HGVS

HUMAN GENOME VARIATION SOCIETY

Copy Number Variation: data collection and analysis

Room J2
Gothia towers
Gothenburg, Sweden

14th June 2019

Scientific Programme Committee

Christophe Béroud (Marseille, France)
Johan T. den Dunnen (Leiden, Netherlands)
Marc Greenblatt (Burlington, VT, USA)
Gary Saunders (Hinxton, UK)

Organising Secretariat

Rania Horaitis (Meeting Makers | www.meeting-makers.com)

Introduction

Copy Number Variations or CNV are major causative contributors in the genesis of rare and common genetic diseases as well as human tumors. Despite their importance, their detection, collection, annotation and interpretation are still in their youth and much remains to be done. With the developments of the Next-Generation Sequencing technologies, especially Whole Genome Sequencing, which is increasingly becoming the primary choice for genomic screening analysis, new algorithms to efficiently detect all CNV types ranging from single exon to large genomic regions, and from germline to somatic events, are needed.

The aim of this Human Genome Variation Society meeting is to provide an overview of the most recent advances on this topic at the international level and to address more specifically;

CNV detection pipelines
availability of reference datasets
data collection and sharing in FAIR environments, including specifications related to data formats and annotations required for efficient data interpretation; and
specific tools for CNV interpretation.
To speed-up international collaboration, a new Human Copy Number Variations Community (h-CNV) was created within ELIXIR. It will work in close interaction with the Global Alliance for Genomic and Health (GA4GH), the European Joint Project for Rare Diseases (EJP-RD) and other international initiatives.

8:00 - 9:00 **Registration**

9:00 - 9:05 **Welcome**

Christophe Béroud

Session 1 Chair: Christophe Béroud

9:05 - 9:50 **KEYNOTE SPEAKER**

ELIXIR, data federation, and the Human Copy Number Variation Community

Gary Saunders
ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

9:50 - 10:15 **KEYNOTE SPEAKER**

Annotation and display of structural variation data in Ensembl

Sarah Hunt
Ensembl, EMBL-EBI, Hinxton, Cambridgeshire, UK

10:15 - 10:25 **Presentation from selected Abstract**

Towards an optimised workflow for the identification of CNVs from heterogeneous WES data, within the scope of the SolveRD project

Steven Laurie
Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

10:25 - 10:35 **Presentation from selected Abstract**

Assessing performance of copy number variation detection tools using simulated data

Iria Roca Otero
Genomes and Disease Group, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela, Spain

10:35 - 10:40	**Rapid Fire Virtual Poster Presentations (*non - CNV*)** • **AnnotSV 2.0: Annotation and Ranking of Human Structural Variations** - *Veronique Geoffroy* • **Human genetics associated with dengue severity in Thailand** - *Unchana Arayasongsak*
10:40 - 11:15	Coffee Break
Session 2	**Chair: Michael Baudis**
11:15 - 12:00	**KEYNOTE SPEAKER** **hCNV in the context of the French Genomic Medicine 2025 Plan and the national BANCCO database** Damien Sanlaville *Claude Bernard University, Lyon, France*
12:00 - 12:10	**Presentation from selected Abstract** **Detection of copy number variations using whole-exome sequencing improves diagnostic yield of patients with rare Mendelian diseases** Paulo Silva *CGPP-IBMC – Centro de Genética Preditiva e Preventiva, Instituto de Biologia Molecular e Celular and i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal*
12:10 - 12:20	**Presentation from selected Abstract** **New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis** Maarja Lepamets *Estonian Genome Center, Institute of Genomics, and Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia*

12:20 - 12:30 **Presentation from selected Abstract**

Identification of mobile retroCNVs during genetic testing: consequences for routine diagnostics

Nicolas Chatron
Service de génétique, CHU Lyon and Equipe GENDEV, CRNL, INSERM U1028, CNRS UMR5292, UCBL1, Lyon, France

12:30 - 13:45 Lunch

Session 3 **Chair: Gary Saunders**

13:45 - 14:30 **KEYNOTE SPEAKER**

CNV detection from exon capture data

Anna Benet-Pages
MGZ- Medical Genetics Centre, Munich, Germany

14:30 - 14:40 **Presentation from selected Abstract**

Detection of copy number variations from NGS data using read depth information: a diagnostic performance evaluation

Olivier Quenez
Dept. of Genetics and CNR-MAJ, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France

14:40 - 14:50 **Presentation from selected Abstract**

Most rare and high-risk CNV carriers do not have major health, cognitive or socioeconomic consequences

Elmo Saarentaus
Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland

14:50 - 15:00 **Presentation from selected Abstract**

Copy number variation detection tool for targeted sequencing data

Ashish Singh
Dept. of Medical Genetics, St. Olavs Hospital and Norwegian University of Science and Technology, Trondheim, Norway

15:00 - 15:20 Coffee Break

Session 4 **Chair: Anna Benet Pages**

15:20 - 16:05 **KEYNOTE SPEAKER**

Association analysis of SVs/CNVs using NGS data

Victor Guryev
European Research Institute for the Biology of Ageing (ERIBA), University Medical Center, Groningen, The Netherlands

16:05 - 16:50 **KEYNOTE SPEAKER**

Implementation Driven Development of Standards for Genomic Data Exchange from Cancer Genome Data Collections

Michael Baudis
Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland

16:50 - 17:00 Closing Remarks

17.00 **MEETING END**

Session 1

ELIXIR, data federation, and the Human Copy Number Variation Community

Gary Saunders

ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

Over the last forty years, we have seen the emergence of large cohorts of human samples from research and national healthcare initiatives. Many countries in Europe now have nascent personalised medicine programmes meaning that human genomics is undergoing a step change from being a predominantly research-driven activity to one funded through healthcare. This is evidenced by the recent Declaration of 20 European countries to sequence and share transnationally at least 1M human genomes by 2022. This initiative will catalyse the transition of genomics from the bench to bedside in Europe. We envisage that a significant subset of these data will be made available for secondary research. However genetic data generated through healthcare is not likely to be shared as widely as research data. Healthcare is subject to national laws, and it is often unacceptable for health data from one country to be exported outside regional or national jurisdictions. Our vision for the ELIXIR human Copy Number Variation Community is to contribute to the ELIXIR federated ecosystem of interoperable services that enables population scale genomic and biomolecular data to be accessible across international borders accelerating research and improving the health of individuals resident across Europe.

In this presentation I shall describe our work within the ELIXIR Human Data Communities - including the Human Copy Number Variation Community - which coordinates the delivery of FAIR compliant metadata standards, interfaces, and reference implementation to support the federated ELIXIR network of human data resources. The result will be a coordinated bioinformatics infrastructure across Europe that enables the transnational access for approved researchers to 1M genomes by 2022.

Annotation and display of structural variation data in Ensembl

Sarah Hunt, Irina Armean, Laurent Gil, Diana Lemos, Andrew Parton, Helen Schuilenburg, Anja Thormann, Fiona Cunningham, Paul Flicek*

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

seh@ebi.ac.uk

Structural variants, by virtue of their size alone, are more likely to have phenotypic impact than short variants. Despite challenges in discovery and analysis which make them more complex to study, increasing numbers of structural variants, in particular copy number variants, have been shown to be implicated in disease. Importantly, population frequency data is also now available.

In the Ensembl project we empower genomic research by aggregating and analysing large scale public data sets and creating comprehensive genomic views. We also provide stable programmatic access to these data. Here we will discuss how structural variation data is integrated into the Ensembl system, annotated and displayed alongside other genomic features in our browser.

The Ensembl Variant Effect Predictor (VEP) is a powerful toolset for the analysis, annotation, and prioritisation of genomic variants. It utilises our regularly updated transcript and regulatory annotations to identify variants which are more likely to impact gene function. VEP is designed to be highly flexible and allow simple extension of functionality. We will describe how it can be used for the annotation of structural variants.

http://www.ensembl.org/

Towards an optimised workflow for the identification of CNVs from heterogeneous WES data, within the scope of the SolveRD project

Steven Laurie[1], Gemma Bullich[1], Lennart Johannson[3], German Demidov[4], Francesco Musacchia[5], Sergi Beltran[1,2], Solve-RD CNV Working Group*

[1] Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3] University Medical Centre, Groningen, Holland
[4] EKUT, Tübingen, Germany
[5] Telethon Institute of Genetics and Medicine (TIGEM), Naples, Italy

**steven.laurie@cnag.crg.eu*

SolveRD is a EU Horizon2020 funded project with the key goal of solving the unsolved rare diseases. One of the primary goals is the comprehensive reanalysis of 19,000 unsolved cases for which exome sequencing (WES) has previously been undertaken, without successful resolution. There are many reasons why resolution may be unsuccessful following WES analysis, one of which is that CNV detection may not have been attempted, or if attempted, the results inconclusive. Therefore within SolveRD we wish to design a pipeline that can rapidly identify rare CNVs that may be clinically relevant.

WES data poses a particular challenge for CNV identification due to the large variation in target capture efficiency across the exome, which hinders efficient normalisation and hence signal detection. Furthermore the signal-to-noise ratio varies between exome enrichment kits, and the amount of sequencing undertaken. Within SolveRD we are receiving data for re-analysis from more than 20 different exome enrichment kits, with sequencing undertaken at a variety of different centres, representing a very heterogeneous dataset.

A large number of bioinformatic tools have been developed to attempt to detect CNVs from WES data, but they tend to display very discordant results, making interpretation challenging. Here we describe in more details the specific challenges we face, and how we are attempting to overcome them within the scope of a large-scale rare disease consortium.

Assessing performance of copy number variation detection tools using simulated data.

Iria Roca Otero[1]; Lorena González-Castro[2]; Helena Fernández[2]; Ana Fernández-Marmiesse[1].*

1: Genomes and Disease Group, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela (Santiago de Compostela, Spain).
2: Galician Research and Development Center in Advanced Telecommunications, GRADIANT (Vigo, Spain).
** Corresponding author: iria.roca@rai.usc.es*

Abstract

Since the development of next-generation sequencing (NGS) technologies, several research groups have been developing copy number variation (CNV) detection tools that compare depth of coverage (DoC) patterns between samples. Most of these CNV detection tools based on DoC comparisons are designed to work with whole-genome sequencing (WGS) or whole-exome sequencing (WES) data. However, few methods developed to date are designed for targeted NGS (tg-NGS) panels, the assays most commonly used for diagnostic purposes. These panels, designed to explore a much smaller proportion of the genome than WGS or WES, offer much greater coverage in the interrogated regions, allowing more accurate and sensitive detection of disease-related small CNVs encompassing one or more exons of one or multiple genes. Tools that can detect small CNVs with a resolution as high as one small exon are essential to effectively tackle diagnostic problems. Moreover, the development and evaluation of these tools is hindered by (i) the scarcity of thoroughly annotated data containing CNVs and (ii) a dearth of simulation tools for WES and tg-NGS that mimic the errors and biases encountered in these data. Ideally, their performance should be evaluated using a large number of validated CNV-positive controls, but these are often unavailable. Although several NGS simulation tools have been developed, none can fully reproduce the complexity of sequencing data generated by NGS technologies. An additional limitation is that most simulation tools were designed to generate WGS data. Very few NGS simulation tools generate synthetic tg-NGS data that faithfully reproduce the specific biases and errors resulting from the restriction of the genome regions being sequenced. In addition to the biases and errors inherent to any NGS experiment, tg-NGS data suffer from other problems related to the size of the sequenced regions and the capture library. These problems need to be reproduced to optimize the development and evaluation of genetic data analysis tools, such as CNV detectors

Here, we demonstrate how to generate simulated tg-NGS data with artificial CNVs by combining two different simulation tools and compare the performance of the most commonly used CNV detection methods using simulated datasets.

AnnotSV 2.0: Annotation and Ranking of Human Structural Variations

Véronique Geoffroy[1*], *Audrey Schalk*[2], *Arnaud Kress*[3], *Hélène Dollfus*[1,4], *Sophie Scheidecker*[1,2]*, *Jean Muller*[1,2]

[1]Laboratoire de Génétique Médicale, U1112, INSERM, IGMA, Université de Strasbourg, Strasbourg, France. [2]Laboratoires de Diagnostic Génétique, IGMA, Hôpitaux Universitaires de Strasbourg, Strasbourg, France. [3]Complex Systems and Translational Bioinformatics, ICube, UMR 7357, University of Strasbourg, CNRS, Strasbourg, France. [4]Centre de référence pour les Affections Rares en Génétique Ophtalmologique, Filière SENSGENE, Hôpitaux Universitaires de Strasbourg, Strasbourg, France.

*: veronique.geoffroy@inserm.fr

Background

Next generation sequencing and array-based techniques are generating a tremendous amount of data including many single nucleotide variations (SNV), small insertions/deletions (indel) and structural variations (SV). However, despite their important role either in genome evolution or in disease pathogenicity, SV are still difficult to reliably detect and annotate.

Objective

To help identifying pathogenic SV in the genome of patients, we have developed AnnotSV.

Results

AnnotSV is a fast and efficient tool to annotate and classify SV identified from the human genome. AnnotSV supports either the VCF (Variant Call Format) or the BED (Browser Extensible Data) formats as input files. Our tool aims at providing annotations useful to i) filter out potential false positive variants from all the SV identified and ii) to interpret SV potential pathogenicity. In particular, AnnotSV can integrate the heterozygous and homozygous counts of called SNV/indel overlapping each SV for the patients of interest. This information can be useful to support or question the existence of a single SV. We also report a computed and unique allelic frequency relative to all benign overlapping SV from the Database of Genomics Variants (DGV, http://dgv.tcag.ca) that is especially powerful to filter out common SV.

Since the publication of AnnotSV in Bioinformatics [1], new substantial developments have been made. First, we have doubled the number of annotations available (60), including along the DGV, the OMIM (https://www.omim.org), the DDD (https://decipher.sanger.ac.uk/) data already included, the ClinGen (haploinsufficiency, triplosensitivity), the ACMG gene list (https://www.acmg.net/), the CNV intolerance score from ExAC (http://exac.broadinstitute.org), the pathogenic SV from dbVar (https://www.ncbi.nlm.nih.gov/dbvar/) and the enhancers/promoters from GeneHancer. Second, in order to help the clinical interpretation of the SV, AnnotSV now provides a systematic classification of each SV into one of the following classes: class 1 (benign), class 2 (likely benign), class 3 (variant of unknown significance), class 4 (likely pathogenic) and class 5 (pathogenic). This new ranking algorithm was implemented based on a dataset of ~23000 SV from ~2500 patients of which 380 were disease causing. Finally, a new web server interface is available to annotate and rank the SV online.

AnnotSV is freely available at the following address: https://lbgi.fr/AnnotSV. The website is well documented with several sections (annotations sources, ranking…), a detailed readme, training resources and a user-friendly web server interface.

Conclusion

We believe the substantial functionally, regulatory and clinically relevant annotations, the integrated SV classification and the new user-web server interface will be of great importance for the audience interested in human genomics including medical geneticists, cytogeneticists, bioinformaticians and genomics scientists.

Human genetics associated with dengue severity in Thailand

Unchana Arayasongsak[a], Jun Ohashi[b], Izumi Naka[b], Jintana Patarapotikul[a], Thareerat Kalambaheti[a], Pornlada Nuchnoi[c], Areerat Sa-Ngasang[d], and Suwanna Chaorattanakawee[e*]

[a]*Department of Microbiology and Immunology, Faculty of Tropical Medicine, Mahidol University,Ratchawithi Road, Bangkok 10400, Thailand*
[b]*Laboratory of Human Genome Diversity, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan*
[c] *Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*
[d] *National Institute of Health, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand*
[e] *Department of Parasitology and Entomology, Faculty of Public Health, Mahidol University,Ratchawithi Road, Bangkok 10400, Thailand*
E-mail: Unchana.ar@gmail.com, *Suwanna.cho@mahidol.ac.th

Dengue is a major life-threatening disease caused by dengue virus[1]. Patients develop different severity ranging from acute febrile to dengue hemorrhagic fever (DHF) and to the most severe form: the dengue shock syndrome (DSS)[2]. Though it has not been clear what causes patients to develop different clinical symptoms, evidences have been found suggesting that host genetics are responsible for the outcome of dengue severity[3,4,5]. However, the information of the genetics that are susceptible to severe dengue is still limited. Here we analyzed the genetic polymorphism of the *IFNL1*gene (rs30461 and rs7247086) regarding their association with dengue severity in Thai population. A case-control association study was performed in 877 patients who were less than 15 years old. They were divided into three groups; DF (n=386), DHF (n=416), and DSS (n=75). Genotyping were done by TaqMan real-time PCR assay.

We found that only SNP rs7247086 of *IFNL1* was associated with DHF but not with DSS. The genotype CT/TT and T allele were shown to be protective against DHF (P= 0.03, OR = 0.62 for CT and OR = 0.13 for TT; and P = 0.01, OR = 0.54 for T allele). Since genetic susceptibility/ protection were found to be different in DHF and DSS groups, the results suggest that DHF and DSS are two different diseases that have their own causes and pathogenesis. Furthermore, *in silico* analysis of rs7247086 demonstrated that polymorphism affects the binding of transcription factor and thus may affect the level of gene expression. This emphasizes that contribution of *IFNL1* to DHF may arise by the increased/reduced level of the cytokine and not by the change of cytokine property since association was not found with SNP rs30461 of *IFNL1.*

Key words: Dengue hemorrhagic fever, *IFNL1* gene, SNP, association

Acknowledgments: This study was supported by the Royal Golden Jubilee Ph.D. Programme (PHD/0169/2556). We thank the Faculty of Tropical Medicine, Mahidol University and Department of Biological Sciences, the Graduate School of Science, the University of Tokyo.

References:

1. Guzman *et al*. (2010) Dengue: a continuing global threat. Nature Reviews Microbiology, 8:S7.
2. World Health Organization & Special Programme for Research and Training in Tropical Diseases. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control. *WHO* (2009)
3. Sa-Ngasang *et al.*(2014) Association of IL1B -31C/T and IL1RA variable number of an 86-bp tandem repeat with dengue shock syndrome in Thailand. *J Infect Dis, 210*(1), 138-145.
4. Stephens *et al*. (2010). HLA and other gene associations with dengue disease severity.Curr Top Microbiol Immunol, 338, 99-114.
5. Sakuntabhai *et al.* (2005) A variant in the CD209 promoter is associated with severity of dengue disease. *Nat Genet, 37*(5), 507-513.

Session 2

hCNV in the context of the French Genomic Medicine 2025 Plan and the national BANCCO database

Damien Sanlaville,

Head of Genomic department, Lyon University Hospital
Medical director of AURAGEN lab (French Genomic Medicine 2025 Plan)

Human Copy Number Variations (CNVs) corresponds to unbalance chromosome structural variants. Firstly, identified on karyotype, the development of molecular techniques such as array Comparative Genomic Hybridization (aCGH) allowed the detection of small CNVs and increased by 100 folds the CNV detection sensibility.

In France, the array CGH technology was transferred to the diagnostic in 2007. Nevertheless, the medical interpretation of rare CNVs is still a challenge. In this context the AChroPuce network of French laboratories performing constitutional aCGH for diagnosis was created. As CNV frequencies depend on ethnic origin, a national CNV database was crucial to improve the CMA diagnostic process, notably by enhancing genotype/phenotype correlation, and new disease-causing genes identification. So, we create a diagnostic CNV database using Cartagenia Bench. Unfortunately, this database only allowed sharing between laboratories which used the Caratgenia Bench solution. In 2016, supported by the French Medical Research Foundation (FRM), we developed the BANCCO database. The Main objectives of this national CNV database are to:

- Provide a secure system to enable upload and storage of genomic data with accompanying phenotypes (anonymized data sharing) from all laboratory;
- Aid in the data interpretation by comparing genomic and phenotypic information from national databases (BANCCO and RDVD) as well as genome annotation resources;
- Identify genes located within any CNV, using coordinates from human genome assemblies;
- Encourage collaboration between molecular geneticists and molecular cytogeneticists
- Facilitate the identification of new syndromes and gene function;
- Determine CNV frequencies from the French population.

The database is available at: https://bancco.fr

More than 15 000 patients are included and 270 000 CNVS reported. Both phenotype and genotype (HPO terms) are available. I will discuss interests of this database in the international context and the difficulties linked to national and European policy.
More recently, the French government launched the French plan for genomic medicine 2025. This plan encompassed four major challenges: Public health issues; genomic medicine; Scientific and clinical; Technological and economical.

This plan wants to position France among the leading countries in personalized and precision medicine and integrate genomic medicine into the care pathway and the management of common diseases. By 2025, some 235,000 genomes will be being sequenced each year.
Among the specific measures proposed for the success of the plan, the establishment of a National Center for Intensive Calculation (CAD, Data Collection and Analysis) was decided. The CAD will be able to process and analyze the huge volumes of generated data and provide primary services for professional health care providers in the framework of their care pathways (*in silico* tests and aids to decision-making in diagnosis, establishing prognosis and designing therapeutic strategies). Moreover, the CAD will collect all genomic data produce in the context of the Plan.

Detection of copy number variations using whole-exome sequencing improves diagnostic yield of patients with rare Mendelian diseases

Paulo Silva[1,2,], Susana Sousa[1,2], Susana Barbosa[1,2], Sara Morais[1,2], Ana Lopes[1,2], Ana Brandão[1,2], Rita Bastos[1,2], Patrícia Arinto[1,2], Jorge Sequeiros[1,2,3], Isabel Alonso[1,2]*

[1] CGPP-IBMC – Centro de Genética Preditiva e Preventiva, Instituto de Biologia Molecular e Celular, Universidade do Porto, Portugal
[2] i3S – Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal
[3] ICBAS – Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal

** Corresponding author: paulo.silva@ibmc.up.pt*

Copy number variations (CNVs) are important elements of human genetic diversity, commonly observed in human populations, and are increasingly being recognised as an important aetiology of disease. Whole-exome sequencing (WES) has become a robust and cost-effective approach for clinical genetic testing of small sequence variants; detection of CNVs within WES data became possible with the development of various algorithms, software programs and statistical methods that enable comparing coverage between probands and controls. In this study, we express the challenges and feasibility of analysing CNVs using WES data.

We performed WES on patients with rare genetic disorders and analysed SNVs using an in-house pipeline, and CNVs using Golden Helix's VarSeq software. Probable causative CNVs were confirmed by qPCR or MLPA. In recessive disorders with a causative heterozygous SNV, fine-tuning of CNV analysis for each gene was performed on a case-by-case basis.

During 2017, we detected 12 CNVs as probably causative in 122 patients, but only one was confirmed by qPCR, a rate of 91.7% false positives (FPs). Using this knowledge, optimising the regions of interest, and using CNVs from public databases – 1kG Project Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/), ClinGen (https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), DECIPHER (https://decipher.sanger.ac.uk/), DGV (http://dgv.tcag.ca/dgv/app/home), ExAC (http://exac.broadinstitute.org) – as well as our internal CNV database, many common CNVs, FPs and sequencing artifacts were filtered. Then, over the past year, we detected 123 CNVs in 1251 patients as likely causative, 41 of which were confirmed by qPCR or MLPA (66.7% FPs).

We identified probable disease-causing CNVs using WES data in 3.3% of patients. Confirmation by orthogonal methodologies validated the software analysis pipeline and the CNVs detected, demonstrating that combining SNV and CNV analysis improves the molecular diagnosis of patients with rare Mendelian diseases.

New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis

Maarja Lepamets[1,2], Kaido Lepik[1,3], Reedik Mägi[1], Zoltán Kutalik[4,5,6]

1: Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
2: Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia
3: Institute of Computer Science, University of Tartu, Tartu, Estonia
4: Center for Primary Case and Public Health, University of Lausanne, Switzerland
5: Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, UK
6: Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
** Correspondence to: maarja.lepamets@ut.ee*

Even though whole-genome sequencing (WGS) is currently the optimal means for copy number variation (CNV) detection, such data is not yet available in large number of samples. As an alternative, abundant data from genotyping arrays could also be used. However, this method is prone to large number of false positive calls and quality filtering is necessary prior to further analyses. So far, only one CNV quality score has been developed (Macé *et al.*, 2016). The aim of our project is to improve this score by incorporation of independent multi-omics datasets.

We detected potential CNV regions by using the most popular CNV detection software PennCNV (Wang *et al.*, 2007) in samples with additional WGS, RNAseq, and/or methylation data available. We assumed that a true CNV affects the gene expression (GE) and the overall probe intensity captured from methylation arrays (MET) in regions within it. Thus, we measured the fluctuation of GE/MET in CNV carriers compared to non-carriers. Additionally, we calculated the fraction of PennCNV calls that were validated by WGS reads. All three independent omics-based quality metrics were concordant: the Pearson correlations between MET and GE metrics ranged between 0.59-0.80 for deletions and 0.33-0.57 for duplications; both had high correlations with WGS metric (R > 0.7). In EGCUT set of 968 samples we saw that approximately 44.2% of PennCNV calls had no effect on GE/MET, were not found from WGS and were, therefore, most likely false positives.

In order to distinguish high and low quality CNVs in samples for which only array data is available, we built a statistical model to predict our quality metric based only on PennCNV output parameters (CNV length, probe-density, number of CNVs per sample, etc). We validated our model on CNVs of close relatives in UK Biobank. We achieved the mean scores of 0.82/0.62 for familial deletions/duplications compared to 0.66/0.27 for non-familial CNVs. Furthermore, we tested our score in a genome-wide association study on four complex diseases (rheumatoid arthritis, inflammatory bowel disease, Crohn's disease and ulcerative colitis). We found 11 novel associations ($P<1.7*10^{-6}$) out of which only one would have been discovered by previous CNV quality metrics. Noteworthy examples are the association between RA and a 6p21.3 deletion ($P=9.97*10^{-7}$) overlapping a known strong GWAS signal ($P<10^{-250}$) and a 10p11.21 deletion associated with IBD ($P=9.10*10^{-7}$) that overlaps the tight junction-related gene *PARD3*, which has been associated with IBD in candidate gene studies.

References

Macé, A., *et al.* (2016). *Bioinformatics* 32:3298–3305. https://doi.org/10.1093/bioinformatics/btw477
Wang, K., *et al.* (2007). *Genome Research* 17:1665–1674. https://doi.org/10.1101/gr.6861907

Identification of mobile retroCNVs during genetic testing: consequences for routine diagnostics

Nicolas Chatron [1,2*], *Kevin Cassinari* [3], *Olivier Quenez* [3], *Stéphanie Baert-Desurmont* [4], *Claire Bardel* [5,6], *Marie-Pierre Buisine* [7], Eduardo Calpena [8],*Yline Capri* [9], Jordi Corominas Galbany [10], *Flavie Diguet* [1,2], *Patrick Edery* [1,2] , *Bertrand Isidor* [11], *Audrey Labalme*[1], *Cedric Le Caignec* [11,12], *Jonathan Lévy* [13], *François Lecoquierre* [4], *Pierre Lindenbaum* [14,15], *Olivier Pichon* [11], *Pierre-Antoine Rollat-Farnier* [1,5], *Thomas Simonet* [16,17], *Pascale Saugier-Veber* [4], *Anne-Claude Tabet* [13,18], *Annick Toutain* [19,20], Andrew O. M. Wilkie [8], *Gaetan Lesca* [1,2], *Damien Sanlaville* [1,2], *Gaël Nicolas* [3], *Caroline Schluth-Bolard* [1,2]

1. *Service de génétique, CHU Lyon, Lyon, France*
2. *Equipe GENDEV, CRNL, INSERM U1028, CNRS UMR5292, UCBL1, Lyon, France*
3. *Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ, F 76000, Normandy Center for Genomic and Personalized Medicine, Rouen, France*
4. *Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics, F 76000, Normandy Center for Genomic and Personalized Medicine, Rouen, France*
5. *Cellule bioinformatique de la plateforme de séquençage NGS-HCL, HCL, France*
6. *Service de biostatistique bioinformatique, HCL, Lyon, France*
7. *Inserm UMR-S 1172, JPA Research Center, Lille University, and Department of Biochemistry and Molecular Biology, Lille University Hospital, Lille, France.*
8. *Clinical Genetics Group, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK*
9. *UF de génétique clinique, Hôpital Universitaire Robert Debré, AP-HP, Paris, France*
10. *Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands*
11. *CHU Nantes, Service de Génétique Médicale, Nantes, France*
12. *Université de Nantes, Nantes, France*
13. *UF de cytogénétique, Hôpital Universitaire Robert Debré, AP-HP, Paris, France*
14. *INSERM, UMR_S1087, Institut du thorax, Nantes, France*
15. *CNRS, UMR 6291, Nantes, France*
16. *Centre de Biotechnologie Cellulaire, HCL, Lyon, France*
17. *Nerve-Muscle Interactions Team, Institut NeuroMyoGène CNRS UMR 5310 - INSERM U1217 - Université Claude Bernard Lyon 1*
18. *Unité Génétique Humaine et Fonction Cognitive, Institut Pasteur, Paris, France*
19. *Service de Génétique, CHU Tours, France*
20. *UMR 1253, iBrain, Université de Tours, Inserm, Tours, France*

Corresponding author : nicolas.chatron@chu-lyon.fr

Human retrocopies, i.e. mRNA transcripts benefitting from the LINE-1 machinery for retrotransposition, may have specific consequences for genomic testing. NGS techniques allow the detection of such mobile elements but they may be misinterpreted as genomic duplications or be totally overlooked. Here, we present eight observations of retrocopies detected during diagnostic NGS analyses of targeted gene panels, exome, or genome sequencing. For seven cases, while an exon-only copy number gain was called, read alignment inspection revealed a depth of coverage shift at every exon-intron junction where indels were also systematically called. Moreover, aberrant chimeric read pairs spanned entire introns or were paired with another locus for terminal exons. The 8th retrocopy was present in the reference genome and thus showed a normal NGS profile. It was identified during the cDNA study performed to validate an intronic SNV by the preferential amplification of the retrocopy. We emphasize the existence of retrocopies and strategies to accurately detect them at a glance during genetic testing. We discuss pitfalls for genetic testing and their potential clinical consequences.

Session 3

Copy-number variation detection from exon capture data

Anna Benet-Pagès

Medical Genetics Center, MGZ, Munich, Germany

Gene dosage abnormalities account for a significant proportion of pathogenic mutations in rare genetic disease related genes. In times of next generation sequencing (NGS), a single analysis approach to detect SNVs and CNVs from the same data source would be of great benefit for routine diagnostics. However, CNV detection from exon-captured NGS data has no standard methods or quality measures so far. The primary strategy of the current bioinformatics methods is based on the read depth of coverage (DOC). The underlying approach is to compare the differences of DOC in particular genomic regions between case and control samples. The DOC-based methods can detect arbitrarily large CNVs and can be effectively used with paired-end, single-end, and mixed read data. Numerous standalone and web-based tools are currently available to detect CNVs based on different features of NGS data, resulting in variation in the prediction of CNVs. In this presentation the advantages of incorporating more than one method for CNV prediction, in addition to the key factors which affect the sensitivity and specificity of CNV pipelines (i.e. size of the reference set, kit performance, normalization approach, single exon calls, variability in the capture efficiency of nearby genomic regions, and low complexity sequences) will be discussed. Furthermore, the experiences of CNV analysis in >5000 patients with hereditary cancer syndromes or rare Mendelian diseases will be presented. Overall parallel analysis of SNVs and CNVs from NGS capture data within a routine diagnostic setting increased the diagnostic yield between 5% and 10% depending on the associated phenotype.

Detection of copy number variations from NGS data using read depth information: a diagnostic performance evaluation

Olivier Quenez[1], Kevin Cassinari[2], Sophie Coutant[2], François Lecoquierre[2], Kilan Le Guennec[1], Stéphane Rousseau[1], Anne-Claire Richard[1], Stéphanie Vasseur[2], Emilie Bouvignies[2], Jacqueline Bou[2], Gwendoline Lienard[2], Sandrine Manase[2], Steeve Fourneaux[2], Myriam Vezain[2], Pascal Chambon[2], Géraldine Joly-Helas[2], Nathalie Le Meur[2], Mathieu Castelain[2], Anne Boland[3], Jean-François Deleuze[3], FREX consortium, Edwige Kasper[2], Thierry Frebourg[2], Pascale Saugier-Veber[2], Stéphanie Baert-Desurmont[2], Dominique Campion[1,4], Anne Rovelet-Lecrux[1], Gaël Nicolas[1]*

1 Department of Genetics and CNR-MAJ, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Normandy Centre for Genomic and Personalized Medicine, Rouen, France.

2 Department of Genetics, Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Normandy Center for Genomic and Personalized Medicine, Rouen, France

3 Centre National de Recherche en Génomique Humaine, Institut de Génomique, CEA, Evry, France.

4 Department of Research, Centre hospitalier du Rouvray, Sotteville-lès-Rouen, France

Corresponding author email : olivier.quenez@inserm.fr

The detection of Copy Number Variations (CNVs) from NGS data is under-exploited and chip-based technologies or targeted techniques are still commonly used for their detection in a diagnostic setting. We assessed the performances of CNV calling using CANOES, a read-depth comparison method applied to gene panels and whole exome sequencing (WES).

We applied CANOES to NGS data obtained from (i) 465 samples with both gene panel and comprehensive Quantitative Multiplex PCR of Short Fluoresent (QMSPF) data available (total of 60 exons assessed), (ii) 95 additional samples with NGS data from 2 different gene panels, (iii) 135 samples with both WES and array CGH (aCGH) data available and (iv) 1,056 additional WES.

From the gene panel data, CANOES detected all 14 events that were previously identified by QMPSF, with neither any false positive nor any false negative among 465 samples (Sensitivity (Se) = specificity = 100%). In addition, CANOES detected 97 candidate CNVs in 95 additional samples, 86 of which were confirmed by a targeted technique (PPV= 90.1% overall). From the WES data, CANOES detected 159 of the 195 exonic events previously detected by aCGH among the 135 samples with WES+aCGH data available (Se=81.5%). Overall, the PPV of CANOES from WES data was 94.8% after the confirmation of 108 of the 123 calls.

Most rare and high-risk CNV carriers do not have major health, cognitive or socioeconomic consequences

Elmo Saarentaus[1], Nina Mars[1], Ari Ahola-Olli[1], Tuomo Kiiskinen[1], Juulia Partanen[1], Sanni Ruotsalainen[1], Mitja Kurki[1,2,3], Lea Urpa[1], Aki S. Havulinna[1,4], Markus Perola[4], Veikko Salomaa[4], Markku Peltonen[4], Jaakko Kaprio[1,5], Olli Pietiläinen[2,6], Mark Daly[1,2,3], Aarno Palotie[1,2,3,]*

[1]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland; [2]Stanley Center for Psychiatric Research, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA; [3]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, USA; [4]National Institute for Health and Welfare, Helsinki, Finland; [5]Department of Public Health, University of Helsinki, Finland; [6]Stem Cell and Regenerative Biology, Harvard University, Cambrige, USA

*aarno.palotie@helsinki.fi

CNVs are associated with syndromic and severe neurodevelopmental and psychiatric disorders (SNPDs), such as intellectual disability (ID), epilepsy, bipolar disorder (BD), and schizophrenia (SZ). Although considered high-impact, CNVs are also observed in the general population. This presents a diagnostic challenge in evaluating their clinical impact. To estimate the impact of CNVs to general health and well-being, we analyzed CNV burden alongside three genome-wide Polygenic Risk Scores (PRS; IQ, Educational Attainment [EA], and SZ) in a Finnish working-age population (FINRISK, n=23,053).

In carriers of high-risk CNVs (Susceptibility CNV, Risk Gene Deletion, or Large [>1Mb] CNV), 95.3% (533/559) had no SNPD diagnosis. The remaining 4.6% (n=26) had been diagnosed with ID (4/26, 15%), Schizophrenic Disorders (10/26, 38%), Epilepsy (13/26, 50%), Bipolar Disorder (3/26, 11.5%), and Behavioral and Emotional Disorders (1, 3.8%). ID in particular was associated with Large Deletions (OR=6.8 [1.6-29]), Large Duplications (OR=5.0 [1.2-21]), ID Gene Deletions (OR=10.6 [1.4-80]), and Susceptibility CNVs (OR=11.0 [1.4-83]). In comparison, the 559 individuals with the highest PRS_{SZ} were enriched for Schizophrenia (OR=4.9 [2.7-8.8]), ID (OR=3.2 [1.2-8.1]) and Bipolar Disorder (OR=2.5 [1.3-5.0]).

We hypothesized that even if CNV carriers might not have a diagnosed SNPD, some of them might have subclinical features that could be associated with their general health, well-being or socioeconomic status. We tested this hypothesis in individuals without SNPD (n=22,210). We observed lower educational attainment for individuals carrying Susceptibility CNVs (1.54 years [0.48-2.6]), Large Deletions (0.70 years [0.10-1.30]), and High pLI Gene Deletions (0.57 years [0.15-0.99]). In the 559 lowest PRS_{EA} individuals, the effect was more severe (1.97 years [1.47-2.26]) than any high-risk CNV type. The 559 lowest PRS_{IQ} individuals had 1.18 years [0.89-1.48] lower EA.

Copy number variation detection tool for targeted s sequencing data

A. K. Singh*[1,2], T. Vold[1], L. A. S. Lavik[1], M. F. Olsen[1]

[1]Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway
[2]Norwegian University of Science and Technology, Trondheim, Norway
Email: ashish.kumar.singh3@stolav.no

Introduction: Copy number variants (CNVs) are often disease causing, hence, of importance for genetic diagnostics. CNV detection is often done using MLPA, which is expensive and lab intensive, also with limitation of detecting CNV for limited number of genes. With more and more genes are being tested, MLPA testing is becoming unmanageable. We at Department of Medical Genetic at St. Olavs Hospital (Norway) have in-house developed In-Silico CNV detection tool for targeted sequencing data.

Materials and Method: CNV detection tool is based on “Depth of coverage” approach. It calculates CNV scores for a test-sample by comparing its coverage-depth for a specific target region with mean coverage-depth for the same target region of pool of normal samples. Tool dynamically selects target regions using sliding window approach with window-length of user-choice. Tool uses both sequencing-run and static pooling approaches for pool creation. For the tool, theoretical CNV-scores are: 0 for normal samples, -1 for deletions and 0.58 for duplications. Validation of the tool has been done using 36 positive controls with CNVs in 12 genes and 11 negative controls.

Results: All the CNVs of the positive controls were successfully detected resulting in a measured sensitivity of 100%. The specificity was measured to be 91% where most of the false positives were due to systematic errors in challenging genomic regions (e.g. high GC –content, low coverage, high homology).

Conclusion: Our CNV detection tool has been validated and established in routine diagnostics of hereditary cancer at our department. CNV detection using targeted NGS data makes it possible to broaden our genetic testing services to also include CNV detection of genes where there is no MLPA analysis available. This tool is shown to be sensitive and specific in addition to time and cost-effective.

Association analysis of SVs/CNVs using NGS data

Victor Guryev

European Research Institute for the Biology of Ageing (ERIBA), University of Groningen, Groningen, UMC Groningen, The Netherlands

Detection of structural genome variants (SVs) using short read next generation sequencing data relies on several algorithmic approaches. Analysis of read coverage, discordant mapping of paired reads, split-mapping of unmapped reads and local *de novo* re-assembly provide ways to discover large sequence variants. However, integration of SV calls identified by different methodologies/tools and merging of SV calls across all individuals is difficult and can complicate further SV-based association studies.

An original way to overcome SV calling difficulties is use hallmarks of structural variants directly in association mapping. We explored association analysis of human phenotypic traits with read coverage, discordant or partial mapping of read pairs. We show that genomic regions identified in such analysis do contain structural variants that correspond to expected type of read discordancy. These SVs might represent the genetic polymorphisms that underlie phenotypic variability in human traits.

This ‘shortcut’ method is incredibly fast and provides an efficient way for association analysis that can identify structural variants that contribute to many phenotypic traits.

Implementation Driven Development of Standards for Genomic Data Exchange from Cancer Genome Data Collections

Michael Baudis

Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland

Cancers are genomic diseases, arising from the clonal propagation of somatic mutation events, with a limited contribution from inherited genomic variants. Genomic copy number variations are major contributors to malignant transformation and progression and constitute - at least in their quantitative extension - the largest contributors to genomic mutation landscapes, in the majority of cancer types. However, while sCNV based alterations of many canonical cancer related genes has been shown, the impact of the widespread, extensive sCNV patterns is poorly understood, not the least due to the limited amount of accessible whole-genome sCNV datasets and the lack of appropriate data standards and exchange formats.

With the Progenetix and arrayMap repositories our group provides 2 of the largest resources for pre-processed CNV data in cancer, mostly based on the curation and re-analysis of data from public repositories and individual publications. As members of the Global Alliance for Genomics and Health (GA4GH) and participants in the ELIXIR h-CNV community and Beacon project, we utilise our resources to design, implement and test variant annotation and storage formats, as well as federated genome variation discovery tools such as the “Beacon” API, with focus on CNV data.

Links:

progenetix.org
beacon-project.io
beacon.progenetix.org
schemablocks.org

HGVS

HUMAN GENOME
VARIATION SOCIETY

Genotypes & Intermediate Phenotypes

Tuesday 15th October 2019

Houston, TX, USA

a satellite to ASHG

events.hgvs.org

meeting site to be changed to this event shortly after this meeting