SCIENTIFIC REPORTS

OPEN

MSeq-CNV: accurate detection of Copy Number Variation from Sequencing of Multiple samples

Received: 12 May 2017
Accepted: 16 February 2018
Published online: 05 March 2018

Seyed Amir Malekpour[1], Hamid Pezeshk[1,2,4] & Mehdi Sadeghi[3]

Currently a few tools are capable of detecting genome-wide Copy Number Variations (CNVs) based on sequencing of multiple samples. Although aberrations in mate pair insertion sizes provide additional hints for the CNV detection based on multiple samples, the majority of the current tools rely only on the depth of coverage. Here, we propose a new algorithm (MSeq-CNV) which allows detecting common CNVs across multiple samples. MSeq-CNV applies a mixture density for modeling aberrations in depth of coverage and abnormalities in the mate pair insertion sizes. Each component in this mixture density applies a Binomial distribution for modeling the number of mate pairs with aberration in the insertion size and also a Poisson distribution for emitting the read counts, in each genomic position. MSeq-CNV is applied on simulated data and also on real data of six HapMap individuals with high-coverage sequencing, in 1000 Genomes Project. These individuals include a CEU trio of European ancestry and a YRI trio of Nigerian ethnicity. Ancestry of these individuals is studied by clustering the identified CNVs. MSeq-CNV is also applied for detecting CNVs in two samples with low-coverage sequencing in 1000 Genomes Project and six samples form the Simons Genome Diversity Project.

Copy Number Variation (CNV) and balanced rearrangements such as inversions and translocations are types of the large structural variations in the human genome and other organisms. In Copy Number Variation, a gene or a genomic region appears in different number of copies in different individuals or even in different cells of the same individual. CNVs are generally referred to as a duplication or deletion of a genomic region with at least 1 kb in length. However, several clinically important CNVs are shorter than 1 kb in length. CNV results in having variations in the gene expressions and abnormalities in the human phenotypes[1]. Moreover, CNV is envisaged to be associated with many human diseases such as autoimmune disease[2], autism[1] and developmental disabilities[3], diabetes, schizophrenia[4], cancer[3] and obesity.

In the last decade, CNVs are studied via Microarray-based Comparative Genomic Hybridization (aCGH) methods[5–10]. However, the current aCGH platforms which benefit of more than 1 million genomic probes have a lower detection limit of CNVs of length ~5 kb to 25 kb[11,12]. In the recent years, Next Generation Sequencing (NGS) has provided new opportunities for the CNV studies with an unprecedented resolution[13–15]. In NGS, millions of single end or mate pair reads are generated from the sample genomes with shotgun sequencing. CNVs are then detected based on the frequency of the reads (read depth) or aberrations in the mate pairs, after mapping the short reads to the reference genome.

The majority of the current CNV detection tools analyze only one sample genome, at a time. These tools which are not capable of the simultaneous analysis of multiple samples rely either on read depth data e.g. CNV-seq[14], rSW-seq[16], m-HMM[17], BIC-seq[18], EWT[19], SegSeq[20], CNVwire[21] and ReadDepth[22] or on mate pair/split reads[23–34]. However, there are benefits in having the capability to analyze several sequencing samples, simultaneously.

Multiple sequencing reduces the effect of the systematic errors and artifacts which are attributed to the library-preparation protocol or individual sample genome characteristics[35]. There are common CNVs which are shared by complex diseases[36] and can be detected from sequencing of multiple samples. Simultaneous analysis of multiple samples allows detecting read counts variations occurring due to the noise across samples, even in genomic positions with constant copy numbers.

[1]School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran. [2]School of Biological Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran. [3]National Institute of Genetic Engineering and Biotechnology, Tehran, Iran. [4]Present address: Department of Mathematics and Statistics, Concordia University, Montreal, Canada. Correspondence and requests for materials should be addressed to H.P. (email: pezeshk@ut.ac.ir)

Therefore, to increase the detection power, more samples should be sequenced with a low sequencing coverage rather than for sequencing a few samples with high-coverage sequencing. Indeed, CNV detection methods which rely on a low-coverage sequencing data are more relevant in the future studies[37,38]. Currently, many individuals are sequenced with a low genome-wide coverage. For example, the 1000 genomes project carried out the whole-genome shotgun sequencing of 179 individuals with 2× to 4× coverages[39–41].

Currently, there are a few tools which are capable of the simultaneous analysis of several sequenced samples[37,42,43]. However, a major drawback of these tools is relying only on read depth data which results in suffering from a low power or a high false positive rate, due to the large noise in read depth signals. Moreover, these tools do not take the observed aberrations in the mate pair reads into account. Indeed, besides read depth, mate pair insertion sizes provide another source of information for the genome-wide CNV detection with an increased resolution.

In this paper, MSeq-CNV is proposed for detecting recurrent genomic deletions and duplications across multiple individuals, by the simultaneous analysis of samples. To the best of our knowledge, MSeq-CNV is the first computational tool which takes both read depth and insertion size signals in several individuals into account. The MSeq-CNV applies a mixture density to model the distribution of the read counts and the distribution of the number of mate pairs with aberrations in insertion size.

Each component in the mixture density applies a Binomial distribution to model the number of mate pairs with insertion size aberrations and a Poisson distribution to model the read counts, in each genomic region. After estimating the model parameters based on Expectation-Maximization (EM) algorithm, the posterior probability of the digitized copy number of each segment in the sample genomes is computed. The resolution of the MSeq-CNV is evaluated on a set of samples with implanted CNVs, which are constructed based on the human reference genome. Compared to the other state of the art methods, MSeq-CNV has reached an unprecedented precision and recall values which allows detecting recurrent genomic variants, accurately.

The MSeq-CNV is also applied for the CNV detection in a set of six HapMap individuals with high-coverage sequencing, including a CEU trio of European ancestry (NA12891, NA12892, and NA12878) and a YRI trio of Yoruba Nigerian ethnicity (NA19238, NA19239, and NA19240).

Methods

Assume that there are k sample genomes which are sequenced using a Next Generation Sequencing platform and mate pair reads are generated. After mapping mate pairs, the reference genome is divided into T segments of length L. Here, we aim at estimating the copy number of each genomic segment in samples 1 to k. To estimate the copy number of each sample in the t^{th} genomic segment where $t = 1, 2, \ldots, T$, this paper relies on two signals: i) number of reads which are mapped to the segment, and ii) information from the mate pair whose insertion (un-sequenced) region is passing the t^{th} genomic segment and its reads are flanking the corresponding genomic segment.

Here, number of reads which are generated from the j^{th} sample and are mapped to the t^{th} segment of reference genome (studied segment) is denoted by f_j. Also, n_j denotes the number of mate pairs which are generated from the j^{th} sample and their insertion region is passing the t^{th} segment, after mapping to it.

Signal characteristics in different genomic states. The studied segment in the j^{th} sample has one of the following copy number states i.e. {homozygous deletion, heterozygous deletion, diploid and duplications}. In each state, the characteristics of the read counts and mate pair signals which are used for the mathematical modeling are described below:

Diploid. in this state, j^{th} sample carries two copies of the corresponding segment in the reference genome. Here, a mate pair which is generated from the j^{th} sample aligns to the reference genome with a normal insertion size, distributed with the clone library insertion size distribution. Also, number of reads which are mapped to the corresponding segment in the reference genome i.e. f_j is assumed to have a Poisson distribution with parameter λ.

Heterozygous deletion. in this state, the j^{th} sample carries only one allele of the corresponding segment in the reference genome. Therefore, some mate pairs which are generated from the j^{th} sample align to the reference with a normal insertion size, distributed with the clone library insertion size distribution. Other mate pairs align to the reference genome much further apart than expected. In this state, read counts are also distributed with a Poisson distribution with a parameter of $\frac{\lambda}{2}$.

Homozygous deletion. in this state, both alleles are deleted from the j^{th} sample. Therefore, a high percentage of the mate pairs which are generated from this region will map to the reference genome much further apart than the expected insertion size distribution in the clone library. Also, we consider a Poisson distribution with a parameter of $\varepsilon\lambda$ for the read counts, where ε is assumed to be a very small value.

Duplication. in this state, the sample genome carries more than two copies of the corresponding segment in the reference genome. However, mate pairs which are generated from duplicated regions will map to the reference with the insertion size distribution of the clone library. Read counts are also distributed with a Poisson distribution with a parameter of $i\ \lambda/2$, for the samples carrying i copies.

Mathematical modeling of mate pair insertion sizes and read counts. Consider a segment of the reference genome whose copy number in the j^{th} sample is of interest, $j = 1, 2, \ldots, k$. Also, let n_j denote the total number of mate pairs which are generated from the j^{th} sample and align to the reference genome with condition

ii, as mentioned before. Also, n_{j1} denotes the number of mate pairs which are mapped to the reference with the insertion size distribution of the clone library and n_{j2} denotes the number of mate pairs which are mapped to the reference much further apart, compared to the insertion sizes in the clone library. Clearly, $n_j = n_{j1} + n_{j2}$. Here, we assumed that n_{j1} is binomially distributed as follows:

$$p(n_{j1}, n_{j2}) = \binom{n_{j1} + n_{j2}}{n_{j1}} \beta_i^{n_{j1}} (1 - \beta_i)^{n_{j2}} \tag{1}$$

where, $n_{j1} = 1, 2, \ldots, n_j$. In the above distribution, β_i indicates the probability of observing a mate pair mapped to the reference with a clone library insertion size distribution, when sample genome is in the i^{th} CNV state. Where, i = 0, 1, 2, 3, …, m corresponds to {homozygous deletion, heterozygous deletion, diploid and duplications}. Also, the maximum copy number of a genomic segment is denoted by m, i.e. $i \leq m$.

When j^{th} sample has i copies of the studied segment of the reference genome, read count f_j follows a Poisson distribution with a parameter of $\theta_i \lambda$:

$$p(f_j) = e^{-\theta_i \lambda} \frac{(\theta_i \lambda)^{f_j}}{f_j!} \tag{2}$$

where, $\theta_0 = \varepsilon\lambda$, and $\theta_i = i\lambda/2$, for $i \geq 1$.

Also, from a total number of k samples, let α_i denote the percentage of samples which have i copies of the studied segment of the reference genome. Taking the above descriptions into account, the probability of observing (f_j, n_{j1}, n_{j2}) in the j^{th} sample genome can be written as follows:

$$\begin{aligned} p(f_j, n_{j1}, n_{j2}) &= \sum_{i=0}^{m} \alpha_i p(f_j, n_{j1}, n_{j2} | state\ i) \\ &= \sum_{i=0}^{m} \alpha_i e^{-\theta_i \lambda} \frac{(\theta_i \lambda)^{f_j}}{f_j!} \left[\binom{n_{j1} + n_{j2}}{n_{j1}} \beta_i^{n_{j1}} (1 - \beta_i)^{n_{j2}} \right] \end{aligned} \tag{3}$$

However, it should be added that in the above formulations n_{j1} and n_{j2} are not known and they depend on the unknown parameter β_i. Also, the estimation of n_{j1} and n_{j2}, requires estimating the probability of each insertion size to be distributed with the clone library insertion size distribution. For this purpose, let o_{jr} denote insertion size of the r^{th} mate pair which was generated from the j^{th} sample and was mapped to the studied segment of the reference genome. Where $r = 1, 2, \ldots, n_j$ and j = 1, 2, …, k. Consequently, a random variable z_{jr} is corresponded to each mate pair insertion size o_{jr}:

$$z_{jr} = \begin{cases} 1 \text{ if } o_{jr} \text{ comes from the insertion size distribution of the clone library} \\ 0 \text{ if } o_{jr} \text{ comes from a shifted insertion size distribution of the clone library} \end{cases}$$

where, $n_{j1} = \sum_{r=1}^{n_j} z_{jr}$ and $n_{j2} = \sum_{r=1}^{n_j} (1 - z_{jr})$. To estimate the expected value of n_{j1} and n_{j2}, we calculate the probability of having a z_{jr} equal to 1, for $r = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$, see Supplementary file 1 for a detailed description.

Now, a Dirichlet prior distribution is defined for the parameter vector $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)$:

$$p(\boldsymbol{\alpha}) \propto \prod_{i=0}^{m} \alpha_i^{\gamma_i - 1} \tag{4}$$

where, $\alpha_0 = 1 - \sum_{i=1}^{m} \alpha_i$ and $\gamma_s = \sum_{i=0}^{m} \gamma_i$. Also, the prior of each β_i is considered to be a beta distribution as follows:

$$p(\beta_i) = \frac{\Gamma(\nu_{i1} + \nu_{i2})}{\Gamma(\nu_{i1})\Gamma(\nu_{i2})} \beta_i^{\nu_{i1} - 1} (1 - \beta_i)^{\nu_{i2} - 1} \tag{5}$$

where, i = 0, 1, 2, …, m. Moreover, the prior distribution of λ is considered to be a uniform distribution, over the interval of $(0, t)$, where t is large enough.

Model parameters. There are a number of parameters in the above mathematical model which have to be estimated. These parameters include λ, the average read counts in a genomic segment of diploid state. The parameter vector $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)$ represents the percentage of samples with copy numbers 0, 1, …, m of the studied segment of the reference genome. Also, for a sample genome with copy number state i, β_i indicates the proportion of the mate pairs which are mapped to the reference genome much further apart than expected under the clone library insertion size distribution.

The parameters of the prior distribution over $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)$ are given values based on information from genome-wide CNV percentage. Since a high percentage of genomic segments in each sample are expected to be in diploid state, γ_2 is given a value much higher than $\gamma_i, i \neq 2$. Also, a beta distribution is defined as a prior distribution over each β_i, i = 0, 1, 2, …, m. The parameters of the beta distribution i.e. ν_{i1} and ν_{i2} are given values

		Real state								
		Heterozygous deletion	Homozygous deletion	Diploid	3 copies	4 copies	5 copies	Sum	Precision	Recall
Predicted state	Heterozygous deletion	26,807	545	4,529	26	12	2	31,921	0.84	0.94
	Homozygous deletion	1,682	86,979	20,474	172	74	24	109,405	0.80	0.99
	Diploid	0	215	954,226	3,106	549	73	958,169	1.00	0.97
	Duplication, 3 copies	0	344	3,124	8,953	554	45	13,020	0.69	0.71
	Duplication, 4 copies	0	0	0	371	5,460	96	5,927	0.92	0.75
	Duplication, 5 copies	0	0	0	0	654	904	1,558	0.58	0.79
	Sum	28,489	88,083	982,353	12,628	7,303	1,144			

Table 1. Precision and recall values of MSeq-CNV are evaluated in 1,120,000 genomic segments of length 150 bp (168 million base pairs are evaluated), for a genome-wide coverage of 10×. In columns 3 to 8, predicted states are reported versus the real states of the genomic segments.

based on the expected number of mate pairs which are mapped to the reference much further apart, compared to the clone library insertion size distribution. In genomic diploid state and segments with an elevated number of copies $\nu_{i1} \gg \nu_{i2}$, for i = 2, 3, …, m. In genomic segments with heterozygous deletion $\nu_{11} \cong \nu_{12}$ and in genomic segments with homozygous deletions $\nu_{01} \ll \nu_{02}$.

Parameter estimation. MSeq-CNV applies the Expectation-Maximization (EM) algorithm, for parameter estimation. The parameter estimation details are given in Supplementary file 1.

Parameter initialization in EM algorithm. λ is initialized based on the number of reads that are expected to be generated from a genomic segment with diploid state, after taking the sequencing coverage into account. For example, for a coverage of 5×, a read length of 100 bp and genomic segments of length 100 bp, λ is initialized with a value of 20. Also, $\alpha_2 = 0.90, \alpha_i = 0.02$ for i = 0, 1, 3, 4, 5, and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 1, \beta_0 = 0$, and $\beta_1 = 0.5$ are taken as the start point of the parameters in the EM algorithm. In the j^{th} sample, μ_{j1}, μ_{j2} are initialized by comparing the mate pair insertion sizes with the clone library insertion size distribution.

In this study, we have considered a segment size of 150 bps in the simulated data analysis and a segment size of 100 bps in the real data analysis. Considering a longer segment size decreases the running time of the algorithm, with the cost of lower resolution. The methods which are compared to MSeq-CNV are also implemented with the same segment size, as MSeq-CNV. However, there are other methods which are specialized in detecting genomic rearrangements and tandem duplications using paired reads which get the nucleotide resolution breakpoint[25,29].

Data availability. BAM (Binary Alignment/Map) files of the alignment of the mate pair reads to the build 36 (hg18) of the human reference genome are available at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/. R program of MSeq-CNV and a detailed procedure for its running is available at https://github.com/CNVdetection/MSeq-CNV. The list of detected CNVs in the studied individuals is also submitted to this webpage.

Results

Implementation To Real Data of The Human Reference Chromosome. We have constructed 40 sample genomes with implanted CNVs, from chromosome 3 of the human reference genome. After duplicating chromosome 3 of the reference genome, it was altered with implanted CNVs of length 250 bp, 500 bp, 750 bp, 1 kb, 1.5 kb, 2 kb, 2.5 kb, 3 kb, 3.5 kb, 4 kb, 4.5 kb, 5 kb. The position of each CNV is randomly chosen so that CNVs do not overlap along chromosome. After determining the CNV positions on the reference genome, CNVs are implanted into each sample genome.

Indeed, for each CNV region which is implanted in the reference genome, distributions and characteristics of CNVs across sample genomes are determined based on a previous analysis of the HapMap individuals.

Based on the characteristics of the HapMap individuals, 80% of the implanted CNVs were of type loss in which deletions occurs in some sample genomes. Also, 15% of the CNV regions were of type gain in which some sample genomes have an elevated number of copies. The other 5% of the implanted CNV regions were of type mixed in which sample genomes may either have copy loss or copy gain.

In each CNV region, the copy number of each sample was also drawn from the copy number distribution in HapMap individuals. For a genomic loss region, a sample has copy numbers 2, 1, and 0 with probabilities 0.8, 0.15 and 0.05, respectively. For a genomic gain region, a sample has copy numbers 2, 3, 4, and 5 with probabilities 0.85, 0.08, 0.06 and 0.01, respectively. Also, for a CNV region of type mixed, a sample has copy numbers 0, 1, 2, 3, 4 with probabilities 0.04, 0.16, 0.67, 0.11 and 0.02, respectively.

After constructing the sample genomes, MAQ is applied for generating mate pair reads from each sample genome. Mate pairs are then mapped to the human reference genome. After dividing the human reference genome into segments of length 150 bp, MSeq-CNV is applied for detecting CNVs in the corresponding segments of the constructed sample genomes. In Table 1, the performance of MSeq-CNV is reported for each CNV state i.e. homozygous deletion, heterozygous deletion, and duplications, for a genome-wide sequencing coverage of 10×.

		Coverage								
		1×			5×			10×		
		precision	recall	F-score	precision	recall	F-score	precision	recall	F-score
Duplication	MSeq-CNV*	0.31	0.36	**0.33**	0.69	0.80	**0.74**	0.79	0.82	**0.80**
	MSeq-CNV (3 copies)	0.01	0.01	0.01	0.49	0.57	0.53	0.63	0.72	0.67
	MSeq-CNV (4 copies)	0.26	0.37	0.31	0.69	0.59	0.64	0.82	0.66	0.73
	MSeq-CNV (5 copies)	0.25	0.55	0.34	0.32	0.80	0.46	0.39	0.81	0.53
	rSW-seq	0.00	0.00	0.00	0.76	0.43	0.55	0.87	0.72	0.79
	CNV-seq	0.12	0.43	0.19	0.72	0.58	0.64	0.97	0.60	0.74
	cn.MOPS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Deletion	MSeq-CNV (heterozygous deletion)	0.84	0.96	0.90	0.85	0.93	0.89	0.94	0.94	0.94
	MSeq-CNV (homozygous deletion)	0.80	0.85	0.82	0.81	0.98	0.89	0.86	0.99	0.92
	MSeq-CNV (hetero + homo)**	0.83	0.90	**0.86**	0.83	0.99	**0.90**	0.89	1.00	**0.94**
	rSW-seq	0.00	0.00	0.00	0.87	0.56	0.68	0.92	0.87	0.89
	CNV-seq	0.66	0.58	0.62	0.98	0.78	0.87	0.99	0.87	0.93
	cn.MOPS	1.00	0.23	0.37	1.00	0.23	0.37	1.00	0.23	0.37
Diploid	MSeq-CNV	0.98	0.96	**0.97**	0.99	0.97	**0.98**	1.00	0.98	**0.99**
	rSW-seq	0.87	1.00	0.93	0.93	0.98	0.95	0.97	0.99	0.98
	CNV-seq	0.94	0.90	0.92	0.97	0.99	**0.98**	0.98	1.00	**0.99**
	cn.MOPS	0.90	1.00	0.95	0.90	1.00	0.95	0.90	1.00	0.95

Table 2. MSeq-CNV in comparison with rSW-seq, CNV-seq and cn.MOPS, for whole-genome sequencing coverage of 1×, 5×, 10×. The average precision and recall values are calculated across five different repeats of the whole study. The highest F-scores are indicated in bold. *Precision, recall and F-score are evaluated over all genomic segments with amplifications i.e. 3, 4 and 5 copies. **Hetero + homo stands for copy losses, both heterozygous and homozygous deletions.

		Coverage		
		1×	5×	10×
All regions**	MSeq-CNV	**0.94**	**0.96**	**0.98**
	rSW-seq	0.87	0.92	0.96
	CNV-seq	0.86	0.95	**0.98**
	cn.MOPS	0.90	0.90	0.90

Table 3. The overall performance of the MSeq-CNV in comparison with rSW-seq, CNV-seq and cn.MOPS. For each method, the average of overall accuracies over five different runs of the whole study is given in each cell. For each coverage, the highest accuracies are indicated in bold. **Number of nucleotides whose states were correctly predicted is divided by the total number of genomic nucleotides (genome length).

The performance of MSeq-CNV is also compared to the central CNV detection tools i.e. rSW-seq, CNV-seq and cnMOPS. These tools are selected for comparisons because of their high resolution and their capability in detecting both genome-wide deletions and duplications[37]. It should be added that rSW-seq and CNV-seq are not capable of detecting the digitized copy number of genomic regions i.e. these tools do not discriminate heterozygous deletions from homozygous deletions. However, MSeq-CNV resembles cn.MOPS in detecting the digitized copy number of each CNV region.

For calculating precisions and recalls, the whole simulation study was repeated five times for each setting and the average results across five repeats are summarized in Table 2. In this table, for a genome-wide sequencing coverage of 1×, 5×, 10×, MSeq-CNV is compared to the other tools in each genomic state i.e. homozygous deletion, heterozygous deletion, diploid and duplications. The F-score which is the harmonic mean of the precision and recall values are also calculated in Table 2, for each CNV state.

As shown in Table 2, for all coverage values and according to the F-score, the performance of MSeq-CNV has been superior to the compared tools in detecting genomic regions with deletions and duplications. In regions with diploid copies, MSeq-CNV outperformed the compared tools for a coverage of 1×. Also for a coverage of 5× and 10×, MSeq-CNV and CNV-seq are both ranked as the best tools in detecting regions with diploid copies.

In Table 3, the overall performance of MSeq-CNV is compared to the other tools. To calculate the overall performance of each tool in estimating the correct copy number state, number of nucleotides whose states were

		Allele Frequency			
		0–5	6–10	11–15	16–25
CNV region	Copy Loss	0.99	0.99	0.99	0.99
	Copy Gain	0.97	0.93	0.86	0.86
	Mix (Copy gain + Copy loss)	0.96	0.96	0.96	0.94

Table 4. The overall accuracy of MSeq-CNV, in terms of allele frequency in genomic regions with CNVs. Allele frequencies are categorized into 4 groups i.e. 0–5, 6–10, 11–15, and 16–25, in the simulated data of 40 samples.

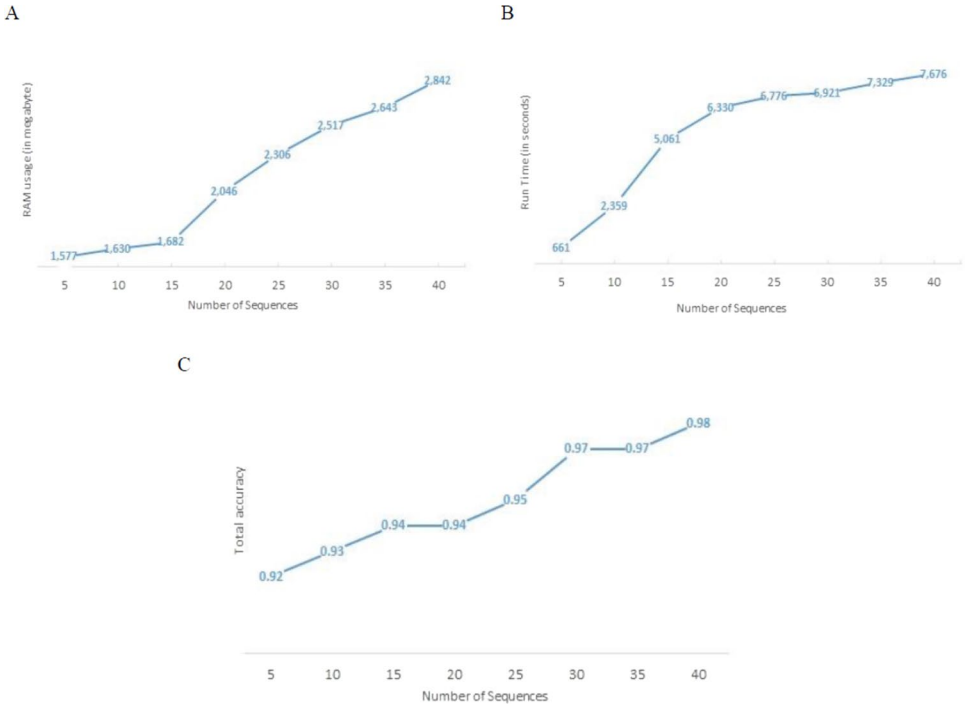

Figure 1. Ram usage, run time and accuracy of MSeq-CNV is reported, in terms of sequence numbers, in a 5 Mega bp of the genomic region and for a coverage of 10×. These results are achieved using parallel programming version of the MSeq-CNV on a 64-bit windows operating system with Intel Core(TM) i7-4710HQ CPU @3.5 GHz processor. (**A**) Ram usage of MSeq-CNV in terms of sequence numbers. **(B**) Run time of MSeq-CNV in terms of sequence numbers. (**C**) The overall accuracy of MSeq-CNV, in terms of number of analyzed individuals.

correctly predicted is divided by the genome length. As indicated in Table 3, the overall performance of the MSeq-CNV is superior to rSW-seq, CNV-seq and cnMOPS, for a coverage of 1× and 5×. Also, MSeq-CNV and CNV-seq outperformed rSW-seq and cn.MOPS with an overall accuracy of 0.98, for a 10× coverage.

Performance of MSeq-CNV is also evaluated in terms of allele frequency of CNVs. As shown in Table 4, in CNV regions with copy loss, accuracy does not change with an increase in allele frequency i.e. MSeq-CNV is accurate in detecting genomic deletions. However, in CNV regions of type copy gain or mixed, overall accuracies decrease with an increase in allele frequency. This is associated with lower accuracies in detecting genomic duplications, compared to the other genomic regions.

Figure 1 shows the RAM usage, running time and overall accuracy of MSeq-CNV in terms of sequence numbers, i.e. number of individuals which are compared to each other, for 10× sequencing coverage. To obtain these results, 4 computer cores were applied for running the parallel programming version of the MSeq-CNV, on a 64-bit windows operating system with Intel Core(TM) i7-4710HQ CPU @3.5 GHz processor. As shown in Fig. 1A,B, RAM usage and running time of MSeq-CNV both increase, with an increase in sequence numbers. However, as shown in Fig. 1C, analyzing more sample genomes at a time has a positive effect on the overall accuracy of MSeq-CNV.

Results From The High-Coverage Data of The 1000 Genomes Project. MSeq-CNV is applied for the CNV detection in the genome of six HapMap individuals. These genomes which are sequenced with a high coverage as part of the 1000 Genomes Project (http://www.1000genomes.org) consist of a CEU trio of European ancestry (NA12891, NA12892 and NA12878) and a YRI trio of Yoruba Nigerian ethnicity (NA19238, NA19239 and NA19240).

	Individual	Number of deletion calls	Total size of deletion calls in Mega bp	Number of duplication calls	Total size of duplication calls in Mega bp	Number of CNVs	Total size of CNV calls in Mega bp
CEU trio, 1000Genomes Project, high-coverage sequencing	NA12891	79,209	143.442	170,195	277.842	249,404	421.284
	NA12892	85,496	151.703	163,393	238.777	248,889	390.479
	NA12878	86,628	155.460	162,534	254.734	249,162	410.194
Average per individual		83,778	150.201	165,374	257.117	249,152	407.319
YRI trio, 1000Genomes Project, high-coverage sequencing	NA19238	116,971	209.685	161,056	249.925	278,027	459.610
	NA19239	81,688	156.520	164,541	296.874	246,229	453.394
	NA19240	76,169	146.339	175,100	342.556	251,269	488.895
Average per individual		91,609	170.848	166,899	296.452	258,508	467.300
1000Genomes Project, low-coverage sequencing	NA12761	56,415	88.675	89,047	102.820	145,462	191.495
	NA12762	47,541	76.609	71,533	92.730	119,074	169.339
Average per individual		51,978	82.642	80,290	97.775	132,268	180.417
American, SGDP	USA, LP6005592-DNA_H03	38,172	65.276	133,238	237.788	171,410	303.064
Asian, SGDP	TAIWAN, LP6005442-DNA_E07	70,114	102.897	127,321	220.741	197,435	323.638
	TAIWAN, LP6005443-DNA_G05	52,834	83.333	130,605	234.064	183,439	317.397
	INDIA, LP6005519-DNA_A04	131,537	186.819	130,868	244.590	262,405	431.409
	INDIA, LP6005519-DNA_A05	204,704	293.091	137,205	252.344	341,909	545.435
Average per individual		114,797	166.535	131,500	237.935	246,297	404.469
European, SGDP	FINLAND, LP6005592-DNA_D01	56,491	88.131	129,346	235.698	185,837	323.829

Table 5. Number of CNV calls which are made by MSeq-CNV in the six high-coverage sequencing data the 1000Genomes Project i.e. NA12891, NA12892, NA12878 (CEU trio), NA19238, NA19239, NA19240 (YRI trio) and in the low-coverage sequencing data of NA12761, NA12762, and six individuals from the Simons Genome Diversity Project (SGDP).

BAM (Binary Alignment/Map) files of the alignment of the mate pair reads to the build 36 (hg18) of the human reference genome are downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/. Mate pair reads with low mapping qualities ($Q < 25$) are then filtered out using SAMtools (samtools.sourceforge.net). Then, MSeq-CNV is applied for detecting deletions and duplications in the genome of CEU and YRI individuals, simultaneously.

In past studies[44], to reduce the false discovery rate, CNVs' callset were heavily pre-filtered and only a high confident set of CNVs were reported. In our Bayesian framework, to reduce the false positive rate, we report a set of CNVs with high posterior probabilities. Indeed, those CNVs with posterior probabilities lower than a fixed threshold i.e. 0.5 are filtered out from the callset.

In Table 5, number of CNVs and their total size (in Mega bp) are reported for deletion and duplication calls with at least 1 kb in size. As indicated in Table 5, numbers of CNVs in CEU trio NA12891, NA12892 and NA12878 are respectively 249, 404 (421.284 Mega bp), 248, 889 (390.479 Mega bp) and 249, 162 (410.194 Mega bp). Also, numbers of CNVs in YRI trio NA19238, NA19239 and NA19240 are respectively 278, 027 (459.610 Mega bp), 246, 229 (453.394 Mega bp) and 251, 269 (488.895 Mega bp). Also, similar to previous estimations[45], CNV calls in NA12891, NA12892, NA12878, NA19238, NA19239 and NA19240 respectively cover 13.02%, 12.07%, 12.68%, 14.21%, 14.02% and 15.11% of the human genome.

The average number of CNVs in YRI trio i.e. 258, 508 (467.300 Mega bp) is slightly higher than the average calls in CEU i.e. 249, 152 (407.319 Mega bp). Also, as indicated in Table 5, the average number of deletion and duplication calls in YRI individuals (91, 609 and 166, 899) are both more than CEU trio (83, 778 and 165, 374), indicating the increased diversity of the African individuals in comparison with CEUs[46]. Moreover, in the studied individuals genomic deletions are less common compared to duplications[44,46,47].

Total number of CNV calls in each chromosome is plotted in Fig. 2A, for each HapMap individual. Numbers of deletion and duplication calls are also given in Table S.1, for each chromosome. See Table S.2 for the size (in Mega bp) of deletion and duplication calls.

To investigate the validity of the CNV calls, their overlap with the Database of Genomic Variants (DGV)[48], http://dgv.tcag.ca/dgv/ is studied. DGV includes 8, 599 CNVs from 40 HapMap individuals which are validated experimentally using aCGH methods.

The overlap of the detected CNVs with DGV are determined by the number of calls and also by the size of overlap, in base pairs. CNV calls in CEU trio NA12891, NA12892 and NA12878 overlap with DGV respectively with a ratio of 0.61, 0.62 and 0.61, for the number of calls. Also, a base which is called as a CNV in NA12891, NA12892 and NA12878 overlap with a base in DGV respectively with a ratio of 0.60, 0.61 and 0.61. The YRI individuals NA19238, NA19239 and NA19240 overlap with DGV respectively with a ratio of 0.62, 0.61 and 0.62 for the number of calls, and 0.61, 0.59 and 0.60 for the base pairs. Therefore, more than a half of CNV calls which are made by MSeq-CNV are previously validated using aCGH methods.

Size distribution of CNVs are also shown in Fig. 3A,B, respectively for the deletion and duplication calls. Clearly, the numbers of deletion and duplication calls decrease exponentially, with an increase in CNV size. As shown in Fig. 3A, deletion size distributions almost overlap in all studied individuals. Duplication size distributions are also very similar in all individuals, with CEU individuals having more CNVs of smaller sizes.

Figure 2. Number of CNV calls which are made by MSeq-CNV. (**A**) In the high-coverage sequencing data of HapMap individuals NA12891, NA12892, NA12878, NA19238, NA19239 and NA19240, from 1000 Genomes Project. (**B**) In the low-coverage sequencing data of NA12761 and NA12762, from 1000Genomes Project. (**C**) In the six individuals from the Simons Genome Diversity Project (SGDP).

Moreover, we applied the hierarchical clustering algorithm to the matrix of CNV regions which are identified in the genome of six HapMap individuals. As shown in Fig. 4, although no information about the individual's identities are used in the hierarchical clustering, the algorithm has correctly segregated the ancestry of the six individuals in two groups. While one group includes the CEU individuals NA12891, NA12892 and NA12878 with European ancestry, the other group includes YRI individuals NA19238, NA19239 and NA19240 with Nigerian ancestry.

Results From The Low-Coverage Data of The 1000 Genomes Project. MSeq-CNV is also applied for the CNV detection form the low-coverage data of two individuals i.e. NA12761 and NA12762, from 1000 Genome project. After downloading BAM files of the alignment of mate pair reads to the human reference genome, mate pairs with low mapping qualities ($Q < 25$) are filtered out.

MSeq-CNV called a total number of 145, 462 (191.495 Mega bp) and 119, 074 (169.339 Mega bp) CNVs in the genomes of NA12761 and NA12762, respectively. Also, in both individuals, genomic deletions are less common compared to the duplications[44,46,47]. Details of deletion and duplication calls are given in Table 5 and Table S.3.

The low number of CNV calls in the genome of NA12761 and NA12762 is potentially associated with lower accuracies in detecting genomic CNVs, especially duplications, from low-coverage sequencing data of NA12761 and NA12762 (see Table 2).

The overall number of detected CNVs in each chromosome is shown in Fig. 2B, for NA12761 and NA12762. Detected CNVs in NA12761 and NA12762 overlap with DGV respectively with a ratio of 0.65 and 0.66 for the number of calls, and 0.66 and 0.67 for base pairs.

Results From The Simons Genome Diversity Project (SGDP). MSeq-CNV is also applied for CNV detection in the genome of six individuals from the Simons Genome Diversity Project[49] i.e. LP6005592-DNA_H03 (USA), LP6005442-DNA_E07 (Taiwan), LP6005443-DNA_G05 (Taiwan), LP6005519-DNA_A04 (India), LP6005519-DNA_A05 (India), and LP6005592-DNA_D01 (Finland).

As indicated in Table 5, in the analyzed individuals from SGDP, the lowest number of CNVs are called in the LP6005592-DNA_H03 (USA)[46] (171, 410 CNVs with a total size of 303.064 Mega bp). Two East Asian individuals LP6005443-DNA_G05, LP6005442-DNA_E07(form TAIWAN) and West Eurasian individual

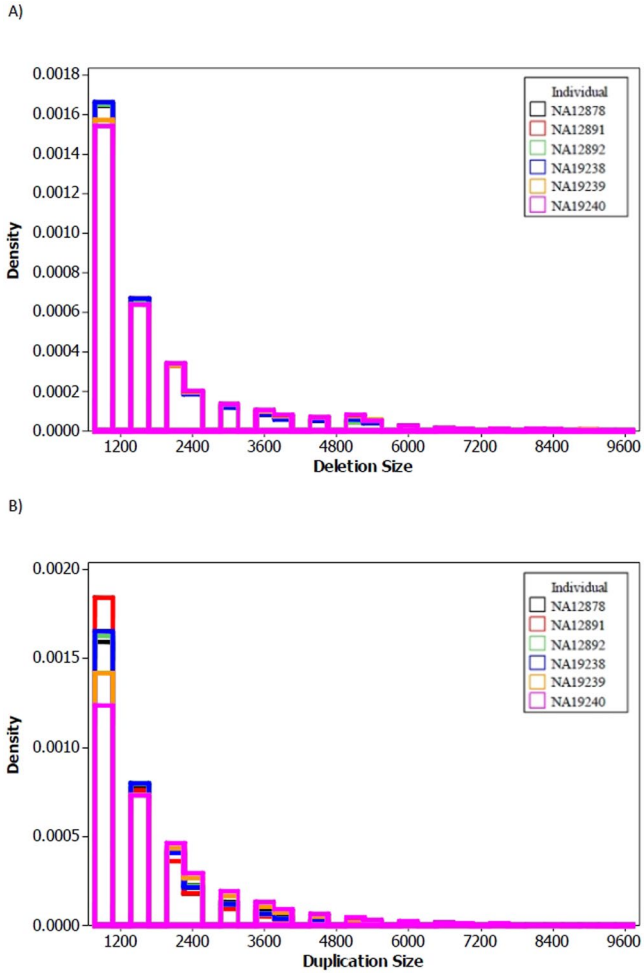

Figure 3. The size distribution of the deletion and duplication calls – made by MSeq-CNV - in six individuals with high-coverage sequencing from 1000Genomes Project. (**A**) Deletion size distributions, (**B**) Duplication size distributions.

LP6005592-DNA_D01 (from FINLAND) are the next, respectively with a total number of 183, 439 (317.397 Mega bp), 197, 435 (323.638 Mega bp), 185, 837 (323.829 Mega bp) calls.

The highest number of CNVs are detected in the South Asian individuals LP6005519-DNA_A04 and LP6005519-DNA_A05 (from INDIA) respectively with a total number of 262, 405 (431.409 Mega bp) and 341, 909 (545.435 Mega bp) calls. Extensive CNVs in Indian individuals, which is as many as YRI trio, were also previously reported in the admixed Indian population of African ancestry[47,50], to adopt with environmental conditions.

Details of deletion and duplication calls are given in Table S.4 and Table S.5. The overall number of detected CNVs in each chromosome is shown in Fig. 2C, for each individual.

Detected CNVs in LP6005592-DNA_H03, LP6005442-DNA_E07, LP6005443-DNA_G05 LP6005519-DNA_A04 and LP6005519-DNA_A05, and LP6005592-DNA_D01 overlap with DGV respectively with a ratio of 0.61, 0.62, 0.61, 0.60, 0.62, and 0.62, for the number of calls and 0.60, 0.60, 0.59, 0.60, 0.59, and 0.61, for the base pairs.

Applications and Limitations. The MSeq-CNV can be applied for detecting the recurrent genome-wide CNVs from NGS data in the diploid genome of human and other organisms, as well. However, the current version of MSeq-CNV is not capable of detecting CNVs in the sequencing data of a haploid genome. The input NGS data for the MSeq-CNV are possibly the mate pair reads which are collected from sequencing with multiple platforms, multiple individuals and experimental conditions.

Although the current version of the MSeq-CNV is limited to the whole genome shotgun sequencing, further work is in progress to adopt MSeq-CNV with the exome or gene panel sequencing data.

Also, as mentioned above, the other attractive feature of the MSeq-CNV is in constructing the ancestry of the sequenced individuals, based on the detected CNV matrix.

Discussion

In this article we proposed MSeq-CNV as a new tool for detecting genome-wide deletions and duplications from sequencing of multiple samples. Simultaneous analysis of multiple samples allows detecting common CNVs

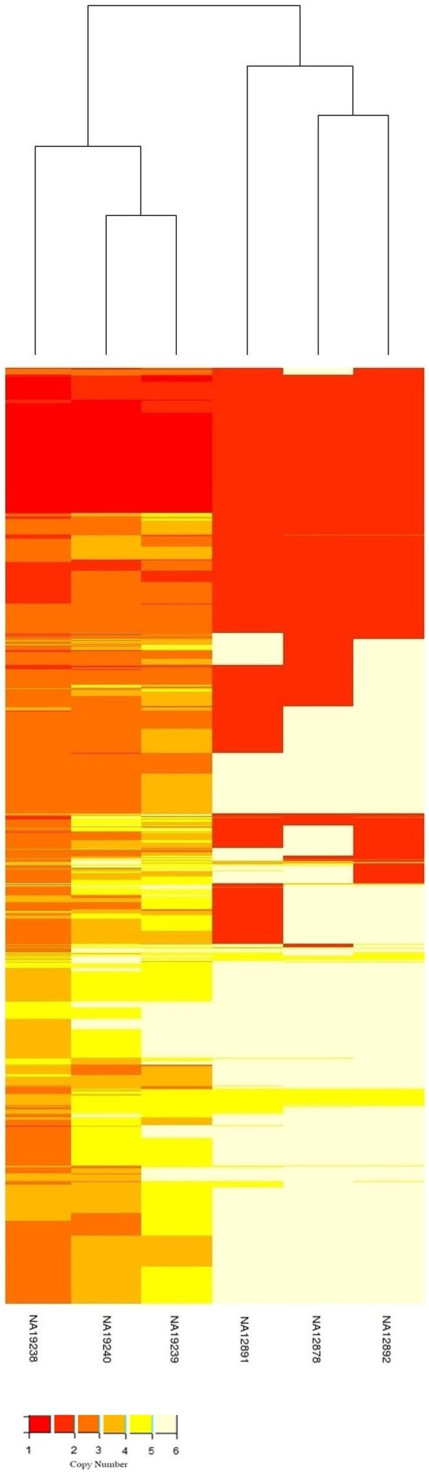

Figure 4. The hierarchical clustering algorithm is applied to the matrix of CNV regions which are detected using MSeq-CNV in the genome of HapMap individuals NA12891, NA12892, NA12878, NA19238, NA19239 and NA19240. Based on the CNV information, the hierarchical clustering algorithm has correctly segregated the ancestry of the six individuals in two groups i.e. a group includes CEU trio and a group includes YRI trio.

which are shared by complex diseases. Also, read count variations which occur due to the sequencing noise can be detected by the analysis of several samples together. MSeq-CNV applies a novel probabilistic framework for modeling the read depth and insertion size signals, together.

The overall performance of MSeq-CNV has been superior to the central CNV detection tools such as rSW-seq, CNV-seq and cnMOPS. Specially, for a coverage of 1× which is fairly low, the overall performance MSeq-CNV has been considerably higher than the compared tools. Reaching a high performance in low coverage data is an

advantage of MSeq-CNV. In future, CNV detection tools which rely on a low-coverage sequencing are more relevant[37,38]. Indeed, a low coverage sequencing is common in many individuals e.g. in the 1000 Genomes Project the shotgun sequencing of 179 individuals is carried out with a coverage of 2× to 4×[41].

The MSeq-CNV works with the empirical distribution of the insertion sizes in clone library. Therefore, MSeq-CNV is robust to deviations from the theoretical insertion size distribution which occurs due to several artifacts, attributed to the library-preparation protocols.

References

1. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**, 437–455, https://doi.org/10.1146/annurev-med-100708-204735 (2010).
2. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855, https://doi.org/10.1038/nature04489 (2006).
3. Albertson, D. G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* **12**(Spec No 2), R145–152, https://doi.org/10.1093/hmg/ddg261 (2003).
4. Cook, E. H. Jr. & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923, https://doi.org/10.1038/nature07458 (2008).
5. Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. & Jain, A. N. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132–153, https://doi.org/10.1016/j.jmva.2004.02.008 (2004).
6. Marioni, J. C., Thorne, N. P. & Tavare, S. BioHMM: A heterogeneous Hidden Markov model for segmenting array CGH data. *Bioinformatics (Oxford, England)* **22**, https://doi.org/10.1093/bioinformatics/btl089 (2006).
7. Shah, S. P., Lam, W. L., Ng, R. T. & Murphy, K. P. Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics (Oxford, England)* **23**, i450–458, https://doi.org/10.1093/bioinformatics/btm221 (2007).
8. Ding, J. & Shah, S. A robust hidden semi-Markov model with application to aCGH data processing. *Int J Data Min Bioinform* **8**, 427–442 (2013).
9. Zhang, Q. *et al.* CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics (Oxford, England)* **26**, 464–469, https://doi.org/10.1093/bioinformatics/btp708 (2010).
10. Park, C., Ahn, J., Yoon, Y. & Park, S. A Multi-Sample Based Method for Identifying Common CNVs in Normal Human Genomic Structure Using High-Resolution aCGH Data. *PLoS ONE* **6**, e26975, https://doi.org/10.1371/journal.pone.0026975 (2011).
11. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166–1174, http://www.nature.com/ng/journal/v40/n10/suppinfo/ng.238_S1.html (2008).
12. Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E. & Nickerson, D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* **40**, 1199–1203, https://doi.org/10.1038/ng.236 (2008).
13. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125–138, https://doi.org/10.1038/nrg3373 (2013).
14. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughtput sequencing. *BMC Bioinformatics* **10**, https://doi.org/10.1186/1471-2105-10-80 (2009).
15. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**, S1, https://doi.org/10.1186/1471-2105-14-s11-s1 (2013).
16. Kim, T. M., Luquette, L. J., Xi, R. & Park, P. J. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* **11**, 432, https://doi.org/10.1186/1471-2105-11-432 (2010).
17. Wang, H., Nettleton, D. & Ying, K. Copy number variation detection using next generation sequencing read counts. *BMC Bioinformatics* **15**, 1–14, https://doi.org/10.1186/1471-2105-15-109 (2014).
18. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci USA* **108**, E1128–1136, https://doi.org/10.1073/pnas.1110574108 (2011).
19. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research* **19**, 1586–1592, https://doi.org/10.1101/gr.092981.109 (2009).
20. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, https://doi.org/10.1038/nmeth.1276 (2009).
21. McCallum, K. J. & Wang, J. P. Quantifying copy number variations using a hidden Markov model with inhomogeneous emission distributions. *Biostatistics* **14**, 600–611, https://doi.org/10.1093/biostatistics/kxt003 (2013).
22. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6**, e16327, https://doi.org/10.1371/journal.pone.0016327 (2011).
23. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677–681, https://doi.org/10.1038/nmeth.1363 (2009).
24. Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics (Oxford, England)* **27**, 595–603, https://doi.org/10.1093/bioinformatics/btq713 (2011).
25. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)* **28**, i333–i339, https://doi.org/10.1093/bioinformatics/bts378 (2012).
26. Yavas, G., Koyuturk, M., Gould, M. P., McMahon, S. & LaFramboise, T. DB2: a probabilistic approach for accurate detection of tandem duplication breakpoints using paired-end reads. *BMC Genomics* **15**, 175, https://doi.org/10.1186/1471-2164-15-175 (2014).
27. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, https://doi.org/10.1186/gb-2014-15-6-r84 (2014).
28. Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10**, R23, https://doi.org/10.1186/gb-2009-10-2-r23 (2009).
29. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)* **25**, 2865–2871, https://doi.org/10.1093/bioinformatics/btp394 (2009).
30. Abel, H. J. *et al.* SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics (Oxford, England)* **26**, 2684–2688, https://doi.org/10.1093/bioinformatics/btq528 (2010).
31. Sindi, S. S., Onal, S., Peng, L. C., Wu, H. T. & Raphael, B. J. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol* **13**, R22, https://doi.org/10.1186/gb-2012-13-3-r22 (2012).
32. Zhang, Z. D. *et al.* Identification of genomic indels and structural variations using split reads. *BMC Genomics* **12**, 375, https://doi.org/10.1186/1471-2164-12-375 (2011).
33. Sindi, S., Helman, E., Bashir, A. & Raphael, B. J. A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)* **25**, i222–230, https://doi.org/10.1093/bioinformatics/btp208 (2009).
34. Malekpour, S. A., Pezeshk, H. & Sadeghi, M. MGP-HMM: Detecting genome-wide CNVs using an HMM for modeling mate pair insertion sizes and read counts. *Mathematical biosciences* **279**, 53–62, https://doi.org/10.1016/j.mbs.2016.07.006 (2016).

35. Ratan, A. *et al.* Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample. *PLoS ONE* **8**, e55089, https://doi.org/10.1371/journal.pone.0055089 (2013).
36. Moreno-De-Luca, D. *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *American journal of human genetics* **87**, 618–630, https://doi.org/10.1016/j.ajhg.2010.10.004 (2010).
37. Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research* **40**, e69, https://doi.org/10.1093/nar/gks003 (2012).
38. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research* **21**, 952–960, https://doi.org/10.1101/gr.113084.110 (2011).
39. The Genomes Project, C. An integrated map of genetic variation from 1, 092 human genomes. **491**, 56, https://doi.org/10.1038/nature11632, https://www.nature.com/articles/nature11632#supplementary-information (2012).
40. The Genomes Project, C. A global reference for human genetic variation. **526**, 68, https://doi.org/10.1038/nature15393 https://www.nature.com/articles/nature15393#supplementary-information (2015).
41. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073, http://www.nature.com/nature/journal/v467/n7319/abs/10.1038-nature09534-unlocked.html#supplementary-information (2010).
42. Duan, J., Deng, H. W. & Wang, Y. P. Common copy number variation detection from multiple sequenced samples. *IEEE transactions on bio-medical engineering* **61**, 928–937, https://doi.org/10.1109/tbme.2013.2292588 (2014).
43. Magi, A., Benelli, M., Yoon, S., Roviello, F. & Torricelli, F. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic acids research* **39**, https://doi.org/10.1093/nar/gkr068 (2011).
44. Sudmant, P. H. *et al.* An integrated map of structural variation in 2, 504 human genomes. *Nature* **526**, 75, https://doi.org/10.1038/nature15394 https://www.nature.com/articles/nature15394#supplementary-information (2015).
45. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454, https://doi.org/10.1038/nature05329 (2006).
46. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science (New York, N.Y.)* **349**, aab3761, https://doi.org/10.1126/science.aab3761 (2015).
47. Veerappa, A. M. *et al.* Global Spectrum of Copy Number Variations Reveals Genome Organizational Plasticity and Proposes New Migration Routes. *PLOS ONE* **10**, e0121846, https://doi.org/10.1371/journal.pone.0121846 (2015).
48. MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**, D986–992, https://doi.org/10.1093/nar/gkt958 (2014).
49. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. **538**, 201, https://doi.org/10.1038/nature18964 https://www.nature.com/articles/nature18964#supplementary-information (2016).
50. Narang, A. *et al.* Extensive copy number variations in admixed Indian population of African ancestry: potential involvement in adaptation. *Genome biology and evolution* **6**, 3171–3181, https://doi.org/10.1093/gbe/evu250 (2014).

Acknowledgements

Hamid Pezeshk and Seyed Amir Malekpour would like to thank department of research affairs at University of Tehran. Hamid Pezeshk is also grateful to School of Biological Sciences at IPM for their supports. Some parts of this study were completed when he was visiting the Department of Mathematics and Statistics of Concordia University during a sabbatical leave. The authors would also like to thank the excellent comments and suggestions of two anonymous referees. The financial support of INSF (No. 95834244) is gratefully acknowledged.

Author Contributions

The data analysis and calculations are done by S.A.M., H.P. and M.S. were involved in the scientific discussions. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-22323-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.