Article info

Article history:
Received 9 May 2019
Received in revised form 20 June 2019
Accepted 24 June 2019
Available online 2 July 2019

Abstract

Genome-wide copy-number association studies offer new opportunities to identify the mechanisms underlying complex diseases, including chronic inflammatory, psychiatric disorders and others. We have used genotyping microarrays to analyse the copy-number variants (CNVs) from 243 Caucasian individuals with Inflammatory Bowel Disease (IBD). The CNV data was obtained by using multiple quality control measures and merging the results of three different CNV detection algorithms: PennCNV, iPattern, and QuantiSNP. The final dataset contains 4,402 CNVs detected by two or three algorithms independently with high confidence. This paper provides a detailed description of the data generation and quality control.
steps. For further interpretation of the data presented in this article, please see the research article entitled 'Copy number variation-based gene set analysis reveals cytokine signalling pathways associated with psychiatric comorbidity in patients with inflammatory bowel disease'.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Data

The presented report is a description of the CNVs identified in 243 IBD patients with Caucasian ethnicity enrolled in the Manitoba IBD Cohort Study [1]. We genotyped 269 individuals with IBD using the Illumina Omni2.5M – 8 microarray. After sample quality control and population stratification analysis, we initially selected 246 IBD patients of Caucasian ethnicity. Three different CNV detection algorithms were applied to analyze the data: PennCNV [2], iPattern [3], and QuantiSNP [4]. The detected CNVs were filtered under stringent quality control criteria for their size, probe content, and algorithm-specific quality score. The quality control workflow is presented in Fig. 1. The quality control criteria and corresponding number of disqualified samples are presented in Table 1. To obtain high-confidence calls, we removed the CNVs detected by only one of the three algorithms while the CNVs detected by two or three algorithms were merged by retaining the outer boundary [5]. Numbers of CNVs detected by different algorithms are presented in Fig. 2. The examples of merging of the results obtained by three algorithms are presented on Fig. 3. Three IBD samples with extremely large number of detected CNVs were removed, which left 243 IBD samples for the further analysis. Of the remaining data, CNVs with significant overlap with the repeat rich regions, such as centromeres and telomeres, segmental duplications, and immunoglobulin regions, were excluded [2,6]. Table 2 contains numbers
of CNVs detected by each algorithm, corresponding numbers of disqualified CNVs, CNVs qualified for merging, and removed due to overlapping with the repeat-rich regions.

After the quality control and filtering, 4,402 stringent CNVs remained for the analysis; of those, 2,872 were deletions and 1,530 duplications. The chromosomal distribution of the stringent CNVs is presented on Table 3; the same data is visually presented on Fig. 4.

2. Experimental design, materials and methods

2.1. Study population

Individuals were enrolled in The Manitoba IBD Cohort Study — a population-based longitudinal study of patients with IBD [7,8]. At enrolment in the Cohort Study, participants were at least 18 years of age with a median disease duration of 4.3 years and maximal disease duration of 7 years. Participants were identified and recruited from a population-based registry, the University of Manitoba IBD Research Registry. The diagnosis of IBD was determined based on surgical, endoscopic, and histologic data. At the time of the cohort study recruitment, there were 3192 participants in the research registry. The Manitoba IBD Cohort
Study was approved by the University of Manitoba Health Research Ethics Board, and participants provided written informed consent. Blood samples for genotyping were obtained from a total of 269 IBD patients.

2.2. Genotyping

Blood samples acquired from the 269 IBD patients in the cohort were genotyped using Illumina Infinium Omni2.5-8 microarray at The Centre for Applied Genomics (TCAG) in Toronto. Rigorous quality control (QC) procedures were performed on the resulting data. The Illumina Infinium Omni2.5-8 microarray contains a total of 2,372,784 markers for SNP and CNV analyses. Samples were processed using the manufacturer’s recommended protocol; BeadChips were scanned on the Illumina BeadArray Reader using default settings. Analysis and intra-chip normalization were performed using Illumina’s GenomeStudio software v.2011.1. Probes reclustering was conducted in the GenomeStudio using the project-specific samples to produce custom cluster file, which was applied to generate LogR ratios (LRR) and B Allele Frequencies (BAF) for the CNV detection.

2.3. Intensity quality control for CNV detection

Quality control (QC) for SNPs was performed at the individual SNP level (Table 1). Samples were excluded from the analysis if they had: i) array call rate <95%; ii) standard deviation (SD) for LRR and BAF values outside mean ± 3SD for SD of an analysis batch. Closely related samples (by identity-by-descent distance for each pair of individuals), duplicates, samples with gender mismatches (by X chromosome homozygosity rate) or Mendelian error rate >1% were excluded from the analysis batch.

Table 1
QC criteria and their cutoffs for sample qualification and the corresponding number of disqualified samples.

| QC criterion     | QC cutoff for qualification | Number of disqualified samples |
|------------------|-----------------------------|-------------------------------|
| Call rate        | >95%                        | 3                             |
| SD for BAF       | Between -0.009 and 0.074    | 2                             |
| SD for LRR       | Between 0.042 and 0.131     | 18                            |
| Population outliers | European ancestry         | 18                            |
| Number of CNVs   | <145                        | 3                             |

Blood samples acquired from the 269 IBD patients in the cohort were genotyped using Illumina Infinium Omni2.5-8 microarray at The Centre for Applied Genomics (TCAG) in Toronto. Rigorous quality control (QC) procedures were performed on the resulting data. The Illumina Infinium Omni2.5-8 microarray contains a total of 2,372,784 markers for SNP and CNV analyses. Samples were processed using the manufacturer’s recommended protocol; BeadChips were scanned on the Illumina BeadArray Reader using default settings. Analysis and intra-chip normalization were performed using Illumina’s GenomeStudio software v.2011.1. Probes reclustering was conducted in the GenomeStudio using the project-specific samples to produce custom cluster file, which was applied to generate LogR ratios (LRR) and B Allele Frequencies (BAF) for the CNV detection.

2.3. Intensity quality control for CNV detection

Quality control (QC) for SNPs was performed at the individual SNP level (Table 1). Samples were excluded from the analysis if they had: i) array call rate <95%; ii) standard deviation (SD) for LRR and BAF values outside mean ± 3SD for SD of an analysis batch. Closely related samples (by identity-by-descent distance for each pair of individuals), duplicates, samples with gender mismatches (by X chromosome homozygosity rate) or Mendelian error rate >1% were excluded from the analysis batch.

![Fig. 2. Number of CNVs detected by different algorithms for each size group.](#) Labels inside the bars indicate the exact number of CNVs detected for each category after sample and CNV quality control but before CNV merging. In the legends, iPN = iPATTERN; PCNV = PennCNV; and QSNP = QuantiSNP.
2.4. Population stratification analysis

The reference population for population stratification analysis using multidimensional scaling (MDS) was obtained from Phase 3 data of 1000 Genomes Project [9]. Population stratification was
**Table 2**

**Overview of the number of CNVs after each step of quality control.** Numbers of CNVs (deletions and duplications separately, and together) detected by each of the three algorithms, number of disqualified and qualified CNVs, number of CNVs detected by each combination of algorithms, and number of CNVs removed due to overlapping with the regions of chromosomal instability.

| Algorithm                        | Initial detected | Disqualified CNVs in QC | Disqualified CNVs for merging | Detected CNVs by two or three algorithms | Remained CNVs due to overlapping with regions of chromosomal instability | Final number of CNVs |
|----------------------------------|------------------|-------------------------|-------------------------------|-----------------------------------------|---------------------------------------------------------------|---------------------|
|                                  | Duplications     | Deletions               | All CNVs                      |                                          | >70% overlap with segmental duplication                       |                     |
|                                  |                  |                         |                               |                                          | >70% overlap with centromere, telomere or immunoglobulin region |                     |
|                                  |                  |                         |                               |                                          | Number of removed CNVs                                        |                     |
|                                  |                  |                         |                               |                                          | Final number of CNVs                                          |                     |
| PennCNV                          | 3776             | 6763                    | 10539                         | 440                                      | 601                                                          | 1041                |
| iPattern                         | 3590             | 10199                   | 13789                         | 225                                      | 228                                                          | 453                 |
| QuantiSNP                        | 11194            | 18413                   | 29607                         | 643                                      | 781                                                          | 1424                |
|                                  |                  |                         |                               |                                          |                                                              |                     |
| PennCNV                          | 1299             | 3066                    | 4365                          | 440                                      | 601                                                          | 1041                |
| iPattern                         | 824              | 5775                    | 6599                          | 225                                      | 228                                                          | 453                 |
| QuantiSNP                        | 7615             | 13747                   | 21362                         | 643                                      | 781                                                          | 1424                |
|                                  |                  |                         |                               |                                          |                                                              |                     |
| PennCNV                          | 2477             | 3697                    | 6174                          | 2173                                     | 3653                                                         | 5826                |
| iPattern                         | 2766             | 4424                    | 7190                          |                                          |                                                               |                     |
| QuantiSNP                        | 3379             | 4666                    | 8245                          |                                          |                                                               |                     |

**Table 3**

**Chromosomal distribution of the stringent CNVs.** Number of deletions and duplications in three size categories, summary numbers of deletions and duplications and the total number of CNVs were presented for each chromosome excluding sex chromosomes. del: deletion and dupl: duplication.

| Chromosome | <20 kbp, del/dupl | 20–200 kbp, del/dupl | >200 kbp, del/dupl | all sizes, del/dupl | all CNV |
|------------|--------------------|-----------------------|--------------------|---------------------|--------|
| 1          | 79/36              | 38/52                 | 4/5                | 121/93              | 214    |
| 2          | 126/88             | 50/29                 | 20/4               | 196/121             | 317    |
| 3          | 141/47             | 79/83                 | 3/4                | 223/134             | 357    |
| 4          | 110/54             | 53/40                 | 1/5                | 164/99              | 263    |
| 5          | 100/36             | 56/39                 | 3/2                | 159/77              | 236    |
| 6          | 84/18              | 98/42                 | 2/8                | 184/68              | 252    |
| 7          | 138/44             | 47/28                 | 17/6               | 202/78              | 280    |
| 8          | 137/34             | 45/23                 | 2/4                | 184/61              | 245    |
| 9          | 111/44             | 30/28                 | 2/2                | 143/74              | 217    |
| 10         | 72/24              | 29/28                 | 3/12               | 104/64              | 168    |
| 11         | 125/62             | 15/13                 | 2/4                | 142/79              | 221    |
| 12         | 149/12             | 20/61                 | 0/1                | 169/74              | 243    |
| 13         | 54/12              | 37/10                 | 0/2                | 91/24               | 115    |
| 14         | 53/10              | 105/12                | 2/23               | 160/45              | 205    |
| 15         | 122/2              | 10/12                 | 0/3                | 132/17              | 149    |
| 16         | 63/50              | 41/45                 | 9/3                | 113/98              | 211    |
| 17         | 53/20              | 16/84                 | 2/4                | 71/108              | 179    |
| 18         | 80/17              | 12/4                  | 0/0                | 92/21               | 113    |
| 19         | 18/28              | 59/54                 | 3/2                | 80/84               | 164    |
| 20         | 57/19              | 36/8                  | 3/1                | 94/28               | 122    |
| 21         | 6/16               | 19/29                 | 1/0                | 26/45               | 71     |
| 22         | 11/4               | 8/26                  | 3/8                | 22/38               | 60     |
| All autosomes | 1889/677         | 903/750               | 80/103             | 2872/1530           | 4402   |
conducted using PLINK version 1.07. All ethnicity outliers were removed so that only samples of European ancestry were used in this study.

2.5. CNV calling algorithms

For comprehensive detection of CNVs in the IBD patients, we ran three CNV calling algorithms, namely, PennCNV, iPatt[3], and QuantiSNP[4]. The required data for CNV analysis, i.e., within-sample normalized fluorescence (i.e., X and Y normalized values), between-sample normalized fluorescence (i.e., Log R ratios (LRR) and B allele frequency (BAF) values) and genotypes for each sample, were exported directly from Illumina’s Genomestudio software. Only autosomal probes were used in the CNV analysis. In summary, 10539, 13789 and 29607 CNVs were detected by PennCNV, iPatt[3] and QuantiSNP, correspondingly (Table 2). We excluded the CNVs if they failed the following quality control criteria: <5 probes, <5000 bp in length and low algorithm-specific confidence score (PennCNV confidential score < 15, QuantiSNP Log Bayes Factor < 10 or iPatt[3] score < 1). After this filtering, 6174, 7190 and 8245 CNVs were identified as high quality CNVs calls for PennCNV, iPatt[3] and QuantiSNP, respectively. Each algorithm performed differently in calling CNVs of different sizes, with PennCNV being the most conservative in detection of CNVs, while QuantiSNP was least conservative except for large (>200 kbp) CNVs (Fig. 2).

2.6. CNV merging

To obtain stringent CNV calls, we merged high quality CNVs detected by at least two of the three algorithms using outer probe boundaries (Fig. 3). All CNVs detected by only one algorithm were excluded from the further analysis. As an additional step of sample QC, we excluded three samples with excessive number of stringent CNVs. We removed the samples with more than 145 CNVs (as the mean

---

Fig. 4. Chromosome view of the detected CNVs. The CNVs were presented as tiles on the corresponding genomic positions. The colour of the tile indicated the CNV type: deletions are red, duplications are blue. The height of tile stack in each genomic region corresponds to the number of CNVs; if a genomic region contains more than 30 CNVs, only 30 tiles were presented. The figure was built using the Circos tool.
number of CNVs plus 3 SD). After CNV merging, 5826 CNVs were considered as stringent. 2173 and 3653 of the CNVs are duplications and deletions, respectively (see Table 2).

2.7. CNV filtering

We further excluded CNVs that: 1) overlapped the centromere (100 kbp regions before and after centromeres) or the telomere (100 kb from the ends of the chromosome); 2) had > 70% of its length overlapping a segmental duplication using the entire segmental duplication dataset downloaded from the University of California, Santa Cruz (UCSC) Genome Browser website [11]; 3) had >70% overlap with immunoglobulin region (susceptible to somatic changes) [2,6].

The final CNV data set includes a total of 4402 CNVs, 2872 and 1530 of which were deletions and duplications, respectively (Fig. 4). The final list of stringent CNVs is available in the dbVar database at NCBI: nstd157. Of all CNVs, 58.3% were smaller than 20 kb, while 4.2% covered more than 200 kb. Chromosomal distribution of the stringent deletions and duplications in three size categories (less than 20 kb, 20–200 kb and more than 200 kb) were presented in the Table 3.

Acknowledgements

The authors wish to acknowledge Drs. Susan Walker and Mehdi Zarrei for their useful comments during the manuscript preparation. This work was supported in part by Health Sciences Centre Foundation, Mitacs, Manitoba Health Research Council and the University of Manitoba.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Frenkel, C.N. Bernstein, M. Sargent, W. Jiang, Q. Kuang, W. Xu, P. Hu, Copy number variation-based gene set analysis reveals cytokine signalling pathways associated with psychiatric comorbidity in patients with inflammatory bowel disease, Genomics (2019), https://doi.org/10.1016/j.ygeno.2019.05.001.

[2] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, M. Bucan, PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, Genome Res. 17 (2007) 1665–1674, https://doi.org/10.1101/gr.6861907.

[3] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A.C. Lionel, B. Thiruvahindrapuram, J.R. Macdonald, R. Mills, A. Prasad, K. Noonan, S. Grible, E. Prigmore, P.K. Donahoe, R.S. Smith, J.H. Park, M.E. Hurles, N.P. Carter, C. Lee, S.W. Scherer, L. Feuk, Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants, Nat. Biotechnol. 29 (2011) 512–520, https://doi.org/10.1038/nbt.1852.

[4] S. Colella, C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, A.S. Bassett, A. Seller, C.C. Holmes, J. Ragoussis, QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data, Nucleic Acids Res. 35 (2007) 2013–2023, https://doi.org/10.1093/nar/gkm076.

[5] D. Pinto, A.T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T.R. Magalhaes, C. Correia, B.S. Abrahams, J. Almeida, E. Bacchelli, G.D. Bader, A.J. Bailey, G. Baird, A. Battaglia, T. Berny, N. Bolshakova, S. Bolte, P.F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S.E. Bryson, A.R. Carson, G. Casallo, J. Casey, B.H.Y. Chung, L. Cochrane, C. Corsello, E.L. Crawford, A. Cossent, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B.A. Fernandez, S.E. Folston, E. Fombonne, C.M. Freitag, J. Gilbert, C. Gillberg, J.T. Glessner, J. Goldberg, A. Green, J. Green, S.J. Guter, H. Hakonarson, E.A. Heron, M. Hill, R. Holt, J.L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S.M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C.M. Lajonchere, J.A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B.L. Leventhal, A.C. Lionel, X.-Q. Liu, C. Lord, L. Lortspeich, S.C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C.R. Marshall, H. McConachie, C.J. McDougle, J. McGrath, W.M. McMahon, A. Merikangas, O. Migita, N.J. Minshew, G.K. Mirza, J. Munson, S.F. Nelson, C. Noakes, F.P. Noonan, G. Nygren, G. Oliveira, K. Papanikolaou, J.R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C.P. Ponting, D.J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. Rutter, L.J. Bierut, J.P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A.F. Sequeira, L. Senman, N. Shah, V.C. Shef, L. Rutter, L.J. Bierut, J.P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A.F. Sequeira, L. Senman, N. Shah, V.C. Sheffield, L. Soorya, I. Sousa, O. Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapuram, A.P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J.B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T.H. Wassink, C. Webber, R. Wexels, K. Wing, K. Wittermeyer, S. Wood, J. Wu, B.L. Yspan, D. Zurawiecki, L. Zwagenbaum, J.D. Buxbaum, R.M. Cantor, E.H. Cook, H. Coon, M.L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D.H. Geschwind, M. Gill, J.L. Haines, J. Hallmayer, J. Miller, A.P. Monaco, J.I. Nurnberger Jr., A.D. Paterson, M.A. Pericak-Vance, G.D. Schellenberg, P. Szatmari, A.M. Vicente, V.J. Vieland, E.M. Wijsman, S.W. Scherer, J.S. Sutcliffe, C. Betancur, Functional impact of global rare copy number variation in autism spectrum disorders, Nature 466 (2010) 368–372, https://doi.org/10.1038/nature09146.
