Sequence Assembly

How to obtain sequences for very long stretches of DNA

The Sanger method for DNA-sequencing we discussed in the previous handout analyzes continuous stretches of sequences of up to 1000 bases. But often the pieces of DNA one is interested in may be many thousand times longer than this (e.g., a whole chromosome or a bacterial genome).

This means that the initial full-length stretch of DNA has to be broken up into parts that are of the appropriate size for sequencing. This step is referred to as **library generation**[1]. Then, after the sequence of each of these smaller pieces has been obtained, these stretches of sequence have to be put back together in the correct order to obtain the sequence of the full-length DNA. This second step is called sequence assembly. Library generation and **sequence assembly** are intimately linked parts of what is called a **sequencing strategy**.

There are two prototypical strategies for sequencing long sequences (e.g., whole genomes) each approaching the task in a very different way: **map-based sequencing** and **shotgun sequencing**. The two strategies each have distinct advantages and disadvantages and in practice most large-scale sequencing projects currently use a hybrid of the two strategies.

Map-based sequencing (a.k.a. BAC-to-BAC sequencing)

The first strategy developed for sequencing very long pieces of DNA we will introduce is called map-based sequencing. In this strategy, the initial, very long piece of DNA is systematically broken up into consecutively smaller fragments, all the while mapping the exact location of each of these fragments in the overall sequence. The result of this process is a library of DNA fragments where each fragment is stored in its own test tube and the location of each of the fragments in the initial full-length piece of DNA is known. The fragments generated in this process are then sequenced individually. The resulting fragment sequences can then be assembled in their known order to obtain the sequence of the full-length DNA.

In practice, restriction enzymes or physical shearing are used to break the full-length DNA from several thousands of cells of a sample into large fragments of about 100'000 to 200'000 bp length. These large fragments are inserted into bacterial artificial chromosomes (**BACs**)[2]. The position of each of these large fragments in the initial full-length DNA sequence is determined by **physical mapping** (figure 1). Based on this information, the experimenter then selects a subset of these BACs that covers the entire sequence.

The selected BACs are then each broken up and the insert is isolated and processed further to yield fragments that are about 10'000 bases long and these fragments are cloned into plasmid vectors. Again, physical mapping is used to determine the position of each of these smaller fragments within their respective BAC and again, a subset of fragments spanning the entire BAC is selected for further analysis.

[1] The term "library" is used here to signify that a large body of information is systematically partitioned into books, and the books into pages and kept in an orderly fashion. As we will see, not all sequencing libraries are organized in this orderly manner.

[2] BACs are circular DNA constructs, which allow the insertion and propagation of long stretches of DNA in a similar way that bacterial plasmids allow the propagation of smaller DNA fragments.

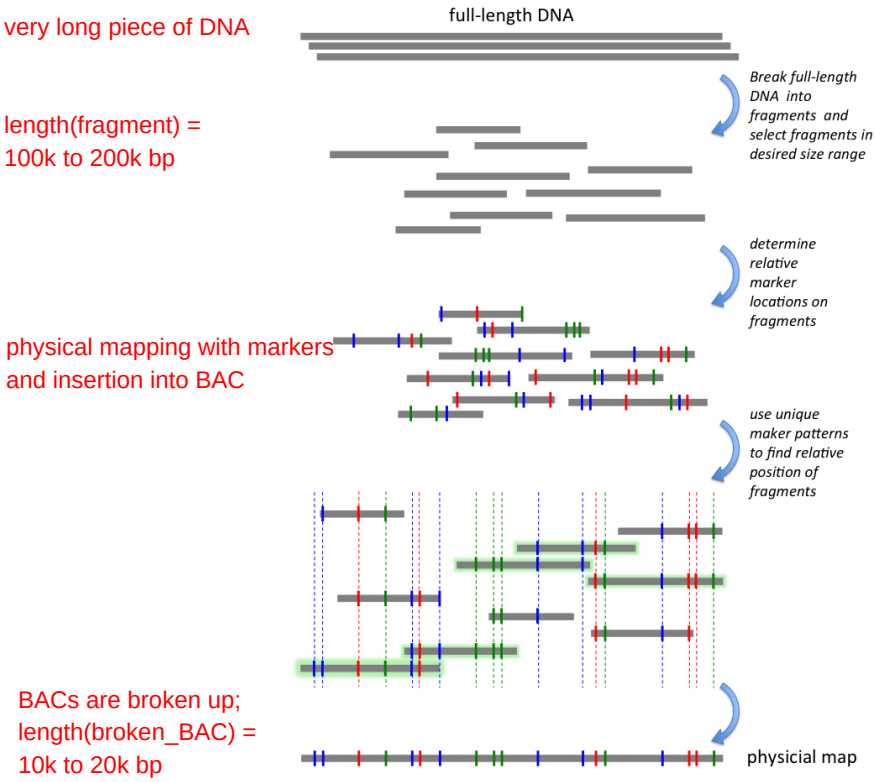

Figure 1: A physical map indicates the position of sequence-specific markers (blue, green, and red) on a continuous piece of DNA (long gray bar). Traditionally, physical maps were often based on naturally occurring restriction-enzyme sites. The positions of these restriction sites were determined by analyzing the sizes of DNA fragments generated by individual and combined digestions with multiple restriction enzymes. Increasingly, fluorescently labeled oligo probes combined with optical-distance measurements are replacing restriction enzymes as the mapping technology of choice. A physical map can be used to determine the relative position of different fragments (short gray bars) within the full-length DNA. Out of all the fragments that were generated, a unique set of fragments (indicated by green glow) spanning the entire DNA sequence is selected. These fragments can then be sequenced or mapped in further detail.

Primer walking can be used to sequence stretches of DNA that are several times longer than the maximal read length of Sanger sequencing

The 10'000-bp fragments in the plasmids are now small enough to be sequenced using a **primer-walking** approach (figure 2). For this the initial sequencing reaction uses a primer that is complementary to a known portion of the plasmid that is directly adjacent to the 5' end of the inserted fragment. The sequence generated in this reaction will then reveal a substantial stretch of the sequence of the fragment. This sequence can then be used to generate the primer for the next sequencing

reaction and so forth. The process is repeated until the appearance of sequence stemming from the plasmid indicates that the entire sequence of the insert has been obtained.

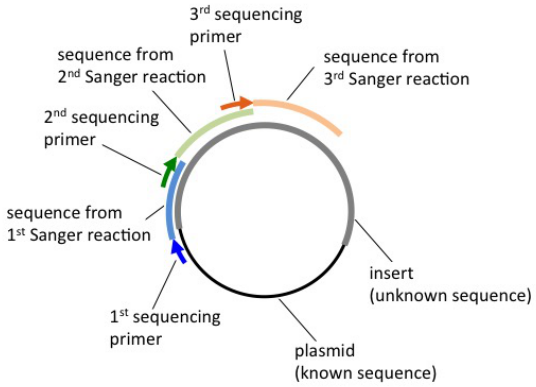

Figure 2: Schematic representation of the primer-walking approach to sequencing stretches of DNA several times longer than the maximal read length from an individual sequencing reaction. The first sequencing reaction uses a primer that is complementary to the portion of the plasmid, for which the sequence is known. This sequencing reaction then progresses into the portion of the plasmid containing the unknown sequence. The end of the sequence determined in this first reaction is used to design the primer for the second reaction and so on.

Map-based sequencing is a labor-intensive, expensive and logistically challenging process

On paper the map-based sequencing approach may appear perfectly straightforward and entirely logical. However, in the laboratory, map-based sequencing on the whole-genome scale turns out to be very labor-intensive and an enormous logistical challenge. Consider the challenge presented by sequencing a diploid human genome containing 6 billion base pairs. These 6 billion base pairs correspond to an absolute minimum of 40'000 BACs that need to be mapped on the genome, a minimum of 600'000 plasmid fragments and over 6 million primers that need to be synthesized.

The reward for all this effort is a very straightforward process, through which the individual fragment sequences can be assembled unambiguously into the full-length genome sequence.

Shotgun sequencing

The alternative to the map-based sequencing strategy is the shotgun strategy (figure 3). In this approach, all of the subcloning, mapping, and primer-walking steps of the map-based sequencing strategy are eliminated. The full-length piece of DNA is mechanically sheared (e.g. by sonication) into fragments that are short enough to be sequenced in one continuous read. These fragments are then sequenced individually.

Obviously, this approach simplifies the library preparation and sequencing process dramatically. But, the resulting library is completely unordered. How can the sequences of the individual fragments in this library be assembled to yield the full-length sequence? The answer lies in the fragment sequences themselves. Because multiple copies of the initial piece of DNA were sheared randomly, most of the

fragments will overlap partially with several of the neighboring fragments. The overlapping portions of the neighboring fragments will therefore have identical sequences. As a result, one can use these overlapping sequence sections to find the neighbors of a particular fragment and then the neighbors of those fragments and so on[3].

Assuming random sequences, an overlap of just 16 bases (4^{16} = 4.3 billion combinations) would be sufficient to uniquely identify a neighboring fragment in a library derived from an entire haploid human genome (3.2 billion bases). With the help of substantial computing power, it should ultimately be possible to assemble the complete sequence for the initial full-length DNA in this way.

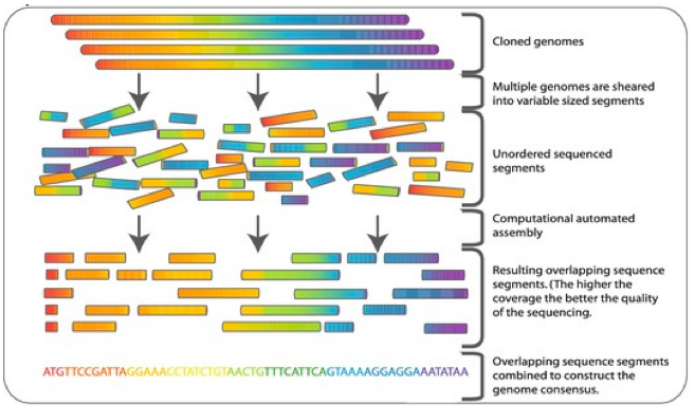

Figure 3: Conceptual workflow for the shotgun sequencing strategy. Multiple copies of the initial full-length DNA (e.g., a full genome) are broken up into small fragments and all fragments are sequenced. Overlaps between the fragment sequences can then be used to assemble the full-length sequence. (adapted from: Commis et al., 2009)

Repeat sequences generate ambiguities in the assembly of shotgun sequencing data

However, many of the DNA sequences of interest (e.g., whole mammalian genomes) contain duplicated genes and long stretches of repeat sequences. The presence of such sequence elements results in ambiguities about the way, in which the fragment sequences should be assembled.

As a result of these ambiguities, it is typically not possible to accurately assemble entire mammalian genome sequences using a pure shotgun approach. The shotgun sequencing data then needs to be complemented by traditional mapping data or special paired-end or mate-pair sequencing data. The latter two are derived from sequencing opposite ends of long fragments with known (approximate) lengths. This then allows the distance between these parts of the fragments to be gauged with reasonable accuracy to either validate an assembly or to reveal areas of the assembly that will need to be investigated further.

Re-sequencing vs. *de novo* sequencing

The map-based and shotgun sequencing strategies (or combinations of the two) allow the sequencing of DNA with a completely unknown sequence. In such cases we speak of de novo sequencing.

[3] The underlying principle of this sequence-overlap-based assembly of fragments is essentially the same as that of the marker-based assembly of large fragments into a physical map. The order and identity of the individual bases take the role of the markers.

Today, most sequencing projects are no longer *de novo* projects but so-called **"re-sequencing"** projects where a genome sequence for a different individual from the same species is already available. This sequence can then be used as a template for the assembly of the individual sequence reads (figure 4).

A typical re-sequencing project would be sequencing the genome of a patient with an inherited disease. We would expect the overall structure of that patient's genome to be very similar to other human genome sequences that are already available. This means that we can use the existing genomes as a template that guides the assembly of the short sequences from individual sequencing reactions into the continuous full-length sequence of the targeted DNA.

Figure 4: In re-sequencing, individual sequence reads are assembled into a full-length sequence by using an already known, highly similar sequence as a template. This template-assisted assembly has been compared to completing a jigsaw puzzle when the final picture is known.

Single-nucleotide polymorphisms or short insertion or deletion mutations (of which there are several millions in a typical human genome) are relatively easy to detect and map using this re-sequencing approach. This is because the overall structure of the template and the new sequence are essentially identical.

By contrast, large-scale structural rearrangements and the elongation (or shortening) of regions with repeat sequences are much harder to detect using this re-sequencing approach, because in these cases the template sequence is, in fact, not correct.