MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 693–702 doi: 10.1111/j.1755-0998.2011.02998.x

Discovery of single-nucleotide polymorphisms (SNPs) in the uncharacterized genome of the ascomycete *Ophiognomonia clavigignenti-juglandacearum* from 454 sequence data

K. D. BRODERS,* K. E. WOESTE,† P. J. SANMIGUEL,‡ R. P. WESTERMAN‡ and G. J. BOLAND*
**School of Environmental Sciences, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada N1G 2W1, †USDA Forest Service Hardwood Tree Improvement and Regeneration Center, Department of Forestry and Natural Resources, Purdue University, 715 W. State Street, West Lafayette, IN 47907, USA, ‡Department of Horticulture & Landscape Architecture, Purdue University, 625 Agriculture Mall Drive, West Lafayette, IN 47907, USA*

Abstract

The benefits from recent improvement in sequencing technologies, such as the Roche GS FLX (454) pyrosequencing, may be even more valuable in non-model organisms, such as many plant pathogenic fungi of economic importance. One application of this new sequencing technology is the rapid generation of genomic information to identify putative single-nucleotide polymorphisms (SNPs) to be used for population genetic, evolutionary, and phylogeographic studies on non-model organisms. The focus of this research was to sequence, assemble, discover and validate SNPs in a fungal genome using 454 pyrosequencing when no reference sequence is available. Genomic DNA from eight isolates of *Ophiognomonia clavigignenti-juglandacearum* was pooled in one region of a four-region sequencing run on a Roche 454 GS FLX. This yielded 71 million total bases comprising 217 000 reads, 80% of which collapsed into 16 125 754 bases in 30 339 contigs upon assembly. By aligning reads from multiple isolates, we detected 298 SNPs using Roche's GS Mapper. With no reference sequence available, however, it was difficult to distinguish true polymorphisms from sequencing error. EAGLEVIEW software was used to manually examine each contig that contained one or more putative SNPs, enabling us to discard all but 45 of the original 298 putative SNPs. Of those 45 SNPs, 13 were validated using standard Sanger sequencing. This research provides a valuable genetic resource for research into the genus *Ophiognomonia*, demonstrates a framework for the rapid and cost-effective discovery of SNP markers in non-model organisms and should prove especially useful in the case of asexual or clonal fungi with limited genetic variability.

Keywords: asexual fungi, butternut canker, high-throughput marker identification, next generation sequencing, non-model organism, reference sequence, single-nucleotide polymorphism

Received 29 October 2010; revision received 6 January 2011; accepted 17 January 2011

Introduction

The fungal plant pathogen *Ophiognomonia clavigignenti-juglandacearum* (Nair, Kostichka, & Kuntz) Broders & Boland (*Oc-j*), formerly known as *Sirococcus clavigignenti-juglandacearum* (Broders & Boland 2011), causes the disease butternut canker and is responsible for range-wide butternut (*Juglans cinerea*) mortality in recent decades and threatens the survival of the species (Ostry & Woeste 2004). Unfortunately, there is limited information on the origin and evolutionary history of this pathogen. The sudden emergence of *O. clavigignenti-juglandacearum* (*Oc-j*), its rapid spread in native North American butternuts and scarcity of resistant trees point to a recent introduction, although *Oc-j* is unknown outside North America. Furnier *et al.* (1999) found that they could not distinguish isolates using RAPDs, and they concluded that the fungus belongs to a single isolate, corroborating the hypothesis that it is a recent invader.

In an attempt to determine the genotypic diversity and evolutionary history of *Oc-j*, we developed markers that provide resolution among closely related isolates. Slowly evolving markers are preferable for determining deeper levels of relatedness, because they are less prone to evolutionary reversals or convergent evolution that can obscure patterns of descent (Keim *et al.* 2004). Single-nucleotide polymorphisms (SNPs) have relatively low

Correspondence: Kirk D. Broders, Fax: (603) 862-3784;
E-mail: kirk.broders@unh.edu

mutation rates and are thus evolutionarily stable and have effectively been used for determining broad patterns of evolution (Brumfield *et al.* 2003; Jakobsson *et al.* 2008; Li *et al.* 2008). SNPs can occur anywhere in the genome; therefore, if entire or large portions of genomes are compared and examined for SNPs, a sufficient number may be found to provide resolution at even short evolutionary scales.

Significant resources have been dedicated to the development of SNPs as high-throughput markers and to SNP discovery. Extensive SNP discovery projects have been undertaken for many species including important crop plants such as rice (Shen *et al.* 2004), maize (Ching *et al.* 2002), soybean (Zhu *et al.* 2003) and wheat (Ravel *et al.* 2006), as well as the model fungus *Neurospora crassa* (Lambreghts *et al.* 2009). For all of these projects, a complete genome or a library of expressed sequence tags (ESTs) was available. The first challenge we faced in developing SNPs for *Oc-j* was that most SNP detection methods require that new sequence be compared to a reference genome. Pre-existing or reference sequences support the assembly of genomes based on data from LifeTech SOLiD, Illumina GA-IIx (Solexa) or Roche 454 next generation sequencing platforms, which have the drawback of producing average read lengths shorter than those of Sanger sequencers. One short-cut method for using pyrosequencing platforms for SNP discovery is to simultaneously sequence multiple genotypes (Meyer *et al.* 2007; Parameswaran *et al.* 2007; Craig *et al.* 2008). This method was used, for example, to identify 297 SNPs in the uncharacterized genome of *Eucalyptus grandis* (Novaes *et al.* 2008), but this method still required the use of reference sequence derived from a cDNA library and a set of over 86 000 ESTs obtained using standard Sanger sequencing.

A second challenge to developing SNP markers using 454 sequencing is separating true polymorphisms from sequencing error. Several software programs are available to assist in the discovery of SNPs found in next generation sequencing data, including Pyrobayes (Quinlan *et al.* 2008), POLYBAYES (Marth *et al.* 1999), and AtlasSNP2 (Shen *et al.* 2010). These software programs may not be suitable for poorly characterized genomes because they all require some prior knowledge of an organism's sequence for comparing resequencing data. A third major challenge is visualizing the enormous volume of data produced by next generation sequencing technologies. Visualization is essential for uncovering errors in sequence, alignment, and assembly that may lead to false SNP calls. Visual inspection of sequence data is especially important when a reference sequence is unavailable for comparison. The software EAGLEVIEW (Huang & Marth 2008) provides a user-friendly viewer with a single-window graphical user interface specifically designed for visualizing next-generation sequencing data. The program is a data integration and visualization tool that facilitates data analyses, visual validation and hypothesis generation. EAGLEVIEW can handle assemblies of millions of reads, display mixed-type sequence reads simultaneously with technology specific trace information and display complex genome annotation.

The benefits from recent improvement in sequencing technologies may be even more valuable in non-model organisms, such as many plant pathogenic fungi, which are of economic importance. One application of this new sequencing technology is the rapid generation of genomic information to identify putative single-nucleotide polymorphisms (SNPs) to be used for population genetic, evolutionary, and phylogeographic studies on non-model organisms. We describe here the sequencing, assembly and discovery of SNPs from the fungal plant pathogen *Ophiognomonia clavigignenti-juglandacearum*, for which virtually no sequence information was previously available.

Material and methods

Fungal material

There was no previous knowledge of the genome of *Oc-j*. Therefore, eight isolates, including the type culture of *Oc-j* (ATCC 36642), were selected for analysis. The isolates were chosen for their geographic distribution and phenotypic characteristics (Table 1). An additional set of 16 isolates from geographically diverse locations in North America were used to validate sequence data, assess allele frequencies, and evaluate the utility of identified SNPs as markers for studies of population structure and disease aetiology. Diseased plant tissue was surface sterilized for 5 min in a 0.6% sodium hypochlorite solution, rinsed in deionized water, allowed to dry on sterile paper towels, plated onto acidified PDA (0.75 ml 50% lactic acid/1L PDA), and incubated at room temperature for 5–7 days. Putative *Oc-j* cultures were transferred to fresh plates of PDA. Single spore cultures were obtained by placing conidia from pycnidia growing *in vitro,* into sterile deionized water and then spreading the suspension onto water agar. The cultures were allowed to incubate at room temperature for 48 h, and individual germinating conidia were transferred to fresh plates of PDA using a stereomicroscope. For DNA extraction, isolates were grown in potato dextrose broth for 14 days, mycelia were collected and ground in liquid nitrogen, and DNA was extracted using the DNEasy Plant Mini Kit (Qiagen Inc., Valencia, CA, USA). DNA was eluted to a final volume of 50 μl, and DNA concentration was estimated using a nanodrop photospectrometer (Wilmington, DE). Total nucleic acid quantities in these samples ranged from 1600

Table 1 Sampling locations and dates for isolates of *Ophiognomonia clavigignenti-juglandacearum* recovered from infected *Juglans cinerea*, *Juglans nigra* and *Juglans ailantifolia* var. *cordiformis* tissue used for the identification and validation of single-nucleotide polymorphisms

Isolate	Origin	Host	Date collected	454 sequencing
SCJ1	WI, USA (type culture)	*J. cinerea*	1979	+
SCJ2	York, Ontario, Canada	*J. cinerea*	1 February 2008	+
SCJ3	York, Ontario, Canada	*J. cinerea*	1 February 2008	+
P-005	Simco Lake, Ontario, Canada	*J. cinerea*	18 July 2008	+
WB-22	Cambridge, Ontario, Canada	*J. cinerea*	8 July 2009	+
Bud2-3	Guelph, Ontario, Canada	*J. cinerea*	14 April 2009	+
GA5-1	Guelph, Ontario, Canada	*J. cinerea*	14 April 2009	+
GA1-1	Guelph, Ontario, Canada	*J. cinerea*	14 April 2009	+
P-013	Guelph Lake, Ontario, Canada	*J. cinerea*	8 August 2008	–
P-017	Conestoga Lake, Ontario, Canada	*J. cinerea*	14 August 2008	–
P-019	Hockley Valley PNR, Ontario, Canada	*J. cinerea*	22 August 2008	–
P-029	Simco Lake, Ontario, Canada	*J. cinerea*	14 August 2008	–
P-034	Bradford, Ontario, Canada	*J. cinerea*	26 August 2008	–
P-037	Brockville, Ontario, Canada	*J. cinerea*	28 August 2008	–
P-043	Charleston Lake, Ontario, Canada	*J. cinerea*	11 September 2008	–
P-045	Big Rideau Lake, Ontario, Canada	*J. cinerea*	16 September 2008	–
1368-1C	Oregon Co., Missouri, USA	*J. cinerea*	16 March 2007	–
AR4534	Lakewood, WI, USA	*J. cinerea*	5 June 2002	–
AR4537	Asheville, North Carolina, USA	*J. cinerea*	2001	–
AR4538	St. Francis National Forest, Arkansas	*J. cinerea*	2001	–
AR4539	Smithville, TN, USA	*J. cinerea*	2002	–
AR4540	Chester, CT, USA	*J. cinerea*	2002	–
70-BW2	Guelph, Ontario, Canada	*J. nigra*	15 June 2009	–
HN-1	Cambridge, Ontario, Canada	*J. ailantifolia* var. *cordiformis*	24 August 2009	–

to 3000 ng for the eight isolates. DNA extracted from eight isolates was pooled in an attempt to maximize the diversity in the genomes sampled.

SNP discovery pipeline

To enable the efficient and inexpensive discovery of novel SNPs in a fungal genome, we developed a process that combines next-generation sequencing, standard sequencing methods and several bioinformatics software programs. This process is referred to as the SNP discovery pipeline (Fig. 1). The following sections describe the steps in the process.

454 sequencing, assembly and SNP detection

The pooled genomic DNA was incubated for 10 min at 37 °C after addition of 10 units of Riboshredder (Epicentre Biotechnologies) followed by removal of degraded RNA using 3 washes of 1 ml of 1×TE through a YM-30 Microcon spin column (Millipore). Library construction was performed as described in Roche's General Library Preparation Method Manual. Briefly, the pool was concentrated to 100 µl and fragmented using a nebulizer for 1 min at 30 PSI, then fractionated away from any residual proteins, low-molecular-weight nucleic acids and other solutes using MinElute (Qiagen) columns. The resulting eluate contained approximately

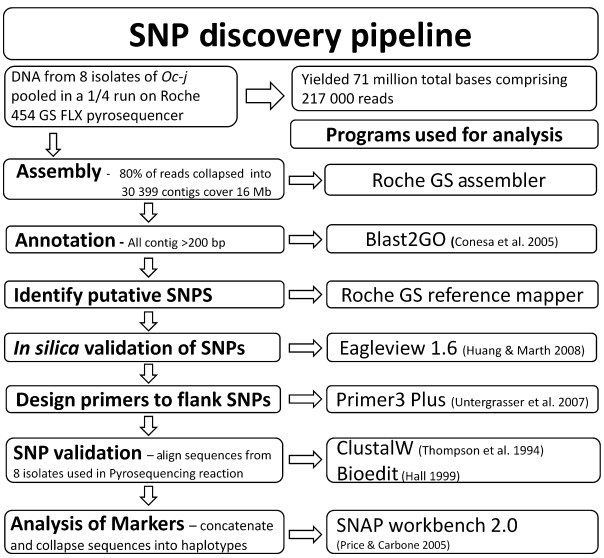

Fig. 1 Steps included in the SNP discovery pipeline used to identify, validate and analyse putative SNPs in an uncharacterized fungal genome using pyrosequencing data.

1.5 μg of sheared genomic DNA as assayed on a NanoDrop 1000 (Thermoscientific). DNA fragments between 550 and 1100 bp were separated and isolated in agarose gel. Roche Titanium PA and PB adaptors were ligated to these fragments after they were end-blunted using a mixture of T4-PNK and T4 DNA polymerase provided in the Roche fragment library construction kit. The biotinylated adaptor-containing constructs were captured using streptavidin conjugated beads, nick translated, and then ssDNA molecules were eluted from the beads. Emulsion PCR was performed to generate templated beads. Beads were enriched and deposited in a 1/4th PicoTiterPlate region and run using Roche Titanium chemistry.

The Roche assembly engine, GS Assembler (454 Life Sciences) was used to assemble the resulting sequence data. The program GS Reference Mapper (454 Life Sciences) was used to detect SNPs from these assembled reads. Single-base substitutions as well as single-base or small-scale, multi-base deletions/insertions were included in our analysis. Without a reference sequence, it was difficult to determine false positives. In SNP discovery, a false SNP call can result from alignment errors, sequencing errors, paralogous genes, or from defects in the SNP detection algorithm. Therefore, contigs which contained putative SNPs were inspected manually using EAGLEVIEW 1.6 (Huang & Marth 2008). The complete assembly file, including all 30 399 contigs, was uploaded into EAGLEVIEW in the standard ACE format, a tag-based format commonly used by genome assembly programs. The drop down menu allows contigs to be viewed one at time.

Primer design and sequencing

After putative SNPs were evaluated by inspection using EAGLEVIEW, primers were designed to span the region containing one or more SNPs of interest using the PRIMER3 PLUS software (Untergrasser *et al.* 2007). Primers were chosen (using default conditions) to amplify a 300–800-bp product that contained one or multiple putative SNPs. Amplification of the targeted regions was completed in a 50-μl reaction consisting of 10 μl of 5 × Green GoTaq reaction buffer (Promega Corp., Madison, WI, USA), 5 μl of 25 mM $MgCl_2$, 1 μl containing 10 mM each dNTP, 0.25 μl of GoTaq *Taq* polymerase, 5 μl each of 5-μM concentration of forward and reverse primers, 2 μl of DNA at a concentration of 10 ng/μl, and 21.75 μl of sterile deionized water. PCR parameters were 95 °C for 5 min; followed by 35 cycles of 95 °C for 1 min, 54 °C for 1 min, 72 °C for 1 min; and completed with 72 °C for 5 min followed by 4 °C. PCR products were purified using Qiaquick spin columns (Qiagen Inc., Valencia, CA, USA). For sequencing, 2 μl of 5 pmoles/μl primer was added to 2 μl of purified PCR product (20 ng/μl). Amplified products were sequenced with the BIGDYE version 3.1 ready reaction kit (Applied Biosystems) on an ABI 3730 automated sequencer at the University of Guelph Genomics Facility. Sequencing chromatograms were visualized, and sequences were aligned and edited using BioEdit (Hall 1999). The target regions were sequenced for each of the eight isolates that were used to create the pooled DNA for the initial 454 pyrosequencing run. Polymorphisms were validated by observing two nucleotides at the same locus, and for the SNP to be verified, the variant allele must have had a chromatogram peak at least 50% higher than the background peaks.

Annotation

Contigs containing potential SNPs along with all contigs >200 bp were used for Blastn and Blastx searches and gene ontology (GO) annotation. The purpose of the BLAST searches was to determine if the putative SNPs were associated with genes, not to provide a complete annotation of these sequences. Blastn searches (e-value cut-off, 10^{-5}) were performed against the non-redundant (nr) nucleotide database at NCBI. The program Blast2GO (Conesa *et al.* 2005) was used for Blastx searches (e-value cut-off, 10^{-5}) against the NCBI nr protein database to extract the GO terms associated with the Blast hits. The annotation parameters were a pre-e-value-Hit-Filter (10^{-6}), annotation cut-off threshold (55) and GO weight (5).

Analysis of markers for use in population genetics studies

To determine whether the markers would be suitable for larger population genetics studies, the 24 isolates were evaluated over all identified SNPs. The sequences of the five contigs in which SNPs were identified were aligned using ClustalW (Thompson *et al.* 1994) and then concatenated and collapsed into unique haplotypes using the SNAP Combine and SNAP Map functions, respectively, in SNAP workbench 2.0 (Price & Carbone 2005). The tests for neutrality were completed to determine whether observed genetic variation was consistent with the hypothesis that the majority of polymorphisms contributing to genetic variability are selectively neutral (Kimura 1983). Tajima's *D*; Fu and Li's *D* and *F*; and Fu's *F* neutrality statistics were tested using DNASP version 5 (Rozas *et al.* 2003). Tests for linkage disequilibrium (LD) were assessed using MULTILOCUS 1.2 (Agapow & Burt 2001) and used to detect the nonrandom association of alleles at different loci (Slatkin 1994). LD was assessed with each haplotype within a contig considered an allele at that locus. Significance was assessed using 1000 randomizations of the data set to create a null hypothesis.

Results

454 sequencing, assembly and annotation of Oc-j DNA

The sequencing run produced some 71 million bases comprising 217 311 filter passing wells. The Roche software, GS Assembler, successfully aligned 80% of the reads into 30 339 contigs comprising 16.1 million bases of sequence. Files containing these sequence reads and quality scores were submitted to the NCBI Short Read Archive (accession SRA023892.1). In addition, this Whole Genome Shotgun project was deposited at DDBJ/EMBL/GenBank under the accession AEGN00000000. The version described in this study is the first version and includes accession numbers AEGN01000000–AEGN01028868. Of the 30 399 contigs, 1452 were shorter than 200 bp and were removed from further analysis. There were 11 297 contigs that were classified as large contigs (>500 bp) comprising 9 588 425 bases. Of the large contigs, the average size was 848 bp and the longest contig was 31 049 bp.

All contigs >200 bp were annotated using Blast2Go, but only data from the 68 contigs that contained putative SNPs is presented here. Of the 68 contigs that contained putative SNPs, nine corresponded to previously described proteins (Blastx e-value $<10^{-5}$) based on nr protein searches. All these hits were to members of the Ascomycota including *Aspergillus oryze, Penicillium chrysogenum, Penicillium marneffei, Podospora anserina, Pyrenophora tritici-repentis, Saccharomyces cervisae, Sclerotinia sclerotiorum,* and *Talaromyces stipitatus* (Table 2). The Blastx search results indicated that all nine sequences represented regions from active or inactive transposable elements, which include both retrotransposons and DNA transposons. Gene descriptions that included the terms endonuclease, integrase, gag, pol, or reverse transcriptase were identified as putative transposable elements. Putative SNPs from four of the nine regions associated with putative transposable elements were visually confirmed with EAGLEVIEW; primers were designed for three of these fragments, and the fragments were sequenced. Loci that were polymorphic among the eight isolates were only found in contig 95 (Accession AEGN01000083). No SNPs were found in Contig 9396 (Accesion AEGN01008947). Polymophisms were found in the individual chromatograms of each of the eight isolates for contig 29332 (AEGN01028189). This information, along with the number of reads (41), probably indicated that this region is present in multiple copies within the genome and therefore is not a good candidate for marker development for phylogeographic studies. The remaining contigs had no significant matches to the Blastx database. Blastn searches using all contigs identified nine sequences as contamination, i.e. as vector, bacterial or human sequence, and these were removed from further analysis.

SNP detection and validation

GS reference mapper was able to detect a total of 422 SNPs and insertion/deletions within 80 contigs, of which, 283 single-nucleotide polymorphic loci within 64 contigs were detected (Table 2). Using EAGLEVIEW, the 64 contigs were manually examined, and 36 (13.7%) of these polymorphic sites within 13 contigs were confirmed (Table 2). Primers were successfully developed using Primer3 to amplify the regions containing polymorphic sites in 10 of the 13 contigs confirmed by EAGLEVIEW (Table 3). Primers could not be designed to amplify four polymorphic sites in three contigs because the polymorphic sites were too close to the end of the contig, or because inadequate data parameters existed for reliable primer design.

The fragments from the eleven contigs were amplified, sequenced and aligned. After manual examination of the alignments, it was observed that only five of the eleven contigs contained true polymorphisms. Therefore, of the 36 putative polymorphic sites within 13 contigs verified using EAGLEVIEW, 12 polymorphic sites (33.3% of EAGLEVIEW SNPs, 4.2% of GS Mapper SNPs) within five contigs were validated using standard sequencing methods (Tables 3 and 4). Four additional SNPs were identified when 16 more isolates of *Oc-j* were genotyped, for a total of 16 SNPs, of which 14 were phylogenetically informative (Tables 3 and 4). An informative single-nucleotide mutation is one that occurs in at least two haplotypes such that one evolutionary relationship among haplotypes can be favoured over the other possible relationship based on parsimony.

Amplification and sequencing of the five polymorphic regions was successful with all isolates evaluated, using the same primer set for both amplification and sequencing reactions (Table 3). Based on these 16 SNP makers, we identified seven distinct multilocus haplotypes within the sample of eight isolates used for 454 pyrosequencing (Table 4). When the additional set of 16 isolates from across North America was included, 17 haplotypes were detected from 24 isolates analysed. Five multilocus haplotypes (H7, H9, H10, H14 and H16) were found in multiple isolates: two isolates each for H7, H9 and H14, and three isolates each for H10 and H16 (Table 4).

All neutrality tests were nonsignificant, indicating these polymorphic loci are selectively neutral. Tests for linkage disequilibrium among the five regions found significant LD overall [index of association (I_A) = 0.59, P = 0.011]. However, in pairwise tests of the five regions, no significant LD was detected (P = 0.224), indicating recombination may have occurred between the five genomic regions in this population from North America.

Table 2 Summary of results from SNP detection performed with GS Reference Mapper on 454 pyrosequencing data of a pooled sample of genomic DNA from eight isolates of *O. clavigignenti-juglandacearum*. The contig accession numbers, number of SNPs detected, visually confirmed using the program EAGLEVIEW and validated by designing primers and sequencing of the SNP containing fragment are also included. Matches to the nonredundant protein (NCBI) database are indicated with accession number and description for the best hit and their E-values

Contig	WGS accession	GS mapper	EAGLEVIEW	Primers designed[a]	Validated	Sequence description	Acc. no.	E-value
58	AEGN01000050	4	1	Y	1			
86	AEGN01000075	1	0	*	0			
95	AEGN01000083	8	5	Y	4	Reverse transcriptase	EDV08562.1	7.00E−12
99	AEGN01000085	14	0	*	0			
347	AEGN01000279	3	0	*	0			
419	AEGN01000334	9	0	*	0			
607	AEGN01000472	18	0	*	0			
634	AEGN01000490	4	0	*	0	Reverse transcriptase	EED11627.1	5.85E−05
1053	AEGN01000804	21	0	*	0			
1481	AEGN01001208	9	0	*	0	Transposase/integrase	XP_001820824.1	6.47E−08
1499	AEGN01001224	2	0	*	0			
1649	AEGN01001369	12	1	N	0			
1650	AEGN01001370	4	2	Y	0			
2974	AEGN01002662	2	0	*	0			
4006	AEGN01003654	1	0	*	0			
4753	AEGN01004387	2	0	*	0			
5389	AEGN01005015	2	4	Y	0			
7219	AEGN01006813	1	0	*	0			
9396	AEGN01008947	1	1	Y	0	Reverse transcriptase	XP_002153683	2.00E−34
10499	AEGN01010034	1	0	*	0			
11350	AEGN01010867	1	0	*	0			
11586	AEGN01011103	10	0	*	0			
15150	AEGN01014614	2	0	*	0			
16083	AEGN01015532	2	0	*	0			
16990	AEGN01016427	3	0	*	0			
18831	AEGN01018240	3	0	*	0			
21414	AEGN01020801	1	0	*	0	Gag-pol polyprotein	ACD86393.1	4.00E−38
23831	AEGN01023193	3	0	*	0			
23927	AEGN01023287	1	0	*	0			
25987	AEGN01025320	1	0	*	0			
27463	AEGN01026742	1	0	*	0	Reverse transcriptase	XP_001937171.1	2.00E−08
27820	AEGN01027062	7	0	*	0			
28204	AEGN01027397	1	0	*	0			
28311	AEGN01027481	1	1	N	0	Rnase H (endonuclease)	CAP79442.1	1.00E−09
28545	AEGN01027633	9	0	*	0			
28700	AEGN01027735	3	0	*	0			
28749	AEGN01027767	1	0	*	0	Reverse transcriptase	XP_001937171.1	1.00E−23
28876	AEGN01027858	4	0	*	0			
28966	AEGN01027928	2	0	*	0			
29085	AEGN01028023	2	0	*	0			
29105	AEGN01028040	1	2	N	0			
29106	AEGN01028041	9	3	Y	2			
29169	AEGN01028083	3	0	*	0			
29332	AEGN01028189	9	8	Y	0	Reverse transcriptase	XP_001588647.1	2.00E−11
29375	AEGN01028221	21	0	*	0			
29380	AEGN01028224	2	0	*	0			
29385	AEGN01028229	10	0	*	0			
29442	AEGN01028260	1	0	*	0			
29664	AEGN01028394	6	0	*	0			
29665	AEGN01028395	2	0	*	0			
29747	AEGN01028437	2	0	*	0			
29760	AEGN01028444	1	0	*	0			

Table 2 Continued

Contig	WGS accession	GS mapper	EAGLEVIEW	Primers designed[a]	Validated	Sequence description	Acc. no.	E-value
29920	AEGN01028549	5	0	*	0			
30024	AEGN01028638	1	0	*	0			
30086	AEGN01028699	4	0	*	0			
30105	AEGN01028716	2	3	Y	2			
30109	AEGN01028719	1	0	*	0			
30206	AEGN01028775	1	2	Y	0			
30257	AEGN01028813	10	0	*	0			
30281	AEGN01028829	4	0	*	0			
30300	AEGN01028839	3	0	*	0			
30307	AEGN01028845	4	0	*	0			
30326	AEGN01028860	3	3	Y	3			
30337	AEGN01028867	1	0	*	0			
Totals	66 contigs	283	36		12			

[a]The (*) indicates that primers were not designed for these regions as no putative SNPs were visually validated for these contigs.

Table 3 Characterization and description of 16 single-nucleotide polymorphic (SNP) markers for *Ophiognomonia clavigignenti-juglandacearum*, including the location, locus name, primer sequences, annealing temperatures, fragment size, Genbank accession number and polymorphism description including no. of SNPs, position within the amplified fragment and frequency of the rare allele. For SNP identity, the most frequent allele is listed first

Contig	Locus	Primer sequence (5′–3′)	T_m	Product size (bp)	No. of SNPS	Position	Identity	Rare allele frequency	GenBank acc. no.
58	OCJ1	F: CAAGAAGCCGGTATAACGAGA	59 °C	276	1	163	G/C	0.083	AEGN01000050
		R: AATAGAGAATAGATCCCAAGTTTTT	55 °C						
95	OCJ2	F: TGAAACTGGAATAAACGCTCTA	56 °C	549	6	136	A/G	0.166	AEGN01000083
	OCJ3	R: GCCTTATTAACGAAGGCTTTA	55 °C			154	T/G	0.291	
	OCJ4					243	C/T	0.125	
	OCJ5					289	T/C	0.166	
	OCJ6					373	T/C	0.083	
	OCJ7					470	T/C	0.166	
29106	OCJ8	F: ATCGAGATATAACTATTAATGCAA	52 °C	177	3	40	C/T	0.458	AEGN01028041
	OCJ9	R: GGGCGTGGAAAGTCTGAGT	60 °C			79	G/A	0.208	
	OCJ10					138	T/C	0.125	
30105	OCJ11	F: TGCGGAATACAGCTTTTCCTA	60 °C	427	3	115	A/T	0.125	AEGN01028716
	OCJ12	R: ATGCCACGTCTCTGTACGAC	59 °C			223	G/A	0.208	
	OCJ13					329	A/G	0.250	
30326	OCJ13	F: GCTGCTATTATCGGCTATGGA	59 °C	323	3	176	A/G	0.042	AEGN01028860
	OCJ15	R: TTTTCTTCAGAATCTACTGTTTCAAAG	59 °C			180	G/A	0.042	
	OCJ16					206	G/A	0.125	

Further analysis of more isolates will be required to determine whether *Oc-j* has undergone sexual recombination.

Discussion

Our analysis of 16 Mbp-consensus sequence of *Oc-j* genomic DNA sequenced with ¼ of a 454 sequencing run demonstrates that short reads produced with pyrosequencing technology can be assembled *de novo* into reasonably long contigs (>200 bp). We were able to demonstrate that the detection of valid SNPs is possible by sequencing a pooled sample of polymorphic genotypes. By aligning the reads of genomic DNA from eight isolates of *Oc-j*, we were able to detect 283 SNPs and validate 36 (13%) *in silica*. After eliminating 87% of the putative SNPs, we were able to validate 16 (5%) using standard Sanger sequencing. Therefore, approximately 267 SNPs may have been false positives, possibly arising from sequencing errors or alignment of paralogs. Paralog sequences may have been

Table 4 Haplotypes characterized using 16 single-nucleotide polymorphic markers on 24 isolates of *O. clavigignenti-juglandacearum* from across North America, including the type culture from Wisconsin (SCJ1)

		Contig 58	Contig 00 095						Contig 29 106			Contig 30 105			Contig 30 326		
		163*	136	154	243	289	373	470	40	79	138	115	223	329	194	198	224
		G†	A	T	C	T	T	T	C	G	T	A	G	A	A	G	G
Isolate	Haplotype	I‡	i	i	i	i	i	i	i	i	i	i	i	i	–	–	i
SCJ1	H1	–	–	G	–	–	–	–	T	–	C	–	–	–	–	–	–
SCJ2	H2	–	–	G	–	C	–	–	T	–	–	–	–	–	–	–	–
SCJ3	H3	C	G	G	–	–	–	C	T	–	C	–	A	–	–	–	–
P005	H4	–	–	–	–	–	–	–	–	–	–	–	A	–	–	–	–
WB22	H5	–	–	–	–	C	–	–	–	–	–	–	–	–	G	A	–
GA11	H6	–	–	G	–	–	–	–	–	–	–	–	–	G	–	–	–
Bud23	H7	–	G	G	–	–	–	C	–	–	–	–	–	G	–	–	A
GA51	H7	–	G	G	–	–	–	C	–	–	–	–	–	G	–	–	A
P013	H8	C	–	–	–	–	–	–	T	–	–	T	A	–	–	–	–
P017	H9	–	–	–	–	–	–	–	T	–	–	T	A	–	–	–	–
P045	H9	–	–	–	–	–	–	–	T	–	–	T	A	–	–	–	–
P019	H10	–	–	–	–	–	–	–	T	–	–	–	–	–	–	–	–
P029	H10	–	–	–	–	–	–	–	T	–	–	–	–	–	–	–	–
P037	H10	–	–	–	–	–	–	–	T	–	–	–	–	–	–	–	–
P034	H11	–	G	G	–	–	–	C	T	–	–	–	–	G	–	–	A
P043	H12	–	–	–	–	–	–	–	T	–	C	–	–	–	–	–	–
70BW2	H13	–	–	–	T	–	–	–	–	A	–	–	–	–	–	–	–
1368-1C	H14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
AR4534	H14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
AR4537	H15	–	–	–	T	C	C	–	–	A	–	–	–	G	–	–	–
AR4538	H16	–	–	–	–	–	–	–	–	A	–	–	–	–	–	–	–
AR4539	H16	–	–	–	–	–	–	–	–	A	–	–	–	–	–	–	–
HN1	H16	–	–	–	–	–	–	–	–	A	–	–	–	–	–	–	–
AR4540	H17	–	–	–	T	C	C	–	–	–	–	–	–	G	–	–	–

*The position within the amplified fragment where the variable allele is located.
†Most frequent or consensus nucleotide for each variable position: A, C, G or T.
‡Character type indicates whether each variable position is phylogenetically informative (i) or noninformative (–). To be informative, a position must have mutations that appear in at least two haplotpyes.

assembled in the same contig because they could not be distinguished because of the short read length typically produced by 454 pyrosequencing.

A significant number of polymorphisms in our sample may have been overlooked because of the stringent methodology used to call SNPs. For contigs with reasonable sequence coverage (>200 bp) and depth (>10 reads on average per nucleotide), we detected one SNP for every 599 276 bp on average. This frequency is much lower than was present in other fungi including *Fusarium graminearum* which was found to have 10 495 SNPs between two isolates with a SNP density of 0–17.5 SNPs per kb (Cuomo *et al.* 2007). There are several likely reasons we identified such a small number of SNPs in *Oc-j*. First, the complete genome (36.1 Mb) of *F. graminearum* was sequenced using Sanger sequencing, which produces fewer false-positives for SNPs. Second, 454 pyrosequencing was used to randomly sequence genomic DNA, not all genotypes were sequenced for every SNP locus. The average sequencing depth was 5.4 reads per base pair, which was less than the number of possible haplotypes in our sample (eight). Additionally, in the case of *F. graminearum*, it was found that the distribution of SNPs was biased, i.e. 25% of SNPs were found in 5% of the genome, and 50% of SNPs were found in 13% of the genome sequence (Cuomo *et al.* 2007). Therefore, a large number of SNPs may have been lost simply because they were not included in the analysis. Finally, the genetic relatedness of the eight isolates used in this study was unknown. As previously mentioned, this organism is believed to be recently introduced into North America and clonally propagated. Additionally, seven of the eight isolates were collected from Ontario, possibly limiting the amount of genetic diversity sampled compared to

what exists throughout the range of the fungus in North America and potentially leading to a SNP discovery bias (Pearson *et al.* 2004).

Compared with traditional methods of SNP discovery, the random whole-genome shotgun approach was an inexpensive and rapid technique to generate a large number of putative SNPs and a method for *in silico* evaluation to eliminate false SNPs, reducing the need for validation using standard sequencing protocols. This method of essentially screening the entire genome for candidate SNPs may provide a preferable alternative to other (more targeted) methods that focus on specific genes or coding regions which may be under evolutionary selection. These anonymous nuclear markers (ANMs) may be preferential to nuclear protein coding loci for studies directed at understanding the phylogeography and population genetics of a particular organism (Thomson *et al.* 2010). For phylogeographic studies and phylogenetic analysis of rapid radiations, finding nuclear markers with sufficient variation remains a major challenge (Brito & Edwards 2009). Two methods currently used to develop ANMs are the development of small insert libraries from sheared genomic DNA, followed by sequencing clone inserts from the library (Rosenblum *et al.* 2007) or through the conversion of AFLP bands into single locus markers (Brugmans *et al.* 2003). However, both these methods require more time and generate fewer markers than the method developed in this study. One of the great benefits of developing ANMs using the whole-genome screen method is the auxiliary data generated. This includes the approximately 16 Mb of genome data that currently is being used to assemble the *Oc-j* mitochondrial genome and to identify microsatellite markers, transposons and gene homologs that were identified through BLAST searches.

The markers developed in this study will be of value to studies of the population structure and phylogeography of *Oc-j*, an invasive fungus found throughout eastern North America. This will be an important advancement, as there was no previous knowledge of the genotypic diversity of this fungus prior to this study. In addition, the only previous study to evaluate the genetic diversity of this pathogen used RAPD markers and found no genetic difference among any of the isolates evaluated (Furnier *et al.* 1999). Finally, the protocol developed in this study to identify SNP markers in *Oc-j* is flexible and provides a framework for the inexpensive and rapid development of makers for population genetic studies for a number of nonmodel organisms.

Acknowledgements

We thank Richard Wilson, Ontario Ministry of Natural Resources (OMNR) for assistance in locating infected trees in Ontario, OMNR and the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding of the project, the R. J. Hilton Centre of the University of Guelph Arboretum and the RARE Charitable Research Reserve for access to butternut trees.

References

Agapow PM, Burt A (2001) Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, **1**, 101–102.

Brito P, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.

Broders KD, Boland GJ (2011) Reclassification of the butternut canker fungus, *Sirococcus clavigignenti-juglandacearum*, into the genus *Ophiognomonia*. *Fungal Biology*, **115**, 70–79.

Brugmans B, van der Hulst RGM, Visser RGF, Lindhout P, van Eck HJ (2003) A new and versatile method for the successful conversion of AFLP markers into simple single locus markers. *Nucleic Acids Research*, **31**, e55–e55.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Ching A, Caldwell KS, Jung M *et al.* (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics*, **3**, 14.

Conesa A, Gotz S, Garcia-Gomez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Craig DW, Pearson JV, Szelinger S *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods*, **5**, 887–893.

Cuomo CA, Gueldener U, Xu JR *et al.* (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.

Furnier GR, Stolz AM, Mustaphi RM, Ostry ME (1999) Genetic evidence that butternut canker was recently introduced into North America. *Canadian Journal of Botany*, **77**, 783–785.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Window 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Huang WC, Marth G (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Research*, **18**, 1538–1543.

Jakobsson M, Scholz SW, Scheet P *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.

Keim P, Van Ert MN, Pearson T *et al.* (2004) Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infection Genetics and Evolution*, **4**, 205–213.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Lambreghts R, Shi M, Belden WJ *et al.* (2009) A high-density single nucleotide polymorphism map for neurospora crassa. *Genetics*, **181**, 767–781.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

Marth GT, Korf I, Yandell MD *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, **23**, 452–456.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, 5.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 14.

Ostry ME, Woeste K (2004) Spread of butternut canker in North America, host range, evidence of resistance within butternut populations and conservation genetics. *General Technical Report NC*, **243**, 114–120.

Parameswaran P, Jalili R, Tao L *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, **35**, 9.

Pearson T, Busch JD, Ravel J *et al.* (2004) Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proceedings of the National Academy of Sciences, USA*, **101**, 13536–13541.

Price EW, Carbone I (2005) SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics*, **21**, 402–404.

Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**, 179–181.

Ravel C, Praud S, Murigneux A *et al.* (2006) Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). *Genome*, **49**, 1131–1139.

Rosenblum EB, Belfiore NM, Moritz C (2007) Anonymous nuclear markers for the eastern fence lizard, Sceloporus undulatus. *Molecular Ecology Notes*, **7**, 113–116.

Rozas J, Sánchez-DelBarrio JC, Messenguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Shen YJ, Jiang H, Jin JP *et al.* (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiology*, **135**, 1198–1205.

Shen YF, Wan ZZ, Coarfa C *et al.* (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, **20**, 273–280.

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. *Genetics*, **137**, 331–336.

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalities and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Thomson RC, Wang IJ, Johnson JR (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology*, **19**, 2184–2195.

Untergrasser A, Hijveen H, Rao X *et al.* (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, **35**, 71–74.

Zhu YL, Song QJ, Hyten DL *et al.* (2003) Single-nucleotide polymorphisms in soybean. *Genetics*, **163**, 1123–1134.