DuplicationDetector, a light weight tool for duplication detection using NGS data

G. Djedatin, C. Monat, S. Engelen, Francois Sabot

Current Plant Biology 9–10 (2017) 23–28

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

DuplicationDetector, a light weight tool for duplication detection using NGS data☆

Gustave Djedatin[a,b,**], Cécile Monat[b,c,d,1], Stefan Engelen[e], Francois Sabot[b,c,d,*]

[a] *BIOGENOM Laboratory, FAST/DASSA, BP 14 Dassa-Zoumé, Benin*
[b] *DIADE UMR IRD/UM–Centre IRD de Montpellier, 911 av Agropolis BP 604501, F-34 394 Montpellier Cedex 5, France*
[c] *South Green Bioinformatics Platform, Agropolis Campus, Montpellier, France*
[d] *Université de Montpellier, Place Eugène Bataillon, 34000, Montpellier, France*
[e] *Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, BP5706, F-91057 Evry, France*

ARTICLE INFO

Article history:
Received 19 May 2017
Received in revised form 18 July 2017
Accepted 19 July 2017

Keywords:
Duplication
NGS
Rice

ABSTRACT

Duplications are one on the main evolutionary forces in angiosperm, especially in *Poaceae*. A large number of genes involved in various metabolisms and pathways originate from such duplications (whole genome, segmental or single gene). However, to detect such duplication may be complicated, costly and generally requires heavy human and material investments. Here, we propose an alternative approach for detecting putative recent segmental duplications in haploid or diploid homozygous organisms based on NGS data. We rely on abusive mappings of paralogous sequences that increase apparent heterozygous points at a given locus to identify such duplicated genomic regions. We test our tool on simulated data, then on true rice genomic sequences and were able to identify about 200 candidate duplicated genes in African rice (*Oryza glaberrima*) lineage compared to Asian one (*O. sativa*).

1. Introduction

Duplication is an important feature of the plant genome architecture, and can involve a single gene, a chromosome segment, an entire chromosome or even the whole genome [1]. It was shown for instance that angiosperms undergone large scale duplications and multiple whole genome duplications all along their evolution [2]. Whole genome duplication, *i.e.* doubling the amount of the complete genetic material of an individual without crossing, appears as a source of evolution and biological complexity [1,3,4]. In the same way, segmental duplications create local variations offering new opportunities for natural selection to occur [5]. Therefore, gene and genomic duplications play an important role in the evolution of plant phenotypes [6], and duplicated genes could undergo different behaviors: (*i*) neofunctionalization – retention of both divergent copies but with a new function for one of them –, (*ii*) subfunctionalization – retention of both copies with conserved function but in another tissue/organ/timeframe for one –, or (*iii*) nonfunctionalization/pseudogenization – large number of mutations accumulation in one of the copies [1]. The two first case (neo- and subfunctionalization) may lead to new expression pattern, or even new regulatory pathway [7].

In cultivated Asian rice (*Oryza sativa*), for instance, genome duplication provided improved root resistance [8], seed germination and seedling growth to salt stress [8,9]. In addition, tandem duplications were evidenced, amplifying adaptively important resistance genes encoding membrane proteins and function related to abiotic and biotic stress [10]. Hence, segmental duplication and tandem duplication lead to HAP (Heterotrimeric Heme Activator Protein) gene duplication regulating rice heading date [11].

However, detecting genome or segmental duplications is a complex task. Different approaches and techniques are used, such as molecular ones that gather (time-consuming) techniques *e.g.* comparative genome hybridization (CGH) [12], FISH, and array CGH [13]. Recently, due to the availability of Next Generation Sequencing technologies and of their low cost [14–16], more computational sequencing-based approaches were developed (*e.g.* [17]). These methods rely mainly on Depth of Coverage variations (DoC) to identify duplications relative to a reference genome, as a duplicated regions are expected to be twice more sequenced than

☆ This article is part of a special issue entitled "Plant Development".

* Corresponding author at: DIADE UMR IRD/UM–Centre IRD de Montpellier, 911 av Agropolis BP 604501, F-34 394 Montpellier Cedex 5, France.
** Corresponding author at: BIOGENOM Laboratory, FAST/DASSA, BP 14 Dassa-Zoumé, Benin.
E-mail addresses: djedatingustave@yahoo.fr (G. Djedatin), francois.sabot@ird.fr (F. Sabot).
[1] Current address: Domestication Genomics Group, IPK Gatersleben, OT Gatersleben, Corrensstrasse 3, D-06466 Seeland, Germany.

https://doi.org/10.1016/j.cpb.2017.07.001

non-duplicated regions [18]. However, molecular approaches such as DoC methods require highly precise experiments, repetitions and heavy computational times, are designed to compare one target individual to a given reference one, and thus cannot be applied on a large number of individuals.

In this study, we propose a new methodology based on the use of apparent excess of heterozygous loci (AEH) on genomic intervals in autogamous diploid species. This methodology was implemented in a tool called *DuplicationDetector*, and was tested on a set of simulated data and shown to be fast and robust. In addition, we applied it on cultivated and wild African rices, respectively *Oryza glaberrima* and *O. barthii*, to detect duplicated genes compared to the Asian rice *Oryza sativa*.

2. Materials and methods

2.1. Material

2.1.1. Simulated sequence data for validation

A fragment of chromosome 7 (1Mb) of *Oryza sativa* ssp *japonica* cv NipponBare IRGSP1.0 was extracted from position 1,000,000 to 1,999,999 using the *extractseq* tool from *EMBOSS* [19]. Three duplications (namely duplications 1–3) were artificially created within this sequence using home-made *Perl* script (available on demand), respectively at positions 300–1500; 350,000–390,000 (containing another initial duplication) and 800,000 to 810,000. A total of nine virtual 'clones' were constructed. Clone1, without duplicated sequences, was considered as identical to the reference. Clone2 has the three duplicated sequences without mutation. Clones 4–6 have the duplicated sequences and mutations with 3% of supplementary divergence, while clones 7–9 have the duplicated sequences with the same mutations and 6% of divergence. In addition we add in clones 3–9 additional common mutations in the non duplicated sequence. All mutations were induced using the *mutate_dna* tool from the *SMS2* suite [20]. The sequences from each virtual clone were then used to simulate FASTQ data using a*RT* simulation tool [21], specifying as options *HiSeq2500* machine, 100 pair-end fragment, depth of 35, insert size of 200, 10% of insert size divergence. a*RT* includes an empiric error model that allows a very good simulation of sequencing data [21].

2.1.2. Sequence data for Oryza glaberrima and O. barthii and initial quality control

Eight accessions of African cultivated rice *Oryza glaberrima* (TOG5307, TOG5307f, TOG5321, TOG5666, TOG5887, TOG7291, UB06, UG26), and six wild relatives *Oryza barthii* (B88, IG05, IRGC106302, MB323, TB41, TG57) were used in this study (see [22] for more informations about those accessions). Asian cultivated rice *O. sativa* IR64 (ssp *indica*) and Azucena (ssp *japonica*) were also included as control. All samples were sequenced at Genoscope (France), in the frame of the IRIGIN project (http://irigin.org), as follows:

Sequencing: Libraries were prepared using the *NEBNext* DNA Modules Products (New England Biolabs, MA, USA) with a 'on beads' protocol developed at the Genoscope, thus reducing the costs and increasing the yields. Briefly, after gDNA fragmentation with the *E210 Covaris* instrument (Covaris, Inc., USA), end repair, A-tailing and ligation with adapted concentrations of *Nextflex* DNA barcodes (Bioo Scientific, Austin, TX,) were performed on the same *AMPure XP* beads that was used for the first purification after end repair. After two consecutive 1x *AMPure XP* clean up, the ligated product was amplified by 12 cycles PCR using *Kapa Hifi Hotstart* NGS library Amplification kit (Kapa Biosystems, Wilmington, MA), followed by 0.6x *AMPure XP* purification. Libraries traces were validated on *Agilent 2100 Bioanalyzer* (Agilent Technologies, USA) and quantified by qPCR using the *KAPA* Library Quantification Kit (KapaBiosystems) on a *MxPro* instrument (Agilent Technologies, USA). Libraries were sequenced on an *Illumina HiSeq2000* or *HiSeq4000* instrument (Illumina, USA), at 2×101 bp or 2×151 bp. respectively. About 50 billion useful paired-end reads were obtained per run.

QC and initial treatments: Low quality clusters were filtered during the sequencing run by *Real Time Analysis* (*RTA*) software. Filtering steps were performed on whole paired FASTQ files: *Illumina* adapters and primers were removed, nucleotides with quality value lower than 20 were trimmed from both ends and sequences between the second unknown nucleotide (N) and the end of the read were trimmed. Reads shorter than 30 nucleotides after trimming were discarded. These trimming steps were achieved using *fastxtend* (http://www.genoscope.cns.fr/fastxtend/), a software based on the FASTX library [23]. The filtered reads and their mates that mapped onto run quality control sequences (*PhiX* genome) were removed using *SOAP* aligner [24].

2.1.3. Reference sequence & annotation

The reference sequenced genome IRSGP-1.0/MSU7.0 and its annotation from MSU v7 [25] were used for analysis as described above. The initial VCF (Variant Call Format) files are available at http://bioinfo-storage.ird.fr/2017/CPB/Djedatin/.

2.2. Methods

2.2.1. Mapping approach & initial SNP calling

For VCF creation, cleaned paired FASTQ data were mapped upon the reference initial sequence using *BWA* 0.7.12 (*aln*/*sampe* legacy algoritm) [26]. SAM (Sequence Alignment/Map) files were cleaned and filtered for low quality mapping and abnormal mapping, merged and realigned using combination of *SAMtools* [27] and *PicardTools* [28]. After realignment, SNP were called using the *GATK HaplotypeCaller* [29] under standard conditions. Calling was performed per individual chromosome to optimize calculation time. All steps were performed using the *TOGGLE* pipeline [30] to ensure repeatability and traceability. The standard default values were chosen respecting two criteria: I) THE IRIC/3K genomes standards for mapping/calling and II) numerous home made tests and evaluations of conditions using control samples in different analyses (such as in Monat et al. [30], GBE) that provide the best results. Detailed options are shown in suppData, as well as *TOGGLE* configuration file.

2.2.2. Heterozygous SNP recovery

VCF were filtered out for recovery of lines containing heterozygous SNPs based on standard default filters (depth for each sample, maximum number of missing data, minimal calling quality value, maximum MQ0 value, homozygous controls).

2.2.3. Genomics intervals recomposition

Extracted VCF lines were recompiled in genomics intervals respecting a specified maximal distance between 2 SNPs to be considered as related, a minimal block size, and a minimal heterozygous SNP density. Resulting files are 3 columns BED-like files.

2.2.4. Duplicated genes identification and SNP potential effect identification

Genomics interval files were crossed with GFF file containing annotation using *intersectBED* from the *BEDtools* suite [31]. Selected heterozygous SNPs were then annotated for their potential effect using *snpEff* software [32]. A schematic view of the whole pipeline is detailed in Suppdata.

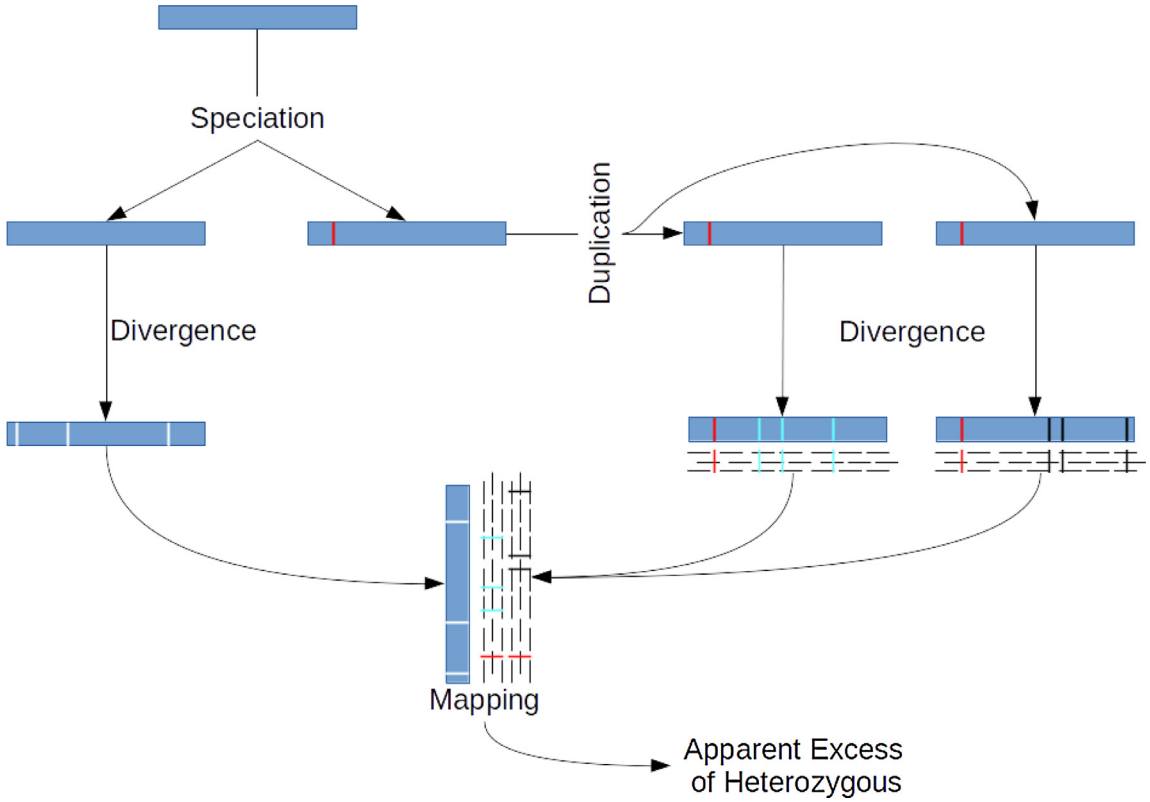

Fig. 1. Apparent Excess of Heterozygous will appear if reads coming from a duplicated region are abusively mapped on a reference genome without the duplication.

2.3. Availability

All codes, installation instructions and manual for *DuplicationDetector* are available, under the GPLv3/CeCiLL-B double licenses, on the GitHub of the project: https://github.com/SouthGreenPlatform/duplicationDetector.

3. Results

3.1. Description of the approach

We rely on AEH genomic intervals to detect duplicated sequences (Fig. 1). Basically, we detect abusive mapping of reads coming from duplicated regions in a sequenced individual when they are mapped on a reference genome without the duplication (*i.e.* harboring only a single copy). Such abusive mapping will lead the SNP calling to produce an Apparent Excess of Heterozygous locus, *i.e.* too many heterozygous loci in a short region. If many individuals are sequenced and mapped in the same experiment, such AEH loci will appear for (almost) each individual in the same location, indicating thus that the region is duplicated in the sequenced genomes compared to the reference one. Users can manage in *DuplicationDetector* the level of stringency for selection of AEH loci using different criteria:

- Minimal depth per individuals (default at 30)
- Minimal number of individuals to be heterozygous for a point to be chosen (default at 10)
- Maximal number of MQ0 reads (that can be mapped at two positions with identical score; default at 0)
- Control individuals (some samples that may not be heterozygous)

The genomic interval creation can also be set up using following criteria:

- Maximum size between two heterozygous loci (default at 1kb)
- Minimum size of the genomic interval (default at 100bp)
- Minimum density of heterozygous loci in the genomic interval (default at 25 bases between each SNP).

If user provide a GFF file for gene annotation, duplicated genes will be identified through overlapping with identified genomic intervals.

In terms of speed, a complete scan from a raw VCF of 16 African rice individuals (12 chromosomes, ~380Mb) at high sequencing depth (~35x, see Materials and methods), with two Asian rice individuals as control, will be performed on a single recent 64-bits core in less than 2 h.

3.2. Results on simulated data

On simulated data, we were able to partially (~40%) identify the duplication 1 directly and almost entirely (~95%) the duplication 3 (Table 1) as fragmented blocks. We were able to limit those two duplications with a quite good resolution, *i.e.* capacity of correctly identify the borders (max 500bp of error in limitating; see Table 1). The difference of recovery level between the two duplications is mainly due to their size, as for duplication 1 (1.2kb), the non recovered fragment is of 193bp in 5′ and 522bp in 3′ on 1200, while for duplication 3 the non-recovered size is of 142/355 bp. The fragmentation effect may be due to the fact that duplication 3 is a large block but with a low variation density, and AEH loci density parameter is quite conservative.

Duplication 2 was not detected, as it contains another older (non-simulated) nested duplication, and AEH loci in this region were removed based on the maximum MQ0 parameter. Indeed, reads coming from a region already duplicated in the reference genome could be mapped on any of the two copies on the reference with the same probability. This will increase the MQ0 level,

Table 1
Statistics about simulated data. Positions are given in base pair.

Duplication	Start	Stop	Start (recovered)	Stop (recovered)	Resolution in bp/Nb blocks	Recovery
1	300	1500	493	978	193 < - > 522/1 block	40.42%
2	3,50,000	3,90,000	NA	NA	NA	NA
3	8,00,000	8,10,000	8,00,142	8,09,675	142 < - > 355/8 blocks	95.33%

Table 2
Statistics about duplicated regions in African rices compared to Asian. Sizes are given in base pair.

	Block Size			AEH loci			AEH loci freq		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Chr1	142	795	2770	9	37	111	9.33	10.99	13.03
Chr2	106	623	2031	5	31	106	8.58	10.75	12.75
Chr3	110	293	825	5	15	40	8.78	10.67	13.3
Chr4	103	794	2079	6	40	132	8.33	10.31	13
Chr5	119	719	2668	6	37	113	9.62	11.07	13.25
Chr6	107	1057	3560	7	50	169	8.44	11.02	13.15
Chr7	147	903	3198	7	49	154	8.94	11.19	13.17
Chr8	111	925	3336	8	44	158	8.1	11.29	13.36
Chr9	127	634	1057	7	35	64	10.07	11.37	13
Chr10	127	937	2817	7	47	170	9.14	11.21	13.29
Chr11	124	716	2055	6	35	92	8.68	10.91	12.89
Chr12	101	744	1804	5	37	85	8.79	10.62	12.05

and thus those regions will be filtered out with the current version of *DuplicationDetector*.

No false positives regions were obtained from the simulated data.

3.3. Results on experimental data

We then test our tool on an experimental set of African rice individuals sequenced at relatively high-depth (see Materials and methods). African cultivated rice *Oryza glaberrima* and its wild relative *O. barthii* diverged from Asian cultivated rice *O. sativa* ancestor almost 1 million year ago [33,34], and may harbor duplicated regions compared to Asian rice. We thus tested 8 cultivated and 6 wild samples to evaluate our tool on real data. To avoid individual effect due to the reference genome (*i.e.* identification of AEH loci only because of a loss of duplication in Nipponbare), we add as control two other *O. sativa* sample, IR64 (*indica* subspecies) and Azucena (*japonica* subspecies), sequenced as described in 2.2.1. Those two individuals must be homozygous and similar to the reference for an AEH locus on the African samples to be validated as such.

The whole analysis on the raw VCF representing variations on all the 12 chromosomes of Nipponbare reference spend less than 2 h using a single core, with a very short memory footprint (max 2Mbytes per file), and provided 786 putatively duplicated regions in African rice lineage relative to Asian one (see Table 2). From those putatively duplicated regions, we can identify 200 annotated gene feature (*i.e.* tagged as 'gene' in the MSU7 GFF file; see SuppData). The blocks ranged from 102 to 3560 bases, with 5–170 SNP (11SNP/kb on average) with AEH in each block (Table 2).

The duplicated regions are widely spread all along the chromosomes, without any main locations (Fig. 2 and SuppData). The putatively duplicated genes themselves seems to be related partially to stress response, but no general trend were observe concerning Gene Ontology (data not shown). We observed a mean value of 38 AEH loci per region, showing that these duplications are quite recent, but dating from before the radiation between *O. glaberrima* and *O. barthii*. Indeed, if applying a mean value of 1.3×10^{-8} mutation per million years (as calculated in [35]), the mean age of the putative duplicated regions is of ~800,000 years, spanning from ~605,000 to 972,000 years, as expected from the two group separation [33,34].

Detailed analysis of mutation induced by the AEH loci in duplicated genes showed that most of the mutations occurred outside of the gene coding regions (7–20% only in exonic sequences; SuppData). For exonic mutation, a mean *Ka*/*Ks* ratio of 1.53 ± 0.45 (1.01–2.75) was observed, indicating a low level of global divergent selection. The mean observed Transition *vs* Transversion

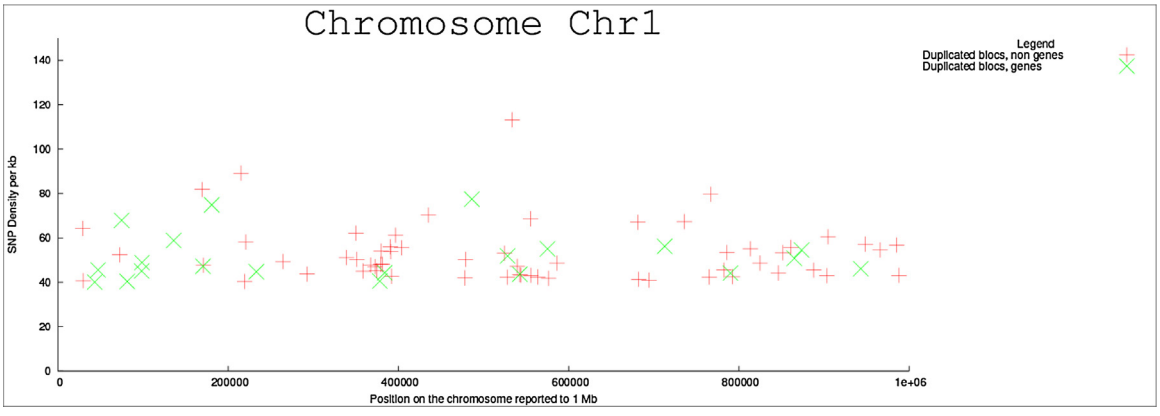

Fig. 2. Chromosomal location of duplications in African rices compared to Asian on Chromosome 1. In red are symbolized duplication involving annotated genes, in green region without annotated genes.

ratio is around 2 (1.7–2.3), as expected for genomic mutation [36,37].

When focusing on individual genes, we were able to identify as putatively duplicated the *LOC_Os07g09900* gene (Chromosome 7, from 5,263,409 to 5,267,310), *Disease resistance protein RPM1*, located under a major QTL of resistance to African strain of *Xanthomonas* (*qABB*-7, from [38]). As it has been shown in other plant species, duplication of resistance genes may increase disease resistance.

3.4. Limits of the approach

In a recent paper, Hutin et al. [39] shown a recent partial duplication of 3.2kb of the *LOC_Os11g31190* gene *OsSWEET14* [39] (Chromosome 11, from 18,171,678 to 18,174,478), involved in an increased resistance to *Xanthomonas oryzae pv oryzae*. A set of 12 high level variant positions, including the 18bp deletion between the original copy and the duplicated one [39], were identified in the variant selection step. However, post-filtrations (especially the minimal SNP density, here of 177bp between two SNP instead of 25) did not allow to recover this duplicated block in our current test.

4. Discussion

Detection of duplicated regions between individuals is a challenging task using molecular tools, and a costing computing task with sequence data. Up to now, use of the latter approach (mainly using NGS) was based on raw mapping, DoC divergence computation and Copy Number Variation (CNV) analyses. Numerous tools experimented such approaches, with more or less efficiency [14–16], but all of them requires a long computation time and cannot compare massively different individuals.

In the present study, we rely on abusive mapping and subsequent AEH loci to identify the duplicated regions. Moreover, our tool does not require intense re-calculation or mapping, as it relies directly on the raw VCF data (already generated in numerous genetic analyses) to identify those AEH loci. This approach is fast and allows to work on large samples; in addition, it offers the possibility to include negative controls which allow users to identify duplications existing in only a subsample of their sequenced individuals. Our approach is however quite conservative, as shown on simulated data, and will not detect for instance new copies of an already existing repeated sequence in the reference genome. In the same way, it cannot identify too recent duplications, due to the low number of mutations between the two copies (as for *OsSWEET14*). However, applied to the divergence between African and Asian rices, we were able to identify more than 200 putatively duplicated genes and almost 780 total regions. The detected genes are widely spread all along the chromosomes, generally related to stress response, at least marginally, and under a quite low divergent selection.

5. Conclusion

DuplicationDetector is thus a very efficient tool to detect duplication in haploid and highly homozygous diploid organisms, such as rice (tested here), but also bacteria, yeasts, autogamous plants, haploid fungi, and so on. The future development of our tool will include the implementation of detection in heterozygous or polyploid organisms or both, as well as additional criteria for filtering (such as hard-clipping level).

Author's contribution

GD and FS manage the whole study and wrote the whole pipeline. SE & Genoscope performed the sequencing and initial data treatments and QC. FS performed the simulation, GD and CM performed the basic data analyses, and GD analyzed the results. GD and FS wrote the manuscript, and all authors corrected and approve the current version.

Acknowledgments

GD was supported by an IRD grant (2013–2017 BEST FellowShip). CM was supported by ANR (AfriCrop project #ANR-13-BSV7-0017) and NUMEV labex (LandPanToggle #2015-1-030-LARMANDE). Authors want to thanks Genoscope members for the sequencing of all rice data. This work was supported by *France Génomique* French National infrastructure, funded as part of "Investissement d'avenir" program managed by ANR (#ANR-10-INBS-09), in the frame of the IRIGIN project (http://irigin.org).

Conflict of interest statement

The authors have no conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cpb.2017.07.001.

References

[1] S. Ohno, The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication, Trans. Am. Fish. Soc. 99 (1970) 120–130.
[2] S. De Bodt, S. Maere, Y. Van de Peer, Genome duplication and the origin of angiosperms, Trends Ecol. Evol. 20 (2005) 591–597.
[3] R. Aburomia, O. Khaner, A. Sidow, Functional evolution in the ancestral lineage of vertebrates or when genomic complexity was wagging its morphological tail, in: Genome Evolution, Springer, Netherlands, Dordrecht, 2003, pp. 45–52.
[4] J.S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old ideas, Annu. Rev. Genet. 38 (2004) 615–643.
[5] R. Chandan, D. Indra, Gene duplication: a major force in evolution and bio-diversity, Int. J. Biodivers. Conserv. 6 (2014) 41–49.
[6] L.E. Flagel, J.F. Wendel, Gene duplication and evolutionary novelty in plants, New Phytol. 183 (2009) 557–564.
[7] C. Feschotte, Transposable elements and the evolution of regulatory networks, Nat. Rev. Genet. 9 (2008) 397–405.
[8] Y. Tu, A. Jiang, L. Gan, M. Hossain, J. Zhang, B. Peng, Y. Xiong, Z. Song, D. Cai, W. Xu, et al., Genome duplication improves rice root resistance to salt stress, Rice 7 (2014) 15.
[9] A. Jiang, L. Gan, Y. Tu, H. Ma, J. Zhang, Z. Song, Y. He, D. Cai, X. Xue, The effect of genome duplication on seed germination and seedling growth of rice under salt stress, Aust. J. Crop Sci. 7 (2013) 1814–1821.
[10] C. Rizzon, L. Ponger, B.S. Gaut, S. Maere, S. Bodt, J. De Raes, T. Casneuf, M. Montagu, G. Van Blanc, K. Wolfe, et al., Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice, PLoS Comput. Biol. 2 (2006) e115.
[11] Q. Li, W. Yan, H. Chen, C. Tan, Z. Han, W. Yao, G. Li, M. Yuan, Y. Xing, Duplication of OsHAP family genes and their association with heading date in rice, J. Exp. Bot. 67 (2016) 1759–1768.
[12] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Döhner, T. Cremer, P. Lichter, Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances, Genes Chromosom. Cancer 20 (1997) 399–407.
[13] C. Shaw-Smith, R. Redon, L. Rickman, M. Rio, L. Willatt, H. Fiegler, H. Firth, D. Sanlaville, R. Winter, L. Colleaux, et al., Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features, J. Med. Genet. 41 (2004) 241–248.
[14] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nat. Rev. Genet. 17 (2016) 333–351.
[15] T.C. Glenn, Field guide to next-generation DNA sequencers, Mol. Ecol. Resour. 11 (2011) 759–769.
[16] T.C. Glenn, <http://molecularecologist.com//>.
[17] S. Newman, K.E. Hermetz, B. Weckselblatt, M.K. Rudd, Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints, Am. J. Hum. Genet. 96 (2015) 208–220.
[18] P. Guan, W.-K. Sung, Structural variation detection using next-generation sequencing data: a comparative technical review, Methods 102 (1 June) (2016) 36–49, http://dx.doi.org/10.1016/j.ymeth.2016.01.020.

[19] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet. 16 (2000) 276–277.
[20] P. Stothard, The sequence manipulation suite: *JavaScript* programs for analyzing and formatting protein and DNA sequences, Biotechniques 28 (2000).
[21] W. Huang, L. Li, J.R. Myers, G.T. Marth, ART: a next-generation sequencing read simulator, Bioinformatics 28 (2012) 593–594.
[22] J. Orjuela, F. Sabot, S. Chéron, Y. Vigouroux, H. Adam, H. Chrestin, K. Sanni, M. Lorieux, A. Ghesquière, An extensive analysis of the African rice genetic diversity through a global genotyping–Springer, Theor. Appl. Genet. 127 (10) (2014) 2211–2223, http://dx.doi.org/10.1007/s00122-014-2374-z.
[23] FASTX-Toolkit, <http://hannonlab.cshl.edu/fastx_toolkit/index.html/>.
[24] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics 25 (2009) 1966–1967.
[25] Y. Kawahara, M. de la Bastide, J.P. Hamilton, H. Kanamori, W.R. McCombie, S. Ouyang, D.C. Schwartz, T. Tanaka, J. Wu, S. Zhou, et al., Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data–Springer, Rice 6 (2013) 4.
[26] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows–Wheeler transform, Bioinformatics 26 (2010) 589–595.
[27] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and *SAMtools*, Bioinformatics 25 (2009) 2078–2079.
[28] Picard Tools–By Broad Institute.
[29] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., The *Genome Analysis Toolkit*: a *MapReduce* framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[30] C. Monat, C. Tranchant-Dubreuil, A. Kougbeadjo, C. Farcy, E. Ortega-Abboud, S. Amanzougarene, S. Ravel, M. Agbessi, J. Orjuela-Bouniol, M. Summo, et al., TOGGLE: toolbox for generic NGS analyses, BMC Bioinform. 16 (2015) 374.
[31] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[32] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, *SnpEff*, Fly (Austin) 6 (2012) 80–92.
[33] D.a. Vaughan, B.-R. Lu, N. Tomooka, The evolving story of rice evolution, Plant Sci. 174 (2008) 394–408.
[34] D.a. Vaughan, K. Kadowaki, A. Kaga, N. Tomooka, On the phylogeny and biogeography of the genus *Oryza*, Breed. Sci. 55 (2005) 113–122.
[35] J. Ma, J.L. Bennetzen, Rapid recent growth and divergence of rice nuclear genomes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 12404–12410.
[36] Z. Yang, D. Yoder a, Estimation of the transition/transversion rate bias and species sampling, J. Mol. Evol. 48 (1999) 274–283.
[37] S. Duchêne, S. Ho, E.C. Holmes, T. Jukes, C. Cantor, W. Brown, E. Prager, A. Wang, A. Wilson, R. Lewontin, et al., Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models, BMC Evol. Biol. 15 (2015) 36.
[38] G. Djedatin, M.-N. Ndjiondjop, A. Sanni, M. Lorieux, V. Verdier, A. Ghesquiere, Identification of novel major and minor QTLs associated with *Xanthomonas oryzae* pv. *oryzae* (African strains) resistance in rice (*Oryza sativa* L.), Rice (N. Y.) 9 (2016) 18.
[39] M. Hutin, F. Sabot, A. Ghesquière, R. Koebnik, B. Szurek, A knowledge-based molecular screen uncovers a broad spectrum OsSWEET14 resistance allele to bacterial blight from wild rice, Plant J. 84 (4) (2015) 694–703, http://dx.doi.org/10.1111/tpj.13042.