UNCLASSIFIED

SAND2019-4923PE

Genomic and Synthetic Biology Cybersecurity

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.

Support

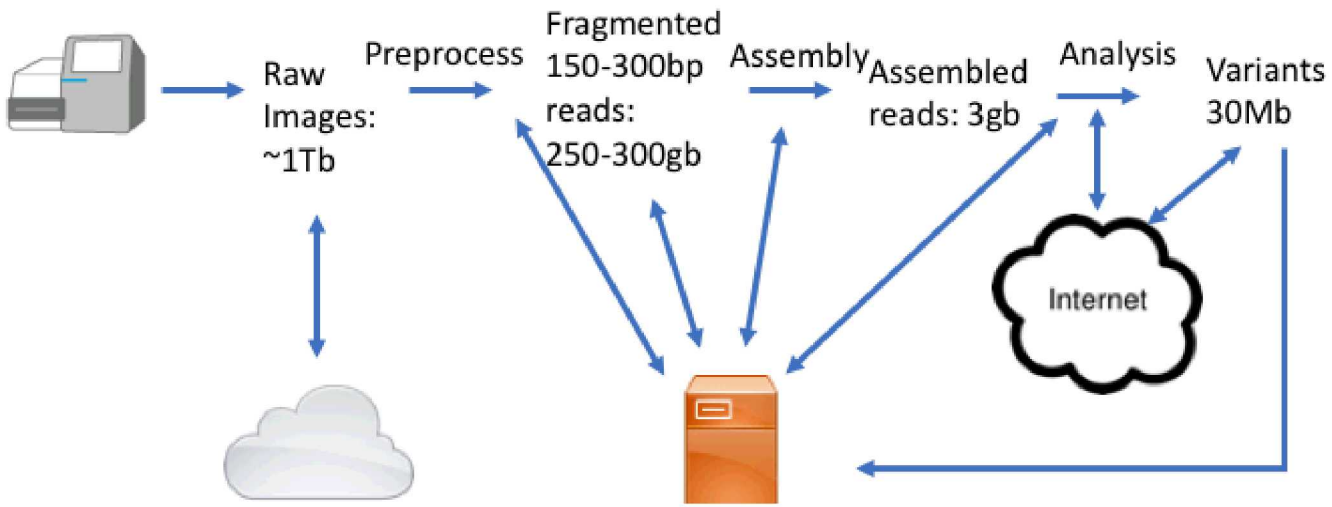
Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

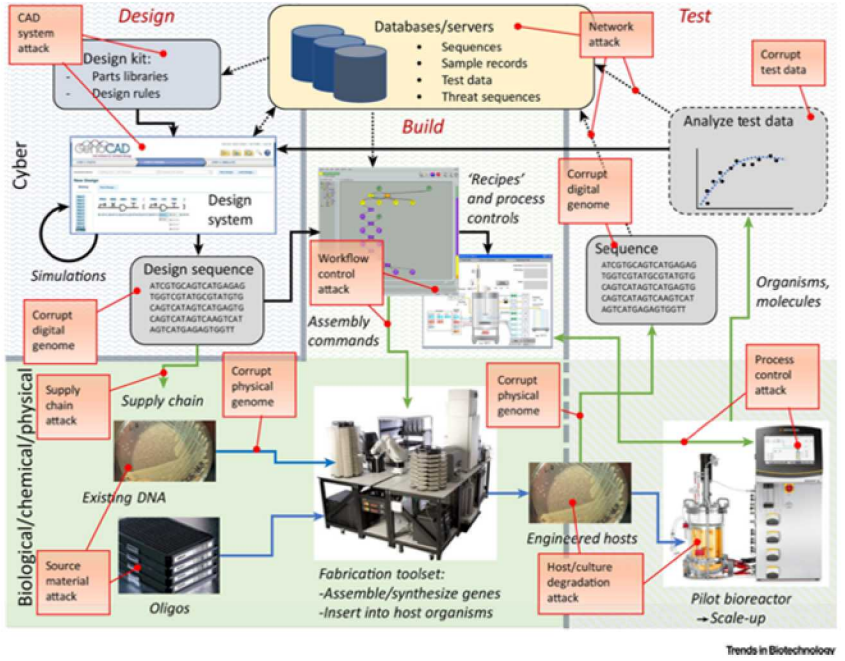
Two areas of focus: Genomic and Synthetic Biological

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

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.

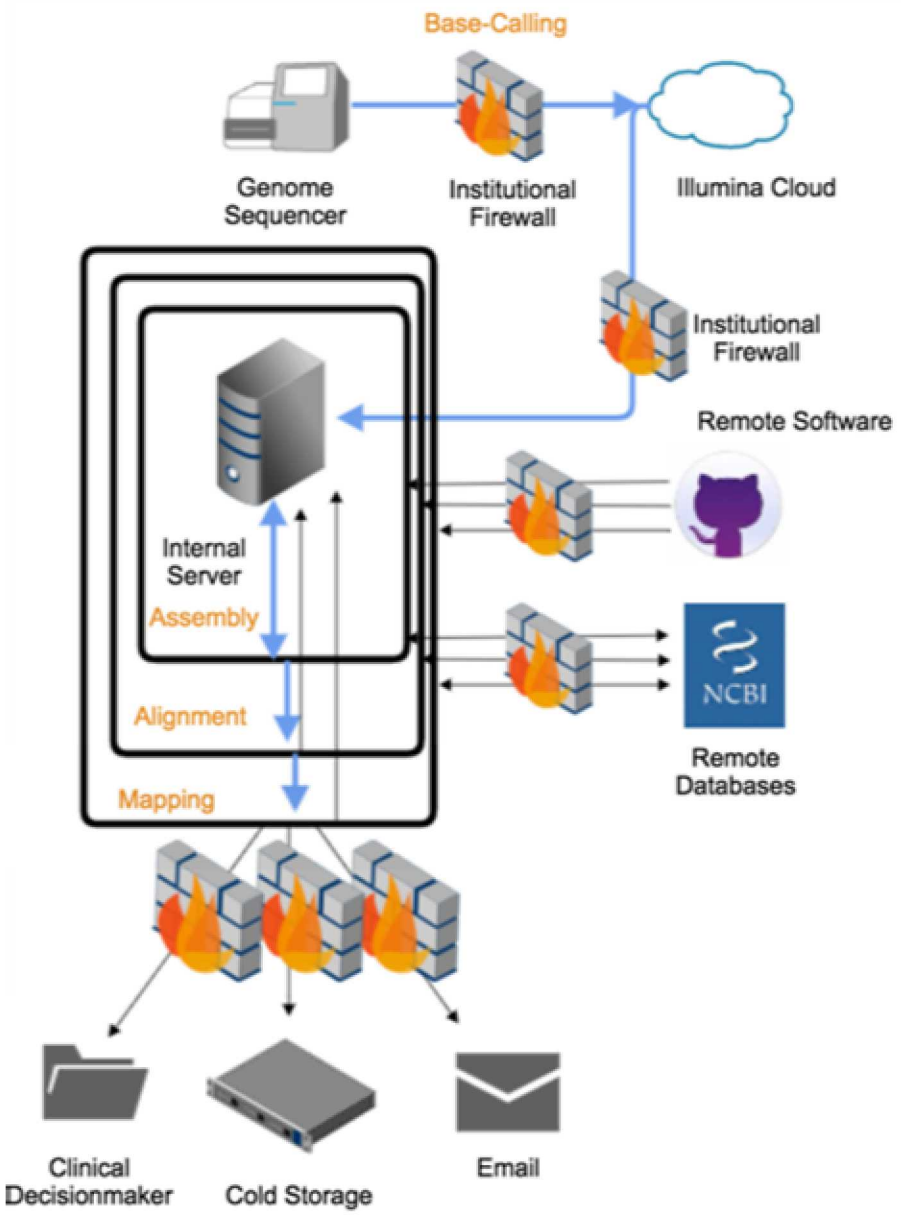
Realistic models of genomic security

Genomic systems rely on internet connected equipment, large numbers of proprietary and open-source software and interactions with the cloud.

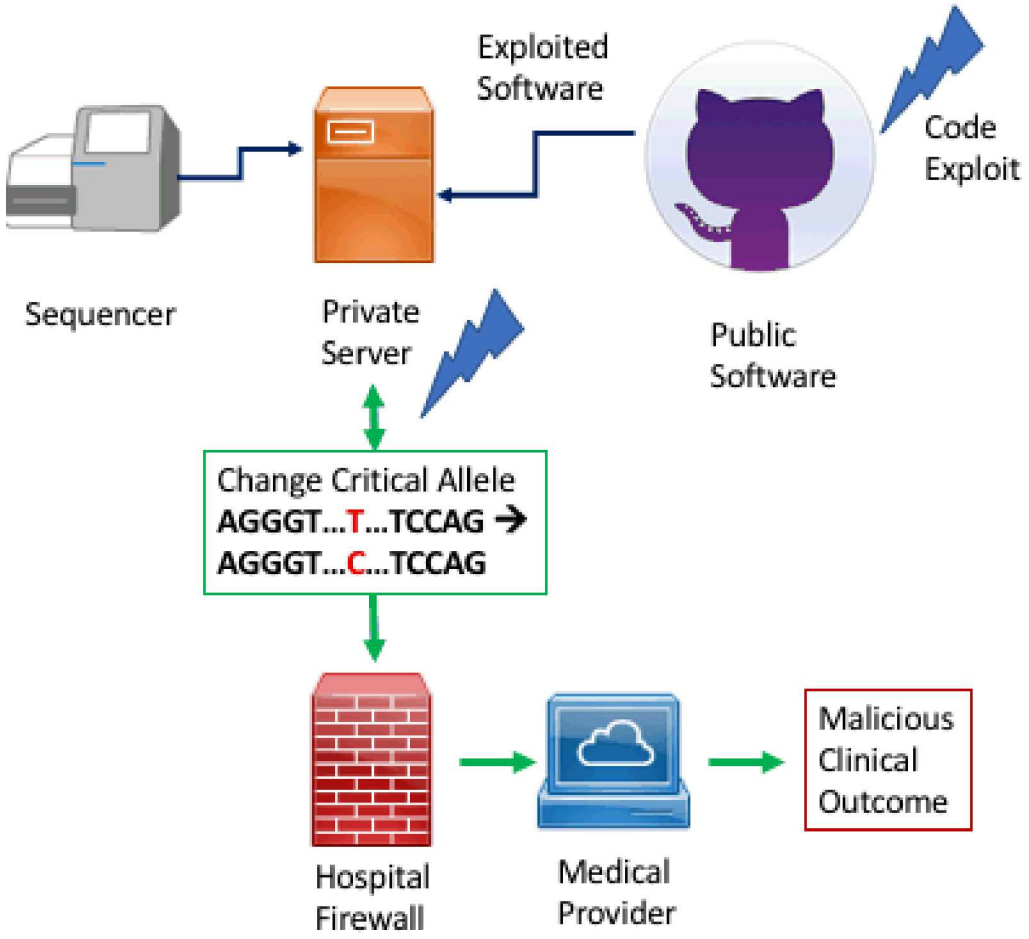
Where these interactions occur relative to an institutional firewall are variable

The vetting of software is variable

And the protocols for internet/remote interactions are variable



Case study: Public Software, Databases and Raw Data



Exploring implications of genomic software vulnerabilities under realistic security assumptions

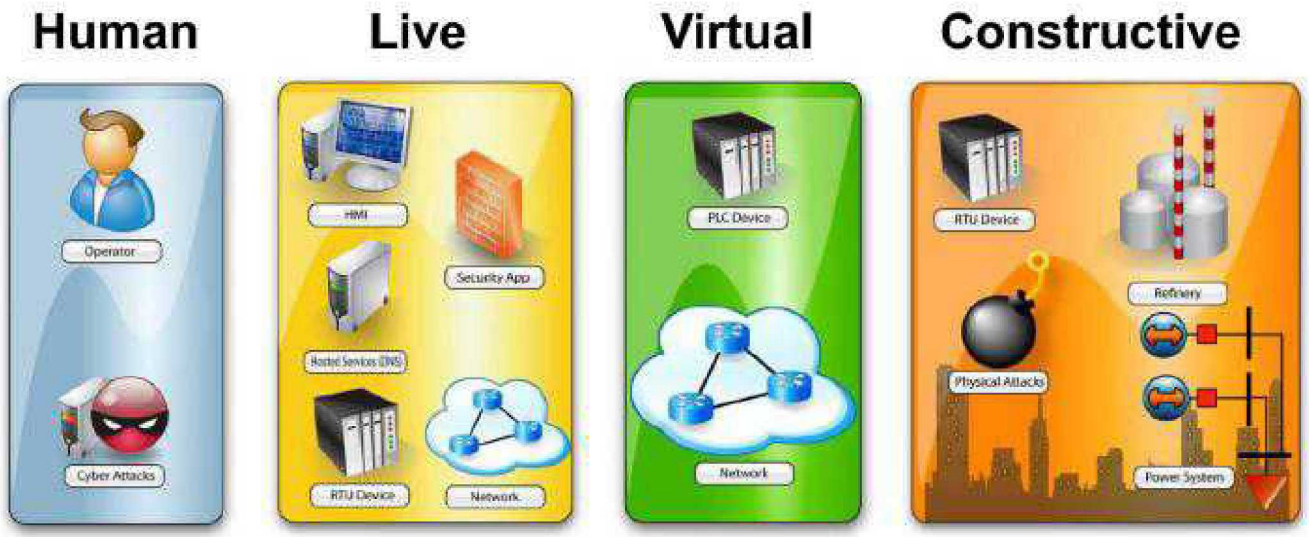
Goals for modeling genomic security

1. *Only allow **standard** simplifying security assumptions*
2. *Use existing knowledge of genomics systems to frame assumptions*
3. *Use a vulnerability that exists* in the wild
4. *Use a standard best-practices genomics pipeline*
5. *Manipulate raw data, prior to analysis and complete analysis without issuing errors*

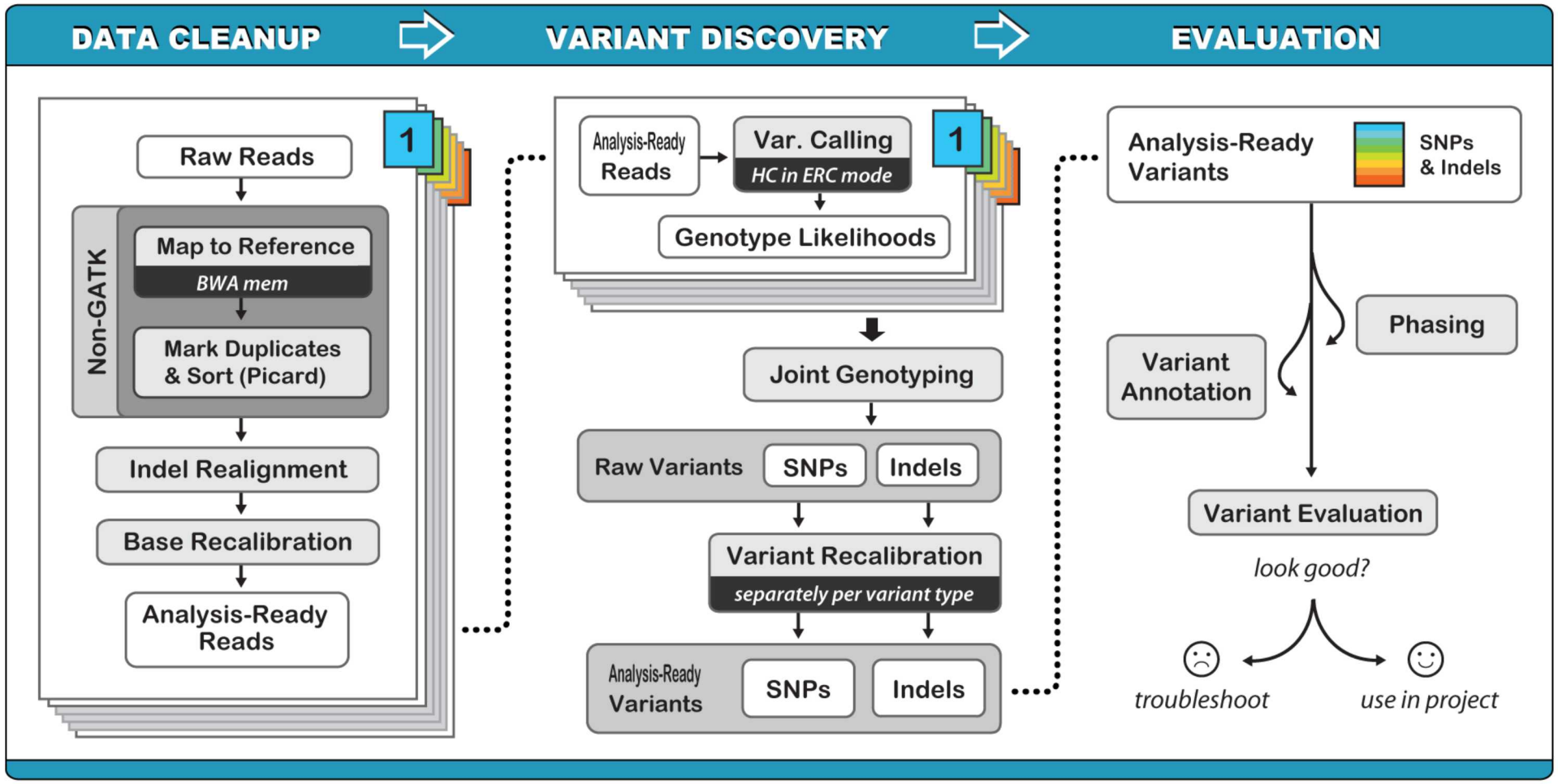
Building a Realistic Genomics Pipeline

Collaboration between University of Illinois' Genomic Pipeline and Sandia National Laboratories' Emulytics Capacity for Simulation and Assessment

UIUC has designed a genomics pipeline capable of handling hundreds-thousands of simultaneous assessments of genomic variability in a clinical context.



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

First piece of software in best-practices pipeline is BWA

1. BWA takes FASTQ files as input and maps these to a reference genome, creating a SAM file
2. In 2014, BWA developers added the ALT-aware capacity – which allowed users to map reads to a population, rather than canonical single reference
3. Since the population is always changing and requires up-to-date knowledge, the reference is hosted at a central repository
4. BWA provides a tool – bwa.kit, which accesses this data from the US National Center for Biotechnology Information (NCBI), which has provided resources for the storage and delivery of these files as a tarred and gzipped directory of indices: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/
5. The user then unzips and stores the indices provided by NCBI
6. A .**alt** file is used to index the genome and make it alt-aware

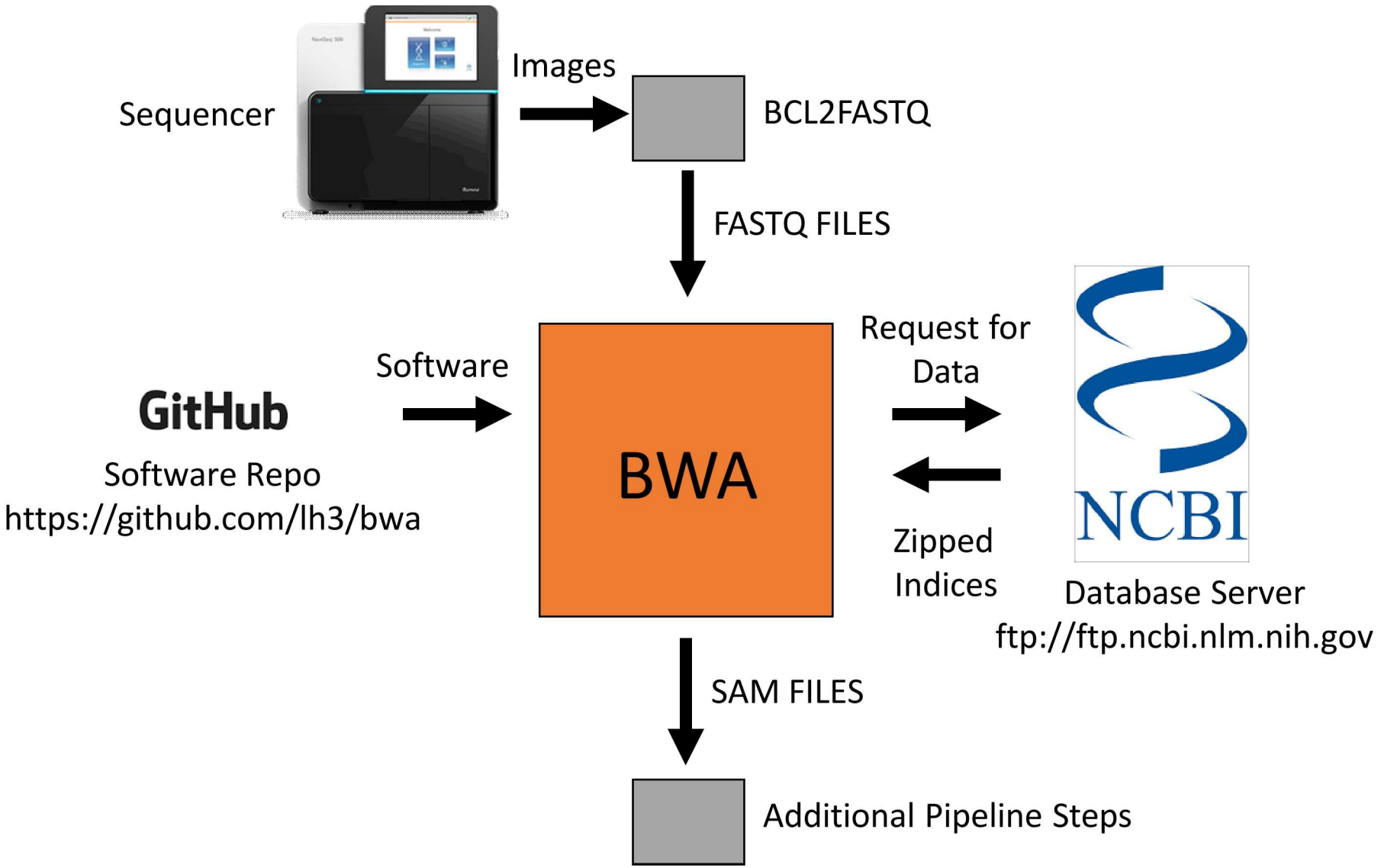
BWA has a vulnerability in its native codebase

```
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

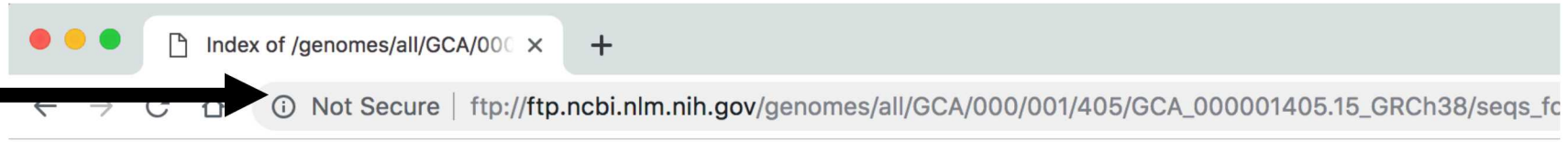
If a .alt file has a line >1024 bytes it will overflow here

BWA interactions with outside data



.ALT files are delivered over unencrypted channels

FTP Protocol



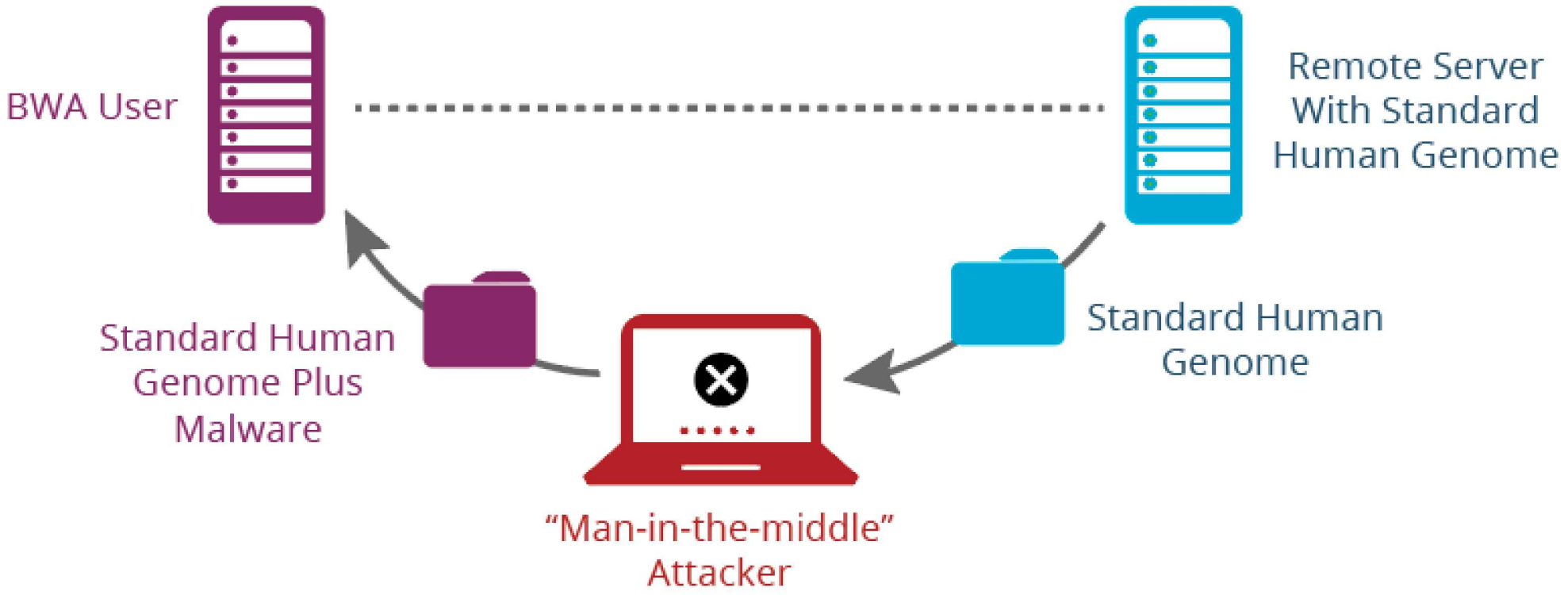
No checksums to validate data transfer

Index of /genomes/all/GCA/000/001/405/GCA_000001405.15_GR

[parent directory]

Name	Size	Date Modified
GCA_000001405.15_GRCh38_full_analysis_set.fna.bowtie_index.tar.gz	3.6 GB	11/18/14, 4:00:00 PM
GCA_000001405.15_GRCh38_full_analysis_set.fna.bwa_index.tar.gz	3.3 GB	1/27/15, 4:00:00 PM
GCA_000001405.15_GRCh38_full_analysis_set.fna.fai	19.0 kB	11/17/14, 4:00:00 PM
GCA_000001405.15_GRCh38_full_analysis_set.fna.gz	861 MB	1/10/14, 4:00:00 PM
GCA_000001405.15_GRCh38_full_analysis_set.refseq_annotation.gff.gz	24.9 MB	11/14/14, 4:00:00 PM
GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set.fna.bowtie_index.tar.gz	3.6 GB	1/27/15, 4:00:00 PM
GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set.fna.bwa_index.tar.gz	3.3 GB	1/27/15, 4:00:00 PM
GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set.fna.fai	132 kB	1/22/15, 4:00:00 PM
GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set.fna.gz	863 MB	1/21/15, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bowtie_index.tar.gz	3.5 GB	11/18/14, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.bwa_index.tar.gz	3.2 GB	6/30/14, 5:00:00 PM
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.fai	7.6 kB	11/17/14, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz	833 MB	1/10/14, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.bowtie_index.tar.gz	3.5 GB	2/18/16, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.bwa_index.tar.gz	3.2 GB	2/18/16, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.fai	120 kB	2/17/16, 4:00:00 PM
GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.gz	834 MB	2/16/16, 4:00:00 PM
README_analysis_sets.txt	12.5 kB	11/16/17, 4:00:00 PM
unmasked_cognates_of_masked_CEN_PAR.txt	6.6 kB	11/15/17, 4:00:00 PM

Modelling Network Vulnerability Using Emulytics



Steps in Attack

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

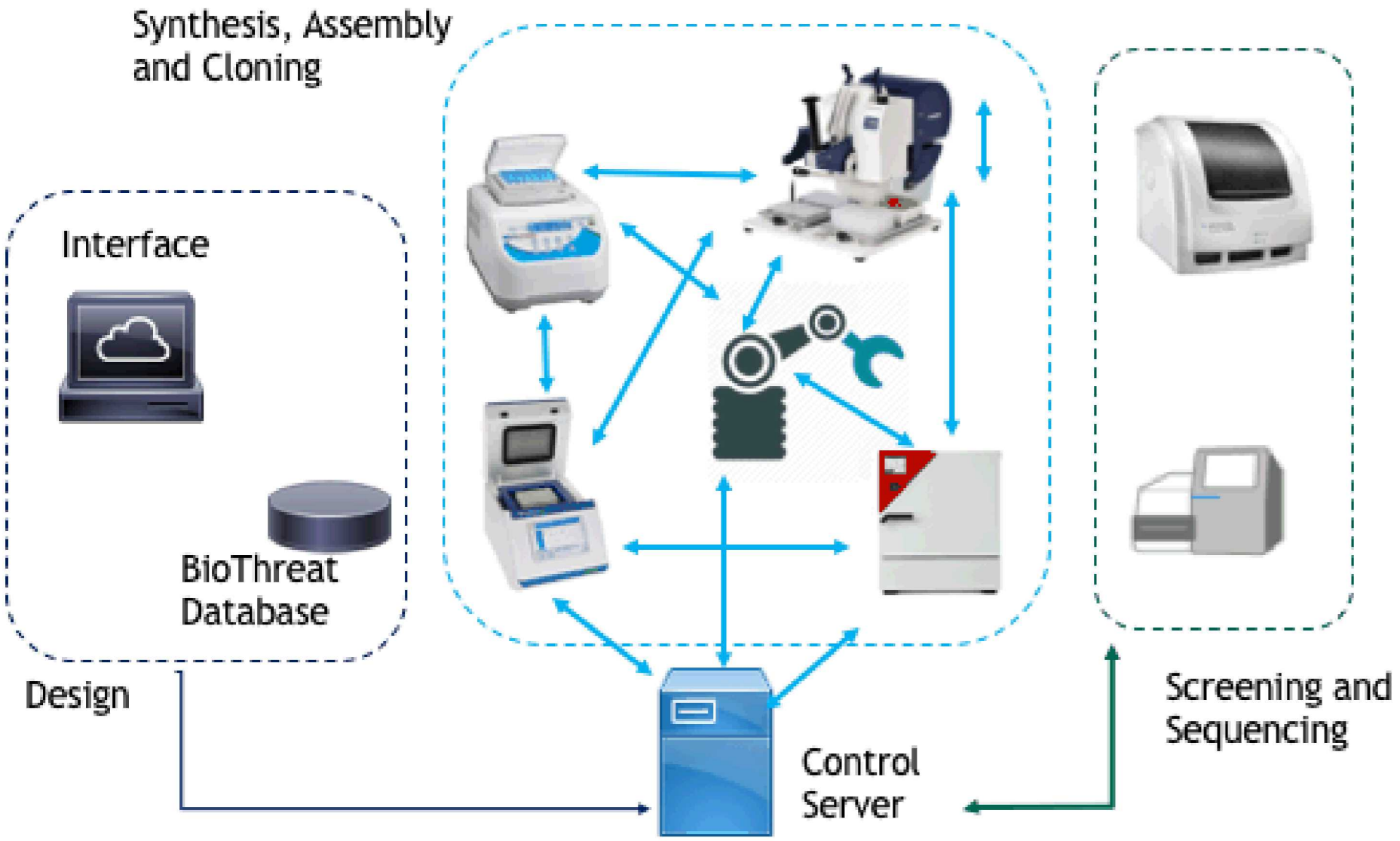
2. Replace with new sequence differing by one character:

CACAGAA**C**GCTAATGGG

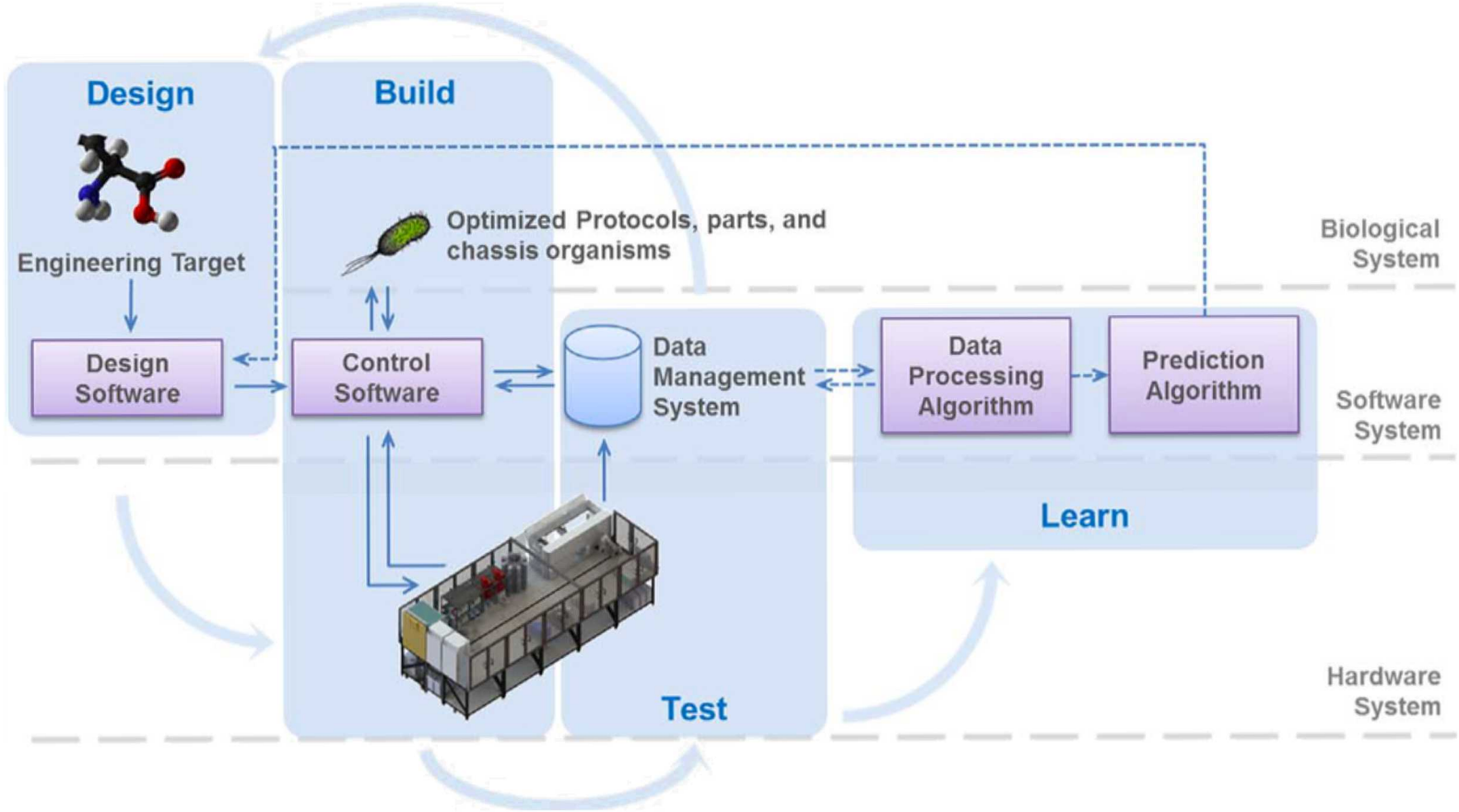
3. Empirical Result in VCF for all files:

- 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

Synthetic Biology and Cybersecurity



Automation and Computational Control



Chao et al., (2017) "Engineering biological systems using automated biofoundries" *Metabolic Engineering.*

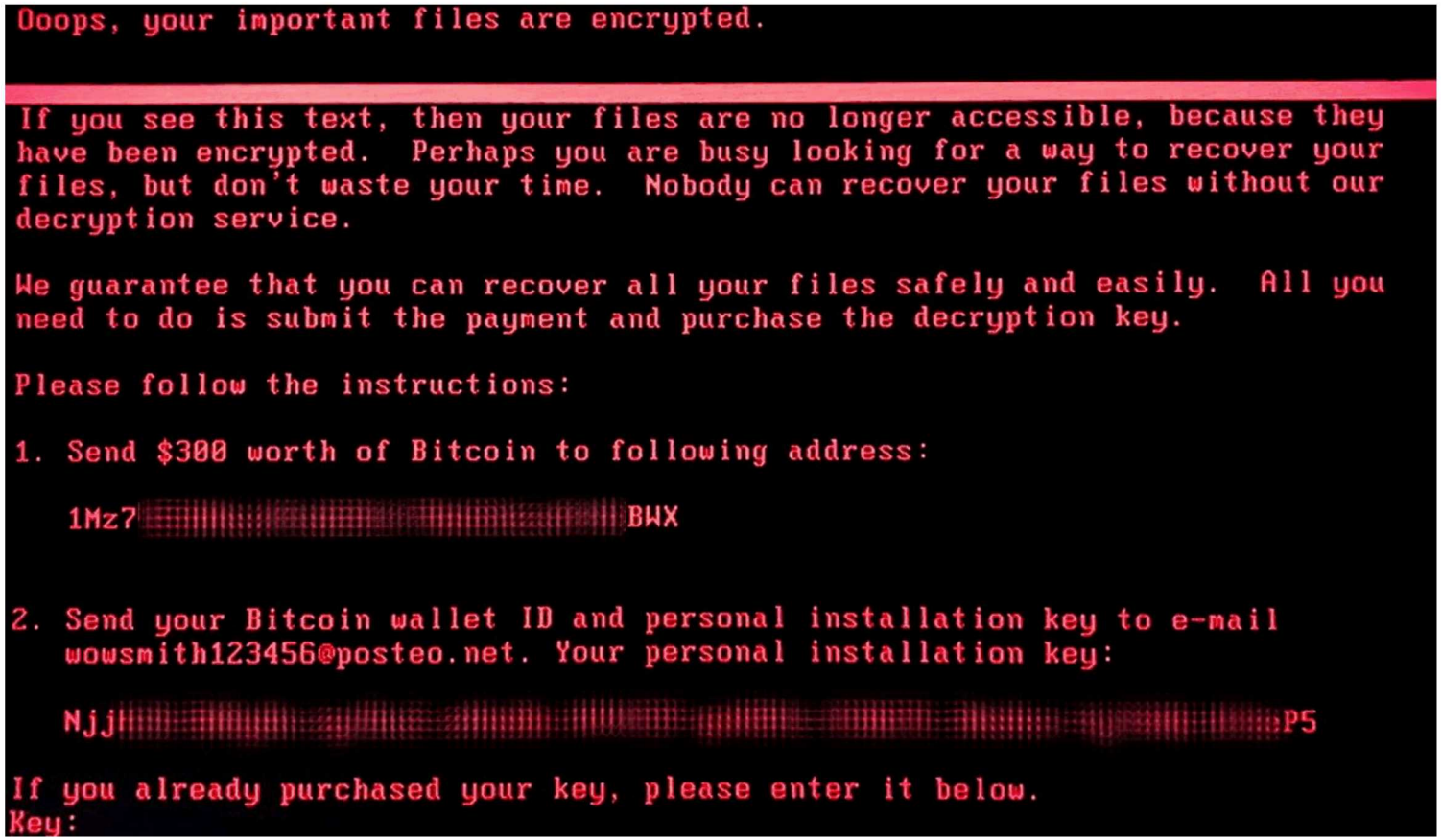
What are the cybersecurity risks?

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

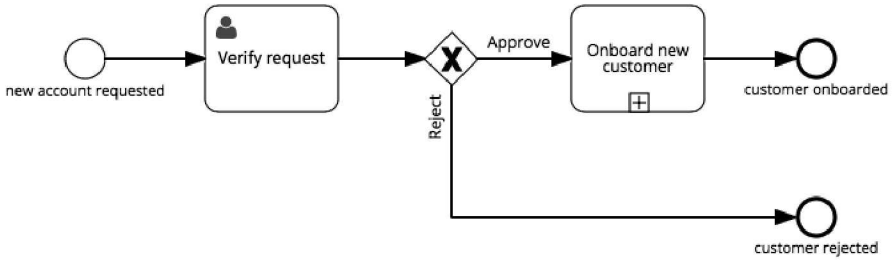
Impact of cyberattack on operations



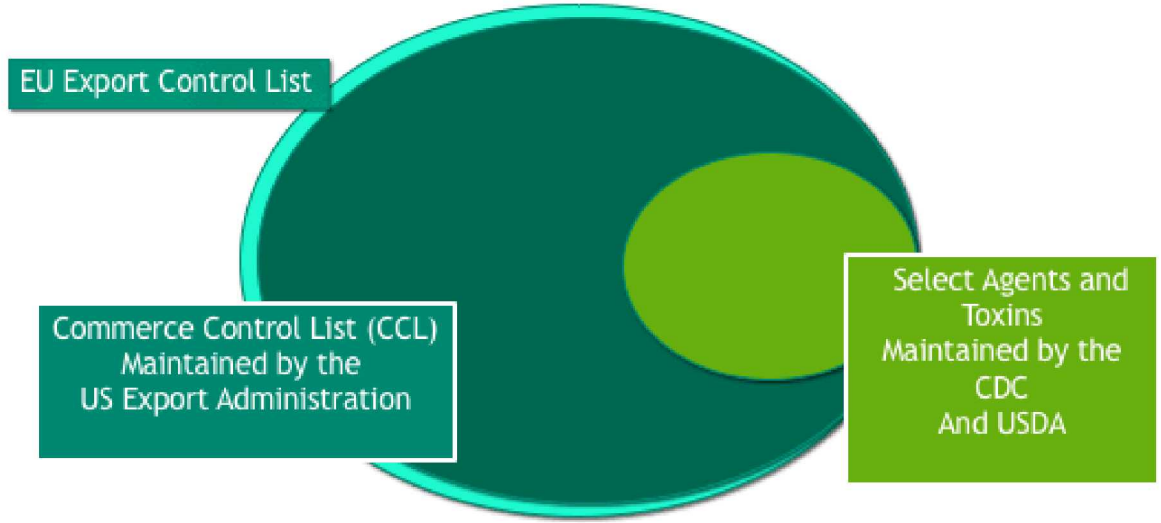
2017 – Merck was hit by NotPetya Ransomware attack. Cost estimates >$700M in damages

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

Front-end risks – malicious job submission



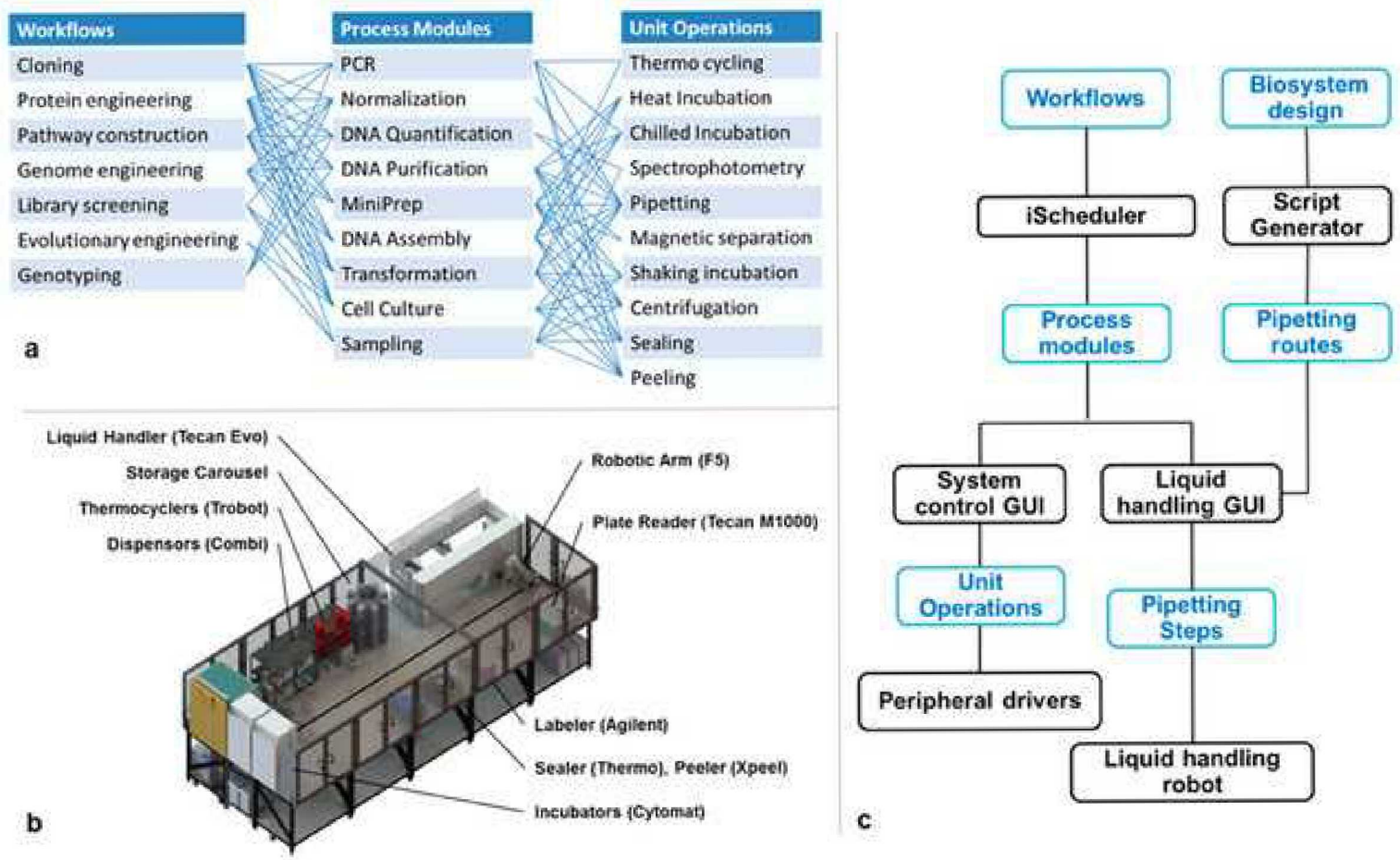
Know Your Customers – Limit Manufacture
(scalability issues)



Select Agent Databases to Prevent Manufacture
(cybersecurity issues)

Process control risks

Lab equipment is increasingly connected to a cloud or interacts with devices – how these also interact with process control poses risks



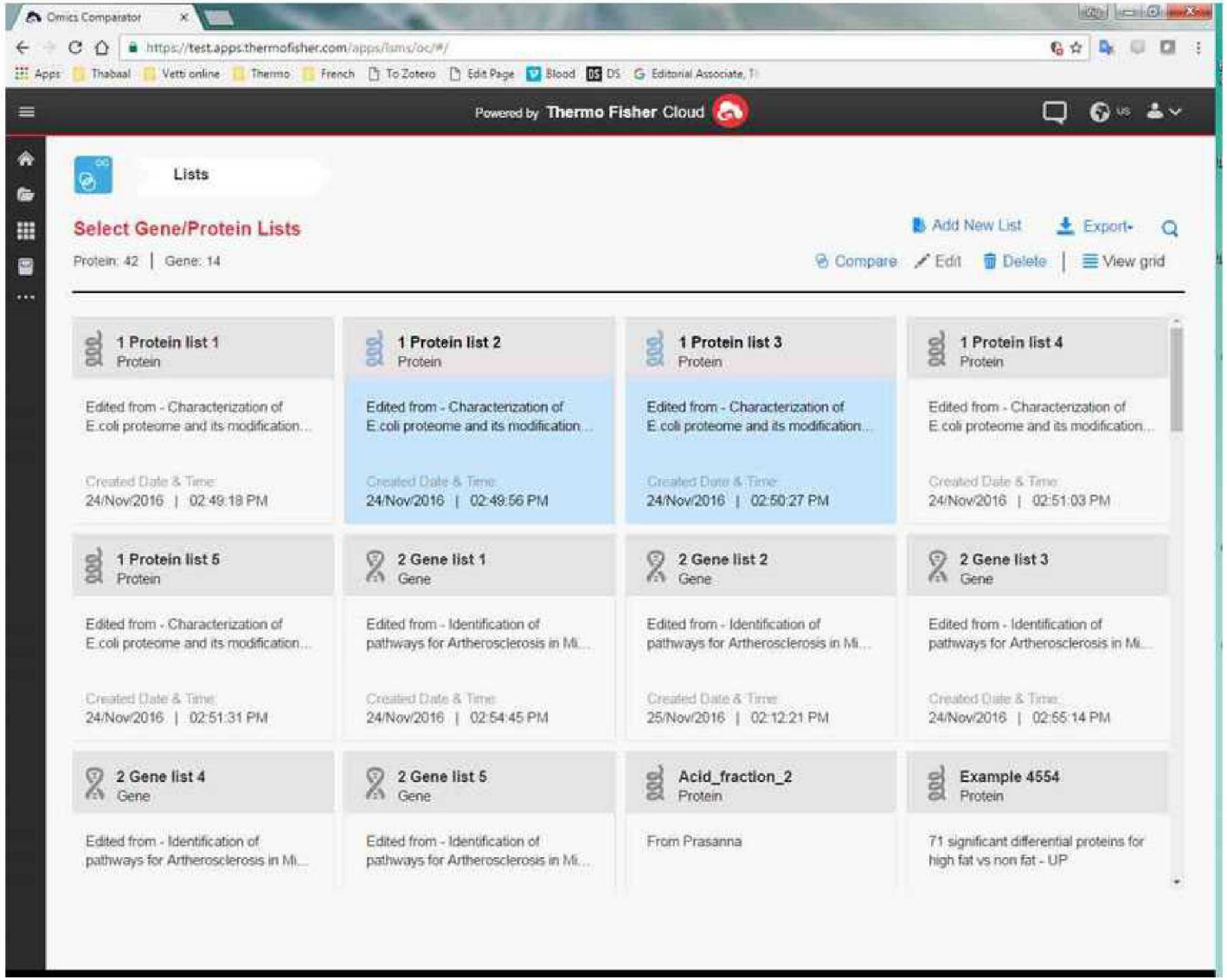
Chao et al., (2017) “Fully automated one-step synthesis of single-transcript TALEN pairs using biological foundry” *ACS SynthBio.*

Network Risks

Lab equipment is increasingly connected and these connections occur on a network, which may be vulnerable

Validation Risks

Final sanity check of process is increasingly handled by software



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 analytics and cyberweapons
- Genomic cybersecurity issues provide a lens for viewing biological cybersecurity generally
- Synthetic biology cybersecurity has several main sources of vulnerabilities – front end, process control, network and validation
- These vulnerabilities may be used to sabotage, steal, threaten systems or manufacture malicious material / bioweapons unbeknownst to the manufacturer

Thank you!