MGP-HMM: Detecting genome-wide CNVs using an HMM for modeling mate pair insertion sizes and read counts

Seyed Amir Malekpour, Hamid Pezeshk, Mehdi Sadeghi

1. Introduction

Copy Number Variation (CNV) is envisaged to be a major source of the genetic variations and aberrations in the genome of human and other organisms. In CNV, number of copies of a gene or a segment of the genome differs from one person to the next. Common types of CNVs include segmental insertions and deletions (indels), and duplications.

Since CNV's introduction, its functionalities and consequences are investigated in many research articles. Association of the CNV with diseases such as diabetes [1], schizophrenia [2–5], autism [6–9] and developmental disabilities [10] is already confirmed. Moreover, it is associated with fatal diseases such as breast [11] and lung [12] cancer.

Motivation: Association of Copy Number Variation (CNV) with schizophrenia, autism, developmental disabilities and fatal diseases such as cancer is verified. Recent developments in Next Generation Sequencing (NGS) have facilitated the CNV studies. However, many of the current CNV detection tools are not capable of discriminating tandem duplication from non-tandem duplications.

Results: In this study, we propose MGP-HMM as a tool which besides detecting genome-wide deletions discriminates tandem duplications from non-tandem duplications. MGP-HMM takes mate pair abnormalities into account and predicts the digitized number of tandem or non-tandem copies. Abnormalities in the mate pair directions and insertion sizes, after being mapped to the reference genome, are elucidated using a Hidden Markov Model (HMM). For this purpose, a Mixture Gaussian density with time-dependent parameters is applied for emitting mate pair insertion sizes from HMM states.

Indeed, depending on observed abnormalities in mate pair insertion size or its orientation, each component in the mixture density will have different parameters. MGP-HMM also applies a Poisson distribution for modeling read depth data. This parametric modeling of the mate pair reads enables us to estimate the length of CNVs precisely, which is an advantage over methods which rely only on read depth approach for the CNV detection. Hidden state of the proposed HMM is the digitized copy number of a genomic segment and states correspond to the multipliers of the mixture Gaussian components. The accuracy of our model is validated on a set of next generation sequencing real and simulated data and is compared to other tools.

© 2016 Elsevier Inc. All rights reserved.
inhomogeneous transition probabilities is used to model aCGH data. While these tools are proposed for detecting CNVs based on one sample genome, [19] applies an HMM for detecting CNVs from several sample genomes.

However in recent years, development of NGS has provided an unprecedented opportunity for genome-wide variation studies [20–21], with more details and lower costs. Most studies that rely on NGS benefit of modeling read counts [22–34]. CNV-seq [22] and BIC-seq [23] compare one sample genome with the reference genome for the CNV detection. In [24] a parametric statistical method, and in [25] an HMM with inhomogeneous emission probabilities are applied for modeling read counts and detecting CNVs. On the other hand, CMDS [26], cn.MOPS [27], rSW-seq [28], and CNaseg [29] take several sample individuals into account and predict CNVs based on information in all samples.

Other approaches utilize NGS mate pair and split reads [35–47] which have more capabilities in detecting genome-wide CNVs and genomic rearrangements, compared to the tools that rely on read depth approach. Therefore, more methods apply mate reads in recent years. Peculiarly, an attractive feature of mate reads is ability in predicting genome-wide tandem duplications [48].

In this paper, we propose MGP-HMM as a tool for detecting genome-wide deletions and duplications which discriminates tandem and non-tandem duplications, from each other. MGP-HMM applies a mixture Gaussian density for modeling variations in the mate pair’s orientation and insertion size, after being mapped to the reference genome. The parameters of Gaussian densities model the length of deleted/duplicated genomic region and therefore they depend on genomic loci. This position-specific parametric modeling allows us to estimate the length of genome-wide CNVs accurately and this is an advantage over methods that rely only on modeling read counts. Besides modeling mate pairs abnormalities, in MGP-HMM read counts are modeled using a Poisson distribution.

2. Methods

We assume that the reference genome is divided into $T$ segments of length $L$ and mate pairs are mapped to the reference genome. In $t$th genomic segment, observations are composed of all those mate pairs whose reads are flanking the segment and their un-sequenced (insertion) regions are spanning the segment. In the $t$th genomic segment with read count $c_t$ and $n_t$ mate pairs satisfying above condition, observation vector is shown by $o_t = \{\delta_{t,1}, \delta_{t,2}, \ldots, \delta_{t,n_t}, c_t\}$, in which information relevant to mate pair insertion size and orientation are indicated by $\delta_{t,v}$, where $v$ represents the mate pair index. For genomic segments $1$ to $T$, observation vector is consequently shown by $O = \{o_1, o_2, \ldots, o_T\}$.

Based on above observation vector, we aim at predicting the hidden state of each genomic segment that can be of type: (non-tandem duplication, tandem duplication, homozygous deletion, heterozygous deletion, and diploidy). Each of these CNV types is considered as a different state in Hidden Markov Model (HMM), see Fig. 1. Corresponding to each possible copy number of a genomic segment, a different state is considered for HMM and the maximum number of tandem or non-tandem copies is $v$.

Transition and emission probabilities are two major components of HMMs. Transition probability is the probability of moving from one state to another in a single step. As shown in Fig. 1, from diploidy state we can reach any other state i.e. non-tandem duplication, tandem duplication, homozygous deletion, heterozygous deletion and it’s possible to get back to the diploidy state, from other states.

Also, emission probabilities define the probability of emitting observations from HMM states. Before defining emission probabilities in Section 2.2, different types of aberrations in the mate pair reads are discussed in Section 2.1.

2.1. Aberrations in mate pair reads

After mapping a mate pair to the reference genome, aberrations may occur in insertion size or orientation of the mate pair. In the human diploid genome, each genomic segment has two copies which each of them are located on a separate homologous chromosome. All mate pairs pertaining to this state, map to the reference genome with a standard insertion size and direction. Since sequencing platform generates mate pairs with normally distributed insertion sizes with mean $\mu$ and variance $\sigma^2$, i.e. $N(\mu, \sigma^2)$, insertion size of mate pairs which pertain to the genomic diploid state are distributed with $N(\mu, \sigma^2)$. Mate pairs from the genomic non-tandem duplication states will also map to the reference genome with no aberrations, and then have insertion sizes distributed with $N(\mu, \sigma^2)$.

Genomic segments with heterozygous deletion state have only one allele on the sample genome. Therefore, approximately half of the mate pairs after mapping to the reference genome will have insertion sizes distributed with $N(\mu, \sigma^2)$ and the other half of the mate reads that correspond to the deleted allele will map to the reference genome much further apart than expected, with a $N(\mu + \text{deletion size}, \sigma^2)$ insertion size distribution.

In genomic segments with homozygous deletion state, both alleles are deleted from the sample genome. Therefore, all mate pairs corresponding to the deleted alleles will map to the reference genome with a $N(\mu + \text{deletion size}, \sigma^2)$ distribution.

Tandem duplication state models genomic segments with more than two copies on the sample genome, at least two copies located next to each other on a homologous chromosome, see Fig. 2. For a tandem duplication of length $L$, mate pairs that are spanning the tandem duplication breakpoints will map to the reference genome with an “everted” orientation and insertion sizes distributed with $N(X - \mu - 2\times(\text{read length}), \sigma^2)$. However, mate pairs that do not span the breakpoint, map to the reference genome with a standard direction and insertion size. In Table 1, above analysis on the distribution of mate pair insertion size is summarized.

It should be added that according to [49–50], a perfect distribution for the insertion size is Gaussian with a peak at the expected position. In [49], distribution of insertion size is studied for different illumina sequencing configurations and the optimal conditions are proposed to achieve a normally distributed insertion size. Although in some cases distribution gets a little bit skewed or fat-tailed (due to the several artifacts that can be attributed to the library-preparation protocol), a normal distribution is still a reasonable assumption.

Indeed, Gaussian distributions have been frequently used in previous studies for modeling insertion sizes. In [51], a Gaussian distribution is applied for modeling insertion sizes in genomic regions with deletions. Also in [52], to detect short tandem repeats, insertion size is assumed to be normally distributed.

2.2. Emission probability matrix of HMM

As discussed before, observations in the $t$th genomic segment are denoted by $o_t = \{\delta_{t,1}, \delta_{t,2}, \ldots, \delta_{t,n_t}, c_t\}$, and hidden state corresponding to this segment is shown by $q_t$, where $1 \leq t \leq T$ and $q_t$ is in {homozygous deletion, heterozygous deletion, diploidy, tandem duplication, non-tandem duplication}. Using Table 1, the probability of emitting a mate pair insertion size in state $q_t$ follows a mixture Gaussian density which is:

\[
\begin{align*}
    f(\delta_{t,k} | q_t) = \frac{3}{z=1} \sum \alpha_{q_{t,z}} f_z(\delta_{t,k} | q_t), & \text{ with } 0 \leq \alpha_{q_{t,z}} \leq 1, \sum_{z=1}^{3} \alpha_{q_{t,z}} = 1
\end{align*}
\]
In above equations, \( \alpha_{q_i, z} \) is the probability of \( \delta_{i, k} \) coming from \( z \)th component of the mixture Gaussian density, for \( 1 \leq z \leq 3 \) and \( 1 \leq k \leq n_t \). Each of these densities corresponds to one density in Table 1. Also, \( f_z(\delta_{t, k}|q_t) \) has the following form:

\[
f_z(\delta_{t, k}|q_t) = \frac{1}{\sqrt{2\pi}\sigma^2_{iz}}\exp\left\{-\frac{1}{2\sigma^2_{iz}}(\delta_{t, k} - \mu_{iz})^2\right\},
\]

where, \( \mu_{iz} \) and \( \sigma^2_{iz} \) are the mean and variance of the \( z \)th component of the mixture Gaussian density. \( f_1(\delta_{t, k}|q_t) \) models the insertion size emission probability in mate pairs with a clone library insertion size distribution which are mapped to the reference genome with standard direction. \( f_2(\delta_{t, k}|q_t) \) models insertion size emission probability in mate pairs that are mapped to the reference genome much further apart, compared to the clone library insertion sizes, and have a standard orientation. As indicated in Table 1, insertion size for these mate pairs is \( \mu_{iz} + \text{deletion size} \). \( f_3(\delta_{t, k}|q_t) \) is used for emitting insertion sizes in mate pairs that are mapped to the reference genome with aberrations in orientation and insertion sizes, compared to the clone library insertion sizes.

Also, number of read counts in the \( r \)th genomic segment is indicated by \( c_r \). As shown in previous studies [27], Poisson distribution is a reasonable choice for emitting read counts. The number of reads which are mapped to the \( r \)th genomic segment depends on its hidden state \( q_t \) i.e. 0, 1, 2, 3, ... which correspond to \{homozygous deletion, heterozygous deletion, diploid, tandem duplication, and non-tandem duplication\} states, respectively. Then, we use the following distribution for modeling \( c_r \):

\[
P(c_{r}|q_t) = e^{-\lambda_{q_t}} \frac{\lambda_{q_t}^{c_r}}{c_r!}
\]

In which \( \lambda_{q_t} \) is the average number of reads that are mapped to the genomic segments with diploid state and it is estimated based on genome-wide data.

### 2.3. Parameter estimation

As mentioned above, transition and emission probabilities are two components of an HMM which are estimated via Expectation-Maximization (EM) algorithm. For this purpose, we correspond a vector of variables \( (m_{q_1, 1}, m_{q_2, 2}, \ldots, m_{q_t, n_t}) \) to each observation vector \( q_t = \{\delta_{t, 1}, \delta_{t, 2}, \ldots, \delta_{t, n_t}, c_t\} \), in which \( m_{q_t,k} \) indicates the component index of the mixture Gaussian density that \( k \)th observation is emitted from, \( 1 \leq k \leq n_t \) and \( m_{q_t,k} = 1, 2, 3 \). Further, the following Q function is defined:

\[
Q(\theta, \theta^\delta) = \sum_q \sum_m \log(P(O, q, m|\theta)) \log(P(O, q, m|\theta^\delta))
\]

In which \( m = \{m_{q_1, 1}, m_{q_2, 2}, \ldots, m_{q_1, n_1}, m_{q_2, n_2}, \ldots, m_{q_2, 1}, \ldots, m_{q_t, n_t}\} \), and \( \theta \) is the set of all parameters whose estimates in
the $g^{th}$ iteration of the EM algorithm is denoted by $\theta^g$, see [53] for more details. In above equation, we sum over all possible values of $m$ and $q = \{q_1, q_2, \ldots, q_T\}$. If $b_{0i}(\theta)$ denotes the density of emitting $q_i$ in state $q_t$, the joint density of $O = \{o_1, o_2, \ldots, o_T\}$ and $q = \{q_1, q_2, \ldots, q_T\}$ is:

\[
P(O, q|\theta) = \sum_{t=1}^{T} \prod_{i=q_{t-1}}^{q_t} b_{0i}(\theta).
\]

In which $\pi_{0q}$ is the probability of starting a sequence of observations $\{o_1, o_2, \ldots, o_T\}$ in state $q_0$, and $a_{0i}q_1 = 1$ is the probability of moving from state $q_{t-1}$ to $q_t$ at time $t$ and $a_{0q_1} = 1$. Therefore,

\[
Q(\theta, \theta^g) = \sum_q \sum_m \log(\pi_{0q})P(O, q, m|\theta^g) + \sum_q \sum_m \sum_t \log(a_{q_{t-1}q_t})P(O, q, m|\theta^g) + \sum_q \sum_m \sum_t \log(b_{0i}(\theta))P(O, q, m|\theta^g)
\]

The first term in the $Q(\theta, \theta^g)$ can be denoted by:

\[
\sum_q \sum_m \sum_t \log(\pi_{0q})P(O, q, m|\theta^g) = \sum_q \log(\pi_{0q})P(O, q_0 = i|\theta^g).
\]

In which, possible number of HMM states is denoted by $S$. Then using the Lagrange multiplier $\gamma$ to introduce the constraint $\sum_{i=1}^{T} \pi_i = 1$, we have

\[
\partial / \partial \pi_i \left( \sum_{i=1}^{S} \log(\pi_i)P(O, q_0 = i|\theta^g) + \gamma \left( \sum_{i=1}^{S} \pi_i - 1 \right) \right) = 0
\]

The following formula for updating $\pi_i$, in each iteration of the EM algorithm is obtained:

\[
\pi_i^{g+1} = \frac{P(O, q_0 = i|\theta^g)}{P(O|\theta^g)}.
\]

For the second term in $Q(\theta, \theta^g)$ we have:

\[
\sum_q \sum_m \sum_t \log(a_{q_{t-1}q_t})P(O, q, m|\theta^g) = \sum_{t=1}^{T} \sum_{j=1}^{S} \sum_{i=1}^{S} \log(a_{ij})P(O, q_{t-1} = i, q_t = j|\theta^g).
\]

Again, after introducing a Lagrange multiplier for $\sum_{j=1}^{S} a_{ij} = 1$, and taking derivative with respect to $a_{ij}$ we obtain:

\[
a_{ij}^{g+1} = \frac{\sum_{t=1}^{T} P(O, q_{t-1} = i, q_t = j|\theta^g)}{\sum_{t=1}^{T} P(O, q_{t-1} = i|\theta^g)}
\]

Also, for the third term in $Q(\theta, \theta^g)$:

\[
\sum_q \sum_m \sum_t \log(b_{0i}(\theta))P(O, q, m|\theta^g) = \sum_q \sum_m \sum_t \log \left( \prod_{k=1}^{n_i} b_{0i}(\delta_{t,k}, m_{q_k}) \right)P(O, q, m|\theta^g) + \sum_q \sum_m \sum_t \log \left( P \left( \frac{\lambda_0}{2} \right) \right)P(O, q, m|\theta^g) + \sum_q \sum_m \sum_t \log b_{0i}(\delta_{t,k}, m_{q_k})P(O, q, m|\theta^g)
\]
<table>
<thead>
<tr>
<th>Real state</th>
<th>Heterozygous deletion</th>
<th>Diploid</th>
<th>Homozygous deletion</th>
<th>3 non-tandem copies</th>
<th>4 non-tandem copies</th>
<th>5 non-tandem copies</th>
<th>6 ≤ non-tandem copies</th>
<th>3 tandem copies</th>
<th>4 tandem copies</th>
<th>5 tandem copies</th>
<th>6 ≤ tandem copies</th>
<th>Sum</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted state</td>
<td>Heterozygous deletion</td>
<td>1812</td>
<td>1409</td>
<td>33</td>
<td>42</td>
<td>1</td>
<td>16</td>
<td>14</td>
<td>19</td>
<td>2</td>
<td>16</td>
<td>15</td>
<td>3379</td>
<td>0.54</td>
</tr>
<tr>
<td>Diplod</td>
<td>79</td>
<td>15,697</td>
<td>28</td>
<td>149</td>
<td>107</td>
<td>88</td>
<td>83</td>
<td>13</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>16,263</td>
<td>0.97</td>
<td>0.88</td>
</tr>
<tr>
<td>Homozygous deletion</td>
<td>33</td>
<td>274</td>
<td>1904</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2229</td>
<td>0.85</td>
<td>0.97</td>
</tr>
<tr>
<td>3 non-tandem copies</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>528</td>
<td>47</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>603</td>
<td>0.88</td>
<td>0.70</td>
</tr>
<tr>
<td>4 non-tandem copies</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>1001</td>
<td>90</td>
<td>8</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1129</td>
<td>0.89</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>5 non-tandem copies</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>644</td>
<td>120</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>789</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>6 ≤ non-tandem copies</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>584</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>592</td>
<td>0.99</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>3 tandem copies</td>
<td>4</td>
<td>249</td>
<td>0</td>
<td>31</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>973</td>
<td>300</td>
<td>2</td>
<td>56</td>
<td>1632</td>
<td>0.60</td>
</tr>
<tr>
<td>4 tandem copies</td>
<td>0</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>889</td>
<td>544</td>
<td>70</td>
<td>1574</td>
<td>0.56</td>
<td>0.74</td>
</tr>
<tr>
<td>5 tandem copies</td>
<td>0</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>662</td>
<td>467</td>
<td>1195</td>
<td>0.55</td>
</tr>
<tr>
<td>6 ≤ tandem copies</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>529</td>
<td>615</td>
<td>865</td>
<td>0.86</td>
</tr>
<tr>
<td>Sum</td>
<td>1928</td>
<td>17,831</td>
<td>1965</td>
<td>750</td>
<td>1173</td>
<td>845</td>
<td>912</td>
<td>1005</td>
<td>1208</td>
<td>1232</td>
<td>1151</td>
<td>30,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2

Precision and recall values of MGP-HMM are evaluated for the constructed diploid genome with implanted CNVs, using 10× depth of coverage. For 30,000 genomic segments (45 million bp), predicted state is shown versus the real state of the genomic segments, and number of segments is indicated in the corresponding cell.

By taking derivative of the above equation with respect to $\mu_2$, we have:

$$
\lambda \left( \sum_{k=1}^{S} \log \left( \frac{b_i(t_k)}{b_i(t_{k+1})} \right) \right) - \mu_2 \left( \sum_{k=1}^{S} \frac{1}{a_k} \right) = 0
$$

In the above equation, $\lambda$ is the prior probability of the state $i$ being in state $i$ and is calculated via Forward-Backward (FB) algorithm. $a_k$ is the joint probability of the $k$th state pair coming from mixture component $2$. To estimate $\mu_2$ and $q_2$, for $i, j, k = 1, 2, \ldots, S$ and setting it equal to zero we get:

$$
\sum_{k=1}^{S} \lambda \left( \sum_{i=1}^{S} \log \left( \frac{b_i(t_k)}{b_i(t_{k+1})} \right) \right) - \mu_2 \left( \sum_{i=1}^{S} \frac{1}{a_k} \right) = 0
$$

Since $\sum_{k=1}^{S} \lambda = 1$, by introducing $\lambda$ as the Lagrange multiplier, we get:

$$
\sum_{i=1}^{S} \log \left( \frac{P(O_i | q_i \cdot m_i)}{P(O_i | q_i \cdot m_i)} \right) + \lambda \left( \sum_{i=1}^{S} \log \left( P(O_i | q_i \cdot m_i) \right) \right) = 0
$$

where:

$$
\sum_{i=1}^{S} \log \left( \frac{P(O_i | q_i \cdot m_i)}{P(O_i | q_i \cdot m_i)} \right)
$$

is the prior probability of the state $i$ being in state $i$ and is calculated via Forward-Backward (FB) algorithm. $a_k$ is the joint probability of the $k$th state pair coming from mixture component $2$.
The Gaussian density functions are fitted to both histograms.

\[
\mu_{iz}^{(1)} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{n_z} \delta_{t,k} P(O, q_t = i, m_{q,k} = z|\theta)}{\sum_{t=1}^{T} \sum_{k=1}^{n_z} P(O, q_t = i, m_{q,k} = z|\theta)}.
\]

Using the same arguments to estimate \(\sigma_{iz}\):

\[
\sigma_{iz}^{(1)} = \left[ \frac{\sum_{t=1}^{T} \sum_{k=1}^{n_z} \left( \delta_{t,k} - \mu_{iz} \right)^2 P(O, q_t = i, m_{q,k} = z|\theta) \sigma_{iz}^{(0)}}{\sum_{t=1}^{T} \sum_{k=1}^{n_z} P(O, q_t = i, m_{q,k} = z|\theta)} \right]^{0.5}.
\]

Further, to estimate \(\lambda_0\) which corresponds to average read counts in genomic segments with diploid state we have:

\[
\sum_{q} \sum_{m} \sum_{t=1}^{T} \log \left( \frac{P\left( \delta_{t,k} \right)}{P(O, q, m|\theta)} \right) P(O, q_t = i|\theta) + \sum_{q} \sum_{m} \sum_{t=1}^{T} \log \left( \frac{P\left( \delta_{t,k} \right)}{P(O, q, m|\theta)} \right) P(O, q_t = i|\theta)
\]

\[
= \sum_{i=1}^{S} \sum_{t=1}^{T} \log \left( \frac{e^{-\frac{1}{2} \left( \delta_{t,k} - \mu_{iz} \right)^2}}{c_1 !} \right) P(O, q_t = i|\theta) + \sum_{i=1}^{S} \sum_{t=1}^{T} \left[ -\frac{\lambda_0}{2} + c_1 \log(2\pi) + \log\left( i \right) - \log(c_1 !) \right] P(O, q_t = i|\theta)
\]

After taking the derivative of the above equation with respect to \(\lambda_0\) we reach:

\[
\sum_{i=1}^{S} \sum_{t=1}^{T} \left[ -\frac{i \lambda_0}{2} + c_1 \right] P(O, q_t = i|\theta) = 0
\]

Then, the following formula is used to re-estimate the value of \(\lambda_0\), in each iteration of the EM algorithm:

\[
\lambda_0^{(1)} = \frac{\sum_{i=1}^{S} \sum_{t=1}^{T} c_1 P(O, q_t = i|\theta)}{\sum_{i=1}^{S} \sum_{t=1}^{T} \frac{1}{2} P(O, q_t = i|\theta)}.
\]

3. Results from simulated data

In this section, performance of MGP-HMM is tested on a simulated human genome with implanted CNVs. To construct a human diploid genome, the forward strand of chromosome 3 of the human reference genome is duplicated. This diploid genome was then altered with CNVs of type heterozygous deletion, homozygous deletion, tandem duplications and non-tandem duplications.

Then, MAQ [54] is used for generating mate pairs from whole-genome shotgun sequencing of the constructed sample genome with implanted CNVs. MAQ was calibrated to generate mate pairs with normally distributed insertion sizes i.e. N(170, 20^2), and each read having a length of 100 bp. The simulated mate pairs were then mapped to the reference genome.

After partitioning the reference genome into segments of length e.g. 150 bp, observation vector in \(r\)th genomic segment is defined with each mate pair whose reads are mapped to the flanking regions of the \(r\)th genomic segment and its insertion size is spanning the segment. See Section 2 for more details.

In Table 2, precision and recall values of MGP-HMM in detecting CNVs in the constructed sample genome is calculated for each CNV state, using 10x depth of coverage. As shown, precision and recall values are high and very close to the theoretical prediction power of the computational methods [45].

A major advantage of MGP-HMM over methods that rely only on a read depth approach for the CNV detection, is in estimating the length of the duplicated or deleted regions. In MGP-HMM, this length is estimated by measuring the deviations of the mate pair insertion sizes from the clone library insertion size distribution. Indeed, insert size modeled with parameter \(\mu_{iz}\) of the mixture Gaussian density, \(1 \leq t \leq T\) and \(1 \leq z \leq 3\). Therefore, for heterozygous and homozygous deletion states, the average deletion size is estimated to be \(\mu_{iz} - \mu\), which indicates the average increase in the mate pairs insertion sizes, after mapping to the reference genome. For a genomic region with copy gain, the length of the duplicated region is calculated by \(\mu_{iz} + 2 \times \text{(read length)} + \mu\).

In Fig. 3, the histograms of real minus predicted length of CNVs are shown for regions with copy loss and copy gain, separately. Fig. 3a shows this histogram for 180 genomic regions with copy loss whose states were correctly predicted by the model. Other 25 regions with copy loss whose states were not correctly predicted, are excluded. Fig. 3b shows the histogram of the real minus predicted length of 230 genomic CNV regions with copy gain whose
states were correctly predicted by the model. 7 CNV regions with copy gain whose states were not correctly predicted are excluded from this histogram.

As shown in Fig. 3, histograms are normally distributed with a mean close to zero and in all cases the difference of estimated and real length of CNVs are a few base pairs. This indicates that MGP-HMM has been accurate in detecting the length of implanted CNVs which were at least 1 kb.

MGP-HMM is compared to central CNV detection methods i.e. CNV-seq, Pindel and Delly. Among these tools, CNV-seq relies on read depth approach for the CNV detection and does not discriminate tandem duplications from non-tandem duplications. However, Pindel and Delly apply mate pair reads which enable them to identify genomic CNV regions with tandem duplication. It should be noted that Pindel and Delly are not capable of detecting non-tandem duplications, on the other hand. Moreover, a major deficiency of CNV-seq, Pindel and Delly is that none of these tools detect the digitized copy number of CNV regions. However, MGP-HMM detects both tandem and non-tandem duplications and discriminates them from each other and also estimates the digitized copy number of genomic regions.

For comparisons, precision, recall and F-score (the harmonic mean of the precision and recall) are calculated for each type of CNVs i.e. tandem duplications, non-tandem duplications, copy gain, and diploid, separately. Also, to measure the dependency of accuracies to the data, the whole simulation study is repeated five times. Then, average and standard deviation of accuracies over five repeats of the whole study are reported in Tables 3 and 4. It should be noted that we have considered different sequencing coverages i.e. 1×, 5×, and 10×, for comparisons.

Table 3
The overall accuracy of MGP-HMM is compared with CNV-seq, for detecting the genomic CNV regions.

<table>
<thead>
<tr>
<th></th>
<th>Sequencing coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1×</td>
</tr>
<tr>
<td></td>
<td>CNV region†</td>
</tr>
<tr>
<td>MGP-HMM</td>
<td>0.87/0.02</td>
</tr>
<tr>
<td>CNV-seq</td>
<td>0.58/0.02</td>
</tr>
</tbody>
</table>

† Number of nucleotides in CNV regions whose state were correctly predicted is divided by the total length of the genomic CNV regions.
‡ Number of nucleotides whose states were correctly predicted is divided by the genome length.
Table 4

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Method</th>
<th>Tandem duplications</th>
<th>Non-tandem duplications</th>
<th>Loss</th>
<th>Diploid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1×</td>
<td>MGP-HMM</td>
<td>0.88/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Pindel</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>CNV-seq</td>
<td>0.60/0.02</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Delly</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td>5×</td>
<td>MGP-HMM</td>
<td>0.86/0.02</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Pindel</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>CNV-seq</td>
<td>0.60/0.02</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Delly</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td>10×</td>
<td>MGP-HMM</td>
<td>0.86/0.02</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Pindel</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>CNV-seq</td>
<td>0.60/0.02</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
<tr>
<td></td>
<td>Delly</td>
<td>0.80/0.01</td>
<td>0.96/0.01</td>
<td>0.90/0.01</td>
<td>0.99/0.01</td>
</tr>
</tbody>
</table>

As shown in Table 3, the overall performance of MGP-HMM is superior to CNV-seq, for all coverage values. Especially according to Table 4 in genomic diploid states and regions with copy gain, MGP-HMM reaches accuracies which are superior to CNV-seq in terms of F-score.

The performance of MGP-HMM is superior to Pindel in detecting all CNV states, Table 4. It also outperformed Delly in detecting genomic diploid states and regions with copy loss. The drastic performance of Delly and Pindel in predicting genomic regions with copy loss is remarkable, which both reached an F-score lower than 50% for all coverage values. However, Delly is ranked as the best method in predicting tandem-duplications and MGP-HMM is ranked second.

MGP-HMM is also compared with CNV-seq, Pindel, and Delly in terms of the running time and RAM usage. For this purpose, we applied a 64-bit operating system with an Intel Core(TM) i7-4710HQ CPU @ 3.50 GHz processor and 16 GB RAM. In Table 5, the running time and RAM usage are reported for the CNV detection in the same simulated dataset that is used in this section. As shown in Table 5, due to the iterative parameter learning in MGP-HMM, its running time is higher than other CNV detection methods. Also, CNV-Seq and Delly have the lower RAM usage in comparison with MGP-HMM and Pindel. The RAM usage of MGP-HMM is lower than Pindel for a genome-wide sequencing coverage of 5×, and 10×.

4. Implementation in real data

BAM files of the alignment of mate pair reads to the human reference genome (hg18) for a Hap Map individual NA12878, is downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/data/NA12878/alignment/. This is a high coverage whole-genome shotgun sequencing of illumina platform. Then, SAMtools (samtools.sourceforge.net) is applied for filtering out mate pairs of low mapping quality (<Q25).

MGP-HMM is applied for the genome-wide CNV detection in NA12878 and it found a total number of 50,906 CNVs with length in the range of 110–84,287 bp. The majority of these calls are non-tandem duplications with a total number of 50,143 CNVs and length in the interval of 110–4950 bp. Also, the number of deletion (either heterozygous or homozygous) and tandem-duplication calls are 714 and 49, whose length varies in the interval of 111–84,287 and 361–68,228 bp, respectively.

The total length of CNV calls is 26,176,358 pb which covers 17.93% of the chromosome 8 of Hap Map individual NA12878. In more details, the total length of non-tandem duplication calls is 23,436,160 bp, covering 16.05% of the studied autosomal chromosome. Moreover, the total length of deletion and tandem-duplication calls are 1703,936 and 1036,262 bp which cover 1.17% and 0.71% of the chromosome, respectively.
The authors would like to thank associate editor and two anonymous referees for comments. The financial support of INSF (No. 95825345) is gratefully acknowledged.

**References**

[22] G. Te, M.T. Tamimi, N. Ben-Dor, A new method to detect copy number variation using high-throughput sequencing, BMC Bioinform. 10 (2009) 80.
[34] A. Gusnanto, et al., Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data, Bioinformatics 28 (1) (2012) 40–47.

**Funding**

This work was supported by a grant from the School of Biological Sciences at Institute for Research in Fundamental Sciences.

**Acknowledgments**

Hamid Pezeshk and Seyed Amir Malekpour would like to thank department of research affairs at University of Tehran.

**Table 6**

<table>
<thead>
<tr>
<th>Number of CNV calls</th>
<th>Overlap against DGV (by calls) (%)</th>
<th>Overlap against DGV (by bases) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-tandem duplications</td>
<td>50,143</td>
<td>54.85</td>
</tr>
<tr>
<td>Deletions</td>
<td>714</td>
<td>90.23</td>
</tr>
<tr>
<td>Tandem duplications</td>
<td>49</td>
<td>81.82</td>
</tr>
<tr>
<td>All CNVs</td>
<td>50,906</td>
<td>58.22</td>
</tr>
</tbody>
</table>


