UNIVERSITÁ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

DOTTORATO DI RICERCA IN

FISIOLOGIA MOLECOLARE E BIOLOGIA STRUTTURALE

CICLO XX

TOMATO (*Solanum lycopersicum*) GENOME PROJECT: SEQUENCING AND ANALYSIS OF CHROMOSOME 12

Coordinatore: Ch.mo Prof. BENEDETTO SALVATO

Supervisore: Ch.mo Prof. GIORGIO VALLE

Dottoranda: SARA TODESCO

31 GENNAIO 2008

1.

Introduction

Contents:

In the genomics world, plants are second-class citizen. Researchers have completed the genome of hundreds of microbes and dozen of animals, yet they have deciphered the genomes of just four plants, *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Oryza sativa* (International Rice Genome Sequencing Project, 2005), *Populus trichocarpa* (Tuskan GA *et al.*, 2006) and *Vitis vinifera* (The French-Italian Public Consortium for Grapevine Genome Characterization, 2007). Genomic comparisons can yield tremendous insights into the evolutions of these organisms. To better understand the complex genetic system of diverse higher plant species, it is necessary to analyze plants in different taxa with characteristic feature.

I.1 The *Solanaceae* family

The *Solanaceae* family is the third most valuable crop family exceeded only by the grasses (e.g. rice, maize, wheat) and legumes (e.g. soybean), and the most valuable in terms of vegetable crops. The family is composed of more than 3000 species, including the tuber-bearing potato (*Solanum tuberosum*), a number of fruit-bearing vegetables (tomato [*Solanum lycopersicum*], eggplant [*Solanum melongena*], and peppers [*Capsicum annuum*]), ornamental plants (petunia [*Petunia hybrida*]), plants with edible leaves (*Solanum aethiopicum*, *Solanum macrocarpon*) and medicinal plants (*Datura*, *Capsicum*). Multiple important species in the family are major contributors to fruit and vegetable consumption and thus human health.

In addition to their role as important food service, many solanaceous species have a role as scientific model plants, such as tomato and pepper for the study of fruit development (Gray JP *et al.*, 1992; Fray RG *et al.*, 1993; Hamilton AJ *et al.*, 1995; Brummell DA *et al.*, 2001; Alexander L *et al.*, 2002; Adams-Philips L *et al.*, 2004; Giovannoni JJ, 2004; Tanksley SD, 2004), potato for tuber development (Prat S *et al.*, 1990; Fernie AR *et al.*, 2001), petunia for the analysis of anthocyanin pigments, and tomato and tobacco (*Nicotiana tabacum*) for plant defence (Bogdanove AJ *et al.*, 2000; Gebhardt C *et al.*, 2001; Li L *et al.*, 2001; Pedley KF *et al.*, 2003).

For several thousand years, solanaceous crops have been subjected to intensive human selection. This had led to an enormous phenotypic diversity within species and to the

adaptation of individual varieties to widely different habitats. *Solanaceae* species thrive in some of the most diverse natural habitats that include rain forests, deserts and the high altitudes of Andean mountains.

I.2 The *Solanaceae* genomes are highly conserved

Comparative genomic mapping showed that the *Solanaceae* genomes have undergone relatively few genome rearrangements and duplications and therefore have a high level of conservation of organization at macro and micro level (Tanksley SD *et al.*, 1992; Livingstone KD *et al.*, 1999; Doganlar S *et al.*, 2002).

Numerous events of polyploidy within both the grasses and *Brassicaceae* have led to segmental duplication, selective gene losses and significant genome reshuffling. As a result, species in the grasses and crucifers are characterized by different chromosome numbers coupled with extensive loss of microsinteny between the paralogous segments of *Brassica* chromosomes, and between those and their *Arabidopsis* homologs. The *Solanaceae* family is unique in that there have been no large-scale duplication events (e.g. polyploidy) early in the radiation of the family; most species possess the same number of chromosomes (2n=2x=24). The polyploidy events (e.g. tetraploid potatoes and tobacco) are all recent events and the diploid forms of both these species are still in existence. Therefore, microsinteny conservation amongst *Solanaceae* genomes is very high.

This high level of genome conservation makes the *Solanaceae* family a model to explore the basis of phenotypic diversity and adaptation to natural and agricultural environments.

I.3 The International Solanaceae Genome Project (SOL)

To meaningfully analyze the gene-to-phenotype relationships, a large amount of sequencing information is necessary. As the high cost of sequencing prohibits direct comparison between full *Solanaceae* genomes, the most cost-effective way to get sufficient information is to sequence a high-quality reference genome and then map sequence (e.g. ESTs) from other organisms onto the reference genome. To fulfil this objective, on November 2003, researchers from more than 10 countries, representing academic and industry laboratories with interest in the *Solanaceae*, met to kick off the initiative called 'The International Solanaceae Genome Project' (SOL) (http://sgn.edu/Solanaceae-project/). As central part of its systems approach to increase diversity and adaptation in

crop plants, the SOL launched the initiative to sequence the full euchromatic portion of the tomato (*Solanum lycopersicum*) genome (Mueller LA *et al.*, 2005). The tomato genome sequencing will be a worthy reference for comparative mapping with other member of the *Solanaceae* family, like potato, eggplant, pepper, tobacco and petunia, and with other dicots and monocots through the fully sequenced genomes of *A. thaliana*, rice, grapevine and poplar.

I.4 Tomato (*Solanum lycopersicum*): a reference for *Solanaceae* genomes

Tomato (*Solanum lycopersicum*) is the most intensively studied *Solanaceae* genome due to its simple diploid genetics, short generation times, routine transformation technology, and available rich genetic and genomic resources. Tomato has been chosen as the model for the *Solanaceae* family because it has a relatively small genome size of 950 Mb (Arumuganathan K *et al.*, 1991) for which homozygous inbreds, a dense genetic map (Tanksley SD *et al.*, 1992) and an advanced BAC-physical map are available to initiate the sequencing project (http://www.sgn.cornell.edu/solanaceae-project/).

The tomato nuclear genome size (1C) is generally considered as approximately 0.95 pg of DNA, corresponding to 950 Mb of DNA organized into 12 acrocentric to metacentric chromosomes (n=x=12).

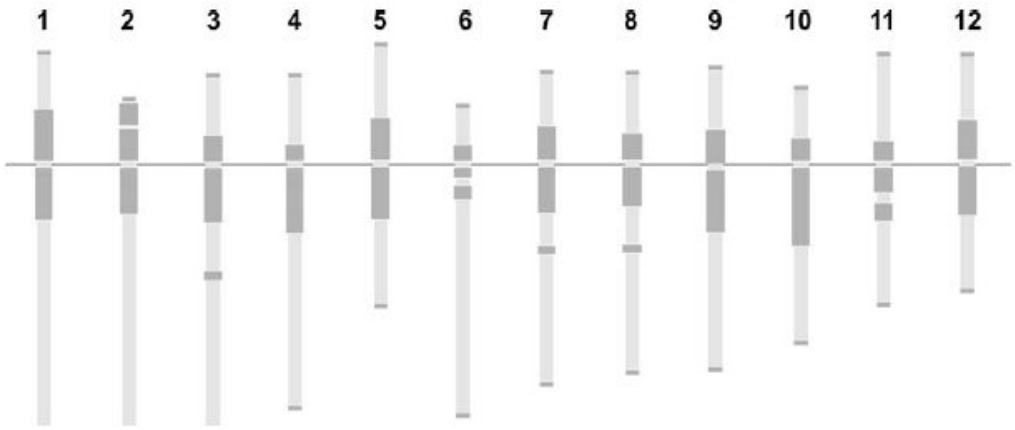


Fig 1.1. The morphology of tomato chromosomes at pachytene. Centromeres are aligned to a line, blocks of dark grey represent heterochromatin and light ones represent euchromatin. (Zhong XB *et al.*, 1998)

Unlike the chromosomes of maize and rice, in which heterochromatin and euchromatin are interspersed, each tomato chromosome has well-differentiated euchromatin and heterochromatin regions, with heterochromatin concentrated around the centromere. While the heterochromatin regions are largely devoid of genes and constitute approximately the 77% of the genome (Peterson DG *et al.*, 1996; van der Hoeven R *et al.*, 2002; Wang Y *et al.*, 2006), the remaining 23% of the DNA is organized into long continuous stretches of gene-rich euchromatin, located in the distal portion of each chromosome arm (Fig 1.1).

Rather than sequencing the entire tomato genome, the SOL committee proposed to sequence the approximately 220 Mb of euchromatin that contains the majority of protein-coding genes. For this purpose, *Hin*dIII, *Mbo*I and *Eco*RI-digested BAC (Bacterial Artificial Chromosome) library were prepared from *Solanum lycopersicum* var Heinz 1706 (Budiman MA *et al.*, 2000). Sequencing follows a BAC-by-BAC strategy that is to perform shotgun sequencing on a minimal tiling path of BAC clones through the 220 Mb of euchromatin.

To obtain this, the SOL pooled the resources of sequencing groups in 10 countries: Korea (chromosome 2), China (chromosome 3 and 11), Great Britain (chromosome 4), India (chromosome 5), The Netherlands (chromosome 6), France (chromosome 7), Japan (chromosome 8), Spain (chromosome 9), The United States (chromosomes 1 and 10), and Italy (chromosome 12).

I.5 The tomato chromosome 12 project

The Italian SOL was established in 2004 and aims to complete the sequencing of euchromatin of chromosome 12.

Metacentric chromosome 12 is the smallest chromosome in the tomato genome, with an estimated size of 76 Mb. DAPI-stained pachytene chromosome 12 is relatively rich in heterochromatin with large characteristic symmetrical pericentromeric blocks on both arms interrupted with a variable number of small weakly fluorescency gaps; both arms have small heterochromatin knobs at the distal ends (Fig 1.2).

The Italian project is supported by the Agronanotech Project (MIPAF, Italy), by the FIRB project (MUR, Italy) and by the EU-SOL project (European Union) and is a collaboration between different laboratories, each contributing with specific competence. Besides sequencing the euchromatic portion of tomato chromosome 12 (G. Valle group, University of Padua), key activities of the Italian SOL include annotation and analysis of the sequenced portion of the tomato genome (G. Valle group, University of Padua; ML

Chiusano group, University Federico II, Naples), identification of genes related to biotic and abiotic stress response and of genes involved in fruit development (L. Frusciante group, University Federico II, Naples; G. Giuliano group, ENEA, Rome).

Fig 1.2. The pachytene tomato chromosome 12 stained with DAPI. DAPI staining reveals the chromatin morphology including the distal heterochromatin blocks of short and long arms and the two pericentromeric heterochromatin regions. (Chang SB, unpublished)

I.6 Scope of this PhD thesis

The International SOL initiative started a global collaboration in which tomato (*Solanum lycopersicum*) as model for all *Solanaceae* species is sequenced. As part of this project, my PhD study focused on the sequencing of tomato chromosome 12 following a BAC-by-BAC basis in the minimum tiling path strategy. In paragraph § 3.2 I present an overview of the sequencing project. I highlight the complexity of the efforts and the further difficulties arisen with the progress of the project; as consequence, the data presented in this thesis are still preliminary.

In paragraph § 3.1 I present an informatics tool called PABS (Platform Assisted BAC-by-BAC Sequencing) that we developed to optimize the BAC-by-BAC sequencing strategy and to try to overcome some of the occurred difficulties. This work received attention by the researcher in the SOL project that can access to the program from our web page (http://tomato.cribi.unipd.it/files/bioinformatics.html).

Paragraph § 3.3 presents the development of high-resolution BAC-FISH mapping technique. By using multi-colour FISH on combed tomato genomic DNA molecules, the distances, overlaps and orientation of the selected BAC clones can be accurately evaluated.

Finally in paragraph § 3.4 I present a preliminary analysis of the available tomato genomic sequences. We annotated the available fully sequenced BAC clones using a combination of bioinformatics methods, relaying on sequence homology detection and *ab initio* gene prediction.

2.

Material and Methods

Contents:

II.1 Protocol 1: preparation of BAC DNA

II.1.1 Materials

Reagents and Solutions

3 M NaAc pH 5.2

Ethanol and 70% ethanol

Isopropanol

CsCl (solid)

Ethidium bromide (10 mg/ml)

1-butanol, saturated with H_2O

Chloramphenicol 25 mg/ ml

Media

LB medium

II.1.2 Purification of BAC DNA

BACs are supplied as frozen cultures stored at -80°C in 384 well-plates.

BAC DNA is purified from bacterial culture that has been inoculated with a single transformed colony picked from a freshly streaked agar plate.

1. Inoculate in a 15 ml Falcon 3 ml of LB medium containing 25 µg/ml of chloramphenicol with a single BAC colony. Incubate the culture O/N at 37°C with vigorous shaking (230-250 rpm).
2. In a 5 l flask inoculate 1.5 l of LB medium containing 25 µg/ml of chloramphenicol with 2 ml of the saturated O/N culture.
3. Incubate the culture O/N at 37°C with vigorous agitation (200-230 rpm).
4. Harvest the cell from the cultures by centrifugation at 6000 x g for 15 minutes at 4°C. Pour off the supernatant, and invert the open centrifuges bottles to allow the last drops of the supernatant to drain away.
5. Follow the suggested protocol of the commercial kit NucleoBond® PC 2000 (Macherey-Nagel). NucleoBond® PC employ a modified alkaline/SDS lysis

procedure, where both chromosomal and BAC DNA are denatured. Potassium acetate is then added to the denatured lysate, which causes the formation of a precipitate containing chromosomal DNA and other cellular compounds. The potassium acetate buffer also neutralizes the lysate. BAC DNA, which remains in solution, can revert to its native supercoiled structure. After equilibrating the NucleoBond® column with equilibration buffer, BAC DNA is bound to the anion-exchange resin and finally eluted after washing of the column.

▲ *Mix by inversion very gently to prevent BAC DNA damage. After alkaline lysis, filtrate the solution through the supplied the NucleoBond® folder filters to clarify the solution from the cell debris in order to prevent clogging of the column.*

6. In a corex tube precipitate the eluted DNA adding 0.8 volumes of room-temperature isopropanol. Centrifuge at x g for 60 minutes at 4°C.
   Carefully discard the supernatant and wash with 7 ml of room-temperature 70% ethanol. Centrifuge at x g for 20 minutes at 4°C.
   Allow the pellet to dry at room-temperature.
   Redissolve the DNA pellet in 5 ml of 1X TE pH 8.0.
7. Determine the BAC yield and quality (as quotient 260 nm/280 nm) by UV spectrophotometry and confirm DNA integrity by 0.8% agarose (SeaKem LE) gel electrophoresis.

II.1.3 Purification of closed circular BAC DNA by equilibrium centrifugation in CsCl-ethidium bromide continuous gradients

I choose equilibrium centrifugation in CsCl-ethidium bromide continuous gradient as the method to separate BAC DNA from chromosomal DNA.

Ethidium bromide and BAC DNA are mixed with a CsCl solution. When the mixture is centrifuged at high speed, the centrifugal force is sufficient to generate and maintain a gradient of cesium atoms. During formation of the gradient, DNAs of different buoyant densities migrate to positions in the tube at which the density of the surrounding CsCl solution equals that of the DNA itself. During centrifugation to equilibrium superhelical closed circular plasmid DNA and non-superhelical DNAs form bands at different buoyant densities.

1. In a corex tube for 4.7 ml of BAC DNA solution, add 4.7 g of solid CsCl. Mix the solution gently until the salt is dissolved.
2. Add 300 µl of 10 mg/ml ethidium bromide and mix gently.
3. Centrifuge the solution at 5000 rpm (Beckman) for 10 minutes at room temperature (20°C).

4. Use a disposable hypodermic syringe to transfer the solution into a Quick-Seal tube, avoiding air bubbles. Make sure that the weights of tubes opposite each other in the rotor are equal.
5. Seal the tubes.
6. Centrifuge at 20°C for 16 hours at 50000 rpm using Beckman VTI 65.2 rotor, without brake.
7. At the end of the centrifuge run, gently remove the rotor from the centrifuge. Carefully remove each tube and place in a tube rack covered with a tin foil.
8. In a dimly lit room mount one tube in a clamp attached to a ring stand.
9. Collect the band of closed circular BAC DNA:
    a. Use a 21-gauge hypodermic needle to make a small hole in the top of the tube to allow air to enter when the fluid is withdrawn.
    b. Attach a 1 ml disposable syringe to a sterile 26-gauge hypodermic needle and insert the needle (beveled side up) into the tube just below the lower DNA band (closed circular BAC DNA).
    c. Slowly withdraw the BAC DNA, taking care not to disturb the upper band of chromosomal DNA.

    ▲ *To avoid contamination with the chromosomal DNA, do not attempt to remove every visible trace. The upper band consists of chromosomal DNA and linearized and/or single-strand nicked BAC DNA: I collect even the upper band to use as template for sequencing, PCR or fingerprinting reactions.*

II.1.4 Removal of Ethidium bromide from the DNA solution

1. Ethidium bromide is removed from DNA purified through a CsCl gradient by repeated extraction with organic solvent.
    a. In a 1.5-2 ml eppendorf, to the DNA solution add an equal volume of water-saturated 1-butanol.
    b. Mix the organic and aqueous phases by inversion.
    c. Centrifuge the mixture at for 3 minutes at room temperature.
    d. Remove the upper (organic) phase.
    e. Repeat the extraction (Steps 1-4) four to six times until all pink colour disappear from both the aqueous phase and the organic phases.

II.2 Protocol 2: pUC19_*Bst*XI vector

The pUC19_*Bst*XI vector is a derivative of pUC19. It has been constructed by cloning a 500 bp stuffer fragment into pUC19 vector; two new *Bst*XI sites has been inserted at positions flanking the stuffer fragment to be removed when creating a plasmid library.
As pUC19, the pUC19_*Bst*XI vector is a high copy number plasmid with the pMB1 replicon rep responsible for the replication of plasmid the bla gene, coding for β-lactamase that confers resistance to ampicillin. It differs from pUC19 in the multiple cloning site (pUC19_*Bst*XI has loose the pUC19 polylinker) and in the blue-white colonies screening (the insertion of the 500 bp stuffer has brought out of frame the N-terminal fragment of β-galactosidase).

II.2.1 Materials

Reagents and Solutions
Ampicillin 50 mg/ ml
Glycerol
3.5 M NaAc pH 5.2
Ethanol and 70% ethanol
Ethidium bromide (10 mg/ml)
50X TAE

Media
LB medium

Vectors and Hosts
pUC19
Chemical competent DH10B *E. coli*

II.2.2 pUC19 plasmid purification

pUC19 plasmid DNA is purified from a bacterial culture that has been inoculated with a single pUC19-transformed colony picked from a freshly streaked agar plate. To optimize the quality of plasmid DNA, the bacterial culture is grown until it reaches an OD550 of 0.9-1; the DNA is purified using Nucleospin Plasmid DNA Purification Kit (Macherey-Nagel). The DNA is eluted with H2O mQ AF.
Determine the plasmid yield and quality (as quotient 260 nm/280 nm) by UV spectrophotometry and confirm DNA integrity by 1% agarose (SeaKem LE) gel

electrophoresis.

II.2.3 pUC19 digestion with *Hin*dIII and *Eco*RI

The pUC19 vector is double digested with *Hin*dIII and *Eco*RI, and then separated in a 0.8% low-melting point agarose gel.

1. Digest 2 µg of pUC19 DNA in a small volume (possibly 20 µl) with 20 units of *Hin*dIII (Neb; 20/µl) and 20 units of *Eco*RI (Neb; 20/µl) for 90 minutes at 37°C.
2. Inactivate the reaction by heating at 80°C for 20 minutes. Then cool to room temperature and place reaction on ice.
3. To verify complete digestion, run an aliquot of the digestion on a 1% agarose gel.
4. Run the double digested plasmid on a 0.8% low-melting point agarose (Invitrogen) gel in a four adjacent wells, at 50V for 3 hour in a 4°C room.
5. At the end of electrophoresis, the gel portions corresponding to the marker is cut and stained with a fresh ethidium-bromide solution (10 mg/ml EtBr in TAE 1X) for approximately 20 minutes.
    a. Place the gel on a UV transilluminator to reveal the stained bands of the marker. Apply little cuts in correspondence of the band of linearized vector.
    b. Rebuild the gel. With a clean sharp blade, excise the gel slice containing the linearized vector, using the cuts as references. Minimize the amount of surrounding agarose excise. Place the gel slices in a 2 ml eppendorf and weight the gel slice.
6. Purify the DNA using PureLink Quick Gel Extraction Kit (Invitrogen). Follow the protocol given and at the end elute the sample with 50 µl of H2O mQ AF (pre-worm at 65°C).
7. Take 5 µl out and quantify on a 1% agarose gel.

II.2.4 Stuffer fragment preparation

The stuffer fragment is designed to introduce in pUC19 vector two *Bst*XI site.
*Bst*XI recognizes a 6 bp palindrome interrupted by an arbitrary six base-pair sequence (CCANNNNNNTGG). The newly inserted *Bst*XI sites contain different arbitrary sequence. Digestion of the vector with *Bst*XI results in two fragments: a 2.6 kb vector and a 500 bp stuffer fragment. The two 5′ ends of the cloning vector are not complementary to each other, which suppress vector self-ligation, but are complementary to the adapter-ends ligated to the sheared DNA fragments.

1. The stuffer fragment has been PCR amplified using two primers containing:
   - ✓ both the *Bst*XI site (CCANNNNNTGG),
   - ✓ both a **3′ portion** overlapping the DNA template,
   - ✓ the *Hin*dIII site (AAGCTT) (*Hin*dIII_primer) or the *Eco*RI site (GAATTC) (*Eco*RI_primer).

   *Eco*RI_primer

   5′ – TACGAATCCAAGTG**TATGGAACCTGACTTACTAG** – 3′

   *Hin*dIII_primer

   5′ – GCCAAGCTTCCAAGT**GTATGGTAGAAAGATCATCT** - 3′

   As template for the PCR reaction I used a plasmid clone of the shotgun library of *Solanum tuberosum* cultivar Desiree chloroplast DNA sequencing project. The stuffer fragment matches on the *Solanum tuberosum* cultivar Desiree chloroplast sequence (DQ386163) at 30316-30798 bp.

   Set up the following reaction:

plasmid DNA	20 ng
*Eco*RI_primer (10 µM)	0.4 µl
*Hin*dIII_primer (10 µM)	0.4 µl
dNTPs (10 mM)	0.4 µl
$MgCl_2$ (50 mM)	0.6 µl
10X buffer (Polymed)	2 µl
Taq polymerase (Polymed)	0.1µl
H_2O mQ AF	15.1
Total volume	20 µl

   Set up the following PCR program on the thermocycler:

   95°C 5 min.

   95°C 15 sec.
   60°C 20 sec. } x 30 cycles
   72°C 1 min.

   72°C 5 min.

   4°C hold

   Load 5 µl of sample onto a 1% agarose gel, using 1-kb DNA ladder. Run the gel for 30 minutes at 120 V and check for insert.

2. Purify the PCR product with PureLink PCR Purification Kit (Invitrogen), following the user manual procedure. At the end, elute the sample with 50 µl of H2O mQ AF (pre-worm at 65°C).

3. Double digest 4 5 µl of PCR product with *Eco*RI and *Hin*dIII, then separated in a 0.8% low-melting point agarose (Invitrogen) gel. Follow the steps of II.2.3.

II.2.5 Stuffer-vector ligation

1. The stuffer fragment is cloned into the double-digested pUC19 vector, using a molar ratio 1:3 vector DNA to stuffer DNA. The reaction is performed in 20 µl using 50 ng of plasmid vector and 1 µl of T4 DNA polymerase (Neb); the mixture is incubated O/N in a water bath at 16°C, and then inactivated at 65°C for 20 minutes.
2. Transform 4 µl of ligation mix into 200 µl of chemical competent DH10B *E. coli* (§ II.3.10). After transformation, plate 10 µl and 100 µl on LB-ampicillin 50 µg/ ml plates.

II.2.6 Screening

1. Determine some clones harbouring the 500 bp stuffer as insert by PCR with the primer flanking the cloning site:

40 M13 primer
5′ – GTTTTCCCAGTCACGAC – 3′
-20 M13 Rev primer
5′ – GTGGAATTGTGAGCGGA – 3′

-40 M13 primer (10 µM)	0.4 µl
-20 M13 Rev primer (10	0.4 µl
dNTPs (10 mM)	0.4 µl
$MgCl_2$ (50 mM)	0.6 µl
10X buffer (Polymed)	2 µl
Taq polimerase (Polymed)	0.1µl
H_2O mQ AF	15.1 µl
Total volume	19 µl

Using pipette tips, pick the colonies from the agar plate into 10 µl of 1x buffer of Taq Polimerase. Mix tips before through them away.
Add 1 µl of the 1x buffer of Taq polimerase containing the colonies as template of the PCR reaction.

Set up the following PCR program on the thermocycler:

95°C 5 min.

95°C 15 sec.
55°C 15 sec.
72°C 1 min.
} x 34 cycles

72°C 5 min.

4°C hold

Load 5 µl of sample onto a 1% agarose gel, using 1-kb DNA ladder. Run the gel for 30 minutes at 120 V. Image the gel and check for inserts.

2. Isolate mini preparative DNA from 2–5 PCR positive colonies. Use Nucleospin Plasmid DNA Purification Kit (Macherey-Nagel), starting from a 3 ml O/N culture. For confirming the correct insertion of the stuffer:
   - **a.** Approximately 500 ng of vector DNA is digested with 10 units of *Bst*XI (Neb; 10 U/ µl) in 10 µl reaction mixture at 37°C for 1 hour. Gel electrophorese restriction digests.
   - **b.** Approximately 500 ng of vector DNA is double-digested with 5 units of *EcoR*I (Neb; 20 U/ µl) and 5 units of *Hin*dIII (Neb; 20 U/ µl) in 10 µl reaction mixture at 37°C for 1 hour. Gel electrophorese restriction digests.
   - **c.** Sequence the plasmid DNA using the 40 M13 primer and -20 M13 Rev primer.
3. Prepare two-four 20% glycerol stocks of the recombinant clone containing the pUC19_*Bst*XI plasmid vector. Store at -80°C freezer.

II.2.7 pUC19_*Bst*XI vector preparation for shotgun library construction

1. Streak out the strain transformed with pUC19_*Bst*XI onto an LB plate with ampicillin 50 µg/ ml. Incubate plate O/N at 37°C.
   Inoculate 30 ml LB-ampicillin with a single colony, shaking ~230 rpm at 37°C until it reaches OD550 1. Prepare DNA using Nucleospin Plasmid DNA Purification Kit (Macherey-Nagel); follow the protocol and elute the sample with H2O mQ AF.
   Determine the plasmid yield and quality (as quotient 260 nm/280 nm) by UV spectrophotometry and confirm DNA integrity by 1% agarose (SeaKem LE) gel electrophoresis.
2. Confirm the correctness of the pUC19_*Bst*XI vector by sequencing with -40 M13 primer and -20 M13 Rev primer.
3. Digest 3 µg of pUC19_*Bst*XI plasmid DNA with *Bst*XI (Fermentas). Since it is very

important to cut the DNA to completion, use an excess of restriction enzyme; often however the use of excess restriction enzyme can lead to other problems. Therefore it is best to use 8 units of enzyme for cutting 1 µg of plasmid DNA.

Set up the following program on the thermocycler:

1. 50°C 90 min (digestion)

2. 65°C 20 min (inactivation)

3. 4°C hold

Run an aliquot (~100 ng) on a 0.8 % gel to confirm complete digestion.

4. Load onto a 0.8% low-melting point agarose (Invitrogen), and electrophoreses for 3 hours at 50 V. Following the protocol previously described (§ II.2.3) dissect just the linearized plasmid DNA band away from all other contaminating plasmid fractions (open circular and supercoiled). Also recover the 500 bp band of stuffer fragment.

Purify the *Bst*XI-linearized vector DNA and the stuffer DNA using PureLink Quick Gel Extraction Kit (Invitrogen). Follow the protocol given and at the end elute the samples with 50 µl of H2O mQ AF (pre-worm at 65°C).

For each sample, take 5 µl out and quantify on a 1% agarose gel. Also determines the DNA concentration and quality by UV spectrophotometry.

5. Aliquot the *Bst*XI-linearized pUC19_*Bst*XI vector at 15-25 ng/µl: usually, ligation is done with 25 ng of vector. Store even the 500 bp stuffer in aliquot useful to set up a control ligation with a molar ratio vector DNA to stuffer DNA 1:3.

Store at -20°C.

II.2.8 Adapters preparation

The *Bst*XI adapters are phosphorylated oligonucleotides:

***Bst*XI adapter primer-For**

5' - GCGGCCGCACACAC - 3'

***Bst*XI adapter primer-Rev**

5' - GTGCGGCCGC - 3'

1. Resuspend dry oligonucleotides to a final concentration of 100 µM with H2O mQ AF. Vortex well.

2. In a PCR tube, mix oligonucleotides For and Rev such that an adapter is obtained at a concentration of 50 µM.

3. Anneal the pair of oligonucleotides by cooling the mixture from 95°C to 4°C. On a thermocycler set up a program where:

- ✓ Temperature gradient = from 95°C to 4°C
- ✓ Temperature increasing = -1°C
- ✓ Time of each temperature steps = 30 seconds
- ✓ Hold = 4°C

4. At the end of the cycle, immediately put on ice. Prepare aliquots of 10 µl in PCR tubes and store at -20°C.

II.3 Protocol 3: Shotgun library

The goal is to create a shotgun library that provides a tenfold sequence redundancy over the BAC clone. The success and efficiency of this process is dependent on random fragmentation of the DNA and unbiased cloning of these fragments to generate a random shotgun library.

II.3.1 Materials

Reagents and Solutions

Ampicillin 50 mg/ ml

NaOH 10 M

HCl 1 M

Ethanol and ethanol 70%

NaAc 3 M pH 5.2

Ethidium bromide (10 mg/ml)

TAE 50X

dNTPs 10 mM

Media

LB medium

SOC medium

Vectors and Hosts

pUC19_*Bst*XI (prepared as described in § 3)

Chemical competent DH10B *E. coli*

II.3.2 Fragmentation of BAC DNA

Two methods are used to cleave double-stranded DNA into fragments of a suitable size for shotgun sequencing: mechanical and enzymatic cleavage. Mechanical methods of DNA fragmentation (e.g., sonication, hydrodynamic shearing) are often preferred over enzymatic methods, as they are more random and reduce the bias of sequencing projects. For the generation of shotgun library I used two different hydrodynamic shears, the sonicator and the Hydroshear (GeneMachines), to shear 3 µg of BAC DNA.

To fragment the DNA by sonication:

a. Before use, wash the sonicator tip (Sonic Dismembrator Model 300, Fisher) with three washes with NaOH 0.2 M and then with, at least, 4 washes with H_2O mQ AF.

b. Place the DNA solution in a 2 ml eppendorf and bring to a final volume of 400 µl with H_2O mQ AF. Vortex briefly to mix the solution.

c. Place the eppendorf containing the DNA in the sonicator such that the bottom of the tube is 1-2 mm above the hole in the center of the cup horn probe.

d. Sonicate the DNA. For most DNA sample, a 1-second/ µg DNA pulse at a power setting of 30 typically produce fragments of 1500-2500 bp. Keep the sample in ice.

▲ *Estabilish the appropriate conditions (number and duration of pulses) for sonication by sonicating a test sample.*

e. Centrifuge briefly to collect the sonicated DNA sample at the bottom of the tube and place it on ice.

To fragment the DNA by Hydroshear:

▲ *All solutions (NaOH 0.2 M, HCl 0.2 M, H_2O mQ) must be filtrated through a 0.22 µm filter. Wash the device before ad at the end of the procedure.*

a. In a 1.5 ml eppendorf, bring the BAC DNA sample to a final volume of 320 µl with H_2O mQ AF.

b. Vortex and incubate for 30 minutes at 37°C under agitation.
Spin for 20 minutes at 12000 rpm.

c. Pipette 300 µl into a new 1.5 ml eppendorf being careful to not take from bottom of tube.

d. Set the shearing parameters as follows:

- ✓ DNA volume = 300 µl
- ✓ Number of cycles = 20
- ✓ Speed code = 9 (check each Hydroshear shearing device for size)
- ✓ Wash cycles = 3x with HCl 0.2 M, 3x with NaOH 0.2 M, 8x with with H_2O mQ AF

e. After shearing, collect the sample into a 1.5 ml eppendorf and place on ice immediately.

In both case, analyze an aliquot of the fragmentated DNA by electrophoresis through a 1% agarose gel (SeaKem LE).

Concentrate the DNA sample to a final volume of 20-30 µl with a Microcon YM-100 (Amicon). Do repeated concentrations, typically 6; avoiding touching the membrane, at each step resuspend the DNA with H_2O mQ AF.

Determine the BAC concentration by UV spectrophotometry and by 1% agarose (SeaKem LE) gel electrophoresis.

II.3.3 Blunt end repair

The physical shearing methods produce a heterogeneous mixture of DNA fragments with blunt ends, 5′ overhangs and 3′ overhangs of varying lengths; fragments ends occur with or without phosphate residues. Various enzymatic treatments can be used to generate blunt ends, which are effective substrates for the ligation reaction to adapter. One approach for repairing fragments ends involves treatment with two enzymes. T4 DNA polymerase possesses a potent 3′→5′ exonuclease activity in addition to its 5′→3′ polymerase activity; thus, it can fill 5′ overhangs and digest 3′ overhangs. The Klenow fragment of *E. coli* DNA polymerase I is used to ensure that all of the 5′ overhangs are repaired; the Klenow fragment retains polymerization and 3′→5′ exonuclease activity, but has lost 5′→3′ exonuclease activity.

1. In a PCR tube set up the reaction following the conditions:
   - **a.** DNA final concentration = ~50 ng/µl
   - **b.** T4 DNA polimerase (Neb) = 1 U/µg of DNA
   - **c.** Klenow fragment = 1 U/µg of DNA
   - **d.** dNTPs = 0.25 M
2. Mix well and quick spin the tube.
3. Allow the reaction to proceed at room temperature (on bench top) for 30 minutes. After the 30 minutes, place the tube on ice and purify to remove the enzyme and dNTPs and change the buffer by binding to a purification column **PureLink PCR Purification Kit** (Invitrogen). I follow the user manual procedure; at the end I elute the sample with 100 µl of H_2O mQ AF (pre-worm at 65°C).
4. Concentrate the DNA by precipitation with 1/10 volumes of NaAc 3 M pH 5.2 and 2.5 volumes of EtOH abs. After a wash with EtOH 70%, resuspend the pellet with 15 µl of H_2O mQ AF (pre-warm at 65°C).
5. Determine the DNA concentration by UV spectrophotometry and by 1% agarose (SeaKem LE) gel electrophoresis.

II.3.4 Attaching adapters to protruding termini

To increase the efficiency of the ligation into vector step, adapters are ligated to the blunt termini of the DNA fragments. Adapters are phosphorylated at their 5′ termini and a four-bases 3′ protruding termini complementary to the *Bst*XI linearized vector.

1. Calculate the number of pmoles of ends:

$$\frac{\text{ng of DNA}}{\text{fragment medium length (bp)}} \times 3.04 = \text{pmoles of ends}$$

Set up the ligation reaction as follows:

- **a.** To achieve the maximum efficiency of ligation, set up the reaction in as small a volume as possible (20 µl)
- **b.** T4 DNA ligase (Neb) = 1 U/20 µl reaction
- **c.** Use a 50x pmoles of adapters respect to the calculated pmoles of DNA ends
- **d.** The adapters solution 50 µM must be defrost in ice.
- **e.** Incubate the reaction 0/N at 16°C and then inactivate by incubate the ligation mixture at 65°C for 20 minutes.

II.3.5 Purification from adapters dimers

During the O/N incubation, the ligation can occur between adapters and fragment ends, but also between the blunt ends of two adapters with the creation of adapters dimers that would compete for the cloning into the vector.

To remove adapters dimmers, I purify the ligation reaction with **PureLink PCR Purification Kit** (Invitrogen). I follow the user manual procedure for removal of primer; at the end I elute the sample with 50 µl of H_2O mQ AF (pre-worm at 65°C).

II.3.6 Size fractionation

Size fractionation of the DNA sample is performed to select the range of fragment sizes for cloning and to eliminate undesired small fragments (including residual adapters dimers) that would be preferentially cloned into the vector.

1. Place the end-repaired DNA sample in a three adjacent wells of a 0.8% low melting agarose (Invitrogen) gel. Also load 5 µl of 1-kb DNA ladder.

   Perform electrophoresis at 100 mA for 10 minutes and then at 50 mA for approximately 3 hours.

2. At the end of electrophoresis, the gel portions corresponding to the marker is cut and stained with a fresh ethidium-bromide solution (10 mg/ml EtBr in TAE 1X) for approximately 20 minutes.
3. Place the gel on a UV transilluminator to reveal the stained bands of the marker. Apply little cuts in correspondence of the bands of interest size.
4. Rebuild the gel. With a clean sharp blade, excise the gel slice containing the DNA sample in the size range, using the cuts as references. Minimize the amount of surrounding agarose excise with the desired DNA.
5. Place the gel slices in a 2 ml eppendorf and weight the gel slice.

II.3.7 Purification of DNA fragments from agarose gel

1. For the extraction of the size-fractionated DNA from low-melting point agarose gel, I use the **PureLink Quick Gel Extraction Kit** (Invitrogen). I follow the user manual; at the end I elute the sample with 100 µl of H_2O mQ AF (pre-worm at 65°C).
2. I perform a subsequent purification by spin dialysis through a micro-concentrator Microcon YM-100 (Amicon) following the manufacturer manual. To optimize the removal of adapters dimers and agarose contaminants, I do repeated concentrations, typically 6. Avoiding touching the membrane, at each step resuspend the DNA with H_2O mQ AF.

   At the end elute in as small as possible volume of H_2O mQ AF (20-25 µl).
3. Determine the BAC concentration by UV spectrophotometry; confirm DNA size and quantity by 1% agarose (SeaKem LE) gel electrophoresis.

II.3.8 Vector ligation

Size-selected DNA fragments are cloned into pUC18_*Bst*XI linearized vector. The optimal conditions for ligation depends from the methods used to generate the fragments and the chosen vector. One factor influencing the success of the ligation is the percentage of randomly sheared fragments possessing ends ligated to an adaptor. For experience, I have found that a molar ratio of vector DNA to DNA fragments included between 1:4 to 1:8 is an appropriate ratio when is used pUC18_*Bst*XI vector.

In a PCR tube on ice, prepare the ligation mixture

- **a.** To achieve the maximum efficiency of ligation, set up the reaction in as small a volume as possible (20 µl)
- **b.** T4 DNA ligase (Neb) = 1 U/20 µl reaction

c. Use 25 ng of linearized pUC19_*Bst*XI vector for reaction.

d. The vector solution must be defrosted in ice.

e. Incubate the reaction 0/N at 16°C. Then inactivate by incubate the ligation mixture at 65°C for 20 minutes. The reaction can stay O/N at 4°C if needed, but should be stores at -20°C for long-term storage.

▲ *Proper control experiments should be performed. For these controls, prepare the appropriate ligation above but 1) replace the DNA fragments used as inserts with H_2O mQ AF to determine the amount of undigested vector DNA, 2) replace the DNA fragments with a control insert (e.g. a EcoRV-digested fragments prior ligated to adapters) to determine the overall efficiency of the ligation.*

II.3.9 Preparation of competent DH10B *E. coli* using chemical treatments

For the preparation of chemical competent DH10B *E. coli* I use the protocol 'Calcium/Manganese-based (CCMB)' described by Hanahan (Hanahan D *et al.*, 1999). To induce a state of competence, this procedure uses calcium chloride and, in addition, manganese and potassium.

The typical competence obtained is 1-5 x 10^8 colonies per microgram of plasmid pUC19.

II.3.10 Transforming bacteria with ligated DNA

For transformation I follow the protocol provided by Hanahan for Calcium-Manganese-based transformation:

1. Remove DHIOB cells from -80°C freezer and place on ice until they thaw completely.
2. When cells are thawed, mix them by tapping gently and then aliquot 200 µl volumes of competent cell into chilled (at -20°C) Falcon.
3. Add DNA and mix by tapping gently. Incubate on ice for 30 minutes.

   ▲ *For each transformation I use 8 µl of a 20 µl ligation reaction. Then, for successive transformation of the same reaction, I use 4 µl of sample.*

   Include a transformation control as a transformation with 10 pg of supercoiled pUC19 DNA.
4. Heat-shock in a water bath at 42°C for 90 seconds. Place on ice for 5 minutes.
5. Add 800 µl of SOC medium. Incubate at 37°C in a shaking incubator at 230 rpm for 30 minutes.
6. Plate onto LB agar plate with ampicillin 50 µg/ ml and incubate O/N at 37°C. Then count the colonies and determine the efficiency of the library.

   ▲ *For each transformation, plate different volumes (typically 20 µl and 200 µl).*

7. For long-term storage, resuspend the transformed cell in a mixture of 80% cell - 20% sterile glycerol; aliquot (typically 400 µl) the cell suspension into a 2 ml eppendorf and store at -80°C.

II.3.11 Screening of recombinant colonies

The pUC19_*Bst*XI vector has lost the blue-white screening of colonies: the engineering made in the multiple cloning site have result in a not in-frame coding sequence of β-galactosidase.

The screening of colonies is made by Polymerase Chain Reaction (PCR) on 96 clones.

-40 M13 primer

5′ – GTTTTCCCAGTCACGAC -3′

-20 M13 Rev primer

5′ - GTGGAATTGTGAGCGGA – 3′

1. Make up the following PCR mix and aliquote 19 µl for each well of a 96-well plate:

	1x	**100x**
-40 M13 primer (10	0.4 µl	40 µl
-20 M13 Rev primer	0.4 µl	40 µl
dNTPs (10 mM)	0.4 µl	40 µl
$MgCl_2$ (50 mM,	0.6 µl	60 µl
10X buffer (Polymed)	2 µl	200 µl
Taq polimerase	0.1µl	10 µl
H_2O mQ AF	15.1 µl	1510
Total volume	19 µl	1900

2. Dispense 19 µl of the PCR mix into each well of the PCR plate. Keep on ice.
3. Using pipette tips, pick the colonies from the agar plate into a new 96-well plate with 10 µl of 1x buffer of Taq Polimerase (20 µl if the colonies are big) in each well. Mix tips before through them away.
4. Add 1 µl of the 1x buffer of Taq Polimerase containing the colonies into the corresponding well in the PCR 96-well plate. Mix and quick spin the plate.

5. Set up the following PCR program on the thermocycler:

95°C 5 min.

95°C 30 sec.
55°C 30 sec.
72°C 3 min.
} x 30 cycles

72°C 10 min.

4°C hold

6. Load 5 µl of sample onto a 1% agarose gel, using 1-kb DNA ladder. Run the gel for 30 minutes at 120 V.

7. Image the gel and check for inserts.

II.4 Protocol 4: Shotgun sequencing

II.4.1 Materials

Reagents and Solutions

Ampicillin 50 mg/ ml

Glycerol

Media

LB medium

LB agar medium

II.4.2 Prepare DNA sample for sequencing

The quality of DNA templates is key to good sequencing results: the presence of excessive amounts of template, protein contaminants or carbohydrates can coat the capillary walls of the sequencer resulting in poor data resolution. There are many good commercially available kits for DNA preparation. We use **Montage Plasmid$_{384}$ 384-well Plasmid Miniprep Clearing Plates Kit** (Millipore) for plasmid prepping because we find it lets a good signal strength and read lengths. The method allows a small-scale purification of plasmid DNA in 384-well plates, and typically results in sample DNA for standard bi-directional sequencing.

1. Plate onto LB agar plate with ampicillin 50 µg/ ml shotgun library transformation stocks stored at -80°C and incubate O/N at 37°C.
2. Manually pick colonies and inoculate each colony onto a separate well of a 384-well plate containing 50 µl of LB with ampicillin 50 µg/ ml. Grow the plate O/N; then seal the plate and store at -80°C with 20% glycerol.

   ▲ *The number of plates required for a 10x coverage is one plate every 50 Kb of DNA to be sequenced. The amount of plates must be increased for libraries with a high percentage of E. coli genomic DNA contamination and/or with high rates of vector and insertless.*
3. The plasmid DNA templates are purified with **Montage Plasmid$_{384}$ 384-well Plasmid Miniprep Clearing Plates Kit** (Millipore) using a robotics system. DNA is prepared from a 384-well plate replica of the plate stored at -80°C, using 150 µl of LB with ampicillin 50 µg/ ml.
4. The purify DNA can be stored at -20°C until it get in the sequencing pipeline.

II.4.3 DNA sequencing

The sequencing is performed by the Bio Molecular Research (BMR) centre of University of Padua. Sequencing reads are obtained by preparing two 384-well cycle sequencing reactions plates from each plasmid template DNA using the BigDye Terminator chemistry (ABI, Applied Biosystems) and standard -40M13 and -20M13 Rev primers, both flanking the pUC19_*Bst*XI cloning site (§ II.3.11).

II.5 Data management and sequencing assembly

Each BAC clone represents an individual project.

The shotgun sequences have been assembled by *phred/phrap/consed* (http://www.phrap.org/phredphrapconsed.html). The aim is to obtain a complete sequence of the BAC insert with less than 1/10000 sequence errors/bp.

II.5.1 Data management

An informatics pipeline has been developed and used for the data management, in a way to assure the traceability of each shotgun clone and to know the sequencing status of each BAC clone. The program *New*, developed by F. Levorin (CRIBI, University of Padua), allows to manage the tomato chromosome 12 sequencing project so as to:

- ✓ create the chromosome 12 tiling path,
- ✓ for each sequenced BAC clone, report the phase of the project (DNA extraction, shotgun library construction, sequencing, finishing),
- ✓ memorize the status and position of each 384-well plates for each BAC project.

By this, it is possible to maintain the data flow, trying to avoid manual errors, and every stage is recorded in the database.

II.5.2 The program *trim_blast.pl*

The reads generated by the sequencing core, are analyzed using the *trim_blast.pl* program, developed in our laboratory.

Initially this program checks the quality of all the electropherograms generated in the high throughput sequencing phase. Low quality reads are store in a directory called *N* and ruled out from the assembly phase.

As a successive step, *trim_blast.pl* analyzes the included sequences, in a temporary multifasta format, by a standard blastN search against the *E. coli* genome. All the reads having a high similarity ($<e^{-5}$) with *E. coli* are accounted as contaminants, stored in a directory named *COLI* and omitted from the project.

All the reads that have successfully passed the two checks are sent in the *chromat_dir* directory to be assembled.

II.5.3 *Phrap* and *PhredPhrap*: sequence assembly programs

For each BAC clone, to align the obtained reads, it has been used a program called *Phrap* (*phr*agment *a*ssembly *p*rogram), which was developed by Phil Green from the University of Washington (Ewing B et al., 1998).

Before starting, it is necessary to create three folds inside the work directory: *chromat_dir* (where the chromatograms are stored), *ph_dir* (where the sequence quality values will be stored) and *edit_dir* (where the output files of the assembly will be stored).

When *phredPhrap* script works:

1. The *phred* software reads DNA sequencing trace files, calls bases, and assigns a quality value to each called base: *phred* uses trace parameters to produce error probabilities associated to each called read base
2. Then runs *cross_match* to mask vectors sequence. It is used to compare the reads to a set of vector sequences (BAC cloning vector and pUC19_*Bst*XI plasmid vector) and produce vector-masked versions of the reads.
3. Finally *phrap* works. It is a program for assembling shotgun DNA sequence data: it aligns the reads depending on overlaps, to create contigs. The output file of *phrap* is saved in *edit_dir* directory as .ace file.

II.5.4 *Consed*

Consed is a tool for viewing, editing and finishing sequence assemblies created with *phrap* (Gordon D *et al.*, 1998). The consensus sequences determined by *Phrap* are viewed using the program *Consed*. *Consed* displays a window where the top line gives the contig sequence, and below it are the read sequences for the top strand (right-pointing arrows) and bottom strand (left-pointing arrows). *Phred* scores are denoted with upper-case (high quality) or lower-case (low quality) letters. More precise scoring is highlighted with a background colour gradient from white to black, white being high quality. Mismatches with the consensus are highlighted in red, and inserted bases are noted with an asterisk. *Consed* also allow displaying the trace of the reads.

The finishing capabilities include allowing the user to pick primers to use in additional sequencing reaction, and facilitating checking the accuracy of the assembly using digest and forward/reverse pair information.

II.5.5 Manual finishing

The initial assembly of the sequence reads generated from a BAC typically yields a certain number of contigs. In the finishing phase the aim is to produce a highly accurate consensus sequence resolving discontinuities between contigs (gaps), areas of low sequence quality, ambiguous bases in the consensus sequence and contig misassemblies. The process of finishing can, in principle, be divided into three stages: viewing the assembly and data for the purpose of deciding where additional data or editing are necessary or identifying other anomalies; obtaining additional read data; and editing to correct errors in the assembly or consensus sequence.

In accordance with the objective of reducing the expected number of errors below one every 10000 base pair, the strategies used has been:

- ✓ re-sequencing failed sequencing reactions if the reads fall into gaps,
- ✓ primer walking directly on shotgun clones (mini preparative of plasmid DNA) or PCR products,
- ✓ primer walking directly on BAC clone,
- ✓ sequencing with specific primers to improve sequence quality of low accuracy regions or to order contigs,
- ✓ modified sequencing chemistry to read through GC-rich regions.

II.5.6 Strategies to read through AT-polymeric regions

A significant amount of plasmid DNA templates submitted to the sequencing core contain poly(AT) polymer, that impedes *Taq* during sequencing reaction and causes slippage and deterioration of the sequence data.

To get through long poly(AT) tails, use a different thermal cycling condition:

Standard Cycle			**Poly(AT) Cycle**		
96°C	1 min		96°C	3 min	
96°C	10 sec	x 25cycles	96°C	10 sec	x 25 cycles
50°C	5 sec		58°C	4 min	
60°C	4 min				

The poly(AT) thermal cycling procedure is a two-steps cycle where the higher annealing temperature used allows for more primers specificity and less slippage in polymer tract. This procedure lifts the annealing step to keep these problem templates denatured longer.

Sequencing primers designed to anneal specifically near this regions are able to produce good-quality sequence when combined with the poly(AT) thermal cycle.

II.6 Protocol 5: BAC clone fingerprinting

One of the most important criterion for selecting a BAC is the evidence that the clone is authentic. The *Hin*dIII restriction digest fingerprinting is compared with that of the previous BAC clone: a BAC clone is selected for sequencing if some restriction fragments are present also in the *Hin*dIII restriction digest fingerprinting of the previous clone. When available a FPC (FingerPrinted Contigs) data, the comparison of the BAC DNA fingerprinting with the FPC fingerprinting is used to check for anomalies, such as internal deletion or chimeric insert, that make that BAC inappropriate for sequencing.
After final assembly, the BAC clone finished sequence is analyzed for its concordance with restriction enzyme digest-based fingerprints. This involves comparing the *in silico* restriction digests of the assembled sequence with the fingerprinting of the BAC clone DNA, with any discrepancy indicating the possibly presence of a sequence misassembly.

1. Prepare BAC DNA with a standard lysis methods using commercial kit **NucleoBond® PC** (Macherey-Nagel) as described in § I.1.1.
2. Digest 500-700 ng of DNA with 8 units of *Hin*dIII (NEB; 20 U/µl) or 8 units of *Bam*HI (NEB; 20 U/µl) in 15 µl volume. Incubate for 90 minutes at 37°C, then inactivate by heating at the proper temperature.
3. Analyze the digestion on a 0.8% agarose (SeaKem LE) gel; load 3.5 µl of 1-kb DNA ladder and 0.8 µl of 1-kb Extension DNA ladder.
   In a 4°C room, perform electrophoresis at 3 volt/cm for approximately 5 hours and then at 4-5 volt/cm for 1 hours.
4. At the end, stain the gel with a fresh ethidium-bromide solution (10 mg/ml EtBr in TAE 1X) for approximately 10 minutes.
5. Acquire and analyze the gel image.

II.7 Protocol 6: Tomato protoplast preparation and HMW DNA extraction

The plant material I use is *Solanum lycopersicum cv* cherry; the enzyme concentration and incubation times have to be tested and adjusted for each new cultivar. Additionally, I have found significant variation among different batches of the cell-wall degrading enzymes employed.

II.7.1 Materials

Reagents and Solutions

0.5 M EDTA pH 8.0

0.1 M MES pH 6.5

1X TE pH 8.0

K3

3.2 g/l	Gamborg's medium B5
750 mg/l	$CaCl_2$ x 2 H_2O
250 mg/l	NH_4NO_3
0.4 M	sucrose
1 mg/l	6-benzylaminopurine (BAP)
1 mg/l	α-naphtalenacetic acid (NAA)

Adjust the pH to 5.6 with KOH. Sterilize by filtration with 0.22 µm filter. Store at -20°C.

Proteinase K (20 mg/ml)

Dissolve the lyophilized powder (Gibco, Invitrogen) at a concentration of 20 mg/ml in sterile 50 mM Tris-HCl pH 8.0, 10 mM $CaCl_2$. Store -20 °C in 1 ml eppendorf.

Proteinase K digestion solution (for 20 plugs)

8.5 ml	EDTA 0.5 M pH 8.0
0.5 ml	N-Laurosylsarcosine 20 % (Sigma)
1 ml	proteinase K (20mg/ml)

Protoplast buffer

0.6 M	Mannitol
0.02 M	2[N-morpholino]-ethenesulfonic acid (MES)

Adjust the pH to 5.5 with KOH. Sterilize by filtration with 0.22 µm filter. Store at -20°C.

Enzymatic mix for protoplast

Protoplast buffer

0.75% Cellulase 'Onozuka R-10' from *Trichoderma viride* (Yakult; Tokyo, Japan)

0.25% Macerozima R-10 from *Rhizopus* sp (Yakult; Tokyo, Japan)

Sterilize by filtration with 0.22 µm filter. Store in falcon in dark at -20°C.

Prepare just before use.

W5

154 mM NaCl

5 mM KCl

125 mM $CaCl_2$ x 2 H_2O

5 mM glucose

Sterilize by filtration with 0.22 µm filter. Store at -20°C.

II.7.2 Protoplast preparation

Day one:

▲ *Work under sterile conditions, in a sterile laminar flow bench with sterile tool.*

1. Grow plants in a growth chamber and use them until they reach the flowering stage. Harvest approximately 2 gr of young leaves (2 to 4 cm in length) with a scissor.

2. If working with soil grown plants, wash the leaves

3. With a razor blade make many cuts from the midvein (1-2 mm from each other) at the surface in the underside without cutting through the whole leaf.

4. Place the leaves, underside down, on the surface of 7 ml of Enzymatic mix for protoplast in a 9 cm Petri dish, without wetting the upper side.

▲ *One Petri dish with 7 ml of solution can accommodate about 4-6 leaves.*

5. Incubate the Petri at room temperature in the dark for 15 hours with gentle shaking.

Day two:

▲ *At all stages when handling the protoplasts, pipette only with wide bores plastic pipette (cut the end to give a wider bore), use blue tip cut at the end and do not drop protoplasts into tubes from any height.*

6. After 15 hours check the release of protoplasts under a microscope (objective 20x or 40x). If there are not many protoplasts incubate for another hour.

7. Remove the enzyme mix with a 50 ml pipette and recover in a falcon.

▲ *The recovered enzyme mix contains protoplasts. You can follow steps 9, 11-15 to recover more protoplasts.*

8. Wash with drops of K3 solution (about 10 ml). Slightly agitate the dish for 20 minutes to detach protoplast from the leaves. Recover the solution with a 25 ml pipette in a falcon.
9. Filter the solution through a large pore size strainer. Strainer is placed few centimetres in height on a new Petri dish; just before using, wet the strainer with few drops of K3.
10. Repeat step 7 to maximize protoplast yield.
11. Move the protoplast resuspension into a 15 ml falcon and centrifuge at 500 g for 5 minutes at room temperature. Vital protoplasts float, dead cells are pelleted.

    ▲ *Do not use 50 ml falcon because it is difficult to recover protoplasts at the end. If you have large volumes, divide in small tubes.*
12. With a cut blue tip, recover the band corresponding to vital protoplast into a new 15 ml falcon.
13. Slowly add 4 volumes of W5, mix very gently and centrifuge at 60 g for 20 minutes at room temperature. Remove supernatant (protoplasts float in K3 and sink in W5).
14. Repeat washing with W5 a second time.
15. Resuspend in 1 ml of protoplast buffer.
16. Store on ice.
17. Count cells in an aliquot using Bürker chamber under a microscope (objective 20x).
18. Adjust the final concentration of approximately 5 x 10^5 protoplasts per 100 µl of protoplast buffer; leave on ice.

II.7.3 High Molecular Weight (HMW) DNA extraction in agarose block

1. Prepare 20 ml of 1.8% low-melting point agarose (InCert Agarose, Cambrex) in 0.1 M MES pH 6.5. Leave the gel molten in a 50°C water bath.
2. Cover the bottom of plug molds with tape and pre-chill at 4°C.
3. Put protoplasts suspension (approximately 5 x 10^5 protoplasts per 100 µl of protoplast buffer) in a 37°C water bath for 5 minutes.
4. Add an equal volume of 50°C molten 1.8% agarose gel, and mix gently but thoroughly with a pipette cut at the end.
5. Fill plug molds with approximately 90 µl of protoplasts/agarose mix.
6. Put the mold at 4°C for 30 minutes to allow blocks to solidify.
7. Eject plugs in a tube with a conical end. Add an adequate amount of proteinase K digestion solution so that there is a minimum of 250 µl per block.
8. Leave O/N at 50°C.

9. The following day, add fresh proteinase K and incubate for other four hours.

10. Remove the proteinase K solution and start washing the plugs. Perform 5 washes with 1X TE pH 8.0 for 1 hour each under gently shaking on a rotating wheel.

11. Store the plugs in sterile tube filled with 0.5 M EDTA pH 8.0. Store at 4°C.

II.8 Protocol 7: DNA Fibre-FISH on combed DNA molecules

II.8.1 Materials

Reagents and Solutions

YOYO-1

VECTASHIELD® Mounting Medium (Amersham)

glycogen 20 µg/µl (Invitrogen)

5 M NaCl

10 M NaOH

3 M NaAc pH 5.2

0.1 M MES pH 6.5

Ethanol, ethanol 70% and ethanol 90%

Ethidium bromide (10 mg/ml)

phenol-chloroform-isoamyl alcohol (25:24:1)

chloroform

1X PBS pH 7.4

50X TAE

1X TE pH 8.0

1 M Tris-HCl pH 7.6

II.8.2 Preparation of tomato $C_o t$-1 DNA

$C_o t$-1 DNA is enriched for repetitive DNA elements, high or moderate in copy number. In a in situ hybridization experiment, it can therefore be used to compete for repetitive sequence hybridization sites of the probe or the target, to eliminate not-specific binding. Commercial source of $C_o t$-1 DNA currently exist for at least three mammalian species (hamster, human and mouse) but are unavailable for most species, including plants.

Preparation of genomic DNA

1. Quantify genomic DNA by UV spectrophotometry and by 0.8% agarose (SeaKem LE) gel electrophoresis. Concentration is the key to determining C_ot-1 reannealing time.
2. Dilute the genomic DNA to a concentration between 100-500 ng/µl, using 5 M NaCl and H_2O mQ AF to a final concentration of 0.3 M NaCl.

DNA shearing

3. Aliquot the DNA sample in a 2 ml eppendorf and share DNA by sonication (§ II.3.2) to obtain fragments with a size ranging from 100 bp to 1000 bp. Determine the fragments size by electrophoresis in a 1% agarose gel.

4. Once sheared, put the sample on ice.

Reannealing

5. Calculate the time of reannealing using the formula

$$C_o t = 1 = \text{mol/l} \times T_s$$

where the initial concentration (C_ot) is calculated in moles of nucleotides per liter and times is in seconds. Assume an average molecular weight for a deoxynucleotide monophosphate to be 339 g/mol.

6. Based on the volume of DNA, calculate the amount of 10X S1 nuclease buffer (Promega) needed to equal 1X final working volume (including enzyme). Calculate this using an increased volume from that of the DNA volume.

7. Once all the calculation has been made, denature the DNA by placing the sample in a 95°C water bath for 10 minutes.

8. Remove the sample; cool it by swirling in ice for 10 seconds and place in a 65°C water bath. Start the reannealing period for the calculated $C_o t$-1 time.

S1 nuclease digestion

10. Following the time for reannealing, remove the sample, add the calculated 10X S1 buffer (Promega) and mix thoroughly. Add 1 units of S1 nuclease (Promega) per microgram of DNA and mix again. Immediately place the sample in a 37°C water bath for 10 minutes.

11. Stop the reaction by immediate phenol extraction using equal volumes of phenol-chloroform-isoamyl alcohol (25:24:1). Repeat the extraction twice. Then extract the supernatant with an equal volume of chloroform.

12. Precipitate the DNA O/N using 2.5 volumes of EtOH abs and 1/10 3 M NaAC pH 5.2. Wash the pellet with 70% ethanol and then resuspend it in approximately 100 – 200 µl of H_2O mQ AF.

13. Quantify the C_ot-1 DNA by UV spectrophotometry and by 1% agarose (SeaKem LE) gel electrophoresis.

14. Store DNA at -20°C in sub-aliquots until needed.

II.8.3 Preparation of target DNA for molecular combing

Solanum lycopersicum protoplasts has been embedded in 0.8 % low-melting agarose blocks (500000 cells /90 µl of block) (§ II.7.3). Genomic DNA is combed on glass cover slip coated with vinyl-silane. Each protoplasts plug is used to comb DNA molecules to check the DNA molecules size and stretch prior to FISH.

1. Remove one agarose block from EDTA storage buffer. In a 50 ml falcon, wash 5 times with 1X TE pH 8.0 on a rotating wheel (1 hour for each wash).

2. Transfer agarose block to a sterile 2 ml eppendorf and add 2 ml of 0.1 M MES pH 6.5.

3. Heat the block at 72°C for 30 minutes to melt the low-melting point agarose.

4. Put the eppendorf in a 42°C water bath for 15 minutes.

5. Add 4 units of β-agarase (NEB, 1 U/ µl) and gently invert the eppendorf to mix the sample. Incubate O/N at 42°C.

6. The following day, leave the eppendorf at room temperature for at least 1 hour to reduce the temperature.

7. Gently poor the DNA solution in a Teflon combing reservoir. Once the DNA solution is in the reservoir, prevent the evaporation when not combing by covering it with a parafilm.

▲ *The combing reservoir must be sterilized prior to use by boiling for 1 hour in ddH_2O.*

8. Store the molten DNA at room temperature or for long time at 4°C.

II.8.4 DNA combing

1. Place a silanised surface in the coverslip holder attached to the combing machine.

2. Let the machine lower the coverslip into the combing reservoir that contain the DNA sample. Wait 5 minutes to allow DNA to bind to the surface.

3. Release the ascent function of the combing machine to remove the coverslip. During removal the meniscus moving along the hydrophobic surface is combing the DNA.

4. As a combing quality check, visualize the DNA by staining with YOYO-1 and analyze under a fluorescence microscope:

- **a.** Let the machine lower the coverslip with the combed DNA into a reservoir containing YOYO-1. Wait 30 seconds to allow DNA to stain
- **b.** Remove the coverslip using the ascent function of the machine.
- **c.** Use superglue to stick the coverslip to a slide.
- **d.** Mount the coverslip with VECTASHIELD® Mounting Medium (Amersham) and analyze it.

5. Prepare coated coverslip following steps 1-3. Then use superglue to stick the coverslip to a slide.
6. Incubate the combed DNA at 60°C for 4 hours, with the coverslip surface facing up.
7. Allow to reduce temperature leaving the slide at room temperature for at least 1 hour. Then store at -20°C.

II.8.5 Labeling probes with biotin or digoxigenin

As probes I use the BAC clones. The DNA has been isolated by alkaline lysis and the purified by equilibrium centrifugation in CsCl-ethidium bromide continuous gradients (§ 1). Five hundreds nanograms of each probe has been labeled by nick-translation with biotin or digoxigenin: random primers (octamers) are annealed to the denatured DNA template and extended by Klenow fragment in the presence of biotin-16-dUTP or digoxigenin to produce sensitive labeled-DNA probes.

Probe labeling with biotin-16-dUTP:

All components are included in the **BioPrime DNA labeling** kit (Invitrogen). Thaw components and keep on ice.

1. In a 0.5 ml eppendorf on ice, to 500 ng DNA add 20 µl 2.5X Random Primers and H_2O mQ AF to a total volume of 44 µl.
2. Denature by heating for 8 min in a boiling water bath; immediately cool on ice for 5 minutes.
3. Centrifuge 15-30 sec.
4. On ice, add 5 µl 10X dNTP Mixture (includes biotin-16-dUTP) and 1 µl Klenow fragment.
5. Mix gently but thoroughly; then centrifuge 15-30 sec.
6. Incubate at 37°C O/N.

Probe labeling with digoxigenin:

I use 2.5X Random Primers and Klenow Fragment of the **BioPrime DNA labeling** kit (Invitrogen).

1. In a 0.5 ml eppendorf on ice, to 500 ng DNA add 20 µl 2.5X Random Primers and H_2O mQ AF to a final volume of 39 µl.
2. Denature by heating for 8 min in a boiling water bath; immediately cool on ice for 5 minutes.
3. Centrifuge 15-30 sec.

4. On ice, add 10 µl 5X DIG dNTPs Mix (0.35 mM DIG-11-UTP, 0.65 mM dTTP, 1 mM (dATP, dGTP, dCTP), 5 mM Tris-HCl pH 7.5, 0.5 mM Na_2EDTA pH 8.0) and 1 µl Klenow Fragment.
5. Mix gently but thoroughly; then centrifuge 15-30 sec.
6. Incubate at 37°C O/N.

In both cases, once completed analyze 5 µl of the labeled DNA by electrophoresis through a 1% agarose gel (SeaKem LE), using HindIII-digested λ DNA as ladder.

II.8.6 Probes hybridization on combed DNA

II.8.6.1 Slide preparation

1. Remove slides to be used from -20°C and leave at room temperature for at least 1 hour.
2. To denature combed DNA, place slides in a solution 0.05 M NaOH, 1 M NaCl for 15 minutes in dark and under gently shaking.
3. To neutralize NaOH, rinse very quickly three times in 0.01 M Tris-HCl pH 7.6 (4°C).
4. To help fix DNA, put slides through 70%, 90% and 100% ethanol (from -20°C) series for 3 minutes at each dilution. Keep in dark and under gently shaking.
5. Remove excess of ethanol with a drier kept at a certain distance from the coversplip. Use slides immediately once dried (go to § II.9.5.3 step3).

II.8.6.2 Hybridizations

1. Prepare the probes:
    a. In a 1.5 ml eppendorf add 4 µl of each probes, 10X of the total probe amount of tomato 1X C_ot-1 and 1 µl of Salmon Sperm (Invitrogen, 10 mg/ml). Bring to a final volume of 100 µl with H_2O mQ AF. Mix briefly.
    b. To precipitate DNA, add 1/10 of the volume of 3M NaAc pH 5.2, 2.5 volumes of EtOH abs and 1 µl of glycogen (Invitrogen, 20 µg/µl). Mix and put at -80°C for at least 1 hour.
    c. Centrifuge at 15000 rpm at 4°C for 30 minutes. Then remove the supernatant and wash with 70% EtOH.
    d. Resuspend very well the dried pellet with 20 µl of H3 buffer. Leave the sample at 37°C until it is used for hybridization.
2. Place the probe in a 80°C water bath for 10 minutes.

3. Take slide from step 5 § II.8.5.1. Place 20 µl of the probe mix on the surface and cover with a glass coverslip. Seal with silicon.
4. Place the slide in a wet chamber and incubate O/N at 37°C.
5. Remove the slide from the wet chamber and carefully remove the glass coverslip with tweezers.
6. In a fume hood, wash three times in 50% formamide/ 2X SSC pH 7 for 5 minutes each.
7. Wash three times in 2X SSC pH 7 for 3 minutes each.
8. Wash in 1X PBS pH 7.4 for 5 minutes with gentle shaking.

▲ *The day before use, remove 2X SSC pH 7, 1X PBS pH 7.4 and water from 4°C, formamide from -20°C and leave at room temperature. Perform all the washes in the dark.*

II.8.6.3 Detection with antibodies

The following steps are repeated for each layer. Exceptions are noted.

1. Immediately prior to use, thaw antibody aliquots on ice and spin briefly.
2. Dilute each antibody used for a layer in 30 µl of 1X Block AID. For the dilutions used see table II.8.1.
3. Mix antibody mix with a pipette and spin briefly.
4. Place the antibody mix on the surface and cover with a glass coverslip.
5. Put in a humid chamber and incubate at 37°C for 30 minutes.
6. Gently remove the glass coverslip by giving a shake to the slide.
7. Wash three times in 1X PBS pH 7.4 for 5 minutes each, in the dark and with gentle shaking.
8. Proceed with successive layer.
9. When finished washing the final layer, mount the slide with VECTASHIELD® Mounting Medium (Amersham) and seal with nail polish.

A: Biotin labeled probes: one colors detection:

Layers	Biot-16-dUTP (green)	
1	SAV488	1:50
2	anti-SAV biot rabbit	1:50
3	SAV488	1:50

B: Digoxigenin labelled probes: one color detection:

Layers	Digoxigenin (green)	
1	Enhancer kit 1° antibody	1:25
2	488 goat anti-mouse	1:50

C: Biotin and digoxigenin labelled probes: two colors detection:

Layers	Biot-16-dUTP (red)		Digoxigenin (green)	
1	extrAvidin-Cy3	1::200	DIG anti-mouse	1:25
2	anti-avidin-biot rabbit	1:50	488 goat anti-mouse	1:50
3	extrAvidin-Cy3	1:200		

Table 2.8.1 Detection with fluorescent antibodies condition.
Abbreviations: Biot = biotin labelled probe, Dig= digoxigenin labelled probe, SAV = streptavidin, 488 = Alexafluor 488.

II.9 Bioinformatics analysis of sequenced BAC clones

II.9.1 Gene prediction and annotation

In data October 2007, the sequences of 357 BAC clones were submitted in HTGS3 phase (ftp://ftp.sgn.cornell.edu/tomato_genome/bacs). For chromosome 4 and 12, the TPF information (ftp://ftp.sgn.cornell.edu/tomato_genome/tpf) were available and used to construct pseudomolecules by joining neighbouring clone sequences without redundancy in the overlapping regions.

JIGSAW (Allen JE *et al.*, 2005) was used to predict gene models from the following evidences:

- *ab initio* gene finder programs, SNAP and geneid (Korf I, 2004; Guigo R, 1998) trained with 112 *Solanaceae* full length cDNA,
- whole genome alignment with MUMmer (Kurtz S *et al.*, 2004) against the *Arabidopsis thaliana*, *Oryza sativa* and *Populus trichocarpa* genome,
- EST sequences from different plant collection (*Solanaceae* and eudicotyledons species) from dbEST (http://ncbi.nlm.nih.gov/dbEST/) and SGN database (http://sgn.cornell.edu) aligned with GMAP software (Wu TD *et al.*, 2005),
- Swiss-Prot and TREMBL proteins (Bairoch A *et al.*, 2005) alignment using BLAT (Kent WJ., 2002) and GeneWise (Birney E *et al.*, 2004).

Repetitive sequences, including transposable elements (TEs), were identified using the TIGR Solanaceae Repeats database (October 2007) (Ouyang S *et al.*, 2004) and removed from the predicted gene set.

The 5123 predicted genes are provisionally named with BAC(or pseudomolecula)-based names. The BAC (or pseudomolecula)-based names consist of the BAC name and a sequential number, that starts at 5′ end of the BAC (or pseudomolecula) sequence.

All predicted gene models were annotated using BLASTP search against Swiss-Prot and TrEMBL databases and the best-hit method (>e^{-5}) with >30% identities.

Gene family were identified by a BLASTP search of the predicted proteome against itself with a e-value threshold of <e^{-5}, identity >30% and a length match >65% of the entire protein length.

II.9.2 Phylogenetic analysis

Protein sequences showing similarity with a tomato gene family were identified by a local BLASTP search of the corresponding tomato predicted proteins against the protein data set of:

- *Arabidopsis thaliana* (ftp://ftp.arabidopsis.org/home/tair/home/tair/),

- *Medicago truncatula* (release version 2.0; http://www.medicago.org/genome/downloads/Mt2/),
- *Oryza sativa* (http://rgp.dna.affrc.go.jp/IRGSP/),
- *Populus trichocarpa* (release version 1.1; http://genome.jgi.-psf.org/Poptr1/Poptr1.home.html),
- *Vitis vinifera* (http://www.vitisgenome.it/).

The sequences were aligned using ClustalW (Thompson JD *et al.*, 1994) and the Phylogenetic tree was constructed using the maximum-likelihood method implemented in PhyML software (Guindon S *et al.*, 2003). Statistical support of the branches was tested with 1000 bootstrap resamples.

II.10 *Appendix*

A. Reagents and Solutions

10 mM dNTPs

dATP

dTTP

dCTP

dGTP

0.5 M EDTA pH 8.0

Add 186.12 g of EDTA to 900 ml of H_2O. Stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 with 10M NaOH (the salt of EDTA will not go into solution until the pH of the solution is adjusted to approximately 8.0). Sterilize by autoclaving and dispense into aliquots.

0.5 M MES pH 6.5

Dissolve 4.88 g of MES (Sigma) in 50 ml of H_2O. Adjust the pH to 6.5 with10 M NaOH. Sterilize by autoclaving; store at RT.

1X PBS

137 mM	NaCl
2.7 mM	KCl
10 mM	Na_2HPO_4
2 mM	KH_2PO_4

Dissolve 8 g of NaCl, 0.2 g of KCl, 1.44 g of Na_2HPO_4 x 2 H_2O (or 2.48 g of Na_2HPO_4 x 12 H_2O), 0.24 g of KH_2PO_4 in 800 ml of distilled H_2O. Adjust the pH 70 7.4 with HCl. Add H_2O to 1 liter. Sterilize by autoclaving.

50X TAE

2 M	Tris-HCl pH 8.0
0.05 M	EDTA pH 8.0
2 M	Glacial acetic acid

10X TE pH 8.0

100 mM	Tris-HCl pH 8.0
10 mM	EDTA pH 8.0

B. Media

LB

Per liter:

To 950 ml of deionized H_2O add:

10 g	tryptone
5 g	yeast extract
10 g	NaCl

Adjust the pH to 7.0 with 5 M NaOH. Adjust the volume of the solution to 1 liter with deionized H_2O. Sterilize by autoclaving.

SOB

Per liter:

To 950 ml of deionized H_2O add:

20 g	tryptone
5 g	yeast extract
0.5 g	NaCl
0.25 M	KCl

Adjust the pH to 7.0 with 5M NaOH. Adjust the volume of the solution to 1 liter with deionized H_2O. Sterilize by autoclaving.

SOC

SOC medium is identical to SOB medium, except that it contains 20 mM glucose and 20 mM solution of $MgCl_2$ – $MgSO_4$. After the SOB medium has been autoclaved, allow it to cool. Then add a proper volume of 1 M solution of glucose to obtain a final concentration of 20 mM, and of 1 M solution of $MgCl_2$ – $MgSO_4$.

Media containing agar

Prepare liquid media according to the recipes given above. Just before autoclaving add Agar 15 g/ l. Sterilize by autoclaving.

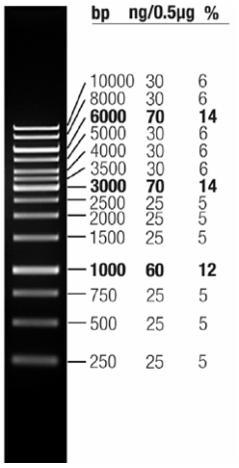
C. Bacterial cells

DH10B *E. coli* with genotype:

F- *mcr*A Δ (*mrr-hsd*RMS-*mcr*BC) ϕ80d*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K *rps*L *end*A1 *nup*G.

D. DNA labber

GeneRuler™ 1 kb DNA Ladders (Fermentas)



E. Abbreviations

abs = absolute

AF = autoclaved and filtered

bijoux = 6 ml sterile container

BSA = serum bovine albumin

dNTPs = dATP + dTTP + dCTP + dGTP

dATP = 2′ - deoxyadenosine 5′ - triphosphate

dTTP = 2′ – deoxythymidine 5′ - triphosphate

dCTP = 2′ – deoxycytidine 5′ - triphosphate

dGTP = 2′ - deoxyguanosine 5′ - triphosphate

EDTA = ethylendiammine-tetra-acetic acid

eppendorf = 1.5 ml or 2 ml polypropylene vial

EtBr = Ethidium bromide

EtOH = ethanol

EtOH abs = absolute ethanol

falcon = graded polypropylene test-tube

h = hour

H_2O mQ = purified water by Milli RO 15 (Millipore) or similar system

HCl = hydrochloric acid

kb = kilo base pair

KCl = potassium chloride

min = minutes

$MgCl_2$ = magnesium chloride

$MgSO_4$ = magnesium solfate

NaAc = sodium acetate

NaCl = sodium chloride

NaOH = sodium hydroxide

O.D. = optical density

O/N = over night

pb = base pair

petri = sterile slab for bacterial cultures

rpm = revolutions for minute

TAE = Tris-acetate EDTA

TE = Tris-EDTA

Tris = Tris-hydroximetylamino-methane

w/v = weight / volume

3.

Results and Discussion

CONTENTS:

III.1 BAC-by-BAC sequencing strategy

The BAC-based shotgun sequencing approach involves obtaining a collection of BAC clones covering the euchromatin region of the tomato genome and performing shotgun sequencing on each clone. The achievement in the key step of determining a reliable minimal tiling path depends on the data available.

For the Tomato Genome Project, the Cornell University (United States) has made available a *Hin*dIII BAC library based on *Solanum lycopersicum* Heinz 1706, covering the target with approximately 15 genome equivalents, and latter an *Mbo*I and an *Eco*RI BAC library (Budiman MA *et al.*, 2000). Furthermore a large number of tomato BAC-end sequences (BES) have been made available (Table 3.1.1).

BAC library	Average insert size (kb)	Genome coverage	BAC-ends sequences available
*Hin*dIII	117.5	15 X	144307
*Mbo*I	135	7 X	77141
*Eco*RI	n.d.	9 X	89132

Table 3.1.1. General statistics on the three tomato BAC libraries.

As a part of the project, a high density genetic map, based on an *S. lycopersicum* x *S. pennellii* F2 population (referred to as the Tomato-EXPEN 2000), has been completed and contains 2500 sequenced markers (1500 ESTs and 1000 AFLP).

A fingerprint contig physical map (FPC) of the *Hin*dIII BAC library has been constructed by the Arizona Genomics Institute (http://www.genome.arizona.edu/fpc/tomato/). Recently a Sanger Initiative was focused on the generation of additional fingerprint data from the *Mbo*I BAC library (ftp://ftp.sanger.ac.uk/pub/tomato/map) in order to integrate the available dataset.

By the Dutch group, a subset of BACs have been localized on pachytene chromosomes via FISH (Fluorescence in Situ Hybridization) to determine the chromosome localization and if the clone belongs to a euchromatin region.

The strategy adopted for selecting the BAC clones to be sequenced is based on the BAC-end sequencing, or sequence-tagged-connector (STC), approach (Batzoglou S *et al.*, 1999). The process starts by sequencing an initial collection of anchored clones (referred to as seed clones) and then 'walks' the genome by iteratively selecting minimally overlapping clones.

III.1.1 Seed BAC selection and validation

We started to work on the tomato chromosome 12 project by choosing suitable candidate BAC clones to be used as 'seeds'. A first selection was done by the Solanaceae Genome Network (SGN) by the overgo strategy, using probes for Conserved Orthologous Set (COS) markers of the genetic map Tomato-EXPEN 2000, hybridized against BAC clones densely arrayed on filters. The overgo selection led to a subset of 116 candidate BACs, possibly belonging to chromosome 12.

Since the actual correspondence of these BACs to chromosome 12 was not certain, a series of multiple controls was set up to verify the BACs and to choose the most suitable seeds:

1. sequence analysis of the PCR products amplified with primers designed on marker sequences;
2. IL (Introgression Lines) analysis;
3. comparison of *Hin*dIII BAC DNA digestion with FPC fragments, if available.

The validation of BAC clones via IL is performed by the University of Naples research group led by L. Frusciante. They use a population consisting of 50 *S. esculentum* introgression lines, each containing a single homozygous restriction fragment length polymorphism (RLFP)-defined chromosome segment, introduced from the species *S. pennellii*. The IL are nearly isogenic to the recipient genotype, and all the genetic variation that differentiate them can be associated with the introgressed fragment. The BAC-end sequences of a candidate seed clone are used to develop specific PCR primers for the screening of chromosome 12-specific IL lines. The sequence of the PCR product is then analyzed and, if a polymorphism is observed, the BAC clone can be mapped on a specific chromosome segment.

We verify and choose 32 seed BAC clones (Table 3.1.2). To optimize the project and due to the time required to complete a BAC clone, we process many BACs in parallel to simultaneously walk from many seed clones. Parallel sequencing, however, introduces a problem: the various walks may join with large overlaps. To avoid redundant sequencing we start from non-overlapping seed clones mapping at a significant distance on the genetic map.

Seed BAC	Marker	Chr12 localization (cM)	Round of 5′ extension	Round of 3′ extension
C12HBa0140M01	C2_At4g03280	12.50	1	1
C12HBa0026C13	cLPT-6-E09	14.00	1	1
C12HBa0260C13*		24.00		
C12HBa0206G16	T1487	24.00	1	0, stopped
C12Hba0075C18*		32.00		
C12HBa0163O04	T0028	33.00	2	1
C12HBa0032K07	T0989	36.00	1	1
C12HBa0244C09*		39.00		
C12HBa0146I19	T1667	39.00	2	0, stopped
C12HBa0180O10	cLET-8-k4	41.00	0, stopped	2
C12HBa0161H10	T1045	51.00	1	1
C12HBa0021L02	T1211	53.00	3	2
C12HBa0049J09	C2_At5g42740	54.50		
C12HBa0105J24	T1078	54.60		
C12HBa0024A16	cLET-5-M3	57.00		
C12HBa0047D08	P62	57.20		
C12HBa0081D06	TG406	57.40		
C12HBa0062P09	cLET-8-E15	57.70		
C12HBa0009J11	T1185	57.80	1	1
C12HBa0150C12	SSR20	58.20		
C12HBa0144B17	CT189	59.00		
C12HBa0059A05*	SSR124	60.00		
C12HBa0266F15	cLET-8-G15	60.00		
C12HBa0077H15	T1947	65.00		
C12HBa0165B12	TG394	68.00	1	1
C12HBa0302G23	G367	68.50		
C12HBa0193C03	T1266	71.00	1	1
C12HBa0326K10*	TG468	85.70	1	1
C12HBa0115G22	T1676	86.00	1	1
C12HBa0093P12	T0882	97.00	0, stopped	2
C12HBa0183M06	T0770	101.00		
C12HBa0147G13	T1504	118.00	2	1

Table 3.1.2. Status of the BAC-by-BAC extension (January 2008). The table indicates the localization on the genetic map (marker and cM) and the extension status for each seed BAC. The extension has been exhausted for 4 seed points.

III.1.2 Extension BAC

Once a BAC clone is sequenced, it becomes a 'nucleation' point from which walking out in the genome. At each step, a minimally overlapping clone is identified by comparing the database of BAC-end sequences with the complete nucleation sequence, to find the BAC-end sequence that lies closest to the growing end and point outward.

A key step in the BAC-by-BAC sequencing is the identification of reliable neighbouring BACs. One important issue to be considered is the possible presence of repeats, that my compromise the success of the project. A BAC-end sequence entirely contained within a repeat element can connect non-contiguous regions of the genome, leading to misalignment of BACs and possible 'jumps' along the genome. The analysis of repeats can be performed using RepeatMasker (http://www.repeatmasker.org) or similar tools able to identify known repeats. However, when genomes are not yet extensively studied, as the tomato genome, the repeated regions are not well characterized and their direct identification is impossible.

With the progress of the project, the tomato genetic map has proved not to be robust, the physical map not to be detailed and the BAC-end database revealed a high percentage of low-quality reads. As a consequence, during the International meetings that accompanied the project, a real high risk of false positive overlaps emerged from the observation of all the participating groups.

Therefore, a challenge for our group was to develop an efficient strategy to make use of the BAC-end sequences for selecting reliable minimal overlapping clones. We have developed an informatics pipeline called PABS (Platform Assisted BAC-by-BAC Sequencing) that we made freely available to the community at our web site (http://tomato.cribi.unipd.it/files/bioinformatics.html) (Todesco S. *et al.*, 2008). The fulfilment of this tool came from the cooperation between different expertises, of genomic (Dr A. Vezzi and I, CRIBI, University of Padua), of genome repeat analysis (D. Campagna, CRIBI, University of Padua) and of informatics (F. Levorin, CRIBI, University of Padua).

PABS has two main functions: 1) PABS-Select, to choose suitable overlapping clones for the sequencing walk; 2) PABS-Validate to verify whether a BAC under analysis is actually overlapping the preceding BAC (Fig. 3.1.1).

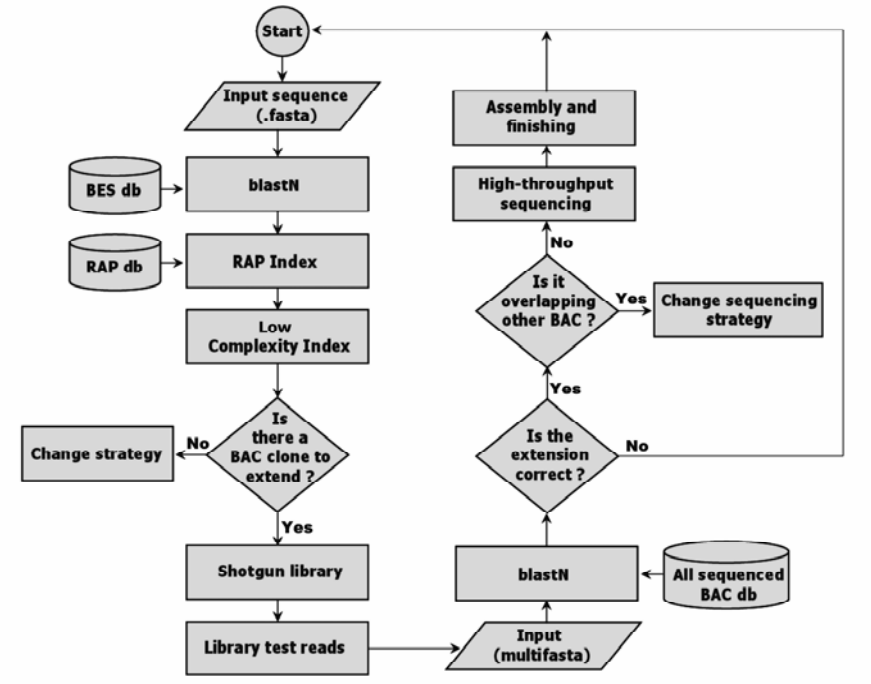


Fig. 3.1.1. Schematic overview of the dataflow used in PABS (Todesco S. *et al.*, 2008).

PABS-Select uses blastN search for high-throughput screening of BES database. Furthermore, the approach we have taken does not depend on the full closure of the initial BAC, reducing the bottleneck in waiting for the sequence of clones to be finished. In practice, PABS-Select takes as 'input sequence' the initial BAC, either the complete sequence or the contigs containing the end under investigation, and it returns a graphical representation of the position and orientation of the BESs (represented as oriented arrows) overlapping the input sequence (Fig 3.1.2 A).

An innovative feature of PABS is its ability to integrate the BES analysis with the presence of repetitive sequences. In particular, PABS identifies repeated regions with RAP and calculates the Low Complexity Index as one minus the Linguistic Complexity Index. The RAP index had been developed in our laboratory by D. Campagna; it gives an estimate of the 'repetitiveness' of a DNA region. It is calculated for each position of the input sequence by means of a *de novo* analysis that does not require any previous knowledge about repeats. PABS displays the results of blastN and RAP, thus allowing a more reliable selection of adjacent clones. We address the choice to BACs with a suitable overlap to the initial BAC and with the aligned BES positioned in a low-repeat region. Yet in some cases we have selected clones with a significant overlap (>20 kb) but with a lower risk of false overlap (low RAP index). At the same time, the BAC clones C12HBa0090D09 and C12Mbo0126D24 have an overlap, respectively, of 624 bp and 1020 bp with the

corresponding seed: it is possible to identify BACs for walking having only a few hundred base pair overlap.

As mentioned before the BAC-end database reveals a high percentage of low-quality reads. To make the selection easier and faster, PABS allows a direct visualization of the BES electropherogram aligned with the input sequence (Fig 3.1.2 B). In this way we can quickly evaluate sequences of poor quality that may be the cause of misleading blastN results. In addition by an automated procedure, we collects from SGN databases and summarizes all the available information on the candidate BACs (insert length, genetic markers, FISH data, sequencing status) to optimize the selection for the extension.

In our strategy, at each extension round, we try to identify two types of BAC clones:

1. candidates clones with a minimal overlap to be used to extend the walk on the genome,
2. one or more clones with a high overlap; we use this clones as bridge between the initial BAC and candidate extension clones, to verify candidate clones.

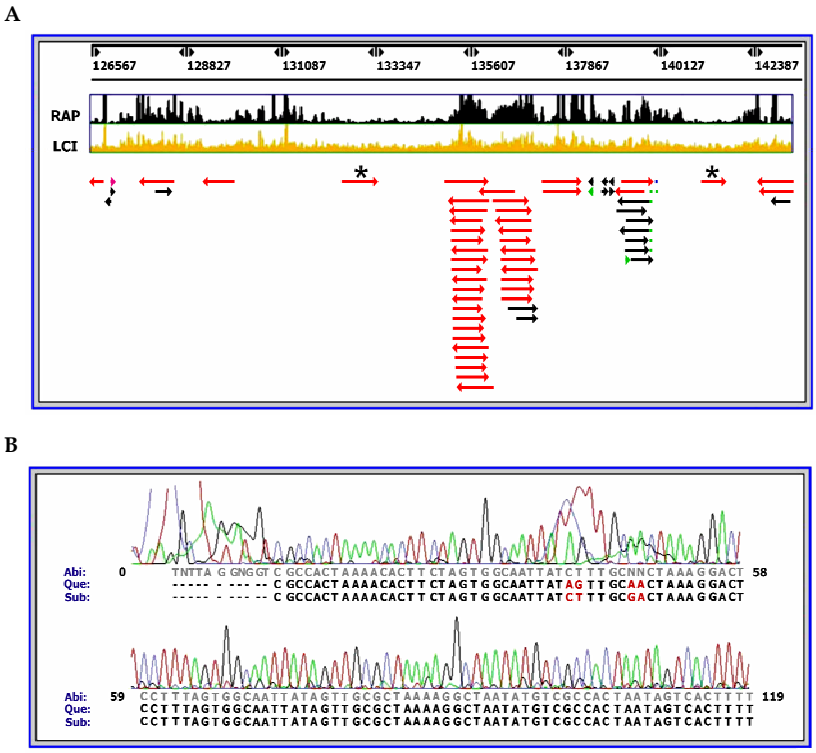


Fig 3.1.2. Screenshot from PABS-Select. **A)** After uploading the input sequence (typically the sequence of the BAC or the contig containing the end to be extended) the application returns a graphical representation of the sequence, including the RAP index (reflecting the repetitiveness of a region) and the Low Complexity Index (LCI, indicating the presence of low complexity regions). To simplify the figure, only the terminal 16 kbp of a 143 kbp BAC insert are shown. The entire database of BAC-end sequences (BES) is preloaded on the system, thus allowing an automatic blastN search to align on the BAC all the matching BES, represented by arrows in the figure. This gives an immediate view of the possible overlapping BACs, the arrows pointing to the direction of the overlap. The final aim is to find, at each end of the input sequence, a suitable overlapping BAC; therefore, the best candidates will be those

corresponding to BES aligned near the end of the original BAC, oriented towards the end and positioned in regions with low RAP and LCI indexes. The asterisks indicate two suitable candidates. By clicking on an arrow, the electropherogram aligned to the input sequence is displayed, as partially shown in **(B)**. The query sequence (Que) corresponds to the BAC taken as input, while the subject (Sub) is the aligned BES as stored in the database. Moreover, the 'Abi' sequence refers to the same BES, generated with the standard Applied Biosystems base caller. This allows a accurate inspection of any discrepancy between the two aligned sequences; for instance, the mismatching bases between query and subject (red coloured) would indicate considerably different sequences, but the analysis of the elecrtopherogram shows a likely perfect match of the two sequences. (Todesco S. *et al*., 2008).

In practical terms, we digest the initial clone, the candidates and the bridging clones with *Hin*dIII to produce restriction maps (generally called fingerprints) on agarose gels. Fingerprintings of the initial clone, of the candidate clones for extension and of the bridging clones are then compared to ensure internal consistency. Restriction fragment sizes are used to select the most suitable clones for overlap (minimal) and insert size (bigger). Furthermore the insert size information (available at SGN just for *Hin*dIII BAC clones) has several uses, including the estimation of the number of sequencing reactions required for a BAC in the shotgun phase and provide the ability to check the assembly of finished shotgun BACs.

Before start sequencing, the localization on chromosome 12 of the selected BAC clone/s must be confirmed using IL (this is done by the University of Naples research group) (Fig 3.1.3).

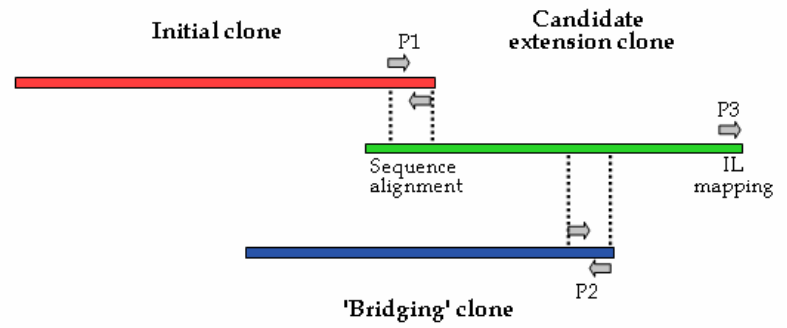


Fig 3.1.3. Strategy for IL validation of candidate extending BAC selected with PABS. Specific primer pairs are designed on initial BAC-end (P1 pair), on 'bridging' BAC-end (P2 pair), and on candidate extending BAC-end (P3 pair). The sequences of the fragments generated with P1 and P2 primer pairs on candidate extension BAC, are aligned on the initial BAC and on the 'bridging' BAC sequences to confirm both overlapping and direction of extension. IL mapping through specific IL-12 lines confirms the position of the selected extending BAC on chromosome 12.

The experience acquired with the progress of the project brought us to the conclusion that despite the lacking of a accurate genetic and physical maps, the multiple constrains of sequence similarity, orientation, low RAP index, fingerprinting and IL analysis provide a very strong evidence for the correctness of the overlap. Nevertheless, to further validate the selection, we have designed PABS-Validate. Typically, before starting the high-throughput sequencing phase, we test the shotgun library produced from the selected BAC by generating 96 random sequencing reads. This first set of 96 shotgun sequences are submitted as a multifasta file to PABS-Validate and analyzed using blastN against three databases:

1. the initial BAC,
2. the finished BACs (i.e. all the finished BACs of the Tomato Genome Project),
3. the partially sequenced BACs (i.e. the BACs under sequencing).

We can make three types of controls:

1. some of the reads should fall into the overlapping region of the initial BAC, thus confirming a correct extension;
2. no reads should significantly match other sequenced BACs belonging to different genomic regions, because this would indicate a possible jump to another region;
3. as an exception to the previous point, when several extensions are carried out simultaneously from different seeds, we expect that eventually the different walks could merge; therefore we must also consider this event and the consequent possibility to work out the extent of the overlap at the two ends of a bridging BAC.

In this way, we are able to make multiple validations at the beginning of the shotgun sequencing phase of each BAC, trying to minimize the possibility of mistakes and optimize the merging of overlapping BACs.

III.1.3 Chromosome 12 sequencing status

At date, 70 BAC clones are in diverse sequencing phases, and 25 of them are already available on public databases. For 16 seed BACs at least one round of extension was performed; in some cases two or three rounds of extension were performed allowing overlapping BACs to merge in sequencing island of more than 300 kb. Despite a non-uniform distribution of anchored seed BACs on chromosome 12, small contigs of overlapping BACs started to merge. Progress can be viewed through the development of the TPF and AGP files, available from SGN repository (ftp://ftp.sgn.cornell.edu/tomato_genome). The TPF is an ordered list of sequenced BAC

clones along the chromosome; the AGP file describes how the BACs can be assembled to obtain a non-redundant, contiguous sequence.

The limiting step for a soon completion of the sequencing project is the selection of the tiling path. The available genetic map has not the resolution and density useful to identify new candidate seed BACs and the BAC-end database does not have accuracy in terms of quality reads necessary for a certain identification of candidate extension clones.

III.2 BAC-by-BAC shotgun sequencing

The first aim of the Tomato Genome Project is the sequencing of the 220 Mb of the tomato euchromatin by a BAC-by-BAC approach. It involves obtaining a collection of large-insert clones (BAC) covering the euchromatin regions and performing shotgun sequencing on each individual clone.

As the project lacks a complete and reliable physical map covering the genome, it proceeds directly to sequencing. One starts by sequencing an initial collection of non-overlapping BACs (called seed BACs) and then 'walks' by iteratively selecting minimally overlapping clones.

The BAC-by-BAC strategy makes possible for many laboratories to cooperate in the effort, allowing international collaboration between large genome centre and small groups. The chromosome 12 has been allotted to an Italian team constituted by different teams of the University of Naples, Padua and the ENEA Institute of Rome; the sequencing is mainly carried out by the G. Valle research group, University of Padua.

In designing the DNA-sequencing process, we focused on developing a system that could be implemented in a robust and reproducible manner and monitored effectively. The process has been designed in a modular form (Fig. 3.2.1), with five principal modules able to operate independently:

4. BAC (seed clones as well as extending clones) selection and validation (§ III.1);
5. BAC DNA preparation;
6. subclone library construction, transformation, plating and colony picking;
7. plasmid DNA template preparation and dideoxy sequencing reaction;
8. finishing phase, assembly authentication and release.

A team of 6 people was trained and structured into the five modules. A central laboratory information management system (LIMS) has been developed to track all sample, both BAC clones and shotgun plasmid subclones (Fig 3.2.2).

Critical to the success of the project is the continuous monitoring and validation of all procedures and software. The tomato chromosome 12 has been a challenging project for our laboratory and has required the refinement of existing strategies as well the development of special approaches. During the early sequencing efforts several difficulties came to light and different improvements have been made, as described below.

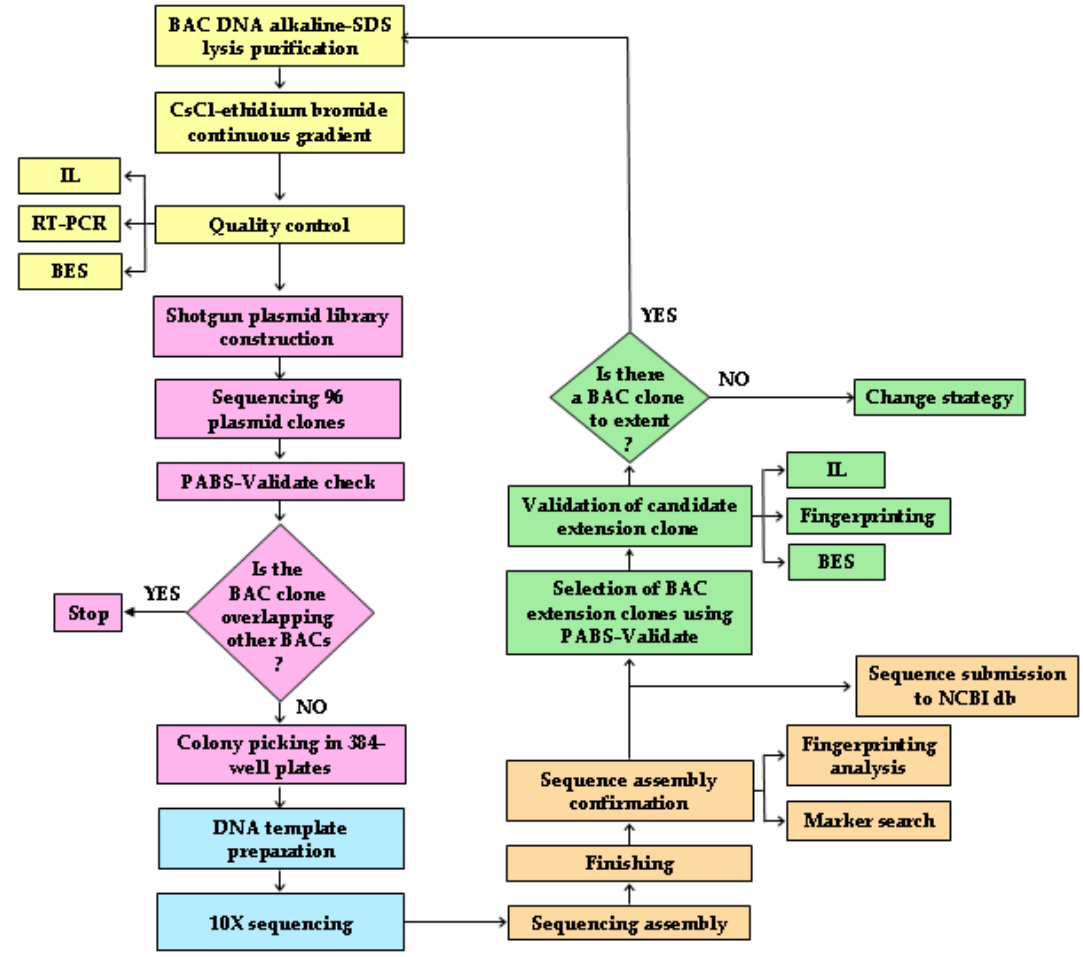


Fig. 3.2.1 Schematic overview of the sequencing pipeline that we had specifically developed to carry out the chromosome 12 sequencing project. The colours point out the modular structure of the process and mark the five modules in which the sequencing work has been distributed: **1.** BACs selection and validation (yellow), **2.** BAC DNA preparation (orange), **3.** subclone library construction (violet), **4.** plasmid DNA template preparation and sequencing (blue), and **5.** finishing and sequence release (green).

	Funding	Project	Extension	overlapping	cM	marker	real_size	estimated_size	declaration	seq_status	status_unipd
SP6 extension	Agronanotech		C12HBa0061F16	19870				88561	being sequenced	in progr	sequencing
Seed BAC	Agronanotech	C12HBa0140M01			12.50	C2_At4g03280	112743	86988	being sequenced	complete	submitted
T7 extension	Agronanotech		C12HBa0221M09	21940				101414	being sequenced	in progr	finishing HTGS 2
SP6 extension	Agronanotech		C12HBa0073O10	9100					being sequenced	in progr	finishing HTGS 1
Seed BAC	Agronanotech	C12HBa0026C13			14.00	cLPT-6-E09	174433	151214	being sequenced	complete	submitted
T7 extension	Agronanotech		C12HBa0090D09	624				41865	being sequenced	complete	submitted

Fig. 3.2.2 Screenshot of the LIMS web-interface implemented to track all BACs selected for the tiling path of chromosome 12. For each BAC, we summarize the main information, as clone features (insert length, marker) and sequencing status. At the same time the relationship between clones is graphically represented: white rows refers to seed clones and grey rows to the extension clones. For each seed, its SP6-extension is the BAC above while its T7-extension is the BAC below.

III.2.1 BAC DNA preparation

Central to the BAC-by-BAC sequencing process is the preparation of high-quality shotgun plasmid libraries in a consistent manner to all the selected BAC clones. High-quality libraries must have an equal representation of all parts of the BAC clones, a small number of clones without inserts, and no contamination from *Escherichia coli* genomic DNA.

The initial method we used to purify the BAC DNA was a standard alkaline-SDS lysis followed by purification from impurities such as RNA, protein, carbohydrates, and small metabolites using DEAE anion-exchange resin, commercially available. Following this procedure, the DNA of the first selected seed clone, C12HBa0032K07, was prepared. To make shotgun library, DNA was randomly sheared, end-polished, size selected by gel electrophoresis; the fragments were inserted into *Eco*RV-linearized pZErO plasmid vector. Based on the estimated insert size, I generated a 10-fold coverage in paired end-sequences. After quality and vector trimming, each trimmed sequence was screened for matches with contaminant *E. coli* genomic DNA.

BAC clone	n° of 384 plates	total n° of reads	n° of reads matching *E. coli*	n° of BAC specific reads
C12HBa0032K07	3	2304	833 (36.15%)	1132

Table 3.2.1. General statistics on the shotgun sequencing of BAC C12HBa0032K07. BAC DNA was prepared with a standard alkaline-SDS lysis using DEAE anion-exchange resin, commercially available. Fragments of 2-3 kb were cloned into *Eco*RV-linearized pZErO plasmid vector and a 10-fold sequencing coverage in paired end was performed.

The results reported in table 3.2.1 underline an unexpected high percentage (36.15%) of reads matching *E. coli* genomic DNA. At the same time, this data are consistent with the ones reported from the other sequencing groups and showing that BAC DNA prepared by traditional methods contains *E. coli* genomic DNA contamination up to 50%. This high percentage of contamination leads to significant rise of the number of sequencing reads to obtain the desired clone coverage, and, as a consequence, to an increase of the total sequencing cost and labor required.

During the alkaline lysis preparation, the contaminating chromosomal DNA is generally nicked and sheared. Based on this principle, I tried two different approaches to 'clean-up' the BAC DNA. Initially I performed digestion with Plasmid-Safe ATP-Dependent DNase (Epicentre). Plasmid-Safe DNase is an exonuclease that selectively hydrolyses linear double-stranded (ds) DNA to deoxinucleotides and with a lower efficiency linear and

closed-circular single-stranded DNAs; it does not affect closed circular supercoiled or nicked circular dsDNAs.

Due to high residual *E. coli* contamination even after Plasmid-Safe treatment, I optimized a protocol for BAC purification with a CsCl-ethidium bromide continuous gradient. Although it is much more time-consuming respect to Plasmid-Safe ATP-Dependent DNase treatment, this protocol yields extremely pure BAC DNA (table 3.2.2). I developed a specific procedure in order to remove firstly *E. coli* genomic DNA and then residual cesium-chloride salt that I proved to inhibit successive enzymatic reactions.

BAC clone library	SDS-alkaline lysis	Plasmid-Safe	CsCl gradient
C12HBa0026C13	67.97%	56.27%	8.17% (9.10%)
C12HBa0140M01	60.07%	47.65%	2.32% (3.71%)
C12HBa0146I109	62.76%	52.40%	3.61% (1.39%)

Table 3.2.2 Comparison of the quality of BAC DNA preparation in terms of percentage of *E. coli* DNA contamination. Using RT-PCR the *E. coli* DNA contamination was estimated after alkaline lysis extraction, Plamid-Safe treatment and CsCl-ethidium bromide continuous gradient purification. The values reported refer to the RT-PCR measurements. For the sample purificated by Cs-Cl gradient the shotgun library was constructed making able a comparison with the percentage of sequenced clones matching with *E. coli* (value reported between parenthesis).

Moreover, we developed a procedure based on Real Time PCR (RT-PCR) technique useful to estimate the BAC DNA in respect to the contamination of *E. coli* DNA in a DNA sample preparation. We had designed two different pair of primers:

1. primers to amplify a region present in all the three BAC vectors used, pBeloBAC11 (*Hin*dIII BAC library), pECBAC1 (*Mbo*I BAC library) and pIndigoBAC-5 (*Eco*RI BAC library);
2. primers to amplify the 16S rDNA of *E. coli.*

We introduced this step in our sequencing pipeline to be able to estimate the quality of a BAC DNA preparation before the shotgun library construction. We tested the reliability of our procedure comparing the RT-PCR data with the sequencing results (as number of shotgun reads matching *E. coli* genome for each BAC sequencing project). As reported in table 3 (§ Appendix), the sequencing data confirms the RT-PCR assessments. As a result, we succeeded in preparing high quality BAC DNA and using RT-PCR we can optimize the sequencing efforts, avoiding redundant sequencing of *E. coli*.

III.2.2 Subclone library construction

The success and efficiency of a large-scale DNA sequencing project are highly dependent on the quality of the subclones libraries. The libraries need to yield a sufficient number of clones for sequencing, as well as must provide an even coverage of the target DNA. In addition, to ensure efficient sequencing and sequence assembly, the background of clones without inserts and with chimeric inserts needs to be as low as possible.

We subjected each selected BAC to random fragmentation by physical shearing methods. After enzymatic repair of broken ends and size fractionation, the DNA fragments in a defined size range (2-3 kb) are recovered and subcloned into a plasmid vector. The main advantages of plasmid subclones are that the resulting templates can be used for deriving sequence reads from both ends of the subcloned fragments (at the cost of purify only one template) and that the pair of sequence reads from each subclone ('read pairs' or 'mate pairs') can be used to facilitate and/or assess the subsequent sequence assembly. The current sequencing protocols are based on the dideoxy sequencing methods using capillary-based instruments, and typically provide ~600-800 bases of high quality sequence per read. The average insert length is 2-3 kb in such a way to maximize the usage of each subclone, avoiding at the same time any sequence redundancy between the read pairs.

A widely used method for shotgun library construction is blunt-end cloning, where end-repaired inserts are directly ligated to linearized vector. At the beginning of the project we used a protocol that exploits an in-house *Eco*RV-linearized pZErO vector. The main disadvantage was a high background of circularized vector and a low efficiency of ligation.

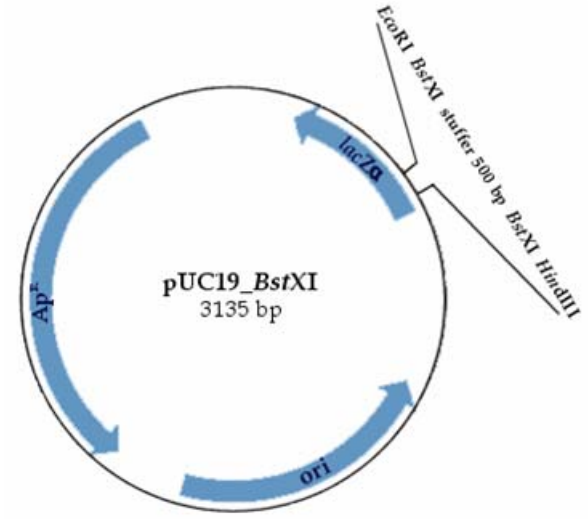
To exploit the higher efficiency of sticky-ended ligation, we used two different protocols, both relying on TA cloning strategy and requiring the addition of a 3' A overhangs (A-taling) to end-repaired fragments using dATP and Taq DNA Polymerase. The first method utilized the ligation into a 3'-T overhangs pGem-T vector (Promega); the second used the ligation into pCR4-TOPO vector (Invitrogen), which is provided with a single 3' thymidine (T) overhangs and with a topoisomerase covalently bound to the vector. Both the protocols resulted in a low number of recombinant clones and in a high background of re-ligated non-T-tailed vector possibly due to the degradation of the single T-overhangs allowing blunt-ended ligation of the vector.

In order to reduce the background and increase cloning efficiency, I developed a protocol using oligonucleotide adaptor ligated to the inserts. In this method, randomly sheared and end-repaired fragments are ligated to oligonucleotide adaptors creating 4-base overhangs (CACA). I use 5' phosphorylated oligonucleotides, to ensure high efficient

ligation of adaptor to insert. The vector is prepared from a modified pUC19 vector, by *Eco*RI/*Hin*dIII digestion followed by ligation to a 500 bp DNA fragment carrying at both extremities a *Bst*XI site (Fig 3.2.3) (for details § 2.2). The digestion of pUC19_*Bst*XI vector with *Bst*XI generates two cohesive ends (TGTG) which are not complementary to each other, thus avoiding self-ligation of the vector, but are compatible with ligation with the cohesive ends of the inserts. A critical aspect in the development of this protocol was the definition of a set of procedures and good senses in order to remove adaptor dimers that compete with the insert in the ligation into the vector.

This protocol is robust and shows a higher yield of clones compared to previous protocols. With this method the libraries constructed have a number of clones that exceed the requirements for the completion of BAC-sequencing projects. Moreover the libraries are of high quality with negligible levels of chimeric clones, and the frequency of clones containing insert has been increased up to 90%. Minimal is the background, represented by clones containing contaminating vector that was undigested by *Bst*XI and clones with adaptor dimers as insert. With the previous methods, we needed to examine the DNA using agarose gel electrophoresis to exclude background clones. With the low background of this method, this time-consuming step can be avoided. The vector can be prepared in batches, sufficient for a discrete number of libraries.

A



B

```
                40 M13 Forward priming site   M13 Forward priming site   EcoRI      BstXI
358 GTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTCCAAGTGTATGGTAGAAAGATTCA 423

424 TCTATTTCTTTCTACCATACTATCTATCTATTAGAATACTGCCGATTCTAGTCCACTCATTTCATTTAA 492

493 GACATGAAATTGGAATCTTTTTCATTTTATTTCGTCAATTTTGGCTAAGAACTCAGAAGTCAAGTTTC 560

561 ATTCAAATTAGTTAATAATTAATCGTTTTGACTGACTGTTTTTACATAAATGATAAGTAGAAAAGCGG 628

                                  stuffer 500 bp
629 TAGGAACTAGAATAAATAGTGCAGTAGCAATAAATGCAAGAATATTTACTTCCATAATCTCATCGGT 695

696 TTTTTACTTCGCAATAACTCGGGATTTAATCCCATAGAGATGATAAATCTTTGGCCTGTAAATTCAAT 763

764 GAATGAATATTACCTCTCGATGATCTTGAATCAGATCAATATCATGAATAACAATATCTGAACTATCA 831

                                                          BstXI      HindIII
832 AATAAATTCGTCGTCGAGAATTGAATAGTATAACATAGGAAGTTCTTTTATCCATACACTTGGAAGCT 899

                M13 Reverse priming site                -20 M13 Reverse priming site
900 TGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCAC 960
```

Fig. 3.2.3 pUC19_*Bst*XI vector **A)** Vector circle map. **B)** Detail of the cloning site and surrounding sequences. Restriction sites are marked: the digestion of pUC19_*Bst*XI vector with *Bst*XI releases the linearized vector with two cohesive ends (TGTG) not complementary to each other and the 500bp stuffer. The stuffer fragment corresponds to the *Solanum tuberosum* cultivar Desiree chloroplast sequence (DQ386163) at 30316-30798 bp.

III.2.3 Random shotgun phase

In the initial sequencing phase subclones are picked at random and sequence reads are derived from the universal priming sites locating at both sides of the cloned insert.

For most of the BAC projects, with the random sequencing phase we have generated a ten-fold average redundancy (coverage) of the consensus sequence (the number of times the consensus sequence is represented by sequence reads). The main reason for a ten-fold redundancy is that lower coverage could leave assemblies more incomplete thus demanding a more time-consuming finishing phase.

To reach a ten-fold coverage, the calculation of the number of clones to pick up involves four factors: 1) the estimated BAC insert size; 2) the average insert size of the contributing subclones; 3) the number of non-contributing subclones (chimeric clones, non-recombinant clones, clones containing *E. coli* genomic DNA as insert) within the library; 4) the average length of sequencing reads with high quality bases (Q20 as reported by *phred*). As a general rule, with an average length of 700 high quality bases, one 384-well plate, to be sequenced in for and rev, is required for a 10-fold coverage every 50 kb of DNA to be sequenced.

As the project proceeded, three different protocols for DNA template preparation for the sequencing reaction have been used in our laboratory. By time:

1. **PCR (Polymerase Chain Reaction).**

   The DNA template is obtained by performing a 384-well PCR reaction using universal M13 forward and reverse primers, a procedure that takes 2,5 hours. Then a step of gel electrophoresis analysis of the PCR products is required in order to identify positive PCRs. To optimize the sequencing efforts, we developed an application that creates a 384-well PCR plate of positive samples. For each 384-well PCR plate, a sample sheet is made as a list of the samples with a PCR product visible on agarose gel. The application positions the positive selected samples in a new 384-well plate, retrieves sample information from the central LIMS and maintains for each sample the name referred to the original 384-well bacterial plate. After these data have been generated, a robotic work-station (Microlab STAR, Hamilton Robotics) is used to pool positive PCR into clean 384-well PCR plate. In this way, only positive sample are processed (PCR-purification and sequencing reaction).

   The PCR method has several disadvantages:

   1. it is time-consuming and requires laboratory personnel hands-on time for the selection of the right PCR products;

   2. it produces sequences of low quality base in presence of homo/di-polymers or GC-rich regions (Fig. 3.2.4 A);

3. finally, after sequencing random subclones to a 10-fold redundancy, we frequently noticed gaps in the sequence assembly which also are gaps in the subclone representation. For some subclone samples with no PCR product, I prepared DNA template for sequencing reaction via plasmid minipreps. Up to 90% of these samples occurred in assembly gaps or within homo/di-polymeric low-quality bases regions. As a consequence the observed gaps in the assembly was not due to an uneven representation of the subclone library but to the low processivity of the *Taq* DNA polymerase through long stretch of TA (up to 100 bases) or regions making stable secondary structure. As a result, gap closure of the consensus sequence requires a more time-consuming direct sequencing approach.

2. TempliPhi (Amersham Biosciences).

The TempliPhi™ HT DNA Amplification Kit (Amersham Biosciences) utilizes bacteriophage Phi 29 DNA polymerase enzyme and random hexamer primers to exponentially amplify DNA by rolling circle amplification. Phi 29 DNA polymerase has a proofreading activity with an error frequency of 1×10^{-6}-10^{-7}.

Several features make this method quite promising:

1. it does not require culturing of a replica since it amplifies DNA directly from bacterial cultures or glycerol stock (as PCR strategy);
2. the amplification is isothermal (30°C) and can be performed in a heat block for 4 hours.
3. amplified template is directly sequenced without additional purification.

Respect to PCR, the sequence quality through short (up to 20 base pairs) stretch of homo/di-polymers is enhanced. Low quality reads are still obtained for GC-rich regions (Fig. 3.2.4 B).

3. Plasmid minipreps.

Template DNA is extracted from liquid bacteria culture using a procedure based upon alkaline lysis minipreps method adapted for high throughput processing in 384-well plate. Reagents are home-made and the dispensing operations are accomplished using a robotic work-station (Microlab STAR, Hamilton Robotics). Initially we recovered the DNA by isopropanol precipitation but this procedure resulted in short sequence reads due to sequencer capillary blockage. Therefore we have now improved the protocol, and we purify and concentrate the lysate using Montage Plasmid Miniprep Kit (Millipore) and the robotic work-station.

Compared to TempliPhi amplification, miniprep is a multistep procedure that requires overnight growth of bacterial replica plate and more labor-intensive efforts.

Despite of this, we have decided to adopt miniprep as the method for preparing DNA template for sequencing; its application has significantly increased sequencing reads lengths and base quality even within homo/di-polymer (up to 20 bases) and GC-rich regions. As a result it has reduced the efforts required in the successive finishing phase.

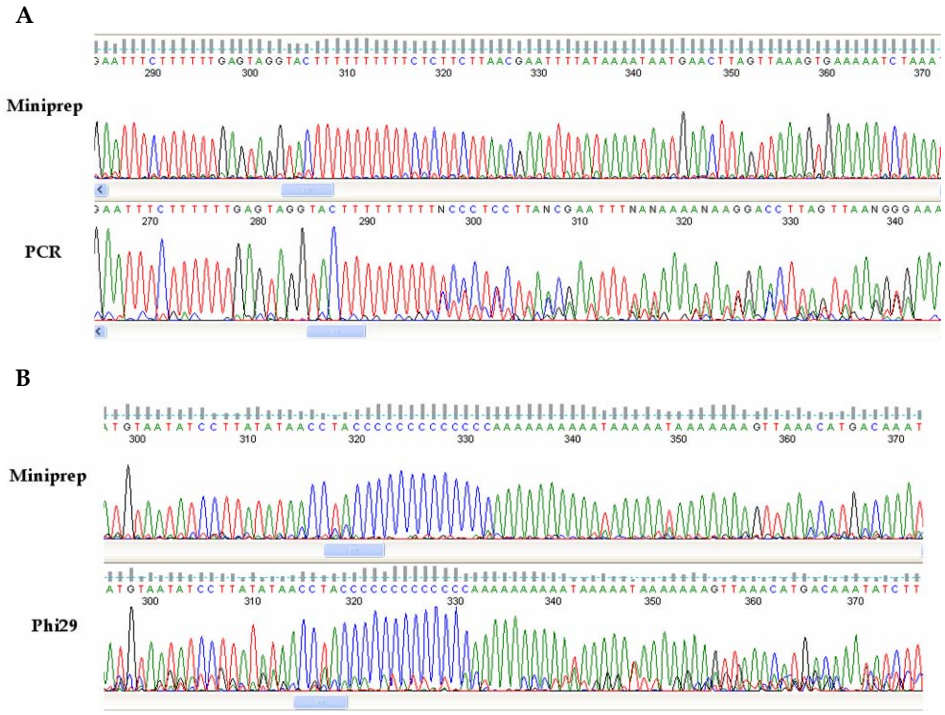


Fig 3.2.4. Comparison of same portions of reads obtained using miniprep DNA or PCR DNA and miniprep DNA or Phi29. In both the examples reported, the same plasmid clone had been prepared for the sequencing with the two methods, allowing a direct comparison. **A)** Low quality base after the poly(T) for the PCR template. **B)** Low quality base after the poly(C) for the Phi29 template.

The optimization of the procedure for the preparation of the sequencing templates had proceed with the improvement of the shotgun library protocol. Initially the shotgun library prepared using pZERrO and the TA-cloning vectors presented a high background of re-ligated vector; PCR allowed us to verify the quality of the shotgun library and to avoid the processing of non-recombinant clones. High quality shotgun libraries were obtained with the pUC19_*Bst*XI-based protocol and this had led us the possibility to decide for miniprep as the method for preparing DNA template.

In our modular sequencing pipeline, the template DNA is prepared and then is submitted to the sequencing core where all sequence data are generated using the ABI PRISM 3730 DNA Analyzer.

A central laboratory information management system (LIMS) has been developed to

track all sample plates throughout processing. Parent-child plate relationship and, in extension, forward-reverse sequence mate pairs are also recorded by the LIMS.

III.2.4 BAC sequence assembly and directed finishing phase

This phase of the BAC sequencing strategy begins with the production of an initial assembly based on shotgun sequence data.

Initially to each base of a read is assigned a quality value by means of the *phred* program (Ewing B *et al.*, 1998). Every high-quality read is further checked for matches with contaminants including sequences of vector (BAC and plasmid cloning vector), adapter dimers and *E. coli* genomic DNA; if a match is found, the read is removed from the assembly process. Finally, any match to the 5' plasmid vector junction in the initial part of the reads is removed.

In the successive step the included reads are computationally assembled by *phrap* (http://www.phrap.org/phredphrapconsed.html) on the basis of detected sequence overlaps. The resulting assembly typically yields a series of contigs, each of which consists of a collection of overlapping reads and a resulting consensus sequence. Generally, this preliminary assembly has gaps and low-quality regions; these regions are highly enriched in DNA stretches that are difficult to clone or sequence and thus are not represented even after a depth of 10-fold coverage with random reads.

The process of finishing converts the initial draft assembly into a high-quality continuous sequence and involves iterative cycles of computational analysis and laboratory work. The finishing of contigs is a time-consuming process, and needs expert knowledge to evaluate base calls, design primers for gap closure, and untangle complex sequences that obstruct a proper assembly.

The first step is to inspect the draft assembly for evidence of mis-assembly, arising from inappropriate merging of repeated sequences. In general, most clones passed assembly inspection since the use of BAC clones, instead of a whole-genome shotgun strategy, avoids problems arising from polymorphism and from different copies of repeated regions in the genome. In a few cases the presence of very similar local dispersed repeats required specific strategy. One approach was to isolate distinct copies of the repeat in subclones and primer-directed sequence these subclones; the final BAC sequence was manually assembled. When the repeat was longer than the average subclone insert length, a different strategy was used, making library of different insert size.

The second step is the gap closure. Because gaps tend to be associated to problematic sequences, gap closure is often a challenging process; it may require multiple attempts

using a variety of alternative methods. Spanned gaps are gaps where the two flanking contig ends are linked by one or more end-sequenced plasmid. Most of such gaps can be closed by primer-directed sequencing of the plasmid. Gap sequences are often recalcitrant to the standard sequencing protocol, making necessary the use of different alternative protocols (different buffer or temperature conditions). Specialized sequencing cycle has been optimized to overcome problems arising from secondary structure or from long homo/di-polymeric stretches (up to 100 bases) (§ II.5.6). In practice, we use a two-steps thermal cycle with a single step of annealing and extension that allow for more primer specificity and less slippage in the repeat or homo/di-polymeric tract during the thermal cycling (Fig 3.2.5). Techniques including sequencing amplified PCR product and primer walking directly on the BAC DNA are used to resolve unspanned gaps, where a contig end is not linked to any other contig.

The third step is the resolution of low-quality regions. This is accomplished by obtaining additional sequence reads from resequencing of existing shotgun subclones or from primer-directed sequencing, using in most cases alternative sequencing protocols.

Switching the DNA template preparation for sequencing process from PCR to TempliPhi and finally to plasmid minipreps, allowed us to generate sequences with a longer average length of high quality bases and to avoid removing from shotgun data the subclones containing regions recalcitrant to PCR amplification. This had greatly facilitated the finishing process, increasing its efficiency and lowering the number of experiments and, as consequence, the overall cost of this phase.

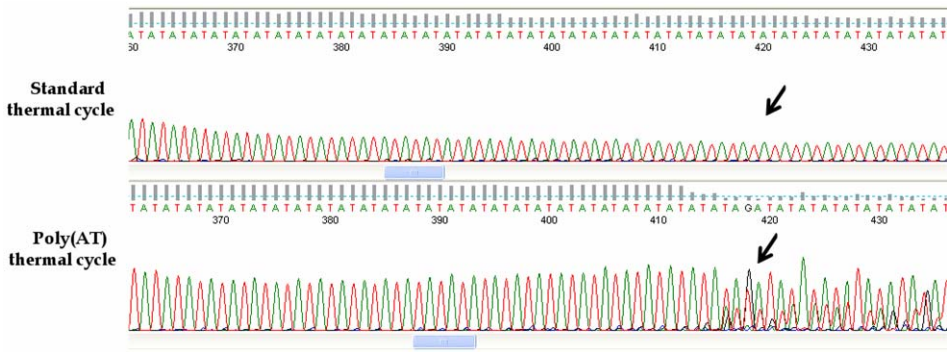


Fig. 3.2.5 Comparison of the quality of the same template through a 76 bp poly(TA). With the standard thermal cycle the *Taq* DNA polymerase slips and the sequence goes in roll-over. The poly(TA) thermal condition (§ II.5.6) allows to go over the poly(TA) with a readable G (pointed by the arrow) that marks the boundary of the polymer. In the assembly other reads contribute to the quality of the sequence after the poly(TA).

The final step involves quality control. To confirm the accuracy of the overall assembly, the restriction digestion pattern of the BAC predicted from the finished sequence is compared with the pattern observed experimentally, with any discrepancy indicating the possible presence of a sequence miss-assembly or a rearrangement of the cloned insert before sequencing. For seed BAC clones, the finished sequence is also analyzed for the presence of the associated marker sequence.

To consider a BAC sequence 'finished' we follow the standards adopted for the Tomato Genome Project and derived from the Medicago Sequencing Project, that are:

1. a single contig;
2. less than 3% of the sequence should be derived from multiple subclones sequenced from the same strand with the same chemistry. Less than 1% of the sequence should be derived from a single subclone;
3. more than 99% of the sequence should have less than one error in 10000 bases as reported by *phrap*;
4. the assembled sequence must be confirmed by restriction enzyme digestion.

In practice, we follow such standards for the finishing of the seed BAC clones and of the extending BAC with a small overlap. For extending BAC clones with a significant overlap (>2 kb), only the non-overlapping BAC sequence is finished following the standards in order to reduce sequencing redundancy and cost.

The sequence release policy is conducted in the spirit of the Bermuda agreement and, once a BAC clone is finished, the sequence is deposited in public databases. In practice, we submit the BAC sequences to the HTGS division of GenBank and also archive all the data of each BAC on SGN database (table 3.2.3). The submission to GenBank is made by using Sequin, UNIX version.

Status	Definition
HTGS Phase 0	Unassembled sequencing reads from a very light shotgun (no contigs)
HTGS Phase 1	Unordered and unoriented assembly of contigs
HTGS Phase 2	Ordered and oriented assembly of contigs, with or without gaps
HTGS Phase 3	Complete sequence, no gap

Table 3.2.3. Definition of the terms used to describe BAC clone sequencing projects. (for details http://www.ncbi.nlm.nih.gov/genome/guide/glossary.htm#HTGS). This nomenclature allows to distinguish all genomic sequence generated in a high-throughput manner.

III.3 Physical mapping of BAC clones

In order to facilitate the sequencing task, marker analysis strategies, cytogenetic protocols and a number of bioinformatics and molecular tools have been developed. The starting points for sequencing are BAC clones selected on the basis of markers from the tomato genetic map. Each sequenced anchored BAC serves as a seed from which expand in both directions. The identification of the most suitable neighbouring BACs in the euchromatin minimum tiling path is based on the use of a BAC-end database as well as on fingerprint contig physical map (FPC).

The chromosome localization of selected BACs (seeds and extending clones) is experimentally confirmed using different strategies, such as IL mapping, fingerprinting analysis and FISH. Currently, a subset of BAC clones has been localized on pachytene chromosome via FISH (Fluorescence In Situ Hybridisation) to confirm chromosome localization and to determine the boundaries between euchromatin and heterochromatin. The FISH analysis is performed by the cytogenetics group of Wageningen University, in the Netherlands.

With the progress of the project, the tomato genetic map has proved to have insufficient resolution and accuracy to guide a BAC-by-BAC sequencing project. Trying to overcome to this limitation, physical mapping of BAC clones with FISH has been an invaluable help to locate clones on the tomato genome. The Dutch group performs FISH on pachytene chromosome obtaining a spatial resolution of 1 Mb (Valárik M *et al.*, 2004). In most cases they confirmed BAC clones position based on genetic marker. In some other experiments FISH pachytene map revealed strikingly discrepancies between the chromosome position of BAC on the chromosome and their theoretical position as determined by DNA markers. So far, FISH has confirmed the position of only 4 BAC clones on chromosome 12, while it has revealed that the already sequenced BAC clone C12HBa0032K07 is located on chromosome 7. Furthermore FISH-based mapping can support the BAC-by-BAC sequencing providing additional information about BAC contig orientation, overlap of contig elements, and extent of gaps. These information can not be obtained using other genomics techniques, such as FPC map (Weier HU, 2001).

To improve spatial resolution, different strategies have been developed. In the Netherlands, seeds selection and BAC walking are supported by FISH but also by DNA fiber-FISH. This technique combines stretching of genomic DNA with fluorescent hybridization and allows visualizing the relative position of probes with a spatial resolution of 1 to 5 kb. In our laboratory, to assist the identification of the chromosome 12 minimal tiling path I have carried out the molecular combing, an innovative fiber-FISH method.

III.3.1 Molecular combing

In molecular combing deproteinised DNA molecules in solution attach with non-sequence specificity to a silanised hydrophobic glass surface by their extremities (Lebofsky R. *et al.*, 2003; Monier K *et al.*, 2001; Allemand JF *et al.*, 1997). The combing process produces a high-density array of DNA molecules that are between 200 and 700 kb in length. DNA fibers are uniformly and in the same orientation stretched along their length regardless of sequence content. This uniform stretch provides a length scale relating physical distance on the silanised surface to genomic length, so that 1 µm = 2 kb. For this cytogenetic investigation I used the diploid cherry tomato cultivar and improvements were needed to adapt the technique for the analysis of the tomato genome. High resolution genomic studies depend to a large extent on visualizing probes on one DNA molecules. A critical step to obtain high quality combed DNA in the megabase range is to prepare high-molecular-weight (HMW) DNA, due to the presence in plants of a hard cell wall. So as first aim, I optimized a protocol for efficiently preparing HMW DNA embedded in agarose plug from tomato leaves protoplasts (Ganal MW *et al.*, 1989). To produce an array of high density but well-separated combed DNA molecules 250000 tomato protoplasts are combed onto a 22 x 22 mm slides (Fig. 3.3.1).

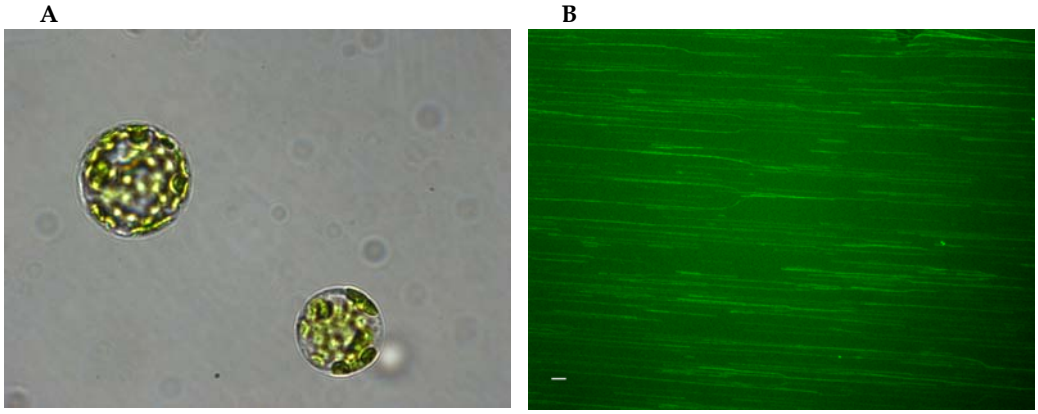


Fig. 3.3.1 Process of preparing HMW combed DNA molecules from tomato leaves. **A)** Isolated tomato protoplasts (x400) from diploid cherry tomato cultivar leaves. **B)** A high density array of tomato DNA molecules combed onto a 22 X 22 mm slides obtained from 250000 protoplasts. The molecules are visualized with YOYO1. Bar=10 µm=20 kb.

I applied FISH on combed DNA molecules to confirm molecular size and overlap of the tiling BACs. I focused my attention on three BACs (C12HBa0115G22, C12HBa0165F06 and C12HBa0326K10) corresponding to two genetically linked contigs, with a genetic distance of 0.3 cM (based on Tomato-EXPEN 2000 *S. lycopersicum* LA925 x *S. pennellii* LA716 type F2.2000). The BACs C12HBa0115G22 and C12HBa0326K10 are two seed

clones anchored to the genetic map with the molecular markers T1676 (86 cM on chromosome 12) and TG468 (85.7 cM on chromosome 12), respectively. The BAC C12HBa0165F06 was selected using PABS tool as extending clone of the SP6-end of C12HBa0115G22 with an overlap of 1636 bp.

These BACs were hybridised to combed DNA in a multi-colour FISH experiment. The experiment scheme was designed so that the three BACs were simultaneously hybridised. For this combinatorial labeling I used two labeling/detection combinations with Cy3 and 488 as fluorescence detection systems in order to produce images in the red and green fluorescence channels. Moreover the presence of overlap gives additional information that permits distinguishing between the two seeds clones. I also optimised the BAC-FISH experiment by using the repeated fraction of genomic DNA, *C_ot-1*, to suppress FISH signals from repetitive sequences (Zwick MS *et al.*, 1997). The using of *C_ot-1* as blocking reagent depends on probe sequence composition in repetitive elements.

III.3.2 Results

The three BACs were unambiguously oriented and the gap between the two contigs was accurately measured (Fig. 3.3.2; Table 3.3.1). The length of each individual BAC was estimated: the molecular sizes of C12HBa0115G22, C12HBa0165F06 and C12HBa0326K10 are 143 kb, 150 kb and 133.5 kb, respectively. The measured lengths of BAC C12HBa0326K10 and C12HBa0115G22 are in perfect agreement with the size of the completed consensus sequence (the C12HBa0165F06 clone is still in HTGS1 phase). I have also shown the 1.6 kb overlapping region between clones C12HBa0115G22 and C12HBa0165F06. Therefore this experiment has confirmed that BAC clone C12HBa0165F06 is a correct extension of C12HBa0115G22 and that the two contigs (C12HBa0326K10 and the pseudomolecula C12HBa0115G22-C12HBa0165F06) belong to the same chromosome 12 region.

	C12HBa0115G22	**C12HBa0165F06**	**C12HBa0326K10**
BAC size estimates by FISH (bp)	143000	150000	133500
Length of the finished sequence (bp)	144653		134225
Difference with sequencing	1.1%		0.5%

Table 3.3.1. Comparison of the size of BAC clones hybridized on combed DNA molecules and the length of the finished sequence. The BAC C12HBa0165F06 currently is in HTGS1 sequencing phase.

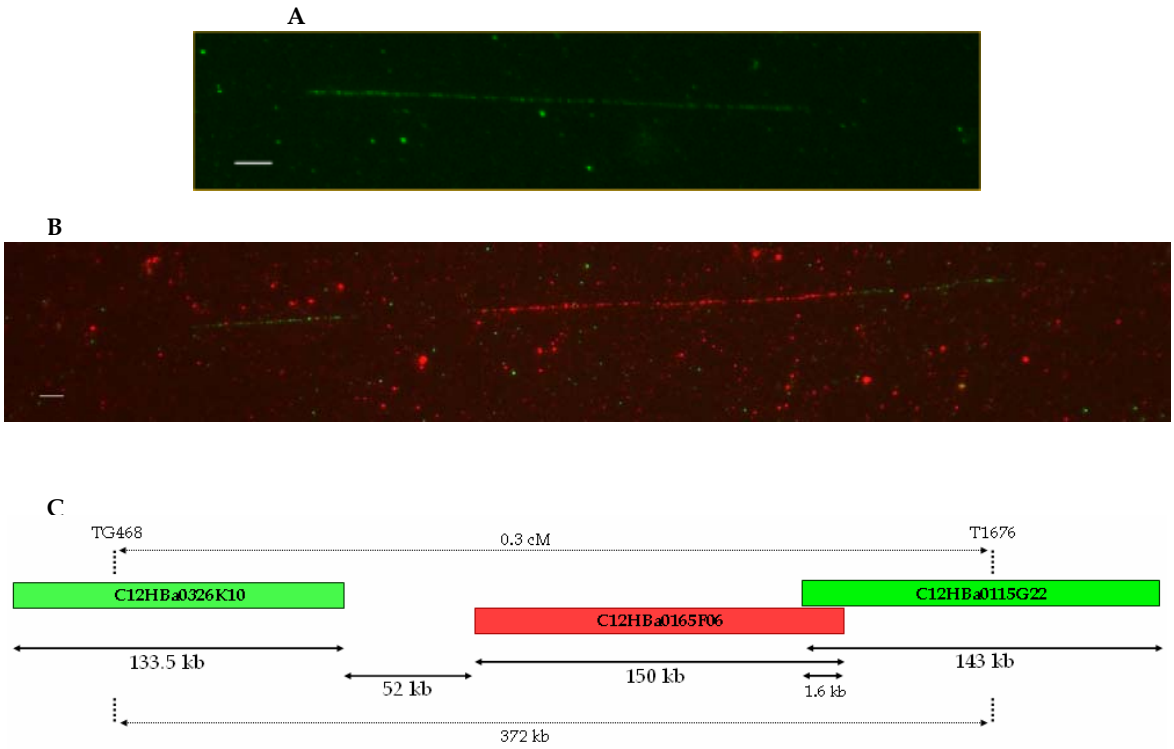


Fig. 3.3.2 Physical map of the BAC clones C12HBa0326K10 (seed), C12HBa0165F06 and C12HBa0115G22. The clones C12HBa0326K10 and C12HBa0115G22 are seed clones anchored to the *S. lycopersicum* genetic map Tomato-EXPEN 2000 and the corresponding genetic marker are indicated. The BAC C12HBa0165F06 is an extending clone with an overlap of 1636 bp. **A)** C12HBa0326K10 hybridization on combed tomato genomic DNA. **B)** Two color FISH on combed DNA of the three BACs: the orientation of the two contigs is unambiguously shown and an intern gap of 52 kb is measured. Bar=10 µm=20 kb. **C)** Schematic representation of the inferred BACs orientation and comparison of physical and genetic distance between molecular markers.

III.3.3 Comparison of FISH on combed DNA and molecular genetic map

The physical distance between the two contigs measured by means of FISH on combed DNA is 52 kb. On the molecular linkage map the distance of the two contigs is 0.3 cM, with an average genetic distance of 1 cM on the tomato map corresponding theoretically to approximately 750 kb (Tanksley SD *et al.*, 1992). Our data show a physical distance between the genetic markers associated with the two seed clones of 372 kb. This indicates a ratio of 1.24 Mb/cM between the two contigs (table 3.3.2) and it implies a far higher ratio than the average ratio of 750 kb/cM. While Ganal M *et al.* (1989) calculated 4 Mb/cM near the centromere of tomato chromosome 4, the higher ratios of 21.74 Mb/cM and 100 Mb/cM were found respectively on the short and long arm pericentromeric heterochromatin region of chromosome 12 (Budiman MA *et al.*, 2004). Sherman JD *et al.* (1995) attributed decreased recombination values in the centromeric region of tomato chromosome to the suppression of recombination. Our observed ratio of 1.24 Mb/cM

between the two contigs implies a far lower ratio in this euchromatin regions. Tor M *et al* (2002) have instead determined 330 kb/cM in the euchromatin regions of chromosome 2L.

	Physical length (Mb)	Genetic length (cM)	Mb/cM
TG468-T1676	0.372	0.3	1.24

Table 3.3.2 Ratio of Mb/cM for the analyzed region of chromosome 12.

III.3.4 Conclusion

This data indicates that FISH analysis on combed DNA molecules of BAC clones is an accurate and efficient method to validate the minimal tiling path with a resolution in the order of few kilobases. This technique can also be used to discover discrepancy in the molecular genetic map and to estimate the distance between clones in base pair. Furthermore it can provide valuable information for quality control of the sequence assembly and for measuring the size of selected BAC clones in order to guide the shotgun sequencing process.

At the same time, when using DNA-combing, it must be considered that the chromosome identification is not possible without the co-localisation with a known marker or clone. Likewise, structural features of chromosome as eu/heterochromatin, centromeres and telomeres are not distinguishable. Moreover there is a technical limit due to the length of the molecules that can be combed. Because deproteinised molecules obtained are fragile and sensitive to mechanical stress, combed DNA molecules normally do not exceed 600-700 kb.

III.3 Bioinformatics analysis

A central goal of genome analysis is the comprehensive identification of functional, regulatory and structural elements. This task remains challenging, and is greatly dependent on the availability of the genomic sequence together with other resources such as cDNA collections, availability of other genomes for sequence comparison and improved computational methods.

The preliminary effort of the bioinformatics centres in the SOL network is mainly focused on setting up procedures and methodologies to provide a reliable tomato genome annotation. In order to contribute to this task, we set up a procedure for a preliminary annotation of the BAC sequences. The collection (October 2007) comprises 358 BAC sequences (in HTGS3 phase) which have been uploaded to the SGN database by all the sequencing centres belonging to the consortium. The annotation process based on individual BAC sequences has some limitation because of incomplete genes at the ends of BACs and duplicated annotations on the overlapping regions. So, when possible, the 'pseudomolecules' (i.e. merged sequences of overlapping BACs) were constructed using the information of the TPF (tiling path format) files, available from the SGN repository (ftp://ftp.sgn.cornell.edu/tomato_genome/tpf). As a result, we analyzed 22 pseudomolecules, mostly derived from the merging of two BACs, and 306 individual BAC clones, for a total of 40 Mb.

This part of my PhD work has been done in collaboration with Dr. N. Vitulo (CRIBI, University of Padua).

III.4.1 Gene prediction and BAC annotation

The primary task of genome annotation is the identification of gene location and the definition of gene structures on the genomic sequence; currently, no gene finders programs specifically calibrated on tomato are available. Researchers of University of Naples (Italy) and University of Ghent (Belgium) within the SOL project are focusing their efforts in the identification of a sufficiently sized training set and in the calibration of computational methods for tomato gene prediction. With the knowledge and the expertises developed during the underway *V. vinifera* genome annotation project, we identified a first-draft reference set of 5123 gene loci in the BAC sequences using a combination of *ab initio*, homology-based and expressed sequence tag (EST)-based methods. Repetitive sequences, including transposable elements (TEs), were identified using the TIGR Solanaceae Repeats database (Ouyang S *et al.*, 2004) and removed from the predicted gene set.

On the basis of available evidences, the main features of the BACs annotation are summarized in Table 3.4.1, where the discrepancy between the reported average and median values are a consequence of the data distribution as shown in Fig 3.4.1. Our preliminary results show the effect of the limited size of the training set we used, that contributes to reduce the accuracy in the prediction of the precise gene structure, in terms of boundaries of all the exons and of the coding sequence. With the progress of tomato genome project, an increased number of full-length cDNA sequences will be made available to the community, so that we aim to produce a more accurate predictions of the gene structure. Otherwise, in our opinion we identified the majority of BAC coding regions, even if these are still not exactly organised in a reliable gene architecture.

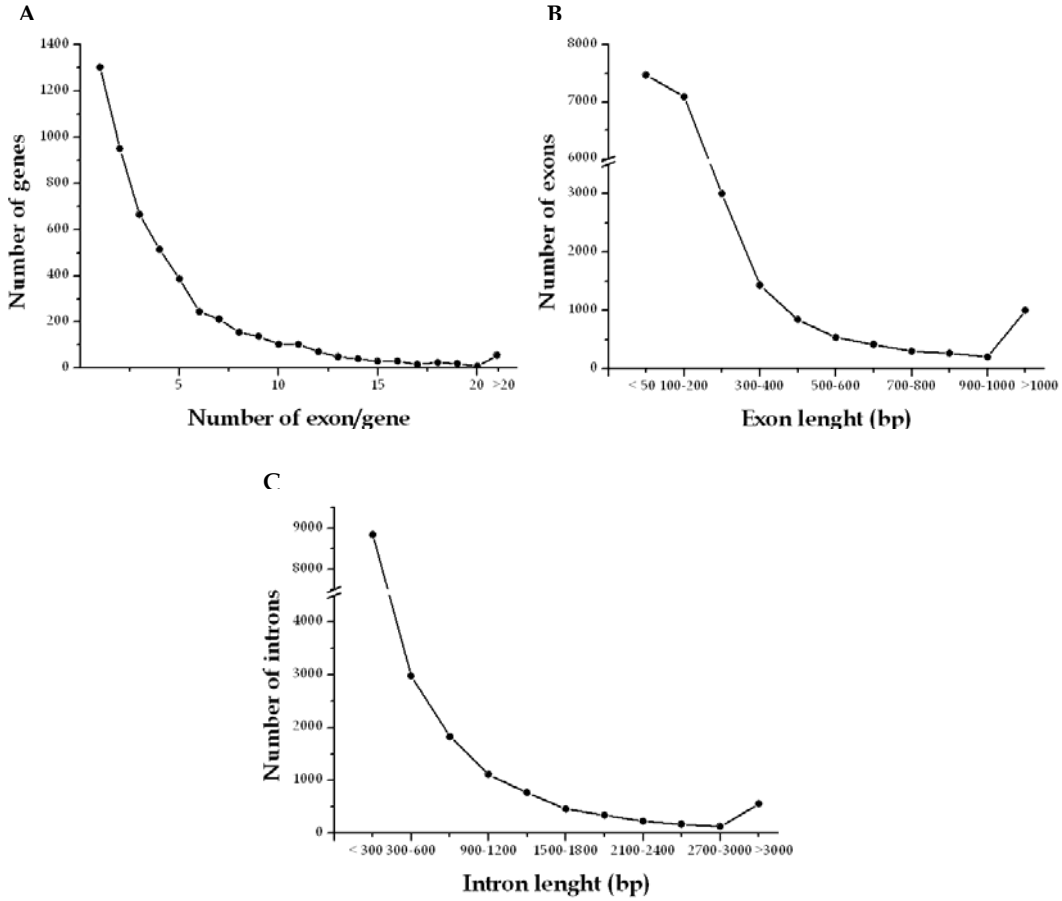


Fig. 3.4.1. Frequency of number of exons per gene (**A**), exons (**B**) and introns (**C**) lengths found in the our BACs prediction data. The values of exon length are grouped in a 100 bp window; the values of intron length are grouped in a 300 bp window.

Features per gene	*S. lycopersicum*		*A. thaliana*	*O. sativa*	*P. trichocarpa*	*V. vinifera*
	Average	Median	Average	Average	Average	Median
Gene lenght (bp)	3393	2244	1992	2699	2300	3399
Exons per gene	4.4	3	5.2	4.7	4.3	4.9
Exon lenght (bp)	265.6	141	250	254	254	130
Intron per gene	3.4	2	4.2	3.7	3.3	3.9
Intron lenght (bp)	653.7	290	168	413	379	213

Table 3.4.1. Statistics of the predicted genes in the *S. lycopersicum* BACs and pseudomolecules; comparison with *A. thaliana* (Arabidopsis Genome Initiative, 2000), *O. sativa* (International Rice Genome Sequencing Project, 2005), *P. trichocarpa* (Tuskan GA *et al.*, 2006) and *V. vinifera* (The French-Italian Public Consortium for Grapevine Genome Characterization, 2007).

III.4.2 Analysis of gene content and organization

The analysis of the sequenced BACs revealed a average coding percentage of 14.8% and a average gene density of one gene every 7.7 kb. Considering that this is a preliminary annotation and that improvements are necessary to better define the gene structure and boundaries, I think that the percentage of coding sequence of each BAC and pseudomolecules gives a much more reliable description of the tomato genome rather than the value of gene density. Yet, gene density allows me to compare our data with the ones reported in literature.

Of the 328 annotated sequences, 21 BACs and 7 pseudomolecules were localized via FISH on tomato pachytene chromosomes, both on euchromatin and on heterochromatin. The comparative analysis of these BACs provides a first general insight into the differential organization of tomato euchromatin and heterochromatin. The analysis of the 18 BACs and 5 pseudomolecules assigned by FISH to euchromatin indicates that the euchromatin have an average coding percentage of 16% and contains, on average, one gene every 6.9 kb. Van der Hoeven R *et al.* (2002) also estimate an average gene density of 7 kb/gene in tomato euchromatin, and Wang Y et *al.* (2006) of 6.7 kb/gene. Two BACs and one pseudomolecula were located on the boundaries between euchromatin and heterochromatin. They have a lower coding percentage (3.88%, 6.89% and 8.58%) and gene density (13.7, 16.8 and 23.5 kb/gene), and also contain many retrotransposon-like sequences. Finally the remaining one BAC and one pseudomolecula, derived from heterochromatin, contain only transposon-related genes similar to *copia*- and *gypsy*- like retrotransposons.

The estimated non-transposon gene densities for euchromatin is slightly higher than the 4.5 kb/gene of *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000) and similar to the 6.9 kb/gene of *Oriza sativa* (International Rice Genome Project, 2005). At the same

time, striking differences are found in heterochromatin. Rice heterochromatin has a gene density only slightly lower than euchromatin (11 kb/gene) (Jiao Y *et al.*, 2005). In contrast, in tomato the analyzed heterochromatic BACs suggest a non-transposon gene density in heterochromatic regions dramatically lower than that in euchromatin. Considering that the three BACs located in heterochromatin together encompass 384 kb with only transposon-related genes, it seems realistic that the tomato genome has a gene density in the heterochromatic regions much more similar to the *Arabidopsis thaliana* (256 kb/gene) heterochromatin.

The tomato genome is composed of ~950 Mb of DNA, 23% of which is euchromatin (Arumuganathan K *et al.*, 1991; Peterson DG *et al.*, 1998). With a coding percentage of 16% and a average coding gene length of 1168 bp (table 3.4.1), we can thus estimate that the euchromatin contains ~30000 genes. This result is highly similar to the ~35000 genes estimated for the entire genome on the basis of EST database (Van der Hoeven R *et al.*, 2002). Thus the sequencing of the euchromatic regions and of the boundaries between euchromatin and heterochromatin would reveal the majority of the genes.

III.4.3 Implication of this study to the sequencing of the tomato genome

Currently the established strategy for the sequencing of the tomato genome is a BAC-by-BAC approach and involves the sequencing of a minimal tiling path of BAC clones covering the approximately 220 Mb of euchromatin. Different molecular and bioinformatics tool have been developed to assist the BAC clones selection and extension but also to focus the sequencing effort to the gene dense euchromatin regions of each chromosome. Fig 3.4.2 describes for each chromosome sequencing project the distribution of sequenced BACs in terms of percentage of bases predicted as non-TE coding.

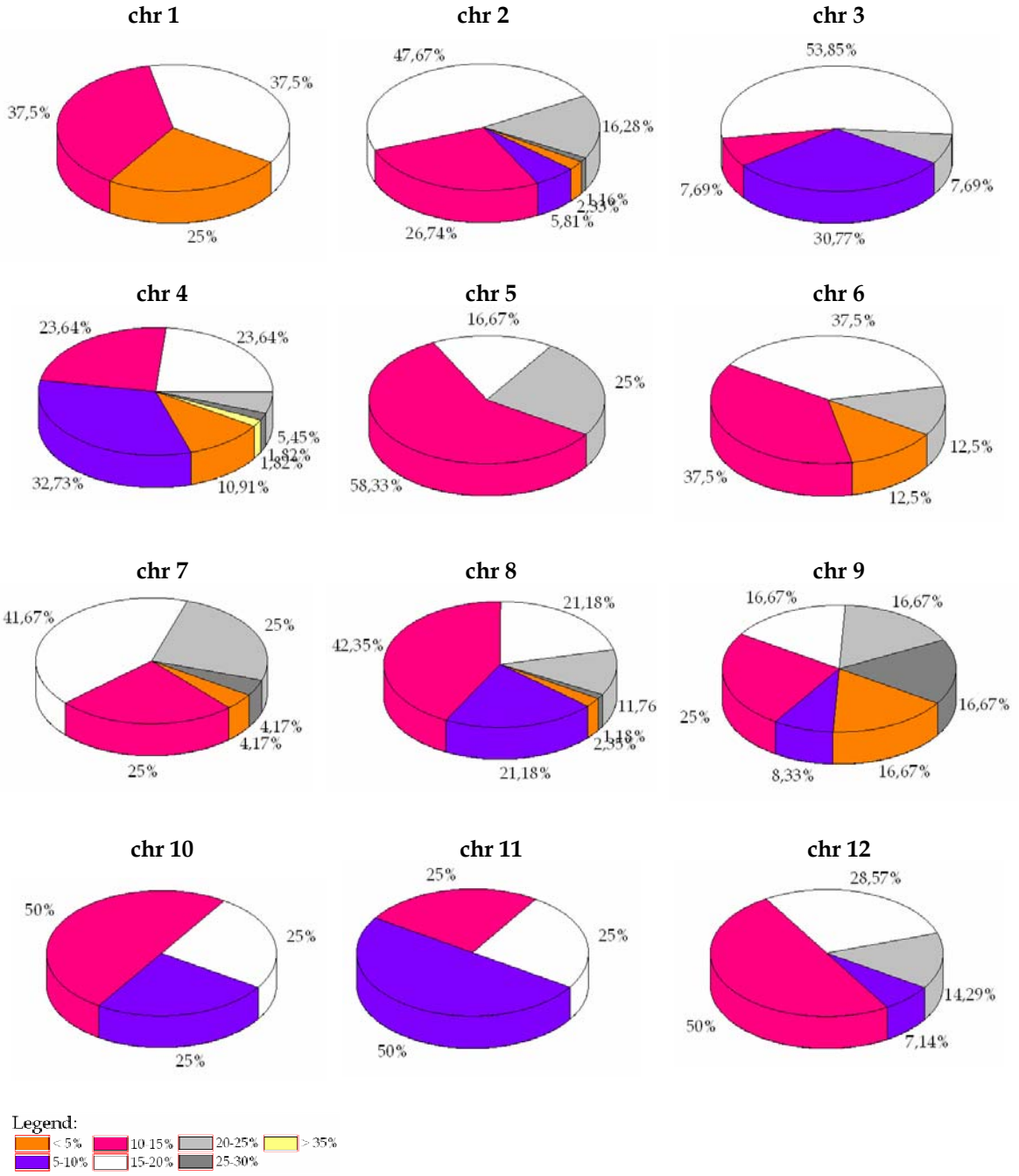


Fig 3.4.2. For each tomato chromosome sequencing project, distribution of HTGS3 BACs based on predicted coding percentage. General statistics are summarize in table 3.4.2.

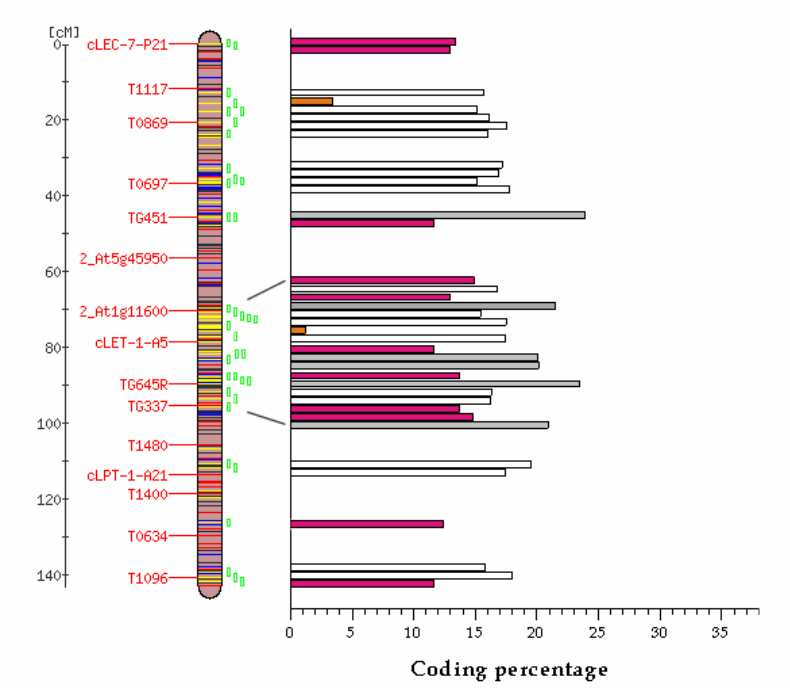
Chromosome	Number of HTGS3 BACs	Sequenced kb (kb)	Average coding percentage	Average gene density (kb/gene)
1	8	1104	12.14%	7.3
2	86	10131	15.65%	7.7
3	13	1574.5	15.7%	8.5
4	77	8848.5	12.5%	8.7
5	12	1253.7	14.8%	7.4
6	8	906.1	15.4%	8.1
7	24	2051.3	16.6%	7.7
8	85	9528.9	13.9%	7.6
9	12	1208	15.1%	7.8
10	4	487.3	13.1%	6.7
11	4	452.2	11.4%	8.2
12	20	2123	12.9%	7.4
Total	353	39668.5		

Table 3.4.2. Sequencing statistics of the tomato genome project (October 2007) and chromosomal distribution of predicted genes.

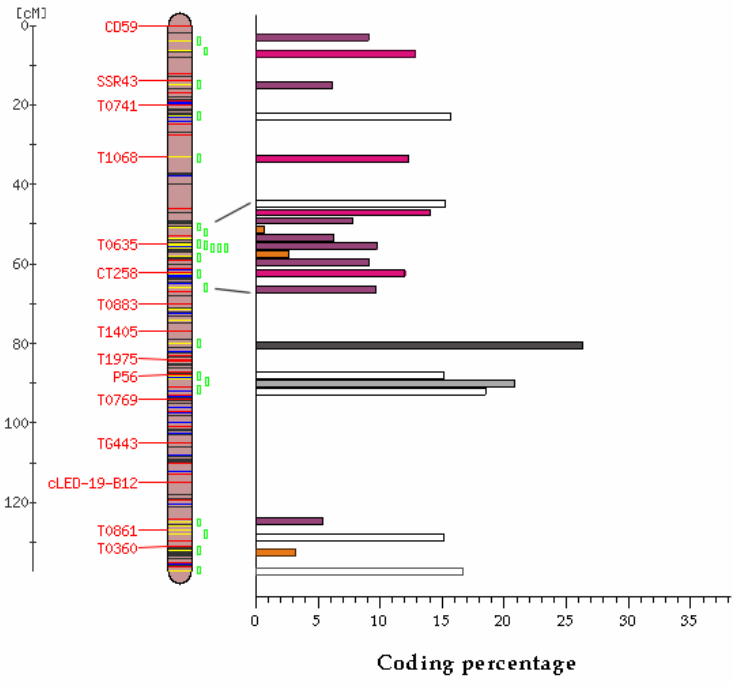
The best coverage is on chromosome 2 (with 86 HTGS3 BACs), on chromosome 4 (with 77 HTGS3 BACs) and on chromosome 8 (with 85 HTGS3 BACs). This data allows comparing the quality of the different strategies adopted from the three sequencing centre in accordance with their specific expertise for the identification of the euchromatic minimal tiling path. The Sanger Centre has mainly based the sequencing of chromosome 4 on physical mapping, using restriction digestion to characterize each clone and to infer the order of clones. To integrate the initial datasets (FPC map constructed on the *Hin*dIII BAC library) a Sanger Initiative was focused on the generation of additional fingerprint data from the *Mbo*I library (ftp://ftp.sanger.ac.uk/pub/tomato/map/). This strategy results in the sequencing of a large portion of BACs with a low coding percentage, and these clones are probably located in transition regions between euchromatin and heterochromatin. Thus, the chromosome 4 minimal tiling path likely ranges even in heterochromatin regions closed to heterochromatin-euchromatin borders. An example is the coding density distribution of BACs between 56 cM (T0635) and 62 cM (CT258) (Fig 3.4.3).

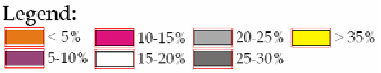
On the other hand, FISH has been extensively used to validate the extension of the tiling path through the euchromatin arms of chromosome 2 and 8. As a result, the majority of the sequenced BACs are characterized by a much higher non-TE coding portion of their sequence.

chr2


[cM]
0
20
40
60
80
100
120
140
cLEC-7-P21
T1117
T0869
T0697
TG451
2_At5g45950
2_At1g11600
cLET-1-A5
TG645R
TG337
T1480
cLPT-1-A21
T1400
T0634
T1096
0
5
10
15
20
25
30
35
Coding percentage

chr4


[cM]
0
20
40
60
80
100
120
CD59
SSR43
T0741
T1068
T0635
CT258
T0883
T1405
T1975
P56
T0769
TG443
cLED-19-B12
T0861
T0360
0
5
10
15
20
25
30
35
Coding percentage


Legend:
< 5%
5-10%
10-15%
15-20%
20-25%
25-30%
> 35%

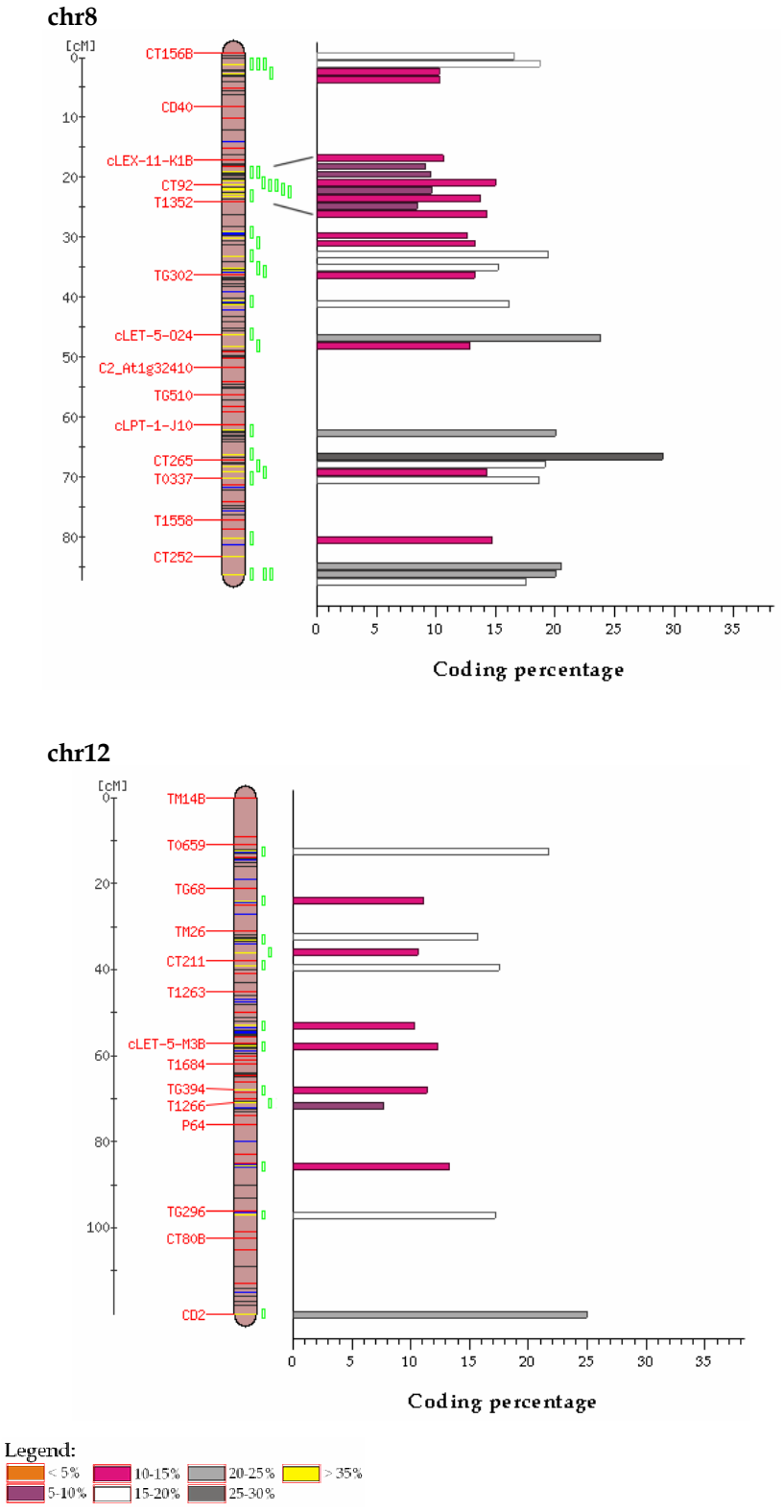


Fig 3.4.3. Histograms of the predicted coding percentage along chromosome 2, 4, 8 and 12. Each bar is a sequenced BAC clone mapped to a chromosome using genetic markers (Tomato-EXPEN 2000). The bar colour represents the predicted coding percentage of the BAC clone (for details see table 3, § Appendix).

III.4.4 Gene family

Gene duplication is the major determinant of the size and gene complement of eukaryotic genomes (Lockton S *et al.*, 2005). During evolution the process of gene duplication and divergence play an important role in genome evolution, providing the opportunity for the development both of novel gene function and functional redundancy. Furthermore the study of the molecular process by which functional innovation is associated with gene duplication takes the interest not only of evolutionary biologists but also of agricultural biologists for the improvement of specific trait. Thus the identification of the members of a known gene family as well as of new gene families has become an increasingly important step in genomics studies.

At date, the 358 sequenced tomato BACs represent only the ~18% of the total euchromatin. On this subset, we performed a preliminary analysis of gene organization. Protein families were identified using as parameters sequence similarity exceeding a BLASTP value of $E< 10^{-5}$ and extending over at least 65% of the protein length and identity of >30%. The majority of the resulting families are composed of two members probably because of the incomplete coverage of the genome so that not all the members of any gene family are identified. Besides that, as a pilot study, within the most abundant genes we identify two family that may be of some interest for plant researchers. By means of a phylogenetic analysis we tried to understand if these genes are organized into similar gene families in other plant species and which is the degree of conservation of the family size.

III.4.4.1 Identification of Aurora-like kinases family

A manual inspection of preliminary data allowed us to identify the gene family that contains the three tomato genes encoding for proteins annotated in *Arabidopsis thaliana* as Aurora-like kinases.

A crucial process in cell division concerns the dynamic restructuring and segregation of chromosomes and a major role in the regulation of this process is played by reversible protein phosporilation. Aurora kinases belong to the serine/threonine protein kinase family that regulates different processes occurring during mitotic events through phosporilation (Andrews PD *et al.*, 2003; Carmena M *et al.*, 2003) in yeast, plant and animal systems. The number of Aurora kinase paralogs are different among organisms. *Saccharomices cerevisiae* and *Schizosaccharomices pombe* have only one Aurora kinase gene in their genome, while animals have three Aurora kinase genes; in *Arabidopsis thaliana* three Aurora kinases (AtAurora1 [Accession n° AB196733], AtAurora2 [Accession n° AB196734], and AtAurora3 [Accession n° AB196735]) were characterized. All plant and

non-plant Aurora-like kinases share a similar structure (Deminov D *et al*, 2005; Kawabe A *et al*., 2005) emphasizing the ancient nature of the process controlling cell division. This conservation allowed a sequence-based identification of Aurora kinases from diverse plants.

With this aim, we compared the predicted tomato Aurora-like kinase family with the protein data set of *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, *Populus trichocarpa* and *Vitis vinifera*, each belonging to different plant family (*Solanaceae*, *Brassicaceae*, *Fabaceae*, *Poaceae*, *Saliceae* and *Vitaceae*, respectively) (Fig 3.4.4).

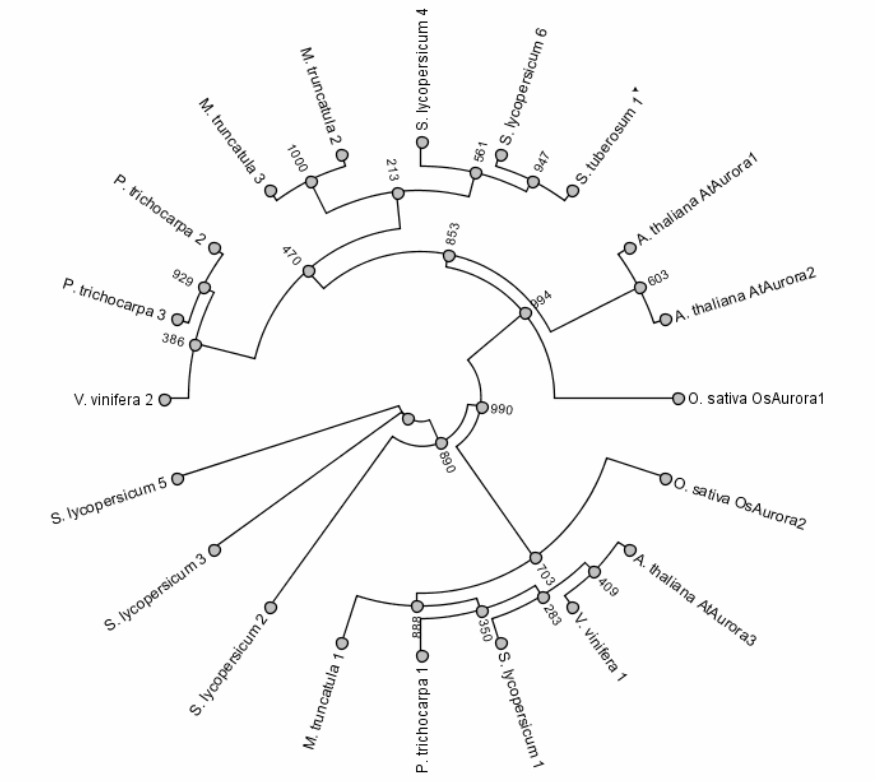


Fig. 3.4.4. Phylogenetic tree showing the relationship of Aurora kinases using the neighbour-joining method. Numbers are bootstrap values derived from 1000 resamples. Aurora derived from EST sequences are indicated by ⅄.
The accession number of the proteins in the tree are reported in table 1, § Appendix.

Phylogenetic analysis suggests that plant Aurora kinases have been separated early in plant evolution into two major subgroups, indicating that plant species possess two different Aurora kinase proteins, that can be classified on the basis of their similarities to *Arabidopsis thaliana*. One subgroup has similarity with *Arabidopsis* AtAurora3, while the other one with AtAurora1. The tree reveals that the AtAurora3 orthologous gene is found in all the analyzed plant genomes. With the exception of *Vitis vinifera* and *Oryza sativa*,

the gene belonging to the other subgroup has undergone a duplication event in the evolutionary lineage of each species. The identification of only one gene orthologous to AtAurora1 in *Vitis vinifera* may represent the real genome composition or may be a consequence that for the analysis we use gene prediction release based on the 8-fold draft sequence of the grapevine genome (The French-Italian Public Consortium for Grapevine Genome Characterization, 2007).

In tomato the three isoforms of the Aurora-like kinases exist.

The phylogenetic tree also shows three tomato predicted genes (*S. lycopersicum* 2, 3 and 5) that belongs to distinct branches. These genes contain a kinase domain that was recognized by BLASTP in the protein family construction.

The precise function of Aurora kinases is still unknown but the phylogenetic analysis suggests that paralogs maintained a conserved role in cell cycle-related signal transduction pathways.

III.4.4.1 Identification of vacuolar processing enzyme, VPE, family

Another family with a significant number of members corresponds to vacuolar processing enzyme family.

Vacuolar processing enzyme (VPE) is a Cys protease that has substrate specificity toward Asn and Asp residues, and VPE homologs are found in various organisms, including plants (Hara-Nishimura I *et al*., 1998) and mammals (Chen JM *et al*, 198; Shirahama-Noda K *et al.*, 2003). Plant VPEs are separated into three types: seed-type VPEs, vegetative-type VPEs and uncharacterized-type. VPE was originally identified as protease responsible for the maturation of seed storage proteins (Yamada K *et al.*, 1999), and successive research has shown that it is a key protease responsible for the maturation of various vacuolar proteins also in vegetative tissues. Vegetative-type VPEs are expressed during senescence and pathogen-induced hypersensitive response so that VPEs play an essential role in the molecular mechanism of vacuole-mediated cell death in both defence and development. Four VPE genes, αVPE, βVPE, γVPE, and δVPE were found in *Arabidopsis thaliana*. βVPE is expressed in seeds and is essential for the proper processing of storage proteins (Kinoshita T *et al*., 1999). αVPE and γVPE are expressed in vegetative organs and are upregulated in association with various types of plant cell death and under stress conditions (Hara-Nishimura I *et al.*, 2000).

The phylogenetic tree of plant VPEs confirms the existence of the four sub-class of plant VPEs (Fig 3.4.5). With the available partial sequence of the tomato genome, we identified tomato homologous of the Arabidopsis genes αVPE, βVPE and γVPE. We also identified

tomato genes that can be assigned to the branches of the *Nicotiana tabacum* vacuolar processing enzyme NtPB1 and NtPB3 subfamily. NtpPB1, NtPB2 and NtPB3 are a novel subfamily of VPE found in tobacco. NtPB1-3 is expressed during embryo- and microsporogenesis, and NtPB3 also in vegetative organs (Zakharov A *et al.*, 2004). Examination of the phylogenetic relationship of the VPE protein family indicated that it is related with the *Arabidopsis thaliana* glycosylphosphatidyl inositol (GPI)-anchor transamidase family, of witch the genes *A. thaliana* 1, 2 and 3 are members.

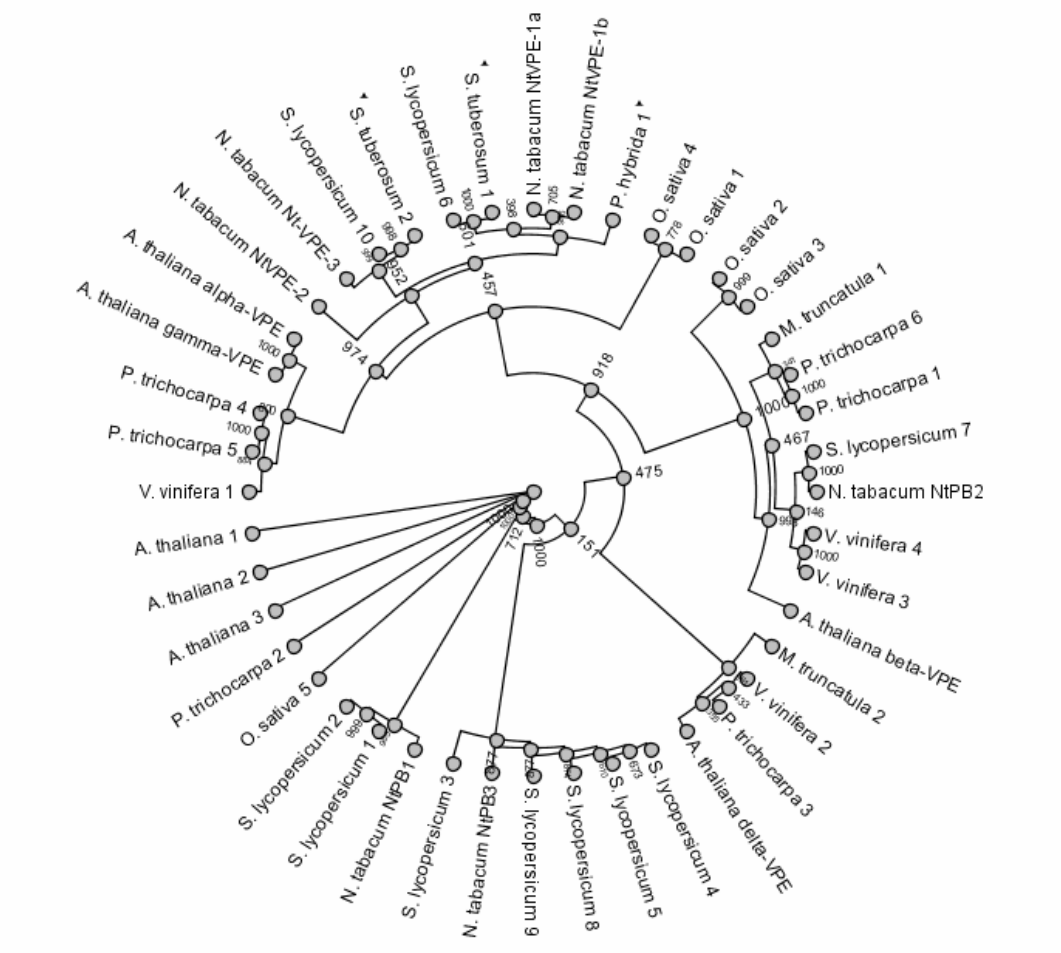


Fig. 3.4.5. (**A**)Phylogenetic tree showing the relationship of VPE proteins using the neighbour-joining method. Numbers are bootstrap values derived from 1000 resamples. Proteins derived from EST sequences are indicated by ⅄.
The accession number of the proteins in the tree are reported in table 2, § Appendix.

4.

Conclusions

Tomato (*Solanum lycopersicum*) is a preeminent model system for genetic studies in plants in addition to its worldwide agricultural and economic importance as a crop. To better understand the functional and structural aspect of its genome, the 'International Solanaceae Genome Project' (SOL) was launched in 2003 with the goal of obtaining a highly accurate sequence of the euchromatin portion of the genome (http://www.sgn.cornell.edu/solanaceae-project/). This project is a collaboration involving sequencing centres in ten countries, and Italy is responsible of the sequencing of chromosome 12.

The tomato chromosome 12 project is funded by MUR (FIRB, 'Tomato Genome Project'; Italy), by the European Union ('EU-SOL Project') and by MIPAF ('Agronanotech Project'; Italy).

The sequencing proceeds with a BAC-by-BAC strategy, choosing minimally overlapping BAC clones covering the euchromatin regions and performing shotgun sequencing of each clone. We started to construct the sequence scaffold of chromosome 12 according to mapping information available at SGN. Seed BAC clones anchored to the genetic map (Tomato-EXPEN 2000) were selected using different strategies, such a IL mapping and internal sequencing. The progressive construction of the minimal tiling path was performed comparing each completed BAC with the available BAC-end database.

With the progress of the project, this approach revealed several advantages but also numerous limitations. The main advantage is that the BAC-based sequencing simplifies the finishing process because each clone is assembled individually and so the possibility of large scale miss-assembly is reduced. Furthermore BAC clones can be made available worldwide and large genome centres (as the Sanger Centre, involved in the tomato genome project with the sequencing of chromosome 4) can cooperate and share expertises with smaller groups.

The generation of the minimal tiling path of BAC clones is the limiting step. With the progress of the project, the tomato genetic map has proved not to be robust, the physical map to be not detailed and the BAC-end database to contain a high percentage of low-quality reads. As a consequence, false overlaps become a real risk in the extension process. In order to improve and assist the minimal tiling path construction, different

strategies have been developed. We have made available to the SGN community an informatics tool called PABS specifically designed to assist the selection of reliable neighbouring BACs, trying to minimize the possibility of mistakes and optimize the merging of overlapping BACs (Todesco S. *et al.*, 2008).

Furthermore, in my PhD thesis I show the significance of BACs hybridization on combed DNA molecules as a molecular cytogenetic tool for supporting chromosome walking. This technique allows the physical mapping of BAC clones with a spatial resolution of a few kilobases (Lebofsky R. *et al.*, 2003).

Despite the enormous efforts at international level, the available genetic map has neither the density nor the resolution to provide useful indications for completing the sequencing, since there are large chromosome regions which are not yet targeted with markers. A recent release of markers from Syngenta to the SGN repository (ftp://ftp.sgn.cornell.edu/tomato_genome/bacs/syngenta/) allowed the identification of new candidate seed BACs. Moreover a fosmid library has been created as a part of the project and the sequencing of fosmid-ends has been launched in the last few months, with our laboratory contributing to the project.

Regarding to the tomato genome project, the advantages of the clone-based strategy are obscured by the difficulties and the intensive work required for clones validation. With the availability of new generation sequencing technologies, including 454/Roche's sequencer FLX, Solexa's Sequencing System and ABI's SOLiD, I'm wondering if a more suitable strategy for completing the tomato genome could be a whole genome shotgun sequencing. The genome sequence could be completed combining a high coverage whole-genome-shotgun with the already available BACs and the genetic and physical maps. Furthermore, the pair-end information obtained from BAC-ends and fosmid-ends could contribute to the whole genome assembly, together with high-density pair-ends produced by the new generation sequencers. Moreover, the recent completion of the grapevine genome by the French-Italian Consortium demonstrates that even a whole-genome-shotgun initiative can be divided between different genomic centres.

I.

References

Allemand JF, Bensimon D, Jullien L, Bensimon A, Croquette V. (1997). **pH-dependent specific binding and combing of DNA.** Biophysical journal 73(4):2064-2070.

Allen JE, Salzberg SL. (2005). **JIGSAW: integration of multiple sources of evidence for gene prediction.** Bioinformatics *21*(18):3596-3603.

Andrews PD, Knatko E, Moore WJ, Swedlow JR. (2003). **Mitotic mechanics: the auroras come into view.** Current opinion in cell biology *15*(6):672-683.

Arabidopsis Genome Initiative. (2000). **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** Nature *408*:796-815.

Arumuganathan, K and Earle, ED. (1991). **Estimation of nuclear DNA content of plants by flow cytometry.** Plant Molecular Biology Reporter *9*:229-233.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, *et al.* (2005). **The Universal Protein Resource (UniProt).** Nucleic Acids Research *33*:D154-159.

Batzoglou S, Berger B, Mesirov J, Lander ES. (1999). **Sequencing a genome by walking with clone-end sequences: a mathematical analysis.** Genome Research *9*(12):1163-1174.

Birney E, Clamp M, Durbin R. (2004). **GeneWise and Genomewise.** Genome Research *14*(5):988-995.

Budiman MA, Chang SB, Lee S, Yang TJ, Zhang HB, de Jong H, Wing RA. (2004). **Localization of jointless-2 gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping.** Theoretical and applied genetics *108*(2):190-196.

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 Research *10*(1):129-136.

Campagna D, Romualdi C, Vitulo N, Del Favero M, Lexa M, Cannata N, Valle G. (2005). **RAP: a new computer program for de novo identification of repeated sequences in whole genomes.** Bioinformatics *21*(5):582-588.

Carmena M, Earnshaw WC. (2003). **The cellular geography of aurora kinases.** Nature reviews Molecular cell biology *4*(11):842-854.

Chen JM, Dando PM, Stevens RA, Fortunato M, Barrett AJ. (1998). **Cloning and expression of mouse legumain, a lysosomal endopeptidase.** The Biochemical journal *335* (1):111-117.

Conti C, Bensimon A. (2002). **A combinatorial approach for fast, high-resolution mapping.** Genomics *80*(2):135-137.

Demidov D, Van Damme D, Geelen D, Blattner FR, Houben A. (2005). **Identification and dynamics of two classes of aurora-like kinases in Arabidopsis and other plants.** The Plant cell *17*(3):836-848.

Demidov D, Van Damme D, Geelen D, Blattner FR, Houben A. (2005). **Identification and dynamics of two classes of aurora-like kinases in Arabidopsis and other plants.** The Plant Cell *17*(3):836-848.

Doganlar S, Frary A, Daunay MC, Lester RN, Tanksley SD. (2002). **A comparative genetic linkage map of eggplant (Solanum melongena) and its implications for genome evolution in the solanaceae.** Genetics *161*(4):1697-1711.

Engler FW, Hatfield J, Nelson W, Soderlund CA. (2003). **Locating sequence on FPC maps and selecting a minimal tiling path.** Genome Research *13*(9):2152-2163.

Eshed Y, Zamir D. (1995). **An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL.** Genetics *141*(3):1147-1162.

Ewing B, Green P. (1998). **Base-calling of automated sequencer traces using phred. II. Error probabilities.** Genome Research *8*(3):186-194.

Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ. (2004). **Comprehensive EST analysis of tomato and comparative genomics of fruit ripening.** The Plant Journal *40*(1):47-59.

Fransz PF, Alonso-Blanco C, Liharska TB, Peeters AJ, Zabel P, de Jong JH. (1996). **High-resolution physical mapping in Arabidopsis thaliana and tomato by fluorescence in situ hybridization to extended DNA fibres.** The Plant journal *9*(3):421-430.

Ganal M, Young ND, Tanksley SD. (1989). **Pulsed field gel electrophoresis and phyiscal mapping of large DNA fragments in the Tm-2a region of chromsosome 9 in tomat**o. Molecular and general genetics *215*:395-400.

Ganal MW, Czihal R, Hannappel U, Kloos DU, Polley A, Ling HQ.(1998). **Sequencing of cDNA clones from the genetic map of tomato (Lycopersicon esculentum).** Genome Research *8*(8):842-7.

Ganal MW, Lapitan NL, Tanksley SD. (1991). **Macrostructure of the tomato telomeres.** The Plant Cell *3*(1):87-94.

Ganal MW, Tanksley SD. (1989). **Analysis of tomato DNA by pulsed field gel electrophoresis.** Plant Molecular Biology Reporter *7*(1): 17-27.

Gordon D, Abajian C, Green P. (1998). **Consed: a graphical tool for sequence finishing.** Genome Research *8*(3):195-202.

Gordon D, Abajian C, Green P. (1998). **Consed: a graphical tool for sequence finishing.** Genome Research *8*(3):195-202.

Gordon D, Desmarais C, Green P. (2001). **Automated finishing with autofinish.** Genome Research *11*(4):614-25.

Green ED. (2001). **Strategies for the systematic sequencing of complex genomes.** Nature reviews Genetics *2*(8):573-583.

Guigo R. (1998). **Assembling genes from predicted exons in linear time with dynamic programming.** Journal of Computational Biology *5*:681-702.

Guindon S, Gascuel O. (2003). **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** Systematic Biology *52*(5):696-704.

Hanahan D, Joel Jessee J, Bloom FR. (1999). **Plasmid transformation of Escherichia coli and other bacteria.** Methods in Enzymology *204*:63-113.

Hara-Nishimura I, Hatsugai N, Nakaune S, Kuroyanagi M, Nishimura M. (2005). **Vacuolar processing enzyme: an executor of plant cell death.** Current Opinion in Plant Biology *8*(4):404-408.

International Rice Genome Sequencing Project. (2005). **The map-based sequence of the rice genome.** Nature *436*(7052):793-800.

Jackson SA, Cheng Z, Wang ML, Goodman HM, Jiang J. (2000). **Comparative fluorescence in situ hybridization mapping of a 431-kb Arabidopsis thaliana bacterial artificial chromosome contig reveals the role of chromosomal duplications in the expansion of the Brassica rapa genome.** Genetics *156*(2):833-838.

Jackson SA, Wang ML, Goodman HM, Jiang J. (1998). **Application of fiber-FISH in physical mapping of Arabidopsis thaliana.** Genome *41*(4):566-572.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, *et al.* (2007). **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** Nature *449*(7161):463-467.

Kawabe A, Matsunaga S, Nakagawa K, Kurihara D, Yoneda A, Hasezawa S, Uchiyama S, Fukui K. (2005). **Characterization of plant Aurora kinases during mitosis.** Plant molecular biology *58*(1):1-13.

Kent WJ. (2002). **BLAT--the BLAST-like alignment tool.** Genome Research *12*(4):656-64.

Kinoshita T, Yamada K, Hiraiwa N, Kondo M, Nishimura M, Hara-Nishimura I. (1999). **Vacuolar processing enzyme is up-regulated in the lytic vacuoles of vegetative tissues during senescence and under various stressed conditions.** The Plant journal *19*(1):43-53.

Korf I. (2004). **Gene finding in novel genomes.** BMC Bioinformatics *5*:59.

Ku HM, Vision T, Liu J, Tanksley SD. (2000). **Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny.** Proceedings of the National Academy of Sciences of the United States of America *97*(16):9121-9126.

Kulikova O, Gualtieri G, Geurts R, Kim DJ, Cook D, Huguet T, de Jong JH, Fransz PF, Bisseling T. (2001). **Integration of the FISH pachytene and genetic maps of Medicago truncatula.** The Plant journal *27*(1):49-58.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. (2004). **Versatile and open software for comparing large genomes.** Genome Biology *5*(2):R12.

Lebofsky R, Bensimon A. (2003). **Single DNA molecule analysis: applications of molecular combing.** Briefings in functional genomics and proteomics *1*(4):385-396.

Li L, Yang J, Tong Q, Zhao L, Song Y. (2005). **A novel approach to prepare extended DNA fibers in plants.** Cytometry Part A *63*(2):114-117.

Lin C, Mueller LA, Mc Carthy J, Crouzillat D, Pétiard V, Tanksley SD. (2005). **Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts.** Theoretical and applied genetics *112*(1):114-130.

Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK. (1999). **Genome mapping in capsicum and the evolution of genome structure in the solanaceae.** Genetics *152*(3):1183-1202.

Lockton S, Gaut BS. (2005). **Plant conserved non-coding sequences and paralogue evolution.** Trends in genetics *21*(1):60-65.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH. (1997). **High throughput fingerprint analysis of large-insert clones.** Genome Research *7*(11):1072-1084.

Monier K, Heliot L, Rougeulle C, Heard E, Robert-Nicoud M, Vourc'h C, Bensimon A, Usson Y. (2001). **Improvement of FISH mapping resolution on combed DNA molecules by iterative constrained deconvolution: a quantitative study.** Cytogenetics and Cell Genetics *92*(1-2):59-62.

Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, *et al*. 2005. **The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond.** Plant Physiology *138*:1310-1317.

Ouyang S, Buell CR. (2004). **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants.** Nucleic Acids Research *32*:D360-363.

Peterson DG, Pearson WR, Stack SM. (1998). **Characterization of the tomato (Lycopersicon esculentum) genome using in vitro and in situ DNA reassociation.** Genome *41*:346–356.

Peterson DG, Price HJ, Johnson JS, Stack SM. (1996). **DNA content of. heterochromatin in tomato (Lycopersicon esculentum) pachytene chromosomes.** Genome *39*:77-82.

Sherman JD, Stack SM. (1989). **Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (Lycopersicon esculentum).** Genetics *141*(2):683-708.

Shirahama-Noda K, Yamamoto A, Sugihara K, Hashimoto N, Asano M, Nishimura M, Hara-Nishimura I. (2003). **Biosynthetic processing of cathepsins and lysosomal degradation are abolished in asparaginyl endopeptidase-deficient mice.** The Journal of biological chemistry *278*(35):33194-33199.

Siegel AF, Trask B, Roach JC, Mahairas GG, Hood L, van den Engh G. (1999). **Analysis of sequence-tagged-connector strategies for DNA sequencing.** Genome Research *9*(3):297-307.

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J. (2001). **Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats.** Proceedings of the National Academy of Sciences of the United States of America *98*(9):5099-5103.

Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, *et al.* 1992. **High density molecular linkage maps of the tomato and potato genomes.** Genetics *132*:1141-1160.

Thompson JD, Higgins DG, Gibson TJ. (1994). **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** Nucleic Acids Research *22*: 4673–4680.

Todesco S, Campagna D, Levorin F, D'Angelo M, Schiavon R, Valle G, Vezzi A. (2008). **PABS: An online platform to assist BAC-by-BAC sequencing projects.** BioTechniques *44*(1):60-64.

Tor M, Manning K, King GJ, Thompson AJ, Jones GH, Seymour GB, Amstrong SJ. (2002). **Genetic analysis and FISH mapping of the Colourless non-ripening locus of tomato.** Theoretical and applied genetics *104*:165-170.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, *et al.* (2006). **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** Science *313*(5793):1596-1604.

van Der Hoeven RS, Monforte AJ, Breeden D, Tanksley SD, Steffens JC. (2000). **Genetic control and evolution of sesquiterpene biosynthesis in Lycopersicon esculentum and L. hirsutum.** The Plant Cell *12*(11):2283-2294.

van der Knaap E, Sanyal A, Jackson SA, Tanksley SD. (2004). **High-resolution fine mapping and fluorescence in situ hybridization analysis of sun, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements.** Genetics *168*(4):2127-2140.

Wang Y, Tang X, Cheng Z, Mueller L, Giovannoni J, Tanksley SD. (2006). **Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome.** Genetics *172*(4):2529-2540.

Wang Y, van der Hoeven RS, Nielsen R, Mueller LA, Tanksley SD. (2005). **Characteristics of the tomato nuclear genome as determined by sequencing undermethylated EcoRI digested fragments.** Theoretical and applied genetics *112*(1):72-84.

Weier HU. (2001). **DNA fiber mapping techniques for the assembly of high-resolution physical maps.** The journal of histochemistry and cytochemistry *49*(8):939-948.

Wendl MC, Marra MA, Hillier LW, Chinwalla AT, Wilson RK, Waterston RH. (2001). **Theories and applications for sequencing randomly selected clones.** Genome Research *11*(2):274-280.

Wu TD, Watanabe CK. (2005). **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** Bioinformatics *21*(9):1859-1875.

Yamada K, Shimada T, Kondo M, Nishimura M, Hara-Nishimura I. (1999). **Multiple functional proteins are produced by cleaving Asn-Gln bonds of a single precursor by vacuolar processing enzyme.** The Journal of biological chemistry *274*(4):2563-2570.

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D. (2005). **Expressed sequence tags from the laboratory-grown miniature tomato (Lycopersicon esculentum) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars.** Gene *356*:127-134.

Yang TJ, Lee S, Chang SB, Yu Y, de Jong H, Wing RA. (2005). **In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotranposons.** Chromosoma *114*(2):103-117.

Zakharov A, Müntz K. (2004). **Seed legumains are expressed in stamens and vegetative legumains in seeds of Nicotiana tabacum L.** Journal of experimental botany *55*(402):1593-1595.

Zhong XB, Fransz PF, Wennekes-Eden J, Ramanna MS, van Kammen A, Zabel P, Hans de Jong J. (1998). **FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato.** The Plant journal : for cell and molecular biology *13*(4):507-517.

Zwick MS, Hanson RE, McKnight TD, Islam-Faridi MN, Stelly DM, Wing RA. (1997). **A rapid procedure for the isolation of Cot-1 DNA from plants.** Genome *40*:138-142.

II.

Appendix

b

Table 1. Accession number of the proteins used for the construction of the Aurora kinases phylogenetic tree.

Protein	Accession number
A. thaliana AtAurora1	BAE00019
A. thaliana AtAurora2	BAE00020
A. thaliana AtAurora3	BAE00021
M. truncatula 1	ABE85513
M. truncatula 2	ABE80792
M. truncatula 3	ABE90417
O. sativa OsAurora1	BAE00022
O. sativa OsAurora2	BAE00023
P. trichocarpa 1 ⋆	estExt_fgenesh4_pm.C_LG_VI0258
P. trichocarpa 2 ⋆	estExt_fgenesh4_pg.C_LG_VI1609
P. trichocarpa 3 ⋆	eugene3.00180136
S. tuberosum 1 ▲	SGN-U277473
S. lycopersicum 1	C04HBa0049A17_C04SLm0040B16.2
S. lycopersicum 2	C04HBa0114G11_C04HBa0050I18_C04HBa0036C23_C04HBa0008H22.39
S. lycopersicum 3	C04HBa0289C05_C04SLm0059M16.23
S. lycopersicum 4	C08HBa00069E09_1.14
S. lycopersicum 5	C12HBa0133N05_C12HBa0163O04.22
S. lycopersicum 6	C12HBa0326K10.8
V. vinifera 1 ⋆⋆	GSVIVP00026259001
V. vinifera 2 ⋆⋆	GSVIVP00032134001

▲ sequences derived from EST

⋆ gene model names in the *Populus trichocarpa* genome browser v. 1.1 (http://genome.jgi.-psf.org/Poptr1/Poptr1.home.html)

⋆⋆ gene model names in the *Vitis vinifera* repository v 1 (http://www.vitisgenome.it/)

Table 2. Accession number of the proteins used for the construction of the VPE proteins phylogenetic tree.

Protein	Accession number
A. thaliana α-VPE	NP_180165
A. thaliana β-VPE	NP_176458
A. thaliana γ-VPE	NP_195020
A. thaliana δ-VPE	NP_188656
A. thaliana 1	NP_563825
A. thaliana 2	NP_849616
A. thaliana 3	NP_973794
M. truncatula 1	ABE91043
M. truncatula 2	ABE93501
N. tabacum NtVPE-1a	BAC54827
N. tabacum NtVPE-1b	BAC54828
N. tabacum NtVPE-2	BAC54829
N. tabacum Nt-VPE-3	BAC54830
N. tabacum NtPB1	CAB42650
N. tabacum NtPB2	CAB42651
N. tabacum NtPB3	CAE84598
O. sativa 1	AP008207
O. sativa 2	AP008208
O. sativa 3	AP008210
O. sativa 4	AP008211
O. sativa 5	NP_001046312
P. hybrida 1 ▲	SGN-U207511
P. trichocarpa 1 ★	gw1.127.139.1
P. trichocarpa 2 ★	gw1.201.52.1
P. trichocarpa 3 ★	gw1.VIII.2629.1
P. trichocarpa 4 ★	grail3.0013022501
P. trichocarpa 5 ★	estExt_Genewise1_v1.C_LG_XVIII0730
P. trichocarpa 6 ★	estExt_fgenesh4_pg.C_LG_III0908
S. tuberosum 1▲	SGN-U268931
S. tuberosum 2▲	SGN-U268932
S. lycopersicum 1	C08HBa0086C12_1.9
S. lycopersicum 2	C08HBa0086C12_1.10
S. lycopersicum 3	C08HBa0197A05_1.3
S. lycopersicum 4	C08HBa0197A05_1.7
S. lycopersicum 5	C08HBa0197A05_1.8
S. lycopersicum 6	C08HBa0197A05_1.10
S. lycopersicum 7	C08SLm0012O12_1.5
S. lycopersicum 8	C08SLm0019J03_1.9
S. lycopersicum 9	C08SLm0019J03_1.11
S. lycopersicum 10	C12HBa0326K10.4
V. vinifera 1 ★★	GSVIVP00032155001
V. vinifera 2 ★★	GSVIVP00029598001
V. vinifera 3 ★★	GSVIVP00012129001
V. vinifera 4 ★★	GSVIVP00007285001

▲ sequences derived from EST

★ gene model names in the *Populus trichocarpa* genome browser v. 1.1 (http://genome.jgi.-psf.org/Poptr1/Poptr1.home.html)

★★ gene model names in the *Vitis vinifera* repository v 1 (http://www.vitisgenome.it/)

Benchmarks

PABS: An online platform to assist BAC-by-BAC sequencing projects

Sara Todesco[1], Davide Campagna[2], Fabrizio Levorin[2], Michela D'Angelo[2], Riccardo Schiavon[2], Giorgio Valle[1,2], and Alessandro Vezzi[1]

[1]*Department of Biology, University of Padova and* [2]*CRIBI Biotechnology Centre, University of Padova, Padova, Italy*

BioTechniques 44:60-64 (January 2008)
doi 10.2144/000112686

Genome sequencing projects are either based on whole genome shotgun (WGS) or on a BAC-by-BAC strategy. Although WGS is in most cases the preferred choice, sometimes the BAC-by-BAC approach may be better because it requires a much simpler assembly process. Furthermore, when the study is limited to specific regions of the genome, the WGS would require an unjustified effort, making the BAC-by-BAC the only feasible strategy. In this paper we describe an informatics pipeline called PABS (Platform Assisted BAC-by-BAC Sequencing) that we developed to provide a tool to optimize the BAC-by-BAC sequencing strategy. PABS has two main functions: (i) PABS-Select, to choose suitable overlapping clones; and (ii) PABS-Validate, to verify whether a BAC under analysis is actually overlapping the neighboring BAC.

The whole genome shotgun (WGS) strategy (1) is in most cases the preferred choice for genomic sequencing; however, in some cases the BAC-by-BAC approach (2) may be a better choice, especially when complex repeated regions must be resolved or when the study is limited to specific regions of the genome.

The BAC-by-BAC strategy consists of shotgun sequencing of individual adjacent BACs that cover the region of interest with a minimal but at the same time significant overlap between clones. To generate the minimal "tiling path," two approaches have been proposed (3): (*i*) the physical mapping approach, which requires the complex and laborious construction of a physical map (typically by BAC fingerprinting) to sort and select a series of clones (the "tiling path") before starting the sequencing process; and (*ii*) the walking approach, which requires direct sequencing without a priori knowledge of the clone position in the genome. In the latter case, the BAC library must be characterized by sequencing the ends of each insert, resulting in a database of BAC-end sequences (BES). After sequencing a BAC, it is possible to identify all the overlapping BES. Therefore the walking can start from "seed" BACs to extend bidirectionally on overlapping clones identified by their BES.

A key step in the BAC-by-BAC sequencing is the identification of reliable neighboring BACs. Often this process is difficult due to the presence of repeats, leading to misalignment of BACs and possible "jumps" along the genome. The analysis of repeats can be performed using RepeatMasker (www.repeatmasker.org) or similar tools able to identify known repeats. However, for those genomes not yet extensively studied, the repeated regions are not well characterized and their direct identification is impossible.

In this paper we describe the implementation of PABS for the International

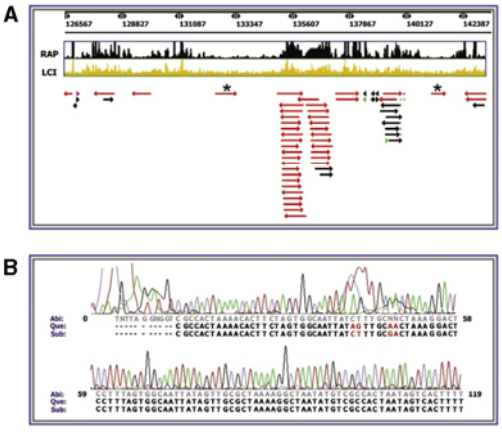


Figure 1. Screenshots from PABS-Select. (A) After uploading the initial sequence (typically the sequence of the BAC or the end to be extended) the application returns a graphical representation of the sequence, including the Repeat Analysis Program (RAP) Index (reflecting the repetitiveness of a region) and the Low Complexity Index (LCI, indicating the presence of low complexity regions such as homopolymers and microsatellites). To simplify the figure, only the terminal 16 kb of a 143 kb BAC insert are shown. The entire database of BAC-end sequences (BES) is preloaded on the system, thus allowing an automatic BLASTn search to align on the initial BAC all the matching BES, represented by arrows in the figure. This gives an immediate view of the possible overlapping BACs, the arrows pointing to the direction of the overlap. The extent of each arrow represents the region of overlap, while the color indicates the BLASTn score: red = >200; violet = 200–80; green = 80–50; blue = 50–40 and black = <40. The final aim is to find at each end of the input sequence a suitable overlapping BAC. Therefore, the best candidates will be those corresponding to BES with the following features: (*i*) direction toward the end of the initial BAC; (*ii*) position in a region with low RAP and LCI indexes; and (*iii*) appropriate extent of the overlap. The asterisks indicate two suitable candidates. By clicking on an arrow, the BES electropherogram aligned to the input sequence is displayed, as partially shown in (B). The query sequence (Que) corresponds to the initial BAC taken as input, while the subject (Sub) is the aligned BES as stored in the database. Moreover, the "Abi" sequence refers to the same BES, generated with the standard Applied Biosystems (Foster City, CA, USA) base caller. This allows an accurate inspection of any discrepancy between the two aligned sequences; for instance, the mismatching bases between query and subject (red colored) would indicate considerably different sequences, but the analysis of the electropherogram shows a likely perfect match of the two sequences.

e

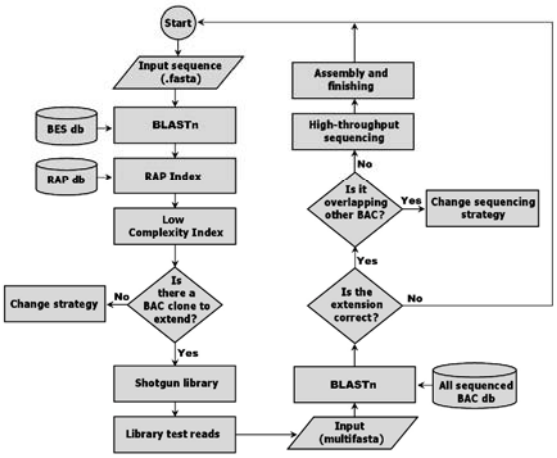


Figure 2. Schematic overview of the dataflow used in PABS. Databases are drawn as bins, rectangles represent applications; direction of dataflow is indicated by connectors. The identification of candidate extension clone is based on BLASTn analysis of the input sequence against the BAC-end sequences (BES) database, and on calculation of Repeat Analysis Program (RAP) Index and Low Complexity Index. The candidate BAC clones are then shotgun sequenced. A first set of 96 clones from the shotgun library is sequenced and a multifasta format of these is processed by the PABS-Validate, using BLASTn against different types of databases.

Tomato Genome Project. The project is based on a BAC-by-BAC sequencing strategy and relies on a BES database (more than 310,000 sequences), but lacks a robust physical map (4,5). Our group is involved in the sequencing of chromosome 12; at the time of writing, it has successfully used PABS for 33 rounds of walking, without any error in the extension process.

PABS uses BLASTn (6) to analyze a fully or partially sequenced clone (hereafter referred to as "initial BAC") against the BES database. PABS-Select takes as "input sequence" the initial BAC (either the complete sequence or the end under investigation) and returns a graphical representation of the position and orientation of the BES (represented as oriented arrows) overlapping the input sequence (Figure 1A).

An innovative feature of PABS is its ability to integrate the BES analysis with the presence of repetitive sequences. In particular, PABS identifies repeated regions with the Repeat Analysis Program (RAP) (7) and calculates the Low Complexity Index as one minus the Linguistic Complexity Index (8). The RAP Index gives an estimate of the "repetitiveness" of a DNA region. It is calculated for each position of the input sequence by means of a de novo analysis that does not require any previous knowledge about repeats. PABS displays the results of BLASTn and RAP, thus allowing a more reliable selection of adjacent clones. The choice will be addressed to BACs with a suitable overlap to the initial BAC and with the aligned BES positioned in a low-repeat region.

To make the selection easier and faster, PABS allows a direct visualization of the BES electropherogram aligned with the input sequence (Figure 1B). In this way the user can quickly evaluate sequences of poor quality that may be the cause of misleading BLASTn results. In addition, an automated procedure collects and summarizes all the available information on the candidate BACs (insert length, genetic markers, FISH data, sequencing status) to optimize the selection for the extension.

The selected BAC is then sequenced with a shotgun approach. To further validate the selection, we have designed PABS-Validate. Typically, the first set of 96 shotgun sequences produced from the selected BAC are submitted as a multifasta file to PABS-Validate and analyzed using BLASTn against three databases: the initial BAC, the finished BACs (i.e., all the finished BACs of the Tomato Genome Project), and the partially sequenced BACs (i.e., the BACs under sequencing). Three types of controls can be made: (*i*) some of the reads should fall into the overlapping region of the initial BAC, thus confirming a correct walking; (*ii*) no reads should significantly match other sequenced BACs belonging to different genomic regions, because this would indicate a possible jump to another region; and (*iii*) as an exception to the previous point, when several extensions are carried out simultaneously from different seeds, we expect that eventually the different walks could merge; therefore we must also consider this event and the consequent possibility to work out the extent of the overlap at the two ends of a bridging BAC.

A complete scheme of the PABS flowchart is represented in Figure 2.

In conclusion, PABS offers two main features:

- it makes the process of generating a reliable minimal tiling path of BACs more robust since it is specifically designed to deal with repetitive sequences;
- it allows a series of validations at the beginning of the shotgun sequencing of each BAC, minimizing the possibility of mistakes and optimizing the merging of overlapping BACs.

PABS is freely accessible at http://tomato.cribi.unipd.it/files/bioinformatics.html, where further detailed instructions are also available. At the moment, the pipeline has been implemented only for the Tomato Sequencing Project but its modular structure would allow easy adaptation to other projects

based on a clone-by-clone sequencing strategy.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This research is supported by the Fondo per gli Investimenti della Ricerca di Base (grant no. RBLA0345SF).

REFERENCES

1. **Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, et al.** 2001. The sequence of the human genome. Science *291*:1304-1351.
2. **Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, et al.** 2001. Initial sequencing and analysis of the human genome. Nature *409*:860-921.
3. **Batzoglou, S., B. Berger, J. Mesirov, and E.S. Lander.** 1999. Sequencing a genome by walking with clone-end sequences: a mathematical analysis. Genome Res. *9*:1163-1174.
4. **Mueller, L.A., T.H. Solow, N. Taylor, B. Skwarecki, R. Buels, J. Binns, C. Lin, M.H. Wright, et al.** 2005. The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol. *138*:1310-1317.
5. **Budiman, M.A., L. Mao, T.C. Wood, and R.A. Wing.** 2000. A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. Genome Res. *10*:129-136.
6. **Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*:3389-3402.
7. **Campagna, D., C. Romualdi, N. Vitulo, M. Del Favero, M. Lexa, N. Cannata, and G. Valle.** 2005. RAP: a new computer program for de novo identification of repeated sequences in whole genomes. Bioinformatics *21*:582-588.
8. **Orlov, Y.L. and V.N. Potapov.** 2004. Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. *32*:W628-W633.

Received 21 September 2007; accepted 26 October 2007.

Address correspondence to Alessandro Vezzi, Department of Biology, University of Padova, via Ugo Bassi 58/B, Padova, Italy. e-mail: sandrin@cribi.unipd.it

g

Università degli Studi di Padova
Dottorato di Ricerca in Fisiologia Molecolare e Biologia Strutturale
Ciclo XX

PhD Student: Sara Todesco
Supervisor: Ch.mo Prof. Giorgio Valle

Tomato (*Solanum lycopersicum*) genome project: sequencing and analysis of chromosome 12

The *Solanaceae* family includes a number of closely related plant species with diverse phenotypes that have been exploited for agronomic, pharmaceutical and ornamental purposes. In 2003 'The International Solanaceae Genome Project' (SOL) launched the initiative to sequence the 220 Mb of euchromatin of the tomato (*Solanum lycopersicum*) genome as the central part of a wider project aiming to increase our knowledge about diversity and adaptation in crop species (http://www.sgn.cornell.edu/solanaceae-project/). The sequencing proceeds on a BAC-by-BAC basis with the 12 chromosomes divided over several genomic laboratories of ten different countries. As a member of the project, the Italian research team is involved in the sequencing of the euchromatin portions of chromosome 12.

During my PhD project, I had the opportunity to face this challenging project from different points of view including molecular, cytogenetic and bioinformatic analysis.

A large part of my effort was focused in setting up a sequencing pipeline and starting the construction of a minimal subset of BAC clones covering the chromosome 12 euchromatin with minimal overlaps. The progress can be viewed through the development of the TPF and AGP files, available from the SGN repository (http://www.sgn.cornell.edu/).

A key step for the success of the sequencing project is the identification of a reliable minimal tiling path of neighbouring BAC clones. To improve this process, I contributed to the development of a informatics pipeline called PABS (Platform Assisted BAC-by-BAC Sequencing), freely available to the community at our web site (http://tomato.cribi.unipd.it/files/bioinformatics.html) (Todesco S. *et al.*, 2008). PABS has been specifically designed to minimize the negative impact of genomic repeats, considering that a repeat element can connect non-contiguous regions of the genome,

leading to misalignment of BACs and possible 'jumps' along the genome. PABS has two main functions: 1) PABS-Select, to choose suitable overlapping clones for the sequencing walk; 2) PABS-Validate to verify whether a BAC under analysis is actually overlapping the preceding BAC.

A BAC-based physical map is a fundamental tool to further assist the sequencing work but also to connect the minimal tiling path of BACs. In my study, I improved the molecular combing technique (Lebofsky R. *et al.*, 2003; Monier K *et al.*, 2001; Allemand JF *et al.*, 1997) for producing multicolour FISH on stretched genomic DNA molecules. This technique allows accurate mapping of BAC clones and precise measurement of physical distances between contigs with a spatial resolution of 1 to 5 kb.

Finally, to explore the data generated by the BAC-by-BAC sequencing I contributed to a preliminary annotation of the tomato BACs sequences. As a result of this analysis, we outlined some features of the gene organization in the tomato genome.

Università degli Studi di Padova
Dottorato di Ricerca in Fisiologia Molecolare e Biologia Strutturale
Ciclo XX

PhD Student: Sara Todesco
Supervisor: Ch.mo Prof. Giorgio Valle

Tomato (*Solanum lycopersicum*) genome project: sequencing and analysis of chromosome 12

La famiglia delle *Solanaceae* comprende oltre 3000 specie di interesse agronomico (pomodoro, patata, melanzana, peperone), farmaceutico (belladonna), ornamentale (petunia) ed anche scientifico come organismi modello (pomodoro, patata, petunia).
Nel 2003 è iniziato il progetto 'The International Solanaceae Genome Project' (SOL), che si è posto come obiettivo lo studio della famiglia delle *Solanaceae* nel tentativo di investigarne i meccanismi di adattamento, di evoluzione, le caratteristiche biochimiche e i sistemi di difesa (http://www.sgn.cornell.edu/solanaceae-project/). Una delle linee di ricerca del progetto SOL è sequenziare il genoma di pomodoro (*Solanum lycopersicum*). Il pomodoro è stato scelto come modello in quanto possiede un genoma diploide relativamente piccolo (950 Mbp per nucleo aploide), un ciclo vitale breve e per il quale sono disponibili mappa fisica basata su BAC (*Bacterial Artificial Chromosome*) e una mappa molecolare con cui poter iniziare il progetto di sequenziamento.
La strategia scelta è di sequenziare solo le circa 220 Mbp di eucromatina 'BAC by BAC', ovvero di selezionare un insieme di cloni BAC ('*minimal tiling path*') che permetta di coprire la porzione di eucromatina con il minor grado di sovrapposizione possibile e il sequenziamento *shotgun* dei singoli cloni.
L'Italia partecipa al progetto internazionale SOL tramite il sequenziamento del cromosoma 12.
Durante il mio progetto di Dottorato, ho avuto la possibilità di affrontare le problematiche emerse nel corso del progetto di sequenziamento da diversi punti di vista, molecolare, citogenetico e bioinformatico. Una componente significativa della mia attività è stata dedicata alla definizione di un insieme di protocolli e procedure necessarie per la gestione del progetto, che hanno consentito l'avvio della costruzione del percorso di cloni BAC ('*minimal tlling path*') sul cromosoma 12.

Con il procedere del progetto, infatti, sono emersi diversi punti deboli per il proseguimento del progetto, quali una mappa fisica e genetica poco accurate e un database di BAC-ends con molte sequenze di bassa qualità.

Allo scopo di assistere la fase critica di scelta del percorso di cloni da sequenziare, ho contribuito alla creazione di un *tool* informatico, PABS (Platform Assisted BAC-by-BAC Sequencing), che abbiamo reso disponibile alla comunità scientifica sul nostro sito web (http://tomato.cribi.unipd.it/files/bioinformatics.html) (Todesco S. *et al.*, 2008). PABS è stato specificatamente progettato allo scopo di minimizzare le possibilità di errore nella scelta dei cloni di estensione derivanti prevalentemente dalla presenza di elementi ripetuti nel menoma.

Inoltre ho applicato la tecnologia del DNA combing (Lebofsky R. *et al.*, 2003; Monier K *et al.*, 2001; Allemand JF *et al.*, 1997) per poter mappare i cloni BAC con una risoluzione nell'ordine di 1-5 kb. Elevate sono le potenzialità di questa tecnica per il mappaggio di cloni BAC e l'orientamento delle isole di sequenziamento che si vengono progressivamente formando,

Infine, ho cercato di dare un senso ai dati di sequenziamento che si stanno accumulando. Ho potuto individuare alcune caratteristiche relativa all'organizzazione genica del genoma di pomodoro attraverso uno studio preliminare di predizione genica e annotazione.