BIOINFORMATICS

Vol. 30 ECCB 2014, pages i639–i645
doi:10.1093/bioinformatics/btu475

cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data

Evangelos Bellos[1,*] and Lachlan J. M. Coin[1,2]

[1]Department of Genomics of Common Disease, Imperial College London, London W12 0NN, UK and [2]Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia

ABSTRACT

Motivation: Exome sequencing technologies have transformed the field of Mendelian genetics and allowed for efficient detection of genomic variants in protein-coding regions. The target enrichment process that is intrinsic to exome sequencing is inherently imperfect, generating large amounts of unintended off-target sequence. Off-target data are characterized by very low and highly heterogeneous coverage and are usually discarded by exome analysis pipelines. We posit that off-target read depth is a rich, but overlooked, source of information that could be mined to detect intergenic copy number variation (CNV). We propose cnvOffseq, a novel normalization framework for off-target read depth that is based on local adaptive singular value decomposition (SVD). This method is designed to address the heterogeneity of the underlying data and allows for accurate and precise CNV detection and genotyping in off-target regions.

Results: cnvOffSeq was benchmarked on whole-exome sequencing samples from the 1000 Genomes Project. In a set of 104 gold standard intergenic deletions, our method achieved a sensitivity of 57.5% and a specificity of 99.2%, while maintaining a low FDR of 5%. For gold standard deletions longer than 5 kb, cnvOffSeq achieves a sensitivity of 90.4% without increasing the FDR. cnvOffSeq outperforms both whole-genome and whole-exome CNV detection methods considerably and is shown to offer a substantial improvement over naïve local SVD.

Availability and Implementation: cnvOffSeq is available at http://sourceforge.net/p/cnvoffseq/

Contact: evangelos.bellos09@imperial.ac.uk or l.coin@imb.uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The advent of whole-exome sequencing (WES) has led to a renaissance in the field of Mendelian genetics (Bamshad *et al.*, 2011; Hoischen *et al.*, 2010; Ng *et al.*, 2010a,b). WES offers deep coverage for protein-coding regions at a lower cost than whole-genome sequencing (WGS) and has thus sparked renewed interest in elucidating rare diseases. Exome sequence analysis typically focuses on detecting previously unobserved (or very low-frequency) coding single-nucleotide polymorphisms (SNPs) and small frame-shift indels that are absent from a reference set. This approach is based on the prior hypothesis that such loss-of-function mutations are more likely to cause severe phenotypic effects commonly seen in rare diseases. A number of studies have also identified exonic copy number variation (CNV) as the underlying genetic basis for various Mendelian disorders (Lango Allen *et al.*, 2014; Rohrer *et al.*, 2013).

Exome-based CNV detection is complicated by the presence of strong batch effects introduced by the enrichment process. Various approaches have been developed to detect, genotype and perform association testing on exonic CNVs in population exome datasets (Coin *et al.*, 2012; Krumm *et al.*, 2012; Sathirapongsasuti *et al.*, 2011). The majority of these approaches operate under the assumption that most of the variation in read depth (RD) is driven by systematic bias which swamps the real signal in global noise. Exome CNV methods attempt to mitigate such bias either through control-based normalization or by identifying and removing the strongest components of variation across the population.

Most exome sequencing technologies rely on a hybridization step to enrich for DNA fragments arising from specific genomic targets. Hybridization is a sensitive but imperfect process that captures large amounts of off-target fragments along with the intended exonic regions. Although hybridization efficiency is highly variable and platform-specific, off-target sequence has been consistently shown to comprise as much as 50% of whole-exome datasets (Hedges *et al.*, 2011). Despite its potential to generate high-quality SNP genotypes in non-coding regions (Guo *et al.*, 2012), non-target reads are almost always treated as a contaminant and ignored by exome analysis pipelines. We postulate that off-target data are a rich, but overlooked, source of information that could be mined to detect intergenic CNV.

The off-target sequence coverage from a WES experiment can range from 0.5× to 3×. We have previously shown (Bellos *et al.*, 2012) that it is possible to accurately identify and genotype CNVs longer than 1 kb from low-coverage WGS data, which itself can be as little as 2× coverage. Despite the similar levels of coverage, analysis of off-target exome sequence data is confounded by the highly uneven nature of off-target coverage. The mechanism behind off-target enrichment is highly complex and contingent on both sequence properties and stochastic processes. Different parts of the genome are subject to distinct biases, such that off-target RD behaves like a combination of whole-genome and whole-exome data and cannot be handled uniformly.

To address the heterogeneity of off-target sequencing coverage, we have developed a dynamic RD normalization pipeline called cnvOffSeq. The pipeline is based on a modified version of singular value decomposition (SVD) that allows for flexible, region-specific noise reduction and signal enhancement. Global (or exome-wide) SVD has been established as a robust framework for on-target RD normalization and CNV calling using WES data. Here we propose a local adaptive SVD approach for detection and genotyping of intergenic CNVs based solely on off-target reads of WES experiments. In principle, introns could also be analysed using our framework, but we focused

[*] To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

on intergenic regions to avoid contamination from exome targets.

2 METHODS

2.1 Local adaptive SVD

The enrichment step that is essential for most targeted sequencing technologies gives rise to biases that heavily affect the depth of coverage. The resulting read depth is highly variable across target regions, rendering CNV detection problematic. Global SVD can alleviate the effects of enrichment bias and has formed the basis of numerous WES normalization strategies for CNV identification (Fromer *et al.*, 2012; Krumm *et al.*, 2012). In this paradigm, the noise is assumed to be non-random and highly correlated across samples, as the confounders (such as capture specificity and capture probe design) are shared among them. Therefore, by exome-wide application of SVD and elimination of the strongest singular components, CNV methods can effectively de-noise RD and achieve accurate segmentation.

Contrary to on-target exome data, off-target reads are unintentional by-products of the enrichment technologies and are thus subject to a mixture of sequencing biases that are highly variable and dependent on genomic context. Due to the repetitive nature of the human genome, some non-coding regions may be enriched if they have a certain degree of homology with exonic sequence. Such off-target regions will share properties with on-target data and exhibit a similar pattern of highly correlated noise. Hybridization-based whole-exome enrichment involves a washing step to remove uncaptured and poorly hybridized fragments. The efficiency of this washing step varies according to the desired level of hybridization stringency and can therefore also introduce varying amounts of off-target sequence into the results. In this case, the off-target noise is expected to be random and decorrelated among samples due to the stochastic nature of its origin. Consequently, off-target data are highly diverse and not amenable to global normalization approaches that treat noise as either entirely random or entirely correlated. To that end, we introduce a localized variant of the SVD, which segments the non-coding genome according to the observed RD noise pattern (Fig. 1). To maintain contiguity and facilitate CNV calling, we retain the coding regions in our analysis but substitute their RD with the genome-wide off-target coverage.

Our normalization framework divides the non-coding genome into regions with distinct RD noise profiles through an iterative application of SVD to genomic windows of varying size:

$$RD\_window_{size\times samples} = U\Sigma V^{*} = U\begin{bmatrix} \sigma_1 & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}V^{*}$$

The columns of matrix U represent the left-singular vectors, the rows of V^* represent the right-singular vectors and Σ is a diagonal matrix that contains the singular values in decreasing order. Each normalization window shrinks repeatedly by 100 bp until it achieves a local maximum for the largest singular value, σ_1. When such a maximum has been detected, we have identified the most 'singular' or degenerate region and the window shifts forward by the size of the shrunken window and repeats the process until the whole genome is covered. The resulting segmentation provides the means for differentiating between regions of correlated and random noise. This can be achieved by examining the relative contribution of each region's maximum singular value as defined by

$$RC_{\sigma_1} = {\sigma_1}^2 \Big/ \textstyle\sum_i {\sigma_i}^2$$

The larger the contribution the more likely the region resembles an exonic target with a highly consistent noise pattern across samples. The lower the contribution, the more randomly distributed the noise will tend to be across samples. We applied heuristic thresholds of 30 and 70% for the $RC_{\sigma 1}$ to define three region classes:

- If $RC_{\sigma 1}$<30%, then the noise appears to be random and highly represented in the lower singular components. Thus, we normalize these regions by keeping the first singular component and eliminating the rest.
- If 30% ≤ $RC_{\sigma 1}$ ≤ 70%, the noise remains mostly random but there appear to be some signs of some systematic bias. Thus, we eliminate all but the first two singular components.
- If $RC_{\sigma 1}$>70%, then the RD is dominated by systematic bias and is therefore normalized by removing the first singular component following the paradigm of WES normalization.

The local RD matrices are then reconstructed using either the truncated or the modified Σ matrix to obtain the filtered signal that can now be used for CNV detection.

2.2 Data modelling and CNV calling

Following normalization, we use the hidden Markov model (HMM) framework described in cnvHiTSeq (Bellos *et al.*, 2012) to perform

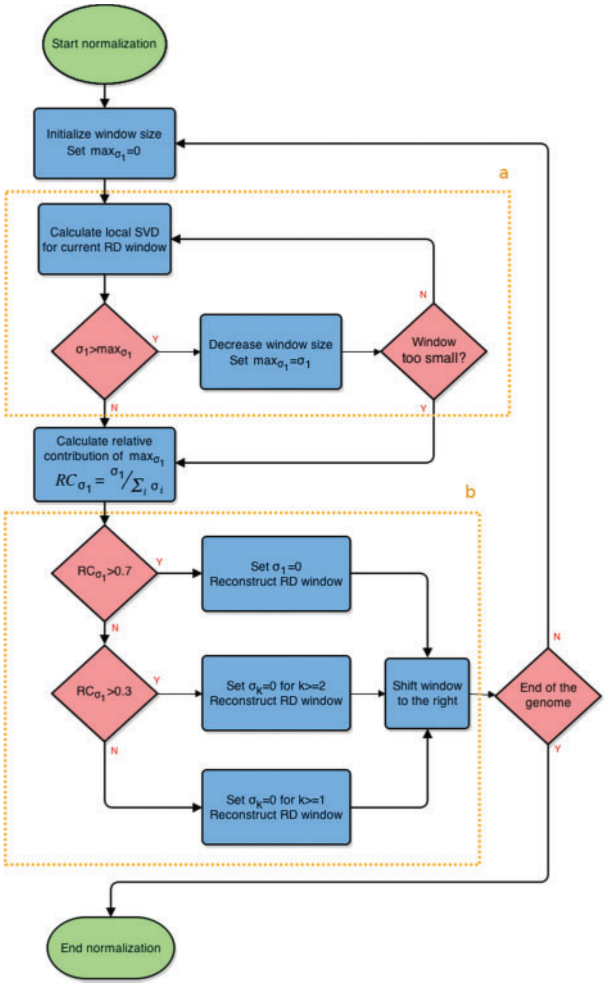
Start normalization
Initialize window size
Set max$_{\sigma_1}$=0
a
Calculate local SVD for current RD window
σ_1>max$_{\sigma_1}$
Y
Decrease window size
Set max$_{\sigma_1}$=σ_1
Window too small?
N
Y
N
Calculate relative contribution of max$_{\sigma_1}$
$RC_{\sigma_1} = \sigma_1 / \Sigma_i \sigma_i$
b
RC$_{\sigma_1}$>0.7
Y
Set σ_1=0
Reconstruct RD window
N
RC$_{\sigma_1}$>0.3
Y
Set σ_k=0 for k>=2
Reconstruct RD window
Shift window to the right
End of the genome
N
Y
N
Set σ_k=0 for k>=1
Reconstruct RD window
End normalization

Fig. 1. Flowchart describing the local adaptive SVD normalization process. (**a**) Iterative segmentation process. (**b**) Differential normalization of each segment according to the observed RD pattern

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

CNV segmentation and genotyping. Like its predecessor, cnvOffseq models the RD at the population level to achieve optimal results. The observed continuous RD is considered to be generated by the hidden underlying discrete copy number states. The emission distribution of RD is modelled using the negative binomial distribution to account for its documented overdispersed nature (Bentley *et al.*, 2008). The transition probabilities of our HMM are determined by the combination of a global transition rate matrix and a local transition rate, to capture the overall transition between copy number states across the region as well as position-specific changes. Despite the substantial de-noising achieved through our local adaptive SVD, the very low off-target coverage combined with the high RD variability informs a more stringent prior on the transition rate than previously used for WGS. As previously described, the parameters of the HMM are estimated using a generalized expectation-maximization algorithm and the most likely CNV segmentation for the trained model is obtained by the Viterbi algorithm.

2.3 Samples and datasets

Our normalization method was developed and evaluated using whole-exome sequencing samples from the 1000 Genomes Project. We only considered samples for which the 1000 Genomes consortium has also generated genome-wide CNV calls, based on the low-coverage and trio phase datasets. Furthermore, we excluded samples that exhibited off-target coverage <1×. The aforementioned criteria were fulfilled by 50 samples spanning 7 populations (Supplementary Table S1) that were sequenced either by the Broad Institute (BI) or Washington University, Genome Sequencing Center (WUGSC). For target enrichment, BI used the Agilent SureSelect All Exon assay, while WUGSC also used the NimbleGen SeqCap EZ Exome assay. The samples were sequenced either on Illumina Genome Analyzer II or HiSeq 2000 using a paired-end protocol, with the insert size varying between 100 and 400 bp across samples. Sequencing reads were aligned to the human reference genome (assembly NCBI37) using BWA (Li and Durbin, 2009). Because we are interested only in off-target data, we excluded regions that were reported as captured by either enrichment assay along with the surrounding 5 kb to avoid possible contamination. In our 50 samples, the resulting on-target coverage ranges from 49× to 248× with a mean of 96×, while the off-target coverage ranges from 1.07× to 2.66× with a mean of 1.97×. Our off-target CNV calls were compared against the 1000 Genomes gold standard set of genotyped deletions that were created by combining the predictions of five computational methods to maximize confidence.

2.4 cnvOffSeq implementation

cnvOffSeq is implemented as a collection of JAR-packaged Java tools and UNIX shell scripts. The main input of the algorithm is BAM alignment files, which are then pre-processed using SAMtools (Li *et al.*, 2009). cnvOffSeq also requires the coordinates of targeted coding regions (to be excluded) in BED format. Such files are typically provided by capture assay vendors. cnvOffSeq produces normalized RD files in text and binary format that can be disentangled from our CNV calling pipeline and used by third party segmentation algorithms. When used in conjunction with our HMM framework, cnvOffSeq generates CNV calls in text format and optional segmentation plots. The sampling density of RD is a user-specified parameter that determines the CNV breakpoint resolution and the computational requirements of the algorithm. At the default high-resolution setting of 100 bp, the normalization of chromosome 6 in 50 samples required 4GB of memory and one hour of processing time. The software is freely available from http://sourceforge.net/p/cnvoffseq.

3 RESULTS

cnvOffSeq's performance was benchmarked against chromosome 6 gold standard deletions for 50 WES samples. We excluded CNVs located less than 5 kb from the nearest gene target as well as CNVs overlapping regions that were blacklisted by the ENCODE project for anomalous RD patterns. We also eliminated gold standard deletions for which none of the 50 samples passed quality control. Our CNV detection resolution is limited by the very low off-target coverage, so we only considered events larger than 500 bp. The final gold standard dataset consisted of 104 deleted regions harbouring 497 confident calls across all samples (Supplementary Table S2). These deletions range from 606 bp to 69 281 bp with a median length of 2375 bp.

3.1 Normalization

First we sampled our RD data every 100 bp and calculated the off-target coverage for each sample. Then we applied the local adaptive SVD with a starting window size of 50 kb and a minimum window of 2.5 kb. Our dynamic normalization framework divides chromosome 6 into 5354 segments with an average length of 32.5 kb. 582 of these segments, amounting to almost 1 Mb of sequence, exhibit highly correlated RD patterns across samples. These segments resemble on-target data and are dominated by systematic bias, which can be mitigated by removing the first singular component (Fig. 2b). The remaining segments are affected by varying degrees of random noise and would suffer from loss of actual signal if normalized the same way (Fig. 2b). Instead, such segments are de-noised by eliminating low-order singular components (Fig. 2c). Consequently, global SVD methods developed for exome sequencing tend to over-correct off-target RD (Fig. 2b), while low-order component filtering tends to under-correct the most outlying observations (Fig. 2c). Depending on the local RD profile our adaptive SVD algorithm applies either first component filtering or low-order component filtering, thus combining the best of both worlds (Fig. 2d).

3.2 Benchmark

Because off-target regions comprise the majority of the genome, we first set out to compare cnvOffSeq with CNV methods designed for WGS. WGS approaches largely rely on accurate depth of coverage estimations to obtain a baseline for comparison. This is problematic in our case, as the on-target coverage is at least an order of magnitude higher than the off-target. Thus, the genome-wide coverage calculation is confounded by the on-target read depth, leading to coverage estimates that are too high for off-target regions and too low for on-target regions. One plausible solution is to exclude exome target regions from the calculations, but this is not trivial for most existing WGS methods. In fact, when we applied CNVnator (Abyzov *et al.*, 2011) to our whole-exome dataset, the results reflected the skewed coverage estimation, as most of the genome was deemed either deleted or duplicated in every sample. Therefore, to ensure an accurate and fair comparison we modified our previously described WGS framework, cnvHiTSeq, to accommodate discontiguous RD datasets, such as those represented by whole-exome off-target analyses.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

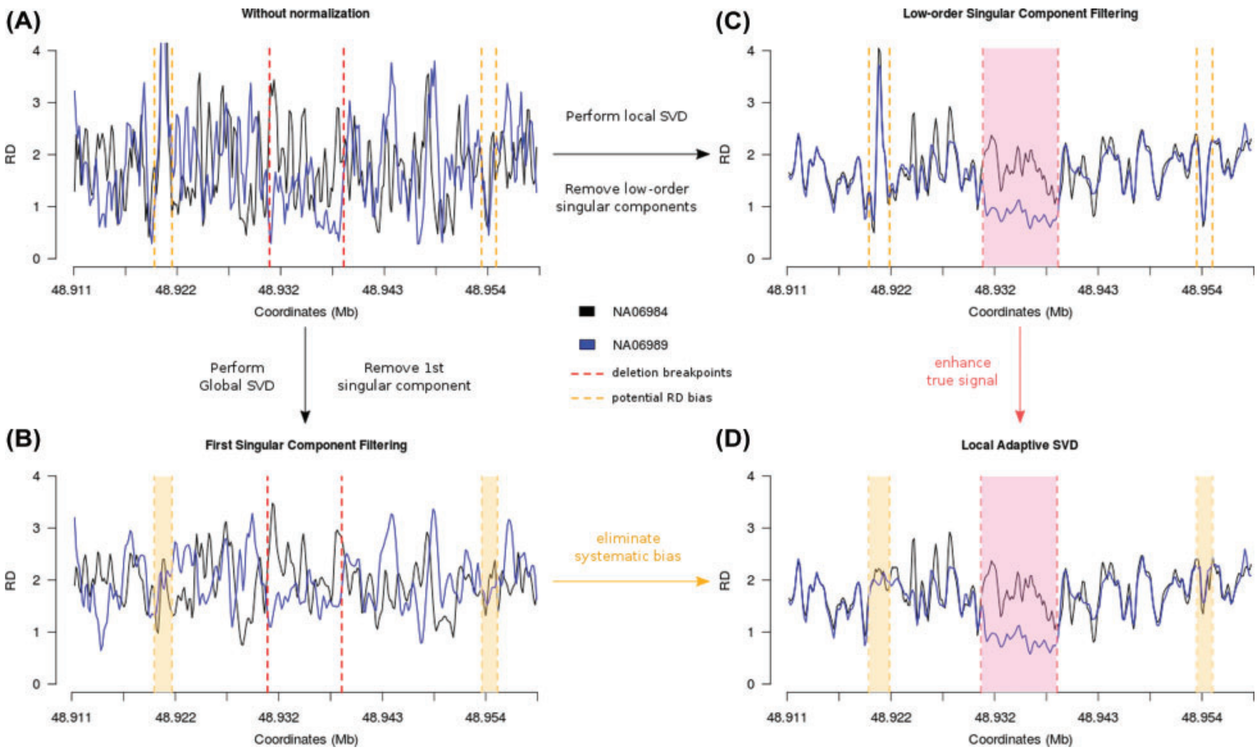

Fig. 2. Comparison of normalization techniques for two samples on chr6:48 911 000–48 959 000. The red dashed lines denote the breakpoints of a gold standard deletion that is present only in sample NA06989. (**a**) Raw RD data. The noisy nature of unnormalized RD makes the deletion difficult to detect in NA06989. The two regions denoted by the yellow dashed lines, show highly correlated RD profiles that most likely correspond to systematic bias. (**b**) Removing the first singular component mitigates the systematic bias, but also suppresses the signal of the true deletion. (**c**) Filtering low-order singular components de-noises the signal and enhances the true deletion but leaves systematic bias unaffected. (**d**) Our local adaptive SVD algorithm achieves the best results by combining the two approaches in one cohesive framework. The yellow shaded regions were normalized by filtering the first singular component (as in b) while the red shaded region was normalized by filtering low-order components (as in c)

cnvHiTSeq utilizes LOESS smoothing and GC/alignability correction to mitigate sequencing biases, but relies on the same HMM framework as cnvOffSeq. Thus, by comparing cnvHiTSeq with cnvOffSeq we provide a benchmark of 'naïve' smoothing versus local adaptive SVD. Although cnvHiTSeq detects only 26 gold standard deletions fewer than cnvOffSeq (55.0% versus 57.5% sensitivity), it suffers from much higher rates of false-positive calls (82.7% versus 5.0% false discovery rate), resulting in considerably lower accuracy (Table 1). This demonstrates the advantage of local adaptive SVD in minimizing aberrant CNV calls while enhancing the true RD signal (Fig. 3).

Next, we examined the performance of CoNIFER (Krumm *et al.*, 2012), as a representative of the global SVD approaches for whole-exome CNV detection. Like all exome CNV methods, CoNIFER requires an explicit list of target coordinates to function. In order to expand CoNIFER's functionality to off-target data, we created pseudo-targets comprising the 104 gold standard regions (extended by 50 kb on either side). These pseudo-targets were then broken down into pseudo-probes of 100 bp and 1000 bp to allow for higher resolution. The best results were obtained by using 1000 bp long pseudo-probes and removing only the first singular component. However, CoNIFER's performance falls far short of cnvOffSeq's in terms of sensitivity. CoNIFER detects only 12 of the gold standard calls (corresponding to a sensitivity of 7.6%), most of which were genotyped as homozygous deletions by the 1000 Genomes Project (Table 1). This highlights the highly stringent nature of global SVD normalization, which renders it unsuitable for off-target CNV detection.

Table 1. cnvOffSeq performance comparison

Metric	cnvOffSeq (%)	cnvHiTSeq (%)	CoNIFER (%)	Local Static SVD
Sensitivity	57.5	55.0	7.6	49.4
Specificity	99.2	79.1	99.9	97.4
PPV	95.0	17.3	96.0	82.2
NPV	89.8	95.7	81.1	88.7
FPR	0.8	20.9	0.1	2.6
FDR	5.0	82.7	4.0	17.8
Accuracy	90.4	77.4	81.3	88.0

Note: PPV, positive predictive value; NPV, negative predictive value; FPR, false positive rate; FDR, false discovery rate.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

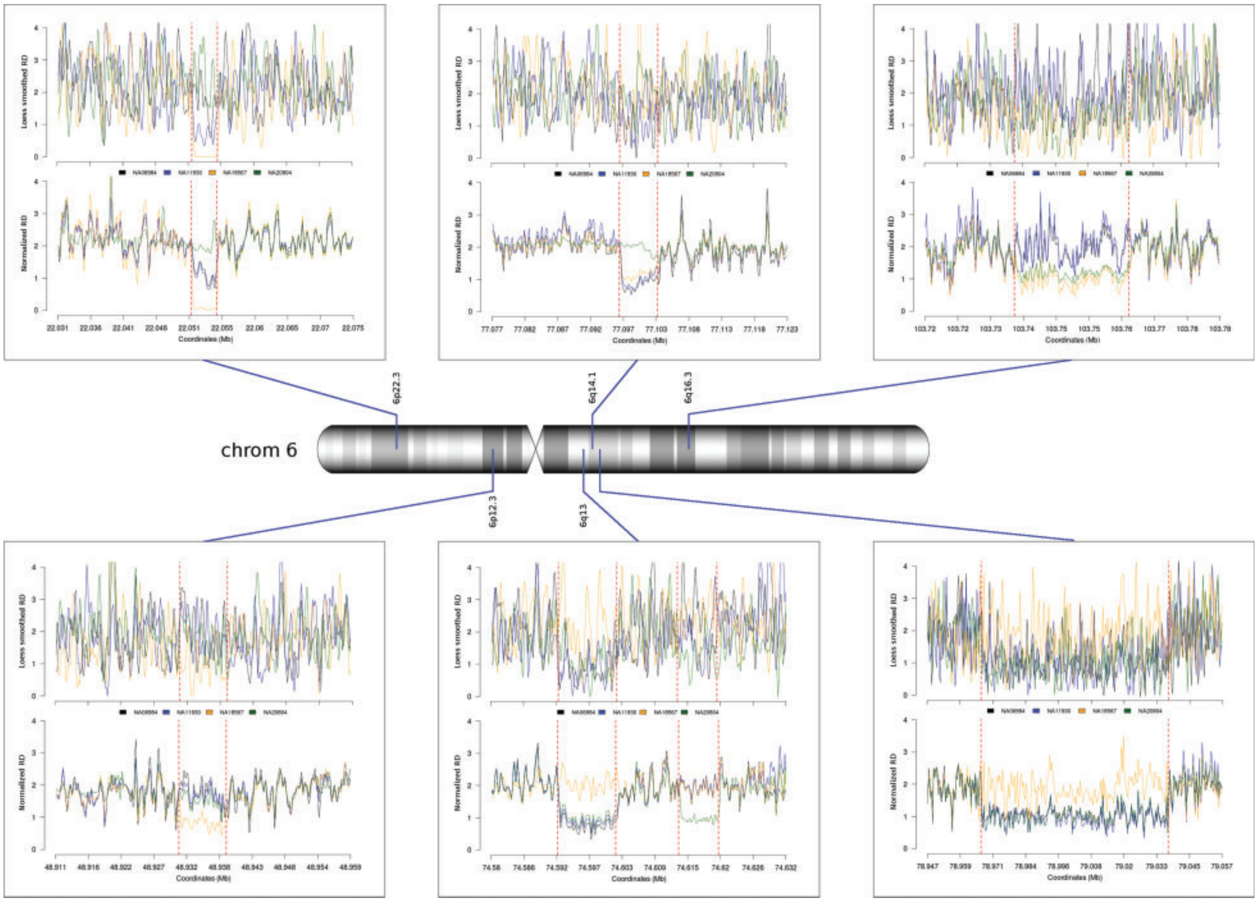

Fig. 3. Normalization results for seven gold standard regions that account for 30% of the total deletion calls. The top panels in each plot represent LOESS smoothed RD. The bottom panels represent RD that is normalized using local adaptive SVD. Each colour corresponds to a different sample while the red dashed lines denote the breakpoints of the deletions as determined by the 1000 Genomes Project

cnvOffSeq uses low-order singular component filtering to denoise a large portion of the off-target RD. To demonstrate why low-order filtering alone is not sufficient for optimal results, we also developed a variation of our normalization algorithm called local static SVD. Like its adaptive counterpart, static SVD segments the genome using patterns of RD, but then applies low-order filtering to every segment. This approach achieves a sensitivity of 49.4%, which is comparable with cnvOffSeq. As expected, however, the non-adaptive normalization does not cope well with segments dominated by systematic bias and is therefore more prone to false positives (Fig. 2b). Thus static SVD exhibits an elevated FDR of 17.8% compared to 5% for adaptive SVD (Table 1).

We also investigated the genotyping performance of cnvOffSeq as compared with cnvHiTSeq and local static SVD. CoNIFER was excluded from this comparison, as it doesn't provide absolute copy numbers. All methods in this analysis generate posterior probabilities for each CNV call, which were used to exclude results of low confidence (designated as missing). cnvOffSeq achieved an overall genotyping accuracy of 96.3% versus 73.0% for cnvHiTSeq and 90.8% for local static SVD (Table 2). Furthermore cnvOffSeq and local static SVD exhibited comparable missing rates, while cnvHiTSeq's was almost twice as high signifying a higher proportion of low quality genotype calls.

Table 2. Genotyping accuracy across methods

Method	Genotyping accuracy (%)	Missing rate (%)
cnvOffSeq	96.3	10.4
cnvHiTSeq	73.0	19.0
Local Static SVD	90.8	8.3

Finally, we set out to explore how our off-target results compare to those obtained using whole-genome data. To that end, we applied cnvHiTSeq and CNVnator to low-coverage WGS data (3×–16×) for the same 50 samples. The higher and more even coverage of the WGS dataset leads to higher overall sensitivity (75.5% for cnvHiTSeq and 62.6% for CNVnator versus 57.5% for cnvOffSeq). The improvement was more pronounced in CNVs smaller than 3 kb for which cnvHiTSeq achieves a

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 3. cnvOffSeq performance across CNV lengths

CNV length threshold (bp)	Sensitivity (%)	Specificity (%)	FDR (%)	Accuracy (%)	Number of CNV loci
>500	57.5	99.2	5.0	90.4	104
>2000	70.2	99.1	5.0	93.2	89
>3000	73.5	98.9	5.0	93.5	67
>4000	83.4	98.6	5.4	95.3	58
>5000	90.4	98.4	5.8	96.7	49
>6000	89.9	98.3	6.8	96.6	39
>7000	88.5	99.0	4.4	96.8	36
>8000	87.4	98.9	4.9	96.6	31
>9000	94.2	99.0	4.0	98.0	25
>10 000	93.9	98.9	4.2	97.8	22
>11 000	94.4	98.7	5.6	97.8	20
>13 000	95.7	98.4	5.7	97.8	16
>14 000	98.5	98.1	5.7	98.2	13
>15 000	100.0	98.2	4.3	98.7	11
>22 000	100.0	97.7	3.9	98.5	5
>25 000	100.0	97.3	9.5	97.8	2

sensitivity of 41.8% compared to 19.5% for cnvOffSeq. The FDR does not appear to benefit and remains comparable between WGS and off-target results (5.5% for cnvHiTSeq and 11.1% for CNVnator versus 5% for cnvOffSeq).

3.3 Performance versus CNV length

The inherently irregular nature of off-target RD may pose limits in the attainable resolution for CNV detection. Therefore, we investigated cnvOffSeq's performance as a function of the gold standard CNV length, by only considering events larger than a certain (variable) threshold. The results indicate that both the sensitivity and the accuracy improve significantly for longer CNVs (Table 3).

Specifically, the sensitivity exceeds 90% for deletions above 5 kb, while the specificity reaches 97%. False positives appear to be evenly distributed across CNV lengths, as the FDR remains consistent throughout. The performance deteriorates for lengths higher than 25 kb, simply because there are only two deletions in the gold standard that exceed this threshold. Thus, we conclude that cnvOffSeq is especially well suited for longer deletions (>5 kb) which are detected with very high sensitivity and consistently low false discovery rate.

4 DISCUSSION

Exome sequencing is a relatively nascent technology that has nevertheless achieved near ubiquity, particularly in the field of Mendelian genetics. Due to its cost- and time-effectiveness, exome sequencing has largely superseded both WGS and more traditional linkage studies for investigating rare genetic disorders. Furthermore, exome sequencing has proven to be a powerful diagnostic tool that has revolutionized clinical genetics. By elucidating disorders of unknown genetic aetiology, exome sequencing can inform custom treatment options, thus ushering in a new era of truly personalized medicine.

Off-target data are an integral, but overlooked, component of exome sequencing. Regardless of the underlying technology, enrichment strategies attempt to strike a balance between on-target stringency and coverage maximization. The fortunate side effect of this delicate equilibrium is off-target reads, which are often regarded as wasteful and dispensable. As exome sequencing is maturing, it will continue to generate increasing amounts of off-target data that represent an unexplored treasure of genomic information. cnvOffSeq is the first method to tap into this valuable resource for the purpose of CNV detection.

Off-target enrichment arises through various processes leading to highly heterogeneous off-target coverage. Off-target read depth is affected by a mixture of deterministic biases and random noise that require individualized treatment, but are difficult to determine a priori. As a result off-target data pose a significant challenge for CNV detection that prompted the development of a tailored data-driven normalization approach, implemented in cnvOffSeq.

We have demonstrated that our local adaptive SVD approach provides a flexible and robust framework for off-target read depth normalization that can enhance true signal while eliminating capture artifacts. cnvOffSeq clearly outperforms CNV detection methods that were designed for WGS. Furthermore, it was shown to improve upon both high-order and low-order singular component filtering, thus amounting to more than the sum of its parts.

cnvOffSeq was tested on data from both Agilent and Nimblegen capture assays but remains platform-agnostic. Like all SVD-based techniques, cnvOffSeq's ability to mitigate systematic bias improves with larger sample sizes. However, even in the absence of sufficiently large datasets, cnvOffSeq retains some of its de-noising capacity by essentially reverting to low-order singular component filtering. Finally, the modular nature of our pipeline allows for the RD normalization to be separated from the CNV calling and incorporated into future applications as a pre-processing step.

5 CONCLUSION

cnvOffSeq offers a new perspective on exome sequencing analysis and provides the tools for repurposing previously discarded surplus into meaningful data. Given the abundance of exome datasets, cnvOffSeq constitutes a powerful, novel method for investigating intergenic CNV and exploring its contribution to disease and phenotypic variation.

Funding: Research leading to these results has received funding from the European Union's seventh Framework program (EC-GA 279185). This work was also supported by the Australian Research Council (FT110100972 to L.J.M.C).

Conflict of interest: none declared.

REFERENCES

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Bamshad,M.J. *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.

Bellos,E. *et al.* (2012) cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.*, **13**, R120.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Coin,L.J. *et al.* (2012) An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. *Bioinformatics*, **28**, i370–i374.

Fromer,M. *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*, **91**, 597–607.

Guo,Y. *et al.* (2012) Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, **13**, 194.

Hedges,D.J. *et al.* (2011) Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PloS One*, **6**, e18595.

Hoischen,A. *et al.* (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.*, **42**, 483–485.

Krumm,N. *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Res.*, **22**, 1525–1532.

Lango Allen,H. *et al.* (2014) Next generation sequencing of chromosomal rearrangements in patients with split-hand/split-foot malformation provides evidence for DYNC1I1 exonic enhancers of DLX5/6 expression in humans. *J. Med. Genet.*, **51**, 264–267.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

Ng,S.B. *et al.* (2010b) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.*, **42**, 790–793.

Rohrer,J.D. *et al.* (2013) Exome sequencing reveals a novel partial deletion in the progranulin gene causing primary progressive aphasia. *J. Neurol. Neurosurg. Psychiatry*, **84**, 1411–1412.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**, 2648–2654.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016