Title

Indel Group in Genomes (IGG) Molecular Genetic Markers

Permalink
https://escholarship.org/uc/item/6dc1r8q4

Journal
Plant Physiology, 172(1)

ISSN
0032-0889

Authors
Toal, Ted W
Burkart-Waco, Diana
Howell, Tyson
et al.

Publication Date
2016-09-01

DOI
10.1104/pp.16.00354

Peer reviewed

Plant Physiology Preview. Published on July 19, 2016, as DOI:10.1104/pp.16.00354

Short title:

IGG Molecular Genetic Markers[a]

Corresponding author:

Siobhan M. Brady[h,l,n]

Article title:

Indel Group in Genomes (IGG) Molecular Genetic Markers

Author names and affiliations:

1. Ted Toal[a,h,j,], ORCID ID:0000-0002-0967-4998, Department of Biochemistry and Molecular Medicine, UC Davis
2. Diana Burkart-Waco[b,i,k], Department of Plant Sciences, UC Davis
3. Tyson Howell[c,k], , Department of Plant Sciences, UC Davis
4. Mily Ron[d,l], Department of Plant Biology, UC Davis
5. Sundaram Kuppu[d,l], Deparment of Plant Biology, UC Davis
6. Anne Britt[e,l], Department of Plant Biology, UC Davis
7. Roger Chetelat[f,i,k], Department of Plant Sciences, UC Davis
8. Siobhan M. Brady[g,h,l,m,n], Department of Plant Biology and Genome Center, UC Davis

One sentence summary:

Genome-wide molecular markers are produced by a bioinformatics pipeline that analyzes pairs of genomic sequences to find primer pairs that amplify indel-containing regions having a targeted amplicon size and size difference.

Footnotes:

List of Author Contributions

[a]T.W.T. conceived the project, developed the original algorithm and wrote the article with contributions of all the authors.

[b]D.B.W. tested and used the markers for introgression line development.

[c]T.H. tested the software on different platforms and reviewed documentation.

[d]M.R. and S.K. performed experimental testing of markers.

[e]A.B. supervised experimental testing.

[f]R.C. supervised the introgression line development.

[g]S.M.B. contributed to experimental design and writing the paper.

[h]S.M.B. and T.T. were funded by NSF-IOS-18085.

[i]Work of D.B.W. and R.C. on *S. sitiens* marker development was supported in part by the USDA Agriculture and Research Initiative grant no. 2013-67013-21131.

[j]Department of Biochemistry and Molecular Medicine, UC Davis, 1 Shields Ave, Davis, CA 95616

[k]Department of Plant Sciences, UC Davis, 1 Shields Ave, Davis, CA 95616

[l]Department of Plant Biology, UC Davis, 1 Shields Ave, Davis, CA 95616

[m]Department of Plant Biology and Genome Center, UC Davis, 1 Shields Ave, Davis, CA 95616

[n]sbrady@ucdavis.edu

Indel Group in Genomes (IGG) Molecular Genetic Markers

Ted W. Toal, Diana Burkart-Waco, Tyson Howell, Mily Ron, Sundaram Kuppu, Anne Britt, Roger Chetelat, Siobhan M. Brady

Abstract

Genetic markers are essential when developing or working with genetically variable populations. IGG (Indel Group in Genomes) markers are primer pairs which amplify single-locus sequences that differ in size for two or more alleles. They are attractive for their ease of use for rapid genotyping and their co-dominant nature. Here we describe a heuristic algorithm that uses a k-mer based approach to search two or more genome sequences to locate polymorphic regions suitable for designing candidate IGG marker primers. As input to the IGGPIPE (IGG pipeline) software, the user provides genome sequences and the desired amplicon sizes and size differences. Primer sequences flanking polymorphic indels are produced as output. IGG marker files for three sets of genomes: *Solanum lycopersicum/S. pennellii*, *Arabidopsis thaliana* Col-0/Ler-0 accessions, and *S. lycopersicum/S. pennellii/S. tuberosum (three-way polymorphic)* are included.

Introduction

Genetic differences or DNA polymorphisms between individuals in a population are a primary cause of phenotypic variation. A critical step in characterizing the genetic basis of such phenotypic variation is the development of molecular genetic markers that enable detection and identification of polymorphisms. Four properties describe a marker: the polymorphism it finds, the assay method used for detecting it, the number of alleles identifiable at one locus, and the number of different loci at which alleles can be found. As new assays revealed increasing numbers of DNA polymorphisms, new types of markers were developed to detect them, each with its own acronym, including these common polymorphisms and representative types of markers: SNPs or Single Nucleotide Polymorphisms (SSCP-Single Strand Conformation Polymorphism markers (Orita et al. 1989; Wenzl et al. 2004)), insertions/deletions (indels) of varying lengths (SCAR-Sequence Characterized Amplified Region markers (Paran and Michelmore 1993; Robarts and Wolfe 2014)), restriction site locations (RFLP-Restriction Fragment Length Polymorphism markers (Botstein et al. 1980; Konieczny and Ausubel 1993; Vos et al. 1995; Miller et al. 2007)), tandem repeat counts (VNTR-Variable Number Tandem Repeat markers(Nakamura et al. 1987)), and differences in polynucleotide repeat counts or lengths (SSR-Simple Sequence Repeat markers (Weber and May 1989; Zietkiewicz, Rafalski, and Labuda 1994; Huang et al. 1991; Dietrich et al. 1992)). A more complete list of markers and their properties is given in **Table 1**.

Historically, visualization of polymorphic markers typically used restriction digests, Southern hybridization, and polyacrylamide gel electrophoresis, augmented later with PCR, agarose gel with ethidium bromide staining, Sanger sequencing, and high-throughput genotyping using microarray technology and next generation sequencing. *Allele-specific* marker assays detect a single allele to provide simple yes-no output, while *codominant* marker assays are able to detect the two different polymorphic states present in a heterozygote at the target locus. A marker is *multi-allelic* if it is able to discriminate between many different polymorphisms in a population. Finally, a marker assay may visualize allele(s) at a *single locus* (used for linkage mapping a locus, for example) or at *multiple loci* simultaneously (used for fingerprinting individuals in a population, for example). The different properties of markers make each type useful in particular applications. The uses of markers span a broad range, from simple genotyping in the lab to areas as diverse as marker-assisted selection (Li et al. 2015), trait association mapping (Nachimuthu et al. 2015), ecology (Pradhan et al. 2015), synteny studies (Guyon et al. 2010), diversity surveys (Salehi, Gottstein, and Haddadzadeh 2015), species authentication (Fu et al. 2015), sex determination (Kafkas et al. 2015), detection of adulterants (Marieschi, Torelli, and Bruni 2012), ingredient traceability (Ahmed, Ferreira, and Hartskeerl 2015), and forensics (Diegoli 2015).

Marker assays can vary in scoring complexity. For instance, CAPS (Cleaved Amplified Polymorphic Sequence, (Konieczny and Ausubel 1993)) markers allow for characterization of multi-allelic polymorphisms, but are relatively low-throughput, as they require an additional digestion step after PCR amplification. Indel markers (Rafalski 2002; Shen et al. 2004) are usually described as a pair of PCR primers binding to single-copy (unique in the genome) sites flanking a single small indel whose size ranges from one to 100 base pairs. The assay requires PCR followed by either a high-percentage agarose gel or (especially for very small indels) a high-resolution polyacrylamide gel. As an example of indel marker amplicon sizes, over 100,000 rice genome indel markers were designed by (Liu et al. 2015) using an exhaustive genome search for single-copy primers which were then aligned to sequence reads to identify polymorphic primer pairs in different rice varieties. These markers have amplicon sizes of no more than 300 bp (mean 218 bp using 150-300 bp table) and size differences between target genotypes of 6 to 100 bp (mean 51 bp). While indels make attractive marker targets because of ease of scoring and the absence of a digest step, the current set of available markers in *Arabidopsis*, tomato, and rice is lacking because often small differences in amplicon sizes can make resolution of genotyping difficult. Thus, there is a need for more high-throughput markers, with easy-to-score length polymorphisms between target genotypes.

Historically, single-locus molecular genetic markers have been developed a few at a time for a specific species and genetically segregating population, often starting with a search for genetic polymorphism using a technique such as random amplification of polymorphic DNA (RAPD) followed by sequencing of amplicons and designing primers specific to them. With the advent and increasing use of next generation sequencing, the number of organisms with sequenced genomes is rising rapidly (Reddy et al. 2015). When genomes (or portions thereof) are available for two or more genetically different but crossable species, subspecies, or accessions, they can be searched *in silico* for polymorphic regions suitable for making genetic markers to genotype polymorphic regions in progeny.

Custom software tools have been used to develop marker sets from whole genome data; however, general-use open-access community software for whole genome marker development is limited. Available tools include IMDP-Indel Markers Development Platform, for indel markers (Lu et al. 2015), PolyMarker for generating SNP-specific primers around known SNPs (Ramirez-Gonzalez, Uauy, and Caccamo 2015), ESMP-EST SSR Marker Pipeline, for finding short sequence repeats for designing SSR markers

(Sarmah et al. 2012), CapsID-CAPS IDentifier, for CAPS markers (Taylor and Provart 2006), and a wet-lab based marker array protocol using unsequenced whole genome data to make RAD (Restriction site-associated DNA) markers (Miller et al. 2007).

In principle, high throughput sequencing-could be used for genotyping purposes as opposed to PCR-based markers. However, there are several limitations on genotyping by sequencing that make PCR-based marker genotyping an equally efficient and affordable method. Next generation sequencing (NGS) is currently suited for generating extensive SNP data, potentially on hundreds to thousands of individuals. However, it can be cost-prohibitive due to the computational power needed for demultiplexing and performing parallel alignments, and in some cases due to a need for extensive bioinformatics support which is not feasible in terms of skills or finances for all labs. Making non-reference based alignments for genome or contig assembly and subsequent marker identification using NGS requires memory resources and computational power that often exceeds resources available (Kleftogiannis, Kalnis, and Bajic 2013; Salzberg et al. 2012). Furthermore, fine-mapping genes from QTL, even with NGS as a tool, still requires validation with PCR-based markers. Finally, generation of a mapping population rapidly with fixed genomic intervals could be done much more quickly through the use of PCR-based markers, rather than preparing multiple sequencing libraries and waiting sometimes months for sequencing data to return and be analyzed.

We present IGGPIPE (IGG pipeline), a command-line based pipeline that uses a search algorithm and common, unique (single-copy) k-mers to sift through multiple target genomes and identify up to thousands of candidate IGG markers, in some cases multi-allelic, *in silico*. IGG markers are benchmarked using cultivated tomato (*Solanum* *lycopersicum*) and *S. pennellii*, and *Arabidopsis thaliana* accessions Col-0 and Ler-0. We further present IGG marker sets polymorphic between *S. lycopersicum/S. pennellii; A. thaliana* accessions Col-0/Ler-0; and *S. lycopersicum*/*S. pennellii/S. tuberosum* and describe how the latter set is being utilized to develop a *S. lycopersicum* × *S. sitiens* introgression line population.

Results

Development of the IGGPIPE pipeline

Identification of unique k-mers

The premise underlying IGG markers is that k-mers of sufficient size should often occur as a single copy in a genome, and when occurring in conserved locations, will often occur as a single copy in both (or all) genomes under consideration. We call these *common unique k-mers,* common to both genomes, and unique (single copy) within each genome. We reasoned that common unique k-mers could be used to identify conserved regions within contigs in all species, and by testing for differences in distance between same-contig common unique k-mers among the genomes, we could discover regions containing length polymorphisms flanked by conserved sequences. These polymorphic regions must contain one or more indels, the requirement for designing a length-polymorphic PCR marker. The IGGPIPE pipeline (**Figure 1A**) was built around this concept.

We began by assessing the number of unique k-mers in a genome as a function of k (**Figure 1B-D**). Regardless of the value of k, a genome contains about the same total number of k-mers as nucleotides, since a k-mer starts at every base pair except those less than k from the end of a chromosome or contig (**Figure 1B**). Using the *S. lycopersicum* (tomato) and *S. pennellii* (a tomato wild relative) genomes (Tomato Genome 2012; Bolger et al. 2014; Bombarely et al. 2011), we counted unique k-mers and common unique k-mers for k ranging from 10 to 20 (**Figure 1C**). The closely related genomes had about the same number of unique k-mers and the number common between the two was roughly 1/3 of the total.

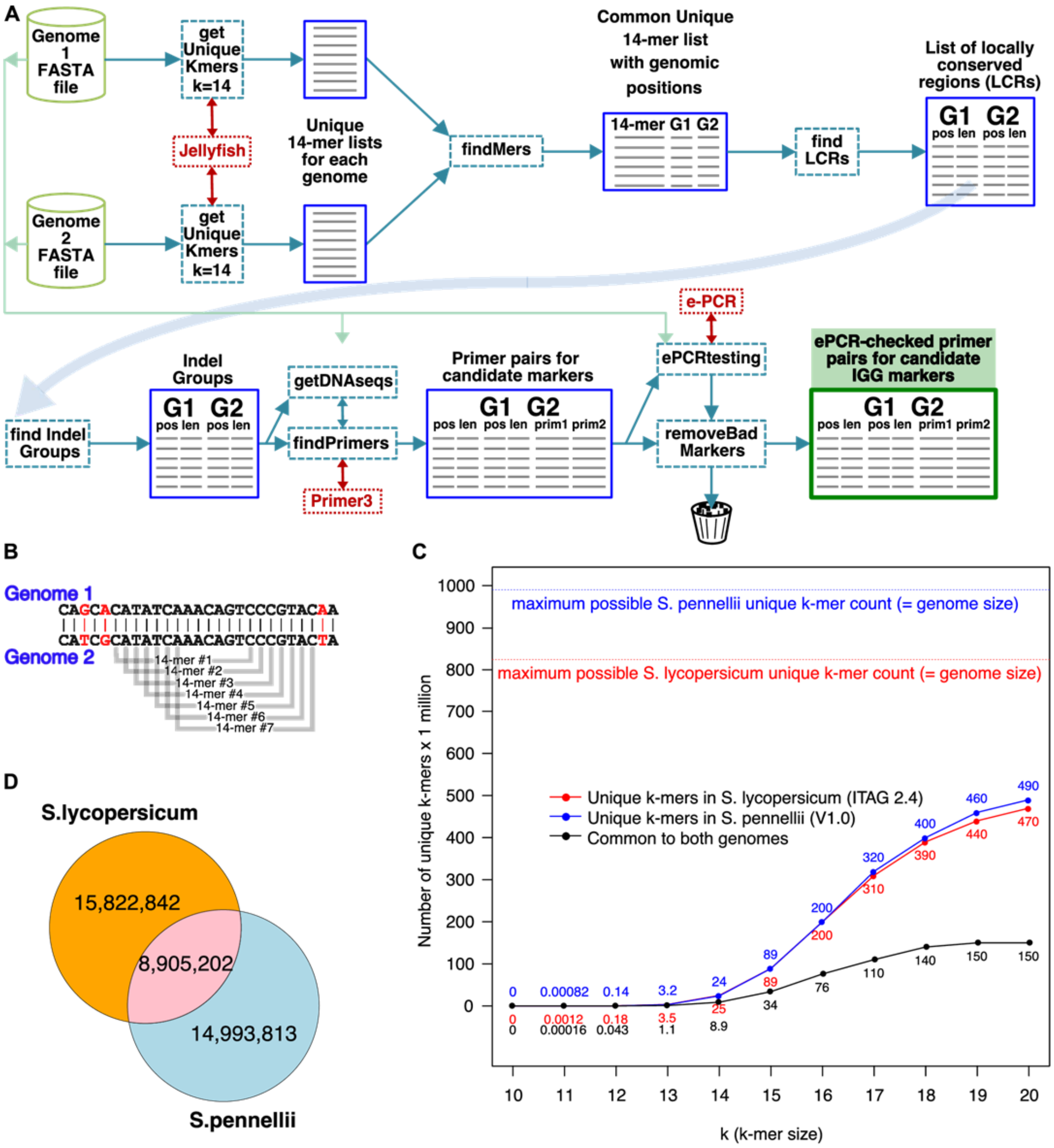

Figure 1 A. IGGPIPE: an IGG (Indel Group in Genomes) marker finder software pipeline. Two genome sequences (G1 and G2) are analyzed for common unique k-mers that identify locally conserved regions (LCRs), some of which are polymorphic for length, containing one or more indels between flanking conserved sequences, making them *Indel Groups*. Primers are designed in the flanking conserved regions and verified with e-PCR to produce candidate IGG markers. Pipeline software is shown in dashed boxes, data in solid line boxes. **B.** A new k-mer starts at each base position. Shown here are seven consecutive 14-mers common to two genomes. **C.** Number of unique k-mers in tomato (*S. lycopersicum*) and closely related *S. pennellii* species as a function of k, and number of unique k-mers common to both species. As k increases, the number of unique k-mers increases, gradually approaching the genome size limit. The common unique k-mer count does not keep increasing, but at some value of k will reach a peak, here around k=19 or k=20. **D.** With k=14, *S. lycopersicum*) and *S. pennellii* have almost 9 million unique k-mers in common between them.

By testing increasing values of k, we found that k=14 provided 8.9 million common unique k-mers (**Figure 1D**) between *S. lycopersicum* / *S. pennellii*, and that this number was sufficient to produce a few thousand IGG markers at the end of the pipeline, while k=14 was small enough to reduce computational and memory load to satisfactory levels for our needs.

The identification of conserved regions is complicated by several features of genome architecture, some of which are illustrated in **Figure 2A**, where each small black line represents a common unique k-mer. One or two k-mers lying on the same genome 1 contig and the same genome 2 contig may not indicate a contiguous length of conserved sequence but may be a random occurrence (see **Supplemental Materials and Methods** for an estimate of the random frequency of occurrence of common unique k-mers), illustrated by the shaded red boxes (a, b, e). When there is more than one k-mer, if their ordering in one genome matches the ordering in the other genome, then as the number of such k-mers increases, so too does the likelihood that they lie within conserved sequence. When a group of at least KMIN (a user-settable parameter typically set to a value from two to four) common unique k-mers has the same ordering on a single contig in each genome, we call the region containing them an LCR, or *locally conserved region*. LCRs are illustrated by shaded blue boxes (c, d, f, g, h, i, j) in **Figure 2A**. A group of k-mers in an LCR may encompass regions of equal lengths in both genomes (c, d, j), or the lengths may be unequal because the genomes contain indels, whose locations are shown with loop-outs of the DNA (f, g, h, i). These LCRs containing indels are the length-polymorphic regions used for generating IGG markers, and are shown as shaded blue boxes with borders.

Our algorithm for LCR identification, findLCRs, seeks groups of common unique k-mers in consecutive order on the same contig pair and satisfying parameter constraints, while *ignoring* all other common unique k-mers (even if they are interspersed among the group

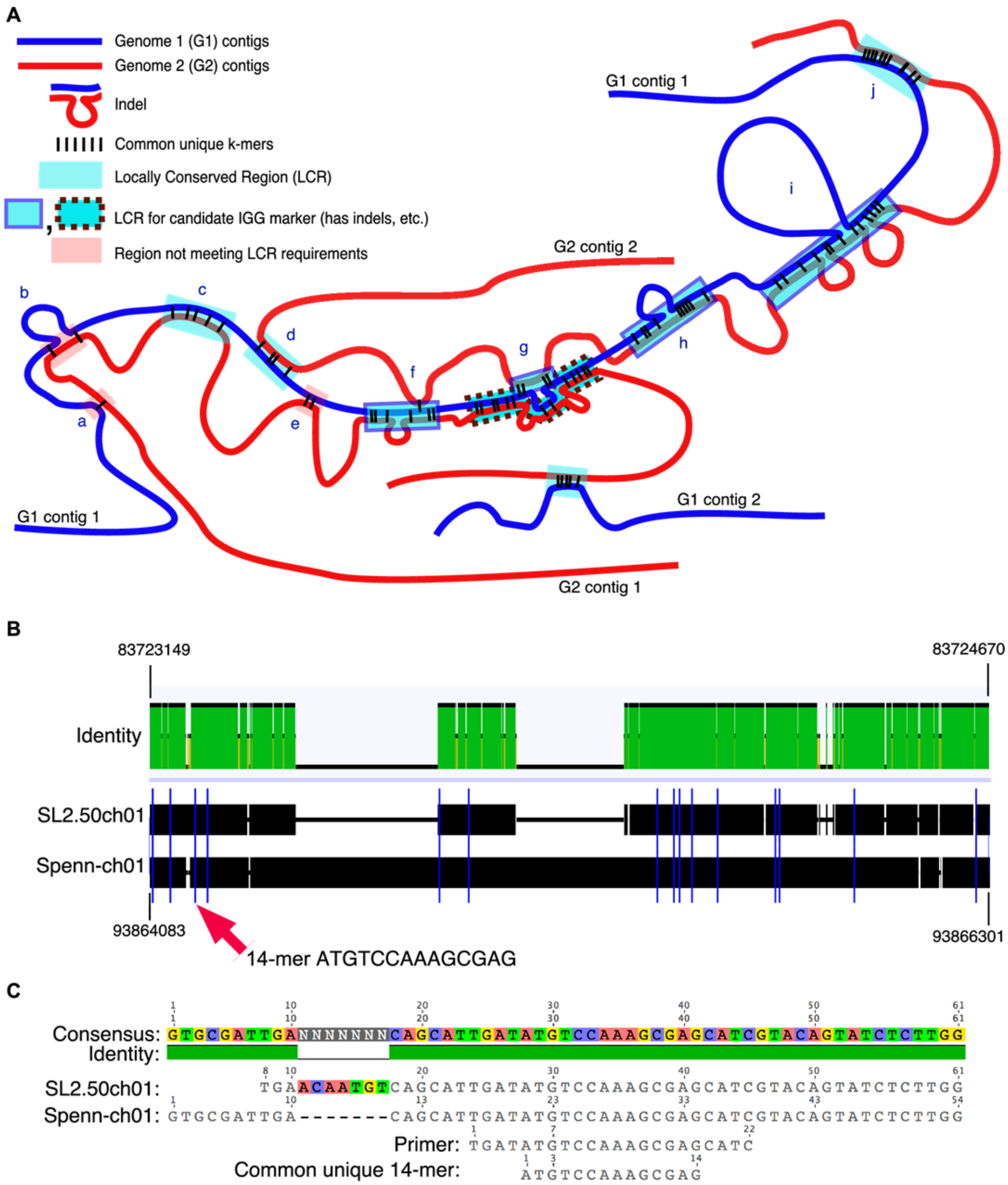
A
Genome 1 (G1) contigs
Genome 2 (G2) contigs
Indel
Common unique k-mers
Locally Conserved Region (LCR)
LCR for candidate IGG marker (has indels, etc.)
Region not meeting LCR requirements
G1 contig 1
G2 contig 2
G1 contig 2
G2 contig 1
a
b
c
d
e
f
g
h
i
j
B
83723149
83724670
Identity
SL2.50ch01
Spenn-ch01
93864083
93866301
14-mer ATGTCCAAAGCGAG
C
Consensus: GTGCGATTGANNNNNNNCAGCATTGATATGTCCAAAGCGAGCATCGTACAGTATCTCTTGG
Identity:
SL2.50ch01: TGAACAATGTCAGCATTGATATGTCCAAAGCGAGCATCGTACAGTATCTCTTGG
Spenn-ch01: GTGCGATTGA-------CAGCATTGATATGTCCAAAGCGAGCATCGTACAGTATCTCTTGG
Primer: TGATATGTCCAAAGCGAGCATC
Common unique 14-mer: ATGTCCAAAGCGAG

Figure 2. A. Locally conserved regions (LCRs) are regions of paired contigs within the genomes under consideration (here G1 and G2) having a sufficient number and spacing of unique k-mers in common between the contigs. When indels are present within LCRs, they form the basis for creating candidate IGG markers. Common unique k-mers can connect pairs of contigs in many ways. The parameter DMAX is the maximum spacing between two adjacent k-mers of the same LCR, and k-mers farther apart than that are assigned to different LCRs. If the number of k-mers is less than parameter KMIN (here assumed to be 4), the k-mers are assumed to be random common unique k-mers not signifying a conserved region, and no LCR is called for that region (a, b, e). LCRs may have no indels in them (c, d, j) or there may be a single indel (b, f, h) or more than one (i). Different LCRs along a contig of one genome might include *different* contigs in the other genome (a, b, c, and e versus d). Some LCR regions may have one or more random interspersed k-mers connecting a contig pair that is different from the contig pair of the LCR (f). Some regions may have complex overlapping of more than one LCR (g). **B.** An alignment of *S. lycopersicum* and *S. pennellii* genomes in the region of an LCR on chromosome 1. Blue vertical lines are positions of common unique 14-mers. An indel is visible that might provide sufficient length polymorphism for an IGG marker surrounding this area. Red arrow points to one 14-mer whose region is enlarged below. **C.** Enlargement of the region around the third 14-mer in the above figure, showing a multiple alignment of the *S. lycopersicum* and *S. pennellii* genome sequences in this region, the primer generated by IGGPIPE, and the 14-mer itself. Alignments made with Geneious (Kearse et al. 2012).

being considered). When such a k-mer group is found, an LCR is called for the group and those common unique k-mers are removed from the pool under consideration. An alignment of part of an LCR and respective common unique kmers in the tomato SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0 genomes is shown in **Figure 2B and C**.

The LCR algorithm found 72,533 LCRs between the tomato SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0 genomes using parameter settings that included 1500 bp maximum k-mer spacing (DMAX) and 400 bp minimum LCR length (LMIN) (**Table 2**). The number of common unique k-mers per LCR ranged from 2 to 642. We tested whether the LCRs truly represented common conserved regions by making a dot plot between the two genomes using the LCRs as data (**Figure S-1**). The plot closely matches a similar dot plot made using data from a whole genome alignment of the same genomes using the progressiveMauve (Darling, Mau, and Perna 2010) whole genome aligner (**Figure S-2**), confirming that LCRs include conserved regions found in whole genome alignments.

Identification of Indel Groups

After LCRs are identified, the next step in the IGG marker pipeline is to examine each LCR's common unique k-mers to find pairs whose separation distance is unequal in the two (or more) genomes and satisfies user-specified parameters. We use the name *Indel Group* for the interval between such a k-mer pair. The name includes *group* because the interval must contain at least one indel but may have more than one. A single LCR can contain more than one Indel Group, each one bounded by a different pair of k-mers. The Indel Group algorithm found 249,635 overlapping Indel Groups within the LCRs between the tomato SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0 genomes using parameter settings that included amplicon size between 400 and 1500 bp and amplicon size difference between 50 bp (at 400 bp amplicon size) and 300 bp (at 1500 bp amplicon size). Counting only one Indel Group from each set of *overlapping* Indel Groups reduced that total to 31,621 non-overlapping Indel Groups between these genomes.

The number of indels within an Indel Group was confirmed to have a broad range (**Figure 3A**), and the length of the indels also spans a broad range (**Figure 3C**), though concentrated at smaller sizes. The number of indels of different sizes decreases approximately exponentially as the indel length increases (**Figure 4A, S-3**). Indels within Indel Groups can be found in all of the gene features and intergenic regions (**Figure 4C**). The density in coding regions is lowest, followed by intron, intergenic, 5'UTR, 3'UTR, and finally upstream and downstream with approximately equal density. We compared these Indel Group count and density results with those from a similar analysis between *Arabidopsis thaliana* accessions Col-0 and Ler-0, shown side-by-side with tomato in **Figures 3B, 3D, 4B, 4D**. Results are similar, although in Arabidopsis the densities ranked somewhat differently, with coding regions again lowest, then 5'UTR, intron, 3'UTR, intergenic, upstream, and downstream. Another difference between the species is that Ler-0 had a slower rate of decline in number of deletions of different sizes with increasing deletion length at indel sites, while Col-0 was similar to that seen in tomato (**Figure 4B**).

Primer Creation

After Indel Groups are identified, IGGPIPE extracts DNA sequence around the pair of common unique k-mers defining each Indel Group and executes Primer3 (Untergasser et al. 2012) to design primers at each of the k-mers, using as Primer3 input the concatenation of the two short DNA sequences, one surrounding each of the two k-mers, omitting the intervening region, which varies between genomes.

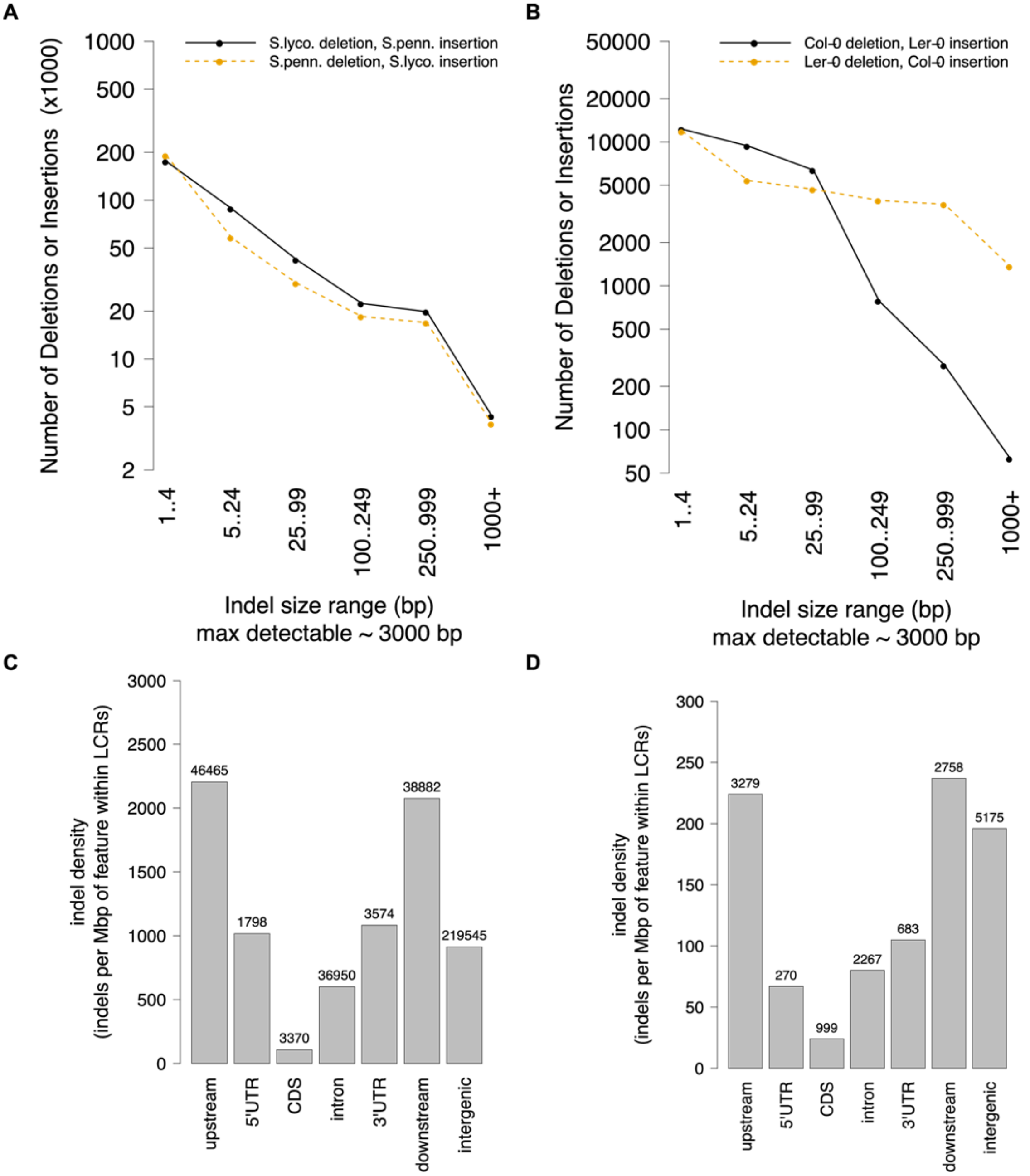

Figure 4. Additional characteristics of indels found within Indel Groups, from the same analysis cited in **Figure 3**. **A,C:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0; **B,D:** *A. thaliana* accessions Col-0/Ler-0. **C.** The number of indels of different sizes decreases approximately exponentially as the indel length increases. H: Heinz (*S. lycopersicum*), P: PENN (*S. pennellii*). **D.** Density of Indel Group indels within genomic features found in the LCRs containing the Indel Groups. Upstream is defined as within 1000 bp 5' of the 5'UTR, and downstream is within 1000 bp 3' of the 3'UTR of a gene, while intergenic is any position not falling into any of the other categories.

In silico PCR Testing

One of the final IGGPIPE steps is to run the *in silico* PCR program e-PCR (Schuler 1997) to test each primer pair, eliminating those not having the predicted amplicon sizes or

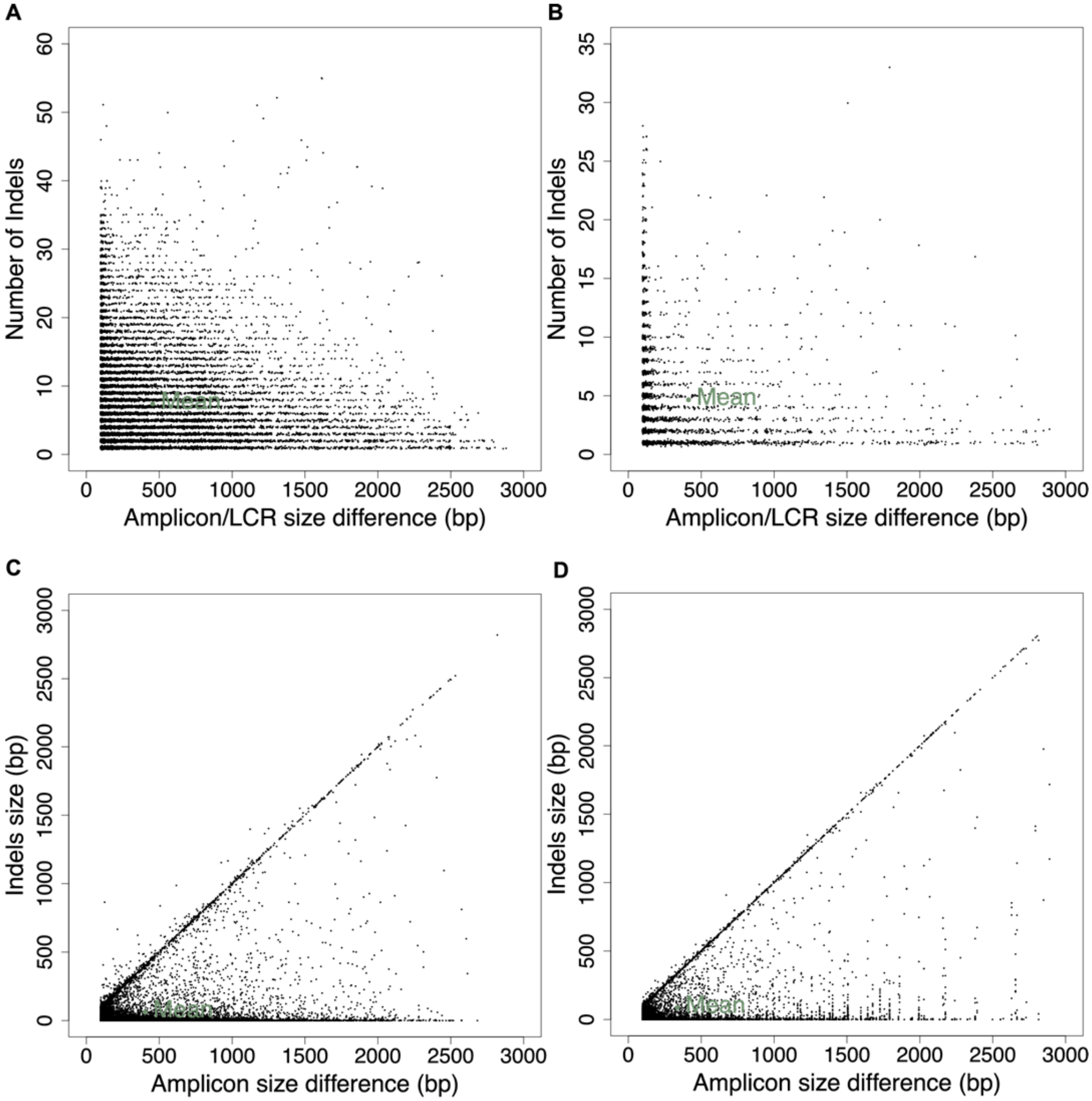

Figure 3. Characteristics of indels found within Indel Groups, from an IGGPIPE analysis of: **A,C:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0 (K=14, AMIN=100, AMAX=3000, ADMIN=ADMAX=100); **B,D:** *A. thaliana* accessions Col-0/Ler-0 (K=13, other parameters the same). **A, B.** Each Indel Group from was plotted as a point, where the x-axis is the predicted amplicon size difference and the y-axis is the number of indels found in the Indel Group after aligning the two sequences. **C,D.** Similar plot but y-axis is indel size. The 45º line is Indel Groups containing a single indel that is responsible for the amplicon size difference. Some points lie above the line because a single Indel Group can have deletions in both genomes, at different places.

amplifying at multiple loci. An alignment is shown in **Figure 2C** of a primer sequence, a common unique k-mer sequence, and k-mer and flanking DNA in the tomato and *S. pennellii* genomes.

IGGPIPE marker assessment testing: two genome polymorphism detection

We assessed the performance of IGG markers generated with IGGPIPE in a pairwise, two-genome fashion – first using the inter-crossable species tomato (*S. lycopersicum*) and *S. pennellii*, and second, a within-species evaluation using *Arabidopsis thaliana* accessions Col-0 and Ler-0. Computer resource usage metrics are provided in **Table 3** and **Tables S-1** and **S-2**.

***Assessment in* S. lycopersicum *and* S. pennellii**

We applied the IGG marker pipeline to the *S. lycopersicum* SL2.50/ITAG2.4 chromosome-based genome (Tomato Genome 2012; Bombarely et al. 2011) and the new *S. pennellii* (inter-crossable wild relative) V2.0 genome (Bolger et al. 2014). First, four different runs were performed with k varying from 12 to 15 and all other parameters remaining constant (**Table 3**). From these runs, a k-mer size of k=14 was chosen for further runs, using a balance between number of IGG markers generated and total computation time as selection criteria. Next, four more runs were performed, using different parameter settings for each, but all using k=14 (**Table 2**). The number of overlapping IGG markers generated ranged from 7,163 to 91,947 and the number of non-overlapping markers ranged from 2,332 to 16,442. In the fourth run (400-1500/50-300), the number of markers was largest at the minimum difference of 50 bp, decreasing in number up to the maximum possible difference of 1100 bp (**Figure 5A**). The marker density closely matches gene density (**Figure 5C**). Markers in *A. thaliana* accessions Col-0 and Ler-0 show similar distribution (**Figure 5B**) but a very different density across chromosomes (**Figure 5D**). A random selection of 24 IGG markers (two per chromosome) was tested molecularly and 21 (87.5%) were found to give a single

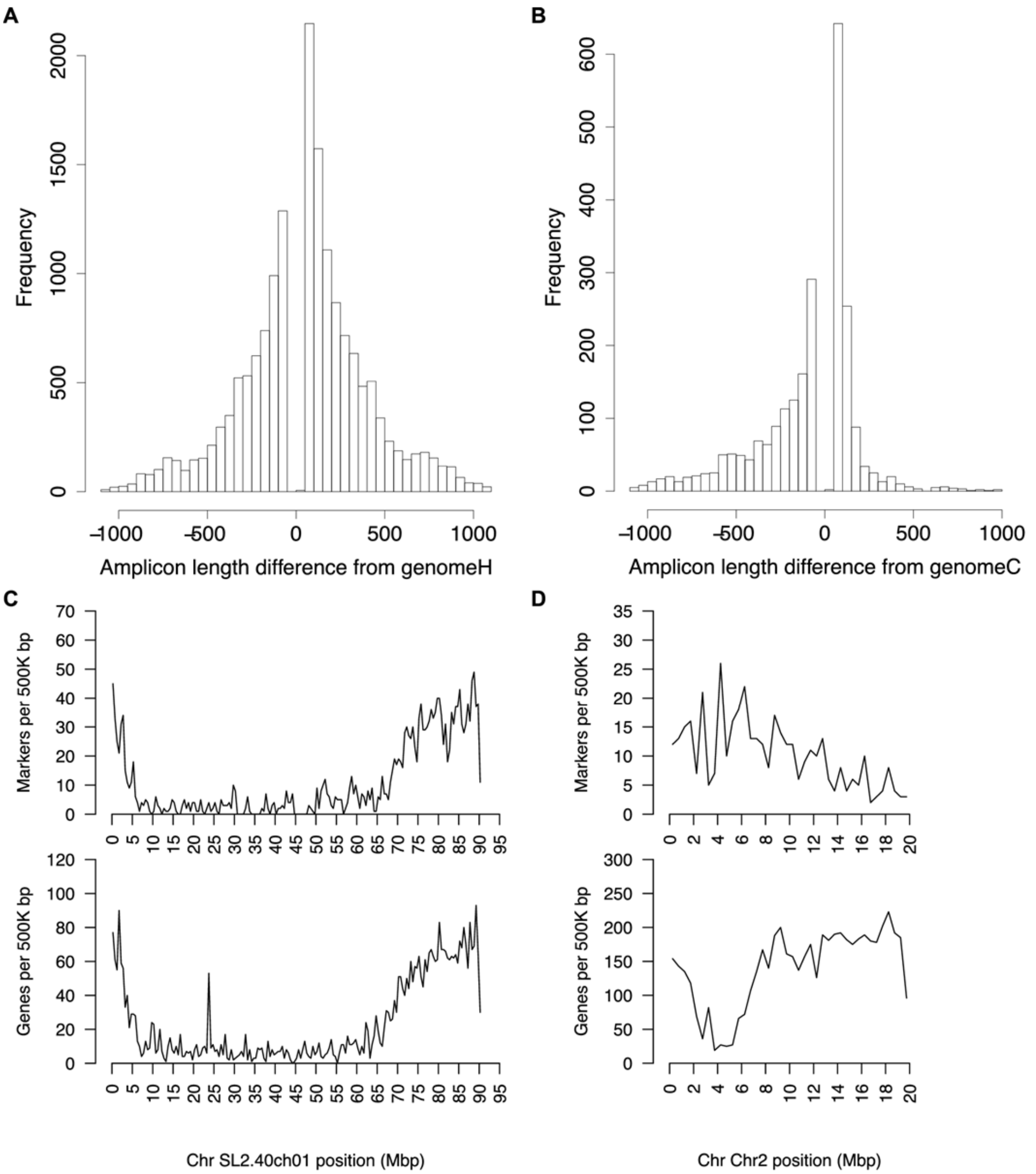

Figure 5. A, B. Distribution of differences in IGG marker amplicon sizes between the two analyzed genomes, from an IGGPIPE analysis of: **A:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0 (K=14, AMIN=400, AMAX=1500, ADMIN=50, ADMAX=300); **B:** *A. thaliana* accessions Col-0/Ler-0 (K=13, other parameters the same). A positive difference means the *S. lycopersicum or Col-0* amplicon is the larger, and negative means the *S. pennellii or Ler-0* amplicon is the larger. **C, D.** Density of IGG markers (top graph) and genes (bottom graph) along a representative chromosome, from the same analysis as above**.** **C:** Chromosome 1 of *S. lycopersicum* (tomato). Note positive correlation. **D:** Chromosome 2 of *A. thaliana* Col-0 accession. Note negative correlation.

amplicon of the predicted size in each of the two species (**Table 4, Figure 6**). Four IGG markers were used to successfully genotype 28 F2 individuals at four loci (**Figure S-3**). Markers cover all chromosomes, with greatest density in the less heterochromatic regions

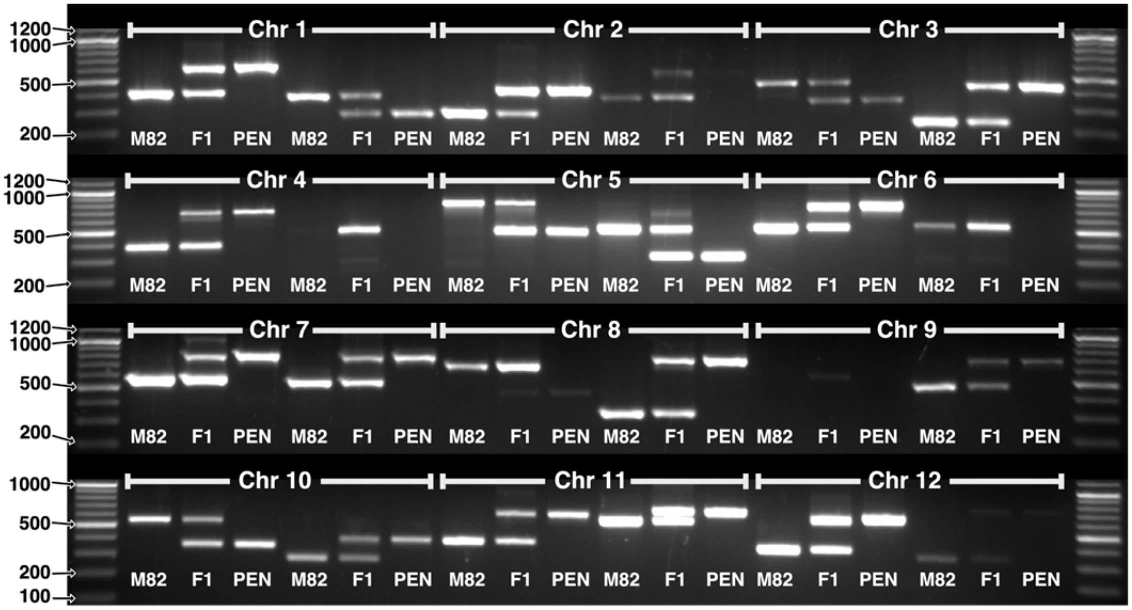
1200
1000
500
200
Chr 1
Chr 2
Chr 3
M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN
1200
1000
500
200
Chr 4
Chr 5
Chr 6
M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN
1200
1000
500
200
Chr 7
Chr 8
Chr 9
M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN
1000
500
200
100
Chr 10
Chr 11
Chr 12
M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN M82 F1 PEN

Figure 6. Twenty four IGG markers, two per chromosome at locations within the first or last 15% of each chromosome, were chosen randomly from three different IGGPIPE runs using different sets of parameters and all analyzing the *S. lycopersicum* (SL2.50/ITAG2.4 pseudomolecules) and *S. pennellii* (V2.0 pseudomolecules) genomes. In 21 of the 24 markers (87.5%) amplifying *S. lycopersicum* cv. M82, *S. pennellii* (PEN), and F1 DNA, two bands of the expected amplicon sizes are seen (**Table 4**), one in each species. In two cases, no band is seen in either species, and in another case, only an *S. lycopersicum* band is seen.

(**Figure S-4, S-5A, B**). The overlapping and non-overlapping IGG marker files from all four of these runs are provided (**Supplemental Data SD-IGGmarkers_HP14.zip**).

Assessment in **A. thaliana** *accessions Col-0 and Ler-0*

Length polymorphisms between Landsberg erecta and Col-0 accessions can be identified using the TAIR Search Polymorphisms/Alleles tool at arabidopsis.org (TAIR 2015; Lamesch et al. 2012). Unfortunately, many of these markers only allow identification of the presence of a PCR fragment in one accession versus its absence in the other. In those markers where there is a PCR fragment length polymorphism, the size difference is very small. **Table 5** shows the best available polymorphisms found (two per chromosome) for maximum product separation within several hundred markers. All markers have a very small (mean 37 bp) difference in size.

We applied the IGG marker pipeline to the *Arabidopsis thaliana* Col-0 accession TAIR10 genome (Lamesch et al. 2012) and the *A. thaliana* Ler-0 accession V0.7 genome (Gan et al. 2011). Parameter settings included k=13, amplicon size of 400 to1500 bp, and a minimum difference in size between amplicons of 50 to 300 bp. Relative to the inter-species marker run, we predicted that the number of polymorphisms between the two Arabidopsis accessions would be much smaller. However, this marker set contains 28,031 overlapping and 2,072 non-overlapping IGG markers all confirmed with e-PCR (**Table 6**, **Figure S-5C, D**). Ten of these markers were tested experimentally, and eight (80%) had the expected amplicon sizes in the two accessions (**Table 7, Figure 7A**). These markers had larger size differences and differences between the accessions were easier to distinguish compared to the TAIR Polymorphism/Search markers in **Table 5**. The entire marker set is provided (**Supplemental Data SD-IGGmarkers_CL13.zip**).

IGGPIPE marker assessment testing: three genome polymorphism detection

S. lycopersicum × S. sitiens ***introgression line development using IGG markers***

Cultivated tomato (*S. lycopersicum*) is an economically important crop, but genetic diversity for key agronomic traits needed for growth in a changing climate, such as abiotic stress tolerance, is lacking in the widely used inbred germplasm. Wild relatives such as *S. sitiens*, endemic to the Atacama Desert of Chile, are of interest because of adaptation to minimal rainfall, cold temperatures, and high soil salinity. Utilization of this genetic variation for breeding purposes can be facilitated by development of an introgression line (IL) population of *S. sitiens* in the background of cultivated tomato. No reference genome sequence is available for *S. sitiens*, and the majority of genomic markers available are SNPs (SolCAP Solanaceae Coordinated Agricultural Project'

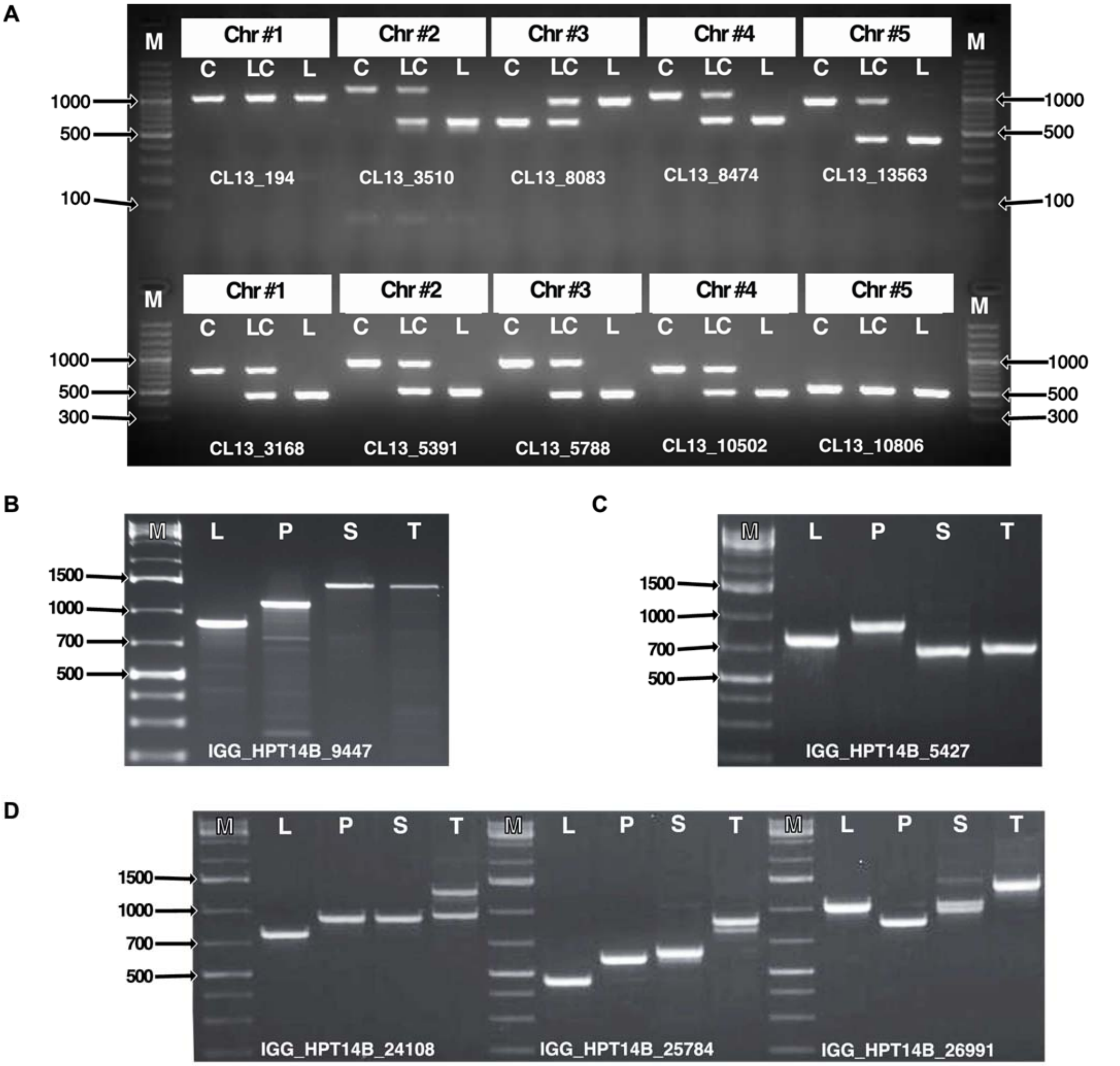

Figure 7. Gel electrophoresis of PCR products of several candidate IGG markers from two IGGPIPE runs. **A.** Testing primers generated against *Arabidopsis thaliana* accessions Landsberg and Columbia. PCR product resolved on 2% gel. **M:** BioLabs QuickLoad 100 bp Ladder; **C:** Columbia-0; **LC:** Landsgerg-Columbia hybrid; **L:** Landsberg-0. Eight of 10 show expected product sizes (**Table 7**). **B-D.** PCR products by gel electrophoresis using IGG markers from triallelic marker run with *S. lycopersicum, S. pennellii,* and *S. tuberosum* genomes. **M:** O'GeneRuler 1Kb Plus Ladder; **L:** *S. lycopersicum*; **P:** *S. pennellii*; **S:** *S. sitiens*; and **T:** *S. tuberosum*. **B**. IGG marker #B_9447 shows three-way polymorphism between the three genomes of interest and amplicons are of predicted size (**Table 9**). In addition, *S. tuberosum* and *S. sitiens* share the same allele. **C.** Marker #B_5427 also shows three-way polymorphism between the three genomes of interest. In this case, the S. tuberosum amplicon is closer to 700 bp than the predicted 527 bp. *S. lycopersicum* and *S. pennellii* have predicted amplicon sizes. In addition, *S. tuberosum* and *S. sitiens* have a very small or zero size difference. **D.** Markers #B_24108, B_25784, and B_26991 also indicate three-way polymorphism between *S. lycopersicum*, *S. pennellii*, and *S. tuberosum.* However, *S. sitiens* shares an allele with either *S. pennellii* (B_24108) or *S. lycopersicum* (B_26991). Presence of multiple bands is observed for select genotypes.

2015; Sim et al. 2012). Here we describe how we utilized the IGG pipeline to identify polymorphisms between three genomes which can be useful in cases where populations are developed between species with varying levels of self-incompatibility or where one parent’s genome is unsequenced but a closely related sequenced genome exists.

Pre-and post-zygotic hybrid incompatibility between cultivated tomato and *S. sitiens* has made introgression line development a challenge (Pertuze, Ji, and Chetelat 2002; Peters et al. 2012; DeVerna et al. 1990; Pertuze, Ji, and Chetelat 2003). To aid in the production of *S. lycopersicum* and *S. sitiens* hybrids, an interspecific bridging line, F1 *S. lycopersicum* × *S. pennellii,* was employed. Hybrids of [*S. lycopersicum* × *S. sitiens] X [S. lycopersicum* × *S. pennellii]* were backcrossed (BC) to cultivated tomato. While the majority of the *S. sitiens* genome was transferred as determined using Cleaved Amplified Polymorphic Sequence (CAPs) markers, repeated backcrossing was needed to eliminate residual background noise and to retain individual introgressed segments. We ran IGGPIPE with three genomes (tomato, *S. pennellii*, and potato) to develop tri-allelic markers for genotyping these crosses. The *S. tuberosum* (potato) sequence was used as a stand-in for the unsequenced *S. sitiens* genome, as the two species share the same chromosomal configuration (Peters et al. 2012; Pertuze, Ji, and Chetelat 2002).

The three-genome IGGPIPE analysis used the *S. lycopersicum* SL2.50/ITAG2.4 (Heinz) genome (Tomato Genome 2012; Bombarely et al. 2011), the *S. pennellii* V1 genome (Bolger et al. 2014), and the *S. tuberosum* (potato) Phureja group clone DM1-3 V4.03 genome (Potato Genome Sequencing et al. 2011) with parameter settings including k=14, an amplicon size of 400 to 1500 bp, and a minimum difference in size between amplicons of 50 to 300 bp. A total of 951 overlapping (278 non-overlapping) IGG markers were generated that were predicted to display three-way polymorphism between *S. lycopersicum* (tomato), *S. pennellii*, and *S. tuberosum* (potato) (**Table 8, Figure S-6**). Of these, 32 markers were selected for further characterization and tested on DNA from the parents of our introgression line population (*S. lycopersicum*, *S. pennellii*, and *S. sitiens*) and *S. tuberosum*. We found that all 32 amplified in cultivated tomato and *S. pennellii,* 30 in *S. tuberosum*, and 28 in *S. sitiens,* with an annealing temperature of 55ºC (success rate of ~88-94%) (**Table 9**). The genetic difference between potato and *S. sitiens* could explain this result (Peters et al. 2012; Pertuze, Ji, and Chetelat 2002). We found that 30 (93.8%) of these 32 markers were triallelic relative to potato, displaying scorable band differences between cultivated tomato, *S. pennellii*, and *S. tuberosum* (**Table 9**). Some non-specific amplification was observed for all species tested. For example, of the 28 amplicons from cultivated tomato, four primer pairs yielded two or more bands (**Figure 7B**). However, the intensity of these products was considerably lower and overall did not affect parent identification.

Potato is a good predictor of the presence of indels in **S. sitiens***, but not of product size*

To determine whether novel *S. sitiens* markers could be used in IL characterization, the 28 *S. sitiens* primer pairs were scored by whether they displayed two-way polymorphism (*S. pennellii* and *S. sitiens* shared allele) or three-way polymorphism (no shared alleles between tomato vs. *S. pennellii* vs. *S. sitiens*) and 20 out of 28 (71.4%) appeared polymorphic between the three parents (for example, IGG15, **Figure 7C, Table 9**). Next we wanted to test whether *S. sitiens* and *S. tuberosum* shared the same allele sizes. We found that only seven markers had a shared allele size with *S. sitiens* (for example IGG15, **Figure 7C, Table 9** and IGG25784, **Figure 7D**).

Taken together, these results indicate that it is possible to identify polymorphisms in *S. sitiens* using potato as a genome reference. Having even a rough *de novo S. sitiens* genome assembly would likely improve marker success. While our observed failure rate is not ideal for marker design, and far below that observed between *S. lycopersicum* and *S. pennellii*, it is quite close to the failure rate of other marker design studies such as those observed for Single Copy Orthologous Genes (Wu et al. 2006) and IGG markers are substantially easier to use than existing CAPs markers. Two sets of IGG markers used in this project are provided (**Supplemental Data SD-IGGmarkers_HPT14.zip**).

Comparative assessment of IGGPIPE against other marker software

We compared features and performance of IGGPIPE to two other marker creation tools, IMDP (Lu et al. 2015) and PolyMarker (Ramirez-Gonzalez, Uauy, and Caccamo 2015), that also process whole genome data *in silico* (**Table 10**). Markers are *discovered* by IGGPIPE and IMDP, whereas PolyMarker requires SNPs as input and generates primers. IGGPIPE differs from these algorithms as it can generate IGG markers having much larger amplicon sizes and size differences, allowing use of a 1% agarose gel assay instead of higher percentage gels or polyacrylamide gels. IGGPIPE generated an order of magnitude more markers with primers in a pair of test species that included tomato than IMDP did using rice cultivars as a test species, while PolyMarker's SNP marker primers using polyploid wheat as a test species were similar in number to IGGPIPE but no PCR testing was done and the majority generated amplicons with size differences of a few bp or less. IGGPIPE has a distinct advantage relative to these tools as it allows the user to generate multi-allelic markers enabling differentiation between two or more genomes.

IGGPIPE is operated from the command line with manually edited user configuration files, whereas IMDP uses a user-friendly graphical tool LONI (Dinov et al. 2009) and PolyMarker has a user-friendly web interface. IGGPIPE includes a set of IGG markers in *tomato/S. pennellii* and *A. thaliana* Col-0/Ler-0 and IMDP included a rice marker public web-based database (RIMD-Rice Indel Marker Database) as part of its release. IGGPIPE includes a utility for annotating markers with information from other genome sources that overlap the markers, and can generate files suitable for custom genome browser tracks.

In silico marker identification using IGGPIPE

IGGPIPE is available as an open-source command line pipeline run via a Mac OSX, Linux-compatible, or Windows/Cygwin terminal interface. It is run from the command line using a 'make' utility, and includes detailed installation and run instructions. The only input required to run the pipeline is a FASTA file for each genome to be analyzed. IGGPIPE is available in open source form in the BradyLab/IGGPIPE GitHub(GitHub 2015) repository at https://github.com/BradyLab/IGGPIPE.

Discussion

IGG markers are similar to common indel markers in that they use a pair of PCR primers binding to regions flanking single-copy sites, but differing in that the amplified region may be larger and may encompass multiple indels whose lengths may range up to 1500 or more bp. Testing pairs of known unique primers for amplicon size differences is done within other indel marker programs (Lu et al. 2015; Liu et al. 2015; Zhou et al. 2015), but is normally limited to very short distances and single indel spans. The actual limits on amplicon sizes and size differences are parameters specified when IGGPIPE is run to generate the markers, providing user flexibility while also allowing length limits like those of traditional indel markers to be obtained when desired. IGG markers are of interest because they are built around an abundant source of polymorphism (indels ranging from a few base pairs to several hundred in size), can be scored easily, and have potential for multi-allelism. The number of markers found using IGGPIPE depends not only on the degree of polymorphism between the genomes, but also on the setting of search parameters, which include minimum and maximum amplicon sizes and minimum difference in their sizes between genotypes. Settings can be optimized to speed post-PCR gel electrophoresis by permitting use of rapidly prepared 1% agarose gels with easily scoreable large amplicon size differences. The IGGPIPE algorithm is flexible enough to make use of whole genome sequence data in either fully assembled chromosome form, or partially assembled scaffold form, as markers have been generated and tested using both reliable scaffolds and fully assembled genomes. Assemblies with substantial redundancy may not be good data sources for IGGPIPE, as they will result in an absence of unique k-mers in the redundant region and therefore fewer IGG markers, but this is advantageous in that it produces a low marker false positive rate. Furthermore, use of assembled genomes with substantial redundancy in addition to scaffold misassembly will have greater false positive and false negative rates than assemblies with substantial redundancy alone. Some pipeline steps, such as e-PCR, can be extremely slow when there are hundreds of thousands of scaffolds, so it is recommended that very short scaffolds that are unlikely to contribute markers be removed.

One distinct advantage of the IGGPIPE algorithm is that it is sufficiently flexible to identify multi-allelic markers, allowing the differentiation of more than two genomes. A parameter (NDAMIN) specifies the number of distinct amplicon sizes that must be present among the genomes being analyzed in order for a marker to be valid. In the 3-way test using tomato, *S. pennellii*, and potato, we used a value of two for this parameter, and the non-overlapping markers included 239 triallelic and 5166 biallelic markers. The pipeline has not been tested with more than three genomes, although it is written with no hard limit. A possible use would be to run a several dozen related genomes, for instance with several related land races of a particular species, with NDAMIN set to five. This would generate markers for loci having at least five distinct alleles among the genomes. A series of such markers could be used as fingerprinting markers. Future IGGPIPE enhancements could include population genomics features such as assessment of information content at multi-allelic loci, which assists in choosing the best markers for studies such as assessment of population-wide variation. A number of usage cases are illustrated in Table 11.

The method can also be used with polyploid species. The additional redundancy in the genomes means that the value used for k may need to be increased so that a sufficient number of unique k-mers is found. Another twist on polyploid analysis is to separate the subgenomes and run them through IGGPIPE as if they are separate genomes. Resulting markers may be used to distinguish between homeologous chromosomes. Another polyploid technique enables one to find IGG markers where a single primer pair produces one amplicon of unique size for a target chromosome region in one subgenome, and a second amplicon of unique size from one chromosome of any of the other subgenomes, permitting a single primer pair PCR to test for presence of a target region while using the second amplicon as a positive control. Alternatively, multiple genomic locations could be screened simultaneously, effectively allowing a single primer set to behave as a multiplex PCR. If detailed indel information is available for a diploid genome, it can be applied to construct a second genome containing the indels which, when used as IGGPIPE input, would produce markers for genotyping loci. Finally, IGGPIPE could be used to compare and generate "cDNA" IGG markers, using sequenced and assembled cDNA libraries of two related genotypes. Such markers could be used to amplify regions from cDNA libraries.

A strong positive correlation between IGG marker density and gene density is visible in marker/gene plots for tomato (**Figure 5C**) where the Pearson correlation of marker density and gene density measured in 5 Mbp windows was 0.83. In contrast, in the *A. thaliana* Col-0 genome a negative correlation of -0.34 is observed, and for all chromosomes (except perhaps chromosome 1), the marker density is highest in the heterochromatic regions where gene density is lowest (**Figure 5D**). The *A. thaliana* analysis was between accessions, whose intergenic regions retain enough sequence identity that LCRs are found within most of the region, and the rapid evolution of polymorphisms in the heterochromatic region likely leads to a high indel density (**Figure 4D, S-7**). We hypothesize that between species, such as tomato and *S. pennellii*, intergenic regions have had sufficient time to thoroughly diverge from one another, and LCRs can no longer be found throughout a majority of the region, leading to an overall low indel density in these regions (**Figure 4C, S-7**). Nevertheless, enough LCRs are found in intergenic regions of tomato to cover about 40% of the region, and within those locally conserved regions, indel density is on a par with UTR indel density (**Figure 4C, S-7)**

IGGPIPE includes a code module alignAndGetIndelsSNPs that extracts DNA sequence around markers, Indel Groups, or LCRs, aligns it, and examines it for indels and SNPs. SNPs, for example, are a rich source of polymorphisms that are often used in GWAS studies. The LCR file from the IGGPIPE comparison of the *S. lycopersicum* SL2.40/ITAG2.3 and *S. pennellii* V2.0 genomes produced 391,968 putative indels and 2.41 million putative SNPs with parameters that included amplicon size range 100-3000 bp and amplicon difference of 100 bp.

The LCRs, Indel Groups, and IGG markers themselves are also of use as they are essentially a form of whole genome alignment. The good match between a dot plot of LCRs produced by the IGGPIPE module dotPlot (**Figure S-1**) and a dot plot of locally colinear blocks produced by progressiveMAUVE(Darling, Mau, and Perna 2010) (**Figure S-2**) illustrates the accuracy of the IGGPIPE alignment. The data might therefore be useful for other purposes, such as mapping features between the genomes that were analyzed, including translocations or inversions, or even local duplications.

Finally, in the future, the uses of unique k-mers could be extended. Unique k-mers in one genome that do not occur in another genome (*genome-unique* k-mers) could be used as primer design sites to make allele-specific markers, amplifying only when one particular allele is present. Both *genotype-unique* and *common-unique* k-mers can be used together to make an alternative type of allele-specific marker that includes a second PCR diagnostic band. The method would be to design three primers, the first at a genome-unique k-mer near the target site in the target genome, the second at a common-unique k-mer near the first, and the third at a nearby genome-unique k-mer in the non-target genome, and then run a three-primer PCR reaction. Combined *genotype-unique* and *common-unique* k-mers could also be employed as an alternative way of measuring gene expression in an RNA-seq experiment, by identifying genes containing these k-mers and counting the number of reads containing k-mers found in each gene. This method might prove faster than mapping reads to a reference, and could be just as accurate. Finally, k-mers could be used to search for duplicated regions by looking for clustering of k-mers that all occur the same number of times in a genome, and primers designed around these k-mers would amplify the duplicated regions.

Conclusions

Common unique k-mers can be used to effectively identify large numbers of groups of one or more adjacent indel polymorphisms in two or more species or populations, flanked by conserved regions where IGG marker primers can be designed to amplify the intervening region, which can be selected for a preferred size range and preferred minimum size difference between species. The method can be extended to use genome-unique k-mers to create allele-specific markers and to create markers that can amplify regions present in specific copy numbers. The method for choosing an Indel Group spanning multiple indels in order to achieve flexibility in amplicon sizes can be extended to any *in silico* indel marker algorithm as long as contig boundaries are honored. Sets of k-mers present in each genome in varying copy numbers may be useful in whole genome alignment or copy number analysis.

Materials and Methods

IGGPIPE pipeline

The IGGPIPE pipeline (**Figure S-8**) uses existing bioinformatic tools as much as possible: Jellyfish (Marcais and Kingsford 2011) for extracting single copy (unique) k-mers from genomes, Primer3 (Untergasser et al. 2012) for designing primers, e-PCR (Schuler 1997) for *in silico* testing of final IGG marker primers, and MUSCLE (Edgar 2004) for aligning DNA sequences to find indels and SNPs. For those portions of the pipeline requiring custom software, three different programming languages were used: C++, Perl (Wall 1987-2012), and R (R Core Team 2014). Details on the custom software of the IGGPIPE pipeline can be found in the **Supplemental Materials and Methods**.

IGGPIPE was developed on a Mac OSX operating system, but it has also been tested on Linux and Windows systems. A C++ compiler (e.g. Apple XCode ('Mac App Store - Xcode' 2015), included with OSX) is required to compile C++ code.

Choosing k-mer size

The value of k is a user-defined parameter in IGGPIPE and must be chosen carefully. The larger the value, the more common unique k-mers will be found, up to a point, beyond which the number will saturate because unique k-mers will begin to be long enough to no longer be in common with the other genome (**Figure 1C**). The number of IGG markers generated by the pipeline will also tend to rise as the number of common unique k-mers increases, because the k-mers are candidate anchors for IGG marker primers. A user manual (**Supplemental Data SD-RUN.html**) included with the IGGPIPE software provides guidance for assessing different values of k when analyzing a set of genomes, using total computational time, number of common unique k-mers, and number of IGG markers generated as criteria for comparing values.

findLCRs algorithm

The list of common unique k-mers was annotated with genome position (chromosome or scaffold) and contig identifier within each genome, and subsequently processed all k-mers in the same contig as a group when searching for locally conserved regions. *Contig* in this case is defined as a continuous sequence of ATCG nucleotides bounded on either side by the end of the sequence or by an "N" unknown nucleotide designator.

Knowledge of contigs is important to avoid creating an IGG marker whose two ends are in two different contigs on opposite sides of a sequence of N's. The N designator implies that the region containing the N's was not sequenced, and the number of N's is not a reliable indicator of the actual sequence length.

Translocations of DNA segments can cause a given contig in one genome to pair up with more than one contig in the other genome (**Figure 2A**, LCRs a, b, c, and e vs. d). Random k-mers that pair with a different contig may occur within an LCR sequence of several k-mers pairing two contigs (f). That single interrupting k-mer should not cause the LCR to be split into two separate LCRs, which might remove an opportunity to use a length polymorphism for a marker. A translocation could even create interruptions of long pairings of two contigs with short pairings with alternate contigs (g). The two pairings in that case should be evaluated independently to see if they qualify as an LCR, while not splitting the larger pairing into two separate LCRs. The LCR algorithm should be tolerant of these and other possibilities introduced by translocations.

Our algorithm for LCR identification (**Figure S-9**) tolerates translocations by temporarily ignoring incompatible k-mers, setting them aside when they are identified while confirming an LCR, then using them again as candidates for the next LCR. The LCR parameter constraints, which are set by the user, include minimum number of k-mers per LCR (KMIN), minimum LCR length (LMIN), and maximum k-mer spacing within a single LCR (DMAX). If the value of KMIN is too small, LCRs may be called which are random occurrences of common unique k-mers close together on the same contigs. This is not the problem it might seem, as markers produced from the miscalled LCRs will be rejected later during the *in silico* PCR phase, and setting KMIN to two is usually adequate. However, if the IGGPIPE indel finder utility is used on the LCR data, it may call spurious indels within the miscalled LCRs, so if accurate indel calls are desired, a KMIN value of four would be better.

The empirical results we obtained with the genomes we worked with lead us to advise setting the minimum LCR length, LMIN, to the minimum desired amplicon size, and the maximum spacing between two adjacent k-mers in an LCR, DMAX, to the maximum desired amplicon size.

A more detailed discussion of the findLCRs algorithm is presented in the **Supplemental Materials and Methods.**

Identification of Indel Groups

The algorithm for identifying Indel Groups tests all possible pairs of k-mers within an LCR (all pairs of blue vertical lines in the LCR of **Figure 2B**) to find all that satisfy the parameter constraints, including Indel Groups that overlap one another. The most important parameter constraints are minimum and maximum amplicon size (AMIN and AMAX) and minimum amplicon size difference (ADMIN). ADMIN and ADMAX define the minimum acceptable amplicon size difference as a function of the amplicon size. The *minimum* amplicon size difference for *smallest* amplicons is ADMIN and the *minimum* size difference for *largest* amplicons is ADMAX. When the smallest amplicon size is AMIN, the next larger one must be at least AMIN+ADMIN, and when the largest amplicon size is AMAX, the next smaller one must be no more than AMAX-ADMAX, and for amplicons with sizes in between the limits, it scales linearly from ADMIN to ADMAX. It is this simple testing of k-mer pairs, more than the details of identifying LCRs, that is at the core of allowing an IGG marker to flexibly acquire the amplicon size characteristics desired by the user. Any indel marker produced using such an indel grouping algorithm can be called IGG markers.

Quantifying indel density within Indel Groups

The number, size, and position of indels within Indel Groups was examined in an IGGPIPE comparison of *S. lycopersicum* SL2.50/ITAG2.4 and *S. pennellii* V2.0 genomes. After running IGGPIPE with parameters that attempted to find as many LCRs as possible (k=14, AMIN=100, AMAX=3000 bp, and ADMIN=ADMAX=100 bp), the sequences for each non-overlapping Indel Group were extracted from the two genomes, aligned, and indels counted, with position noted for each one relative to a gene CDS, 5'UTR, 3'UTR, introns, 1000 bp upstream or downstream of UTRs, and intergenic regions. Density was computed by dividing the number of indels within a type of gene region by the total length of those regions within the LCRs from which the Indel Groups were extracted.

Primer creation

After extracting DNA sequence around the k-mer pair for an Indel Group, wherever base pair positions flanking the k-mers don't match in all genomes, the base pairs are replaced by the nucleotide designator "N", which forces Primer3 to disallow primer overlap at that position. The primers always include at least some bases of the common unique k-mers, extending beyond them by no more than a limited amount which itself is a parameter called EXTENSION_LEN.

If EXTENSION_LEN is set to the approximate primer length minus the k-mer length k, then each primer will include most or all bases of the k-mer. The advantage of including the k-mer in the primer is that it is already known to be unique in the genome. However, even if primers are designed off to one side of the k-mer and the region happens to occur multiple times in the genome, the next IGGPIPE step, *in silico* PCR testing, will catch and reject bad primer pairs that amplify multiple amplicons. The Primer3 parameter file can optionally be modified by the user to specify user-preferred primer design parameters.

Sub-pipeline for finding indels and SNPs

An IGGPIPE sub-pipeline, invoked using a different argument on the "make" command line, reads a file of LCRs, Indel Groups, or IGG markers, extracts DNA sequence from each genome around each element, aligns them, and examines the alignments for indels and SNPs, writing them to a file (**Figure S-10**). The aligner currently used is MUSCLE (Edgar 2004) because of its high speed and satisfactory alignments, but the code can easily be changed to use a different aligner. Parameters MAX_INDELS_PER_KBP and MAX_SNPS_PER_KBP are used to detect poor alignments or alignments of unalignable regions. If the number of SNPs in an alignment is more than that fraction of the total sequence length in any genome, the alignment is ignored.

Marker file output

After *in silico* PCR testing, the final sets of markers are written to two files, one containing those whose amplicon regions may overlap, and a second with only non-overlapping markers. A parameter (OVERLAP_REMOVAL) selects whether the marker with the shortest or longest amplicon should be retained among a group of two or more markers that overlap.

Plotting utilities

Several plotting utilities are provided with the IGGPIPE pipeline, which plot marker number (**Figure S-11, S-12**) and density per chromosome (**Figure S-4, S-5, S-6**), indel size distribution and density (**Figure 3, 4, S-13**), and a dot plot of LCR positions in each genome (**Figure S-1**).

Position-based file merge utility

An additional useful utility in IGGPIPE is annotateFile.R, which is able to read any text-based data file containing columnar data that includes sequence position information. This module searches such a file, A, for data rows whose position intersects positions within rows of another such file, B, and outputs a new file A' containing new columns with data from the rows of B that intersect each row of A. This can be used for many purposes. We have used it to add a column to marker files containing the M82 x PENN introgression lines (Eshed et al. 1992) whose introgressions contain each marker, and the marker's approximate location within the introgression. Another use is to read the position information from a gene model .GFF file to annotate marker files with a column giving the nearest gene or gene feature. We used this technique to annotate an indel output file from the indel finder program with intron and exon information which was then used to assess indel frequency in genomic areas (**Figure 4A, B, Figure S-13**). The same module can also generate .GFF files from other data file types (such as marker files, which are in tab-separated format), and this can be used to add a new track to a genome browser that displays the markers in their appropriate genomic position (**Figure S-14**).

Plant material

The tomato plant material was provided by the Tomato Genetic Resources Center (TGRC) and was composed of the parental genotypes of the introgression line population: *S. lycopersicum* (NC84173), *S. pennellii* (LA716), and *S. sitiens* (LA716). *S. tuberosum* (cultivated potato) DNA was use as a marker control. DNA was isolated in a 1.5 mL Eppendorf tube from a single, 3-week old leaflet, following a method outlined in (Li, Royer, and Chetelat 2010).

Testing IGG markers for tomato/*S. pennellii*/potato

A set of 32 markers were picked at random from the list of 857 IGG primer pairs that were predicted to be polymorphic between tomato, *S. pennellii*, and potato by the IGGPIPE pipeline without e-PCR verification. Fragments were amplified in a 20µL PCR reaction using AmpliTaq (Life Technologies, Carlsbad, California), following manufacturer's recommended procedure with 2 µL (100ng) template DNA. The thermal cycling conditions were as follows: denaturation for 2 minutes at 94ºC, followed by 35 cycles of 94ºC for 30 s, 55ºC for 30 s, and 72ºC for 2 minutes, with a final extension of 72ºC for 5 minutes. PCR reactions were run on a 400 mL 2% agarose gel (containing 15 µL of a 10 mg/mL ethidium bromide stock) at 160V for 60 minutes. All 32 markers amplified with these conditions in at least two of the four parental species. The image was annotated using Affinity Designer software (Serif Europe 2015).

Testing IGG markers for* A. thaliana *accessions Col-0/Ler-0

Genomic DNA from *Arabidopsis thaliana* accessions Columbia, Lanbsberg Erecta, and the hybrids were extracted individually by CTAB method (Doyle 1987). The final DNA pellet was dissolved in 100ul of sterile double distilled water. One ul of the genomic DNA was used as a template in the PCR reactions. PCR master mix was made with TAKARA EX-TAQ DNA polymerase (Cat. #RR001A, Clontech Laboratories, Inc.). The PCR was programmed as: 37 cycles of 10 sec at 98° C (denaturation), 30 sec at 55° C (annealing) and 1 min at 72° C (extension) followed by final extension for 5 min at 72°C. The PCR products were resolved in 2.5% agarose gel and imaged using an AlphaImager gel documentation system. The image was annotated using Affinity Designer software (Serif Europe 2015).

Supplemental Material

The following supplemental materials are available.

- **Supplemental Figures:**
  - **Supplemental Figure S-1.** Dot plot produced with the dotplot.R utility, showing *S. lycopersicum* SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0 genomes.
  - **Supplemental Figure S-2.** Dot plot from a progressive MAUVE whole genome alignment of *S. lycopersicum* SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0 genomes.
  - **Supplemental Figure S-3.** Gel electrophoresis of PCR products of several candidate IGG markers identified by IGGPIPE using genomes *S. lycopersicum* cv. M82 and *S. pennellii.*
  - **Supplemental Figure S-4.** Density of non-overlapping candidate IGG markers on chromosome 1 of both *S. lycopersicum* cv. M82 and *S. pennellii*.
  - **Supplemental Figure S-5.** Density of candidate IGG markers found in two different runs of the IGGPIPE pipeline with two different genome sets.
  - **Supplemental Figure S-6.** Density of candidate 2-way and 3-way IGG markers found using IGGPIPE pipeline to analyze three genomes together.
  - **Supplemental Figure S-7.** Fraction of different genomic regions covered by LCRs, and density of indels within those LCRs, for analysis of both *S. lycopersicum* / *S. pennellii* and *A. thaliana* Col-0 / Ler-0.
  - **Supplemental Figure S-8.** IGGPIPE pipeline elements flowchart.
  - **Supplemental Figure S-9.** Details of algorithm used by the R code module findLCRs to use common unique k-mers to call LCRs.
  - **Supplemental Figure S-10.** IGGPIPE pipeline optional elements for aligning Indel Groups and locating the actual indels of each one.
  - **Supplemental Figure S-11.** Number of candidate IGG markers per million base-pairs found in *S. lycopersicum* cv. M82 and *S. pennellii* chromosomes.
  - **Supplemental Figure S-12.** Number of candidate IGG markers per million base-pairs found in *A. thaliana* Col-0 and Ler-0 chromosomes.
  - **Supplemental Figure S-13.** The number of indels of different sizes, overlapping or upstream or downstream of genes, within the Indel Groups resulting from IGGPIPE analyses of genomes *S. lycopersicum* SL2.50/ITAG2.4 and *S. pennellii* V2.0, and genomes *A. thaliana* Col-0 and Ler-0, is shown.
  - **Supplemental Figure S-14.** A custom browser track was added to the SolGenomics.net JBrowse browser using a GFF3 file produced by an IGGPIPE utility from an overlapping IGG markers file generated from *S. lycopersicum* and *S. pennellii* genomes.

- **Supplemental Tables**
  - **Supplemental Table S-1.** Computer resource usage of each IGGPIPE module is shown for the analysis of two genome pairs: Tomato/*S. pennellii* and *A. thaliana* Col-0/Ler-0 using the same parameter settings as shown in main text Table 3 for K=14 and K=13 respectively. The findLCRs module uses a bigger fraction of the total CPU time than any other module. Memory usage is minor for these genomes and parameters, and total CPU time is about 2 hours in the first case and half an hour in the second case. This table groups modules into three sets and subtotals each set. All statistics were gathered with the BSD time utility running on a system with an Intel 2.4 GHz Core 2 Duo CPU, 16 Gb DRAM, and Mac OSX 10.11.4.
  - **Supplemental Table S-2.** Computer resource usage of each IGGPIPE module subset from **Supplemental Table S-1**, for a baseline parameter set (first entry) and seven other cases where a single parameter is changed from the baseline set. Increases in maximum amplicon length (AMAX/DMAX) and reductions in minimum amplicon sizes difference (ADMAX) have dramatic effects on CPU time, which is expected because this increases the number of potential good markers that must be generated and tested. All cases are for the analysis of Tomato/*S. pennellii* genomes. Statistics program and system are as in **Supplemental Table S-1**.
- **Supplemental Materials and Methods.** Additional detail on IGGPIPE algorithms. (Same file as supplemental figures).
- **Supplemental Data SD-INSTALL.html.** The IGGPIPE installation manual, also part of the IGGPIPE GitHub repository.
- **Supplemental Data SD-RUN.html.** The IGGPIPE user manual, also part of the IGGPIPE GitHub repository.
- **Supplemental Data SD-IGGmarkers_HP14.zip.** A zip file containing files of IGG markers and associated data for the genomes *S. lycopersicum* SL2.50/ITAG2.4 (Heinz) and *S. pennellii* V2.0.
- **Supplemental Data SD-IGGmarkers_CL13.zip.** A zip file containing files of IGG markers and associated data for the genome *A. thaliana* accessions Col-0 TAIR10 and Ler-0.
- **Supplemental Data SD-IGGmarkers_HPT14.zip.** A zip file containing files of IGG markers and associated data for the genomes *S. lycopersicum* SL2.50/ITAG2.4 (Heinz), *S. pennellii* V2.0, and *S. tuberosum* Phureja group clone DM1-3 V4.03.
- **IGGPIPE** is available in open source form in the BradyLab/IGGPIPE GitHub(GitHub 2015) repository at https://github.com/BradyLab/IGGPIPE.

Acknowledgements

We thank Tony Bolger, Bjoern Usadel, and Alisdair Fernie for sharing *S. pennellii* genome scaffolds before publication. SMB and TWT were funded by NSF-IOS-18085. Work on *S. sitiens* marker development was supported in part by the USDA Agriculture and Research Initiative grant no. 2013-67013-21131. Additional thanks to Luca Comai and Julin Maloof who provided critical discussion.

Tables

year	acronym[a]	name[b]	polymorphism[c]	visualization technique	codom[d]	loci visualized[e]	same as[f]	*reference[g]
1980	RFLP	Restriction Fragment Length Polymorphism	variable length of restriction digest fragments	Southern hybridization to random probe	mostly	one		(Botstein et al. 1980)
1987	VNTR	Variable Number Tandem Repeat	variable numbers of tandem repeats of short sequences	Southern hybridization to custom probe	yes	variable		(Nakamura et al. 1987)
1989	SSCP	Single Strand Conform-ation Polymorphism	single nucleotide polymorphisms and indels	electrophoretic mobility shift of hybridized probe	yes	one		(Orita et al. 1989)
1989	STS	Sequence Tagged Site	any polymorphism	make tag by sequencing a contig from any marker type	N/A	N/A		(Olson et al. 1989)
1989	SSR	Simple Sequence Repeat	variable numbers of short polynucleotide repeats	PCR using primers for flanking sequence, polyacrylamide gel	yes	one	STR, SSLP, microsatellites	(Weber and May 1989; Jacob et al. 1991)
1990	RAPD	Random Amplified Polymorphic DNA	random presence/absense of primer sites in DNA	PCR with random primers discovered to flank polymorphisms, then gel	no	many (fingerprinting)		(Williams et al. 1990)
1991	STR	Short Tandem Repeat	see SSR	see SSR	yes	one	SSR, SSLP, microsatellites	(Huang et al. 1991)
1992	SSLP	Simple Sequence Length Polymorphism	see SSR	see SSR	yes	one	SSR, STR, microsatellites	(Dietrich et al. 1992)
1993	SCAR	Sequence Characterized	RAPD marker sites and length	PCR with primers specific to internal or	sometimes	one	indel	(Paran and Michelmore 1993)

		Amplified Region	polymorphic sites	flanking sequence, then gel				
1993	CAPS	Cleaved Amplified Polymorphic Sequence	restriction site location variation	PCR using primers for unique flanking sequence, then digestion and gel	yes	one		(Konieczny and Ausubel 1993)
1994	ISSR	Inter-Simple Sequence Repeat	See SSR	PCR using short primers matching many flanking sequences, polyacrylamide gel	yes	many (fingerprinting)		(Zietkiewicz, Rafalski, and Labuda 1994)
1995	AFLP	Amplified Fragment Length Polymorphism	variable length of restriction digest fragments	digest, ligate adapters, PCR with primers partially specific to sequence, gel	yes	many (fingerprinting)		(Vos et al. 1995)
1998	RGA	Resistance Gene Analog	plant disease resistance gene polymorphism	PCR with primers specific to disease resistance gene, polyacrylamide gel	yes	many (fingerprinting)		(Ellis and Jones 1998; Meyers et al. 1999; Chen, Line, and Leung 1998)
2001	SRAP	Sequence Related Amplified Polymorphism	indels in exons and introns	PCR, special primers with permissive temperature, polyacrylamide gel	yes	many (fingerprinting)		(Li and Quiros 2001)
2004	DArT	Diversity Arrays Technology	SNPs and indels	Semi-random sequence microarray hybridization and scanning	sometimes	variable		(Wenzl et al. 2004)
2006	SFP	Single Feature Polymorphism	variation in annealing affinity to 25-bp oligo	Specific sequence microarray hybridization and scanning	sometimes	variable		(Borevitz et al. 2003)
2006	GEM	Gene Expression Marker	gene transcript level variation	Hybridize transcript library to specific sequence microarray, scan	yes	one		(West et al. 2006)
2007	RAD	Restriction-site Associated DNA	sequence variation adjacent to restriction sites	Specific sequence microarray hybridization and	no	one		(Miller et al. 2007)

				scanning				
-	Indel	Insertion Deletion	insertion/deletion (indel)	PCR with primers specific to internal or flanking sequence, then gel	sometimes	one	SCAR	(See (Rafalski 2002)
-	SNP	Single Nucleotide Polymorphism	single nucleotide polymorphism (SNP)	Specific sequence microarray hybridization and scanning	sometimes	one		(See (Rafalski 2002)

Table 1: Genetic markers and their properties.

[a] Commonly used acronym for the marker.

[b] Expanded acronym name.

[c] Description of the polymorphism (and in some cases the visualization technique, which may be closely tied to the marker method).

[d] Codominance status of the marker.

[e] Number of different loci *usually* visualized by the marker (*one* for markers that assess a single locus; *many* for a fingerprint type of marker).

[f] Other markers that are fundamentally the same type of marker despite having different names.

[g] Reference to the paper defining the marker technique and in some cases to the paper first using the marker acronym.

* The marker acronym may have originated later than the invention of the marker technique, and the identification of the polymorphism upon which the marker is based may have occurred earlier than the invention of the technique.

Metric	**Run #1 (A)**	**Run #2 (B)**	**Run #3 (C)**	**Run #4 (D)**
Marker ID prefix (ID_PREFIX)	IGG_HP14A_	IGG_HP14B_	IGG_HP14C_	IGG_HP14D_
Genome 1	S. lycopersicum SL2.50	same	same	same
Genome 2	S. pennellii V2.0	same	same	same
k	14	same	same	same
Genome sizes (= number of k-mers)	824/990 Mbp	same	same	same
Unique k-mers	24.7/23.9 M	same	same	same
Common unique k-mers	8.9 M	same	same	same
Minimum k-mers per LCR (KMIN)	4	2	4	2
Minimum amplicon size (AMIN)	200	250	300	400
Maximum amplicon size (AMAX)	700	800	800	1500
Min. ampl. size diff. at AMIN (ADMIN)	100	100	200	50
Min. ampl. size diff. at AMAX (ADMAX)	100	200	200	300
LCRs	102 K	106 K	90.4 K	72.5 K
Non-overlapping Indel Groups	11 K	9.2 K	5.0 K	32 K
Overlapping Indel Groups	333 K	31.3 K	113 K	250 K
Overlapping unvalidated markers	26.6 K	11.7 K	9.3 K	97.6 K
Overlapping ePCR-validated IGG markers	21,654	9,437	7,163	91,947
Non-overlapping ePCR-validated IGG markers	5,526	3,720	2,332	16,442

Table 2. Metrics for four separate runs of IGGPIPE on *S. lycopersicum/S. pennellii* genomes using four different sets of parameters. Note how the initial unique k-mer pool (metric "Unique k-mers") is filtered down at each step of the IGGPIPE pipeline until finally converging at non-overlapping validated candidate IGG markers. Each run uses a different marker ID prefix to distinguish the markers. The IGG markers from these runs are provided as a supplemental data file. The metrics k, KMIN, AMIN, AMAX, ADMIN, and ADMAX are all user-specified parameters.

Metric	**Tomato / S. pennellii**				**A. thaliana Col-0 / Ler-0**		
Run number	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)
k	12	13	14	15	12	13	14
Minimum k-mers per LCR (KMIN)	2	same	same	same	4	same	same
Min. k-mer-to-k-mer distance in bp (DMIN)	10	same	same	same	15	same	same
Minimum amplicon size (AMIN)	250	same	same	same	400	same	same
Maximum amplicon size (AMAX)	800	same	same	same	1500	same	same
Min. ampl. size diff. at AMIN (ADMIN)	100	same	same	same	50	same	same
Min. ampl. size diff. at AMAX (ADMAX)	200	same	same	same	300	same	same
Genome sizes (= number of k-mers)	824/990 Mbp	same	same	same	120/118 Mbp	same	same
Total CPU time (BSD time)	13 min	27 min	113 min	906 min	6 min	30 min	136 min
Maximum memory usage (BSD time)	1.5 Gb	1.5 Gb	2.4 Gb	6.3 Gb	0.37 Gb	1.2 Gb	3.2 Gb
Operating system waits (BSD time)	24	180	4204	11357	104	928	5418
Unique k-mers	0.18/0.14 M	3.5/3.2 M	25/24 M	89/89 M	0.70/0.71 M	6.8/6.9 M	27/27 M
Common unique k-mers	43 K	1.1 M	8.9 M	34 M	563 K	5.7 M	23 M
LCRs	330	42 K	106 K	122 K	10.2 K	7.6 K	8.2 K
Overlapping Indel Groups	0	835	31 K	209 K	2.3 K	66 K	150 K
Overlapping unvalidated IGG markers	0	477	12 K	35 K	1.6 K	28 K	26 K
Overlapping ePCR-validated IGG markers	0	376	9437	28379	1588	28031	25201
Non-overlapping ePCR-validated IGG mrkrs	0	295	3720	7665	528	2392	2523

Table 3. Metrics for four separate runs of IGGPIPE on *S. lycopersicum/S. pennellii* genomes using four different values of k, and three separate runs on *A. thaliana* Col-0/Ler-0 accessions. All other parameters besides k were unchanged. The metrics k, KMIN, DMIN, AMIN, AMAX, ADMIN, and ADMAX are all user-specified parameters. Three measurements of computer resource usage are provided: CPU time, memory usage, and number of operating system waits, all gathered with the BSD time utility running on a system with an Intel 2.4 GHz Core 2 Duo CPU, 16 Gb DRAM, and Mac OSX 10.11.4. IGGPIPE memory requirements are modest (but increase with increasing K and increasing genome size), and CPU time increases dramatically with increasing K. For the genomes and parameters shown here, the IGGPIPE software can be run on a personal laptop computer.

#	IGG ID[a]	Chr[b]	Expected Size[c]		Dif Size[d]	Bands[e] M82/PENN	Correct Size[f] M82/PENN	Codom[g]	PrimerFwd	PrimerRev
			M82	PENN						
1	IGG_HP14B_179	1	405	616	-211	1/1	YES/YES	YES	GACACTCAGCCTAAGTTGCAG	TACACTGAGGCATCGTCTCC
2	IGG_HP14A_882	1	377	275	102	1/1	YES/YES	YES	CCTACCTGGGACTCAATCTGT	TCAGTGTATAAGCTTGACCTCCA
3	IGG_HP14B_1342	2	281	419	138	1/1	YES/YES	YES	ATTATCAGCTCCCAGACCCC	TGAGGATGCTTCATATCGCC
4	IGG_HP14B_2155	2	371	554	183	1/1	YES/YES	YES	AAGCAGTGGTCGGTGATCAG	CGTTCCACATGACTATCGGAC
5	IGG_HP14B_2564	3	467	346	-121	1/1	YES/YES	YES	TAAAGCTTCCGAGGCCTATG	TTTCACCCTCGTCGAGTCTC
6	IGG_HP14A_7145	3	234	442	-208	1/1	YES/YES	YES	TCGGGTCTGTTCTACTGCTT	CCTCCTGGTGTGTATGGGAG
7	IGG_HP14B_3418	4	389	681	292	1/1	YES/YES	YES	TTATGCACGTCTCCTCAAGG	GAGAGGTTCTTGGTGGATGAC
8	IGG_HP14B_3934	4	495	749	254	0/0	NO/NO	NO	CGTCCCTTTGTCACGTGTC	GGAGCGTAAATTTGAGCTACTTG
9	IGG_HP14B_4268	5	799	489	-310	1/1	YES/YES	YES	CCCCTAAAGATCTGCTCGAAATC	TGACCACGTTTCCCTTCTAATG
10	IGG_HP14B_4544	5	527	325	-202	1/1	YES/YES	YES	CCTCTGGCAATCTTCAGGTG	TCCTGCCTATTTTGCTTGCTG
11	IGG_HP14B_4721	6	531	767	236	1/1	YES/YES	YES	ACCAGAGAGAACCCTTGATCC	GCTCTTTCAACTTTGCCTGTG
12	IGG_HP14B_5488	6	543	741	198	1/0	YES/NO	NO	TCATAATGGCCAGAAACCCG	CACGCAACAATCAACATTTAGGG
13	IGG_HP14B_6105	7	523	753	-230	1/1	YES/YES	YES	GGCTACCAGTCCTGTCGAG	TTTCGCGCTGATGAACACC
14	IGG_HP14C_4527	7	558	783	-225	1/1	YES/YES	YES	GACAGTGGCGGAGTGAGATA	AAGTACGCTATGGTTCGGGG
15	IGG_HP14B_6403	8	670	448	-222	1/1 faint	YES/YES	YES	AACAACCAGTCAATAAGCTGC	TCAAGGAATCAACTGTGCCTC
16	IGG_HP14A_15764	8	329	720	-391	1/1	YES/YES	YES	AATCTTGATGAGTGTCCGCG	GCACAAAGCGGGTCTAGAAA
17	IGG_HP14B_7175	9	790	563	-227	0/0	NO/NO	NO	GACACAGCTGTTAATTGGACATC	CAAAGAAGATGCACGTGGAAC
18	IGG_HP14A_16708	9	491	697	-206	1/1	YES/YES	YES	GTTTGATCCTGCGCACACC	CCAGTTAACAGAGGTAAAAGCCA
19	IGG_HP14A_17903	10	555	357	198	1/1	YES/YES	YES	ACATTCACACAAACCGCACA	TGTAGCGCTGGTAATGCTTA
20	IGG_HP14A_18108	10	282	411	-129	1/1	YES/YES	YES	ACCGAACTAGCCAGACCAAA	TTTTGCTTTGGTGCTCGTCA
21	IGG_HP14B_8438	11	406	650	244	1/1	YES/YES	YES	TCATCAGCTTGTT	GACGGTGGAGTT

									GGGTATGTG	GTGATATGG
22	IGG_HP14A_20373	11	600	703	-103	1/1	YES/YES	YES	CGCTTGCCTTCTT CGTTAGA	GACCACGATTCT GCTTTGGT
23	IGG_HP14B_9081	12	377	646	-269	1/1	YES/YES	YES	ACCCTAAGCTGC TGTAGTGC	AACCCGCAGCCT TCAAAAC
24	IGG_HP14B_9341	12	331	723	392	faint	YES/YES	YES	TCTACAAGCATG CGATCAAGTC	TCAACAAGGAGG CTTTAACCC

Table 4. IGG markers tested in *S. lycopersicum* and *S. pennellii*. PCR gel results are shown in **Figure 6**. Out of 24 markers tested, 21 (87.5%) amplified with the predicted amplicon sizes in both species, two failed to amplify in either species, and one didn't amplify in *S. pennellii*.

[a] IGGPIPE spreadsheet ID number for IGG marker.

[b] Chromosome number.

[c] Expected amplicon sizes in M82 (*S. lycopersicum*) and PENN (*S. pennellii*).

[d] Expected difference in size of the amplicons.

[e] Number of bands observed for M82 and PENN.

[f] Was the observed band size the predicted size?

[g] Was the marker co-dominant (different amplicon size in both species)?

Chr #[a]	Polymorphism name	Col-0 amp[b]	Ler-0 amp[c]	Difference in size[d]	Primer Fwd	Primer Rev
1	nga111	128	162	34	TGTTTTTTAGGACAAATGGCG	CTCCAGTTGGAAGCTAAAGGG
1	F16J7-TRB	165	114	51	TGATGTTGAGATCTGTGTGCAG	GTGTCTTGATACGCGTCGAT
2	nga168	150	135	15	GAGGACATGTATAGGAGCCTCG	TCGTCTACTGCACTGCCG
2	ciw3	230	200	30	GAAACTCAATGAAATCCACTT	TGAACTTGTTGTGAGCTTTGA
3	ciw11	180	230	50	CCCCGAGTTGAGGTATT	GAAGAAATTCCTAAAGCATTC
3	nga172	162	136	26	CATCCGAATGCCATTGTTC	AGCTGCTTCCTTATAGCGTCC
4	JV30/31	195	165	30	CATTAAAATCACCGCCAAAAA	TTTTGTTACATCGAACCACACA
4	nga8	154	198	44	TGGCTTTCGTTTATAAACATCC	GAGGGCAAATCTTTATTTCGG
5	ciw8	100	135	35	TAGTGAAACCTTTCTCAGAT	TTATGTTTTCTTCAATCAGTT
5	ciw15	177	120	57	TCCAAAGCTAAATCGCTAT	CTCCGTCTATTCAAGATGC

Table 5. Length polymorphism markers for Arabidopsis thaliana accessions Col-0 and Ler-0, found with Polymorphism/Allele search tool at arabidopsis.org.

[a] Chromosome number.

[b, c] Expected amplicon sizes in Col-0 and Ler-0, respectively.

[d] Expected difference in size of the amplicons.

Metric	**A. thaliana Col-0 / Ler-0**
Marker ID prefix (ID_PREFIX)	IGG_CL13_
k	13
Minimum number k-mers per LCR (KMIN)	10
Minimum LCR k-mer spacing in bp (DMIN)	30
Minimum amplicon size (AMIN)	400
Maximum amplicon size (AMAX)	1500
Min. ampl. size diff. at AMIN (ADMIN)	50
Min. ampl. size diff. at AMAX (ADMAX)	300
Genome sizes (= number of k-mers)	120/118 Mbp
Unique k-mers	6.8/6.9 M
Common unique k-mers	5.7 M
LCRs	6.2 K
Overlapping Indel Groups	34 K
Overlapping unvalidated IGG markers	14 K
Overlapping ePCR-validated IGG markers	14 K
Non-overlapping ePCR-validated IGG markers	2072

Table 6. Parameters and statistics for IGGPIPE run using *A. thaliana* accessions Col-0 and Ler-0. The IGG markers from this run are provided as a supplemental data file. The metrics k, KMIN, DMIN, AMIN, AMAX, ADMIN, and ADMAX are all user-specified parameters.

#	IGG_ID[a]	Chr[b]	Expected Size[c]		Dif Size[d]	Bands[e] Col-0/Ler-0	Correct Size[f] Col-0/Ler-0	Codom[g]	Primer Fwd	Primer Rev
			Col-0	Ler-0						
1	IGG_CL13_194	1	1033	556	477	1/1	YES/NO	NO	GGGTCAATCATCGGTGTTTTG	TGCATGCCTCTGTTCAACTG
2	IGG_CL13_3510	2	1282	655	627	1/1	YES/YES	YES	TTCATCCGACTCAATTGGCG	TCGTTTATTCAGGACAGCTGC
3	IGG_CL13_8083	3	643	997	354	1/1	YES/YES	YES	AAGAGACAGAGACGGGTTGC	CGTTGACTGAAGCTCAAGGG
4	IGG_CL13_8474	4	1119	652	467	1/1	YES/YES	YES	GTAGAATCAGCGAACAATGTAGC	TCAAAACAACAAAATAAGGCCGG
5	IGG_CL13_13563	5	956	445	511	1/1	YES/YES	YES	GTCGATTAGGTCAACGGCTG	GGTTTGACCCCTTTGCATCG
6	IGG_CL13_3168	1	765	450	315	1/1	YES/YES	YES	TCTCTTTCGTGGACAGAGCC	TCGCACTTCAATTTCAGACCG
7	IGG_CL13_5391	2	883	492	391	1/1	YES/YES	YES	GTCAGTAAATTAACACACGTCCG	CGACTGAAAGATGTTGAAATGGG
8	IGG_CL13_5788	3	897	457	440	1/1	YES/YES	YES	CATCCAGACATAAACATCATGCG	GAGAAGGCACAGCAGACAAG
9	IGG_CL13_10502	4	762	466	296	1/1	YES/YES	YES	AATGGATTCCTGCGACGGAG	TCTTCGGATCAGAGCCAAGC
10	IGG_CL13_10806	5	507	908	401	1/1	YES/NO	NO	GTCGATTAGGTCAACGGCTG	GGTTTGACCCCTTTGCATCG

Table 7. IGG markers tested in *Arabidopsis thaliana* accessions Col-0 and Ler-0. PCR gel results are shown in **Figure 7A**. Out of ten markers tested, eight (80%) amplified with the predicted amplicon sizes in both species.

[a] IGGPIPE spreadsheet ID number for IGG marker.

[b] Chromosome number.

[c] Expected amplicon sizes in Col-0 and Ler-0, respectively.

[d] Expected difference in size of the amplicons.

[e] Number of bands observed for Col-0 and Ler-0.

[f] Was the observed band size the predicted size?

[g] Was the marker co-dominant (different amplicon size in both accessions)?

Metric	**Tomato / S. pennellii / S. tuberosum Run #1 (A)**	**Tomato / S. pennellii / S. tuberosum Run #2 (B)**
Marker ID prefix (ID_PREFIX)	IGG_HPT14A_	IGG_HPT14B_
Genome 1	*S. lycopersicum* ITAG2.4	same
Genome 2	*S. pennellii* V2.0	*S. pennellii* V1.0
Genome 3	S. tuberosum DM V4.03	same
k	14	14
Minimum number k-mers per LCR (KMIN)	2	4
Minimum LCR k-mer spacing in bp (DMIN)	5	1
Minimum amplicon size (AMIN)	300	400
Maximum amplicon size (AMAX)	1500	1500
Min. ampl. size diff. at AMIN (ADMIN)	50	50
Min. ampl. size diff. at AMAX (ADMAX)	300	300
Max. bp beyond k-mer primer extends (EXTENSION_LEN)	15	10
Primer GC clamp	Yes	No
Genome sizes (= number of k-mers)	120/118 Mbp	120/118 Mbp
Unique k-mers	24.7/23.9 M	24.7/23.9 M
Common unique k-mers (all 3 genomes)	2.8 M	2.8 M
LCRs	27.9 K	24.6 K
Overlapping Indel Groups	61 K	240 K
Overlapping unvalidated IGG markers	18 K	29 K
Overlapping ePCR-validated IGG markers	18.2 K	29.5 K
- 2 distinct alleles (two genomes have same size alleles)	17665	28505
- 3 distinct alleles	534	951
Non-overlapping ePCR-validated IGG markers	5203	5549
- 2 distinct alleles (two genomes have same size alleles)	4975	5271
- 3 distinct alleles	228	278

Table 8. Parameters and statistics for two runs, designated A and B, of IGGPIPE using a three-way genome analysis of *S. lycopersicum* (tomato), *S. pennellii,* and *S. tuberosum* (potato). The introgression line development and 3-allele marker testing using *S. sitiens*, described in the text, used the run B markers. The two runs use a different marker ID prefix to distinguish the markers. The

IGG markers from these runs are provided as supplemental data files. The metrics k, KMIN, DMIN, AMIN, AMAX, ADMIN, and ADMAX are all user-specified parameters.

		Polymorphism Type								Amplicon size[e]				
		2-way[c]						3-way[d]						
Marker[a]	IGG ID[b]	L v. P	L v. T	P v. T	L v. S	P v. S	T = S	LPT	LPS	L	P	T	Primer Fwd	Primer Rev
M-1	IGG_HPT14B_754	Yes	Yes	Yes	Yes	No	No	Yes	No	991	737	584	AGAGAACTTAGTGCAGGCAG	TGCTCTGGGTCTCCTAGTTC
*M-2	IGG_HPT14B_926	Yes	Yes	Yes	N/A	N/A	N/A	Yes	N/A	1247	1519	765	TCACAATCATCACGGAGCAAC	ACCACAGCTTCTACGCCTTA
M-3	IGG_HPT14B_1105	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1103	809	682	TGAAACACGAAAGGAGCTTGT	AGCCGTTCATCAGCAATCAA
M-4	IGG_HPT14B_1608	Yes	Yes	Yes	Yes	No	No	Yes	No	1093	1508	725	TACTCGCTCTTCATGACGCT	CTAATTCGCAGCAAATCGAAAC
M-5	IGG_HPT14B_4592	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	943	1121	751	ATTGCATACCCACTGCGAGG	CAGGCGGATGTGTGAGTTAT
M-6	IGG_HPT14B_4936	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1327	582	671	AAGAGAGGCATTCGAGGGAG	CATGCGCCACGTGTACTC
M-7	IGG_HPT14B_5427	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	792	932	527	TGTTGAGGGCTGGTGGATAC	CTGTAGCAGGCTCATCTTAAAAC
M-8	IGG_HPT14B_6347	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	969	1196	1449	CTCATGGCCACGAATGTCTG	GGTGGTGGCAGTAACGTTTC
M-9	IGG_HPT14B_8121	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	824	1150	1366	AGAGCCGCCTTTCCTCCTA	TTTCAAGCTGGCATTCGAGC
*M-10	IGG_HPT14B_8235	Yes	Yes	Yes	N/A	N/A	N/A	Yes	N/A	1056	649	885	TTACCACGTTCTCCAGCAGG	CTCATGAAAACCTCCGACCTG
M-11	IGG_HPT14B_8264	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	898	1097	1358	GCCGCTACTTCTCGATCAAA	TTGTTCAGGTGCCTCGTG
M-12	IGG_HPT14B_8563	Yes	N/A	N/A	Yes	Yes	N/A	N/A	Yes	651	430	913	AAAAGGAAGCGCGAGATGAG	CCAGTGGAGCAGGTTACTC
*M-13	IGG_HPT14B_8811	Yes	Yes	Yes	N/A	N/A	N/A	Yes	N/A	579	1012	815	CAAGGATCTGGCTGGGTAGT	GGTACCCTTGCTCGATTAGATAG
M-14	IGG_HPT14B_8853	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	841	1247	1051	TCCAACTCCGGACAAAGGT	TCTCACGGTATAAGCAGAGCA
M-15	IGG_HPT14B_9447	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	875	1111	1368	AGGGCACGTACCAGCATAAA	ATGATGGGATGCTGTCGACA
M-16	IGG_HPT14B_10635	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	963	786	615	TAAGCTGTAACGCAATCCCG	CCCTGTGGAGCCAACAAT
*M-17	IGG_HPT14B_11038	Yes	Yes	Yes	N/A	N/A	N/A	Yes	N/A	1371	1120	907	TGACAGTTCAAGCCCACAG	GTGAACACTCCCTGACTTTGT
M-18	IGG_HPT14B_15532	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1425	1148	611	TTATCTTGCTGTGCTTGCCC	CAAGTTTATGGGGTGGCACA
M-19	IGG_HPT14B_15683	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	464	528	846	CGTTTGATGGTTGGTGCGTA	TAGTCTACGCGGCGCATC

M-20	IGG_HPT14B_15777	Yes	Yes	Yes	Yes	No	No	Yes	No	877	718	1225	GGCACTTGTGAGCAGTATCC	TGCAAGTCGACAGTATCTAACA
M-21	IGG_HPT14B_21272	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	469	554	899	GGGATCTTCGCACCTAAATCC	ATTCCGACTGCCTGGTGTTT
M-22	IGG_HPT14B_21501	Yes	Yes	Yes	Yes	No	Yes	Yes	No	1290	983	765	CTTCCCTCATCTCGTCGGG	AATGCGTGCAGAAGAAGACG
M-23	IGG_HPT14B_23704	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1034	783	1406	GCGGCGGATTGGGAAATC	CGGCGAGTAGGAGAACTGAG
M-24	IGG_HPT14B_24108	Yes	Yes	Yes	Yes	No	No	Yes	No	779	933	1504	GCTTATGCGGGTTTGTTAGAAA	CGGTATAACTTCACGGCATTAAG
M-25	IGG_HPT14B_25784	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	463	590	902	GCATCTTCTCAACGTACCTCTC	CCAGTTTTACCACCTAAACCGG
M-26	IGG_HPT14B_26897	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1393	1104	897	TGTCACCAGCATACTTTGTCA	ACTGATAACTGGGTGAAAGGTG
M-27	IGG_HPT14B_26991	Yes	Yes	Yes	No	Yes	No	Yes	No	1051	866	1333	CTGGAAGCAGCAGGTATTCT	GCTCGGATTGCATTCACTTG
M-28	IGG_HPT14B_27175	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	863	1050	728	AGGAGAAGACTGGCGGAAAG	TGGAAAGCACAGAAACAGATGA
M-29	IGG_HPT14B_27897	Yes	N/A	N/A	Yes	Yes	N/A	N/A	Yes	972	480	1339	AAGTGCTGGCGTAAATTCAC	AGTGTGTTTGTGAGTGAAGCA
M-30	IGG_HPT14B_28355	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	1469	1148	964	TGGACCCCTATTGACTTAGTTGT	GTAGGGAGGGGCACATAACC
M-31	IGG_HPT14B_28659	Yes	Yes	Yes	No	Yes	No	Yes	No	1234	1026	758	TGGTTGCCTTGGCTTAGAAG	TGAACCACTCAACGCGGG
M-32	IGG_HPT14B_29367	Yes	Yes	Yes	Yes	No	No	Yes	No	830	1207	1011	CCTGAATCCCTGAGAATCCCA	AACACTGTTTAGAAGCCGGT

Table 9. PCR testing of IGG markers for three-way genome analysis. A set of 32 IGG markers were selected for PCR testing. DNA from *S. lycopersicum* (**L**), *S. pennellii* (**P**), *S. tuberosum* (**T**), and *S. sitiens* (**S**) was amplified and polymorphism scored. PCR gel results are shown in **Figure 7B, C, D.**

[a] Marker number used in PCR experiments.

[b] IGGPIPE spreadsheet ID number for IGG marker.

[c,d] 2-way markers have two distinct amplicon sizes in two species, 3-way markers have three distinct sizes in three species

[e] Predicted amplicon sizes. *S. sitiens* has no predicted size due to absence of a reference genome.

'yes': base pair difference (on 2% agarose gel) between genotypes was easily identified

'no': base pair difference was not easily identified.

'N/A': comparison could not be made because one or more genotypes did not amplify under conditions tested.

'*': Did not amplify in *S. sitiens*

feature/item	IGGPIPE	IMDP	PolyMarker
Reference	(this paper)	(Lu et al. 2015)	(Ramirez-Gonzalez, Uauy, and Caccamo 2015; Ramirez-Gonzalez et al. 2015)
Polymorphism type	Indel Groups	Small indels	Small indels
Assay method	PCR and agarose 1% gel	PCR and polyacrylamide gel	KASP proprietary method and PCR with polyacrylamide gel
Co-dominant	yes	yes	yes
Multiple alleles detectable?	yes, can force discovery of multiallelic-only markers	yes	yes
Input	Two genome sequences	Genome sequences or NGS resequencing data	Reference genome and known SNPs
Output	File of IGG markers with positions and primer sequences	Indel markers with primers	Primers for SNP markers
Sample run species	*S. lycopersicum/S. pennellii*	*Oryza sativa* japonica/indica varieties	Polyploid wheat
# markers from sample run	87,351 overlapping, 16,548 non-overlapping	1,042	81,587
Mean amplicon size (bp)	745 (parameters=400 to 1500) (of non-overlapping markers)	159 (of 95 tested markers)	None given, 100 bp mentioned in text
Mean amplicon size difference (bp)	284 (parameters=50 to 300) (of non-overlapping markers)	15 (of 95 tested markers)	None given
# markers tested	55 (tentative)	95	35
# markers work as predicted	48 (87%) (tentative)	93 (multiple cultivars) (98%)	28
# cultivars tested at one time	3	12	38
Open access	TBD	yes	yes
Platform	Unix-based, tested on OSX	Linux (tested on Ubuntu 64)	BioGem
Operating Environment	Command line	LONI pipeline processing environment, graphical	Web interface
External tools used	Jellyfish, Primer3, e-PCR	MUMmer3, Pindel, Primer3, MFEprimer2, LONI, BWA,	Primer3, MySQL

		samtools, FastQC, QualiMap, Trimmomatic, LAMP (Linux, Apache, MySQL, PHP), LastZ	
Language environments used	C++, R, Perl, bash	R, perl, bash	BioGem, bioruby, Java
Installation	Command line installation, install and run guides provided	LONI installation	Install private web server
Additional data provided	tomato/*S. pennellii* IGG marker files, *A. thaliana* Col-0/Ler-0 IGG marker files.	Rice Indel marker database on the web	none
Additional utilities provided	Dot plot of markers; convert between tsv, csv, gff3, gtf; merge data between two files based on genomic position overlap/proximity.	N/A	N/A

Table 10. Feature and performance comparison of IGGPIPE and two other *in silico* marker creation packages, IMDP (Lu et al. 2015) and PolyMarker(Ramirez-Gonzalez, Uauy, and Caccamo 2015).

Description	Genomes	FASTA	k	NDA - MIN	LMIN = AMIN	KMIN / DMIN	DMAX = AMAX	ADMIN / ADMAX	Comments
Regular 2-accessions/2-species diploid markers	Two sequenced genomes (accessions or highly syntenic species). At least one can be chromosomal if chromosomal position coordinates are needed.	2 FASTA files. If non-chromosomal assembly, remove all small scaffolds.	13.. 16	2	>= 100	2..4 / 1..10	AMIN+ 10 ..5000	1..4900 / ADMIN ..4900	Each marker's 2 primers produce 1 uniquely sized amplicon in each species.
Multi-accession/ multi-species multiallelic markers	Three (or more) sequenced genomes, say N of them.	N FASTA files, one per genome, with unwanted sequences removed.	14.. 17	2..N	"	"	"	"	There are between NDAMIN and N unique amplicon sizes per marker (NDA column). If fewer than N, some genomes share the same amplicon size.
Fingerprinting markers	Numerous sequenced genomes, say N of them. Perhaps N > 10, but this is untested.	N FASTA files	13.. 17	Say 5	"	"	"	"	There are between 5 and N unique amplicon sizes per marker, some species may share. Use 2 or more markers to obtain unique sets of amplicon sizes for each species.
2-accessions/2-species polyploid markers	Two good-quality polyploid genomes.	2 FASTA files, each containing all subgenomes.	15 .. 17^a	2	"	"	"	"	Each marker's 2 primers produce 1 uniquely sized amplicon in each species. Marker density is lower because of subgenome similarity.
Polyploid sub-genome markers	One good-quality polyploid genome, chromosomal, not scaffold-based.	Split into N FASTA files, each with one subgenome. N=number of subgenomes.	15 .. 17^b	2..N	"	"	"	"	There are between NDAMIN and N unique amplicon sizes per marker (NDA column). If N, each sub-genome produces its own unique amplicon size.
Polyploid (or diploid^c) presence/absence marker with control	One good-quality polyploid genome, chromosomal. ^{c}Or, diploid genome.	2 FASTA files, one with target subgenome, one with other subgenomes.	15 .. 17^a	2	"	"	"	"	Each marker's 2 primers produce 1 uniquely sized amplicon in the target subgenome and one in one of the other subgenomes.
Two target regions on different chromosomes, polyploid (or diploidc)	One good-quality polyploid genome, chromosomal. cOr, diploid genome.	d2 FASTA files, one with target 1 chromosome, the other with target 2 chromosome.	"	2 (or 3d)	"	"	"	"	Each marker's 2 primers produce 1 uniquely sized amplicon in target chromosome 1 and one in target chromosome 2.

cDNA markers	Two good-quality assembled transcriptomes of accessions or related species.	2 FASTA files with transcriptomes. Best to remove contigs smaller than LMIN+ADMIN	12 .. 15	2	"	"	"	"	Each marker's 2 primers when amplifying from a cDNA library produce 1 uniquely sized amplicon in each species.
Diploid genotyping markers	One genome with a large database of indels commonly found within it.	2 FASTA files, one the main genome, the other the same genome but modified to apply the indels.	13 .. 16	2	"	"	"	"	Each marker's 2 primers produce 1 uniquely sized amplicon from the main genome and 1 from the modified genome.
Identify major structural variation	Two sequenced genomes (accessions or highly syntenic species), both chromosomal, not scaffold-based.	2 FASTA files, one for each genome.	14 .. 16	NA	100	4 / 1	3000	100 / 100	'make findLCRs' clone dotplot.template and edit it 'Rscript code/R/dotplot.R <myfile>'

Table 11: IGGPIPE usage cases. Parameter values are meant to provide a rough guide to what is reasonable, but other values can also be used. Memory usage increases dramatically with k, so smaller values of k may be runnable on a personal computer, while larger values may require servers with more memory.

Notes:

[a] Larger k may be needed for more unique k-mers, to increase odds of finding markers that amplify uniquely in only one subgenome.

[b] Larger k may be needed for more unique k-mers, to increase odds of finding markers that amplify uniquely in each subgenome.

[c] Same technique works with diploid genomes, treating one target chromosome as a subgenome, but density of markers will be lower than with a polyploid since the chromosomes have less redundancy between them than polyploid subgenomes.

[d] This technique can be combined with the one on the previous row to generate markers that have a third amplicon that serves as a PCR control, by putting the remaining chromosomes into a third FASTA file and using NDAMIN=3. The density of markers will be lower.

Figure Legends

Figure 1 A. IGGPIPE: an IGG (Indel Group in Genomes) marker finder software pipeline. Two genome sequences (G1 and G2) are analyzed for common unique k-mers that identify locally conserved regions (LCRs), some of which are polymorphic for length, containing one or more indels between flanking conserved sequences, making them *Indel Groups*. Primers are designed in the flanking conserved regions and verified with e-PCR to produce candidate IGG markers. Pipeline software is shown in dashed boxes, data in solid line boxes. **B.** A new k-mer starts at each base position. Shown here are seven consecutive 14-mers common to two genomes. **C.** Number of unique k-mers in tomato (*S. lycopersicum*) and closely related *S. pennellii* species as a function of k, and number of unique k-mers common to both species. As k increases, the number of unique k-mers increases, gradually approaching the genome size limit. The common unique k-mer count does not keep increasing, but at some value of k will reach a peak, here around k=19 or k=20. **D.** With k=14, *S. lycopersicum*) and *S. pennellii* have almost 9 million unique k-mers in common between them.

Figure 2. A. Locally conserved regions (LCRs) are regions of paired contigs within the genomes under consideration (here G1 and G2) having a sufficient number and spacing of unique k-mers in common between the contigs. When indels are present within LCRs, they form the basis for creating candidate IGG markers. Common unique k-mers can connect pairs of contigs in many ways. The parameter DMAX is the maximum spacing between two adjacent k-mers of the same LCR, and k-mers farther apart than that are assigned to different LCRs. If the number of k-mers is less than parameter KMIN (here assumed to be 4), the k-mers are assumed to be random common unique k-mers not signifying a conserved region, and no LCR is called for that region (a, b, e). LCRs may have no indels in them (c, d, j) or there may be a single indel (b, f, h) or more than one (i). Different LCRs along a contig of one genome might include *different* contigs in the other genome (a, b, c, and e versus d). Some LCR regions may have one or more random interspersed k-mers connecting a contig pair that is different from the contig pair of the LCR (f). Some regions may have complex overlapping of more than one LCR (g). **B.** An alignment of *S. lycopersicum* and *S. pennellii* genomes in the region of an LCR on chromosome 1. Blue vertical lines are positions of common unique 14-mers. An indel is visible that might provide sufficient length polymorphism for an IGG marker surrounding this area. Red arrow points to one 14-mer whose region is enlarged below. **C.** Enlargement of the region around the third 14-mer in the above figure, showing a multiple alignment of the *S. lycopersicum* and *S. pennellii* genome sequences in this region, the primer generated by IGGPIPE, and the 14-mer itself. Alignments made with Geneious (Kearse et al. 2012).

Figure 3. Characteristics of indels found within Indel Groups, from an IGGPIPE analysis of: **A,C:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0 (K=14, AMIN=100, AMAX=3000, ADMIN=ADMAX=100); **B,D:** *A. thaliana* accessions Col-0/Ler-0 (K=13, other parameters the same). **A, B.** Each Indel Group from was plotted as a point, where the x-axis is the predicted amplicon size difference and the y-axis is the number of indels found in the Indel Group after aligning the two sequences. **C,D.** Similar plot but y-axis is indel size. The 45º line is Indel Groups containing a single indel that is responsible for the amplicon size difference. Some points lie above the line because a single Indel Group can have deletions in both genomes, at different places.

Figure 4. Additional characteristics of indels found within Indel Groups, from the same analysis cited in **Figure 3**. **A,C:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0; **B,D:** *A. thaliana* accessions Col-0/Ler-0. **C.** The number of indels of different sizes decreases approximately exponentially as the indel length increases. H: Heinz (*S. lycopersicum*), P: PENN (*S. pennellii*). **D.** Density of Indel Group indels within genomic features found in the LCRs containing the Indel Groups. Upstream is defined as within 1000 bp 5' of the 5'UTR, and downstream is within 1000 bp 3' of the 3'UTR of a gene, while intergenic is any position not falling into any of the other categories.

Figure 5. A, B. Distribution of differences in IGG marker amplicon sizes between the two analyzed genomes, from an IGGPIPE analysis of: **A:** *S. lycopersicum* SL2.50/ITAG2.4 / *S. pennellii* V2.0 (K=14, AMIN=400, AMAX=1500, ADMIN=50, ADMAX=300); **B:** *A. thaliana* accessions Col-0/Ler-0 (K=13, other parameters the same). A positive difference means the *S. lycopersicum or Col-0* amplicon is the larger, and negative means the *S. pennellii or Ler-0* amplicon is the larger. **C, D.** Density of IGG markers (top graph) and genes (bottom graph) along a representative chromosome, from the same analysis as above**. C:** Chromosome 1 of *S. lycopersicum* (tomato). Note positive correlation. **D:** Chromosome 2 of *A. thaliana* Col-0 accession.

Figure 6. Twenty four IGG markers, two per chromosome at locations within the first or last 15% of each chromosome, were chosen randomly from three different IGGPIPE runs using different sets of parameters and all analyzing the *S. lycopersicum* (SL2.50/ITAG2.4 pseudomolecules) and *S. pennellii* (V2.0 pseudomolecules) genomes. In 21 of the 24 markers (87.5%) amplifying *S. lycopersicum* cv. M82, *S. pennellii* (PEN), and F1 DNA, two bands of the expected amplicon sizes are seen (**Table 4**), one in each species. In two cases, no band is seen in either species, and in another case, only an *S. lycopersicum* band is seen.

Figure 7. Gel electrophoresis of PCR products of several candidate IGG markers from two IGGPIPE runs. **A.** Testing primers generated against *Arabidopsis thaliana* accessions Landsberg and Columbia. PCR product resolved on 2% gel. **M:** BioLabs QuickLoad 100 bp Ladder; **C:** Columbia-0; **LC:** Landsgerg-Columbia hybrid; **L:** Landsberg-0. Eight of 10 show expected product sizes (**Table 7**). **B-D.** PCR products by gel electrophoresis using IGG markers from triallelic marker run with *S. lycopersicum, S. pennellii,* and *S. tuberosum* genomes. **M:** O'GeneRuler 1Kb Plus Ladder; **L:** *S. lycopersicum*; **P:** *S. pennellii*; **S:** *S. sitiens*; and **T:** *S. tuberosum*. **B**. IGG marker #B_9447 shows three-way polymorphism between the three genomes of interest and amplicons are of predicted size (**Table 9**). In addition, *S. tuberosum* and *S. sitiens* share the same allele. **C.** Marker #B_5427 also shows three-way polymorphism between the three genomes of interest. In this case, the S. tuberosum amplicon is closer to 700 bp than the predicted 527 bp. *S. lycopersicum* and *S. pennellii* have predicted amplicon sizes. In addition, *S. tuberosum* and *S. sitiens* have a very small or zero size difference. **D.** Markers #B_24108, B_25784, and B_26991 also indicate three-way polymorphism between *S. lycopersicum*, *S. pennellii*, and *S. tuberosum.* However, *S. sitiens* shares an allele with either *S. pennellii* (B_24108) or *S. lycopersicum* (B_26991). Presence of multiple bands is observed for select genotypes.

Parsed Citations

Ahmed, A., A. S. Ferreira, and R. A. Hartskeerl. 2015. 'Multilocus sequence typing (MLST): markers for the traceability of pathogenic Leptospira strains', Methods Mol Biol, 1247: 349-59.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bolger, A., F. Scossa, M. E. Bolger, C. Lanz, F. Maumus, T. Tohge, H. Quesneville, S. Alseekh, I. Sorensen, G. Lichtenstein, E. A. Fich, M. Conte, H. Keller, K. Schneeberger, R. Schwacke, I. Ofner, J. Vrebalov, Y. Xu, S. Osorio, S. A. Aflitos, E. Schijlen, J. M. Jimenez-Gomez, M. Ryngajllo, S. Kimura, R. Kumar, D. Koenig, L. R. Headland, J. N. Maloof, N. Sinha, R. C. van Ham, R. K. Lankhorst, L. Mao, A. Vogel, B. Arsova, R. Panstruga, Z. Fei, J. K. Rose, D. Zamir, F. Carrari, J. J. Giovannoni, D. Weigel, B. Usadel, and A. R. Fernie. 2014. 'The genome of the stress-tolerant wild tomato species Solanum pennellii', Nature Genetics, 46: 1034-8.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bombarely, A., N. Menda, I. Y. Tecle, R. M. Buels, S. Strickler, T. Fischer-York, A. Pujar, J. Leto, J. Gosselin, and L. A. Mueller. 2011. 'The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl', Nucleic Acids Research, 39: D1149-55.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Borevitz, J. O., D. Liang, D. Plouffe, H. S. Chang, T. Zhu, D. Weigel, C. C. Berry, E. Winzeler, and J. Chory. 2003. 'Large-scale identification of single-feature polymorphisms in complex genomes', Genome Research, 13: 513-23.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis. 1980. 'Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms', American Journal of Human Genetics, 32: 314-31.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Chen, X. M., R. F. Line, and H. Leung. 1998. 'Genome scanning for resistance-gene analogs in rice, barley, and wheat by high-resolution electrophoresis', Theoretical and Applied Genetics, 97: 345-55.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Darling, A. E., B. Mau, and N. T. Perna. 2010. 'progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement', PLoS One, 5: e11147.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

DeVerna, J. W., C. M. Rick, R. T. Chetelat, B. J. Lanini, and K. B. Alpert. 1990. 'Sexual hybridization of Lycopersicon esculentum and Solanum rickii by means of a sesquidiploid bridging hybrid', Proc Natl Acad Sci U S A, 87: 9486-90.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Diegoli, T. M. 2015. 'Forensic typing of short tandem repeat markers on the X and Y chromosomes', Forensic Sci Int Genet, 18: 140-51.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Dietrich, W., H. Katz, S. E. Lincoln, H. S. Shin, J. Friedman, N. C. Dracopoli, and E. S. Lander. 1992. 'A genetic map of the mouse suitable for typing intraspecific crosses', Genetics, 131: 423-47.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Dinov, I. D., J. D. Van Horn, K. M. Lozev, R. Magsipoc, P. Petrosyan, Z. Liu, A. Mackenzie-Graham, P. Eggert, D. S. Parker, and A. W. Toga. 2009. 'Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline', Front Neuroinform, 3: 22.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Doyle, Jeff J. 1987. 'A rapid DNA isolation procedure for small quantities of fresh leaf tissue', Phytochem bull, 19: 11-15.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Edgar, R. C. 2004. 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', Nucleic Acids Research, 32: 1792-7.

Pubmed: Author and Title

CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ellis, J., and D. Jones. 1998. 'Structure and function of proteins controlling strain-specific pathogen resistance in plants', Current Opinion in Plant Biology, 1: 288-93.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Eshed, Y., M. Abu-Abied, Y. Saranga, and D. Zamir. 1992. 'Lycopersicon esculentum lines containing small overlapping introgressions from L. pennellii', Theoretical and Applied Genetics, 83: 1027-34.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Fu, J. J., Z. Q. Mei, M. Tania, L. Q. Yang, J. L. Cheng, and M. A. Khan. 2015. 'Development of RAPD-SCAR markers for different Ganoderma species authentication by improved RAPD amplification and molecular cloning', Genet Mol Res, 14: 5667-76.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Ratsch, and R. Mott. 2011. 'Multiple reference genomes and transcriptomes for Arabidopsis thaliana', Nature, 477: 419-23.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

GitHub. 2015. 'About GitHub', Accessed June 24, 2015. https://github.com/about.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Guyon, R., F. Senger, M. Rakotomanga, N. Sadequi, F. A. Volckaert, C. Hitte, and F. Galibert. 2010. 'A radiation hybrid map of the European sea bass (Dicentrarchus labrax) based on 1581 markers: Synteny analysis with model fish genomes', Genomics, 96: 228-38.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Huang, T. H., J. F. Hejtmancik, A. Edwards, A. L. Pettigrew, C. A. Herrera, H. A. Hammond, C. T. Caskey, H. Y. Zoghbi, and D. H. Ledbetter. 1991. 'Linkage of the gene for an X-linked mental retardation disorder to a hypervariable (AGAT)n repeat motif within the human hypoxanthine phosphoribosyltransferase (HPRT) locus (Xq26)', American Journal of Human Genetics, 49: 1312-9.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jacob, H. J., K. Lindpaintner, S. E. Lincoln, K. Kusumi, R. K. Bunker, Y. P. Mao, D. Ganten, V. J. Dzau, and E. S. Lander. 1991. 'Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat', Cell, 67: 213-24.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kafkas, S., M. Khodaeiaminjan, M. Guney, and E. Kafkas. 2015. 'Identification of sex-linked SNP markers using RAD sequencing suggests ZW/ZZ sex determination in Pistacia vera L', BMC Genomics, 16: 98.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. 2012. 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', Bioinformatics, 28: 1647-9.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kleftogiannis, D., P. Kalnis, and V. B. Bajic. 2013. 'Comparing memory-efficient genome assemblers on stand-alone and cloud infrastructures', PLoS One, 8: e75505.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Konieczny, A., and F. M. Ausubel. 1993. 'A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers', Plant Journal, 4: 403-10.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lamesch, P., T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-

Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. 2012. 'The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools', Nucleic Acids Research, 40: D1202-10.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, G., and C. F. Quiros. 2001. 'Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in Brassica', Theoretical and Applied Genetics, 103: 455-61.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, M. W., H. J. Yu, X. L. Yi, J. Li, F. Y. Dai, and C. X. Hou. 2015. 'Marker-assisted selection in breeding silkworm strains with high tolerance to fluoride, scaleless wings, and high silk production', Genet Mol Res, 14: 11162-70.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, W., S. Royer, and R. T. Chetelat. 2010. 'Fine mapping of ui6.1, a gametophytic factor controlling pollen-side unilateral incompatibility in interspecific solanum hybrids', Genetics, 185: 1069-80.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Liu, J., J. Li, J. Qu, and S. Yan. 2015. 'Development of Genome-Wide Insertion and Deletion Polymorphism Markers from Next-Generation Sequencing Data in Rice', Rice (N Y), 8: 63.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lu, Y., X. Cui, R. Li, P. Huang, J. Zong, D. Yao, G. Li, D. Zhang, and Z. Yuan. 2015. 'Development of genome-wide insertion/deletion markers in rice based on graphic pipeline platform', J Integr Plant Biol.

'Mac App Store - Xcode'. 2015. Accessed June 24, 2015. https://itunes.apple.com/us/app/xcode/id497799835.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Marcais, G., and C. Kingsford. 2011. 'A fast, lock-free approach for efficient parallel counting of occurrences of k-mers', Bioinformatics, 27: 764-70.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Marieschi, M., A. Torelli, and R. Bruni. 2012. 'Quality control of saffron (Crocus sativus L.): development of SCAR markers for the detection of plant adulterants used as bulking agents', Journal of Agricultural and Food Chemistry, 60: 10998-1004.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Meyers, B. C., A. W. Dickerman, R. W. Michelmore, S. Sivaramakrishnan, B. W. Sobral, and N. D. Young. 1999. 'Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily', Plant Journal, 20: 317-32.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. 'Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers', Genome Research, 17: 240-8.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Nachimuthu, V. V., R. Muthurajan, S. Duraialaguraja, R. Sivakami, B. A. Pandian, G. Ponniah, K. Gunasekaran, M. Swaminathan, K. S. K, and R. Sabariappan. 2015. 'Analysis of Population Structure and Genetic Diversity in Rice Germplasm Using SSR Markers: An Initiative Towards Association Mapping of Agronomic Traits in Oryza Sativa', Rice (N Y), 8: 30.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Nakamura, Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and et al. 1987. 'Variable number of tandem repeat (VNTR) markers for human gene mapping', Science, 235: 1616-22.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Olson, M., L. Hood, C. Cantor, and D. Botstein. 1989. 'A common language for physical mapping of the human genome', Science, 245: 1434-5.

Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi, and T. Sekiya. 1989. 'Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms', Proc Natl Acad Sci U S A, 86: 2766-70.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Paran, I., and R. W. Michelmore. 1993. 'Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce', Theoretical and Applied Genetics, 85: 985-93.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pertuze, R. A., Y. Ji, and R. T. Chetelat. 2003. 'Transmission and recombination of homeologous Solanum sitiens chromosomes in tomato', Theoretical and Applied Genetics, 107: 1391-401.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pertuze, R. A., Y. F. Ji, and R. T. Chetelat. 2002. 'Comparative linkage map of the Solanum lycopersicoides and S-sitiens genomes and their differentiation from tomato', Genome, 45: 1003-12.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Peters, S. A., J. W. Bargsten, D. Szinay, J. van de Belt, R. G. Visser, Y. Bai, and H. de Jong. 2012. 'Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper', Plant Journal, 71: 602-14.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Potato Genome Sequencing, Consortium, X. Xu, S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, G. Zhang, S. Yang, R. Li, J. Wang, G. Orjeda, F. Guzman, M. Torres, R. Lozano, O. Ponce, D. Martinez, G. De la Cruz, S. K. Chakrabarti, V. U. Patil, K. G. Skryabin, B. B. Kuznetsov, N. V. Ravin, T. V. Kolganova, A. V. Beletsky, A. V. Mardanov, A. Di Genova, D. M. Bolser, D. M. Martin, G. Li, Y. Yang, H. Kuang, Q. Hu, X. Xiong, G. J. Bishop, B. Sagredo, N. Mejia, W. Zagorski, R. Gromadka, J. Gawor, P. Szczesny, S. Huang, Z. Zhang, C. Liang, J. He, Y. Li, Y. He, J. Xu, Y. Zhang, B. Xie, Y. Du, D. Qu, M. Bonierbale, M. Ghislain, R. Herrera Mdel, G. Giuliano, M. Pietrella, G. Perrotta, P. Facella, K. O'Brien, S. E. Feingold, L. E. Barreiro, G. A. Massa, L. Diambra, B. R. Whitty, B. Vaillancourt, H. Lin, A. N. Massa, M. Geoffroy, S. Lundback, D. DellaPenna, C. R. Buell, S. K. Sharma, D. F. Marshall, R. Waugh, G. J. Bryan, M. Destefanis, I. Nagy, D. Milbourne, S. J. Thomson, M. Fiers, J. M. Jacobs, K. L. Nielsen, M. Sonderkaer, M. Iovene, G. A. Torres, J. Jiang, R. E. Veilleux, C. W. Bachem, J. de Boer, T. Borm, B. Kloosterman, H. van Eck, E. Datema, Bt Hekkert, A. Goverse, R. C. van Ham, and R. G. Visser. 2011. 'Genome sequence and analysis of the tuber crop potato', Nature, 475: 189-95.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Pradhan, S. K., S. R. Barik, J. Sahoo, E. Pandit, D. K. Nayak, D. R. Pani, and A. Anandan. 2015. 'Comparison of Sub1 markers and their combinations for submergence tolerance and analysis of adaptation strategies of rice in rainfed lowland ecology', Comptes Rendus Biologies, 338: 650-9.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

R Core Team. 2014. "R: A language and environment for statistical computing." In. Vienna, Austria: R Foundation for Statistical Computing.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Rafalski, A. 2002. 'Applications of single nucleotide polymorphisms in crop genetics', Current Opinion in Plant Biology, 5: 94-100.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ramirez-Gonzalez, R. H., V. Segovia, N. Bird, P. Fenwick, S. Holdgate, S. Berry, P. Jack, M. Caccamo, and C. Uauy. 2015. 'RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat', Plant Biotechnology Journal, 13: 613-24.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ramirez-Gonzalez, R. H., C. Uauy, and M. Caccamo. 2015. 'PolyMarker: A fast polyploid primer design pipeline', Bioinformatics.

Reddy, T. B., A. D. Thomas, D. Stamatis, J. Bertsch, M. Isbandi, J. Jansson, J. Mallajosyula, I. Pagani, E. A. Lobos, and N. C. Kyrpides. 2015. 'The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification', Nucleic Acids Research, 43: D1099-106.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Robarts, D. W., and A. D. Wolfe. 2014. 'Sequence-related amplified polymorphism (SRAP) markers: A potential resource for studies in plant molecular biology(1.)', Appl Plant Sci, 2.

Salehi, N., B. Gottstein, and H. R. Haddadzadeh. 2015. 'Genetic diversity of bovine Neospora caninum determined by microsatellite markers', Parasitology International, 64: 357-61.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Salzberg, S. L., A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke. 2012. 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', Genome Research, 22: 557-67.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Sarmah, R., J. Sahu, B. Dehury, K. Sarma, S. Sahoo, M. Sahu, M. Barooah, P. Sen, and M. K. Modi. 2012. 'ESMP: A high-throughput computational pipeline for mining SSR markers from ESTs', Bioinformation, 8: 206-8.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Schuler, G. D. 1997. 'Sequence mapping by electronic PCR', Genome Research, 7: 541-50.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Serif Europe. 2015. 'Mac App Store - Xcode', Accessed June 24, 2015. https://itunes.apple.com/us/app/xcode/id497799835.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Shen, Y. J., H. Jiang, J. P. Jin, Z. B. Zhang, B. Xi, Y. Y. He, G. Wang, C. Wang, L. Qian, X. Li, Q. B. Yu, H. J. Liu, D. H. Chen, J. H. Gao, H. Huang, T. L. Shi, and Z. N. Yang. 2004. 'Development of genome-wide DNA polymorphism database for map-based cloning of rice genes', Plant Physiology, 135: 1198-205.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Sim, S. C., G. Durstewitz, J. Plieske, R. Wieseke, M. W. Ganal, A. Van Deynze, J. P. Hamilton, C. R. Buell, M. Causse, S. Wijeratne, and D. M. Francis. 2012. 'Development of a large SNP genotyping array and generation of high-density genetic maps in tomato', PLoS One, 7: e40563.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

'SolCAP Solanaceae Coordinated Agricultural Project'. 2015. Accessed 2-Sep-2015. http://solcap.msu.edu/index.shtml.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

TAIR. 2015. 'Polymorphism/Allele', TAIR. http://www.arabidopsis.org/servlets/Search?action=new_search&type=polyallele.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Taylor, J., and N. J. Provart. 2006. 'CapsID: a web-based tool for developing parsimonious sets of CAPS molecular markers for genotyping', BMC Genetics, 7: 27.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tomato Genome, Consortium. 2012. 'The tomato genome sequence provides insights into fleshy fruit evolution', Nature, 485: 635-41.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen. 2012. 'Primer3--new capabilities and interfaces', Nucleic Acids Research, 40: e115.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and et al. 1995. 'AFLP: a new technique for DNA fingerprinting', Nucleic Acids Research, 23: 4407-14.
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wall, L. 1987-2012. "Perl 5." In.

Weber, J. L., and P. E. May. 1989. 'Abundant Class af Human DNA Polymorphisms Which Can Be Typed Using the Polymerase Chain Reaction', American Journal of Human Genetics, 44: 388-96.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wenzl, P., J. Carling, D. Kudrna, D. Jaccoud, E. Huttner, A. Kleinhofs, and A. Kilian. 2004. 'Diversity Arrays Technology (DArT) for whole-genome profiling of barley', Proc Natl Acad Sci U S A, 101: 9915-20.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

West, M. A., H. van Leeuwen, A. Kozik, D. J. Kliebenstein, R. W. Doerge, D. A. St Clair, and R. W. Michelmore. 2006. 'High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis', Genome Research, 16: 787-95.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Williams, J.G.K., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey. 1990. 'DNA polymorphisms ampilified by arbitrary primers are useful as genetic markers', Nucleic Acids Research, 18: 6531-35.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wu, F., L. A. Mueller, D. Crouzillat, V. Petiard, and S. D. Tanksley. 2006. 'Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade', Genetics, 174: 1407-20.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zhou, G., Q. Zhang, C. Tan, X. Q. Zhang, and C. Li. 2015. 'Development of genome-wide InDel markers and their integration with SSR, DArT and SNP markers in single barley map', BMC Genomics, 16: 804.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Zietkiewicz, E., A. Rafalski, and D. Labuda. 1994. 'Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification', Genomics, 20: 176-83.

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title