Building ultra-high density linkage maps based on efficient filtering of trustable markers

1Ronin Y.I., 1Mester D.I., 1Minkov D.G., 2Akhunov E., 1*Korol A.B.

1Institute of Evolution and Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 3498838, Israel

2Kansas State University, Department of Plant Pathology, Manhattan KS 66506-5502, USA
Building ultradense linkage maps

Key words: marker space, co-segregating markers, twin groups, skeletal markers; marker clustering, genotyping errors, missing data, marker filtration

*Correspondence:
Phone: (972)-48240-449, Fax: (972)-48288-788
korol@research.haifa.ac.il
Abstract

The study is focused on addressing the problem of building genetic maps in the presence of \(10^{3}-10^{4}\) of markers per chromosome. We consider a spectrum of situations with intrachromosomal heterogeneity of recombination rate, different level of genotyping errors, and missing data. In the ideal scenario of the absence of errors and missing data, the majority of markers should appear as groups of co-segregating markers ("twins") representing no challenge for map construction. The central aspect of the proposed approach is to take into account the structure of the marker space, where each twin group (TG) and singleton markers are represented as points of this space. The confounding effect of genotyping errors and missing data leads to reduction of TG size, but upon a low level of these effects surviving TGs can still be used as a source of reliable skeletal markers. Increase in the level of confounding effects results in considerable decrease in the number or even disappearance of usable TGs and, correspondingly, of skeletal markers. Here, we show that the paucity of informative markers can be compensated by detecting kernels of markers in the marker space using a clustering procedure, and demonstrate the utility of this approach for high-density genetic map construction on simulated and experimentally obtained genotyping datasets.
INTRODUCTION

The problem: In recent years, new genotyping technologies based on DNA-arrays (chips) and next generation sequencing (NGS) have become widely available for scoring thousands of single nucleotide polymorphic markers (SNPs) in a wide spectrum of model and non-model organisms. These datasets pose new challenges for building high-density genetic maps. With large-scale chip-based SNP genotyping data, genotypic-by-sequencing (GBS) or specific-locus-amplified-fragment-sequencing data (SLAF-Seq) (e.g. Qi et al. 2014), building genetic maps with $10^5$-$10^6$ markers per genome (or $10^3$-$10^4$ markers per chromosome) requires new algorithms. Indeed, the dramatic increase in the number of markers is only one of the challenges. Among other difficulties with such an amount of markers are genotyping errors, missing data, and small population size. If the mapping algorithms cannot efficiently cope with these problems, generating big SNP marker sets for building ultra-dense maps will not achieve the goal. Obviously, the population size sets an upper limit to the number of markers per chromosome that can be resolved by recombination; genotyping errors and missing data calls may complicate deducing the correct marker order in the chromosome.

Some limitations of recent mapping algorithms when dealing with high throughput data with genotyping errors and missing scores: Usually a two-phase approach is applied for genetic mapping: clustering of all markers into linkage groups (LGs) and ordering the markers within each LG. Earlier algorithms and software packages for genetic mapping were based on a few approaches suitable in a situation when the number of markers per population was relatively small, e.g., a few tens or hundreds per chromosome. In both phases, a full distance matrix for the chromosome markers is required. In case of a significantly increased dimension of the problem ($n \sim 10^4$-$10^6$), the existing algorithms for genetic mapping cannot solve the problem in reasonable computer time, i.e., even using simple optimization algorithms of order $O(n^2)$. Moreover, a huge computer memory (RAM) for the distance matrix is required on the
clustering and map construction phases (but see Strandova-Neeley et al. 2015). With big data, even more challenging are the difficulties caused by missing scores and genotyping errors. Usually markers with considerable missing (e.g., 10-20%) are removed from the dataset at the pre-treatment stage, while markers with genotyping errors are not removed. Instead, their positions are slightly corrected by corresponding maximum likelihood (ML) algorithms (Wu et al. 2008; Rastas et al. 2013, 2016). The inability of existing map construction algorithms to cope with these factors, rather unexpectedly for the wide genetics community, posed a serious problem. In some cases maps of 400-800 cM have been obtained and required additional "re-scaling" correction in order to correspond to the expectations based on cytogenetic analysis of meiosis (Wang et al. 2014).

**New approaches to address the problem:** Several algorithms, proposed to deal with big mapping datasets, employ the ideas of k-means (Arthur and Vasilevskii 2007) or k-nearest neighbors (Dongyuan et al. 2014) for marker clustering into LGs followed by the ordering of each LG. The genetic map is represented as a linear sequence of ordered marker subsets $S_i$ (bins), with the number of markers in each subset defined by a neighborhood of radius $r_i$ around a center $c_i$ (Strandova et al. 2014; Dongyuan 2014). Ordering is conducted for bins rather than markers. In certain cases, genetic mapping can be reduced to the Minimum Spanning Tree (MST) problem (Wu et al. 2008) instead of a more traditional reduction to the Traveler Salesperson Problem (TSP) (Mester et al. 2003). The MST algorithm gives fast and good solutions for low-noise data and simple geometry of the spanning tree, i.e., when the majority of 'leaves' of the tree are interconnected (via linkage) in a linear-like structure corresponding to the organization of eukaryotic chromosome and only a small part of markers appear in the tree branches (Rastas et al. 2013). But if the number of markers in the branches is large, the maximal MST path may inadequately represent the chromosome. In such cases, MST can serve only as a source of an initial solution that should be complemented by markers from the branches. Thus,
the Lep-Map algorithm (Rastas et al. 2013, 2016) imitates MST construction in finding a feasible initial order (path of maximum length) and then inserts markers from MST branches into the path via TSP heuristics. After this step, local changes in the order are applied to maximize the likelihood of the final order. However, the MST approach cannot manage situations with large number of markers in the presence of genotyping errors and missing data.

Another approach to solve the problem was first described in our short report (Ronin et al. 2015). Its central idea is to take into account the structure of the marker space of the mapping problem, where each point represents a marker with \( n \) coordinates corresponding to the marker alleles of \( n \) genotypes of the population. With this approach, in addition to routine filtering of markers based on segregation distortion and level of missing data, we suggested a heuristic procedure of selecting high-quality markers. It is based on the assumption that error-free markers are more abundant among groups of co-segregating (twin) markers, which should have priority during the selection of 'skeletal markers' for inclusion into the genetic map. If the error rate is low (e.g., \( p_e \sim 0.01-0.02 \)), a sufficient amount of such markers can be selected to build a high quality map. Here we propose a new approach for constructing genetic maps using big genotyping data (with up to \( 10^3-10^4 \) markers per chromosome), which extends the method by Ronin et al. (2015) and includes an additional filtering step to cope with a higher level of errors (say, \( p_e \sim 0.02-0.04 \) or more). Obviously, with the higher error rates, the quality of the maps is supposed to decrease. We show that the increase in the error rate can be compensated by the availability of a large number of markers allowing for building good quality maps. In our algorithm, the procedure of choosing reliable marker candidates from noisier subsets of markers is applied after the best candidates, representing twin groups (TGs), have already been selected. The remaining markers are clustered and the representatives of such clusters, satisfying certain conditions, are appended to the set of the best candidates representing the TGs. The choice between the two approaches or usage of a hybrid strategy integrating both approaches, for any
dataset can be based on preliminary filtering/clustering cycles as described in the paper. The chromosomal distribution of markers suitable to be tried as candidates for the skeletal map at the consequent stages of analysis is shown in Fig. 1.

**Figure 1**

For the ordering the selected candidate markers various optimization algorithms can be applied, for example, GES (Mester *et al.* 2010; Ronin *et al.* 2010), L&K (Lin and Kernighan 1973; Helsgaun 2000; Appletgate *et al.* 2003) and Concord (Appletgate *et al.* 2001). The efficiency of our approach for the selection of the most informative candidates was studied here on simulated and real datasets. Ordering the selected candidates, testing and stepwise improving of the genetic map is then conducted using the effective scheme described in our previous publications (Mester *et al.* 2003, 2004, 2010; Korol *et al.* 2009; Ronin *et al.* 2010, 2012, 2015).

**MATERIALS and METHODS**

**Simulation of mapping data:** For testing the algorithms, we employed simulated and real mapping populations of doubled haploid (or backcross) type, with a population size $n=200$, and the number of markers $N_m=10000$, 20000 and 40000 per chromosome. In many organisms the distribution of recombination events along chromosomes is highly heterogeneous, with very high differences between peri-centromeric and sub-telomeric regions due to the centromere and heterochromatin effects on recombination (Korol *et al.* 1994; Akhunov *et al.* 2003; Backstrom *et al.* 2010; Roesti *et al.* 2013; Sharma *et al.* 2013; Hill *et al.* 2015; Wang *et al.* 2015; Nambiar and Smith 2016; Tsai *et al.* 2016). Another contributing factor may be correlation between recombination rate and chromosomal variation in GC content along chromosomes (Duret and Arndt 2008). Obviously, upon an even distribution of polymorphic markers with respect to DNA physical length, the density of markers per unit of recombination in the regions with low recombination rate will be much higher than in high-recombination regions. Therefore, in our
simulations, three different regions were considered with respect to the proportion of simulated markers and genetic map length \( L \) (cM). Namely, the peri-centromeric and the two sub-telomeric regions included 80% and 20% of \( N_m \), while the contribution of the peri-centromeric part to the genetic map length was much lower compared to the sub-telomeric parts (Table S1). Simulation of recombination distances between adjacent markers for each region was conducted by sampling the distance values from the pre-set region specific ranges of very small, small, and moderate distances \( (d_{vs}, d_s, \text{ and } d_m, \text{ respectively}) \) (Table S2). The average characteristics of the resulting maps constructed for error-free data are presented in Table S3. For all notations in the text and the tables and figures see supplementary File S1 (Glossary).

**Construction of skeletal map in the case of low level of genotyping errors using the "twin" method**: The major part in the proposed approach for building dense genetic maps is selecting informative skeletal markers. Depending on the level of genotyping errors (unknown a priori), we developed and evaluated two procedures to address this problem. The first procedure is based on the fact that under low levels of genotyping errors (e.g., \( p_e = 0.01-0.02 \) per marker locus), large number of markers per LG (e.g., \( N_m \sim 10^4 \)) and relatively small sizes of the mapping population \( (n \sim 10^2) \), a considerable proportion of markers will appear as large groups of twins groups (TGs) (Ronin et al. 2015). For such situations, a simple and efficient principle for selecting reliable skeletal markers is to take representative (‘delegate’) markers with the minimal rate of missing data for each TG (see File S2).

The process of constructing a skeletal map includes three stages. The first stage is to select a threshold value \( t_{s0} \) for the size \( ts \) of TGs, which will be represented by their delegate markers in the initial variant of the skeletal map (Ronin et al. 2015). In our approach, the selected markers are then ordered based on the reduction of the mapping problem to TSP using the Evolutionary Strategy heuristic optimization (Mester et al. 2003, 2004, 2010; Ronin et al. 2010). The second step is testing map quality using jackknife re-sampling followed by the
deletion of markers violating local map stability and monotonicity (i.e., increase in recombination rate between a marker and its subsequent neighbors) (Ronin et al. 2010). After this step, we can insert in the resulting map additional markers representing TGs with smaller sizes compared to the chosen $t_s0$ (as well as suitable singleton markers not causing map inflation), and then check the map quality again. This step may be helpful for filling in the gaps in the genetic map. Such cycles can be applied repeatedly (Fig 2).

**Figure 2**

**Skeletal map construction in the presence of high genotype calling error rates using clustering:** The second approach is designed to manage situations with higher levels of errors (e.g., up to $p_e=0.02-0.04$). The decrease in the proportion of error-free markers leads to the erosion of most of the TGs. Thus, for a marker with $p_e=0.03$, the probability that in a population with $n=200$ genotypes none of the marker scores could be erroneous is $P = (1-0.03)^{200} \approx e^{-6} \approx 0.0024$. Here the average number of errors is 6, implying that two markers inseparable by recombination (upon an error-free situation) will show a "distance" of ~6 cM. In order to extend the twin-based filtration idea to such situations, when the number of TGs is not sufficient for covering the chromosome even at a low-marker density, we employ marker clustering. An important geometrical fact is that in the $n$-dimensional space of markers, many groups of co-segregating markers that should be represented in this space as one point per group, as a result of errors will turn into clouds (clusters) of close points. The mid-point of such a cloud is geometrically close to the position of the corresponding (error-free) set of completely linked markers, or twins. Thus, in the space of markers, noisy markers geometrically represent a "fuzzy" set of varying density, with higher density in the vicinity of the residence point of the original error-free markers. Bearing this in mind, we complement the procedure based on using TGs with zero intra-group distances, by a clustering procedure that dissects the entire set of markers into clusters in such a way that the sum of distances of markers to the means of their
clusters, taken over all clusters, will be minimal (File S3). This enables covering the chromosomal regions with relatively high recombination rates, where the joint effect of an increased proportion of false (due to errors) "recombinants" and lower real density of markers per cM may lead to a negligible proportion of the remaining TGs. With this approach, we select a representative marker from each cluster (the marker closest to the centroid) and use these representatives as additional candidate markers. This approach enables to build good quality maps even under the paucity or absence of TGs (due to genotyping errors) in datasets with high error rate. After selection of markers representing the obtained clusters, the next steps are identical to those described above for the twin approach, i.e., resampling-based detection and removing markers violation map stability and monotonicity combined with saturating the skeletal map by filling in the gaps wherever possible.

Although a priory we cannot know which of the two foregoing situations (low or high level of genotypic errors) is characteristic of the target mapping project, it is easy to address this question by a trial analysis and evaluate the number of TGs and the chromosomal distribution of twin sizes, and thereby to assign the project to the first or second class. Obviously, before starting the mapping analysis, trivial preliminary removing of low quality data is needed based on the level of missing data and deviation of segregation ratios from the expected ones. In reality, the foregoing two situations, with high vs negligible number of twins, may take place simultaneously within the same chromosome. As noted above, this may happen due to the effect of centromere and heterochromatin blocks on recombination. To deal with such situations, we employ a hybrid procedure (see Fig. 2).

**Characterizing the quality of constructed maps:** In the analysis of simulated data, to characterize the quality of maps constructed using different algorithms or different parameter settings of the same algorithm, we employ as a reference the map representing the “true” (simulated) order of markers, corresponding to the ideal error-free case with no missing data
points. Then, the best algorithm is the one that generates a solution closest to the simulated order. Several parameters were used to assess the map quality. Bearing in mind that in a simulation study we do know the true order, the simplest score of map quality would be the coefficient of recovery ($C_r$), as described in Mester et al. (2003). Upon error-free genotyping, many markers are expected to co-segregate due to a high ratio of the total number of markers to population size, leading to the limited use of $C_r$ because it does not take into account the fact that genotyping errors and missing marker scores can lead to fissions of TGs into groups of a smaller size (see Results). Thus, instead of $C_r$ we employ two other characteristics: (a) $n_e$, the number of errors in the order of markers compared to the simulated order (we consider as an error each situation when the marker's original rank in the constructed skeletal map is higher than the rank of the next marker in the map); and (b) $n_r$, the number of "repeats" resulting from the separation of the initial groups of co-segregating markers into sub-groups due to genotyping errors and missing marker calls; such repeats will appear in the constructed skeletal map at separate (usually, but not necessarily, adjacent) positions. Fig. 3 (a and b) illustrates the calculation of $n_e$ and $n_r$. Numbers from 1 to 9 are the numbers of co-segregating groups in the simulated map; the figure shows the estimated order of noisy marker data. But the degree of deviation from the true order in the skeletal map is only a partial characteristic of the map quality; map coverage ($m_c$) is another important score (Fig. 3c,d). Additional quality scores used in our analysis included: loss factor $l_f$ (%), which is the percentage of lost (non-characterized) unique marker positions in the constructed map compared to the simulated map: $l_f = 100 \frac{(N_{skef} - (N_{sk} - n_3))}{N_{skef}}$, where $N_{skef}$ and $N_{sk}$ are the number of intervals in the skeletal maps built for the simulated error-free and noisy markers.

Figure 3
In addition to simulated data, we demonstrated our approaches on several wheat chromosomes using data generated using the 90K iSelect SNP genotyping assay for 150 doubled haploid (DH) wheat lines (Wang et al. 2014).

RESULTS

Both procedures considered in this paper are based on selecting a subset of most informative markers, referred to as 'skeletal' markers, as a basis for constructing a quality genetic map (Mester et al. 2003, 2004, 2010; Ronin et al. 2010). In the first procedure, the selected subset of markers for the skeletal map is comprised of 'delegate' markers representing TGs. In the second procedure, when the proportion of TGs becomes small, we cluster markers into non-overlapping sub-sets, e.g., using k-means or other algorithms (Bishop 1995; Strandova-Neeley et al. 2015); markers with minimal average distance to all other markers of such a subset serve as delegates. In reality, upon moderate level of genotyping errors the two situations may coexist within the same project, e.g. with considerable and very low proportions of co-segregating markers in near-centromeric regions (due to low cM/Mbp) and sub-telomeric regions (due to high cM/Mbp), respectively. Therefore, in the present examples we illustrate the proposed analytical schemes on the two types of practically important situations: (i) with the low level of genotyping errors (hence high proportion of TGs); and (ii) with a relatively high level of errors (hence much lower proportion of TGs) justifying the efforts to look for the representatives of "fuzzy TGs" by using cluster analysis. For both situations, we consider simulated data as well as real data on wheat.

Analysis of simulated data: Although one of the effects of missing data in error-free genotyping data is the reduction in the number of TGs, non-zero but small level of errors leads to an increase in the number of TGs (Table S7; for comparison see File S5 and Table S4 with
results for 'no-missing' situations). Nevertheless, further increase in the error rate reverse the
direction of changes in the number and size of TGs, and thereby in the number of and
confidence in the selected candidate skeletal markers. Thus, qualitatively the effect of missing
genotype calls does not change the main conclusions reached for 'no missing ' situation. With
the increased level of genotyping errors a more liberal threshold $t_{s0}$ should be chosen compared
to the situations with low error rate. In Table 1, for a population of size $n=200$ with two levels
of errors (1% and 2%), we show how the results of the described map construction procedure
depend on the choice of $t_{s0}$ for different numbers of available markers per chromosome. As
expected, for datasets with higher $p_e$ more liberal choice of $t_{s0}$ should be recommended
compared to the datasets with lower $p_e$. The observed distribution of the numbers of groups of
size 2, 3, 4 etc. can serve as diagnostic characteristic of the underlying real situation.

Table 1

<table>
<thead>
<tr>
<th>$t_{s0}$</th>
<th>$N_{m}$</th>
<th>$L$ (cm)</th>
<th>$L_1$ (cm)</th>
<th>$L_2$ (cm)</th>
<th>$L_c$ (cm)</th>
<th>$N_{sk}$</th>
<th>$N_{t1sk}$</th>
<th>$N_{t2sk}$</th>
<th>$N_{csk}$</th>
<th>$l_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>$10^4$</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>10.0</td>
</tr>
<tr>
<td>0.010</td>
<td>$10^4$</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>9.0</td>
</tr>
<tr>
<td>0.020</td>
<td>$10^4$</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>8.0</td>
</tr>
</tbody>
</table>

The observed distribution of the numbers of groups of
size 2, 3, 4 etc. can serve as diagnostic characteristic of the underlying real situation.
Map coverage ($m_c$): For the entire range of parameters, $m_c$ was higher than 97%, excluding the cases with simultaneously maximal values of $p_e$ and $t_s_0$. In such cases, the proportion of TGs may become very small (not sufficient to recruit candidates for skeletal map), especially in regions with high cM/Mb. Such situation requires the complementation of twin markers by markers representing the clusters of tightly-linked but not co-segregating markers (see below).

Map density: map density decreased ($\delta$ increased) with the increased level of genotyping errors $p_e$ and the chosen threshold $t_s_0$, especially in the sub-telomeric regions, due to a higher level of error-driven degradation of TGs. On average, a 2-2.5 increase in $\delta$ was observed for the maximal considered level of genotyping errors $p_e=0.02$.

Number of "repeats" ($n_r$) in the map: as explained in section Methods and illustrated in Fig. 3, genotyping errors and missing data calls may lead to dissociation of a part of the initial TGs into sub-groups; representatives of such sub-groups appear in the constructed skeletal map at separate (usually, but not necessarily, adjacent) positions. Table S8 shows that increasing in the chosen $t_s_0$ results in lower $n_r$, implying more efficient filtration in favor of better markers.

Errors in the order of markers ($n_e$): in the resulting skeletal maps, $n_e$ was rather small. Yet, attempts to compensate the effect of high $p_e$ by choosing too stringent $t_s_0$ may lead to a considerable increase of $n_e$ due to a drop in the number of skeletal markers and map coverage.

Two approaches can be used if small number of TGs (i.e., candidate skeletal markers) and low coverage were obtained for the chosen $t_s_0$: (i) reduce $t_s_0$, thereby increasing the number of TGs; and (ii) recruit additional candidate markers based on the k-means or similar clustering procedures (see section Methods). With a low level of genotyping errors, the first approach may be sufficient, at least for genomic regions with relatively low cM/Mb ratio; in combination with the second approach it may enable solving the problem for the whole genome. However, with high level of errors, the number of TGs may be too small even at $t_s_0=2$. In this case, the second approach can be used as a major source of candidate skeletal markers. These considerations are
reflected in the examples present in Table 2, where combined analysis enabled considerably improving the map coverage and reducing the proportion of lost map unique positions.

Table 2

The decrease in the number of TGs may also be caused by increase in the population size \((n)\). Indeed, we have noted above that even for the small level of genotyping errors, increase in \(n\) leads to decrease in the size of the TGs. In Table 3, we show the effect of population size on map characteristics for error-free data and moderate level of genotyping errors \((p_e=0.01)\) and missing data points \((ms=20\%)\). For \(n=200\), twin approach was sufficient to build a good quality map, but for \(n=500\) it became impractical due to a catastrophic decrease in the number of TGs in the sub-telomeric regions. Therefore, the map was constructed using the combination of twin approach and clustering. And finally, for \(n=1000\) only clustering approach was suitable. It is noteworthy that despite the growing proportion of lost markers \(l_f(\%)\), the number of markers in and the genetic length of the constructed skeletal maps also grow with the population size. The last effect results from the known fact that genotyping errors lead to map length inflation.

Table 3

Analysis of empirical datasets: As was shown above, genotyping errors can result in decrease in the number of TGs reducing the number of candidate markers for the skeletal map. Therefore, the usage of some of the standard mapping algorithms with the large number of markers can result in maps with inflated inter-marker distances. The ability of our approaches to effectively deal with the high-density marker datasets was demonstrated by comparing the maps constructed for several wheat chromosomes using the MST algorithm (Wang et al. 2014) and the twin approach (Table 4). The lengths of MST-maps were 2-3 times longer than those constructed using the twin method and strongly disagreed with chiasma frequencies (~1.7-2.5 per bivalent) as cytogenetics characteristics of meiotic recombination in wheat (e.g., Feldman 1966; Koul et al. 2000). For chromosome 2A, in spite of the high estimated coefficient of MST-map coverage (0.994) by the markers from the UDM-map, we observed substantial differences
in the estimates of genetic distances between the markers (Table S9). Moreover, the number of the identified unique recombination map intervals (bins) was substantially higher in the MST-map (220 bins) than in the UDM-map (125 bins). More than half of the MST-map bins contained a single marker, whereas on the UDM-map only 3 bins had single markers. Considering the size of the DH population, it is unlikely that the inferred number of unique recombination bins on the MST-map is real; most likely it is caused by genotyping errors resulting in the over-estimation of the recombination rate by MST. Therefore, more than hundreds of bins on the MST-map are represented by the replicated co-segregating markers that should be excluded from the map. While it is possible that some of these single markers do capture unique recombination events not be accounted for in the final UDM-map, the usage of our approach would exclude hundreds of erroneously identified unique recombination events.

Table 4

As a further complication of the considered example, the set of 26,000 markers generated based on 90K iSelect platform was merged with a set of ~421,000 markers obtained using genotyping by sequencing (GBS) for the same population (Saintenac et al. 2013). Some of GBS markers were used as usual two-allele SNPs, but the majority were of presence-absence type, representing either M6 or Opata alleles. The combined dataset was filtered to exclude markers with too high level of missing scores (more than 40) and too high segregation distortion ($\chi^2>35$), leaving for further analysis ~130,400. This set was analyzed using twin approach with $ts_0=4$. After removing from the map markers violating map stability and monotonicity (Ronin et al. 2010), followed by map saturation with markers from smaller-size TGs and singleton markers, the total number of skeletal markers was 1481. As illustration, we provide here only the results for 2B chromosome. The number of skeletal markers in the obtained map of 158.5 cM length was 81 (total number of markers was 526 when cosegregating markers were taken into account) (Fig. S1). If we also attach markers for which intervals of 2B are their best location
(but their inclusion to the skeletal would considerably reduce the map quality), then the total number of markers associated with 81 markers of 2B skeletal map will be 2241 (Table S10).

**Comparing MST and ULD algorithms on simulated data:** For comparisons we employed a double haploid mapping population with $M=2000$ markers positioned on 84 separable by recombination positions (for population size $n=200$) of the simulated chromosome of $L=136.8$ cM (Kosambi metric). As can be seen from the results in Table 5, for both levels of missing data (0 and 10%), the MST map undergoes an increase in the number of map bins and inflation of the map length growing with the rate of genotyping errors. This was the case even for error rates as small as 0.001: i.e. 30-35% for $L$ and 70-110% for bin number increases as compared to simulated parameters. For a more realistic error level (1%), the corresponding numbers were: 400-450% for $L$ and 750-900% for bin number. Unlike MST, the length of maps constructed with our approach practically do not vary with the error rate and remains remarkably close to the simulated map despite one order of magnitude in variation of genotyping error rate. Interestingly, the obtained results fit rather well the patterns observed in the above examples on real data on wheat, especially for chromosomes 2B and 5A (see Table 4), suggesting that the rate of genotyping errors in that data could be approximately 1-2%. Additional important criteria that we used to assess the quality of maps constructed for simulated data were $n_r$ (the number of "repeats") and $l_f$ (percentage of lost unique positions in the constructed skeletal map compared to the simulated map) (see Materials and Methods and File S1). As can be seen from Table 5, even in the worst cases, the number of repeats does not exceed the rate of one repeat per 3-4 skeletal markers. For MST, a lower bound estimation of $n_r$ varies with the number of bins, from two to 10 repeats per marker. With our approach, the proportion of lost unique positions in the considered examples is a growing function of the error rate and missing data calls ($l_f$ varied from 5% to 40%). Obviously, the increase of $l_f$ is an unavoidable cost for the noise caused by errors when our attitude is to employ error filtration for getting maps with
minimum number of errors in the reconstructed order of markers. An attempt to keep as many markers as possible leads to maps with unrealistic length and rather questionable marker order.

Table 5

**DISCUSSION**

The increase in the number of markers by orders of magnitude achievable by the new technologies (GBS, RAD-seq, RNA-seq, etc.) was perceived as a breakthrough that enables building quality ultra-dense genetic maps. This expectation, being basically correct, may practically be far from reality, due to difficulties caused by genotyping errors, missing marker calls, strong intra-chromosomal variation in recombination density, etc. These factors may lead to biased estimates of recombination rates and wrong marker orders, especially for markers with increased error load. Obviously, if the marker set is small, there are no much possibilities for filtering and most of the markers will have to remain in the map even if the map lengths are inflated. However, high-throughput genotyping provides ample amount of data that can be filtered to obtain high-quality datasets. The application of technology-specific filtering and the usage of appropriate quality controls is the first step for generating reliable data usable for map construction. At the level of genotyping datasets, data filtering also imply not just the detection and removal of markers with high segregation distortion and massive losses of data. It is also possible to detect and remove markers violating local map stability by using jackknife re-sampling (Mester et al. 2003; Ronin et al. 2010). Another quality test detection of non-monotonicity of recombination rates along the map, although in some cases such deviation may be caused by negative interference (Denell and Keppy 1979; Peng et al. 2000; Korol et al. 2009; Aggarwal et al. 2015). The method of pre-mapping filtering described in this paper is based on the idea that with very large numbers of scored markers, many markers remaining irresolvable by recombination and appearing as twin groups (TGs) exceeding certain minimal preset threshold size $t_{so}$, can be trusted more than singleton markers or markers from smaller-size TGs.
Simple analysis shows the importance of proper decision about $ts_0$ for selecting trustable markers for the skeletal map. We are rather skeptical with respect to some alternative approaches that first build a trial genetic map and then apply different ways of marker correction followed by subsequent map correction. With higher level of errors, the proportion of error-free markers may become negligible. To cover such situations as well, we employ a generalization of the twin approach based on analysis the geometry of the $n$-dimensional space of markers of the mapping population with each marker being presented as a point. With such presentation, the genotyping errors lead to dissipation of TGs, so that the resulting marker agglomerations are “blurred” around the positions of the (unobservable because of errors) initial points corresponding to error-free situation. Therefore, with higher level of errors we employ an additional filtration: after the TGs exceeding the threshold size $ts_0$ are selected as candidate for the skeletal map, we conduct clustering of the remaining markers by a procedure similar to k-means algorithm. Then, representative markers of clusters are added to the set of selected candidates for building the skeletal map. The next steps include resampling-based detection and removing markers violation local map stability and monotonicity combined with saturation of the skeletal map by filling in the gaps wherever possible (Ronin et al. 2010). An important factor in getting a high quality map from the available data given the known parameters (population size, total number of markers and missing) is the choice of threshold value TG size $ts_0$ and initial radius $r$ in the clustering procedure. Although both these parameters should depend on $a priory$ unknown rate of genotyping errors $p_e$, several trials should usually be enough to clarify the situation and make a rational choice (as illustrated in the section Results).

Like other approaches, our analysis starts from pre-mapping filtering based of simple criteria (segregation distortion and missing-data). Like other approaches, we also reduce the size of the target dataset by removing redundancy (by representing TGs by single markers). However, unlike others, we take advantage of the information on marker quality hidden in the
structure of the multidimensional marker space, in particular in the sizes of TGs as well as in "derivatives" of such groups resulted from dissipation due to genotyping errors and missing data. Depending on genotyping quality, the number of available markers, and the population size, the number of confidently ordered skeletal markers may vary from several tens to several hundred per chromosome. Yet, markers 'represented' by the skeletal markers (i.e. their cosegregants) plus centroid markers from the blurred clusters and markers attached to the closest intervals of the skeletal map (but not included, to prevent map length inflation) may reach tens or even hundreds of thousands. The described system is implemented in an interactive user-friendly software MultiPoint-ULD (MULD) for building ultradense genetic maps for controlled crosses (backcross, doubled haploids, F2, RIL populations); further development will also include F1 progeny of outbred species and multi-parental populations. Our approach for pre-mapping filtering, together with previously developed algorithms (Mester et al. 2003; Ronin et al. 2010) implemented in the MultiPoint software enable building quality genetic maps, with realistic map length and reliable marker orders (e.g., Avni et al. 2014; Reddy et al. 2014). A trial version for Windows with simulated examples can be downloaded using the link: http://evolution.haifa.ac.il/images/stories/Software/MultiPoint-ULD_Demo.zip

The proposed approach for building high-quality dense genetic maps with a possibility to deal with big datasets of SNPs may be helpful in addressing a wide range of genetic and genomic problems. We list below just a few for illustration. Although several tens of markers per chromosome may be enough for usual applications of linkage mapping of trait loci (Mendelian or quantitative) based on bi-parental mapping populations, for association mapping and genomic selection the requirements are much more challenging, especially in situations with steep decay of linkage disequilibrium. Similarly, for map-based cloning, high density of markers is needed to get as close as possible to the target candidate gene, implying the availability of accurate dense map. Dense genetic maps were successfully used for anchoring
physical contigs (Raats et al. 2013) and sequence scaffolds (Mascher and Stein 2014) to LGs and controlling the quality of sequence assembly (Hedgecock et al. 2015; Zeng et al. 2017). The high-coverage shotgun sequencing in combination with new analytical tools of sequence scaffolding, ultradense mapping and 3-dimensional chromosome-conformation-capture-sequencing data was successfully used for high-quality sequencing of such a big and complex genome as wheat (https://www.wheatgenome.org/News2/RefSeq-v1.0-URGI). High quality dense genetic maps become a powerful tool for detailed analysis of recombination genomic distribution, sex dependence, genetic variation and genetic control (Bauer et al. 2013; Rodgers-Melnicka et al. 2015; Ross et al. 2015; Li et al. 2016; Tsai et al. 2016), and genome comparisons of related species (Hill et al. 2015). Obviously, high quality of dense genetic maps is vital for successful using of genome mapping in these and numerous other applications in non-model organisms, where validated genome sequence is currently not available. We believe that such studies will benefit from using the described here approach.

**Acknowledgments:** We are thankful to X. Chen, S. Lonardi and M. Moscou for providing us with original sources of the MergeMap and IPLMap programs. We acknowledge with thanks G. Churchill and two anonymous referees for helpful comments and suggestions on the first version of the manuscript. This work was partially supported by the Israel Science Foundation (grant #800/10), US-Israel BARD foundation (grant #IS-4137-08), European FP7 Programme TriticeaeGenome (grant agreement number FP7-212019), and MultiQTL Ltd.

**REFERENCES**


42. Tsai, H.Y., D. Robledo, N.R. Lowe, B. M. Bekaert, J.B. Taggart et al., 2016 Construction and annotation of a high density SNP linkage map of the atlantic salmon (Salmo salar) genome. G3 6: 2173-2179.


Table 1. Effect of the selected threshold for minimal twin group size $ts_0$ on the skeletal map length under low and moderate rates of genotyping errors and 20% of missing data points.

<table>
<thead>
<tr>
<th>$p_e$</th>
<th>$N_m$</th>
<th>$ts_0$</th>
<th>$N_{sk}$</th>
<th>$L$</th>
<th>After cleaning</th>
<th>After saturation</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>$10^4$</td>
<td>4</td>
<td>349</td>
<td>384.2</td>
<td>81</td>
<td>172.1</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>128</td>
<td>221.0</td>
<td>65</td>
<td>173.8</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>108</td>
<td>237.0</td>
<td>47</td>
<td>175.5</td>
<td>65</td>
</tr>
<tr>
<td>0.02</td>
<td></td>
<td>3</td>
<td>237</td>
<td>397.8</td>
<td>58</td>
<td>207.3</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>102</td>
<td>213.8</td>
<td>38</td>
<td>128.8</td>
<td>47</td>
</tr>
<tr>
<td>0.01</td>
<td>$2\times10^4$</td>
<td>6</td>
<td>398</td>
<td>472.1</td>
<td>97</td>
<td>216.1</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>256</td>
<td>357.9</td>
<td>91</td>
<td>219.9</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>178</td>
<td>307.3</td>
<td>71</td>
<td>205.8</td>
<td>83</td>
</tr>
<tr>
<td>0.02</td>
<td>$10^4$</td>
<td>4</td>
<td>305</td>
<td>406.4</td>
<td>85</td>
<td>208.0</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>175</td>
<td>333.6</td>
<td>57</td>
<td>207.6</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>110</td>
<td>200.6</td>
<td>42</td>
<td>135.1</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>377</td>
<td>461.7</td>
<td>140</td>
<td>244.7</td>
<td>150</td>
</tr>
<tr>
<td>0.01</td>
<td>$4\times10^4$</td>
<td>12</td>
<td>281</td>
<td>399.4</td>
<td>110</td>
<td>244.7</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>195</td>
<td>339.0</td>
<td>95</td>
<td>245.0</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>289</td>
<td>446.0</td>
<td>89</td>
<td>262.4</td>
<td>115</td>
</tr>
<tr>
<td>0.02</td>
<td>$10^4$</td>
<td>8</td>
<td>157</td>
<td>364.9</td>
<td>67</td>
<td>272.9</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>94</td>
<td>323.6</td>
<td>50</td>
<td>283.0</td>
<td>67</td>
</tr>
</tbody>
</table>

Simulated map length was $L=191.3$ cM, 214.6 cM, and 273 cM for populations with $N_m=10^4$, $2\times10^4$, and $4\times10^4$ markers, respectively;

$p_e$ - level of genotyping errors per marker locus;

$ts$ - twin group size, $ts_0$ - threshold $ts$ value: skeletal markers should obey the condition $ts\geq ts_0$;

$N_{sk}$ - number of intervals in the map;

$N_m$ - number of markers per LG in the initial data set;

$\delta$ - map density (the ratio of the entire chromosome map length to the number of intervals).
## Table 2. Improvement of the skeletal map using the hybrid approach (*tw_cl*) for selecting candidate markers as compared to twin selection

<table>
<thead>
<tr>
<th>Method</th>
<th>$p_e$</th>
<th>$t_s$</th>
<th>$L_{t1}$</th>
<th>$N_{t1sk}$</th>
<th>$\delta_{t1}$</th>
<th>$L_e$</th>
<th>$N_{esk}$</th>
<th>$\delta_e$</th>
<th>$L_{t2}$</th>
<th>$N_{t2sk}$</th>
<th>$\delta_2$</th>
<th>$L$</th>
<th>$N_{sk}$</th>
<th>$\delta$</th>
<th>$n_e$</th>
<th>$n_r$</th>
<th>$m_c%$</th>
<th>$l_f%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$N_m=10^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>tw</em></td>
<td>0.01</td>
<td>8</td>
<td>57</td>
<td>11</td>
<td>5.2</td>
<td>38</td>
<td>34</td>
<td>1.1</td>
<td>79</td>
<td>19</td>
<td>4.2</td>
<td>174</td>
<td>64</td>
<td>2.7</td>
<td>2</td>
<td>3</td>
<td>0.96</td>
<td>57</td>
</tr>
<tr>
<td><em>tw_cl</em></td>
<td>0.01</td>
<td>8</td>
<td>61</td>
<td>20</td>
<td>3.1</td>
<td>41</td>
<td>39</td>
<td>1.0</td>
<td>82</td>
<td>25</td>
<td>3.3</td>
<td>184</td>
<td>84</td>
<td>2.2</td>
<td>2</td>
<td>8</td>
<td>0.98</td>
<td>47</td>
</tr>
<tr>
<td><em>tw</em></td>
<td>0.02</td>
<td>4</td>
<td>30</td>
<td>7</td>
<td>4.3</td>
<td>48</td>
<td>33</td>
<td>1.4</td>
<td>49</td>
<td>6</td>
<td>8.2</td>
<td>127</td>
<td>46</td>
<td>2.8</td>
<td>1</td>
<td>8</td>
<td>0.72</td>
<td>74</td>
</tr>
<tr>
<td><em>tw_cl</em></td>
<td>0.02</td>
<td>4</td>
<td>74</td>
<td>17</td>
<td>4.4</td>
<td>47</td>
<td>33</td>
<td>1.4</td>
<td>87</td>
<td>22</td>
<td>4.0</td>
<td>208</td>
<td>72</td>
<td>2.9</td>
<td>1</td>
<td>15</td>
<td>0.97</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>$N_m=2\times10^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>tw</em></td>
<td>0.02</td>
<td>6</td>
<td>58</td>
<td>12</td>
<td>4.8</td>
<td>49</td>
<td>36</td>
<td>1.4</td>
<td>26</td>
<td>7</td>
<td>3.7</td>
<td>133</td>
<td>55</td>
<td>2.4</td>
<td>6</td>
<td>4</td>
<td>0.72</td>
<td>72</td>
</tr>
<tr>
<td><em>tw_cl</em></td>
<td>0.02</td>
<td>6</td>
<td>62</td>
<td>14</td>
<td>4.4</td>
<td>48</td>
<td>34</td>
<td>1.4</td>
<td>89</td>
<td>20</td>
<td>4.5</td>
<td>199</td>
<td>68</td>
<td>2.9</td>
<td>5</td>
<td>6</td>
<td>0.98</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>$N_m=4\times10^4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>tw</em></td>
<td>0.02</td>
<td>10</td>
<td>87</td>
<td>15</td>
<td>5.8</td>
<td>57</td>
<td>38</td>
<td>1.5</td>
<td>108</td>
<td>13</td>
<td>8.3</td>
<td>252</td>
<td>66</td>
<td>3.8</td>
<td>16</td>
<td>5</td>
<td>0.92</td>
<td>71</td>
</tr>
<tr>
<td><em>tw_cl</em></td>
<td>0.02</td>
<td>10</td>
<td>92</td>
<td>21</td>
<td>4.4</td>
<td>74</td>
<td>37</td>
<td>2.0</td>
<td>106</td>
<td>25</td>
<td>4.2</td>
<td>255</td>
<td>83</td>
<td>3.1</td>
<td>2</td>
<td>6</td>
<td>0.98</td>
<td>64</td>
</tr>
</tbody>
</table>

$p_e$ - level of genotyping errors per marker locus; $N_m$ - number of markers per LG in the initial data set;

$t_s$ - twin group size, $t_{s0}$ - threshold $t_s$ value: skeletal markers should obey the condition $t_s\geq t_{s0}$;

in all *tw_cl* variants, the threshold kernel radius was $r=0.04$ (for description see File S3);

$L_s$, $L_{t1}$, $L_{t2}$ and $L_e$ - the estimated genetic lengths (in cM) of the entire chromosome map and its sub-telomeric and near-centromeric regions, while
$N_{sk}$, $N_{t1sk}$, $N_{t2sk}$, and $N_{csk}$ are the corresponding numbers of intervals of the entire map and its sub-telomeric and peri-centromeric regions;

$\delta, \delta_{t1}, \delta_{t2}$ and $\delta_{c}$ - map density (cM/interval) of the entire map and its sub-telomeric and peri-centromeric regions;

$n_e$ - the number of errors in the estimated order of markers compared to the simulated order;

$n_r$ - the number of "repeats" caused by fission of the initial TGs into sub-groups due to genotyping errors and missing marker scores;

$m_c$ (%) - map coverage, which represents the proportion of the constructed skeletal map length relative to the simulated map length;

$l_f$ (%) - loss factor, the percentage of lost (non-characterized) map unique positions in the constructed skeletal map compared to the simulated map; it is calculated as $l_f = 100 \frac{(N_{skf} - (N_{sk} - n_r))/N_{skf}}{100 \frac{(N_{skf} - (N_{sk} - n_r))/N_{skf}}{N_{sk}}}$, where $N_{skf}$ represents the number of intervals in the skeletal maps built for the simulated error-free data, while $N_{sk}$ and $(N_{sk} - n_r)$ represent the number of noisy markers in the skeletal map, non-corrected and corrected for the number of repeats, respectively.
Table 3. The effect of population size (n) on map characteristics

<table>
<thead>
<tr>
<th>$ms%$</th>
<th>$p_e$</th>
<th>$ts_0$</th>
<th>$r$</th>
<th>$L_{tl}$</th>
<th>$N_{t1sk}$</th>
<th>$\delta_{1}$</th>
<th>$L_{c}$</th>
<th>$N_{csk}$</th>
<th>$\delta_{c}$</th>
<th>$L_{t2}$</th>
<th>$N_{t2sk}$</th>
<th>$\delta_{2}$</th>
<th>$L$</th>
<th>$N_{sk}$</th>
<th>$\delta$</th>
<th>$n_e$</th>
<th>$n_r$</th>
<th>$m_c%$</th>
<th>$l_f%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Nm = 10^4$, $n = 200$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>67</td>
<td>41</td>
<td>1.6</td>
<td>41</td>
<td>64</td>
<td>0.6</td>
<td>82</td>
<td>39</td>
<td>2.1</td>
<td>191</td>
<td>144</td>
<td>1.3</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>6</td>
<td></td>
<td>58</td>
<td>18</td>
<td>3.2</td>
<td>39</td>
<td>40</td>
<td>1.0</td>
<td>78</td>
<td>18</td>
<td>4.3</td>
<td>175</td>
<td>76</td>
<td>2.3</td>
<td>3</td>
<td>7</td>
<td>99.5</td>
<td>52</td>
</tr>
<tr>
<td>$Nm = 10^4$, $n = 500$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>84</td>
<td>47</td>
<td>1.8</td>
<td>37</td>
<td>118</td>
<td>0.3</td>
<td>85</td>
<td>52</td>
<td>1.6</td>
<td>206</td>
<td>217</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>3</td>
<td></td>
<td>94</td>
<td>33</td>
<td>2.8</td>
<td>38</td>
<td>42</td>
<td>0.9</td>
<td>89</td>
<td>28</td>
<td>3.2</td>
<td>238</td>
<td>103</td>
<td>2.3</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>57</td>
</tr>
<tr>
<td>$Nm = 10^4$, $n = 10^3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>73</td>
<td>55</td>
<td>1.3</td>
<td>40</td>
<td>194</td>
<td>0.2</td>
<td>85</td>
<td>65</td>
<td>1.3</td>
<td>198</td>
<td>314</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>2</td>
<td></td>
<td>86</td>
<td>27</td>
<td>3.2</td>
<td>85</td>
<td>66</td>
<td>1.3</td>
<td>107</td>
<td>38</td>
<td>2.8</td>
<td>278</td>
<td>131</td>
<td>2.1</td>
<td>2</td>
<td>12</td>
<td>100</td>
<td>62</td>
</tr>
</tbody>
</table>

$p_e$ - level of genotyping errors per marker; $ms$ - simulated rate of missing data per marker; $N_m$ - number of markers per LG in the initial data set; $ts$ - twin group size, $ts_0$ - threshold $ts$ value: skeletal markers should obey the condition $ts \geq ts_0$; $r$ - kernel radius; $L$, $L_{tl}$, $L_{t2}$ and $L_{c}$ - the estimated genetic lengths (in cM) of the entire chromosome map and its sub-telomeric and near-centromeric regions, while $N_{sk}$, $N_{t1sk}$, $N_{t2sk}$, and $N_{csk}$ are the corresponding numbers of intervals of the entire map and its sub-telomeric and peri-centromeric regions; $\delta$, $\delta_{t1}$, $\delta_{t2}$ and $\delta_{c}$ - map density (cM/interval) of the entire map and its sub-telomeric and peri-centromeric regions; $n_e$ - the number of errors in the estimated order of markers compared to the simulated order; $n_r$ - the number of "repeats" caused by fission of the initial TGs into sub-groups due to genotyping errors and missing marker scores; $m_c$ (%) - map coverage, which represents the proportion of the constructed skeletal map length relative to the simulated map length;
$l_f(\%)$ - loss factor, the percentage of lost (non-characterized) map unique positions in the constructed skeletal map compared to the simulated map; it is calculated as 

$$l_f = 100 \left( \frac{N_{skef} - (N_{sk} - n_r)}{N_{skef}} \right) = 100 \left( \frac{N_{skef} - N_{sk} + n_r}{N_{skef}} \right)$$

where $N_{skef}$ represents the number of intervals in the skeletal maps built for the simulated error-free data, while $N_{sk}$ and $(N_{sk} - n_r)$ represent the number of noisy markers in the skeletal map, non-corrected and corrected for the number of repeats, respectively.
Table 4. Comparison of genetic maps constructed using the MST and twin approach implemented in the MultiPoint-ULD (MULD) software

<table>
<thead>
<tr>
<th>Chr</th>
<th>$N_m$</th>
<th>all</th>
<th>singletons</th>
<th>MST bins $L_cM$</th>
<th>MULD $L_cM$</th>
<th>$N_{sk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A</td>
<td>862</td>
<td>217</td>
<td>125</td>
<td>813.9</td>
<td>181.7</td>
<td>60</td>
</tr>
<tr>
<td>2B</td>
<td>1674</td>
<td>317</td>
<td>176</td>
<td>795.4</td>
<td>190.0</td>
<td>81</td>
</tr>
<tr>
<td>5A</td>
<td>1556</td>
<td>296</td>
<td>167</td>
<td>787.7</td>
<td>156.7</td>
<td>67</td>
</tr>
<tr>
<td>7A</td>
<td>930</td>
<td>269</td>
<td>151</td>
<td>633.8</td>
<td>225.7</td>
<td>78</td>
</tr>
</tbody>
</table>

$L$ - map length using Kosambi metric; $N_{sk}$ number of intervals of the map

$N_m$ - number of markers per LG in the initial data set

Table 5. Comparison of the proposed approach implemented in MultiPoint-ULD (MULD) software with MST on a simulated double haploid population (of size 200 with 2000 markers per chromosome)

<table>
<thead>
<tr>
<th>$ms%$</th>
<th>0</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_e$</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>$L_{cM}$ bins</td>
<td>180</td>
<td>294</td>
</tr>
<tr>
<td>$L_{cM}$</td>
<td>134</td>
<td>134</td>
</tr>
<tr>
<td>$N_{sk}$</td>
<td>81</td>
<td>75</td>
</tr>
<tr>
<td>$n_r$</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$l_f %$</td>
<td>4.8</td>
<td>14.3</td>
</tr>
</tbody>
</table>

$p_e$ - level of genotyping errors per marker; $ms$ - simulated rate of missing data per marker; $L_{cM}$ - map length (in cM) of a chromosome or LG;

$l_f (\%)$ - loss factor, which represents the percentage of lost (non-characterized) map unique positions in the constructed skeletal map compared to the simulated map; it is calculated as $l_f = 100 \frac{(N_{skef} - (N_{sk} - n_r))}{N_{skef}}$, where $N_{skef}$ is the number of intervals in the map built for the simulated error-free data, while $N_{sk}$ and $(N_{sk} - n_r)$ are the number of noisy markers in the skeletal map, non-corrected and corrected for the number of repeats, respectively; $n_r$ - the number of "repeats" resulting from fission of the initial groups of co-segregating markers into sub-groups due to genotyping errors and missing marker scores; such repeats will appear in the constructed map at separate (usually, but not necessarily, adjacent) positions.
Figure legends

Fig. 1. Selecting candidate markers for the skeletal map in the presence of strong regional heterogeneity of recombination, genotyping errors and missing data:
(a) - marker positions along the chromosome in an error-free situation. The continuous black horizontal bars are representing groups of co-segregating markers (twins) unresolvable by recombination due to low local recombination rate and small population size. In the simulated example, much higher marker density is shown for the peri-centromeric (C) compared to subtelomeric regions (T1 and T2); (b) - disturbed distribution of markers due to genotyping errors that lead to disruption of some twin groups; markers with genotyping errors (shown in grey) may cause map length inflation if included in the map; (c) - skeletal map based on representatives of twin groups remained despite the "losses" of markers caused by genotyping errors; (d) - recovering a part of dissipated twin groups by cluster analysis.

Fig. 2. The scheme of the proposed strategy for building ultra-dense genetic maps. The major difference from other approaches and our previous algorithm is the proposed pre-mapping filtering (Filter 2) based on twins and clusters of tightly linked markers.

Fig. 3. Illustration of complicating factors causing the need for introduced parameters \( (n_e, n_r, m_c) \) as characteristics of map quality for simulated data: errors and repeats as indicated by thin and thick lines, respectively (here, \( n_e = 12 \) and \( n_r = 3 \)).
(a) A fragment of a simulated map with nine ordered markers each representing a twin group.
(b) The simulated noise (missing datapoints and genotyping errors) cause the following changes: (i) some twin groups undergo week or zero disturbance implying stability of the corresponding skeletal markers (markers #1 and #5); (ii) for some markers (#2 and #3) the disturbance caused a slight change in the order; (iii) the noise caused disintegration of twin groups (represented by markers #7 and #8 in the simulated map) into sub-groups with sizes
fitting the condition $t_s \geq t_{s0}$ for selection of candidate skeletal markers; the generated repeat markers contributed to the number of errors in the map; (iv) disintegration of twin groups in such a way that none of the resulting sub-group obey the condition $t_s \geq t_{s0}$ (markers #4 and #6); (c) Simulated genetic map (with no disturbances); $L_1$ and $L_2$ represent the distances (in cM) between markers $m_i$ and $m_j$, and between $m_1$ and $m_n$, respectively. (d) Constructed genetic map, where the first position is represented by a marker that in the simulated map was at the $i^{th}$ position (marker $m_i$), while the last position is represented by a marker that in the simulated map was at the $j^{th}$ position (marker $m_j$). Map coverage for a map constructed either for error-free or noisy simulated data is calculated as $m_c(\%) = 100L_1/L_2$.

Figures

Fig.1