Genome **Biology**

Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*

Michael C Schatz[1†], Lyza G Maron[2†], Joshua C Stein[1†], Alejandro Hernandez Wences[1,3], James Gurtowski[1], Eric Biggers[1,4], Hayan Lee[1,5], Melissa Kramer[1], Eric Antoniou[1], Elena Ghiban[1], Mark H Wright[2], Jer-ming Chia[1], Doreen Ware[1,6], Susan R McCouch[2*] and W Richard McCombie[1*]

Abstract

Background: The use of high throughput genome-sequencing technologies has uncovered a large extent of structural variation in eukaryotic genomes that makes important contributions to genomic diversity and phenotypic variation. When the genomes of different strains of a given organism are compared, whole genome resequencing data are typically aligned to an established reference sequence. However, when the reference differs in significant structural ways from the individuals under study, the analysis is often incomplete or inaccurate.

Results: Here, we use rice as a model to demonstrate how improvements in sequencing and assembly technology allow rapid and inexpensive *de novo* assembly of next generation sequence data into high-quality assemblies that can be directly compared using whole genome alignment to provide an unbiased assessment. Using this approach, we are able to accurately assess the 'pan-genome' of three divergent rice varieties and document several megabases of each genome absent in the other two.

Conclusions: Many of the genome-specific loci are annotated to contain genes, reflecting the potential for new biological properties that would be missed by standard reference-mapping approaches. We further provide a detailed analysis of several loci associated with agriculturally important traits, including the *S5* hybrid sterility locus, the *Sub1* submergence tolerance locus, the *LRK* gene cluster associated with improved yield, and the *Pup1* cluster associated with phosphorus deficiency, illustrating the utility of our approach for biological discovery. All of the data and software are openly available to support further breeding and functional studies of rice and other species.

Background

Rice (*Oryza sativa*) provides 20% of the world's dietary energy supply and is the predominant staple food for 17 countries in Asia, 9 countries in North and South America and 8 countries in Africa. Within *O. sativa*, there are two major varietal groups, *Indica* and *Japonica*, that can be further subdivided into five major subpopulations: *indica* and *aus* share ancestry within the *Indica* varietal group, and *tropical japonica*, *temperate japonica* and *aromatic* (*Group V*) share ancestry within the *Japonica* varietal

group (Figure 1) [1-3]. The subpopulation structure of *O. sativa* is deep and ancient, with estimates of divergence showing average pairwise Fst values of 0.375 to 0.45 [1-3], compared with Fst values of 0.25 for dogs [4], around 0.10 to 0.12 across human populations [5], or 0.08 to 0.09 for heterotic groups in maize [6].

The time since divergence of the ancestral *Indica* and *Japonica* gene pools is estimated at 0.44 million years, based on sequence comparisons between cv Nipponbare (*Japonica*) and cv 93-11 (*Indica*) [8]. This time estimate pre-dates the domestication of *O. sativa* by several hundred thousand years, suggesting that rice cultivation proceeded from multiple, pre-differentiated ancestral pools [1,9-13]. This is consistent with genome-wide estimates of divergence based on gene content [14], transcript levels [15], single nucleotide polymorphisms (SNPs) [3,16], and
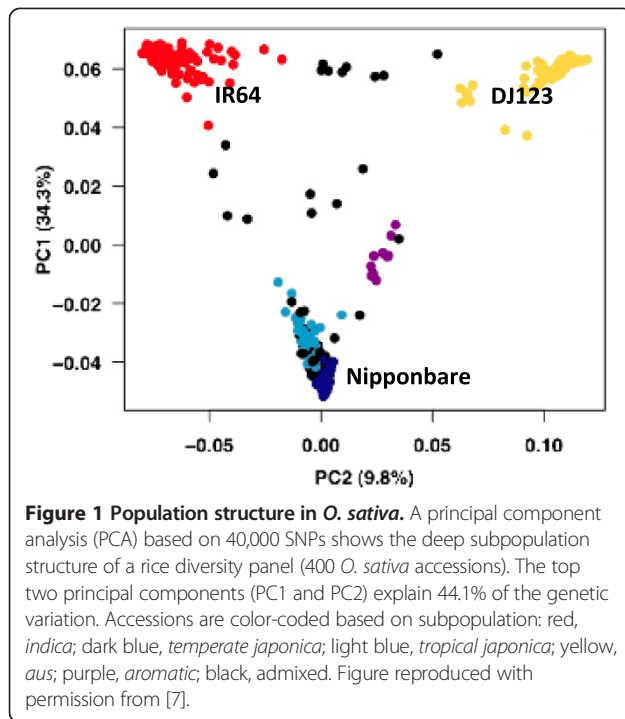
* Correspondence: srm4@cornell.edu; mccombie@cshl.edu
†Equal contributors
2Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853, USA
1Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
Full list of author information is available at the end of the article

Figure 1 Population structure in *O. sativa*. A principal component analysis (PCA) based on 40,000 SNPs shows the deep subpopulation structure of a rice diversity panel (400 *O. sativa* accessions). The top two principal components (PC1 and PC2) explain 44.1% of the genetic variation. Accessions are color-coded based on subpopulation: red, *indica*; dark blue, *temperate japonica*; light blue, *tropical japonica*; yellow, *aus*; purple, *aromatic*; black, admixed. Figure reproduced with permission from [7].

transposable elements [17]. This is also consistent with evidence from the cloning of dozens of genes underlying diverse quantitative trait loci (QTLs) [2,10,18-21]. Despite ongoing debate about the precise moment and location of the first domestication 'event' in rice, these studies all demonstrate that natural variation in the rice genome is deeply partitioned and that divergent haplotypes can be readily associated with major varietal groups and subpopulations. The course of domestication, as rice transitioned from its ancestral state as a tropical, outcrossing, aquatic, perennial species to a predominantly inbreeding, annual species adapted to a wide range of ecologies, was punctuated by persistent episodes of intermating among the different subpopulations. This resulted in both natural and human-directed gene flow between the different gene pools, but the essential differentiation that distinguishes the *Indica* and *Japonica* genomes was maintained and reinforced over time as a result of numerous partial sterility barriers scattered throughout the genome [22-25].

A better understanding of the nature and extent of genome variation within the *Oryza* clade is critical for both practical and scientific reasons. While the OMAP project [26] is focused on documenting structural variation across 21 wild species of *Oryza*, relatively little effort has been made to explore the nature of structural variation within and between subpopulations of *O. sativa*. The high quality, bacterial artificial chromosome (BAC)-by-BAC sequence of the *temperate japonica* rice variety Nipponbare, generated by the International Rice Genome Sequencing Program (IRGSP) [27], and the shotgun

assembly of an *indica* rice genome, cv 93-11, by Chinese scientists in 2005 [28,29] have served as 'reference genomes' for the rice research community. The availability of these reference genomes helped catalyze and unify rice research efforts for over a decade, and continue to serve as the backbone for re-sequencing efforts today [2,30-33].

Recently, the resequencing of hundreds of wild and cultivated rice genomes using next generation sequencing (NGS) and various complexity-reduction and genotype-by-sequencing strategies have enriched the pool of sequence information available for rice [30,34,35]. However, the vast majority of resequenced genomes are aligned to and compared with the Nipponbare reference rather than being assembled *de novo*, including in our own previous work [35] and in the current 3,000 rice genomes project [36]. This introduces a potential bias due to significant differences in genome size [37,38] and structure [14,17,29,39] that characterize the different subpopulations and varieties of rice. Alignment to a single reference is particularly problematic when NGS data from *indica*, *aus* or divergent wild species genomes from the center of diversity of *Oryza* are aligned to the genetically and geographically divergent Nipponbare (*temperate japonica*) reference because of the potential for misalignment, and for elimination of critical sequences that cannot be aligned with confidence.

The type and distribution of structural variation that distinguishes one rice genome from another, both within and between the five subpopulations of *O. sativa*, remain largely unknown. Yet it is essential to understanding the genetic basis of heterosis, as well as to identify genes underlying many of the most significant phenotypic differences that are critical to global food security, including a plant's ability to grow in stressful environments afflicted by drought, submergence, low phosphorus and/or disease. The only practical way to fully understand the genomic diversity of rice is to carry out whole genome shotgun sequencing and *de novo* assembly. This has been problematic until recently due to the difficulties in assembling the short reads initially provided by NGS. However, recent advances in NGS chemistry and in computational approaches to sequence assembly have significantly improved the power and reliability of *de novo* assembly of NGS data.

In this study we use these advances to *de novo* assemble three divergent rice genomes representing the *indica* (IR64), *aus* (DJ123) and *temperate japonica* (Nipponbare) subpopulations and to determine the extent and distribution of structural variation among them. These varieties were chosen for both biological interest and to facilitate evaluation of assemblies. On the biological side, different subpopulations of rice are adapted to different ecologies and geographies, and harbor different alleles and traits of interest for plant improvement [3,19,20,40-43]. The *aus* subpopulation is of particular interest because it is the

source of important alleles conferring disease resistance [44], tolerance to submergence [33], deep water [45], low-phosphorus soils [41], and drought [46]. *Indica* rice harbors the greatest amount of genetic variation [1,30] and accounts for the largest contribution to rice production globally. Our choice to sequence Nipponbare was due to the fact that it provided a high quality BAC-by-BAC sequence assembly [27] that served as a solid benchmark for assessing the quality of our three NGS assemblies and provided a context for understanding the impact of varying data sets and parameters used in the assemblies.

Results and discussion

De novo genome assemblies and functional annotation

The three rice varieties were assembled using the ALLPATHS-LG whole genome assembler [47] using approximately 50× coverage of a 180 bp fragment library, approximately 30× coverage of a 2 kbp jumping library, and approximately 30× coverage of a 5 kbp jumping library (see Materials and methods). We selected this assembler based on its performance with these data compared with other assemblers and its high ranking in the Assemblathon I and II and GAGE evaluations [48-50]. The three assemblies were named Os-Nipponbare-Draft-CSHL-1.0, Os-IR64-Draft-CSHL-1.0, and Os-DJ123-Draft-CSHL-1.0, following nomenclature proposed by [51].

All three of our assemblies had excellent results: approximately 90% of each of the genomes were assembled into scaffolds at least 1 kbp long, with scaffold N50 sizes ranging from 213 kbp to 323 kbp, and contig N50 sizes ranging from 21.9 kbp to 25.5 kbp (Table 1). It is notable that an earlier assembly of the Nipponbare genome prior to sequencing the 5 kbp jumping library achieved a

similar contig N50 size (21.2 kbp versus 21.9 kbp), but a substantially smaller scaffold N50 size (99 kbp versus 213 kbp) (also see Materials and methods). Improved scaffold sizes from including the larger library were expected, although the magnitude depends on the specific genome characteristics. Since the scaffolds were more than twice as large for Nipponbare with the larger library, this prompted us to sequence the 5 kbp jumping library for all three genomes to maximize our ability to identify genes and other features, as well as to structurally compare the genomes.

The assemblies were repeat-masked and annotated for protein-coding genes using the MAKER-P automated pipeline [52], combining both evidence-based and *ab initio* methods (Table S1 in Additional file 1). In addition to EST and full-length cDNA, we included as evidence the two published annotations of Nipponbare [51], and the published annotations of strains 93-11 and PA64s [28], thereby maximizing consistency and reducing bias of annotation across the three assemblies. Putative transposon-encoded genes were screened following analysis of InterPro domains (see Materials and methods), which flagged approximately 1% of initial gene calls in each of the three assemblies. Summary statistics for remaining genes are provided in Table 1 and in Table S2 in Additional file 1. Gene counts ranged from 37,758 (IR64) to 39,083 (Nipponbare), similar to the numbers reported by the Michigan State University (MSU) Rice Genome Annotation Project and Rice Annotation Project for the Os-Nipponbare-Reference-IRGSP-1.0 (39,102 and 35,681 respectively) [51]. Overall statistics for structural features, such as exons, introns, and coding regions were highly consistent between the three assemblies and with published annotations. For instance, average translated protein lengths compared across MSU, Rice Annotation Project, and the three *de novo* assemblies ranged from 280 to 288 amino acids (median values: 268 to 291 amino acids), suggesting that contiguity of the *de novo* assemblies did not limit ability to identify protein-coding genes. For each assembly, 61 to 62% of annotated loci possessed one or more InterPro domains and 77% showed homology to plant NCBI RefSeq genes.

Whole genome comparison to Nipponbare reference genomes

We evaluated the agreement between our *de novo* assemblies to the Nipponbare reference sequences using the GAGE assembly evaluation algorithm [50]. As expected, the *de novo* Nipponbare assembly very closely matches the reference Nipponbare sequence, with a 99.94% average identity and only 0.31% of the assembly not aligning to the reference (Tables 2, 3 and 4). Even at this very high agreement, there are several tens of thousands of small variations, and several hundred larger

Table 1 Assembly and annotation statistics of the three *de novo* assemblies used in this study

	Nipponbare	IR64	DJ123
Total span	355.6 Mbp	345.2 Mbp	345.9 Mbp
Total bases	318.2 Mbp	316.3 Mbp	321.2 Mbp
Genome coverage[a] (span/bases)	91.2%/81.8%	88.5%/81.3%	88.6%/82.5%
Number of scaffolds	4,110	2,919	2,819
N50 scaffold span	213 kbp	293 kbp	323 kbp
Max scaffold span	1.37 Mbp	2.85 Mbp	2.38 Mbp
Number of contigs (>1,000 bp)	27,486	26,160	23,902
N50 contig size	21.9kbp	22.2 kbp	25.5 kbp
Max contig size	133 kbp	160 kbp	252 kbp
Total genes	39,083	37,758	37,812
Median gene length	2,224	2,275	2,285
Mean exons per transcript	4.8	4.8	4.9

[a]Assumes the total genome size is 389 Mbp, as according to IRGSP [27].

Table 2 Comparison of the three *de novo* assemblies to the Nipponbare reference (IRGSP-1.0)

	Unaligned reference bases	Unaligned assembly bases	Average ID	SNPs and small indels	Indels > 5 bp	Inversions	Relocations	Translocations
Nipponbare	3.14%	0.31%	99.94%	57,459	3,445	131	252	617
IR64	11.08%	8.94%	98.91%	2,917,780	80,631	1,004	1,721	7,060
DJ123	9.85%	8.55%	98.93%	2,933,257	80,346	1,007	1,615	6,683

variations. These variations are a combination of true variations from our sample relative to the reference genome, of which we expect there to be few, and errors from ALLPATHS-LG when used with these libraries and coverage levels. Consequently, considering that the assembly has a 99.94% overall similarity, the upper-bound on the error rate of sequencing and assembling with ALLPATHS-LG is at most 0.06%.

The portions of the reference genome without any alignments from our Nipponbare assembly are scattered throughout the genome in 57,821 segments averaging 203 bp long. However, of this, only 301,525 bp are annotated to be within the coding sequence (CDS; 0.72% of the total CDS), and another 12,344 bp are annotated to be within non-coding exons. We further evaluated the unaligned regions by computing their read k-mer coverage from a sample of 400 million unassembled Nipponbare reads, and found the mean k-mer coverage of these regions exceeds 12,000×, while the mode k-mer coverage of the set is less than 100× (Figure S1 in Additional file 1). A full two-thirds (38,373/57,821) of these regions exceed 1,000× k-mer coverage, more than 10 times higher than unique segments of the genome. This implies the unassembled/unaligned regions are highly enriched for high copy repeats too complex to be assembled. In contrast, the genic regions are very well represented, suggesting it would be possible for a detailed analysis of the 'gene space' of the accessions from these assemblies.

Tables 2, 3 and 4 summarize the alignments of the three *de novo* assemblies relative to the reference IRGSP-1.0 Nipponbare assembly. As expected, the IR64 and DJ123 assemblies show noticeably lower overall identity, and have considerably more unaligned bases. The average k-mer coverage of the unaligned bases indicates most regions are unassembled high copy repeats, although there

Table 3 Summary of unaligned reference regions

	Total bp	Regions	Average size	Maximum size	Mean k-mer coverage
Nipponbare	11,750,969	57,821	203 ± 350	12,773	12,210×
IR64	41,639,095	133,536	311 ± 921	27,087	7,938×
DJ123	37,010,281	122,589	302 ± 931	26,236	6,121×

Mean k-mer coverage was evaluated by counting the k-mers in a sample of 400 M unassembled reads in each of the three genomes, and evaluating those counts along the reference sequence.

are 11.8 Mbp and 12.3 Mbp unaligned reference bases in IR64 and DJ123, respectively, that are not repetitive based on the k-mer analysis. This suggests there may be megabases of sequence specific to each of the three genomes.

Whole genome comparison to *indica* reference genomes

Using the same methods used for comparing to the reference Nipponbare genome, we also evaluated the three genomes relative to the reference *indica* genome (cv 93–11) [27] (Tables 5, 6 and 7). The agreement between the *de novo* IR64 assembly and the reference *indica* sequence is appreciably less than the Nipponbare-Nipponbare alignment; 4.31% of the IR64 assembly does not align to the 93–11 reference and the aligned regions have only 99.52% identity between these two *indica* varieties. Since the assemblies and alignments were computed with the same sample preparation and analysis algorithms, this suggests there are more true biological variations between IR64 and 93–11 (as would be expected from two different varieties), and/or that the 93–11 reference assembly is not as complete nor as accurate as the reference Nipponbare assembly. The later explanation is quite likely to be a contributing factor, given the fact that the 93–11 genome represents a whole genome shotgun assembly, while the Nipponbare genome utilized a combination of BACs and whole genome shotgun sequencing. For example, the 93–11 assembly has 14.1 million unresolved ('N') bases, while the Nipponbare reference has only 118,200. As seen with Nipponbare, most of the unassembled/unaligned bases between the 93–11 reference and our assemblies are repetitive with mean k-mer coverage over 14,000×. A quarter of the unaligned references bases (7.75 Mbp/31 Mbp) are non-repetitive from the k-mer analysis, while less than 900 kbp of the unaligned reference Nipponbare genome are not repetitive. This underscores the fact that there are substantially more true biological differences between IR64 and the reference 93–11 *indica* assembly than our Nipponbare sample and reference.

Finally, the comparison between the IR64 and DJ123 assemblies shows that they differ from each other nearly as much as either one differs from the Nipponbare reference sequence. These results suggest that the *aus* genome harbors a greater amount of novel variation than previously recognized. It also highlights the value of taking an unbiased, *de novo* assembly approach when evaluating

Table 4 Summary of unaligned bases by reference annotation

	Non-coding exon	5' UTR	3' UTR	mRNA	Coding sequence	Repetitive bp (>100× k-mer coverage)
Nipponbare	12,344	75,443	80,518	984,489	301,525	10,863,120
IR64	129,428	902,520	638,856	6,827,679	1,706,454	29,685,919
DJ123	114,735	865,470	601,928	6,418,865	1,519,686	24,505,588

Repeats were evaluated by analyzing their k-mer coverage using the same method as above.

genomic variation among varieties and subpopulations to capture genome-specific variations.

Pan-genome analysis

We next evaluated the 'pan-genome' of the three *de novo* assemblies to identify sequences that were conserved across the genomes as well as sequences specific to just one genome (see Materials and methods). Using the whole genome alignment information, we classified each base of each genome as being specific to that genome (unaligned to either other genome), or shared by one or both genomes. The majority of the assembled sequences (approximately 302 Mbp per genome) and exonic sequences (approximately 55.5 Mbp per genome), were shared among the three genomes, although 4.8 Mbp to 8.2 Mbp (423 kbp to 930 kbp exonic) were found to be genome-specific (Figure 2A). Since a gene sequence may be partially shared or partially genome-specific, we assigned each gene to the sector on the Venn diagram for which the majority of the exonic bases were assigned over all transcripts associated with each gene. For example, if 90% of a gene is shared among all three genomes, but 10% is genome-specific, we would assign it to the center (fully shared) sector under the majority rule. This will not necessarily characterize changes in gene function if critical protein domains are shared/unshared, but highlights the major trends between the lineages and discovers 297 to 786 genome-specific loci.

Using the same k-mer analysis techniques we applied for the reference analysis, we further classified the genome-specific bases as being unique or repetitive, using a threshold of 100× average k-mer coverage to classify unique sequences. From this, we identified only 1.2 Mbp to 1.5 Mbp of non-repetitive sequence specific to each genome, meaning that most of the genome-specific bases were actually repetitive (Table 8). Since repetitive sequences are also the most likely to be unassembled, as observed in our comparison to the reference genomes, we further

examined the genome-specific exonic bases and refined our initial estimates to 555 kbp to 760 kbp of non-repetitive, genome-specific sequences intersecting annotated genes by at least 100 bp (Table 9). Note these segments may include flanking promoter and other regulatory regions in addition to the exons themselves. From this catalog, we selected 10 of the largest regions in each of the genomes for PCR validation, and were able to confirm the computational analysis with 100% success (Figure 3; Table S4a-c in Additional file 1).

For Nipponbare and IR64, we determined the positions of the non-repetitive segments along the different reference chromosomes, and found the segments were broadly distributed. For Nipponbare, we could localize 2,208 of the genome-specific regions, and found that one region occurred, on average, every 162 ± 362 kbp, following an approximately exponential distribution (data not shown). For IR64, we could localize 1,074 of the genome-specific regions, and found one region occurred, on average, every 338 ± 752 kbp, also from an approximately exponential distribution. The distributions suggest that the genome-specific bases are not highly localized, as an exponential distribution in spacing can occur if there is a uniform probability distribution of a site occurring at any position at random.

Genome-specific loci, as well as those shared between two genomes but not the third, exhibited shorter CDSs and greater novelty compared with genes shared among all three genomes (Figure 2B). For example, loci common to all genomes had a median coding length of 888 bp compared with median values ranging from 483 to 654 bp for the genome-specific gene sets. Likewise, the core, fully shared set of genes averaged 4.9 exons per transcript compared with a range of 2.9 to 3.0 genes per transcript amongst genome-specific genes. A smaller fraction of genome-specific loci contained InterPro domains compared with the core set (40% versus 63%), and fewer showed homology to plant RefSeq proteins (57%

Table 5 Comparison of the three *de novo* assemblies to the *Indica* reference (93-11 from [28])

	Unaligned reference bases	Unaligned assembly bases	Average ID	SNPs and smal indels	Indels > 5 kbp	Inversions	Relocations	Translocations
Nipponbare	16.95%	8.90%	98.91%	2,813,076	75,944	1,162	11,627	11,030
IR64	13.29%	4.31%	99.52%	1,228,732	35,762	644	6,985	7,101
DJ123	14.37%	6.78%	99.16%	2,264,541	62,191	974	8,582	10,170

Table 6 Summary of unaligned reference regions relative to the *Indica* reference

	Total bp	Regions	Average size	Maximum size	Mean k-mer coverage
Nipponbare	46,101,370	176,504	261 ± 796	53,066	6,945
IR64	31,006,053	139,336	264 ± 718	54,056	14,215
DJ123	49,562,877	152,378	325 ± 845	54,307	9,605

versus 79%). However, artifacts of inaccurate annotation may contribute to this trend [53], so we investigated if these differences were negatively influenced by assembly quality, especially if genome-specific genes tended to terminate in scaffold gaps more frequently than core genes. We observe a modest effect, and genes shared by all three strains have a median distance of 12 to 14 kbp (5′ and 3′ flanking distances), whereas genome-specific genes have a median distance of 8 to 10 kbp (Table S7 in Additional file 1). Only 7 to 12% of the genes have a flanking distance of less than 1 kbp, suggesting this is not a major factor in our analysis. Indeed, our results are similar to studies in yeast, *Drosophila*, and vertebrates that have found that novel and recently evolved genes tend to encode smaller proteins than conserved or ancient genes [53-55].

To characterize potential function of genome-specific genes we further examined genes with annotated Inter-Pro domains. Notably, genes with domains related to disease resistance were the most prevalent type among genome-specific genes. For example, 12% of genes specific to IR64 possessed the NB-ARC motif (IPR002182), the central nucleotide-binding domain of plant R-genes. This domain, and others associated with R-genes, also prevailed among the DJ123-specific and Nipponbare-specific gene sets, accounting for 9% and 5% of genes, respectively. In contrast, only 0.35% of genes shared among all three genomes encode the NB-ARC domain. Genes shared between just two genomes showed intermediate frequencies of disease resistance genes (1.5 to 2.5%). Similar distributions were seen for genes classified with the Gene Ontology (GO) term 'defense response' (GO:0006952). These results are consistent with Ding *et al.* [14], who showed high levels of 'genome asymmetry' among R genes when comparing the Nipponbare

Table 7 Summary of unaligned bases by reference annotation relative to the *Indica* reference

	mRNA	CDS	Repetitive bp (>100× kmer coverage)
Nipponbare	9,768,022	4,809,510	33,934,310
IR64	7,629,670	3,993,153	23,255,924
DJ123	5,999,753	1,947,112	23,805,398

Note that only CDS and mRNA annotations are available for the reference *Indica* assembly.

Figure 2 Venn diagrams of the shared sequence content between Nipponbare (*temperate japonica*), IR64 (*indica*) and DJ123 (*aus*). (A) overall sequence content. In each sector, the top number is the total number of base pairs, the middle number is the number of exonic bases, and the bottom is the gene count. If a gene is partially shared, it is assigned to the sector with the most exonic bases. **(B)** Genic content. In each sector, the top number is the median CDS length, the middle number is the average number of exons per gene, and the bottom is the percentage InterPro/homology.

and 93-11 reference assemblies. A large diversity of other protein domain classes, such as those associated with receptor and non-receptor protein kinases, transcription factors, metabolic enzymes, proteases, and transporters, were also found in the genome-specific gene sets. A complete listing of putative strain-specific genes, their InterPro domains, GO terms, and summary of homology search results are provided in Additional file 2. We anticipate these findings will greatly enhance the ongoing 3,000 rice genomes project [36] and other resequencing projects that

Table 8 Genome-specific non-repetitive bases

Genome	Genome-specific	Regions	Mean ± standard deviation	Maximum size
Nipponbare	1,574,801 bp	2,250	699 ± 904	8,463
IR64	1,336,650 bp	1,702	785 ± 1066	10,400
DJ123	1,263,681 bp	1,569	805 ± 1,058	8,960

Identified sequences must be at least 100 bp long with no alignments to the other genomes, average between 10× and 100× k-mer coverage in that genome, and average below 10× k-mer coverage using the reads from the other two genomes.

had previously focused on single nucleotide variations relative to the Nipponbare reference.

Detailed regions

We chose four agronomically relevant regions of the rice genome that were previously reported to harbor differences among the three varieties or subpopulations to illustrate the utility of these high quality whole genome assemblies for understanding the variation in genome structure underlying salient phenotypic variants.

S5 hybrid sterility locus

S5 is a major locus for hybrid sterility in rice that affects embryo sac fertility. Genetic analysis of the *S5* locus documented three alleles: an *indica* (*S5*-i), a *japonica* (*S5*-j), and a neutral allele (*S5*-n) [23,56]. Hybrids of genotype *S5*-i/*S5*-j are mostly sterile, whereas hybrids of genotypes consisting of *S5*-n with either *S5*-i or *S5*-j are mostly fertile. The *S5* locus contains three tightly linked genes that work together in a 'killer-protector'-type system [57,58]. During female sporogenesis, ORF5+ (killer) and ORF4+ (partner) cause endoplasmic reticulum stress. ORF3+ prevents endoplasmic reticulum stress and allows the production of normal gametes, whereas the ORF3- allele cannot prevent it, resulting in embryo sac abortion. The ORF3- allele has a 13-bp deletion; the ORF4- allele carries an 11-bp deletion that causes a premature stop codon [58]. The ORF5 *indica* (ORF5+) and *japonica* (ORF5-) alleles differ by only two nucleotides, whereas the wide compatibility allele *S5*-n (ORF5n) has a large deletion in the amino terminus of the predicted protein, rendering it presumably non-functional [57]. The typical *indica* haplotype is ORF3+/ORF4-/ORF5+, while the typical *japonica* haplotype is ORF3-/ORF4+/ORF5-.

Table 9 Genome-specific non-repetitive gene sequences

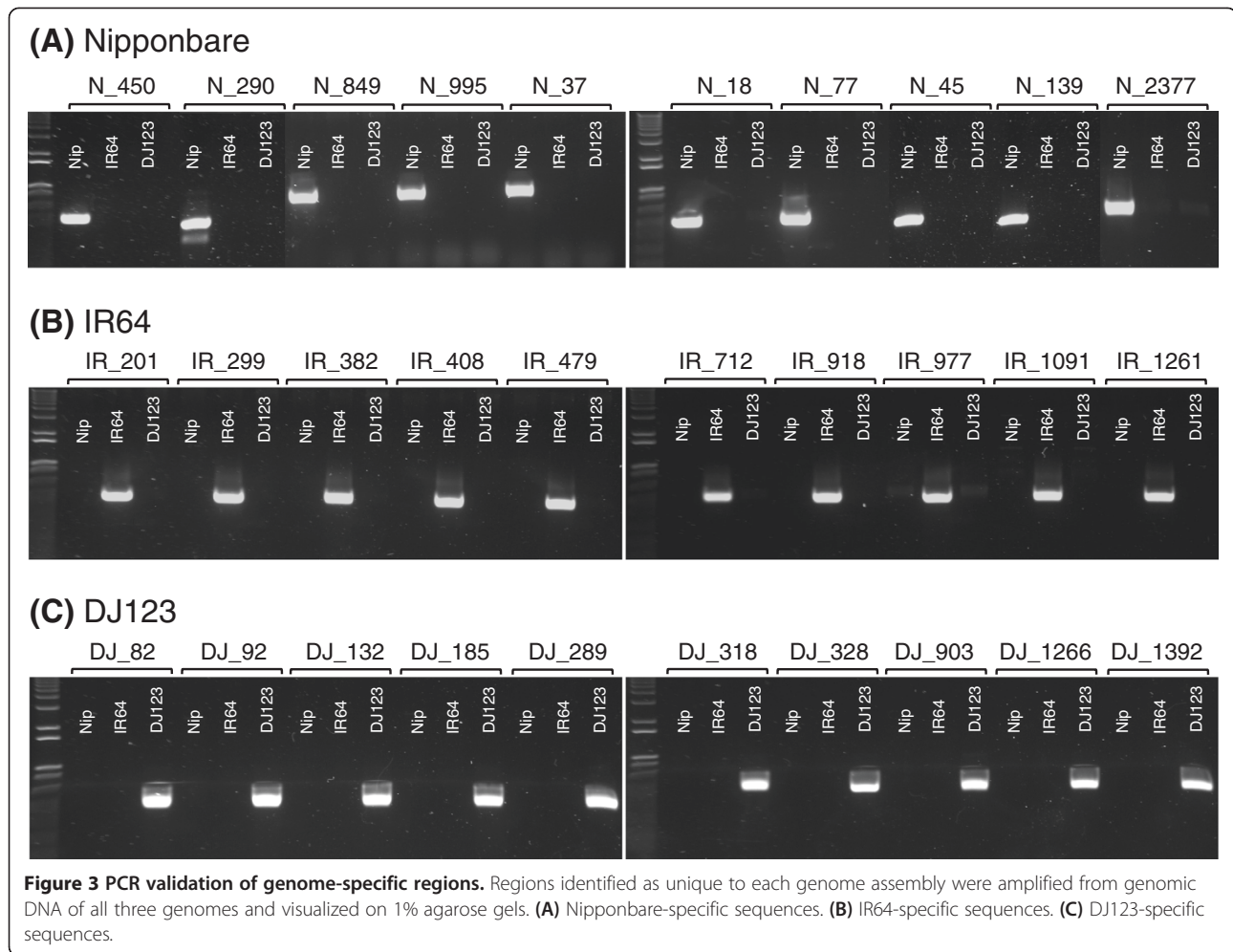| Genome	Genome-specific	Regions	Mean ± standard deviation	Maximum size
Nipponbare	760,064 bp	779	975 ± 1,197	8,463
IR64	637,470 bp	583	1,093 ± 1,430	10,400
DJ123	555,507 bp	492	1,129 ± 1,409	8,960

Regions identified to be specific to a given accession using the criterion used in Table 8 and that intersect an annotated gene region by at least 100 bp.

In each of the three *de novo* assemblies reported here the *S5* locus containing the three genes lies within a single scaffold and haplotypes can be easily identified (Table S5 and Figure S2 in Additional file 1). The identity of the *ORF5* alleles in Nipponbare, IR64 and DJ123 were also confirmed by Sanger sequencing from genomic DNA, and perfectly confirm the assembly results (Figures S3 and S4 in Additional file 1). The Nipponbare assembly is in agreement with the Nipponbare IRGSP-1.0 reference sequence for the region and shows that it carries the typical *japonica* haplotype ORF3-/ORF4+/ORF5-. The IR64 assembly shows that this accession carries the typical *indica* haplotype ORF3+/ORF4-/ORF5+. In the case of DJ123, our *de novo* assembly revealed that this *aus* accession carries the 136-bp deletion characteristic of the neutral allele, *ORF5*n. However, the DJ123 *ORF5*n allele is novel, as it differs from the reported *ORF5*n allele by two SNPs and one 10-bp deletion within the coding region of the gene (also confirmed by Sanger sequencing). The DJ123 haplotype for the locus is ORF3-/ORF4-/ORF5n, a haplotype previously identified by Yang *et al.* [58] in four accessions from Bangladesh. Although the accessions bearing this haplotype were referred to as *indica* in this study, they almost certainly belonged to the *aus* subpopulation.

Sub1 locus

The *Submergence 1* (*Sub1*) locus on chromosome 9 is a major QTL determining submergence tolerance in rice [33]. The *Sub1* locus is a cluster of three genes encoding putative ethylene response factors. *Sub1B* and *Sub1C* are present in all rice accessions tested to date, while *Sub1A* may be present or absent. Originally identified in the *aus* accession FR13A, *Sub1A* appears to be found only within the *Indica* varietal group [33]. *Sub1A* has two alleles: *Sub1A-1* is found in submergence-tolerant varieties, while *Sub1A-2* is found in intolerant varieties. A haplotype survey in *O. sativa* varieties also identified nine *Sub1B* and seven *Sub1C* alleles [33].

In the IR64 and DJ123 *de novo* assemblies reported here the *Sub1* locus lies within a single scaffold and haplotypes can be easily identified (Table S5 in Additional file 1). In the IR64 assembly the *Sub1A* gene is present as the *Sub1A-2* allele, previously identified in submergence-intolerant accessions including IR64 [33]. For the *Sub1B* and *C* genes, IR64 carries the alleles *Sub1B-1* and *Sub1C-3*, as reported [33]. *Sub1A* is absent from the DJ123 assembly, suggesting that this *aus* variety is not submergence tolerant. DJ123 carries a novel *Sub1B* allele (*Sub1B-10*), and the previously identified *Sub1C-6* allele. In the Nipponbare assembly, *Sub1B* and *Sub1C* lie within a single scaffold, and the alleles identified are in agreement with published results [33]. Nipponbare is not submergence tolerant and the *Sub1A* gene is absent in Nipponbare

Figure 3 PCR validation of genome-specific regions. Regions identified as unique to each genome assembly were amplified from genomic DNA of all three genomes and visualized on 1% agarose gels. **(A)** Nipponbare-specific sequences. **(B)** IR64-specific sequences. **(C)** DJ123-specific sequences.

according to previous reports. Our *de novo* assembly is unresolved in the region that corresponds to the *Sub1A* gene, but a k-mer analysis using the methods and data applied above clearly shows a lack of coverage in the DJ123 and Nipponbare sequencing reads across the locus except for high copy repeats dispersed in the sequence (Figure 4, top and bottom). Conversely, the k-mer coverage of the IR64 assembly is uniformly at the single-copy coverage level (approximately 100×), except for a small number of localized gaps in coverage, corresponding to SNPs in IR64 relative to the reference *Sub1A* sequence, and the high frequency repeats (Figure 4, middle). In contrast, the k-mer coverage across *Sub1B* and *Sub1C* is consistently approximately 100×, except for isolated sharp gaps corresponding to variations relative to the reference sequences (Figures S5 and S6 in Additional file 1).

LRK gene cluster
Fine-mapping of a yield-improving QTL on rice chromosome 2 identified a cluster of leucine-rich repeat receptor kinase genes [59], consisting of seven or eight intronless gene copies contained within a 40 to 50 kb genomic region. The QTL, originally introgressed from a wild rice accession (Dongxiang), was shown to increase grain yield of the recurrent parent Guichao2 (*indica*) by about 25%. The *LRK* locus in Dongxiang carries an extra gene, *LRK1*, absent from Guichao2. A survey of haplotype divergence in 13 rice accessions showed that *LRK1* is absent in only three *indica* accessions, suggesting that these haplotypes may have originated via gene loss.

In each of the three *de novo* assemblies reported here the *LRK* locus lies within a single scaffold and haplotypes can be easily identified (Table S5 in Additional file 1). The Nipponbare assembly is in agreement with the reference sequence, with the exception of regions that the *de novo* assembly was not able to resolve because of high copy repeats (Figure S7 in Additional file 1). *LRK1* is absent in the IR64 assembly as evident in the k-mer plot (Figure S8 in Additional file 1), indicating that IR64 carries the seven-gene haplotype identified in other *indica* accessions [59]. According to our assembly and the corresponding k-mer analysis, the *aus* accession DJ123 carries *LRK1*. Based on
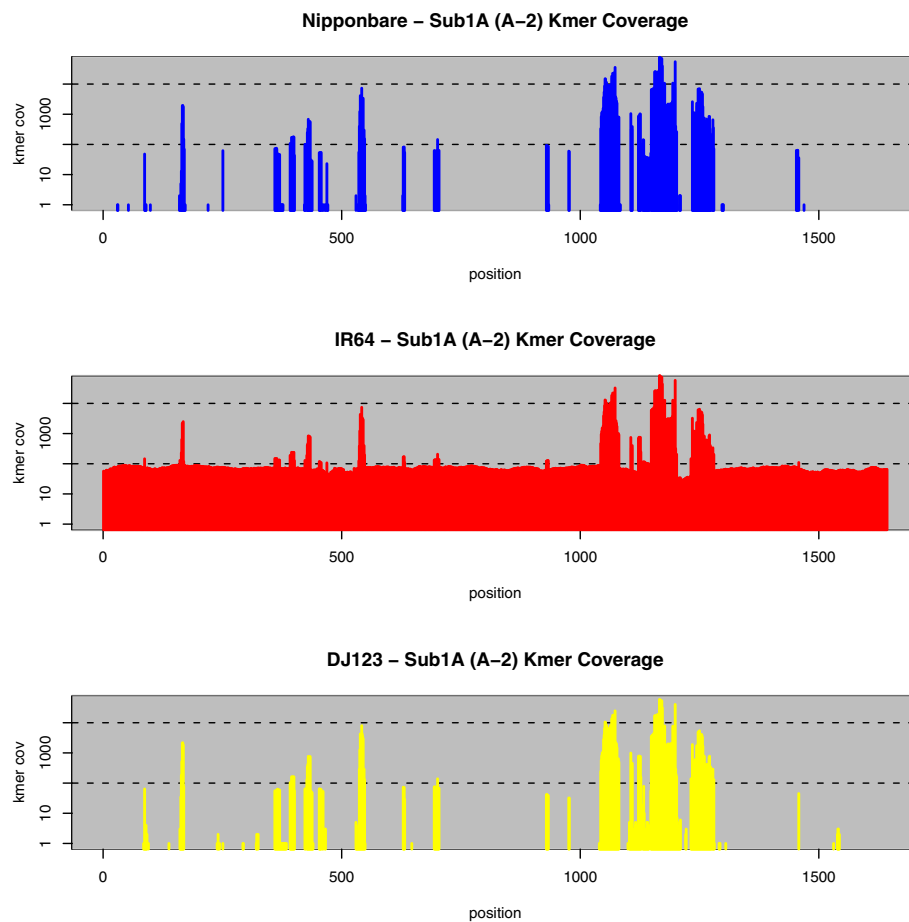
Figure 4 K-mer coverage in the three assemblies across the *Sub1A* gene. In each panel, the k-mer coverage of the sequence reads of the three respective genomes are plotted according to the sequence of the Sub1A A-2 allele. Only IR64 has consistent coverage across the gene, while the other two genomes have sparse coverage of a few repetitive k-mer sequences. For clarity, the k-mer coverage range 1× to 50,000× (log scale) is displayed in all the plots.

sequence variation on the 5′ upstream region of *LRK4* and *LRK6*, we can predict that the DJ123 haplotype for the LRK gene cluster is closest to the haplotypes identified in *indica* accessions in which *LRK1* is present (haplotypes A, B and C in Figure 3 of [59]).

Pup1 region

Phosphorus uptake1 (*Pup1*) is a major rice QTL associated with tolerance to phosphorus deficiency in soils [60,61]. The *Pup1* locus is a large, 90 kb region originally identified in Kasalath, an *aus* variety that is tolerant to phosphorus deficiency, but is absent in phosphorus starvation-intolerant varieties, including Nipponbare [62]. A gene encoding a protein kinase, *Pstol1*, located within the 90 kb indel, is responsible for the P-uptake efficiency phenotype [41].

Of the three *de novo* assemblies reported here, the 90 kb indel is absent from both Nipponbare and IR64, but a large portion of it, including the *Pstol1* gene, is present in the *aus* variety DJ123 (Figure 5; Table S5 in Additional file 1). Although it is at least partially present, the region of the 90 kb indel described in Kasalath could not be fully resolved in our DJ123 assembly. This suggests that the 90 kb indel may be truncated and/or rearranged in some *aus* varieties. Interestingly, as shown in Figure 5, the Kasalath reference sequence contains unresolved gaps flanking regions of very high k-mer coverage; therefore, longer reads may be necessary to assemble this region with confidence. The *Pstol1* gene sequence is complete in DJ123, and shows six SNPs relative to the Kasalath sequence (also apparent as abrupt drops in coverage in the k-mer coverage plot; Figures S9 and S10 in Additional file 1). These SNPs were confirmed via Sanger sequencing on genomic DNA (Figure S11 in Additional file 1). One of these SNPs introduces a premature stop codon, resulting in a protein that is only 136 amino acids long (the intact PSTOL1 protein is 324 amino acids) and, therefore, presumably non-functional.
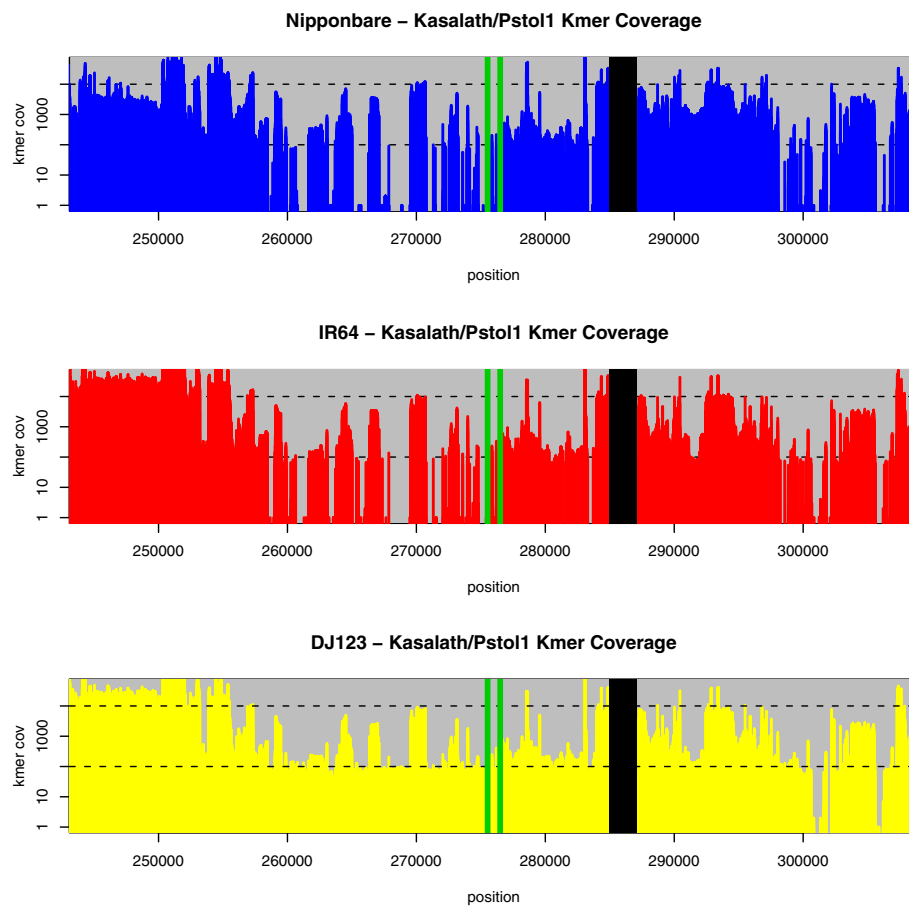
Figure 5 K-mer Coverage across the Kasalath/*Pstol1* gene in the three genomes, with 30 kbp of upstream and downstream flanking sequence. The k-mer coverage is plotted with respect to the reference Kasalath sequence (AB458444.1). The position of the *Pstol1* gene is indicated with green vertical bars. Also see Figure S9 in Additional file 1 for a detailed view of the *Pstol1* coverage, and Figure S10 in Additional file 1 for a plot of the entire Kasalath sequence. Unresolved gaps in the reference sequence are indicated with black vertical bars. Only DJ123 has consistent coverage across this region, especially upstream of the gene, while the other two genomes show complete gaps in coverage.

Conclusions

In this study we wanted to overcome the limitation on sequencing and comparison to a reference genome by instead analyzing high quality *de novo* assemblies of multiple rice genomes to observe biologically significant changes between them. The rice accessions sequenced were selected to represent the *indica* (cv IR64), *aus* (DJ123) and *temperate japonica* (Nipponbare) subpopulations (Figure 1). The inclusion of the high quality, BAC-by-BAC assembly of Nipponbare and the shotgun assembly of 93–11 provided a control that allowed us to assess the quality of the different datasets and *de novo* assembly strategies. It is apparent from comparing different assembly software that ALLPATHS-LG gave the best results in our hands (Table S3 in Additional file 1). It is also apparent that the use of k-mer frequencies is a robust technique for characterizing repetitive regions, and enabled us to correctly characterize and validate genome-specific regions.

The three-way comparison among the different genomes was informative in identifying major shared and structurally variable regions of the rice genome. We were particularly interested in regions that were structurally unique to either the *indica* and/or the *aus* genome because they would likely have been discarded in previous re-sequencing efforts due to difficulties aligning their sequencing reads to the Nipponbare reference genome. This would be particularly true for longer genome-specific sequences, which would be completely absent in the alignments to the reference. We anticipate the on-going 3,000 rice genomes project [36] will benefit greatly from having our assemblies available, especially so that they can map variations within regions not present in the Nipponbare reference as they have currently done. We also anticipate future studies will systematically perform follow-up functional studies of genome-specific gene loci as being likely candidates for phenotypic differences observed between the genomes.

Our analysis clearly demonstrates that the *indica* and the *aus* genomes are more distantly related than previously known. Because the *aus* subpopulation is phenotypically so similar to *indica*, the degree of genetic differentiation has been underappreciated by breeders and geneticists alike [43,63,64]. The unusual characteristics of the *aus* subpopulation, combined with evidence of unique *aus* alleles at loci such as *Rc*, conferring white versus colored pericarp [19], the *Snorkel* locus conferring deep water ability [45], the *Pstol1* locus conferring phosphorus-update efficiency [41], or the *Sub1* locus conferring submergence tolerance [33], all support the hypothesis that *aus* may have a unique domestication history compared to *japonica* and *indica*. These findings underscore the importance of recognizing genetic subpopulation structure to guide plant breeders in identifying novel sources of variation for traits of interest. In recent years, many key biotic and abiotic stress tolerance genes have been discovered in *aus* varieties [33,41,44-46]. It is interesting to note that in several cases, the donor *aus* germplasm is referred to as *indica*, underscoring how *indica* and *aus* are often confused, as noted for the DJ123 haplotype of the S5 hybrid sterility locus (see above).

The overall annotation of our Nipponbare assembly is quite close to that of the reference Nipponbare genome. This illustrates that the approach we describe here provides a genome sequence of considerable vitality for further research. However, our contig N50 sizes (as opposed to scaffold N50) are still fragmented by the presence of repeats too long and too complex to be fully resolved in the short read assemblies. This somewhat limits application when studying large structural rearrangements, as exemplified by the *Pup1* region that remains partially unassembled in the *aus* variety DJ123 (Figure 5), and the modest differences we observed in annotation quality between core genes and genome-specific genes. We anticipate that some combination of short-read NGS sequencing and newly emerging long read sequences, such as Pacific Biosciences Single Molecule Real Time Sequence [65], which can now produce reads approaching 100 kbp long, will soon overcome this limitation and provide assemblies approaching, or perhaps even surpassing, those provided by the vastly more expensive and time consuming BAC-by-BAC approach. Once this occurs it should spark an outburst of genomics studies of agronomically important plant genomes, greatly enriching our potential to understand their many unique qualities and characteristics and paving the way for enhanced utilization of natural variation in plant improvement.

Materials and methods
Plant material
Three rice (*Oryza sativa*) accessions (Nipponbare, DJ123) were used in the study. Accession information

(that is, Genetic Stocks *Oryza* (GSOR) identifier, accession name, country of origin, subpopulation) is summarized in Table 10 [63]. The plants were grown in the Guterman greenhouse facility at Cornell University, leaf tissue was harvested from one-month-old seedlings, ground in a mortar and pestle, and DNA was extracted using the Qiagen Plant DNeasy kit (Qiagen, Valencia, CA, USA).

DNA sequencing
The DNA sequencing was performed in the Cold Spring Harbor Laboratory Genome Center using Illumina HiSeq 2000 instruments. For each of the three varieties, three libraries were sequenced following the requirements and recommendations of the ALLPATHS-LG whole genome assembler: (1) a 180 bp fragment library sequenced as 2×100 bp reads; (2) an approximately 2 kbp jumping library sequenced as 2×50 bp reads; and (3) an approximately 5 kbp jumping library sequenced as 2×50 bp reads.

For the 180-bp overlap library the sample was mechanically fragmented by using the Covaris S2 System and then prepared based on the New England Biolabs NEB-Next Illumina library protocol and ligated to standard Illumina paired-end adapters. To maximize sample throughput the samples were size-selected in 50-bp windows between 290 and 310 bp using the Caliper XT instrument. Each library was PCR enriched for 12 cycles and quantified using the Bioanalyzer.

For the jumping libraries, the Illumina mate-pair library protocol was used. The DNA was fragmented into 2 kb and 5 kb segments. We again used the Covaris S2 System using programs that we developed in the lab. The fragmented DNA was then end-repaired with biotin-labeled dNTPs. The labeled fragments are circularized and fragmented again into 400 bp pieces. Fragments with the biotin labels are enriched, end-repaired, and ligated with adapters used for downstream processes. Each library was PCR enriched for 18 cycles and size-selected for 350 to 650 bp fragments. The final library consists of fragments made up of two DNA segments that were originally separated by approximately 2 kbp or approximately 5 kb. Each of the libraries was sequenced to 30× to 80× sequence coverage, as recommended by the assembler.

Table 10 Accession information for the three rice genomes in the Genetic Stocks *Oryza* (GSOR) stock center

GSOR ID	Accession name	Country of origin	Subpopulation
301164	Nipponbare	Japan	*temperate japonica*
312010	IR64	Philippines	*indica*
301307	DJ123	Bangladesh	*aus*

Libraries were sequenced on one or more lanes of an Illumina HiSeq 2000 using paired-end 50- or 100-bp runs. Image processing and base calling were performed as the runs progressed with Illumina's Real Time Analysis (RTA) software. The binary base call files were streamed to a shared Linux server for further processing. The Illumina Casava pipeline (v1.8) was used to process the binary files to fastq files containing the base-called reads and per base quality scores. Only reads passing the standard Illumina quality filter were included in the output files.

Genome assembly

The ALLPATHS-LG version R41348 assembly algorithm was used for the assemblies. It consists of five major phases: (1) pre-assembly error correction, (2) merging of the overlapping fragment reads into extended reads, (3) constructing the unipath graph from the k-mers present in the reads, (4) scaffolding the unipaths with the jumping libraries, and (5) gap closing. To complete the five phases, the algorithm requires an overlapping pair fragment library and at least one jumping library, although the authors recommend at least two jumping libraries of approximately 2 kbp and approximately 5 kbp or larger. We assembled each of the genomes using approximately 50× coverage of the fragment library and approximately 30× coverage of each of the two jumping libraries using the recommended parameters, except we lowered the MIN_CONTIG size to 300 bp from the default 1,000 bp. This parameter controls the minimum contig size to be used for scaffolding, and our previous testing determined this change leads to (modestly) improved contig and scaffold statistics.

We also evaluated using SOAPdenovo2 [66] and SGA [67] for the assemblies (Table S3 in Additional file 1), using the same fragment, 2 kbp, and 5 kbp libraries but both assemblers had substantially worse contiguity statistics under a variety of parameter settings. For SOAPdenovo2, we corrected the reads using the Quake error correction algorithm [68], and then ran seven assemblies with the de Bruijn graph k-mer size set to k = 31 through k = 45 (odd values only, as required). In every attempt the scaffold N50 size was below 10 kbp compared with >200 kbp for our best ALLPATHS-LG assembly. For SGA, we evaluated four assemblies with the string graph minimum overlap length of k = 71 through k = 77 (odd values only, as required), but the scaffold N50 size was below 15 kbp in every attempt. We hypothesize that ALLPATHS-LG achieved superior results because the algorithm automatically measures many of the properties of the sequencing data, and could therefore self-adjust the various cutoffs used by the algorithm for error correction, contiguing, and scaffolding.

Applying nomenclature proposed by [51], we have named these assemblies to convey accession, quality, origin, and iteration as follows: Os-Nipponbare-Draft-CSHL-1.0, Os-IR64-Draft-CSHL-1.0, Os-DJ123-Draft-CSHL-1.0.

Genome annotation

Repeat elements were masked using RepeatMasker [69] with a rice repeat library available from the Arizona Genome Institute. Protein-coding genes were annotated using MAKER-P version 2.30, installed on the Texas Advanced Computer Center Lonestar cluster and provisioned through an iPlant Collaborative allocation [52,70-72]. Sequence evidence used as input for MAKER-P included *Oryza* expressed sequences (EST, cDNA, and mRNA) downloaded from the National Center for Biotechnology Information (NCBI), and annotated coding and protein sequences available for Nipponbare (IRGSP1.0 and MSU release 7) [51], 93-11 [28], and PA64s (Table S1 in Additional file 1). *Ab initio* gene predictions made using FGENESH [73] were incorporated exogenously into the MAKER-P pipeline using the pred_gff parameter. The SNAP [74] *ab initio* predictor was run within MAKER-P using the O.sativa.hmm parameter provided with SNAP. To annotate protein domain structure and assign GO terms we used InterProScan 5 software [75], available within the iPlant Discovery Environment [76]. Among resulting InterPro domains we curated 21 as being associated with transposon-encoded genes and screened out MAKER-P annotations with these domains (IPR000477, IPR001207, IPR001584, IPR002559, IPR004242, IPR004252, IPR004264, IPR004330, IPR004332, IPR005063, IPR005162, IPR006912, IPR007321, IPR013103, IPR013242, IPR014736, IPR015401, IPR018289, IPR026103, IPR026960, IPR027806). To identify homologies we conducted BLASTP alignment to the plants subsection of NCBI RefSeq (release 63), using an e-value threshold of 1e-10.

Whole genome comparisons

We used the MUMmer [77] whole genome alignment package and the GAGE assembly comparison scripts to compare the *de novo* assemblies to the reference Nipponbare and *Indica* genomes. Briefly, we aligned the assemblies to the genomes using *nucmer* using sensitive alignment settings (-c 65 -l 30 -banded -D 5). For base level accuracy evaluations, we used the GAGE assembly comparison script, which further refines the alignments by computing the best set of one-to-one alignments between the two genomes using the dynamic programming algorithm *delta-filter.* This algorithm weighs the length of the alignments and their percentage identity to select one-to-one non-redundant alignments. This effectively discards spurious repetitive alignments from consideration, allowing us to focus on the meaningful differences

between the genomes. Finally, the evaluation algorithm uses *dnadiff* to scan the remaining, non-repetitive alignments to summarize the agreement between the sequences, including characterizing the nature of any non-aligning bases as substitutions, small indels, or other larger structural variations. To characterize the unaligned regions of the reference genome, we converted the whole genome alignments into BED format. For this we did not exclude repetitive alignments, so that we could focus on novel sequence instead of copy number differences. We used BEDTools [78] to intersect the unaligned segments with the reference annotation, and summarized the size distributions of the unaligned segments using AMOS [79].

K-mer analysis

To evaluate the repeat composition, we selected a random sample of 400 million unassembled reads from each of the three genomes and used *Jellyfish* [80] to count the number of occurrences of all length 21 k-mers in each read set. Length 21 was selected to be sufficiently long so that the expected number of occurrences of a random k-mer was below 1, but short enough to be robust to sequencing errors. The modes of the 3 k-mer frequency distributions, excluding erroneous k-mers that occurred less than 10 times, were 60× (Nipponbare), 64× (DJ123), and 73× (IR64) drawn from an approximately negative binomial distribution (Figure S1 in Additional file 1). These values correspond to the average k-mer coverage for single copy, non-repetitive regions of the genome. See Kelly *et al.* [68] for a discussion of k-mer frequencies. We then used the AMOS program *kmer-cov-plot* [79] to report the kmer coverage along the two reference genomes using the three databases of read k-mer frequencies. Unlike read alignments, which may be sensitive to repeats and variations, evaluating k-mer coverage is very robust to determine repetitive content [81,82]. Single nucleotide variants are also readily apparent in these plots as abrupt gaps in coverage kilobase pairs long, while indels will be present as longer gaps in coverage [83].

Pan-genome analysis

The pan-genome analysis followed the reference-based analysis above, using *nucmer* to align the genomes to each other, *BEDTools* to find the genome-specific and shared regions of the genomes, and the *jellyfish/AMOS* k-mer analysis as described above to classify unique and repetitive sequences. We also used *BEDTools* to intersect the genome-specific/shared regions against their respective annotations to determine how the exonic bases were shared across the genomes. We summarized the genome-specific/shared exonic bases into gene counts by counting the total number of shared or specific exonic bases across all possible transcripts for a gene, and assigned the gene to

Table 11 NCBI Sequence Read Archive accession codes for sequencing data used in this study

Genome	Library type	Read length	SRA accession
Nipponbare	180 bp fragment	2 × 101	SRX734432
	2 kbp jump	2 × 50	SRX179260
	5 kbp jump	2 × 50	SRX179265
IR64	180 bp fragment	2 × 101	SRX180537
	2 kbp jump	2 × 50	SRX180555
	5 kbp jump	2 × 50	SRX180597
DJ123	180 bp fragment	2 × 101	SRX180718
	2 kbp jump	2 × 50	SRX180822
	5 kbp jump	2 × 50	SRX180892

the sector of the Venn diagram with the most bases associated with it. For the purposes of the Venn diagram (Figure 2A), wherever possible, the Nipponbare base or gene counts were used, followed by the values from IR64, and then followed by the DJ123 specific values, although the values were all largely consistent.

PCR and sequencing validation of specific regions

The same algorithms and parameters as the pan-genome analysis were also used to characterize the specific regions identified in the paper. PCR and/or sequencing validation were performed on genomic DNA extracted from tissue collected from independently grown plants obtained from the same seed source used for Illumina sequencing. Genomic DNA was extracted from young leaf tissue using the Qiagen Plant DNeasy Mini kit. Primers used for validation of 10 of the longest genome-specific sequences from each rice line, and of the *S5* and *Pup1* loci, are listed in Table S4a-c in Additional file 1. Sanger sequencing was performed at the Biotechnology Resource Center at Cornell University.

Data access

The read data, assemblies, annotations, and pan-genome alignments are posted on the CSHL website at [84]. The NCBI Sequence Read Archive (SRA) accession numbers for the short read data used in this study are listed in Table 11. Analysis software packages are available open source from the websites for ALLPATHS-LG [85], MUMmer [86], AMOS [87], Jellyfish [88], and BEDTools [89].

Additional files

Additional file 1: Contains the supplementary tables and figures described in the manuscript.

Additional file 2: Is an excel spreadsheet listing the putative strain-specific genes.

Abbreviations
BAC: bacterial artificial chromosome; bp: base pair; CDS: coding sequence; EST: expressed sequence tag; GO: Gene Ontology; IRGSP: International Rice Genome Sequencing Program; MSU: Michigan State University; NCBI: National Center for Biotechnology Information; NGS: next generation sequencing; ORF: open reading frame; PCR: polymerase chain reaction; QTL: quantitative trait locus; SNP: single nucleotide polymorphism; UTR: untranslated region.

Authors' contributions
SRM, WRM, and DW designed the study. MCS, LGM, JCS, AHW, JG, EB, HL, MK, and JMC performed the computational analysis. LGM, MK, EA, EG, MHW, and SRM performed the experimental analysis. MCS, LGM, JCS, JMC, DW, SRM, and WRM wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [2]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853, USA. [3]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca 62210, Morelos, Mexico. [4]Macalester College, St Paul, MN 55105, USA. [5]Stony Brook University, Stony Brook, NY 11794, USA. [6]USDA-ARS NAA Plant, Soil and Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY 14853, USA.

References
1. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S: **Genetic structure and diversity in Oryza sativa L.** *Genetics* 2005, **169**:1631–1638.
2. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, *et al*: **A map of rice genome variation reveals the origin of cultivated rice.** *Nature* 2012, **490**:497–501.
3. Zhao KY, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante C, McCouch SR: **Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome.** *Plos One* 2010, **5**:e10780.
4. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, Cargill M, Auton A, Reynolds A, Elkahloun AG, Castelhano M, Mosher DS, Sutter NB, Johnson GS, Novembre J, Hubisz MJ, Siepel A, Wayne RK, Bustamante CD, Ostrander EA: **A simple genetic architecture underlies morphological variation in dogs.** *PLoS Biol* 2010, **8**:e1000451.
5. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG: **Measures of human population structure show heterogeneity among genomic regions.** *Genome Res* 2005, **15**:1468–1476.
6. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J: **A single domestication for maize shown by multilocus microsatellite genotyping.** *Proc Natl Acad Sci U S A* 2002, **99**:6080–6084.
7. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR: **Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa.** *Nat Commun* 2011, **2**:467.
8. Ma J, Bennetzen JL: **Rapid recent growth and divergence of rice nuclear genomes.** *Proc Natl Acad Sci U S A* 2004, **101**:12404–12410.
9. Cheng CY, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E: **Polyphyletic origin of cultivated rice: Based on the interspersion pattern of SINEs.** *Mol Biol Evol* 2003, **20**:67–75.
10. Kovach MJ, Sweeney MT, McCouch SR: **New insights into the history of rice domestication.** *Trends Genet* 2007, **23**:578–587.
11. Roy SC: **A preliminary classification of the wild rices of the Central Province and Berar.** *Agric J India* 1921, **16**:365–380.
12. Second G: **Origin of the genic diversity of cultivated rice (Oryza-spp) - study of the polymorphism scored at 40 isoenzyme loci.** *Jpn J Genet* 1982, **57**:25–57.
13. Second G: **Molecular markers in rice systematics and the evaluation of genetic resources.** *Biotechnol Agric For* 1991, **14**:468–494.
14. Ding J, Araki H, Wang Q, Zhang P, Yang S, Chen JQ, Tian D: **Highly asymmetric rice genomes.** *BMC Genomics* 2007, **8**:154.
15. Liu XH, Lu TT, Yu SL, Li Y, Huang YC, Huang T, Zhang L, Zhu JJ, Zhao Q, Fan DL, Mu J, Shangguan YY, Feng Q, Guan JP, Ying K, Zhang Y, Lin ZX, Sun ZX, Qian Q, Lu YP, Han B: **A collection of 10,096 indica rice full-length cDNAs reveals highly expressed sequence divergence between Oryza sativa indica and japonica subspecies.** *Plant Mol Biol* 2007, **65**:403–415.
16. Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH: **An SNP resource for rice genetics and breeding based on subspecies Indica and Japonica genome alignments.** *Genome Res* 2004, **14**:1812–1819.
17. Huang XH, Lu GJ, Zhao Q, Liu XH, Han B: **Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice.** *Plant Physiol* 2008, **148**:25–40.
18. Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M: **Deletion in a gene associated with grain size increased yields during rice domestication.** *Nat Genet* 2008, **40**:1023–1028.
19. Takano-Kai N, Jiang H, Kubo T, Sweeney M, Matsumoto T, Kanamori H, Padhukasahasram B, Bustamante C, Yoshimura A, Doi K, McCouch S: **Global dissemination of a single mutation conferring white pericarp in rice.** *PLoS Genet* 2007, **3**:e133.
20. Takano-Kai N, Jiang H, Kubo T, Sweeney M, Matsumoto T, Kanamori H, Padhukasahasram B, Bustamante C, Yoshimura A, Doi K, McCouch S: **Evolutionary history of GS3, a gene conferring grain length in rice.** *Genetics* 2009, **182**:1323–1334.
21. Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, Fu Y, Cai H, Wang X, Xie D, Sun C: **Control of a key transition from prostrate to erect growth in rice domestication.** *Nat Genet* 2008, **40**:1360–1364.
22. Harushima Y, Nakagahra M, Yano M, Sasaki T, Kurata N: **Diverse variation of reproductive barriers in three intraspecific rice crosses.** *Genetics* 2002, **160**:313–322.
23. Lin SY, Ikehashi H, Yanagihara S, Kawashima A: **Segregation distortion via male gametes in hybrids between Indica and Japonica or wide-compatibility varieties of rice (Oryza-sativa L).** *Theor Appl Genet* 1992, **84**:812–818.
24. Oka HI: **Functions and genetic base of reproductive barriers.** In *Origin of Cultivated Rice.* Amsterdam: Tokyo/Elsevier Science/Japan Scientific Societies Press; 1988:181–209.
25. Sano Y: **Constraints in using wild relatives in breeding: lack of basic knowledge on crop gene pools.** *Int Crop Sci* 1993, **1**:437–443.
26. Ammiraju JSS, Song XA, Luo MZ, Sisneros N, Angelova A, Kudrna D, Kim H, Yu Y, Goicoechea JL, Lorieux M, Kurata N, Brar D, Ware D, Jackson S, Wing RA: **The Oryza BAC resource: a genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives.** *Breeding Sci* 2010, **60**:536–543.
27. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793–800.
28. Gao ZY, Zhao SC, He WM, Guo LB, Peng YL, Wang JJ, Guo XS, Zhang XM, Rao YC, Zhang C, Dong GJ, Zheng FY, Lu CX, Hu J, Zhou Q, Liu HJ, Wu HY, Xu J, Ni PX, Zeng DL, Liu DH, Tian P, Gong LH, Ye C, Zhang GH, Wang J, Tian FK, Xue DW, Liao Y, Zhu L, *et al*: **Dissecting yield-associated loci in**

super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc Natl Acad Sci U S A* 2013, 110:14492–14497.

29. Yu J, Wang J, Lin W, Li SG, Li H, Zhou J, Ni PX, Dong W, Hu SN, Zeng CQ, Zhang JG, Zhang Y, Li RQ, Xu ZY, Li ST, Li XR, Zheng HK, Cong LJ, Lin L, Yin JN, Geng JN, Li GY, Shi JP, Liu J, Lv H, Li J, Wang J, Deng YJ, Ran LH, Shi XL, et al: The Genomes of Oryza sativa: A history of duplications. *PLoS Biol* 2005, 3:266–281.

30. Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y, Li CY, Zhu CR, Lu TT, Zhang ZW, Li M, Fan DL, Guo YL, Wang A, Wang L, Deng LW, Li WJ, Lu YQ, Weng QJ, Liu KY, Huang T, Zhou TY, Jing YF, Li W, Lin Z, Buckler ES, Qian QA, Zhang QF, Li JY, Han B: Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 2010, 42:961–U76.

31. McCouch SR, Zhao KY, Wright M, Tung CW, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML, McClung A, Eizenga G, Bustamante C: Development of genome-wide SNP assays for rice. *Breeding Sci* 2010, 60:524–535.

32. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell CR, Leung H, Leach JE: Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A* 2009, 106:12273–12278.

33. Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, Ismail AM, Bailey-Serres J, Ronald PC, Mackill DJ: Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 2006, 442:705–708.

34. Huang XH, Feng Q, Qian Q, Zhao Q, Wang L, Wang AH, Guan JP, Fan DL, Weng QJ, Huang T, Dong GJ, Sang T, Han B: High-throughput genotyping by whole-genome resequencing. *Genome Res* 2009, 19:1068–1076.

35. Xu X, Liu X, Ge S, Jensen JD, Hu FY, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li JX, He WM, Zhang GJ, Zheng XM, Zhang FM, Li YR, Yu C, Kristiansen K, Zhang XQ, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W: Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* 2012, 30:105–U57.

36. Li JY, Wang J, Zeigler RS: The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience* 2014, 3:8.

37. Han B, Xue YB: Genome-wide intraspecific DNA-sequence variations in rice. *Curr Opin Plant Biol* 2003, 6:134–138.

38. Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA: Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC Evol Biol* 2007, 7:152.

39. Yu P, Wang CH, Xu Q, Feng Y, Yuan XP, Yu HY, Wang YP, Tang SX, Wei XH: Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* 2011, 12:372.

40. Famoso AN, Zhao K, Clark RT, Tung CW, Wright MH, Bustamante C, Kochian LV, McCouch SR: Genetic architecture of aluminum tolerance in rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. *PLoS Genet* 2011, 7:e1002221.

41. Gamuyao R, Chin JH, Pariasca-Tanaka J, Pesaresi P, Catausan S, Dalid C, Slamet-Loedin I, Tecson-Mendoza EM, Wissuwa M, Heuer S: The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* 2012, 488:535.

42. Uga Y, Sugimoto K, Ogawa S, Rane J, Ishitani M, Hara N, Kitomi Y, Inukai Y, Ono K, Kanno N, Inoue H, Takehisa H, Motoyama R, Nagamura Y, Wu J, Matsumoto T, Takai T, Okuno K, Yano M: Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. *Nat Genet* 2013, 45:1097–1102.

43. Liakat Ali M, McClung AM, Jia MH, Kimball JA, McCouch SR, Susan R, Georgia CE: A rice diversity panel evaluated for genetic and agro-morphological diversity between subpopulations and its geographic distribution. *Crop Sci* 2011, 51:2021–2035.

44. Garris AJ, McCouch SR, Kresovich S: Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (Oryza sativa L.). *Genetics* 2003, 165:759–769.

45. Hattori Y, Nagai K, Furukawa S, Song XJ, Kawano R, Sakakibara H, Wu J, Matsumoto T, Yoshimura A, Kitano H, Matsuoka M, Mori H, Ashikari M: The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* 2009, 460:1026–1030.

46. Bernier J, Kumar A, Venuprasad R, Spaner D, Verulkar S, Mandal N, Sinha P, Peeraju P, Dongre P, Mahto RN, Atlin G: Characterization of the effect of a

QTL for drought resistance in rice, qtl12.1, over a range of environments in the Philippines and eastern India. *Euphytica* 2009, 166:207–217.

47. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 2011, 108:1513–1518.

48. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, et al: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2013, 2:10.

49. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S: Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 2011, 21:2224–2241.

50. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA: GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* 2012, 22:557–567.

51. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T: Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* 2013, 6:4.

52. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, Ware D, Shiu SH, Childs KL, Sun Y, Jiang N, Yandell M: MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* 2014, 164:513–524.

53. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA: The relationship of protein conservation and sequence length. *BMC Evol Biol* 2002, 2:20.

54. Capra JA, Pollard KS, Singh M: Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol* 2010, 11:R127.

55. Cai JJ, Petrov DA: Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* 2010, 2:393–409.

56. Yanagihara S, Mccouch SR, Ishikawa K, Ogi Y, Maruyama K, Ikehashi H: Molecular analysis of the inheritance of the S-5 locus, conferring wide compatibility in Indica-Japonica hybrids of rice (Oryza-sativa L). *Theor Appl Genet* 1995, 90:182–188.

57. Chen JJ, Ding JH, Ouyang YD, Du HY, Yang JY, Cheng K, Zhao J, Qiu SQ, Zhang XL, Yao JL, Liu KD, Wang L, Xu CG, Li XH, Xue YB, Xia M, Ji Q, Lu JF, Xu ML, Zhang QF: A triallelic system of S5 is a major regulator of the reproductive barrier and compatibility of indica-japonica hybrids in rice. *Proc Natl Acad Sci U S A* 2008, 105:11436–11441.

58. Yang J, Zhao X, Cheng K, Du H, Ouyang Y, Chen J, Qiu S, Huang J, Jiang Y, Jiang L, Ding J, Wang J, Xu C, Li X, Zhang Q: A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science* 2012, 337:1336–1340.

59. He GM, Luo XJ, Tian F, Li KG, Zhu ZF, Su W, Qian XY, Fu YC, Wang XK, Sun CQ, Yang JS: Haplotype variation in structure and expression of a gene cluster associated with a quantitative trait locus for improved yield in rice. *Genome Res* 2006, 16:618–626.

60. Wissuwa M, Wegner J, Ae N, Yano M: Substitution mapping of Pup1: a major QTL increasing phosphorus uptake of rice from a phosphorus-deficient soil. *Theor Appl Genet* 2002, 105:890–897.

61. Wissuwa M, Yano M, Ae N: Mapping of QTLs for phosphorus-deficiency tolerance in rice (Oryza sativa L.). *Theor Appl Genet* 1998, 97:777–783.

62. Chin JH, Gamuyao R, Dalid C, Bustamam M, Prasetiyono J, Moeljopawiro S, Wissuwa M, Heuer S: Developing rice with high yield under phosphorus deficiency: Pup1 sequence to application. *Plant Physiol* 2011, 156:1202–1216.

63. Eizenga GCAM, Bryant RJ, Yeater KM, McClung AM, McCouch SR: Registration of the rice diversity panel 1 for genomewide association studies. *J Plant Reg* 2013, 8:109–116.

64. Bin Rahman AN, Zhang J: Rayada specialty: the forgotten resource of elite features of rice. *Rice* 2013, 6:41.

65. Roberts RJ, Carneiro MO, Schatz MC: **The advantages of SMRT sequencing.** *Genome Biol* 2013, **14**:405.

66. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.** *Gigascience* 2012, **1**:18.

67. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures.** *Genome Res* 2012, **22**:549–556.

68. Kelley DR, Schatz MC, Salzberg SL: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biol* 2010, **11**:R116.

69. Smit AFA, Hubley R, Green P: RepeatMaster Open-3.0. 1996–2010. http://www.repeatmasker.org

70. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188–196.

71. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, Muir A, Merchant N, Lowry S, Mock S, Helmke M, Kubach A, Narro M, Hopkins N, Micklos D, Hilgert U, Gonzales M, Jordan C, Skidmore E, Dooley R, Cazes J, McLay R, Lu Z, Pasternak S, Koesterke L, Piel WH, *et al*: **The iPlant Collaborative: cyberinfrastructure for plant biology.** *Front Plant Sci* 2011, **2**:34.

72. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC bioinformatics* 2011, **12**:491.

73. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila genomic DNA.** *Genome Res* 2000, **10**:516–522.

74. Korf I: **Gene finding in novel genomes.** *BMC bioinformatics* 2004, **5**:59.

75. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30**:1236–1240.

76. Oliver SL, Lenards AJ, Barthelson RA, Merchant N, McKay SJ: **Using the iPlant collaborative discovery environment.** *Curr Protoc Bioinformatics* 2013, **Chapter 1**:Unit1 22.

77. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.

78. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841–842.

79. Schatz MC, Phillippy AM, Sommer DD, Delcher AL, Puiu D, Narzisi G, Salzberg SL, Pop M: **Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies.** *Brief Bioinform* 2013, **14**:213–224.

80. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27**:764–770.

81. Kurtz S, Narechania A, Stein JC, Ware D: **A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.** *BMC Genomics* 2008, **9**:517.

82. Phillippy AM, Schatz MC, Pop M: **Genome assembly forensics: finding the elusive mis-assembly.** *Genome Biol* 2008, **9**:R55.

83. Reyes J, Gomez-Romero L, Ibarra-Soria X, Palacios-Flores K, Arriola LR, Wences A, Garcia D, Boege M, Davila G, Flores M, Palacios R: **Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs.** *Proc Natl Acad Sci U S A* 2011, **108**:15294–15299.

84. New whole genome *de novo* assemblies of three divergent strains of rice (O. sativa) documents novel gene space of aus and indica. [http://schatzlab.cshl.edu/data/rice]

85. ALLPATHS-LG. [http://www.broadinstitute.org/software/allpaths-lg/blog/?page_id=12]

86. MUMmer. [http://mummer.sourceforge.net]

87. AMOS. [http://amos.sourceforge.net]

88. Jellyfish. [http://www.genome.umd.edu/jellyfish.html]

89. BEDTools. [https://github.com/arq5x/bedtools2]