CSH PRESS

GENOME RESEARCH

Sequence-based physical mapping of complex genomes by whole genome profiling

Jan van Oeveren, Marjo de Ruiter, Taco Jesse, et al.

Genome Res. published online February 1, 2011
Access the most recent version at doi:10.1101/gr.112094.110

SEQUENCE-BASED PHYSICAL MAPPING OF COMPLEX GENOMES BY WHOLE GENOME PROFILING

Jan van Oeveren[1], Marjo de Ruiter[1], Taco Jesse[1], Hein van der Poel[1], Jifeng Tang[1], Feyruz Yalcin[1], Antoine Janssen[1], Hanne Volpin[1], Keith E. Stormo[2], Robert Bogden[2], Michiel J.T. van Eijk[1] and Marcel Prins[1]

[1] Keygene N.V., Agro Business Park 90, 6708 PW, Wageningen, The Netherlands
[2] Amplicon Express Inc, 2345 NE Hopkins Court, Pullman, WA 99163, USA

Corresponding author:
Dr. Marcel Prins
Keygene N.V.
PO Box 216
6700 AE
Wageningen
The Netherlands
Tel. (+31) 317 46 68 66
Fax. (+31) 317 42 49 39
marcel.prins@keygene.com

Running title: Whole Genome Profiling

Keywords:
Sequence-based physical mapping, next generation sequencing, BAC library, whole genome profiling, genome assembly

Abstract

We present Whole Genome Profiling (WGP), a novel next-generation sequencing based physical mapping technology for construction of BAC contigs of complex genomes, using *Arabidopsis thaliana* as an example. WGP leverages short read sequences derived from restriction fragments of two-dimensionally pooled BAC clones to generate sequence tags. These sequence tags are assigned to individual BAC clones, followed by assembly of BAC contigs based on shared regions containing identical sequence tags. Following *in silico* analysis of WGP sequence tags and simulation of a map of Arabidopsis chromosome 4 and maize, a WGP map of *Arabidopsis thaliana* ecotype Columbia was constructed *de novo* using a six genome equivalent BAC library. Validation of the WGP map using the Columbia reference sequence confirmed that 350 BAC contigs (98%) were assembled correctly, spanning 97% of the 102 Mb calculated genome coverage. We demonstrate that WGP maps can also be generated for more complex plant genomes and will serve as excellent scaffolds to anchor genetic linkage maps and integrate whole genome sequence data.

Introduction

Physical clone maps are indispensable tools that form the intermediate layer between local (gene) sequences, genetic maps and whole genome sequences. Physical maps are widely used for a range of purposes including positional (map-based) cloning (Bakker et al. 2003), anchoring chromosomes using FISH (Islam-Faradi et al. 2002), repeat classification (Cardle et al. 2000), draft genome sequence assembly (Sasaki et al. 2005), local marker development (van der Vossen et al. 2000) and analysis of structural variation in the genome (Kidd et al. 2008). Despite advances in next-generation sequencing (NGS) technologies which have accelerated (re)sequencing complete genomes (Hillier et al 2008; Wheeler et al. 2008), the need for high-quality physical maps remains (Lewin et al. 2009). For example, *de novo* sequencing and assembly of complex genomes containing large regions of repeated sequences will not be easily addressed by NGS alone, and requires additional procedures to provide anchor points to link sequence contigs and bridge large repeat regions. An efficient way to provide these anchor points is by construction of a whole genome physical map from Bacterial Artificial Chromosome (BAC) clones (Shizuya et al. 1992; Rounsley et al. 2009), in combination with BAC-end sequencing (Nelson et al. 2009). BAC insert clones are relatively easy to generate and store and have proven to be effective for genome-wide physical map construction (Gregory et al. 1997; Marra et al. 1997; Klein et al. 2000; Wu et al. 2004). Hence, BAC-based physical maps combined with BAC end sequencing have formed the basis of several whole genome sequencing projects (Sasaki et al. 2005; Wei et al. 2007).

In the current gold standard for physical mapping, SNaPshot (Luo et al. 2003) and alternative methods such as AFLP® (Vos et al. 1995), BACs are characterized by means of DNA fingerprinting (Srinivasan et al. 2003; Borm 2008). The general principle behind these approaches is the characterization of individual BACs by means of specific tags, such as restriction fragments of specific lengths. These fragments are visualized by gel- or capillary electrophoresis and fingerprint patterns are scored based on the size of unique bands, with an assumption that bands of identical length represent identical fragments. To provide sufficient power for distinction and correct assembly, around 100 fragments are typically scored per BAC and BACs originating from the same region of the genome will be linked into a contig based on shared fragments. One of the assumptions for contig building is that the majority of these fragments are uniquely identifiable, i.e. that they represent a single location in the genome. However, in practice this is not always the case because the assessment of fragment length is known to suffer from both scoring inaccuracy as well as occasional co-migration of non-identical or duplicated fragments (Koopman et al. 2004). The latter may lead to false linkage between BACs. For example, in the maize HICF map (Nelson et al. 2005), the average number of erroneously shared bands was reported to be 10.8 from an average of 98 bands per clone, i.e. a random overlap of 11% of bands (Nelson et al. 2009). This observation underscores the need to apply stringent assembly criteria to prevent formation of contigs comprising non-contiguous BAC clones when using DNA fingerprint data for physical map building.

As an alternative, optical mapping (Schwartz et al. 1993) has been described for constructing ordered restriction maps and has been used in sequencing projects of several whole genomes (e.g. Zhou et al. 2009, Church et al. 2009). In contrast to SNaPshot it preserves the order of the restriction fragments in a given DNA fragment, however both methods share the inaccuracy of restriction fragment (length) calling and are not sequence-based.

AFLP® is a robust complexity reduction technology, which uses restriction enzyme digestion and consecutive adapter ligated restriction fragment amplification. Sample preparation of short read NGS platforms such as the Illumina Genome Analyzer (GA) use similar steps. Extending on this similarity we envisaged the use of NGS to uniquely characterize restriction enzyme fragments not by their size but by their unique sequence content. Hence, we developed WGP as a high resolution sequence-based physical mapping technology based on short read NGS with sequence tags placed along the entire BAC clone. WGP incorporates the use of sample identification tags (also known as barcodes) and a two-dimensional BAC clone pooling strategy to make optimal use of sequencing capacity and reduce costs for BAC DNA preparation. WGP tags uniquely characterize individual BACs. Shared WGP tags are consecutively used for BAC contiging (Fig. 1 and Supplemental Fig. S1).

To provide proof of concept of the WGP technology a wet lab experiment was conducted involving genome-wide assembly of an *Arabidopsis thaliana* ecotype Columbia BAC library. The WGP results were validated using the publicly available reference genome sequence of this ecotype. Subsequent application of WGP in other plant genomes and a simulation study performed in maize demonstrate the scalability of WGP to complex plant genomes.

Results

In silico analysis

To investigate the feasibility of the WGP approach and the requirements for the WGP tags, the complete *Arabidopsis thaliana* Columbia sequence was *Eco*RI/*Mse*I 'digested' *in silico*. This produced 69,275 fragments representing 64,059 unique sequences flanking the *Eco*RI sites (Table 1). The length of the sequence tag influences its uniqueness and showed a saturation point around 20 nt, where over 87% of all restriction fragments are represented by a single unique tag (Supplemental Fig. S2). Such tags can be easily generated with short read NGS platforms such as Illumina's Genome Analyzer (GA) or Life Technologies' SOLiD.

In silico WGP simulation analysis on Arabidopsis chromosome 4 and the complete genome of maize

Two WGP simulations, one on Arabidopsis chromosome 4 and another on the entire genome of maize, were carried out to substantiate the hypothesis that short sequence tags adjacent to restriction enzyme sites can be used for unique identification of BAC clones and subsequent BAC-based whole genome physical mapping. Both analyses demonstrated that pooling of BACs in a 2D format allowed satisfactory deconvolution of tags to individuals BACs and the ability to use such information to generate high quality physical maps ranging from relatively simple single chromosomes up to complex large genomes such as maize. For details see supplemental information.

WGP tag generation and deconvolution

An existing six genome equivalent library of *Arabidopsis thaliana* ecotype Columbia was subjected to WGP. The BAC library of 6,144 clones was divided over sixteen 384-well plates which were each pooled by row (24 clones) and column (16 clones) to result in 40 pools per plate. DNA isolated from these pools was digested with *Eco*RI and *Mse*I, barcoded adaptors were ligated and restriction fragments were amplified. Amplified fragments were subsequently sequenced from the *EcoRI* restriction site end using the Illumina GA. In total this yielded 30.3 million passed filter reads with 31 nt read length (NCBI Sequence Read Archive: SRA026464.1). Of these reads, 93% (28.2 M) contained a valid (100% matching) sample identification tag and *Eco*RI restriction site sequence. In table 2 the data are specified for each of the 8 lanes of the experiment, which demonstrate the high consistency of the data quality and percentages of reads that could be used for assignment to individual BACs.

Next, the reads were analyzed with custom made Perl scripts. Using these scripts, the sample (pool) identification tag and the restriction site sequence were identified and removed after binning in the appropriate pool. Clustering on 100% identity was done to obtain unique WGP tags. Only those WGP tags which were present in a single row and a single column pool per plate were used for further analysis as they could be unequivocally assigned to a single BAC. In total this encompassed 12.1 M reads corresponding to 43% of the sequence read data (% deconvolution; Table 2). With these reads 4,599 BACs were identified having at least one WGP tag. An average of 40 WGP tags was identified per BAC, each WGP tag being sequenced on average 66 times (Table 2; Fig. 2).

Experimental data confirms uniqueness of 20 nt tag length

To substantiate the *in silico* analysis of tag length required for uniqueness, an analysis of required read length was performed on the data derived from the Illumina GA reads. Reads were trimmed to lengths ranging from 11 to 26 nt. The number of resulting deconvolutable tags showed a saturation point around 20 nt similar to the *in silico* data (Supplemental Fig. S2). This suggests that generating sequence tag reads longer than 20 nt will not improve the WGP map and confirms the practical use of short read lengths as provided by the Illumina GA sequencer for WGP in *Arabidopsis thaliana*.

Verification of WGP tag quality

The total of 183,366 tags that were assigned to single BACs represented 65,734 different sequences (Tables 1 and 2). These sequences were compared to the Arabidopsis genome sequence and only tags with sequences that were 100% identical were retained for further comparison, yielding 61,638 hits in the genome from 56,513 different tags (82%; Table 1). The majority of these tags showed a single hit to the genome: 54,271 (96%; Table 1). The remaining 2,242 tags matched two or more genomic regions, likely indicating genuine low copy repeats in the genome that were nevertheless deconvoluted. The average distance between the uniquely mapped tags equaled 2,100 bp. As each *Eco*RI site can theoretically generate a forward and a reverse WGP tag, this is an accurate reflection of the approximately 4.5 kb average distance between the *Eco*RI sites in the genome, taking into account that not all possible tags are indeed recovered. The *in silico* *Eco*RI/*Mse*I digest of the Columbia genome that was generated for validation of the experimental data yielded 69,275 fragments. When trimmed to 26 nt to match the experimental WGP analysis, these fragments represented 60,676 different tags, 57,992 (96%) of which represented a unique position in the genome. Eighty six percent (51,935 tags) of the theoretically possible 60,676 unique *in silico* tags were also retrieved in the wet-lab WGP experiment (Table 1).

Contig building using FPC and map quality assessment

A cut-off value of $1.0\text{x}10^{-6}$ and a consecutive DQ step was used in the FingerPrinted Contig (FPC) software (Soderlund et al. 1997) to generate the WGP map as it produced a high level of BACs in contigs with a small number of questionable clones that may cause false overlaps. This resulted in 273 contigs representing 4,048 BACs and 61,514 WGP tags. The distribution of contig sizes and number of BACs is shown in supplemental Fig. S3 and these contigs covered approximately 101 Mb (78%) of the genome, using a 2,100 bp average distance between WGP tags as basis for the coverage calculation. Next to the contigs, 551 singleton BACs remained including the majority of the clones containing less than five tags (Fig. 2). At the FPC settings used in this experiment, the minimum number of WGP tags on a contiged BAC equaled two, even though only a limited number of such BACs were incorporated.

To validate correctness of the BAC contigs that together form the whole genome physical map, the 65,734 WGP tag sequences were mapped to the *Arabidopsis* genome sequence using a perfect string match algorithm. Ninety-one percent of the 270 contigs containing two or more unique WGP tag hits mapped to contiguous regions of the genome and covered 82 Mb, representing 81% of the contig coverage (Table 3). However, twenty three contigs mapped to two distinct regions and one contig mapped to three regions. These 24 contigs (9%) comprised 19% of the contig coverage. To address this issue a generic identification and purging tool was applied to remove 230 potentially problematic BACs from the data set including 177 BACs that contained more than 80 WGP tags. With this filtered BAC data set another FPC run was performed using the same settings. This resulted in 3,813 BACs assembled into 362 contigs, and 556 singleton BACs (Table 3). The 362 contigs represented a calculated genome coverage of 102.4 Mb, which in spite of the removal of 230 BACs is nearly the same as the unfiltered data set (100.8 Mb), albeit split over more contigs. Of these 362 contigs, 357 contained two or more WGP tags with unique hits to the genome, of which 350 (98%) mapped to contiguous genome regions and covered 99 Mb (97%). The remaining seven contigs (3%) mapped to multiple regions and comprised 3% contig coverage (Table 3). Figure 3 shows the alignment of a representative contig mapping to chromosome 3.

Scalability of WGP to larger and more complex genomes

To further substantiate the robustness of the WGP method, four additional plant genomes were subjected to WGP, comprising of: melon (*C. melo,* 450 Mb genome size), tomato (*S. lycopersicum,* 950 Mb), the allotetraploid rape seed (*B. napus,* 1200 Mb) and lettuce (*L. sativa*, 2,600 Mb) Data presented in Supplemental Table S1 demonstrate that as for Arabidopsis these genomes were amenable to WGP and thus generated high quality high coverage whole genome physical maps.

Discussion

In this manuscript Whole Genome Profiling is demonstrated as a novel approach to construct sequence-based genome-wide physical maps of complex genomes. Restriction fragment-based sequence data from pooled BAC clones were generated using the Illumina GA NGS platform and resulted in the assignment of an average of 40 unique WGP tags per BAC clone. Using these tags a whole genome physical map of *Arabidopsis thaliana* was created using a 6x BAC library. Validation of the WGP map using the Columbia reference genome sequence confirmed that 350 BAC contigs (98%) were assembled correctly, which comprised 97% of 102 Mb calculated genome coverage, equaling approximately 80% of the genome of Arabidopsis.

A first version of the Arabidopsis WGP map was produced using stringent cut-off settings in the FPC software. As the actual position of the WGP tags in the released genome sequence of Arabidopsis could easily be established, the quality of the contigs produced by WGP was investigated in detail. This demonstrated that tags belonging to the vast majority of the contigs originated from a single region in the genome (see e.g. Fig. 3). However, tags present in 24 of the 273 contigs mapped to more than one region in the genome. Careful inspection of these contigs showed that in all these cases single BACs were responsible for the merger of two distinct genomic regions. The Arabidopsis BAC library that was used in this study contained either a chimeric BAC or multiple BACs per well in less than 0.6% of the plate wells. However, due to the integrating power of the FPC assembly this had a profound effect on the WGP map, as nearly 9% of the contigs obtained from the first round of FPC using unfiltered BACs was not contiguous and represented 19% of the contig coverage. In order to minimize chimeric contig formation, a generic method was developed that does not rely on a reference genome sequence to recognize and eliminate problematic BACs. The selection of these BACs was based on two observations: firstly that many of the disturbing BACs produced a significantly higher number of tags compared to the genome fraction they contained, suggesting the presence of two separate genome regions. Secondly, chimeric contigs showed a typical tag distribution, which allowed pin pointing potentially chimeric contigs and to identify BACs causing the chimerism and purge them from the data set.

As a result, the final physical map based on the filtered BAC data set retained its coverage, but was split into more contigs. Seven contigs, containing 130 BACs, were still identified as split over two genome regions. However, this covered no more than 3% of the genome, a marked improvement over the 19% coverage of chimeric contigs prior to filtering (Table 3). It should be noted that the genome coverage of this experiment (6x) is relatively low and that contig building with FPC of deeper coverage BAC libraries under high stringency settings will prevent formation of chimeric contigs more easily, particularly in combination with the purge tool.

A number of general features of the WGP technology are worthwhile addressing in more detail, such as the influence of sequence read length, pooling schemes, restriction enzyme choice and genome size on map resolution and cost effectiveness.

In silico data, confirmed by experimental data, demonstrated that a tag length of 20 nt was sufficient to define more than 90% of the unique sequence tags in the *Arabidopsis* genome. This is in agreement with results from previous studies on the effect of read length on determining unique positions in the genome (Whiteford et al 2005;

Chaisson et al. 2009). A 20 nt tag length even when adding 6 nt for a pool identifier is well within the sequence length output of contemporary high throughput sequencing machines. Additional analyses on other (partially) sequenced genomes and our simulation of the maize genome demonstrated that also for significantly larger genomes tag lengths between 26 and 31 nt are sufficient for WGP.

The 2D pooling strategy that was chosen to process the limited number of 6,144 Arabidopsis BACs was a trade-off between costs of sample preparation and sequencing using the Illumina GA and genome size. Since sample preparation of single BACs was deemed too costly, it was decided to pool BACs prior to DNA isolation and then apply the restriction, ligation and fragment amplification to pooled DNAs in combination with deconvolution of sequencing results. The backbone of the contiging strategy is the occurrence of unique WGP tags in two or more BACs that share the same portion of the genome. If, however, these two BACs or fragments thereof appear in the same pooling set ("deconvolution space"), the WGP tag will be lost in the deconvolution process as it will show up in three or more BAC pools. The 6x BAC library we used was divided over sixteen 384-well plates. A pooling set size of a full plate (384 BACs with a 125 kb average insert size) thus covers approximately 48 Mb (40%) of the Arabidopsis genome. The chance of finding multiple BACs that contain fragments from the same genomic region within a single pool is therefore relatively high. In general, the probability of encountering the same genomic region in a single pool will depend on the size of the genome related to the complexity of the pools: thus WGP analyses on small genomes require less complex pools, whereas large genomes allow larger numbers of BACs per pool. Furthermore, for each WGP tag on each BAC a minimum redundancy is required to compensate for random sampling variation. This implies that a 2D pooling set-up requires twice the number of sequence reads per BAC compared to individual BAC clone sequencing (1D), whereas a 3D pooling scheme requires three times as many reads. Depending on relative costs of sequencing compared to the costs of sample preparation, an optimal pooling scheme can be designed. As the cost of sequencing per nucleotide is rapidly dropping in the NGS systems used to date it is likely that higher level pooling schemes will become cost effective in the near future.

A particular advantage of using BAC DNA pooling and the consecutive deconvolution procedure in WGP is the selection against tags derived from repetitive regions, which often hamper building accurate contigs in fingerprint-based physical mapping methods. As repeat tags are likely to occur in multiple BACs within a pooling set, the majority will not be assigned to BACs by deconvolution and excluded from the input file for contig building. As a result, only unique or low copy regions of BACs are used for contig building and this is beneficial for utility of the map for integrating sequence scaffolds or marker development. On the other hand, this also implies that BAC clones containing predominantly repeated sequences may end up with too few unique WGP tags for contiging and not be represented in the WGP map. Indeed, it was our observation that singleton BACs were characterized by far fewer WGP tags per BAC (Fig. 2).

*Eco*RI/*Mse*I was used for the Arabidopsis WGP, but the use of amplified restriction fragments as a starting point in the WGP concept is flexible in the choice of restriction enzyme(s). Commonly used restriction enzymes such as *Eco*RI or *Hin*dIII typically generate between 30 and 50 fragments per BAC clone. Depending on the abundance of restriction sites in a given species, the use of alternative enzymes will allow additional fine tuning in the number or distribution of WGP tags per BAC. Depending on the complexity of a target genome, the use of specific restriction enzymes in combination with pooling complexity assures an optimal use of sequencing capacity to further improve cost effectiveness. *In silico* analysis of available sequence data can aid in providing the required information for novel WGP projects.

Additional results of WGP applied to melon, tomato, rape seed and lettuce substantiate the robustness of this new technology (Supplemental Table S1). These data are complemented with the results from the *in silico* maize analysis (Supplemental data). Results show that even for large and complex genomes that are known to contain a high proportion of repeated regions WGP is efficient in tagging 70-90% of the BAC clones and yields high coverage physical maps.

In conclusion, Whole Genome Profiling was demonstrated to be an efficient novel NGS-based technique for whole genome physical map construction. In addition, WGP maps are an excellent scaffold to complement *de novo* whole genome sequencing efforts, as they consist of densely spaced unique sequence tags. To date, especially random shotgun whole genome sequencing (WGS) efforts rely on high sequencing redundancies in combination with paired-end sequencing strategies to achieve sufficiently large sequence contigs and scaffolds. WGP maps, as they consists of sequence-tagged contiged BACs, provide multiple anchor points over distances that by far exceed those that can be bridged by current paired-end sequencing strategies. It can thus be expected that WGP will contribute to significantly improved genome sequence assembly metrics and cost reduction for WGS efforts, because very high sequencing redundancy levels are no longer needed to obtain large scaffolds. In addition, like fingerprint-based physical methods, WGP maps provide a ‘minimum tiling path’ of BAC clones for targeted sequencing of genome regions of particular interest and direct access to individual BAC clones. These features of WGP and the widespread availability of BAC libraries underscore the power of WGP to advance genome analysis in a broad range of species.

Methods

In silico analysis

The complete *Arabidopsis thaliana* ecotype Columbia sequence (TAIR8; http://www.arabidopsis.org) was digested *in silico* with *Eco*RI/*Mse*I to obtain all possible restriction fragments. An analysis was performed to obtain the number of unique WGP tags after trimming them to varying read lengths from 10 to 26 nt from the *Eco*RI site. Data were compared to the number of tags resulting from the *in vitro* WGP experiment.

BAC library and pooling strategy

A 6,144 clone BAC library of the *Arabidopsis thaliana* Col-1 ecotype (available at the ABRC stock center, stock # CS6000) and two-dimensionally (2D) pooled BAC DNAs were made available for this project by Amplicon Express Inc. (Pullman, WA, USA). The 2D pooling strategy was based on a simple 384-well plate by rows and by columns pooling strategy implemented with a liquid handling robot (Beckman Coulter, Biomek 2000, Brea, CA, USA). The robot pooled 75 µl of freshly grown bacterial culture in 2xYT media from all of the clones from each row (24 BACs) on each plate and 112 µl all of the clones from each column (16 BACs). The set of 16 row pools and 24 column pool is termed a SuperPool (SP). The DNA tray format allows culture liquid to be collected from two individual 384-well library plates (2 SPs) and have all 80 wells processed independently for high quality DNA extraction in a 96-well deep block format. The collected culture fluid is centrifuged to pellet the cells and discard the supernatant. The blocks of cell pellets are then frozen at -20°C until processed with a modified alkaline lysis DNA extraction protocol (Maniatis et al. 1982). The average yield of plasmid DNA is approximately 200 ng per well of high quality, Illumina sequencing grade DNA.

Sequencing sample preparation

AFLP templates were prepared from the pooled BAC clone DNA as described by Vos and co-workers (Vos et al. 1995). Twenty ng of pooled BAC DNA was digested using 5 units *Eco*RI and 2 units *Mse*I for at least 1 hour at 37°C. Next, adapter ligation using a universal P7 *Mse*I adapter and a sample specific tagged *EcoRI* P5 adapter was carried out for 3 hours at 37°C. PCR was performed in 20 μl and contained 5 µl 10-fold diluted restriction ligation mixture, 30 ng Illumina P5 primer (5'-AATGATACGGCGACCACCG-3'), 30 ng Illumina P7 primer (5'-CAAGCAGAAGACGGCATACGA-3'), 0.2 mM dNTPs, 0.4 U AmpliTaq® (Applied Biosystems) and 1x AmpliTaq buffer. PCR was performed with the following profile: 2 minutes 72°C followed by 22 cycles of 30 sec 94°C, 60 sec 56°C, 60 sec 72 °C, and finally held at 4°C. Next, equal amounts of SP samples were purified using the QIAquick PCR Purification Kit (Qiagen). All 80 sample specific *Eco*RI P5 adapters included a unique five nucleotide sample identification tag adjacent to the *EcoRI* restriction site overhang for identification of individual BACs by deconvolution.

Sequencing

Two Illumina Genome Analyzer runs were performed, a titration run and a full scale run. Each run was done on a flow cell divided into 8 lanes for physical separation of samples, such that the same set of sample tags were used for each lane. The first run contained only 4 SPs, with 2 sets of 2 SPs replicated on 3 lanes each in a range of increasing concentration (1.5, 2.0 and 2.5 pM). The 2.5 pM lanes were selected as the optimal DNA concentration and a second GA run was performed comprising six lanes with Arabidopsis WGP samples each covering two SPs represented in 80 row- or column pools. A total of 16 SPs were sequenced equaling 16 x 384 = 6,144 BACs.

The Illumina pipeline software (GA pipeline v0.3) was used to analyze images into sequence reads of 31 nt length. An additional quality filter was applied to select only those reads with all base calls being at least 0 on the Illumina GA scale. All sequence data have been deposited in the NCBI Sequence Read Archive (SRA) SRA026464.1.

Deconvolution

Sequence reads were split into three parts to enable assignment of unique tags to pools and to allow for consecutive deconvolution into individual BACs: the first 5 nucleotides (nt) represent the sample (i.e. BAC pool) identification tag or barcode; the next 6 nt match the *Eco*RI restriction site of the adapter and the remaining 20 nt define the WGP tag of Arabidopsis genomic DNA sequence from BAC clone inserts. The assignment of unique WGP tags to individual BACs was based on the following criteria: 1) A specific WGP tag must occur in two pools to indicate its unique position on the plate: one column and one row pool with both being represented by at least two reads, and 2) if WGP tags are inadvertently observed in a third or fourth pool, the number of reads in these other pools must be less than a tenth of those in the smallest (correct) pool. All WGP tags not matching these criteria were discarded. Perl scripts were used to recognize and trim the sample identification tags and the restriction site part of the sequence reads and to perform the deconvolution. Unique WGP tags were defined by grouping them in 100% identical read sets. The output of this procedure consisted of a list of all WGP tags, the corresponding number of reads, and the identification number of the unique BAC to which they were assigned.

Contig building

Contiging was performed using the FingerPrinted Contig (FPC) program (v9.4; Soderlund et al. 1997; http://www.agcol.arizona.edu/software/fpc/). This software tool was originally developed for analyzing BAC fingerprint data: restriction fragments determined by their length. The WGP tags were adapted for use in FPC by converting each unique sequence tag into a number, yielding pseudo restriction fragment sizes for which the FPC software was originally designed. As the WGP tags are uniquely defined by their sequence composition, FPC could be used at the highest stringency setting of tolerance (value = 0). Normally this setting compensates for variation in band mobilities of restriction fragments, which is not an issue here. Different cut-off values were tested, specifying the threshold on the probability of BAC coincidence, i.e. the likelihood that different BACs share overlapping WGP tags, while not originating from the same genomic region. A DQ analysis was performed to further split problematic contigs with so-called Q-clones: clones which cause potential false overlap. The output of FPC consisted of a list of contigs and the corresponding order of BACs within each contig.

Map validation

WGP tag data, including the adjacent *Eco*RI restriction site, were mapped to the *Arabidopsis* genome sequence. All hits with a 100% identity and full-length alignment were recorded and the corresponding WGP tags were consequently characterized by their chromosome number and base pair position. Visual alignments were made from the sorted position of the WGP tags, based on their location in the genome compared to their orientation on the BAC contigs. To check the quality of the resulting FPC contigs, all tags were included which showed a single hit to the genome reference. For each contig with two or more of such unique tags the tag positions were verified in the following way: the median genome position was calculated for the set of tags, tags were classified as outliers if their position was more distant from the median than a given threshold (the number of unique tags in this contig multiplied by 5,000 bp – approximately twice the mean distance between tags). Contigs with more than five outliers were identified as hybrid contigs, mapping to multiple regions on the genome, and checked as such. All other contigs with two or more unique tags were defined as good quality contigs.

Problematic BAC identification and purging

To overcome false (chimeric) contig formation caused by problematic BACs, e.g. BACs with chimeric inserts or two BACs which were accidentally deposited in the same well of a 384-well plate, a three step procedure was developed. First, BACs containing more than twice the mean number of tags were excluded from the analysis as they represent a large fraction of problematic clones relative to their total number. Secondly, FPC contig building was performed and the resulting contigs were scored on two metric parameters which correlated with chimeric contig features. These metrics are the fraction of BAC pairs within a contig sharing at least one WGP tag (C1) and the average tag density in a contig as defined by the number of unique WGP tags in the contig divided by the number of BACs in the contig (C2). Empirically, the square of C1 divided by C2 provided a value that effectively discriminated chimeric contigs from contiguous contigs at a threshold of 0.003 and lower. The rationale for C1 was that chimeric contigs were expected to represent two proper contigs that were incorrectly linked by a single problematic BAC, causing the fraction of BAC pairs that share at least one WGP tag (C1) to drop. At the same time the total number of WGP tags represented in the contig increases compared to the genome coverage it represents (C2). Thirdly, to identify the problematic BACs within putatively chimeric contigs, an iterative approach of removing individual BACs from these contigs based on transitivity clustering was performed for all BACs. BACs were eliminated from the dataset if their removal led to breaking up the contig. The filtered BAC dataset was used for a final FPC analysis using the same stringency settings.

All scripts used to perform the WGP experiments as described are freely available after registering at http://www.keygene.com/research/WGPpublication.php

Acknowledgements

This work was supported by Enza Zaden, Rijk Zwaan, Vilmorin & Cie and Takii & Co. We are very grateful to Drs. William Nelson and Carol Soderlund of the University of Arizona for adapting the FPC program for WGP data processing. The AFLP® and Whole Genome Profiling (WGP) technologies are covered by patents and patent applications owned by Keygene N.V. AFLP and KeyGene are registered trademarks of Keygene N.V.

References

Bakker E, Butterbach P, Rouppe van der Voort JN, van der Vossen EA, van Vliet J, Bakker J, Goverse A. 2003.Genetic and physical mapping of homologues of the virus resistance gene Rx1 and the cyst nematode resistance gene Gpa2 in potato. *Theor Appl Genet* **106**, 1524–1531.

Borm TJA. 2008. Construction and use of a physical map of potato. PhD thesis, Wageningen University, 139p.

Cardle L, Ramsaya L, Milbournea D, Macaulaya M, Marshalla D, Waugha R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* **156**, 847-854.

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* doi:10.1101/gr.079053.108.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**:e1000112.

Gregory SG, Howell GR, Bentley DR. 1997. Genome Mapping by Fluorescent Fingerprinting. *Genome Res* **7**, 1162-1168.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, et al. 2008. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods*. **5,** 183–188.

Islam-Faridi MN, Childs KL, Klein PE, Hodnetta G, Menz MA, Klein RR, Rooney WL, Mullet JW, Stelly DM, Price HJ. 2002. A Molecular Cytogenetics Map of Sorghum Chromosome 1: FISH Analysis with Mapped BACs. *Genetics* **161,** 345-353.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64.

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, et al. 2000. A high-throughput AFLP-based method for constructing integrated genetic and physical maps: Progress towards a sorghum genome map. *Genome Res* **10**, 789–807.

Koopman WJM, Gort G. 2004. Significance tests and weighted values for AFLP similarities, based on Arabidopsis in silico AFLP fragment-length distributions. *Genetics* **167**, 1915-1928.

Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. *Genome Res* **19**, 1925-1928.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378–389.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**, 1072-1084.

Maniatis T, Fritsch EF, Sambrook, J. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Nelson W, Soderlund C. 2009. Integrating sequence with FPC fingerprint maps. *Nucleic Acids Res* **37**, e36 doi:10.1093/nar/gkp034.

Nelson WM, Bharti AK, Butler E, Wei F, Fuks G, Kim H, Wing RA, Messing J, Soderlund C. 2005. Whole-Genome Validation of High-Information-Content Fingerprinting. *Plant Physiol* **139**, 27-38.

Rounsley S, Marri PR, Yu Y, He R, Sisneros N, Goicoechea JL, Lee SJ, Angelova A, Kudrna D, Luo M, et al. 2009. De novo Next Generation Sequencing of plant genomes. *Rice* **2**, 35-43.

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, et al. 2005. The map-based sequence of the rice genome. *Nature* **436**, 793-800.

Schwartz DC, Li X, Hernandez L, Ramnarain SP, Huff EJ, Wang YK. 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **262**, 110–114.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci USA* **89**, 8794-8797.

Soderlund C, Longden I, Mott R. 1997. FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci. **13**, 523–535.

Srinivasan J, Sinz W, Jesse T, Wiggers-Perebolte L, Jansen K, Buntjer J, van der Meulen M, Sommer RJ. 2003. An integrated physical and genetic map of the nematode Pristionchus pacificus. *Mol Genet Gen* **269**, 715-722.

van der Vossen EA, van der Voort JN, Kanyuka K, Bendahmane A, Sandbrink H, Baulcombe DC, Bakker J, Stiekema WJ, Klein-Lankhorst RM. 2000. Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. *Plant J* **23**, 567–576.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**, 4407–4414.

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S, et al. 2007. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* **3**, e123.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876.

Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Mark Bradley M, Cameron Neylon C. 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res* **33**, e171.doi:10.1092/nar/gni171.

Wu CC, Nimmakayala P, Santos FA, Springman R, Scheuring C, Meksem K, Lightfoot DA, Zhang HB. 2004. Construction and characterization of a soybean bacterial artificial chromosome library and use of multiple complementary libraries for genome physical mapping. *Theor Appl Genet* **109**, 1041–1050.

Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas, C, Pasternak S, et al. 2009. A single molecule scaffold for the maize genome. *PLoS Genetics* **5**, e1000711.

Figure legends

Figure 1. Overview of the Whole Genome Profiling technology.
(A) BAC library: BAC clones are available in a 384-wells plate format. (B) BAC pooling and DNA extraction: DNA is extracted after pooling each row (24 BACs) and each column (16 BACs). (C) Template preparation and sequencing: pooled BAC DNA is digested (*Eco*RI/*Mse*I) and amplified after barcoded adaptors are ligated for pool identification after sequencing on the Illumina GA platform. (D) Deconvolution: sequence tags (30-50 per BAC) are assigned to individual BACs based on presence in 1 row and 1 column pool. (E) Contiging: overlapping (sets of) sequence tags from individual BAC clones generate contigs. Together these contigs represent a sequence-based physical map of the genome.

Figure 2. Distribution of the number of tags per BAC, specified for all 4,599 deconvoluted BACs and for the subset of 551 BACs that were not assembled in contigs (singleton BACs).

Figure 3. Overview (**a**) and zoomed in detail (**b**) of a typical BAC contig located on Arabidopsis chromosome 3.
The sequence of the WGP tags, the chromosome number and the base pair position in the chromosome of the first base of the WGP tag are shown. In each color-coded BAC, the presence of a WGP tag is indicated by 'x'. Gaps in BACs represent missing tags due to insufficient deconvoluted reads.

Table 1. Overview of the number of sequence tags derived from the *Arabidopsis* WGP experiment and the *in silico* simulation analysis.

	WGP tags	
	Experimental	***in silico***
Total nr fragments	n.a.	69,275
Total number of different fragments	n.a.	64,059
Total number of different WGP tags (26 nt)*	65,734	60,676
100% hit with genome	56,513	60,676
Unique genome position	54,271	57,992
Overlap (found both *in silico* and real)	51,935	51,935

* 6 nt *Eco*RI + 20 nt fragment sequence

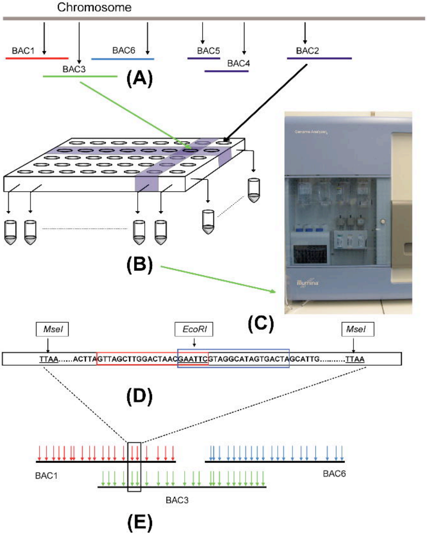
Table 2. Number of reads, deconvolutable reads, tags and BACs for the eight Illumina lanes.

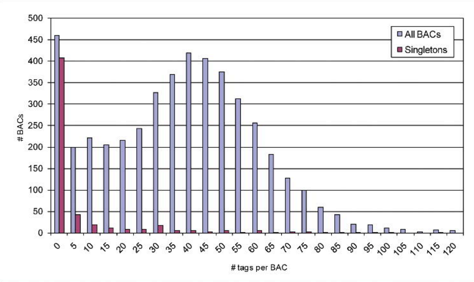
Pooling set	# OK reads	Deconvolutable			
		# tags	# reads	% reads	# BACs with tags
Plates 1 & 2	2.6 M	25,598	1.1 M	42 %	585
Plates 3 & 4	4.3 M	22,751	1.7 M	40 %	593
Plates 5 & 6	3.5 M	21,179	1.5 M	44 %	582
Plates 7 & 8	3.8 M	29,027	1.8 M	48 %	591
Plates 9 & 10	3.1 M	21,626	1.4 M	46 %	569
Plates 11 & 12	3.6 M	20,060	1.4 M	40 %	549
Plates 13 & 14	3.3 M	22,019	1.5 M	44 %	576
Plates 15 & 16	3.9 M	21,106	1.6 M	42 %	554
Total	**28.2 M**	**183,366**	**12.1 M**	**43 %**	**4,599**
Average		**40 tags/BAC**	**66 reads/tag**		

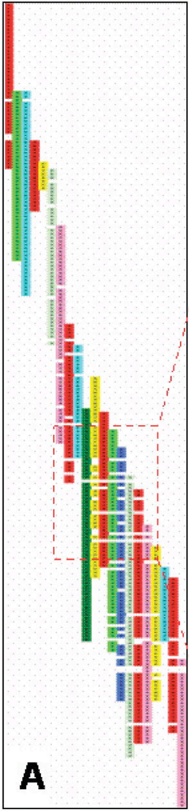
Table 3. FPC results for contig building of WGP Arabidopsis for the initial physical BAC map (All BACs) and after filtering problematic BACs (Filtered BACs). Indicated is whether contigs map to a single region or have multiple hits

	All BACs		**Filtered BACs**	
Nr of input BACs	4,599		4,369	
Nr of input WGP tags	65,734		62,829	
Nr of BACs placed in contigs	4,048		3,813	
Nr of contigs	273		362	
Nr of Q-contigs (Q>5)	16		1	
Nr of singleton BACs	551		556	
coverage (Mb)	100.8		102.4	
WGP tags with hits	56,513	86%	54,877	87%
Nr of contigs > 1 unique WGP tag hit	270		357	
Nr of contigs mapping to 1 region (= single region contigs)	246	91%	350	98%
Nr of BACs in single region contigs	3,071	76%	3,649	96%
coverage (Mb) single region contigs	82	81%	99	97%
Nr of contigs mapping to multiple regions	24		7	

Parameter setting: tolerance = 0; cut-off = $1.0x10^{-6}$; DQ step 1.


Chromosome
BAC1
BAC6
BAC5
BAC4
BAC2
BAC3
(A)
Genome Analyzer
illumina
(B)
(C)
MseI
EcoRI
MseI
TTAA......ACTTAGTTAGCTTGGACTAACGAATTCGTAGGCATAGTGACTAGCATTG..........TTAA
(D)
BAC1
BAC6
BAC3
(E)


500
450
400
350
300
250
200
150
100
50
0
# BACs
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
# tags per BAC
All BACs
Singletons



B

BACs in order of their FPC map position														
BAC852	BAC4124	BAC1373	BAC285	BAC2544	BAC704	BAC3536	BAC2070	BAC4237	BAC5328	BAC3912	BAC1461	Sequence	Chrom	bp
X	X	X	X	X	X							GAATTCAAGAGTACCTTTCAAGGGAG	Chr3	17644262
X	X	X	X	X	X	X						GAATTCCAGTGTATCCATTAGGCCCT	Chr3	17648937
X	X	X	X	X	X	X						GAATTCCAAGTTCTTGTTGCAGCCAT	Chr3	17648957
X	X	X	X	X	X	X						GAATTCAATCGAGTAAACTCTTCGCA	Chr3	17652220
X	X	X	X	X	X	X						GAATTCTCCCTGAGGAACTATAATTG	Chr3	17652240
	X	X	X	X	X	X						GAATTCAGAAGAACCCTAGACTAAAT	Chr3	17674086
	X	X	X	X	X	X	X					GAATTCAATGCATTTTTGATTTTCCA	Chr3	17674106
	X	X	X	X	X	X	X					GAATTCTATCCCTAAGTGCTACAACA	Chr3	17676593
	X	X	X	X	X	X	X					GAATTCCATAAAGTTCTCGGATCACA	Chr3	17676613
			X		X							GAATTCGCTAGTTTTAAGATCATTAT	Chr3	17680904
	X		X		X	X	X					GAATTCGGATTTAAACGCGTTCTCGA	Chr3	17680924
	X		X	X	X	X	X					GAATTCAACACGGTATCAATGAACAA	Chr3	17681881
	X		X	X	X	X	X					GAATTCACGGTAATGTTGAGCTTGCA	Chr3	17683056
			X		X							GAATTCGGAGATGAATCTTTGGTTTC	Chr3	17683621
	X		X	X	X	X	X	X				GAATTCAGCATGGAAAAAGTGGTGCT	Chr3	17691042
	X											GAATTCACTAAATTAATCAAACCTCA	Chr3	17691062
			X	X	X	X	X	X				GAATTCTATATAAACCTTTTTTTGTG	Chr3	17694949
			X			X		X				GAATTCATGGTTAATTTGTATAGATT	Chr3	17694969
			X	X	X	X	X	X	X			GAATTCTATGATACACTTATGTAGTT	Chr3	17697899
			X	X	X	X	X	X	X			GAATTCCTCTTGTCAAAAAATTTATC	Chr3	17697919
			X		X	X		X				GAATTCAGGTATTCGATGGTTAATTT	Chr3	17698336
			X	X	X	X	X	X	X			GAATTCTACACTACACTAATGAGGTC	Chr3	17698356
			X	X	X	X	X	X	X			GAATTCGCCACCAGAACTACTCAGGT	Chr3	17698722
			X	X	X	X	X	X	X			GAATTCAACACCAATAGTGGATTTAG	Chr3	17698742
			X					X				GAATTCGGTTTATTAATTATGGCAGC	Chr3	17701063
			X	X	X	X	X	X	X			GAATTCAGAATATACATTCCTTACTT	Chr3	17701083
			X	X	X	X	X	X	X			GAATTCCGTCAGTTGTGCACCCATCG	Chr3	17702722
			X	X	X	X		X	X			GAATTCCGCAGGAAACAGTGGTCCAG	Chr3	17702887
			X	X	X	X	X	X	X	X		GAATTCTACTATGGGTCCAACGTATG	Chr3	17705872
			X		X	X	X	X	X	X		GAATTCGTTTTCTACCTTACACATTC	Chr3	17705892
			X		X							GAATTCTTGATCGATATATAGACATG	Chr3	17707204
			X	X	X	X	X	X	X	X		GAATTCATAGAACCTCTAACAAATGT	Chr3	17707224
			X									GAATTCCATCAGATGTGCACCTTATG	Chr3	17708033
			X	X	X	X	X	X	X	X		GAATTCTAGCCGCATTTGATGATGCC	Chr3	17708053
			X	X	X	X	X	X	X	X	X	GAATTCCCCATAAACTAAGCATATAT	Chr3	17718499
			X	X	X	X	X	X	X	X	X	GAATTCCCAAAAGAGTAAGGAAAAAG	Chr3	17718519
			X	X	X	X	X	X	X	X	X	GAATTCGAATCCTTTTGTGCGGTTTC	Chr3	17721809
			X	X	X	X	X	X	X	X	X	GAATTCAACATGTGATCTTCATCTAA	Chr3	17721829