(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2022 年 8 月 4 日 (04.08.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/160700 A1

(51) 国际专利分类号:
G16B 30/00 (2019.01) *G16B 20/00* (2019.01)

(21) 国际申请号: PCT/CN2021/115146

(22) 国际申请日: 2021 年 8 月 27 日 (27.08.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110131330.4 2021年1月30日 (30.01.2021) CN

(71) 申请人: 中国科学院分子植物科学卓越创新中心 (CAS CENTER FOR EXCELLENCE IN MOLECULAR PLANT SCIENCES) [CN/CN]; 中国上海市徐汇区枫林路 300 号 4 号楼, Shanghai 200032 (CN)。

(72) 发明人: 韩斌 (HAN, Bin); 中国上海市徐汇区枫林路 300 号 4 号楼, Shanghai 200032 (CN)。 朱舟 (ZHU, Zhou); 中国上海市徐汇区枫林路 300 号 4 号楼, Shanghai 200032 (CN)。 王阿红 (WANG, Ahong); 中国上海市徐汇区枫林路300号4号楼, Shanghai 200032 (CN)。

(74) 代理人: 上海一平知识产权代理有限公司 (XU&PARTNERS, LLC.); 中国上海市普陀区真北路 958 号天地科技广场 1 号楼 106 室, Shanghai 200333 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告 (条约第21条(3))。

(54) **Title:** GENOTYPE IDENTIFICATION OF MULTI-PARENT CROP ON BASIS OF HIGH-THROUGHPUT WHOLE GENOME SEQUENCING

(54) 发明名称: 基于高通量全基因组测序的多亲本作物基因型鉴定

(57) **Abstract:** Genotype identification of a multi-parent crop on the basis of high-throughput whole genome sequencing. Specifically, the method comprises: (a) for n parents and the progeny thereof, providing sequencing data Df of a progeny crop to be identified and sequencing data Dp of a parent crop corresponding to the progeny crop, with n being a positive integer of ≥ 3; (b) determining SNP site information of the parent and the progeny on the basis of the sequencing data Df and the sequencing data Dp; (c) determining the genotype of the progeny on the basis of the SNP site information, thereby obtaining an evaluation result of each SNP of the progeny and distribution information of the recombinant breaking point on each chromosome of the whole genome of the progeny; and (d) constructing and/or drawing a genotype map of the progeny, thereby obtaining a genotype identification result of a multi-parent crop. The method can be used for identifying the genotype of the multi-parent crop with high throughput, rapidness and accuracy.

(57) 摘要: 一种基于高通量全基因组测序的多亲本作物基因型鉴定。具体地，包括：(a)对于n个亲本及其子代，提供待鉴定的子代作物的测序数据Df，以及与所述子代作物相应的亲本作物的测序数据Dp，其中n为≥3的正整数；(b)基于所述测序数据Df和所述测序数据Dp，确定亲代和子代的SNP位点信息；(c)基于所述的SNP位点信息，对子代的基因型进行判断，从而获得所述子代的各个SNP的评定结果以及所述子代的全基因组的各染色体上的重组断裂点的分布信息；(d)构建和/或绘制所述子代的基因型图谱，从而获得所述多亲本作物的基因型鉴定结果。本方法可高通量、快速、准确的对多亲本作物基因型进行鉴定。

WO 2022/160700 A1

GENOTYPE IDENTIFICATION OF MULTI-PARENT CROP ON BASIS OF HIGH-THROUGHPUT WHOLE GENOME SEQUENCING

TECHNICAL FIELD

The present invention relates to the technical field of biological information processing, specifically to genotype identification of multi-parent crop on basis of high-throughput whole genome sequencing. More specifically, the present invention provides a method and device for identifying the genotype of multi-parent crops based on high-throughput whole genome sequencing data.

BACKGROUND OF THE INVENTION

At the end of the last century, the use of DNA molecular markers greatly promoted the development of reverse genetics. With advances in molecular biology technology, the types of markers and methods of constructing genetic maps are also developed and perfected step by step. The occurrence of polymerase chain reaction (PCR) causes an era of a molecular marker explosive application, because PCR can greatly simplify the experimental steps of marking design and result analysis. These DNA molecular markers are still widely used, but also show more and more limitations in genome coverage, time consumption, and expense. Currently, the development of genomics and the gradual maturation of related technical methods provide a basis for replacing a labeled-based mapping method for a genome-based high-throughput policy.

The genomic sequence opens gates of high-throughput genotyping. Initially, this was achieved using microarray chip technology, which hybridized genomic DNA with oligonucleotides on a gene chip to detect single nucleotide polymorphisms (SNPs). Due to one-time hybridization, hundreds to thousands of markers can be detected, and this genotype identification method fully improves efficiency [1]. This method is applied to some mode biological systems such as humans, Arabidopsis, and rice [2-4]. While high throughput targets have been achieved, methods based on micro-arrays also have severe limitations, such as laborious, time-consuming, and design, production, and

high expense in the use of microarray processes.

The appearance of the second-generation sequencing technology results in a method-learned leap progress for genotype identification and genetic mapping. The new sequencing technique not only increases the sequencing flux of several orders, but also allows for parallel sequencing of many samples [5-6]. The progress of these technologieshas paved the way for the development of a high-throughput genotype identification method based on sequencing. The new genotype identification method combines the following advantages: fast and cheap, high density marker coverage, high accuracy and high resolution, and is also suitable for constructing comparative genomes and genetic maps between more mapping populations and species.

Although some methods have been available for the genotype identification of two parent plants, for the identification of multi-parent plant genotypes involving 3 or more parents, the present methods have significant deficiencies, for example, lower accuracy, long analysis time consumption, and the like.

Therefore, there is an urgent need in the art to provide a method and a device for identifying the genotype of a multi-parent plant related to more than 3 parents quickly and accurately.

SUMMARY OF THE INVENTION

The object of the present invention is to provide a method and a device for identifying the genotype of a multi-parent plant with rapid analysis and accurate results, so that the population genotype constructed by the multi-parent can be quickly and accurately analyzed.

In the first aspect of the present invention, a method for identifying the genotype of a multi-parent plant (such as a crop) is provided, wherein the method comprises:

(a) for n parents and the progeny thereof, providing sequencing data Df of a progeny plant to be identified and sequencing data Dp of a parent plant corresponding to the progeny plant, with n being a positive integer of ≥ 3;

(b) determining SNP site information of the parent and the progeny on

the basis of the sequencing data Df and the sequencing data Dp;

(c) determining the genotype of the progeny on the basis of the SNP site information, thereby obtaining an evaluation result of each SNP of the progeny and distribution information of the recombinant breaking point on each chromosome of the whole genome of the progeny;

(d) constructing and/or drawing a genotype map of the progeny on the basis of the evaluation result information of SNP of the progeny and position information of the recombinant breaking point of the whole genome, thereby obtaining a genotype identification result of a multi-parent plant.

In another preferred embodiment, in step (c), the analysis of the recombinant breaking point is performed on the basis of the SNP "string".

In another preferred embodiment, step (c) includes analyzing the recombinant breaking point to obtain an analysis result of the recombinant breaking point,

and the recombinant breaking point analysis includes:

(s1) constructing an SNP "string", wherein genotypes of all SNPs on each chromosome of the parent and the progeny are compressed into a string in sequence;

(s2) determining each sliding window corresponding to the SNP string according to a predetermined window size, and scoring the SNP site in each window, so as to obtain a respective score value P of each parent in the window;

(s3) determining the genotype of each chromosome region corresponding to the progeny on the basis of the score value P obtained in the step (s2).

In another preferred embodiment, in step (s1), regardless of the actual distance between two adjacent SNPs, all the gaps between the SNPs are removed.

In another preferred embodiment, in step (s1), the SNP constituting the string is a parent homozygous SNP site.

In another preferred embodiment, in step (s1), the method further comprises: performing preliminary screening on the SNP sites that are included

in the analysis, so as to eliminate any SNP sites with heterozygous parents.

In another preferred embodiment, in step (s2), scoring is performed according to the scoring rule of table A.

In another preferred embodiment, in step (s3), in step (s3), for each chromosome region of the progeny, the genotype corresponding to each chromosome region of the progeny is determined on the basis of each parent score value or a score value curve.

In another preferred embodiment, in step (s3), the genotype of each chromosome region is determined on the basis of the score value and the standard deviation.

In another preferred embodiment, for the chromosome region of the genotype to be determined, if a certain parent A has a high score value ($\geq$ 80% perfect score, preferably $\geq$ 80% perfect score) close to the perfect score, and the parent is quite stable in score value in this region, and there is no too large numerical fluctuation, and the score value of the rest parents is low ($\leq$ 50% perfect score, preferably $\leq$ 30% perfect score) or there is a large numerical fluctuation, then the genotype of the chromosome region is determined as the genotype of the parent A..

In another preferred embodiment, in step (s3), comprising: the score value of each parent on each chromosome can be obtained by sliding the sliding window on the whole genome SNP site, and the score value is taken as the ordinate, with the position of each sliding window on the chromosome as the abscissa, and the score curve of each parent is drawn.

In another preferred embodiment, in step (S3), comprising a sub-step of evaluating the hybrid region:

(s3a) for a multi-parent hybrid progeny, it is still set that the parent source of a certain segment of chromosome region is two parents at most, and whether the segment of region is a hybrid region is determined based on the score curve of the two parents in the segment of region.

In another preferred embodiment, in step (s3), the similarity between the progeny and each parent in the segment is quantized, and the genotype of each

segment is determined according to the numerical features (value size and standard deviation) of the score curve of each parent.

In another preferred embodiment, in step (s3), genotype assessment is performed in the following manner:

(Z1) if a certain parent is higher in score value in the section and the score thereof is relatively stable (the score curve is close to the platform period), meanwhile, the score curve of the rest parents in the section has large fluctuation, the standard deviation is large (the score curve is up and down fluctuation of the peak shape), and it is judged that the section is the homozygous genotype of the parent;

(Z2) when it is determined that the number of parents is two, it is possible to infer that the region is a hybrid genotype of two parents according to a large numerical fluctuation of both parents in a certain section and a large standard deviation;

at the time of multi-parent determination, it is likely that only possible hybrid section can be found, that is, a high-score and stable parent cannot be found in a certain section; because the score of each parent has a large fluctuation in the section, it is only possible to give the most likely two parents according to the numerical features.

(Z3) If there are two or more parents similar to the progeny in a certain section with two or more parent curves with high scores and smaller standard deviation appearing, it is indicated that some of the analyzed parents are very similar in this section without much difference, and the judgment can be temporarily withheld (marked as an “unknown area”).

In another preferred embodiment, the method further comprises: if the genotypes on both sides of the unknown region are the same, the region can be determined as the genotype; and if the genotypes on both sides are different, the middle position of the unknown region can be regarded as a recombinant breaking point, and the two sides of the unknown region being genotypes on both sides, respectively.

In another preferred embodiment, the progeny is a multi-parent plant.

In another preferred embodiment, n is 3 -6, and more preferably, 3, 4, or 5.

In another preferred embodiment, the sequencing data is selected from the following group: genome sequencing data, RNA sequencing data, and a combination thereof.

In another preferred embodiment, the sequencing data is a fastq format file.

In another preferred embodiment, the size of the sliding window is 170-500 consecutive SNP sites, preferably 200-400 consecutive SNP sites;

and/or the sequencing depth of the sequencing data is 0.1x-10x, preferably 0.2x-5x.

In another preferred embodiment, the sequencing depth of the sequencing data: ≥1, preferably 1-5, and more preferably 1.5-3.

In another preferred embodiment, for each chromosome, each parent score curve is obtained.

In another preferred embodiment, the SNP site is used to determine the genotype of the individual.

In another preferred embodiment, in step (b), sequencing data (such as a fastq file) is compared, and SNP information is obtained through bwa and GATK software processing.

In another preferred embodiment, the SNP site information includes location information and genotype information.

In another preferred embodiment, the SNP site for determining the genotype satisfies the following requirements:

first, the SNP site covers the whole genome as much as possible, and does not have a deletion in some regions;

second, for any one of SNP site, the SNP information (position information and genotype information) of two corresponding parents and progeny are both known, and if any one of the three is unknown, the site should be deleted.

In another preferred embodiment, in step (c), the evaluation result of each SNP of the progeny is recorded in a rlt file, and the rlt file records the genotype

determination condition of each SNP position;

the distribution information of the recombinant breaking points on the chromosomes of the whole genome of the progeny is recorded in the bin file, and the bin file records the distribution condition of the recombinant breaking points on 12 chromosomes of the whole genome.

In another preferred embodiment, in step (c), the SNPWindow script is used to read the genotype and determine the recombinant breaking point.

In another preferred embodiment, in step (d), a genotype map is performed on m individuals of the progeny at the same time.

In another preferred embodiment, in step (d), the construction of the recombination map is carried out through the SNPWindow script, and the gene map of each progeny individual is drawn by using the SNP2png script.

In another preferred embodiment, in step (d), further comprises that a recombination map of each individual is league matched by means of a *Bin2MCD* script to generate a recombination bin map.

In another preferred embodiment, the resolution of the recombination bin map is one bin per 5-200kb, preferably one bin per 10-100kb.

In another preferred embodiment, the method further comprises that: processing the recombination bin map to obtain a genetic map of the progeny.

In another preferred embodiment, the method further comprises: performing QTL analysis on the genetic map.

In another preferred embodiment, the method further comprises: performing visual analysis on the genotype of the whole population of the parent and the progeny, generating genotype data, and constructing a linkage map based on the genotype data.

In another preferred embodiment, the plant comprises crops, preferably gramineous crops.

In another preferred embodiment, the crops comprises rice, wheat, soybean, and tobacco.

In the second aspect of the present invention, it is provided a data analysis

device for identifying the genotype of a multi-parent plant, and the device comprises:

a data input module, which is configured to input data to be processed to be analyzed, the data to be processed comprising: a sequencing data Df of a progeny plant to be identified, and a sequencing data Dp of a parent plant corresponding to the progeny plant;

a multi-parent plant genotype identification module, which is configured to execute the method of the first aspect of the present invention, so as to obtain a genotype identification result of the progeny;

and an output module, which is configured to output the genotype identification result of the progeny.

In another preferred embodiment, the multi-parent plant genotype identification module comprises:

an SNP site information analysis sub-module, which is configured to determine SNP site information of a parent and a progeny based on the sequencing data Df and the sequencing data Dp;

a chromosome recombinant breaking point analysis sub-module, which is configured to determine the genotype of the progeny on the basis of the SNP site information, so as to obtain the evaluation result of each SNP of the progeny and the distribution information of the recombinant breaking points on each chromosome of the whole genome of the progeny;

a genotype map construction sub-module, which is configured to construct and/or draw the genotype map of the progeny based on the SNP evaluation result information of the progeny and the position information of the whole genome recombinant breaking point, so as to obtain the genotype identification result of the multi-parent plant.

In another preferred embodiment, the plant includes crops, preferably gramineous crops.

In another preferred embodiment, the output module comprises: a display, a printer, a pad, and the like.

It should be understood that, in the scope of the present disclosure, the above-mentioned technical features of the present disclosure and the technical features specifically described in the following (for example) may be combined with each other, thereby constituting a new or preferred technical solution. It is not limited to a lute, which is not yet described again herein.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG.1 shows a simulated two-parent material whole genome recombinant breaking point.

FIG.2 shows a simulated four-parent material whole genome recombinant breaking point.

FIG.3 shows genotype identification of two-parent simulated progeny using SNP-based sliding window methods.

FIG.4 shows the genotype identification of two-parent simulated progeny using SEG-MAP software.

FIG.5 shows genotype identification of a four-parent simulated progeny using an SNP-based sliding window method.

FIG.6 shows the effect of different sliding window sizes on the accuracy of genotype determination results.

FIG.7 shows the effect of different sequencing depths on the accuracy of genotype determination results.

FIG.8 shows the analysis framework flow of the SNP-based sliding window genotype identification method.

FIG.9 shows the drawing of the gene map using the SNP2png script.

FIG.10 shows a set map of genotype identification of a rice population.

FIG.11 shows a genotype table of a recombinant inbred line individual recombinant section map in one embodiment.

FIG.12 shows an SNP "string" with a window of 15.

FIG.13 shows four parent score curves of the simulated progeny of the rice chromosome 3.

FIG.14 shows two parent score curves of the simulated progeny of the rice chromosome 11.

FIG.15 shows the scores of each parent when it is determined as a parent tri-homozygous genotype in one embodiment.

FIG.16 shows the scores of each parent when it is determined as two-parent hybrid genotype in one embodiment.

FIG.17 shows the scores of each parent when the genotype is determined to be unknown in one embodiment.

FIG.18 shows a subsequent genotype determination of the unknown region in one embodiment.

FIG.19 shows the genotype identification map of a single individual in the DH population.

FIG.20 shows a set map of genotype identification in DH population.

FIG.21 shows the genotype identification of the tree-parent material in one embodiment.

FIG.22 shows SEG-Map identification of a tree-parent material in one embodiment.

FIG.23 shows the genotype identification of a four-parent simulated material in one embodiment.

FIG.24 shows a true genotype of a four-parent simulated material in one embodiment.

DETAILED DESCRIPTION OF THE EMBODIMENTS

The inventor develops a more rapid and accurate genotype identification method for the first time through extensive and intensive research, thereby realizing more effective genetic mapping and genome analysis. The method of the present invention is particularly suitable for genotype analysis and identification of a multi-parent population of low coverage sequencing. In the present invention, the genotype information of the real SNP of the multi-parent and the progeny in a certain section is directly read, and then the similarity between the progeny and each parent in this section is quantified, and an efficient, simplified and accurate multi-parent plant (or multi-parent crop) genotype identification method is formed according to the numerical features (value size and standard deviation) of the score curve of each parent. The

present invention is completed on this basis.

Specifically, the inventor develops a high-throughput method to identify a recombinant population genotype containing a plurality of parents on the basis of whole-genome low-coverage sequencing data generated by a second-generation sequencing technology. The inventor designs a "sliding window" method, which determines the genotype of this segment by comprehensively analyzing the genotype of a plurality of single nucleotide polymorphism (SNP) in the local region of the genome, and further determines the specific position of the recombinant breaking point to construct a fine recombination map of the multi-parent population.

In order to verify the method, the inventor constructs simulated whole genome sequencing data of the two-parent population and the multi-parent population, constructs a genetic linkage map by using the method, compares the finally identified genotype information with a real simulated data genotype, and the genotype identification accuracy of the parent population can reach 89.61%, which is similar to the accuracy of identifying the parental population genotype by the SEG-Map software method of the present invention (the accuracy of the SEG-Map method is 89.32%). This genotype identification method newly developed by the present inventors can reach 92.10% for the identification accuracy of the multi-parent population, which cannot be achieved by SEG-Map software or methods.

According to the method, the genotype of each individual in the population can be effectively and quickly analyzed, a key guiding effect is achieved in genome design breeding, and rapid and accurate genotype data can also be provided for QTL positioning of different crop multi-parent populations. Meanwhile, the inventor tests the method by using a real rice RIL genetic group, uses genotype identification based on high-throughput sequencing, and finally obtains a quite good high-precision recombination map.

Therefore, with the continuous development of sequencing technology, the genotype identification method based on genome low-coverage sequencing can replace a traditional marker-based genotype identification method, and

provides a powerful tool for large-scale gene exploration and research and solving more complex biological problems. The method of the present invention is more suitable for genotype identification of a multi-parent backcross population with low coverage rate sequencing, provides accurate genotype support for QTL positioning, and also contributes to molecular design breeding application of a multi-parent population.

TERM

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.

As used herein, the terms “contain” or “include (comprise) " may be open, semi-closed, and closed. In other words, the term also includes “composed of…substantially”, or “composed of”.

As used herein, the term “two-parent” indicates that two parents are involved.

As used herein, the term “multi-parent” indicates that three parents and more parents are involved.

As used herein, the term “multi-parent plant” indicates a plant related to three parents and more parents, such as a progeny plant (eg., a crop) relating to 3, 4, or 5 parents.

Method for identifying genotype of multi-parent crop

The present invention provides a method for identifying the genotype of a multi-parent crop. The method of the present invention is a genotype identification method for an SNP site sliding window.

In the genotype identification method based on the SNP site sliding window of the present invention, data processing is optimized. The optimized process can directly analyze and process one-way or two-way terminal short-sequence sequencing results generated by the next-generation sequencing technology, and finally construct the genetic map of the recombinant group.

For a mapping group from two parents, before performing a data analysis

process, the SNPs in the whole genome range of the two parents need to be identified. The SNP identification work can be obtained from high-coverage whole genome deep sequencing, can also be obtained from existing genome SNP information in a rice haplotype map, and can also be obtained by combining low-coverage whole genome sequencing with deletion genotype (SNP) filling. Due to the fact that the SNP identification work between the two parent varieties can be obtained by means of rapid and money-saving approaches, the sequencing-based genotype identification of one recombinant group will mainly rely on subsequent analysis work, including reading genotype, determining a recombinant breaking point, and constructing a genetic linkage map.

Functions, steps, and software (scripts) in data analysis may refer to FIG.8.

The first step includes a number of tasks that can be processed simultaneously. Individuals and parent materials in a certain number of recombinant populations are subjected to second-generation high-throughput sequencing at the same time. The obtained fastq file is processed by bwa and GATK software to obtain high-quality SNP information.

The SNP site used for the final genotype determination should satisfy the following requirements:

1. The SNP site covers the whole genome as much as possible, and does not have a deletion in some regions;

2. For any SNP site, the SNP information (position information and genotype information) of two corresponding parents and offspring are both known, and if any one of the three is unknown, the site should be deleted.

In addition, by taking rice as an example, it is generally considered that the rice parent is a selfing homozygous line, and there is basically no heterozygous site in the genome, and therefore, if a hybrid SNP site is found in the parent, it is generally considered that the site is untrusted, and therefore, any SNP site of which the parent is heterozygous can be deleted.

After screening the high-quality whole genome SNP of the two parents and the progeny, a Python script SNPwindow can be used to determine the

genotype of the offspring. The script output may have two files, an rlt file, and a bin file. The rlt file records the genotype determination of each SNP position, and the bin file records the distribution situation of the recombinant breaking points on 12 chromosomes of the whole genome. These two files are important basis for subsequent mapping and chain analysis.

Referring to FIG.9, a genotype map is generally drawn using rlt and bin files through a perl script SNP2png, and the picture format is a PNG format. The map is drawn according to the determined SNP site genotype information and whole genome recombinant breaking point location information, and different colors are used in the map to represent different genotype types.

In actual work, a large number of descendant populations need to be processed, so that the SNP information of each individual can be extracted, and then the *SNPwindow* script is used for analysis and inspection, and the work of reading genotypes, determining the recombinant breaking points and constructing the recombination map is carried out. Firstly, *SNP2png* is used to draw the gene map of each individual, there is an overall grasp on the genotype of the individual, and then the script *BIN2MCD* is used to combine the recombination maps of all the individuals to generate a recombination bin map.

Referring to FIG.10, a perl script may also be used to visually analyze the genotype of the entire population. The programs and scripts used in the analysis flow are represented by italics, forming a series of analysis steps. The genotype data generated by the analysis process may be directly used for other software (including Mapmaker and JoinMap) to construct the linkage map.

When the output data finally generated by the software is analyzed, the bin is recombined, and the resolution is usually one bin per 100 kb, or even one bin per 10 kb. The genotype result of the mapping population can be imported into a program such as Mapmaker [16] or JoinMap [32] to construct the genetic map. The genetic map is subjected to QTL analysis.

Referring to FIG.11, the genetic map produced by the method of the present invention has a much finer scale than that generated by most traditional molecular markers.

The method of the present invention relates to determining a recombinant breaking point, and its detailed process includes the following steps:

Step 1: Constructing an SNP "String".

An example is taken as a window size of 15 (win=15). The genotypes of all SNPs on two parents and 12 chromosomes are compressed into a string in sequence. Regardless of the actual distance between two adjacent SNPs, all the gaps are removed.

By taking 12 chromosomes as an example, SNP on 12 chromosomes becomes 12 consecutive strings (see FIG. 12). In the figure, the blue represents the genotype of the parent I, the red represents the genotype of the parent II, and for one parent offspring, there are three conditions possible at each SNP site of the parent, the homozygous genotype (blue) of the parent I, the homozygous genotype (red) of the parent II, and the heterozygous genotype (yellow).

It is generally considered that the genome of the artificially cultivated rice parent material is highly pure, and at the same time, for some multi-generation selfing recombinant rice populations, the genome thereof is also homozygous, and only some heterozygous areas exist at some chromosome positions. Therefore, human screening is performed on the SNP sites included in the analysis, any SNP sites of which the parents are heterozygous are excluded, and the sites are not accurately judged and scored. In addition, if the sequencing depth of the progeny is not very high, the SNP site with the progeny being heterozygous can be filtered, because the reliability of the hybrid site determined according to the low depth is not high, it is likely to be a misjudgment caused by a sequencing error.

Step 2: The two parents are scored in one window

Calculating the score of all SNP site in a sliding window according to the Mendelian inheritance laws, and calculating the sum score of each parent as the

score of the parent in the chromosome position where the sliding window is located. Measure the degree of conformity between the progeny and the parent according to the scoring of each parent. Scoring rules are shown in table A or similar scoring rules. Preferably, the scoring rule is formulated according to the genetic law of the organism.

In combination with the theoretical model, the genotype scoring rule in the present invention is further described below.

For a sliding window composed of continuous SNPs, the score value of the progeny to any parent is composed of three parts: 1. SNP site that is the same as the parent; 2. site that is different from the parent but comply with Mendelian inheritance laws; 3. site that is different from the parent and do not conform to Mendelian inheritance laws; and misjudgment site that iscaused by various possible factors.

For a single parent A, the number of SNP sites in the progeny to be detected that are the same as the parent is m, the number of site in the progeny that are different from their parents but comply with the Mendelian inheritance laws is n, and the number of sites that are different from the parent and do not conform to Mendelian inheritance laws and misjudgment sites that are caused by various possible factors is e.

Then, the parent 's score value S_A is :

$$S_A = s_1 * m + s_2 * n + s_3 * e$$

wherein, S_1 is the scoring value of the SNP site of the progeny that is the same as the parent.

S_2 is the scoring value of the site of the progeny that is different from the parent but conforms to the Mendelian inheritance laws.

S_3 is the scoring value of the sites of the progeny that differ from their parents and do not conform to Mendelian inheritance laws, and misjudgment sites caused by various possible factors.

In a continuous SNP box with a size of N, there are i parents to be determined, for the chromosome position k of a given SNP site, the progeny and the parental genotypes on the site are respectively g_k and g'_k. The genotype

of the gene of the rice pure-line parent is generally 0/0, 0|0, 1/1, 1|1, and the genotype of the progeny is generally 0/0, 0|0, 1/1, 1|1, 0/1, 0|1. The frequency of occurrence of alleles of 2 is generally low, so it is not considered temporarily

For the i th individual parent, the probability that the progeny conforms to its genotype is :

$$P_i(g_k, g_k') = C_N^{m+n}\left(\frac{2}{3}\right)^{m+n} * \left(\frac{1}{3}\right)^{e}$$

Use Bayesian methods to find the likelihood that the region of the progeny belongs to that parent:

$$P_i = \sum_{k=1}^{N} P_i(g_k, g_k')$$

Determine the genotype of a certain region of the offspring to the parent genotype with the highest probability of coincidence, that is, calculate the maximum probability of coincidence among the i parents:

$$P_{max} = max\{P_1, P_2, \ldots, P_i\}$$

Preferably, in the present invention, the following table is used for genotype scoring.

Table A Genotype scoring rules table

Progeny genotype	Parental genotype	Scoring
0/0	0/0	+1
1/1	1/1	+1
0/0	1/1	+0
1/1	0/0	+0
0/1	0/0,1/1	+1
0/2	0/0,2/2	+1
1/2	1/1,2/2	+1

According to scoring rules, S_1=1, S_2=1, S_a=0.

If the parent A is scored, obtaining:

$$S_A = \sum_{k=1}^{N} F(g_k, g_k')$$

According to the method, after the score value of each parent is counted

respectively, a standard deviation std is calculated by sliding window for the continuous parent score value.

Preferably, when it is determined that a certain region of the chromosome is a homozygous genotype of A parent, the following conditions need to be met: the score S is the highest, and the standard deviation is minimum.

Step 3: Determining the genotype of the chromosome region according to the score value.

By sliding the sliding window on the whole genome SNP site, the score value of each parent on each chromosome can be obtained. The score is the ordinate, and the position of each sliding window on the chromosome is taken as the horizontal coordinate, and the score curve of each parent is drawn.

The genotype of each segment of chromosome is determined according to features of different parental score curves.

In one example, for example, referring to FIG.13, a sliding window scoring is performed on a progeny from simulating four-parent, and a score curve of four parents is drawn according to the four-parent score values.

The score curves of the four parents are observed, and in one example, it can be seen that the parent curve has two different distribution modes of the platform stabilization period and the fluctuation period in different areas of the 11 chromosome. Therefore, the genotype of the progeny in the region is determined by using the states of different parental curves in the same region.

As in the 1 bp to 10780000 bp region in the figure, it can be observed that the score of the yellow curve (the parent 4) in the region has a high score value close to the perfect score, and the score value of the parent in the region is quite stable, there is no too large numerical fluctuation, and the fluctuation of the score is measured by using the standard deviation in statistics. In this region, the parent 4 has a very high score value and a relatively small standard deviation in the region, and at the same time, the score value of the other three parents in the region fluctuates up and down within the range of 0-200, which has a very high standard deviation, and therefore, it can be determined that the

region is a homozygous genotype of the parent 4. By using a similar method, progeny genotypes of different regions of 12 chromosomes can be determined according to score values.

The rectangular strip corresponding to "true" in the figure corresponds to the real genotype information of the simulated progeny of each chromosome segment, and the rectangular strip corresponding to the "judge" represents the progeny genotype determined by the method of the present invention, and the information of the two is basically consistent.

Determination of Hybrid Area

The determination of the hybrid region is illustrated by the genotype determination of the two-parent simulated progeny. According to the genetic principle, even if a hybrid progeny derived from a multi-parent, it comes from at most two parents in a certain segment of chromosome region. Therefore, whether the region is a hybrid region can be determined according to the score curve of the two parents in the region.

In one example, as shown in FIG.14, genotype identification is performed on the simulated progeny, the rectangular strip corresponding to "true" corresponds to the real genotype information of the simulated progeny of each chromosome segment, and the rectangular strip corresponding to the "judge" represents the progeny genotype determined by the method of the present invention. Similarly, when the score of one parent is relatively high and the standard deviation is small and the score of the other parent has relatively large fluctuation, and when the standard deviation is large, it is determined that the section of region is a homozygous genotype of the former (orange or blue region in the figure.).

When the relative scores of the two parents in a certain section (60000 bp to 15000000 bp) are not large, and both have a certain fluctuation, when the standard deviation is large, this section is determined as the hybrid genotype of the two parents (the yellow region in the figure). This determination result is also consistent with the real information.

A determination is made based on the degree of similarity of quantization

One of the core ideas of the method of the present invention is based on directly reading the genotype information of the real SNP of the parent and the progeny in a certain section, then quantifying the similarity between the progeny and each parent in this section, and forming a relatively simplified analysis model according to the numerical features (value size and standard deviation) of the score curve of each parent, so as to determine the genotype of each section.

The criteria of the determination of the present invention mainly include the following several cases:

1. If a certain parent is higher in score value and stable in score (the score curve approaches the platform period), meanwhile, the score curve of the rest parents in the section has large fluctuation, the standard deviation is large (the score curve is up-and-down fluctuation of the peak shape), and it is judged that the section is the homozygous genotype of the parent (FIG.15).

2. When it is determined that the number of parents is two, it is possible to infer that the region is a hybrid genotype of two parents (FIG.16) according to a large numerical fluctuation of both parents in a certain section and a large standard deviation.

At the time of multi-parent determination, it is likely that only possible hybrid section can be found, that is, a high-score and stable parent cannot be found in a certain section. Because the score of each parent has a large fluctuation in the section, it is only possible to give the most likely two parents according to the numerical features.

3. Since the determination method depends on the degree of similarity between the progeny and the parental genotype to a great extent, if there are two or more parents similar to the progeny in a certain section, two or more parent curves with high scores and smaller standard deviation appear, which indicates that some parents analyzed are very similar in this section, there is no

too large difference, and it can be temporarily not determined.

Referring to FIG.17, in some cases, the region to be determined is first defined as "unknown", while the genotype of the region is determined by the genotype of the regions on both sides.

Referring to FIG.18, if the genotypes on both sides are the same, the region is determined as the genotype, and if the genotypes on both sides are different, the middle position of the region is regarded as a recombinant breaking point, and both sides of the region are genotypes on both sides.

Two sliding windows for genotype determination

In the present invention, it is preferable to perform genotype determination by means of a secondary sliding window.

A sliding window is performed on the genotype of the SNP for the first time, and the parent score value in each window is counted.

A sliding window is performed on the obtained parent score value for the second time, and the height and the standard deviation of the score value are detected.

The final genotype determination depends on the score and standard deviation obtained from the two sliding windows, and is determined by the highest probability of a certain region of the offspring belonging to a certain parent.

In the present invention, genotype determination can be performed more quickly and accurately by using the two sliding windows.

A schematic diagram of the two sliding windows is shown in the following:

Parent	offspring	score	std
00	00	199	0.465
11	11	199	0.474
00	00	198	0.491
11	11	197	0.423
.....		.....	.....
00	01	199	0
11	11	199	0
11	11	199	0
11	12	199	0
00	01	199	0
11	11	199	0
11	11	199	0
11	11	199	0

Identification device for genotype of multi-parent crop

The present invention also provides an identification device or an analysis device for a multi-parent crop genotype for performing the method of the present invention. Typically, the device includes:

a data input module, which is configured to input data to be processed to be analyzed, the data to be processed comprising: a sequencing data Df of a progeny plant to be identified, and a sequencing data Dp of a parent plant corresponding to the progeny plant;

a multi-parent plant genotype identification module, which is configured to execute the method of the present invention, so as to obtain a genotype identification result of the progeny;

and an output module, which is configured to output the genotype identification result of the progeny.

The main advantages of the present invention include:

(a) The present invention provides a multi-parent crop genotype identification method based on high-throughput sequencing data for the first time, and there is no systematic crop multi-parent genotype identification

method prior to the present invention.

(b) The high-throughput genotype identification method of the present invention can greatly simplify and accelerate the genetic positioning of quantity traits in crops [37 -39, 20].

(c) The theoretical method of the present invention can better match a multi-parent population for genotype identification, improve the accuracy and efficiency of QTL positioning, and make full use of rich genetic variation present in a multi-parent population. Meanwhile, the improvement of crop genetic quality and molecular breeding design are also facilitated.

(d) In practical application, the present invention can be used for obtaining molecular markers closely linked to important agronomic trait genes, efficient screening of offspring during breeding, fine identification of improved variety genotype maps, etc., and provides a fast and efficient means and platform for molecular marker-assisted screening breeding, so as to improve the efficiency and accuracy to a new step..

In summary, the sequencing-based high-throughput genotype identification method of the present invention will provide convenience for solving complex biological problems and crop breeding improvement.

The present invention is further described below with reference to specific examples. It should be understood that these embodiments are merely used to illustrate the present disclosure and are not intended to limit the scope of the present disclosure. The following examples do not indicate the experimental methods of specific conditions, usually according to conventional conditions, or according to the conditions suggested by manufacturers. Unless otherwise stated, the percentages and parts are weight percentages and parts by weight.

Example 1

1.1 Whole Genome Simulation Data Fabrication Based on Real Sequencing Data

1.1. 1. Rice Material and Simulated Data

Using Laboratory Existing High Depth Real Rice Material Indica Rice 93 -11 (*Oryza Sativa SSP. indicca cv.* 93 - 11), Shuohui 70, the Wushansimiao and Huangzhanhua as parent materials, and the simulated progeny fastq data is obtained after screening and combining through comparison, screening and combination of the real fastq data of the parent material. In the simulation data of the two parents, the inventor uses 93-11 and Wushansimiao as simulated parents, and each chromosome of the rice simulates the respective homozygous regions of the two parents and the heterozygous region superimposed by the two parents, so as to test the determination of the recombinant breaking point and the determination of the hybrid region of the method of the present invention. In the multi-parent simulation data, 93 -11, Shuohui 70, Wushansimiao and Huangzhanhua are used as simulation parents, 100 times of data simulation is performed, and recombinant breaking points of different combinations are simulated on each chromosome of rice. Each chromosome of the rice has an average of 4 -6 recombinant breaking points, the whole genome has a total of 50-60 recombinant breaking points, and the purpose is to simulate the recombination condition inside the genetic group of the rice multi-parent source as much as possible and verify the accuracy of the method of the present invention.

1.1.2. Identification of Simulated Data SNP

Comparing the four parental sequencing data of 93-11, Shuohui 70, Wushansimiao and Huangzhanhua with complete sequence 12 chromosomes (http://rice.pantbiology.msu.edu/annotation_pseudo_current.shtml) IRGSP 1.0 of japonica cv. Nipponbare which sequenced by International Rice Genome Sequencing Project (IRGSP), and the comparison software is bwa 0.7.17-r1188 [13]. The candidate SNPs of the parent and the simulated progeny were then identified with a Haploypealler program (parameter is -ERC GVCF) in the GATK software package [14]. After obtaining the variant intermediate file g.vcf file of each parent and the simulated progeny, using the GenomicsDBImport program in the GATK software package to merge all the variant intermediate

files, and then using the GenotypeGVCFs program in the GATK software package to derive the merged variant file, using the SelectVariants program to select the required SNP site information, and then filtering all the SNP site by using a VarianttFiltration program (the parameters are --cluster-size 3 -- cluster-window-size 10 -- filter-expression "QD <10.00" –filter-namelowQD –filter-expression "FS > 15.000" –filter-name highFS –genotype-filter-expression "DP > 50 | | DP < 5" –genotype-filter-nameInvalidDP) to obtain a high-quality SNP site. Moreover, genotype identification is performed on the simulated progeny on this basis.

1.2 Program Development of Genotype Identification Process Based on Sequencing

In the process of sequencing-based genotype identification, massive data needs to be processed, multiple different algorithms are applied, and some existing software, such as sequence matching software and QTL analysis software, is used. Therefore, the present inventors have developed a plurality of perl and python scripts to implement the steps described above and make them a complete and easy-to-use process.

After the SNP information of the parent and the progeny is obtained through GATK software, through a Python script, the principle is that the SNP region identified by each individual is subjected to comprehensive analysis along the windows sliding along all the SNP sites, the genotype is read based on a sliding window with a fixed length, and then the judgment of the recombinant breaking point and the construction of the recombination section map are carried out. In addition, by using an intermediate file determined by a Perl script using a program, a PNG format recombinant section map is generated for each individual, thereby facilitating the user to intuitively browse the overall genotype thereof. When plotted, the GD module in the Perl needs to be used.

Next, another script *Bin2MCD* generates a high density map composed of recombinant bin [19] to facilitate subsequent QTL analysis. Once the phenotype

is evaluated and the trait data is prepared, the output file may be directly used by some QTL analysis software packages to identify the QTL, including Windows QTL Cartographer V2.5 [17].

1.3 Genotype Identification Based on Rice DH Population

1.3. 1 Rice DH Population and Tree-parent Population

The rice DH population used in the present research was constructed by Laboratory of National Gene Research Center, Chinese Academy of Sciences, the two parents of which are Kasalath and Japonica *Japonica* cv. Nipponbare. The DH group is a strain generated by the F2 offspring after many years of selfing recombination. The inventor selects dozens of strains to perform genotype identification and analysis. The rice tree-parent plants used in the present research are constructed by Laboratory of National Gene Research Center, Chinese Academy of Sciences. Each of the three parents is Wushansimiao, 93-11, and Shuohui 70. The plants in the population are generated by self-recombination of hybrids offspring of three parents, and there are many recombination information in the genome.

1.3.2. Genotype identification is performed on a rice DH population and a tree-parent population by using a sequencing-based method.

By using the method of the present invention, genotype identification is performed on a DH population of rice, and a high-density map composed of a recombinant bin is generated by means of *Bin2MCD*. At the same time, in order to measure the accuracy of the method, genotype analysis and high-density bin maps are also performed by using the method published in 2010. The high-depth (20-30x) sequencing data of the two parent Kasalath and Nipponbare are compared to the reference genome IRGSP 1.0 by bwa software, then the GATK software is used to respectively find the information of the high-quality SNPs of the two parents, and then the SNP of the specified site is replaced on the Nipponbare genome through a perl script, so that the pseudo reference of the two parents is generated. Then, the low-abundance sequencing

data of the DH population is compared to the pseudo reference of the two parents, and then genotype identification is performed.

Result

2.1 Based on Real Parent Rice Multi-parent Simulation Data and Genotype Identification

2.1. 1. Simulation Data of Rice Genome Information

As shown in FIG.1, the expected genotype information of the two parent-derived progeny simulated by the inventor should be consistent with the expected genotype information. The manufactured simulation data includes three cases: a homozygous region of Wushansimiao, a homozygous region of 93-11, and a hybrid region of the Wushansimiao, 93-11. The corresponding length of each region and the position of the recombinant breaking point are shown in the figure. The simulation data is made on the basis of the real sequencing data of the two parents. Firstly, the fastq data of the two parents are compared to the rice Nipponbare genome, and then the required fastq information is screened according to the comparison information (chromosome and position information) in the obtained sam file, and then these fastq data derived from the two parents reform the simulated hybrid progeny fastq data.

As shown in FIG.2, a similar method is adopted, and simulated progeny data derived from four parents Wushansimiao, Huangzhanhua, 93-11, and Shuohui 70 are manufactured. It is expected that the genome genotype information and the recombinant breaking points conform to the information in the graph.

2.1.2. Genotype Identification of Simulation Data

The fastq data of the simulated progeny is compared with the fastq data of the two parents Wushansimiao and 93-11 to the rice reference genome IRGSP 1.0, and then the GATK software is used to find the whole genome variation information of the two parents and the simulated progeny, and the high-quality SNP site are obtained by filtering and screening.

As shown in FIG.3, after the required SNP is obtained, the SNP of the whole genome is judged by using a "sliding window" method, the two parents are scored and compared in one sliding window, and if a certain parent is higher in score, the section is judged as the homozygous genotype of the parent (represented by red or blue in the figure). When the difference between the two parents is not large, it is determined that the segment is a hybrid region of the two parents (represented by a yellow color in the figure). The inventor designs a quantification method to measure and determine the accuracy, divides the whole genome into thousands of small regions (or 20-200 kb small regions) of 100kb, and then compares the result obtained by the method of the present invention with the standard map to measure the accuracy of the method of the present invention. According to this method, the obtained simulated data genotype information is compared with the real simulated data genotype, and the parent identification accuracy can reach 89.61%.

At the same time, the inventor also uses the published SEG-Map method to perform genotype determination on the simulated progeny data, the fastq file of the simulation data was aligned to the pseudo reference of the two parents, the parent-specific fastq sequence was screened out by using software, then the information of the SNP site according to the position of the sequence alignment was determined, and then the genotype information by using the sliding window method was determined. The method has more detailed theoretical verification and data simulation in published articles, and has high accuracy and feasibility. The accuracy was measured by using the quantized method, the accuracy of the SEG-Map software result is 89.32%, and the difference between the method and the method of the present invention is not large, which indicates that the method of the present invention has high feasibility and accuracy.

The SEG-Map method does have high credibility to the genotype identification of the two parents, and the inventor has also used the method to perform genome analysis of the rice material for a long time. However, the method cannot perform genotype identification on a multi-parent-derived

material, and therefore, the method of the present invention is also to solve the problem of multi-parent genotype identification. As shown in FIG.5, the inventor uses an SNP-based sliding window method to perform genotype identification on the simulated progeny of the four-parent source, scoring the four parents in one window, and if a certain parent score is the highest, the region will be determined as the homozygous region of the parent, and representing the homozygous regions of the four parents by using red, blue, green and yellow respectively. For the four-parent fastq data, 100 simulation is performed, and the accuracy of the method for dividing the genome into small regions is also used, and by comparing with the standard map, the method has an average simulation accuracy of 92.10% for genotype identification of simulated data of four parents.

2.1.3. Determination of Recombinant breaking points

When the "window" slides along the chromosome, the genotype is read by the scores of the two parent SNPs. One genotype is not varied prior to encountering a recombinant break point. The present inventors have found that there are two types of breaking points: one is separating two homozygous genotypes, and the other is separating a segment of homozygous genotype from a segment of heterozygous genotype; the previous situation is in the form of most of the existence in RIL, and most of the latter cases appear in the F_2 population. When a sliding window encounters a "homozygous/homozygous" breaking point, the homozygous genotype transiently becomes a heterozygous genotype, and then becomes homozygous genotype from the heterozygous genotype. When a sliding window encounters a "homozygous/heterozygous" breaking point, the homozygous genotype becomes a heterozygous genotype, and then becomes homozygous genotype from the heterozygous genotype again, and at this time, the boundary point of the homozygous genotype region and the heterozygous genotype region can be determined.

2.1.4. Influence of Different Window Sizes on the Accuracy of

Multi-Parent Determination

When using this sequencing-based method to perform genotype identification research, it is necessary to set appropriate analysis parameters, first considering whether the size of the sliding window will affect the accuracy of genotype detection, for example, each window in a given physical length contains many SNPs.

As shown in FIG.6, the invention adopts different window sizes for genotype analysis of the final SNP information obtained from the four-parent simulation data screening, and finds that the sliding window sizes of different sizes do affect the final analysis accuracy. when the size of the sliding window is small (less than 199), the final accuracy rate is less than 90%, but when the size of the sliding window is increased to 199, the accuracy rate of genotype identification can reach 93.72%, but when the size of the sliding window is increased continually, the final accuracy rate is not greatly changed, which indicates that the accuracy of the determination result is not increased along with the increase of the size of the sliding window. For program operation, the size of a larger sliding window needs more computing resources and operation time, and when a large-scale group needs to be processed, the time cost is more prominent. Therefore, the inventor comprehensively considers the time cost and the accuracy rate, and the size of the sliding window of 199 (or the size of the sliding window of 180 -220) is a relatively reasonable selection.

2.1.5. Influence of Different Sequencing Depths on the Accuracy of Multi-Parent Determination

Considering that the sequencing depth has an important influence on genotype identification and more accurate genotype identification can be performed with SEG-Map software at a lower sequencing depth, and therefore, a depth test is performed on the method of the present invention.

As shown in FIG.7, three different depths of 0.2x, 1.5x, and 3x were tested for genotype accuracy testing. Under the test of each depth, 100 fastq data simulation is performed, then genotype identification is performed according to

the variation information, and finally the accuracy of genotype identification is measured according to the coincidence program with the standard map. The results show that with the improvement of the depth, the accuracy of genotype identification is slightly improved, but the increased amplitude does not reach the expected amplitude, and the final genotype identification accuracy does not reach 95% or above. Therefore, the present inventors have made further improvements.

2.2 Genotype Identification Method Based on SNP Site Sliding Window

2.2.1. Main Steps of Data Analysis Flow

In order to make this new method of genotype identification based on sequencing widely available, the inventor arranges and optimizes the data processing flow. This flow can directly analyze and process one-way or two-way terminal short-sequence sequencing results generated by the next-generation sequencing technology, and finally construct the genetic map of the recombinant group.

For a mapping group from two parents, before performing a data analysis process, the SNPs in the whole genome range of the two parents need to be identified. The SNP identification work can be obtained from high-coverage whole genome deep sequencing, can also be obtained from existing genome SNP information in a rice haplotype map, and can also be obtained by combining low-coverage whole genome sequencing with deletion genotype (SNP) filling. Due to the fact that the SNP identification work between the two parent varieties can be obtained by means of rapid and money-saving approaches, the sequencing-based genotype identification of one recombinant group will mainly rely on subsequent analysis work, including reading genotype, determining a recombinant breaking point, and constructing a genetic linkage map.

Functions, steps, and software (scripts) in data analysis are shown in FIG.8. The first step includes a number of tasks that can be processed simultaneously.

Individuals and parent materials in a certain number of recombinant populations are subjected to second-generation high-throughput sequencing at the same time. The obtained fastq file is processed by bwa and GATK software to obtain high-quality SNP information. The SNP site used for the final genotype determination should satisfy the following requirements: 1.The SNP site covers the whole genome as much as possible, and does not have a deletion in some regions; 2. For any SNP site, the SNP information (position information and genotype information) of two corresponding parents and progeny are both known, and if any one of the three is unknown, the site should be deleted. 3. It is generally considered that the rice parent is a selfing homozygous line, and there is basically no heterozygous site in the genome, and therefore, if a hybrid SNP site is found in the parent, it is generally considered that the site is untrusted, and therefore, any SNP site of which the parent is heterozygous can be deleted.

After screening the high-quality whole genome SNP of the two parents and the progeny, a Python script *SNPwindow* can be used to determine the genotype of the progeny. The script output may have two files, an rlt file, and a bin file. The rlt file records the genotype determination of each SNP position, and the bin file records the distribution situation of the recombinant breaking points on 12 chromosomes of the whole genome. These two files are important basis for subsequent mapping and chain analysis.

As shown in FIG.9, a genotype map is generally drawn using rlt and bin files through a perl script *SNP2png*, and the picture format is a PNG format. The map is drawn according to the determined SNP site genotype information and whole genome recombinant breaking point location information, and different colors are used in the map to represent different genotype types.

In actual work, a large number of descendant populations need to be processed, so that the SNP information of each individual can be extracted, and then the *SNPwindow* script is used for analysis and inspection, and the work of reading genotypes, determining the recombinant breaking points and constructing the recombination map is carried out. Firstly, *SNP2png* is used to

draw the gene map of each individual, there is an overall grasp on the genotype of the individual, and then the script *BIN2MCD* is used to combine the recombination maps of all the individuals to generate a recombination bin map.

As shown in FIG.10, a perl script may also be used to visually analyze the genotype of the entire population. The programs and scripts used in the analysis flow are represented by italics, forming a series of analysis steps. The genotype data generated by the analysis process may be directly used for other software (including Mapmaker and JoinMap) to construct the linkage map.

When the output data finally generated by the software is analyzed, the bin is recombined, and the resolution is usually one bin per 100 kb, or even one bin per 10 kb. The genotype result of the mapping population can be imported into a program such as Mapmaker [16] or JoinMap [32] to construct the genetic map. The genetic map is subjected to QTL analysis.

As shown in FIG. 11, this genetic map is at a much finer scale than that produced by most traditional molecular markers. This software package is compatible with multiple platforms (eg, UNIX, Linux, and Windows). In addition to the perl environment itself, there is also a need to install a GD module because there is a drawing step in the process operation.

2.2.2. Detailed Process for Determining Recombinant breaking point

Step 1: Constructing an SNP "String".

An example is taken as a window size of 15 (win=15). The genotypes of all SNPs on two parents and 12 chromosomes are compressed into a string in sequence. Regardless of the actual distance between two adjacent SNPs, all the gaps are removed.

Therefore, the SNPs on 12 chromosomes become 12 consecutive strings (FIG. 12). In the figure, the blue shows the genotype of the parent I, the red represents the genotype of the parent II, and for one parent, there are three conditions possible at each SNP site of the parent, the homozygous genotype (blue) of the parent I, the homozygous genotype (red) of the parent II, and the heterozygous genotype (yellow).

It is generally considered that the genome of the artificially cultivated rice parent material is highly pure, and at the same time, for some multi-generation self-recombinant rice populations, the genome thereof is also relatively homozygous, and only some heterozygous areas exist at some chromosome positions. Therefore, human screening is performed on the SNP sites included in the analysis, any SNP sites of which the parents are heterozygous are excluded, and the sites are not accurately judged and scored. In addition, if the sequencing depth of the progeny is not very high, the SNP site with the progeny being heterozygous can be filtered, because the reliability of the hybrid site determined according to the low depth is not high, it is likely to be a misjudgment caused by a sequencing error.

Step 2: The two parents are scored in one window

Calculating the score of all SNP site in a sliding window according to the Mendelian inheritance laws, and calculating the sum score of each parent as the score of the parent in the chromosome position where the sliding window is located. Measure the degree of conformity between the progeny and the parent according to the scoring of each parent. The preferred scoring rule is shown in table A, and the scoring rule of the present invention is formulated according to the genetic law of organisms.

Table A Genotype scoring rules table

Progeny genotype	Parental genotype	Scoring
0/0	0/0	+1
1/1	1/1	+1
0/0	1/1	+0
1/1	0/0	+0
0/1	0/0,1/1	+1
0/2	0/0,2/2	+1
1/2	1/1,2/2	+1

Step 3: Determining the genotype of the chromosome region according to the score value.

By sliding the sliding window on the whole genome SNP site, the score value of each parent on each chromosome can be obtained. The score is the ordinate, and the position of each sliding window on the chromosome is taken as the horizontal coordinate, and the score curve of each parent is drawn. The genotype of each segment of chromosome is determined according to features of different parental score curves. In one example, for example, as shown in FIG.13, a sliding window scoring is performed on a progeny from simulating four-parent, and a score curve of four parents is drawn according to the four-parent score values.

The score curves of the four parents are observed, it can be seen that the parent curve has two different distribution modes of the platform stabilization period and the fluctuation period in different areas of the 11 chromosome. Therefore, the genotype of the progeny in the region is determined by using the states of different parental curves in the same region.

As in the 1 bp to 10780000 bp region in the figure, it can be observed that the score of the yellow curve (the parent 4) in the region has a high score value close to the perfect score, and the score value of the parent in the region is quite stable, there is no too large numerical fluctuation, and the fluctuation of the score is measured by using the standard deviation in statistics. In this region, the parent 4 has a very high score value and a relatively small standard deviation in the region, and at the same time, the score value of the other three parents in the region fluctuates up and down within the range of 0-200, which has a very high standard deviation, and therefore, it can be determined that the region is a homozygous genotype of the parent 4. By using a similar method, progeny genotypes of different regions of 12 chromosomes can be determined according to score values. The rectangular strip corresponding to “true” in the figure corresponds to the real genotype information of the simulated progeny of each chromosome segment, and the rectangular strip corresponding to the “judge” represents the progeny genotype determined by the method of the present invention, and the information of the two is basically consistent.

Step 4: Determining Hybrid Area

The determination of the hybrid region is illustrated by the genotype determination of the two-parent simulated progeny. According to the genetic principle, even if a hybrid progeny derived from a multi-parent, it comes from at most two parents in a certain segment of chromosome region. Therefore, whether the region is a hybrid region can be determined according to the score curve of the two parents in the region. As shown in FIG.14, genotype identification is performed on the simulated progeny, the rectangular strip corresponding to "true" corresponds to the real genotype information of the simulated progeny of each chromosome segment, and the rectangular strip corresponding to the "judge" represents the progeny genotype determined by the method of the present invention. Similarly, when the score of one parent is relatively high and the standard deviation is small and the score of the other parent has relatively large fluctuation, and when the standard deviation is large, it is determined that the section of region is a homozygous genotype of the former (orange or blue region in the figure.).

When the relative scores of the two parents in a certain section (6000000 bp to 15000000 bp) are not large, and both have a certain fluctuation, when the standard deviation is large, this section is determined as the hybrid genotype of the two parents (the yellow region in the figure). This determination result is also consistent with the real information.

Step 5: Determination of Criteria

One of the core ideas of the method of the present invention is based on directly reading the genotype information of the real SNP of the parent and the progeny in a certain section, then quantifying the similarity between the progeny and each parent in this section, and forming a relatively simplified analysis model according to the numerical features (value size and standard deviation) of the score curve of each parent, so as to determine the genotype of each section. The criteria of the determination mainly include the following several cases:

1. If a certain parent is higher in score value and stable in score (the score curve approaches the platform period), meanwhile, the score curve of the rest parents in the section has large fluctuation, the standard deviation is large (the score curve is up-and-down fluctuation of the peak shape), and it is judged that the section is the homozygous genotype of the parent (FIG.15).

2. When it is determined that the number of parents is two, it is possible to infer that the region is a hybrid genotype of two parents (FIG.16) according to a large numerical fluctuation of both parents in a certain section and a large standard deviation.

At the time of multi-parent determination, it is likely that only possible hybrid section can be found, that is, a high-score and stable parent cannot be found in a certain section. Because the score of each parent has a large fluctuation in the section, it is only possible to give the most likely two parents according to the numerical features.

3. Since the determination method depends on the degree of similarity between the progeny and the parental genotype to a great extent, if there are two or more parents similar to the progeny in a certain section, two or more parent curves with high scores and smaller standard deviation appear, which indicates that some parents analyzed are very similar in this section, there is no too large difference, and it can be temporarily not determined.

As shown in FIG.17, the region to be determined is first defined as "unknown", while the genotype of the region is determined by the genotype of the regions on both sides.

As shown in FIG.18, if the genotypes on both sides are the same, the region is determined as the genotype, and if the genotypes on both sides are different, the middle position of the region is regarded as a recombinant breaking point, and both sides of the region are genotypes on both sides.

2.3 Sequencing-Based Rice DH Population Genotype Identification

2.3.1 Rice DH Population

The two parents of the DH population of rice used are Kasalath and Nipponbare. It is a population formed by inducing haploid and doubling by the F_1 generation of parental hybridization. A plant thereof is a homozygote, and the selfing offspring is a pure line, which can perform multi-year multi-point repeated experiments, and is an ideal material for studying genotype and environment interaction.

2.3.2 SNP identification between two parents

The high-depth sequencing data (20x-30x) of the two parent Kasalath and Nipponbare materials is compared to the rice reference genome IRGSP1.0 by bwa software, and then the required high-quality SNP information is found through GATK software, and the SNP information of the parent and the SNP information of all offspring are merged into the same vcf file, so that the required variation site information can be conveniently extracted from it.

2.3.3 Genotype Identification of DH Population

The average sequencing depth of each progeny in the DH population is about 0.02x, belongs to sequencing data with a relatively low depth, separately extracts the SNP information and the parent information of each progeny, and then performs judgment by using the *SNPwindow* script to obtain a rlt file and a bin file obtained by judging each progeny.

As shown in FIG.19, a *SNP2png* script is used, and the result file obtained in the previous step is used to visualize the genotype identification result. In the figure, the homozygous genotypes of the two parents can be observed (red is Kasalath and blue is Nipponbare), and for one multi-year selfing group, the reliability of the hybrid region is low, which may be a sequencing error or a program misjudgment caused by a lower polymorphism of the two parents in the region.

Then, the *BIN2MCD* script is used to calculate the map file of the overall genotype distribution with the bin file of the whole population as input, the map file divides the whole genome into a plurality of small bins, and each bin

determines the genotype type of the bin according to the genotype result identified by the individual.

As shown in FIG.20, after the map file is reached, a perl script is used to visualize the genotype information of the whole population, and the proportion of different genotypes of each bin position is also calculated, which is an important parameter of population genetics research. The red and blue scale map of part of FIG.20 represents the proportions of three genotypes of different bins. The visualization of this step is to facilitate a rapid and direct understanding of the population genotype. The map file matched with the phenotype output can directly use the analysis software such as winQTL to perform the positioning analysis of the QTL.

2.3.4. Genotype Identification of Rice Tree-parent Material

As shown in FIG.16, the tree-parent material cultivated in a laboratory is subjected to multi-parent genotype identification. The sequencing data volume of the progeny is about 0.2x, and the three parent materials thereof are Wushansimiao, 93-11, and Shuohui 70, respectively, the sequencing depths of the three parents are about 20x-30x. The progeny and the three parent SNP information are integrated into the same vcf, and then the final high-quality SNP is further screened out. Next, the genotype of the progeny is determined by using the *SNPwindow* script, and in one window, if the score of a certain parent is the highest, the region is determined as the homozygous genotype of the parent.

As shown in FIG.21, the bin file for determining the recombinant breaking point is obtained by utilizing the judgment degree of the multi-parent, and a perl script is used to visualize the judgment result, which can directly find genotype information on 12 chromosomes through the picture. The red region corresponds to parent I Wushansimiao, the blue region corresponds to parent II 93-11, the green region corresponds to the parent III Shuohui 70, and the yellow color is a heterozygous region.

As shown in FIG.22, genotype determination is performed on the material

by using a SEG-Map method, so that genotype determination of these plants before a laboratory is mainly genotype identification according to the two parents of Wushansimiao and 93-11, and three genotypes can be identified: the Wushansimiao homozygous genotype, the 93-11 homozygous genotype and the heterozygous genotype of the two. According to the result of the tree-parent determination, the inventors have found that it is determined that the heterozygous section is likely to correspond to the homozygous genotype of the third parent for the two-parent material. Therefore, the method of the present invention can make up for the defects existing in the previous SEG-Map software under the condition of ensuring the accuracy, and solve the problem of multi-parent genotype determination.

2.3. 5. Genotype Identification of Rice Four-parent Simulation Material

As shown in FIG.23, the inventor uses the real sequencing fastq data of the four real rice materials 93-11, Shuohui 70, Wushansimiao, and Huangzhanhua in the laboratory, and then extracts the reads of the corresponding region according to the comparison result by segment, and the data of the simulated progeny is manufactured through manual combination and screening, and the real genotype information and the recombinant breaking point of the progeny are clear, so that the feasibility and accuracy of the present invention can be evaluated by using the simulation data..

According to the determination method, the inventor identifies dozens of recombinant breaking points in the whole genome, and the determined different chromosome regions are also roughly consistent with the real genotype result..

As shown in FIG.24, this figure shows the real genotype information of the simulated progeny of the present invention.

Comparing the detail areas of the two, the present inventors have found that some local areas have some differences, and the inventors have viewed the intermediate output rlt file of the determination process, and check the cause of the determination difference. The possible reasons are as follows: 1. Because

the sequencing data depth of the progeny is not very high, only a part of the variation information of the whole genome can be captured, which may miss a part of important parent distinguishing sites, resulting in that a real parent cannot be distinguished in some regions. 2. The identified two or more parents are very similar in certain areas, and there is no parental polymorphism. This is also not due to sequencing errors or sequencing depth. For such a high-similarity region, it may temporarily not make a determination, and the genotype of the high-similarity region depends on the genotype information on both sides thereof, so that a partial region cannot be accurately determined, and the genotype is determined to the most likely parent genotypes on both sides.

It can be clearly seen from the analysis result that the most probable real parent of the whole genome can be determined, which is difficult to achieve by the previously published SEG-Map software and the traditional analysis method, which also provides a multi-parent analysis method in the breeding process of different crops.

3. Discussion

The multi-parent population has a great application prospect in genetic analysis, and by selecting a plurality of parents, population genetic diversity can be increased, a plurality of parents are fused to one population by two means of hybridization and selfing (or inbreeding), and the number of times of recombination is increased. The multi-parent population can not only increase the number of recombination occurrences, mine the genetic basis behind complex traits, but also have great potential in breeding applications due to the richness of the selection of parent genetic basis. Compared with the two-parent population, the parent number of the multi-parent population is large, the population variation richness is increased, including allele diversity and phenotypic diversity, the mapping accuracy and accuracy are provided, the QTL detection efficiency is improved, and a large amount of accumulated recombination events can improve the QTL positioning resolution; because the parent screening of the multi-parent population is finer, that is, the standard is

more strict, and the plurality of parents increase the diversity of the genetic basis, so that the QTL result can be applied to breeding research. Compared with a natural group, the multi-parent group is constructed by uniformly mixing multi-parents, and compared with a natural group, due to the fact that a pedigree relationship can be known, detailed information of group construction exists, and from the aspect of experimental design, group layering is avoided, and then the false positive problem of the positioning result is controlled.

The recombinant population is the basis of a Mendel genetics experiment, and is always used as a key factor for gene, genome and genetic variation research. However, the operation of genotype identification for a mapping population is always very laborious, time-consuming, including the costly and tedious process of marker development and genotyping of hundreds of individuals with hundreds of markers. Moreover, the map resolution obtained using such a method is also relatively low [34-36]. By applying the second-generation sequencing technology, the present inventors have developed a rapid, efficient, low-cost, large amount of information, and a reliable genotype identification method. With this new approach, for an ultra-high rate genotype identification operation of a typical plot comprising hundreds of individuals, a genome sequencing service center may be completed within a few weeks without the need for a few months or even several years as previously used for marking.

The present inventors developed a new method for high-throughput genotype identification by detecting SNPs through whole genome low coverage re-sequencing. This type of SNP data differs from conventional genetic markers mainly in two aspects. First, in general, in a recombinant group, not all strains can obtain information on a certain SNP site by means of random sequencing. Second, a single SNP site is not a reliable flag or site for genotype identification because some potential sequence errors exist.

In order to process these SNP data with unique properties generated by the second generation sequencing, the present inventors further develop a new analysis architecture, ie, using a "Sliding Window" method to determine the

genotype of this section at the local location according to the genotype of the plurality of SNPs.

The present inventors also develop a set of program processing analysis flows (pipeline) based on this theory, and the names are called SEG-Map (Sequencing Enabled Genotyping for Mapping recombination populations), meaning a sequencing-based recombinant population mapping process. Using SEG-Map, a one-way or two-way terminal short-sequence sequencing result generated by an Illumina Genome Analyzer II (GAII) can be analyzed and processed, and through multi-step analysis processing, a genetic map of a recombinant population is finally constructed, and the method is applicable to a recombinant group constructed by two- parent.

The present inventors have developed a set of novel program processing analysis processes and corresponding methods and device through research. In addition to being able to optimize the steps in the previous SEG-Map procedure and being compatible with current mainstream bioinformatics analysis software and different types of high-throughput sequencing data, it is most important that it can quickly, accurately and reliably analyze the population genotype of the multi-parent construction.

The establishment of the method of the present invention can help a multi-parent population to be better applied to crop breeding; it is also possible to accurately identify more QTL site in a multi-parent population; and genome prediction is performed for a multi-parent population, which can help them as germplasm resources to be directly applied to varieties to provide a basis.

In the current analysis flow, the program reading the genotype and determining the recombinant breaking point is designed to accommodate a variety of types of mapping populations, and is fully linked to the previous steps of identifying the SNP and then constructing the recombination bin map. After these functions are combined, the analysis software takes the short-sequence generated by the next-generation sequencing technology as an input, and outputs a recombination section through a series of operations, and this output result can be analyzed by the existing software for constructing the

genetic linkage map and the QTL (Quantity Trait Loci) analysis.

The inventor uses a high-throughput sequencing-based method to identify the genotype of the rice recombinant inbred line, showing the advantages of this new genotype identification method with respect to the PCR-based method commonly used. Prior to the development of this rice F_{11} generation recombinant inbred line population based on sequencing-based methods, the inventors have genotype identification of the F_8 generation individuals of this recombinant inbred line population with 287 insertion/deletion markers (including SSR markers). These markers were amplified with PCR and identified on agarose gel electrophoresis. The genetic linkage map constructed with the results of PCR labeling, the average coverage of each marker is approximately the genetic distance of 5 cM, which is equivalent to a physical distance of about 1.4 Mb, which is greater than the majority of the rice genetic map reported previously. The design, screening, and collection of these PCR markers took three researchers more than one year of work time. In the research of the rice recombination inbred line, the inventor uses Illumina GA to obtain the average mark coverage of each SNP of 40 kb in less than two weeks. In this way, the method for high-throughput genotype identification based on sequencing is faster, more efficient and takes much less cost than conventional PCR-based genotype identification methods.

The flux of the re-sequencing can be easily adjusted, which also enables the inventors to obtain a suitable mark density level and the resolution of the recombinant breaking point when selecting the shortest time and the minimum resource investment. When there is a new scientific problem, there is a need to have a higher marking density or more accurately determine the recombinant breaking point, the present inventors can improve the coverage of the re-sequencing for the entire or part of the mapping group. In particular, it should be noted that, with this method, the recombinant breaking point can be determined very precisely, and if there is a sufficiently high re-sequencing coverage, it is theoretically possible to locate within 1 kb. Such a fine resolution enables the detection of "double switching" phenomena that have not

been previously identified with other types of genetic markers. Finally, the method can improve the accuracy of QTL detection and positioning and increase the efficiency and success rate of gene cloning. The accurately identified recombinant breaking points also enable research on genomic regions (such as recombinant hotspots) with special genetic features.

In summary, it can be seen that the high-throughput genotype identification method implemented in combination with the second-generation sequencing technology will greatly simplify and accelerate the genetic positioning of the quantity traits in the crops [37-39, 20]. The theoretical method proposed by the present inventors can better match a multi-parent population for genotype identification, improve the accuracy and efficiency of QTL positioning, and make full use of rich genetic variations present in a multi-parent population. Meanwhile, the improvement of crop genetic quality and molecular breeding design are also facilitated. In practical application, the method can be used for obtaining molecular markers closely linked to important agronomic trait genes, efficient screening of offspring during breeding, fine identification of improved variety genotype maps, etc. It provides a fast and efficient means and platform for molecular marker-assisted screening and breeding, which improves the efficiency and accuracy to a new level. In summary, this sequencing-based high-throughput genotype identification method will provide convenience for solving complex biological problems and crop breeding improvement.

All documents mentioned in the present invention are referred to in the present application as references, as if each document is individually cited as a reference. In addition, it should be understood that after reading the above teaching content of the present disclosure, those skilled in the art may make various changes or modifications to the present disclosure, and these equivalent forms also fall within the scope defined by the appended claims of this application.

References

[1] Winzeler, E.A. et al (1998). Direct allelic variation scanning of the yeast genome. Science, 281: 1194-1197.

[2] Meaburn, E., Butcher, L.M., Schalkwyk, L.C., & Plomin, R. (2006) Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. Nucleic Acids Res., 34: e27.

[3] Singer, T. et al. (2006) A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet., 2: e144.

[4] Jeremy, E. et al. (2008) Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. Plant Methods, 4: 13.

[5] Craig, D.W. et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods, 5: 887-893.

[6] Cronn, R. et al (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res., 36: e122.

[7] Doi K , Iwata N , Yosiiimura A (1997). The construction of chromosome substitution lines of African rice (Oryza glaberrima Steud.) in the background of japonica rice (O. sativa L.). Rice Genet. Newsl., 14: 39–41.

[8] Wan XY , Wan JM , Su CC , Wang CM , Shen WB , Li JM et al. (2004). QTL detection for eating quality of cooked rice in a population of chromosome segment substitution lines. Theor. Appl. Genet., 110: 71–79.

[9] Ebitani T , Takeuchi Y , Nonoue Y , Yamamoto T , Takeuchi K , Yano M (2005). Construction and evaluation of chromosome segment substitution lines carrying overlapping chromosome segments of indica rice cultivar 'Kasalath' in a genetic background of japonica elite cultivar 'Koshihikari'. Breed Sci., 55: 65–73.

[10] Hao W , Jin J , Sun SY , Zhu MZ , Lin HX (2006). Construction of chromosome segment substitution lines carrying overlapping chromosome segments of the whole wild rice genome and identification of quantitative trait

loci for rice quality. J. Plant Physiol. Mol. Biol., 32: 354–362.

[11] Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q, Li J, Han B. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genet., 42: 961-967

[12] Li, H., Ruan, J., & Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18: 1851-1858.

[13] Kurtz, S. et al. (2004) Versatile and open software for comparing large genomes. Genome Biol., 5: R12.

[14] Rice, P., Longden, I., & Bleasby, (2000) A. EMBOSS: The European molecular biology open software suite. Trends in Genetics, 16: 276-277.

[15] Ning, Z., Cox, A.J., & Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases. Genome Res., 11: 1725-1729.

[16] Lincoln, S.E. & Lander, S.L. (1993) Mapmaker/exp 3.0 and mapmaker/qtl 1.1. Technical report. Whitehead Institute of Medical Research, Cambridge, MA.

[17] Wang, S., Basten, C.J. & Zeng, Z.B (2007). Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.

[18] Li, R. et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25: 1966-1967.

[19] Van Os, H. et al. (2006) Construction of a 10,000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. Genetics, 173: 1075-1087.

[20] Xu J, Zhao Q, Du P, Xu C, Wang B, Feng Q, Liu Q, Tang S, Gu M, Han B, Liang G. (2010) Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (Oryza stative L.). BMC Genomics, 11: 656.

[21] Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J,

Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS. (2009) The Sorghum bicolor genome and the diversification of grasses. Nature, 457(7229): 551-6.

[22] Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. (2002) The generic genome browser: a building block for a model organism system database. Genome Res., 12(10): 1599-610.

[23] Rice Annotation Project. (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. Genome Res., 17: 175-83.

[24] Rice Annotation Project. (2008) The rice annotation project database (RAP-DB): 2008 update. Nucleic Acids Res., 36: D1028-D1033.

[25] The Rice Full-Length cDNA Consortium. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice, Science, 301: 376–379.

[26] Liu, X., Lu, T., Yu, S., et al. (2007) A collection of 10,096 indica rice full-length cDNAs reveals highly expressed sequence divergence between Oryza sativa indica and japonica subspecies, Plant Mol. Biol., 65: 403–415.

[27] International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature, 436: 793-800.

[28] Yu, J. et al. (2005) The Genomes of Oryza sativa: A history of duplications. PLoS Biol., 3: 266-281.

[29] Dohm, J.C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res., 36: e105.

[30] Mouse Genome Sequencing Consortium. (2002) Initial sequencing andcomparative analysis of the mouse genome. Nature, 420: 520–562.

[31] Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J.,Morenzoni, M.M., Nilsen, G.B., et al. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature, 448: 1050–1053

[32] Stam P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J., 3:739–44.

[33] Sasaki, A. et al. (2002) A mutant gibberellin-synthesis gene in rice. Nature, 416: 701-702.

[34] Eshed, Y. and Zamir, D (1995). An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics, 141: 1147–1162.

[35] Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D. and Daniel-Vedele, F. (2002) Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor. Appl. Genet., 104: 1173–1184.

[36] Simon, M., Loudet, O., Durand, S., Berard, A., Brunel, D., Sennesal, F.-X., Durand-Tardif, M., Pelletier, G. and Camilleri, C. (2008) Quantitative trait loci mapping in five new large recombinant inbred line populations of Arabidopsis thaliana genotyped with consensus single-nucleotide polymorphism markers. Genetics, 178: 2253–2264.

[37] Huang, X. et al (2009). High-throughput genotyping by whole-genome resequencing. Genome Res., 19: 1068-1076.

[38] Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q. (2010) Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. U S A., 107(23): 10578-83. Epub 2010 May 24.

[39] Zhao Q, Huang X, Lin Z, Han B. (2010) SEG-Map: A novel software for genotype calling and genetic map construction from next-generation sequencing. Rice, 3: 98-102.

Claims

1. A method for identifying the genotype of a multi-parent plant, wherein the method comprises:

(a) for n parents and the progeny thereof, providing sequencing data Df of a progeny plant to be identified and sequencing data Dp of a parent plant corresponding to the progeny plant, with n being a positive integer of ≥ 3;

(b) on the basis of the sequencing data Df and the sequencing data Dp, the SNP site information of the parent and the progenyis determined;

(c) on the basis of the SNP site information, the genotype of the progeny is determined, thereby an evaluation result of each SNP of the progeny and distribution information of the recombinant breaking point on each chromosome of the whole genome of the progeny is obtained;

(d) on the basis of the evaluation result information of SNP of the progeny and position information of the recombinant breaking point of the whole genome, a genotype map of the progeny is constructed and/or drawed, thereby a genotype identification result of a multi-parent plant is obtained.

2. The method according to claim 1, wherein in step (c), an analysis of the recombinant breaking point is performed on the basis of the SNP "string".

3. The method according to claim 1, wherein step (c) includes analyzing the recombinant breaking point to obtain an analysis result of the recombinant breaking point,

and the recombinant breaking point analysis comprises:

(s1) constructing an SNP "string", wherein genotypes of all SNPs on each chromosome of the parent and the progeny are compressed into a string in sequence;

(s2) determining each sliding window corresponding to the SNP character string according to a predetermined window size, and scoring the SNP site in each window, so as to obtain a respective score value P of each parent in the window;

(s3) determining the genotype of each chromosome region corresponding to the progeny on the basis of the score value P obtained in the step (s2).

4. The method according to claim 1, wherein in step (s3), for each chromosome region of the progeny, the genotype corresponding to each chromosome region of the progeny is determined on the basis of each parent score value or a score value curve.

5. The method according to claim 1, wherein in step (s3), comprising: the score value of each parent on each chromosome can be obtained by sliding the sliding window on the whole genome SNP site, and the score value is taken as the ordinate, with the position of each sliding window on the chromosome as the abscissa, and the score curve of each parent is drawn.

6. The method according to claim 1, wherein in step (s3), the similarity between the progeny and each parent in the segment is quantized, and the genotype of each segment is determined according to the numerical features (value size and standard deviation) of the score curve of each parent.

7. The method of claim 3, wherein the size of the sliding window is 170-500 consecutive SNP sites, preferably 200-400 consecutive SNP sites;

and/or the sequencing depth of the sequencing data is 0.1x-10x, preferably 0.2x-5x.

8. The method according to claim 1, wherein the plant comprises crops, preferably gramineous crops.

more preferably, the crops comprise rice, wheat, soybean, and tobacco.

9. A data analysis device for identifying the genotype of a multi-parent plant, the device comprises:

a data input module, which is configured to input data to be processed to be analyzed, the data to be processed comprising: a sequencing data Df of a progeny plant to be identified, and a sequencing data Dp of a parent plant corresponding to the progeny plant;

a multi-parent plant genotype identification module, which is configured to execute the method of claim 1, so as to obtain a genotype identification result of the progeny;

and an output module, which is configured to output the genotype identification result of the progeny.

10. The device according to claim 9, wherein the multi-parent plant genotype identification module comprises:

an SNP site information analysis sub-module, which is configured to determine SNP site information of a parent and a progeny based on the sequencing data Df and the sequencing data Dp;

a chromosome recombinant breaking point analysis sub-module, which is configured to determine the genotype of the progeny based on the SNP site information, so as to obtain the evaluation result of each SNP of the progeny and the distribution information of the recombinant breaking point on each chromosome of the whole genome of the progeny;

a genotype map construction sub-module, which is configured to construct and/or draw the genotype map of the progeny on the basis of the SNP evaluation result information of the progeny and the position information of the whole genome recombinant breaking point, so as to obtain the genotype identification result of the multi-parent plant.

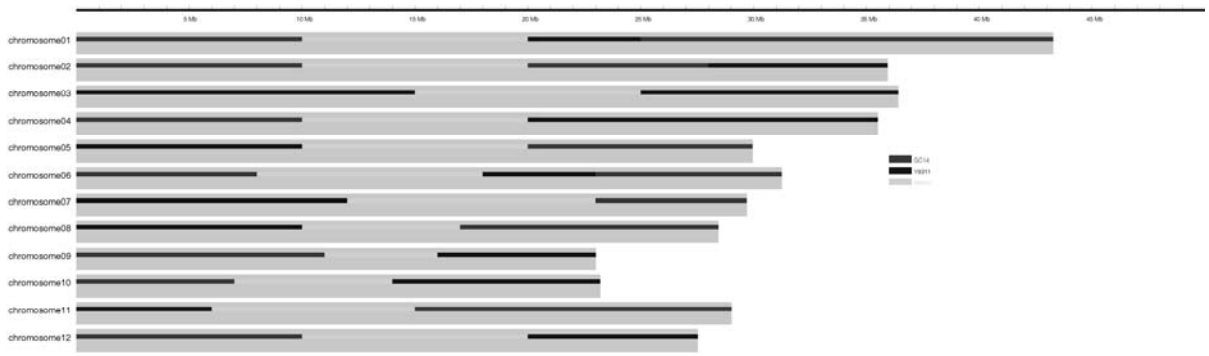

FIG. 1

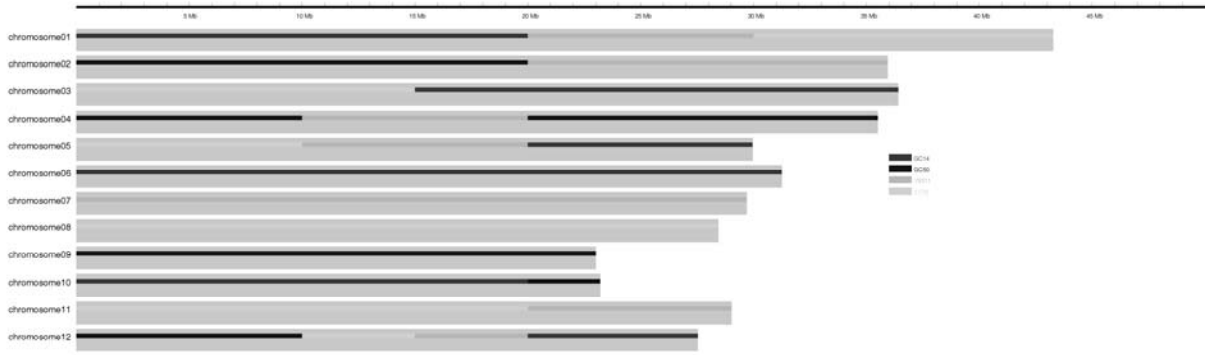

FIG. 2

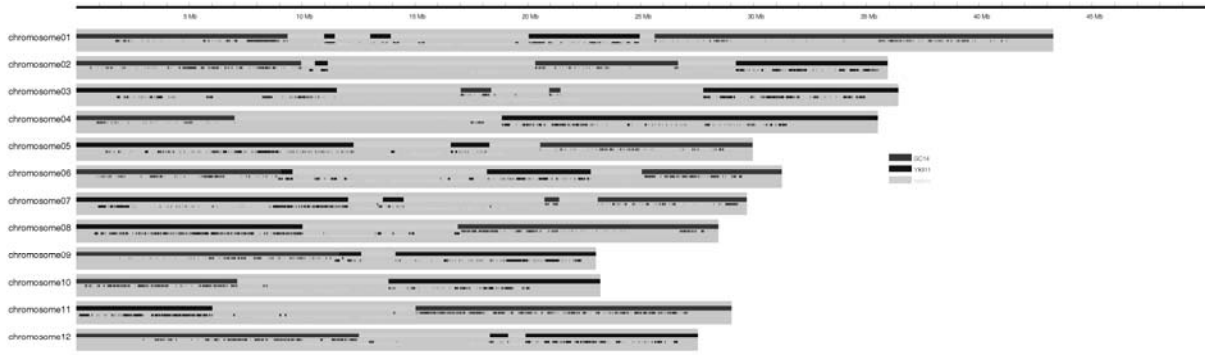

FIG. 3

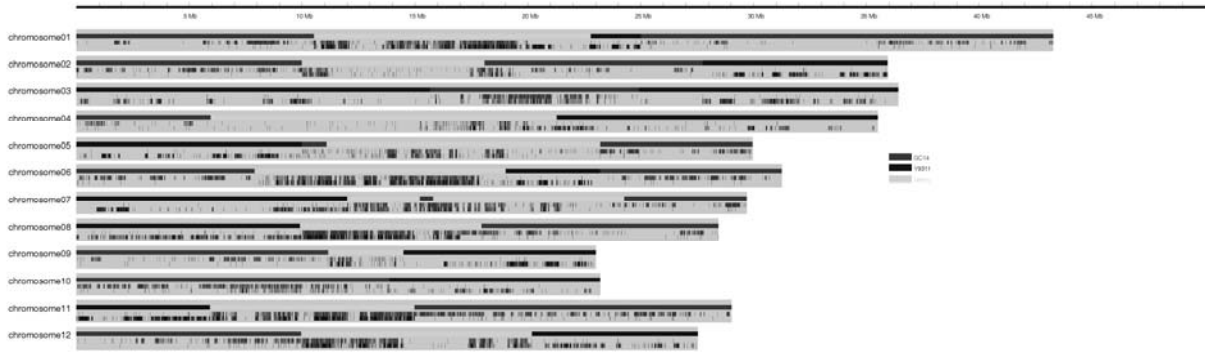

FIG. 4

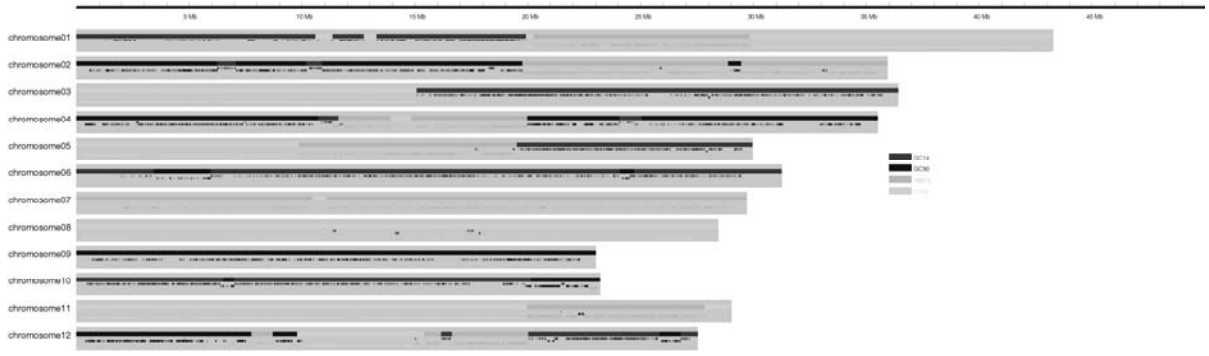

FIG. 5

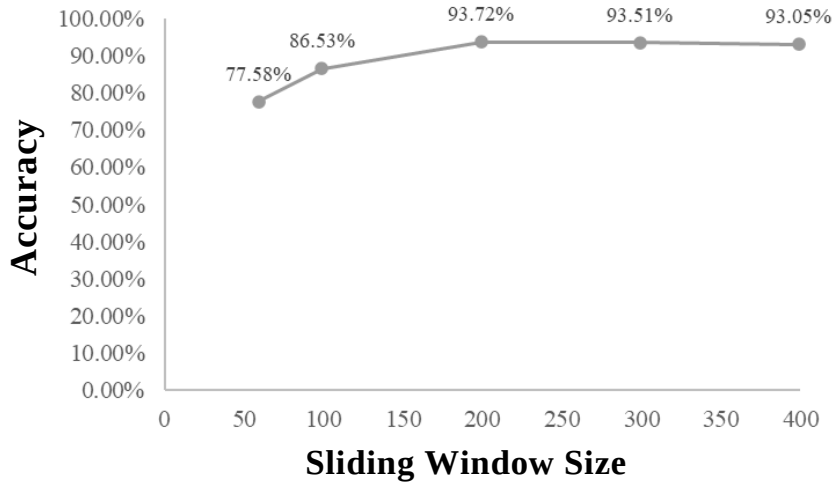

FIG. 6

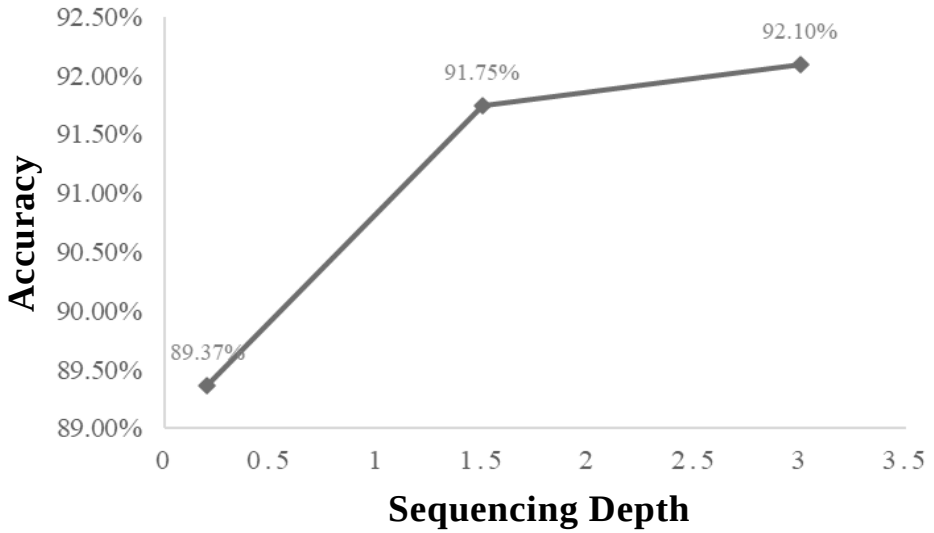

FIG. 7

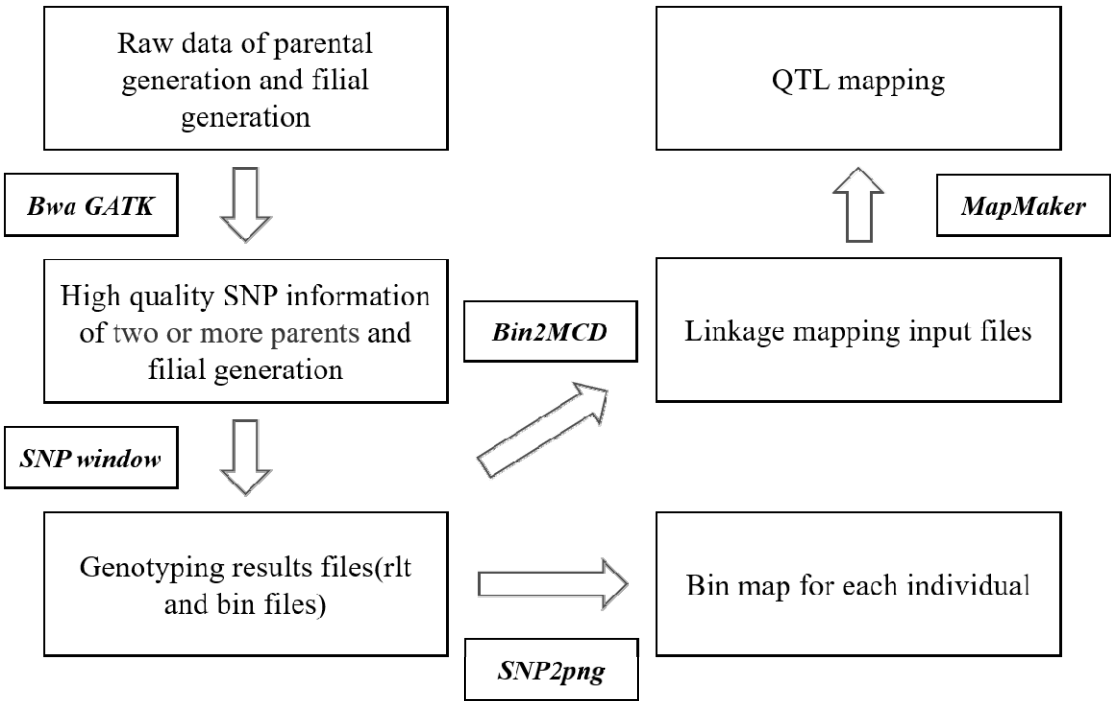

FIG. 8

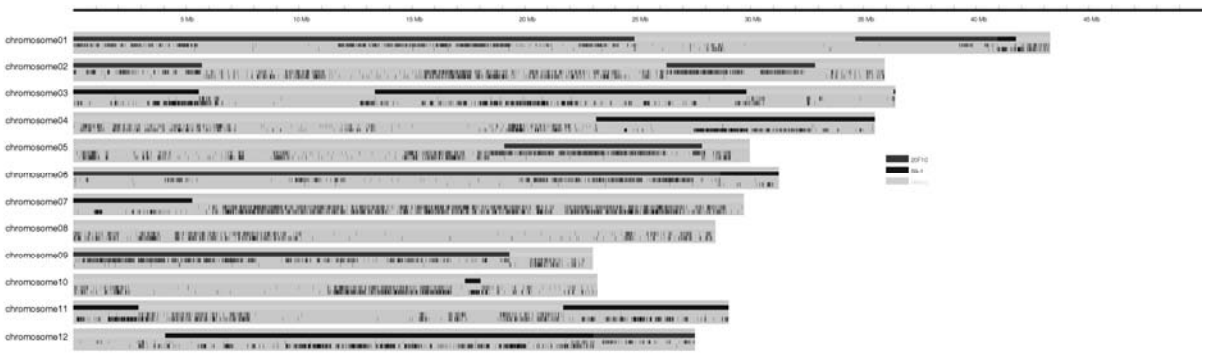

FIG. 9

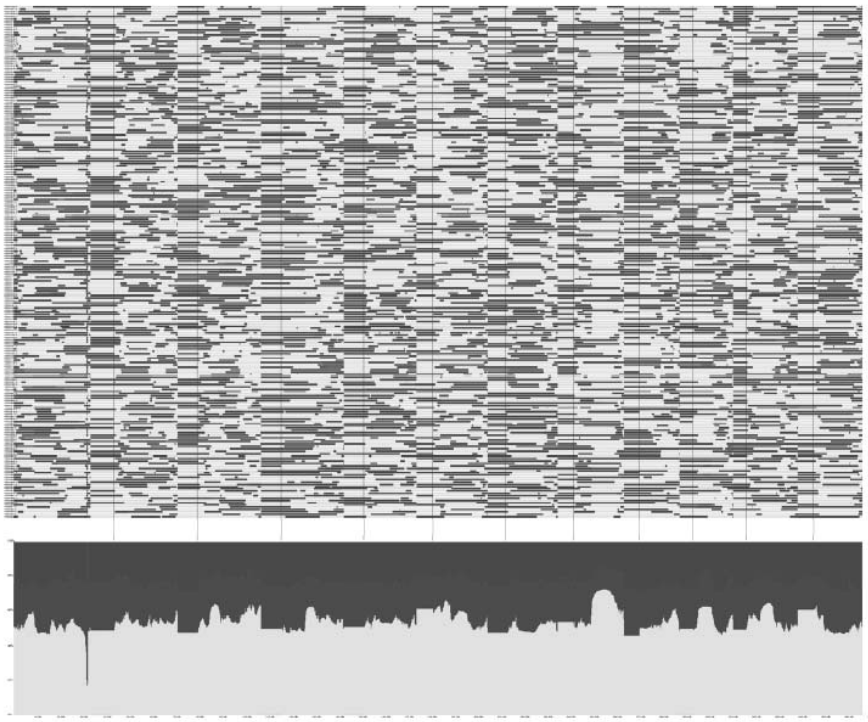

FIG. 10

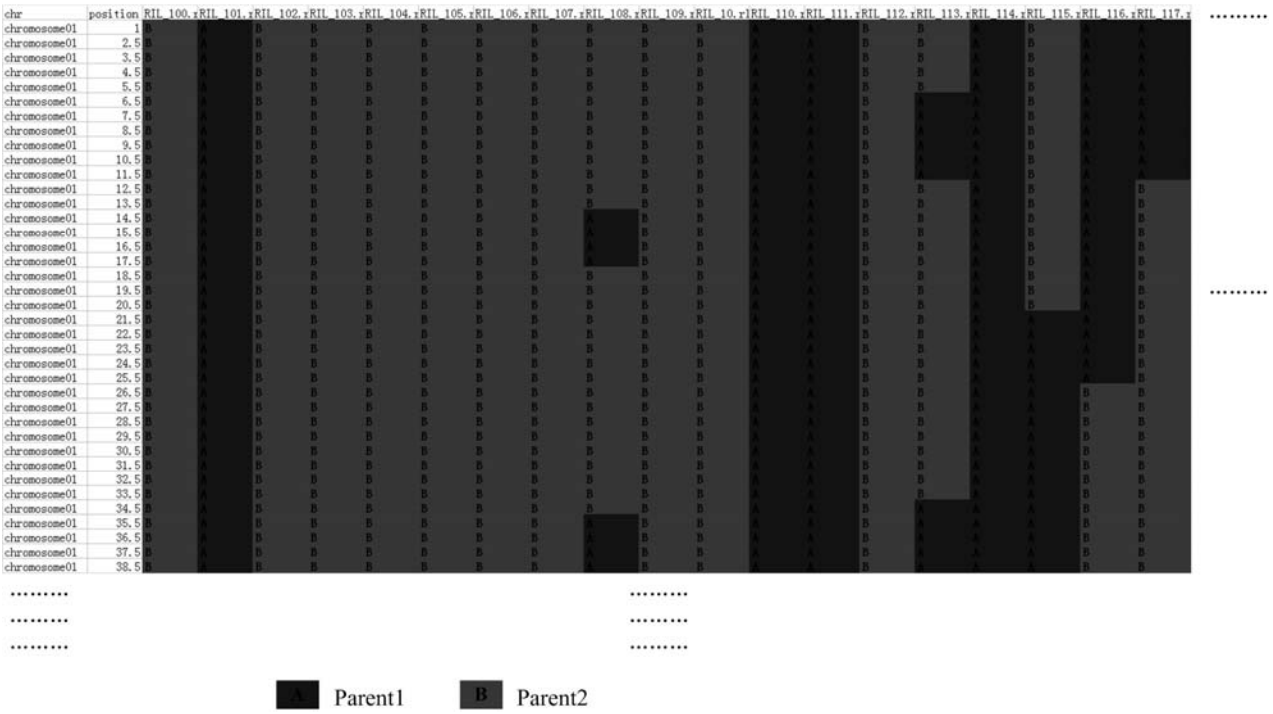

FIG. 11

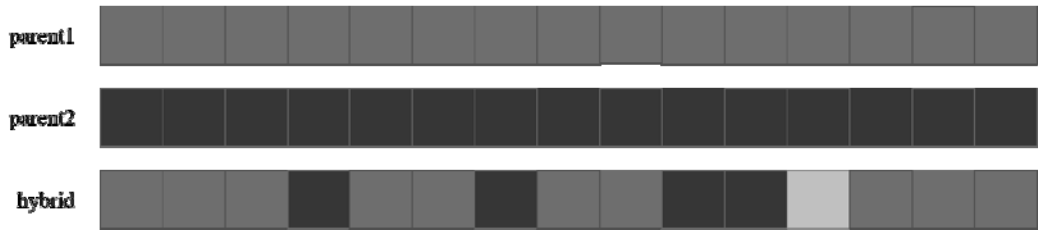

FIG. 12

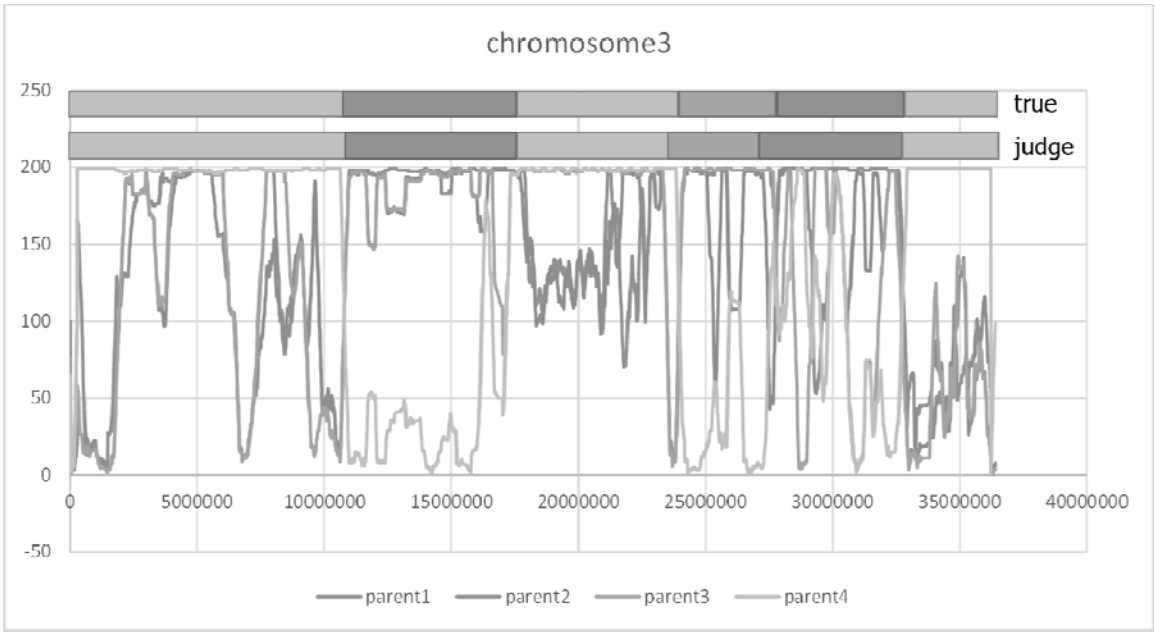

FIG. 13

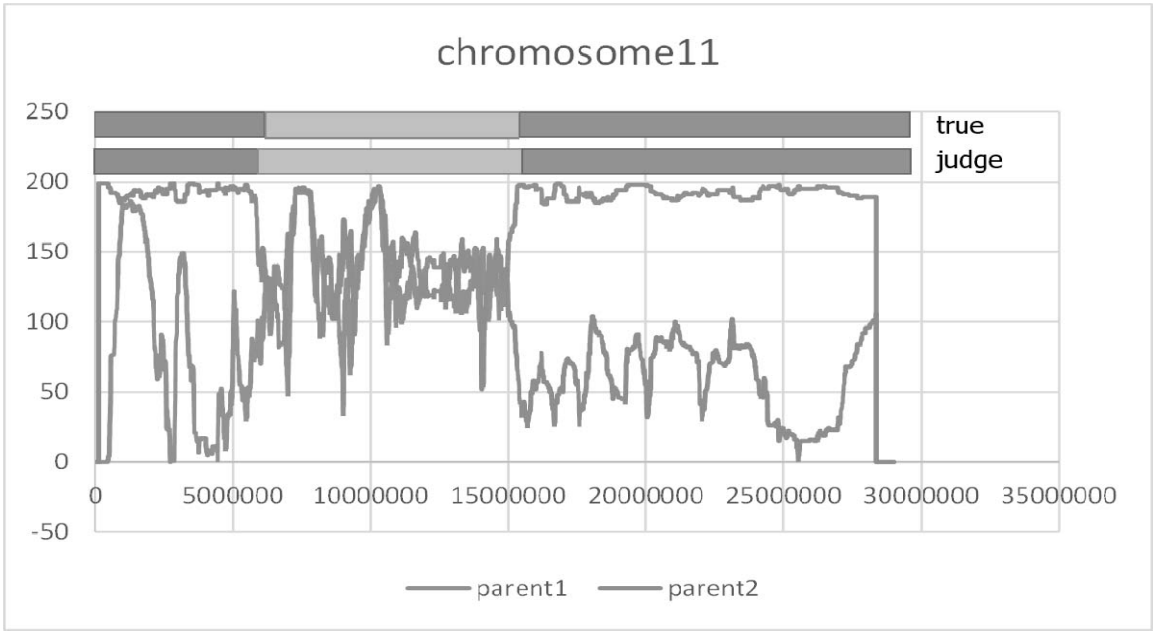

FIG. 14

Genotype	p1_score	p2_score	p3_score	p4_score	p1_Standard Deviation	p2_Standard Deviation	p3_Standard Deviation	p4_Standard Deviation
parent3	96	112	194	89	57.256801175002195	64.19760200487656	1.003074060978707	56.09645759865928
parent3	96	112	194	89	57.23685838079077	64.11827039499099	1.0046375787494466	56.07152691809911
parent3	96	112	194	89	57.21734441718352	64.03818430717145	1.0061971102302039	56.04706901679064
parent3	96	112	194	89	57.19780528320415	63.95736855813137	1.0077526739277365	56.0225769231568 0
parent3	95	111	194	88	57.17872028615144	63.875847991026696	1.0093042881877692	55.998584000158914
parent3	95	111	194	88	57.16011383198409	63.793647478416474	1.0108519711969521	55.97511559965544
parent3	94	110	194	87	57.142010264586474	63.71079192515386	1.0123957409847864	55.95219701776453
parent3	93	109	194	86	57.124433863576364	63.62730627120748	1.0139356154255226	55.929853492317775

FIG. 15

Genotype	p1_score	p2_score	p1-p2	p1_Standard Deviation	p2_Standard Deviation
hetero	109	120	-11	33.789262487309 33	26.812664495940307
hetero	109	120	-11	33.814659708072874	26.83219168618537
hetero	108	121	-13	33.84112047085422	26.852642339272503
hetero	107	122	-15	33.86861200005486	26.873968093011705
hetero	106	122	-16	33.897101283682716	26.896120339076145
hetero	106	122	16	33.92510453387825	26.918005759354894
hetero	106	122	-16	33.95407848417418	26.94067371509561
hetero	106	122	-16	33.98181456888832	26.961165096176792
hetero	106	122	-16	34.010469865113045	26.981412736015773
hetero	105	122	17	34.04217034981059	27.003323984070317

FIG. 16

Genotype	p1_score	p2_score	p3_score	p4_score	p1_Standard Deviation	p2_Standard Deviation	p3_Standard Deviation	p4_Standard Deviation
unknown	195	93	88	197	1.2667761432530134	17.076969643198343	38.18418387822731	1.3452716963866802
unknown	195	92	89	198	1.264476573222197	17.080240119683584	38.08077748427671	1.3438772053179366
unknown	195	92	88	198	1.262167849389228	17.08499530867891	37.97643659578358	1.3424707644498786
unknown	195	92	87	198	1.259849921430016	17.086430279859233	37.87117822297912	1.3410523361851248
unknown	195	92	86	198	1.2575227384480623	17.087695635221205	37.76501937360016	1.3396218824459247
unknown	195	92	85	198	1.2568536547753169	17.085641849073262	37.657977056959666	1.3381793646682012
unknown	195	92	85	198	1.256182967765069	17.081752702736285	37.550068288009356	1.3367247437954688
unknown	195	92	84	198	1.2555106748478364	17.076203800044414	37.441637627967786	1.3352579802726423

FIG. 17

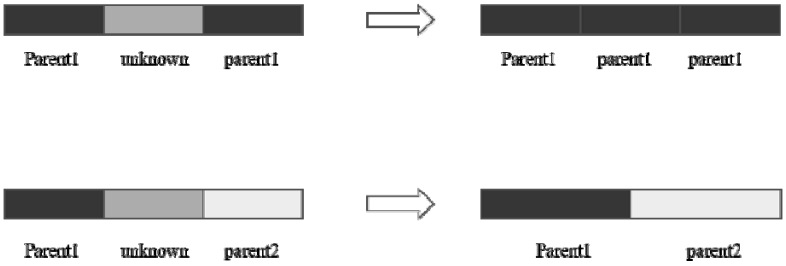

FIG. 18

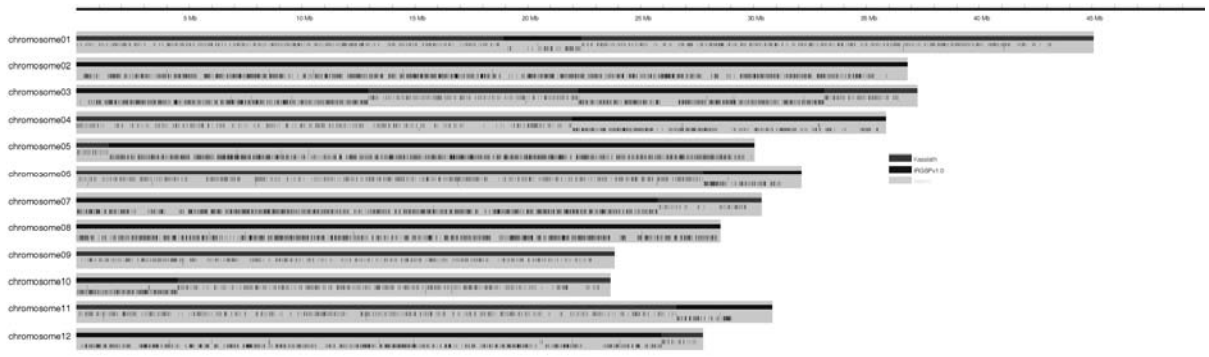

FIG. 19

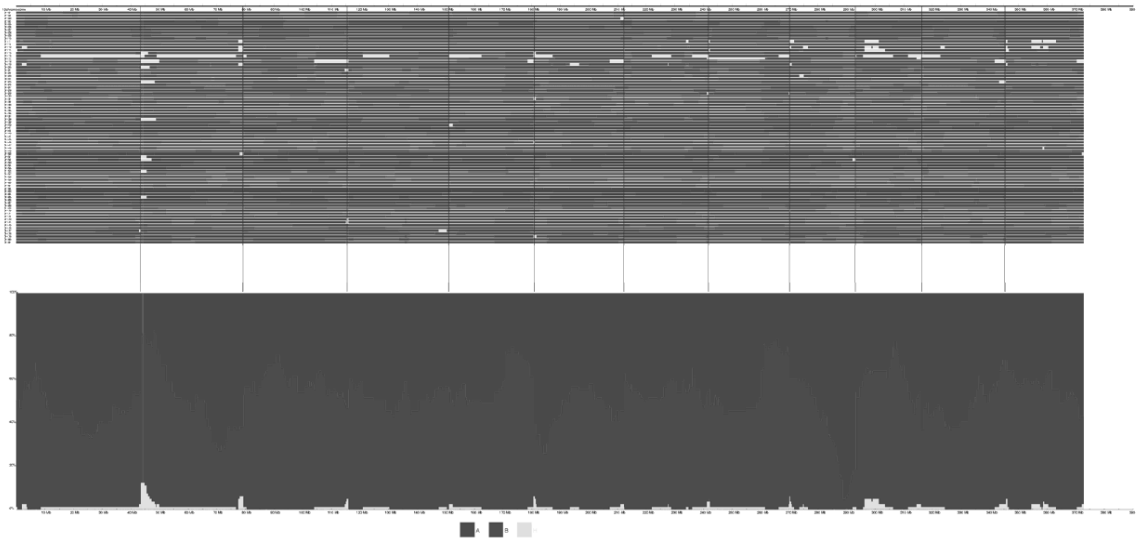

FIG. 20

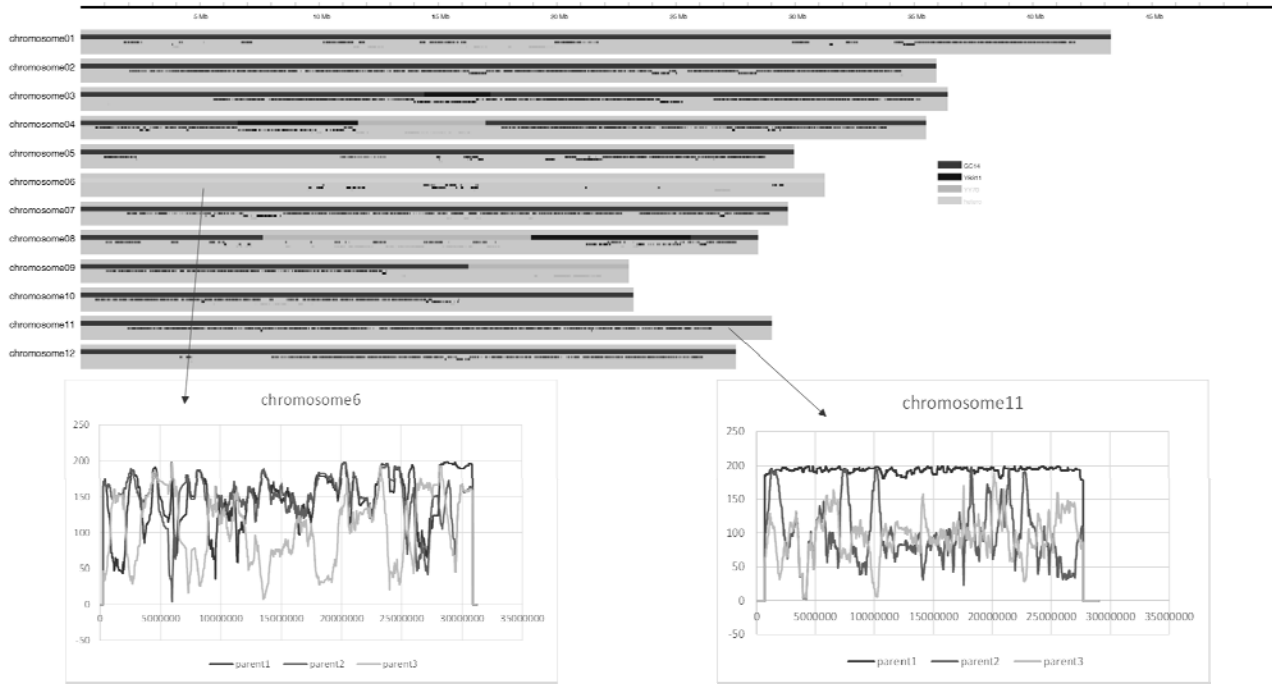

FIG. 21

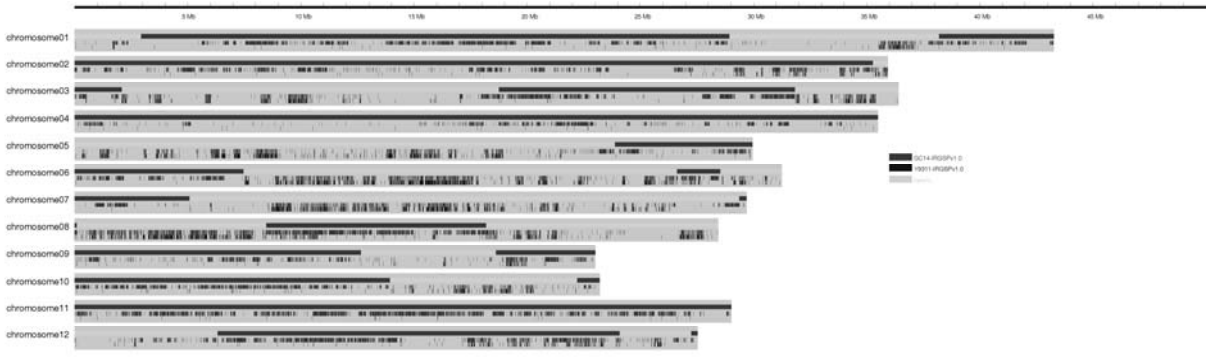

FIG. 22

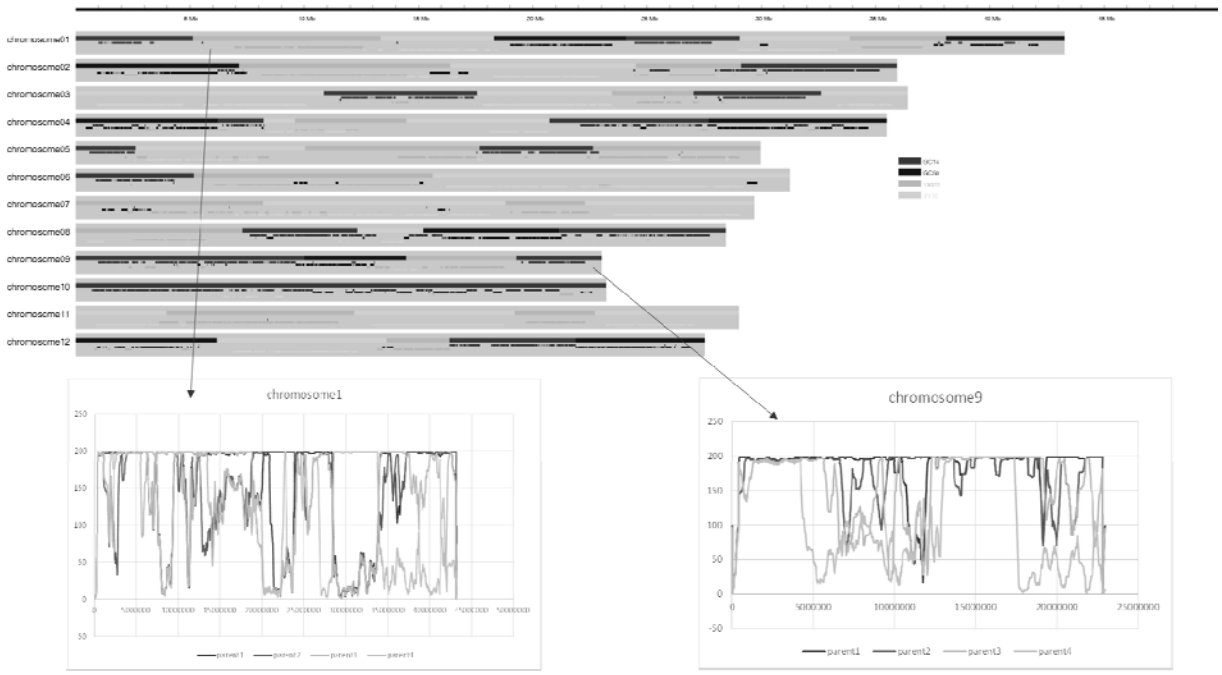
chromosome01
chromosome02
chromosome03
chromosome04
chromosome05
chromosome06
chromosome07
chromosome08
chromosome09
chromosome10
chromosome11
chromosome12
chromosome1
chromosome9
parent1
parent2
parent3
parent4

FIG. 23

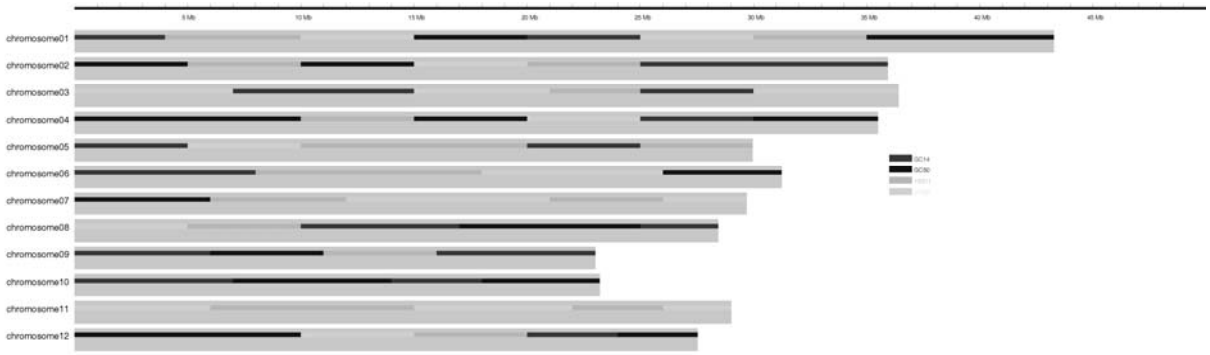
chromosome01
chromosome02
chromosome03
chromosome04
chromosome05
chromosome06
chromosome07
chromosome08
chromosome09
chromosome10
chromosome11
chromosome12

FIG. 24