THE HUMAN VARIOME PROJECT

13th International Symposium on Mutations in the Genome: detection, genome sequencing & interpretation

Holiday Inn Leiden
Leiden, The Netherlands

27th – 30th April 2015

ABSTRACTS

Scientific Organizing Committee

Johan T. den Dunnen, Chair (Nederland)
Richard G. H. Cotton, (Australia)
Ivo Gut (Spain)
Mats Nilsson (Sweden)
Stefan White (Nederland)

Organizing Secretariat

Meeting Makers
www.meeting-makers.com

Monday 27th April

SESSION 1

New technologies applied in the diagnostic laboratory

Joris Vermeesch

Lab. For Cytogenetics & Genome Research, KU Leuven, Leuven, Belgium
Joris.Vermeesch@uzleuven.be

Massive parallel sequencing is transforming genetic diagnostic laboratories. Following an introduction about available high throughput sequencing and array technologies, I will elaborate on how those technologies are transforming preimplantation, prenatal, postnatal and onco-haematological diagnostic testing. I will elaborate on the approaches followed towards introduction into the « routine » diagnostic testing Finally, I will embark on current gaps and challenges for genome wide analysis and potential solutions.

Clinical Exome Sequencing at a Large Academic Medical Center: Diagnostic Yield, Variant Spectrum, and Lessons Learned

Wayne W. Grody, Stanley F. Nelson, Joshua L. Deignan, Naghmeh Dorrani, Negar Ghahramani, Jianling Ji, Rena *Xian*, Sibel Kantarci, Fabiola Quintero-Rivera, Kingshuk Das, Michelle Fox, Eric Vilain, Sam Strom and Hane Lee

Los Angeles, California, USA

Our center has been performing clinical-grade whole-exome sequencing (WES) for the diagnosis of rare Mendelian disorders since January 2012. We report here our experience and important lessons learned from the first sequential 1000 cases tested. We explore indications for test ordering, diagnostic success rates, and practical issues in clinical implementation.

Our WES test was validated and performed under Clinical Laboratory Improvement Amendments (CLIA) regulations and College of American Pathologists (CAP) accreditation as a single comprehensive test from DNA extraction to result reporting. The cases were ascertained between January 17, 2012 and December 31, 2014. Whenever possible, parental DNA was collected and analyzed alongside that of the proband. All work was performed within the University of California, Los Angeles (UCLA) Clinical Genomics Center. The study was approved by the UCLA institutional review board. Exon capture was performed using the SureSelect Human All-Exon V2 Kit (Agilent Technologies) and sequencing was performed using the HiSeq 2000 for a 50-bp paired-end run or HiSeq 2500 for a 100-bp paired-end run (both from Illumina). Our mean coverage was 100X.

The most frequent clinical indication for WES was developmental delay or intellectual disability, usually as a component of a complex syndrome including seizures, hypotonia or dysmorphic features. Despite the nonspecificity of such presentations, mutations in *KMT2A, ZEB2, DYRK1A* and *SCN2A* were found to be causative in multiple independent cases, suggesting these genes are more commonly altered in individuals with syndromic developmental delay/intellectual disability than previously recognized. Other indications included ataxia and neurodegenerative conditions, cardiac anomalies, immunodeficiencies, and familial cancers. Definitive pathogenic variants were identified and reported in 27% of all cases, and likely pathogenic variants were detected in an additional 28% of cases, producing an aggregate diagnostic yield of up to 55%. Approximately 60% of the diagnostic findings were autosomal dominant mutations, most of which were *de novo* (when this could be deduced in the trio cases). About 30% were due to homozygous or compound heterozygous recessive alleles, 4% were X-linked, and 2% were copy-number variants or uniparental disomy. Overall about 4% of our diagnostic findings were pathogenic or likely pathogenic variants in novel disease genes not previously described.

Conclusions: We have shown that clinical exome sequencing can produce a diagnostic yield much higher than more routine (e.g., single-gene or gene-panel) genetic tests, especially when sufficient effort is expended to assure high sequencing coverage and quality and to obtain parental DNA samples when available; the latter enables ascertainment of *de novo* variants and correct phase of compound recessive variants. We propose that sufficient evidence now exists for WES to move up to become a first- or second-tier test for patients with undiagnosed syndromic conditions, rather than as a last resort after many other expensive tests have failed.

PacBio sequencing: improving mutation detection in complex genomic regions

Seyed Yahya Anvar[1,2,*], Michael Liem[2], Daniel Borras[3], Henk PJ Buermans[1,2], Heleen M van der Klift[1,4], Juul Wijnen[1,4], Monique Losekoot[4], Dorien JM Peters[1], Johan den Dunnen[1,2,4]

[1] Department of Human Genetics, [2] Leiden Genome Technology Center, [4] Department of Clinical Genetics, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands. [3] ServiceXS B.V., 2333 BZ, Leiden, The Netherlands
Corresponding author: s.y.anvar@lumc.nl

The emergence of genomic technologies has led to unprecedented amount of insights into the role of genetic variants in human biology. Today, Whole-genome (WGS) and whole-exome (WES) sequencing are transforming clinical diagnostics and are elucidating the genetic bases of diseases and clinically relevant traits much faster and more efficiently than before. Albeit the human genome is arguably the most well-characterized mammalian genome, a substantial part of it remains inaccessible or difficult to resolve due to gaps, highly repetitive sequences, large segmental duplications (SD), copy-number variations (CNVs), and excessive allelic diversity. Yet, there is ample evidence for significant biological and clinical relevance of these regions where annotation and variations are poorly characterized. The performance of NGS technologies is limited by short read-length (typically ≤250bp) and inherent biases to reliably resolve genomic regions with high homology. These regions are important sources of false-negative and false-positive variant calls in WGS and WES experiments. Thus, single-molecule long-read sequencing (SMLR-Seq) technologies (Pacific Biosciences; Oxford Nanopore) have great potentials for resolving inaccessible genomic regions and phasing of observed variants. Here, we show that sequencing three genes (*PMS2*, *PKD1*, *CYP2D6*) that are located in complex and repetitive regions of the genome allows a more accurate identification of genetic variants. Long read-range enabled an improved design of PCR-fragments, which subsequently allows for unique alignment of sequencing reads to targeted regions. For *PMS2*, next to SNVs without WES coverage, all SNVs that were called in WES were also identified in PacBio data. But, many false-negative SNVs were found in the WES data that resulted from alignment ambiguity in repetitive regions of *PMS2*. For *PKD1*, we identified a damaging mutation in a region that was not Sanger sequenced. Moreover, SMLR-Seq exposed false-negative and false-positive calls in other datasets and allowed a more reliable interpretation. We also showed that de novo reconstruction of *CYP2D6* alleles is of great importance for reliable haplotyping since different alleles are associated with different capacities for drug metabolism. These strategies not only improve our understanding of the functional impact and organisation of complex regions of the human genome but will also provide a robust framework for improved disease diagnostics.

SESSION 2

Multiplex targeted long amplicon sequencing method for Cyp2d6 genotyping using the PacBio RSII

Henk Buermans[1], Tahar van der Straaten[2], Rolf Vossen[1], Yahya Anvar[1], Jesse Swen[2], Johan den Dunnen[1]

Leiden University Medical Center
1 Department of Human Genetics; Leiden Genome Technology Center
2 Department of Clinical Pharmacy & Toxicology Leiden, the Netherlands
corresponding author: h.buermans@lumc.nl

The Cytochrome P450 2D6 enzyme encoded by the *Cyp2d6* gene, is among the most important enzymes involved in the metabolism of prescription drugs. Specific variants in the gene are associated with variations in the activity and amount of the Cyp2D6 enzyme between individuals. Different technologies exist to determine these variants, like the AmpliChip CYP450 GeneChip (Roche), Taqman qPCR or Second Generation Sequencing. However, genotyping of *Cyp2d6* with these routinely used methods is hindered because of the high sequence homology between *Cyp* genes, specifically with *Cyp2d7*. In addition, information on which variants reside on the separate alleles cannot be determined with these assays.

Targeted long amplicon sequencing using the PacBio RSII third generation sequencing platform holds several advantages over second generation NGS systems, including the ability to sequence multi-kb amplicons without the need for fragmentation steps, obtain high accuracy consensus sequences (QV50) and delivering fully phased variant information for separate alleles. For optimal implementation of PacBio sequencing for variant profiling, efficient sample barcoding strategies are needed. Standard PacBio methods introduce sample barcodes either via ligation of barcoded SMRTbell adapters to amplicons, or via PCR, by using a set of fusion primers linking the barcode sequence directly to the

locus specific primer. Downside of these barcoding methods is that they are expensive and lack flexibility.

We have setup a new versatile and cost effective PCR based multiplexing strategy for long amplicon variant profiling using the PacBio RSII platform. When applied for *Cyp2d6* variant profiling, for each individual the ~6.6 kb gene is first amplified with a pair of gene-specific primers with forward and reverse M13 sequence tails. A sample barcode is subsequently introduced in a second PCR using a set of re-usable M13-tailed barcode primers. Barcoded samples are then pooled in equimolar amounts and processed for PacBio sequencing. Using this setup, we sequenced the *Cyp2d6* gene for 12 individuals with previously established *Cyp2d6* genotypes as determined by golden standard Roche's Amplichip. Per SMRT cell 4 samples were multiplexed. Full length *Cyp2d6* sequences were obtained for all individuals after sequencing, barcode demultiplexing and processing the data with the Long Amplicon Analysis software (PacBio SMRT Analysis portal software v2.3). Moreover, for four individuals, two different allele sequences were evident from the data, indicating the exact distribution of multiple hetrozygous SNPs over the two separate *Cyp2d6* alleles. Predicted genotypes from the PacBio RSII were in agreement with those obtained with the Roche AmpliChip assay.

In conclusion, we have setup a new, versatile and simple to use multiplexing method for targeted long amplicon sequencing on the PacBio RSII and have successfully applied it to obtain fully phased allele sequences for the *Cyp2d6* gene. Moreover, with minor modifications this method can in principle be applied for targeted long amplicon sequencing of other gene panels.

Unravelling the origin of variants identified which may or may not be part of a patient's genotype

Belinda Chong, Caitlin Barns-Jenkins, Sarah-Jane Pantaleo, John–Paul Plazzer and Desirée du Sart

Molecular Genetics Laboratory, Victorian Clinical Genetics Services, Murdoch Childrens Research Institute, Parkville, Victoria, Australia, 3052

Catecholaminergic Polymorphic Ventricular Tachycardia (CPVT) is a rhythm disorder of the ventricles of the heart that occurs in genetically predisposed individuals. CPVT can be caused by mutations in *CASQ2*, *CALM1*, *KCNJ2*, *RyR2*, and *TRDN* genes. A 6 year old patient underwent genetic testing for a clinical presentation of CPVT with life threatening arrhythmias. A DNA sample extracted from EDTA blood was received and NGS arrhythmia panel testing was performed, which consists of a total of 28 genes, including the genes listed above for CPVT. NGS cardiac testing in our laboratory is carried out using an Agilent SureSelect custom capture design. Sequencing is performed either on the MiSeq or HiSeq Illumina instruments. To complete coverage and to confirm variants identified; all uncovered regions and any variants classified as clinically actionable are Sanger sequenced.

The following variants were identified in the NGS analysis: A likely pathogenic missense mutation in the *KCNH2* gene, NM_000238.3(KCNH2):c.442C>T, a variant of unknown significance (VUS) in the *KCNE1L* gene, NM_012282.2(KCNE1L):c.206_208delTCT and common variants in AKAP9, CACNA1C and CASQ2. No other variants were identified in the other arrhythmia genes. For the likely pathogenic variant, the depth of coverage was 436, the quality score was 77, and the heterozygous call rate was 30% to 70%. The depth of coverage was 561, the quality score was 217 and the heterozygous call rate was about 50% for the VUS. A number of issues were raised with these result: (i) the genes containing the variants were not consistent with clinical presentation in the patient; (ii) the likely pathogenic mutation was not confirmed using Sanger sequencing.
A new sample was requested and the NGS analysis was repeated on the new sample to determine the significance of the variants identified by NGS analysis on the first sample. The following variants were identified in the second NGS analysis: A likely pathogenic missense mutation in the *RYR2* gene, NM_001035(RYR2):c. 12017 C>T, two variants of unknown significance (VUS) - one in *KCNE1L*, NM_012282.2(KCNE1L):c.206_208delTCT and the other in SCN5A, NM_198056(SCN5A):c. 892 G>A and common variant in CASQ2 and RYR2. The depth of coverage was ranged from 460 to 1182, the quality score was 225 and the heterozygous call rate was about 50% for all the variants identified. These results were consistent with clinical presentation in the patient and all variants were confirmed by Sanger sequencing.

The laboratory process to solve this "mystery" of the origin of the variants identified in the first NGS analysis will be presented. We will discuss the critical value of the key parameters to be assessed for identifying true variants present in a sample, namely coverage, quality score and variant call rate as well as the clinical presentation in the patient.

Validating Bioinformatic Pipelines for the Clinic

Kenneth Doig1,3, Andrew Fellowes2, Jason Ellul1, Jason Li1, Anthony Bell2, Stephen Fox1,3

1 Research Division, Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia
2 Molecular Pathology Laboratory, Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia
3 Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC, Australia

Increasingly affordable sequencing platforms has led to their wide spread adoption beyond research groups into clinical pathology. The introduction of desktop sequencers into the clinic has required many institutions to construct ad-hoc bioinformatic pipelines that can process locally generated data. These pipelines refine the voluminous data generated by next generation sequencing (NGS) platforms and transform raw sequencing reads into meaningful biological data suitable for clinical reporting or research analysis. However, there is little consensus across labs on how this should be done and indeed, the vast range of pipeline software components with varying features means there is unlikely to be any standardisation in the near future.

Todays NGS pipelines are typically built as a concatenation of various open source and research derived tools. These tools are often built by researchers, (rather than software engineers), as a lab-centric solution to a local problem. As such, they often share some or all of the following characteristics: not robust with respect to novel input data, poorly supported, poorly documented and often unmaintainable for a production environment.

To mitigate the shortcomings of existing pipelines, we have developed PipeCleaner, a testing framework for pipeline validation. PipeCleaner generates known input reads for a pipeline, then repeatedly runs the pipeline under test and finally collects the actual variants found by the pipeline and compares them to the expected variants in the input reads. Given a VCF file, PipeCleaner generates either synthetic NGS reads or manipulates existing FASTQ files with known variants allowing a pipeline's results to be analytically tested for correctness, false positives and false negatives. Hence the specificity and sensitivity of a pipeline can be quantified independently of any biological variation. The separation of technical errors from biological variation is critical for defining the error properties of an NGS pipeline.

PipeCleaner is driven by a configuration file written in a rich domain specific language (DSL) defining all aspects of the test environment. This language embeds a interpreter giving the tester the power to create many complex scenarios to exercise the pipeline under test. Currently the user can define the error properties of the reads, the variants embedded in the reads (SNPs and/or indels) and the simulated sample cell heterogeneity. The latter is key within a cancer context to effectively simulate tumour samples that typically contain a mix of tumour and normal cells. The structure of the DSL allows the development of pipeline test suites encompassing regression testing, performance testing, conformance testing and exhaustive coverage testing. We have used these test suites to validate and characterise pipelines for clinical amplicon pipelines and large research development panels. PipeCleaner and its suite of tests provide an objective test framework for evaluating pipeline correctness, performance and behaviour.

Keywords: regression testing, NGS, high throughput sequencing, pipeline validation, domain specific language

Combination of MLPA and Illumina MiSeq Amplicon Sequencing provides maximum mutation detection coverage for TSC1/TSC2 gene analyses in patients with Tuberous Sclerosis Complex

[1,3]Teguh Haryo Sasongko, [1,3]Nur Farrah Dila Ismail, [1,3]Nik Mohd Ariff Nik Abdul Malik, [2]Salmi Abdul Razak, [4]Narazah Mohd. Yusoff, [2,3]Zabidi Azhar Mohd. Hussin

[1]Human Genome Center, [2]Department of Pediatrics, School of Medical Sciences, Universiti Sains Malaysia, USM Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia
[3]Center for Neuroscience Services and Research, Universiti Sains Malaysia, USM Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia
[4]Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, Pulau Pinang, Malaysia
Corresponding author : Teguh Haryo Sasongko (tghsasongko@gmail.com, teguhhs@usm.my)

Mutation detection in large genes where there is no hotspots or common type of mutations, such as in TSC1 and TSC2 is always challenging. Here we show our findings on TSC1 and TSC2 mutation analyses in patients with Tuberous Sclerosis Complex (TSC) using a combination of two approaches.

Thirty-seven Malaysian patients diagnosed with TSC were referred to our molecular genetic laboratory in Human Genome Center, Universiti Sains Malaysia. Clinical diagnosis was based on the 2012 revised consensus criteria for TSC [1]. Informed consent was taken prior to blood taking and the study was approved by the USM Human Research Ethics Committee. Genomic DNA was extracted from whole blood. We initially used MLPA (MRC-Holland) to detect large copy number changes. We subsequently employed Amplicon Sequencing using Illumina MiSeq to detect small mutations. For the amplicon sequencing, long-range PCR for 10 amplicons was done to cover the whole genomic portion of TSC1 (4 amplicons) and TSC2 (6 amplicons) with sizes range from 3 to 11 kilobases. Pathogenicity of the missense mutations were determined through LOVD v3.0 [2] or Polyphen 2.0 [3] analysis.

TSC1 mutations were found in 5 patients (13%), TSC2 in 24 patients (65%) and the rest 8 patients (22%) show no mutations in both genes. We identified 3 mutations in TSC1 (all nonsense) and 22 mutations in TSC2 (8 nonsense, 5 missense, 1 splice site, 5 small insertion/deletion and 3 multiple exon deletions). Half of the mutations we found are novel. Total procedure running time for MLPA was 3 days, with reagent cost up to MYR 100/patient. Total procedure running time for Amplicon Sequencing was 14 days, with reagent cost up to MYR 511/patient.

We have shown that combination of MLPA and Illumina MiSeq amplicon sequencing provide maximum detection coverage for TSC1/TSC2 mutation analyses in patients with TSC. In addition, the amount of laboratory works shown here suggested application of the approaches into clinical molecular diagnostics with reasonable cost and turnaround time.

Acknowledgement
This study was funded by the USM Research University grant no. 1001/PPSP/812048 for T.H.S. and USM Neuroscience Excellence grant no. 304.PPSP.652205.K134 for Z.A.M.H.

References
Northrup H, Krueger DA. Tuberous sclerosis complex diagnostic criteria update: recommendations of the 2012 international tuberous sclerosis complex consensus conference. Pediatr Neurol 2013;49:243–54.
Leiden Open Variation Databases (LOVD) v3.0 for TSC1 and TSC genes; http://chromium.liacs.nl/LOVD2/TSC/home.php?action=switch_db
Polyphen 2.0 (http://genetics.bwh.harvard.edu/pph2/)

COMPANY LECTURE

MRC HOLLAND

MLPA 2.0 - Multiplex Ligation-dependent Probe Amplification on Illumina NGS platforms using a 600+ probe assay

Jan Schouten

MRC-Holland

MLPA, the multiplex PCR-based technique that has now become the standard in copy number variation detection, was developed by MRC-Holland in 2002. Although the method is used worldwide for the detection of copy number and methylation changes in human DNA, traditional MLPA has its limitations: an MLPA assay can contain a maximum of 60 probes and requires a minimum of 20-50 ng sample DNA of good quality.

We now introduce MLPA 2.0: a NGS-based MLPA variant enabling the use of 600 probes in a single reaction, with less stringent requirements regarding both sample DNA quantity and quality. This newly developed MLPA variant enables the creation of assays that screen a much larger proportion of the human genome. MLPA 2.0 assays can be used on all Illumina NGS platforms. As NGS-based sequence analysis still has its limitations when it comes to reliable detection of copy number variants, MLPA2.0 fulfills a yet unmet need.

In this presentation, we show the results of an MLPA 2.0 assay designed to complement sequence analysis for finding the cause of a hereditary predisposition to cancer. The assay contains probes for each exon of 26 genes involved in breast, colon, gastric and prostate cancer and melanomas, including BRCA1, BRCA2, MLH1 and MSH2.
Other MLPA 2.0 products that are currently in development include assays for newborn screening and the analysis of tumour-derived DNA.

RAPID FIRE POSTER PRESENTATIONS

RF1

Development of an universal variant classification system and reporting scheme for routine NGS panel diagnostics

Andreas Laner, Anna Benet-Pagés, Melanie Locher, Ulrike Schön, Veronika Mayer, Elke Holinski-Feder

MGZ - Medizinisch Genetisches Zentrum München, Bayerstr. 3-5, 80335 München, Germany
corresponding author: laner@mgz-muenchen.de

NGS is widely used in routine molecular-genetic diagnostics and many laboratories have now access to high quality NGS data resulting in a reliable variant detection. Guidelines [1] are currently formulated to standardize and harmonize many aspects of NGS testing with emphasis on technical issues, nevertheless considerable inter- and intra-laboratory discrepancy concerning classification and reporting of "variants of uncertain clinical significance" (VUS) can be observed.

There are several variant classification systems published. Some are restricted to specific genes or phenotypes (e.g. IARC [2]; HNPCC genes only), some are restricted to specific modes of inheritance (e.g. Ambry [3]; autosomal dominant and X-linked only), and most of them require extensive additional information (IARC, Ambry and Emory [4]; e.g. segregation data, functional assays, RNA data, immunohistochemical data, etc.), often not available at time of report creation. Therefore we aimed to formulate an universal classification algorithm based in structure and content mainly on the IARC and Emory classification schemes, allowing a robust and simple variant classification for a routine diagnostic lab.

Furthermore we created an algorithm for classification of variants possibly affecting RNA splicing, based on published data [5], [6], since these type of variants - although quite common - are not covered by above mentioned systems.

As the number of genes analyzed in a NGS panel rises and/or the available information about a patients clinical phenotype declines, the number of "VUS" inevitably increases. To address this problem, we divided our NGS diagnostic in three distinct groups (Type A-C), as suggested in the draft version of the "Guidelines for diagnostic next generation sequencing" [1]. Depending on the mode of inheritance, clinical information and/or genetic information VUS are reported differently in these three categories.

[1] http://www.eurogentest.org
[2] Plon et al., Hum Mutat. 2008
[3] http://www.ambrygen.com/variantclassification
[4] http://geneticslab.emory.edu/emvclass/EGLClassificationDefinitions.php
[5] Houdayer et al.; Hum Mutat. 2012
[6] Jian et al.; Nucl Acids Res. 2014

RF2

Analysis of circulating cell free DNA (ccfDNA): A promising tool for personalized medicine and cancer therapy

Keup C[1], Mardin W[2], Dworniczak J[3], Dockhorn B[4], Haier J[5], Rijcken E[2], Dworniczak B[1]

1 University Hospital Muenster, Institute for Human Genetics, Germany
2 University Hospital Muenster, Department of General and Visceral Surgery, Germany
3 Klinikum Rechts der Isar, Department of Radiology, Technical University Munich Germany
4 Center for Pathology, Kempten, Germany
5 University Hospital Muenster, Comprehensive Cancer Center, Germany
corresponding author: dwornic@uni-muenster.de

Although significant progress has been made in the development of new therapy approaches, cancer remains one of the leading causes of death worldwide. In most cases cancer remains undetected until its advanced stages because up to now efficient screening techniques for early detection are not still available.
However recently published data indicate that circulating cell-free DNA (ccfDNA) could become a promising biomarker in cancer diagnosis, therapy and prognosis. The use of ccfDNA presents several conceptual advantages compared to classic genetic analysis via tumor-tissue sampling. CcfDNA analysis is non-invasive and enables day-to-day patient follow-up and monitoring of treatment response. Analysis of ccfDNA also allows detection of genetic and epigenetic alterations within the tumor. Careful analysis of these alterations could provide valuable information to tailor the clinician's choice of treatment.

Despite the fact that ccfDNA is known since long time and despite urgent need of secure biomarker in cancer therapy analysis of ccfDNA in the clinics is far from reality. This is mainly due to reported discrepancies and contradictory data on the analysis certainly caused by lack of normalization of the experimental conditions. We therefore started a pilot study with patients suffering from colorectal cancer in order to establish analysis of ccfDNA in our routine laboratory. Optimization and normalization of the Workflow of the pilot study Covers all aspects of the complete procedure: starting with blood sampling, isolation of the ccfDNA, determination of its concentration and determination of tumor-derived ccfDNA part and its fragmentation. Prior to analysis tumor derived DNA-fragments are enriched by cold-PCR and presence of sequence variants are either shown by next generation sequencing (NGS) or - if the mutation is known - by quantitative PCR, digital PCR and by NGS. To validate results DNA is isolated from respective tumor specimen and genes which are known to be frequently mutated in colon Cancer are sequenced by use of appropriate gene panels on Ion Torrent Personal Genome Machine (PJM) or Ion Proton. In our presentation we will show first data concerning the feasibility of the approach.

RF3

Splicing consequences of exonic variations

Christel Vaché[1], Zohor Azher[1], Michel Koenig[1,2], Mireille Claustres[1,2], Anne-Françoise Roux [1]

[1]*Laboratoire de génétique moléculaire, CHU Montpellier, France*
[2]*Université de Montpellier, France*
Corresponding author: vache.christel@inserm.fr

There is growing evidence that misclassification of exonic variants between splicing mutation or not can commonly occur leading to a wrong knowledge of mechanisms implicated in pathogenic phenotype. Indeed, if variations of the weakly conserved signals 5' and 3' splice sites and branch site are commonly predicted and classified as splicing mutations, exonic alterations are frequently only considered for their effects at protein level ignoring their potential effects on the splicing process.

Furthermore, even more and more databases include splice annotations for exonic variants derived from prediction tools, these are usually restricted to the creation of an ectopic splice site or to the strengthening of a cryptic site. As a consequence, exonic variations resulting in true alterations of ESEs, EESs or RNA secondary structures are under-estimated.

In either case, the effect on splicing of an exonic mutation is not systematically envisaged at RNA level and when it is, analyses of transcripts can be complex and interpretation challenging resulting, sometimes, in controversial results.

This challenge, which has a real impact for diagnosis and which is fundamental to the development of adapted therapeutic strategy, will be highlighted with examples of exonic variations implicated in diseases.

RF4

Validation of high-throughput mutation screening of pooled non-indexed samples in a case-control study

Therese Törngren(1), Hans Ehrencrona(2), Åke Borg(1) & Anders Kvist(1), on behalf of *SWE-BRCA*

(1) Oncology and Pathology, Department of Clinical Sciences, Lund University, 22185 Lund, Sweden
(2) Department of Clinical Genetics, Lund University, 22185 Lund, Sweden

Germline loss-of-function mutations in the *BRCA1* and *BRCA2* genes are associated with a high risk of breast cancer and explain about 20% of familial clustering of the disease. Rare variants in at least fifteen other genes confer a moderate to high risk. Until recently, only *BRCA1* and *BRCA2* were routinely screened for mutations in breast cancer families, but with next-generation sequencing extended screening is being introduced in many countries. The cancer risks associated with variants in the additional tested genes are often poorly known. Individually, risk variants are rare in the population, but together 25% of tested individuals may carry potential pathogenic variants in one of these genes. This represents a major issue for genetic counseling and there is an urgent need for additional knowledge.

The SWEA study is a national collaboration involving all cancer genetics clinics in Sweden. Within the SWEA study, index cases from breast cancer families are screened for mutations in 17 known breast

cancer susceptibility genes: *BRCA1, BRCA2, TP53, PTEN, STK11, CDH1, CDKN2A, CHEK2, PALB2, BRIP1, ATM, RAD50, RAD51C, RAD51D, BARD1, NBN* and *MRE11A*. In addition, we screen the protein coding exons of 47 candidate susceptibility genes. To estimate population allele frequencies and gain knowledge of risks associated with variants in these genes, we will perform mutation screening in 5000 healthy controls, 5000 consecutive breast cancer patients and 4000 familial high-risk cases. Pooling of non-indexed DNA samples makes screening of large cohorts possible at significantly lower cost and labor effort. To evaluate variant detection in pooled non-indexed DNA samples, 4 pools consisting of 16 individual samples each were tested and evaluated in this study.

DNA from 16 individuals was pooled in equimolar amounts. DNA from the genomic loci of the 17 known susceptibility genes plus near-gene regions and the coding exons of the additional 47 candidate genes were captured using hybrid selection and sequenced on an Illumina HiSeq 2500. Coding exons and nearby introns were covered to an average depth of 100 reads per individual sample. Freebayes was used for variant calling. The 64 samples studied harbored 13 known small insertions or deletions and multiple known substitution variants in the target genes.

All known variants were detected except one complex variant (NM_000059.3:c.10095delCinsGAATTATATCT). However, this variant is present in aligned reads and is seen when examining the alignments in Integrative Genome Viewer (IGV). The false positive rate in the initial analysis was high and there is a need to fine-tune variant filtering to achieve a better balance between sensitivity and specificity.

High-throughput mutation screening of pooled non-indexed samples is a time and cost-effective way of getting allele frequencies for new genes of interest. At an average read depth of 100 reads per individual we can detect variants at frequencies of 2-3% with good sensitivity.

RF5

Adding value to genetic variant databases by integration of social media functionality

John-Paul Plazzer

Department of Colorectal Medicine and Genetics, The Royal Melbourne Hospital, Melbourne, Australia
johnpaul@variome.org

Locus-Specific databases (LSDBs) exist to publicly list genetic variation and their associated phenotypic, experimental or clinical information. The main source of data is diagnostic laboratory submissions or published literature. However, there are significant disincentives for laboratories to submit data due to the time and effort required. Attempts to encourage submission have largely been based around recognition via publications or more recently with microattributions.
With increasing availability and through-put of current sequencing technology, ways of value-adding are required to make the submission of data to LSDBs worthwhile for submitters. These may include curation to ensure the quality of the information, classification of variants by expert review panels, or integration of data with international databases.

In 2014, InSiGHT added alongside variant classifications an option to allow users to *follow* variants of interest in the mismatch repair genes. This "one-click" function is simple to implement into an existing system, and provides an easy way for users to express clinical interest in a variant, and allowing opportunities for further clinical information to be submitted. To date, this has shown a promising take-up, with 300 variants from over 100 users now listed. Such a system may prove to be a vital component of the international sharing of variant information.

SESSION 3

Is exome sequencing of single patients with intellectual disability an effective diagnostic strategy? And what about whole genome sequencing?

Claudia Ruivenkamp1 , Mariëtte Hoffer1, Antoinet Gijsbers1, Sander Bollen1, Wibowo Arindrarto2, Marlies Laurense-Bik1, Ivo Fokkema3, Saskia Smit1,Michiel van der Wielen1, Emmelien Aten1, Emilia Bijlsma1, Martijn Breuning1, Yvonne Hilhorst-Hofstee1, Nicolette den Hollander1, Sarina Kant1, Arie van Haeringen1, Marjolein Kriek1, Jeroen Laros2,3, Bert Bakker1, Christi van Asperen1, and Gijs Santen1

1) Clinical Genetics, Leiden University Medical Center, Leiden, 2) Sequence Analysis Support Core, Leiden University Medical Center, Leiden, 3) Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; in collaboration with GenomeScan, Leiden, The Netherlands

Trio-sequencing can be used in all disorders, and has particularly proven its value in finding causes of intellectual disability (ID) or multiple congenital anomalies. In contrast, we investigated whether sequencing only the affected patient without parents is sufficient to find the causative mutation, leading to a considerable reduce in costs. In this study, we enrolled 36 patients with unexplained ID, and sequenced the exome. The exome sequences were analysed with a stringent post-sequencing annotation pipeline including an ID gene panel of ~500 genes for filtering of the data. All remaining variants with a potential clinical consequence were validated by Sanger sequencing and tested in the parents for inheritance.

After variant filtering we noticed an average of 13 variants per patient (range 2 to 27) requiring further clinical interpretation. The majority of these variants were inherited from one of the parents. Hitherto, we identified 5 *de novo* mutations and 1 homozygous mutation in 33 patients (18%). For the remaining 27 patients both parents have been sequenced. Further analysis of these trios is currently performed and these results will be presented, as well as results on whole genome sequencing.

Without exome sequencing the parents, a relatively high amount of potentially pathogenic variants remain. All these variants require clinical interpretation which is very time-consuming, while most of these variants were likely benign because they are inherited from one of the parents. With trio-analysis inherited variants can be filtered out suggesting that this strategy, at this moment, is more efficient in identifying the causative variant. In the future when databases are filled with more and more exome data and consequently with more rare benign variants, exome sequencing single patients will become a more realistic diagnostic approach. On the long term, whole genome sequencing will be the most cost-efficient approach. This however, will require a further reduction of sequencing costs.

A novel POLE variant, identified by exome sequencing, in a family with high burden of colorectal- and extra-colonic cancers

Maren F. Hansen1,2, Jostein Johansen1, Inga Bjørnevoll2, Anna E. Sylvander2, Kristin S. Steinsbekk1, Pål Sætrom1, Arne K. Sandvik1,2, Finn Drabløs1, Wenche Sjursen1,2.

1Norwegian University of Science and Technology, Trondheim, Norway.
2St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

A Norwegian family with high burden of colorectal adenomas and adenocarcinomas in addition to extra-colonic cancers was investigated for two decades in order to find the cause for their cancer predisposition. Due to the striking dominant inheritance in this family, we strongly suspected a highly penetrant variant as the cause of cancer predisposition. Several CRC predisposing genes were tested with Sanger sequencing during this period of time, however, no causative mutation was identified. We therefore performed exome sequencing to detect the cancer predisposing mutation in this family

All patient samples and clinical information were obtained with informed written consent and the study was approved by the Regional Committee for Medical and Health Research Ethics of Central Norway (approval 2012/1707). Exome capture was performed using SureSelectXT Human All Exon V5+UTRs. The libraries were sequenced on Illumina HiSeq2500 with 2x100 bp paired end sequencing. Exome sequencing data was aligned to the human genome (hg19, UCSC assembly, February 2009) using the Burrows-Wheeler-Aligner. PCR duplicates were removed with Picard-tools and BAM files were converted with SAMtools . Variant calling was done using GATK version 3.1. Variants were annotated

with ANNOVAR and subsequent filtering was done using the filtering tool FILTUS version 0.99-9.

We identified the novel POLE variant c.1373A>T (p.Tyr458Phe) as the most likely cause of cancer predisposition in this family. POLE and POLD1 encode the catalytic and proofreading subunits of DNA polymerase ε and δ enzyme complexes, respectively. Pathogenic germline mutations in these genes have recently been described to cause Polymerase proofreading-associated polyposis (PPAP). This is a highly penetrant, autosomal dominant syndrome predisposing to development of multiple adenomas and carcinomas. Tyr458 is a highly conserved residue located at the active site of POLE. Studies in microorganisms show increased mutation rate due to reduced exonuclease activity when the residue corresponding to Tyr458 is replaced with Phenylalanine. The POLE mutation segregates with disease and is associated with colorectal cancers and adenomas in addition to cancers of ovaries, small intestine and pancreas. We also observe a large phenotypic variation among the POLE mutation carriers which might be explained by modifying variants in other genes. In addition, we identified variants with potential functional effects which might explain some of the phenocopies observed in this family.

The POLE variant p.Tyr458Phe predisposes to colorectal adenomas and carcinomas in addition to extra colonic cancers.

This work was supported by grants from the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU).

Target Locus Amplification (TLA): A comprehensive new DNA test for detection of fusion genes in leukemia

Birgit Sikkema-Raddatz[1], Eva van den Berg[1], Max van Min [2], Petra Klous[2], Andre Mulder[3], Edo Vellenga[4], Pieter van der Vlies[1], Desiree Weening[1], Cleo van Diemen[1], Erik Splinter[2], Richard Sinke[1], Rolf Sijmons[1]

[1]University of Groningen, University Medical Center, Genetics, Groningen, Netherlands
[2]Cergentis, Utrecht, Netherlands
[3]University of Groningen,University Medical Center, Laboratory Medicin, Groningen, Netherlands
[4]University of Groningen, University Medical Center, Hematology, Groningen, Netherlands
Corresponding author: b.sikkema01@umcg.nl

Leukemia patients carry a wide range of chromosomal abnormalities which affect their prognosis and treatment options. Thus, a full and rapid diagnosis of these abnormalities is essential. At the moment, mostly a combination of tests is used including karyotyping, array, FISH, and/or RT-PCR. These techniques are challenging to perform and at times inadequate. Recently a new technique, Targeted Locus Amplification (TLA) has been developed enabling the detection of chromosomal abnormalities in leukemias[1] which can potentially replace all the current techniques. TLA is a unique DNA capturing technology followed by next generation sequencing (NGS). TLA enables the targeted amplification of up to 100kb surrounding one primer pair complimentary to a small sequence unique to the gene of interest. Therefore all genetic changes and gene-fusions should be detected regardless of the identity of the fusion partner or physical position of the gene-fusion. The aim of this study was to demonstrate the capability of TLA to detect fusion genes in a proof-of–principle study.

A TLA assay was developed for the KMT2A gene, which is involved in gene fusions in AML patients and has more than 50 translocation partners. We used four cell-lines positive for a KMT2A gene fusion, namely KMT2A-AF4, KMT2A -ELN, KMT2A -AF6, and KMT2A -AF9 gene fusion. TLA primers were designed on 5 positions across the KMT2A gene. DNA of the cells was cross-linked, digested and re-ligated. Amplifications were performed using all 5 TLA primers and were then sequenced on an Illumina MiSeq machine. Generated sequences were processed and mapped.

The fusion partner of the KMNT2A gene was correctly identified in all the samples and resulting break-point sequences were identified. TLA was also performed in different mixtures of gene-fusion positive and healthy cells to determine the sensitivity. With the current protocol and existing data-analysis tools a sensitivity of 5% was already established.

Our first results demonstrate the capability of TLA for detection of structural abnormalities. Based on this result we will develop diagnostic assays which will simultaneously detect all types of structural genetic abnormalities and other aberrations relevant to almost all types of leukemia.

[1] de Vree et al.2014.Nat Biotechnol. Aug 17.

Pathway analyses of whole genome sequence data identifies novel candidate Intellectual Disability genes

Farah Zahir1, 2, Leora Lee2, Nancy Makela2, Jan M. Friedman2, Marco A. Marra1,2

1. Canada's Michael Smith Genome Sciences Center, Vancouver, Canada
2. Department of Medical Genetics, University of British Columbia, Vancouver, Canada

Intellectual disability (ID) is the most frequent severe life-long handicap, with a global prevalence of 1-3% (World Health Organization). The cause remains unknown ~ 30% of cases despite undergoing a plethora of clinical testing. We conducted whole genome sequencing (WGS) on a carefully selected cohort of 8 trios – patients with idiopathic ID and morphological brain defects, and both normal parents. Patient pedigrees suggested autosomal dominant inheritance.

Genomes were sequenced on Illumina HiSeq (30X). BWA, SAMtools MpileUp, Bedtools and ANNOVAR were used for re-alignment (hg19), variant calling, selection of de novo variants and annotation, respectively. We initially employed a liberal filter- selecting as possible pathogenic variants all de novo rare (MAF of <0.01 in 1000G, NHLBI-ESP and local database of 1500 genomes) heterozygous coding variants. We found ~30 genes with at least one hit in each patient. We conducted pathway analyses of all candidate genes for enrichment in neurodevelopment (IGA, DAVID, Panther) and refined the list of candidate variants/genes for verification.

We independently verified de novo heterozygous coding SNVs in 6 genes in 5 patients: SPRY4 (non-synonymous SNV), CACNB3 (non-synonymous SNV), SQSTM1 and UPF1 (both in same patient), PHF6 (stop-gain), and ARID1B (frame-shift). Genotype-phenotype correlations for the known ID gene ARID1B confirmed pathogenicity. SQSTM1 and UPF1 each contain two adjacent cis mis-sense SNVs; in SQSTM1 they cause an early termination while in UPF1 they cause an amino-acid substitution. The variants in PHF6, SQSTM1, UPF1 and SPRY4 fit accepted criteria1 to be likely pathogenic while we are unsure of the pathogenicity of the CACNB3 variant.

In a subsequent validation study, we found that predicted damaging variants in the coding sequence of the SPRY4, CACNB3, SQSTM1 and UPF1 genes were each significantly enriched in 2081 patients (from the UK10K project) who presented with neurodevelopmental/neurofunctional phenotypes versus a control 2535 subjects (from the 1000G project). For PHF6, we found one loss of function variant in the test cohort versus no results in the control population, though this result did not reach statistical significance. The known ID gene, ARID1B validated as a positive control; we found a significant enrichment for predicted damaging variants in our test population versus control population. These data indicate the novel genes we identified are important for neurodevelopment. Further pathway analyses revealed that all 6 genes connected via one or two nodes to the ubiquitin proteasome degradation pathway, indicating the importance of this pathway in normal brain development.

Our informed approach to variant filtration has enabled the detection of novel candidate ID genes, and our data highlight an important pathway in neurodevelopment.

[1] Gilissen et al. Nature. Genome sequencing identifies major causes of severe intellectual disability. 2014 Jul 17;511(7509):344-7.

Tuesday 28th April

SESSION 4

RNA: the neglected molecule

Johan T. den Dunnen

Leiden Genome Technology Center (LGTC), Depts. of Human & Clinical Genetics, Leiden University Medical Center, Leiden, Nederland

In diagnostics the focus is on DNA; easy to isolate, store and analyse. When a variant is found in DNA, its consequences on protein level are predicted, evaluated and a classification is made: *disease-causing or not.* In reports regularly only the protein variant is reported, without hesitation and mentioning the actual DNA result, case closed. The tendency to skip the level in between, the RNA, makes RNA the "neglected" molecule. Even databases often do not even have a column to indicate whether RNA was analysed or not.

Working with the giant DMD gene, variants causing Duchenne and Becker Muscular Dystrophy, we started to appreciate the value of the RNA molecule from early on. When the choice is to check either a 79-exon 2.4 Mb gene or a 11.5 kb RNA molecule for

variants that are disease-causing the choice is rather simple. RNA helped us to identify, and intrinsically confirm, the presence of variants affecting splicing (coding and intronic) and to detect deletions/duplications which could be easily missed using DNA-based sequencing. It highlighted cases where the consequences predicted from DNA were clearly wrong, significantly changing the predicted outcome of the disease. Applied to other diseases, RNA analysis helped us to crack several diagnostic puzzles and surprised many times: one should not go blindly from DNA to protein.

RNA plays a prominent part in our research and diagnostics. Blood-derived RNA analysis is used to support data analysis from whole exome/whole genome sequencing; to study the potential consequences of variants on splicing, to reduce the number of variants of unknown significance and simply using the disturbances in the expression profile itself. Ribosome profiling was developed to study actively translated RNA, protein translation and uORF sequences. Pacific Biosciences long read single molecule sequencing is used to study RNA structure, in particular promoter usage and differential splicing. Not surprisingly, using exon skipping, we even try to modulate RNA transcription as a potential treatment for specific diseases.

Traditional vs. Next-Generation Panel Testing of Hereditary Breast and Ovarian Cancer Genes in a Large Clinical Population

Stephen Lincoln[1*], Allison Kurian[2], Andrea Desmond[3], Yuya Kobayashi[1], Shan Yang[1], James Ford[2], and Leif Ellisen[3]

[1]Invitae, San Francisco, USA, [2]Stanford University, Palo Alto, USA, [3]Massachusetts General Hospital, Boston, USA
Corresponding author: steve.lincoln@me.com

Next-generation sequencing (NGS) of gene panels has gained clinical acceptance although questions remain about these tests. Expanding on our recently published work [1] we considered whether NGS can both replace and supplement traditional genetic tests for hereditary breast/ovarian cancer (HBOC). Specifically we evaluated whether the broad spectrum of variants detected by traditional methods (e.g. Sanger, qPCR, MLPA, arrays) can also be detected by NGS. We compared BRCA1/2 variant interpretations produced using only publicly available resources to interpretations produced previously by an independent laboratory using their large proprietary database. Finally, we examined the clinical relevance and potential actionability of the non-BRCA1/2 results.

1062 patients indicated for BRCA testing under clinical guidelines were tested with a 29-gene panel. Most (92%) had previously received traditional testing for *BRCA1* and/or *BRCA2*. 43 reference samples were also included. Sequence and copy-number variants (CNVs) were called from the NGS data using a battery of algorithms (GATK, Freebayes, PolyMNP, CNVitae and split-read analysis). Patient records and family histories were reviewed by medical geneticists and genetic counselors to evaluate the implications of these results.

750 variants could be directly compared between the panel results and the previous tests. No NGS false positives or false negatives were observed. Importantly, 48 of these 750 were of types known to be technically challenging for NGS, e.g. large indels (max 126bp), small CNVs (half of which affected only single exons) and complex events. Of the 260 *pathogenic* variants, 13% were of these challenging types, underscoring the importance of accurate methods to detect them. No single NGS calling algorithm achieved this performance but rather the combination did. Considering BRCA1/2 interpretations, 99.8% concordance was observed with data from the previous testing lab.

5% of the BRCA-negative patients had a pathogenic variant in another dominant-acting cancer risk gene. Most of these findings were in PALB2, ATM, CHEK2 or the Lynch syndrome genes. In 80% of these cases the patient's cancer and/or family history was consistent with the known effects of the gene they carry, suggesting that these findings are not incidental. In 70% of cases the non-BRCA findings would warrant consideration of a change in care under current medical guidelines.

NGS panels can be a viable replacement for traditional genetic tests, even considering the significant fraction of pathogenic variants that are challenging for NGS. The non-BRCA findings we saw were largely consistent with presentation or family history even though many of these patients would not have been eligible for that gene test under current guidelines. Moreover many of the non-BRCA findings could potentially improve care and outcome. Orthogonal confirmation (e.g. by Sanger sequencing) of NGS results remains a recommended practice, although the high concordance of Sanger and NGS data suggests that this merits careful consideration over time.

[1] Kurian et al., J Clinical Oncology, 2014

Next Generation Sequencing of Short Tandem Repeats: additional variation takes Forensic mixture analysis to the next level

Van der Gaag, K.J.[1], de Leeuw, R.H.[1], Laros, J.F.J.[1,2,3], den Dunnen, J.[1,2] and de Knijff, P.[1]

[1]*Department of Human Genetics Leiden, Leiden University Medical Centre, the Netherlands*
[2]*Leiden Genome Technology Centre, Leiden, the Netherlands*
[3]*Netherlands Bioinformatics Centre, Leiden, the Netherlands*
Corresponding author: knijff@lumc.nl

In the last two decades, forensic DNA research almost exclusively used Capillary Electrophoresis (CE) of Short Tandem Repeats (STRs) using fluorescently labeled PCR-products as the preferred genotyping technology. However, although being successful, CE is not without limitations. STR genotypes are identified by PCR product-size only, thereby ignoring any additional sequence-variation in the PCR product. More importantly, CE-based STR genotyping also limits the identification of minor (20% or less) contributions in mixed-DNA samples.

Using the MiSeq (Illumina) we sequenced the STRs used in most forensic CE-assays for 300 samples from three globally dispersed populations to explore additional sequencevariation. We analysed a set of mixed DNA-samples in different ratios ranging from 1:1 to 1:99 to assess the performance of this method for the analysis of challenging forensic samples.

Most STRs revealed substantial additional variation at the sequence-level resulting in an increased number of alleles compared to CE-analysis of STRs in the same samples. This additional variation results in an increased discriminating power and facilitates better deconvolution of mixed DNA-samples. Even in two-donor mixtures with a ratio of 1:99 all alleles from both contributors could be observed. Although analysis of samples with multiple unknown donors is still complex at this level, the additional sequence-variation in the alleles helps to distinguish true alleles from artefacts and greatly improves genotyping of forensic (and other) DNA mixtures.

COMPANY LECTURE

The Irys System; Rapid Genome Wide Mapping at the Single Molecule Level Using Nanochannel Arrays for Structural Variation Analysis and *de novo* Assembly

Dr. Jack Peart, Director of Sales – EMEA

BioNano Genomics UK Ltd

Despite continued reductions in the cost of sequencing, improvements in base-calling accuracy, and recent advances in read length, complete *de novo* assembly and genome wide structural variant analysis of large, complex genomes remains expensive and challenging.

We present a rapid genome wide analysis method based on NanoChannel Array technology (Irys) that dynamically streams and linearizes extremely long DNA molecules for direct image analysis. This high-throughput platform automates the imaging of individual molecules of genomic DNA hundreds to thousands of kilobases in length that have been labeled at specific sequence motifs. High-resolution genome maps are assembled *de novo*, preserving long-range structural variation and haplotyping information that is intractable by current short read NGS platforms.

SESSION 5

De novo Genome Assembly using Diverse Data Types

I. Gut[1], T. Alioto, F. Cruz, L. Frias, P. Ribeca

Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain
Corresponding author: igut@pcb.ub.es

At the CNAG we have carried out *de novo* genome assembly and annotation for many different species (e.g. Iberian lynx, turbot, cedar aphid, almond, wasp, olive). The objective of each *de novo* assembly is to provide high contiguity and scaffolding with correct order and orientation of contigs so that downstream annotation has the best possible chance to recover the genes and gene structures correctly. Over the years we have refined the sequencing strategies that we apply to optimize contiguity and scaffolding while

reducing the overall cost of a *de novo* assemble. Our basic strategies use paired-end whole genome shotgun sequencing together with matepair analysis, fosmid pool shotgun sequencing and fosmid end sequencing in pools that are all run on Illumina sequencing systems. Computational assembly strategies start with *de novo* assembly of fosmid pools, followed by error correction with the whole genome shotgun data. Scaffolding is refined with the matepair and fosmid end sequences. Throughout we apply a contig breaking and re-assembly strategy. Recently, we have started experimenting with including other data types. In particular these are the inclusion of fosmid pool sequencing using Oxford Nanopore Technologies MinIon systems and the Iris system of BioNanoGenomics.

In this presentation we will discuss the strategies and pipelines that we have developed to achieve high quality reference genomes for the different *do novo* assembly projects we have carried out.

Rapid screening for monogenic diseases in severely ill newborns using whole genome sequencing

CC van Diemen[1], TJ de Koning[1,2], B Sikkema-Raddatz[1], JDH Jongbloed[1], KM Abbott[1], PBT Neerincx[1], G de Vries[1], M Meems-Veldhuis[1], M Viel[1], AJ Scheper[1], K de Lange[1], J Dijkhuis[1], J van der Velde[1], M de Haan[1], MA Swertz[1], KA Bergman[2], P Rump[1], M Kerstjens[1], CMA van Ravenswaaij[1], IM van Lange[1], C Wijmenga[1], RH Sijmons[1], RJ Sinke[1]

[1]. Dept of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2]. Beatrix Children Hospital, University Medical Center Groningen, Groningen, The Netherlands
corresponding author: c.c.van.diemen@umcg.nl

Monogenic diseases are frequent causes of neonatal morbidity and mortality, and disease presentations are often undifferentiated at birth. Additionally, many monogenic diseases feature clinical and genetic heterogeneity. Routine molecular testing is time consuming and available for only few of these diseases. For severely ill newborns quick molecular diagnoses is important for clinical decision-making and can prevent unnecessary and sometimes invasive diagnostics.
Here we describe a procedure and present the first results to analyze 2800 genetic disorders in severely ill newborns by whole-genome sequencing (WGS) where we aim to provide a differential diagnosis within 4 weeks, but preferably shorter. The procedure is carried out by a team consisting of physicians from the Neonatal and Pediatric Intensive Care Units, clinical geneticists, technicians and laboratory specialists from the Genome Diagnostics laboratory, researchers and bioinformaticians. Newborn patients with a suspected genetic disease are presented by the treating physician and the clinical geneticist to the project team. Upon inclusion of the patient in the study, parents of the patient are counseled by the clinical geneticist. When consent is obtained, blood is collected from the patient, DNA is isolated, prepared for WGS using Nextera library preparation, and sequenced on an Illumina HiSeq 2500 (50 hours). Mean whole genome coverage of ~30x is generated, as validated in 4 test samples. Using an in-house developed pipeline on a dedicated server environment, sequence reads are aligned and variants are called only for genes present in the Clinical Genomics Database (20 hours). Variants are then filtered using Cartagenia software for population frequencies, patient phenotypes using Human Phenotype Ontology (HPO) terms, zygosity and possible functional effect (2 hours). Resulting variants are further annotated and filtered on a gene level with inheritance information for the corresponding OMIM disease and pathogenicity scores from the CADD database and manually judged for their potential disease-causing effect by the project team (3 hours). Candidate disease causing variants are validated using Sanger sequencing. Validation tests have shown that variant calling is accurate and known pathogenic mutations in the test samples have been detected. The procedure is accredited under CCKL regulations.

Thus far we have included seven patients in the study and have provided a genetic diagnosis for one patient. This patient presented with microcephaly, seizures and developmental delay and appeared to have compound heterozygous mutations in the *EPG5* gene which is associated with Vici syndrome. Currently we are optimizing the procedure by including additional CNV algorithms, extra annotation of variants, and by integrating SNP array results with sequencing data. Ultimately, we aim to decrease the turn-around-time time to 72 hours.

The impact of genetic variations in miRNA binding sites on the miRNA-mediated regulation of genes associated with cardiometabolic traits

Mohsen Ghanbari[1,2], Oscar H Franco[1], Hans de Looper[3], Albert Hofman[1], Stefan Erkeland[3], Abbas Dehghan[1]

[1]Department of Epidemiology, Erasmus University Medical Center, 3000 CA Rotterdam, the Netherland
[2]Department of Genetics, School of medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[3]Department of Hematology, Erasmus University Medical Center, Cancer Institute, 3000 CA Rotterdam, the Netherlands.
Corresponding author: a.dehghan@erasmusmc.nl

Genome-wide association studies (GWAS) have enabled us to discover a large number of genetic variants and loci contributing to cardiovascular and metabolic disorders. However, since the vast majority of the identified variants are thought to merely be proxies for other functional variants, the causal mechanisms remain to be elucidated. Here, we hypothesized that part of the functional variants involved in deregulating cardiometabolic genes are located in microRNA (miRNA) binding sites.

Using the largest GWAS available on glycemic indices, lipid traits, anthropometric measures, blood pressure, coronary artery diseases and type 2 diabetes, we identified 11,353 variants that are associated with different cardiometabolic phenotypes. Of these, 191 variants (at 129 genomic loci) are located in putative miRNA binding sites. Thirty-four out of 191 variants were found to fulfil our predefined criteria for being functional. Ten variants were subsequently selected for experimental validation based on GWAS results, eQTL analyses and evidence for co-expression of their host genes and regulatory miRNAs. Luciferase reporter assays revealed an allele-specific regulation of genes hosting the variants by miRNAs. These co-transfection experiments showed that rs174545 (FADS1:miR-181a-2), rs1059611 (LPL:miR-136), rs13702 (LPL:miR-410), rs1046875 (FN3KRP:miR-34a), rs7956 (MKRN2:miR-154), rs3217992 (CDKN2B:miR-138-2-3p) and rs11735092 (HSD17B13:miR-375) abrogate miRNA-dependent regulation of the transcripts. Conversely, two variants, rs6857 (PVRL2:miR-320e) and rs907091 (IKZF3:miR326), were shown to enhance the activity of the miRNAs on their host transcripts.

We provide evidence for a model in which polymorphisms in miRNA binding sites can both positively and negatively affect miRNA-mediated regulation of cardiometabolic genes. This approach could be applied to a wide range of phenotypes and may contribute to improving the annotation of GWAS findings.

Key words: miRNA binding site, miRNA polymorphism, cardiometabolic phenotypes.

SESSION 6

New Approaches to Pathogenicity Interpretation in the Age of Precision Medicine

Marc Greenblatt

Abstract not available at time of preparation.

Evaluating Inheritance Pattern for Genetic Variant Interpretation: are we generally oversimplistic?

Sobrido MJ, Blanco-Arias P, Castro C, Ordóñez-Ugalde A, Carracedo A, Quintáns B

Fundación Pública Galega de Medicina Xenómica-Instituto de Investigación Sanitaria
Santiago de Compostela, Spain

When interpreting NGS data for diagnosis, compliance with the suspected mode of inheritance is one of the first and main steps generally used to filter the large lists of genetic variants obtained in a patient or family. Causality of a given variant is ruled out if it does not co-segregate with symptoms in the family; genes with homozygous or compound heterozygous variants are sought for when a recessive condition is suspected, and so forth. However, since Mendel's Laws, our knowledge of the mechanisms of genetic inheritance has grown immensely, uncovering many exceptions to the basic rules. For instance, there are degrees of dominance and recessiveness with some traits, while others are determined by the combined effect of two or more genes. Two alleles are co-dominant if both are expressed in heterozygous individuals, i.e., the phenotype is not intermediate between the two. In some cases both homozygous and heterozygous individuals can express the phenotype, only the homozygous state causes a more severe phenotype than the heterozygous state (semidominant). It is also becoming clear that multiple allele series - where three or more alleles are involved in

determining the phenotype - are very common. Linkage disequilibrium and other mechanisms can cause that inheriting one allele increases the chance of inheriting another. This could increase the chance that a trait observed within a family is actually the consequence of more than one genetic variant and, therefore, that one of those variants alone might not be causal in a different family. On the other hand, the effect of some alleles does not occur unless certain environmental factors are present. The influence of modifying and regulator genes, as well as gender-dependent effects and epigenetic marks may underlie pleiotropic manifestations, variable expressivity, incomplete penetrance and anticipation, adding to the complex scenario in which we interpret one individual´s genetic variants. In summary, exceptions to the simple Mendelian rules of inheritance are, in fact, relatively common. We need to take this into account when interpreting NGS data and don't dismiss the fact that there could be more than one or two variants at play in the clinical manifestations of a given patient. These considerations are crucial, given the implications of variant interpretation for genetic counseling, medical and reproductive decision-making.

Using Selective Constraint at the Domain Level to Assess Variant Pathogenicity

Cassa Christopher A[1], Jordan Daniel M[1] and Sunyaev Shamil R[1]

Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA
Corresponding author: cassa@mit.edu

The clinical assessment of genetic variants poses unique challenges when limited phenotypic or population data are available during classification. We extend existing classification approaches using human population sequencing data to measure selective constraint at the domain level. We measure selective effect in two ways: using data from asymptomatic individuals to predict the sensitivity of each domain to novel variation, and using data from symptomatic individuals to predict which types of lesions will functionally impact that domain. This allows for variant classification at the domain level, providing insight into which regions are most likely to harbor missense mutations linked to Mendelian diseases. This method can be used to assess variant pathogenicity and to prioritize variants in genes where clinically important variation has already been observed.

Data from asymptomatic individuals (NHLBI Exome Sequencing Project) provides a distribution for each functional category of variant in each domain. For example, what is the frequency and expected number of nonsense or missense variants in an asymptomatic individual in each domain, and how does it differ from other domains? In asymptomatic individuals, we expect that genes with large depletions of rare nonsense or missense variants will be unlikely to tolerate such variants, especially if predicted to have negative protein or evolutionary effects (Goldstein, et al. 2013). We also include selective signals, including the ratios of nonsense:synonymous and missense:synonymous variants.

Next we include data from symptomatic individuals in two clinical laboratories, HGMD, ClinVar, and OMIM, to understand the functional impact of variation in each domain. In aggregate, we consider the frequency and number of pathogenic variants by functional category, to understand the potential impact of variants in each domain. These domains are then annotated with phenotypic information about associated diseases, including prevalence and mode of inheritance.

Finally, we combine both the asymptomatic and symptomatic data at the gene and domain level, creating a multi-dimensional feature space. We develop a Naïve Bayes classifier to assess pathogenicity of clinically important and benign variants. When trained on variants with existing functional annotations (N=25,317) the classifier accurately predicts variant pathogenicity (AUC=0.853) and can predict well in the separately ascertained HumVar dataset (AUC=0.841).

This application of evolutionary biology can be used to predict clinical importance of variants, and allows for the integration of a large knowledge base of genetic variants in genomic interpretation. This classifier can help filter results that are unlikely to be clinically impactful, protecting patients from invasive and unnecessary interventions and screening.

PON-P2 and PON-Diso - reliable prediction of variation pathogenicity

Abhishek Niroula, Siddhaling Urolagin and Mauno Vihinen

Department of Experimental Medical Science, BMC D10, Lund University, SE-22184 Lund, Sweden
Corresponding author mauno.vihinen@med.lu.se

Reliable prediction methods are needed to analyze NGS datasets. Due to the large volume of identified variations only computational approaches can handle

the datasets. Genomes contain millions of variants and typically over 10 000 of them lead to amino acid substitution. Tools for filtering and prioritizing the cases have to be both reliable and fast. We have developed a completely new machine learning-based method PON-P2 [1], which fulfills these criteria. It has excellent performance (accuracy 0.87, MCC 0.77) and it can predict very large datasets in reasonable time. It is more reliable and faster than our previous predictor, PON-P [2] and competing methods such as CADD; MutationTester2, PolyPhen2, SIFT etc. The method is based on extensive feature selection and training with a large benchmark dataset from VariBench [3]. The method is implemented with random forest, a powerful classifier.

PON-Diso is a method for predicting effects of amino acid substitutions on order/disorder status of proteins [4]. Several proteins contain disordered regions or are completely disordered, i.e. without regularly ordered structure. Changes to the structure in these regions can be related to diseases. We tested the performance of a large number of existing disorder prediction methods, but they were found unsuitable for this task. We developed a novel tool that has a success rate on 70% in cross validation.

These tools have been tested with independent benchmark datasets and they show the highest performance currently available.

Both the predictors are available at http://structure.bmc.lu.se/

References:
1. Niroula, A., Urolagin, S. and Vihinen, M. PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE* (in press).
2. Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. (2012) PON-P: Integrated predictor for pathogenicity of missense variants. Hum Mutat 33, 1166-1174.
3. Nair PS, Vihinen M. VariBench: A benchmark database for variations. (2013) *Hum Mutat* **34**, 42-49.
4. Ali HS, Urolagin S, Gurarslan Ö, Vihinen M. Performance of disorder prediction methods on missense variants. *Hum. Mutat.* **35**, 794–804.

SESSION 7

The European Variation Archive: A New Genetic Variation Resource at EMBL-EBI

Saunders G[1], Medina I[1], Spalding D[1], Y. Gonzalez C[1], Kandasamy J[1] and Paschall J[1]

[1]*EMBL-EBI, Hinxton, Cambridge, UK*
corresponding author: garys@ebi.ac.uk

The European Variation Archive (EVA) is a new genetic variation resource at EMBL-EBI. The central aim of the EVA is to catalogue all types of variation from all species and provide a suite of tools to access and analyze these data. The EVA currently contains over 1.7 billion submitted short variants (less than 50bp) from a range of more than 20 large-scale projects that includes 1000 Genomes (phases 1 and 3), Exome Variant Server, Genome of the Netherlands and UK10K. We shall present the EVA data model that links each variant to sample(s), submitter, methodology and reference genome build. This detailed structuring of our archive allows dynamic querying of the data and permits calculation of allele frequencies across studies and populations, a unique feature of the EVA when compared to other similar resources.
Associated with submission, the EVA accepts submitter assertions of the relevance of a variant to a clinical phenotype. Our workflows here are in agreement with ClinVar and data submitted to the EVA is also shared with our collaborator at NCBI. We shall present our submission process and describe each of the steps from notification of 'intent to submit' to data sharing/mirroring with ENSEMBL, dbSNP and ClinVar.
The EVA also provides direct access to structural variants (more than 50bp) stored at the Database of Genomics Variants Archive (DGVa) at EMBL-EBI. Genomic structural variation accounts for the majority of the individual differences at the base pair level in humans and has been shown to play a role in genomic disorders such as cancer, human evolution, and genetic diversity. The DGVa currently stores over 14 million variants and more than 15 million genotypes. Data submitted to the DGVa is curated at a high granularity permitting a robustly structured archiving of the data. We shall present how this level of detailed curation allows the data to be shared in a user-friendly and informative way via our website and programmatic interface.

The complete repository of the EVA data can be searched, viewed and queried via our EVA Portal application. This web portal provides data mining and visualisation tools that permit views at the granularity of a given submitted study or individual variant. Complex queries can be built based upon entry points including study, gene, genomic location, variation type and/or consequence type. Our web portal is based on modern web technologies and our front-end is made modular and scalable through use of a RESTful web service interface to the backend data store, allowing the EVA data to be accessible

programmatically for a variety of applications such as annotation pipelines. All of our software is released as open source (https://github.com/EBIvariation/eva). The EVA website can be found at www.ebi.ac.uk/eva Please direct questions and submissions to eva-helpdesk@ebi.ac.uk

Funding: The EVA is funded by EMBL, Wellcome Trust, Medical Research Council, The European Union, European Commission Framework Programme 7 and RD-Connect.

Computer analysis of context dependencies of SNP sites in human genome

Safronova Nataly[1,2] and Orlov Yuriy L.[2]

[1]Novosibirsk State University, Novosibirsk, Russia,
[2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
corresponding author: orlov@bionet.nsc.ru

Studying the dependence of emergence of SNPs on surrounding nucleotide context is an important problem of Genetics. Investigating these polymorphisms can allow us to answer a number of important biological questions and to achieve a success in prediction and cure of genetic diseases [1].

Single nucleotide polymorphisms (SNPs) are the most widespread form of sequence variation in the genome, representing about 90% of human DNA. Several studies considering small numbers of di and trinucleotide repeats have found evidence of an increase in mutation rate near the repeat [2-4]. However, such studies used limited SNP data and only estimated mononucleotide repeats without counting overall text complexity.

Such characteristic as complexity can be defined for each sequence. Intuitively, complexity of symbolic sequence reflects an ability to represent a sequence in a compact form based on some structural features of this sequence. To evaluate text complexity there were developed several methods.

In our research were used next methods which allow us to study different features of genetic sequences:
Lempel-Ziv complexity measure [5, 6] is based on text segmentation and it allows finding repeated sequence blocks: in smaller scale - in gene transcription regulatory regions, in larger scale - in complete bacterial genomes.
Shannon entropy provides a measure of the irregularity of oligonucleotide composition. It is calculated by Shannon's formula.

Linguistic measure is defined as the ratio of the number of encountered "words" to the number of all possible "words" in the sequence of a given length (under the "word" we mention the genetic sequence of the fixed length).

Above mentioned measures are part of the complexity measures implemented in the software package developed at the Institute of Cytology and Genetics [7, 8].
Also for our analysis the additional measures were used:
Measure of monomers supplements Shannon's entropy and reflects the proportion of monomers in the text.
Weight measure allows specifying the composition of the sequence, supplements all known measures.

Nucleotide sequences, which contain annotated SNPs, from Human genome were analyzed (sequence length 101nt: 1 for polymorphism and +\- 50 nt for surrounding region). Four classes of SNPs – as for nucleotide variant in genetic sequence - were studied separately, for each class was provided complexity analysis in sliding window (window sizes were 5,7,10 nt). Complexity analysis was provided with program complex which was developed in Institute of Cytology and Genetics [7].
Also were selected and compared oligonucleotides of length 7 which have the highest frequencies of occurrence in SNP point and in remote point of one of the flanks.
Such research was provided for Mouse and Rat SNPs, similar results were obtained.

The analysis of Human SNPs was provided and the reduction of local text complexity in 3-5nt region near the SNP point was shown. The saturation of the polymorphism point with politracts was obtained, which indicates a higher probability of instability of dual DNA helix in this point.

References

Cooper et al., 2006
Vowles, E.J. and Amos, W. (2004) Evidence for widespread convergent evolution around human microsatellites, PLoS Biol, 2, E199
Siddle KJ, Goodship JA, Keavney B, Santibanez-Koref MF. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics. 2011 Apr 1;27(7):895-8.
Lenz C, Haerty W, Golding GB. Increased substitution rates surrounding low-complexity regions within primate proteins. Genome Biol Evol. 2014 Mar;6(3):655-65.
Lempel A. and Ziv J. (1976) On the complexity of finite sequences. IEEE Trans. Inf. Theory, IT-22, 75–81.

Gusev V.D., Kulichkov V.A. and Chupakhina O.M. (1993) The Lempel-Ziv complexity and local structure analysis of genomes. Biosystems, 30(1-3), 183-200.
Orlov Y.L., Potapov V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. 32(Web Server issue):W628-33.
Orlov Y.L., Te Boekhorst R., Abnizova I.I. (2006) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J Bioinform Comput Biol. 4:523-36.

An Efficient Algorithm for the Extraction of HGVS Descriptions

Jonathan Vis[1], Peter Taschner[2] and Jeroen Laros[2]

Depts of Molecular Epidemiology[1] and Human Genetics[2], Leiden University Medical Center, Leiden, Nederland
corresponding author: p.taschner@lumc.nl

The Human Genome Variation Society (HGVS) [1] publishes nomenclature guidelines for unambiguous sequence variant descriptions. The Mutalyzer suite [2] has been built with as main purpose the automatic analysis, checking and interpretation of variant descriptions using HGVS nomenclature. When complex variants are observed, construction of the corresponding descriptions is not always straightforward. Automated extraction of HGVS descriptions by comparison of raw sequences would prevent errors. We propose an efficient algorithm for descriptions of the difference between sequences in HGVS format with three main requirements in mind: minimizing the length of the resulting descriptions, minimizing the computation time and keeping the descriptions meaningful for the user.

The underlying idea of our extraction algorithm is to divide the sequences (strings) from which a description is to be extracted into a list of unaltered regions and regions of change (substrings). In order to minimize the length of the resulting descriptions, we apply a greedy heuristic in choosing the longest possible unaltered regions, thereby conforming to the HGVS guidelines.

In addition to the traditional edit operations we define three additional operators. Inversion matches the reverse complement of the sequence, transpositions are substrings which are copies of sequences found elsewhere in the reference sequence. Finally, we have inverse transpositions: transpositions that match the reverse complement of the sequence. The algorithm is formulated recursively: given two sequences, find the longest sequence that is contained in both sequences. Then remove this sequence and continue recursively with both prefixes and both suffixes. The recursion ends when either one of the two remaining sequences is empty or no common substring could be found. These remaining sequences are described as variants. The traditional dynamic programming approach for finding the longest common substring has a quadratic time complexity. For similar sequences, splitting both into overlapping and non-overlapping k-mers gives us an expected linear time complexity and thus a faster results for large sequences.

We have used complete hg18 and hg19 genome build sequences to extract the HGVS descriptions per chromosome using hg18 as a reference with a total computation time of 38 hours. We observe that the maximum size for any chromosome from hg19 described relative to hg18 is about 1 MB. For most chromosomes, descriptions can be calculated in at most 1 hour. Notable exceptions are chromosomes 5, 7, 8, and X, where a large number of small insertions, just large enough to be considered for transposition extraction, are observed.

We show that the extraction algorithm is able to compute relatively small HGVS descriptions from large DNA sequences in a reasonable amount of computation time. In addition to having a canonical algorithm for generating HGVS descriptions, this approach also supports efficient storage of large (haplotype) sequences.

Acknowledgement
This publication was supported by the Dutch national program COMMIT.

References
[1] Nomenclature for the description of sequence variants, www.hgvs.org/mutnomen, Feb. 20, 2014.
[2] https://mutalyzer.nl

Wednesday 29th April

SESSION 8

Developments in single-cell sequencing

Thierry Voet

[1] Centre for Human Genetics, University Hospital Leuven, Department of Human Genetics, KU Leuven, Belgium
[2] Single-cell Genomics Centre, Welcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

Single-cell genomics enables investigating the extent and nature of genomic and transcriptomic heterogeneity which occurs in both normal development and disease, and in addition provides new tools for clinical application. We have developed various wet-lab and computational methods that allow analyzing the (deoxy)ribonucleic acids of a solitary cell via high resolution microarray, SNP-array and next-generation sequencing methods. Using this toolkit for single-cell genomics, we are investigating the acquisition of (functional) genetic variation in human development; from embryogenesis towards healthy and diseased organs. These and other studies, as well as a visionary on the future, will be presented.

Reverse Diagnosis of Missense Variants in DNA Mismatch Repair Genes in Lynch syndrome

Mark Drost[1], Anne Lützen[2], Fabienne Calléja[1], Daniel Ferreira[1], José Zonneveld[1], Sandrine van Hees[1], Finn Cilius Nielsen[3], Lene Juel Rasmussen[2] and Niels de Wind[1]

[1] Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Center for Healthy Aging, University of Copenhagen, Copenhagen, Denmark
[3] Center for Genomic Medicine, Rigshospitalet, University of Copenhagen, DK-2100 Copenhagen, Denmark

Lynch syndrome (LS) is an inherited cancer susceptibility characterized by a high risk for the development of colorectal and other cancers. LS is caused by mutations in genes involved in the DNA mismatch repair (MMR) pathway. Determination of the disease-causing mutation within families is important, as this confirms diagnosis and enables proper risk assessment. About 40% of all MMR gene alterations found in patients suspected of LS are so-called Variants of Uncertain Significance (VUS), generally missense alterations. This class of alterations poses a problem since it is difficult to distinguish between pathogenic, disease-causing, mutations and benign polymorphisms.

We have developed a method to identify pathogenic missense mutations by lookup in a comprehensive catalog of variants that inactivate MMR *in vivo.* The generation of such a catalog entails the prior identification of all amino acids that are important for MMR protein function *in vivo*, using a large-scale mutagenesis screen. The screen is based on randomly mutating a mouse embryonic stem cell line, heterozygous for one of the four MMR genes, using the point mutagen N-ethyl-N-nitrosourea (ENU). After ENU treatment, clones that have lost MMR activity are selected using the nucleotide analog *6-Thioguanine.* Identification of the mutation in each clone pinpoints amino acid alterations that abolish MMR *in vivo*, and therefore would cause LS when also identified in an individual suspected of LS. These deleterious mutations are included in the "Reverse Diagnosis Catalogs" (RDC).

We have used this method to identify 26 inactivating substitutions in the MMR genes *MSH2* and 51 in *MSH6*, comprising pilot RDC. We have employed *in silico*, cellular and biochemical assays to validate these substitutions. Of these 77 variants, 23 have previously been identified in individuals suspected of LS. Currently we aim to further optimize the screening procedure for the generation of RDC and are generating comprehensive RDC of the major MMR genes. This effort is expected to greatly improve personalized preventive and curative healthcare in LS patients and their relatives in the foreseeable future.

Calibrating an in vitro functional assay for use in the diagnosis of Lynch syndrome due to missense variants in the mismatch repair genes

Niels de Wind, Mark Drost, Bryony A. Thompson, Amanda B. Spurdle, David E Goldgar, Marc S. Greenblatt, Sean V. Tavtigian

Background: Lynch Syndrome (LS), a hereditary condition predisposing to colorectal and other cancers, is caused by a germ-line mutation in one of the DNA mismatch repair (MMR) genes. A significant fraction of the variants detected in suspected LS patients are missense amino acid substitutions. The pathogenicity of many of these variants is difficult to assess. Currently available tools include segregation, in silico assays, and some tumor pathology assessments. Until now, no time- and cost-effective, validated, and widely applicable functional assay for measuring MMR activity has been available. Here we describe the calibration of our rapid, quantifiable, cell-free, in vitro MMR activity assay (CIMRA, Drost 2010, 2012, & 2013).

Methods: We used the CIMRA to test the MMR function of a series of MMR gene variants that have been classified using other evidence as either pathogenic, likely pathogenic, likely neutral, or neutral (Classes 5, 4, 2, or 1, per the schema of Plon, 2008). We performed logistic regression of the CIMRA values, expressed as percentage of the

activity of wild type protein, versus probability of pathogenicity using other evidence (0.1%, 5%, 95%, or 99%) to convert CIMRA % of wild-type activity to odds in favor of pathogenicity so that the data could be combined with sequence analysis-based prior probabilities (Thompson 2013) to estimate posterior probabilities of pathogenicity.

Results: Of 35 variants classified as pathogenic by other evidence, CIMRA activity was <25% of wild type in 29/35 (82.9%), 25-50% of wild type in 3/35 (8.6%), and >50% of wild type in 3/35 (8.6%). Combining CIMRA with the prior probability of pathogenicity from in silico analysis corroborated the assessment of 31/35 (88.6%) as Pathogenic; 3/35 fell into the category of "Uncertain" by multifactorial analysis; and 1/35 (2.9%) resulted in a classification of "likely neutral". Of 15 variants classified as neutral by other evidence, CIMRA activity was >50% of wild type in 14/15 (93.3%), and <50% of wild type in 1/15 (6.7%). Combining CIMRA with the prior probability of pathogenicity from in silico analysis corroborated the assessment of 11/15 (73.3%); 3/15 (20%) fell into the category of "Uncertain" by multifactorial analysis; and 1/15 (6.7%) resulted in a classification of "likely pathogenic".

Conclusion: The CIMRA has strong predictive value and can be used as a tool to evaluate missense MMR variants for pathogenicity. The CIMRA is scalable and can be used to test large numbers of missense variants. We are exploring other in vitro assays to assist in the classification of variants that remain of uncertain significance despite data from the CIMRA, clinical features, and in silico analysis.

SESSION 9

Spatially resolved mutation detection and sequencing *in situ*

Mats Nilsson

Stockholm University, Sweden

We have developed a technology that enables creation of targeted sequencing libraries in morphologically preserved cells and sections of tissues. The sequencing libraries are generated through rolling circle amplification (RCA) of probes that becomes circularized in a target specific way on cDNA synthesized *in situ*. The in situ generated libraries are sequenced by sequencing by ligation chemistry. The technology permits detection of somatic point mutations and genomic rearrangements generating fusion transcripts, and expression profiling of sets of genes. We have applied the technology for detection of oncogenic hotspot mutations and fusion transcripts in cytological preparations as well as in sections of fresh frozen and formalin fixed paraffin embedded tumor tissues. We have also shown that rare mutated cells can be detected in the background of a vast majority of normal cells. We also demonstrate that tumor heterogeneity can be visualized, by targeting multiple somatic mutations, and that molecular phenotypes can be associated with different tumor sub-clones by *in situ* expression profiling.

References:
1. Larsson, C., Grundberg, I., Söderberg, O. and Nilsson, M. (2010) In situ detection and genotyping of individual mRNA molecules. *Nat. Methods*, **7**, 395–397.
2. Grundberg, I., Kiflemariam, S., Mignardi, M., Imgenberg-Kreutz, J., Edlund, K., Micke, P., Sundström, M., Sjöblom, T., Botling, J. and Nilsson, M. (2013) In situ mutation detection and visualization of intratumor heterogeneity for cancer research and diagnostics. *Oncotarget*, **4**, 2407-2418.
3. Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wahlby, C. and Nilsson, M. (2013) In situ sequencing for RNA analysis in preserved tissue and cells *Nat. Methods*, **10**, 857-860.
4. Kiflemariam, S., Mignardi, M., Ali, M.A., Bergh, A., Nilsson, M. and Sjoblom, T. (2014) In situ sequencing identifies TMPRSS2-ERG fusion transcripts, somatic point mutations and gene expression levels in prostate cancers. *J. Pathol.*, **234**, 253-261.

Full-length mRNA sequencing uncovers a widespread coupling between transcription and mRNA processing

Seyed Yahya Anvar[1,2,*], Eleonora de Klerk[1], Martijn Vermaat[1,2], Johan T den Dunnen[1,2], Stephen W. Turner[3], Peter AC 't Hoen[1,*]

[1]Department of Human Genetics and [2]Leiden Genome Technology Center, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands
[3]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA
corresponding author: p.a.c.hoen@lumc.nl

High-throughput RNA sequencing helps deciphering the global landscape of RNA expression. However, a comprehensive survey of transcriptional and posttranscriptional events in the same mRNA molecule is compromised by the short read length of second-generation sequencing platforms. Here, we

analyzed Pacific Biosciences long sequencing reads capturing full-length mRNA molecules in MCF-7 human breast cancer cells. We obtained 7.4 million single-molecule long sequencing reads representing full-length mRNA molecules. From the 14,385 genes with detectable expression, 49% produced multiple transcripts. A total of 93 candidate fusion genes were identified based on the inter-chromosomal or distant intra-chromosomal split-alignment of transcripts to the human reference genome. In addition, 42% of identified transcripts in MCF-7 are potentially novel in comparison with the GENCODE annotation.

Our long read-based survey and quantification of transcripts demonstrates a striking degree of coordination between transcription initiation, splicing and polyadenylation. In nearly half of the genes the selections of alternative transcription start sites, alternative exons or alternative polyadenylation sites are interdependent. Notably these couplings can occur over large distances, and a particular selection of a transcription start site at the 5'-end of a transcript can influence the choice of the polyadenylation at the 3'-end. Interestingly, alternative polyadenylation sites that are coupled with alternative splicing events are depleted for known polyadenylation signals and enriched for binding motifs for RNA binding proteins from the muscle blind (MBNL) family. Our data suggest a coordinating role for MBNL proteins in the regulation of splicing and polyadenylation.
Our findings demonstrate that our understanding of transcriptome complexity is far from complete. Full-length transcript sequencing provides excellent opportunities to study largely unresolved mechanisms that coordinate transcription and mRNA processing and the effect of genetic variants on transcript structure.

How to choose the right predictor for variation interpretation

Mauno Vihinen

Department of Experimental Medical Science, BMC D10, Lund University, SE-22184 Lund, Sweden
mauno.vihinen@med.lu.se

Analysis and interpretation of NGS and other sequencing data is dependent on computational methods. When identifying variations one typically first checks whether the variant is previously known by checking from locus specific variation databases and from central variation databases. However, unique variations are often identified and then the first resource to utilize is computational prediction methods. There are basically two types of tools. Tolerance predictors aim at identifying harmful/harmless variations while mechanism-based tools try to predict the effects and consequences of variations.

For average end user it is difficult to find out the best bioinformatics methods. There are numerous tools, new are published all the time and all they claim to be better than the others! Some guidelines will be presented for selecting best performing tools [1,2]. Most of the methods are utilize machine learning algorithms i. e. kind of artificial intelligence. These tools are basically statistical and trained to distinguish between benign and harmful ones. There are three crucial issues: training and testing data, feature selection, and actual implementation of the predictor. In addition, the results have to be presented in a comprehensive way to provide full picture of the performance.

Recently benchmark datasets have been released and utilized for comparison of prediction methods available in VariBench [3] and VariSNP [4] for different types of methods. Comparisons of variation tolerance [5], protein stability [6], protein disorder [7] and protein localization [8] method performance revealed great differences in method performance. The end users need to pay attention to the methods they are using, not just taking the same one as last time or one widely used in the past literature. There are still some additional pitfalls to be avoided and which one should be aware e. g. how to combine predictions [9].

References

1. Vihinen, M. (2012) How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13**(Suppl 4):S2.
2. Vihinen, M. Guidelines for reporting and using prediction tools. (2013) *Hum. Mutat.* **34**, 275-282.
3. Nair PS, Vihinen M. VariBench: A benchmark database for variations. (2013) *Hum Mutat* **34**, 42-49.
4. Schaafsma, G. and Vihinen, M. VariSNP, a benchmark database for variations from dbSNP. *Hum. Mutat.* (in press).
5. Thusberg, J, Olatubosun, A., and Vihinen, M. (2011) Performance of mutation pathogenicity prediction methods. *Hum. Mutat.* **32**, 358–368.
6. Khan, S. and Vihinen, M. (2010) Performance of protein stability predictors. *Hum Mutat.* **31**, 675-684.
7. Ali, H. S., Urolagin, S., Gurarslan, Ö. and Vihinen, M. (2014) Performance of disorder prediction methods on missense variants. *Hum. Mutat.* **35**, 794–804.
8. Laurila, K. and Vihinen, M. Disease-related mutations affecting protein localization. (2009) *BMC Genomics* **10**:122.

9. Vihinen, M. (2014) Majority vote and other problems when using computational tools. *Hum. Mutat.* **35**, 912-914.

SESSION 10

What data sharing means for patients with rare diseases: knowledge from sharing

Milan Macek, Prof. MD

Department of Biology and Medical Genetics, Charles University Prague and University Hospital Motol, V Uvalu 84, Prague, CZ 15006, Czech Rep.
Milan.Macek.Jr@LFmotol.cuni.cz

Outcomes from next generation sequencing, both in terms of whole-exome (WES) and soon also the whole genome analysis (WGA), would potentially change health care provision and foster introduction of precision (or "personalised", "stratified") medicine. In the foreseeable future the main impact of genomic technologies is realistically envisioned in the field of rare diseases (RD). Research within this group of diseases is substantially enriched by the long-term collaboration of patient support groups with health care providers. This unique partnership enables researchers, clinical geneticists and bioinformaticians to obtain realistic, "patient centric", views on the impact of genomics on diagnosis and management of "their" disorders. This unique partnership enables all stakeholders to achieve better understanding of how genomic variation affects our health and develop new diagnostic methods and eventually customised treatments for many RD. It needs to be stressed that sharing of patient data has key role in this process. The uninitiated public debate usually and somewhat "paternalistically" focuses on concerns related to data security, privacy and access. However, recently patient supports groups, as the end-beneficiaries of genomic research, started to express their views on data sharing. Various national- (eg. Genetic Alliance UK) or international RD patient associations (e.g. Eurordis.org) published outcomes of internal discussion or surveys on this topic. For instance Eurordis published the views of their constituents in the recently published monograph "The Voice of Rare Disease Patients" and in the UK absolute majority of RD patients wanted their RD genomic data utilised for better diagnosis and disease management. Unsurprisingly, patient views are different from the general public due to their unique experience. Many of them already have participated in research and in clinical trials, where they learned about the utility of genomic testing. RD patients generally tend to be knowledgeable and positive on data sharing and their broad use in research. Given the small number of individuals with RD, patients support cross-border collection of data to carry out studies with higher statistical power or where various population-specific confounders could be efficiently accounted for. The RD-Connect.eu project involves a multidisciplinary Patient Advisory Council which has elaborated a number of recommendations in this regard. Finally, it needs to stressed that professionals need to be aware that the time of patients is "ticking faster" than theirs. Thus, we should also be careful not to create unsubstantiated expectations and educate patients that genomic information needs to be used responsibly, and that there are still major challenges ahead of us which could not be solved if we do not work together. Realism, honesty and continuous mutual consultations are decisive, since RD families value establishment of a diagnosis through genomic analysis, even when this does not immediately lead to novel treatment modalities.

A charter and code of practice enabling data sharing: ethics, privacy and legal issues

Mats G. Hansson

Centre for Research Ethics & Bioethics, Uppsala University
corresponding author: mats.hansson@crb.uu.se

Sharing data and bio-specimens is essential for the discovery, new knowledge creation and translation of various biomedical research findings into improved diagnostics, biomarkers, treatment develop- ment, patient care, health service planning and general population health. In order to reach a balance that respects both privacy concerns and effectiveness of translational research we need a common approach. Such an approach is available through documents of fundamental human rights that address both concerns of respect for autonomy and of producing knowledge for the prevention of illness and improvemnet of medical treatment. I will present this approach and its implications for sharing and access policies.

Data integration for rare diseases facilitated by the RD-Connect platform

I. Gut[1], D. Piscia[1], S. Laurie[1], A. Cañada[2,14], J.M. Fernández[2,14], C. Kingswood[1], J.P. Desvignes[3,4], M. Thompson[5], R. Kaliyaperumal[5], E. van der Horst[5], S. Lair[6], P. Sernadela[7], A. Topf[8], I. Zaharieva[9], M. Girdea[10], M. Brudno[10], A. Blavier[6], R. Thompson[8], H. Lochmüller[8], M. Bellgard[11], J. Paschall[12], P. Lopes[7], J.L. Oliveira[7], M. Roos[5], P.A.C. 't Hoen[5], V. de la Torre[2,14], Alfonso Valencia[2,14], D. Salgado[3,4], C. Béroud[3,4,13], S. Beltran[1], on behalf of the RD-Connect Consortium

[1]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain
[2]Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain
[3] Aix-Marseille Université, Marseille, France
[4] Inserm, UMR_S 910, Marseille, France
[5]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[6]Interactive Biosoftware, Rouen, France
[7] DETI/IEETA, University of Aveiro, Portugal
[8]Institute of Genetic Medicine, MRC Centre for Neuromuscular Diseases, Newcastle University, UK
[9]Dubowitz Neuromuscular Centre, UCL Institute of Child Health and Great Ormond Street Hospital for Children, London, United Kingdom
[10]Centre for Computational Medicine, Hospital for Sick Children and University of Toronto, Toronto, ON, Canada
[11]Centre for Comparative Genomics, Murdoch University, Perth, Western Australia
[12] European Molecular Biology Laboratory , European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom
[13] APHM, Hôpital TIMONE Enfants , Laboratoire de Génétique Moléculaire, Marseille, France
[14] Instituto Nacional de Bioinformática (INB), Spain
corresponding author: igut@pcb.ub.es

Rare disease research faces particular challenges because patient populations, clinical expertise, and research communities are small in number and highly fragmented. To overcome these challenges the EU FP7-funded RD-Connect project, in collaboration with Neuromics and EURenOmics, is building a platform to harmonise and securely integrate clinical data with biosample and -omics data. The platform already includes over 360 NGS exomes processed with the same analysis pipeline and linked to detailed phenotypes stored in PhenoTips using the Human Phenotype Ontology. The platform runs on a Hadoop cluster and uses technologies such as ElasticSearch, Postgres, Scala and Angular.js, making it highly configurable and efficient. The exomes can be combined in a very flexible manner and variants can be prioritized through the user-friendly front-end using the most common filters and additional tools such as UMD Predictor, DiseaseCard, Alamut Functional Annotation (ALFA) and gene-disease relationships in nanopublication format. Integration of biobank and patient registry data is underway. The project aims to publicly release the first version of the platform during 2015 for authorized users and will gladly accept submissions from other projects in the future.

Genetic variation interoperability standards: Global Alliance for Genomics and Health data sharing beacon project

Peter Goodhand

The Global Alliance for Genomics and Health, Toronto, Canada

The Global Alliance for Genomics and Health is an international, non-profit alliance formed to help accelerate the potential of genomic medicine to advance human health. Bringing together over 270 leading, global organizations working in healthcare, research, disease and patient advocacy, life science, and information technology, members in the Global Alliance are working together to create a common framework of standards and harmonized approaches to enable the responsible, voluntary, and secure sharing of genomic and clinical data. There are currently four active Working Groups: Regulatory and Ethics, Data, Security, and Clinical. These Working Groups are charged with producing thoughtful, actionable conclusions and products in their respective work areas.

Specifically, the Data Working Group concentrates on data representation, storage, and analysis of genomic data, including working with academic and industry leaders to develop approaches that facilitate interoperability. Data sharing projects, like the Beacon Project, attempt to apply the tools and guidelines that are developed within Working Groups in order to evaluate value and demonstrate relevance.
Learn more at: **http://genomicsandhealth.org**.

Datastewardship for Discovery in Human Genetics and health

Barend Mons

Center for Human and Clinical Genetics, LUMC, Leiden, The Netherlands

Knowledge Discovery across biological data resources is hampered by the lack of standards and the poor adoption of existing standards by stakeholders. Data Interoperability is crucial for modern research as it overcomes the barriers of syntactic access with semantic use in one implementation. Optimal Interoperability is only attained when access and use can be completely automated: programming and interfaces conform to standards that specify consistent syntax and formats; and data are associated with metadata and terminology identifiers and codes that support computational aggregation and comparison of information that resides in separate resources. To support modern knowledge discovery data has to be published in a format that renders it Findable, Accessible, Interoperable and ultimately Reusable (FAIR) for machines as well as humans.

Increasingly, European research programmes and public private partnerships already make significant investments in data infrastructure to make data FAIR , but without coordination such as provided by ELIXIR the large numbers of stakeholders and programmes within Europe will drive fragmentation and overlapping investments in data management, stewardship, analytics and technology approaches. Through the implementation of community adopted and ELIXIR endorsed standards and, importantly, a European wide framework of experts and a credible supporting organisation, ELIXIR will drive the coordination efforts both at national and international level.

For Human Genetics research, the FAIR data principle already yields significant benefits. Examples of enhanced Knowledge Discovery potential through interlinking of FAIR data will be given, including across HPA, UniProt, LOVD, FANTOM5 and Semantic MedLine.

The full connection of all variant information, a.o. from GWAS and clinical diagnostics labs is imminent for optimal progress in human genetics

Thursday 30th April

SESSION 11

Rare, non-coding DNA variation and Disease

Stefan White

Department of Human Genetics and Leiden Genome Technology Center (LGTC), Leiden University Medical Center (LUMC), Leiden, the Netherlands

The development of high-density microarrays and next-generation sequencing technologies has revealed the true extent of human genetic variation. Not only does every genome contain millions of single nucleotide variants, but there are also a large number of copy number variants (CNVs). This variation contributes to what makes each individual unique, but in some cases also plays a role in disease. Determining the effect of a coding change is relatively straightforward in many cases, however predicting the consequence of variants affecting non-coding DNA is more problematic. This presentation will discuss studies of rare CNVs and single nucleotide variants in non-coding DNA that are associated with specific conditions.

GeneMatcher: A Matching Tool for Identification of Individuals with Mutations in the Same Gene

N. Sobreira[1], F. Schiettecatte[2], D. Valle[1], A. Hamosh[1]

1) Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) FS Consulting, Salem, MA.

In the last few years, whole exome sequencing (WES) has been the main method used to search for Mendelian disease genes. Identifying the pathogenic mutation from among thousands to millions of genomic variants in typical WES is a challenge. In more than half of the individuals who have a clinical WES the responsible gene and variants cannot be determined. Reasons for this relatively low yield include phenotypic variation; the current, small number of known disease genes (~3300 or only 15% of the total protein coding genes); uncertainty regarding functional consequences of identified variants; and, limited connections between clinicians with clinical WES data and basic scientists. Structured, comprehensive phenotypic data and its

sharing as well as improvements in searching for patients or model organisms with mutations in specific candidate genes are important for the success of clinical and research WES analysis. Here we describe GeneMatcher (www.genematcher.org), a freely accessible web-based tool designed to enable connections between clinicians and researchers from around the world who share an interest in the same gene or genes. No identifiable data are collected. The site allows investigators to post a gene(s) (by gene symbol, base pair position, Entrez- or Ensembl-Gene ID) of interest and will connect investigators who post the same gene. The match is done automatically, submitters will automatically receive email notification and follow-up is at the discretion of the submitters. There is no way to search the database. Submitters have access to their own data and may edit it or delete it at will. There is also an option to provide diagnosis based upon OMIM® number and match on that, but this is not required. If a match is not identified at the time of submission the genes of interest will continue to be queried by new entries. As part of the Matchmaker Exchange project (http://matchmakerexchange.org/), we have also developed an Application Programing Interface (API, available upon request) that is being implemented and allows submitters to query other databases of genetic variants and phenotype information (e.g. PhenomeCentral (https://phenomecentral.org/), DECIPHER (https://decipher.sanger.ac.uk/, etc). In the future, we plan to enable matching based upon phenotypic features, with or without candidate genes to enhance interpretation of clinical and research exome sequence data. As of February 2015, there are 1553 genes from 288 submitters from 31 countries and 120 matches have been made enabling collaboration between clinicians and researchers from different countries and different backgrounds but with interest in the same genes. We performed a follow up of these matches by personal communication and while most are still in progress, at least 8 successful matches are being followed up by the Baylor-Hopkins Center for Mendelian Genomics.

Regression and overdispersion correction improves sensitivity of non-invasive prenatal testing

Lennart Johansson[1,2], Eddy de Boer[1], Freerk van Dijk[1,2], Dirk de Weerd[1,2], Ron Suijkerbuijk[1], Martin Elferink[3], Lieve Page-Christiaens[4], Heleen Schuring-Blom[3], Richard Sinke[1], Rolf Sijmons[1], Morris Swertz[1,2], Birgit Sikkema-Raddatz[1] and Gerard te Meerman[1]

[1]*Department of Genetics, University of Groningen and University Medical Center Groningen*
[2]*Genomics Coordination Center, University of Groningen and University Medical Center Groningen*
[3]*Department of Genetics, University Medical Center Utrecht*
[4]*Department of Obstetrics, University Medical Center Utrecht*

Amongst pregnant women, non-invasive prenatal testing (NIPT) is rapidly gaining popularity as a screening test for trisomy 13, 18 and 21 as it reduces the amount of unnecessary invasive procedures. NIPT is based on counting cell free DNA fragments, from both fetus and mother, in maternal blood. The fraction of DNA fragments originating from a specific chromosome is compared with the chromosomal fractions in a group of control samples. It is of utmost importance to have the highest possible sensitivity to prevent false negative results.
Even when laboratory procedures are optimal, the NIPT result can be improved by introducing statistical algorithms for reduction of systematic bias and trisomy prediction. Various methods have been reported to reduce sample to sample variability, such as peak correction and GC-correction [1,2] and to increase sensitivity in the Z-score prediction, such as the normalized chromosome value (NCV) [3], the median absolute deviation (MAD) based Z-score [4] and the Bonferroni method [2].
To further improve NIPT sensitivity we added a new prediction algorithm, the regression based Z-score, as well as a novel variation reducing method: the chi-squared based variation reduction.

All possible combinations of the above mentioned algorithms were tested on 128 non-trisomy and 42 trisomy samples sequenced on a SoLiD platform as well as on 164 non-trisomy and 6 trisomy samples sequenced on a HiSeq. For both platforms the sample to sample variation of the chromosomal fractions in non-trisomy samples were calculated as well as the different Z-scores of all samples.

Results show that chi-squared based variation reduction is able to further reduce variation in samples already corrected by other earlier described methods and that the regression based Z-score leads to a lower variability and thus a higher sensitivity for chromosomes 13, 18 and 21 compared to other prediction methods.

NIPT sensitivity can be improved using our statistical approach to reduce variability between samples. As a result NIPT will become a more powerful screening tool that is more suitable for trisomy prediction.

1. Chen, E.Z., et al. (2011).. *PloS One*, *6*(7), e21791.
2. Fan, H. C., & Quake, S. R. (2010). *PloS One*, *5*(5), e10439.

3. Sehnert, A. J. et al. (2011). *Clinical Chemistry*, *57*(7), 1042–9.
4. Stumm, M., et al. (2012). *Prenatal Diagnosis*, *32*(6), 569–77.

SESSION 12

Integrated genome and transcriptome sequencing of the same cell

Siddharth Subhas Dey

Hubrecht Institute, Utrecht, The Netherlands

Understanding the consequences of genomic variation on phenotypic heterogeneity is one of the central questions in biology. Single-cell genomics and single-cell transcriptomics have emerged as powerful tools to address these questions at a genome-wide scale. However, to directly correlate the genome to the transcriptome, a major challenge is to quantify both genomic DNA (gDNA) and mRNA from the same cell.

We have developed a novel method that allows simultaneous quantification of both the genome and transcriptome from the same cell. Importantly, this method does not involve physical separation of the nucleic acids prior to amplification, thereby minimizing losses and allowing us to achieve efficiencies similar to those attained in single-cell gDNA or mRNA sequencing. We applied this new technology to explore the correlation between DNA copy number variation and transcriptome variability among individual SK-BR-3 breast cancer cells. Interestingly, we found that genes that display more cell-to-cell variability in transcript numbers are generally associated with reduced copy number loci and vice-versa, implying that copy number variations could potentially drive variability in gene expression between single cells.

Finally, applying such integrated gDNA and mRNA single-cell sequencing approaches to other biological systems will elucidate direct functional relationships between the genome and transcriptome.

Genome in a Bottle: You may have sequenced, but how well did you do?

Justin Zook[1], Marc Salit[1], and Stephen Lincoln[2], on behalf of the Genome in a Bottle Consortium

[1]*National Institute of Standards and Technology, Gaithersburg, USA*
[2]*Invitae, San Francisco, USA*

Clinical laboratories, research laboratories and technology developers all need DNA samples with reliably known genotypes in order to help validate and improve their methods. The Genome in a Bottle Consortium (www.genomeinabottle.org) has been developing reference DNA standards with high-accuracy whole genome sequences to help support these efforts internationally.

Our pilot reference standard is based on Corriel sample NA12878. To minimize bias and improve accuracy, 11 whole-genome and 3 exome data sets produced using 5 different technologies were integrated using a systematic arbitration method [1]. Because differing variant representations make comparison between data sets difficult (both under-counting and over-counting discordances) we developed a methodology for regularizing data and we manually inspected many potentially discordant calls to produce the final reference data.

We are now adapting these methods to characterize 5 additional samples from 2 consented families, one Asian and one Ashkenazi Jewish. We are collecting a larger and even more diverse data set on these samples including high-depth Illumina, Complete Genomics, and Ion Torrent short-read data, as well as Moleculo, PacBio, and BioNano Genomics long-read data. These data will provide an accurate assessment of not just small variants but also large structural variants in both "easy" regions of the genome and in some "hard" regions (such as segmental duplications, low-complexity sequence, and hyper-variable loci). All of the input data sources are available for download and analysis by the community.

Our arbitration method produced a reference data set of 3,137,725 single nucleotide variants (SNVs) and 201,629 indels (insertions/deletions) covering 85% of the NA12878 genome. We also produced homozygous reference calls, indicating 2.195 billion locations that we confidently believe are invariant. We found that our integrated call set is highly sensitive and specific, showing 100% concordance with clinical lab sequences of genes in NA12878. We also compared our arbitrated calls to high-accuracy SNP genotypes, to fosmid sequences and to independent reference data sets produced using phased pedigree information, in all cases with high concordance. Preliminary analyses of our expanded data set suggests that dozens of medically relevant loci, currently in the "no-called" 15%, will indeed be filled in.

We combined the strengths of each of our input datasets to develop a comprehensive and accurate benchmark set of SNV, indel, and homozygous reference calls for most of NA12878's genome. In the short time it has been available, over 20 published or submitted papers have used our data. Diverse applications include the validation of germ-line sequencing tests as well as the development of spiked controls for tumor sequencing. The challenges that we encountered comparing across variant representations are also encountered by the end users of our reference data, and thus we are working with the Global Alliance for Genomics and Health to develop standardized methods, performance metrics, and software to assist in its use.

[1] Zook et al, Nature Biotech. 2014

The Global Variome

Sir John Burn MD FRCP FRCPE FRCPCH FRCOG FMedSci

Professor of Clinical Genetics, Newcastle University, Genetics Chair, National Institute of Health Research Biomedicine West, Centre for Life, Newcastle NE1 3BZ,UK
john.burn@ncl.ac.uk; john.burn@nuth.nhs.uk;@capp3

The need to pool data in order to understand its clinical significance has been obvious to all who have an interest in the medical application of DNA technology for decades. The complexity of the challenge coupled with the limitations of funding mechanisms and competing interests limited early progress to a few key genes like CFTR in cystic fibrosis. The emergence of the Genome project stimulated the community to support the Human Variome Project which made significant progress in terms of developing a common purpose and addressing issues around ethics, regulation and nomenclature. The concept of "country nodes" recognised the critical importance of involvement of the whole human population and not just those of North European extraction but the emphasis on academically led databases build around specific monogenic phenotypes is not scalable to deal with the surge of data from the more recent developments in genomics.

In 2014, a new initiative, the Global Alliance for Genomics and Health was launched to bring together genomicists to address the need to share knowledge from the growing collection of large scale datasets. At the first plenary assembly in March 2014, it was agreed to make the BRCA Challenge one of the flagship or "vanguard" projects, drawing on the experience of the Human Variome Project and the large numbers of diagnostic molecular geneticists who have yet to become fully engaged with the process of variant curation at the international level. The success of the InSiGHT database provided an obvious model. Over the last 12 months the ideas have crystallised and an international steering committee established. Sub-committees have set out to address the Ethics, Regulatory& Advocacy (ERA) issues and another is addressing data acquisition from the perspective of the need for an Applied Programme Interface (API) able to extract BRCA sequence variants at scale from genomic datasets and the important legacy data stored at national and individual laboratory level. The third committee draws heavily on the ENIGMA group and will form the basis of an international curation committee which will aim to rapidly resolve conflicts around variants considered to be pathogenic by some and labelled Variants of Uncertain Significance (VUSs) by others. With support from Astra Zeneca, whose selective PARP inhibitor Lynparza has now been licensed to treat ovarian cancer in relapse in people with a BRCA 1 or 2 mutation, the international curation effort will be led teams at University of Queensland, Australia and the Huntsman Institute in the USA.

It is hoped that bringing together the resources of the HVP and GA4GH, it will be possible to create BRCA nodes in every country able and willing to share variant data in such a way that we can eliminate the current time wasted by diagnostic laboratories providing individual reports and separately analysing the literature. Pooled data will allow the clinical significance of hypomorphic variants to be better characterised along with founder mutations across different ethnicities. In turn this should form a basis for expansion into a shared knowledge base for all the 100 genes involved in monogenic predisposition to cancer and ultimately to a sustainable international resource for the curation of the Global Variome.

Poster Presentations

M-PO1

Case report: Clinical manifestations of 4 patients with known genetic predisposition to Morbus Wilson and their affinity with genetic findings of alopecia, sclerodermia and lactose intolerance and others

Eva Augste[1], Zenon Lasota[2], Pavla Vanickova[1], Petra Drahosova[1], Eva Jaluvkova[1], Spiros Tavandzis[1], Radmila Richterova[1]

[1]*Laboratory of medical genetics, department of molecular biology, Laboratore AGEL a.s., Novy Jicin, Czech Republic*
[2]*Transfusion department, Hospital Novy Jicin, Novy Jicin, Czech Republic*
corresponding author: eva.augste@lag.agel.cz

We demonstrate thesaurismotic effect of frequency of phenotypic clinical incidence associated with age. Uniform clinical symptom of 4 patients attending ambulance of clinical immunology is association to neuroabnormalities. One man and three women – a child year of birth 2008, a miss year of birth 1985, a young woman born in 1973, and a older woman born in 1956.

All patients have genetic predisposition to Morbus Wilson (MW) - molecular genetic analysis was performed using the Sanger sequencing of selected exons of the gene ATP 7B. Two of predispositions are unknown variants (UV) - c.[3243+31C>A];[3243+31C>A] and c.[1441T>G];[=] not listed in any database, one is heterozygous pathogenic mutation c.[3207C>A];[=] and one heterozygous potentially pathogenic mutation c. [3320-3322del3];[=].
Next step was molecular genetic testing using real-time PCR method for alopecia (EDA2R, AR, rs 1160312, rs2180439, rs6625163 and rs1041668) and sclerodermia has also been using real-time PCR method for analysis of c.-945G>C polymorphism in the promoter region of CTGF gene (connective-tissue growth factor).

In all patients we observe predisposition to MW with mutations typical for alopecia, in three patients with predisposition to sclerodermia. In all patients we randomly found another predisposition for example to lactose intolerance, celiac disease, Crohn disease and hemochromatosis or narcolepsy.

We presume that the genetic predisposition to Morbus Wilson may just be negligible in area of the mental and neurological symptomatology.
We believe that the combination of tesaurismotic and malabsorption assumptions supported by mutations in CTGF gene for regenerating connective tissue and premise of tissue hypoxia is a major predisposing factor for the phenotypic clinical expression increasing with age.

T-PO2

Chromosome Microarray testing: Towards a decade of routine diagnostics

Darmanian AP & Peters GB

Cytogenetics Dept, CHW, Sydney NSW, Australia
Corresponding author:
artur.darmanian@health.nsw.gov.au

Microarray based comparative genomic hybridisation (aCGH) is widely used in cytogenetics laboratories for both pre- and postnatal constitutional genome analysis. For the latter it has been the recommended as the first line test in chromosome analysis (=molecular karyotype), since 2010 [1]. Our lab is part of a specialist paediatric hospital/tertiary referral centre, serving a population of c. 3 million people. In this role, we have been performing aCGH diagnostic tests for nine years. For the last seven of these, we have been using oligomeric CGH microarray platforms (of 60,000 probes), in our routine test procedure. In any diagnostic microarray service, decisions as to which Copy Number Variation (CNV) will be reported, and which will be ignored, are complex. They rely on considerations ranging from clinical genetics to population biology, and the *structural variant disease hypothesis* [2], as we have reviewed elsewhere [3].

In this context: this poster describes aCGH data for c.10,000 probands, 24% of whom carried a "reportable CNV": reported either as a pathological entity, or as a "variant of uncertain significance" (or VOUS). After follow up testing of available parental pairs, the figure of "24% reportable" CNVs was reduced to an estimated "16.5% likely pathological" CNVs. These figures are consistent with findings of other studies in the field, and more recent work suggests that these "likely pathological" CNVs comprise around 1/4 of all clinically significant

mutations, as detectable by whole genome sequencing [4].

Although the sun may soon be setting on the diagnostic microarray: we have learned much along the way, which should prepare us well for the greater adversities to follow. Other issues discussed here include turn-around times, sample flow, success rates, and some specific examples.

References:

1. Miller DT et al (2010) Am J Hum Genet 86:749-764
2. Vissers, LE, et al (2010) Nat Genet 42:1109-1112
3. Peters GB & Pertile MD, *in* Trent, R [ed] *Clinical Bioinformatics,* Ch8, pp117-155 Springer NY 2014
4. Gilissen C et al (2014) Nature: doi:10.1038/nature13394

M-PO3

GJB 2 Mutations in Patients with Nonsyndromic Hearing Loss from Croatia

Jasminka Pavelic[1], Ivona Sansovic[2]

[1] Rudjer Bošković Institute, Zagreb, Croatia, [2] Childrens Hospital, Zagreb, Croatia

Recessive mutations at the DFNB1 locus (13q11-1) are the cause of about 50% nonsyndromic hearing loss (NSHL). Diagnosis DFNB1 is confirmed when mutation in *GJB2* and *GJB6* genes, mapped at the DFNB1 locus is found. The aim of the present study was to: a) determine the carrier frequency of 35delG and 167delT mutations among 342 normal hearing persons, b) the frequency and type of mutation in the coding region of *GJB2* gene, c) the frequency of mutation IVS1+1G>A in the *GJB2* gene, d) the frequency of del(GJB6-D13S1830) in *GJB6* gene, d) genotype-phenotype correlation in 85 patients with NSHL from Croatia. The mutations were analyzed by PCR/RFLP method, sequencing, duplex PCR and MLPA analysis. 49.4% of the patients presented with mutation in *GJB2* gene. We identified ten sequence variations of which two were the novel variants. Similar to the most European populations, 35delG frequency was the highest (38.2%) in patients. The allelic frequencies of other mutations were 1.8% - 0.6%. The 35delG carrier frequency (1.5%) was similar to neighboring countries. The 35delG/35delG genotype was associated with severe to profound HL. High mutation rate indicates that testing of GJB2 gene will clarify the causes in almost half of the cases of recessive NSHL.

T-PO4

Gonadal mosaicism in Wiskott-Aldrich syndrome gene defect carrier

Varlamova T, *Kuzmenko N, Bobrynina V, Shcherbina A*

The Rogachev Federal Research Center for Pediatric Hematology, Oncology and Immunology, Moscow, Russia corresponding author: varltatwell@mail.ru

Wiskott-Aldrich Syndrome (WAS) is a monogenic X-linked primary immunodeficiency, caused by mutations in WASP gene, and characterized by thrombocytopenia, eczema, infections and high susceptibility to develop tumors and autoimmunity. Before the era of hematopoietic stem cell transplantation (HSCT) the average survival age of patients was 8 years.Though HSCT is curative in WAS, it is often accompanied by high morbidity.Therefore, in families with WAS prenatal diagnosis is a valuable option. Yet, since 30% of mutations in WASP gene are sporadic, until recently, females who were not germ-line carriers, were not suggested this option. Here we report a case with somatic mosaicism in a female carrier. The proband was a male patient, suffering from WAS caused by c.1046_1047insT (p.Ile349IlefsTer146) mutation of the WASP gene. The boy underwent HSCT and died due to its complications. His mother on several occasions has tested negative for WASP mutation in hematopoietic cells. Yet, during the next pregnancy, prenatal diagnosis was performed, and the male fetus was found to carry the same mutation in WASP, hence the family chose to terminate the pregnancy. During the third pregnancy via prenatal testing the female fetus was found to be a non-carrier, which was confirmed after the birth of a healthy girl. The evidence above suggests gonadal mosaicism in the mother. This phenomenon is important to consider in all families with X-linked disorders (including WAS), in whom prenatal diagnosis is recommended - irrespective of the mother's carrier status.

M-PO5

Significance of thrombophilic markers in Sickle cell disease

Students Salim Mohammed Al riyami, Ayman alwahaibi
Supervisors Dr Salam Alkindi, Dr Anil Pathare, Mr Shoaib Alzadjali

Sickle cell disease [SCD] is a genetic disorder of significant economic importance, and is highly prevalent in the Sultanate of Oman. The gene in its heterozygous state is present in 6% of the indigenous Omani population. One of the complications of SCD is a thrombophilic state associated with complex haemostatic abnormalities. Intermittant painful crisis often activate the haemostatic system with consumption and reduction of several haemostatic proteins.

Aim: To study the haemostatic alterations in SCD patients during clinical and sub clinical thrombotic manifestations where markers of thrombophilia were investigated.

Methods: We investigated 71 cases of SCD patients in steady state or following an episode of venous thrombosis. Plasma levels of several thrombophillic markers like Antithrombin III,[AT] Protein S and C, Activated protein C resistance, and genetic screening of Factor V Leiden, [FVL] MTHFR mutation(C677T) and CBS mutation(844ins68) was performed in 37 patients.

Results: We observed a significant reduction in Protein S and C in 18(48.6%)and 16(43.2%) cases respectively. Ten patients(27%) showed a reduction of both Protein S and C levels. No patient showed AT deficiency or FVL. MTHFR C677T mutation was seen in 16(43.2%) cases, whereas the CBS 844ins68bp was seen in 12(32.4%) cases in heterozygous state. Two patients had both the mutations. CBS mutation was significantly correlated with low Protein S levels, whereas the MTHFR mutation alone was not correlated with any other thrombophilic marker studied. However, presence of both CBS and MTHFR mutations was significantly correlated with Protein C deficiency.

Discussion: Although this is small study with analysis on 37 patients of SCD, it demonstrates variable alterations in several haemostatic markers of thrombophilia studied. It is important to understand the significance of these alterations as it is often difficult to assign whether these changes represent a cause or effect. However, in almost all cases, as AT levels were normal, it is likely to reflect an underlying cause rather than effect as one would have expected a reduction in the AT levels in case of consumption.

T-PO6

Correlation of Acute Chest Syndrome with e-NOS, ARG1 and GSNOR gene polymorphisms in Omani Sickle cell patients

Aiman Al Wahaibi, Salam Alkindi, Shoaib Al Zadjali, Anil Pathare

Background: Acute chest syndrome (ACS), is the most common pulmonary complication of sickle cell disease (SCD) and is associated with reduced nitric oxide(NO). NO induces vasodilatation and the helps recruitment of neutrophils. However, the T-786C polymorphism significantly reduces eNOS gene promoter activity, whereas the E298D changes an amino acid in the enzyme's oxygenase domain. The expression of eNOS is also related to the number of 27 bp repeat VNTRs in intron 4, with genotype 4bb repeats showing a decrease in eNOS expression whereas, homozygosity for the minor allele of GSNOR SNP rs28730619 is associated with increased risk of asthma.

Objectives: To correlate endothelial nitric oxide synthase (NOS3) gene polymorphisms (T-786C, E298D and Intron 4 VNTR) as well as ARG1 and GSNOR gene polymorphisms with ACS in SCD Omani patients.

Methods: Genomic DNA was isolated using the standard techniques and stored at -20^{0}C pending analysis. DNA sequence polymorphisms for HBB gene(**b**$^{6Glu>Val}$), NOS3 gene polymorphisms (T-786C, E298D and Intron 4 VNTR) as well as ARG1 and GSNOR gene polymorphisms were studied by direct sequencing of the relevant genomic segment amplified by polymerase chain reaction on an ABI PRISM 3100 genetic analyzer using appropriate primers described in literature.

Results: Our results showed that only the eNOS promoter C-786 allele showed a statistically significant association (P=0.001) in ACS cases especially so with the female gender (p=0.005). There was no correlation observed with eNOS polymorphisms E298D and Intron 4VNTR, ARG1 and GSNOR gene polymorphisms studied in this cohort of SCD patients with ACS.

Conclusion: eNOS promoter C-786 variant which reduces *eNOS* gene activity was observed as a genetic risk factor for ACS in adult female sickle cell anemia patients, explained by the fact that eNOS is known to be regulated by oestrogens.

M-PO7

Knowledge and Attitudes of Oman Medical Specialty Board Residents towards Evidence-Based Medicine

Aiman Al Wahaibi, Saada AL-Adawi, Wafa AL-Shehhi, Syed Gauhar A. Rizvi, Nasser Al-Kemyani, Khalfan Al-Amrani and Murtadha Al-Khabori

This study aims to evaluate the knowledge and attitudes of Oman Medical Specialty Board (OMSB) residents towards Evidence-Based Medicine (EBM).

Methods: This cross sectional study was conducted on all OMSB residents through a self-administered online questionnaire between October 2012 and March 2013. An electronic survey was designed to identify and determine residents' knowledge and attitudes toward the use of EBM.
Results: The survey was completed by 93 (21%) OMSB residents, 76 (82%) of whom took part in continuing education courses and 50 (54%) belonged to professional practice-oriented organizations. On average, the residents were reportedly involved in patient care for approximately 70% (Standard Deviation [SD] 17%) of their time, while 14% (SD 12%) participated in research activities. The results showed that 53 respondents (57%) were competent users of medical search engines compared to 23 residents (25%) who rated their skills as neutral. Sixteen percent of the respondents strongly agreed and 46% only agreed that the facility supports the use of current research in practice. Fourteen percent strongly agreed and fifty-three percent only agreed that the foundation of EBM is part of OMSB academic preparation. On the other hand, 17% of the respondents thought that insufficient time is always a barrier against EBM, while another 27% perceived insufficient time as a usual barrier. The lack of information resources was reported to always be a barrier in 11% of the respondents while 32% thought that it usually acts as a barrier.

Conclusion: Time constraints and skills in EBM were found to be the two major obstacles. This study was, however, limited by the low response rate of the survey; thus larger studies with a previously validated questionnaire should be conducted in the future.

T-PO8

Patterns of Similarity and Difference Among Clinical Interpretations from Multiple Laboratories in ClinVar

Shan Yang, Keith Nykamp, Yuya Kobayashi, Stephen Lincoln, and Scott Topper

Invitae, San Francisco, USA
Corresponding author: shan.yang@invitae.com

Purpose: Many germline variants observed in patients are quite rare, and individual laboratories often do not not have access to enough scientific evidence to characterize their clinical impact, if any. For example, in our own diagnostic practice using relatively small gene panels (typically 5 to 30 genes), we see apparently novel variants approximately one every other patient. This rate is not decreasing quickly even after testing thousands of patients, reflecting the known "long tail" of rare human variation. It is always possible that another laboratory may have seen some of these variants but not published the data. Hence, ClinVar (www.ncbi.nlm.nih.gov/clinvar/) has been established as one of the publicly available databases that facilitate sharing of variants and clinical interpretations between diagnostic laboratories. ClinVar has grown steadily, from 31,623 records to 149,954 in only 15 months. We sought to review the data in ClinVar to understand the implications of this growing resource on clinical testing practice.

Methodology: We compared ClinVar records for a list of 216 well-known, clinically relevant genes across a set of medical disciplines (cancer, cardiology, neurology, etc). We further focused on submissions from a list of well-established clinical testing labs that contributed over half of the entries for these genes. Finally, in this preliminary analysis we focused on 1502 variants which had been observed by our laboratory and at least one other ClinVar submitter, and thus are not very rare (a few of these may come from family members). ClinVar submissions from our own laboratory are not counted.

Results: 40% of these variants had ClinVar submissions from multiple labs. In these cases the submitted clinical interpretations are the same 2/3 of the time but are different for the remaining 1/3. Most discrepancies are 1-step differences on the 5-

class scale, most commonly in benign vs. likely benign classifications (17.5% of cases). However in 1.5% of cases the differences were in positive (pathogenic or likely pathogenic) vs. not positive (uncertain, likely benign or benign) results which could be quite significant in terms of clinical actions. 3 of these had both fully pathogenic vs. benign or likely benign interpretations. 7.5% of 119 variants with 3 or more submitters had 3 or more different interpretations.

Conclusion: A significant fraction of variants submitted to ClinVar in these genes have differing interpretations, including some with significant clinical implications. Further investigation is underway into these patterns for many more variants and genes in ClinVar. In addition, comparisons of data in ClinVar against our own blinded interpretations produced using the the very recently (January 2015) released ACMG interpretation guidelines are underway. These results should help the community understand the value and caveats in shared clinical variant databases.

M-PO9

Variant calling for clinical genomics on the cloud with a focus towards disorders of the endocrine system

Charlotte Anderson[1], Sehrish Kanwal[2], Richard Sinnott[2] and Andrew Lonie[1]

[1] Victorian Life Sciences Computation Initiative, The University of Melbourne, Australia
[2] Department of Computing and Information Systems, The University of Melbourne, Australia
corresponding author:
charlotte.anderson@unimelb.edu.au

Purpose: To facilitate the clinical implementation of genomic medicine by next-generation sequencing, it will be critically important to obtain accurate, consistent and reproducible variant calls on personal genomes. Many bioinformatic pipelines have been developed to call variants from NGS data, performance of these pipelines depends crucially on the calling strategies implemented, the alignment methods and the variant callers used.
Reproducibility depends on accurately recording the versions of the tools selected and the parameter setting used. We compare the concordance and discordance between three analysis workflows available for use in the endocrine virtual laboratory.
We explore the values of key metrics related to SNV quality produced from each workflow to use as indicators of performance; asking why some SNVs were called by one workflow, but not others and whether the SNV common to all workflows share traits.

Methodology: We sequenced 6 exomes from Disorders of sex development (DSD) patients using commercial kits, Illumina HiSeq 2000 platform and Agilent SureSelect version 2 capture kit. We analysed the raw data using 3 different bioinformatic pipelines: A) BWA-GATK-Haplotype caller, B) BWA-GATK-Unified Genotyper, and C) BOWTIE2-GATK-SAMtools/Unified Genotyper merge.
Pipelines were run using the Genomic Virtual Lab. as a framework for cloud-based genomics, facilitated by the Nectar Research cloud. Downstream analysis of variants was restricted to a list of genes known to manifest the DSD phenotype. To use as a truth set, Melbourne genomics health alliance sequenced a NA12878 sample from the Coriell institute and the Genome in a Bottle high confident calls were downloaded.

Results: SNV concordance between all 3 pipelines across all 6 exomes was 63.7% on average. We found that for novel SNVs (those not found in dbSNP v137) the overall concordance was 27.6%, much lower than the overall concordance between known SNVs at 72.3%. All pipelines were then benchmarked against a known truth set using data from 100 Genomes sample NA12878 from the Genome in a Bottle consortium.

Conclusion: The concordance between the three variant calling workflows is relatively low for novel variants. Each of the three workflows had merits, however integration of selected tools from the pipelines will improve efficacy of variants called. In future, truth sets of known variants should be used initially to gauge the efficacy of new and novel workflows.

T-PO10

De novo *ZIC2* variant in a child with global developmental delay and multiple congenital anomalies

Maggie Brett[1], Eileen Lim[1], Jiin Ying Lim[2], Angeline Lai[2] and Ene Choo Tan[1],

[1]KK Research Centre, KK Hospital, Singapore,
[2]Department of Paediatrics, KK Hospital, Singapore
corresponding author: Maggie.Brett@kkh.com.sg

Whole exome sequencing has proven to be an effective and cost-effective approach for the identification of causal variants in many children with developmental delay (DD) and multiple congenital anomalies (MCA). We describe the use of exome sequencing in a child with global DD, semi-lobar holoprosencephaly (HPE), microcephaly, eye anomalies, and bilateral hearing loss.

This child was recruited by the Genetic Services, KK Hospital and no abnormal findings were detected by chromosomal microarray. Exome sequencing was carried out using the Agilent SureSelect Exome enrichment kit and sequencing on an Illumina HiSeq. Raw reads were mapped using GATK and variant calling was performed by the Agilent GeneSpring software. Rare variants with population frequency of <1% in the NHLBI Exome Variant Server and 1000 Genomes databases were selected and prioritized for confirmation by Sanger sequencing. Confirmed variants were tested in parental samples.

Over 40 million reads were obtained with 83% occurring in the targeted regions. Approximately 82% of targeted bases had >20x coverage. A *de novo* c.667delC variant in *ZIC2* was confirmed in the child. This variant is novel and is predicted to cause a frameshift and premature protein truncation. *ZIC2* encodes a transcription factor that is important in neurological development and *ZIC2* mutations are common causes of non-syndromic HPE. A *de novo* p.Pro335Leu variant in *SOX5* was also detected in the child. The variant is novel and is of uncertain clinical significance. Hemizygous deletions involving *SOX5* have recently been shown to cause intellectual disability but no *SOX5* pathogenic mutations have been reported. Additional candidates are being followed up as possible causes of her hearing loss.

Exome sequencing has identified a novel *de novo* variant in *ZIC2* which is likely causative for her brain and eye anomalies. This case illustrates that exome sequencing is an efficient and cost effective way for identification of pathogenic variants in clinically heterogeneous disorders.

T-PO11

Which DNA capture method for 2nd generation sequencing in clinical practice?

Baux D.[1], Garcìa-Garcìa G.[2], Koenig M.[1,2], Claustres M.[1,2], Roux A-F.[1,2]

[1]Laboratoire de génétique moléculaire, CHU Montpellier, France
[2]Laboratoire de génétique de maladies rares, Université de Montpellier, France
corresponding author: david.baux@inserm.fr

Along with exome and genome sequencing, gene panels represent a powerful method to explore patients in diagnostic laboratories. There are several technical and financial advantages as larger series of patients can be studied in a single run with a higher coverage and gene panels can be run on medium throughput instrument. Moreover, gene panels may reduce the ethical issue of incidental findings. DNA capture in solution is the most widely used method for 100kb - 5Mb panels, and at least three major suppliers provide different approaches. Some studies have been published, but these methods evolve rapidly, and this work focuses on two recent methods released in 2014, Illumina Nextera Rapid Capture Custom Enrichment, and Agilent SureSelect QXT, which will be compared with an efficient-proven method, Nimblegen SeqCap EZ Choice. The panel of genes to be compared is the same for the three methods, and includes 115 genes involved in deafness and blindness, representing in total 800 kb. As for many panels, most of the selected sequences are exonic and includes flanking intronic sequences (+/-50 bp). The target design includes repeated sequences, which are dealt with different criteria by the manufacturers. The library preparation will be analysed from the lab position, and classical metrics from the run will be compared, including the quality of the raw data, the on-target reads/bases proportions, the mean coverage and the coverage homogeneity, but also the ability of each method to correctly capture mutated regions, such as short indels, which are a frequent cause of inherited diseases.

T-PO12

GensearchNGS: Interactive variant analysis

Beat Wolf[1, 2], Pierre Kuonen[1], Thomas Dandekar[2], David Atlan[3]

[1]HES-SO Fribourg, Switzerland [2]University of Würzburg, Germany [3]Phenosystems SA, Belgium
Corresponding author: beat.wolf@hefr.ch

NGS data analysis is increasingly popular in the diagnostics field thanks to advances in sequencing technologies which improved the speed, quantity and quality of the produced data. Due to those improvements, the analysis of the data requires an

increasing amount of technical knowledge and processing power. Several software tools exist to handle these technical challenges involved in NGS data analysis. We present the latest improvements in one of those software-tools, GensearchNGS 1.6, a NGS data analysis software allowing users to go from raw NGS data to variant reports. We focus on the improvements made in terms of variant calling, annotation and filtering. The variant calling algorithm has been completely rewritten, based on the variant calling model used in Varscan 2, greatly improving its speed (over 10 times faster than Varscan 2, over 5 times faster than GATK) and accuracy, while reducing memory requirements. For the subsequent annotation of the called variants, various new data-sources have been integrated, such as Human Phenotype Ontology and the clinical predictions from Ensembl, which give the user more information about the clinical relevance of the called variants. An initial prototype of the integration of interactome data from different sources, such as CCSB or BioGRID, is also presented, further increasing the available information for variant effect prediction. The addition of annotation data has been accompanied by various optimizations, keeping memory requirements and analysis times stable. The interactive variant filtering, which updates a variant list presented to the user while he changes the filters, has been further optimized, making it possible to filter variants interactively even on computers with limited processing power and memory. Similar improvements have also been made to the visualizer, allowing for a faster visualization requiring fewer resources, while integrating more data, such as the previously mentioned databases.

T-PO13

Distribution of Allele and Genotype Frequencies of Mdr1 Gene C3435T Polymorphism in Three Arab Populations

Abdel Halim Salem[1], Muhalab Ali[2], Amir Ibrahim[3], Mohamed Ibrahim[4]

[1]Anatomy, [2]Medical Biochemistry, College of Medicine and Medical Sciences, Arabian Gulf University, Manama, Bahrain, [3]Central Laboratory, Ministry of Science and Technology, [4]College of Animal Production Science and Technology, Sudan University of Science and Technology, Khartoum, Sudan
Corresponding author: ahaleemfd@agu.edu.bh

Objectives: The aims of this study were to determine the genotype and allele frequencies of MDR1 gene C3435T polymorphism in three Arab populations (Bahrainis, Jordanians and Sudanese), and to compare the results with the frequencies established in various ethnic groups.
Methods: Genotyping was carried out on 184 unrelated Bahraini, 116 Jordanian and 131 Sudanese subjects. The genotypes of polymorphic position C3435T were determined by polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) assay.

Results: Results showed that 34.8% of the studied Bahraini subjects were homozygous for the CC genotype, 45.7%% were heterozygous for the CT genotype and 19.5% were homozygous for the TT genotype. Whereas 20.7% of Jordanians were homozygous for the CC genotype, 51.7% were heterozygous for the CT genotype and 27.6% were homozygous for the TT genotype. Among Sudanese subjects, the genotype frequencies were: CC 52.7%, CT 42.0% and TT 5.3%. The frequencies of the 3435T variant in the MDR-1 Gene among Bahrainis, Jordanians and Sudanese were found to be 0.42, 0.53 and 0.26, respectively. According to the distribution of the C3435T polymorphism, Bahrainis and Jordanians were resemble Asians and Europeans but were different significantly from Sudanese, while Sudanese were similar to Africans.

Conclusion: In conclusion, the observed distributions of the C3435T polymorphism in the three Arab populations studied were within the range detected in other populations. The data obtained may give the basis for predicting effects of drugs that are substrates for MDR-1 in these populations and may be useful for individualized therapy of some diseases and in epidemiological studies of the *MDR-1* gene variation.

T-PO14

Distinctive AGG interruption patterns in the FMR1 gene in the UAE population

Khaled Amiri[1]*,Carolyn M. Yrigollen[2], Stefan Sweha[2], Huda Shaheen[1], Flora Tassone[2,3]*

[1]Department of Biology, College of Science, UAE University, Alain, UAE.
[2]Department of Biochemistry and Molecular Medicine, University of California Davis, School of Medicine, Davis, CA, USA
[3]M.I.N.D. Institute, University of California Davis Medical Center, Davis, CA, USA
corresponding author: ftassone@ucdavis.edu; k.amiri@uaeu.ac.ae

The United Arab Emirates (UAE) is located in the eastern part of the Arabian Peninsula, extends along part of the Gulf of Oman and the southern coast of the Arabian Gulf. The cultural influences such as consanguineous marriages may have influenced the genetic of population structure. The nature of the population structure may call for a modified disease risk assessment and genetic counselling. This research aims to study the distribution pattern of the AGG interruptions within the CGG repeat element of *FMR1* gene in the UAE population. The expansion of trinucleotide repeats is associated with multiple clinical manifestations including fragile X syndrome, fragile X-associated tremor/ataxia syndrome, and primary ovarian insufficiency [1]. Recently we demonstrated differences in the distribution of AGG interruption patterns within nine world populations including UAE population [2]. The presence of AGG interruptions within the CGG repeat is associated with reduction of expansion to a full mutation during transmission, thereby reducing risk [3]. The UAE population (n=263) presents 77 different AGG interruption patterns out which 21 are only observed in the UAE. The distribution of the UAE-specific alleles appears to be specific to UAE regions and their mapping to population structure may be of clinical significance in genetic counselling and risk assessment. Furthermore, the study will serve as an index genetic homogeneity or diversity and it will complement our recent study on Y chromosome diversity in UAE population (unpublished data).

References

1. Maddalena, Anne, Carolyn Sue Richards, Matthew J. McGinniss, Arthur Brothman, Robert J. Desnick, Robert E. Grier, Betsy Hirsch et al. "Technical standards and guidelines for fragile X: the first of a series of disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics." *Genetics in Medicine* 3, no. 3 (2001): 200-205.

2. Yrigollen, Carolyn M., Stefan Sweha, Blythe Durbin-Johnson, Lili Zhou, Elizabeth Berry-Kravis, Isabel Fernandez-Carvajal, Sultana MH Faradz et al. "Distribution of AGG interruption patterns within nine world populations."*Intractable & rare diseases research* 3, no. 4 (2014): 153.

3. Yrigollen, Carolyn M., Blythe Durbin-Johnson, Louise Gane, David L. Nelson, Randi Hagerman, Paul J. Hagerman, and Flora Tassone. "AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome." *Genetics in Medicine* 14, no. 8 (2012): 729-736.

T-PO15

Allele Frequencies of childhood-onset Acute Lymphoblastic Leukaemia-associated SNPs in a Mexican Mestizo Population and their relationship with disease presentation

Aguilera-Guerrero Julia Cecilia[1], Ramírez-Ramírez Edgar Antonio[1,2], Valdes-Morales Karla Leticia[1,2], García-Marín Ana Yuritzen[2], Cerón-Trujillo Berenice[2], González-Galarza Francisco Faviel[1,2], Argüello-Astorga Jesus Rafael[1,2].

[1]Facultad de Medicina, Universidad Autónoma de Coahuila, México. [2]Instituto de Ciencia y Medicina Genomica, México.

Objective: To asses the frequency of the Acute Lymphoblastic Leukaemia-associated SNPs in a well characterized Mexican population, in comparison with the MEX Hap-map population and correlate it with the frequency of the disease in México.

Methodology: In this study, we extracted DNA from saliva samples of 50 healthy volunteers and analyzed them in a microarray plataform with a customized Illumina Omin-expreess chip. 32 SNPs related with Childhood onset Acute Lymphoblastic Leukaemia were selected from NCBIs GWAS catalog (rs3824662 rs10828317, rs10821936, rs6964969, rs4982731, rs7142143, rs17079534, rs17837497, rs10170236, rs6683977, rs1496766, rs9958208, rs7578361, rs41322152, rs546784, rs4132601, rs7089424, rs2239633, rs10821936, rs11978267, rs2089222, rs11155133, rs2191566, rs7554607, rs12621643, rs10873876, rs9290663, rs6428370, rs1881797, rs563507, rs10849033, rs1879352), and the Allele frequencies of them compared with the MEX Hap-map population. The SNPs in which we didn't have the allele frequencies of both datasets were excluded, as well as those in which the Risk allele had a lower frequency than the MAF. A t-student test was performed to compare the allele differences in both populations. Also, Frequency analysis was performed to count the SNPs in Hardy-Weinberg disequilibrium and denote the ones in which the risk allele had a higher frequency than the wild type.

Results: The allele frequencies in our study group were consistent with the MEX Hap-map population; without a significant variation in 9 out of 10 SNPs. From the studied SNPs, 5/10 were not in Hardy-Weinberg equilibrium and of those 5, only 1 had

higher frequencies on the risk allele in comparison with the wild type allele.

Conclusions: Our small cohort closely resembles the one reported in the Hap-Map proyect of Latin Americans in Los Angeles with Mexican Ancestry (MEX). The Mexican population has been reported to have a higher incidence of Childhood-onset Acute Lymphoblastic Leukaemia in comparison with Worldwide incidence. The fact that half of the SNPs are not in Hardy-Weinberg equilibrium and only one of those has a higher frequency in the risk allele than in the wild-type one does not explain the high incidence of the disease in Mexican population. The low coverage and the small number of subjects in both cohorts may not help in making a statistically significant result, but denote the importance of making a larger case-control study in a Mexican population in order to demonstrate the existence (or non existence) of a derivative genetic cause of this disease.

Selected references:
Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 20/02/2015.
Rendón-Macías ME, Reyez-Zepeda NC, Villasis-Keever MA, Serrano-Menses J, Escamilla-Nuñez A. Global trend of survival in pediatric acute lymphoblastic leukemia: a review of the last four decades. Bol Med Hosp Infant Mex. Vol. 69, Mayo-Junio 2012

T-PO16

Describing complex rearrangements using HGVS sequence variation nomenclature, suggested extensions

Peter Taschner[1] and Johan den Dunnen[2]

[1]Generade Center of Expertise Genomics, University of Applied Sciences Leiden,
[2]Depts of Human and Clinical Genetics, Leiden University Medical Center, Leiden, Nederland
taschner@generade.nl

Breakpoints involved in translocation and chromothripsis are traditionally described using ISCN nomenclature based on chromosomal banding patterns (1). The sequence variation nomenclature guidelines of Human Genome Variation Society (HGVS, http://www.hgvs.org/mutnomen) traditionally focused on simple variants not requiring specific rules for detailed description of genetic rearrangements. This changed with the introduction of new technologies allowing rapid discovery of breakpoint sequences from complex structural rearrangements including translocations. The description of such complex variants challenges the existing guidelines. Previously, we suggested extensions for simple translocations (2). Here, we suggest extending the HGVS nomenclature guidelines for unambiguous descriptions of more complex structural rearrangements including chromothripsis. Main requirement: precise chromosomal breakpoint sequences should be derived easily for the descriptions. Their format should provide sufficient flexibility and consistency limiting alternative interpretations and ambiguous descriptions. The new rules can be combined with those proposed previously for complex changes, which included: i) nesting to support description of changes within inversions and duplications, ii) composite changes to support concatenation of inserted sequences (3). We have applied the rules in practice by describing complex cases involving many breakpoints. The specifications should allow easy implementation in sequence variant nomenclature checkers (e.g. Mutalyzer, https://Mutalyzer.nl). We are planning to extend Mutalyzer's functionality to incorporate the latest version of the HGVS sequence variation nomenclature guidelines as part of the development of curational tools for gene variant databases.

1) ISCN (2013). 2013. An International System for Human Cytogenetics Nomenclature. Shaffer LG, McGowan-Jordan J, Schmid M (eds). Basel: Karger.
2) http://www.hgvs.org/mutnomen/SVtrans_HGVS2013_PT.pdf
3) Taschner PE, den Dunnen JT. Hum Mutat. 32:507-511 (2011).