Chapter 4

Genomic assembly using paired-end reads

4.1 Introduction

The *Sequence mapped gap(SMG)* approach was described by Edwards and colleagues [86] as a workable solution to the then-recognized problem of ordering and closing gaps between shotgun-derived DNA sequence fragments. This report was one of the first to suggest the predecessor of mate-pair or paired-end reads as a mechanism for assembling and finishing genomic projects. With the advent of current("next")-generation sequencing methods [87], the mechanism of sequencing DNA has changed significantly, and the mechanism used to generate paired-reads has undergone a commensurate adaptation. We have commercially available technologies that do not rely on cellular cloning and produce 10^5 to 10^8 reads of 500 or 50 nucleotides, respectively.

Today, we make use of mate-pair, or paired-end sequencing to achieve the effect foreseen by Edwards. Each of these sequencing methods produce two sequencing reads from each end of a single DNA fragment or molecular clone. The size of the DNA fragment is usually known within a range of nucleotides, referred to as the *insert* size, and can range between short inserts of 100-200 nucleotides, or long inserts of 6,000-20,000 bases in length. Typical uses for paired-end libraries include: establishing contig order in a scaffold, mapping large rearrangements or insertion/deletions (indels) to a reference genome [88], or establishing sequence within flanking repeats.

Next-generation sequence technologies have become somewhat specialized to optimize the benefits of a specific technology while minimizing its deficiencies. As a result, many of the "short read" technologies are focusing on resequencing efforts, while "long read" technologies are positioned for *de-novo* efforts.

Making use of multiple libraries with different insert sizes offers the advantage of providing fine-scale resolution over non-repetitive sequence with a small-insert library, combined with the ordering, orienting and untangling capability offered by long-insert libraries. This approach may also allow a single sequencing technology to be used for complete genome assembly. Allpaths [89] and Velvet [90] are genomic assemblers that take this approach, while the former also presents the multiple variations present in a genome assembly.

I have combined two distinct sequencing technologies to generate initial contigs, prepare scaffolds, construct draft assemblies and refine a consensus genome. In this chapter, I will describe the underlying sequencing technologies and error modes, and

provide an in-depth discussion of paired-end methods used to construct a draft assembly from contigs generated by 454/Roche sequencing, using paired-end reads established by ABI SOLiD paired-end sequencing .

4.1.1 Next-generation sequencing technologies

4.1.1.1 454 sequencing and error modes

The 454 titanium chemistry is used with GS FLX platform, providing reads of approximately 500 bases in length. Individual DNA fragments are ligated with defined sequence primers at each end. Individual molecules are then captured on sequencing beads and that sequence is amplified using an emulsion PCR reaction that yields a clonal amplification of that sequence over the surface of the sequencing bead. One million beads are then loaded onto a sequencing plate, one bead per well. This technology makes use of pyrosequencing, wherein a microfluidic stream sequentially delivers either ATP, CTP, GTP, or TTP to the individual sequencing reactions utilizing Titanium-Taq DNA polymerase (www.454.com/ products-solutions/how-it-works/sequencing-chemistry.asp). When the appropriate nucleotide arrives, the polymerase extends, yielding a free pyrophosphate(PPi) resulting from incorporation of the nucleotide into the extending chain. This pyrophosphate is detected through a set of enzymatic reactions that combine PPi and 5' adenosine phosphosulfate (APS) to produce ATP. ATP is then detected using luciferase to produce light, proportional to the amount of PPi released during the extension phase. Thus, a polymerase positioned to sequence (N) Cytosine nucleotides in a row will produce (N) PPi molecules which will produce (N) ATP and eventually (N)

photons of light, only when a stream of GTP (Guanosine complements Cytosine) is delivered to the emulsion PCR reaction. This is the method of pyrosequencing, and is the source of its major systematic error, homopolymeric read errors, caused by a difficulty distinguishing light levels with longer repetitions of a single base.

4.1.2 ABI SOLiD sequencing and error modes

ABI SOLiD sequencing utilizes ligation-based sequencing which makes use of fluorescent tagged probe sequences of single-stranded DNA to encode two adjacent bases (dinucleotides) with a fluorescent tag. Each of the 16 possible dinucleotide sequences are encoded with one of four fluorescent tags. Each base is read twice in a shifted reading frame, as a member of a shifted dinucleotide, producing a second encoding which is then combined with the first encoding to produce a "color space" encoding of the sequence. For any given sequence in color-space, there are four possible mappings into base-space. Conversion between color-space and base-space requires knowledge of at least one base in the sequence that the color-space mapping was derived from.

Sequencing errors with SOLiD sequencing are not subject to homopolymeric read uncertainties but may be subject to an intrinsic error rate that is higher than pyrosequencing, especially during de-novo sequencing. Color-space encoding when coupled with a reference genome provides a significant error-correction feature. This mechanism can distinguish read errors from single-nucleotide polymorphisms (SNPs) because a SNP would require adjacent dinucleotide encodings to both change with respect to the reference sequence, while a read error would be expected to have only a single dinucleotide

encoded error. Raw read error rates are not freely available but other reports suggest this may be as high as 3%. If this raw error rate(3%) was correct and we translated 25 base reads to base space, we would expect only 46% of the resulting sequences to be error-free ($(1 - .03)^{25}$). Other studies suggest that only 34% of SOLiD raw reads are alignable to their reference sequence [91]. In cases where no reference is available, such as *de-novo* sequencing, consensus methods must be employed to reduce the overall base error-rate.

4.1.3 Preparation of mate-pairs

The ABI SOLiD version 2 chemistry (Applied Biosystems SOLiD System 2.0 User Guide), when applied in a paired-end, or mate-pair configuration, allows two reads to be generated from the same molecule. Each read is from the same strand and separated by an internal insert of variable size. As is done for the 454 sequencer, genomic DNA is fractured or sheared, but in this case the DNA fragment must be prepared for an eventual cleavage with EcoP15I.

EcoP15I, used for both methylation and cleavage, is a type III restriction modification enzyme that has both a site-specific methylase domain as well as the site-specific, ATP-dependent nuclease. In the presence of S-adenosyl-L-methionine (SAM), the methylase modifies endogenous 5 CAGCAG -3 sites.

Fragmented and end-repaired DNA is methylated with EcoP15I (with SAM and without ATP) to protect sequences of interest from digestion by the restriction enzyme. Methylated DNA is then blunt-end ligated at each end with double-stranded

DNA sequences containing non-methylated 5' CAGCAG -3' restriction sites. These linear molecules are size selected to the desired insert range on an agarose gel, followed by a circularization step that incorporates a central linker with dual 2- base overhangs, compatible with the EcoP15I caps. These circularized elements are then digested with EcoP15I (without SAM and with ATP), cleaving 25 bases downstream of their respective recognition sites, producing linear fragments containing the central linker sequence flanked by the two mate-pairs. Sequencer specific primers are ligated to each end of the resulting digestion product, and the resulting constructs are introduced to the SOLiD sequencing platform, producing two reads of 25 bases in length from the same strand, derived from both ends of the parent molecule.

4.2 Dual method sequencing, assembly and error removal

454 pyrosequencing is particularly well suited to produce high quality contigs in regions with low frequency inversion polymorphisms and with repeat regions smaller than the nominal read length. In the *Pyrobaculum oguniense* assembly, CRISPR arrays approaching twice the read length (450 bases) were correctly assembled in one of three cases (CRISPR1) with an assembled size of 927 bp, suggesting that this may be an upper bound for repetitive region assembly.

ABI SOLiD mate-pair sequencing is well suited to scaffold construction, detection and evaluation of inversion polymorphisms, detection of single nucleotide polymorphisms (SNPs) and small deletion-insertion polymorphisms (DIPs). However, each

of these features relies on an underlying reference sequence.

When the two sequencing methods are combined, we can get the advantages of both while compensating for the shortcomings of each. An assembled "long-read" library generated using 454 pyrosequencing has proven to produce high-quality contigs. These contigs can then be used as the high-quality reference required to attain the advantages of the ABI SOLiD mate-pair method.

Genome assembly proceeds through a process of decontamination, contig assembly, content specific analysis, scaffold assembly using mate-pairs, followed by refinement and assessment of polymorphisms. Initial contig assembly using a 454 "long" library makes use the process described by Roche/454. These initial contigs are then evaluated for possible decontamination and selective removal of contaminating reads by mapping all reads.

Here, I will consider the process of producing a scaffold using Paired-end reads. We begin with a small set of contigs assembled using the Roche/454 assembler (Newbler).

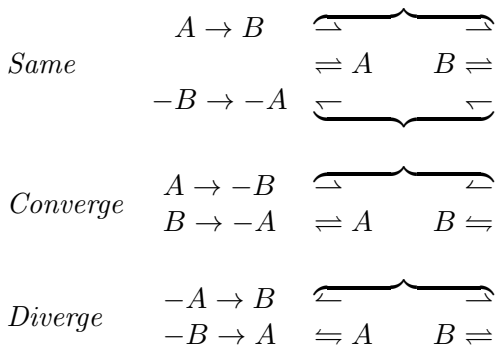
4.2.1 Scaffold mapping using paired-end reads

Paired-end (PE) reads have three properties that are of primary use. For purposes of this discussion, I will be using read pairs prepared for ABI-SOLiD. The conventions with other sequencing technologies are a natural extension of those presented here.

- Both reads are from the same strand,

- the approximate range of distances separating the two reads is known, and
- the forward and reverse reads are identified as such, and as a pair.

Assembled contigs can be from either strand of the parent molecular duplex. When we map PE reads to a set of N contigs, they can be entirely contained within a single contig, or they can span two contigs. When PE reads are completely contained within a single contig and if that contig was properly assembled, then both ends of the read-pair can be expected to be from the same strand. If a PE read-pair spans two contigs, we again know that it came from the same strand in the parent molecule, but contigs can be assembled from either strand. This means that the relationship of a PE read that spans two contigs can tell us about relative contig orientation in one of three relationships, termed *Same, Converge or Diverge* as follows:



These relationships give rise to the *From::To* identity rules of Equations 4.1, 4.2, and 4.3.

$$(a \to b) \implies (-b \to -a) \tag{4.1}$$
$$(a \to -b) \implies (b \to -a) \tag{4.2}$$
$$(-a \to b) \implies (-b \to a) \tag{4.3}$$

We must next map PE reads to assembled contigs. Each of the two sequences in a Paired-end read can be mapped individually to contigs using BLAT [92], followed

by selection of those pairs that have unique mappings to the contig set. We can further qualify read pairs by considering only those pairs that do not violate the insert-size established by initial size selection. Paired reads that both map to a single contig must obey the three properties of specified insert-size, strandedness and order of forward and reverse reads. Pairs that map to two contigs can be qualified by assuring that the summed length of the region downstream(3') of the reverse read and upstream(5') of the forward read are together shorter than the maximum length of the insert-size. The notion of upstream and downstream are always relative to the read itself. If the read mapped to the negative strand of the contig then the implied negative strand of that contig is evaluated.

We then consider the resulting PE mappings to both the contigs involved and their implied orientation to each other. That data is summarized in an $N \times N$ table similar to (Table A.1). Counts of associated relationships between an arbitrary contig-pair[c(relationship)] yield the following triad.

$$(c(Same), c(Converge), c(Diverge))$$

$$(\rightarrow\rightarrow, \rightarrow\leftarrow, \leftarrow\rightarrow)$$

4.2.1.1 Linear relationships derived from paired-end reads

If we consider possible relationships that might exist among N contigs with evidence from PE mapping derived triads, we can see that a scaffold can be established by selecting a sequenced segment (a contig) and incrementally connecting additional contigs in the order and orientation prescribed by the *From::To* triads. The final scaffold

contig	1	2	3	4
1	(150,0,0)[a]	(100,0,0)		
2		(250,0,0)	(0,100,0)	
3			(350,0,0)	(0,0,200)
4				(450,0,0)
	$1 \rightarrow 2 \rightarrow -3 \rightarrow 4$			

[a] (c(Same), c(Converge), c(Diverge)),
$(\rightarrow\rightarrow, \rightarrow\leftarrow, \leftarrow\rightarrow)$

Table 4.1: From::To mapping of contigs. Triads of counts of uniquely mapped PE reads shown. Identity rules are used to combine PE reads such that all "same" counts are mapped to the top strand, and convergent and divergent counts are mapped such that the "From" contig number(row) smaller or equal to the "To" contig (column). For example $-2 \rightarrow -1$ would be remapped to $1 \rightarrow 2$, and $2 \rightarrow -1$ would be remapped to $1 \rightarrow -2$.

used to establish the genome of *Pyrobaculum oguniense* and *PIV* (Figure 3.2) was derived using 20 contigs and an associated 20×20 matrix of PE *From::To* triads (Table A.1). The 20 contigs resulted from an assembly of 454 derived sequence reads and the 20×20 *From::To* matrix was derived from ABI SOLiD PE reads. We interpret those triads by examining patterns of evidence (triads) to resolve the underlying relationships and parent molecules (chromosomes).

Consider the example of Table 4.1. Triads on the diagonal remind us that PE reads exist within contigs, and those pairs should always be in the same orientation if the contig was assembled correctly. This is exemplified by the cell providing *From::To* counts of contig1 to contig1 of (150,0,0). 150 pairs were found and both the forward and reverse reads aligned to the same strand (either top or bottom strand). Additionally, no

contig	1	2	3	4
1	(150,0,0)[a]		(0,100,0)	
2		(250,0,0)	(0,0,200)	(125,0,0)
3			(350,0,0)	
4				(450,0,0)
	$1 \rightarrow -3 \rightarrow 2 \rightarrow 4$			

[a] (c(Same), c(Converge), c(Diverge)),
$(\rightarrow\rightarrow, \rightarrow\leftarrow, \leftarrow\rightarrow)$

Table 4.2: From::To mapping using the Identity equations. Contigs 2 and 3 are divergently oriented with respect to reach other $(-2 \rightarrow 3)$, and with application of the divergence identity rule, this can be rewritten as $(-3 \rightarrow 2)$, allowing construction of the example scaffold (lower panel).

pairs were found contained within contig1 where the read-mates of a pair were arranged in a convergent or divergent relationship.

We find in this example, that contig1 and contig2 are paired in the *Same* orientation, supported by the triad (100,0,0) at cell (1,2), and that contig2 is paired with contig3 in a *convergent* relationship by the triad (0,100,0) at cell (2,3). Furthermore, contig3 is paired with contig4 in a *divergent* relationship by the triad (0,0,200) at cell(3,4). These data provides us with a scaffold of $1 \rightarrow 2 \rightarrow -3 \rightarrow 4$ (negated contigs denote a reverse complement operation), as shown in the lower panel of Table 4.1.

In the next example (Table 4.2), we again see the diagonal triads confirming proper contig construction, and a linkage of contig1 and contig3 in a *convergent* orientation, established by the (0,100,0) triad. This implies a $1 \rightarrow -3$ orientation. Contig2 maps to both contig3 in a *divergent* relationship $(-2 \rightarrow 3)$ and contig4 in the *same* orientation $(2 \rightarrow 4)$. If we make use of the *divergent* identity rule (Equation 4.3),

contig	1	2	3	4
1	(150,0,0)[a]	(100,0,0,0)		
2		(250,0,0)		
3			(350,0,0)	(0,0,200)
4			(0,100,0)	(450,0,0)
	$1 \rightarrow 2, -3 \circlearrowright 4$			

[a] (c(Same), c(Converge), c(Diverge)), $(\rightarrow\rightarrow, \rightarrow\leftarrow, \leftarrow\rightarrow)$

Table 4.3: From::To mapping with two scaffolds. Two distinct contig sets are shown with no PE evidence between the two sets. Contigs 3 and 4 present a circular independent scaffold while contigs 1 and 2 are in a linear configuration.

$(-2 \rightarrow 3) \Longrightarrow (-3 \rightarrow 2)$, we can produce a final scaffold of $1 \rightarrow -3 \rightarrow 2 \rightarrow 4$ (Table 4.2, lower panel).

With independent molecules, we see (Table 4.3) data supporting linkages of $1 \rightarrow 2, -3 \rightarrow 4$, and $4 \rightarrow -3$. This implies that $1 \rightarrow 2$ is independent of $-3 \rightarrow 4$, but contig4 also links back to contig3 in a convergent relationship. Thus, contigs 1 and 2 are linearly connected, while contigs -3 and 4 are arranged in a circle.

4.2.1.2 Inversion relationships derived from paired-end reads

During the analysis of *P. oguniense*, two forms of inversion were found; the *simple* inversion and *loop* inversion. Tables 4.4 and 4.6 show the pattern of triad evidence that yield evidence of these inversions.

In the *simple* inversion example(Table 4.4), cell(1,2) contains (100,100,0), establishing that contig1 is linked to contig2 in both a *same* and a *convergent* orientation. Cell(2,3) implies $2 \rightarrow -3$ and cell(3,2) has $3 \rightarrow 2$, which by identity 4.1, gives $-2 \rightarrow -3$.