Genome UK

Considerations on data security, sharing and analysis

PetaGene

Genome UK — The Three Pillars

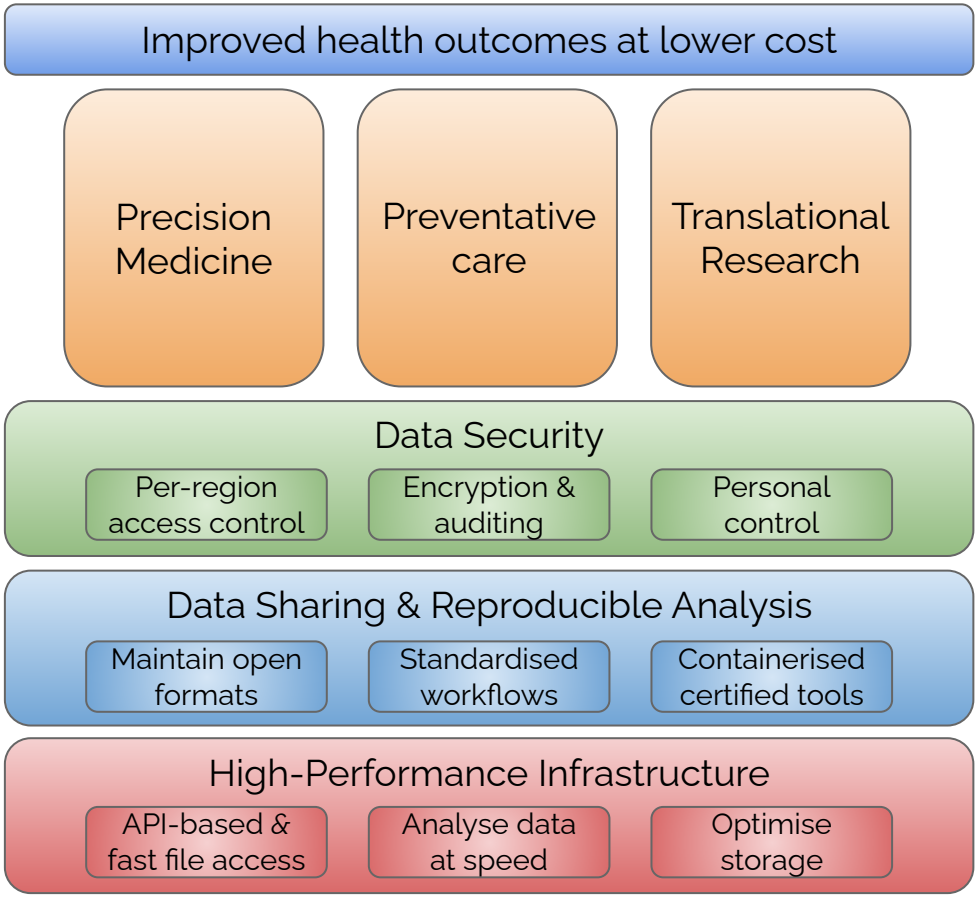
- Pillar 1: Diagnosis and Personalised Medicine
  - Improve the diagnosis, stratification and treatment of illness.
- Pillar 2: Prevention
  - Enabling predictive and preventative care.
- Pillar 3: Research
  - Supporting fundamental and translational research.

☞ **The success of Genome UK will rely on the public's willingness to share highly sensitive data for research and clinical applications.**

Genome sequences: the bulk of the data

- **Genome UK:** Goal of generating WGS data for 1.5 M individuals → 150 Pb data.
  - Hardware and software infrastructure for data security and data sharing security essential.
- **PetaGene:** Experience supporting customers across Life Sciences in managing genomic data.
  - We can help Genome UK ensure successful implementation.

Genome UK – our key recommendations



PetaGene

Key opportunities

Genomic Data is sensitive

- Public willingness to share data is declining (Ghafur, S. *et al* 2020 *Lancet*).
- Controlled, secure data access is fundamental to maintaining public trust.
- Sharing a person's whole genome, exome or chromosome is unnecessarily invasive since usually a very small portion (< 0.1%) is sufficient.
- Data minimisation: share only minimum necessary information for each purpose.
  - Italy, Slovakia, US HIPAA explicitly enforce in research. Collection explicit in UK law.
  - Unnecessary sharing significantly increases de-anonymisation risks.
  - Share only genetic regions relevant for analyses. Audit all access, and by purpose.
- Ideally, give each individual control of data & visibility into how their data is being used.

Key opportunities

Avoid technical barriers

- Life Sciences stakeholders need to combine existing data and pipelines with new data.
- Democratise cost of precision medicine.
  - Implementation of new standards must be *practical, improve performance & not introduce bottlenecks.*
- Public & private sector have invested billions in R&D building existing pipelines.
  - Therefore: crucial to keep supporting them.
  - Exciting new APIs; e.g. GA4GH *htsget* (fully supported by PetaGene).
  - **But** slow & expensive for fine-grained regional access compared to alternatives.

Key opportunities

Computational Reproducibility is essential

- Ability to reproduce results for clinical and research purposes.
  - This is the basis on which scientific advancements are made possible.
- Ability to continue to use existing containerised, certified workflows with Genome UK data.

Preserving data integrity is key

- Need to consider strategies to retain data integrity.
- Minimise loss of information that may become important for accurate analyses in future.

PetaGene

For any questions or inquiries please contact us at: info@petagene.com