MAPPING AND SEQUENCING COMPLEX GENOMES: LET'S GET PHYSICAL!

Blake C. Meyers, Simone Scalabrin[‡§] and Michele Morgante[‡]*

Physical maps provide an essential framework for ordering and joining sequence data, genetically mapped markers and large-insert clones in eukaryotic genome projects. A good physical map is also an important resource for cloning specific genes of interest, comparing genomes, and understanding the size and complexity of a genome. Although physical maps are usually taken at face value, a good deal of technology, molecular biology and statistics goes into their making. Understanding the science behind map building is important if users are to critically assess, use and build physical maps.

BACTERIAL ARTIFICIAL CHROMOSOME (BAC). A cloning vector derived from a single-copy F-plasmid of *Escherichia coli*. Large genomic fragments (100–200 Kb) can be cloned into BACs, making them useful for constructing genomic libraries.

**Department of Plant and Soil Sciences and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19711, USA.*
[‡]Dipartimento di Scienze Agrarie ed Ambientali,
[§]Dipartimento di Matematica ed Informatica, Universita' di Udine, Via delle Scienze 208, I-33100 Udine, Italy.
Correspondence to M.M.
e-mail: michele.morgante@uniud.it
doi:10.1038/nrg1404

The construction of a whole-genome physical map has been an essential component of numerous genome projects initiated since the inception of the Human Genome Project (HGP). The production and integration of genetic, physical, gene and sequence maps was the goal of the HGP[1]. Although genetic mapping has been pursued in plants and animals for decades, it is only relatively recently that advances in cloning and clone fingerprinting have allowed the construction of physical maps. A physical map is an ordered set of DNA fragments, among which the distances are expressed in physical distance units (base pairs). These days, a physical map usually comprises a set of ordered large-insert clones such as BACTERIAL ARTIFICIAL CHROMOSOMES (BACs)[2], which have largely replaced YEAST ARTIFICIAL CHROMOSOMES[3] as the preferred building blocks of a physical map. Physical maps can be independent of genetic information but are more valuable if linked to genetically mapped markers, and are even more powerful if integrated with genomic sequence data.

Much progress has been made in the development of technologies and strategies for whole-genome sequencing, but these strategies still depend on the development of a physical map. In the clone-by-clone whole-genome sequencing method, the physical map is constructed first, and a MINIMAL TILING PATH of clones is then selected for separate shotgun sequencing of each clone in the path[4].

An alternative to the clone-by-clone method is whole-genome shotgun (WGS) sequencing, which uses assembled sequence data generated randomly from the entire genome[4,5]. In theory, WGS sequencing makes obsolete the process of physical mapping because it should construct overlapping contiguous segments (contigs) of sequence data. However, it is not yet clear whether WGS sequencing alone is sufficient to produce a linearly ordered set of sequences if the sequence contigs are not coupled to a robust physical map[4,6–9]. Therefore, a hybrid strategy of the two methods for whole-genome sequencing will probably prove to be most productive[4]. With this hybrid approach, WGS sequence data are aligned with mapped BAC-end sequences, and these assembled contigs are anchored to a physical map scaffold that comprises ordered and orientated BACs that include mapped molecular markers[10,11].

The lack of high-quality physical maps could rapidly become one of the limiting factors in assembling newly generated WGS sequences for large genomes. The productivity of large sequencing centres has already outstripped the ability of physical mapping laboratories to provide ordered sequence maps. Without the linear order that physical maps provide, the marginal advantage that WGS sequencing projects have over a comprehensive EST or a full-length cDNA sequencing effort does not justify the considerable increase in costs.

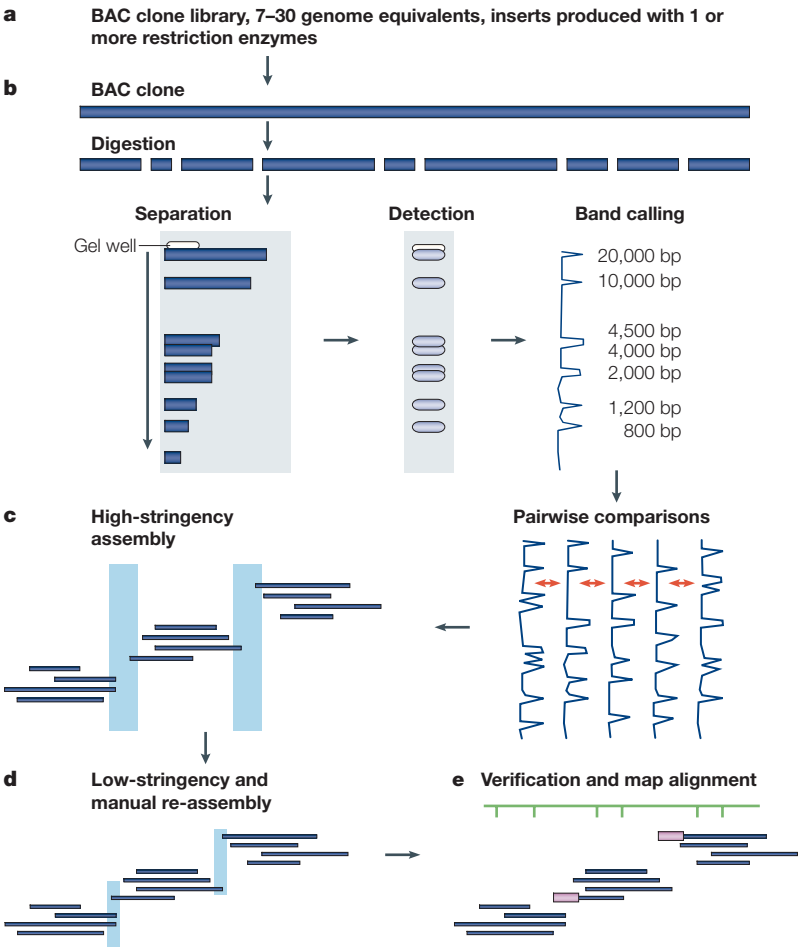


Figure 1 | **The DNA fingerprinting approach to building a whole-genome physical map.** **a** | A bacterial artificial chromosome (BAC) library. A BAC library that represents from 7 to 30 (or more) genome equivalents is constructed. Use of multiple libraries produced with different restriction enzymes will result in better genome coverage. **b** | DNA fingerprinting of BAC clones. Each clone is restriction-enzyme-digested and the resultant fragments are subjected to electrophoresis to produce the DNA fingerprints. Sizes of all DNA fragments detected on gel are estimated for each clone. **c** | Automated assembly. Using appropriate software, a full pairwise comparison of all clones is performed to detect the proportion of shared bands among each pair of clones. Overlapping clones are identified and placed into contigs on the basis of a set threshold (SULSTON CUTOFF SCORE) of a minimum proportion of shared bands. A clone-ordering algorithm is then used to find the most likely relative order of BAC clones within each contig. This high-stringency assembly process results in some overlaps that are not detected (the blue band indicates gaps in the assembly). **d** | Manual curation and assembly. End clones from each contig can be compared with one another at a relaxed cutoff score to detect smaller overlaps that went undetected at the more stringent cutoff score used in the automated assembly (that is, to identify assembly gaps). Misassembled clones can also be detected and removed from the assembly, or contigs can be split if deemed unreliable. **e** | Map alignment and verification. The contigs are aligned to the genetic map or radiation hybrid map using shared markers to verify the map and to further merge contigs.The pink boxes indicate BAC-end sequences that have been used as genetic markers to align contigs to the genetic map.

YEAST ARTIFICIAL CHROMOSOME (YAC). A cloning vector system that can accommodate large genomic fragments (500–1,000 Kb). YACs are grown in yeast, and can be unstable and difficult to isolate in comparison to BACs.

Large-scale mapping and sequencing is underway or planned for many diverse organisms. However, most of these efforts will need to proceed without the vast molecular and financial resources that support organisms such as human, mouse and rat. Physical maps can now be built quickly for many species in which complete genome sequences will not be available soon because a map can be obtained at a fraction of the cost of a whole-genome sequence. Some resources, such as RADIATION HYBRID CELL LINES, were used extensively in the construction of physical maps of mammals, but have so far proved difficult or impossible to develop for other species[12,13]. Several alternative strategies are now being considered to obtain genic sequences in species with large genomes[14]. Two of these strategies, METHYLATION FILTRATION and HIGH C_0T SELECTION, have recently been applied to maize and shown to be valid alternatives to traditional approaches to genomic sequencing[15,16]. However, sequence contigs that are generated by these approaches will have to be ordered on the basis of a genomic scaffold, and this will require a robust physical map. Even in the absence of a whole-genome sequence assembly, a densely populated physical map allows map-based cloning and comparative genomics. Physical maps are also being built for wild relatives of species with a sequenced genome for comparative purposes; this provides a shortcut to address certain questions for which re-sequencing a genome is impractical.

The goal of this review is to provide guidance both in the evaluation of previously constructed physical maps and in the choice of methods used to build a physical map *de novo*. Here, we discuss the different physical mapping techniques and their advantages and disadvantages. In particular, we focus on methods that order large-insert clones rather than those that order markers such as radiation hybrid (RH) mapping[17] or HAPPY MAPPING[18]. Physical maps are often made available through the Internet before publication in refereed journals, and before critical evaluation. Moreover, primary research publications do not evaluate techniques or approaches in a critical or comparative fashion. Here, we aim to address this deficit in critical evaluation to allow potential users to take full advantage of the maps and to help them to understand the science and statistics that lie behind the physical mapping process.

Fingerprinting technologies for physical mapping

Banding patterns on chromosomes might be considered to be the earliest and least detailed form of a physical map, with the complete nucleotide sequence of an organism representing the other extreme. Current physical maps are based on technologies to detect overlaps among BACs. Two distinct approaches are used to identify the overlap among clones, and numerous techniques have been applied for each approach. The first approach is to screen the clones to assess the presence of DNA landmarks. Screening techniques include PCR amplification of short fragments known as 'SEQUENCE-TAGGED SITES' (STSs)[19,20], and hybridization of labelled cDNA clones or short, gene-specific oligonucleotides[21] (see, for example, REF. 22). This approach is laborious, and if used alone to construct a physical map, requires an extremely high density of markers that is impractical for most applications.

Here, we focus on the second approach to physical mapping, which is to use DNA fingerprinting and essentially to perform restriction mapping at a whole-genome level[23]. This approach is better suited to relatively unexplored genomes and is more amenable

to high-throughput methods than the STS/hybridization mapping approach. In the fingerprinting approach (see FIG. 1), each clone is digested into fragments with restriction enzymes, which are then separated and detected. Overlapping clones derived from the same genomic region produce patterns of shared restriction fragments, seen as bands on a gel. The proportion of shared bands is indicative of the degree of overlap. The overlap across numerous clones is then used to order the clones into contigs. Highly repetitive genomes can confound the fingerprinting process, because the repetitive elements can produce identical band sizes and generate false overlaps. Combining information about thousands of DNA landmarks, or markers, that are assigned an order on the chromosomes (through genetic mapping, for example) with the presence of those DNA landmarks on the contigs can allow these contigs to be assembled into a genome-wide physical map. Finishing work to identify clones that span predicted gaps between adjacent contigs will coalesce the contigs into larger scaffolds.

Fingerprinting methods. Modern fingerprinting methods are derivations of classic techniques that used restriction enzymes for early genome projects including *Escherichia coli*, *Saccharomyces cervisiae* and *Caenorhabditis elegans*. The first application of whole-genome fingerprinting was the construction of a physical map of the *C. elegans* genome using cosmid clones[24]. In this study, radioactively labelled restriction fragments were separated on polyacrylamide sequencing gels (FIG. 2a). *Hin*dIII — a 6-bp-recognizing enzyme (a 'rare cutter') — was used for the initial digestion of the clone into fragments, which are then end-labelled. Another digestion with *Sau*3AI — a restriction enzyme that recognizes 4 bp (a 'frequent cutter') — produces smaller fragments that are suitable for separation and detection on sequencing gels. The subset of these fragments that have labelled *Hin*dIII-ends can be detected[24].

Brenner and Livak proposed a second fingerprinting method[25] that uses automated sequencers. This method took advantage of properties of the type IIS restriction enzymes that cut at a precisely defined distance from their recognition site, leaving single-stranded overhangs of variable composition. The overhangs are filled in using unlabelled deoxynucleotides (dNTPs) and fluorescently labelled dideoxynucleotides (ddNTPs) to produce bands that automated sequencers can detect. These machines can resolve band sizes at high resolution and determine the sequence of the 3′ fluorescently labelled bases. The availability of the terminal sequence of these fragments markedly increases the information content of fingerprints compared with the older radioactive methods. This in turn allows more reliable identification of shared fragments.

In a substantially different method[26], large-insert clones are digested with a restriction enzyme — often *Hin*dIII — that recognizes 6 bp, and the resulting fragments are detected on agarose gels stained with ethidium bromide or in a more recent modification with SYBR Green[27] (a highly sensitive DNA dye that is commercially produced by Molecular Probes, Inc., Eugene, Oregon, USA) (FIG. 2b). This third method differs substantially from those described above, because nearly all restriction fragments that are produced from a clone are visible on the agarose gel, whereas the above methods visualize only a subset of fragments that have been labelled and require sequencing gels. The advantage of observing all fragments that result from a clone is that the integrity of the overlap among clones can be verified easily and the size of the overlapping region can be directly estimated rather than just inferred on the basis of the proportion of shared fragments, as with the two other methods. The agarose fingerprinting method has since been widely applied because of its relative simplicity and low costs. This method also has several further advantages that derive from the fact that it is the

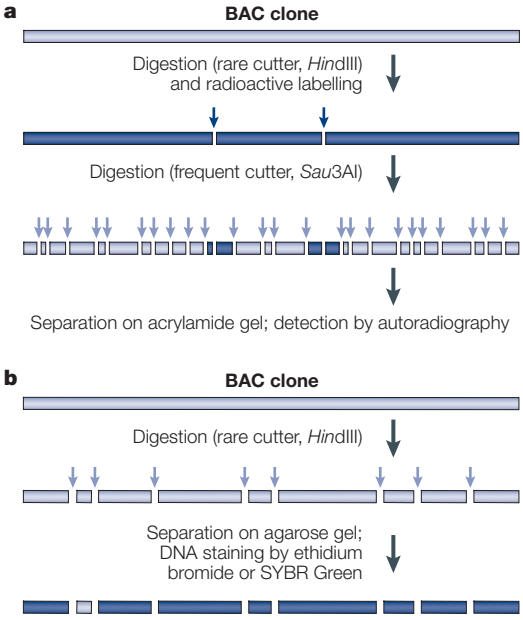


Figure 2 | **Two main DNA fingerprinting methods.** **a** | Schematic overview of the Coulson *et al.* fingerprinting method[24]. BAC clone DNA is digested with the rare cutter *Hin*dIII. The *Hin*dIII-fragment ends are labelled with [^{32}P]dATP using a DNA polymerase or reverse transcriptase. After heat inactivation of the enzymes, fragments are cut again using a frequent cutter, *Sau*3AI, so that they can be separated as single-stranded molecules on a denaturing polyacrylamide gel. Only fragments with at least 1 end produced by *Hin*dIII and that are ~50–600 bp will be visible; fragments with both ends produced by *Sau*3AI (grey fragments) or that are outside this size range will not be detected. A single DNA strand is labelled for each fragment, unless both ends of the fragment are produced by *Hin*dIII, in which case two labelled fragments of the same size are produced. **b** | Schematic overview of the agarose fingerprinting method[24]. BAC clone DNA is digested with the rare cutter *Hin*dIII. Restriction fragments are separated on agarose gels as double-stranded molecules. Detection is achieved by staining the gel with either ethidium bromide[24] or SYBR Green[24]. All fragments will be visible except those that fall outside the resolution range of the gel (grey fragment in the figure), which is between 600 and 25,000 bp. In both methods, external size markers are used to size the fragments. Note that the two panels are not drawn to the same scale.

MINIMAL TILING PATH
A minimal set of overlapping clones that together provides complete coverage across a genomic region.

SULSTON CUTOFF SCORE
A score that expresses the probability that the number of bands matched between any two clones being fingerprinted is a coincidence. Clones are considered to overlap if the score is below a user-supplied threshold (cutoff).

RADIATION HYBRID CELL LINES
A collection of cell lines, each of which is a clonal population of cells that are derived by the fusion of lethally X-irradiated donor cells with mammalian cells. Such cell lines can be used to create a physical map of the donor genome.

METHYLATION FILTRATION
A method that takes advantage of higher DNA methylation in repetitive than in low-single-copy sequences to selectively clone in *Escherichia coli* the latter (hypomethylated) ones that usually represent a gene-enriched fraction.

HIGH C_0T SELECTION
A method that takes advantage of faster re-naturation of repetitive than of low-single-copy sequences to select first and then clone in *Escherichia coli* the latter ones that represent a gene-enriched fraction.

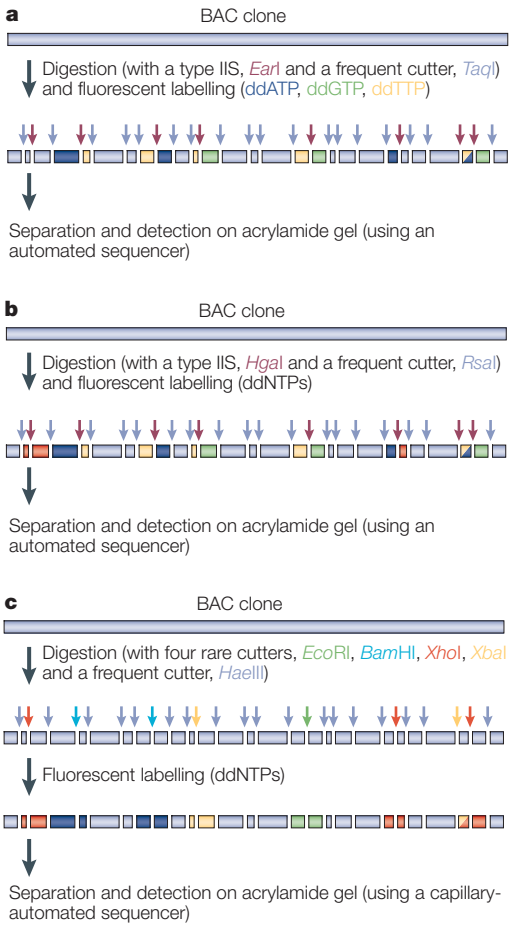
HAPPY MAPPING
A simple method for ordering markers and determining the physical distances between them that uses subhaploid equivalents of randomly sheared DNA and requires the use of whole-genome amplification methods to perform multiple PCR reactions.

SEQUENCE-TAGGED SITES
(STS). Short (for example, <1,000 bp), unique sequence that is associated with a PCR assay that can be used to detect that site in the genome.

Box 1 | Variations on a theme: fingerprinting BAC clones on automated DNA sequencers

Several modifications to the basic fingerprinting methods (see FIG. 2) have been proposed. The possibility of increased throughput, sizing accuracy and/or information content of fingerprinting motivated such developments, as did the possibility of more efficient exploitation of the potential of automated sequencing machines. Initially, the Coulson method was refined to allow high-throughput fingerprinting of rice BAC clones[46,47]. To further refine this method, Klein *et al.*[48] and Tao *et al.*[30] used a frequent cutter that leaves blunt ends (*Hae*III) in place of *Sau*3AI to allow simultaneous digestion and radioactive labelling. Gregory *et al.*[49] adapted the original Coulson method[24] for use on automated sequencers. These authors took advantage of the increased sizing accuracy owing to the use of INTERNAL SIZE STANDARDS and of the possibility of simultaneously digesting with the two enzymes and labelling the fragments. ddATP that was labelled with one of three fluorescent dyes was used to fingerprint three different BAC clones in a single lane, with fragments from each clone labelled with a different dye[49]. This does not increase information content per clone, but reduces the total number of lanes required. Ding *et al.*[50] multiplexed reactions on an automated sequencer: they combined three different double-enzyme digests from a single BAC clone into one lane. The three digests are performed separately, each using *Hin*dIII paired with a different frequent cutter; the fragments from each digestion are differentially labelled and the three reactions are combined before electrophoresis[50]. The information content is not substantially increased because many of the bands are redundant among the three digests due to the fact that the same *Hin*dIII sites are analysed.

The Brenner and Livak[25] method has also been modified. Faller *et al.*[51,52] (see figure part **a**) used a different enzyme combination (*Ear*I as type IIS enzyme and *Taq*I as frequent cutter) and introduced simultaneous labelling and digestion. A single-base extension reaction that involved the variable overhang produced by a type IIS enzyme was used to label the restriction fragments with one of three fluorescently labelled ddNTPs (ddATPs, ddGTPs, ddTTPs). Unlabelled ddCTP is also added to fill in the frequent-cutter ends without the incorporation of dyes that would make these bands visible on gels. The fourth fluorescent colour was used for the internal size standard. This approach varies from the Brenner and Livak method in that only the first base in the overhang, rather than four bases, is sequenced. This lowers the information content per band but increases the number of bands that can be distinguished on an automated sequencer. Ding *et al.*[28] took advantage of the availability of five different fluorochromes to further modify the Faller *et al.* method (see figure part **b**). Five fluorochromes can be used to detect all four nucleotides plus the internal size standard. Luo *et al.*[42] introduced another modification that involved digesting the clones with four rare cutting enzymes and a frequent cutter (see figure part **c**). Each of the rare cutters leaves a different single-stranded overhang that can be filled in with a distinct, labelled ddNTP. The fingerprinting reaction takes place in two steps: digestion followed by labelling. Again, a fifth fluorochrome is used for the internal size standard.



only one that can detect almost all fragments within a clone. First, BAC-clone insert sizes can be determined directly and individually for each clone. This can be important when incorporating BAC-end sequences into genomic assemblies, as well as when assessing the alignment of BAC-end sequence to genome sequence. Second, fingerprint data obtained using this method can be used to verify the accuracy of the sequence assembly, which is an important quality-control step. Finally, deleted or otherwise rearranged BACs might be more reliably detected with agarose gel-based fingerprints, which is pertinent for selecting clones for sequencing that faithfully represent the genome.

Which method? The three methods and their variations (see BOX 1) described above are clearly different in terms of the reaction biochemistry, the information content and the separation medium. The best direct comparison of these physical mapping approaches would involve constructing maps with each method using identical clone libraries and evaluating the resulting assemblies. In lieu of whole-genome experimental comparisons, we carried out simulations using sequenced clones to compare the methods. The *in silico* digestion results of a set of 19 sequenced rice BACs show that all the fluorescent methods produced more than 100 bands on average, and these bands are divided among 3 or 4

INTERNAL SIZE STANDARD
A set of DNA fragments of known size that are run in the same lane as the sample to be sized but distinguishable from the fragments of unknown size. Unlike the external size standards normally used on DNA gels, internal size standards allow for greater accuracy in sizing because they are not affected by lane-to-lane variation in the migration rate.

colours (TABLE 1). The agarose gel method produces the lowest number of bands (36), and the Ding *et al.* method[27] produces the highest (198). The high number of bands produced with the Ding *et al.* method results from the combination of using *Hga*I (recognition site GACGC) and the G+C-rich properties of the rice genome (including low CpG dinucleotide suppression)[28]. By contrast, the low frequency of CpG dinucleotides in human DNA resulted in a much lower number of bands per clone (36) (REF. 28), demonstrating that genome composition will affect the choice of methods and enzymes used in physical map construction. Some empirical data are lower than our calculated averages[29,30], presumably because we did not account for bands of equal or similar size that would mask each other on a gel, because we assumed no other errors in the detection of bands and because we did not adjust the calculation to account for different average BAC sizes.

In our opinion, the average band number in TABLE 1 indicates the difference in information content for each of the fingerprinting methods. Bands can be thought of as anchors that are dispersed along the DNA fragments; the overlap between two BAC clones is estimated from the proportion of shared bands. The accuracy of this estimate directly correlates with the number of anchors observed in each clone, because methods that produce fewer bands generate large relative errors in the estimates of overlap (FIG. 3). Increasing the average number of bands per fingerprinted clone can minimize such errors. However, an excessive number of bands will result in an increased probability of false overlaps that occur if bands randomly match between non-overlapping clones[25] (FIG. 4).

The number of bands is not the only important criterion that determines the rate of false overlaps: an increase in the band resolution or the space that separates the bands can also reduce the error[25]. Labelling the fragments derived from a single clone with three or four different colours effectively increases the gel space by a corresponding factor. We calculated the percentage overlap that is detected at a given Sulston cutoff score (described below) for variable numbers of bands and for different separation methods (FIG. 4). The inverted U-shaped curve for 200 and 300 bands in the agarose gel methods (FIG. 4a) indicates that a high rate of false overlaps is detected if the number of potential fragment sizes is saturated. It is much more difficult to saturate methods that use automated sequencers (FIG. 4b,c). In general, methods that produce a greater number of bands can detect overlaps at a lower cutoff than methods that produce fewer bands (FIG. 4). The benefit of applying the method with higher band numbers is that

Table 1 | **Predicted number of fingerprint bands for rice genomic sequence***

BAC clone Genbank ID	Size of clone (bp)	*Hin*dIII‡	*Hin*dIII and *Hae*III§	*Ear*I and *Taq*I‖	*Eco*RI, *Bam*HI, *Xho*I, *Xba*I and *Hae*III¶	*Hga*I and *Rsa*I#
AP003446	100,635	27	20	76	129	102
AC091680	148,611	44	73	93	117	302
AP003561	183,580	51	81	142	197	235
AC092750	134,933	35	46	106	200	115
AP003853	155,939	34	37	122	119	150
AP003734	154,084	36	56	104	160	263
AP003236	167,399	39	58	126	168	222
AP002912	140,791	33	45	94	150	141
AP003019	153,749	42	54	104	169	223
AP002485	136,150	31	42	92	138	222
AP001129	194,509	45	62	136	167	299
AC078893	81,784	16	21	52	53	111
AC079830	129,655	38	49	111	103	185
AC025296	165,394	40	47	94	164	224
AC084406	140,044	41	56	99	109	222
AL512542	97,076	13	13	85	83	160
AL442110	140,072	40	65	81	157	153
AP001111	175,439	37	59	146	163	248
AC079830	129,655	38	49	111	103	185
Total	2,729,499	680	933	1,974	2,649	3,762
Average	143,658	36	49	104	139	198
bp per band		4,014	2,926	1,383	1,030	726

*In each case, the number of bands was determined by *in silico* analyses of restriction sites. For all methods, only bands within the detectable size ranges are reported as from the respective references. ‡Based on the method described by Marra *et al.*[27]. §Based on the method described by Tao *et al.*[30]. This analysis included both *Hin*dIII/*Hae*III bands as well as *Hin*dIII/*Hin*dIII bands, as both types of fragment would be labelled using the Tao *et al.*[30] methodology. Only bands in the size range of 58–673 bp were counted, corresponding to the range measured by Tao *et al.*[30]. ‖Based on the method described by Faller *et al.*[51,52]. ¶Based on the method described by Luo *et al.*[42]. #Based on the method described by Ding *et al.*[28]. BAC, bacterial artificial chromosome.

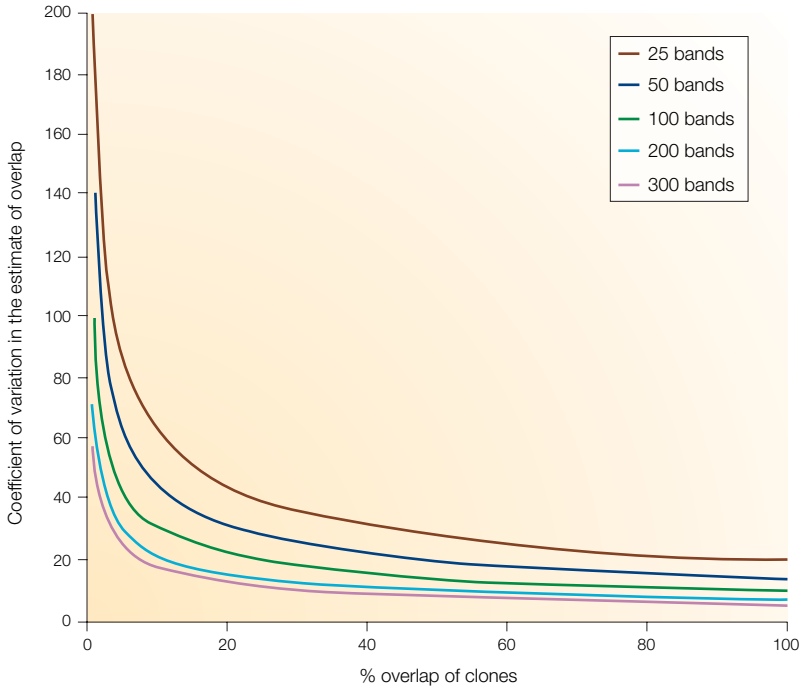


Figure 3 | **Variation in estimates of clone overlap depends on the number of bands in the fingerprint.** The coefficient of variation in the estimate of overlap (*y*-axis) as a function of the real overlap, expressed as a percentage (*x*-axis), is plotted for a range of possible band numbers in the bacterial artificial chromosome (BAC) fingerprints. The proportion of shared fragments/bands was used as an estimate of the percentage of overlap among clones. For this calculation, the total number of bands obtained from a given clone was combined, regardless of the number of fluorochromes used to detect the bands. The curves were calculated by assuming a Poisson distribution of the number of shared bands.

fewer false overlaps are included in the final assembly of the map (discussed below). A corollary to this conclusion is that the ability to detect smaller overlaps at a particular cutoff (the result of increasing the number of bands; FIG. 4) results in a smaller number of gaps in the assembled map. As shown in FIG. 4, if the fingerprinting technique can be modified to generate a particular number of bands at a pre-selected cutoff, this will allow the detection of the smallest amount of overlap among clones. The number of bands could be altered either by selecting restriction enzymes with a specific G+C ratio, or by selecting an entirely different method.

We believe that these data indicate that fluorescence-based fingerprinting has technical advantages over other methods. However, other considerations might also influence the choice of methods. Owing to the differences in efficiency and information content, methods that use radioactivity are clearly being abandoned in favour of those that use fluorescently labelled samples. Agarose separation is not desirable for a high-resolution map; however, it is occasionally chosen instead of more automated methods owing to the high initial cost of automated sequencers or because it has historically been used in many laboratories.

All simulations presented here assume perfect fingerprints with no errors. Experimental errors can arise from partial restriction digestion, STAR ACTIVITY of restriction enzymes (which has been described for some of the restriction enzymes used in the fingerprinting methods in BOX 1), contamination of the BAC preparation with *E. coli* genomic DNA, overlapping fragments (that is, fragments that cannot be separated on the basis of separation medium resolution) and lane mis-tracking. Although a detailed discussion of the errors in fingerprints and their implications for map assembly is beyond the scope of this review, it must be kept in mind that errors will depend on the method chosen, that they will only make the situation worse than our predictions from the simulations and that they will make the assembly more complicated. Many of these errors can be avoided with the use of proper laboratory practices or can be recognized by the use of adequate software tools. Some of these tools are implemented in the GenoProfiler software package (see online links box). These issues should be carefully considered when evaluating a physical map.

Assembly strategies and evaluation

The assembly from the fingerprint data of the contigs that comprise a physical map is a complex statistical construction (FIG. 1c). This assembly demands rigour and proper statistical testing; the parameters used for map assembly are one of the primary gauges of the quality of the contigs that comprise the map. The two primary parameters used in the only currently available automated physical map assembly software — FingerPrinted Contigs; FPC — are the tolerance and the Sulston cutoff score[31–33]. Fingerprinting bands are considered to be 'shared' if they have the same size within a given tolerance; the probability (p) that n bands are coincidentally shared between any two clones is computed from the formula described by Sulston *et al.*[33]. The Sulston cutoff is based on the binomial probability distribution. If this probability is below a user-defined cutoff, the two clones are declared to be overlapping. Two non-overlapping clones might have a coincidence score below the cutoff, producing a false positive (a type I error), or two overlapping clones might have a coincidence score above the cutoff, resulting in a false negative (a type II error). Type I errors will merge non-overlapping clones and create chimeric contigs that represent the most egregious problems in a physical map, whereas type II errors simply create extra gaps in the assembly. The cutoff must be set to minimize both false positives and false negatives[31].

The tolerances and cutoffs in the automated assembly software must be precisely calibrated and verified for a given assembly, depending on the size and complexity of the genome. Genome size affects the Sulston cutoff settings in physical map assemblies because the number of false positives depends on the total number of pairwise comparisons performed among all BAC fingerprints. In other words, the Sulston cutoff settings need to be adjusted for the multiple comparisons made among the BAC clones, and the number of BAC clones needed for a given coverage of a genome will vary depending on the genome size. The overall p-value (Sulston cutoff) is less stringent than the nominal p-value set for the assembly because of the adjustment for multiple comparisons (a BONFERRONI CORRECTION).

STAR ACTIVITY
The activity of restriction endonucleases under non-standard conditions that results in cleavage at sequences that are similar but not identical to their defined recognition sequence. The degree and type of this altered specificity varies among enzymes and reaction conditions.

BONFERRONI CORRECTION
A multiple-comparison correction to the significance level α that is used to avoid many spurious positives (type I errors) when several independent statistical tests are being performed simultaneously.

LONG TERMINAL REPEAT RETROTRANSPOSONS (LTR retrotransposons).
Transposable elements that move through an RNA intermediate, are related to retroviruses and possess direct repeats at their ends (long terminal repeats, LTRs).

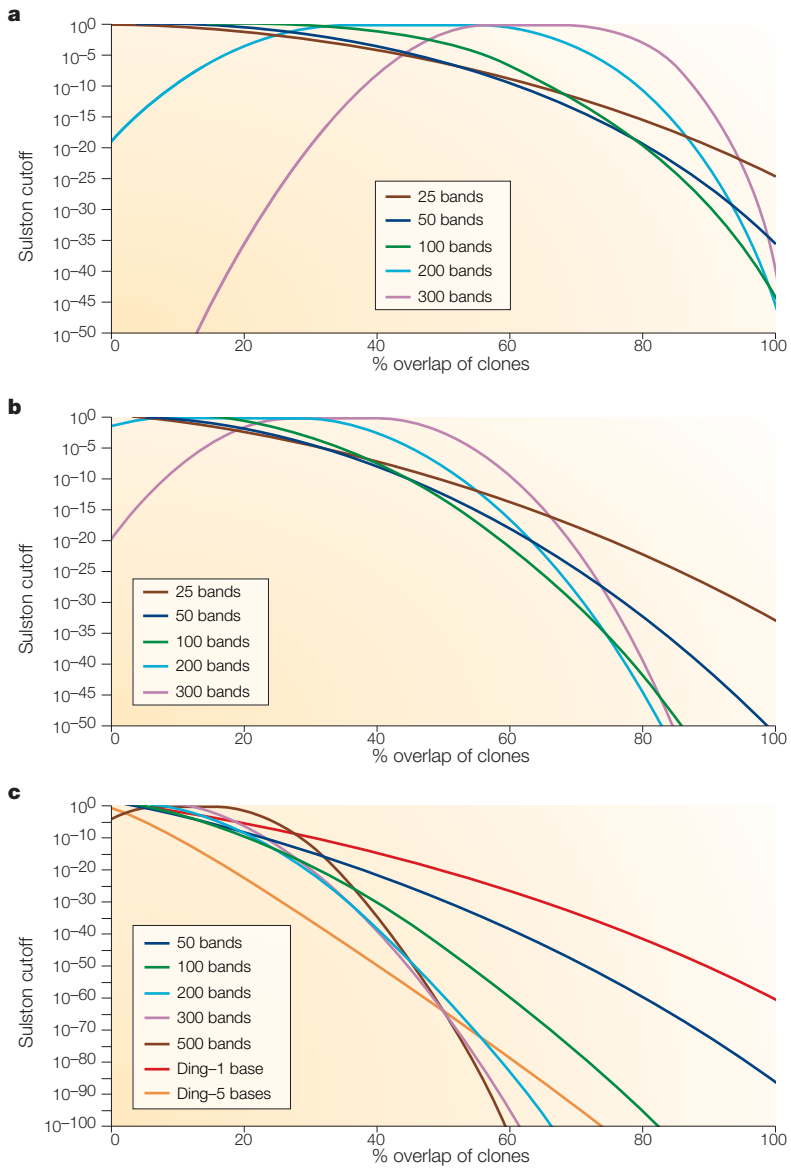


Figure 4 | **Optimal number of bands per clone (or bands per colour per clone) and detectable overlap at a given cutoff.** The Sulston cutoff score that was used for the map assembly (*y*-axis) is shown as a function of the percentage of clone overlap (*x*-axis) required to obtain that cutoff. The proportion of shared fragments/bands was used as an estimate of the percentage of overlap among clones. The cutoff is calculated assuming a comparison between two bacterial artificial chromosome (BAC) clones that have the same number of bands, and the relationship is shown for varying numbers of bands. Different fingerprinting technologies were plotted separately because each uses a different effective gel length and a different tolerance or size range within which two bands are assumed to be coincident. For the methods shown in (**c**), which use multiple fluorochromes for a single clone, the gel length is multiplied by the number of fluorochrome combinations that can be carried by each fragment. The formula that was used for these plots is described in Sulston *et al.*[33]. The gel length and tolerance that were used in the calculations are reported for each method and are derived from the corresponding publications. **a** | Agarose gel fingerprinting method[27,30], calculated with tolerance = 7, gel length = 3,300. **b** | Radioactive method of Tao *et al.* (2001), calculated with tolerance = 3, gel length = 3,300. **c** | Fluorescent method of Faller *et al.*[51,52] using 3 fluorochromes for the fragments, calculated with tolerance = 0.25, gel length = 1,350 (450 for each of 3 fluorochromes). As a comparison, 2 lines were calculated using the method of Ding *et al.*[28] for 4 fluorochromes, assuming either the labelling of only the first base in the overhang (tolerance = 0.5, gel length = 1,700, from a gel length of 425 for each of 4 fluorochromes; labelled 'Ding–1 base') or the labelling of all 5 bases in the overhang (tolerance = 0.5, gel length = 435,200, from a gel length of 425 for each of 1,024 combinations of fluorochromes; labelled 'Ding–5 bases'). For the Ding method, 36 bands per clone were assumed as from the published data on human BACs[28]. The Ding–1-base method results are also representative of the Luo *et al.* method[28] for a similar number of bands.

With an increased genome size, the cutoff needs to be lowered by a factor that is equivalent to the square of the increase factor (for example, a 5-fold difference in genome size requires a 25-fold lower cutoff). This adjustment will increase the occurrence of false negatives (type II errors) and decrease the likelihood of detecting the same amount of overlap between any two BAC clones in the larger genome compared with the smaller genome. Using the same settings for assembling genomes of different sizes will produce different overall error rates in the physical map.

The degree of repetitiveness of the genome should also influence the cutoff setting in physical map assemblies. The Sulston cutoff formula assumes an equal probability of observing any single fingerprint band size. Even in a genome that comprises predominantly single-copy sequences, this assumption might be violated because the size distribution of the restriction fragments is not uniform across the length of the gel. Highly repetitive genomes present further difficulties because of the prevalence of certain band sizes derived from repetitive elements that are shared by non-overlapping clones. The long, recently amplified repeats of LONG TERMINAL REPEAT RETROTRANSPOSONS (LTR retrotransposons) will exacerbate this problem. For example, in maize, the 3 most abundant LTR-retrotransposon families each make up ~10% of the genome[34], so ~30% of fingerprint bands could be shared even between many non-overlapping BAC clones. Shorter or more diverse (ancient) repetitive elements will pose less of a problem. The cutoff must be sufficiently stringent to reject most, if not all, false overlaps and this will vary according to genome complexity.

The products of automated assemblies are traditionally confirmed by manual curation and analysis of the maps (FIG. 1d). Typically, the automated assemblies are performed at high stringency to avoid false assemblies during the combinatorial comparison of fingerprints, which could lead to the production of large, illegitimate contigs that have disastrous consequences for the quality of the final map. Mis-mapped and illegitimate contigs can be detected after the initial binning of clones into contigs owing to the clone-ordering algorithm implemented in FPC. It is, however, standard practice to start the assembly at a high level of stringency, and then lower the stringency to consolidate the map. In the manual curation step, the stringency of assembly is relaxed, but comparisons are limited to those that involve BAC clones that are located at the ends of contigs. This focuses the lower-stringency assembly on the joining of already-assembled contigs. The manual curation process is greatly facilitated if supporting data from genetic or radiation hybrid maps are available together with the location of the corresponding markers on the BAC clones.

Physical map evaluation

Estimating genome coverage of a map. A physical map's genome coverage can be estimated from the total physical size of the contigs that comprise the map after some level of fingerprinting. This physical size is compared

Table 2 | **Assembly parameters for physical maps that are finished and underway***

Map	Number of BACs assembled	BAC coverage‡	Sulston cutoff	Overall *p*-value§	Number of contigs (unedited)	Reference
Published physical maps						
Human	283,287	15x	$3x10^{-12}$	0.1203773	7,133	36
Mouse	305,716	33x	10^{-16}	0.0000047	7,587	35
D. melanogaster	10,253	14x	10^{-10}	0.0052562	Not reported	53
A. thaliana I	20,206	17x	10^{-9}	0.2041412	372	54
A. thaliana II	9,389	7.2x	10^{-12}	0.0000441	196	55
Rice I	21,087	6.9x	10^{-10}	0.0222331	585	30
Rice II	65,287	20x	10^{-12}	0.0021312	1,019	29
Sorghum	22,233	4x	$5x10^{-14}$	0.0000124	3,345	48
Soybean	78,001	9.6x	10^{-30}	3.042×10^{-20}	4,792	56
Rat	189,689	13.1x	10^{-17}	1.799×10^{-7}	11,274	11
B. japonicum	4,608	77x	10^{-13}	1.061×10^{-6}	6	57
Physical maps currently underway						
Maize	291,569	15x	10^{-9} to 10^{-12}	42.506095 0.042506	4,518	58; ¶
Maize II	305,849	16x	10^{-50}	5.0×10^{-40}	4,681	#
Wheat	267,451	7.5x	~10^{-30}	21.0×10^{-19}	13,647	42; **
Cow	~200,000	12x	N/A	N/A	N/A	‡‡
Poplar	~50,000	10x	N/A	N/A	N/A	‡‡

*Includes only maps built from fingerprinted bacterial artificial chromosome (BAC) clones. This is not an exhaustive list, and additional maps are underway. Some of the maps in progress are reported in abstracts from the International Plant and Animal Genome Conference (http://www.intl-pag.org). ‡Based on total library size, including some BACs for which fingerprints failed. §Overall *p*-value = $(n(n-1)/2) \times$ Sulston cutoff, where *n* is the number of BACs. ¶http://www.genome.arizona.edu/fpc/maize; #http://www.genome.arizona.edu/fpc_hicf/maize; **http://wheat.pw.usda.gov; ‡‡http://www.bcgsc.bc.ca. *A. thaliana*, *Arabidopsis thaliana*; *B. japonicum, Bradyrhizobium japonicum*; *D. melanogaster*, *Drosophila melanogaster*; N/A, not available.

with the C-value (or DNA content) of a haploid genome to determine genome coverage. The total number of contigs in the map will result from both 'assembly' and 'physical' gaps. Assembly gaps result from the false-negative rate that is determined with the choice of the cutoff and correspond to our inability to detect existing overlaps between clones. Physical gaps result from regions that are not covered in the clone collection. Because clone distribution is mainly random, the final map will probably contain both densely covered regions and significant gaps. Increasing the number of genome equivalents that are represented in the library will decrease the number of assembly and physical gaps. Physical maps are usually built from libraries that contain at least ten genome equivalents. Methods with low information content might require more genome equivalents in the library because they fail to detect small overlaps. A biased set of clones, such as those produced by restriction enzyme digestion, will have even larger gaps and deeper coverage in some regions of the genome than a set of clones produced in a completely random way, such as through mechanical shearing. The use of different libraries produced by digesting genomic DNA with different enzymes will reduce physical gaps that might result from biased restriction site distribution.

The average insert size in a BAC library is a crucial value in the calculation of library coverage. Too often, this size estimate is obtained on the basis of only a small sample of BACs that have been sized using PULSED FIELD GELS. The pulsed field gel estimates provide only rough estimates of the average insert size, which might be sufficiently accurate for measurements of small sets of clones; however, the extrapolation of the whole-genome coverage of a library can be confirmed and refined using additional data. The physical map itself is a significant resource for estimating BAC insert sizes. The map provides at least three parameters that can be used to verify BAC insert sizes: first, the average number of positives identified by each single-copy probe; second, the proportion of single-copy probes that are found in the libraries; and third, the number of fingerprinting bands observed per clone. The frequency of positive clones is a direct measurement of the representation of a single locus. These results averaged over many single-copy loci provide a robust measurement of genomic coverage. The observed percentage of the markers that are identified in the library can be entered into the Lander and Waterman[6] formula to determine the clone coverage. The average frequency of positive clones per probe is related to the percentage of single-copy probes found in the library, but is a more direct measurement of genome coverage. The insert size of the BAC clones can also be estimated from the number of fingerprinting bands observed per clone. The observed number of bands must be combined with the calculated frequency of restriction sites in genomic DNA, but this can provide an accurate estimate for insert size. Therefore, estimates of BAC insert sizes and library coverage might be derived from diverse and independent data sets, and these

PULSED FIELD GEL
Agarose electrophoresis gel that is run by periodically changing the orientation of the electric field applied to the gel to achieve separation of large fragments of DNA (>20 Kb and up to 10 Mb).

parameters are directly relevant to the construction and evaluation of the physical map.

Assessing physical map quality. There are several robust methods that use genetic data to assess the quality of physical map assemblies. For example, genetically linked markers can be localized on BAC clones by hybridization[21], and these markers should co-localize on large contigs (FIG. 1e). The corollary is also true: genetically unlinked or distant markers should not co-localize on the physical map. Genetic mapping can also be applied *a posteriori* to validate physical maps; by using sequences from both ends of large contigs, markers can be designed that should genetically co-segregate, verifying that the DNA that is contained in the clones is contiguous in the genome. Collinearity between the order of markers placed on the contigs in the physical map and their order on genetic or radiation hybrid maps is a good way to assess the correctness of an assembly. Sequence duplications that involve genes are frequent in higher eukaryotes, and can make comparisons of collinearity more difficult; a sequence or probe might detect multiple locations on the physical map but only those that are polymorphic can be placed on a genetic map.

Physical maps of plant and animal genomes

At least nine whole-genome physical maps have been reported that use fingerprinting for a significant portion of the map construction (TABLE 2). Additional maps are underway in several organisms. The assembly of the human and mouse genomic sequences relied heavily on combined resources of a physical map, genetic maps, radiation hybrid data and SYNTENY analyses for assembly[35,36] (FIG. 1e). For example, after an initial contig assembly based on fingerprint data, 305,716 mouse BAC clones produced 7,587 contigs[35]. Integration of genetic and comparative genomic data allowed these contigs to be further collapsed. BAC-end sequences from these clones were simultaneously obtained and then compared with the assembled human sequences, creating a human–mouse homology clone map. Such maps might be possible in some plant species (for example, rice–maize or *Arabidopsis thaliana*–*Brassica* spp.), but the diversity of most genomes for which physical maps are needed could confound cross-genome comparison efforts. Integration of mapped mouse markers from already-constructed genetic and radiation hybrid maps, followed by manual contig editing, reduced the mouse clone map to 296 contigs[35]. The result of this integrative and comparative effort is striking in that it produced a 25-fold reduction in the number of contigs; however, this might not be possible in many other genomes because it requires the concurrent availability of dense genetic or physical STS maps and genomics resources as well as considerable manual intervention. It is preferable to make the *a priori* decision to use the most information-rich fingerprinting method available to reduce the number of contigs that are produced after fingerprint assembly.

Some genomes present unique mapping opportunities that can take advantage of specific biological resources. For example, FISH, applied to the large POLYTENE CHROMOSOMES of the salivary glands of *Anopheles gambiae* (mosquito), was used to assign the sequence scaffolds that were generated by the whole-genome sequence to chromosomal locations[37]. Chromosome-addition lines are available for some plant species such as oat–maize addition lines[12]. Hexaploid wheat tolerates chromosomal deletions, and these ANEUPLOID lines can similarly facilitate mapping[38,39]. Only plants that have undergone recent polyploidization can generally tolerate these types of chromosomal aberration, so these resources are available in only a small number of important crop plants. The main limitation of any of these chromosomal variants is that the resolution is limited to entire chromosomes, compared with the sub-chromosomal resolution of the radiation hybrid maps available in animal systems[13].

Future prospects

Physical maps are often the starting point for laborious and expensive undertakings such as chromosome walks and genome sequencing. Poorly executed whole-genome efforts are little better than no mapping at all, as mis-assembly of a small fraction of clones can endanger the entire project, potentially requiring it to be repeated. These maps are difficult to verify, and represent 'one-of-a-kind' experiments because of the cost and labour involved. However, owing to advances in technology, it could be possible to add a second level of more accurate fingerprinting data on top of an initial round of fingerprinting. For example, if a low-resolution fingerprinting methodology identifies highly overlapping and redundant clones, a high-resolution method could be used selectively to assemble 'singletons' (unassembled individual clones), contigs with poor coverage and end-clones of robustly assembled contigs. This second phase of fingerprinting would require clone selection, and although further handling of clones always increases error rates, it would improve the assembly. Alternatively, the entire library of clones could be re-fingerprinted at extra cost but with reduced error rates. More focused mapping projects might be possible in localized regions of particular interest, using BACs identified by screening with co-localized markers. The fingerprinting of this subset of BACs could be used to develop sequence-ready contigs in a region of particular interest. Although the focused approaches are useful for specific projects, they might be redundant with whole-genome efforts that could not avoid re-assembling these regions. We believe that the best strategy for physical mapping is to use the most information-rich technology with large-insert BACs that provide deep coverage of a genome, such as the tenfold coverage or more used in most genome projects (TABLE 2; for a more formal discussion of the consequences of different genome coverages on the assembly gaps, see REFS 6,40). A powerful, but as-yet untried strategy would be to take advantage of the sizing accuracy of automated sequencers to combine fingerprint data and band sizes with sequence data to

SYNTENY
The conservation of the relative order of genes (or of other DNA sequences) in the chromosomes of different species.

FISH
(Fluorescence *in situ* hybridization). A technique in which a fluorescently labelled DNA probe is used to detect and localize a particular sequence on a chromosome with the help of fluorescence microscopy.

POLYTENE CHROMOSOMES
A giant chromosome that is formed by many rounds of replication of the DNA. The replicated DNA molecules tightly align side-by-side in parallel register, which creates a non-mitotic chromosome that is visible by light microscopy.

ANEUPLOID
Having an unbalanced chromosome number (owing to extra or missing chromosomes). An example is trisomy.

co-assemble physical maps with contigs from whole-genome shotgun sequencing. Regardless of the method used for characterizing clones, it is essential that physical maps are accurately assembled and that rigorous and explicit criteria are applied for their assessment and evaluation. The time spent in producing and verifying a robust physical map is well-spent because it provides an invaluable tool for molecular studies in a genome of interest that can be readily transferred across laboratories. For example, research in crop plants and domesticated animals will benefit immeasurably from the production of physical maps in key species of interest.

Technological advances in sequencing technologies have provided the means for parallel advances in fingerprinting. These advances, and specifically the development of capillary sequencers using multi-coloured fluorescent dyes for labelling DNA, should make the construction of physical maps much easier, faster and cheaper. Although these advances were recognized and anticipated more than 14 years ago[25], it is only recently that they have been incorporated into large-scale physical mapping projects. Most of the recently published physical maps have used agarose gels and conventional DNA stains. This low-resolution method has been shown to be useful and has recently been improved through the development of software for automated analysis of DNA fingerprinting gels[41]; however, in the era of capillary sequencers, this method is rapidly becoming outdated. The advent of high-throughput fluorescent methods[25,42] is a boon to future physical mapping projects, as the high-resolution data will reduce the number of contigs produced before manual assembly and because the work can be done with only a few people in a small but heavily automated laboratory. A team of 3 people can easily process 1,600 clones per day using a single new-generation automated sequencer such as the Applied Biosystems 3730 (Applera, California) (M.M., unpublished observations). At this rate, a physical map can be constructed for many genomes in a matter of weeks rather than months or years. Such technical advances will decrease costs and improve the efficiency of physical map construction.

Software development has kept pace with technology development. The availability of a parallelized version of the map assembly program FPC represents a clear advantage in terms of the ability to quickly build multiple assemblies at different stringencies for evaluation and comparison purposes[43]. New tools have been developed to facilitate the viewing of fingerprint maps, such as internet Contig Explorer (iCE)[44], but physical map viewers are still a limiting factor, as most of the available genome browsers are sequence-centred and do not allow for an easy link of the physical and genetic maps through shared markers.

Physical maps will be fundamental components of future genome-sequencing projects in many species. For eukaryotic species that do not receive enough support for complete genome sequencing, physical maps will be invaluable resources for the cloning of genes of importance. The molecular basis of quantitative variation is still largely unknown and untapped; the discovery of the genes that underlie this variation is currently feasible only through positional cloning efforts[45]. Physical maps, when not contributing to genomic sequencing efforts, will accelerate positional cloning projects in many genomes.

1. Collins, F. & Galas, D. A new five-year plan for the U.S. Human Genome Project. *Science* **262**, 43–46 (1993).
2. Shizuya, H. *et al.* 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 (1992).
3. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806–812 (1987).
4. Green, E. D. Strategies for the systematic sequencing of complex genomes. *Nature Rev. Genet.* **2**, 573–583 (2001).
5. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
7. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101**, 1916–1921 (2004).
8. Myers, E. W., Sutton, G. G., Smith, H. O., Adams, M. D. & Venter, J. C. On the sequencing and assembly of the human genome. *Proc. Natl Acad. Sci. USA* **99**, 4145–4146 (2002).
9. Waterston, R. H., Lander, E. S. & Sulston, J. E. On the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* **99**, 3712–3716 (2002).
10. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
A good example of the use of multiple types of information, including those obtained from a physical map, in a hybrid approach to assemble the sequence of a complex eukaryotic genome.
11. Krzywinski, M. *et al.* Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. *Genome Res.* **14**, 766–779 (2004).
12. Kynast, R. G., Okagaki, R. J., Rines, H. W. & Phillips, R. L. Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics. *Funct. Integr. Genomics* **2**, 60–69 (2002).
13. Wardrop, J., Snape, J., Powell, W. & Machray, G. C. Constructing plant radiation hybrid panels. *Plant J.* **31**, 223–228 (2002).
14. Mayer, K. & Mewes, H. W. How can we deliver the large plant genomes? Strategies and perspectives. *Curr. Opin. Plant Biol.* **5**, 173–177 (2002).
15. Palmer, L. E. *et al.* Maize genome sequencing by methylation filtration. *Science* **302**, 2115–2117 (2003).
16. Whitelaw, C. A. *et al.* Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**, 2118–2120 (2003).
17. Lawrence, S., Morton, N. E. & Cox, D. R. Radiation hybrid mapping. *Proc. Natl Acad. Sci. USA* **88**, 7477–7480 (1991).
18. Dear, P. H. & Cook, P. R. Happy mapping: linkage mapping using a physical analogue of meiosis. *Nucleic Acids Res.* **21**, 13–20 (1993).
19. Olson, M., Hood, L., Cantor, C. & Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435 (1989).
20. Green, E. D. & Green, P. Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods Appl.* **1**, 77–90 (1991).
21. Ross, M. T., Labire, S. M., McPherson, J. & Stanton, J., V. in *Current Protocols in Human Genetics* (ed. Boyl, A.) 5.6.1–5.6.52 (Wiley, New York, 1999).
22. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 271–275 (1999).
23. Smith, C. L. & Cantor, C. R. Evolving strategies for making physical maps of mammalian chromosomes. *Genome* **31**, 1055–1058 (1989).
24. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
A pioneering paper that outlined the general strategy to produce DNA fingerprints from large-insert clones, and applied it to a small eukaryotic genome, effectively providing the first contig map.
25. Brenner, S. & Livak, K. J. DNA fingerprinting by sampled sequencing. *Proc. Natl Acad. Sci. USA* **86**, 8902–8906 (1989).
Another pioneering paper that exploited the properties of type IIS restriction enzymes to differentiate fingerprint fragments not only by their size, but also by their terminal sequence.
26. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
27. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**, 1072–1084 (1997).
The original paper that described the large-scale application of BAC fingerprinting on agarose gels, which has since found widespread use and produced many physical maps.
28. Ding, Y. *et al.* Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using type IIS restriction endonucleases. *Genomics* **74**, 142–154 (2001).
29. Chen, M. *et al.* An integrated physical and genetic map of the rice genome. *Plant Cell* **14**, 537–545 (2002).
30. Tao, Q. *et al.* Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. *Genetics* **158**, 1711–1724 (2001).
31. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000).
A paper that presents the FPC (FingerPrinted Contigs) software, which is the only software that is used for the assembly of physical maps, and extends the theoretical framework put forward in reference 32.
32. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13**, 523–535 (1997).
33. Sulston, J. *et al.* Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4**, 125–132 (1988).

34. Meyers, B. C., Tingey, S. V. & Morgante, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660–1676 (2001).
35. Gregory, S. G. *et al.* A physical map of the mouse genome. *Nature* **418**, 743–750 (2002).
The construction of this physical map shows the advantages of having multiple types of data available in the assembly and editing of the map, including the availability of the genome sequence of a syntenic species.
36. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).
37. Holt, R. A. *et al.* The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149 (2002).
38. Endo, T. & Gill, B. The deletion stocks of common wheat. *J. Hered.* **87**, 295–307 (1996).
39. Akhunov, E. D. *et al.* The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res.* **13**, 753–763 (2003).
40. Wendl, M. C. & Waterston, R. H. Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. *Genome Res.* **12**, 1943–1949 (2002).
41. Fuhrmann, D. R. *et al.* Software for automated analysis of DNA fingerprinting gels. *Genome Res.* **13**, 940–953 (2003).
42. Luo, M. C. *et al.* High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378–389 (2003).
43. Ness, S. R., Terpstra, W., Krzywinski, M., Marra, M. A. & Jones, S. J. Assembly of fingerprint contigs: parallelized FPC. *Bioinformatics* **18**, 484–485 (2002).
44. Fjell, C. D., Bosdet, I., Schein, J. E., Jones, S. J. & Marra, M. A. Internet Contig Explorer (iCE) — a tool for visualizing clone fingerprint maps. *Genome Res.* **13**, 1244–1249 (2003).
45. Morgante, M. & Salamini, F. From plant genomics to breeding practice. *Curr. Opin. Biotechnol.* **14**, 214–219 (2003).
46. Zhang, H. B. & Wing, R. A. Physical mapping of the rice genome with BACs. *Plant Mol. Biol.* **35**, 115–127 (1997).
47. Hong, G. A rapid and accurate strategy for rice contig map construction by combination of fingerprinting and hybridization. *Plant Mol. Biol.* **35**, 129–133 (1997).
48. Klein, P. E. *et al.* A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Res.* **10**, 789–807 (2000).
49. Gregory, S. G., Howell, G. R. & Bentley, D. R. Genome mapping by fluorescent fingerprinting. *Genome Res.* **7**, 1162–1168 (1997).
50. Ding, Y. *et al.* Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. *Genomics* **56**, 237–246 (1999).
51. Faller *et al*. Construction of a contig-based physical map of corn using fluorescent fingerprinting technology. The Plant and Animal Genome VIII Conference, San Diego [online], <http://www.intl-pag.org/pag/8/abstracts/pag8265.html> (2000).
52. Faller, M. L. *et al.* in *Genome Sequencing and Biology* (eds Bogusk, M., Brown, S. & Gibbs, R.) 72 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000).
53. Hoskins, R. A. *et al.* A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science* **287**, 2271–2274 (2000).
54. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22**, 265–270 (1999).
55. Chang, Y. L. *et al.* An integrated map of *Arabidopsis thaliana* for functional analysis of its genome sequence. *Genetics* **159**, 1231–1242 (2001).
56. Wu, C. *et al.* A BAC- and BIBAC-based physical map of the soybean genome. *Genome Res.* **14**, 319–326 (2004).
57. Tomkins, J. P. *et al.* A marker-dense physical map of the *Bradyrhizobium japonicum* genome. *Genome Res.* **11**, 1434–1440 (2001).
58. Coe, E. *et al.* Access to the maize genome: an integrated physical and genetic map. *Plant Physiol.* **128**, 9–12 (2002).

Acknowledgements

We thank many colleagues at DuPont Crop Genetics–Genomics for helpful discussions on physical mapping. The physical mapping work in M.M.'s laboratory is supported by funding from Provincia Autonoma di Trento.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

FURTHER INFORMATION

FPC: http://www.genome.arizona.edu/software/fpc/
GenoProfiler:
http://wheat.pw.usda.gov/PhysicalMapping/tools/genoprofiler/genoprofiler.html
iCE: http://www.bcgsc.ca/bioinfo/ice
Access to this links box is available online.

Author biographies

Blake Meyers is an assistant professor in the Department of Plant and Soil Sciences at the University of Delaware, USA, with a laboratory in the Delaware Biotechnology Institute. He received his B.A. from the University of Chicago, USA, where he first started research in plant biology in the laboratories of Deborah Charlesworth and Manfred Ruddat. He earned an M.S. and a Ph.D. in genetics at the University of California (UC) at Davis, USA. At UC Davis, he worked in Richard Michelmore's laboratory, chromosome walking and physical mapping in the region of the Dm3 disease-resistance gene in lettuce. He then moved to the DuPont Genomics group in Wilmington, Delaware, and worked on physical mapping and genome composition in maize. He later moved back to Richard Michelmore's laboratory to study disease-resistance genes in the model plant *Arabidopsis thaliana*. At present, his laboratory focuses on gene-expression analysis using massively parallel signature sequencing and his laboratory continues to study plant disease-resistance genes.

Simone Scalabrin received his masters degree in computer science at the Universita' di Udine, Italy, and has worked in bioinformatics on a physical mapping project in the grape genome, developing software for data acquisition and analysis. At present, he is a Ph.D. student in computer science in the Department of Mathematics and Computer Science at the Universita' di Udine, developing methods to identify regulatory motifs in DNA sequences.

Michele Morgante is Professor of Genetics in the Dipartimento di Scienze Agrarie ed Ambientali at the Universita' di Udine, Italy. After graduating from the Universita' di Padova, he spent two years as a post-doctoral researcher in the DuPont Genomics group in Wilmington, Delaware, USA, focusing on microsatellite analysis and applications in plants. He then joined the Dipartimento di Produzione Vegetale at the Universita' di Udine, and after a few years, he moved back to DuPont Genomics as a senior scientist, working on genome organization and physical mapping in the maize genome. Since moving back to Italy once more, his laboratory has focused on genome analysis in plants, including physical mapping and genome-evolution studies, and on sequence-diversity analysis and association mapping. He is particularly interested in new approaches to investigating complex traits and in developing the genomics technologies needed for this.

Online summary

- Whole-genome sequencing, positional cloning and comparative genomics mainly depend on the construction of high-quality physical maps.
- Current physical maps have been developed using agarose gel-based or, more recently, acrylamide gel-based techniques for fingerprinting large-insert clones.
- Fingerprinting methods based on fluorescent labelling are rich in information, have high throughput and produce more robust physical maps than traditional agarose gel-based methods.
- The use of high-throughput capillary electrophoresis machines and fluorescent fingerprinting methods makes physical map construction fast, efficient and largely automated.
- The proper use of statistics is required to produce high-quality physical maps, taking into account genome size and sequence complexity.
- There are methods to properly evaluate the quality and coverage of physical maps. The presence of undetected mis-assembled contigs can represent a serious problem if such methods are not applied.
- There are no non-model animal and plant species for which the same number and types of genomic resource are available as there are for humans, mice, rats, *Arabidopsis thaliana* and rice. This poses new challenges for the construction of whole-genome physical maps that require the adoption of refined clone-fingerprinting technologies.

Online links

Further information

FPC
http://www.genome.arizona.edu/software/fpc/

GenoProfiler
http://wheat.pw.usda.gov/PhysicalMapping/tools/genoprofiler/genoprofiler.html

iCE
http://www.bcgsc.ca/bioinfo/ice