Computational Methods for High-Throughput Omics Data - Genomics

Introduction to NGS Genomics

25.10.2011
Outline

- Sequencing technologies:
  - Sanger, Illumina and other 2nd generation or NGS technologies
  - paired end sequencing, error profiles, quality values
- Applications: RNA-Seq, DNA-Seq, ...
- Computational analysis: Read mapping
  - goal of read mapping, special types of read mapping for special types of reads
  - edit/hamming distance, scores, quality values, mapping qualities
  - problems: sequencing errors, repeats, multi-reads
- Different types of genomic variants:
  - SNPs, small indels, structural variants (SVs), copy number variants (CNVs)
  - functional impact: coding/regulatory region, gene fusions
- Computational analysis: SV detection
  - methods based on read pair, read depth, split-read, de-novo assembly
  - different mapping signatures
Literature & links

Sanger sequencing animation:
   http://www.dnalc.org/resources/animations/sangerseq.html

Illumina sequencing video:
   http://www.youtube.com/watch?v=77r5p8IBwJk

454 sequencing video:
   http://www.youtube.com/watch?v=kYAGFrbl6E

NGS forum:
   www.seqanswers.com

Review papers on wiki page:
- Review of next generation sequencing technologies
- Review of sequence alignment algorithms for NGS
- 2 reviews of structural variant detection methods
Brief history of DNA sequencing

- Sanger sequencing method developed in early 1970s by Frederick Sanger
- First complete genome sequenced in 1981: phage M13
- Late 1980s: first automated Sanger sequencing machines (Applied Biosystems)
- 1990s: sequencing of small genomes and expressed sequence tags (ESTs)
- 2001: First draft sequence of the human genome is published, cost about $1 per base
- Mid 2000s: so-called Next Generation Sequencing (NGS) technologies emerge, e.g. 454 Life Sciences pyrosequencing, Illumina sequencing
- 2008: The 1000 Genomes Project is launched (http://www.1000genomes.org/), cost approaching $1 per Megabase

Problem: NGS produces much shorter reads than Sanger sequencing (50-400bp compared to >1000bp) \(\rightarrow\) mostly reference-guided analyses
Sanger Sequencing

- Chain termination through dideoxynucleotide incorporation
- Fragments with known last nucleotide are size-separated in a gel
- Produces sequencing reads of length ~1 kb
- Used to produce the first draft sequence of the human genome

Sanger sequencing animation:
http://www.dnalc.org/resources/animations/sangerseq.html

The NGS revolution: sequencing-by-synthesis

- 454 pyrosequencing is first NGS technology to become commercially available in 2005
- 300x less expensive than Sanger
- Soon followed by: Illumina (Solexa) reversible terminator sequencing (most popular sequencing technology at the moment) and ABI SOLiD two-base encoding

454 Pyrosequencing

- First NGS technology to be commercially available
- First instrument in 2005: 300x less expensive than Sanger
- Adds only one type of base at once
- Base incorporation emits light → the higher light intensity, the more bases were incorporated

- Sequencing errors mostly indels, especially in homopolymer runs
- Read length ~300-400bp
- ~500K reads per instrument run

http://www.genengnews.com/sequencing/supp_04.htm
Illumina reversible terminator sequencing

- Another order of magnitude less expensive than 454
- Most popular sequencing technology at the moment
- One base incorporated per cycle, color signal to determine which one

- Sequencing errors mostly subsitations
- Read length <30bp in the beginning, now >~100bp
- ~ 300 million reads per instrument run

Sequencing technologies

A bit out-dated:

Table 1 Second-generation DNA sequencing technologies

<table>
<thead>
<tr>
<th>Feature generation</th>
<th>Sequencing by synthesis</th>
<th>Cost per megabase</th>
<th>Cost per instrument</th>
<th>Paired ends?</th>
<th>1° error modality</th>
<th>Read-length</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>Emulsion PCR, Polymerase (pyrosequencing)</td>
<td>~$60</td>
<td>$500,000</td>
<td>Yes</td>
<td>Indel</td>
<td>250 bp</td>
<td>14,20</td>
</tr>
<tr>
<td>Solexa</td>
<td>Bridge PCR, Polymerase (reversible terminators)</td>
<td>~$2</td>
<td>$430,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>36 bp</td>
<td>17,22</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Emulsion PCR, Ligase (octamers with two-base encoding)</td>
<td>~$2</td>
<td>$591,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>35 bp</td>
<td>13,26</td>
</tr>
<tr>
<td>Polonator</td>
<td>Emulsion PCR, Ligase (nonamers)</td>
<td>~$1</td>
<td>$155,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>13 bp</td>
<td>13,20</td>
</tr>
<tr>
<td>HeliScope</td>
<td>Single molecule, Polymerase (asynchronous extensions)</td>
<td>~$1</td>
<td>$1,350,000</td>
<td>Yes</td>
<td>Del</td>
<td>30 bp</td>
<td>18,30</td>
</tr>
</tbody>
</table>

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly out-dated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

Next generation DNA sequencing, Jay Shendure and Hanlee Ji, Nat. Biotech. 2008

Going from Sanger → sequencing-by-synthesis → single-molecule sequencing
Reads and quality values

Each base of a read sequence has a base call quality value:

- Probability $e$ of the base call being wrong (based on signal intensities)
- Log transformed into quality scores:
  \[ Q_{:\text{illumina}} = 10 \log_{10} \left( \frac{e}{1-e} \right) \]
  for low values of $e$ asymptotically identical to
  \[ Q_{\text{phred}} = -10 \log_{10} (e) \]

• Illumina example:

<table>
<thead>
<tr>
<th>Cycle No.</th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>N</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40</td>
<td>-40</td>
<td>-29</td>
<td>-5</td>
<td>-40</td>
<td>-6</td>
</tr>
<tr>
<td>C</td>
<td>-40</td>
<td>40</td>
<td>-40</td>
<td>-5</td>
<td>-40</td>
<td>-3</td>
</tr>
<tr>
<td>G</td>
<td>-40</td>
<td>-40</td>
<td>-35</td>
<td>-5</td>
<td>40</td>
<td>-3</td>
</tr>
<tr>
<td>T</td>
<td>-40</td>
<td>-40</td>
<td>28</td>
<td>-5</td>
<td>-40</td>
<td>-8</td>
</tr>
</tbody>
</table>

→ Called read seq = A C T N G C
Illumina sequencing errors

Mostly miscalls, rarely under- or overcalls (i.e. mostly substitution, rarely insertion or deletion sequencing errors)

Strong quality trend:

Quality of sequencing decreases with increasing cycle number → reads are sometimes trimmed
Paired-ends and Mate-Pairs

Both ends of a fragment are sequenced, different protocols:

- Paired-ends: ~200-500 insert size

- Mate pairs: ~2-5kb insert size
Applications

- DNA-Seq:
  - Whole-genome resequencing
  - Targeted resequencing: genomic variants in specific regions, e.g. exons

- RNA-Seq:
  - mRNA sequencing
  - Small RNA sequencing

- CHiP-Seq:
  - Sequencing of transcription factor binding sites
  - Histone modification profiling

- Bisulfite sequencing:
  - DNA methylation profiling

(… the list goes on)
Fundamental Step: Read Mapping

**Goal:** Map each read to the genomic location it originated from

---

Approximate string matching problem!
And Blast is way too slow...

**Difficulties:**
- Billions of short reads, long genome
- Sequencing errors + genomic variants $\rightarrow$ alignment errors
- Repeats, ambiguous regions $\rightarrow$ multi-reads
Fundamental Step: Read Mapping

**Goal:** Map each read to the genomic location it originated from.

- Approximate string matching problem!
  - And Blast is way too slow...

**Mathematical/computational solutions:**
- Hamming/edit distance alignment, allow up to k errors per read
- Use base quality values to assign ambiguously mapped reads
- Mapping quality: probability of a read being wrongly mapped (Maq)
- Filtering strategies based on index datastructures
Read Mapping Strategies

- **NovoAlign**
- **segemehl**

**Burrows-Wheeler transform based methods**
- **BowTie**
- **BWA**
- **SOAP2**

**build data structure from genome**
- **Razer-S**

**build data structure from reads**
- **ELAND**
- **MAQ**

**simple look-up table**
- **SSAHA**
- **Mosaik**
- **Pattern Hunter**
- **SSAHA2**
- **ZOOM**
- **PerM**

**gapped seeds**
- **SOAP**
- **BLAT**
Read Mapping Strategies: Example

General algorithmic techniques:
- Filtering (avoid having to check each genomic position for a possible match)
- Indexing and hashing → quick access to exact k-mer matches

Basics of Eland-algorithm (allowing up to 2 errors):

- Split each read into 4 segments
- 2 errors can affect at most 2 segments
- At least 2 segments match exactly

Efficient filtering for short reads, but does not extend well to longer reads
RNA-Seq

→ Find out which genes are expressed in a certain tissue at a certain timepoint
→ Quantify gene expression
→ Identify alternative splicing events, i.e. Transcript isoforms

Exon1 Exon2 Exon3

Transcript: Exon1 Exon2 Exon3

Junction reads: GSNAP paper

reads from sample transcriptome

reference genome
RNA-Seq

Find out which genes are expressed in a certain tissue at a certain timepoint
Quantify gene expression
Identify alternative splicing events, i.e. Transcript isoforms

Transcripts:
 Isoform 1: Exon1 Exon2 Exon3
 Isoform 2: Exon1 Exon3

→ Junction reads and exon expression levels are used to estimate isoform-specific expression levels
→ Number of mapped reads are used to estimate expression level
DNA-Seq or Genome resequencing

→ Reference sequence known
→ Identify differences between individuals/cell types, e.g. cancer vs. healthy cells
→ SNPs, short indels, large indels, inversions, translocations, copy number variations

**Variant Detection**

**Goal:** Detect differences between reference and donor DNA
DNA-Seq or Genome resequencing

→ Reference sequence known
→ Identify differences between individuals/cell types, e.g. cancer vs. healthy cells
→ SNPs, short indels, large indels, inversions, translocations, copy number variations

**Goal:** Detect differences between reference and donor DNA
DNA-Seq: Targeted resequencing

Usually of special interest: variants in coding region

→ Targeted resequencing often uses exon capture to reduce sequencing cost

→ Array with exon specific probes to pull out sequence of interest

**Functional impact of small variants:**

SNPs in coding region can be non-synonymous or synonymous, i.e. alter the amino acid that is incorporated or not

Indels in coding region can cause a frameshift and thereby non-sense proteins

Splice sites and regulatory elements can be affected, changing protein sequence or dosage
Types of Structural Variation (SV)

Understand role of SVs in
- Disease
- Complex traits
- Evolution

Difficulty:
They tend to reside in repetitive regions

SVs are classified w.r.t. to a reference genome
BreakSeq paper: detecting SVs and reclassifying acc. to formation mechanism

Definition of SV varies: > 50bp in some articles, >1kb in others

Figure 1 | Classes of structural variation.
Alkan et al. „Genome structural variation discovery and genotyping“
Nature Reviews Genetics, 2011
Special type of SVs: gene fusions

Especially abundant in cancer genome

CimeraScan paper: gene fusions in cancer

Side note: Other experimental SV discovery approaches

Hybridization-based microarray approaches
- Array CGH (BAC/oligo array)
- SNP microarrays

Sequencing approaches
- Fosmid paired-end sequencing
- Next-generation sequencing

Arrays currently still offer higher throughput at lower cost, but this is slowly changing

Experimental SV discovery approaches

**Box 2 | Copy number variant size distribution and discovery methods**

- **NGS:**
  - Less biased towards *a priori* knowledge of reference genome
  - Large range of SV classes can be detected

Alkan et al. „Genome structural variation discovery and genotyping“ Nature Reviews Genetics, 2011
SV sequence signatures (1)

<table>
<thead>
<tr>
<th>SV classes</th>
<th>Read pair</th>
<th>Read depth</th>
<th>Split read</th>
<th>Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deletion</strong></td>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
<td><img src="image4.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Novel sequence insertion</strong></td>
<td><img src="image5.png" alt="Diagram" /></td>
<td>Not applicable</td>
<td><img src="image6.png" alt="Diagram" /></td>
<td><img src="image7.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Mobile-element insertion</strong></td>
<td><img src="image8.png" alt="Diagram" /></td>
<td>Not applicable</td>
<td><img src="image9.png" alt="Diagram" /></td>
<td><img src="image10.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

VariationHunter paper, Lee paper
AGE paper, SRiC paper

Figure 2 | Structural variation sequence signatures. (Part 1)
Alkan et al. „Genome structural variation discovery and genotyping“ Nature Reviews Genetics, 2011
## SV sequence signatures (2)

<table>
<thead>
<tr>
<th>Inversion</th>
<th>Not applicable</th>
<th>Contig/scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Inversion" /></td>
<td><img src="image2" alt="Inversion" /></td>
<td><img src="image3" alt="Inversion" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interspersed duplication</th>
<th></th>
<th>Contig/scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="Interspersed duplication" /></td>
<td><img src="image5" alt="Interspersed duplication" /></td>
<td><img src="image6" alt="Interspersed duplication" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tandem duplication</th>
<th>Contig/scaffold</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="Tandem duplication" /></td>
<td><img src="image8" alt="Tandem duplication" /></td>
</tr>
</tbody>
</table>

**Figure 2** | *Structural variation sequence signatures. (Part 2)*  
Alkan *et al.* „*Genome structural variation discovery and genotyping*“ *Nature Reviews Genetics*, 2011
### Sequencing-based computational methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read pair</td>
<td>Powerful for a wide range of SVs</td>
<td>Not for very small SVs, chimeric reads/fragments lead to wrong mappings</td>
</tr>
<tr>
<td>Read depth</td>
<td>Potential to accurately predict absolute copy numbers of very large SVs</td>
<td>Poor breakpoint resolution, large SVs only, no novel sequence insertions</td>
</tr>
<tr>
<td>Split read</td>
<td>Determines exact size, location and sequence content</td>
<td>Suffers most from hard-to-align repetitive sequence, only detects short insertions</td>
</tr>
<tr>
<td>Assembly</td>
<td>Can theoretically predict all types of SVs</td>
<td>Currently feasible only for local assembly, problematic in repeats</td>
</tr>
</tbody>
</table>
Overlap between methods is low

Indels from 1000 genomes project

Comprehensive methods need to make use of all approaches

Figure 3 | Copy number variant discovery biases.
Alkan et al. „Genome structural variation discovery and genotyping“ Nature Reviews Genetics, 2011
Difficulties/biases

SVs tend to be in repetitive sequence where reads are hard to map

Deletions are „easier“ than insertions → more deletions contained in public databases (e.g. 2:1 ratio in DGV)

With advances in sequencing technology, reads become longer and assembly will become feasible
Conclusions

Different types of sequencing data and applications

→ Different types of read mapping and variant detection algorithms

General difficulties:
- lie in errors and biases of sequencing
- repetitiveness of genome, mapping ambiguities
- computational efficiency