 3953)
  AUTHORS   Hajra,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-NOV-1994) Amitav Hajra, Laboratory of Gene Transfer,
            National Center for Human Genome Research, NIH, Building 49, Room
            3A23, 49 Convent Drive, MSC 4470, Building 49, Room 3A23, 9000
```

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

More about the gene LOC111811965

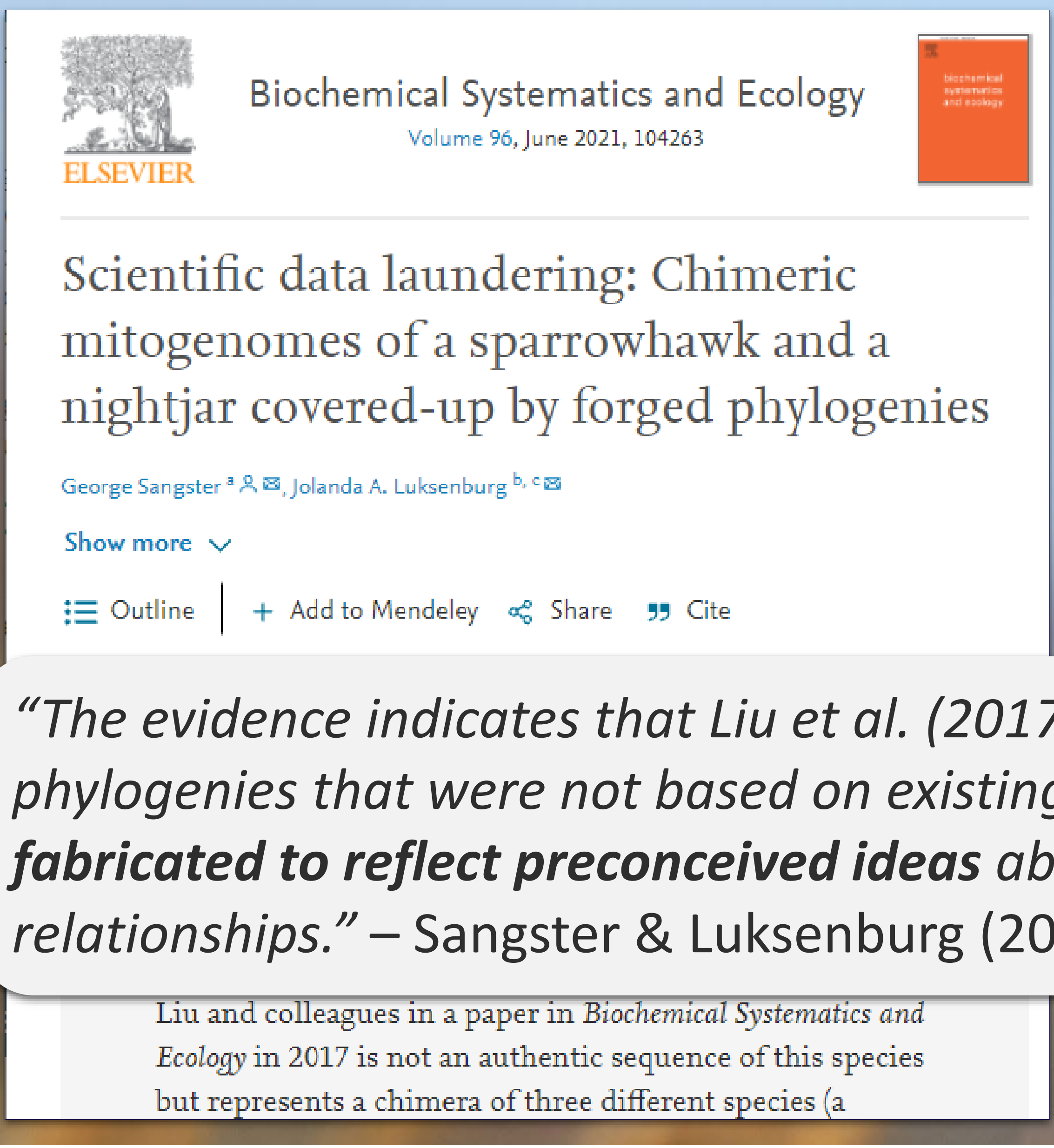
This region represents the pr associated regulatory elemer neurofibromin 1 gene. It inclu sequence downstream of the

Related information

- Protein
- PubMed
- Taxonomy
- Gene
- OMIM
- GEO Profiles

LinkOut to external resou

Falsified sequencing data to support a false phylogeny



ELSEVIER

Biochemical Systematics and Ecology
Volume 96, June 2021, 104263

Scientific data laundering: Chimeric mitogenomes of a sparrowhawk and a nightjar covered-up by forged phylogenies

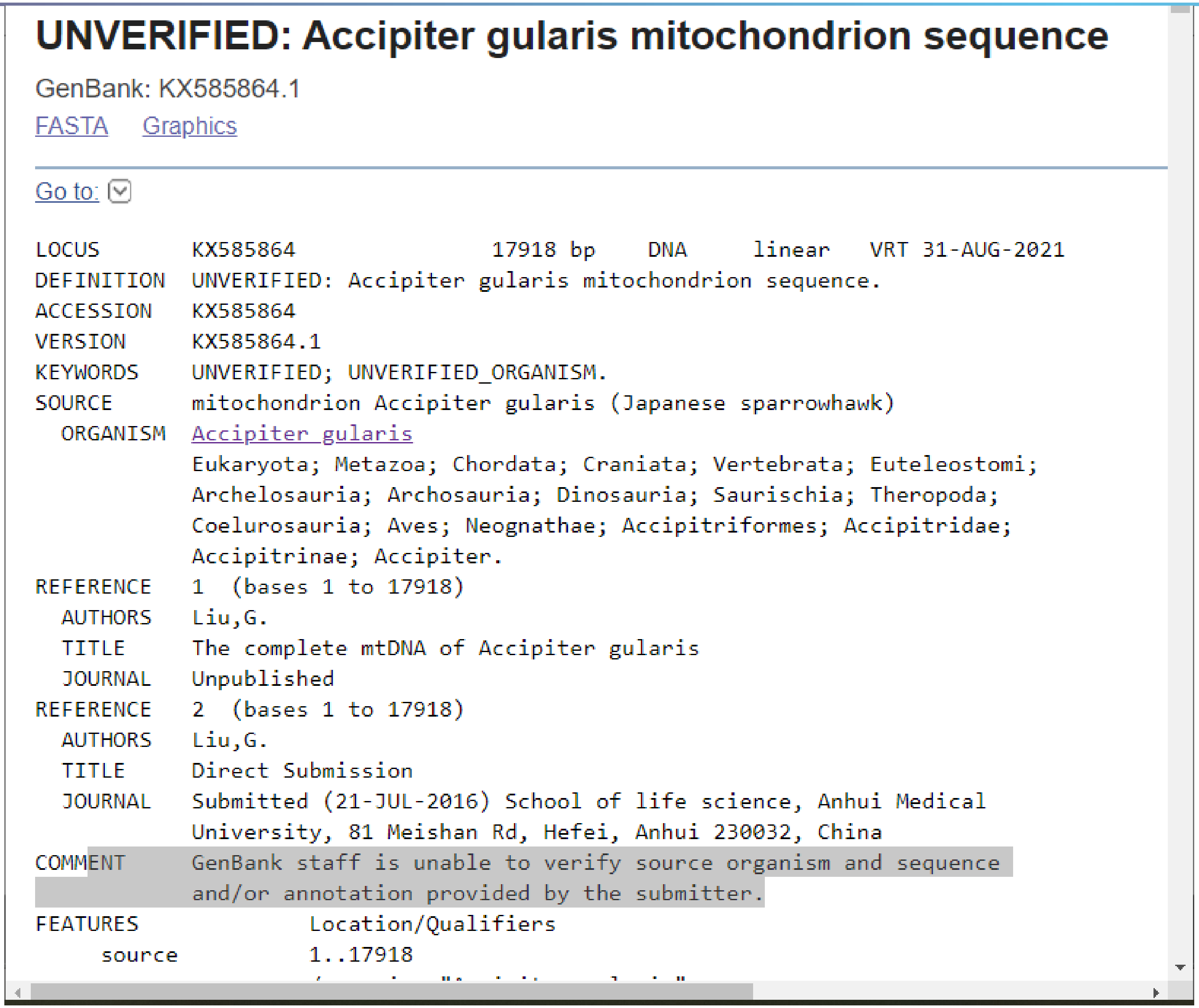
George Sangster [a], Jolanda A. Luksenburg [b, c]

Liu and colleagues in a paper in *Biochemical Systematics and Ecology* in 2017 is not an authentic sequence of this species but represents a chimera of three different species (a

> *"The evidence indicates that Liu et al. (2017) published phylogenies that were not based on existing data **but were fabricated to reflect preconceived ideas** about phylogenetic relationships."* – Sangster & Luksenburg (2021)

Sangster, G. and Luksenburg, J.A. (2021) 'Scientific data laundering: Chimeric mitogenomes of a sparrowhawk and a nightjar covered-up by forged phylogenies', *Biochemical Systematics and Ecology*, 96, p. 104263. doi:10.1016/j.bse.2021.104263.

Unfortunately, the falsified mitogenome is still in GenBank…



UNVERIFIED: Accipiter gularis mitochondrion sequence

GenBank: KX585864.1

FASTA   Graphics

Go to:

```
LOCUS       KX585864               17918 bp    DNA     linear   VRT 31-AUG-2021
DEFINITION  UNVERIFIED: Accipiter gularis mitochondrion sequence.
ACCESSION   KX585864
VERSION     KX585864.1
KEYWORDS    UNVERIFIED; UNVERIFIED_ORGANISM.
SOURCE      mitochondrion Accipiter gularis (Japanese sparrowhawk)
  ORGANISM  Accipiter gularis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Accipitriformes; Accipitridae;
            Accipitrinae; Accipiter.
REFERENCE   1  (bases 1 to 17918)
  AUTHORS   Liu,G.
  TITLE     The complete mtDNA of Accipiter gularis
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 17918)
  AUTHORS   Liu,G.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-JUL-2016) School of life science, Anhui Medical
            University, 81 Meishan Rd, Hefei, Anhui 230032, China
COMMENT     GenBank staff is unable to verify source organism and sequence
            and/or annotation provided by the submitter.
FEATURES             Location/Qualifiers
     source          1..17918
```

Labeled as "Unverified", but the sequence still remains in GenBank and, for example, will come up in a BLAST search…

Intentional falsification is rare… but accidents happen right?

(>2 million times) …

Mukherjee *et al. Standards in Genomic Sciences* 2015, **10**:18
http://www.standardsingenomics.com/content/10/1/18

Standards in Genomic Sciences

COMMENTARY **Open Access**

Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratim Mukherjee[1*], Marcel Huntemann[1], Natalia Ivanova[1], Nikos C Kyrpides[1,2] and Amrita Pati[1]

Abstract

With the rapid growth and development of sequencing technologies, ge
exploring solutions to some of the world's biggest challenges such as sea
exploration of genomic dark matter. However, progress in sequencing ha
that can occur during template or library preparation, sequencing, imagin
screened over 18,000 publicly available microbial isolate genome sequen
database and identified more than 1000 genomes that are contaminated
during Illumina sequencing runs. Approximately 10% of these genomes h
contaminated genomes were sequenced under the Human Microbiome
contamination from various sources and are usually eliminated during do
of PhiX contaminated genomes indicates a lapse in either the application
measures. The presence of PhiX contamination in several publicly availab
errors when such data are used in comparative genomics analyses. Such
far-reaching consequences in the form of erroneous data interpretation a
measures to proofread raw sequences before releasing them to the broa

Keywords: Next-generation sequencing, PhiX, Contamination, Comparati

Background

The ability to produce large numbers of high-quality, low-cost reads has revolutionized the field of microbiology [1-3]. Starting from a meager 1575 registered projects in September 2005, there has been a steady increase in the number of sequencing projects according to the Genomes OnLine Database [4]. As of November 17th 2014, there were 41,553 bacterial and archaeal isolate genome sequencing projects reported in GOLD [4,5]. This explosion of genome sequencing projects especially during the last 5 years has been largely cata-

Despite its
platform doe
need to be
One such ch
used as a qu
runs. PhiX i
with a single
5386 nucleot
sequenced by
defined geno
used as a co

Downloaded from genome.cshlp.org on October 20, 2021 - Published by Cold Spring Harbor Laboratory Press

Research

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,[1] Mihaela Pertea,[1,2] Aleksey V. Zimin,[1,3] and Steven L. Salzberg[1,2,3,4]

[1] Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; [2] Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; [3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; [4] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

shed genomes can cause numerous problems for downstream analyses, partic-
mics projects. Our large-scale scan of complete and draft bacterial and archaeal
s that 2250 genomes are contaminated by human sequence. The contaminant
uman repeat regions, which themselves are not adequately represented in the
The absence of the sequences from the human assembly offers a likely expla-
lies. In some cases, the contaminating contigs have been erroneously annotated
h over time have propagated to create spurious protein "families" across mul-
As a result, 3437 spurious protein entries are currently present in the widely
report here an extensive list of contaminant sequences in bacterial genome as-
em. We found that nearly all contaminants occurred in small contigs in draft
nall contigs from draft genome assemblies may mitigate the issue of contami-
enuine genomic sequences.

rticle.]

y available ge-
s to well over
l resources for
g microbiome
nplex samples
ence databases
but for practi-
today are still
gs or scaffolds
chromosomes
e or "finished"
y chromosome
ne human ge-
her animal ge-
an assembly,
folds that con-
sequence has
ive regions are
g to problems
es vary widely

assemblies in the NCBI and UCSC Genome Browser databases were contaminated with the primate-specific *AluY* repeats (Longo et al. 2011). Although validation pipelines have improved substantially since then (Tatusova et al. 2016; Haft et al. 2018), some contaminants still remain, as we describe below. Furthermore, when open reading frames (ORFs) in the contaminated contigs get annotated as protein-coding genes, their protein sequence may be added to other databases. Once in those databases, these spurious proteins may in turn be used in future annotation, leading to the so-called "transitive catastrophe" problem where errors are propagated widely (Karp 1998; Salzberg 2007; Danchin et al. 2018). Indeed, one study found that the percentage of misannotated entries in the NCBI nonredundant (nr) protein collection, which is used for thousands of BLAST searches every day, has been increasing over time (Schnoes et al. 2009).

Contamination of genomic sequences can be particularly problematic for metagenomic studies. For example, if a genome labeled as species X contains fragments of the human genome, then any sample containing human DNA might erroneously be identified as also containing species X. Since human DNA is virtually al-

Steinegger and Salzberg *Genome Biology* (2020) 21:115
https://doi.org/10.1186/s13059-020-02023-1

Genome Biology

METHOD **Open Access**

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Check for updates

Martin Steinegger[1,2,3*] and Steven L. Salzberg[2,4,5]

*Correspondence:
martin.steinegger@snu.ac.kr
[1] School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea
[2] Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, 21218 Baltimore, Maryland, USA
Full list of author information is available at the end of the article

Abstract

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to "complete" model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/martin-steinegger/conterminator

Poor quality genomes result in taxonomic misclassification

Multiple papers (more than the two listed here) have found widespread misclassification in GenBank

Bioinformatics, 36(18), 2020, 4699–4705
doi: 10.1093/bioinformatics/btaa586
Advance Access Publication Date: 24 June 2020
Original Paper

OXFORD

Sequence analysis

Detecting and correcting misclassified sequences in the large-scale public databases

Hamid Bagheri [1,*], Andrew J. Severin[2] and Hridesh Rajan[1]

[1]Department of Computer Science and [2]Genome Informatics Facility, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on April 2, 2020; revised on June 10, 2020; editorial … June 16, 2020

Abstract

Motivation: As the cost of sequen… deposited into public repositories is increasing rapidly. Public databases rely … h submission that is prone to user error. Unfortunately, most public datab… user input and do not have methods for identifying errors in the provided … propagation. Previous research on a small subset of the NR database analyz… milarity. To the best of our knowledge, the amount of misclassification in the … We propose a heuristic method to detect potentially misclassified taxonomic as… nique and quality control to find the most probable taxonomic … ency of each annotation from manually and computationall… ty.

Results: We found more than two million … database. Using simulated data, we show a high precision of 9… classified proteins. The proposed approach and findings could also be app…

Availability and implementation: Source code, dataset, do… ker container are available at https://github.com/boalang/nr.

Contact: hbagheri@iastate.edu

Supplementary information: Supplementary data are availab…

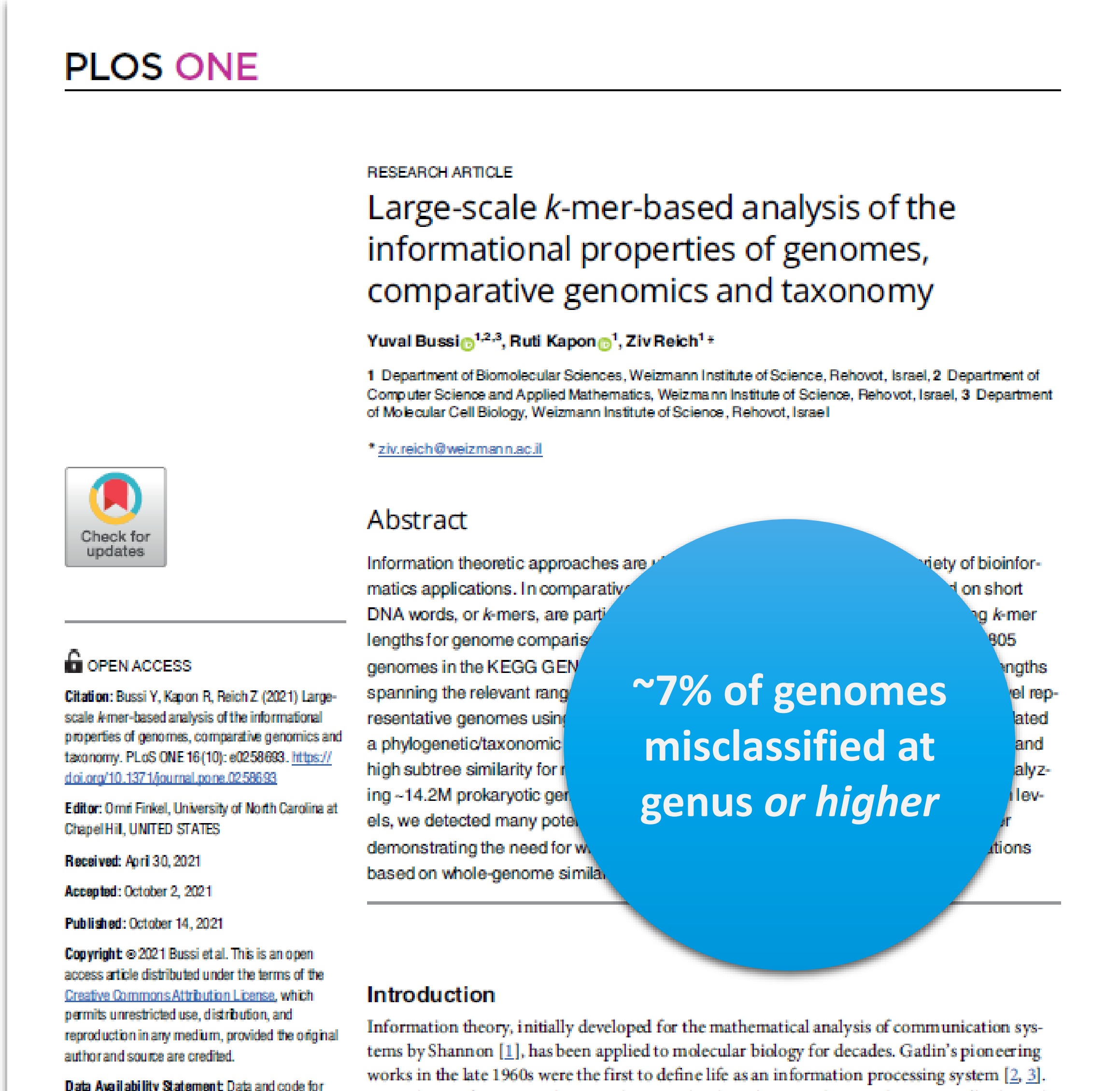
1 Introduction

Researchers use BLAST on the non-redundant (NR) database on a daily basis to identify the source and function of a protein/DNA sequence. The NR database encompasses protein sequences from

are de… sequences were deposited by a … ns obtained from a soybean roots, th… hat metadata will be labeled with the organism … max. If the researcher had in fact been working on Glycine soja then this would result in a misassign-

~7.8% of genomes misclassified at the species level

~4% at the genus level



PLOS ONE

RESEARCH ARTICLE

Large-scale *k*-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy

Yuval Bussi[1,2,3], Ruti Kapon[1], Ziv Reich[1]*

1 Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel, 2 Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, 3 Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel

* ziv.reich@weizmann.ac.il

Check for updates

OPEN ACCESS

Citation: Bussi Y, Kapon R, Reich Z (2021) Large-scale *k*-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. PLoS ONE 16(10): e0258693. https://doi.org/10.1371/journal.pone.0258693

Editor: Omri Finkel, University of North Carolina at Chapel Hill, UNITED STATES

Received: April 30, 2021

Accepted: October 2, 2021

Published: October 14, 2021

Data Availability Statement: Data and code for

Abstract

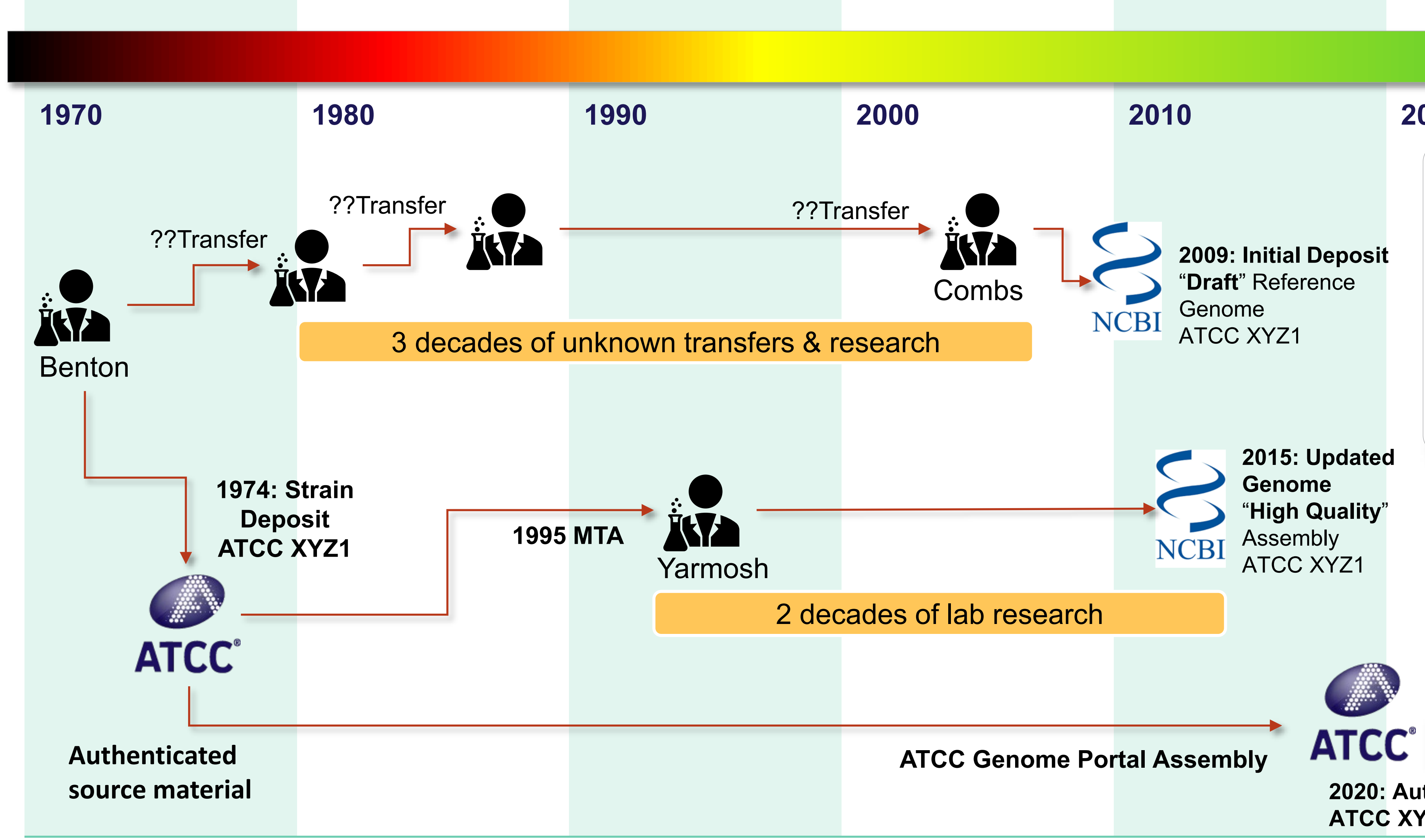
Information theoretic approaches are … riety of bioinformatics applications. In comparativ… d on short DNA words, or *k*-mers, are parti… g *k*-mer lengths for genome comparis… 805 genomes in the KEGG GEN… engths spanning the relevant rang… el representative genomes usin… lated a phylogenetic/taxonomic … and high subtree similarity for … alyzing ~14.2M prokaryotic ge… levels, we detected many pote… r demonstrating the need for w… ations based on whole-genome simila…

Introduction

Information theory, initially developed for the mathematical analysis of communication systems by Shannon [1], has been applied to molecular biology for decades. Gatlin's pioneering works in the late 1960s were the first to define life as an information processing system [2, 3].

~7% of genomes misclassified at genus *or higher*

Challenging traceability of most public genomics data



3 decades of unknown transfers & research

2009: Initial Deposit
“**Draft**” Reference
Genome
ATCC XYZ1

1974: Strain
Deposit
ATCC XYZ1

2015: Updated
Genome
“High Quality”
Assembly
ATCC XYZ1

2 decades of lab research

Authenticated
source material

ATCC Genome Portal Assembly

Potential issues with genomic source material

- “Lab adaptation”
- Loss of plasmids
- Sample mix ups
- Unknown chain of custody
- Differences in sequencing technology and bioinformatics

SNPs differences	ATCC Reference
Draft genome	>13,000
“High-quality” reference genome	>10,000

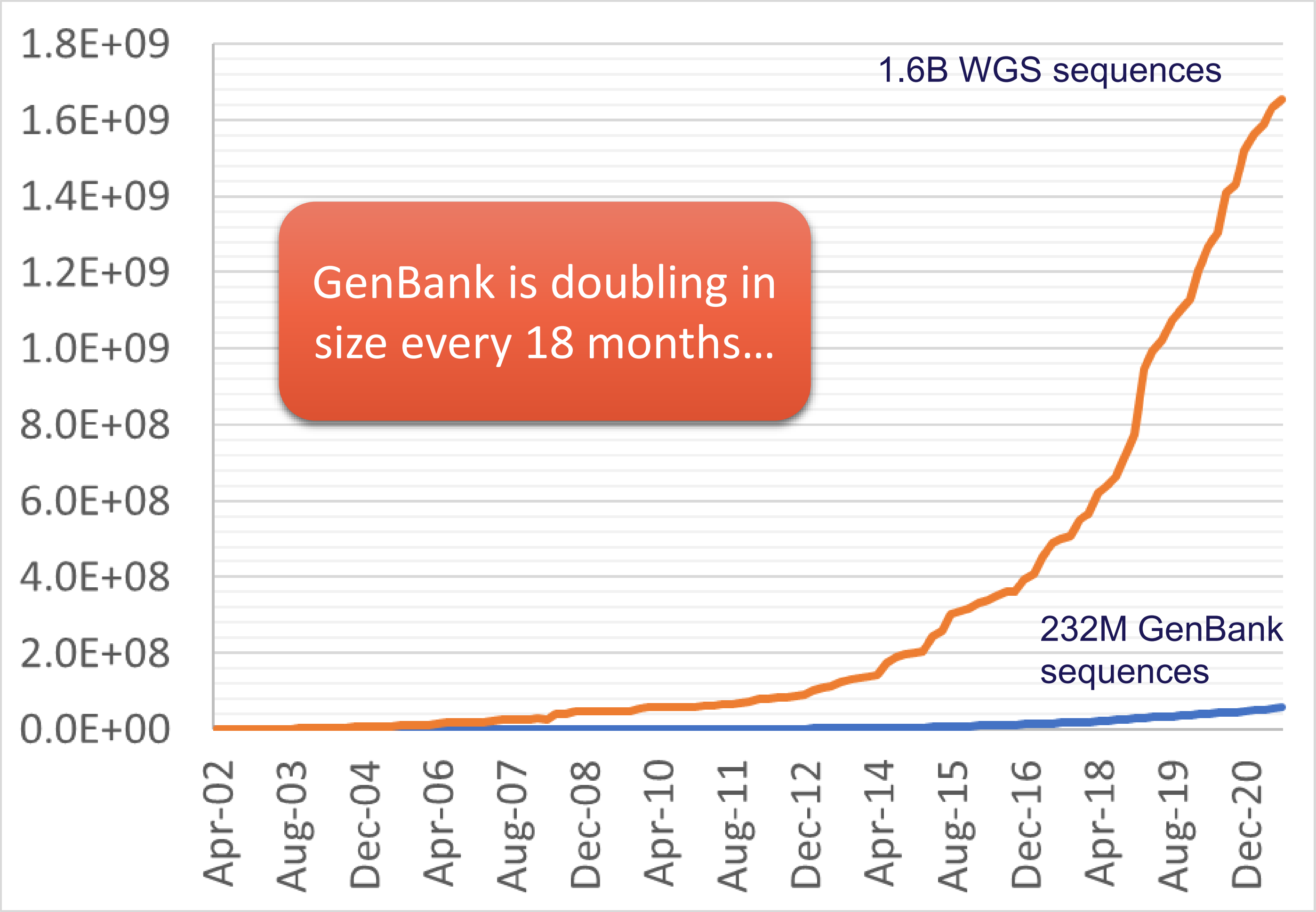
100% authenticity and traceability

2020: Authenticated Genome Reference
ATCC XYZ1

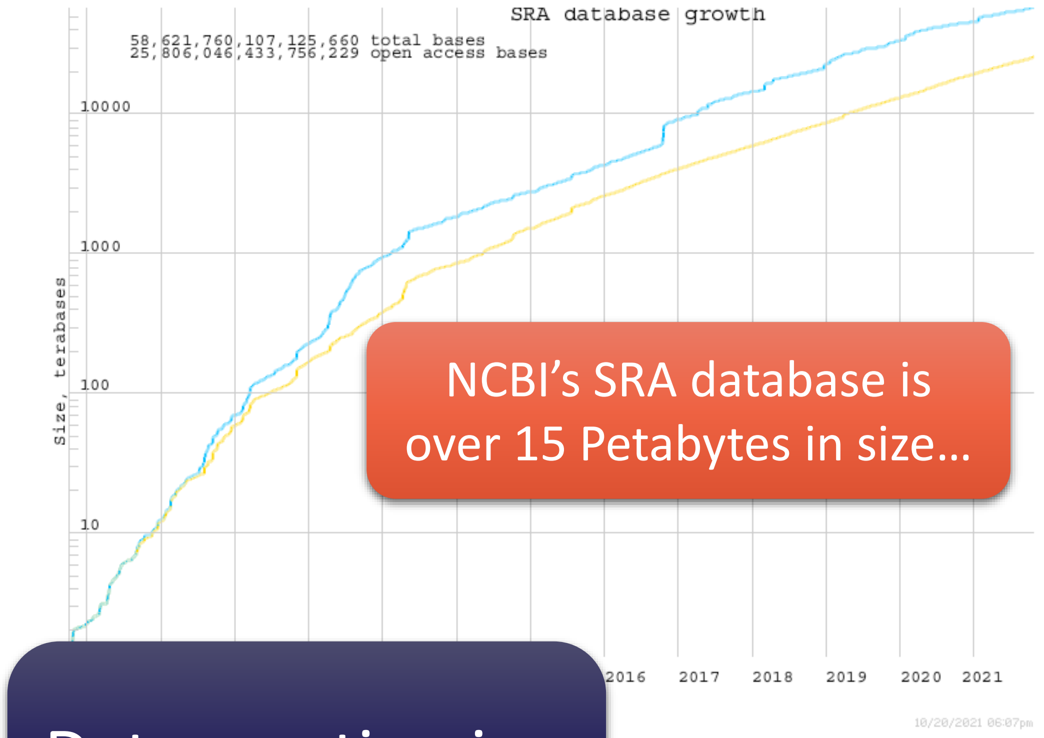
Names Changed To Protect The Innocent.

A reminder on the growth of public genomics data

1.6B sequences in WGS
232M sequences in GenBank



GenBank is doubling in size every 18 months…



NCBI's SRA database is over 15 Petabytes in size…

Data curation is a huge challenge

Genomics data quality issues impact many disciplines

FACTORS

- Misclassification of sequences
- Chimeric genome assemblies
- Sample contamination
- Sequencing errors
- Mislabeling or data errors
- Data omission
- Data obfuscation
- Intentional misconduct

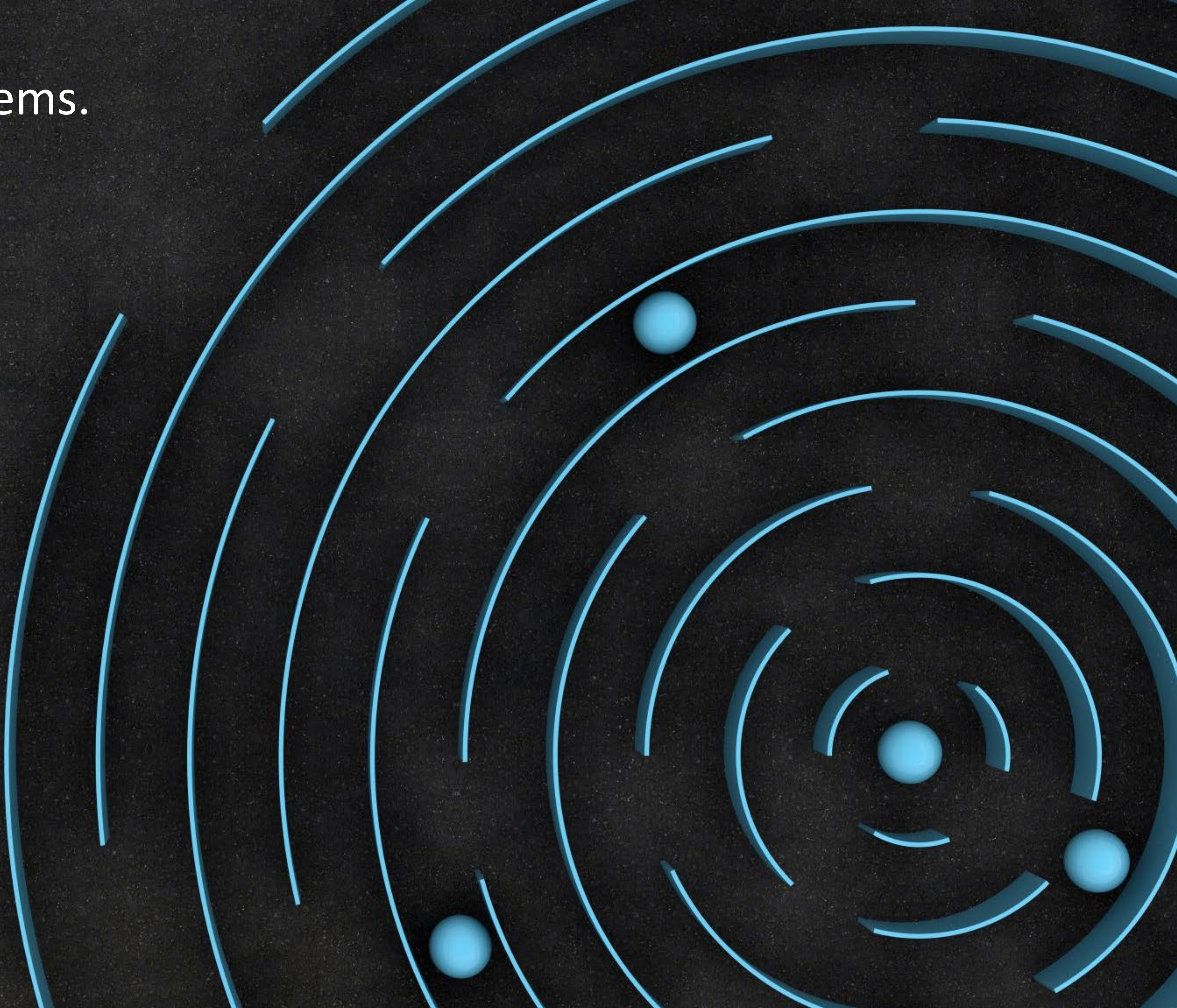
Critically Impacted Areas

- Basic research (hypothesis generation)
- Biodiversity and environmental sciences
- Diagnostics & epidemiology
- Forensics
- Food safety
- Biodefense
- Many other areas…

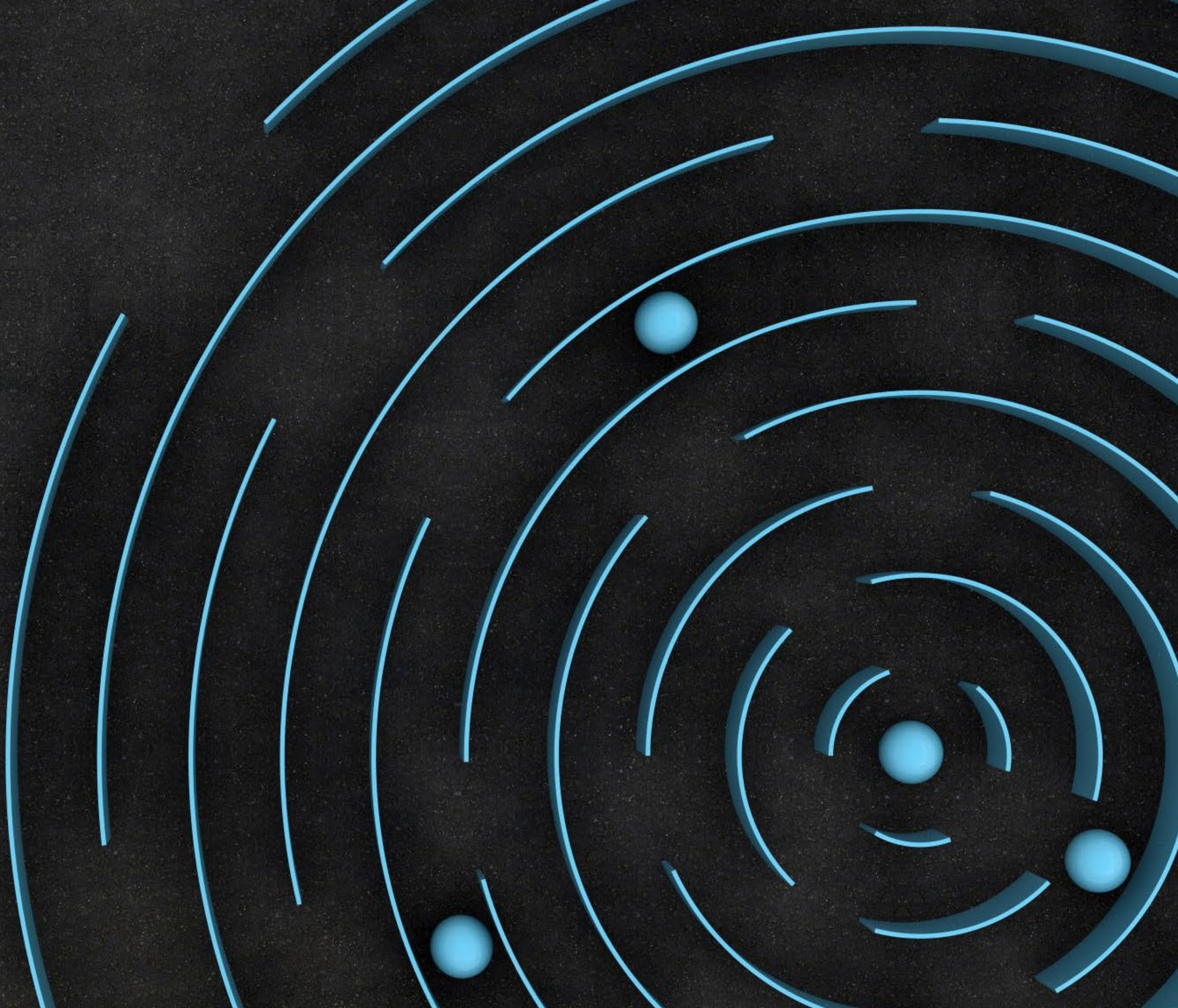
These are not “new” problems.

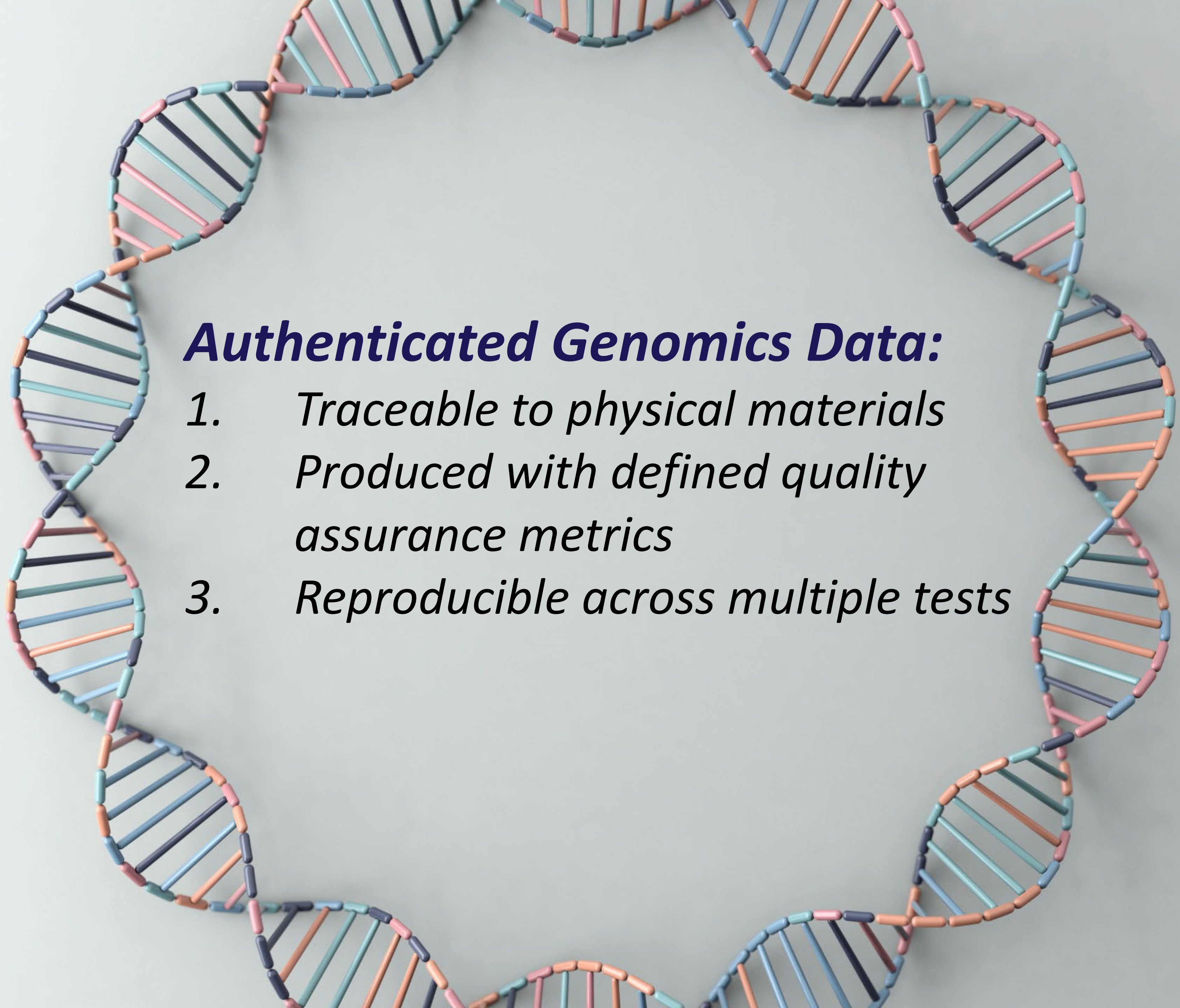
Many groups have sought solutions.

None, however, have sought to create **Authenticated Genomics Data**



What is “Authenticated Genomics Data”?



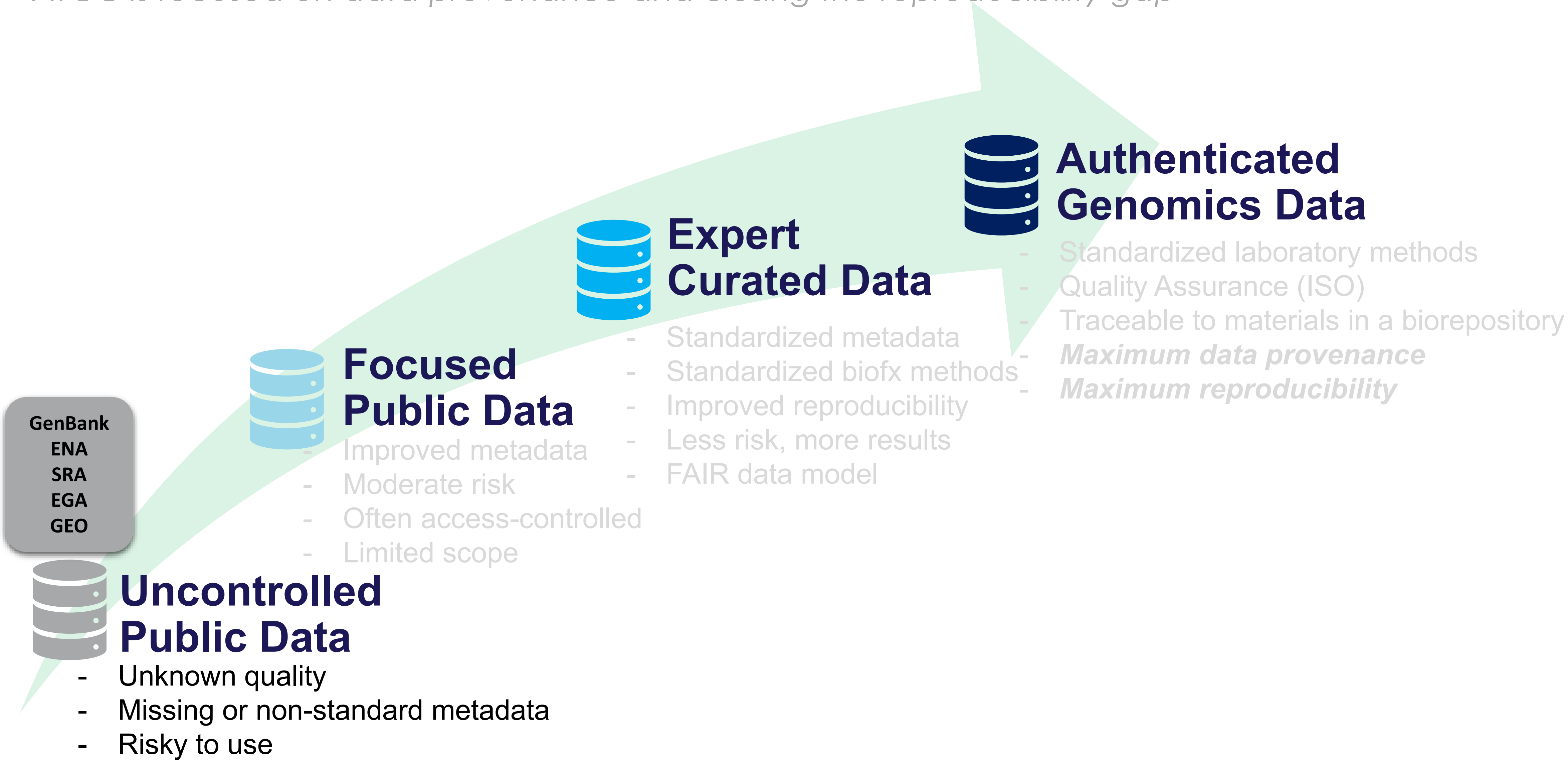


Authenticated Genomics Data:

1. *Traceable to physical materials*
2. *Produced with defined quality assurance metrics*
3. *Reproducible across multiple tests*

Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap



GenBank
ENA
SRA
EGA
GEO

Uncontrolled Public Data

- Unknown quality
- Missing or non-standard metadata
- Risky to use

Focused Public Data

- Improved metadata
- Moderate risk
- Often access-controlled
- Limited scope

Expert Curated Data

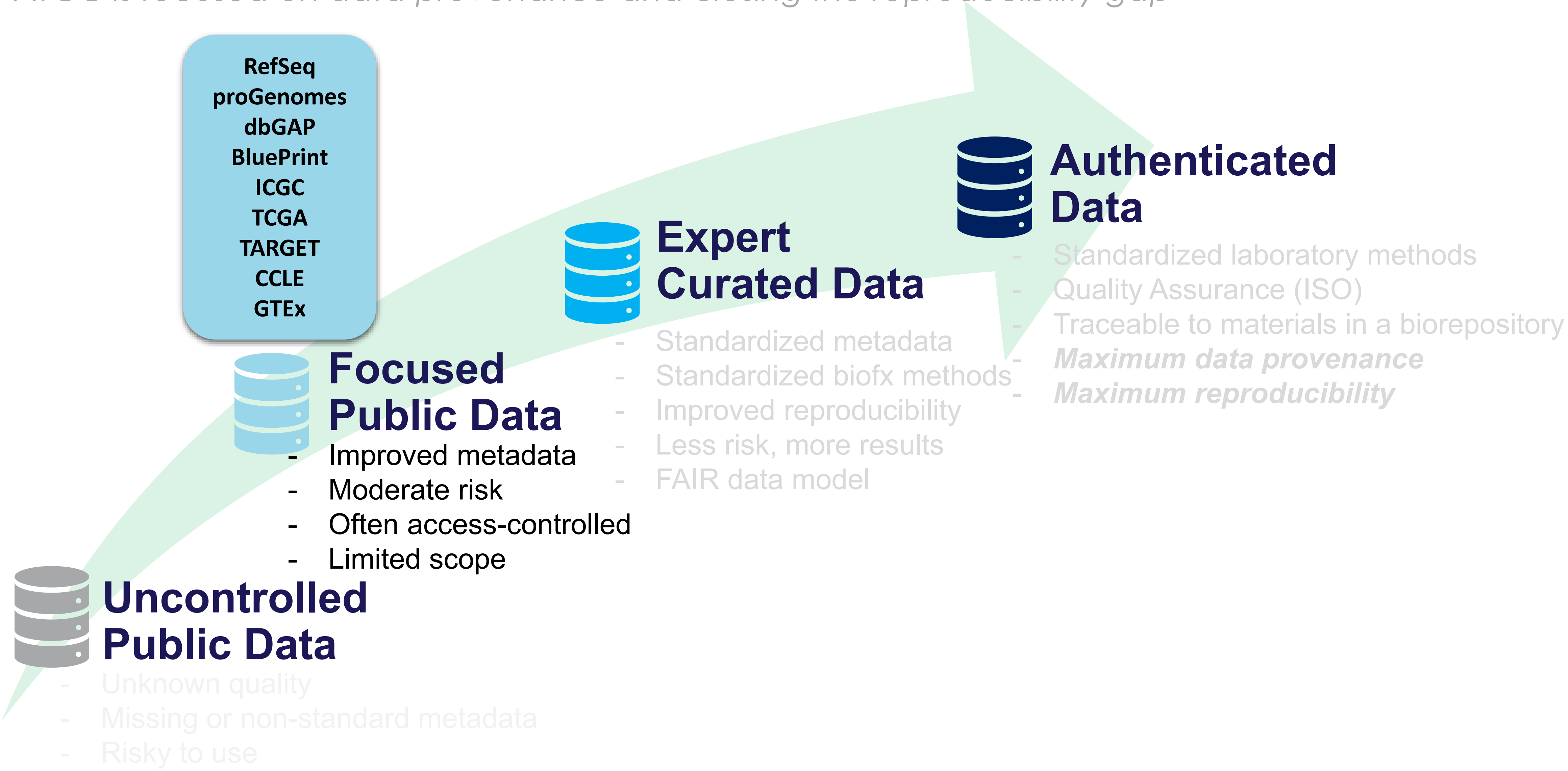
- Standardized metadata
- Standardized biofx methods
- Improved reproducibility
- Less risk, more results
- FAIR data model

Authenticated Genomics Data

- Standardized laboratory methods
- Quality Assurance (ISO)
- Traceable to materials in a biorepository
- ***Maximum data provenance***
- ***Maximum reproducibility***

Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap



RefSeq
proGenomes
dbGAP
BluePrint
ICGC
TCGA
TARGET
CCLE
GTEx

Uncontrolled Public Data

- Unknown quality
- Missing or non-standard metadata
- Risky to use

Focused Public Data

- Improved metadata
- Moderate risk
- Often access-controlled
- Limited scope

Expert Curated Data

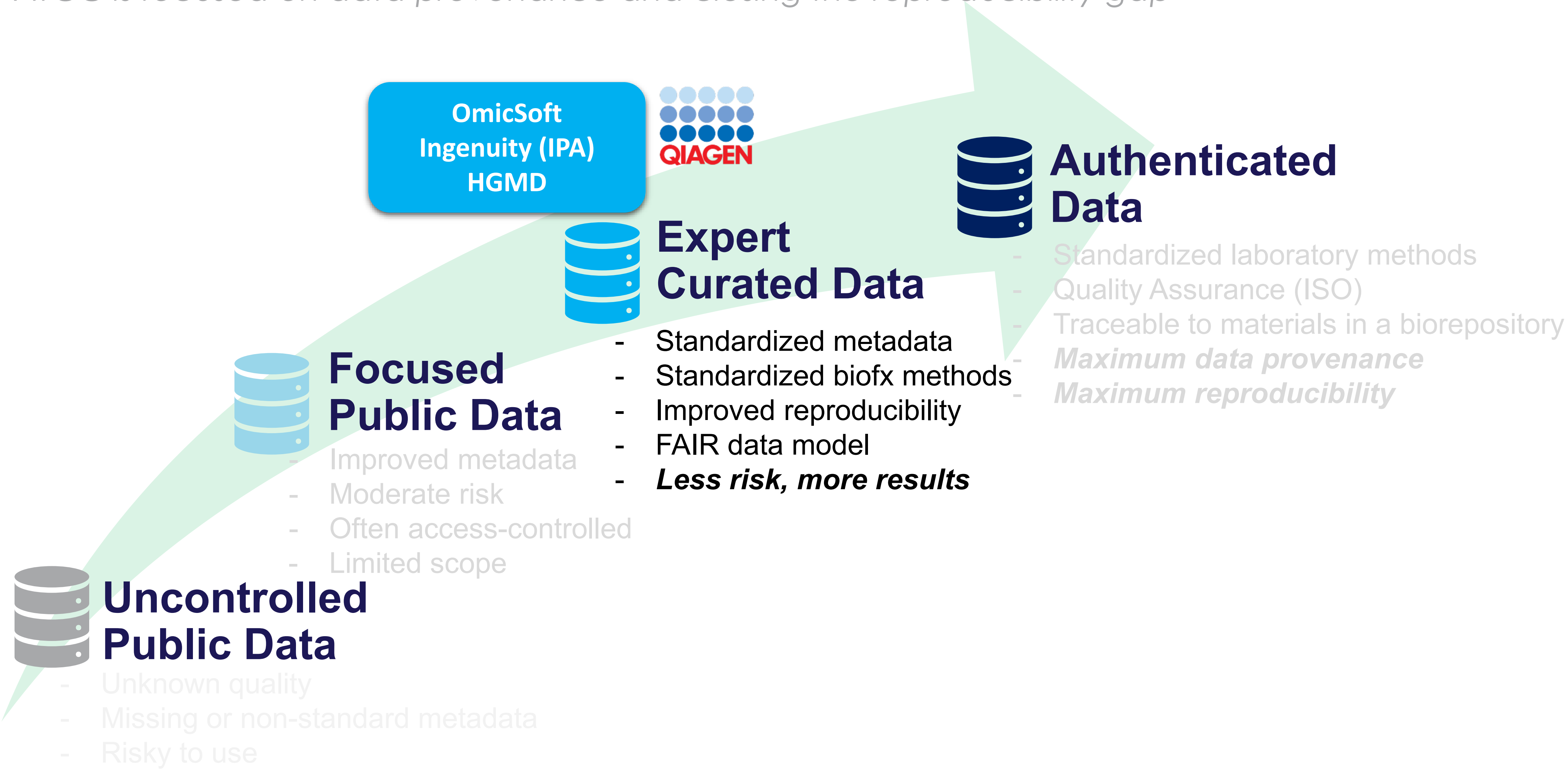
- Standardized metadata
- Standardized biofx methods
- Improved reproducibility
- Less risk, more results
- FAIR data model

Authenticated Data

- Standardized laboratory methods
- Quality Assurance (ISO)
- Traceable to materials in a biorepository
- ***Maximum data provenance***
- ***Maximum reproducibility***

Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap



OmicSoft
Ingenuity (IPA)
HGMD

QIAGEN

Uncontrolled Public Data

- Unknown quality
- Missing or non-standard metadata
- Risky to use

Focused Public Data

- Improved metadata
- Moderate risk
- Often access-controlled
- Limited scope

Expert Curated Data

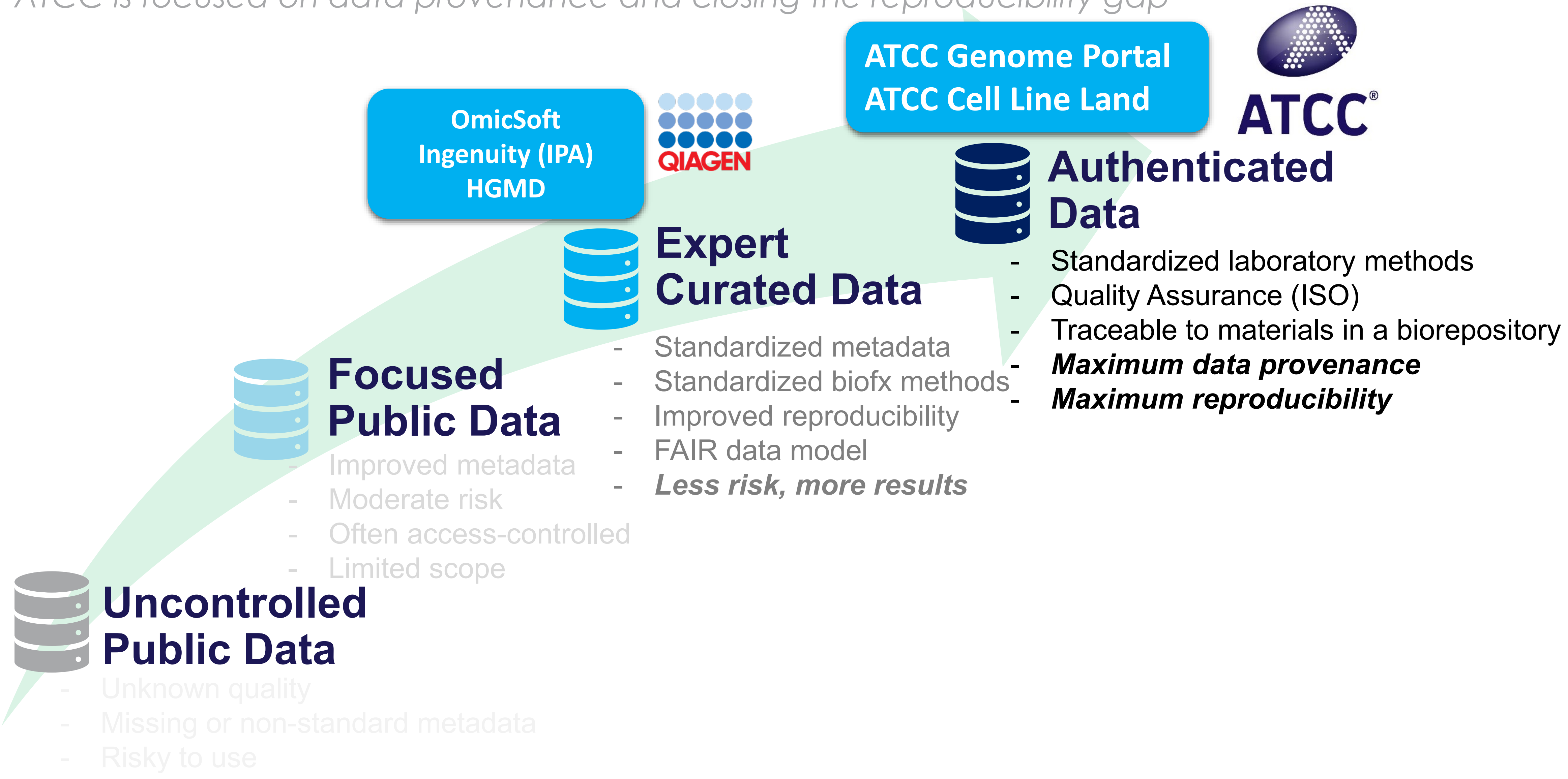
- Standardized metadata
- Standardized biofx methods
- Improved reproducibility
- FAIR data model
- ***Less risk, more results***

Authenticated Data

- Standardized laboratory methods
- Quality Assurance (ISO)
- Traceable to materials in a biorepository
- ***Maximum data provenance***
- ***Maximum reproducibility***

Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap



ATCC Genome Portal
ATCC Cell Line Land

Authenticated Data

- Standardized laboratory methods
- Quality Assurance (ISO)
- Traceable to materials in a biorepository
- ***Maximum data provenance***
- ***Maximum reproducibility***

OmicSoft
Ingenuity (IPA)
HGMD

Expert Curated Data

- Standardized metadata
- Standardized biofx methods
- Improved reproducibility
- FAIR data model
- ***Less risk, more results***

Focused Public Data

- Improved metadata
- Moderate risk
- Often access-controlled
- Limited scope

Uncontrolled Public Data

- Unknown quality
- Missing or non-standard metadata
- Risky to use

The ATCC Genome Portal

Tackling the reproducibility gap in microbial genomics

ATCC Genome Portal

The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal

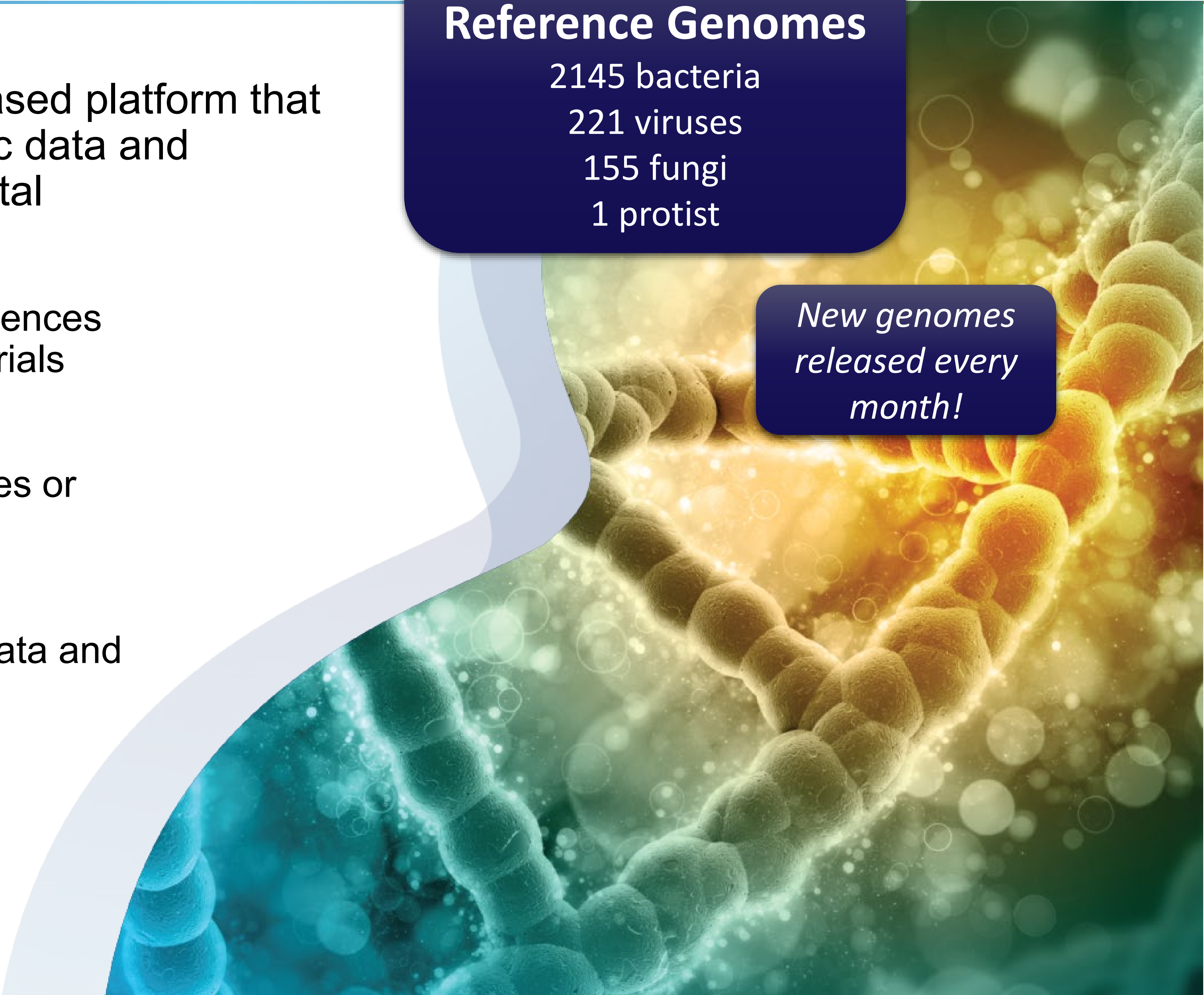
- Download whole-genome sequences and annotations of ATCC materials
- Search for nucleotide sequences or genes within genomes
- View genome assembly metadata and quality metrics

genomes.atcc.org

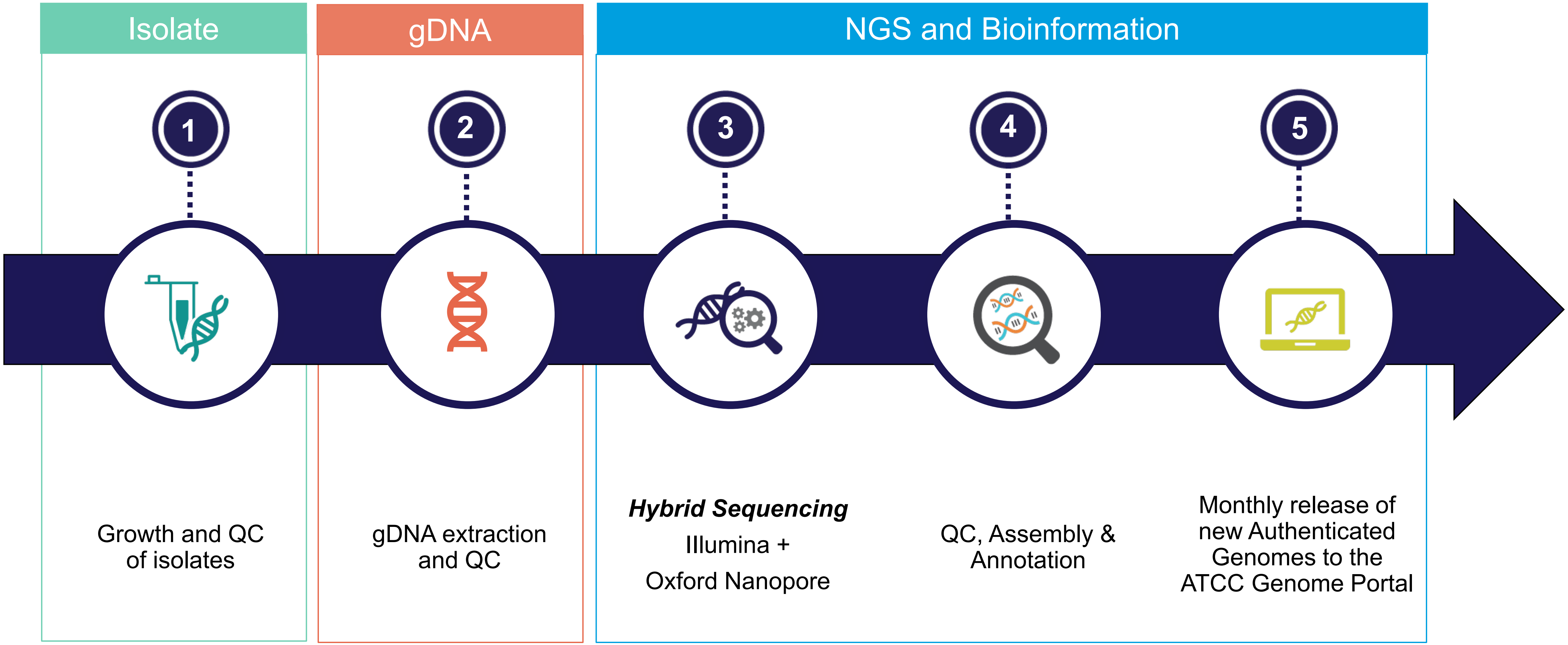
2,522 Authenticated Reference Genomes

2145 bacteria
221 viruses
155 fungi
1 protist

New genomes released every month!



Authenticated physical material coupled with reference-quality genome sequences



Isolate

gDNA

NGS and Bioinformation

1

Growth and QC of isolates

2

gDNA extraction and QC

3

Hybrid Sequencing

Illumina +

Oxford Nanopore

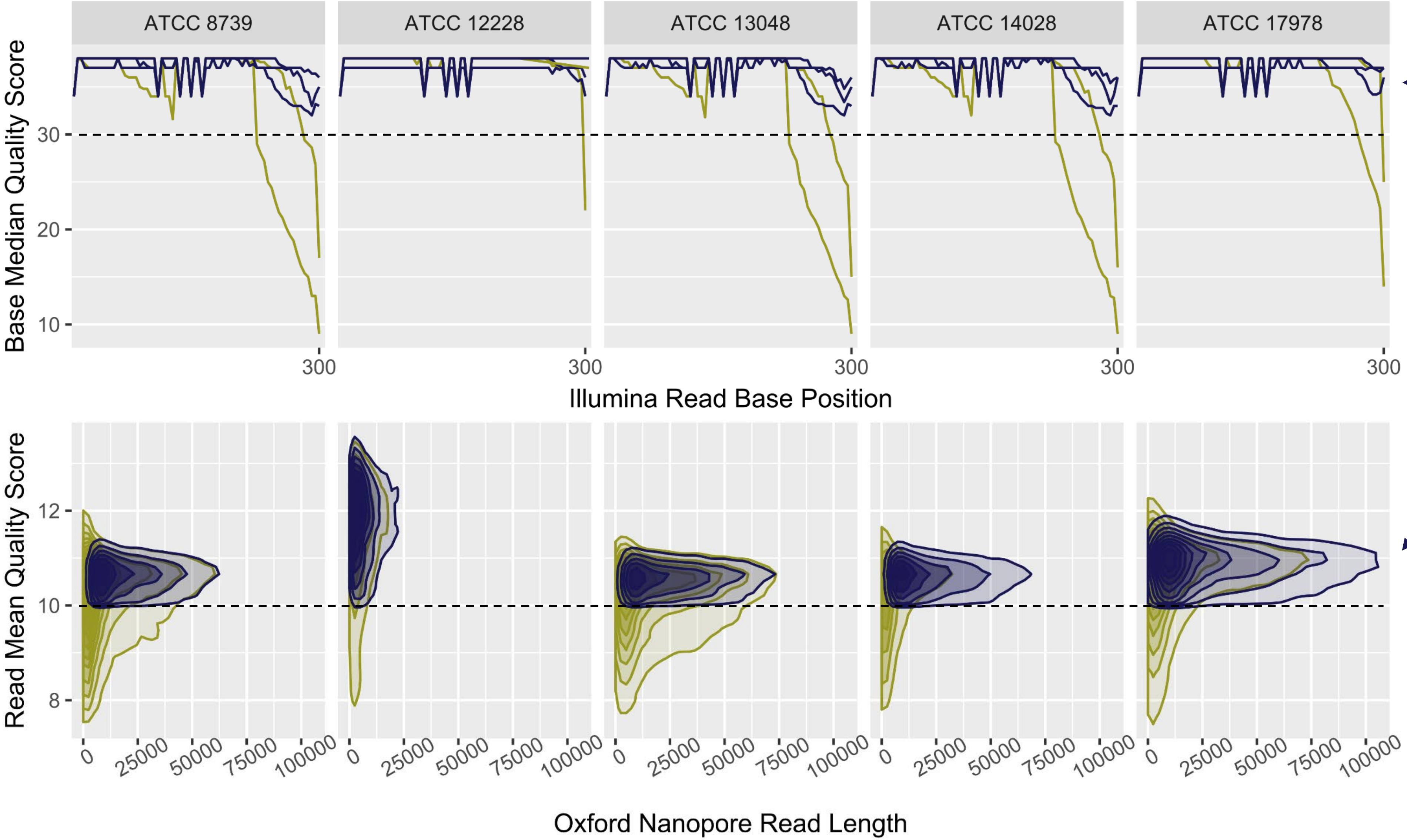
4

QC, Assembly & Annotation

5

Monthly release of new Authenticated Genomes to the ATCC Genome Portal

Sequencing QC – Read trimming/filtering



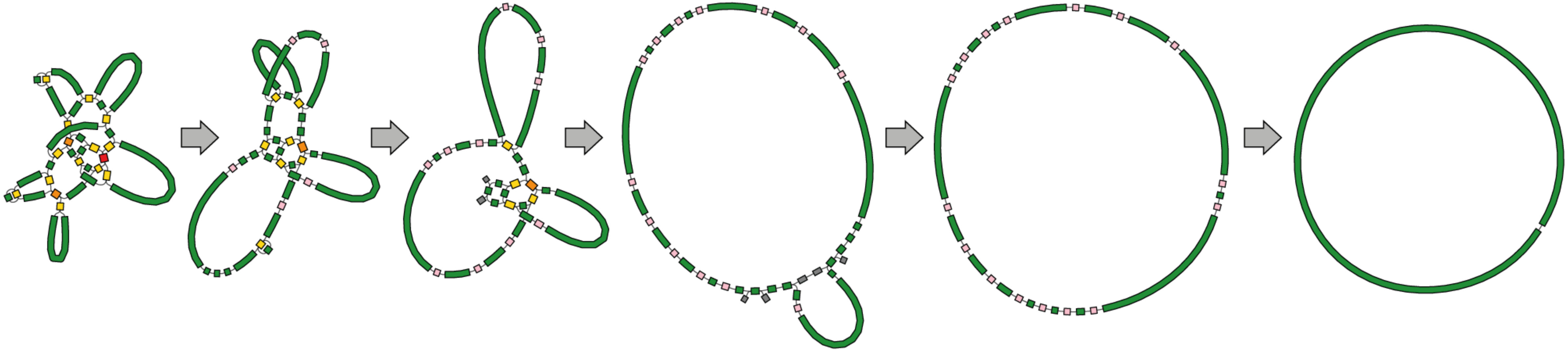
Only keep high-quality base calls

Only keep long, good quality reads

■ Quality controlled data

■ Raw data

Hybrid genome assembly



Illumina-only genome assembly
150 bp reads

Long reads mapped to a tangled region creates a resolved bridge
Successively applying bridges resolves the structure of the genome

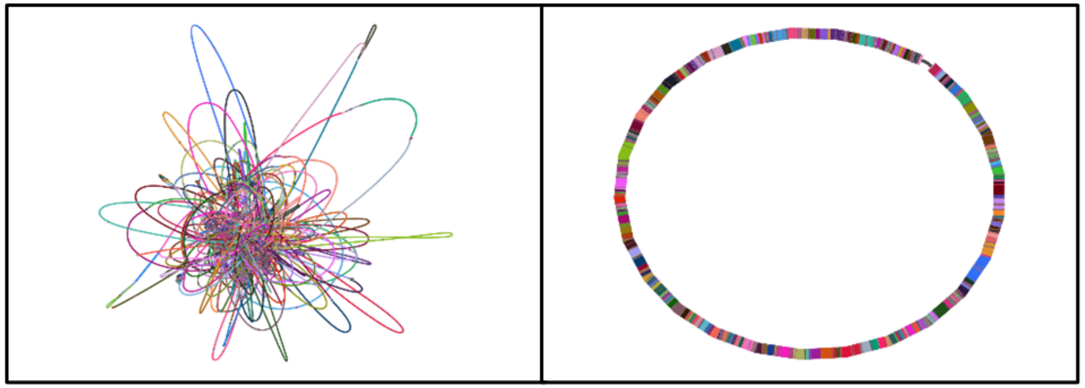
Completed hybrid assembly

Image reproduced from https://github.com/rrwick/Unicycler

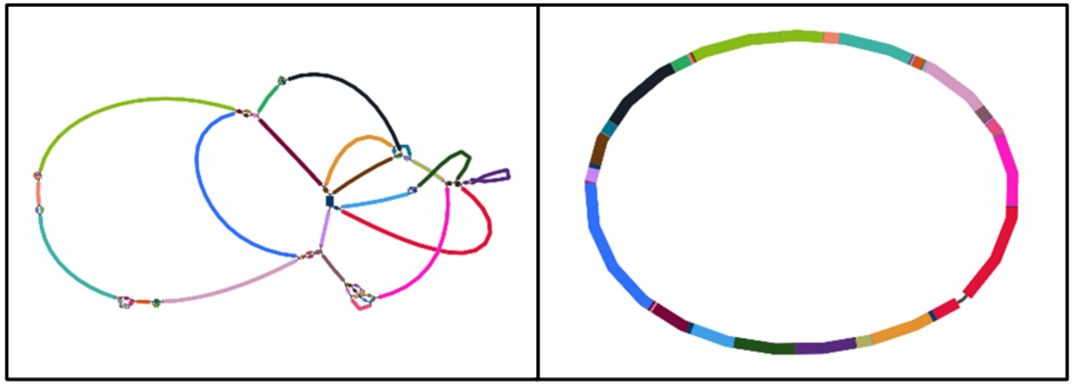
Advantage of hybrid assemblies

Illumina-only assembly | Hybrid assembly

Neisseria meningitidis (ATCC® 53417™)

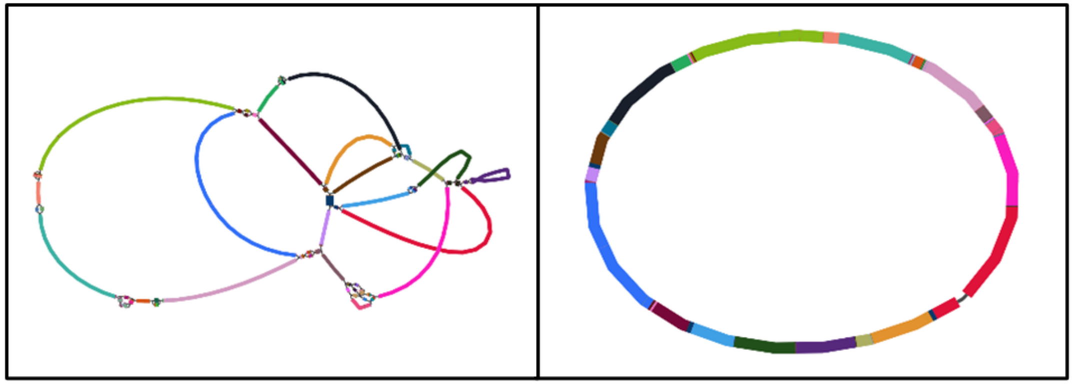


Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)

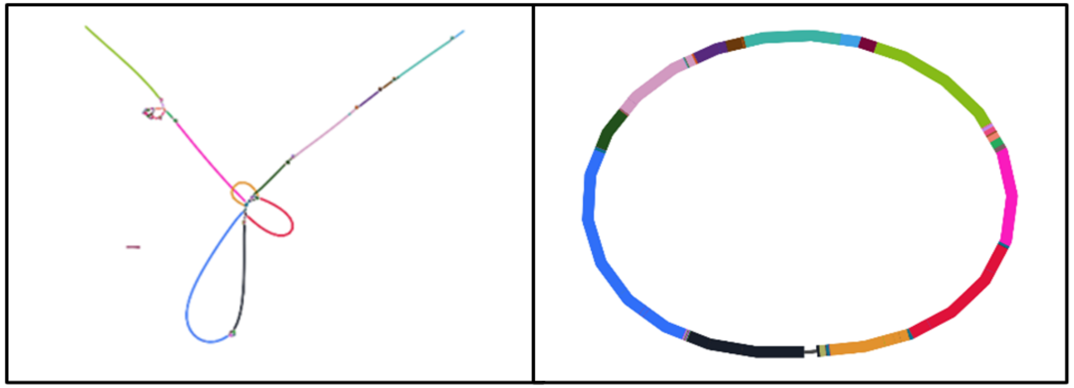


Illumina-only assembly | Hybrid assembly

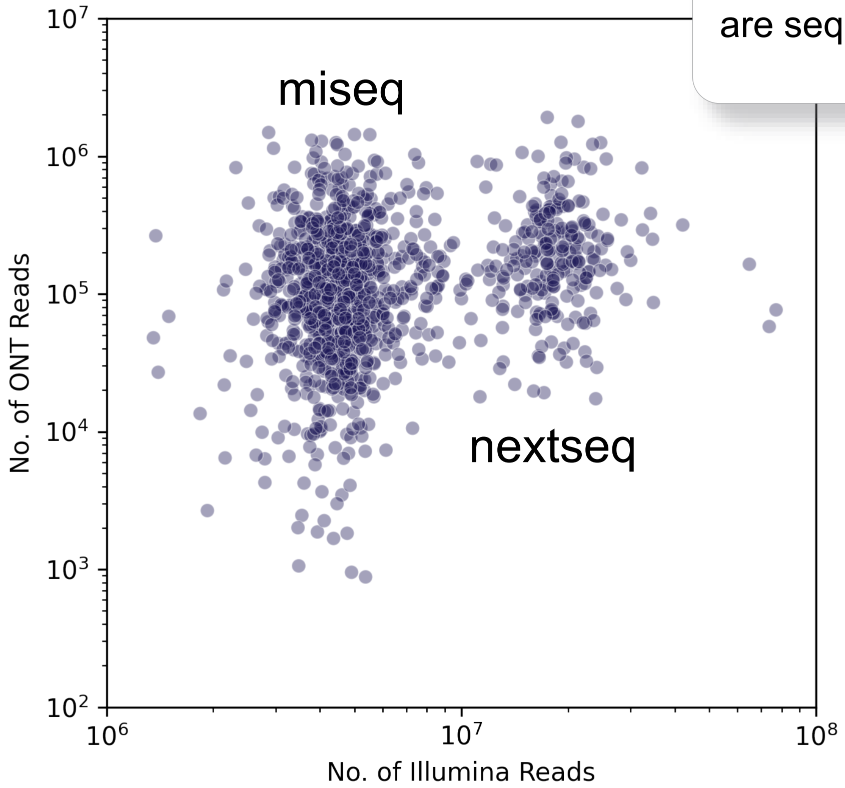
Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)



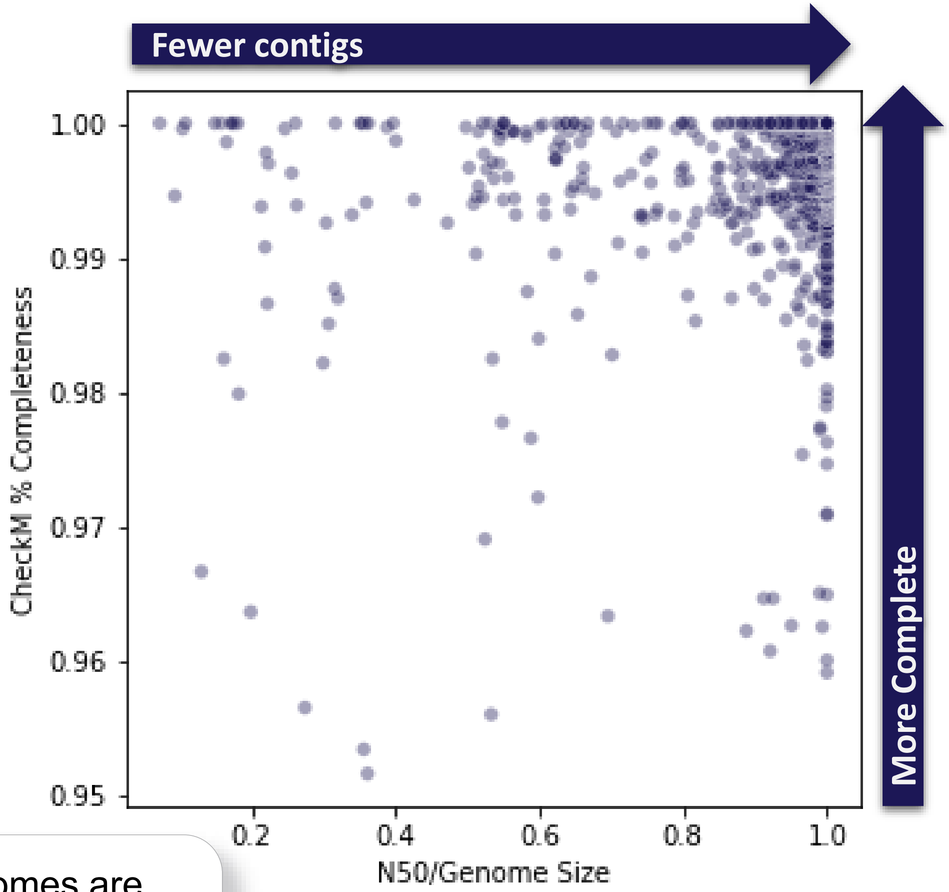
Streptococcus gordonii (ATCC® 35105™)



Quality of ATCC Genome Portal assemblies



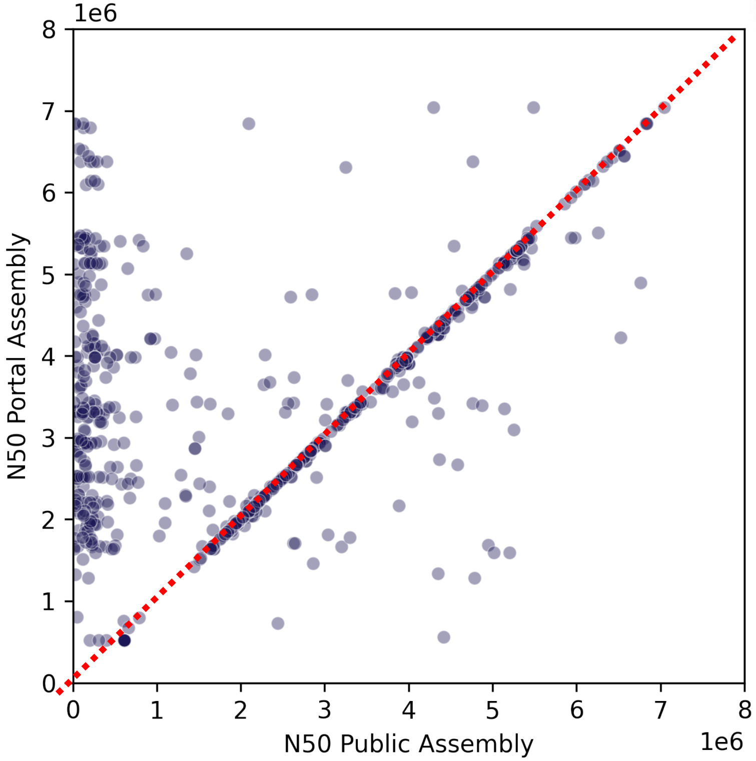
All bacterial and fungal genomes are sequenced on both Illumina and Oxford Nanopore

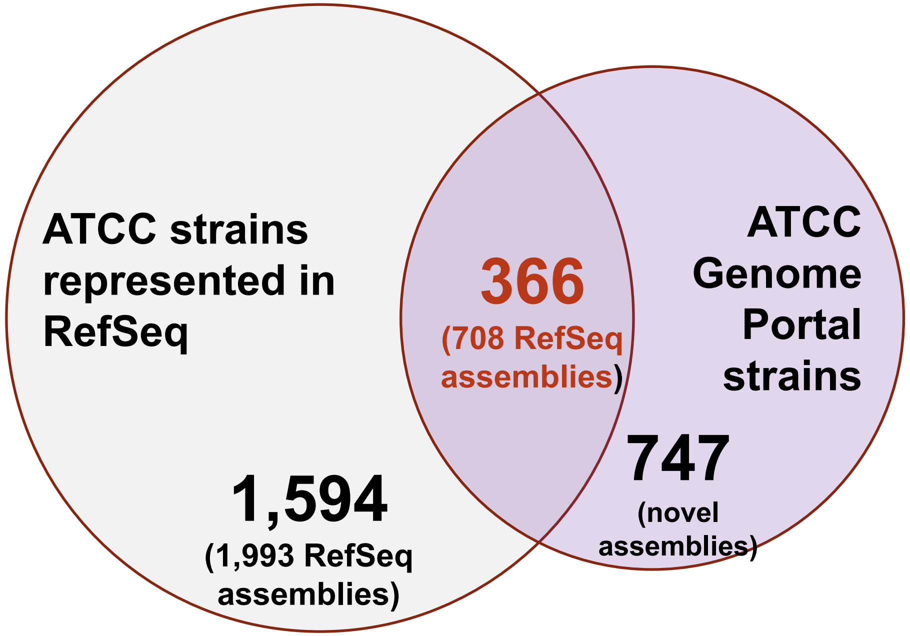


All genomes are required to be at least 95% complete (CheckM/Busco)

Comparison of ATCC vs. RefSeq bacterial assemblies

>98% of our assemblies are more complete and of higher quality than RefSeq





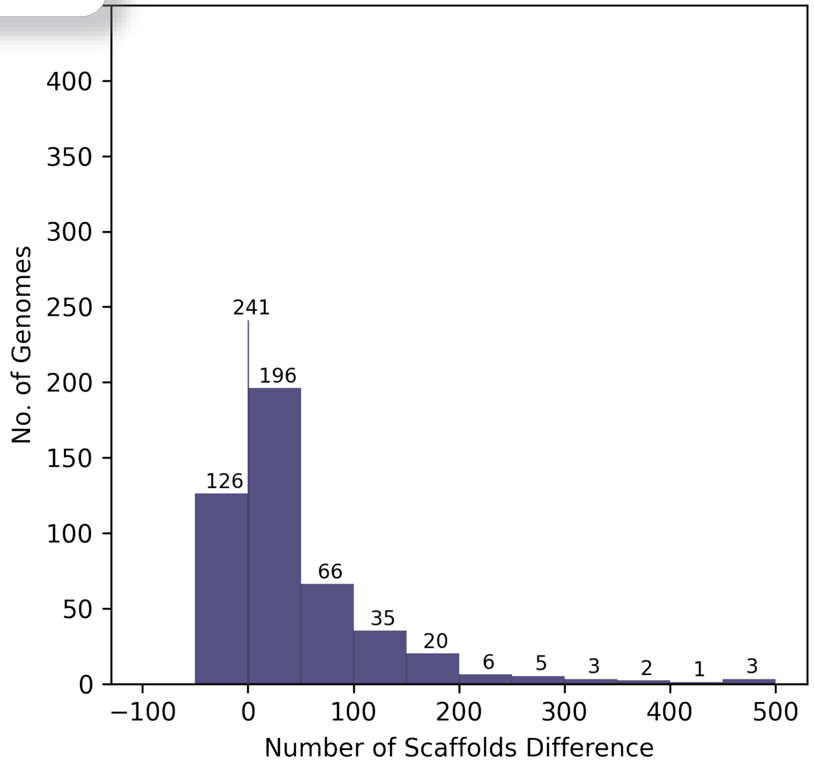
ATCC strains represented in RefSeq

366
(708 RefSeq assemblies)

ATCC Genome Portal strains

1,594
(1,993 RefSeq assemblies)

747
(novel assemblies)



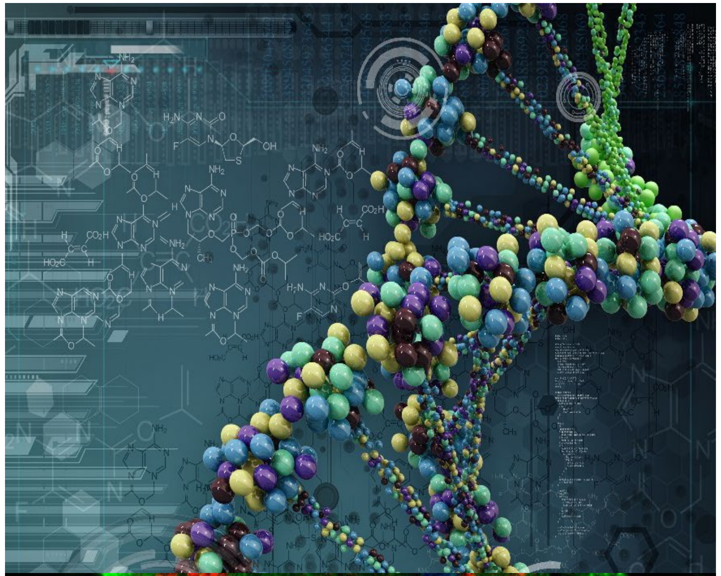
Yarmosh DA et. al. ***Comparative Analysis and Data Provenance for 1,113 Bacterial Genome Assemblies.*** *mSphere* **2022**, e00077-22. https://doi.org/10.1128/msphere.00077-22.

ATCC Cell Line Land

A partnership with QIAGEN Digital Insights

ATCC cell biology collection

ATCC has **3,000+ authenticated** mammalian cell lines, genetic engineered cell lines, primary cells, stem cells, iPSCs, hTERT-immortalized cells, and tumor organoids representing various species, cell types, tissues origins, and diseases.

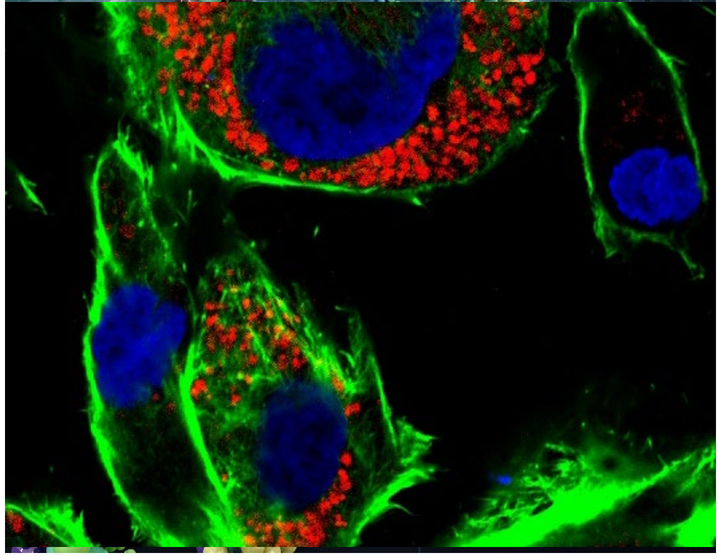


70+
Species

100+
Cell types

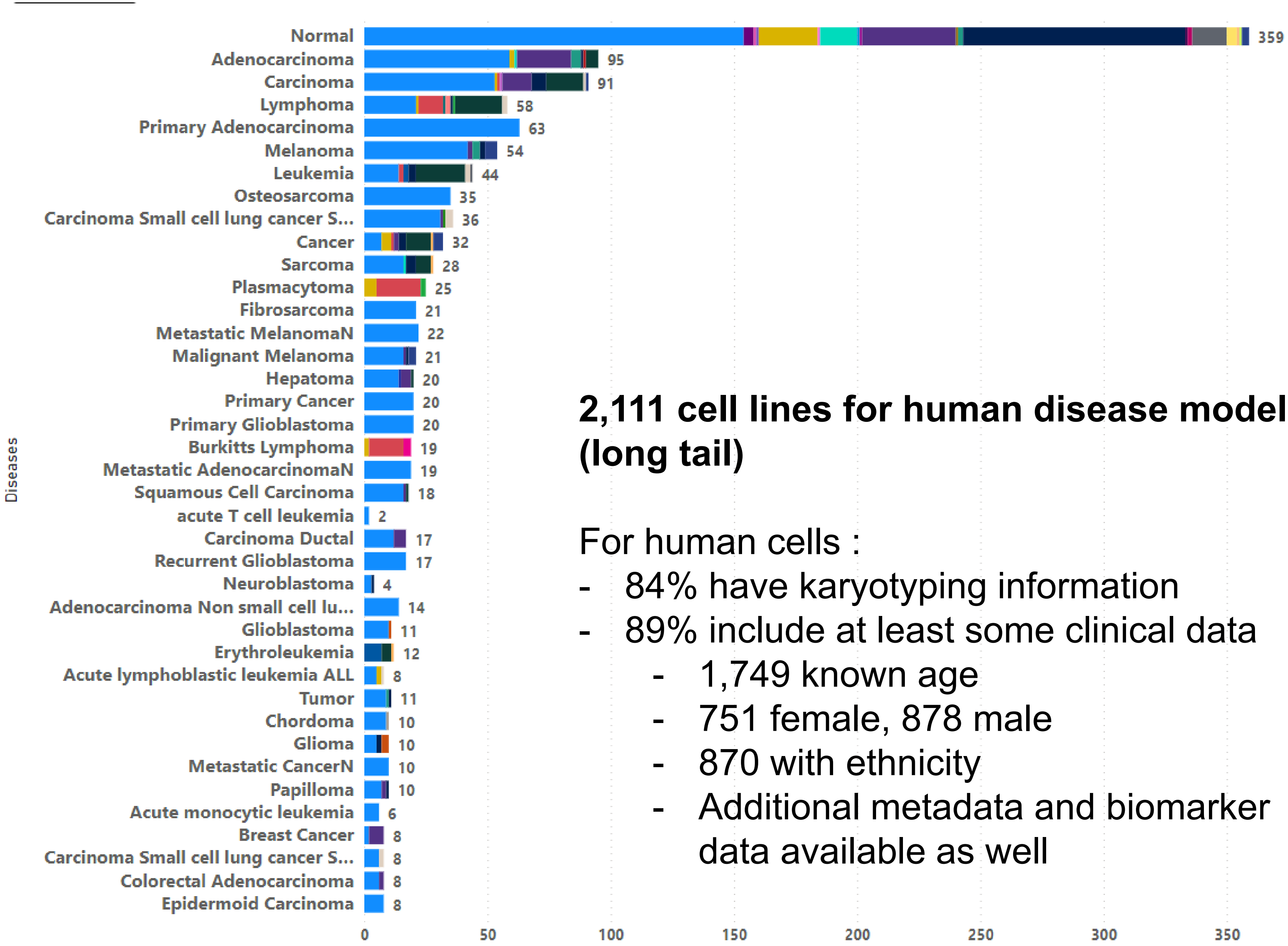
100+
Tissue types

400+
Diseases types



ATCC Cell Biology Collection (by disease type)

Cell line models for over 400 disease types



2,111 cell lines for human disease models (long tail)

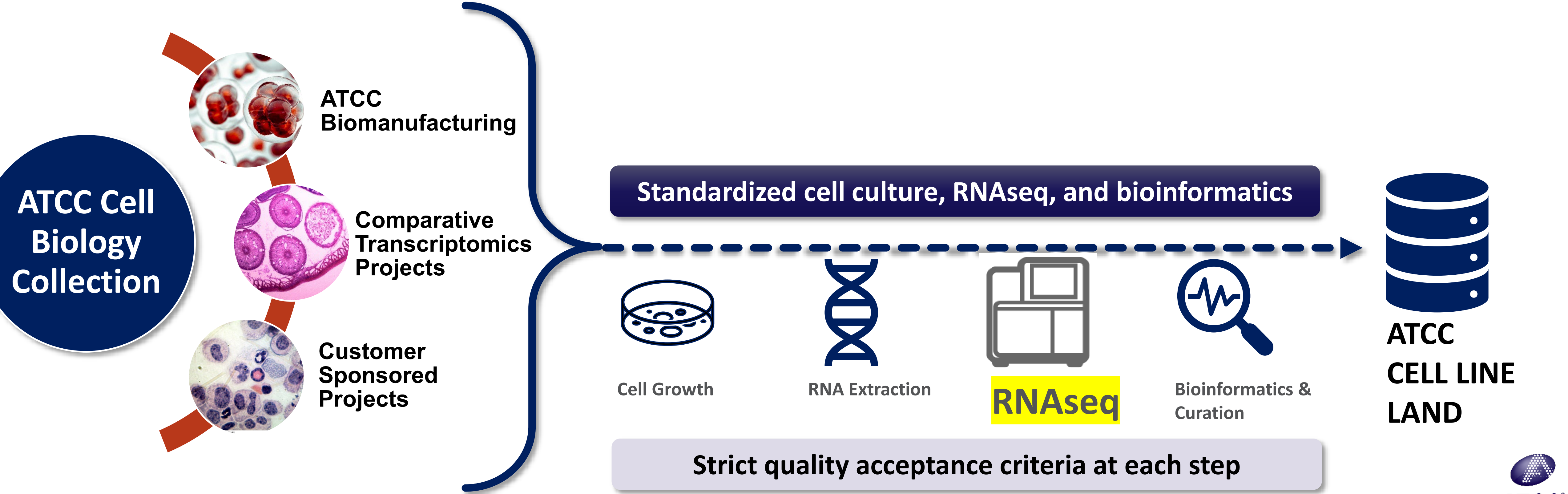
For human cells :

- 84% have karyotyping information
- 89% include at least some clinical data
    - 1,749 known age
    - 751 female, 878 male
    - 870 with ethnicity
    - Additional metadata and biomarker data available as well

ATCC Cell Line Land

KEY FEATURES

1. Repository of ***authenticated 'omics data traceable to physical materials***
2. Data production, curation, and analysis uniformly standardized
3. Enables the highest level of **scientific reproducibility**
4. End-to-end **data provenance**



ATCC Cell Biology Collection

ATCC Biomanufacturing

Comparative Transcriptomics Projects

Customer Sponsored Projects

Standardized cell culture, RNAseq, and bioinformatics

Cell Growth

RNA Extraction

RNAseq

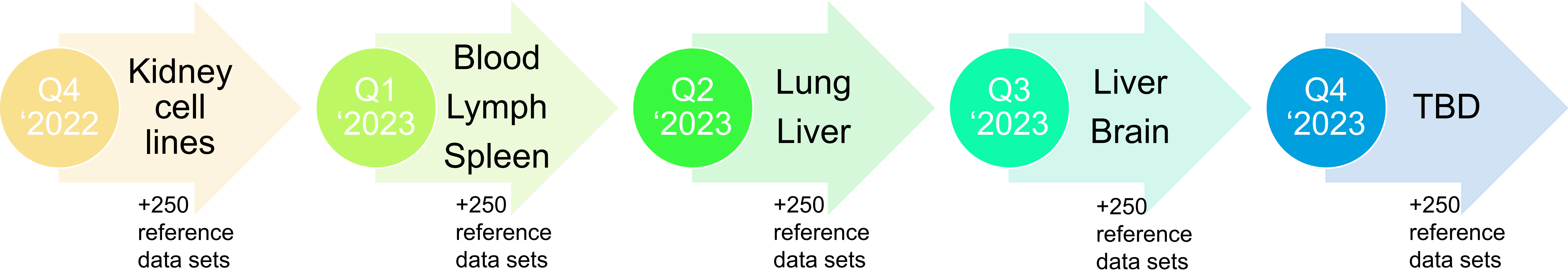
Bioinformatics & Curation

ATCC CELL LINE LAND

Strict quality acceptance criteria at each step

ATCC Cell Line Land

A partnership with QIAGEN Digital Insights



Q4 ‘2022 — Kidney cell lines — +250 reference data sets

Q1 ‘2023 — Blood Lymph Spleen — +250 reference data sets

Q2 ‘2023 — Lung Liver — +250 reference data sets

Q3 ‘2023 — Liver Brain — +250 reference data sets

Q4 ‘2023 — TBD — +250 reference data sets

- Current road-map for data production is subject to change
- Based on customer feedback

- **1,000+ traceable, authenticated RNAseq datasets per year**

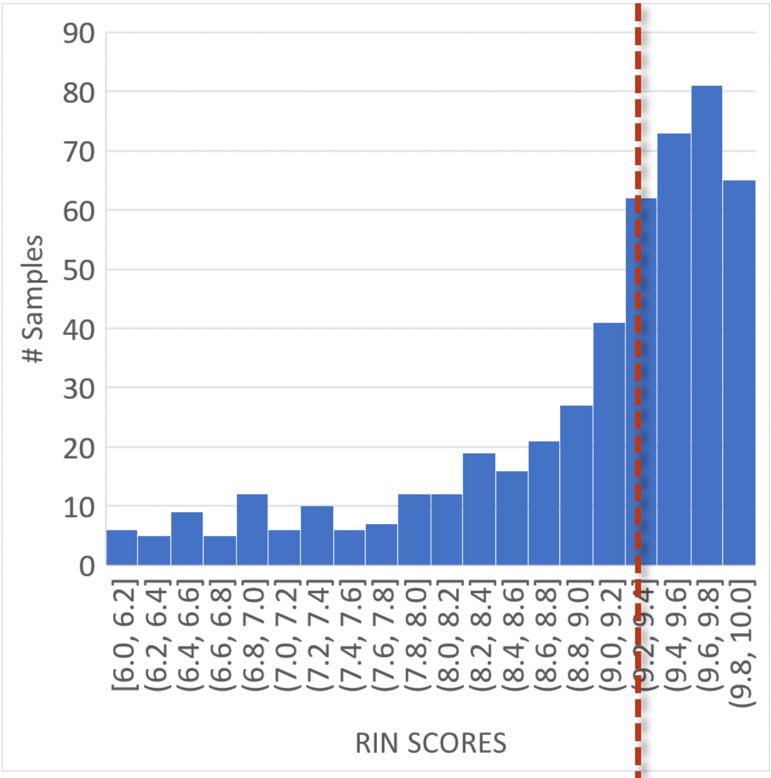
ATCC Cell Line Land

A partnership with QIAGEN Digital Insights

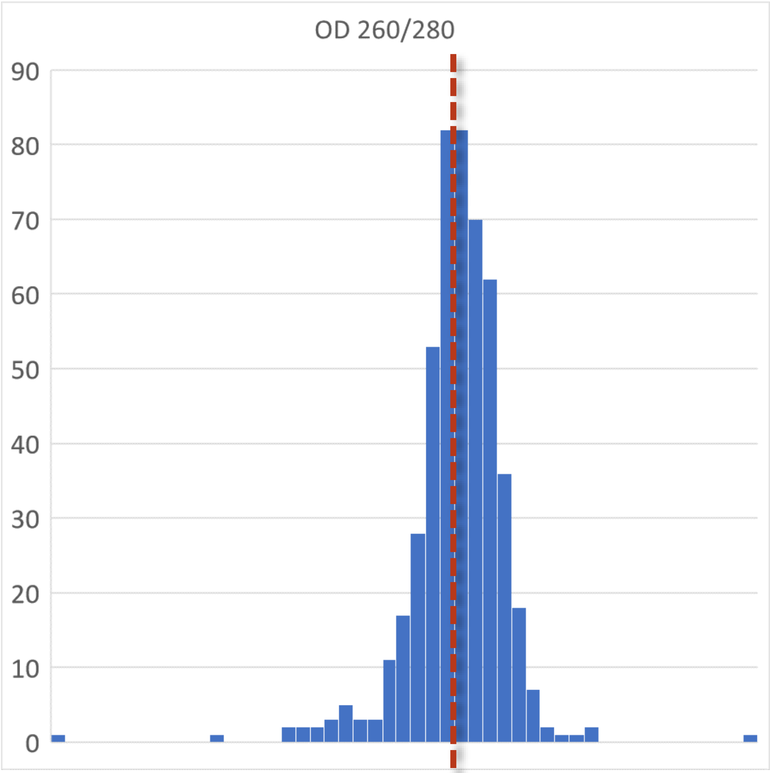
- Quality-controlled data from ATCC cell lines
- Over 1000 new datasets added each year, released quarterly
- Careful metadata curation with controlled vocabulary
- Reprocessed and normalized RNAseq expression
- Metadata include standard culture conditions, extraction protocols, sample preparation, and library preparation
- Data grows based on what you, as a researcher, need most:
  - Our team takes your requests to prioritize the cell lines you want added to our ATCC Cell Line Land collection, as well as the type of experimental data you want curated

ATCC Cell Line Land

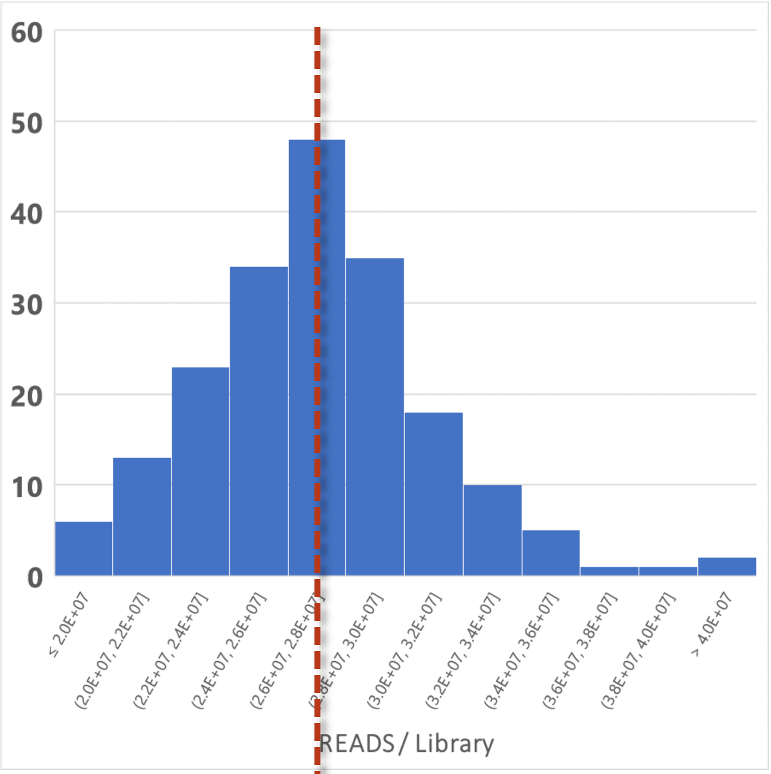
A partnership with QIAGEN Digital Insights



9.3 median RIN score



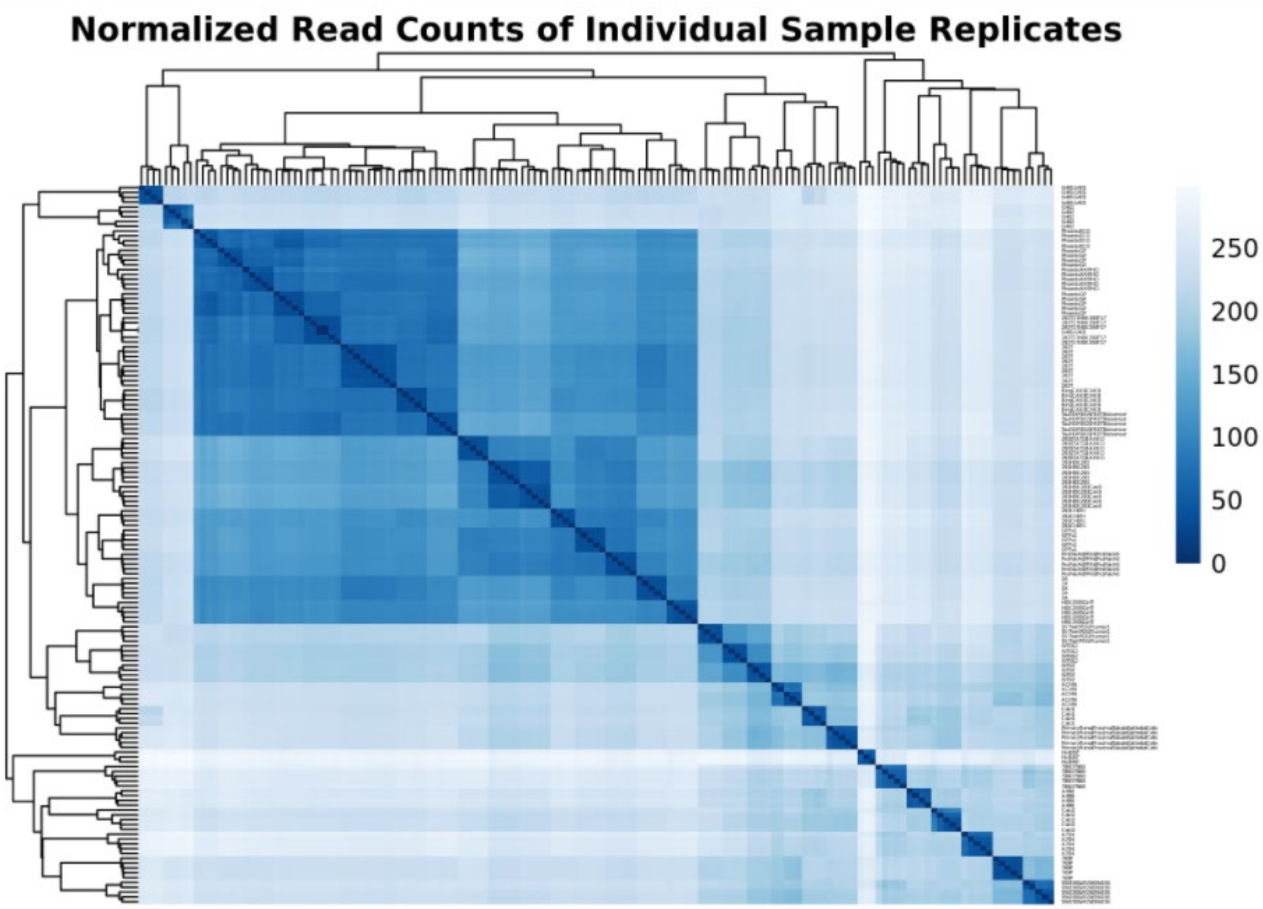
2.071 median OD260/280



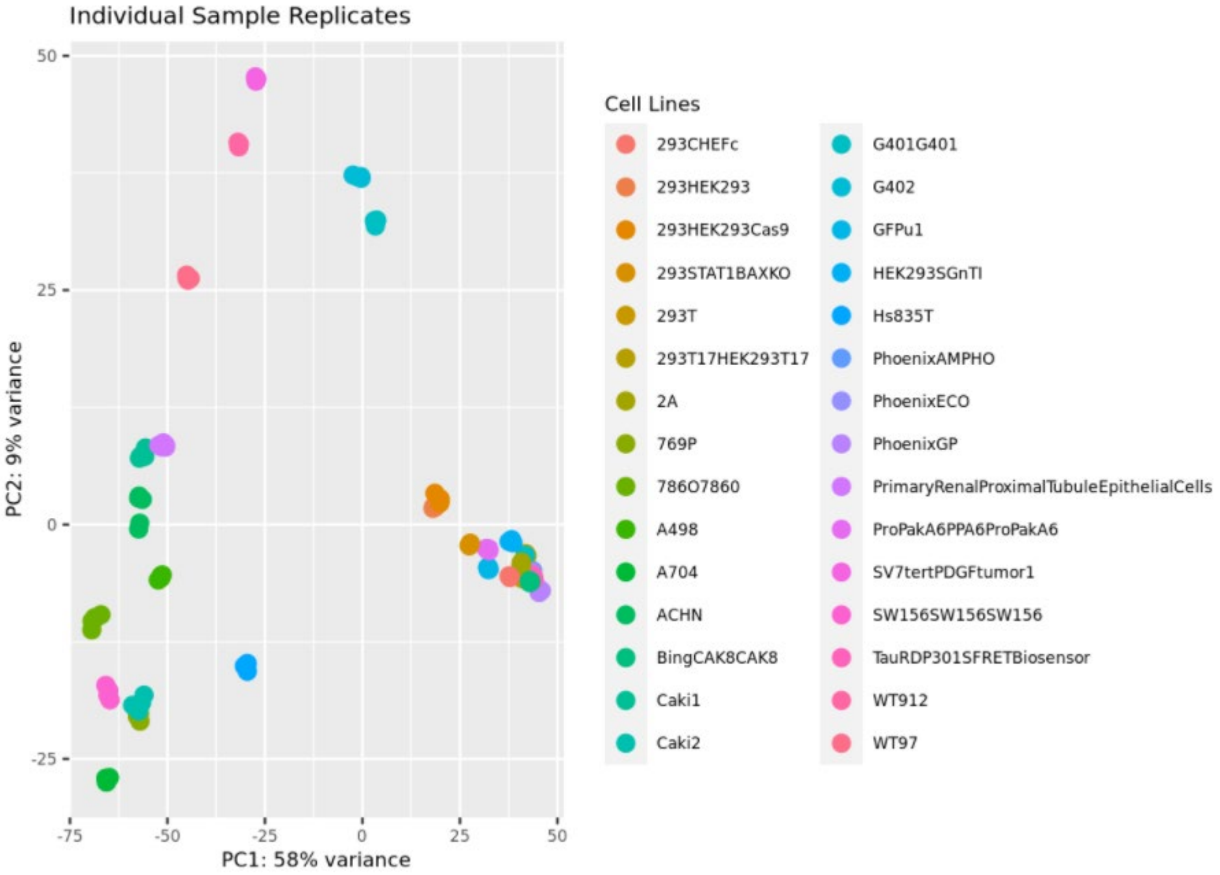
26.9M reads per library (median)

ATCC Cell Line Land – Example (kidney cell lines)

A partnership with QIAGEN Digital Insights



Normalized Read Counts of Individual Sample Replicates

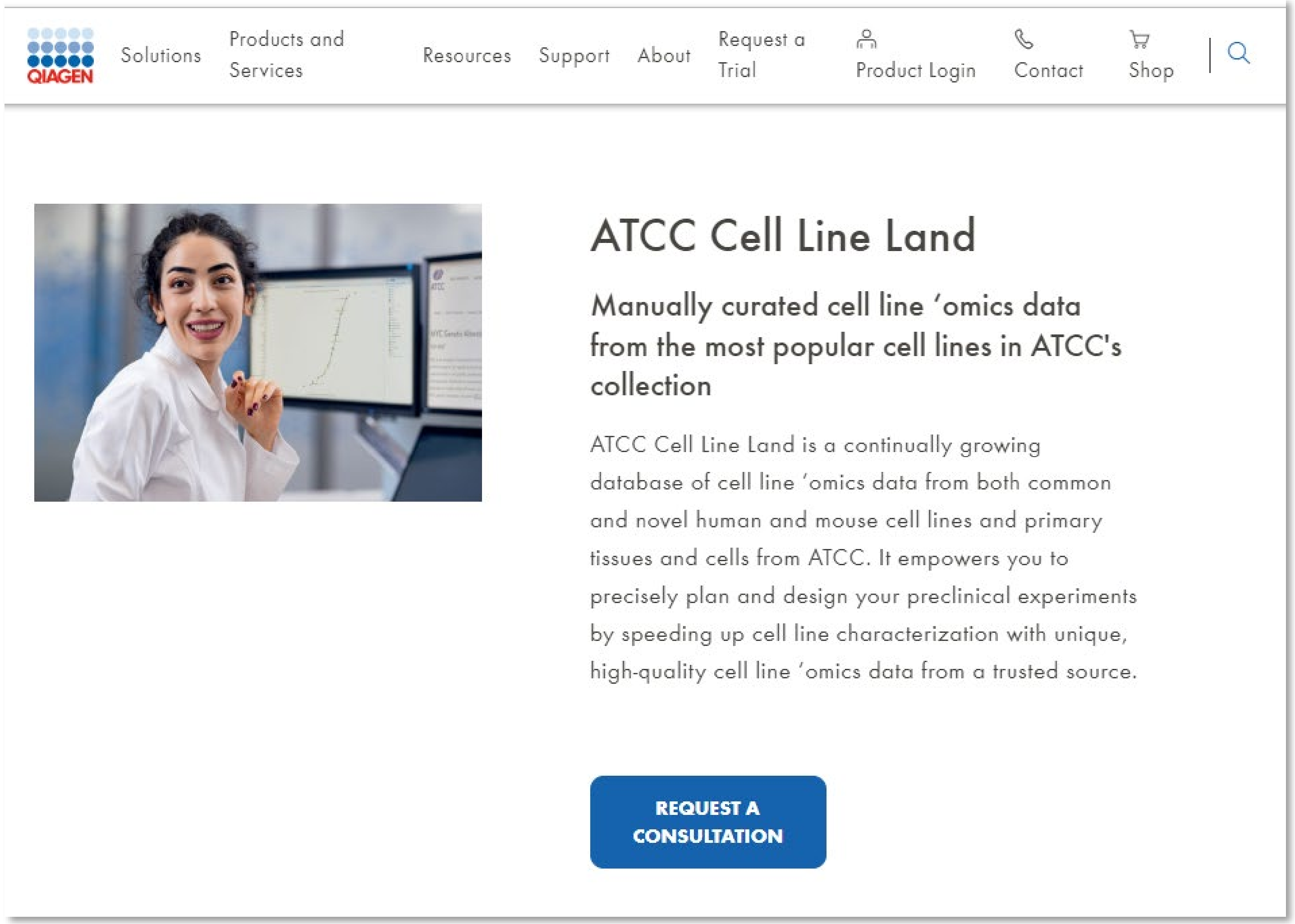


Individual Sample Replicates

Full data for over 60 kidney cell lines will be presented at the American Society of Cell Biology (December 2022)

ATCC Cell Line Land – Available through QIAGEN

A partnership with QIAGEN Digital Insights



ATCC Cell Line Land

Manually curated cell line ʻomics data from the most popular cell lines in ATCC's collection

ATCC Cell Line Land is a continually growing database of cell line ʻomics data from both common and novel human and mouse cell lines and primary tissues and cells from ATCC. It empowers you to precisely plan and design your preclinical experiments by speeding up cell line characterization with unique, high-quality cell line ʻomics data from a trusted source.

Currently includes **Authenticated RNAseq Data** for over 200 ATCC cell lines.

https://digitalinsights.qiagen.com/atcc-cell-line-land/

The ATCC Genomics Team

@ATCCgenomics

Jonathan Jacobs, PhD

Senior Director, Bioinformatics

BioNexus Principal Investigator

✉ jjacobs@atcc.org **_@bioinformer_**

<u>Genomics Team</u>	**<u>Bioinformatics Team</u>**
Briana Benton, PMP	**John Bagnoli**
Ana Fernandes	Scott Nguyen, PhD
Ajeey Singh, PhD	David Yarmosh, MSc
Stephen King, MSc	Nikhita Putheveetil, MSc
James Duncan, MSc	P. Ford Combs, PhD
Samuel Greenfield, MSc	Amy Reese, MSc
Corina Tabron, MSc	
Noah Wax, MSc	
Rula Khairi, MSc	
Robert Marlow	
Jade Kirkland	

PARTNERS

Marco Riojas (ATCC / BEI)

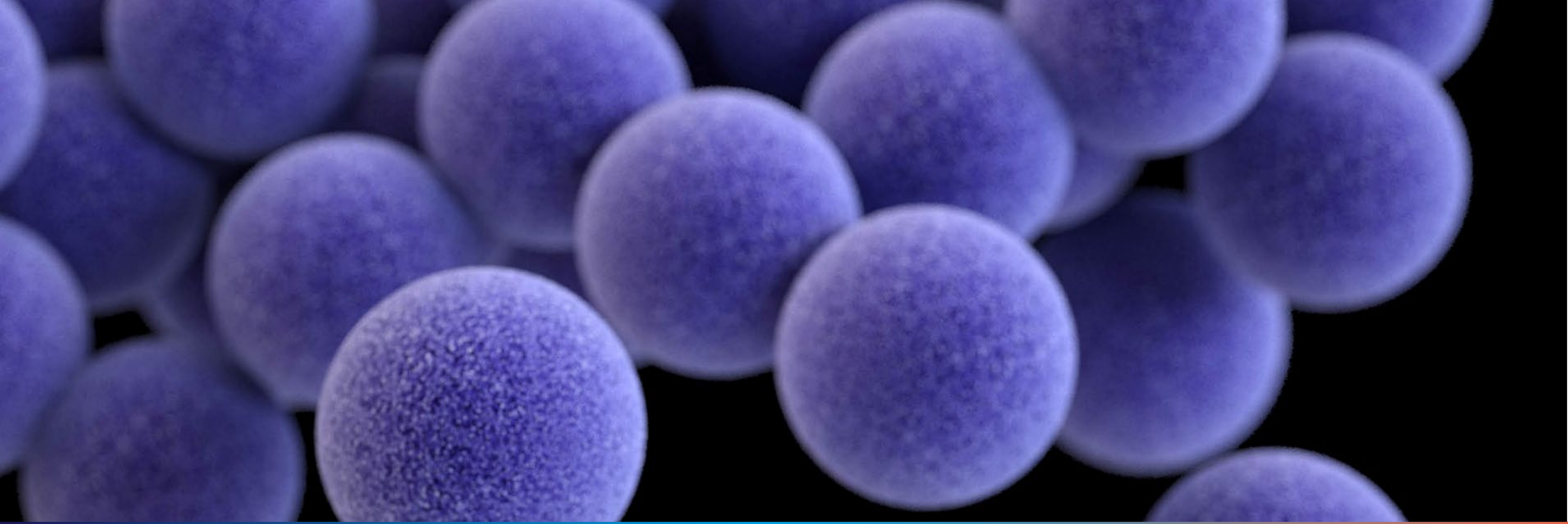
James Crill (Syracuse University)

Kylene Kehn-Hall (Virginia Tech)

Vishnu Chaturvedi (Wadsworth)

QIAGEN Digital Insights

One Codex



Thank you!

ATCC Genome Portal	https://genomes.atcc.org
ATCC Cell Line Land	https://digitalinsights.qiagen.com/atcc-cell-line-land/

ATCC®