Measuring Mutations: DNA Resequencing

CSCI 2950-C
9/26/2011
Anna Ritz

Today’s Topic

Genome
Millions -billions nucleotides

... CATTACAG ...
... GCTAAGTTAG ...
... ATATAATTAG ...
... CTGGTACCTAG ...
... CATTCAGTAG ...

10-100’s million reads
Reads: 30-1000 nucleotides

Next-generation DNA sequencing

1. “Resequencing” algorithms
   (also RNA-Seq, ChIP-Seq, *-Seq)

2. De novo assembly algorithms
**Genome Sequencing and Comparison**

- **Comparative genomics**
  - Differences between species?

- **Personal genomics**
  - Genetic basis for traits of individuals?

- **Cancer genomics**
  - Which somatic mutations lead to cancer?
  - Which somatic mutations are *shared* by multiple people w/ same cancer?

---

**Resequencing**

- Instead of assembly, can we identify differences between a test genome and a closely related *reference genome*?  
  - Reference: sequence is known and often carefully annotated

- **Human genome** (first draft in 2001)
Resequencing

• Generate reads from test genome. Align to reference genome and identify differences.

Alignment Programs for Resequencing

There are lots of programs that align sequences to a reference.

BLAST (Basic Local Alignment Search Tool) is the most widely-used.

MAQ (Mapping and Assembly with Quality) was used for the BAM file in the homework.
Types of Mutations

- Single Nucleotide Variants
- Copy Number Variants
- Structural Variants

Resequencing

Array Comparative Genomic Hybridization (aCGH)
Resequencing with Paired Reads

How can a mutation happen?

Inherited Germline SVs
The Database of Genomic Variants reports 101,923 SVs (March 2011)

Somatic SVs

CANCER
Mutation in Down Syndrome

Trisomy 21

Copy Number Variants in Autism

[Marshall e al 2007]
**Somatic Mutations and Cancer**

Clonal Theory (Nowell 1976)

- **Founder cell**
- **Passenger mutations**
- **Driver mutation**

```
"typical tumor": ~10 driver mutations
100's – 1000's of passenger mutations
```

**Translocations in Cancer**

- **Leukemia**
- **Breast Cancer**

*The Philadelphia Chromosome and Chronic Myelogenous Leukemia (CML)*

- Normal Chromosomes
- Translocated Chromosomes
- Trans... location
- 9 elongated
- 22 (Philadelphia chromosome)*
Resequencing: Single Nucleotide Variants

Test Genome

Reference Genome

Problems:
1) Sequencing errors
2) Alignment difficulty (repeats)

Resequencing: Copy Number Variants (Deletions)

Test Genome

Reference Genome

Read Depth

(coverage c)

(unaligned)
Resequencing: Copy Number Variants (Duplications)

Test Genome (coverage c)

Reference Genome

Read Depth

Measuring DNA Mutations

Single Nucleotide Variants

Copy Number Variants

Resequencing (Consensus)

Resequencing (Read Depth)

Array Comparative Genomic Hybridization (aCGH)
DNA Microarrays

Microarray

Well

Probe: ATCATG

Array Comparative Genomic Hybridization (aCGH)

Test DNA

Reference DNA

Hybridized to an Array

Copy Number Profile

Problems:
(1) Error in the ratio measurement
(2) Placement of the probes
Array Comparative Genomic Hybridization (aCGH)

Copy Number Analysis
Divide genome into segments of equal copy number

- Deletion
- Amplification
aCGH Segmentation Algorithm #1

- Circular Binary Segmentation (CBS) [Olshen et al. 2004]

Copy Number Profile
\[ X = X_1, X_2, X_3, \ldots, X_{L-2}, X_{L-1}, X_L \]

Q: Are there points \( m \) and \( n \) where mean(\( X_m, \ldots, X_n \)) \( \neq \) mean(\( X_{n+1}, \ldots, X_{m-1} \))?

Use the t-test: Given two normal distributions with equal variance, what is the probability that they have the same mean?

aCGH Segmentation Algorithm #2

Use Hidden Markov Model (HMM) to “parse” sequence of probes into copy number states

Deletion Amplification

Genomic position
Hidden Markov Models

**Alphabet**
\[ \Sigma = \{ b_1, b_2, \ldots, b_M \} \]

**Set of states**
\[ Q = \{ 1, \ldots, K \} \]

**Transition probabilities** \( a_{ij} \)
between any two states

**Start probabilities**
\[ a_{0i}, \ldots, a_{0K} \]

**Emission probabilities** \( e_k(b) \)
within each state

**Memory-less Property**
At each time step \( t \), the only thing that affects future states is the current state \( \pi_t \)

\[
P(\pi_{t+1} = k \mid \text{"whatever happened so far"}) =
\]
\[
P(\pi_{t+1} = k \mid \pi_1, \pi_2, \ldots, \pi_t, x_1, x_2, \ldots, x_t) =
\]
\[
P(\pi_{t+1} = k \mid \pi_t)
\]
An HMM for aCGH data [Guha et al 2008]

States: copy numbers

- Homozygous Deletion (copy = 0)
- Heterozygous Deletion (copy = 1)
- Normal (copy = 2)
- Duplication (copy > 2)

Emissions: Gaussians

- $N(\mu_1, \sigma_1)$
- $N(\mu_2, \sigma_2)$
- $N(\mu_3, \sigma_3)$
- $N(\mu_4, \sigma_4)$

$\mu, \sigma, A$ are hyperparameters

aCGH Segmentation Algorithm #3

[Ritz et al 2010] (from Liu & Lawrence 1999)

Copy number profile $X = (X_1 X_2 \ldots X_n)$

- $k$ segments $\leq k_{max}$
- $X_i \sim N(\mu_k, \sigma^2)$

Binary variable $b_i = 1$ if breakpoint at $i$

Compute $P( b_i = 1 | X )$

Computes $P(b_i = 1)$ over all possible segmentations of the $X$
Resequencing (Read Depth) vs. aCGH (Copy Number Profile)

Finding Recurrent Variants

Goal: Find copy number variants that appear in many individuals with the same type of cancer.

26 lung tumor samples on Chromosome 3
Assumption: CNV alters a locus/gene that lies within the recurrent segment
TMPRSS2-ERG Fusion [Tomlins et al. 2005]

5 Prostate Tumors (UCSF,Baylor)

Identify Recurrent Breakpoints [Ritz et al. 2011]
Where’s the Breakpoint?

There may be many reasonable places for a breakpoint.

CBS: Outputs a single segmentation.

HMMs: computes 
\( P(\text{transition from state } S \text{ to } S' \mid X) \)

Change-Point Algorithms:
Computes \( P(b_i=1 \mid X) \) for all segmentations

Recurrent Breakpoint Detection
[Ritz et. al. 2010]

Bayesian Method to Compute Breakpoint Probabilities

Input: aCGH data

Detect Recurrent Breakpoints

\( P(b_i=1 \mid S_j) \)
TMPRSS-ERG Result [Ritz et al. 2011]

36 Prostate Tumors (UCSF, Baylor)

Breakpoint in 5 Patients

False Discovery Rate Hypothesis Correction: \( p(\text{ERG, TMPRSS2}) = 2.7 \times 10^{-10} \)

TMPRSS-ERG Result for CBS
DNA Mutations

- Single Nucleotide Variants
- Copy Number Variants
- Structural Variants

Resequencing
Resequencing
Array Comparative Genomic Hybridization (aCGH)
Resequencing with Paired Reads

Paired-End DNA Sequencing

- Test Genome (sequence is unknown)
- Fragment DNA
- Sequence Ends
- Concorant pair: C
- Discordant pair: D

Reference Genome (sequence is known)
Aligned distances
### Paired-End DNA Sequencing Platforms

<table>
<thead>
<tr>
<th>Sequencing Platform (Company)</th>
<th>Fragment Length L</th>
<th>Read Length</th>
<th>Primary Error Type</th>
<th>Error (Approximate)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Illumina</strong> (Formerly Solexa)</td>
<td>200-500bp</td>
<td>100-150bp</td>
<td>Substitution</td>
<td>0.1%</td>
</tr>
<tr>
<td><strong>SOLiD</strong> (Life Tech.)</td>
<td>~3Kb</td>
<td>Read 1: 75bp, Read 2: 35bp</td>
<td>A-T bias</td>
<td>0.06%</td>
</tr>
<tr>
<td><strong>454</strong> (Roche)</td>
<td>~3Kb</td>
<td>400-500bp (nearing 800bp)</td>
<td>Indel</td>
<td>1%</td>
</tr>
</tbody>
</table>

[Reference: T. Glenn, Molecular Ecology Resources (2011)]

---

### What does Coverage Mean Now?

**Sequence Coverage:**
- Avg. # of reads that span a base
  - (Avg. # of times a base is sequenced)

**Physical Coverage:**
- Avg. # of fragments that span a base
  - (Avg # of times a base is sequenced OR spanned)

Reference Genome (sequence is known)
Resequencing: Structural Variants

Use multiple measurements to predict real SVs

The issue: Repeats

Where’s the Deletion?

breakpoints (a, b)

\[ L_{\text{min}} \leq (a - x_1) + (y_1 - b) \leq L_{\text{max}} \]
Where’s the Inversion?

• (On board)

Distinguish Simple from Complex Structural Variants

Simple Variant: Cluster
There exists a single breakpoint \((a,b)\)
consistent with all fragments.

Complex Variant: Non-Cluster
No single breakpoint
consistent with all fragments.

Maybe this was a BAD alignment.
Split Reads

Deletion
Test

If is mapped

If is not mapped

Right breakpoint

Left breakpoint

DNA Mutations

Single Nucleotide Variants
Resequencing
Resequencing

Copy Number Variants
Resequencing

Structural Variants

Array Comparative Genomic Hybridization (aCGH)

Resequencing with Paired Reads
Resequencing with Strobes
Strobe Sequencing by Pacific Biosciences

Fragment Length: much longer
Subread Length: a little longer, variable
# Subreads: user defined
Error Rate: much higher

Strobes Generalize Pairs (at a high level)

Resequeencing with Strobes

Test Genome (sequence is unknown) → Fragment DNA → Strobos

Reference Genome (sequence is known)
Discordance for Strobes

**Deletion**

Test Genome

Reference Genome

\[ l \leq x_2 - y_1 \leq u \]
\[ l \leq (x_2 - y_1) + (y_3 - b) \leq u \]

**Inversion**

\[ l \leq (a - y_1) + (b - x_3) \leq u \]
\[ l \leq (y_2 - a) + (x_4 - b) \leq u \]

DNA Sequencing Platforms

<table>
<thead>
<tr>
<th>Sequencing Platform (Company)</th>
<th>Fragment Length ( L )</th>
<th>Read Length</th>
<th>Primary Error Type</th>
<th>Error (Approx)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Illumina</strong> (Formerly Solexa)</td>
<td>200-500bp</td>
<td>100-150bp</td>
<td>Substitution</td>
<td>0.1%</td>
</tr>
<tr>
<td><strong>SOLID</strong> (Life Tech.)</td>
<td>~3Kb</td>
<td>Read 1: 75bp Read 2: 35bp</td>
<td>A-T bias</td>
<td>0.06%</td>
</tr>
<tr>
<td><strong>454</strong> (Roche)</td>
<td>~3Kb</td>
<td>400-500bp (nearing 800bp)</td>
<td>Indel</td>
<td>1%</td>
</tr>
<tr>
<td><strong>Strobe Sequencing</strong> (Pacific Biosciences)</td>
<td>~10Kb</td>
<td>~1Kb Total (2 subreads @ 500bp) (3 subreads @ 300bp etc.)</td>
<td>Indel</td>
<td>15%</td>
</tr>
</tbody>
</table>
Can Strobes Help Detect Repeats?

Subread is “anchored” in a unique region

**Question:** Do strobes provide enough pairing information to detect SVs, despite high error rates?

Predicting Structural Variants with Strobes

[Ritz et al 2009]

1. Formalize the problem
2. Construct a graph representing all possible solutions
3. Solve an optimization problem on the graph
1. Formalize the problem

(Ambiguous alignments due to repetitive regions and high error rate)

Goal: Pick an alignment for each subread that minimizes the number of structural variants predicted
1. Formalize the problem
2. Construct a graph representing all possible solutions

Ref. [Ritz et al 2009]

Vertices:
- Pairs
- Source $\alpha_i$
- Sink $\beta_i$

Edges:
- Subread alignments

Find a subgraph with (1) fewest # of vertices and (2) a path from $\alpha$ to $\beta$ for all strobes.

Goal: Pick an alignment for each subread that minimizes the number of structural variants predicted.
ILP Formulation [Ritz et al. 2010] “Network Flow”

Goal: Find a subgraph $H$ of $G$ with
1. Fewest number of vertices
2. Path exists from $\alpha$ to $\beta$ for all strobes

$$\min \sum_i p_i$$

s.t.
$$p_i \in \{0, 1\} \forall i$$
$$0 \leq q_{ij} \leq p_i \forall i, j, k$$
$$0 \leq q_{ij} \leq p_j \forall i, j, k$$

$$\sum_{j \in N_k^+} q_{ij} - \sum_{j \in N_k^-} q_{ij} = \begin{cases} 1 & \text{if } i = \alpha S_k \\ -1 & \text{if } i = \beta S_k \\ 0 & \text{otherwise} \end{cases}$$

Is this a Standard Flow Problem?

Goal: Find a subgraph $H$ of $G$ with
1. Fewest number of vertices
2. Path exists from $\alpha$ to $\beta$ for all strobes

Standard Flow:
1. Fewest number of vertices
2. Sources have flow +1, sinks have flow -1
3. Flow in = flow out for all others
Split Reads for Strobes

Nearby SVs (example: deletions)

Test

Reference

(split)

breakpoint

breakpoint

DNA Mutations

Single Nucleotide Variants

Copy Number Variants

Structural Variants

Resequencing

Resequencing

Resequencing with Paired Reads

Resequencing with Strobes

Resequencing

Array Comparative Genomic Hybridization (aCGH)
Structural Variation Signatures

<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>Deletion</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>Inversion</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>
</tbody>
</table>

Structural Variation Signatures

<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>Tandem duplication</td>
<td><img src="image8.png" alt="Diagram" /></td>
<td><img src="image9.png" alt="Diagram" /></td>
<td><img src="image10.png" alt="Diagram" /></td>
<td><img src="image11.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Interspersed duplication</td>
<td><img src="image12.png" alt="Diagram" /></td>
<td><img src="image13.png" alt="Diagram" /></td>
<td><img src="image14.png" alt="Diagram" /></td>
<td><img src="image15.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Nonad sequence insertion</td>
<td><img src="image16.png" alt="Diagram" /></td>
<td>Not applicable</td>
<td><img src="image17.png" alt="Diagram" /></td>
<td><img src="image18.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>
Combining Measurements

Combining Measurements