Characterizing Copy Number Variations using Next- and Third-Generation Sequencing and their Association with Plasma Biomarkers

Daniel Schmitz[1], Zhiwei Li[2], Nima Rafati[3] and Åsa Johansson[1]

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University

[2]Twelve Bio ApS, Copenhagen, Denmark

[3]National Bioinformatics Infrastructure Sweden, Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University

Abstract

Structural variations (SVs), including copy number variations (CNVs), affect approximately 20 million bases in a typical human genome. Recently, there has been an increasing interest in the role of CNVs in human development and diseases. However, most commonly used CNV detection methods, such as array comparative genomic hybridization, lack the ability to detect novel signals. This study aims to explore the role of CNVs in regulating plasma protein

biomarkers. We identified CNVs, using CNVnator, from high-coverage (30x) Illumina next-generation sequencing data, in more than 1,000 individuals. The identified CNVs were summarized and filtered as a population copy number matrix to 23,381 non-overlapping CNV regions (CNVRs) that were polymorphic in at least three individuals. Using a total of 438 plasma protein biomarkers, that were available in 872 individuals with WGS data, we conducted linear regression analyses . We identified CNVs at 19 CNVRs to be significantly associated with 22 plasma proteins ($\boldsymbol{p < 4.79 \times 10^{-9}}$). Lastly, we selected a set of five polymorphic CNVRs for validation using Pacific Bioscience SMRT sequencing in 15 samples. Two CNVRs replicated and we identified two more CNVRs to be clusters of many short repetitive elements. Our findings provide insight into the involvement of CNVs on human disease as well as the application of novel sequencing approaches for SV detection.

Introduction

Genome-wide association studies (GWAS) have enabled the discovery of thousands of associations between single-nucleotide polymorphisms (SNPs) and common traits as well as diseases. Despite their success, GWAS can only explain a small part of observed heritability. For instance, a recent study estimated the variance of 32 complex traits explained by common SNPs to lie between 9.8% and 48.9%[1]. In 2015, the 1000 Genomes project consortium found 99.9% of all identified variants to be SNPs or short indels[2]. However, they also emphasize the importance of structural variants (SVs), which are rarer but cover more bases, with a typical genome containing 2,100 to 2,500 SVs covering 20 million bases. A meta-analysis of 55 population studies using the Database of Genomic Variants (DGV) estimated CNVs to cover 4.8 – 9.5% of the genome and observed complete deletions of about 100 genes without apparent phenotypic effects from the copy number variations (CNVs)[3] . Another study identified SVs in 2,504 human genomes to DGV and found that 43% of their CNVs were novel discoveries. Furthermore, they also observed multiple breakpoints in the CNV regions (CNVRs) across the population likely to be caused by individual mutational events.

Traditionally, array-based methods, such as SNP arrays and array comparative genomic hybridization (aCGH), have been used to detect genetic variations, including CNVs. Their high accuracy and reasonable price made them a popular choice for association studies. However, their limitation to previously identified polymorphisms makes it less powerful to identify novel signals. Furthermore, their resolution is limited, with common arrays being

limited to CNVs of at least 8 kbp in size[5]. Recent developments in sequencing technologies have enabled unprecedented insights into the genetic architecture of the human genome. With improvements in data quality and cost, whole-genome sequencing (WGS) has become more popular in large-scale genomic studies, including the identification of SVs. Thanks to continuously improving cost and quality of high-throughput sequencing, interest in these variants has been invigorated. There is, however, no widely accepted pipeline to detect SVs with high recall for NGS data[6].

Even though SVs identification is state-of-the-art when searching for genetic causes of monogenetic diseases, few association studies have considered the effect of SVs on common diseases and complex traits. Previously, CNVs have been associated with evolutionary fitness and embryonic lethality[7], psychiatric disorders[8,9], Crohn's diseases, type 1 diabetes, and multiple developmental diseases[9]. A previous study using data from UK Biobank found CNVs in 28 genes to be associated with 13 blood biomarkers[10]. Several studies linked both germline and somatic CNVs to multiple types of cancer, including an integrated analysis of CNVs, SNPs and expression data[11–14].

Biomarkers are well-studied traits in GWAS as they are often measured in large cohorts and are quantitative measures which increases the power to find associations. More specifically, protein biomarkers, expressed by one single gene, is usually more or less monogenic, which is beneficial for statistical power in studies in smaller cohorts[15,16]. In this project, we focused on characterizing CNVs from high coverage WGS data of over 1,000 individuals from the

Northern Sweden Population Health Study (NSPHS)[17]. We called CNVs using CNVnator, and tested for association between copy number polymorphisms (CNPs) and the variation in protein levels for a large set of proteins (N=438) that has been selected to be established or exploratory biomarkers of disease[18]. Subsequently, we re-sequenced 15 individuals from our cohort using Pacific Bioscience (PacBio) Single-Molecule Real Time (SMRT) technology, to verify the CNPs.

Materials and methods

Study cohort

The *Northern Swedish Population Health Study* (NSPHS) was a cohort study conducted in two municipalities in the region of Norrbotten, Sweden. Blood samples were taken and immediately frozen at -70°C. WGS was performed at SciLifeLab in Stockholm using Illumina technology to 30x coverage, and mapped to GRCh37, as described previously[19,20]. After variant calling and QC 1,021 samples remained for analysis.

Protein biomarkers had been measured in 903 individuals using the Olink Protein Extension Assay (PEA) and five Proseek panels (CVD2, CVD3, NEU1, ONC2, INF1), as described previously[21]. In short, PEA is an affinity-based assay that uses oligonucleotide-labelled pairs of antibodies that bind to the target proteins in close proximity to each other. If both antibodies bind, they produce a PCR target sequence, which can be quantified using

standard real-time PCR. The analysis was performed on plates with 96 wells, allowing for 92 individuals as well as three positive and one negative controls per batch, which serve to determine the lower detection limit and normalize the protein measurements. Signals below detection limit were removed and the remaining measurements were normalizes using the rank-based inverse normal transformation ($\mu = 0, \sigma = 1$). A total of 438 biomarkers and 892 samples and passed protein QC and 872 samples passed both genotyping and biomarker QC.

CNP calling using CNVnator

CNV calling was done using CNVnator. We first used CNVnator to estimate the optimal bin size for each sample, as the lowest value of (70, 85, 100, 150, 200, 250 bp) at which the ratio of the bins' read depth (RD) mean values to the standard deviation was between 4 and 5. The reason for having the mean value of the converted RD signal 4~5 times greater than its standard deviation is to preserve enough statistical power for detecting deletions by t-test between the regional and global read depth (RD) signal while detecting the variations with smallest bin size possible to enable higher breakpoints resolution. CNVs were then identified in each sample separately and filtered in accordance with the recommendations set by CNVnator's authors[18] . CNVnator provides P-values for each detected CNV calculated from a one-sample t-test between the local and global RD signal. In the initial CNV calling, CNVs are considered high-quality detections if they pass the significance threshold $p < 0.05$ after Bonferroni correction. Additionally, we excluded CNVs where the fraction of reads with ambiguous alignments (q_0) was 0.5 or more.

After detecting CNVs in all samples, a population CNP matrix was created using bedtools. We split the genome into non-overlapping windows of 200 bp each and identified all windows where any sample had a high quality CNV detected. Then, we recorded all CNVs that were detected in each of the windows for all samples. We applied less stringent QC for inclusion of copy numbers in the CNP matrix and only applied the q_0 threshold. Samples for which no CNV had been identified within a 200bp window were assigned CN=2 (wildtype), and samples that failed QC were set as missing. Finally, adjacent 200bp windows with identical genotypes across all samples were merged. The merged windows represented the final CNV regions (CNVRs) used in our downstream analyses.

Association analysis

We have previously shown that we need a minimum of four individuals with one copy of the minor allele to reach the genome wide significance threshold in a GWAS in the cohort[19]. We therefore excluded all CNVRs, for which less than three individuals had a copy number different from 2. To test for association between the CNVRs and the biomarker levels, we used a linear regression model in the glm function in R version 4.3.4 with the CNs in the CNVRs as predictors and biomarker measurements as responses. We included age and sex as covariates. We applied a Bonferroni adjustment for multiple testing and adjusted for the number of biomarkers × the number of CNVRs analysed.

Resequencing with PacBio SMRT

We selected 15 individuals for resequencing with Pacific Bioscience SMRT technology. We performed whole-genome sequencing on a PacBio SEQUEL II system in continuous long read (CLR) mode. Libraries were prepared according to manufacturer specification. The resulting reads underwent standard QC procedures and were automatically mapped to human genome assembly GRCh38. These mappings were delivered as binary alignment map (BAM) files. We extracted the reads from the mappings using BEDTools bam2fastq version 2.29.2 and aligned them to GRCh37 (the same build as the Illumina data) using pbmm2 version 1.4.0. We then called structural variation using SVIM v1.4.2, Sniffles v1.0.12 and PBSV v2.4.0 and visualized the coverage and reads supporting the CNVs in the regions of interest using samplot v1.1.0[22]. Variants called by SVIM, Sniffles and PBSV were considered to correspond to the previously called CNVs if they overlapped reciprocally by at least 50%.

Results

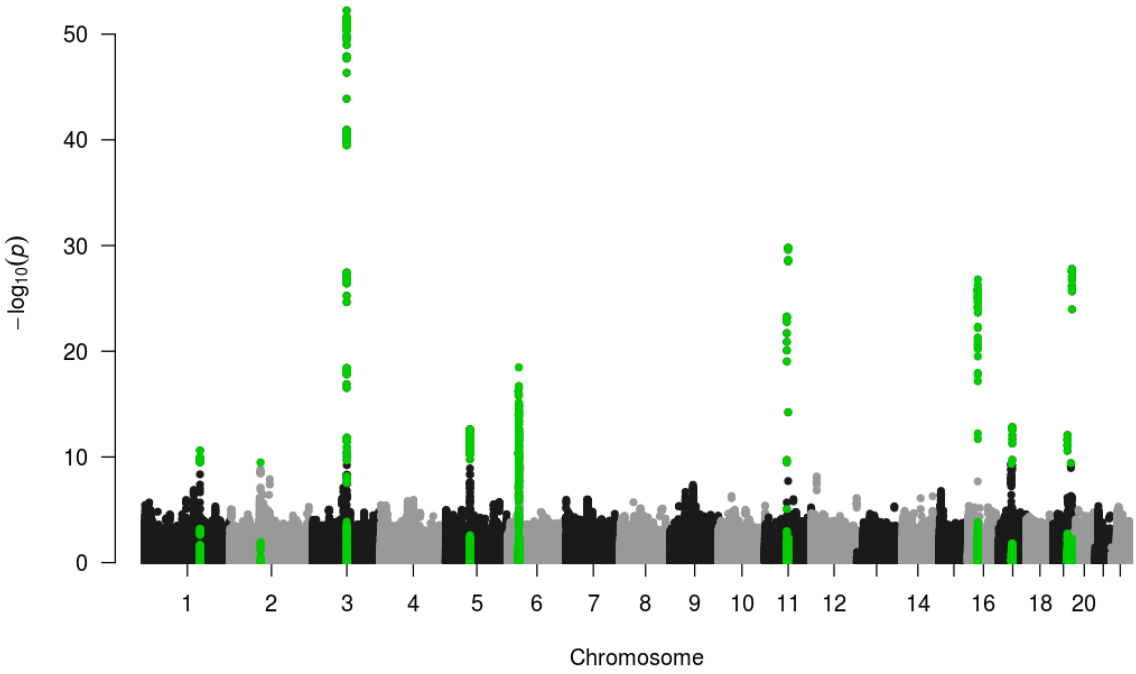
CNV Detection

We selected different optimal bin sizes for each sample. The mean optimal bin size was 92 bp. We expected the resolution of CNVnator to be around the bin size for each sample, which is around 92 bp. We observed that the size of structural variations reported by CNVnator ranged from 140 bp to 20 Mbp. Our observation of the smallest variations reported by CNVnator in the population is 140 bp, which agreed with our expectation. An average 774 deletions and 641 insertions were detected per sample. The final CNP matrix

contained genotypes (CNs) of 1,021 individuals in 243,987 windows. After merging of adjacent loci with consistent genotypes, 23,381 CNVRs remained.

Association Analysis

In the CNV-phenotype association analysis with 438 plasma proteins, we detected 17 biomarkers with at least one association passing the significance threshold ($p < \frac{0.05}{438 \times 243987} \approx 4.68 \times 10^{-10}$), when adjusting for all 243,987 200-bp windows (Figure 1). Since some of the 17 biomarkers are clustered in the nearby regions in chromosome 3, 6, 17 and 19 only nine peaks of significant associations between CNVs and 17 protein biomarkers. The nine peaks consist of 382 significant 200-bp windows. Considering that the 200-bp windows were not independent and could be merged into 23,381 independent CNVRs, a more liberal adjustment for multiple testing might be appropriate, which resulted in a total of 19 significant CNVRs (Table 1) that are significantly associated with the 22 protein markers.



167

168 *Figure 1. Manhattan plot for the CNPs-biomarker association with 200-bp CNV windows. A total of 17 protein biomarkers*
169 *had at least one hit passing the adjusted significant threshold* $\boldsymbol{p < 4.68 \times 10^{-10}}$*. However, a total of 382 significant 200-*
170 *bp windows, distributed over nine loci, were associated with any of the 17 biomarkers. The CNPs with at least one hit are*
171 *highlighted in green.*

SMRT Sequencing

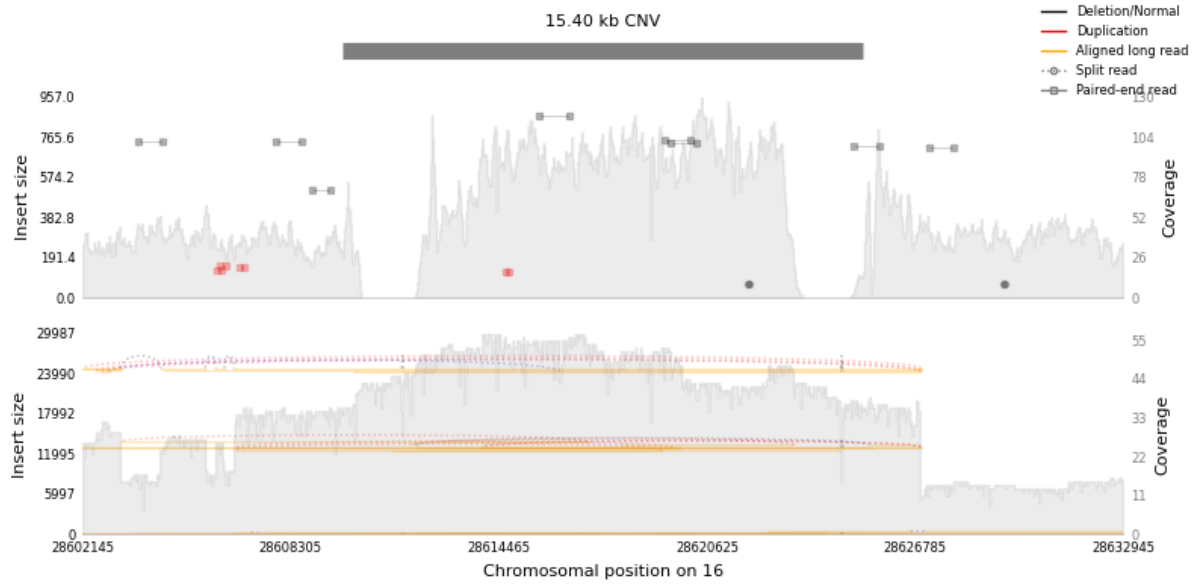
We selected five of the CNVR regions (Table 1) which were highly polymorphic in the population and strongly associated with any of the biomarkers for validation using long-read sequencing. In addition, theses five CNV-associations did not overlap with, or were more strongly associated, than previous GWAS findings and were of an optimal size to be detected by long read sequencing (close to 5000bp). The 15 individuals that were prioritized for long read sequencing, were selected to display different alleles of the five CNVRs. The number of high-quality reads – and therefore the coverage – varied wildly between samples. The system reported that less than half of all reactions produced high-quality reads for three samples. These sample were therefore excluded from further analyses. In general, deletions were consistently called among the three CNV-callers used, whereas duplications could not be replicated between the callers. SVIM called the most of our target CNVs in agreement with CNVnator , while PBSV detected the fewest.

The CNV on chromosome 2 was not detected by any of our callers in the long-read sequencing data. The coverage data in this region also lacked evidence for CNVs in the samples where CNVnator called a CNV in the short-read sequencing data. This led us to suspect this CNV to be an artifact. Interestingly, the region downstream of this CNV, as well as a short one within the CNV, received no coverage at all in any sample. This might have affected CNVnator's ability to accurately detect structural variation.

The CNV on chromosome 3 was detected by SVIM, only, which called it consistently with the Illumina data set. Neither PBSV or Sniffles called variants in this region consistent with previous results. An optical inspection of the region revealed strong evidence for a large deletion where CNVnator called this CNV.

The CNV on chromosome 5 could not be replicated with any caller in the long-read data. The per-base coverage in this region turned out to be highly variable with many low-quality alignments. This may be due to the high frequency of repetitive elements, in this region. Genome Multitool reported a low mappability of this region, as well.

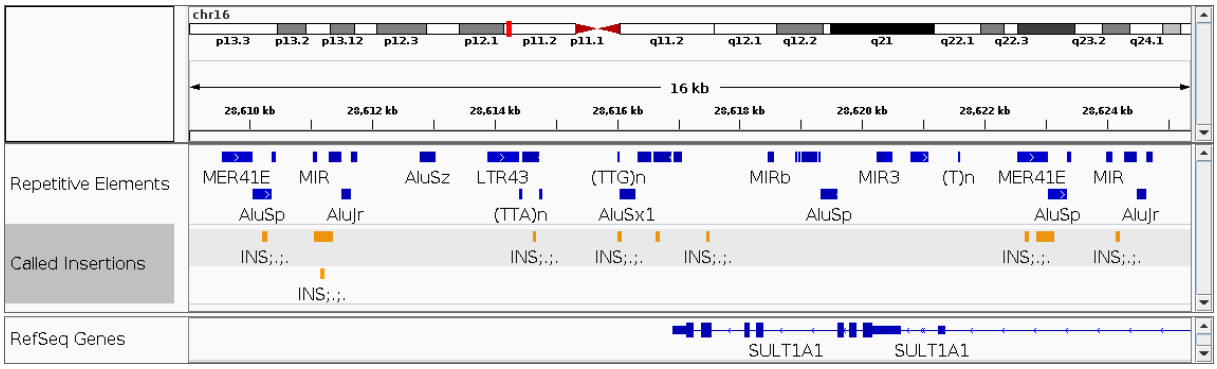
The CNV on chromosome 16 could not be replicated in the long-read data (Figure 2). However, SVIM called several small insertions in this region, corresponding to CN gains reported by CNVnator. Additionally, the flanking regions of this CNV exhibited low mapping quality. Both observations might be caused by the repetitive elements in this region, which map well to them (Figure 3). This provides a possible explanation for the reported CNV. CNVnator might have called these small insertions, which were then merged into a single larger window by the analysis pipeline.



211

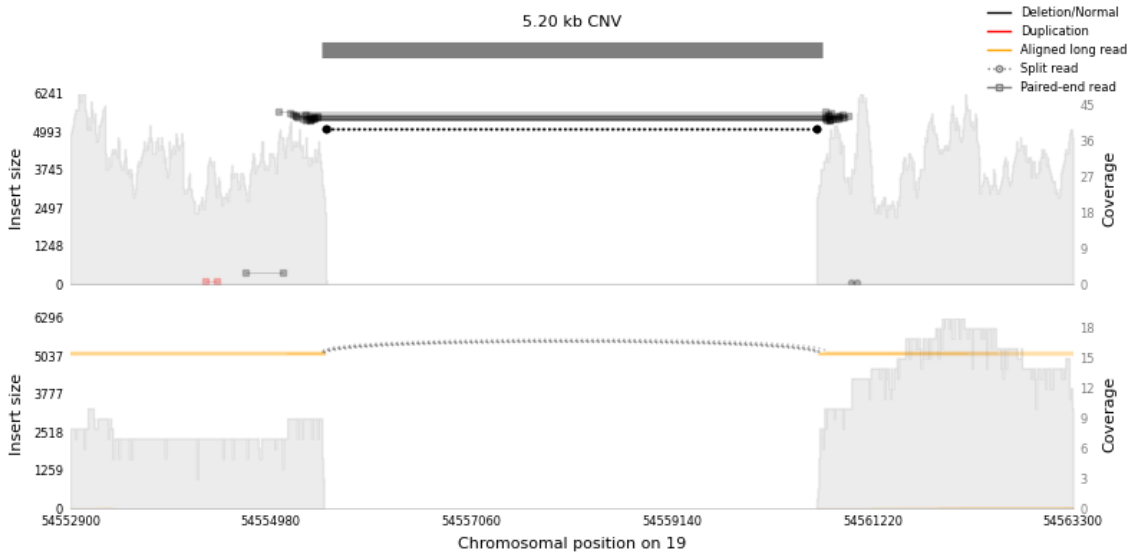
212 *Figure 2: Coverage plot of the CNV on chromosome 16 in one individual. The top shows Illumina, the bottom SMRT data. The*
213 *gray area represents the per-base coverage in the area. Lines in the plot show read-level evidence for SV. In this individual,*
214 *this CNV was called as a duplication by CNVnator, which is illustrated by the higher coverage. However, there was no clear*
215 *evidence a duplication event with the same breakpoints in the long-read sequencing data*

216



217

218 *Figure 3. Repetitive elements on chromosome 16 and insertions called by SVIM. The insertions mostly map to the repetitive*
219 *elements reported by RepeatMasker. This suggests that CNVnator actually detected these smaller CNVs and merged them*
220 *because of its binning approach.*



221

222 *Figure 4. Coverage plot of the CNV on chromosome 19 in one individual. The top shows Illumina data and the bottom SMRT*

223 *data. This was called as a homozygous deletion (CN 0). There is very clear evidence of this in both short- and long-read data.*

224 The CNV on chromosome 19 was consistently detected by all callers. SVIM confirmed its

225 presence, including zygosity, in all samples. Sniffles and PBSV called the variant in

226 accordance with the Illumina data in all but two samples. A visual inspection of the area also

227 confirmed its presence (Figure 4,Figure 5)

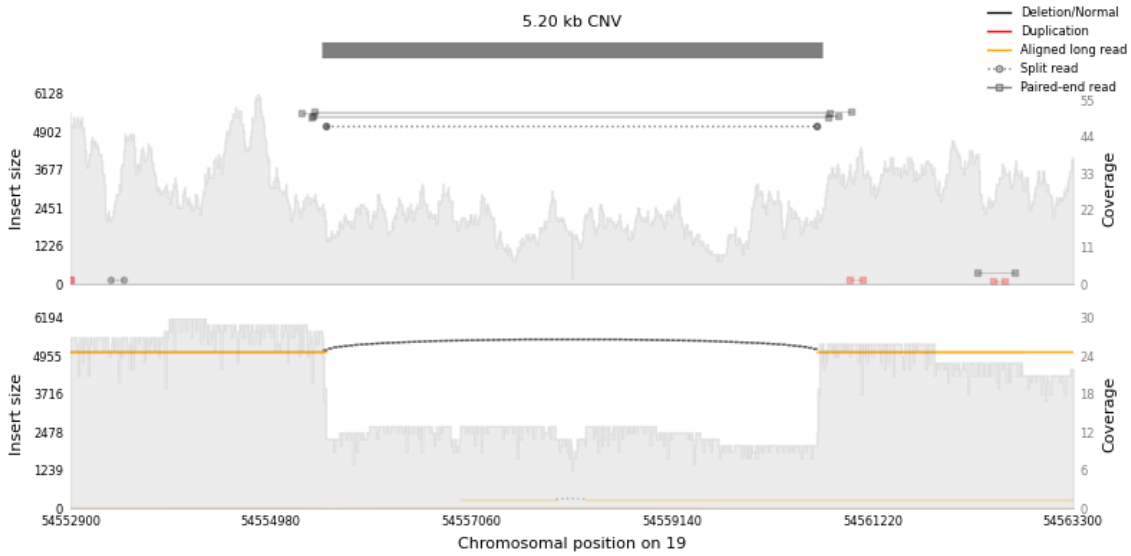


Figure 5. Coverage plot of the CNV on chromosome 19 in one individual. This was called as a heterozygous deletion (CN1). There is clear evidence in both long- and short-read data.

Discussion

We have identified CNVs in a population-based cohort and identified associations between CNVs and protein biomarkers. For the 872 samples with both genotype and phenotype data, we identified a total of 19 CNVR regions to be associated with 22 protein biomarkers. This is clearly a smaller number compared to our previous GWAS in the same cohort[19,21]. However, this agrees with that polymorphic CNVs are numerically fewer compared to SNPs that are commonly used in GWAS.

There have not been many studies focusing on identifying CNPs using high-throughput sequencing and downstream CNV-phenotype association analysis previously. A recent study developed a novel pipe-line for CNV discovery at population level with 1,364 individuals and tested for association with 275 protein biomarkers[23]. They report four significant associations ($\boldsymbol{p < 1.79 \times 10^{-6}}$, resolution: 15kb). Only one of our significant CNVRs (chr6: 30994100 — 35629600) overlapped with the result of their study. However the proteins measured by the two studies are not identical which might explain the low degree of overlap. By identifying 19 CNVs to be associated with the expression level of protein biomarkers, our study contributes to increasing the number of CNV-biomarker associations compared to previous studies.

Our long-read sequencing approach showed that CNV calling results from short- and long-read technologies may not agree in many cases. While the deletions among our target CNVs mostly could be validated, the characterization of the duplications was not immediately clear. For instance, the target CNV on chromosome 16 manifested as many smaller insertion events, rather than one large duplication as indicated form the short read sequencing. The binning approach employed by CNVnator might have been responsible for them appearing as a single variant in the short-read data. SMRT sequencing provided a way to accurately resolve this region. On the other hand, we could not detect the target CNV on chromosome 2 in the long-read data, despite the presence of a strong association with levels of GPNMB, which would suggest that there is indeed an underlying genetic effect causing this association.

One limitation in the current study is that we did not include other SVs such as inversions and translocations. A recent study of haplotype-resolved SVs discovery in the human genome integrated long-read, short-read, strand-specific sequencing technologies and numerous variations calling algorithms in three parents-child trios and detected an average of 156 inversions out of 27,622 SVs per sample[24]. The lower frequency of other SVs and lack of long-read strand-specific information cause difficulty in detecting them by only WGS data.

In this project, we focus on CNV discovery based on the alignments of short-read WGS reads to the human genome reference. Although the current human genome references (GRCh38 and GRCh37) claim to resolve 99% of the human euchromatic genome, a study constructed a de novo assembly of two Swedish genomes by long-read sequences and reported around 10 Mbp novel sequences missing from the GRCh38 mainly located in the centromeric or telomeric regions[25]. The misalignments of the reads from the unresolved regions on the current human genome reference can limit the discovery of true signals and lead to false positive discoveries.

Author contributions. ÅJ conceived the study. NR and ÅJ planned and designed the study. DS and ZL performed analyses and DS produced the figures. DS, ZL and ÅJ wrote the first draft if the manuscript and all authors have read and critically reviewed the manuscript.

Competing interests.

The author declares no conflict of interests.

Acknowledgement

We acknowledge all the participants and staff involved in NSPHS for their valuable contribution. The NSPHS was funded by the Foundation for Strategic Research (UG) and the European Commission FP6 (UG). Sequencing was funded by the Science for Life Laboratory (SciLifeLab), Swedish Genomes Program, which has been made available by support from the Knut and Alice Wallenberg Foundation. Sequencing was performed by NGI (National Genomics Infrastructure). Protein measurements were carried out by Olink Proteomics AB in Uppsala, Sweden. The computations and data handling were enabled by resources in project SNIC 2018/8-372, sense2016007 provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was also funded by the SciLifeLab´s Technology Development Project (TDP) program, the Swedish Medical Research Council (2019-01497) and the Swedish Heart-Lung foundation (nr. 20200687).

References

1. Nolte, I. M. *et al.* Missing heritability: Is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur. J. Hum. Genet.* **25**, 877–885 (2017).

2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
4. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
5. Quenez, O. *et al.* Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *Eur. J. Hum. Genet. 2020 291* **29**, 99–109 (2020).
6. Guan, P. & Sung, W. K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102**, 36–49 (2016).
7. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.* **8**, 639–646 (2007).
8. Malhotra, D. & Sebat, J. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
9. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur. J. Hum. Genet.* **20**, 801–805 (2012).
10. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).

11. Momtaz, R., Ghanem, N. M., El-Makky, N. M. & Ismail, M. A. Integrated analysis of SNP, CNV and gene expression data in genetic association studies. *Clin. Genet.* **93**, 557–566 (2018).
12. Park, R. W., Kim, T.-M., Kasif, S. & Park, P. J. Identification of rare germline copy number variations over-represented in five human cancer types. *Mol. Cancer 2015 141* **14**, 1–12 (2015).
13. Chattopadhyay, A. *et al.* CNVIntegrate: the first multi-ethnic database for identifying copy number variations associated with cancer. *Database* **2021**, 1–12 (2021).
14. Brezina, S. *et al.* Genome-wide association study of germline copy number variations reveals an association with prostate cancer aggressiveness. *Mutagenesis* **35**, 283–290 (2020).
15. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat. Commun.* **5**, 4684 (2014).
16. Enroth, S., Bosdotter Enroth, S., Johansson, Å. & Gyllensten, U. Effect of genetic and environmental factors on protein biomarkers for common non-communicable disease and use of personally normalized plasma protein profiles (PNPPP). *Biomarkers* **20**, 355–364 (2015).
17. Igl, W., Johansson, A. & Gyllensten, U. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* **10**, 1363 (2010).

18. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
19. Höglund, J. *et al.* Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Sci. Rep.* **9**, 16844 (2019).
20. Ameur, A. *et al.* SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur. J. Hum. Genet. 2017 2511* **25**, 1253–1260 (2017).
21. Enroth, S. B. S. *et al.* Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases. *Sci. Rep.* **8**, 5531 (2018).
22. Belyeu, J. R. *et al.* Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol.* **22**, 161 (2021).
23. Png, G. *et al.* Population-wide copy number variation calling using variant call format files from 6,898 individuals. *Genet. Epidemiol.* **44**, 79–89 (2020).
24. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).
25. Ameur, A. *et al.* De novo assembly of two swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data. *Genes* **9**, 486 (2018).

Table 1. CNVRs detected to be significantly associations with a biomarker when adjusting for the number of CNVRs identified. The table show the summary statistics for the most significant 200-bp window in each CNVR.
* CNVRs selected for validation using long-read sequencing.

CNVR #	Biomarker	Chr	Start	End	Size	Lead CNV	Beta	SE	P Value
1	CD48	1	158867600	158867800	200	1:158867600-158867800	-0.282444	0.04714985	3.1473E-09
2	FCRLB	1	161640580	161642980	2400	1:161640580-161640780	0.61738819	0.09112364	2.4227E-11
3	LY9	1	179455600	179455800	200	1:179455600-179455800	0.53598439	0.0885526	2.1686E-09
4*	GPNMB	2	89610400	89613200	2800	2:89613000-89613200	-0.9906326	0.15575469	3.3349E-10
5*	PD-L2	3	98410600	98414800	4200	3:98411600-98411800	0.40933966	0.04469867	3.9228E-19
5*	ICAM-2	3	98410600	98414800	4200	3:98410600-98410800	0.64827303	0.03946845	5.665E-53
5*	Siglec-9	3	98410600	98414800	4200	3:98411800-98413400	0.68180291	0.04171774	3.4758E-52
5*	CD200R1	3	98410600	98414800	4200	3:98411800-98413400	0.50824674	0.04448441	3.708E-28
5*	VEGFR-3	3	98410600	98414800	4200	3:98414600-98414800	0.59647014	0.04166388	1.1011E-41
5*	ICAM-3	3	98899100	98902300	3200	3:98899900-98900100	0.39646955	0.05517257	1.4524E-12
6*	AMBP	5	737870	746270	8400	5:745070-745270	-0.2838274	0.0451137	5.0537E-10
7	IL-18	5	70303300	70395300	92000	5:70393100-70393300	0.41790501	0.05609865	2.3753E-13
8	MIC-AB	6	30994100	35629600	4635500	6:32496600-32496800	0.61926004	0.06689785	3.3464E-19
9	CCL19	6	32501000	32541400	40400	6:32522200-32522400	-0.4665064	0.05366892	1.9287E-17
10	FR-gamma	11	63442300	67332555	3890255	11:67331355-67331955	-1.2325907	0.10305024	1.6409E-30
11	FR-gamma	11	67330155	67332555	2400	11:63443100-63445300	-1.0251322	0.09829072	5.1728E-24
12	CNTN1	12	45903400	45909800	6400	12:45909600-45909800	-0.2854559	0.04690077	1.7331E-09
13*	ST1A1	16	28609845	28625245	15400	16:28613645-28613845	0.78565085	0.06931645	1.6051E-27
14	CCL4	17	36387670	36399670	12000	17:36392670-36394670	0.14458813	0.02294042	4.649E-10
15	CCL15	17	39203400	39211600	8200	17:39210800-39211000	-0.9207193	0.12255953	1.4473E-13
16	SMPD1	19	35863600	35863800	200	19:35863600-35863800	0.3688639	0.0607791	1.9663E-09
17	MIA	19	41381725	41387525	5800	19:41381925-41385125	-1.2302332	0.1693139	8.578E-13
17	hK11	19	51508740	51510940	2200	19:51508940-51510740	1.40588574	0.22183536	3.8361E-10
18*	hOSCAR	19	54555500	54560700	5200	19:54558900-54559100	-0.7211762	0.06273603	1.6099E-28
19	WFDC2	20	44204035	44206635	2600	20:44204435-44205035	0.41745715	0.06811548	1.3725E-09