

Evaluating physical maps by clone location comparisons.

Jeffry Shultz*, Khalid Meksem, David Lightfoot.

Center for Excellence in Soybean Research, Teaching and Outreach, Genomics Core-Facility, Southern Illinois University at Carbondale, Carbondale, IL 62901.

* Author for correspondence

Department of Plant, Soil and General Agriculture

Southern Illinois University

Carbondale, IL 62901-4415

Tel: 618-453-1797 Fax: 618-453-7457

E-mail: ga4082@siu.edu, jshultz@siu.edu

Abstract

Pairs of large insert clones located in proximity within 384 well storage plates should join contiguous overlapping sets of clones (hereafter contigs) at a predictable frequency. The frequency of clone pairs joining single contigs should be related to distance apart in the plate. Biased frequencies will be caused by errors such as clone contamination, clone duplication, clone insert size variation and local variations in genome representation. The aim of this study was to detect potential clone contamination errors by biases in the frequencies at which clone pairs that were within the plate, adjacent rows and columns or neighboring wells joined the same contigs. Simple statistical analyses of three biases in the frequency of pairs of clones coalescing into the same contig were used to identify potential contaminants. Manual editing determined whether the potential contaminants caused gaps or branches within contigs. BAC and YAC contigs of four plants and two animals were examined including *Glycine max* (soybean), *Arabidopsis thaliana* (thrale cress), *Oryza sativa* (rice), *Zea mays* (maize), *Sorghum bicolor* (sorghum) *Mus musculus* (mouse) and *Homo sapiens* (human). Clone pairs that formed contigs significantly ($0.05 < P < 0.0001$) exceeded expectation with plate pair clones (up to 5 fold, mouse) neighboring clones (up to 3 fold, mouse) and adjacent clones (up to 4 fold, soybean). After manually examining clone pairs contigs that contained fewer errors were built.

Keywords: physical map, clone-contamination, contig-contamination, contig-merge

Introduction

The large genome size of many eukaryotic organisms (0.1 to ~50 Tbp) present problems of scale for genomic methods. Genome encompassing physical maps derived from large insert libraries have been or are being developed for eukaryotic organisms of compelling scientific and phylogenetic value [1-4], or of economic importance [5-10]. The development of physical maps for large genomes has facilitated or assisted with the finishing of whole genome sequences for many higher eukaryotes [11-14]. Physical map builds proceed by iterative cycles of building contiguous overlapping sets of BAC clones (hereafter contigs) at high stringency; merging contigs at a lower frequency; and manual editing to identify and remove spurious BACs within contigs [10,15]. Therefore, physical maps represent large sets of data with clone overlap probabilities of variable robustness that are time consuming verify manually. Algorithms to automatically edit contigs within physical maps have been developed [16,17]. However, map-as you go sequencing is often the most efficient method for whole genome sequencing using nascent physical maps [11]. As genomic fingerprint data sets grow in size; increase genome coverage or represent larger genomes the importance of automated editing increases [3,5,10]. Low cost high-through-put builds of physical maps for large genomes will require techniques to identify potential problems as a precursor to tiling path DNA sequencing or the map-based isolation of economically important genes [18].

A contaminated well contains more than one unique clone. Contaminated wells may be a significant source of error in physical map builds [5,19]. Fingerprints of BACs located in close physical proximity to each other within 384 well storage plates should join contigs at a predictable frequency determined by their distance apart. Biases in this frequency will include several sources of error that have different consequences. Cross-contamination is the most damaging form of error for contig builds. After clone storage the cross-contamination of wells during replicating cannot be completely avoided [19]. During DNA preparations neighboring clones may be mixed. During gel loading prior to electrophoresis wells may be mixed. Other events with lesser effects on contig quality are expected. Cell division before plating for colony formation occurs at a predictable frequency that increases as incubation time after transformation increases [20]. Repeat colony picking by robot or human error is common where colony morphology is unusual [21]. These errors are likely to result in duplicate clones within the same plate (Fig. 1). During electrophoresis or editing lanes may be miss-identified. These errors are likely to result in duplicate clones that are immediate (Fig. 1; blue wells) or adjacent neighbors (Fig. 1; red wells) within the same plate. Contaminated wells pre- or post- storage will cause either gaps or branches within contigs whereas duplicate clones will not. Although contaminated BAC pairs should be easy to eliminate in small data sets in large data set their identification becomes an onerous task because other errors mimic contamination.

Biased frequencies will be caused by errors such as clone duplication, clone insert size variation and local variations in genome representation or duplication [15,19,22,23]. Pairs of clones that are duplicated during physical map development are relatively easy to

edit as they absolutely overlap and do not affect contig quality greatly. Large insert clones tend to produce many bands or hybridizations can join many contigs at high stringency and present a problem for editors identifying the correct contig for assignment. Regions of the genome that are over-represented, duplicated or are cyclically over- and under –represented by clones across a short physical distance are also likely to bias clone pair frequencies. In each case these biases should be proportional to distance, unlike contaminated clones.

The three types of clone pair biases may be diagnostic of the source of problems. The first source of pair generation is general handling and use of the library including picking, allowing the transformed cells to replicate before spreading, and miscellaneous other sources results in duplication of a well within a plate (Fig. 1). These pairs are most frequently plate pairs. The second type of pair generation is contamination between adjacent wells (Fig. 1; red wells) that results from the fingerprinting procedure that requires the replication of the stored BACs from low-media, high density storage (384 well plates) to a high media, low density growth platform (four, 2 ml-96 well plates for each 384 well storage plate). The transfer gives each BAC new “ row or column neighbors” (Fig. 1; red wells). The third type of pair generation is contamination, between adjacent storage plate wells during 384 well plate library storage and manipulation. These pairs are most frequently formed between immediate neighbors (Fig.1; blue wells).

In this study we report a method for the detection of potential errors in nascent and published physical map builds. It is important to note that a contaminated pair of wells or clones is not confirmed by the following analyses. This analysis only reports on the frequency of within plate linkages in comparison to their likelihood.

Materials and Methods

A. Libraries, Maps and Contigs Used

Publicly available large insert library derived contig databases were analyzed: *Glycine max* (soybean)[24], *Arabidopsis thaliana* (thrale cress)[27], *Oryza sativa* (rice) [6,10], *Zea mays* (maize)[9], *Sorghum bicolor* (sorghum)[25] *Mus musculus* (mouse)[3] and *Homo sapiens* (human)[5]. Each library database is unique in site of construction and/or method of physical map development.

The soybean library [24] database consisted of 78,001 BAC fingerprints encompassing 11 haploid genome equivalents. Fingerprints derived from restriction digestion, end-labeling and acrylamide gel electrophoresis (download from http://www.siu.edu/~pbgc/contig/soy_fpc_data/). There were 56,417 of these clones represented in 5,597 contigs (view at <http://hbz.tamu.edu/- Physical Mapping/Soy Map>).

The rice library [10] analyzed consisted of 92,160 BAC clones encompassing 25 haploid genome equivalents. Exactly 73,728 clones from this library have been fingerprinted using restriction digestion and agarose gel electrophoresis. There were 62,509 clones in 438 contigs.

The mouse library [3] analyzed consisted of a working copy of 24,000 YAC clones encompassing 13 haploid mouse genome equivalents. Hybridization was used to create contigs, thus allowing each clone to be listed in more than one contig. After single

groups were removed, 57,971 comparisons in 9,371 contigs (6.186 per group) remained. Comparisons were made within contigs, eliminating between contig duplication of YACs.

In addition to the three libraries above, publicly available contig data was sampled. Contigs from Arabidopsis [27], rice [6], maize [9], sorghum [25] and human [5] physical map builds were examined.

B. Data preparation and analysis

The data from each analyzed library was in different formats. Soybean was derived from HTML source code, mouse was downloaded in table format and rice was available in .FPC format. All data types were converted to and between standard spreadsheet formats during analysis. A Dell Dimension XPS R450 was used in all analyses. EZMacros ver 5.0 was used for repetitive user input (www.americansys.com).

Extraneous information was removed from the initial data and a list of contigs with associated clones is created. As necessary (see below) clones were converted to either a complete numeric form; i.e. B002J12 becomes 10021012 ($1^{\text{BamHI}}002^{\text{plate2}}10^{\text{Row}}$ J12^{Column} 12) or a simple plate form 1002 ($1^{\text{BamHI}}002^{\text{plate2}}$) The resulting list was ordered by contig ready for manual comparisons (Fig. 2).

Three types of clone pair comparisons were made: neighboring wells; adjacent wells; and adjacent plates (Fig. 1). In neighboring well comparisons, a complete numeric

form conversion is required. The comparison allows the interrogator to compare one well with another well relative to the first within all plates listed. The analysis is valuable for genomes that have already undergone extensive editing. Adjacent well analysis also requires the full numeric form of the clone ID, but can be made more general – for instance, all clones within 5 rows can be identified using this comparison. The most useful initial comparison is plate to plate that creates a list that encompasses all clones that are in the same contig that come from the same plate. The user can rapidly list clones to be checked during manual editing.

C. Error rate prediction

Assuming no biases the probability of two clones randomly joining a common contig (P_{CT}) may be predicted using the following formulae:

$$\text{Well to Well Contamination} \quad \frac{1^*}{\text{Clones in 1X coverage}} \quad \times \quad CC^{**} = \quad P_{CTW}$$

$$\text{Adjacent Well Contamination} \quad \frac{16^*}{\text{Clones in 1X coverage}} \quad \times \quad CC = \quad P_{CTA}$$

$$\text{Within Plate contamination} \quad \frac{384^*}{\text{Clones in 1X coverage}} \quad \times \quad CC = \quad P_{CTP}$$

* Number of wells containing possible match.

**Calculation of Contig Constant (CC).

Average clones per contig = clones in contigs/contigs.

Average clones per contig/genome coverage.

D. Statistical Analysis.

In the simplest form the Chi squared statistic with one degree of freedom can be used to examine each type of contamination independently. A more sophisticated analysis uses ANOVA (SAS Institute, Cary, NC) to examine the bias of each of the three metrics simultaneously. Using ANOVA the independence of the biases can be examined.

E. Manual Editing.

The data from 78,001 soybean BAC clones was modified by removing the larger of the pair of contaminated clones. This dataset was then used to perform a series of contig builds and compared to the original dataset.

Results

A. Soybean (PAGE fingerprinted)

Using the 56,417 BACs and BIBACs that remained in 5,597 contigs after eliminating singletons from 83,026 clones fingerprinted expected rates of coincidence were calculated. Clones in 1 fold genome coverage was assumed to be 8,000. The average number of clones per contig was 10.08 (56,417/5,597). The average number of clones per contig per genome represented was 1.44 (10.08/7). Therefore P_{CTW} was 0.00018, P_{CTA} was 0.0029 and P_{CTP} was 0.0691. There were significantly more plate pairs and adjacent pairs than expected (Table 1; Figure 1). When combined neighboring pairs did not differ as significantly from expected as adjacent pairs. However, some neighboring pair positions were significantly under-represented and others over-represented so ANOVA was strongly significant.

B. Mouse (YAC STS hybridized)

The mouse data is hybridization based, therefore clones can be represented more than once. In mouse, the number of comparisons within contigs is measured. After single groups were removed, 57,971 comparisons in 9,371 groups (6.186 per group) remained. These clones are in 96 well format, which reduces the chance of within plate contamination by 75%. Each YAC is approx. 800 kbp, with 1,840 YACs per genome coverage. The current data is based on 24,000 clones, representing 13x coverage.

Comparisons were made within contigs, eliminating between contig duplication of YACs (<http://www.resgen.com/products/WIMITMYAC.php3>). Expected rates of coincidence was calculated. Clones in 1 fold coverage was 1,840. The average number of clones per contig was 6.186 (57,971/9,371). The average number of clones per contig per genome represented was 0.476 (6.186/13). P_{CTW} was 0.0005, P_{CTA} was 0.0041 and P_{CTP} was 0.0248. There were significantly more plate pairs and neighboring pairs than expected (Table 1). Surprisingly for a library stored in 96 well plates, adjacent pairs were under-represented so ANOVA was not strongly significant.

C. Rice (agarose fingerprint)

Exactly 65,287 fingerprinted BAC clones were analyzed, with 2,778 singletons and 62,509 clones in contigs that were composed from BACs that represented about 20 haploid genomes. Due to the low number of contigs, once data was converted, comparison values were set to measure the clones directly adjacent to each other or in wells adjacent during fingerprinting (P_{CTA}). Expected rates of coincidence was calculated (below). Clones in 1 fold coverage was 3,264. Average clones per contig was 143 (62,509/438). Average clones per contig per genome coverage was 7.15 (143/20). P_{CTW} was 0.002, P_{CTA} was 0.036 and P_{CTP} was 0.841. There were not significantly more neighboring pairs than expected (Table 1). However, plate pairs were under-represented and adjacent pairs were not significantly different so ANOVA was not significant.

However, in order to generate an accurate assessment of the rice fingerprint data, it was necessary to check all clones using well to well comparisons, then adding the comparisons together to indicate only the wells that are adjacent during storage or fingerprinting. The P_{CTA} of the rice fingerprints is an expected value of 2,190 clones randomly assigned to the same contig. Data comparisons identified 2,197 possible contaminants, indicating that the rice fingerprint database was relatively accurate and well edited prior to this analysis.

D. Public Databases

In addition to the three libraries above, publicly available databases were examined by application of this test. A sample contig from each library is listed, along with clones that would be indicated by a stringent comparison test (Table 2).

Discussion

Factors that affect physical map quality analysis are the insert size and variability; genome size, coverage depth and local representation; numbers of contigs and super-tigs; and the tolerance and cutoff used for builds [15,19,23]. Only the average insert size and genome size are taken into account in the formulae underlying this method. Therefore, contaminated clone pairs identified by this method should be considered indicative not conclusive.

A positive feature of the physical map quality analysis method reported here was that the fingerprints, band or hybridization scores themselves are not compared [16]. Since only within contig comparisons can be made, the fewer the contigs and the greater the number of clones the more likely that there will clone pairs. If the probability of two clones from the same plate being placed on the same contig is high (eg. few contigs and many clones; rice) [10], more stringent within plate comparison must be made. If the probability is low (eg. many contigs and few clones; soybean)[26], plate to plate comparisons gave acceptable results. Assuming this method was incorporated to the iterative process of contig building, tolerance and cutoff can be accounted for by following a high cut-off and comparison of clones, with a lowering cutoff, and repeating until all contaminated wells are identified.

Were problematic clones identified?

In every library tested, the analysis identified contigs with improbable clone linkages. Some libraries had more potential contaminant pairs than others. The high singleton rate in soybean was probably due to the very high cutoff score ($1e^{-20+}$) used to generate the contigs. The number of singletons and contigs may drop once the contaminated wells are identified and removed, and a lower cutoff score used.

Which fingerprints are good, and which are not?

There are three profiles commonly observed when manually comparing fingerprints; pairs of clones are identical, one is larger than the other in the pair or both clones in the pair are dissimilar [15,19,22,23]. Identical fingerprints most likely result

from miss loading during gel electrophoresis. The danger of this profile is that the user does not know which clone is represented. If further analysis is performed on one of these clones based on its fingerprint, the researcher has a 50% chance of using a completely unrelated clone. In the case of identical fingerprints, the safest procedure is to clearly “label” both clones, so that the correct clone can be identified prior to research. When one of the two fingerprints contains more bands than its neighbor, it is likely to be the contaminated clone and should be removed from analysis. Since minimum tiling paths are usually generated from the longest clones possible (McPherson et al., 2000), there is an increased risk of choosing a chimerical clone. If the fingerprint profiles are dissimilar, the clone comparison can be ignored. For hybridization, the same three tests are applied, although probe positives are used instead of bands.

Within plate, and well-to-well contamination are serious detractors to the accuracy of contig data. With the reduction or elimination of these contaminations, the required stringency for contig creation is reduced along with the number of clones required for the creation of an accurate physical map. Reductions in the number of clones required for a physical map relate directly to the cost of each project. Identifying and removing just 5-10% of the clones from a library would greatly reduce forward and reverse end sequencing costs, along with commensurate time to analyze duplicate/erroneous data.

Manual editing improves the map in two ways, identifying potential joins between contigs, performing merges and increasing contiguity (Fig. 3). Secondly, it identifies

potential chimerical contigs by revealing incorrectly overlapped fingerprint data or by high-lighting conflicting marker data [10]. The major problem with human manual editing is the near impossibility of identifying possibly contaminated wells. As contigs become larger, the task of the editor becomes more difficult. The method we report provides an automated method for identifying clones to be examined manually

Conclusion

In a series of tests using our own physical map builds [28-31] and those of others [3,10,26] well contamination appeared to be present. Contamination seems to be an effect of processing tens or hundreds of thousands of highly transferable organisms, the bacterial strains. The method presented allows map editors and map users to rapidly evaluate the quality of physical maps and to identify clones in contigs that should be re-examined before merging to form super-tigs.

References

1. Waterston R, Sulston J. 1995. The genome of *Caenorhabditis elegans*. Proc Natl Acad Sci U S A. 92:10836-10840
2. Marra M, Kucaba T, Sekhon M, Hillier L, Martienssen R, Chinwalla A, Crockett J, Fedele J, Grover H, Gund C, McCombie WR, McDonald K, McPherson J, Mudd N, Parnell L, Schein J, Seim R, Shelby P, Waterston R, Wilson R. 1999. zA map for sequence analysis of the *Arabidopsis thaliana* genome. Nat Genet. 22:265-270.
3. Lindblad-Toh K, Lander ES, McPherson JD, Waterston RH, Rodgers J, Birney E. 2001. Progress in sequencing the mouse genome. Genetics 31:137-141.
4. Hoskins RA, Nelson CR, Berman BP, Lavery TR, George RA, Ciesiolka L, Naeemuddin M, Arenson AD, Durbin J, David RG, Tabor PE, Bailey MR, DeShazo DR, Catanese J, Mammoser A, Osoegawa K, de Jong PJ, Celniker SE, Gibbs RA, Rubin GM, Scherer SE. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. Science.287:2271-2274
5. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, Fulton R, Kucaba TA, Wagner-McPherson C, Barbazuk WB, Gregory SG, Humphray SJ, French L, Evans RS, Bethel G, Whittaker A, Holden JL, McCann OT, Dunham A, Soderlund C, Scott CE, Bentley DR, Schuler G, Chen HC, Jang W, Green ED, Idol JR, Maduro VV, Montgomery KT, Lee E, Miller A, Emerling S, Kucherlapati, Gibbs R, Scherer S, Gorrell JH, Sodergren E, Clerc-Blankenburg K, Tabor P, Naylor S, Garcia D, de Jong PJ, Catanese JJ, Nowak N, Osoegawa K, Qin S, Rowen L, Madan A, Dors M, Hood L, Trask B, Friedman C, Massa H, Cheung VG, Kirsch IR, Reid T, Yonescu R, Weissenbach J, Bruls T, Heilig

- R, Branscomb E, Olsen A, Doggett N, Cheng JF, Hawkins T, Myers RM, Shang J, Ramirez L, Schmutz J, Velasquez O, Dixon K, Stone NE, Cox DR, Haussler D, Kent WJ, Furey T, Rogic S, Kennedy S, Jones S, Rosenthal A, Wen G, Schilhabel M, Gloeckner G, Nyakatura G, Siebert R, Schlegelberger B, Korenberg J, Chen XN, Fujiyama A, Hattori M, Toyoda A, Yada T, Park HS, Sakaki Y, Shimizu N, Asakawa S, Kawasaki K, Sasaki T, Shintani A, Shimizu A, Shibuya K, Kudoh J, Minoshima S, Ramser J, Seranski P, Hoff C, Poustka A, Reinhardt R, Lehrach H. 2001. A physical map of the human genome. *Nature* 409:934-941.
6. Tao Q-Z, Chang Y-L, Wang J, Chen H, Schuering C, Islam-Faridi MN, Wang B, Stelly DM and Zhang H-B. 2001. Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. *Genetics*. 158:1711-1724.
7. Zobrist K, Meksem K, Wu C, Tao Q, Zhang H, and Lightfoot DA (2000) Integrated physical mapping of the soybean genome: A tool for rapid identification of economically important genes. *Soybean Genetics Newsletter* 27: 10-15.
8. Zobrist K, 2000a. Integrative physical mapping of the soybean genome. MS thesis SIUC Carbondale pp213.
9. Coe E, Cone K, McMullen M, Chen SS, Davis G, Gardiner J, Liscum E, Polacco M, Paterson A, Sanchez-Villeda H, Soderlund C, Wing R. 2001. Access to the maize genome: an integrated physical and genetic map. *Plant Physiol*. 128:9-12.
10. Chen M., G Presting, W Barbazuk, JL Goicoechea, B Blackmon, G Fang, H Kim, D Frisch, Y Yu, S Sun, S Higingbottom, J Phimphilai, D Phimphilai, S Thurmond, B Gaudette, P Li, J Liu, J Hatfield, D Main, K Farrar, C Henderson, L Barnett, R Costa, B

Williams, S Walser, M Atkins, C Hall, MA Budiman, JP Tomkins, M Luo, I Bancroft, J Salse, F Regad, T Mohapatra, NK Singh, AK Tyagi, C Soderlund, RA Dean, and RA Wing. 2002. [An Integrated Physical and Genetic Map of the Rice Genome](#). *The Plant Cell* 14:537-545.

11. TAGI 'The Arabidopsis Genome Initiative' 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815.
12. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays, AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko, P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G,

- Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195.
13. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyas E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer

- T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 420:520-562.
14. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit AF, Sollewijn Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef, F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ,

- Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. 2002. Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. *Science* 297:1301-1310.
15. Soderlund C., Humphrey S., Dunhum A., and French, L. (2000). Contigs built with fingerprints, markers and FPC V4.7. *Genome Research* 10:1772-1787.
16. Thayer EC, Olson MV, Karp RM. 1999. Error checking and graphical representation of multiple-complete-digest (MCD) restriction-fragment maps. *Genome Research* 9:79-90.
17. States DJ, Nowotny V, Blackwell TW. 2001. Probabilistic approaches to the use of higher order clone relationships in physical map assembly. *Bioinformatics*. 17: Suppl 1:S262-269.
18. Zhang H-B and Wu CC. 2002. BACs as tools for genome sequencing. *Plant Physiology and Biochemistry* 39:195-209
19. Mott R, Grigoriev A, Lehrach H. 1998. Long range physical mapping construction and the integration of genetic and physical maps. In "ICRF Handbook of genome analysis" Eds NK Spurr B, Young and SP Bryant. Blackwell Science Ltd., Oxford, UK. pp421-439.
20. Boyl A. 1999. Strategies for Large insert cloning and analysis. In "Current Protocols in Human Genetics" (ed. Ann Boyl), Wiley, New York.
21. Uber DC, Jaklevic JM, Theil EH, Lishanskaya A, McNeely MR. 1991. Application of robotics and image processing to automated colony picking and arraying. *Biotechniques*. 11:642-647

22. Sulston J, Mallett F, Durbin R, Horsnell T. 1989. Image analysis of restriction enzyme fingerprint autoradiograms. *Comput Appl Biosci.* 5:101-106.
23. Sulston, J. Sulston J, Mallett F, Staden R, Durbin R, Horsnell T, Coulson A. 1988. Software for genome mapping by fingerprinting techniques. *Comput Appl Biosci.* 4:125-132.
24. Meksem K, Ruben E, Zobrist K, Zhang H-B, Lightfoot DA. 2000. Two large insert libraries for soybean: Applications in cyst nematode resistance and genome wide physical mapping. *Theor Appl Genet* 101: 747-755.
25. Klein PE, Klein RR, Cartinhour SW, Ulanich PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, Mullet JE. 2000. A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Res* 10:789-807.
26. Wu C, Nimmakayala P, Santos P, Springman R, Meksem K, Lightfoot DA, Zhang HB. 2003. A BAC and BIBAC based physical and genetic map for the soybean genome. *Research Genetics* (in review). www.hbz.tamu.edu
27. Chang Y-L, Tao Q, Schuering C, Meksem K and Zhang H-B. 2001. An integrated map of *Arabidopsis thaliana* for functional analysis of genome sequence. *Genetics* 159: 1231-1241.
28. Shultz J, Meksem K, Lightfoot DA. 2003. A comparison of fingerprint and hybridization physical maps of the *Ustilago maydis* strain 512 genome. *Molec. Gen. Genomics* (in press).

29. Shultz J, Wu C, Zobrist K., LaMontague C., Meksem K., Zhang H-B., and D.A. Lightfoot. 2003a. A revised physical and genetic map for the soybean genome linkage group G. *Molecular and General Genomics* 266:xxx-xxx (in review).
30. Shultz J., Wu C, FA Santos, P Nimmakayala, R Springman LaMontague C., Zobrist K., Meksem K., Zhang H-B., and D.A. Lightfoot. 2001. A users guide for the physical map for the soybean genome. *Soybean Genetics Newsletter* 28:5-10.
31. Lightfoot DA, Meksem K., Zhang H-B. 2003. An integrated physical and genetic map for the soybean genome. In “Summaries of legume genomics projects from around the globe. Community resources for crops and models.” Eds K Van den Bosch and G Stacey. *Plant Physiology* 131:840-865

ACKNOWLEDGEMENTS

This research was funded in part by a grant from the NSF 9872635, ISPOB 02-127-03 and USB 2228. The continued support of SIUC, College of Agriculture and Office of the Vice Chancellor for Research to DAL and KM is highly appreciated. We thank Dr. HongBin Zhang of Texas A&M for releasing the fingerprint database to the soybean community and Dr. Chencang Wu for developing the database.

Table 2: Examples of clone pairs from five libraries identified by comparison of the probability of clone pairs joining the same contigs that were within the same plate, in adjacent rows and columns or in neighboring wells.

Genome	Source of Contig	Contig	Clones
Arabidopsis	http://hbz.tamu.edu/bacindex1.html	1013	T02H20/G22/G18 B12F20/D16
	http://www.mpimp-golm.mpg.de/mpimp-map/bacmap/chr1/chr1_1.html	ch1;contig1	6K7/F3; 22M8/G18
Rice	http://hbz.tamu.edu/bacindex1.html	400	R02E11/K19 B13A01/M01 H31D11/D12
Maize	http://www.genome.clemson.edu/	2	c0117P05/L23; c0042L02/J08
Sorghum	http://www.genome.clemson.edu/	7	c0030L20/F07
Human	www.ncbi.nlm.nih.gov	ch1	61A13/14 239E10/D12

Figure 1. An example of a section of a 384 well plate showing the positions of clone pairs expected from contamination by neighbor clones (blue) or adjacent clones (red). The well labeled D12 is the contamination source well. Panel A shows the coordinate position of wells. Panel B shows the number of clone pairs detected at each position when examining the mouse YAC contigs. Panel C shows the number of clone pairs detected at each position when examining the soybean BAC contigs. Panel D shows the number of clone pairs detected at each position when examining the rice BAC contigs.

A, Example

B10	B11	B12	B13	B14
C10	C11	C12	C13	C14
D10	D11	D12	D13	D14
E10	E11	E12	E13	E14
F10	F11	F12	F13	F14

B, Mouse

14	19	54	47	31
26	67	177	94	53
79	243	D12	243	79
53	94	177	67	26
31	47	54	19	14

C, Soybean

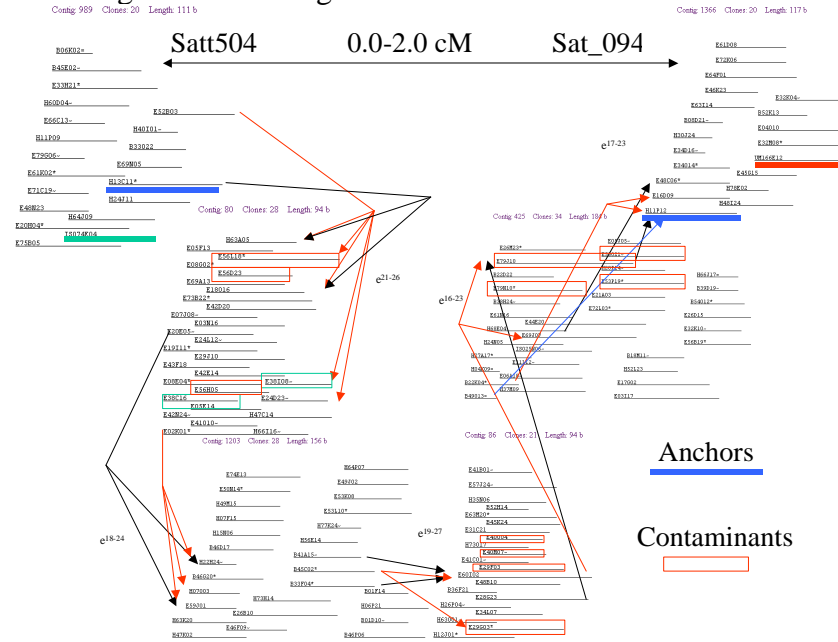
11	8	259	12	25
16	17	75	25	12
74	96	D12	97	74
13	25	74	18	15
25	13	260	9	12

D, Rice

95	90	119	96	99
112	114	148	154	114
157	219	D12	219	158
114	154	148	114	112
99	97	118	90	95

Figure 3: Examples of contigs built before and after contaminated clone removal. Panel A: Five contigs formed a set that would encompass the 0-2.0cM from genetic markers BARC-Satt504 to Sat_094. Anchored BAC clones are shown by blue line for “Forrest” BACs or green lines for ‘Williams’ BACs or red lines for “Faribault” BACs. The contigs contained 10 potential contaminants indicated by red boxes. Red and black arrows show the merges predicted when contaminant pairs are edited. There were 270 bands between markers (after editing) so the interval is expected to represent about 710 kbp

A. Contigs before editing.



B. Edited Contig

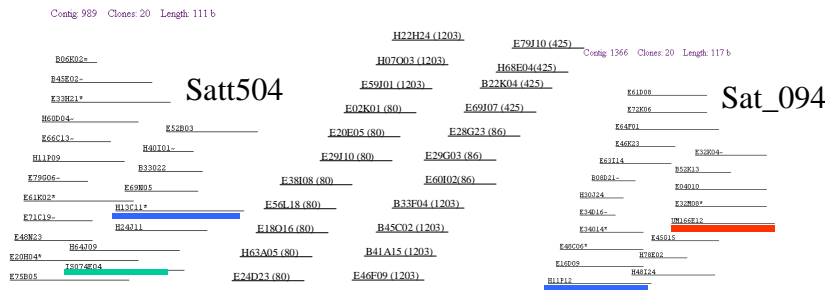


Table 1: Comparison of the probability of clone pairs joining the same contigs that were within the same plate or in adjacent rows and columns or in neighboring wells in three libraries.

Genome	Contigs/ Groups	Clones	Plate pairs	E#	P*	Combined pairs	E	P	Neighbor pairs	E	P	Adjacent pairs	E	P	Anova+
Soybean	5597	56417	8887	3898	0.0001	1167	960	0.0001	427	480	0.011	740	480	0.0001	0.0001
Mouse	9371	57971*	7094	1437	0.0001	1518	1312	0.0001	1162	656	0.0001	356*	656	0.0001	0.027
Rice	438	71575	30198	32581	0.0001	2197	2190	0.83	1270	1095	0.0001	927	1095	0.0001	0.058

- E is the expected number of pairs assuming no biases in clone size, genome representation or data content.

* - P is the probability of a significant difference by Chi squared.

+ - Anova P presents the probability that variance is unequal among the adjacent and neighbor pairs.

@ - note that Well/Well Pairs Fingerprinting mouse data may serve as a negative control for 384 to 96 well re-arraying errors because this library was not stored in 384 well plates.