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]

---


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and fabricating research data in five published research papers, two published review articles, one submitted but unpublished paper, in his doctoral dissertation, and in a submission to the GenBank computer database.” – The Federal Register, v62, n135 (1997)
24 years later, this falsified data still being cited...

Pathogenic noncoding variants in the neurofibromatosis schwannomatosis predisposition genes

Cristina Perez-Becerril

Division of Evolution and Genomic Science, Manchester Centre for Genomic Medicine, Manchester, UK

Comparison of the full human and murine neurofibromatosis genes revealed a high degree of similarity (>75%) and high conservation levels across 5'- and 3'-UTRs (Bernards et al., 1994a; Hata et al., 1994b). A subsequent in situ hybridization study compared the S- and L-chain introns 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

NF1 and NF2 loci, respectively. To date, most variants associated with NF1 and NF2 have been identified in the SMARC1 and LTR1 genes, and the D6G RNase recently reported to predispose to schizophrenia for high enzyme activity (Perez-Becerril et al., 2017). 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

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

Dated: July 9, 1997.
David F. Garrison,
Principal Deputy Assistant Secretary for Planning and Evaluation.

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:

Amitava 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, a former graduate student and University of Michigan, engaged in scientific misconduct by falsifying and fabricating research data in five published research papers, two published review articles, and one submitted but unaudited paper, in his doctoral dissertation, and in his submission to the GenBank computer data base.

Mr. Hajra submitted a falsified nucleotide sequence in computer data base.


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


Mr. Hajra submitted a falsified nucleotide sequence in computer data base.
After 42 citations... the data is still in GenBank...
Falsified sequencing data to support a false phylogeny

“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)
Unfortunately, the falsified mitogenome is still in GenBank...

Labeled as “Unverified”, but the sequence still remains in GenBank and, for example, will come up in a BLAST search...
Large-scale contamination of microbial isolate genomes by Illumina PhiX control

Supratik Mukherjee,1,2,3,4,5 Matthew Huntteman,1 Natalia Varennikova1, Nikos C. Vymades1,2,5 and Amita Pan1

Abstract

While the rapid growth and development of sequencing technologies are exploiting solutions to some of the world’s biggest challenges such as the exploration of genomic dark matter. However, progress in sequencing that can occur during sample or library preparation, sequencing, many scenarios remain that make publicly available microbial isolate genome sequence databases and identified more than 1000 genomes that are contaminated during Illumina sequencing runs. Approximately 10% of these genomes showed sequence similarity with PhiX contaminated genomes were sequenced under the Human Microbiome Project (HMP) database and identified more than 1000 genomes that are contaminated during Illumina sequencing runs. Approximately 10% of these genomes showed sequence similarity with PhiX.

METHOD

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

Martin Steinegger1,2,3 and Steven L. Salzberg1,2,3

Abstract

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Contaminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2164746, 1140036, and 14146 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 1.3 TB in 12 days on a 32-core computer. Contaminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/marxin/contaminator

Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser,1 Mihaela Pertea,1,4,5 Aleksey V. Zimin,1,5 and Steven L. Salzberg1,2,3,4

1 Center for Computational Biology, Missouri–Kansas City Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA; 2Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; 3Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA; 4Department of Biomedical Informatics, School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA

Human contamination in bacterial genomes can create thousands of spurious proteins.

Research
Poor quality genomes result in taxonomic misclassification

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

- Approximately 7.8% of genomes misclassified at the species level.
- Approximately 4% at the genus level.
- Approximately 7% of genomes misclassified at genus or higher.
Challenging traceability of most public genomics data

Potential issues with genomic source material
- “Lab adaptation”
- Loss of plasmids
- Sample mix ups
- Unknown chain of custody
- Differences in sequencing technology and bioinformatics

<table>
<thead>
<tr>
<th>SNPs differences</th>
<th>ATCC Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draft genome</td>
<td>&gt;13,000</td>
</tr>
<tr>
<td>“High-quality” reference genome</td>
<td>&gt;10,000</td>
</tr>
</tbody>
</table>
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

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

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

**Uncontrolled Public Data**
- Unknown quality
- Missing or non-standard metadata
- Risky to use

**Authenticated Genomics Data**

GenBank
ENA
SRA
EGA
GEO
Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap

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

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

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

RefSeq
proGenomes
dbGAP
BluePrint
ICGC
TCGA
TARGET
CCE
GTEx
Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap

**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

OmicSoft
Ingenuity (IPA)
HGMD
Authenticated genomics data at ATCC

ATCC is focused on data provenance and closing the reproducibility gap

**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*
The ATCC Genome Portal

Tackling the reproducibility gap in microbial genomics
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

2,522 Authenticated Reference Genomes

- 2145 bacteria
- 221 viruses
- 155 fungi
- 1 protist

New genomes released every month!

[genomes.atcc.org]
**Authenticated physical material coupled with reference-quality genome sequences**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>gDNA</th>
<th>NGS and Bioinformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth and QC of isolates</td>
<td>gDNA extraction and QC</td>
<td>Hybrid Sequencing&lt;br&gt;Illumina + Oxford Nanopore&lt;br&gt;QC, Assembly &amp; Annotation</td>
</tr>
</tbody>
</table>

**Hybrid Sequencing**

Illumina + Oxford Nanopore
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

<table>
<thead>
<tr>
<th></th>
<th>Illumina-only assembly</th>
<th>Hybrid assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Neisseria meningitidis</em> (ATCC® 53417™)</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td><em>Campylobacter jejuni subsp. jejuni</em> (ATCC® 43446™)</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td><em>Campylobacter jejuni subsp. jejuni</em> (ATCC® 43446™)</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td><em>Streptococcus gordonii</em> (ATCC® 35105™)</td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
</tr>
</tbody>
</table>
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
1,594 (1,993 RefSeq assemblies)

ATCC Genome Portal strains
366 (708 RefSeq assemblies)

747 (novel assemblies)

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.

<table>
<thead>
<tr>
<th>Species</th>
<th>Cell types</th>
<th>Tissue types</th>
<th>Diseases types</th>
</tr>
</thead>
<tbody>
<tr>
<td>70+</td>
<td>100+</td>
<td>100+</td>
<td>400+</td>
</tr>
</tbody>
</table>
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 Bioinformatics & Curation

Strict quality acceptance criteria at each step
ATCC Cell Line Land
A partnership with QIAGEN Digital Insights

- 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

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

Manually curated cell line ‘omics data from the most popular cell lines in ATCC’s collection

ATCC Cell Line Land is a continually growing database of cell line ‘omics 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 ‘omics 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

Jonathan Jacobs, PhD
Senior Director, Bioinformatics
BioNexus Principal Investigator

Genomics Team
Briana Benton, PMP
Ana Fernandes
Ajeey Singh, PhD
Stephen King, MSc
James Duncan, MSc
Samuel Greenfield, MSc
Corina Tabron, MSc
Noah Wax, MSc
Rula Khairi, MSc
Robert Marlow
Jade Kirkland

Bioinformatics Team
John Bagnoli
Scott Nguyen, PhD
David Yarmosh, MSc
Nikhita Putheveetil, MSc
P. Ford Combs, PhD
Amy Reese, MSc

PARTNERS
Marco Riojas (ATCC / BEI)
James Crill (Syracuse University)
Kylene Kehn-Hall (Virginia Tech)
Vishnu Chaturvedi (Wadsworth)
QIAGEN Digital Insights
One Codex

@ATCCgenomics

© 2022 American Type Culture Collection. The ATCC trademark and trade name, and any other trademarks listed in this publication are trademarks owned by the American Type Culture Collection unless indicated otherwise. These products are for laboratory use only. Not for human or diagnostic use. ATCC products may not be resold, modified for resale, used to provide commercial services, or to manufacture commercial products without prior ACC written approval. Illumina is a registered trademark of Illumina, Inc. Oxford Nanopore is a registered trademark of Oxford Nanopore Technologies Limited.
Thank you!

<table>
<thead>
<tr>
<th>ATCC Genome Portal</th>
<th><a href="https://genomes.atcc.org">https://genomes.atcc.org</a></th>
</tr>
</thead>
</table>