Integrated physical mapping of the soybean genome: A tool for rapid identification of economically important genes.

Jeffry Shultz 1,
Chad LaMontagne 1
Kimbely Zobrist 1,
Khalid Meksem 1,
Chengcang Wu2,
Quanzhou Tao2,
Padmavathi Nimmakayala 2,
Filip Santos 2,
Hongbin Zhang 2,
David A. Lightfoot 1

1 Department of Plant Soil and General Agriculture, Southern Illinois University, Carbondale, IL 62901-4415. E-mail: ga4082@siu.edu http://www.coalab.siu.edu/genome/
2 Department of Soil and Crop Sciences and Crop Biotechnology Center, Texas A&M University, College Station, Texas 77843-2123.

The complex genomes of crops like soybean are rich in highly repeated sequences (55%) that block chromosome walking from linked DNA markers to target genes (Danesh et al., 1998) and confound shotgun approaches to complete genome sequencing (Boysen et al., 1999). In addition, much of the rest of the soybean genome is moderately repeated (25%) or unique but duplicated in two homeologous loci (20%) (Grant et al., 2000). Success of map-based cloning in crop plants depends on the target gene being localized to such a short genetic interval (0.01-0.1 cM) that the markers and target gene are separated from regions rich in repeated sequences (Meksem et al., 2000; 2001). Fine mapping to this degree can be accomplished 10-50 fold more efficiently with a physical map that encompasses the whole genome than by marker saturation and chromosome walking or landing (Zhang and Wing, 1997; Meksem et al., 2001a).

Three types of approach can be used singly or in combination for physical map construction. Physical maps may be constructed by the sequence tagged site (STS) based hybridization to generate sets of contiguous overlapping clones (hereafter contigs; Chumakov et al. 1995; Mozo et al., 1999; Ross et al., 1999). STS based physical maps of metazoans are often refined from radiation hybrid map construction, that preliminarily produce low resolution gene maps (Schuler et al., 1996; Deloukas et al., 1998) but form the basis and provide tools for high density sequence

ready physical map development (McPherson et al., 2001). However, hybridization may provide low resolution and non-robust physical maps when the genome being studied is polyploid or diploidized tetraploid as in the case of soybean. Although preliminary work is promising the ability of overgo probes to be homeolog specific is not yet clearly proven (Han et al., 2000).

Physical maps may be developed from random large insert clones by fingerprinting with restriction enzyme digestion and fragment size estimation by gel electrophoresis (Coulson et al. 1986; Marra 1997; 1999). Fragment size estimation may also be determined by optical mapping (Lai et al., 1999) at least for small genomes. Early researchers used YACs, cosmids and lambda clones for fingerprinting small (Hodgkin et al. 1995) and large eukaryotic genomes (Heding et al. 1992). BAC libraries have facilitated the method by simplifying large insert library development (Marra et al., 1997; 1999; Zhang and Wing 1997; Hong 1997; Gregory et al., 1997; Ding et al., 1999; Klein et al., 2000). Our group (Zhang and Wing 1997; Tao and Zhang 1998; Zobrist et al., 2000; 200a; Tao et al. 2001) and others (Marra et al., 1997; 1999; Klein et al., 2000; Mao et al., 2000; Tingey et al., 2001; McPherson et al., 2001) have demonstrated the feasibility of genome physical map construction from BAC fingerprints for large genomes. Among plants genome encompassing physical maps have been developed for Arabidopsis, rice, sorghum, soybean and corn, in ascending genome size order. Local physical maps around specific loci have been constructed or evaluated by fingerprinting in many large genome crops (Meksem et al., 2000; 2001; Pattochi et al., 2000; Wu et al., 2000) enabling comparison of the method with hybridization and PCR methods for contig building. Using these fingerprints we can distinguish among regions where sequences are repeated in rice (Tao et al., 2001) and soybean (Meksem et al., 1998; 2000; Zobrist et al., 2000; 2000a). Fingerprinting has the advantage of requiring relatively little labor and therefore has low cost load and is useful as an independent test of contigs generated with other methods.

Physical maps can be developed by sequencing the two ends of enough BACs to cover 15 x haploid genomes (Venter et al., 1996; Boysen et al., 1997; Budiman et al., 2000; Marek et al., 2000). A seed BAC is then selected, sequenced in toto, and compared with all BAC end sequences in the database. Putative overlapping

BACs are fingerprinted and used to construct the regional BAC contig physical map. Therefore, the sequence-ready physical map development and genome sequencing are conducted simultaneously by the Sequence-Tag Connector strategy (Boysen et al., 1997; 1997a).

To develop the proposed integrated physical map for soybean, we have developed three large-insert soybean DNA libraries from 'Forrest' (Meksem et al., 2000). We have generated restriction enzyme fragment fingerprints for 90,000 of 112,600 clones to date. Each clone was tested for the content of known genetic markers (Cregan et al., 1999; 1999a; Kassem 2001) and some ESTs and SNPs. We have shown that fingerprints of BACs from other cultivars can be integrated with this data set. The integrated genetic and physical map we are developing will inexpensively and rapidly enable identification of a number of genes underlying QTL of economic importance. Here we report progress toward genetic and physical map integration using microsatellite markers and the web-site at which data can be retrieved, clones submitted for fingerprinting and clones requested.

Materials and Methods

Pools and super-pools of the two BAC libraries: Pools and super-pools of mixed DNA were constructed as Zobrist et al., 2000)

Microsattelite markers screening of the two BAC libraries: Two methods were used as desribed by Zobrist et al., (2000).

Fingerprinting of BACs for physical mapping: Individual BAC DNAs were fingerprinted using the *Hind*III and *Hae*III digestion kit developed for large-insert bacterial clones (the Fpase kit; H.-B.Zhang and Q.Tao, Invention no. TAMUS1228) by methods described previously (Zobrist et al., 2000; Klein et al 2001)

Hybridization with RGAs and overgo probes.

Hybridization probes were prepared from RGAs by amplification of soybean DNA with the gene class specific primers described by Kanazin et al., (1996) and labeled with 32PdATP labeling buffer at 37 c for 30 min (Ross et al., 1999).Cold dATP and dCTP (200 uM) was added and incubation continued for 30 min. Unincorporated nucleotides were removed by Sephadex G50 spin columns, probes

denatured at 94 C for 4 min and added to the hybridization medium. Hybridization was performed at 58 C for 18 h in hybrization solution. The 38,500 clones of the *Eco*RI BAC library were spotted onto nylon filters at 13,824 (36 x 384) clones on seven filters of 24 x 24 cm. After hybridization filters were washed and used to expose film at 70 C overnight.

Overgo primers were selected following a BLAST search to exclude repeat sequences. Oligomers were purchased from Research Genetics and labeled with 32PdATP and 32PdCTP in labeling buffer at 37 c for 30 min (Ross et al., 1999).Cold dATP and dCTP (200 uM) was added and incubation continued for 30 min. Background overgo for E.coli K12 detection were synthesized after Han et al., (2000). Unincorporated nucleotides were removed by Sephadex G50 spin columns, probes denatured at 94 C for 4 min and added to the hybridization medium. Hybridization was performed at 58 C for 18 h in hybrization solution. Filters were washed and used to expose film at 70 C overnight.

Results and Discussion

The soybean bacterial artificial chromosomes (BACs) were generated from cultivar 'Forrest'with *Eco*RI, and the large insert plasmid clones (LICs) with *Hind*II and *Bam*HI. We identified BAC clones that contained microsatellite markers by PCR on pooled DNA samples to generate local integrated genetic and physical maps Figure 1 and 2). Separately we identified sequence tagged sites in groups of 384 by multiplexed hybridization methods. We examined the robustness of these local maps by fingerprinting of DNA from LICs and BACs after digestion and end-labeling with an enzyme kit that contains *Hin*dIII and *Hae*III. The satellite physical map was developed from the *Hin*dIII LIC library. Super-pools were screened with 238 SSR primer pairs that identified 702 positive clones, an average of 2.95 clones per primer from from LICs representing 4.5 genome equivalents. Seventy-three SSR primer pairs were used to identified 163 single positive clones from 38,400 clones and a further 385 primer pairs were used to identify single positive clones from 7,680 clones (about one genome

equivalent). Therefore there are 485 microsattellite anchor available for the soybean physical amp.

Fingerprinting the 163 clones showed that large microsatellite-anchored groups of contiguous overlapping clones (contigs) were formed at the overlap statistic of e^{-18} (about 50% of bands required to overlap). However, many of the clones with large numbers of bands (>50 bands) compared to the mean (35 bands) were not true members of the contigs and should be removed from consideration. Edited anchored contigs would separate into two or three groups at overlap statistic of e^{-23} (about 60% of bands required to overlap). At this stringency fingerprinting was also able to incorporate BAC clones from soybean cultivar Williams in to Forrest derived anchored contigs Figure 3). Integration of these markers provides more than 600 RFLP anchors for the physical map and duplicates many of the microsatellite anchors.

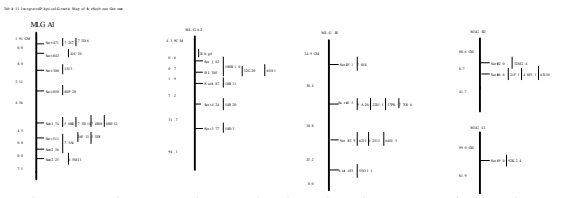
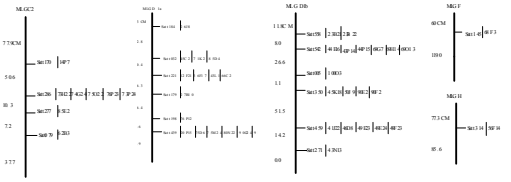


Fig. 1: The anchored physical map of Soybean



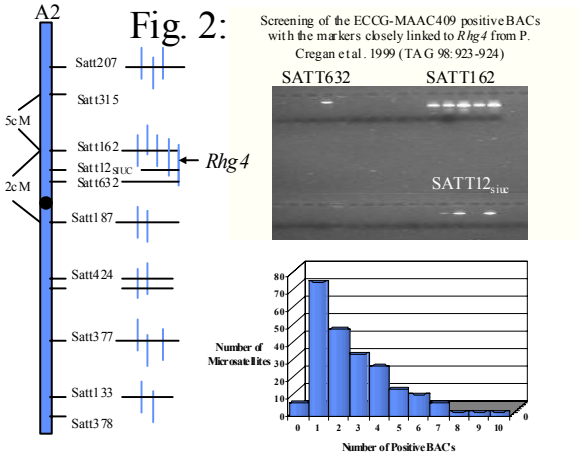
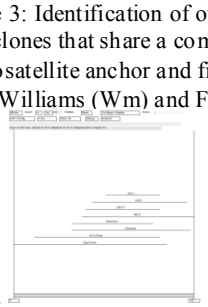
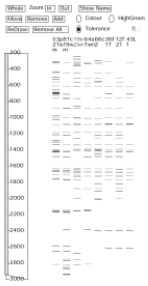


Fig. 2:

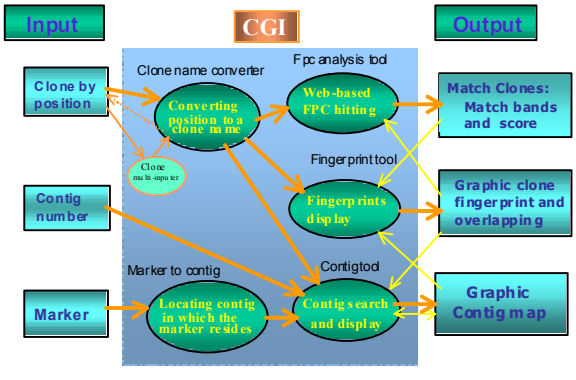
Figure 3: Identification of overlapping clones that share a common microsatellite anchor and fingerprint from Williams (Wm) and Forrest (F).





Group	MLGD1a	MLG G	MLG G	MLG J
Marker	Satt221	Satt217	Satt288	Satt183
Forrest Fpase positive	4	1	2	1
Williams Fpase positive	5	3	2	2
Anchored but Fpase negative	2	2	1	2

Figure 4: A Bioinformatic Tool Showing Input-Output Paths for Contig Access



From 50,000 fingerprinted BACs and LICs we generated 4,500 contigs that were predominantly short, containing 2-4 clones and encompassed about 500 Mbp (Figure 4) . Hybridization to the BAC library with eight classes of resistance gene analogs identified 1,860 hybridizing clones, that formed 99 contigs from 2-46 BACs each that jointly encompass 22 Mbp (wu et al., 2001). Overgo probes were effective in identifying 50% of the RGAs as unique genes on specific contigs (not shown). End sequencing of BACs and LICs identified new gene fragments in soybean that could be positioned on the genetic map. Therefore, the BACs and LICs that are anchored and located on the physical map are ready to nucleate marker saturation, genome sequencing and gene discovery methods for gene rich regions.

Updates to the physical map, the clones and the libraries are available through requests to SIUC at http://www.coalab.siu.edu/genome/ and TAMU at Physical Mapping sections of http://hbz.tamu.edu/. This work was supported by the NSF project 99872635 and the ISPOB project 98-24-198-3. We thank Dr R. Shoemaker for making available the BAC clones from the cultivar Williams.

References

Boysen C, Simon MI, Hood L.1997. Analysis of the 1.1-Mb human alpha/delta T-cell receptor locus with bacterial artificial chromosome clones. Genome Res. 7:330-338.

Boysen C, Simon MI, Hood L.1997a. Fluorescence-based sequencing directly from bacterial and P1-derived artificial chromosomes. Biotechniques. 23:978-982.

Budiman MA, Mao L, Wood TC, Wing RA. 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res. 10:129-136.

Chumakov IM, Rigault P, Gall IL . 1995. A YAC contig map of the human genome. Nature 377: 157-298.

Coulson A, Sulton J, S Brenner S, Karn J. 1986: Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA 83:7821-7825.

Cregan P.B., Shoemaker, R.C. and Specht, J.E. 1999. An Integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464-1490.

Cregan P.B. Mudge J., Fickus E.W., Denny R., Danesh R., and Young, N.D. 1999a. Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor Appl. Genet. 99:918-928.

Danesh, D., S. Penuela, J. Mudge, R.L. Denny, H. Nordstrom, J.P. Martinez, Young, N.D. 1998. A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theor. Appl. Genet. 96:196-202.

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Bentley DR. 1998. A physical map of 30,000 human genes. Science 282:744-746.

Ding Y, Johnson MD, Colayco R, Chen YJ, Melnyk J, Schmitt H, Shizuya H. 1999. Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. Genomics 56:237-246.

Grant D, Cregan P, Shoemaker RC. 2000. Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci USA.97:4168-4173.

Gregory SG, Howell GR, and Bentley DR. 1997: Genome mapping by fluorescent fingerprinting. Genome Research 7: 1162-1168.

Han CS, Sutherland RD, Jewett PB, Campbell ML, Meincke LJ, Tesmer JG, Mundt MO, Fawcett JJ, Kim UJ, Deaven LL, Doggett NA. 2000. Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization. Genome Res. 10(5):714-721.

Heding IJJP, Ivens A, Wilson J and et al 1992: The generation of ordered sets of cosmid DNA clones from human chromosome region 11p. Genomics 13:89-94.

Hong G. 1997. A rapid and accurate strategy for rice contig map construction by combination of fingerprinting and hybridization. Plant Mol Biol 35:129-133.

Hodgkin J, Plasterk RHA, Waterston RH. 1995. The nematode *Caenorhabditis elegans* and its genome. Science 270:410-414.

Kassem AM Iqbal J. Meksem K., Njiti V. and Lightfoot DA. 2001. A saturated microsatellite markers map for Essex x Forrest RILs identify three additional quantitative trait loci for yield and phytoestrogen content. Theor. Appl. Genet. 102:xxx-xxx.

Klein PE, Klein RR, Cartinhour SW, Ulanch 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.

Lai Z, Jing J, Aston C, Clarke V, Apodaca J, Dimalanta ET, Carucci DJ, Gardner MJ, Mishra B, Anantharaman TS, Paxia S, Hoffman SL, Craig Venter J, Huff EJ, Schwartz DC. 1999. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nat Genet. 23:309-313.

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA. 2000. Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res. 10:982-990.

Marek L, Mudge J, Darnielle L, Grant D, Hanson N, Paz M, Huihuang Y, Denny R, Larson K, Foster-Hartnet D, Cooper A, Danesh D, Raph T, Staggs R, Crow JA, Retzel E, Young N, Shoemaker R. 2000. Soybean genomic survey: What we are learning about the soybean genome from BACs identified by RFLP and SSR markers. Proceeding of Soy2000, 8 th Biennial Conference of the Cellular and Molecualr Biology of the Soybean. Lexington, Kentucky, PIII 04.

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

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. 1999 22:265-270

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(6822):934-941.

Meksem K., Zhang H-B., and D.A. Lightfoot. 1998. A plant transformation ready bacterial artificial chromosome library for soybean: Applications in chromosome walking and genome wide physical mapping. Soybean Genetics Newsletter 25:108-111.

Meksem K., Doubler T.W., Chang S.J.C., Chancharoenchai K., R. Suttner, P. Cregan, P. Rao Arelli, P.T. Gibson and D.A. Lightfoot. 1999. Clustering among genes underlying QTL for field resistance to Sudden Death Syndrome and cyst nematode race 3. Theor. Appl. Genet. 99:1131-1142.

Meksem K. Zhang H.B. and D.A. Lightfoot. 2000. Two transformation ready large insert clone libraries for soybean: Physical mapping of resistance to Soybean Cyst Nematode and Sudden Death Syndrome. Theor. Appl. Genet. 100:747-755.

Meksem K., Ruben E., Hyten D., Triwitayakorn K., Lightfoot DA. 2001. Conversion of AFLP bands to high-throughput DNA markers. Mol Gen Genomics 265:207-214.

Meksem K., Ruben E, Hyten D Schmidt M and Lightfoot DA. 2001a. High-throughput detection of polymorphism physically linked soybean cyst nematode resistance gene *Rhg4* using Taqman Probes. Molecular Breeding 7: 19-28.

Mozo, T., Fischer, S., Meier-Ewert, S., Lehrach, H. & Altmann, T. 1997. Use of the IGF BAC library for physical mapping of the *Arabidopsis thaliana* genome. Plant J. 16:377-384.

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S, Altmann T. 1999. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat Genet.22:271-275.

Patocchi A, Vinatzer BA, Gianfranceschi S, Zhang H-B, Sansavini S and Gessler C. 1999. Construction of a 550-kb BAC contig spanning the genomic region containing the apple scab resistance gene *Vf*. Mol. Gen. Genet. 262:884-891.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al.1996. A gene map of the human genome. Science. 274:540-546.

TAGI 'The Arabidopsis Genome Initiative' 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796-815.

Tao Q-Z, Zhang H-B. 1998. Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res. 26: 4901-4909.

Tao Q-Z, Zhang H-B. 2001. Genome wide physical mapping with large insert bacterial clones and a BAC physical map of the rice genome. Nature Genet. (In press).

Tingey S. 2001. Genetic, Physical and Diversity Maps of Maize as a Foundation for a Molecular Breeding Approach to Germplasm Improvement. Plant and Animal Genome IX.

Venter JC, Smith HO, and Hood L.1996. A new strategy for genome sequencing. Nature 381: 364-366.

Wu Y, Tulsieram L, Tao Q, Zhang H-B and Rothstein SJ. 2000. A binary vector-based large insert library for *Brassica napus* and identification of clones linked to a fertility restorer locus for *Ogura* cytoplasmic male sterility (CMS). Genome 43:102-109.

Wu C, FA Santos, P Nimmakayala, R Springman K Meksem, DA Lightfoot† and H-B Zhang 2001. The NBS-LRR gene family in the soybean genome is organized and evolves by a process of birth-and-death, supercluster formation and adaptive selection**.** Proc. Natl Acad Sci USA (submitted).

Zhang H-B, Wing RA: 1997. Physical mapping of the rice genome with BACs. Plant Mol Biol. 35:115-127.

Zhang H-B, Woo S-S, Wing RA: 1996. BAC, YAC and cosmid library construction. In Foster GD, Twell D (eds), Plant Gene Isolation: Principles and Practice, pp.75-99. John Wiley & Sons Ltd.

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.

Zobrist K, 2000a. Integrative physical mapping of the soybean genome. MS thesis SIUC Carbondale pp213.