UNCLASSIFIED

SAND2019-11338PE

Cybersecurity in DNA Design and Verification Tools: Risks and Solutions

PRESENTED BY

Corey Hudson
Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Cyberbiosecurity is not scenario independent

> If you can't measure it, you can't improve it.
> -Peter Drucker

Questions for scenario planning in cyberbiosecurity:

- What actual events and activities have occurred?
- What makes the synthetic biology and genomics industries unique?
- What are high risk vulnerabilities?

Mitigating cyberbiosecurity vulnerabilities through Emulation

- **Genomic research:**
  - Modeling an at-scale genomics facilities – designed to handle high throughput human genomic data
  - Identify important vulnerabilities, common in genomics pipelines
- **Synthetic biology research:**
  - Modeling research synthetic biology facilities
  - Working to identify critical industry-wide vulnerabilities

What actual events and activities have occurred?

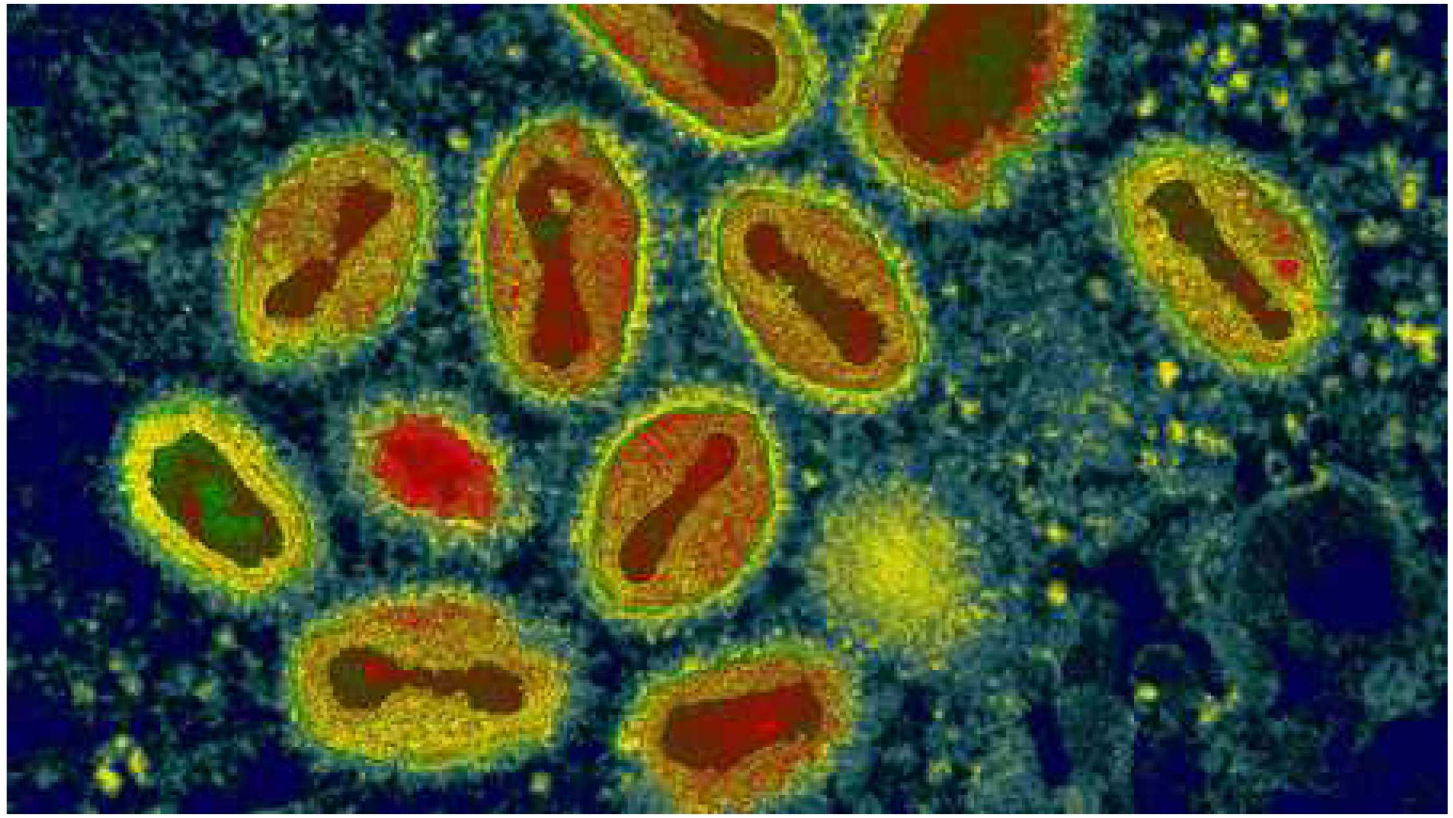


Image from Science 2017

2017 – Reconstitution of poxvirus for $100K

Led by virologist David Evans @ University of Alberta in Edmonton in collaboration with Geneart

Demonstrated the relative ease of stitching together 212K bp variola genome

What actual events and activities have occurred?

Image from Ney et al. 2017 *USENIX*

Demonstrated the use of DNA synthesis and DNA sequencing in combination to hack a system

Required a vulnerability in an exploited utility (fqzcomp) induced by the authors

Extremely sophisticated, yet highly fragile attack vector

What actual events and activities have occurred?

2017 – Merck was hit by NotPetya Ransomware attack. Cost estimates >$700M in damages

Disrupted Gardasil (HPV Vaccine) production so severely – forced to borrow vaccine from CDC stockpiles

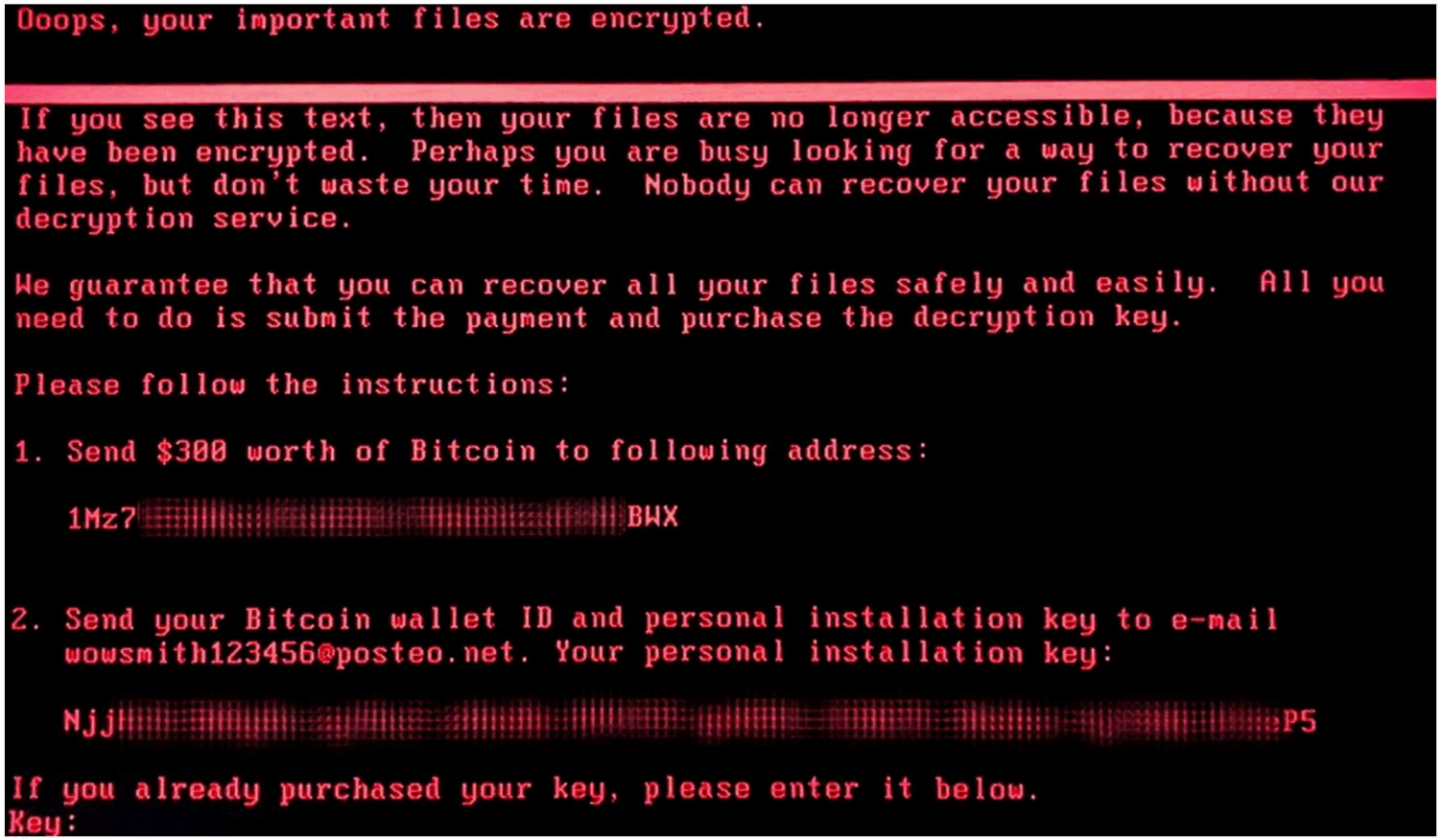
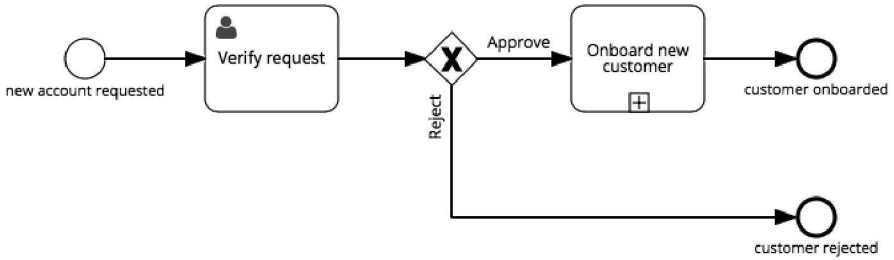


Image from NotPetya Ransomeware Front Page

What is the current threat model in DNA Design and Verification?



Know Your Customers – Limit Customers to Known Entities

Select Agent Databases to Prevent Manufacture

Advantages and Limitations of Know Your Customers Rules

Follows on regulations in the financial industry developed in the US in 2001 to deter terrorist behavior

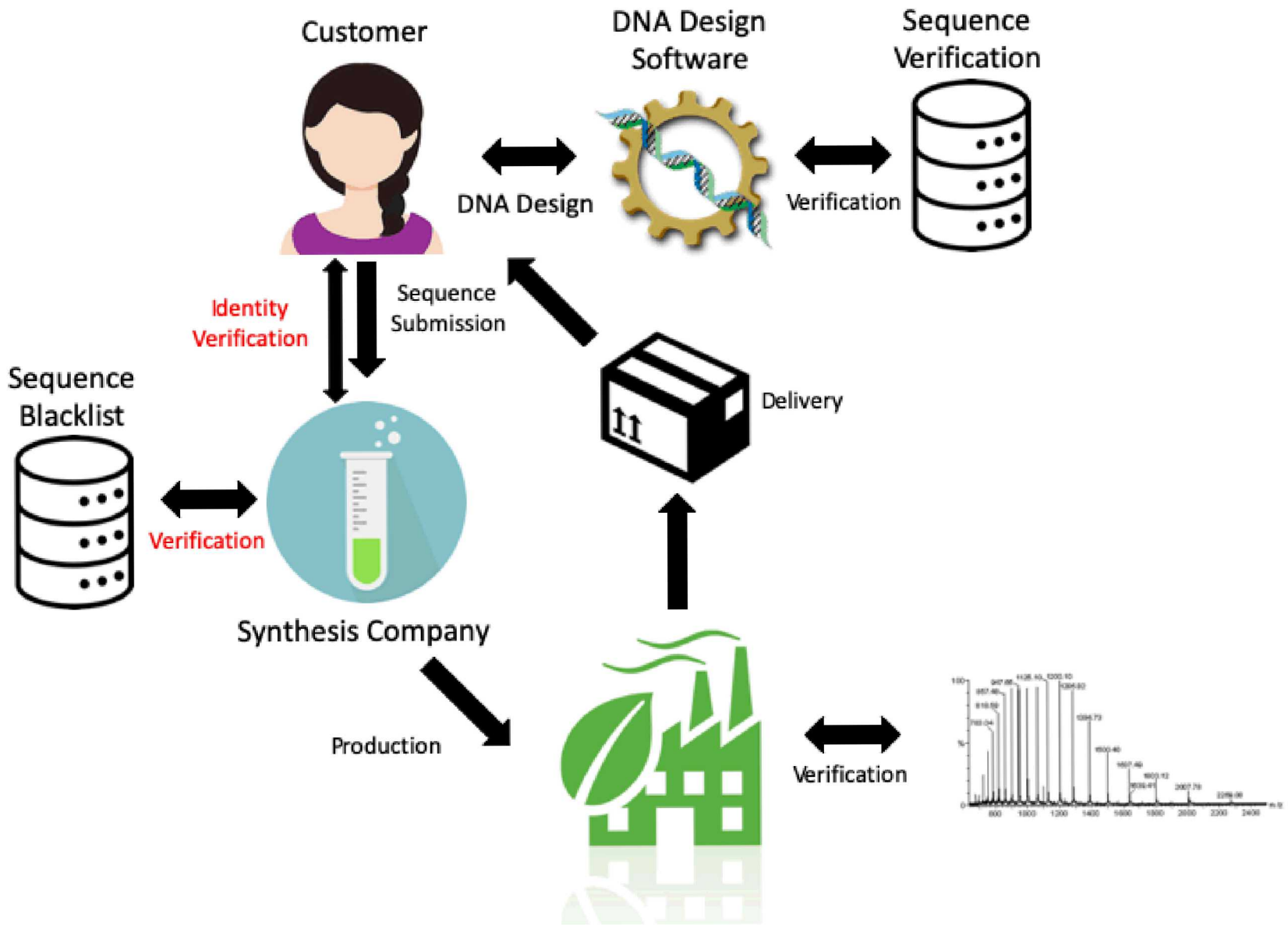
Know Your Customers limits anonymous suspicious behavior and provides a means of punitive action and regulation sufficiency

Insider threat, social engineering, extortion, threats and various techniques of *phishing* are possible issues

Ideal cases for a select agents database

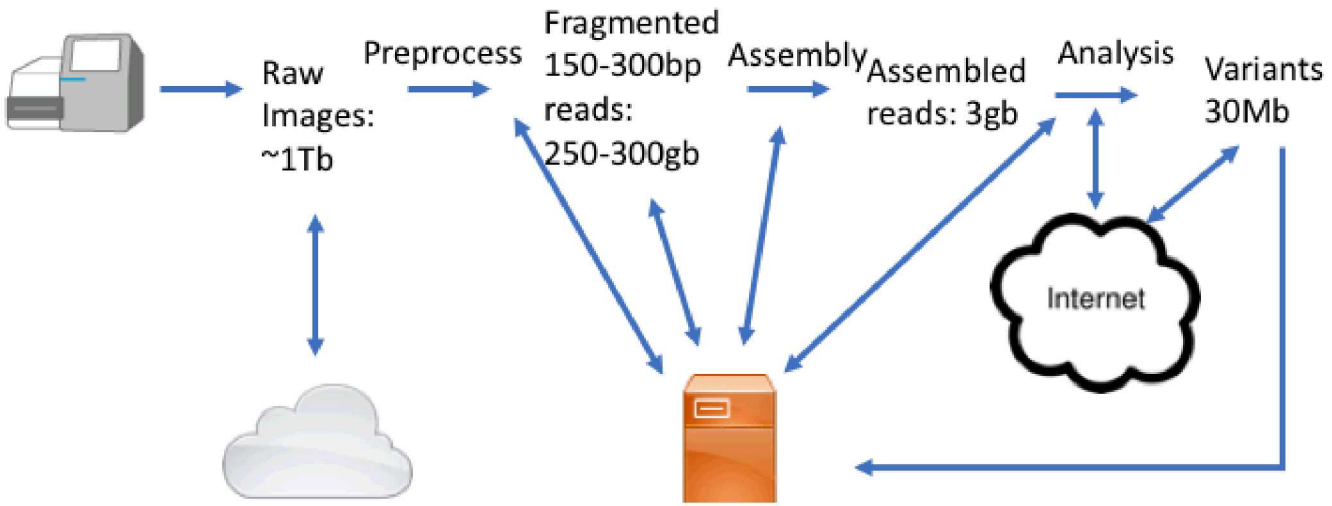
1) **Preventing accidental manufacture** – customer mistakenly ordered something dangerous
2) **Preventing BSL escalation** – limit the synthesis lab's scope of work
3) **Preventing manufacture with malicious intent** – a threat actor wishes to manufacture something dangerous

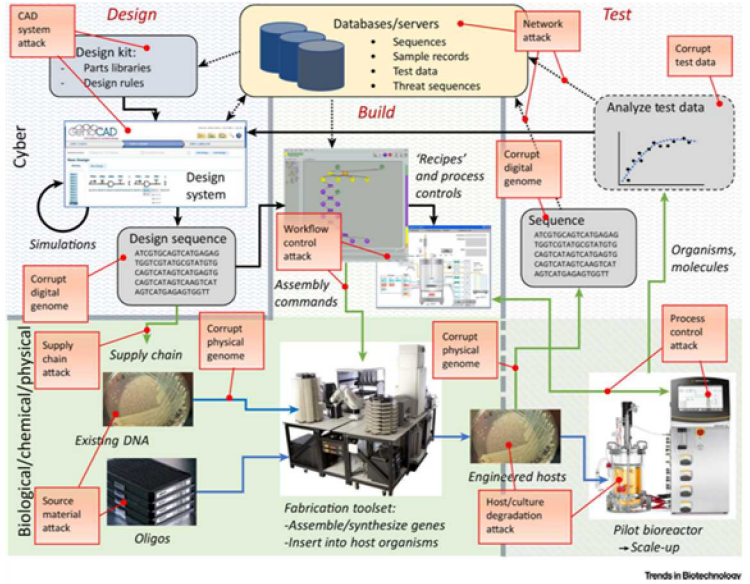
Standard Practice of DNA Design, Verification and Synthesis



Genomics and Synthetic Biology are Increasingly Computational Fields

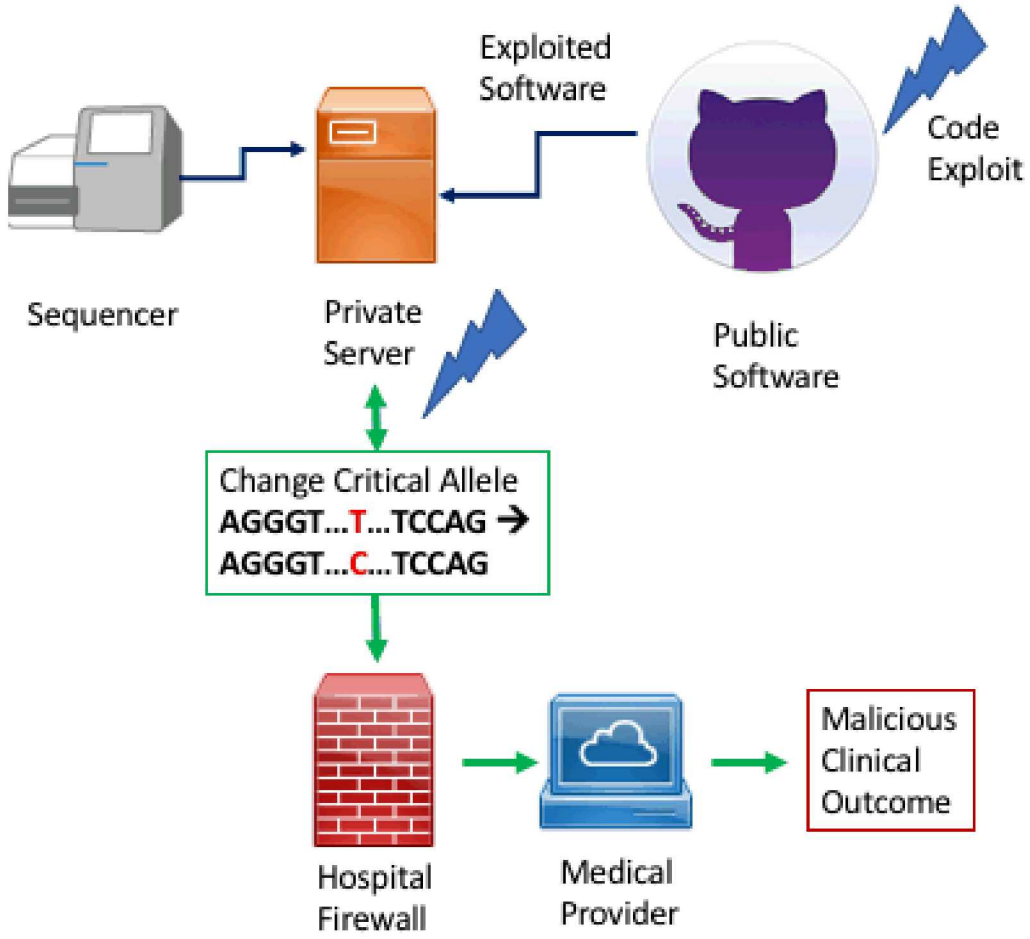
- **Genomics** – Computational complexity and data bigness
- **Synthetic biology** – Automation driven (operations, robotics, validation)





Peccoud et al. (2018) Cyberbiosecurity: From Naïve Trust to Risk Awareness. *Trends in Biotechnology* 36(1): 4-7.

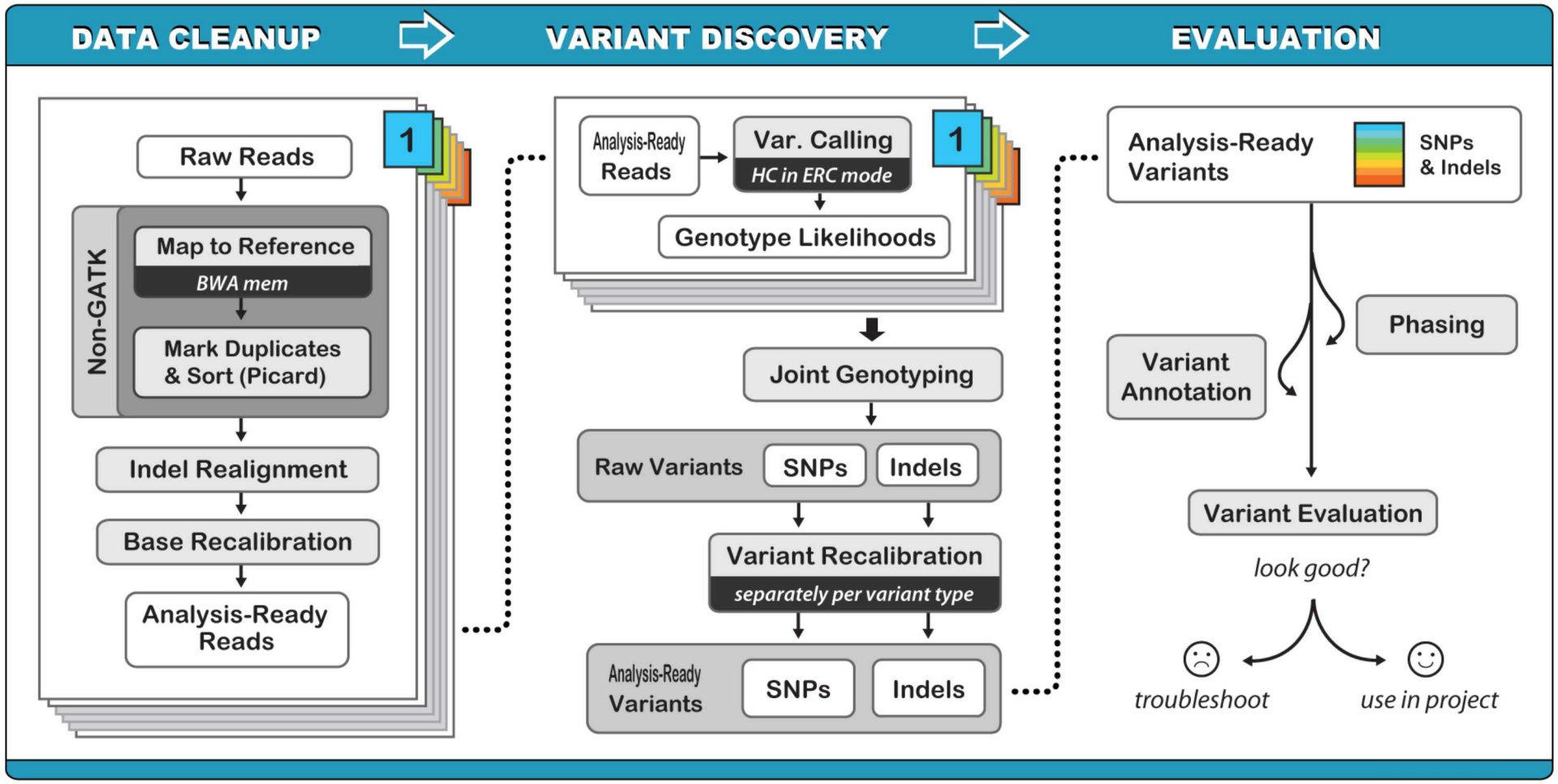
Threat Model Case study: Public Software, Databases and Raw Data



Demonstrates three risks:

1) Behind the scenes threat
2) Environmental risk
3) Irrecoverable and suspect raw data

Standard Best-Practices for Genomic Variant Detection



Best Practices Pipeline for Variant Detection, per BROAD Institute:
https://gatkforums.broadinstitute.org/gatk/discussion/3238/best-practices-for-variant-discovery-in-dnaseq

BWA had a vulnerability in its native codebase

```
bntseq_t *bns_restore(const char *prefix)
{
    char ann_filename[1024], amb_filename[1024], pac_filename[1024], alt_filename[1024];
    FILE *fp;
    bntseq_t *bns;
    strcat(strcpy(ann_filename, prefix), ".ann");
    strcat(strcpy(amb_filename, prefix), ".amb");
    strcat(strcpy(pac_filename, prefix), ".pac");
    bns = bns_restore_core(ann_filename, amb_filename, pac_filename);
    if (bns == 0) return 0;
    if ((fp = fopen(strcat(strcpy(alt_filename, prefix), ".alt"), "r")) != 0) { // read .alt file if present
        char str[1024];
        khash_t(str) *h;
        int c, i, absent;
        khint_t k;
        h = kh_init(str);
        for (i = 0; i < bns->n_seqs; ++i) {
            k = kh_put(str, h, bns->anns[i].name, &absent);
            kh_val(h, k) = i;
        }
        i = 0;
        while ((c = fgetc(fp)) != EOF) {
            if (c == '\t' || c == '\n' || c == '\r') {
                str[i] = 0;
                if (str[0] != '@') {
                    k = kh_get(str, h, str);
                    if (k != kh_end(h))
                        bns->anns[kh_val(h, k)].is_alt = 1;
                }
                while (c != '\n' && c != EOF) c = fgetc(fp);
                i = 0;
            } else str[i++] = c; // FIXME: potential segfault here
        }
        kh_destroy(str, h);
        fclose(fp);
    }
    return bns;
}
```

1024 byte buffer

If a .alt file has a line >1024 bytes it will overflow here

BWA interactions with outside data

Sequencer

Images

BCL2FASTQ

FASTQ FILES

GitHub

Software Repo
https://github.com/lh3/bwa

Software

BWA

Request for Data

Zipped Indices

NCBI

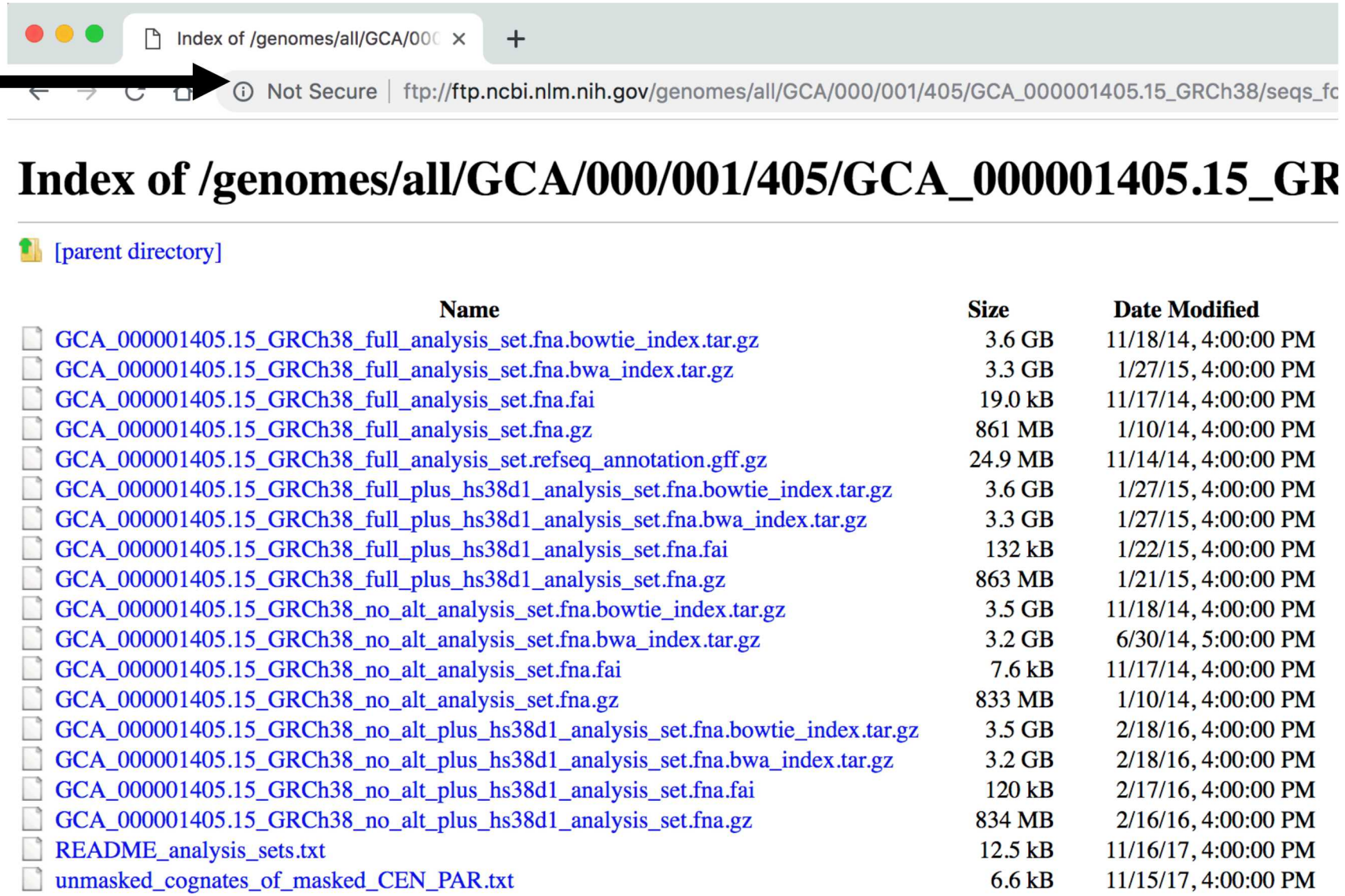
Database Server
ftp://ftp.ncbi.nlm.nih.gov

SAM FILES

Additional Pipeline Steps

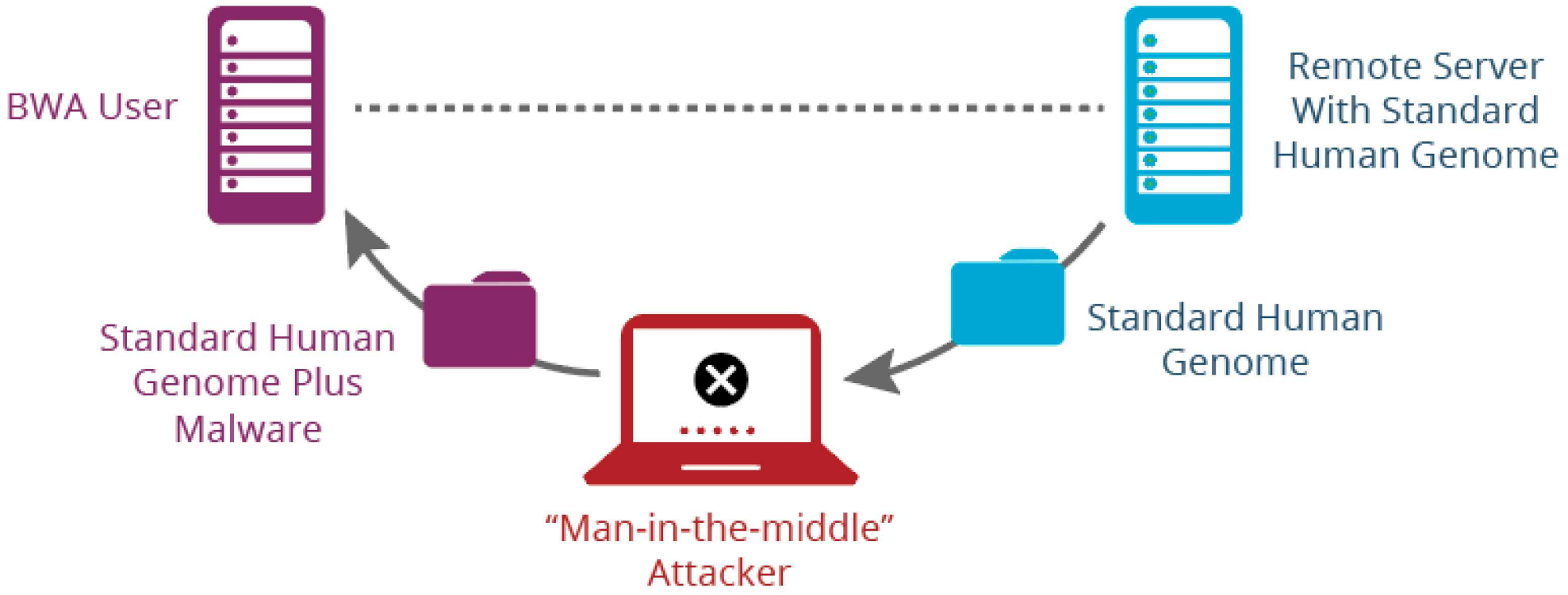
Indices are delivered over unencrypted channels

FTP Protocol



No checksums to validate data transfer

Modelling Network Vulnerability



Anatomy of the threat

Setup

1. Get presence of host network
2. Spoof FTP data transfer
3. Have remote machine stop database transfer from NCBI and deliver poisoned .ALT file
4. When BWA reads poisoned .ALT file it will trigger a buffer overflow
5. Use overflow to issue a command to overwrite all .FASTQ files in the system to change one sequence, to another sequence and complete analysis

Outcome

1. Continue to process files using standard workflow
2. Result will be a different genotype for all files in the system
3. Final .vcf files (standard genotype format) will report new genotypes

Proof of Concept: In-Place Data Manipulation

1. Search for unique sequence in all .fastq files:

CACAGAA**A**GCTAATGGG

2. Replace with new sequence differing by one character:

CACAGAA**C**GCTAATGGG

3. Empirical Result in VCF for all files:

- Statistically significant difference between files – with and without exploit
- **Without exploit** – Genotype **AA** at chromosome 12 position 64544989
- **With exploit** – Genotype **AC** ($P<10^{-200}$) at same position

Responsible Disclosure of Bug

1. Developed patch for the bug
2. Contacted developers
3. Deployed patch to main source code
4. Contacted DHS US-CERT, NIST & MITRE to report bug
5. Developed a proof of concept
6. Identified mitigations to similar bugs
7. Publicized and discussed bugs with other developers

Mitigations and Lessons Learned for Synthetic Biology

1. Need an evolving trust model in genomics and synthetic biology
2. Validation process and environment need to be vetted
3. Trusted relationships need to be verified
4. Databases need to be securely handled
5. Software tools need to be evaluated continuously and there needs to be a community
6. Maintain limited access
7. Protect raw data as much as possible

What are the cybersecurity risks?

Automated Synthetic Biology Systems have Unique Cyberbiosecurity Risks

1. Proprietary Data Risks
2. Operational and system security risks
3. Sabotage
4. Reduced ability to monitor bioproduction
5. Unintentional and stand-off manufacture risks, especially of manufacture of select agents or bioweapons

Automated systems lower the level of sophistication necessary to carry out malicious activities

Summary

- Our assumptions about the misuse of synthetic biology and its use in malicious behavior has been historically determined by wetlabs
- The barriers to entry are lower in cybersecurity because of the wide availability of cybersecurity threats
- Genomic cybersecurity issues provide a lens for viewing cyberbiosecurity generally
- DNA designs are a source of system input, meaning that unless the input is validated output can be manipulated and systems can be damaged

Thank you!