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 dupli-

[☆] This article is part of a special issue entitled "Plant Development", published in the journal Current Plant Biology 9–10, 2017.

[*] 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.

http://dx.doi.org/10.1016/j.cpb.2017.07.001
2214-6628/

cated regions are expected to be twice more sequenced than 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.