1

CNVmap: a method and software to detect and map copy number variants from segregation data

4

5

6 Matthieu Falque [1]*, Kamel Jebreen [1,2], Etienne Paux [3], Carsten Knaak [4], Sofiane Mezmouk [4], and

7 Olivier C. Martin [1]

8

9 * Corresponding author: matthieu.falque@inra.fr

10 [1] GQE - Le Moulon, Université Paris-Saclay, INRA, CNRS, AgroParisTech, 91190 Gif-sur-Yvette,

11 France

12 [2] Department of Mathematics, An-Najah National University, Nablus, Palestine

13 [3] GDEC, INRA, Université Clermont Auvergne, 63000 Clermont-Ferrand, France

14 [4] KWS SAAT SE & Co.KGaA, Grimsehlstr. 31, 37574 Einbeck, Germany

15

16

17

Abstract

Single nucleotide polymorphisms (SNPs) are widely used for detecting quantitative trait loci or for searching for causal variants of diseases. Nevertheless, *structural* variations such as copy-number variants (CNVs) represent a large part of natural genetic diversity and contribute significantly to trait variation. Over the past decade, numerous methods and softwares have been developed to detect CNVs. Such approaches are based on exploiting sequencing data or SNP arrays, but they bypass a wealth of information such as genotyping data from segregating populations, produced *e.g.* for QTL mapping. Here we propose an original method to both detect and *genetically map* CNVs using mapping panels. Specifically, we exploit the apparent heterozygous state of duplicated loci: peaks in appropriately defined genome-wide allelic profiles provide highly specific signatures that identify the nature and position of the CNVs. Our original method and software can detect and map automatically up to 33 different predefined types of CNVs based on segregation data only. We validate this approach on simulated and experimental bi-parental mapping panels in two maize and one wheat populations. Most of the events found correspond to having just one extra copy in one of the parental lines but the corresponding allelic value can be that of either parent. We also find cases with two or more additional copies, especially in wheat where these copies locate to homeologues. More generally, our computational tool can be used to give additional value, at no cost, to many datasets produced over the past decade from genetic mapping panels.

Keywords: Copy number variation (CNV); Segregating populations; Allele frequency profiles; Non-Mendelian markers

Introduction

Single Nucleotide Polymorphisms (SNPs) are typically exploited *via* genotyping technologies, such as arrays or Genotyping by Sequencing, leading to high-density information on such polymorphisms. The wide availability of such tools explains why polymorphisms are principally characterized at this SNP level, even though it is known in many species that there is also a great deal of *structural polymorphism* across genomes. Generally one uses the terminology "structural variation" (SV) when there are inversions, translocations, insertions, deletions, or duplications involving segments of over 1000 base pairs. Identifying such SVs is a current challenge, rendered difficult by the so-called "complexity" of large genomes that involve many repetitive sequences, generally associated with transposable elements. Work on identifying SVs includes in particular searching for Copy Number Variations (CNVs) (Redon *et al.* 2006) and Presence/Absence Variations (PAVs) (Alkan *et al.* 2011). In CNVs, a gene or segment of DNA is present in different numbers of copies in two genomes. CNVs have been discovered in numerous species, and in particular in mammals (Guryev *et al.* 2008; Conrad *et al.* 2010). In PAVs, a gene or DNA segment is present in one genome and missing in the other. It has been claimed that more nucleotide bases are affected by SVs than by SNPs (Zhang *et al.* 2009). Both CNVs and PAVs are associated with phenotypic variations and diseases (Beckmann *et al.* 2007; Zhang *et al.* 2009), making them a focus of much current research. A large number of approaches have been used to detect structural variations. Perhaps the oldest is based on competitive hybridizations to oligonucleotide probes as occurs in comparative genomic hybridization (CGH) arrays (Beló *et al.* 2009; Springer *et al.* 2009). A somewhat different technique is used in SNP arrays where individual samples are genotyped (no competition between samples). In such arrays, a fluorescence intensity is associated with each allele of a SNP; by following these intensities for successive SNPs along the genome, it is

3

61 possible to identify regions in which the signal is anomalously low for all alleles, indicative of a PAV,

62 or in which some intensities are a few fold higher than expected, indicative of a CNV (Colella *et al.*

63 2007; Cooper *et al.* 2008). Later approaches involve exploiting sequencing read depth (Bailey *et al.*

64 2002; Yoon *et al.* 2009; Alkan *et al.* 2009), unexpected mappings of read pairs (Chen *et al.* 2009; Pang

65 *et al.* 2010), split reads (Mills *et al.* 2006; Ye *et al.* 2009), and sequence-based reassembly (Zerbino and

66 Birney 2008; Chaisson *et al.* 2009; Simpson *et al.* 2009; Li *et al.* 2010). Each approach has its

67 advantages (Alkan *et al.* 2009; Sudmant *et al.* 2010) and continues to undergo optimization (Zare *et al.*

68 2017). Our work takes a novel approach, different from all that we listed above, and gives "a second

69 life" to SNP genotyping technologies in the specific context of populations in segregation. Indeed, SNP

70 arrays have been available at low cost for several years now and so have been used rather extensively

71 to genotype segregating populations. Such populations are typically constructed for mapping

72 quantitative trait loci (QTLs) for fundamental research or for breeding programs. Interestingly, such

73 technologies provide, in the context of segregating populations, information on structural variation

74 between the founding parents used to build the population. However, that information is hidden within

75 the markers showing non-Mendelian segregation patterns, markers that generally are discarded early-on

76 in the linkage mapping analyses. The present work provides a methodology for inferring certain types

77 of SVs based on those non-Mendelian markers in bi-parental mapping populations. We will stress the

78 CNV cases, because in such situations the previously unknown copies of the region can also be

79 *mapped* based on the linkage disequilibrium analyzed from the segregation data. Indeed, our approach

80 exploits particular *profiles* of allele frequencies arising along the genome, somewhat analogously to

81 what is done in genome-wide association studies. However, in our case, instead of working with the

82 genotype at a single marker at a time, we work with compound genotypes involving multiple markers.

83 The profiles built in this way exhibit peaks at the loci where the extra copies arise and provide

4

84 signatures allowing the identification of the type of SV involved. These additional loci should not be

85 too close in genetic distance to the reference locus because, as we shall see, it is the recombinations

86 between these loci in the population that lead to the characteristic signal.

87 Any marker that deviates strongly from Mendelian behavior in a segregating population points to a

88 potential SV. The simplest context in which to understand how to exploit a non-Mendelian signal is to

89 consider the case of doubled-haploid (DH) plant populations in segregation. Indeed, in that type of

90 population where homozygous individuals are derived from single meiotic products by chromosome

91 doubling, all of the Mendelian markers should show full homozygosity while SVs and more

92 specifically CNVs will lead some individuals to appear as being heterozygous at markers involved in

93 rearrangements. Therefore, we begin our study here using a doubled haploid (DH) population

94 consisting of 625 maize individuals; we refer to it as the GABI population. We will also show how our

95 approach can be used in the context of RIL (Recombinant Inbred Line) and IRIL (Intermated

96 Recombinant Inbred Line; (Lee *et al.* 2002)) populations. The maize IRIL population studied in this

97 paper is the IBM (Ganal *et al.* 2011) mapping population. We also study a RIL wheat population

98 (Choulet *et al.* 2014; Rimbert *et al.* 2018), hereafter referred to as WHEAT. The markers in these RIL

99 and IRIL populations have residual heterozygosity because the number of generations of selfing used is

100 not so large. Such blocks of markers with residual heterozygosity might be expected to swamp any SV

101 signal; interestingly this turns out not to be true. We will (1) demonstrate the efficiency of our approach

102 on our three populations that include two situations with this potential problem, (2) show that the allele

103 frequency profiles are fully compatible with the inferred CNVs by comparing those profiles to the ones

104 produced from simulated populations, and (3) compute p-values associated with the H_0 hypothesis that

105 the observed profiles occurred by chance in the absence of SV. As a first illustration, we apply our

106 method and software to the GABI maize DH population, revealing striking signals of duplications and

5

107 triplications, the corresponding copies arising on a *common* chromosome not more often than by

108 chance. Then we examine the case of recombinant inbred lines (RILs or IRILs); as expected, having

109 residual heterozygosity makes the problem more challenging but our method generalizes well. Indeed,

110 in the IBM and WHEAT experimental populations, we are able to identify unambiguous signatures of

111 copy number variations. Interestingly, in the WHEAT data set we find a large number of triplicated loci

112 that involve the homeologous chromosomes. Finally, we assess the candidate CNVs in the IBM

113 population using reference genome sequences of the parents.

114

Materials and Methods

116 *Aim and design of the study*

117 The locus-specificity of genetic markers used for genotyping, be they PCR-based or array-based,

118 mostly relies on oligonucleotides hybridizing to sequences flanking the SNP of interest. In the case of

119 duplicated regions, such oligonucleotides may find alternative targets, messing up the interpretation of

120 the observed raw data (mostly fluorescence intensities at two wavelengths), which usually assumes a

121 single target locus. This results in apparent non-Mendelian behaviours of some markers, which are

122 usually filtered out from data sets before using them for mapping or QTL analyses. The goal of this

123 work was to exploit these types of data sets to infer markers involved in SVs and genetically map the

124 previously unknown copies, providing a software package to do so automatically. This software is

125 provided as a R package named CNVmap, provided in Supplementary File S1 (see Supplementary

126 File S6 for installation). More detailed explanations about the functionalities and procedures of this

127 software are provided in the package *via* the embedded documentation of the associated functions.

6

128   In our approach, we focus on non-Mendelian markers to provide and test appropriate hypotheses,

129   interpreting the observed segregations as CNV events polymorphic between the parents used to

130   generate segregating populations. The main signature of such events being apparent heterozygous calls,

131   we therefore worked with segregating populations in which the individuals are almost fully

132   homozygous, like doubled-haploid or recombinant inbred populations. In such populations, high levels

133   of heterozygous calls in some markers strongly point to possible CNVs containing those markers.

134   Because some systematic or random errors in the genotyping process can also lead to unexpected

135   heterozygotes, we extended our approach to a joint analysis of each candidate marker with its local

136   allelic context, which provides multiple and unambiguous signatures that make up strong evidence for

137   the reality of the event. Moreover, this approach (implemented in our software) also provides the

138   genomic localization of the other loci involved in the CNV.

139   *Populations and segregation data used*

140   The population mainly used in this study is a Doubled Haploid (DH) maize population called GABI

141   (Presterl *et al.* 2007), from which genotyping data were kindly provided by KWS SAAT SE &

142   Co.KGaA (Einbeck, Germany). The GABI population contains 625 DH lines, which were genotyped

143   using the Illumina MaizeSNP50 array (Ganal *et al.* 2011). Second, we also studied the maize IBM

144   population (Ganal *et al.* 2011) which is is an Intermated Recombinant Inbred Line (IRIL) population

145   obtained by intermating F2 individuals for four generations to accumulate recombination, before

146   beginning the selfing generations used to fix the material (Lee *et al.* 2002). The IBM population

147   contains 239 RILs, which were genotyped using the same SNP array and cluster file as for GABI.

148   Lastly, we studied a wheat population consisting of, 406 F6 individuals derived from a cross between

149   Chinese Spring and Renan; these were genotyped using the TaBW280K SNP array (Rimbert *et al.*

7

150   2018). Marker segregation data for populations GABI, IBM, and WHEAT are respectively provided in

151   Supplementary Files S2, S3, S4.

152   *Linkage map construction*

153   Our method and software to detect CNVs requires prior knowledge of the order and genetic position of

154   the markers. The genetic maps used to analyze the GABI and IBM segregation data were thus first

155   obtained using a seriation approach implemented in R scripts calling functions from the CartaGene

156   software (de Givry *et al.* 2004), as described in Ganal et al. (Ganal *et al.* 2011). The genetic map

157   used to analyze the segregation data of the WHEAT population were produced using the software

158   MSTmap (Wu *et al.* 2008) with the following default parameters: population type: RIL6; distance

159   function: Kosambi; cut-off: 0.00000000001; map dist.: 15; map size: 2; missing threshold: 0.20;

160   estimation before clustering: yes; detect bad data: yes; objective function: ML (Rimbert *et al.* 2018).

161   Linkage map data for populations GABI, IBM, and WHEAT are respectively provided in

162   Supplementary Files S2, S3, and S4.

163   *Raw data filtering*

164   In DH populations, each normal marker should be homozygous in every offspring. The possible calls

165   for any marker are then "A" (the allele attributed to parent 1) or "B" (the allele attributed to parent 2). It

166   is important to keep in mind that a call of a SNP is the result of an elaborate identification process

167   which is not 100% reliable so that one cannot exclude a low proportion of errors, leading for instance

168   to some small proportion of "H" calls (heterozygous, which should almost never happen in a DH

169   population if there is no genotyping error) or "-" for missing data if the calling is too ambiguous. In

170   practice, we do see heterozygous markers in spite of each individual being homozygous. Such cases

8

171   might be erroneous calls or indicative of regions belonging to SVs. Beyond possible errors in the calls

172   for certain markers, some individuals of the population may themselves be corrupt (e.g. through pollen

173   contamination). A strong indication of this is if an offspring has an anomalously large number of

174   markers that are called "H". We thus apply some quality filtering on the data sets, both at the level of

175   the individuals (e.g. we cast out individuals having too high heterozygozity rates) and at the level of

176   markers. In the R code, the user can change the thresholds for such filterings, for instance based on

177   minor allele frequency or genotype frequencies (see all parameter descriptions in the R package).

178   *Defining the Mendelian and candidate markers*

179   Given the markers passing the previous test, we now divide them into three classes: "Mendelian",

180   "candidate", and "other". Our procedure for defining the first or second class of markers is based on

181   forcing them to be respectively "typical" and "atypical" for some statistic while the "other" markers are

182   all the ones that do not pass these tests. Our first statistic is the fraction of individuals that are

183   heterozygous. We require that this fraction be in the bottom X_H percentile for "Mendelian" and in the

184   top Y_H for "candidate" markers. We do the same thing (but with different thresholds) using the fraction

185   of individuals that are called missing for that marker. Our use of percentiles has the advantage that it

186   automatically takes into account the properties of the population, such as the low numbers of

187   heterozygotes in DH populations and the significantly larger numbers arising in RILs that have not

188   reached fixation. Clearly, as the threshold Y_H is lowered, the number of candidate markers will increase

189   so if one wants to find as many events as possible pointing to SVs it is good to not take Y_H too large. In

190   contrast, the potential number of Mendelian markers is quite substantial so it is not a problem to be

191   rather stringent for the value of X_H.

9

192   *Automatic detection of peaks in the allele profiles*

193   For each candidate marker M* we identify its flanking Mendelian markers M_L and M_R from which we

194   identify the individuals in the population belonging to each of six associated 3-marker genotype classes

195   (see Supplementary File S5). Then for each of these six classes we compute the corresponding genome-

196   wide allele profile using the *Mendelian* markers only, each marker leading to a frequency defined as the

197   number of individuals carrying the A allele divided by the number of individuals carrying either the A

198   or B alleles. These six genome-wide allele profiles along chromosomes are then analyzed for

199   occurrences of peaks. Roughly a peak can be defined *via* a region on the genome in which the allele

200   frequency curve has a pointed shape and approaches very close to 0 or 1. In practice, to avoid being

201   sensitive to noisy or erroneous data, we get rid of outliers by a first filter. That means producing a first

202   smoothed version of the allele curve using splines (*smooth.spline()* function in R) and throwing out the

203   data points that are outliers with respect to that curve. Second, a new smoothed curve is generated

204   using the remaining markers. Then all regions for which this new smoothed curve is close enough to 0

205   or 1 are identified. A linear fit of the data (outliers excluded) is performed on each side of the putative

206   peak to determine its expected position, and also to assess the quality and slope of the linear regression

207   on both sides of the peak (or on one side only if the peak is at the extremity of a chromosome, or close

208   to another peak). Then the list of all peaks for all six classes are compared to see whether peaks co-

209   localize. This leads to a list of peaks (genetic positions on the genome) with each peak being called as

210   "present" or "absent" for each of the six allele frequency curves. Note that if a class contains no

211   individuals it is just ignored (see Supplementary File S5). Furthermore, when there are few individuals

212   in a class, the associated allele frequency curves are noisy and thus will have peaks by chance. We thus

213   only consider classes having a minimum number of individuals, this minimum being determined so that

10

214   by the Bonferroni test one has a false discovery rate for peak detection that is 5% under the hypothesis

215   that there is no structural variation present.

216   *Automatic assignment to a type of CNV event*

217   Once all peaks have been detected, for the associated locus the presence or absence of peaks or troughs

218   for the list of 6 different 3-marker genotype classes of individuals was encoded with a 6-character

219   string. The list of these strings (one per locus) provided the observed signature of the event. This

220   signature was then compared by the software with a list of 33 predefined signatures (details provided as

221   Supplementary File S5), and in case of a match between the observed signature and a predefined one,

222   then the event was assigned to the corresponding type. The predefined signatures were based on

223   theoretically expected patterns arising from CNVs involving additional copies at one or two loci. Such

224   signatures depend on the allelic content at these different loci, leading us to introduce below a

225   schematic notation for CNV events.

226   *Nomenclature used for the different types of CNVs*

227   In the following, a CNV involving in parent 1 X doses of the genomic region of interest and Y doses in

228   parent 2 will be referred to as a "X:Y CNV" (X and Y being equal to 1, 2, or 3). Moreover, each CNV

229   category is encoded as a string of 2 to 3 groups of 3 characters, there being one group per locus, each

230   separated by an underscore. Each group contains the parental alleles separated by a slash, so the result

231   takes the following form: A group of 3 characters specifies the alleles carried at the considered locus by

232   parents 1 and 2, in that order, separated by a slash, A being the reference allele of M* in parent 1 and B

233   being the reference allele of M* in parent 2. The different groups are further concatenated using the

234   underscore as a separator for the successive loci: locus1(P1/P2)_locus2(P1/P2)_locus3(P1/P2). The

11

235   first group is always encoded "A/B" and indicates the reference locus, located at the position where the

236   candidate marker was initially mapped. Further groups indicate the different additional loci carrying

237   copies of the region targeted by the candidate marker. So for instance for a 2:1 CNV of type A/B_B/-

238   the Parent 1 has two copies of the considered genomic region but the copy at the second locus carries

239   the allele B, while the Parent 2 has only one copy.

240   *Analyzing candidates based on missing data*

241   Our method is based on the detection of candidate markers for which the number of H calls is

242   anomalously high, followed by an analysis of each associated genome-wide allelic profile. However,

243   when for completeness we tried to analyze allelic profiles for *all* markers, we discovered clear CNV-

244   like signatures for some markers with little or no H calls but with large numbers of missing data. In

245   such cases, instead of the AHA or BHB 3-marker genotypic class, the peaks were observed on the

246   allelic profiles associated with A-A or B-B classes suggesting that one had a CNV but where, for

247   unknown reasons, the H calls for the candidate marker were transformed into missing data calls. So we

248   specified in the software the signatures that would arise from having H calls be erroneously modified

249   and denoted them by adding a suffix to their putative CNV type. The suffix was "|Hm" when (part of)

250   the expected H calls have been turned into missing data calls (with probability p_{Hm}), and "|HmHa"

251   respectively "|HmHb" when the expected H calls have been turned partly into missing data calls (with

252   probability p_{Hm}) and partly into A respectively B homozygote calls (with probabilities p_{Ha} respectively

253   p_{Hb}).

254   To simulate such events, we first estimated the probabilities p_{Hm}, p_{Ha}, and p_{Hb} from the data for the

255   marker considered based on the allelic profiles at the inferred peaks. We then simulated the CNV event

256   as explained below. Finally, we introduced systematic "errors" to the resulting candidate marker

12

257   genotype depending on the type of the putative CNV event with its suffix. Specifically, we randomly

258   transformed the H calls into missing data, A, or B calls according to the probabilities p_{Hm}, p_{Ha} and p_{Hb}.

259   *Producing simulated datasets*

260   We produced simulated data staying as close as possible to the experimental population parameters,

261   keeping the same marker positions on the genetic maps of each chromosome and the same population

262   size. We simulated the exact same scheme of crossings as the one used to produce the experimental

263   populations, implementing *in silico* crossovers that can arise in each marker interval during each

264   meiosis based on the experimental genetic map. Crossover interference was also implemented using the

265   Gamma model (McPeek and Speed 1995) whose parameter *nu* can be set as a parameter in our

266   software (for typical values of *nu* in maize, see (Falque *et al.* 2009)). This implementation of

267   interference proved to be important for having comparable peak width between experimental and

268   simulated profiles. To simulate any particular CNV hypothesis, we implemented into the parental

269   genomes the associated duplications or triplications of the marker M* of interest, using positions of loci

270   inferred from the analysis of the actual experimental population. The corresponding modified parental

271   genomes thus had extra fictitious markers each tagged with the parental allelic value (and thus

272   independent of the CNV hypothesis). For instance in the case of a duplication in Parent A but with

273   opposite allele (CNV of the type A/B_B/-), the extra marker had nevertheless allele A in parent A and

274   allele B in parent B. Then the scheme of crossings was simulated based on these modified parental

275   genomes (note that the genetic map was also modified but just by the inclusion of the extra markers at

276   their inferred positions). Lastly, the individuals in the resulting population were "genotyped" *in silico*.

277   For the markers that were not involved in the CNV, this was straightforward. However, to genotype an

278   individual for the marker M*, it was necessary to take into account allelic values not only at M* but

13

279 also at the extra copy or copies of the marker, to mimic the fact that oligonucleotides used for

280 genotyping M* would hybridize on all copies. This was where the actual CNV hypothesis intervened

281 because the "raw" genotypes at each extra marker as produced by the simulation had to be reinterpreted

282 using the CNV allelic content. Specifically, the call of the marker M* had to be changed to H if and

283 only if the reference locus was not already H and both A and B alleles were present in the reinterpreted

284 individual when considering M* and all of its copies. As an illustration of this rule, consider again the

285 CNV of the type A/B_B/-. The only situation requiring that a call of M* be changed to H is when the

286 raw genotype is A at the first locus and also A at the second. In practice we apply such "transformation"

287 rules using successively each of the extra copies of the marker, each time testing whether the

288 genotyping should be changed to H. Once that is done, the extra copies are removed from the data set

289 and only the original markers and associated modified calls are used as input to the analysis program,

290 leading to production of corresponding genome-wide allelic profiles based on these simulated data sets.

291 A good agreement between profiles produced from the experimental and simulated data sets then

292 provides strong support for the hypothesized CNV.

293 *Calculation of* p-*values associated with the hypothesis H$_0$ of no structural variation*

294 Although having the simulated profiles allows one to get a feeling for whether a proposed CNV is

295 plausible through consistency between theory and experiment, it is appropriate to also compare to the

296 null hypothesis H$_0$ whereby there is no CNV and the marker M* is present in only one copy in both

297 parents. Under that hypothesis, the additional peaks in the experimental allelic profiles are simply due

298 to stochasticity in the segregation, a situation that will be a problem whenever relatively few

299 individuals contribute to these profiles. The CNVmap package provides a test of H$_0$ in the form of a *p-*

300 value that is computed as follows. Let M* be the considered marker that is a candidate for belonging to

14

301   a region involved in a CNV. In our first step we identify within the whole population two sub-classes of

302   individuals: the ones for which the flanking (Mendelian) markers of M* are both called as A alleles,

303   and the ones for which those markers are both called as B alleles. Not all individuals fall within one of

304   these classes, so for instance if for an individual one flanking marker is heterozygous, or if one is A and

305   the other is B, then the individual is not further considered. Within each sub-class, the errors

306   (heterozygous and/or missing data calls) under H_0 are random. Thus the second step of our procedure is

307   to produce a simulated dataset by shuffling the calls of M* separately in each of the two sub-classes of

308   individuals defined in the first step. Under H_1 (presence of a CNV) the M* calls are correlated with the

309   calls at the second locus, while under H_0 (M* is single-locus in both parents) there is no such second

310   locus. The third step is to apply our analysis pipeline to this shuffled dataset and identify the peaks in

311   the allelic profiles. The second and third step are repeated a large number of times (this number is

312   specified by the user and computed *via* parallelization). Lastly, the *p*-value for rejecting the hypothesis

313   H_0 is obtained from the fraction of the shufflings that lead to having additional peaks in the allelic

314   profiles.

315   *Use of parental genome sequences for validating CNVs predicted from the IBM population*

316   To provide independent validation of CNVs detected with our software, we examined the whole-

317   genome sequence assemblies of the two parents (B73 and Mo17) used to produced the IBM population.

318   First, for each non-Mendelian marker M* identified with our software as being located in a 1:2 or 2:1

319   CNV, we extracted from the B73 sequence *three* regions 201 bp long (100bp before and 100bp after the

320   SNP) flanking not only the marker M* (indicating the reference locus) but also each of the two markers

321   (M_{left_peak} and M_{right_peak}) delimiting the second locus (identified automatically in our software *via* the

322   corresponding fitted peak positions). To do that, we used the V2 version of the B73 genome assembly

15

323   (AGPv2 RefGen_v2 https://www.maizegdb.org/genome/genome_assembly/B73%20RefGen_v2)

324   because the physical coordinates of the MaizeSNP50 SNPs are given on that V2 version (Ganal *et al.*

325   2011). Then, for our CNV validation, we BLASTed these three 201bp sequences against the B73

326   AGPv4 RefGen_v4 Maize genome assembly (the most recent available assembly

327   https://www.maizegdb.org/genome/genome_assembly/Zm-B73-REFERENCE-GRAMENE-4.0) using

328   default parameters. We then considered that the presence of a second copy in B73 was validated if one

329   of the high-scoring pairs (HSP) returned by BLAST for the M* flanking region was on the same

330   chromosome as the second locus and was included in the confidence interval of that locus (based on

331   coordinates of HSPs obtained when BLASTing M_{left_peak} and M_{right_peak} flanking regions). Presence of

332   the second locus in B73 is expected in 2:1 CNVs but not in 1:2 CNVs. We proceeded similarly for

333   testing the presence of both loci in the Mo17 parent, except that we first extracted the three 201bp

334   regions of Mo17 corresponding to M*, M_{left_peak} and M_{right_peak} markers by BLASTing the 201bp B73

335   regions against the Mo17 genome (https://www.maizegdb.org/genome/genome_assembly/Zm-Mo17-

336   REFERENCE-CAU-1.0). We then BLASTed those 201bp Mo17 sequences against the Mo17 genome

337   assembly. The second locus is then expected to be present in Mo17 in 1:2 CNVs but not in 2:1 CNVs.

338   *Availability of data and material*:

339   All data generated or analysed during this study are included in this published article and its

340   supplementary information files (software provided as Supplementary File S1 and data sets provided as

341   archives in Supplementary Files S2, S3, and S4). All Supplementary Figures, Tables, and Files have

342   been uploaded to FigShare.

343

16

344

Results

346 *Genome-wide allele frequency profiles identify the loci involved in CNVs*

347 *Strikingly clean signatures for 1:2 or 2:1 CNVs*

348 What should be expected in a segregating population if only one of the parents has a marker

349 duplicated? The simplest situation is schematically represented in Fig. 1A where in parent 1 (with

350 alleles denoted "A") a DNA segment carrying the SNP has been duplicated producing an insertion in

351 some other place in the genome. The marker involved in this duplication is labeled M* and can be

352 thought of as having been identified as a "candidate" marker given its non-Mendelian behavior in terms

353 of heterozygote calls (see Materials and Methods) while M_L and M_R correspond to its flanking

354 Mendelian markers that are thus *not* part of the duplication. The region where the M* locus was

355 initially mapped will hereafter be referred to as the "reference locus". In Fig. 1A we assume that only

356 Parent 1 carries the duplication and this duplicate copy has the allele of that same parent. For the

357 purposes of the figure, we only represent M* in this duplication but other markers can very well be

358 implicated too and if this is so one has even more evidence that there is a CNV. After crossing these

359 two homozygous parents to produce an F1 individual, meiosis of the F1 leads to gametes that may

360 shuffle the alleles of the parental chromosomes. In the case of a DH population, these gametes are used

361 to produce diploid plants whose genomic content is that of a gamete but simply doubled. For the

362 situation depicted in Fig. 1A where we focus on the reference locus and the duplication, the (gametic or

363 DH) associated segregation patterns fall into 4 categories. Assuming that these two loci are on different

364 chromosomes (or far enough away from each other on the same chromosome), the genotyping of these

17

365 plants will generate a call for M* that will be "A" 50% of the time, "B" 25% of the time and "H" 25%

366 of the time. Thus the marker will be detected as anomalous (non-Mendelian) in this mapping

367 population, having too many "H" calls, and this is the simplest situation for which our method allows

368 one to map the second locus. As indicated in Fig. 1A, we introduce the associated 3-marker genotype

369 classes based on the alleles arising for the M_L, M*, and M_R markers. The CNV situation depicted in

370 Fig. 1A will lead to a characteristic signature when considering the genome-wide allele frequency

371 profiles. To illustrate this, we simulated a DH population with the same characteristics as GABI, taking

372 a marker M* from chromosome 4 (specifically marker PZA-000492026) and then we duplicated it onto

373 chromosome 5. The resulting allele frequency profiles are displayed in Fig. 1B. To construct the

374 profiles, we first assigned the individuals of the simulated population to one of the 6 classes defined via

375 the 3-marker genotypes M_L M* M_R. There are 6 classes because M* can be A, B or H while we impose

376 M_L and M_R to be of the same parental type because in practice these markers are very close on the

377 chromosome (because of that proximity, almost all individuals in the population will satisfy the

378 imposed property and so in practice this restriction serves really to filter out cases that have been

379 improperly mapped). Then for each class of individuals, we determine the allele frequency of all the

380 Mendelian markers genome-wide (0 means only the B allele arises for the considered marker, 1 means

381 only the A allele arises, *cf.* Materials and Methods), and plotting these leads to the allele frequency

382 profiles as displayed in Fig. 1B. The x-axis is the cumulated genetic position for each of these

383 Mendelian markers. Also displayed are the corresponding smoothed frequency curves as well as the

384 allele frequency obtained without separating the individuals into the 3-marker genotype classes (dashed

385 black curve). In this example the BHB curve has a peak (pointing down) on chromosome 4 as expected

386 (the reference locus for M*) but also a second peak pointing up on chromosome 5. This peak is

387 corroborated with that of the BBB curve (down) at that same position. We can thus say that the BHB

18

388   and BBB curves together provide strong evidence for a 2:1 CNV of the A/B_A/- type, where the

389   reference locus is normal (A/B; parent 1 having allele A and parent 2 having allele B) while the second

390   locus involves a duplicate copy (A/-; carried by parent 1 only and where the copy has the allele of

391   Parent 1 for the reference marker M*; see detailed explanation of the encoding in Materials and

392   Methods). In such a notation, four different possible 1:2 or 2:1 CNV types are enumerated in the form

393   A/B_-/A, A/B_A/-, A/B_-/B, and A/B_B/-.

394   Analysis of the GABI data leads to many markers M* compatible with scenarios like that of Fig. 1 or

395   their analogs under parental or allele exchange. For instance in Fig. 2A we show the profiles for a case

396   that was detected as a 1:2 CNV with the duplicated locus within parent 2 but carrying the allele A. For

397   completeness, we show further examples in Supplementary Figure S1 to cover all four types of 1:2 or

398   2:1 CNVs. In all these cases, the hashed rectangles at the peaks delimit the regions where the software

399   localized each of the two loci by using the profile shapes in the neighborhood of these peaks (see

400   Materials and Methods). Furthermore, to add credence to the different CNV claims when analyzing the

401   data, we systematically provide simulations to determine the *expected* profiles under the hypothesis of

402   the inferred scenario. Specifically, our software produces a simulated segregating population using the

403   same number of individuals and the same marker positions as in the experimental data set but including

404   one or more duplicate copies of M* at the position(s) predicted by the scenario (see Materials and

405   Methods for a detailed explanation). Fig. 2B thus shows the expected profiles in the 1:2 CNV inferred

406   from Fig. 2A while Supplementary Figure S1 includes the simulation for each of the four types of 1:2

407   or 2:1 CNV. If the result of a simulation shows patterns of peaks very close to the experimental ones,

408   then one can have high confidence in the proposed CNV hypothesis.

19

409   The computer generation of all the profiles presented in this paper were obtained using the R package

410   CNVmap available as Supplementary File S1.

411   *The two loci of 1:2 or 2:1 CNVs sometimes arise on the same chromosome*

412   The examples just shown had the duplicated locus on a different chromosome from the reference locus,

413   but we also found other cases where the two peaks lay on the *same* chromosome. For illustration, we

414   show such a case in the GABI population in Fig. 2C, the candidate marker being SYN7974. As in the

415   previous case, our software produces a plot of the allele frequency profiles both for the experimental

416   data and for a simulated data set given the inferred scenario which for Fig. 2C is A/B_B/-, both loci

417   lying within chromosome 2. Clearly, the simulated profiles that are shown in Fig. 2D have all the

418   qualitative properties seen in the experimental data, providing strong evidence that the parent 1 really

419   does have a duplication of the region containing the SYN7974 marker and that the corresponding allele

420   is that of parent 2. Compared to the case where the two loci are on different chromosomes, the expected

421   proportion of individuals carrying the heterozygote signal is reduced: instead of the theoretical 25%, it

422   is $r/2$ where r is the recombination rate between the two loci. Whenever these two loci are very close,

423   the number of such recombinant individuals will be low and so it will be much more difficult to argue

424   that there is a real CNV *vs* simply a few genotyping errors.

425   *Allele frequency profiles reveal different types of 1:3 or 3:1 CNVs*

426   Our software is set up to detect any number of peaks in the allele frequency profiles. Thanks to this

427   feature, we found multiple cases where there were three separate loci. We illustrate such a situation in

428   Fig. 2E in the GABI population for marker PZE-105075897, where the reference locus is on

429   chromosome 5, and two additional copies were detected on chromosomes 3 and 7. The software

430   automatically identifies this as an A/B_-/B_-/A, which means that the parent 1 has no additional copy

20

431   while the parent 2 has two additional copies: one on chromosome 3 with allele "B" and one on

432   chromosome 7 with allele "A". Note that the peaks localizing these additional copies arise from the 3-

433   marker genotypic classes AAA and AHA on chromosome 3, and the 3-marker genotypic classes BBB

434   and BHB on chromosome 7. Again, to have a high level of confidence that the patterns observed have

435   been properly interpreted, one can compare with the results of a simulation as was done in the previous

436   figures. The result of simulating the triplication inferred from Fig. 2E is displayed in Fig. 2F, showing

437   that the 1:3 CNV hypothesis is indeed strongly supported by the experimental data because of the high

438   similarity between the Figs. 2E and 2F. Note that it is possible to show that the theoretical frequencies

439   of the AAA, AHA, BBB, and BHB genotypes are 1/4, and of course this result agrees with what we see

440   at the top of Fig. 2F and is not far from what is observed in the experimental case. In Supplementary

441   Figure S2 we display similar cases but arising this time in the WHEAT population, corresponding to

442   three-locus events of the types A/B_-/A_-/A, A/B_A/-_-/A, A/B_A/-_A/-, A/B_B/-_-/B, and A/B_B/-

443   _B/-. Note that in all these last cases for which one of the alleles arises solely at the reference locus, the

444   additional loci are identified only through one of the 3-marker genotypic classes, the classes having

445   heterozygotes giving rise to enhanced frequencies at those two loci but not reaching the 100% value

446   (see Supplementary Figure S2). The reasons one has enhancement but not a saturated peak or trough is

447   that the constraint of capturing the multiple-copy allele can be satisfied at *either* of the two additional

448   loci.

449   *Missing data also can provide convincing signatures for 1:2 or 2:1 CNVs in the presence of systematic*

450   *genotyping errors*

451   In Fig. 3A we show a case arising within the GABI population, constructed based on the M* marker

452   PZE-104096422. The patterns of the profiles resemble those of a 2:1 CNV except that the "BHB"

21

453   profile is replaced by a similar one labeled "B-B" where the "-" means the call of the M* marker is

454   "missing data". We denote this case A/B_A/-|Hm to indicate that "H" calls were erroneously and

455   systematically turned into missing data. Because this situation happens many times in the GABI

456   population, we investigated a few cases in detail by examining the fluorescence data, the calls and the

457   cluster file used with the Illumina array data. Given the two clouds of points produced from the

458   fluorescence data for the cases of A and B calls, we find that the "-" calls typically correspond to a

459   region that lies between those two clouds. Thus it is plausible that these cases, called as "-", are in fact

460   H, the discrepancy being due to a miscalibration of the cluster file. Based on this observation, we

461   implemented in our code a procedure whereby the peaks of missing data detected in the allele

462   frequency profiles could be interpreted as being due to such a "rule" according to which some

463   proportion or even all of the H calls of M* become transformed into "-" calls (see details in

464   Supplementary File S5). If only a fraction becomes transformed, both the BHB and B-B profiles

465   provide a peak but if all H calls are transformed into "-" calls as seems to be the case in Fig. 3A, then

466   the BHB curve will be absent. This reconsideration of the data in effect introduces a way to overcome

467   the technical problem of inadequate cluster files that we observed to arise in the GABI population data.

468   We also implemented the possibility of applying that transformation rule on *simulated* data, dependent

469   on the probability of transforming an H call into a "-" call. That probability was estimated from the

470   data. The resulting simulated profiles based on the inference of a 2:1 CNV in Fig. 3A are displayed in

471   Fig. 3B, showing an excellent agreement between theory and experiment. Furthermore, this new class

472   of events leads us to define a signature to be "strong" if each locus that is inferred to be involved in a

473   CNV is identified by at least one peak from a 3-marker genotypic class without missing data, i.e.,

474   AAA, AHA, BBB or BHB. As seen in Figs. 3A and 3B, the locus carrying the putative duplication is

22

475 identified by a peak for B-B but also by a peak for BBB and thus this event is associated with a strong

476 signature. Clearly, all of the events illustrated in the previous sections correspond to strong signatures.

477 We now move on to a more complex case where the second locus contains peaks but only for missing

478 data and thus corresponds to a *weak* signature. As motivation for this more complex case, note that a

479 miscalibration of the cluster file may be sufficiently severe that the H genotypes are called not only as

480 "-" but also as either A or B. If that is the case, the peak in the previous case arising for the BBB 3-

481 marker genotypic class no longer reaches the allele frequency zero at the second locus because some of

482 the individuals contributing to the BBB class correspond in fact to BHBs. The result of these miscalls is

483 the increase of the BBB frequency up from zero and thus the more or less disappearance of the BBB

484 peak. Although the second locus of the CNV can be localized by the B-B curve, it is no longer detected

485 *via* the BBB curve and this can raise some doubts as to the veracity of a 1:2 or 2:1 CNV interpretation.

486 In Fig. 3C we show such a case, produced from the GABI data set for the candidate marker PZE-

487 104127025. Because we have implemented the rule of transforming H calls into both "-" and "B" calls,

488 the software detects this event and classifies it as a 2:1 CNV, denoted as A/B_A/-|HmHb to reflect the

489 fact that "H" calls were erroneously systematically turned into either missing data or "B" calls. For

490 these types of events also, our software provides a simulation of what should be expected under the

491 corresponding hypothesis, estimating from the data the error rates turning H calls into "-" or into B;

492 Fig. 3D shows the corresponding result, from which one may conclude that probably the weak signal in

493 Fig. 3C is indeed indicative of a A/B_A/-|HmHb event.

494

495

23

496 *Analyses of all events across all 3 mapping populations*

497 We now move on and summarize what comes out of the analyses of each population when considering

498 all of the corresponding candidate markers. Some characteristics of these populations, in particular

499 their size and number of markers, are given in TABLE 1.

500

501 **TABLE 1. Results of automatic CNV detection in three different mapping populations**

Population name	GABI	IBM	WHEAT
Population type	DH	IRIL	RIL
Individuals	625	239	406
Total markers	13160	20913	83721
Candidate markers	746	938	10754
Single loci	489	515	2807
1:2 or 2:1 CNVs (strong signature)	77	35	47
1:2 or 2:1 CNVs (weak signature)	42	115	278
Estimated true total 1:2 or 2:1 CNVs	102	56	47
3-locus events (strong signature)	5	1	50

502 DH: doubled haploids, RIL: recombinant inbred lines, IRIL: intermated recombinant inbred lines. Candidate markers are

503 identified by their higher rate of Heterozygous or Missing Data calls. Signatures involving for each locus at least one peak

504 from a curve based on non-missing data ("AAA", "AHA", "BBB", or "BHB") are named "strong", otherwise the signatures

505 are considered "weak". Estimation of the number of "true" events was based on visual examination of the candidates for

506 which the software produced allele frequency profiles, leading to our calling the events either true or false positives. Three-

507 locus events correspond to situations where three distinct genomic regions show peaks of allele frequency profiles for

508 classes of individuals, indicating three copies of a region targeted by the candidate marker. Such three-locus events were

509 analyzed based on strong signatures only. "Single loci" correspond to candidate markers for which only the reference locus

510 shows allele frequency peaks, with no other peaks elsewhere in the genome.

511

24

512   *The GABI DH population*

513   This population is very large (625 individuals, *cf*. TABLE 1) and there is no issue of residual

514   heterozygosity coming from incomplete fixation because of the genome doubling. Given the thresholds

515   used to classify the markers into the classes Candidate, Mendelian and "Other" (*cf*. Supplementary

516   Table S1), we obtain 746 candidate markers (out of the total 13160). The software automatically

517   analyzes the profiles associated with these markers to identify peaks and corresponding loci. Of these

518   markers, 489 (a wide majority) lead to profiles involving a single locus (TABLE 1). In effect, these

519   markers were assigned to the class Candidate because of technical problems with the genotyping,

520   producing too many H or "-" calls, presumably because of some issues with the cluster file calibration

521   rather than a presence of CNVs. One such case arises for marker SYN12874; it is presented in

522   Supplementary Figure S3A and detected as "single locus" by our software. However, the remaining

523   candidate markers lead to profiles having at least two loci (see TABLE 1).

524   About half of such multilocus cases are identified by the software as being proper 1:2 or 2:1 CNVs but

525   their signatures are split between strong and weak. Visual examination of these profiles allowed us to

526   validate or not these events, leading to an estimate of a total of 102 true 1:2 or 2:1 CNVs in this data set

527   (*cf*. TABLE 1); 17 events were not validated by this visual inspection (putative false positives),

528   corresponding to 16 with weak signatures and only one with a strong signature (Supplementary

529   Table S2). Furthermore, in TABLE 2 we give the number of 1:2 or 2:1 CNVs found for each of the 4

530   possible cases, A/B_-/A, A/B_A/-, A/B_-/B and A/B_B/-. In a duplication-divergence scenario, one

531   could hypothesize that a distant ancestor of one of the individuals formed an additional copy that

532   subsequently diverged by mutation at a single base (Ohno 1970; Lynch and Conery 2000). In such a

533   scenario, one might naively expect enrichments of A/B_A/- and A/B_-/B over the other two classes.

25

534   However, in view of the numbers in TABLE 2, there is no such enrichment as all four classes have

535   occurrence numbers of similar magnitude, a point that will be justified in the discussion. It is possible

536   to further analyze the 1:2 or 2:1 CNVs by considering the separation into strong and weak signatures.

537   Supplementary Table S1 gives the numbers for all types of automatically detected events. As indicated

538   in Supplementary Table S2, many events with weak signatures are *not* validated by our visual

539   inspection, so it is best to concentrate on the events providing strong signatures.

540   Of the other multilocus events, only 5 have a strong signature involving three or more loci (*cf.*

541   TABLE 1). These 5 events belong to just a few of the different types with respect to allelic content as

542   shown in TABLE 2; although it may be tempting to argue for an enrichment of the A/B_-/B_-/B type,

543   the numbers are very small so it is not useful to go into such speculations.

544   Lastly, the software identifies 131 events in which there are multiple loci but with patterns of profiles

545   that are "unknown" because they differ from those produced by CNVs in the list provided in

546   Supplementary Table S1. Might some of these cases reveal true CNVs that are novel compared to what

547   we considered so far? To get some insight into that possibility, we examined the corresponding profiles.

548   Some cases provide no compelling evidence for a CNV, the profiles are simply very noisy and peaks

549   may be presumed to be non-significant. For other cases, as in Supplementary Figures S3B and S3C,

550   there is clearly an additional locus but the profiles are not as expected from our list of standard 1:2 or

551   2:1 CNVs. We also find cases of more than one additional locus as in Supplementary Figures S3D and

552   S3E, where again the signature is not compatible with any of our standard 3 locus events or extensions.

553   *The IBM IRIL population*

26

554   Similar statistics for the IBM RIL population are given in TABLE 1 and TABLE 2, and in

555   Supplementary Tables S1 and S2 and differ mainly quantitatively when compared to the GABI

556   population. Nevertheless, from the conceptual point of view, the main new feature when going from

557   GABI to IBM is the presence of residual heterozygosity in Mendelian markers. Indeed, since IBM is an

558   intermated RIL population, fixation can be incomplete because either not enough generations of selfing

559   have been applied or because there is a selective force impeding the homozygous state, situations that

560   do not occur with doubled haploids. But once appropriate thresholds are set for the minimum number

561   of heterozygote calls to select a marker as a candidate, our method was also efficient to discover CNVs

562   in that population. As with the GABI population, the rate of true over false positives was extremely

563   high (100% here, see Supplementary Table S2) when considering events with strong signatures. On the

564   other hand, events with weak signatures gave a higher proportion of false positives in IBM as

565   compared to GABI. Finally, there was only a single event associated with three loci. It should be noted

566   that both IBM and GABI populations were genotyped with the same Illumina Maize SNP50 array, and

567   deal with the same species, which is consistent with the qualitatively similar results obtained.

568   *The WHEAT RIL population and importance of homeologues*

569   The case of the WHEAT population is *a priori* quite different from the two first populations because

570   bread wheat is hexaploid and also because the population was genotyped with a different SNP array.

571   Not so surprisingly, we observed quite different results (see TABLE 1 and TABLE 2, and

572   Supplementary Tables S1 and S2). This population is quite large (406 individuals, *cf*. TABLE 1) and it

573   has far more markers (83721) than the other populations, justifying that the number of candidate

574   markers, 10754, is also much higher (see Supplementary Table S1 for the thresholds used to define

575   Candidate and Mendelian markers). In contrast to the other populations, the great majority of these

27

576   candidates do *not* give rise to any profile, even the single-locus one, which is indicative of potential

577   mapping problems. Nevertheless, a fair fraction of the candidates does give profiles. A large number of

578   these, 2807 specifically, are identified as having just the reference locus and thus are not of interest.

579   Such quite frequent cases are expected in WHEAT as in IBM because residual heterozygosity produces

580   false candidates.

581   Of the remaining candidates, some are identified by the software as associated with 2 or more loci. For

582   the events detected as being of the 1:2 or 2:1 CNV type, 47 have a strong signature and 278 have a

583   weak signature. Validation by inspecting all of these events suggests that only the strong signatures

584   provide true positives. In TABLE 2 we give the number of 1:2 or 2:1 CNVs found for each of the 4

585   types, A/B_-/A, A/B_A/-, A/B_-/B, and A/B_B/-. Though these classes are less balanced than in the

586   GABI population, the evidence for enrichment of particular classes is not very strong. Supplementary

587   Table S1 gives the numbers for all types (including thus the rules to take into account genotyping

588   errors) of automatically detected events. However, as indicated in Supplementary Table S2 and

589   previously mentioned, the events with weak signatures are not validated by our visual inspection, so it

590   is best to focus on the events with strong signatures only.

591   This brings us to the strong signatures for events involving three or more loci (*cf.* TABLE 1). The types

592   of these events are given in TABLE 2. Clearly the main types seen have one allele in three copies and

593   the other in a single copy. It is relevant here to recall that wheat is a hexaploid which contains three

594   genomes (A, B, and D) with seven chromosomes each. We found 20 cases where the three copies are

595   located on three homeologous chromosomes (e.g. in Supplementary Figures S2C and S2E on

596   chromosomes 2 and 6), 8 cases with two copies on two homeologous chromosomes and the third one

597   on a different (non-homeologous) chromosome (e.g., in Supplementary Figures S2A and S2I), and

28

598   finally 22 cases where the three loci are on three non-homeologous chromosomes (e.g., in

599   Supplementary Figure S2G). Although enrichment amongst homeologues is expected, it is appealing to

600   have it come out from our automatized software.

601

602   **TABLE 2. Number of each type of event found in three segregating populations.**

Population name	GABI	IBM	WHEAT
A/B_-/A	21	0	3
A/B_A/-	17	13	12
A/B_-/B	18	22	15
A/B_B/-	20	0	17
A/B_-/A_-/A	0	0	5
A/B_A/-_A/-	0	0	1
A/B_-/B_-/B	3	1	0
A/B_B/-_B/-	0	0	9
A/B_-/A_A/-	0	0	0
A/B_A/-_-/A	1	0	16
A/B_-/A_-/B	0	0	0
A/B_-/B_-/A	1	0	0
A/B_-/A_B/-	0	0	0
A/B_B/-_-/A	0	0	0
A/B_A/-_-/B	0	0	0
A/B_-/B_A/-	0	0	0
A/B_A/-_B/-	0	0	0
A/B_B/-_A/-	0	0	0
A/B_-/B_B/-	0	0	0
A/B_B/-_-/B	0	0	19

603   Number of events found for each category of 1:2 or 2:1 CNVs (upper part) or 3-locus events (lower part) in two Maize

604   populations (GABI and IBM), and one wheat population (WHEAT). Every event was automatically detected by the

605   software, and also visually checked by looking at the corresponding allele frequency profiles. Each category is encoded as a

606   string of 2 to 3 groups of 3 characters each, separated by an underscore. The first group is always encoded "A/B" and

607   indicates the reference locus, located at the position where the candidate marker was initially mapped. Further groups

608   indicate other copies of the region targeted by the candidate marker. For all groups (loci), the letters just before and just after

609   the slash represent respectively the haplotypes of the first parent (alleles denoted "A"), and of the second parent (alleles

610   denoted "B").

611

29

612   *Sequence-based validation of candidates with strong but not weak signatures for IBM*

613   CNVmap provides candidate CNVs and predictions for the associated loci. In the case of the IBM

614   population, the parental genomes are fully assembled and so our predictions can be checked by

615   searching in those reference genomes for multiple occurrences of the specific sequences flanking the

616   candidate SNPs *via* BLAST (see Methods for details). The results of those analyses are as follows.

617   First, concerning events having strong signatures, the majority of the predictions are validated.

618   Specifically, of the 13 predicted 2:1 CNVs (A/B_A/- or A/B_B/-; two copies in parent B73), all are

619   validated, while of the 22 predicted 1:2 CNVs (A/B_-/A or A/B_-/B; two copies in parent Mo17), 12

620   are validated (Fig. 4). Not surprisingly, when testing the hypothesis H_0 of no CNV in these strong

621   signature events, all but one of the *p*-values (for events validated or non-validated by BLAST) were

622   below 0.05 and most of them were below 10^{-3} (Fig. 4, see Materials and Methods for the calculation of

623   these *p*-values). Second, concerning events having weak signatures, essentially none of them are

624   validated; furthermore, Fig. 4 shows a broad distribution of *p*-values, calling in doubt the credibility of

625   these weak candidate CNVs.

626

Discussion

628   *An original method based on linkage to detect and genetically map CNVs*

629   We presented a new method for revealing and genetically mapping copy number variations in bi-

630   parental segregating populations made of homozygous individuals. The heart of the method is the fact

631   that a marker participating to a CNV will lead to an excess of heterozygotes in the segregating

30

632 population with associated signatures in genome-wide allele frequency profiles. We validated this

633 method with maize doubled-haploid lines, maize intermated recombinant inbred lines, and wheat

634 recombinant inbred lines, revealing CNVs even in these last two types of populations in spite of the

635 presence of their residual heterozygosity. The approach does not involve any *a priori* knowledge about

636 the type or location of the events, rather it is based on signatures in genome-wide allele frequency

637 profiles, assuming that the individuals therein have been genotyped. Such genotyping might have been

638 done for instance for detecting QTLs or for breeding purposes (as in genomic selection), and thus our

639 approach can "piggy-back" *for free* after the production of such genetic material. In this context, our

640 detection of CNV loci has a spatial resolution that depends on the local recombination rate, so the

641 larger the population and the larger the recombination rate the better. Nevertheless, detecting the

642 *existence* of duplicated loci and finding their approximate localization is relatively easy: 239

643 individuals are sufficient for the RILs (F6) we studied, and lower numbers can also give good results.

644 The major limitation of our approach is that the duplicated loci cannot be too close to the original

645 locus, and thus we cannot easily detect tandem duplications. Another requirement of course is that the

646 markers be robustly ordered, so the quality of the genetic map is important. In usual genotyping arrays,

647 SNPs have been included only if they were found to be exploitable on a reference panel, and thus SNPs

648 with heterozygous signals have little chance of having been kept for inclusion on that array. As a result,

649 we certainly strongly underestimate the real number of CNVs. Consequently, our approach, when used

650 on data produced with SNP arrays, should not be considered as a way of surveying the *number* of

651 structural variation events between two parents, but rather as a cost-free means of getting, for a subset

652 of such events, detailed information on their nature and in particular the genomic locations of the

653 associated copies.

654 *Different rate of success of sequence-based validations in B73 and* Mo17 *parents*

655 In the IBM population, all 2:1 CNV events (with two copies in parent B73) were confirmed by BLAST

656 analysis, whereas only 55% of 1:2 CNVs (with two copies in parent Mo17) led to successful validation.

657 This difference may be due to a less good quality of the genome assembly of Mo17. Indeed, the quality

658 of B73 assembly is most probably higher because that inbred line was chosen for the first Maize

659 reference genome sequence, so more sequencing and assembly effort was dedicated to this line.

660 Another explanation may be that the level of sequence divergence between B73 and Mo17 leads some

661 loci to escape our BLAST search on Mo17 because the oligonucleotides used in the MaizeSNP50

662 Illumina array were designed based on the B73 reference sequence, but nevertheless the markers would

663 still be able to hybridize on Mo17 DNA.

664 *Applicability to non-fixed populations*

665 Our method is primarily based on the detection of apparent heterozygosity, so the presence of

666 heterozygous loci due to incomplete fixation, as occurs in RIL populations, is expected to greatly lower

667 the efficiency of detection. So we adapted the criterion for a marker to be a non-Mendelian candidate

668 by enforcing its level of heterozygosity to be higher than a given threshold, thereby limiting the number

669 of candidate markers to analyze. And in fact, the method proved to be sufficiently powerful to detect

670 CNVs in recombinant inbred line populations corresponding to six generations of selfing.

671 *Detecting CNVs in the presence of systematic genotyping errors*

672 We also found clear signatures of CNVs based on missing data. Typically, such missing data arise from

673 technical systematic biases in the genotyping (e.g. systematic mis-calling of heterozygotes as missing

674 data and/or as homozygotes), and thus can be put on a similar footing with the more standard signatures

32

675   of CNVs such as A/B_B/-. Thus, in addition to being able to detect non-Mendelian SNPs in a

676   genotyping array (in linkage analyses like QTL mapping, it is useful to remove them), our method is

677   also able to reveal some flaws in the cluster files used for analyzing Illumina array data. Such cluster

678   files, which determine the way fluorescence levels at two different wavelengths are converted into a

679   genotype call, may be more or less appropriate depending on the genetic origin of the material being

680   genotyped, or may be sensitive to some variations of the experimental conditions during the

681   hybridization of the arrays. In our results, we could clearly demonstrate that a large number of markers

682   suffered from such systematic genotyping errors transforming heterozygote calls into missing data and/

683   or into homozygotes.

684   *Most detected events correspond to four types of parent-specific duplications*

685   For the two maize datasets studied, the vast majority of the events involve just two loci. Furthermore,

686   most of these can be identified as parent-specific duplications, corresponding to the four types

687   A/B_A/-, A/B_-/A, A/B_B/-, or A/B_-/B. Among them, A/B_A/- and A/B_-/B involve haplotypes with

688   two copies of the same allele in the parent carrying the duplication. On the other hand, A/B_-/A and A/

689   B_B/- involve two different alleles at the additional locus in the haplotype of the parent carrying the

690   duplication, so that parent -- although it is an inbred line supposed to be almost fully homozygous -- is

691   expected to be genotyped as heterozygous for marker M*. We thus analyzed genotyping data obtained

692   from the parents, and indeed, in all such cases, we observed heterozygous calls in the genotype data

693   associated with the corresponding parent of the GABI population. From an evolutionary point of view,

694   the latter situations (A/B_-/A and A/B_B/-) might seem to suggest a temporal order, namely a

695   duplication followed by a divergence at the reference locus. However, it is just as likely that the

696   divergence happened first and the duplication later: since the two loci are not tightly linked,

33

697   recombination between them can very well produce the A/B_B/- haplotype starting with the A/B_-/B

698   haplotype. Thus there is no *a priori* expectation that two of the four types of 1:2 or 2:1 CNV should be

699   much rarer than the other two types, and this is in line with what comes out of the summary statistics as

700   can be seen in TABLE 2 for GABI and WHEAT populations. However, in IBM, all 1:2 or 2:1 CNVs

701   detected had the same allele on both copies for a given haplotype, which suggests that SNPs in the IBM

702   mapping data set may have been selected to remove markers with heterozygous calls on the parents.

703   *The special case of wheat homeologous chromosomes*

704   In the case of wheat which is a hexaploid species containing three diploid genomes, one has the further

705   issue of homeologous chromosomes. Because these chromosomes have diverged from a common

706   ancestor, the gene content is quite well conserved and chromosomes display good collinearity with

707   limited rearrangements (Consortium (IWGSC) *et al.* 2018). A consequence of this is that SNPs may not

708   necessarily be genome-specific and may therefore hybridize on two or three homeologous loci

709   (Rimbert *et al.* 2018). Such similar sequences may generate signals of CNVs in the allelic profiles and

710   so we asked the question of whether the duplicated loci we found in wheat were more often than

711   expected on the homeolog. The analysis of the two- and three-locus events in our WHEAT dataset in

712   fact shows a huge enrichment for favoring the homeologous chromosomes. Our method can thus

713   provide a useful way to assess the level of genome-specificity of the SNPs of a given genotyping array,

714   and help validating the selection of subsets of purely Mendelian markers.

715   *Only a tiny minority of the allelic profiles involve three or more loci in the maize populations*

716   Because of the hexaploid nature of wheat, this plant was expected to reveal many triplication events if

717   markers were not perfectly genome-specific, and this is actually what we found. On the other hand, in

718   maize the ancient allotetraploid origin of the species is old enough for most markers to behave as

34

719   single-copy, so one may expect far fewer three-locus events. And indeed, as can be seen from

720   TABLE 2, there are some candidate markers that generate profiles with three loci but they are quite rare

721   and arise mostly within the GABI population. This difference may be due to the GABI population

722   being much larger, allowing our method to be more powerful on that data.

723   *Possible evolutionary scenarios for triplications*

724   Some entangled events such as A/B_B/-_-/B may seem unexpected because they involve the allele

725   from the opposite parent. However, just as we explained for the two-locus case, recombination can

726   scramble the assignment of alleles and so a posteriori such events are not surprizing. But there is

727   another possibility for justifying such an entangled CNV without appealing to recombination. Indeed,

728   imagine that an ancestral triplication arose so that the allele B was present at all three loci. Parent 1 and

729   Parent 2 may be identical by descent for that triplication for all of their homologues. If so, today's

730   situation can very well be due to subsequent divergence only: the divergence at the reference locus

731   would produce a SNP while the divergence at the other two loci would be more severe, for instance

732   corresponding to a deletion or appearance of other SNPs in the flanking sequences of the two other

733   loci, thereby preventing the hybridization of oligonucleotides. Clearly such a scenario can also be

734   responsible for entangled 1:2 or 2:1 CNVs.

735   *Conclusion*

736   We developed an original *linkage-based* method to detect CNVs and genetically map the associated

737   previously unknown copies from genotype data of segregating populations. Our software based on this

738   method makes it possible to perform fully automatic mining of segregation data to extract a list of high

739   confidence CNVs, including the detailed type of event and the genomic location(s) of the initially

740   unknown locus or loci. It is thus a costless and easy way to generate additional added value from

35

741   genotyping efforts initially dedicated to genetic map construction or QTL analyses. Because of its ease

742   of use, our tool for detecting CNVs could be applied to other kinds of populations. First, going from bi-

743   parental to multi-parental RILs as used in MAGIC (Dell'Acqua *et al.* 2015) or NAM (McMullen *et al.*

744   2009) populations should be straightforward, our computer program can be used as such for all biallelic

745   SNPs. Second, it seems possible that CNVs could be detected by our approach when using the kinds of

746   panels exploited in GWAS when the individuals in the panel are homozygous (e.g. inbred lines); the

747   method would then correspond to searching genome-wide for associations between allele frequency

748   and the particular 3-loci genotype (e.g., AHA) detected at a reference locus (that is for the non-

749   Mendelian marker of interest and its two flanking markers). Such an approach, using a diversified

750   panel, might in fact allow one to identify duplicated loci with a high level of resolution.

751

752

753   **Declarations**

754   *Ethics approval and consent to participate*: 'Not applicable'

755   *Consent for publication*: 'Not applicable'

756   *Competing interests:*

757   The authors declare that they have no competing interests

758   *Funding*:

36

765 *Authors' contributions*:

766 MF and OM conceived the method, developed the computer programs, analyzed and interpreted the

767 results, and wrote the final manuscript. MF, KJ and OM worked on the computer program, performed

768 analyzes on the data sets, and produced the first draft. EP, CK, and SM provided data sets. All authors

769 read, edited and approved the manuscript.

37

775

References

Alkan C., J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, *et al.*, 2009 Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics 41: 1061–1067. https://doi.org/10.1038/ng.437

Alkan C., B. P. Coe, and E. E. Eichler, 2011 Genome structural variation discovery and genotyping. Nature Reviews Genetics 12: 363–376. https://doi.org/10.1038/nrg2958

Bailey J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, *et al.*, 2002 Recent Segmental Duplications in the Human Genome. Science 297: 1003–1007. https://doi.org/10.1126/science.1072047

Beckmann J. S., X. Estivill, and S. E. Antonarakis, 2007 Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nature Reviews Genetics 8: 639–646. https://doi.org/10.1038/nrg2149

Beló A., M. K. Beatty, D. Hondred, K. A. Fengler, B. Li, *et al.*, 2009 Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120: 355–367. https://doi.org/10.1007/s00122-009-1128-9

Chaisson M. J., D. Brinza, and P. A. Pevzner, 2009 De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 19: 336–346. https://doi.org/10.1101/gr.079053.108

Chen K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, *et al.*, 2009 BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6: 677–681. https://doi.org/10.1038/nmeth.1363

Choulet F., A. Alberti, S. Theil, N. Glover, V. Barbe, *et al.*, 2014 Structural and functional partitioning of bread wheat chromosome 3B. Science 345: 1249721. https://doi.org/10.1126/science.1249721

Colella S., C. Yau, J. M. Taylor, G. Mirza, H. Butler, *et al.*, 2007 QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013–2025. https://doi.org/10.1093/nar/gkm076

Conrad D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, *et al.*, 2010 Origins and functional impact of copy number variation in the human genome. Nature 464: 704–712. https://doi.org/10.1038/nature08516

Consortium (IWGSC) T. I. W. G. S., R. Appels, K. Eversole, N. Stein, C. Feuillet, *et al.*, 2018 Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361: eaar7191. https://doi.org/10.1126/science.aar7191

Cooper G. M., T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson, 2008 Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40: 1199–1203. https://doi.org/10.1038/ng.236

Dell'Acqua M., D. M. Gatti, G. Pea, F. Cattonaro, F. Coppens, *et al.*, 2015 Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. Genome Biology 16: 167. https://doi.org/10.1186/s13059-015-0716-z

Falque M., L. K. Anderson, S. M. Stack, F. Gauthier, and O. C. Martin, 2009 Two Types of Meiotic Crossovers Coexist in Maize. The Plant Cell 21: 3915–3925. https://doi.org/10.1105/tpc.109.071514

39

Ganal M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler, *et al.*, 2011 A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome, (L. Lukens, Ed.). PLoS ONE 6: e28334. https://doi.org/10.1371/journal.pone.0028334

Givry S. de, M. Bouchez, P. Chabrier, D. Milan, and T. Schiex, 2004 CarthaGene: multipopulation integrated genetic and radiation hybrid mapping. Bioinformatics 21: 1703–1704. https://doi.org/10.1093/bioinformatics/bti222

Guryev V., K. Saar, T. Adamovic, M. Verheul, S. A. A. C. van Heesch, *et al.*, 2008 Distribution and functional impact of DNA copy number variation in the rat. Nature Genetics 40: 538–545. https://doi.org/10.1038/ng.141

Lee M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt, *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant molecular biology 48: 453–61.

Li R., H. Zhu, J. Ruan, W. Qian, X. Fang, *et al.*, 2010 De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20: 265–272. https://doi.org/10.1101/gr.097261.109

Lynch M., and J. S. Conery, 2000 The Evolutionary Fate and Consequences of Duplicate Genes. Science 290: 1151–1155. https://doi.org/10.1126/science.290.5494.1151

McMullen M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, *et al.*, 2009 Genetic Properties of the Maize Nested Association Mapping Population. Science 325: 737–740. https://doi.org/10.1126/science.1174320

McPeek M. S., and T. P. Speed, 1995 Modeling Interference in Genetic Recombination. Genetics 139: 1031–1044.

Mills R. E., C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, *et al.*, 2006 An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 16: 1182–1190. https://doi.org/10.1101/gr.4565806

Ohno S., 1970 *Evolution by Gene Duplication*. Springer Science & Business Media.

Pang A. W., J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq, *et al.*, 2010 Towards a comprehensive structural variation map of an individual human genome. Genome Biology 11: R52. https://doi.org/10.1186/gb-2010-11-5-r52

Presterl T., M. Ouzunova, W. Schmidt, E. M. Möller, F. K. Röber, *et al.*, 2007 Quantitative trait loci for early plant vigour of maize grown in chilly environments. Theor Appl Genet 114: 1059–1070. https://doi.org/10.1007/s00122-006-0499-4

Redon R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, *et al.*, 2006 Global variation in copy number in the human genome. Nature 444: 444–454. https://doi.org/10.1038/nature05329

Rimbert H., B. Darrier, J. Navarro, J. Kitt, F. Choulet, *et al.*, 2018 High throughput SNP discovery and genotyping in hexaploid wheat. PLOS ONE 13: e0186329. https://doi.org/10.1371/journal.pone.0186329

Simpson J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, *et al.*, 2009 ABySS: A parallel assembler for short read sequence data. Genome Res. 19: 1117–1123. https://doi.org/10.1101/gr.089532.108

Springer N. M., K. Ying, Y. Fu, T. Ji, C.-T. Yeh, *et al.*, 2009 Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLOS Genetics 5: e1000734. https://doi.org/10.1371/journal.pgen.1000734

Sudmant P. H., J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, *et al.*, 2010 Diversity of Human Copy Number Variation and Multicopy Genes. Science 330: 641–646. https://doi.org/10.1126/science.1197005

Wu Y., P. R. Bhat, T. J. Close, and S. Lonardi, 2008 Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph, (L. Kruglyak, Ed.). PLoS Genetics 4: e1000212. https://doi.org/10.1371/journal.pgen.1000212

Ye K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Yoon S., Z. Xuan, V. Makarov, K. Ye, and J. Sebat, 2009 Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19: 1586–1592. https://doi.org/10.1101/gr.092981.109

Zare F., M. Dow, N. Monteleone, A. Hosny, and S. Nabavi, 2017 An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics 18: 286. https://doi.org/10.1186/s12859-017-1705-x

Zerbino D. R., and E. Birney, 2008 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829. https://doi.org/10.1101/gr.074492.107

Zhang F., W. Gu, M. E. Hurles, and J. R. Lupski, 2009 Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics 10: 451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

777

778

Figure legends

780

Figure 1. Consequences of parent-specific locus duplication on allele frequency profiles

A: Duplication in one parent leads to apparent heterozygotes in the gametes for the non-Mendelian marker M*. B: simulated genome-wide allele frequency profiles using subsets of individuals belonging to the three-markers genotype classes AAA, BBB, or BHB at markers M_L, M*, and M_R (M_L and M_R being the Mendelian markers flanking M*; see text). Such profiles reveal the loci involved in duplications. The allele of parent 1 is called "A", the allele of parent 2 is called "B", heterozygotes are called "H", and missing data are called "-". Each curve shows the frequency of the allele "A" along the genome (X-axis indicates cumulated genetic positions), when considering different subsets of individuals of the population as follows: cyan dots and curve for individuals (denoted "BHB") genotyped "H" at the candidate marker M* and "B" on both non-candidate flanking markers M_L and M_R indicating the allelic context of the region, and similarly red for "AAA" individuals, dark blue for "BBB" individuals. Hatched rectangles indicate the estimated confidence intervals on the position of the detected loci involved in the event. The rectangle is black for the reference locus (see text) and red for the secondary locus. Dots represent values of individual markers and associated curves show the result of the smoothing procedure used to detect the peaks. Lastly, the black dashed line indicates the frequency of "A" allele based on all individuals of the population.

797

Figure 2. Examples of signatures of events involving two or three loci

43

799   Data are from the Maize doubled-haploid population GABI. Dots and curves have the same meaning as

800   in Figure 1. Panels A, C, and E show experimental profiles for respectively a 1:2 CNV event with both

801   copies on different chromosomes, a 2:1 CNV event with both copies on the same chromosome, and a 3-

802   locus event with copies on three different chromosomes. Panels B, D, and F show simulation results

803   reproducing the CNV situation inferred from panels A, C, and E respectively (see text). The allele of

804   parent 1 is called "A", the allele of parent 2 is called "B", heterozygotes are called "H", and missing

805   data are called "-". Each curve shows the frequency of the allele "A" along the genome (X-axis

806   indicates cumulated genetic positions), when considering different subsets of individuals of the

807   population as follows: pink dots and curve for individuals (denoted "AHA") genotyped "H" at the

808   candidate marker and "A" on both non-candidate flanking markers indicating the allelic context of the

809   region, and similarly cyan for "BHB" individuals, red for "AAA" individuals, dark blue for "BBB"

810   individuals. Curves generated by the software for classes based on missing data (light grey for "A-A"

811   individuals, and black for "B-B" individuals) were hidden here for better clarity of the profiles.

812   Hatched rectangles indicate the estimated confidence intervals on the position of the detected loci

813   involved in the event. They are black for the reference locus (see text) and red for the secondary locus

814   (or red or green for the two secondary loci in the case of the 3-locus event in panels E and F). The

815   name of the candidate (non-Mendelian) marker considered is given in the header of each panel, as well

816   as numbers of individuals counted for each three-locus genotype class.

817

818   **Figure 3. Examples of profiles showing characteristic signatures of CNVs in the presence of**

819   **systematic genotyping errors**

44

820   Data are from the GABI population. Dots and curves have the same meaning as in Figure 1. Panel A

821   shows a typical "strong" signature, with a 2:1 CNV event in the case where "H" calls of the candidate

822   marker were systematically called missing data ("-"). Panel C shows a typical "weak" signature, where

823   a 2:1 CNV event in the case where "H" calls of the candidate marker were systematically called either

824   as missing data ("-") or as "B" in non-zero proportions. The software provides estimated systematic

825   error rates for each such candidate. Panels B and D show simulation results reproducing the CNV

826   situation inferred from A and C respectively (see text). The allele of parent 1 is called "A", the allele of

827   parent 2 is called "B", and missing data are called "-". Each curve shows the frequency of the allele "A"

828   along the genome (X-axis indicates cumulated genetic positions), when considering different subsets of

829   individuals of the population as follows: pink dots and curve for individuals (denoted "AHA")

830   genotyped "H" at the candidate marker and "A" on both non-candidate flanking markers indicating the

831   allelic context of the region, and similarly cyan for "BHB" individuals, red for "AAA" individuals, dark

832   blue for "BBB" individuals, grey for "A-A" individuals, and black for "B-B" individuals. Hatched

833   rectangles indicate the estimated confidence intervals on the position of the detected loci involved in

834   the event. The rectangle is black for the reference locus (see text) and red for the secondary locus. The

835   name of the candidate (non-Mendelian) marker considered is given in the header of each panel, as well

836   as numbers of individuals counted for each three-locus genotype class.

837

Figure 4. Validation of CNVs found in the IBM population

839   All 1:2 and 2:1 CNVs found in the IBM population (obtained from the cross B73xMo17), based on

840   Strong or Weak signatures, were submitted to two different types of validation (see Materials and

841   Methods): (1) a p-value (Y-axis) was computed using 1000 simulations for the H_0 hypothesis: "the

45

842 marker M* is present as a single copy in both B73 and Mo17 parents", and (2) the presence of the

843 second copy in the reference genome of the expected parent (B73 for 2:1 CNVs and Mo17 for 1:2

844 CNVs) was checked using BLAST search against whole genome sequence assemblies of both parents.

845 On X-axis, events are denoted 'Blast-OK' or 'No-Blast' according to the success of the sequence-based

846 validation, and 'B73 (or Mo17)-Strong (or Weak)' according to the types of events considered (strength

847 of the signature and B73 or Mo17 having two copies). Numbers below the line Y=0 indicate the

848 number of events in each category.

849

850 **List of Supplementary Material**

851 **Supplementary Table S1**. Detailed statistics of the three populations GABI, IBM, and WHEAT

852 **Supplementary Table S2**. Assessment of rates of true and false positives by visual inspection

853 **Supplementary Figure S1**. Examples of allele frequency profiles showing characteristic signatures of

854 four different types of 1:2 or 2:1 CNVs.

855 **Supplementary Figure S2**. Examples of allele frequency profiles for each type of three-locus events

856 found in wheat.

857 **Supplementary Figure S3**. Examples of allele frequency profiles showing different situations where

858 no CNV was identified by the software.

859 **Supplementary File S1.** CNVmap software as an R package (CNVmap12.tar.gz archive to open in R

860 with the function install.packages()).

861 **Supplementary File S2**. Compressed (.zip) archive of GABI data sets to run CNVmap.

862 **Supplementary File S3**. Compressed (.zip) archive of IBM data sets to run CNVmap.

863 **Supplementary File S4**. Compressed (.zip) archive of WHEAT data sets to run CNVmap.

864   **Supplementary File S5**. Supplementary Methods: specification of all CNV signatures recognized by

865   the software CNVmap.

866   **Supplementary File S6**. Installing with Rstudio. Particular procedure for some versions of R/Rstudio.
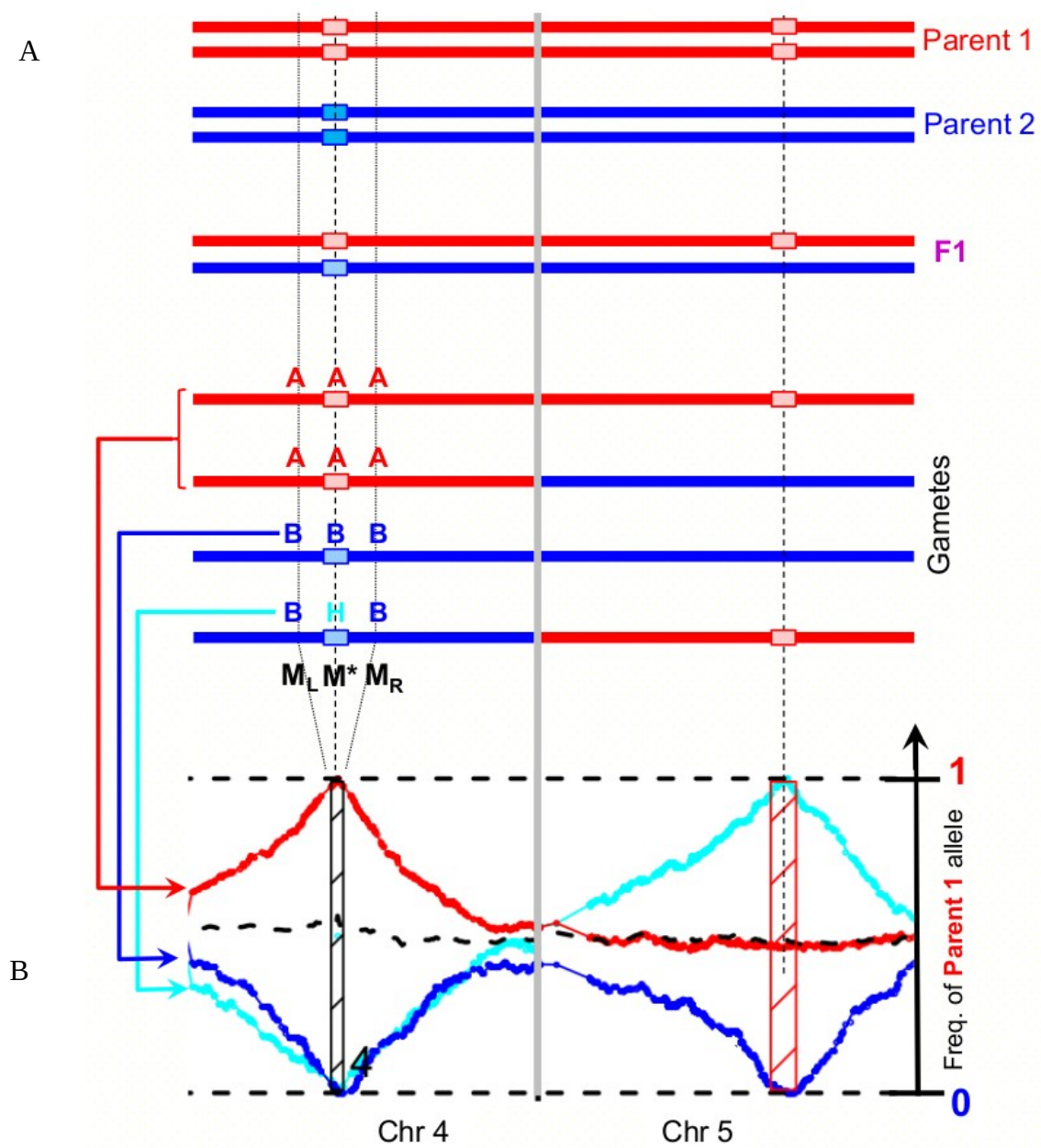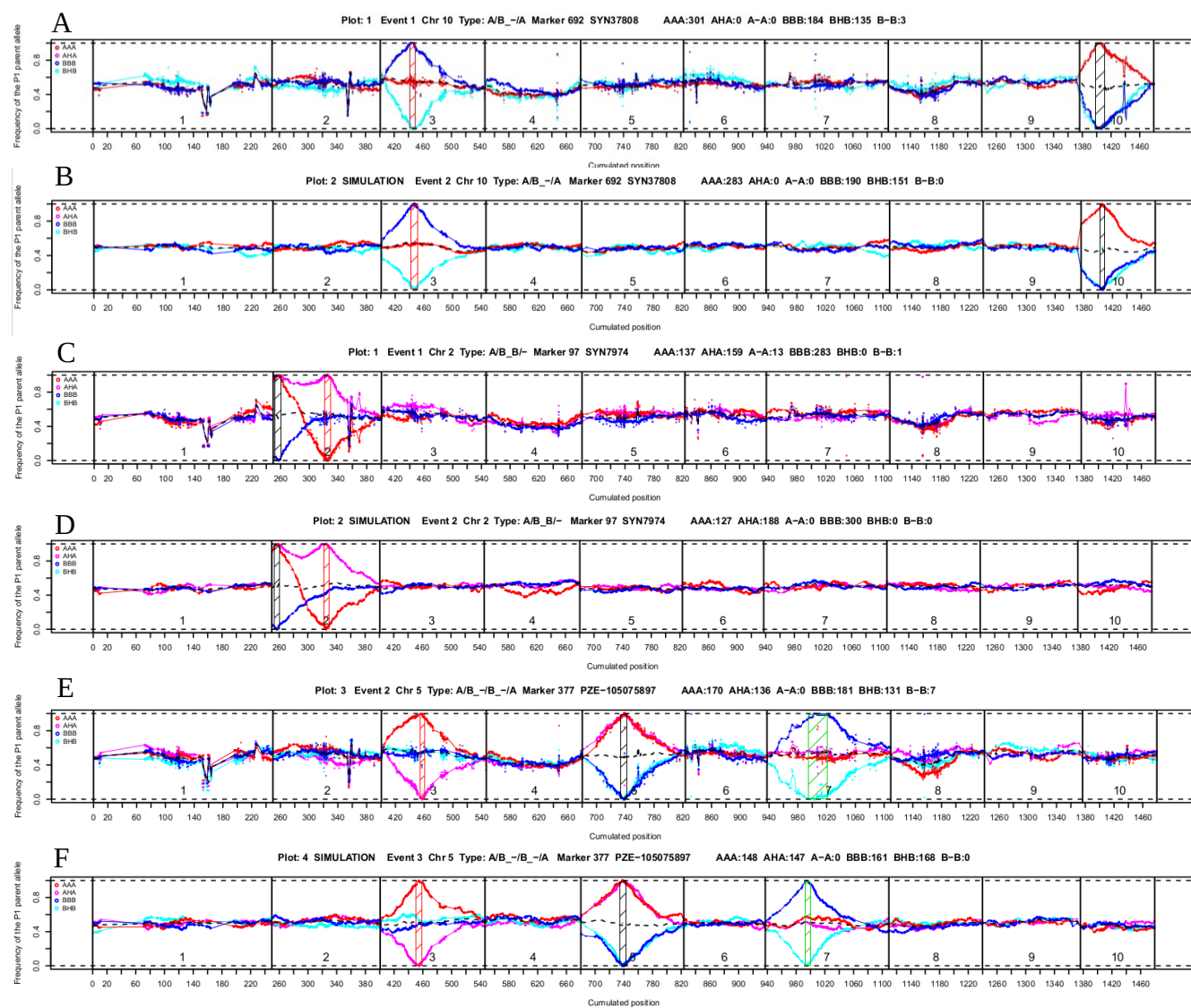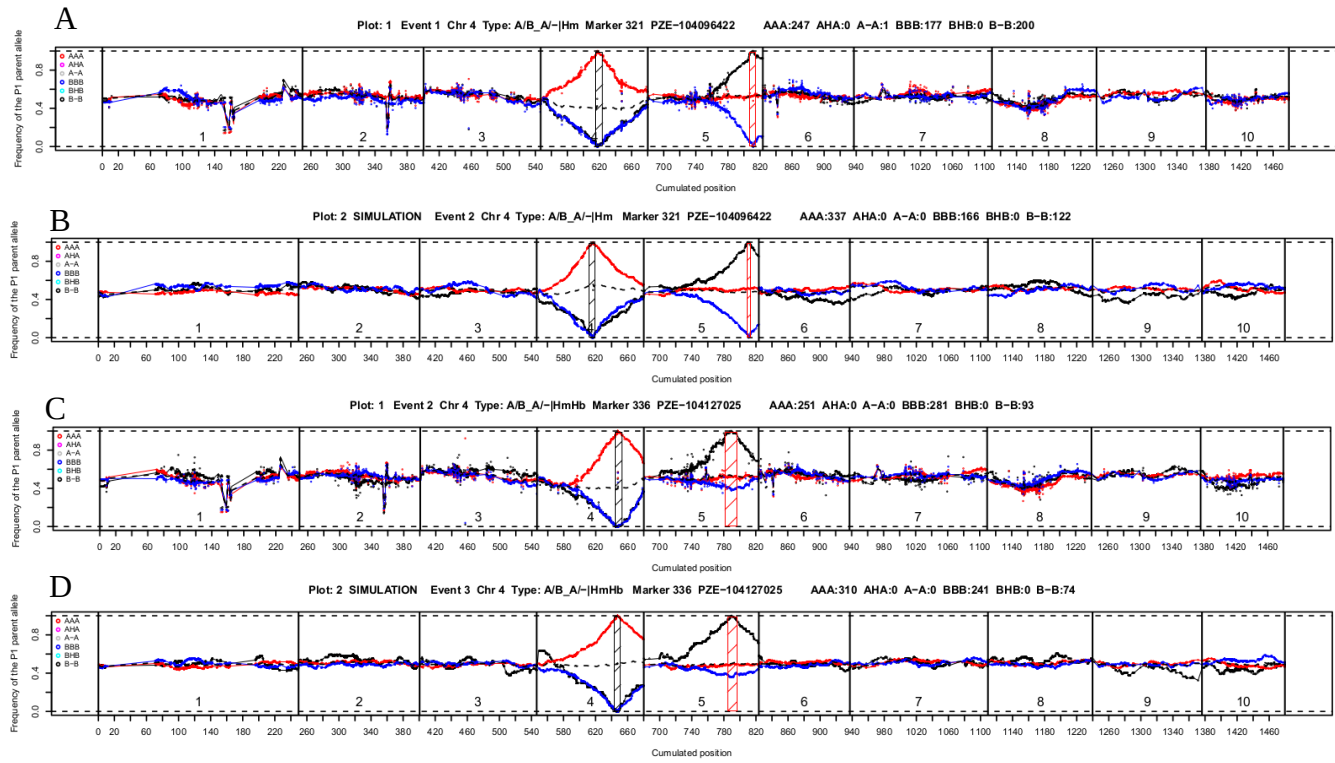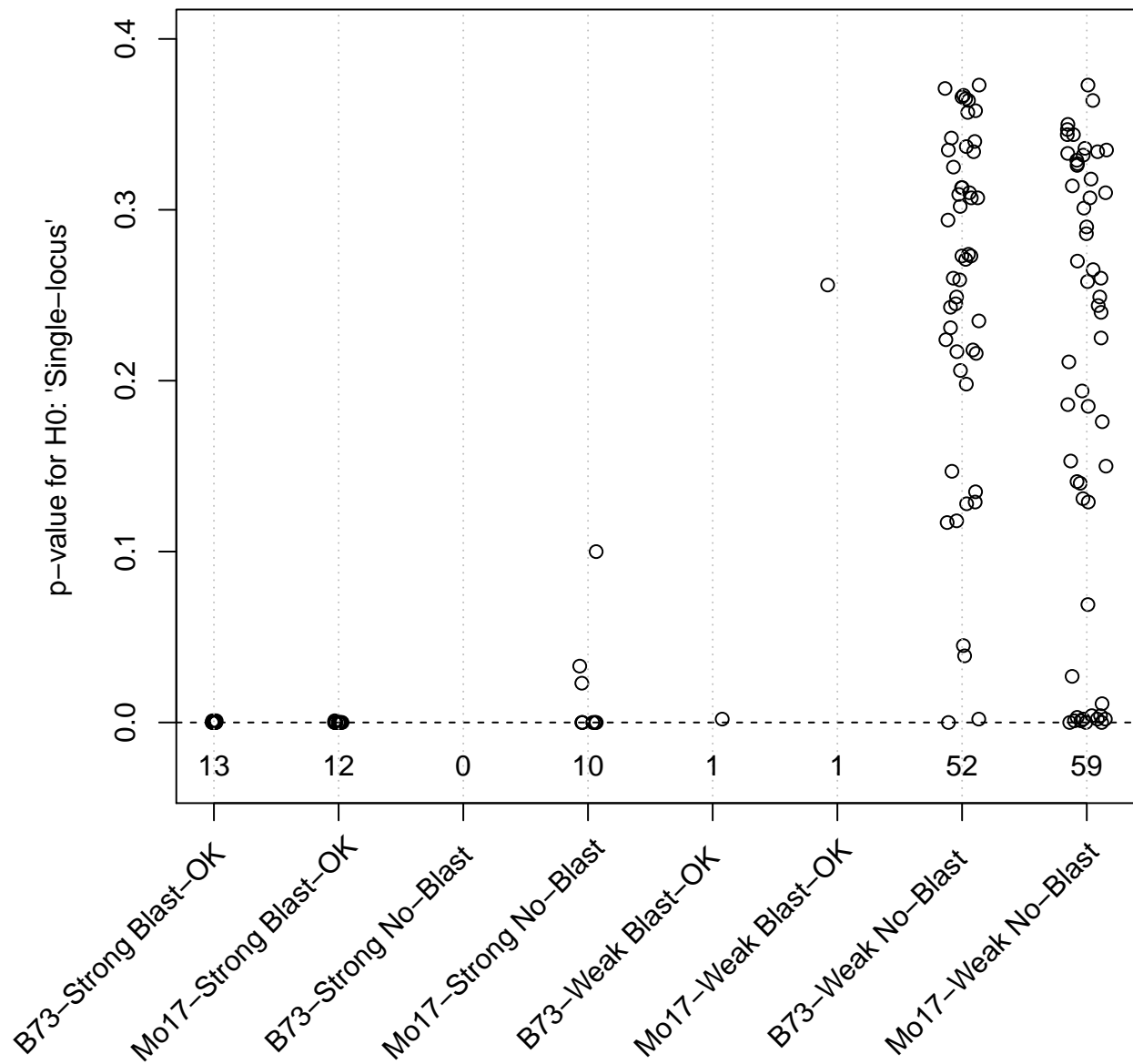
867

Figure 1.

Figure 2.

Figure 3.

Figure 4.