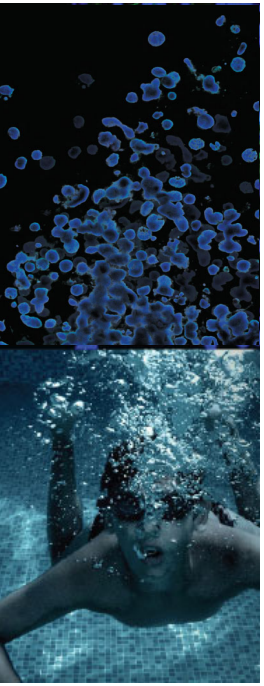
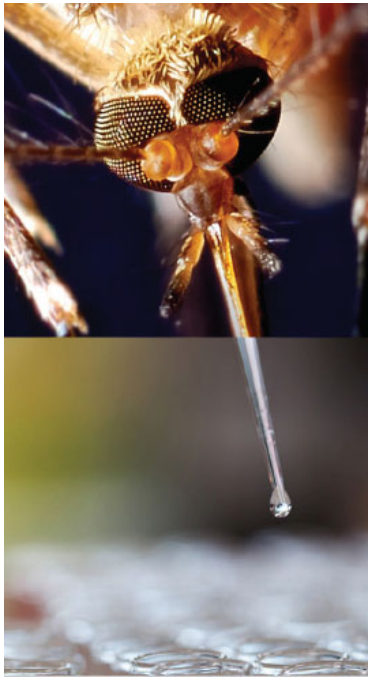
Genomic Data Quality:

Connecting the Dots Between Bioinformatics and Physical Materials



Jonathan Jacobs, PhD
Senior Director, Bioinformatics
BioNexus Principal Scientist
ATCC



Credible Leads to Incredible™

About ATCC…

- American Type Culture Collection founded in 1925
- *Non-profit institution with a mission to develop biomaterials, resources, and standards critical for life science research.*
- World's largest, most diverse biological materials and information resource for microbes and cell lines (BEI & ATCC)
  - 32,000 bacterial strains
  - 46,000 mycology strains
  - 11,000 human / animal cell lines
  - 5,300 virus strains
  - 3,400 protistology strains
- cGMP biorepository & biomanufacturing
- Global supplier of authenticated cell lines, microorganisms, and molecular standards
- Innovative R&D company focused on biomaterial and genome engineering, cell-based model systems development and cryopreservation technologies.
- Sales and distribution to 150+ countries, with 19 international distributors

https://www.atcc.org/about-us/what-we-do

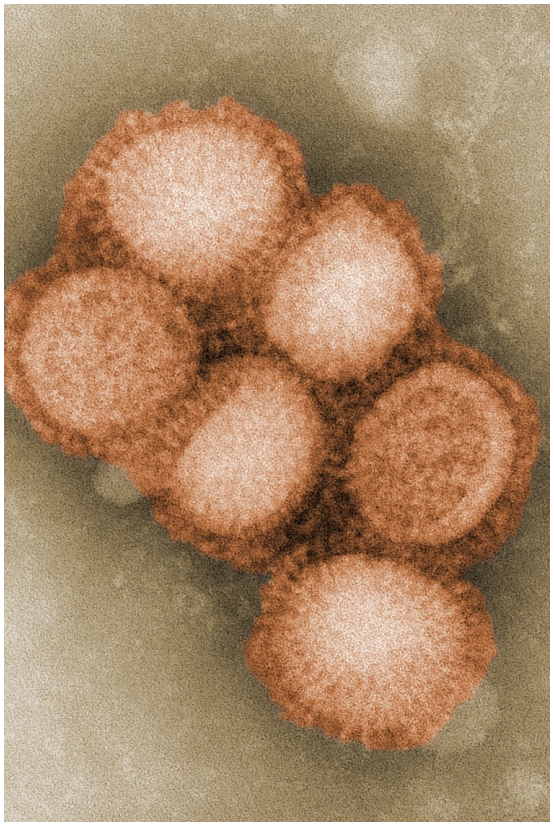
https://genomes.atcc.org

Microbiology Resources

A vast collection of microbial reference materials for molecular diagnostics

Well-characterized bacteria, viruses, fungi, protists, and derivatives

- **SARS-CoV-2 (and other pathogens) molecular Dx materials**
  - Genomic RNA from clinical isolates
  - Synthetic nucleic acids for use in BSL-1 facilities
  - Microbial strains for cross-reactivity testing
  - Heat-inactivated preparations for use in molecular assays or as a process control
- **Microbiome Standards**
  - Fully sequenced, characterized, and authenticated mock microbial communities
  - Mixed whole-cell or genomic material
  - Even or standard mixes
  - Bacteriome, virome, or mycobiome
- **Drug resistant bacterial strains**



www.atcc.org/microbes

ATCC

Custom Services

Providing secure & reliable biomaterials management, storage & distribution

Partner with the global biological resource leader

- **cGMP & cGTP Cell Banking**:
  - 21 CFR 600, 610, Good Manufacturing Practices (cGMP)
  - 21 CFR Part 1271, Good Tissue Practices (cGTP)
  - Mammalian and stem cells
  - Primary cell derivation and expansion
  - Custom-built, designated cell processing suites
  - Healthy cells and cells derived from diseased tissues
  - Master and working cell banks (MCB and WCB)
- **cGMP Biorepository**
  - ISO 9001:2008, cGMP-compliant
  - LN_2, -80°C, -20°C, and 2-8°C storage available
  - Cell, microbe, protein, and nucleic acid storage options
  - Cell line and microbe expansion (MCB & WCB)

www.atcc.org/cGMP

ATCC

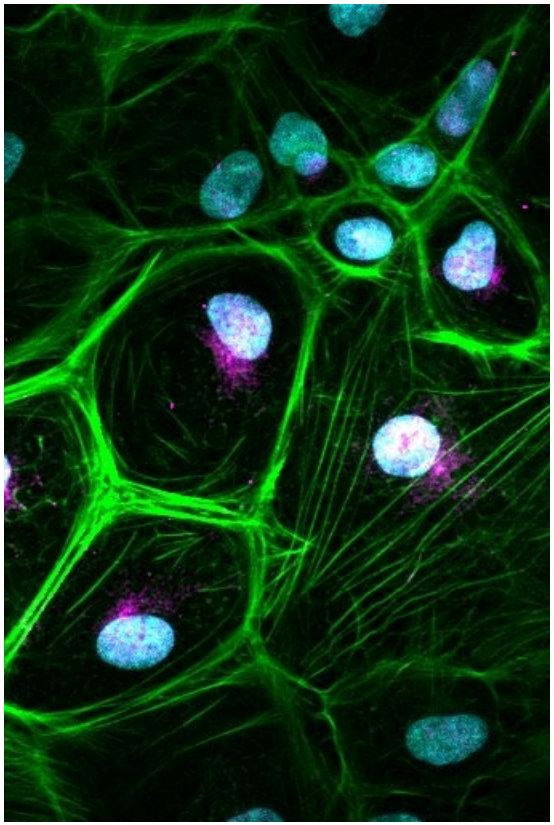
Cell Biology Resources

The world's largest and most extensive product catalog of human and animal cells

Authenticated cell lines, advanced models, and derivatives

- Reference samples for detecting somatic 2 mutations
  - Characterized triple-negative breast cancer cell line and its B lymphocyte-derived normal cell line
- Tumor/normal matched pairs
  - Matched normal and tumorigenic or metastatic cell lines
- Isogenic cell models that contain mutations in key oncogenes
  - KRAS G13D, NRAS Q61K, MEK 1Q56P, IDH1 R132H, IDH2 R140Q, and EML4/ALK fusion mutated cell lines available
  - Luciferase-labeled models for easy bioluminescence detection
- Epithelial/mesenchymal transition reporter cell lines for real time imaging of phenotypic transition
- Quantified cell line genomic DNA isolated from normal and tumor cell lines

www.atcc.org/cancer



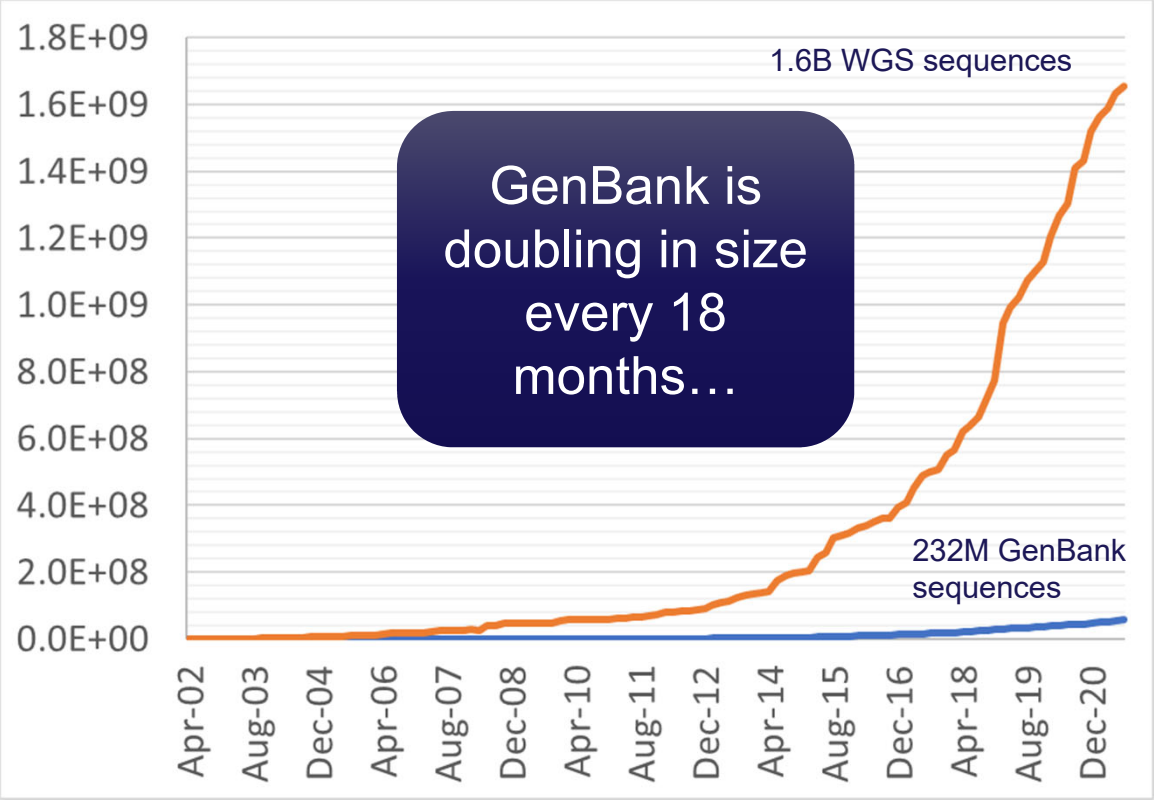
ATCC

Today's Focus

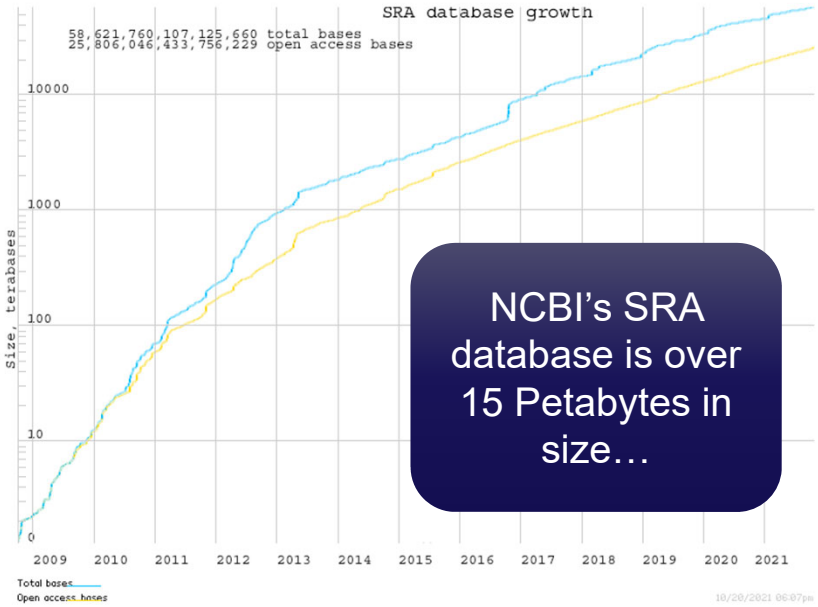
- Open questions on genomics data quality
- Examples
- Why it matters
- What can you do about it
- What are we doing about it at ATCC

First – a reminder on the growth of GenBank –

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…

How sustainable is the growth of GenBank?

How reliable is the data in GenBank?

How do you build trust in a data ecosystem this large?

- “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 from the publication to the [data] repository.” [2]

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

1. 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.

2. 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.

3. 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.

ATCC

#1: Fake data was first deposited into GenBank in 1995

"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)*

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 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

24 years later, it's still being cited…

Received: 25 March 2021 | Revised: 16 June 2021 | Accepted: 13 July 2021
DOI: 10.1002/humu.24261

REVIEW

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Pathogenic noncoding variants in the neurofibromatosis and 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 M13 9WL, UK.
Email: miriam.smith@manchester.ac.uk

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

…atosis are a group of …nent of nerve sheath tumors. Pathogenic variants (PVs) that cause NF1 and NF2 are located in the *NF1* and *NF2* loci, respectively. To date, most variants associated with schwannomatosis have been identified in the *SMARCB1* and *LZTR1* genes, and a missense variant in the *DGCR8* gene was recently reported to predispose to schwannomas. In spite of the high detection rate for PVs in NF1 and NF2 (over 90% of non-mosaic germline variants can be identified by routine genetic screening) underlying PVs for a proportion of clinical cases remain undetected. A higher proportion of non-NF2

…tion is identified …e in the project is …Experience. The …s the qualifying …anization to …icant's ability to …ently administer …lication specifically …nt as a nationally- …ion, institution, or …rd of study and …special …Previous specific …s similar to the …early and …d. The relationship …and other work …, or underway by …ibed, including a …elated Federal …within the last five …nsortium of …d, the project …work should be …us Federal …d by project …cy, and grants of …5 points

…mplete Application …ation consists of the …is 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

ATCC

And after 42 citations… the data is still in GenBank…

DNA Sequences in the Promoter

https://www.sciencedirect.com/science/article/abs/pii/S08887543...

View PDF | Access through your institution | Purchase PDF | Search ScienceDirect

Article preview

Abstract

Cited by (42)

Recommended articles (6)

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. Tarle, Matthew Freedman, Susan Wilson-Gunn, Andre Bernards, Francis S. Collins

Show more

+ Add to Mendeley | Share | Cite

https://doi.org/10.1006/geno.1994.1328 | Get rights and content

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) ge

https://www.ncbi.nlm.nih.gov/nuccore/U17084

GenBank | Send to:

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

GenBank: U17084.1

FASTA | Graphics

Go to:

```
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
```

Change region shown

Customize view

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

ATCC

#2: Falsified sequencing 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]

Show more

Outline | + Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.bse.2021.104

Highlights

- This manuscript pre
mitochondrial genor
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
```

NCBI staff labeled this as “Unverified”, but the sequence still remains in GenBank…

#3: 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 Amrit

Abstract

With the rapid growth and development of sequencing technologies, genomes have becom exploring solutions to some of the world's biggest challenges such as searching for alternativ exploration of genomic dark matter. However, progress in sequencing has been accompanie that can occur during template or library preparation, sequencing, imaging or data analysis. screened over 18,000 publicly available microbial isolate genome sequences in the Integrate database and identified more than 1000 genomes that are contaminated with PhiX, a contro during Illumina sequencing runs. Approximately 10% of these genomes have been publishe contaminated genomes were sequenced under the Human Microbiome Project. Raw sequen contamination from various sources and are usually eliminated during downstream quality c of PhiX contaminated genomes indicates a lapse in either the application or effectiveness of measures. The presence of PhiX contamination in several publicly available isolate genomes errors when such data are used in comparative genomics analyses. Such contamination of p far-reaching consequences in the form of erroneous data interpretation and analyses, and ne measures to proofread raw sequences before releasing them to the broader scientific comm

Keywords: Next-generation sequencing, PhiX, Contamination, Comparative genomics

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 catalyzed by the development of several next-generation sequencing platforms offering rapid and accurate genome information at a low cost. Among the different NGS technologies available commercially, the sequencing by synthesis technology [6] championed by Illumina [7] is the most widely used.

Despite its high accuracy, t platform does come with its sha need to be addressed by the u One such challenge is the pro used as a quality and calibratio runs. PhiX is an icosahedral, with a single-stranded DNA. It 5386 nucleotides and was the f sequenced by Fred Sanger [9]. defined genome sequence, Phi used as a control for Illumina s majority of its library preparatio using PhiX at a low concentrati raised up to 40% for low diver on the concentration of PhiX u the same lane along with the sa ate lane. Addition of PhiX as a cessitates subsequent quality co sequences such that they do not the target genome.

*Correspondence: supratimmukherjee@lbl.gov
[1]DOE Joint Genome Institute, Walnut Creek, CA, USA
Full list of author information is available at the end of the article

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

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

Keywords: Genomes, Contamination, Software, RefSeq, GenBank

Introduction

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,
artment of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore,
artment of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;
Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

that appear in published genomes can cause numerous problems for downstream analyses, partic- studies and metagenomics projects. Our large-scale scan of complete and draft bacterial and archaeal efSeq database reveals that 2250 genomes are contaminated by human sequence. The contaminant rily from high-copy human repeat regions, which themselves are not adequately represented in the e genome, GRCh38. The absence of the sequences from the human assembly offers a likely expla- e in bacterial assemblies. In some cases, the contaminating contigs have been erroneously annotated oding sequences, which over time have propagated to create spurious protein "families" across mul- tikaryotic genomes. As a result, 3437 spurious protein entries are currently present in the widely rotein databases. We report here an extensive list of contaminant sequences in bacterial genome as- ns associated with them. We found that nearly all contaminants occurred in small contigs in draft s that filtering out small contigs from draft genome assemblies may mitigate the issue of contami- g nearly all of the genuine genomic sequences.

is available for this article.]

e number of publicly available ge- a handful of species to well over e genomes are pivotal resources for questions, including microbiome ntify species in complex samples , all genomes in reference databases e (Fraser et al. 2002), but for practi- of genomes available today are still sts of multiple contigs or scaffolds nd not assigned into chromosomes e is not truly complete or "finished" determined for every chromosome th no gaps. Even the human ge- plete than most other animal ge- The current human assembly, , 2019), has 473 scaffolds that con- most of the human sequence has some highly repetitive regions are et al. 2014), leading to problems nomes of other species vary widely y, with some having thousands of uch smaller number. e assemblies with sequences from mon, especially in draft genomes et al. 2014; Delmont and Eren 2016; Lu and Salzberg 2018). In over 10% of selected nonprimate

hw@gmail.com, salzberg@jhu.edu
Article, supplemental material, and publi- me.org/cgi/doi/10.1101/gr.245373.118.

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 always present in the environment of sequencing laboratories, human contamination is very common in sequencing experiments of all types. Contamination of laboratory reagents with DNA from other organisms can also lead to serious misinterpretations, such as the supposed detection of the novel virus

ATCC

#4: 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 decision on June 11, 2020; accepted on June 16, 2020

Abstract

Motivation: As the cost of sequen… of data being deposited into public repositories is increasing rapidly. Public databases … ata for each submission that is prone to user error. Unfortunately, most public d… rely on user input and do not have methods for identifying errors in the pr… error propagation. Previous research on a small subset of the NR databas… e similarity. To the best of our knowledge, the amount of misclassificati… d. We propose a heuristic method to detect potentially misclassified ta… plied a curation technique and quality control to find the most probab… s provenance and frequency of each annotation from manually and … information at 95% similarity.

Results: We found mor… classified proteins in the NR database. Using simulated data, we show… or detecting taxonomically misclassified proteins. The proposed appr… ses.

**Availability and implemen… notebooks and Docker container are available at https://github…

Contact: hbagheri@iastate.edu

Supplementary information: Suppl…

~7.8% of genomes misclassified at the species level

~4% at the genus level

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 non-curated (low quality) and curated (high quality) databases. It contains NR sequences from GenBank translations (i.e. GenPept) together with sequences from other databases [Refseq (Pruitt *et al.*, 2007), PDB (Berman *et al.*, 2003), SwissProt (Boeckmann *et al.*, 2003), PIR (Wu *et al.*, 2003) and PRF]. NR removes 100% identical sequences and merges the annotations and sequence IDs.

We have identified three root causes for annotation errors in the public databases: user metadata submission, contamination error in … a for DNA sequences were deposited … ng soybeans obtained from a soy… ied to that metadata will be labeled … *ine max*. If the researcher had in fact … *soja* then this would result in a misassignment of … *max* sequences.

The second key challenge that all large databases have is the issue of contamination (Schnoes *et al.*, 2009). For example, if the aforementioned hypothetical soybean research did not remove the soybean root nodules during sample processing, then the tissue sample could also contain DNA from Nitrogen fixating soil bacteria that infect nodules leading to contamination of the sequences and ultimately the sequence database. NCBI is aware of the potential for

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 this project are available at https://github.com/zreichlab/LargeScaleKmerAnalysis. All relevant data and code come from public repositories and

Abstract

Information theoretic approaches are ubiquitous and effective in a wide variety of bioinformatics applications. In comparative genomics, alignment-free methods, based on short DNA words, or *k*-mers, are particularly powerful. We evaluated the utility of varying *k*-mer lengths for genome comparisons by analyzing their sequence space coverage of 5805 genomes in the KEGG GENOME … nt analyses on four k-mer lengths spanning the relevant range (… tering of 1634 genus-level representative genomes usin… larities best recapitulated a phylogenetic/taxonom… rkingdom domains and high subtree similarity … gh phylum). By analyzing ~14.2M prokaryot… n-ancestor taxon levels, we detected man… database, further demonstrating the ne… mic classifications based on whole-geno…

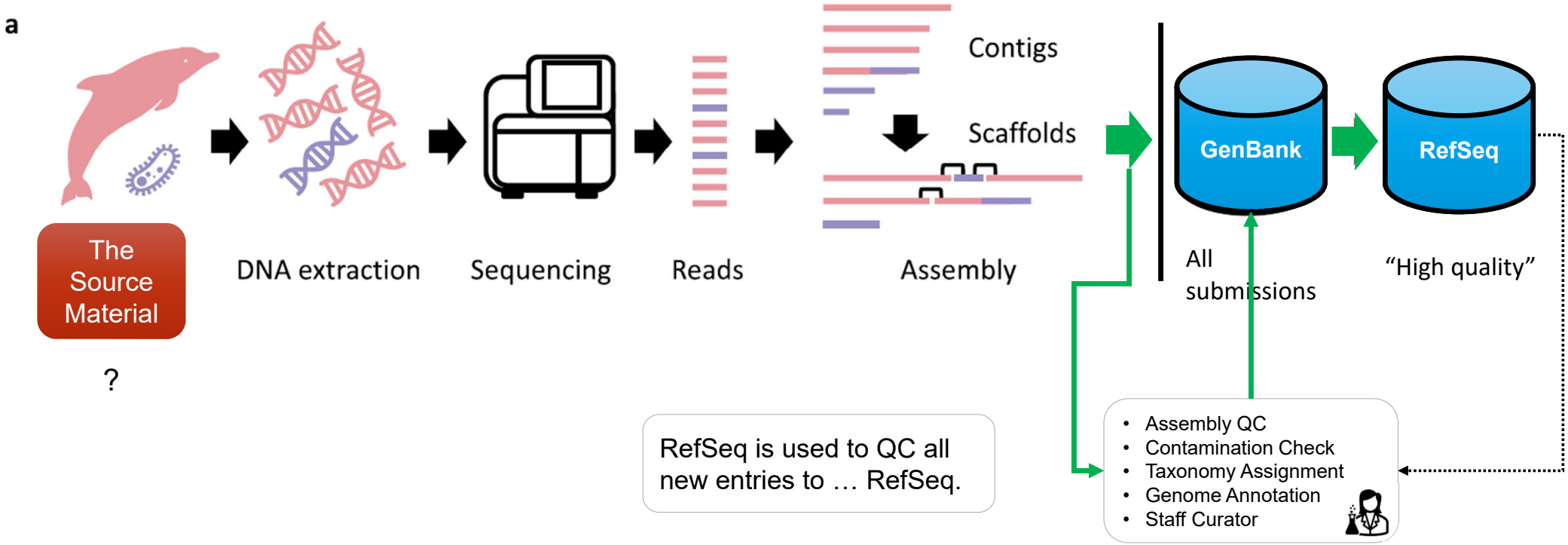
~7% of genomes misclassified at genus or higher

Introduction

Information theory, initially developed for the mathematical analysis of communicatio… tems by Shannon [1], has been applied to molecular biology for decades. Gatlin's pione… works in the late 1960s were the first to define life as an information processing system… Since then, information-theoretical approaches have been used in a wide variety of bio… sequence analyses, such as in the study and prediction of protein structure, protein-pro… interactions, transcription factor binding motifs, gene identification, as well as for seq…

ATCC

A Genomic Catch 22



Adapted from: Steinegger, M. and Salzberg, S.L. (2020) 'Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank', Genome Biology, 21(1), p. 115. doi:10.1186/s13059-020-02023-1.

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

ATCC

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…

ATCC

Open questions…

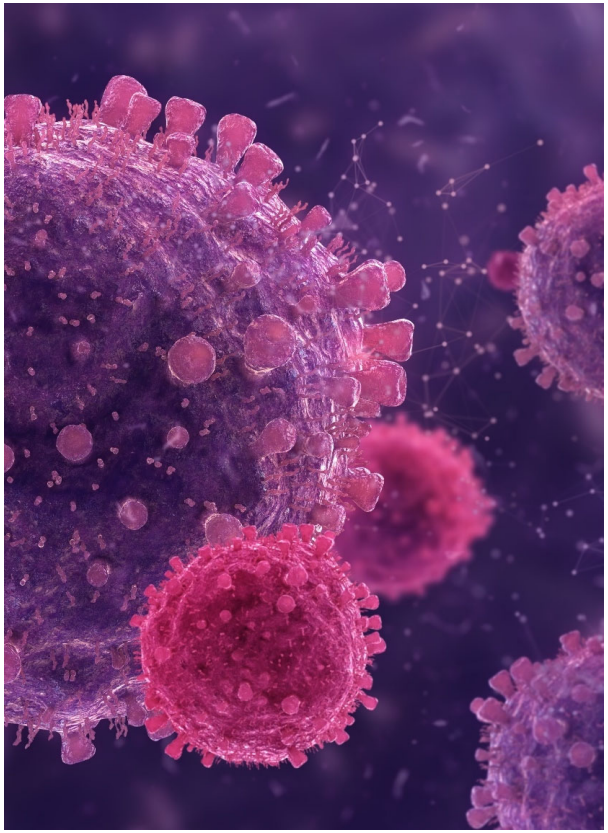
What is the *cost* of poor-quality data in public genome databases?

What are the consequences for sharing "bad data"?

What can you do?

ATCC

4 Ways to Improve the Quality of your Genomics Research

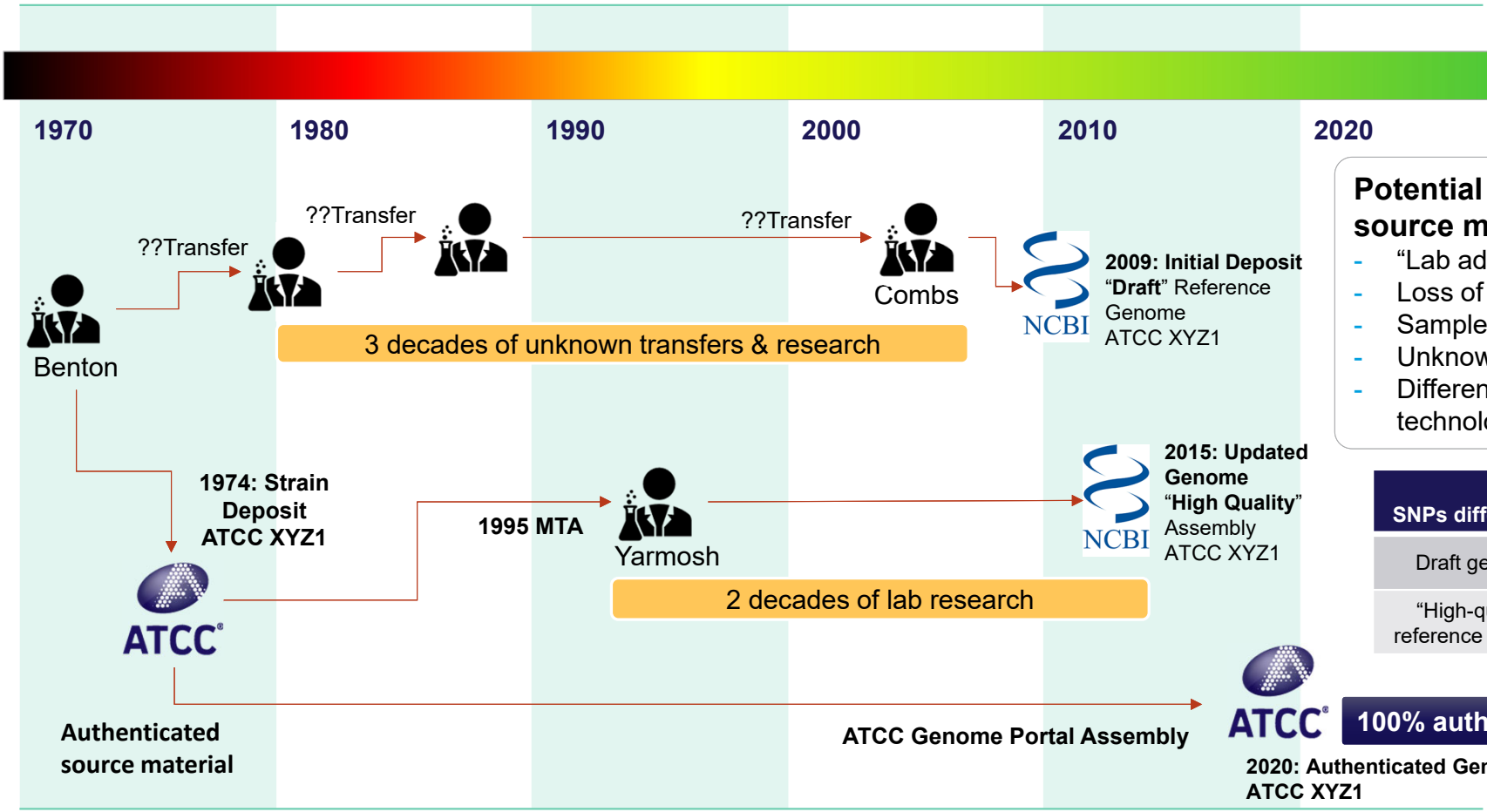


1. **Trust but Verify**
   - Use authenticated source materials whenever possible.
   - Be curious and investigate origins of data from outside your lab.
2. **Be a Standards Champion**
   - Know what material or data standards are available for you.
   - Get involved in defining new ones.
3. **Assume It's Dirty**
   - Data is rarely "clean".
   - Public data is often not "correct" and almost never "perfect".
4. **Adopt a "Digital First" Mindset**
   - Involve bioinformatics and data science early, not "after the data is produced".
   - Standardize your pipelines ahead of time.
   - Have an eye for quality and reproducibility.

ATCC

Trust but Verify

Use authenticated source materials whenever possible.



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

Names Changed To Protect The Innocent.

Champion Standards

Know what standards are available for you. Get involved in defining new ones.

PERSPECTIVE

nature biotechnology

"Source material identifier" is an exception; the *GSC recommends this be a core descriptor*, *but as yet, physical archives are not yet routinely created for all cases or types of biological material subjected to genome sequencing …*

The minimum information about a genome sequence (MIGS) specification

Dawn Field*[1], George Garrity[2], Tanya Gray[1], Norman Morrison[3,4], Jeremy Selengut[5], Peter Sterk[6], Tatiana Tatusova[7], Nicholas Thomson[8], Michael J Allen[9], Samuel V Angiuoli[5,10], Michael Ashburner[11], Nelson Axelrod[5], Sandra Baldauf[12], Stuart Ballard[13], Jeffrey Boore[14], Guy Cochrane[6], James Cole[2], Peter Dawyndt[15], Paul De Vos[16,17], Claude dePamphilis[18], Robert Edwards[19,20], Nadeem Faruque[6], Robert Feldman[21], Jack Gilbert[9], Paul Gilna[22], Frank Oliver Glöckner[23], Philip Goldstein[24], Robert Guralnick[24], Dan Haft[5], David Hancock[3,4], Henning Hermjakob[6], Christiane Hertz-Fowler[8], Phil Hugenholtz[25], Ian Joint[9], Leonid Kagan[5], Matthew Kane[26], Jessie Kennedy[27], George Kowalchuk[28], Renzo Kottmann[23], Eugene Kolker[29–31], Saul Kravitz[5], Nikos Kyrpides[32], Jim Leebens-Mack[33], Suzanna E Lewis[34], Kelvin Li[5], Allyson L Lister[35,36], Phillip Lord[35], Natalia Maltsev[20], Victor Markowitz[37], Jennifer Martiny[38], Barbara Methe[5], Ilene Mizrachi[7], Richard Moxon[39], Karen Nelson[5,40], Julian Parkhill[8], Lita Proctor[26], Owen White[10], Susanna-Assunta Sansone[6], Andrew Spiers[42], Robert Stevens[3], Paul Swift[1], Chris Taylor[6], Yoshio Tateno[43], Adrian Tett[1], Sarah Turner[1], David Ussery[44], Bob Vaughan[6], Naomi Ward[45], Trish Whetzel[46], Ingio San Gil[41], Gareth Wilson[1] & Anil Wipat[35,36]

With the quantity of genomic data increasing at an exponential rate, it is imperative that these data be captured electronically, in a standard format. Standardization activities must proceed within the auspices of open-access and international working bodies. To tackle the issues surrounding the development of better descriptions of genomic investigations, we have formed the Genomic Standards Consortium (GSC). Here, we introduce the minimum information about a genome sequence (MIGS) specification with the intent of promoting participation in its development and discussing the resources that will be required to develop improved mechanisms of metadata capture and exchange. As part of its wider goals, the GSC also supports

can manipulate it to provide new solutions to critical problems. Such solutions include therapies and cures for disease, industrial products, approaches for biodegradation of xenobiotic compounds and renewable energy sources. With improvements in sequencing technologies, the growing interest in metagenomic approaches and the proven power of comparative analysis of groups of related genomes, we can envision the day when it will be commonplace to sequence tens to hundreds of genomes or more as part of a single study. At current rates of genome sequencing, it has been estimated that >4,000 bacterial genomes will be available soon after 2010 (ref. 1).

Given the importance of the growing genome collection, the capital investment in its creation and the benefits of leveraging its value

This was in 2008.

We agree, but…

12 years later "***physical archives are [still] not yet routinely created***" by groups doing whole genome sequencing.

Data provenance for genomics data and the chain of custody for the original biomaterials is poorly documented (if at all).

Field, D. *et al.* (2008) 'The minimum information about a genome sequence (MIGS) specification', *Nature Biotechnology*, 26(5), pp. 541–547. doi: 10.1038/nbt1360.

ATCC

Assume It's Dirty

Data is rarely "clean". Public data is often not "correct".

Product	NCBI existing reference genomes	NCBI assembly level (plasmids)	Sequencing technology and coverage	# of SNPs	# of indels	Average coverage (variants)
Acinetobacter baumannii (ATCC® 17978™)	GCA_001593425.2	Complete Genome	Illumina (300.0x)	14	5	210.1
	GCA_000015425.1*	Complete Genome (2)	Not available	118	656	152.7
	GCA_014672775.1	Complete Genome (1)	PacBio (399.24x)	15	87	170.4
	GCA_013372085.1	Complete Genome (2)	Illumina, Nanopore (80x)	14	2	210.2
	GCA_004797155.2	Complete Genome (2)	PacBio (247.19x)	28	62	162.1
	GCA_001077675.1	Complete Genome (1)	Illumina, PacBio (153x)	15	6	135.9
	GCA_011067065.1	Complete Genome (2)	PacBio (231.08x)	60227	2486	165.6
Candida albicans (ATCC® 10231™)	GCA_015227795.1	3, 081 Contigs	NovaSeq (16x)	10174	1573	265.6
	GCA_002276455.1	2,219 Scaffolds	HiSeq (95x)	13408	2390	274.6
Meyerozyma guilliermondii (ATCC® 6260™)	GCF_000149425.1	9 RefSeq Scaffolds	Not available	505	1973	278.2
	GCA_006942155.1	9 Contigs	ONT+MiSeq (240x)	74	386	223.3
Clavispora lusitaniae (ATCC® 42720™)	GCF_000003835.1	9 RefSeq Scaffolds	Not available	587	2336	265.6
	GCA_003675505.1	109 Scaffolds	NextSeq (182x)	102	5142	236.9

1 strain

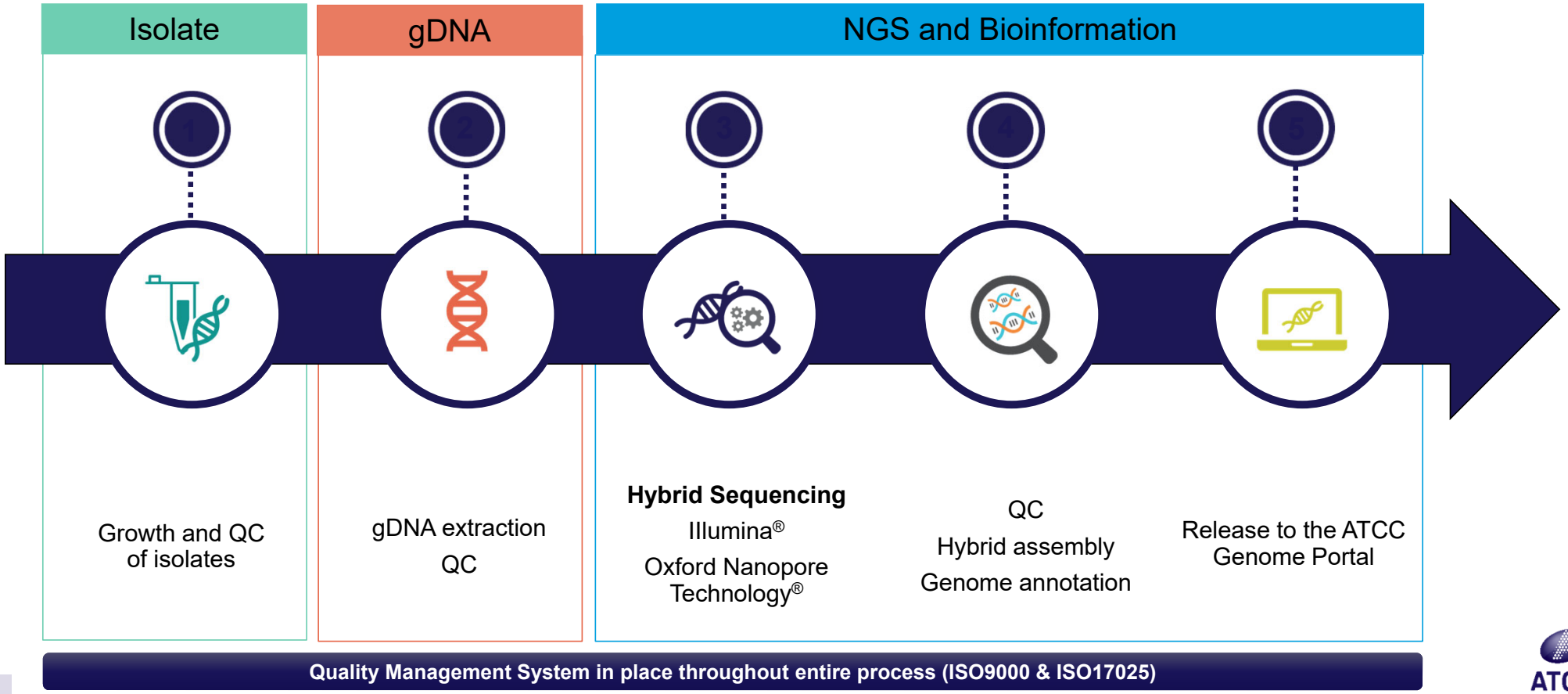
7 assemblies

Unknown origin for all source materials

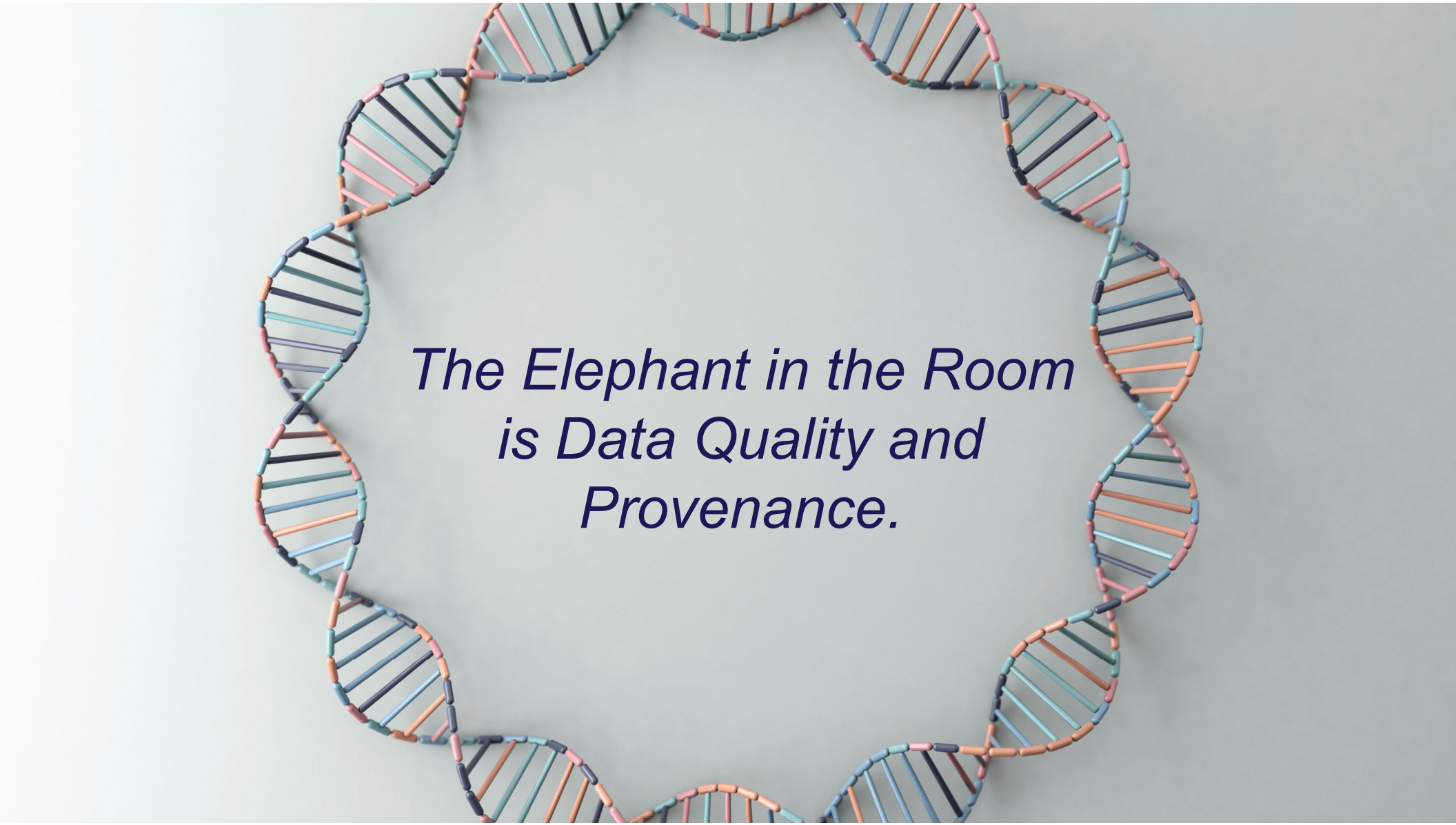
ATCC

Adopt a "Digital First" Mindset

Involve bioinformatics and data science early. Standardize your pipelines ahead of time.

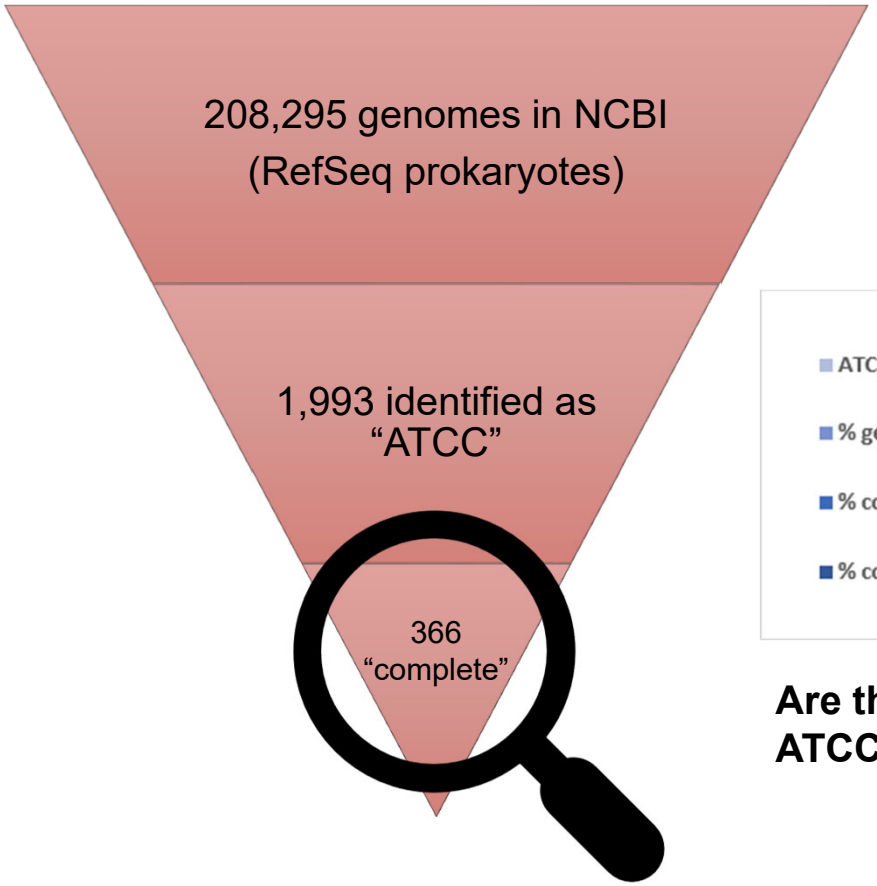


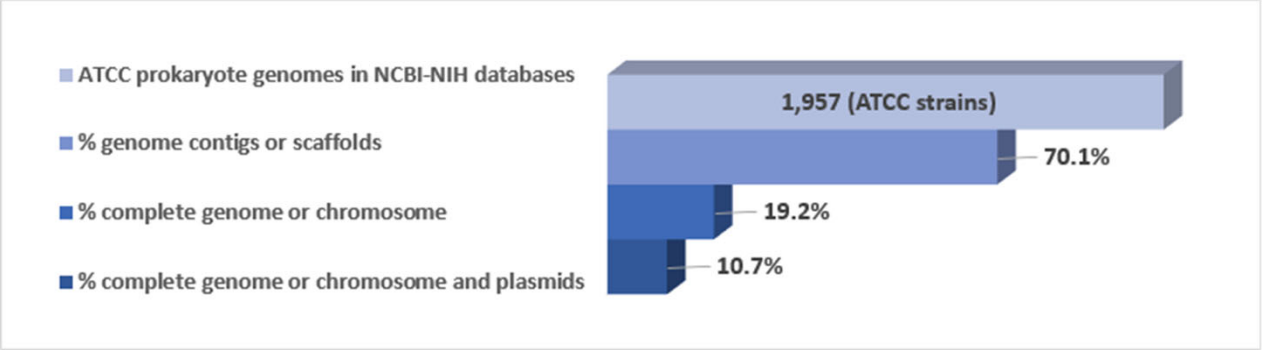
ATCC


The Elephant in the Room
is Data Quality and
Provenance.

The ATCC Genome Portal

How many "ATCC" strains are in RefSeq?



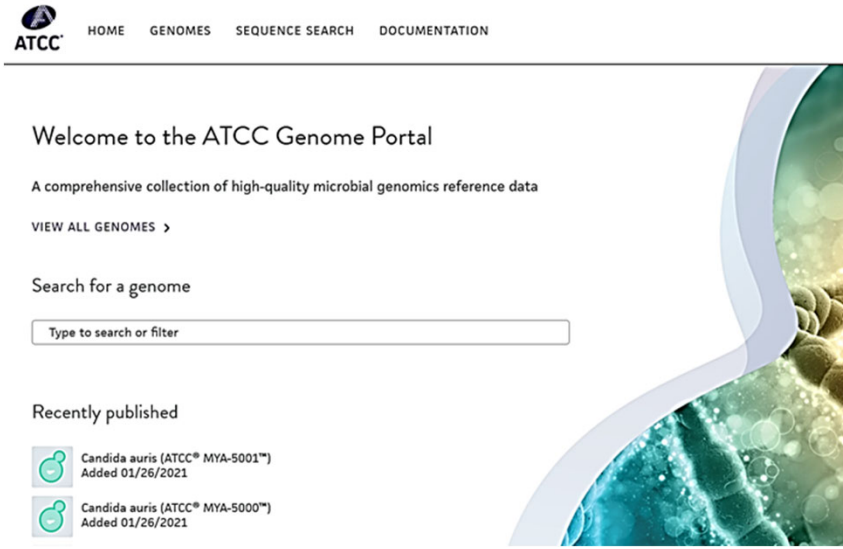


Are these 585 RefSeq genomes traceable back to authenticated ATCC materials?

ATCC's Enhanced Authentication Initiative

Our approach to producing authenticated reference genomes

- **2017-2018** – Planning and proof-of-concept experiments
- **2018** – Commitment
  - Laboratory and staffing resources
  - Instrumentation
  - Bioinformatics pipelines
- **2019** – Launch of the Enhanced Authentication Initiative
  - June 2019 – *beta* launch at ASM Microbe
  - Sept 2019 – formal launch of the ATCC Genome Portal
    - Provide our customers with the whole-genome sequences of the specific, authenticated materials researchers need to generate credible data
    - genomes.atcc.org



ATCC Genome Portal

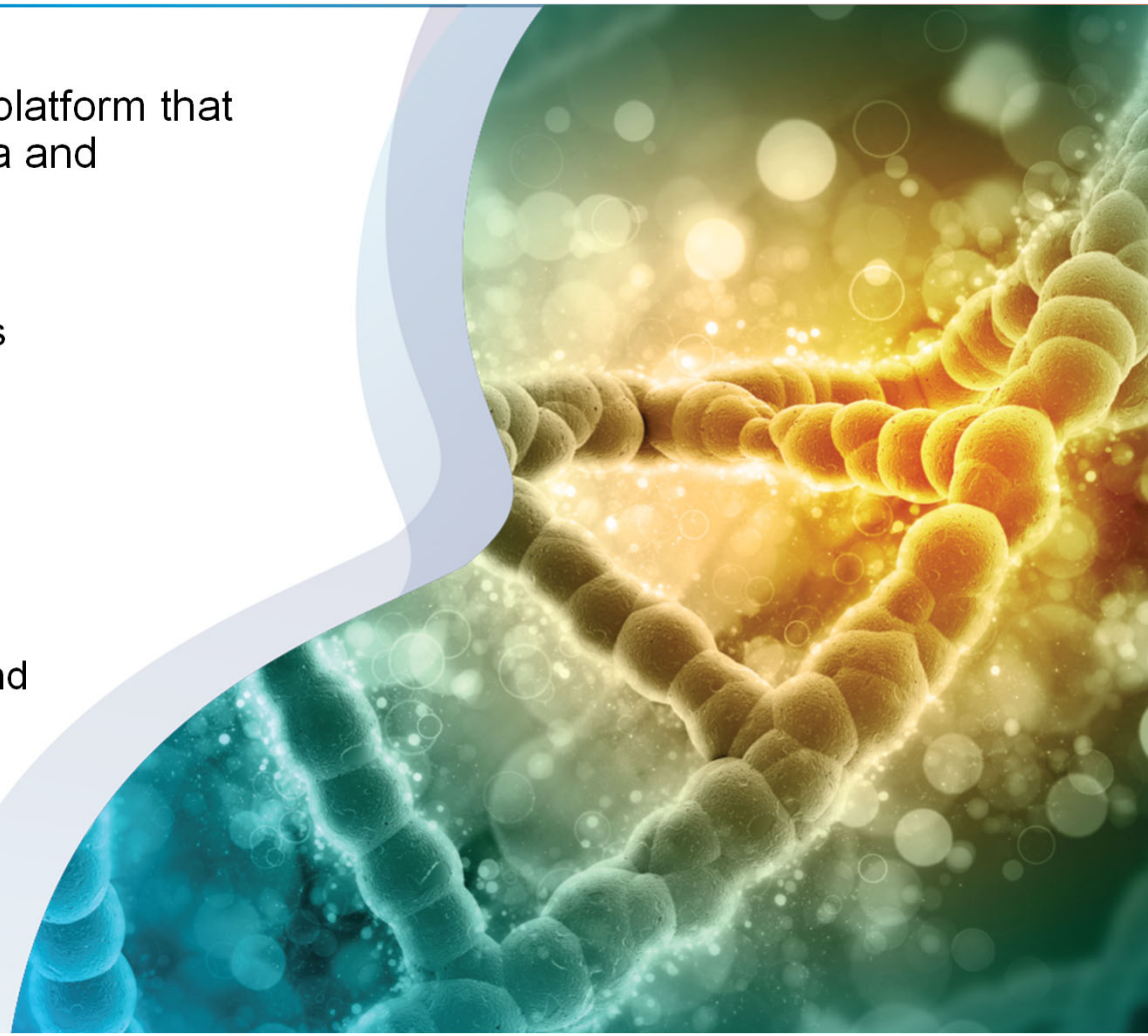
The ATCC Genome Portal is a cloud-based platform that enables users to easily browse genomic data and metadata by simply logging into the portal

Download whole-genome sequences and annotations of ATCC materials

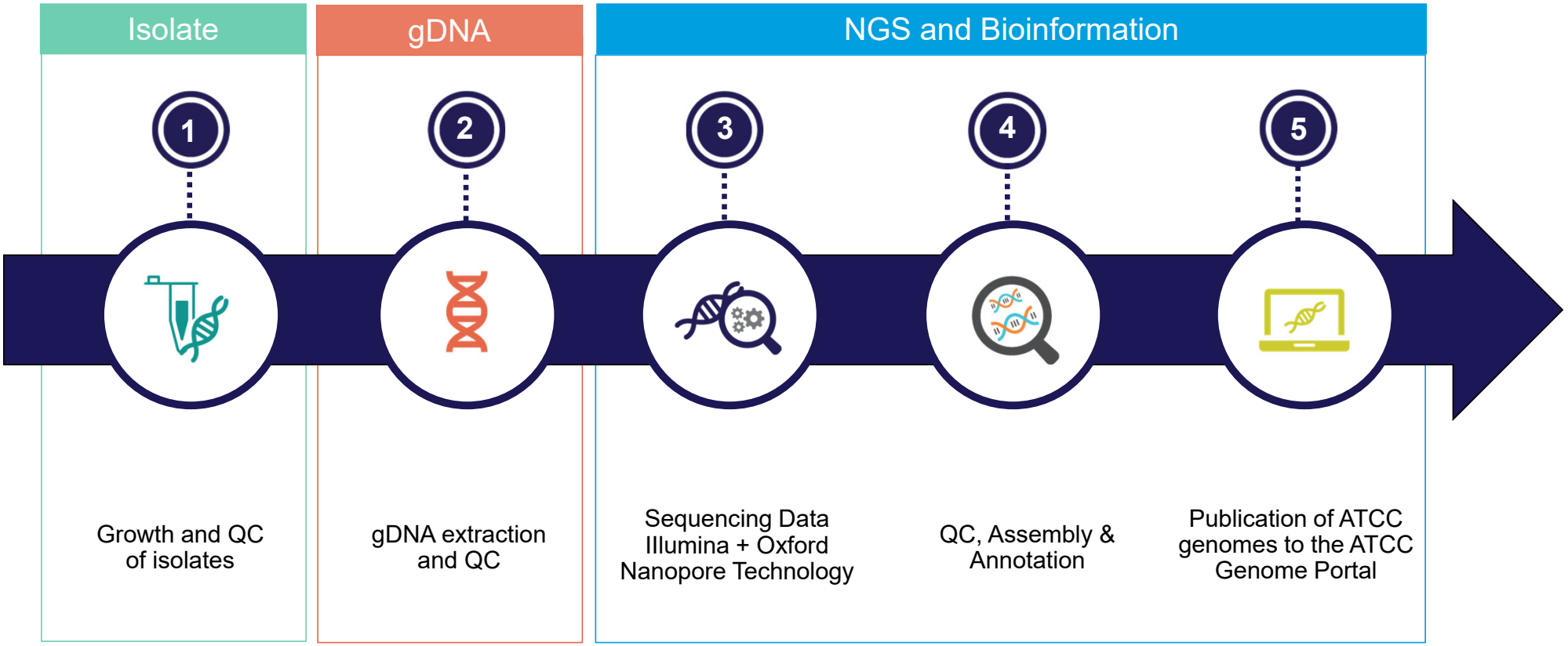
Search for nucleotide sequences or genes within genomes

View genome assembly metadata and quality metrics

genomes.atcc.org



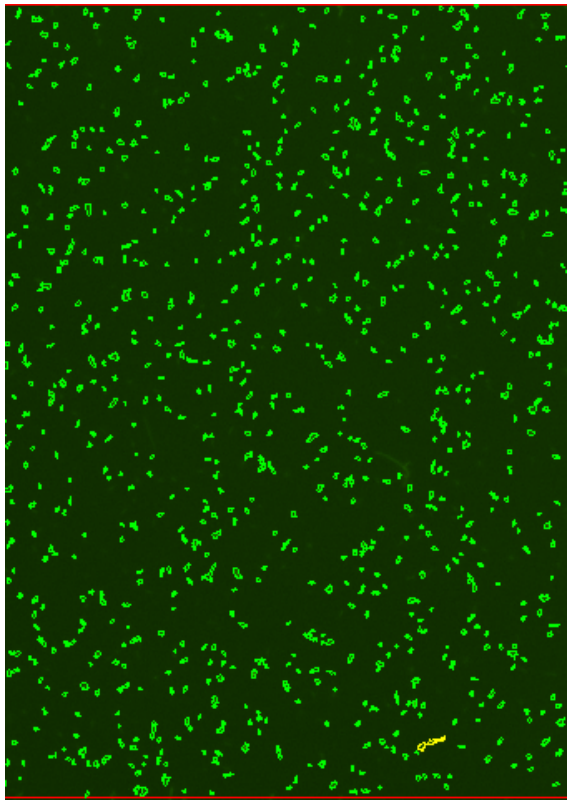
Authenticated physical material coupled with reference-quality genome sequences



Processes for producing reference-quality genomes

Extraction of gDNA

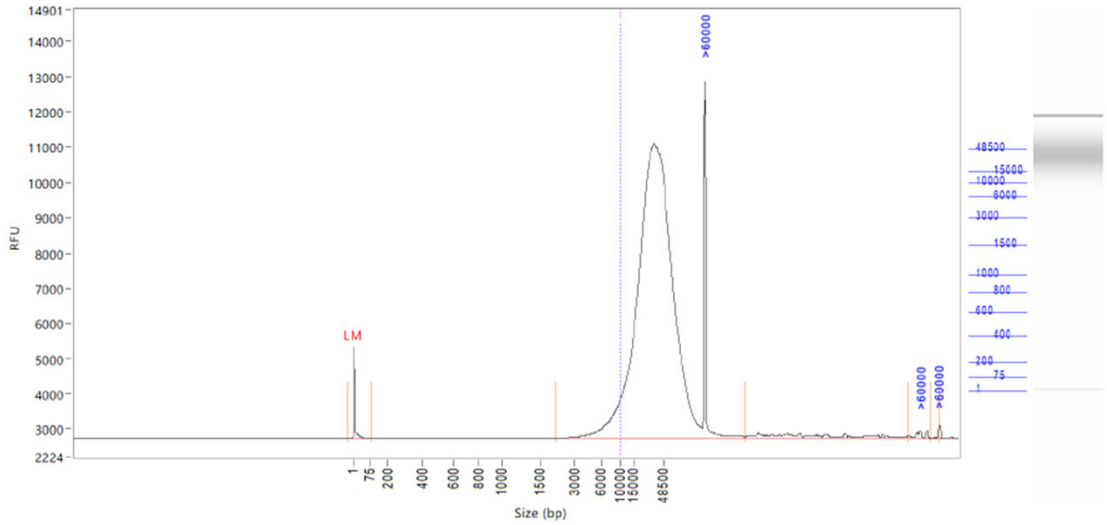
- Start with a fresh culture grown according to ATCC's item-specific manufacturing process
- Determine the cell count
  - Typically start with ≥10^9 cells/mL
- The "best" extraction method depends on the organism
- Simply recovering DNA is not good enough
  - Concentration
    - Measured by Qubit™ or Picogreen®
  - Purity
    - Measured with NanoDrop™
    - $A_{260/280}$ ≥1.7 to ≤ 2.1
  - Quality and Integrity
    - Fragment size is measured by Fragment Analyzer™



Fusobacterium nucleatum ATCC® 25586™
6.58 x 10^8 cells /mL

Processes for producing reference-quality genomes

Fragment analysis of gDNA



Peak	Size (bp)	Conc. (ng/uL)	From (bp)	To (bp)	Avg. Size (bp)	CV%	RFU	Corr. Peak Area
1	1 (LM)	0.0154	0	79	3	271.10	2572	15.706
2	>60000	231.7799	2177	>60000	42443	53.22	10128	1178.612
3	>60000	0.7996	>60000	>60000	>60000	1.85	206	4.066
4	>60000	0.1158	>60000	>60000	>60000	0.30	253	0.589

TIC:	232.6953	ng/uL
TIM:	8.9951	nmole/L
Total Conc.:	237.9093	ng/uL

GQN:	9.6
Threshold:	10000

- *Corynebacterium tuberculostearicum* (ATCC® 35692™)
- Total concentration: 234 ng/µL
- Average fragment size: ≥42,000bp
- GQN: 9.6 with a threshold of 10,000bp
  - "Genomic Quality Number"
  - 96% of the sample contains fragments larger than 10,000 bp

Processes for producing reference-quality genomes

ATCC extraction quality control

ATCC® no.	Species	Qubit (ng/µL)	A_{260}/A_{280}	DNA fragment size (range)**
8739™	*Escherichia coli*	101.9	1.92	49.5 kb (1.5 – >60 kb)
13048™	*Klebsiella aerogenes*	98.1	1.86	49.5 kb (1.6 – >60 kb)
11828™	*Cutibacterium acnes*	197.7	1.84	29.8 kb (0.8 – >60 kb)
6538™	*Staphylococcus aureus*	97.8	1.85	32.9 kb (2.7 – >60 kb)
BAA-2797™	*Pseudomonas aeruginosa*	153.3	1.99	44.1 kb (1.1 – >60 kb)
824™	*Clostridium acetobutylicum*	73.8	2.05	12.5 kb (4.6 – 57.8 kb)
6538™	*Staphylococcus aureus*	37.1	2.00	26.2 kb (6.9 – >60 kb)
27774™	*Desulfovibrio desulfuricans*	69.2	1.99	58.5 kb (13.3 – >60 kb)
11842™	*Lactobacillus delbrueckii*	64.8	2.02	41.9 kb (6.1 – >60 kb)
15697™	*Bifidobacterium longum*	76.2	1.95	51.3 kb (10.5 – >60 kb)

** Main peak reported

ATCC

Processes for producing reference-quality genomes

Library preps for both Illumina® and Oxford Nanopore Technologies®

Illumina

- DNA libraries are prepared using Illumina's DNA Prep kit and unique dual indexes (Cat. # 20018705)
- RNA libraries are prepared using NEBNext Ultra II RNA Library Prep Kit (Cat # E7770S)
- Sequenced on the MiSeq® or NextSeq® instrument
  - Paired-end read set per sample
  - Multiplexing is based on the estimated genome size
  - Data necessary to generate at least 100X coverage of the genome
- Reads are adapter trimmed using the adapter trimming option on the Illumina instrument

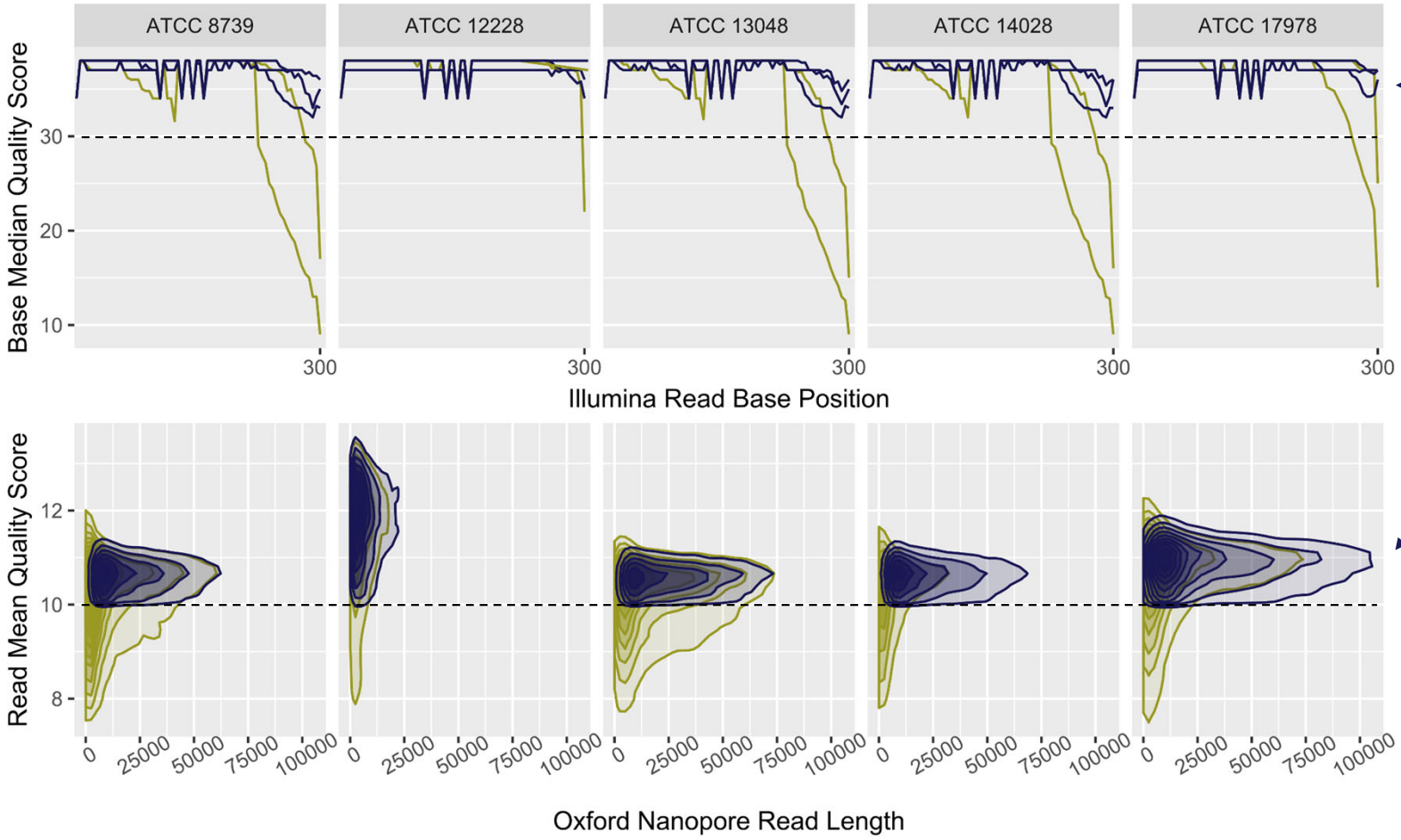
+

Oxford Nanopore Technologies

- Libraries are prepared using ONT's Ligation Sequencing Kit (SQK-LSK109) with the Native Barcoding Expansion kit (EXP-NBD104 or EXP-NBD114)
- Sequenced on the GridION using the version 9.4.1 flow cell
- The quantity of samples typically multiplexed is based on the estimated genome size of the given organism.
- Flow cells are run for 48-72 hours
- Barcode detection, demultiplexing, and barcode trimming are completed on the instrument, parallel to the run

ATCC

Sequencing QC – Read trimming/filtering



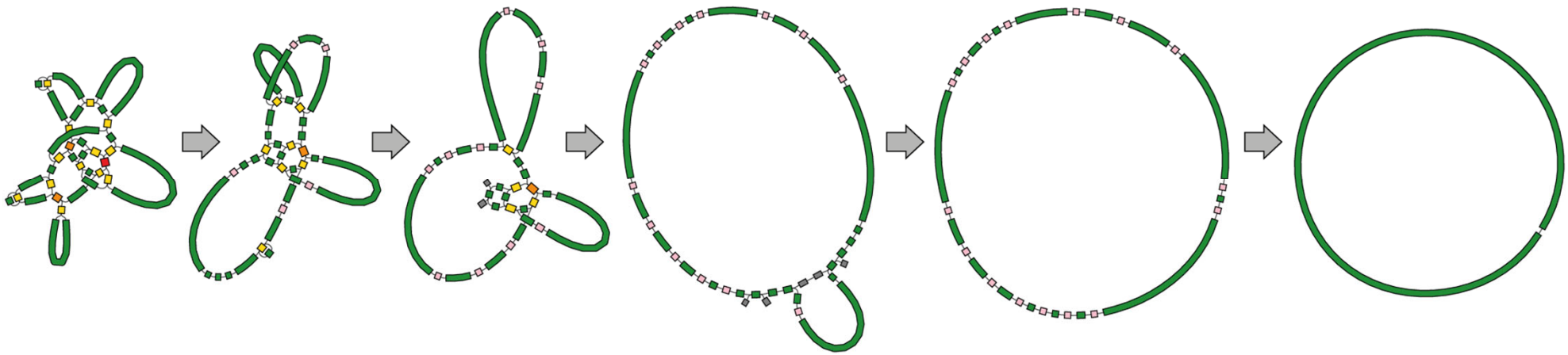
Only keep high-quality base calls

Only keep long, good quality reads

Quality controlled data

Raw data

Hybrid genome assembly approach



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

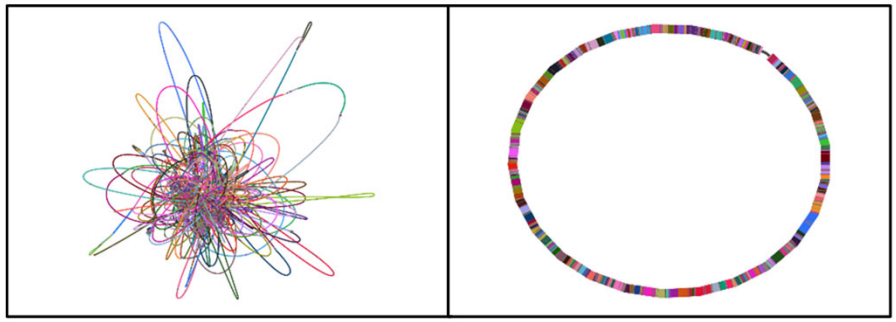
Completed hybrid assembly

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

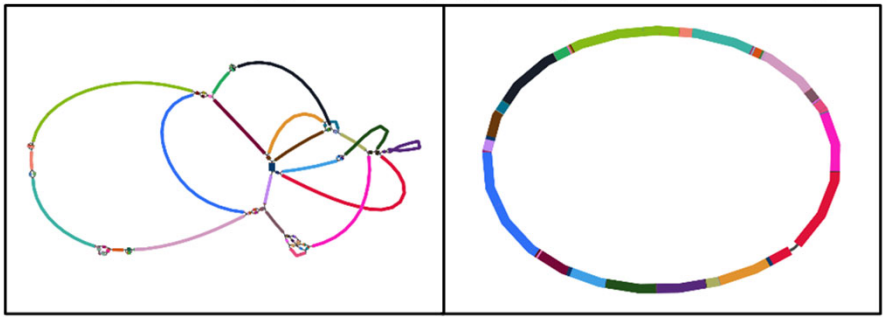
Improved assemblies via hybrid sequencing

Illumina-only assembly	Hybrid assembly

Neisseria meningitidis (ATCC® 53417™)

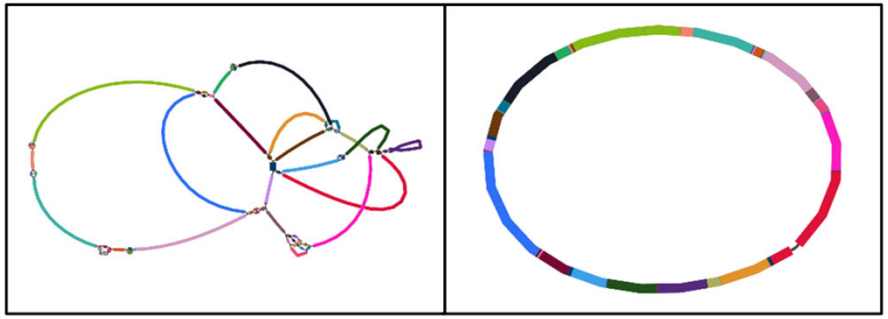


Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)

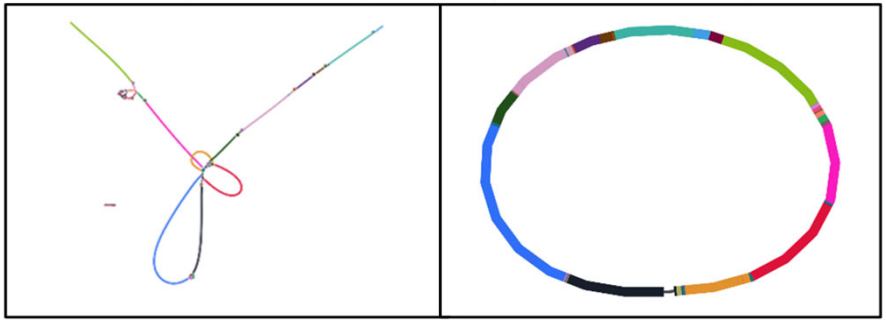


Illumina-only assembly	Hybrid assembly

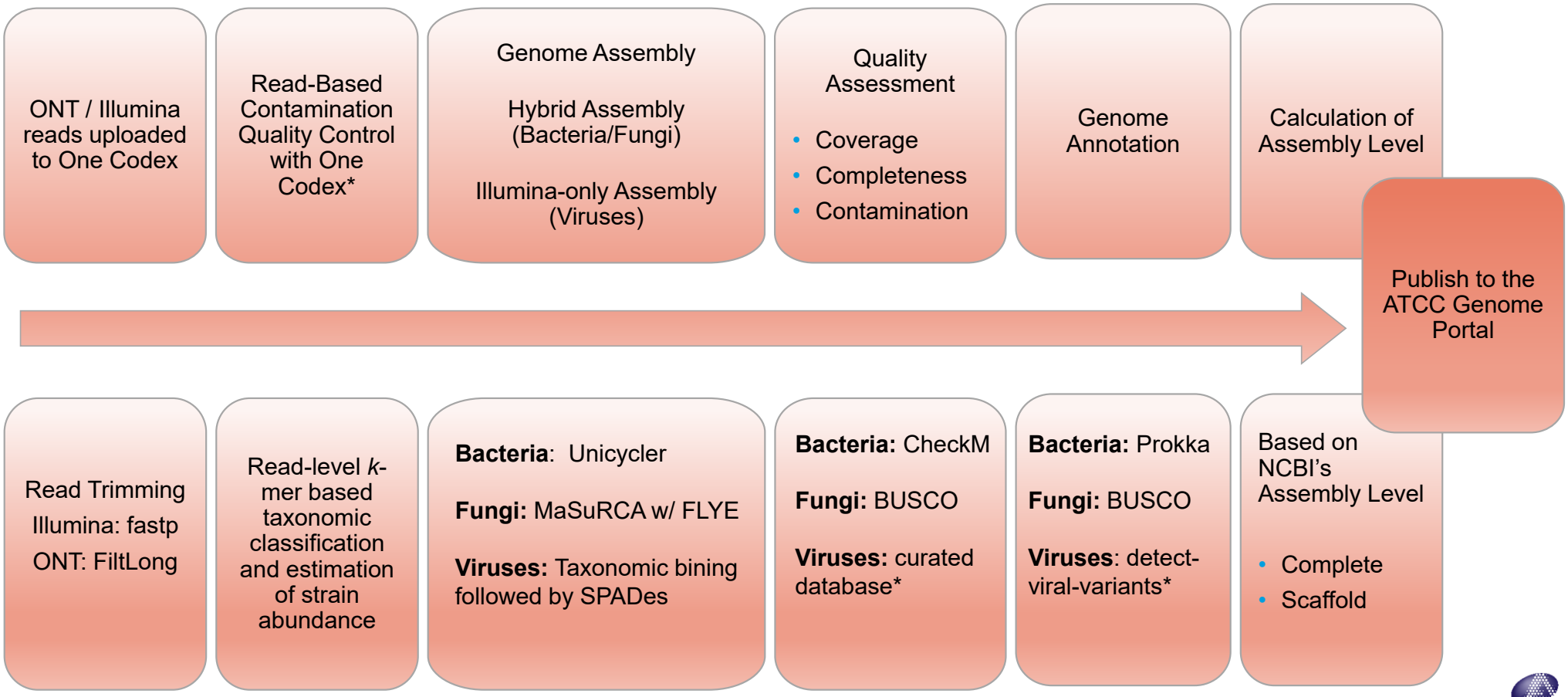
Campylobacter jejuni subsp. *jejuni* (ATCC® 43446™)



Streptococcus gordonii (ATCC® 35105™)



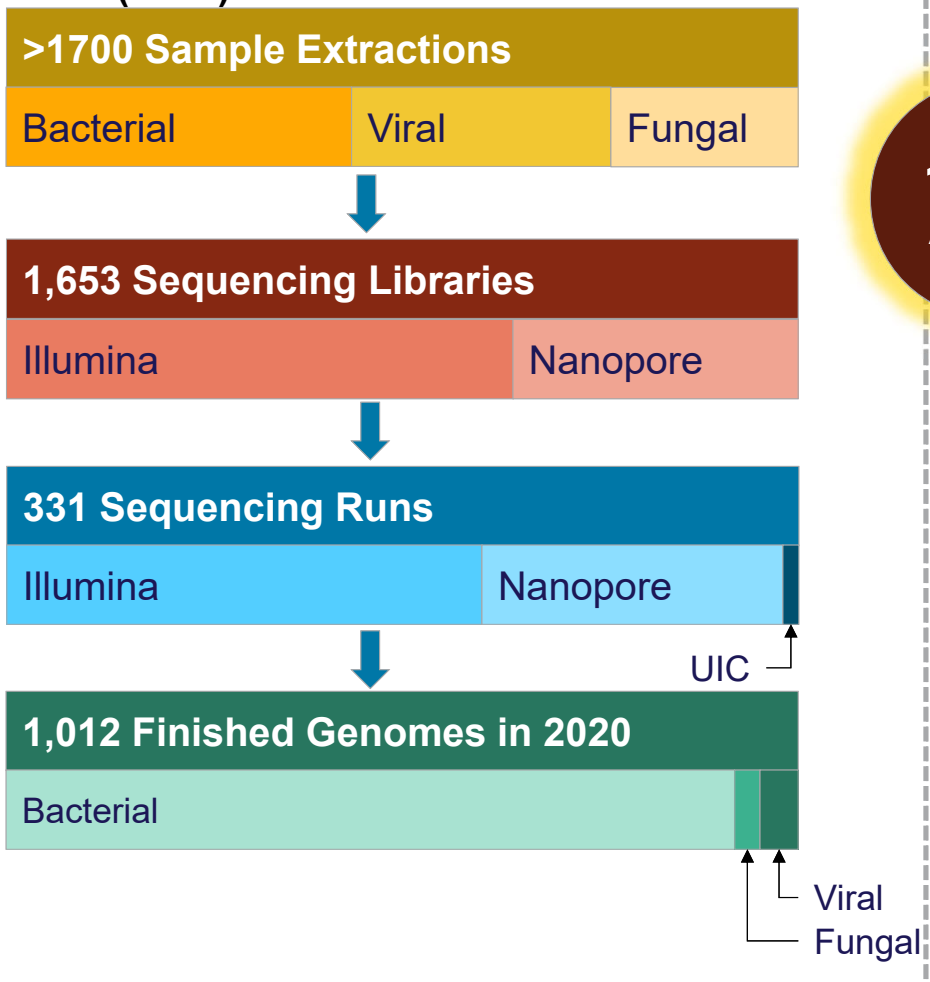
ATCC genome assembly process



* One Codex proprietary software

ATCC Standard Reference Genomes (ASRGs)

Last Year (2020)



1,579 ASRGs

← As of today

- 1,396 bacterial references
  - 900+ complete circularized genomes
  - 436 Type Strains
  - 147 clinical MDR (GPS) strains
  - 56 Microbiome Standards references
- 74 mycology references
- 128 viral reference genomes

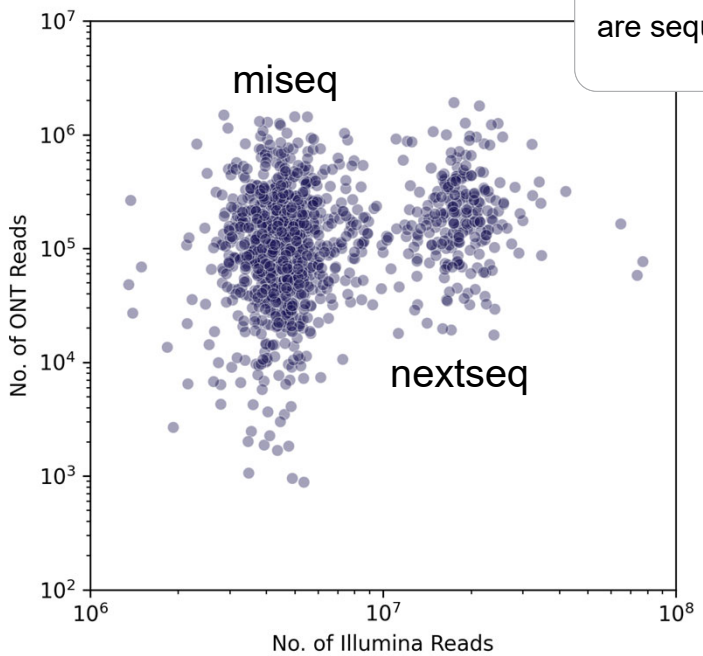
- References can be downloaded for free for Research Use Only purposes
- Commercial licenses available (inquire)

- Monthly updates
- All genomes are 100% traceable to ATCC's biomaterials
- Hybrid assembly for all bacterial & fungal genomes
- Genomes annotations included

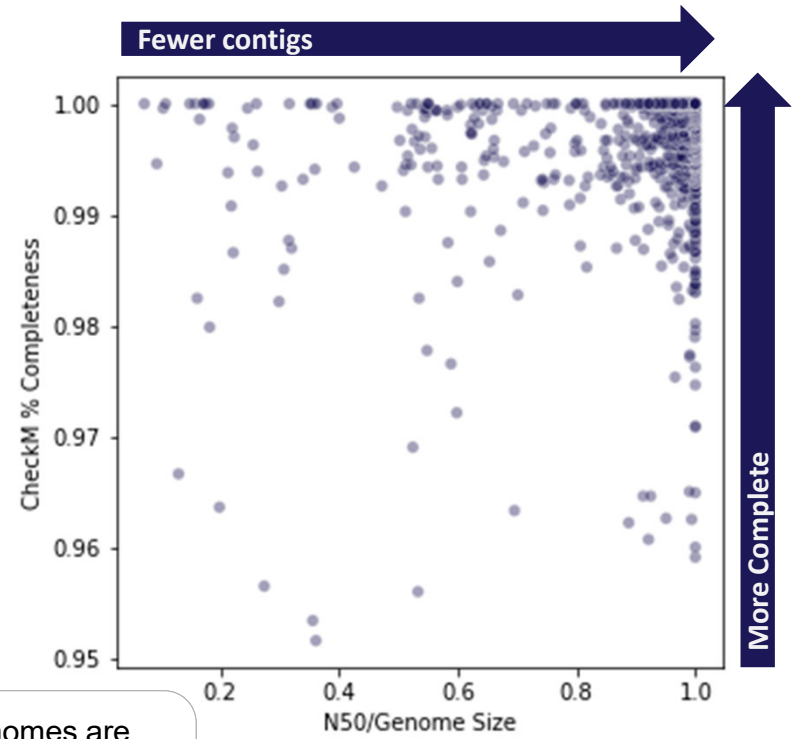
- Additional content and site improvements coming …

ATCC

Quality of ATCC Genome Portal Assemblies

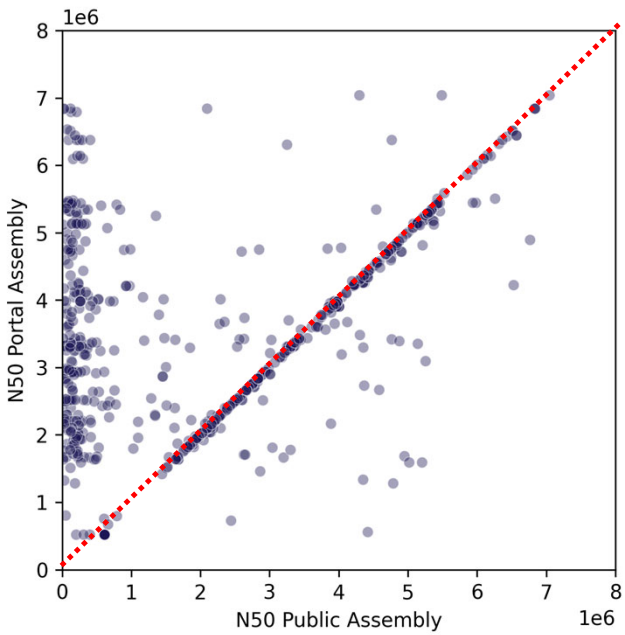


All bacterial and fungal genomes are sequenced on both Illumina and Oxford Nanopore

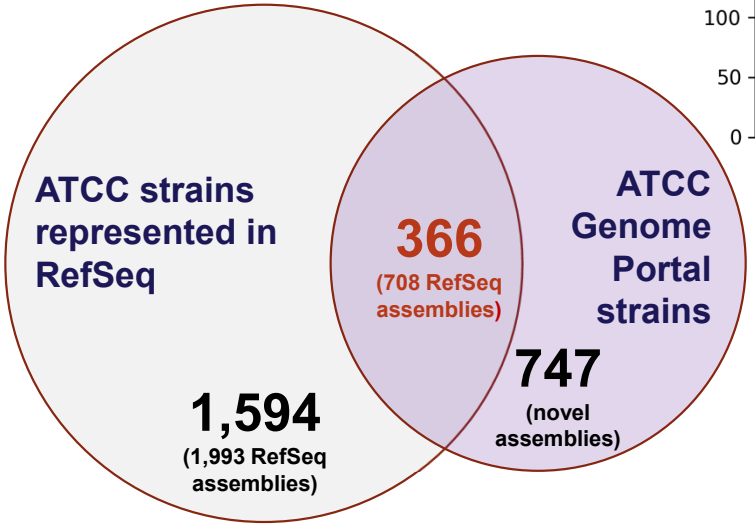


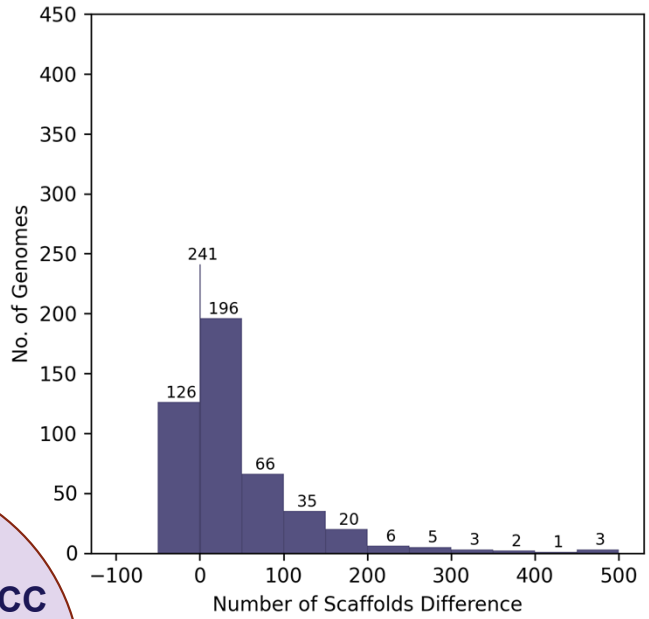
All genomes are required to be at least 95% complete (CheckM/Busco)

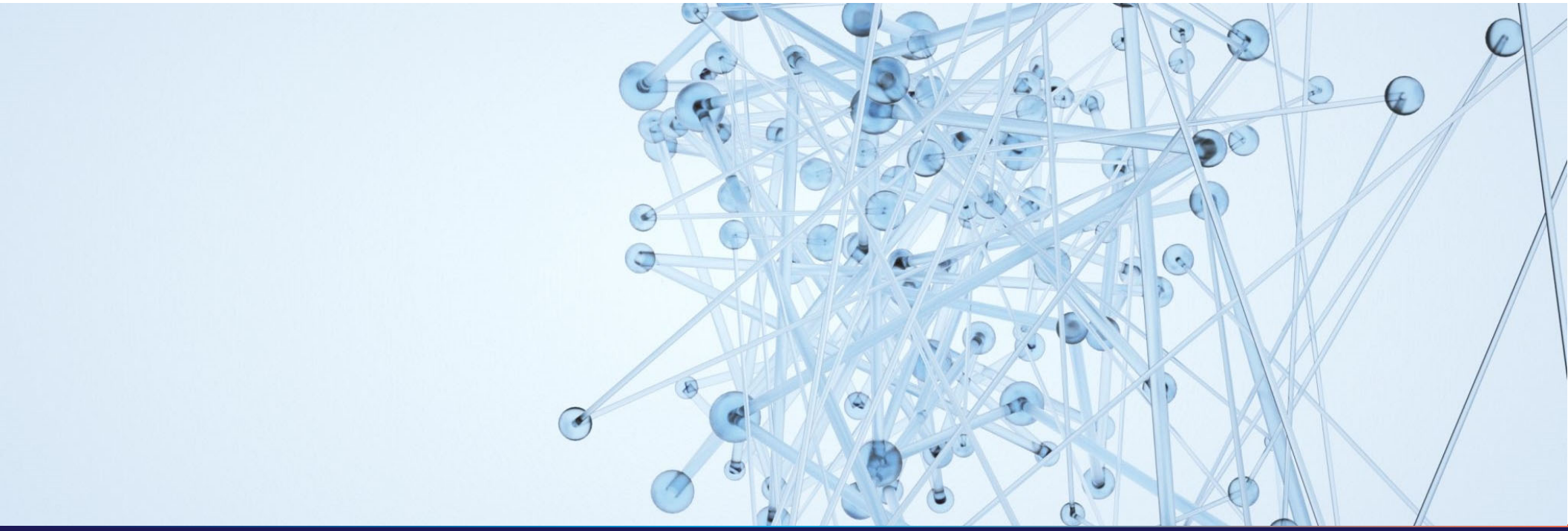
Comparison of ATCC vs. RefSeq bacterial assemblies



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







In Summary…

Summary & Questions?

- **Public genomics data should be handled with care**
  - Usually “OK”, but often has errors or omissions
  - Examples of data falsification persist in public databases
  - Ineffective data curation and control
  - ~50% of RefSeq does not have clear provenance to source materials

- **ATCC Genome Portal**
  - The only genomic database with 100% data provenance
  - Over 98% of our assemblies are superior to RefSeq
  - Adding 100+ new genomes per month
  - All source materials are available from ATCC
  - All methods and procotols are traceable and controlled.

Visit us at https://genomes.atcc.org

ATCC

The ATCC Genome Portal Team @ATCCgenomics

Jonathan Jacobs, PhD
Senior Director, Bioinformatics
BioNexus Principal Investigator
jjacobs@atcc.org @bioinformer

Genomics Lab	Bioinformatics Lab	Partners
Briana Benton, PMP	**John Bagnoli**	Marco Riojas, PhD
Stephen King, MS	David Yarmosh, MS	
James Duncan, MS	Nikhita Putheveetil, MS	
Robert Marlow	P. Ford Combs, MS	
Samuel Greenfield	Amy Reese, MS	
Corina Tabron		
Amanda Pierola		*… and OneCodex*
Shanice Corlette		

***JOIN OUR TEAM!** We're hiring!*