nature genetics

The genome of the domesticated apple (*Malus* × *domestica* Borkh.)

Riccardo Velasco[1,20], Andrey Zharkikh[2,20], Jason Affourtit[3], Amit Dhingra[4], Alessandro Cestaro[1], Ananth Kalyanaraman[5], Paolo Fontana[1], Satish K Bhatnagar[2], Michela Troggio[1], Dmitry Pruss[2], Silvio Salvi[1,6], Massimo Pindo[1], Paolo Baldi[1], Sara Castelletti[1], Marina Cavaiuolo[1], Giuseppina Coppola[1], Fabrizio Costa[1], Valentina Cova[1], Antonio Dal Ri[1], Vadim Goremykin[1], Matteo Komjanc[1], Sara Longhi[1], Pierluigi Magnago[1], Giulia Malacarne[1], Mickael Malnoy[1], Diego Micheletti[1], Marco Moretto[1], Michele Perazzolli[1], Azeddine Si-Ammour[1], Silvia Vezzulli[1], Elena Zini[1], Glenn Eldredge[2], Lisa M Fitzgerald[2], Natalia Gutin[2], Jerry Lanchbury[2], Teresita Macalma[2], Jeff T Mitchell[2], Julia Reid[2], Bryan Wardell[2], Chinnappa Kodira[3], Zhoutao Chen[3], Brian Desany[3], Faheem Niazi[3], Melinda Palmer[3], Tyson Koepke[4], Derick Jiwan[4], Scott Schaeffer[4], Vandhana Krishnan[5], Changjun Wu[5], Vu T Chu[7], Stephen T King[7], Jessica Vick[7], Quanzhou Tao[8], Amy Mraz[8], Aimee Stormo[8], Keith Stormo[8], Robert Bogden[8], Davide Ederle[9], Alessandra Stella[9], Alberto Vecchietti[9], Martin M Kater[10], Simona Masiero[11], Pauline Lasserre[12], Yves Lespinasse[12], Andrew C Allan[13], Vincent Bus[14], David Chagné[15], Ross N Crowhurst[13], Andrew P Gleave[13], Enrico Lavezzo[16], Jeffrey A Fawcett[17,18], Sebastian Proost[17,18], Pierre Rouzé[17,18], Lieven Sterck[17,18], Stefano Toppo[19], Barbara Lazzari[9], Roger P Hellens[13], Charles-Eric Durel[12], Alexander Gutin[2], Roger E Bumgarner[7], Susan E Gardiner[15], Mark Skolnick[2], Michael Egholm[3], Yves Van de Peer[17,18], Francesco Salamini[1,9] & Roberto Viola[1]

We report a high-quality draft genome sequence of the domesticated apple (*Malus* × *domestica*). We show that a relatively recent (>50 million years ago) genome-wide duplication (GWD) has resulted in the transition from nine ancestral chromosomes to 17 chromosomes in the Pyreae. Traces of older GWDs partly support the monophyly of the ancestral paleohexaploidy of eudicots. Phylogenetic reconstruction of Pyreae and the genus *Malus*, relative to major Rosaceae taxa, identified the progenitor of the cultivated apple as *M. sieversii*. Expansion of gene families reported to be involved in fruit development may explain formation of the pome, a Pyreae-specific false fruit that develops by proliferation of the basal part of the sepals, the receptacle. In apple, a subclade of *MADS-box* genes, normally involved in flower and fruit development, is expanded to include 15 members, as are other gene families involved in Rosaceae-specific metabolism, such as transport and assimilation of sorbitol.

The domesticated apple (*Malus* × *domestica* Borkh., family Rosaceae, tribe Pyreae) is the main fruit crop of temperate regions of the world. Here we describe a high-quality draft genome sequence of the diploid apple cultivar 'Golden Delicious'. Domesticated apple genotypes are all highly heterozygous, imposing technical challenges in genome sequencing and assembly[1] while allowing identification of a very large set of SNPs[2].

Rosaceae belong to the rosids, which include one-third of all flowering plants[3]. Whereas the haploid (x) chromosome numbers of most Rosaceae are 7, 8 or 9, Pyreae have a distinctive $x = 17$. Pyreae have

[1]Istituto Agrario San Michele all'Adige Research and Innovation Centre, Foundation Edmund Mach, Trento, Italy. [2]Myriad Genetics, Salt Lake City, Utah, USA. [3]454 Life Sciences, A Roche Company, Branford, Connecticut, USA. [4]Department of Horticulture and Landscape Architecture, Washington State University, Pullman, Washington, USA. [5]School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington, USA. [6]Department of Agroenvironmental Sciences and Technologies, University of Bologna, Bologna, Italy. [7]Department of Microbiology, University of Washington, Seattle, Washington, USA. [8]Amplicon Express, Pullman, Washington, USA. [9]Parco Tecnologico Padano, Lodi, Italy. [10]Department of Biomolecular Science and Biotechnology, University of Milan, Milan, Italy. [11]Department of Biology, University of Milan, Milan, Italy. [12]French National Institute for Agricultural Research Angers-Nantes, UMR1259 Genetics and Horticulture, Genhort, IFR149 QUASAV, Beaucouze Cedex, France. [13]The New Zealand Institute for Plant & Food Research, Mt. Albert Research Centre, Auckland, New Zealand. [14]The New Zealand Institute for Plant & Food Research, Hawke's Bay Research Centre, Havelock North, New Zealand. [15]The New Zealand Institute for Plant & Food Research, Palmerston North Research Centre, Palmerston North, New Zealand. [16]Department of Histology, Microbiology, and Medical Biotechnologies, University of Padua, Padua, Italy. [17]Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, Flanders Institute for Biotechnology (VIB), Ghent University, Ghent, Belgium. [18]Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium. [19]Department of Biological Chemistry, University of Padua, Padua, Italy. [20]These authors contributed equally to this work. Correspondence should be addressed to R. Velasco (riccardo.velasco@iasma.it).

Received 19 November 2009; accepted 3 August 2010; published online 29 August 2010; doi:10.1038/ng.654

Table 1 Summary of genome assembly of the apple variety 'Golden Delicious'

		Size (Mb)	No.	N_{50}	L_{50}	% of sequence assembled	% of sequence anchored
Contigs	All	**603.9**[a]	122,146	16,171	13.4	81.3	
	Singletons	26.6	19,070	5,327	3.4		
	In metacontigs	577.3	103,076	13,031	15.2		
	Anchored	550.3	95,716	12,149	15.4		
Metacontigs	All	598.3	1,629	102	1,542.7		
	Anchored	528.3	439	80	1,971.0		71.2
Repetitive	Assembled	362.3 (48.8%)					
Sequences	Not assembled	**138.4** (18.6%)					
	Total	500.7 (67.4%)				72.3	
Estimated genome size[b]		**742.3**					
No. of genes[c]			95,216/57,386				90.2
Anchoring markers			1,643				

N_{50}, minimum number of contigs required to represent 50% of the genome. L_{50}, length of the smallest N_{50} contig. [a]The value reported was reduced to take into account the haplotype overlapping and hemizygous DNA. See details in **Supplementary Note**. [b]Calculated as the sum of 603.9 and 138.4 Mb (see **Supplementary Note**). [c]Before/after excluding transposable element–related genes, alleles, short predictions and low-level functional annotate genes.

gene families among ten sequenced plant species revealed apple-specific subclades of genes encoding MADS-box transcription factors and overrepresented sorbitol-related genes, which may contribute to specific aspects of apple development and carbohydrate metabolism (**Table 2** and **Supplementary Table 7**, and see Discussion). The 71.2% of the genomic sequences that were anchored represent the gene-rich part of the genome, which covers as many as 90.2% of the genes assigned to the chromosomes. The distribution of transposable elements and predicted genes along the linkage groups is reported in **Supplementary Figure 6**. The total number of genes predicted for the apple genome (57,386, including some genes that may be present only in one of the two chromosomes of a pair) is the highest reported among plants so far (**Supplementary Note**).

long been considered an example of allopolyploidization between species related to extant Spiraeoideae ($x = 9$) and Amygdaleoideae ($x = 8$), although a within-lineage polyploidization event has also been hypothesized[4].

In addition, we examine the genetic variability in Rosaceae and related taxa, comparing Pyreae species, Rosaceae tribes and two rosid families. Gene content and order of the assembled chromosomes indicate that both recent and old GWDs have occurred. We provide a model describing the evolution of the Pyreae genome, including *Malus*, and offer insights into the origin of the domesticated apple.

RESULTS

Sequencing, assembling and anchoring the apple genome

Sequencing and assembly of the 'Golden Delicious' apple genome followed the whole-genome shotgun approach. Of the 16.9-fold genome coverage, 26% was provided by Sanger dye primer sequencing of paired reads, and the remaining 74% was from 454 sequencing by synthesis of paired and unpaired reads (**Supplementary Table 1** and **Supplementary Note**). An iterative assembly approach, previously used to assemble the highly heterozygous grape genome[1], produced 122,146 contigs, 103,076 of which were assembled into 1,629 metacontigs (**Table 1**, **Supplementary Fig. 1** and **Supplementary Note**). The total contig length (603.9 Mb) covers about 81.3% of the apple genome (**Table 1** and **Supplementary Note**). Anchoring of metacontigs (598.3 Mbp, or 71.2% of genome) was based on the high-quality genetic map with 1,643 markers (**Supplementary Table 2** and **Supplementary Note**). In total, 17 linkage groups, or chromosomes, were reconstructed. In the genome, repetitive elements correspond to 500.7 Mb (67%; **Supplementary Note**). The unassembled part of the genome is 98% repetitive (138.4 Mb), and the estimated genome size is 742.3 Mb (**Table 1** and **Supplementary Note**). We compared repetitive elements among ten plant species (**Supplementary Tables 3–6**). Information on relevant genes and genome parameters is provided in **Tables 1** and **2**, **Supplementary Figures 2–5** and **Supplementary Tables 7–19**. Comparing

Genome-wide duplications and the origin of the Pyreae

Pairwise comparison of 17 apple chromosomes highlighted strong collinearity between large segments of chromosomes 3 and 11, 5 and 10, 9 and 17, and 13 and 16, and between shorter segments of chromosomes 1 and 7, 2 and 7, 2 and 15, 4 and 12, 12 and 14, 6 and 14, and 8 and 15 (**Fig. 1a**). The distribution of synonymous substitution rates (K_S)—an indication of the relative age of duplication, based on the number of synonymous substitutions in the coding sequences—peaked around 0.2 for recently duplicated genes (**Fig. 1b**), indicating that a (recent) GWD has shaped the genome of the domesticated apple.

Dating of this GWD (**Supplementary Note**) was based on the construction of penalized likelihood trees, as described previously[5]. Given a node of grape to rosids fixed at 115 million years ago (Mya), the GWD has been dated to between 30 and 45 Mya[5]. If similar rates of protein evolution are assumed for apple and poplar (**Fig. 1c**), the recent apple GWD may be as old as that of poplar, about 60 to 65 Mya[6].

Remnants of older large-scale gene duplications or GWDs were also evident (**Supplementary Fig. 7a,b**). Genes in these duplicated regions had average K_S values around 1.6, as expected for paleoduplication events (**Fig. 1b**). Most remnants of these older duplications

Table 2 Comparison of the apple genome to other sequenced plant genomes

	Genes/gene density[a]	Transposable elements[b]	Transcription factors[c]	miRNAs[d]	Resistance genes[e]	Biosynthetic genes[f]
Apple	57,386/0.78	42.4	4,021	178	992 (58, 27)	1,246
Cucumber	26,682/0.73	14.8	ND	ND (171)	61 (ND)	ND
Soybean	46,430/0.42	50.3	5,671	41 (85)	392 (61, 32)	958
Poplar	45,654/0.94	35.0	2,758	174 (234)	402 (59, 20)	1,034
Arabidopsis	27,228/2.2	18.5	2,437	89 (199)	178 (32, 52)	719
Grape	33,514/0.66	21.5	2,080	130 (137)	341 (57, 11)	1,121
Rice	40,577/0.97	39.5	2,798	140 (447)	535 (89, 0)	910
Brachypodium	25,532/0.94	28.1	2,187	62 (129)	238 (89, 0)	390
Sorghum	34,496/0.47	62.0	2,312	116 (148)	245 (75, 0)	555
Maize	32,540/0.15	84.2	5,246	153 (170)	129 (74, 0)	457

Only statistics showing major or important differences among species are reported (ND, not determined). Data were taken from web sites listed in the 'URLs' section in the Online Methods.
[a]Gene density expressed in number of genes per 10 kb. [b]Expressed in % of total genome size. For details, see **Supplementary Table 7**. [c]Total number of transcription-factor genes identified in plant genomes. For details, see **Supplementary Tables 7** and **8**. [d]Number of microRNAs (miRNAs) from miRNA families present in apple is given outside brackets; total number of miRNAs found in the literature is in brackets. For details, see **Supplementary Tables 13–16**. [e]Total number of *NBS* resistance genes (numbers in brackets respectively indicate % *NBS-LRR*, % *TIR-NBS-LRR*). For details see **Supplementary Tables 9** and **10**. [f]Total number of genes involved in volatile, aromatic-compound, pigment, antioxidant and sorbitol biosynthetic pathways. For details see **Supplementary Table 7**.

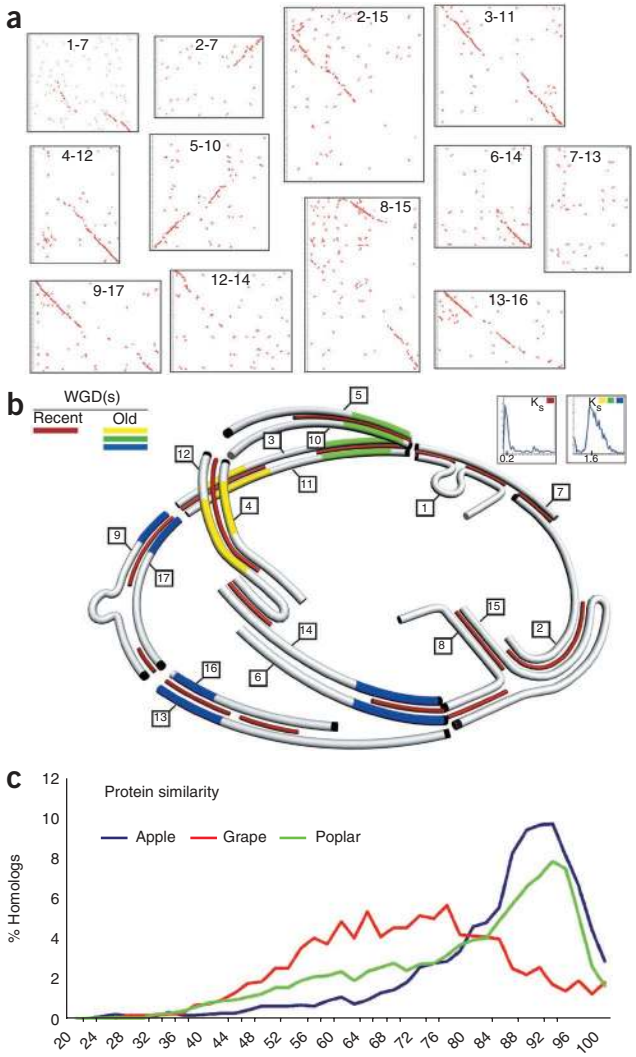

Figure 1 Genome-wide duplications in the apple genome. (**a**) Alignment of apple chromosomes shown by pairwise dot plots based on gene homology. Strong collinearity of members of chromosome doublets, or of large chromosome segments, indicates a recent GWD (red dots and bars in **a** and **b**, respectively). Unrelated chromosomes 7 and 13 were compared as a negative control. (**b**) Reconstruction of the relationships among apple chromosomes based on the most recent and the older GWD. The model derives from data in **a** for the recent GWD and from data in **Supplementary Figure 6b** for the oldest GWD. The chromosomes ends represented at bottom right corners in **a** are marked in black in **b**. Red bars, regions of synteny that support the recent GWD. Size of chromosomes is proportional to their DNA content in megabases. Segments of chromosomes 1, 5, 6, 8, 10, 13, 14 and 15 have no syntenic counterparts. Chromosome segments predicted to be the outcome of the older duplication are highlighted with blue, green and orange. Chromosomes 1, 2, 7, 8 and 15 do not show obvious signs of the older duplications, although they may contain short blocks of genes that reveal old paleopolyploid events. Inset graphs show that K_s from the comparisons between paralogous genes has a peak at 0.2 when the recent duplication is considered, and between 1.4 and 1.6 for the older paleopolyploid events. (**c**) Distributions of protein similarities for duplicated genes in duplicated segments compared with grape (red), poplar (green) and apple (blue).

are found between chromosomes 5 and 10 and chromosomes 3 and 11, between chromosomes 3 and 11 and chromosomes 4 and 12, and between chromosomes 6 and 14, 13 and 16, and 9 and 17 (**Fig. 1a**,**b**). Chromosomes 1, 2, 7, 8 and 15 seem relatively devoid of older duplicated blocks; however, short blocks of genes showing old polyploidy events were found on all chromosomes. One region in the apple genome with an approximate size of 4 to 7 Mbp seems to be clearly present in six copies (regions in blue, **Fig. 1a**,**b**). Remapping those to the ancestral state reveals a triplicate structure among parts of chromosomes 9 and 17, 6 and 14 and 13 and 16. Notably, we found that these regions are collinear with chromosomes 1, 14 and 17 of grape (**Fig. 2**), which have been demonstrated to be homologous because of an ancient hexaploidy[7]. Additional chromosomal fragments that we found to be duplicated in apple (green and yellow bars in **Fig. 1b**) can also be interpreted as remains of a paleohexaploid state of the eudicot progenitor on the basis of dot-plot comparisons among other grape and apple chromosomes (**Supplementary Fig. 8a**,**b**). This provides further evidence for a paleohexaploid state shared by most eudicots[8,9].

The chromosome homologies derived from the recent GWD allow inference of the cytological events that have led to the number and composition of the extant apple chromosomes, starting from a putative nine-chromosome ancestor (**Fig. 3**). Each doublet of the eight apple chromosomes (3-11, 5-10, 9-17 and 13-16) is derived principally from one ancestor, although minor interchromosomal rearrangements have occurred (**Supplementary Fig. 9a–k**). Chromosomes 4, 6, 12 and 14 originate from duplications of the ancient chromosomes V and VI, followed by a translocation and a deletion event. Similar events have generated chromosomes 1, 2, 7, 8 and 15 from chromosomes VII, VIII and IX. Chromosome 15 could have been produced from the translocation of an entire copy of chromosome IX into the centromeric region of chromosome VIII, following a model of dysploidy (reduction of chromosome number) common in cereals[10]. The second copy of ancient chromosome VIII has evolved into the extant chromosome 8. A conservative estimate of the number of large chromosome rearrangements since the divergence of the Pyreae subtribe, corresponding to the recent chromosome duplication, includes one chromosome fusion (extant chromosome 15), three translocations

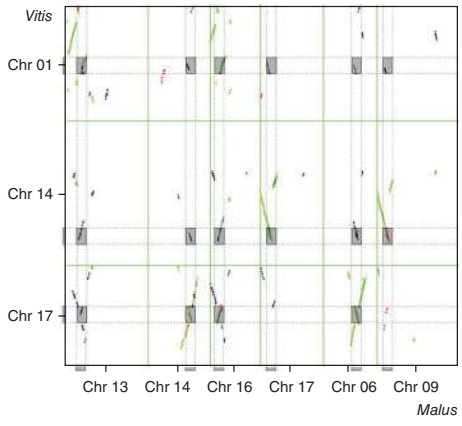

Figure 2 Dot-plot comparisons between apple and grape chromosomes. Dot plots are based on gene homology. The apple chromosomes are those with the segment triplication deriving from an old GWD (shown in blue in **Fig. 1b**). Grape chromosomes 1, 14 and 17 constitute a triplet having the same ancestor in common[7]. Chromosome segments with homologous genes common both to grape and apple (16 of a total of 18 comparisons) are indicated by gray boxes connected with dashed lines. Green, red and blue dots indicate increasing K_s values, in that order. Perpendicular lines on the *x* and *y* axes mark the middle of each chromosome. Green grid separates chromosomes.

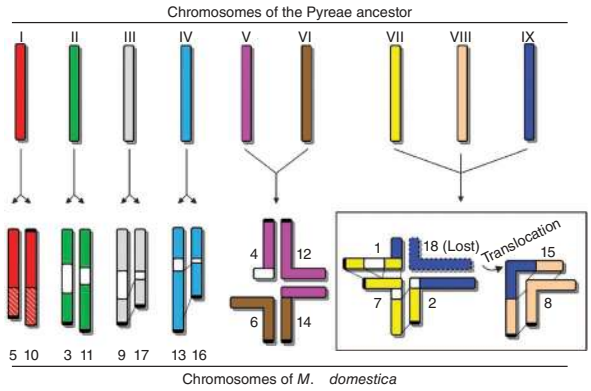

Figure 3 A model explaining the evolution from a 9-chromosome ancestor to the 17-chromosome karyotype of extant Pyreae, including the genus *Malus*. A GWD followed by a parsimony model of chromosome rearrangements is postulated. Shared colors indicate homology between extant chromosomes. White fragments of chromosomes indicate lack of a duplicated counterpart. The white-hatched portions of chromosomes 5 and 10 indicate partial homology (see also **Supplementary Fig. 9**). Black marks at chromosome ends correspond to those in **Figure 1b**.

(involving extant chromosomes 1, 2 and 14), six deletions defined by telomeres that are not currently duplicated (chromosomes 4, 6, 8, 10, 11 and 13), one intrachromosome deletion (within chromosome 7, according to the chromosome 1–chromosome 7 comparison) and a deletion of a centromere (from ancient chromosome IX).

Molecular distances, taxonomy and phylogeny of Rosaceae

Available Rosaceae molecular data allow intrafamily comparisons of apple with pear and of a consensus of apple and pear with peach. Further comparisons with grape—a species basal to rosids but belonging to the Vitaceae, a strictly different, although related, family—introduce the possibility of comparing interfamily molecular distances. DNA sequences used in this molecular phylogeny consist of those from EST databases and, for apple, the genomic data as described in detail in the **Supplementary Note**. Data from a three-way sequence alignment between predicted gene space in apple (~84 Mb) and experimentally derived EST data from pear (~14.9 Mb) and peach (~18 Mb), performed as in ref. 11, indicates that the genetic distance, based on DNA sequence divergence per base pair between members of Rosaceae, increases from apple to pear to peach (**Supplementary Table 20**). When predicted gene spaces of apple and pear were compared, a value of 96.35% nucleotide identity was calculated between these two species of the tribe Pyreae. The estimate for nucleotide identity between the tribes Pyreae and Amygdaleae (apple and peach) was 90.64%. When grape was compared with apple and pear, nucleotide identity was estimated at 85.31%. When the frequency of transitions and transversions was considered (**Fig. 4**), the ratio *R* (transitions/transversions) was similar for apple-specific and pear-specific mutations. For peach-specific mutations, the *R* value is more difficult to interpret, as it is probably biased by the existence of recent GWD in apple and pear. The comparison of apple and pear with grape showed that although transitions were only 20% more frequent than transversions, T-to-G transversions represented 12% of the total number of mutations observed (**Fig. 4d**), implying that Vitaceae is strongly divergent taxonomically from core members of the Rosaceae.

The granule-bound starch synthase (*gbss*) genes, also known as waxy (*Wx*) genes (divided in two groups, *Wx1* and *Wx2*), were also used[4] as a tool to study molecular taxonomy of Rosaceae (**Supplementary Table 21**). We identified six *Wx* genes in the apple genome, located on chromosomes 7, 9 and 16 (*Wx1* type) and 8, 6 and 14 (*Wx2* type) (**Supplementary Fig. 10**). After counting *Wx* genes of apple, including putative gene losses in syntenic chromosomal segments, we were able to identify eight two-by-two syntenic regions containing or expected to contain *Wx* loci. If *Wx1-1* on chromosome 7 is not considered (because neither a syntenic *Wx-1-1* region nor a paralogous *Wx-1-1* copy was found), four *Wx* loci should have been present in the nine-chromosome Pyreae ancestor, a result that is consistent with an ancestral paleopolyploid state. When the genomic *Wx* gene sequences were integrated in the phylogenetic analysis based on sequences present in the Rosaceae database[12], the three *Wx-1* and the three *Wx-2* genes were mapped to two separate clades, both of which also included *Wx* genes of *Gillenia* (**Supplementary Fig. 11**). However, *Prunus* and *Spiraea* sequences clustered in separate clades, supporting the conclusions that the tight relationships between apple and *Gillenia Wx1* genes, as well as between apple and *Gillenia Wx2* genes, were probably generated by the recent GWD (the Pyreae event)—the founding step of the Pyreae genome—and that *Prunus*- and *Spiraea*-related species are less likely to have contributed to the Pyreae genome. Hence, we tested the Rosaceae molecular taxonomy[12] by Bayesian analysis of the sequences of seven nuclear and chloroplast genes. A major clade with the maximum statistical support included all Pyreae ($x = 17$) as well as *Gillenia* ($x = 9$) (**Supplementary Fig. 12**). Notably, the genera *Spiraea* ($x = 9$) and *Prunus* ($x = 8$) were not included in this clade.

Apple domestication

Although *M. sieversii* has been considered to be the ancestor of the domesticated apple[13], this has been challenged by the identification of molecular similarities between domestic apple and *M. sylvestris*[14]. To test these two hypotheses, we surveyed molecular differences at 23 genes across the genus *Malus* (**Supplementary Table 22**). The 74 accessions we considered included 12 *M.* × *domestica* cultivars, 10 *M. sieversii*, 21 *M. sylvestris*, all major wild apple species and two *Pyrus* species (**Supplementary Table 23**). For *M.* × *domestica*, we included the cultivars 'Cox's Orange Pippin', 'Golden Delicious', 'McIntosh', 'Red Delicious' and 'Jonathan', the most important 'founders' of

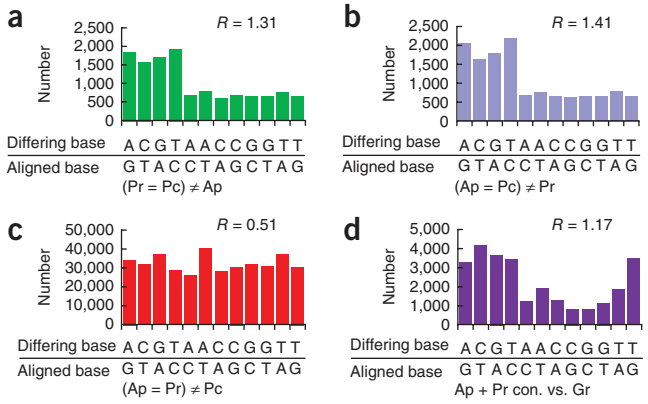

Figure 4 Molecular distances among Rosaceae species and their comparison with grape. (**a**–**c**) Mutations identified in a three-way comparison of apple, pear and peach. Numbers of transitions and transversions where apple (Ap) differs from pear (Pr) and peach (Pc) (**a**; 12,273 total), where pear differs from apple and peach (**b**; 13,124 total), and where peach differs from apple and pear (**c**; 381,619 total). (**d**) Number of transitions and transversions in a two-way alignment of grape DNA (Gr) to a consensus sequence (con.) of apple and pear (26,693 total). Note the high rate of T-to-G transversions. *R* is the transitions/transversions ratio. Methods and computer calculations were similar to those in ref. 11 (**Supplementary Note**).

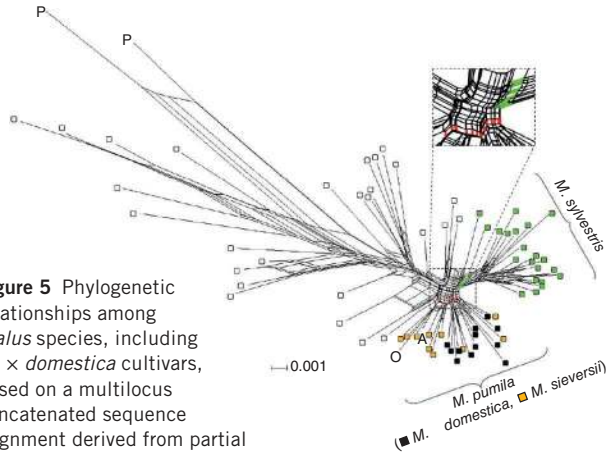

Figure 5 Phylogenetic relationships among *Malus* species, including *M.* × *domestica* cultivars, based on a multilocus concatenated sequence alignment derived from partial resequencing of 23 apple genetic loci. Black, orange and green, accessions of *M.* × *domestica* cultivars, *M. sieversii* and *M. sylvestris*, respectively; A, O and P, accessions of *M.* × *asiatica*, *M. orientalis* and *Pyrus*, respectively; squares, accessions of all other wild species. Full information on accessions is provided in **Supplementary Table 23** and **Supplementary Figure 13**. The split separating the *M.* × *domestica*–*M. sieversii*–*M. orientalis*–*M.* × *asiatica* complex from other species is highlighted in red, and the split separating *M. sylvestris* is highlighted in green. Genetic distances were obtained as Hamming distances, with pairwise alignment of nucleotide positions. The planar graph was constructed with Splits-Tree 4.10.

modern apple breeding[15] (**Supplementary Note**). For each gene and accession, a PCR amplicon was resequenced and the data were analyzed as a concatenated data set with a total length of ~11,300 bp, with 1,507 polymorphic informative sites. A neighbor-net planar graph[16] was constructed from the molecular differences among accessions (**Fig. 5** and **Supplementary Fig. 13**). Although the clade containing *M. sylvestris* was well separated from the clade with *M.* × *domestica*, *M. sieversii* and *M.* × *domestica* genotypes shared a large common clade that also included accessions of *M. orientalis* and *M.* × *asiatica*. The average polymorphism rate within the domestic cultivars was 4.8 SNPs per kb, with 5.7 SNPs per kb between 'Golden Delicious' and *M. sieversii*, and 9.6 SNPs per kb between and *M. sylvestris* (**Supplementary Table 24**). The genetic differentiation was categorized as 'moderate' between *M.* × *domestica* and *M. sieversii* (F_{st} = 0.14), and 'great' between *M.* × *domestica* and *M. sylvestris* and between *M. sieversii* and *M. sylvestris* (F_{st} = 0.17 and F_{st} = 0.21, respectively)[17]. The mean numbers of haplotypes per gene were 6.4, 5.8 and 10.0 for *M.* × *domestica*, *M. sieversii* and *M. sylvestris*, respectively (**Supplementary Table 25**).

DISCUSSION

The putative gene content in apple (57,386 putative genes plus 31,678 transposable element–related ORFs) is high compared to *Arabidopsis thaliana* (27,228), poplar (45,654), papaya (28,027), *Brachypodium distachyon* (25,532), grape (33,514), rice (40,577), sorghum (34,496), cucumber (26,682), soybean (46,430) and maize (32,540). Putative apple-specific genes, identified as described in **Supplementary Note**, totaled 11,444. The gene density in apple (**Table 2**) is within the range of those in poplar and grape, but lower than those in *Arabidopsis*, *Brachypodium* and rice. The existence of hemizygous DNA in the heterozygous variety 'Golden Delicious' may have contributed to this gene number, as has also been noted for grape[2].

The apple genome has a relatively high number of repeated sequences, which are difficult to assemble or anchor. As seen in grape and cereals, retrotransposons represent the most abundant transposable-element fraction, comprising 38% of the total genome and 89% of all transposable elements (**Table 2** and **Supplementary Table 7**). In contrast, apple has the lowest content of DNA transposons (including the CACTA superfamily) among the reported plant genomes.

The number of transcription factors identified (4,021; **Supplementary Table 7**) was among the highest of the sequenced plant genomes (**Table 2**), although the allocation of transcription factor genes to gene families was similar to other sequenced plant species (**Supplementary Fig. 3**). Partial exceptions were the families *C2H2*, *CCAAT* and *NAC*, which were notably more represented in apple.

The fraction of nucleotide-binding site–leucine-rich repeat (*NBS-LRR*) resistance genes is considerably higher in eurosids II (apple, poplar and grape) than in eurosids I (*Arabidopsis*). In monocotyledons (rice), this class of genes predominates. The content of Toll/interleukin region (*TIR*)-*NBS-LRR* genes is highest in *Arabidopsis* (52%), lower in other eurosids (11–32%) and absent in monocots (**Table 2**). In addition to *NBS* genes, the apple genome contains 575 *LRR-kinase* genes.

As seen in other genomes, different classes of apple genes differ greatly in their degree of duplication (**Supplementary Table 11** and **Supplementary Fig. 4**). Across the ten genomes considered, there are gene families with either low or high numbers of paralogous copies. This is particularly evident for genes likely to be involved in metabolism of anthocyanins and flavonoids, isoflavones and isoflavonones, and terpenes (**Supplementary Table 7**). Relevant cases in each pathway are *flavonone 3-hydroxylase* (2–13 copies in nine plant genomes) and *isoflavone reductase* (3–19 copies) compared to *isoflavone synthase* (54–151 copies); *squalene synthase* (13 copies) compared to *squalene monooxygenase* (1–27 copies). It seems that, for some gene classes, the number of paralogous copies may already have been established in the genome of common progenitor(s) of higher plants.

An intriguing aspect of the apple's biology concerns its characteristic fruit, the pome, which is found only in the Pyreae tribe[12]. This indicates that the pome probably evolved after a relatively recent Pyreae-specific GWD, a polyploidization step that we hypothesize has contributed to the apple's developmental and metabolic specificity (**Supplementary Table 7**). Pome fruit is derived by enlargement of the receptacle, which is the region below the whorl of sepals in the apple flower. *MADS-box* genes may regulate pome development, as they determine the eventual fate of floral tissues in all plant species analyzed so far[18]. For example, it has recently been shown that an apple *MADS-box* gene that is a member of the *AP1* clade, common to all flowering plants[19] and closely related to *Arabidopsis FRUITFULL* (*FUL*), is differentially expressed during pome development[20]. In addition, a substantial number of apple type II *MADS-box* genes belong, phylogenetically, to the *StMADS11* subclade, a group named for its first reported member, which was isolated from potato (**Supplementary Fig. 14a**)[21]. This subclade includes only two *Arabidopsis* genes, *SVP* and *AGL24*. Ectopic overexpression of *SVP* and related genes in *Arabidopsis* leads to foliose sepal syndrome—that is, the formation of large sepals[22]. In apple, this specific subclade not only includes two genes expressed in the pome but is also expanded to include 15 other genes.

Carbohydrate metabolism is another important aspect of fruit composition. In Rosaceae, photosynthesis-derived carbohydrates are transported mainly as sorbitol[23,24]. Compared with other plant genomes, apple has considerably more copies of key genes related to sorbitol metabolism. These include *aldose 6-P reductase* (*A6PR*), which is rate-limiting for sorbitol biosynthesis, *sorbitol-dehydrogenase* (*SDH*), which converts sorbitol to fructose in the fruit[25], and *sorbitol transporter PcSOT2*, which is specific to Rosaceae fruit[26,27]. In total,

there are 71 sorbitol metabolism genes in apple; in other species, the number ranges between 9 and 43 (**Supplementary Tables 7** and **26**, and **Supplementary Fig. 14b–d**). In the Rosaceae, an evolutionary trend toward fruit organ specialization may have been partially based on gene duplication, which has created large families of specific paralogous genes (particularly evident for *SDH*; **Supplementary Fig. 14c**). Gene families expanded in apple, such as *StMADS11*-like and *SDH*-like, have yet to be tested functionally for their involvement in fruit characteristics.

A number of models have been proposed to explain the uniquely high number of chromosomes in Pyreae, the most popular being the 'wide-hybridization' hypothesis based on an allopolyploidization event between spireoid ($x = 9$) and amygdaloid ($x = 8$) ancestors[28,29]. More recent molecular phylogeny studies point to the possibility that Pyreae originated by autopolyploidization or by hybridization between two sister taxa with $x = 9$ (similar to extant *Gillenia*), followed by diploidization and aneuploidization[4] to $x = 17$. This hypothesis takes into account that *Gillenia* and related taxa are New World species and that the earliest fossil evidence of specimens belonging to extant genera of Pyreae are from North America.

Our results support the autopolyploidization hypothesis[4], as the derivation from a *Gillenia*-like taxon best fits the available data. First, the apple genome derives from a relatively recent duplication. Relationships between its homologous chromosomes based on genome sequence extend observations based on synteny and collinearity of molecular markers[30,31]. The timing of such a GWD, as estimated from our genomic data (**Fig. 1c** and **Supplementary Figs. 15** and **16**), agrees with archeobotanical dates of 48–50 Mya[32].

Second, molecular phylogeny of *Wx* genes in the apple genome confirms the close relationship of *Gillenia* ($x = 9$) with the Pyreae ($x = 17$) lineage, as the *Wx* gene sequences of *Prunus*, *Spiraea* and other Rosaceae genera belong to a different phylogenetic cluster (**Supplementary Fig. 11**). The monophyletic origin of Pyreae and *Gillenia* was confirmed by a molecular phylogeny of a broader set of genes (**Supplementary Fig. 12**).

In addition, a simple and parsimonious pattern of chromosome breakage and fusion explains the derivation of the current $x = 17$ Pyreae karyotype from a polyploidization event of two $x = 9$ genomes (**Fig. 3**). The rate of chromosome rearrangements after polyploidization (12 chromosome events in 60 My) is similar to that for poplar (~16 events in 60 My)[6] and lower than in maize (at least 17 chromosome fusion events in 5 My)[33] or in artificial neopolyploids[34]. In this sense, molecular clocks of perennial woody species seem slower than those of annual species, in terms of both nucleotide substitutions and chromosome rearrangements[9]. For the genus *Helianthus*, a similar observation that only some of the ancestor chromosomes are rearranged in the extant chromosomes has been discussed in detail. In this genus, such rearrangement was associated with chromosomal differences between two sister species contributing to a GWD allopolyploid event[35].

Similarly, the collinearity between *Pyrus* and *Malus* genetic maps[31,36] suggests that the Pyreae genome reorganization occurred before the divergence of the two genera. A rapid genome rearrangement after polyploidization is expected in species lacking the *Ph1*-like function that prevents the pairing of homologous chromosomes in wheat[37].

It has been proposed that central Asia is the center of origin of domesticated apple[38]. Between 25 and 47 different *Malus* species, including *M.* × *domestica*, are currently recognized[39]. As asiatic *M.* × *asiatica*, *M. baccata*, *M. micromalus*, *M. orientalis*, *M. prunifolia* and *M. sieversii*, and European *M. sylvestris*, are the species taxonomically closest to *M.* × *domestica*[39], they are considered to have contributed, to differing extents, to the domestic gene pool. *M. sieversii*, common in the Tian Shan region of central Asia, is the only wild species sharing all the qualities of the domesticated apple in terms of fruit and tree morphology[40].

Apples are known to have been gathered in the Neolithic and Bronze Age in the Near East and Europe, and all archaeological findings indicate a fruit size compatible with those of the wild *M. sylvestris*[41], a species bearing small astringent and acidulate fruits. Sweet apples corresponding to extant domestic apples appeared in the Near East around 4,000 years ago[41], at the time when the grafting technology used to propagate the highly heterozygous and self-incompatible apple was becoming available. From the Middle East, the domesticated apple passed to the Greeks and Romans, who spread fruit cultivation across Europe[13,41].

On the basis of our molecular results, *M.* × *domestica* cultivars appear more closely related to accessions of the wild species *M. sieversii* and less closely related to accessions of *M. sylvestris*, *M. baccata*, *M. micromalus* and *M. prunifolia*. The already known[42,43] genetic similarity of *M. sieversii* to *M. orientalis* and to *M.* × *asiatica* (a Chinese cultivated apple form) is also confirmed by our data.

The data support the formation of the *M.* × *domestica* gene pool from *M. sieversii*. Once grafting was introduced, the crop passed through a process described as 'instant domestication'[44]. This could explain apple's lack of domestication syndrome, which is the loss of sexual reproduction, seed dispersion and seed dormancy. Despite evidence of intrageneric hybridizations[14,45], the possibility of substantial genetic contributions to the domestic gene pool of other wild *Malus* species, such as *M. sylvestris*[14], was rejected in our analysis.

Our study also fully supports the proposal that *M.* × *domestica* and *M. sieversii* are the same species, for which the more appropriate nomenclature of *M. pumila* Mill. could be adopted[13,46].

A practical goal of sequencing the complex heterozygous apple genome is to accelerate the breeding of this economically important perennial crop species. Many genes related to disease resistance, aroma and taste, plant development and reaction to the environment have been identified and mapped to the chromosomes. In addition, SNP molecular markers have been made available at a frequency of 4.4 SNPs per kb. These markers are currently being used in advanced breeding programs and comparative genetic studies[31] that should speed cultivar development. The anchored sequence of the apple genome will be a tool to initiate a new era in the breeding of this crop. The availability of nearly all apple gene sequences should benefit apple researchers by enabling genome-wide functional studies and accelerating establishment of gene-trait relationships.

URLs. *Arabidposis thaliana* (TAIR Release 8.0), ftp://ftp.arabidopsis.org; *Carica papaya*, ftp://asgpb.mhpcc.hawaii.edu/papaya/annotation/; *Populus trichocarpa* (assembly release v1.0, annotation v1.1.), http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html; *Vitis vinifera* (assembly release v1.0, annotation v2.0), http://genomics.research.iasma.it; *Oryza sativa* (MSU Rice Genome Annotation Project Release 6.0 assembly), http://rice.plantbiology.msu.edu/index.shtml; *Sorghum bicolor* (assembly release v1.0, annotation v1.4), http://www.phytozome.net/sorghum; *Cucumis sativus* (assembly release v1.0, annotation v1.0), http://cucumber.genomics.org.cn/page/cucumber/index.jsp; *Glycine max* (assembly release Glyma1, annotation Glyma1.0), http://genome.jgi-psf.org/soybean/soybean.home.html; *Zea mays* (assembly release B73 RefGen_v1), http://www.maizesequence.org/Zea_mays/Info/Index; *Brachypodium distachyon* (assembly release v1.0, annotation v1.0), http://www.brachybase.org; integrated genetic map, http://genomics.research.iasma.it; RepBase14.01, http://www.girinst.org.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The Italian apple genome project was supported by the research office of the Provincia Autonoma di Trento. The US apple genome project was supported by Washington State University Agriculture Research Center, Washington Tree Fruit Research Commission and US Department of Agriculture National Research Initiative (USDA-NRI) grant 2008 –35300-04676 to A.D., A.K. and R.E.B. V.K. and C.W. received support from the USDA-NRI grant. S. Schaeffer and T.K. were supported by the US National Institutes of Health Protein Biotechnology Training Program and an Achievement Rewards for College Scientists fellowship. A.C.A., V.B., D.C., A.P.G., S.E.G., R.P.H. and R.N.C. were partially supported by the New Zealand Foundation for Research Science and Technology, contract no. C06X0812. We thank S. Attiya, E. Buglione and C. Celone from 454 Life Sciences-Roche Company as well as E. Stefani, A. Castelli and E. Potenza for technical support and V. Sgaramella for critical reading of the manuscript. Fosmid and shotgun libraries were prepared following the method developed by R. Meilan (Oregon State University).

AUTHOR CONTRIBUTIONS

R. Velasco, A.D., A.K., R.E.B., M.S., M.E., F.S. and R. Viola managed the project. R. Velasco, A.Z., J.A., A.D., A.C., A.K., P.F., S.K.B., M.T., D.P., S. Salvi, J.L., A.G., R.E.B., S.E.G., M.S., M.E., Y.V.d.P. and F.S. designed the analyses. J.A., S.K.B., D.P., M. Pindo, G.C., D.M., G.E., L.M.F., N.G., T.M., J.T.M., J.R., B.W., C.K., Z.C., B.D., F.N., M. Palmer, T.K., D.J., S. Schaeffer, V.T.C., S.T.K., J.V., Q.T., A.M., A. Stormo, K.S. and R.B. performed DNA preparation and sequencing. R. Velasco, A.Z., A.D., A.C., A.K., P.F., M.T., D.P., P.B., V.C., A.D.R., M.K., P.M., D.M., P.L., Y.L., V.B., D.C., R.N.C., S.T., C.-E.D., A.G., R.E.B. and S.E.G. contributed to sequence assembly and anchoring to chromosomes. A.Z., A.D., A.C., A.K., P.F., M.T., S. Salvi, M. Pindo, S.C., M.C., F.C., V.G., S.L., G.M., M. Malnoy, D.M., M. Moretto, M. Perazzolli, A.S.-A., S.V., E.Z., V.K., C.W., D.E., A. Stella, A.V., M.M.K., S.M., A.C.A., R.N.C., A.P.G., E.L., J.A.F., S.P., P.R., L.S., S.T., B.L., R.P.H., Y.V.d.P. and F.S. contributed to automatic and manual genome annotation, genome structure and evolution analyses. R. Velasco, A.Z., J.A., A.D., A.K., P.F., M.T., S. Salvi, F.C., M. Malnoy, A.S.-A., S.V., R.E.B., A.V., S.M., J.A.F., L.S., S.T., B.L. and F.S. wrote the paper. M.M.K., A.C.A., R.N.C., A.P.G., R.P.H., C.-E.D., A.G., R.E.B., S.E.G., M.S., Y.V.d.P. and R. Viola revised the paper. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Zharkikh, A. *et al.* Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: problems and solutions. *J. Biotechnol.* **136**, 38–43 (2008).
2. Velasco, R. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
3. Hummer, K.E. & Janick, J. Rosaceae: taxonomy, economic importance, genomics. in *Genetics and Genomics of Rosaceae* (eds. Folta, K.M. & Gardiner, S.E.) 1–17 (Springer, New York, 2009).
4. Evans, R.C. & Campbell, C.S. The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. *Am. J. Bot.* **89**, 1478–1484 (2002).
5. Fawcett, J.A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. USA* **106**, 5737–5742 (2009).
6. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
7. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
8. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486 (2008).
9. Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends Plant Sci.* **14**, 680–688 (2009).
10. Luo, M.C. *et al.* Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl. Acad. Sci. USA* **106**, 15780–15785 (2009).
11. Green, R.E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).
12. Potter, D. *et al.* Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* **266**, 5–43 (2007).
13. Juniper, B.E. & Mabberley, D.J. *The Story of the Apple* (Timber Press, Portland, Oregon, USA, 2006).
14. Coart, E., Van Glabeke, S., De Loose, M., Larsen, A.S. & Roldan-Ruiz, I. Chloroplast diversity in the genus *Malus*: new insights into the relationship between the European wild apple (*Malus sylvestris* (L.) Mill.) and the domesticated apple (*Malus domestica* Borkh.). *Mol. Ecol.* **15**, 2171–2182 (2006).
15. Noiton, D.A.M. & Alspach, P.A. Founding clones, inbreeding, coancestry, and status number of modern apple cultivars. *J. Am. Soc. Hortic. Sci.* **121**, 773–782 (1996).
16. Huson, D.H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
17. Wright, S. *Evolution and the Genetics of Populations* Vol. 4 (University of Chicago Press, 1978).
18. Ng, M. & Yanofsky, M.F. Activation of the *Arabidopsis* B class homeotic genes by APETALA1. *Plant Cell* **13**, 739–753 (2001).
19. Shan, H. *et al.* Evolution of plant MADS Box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Mol. Biol. Evol.* **26**, 2229–2244 (2009).
20. Janssen, B.J. *et al.* Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. *BMC Plant Biol.* **8**, 16 (2008).
21. Becker, A. & Theissen, G. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.* **29**, 464–489 (2003).
22. Masiero, S. *et al. INCOMPOSITA*: a MADS-box gene controlling prophyll development and floral meristem identity in *Antirrhinum*. *Development* **131**, 5981–5990 (2004).
23. Newcomb, R.D. *et al.* Analyses of expressed sequence tags from apple. *Plant Physiol.* **141**, 147–166 (2006).
24. Bieleski, R.L. Sugar alcohols. in *Encyclopedia of Plant Physiology New Series* Vol. 13 (eds. Loewus, F.A. & Tanner, W.) 158–192 (Springer-Verlag, Berlin, 1982).
25. Loescher, W.H., Marlow, G.C. & Kennedy, R.A. Sorbitol metabolism and sink-source interconversions in developing apple leaves. *Plant Physiol.* **70**, 335–339 (1982).
26. Watari, J. *et al.* Identification of sorbitol transporters expressed in the phloem of apple source leaves. *Plant Cell Physiol.* **45**, 1032–1041 (2004).
27. Gao, Z. *et al.* Cloning, expression, and characterization of sorbitol transporters from developing sour cherry fruit and leaf sink tissues. *Plant Physiol.* **131**, 1566–1575 (2003).
28. Chevreau, E., Lespinasse, Y. & Gallet, M. Inheritance of pollen enzymes and polyploid origin of apple. (*Malus* x *domestica* Borkh.). *Theor. Appl. Genet.* **71**, 268–277 (1985).
29. Phipps, J.B., Robertson, K.R., Rohrer, J.R. & Smith, P.G. Origins and evolution of subfam. Maloideae (Rosaceae). *Syst. Bot.* **16**, 303–332 (1991).
30. Maliepaard, C. *et al.* Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers. *Theor. Appl. Genet.* **97**, 60–73 (1998).
31. Celton, J.M., Tustin, D.S., Chagne, D. & Gardiner, S.E. Construction of a dense genetic linkage map for apple rootstocks using SSRs developed from *Malus* ESTs and *Pyrus* genomic sequences. *Tree Genet. Genomes* **5**, 93–107 (2009).
32. Wolfe, J. & Wehr, W.J.A. Rosaceous *Chamaebatiaria*-like foliage from the Paleogene of western North America. *Aliso* **12**, 177–200 (1988).
33. Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24 (2008).
34. Doyle, J.J. *et al.* Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461 (2008).
35. Rieseberg, L.H., Sinervo, B., Linder, C.R., Ungerer, M.C. & Arias, D.M. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. *Science* **272**, 741–745 (1996).
36. Yamamoto, T. *et al.* Genetic linkage maps of Japanese and European pears aligned to the apple consensus map. *Acta Hortic.* **663**, 51–56 (2004).
37. Griffiths, S. *et al.* Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749–752 (2006).
38. Vavilov, N.I. Wild progenitors of the fruit trees of Turkestan and the Caucasus and the problem of the origin of fruit trees. in *Proceedings of the 9th International Horticultural Congress* 271–286 (The Royal Horticultural Society, London, 1930).
39. Robinson, J.P., Harris, S.A. & Juniper, B.E. Taxonomy of the genus *Malus* Mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* Borkh. *Plant Syst. Evol.* **226**, 35–58 (2001).
40. Forsline, P.L., Aldwinckle, H.S., Dickson, E.E., Luby, J.J. & Hokanson, S.C. Collection, maintenance, characterization, and utilization of wild apples of Central Asia. *Hortic. Rev. (Am. Soc. Hortic. Sci.)* **29**, 1–61 (2003).
41. Zohary, D. & Hopf, M. *Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Asia, Europe and the Nile Valley* (Clarendon Press, Oxford, 1994).
42. Gharghani, A. *et al.* Genetic identity and relationships of Iranian apple (*Malus* x *domestica* Borkh.) cultivars and landraces, wild *Malus* species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genet. Resour. Crop Evol.* **56**, 829–842 (2009).
43. Luby, J. Taxonomy, classification and brief history. in *Apples: Botany, Production and Uses* (eds. Ferree, D.C. & Warrington, I.J.) 1–14 (CABI, Cambridge, Massachusetts, USA, 2003).
44. Zohary, D. & Spiegelroy, P. Beginnings of fruit growing in the Old World. *Science* **187**, 319–327 (1975).
45. Korban, S.S. Interspecific hybridization in *Malus*. *HortScience* **21**, 41–48 (1986).
46. Korban, S.S. & Skirvin, R.M. Nomenclature of the cultivate apple. *HortScience* **19**, 177–180 (1984).

ONLINE METHODS

Plant material. The DNA of *Malus* × *domestica*, variety 'Golden Delicious', was extracted from young leaves of a two-year-old plant grown in the greenhouse at Fondazione Edmund Mac–Istituto Agrario di San Michele all'Adige. The dihaploid 'Golden Delicious' derivative genotype used at Washington State University and the University of Washington to produce 1.5× of 454 sequence was developed by the French National Institute for Agricultural Research[47] after a spontaneous duplication of a haploid individual selected in the progeny of a selfed derivative from 'Golden Delicious'[47].

'Golden Delicious' was chosen for genome sequencing because of its extensive use in apple breeding programs worldwide. Its heterozygous status did not hamper the genome assembly, thanks to expertise gained in heterozygous grape sequencing[1]. Indeed, it allowed the inference of both haplotypes, thus giving access to both allelic versions for further genomic projects, and the development of SNP markers. The dihaploid genotype was important for a more accurate haplotype phase determination.

Bacterial artificial chromosomes, shotgun libraries and Sanger sequencing. The apple bacterial artificial chromosome (BAC) library was from high–molecular weight genomic DNA (Amplicon Express), prepared as described[48]. The fosmid and shotgun libraries were from genomic DNA provided by R. Meilan (Oregon State University). The shotgun libraries were from DNA sheared with a Gene Machines Hydroshear device. The DNA was size-selected for inserts from 2 to 12 kb to produce libraries of 2, 3, 6, 9 and 11 kb (average sizes). DNA was amplified with the Templiphi kit (GE Healthcare) and sequenced with the Sanger method.

Libraries and 454 pyrosequencing. Two random shotgun genomic libraries were created by fragmentation of 10 µg of genomic DNA with the GS FLX Titanium library preparation kit (454 Life Sciences). Sequencing was performed with the GS FLX instrument (454 Life Sciences). Further details on library construction and pyrosequencing are in the **Supplementary Note**.

Genome assembly and anchoring. From 27 libraries, 39.2 million reads (11.6 billion Q20 bases) were produced by Sanger sequencing and sequencing by synthesis (**Supplementary Table 1**). Chloroplast and mitochondrial sequences were identified with 847× and 168× coverage, respectively. Chloroplast (160,068 bp) and mitochondrial (396,947 bp) genomes were used to assess sequence quality and clone size in each library. Preliminary estimates of one to two SNPs per 1,000 bp were adopted in the assembly process. The actual SNP rate (4.4 SNPs per 1,000 bp) indicates that the preliminary value was conservative. Metacontigs were constructed on the basis of paired reads matching to nonrepetitive parts of contigs. Merging of contigs into metacontigs accepted a maximum total average coverage of 20×. Fifteen BAC clones were sequenced and individually assembled for quality assessment of sequencing accuracy and genome assembly.

Genetic maps used in metacontig anchoring were derived from six F_1 populations totaling 720 individuals (**Supplementary Note**). Simple sequence repeat primer sequences[49–52] enabled detection of 196 polymorphic markers. Thirty-four SNP-based markers were from apple EST sequences, and 1,489 from genomic electronic SNPs, deduced by genomic sequence comparison between the two haplotypes present in the heterozygous genotype of 'Golden Delicious'. The consensus genetic maps for the six populations were used to generate an integrated genetic map (**Supplementary Fig. 1**) with TMAP[53] and a minimum logarithmic odds of 10.

Repetitive elements. The highest-coverage sequences were characterized as repetitive elements. Identified elements were iteratively masked, and the remaining sequences were searched for the next highest–coverage sequence. For each type, members were searched (BLASTN and BLASTX) against RepBase14.01, the NCBI databases and the Uniprot database[54,55].

Gene prediction and annotation. FgenesH[56], Twinscan[57], GlimmerHMM[58] and GeneWise[59] were used. The predicted protein sequences were searched with BLAST against Uniprot, protein domain data banks and plant protein databases annotated with GO terms. The GO terms were extracted by Argot[60] and InterproScan[61]. Unique genes were searched against proteins from rice, poplar, papaya, barrel medic, sorghum, *Arabidopsis* and grape by BLAST with an *e*-value cutoff of e^{-10}.

All-versus-all BLAST. Protein sequences from apple, poplar[6] and grape[2] were extracted from a BLAST database, and pairwise similarities between all genes was obtained by BLASTP *e*-value (cutoff e^{-5}; 500 hits)[62].

Gene families. Tribe-MCL[63] was adopted, with parameter *I* set to 2 and parameter 'scheme' to 4; other parameters were at default values.

Detection of collinearity. Metacontig anchoring generated lists of apple genes, from which transposable element–related sequences were removed. Poplar and grape gene lists were as described[2]. Colinearity in the gene order was detected with i-ADHoRe 2.4 (ref. 64), with the following parameters: family blast type; alignment method, gg; gap size, 30; cluster gap, 35; *q* value, 0.9; prob cutoff, 0.0001; anchor points, 4; level 2 only, false.

K_s dating. Homologous genes were aligned with CLUSTALW[65]. K_s dating was based on codeml[66] with the following parameters: verbose, 0; noisy, 0; runmode, –2; seqtype, 1; model, 0; NSsites, 0; icode, 0; fix_alpha, 0; fix_kappa, 0; RateAncestor, 0.

Molecular distances, taxonomy and phylogeny. Molecular distances were analyzed with EST data sets. A two-way alignment between apple and pear contigs (cDNA sequences, data not shown) was first generated. Sequences from apple and pear were combined with the peach sequence (EST databases) in three-way alignments. Phylogenetic analysis of the *Wx* genes included *gbss1* (*Wx1*) and *gbss2* (*Wx2*) sequences from the apple genome and from ref. 12. Sequences were aligned by T-coffee[67], and phylogenesis was by a Bayesian inference approach (MrBayes program). Phylogeny of Rosaceae was based on four chloroplast DNA sequences and on the nuclear internal transcribed spacer region. The data set included 6,308 positions in 85 operational taxonomic units, each representing one genus, aligned by a Bayesian method[68].

Apple domestication. A set of 74 *Malus* accessions, including 12 accessions of *M.* × *domestica* cultivars[15], 10 of *M. sieversii* and 21 of *M. sylvestris*, was assembled. This included 31 of 34 recognized *Malus* species[69]. Twenty-three genes were resequenced and, after alignment[67], a concatenated 11,300-bp multilocus sequence was generated for each accession. Genetic relationships analysis used Splits-Tree v4.10 (ref. 16) and Hamming distance per pair of accessions. Haplotypes were computed with Phase v2.1 (ref. 70). Nucleotide diversity (π), H_e, H_o and F_{st} values[17] were computed with Arlequin 3.1 (ref. 71).

47. Lespinasse, Y., Bouvier, L., Djulbic, M. & Chevreau, E. Haploidy in apple and pear. *Acta Hortic.* **538**, 49–54 (1999).
48. Tao, Q., Wang, A. & Zhang, H.B. One large-insert plant-transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses. *Theor. Appl. Genet.* **105**, 1058–1066 (2002).
49. Guilford, P. *et al.* Microsatellites in *Malus* X *domestica* (apple): abundance, polymorphism and cultivar identification. *Theor. Appl. Genet.* **94**, 249–254 (1997).
50. Liebhard, R. *et al.* Development and characterisation of 140 new microsatellites in apple (*Malus* x *domestica* Borkh.). *Mol. Breed.* **10**, 217–241 (2002).
51. Gianfranceschi, L., Seglias, N., Tarchini, R., Komjanc, M. & Gessler, C. Simple sequence repeats for the genetic analysis of apple. *Theor. Appl. Genet.* **96**, 1069–1076 (1998).
52. Silfverberg-Dilworth, E. *et al.* Microsatellite markers spanning the apple *(Malus* x *domestica* Borkh.) genome. *Tree Genet. Genomes* **2**, 202–224 (2006).
53. Cartwright, D.A., Troggio, M., Velasco, R. & Gutin, A. Genetic mapping in the presence of genotyping errors. *Genetics* **176**, 2521–2527 (2007).
54. Uniprot Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* **36**, D190–D195 (2008).
55. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
56. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, S10 (2006).
57. Korf, I., Flicek, P., Duan, D. & Brent, M.R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, S140–S148 (2001).
58. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
59. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).
60. Fontana, P., Cestaro, A., Velasco, R., Formentin, E. & Toppo, S. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. *PLoS ONE* **4**, e4619 (2009).

doi:10.1038/ng.654

61. Mulder, N. & Apweiler, R. InterPro and InterProScan—Tools for protein sequence classification and comparison. *Methods Mol. Biol.* **396**, 59–70 (2007).
62. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
63. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
64. Simillion, C., Janssens, K., Sterck, L. & Van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24**, 127–128 (2008).
65. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
66. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
67. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
68. Huelsenbeck, J.P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
69. Janick, J., Cummins, J.N., Brown, S.K. & Hemmat, M. Apples. in *Fruit Breeding* Vol. 1 (eds. Janick, J., Moore, J.J.) 1–77 (Wiley, New York, 1996).
70. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
71. Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1**, 47–50 (2005).

PRDM9 marks the spot

Gil McVean & Simon Myers

A new study demonstrates that *PRDM9* variation in humans leads to profound differences in the activity of hotspots for both allelic recombination and genomic instability. Although *PRDM9* is found to play a role in many more human hotspots than previously suspected, the search remains for additional, undetermined factors involved in defining hotspot locations and intensities.

During mammalian meiosis, crossing over following programmed double strand breaks (DSBs) is critical for the correct segregation of chromosomes into gametes in both sexes and serves to shuffle genetic variation between the two chromosomes. Although DSBs are typically repaired with high fidelity, those occurring at certain genomic locations, particularly within repeat DNA, are prone to errors in repair. Such errors are thought both to drive mutational processes at hypermutable minisatellites and to produce recurrent *de novo* nonallelic homologous recombination (NAHR) events, which are megabase-scale duplications, deletions or rearrangements that are responsible for many genomic disorders. Although the molecular machinery controlling programmed DSB formation during meiosis is highly conserved across eukaryotes, recombination hotspots (particular narrow regions of DNA showing elevated rates of recombination) can vary markedly between taxa[1,2] and even among members of the same species[3,4]. In the last year, the chromatin-modifying protein PRDM9 was identified as a key player in this process, regulating recombination through binding to recombination hotspots in both mice and humans[3–5]. On page 859 of this issue, Alec Jeffreys and colleagues[6] demonstrate that *PRDM9* variation influences crossover activity at individual human recombination hotspots, as well as genome instability both at minisatellites and at pathological NAHR rearrangements, directly linking variation in this gene to human disease.

Recombination hotspots

Most known human recombination hotspots have been initially identified using linkage disequilibrium (LD) data and therefore reflect positions of elevated mean recombination rate in ancient, ancestral human populations. In humans, a degenerate 13-bp motif is highly enriched in these LD-identified recombination hotspots[7] as well as in both hypervariable minisatellites and NAHR hotspots, whose mutational processes are closely tied to meiotic recombination. This motif has been estimated to play a critical role in recruiting crossover events at 40% of hotspots, with evidence that a zinc finger (ZnF)-containing protein is responsible for binding the motif[7]. Experiments in yeast[8] and mouse[9] have also identified a role for epigenetic marks, specifically the trimethylation of lysine 4 in histone 3 (H3K4me3), in specifying hotspot locations. These studies have recently been linked, with the discovery that PRDM9, a meiosis-specific protein with H3K4me3 activity, binds to the 13-bp motif in humans[3,5]. PRDM9 is the unique ZnF protein predicted by a bioinformatic screen[5], and shown *in vitro*[3], to bind the motif. PRDM9 has an array of 13 ZnFs, encoded by a minisatellite, which bind to the motif and which show exceptionally rapid evolution between humans and chimpanzees, consistent with the 13-bp motif not generating hotspots in chimpanzee[6]. This ZnF array shows substantial variation in humans[3,4], and analysis of crossover events in human pedigrees showed that individuals of European ancestry that carry a rare variant of the protein with 16 zinc fingers show very little, if any, use of recombination hotspots defined by LD data[3]. In contrast, in individuals carrying common *PRDM9* alleles predicted to bind the 13-bp motif, the majority of recombination events occur within LD-based recombination hotspots[3], suggesting ancient hotspot activity is strongly linked to these or similar *PRDM9* alleles.

Berg *et al.*[6] now bring a more detailed understanding of the role of *PRDM9* by exploring the effect of *PRDM9* variation at individual human hotspots (**Fig. 1**). First, they carried out a comprehensive survey of *PRDM9* variation within a study sample of 74 African and 156 European semen donors, finding 16 different forms of *PRDM9* containing between 8 and 18 zinc fingers. They next assayed recombination activity in sperm from 15 of these individuals, selected to show variability in the PRDM9 ZnF array. They examined 10 highly active hotspots (as inferred from LD data), including 5 hotspots with the sequence motif. All 10 of these showed activity dependent on PRDM9, a surprising result given that 5 of the examined hotspots contain no clear match

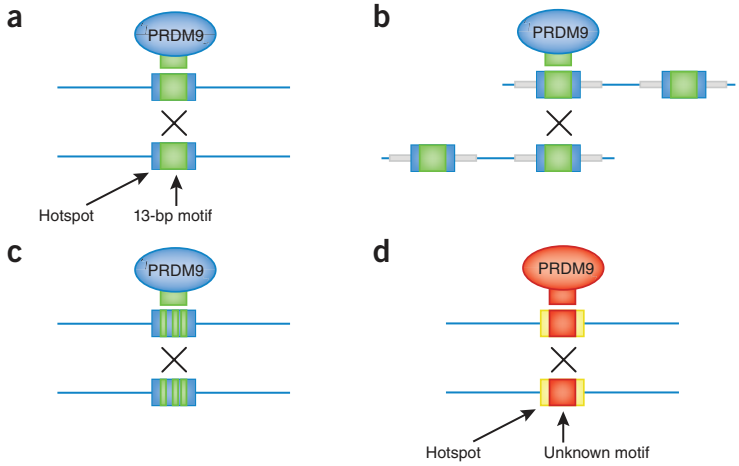

Figure 1 View of how *PRDM9* variation influences both allelic and nonallelic recombination at individual hotspots. The common form of PRDM9 (*PRDM9* allele A, blue circles) activates recombination at a particular class of hotspots (indicated by the blue boxes) at both allelic crossover (**a**) and nonallelic homologous recombination (**b**) hotspots (the gray bars indicate the extended low-copy repeats in which the NAHR breakpoints lie) through binding to a degenerate motif (green bar). Surprisingly, the common form of PRDM9 activates hotspots even when there is no clear motif (**c**), suggesting it can bind to highly degenerate forms (the shattered green bar). (**d**) Berg *et al.*[6] also identified two hotspots activated only by other variants of PRDM9 (red circle), which presumably bind as-yet unidentified motifs in a different class of hotspots (the red box and yellow bar, respectively, in **d**).

Gil McVean and Simon Myers are at The Wellcome Trust Centre for Human Genetics, Oxford, UK.
e-mail: mcvean@stats.ox.ac.uk

to the motif. The results of Berg *et al.* imply that PRDM9 is involved in a much higher fraction of hotspots than the 40% previously estimated to be activated by the motif, and that PRDM9 may be involved in all hotspots. Berg *et al.* further found that these hotspots are primarily activated by the common reference *PRDM9* allele (referred to as allele A). However, in an additional hotspot cluster, one hotspot was activated more strongly by a non-A allele and the other hotspot was activated only by non-A alleles, with subtle amino acid changes within the array strongly altering hotspot activity.

The connections between recombination, minisatellite and NAHR rearrangement events suggest that *PRDM9* variation is likely to influence rearrangement frequencies, at least where the 13-bp motif is present. Berg *et al.*[6] directly tested this prediction, finding that for three hypervariable minisatellites where the repeated element contains a close match to the hotspot motif[7], and for a recurrent NAHR rearrangement where there is also a crossover hotspot likely driven by the hotspot motif[7,10], variation in *PRDM9* has a profound effect on the rate of mutation. In contrast, at a recurrent translocation site where there was no previous evidence for the involvement of recombination or the hotspot motif, *PRDM9* variation had no influence on the mutation rate.

Motifs and modifiers

Berg *et al.*[6] demonstrate that in some cases, PRDM9 may define hotspot location without binding to the known 13-bp hotspot motif. This may suggest that different zinc fingers are required for binding to different hotspots[11]. However, this seems unlikely given the consistent inactivity for almost all non-A *PRDM9* alleles. More likely, PRDM9 may be capable of binding highly degenerate copies of the motif, and it is possible that a favorable flanking sequence[7] strengthens binding to highly degenerate motifs. Consistent with this idea, sequences containing three mismatches to the eight non-degenerate bases in the 13-bp motif occur near the center of each of the five non-motif hotspots examined by Berg *et al.*[6] Another possibility is that PRDM9 does bind in a sequence-specific manner but also exerts H3K4me3 activity in *cis* at some distance from the binding location.

Although PRDM9 is an important player in recombination, additional factors influence binding and hotspot activity. Modifiers of recombination rate have been identified at the genome-wide level[12,13], and there are differences between recombination rates in males and females at megabase scales[14]. Within several hotspots, specific SNPs influencing recombination activity have been identified[15], and these SNPs often occur in sequences lacking the 13-bp motif. Further, previous statistical analysis[7] has identified multiple additional motifs enriched in recombination hotspots but which bear no homology to the predicted PRDM9 binding sequence. For example, one of the motif-containing hotspots studied by Berg *et al.*[6] is centered within a THE1B retrotransposon. On this specific repeat background, the presence of an 8-bp motif, 129 bp upstream from the 13-bp motif and therefore outside the region likely bound by PRDM9, leads to a twofold increase in the average recombination rate[11]. The apparent background specificity of this motif and others suggests that although PRDM9 binding is a shared feature across human hotspots, other gene products are likely to play hotspot-specific roles.

Berg *et al.*[6] confirm that *PRDM9* plays a key role in both allelic recombination and certain forms of genome instability and demonstrate the remarkable effect that variation in one gene can have on specific recombination and mutation events. Furthermore, they raise the question of which, if any, recurrent genomic mutations are activated by individuals lacking the common *PRDM9* allele and whether, because there are strong population differences in *PRDM9* alleles, NAHR disorders may also have large differences in frequency between populations. Future studies are needed to search for the molecular partners of PRDM9 in recruiting recombination and to characterize which of these partners play general versus hotspot-specific roles.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Ptak, S.E. *et al. Nat. Genet.* **37**, 429–434 (2005).
2. Winckler, W. *et al. Science* **308**, 107–111 (2005).
3. Baudat, F. *et al. Science* **327**, 836–840 (2010).
4. Parvanov, E.D., Petkov, P.M. & Paigen, K. *Science* **327**, 835 (2010).
5. Myers, S. *et al. Science* **327**, 876–879 (2010).
6. Berg, I.L. *et al. Nat. Genet.* **42**, 859–863 (2010).
7. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. *Nat. Genet.* **40**, 1124–1129 (2008).
8. Borde, V. *et al. EMBO J.* **28**, 99–111 (2009).
9. Buard, J., Barthes, P., Grey, C. & de Massy, B. *EMBO J.* **28**, 2616–2624 (2009).
10. Lindsay, S.J., Khajavi, M., Lupski, J.R. & Hurles, M.E. *Am. J. Hum. Genet.* **79**, 890–902 (2006).
11. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. *Science* **310**, 321–324 (2005).
12. Kong, A. *et al. Science* **319**, 1398–1401 (2008).
13. Chowdhury, R., Bois, P.R., Feingold, E., Sherman, S.L. & Cheung, V.G. *PLoS Genet.* **5**, e1000648 (2009).
14. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
15. Jeffreys, A.J. & Neumann, R. *Hum. Mol. Genet.* **14**, 2277–2287 (2005).

Harvesting the apple genome

James Giovannoni

The genome sequence of the domesticated apple has been assembled and compared to previously sequenced plant genomes. The genetic sequence of the 17 apple chromosomes shows evidence of a recent genome duplication that may have spawned the additional gene family members needed for the evolution and development of the unique fruit structure of the apple termed the pome.

On page 833 of this issue, an international consortium of plant scientists led by the Istituto Agrario di San Michele all'Adige (IASMA) Research and Innovation Center in Trento, Italy report the genome sequence of the cultivated apple (*Malus* × *domestica*)[1]. Apples are among the most widely grown and consumed fruits in temperate regions of the world. This is in part due to years of extensive worldwide breeding and selection resulting in a treasure trove of apple colors, flavors and textures with broad versatility for the creation of numerous fresh and processed foods. Equally important to the apple's prominence in the marketplace (though less appreciated) is the fact that its unique fruit structure, termed a pome, has proven amenable to long-term controlled-atmosphere storage, facilitating year-round availability of high quality fruit from a crop

James Giovannoni is at the United States Department of Agriculture, Agricultural Research Service and the Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, New York, USA.
e-mail: jjg3@cornell.edu

that is harvested in the fall of the year. The apple genome sequence provides insight into the evolution of this agronomically important species and uncovers clues regarding the genetic basis of pome development. As such, it will provide a foundation from which further experiments can be designed to more fully understand apple genetics and biology, in addition to serving as a reference for important related crop species in its family, Rosaceae, which includes pear, peach, apricot, plum, cherry, quince, almond, strawberry, raspberry and the most popular of cut ornamentals, roses.

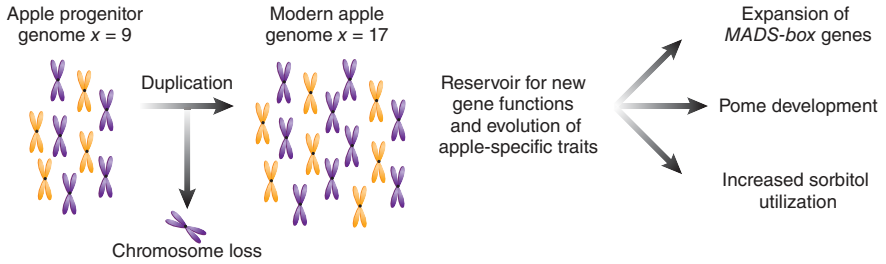

Figure 1 The evolution of the cultivated apple genome. Analysis of the apple genome suggests that a whole-genome duplication event in an ancestral genome, followed by loss of a single chromosome, led to the 17-chromosome karyotype of the cultivated apple. Expansion of particular gene families may have served as a reservoir for new gene functions, underlying the genetic basis of apple-specific traits.

Apple genome sequence

The apple genome was assembled using a combination of traditional and next generation sequencing technologies that produced approximately 600 Mb of the estimated 743.2-Mb genome of the popular diploid variety 'Golden Delicious'. This variety was selected in part because it is a member of the core set of varieties representing most cultivated apple germplasm. Cultivated apples are generally diploid and highly heterozygous, whereas some varieties, such as 'Gravenstein' and 'Jonagold', are triploid. The authors demonstrate that the vast majority of cloned apple genes (>90%) are found in this genome sequence, and that the bulk of the missing sequence, as is the case for most sequenced genomes, is highly repetitive and likely to contain relatively few functional gene sequences.

Evolutionary analyses of sequence duplications found in the apple genome suggest that a relatively recent genome duplication occurring around 50 million years ago contributed to the 17-chromosome karyotype of the Pyreae sub-tribe of Rosaceae, members of which are characterized by the presence of pome fruit and include apple and pear. Additional and more highly diverged duplicated sequences suggest ancient duplication events occurred in progenitor species, and analysis of inter- and intra-chromosomal organizational similarity (synteny) supports the hypothesis[2] that the modern apple genome resulted from autopolyploidization of a 9 chromosome progenitor to 18 chromosomes, followed by loss of a chromosome, resulting in the current 17-chromosome karyotype (**Fig. 1**). Phylogenetic reconstructions using the apple genome and gene (cDNA) sequences from related species clarify conflicting hypotheses regarding the progenitor of the cultivated apple, which this work suggests is the central Asian species *Malus sieversii*.

Evolution of pome fruit

The botanical definition of a fruit is the tissues that surround developing seeds. In many species, the fruit is derived primarily from carpel tissue that undergoes expansion in response to hormonal signals associated with pollination. Most fruits can be classified as dry (for example, beans or peanuts) or fleshy (for example, apple, banana, melons, etc.); the latter have evolved to produce succulent, sweet and colorful fruits to attract birds or animals that consume the fruits and disperse the seeds. The pome is a unique fruit structure in which the bulk of the fruit tissue (termed the cortex) is derived from expansion of tissue at the base of the floral organs known as the receptacles. MADS-box transcription factors are conserved among eukaryotes, and in plants, type II subfamily members play important roles in regulating floral organ identity and development[3]. Several of these genes have also been shown to be involved in fleshy fruit development and ripening[4–6].

Velasco *et al.* show that the type II subfamily of apple is expanded compared to most sequenced plant genomes. Other researchers have shown that some of these genes are differentially expressed during floral development and in the pome[7], leading to the tantalizing possibility that this family of transcription factors may play a prominent role in the unique fruit development of apple (**Fig. 1**). Velasco *et al.* highlight the fact that an apple *MADS-box* gene homologous to the *Arabidopsis FRUITFUL* carpel development gene is differentially expressed in the pome, and expansion of this family in apple is evident for genes related to the *Arabidopsis* clade containing *SVP* and *AGL24*, suggesting these genes as promising candidates for insights into pome biology. It will also be interesting to pursue apple homologs of other *MADS-box* genes that are more directly associated with floral organ expansion and fleshy fruit development, including *APETALA1* and members of the *AGAMOUS* clade. These genes have been shown to result in tomato sepal expansion[6] and carpel development in both tomato and the Rosacea family member peach[4,5,8], respectively.

Velasco *et al.* also observed that an additional functional class of expanded genes in apple is that related to sorbitol metabolism and transport. Although most plant species use sucrose as their primary means of photosynthetic carbohydrate transport, sorbitol is widely transported in apple, including into the pome, making this class of genes an additional logical target for pome evolutionary analysis. With the apple genome now available, its sequence can be harvested for DNA markers to assist in breeding and for explaining the genetic basis of the domestication of this important crop species and its unique and versatile fruit.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Velasco, R. *et al. Nat. Genet.* **42**, 833–839 (2010).
2. Velasco, R. *et al. PLoS ONE* **2**, e1326 (2007).
3. Irish, V.F. & Litt, A. *Curr. Opin. Genet. Dev.* **15**, 454–460 (2005).
4. Vrebalov, J. *et al. Plant Cell* **21**, 3041–3062 (2009).
5. Itkin, M. *et al. Plant J.* **6**, 1081–1095 (2009).
6. Vrebalov, J. *et al. Science* **296**, 343–346 (2002).
7. Janssen, B. *et al. BMC Plant Biol.* **8**, 16 (2008).
8. Tadiello, A. *et al. J. Exp. Bot.* **60**, 651–661 (2009).