REVIEW ARTICLE

Genomics Vault: A framework for precision medicine data management

[*]Bhupinder Bhullar[1], Guy Gross[2], Hakan Akozek[3]

[1] Managing Director, BasePort Inc., London, United Kingdom
[2] MBBS, CEO, Go2Health, London, United Kingdom
[3] Tip Doktoru, Director, Technologio LTD., London, United Kingdom

Abstract: A mixture of fumaric acid esters (FAEs) is approved for the oral therapy of psoriasis. However, for a long time the active ingredient of this mixture was unknown. We reviewed the *in vitro* data available for the different FAEs present in the multi compound drug and elaborate how they may contribute to possible clinical effects. Although helpful overall, many in vitro data must be viewed critically because the concentrations used in the experiments exceed the plasma levels reached in patients. The data suggest that dimethylfumarate (DMF) is the most active compound, mediating the major therapeutic effect after metabolization into monomethylfumarate (MMF) *via* an according receptor expressed on target cells. Identifying the active pharmaceutical ingredient within a mixture of compounds helps to subsequently eliminate unnecessary, potentially harmful compounds. This provides a promising example for an alternative precision medicine approach in clinical practice.

Keywords: dimethylfumarate (DMF); monoethylfumarate (MEF); monomethylfumarate (MMF); fumaric acid esters (FAEs); psoriasis

*Correspondence to: Bhupinder Bhullar, BasePort Inc., London, United Kingdom; b.bhullar@baseport.ch

Received: July 24, 2017; **Accepted**: March 20, 2018; **Published Online**: April 24, 2018

Citation: Bhullar B, Gross G, Akozek H, 2017, Genomics Vault: A framework for precision medicine data management. *Advances in Precision Medicine*, vol.2(2): 267. http://dx.doi.org/10.18063/APM.v2i2.267

Introduction

On March 9, 2017, the Canadian parliament overwhelmingly passed the Genetic Non-Discrimination Act to prevent the use of genetic data to deny individuals health insurance, employment, housing or influence their child custody or adoption decisions. Supporters of the law noted the reluctance of Canadians to take genetic tests during clinical care for fear of the data being used against them[1]. Canada is the latest country to enact such laws, but are not the only ones struggling to balance the risks and benefits of genomic based medicine. Figure 1 shows more than 20 countries which have population genome sequencing programs. As genome sequencing data becomes integrated into healthcare, a serious risk to individuals and their extended families could result from inadequately protecting such information. Although there are many technical challenges with implementation of national clinical genomics programs, governments are backing programs to adopt Precision Medicine due to the promise to save health institutions money and improve national economic output with healthier, productive, longer-living populations. With Precision Medicine comes a tsunami of personal data that will impact our understanding of disease, how we can stay healthy, and how we deliver healthcare.

The information sources that will be used to determine an individual's health and health risk will be drawn from two main data types:

a. Benchmarks of objective risk - Baseline genome sequence, and how the genes are expressed in the individual, *i.e.*, genotype and a cumulative list of

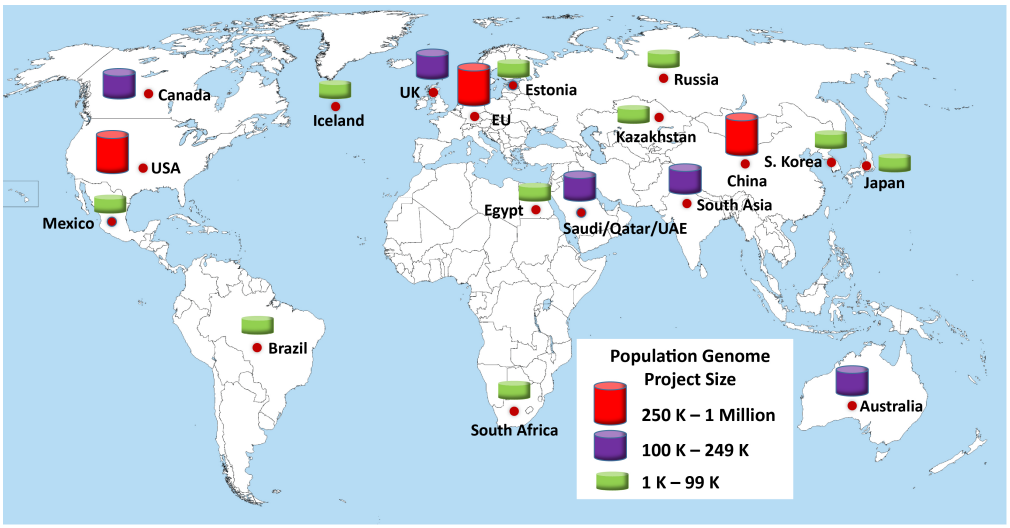

Figure 1. Regional Population Genome Sequence programs announced in over 20 countries. *Data source: www.phgfoundation.org*

irreversible conditions resulting from the genotype, and

b. Benchmarks of subjective risk - An omni-channel real-world behavioural and medical data set (*e.g.* wearables, EMR, imaging, AI, *etc.*) that allows us to quantify lifestyle, environment, healthcare factors that impact outcomes.

The data management strategies that will be adopted by the industry will determine the efficiency with which the healthcare sector will transition into implementing precision medicine. This report aims to assess the views of practitioners, *i.e.*, the data scientists, directors of clinical data, and IT administrators, on the future growth of the objective datasets, and understand the most pressing issues for healthcare institutions in terms of management of this data and its integration into the clinical framework. Subjective risk data is not included in this article due to the exponential proliferation of models and fragmentation of the marketplace (the reader is advised to view articles on the growing "wearables" economy[2]).

There is an unarticulated need to have a framework that, at its core, will protect consumers, be fully interoperable within existing healthcare infrastructure but will not limit the speed or direction of industry growth and innovation. However there are several challenges ahead for the healthcare industry:

a. Data size - The growth of genome data alone is predicted to surpass Youtube, Twitter, and Astronomy data in size by 2022[3],

b. Cost - There are financial implications to setting up and managing data centres that can handle this data volume,

c. Privacy - Genome data is the ultimate blueprint for individuals and the information exposes patients to as-yet unquantifiable risk (*e.g.* insurance, employment, healthcare, *etc.*)

d. Longevity - Patients are likely to live for 100+ years, maintaining records for a lifetime can be a technical challenge particularly in the context of variation in quality of data, long-term compatibility of software/hardware systems, and data storage models.

These challenges mean, similar to the experiences of other industries (*e.g.* insurance, financial services, marketing, media *etc.*), data management by healthcare institutions needs to evolve from a fringe activity into a core function that drives business strategy & activity, creates accountability, and is used to measure success and impact.

Assessing the requirements and capabilities of healthcare IT to address these challenges will help determine the timing and impact of Precision Medicine as routine care in the healthcare sector.

Genome Data Management

It has been conventionally accepted that sequencing base costs are exponentially decreasing compared to Moore's law. Moore's law, which projects the rate of doubling of transistors in an integrated circuit, roughly reflects the linear decreasing cost for data storage disks. The logarithmic decrease in sequence costs,

outstripping Moore's law, has been used as an argument that data storage costs will be the limiting factor for further growth[4]. In Figure 2, we mapped the cost for sequencing and storage per genome (assuming 30x coverage, 100 GigaByte file size). What is evident from this data is that the current chemistry cost to sequence a human genome outstrips the cost for disk storage by 1000X. The trend line for data storage cost decreases predictably. The sequencing chemistry cost, however, may plateau out, but is predicted to decrease to a "few hundred dollars" per genome in a few years.

Neither the sequencing cost or storage cost depicted in the graph, show the true cost to generate and manage a genome medical record. For sequencing, factoring in lab costs for sample preparation & operation costs, data analysis and reporting, the average cost per whole genome record is estimated to be 10,000 USD[5]. For data storage, disk purchase costs are low compared to the operation costs (power consumption, data operations, personnel, *etc*) to manage a data center. It is estimated that average annual cost to manage a 100 GB record, maybe anywhere near $200–$400/genome record, based on current annual cost to run a 1 PetaByte data center, which is a typical size for most healthcare centers making initial moves into genomics data storage[6]. While we assume the unit IT costs amortized over a large & growing data center will steadily decrease gradually over time, the operational cost for this storage is an on-going cost for the lifetime of the record. The overall cost continually increase as new records are continually aggregated.

Future estimates predict a reduction in both cost to sequence a genome sample and the size of the genomic data record maintained by clinical centers. From our interviews of leading genome centers, the next decade will see a dramatic increase in genome data. By 2030, it is estimated 500 Million human genomes will be sequenced world-wide and whole genome sequencing will be the routine procedure due to the cost reduction for sample preparation and sequencing (Dr. Torsten Schwede, Director, SciCore, University of Basel, CH, and Dr. Ewan Birney, Director GA4GH, European Bioinformatics Institute, UK, personal communications). The predicted growth in number of clinical samples will stress the economics and efficiencies of data centers to manage the volume of data, despite any reductions in cost to sequence or size of data files.

It is noted that commercial Cloud storage not only offers convenience for data management, but at a rate of $0.01/GigaByte/month, the static storage costs are low. However, cloud storage vendors have tiered costs connected with data transfer, CPU operations for analysis, and security levels, which bring the cost significantly higher. These associated costs are included in data operations for on-site data storage centres. Although some facilities may opt to use cloud based storage, data transfer on/off cloud based facilities constrains data integrity and security for 100 GigaByte records. Some vendors propose tape drives to enable long term storage data records, and these further decrease the costs. We hypothesize that for the next 10 years in the clinical genomics field, patient genome sequence data will be analysed in multi-omic aggregate studies to elucidate health status biomarkers.

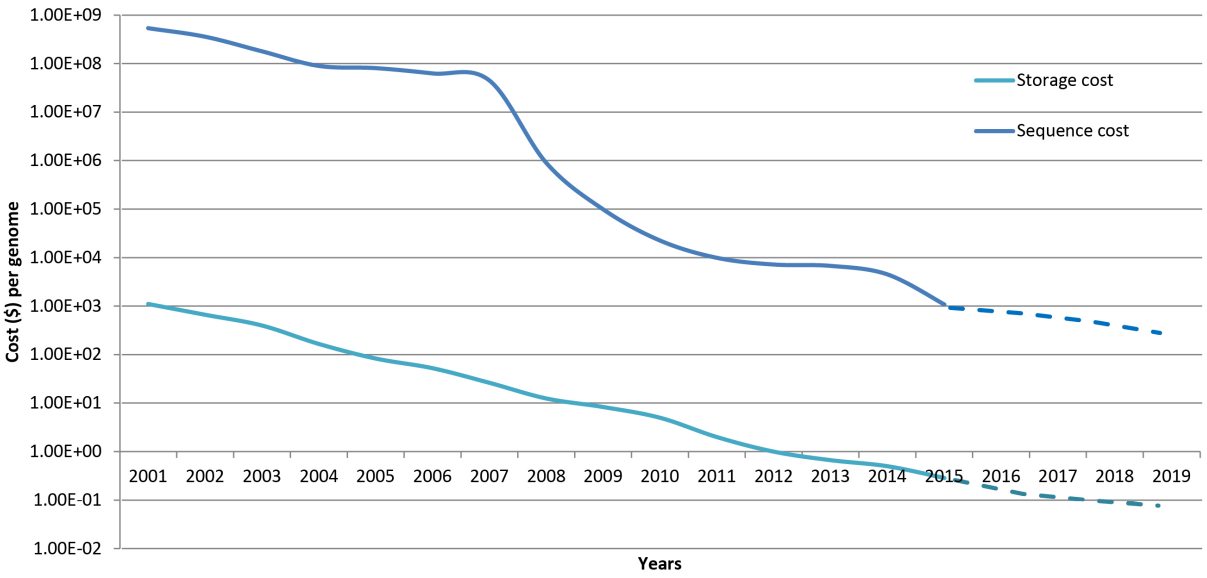

Figure 2. Approximated costs to sequence and store the data for a human genome (whole genome sequence, 30X coverage, 100 GigaBytes).

As such, tape drives will not be the primary media as it is passive storage, and hurdles with cloud storage will make the management more complex than on-premise administration of the data.

Another decision driver that is emerging in this field is the need for data privacy. One of the unique privacy challenges brought by the genome sequence data is the "known unknown". Unlike medical records which hold only historic information at any given time, human genome is a database of unknown number of markers that are yet to be discovered. Anyone who obtains a copy of the identifiable raw genome data can use it to unlock this information about the individual as and when new markers are discovered. As a result, it is practically impossible to understand the future impact of information security breaches relating to genome data and strong safeguards will need to be in place to protect it. It is one of the hypothesis of this paper that data privacy will be a driving factor in decision making for management of this data.

To ascertain the decision drivers in practice, we interviewed 14 people identified in organizations that are advancing precision medicine in clinical and research organizations. These individuals (CEOs, data scientists, clinical directors, IT managers *etc*.) have first-hand knowledge for the growth of this field, with direct experience executing the processes in the pipeline, or working with vendors. The individuals participated in a 30-minute interview, and subsequently were sent an online questionnaire to gauge relative importance of decision making criteria and approach to data management. The subject number is small but justifiably so because there are only a few centers that are currently executing whole genome sequencing programs–which is the testing ground for how genomics based medicine might be implemented in the future. These early adopters will ultimately set models that will be used by others as the innovations related to genomic medicine diffuse through the healthcare industry.

The results of the interviews indicate that all participants agree unanimously that whole genome sequencing will be a standard practice in healthcare delivery in the future. The interviewees, who worked directly on clinical genomics pipelines, had examples of individuals who benefited directly from the data they generated. The benefit to patients will be the impetus for widespread growth of genomic services in the future. The support from national health programs will continue because of the potential breakthroughs for predictive health biomarkers and the overall decrease in medical expenditures with more accurate diagnosis.

The current limitations experienced by all the centres interviewed, is insufficient biomarkers linking genotypic mutations to phenotype, or treatment options, limiting the number of patients that can benefit from this information. Thus, a critical hurdle in the delivery of the genomic testing in the clinic is interpreting the genetic information. The most time-consuming part of the pipeline is the manual curation of the patient's genome data. Because of a lack of definitive genotype-phenotype correlations, some cases require further testing to validate the genotype, and in other cases, it is difficult to identify any variant for clinical follow-up. The fix for these limitations is to continue to populate databases with enriched meta information about patients, and to facilitate means to share these findings for clinical decision making. Indeed, as databases grow, facilitating researchers access to these genomes for population studies will elucidate new biomarkers.

The second insight from interviews was that the size of the genome data files is a critical concern for all interviewees who managed or interacted with IT issues. The need to store a 100 Gigabyte file has meant that most of the institutions have setup data centres reaching PetaBytes in size. Because the research community is actively using this data, some of the growth was in secondary analysis and storage of genome information related to research methodology. The data sizes are already getting difficult to manage, and one center went as far as to suggest that they are considering deleting the full genome data, once the relevant data is extracted, and saving only the bio-sample (*e.g*. patient's blood, or tissue). The intention is that, because sequencing is getting cheaper, the bio-sample can be sequenced again when the data is required. The limitation here is that if the sample is damaged by loss of freezing, the loss is permanent, and irreversible There was an interesting split between the European and North American interviewees with regards to cloud based or on-premises data storage. The strong privacy rules in Europe meant that data cannot be held on public cloud infrastructure. However, the North American participants (n = 3), were approving of, or already testing public cloud storage for genome data.

To follow-up, the interviewees participated in a short on-line survey to quantify criteria for how they would setup operations. All the participants choose data security as the more important factor when choosing to setup a clinical genomics facility. The next 3 factors, in order of importance, were ease of operations, low overhead costs, and purchase price. This is interesting because, during the interviews, considerations for data privacy weighed heavily on the operations of the facility, and the on-line survey verified this observation. If data security is a driving decision making criteria, we are approaching a critical point where the associated data management costs are the controlling factor over whether genetic mapping/genomic testing can become

scalable. The need for data privacy, concomitant with the societal need to study these genomes to find lifesaving markers for human diseases, poses challenges that are largely unique to healthcare. This cannot take place without significantly changing the approach to data management. Here we consider several factors that can enable building a framework for managing the privacy of personal genomes, decrease associated storage costs by streamlining data management operations, and, at the same time, facilitate the industry moving forward.

Working towards a common national (or international) data management framework would go a long way to facilitate the practice of precision medicine. Considering the conflicting needs for the different stakeholders, a new data architecture model is proposed below for how all of the stakeholders, in particular, patients, physicians and scientists, can interact with genomic data. (N.B. The framework could be employed for any type of healthcare data.)

Liberating Genome Information for Precision Medicine

When genome sequencing and precision medicine becomes a routine practice, a number of new challenges are likely to emerge, including ensuring the safety of the genome data whilst making the clinically valuable information available and locating and accessing patients' marker information where sequencing is done by a multitude of providers.

Figure 3 below outlines a potential high-level architecture to help address some of these challenges and enable practicing precision medicine as part of routine healthcare. The proposed architecture includes four key components implemented at two different levels:

Sequencing Provider Level

Genome Vaults: These are highly secure (both physically and digitally) facilities controlled by sequencing providers. Made up of data processing and storage facilities as well as interlinked sampling equipment, these Genome Vaults sample, sequence, store, and analyse citizens' genomes for all known markers. Once all known markers are identified, a citizen specific report is published to the Report Library hosted by the sequencing provider. The report essentially contains a list with presence or absence of all known markers in the citizen's genome.

When a new marker is discovered, all genomes hosted in the vault are re-processed to search for this marker and all reports in the report library are updated.

In an ideal scenario, each citizen would only have a single copy of their raw genome data hosted by a trusted provider.

Consent Engine: The Consent Engine, hosted by the citizen's trusted sequencing provider, holds all consent given by the citizen past and present. Granular consent levels enable the citizen to provide generic consent, such as "disclose all marker information to all valid clinical providers" or specific consent given in response to a request such as "disclose ovarian cancer risk markers to Dr X just this once".

Consent Engines can also be used to record citizens' consent to participate in genome research.

Report Libraries: Report Libraries contain a list of all known markers for each citizen whose genome is deposited in the sequencing provider's Genome Vault. On request from a clinician and subject to citizen's consent, the presence or absence of the marker is confirmed to the clinician. Report Libraries are also accessible to citizens to review their own reports.

Each time a new report is published, the Report Library also sends a consent request to the citizen. This acts a notification to the patient to review any new information as well as a reminder to update any existing generic consent.

To avoid accidental disclosure, the Report Engine will need to reject any requests for marker information where the citizen's consent to provide the information is older than the date the marker was identified.

National Level

Request Engine: A national component will also be required to hold a registry of all citizens who have genome sequences, together with the details of the trusted sequencing provider that holds this information. The Request Engine will handle all marker information requests from authorised clinicians practicing precision medicine and will keep an audit trail of all requests, responses and the citizen's consent at the time of the response.

The Request Engine also facilitates citizen's access to their marker reports either for review or for using this information to receive care outside national boundaries.

Facilitating Genome Research

This new approach can also be further developed to significantly increase the speed and accuracy of genome research with some modifications as outlined in Figure 4, particularly where there is a national shared clinical record either in full or in summary form. To preserve the patient's privacy and trust, a safe pseudonymisation function would need to be added to the Request Engine which would also facilitate the identification of the subject cohort and coordination of the responses from various sequencing providers. Sequencing providers would also need to host specific functionality to manage requests and responses which we called the Research Engine. Use of pseudonymisation would require

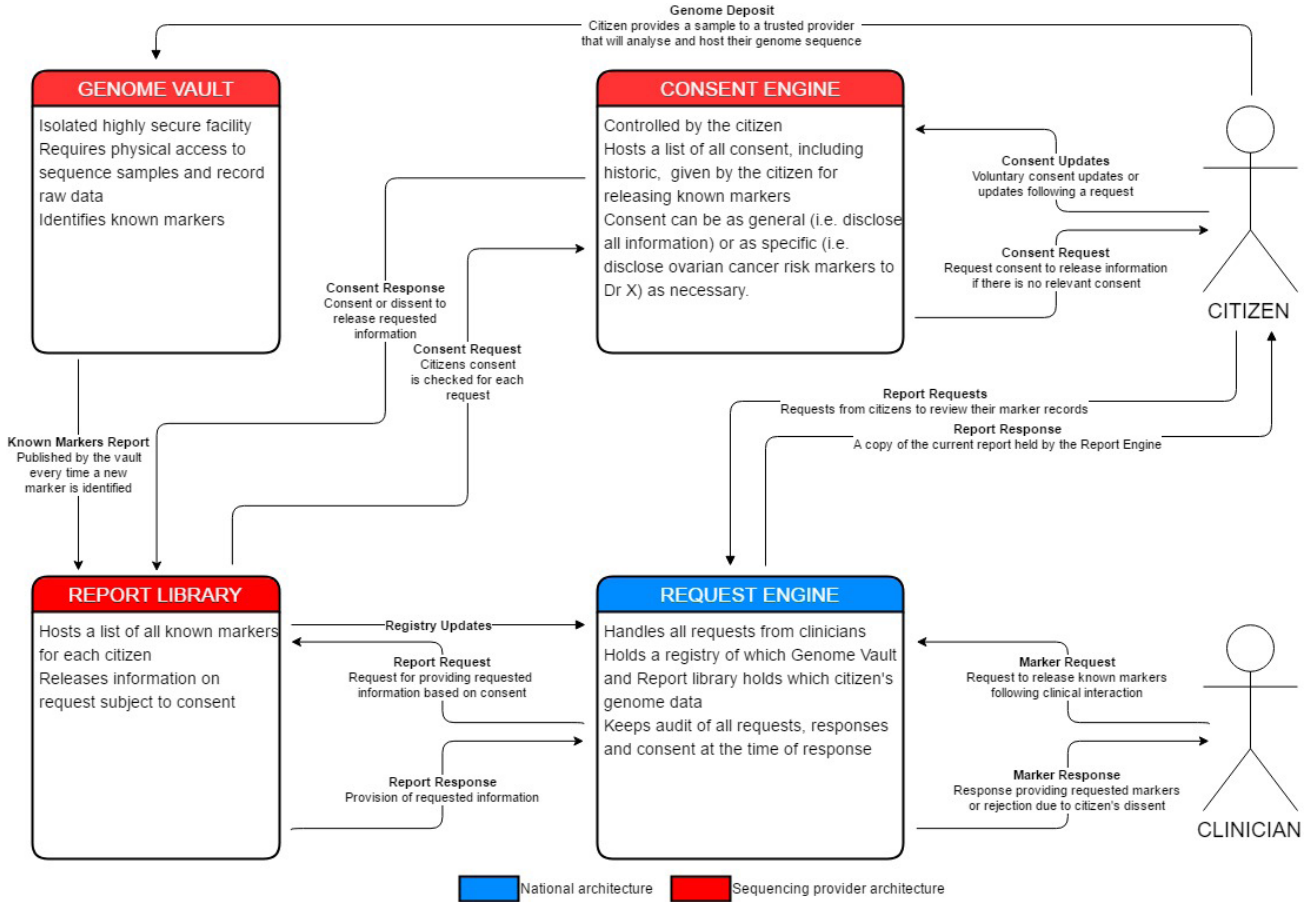

Figure 3. Patient and Clinician information flow with a "Genome Vault" architecture.

strong safeguards to be in place for approving research boundaries and associated queries to protect against malicious or unintentional identification of citizens.

Facilitating Genome Research

This new approach can also be further developed to significantly increase the speed and accuracy of genome research with some modifications as outlined in Figure 4, particularly where there is a national shared clinical record either in full or in summary form. To preserve the patient's privacy and trust, a safe pseudonymisation function would need to be added to the Request Engine which would also facilitate the identification of the subject cohort and coordination of the responses from various sequencing providers. Sequencing providers would also need to host specific functionality to manage requests and responses which we called the Research Engine. Use of pseudonymisation would require strong safeguards to be in place for approving research boundaries and associated queries to protect against malicious or unintentional identification of citizens.

In this scenario, after obtaining necessary approvals, the scientist provides a query with the acceptance criteria for the research cohort (*i.e.*, males who have had their first heart attack before the age of 30) and a request for identifying a specific genetic marker.

The Request Engine then sends the national Shared Clinical Record a list of citizens who are known to have sequencing providers and obtains a list of subjects who fit the criteria. When the Shared Clinical Record responds with a list of citizens who fit the criteria, the Request Engine assigns each citizen a pseudoidentifier unique to the specific research. The Request Engine then sends a genetic marker request to each sequencing provider with the list of citizens against which the marker should be checked.

After receiving the research request, each provider's Research Engine checks the citizen's consent for research and compiles a work schedule for each consented citizen. To preserve the Genome Vault's "outbound traffic only" principle, the provider's scientists work within the Genome Vault and use the schedule to check each genome against the requested marker. They then publish a report to the Research Engine with the presence or the

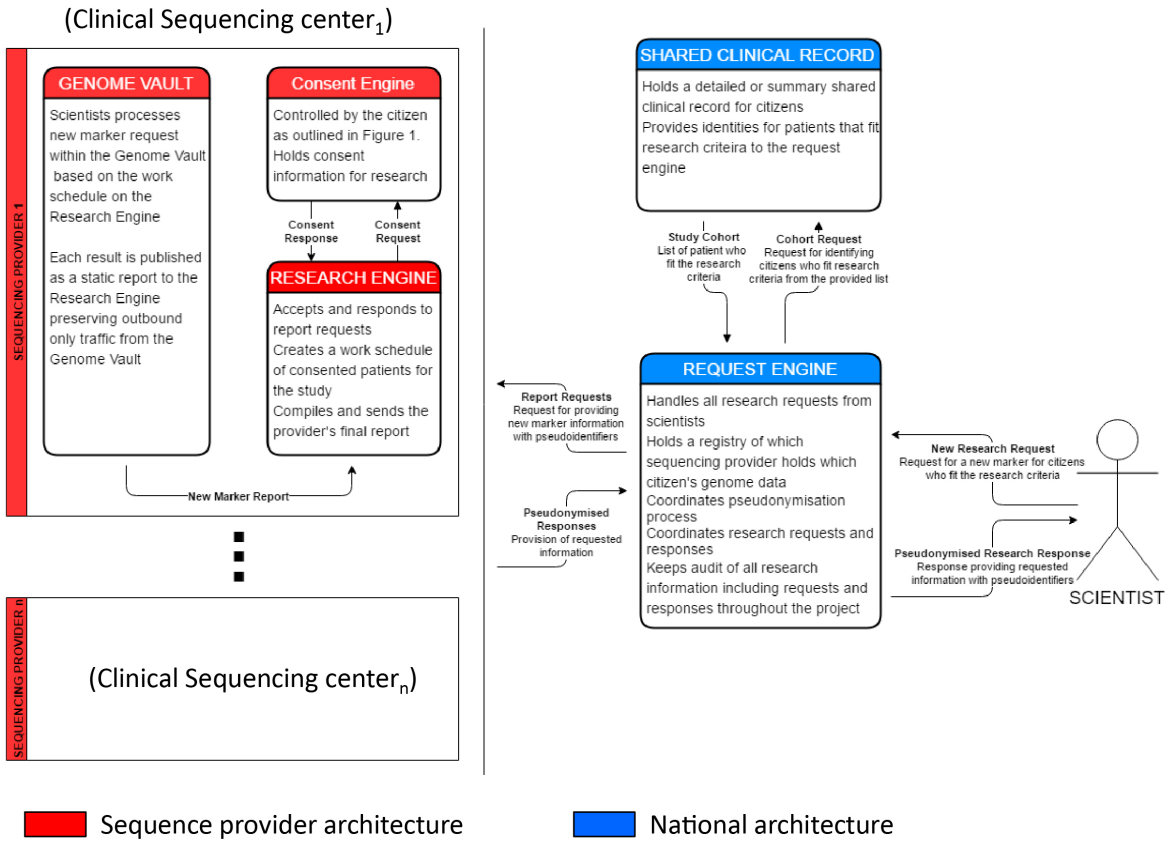

Figure 4. Research information flows to study aggregated patient genome data from Genome Vaults.

absence of the marker. When all the genomes within the work schedule are received from the Genome Vault, the Research Engine sends a final report to the Request Engine with the requested genetic marker information. This report only uses the pseudoidentifiers and does not include the citizen's identity

The Request Engine, compiles the responses from all sequencing providers and sends the final report to the scientist using pseudoidentifiers. The scientist can use these pseudoidentifiers for the duration of their research to ask specific questions about the individual citizens within the cohort without knowing their identity. Once the research project is closed, the Request Engine would delete all pseudonymisation information irretrievably.

There are other components that we have not included in our diagrams such as identity management and access control systems that will be required for the safe and secure operation of this architecture. Where there is a national initiative for sharing clinical information across the healthcare system, these are likely to be in place already in some shape or form and there are obvious benefits to sharing the overlapping components between the two architectures as well as enabling information exchange between these using an agreed standard, such as HL7. Where this is the case, if there is enough public trust in national structures, it would also be beneficial to implement the Consent Engine at a national level to include all consent to share medical records and to participate in research.

Moving Forward

The basic framework presented here does not cover all the details for the functioning elements of this platform. For example, it would be imperative that sequencing centers setup their operations to directly deposit the clinical sequence data into the "Genomics Vault". The utility of the vault functions to protect the data–which, from our opinion, is the "common currency" at the core of the discussion. This will facilitate the clinical genomics field to move forward by empowering more participants to opt-in to genetic testing, and enable more research discoveries from the rich diversity of the data. Failure to adequately protect this information could lead to societal and regulatory repercussions, hindering large-scale genomic research projects. Genomic technologies are making the most significant advance

towards precision medicine, and as other "omics" (*e.g.* proteomics, epigenetics) fields progress, the security and privacy standards that are developed will help integrate data from these new platforms.

References

1. Wayne Kondro, 2017, Canada's new genetic privacy law is causing huge headaches for Justin Trudeau. *Science*. Avaliable from: http://www.sciencemag.org/news/2017/03/canada-s-new-genetic-privacy-law-causing-huge-headaches-justin-trudeau. http://dx.doi.org/10.1126/science.aal0901
2. Alexandros Pantelopoulos A, Bourbakis N G, 2010, A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems Man and Cybernetics*, vol.40(1): 1–12. http://dx.doi.org/10.1109/TSMCC.2009.2032660
3. Stephens Z D, Lee S Y, Faghri F, et al., 2015, Big data: Astronomical or genomical? *PLoS Biology*, vol.13(7): e1002195. http://dx.doi.org/10.1371/journal.pbio.1002195
4. Stein L, 2010, The case for cloud computing in genome informatics. *Genome Biology*, vol.11: 207. http://dx.doi.org/10.1186/gb-2010-11-5-207
5. Rehm H L, 2017, Evolving health care through personal genomics. *Nature Reviews Genetics*, vol.18: 259–267. http://dx.doi.org/10.1038/nrg.2016.162
6. Buffington J, DeMattia A, 2016, *Whitepaper: Enterprise Strategy Group*, http://www.esg-global.com