CLINICAL GENOMICS DATA SHARING

Workshop summary report

Oslo Universitetssykehus
SciLifeLab
Karolinska Universitetssjukhuset
University Hospital Copenhagen (Rigshospitalet)
DNV GL

Report No.: 0, Rev. 0
Document No.: NA
Date: 2016-10-19

Project name:	Clinical Genomics Data Sharing
Report title:	Workshop summary report
Participants:	Oslo Universitetssykehus
	SciLifeLab
	Karolinska Universitetssjukhuset
	University Hospital Copenhagen (Rigshospitalet)
	DNV GL
Date of issue:	2016-10-19
Revision No.:	0, Rev.0
	NA

The objective of this summary report is to document the workshop taking place between the above participants in Oslo 30.-31. May 2016. The scope of the workshop was to:

- review current clinical variant pipelines in the three laboratories; discuss common challenges and identify areas where standardisation/harmonisation could be beneficial.
- Identify what specific data would be valuable for laboratories to be able to share in the short, medium and long term as well as current technical, legal and ethical barriers that hinder sharing this data today.
- Discuss potential models for future cooperation and agree on next steps.

Prepared by:

Guro Meldre Pedersen
Principal Researcher
DNV GL
Guro.Meldre.Pedersen@dnvgl.com

Sharmini Alagaratnam
Principal Researcer
DNV GL
Sharmini.Alagaratnam@dnvgl.com

Rev. No.	Date	Reason for Issue	Prepared by	Verified by	Approved by
0	2016-10-19	First issue			

Table of contents

List of tables

List of figures

GLOSSARY / ACRONYMS

Acronym	Meaning
ACMG	American College of Medical Genetics (guidelines)
BAM	Binary (sequence) Alignment/Map (file format)
BCF	Binary (variant) call format (file format)
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CMM	Center for molecular medicine (Centrum för Molekylär Medicin)
CMMS	Center for inherited metabolic diseases (Centrum för medfödda metabola sjukdomar)
CUH	Copenhagen University Hospital
DIPS	Supplier of electronic medical journal
DMG	Department of Medical Genetics
DNA	Deoxyribonucleaic acid
EMQN	European Molecular Genetics Quality Network
EMR	Electronic medical record
EPJ	Electronic Patient Journal
GA4GH	Global Alliance for Genomics & Health
GATK	The Genome Analysis Tool Kit (Broad Institute)
GIAB	Genome in a bottle
HGMD	The Human Gene Mutation Database
HGNC	Huge Gene Nomenclature Committee
HPO	Human Phenotype Ontology
HTS	High Throughput Sequencing
HW	Hardware
ICT	Information and Communication Technology
ID	Identification
LIMS	Laboratory Information Management System
LIS	Laboratory Information System
MD	Doctor of Medicine
MIP	SciLifeLab's own developed pipeline consisting of in house and open tools
MLPA	Multiplex ligation-dependent probe amplification
NGS	Next generation sequencing
NIPT	Non invasive prenatal testing
NIST	National Institute of Standards and Technology
NSC	Norwegian Sequencing Centre
OMIM	Online Mendelian Inheritance in Man – database
OUS	Oslo University Hospital
PCR	Polymerase chain reaction
PFAM	Protein Families (database)
QA	Quality assurance
QC	Quality control
qPCR	Quantitative (real time) PCR
SNP	Single nucleotide polymorphism
SNPEFF	Genetic variant annotation and effect prediction toolbox
SO	The Sequence Ontology
SW	Software
TSD	Service for sensitive data (Tjenester for Sensitive Data), USIT, University of Oslo
VCF	Variant call format (file format)
VEP	Variant Effect Predictor
VUS	variants of uncertain significance
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 EXECUTIVE SUMMARY

Moving from research and into the clinic, whole genome and exome sequencing technologies offer new opportunities for identification of biological basis for disease and treatment. While the challenge has shifted from generation of data to interpretation of data, the full utilization of the technology will require access to aggregated information on genetic variants as source of scientific evidence to support the clinical validity of variant interpretation processes. The clinical implementation is in its development, and the analytical pipeline from sample taking to clinical use of conclusion is not standardized.

The Clinical Genomic Data Sharing workshop was initiated to learn about similarities and differences in clinical variant interpretation pipelines between the three involved institutions, in order to evaluate the potential for sharing of data between the laboratories based on a common understanding of the current situation, with the ultimate goal of sharing interpreted genetic variants. The objectives of the workshop were:

- To review current clinical variant pipelines in the three laboratories; discuss common challenges and identify areas where standardisation/harmonisation could be beneficial.
- Identify what specific data would be valuable for laboratories to be able to share in the short, medium and long term as well as current technical, legal and ethical barriers that hinder sharing this data today and
- Discuss potential models for future cooperation and agree on next steps

The workshop included representatives from the Department of Medical Genetics, Oslo University Hospital, Oslo, Norway, Department of Clinical Genetics, University Hospital Copenhagen (Rigshospitalet), Denmark, Clinical Genomics facility, SciLifeLab, Karolinska Institutet, Stockholm, Sweden, Karolinska University Hospital, CMMS, Stockholm, Sweden and DNV GL.

During the workshop the three pipelines were mapped according to agreed principal process steps, and steps including quality control and/or reference guidelines and standards were identified. The participants then discussed and prioritized types of data they would like to have access to from each other, and functional requirements for these data.

The workshop participants would like to have access to information related both to the genetic variants and to the variant interpretation pipeline, including sharing experiences on gene panels, operating procedures, databases and tool development. It was recognized that further work is needed on understanding prerequisites for exchange of data, such as standardization of process and harmonization of accept criteria. A technical benchmarking of the variant interpretation pipelines was agreed as an efficient exercise for understanding impact of set-ups on final outcomes.

Discussion on access to data related to variants ranged from population variant frequencies, databases of curated classified variants and linked information on patients' genotype and phentotype to full access to FastQ files and patient phenotype descriptions. While the participants ultimately would like to have full access to databases, this may be more complex to achieve technically and with respect to societal accept. Sharing of population variant frequencies was agreed as a possible first step, while exploring the legal basis and barriers for sharing data and establishing a common database.

To continue the work, appropriate actions and responsibilities for follow-up were agreed as summarized in section 11. The workshop was agreed to be a first step towards sharing of genetic information and information related to the genetic variant interpretation pipelines between the participating laboratories, and a follow-up workshop was scheduled for November 2016 to summarize achievements and agree on further work.

2 PARTICIPANTS

The workshop included participants from the following institutions, with the list of participants provided in section 14.1:

- Department of Medical Genetics, Oslo University Hospital, Oslo, Norway.
- Department of Clinical Genetics, University Hospital Copenhagen (Rigshospitalet), Denmark.
- Clinical Genomics facility, SciLifeLab, Karolinska Institutet, Stockholm, Sweden.
- Center for inherited metabolic disorders (CMMS), Karolinska University Hospital, Stockholm, Sweden
- DNV GL

2.1 Oslo University Hospital (OUS)

OUS has 18000 employees and holds the largest medical genetics department in Norway, covering 3 million people, 60% of the Norwegian population, as part of the South East Norway health region (Helse Sør-Øst).

OUS operates the Norwegian Sequencing Centre (NSC) is a national core facility for high throughput sequencing (HTS) providing services to researchers. NSC is currently organized as a project with 15 employees, and will be a separate unit within end of 2016.

HTS is used in diagnostics. Exome sequencing with preselected gene panels is used for inherited disorders, currently using ~12 panels. Activities also include TRIO sequencing and targeted sequencing for cancer and cardio, but these processes are not used as basis for the mapping in this workshop. A separate group does tumor sequencing for research; this is not done in diagnostics yet. Computation takes place in a data cluster in TSD.

2.2 University Hospital Copenhagen (Rigshospitalet)

Lab activities of the Department of Clinical Genetics at Rigshospitalet include five laboratories; metabolic lab, cytogenetic lab, haematology/oncology lab and a molecular genetic lab at campus and a medical genetic lab at the Kennedy Institute. The NSG technology is merging the five existing labs, and there is a need to reorganize the department centred around the technology.

The capacity for exome sequencing is not sufficient, and the hospital is considering outsourcing. Funding for equipment and increased throughput must be secured. Sequencing is currently done using the IonTorrent system (Thermofisher) at campus and MiSeq (Illumina) at the Kennedy Institute. The hospital coor facility is used for research and clinical testing. The technology technology is applied for exome sequencing (Ion Proton), panels, and NIPT, as criteria for testing includes functional evidence.

Bioinformatic analysis is basically done using equipment standard setting, and there is no in-house development of tools. Evaluations are done based on adaption of ACMG[1] guidelines. Development of internal database is in process.

2.3 SciLifeLab – Clinical Genomics Unit SLL

Science for Life Laboratory, SciLifeLab, is a national center for molecular biosciences with focus on health and environmental research. SciLifeLab is a national resource and collaboration between four universities: Karolinska Institutet, KTH Royal Institute of Technology, Stockholm University and Uppsala University. It

[1] Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Richards, S. et al. Genetics in Medicine (2015) 17, 405-423.

is not a separate legal entity, and the cooperation clinics have the medical responsibility. Selected analyses at Clinical Genomics, SciLifeLab, are accredited to the ISO 17025 standard.

The Clinical Genomics unit belongs to the Next Generation Diagnostics platform at SciLifeLab. The primary affiliation is with the Department of Microbiology, tumour and cell biology at the Karolinska Institute, but the unit delivers services to among other the CMMS at Karolinska University Hospital, also participating at the workshop. The mission is to create a state-of-the-art infrastructure for translational research enabling clinical diagnostic analyses of patient samples, to introduce and validate the utility in routine medical care in collaboration with Swedish healthcare, and to carry out the work in a manner that is competitive on an international level.

High Throughput Sequencing (HTS) in diagnosis

SciLifeLab switched from WES to WGS in mid-2015. The split is currently 98% WGS and 2% WES. The clinical tests are done based on panel approaches, i.e. only variants in a selected list of genes are reported. The gene content of the panels are set up, maintained and owned by the clinical collaborators. Work is done on both time critical and normal cases; samples are submitted from clinics with different priority levels based on clinical need.

Variants in entire genome or exome are analyzed bioinformatically, but results are only reported for the disease specific panel. The entire analysis is fully automated until the point of clinical interpretation. Results are reported to collaborating clinics while still waiting for the confirmation of sample identity check. The entire genome or exome is available on request; if needed results can be made available within a few hours. Structural variance (insertions, deletions) will be analyzed from next release, estimated to autumn 2016. For acute setting scenario the team has demonstrated in a publication the principle of 15 hours whole-genome sequencing and analysis and is currently trying to lower the technical and financial barriers in order to make this available on the routine basis.

Process and infrastructure

Sample preparations are done in custom-built labs with high level of automation and controlled using Illumina BaseSpace Clarity LIMS. Sequencing is done using Illumina HiSeq X, HiSeq 2500 and MiSeq. Analysis is done through in-house developed pipeline (MIP, lead developer Henrik Stranneheim, CMMS, KS) for alignment, variant calling and functional annotation to produce a list of ranked variants for clinical grade analysis.

SciLifeLab currently holds an in-house computational cluster which is considered sufficient for now and for the short term. For future upgrade, a private cloud solution is considered to be more scalable and cost-efficient, but this is regarded to be more of a storage issue than a computation issue. Medium-term alternatives include hardware or software accelerated solutions, for example Genalice MAP or Edico Genome DRAGEN.

WES / WGS is used to diagnose patients with inborn errors of metabolism, primary immunodefincies, skeletal dysplasia, syndromes, neuromuscular disorders etc. More than 1200 samples were analysed during the last 27 months using the pipeline developed and run in cooperation with CMMS.

2.4 Karolinska Universitetssjukhuset – Centre for Inherited Metabolic Diseases (CMMS)

Karolinska University Hospital is one of the largest University hospitals in Europe, employing 15800 people serving the 1.6 million annual patient visits. The Centre for Inherited Metabolic Diseases (CMMS) headed by Anna Wedell serves all of Sweden through cross-disciplinary teams with expertise on all aspects of metabolic disorders, such as pediatrics, neurology, endocrinology, clinical genetics, clinical

chemistry, biochemistry, analytical chemistry, metabolism, molecular biology and bioinformatics. CMMS is also responsible for the national neonatal screening program for about 115000 newborns per year, currently screening for 24 diseases.

WES / WGS is used to diagnose patients with inborn errors of metabolism, primary immunodefincies, skeletal dysplasia, syndromes, neuromuscular disorders etc. More than 1200 samples were analysed during the last 27 months using the pipeline developed and run in cooperation with SciLifeLab.

The cooperation between CMMS and SciLifeLab is described as the development of a new unit, beyond the traditional thinking of segregation between research and clinic. Currently this is in the stage of "Proof of concept" – the idea to be transferred to other medical groups by showing what can be done.

2.5 DNV GL

Driven by the purpose of safeguarding life, property and the environment, DNV GL combines core competencies of technical and operational expertise, risk methodology and in-depth industry knowledge to enable organizations to advance the safety and sustainability of their business and build trust and confidence in their operations. DNV GL continuously invests in research and collaborative innovation to provide customers and society with operational and technological foresight. The company is a global leader in the Maritime, Oil & Gas, Energy, Business Assurance and Software business areas, operating in more than 100 countries. DNV GL Business Assurance is currently certifying and accrediting 2400 healthcare providers worldwide, and is involved in testing and certification of the quality and safety of medical devices. According to the DNV GL strategy, Life Sciences will be a new vertical for the group, the main themes being "preserving health" and "providing food". The strategic focus on preserving health includes two main industry segments; health providers and health suppliers. A dedicated multi-disciplinary research group is focused on understanding patient safety issues and risk, and on risk related to implementation of genomics in clinics.

DNV GL engages in standardisation and harmonisation as a key approach towards managing risks. This includes participating in international committees, national initiatives and working groups / industry consortiums, and also developing DNV GL rules and standards where there is a need. The company has extended experience with standardization work across industries. Involvement of both experts and users is essential in standardisation processes, where effectiveness of work processes must be balanced with the increasing consensus of a larger group of stakeholders involved. Aligning expectations through agreement on target for the outcome is another success factor.

DNV GL has 15 years' experience with healthcare assurance, and has been focusing on the genomics area for the last 1.5 years. The motivation for looking into genomics was based on need for major healthcare clients to understand the technology risk and opportunities, and focus is now on understanding the status of clinical implementation of NGS, and needs for quality assurance, verification and validation. Risks related the implementation of this technology is high and complex, there is a need to build trust between stakeholders. The DNV GL Business Assurance in Next Generation Genomics (BANGG) model is focusing on three different aspects of clinical implementation of NGS; quality assurance of process, governance of genomic database(s), facilitation of sharing.

DNV GL does not have any ambition to own data, but is in line with the 3rd party role of the company working across industries to develop the role as a data custodian, focusing on governance and facilitation of regulated sharing.

2.6 Other relevant projects: BigMed

OUS, Karolinska / SciLifeLab and DNV GL are partners and participants in BigMed, a research project recently funded by the Norwegian Research Council, with some overlap with the objectives of the work in the workshop.

3 INTRODUCTION

The workshop was initialized by OUS through contacts between the participating parties through other clinical and research collaboration processes. The workshop took place in Oslo 30.-31. May 2016 and included participants from the four clinical entities presented in the next section and DNV GL.

3.1 Objectives and agenda

The objectives for the workshop were proposed in the workshop invitation and confirmed at the beginning of the workshop to include:

- Review current clinical variant pipelines in the three laboratories; discuss common challenges and identify areas where standardisation/harmonisation could be beneficial.
- Identify what specific data would be valuable for laboratories to be able to share in the short, medium and long term as well as current technical, legal and ethical barriers that hinder sharing this data today.
- Discuss potential models for future cooperation and agree on next steps

Expectations to / motivation for the workshop included networking, learning from other participants, improvement opportunities, understanding benefits of sharing for institutions and for patients, exploring of opportunities for and barriers to sharing, learning about experiences on practical implementation of ACMG and AMP guidelines[2] and best practices in using these guidelines.

During the workshop, the six process steps of the clinical WGS / WES pipelines (Figure 1) of the participating institutions were mapped with respect to what is done, how is it done (hardware / software used) and who does it (institution or competences). Steps including quality controls were identified, as well as reference standards / guidelines applied. The workshop participants discussed what type of data they would like to have access to from the other institutions, and functional requirements to these data. The workshop concluded with discussions on opportunities for further cooperation, and agreeing on next steps for cooperation and follow-up actions assigned to action responsible. The detailed agenda is provided in section 14.2.

[2] Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Richards, S. et al. Genetics in Medicine (2015) 17, 405-423.

4 EXOME / GENOME SEQUENCING & CLINICAL PIPELINES

The tables in section 4 subsections are made available in separate excel-file.

Mapping of clinical pipelines for exome / genome sequencing for the three clinical entities was done to identify similarities and differences in design and operations. The clinical process was mapped according to the principal process steps as described in Figure 1, under the following assumptions:

- The target is WES and WGS
- There may be small differences in the exome / genome processes, but they were treated as one process during the workshop

In general, whole exome / genome sequencing is organized to focus on gene panels with established relevance for specific diseases. Analysis are made based on in silico panels.

Figure 1 Mapping of WES / WGS variant pipelines - process steps

The process steps were mapped to identify what is done, how is it done, and who does it as described in Table 1. The outcome of the mapping exercise is presented in tables in sections 0 to 4.6., including identification of steps where quality control is carried out and/or where a reference law / guideline / standard is used. The information presented in the tables below reflect the key words provided by the workshop participants, complemented with notes from discussions. Summary photos of the session are provided in section 14.3.1.

Table 1 Explanation of colour codes in process mapping

	What is done
	How is it done (software / hardware)
	Who does it (institution, competence)
	Notes
	Steps where QC is included or a reference law / guideline / standard is used

4.1 Clinical pipeline - Consent and sample taking

Table 2 Clinical pipeline - Consent and sample taking

		Consent and sample taking					
		Clinical decision	**Information to patient**	**Consent**		**Sample**	**Extraction**
Karolinska/ SciLifeLab	What	Clinical decision	Information to patient	Document consent	Consent (written informed consent)	Blood sample in 99% of the cases.	Extraction
Karolinska/ SciLifeLab	How (SW)			Medical records	Document with patient written consent is scanned	LIS	
Karolinska/ SciLifeLab	How (HW)						
Karolinska/ SciLifeLab	Who	Testing is always ordered from within own team, e.g. for epilepsy. Multidisciplinary team / treating physician – there is always an MD in the team who decides on sequencing.	Treating physician (usually at Karolinska, but could be from all of Sweden e.g. for mitocondrial patients)	Written consent not needed, but the patient agrees and the physician records agreement			Lab
Karolinska/ SciLifeLab	Notes	Will refer the patient elsewhere if other entities have better competence on the specific case.		The consent form states that the samples may be re-analysed after two years. This is typically done upon request from the clinician who is investigating an unsolved case. It is easy to maintain overview as the community working with metabolic disease in Sweden is small.	If there is a need to revisit sequencing data a new consent must be acquired. This includes cases where gene panels are updated later which may lead to new analysis. This is implemented but not systematized. By High suspicion negative findings filters can be lifted (i.e. the analysis broadened), and one must go back to the patient for written consent.		
Rigs-hospitalet	What	Clinical decision	Information to patient	Informed written consent.		Blood sample in 99% of the cases.	Results
Rigs-hospitalet	How (SW)						
Rigs-hospitalet	How (HW)						
Rigs-hospitalet	Who	Team - clinical group. Includes clinical geneticist (MD with specialization in genetics).	Close cooperation with pediatrician as patients are mostly children. Parents informed by clinical geneticist.	Team – clinical group	Team – clinical group	Team – clinical group	

Rigs-hospitalet	Notes	Decision criteria: Nice to know not sufficient.		Written consent according to DSMG guidelines / policy papers. Select level of information: All, only treatable, very limited. There is a clause stating that patient in some cases may be informed even if he/ she has asked not to be informed.			
OuS	What	Clinical decision	Info to the patient / parents	Consent not required for diagnostic samples		Sample – mostly blood, some other samples have been used (foetuses)	Extraction of DNA
OuS	How (SW)					Swisslab LIMS	
OuS	How (HW)						
OuS	Who	Clinical specialist					Reception of samples at AMG
OuS	Notes	Trials: samples from neurologists and paediatricians. Mitochondrial disease has a special cooperation with NSC.		Refers to the Biotechnology act; consent is not needed for diagnostics. Parents must sign a consent form whether they also want information about incidental findings. Future: Specific consent will be collected for inclusion in Norvariom database. Licence to store genomic data has been granted by the Norwegian Data Protection Authority. Desire to store more data as reference for diagnostics and for research. Relevant IKTPLUSS (ICT) project on developing solutions for dynamic consent related to the EPJ MinJournal. To access the exome a research group must be involved and the use must be justified. Process for re-access when gene panels are revised is not defined.			

4.2 Clinical pipeline - Sample preparation and raw data generation

Table 3 Clinical pipeline - Sample preparation and raw data generation

		Sample preparation and raw data generation						
		QC		Library preparation	QC	Sequencing	Storage of raw data	Post sequencing QC
Karolinska/ SciLifeLab	What	DNA QC at reception at SciLifeLab	Control ID typing	Library prep	QC	Sequencing	Backup	Demux
Karolinska/ SciLifeLab	How (SW)	Illumina BaseSpace Clarity LIMS (Genologics)	Third party (Massarray panel)					Casava, in-house databases
Karolinska/ SciLifeLab	How (HW)	Qubit or QuantIT		Bravo Robot or manually. WES and other targeted analyses: Sureselect XT (CRE bait set for WES). WGS: TruSeq DNA PCR free (standard input).	WES: Qubit/QuantIT + Fragment Analyzer WGS: qPCR	WGS: HiSeq X (Illumina) WES, other targeted panels: HiSeq2500 (Illumina) Small targeted panels: MiSeq (Illumina)	Tape	
Karolinska/ SciLifeLab	Who	Lab		Lab	Lab	Lab	Automated	Automated
Karolinska/ SciLifeLab	Notes						Backup encrypted and taped	
Rigs-hospitalet	What	DNA Sample QC		Library prep	QC	Sequencing		
Rigs-hospitalet	How (SW)	Nanodrop AO-Q-CT			BioAnalyzer			#reads #uniformity % on target ReadLength Equal representation of each PCR pool
Rigs-hospitalet	How (HW)			Manual AmpliSeq	qPCR (SDS 7500)	ThermoFisher IonTorrent system; Ion Chef; Ion Proton		Manual
Rigs-hospitalet	Who			Laboratory technicians	Laboratory technicians			Lab technician and scientist
Rigs-hospitalet	Notes			Reference: Assay guideline		Use the ThermoFisher Ion Proton custom standard settings throughout the pipeline, which are updated every 3 months. Updates are not subject to quality checks. The unit does not include bioinformaticians and has limited capabilities in changing / tuning parameters.		
OuS	What	QC DNA	SNP-ID	Library prep	QC	Sequencing	Write raw sequencing data (bcl files) to disk	De-multiplexing and post-sequencing QC

OuS	How (SW)				Swisslab Clarity LIMS	Swisslab Clarity LIMS	Isilon storage system	Clarity LIMS automatically starts de-multiplexing script which runs Casava?
OuS	How (HW)	Quant-IT	TaqMan probes 23SNP library prep ArrayCard	Bravo robot, SureSelect 50 Mb v5	Tapestation qPCR	HiSeq 2500 MiSeq		FastQC to generate QC plots (PDF). Bioinformatician eyeball's the plots.
OuS	Who		HTS lab engineer			NSC		NSC
OuS	Notes	Reference: SureSelect recommendations	Reference: Internal	Reference: Sure Select recommendations	Reference: Illumina recommendations	Increasingly the diagnostic lab engineers will take over sequencing machine work from NSC staff		Aiming to automate post-sequencing QC (the eyeballing of FASTQC plots)

4.3 Clinical pipeline – Sequence alignment and variant calling

Table 4 Clinical pipeline – Sequence alignment and variant calling

		Sequence alignment and variant calling								
			Alignment		**Variant calling**		**QC**	**Validation and control**		
Karolinska/ SciLifeLab	What		Alignment	BAM Recalibration	Variant calling		QC	NIST QC	Match with control ID typing	Coverage %>10x
Karolinska/ SciLifeLab	How (SW)		BWA		GATK (SNV) SAM tools (SNV) Freesbaves (SNV) Manta (CNV,SV) Delly (CNV, SV)		FASTQC / MIP Picard Tools GATK Sambamba	After each update of bioinformatic pipeline. Old data reanalysed.	An aliquot of each sample send to third party for analysis using two 29plex MassArray panels (in-house designed to maximise information content for samples of Swedish origin). Obtained results matched with NGS data. Delivery of results also before QC data received.	Chanjo (in-house developted QC tool for coverage estimates). Uses Sambamba in the background. Calculates % of bases covered at 10x or more in the selected gene panel. Also calculates the number of transcripts for genes in panel that are not fully covered (ie every base >10x).
Karolinska/ SciLifeLab	How (HW)		In-house computing cluster	In-house computing cluster	In-house computing cluster			In-house computing cluster	In-house computing cluster	In-house computing cluster
Karolinska/ SciLifeLab	Who		Part of automated pipeline	Part of automated pipeline	Part of automated pipeline			Part of automated pipeline	Part of automated pipeline	Chanjo as part of automated pipeline.
Karolinska/ SciLifeLab	Notes	The entire alignment and variant calling pipeline is set up according to GATK best practice, and modified to include additional callers.			For SNVs three different callers are used, and variants identified in any of these are retained. The purpose is to maximise sensitivity. Reduced specificity is addressed through variant prioritisation (ranking).		"For each sample following quality checks are carried out: 1) Coverage at genome level as well as in targeted panel 2) Gender should match expected (XY coverage, SNPs on X chromosome) 3) Sample mixup; analysis of 56 SNPs using third party and results should match SNP calls in NGS data Overall	"References: GIAB / NIST. Two samples."		

							bioinformatic pipeline is QC using external samples (NIST) as well as ca 20 previously analysed in-house samples. This is done at each update of clinical pipeline, every three months."			
Rigs-hospitalet	What		Alignment				QC #reads #uniformity % on target			
Rigs-hospitalet	How (SW)		Ion-Torrent, standard settings							
Rigs-hospitalet	How (HW)									
Rigs-hospitalet	Who			Reference: Specifications for service providers (Axxx RDY)	Analysis is based on a closed system. Alignment is based on three different algorithms. Performance analysis not carried out, but some consistency checks. Control sample included to filter out sequencing mistakes Run 8 samples each time, all interesting findings are confirmed with Sanger Sequencing. Mean [coverage] rate is 80-100X. 58 MegaBases. Almost 100% (At least 96% of at least 20x) False positives Sanger will pick up. False negatives are always a problem. Always look at single genes.					
Rigs-hospitalet	Notes									
OuS	What	Mapping	Realignment / BQSR		Variant calling	VQSR	QC	Trend analysis	Control sample	Gold standard performance validation
OuS	How (SW)	BWA	GATK - BQSR		GATK Haplotype caller	GATK VQSR - Best practice settings	In house tools + Picard Tools	QC data from each sample is stored and can be periodically plotted	- NA12878 - GIAB - Inhouse tool	
OuS	How (HW)									
OuS	Who	Variant calling (vc) pipeline (automated)	Variant calling (vc) pipeline (automated)	Variant calling (vc) pipeline (automated)	Variant calling (vc) pipeline (automated)	Variant calling (vc) pipeline (automated)		Bioinformatician	Bioinformatician	Bioinformatician
OuS	Notes	Reference: GATK – Best practice, vc pipeline						Genome in a bottle (GIAB) and benchmarking tools		

4.4 Clinical pipeline – Variant annotation

Table 5 Clinical pipeline – variant annotation

		Variant annotation								
Karolinska/ SciLifeLab	What	Functional	Frequency	Conservation	Inheritance	Impact	Clinical significance	Domain	Ranking	Upload to Scout
Karolinska/ SciLifeLab	How (SW)	VEP / SNPEFF Annovar / VT Genmod	ExAc 1000 Genomes Browser Local Genbank (MT)	Phylop PhasdCons GerP	MIP Genmod OMIM	MIP CADD SIFT Polyphen ClinVar Spidex Consequence (SO terms)	OMIM ClinVar Local Gene panels HPO Pathogenic Transcript Red. penetrance	PFAM	Genmod	Scout
Karolinska/ SciLifeLab	How (HW)	In-house computing cluster	In-house computing cluster	In-house computing cluster	In-house computing cluster	In-house computing cluster	In-house computing cluster	In-house computing cluster	In-house computing cluster	Hosted resource at SciLifeLab
Karolinska/ SciLifeLab	Who	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Part of automated pipeline	Bioinformatician, after final manual QC
Karolinska/ SciLifeLab	Notes	Reference: SO-terminology		Reference: HPO			Ranking is done within the in silico filter, the 5-10.000 highest ranking variants are loaded into Scout – more can be requested later.			
Rigs-hospitalet	What	Control exome (vcf file) - HGMD®-genes - UCSC common SNP Frequency - Allele read count>7 -5000 Exomes MAF < 0,01 Location - Slicesite (10bp), exonic								
Rigs-hospitalet	How (SW)	Ion Reporter (Thermo Fisher) with customized annotation tools (HGMD genes, or panel subsets). Filtering against control sample and in-house database.								
Rig-shospitalet	How (HW)									
Rig-shospitalet	Who	Clinical laboratory geneticist								
Rig-shospitalet	Notes								Hard filtering based on indication, usually left with 200-300 variants.	

OuS	What							Filter variants on gene panel	Create excel-report on ”all” variants and frequency-filtered variants (>=1%)	
OuS	How (SW)	VEP - Ensembl + Refseq transcripts - All data	ExAc 1000 Genomes Browser In-house freq.	ClinVar HGMD Pro®	OMIM				External databases (ExAc, 1000 Genomes Browser): < 1% Internal database: <5%	
OuS	How (HW)									
OuS	Who	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	Variant calling pipeline (automated)	
OuS	Notes	Reference: VEP: Sequence ontology (efforts)							The hard filtering will be loosened (threshold raised) if no finding; 1 →10%	

4.5 Clinical pipeline – Variant interpretation

Table 6 Clinical pipeline – variant interpretation

		Variant interpretation					
		Determination of clinical significance, classification			**Verification**		**Clinical report**
Karolinska/ SciLifeLab	What	Determine clinical significance			Sanger functional validation		
Karolinska/ SciLifeLab	How (SW)	Scout (Puzzle) visualisation tool	Literature, HGMD®, Etc	Multidisciplinary meetings – weekly addressing about 20 patients.	Biochemical / molecular lab		
Karolinska/ SciLifeLab	How (HW)	Hosted resource at SciLifeLab					
Karolinska/ SciLifeLab	Who	Each patient analysed by appropriate expert team, consisting of geneticist and physicians (MD). Typicallt independent assessments. Looks at data within defined scope / panel. The most common clinical panels are in the tool and can be used for filtering. HPO terms can be used as filters. Analysis can be done without clinical data and anonymized. Easy to access positive and negative findings. If negative finding and convincing phenotype, may also look at heterozygotes. Comments / annotation in the Scout tool. Access permissions managed within the Scout tool; can include groups with alternative competence if needed. Ability to share case with other expert teams. Incidental findings will not occur in research; will only get incidental findings when actively looking at genes. Not relevant / ethical to report to patient in acute setting.		Multidisciplinary team (Friday meeting)			
Karolinska/ SciLifeLab	Notes						
Rigshospitalet	What	Manual curation	Classification 1-5		Sanger verifications	Segregation analysis	Clinical report
Rigshospitalet	How (SW)						Ion Reporter connected to an in house database containing own experience data. All variants in the BAM file considered.
Rigshospitalet	How (HW)						
Rigshospitalet	Who	Clinical lab geneticist	WES team incl. clinical geneticist,				

			clinical lab geneticist				
Rigshospitalet	Notes	Reference: Clinical lab geneticist (EU certification)	Reference: ACMG				Reference: ACMG, EMQN
OuS	What		Classification of all variants 1-5		Verify variant Only for low quality SNV and deletions/ duplications/ insertions.		Lab report to clinician including coverage report
OuS	How (SW)		ACMG based in-house procedure, in-house database. All variants filtered out: class 1. Remaining variants scored 2-5. Rationale recorded for scoring of variants class 3-5. Small panels, usually 2-3 variants class 3-5. Recessive findings are assigned class 4-5 even if only one finding.		Sanger MLPA ArrayCGH		Swisslab LIMS Coverage: Inhouse script (pdf)
OuS	How (HW)						
OuS	Who		Independent laboratory engineers (manual)				Laboratory engineer and laboratory clinician
OuS	Notes		Reference: ACMG		Reference: Criteria for Sanger sequencing confirmation (from published paper)	Coverage report (PDF) is issued to see if a gene is lacking completely and to see if some regions have coverage less than ten times.	Reference: Recommendations EMQN Look to guidelines

4.6 Clinical pipeline – Clinical use

Table 7 Clinical pipeline – clinical use

		Clinical use				
		Reporting	**Treatment**	**Genetic counselling**		
Karolinska/ SciLifeLab	What	Report writing	Clinical action (treatment)	Genetic counselling		
Karolinska/ SciLifeLab	How (SW)	LIS				
Karolinska/ SciLifeLab	How (HW)					
Karolinska/ SciLifeLab	Who	Geneticist Physician	Treating physician	Physician		
Karolinska/ SciLifeLab	Notes	Clearly defined format for reporting.				
Rigshospitalet	What			Genetic counselling	More solved cases	Prenatal diagnosis / treatment
Rigshospitalet	How (SW)					
Rigshospitalet	How (HW)					
Rigshospitalet	Who			Clinical geneticist	Treating physician	
Rigshospitalet	Notes	Findings to be reported must have a consequence for either patient or family. Many of the analyses are prenatal.		Communication to treating physician (often paediatrician, neuro-paediatrician) is through clinical geneticist. It is regarded as important that clinical geneticist explains findings to physician.		
OuS	What	Reporting		Genetic counselling	Offer family testing	
OuS	How (SW)	EMR (Swisslab or DIPS)				
OuS	How (HW)					
OuS	Who	Geneticist Physician		Clinical geneticist Genetic counsellor		
OuS	Notes	There is always a written report. There is no multidisciplinary reporting, but there is a plan to do this for intellectual disabilities. Level 3-5 findings are reported. Unsolved cases → open exome (Filtus) → research		Sometimes there is referral for genetic counselling.		

5 TECHNOLOGIES USED (HARDWARE)

Table 8 Technology platforms used for data generation

	Library prep	Sequencing platform
Karolinska / SciLifeLab	Bravo Robots in both PRE and POST PCR steps (Agilent Technologies)[3] WES: SureSelect^{XT} (Agilent Technologies)[4] + CRE WGS: TruSeq PCR Free (Illumina)[5]	WGS: HiSeq X System (Illumina)[6] WES: HiSeq 2500 System (Illumina)[7] Targeted WES: MiSeq (Illumina)[8]
Rigshospitalet	7500 Fast System SDS (ThermoFisher)[9]	Ion Torrent (ThermoFisher)[10]* - Ion Chef™ System - Ion Proton™
OUS	Bravo Robot (Agilent Technologies)[11] SureSelect^{XT} (Agilent Technologies)[12] 50 Mb v5	HiSeq 2500 System (Illumina)[13]

*Custom standard settings are used throughout the pipeline, updated every 3 months. Updates are not subject to quality checks as there are no bioinformaticians in the unit.

6 DATA FILE FORMATS USED

The following file formats are used in the clinical pipelines based on WES / WGS sequencing:

Data types	File format	Description
Raw reads, processed reads	FastQ[14]	the FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the de facto standard for storing the output of high-throughput sequencing instruments such as the Illumina Genome Analyzer.[15]
Read alignments (mapped to reference genome, post-alignment processing)	BAM[16]	The Binary (sequence) Alignment/Map (BAM) is a binary format for storing sequence data, including basecalls (reads), quality scores, alignment data, etc. The corresponding SAM Format can be used to store sequence data, both aligned as well as unaligned, in a human readable format.
Variant calls	VCF[16]	The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations including metadata. The VCF specification is now maintained by Global Alliance for Genomics and Health Data Working group file format team[17]. The main version of the VCF specification can be found on the HTS-spec GitHub page. The Bcf format is the binary counterpart of the vcf file format.

[3] https://www.agilent.com/en-us/products/automation-solutions/automated-liquid-handling/bravo-automated-liquid-handling-platform
[4] http://www.genomics.agilent.com/en/SureSelect-DNA-Library-Preps/SureSelectXT-Reagent-Kits/?cid=AG-PT-177&tabId=AG-PR-1302
[5] http://www.illumina.com/products/truseq-dna-pcr-free-library-prep-kits.html
[6] http://www.illumina.com/systems/hiseq-x-sequencing-system/x-five-system.html
[7] http://www.illumina.com/systems/hiseq_2500_1500.html
[8] http://www.illumina.com/systems/miseq.html
[9] http://www.thermofisher.com/no/en/home/life-science/pcr/real-time-pcr/real-time-pcr-instruments/7500-fast-real-time-pcr-system.html
[10] https://www.thermofisher.com/no/en/home/brands/ion-torrent.html
[11] https://www.agilent.com/en-us/products/automation-solutions/automated-liquid-handling/bravo-automated-liquid-handling-platform
[12] http://www.genomics.agilent.com/en/SureSelect-DNA-Library-Preps/SureSelectXT-Reagent-Kits/?cid=AG-PT-177&tabId=AG-PR-1302
[13] http://www.illumina.com/systems/hiseq_2500_1500.html
[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/
[15] https://en.wikipedia.org/wiki/FASTQ_format
[16] SAM/BAM and related specifications including vcf/bcf are available at http://samtools.github.io/hts-specs/
[17] http://ga4gh.org/#/fileformats-team

7 TOOLS USED (SOFTWARE)

Table 9 LIS and EPJ used for managing traceability, storage and reporting

Institution	Laboratory Information (management) System (LIS)
Karolinska / SciLifeLab	Illumina BaseSpace Clarity LIMS
Rigshospitalet	IonReporter, Logos og In-house longtime storage of datafiles
OUS	SwissLab LIMS, DIPS, Clarity LIMS

Table 10 Variant calling / annotation / interpretation pipeline

Institution	Laboratory Information (management) System (LIS)
Karolinska / SciLifeLab	MIP - SciLifeLab's own developed pipeline consisting of in house and open tools
Rigshospitalet	TorrentSuite for variant calling, IonReporter for annotation and interpretation
OUS	Variant calling pipeline (vcpipe – internally developed pipeline)

Table 11 Tools used for sequence alignment, variant calling and variant annotation

Process step	Process substep	Tool	Description	Institution
Sequence alignment and variant calling	Alignment	BWA[18]	Burrows-Wheeler Aligner (BWA) is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.	Karolinska / SciLifeLab OUS
Sequence alignment and variant calling	Alignment QC	Picard Tools[19]	Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/ BAM/ CRAM and VCF.	Karolinska / SciLifeLab OUS
Sequence alignment and variant calling	Alignment QC	Sambamba[20]	sambamba view allows to efficiently filter SAM/BAM/CRAM files for alignments satisfying various conditions, as well as access its SAM header and information about reference sequences.	Karolinska / SciLifeLab
Sequence alignment and variant calling	Variant calling	SAM tools[21,22]	SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.	Karolinska / SciLifeLab
Sequence alignment and variant calling	Variant calling	freebayes[23]	a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs, indels, MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment	Karolinska / SciLifeLab

18 http://bio-bwa.sourceforge.net/
19 https://broadinstitute.github.io/picard/
20 http://lomereiter.github.io/sambamba/
21 http://samtools.sourceforge.net/
22 https://github.com/samtools/samtools
23 https://wiki.gacrc.uga.edu/wiki/Freebayes

Process step	Process substep	Tool	Description	Institution
Sequence alignment and variant calling	Variant calling	Manta[24]	Structural variant and indel caller for mapped sequencing data	Karolinska / SciLifeLab
Sequence alignment and variant calling	Variant calling	Delly (SV)[25]	Structural variant discovery by integrated paired-end and split-read analysis	Karolinska / SciLifeLab
Sequence alignment and variant calling	Variant calling	HaplotypeCaller[26]	calls germline SNPs and indels via local de-novo assembly of haplotypes in an active regionA component of GATK best practices.	OUS
Variant annotation	Functional	VEP[27]	The Variant Effect Predictor (VEP) tool determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions	Karolinska / SciLifeLab OUS
Variant annotation	Functional	SnpEff[28]	Genetic variant annotation and effect prediction toolbox: annotates and predicts the effects of variants on genes and proteins	Karolinska / SciLifeLab
Variant annotation	Functional	Annovar[29]	Utilizes update-to-date information to functionally annotate genetic variants detected from genomes	Karolinska / SciLifeLab
Variant annotation	Variant calling	Vt[30]	A variant tool set that discovers short variants from NGS data. Mainly used for normalising and de-composing VCF files.	Karolinska / SciLifeLab
Variant annotation	Functional / Inheritance /	genmod[31, 32]	Analyzes and annotates genomic variations and genetic patterns of inheritance in VCF files	Karolinska / SciLifeLab
Variant annotation	Conservation	Phylop[33]	PhyloP scores measure evolutionary conservation at individual alignment sites.	Karolinska / SciLifeLab
Variant annotation	Conservation	PhastCons[34]	Identify conserved elements or produce conservation scores	Karolinska / SciLifeLab
Variant annotation	Conservation	GERP[35]	Genomic Evolutionary Rate Profiling (GERP) identifies constrained elements in multiple alignments by quantifying substitution deficits.	Karolinska / SciLifeLab
Variant annotation	Interpretation	CADD[36]	Combined Annotation Dependent Depletion (CADD) is a tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome.	Karolinska / SciLifeLab

24 https://github.com/Illumina/manta
25 https://github.com/tobiasrausch/delly
26 https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php
27 http://www.ensembl.org/info/docs/tools/vep/index.html
28 http://snpeff.sourceforge.net/
29 http://annovar.openbioinformatics.org/en/latest/
30 http://genome.sph.umich.edu/wiki/Vt
31 http://opensource.scilifelab.se/projects/genmod/
32 https://github.com/moonso/genmod
33 http://compgen.cshl.edu/phast/help-pages/phyloP.txt
34 http://compgen.cshl.edu/phast/help-pages/phastCons.txt
35 http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html

Process step	Process substep	Tool	Description	Institution
Variant annotation	Functional	SIFT[37, 38]	Sorting Intolerant from Tolerant (SIFT) predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. SIFT can be applied to naturally occurring nonsynonymous polymorphisms and laboratory-induced missense mutations.	Karolinska / SciLifeLab
Variant annotation	Functional	Polyphen[39]	PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.	Karolinska / SciLifeLab
Variant annotation	Functional	Spidex[40]	SPIDEX™ predicts how SNVs affect RNA splicing in humans	Karolinska / SciLifeLab
Variant interpretation	Interpretation	Scout[41]	VCF visualization interface, clinical reporting of ranked variants. Contains a number of tools within the tool.	Karolinska / SciLifeLab

[36] http://cadd.gs.washington.edu/
[37] http://sift.bii.a-star.edu.sg/
[38] http://sift.jcvi.org/
[39] http://genetics.bwh.harvard.edu/pph2/
[40] http://www.deepgenomics.com/spidex/
[41] https://github.com/Clinical-Genomics/scout

8 DATABASES USED

Table 12 Databases[42] used as references for the clinical variant interpretation pipelines

Process step	Process substep	Database	Content	Institution
Variant annotation		RefSeq[43]	The NCBI Reference Sequence (RefSeq) Database is a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.	OUS
Variant annotation	Frequency	ExAc[44]	The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community. The data set provided on this website includes 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies.	Karolinska / SciLifeLab OUS
Variant annotation	Frequency, ranking	1000 Genomes browser[45]	Variant calls and supporting sequence and read alignments produced by the 1000 Genomes Project[46] is made available through the 1000 Genomes browser. The project was an international research effort to establish a detailed catalogue of human genetic variation.	Karolinska / SciLifeLab OUS
Variant annotation	Frequency	UCSC Common SNPs[47]	A track within the UCSC Genome Browser derived from NCBI dbSNP	Rigshospitalet
Variant annotation	Frequency	Genbank (MT)	?	Karolinska / SciLifeLab
Variant annotation	Interpretation	ClinVar[48]	Clinical Variants (ClinVar) database of aggregated information about genomic variation and its relationship to human health.	Karolinska / SciLifeLab OUS
Variant annotation	Interpretation	OMIM[49, 50]	Online Mendelian Inheritance in Man, An online catalogue of human genes and genetic disorders	OUS Karolinska / SciLifeLab
Variant annotation	Functional	PFAM[51]	The Protein family (Pfam) database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)	Karolinska / SciLifeLab
Variant annotation	Annotation	HGMD[52]	Human Gene Mutation Database - resource for comprehensive data on published human inherited disease mutations	OUS

[42] The involved laboratories are using databases for investigation of variants, but are currently not actively contributing to databases with new information. Entering into partnership with the Global Alliance for Genomics & Health (GA4GH, http://genomicsandhealth.org/) requires contribution to public databases such as ClinVar.

[43] http://www.ncbi.nlm.nih.gov/refseq/

[44] http://exac.broadinstitute.org/

[45] http://browser.1000genomes.org/index.html

[46] http://www.1000genomes.org/

[47] https://genome.ucsc.edu/cgi-bin/hgGateway

[48] http://www.ncbi.nlm.nih.gov/clinvar/

[49] http://omim.org/

[50] http://www.ncbi.nlm.nih.gov/omim

[51] http://pfam.xfam.org/

9 RECOMMENDATIONS, GUIDELINES AND STANDARDS

Table 13 Recommendations, guidelines and standards used

Process step	Reference	Introduction	Institution
Consent	DSMG guidelines[53]	The Danish Society for Medical Genetics (DSMG) has developed guidelines to ensure quality in genetic counselling and testing in accordance with international standards.	Rigshospitalet
Consent	Biotechnology Act[54,55]	The Norwegian Biotechnology Act regulates medical use of biotechnology, including requirements for consent to genetic testing	OUS
Variant calling	GATK best practices[56]	The GATK Best Practices provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data. There are several different Best Practices workflows tailored to particular applications depending on the type of variation of interest and the technology employed. The Best Practices documentation attempts to describe in detail the key principles of the processing and analysis steps required to go from raw reads coming off the sequencing machine, all the way to an appropriately filtered variant callset that can be used in downstream analyses. Wherever we can, we try to provide guidance regarding experimental design, quality control (QC) and pipeline implementation options, but please understand that those are dependent on many factors including sequencing technology and the hardware infrastructure that are at your disposal, so you may need to adapt our recommendations to your specific situation.	Karolinska / SciLifeLab OUS
Variant annotation	SO[57]	The Sequence Ontology (SO) is a collaborative ontology project for the definition of sequence features used in biological sequence annotation. A structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases. The SO is part of the OBO foundry[58], a collective of ontology developers.	Karolinska / SciLifeLab
Variant annotation	HPO[59]	The Human Phenotype Ontology (HPO) aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality. The HPO is currently being developed using the medical literature, Orphanet, DECIPHER, and OMIM. HPO currently contains approximately 11,000 terms (still growing) and over 115,000 annotations to hereditary diseases. The HPO also provides a large set of HPO annotations to approximately 4000 common diseases.	Karolinska / SciLifeLab
Variant interpretation – classification	ACMG[60]	American College of Medical Genetics (ACMG) has developed standards and guidelines for the interpretation of sequence variants, including classification of variants identified in genes causing Mendelian disorders. The recommendation describes a process for classifying variants into five categories (pathogenic, likely pathogenic, uncertain significance, likely benign and benign) based on criteria using typical types of variant evidence.	Rigshospitalet OUS

[52] http://www.biobase-international.com/product/hgmd

[53] http://dsmg.dk/jannes/index.php/dsmg-guidelines

[54] https://helsedirektoratet.no/lover/bioteknologiloven

[55] https://lovdata.no/dokument/NL/lov/2003-12-05-100?q=bioteknologiloven

[56] https://www.broadinstitute.org/gatk/best-practices/

[57] http://www.sequenceontology.org/

[58] http://www.obofoundry.org/

[59] http://human-phenotype-ontology.github.io/

Process step	Reference	Introduction	Institution
Variant interpretation – Clinical reporting	EMQN[61]	The European Molecular Genetics Quality Network is a not-for-profit organisation promoting quality in genetic testing by establishing, harmonising and disseminating best practice. UKAS accredited provider of External Quality Assessment (EQA) services.	Rigshospitalet OUS

[60] Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Richards, S. et al. Genetics in Medicine (2015) 17, 405-423.

[61] http://www.emqn.org/emqn/Home

10 BASIC QC STEPS - QA / QC / REFERENCE STANDARDS USED

10.1 Quality controls

Table 14 Quality control tools and reference materials used, including benchmarking[62]

Process step	Process substep	QC / tool	Description	Institution
Sequence alignment and variant calling	Alignment	BQSR[63]	Base Quality Score Recalibration is a data pre-processing step that detects systematic errors made by the sequencer when it estimates the quality score of each base call. Included in The Genome Analysis Tool Kit (GATK) (Broad Institute).	OUS
Sequence alignment and variant calling	Variant calling	VQSR[64]	Variant quality score recalibration (VQSR) calculates a new quality score to enable variant filtering in a way that allows analysts to balance sensitivity and specificity as finely as possible. Included in GATK.	OUS
Sequence alignment and variant calling	Variant calling QC	FASTQC / MIP[65]	A quality control tool for high-throughput sequence data. FastQC aims to provide a simple way to perform QC on raw sequence data coming from NGS pipelines.	Karolinska / SciLifeLab
Variant calling	Variant calling validation and control	NIST[66] / GIAB[67] and benchmarking tool	The National Institute of Standards and Technology (NIST) has organized the "Genome in a Bottle Consortium" (GIAB) to develop the reference materials, reference data, and reference methods needed to assess performance of human genome sequencing.	Karolinska / SciLifeLab OUS
Sequence alignment and variant calling	Variant calling validation and control	chanjo[68]	Coverage analysis for clinical sequencing, sample-specific quality measurement	Karolinska / SciLifeLab

[62] Further details of QC operations not available; methods and accept criteria not mapped in detail during the first workshop

[63] http://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr

[64] http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr

[65] http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

[66] www.nist.gov

[67] http://jimb.stanford.edu/giab

[68] http://www.chanjo.co/en/latest/

11 DATA SHARING

The workshop participants were invited to put forward wishes for data they wanted access to from other laboratories, and to discuss barriers and functional requirements for sharing of such data. Data sharing was discussed for the data types summarized in Table 15 and outlined in more detail in the sections below. A summary photo of the session is provided in section 14.3.2.

The data types discussed could be roughly separated in two groups; data related to the variants and data related to the quality of the interpretation pipeline (Table 15). Discussions on data to share, functional requirements and possible barriers were initialized and are preliminary comments summarized in Table 16 and Table 17.

It was generally agreed among the participants that basic principles for any sharing should be transparency, openness and flexibility; sharing and cooperation should not be limited to the workshop group.

Table 15 Data sharing – data types discussed and prioritized[69]

Priority#	Detailed information	Data	Related to variants or pipeline	Group leader detailed discussion
1	Table 17	Population frequencies	Variants	Morten
2	Table 17	Curated variant classification database	Variants	Dag
3	Table 17	Genomes and patient phenotype	Variants	Morten
4	Table 17	Genotype (full genomes) and patient phenotype database	Variants	Dag
5	Table 16	Matchmaking through accurate and standardized phenotype descriptions	Pipeline	Eidi
6	Table 17	"Everything" – FastQ files with phenotypes	Variants	
7	Table 16	Benchmarking	Pipeline	Dag
8	Table 16	Classification / ranking of variants , variant interpretation procedure (Application of ACMG)	Pipeline	Eidi
9	Table 16	Gene panels	Pipeline	Morten
10	Table 16	QC procedure: coverage mapping, verification, etc.	Pipeline	Morten
11	Table 16	Variants in hypernormal controls	Pipeline	Eidi
12	Table 16	Ability to query variant database by position	Pipeline	

11.1 Data sharing – access requested for variant interpretation pipeline data

The workshop prioritization of data related to the variant interpretation pipeline is listed in Table 16.

[69] Prioritization does not necessarily reflect expected value, but also the different interests of the individuals in the group and which type of data is seen as practical starting point and more long term goals.

Table 16 Data sharing – data relevant for the variant interpretation pipeline

Priority# (Table 15)	Data sharing requested	What to share	Functional requirements	Barriers identified	Notes from discussion	Group leader detailed discussion
5	Matchmaking through accurate and standardized phenotype descriptions	HPO-based phenotype description. Tools Referral		Challenge if phenotype is not available at referral. Need to get the physicians on board in using HPO terminology. Knowhow is low – awareness & rewarding?	Motivation: to be able to find out if there are patients with the same phenotype. Currently large volumes of unsystematic data. Should be standardized, e.g. HPO based, which reaches beyond genetics. Standardization of phenotype links to transformation of medicine across all medical disciplines, education and training needed. Implementation will have to include non-geneticists. A transformation to standardized phenotype description is a lot bigger than the genetics and will transform medicine. Some emerging solutions for such match making exist; MyGene2 (patient oriented), Geno2MP (professionally oriented). Need a project to get started.	Eidi
7	Benchmarking	"Gold standard" bench-marking data, "NIST data" for different lab flows, QC thresholds/ metrics.	Define steps for QC.	Resources: labor intensive – cost / benefit to be considered.	Propose to run same samples through the pipeline and compare. External benchmarking part of ISO requirements	Dag
8	Classification /ranking of variants , variant interpretation procedure (Application of ACMG)	Application of ACMG guidelines; documentation of ACMG guidelines classification used.		Different local / national interpretation on application of ACMG guidelines. How to trust each other's interpretations?	Share variants to align variant ranking / interpretations; Proposal: share a set, meet and discuss.	Eidi

Priority# (Table 15)	Data sharing requested	What to share	Functional requirements	Barriers identified	Notes from discussion	Group leader detailed discussion
9	Gene panels	Increasing complexity and value: - Simple list of genes - Definition of gene with rationale - Rationale for inclusion of genes in gene panels	The information must be standardized, accessible and updated. - Contact information for sharing laboratory - Standardized information - Accessible - Updated - Gene reference according to HGNC (Hugo Gene Nomenclature Committee) - Versioning - Documentation with references	Time Cost Dynamic gene panels, adjusted for new patient groups.	*In silico* panels Consensus gene panels (minimum for clinical incident), e.g. Genomics England app. Must be agile and not introduce additional step of approval – not prevent freedom to work.	Morten
10	QC procedure	Output from variant calling pipeline	Coverage mapping Verification (e.g. Sanger verification) Metadata - Information - Wetlab	Different pipelines		Morten
11	Variants in hypernormal controls	Share "pathogenic" variants in normal persons		Costs	Reference variants	Eidi
12	Ability to query variant database by position					

11.2 Data sharing – access requested for data related to variants

There are several layers of information on variants which could be shared with increasing complexity and value to the interpreting clinical laboratories (Figure 2). The workshop prioritization of data related to variants together with perceived complexity is listed in Table 17.

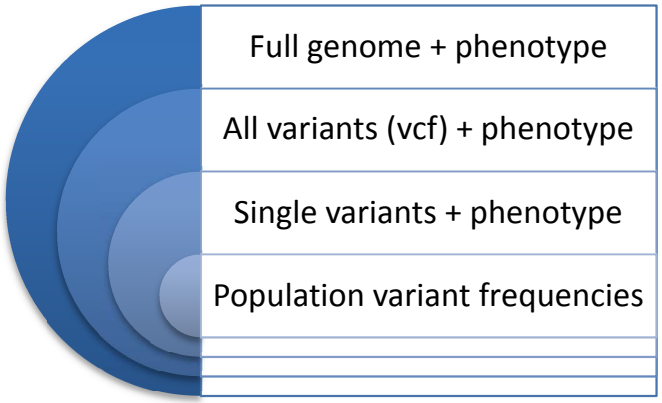


Figure 2 Increasing complexity and value of data related to genetic variants

Table 17 Data sharing – data relevant for variants

Priority# (Table 15)	Data sharing requested	What to share	Functional requirements	Barriers identified	Notes from discussion	Group leader detailed discussion
1	Population frequencies	Increasing complexity and value: - Aggregated data - Filtered: QC - Filtered: phenotype, proband, etc.	- Filters - QC threshold - Baseline - versioning	Need to agree on what to filter on and QC threshold/ baseline, versioning	Significant differences between the Nordic countries, and also some difference between northern and southern parts of Sweden and Norway. Frequencies of alleles, variants in the different genes; population frequencies. Database of frequencies. May be a "low hanging fruit", but there is a need to define quality criteria / thresholds.	Morten

Priority# (Table 15)	Data sharing requested	What to share	Functional requirements	Barriers identified	Notes from discussion	Group leader detailed discussion
2	Curated variant classification database	Guidelines Technical platform Standardized • Clinically annotated variants including justification for ACMG classification (1-5) with phenotype description. • To include pathogenic and non-pathogenic variants, and variants of uncertain significance (VUS).		Mainly technical Ethical – patient recognition	Short term. Variants of low frequency are not published anymore. For pathogenic variants this is useful. Variants could be submitted to ClinVar but this is not done much since it is time consuming. ClinVar may offer the needful, to be explored. Causative/pathogenic variants (cat 3-5) with standardized suspected/confirmed clinical phenotype description (HPO).	Dag
3	Genome + phenotype	Vcf (phased) Phenotypes	Query tools HPO phenotype	Legal Consent Data security Scalability Cost of storage Ethical – patient recognition		Morten
4	Genotype (full genomes) and patient phenotype database	Database of genomes and associated phenotypes = variants in context.	Simplest vcf → phased vcf. Phenotype	Legal Consent Data security Scalability Cost of storage Ethical – patient recognition	Medium-long term	Dag
6	"Everything" – FastQ files with phenotypes			Legal Consent Data security Scalability Cost of storage Ethical – patient recognition	Raw data: comparison also of false negative rates. Possible to build query tools that allow extraction of needed information without breaching privacy? Will take time to get real quality phenotype; need to change clinical practice.	Dag

12 CONCLUSIONS AND NEXT STEPS FOR COOPERATION

During the workshop, the clinical variant pipelines in the three laboratories were mapped to identify commonalities and differences, and areas where standardization and harmonization could be beneficial.

The workshop participants discussed specific sets of data which would be valuable to have access to. The sets of data identified were related both to the genetic variants and to the variant interpretation pipeline, including sharing experiences on gene panels, operating procedures, databases and tool development. It was recognized that further work is needed on understanding prerequisites for exchange of data, such as standardization of process,harmonization of accept criteria and further clarifications of legal aspects. A technical benchmarking of the variant interpretation pipelines was agreed as an efficient exercise for understanding impact of set-ups on final outcomes. The workshop participants agreed to start benchmarking on selected aspects of the mapped processes to start sharing and learning.

Discussion on access to data related to variants ranged from population variant frequencies, databases of curated classified variants and linked information on patients' genotype and phentotype to full access to FastQ files and patient phenotype descriptions. While the participants ultimately would like to have full access to databases, this may be more complex to achieve technically and with respect to societal accept. Sharing of population variant frequencies was agreed as a possible first step, while exploring the legal basis and barriers for sharing data and establishing a common database.

To continue the work, appropriate actions and responsibilities for follow-up were agreed as summarized in Table 18. The workshop was agreed to be a first step towards sharing of genetic information and information related to the genetic variant interpretation pipelines between the participating laboratories, and a follow-up workshop was scheduled for November 2016 to summarize achievements and agree on further work.

Table 18 Agreed next steps for cooperation including identified actions and action responsibles

Topic: Organization of cooperation	**Action responsible**	**Notes**
Establish platform for sharing of information	SciLifeLab / Måns	- Github account? - Send information email to all participants
Funding	OUS / Dag	- Explore opportunities for funding
Stakeholder involvement	All	- Open approach; units at liberty to invite additional participants from own countries
Topic: Organization of follow-up workshop	**Action responsible**	**Notes**
Practical organization - Find date in November (doodle) - Practical planning of workshop II	DNV GL / Guro - connect with Anna	- November in Stockholm
Workshop agenda planning - coordinate input from all on agenda	DNV GL / Guro	- legal, data sharing, benchmarking - Separate work streams on bioinformatic and legal issues
Topic: Legal	**Action responsible**	**Notes**
Understand legal basis and barriers for sharing data and establishing a common database	DNV GL / Guro (agreed with Dag)	- initiate and coordinate with OUS, Karolinska/SciLifeLab (Anna, Valtteri) and Rigshospitalet (Karin, Morten)
Topic: Data / information sharing **Goal:** - **First step: Establish basic level of**	**Action responsible**	**Notes**

sharing - **Vision: Nordic sharing / database**		
Sharing of gene panels	SciLifeLab / Måns	- Establish basis in repository
Share procedural docs	OUS / Morten	- Standard operating procedures - Sweden: Rank score definition files
Explore applicability of existing tools	OUS / Tony?	- Do existing tools serve the need for exchange of information in a curated database? o ClinVar o Matchmaker / Beacon - Check legal?
Frequency database	OUS / Svein Tore	- Variant frequencies / counting - Include research data? - Check legal?
Topic: Benchmarking	**Action responsible**	**Notes**
Technical benchmarking exercise	SciLifeLab / Valtteri	- Easy first step to get started - Scope of benchmarking to be defined - Clinical genomics focus - Variant identification - Variant calling - Variant interpretation - Testing of 25-30 variants? - Benchmarking outcome to be presented at Workshop II.
Topic: Tools	**Action responsible**	**Notes**
Co-development of tools	SciLifeLab / Måns?	- To be followed up in Workshop II – separate stream for bioinformatics
Topic: Standardization	**Action responsible**	**Notes**
Concept document	DNV GL / Guro	- prepare draft, circulate
Process description	DNV GL / Guro	- Circulate for feedback - Process risk assessment → QC, QA, validation - Identify common standards / guideline references / etc. - What part of the process needs harmonizing to facilitate data sharing (based on data sharing exercise)

13 OTHER ISSUES (PARKING LOT)

During the workshop there were discussions related to the below topics, which were not concluded on as they were not part of the workshop scope.

Criteria for selection of in silico gene panels: selection of gene panels and of extent of panels upstream / downstream of gene.

Handling of incidental findings: Discussion related to handling of incidental findings in research/clinical setting and relevance when applying selected panel approach. Reporting of incidental findings not regarded ethical or relevant in an acute setting.

Mosaicism in WGS / WES: "In genetics, a mosaic or mosaicism denotes the presence of two or more populations of cells with different genotypes in one individual who has developed from a single fertilized egg."[70] This can be a real and relatively frequent problem in WGS for some genes with many dominant de novo mutations. Mosaicism can look like reading errors. E.g.: Child de novo, very few reads. Maybe

[70] https://en.wikipedia.org/wiki/Mosaic_(genetics)

run tissue biopsies on parents to be more certain. De Novo findings are usually positive compared to findings from parents. There is a risk of filtering it out.

Reporting: Ethical discussion on reporting; in Denmark incidental findings of class 3 BRCA variants would not be reported back, while in Norway class 3 BRCA variants would be reported and patient called in for counselling. This is a recent development in Norway; if something is in the journal, it should also be available to the patient. Would it be useful to harmonize reporting requirements?

14 APPENDIXES

14.1 List of Participants

Table 19 Genomics data sharing workshop – list of participants

Institution	First name	Last name
SciLifeLab	Måns	Magnussen
SciLifeLab	Valtteri	Wirta
Rigshospitalet	Karin	Wadt
Rigshospitalet	Morten	Dunø
OUS	Svein Tore	Seljebotn
OUS	Tony	Håndstad
OUS	Yngve	Sejersted
OUS	Dag Erik	Undlien
OUS	Eidi	Nafstad
OUS	Knut Erik	Berge
OUS	Morten	Eike
Karolinska University Hospital	Henrik	Stranneheim
Karolinska University Hospital	Nicole	Lesko
Karolinska University Hospital	Anna	Wedell
DNV GL	Arild Braathen	Torjusen
DNV GL	Erik	Stensrud
DNV GL	Guro Meldre	Pedersen
DNV GL	Sharmini	Alagaratnam
DNV GL	Stephen	McAdam
DNV GL	Vibeke Binz	Vallevik

14.2 Genomics data sharing workshop – agenda

Clinical Genomics Data sharing workshop

Time: May 30-31

Place: Meeting room Parken, DNV GL offices / Veritas 2 upper reception,

Veritasveien 1, 1363 Høvik, Norway.

Participants:

- Department of Medical Genetics, Oslo University Hospital, Oslo, Norway.
- Department of Clinical Genetics at The Juliane Marie Centre, Copenhagen University Hospital and the University of Copenhagen, Denmark.
- Clinical Genomics facility, SciLifeLab, Karolinska Institutet, Stockholm, Sweden.
- DNV GL

Meeting Objectives:

- Review current clinical variant pipelines in the three laboratories; discuss common challenges and identify areas where standardisation/harmonisation could be beneficial.
- Identify what specific data would be valuable for laboratories to be able to share in the short, medium and long term as well as current technical, legal and ethical barriers that hinder sharing this data today.
- Discuss potential models for future cooperation and agree on next steps

Time	Topic	Who
Day 1. Meeting Room: Parken		
11:45	Sandwiches available in meeting room	All
12:00	Intro & practicalities	DNV GL / Vibeke B Vallevik
12.05	Welcome, review of objectives and agenda, round the table introduction	OUS / Dag Undlien
12.30	DNV GL – Who are we and why are we interested in building trust in genomics?	DNV GL / Stephen McAdam
12.40	Short overview of OUS Clinical Genetics Dept	OUS
12:50	Short overview of CUH Clinical Genetics Dept	CUH
13:00	Short overview of SciLifeLab Clinical Genomics Unit	SLL
3:10	Break	
13:20	Mapping of WES / WGS variant calling pipeline - Consent - Sample taking - Raw data generation - Sequence alignment and variant calling	All

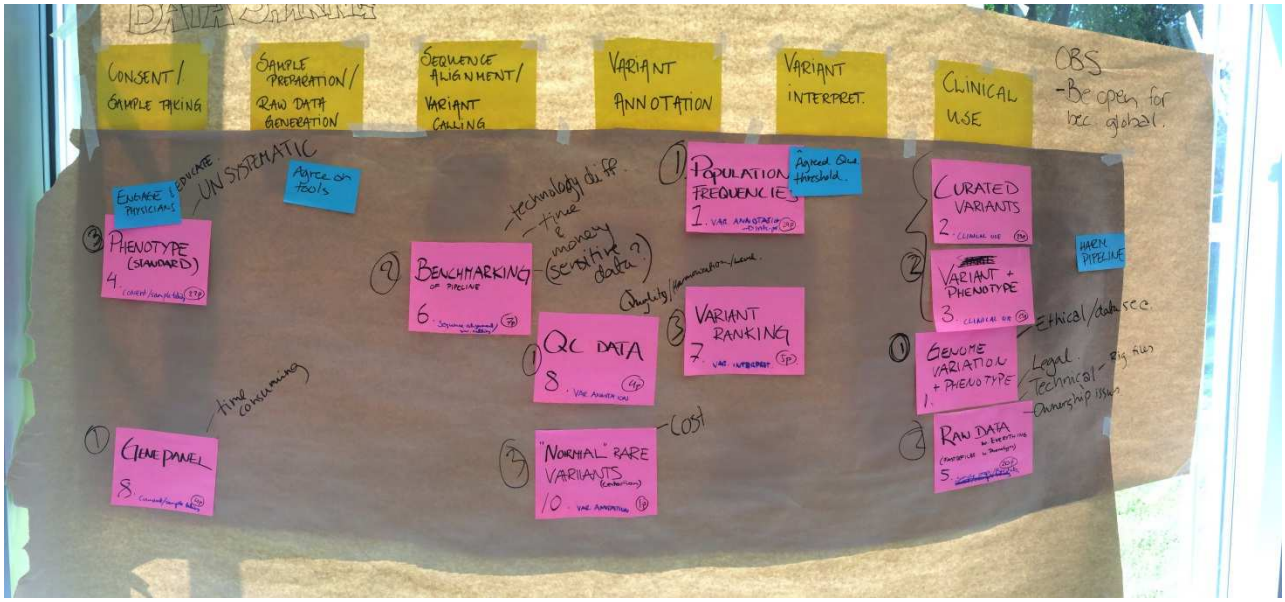
	- Variant annotation	
15:50	Break	
16.00	Sharing of data - What to share - Functional requirements - Short, medium and long term barriers	Brainstorming
17.00	30 min review of the day	All
19.00	Dinner	

Time	Topic	Comment
Day 2. Meeting Room: Parken		
08:30	Standardisation/harmonisation of variant calling and clinical variant classification pipelines – what and how?	All
10.00	Models for future cooperation – discussion - Consensus building - Other relevant initiatives - Funding needs/opportunities?	All
11.00	Next steps	All
11:30	Lunch for those who can stay	All

14.3 Workshop documentation photos

14.3.1 Exome / genome – clinical process

14.3.2 Data sharing



14.3.3 Next steps