Verification of Drosophila Sequence Assembly
using restriction digest BAC fingerprints derived from multiple enzymes

Krzywinski M, Schein J, Chiu R, Bosdet I, Mathewson C, Wye N, Barber S
Brown-John M, Chand S, Cloutier A, Masson A, Mayo M, Olson T
Jones S, Hoskins R[†], Celniker S[†], Rubin G[‡], Marra M

British Columbia Cancer Agency
Vancouver, British Columbia, Canada
Genome Sciences Centre
www.bcgsc.ca
info@bcgsc.ca

1. Abstract

The annotated *D. melanogaster* genomic sequence is currently in its third revision (Release 3) and covers nearly all of the 120 Mb euchromatic DNA. The sequence assembly is curated by the Berkeley Drosophila Genome Project (BDGP) and comprised of data produced at Celera, Genoscope, Lawrence-Berkeley National Labs, Baylor College of Medicine and the European Drosophila Genome Project (EDGP). Drosophila continues to play a major role in providing a model for inheritance and gene interaction and a high quality assembly is required to ensure accuracy of sequence-based analysis. To this end, we have developed an automated data analysis pipeline for verification of the sequence assembly using multiple restriction enzyme digests of tiling path BAC clones. Various types of repeat regions produce incorrect, but self-consistent, sequence assemblies. These errors are very difficult to spot without an external validation method. The fingerprint verification method offers several benefits: the sequence is verified by an independent laboratory process, the fingerprints are robust in elucidating repeat elements and the data processing pipeline is extensible and can be adapted to any sequence data.

A set of 1,056 tiling path clones spanning the euchromatic portion of the genome were selected. Each clone was independently fingerprinted using 5 restriction enzymes. The enzymes were chosen to maximize coverage of the sequence with fragments in the size range of 1-20 kb to facilitate detection. The enzymes selected were ApaLI, BamHI, EcoRI, HindIII and XhoI. This combination provides coverage by at least two, three and four optimally-sized fragments for 99.9%, 98% and 87% of the sequence, respectively.

An in-silico fingerprint of each clone was derived from the sequence and compared to its experimental counterpart using a Needleman-Wunsch alignment and a 2% fragment size tolerance. Each base of the sequence is assigned a verification depth that corresponds to the number of experimentally verified in-silico fragments containing that sequence location. The average verification depth is used as a measure of overall verification. We have devised various figures of merit to identify clones with unverified subsequences and to categorize the discrepancies. An interactive web-based system has been created to visualize verification coverage.

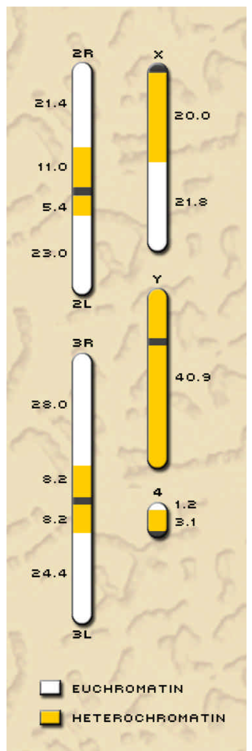
We are currently analyzing the tiling clone verification data and identifying potential authentic inconsistencies between the sequence-derived and experimental restriction maps. The method described here will also be applied to verification of heterochromatic DNA sequence, which is being generated using smaller clones. We anticipate that this fingerprint-based sequence verification methodology can positively impact the final sequence assembly quality of other organisms such as human, mouse and rat.

Drosophila melanogaster

Commonly known as the fruit fly, drosophila continues to play a significant role in the formulation of genetic inheritance and gene interaction models.

The fly's initial use in genetics hinged on polytene chromosomes. The larva maintains a constant cell count and each of its chromosomes divides 100's of times. All the strands stay attached producing a massively thick polytene chromosome, easily seen under the microscope.

Drosophila has 4 pairs of chromosomes: 2,3,4 and X/Y. The genome is 180Mb in size with approximately 14,000 genes and GC=0.46.



2. Verification Methodology

Independent verification of sequence assembly requires employing a quantitative method which is independent of any sequence-based results. The placement of end sequence hits on the genomic assembly can be used to provide a scaffold which is used by assembly algorithms to resolve inconsistencies and create larger contigs. Repeats and difficult-to-sequence regions, such as the heterochromatin, can cause the assembly to contain consistently sized but incorrect subsequences.

Enzyme Selection

Five restriction enzymes were chosen to maximize the depth of validation (Figure 3) and minimize the effect that undetectable fragments have on the validation. Fragments which are <600bp or >30kb are not reliably identified by our agarose electrophoresis method. The enzyme combination was therefore chosen to maximize coverage by optimally sized fragments (1-20kb).

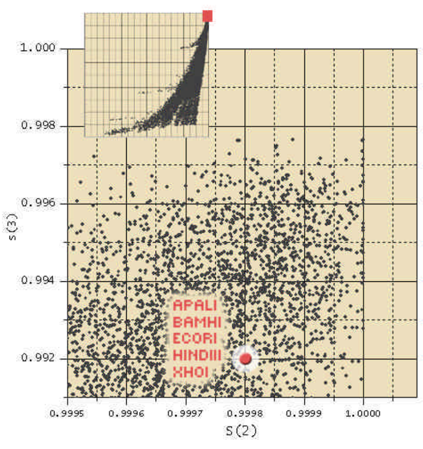


Figure 1 Profile of S(3) vs S(2) for all simulated combinations (*inset*). The extent of coverage by our 5 enzyme combination is highlighted in the zoomed part of the plot.

S(*i*) = fraction of genomic sequence in which every base pair is covered by at least *i* optimally sized fragments for a given enzyme combination.

The 5-enzyme combination was selected from about 200,000 computationally simulated combinations. The combinations were scored on the merits of ease of use and *S*(*i*) values. Some enzymes which contributed to better coverage are either not available in high concentration or require specialized laboratory conditions for reproducible digests.

The best practical choice of enzymes which yield consistent high-quality fingerprints and provide optimum coverage for Drosophila sequence is **ApaLI** (g.tgcac), **BamHI** (g.gatcc), **EcoRI** (g.aattc), **HindIII** (a.agctt) and **XhoI** (c.tcgac). The cut siteGC content is 53%.

In Silico Digest and Fingerprint Comparison

The in silico digest of each tiling set BAC was incorporated into the BACs library vector sequence in two orientations, to reflect the uncertainty of the orientation of the BAC insert. Each of these in silico fingerprints was compared with the experimental fingerprint to determine the orientation of the BAC insert and to allow verification of the junction fragments. Two fingerprints are compared using the Sulston score, which provides a quantitative measure of the probability that two fingerprints share a given number of bands by chance. The number of shared bands is computed using a global alignment algorithm with a uniform, relative size tolerance of 2%. The alignment attempts to match as many fragments as possible that are within 2% of their size between the fingerprints while minimizing the sum of differences for all matched fragments.

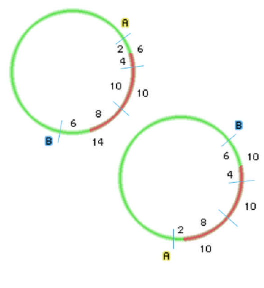


Figure 2 BAC insert can be incorporated into its vector in one of two orientations.

The analysis of a hypothetical sequence region with 3 enzymes is shown in Figure 3.

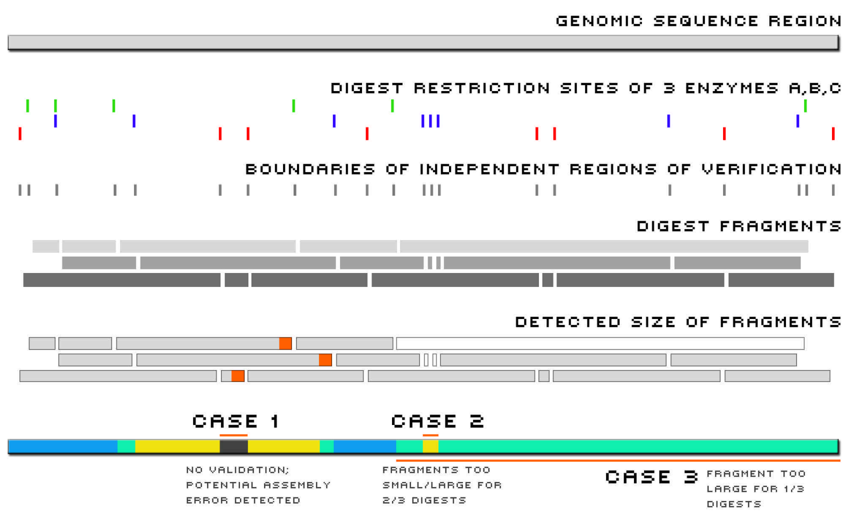

GENOMIC SEQUENCE REGION
DIGEST RESTRICTION SITES OF 3 ENZYMES A,B,C
BOUNDARIES OF INDEPENDENT REGIONS OF VERIFICATION
DIGEST FRAGMENTS
DETECTED SIZE OF FRAGMENTS
CASE 1
CASE 2
NO VALIDATION; POTENTIAL ASSEMBLY ERROR DETECTED
FRAGMENTS TOO SMALL/LARGE FOR 2/3 DIGESTS
CASE 3
FRAGMENT TOO LARGE FOR 1/3 DIGESTS
VERIFICATION DEPTH (DIGESTS)
3
2
1
0

Figure 3 Illustration of the fingerprint validation methodology. A genomic sequence region (e.g. tiling set BAC clone) is digested independently with multiple enzymes (e.g. 3). The density of restriction sites, found by performing an in-silico digest of the corresponding assembled sequence, determines the resolution of the method. In this example, some of the digest fragments are too large (*case 3*) or too small (*case 2*) to be detected. In addition, three overlapping fragments in every digest are found to be smaller than the in-silico prediction (*case 1*). By examining the validation depth across the sequence region, potential assembly errors can be identified.

Assembly History

Libraries

RPCI-98 partial EcoRI digest from Pieter de Jong

additional libraries (Hind, Nde) from EDGP

Release 1

March 2000

Celera provides 12X coverage. BDGP provides BAC physical map, 26 Mb finished sequence and 1.5X shotgun for each tiling set BAC

Release 2

October 2000

Celera/BDGP fill 330 gaps, leavling 1300 remaining gaps in the assembly

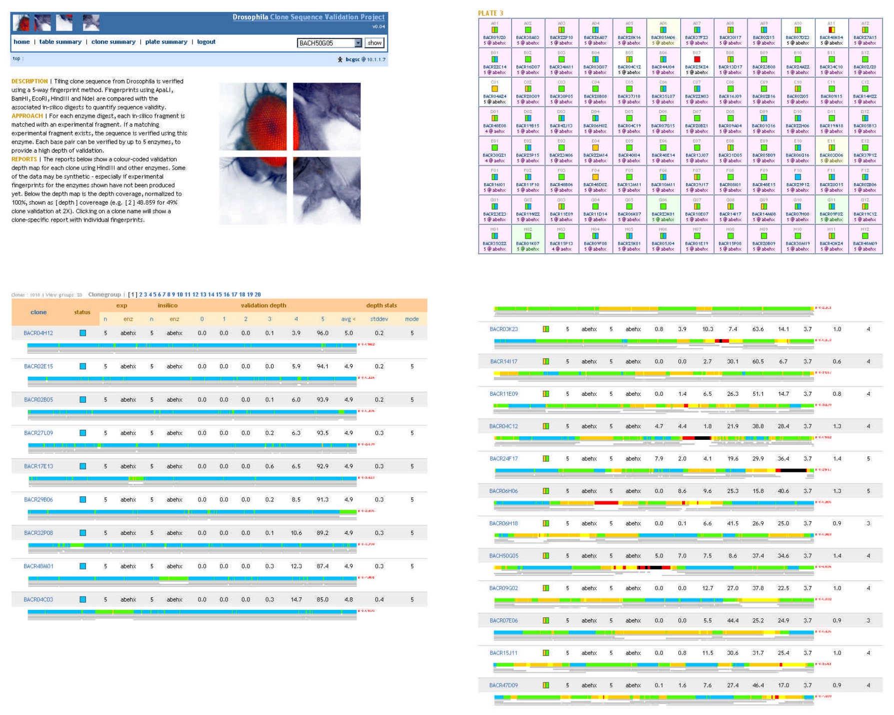
Release 3

July 2002

2L,2R,3R,4 and 12-20X finished by BDGP. 3L and 1-11X finished by Baylor

3. Application of Methodology to Fly BACs

A web-based front-end facilitates collaborative analysis and permits investigators to evaluate clones or sequence regions.


Drosophila Clone Sequence Validation Project
v0.04
home | table summary | clone summary | plate summary | logout
BACH50G05
show
top
bcgsc @ 10.1.1.7
PLATE 3

DESCRIPTION | Tiling clone sequence from Drosophila is verified using a 5-way fingerprint method. Fingerprints using ApaLI, BamHI, EcoRI, HindIII and NdeI are compared with the associated in-silico digests to quantify sequence validity.
APPROACH | For each enzyme digest, each in-silico fragment is matched with an experimental fragment. If a matching experimental fragment exists, the sequence is verified using this enzyme. Each base pair can be verified by up to 5 enzymes, to provide a high depth of validation.
REPORTS | The reports below show a colour-coded validation depth map for each clone using HindIII and other enzymes. Some of the data may be synthetic - especially if experimental fingerprints for the enzymes shown have not been produced yet. Below the depth map is the depth coverage, normalized to 100%, shown as [depth] covereage (e.g. [2] 48.859 for 49% clone validation at 2X). Clicking on a clone name will show a clone-specific report with individual fingerprints.

Figure 4 | *top left* | web-based data mining and visualization tool for verification results | *top right* | summary of verification for clones on a 96-well plate showing the average validation depth for each clone by colour (clone in B07 was associated with the wrong sequence) | *bottom left* | table view showing complete validation of BACs (5X blue, 4X green) | *bottom right* | similar table view showing BACs which contain subsequences which were not validated (3X orange, 2X yellow, 1X red, 0X black).

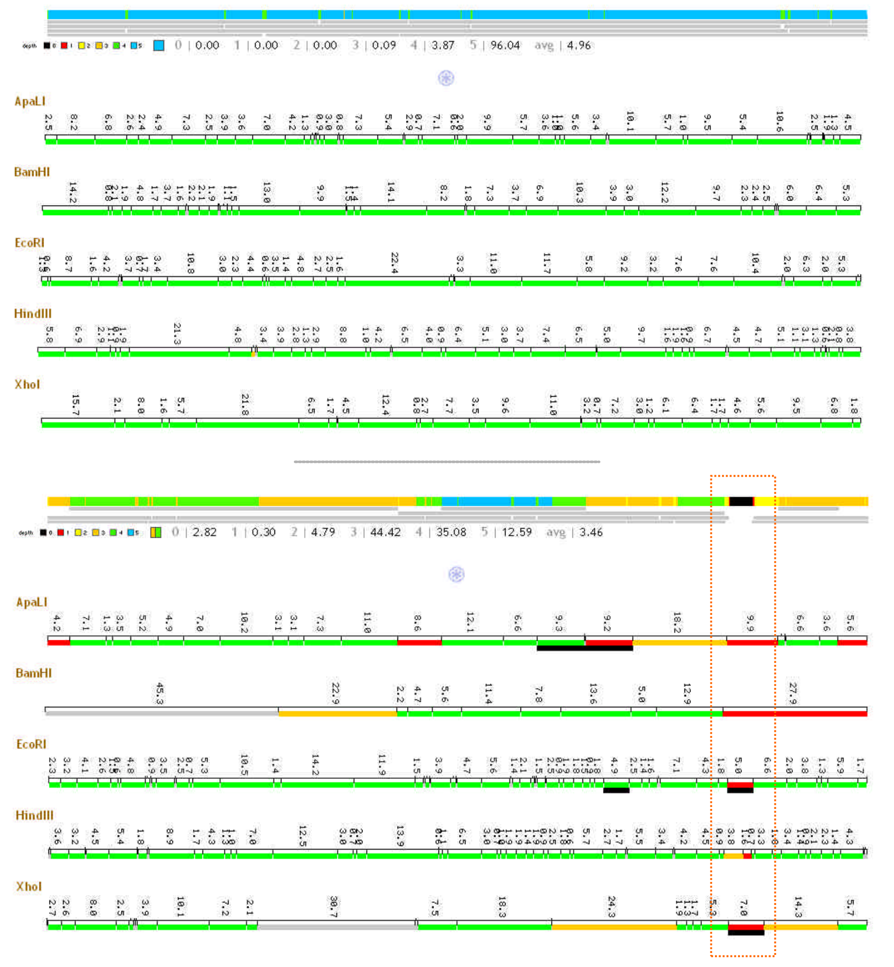


Figure 5 | *top* | restriction map view of digests of the clone BACR04H12 where every experimental fragment in each digest matches the in silico fragments | *bottom* | similar view of assembly of the heterochromatic clone BACR48M17 indicating a possible misassembly | **green** fragments indicate that the experimental fragment matched the in silico fragment at 2% size tolerance; **yellow** indicates a match at 10%; **red** indicates that no experimental fragment of the expected size was found; **grey** is reserved for fragments <600bp and >30kb, which fall outside of the reliable detection limit; **black** corresponds to ambiguous validation where an excess of in silico fragments are found.

Figure 6 shows one method to categorize BACs by their validation profile. Using the *S*(*i*) quantity, BACs with a large *S*(i) average but showing significant *S*(0) and *S*(1) validation are flagged as having potential assembly errors.

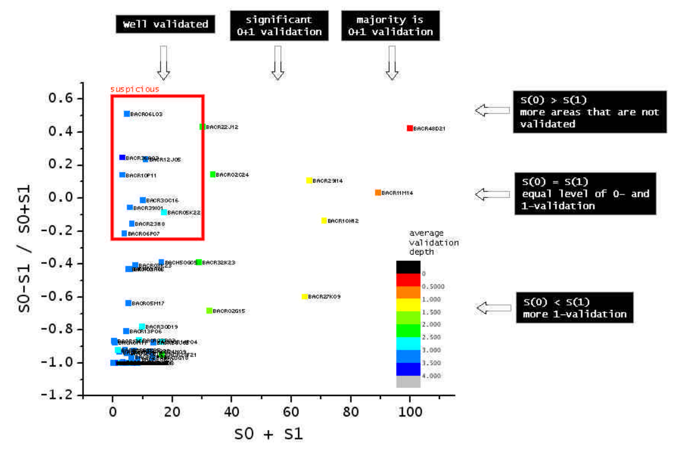
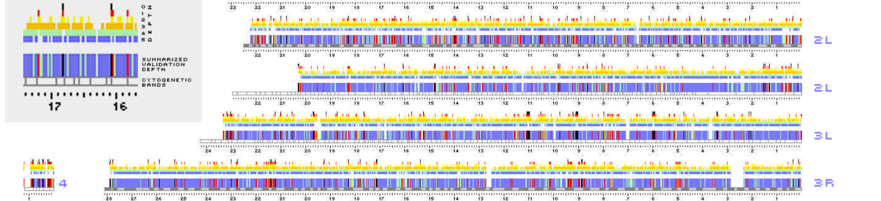


Figure 6 Using *S(i)*, BACs can be categorized. For most bacs the quantity $x=S(0)+S(1)$, corresponding to fraction of their sequence not validated at all or validated by only one digest, is expected to be small. The proportion of *S*(0) validation is found by using $y=(S(0)-S(1))/x$. BACs for $(x,y) = (0,-1)$ are perfectly validated and those with $(x,y)=(u,v)$ for small u and $v > -0.5$ contain local inconsistencies.

4. Genomic Summary

To date, we have applied the verification method to 119 Mb of the genome. Approximately 330kb (Figure 7) of the total assembly was found to be inconsistent with the fingerprints derived from these regions.



CHR	ANALYZED		LACKS VALIDATION	
2L	22.6MB	97%	47KB	0.21%
2R	20.7MB	97%	50KB	0.25%
3L	23.8MB	95%	83KB	0.37%
3R	28.4MB	95%	29KB	0.11%
4	1.3MB	70%	10KB	1.09%
X	22.2MB	68%	144KB	0.95%

Figure 7 Levels of validation for each chromosome assembly. Colour code is the same as used in the clone summary views in Figure 4. The scale is shown in Mb.

Acknowledgments

Clone libraries | RPCI-98 BAC PAC Resource Centre www.chori.org/bacpac | *BAC physical map and sequence* | Berkeley Drosophila Genome Project www.fruitfly.org

[†]Berkeley Drosophila Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA, USA | [‡]Dept of Molecular and Cellular Biology and Howard Hughes Medical Institute, UC Berkeley, Berkeley, CA, USA

Genome Sciences Centre • BC Cancer Agency • 600 W 10th Ave • Vancouver BC V5Z 4E6 • Canada • tel.604.877.6086 • fax.604.877.6085 • www.bcgsc.ca