Qualitas new SNP-archive: TheSNPpit – a preliminary report

SABRE-TP Workshop

F. Seefried, P. von Rohr

10.11.2021

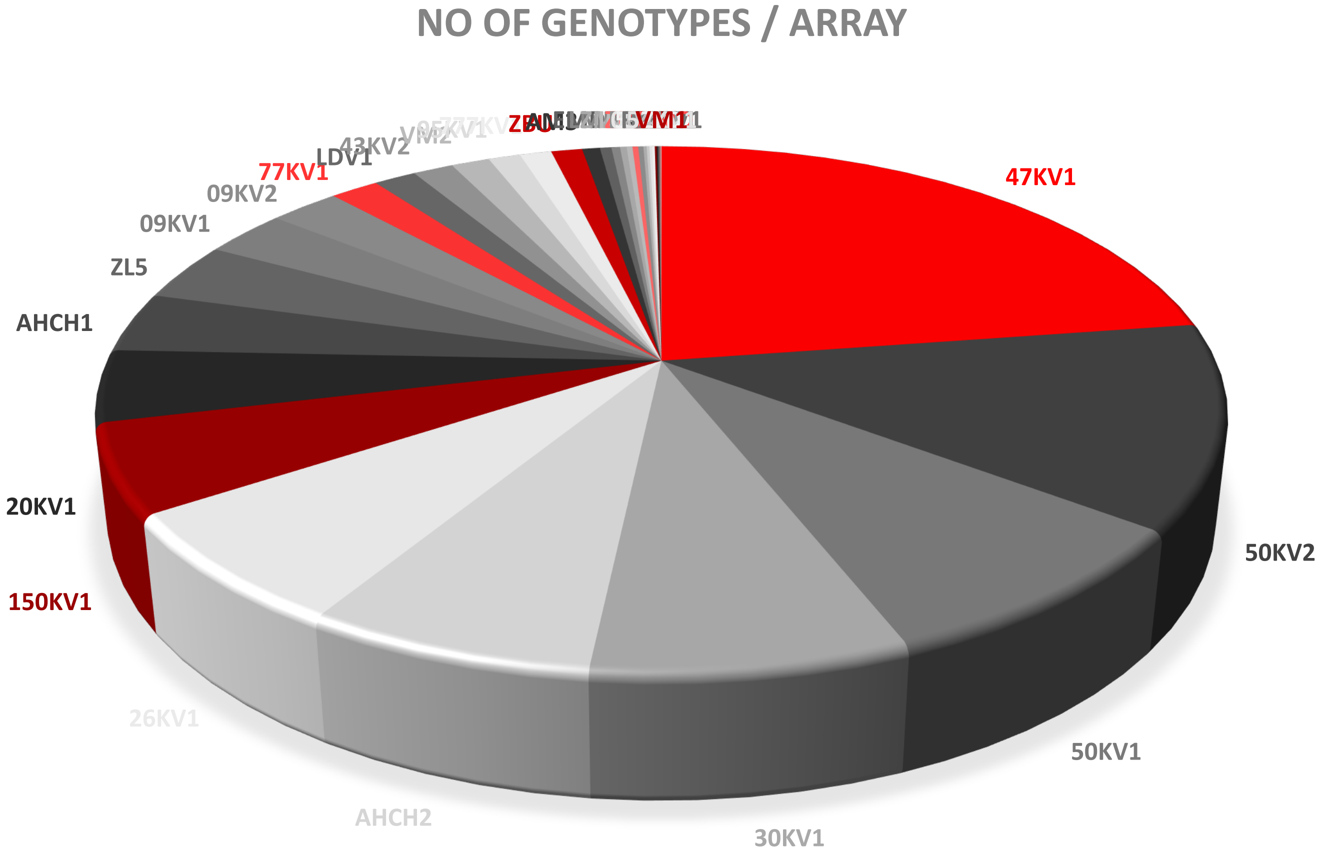
Outline

- General overview
- New archive – snppit
- Performance Test
- Pro’s / Contra’s
- Summary

No. of Arrays @Qualitas

- Cattle: 43

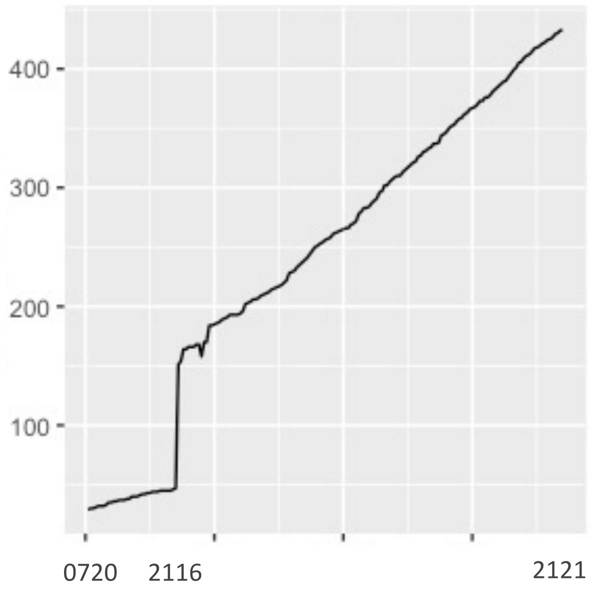
- Goats: 2



No. of Genotypes @Qualitas

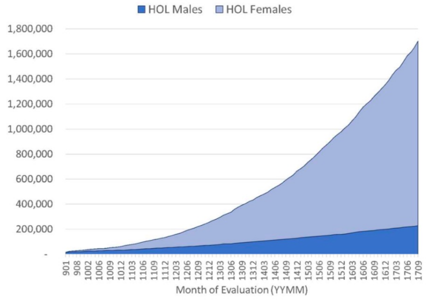
- Trend in No. Of Genotypes:
  - Linear
  - Oct. 21:
    - ~ 430'K genotyped samples (cattle)
    - ~ 3.4K genotyped samples (goats)

No. of Genotypes (1000)



- CDCB:
  - 1.8 M (Sep. 2009)
  - 5.5 M (Oct. 2021)

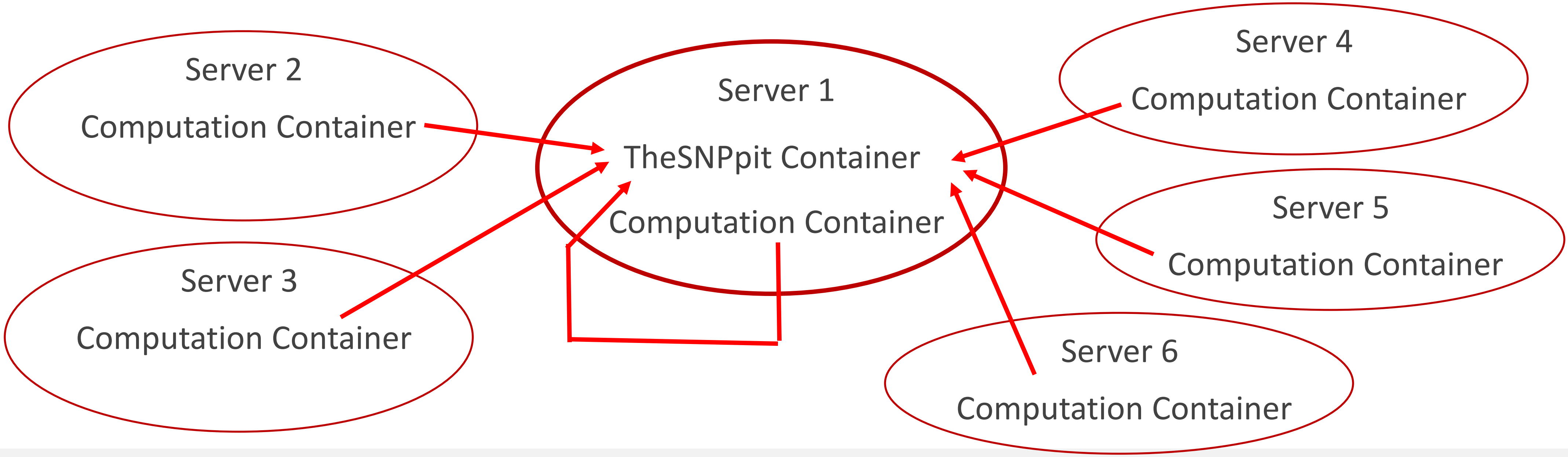
FIG 2 - Holstein genotypes added monthly to CDCB database since January 2009



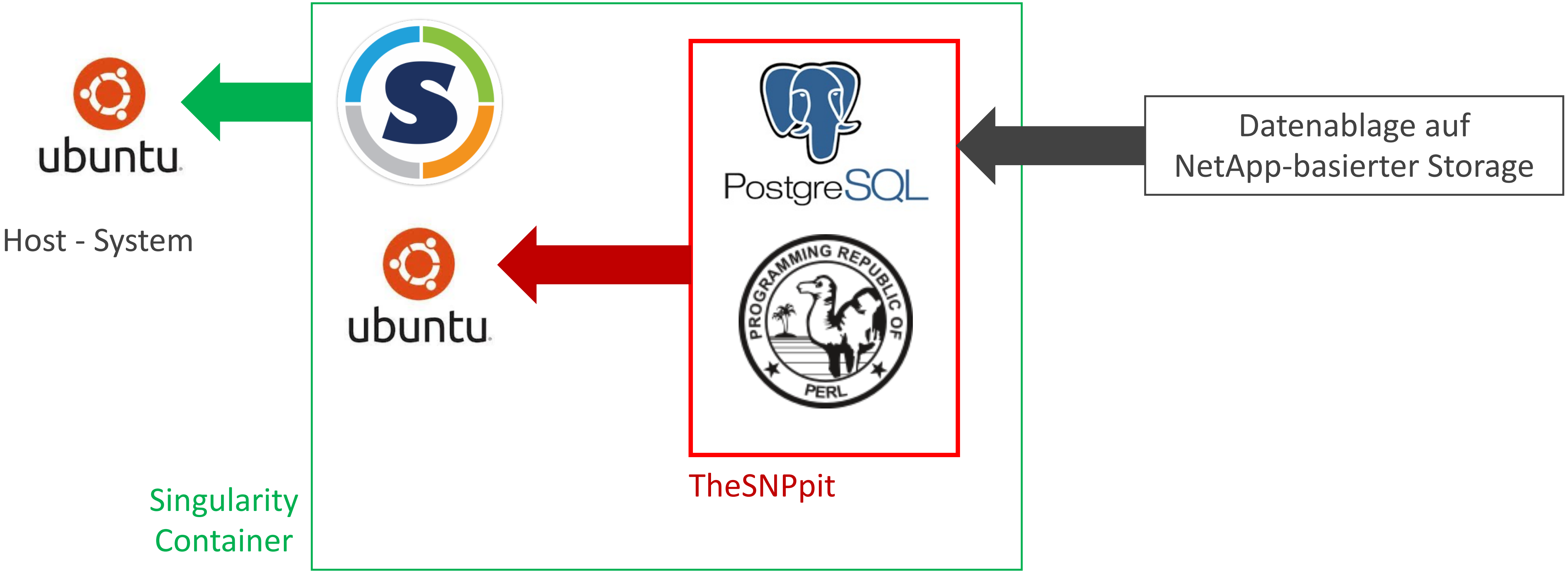
- Human genetics:
  - «Computational and functional gene prioritization from a saturated GWAS of adult height in 5 million people" 5 MILLION!» #ASHG

Environment

- 6 Linux Server
- Singularity Container Image System since 2018
- Installation of TheSNPpit container on one server
- Access of computation container on 6 server via ssh



TheSNPpit - Container



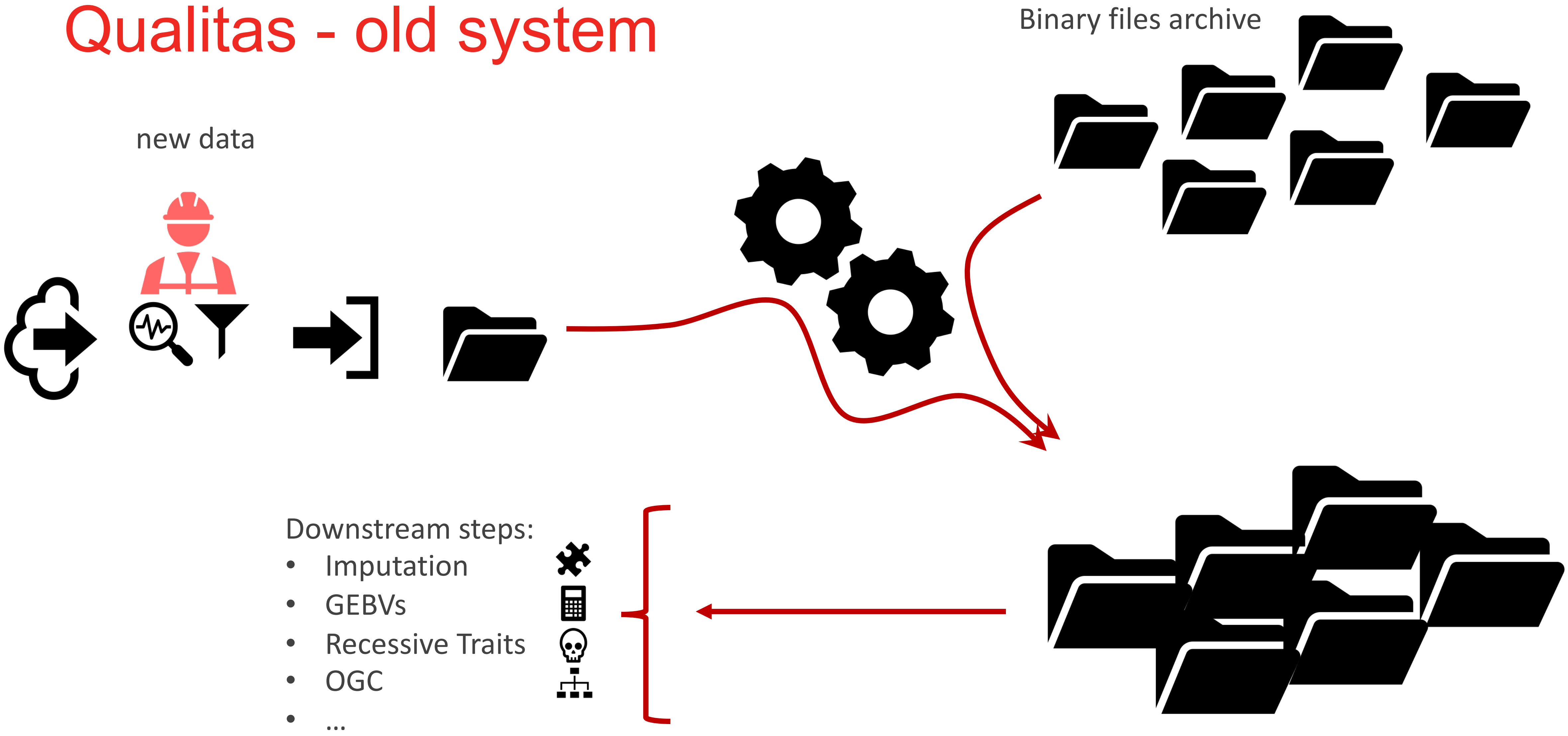
Host - System

Singularity Container

TheSNPpit

Datenablage auf NetApp-basierter Storage

Qualitas - old system



new data

Binary files archive

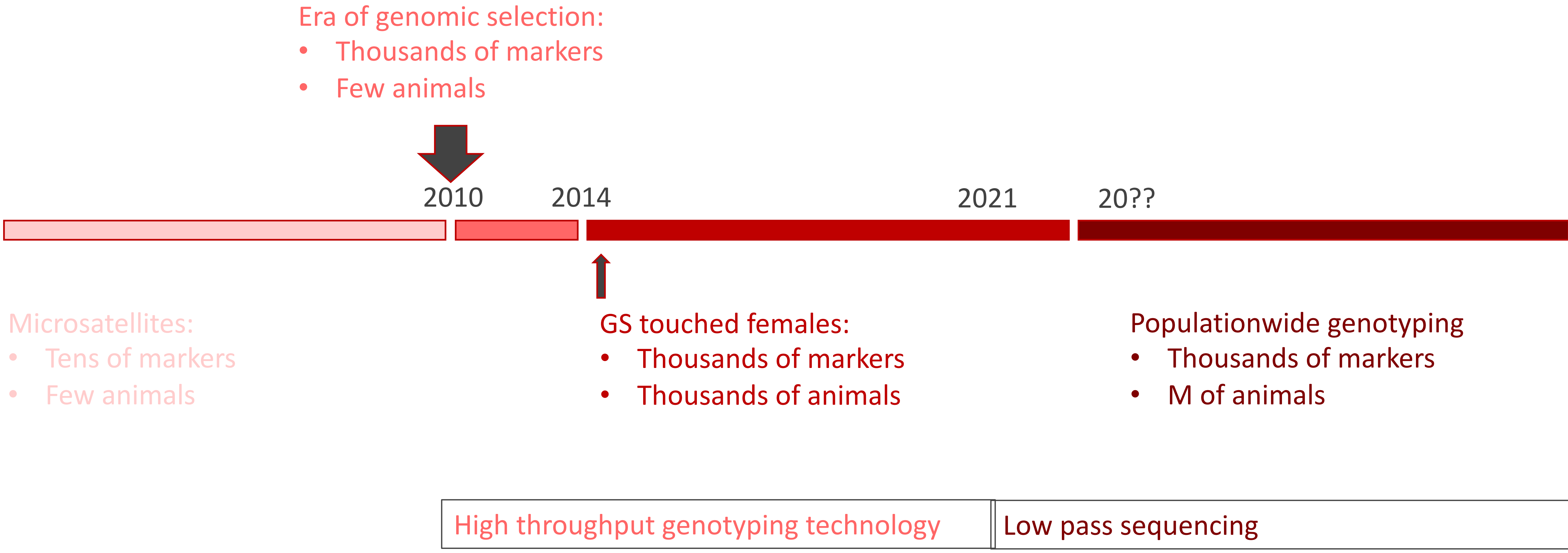
Downstream steps:

- Imputation
- GEBVs
- Recessive Traits
- OGC
- …

TheSNPpit - background

- Published in 2016 PLOS ONE
- Authors: E. Gröneveld & H. Lichtenberg
- Workshop 2018
  - Mariensee FLI
  - Retirement Gröneveld & Lichtenberg
  - Creating a user group
    - Open access
    - Intention: discussion, sharing experiences, development of further needs…
    - Sound: very silent
- Qualitas took over hosting TheSNPpit 2019
- Homepage

When we look back and forward:



Era of genomic selection:

- Thousands of markers
- Few animals

2010

2014

2021

20??

Microsatellites:

- Tens of markers
- Few animals

GS touched females:

- Thousands of markers
- Thousands of animals

Populationwide genotyping

- Thousands of markers
- M of animals

High throughput genotyping technology

Low pass sequencing

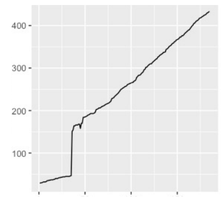
TheSNPpit - basics

- Database system for managing large scale SNP genotype data from
  - any genotyping platform
  - and numerous genotyping arrays

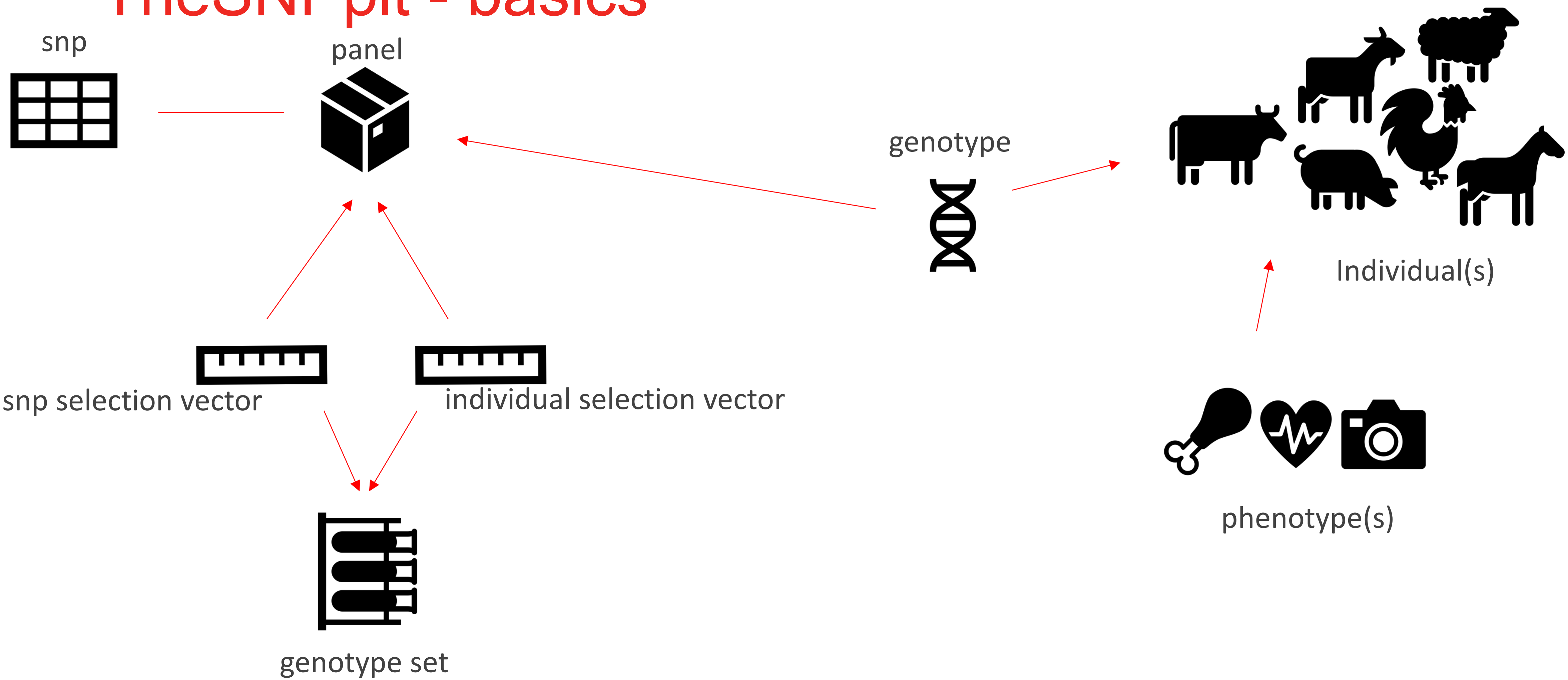
- Key ideas:
  - Highly compressed vector storage in a relational database
  - Focus on a fast export written in C, Perl and PostgreSQL as database backend
  - Few interfaces with other software, data formats
  - Focus on organisation and storage of data

TheSNPpit - basics

- Vector conversion
  - Each panel comprises a set of SNPs
  - 1 byte per SNP (0-1-2-5)
  - Using SNPs position in the panel as it's position in the genotype bit vector enables access to each SNP
  - So called: genotypes are treated as vectors
- (Different) genotype_sets defined by:
  - (different) SNP_selection_vectors: SNP_sel_vec
  - Individual selection vectors: Indiv_sel_vec



TheSNPpit - basics



snp

panel

genotype

Individual(s)

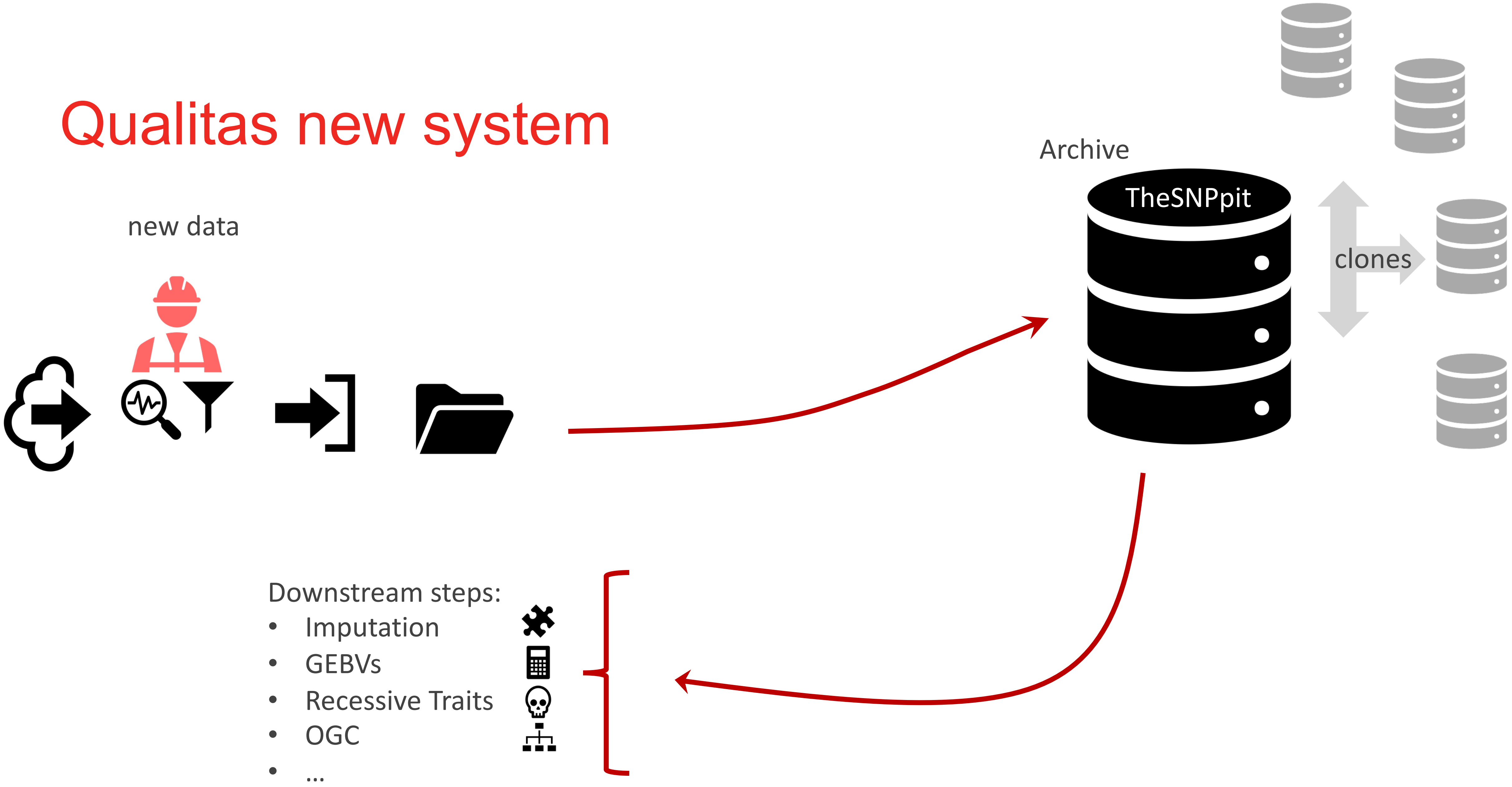
snp selection vector

individual selection vector

phenotype(s)

genotype set

Qualitas new system



Downstream steps:
- Imputation
- GEBVs
- Recessive Traits
- OGC
- …

Performance test - Numbers

- Repeated import:
  - 500x
  - 8.1K genotyped samples on SWISScow Array (310K marker)
  - 4.1M records (310K marker)
  - Disk storage database: 348 GB

- Export:
  - 4.1M records
  - Binary file: 308 GB
  - CPU time: 180’

Summary: Pro’s – Contra’s

- Efficient storage system
- Fast
- 2 bytes per genotype

- Open source

- Key role: panel
  - Storage of WGS data
  - GBS data