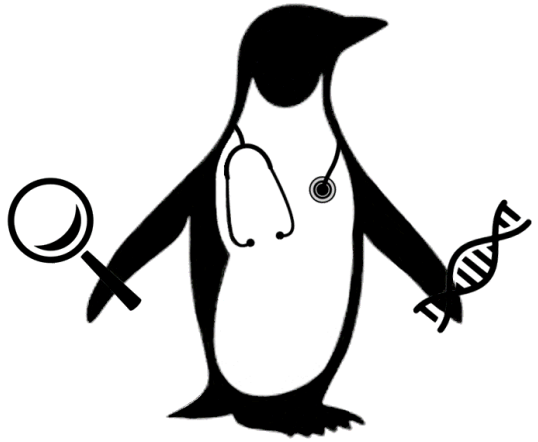
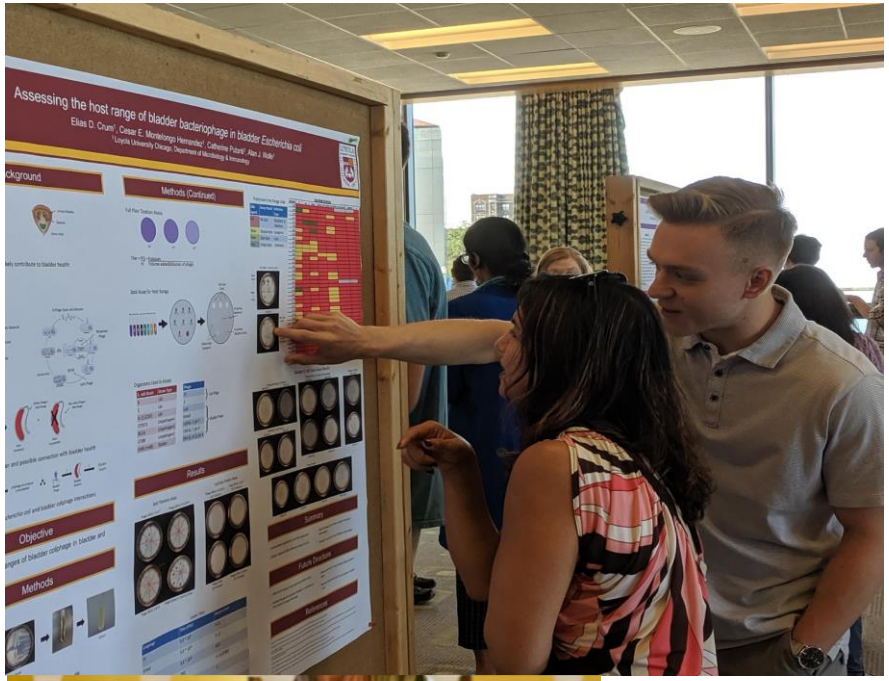
PENGQUIN:

PErsoNal Genome QUery IN healthcare and clinical practice



Elias Crum

FWO Strategic Basic PhD Fellowship Interview
September 5th, 2024
Pannel: SBWT5B



Personal Genome Data Usage?

My Journey

LOYOLA UNIVERSITY CHICAGO 1870 AD MAJOREM DEI GLORIAM

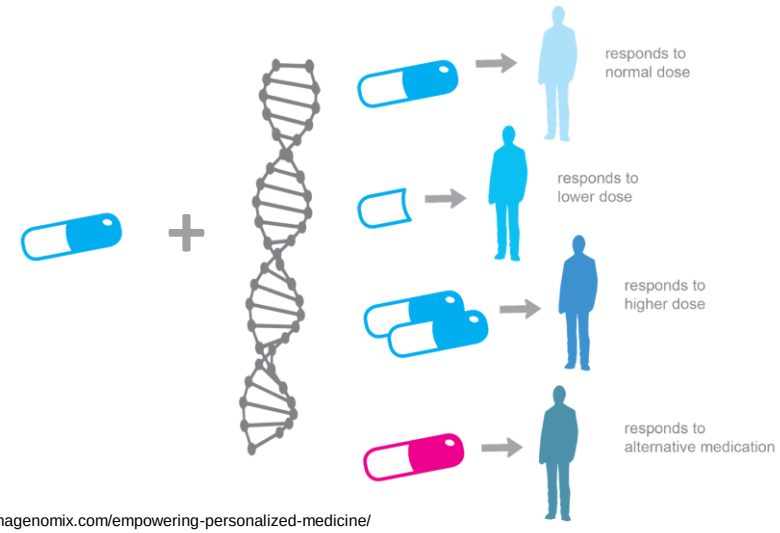
Medicine + Bioinformatics

Genomics

How can we increase **genomics** use in **medicine**?

Genomics + Healthcare

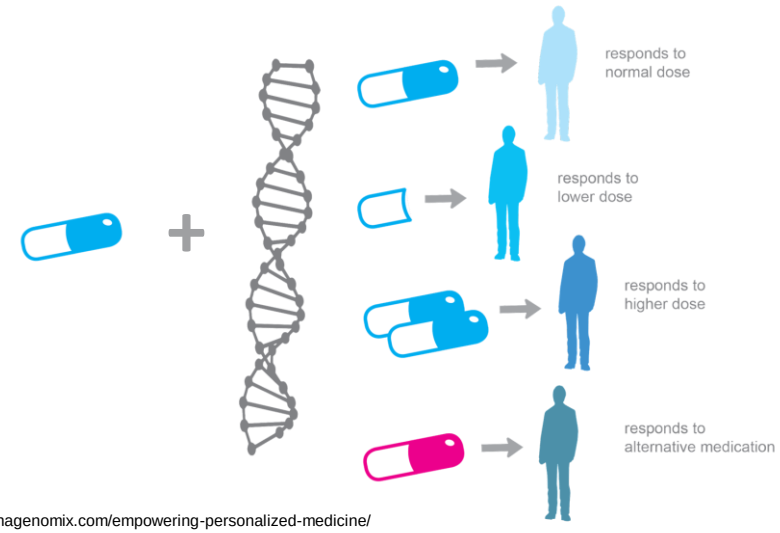
A *Example: Pharmacogenomics*



https://alphagenomix.com/empowering-personalized-medicine/

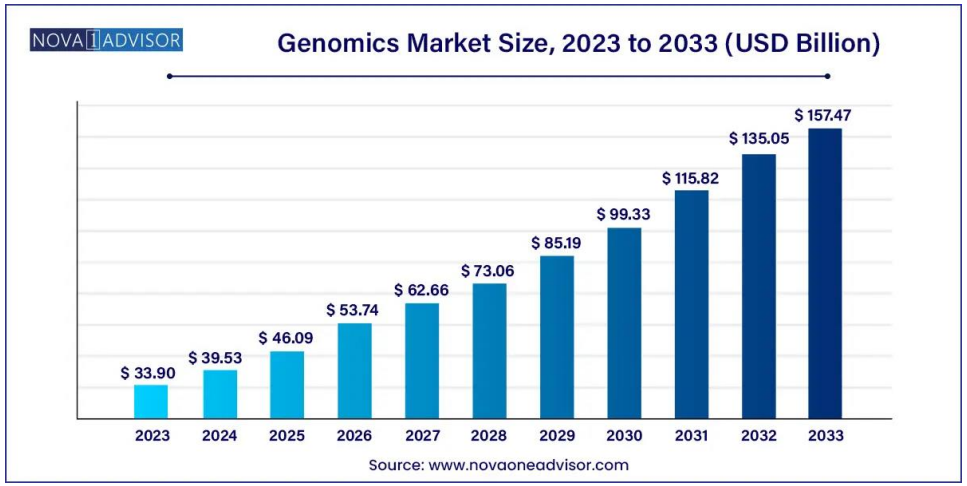
Genomics + Healthcare

A *Example: Pharmacogenomics*



https://alphagenomix.com/empowering-personalized-medicine/

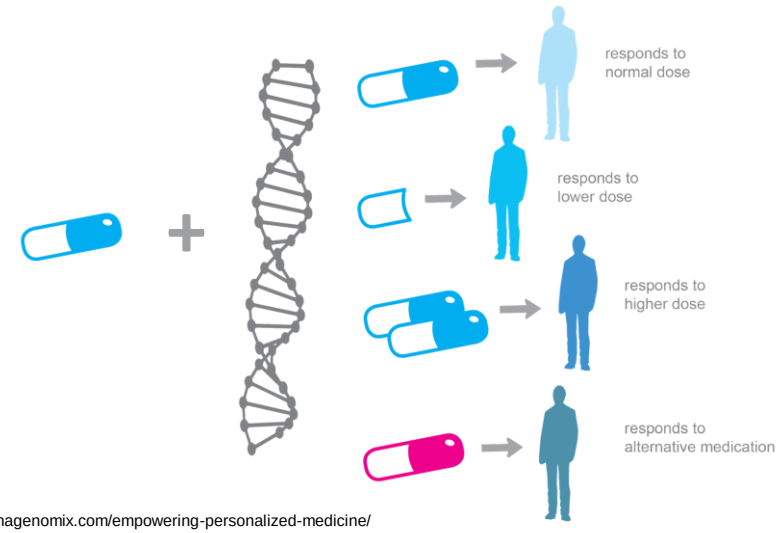
B *Genomics market is growing RAPIDLY*



https://www.novaoneadvisor.com/report/genomics-market

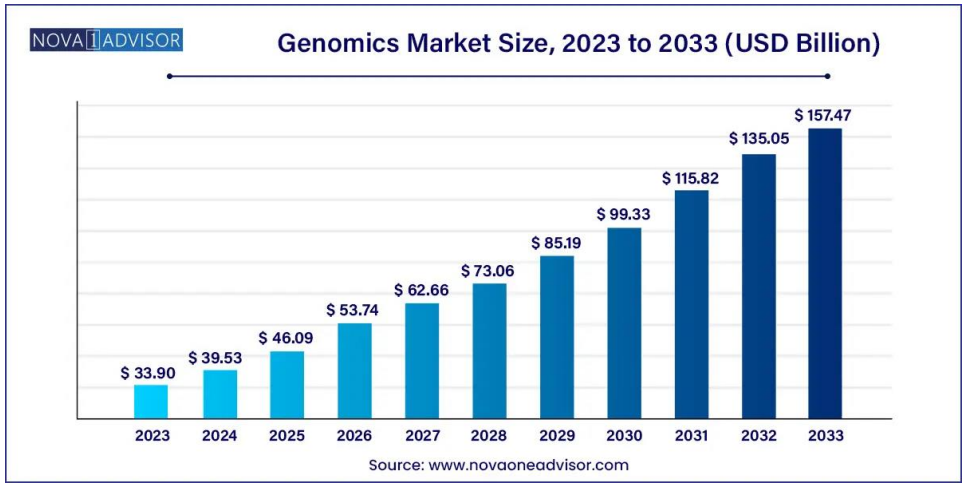
Genomics + Healthcare

A *Example: Pharmacogenomics*



https://alphagenomix.com/empowering-personalized-medicine/

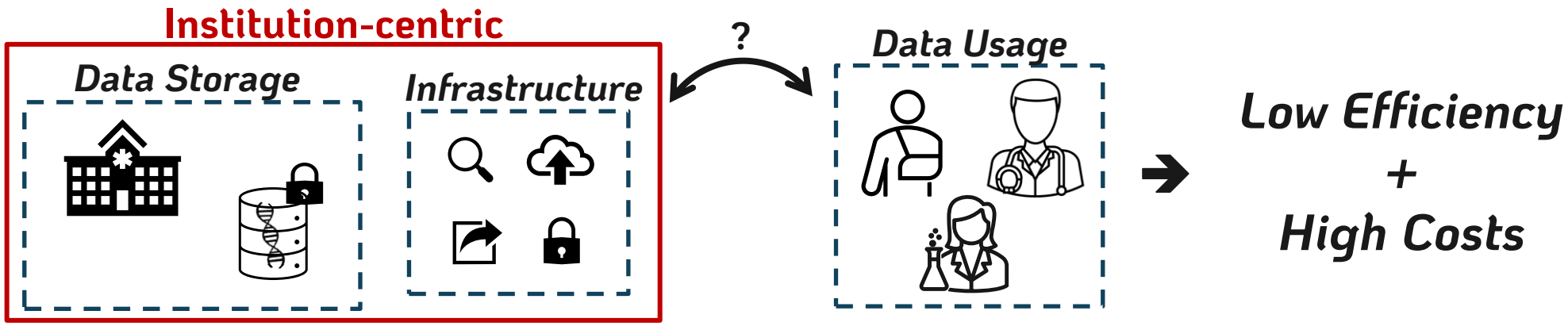
B *Genomics market is growing RAPIDLY*



https://www.novaoneadvisor.com/report/genomics-market

C

Problem:



Personal Genome Sequence Data Challenges

Size considerations:

*Variant Call Format (**VCF**)* file = **~200 MB**

Represents ~3 million mutations

Formatted as a flat (.txt) file

Other considerations:

- Governance
- Formatting
- Privacy
- Storage

PENGQUIN Goals

Improve genomic data USAGE and SCALABILITY

Improve genomic data store accessibility (while maintaining privacy)

Represent genomic data semantically as Linked Data

Query genomic data and linkages

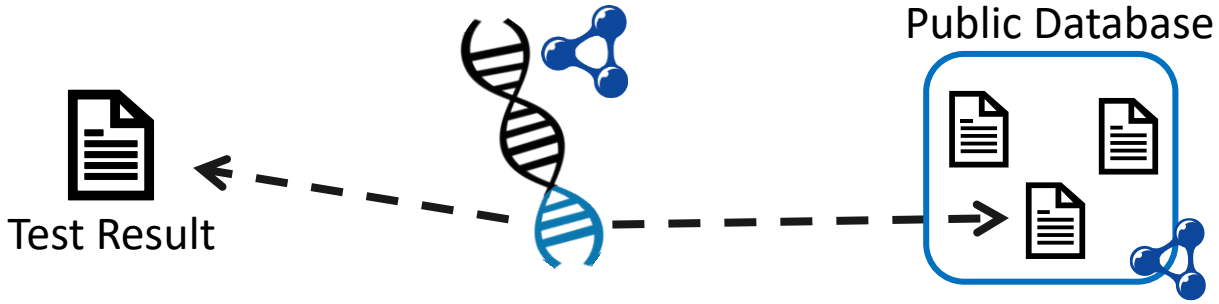
Connect to existing initiatives / applications

Proof-of-concept

Develop a framework and web application

Primary Research Challenges

WP1. Linking Genomic data to clinically-relevant data

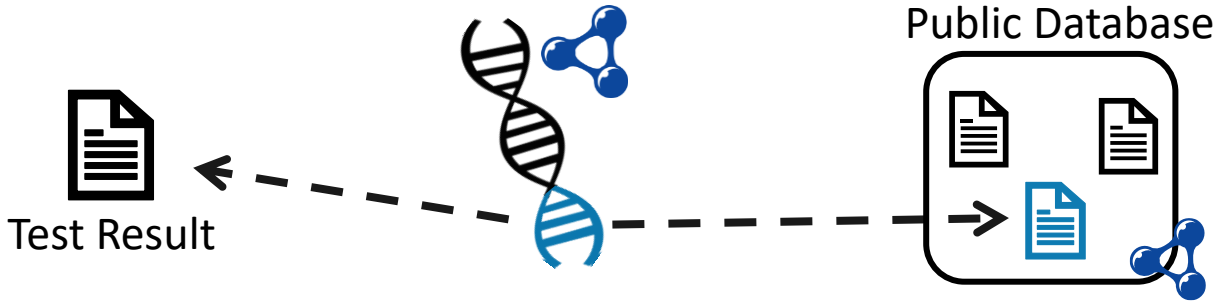


Genomic data to RDF?

Automation of Linkages?

Primary Research Challenges

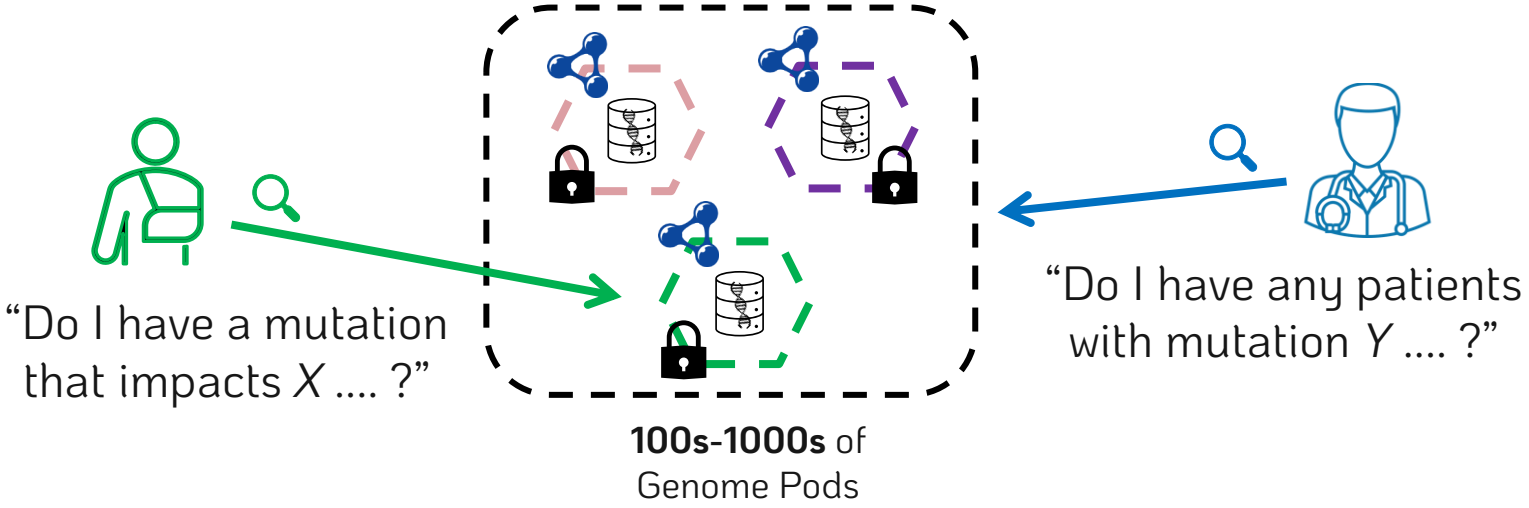
WP1. Linking Genomic data to clinically-relevant data



Genomic data to RDF?

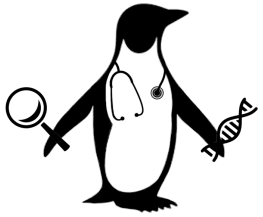
Automation of Linkages?

WP3. Efficiently finding data through querying



MANY, LARGE sources?

PENGQUIN



Genomic Data → Clinical Care

Semantic, Linked Data

Query Capabilities

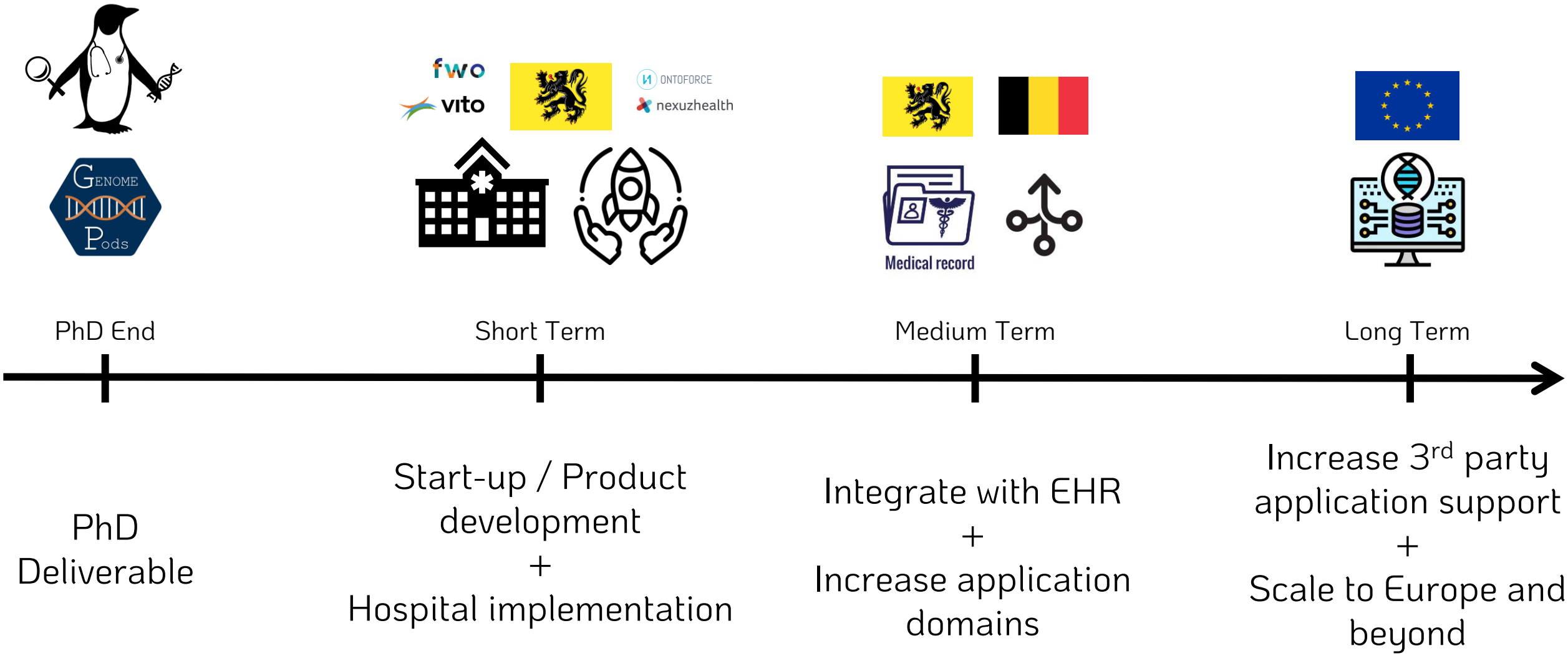
Framework + App

Promoters: Dr. Gökhan Ertaylan[1], Dr. Bart Buelens[1], Dr. Ruben Taelman[2], Prof. Dr. Ruben Verborgh[2]

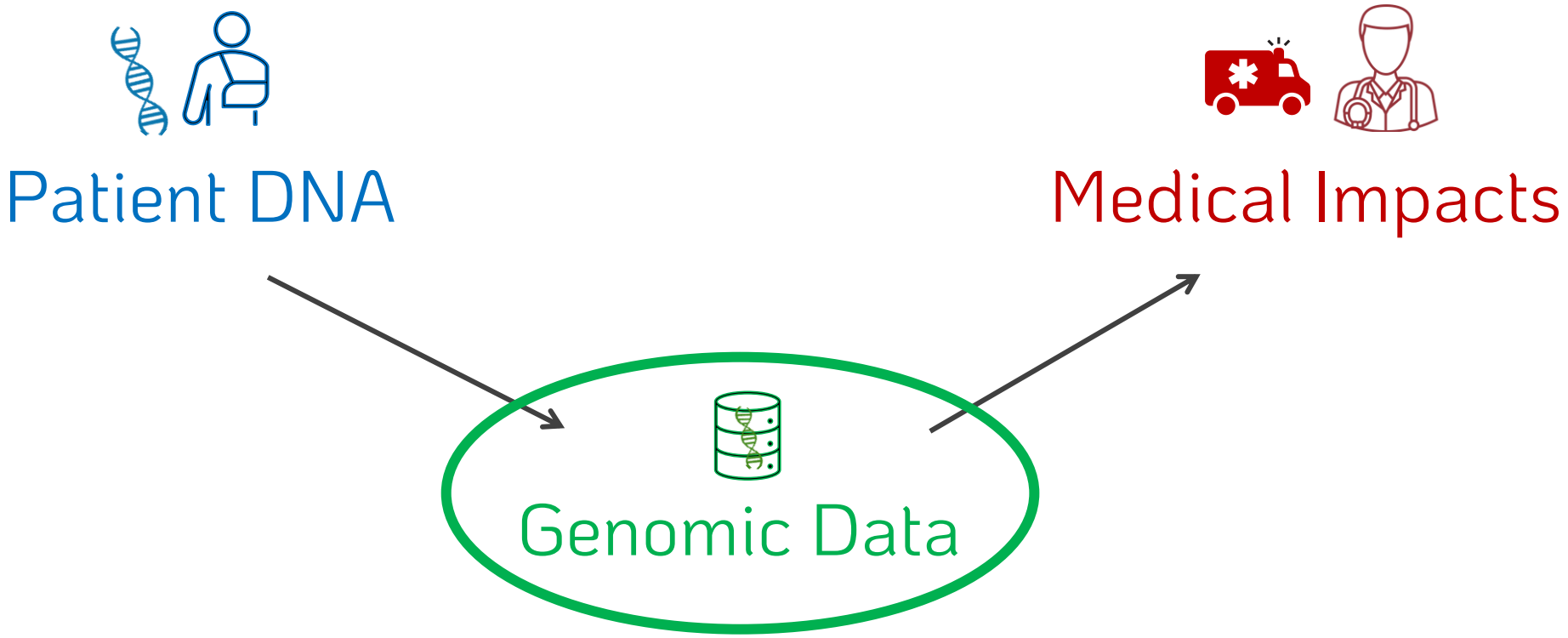
[1] = VITO NV
[2] = Ghent University

Supplemental Slides

Envisioned Product Evolution



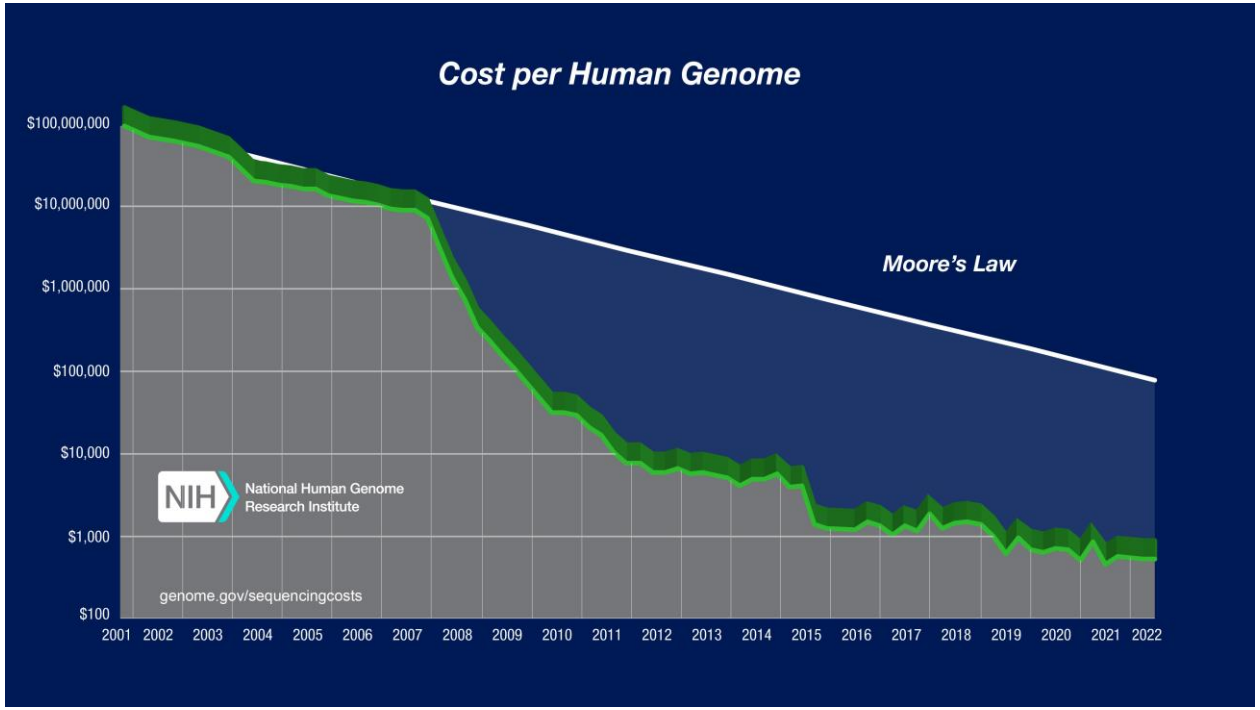
Why this Panel?



PErsoNal Genome QUery IN healthcare and clinical practice

Genome Sequencing in Health Care

Sequence Generation Costs



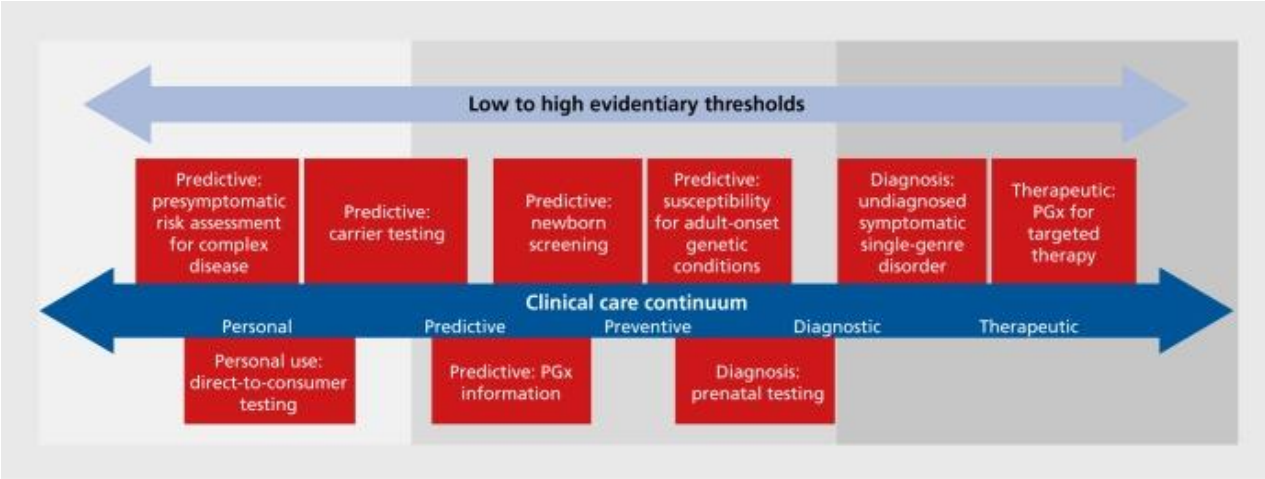
Wetterstrand KA. Accessed 12/04/24.

Currently getting close to $100 per genome[4]

Per-base cost of sequencing has been dropping by half every ~5 months

Per-base cost of sequence data storage is dropping by half every ~14 months

Use-cases



10.31887/DCNS.2016.18.3/jkrier

Examples include:

- Medication prescription (pharmacogenomics)
- Genetic disease screening
- Cancer diagnosis & treatment

Personal Genome Sequence Data

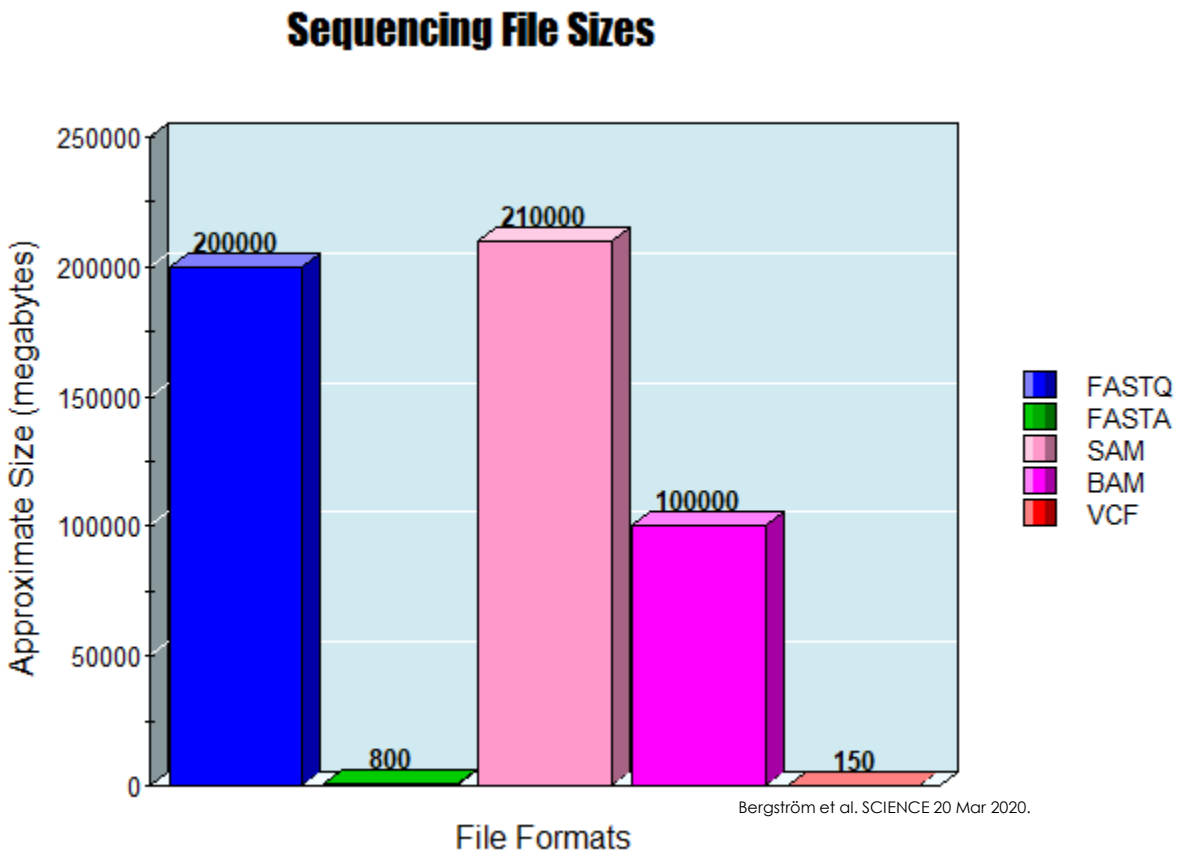
Human genome data is big…

~3.2 BILLION base pairs ----> ~1 GB

Sequencing output file ----> ~200 GB

Variant Call Format (***VCF***) file ----> ~200 MB

- Individual humans exhibit ~3 million mutations
- VCF file **records only those differences**
- Flat file



Bergström et al. SCIENCE 20 Mar 2020.

VCF file semantic representation

- **~ >27 MILLION** Triples when represented as RDF
- Ontology for conversion
- Header Dictionary Triples (for compression)

Sequence Data Formats

FASTQ ✚ **FASTA** → **SAM/BAM** → **VCF**

Raw Sequencing Outputs

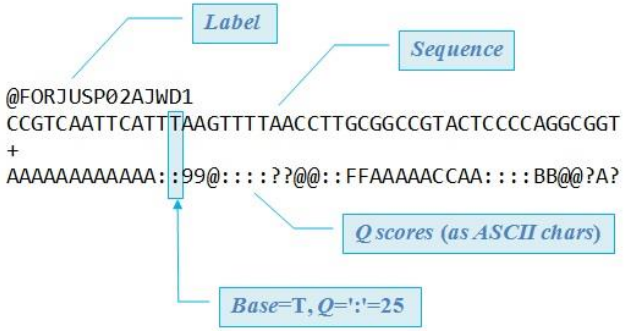
Contiguous DNA sequence

FASTQ reads aligned to a FASTA reference

Areas in SAM where sample and reference differ

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```




```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Sequencing Storage

The human genome = ~3.2 billion base pairs

- 3 billion letters -- **~700 megabytes**

Representative of a **single strand** of the whole human genome sequence
E.g. : …ATGCCGTAAACAGATGTCA …

Format: *TXT* file

- Raw human whole genome sequence -- **~200 gigabytes**

Whole genome sequences typically have **>=30x coverage at each base position**. (i.e. each base of the genome is represented in the sequence file a >= 30 times)

Format: *FASTQ* file

- Aligned human whole genome sequence -- **~210 gigabytes // ~100 gigabytes // ~15 gigabytes**

Whole genome sequence **is mapped to a reference human genome** and can be stored in various formats. Recent advances have allowed for improved compression.

Format*: SAM* file // *BAM* file // *CRAM* file

- Human genome variant file (list of mutations) -- **~125 megabytes**

Individual humans only exhibit **~0.1% variation** comparatively (~3 million mutations). Variant file **records those differences** in one individual's genome compared to the "normal" reference genome.

Format: *VCF* file

Variant Call Format (VCF)

- tab delimited .txt file

- Rows represent individual mutations in a sequence (when compared to reference)

- Columns store information about that mutation

```
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT
```

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   rs4477212   a   .   .   .   .   GT  0/0
chr1    752566  rs3094315   g   A   .   .   .   GT  1/1
chr1    752721  rs3131972   A   G   .   .   .   GT  1/1
chr1    798959  rs11240777  g   .   .   .   .   GT  0/0
chr1    800007  rs6681049   T   C   .   .   .   GT  1/1
chr1    838555  rs4970383   c   .   .   .   .   GT  0/0
chr1    846808  rs4475691   C   .   .   .   .   GT  0/0
chr1    854250  rs7537756   A   .   .   .   .   GT  0/0
chr1    861808  rs13302982  A   G   .   .   .   GT  1/1
chr1    873558  rs1110052   G   T   .   .   .   GT  1/1
chr1    882033  rs2272756   G   A   .   .   .   GT  0/1
chr1    888659  rs3748597   T   C   .   .   .   GT  1/1
chr1    891945  rs13303106  A   G   .   .   .   GT  0/1
```

Experimental VCF Data

Primary dataset from publicly available repositories (>100 genomes)

- Personal Genome Project (https://my.pgp-hms.org/public_genetic_data)
- NIST's Genome in a Bottle public dataset (https://data.nist.gov/od/id/mds2-2336)
- Illumina Platinum Genomes repository (https://emea.illumina.com/platinumgenomes.html)

If more are needed (>1000 [research grade] genomes):

- 1000 genomes project (https://www.internationalgenome.org/)
- NCBI's ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)

Solid Technology

- Does not solve ALL problems, but offers an environment where improvements are possible
  - ✓ Easier sharing (web-facilitated)
  - ✓ Granular privacy controls
  - ✓ Direct patient data access
  - ✓ Data representation flexibility (RDF)
  - ✓ Possibility of data-linking and querying

https://solidproject.org/TR/protocol

Genome Data in RDF

VCF genomic data can be represented as RDF.

VCF to RDF Vocabularies exist → SPHN Semantic Interoperability Framework[4]
+ FHIR HL7 Format[5]

VCF genomic data can be Linked to other data.

How to automate this?
What vocabularies could be used for the linkages?

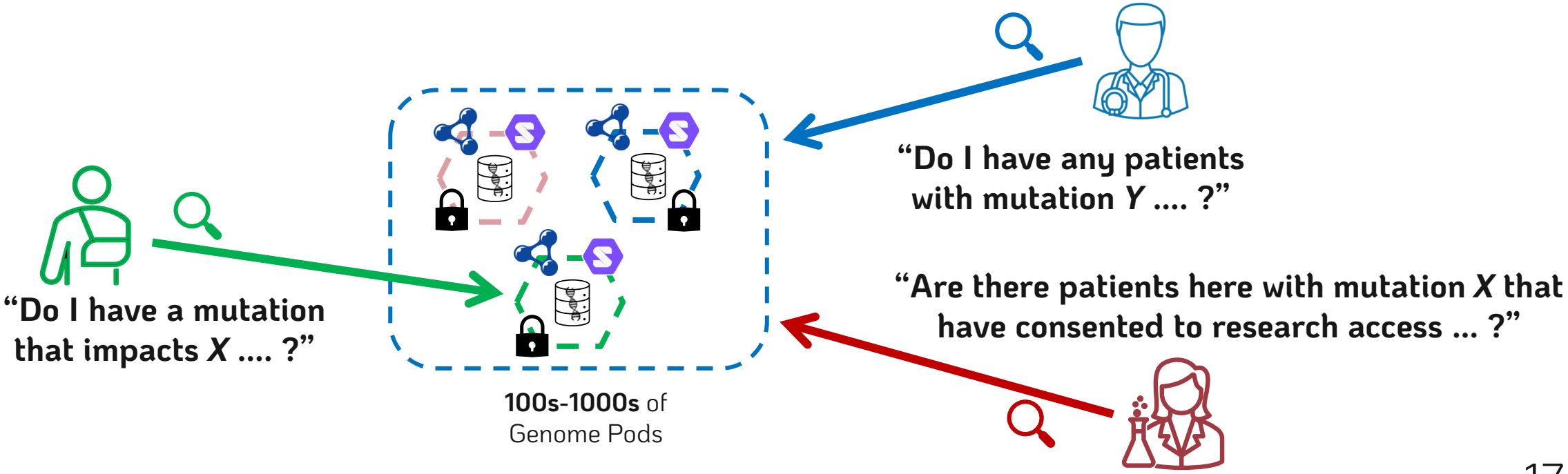
FHIR SPHN

[4] van der Horst E, et al. Bridging Clinical and Genomic Knowledge: An Extension of the SPHN RDF Schema for Seamless Integration and FAIRification of Omics Data. 2023.
[5] Dolin RH, et al. vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration. 2021.

Querying Genome Pods

RDF genomic data pods can be queried.

Query Engine → Custom Comunica implementation/engine[6]



[6] https://comunica.dev/

Querying Genome Pods (possibilities)

Single Pod Query

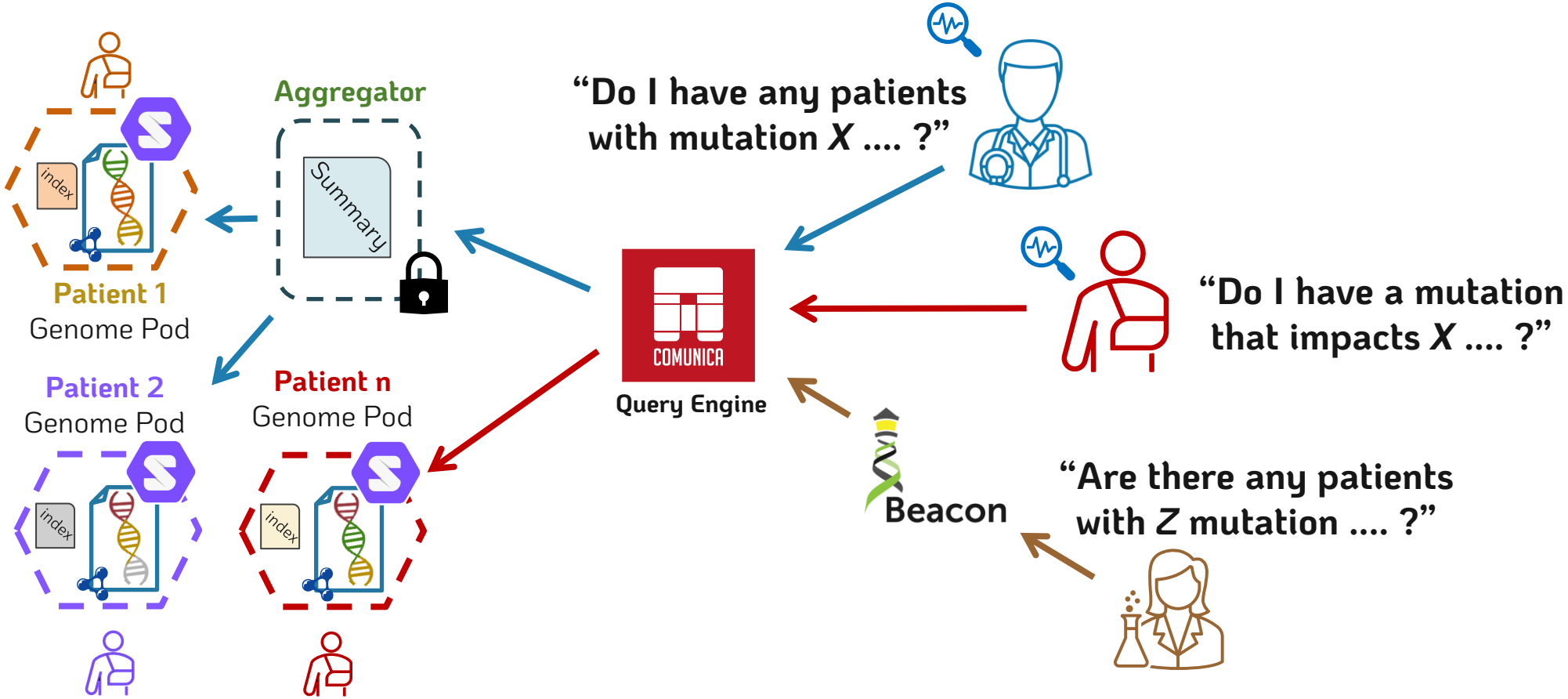
Utilizes an intra-pod index

Many Pod Query

First, utilizes an extra-pod aggregator
Then, utilizes the intra-pod indexes

Beacon API for Researcher Queries

Query is translated from Beacon API to Comunica
Then, the **Many Pod** query approach is followed



Implementation Specifics

Solid pod instances → Community Solid Server[1]

Genomic Data → Publicly available *VCF* data[2]

RDF Conversion → Aim to use SPHN RDF vocabulary[4]

Data-linking → Actively exploring ontologies / methodologies

Data Querying → Custom instance of Comunica[6]