An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis

Lachlan J.M. Coin1,2,*, Dandan Cao1, Jingjing Ren1, Xianbo Zuo3,4, Liangdan Sun3,4, Sen Yang3,4, Xuejun Zhang3,4, Yong Cui3,4, Yingrui Li1, Xin Jin1,5 and Jun Wang1,6

1BGI-Shenzhen, Shenzhen, 518083, China and 2Department of Genomics of Common Disease, Imperial College London, London W12 ONN, UK, 3Institute of Dermatology and Department of Dermatology, No.1 Hospital, Anhui Medical University, Hefei, Anhui, 230032, China, 4State Key Laboratory Incubation Base of Dermatology, Ministry of National Science and Technology, Hefei, Anhui, China and 5School of Bioscience and Biotechnology, South China University of Technology, Guangzhou 510641, China

1 INTRODUCTION

Common copy number variation (CNV) in the human genome has been extensively catalogued using both array comparative hybridization (aCGH) as well as more recently with next-generation whole genome sequencing (Conrad et al., 2010; Mills et al., 2011; Redon et al., 2006). Studies have demonstrated that common CNV is prevalent in genic as well as non-genic regions of the genome and accounts for ~18% of detected genetic variation in gene expression (Stranger et al., 2007). While common copy number variants associated with complex disease have been identified (Aitman et al., 2006; de Cid et al., 2009; Diskin et al., 2009; McCarroll et al., 2008), they are vastly fewer than reported single-nucleotide polymorphism (SNP) associations. In a landmark study, the Wellcome Trust Case Control Consortium identified the majority of common CNV associated with disease. Our pipeline to identify disease-associated CNVs and to generate absolute copy number genotypes at putatively associated loci. Our method re-discovered the LCE3B_LCE3C CNV association with psoriasis (P-value = 5 × 10⁻⁶) while controlling inflation of test statistics (λ < 1). ExoCNVT est-derived absolute CNV genotypes were 97.4% concordant with PCR-derived genotypes at this locus.

Availability and implementation: ExoCNVT est has been implemented in Java and R and is freely available from www .imperial.ac.uk/medicine/people/l.coin .

Contact: wangj@genomics .org .cn or Lachlan .J .M .Coin@genomics .org .

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Exome sequencing pipeline for identifying and genotyping CNV

2 METHODS

2.1 Exon capture and sequencing

Genomic DNA of each individual was hybridized with NimbleGen 2.1 M-probe sequence capture array to enrich the exonic DNA in each library. The array captures 18,654 of the 20,091 genes deposited in Consensus Coding Region database.

DNA was randomly fragmented by Covaris to an average size of 200–300 bp, and adaptors were ligated to both ends of the resulting fragments. The adaptor-ligated DNA products were amplified with different index primers. The same amount of PCR product from pairs of samples was pooled then hybridized to each capture array following NimbleGen’s protocol, after which the exome-enriched DNA fragments were eluted from the array and amplified by ligation-mediated PCR, and non-hybridized fragments were then washed out.

Sequencing was carried out on the Illumina HiSeq 2000 platform for each captured library independently to ensure each sample had at least ~15-fold coverage. Raw image files were processed by Illumina Pipeline (version 1.3.4) for base-calling with default parameters and the sequences of each individual were generated as 90-bp reads. SOAPaligner (v2.21) (Li et al., 2008) was used to align the sequencing reads to the NCBI human genome assembly reference (build 36.3) with a maximum of three mismatches and the parameters set as ‘-a -b -D -o -u -r 1 -l 33’.

2.2 Read depth normalization

Within each capture region i, and for each individual k = 1,…,N, we record the total aligned read depth divided by the average read depth for that sample, which we label \( r_{ik} \) for each base pair \( j = 1,…,M \). We then calculate the first ‘local’ principal component \( \text{FPC}_{1k} \) for each matrix \( R_i = \{r_{ik}\} \), using an iterative scaling algorithm outlined below. This algorithm has a very low memory footprint and is highly computationally efficient. We then calculate the first 50 ‘global’ principal components of the matrix comprised of all of the local first principal components \( \{\text{FPC}_{1k}\} \), which we label \( \text{FPC}_{1}…\text{FPC}_{50} \).

This calculation is carried out using the singular value decomposition, which is not as fast or memory efficient as the iterative scaling algorithm, but is more progressive in its calculation of the covariance matrix \( R_i^{T}R_i \) Moreover, it has a very low memory footprint as the calculation in Equation (3) can be calculated via streaming one column of the depth matrix at a time. This algorithm accurately calculates the FPC but becomes progressively less accurate for lower order PCs.

2.3 Iterative scaling algorithm to calculate FPC

We have implemented an algorithm described in (Roweis, 1998) for computationally efficient calculation of the first principal component (FPC). We first centralize the matrix \( R_i \) so that each column has zero mean. We first FPC to be a vector of length \( N \) with entries randomly chosen from the interval \((0,1)\). We then iteratively update FPC via the following equations:

\[
t = \sum_{i=1}^{N} \frac{r_{i}}{\sqrt{\sum_{i=1}^{N} r_{i}^{2}}} R_{ij} \\
FPC_{i} = t_{i} \cdot R_{ij}
\]

until convergence of FPC. This algorithm avoids the computationally expensive calculation of the covariance matrix \( R_i^{T}R_i \) Moreover, it has a very low memory footprint as the calculation in Equation (3) can be calculated via streaming one column of the depth matrix at a time. This algorithm accurately calculates the FPC but becomes progressively less accurate for lower order PCs.

2.4 Association with disease status

For a fixed level of adjustment, \( H \), we then carry out association of \( FPC_{1}^{(H)} \) over all regions \( i \), with the control phenotype. The association is modelled with logistic regression, with \( P \)-values calculated using the GLM package in \( R \). We plot quantile-quantile plots for different values of \( H \) and also calculate the genomic control inflation factor \( \lambda \) in order to inspect for inflation of test statistics.

2.5 Calculating absolute copy number genotypes

For a given putative association region \( i \), and for each sample \( k \), we calculate \( r_{ij} \), the average depth in non-overlapping 100-bp windows \( j \) divided by the genome-wide average depth. We then project out the first \( H \) global PCs to calculate

\[
r_{ij}^{(H)} = r_{ij} - \sum_{h=1}^{H} \text{proj}(r_{ij}, \text{FPC}_{h})
\]

where the projection \( \text{proj}(FPC, \text{FPC}_{h}) = \left( < FPC, \text{FPC}_{h} > / < \text{FPC}, \text{FPC}_{h} > \right) \text{FPC}_{h} \)

is the projection of the FPC onto the \( h \)th global PC, using \( < > \) to denote the inner product in a \( N \) dimensional vector space.

3 RESULTS

We applied ExoCNVTest to exome sequence data collected on 800 controls and 700 psoriasis cases. Association of the uncorrected FPC was highly inflated with inflation factor \( \lambda \) of 8.5 (Fig. 1 and Table 1). In this analysis the known association \( LCE3B\_LCE3C \) was ranked 3750 out of 105,088 regions. Correction for 5 and 20 GPCs progressively reduced inflation of test statistics and also increased the rank of \( LCE3B\_LCE3C \). However, correction for 40 GPCs was required to entirely remove inflation of test statistics and reduce \( \lambda < 1 \).

In this analysis the \( LCE3B\_LCE3C \) cluster was ranked 25th, with \( P \)-value of 7.2e-5.

In order to investigate whether our approach was accurately capturing the underlying CNV present in \( LCE3B\_LCE3C \), we
investigated the correlation between the adjusted FPC in these capture regions and gold standard copy number genotypes obtained from PCR in these samples. We observed increasing correlation between the adjusted FPC and the gold-standard PCR genotypes up to 20 GPC, but not beyond (Table 2). This demonstrates that adjusting for GPCs has two effects—the first being to reduce spurious signals and the second being to increase the correlation between real and detected signals at genuine associated loci, although the optimal number of GPC correction may be different for these two goals. We also observed substantial GC biases in uncorrected association statistics, with an over-representation of significantly associated signals at $\sim$33% and 60% GC content (Fig. 2A). Correcting for 40 GPCs removed this bias (Fig. 2B).

Finally, we attempted to obtain absolute copy number genotypes at the LCE3B_LCE3C loci. We did this by averaging the per-sample read depth in 100-bp windows divided by average genome wide read depth and then projecting out the first 40 global PCs.

These normalized data were presented to cnvHap, with initial copy number cluster positions at 0, 0.5, 1, 1.5, 2 normalized read depth for 0, 1, 2, 3, 4 copy numbers, respectively. cnvHap updates the positions of these cluster positions within a generalized expectation maximization framework and then reports the per-sample copy number genotypes (Fig. 3). The reported CNV genotypes were found to be 97.4% concordant with gold standard CNV genotypes, with a missing rate of 14.7% (Table 3). These CNV genotypes had stronger association with case-control status ($P$-value of $5e^{-6}$) than FPC, demonstrating the value of using these genotypes to refine the association signal.

**Table 1.** Association of FPC with psoriasis

| GPCs | $\lambda$ | $P$-value | Rank |
|------|-----------|-----------|------|
|      | LCE3B | LCE3C | LCE3B | LCE3C |
| 0    | 8.51 | 1.3e-6 | 4.5e-7 | 5048 | 3750 |
| 5    | 1.88 | 1.3e-5 | 9.8e-6 | 238 | 204 |
| 20   | 1.28 | 2.0e-5 | 4.4e-5 | 56 | 82 |
| 40   | 0.92 | 7.2e-5 | 1.1e-4 | 25 | 33 |
| 50   | 0.91 | 1.4e-4 | 2.0e-4 | 31 | 35 |

Association $P$-value with case-control status of FPC after adjustment for 0, 5, 20, 40 and 50 GPCs. Inflation factor $\lambda$ calculated from median $\chi^2$ test statistic over all regions.

**Table 2.** Correlation of FPC with PCR genotypes

| #GPCS | Correlation with PCR |
|-------|----------------------|
|       | LCE3B | LCE3C |
| 0     | 0.857 | 0.884 |
| 5     | 0.859 | 0.885 |
| 20    | 0.894 | 0.912 |
| 40    | 0.892 | 0.908 |
| 50    | 0.888 | 0.901 |

**Fig. 1.** Quantile-quantile plot of association statistics. Observed versus expected $-\log_{10}(p)$ test-statistic quantiles. Grey line indicates equality of observed and expected under the null hypothesis of no association. Pink, blue, green, red and maroon lines indicate observed quantiles after correcting for 0, 5, 20, 40 and 50 GPCs, respectively.

**Fig. 2.** Significance versus GC-content before and after correction for 40 GPCs. Values for each of 105 088 regions indicating GC content and significance ($-\log_{10} P$-value). (A) without correction, (B) correcting for 40 GPCs.
While approaches for finding common SNP association with correction for 40 genome-wide PCs between LCE3B_LCE3C to independently replicate. Array CGH was used to discover the arrays remains very challenging, and attempts to associate CNV CNVs in large case–control cohorts using data from SNP genotyping complex disease has proven more elusive. Accurately genotyping population. cnvHap has been shown to be one of the most sensitive CNV applied cnvHap to perform joint CNV segmentation and genotyping. copy number genotypes at associated loci. To do this, we have been able in this way to easily identify CNV associations with disease. We might speculate that CNV of genes, and more specifically exons, may generate substantial phenotypic variation. Our ability to assess this effect to date has been limited by the technical challenge of array technologies. Exome sequencing, however, as it is measuring precisely this type of variation in the genome, provides an exciting opportunity to finally assess the association between gene copy number and phenotype.

ACKNOWLEDGEMENT

We would like to thank Haojing Shao for valuable discussions on Exome Sequence data.

Funding: This work was supported by Major State Basic Research Development Program of China 973 Program (2011CB809201, 2011CB809202, 2011CB809203), Major Program of National Natural Science Foundation of China (30890032) and Shenzhen Key Laboratory of Transomics Biotechnologies (CXB201108250096A). L.J.M.C. acknowledges support from the BBSRC research grant number BB/H024808/1.

Conflict of Interest: None declared.

REFERENCES

Aitman,T.J. et al. (2006) Copy number polymorphism in Fgr3 predisposes to glomerulonephritis in rats and humans. Nature, 439, 851–855.

Bonnefond,A. et al. (2012) Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat. Genet., 44, 297–301.

Conrad,D.F. et al. (2010) cnvHap: an integrative population and haplotype-based multiparameter model of SNPs and CNVs. Nat. Methods, 7, 541–546.

Conrad,D.F. et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature, 464, 704–712.

Cradick,N. et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 1,000 shared controls. Nature, 464, 713–720.

de Coimbra et al. (2009) Deletion of the late crystalline envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat. Genet., 41, 211–215.
L.J.M.Coin et al.

Diskin,S.J. et al. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature, 459, 987–991.

Hollox,E.J. et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet., 40, 23–25.

Huffmeier,U. et al. (2010) Replication of LCE3C-LCE3B CNV as a risk factor for psoriasis and analysis of interaction with other genetic risk factors. J. Invest. Dermatol., 130, 979–984.

Li,M. et al. (2011) Deletion of the late cornified envelope genes LCE3C and LCE3B is associated with psoriasis in a Chinese population. J. Invest. Dermatol., 131, 1630–1634.

Li,R. et al. (2008). SOAP short oligonucleotide alignment program. Bioinformatics, 24, 713–714.

McCarroll,S.A. et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn’s disease. Nat. Genet., 40, 1107–1112.

Mills,R.E. et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature, 470, 59–65.

Ng,S.B. et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet., 42, 30–35.

O’Roak,B.J. et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet., 43, 585–589.

Palmer,N.D. et al. (2011) Resequencing and analysis of variation in the TCF7L2 gene in African Americans suggests that SNP rs7903146 is the causal diabetes susceptibility variant. Diabetes, 60, 682–688.

Redon,R. et al. (2006) Global variation in copy number in the human genome. Nature, 444, 444–454.

Rivas,M.A. et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet., 43, 1066–1073.

Riveira-Munoz,E. et al. (2011) Meta-analysis confirms the LCE3C_LCE3B deletion as a risk factor for psoriasis in several ethnic groups and finds interaction with HLA-Cw6. J. Invest. Dermatol., 131, 1105–1109.

Roweis,S. (1998) EM algorithms for PCA and SPCA. In Jordan,M.I., Kearns, M.J. and Solla,S.A. (eds) Advances in Neural Information Processing Systems. The MIT Press, Cambridge Massachusetts.

Sathirapongsasuti,J.F. et al. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics, 27, 2646–2654.

Stauch,B.E. et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science, 315, 848–853.

Wiwantikul,V. (2010) LCE3C-LCE3B-del genotype and psoriasis: a summative meta-analysis. Acta Dermato-Venereol. Croat., 18, 150.

Xu,L. et al. (2011) Deletion of LCE3C and LCE3B genes is associated with psoriasis in a northern Chinese population, Br J. Dermatol., 165, 882–887.
Author/s:
Coin, LJM; Cao, D; Ren, J; Zuo, X; Sun, L; Yang, S; Zhang, X; Cui, Y; Li, Y; Jin, X; Wang, J

Title:
An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis

Date:
2012-09-15

Citation:
Coin, L. J. M., Cao, D., Ren, J., Zuo, X., Sun, L., Yang, S., Zhang, X., Cui, Y., Li, Y., Jin, X. & Wang, J. (2012). An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. BIOINFORMATICS, 28 (18), pp.I370-I374. https://doi.org/10.1093/bioinformatics/bts379.

Persistent Link:
http://hdl.handle.net/11343/249650

File Description:
published version

License:
CC BY