TWIST BIOSCIENCE

# DNA synthesis screening

JAMES DIGGANS

# We Are Writing the Future

558k genes in FY22
634k genes in FY23

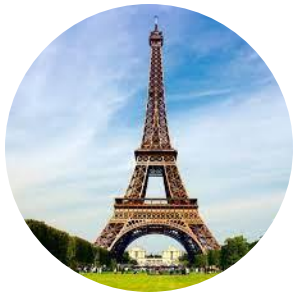
TWIST BIOSCIENCE

DNA ON SILICON
Nasdaq: TWST
~1000 Employees

**South San Francisco* Headquarters**

**Portland, OR***

**San Diego, CA**



**Europe**

**Singapore**

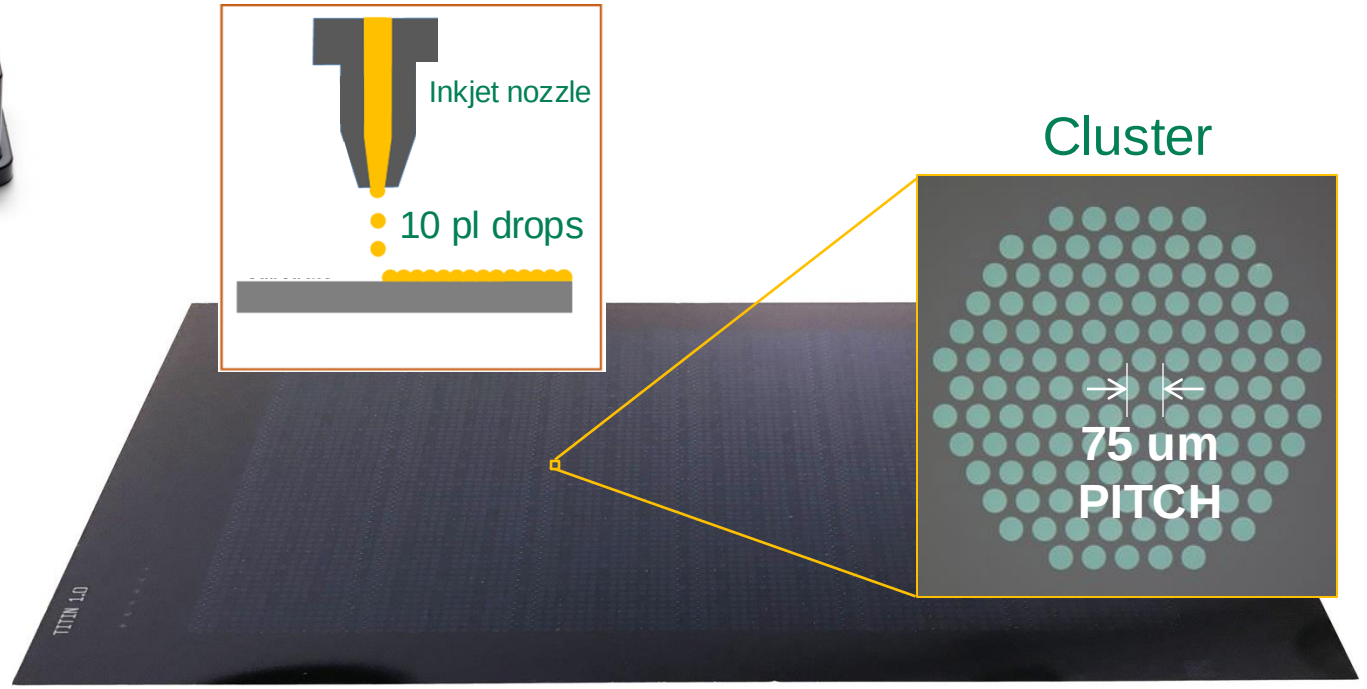
**Guangzhou China**

*Manufacturing Facilities

# Massively Parallel Synthetic DNA Synthesis

96-well plastic plate

- **700k+ unique oligo sequences** in 6,144 clusters per array
- **600 million DNA strands** synthesized per sequence on each 50-um spot




Passive silicon substrate with 6,144 clusters

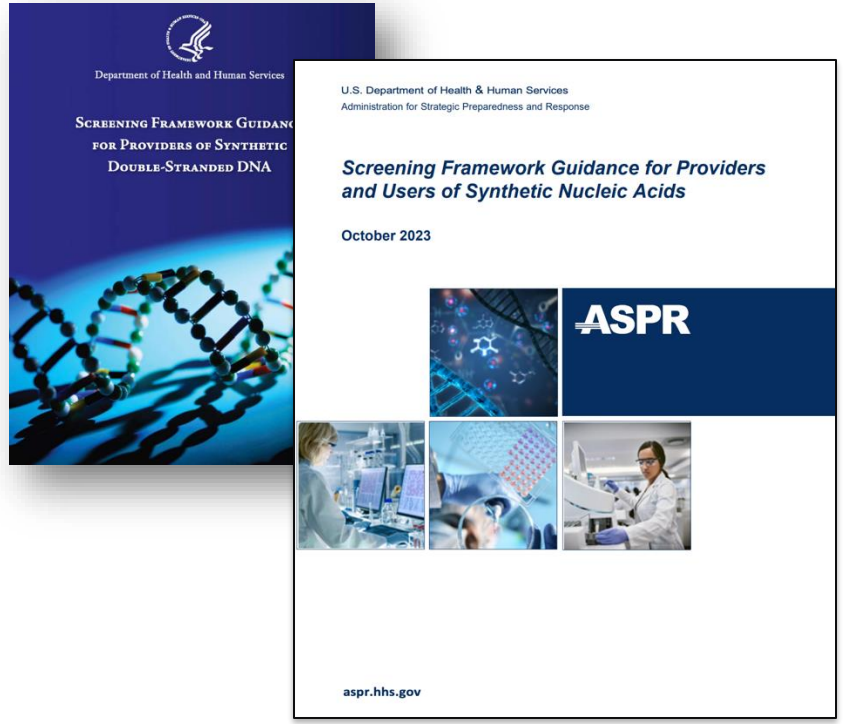
# Nucleic Acid Synthesis & Biosecurity

- **<u>DNA is a dual-use technology</u>**

- Orders are screened to ensure <u>responsible use</u>
  - Customer screening
  - Sequence screening – homology-based

- IGSC: 40+ companies working together to build best practices to reduce the risk of misuse

- But technology moves quickly:
  - Scale of synthesis and turn around time
  - Synthetic biology blurs species lines
  - **AI biodesign tools <u>reduce</u> homology to naturally occurring proteins**




International Gene Synthesis Consortium

# Microsoft & the IGSC: Building AI resistant screening

- Microsoft approached the IGSC:
  - Identified a potential AI bio tool-related screening vulnerability
- MSFT and IGSC companies including 4 screening tool makers:
  - Followed CERT vuln. disclosure
  - MSFT protein reformulation pipeline
  - Assessed the vulnerability
  - Developed mitigations
- Wrote up the results for publication
  - *Extensive* pre-publication review

## Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations

Bruce J. Wittmann[1,*], Tessa Alexanian[2,3,†], Craig Bartling[2,4,†], Jacob Beal[2,5,†], Adam Clore[2,6,†], James Diggans[2,7,†], Kevin Flyangolts[2,8,†], Bryan T. Gemler[2,4,†], Tom Mitchell[5,†], Steven T. Murphy[5,†], Nicole E. Wheeler[2,9,†], Eric Horvitz[1,*]

[1]Microsoft, Office of the Chief Scientific Officer; Redmond, WA 98052, USA.

[2]International Gene Synthesis Consortium.

[3]The International Biosecurity and Biosafety Initiative for Science (IBBIS); Geneva, 1207, Switzerland.

[4]Battelle; Columbus, OH 43201, USA.

[5]RTX BBN Technologies; Cambridge, MA 02138, USA.

[6]Integrated DNA Technologies Inc.; Coralville, IA 52241, USA.

[7]Twist Bioscience; South San Francisco, CA 94080, USA.

[8]Aclid; New York, NY 10018, USA.

[9]Institute of Microbiology and Infection, Department of Microbes, Infections and Microbiomes, School of Infection, Inflammation and Immunology, College of Medicine and Health, University of Birmingham; Birmingham, B15 2TT, United Kingdom.

*Corresponding Authors: bwittmann@microsoft.com and horvitz@microsoft.com

[†]These authors contributed equally

Fast-moving advances in AI-assisted protein engineering are enabling breakthroughs in the life sciences that promise numerous beneficial applications. At the same time, these new capabilities are creating potential biosecurity challenges by providing new pathways to intentional or accidental synthesis of genes that encode hazardous proteins. The synthesis of nucleic acids is a key choke point in the AI-assisted protein engineering pipeline as it is where digital designs are transformed into physical instructions that can produce potentially harmful proteins. Thus, one focus for efforts to enhance biosecurity in the face of new AI-enabled capabilities is on bolstering the screening of orders by nucleic acid synthesis providers. We describe a multistakeholder, cross-sector effort to address biosecurity challenges with uses of AI-powered biological design tools to reformulate naturally occurring proteins of concern to create *synthetic homologs* that have low sequence identity to the wild-type proteins. We evaluated the abilities of traditional nucleic acid biosecurity screening tools to detect these synthetic homologs and found that, of tools tested, not all could previously detect such AI-redesigned sequences reliably. However, as we report, patches were built and deployed to improve detection rates over the course of the project, resulting in a final mean detection rate over tools of 97% of the synthetic homologs that were determined, using in-silico metrics, to be more likely to retain wild-type-like function. Finally, we make recommendations on approaches for studying and addressing the rising risk of adversarial AI-assisted protein engineering attacks like the one we identified and worked to mitigate.

# What is an information hazard?

- Pre-publication reviewers had broadly **divergent** opinions

- The least risk tolerant often end up having outsized impact
- Publications often end up removing information until all reviewers are satisfied

- We need better vocabulary and categorization:
  - “That is a thing you can get via trivial web search”
  - “That might take you an hour to find on the web”
  - “You could understand this idea by reading these three published papers”
  - “No one knows that except deep experts”

- We need more formal norms around drawing & defending specific lines