Report of DTL focus meeting on

Next Generation Sequencing Data Storage and Sharing

(February 3rd, 2014)

Johan den Dunnen, Leon Mei (LUMC)

1 Meeting program

The goal of the workshop is to discuss the best practice on storing, sharing and publishing NGS data, identify the common questions and issues, and come up with a list of action points that can be tackled by the sequencing community within DTL. The following speakers presented latest technical and organizational efforts, followed by active questions and discussions among the attendees.

- Ies Nijman (UMC Utrecht): "Joint NGS variant database in Center For Personalized Cancer Treatment" (http://www.cpct.nl/)
- Ivo Fokkema/Martijn Vermaat (LUMC): "LOVD and Shared Diagnostic Variant Database"
- Morris Swertz (UMCG): "Dutch medical research and clinical data infrastructure coordinated by NFU and VKGL"
- Justin Paschall (EBI): "Sharing NGS data, variants in European Genome Archive (EGA)"
- Hendrik-Jan Megens (WUR): "Sharing ag-genomics data"
- Anthony J Brookes (Leicester University): "Data Discovery and Knowledge Sharing: parts of the Data Sharing Continuum"

2 Summary

In the last several years, we have seen a fast adoption of NGS technology in all eight Dutch University Medical Centers and many University research groups. Vast amount of NGS data (raw and interpreted) are generated from both diagnostic work and research projects. Although many people in the field agree that sharing these data will greatly benefit research and diagnostics and directly help patients, there are still technical challenges and organizational difficulties to overcome first.

Besides the discussion notes in Section 3, below are a list of overlapping interests touched by multiple speakers or questions.

- Funding opportunity to support developing NGS variant sharing software. Hard to get dedicated funding now. DTL hotel call is a possible model or ask healthcare insurance. In any case, this should be included in the request/work plan we are writing to NFU. Support for data curation is largely non-existing.
- How to contribute to the next step plan of NFU/VKGL? Join the planned hackathon of Morris/Kees van Bochove. Plan some sort of follow-up steps based on the similarity in the systems we have presented today.
- It is hard to rely on regulation/policy to force sharing. We should try to support access granularity in our systems now (refer to the 5 level model proposed by Morris in Appendix A). Of course we should keep on pushing the full publicly accessible solutions.
- Role and the importance of EBI? EBI has been in the center of many software and infrastructure

project. We could try to look for more collaboration and support from EBI on adopting and sustaining some of our software projects.

- In the ideal world, we should be able to get full data sharing support from the real data owner (patient consent) and that should be sufficient to share all data. However, in the real world, it is still a challenge.

3 Discussion notes

3.1 LOVD and Shared Diagnostic Variant Database by Ivo Fokkema/Martijn Vermaat

LOVD and Varda were originally built as separate systems. LOVD is a database to store disease/gene specific variants and phenotypes. Varda is to support variant frequency calculation. Both systems are now part of a clinical diagnostic pipeline at LUMC (under development).

- Will the variants discovered by the clinical genetics pipeline become available for public? Yes if the data owner agrees. Aggregate frequencies should not be a problem. Technically it is already supported.
- Can the pipeline be adapted for other center? Yes, we focus at LUMC at the moment and we are open to share the experience with other centers.
- What's the difference with commercial solution (e.g. Cartagenia)? We have full IP and control so it is very easy to add new features. As we experienced, commercial solutions are costing more money and time.
- Can the system be used for other species? Yes, if good reference and annotation is available.
- Ideally financial support for trustworthy gene/variant databases should be covered by the diagnostic budget (e.g., via healthcare insurance).
- DVD stores data on the regions covered to calculate accurate variant frequencies. Using a flat coverage profile to filter variants is sometimes not ideal. E.g., regions with very high coverage but lots of junks and regions with low coverage but high quality alignment and variant calling. What's your thought on this? Indeed an issue, gVCF might help here.
- Can phenotype info be stored in Varda/DVD? Not directly in Varda, but the system can be connected to a phenotype DB. Furthermore, this might cause people to hesitate to share.
- Quite a number of system exist to provide variant frequency information. How to compare their results will be a challenge.

3.2 Joint NGS variant database in Center For Personalized Cancer Treatment by Ies Nijman

A number of institutes joined CPCT (UMCU, VUMC, NKI, ErasmusMC, etc). All biopsy are collected locally and sent to Utrecht for sequencing and analysis. CPCT supports both clinical track (gene panel approach, ~2 week turn around time) and research track (exome, WGS). Discovered variants are stored in a closed DB which is only accessible to participating groups. There are two types of supported queries: population centric and patient centric. Besides dealing with data sharing requirement, data integration (with clinical data, pathology, etc) is also a goal of the system. Pharma companies are potential users of this system, so granular authentication is required.

- The system is using external annotation source, e.g, cBioPortal, Cosmic, but will its result being shared? No, this DB is only accessible to partners.
- There are many impressive features in this system, is there a potential to jointly develop or share some technical features with other teams? Yes, we plan to make the system open source.

Also we have technical discussion with LOVD/Ivo.

3.3 Dutch medical research and clinical data infrastructure coordinated by NFU and VKGL by Morris Swertz

Morris presented the on-going discussion and next step plans at NFU and VKGL. There are many themes in the discussion, hence prioritization will be a challenge. The NGS community, incl. this group, should contribute and influence it now to make sure our expertise and concerns are known to the decision body at NFU. For example, members in the “schrijfgroep” can be contacted for suggestion or information. Furthermore, Morris also shared the latest development at UMCG on SOP (quality data management flow), Molgenis-data (unified modeling of NGS data), eDAS (enhanced DAS protocol, from the Brookes lab).

- Is there an equivalent body to VKGL in Pathology? Someone in the audience mentioned that all UMC's pathology labs agreed to buy the same type of sequencer (IonTorrent). So some sort of collaboration is in place. Jeroen Belien might have contact for this.
- Step 3 of the VKGL work plan is to evaluate existing software and architectures used. It will be around June 2014. A hackathon is planned between Morris' team and the Hyve. There is a possibility for other teams/developers to contribute or suggest new joint demonstrator.

3.4 Sharing NGS data, variants in European Genome Archive (EGA) by Justin Paschall

EGA is a kind of equivalent to dbGAP of NCBI with extensive possibilities to regulate data access. During submission people define which people under which conditions can get access to the data. . Data Access Committee (DAC) is required to form to govern the data access. This archiving service is for free. In near future, there will be also a Cloud connected to it to support data analysis for smaller user groups and distributed archiving service will be provided by CSC, CRG, FIMM as well.

- Can Justin write a summary about which services are already ready? Linking to effort of BiomedBridges?
- Can pure closed project/dataset be hosted at EGA? If the free service of EBI is used, then the dataset should have a perspective to become public eventually. Or at least certain level of aggregated information should become public.

3.5 Data Discovery and Knowledge Sharing: parts of the Data Sharing Continuum by Anthony J Brookes

Tony has been part of many data sharing discussions (e.g., Gen2Phen). He learned that sharing data often involves too many political discussions. As additional necessary strategies he suggests sharing both knowledge and sharing the existence of data and directing users to the source. To this end he developed (a) 'OmicsConnect' - an self-standing genome browser for local and networked visual exploration of knowledge in multi-omics datasets, using 'eDAS' to feed data from files and databases enhanced by embedding user permissions, and (b) a flexible platform (Cafe Variome, CV) that allows discovery searches driven from dta rather than metadata, also incorporating phenotype ontologies, NGS data capabilities, and follow-on facilitated data sharing regulated by each data source. It can offer more control to the data owner and support granularity of access. Very important to note is that CV is not a database solution, it is just providing a level on top of existing databases to support searching.

- How to upload data to CV? The minimal interrogated dataset needs only 3 columns, but any number and type of extra field can be added.
- Does CV support data versioning? The question is misguided. The data, whether versioned or

not remain in the sources own databasae. CV then facilitates searching across some depth of that content, either remotely via an API or by mirroring the needed minimal content in the CV tool.

3.6 Sharing AG-genomics data by Hendrik-Jan Megens

Hendrik-Jan started his presentation with an interesting remark of Ewan Birney: “cheap genomic data might have a bigger impact on agriculture than human health in the long term”. There are a number large animal genome sequencing consortiums that are interested in building a data storage/sharing platform. E.g, in pig genome project, they have been comparing both open source and commercial solution. Like in the analysis pipeline, the animal/plant field is pretty much copying the best practice of human sequencing projects, e.g., 1000Genomes, GATK, gVCF, imputation. HJ likes the idea proposed by Tony on sharing the existence of data instead of data itself. The main reason is that although privacy (of pigs ;-) is not a great concern, animal breeding companies often have a big say in those genome consortiums and are also reluctant to share any data.

- If there a public DB to access pig data? Yes, from pig genome consortium.
- If there any funding opportunity from the breeder companies to support data sharing software? No, that's not their primary interest.
- Is there an easy way to get from a variant in the pig genome to the comparable position in the human genome to learn about potential consequence? No, not really. It is possible but takes expertise and work.

Appendix A: Five data access levels in biobanking (contributed by Morris Swertz)

Level 1: summary of Cohorts / Sample collections / Studies

General data set descriptions such as ‘size’, ‘data topics’, ‘diagnoses’, ‘material types’.

Level 2: attributes of data / protocols / templates

information on data items available in each cohort such as questionnaire questions, lab measurements, diagnoses or GWAS

Level 3: aggregate cohort/collection level information

Non-identifiable aggregate information over the collection such as ‘no samples for high blood pressure cases’, ‘variant frequencies’

Level 4: individual data records, anonymized

complete set of (pseudonomized) phenotypes/genotypes for each sample/individual (‘the data’)

Level 5: linkable data identifiers per Individual

Identifiers that enable record linkage with data from other sources, such as the national cancer registry, via pseudonym mapping.

Appendix B: Statement in Dutch law on sharing patient data (contributed by Jeroen Belien)
English summary: if you are in the line of treatment of a patient you are allowed to share the medical data but only with the people involved in the line of treatment.

WGBO artikel 7:457 lid 1 en 2

Artikel 457
1. Onverminderd het in artikel 448 lid 3, tweede volzin, bepaalde draagt de hulpverlener zorg, dat aan anderen dan de patiënt geen inlichtingen over de patiënt dan wel inzage in of afschrift van de bescheiden, bedoeld in artikel 454, worden verstrekt dan met toestemming van de patiënt. Indien verstrekking plaatsvindt, geschiedt deze slechts voor zover daardoor de persoonlijke levenssfeer van een ander niet wordt geschaad. De verstrekking kan geschieden zonder inachtneming van de beperkingen, bedoeld in de voorgaande volzinnen, indien het bij of krachtens de wet bepaalde daartoe verplicht.
2. Onder anderen dan de patiënt zijn niet begrepen degenen die rechtstreeks betrokken zijn bij de uitvoering van de behandelingsovereenkomst en degene die optreedt als vervanger van de hulpverlener, voor zover de verstrekking noodzakelijk is voor de door hen in dat kader te verrichten werkzaamheden.
3. Daaronder zijn evenmin begrepen degenen wier toestemming ter zake van de uitvoering van de behandelingsovereenkomst op grond van de artikelen 450 en 465 is vereist. Indien de hulpverlener door inlichtingen over de patiënt dan wel inzage in of afschrift van de bescheiden te verstrekken niet geacht kan worden de zorg van een goed hulpverlener in acht te nemen, laat hij zulks achterwege.