Next Generation Sequencing
Applications

Sylvain Forêt

March 2010

http://dayhoff.anu.edu.au/~sf/next_gen_seq
1. Genome sequencing
2. Transcriptome sequencing
3. Bisulfite sequencing
4. ChIP-seq
1. **Genome sequencing**
   - De novo genome sequencing
   - Genome resequencing

2. **Transcriptome sequencing**

3. **Bisulfite sequencing**

4. **ChIP-seq**
Definition

Sequencing a genome *from scratch*, without any pre-existing template.
Coverage depth

\[ \text{coverage} = a = \frac{NL}{G} \]

where \( N \) is the number of reads, \( L \) the read size, and \( G \) the genome size.

Assuming that reads are uniformly distributed, and ignoring end effects, the probability of a read starting in an interval \([x, x + h]\) is \( h/G \).

The number of reads falling in this interval is this a binomial distribution of mean \( Nh/G \).

For large \( N \) (many reads) and small \( h \) (\( h = L \), reads are small), the number of reads covering a segment of size \( L \) can be approximated with a Poisson distribution of mean \( a \).
### How many sequences?

<table>
<thead>
<tr>
<th>Coverage</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected proportion</td>
<td>0.864</td>
<td>0.981</td>
<td>0.997</td>
<td>0.999</td>
</tr>
<tr>
<td>Expected contig size</td>
<td>1,600</td>
<td>6,700</td>
<td>33,500</td>
<td>186,000</td>
</tr>
</tbody>
</table>

NB: the Poisson approximation usually overestimates the actual proportion covered.
Genome Assembly

Reads

Contigs

Scaffolds

Super-Scaffolds
Alignments: theory

- Aligning 2 sequences of size $n$ has complexity $o(n^2)$
- Aligning $m$ sequences has complexity $o(n^m)$
- $\Rightarrow$ Need faster algorithms

Alignments: heuristics

- Find ‘similar’ reads by looking for common words ($o(n)$)
- Align clusters of similar reads
- Allow for more mismatches at the ends of the reads
Building Scaffolds

Physical map
- For instance: micro-satellites
- One marker on the contig: located
- Two markers on the contig: oriented

Mate pairs
- One mate pair: oriented with other contig
- Can provide accurate distance between contigs
- Long insert libraries (cosmids, fosmids) are usually part of genome sequencing projects
Super-Scaffolds

Any other type of information ...

- Weak matches (eg poor quality reads)
- ESTs
- Protein homology
- Long range PCR
- ...
- Often a manual (and tedious) process
Next Generation Sequencing

Which Technique?
- The curse of repeats and low complexity
- 454 is a reasonable choice
- Other technologies mainly applied to prokaryotes
- However: Panda genome sequencing with Illumina (!!!)
1. **Genome sequencing**
   - De novo genome sequencing
   - Genome resequencing

2. **Transcriptome sequencing**

3. **Bisulfite sequencing**

4. **ChIP-seq**
Genome Resequencing

Definition
Sequencing the genome of a species with a sequenced genome. Reads are *mapped* onto this template, no assembly is involved.
Genome Resequencing

Looking for differences

- Single nucleotide polymorphisms (SNPs)
- Insertions and deletions
- Other molecular markers: micro-satellites, mini-satellites, ...
- Segmental duplications and other genomic re-arrangements
- ...

...
SNPs

Homozygous SNP
Heterozygous SNP

Template
Reads

Source: http://solid.appliedbiosystems.com
Genomic Re-Arrangements

Source: http://solid.appliedbiosystems.com
Applications

Resequencing applications

- Comparing closely related species (e.g., *Homo sapiens* vs *H. neandertalis*)
- Genome wide association studies (GWAS)
- Tumor-associated mutations
- …
Targeted Resequencing

1. DNA extraction
2. DNA fragmentation
3. Hybridize to microarray
4. Elute microarray
5. Sequence
Which Technique?

- To resequence the same species, small reads are more cost-effective.
- For different species, 454 may be preferable.
Resequencing Example: Myeloid Leukaemia

3,813,205 tumour SNVs (Maq15)

2,647,695 well supported SNVs (decision tree)

2,584,418 present in skin (SNPs)

63,277 tumour-specific SNVs

20,440 in non-genic regions

31,632 new SNVs

10,735 intronic

11,192 SNVs in genic regions

241 SNVs in coding sequence

31,645 in dbSNP/Watson/Venter

216 in UTR

181 SNVs predicted to alter gene function (non-synonymous and splice junctions)

60 synonymous

7 unable to be validated (technical failures)

14 validated as germline SNVs (SNPs)

8 validated as somatic SNVs (acquired mutations)

152 validated as wild type (false positives)

From Ley et al, Nature 2008
Resequencing Example: Maternal Blood

From Chiu et al, PNAS 2008
1. Genome sequencing

2. Transcriptome sequencing
   - De novo transcriptome sequencing
   - Transcriptome profiling
   - Differential gene expression

3. Bisulfite sequencing

4. ChIP-seq
De Novo Transcriptome Sequencing

Pros

- ‘Genome of the poor’: Only a small proportion of eukaryotic genomes is protein coding. Therefore sequencing a transcriptome is cheaper than a genome.
- Can give more information than a genome: genes can be hard to predict in silico. Here, no need for prediction.
- Provides access to alternative splicing.

Cons

- No insight into the non-expressed functional elements
- Adequate coverage is difficult for genes expressed at low level
- Long transcripts can be difficult to sequence entirely
Transcriptome Assembly

Assembly

- Same basic procedure as for genomes (reads $\rightarrow$ contigs)
- **BUT:**
  - Genomes are linear segments (or circular)
  - Transcripts are graphs of alternatively spliced exons
  - No assembler can currently handle this
Next Generation Sequencing

Which Technique?

- Longer reads make assembly easier
- Short reads, especially with mate pairs can be useful to complement an existing assembly
1 Genome sequencing

2 Transcriptome sequencing
   - De novo transcriptome sequencing
   - Transcriptome profiling
   - Differential gene expression

3 Bisulfite sequencing

4 ChIP-seq
Combining a high-quality genome assembly with high-throughput transcriptome sequencing has provided unprecedented insight into the complexity of eukaryotic transcriptomes.
Mapping Reads

From Cloonan et al, Nature Methods 2008
<table>
<thead>
<tr>
<th>Read size</th>
<th>M1</th>
<th>M5</th>
<th>M10</th>
<th>M100</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>62%</td>
<td>33%</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>30</td>
<td>73%</td>
<td>20%</td>
<td>5%</td>
<td>2%</td>
</tr>
<tr>
<td>35</td>
<td>79%</td>
<td>17%</td>
<td>4%</td>
<td>2%</td>
</tr>
</tbody>
</table>

From Cloonan et al, Nature Methods 2008
Saturating the Transcriptome

Discovery plot for ESEB SQRL libraries

From Cloonan et al, Nature Methods 2008
Recent Discoveries

Transcriptome profiling breakthroughs

- Alternative splicing: 92-94% of human genes undergo alternative splicing
- Patterns of alternative splicing are highly dynamic
- Discovery of many non-coding RNAs (ncRNA)
Next Generation Sequencing

Which Technique?

- Short reads are more cost-effective
- Mate pairs can improve mapping
- Mate pairs impose restrictions on sequence size
1. Genome sequencing

2. Transcriptome sequencing
   - De novo transcriptome sequencing
   - Transcriptome profiling
   - Differential gene expression

3. Bisulfite sequencing

4. ChIP-seq
Differential Gene Expression

**Definition**

Identifying genes expressed at different levels in different conditions.

Examples:

- Diseased vs healthy
- Treated vs non-treated
- Mutant vs wild-type
- Dose response
- More complex, factorial designs
Differential Gene Expression

Assumptions

- Number of reads from a given transcript is proportional to:
  - molar concentration
  - length of transcript

- A possible unit of measurement is: reads of per kilobase of exon model per million mapped reads (RPKM, Mortazavi et al, Nature Methods 2008)

From Mortazavi et al, Nature Methods 2008
Statistical modelling

The model

- Hypothesis: number of reads mapping to a given gene is a Poisson random variable
- Recall Poisson is the limit of binomial as the number of ‘trials’ gets big but the probability of ‘success’ gets small

\[ \text{bin}(n, p) = \text{Pois}(\mu) \text{ as } n \to \infty, p \to 0, np = \mu \]

Here, \( n \sim 10^8 \),
and for a given gene ‘j’:

\[ p_j = \frac{\text{number of transcripts from gene j in flow cell}}{\text{total number of transcripts in flow cell}} \sim 10^{-3} - 10^{-6} \]

Then number of reads of gene j is Poisson with mean

\[ \mu_j = np_j \sim 10^2 - 10^5 \]
Poisson distribution and empirical distribution

From Marioni et al, Genome Research 2008
Hypothesis testing

- Null hypothesis: \( \mu_{j1} = \mu_{j2} \)
- Alternate hypothesis: \( \mu_{j1} \neq \mu_{j2} \)

Procedure

\[ x_{jk} \sim \text{Pois}(\mu_{jk}) \text{ where } \hat{\mu}_{jk} = C_k p_j \]

Note: \( \hat{\mu} \) means estimate of \( \mu \).

If the reads are distributed randomly amongst the \( N \) samples:

\[ X_j = \sum_{k=1}^{N} \frac{(x_{jk} - \hat{\mu}_{jk})^2}{\hat{\mu}_{jk}} \sim \chi^2_{N-1} \]
Which Technique?

- Multiplexing can be very useful:
  - Technical or biological replicates
  - Complex factorial designs
  - Cost savings
1. Genome sequencing

2. Transcriptome sequencing

3. Bisulfite sequencing
   - CpG methylation
   - Genome-wide CpG profiles

4. ChIP-seq
DNA Methylation

**Biological significance**
- DNA methylation involves the addition of a methyl group to some nucleotides
- Present in all realms of life
- Involved in various functions
- Can be inherited

**DNA methylation in animals**
- Mostly CpG dinucleotides
- Gene silencing (chromatin remodelling)
- Imprinting
- Widespread in mammals
- Involved in a number of diseases: cancer, obesity, ...
- Poorly understood
DNA Methylation: An Example

DNA methylation in the honeybee

- Some insects have a mammalian-like methylase gene set
- For instance, the honeybee
- Workers and queens, same genome
- Dnmt3 knockdown ⇒ queens
- This illustrates the importance of methylation in the integration of environmental clues
1. Genome sequencing

2. Transcriptome sequencing

3. Bisulfite sequencing
   - CpG methylation
   - Genome-wide CpG profiles

4. ChIP-seq
Bisulfite Sequencing

ATCACGATCAGCTCGATA

↓ Treat with bisulfite

ATCACGATTAGGTTTGATA

↓ Sequence

↓ Align to genome
Next Generation Sequencing

Which Technique?

- 454: longer is better (loss of complexity), but more homopolymers
- Short reads are more cost-effective
- Mate pairs can improve mapping
- SOLiD has the advantage of color-space
#### Bisulfite Sequencing Example: Leukaemia

From Taylor et al, Cancer Research 2008
1. Genome sequencing

2. Transcriptome sequencing

3. Bisulfite sequencing

4. ChIP-seq
   - Method
   - Example
   - Ribosome profiling
DNA-Protein Interactions
ChIP-seq

From Mardis, Nature Methods 2007
ChIP-seq

1. Cross-link bound proteins to DNA.
2. Isolate chromatin and shear DNA.
4. Reverse cross-link and digest protein.
5. Ligate P1 and P2 adaptors to construct fragment library.

Source: http://solid.appliedbiosystems.com
1. Genome sequencing

2. Transcriptome sequencing

3. Bisulfite sequencing

4. ChIP-seq
   - Method
   - Example
   - Ribosome profiling
Histone Profiles

From Barski et al, Cell 2007
1. Genome sequencing
2. Transcriptome sequencing
3. Bisulfite sequencing
4. ChIP-seq
   - Method
   - Example
     - Ribosome profiling
### Problems with RNA-based methods

- RNA-based methods (RNA-seq, microarrays, quantitative PCR, ...) provide a proxy to protein concentration
- However, these methods ignore post-transcriptional events
- Ribosome profiling provides a better proxy to protein concentration

### Ribosome profiling

- Technology similar to ChIP-seq
- Measures RNA sequences attached to ribosomes
- Very new, might or might not be practical
Ribosome Profiling

Stalled translation

Active translation

Number of Reads

Location on transcript

Location on transcript