New Sequence Variants Identified by Region-Specific Extraction

PAG XX Workshop Presentation – Mutation Screening
January 14, 2012

Johannes Dapprich
Generation Biotech – Company Background

- Generation Biotech, LLC Established (2000)
- Initial Development HSE – Technology (2001)
- Phase I NIAID SBIR Award (2002)
- Phase I Commercialization – Qiagen acquires HSE License HLA (2003)

- Number of Staff: 5
- Revenue through Licensing, Services & 3rd Party Funding

Generation Biotech Launches RSE / HSE Extraction Service December 2009
The Problem

NGS and short-fragment-based enrichment methods cannot always resolve more complex or unknown sequence regions:

- homologous regions, repeat elements, structural variations
- highly polymorphic / polyploid genome of interest
- incomplete reference sequence
- gaps of unknown sequence
- new alleles
Generation Biotech Technology:

Multiplexed, magnetic particle-based purification of genomic DNA

**RSE: Region-Specific Extraction**
- Large DNA segments (>20kb per primer)
- Can identify unknown sequence space
- Productivity and resolution improvement when combined with NGS

**HSE: Haplotype-Specific Extraction**
- SNP-specific resolution of ambiguous allele combinations
- Associated mutations can be discovered and/or resolved
- Possible improvement of coverage for low-abundance SNPs
What We Provide

RSE and HSE -

Enrichment Service or Technology Platform:

• Primer Design, Amplification, Tech Support & QC
• Sample extraction
• Interface with NGS if desired (all major providers)

Solution Based Approach

EZ1

BS96

Generation Biotech
How does it work

- Flexible - any regular primer can be used
- Sequence- or SNP-specific capture
- 10-20 kb DNA fragments per oligo for NGS, can be larger for other applications
- Unique enzyme-based technology
- Only method to obtain large fragments from a single SNP, enrichment factor ~30x - 100x
- Scalable – can cover hundreds of regions in parallel for hundreds of samples
  - **EZ1**: 100 samples per day (6 samples / 1h)
  - **BS96**: 1000 samples per day

**Generation Biotech**
Haplotype- / Region-Specific Extraction, HSE / RSE
Haplotype-/Region-Specific Extraction, HSE/RSE
Haplotype-/Region-Specific Extraction, HSE/RSE
Haplotype- / Region-Specific Extraction, HSE / RSE
Haplotype- / Region-Specific Extraction, HSE / RSE
Haplotype-/Region-Specific Extraction, HSE/RSE
Benefit – Fragment Size

- Large fragment size captured per primer / probe
- Retains linkage information & avoids gaps
- Creates redundancy & robustness across target sites
- Allows detection of structural variations and unknown sequence
Total region size: 142,304 bp

7 capture primers

Large fragment capture
Total region size: 142,304 bp

7 capture primers

Large fragment capture

Small fragment capture

GENERATION Biotech
Discovery of Unexpected Structural Variants (Zebrafish)

Capture oligos (41)
qPCR assays

300kb
Target region

10^6 bp

Gupta et al. PLoS Genetics 2010
Discovery of Unexpected Structural Variants (Zebrafish)

First biologically relevant identification of a causative developmental variant by targeted capture & NGS

Gupta et al. PLoS Genetics 2010
Simple Workflow

**Design**
- Select region(s) of interest
- Repeat mask available sequence
- Design capture primers

**Capture**
- Combine 600 ng DNA, primers and reaction mix & denature
- Place on robot - incubation, capture and washing (1 hour)
- Targeted DNA is retained on magnetic particles

**NGS**
- Perform WGA from extracted DNA (i.e. REPLI-g, overnight)
- Use 1-2 ug of WGA as input for NGS library preparation
  (use like genomic DNA)
RSE Application areas

- Mutation analysis of hundreds of genes across large sample sets
- Gap closure, resolution of duplications / repeat elements
- Determination of unambiguous linkage for unknown sequence
- Insertion point mapping - e.g. transposons / T-DNA / viral DNA
- Sequencing of cloning vectors
- Alternative to Southern, e.g. transgenics production and QC
RSE Captures Multiple Regions of Interest

- 10 separate regions (size ≈ 1 Mb) captured with 81 specific oligos
  - cost effective
  - only sequence what you need
  - simplified analysis of data
  - less coverage needed

- Quick adaptation to any other target region or pathway of interest

GENERATION Biotech
Example - Targeting 100 Genes for 10 Samples

10 samples = 10 separate extractions
100 genes = ~200-500 capture primers

Barcode captured samples after extraction

Combine barcoded samples

Sequence

Bioinformatics

~$500 per sample

~$250 per sample

~$2,000 combined

~$1,000 ? (simplified)

~ $10,000 total cost for 10 samples
HSE applications – not achievable with other methods

• Capture inserts or deletions or other structural variants that are associated with SNPs - particularly if rare SNPs
  – *Is the SNP pointing to a structural variant?*
  – *Does the SNP regulate a downstream unknown target?*

• Find minor variants associated with a SNP that would otherwise be missed or interpreted as sequencing error

• Resolve phase information of multiple SNPs
  – *Are two SNPs on the same allele?*
  – *Can two SNPs interact with one another? (Target-Passenger)*

• SNP-based resolution of specific alleles & complex regions
Human MHC:

- Well characterized, highly polymorphic genomic region
- Numerous known disease associations
- Many HLA allele combinations are ambiguous

![Eletrophoretogram image with two heterozygous positions highlighted: T + G & A + G]
<table>
<thead>
<tr>
<th>Pair 1</th>
<th>B*07021</th>
<th>C TCCG TGTCC...249bp... GAGCATGTA...46bp.. CCAGTACG</th>
<th>B*1501</th>
<th>CGCCATGTCC...249bp... GAGGATGTA...46bp.. CCAGTCCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes:</td>
<td>T/G G/A</td>
<td>C/G</td>
<td>C/A</td>
<td></td>
</tr>
</tbody>
</table>

- OR -

<table>
<thead>
<tr>
<th>Pair 2</th>
<th>B*0707</th>
<th>C TCCG TGTCC...249bp... GAGGATGTA...46bp.. CCAGTACG</th>
<th>B*1507</th>
<th>CGCCATGTCC...249bp... GAGCATGTA...46bp.. CCAGTCCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes:</td>
<td>T/G G/A</td>
<td>C/G</td>
<td>C/A</td>
<td></td>
</tr>
</tbody>
</table>

- OR -

<table>
<thead>
<tr>
<th>Pair 3</th>
<th>B*0709</th>
<th>C TCCG TGTCC...249bp... GAGCATGTA...46bp.. CCAGTCCG</th>
<th>B*1563</th>
<th>CGCCATGTCC...249bp... GAGGATGTA...46bp.. CCAGTACG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes:</td>
<td>T/G G/A</td>
<td>C/G</td>
<td>C/A</td>
<td></td>
</tr>
</tbody>
</table>

**Generation Biotech**

(ARC sample CB-396)
### Pair 1

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Sequence Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>B*07021</td>
<td>C’TCCG’TGTCC…249bp… GAGCATGTA…46bp.. CCAGTACG</td>
</tr>
<tr>
<td>B*1501</td>
<td>CGCCATGTCC…249bp… GAGGATGTA…46bp.. CCAGTC CG</td>
</tr>
<tr>
<td>Genotypes:</td>
<td>T/G  G/A  C/G  C/A</td>
</tr>
</tbody>
</table>

### Pair 2

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Sequence Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>B*0707</td>
<td>C’TCCG’TGTCC…249bp… GAGGATGTA…46bp.. CCAGTACG</td>
</tr>
<tr>
<td>B*1507</td>
<td>CGCCATGTCC…249bp… GAGCATGTA…46bp.. CCAGTC CG</td>
</tr>
<tr>
<td>Genotypes:</td>
<td>T/G  G/A  C/G  C/A</td>
</tr>
</tbody>
</table>

### Pair 3

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Sequence Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>B*0709</td>
<td>C’TCCG’TGTCC…249bp… GAGCATGTA…46bp.. CCAGTCC CG</td>
</tr>
<tr>
<td>B*1563</td>
<td>CGCCATGTCC…249bp… GAGGATGTA…46bp.. CCAGTACG</td>
</tr>
<tr>
<td>Genotypes:</td>
<td>T/G  G/A  C/G  C/A</td>
</tr>
</tbody>
</table>

**GENERATION Biotech**

(ARC sample CB-396)
Pair 1

B*07021
CTCC\textcolor{red}{G}TGTCC\ldots249bp...\ GAG\textcolor{red}{C}ATGTA\ldots46bp..\ CCAGT\textcolor{red}{A}CG

B*1501
CGCC\textcolor{red}{A}TGTC\ldots249bp...\ GAGG\textcolor{red}{G}ATGTA\ldots46bp..\ CCAGT\textcolor{red}{C}CG

Genotypes: T/G G/A C/G C/A

- OR -

Pair 2

B*0707
CTCC\textcolor{red}{G}TGTC\ldots249bp...\ GAGG\textcolor{red}{G}ATGTA\ldots46bp..\ CCAGT\textcolor{red}{A}CG

B*1507
CGCC\textcolor{red}{A}TGTC\ldots249bp...\ GAGG\textcolor{red}{C}ATGTA\ldots46bp..\ CCAGT\textcolor{red}{C}CG

Genotypes: T/G G/A C/G C/A

- OR -

Pair 3

B*0709
CTCC\textcolor{red}{G}TGTC\ldots249bp...\ GAGC\textcolor{red}{A}TGTA\ldots46bp..\ CCAGT\textcolor{red}{C}CG

B*1563
CGCC\textcolor{red}{A}TGTC\ldots249bp...\ GAGG\textcolor{red}{A}TGTA\ldots46bp..\ CCAGT\textcolor{red}{A}CG

Genotypes: T/G G/A C/G C/A

GENERATION Biotech

(ARC sample CB-396)
<table>
<thead>
<tr>
<th>Pair 1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B*07021</strong></td>
<td><strong>C</strong></td>
<td><strong>TGTCC</strong>...249bp...</td>
</tr>
<tr>
<td><strong>B*1501</strong></td>
<td><strong>C</strong></td>
<td><strong>G</strong></td>
</tr>
<tr>
<td>Genotypes:</td>
<td><strong>T/G  G/A</strong></td>
<td><strong>C/G  C/A</strong></td>
</tr>
</tbody>
</table>

- OR -

<table>
<thead>
<tr>
<th>Pair 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B*0707</strong></td>
<td><strong>C</strong></td>
<td><strong>TGTCC</strong>...249bp...</td>
</tr>
<tr>
<td><strong>B*1507</strong></td>
<td><strong>C</strong></td>
<td><strong>G</strong></td>
</tr>
<tr>
<td>Genotypes:</td>
<td><strong>T/G  G/A</strong></td>
<td><strong>C/G  C/A</strong></td>
</tr>
</tbody>
</table>

- OR -

<table>
<thead>
<tr>
<th>Pair 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B*0709</strong></td>
<td><strong>C</strong></td>
<td><strong>TGTCC</strong>...249bp...</td>
</tr>
<tr>
<td><strong>B*1563</strong></td>
<td><strong>C</strong></td>
<td><strong>G</strong></td>
</tr>
<tr>
<td>Genotypes:</td>
<td><strong>T/G  G/A</strong></td>
<td><strong>C/G  C/A</strong></td>
</tr>
</tbody>
</table>

(ARC sample CB-396)
Human MHC:

\[ K = T + G \]
\[ R = A + G \]
Identification of flanking sequences

- Insertion site mapping / gene transfection / T-DNA
- Target DNA via conserved & unique sequence in inserted element

- Label captured segments with Cy3 / Cy5 for arrays or:

- Amplify via WGA and proceed to library preparation for NGS
Global Mapping of Transposon Location

Abram Gabriel1, Johannes Dapprich2, Mark Kunkel2, David Gresham3,4, Stephen C. Pratt5, Maitreya J. Dunham3

1 Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey, United States of America. 2 Generation Biotech, Lawrenceville, New Jersey, United States of America. 3 Lewis-Sigler Institute, Princeton University, Princeton, New Jersey, United States of America. 4 Department of Molecular Biology, Princeton University, Princeton, New Jersey, United States of America. 5 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, United States of America.

Transposable genetic elements are ubiquitous, yet their presence or absence at any given position within a genome can vary between individual cells, tissues, or strains. Transposable elements have profound impacts on host genomes by altering gene expression, assisting in genomic rearrangements, causing insertional mutations, and serving as sources of phenotypic variation. Characterizing a genome’s full complement of transposons requires whole genome sequencing, precluding simple studies of the impact of transposition on interindividual variation. Here, we describe a global mapping approach for identifying transposon locations in any genome, using a combination of transposon-specific DNA extraction and microarray-based comparative hybridization analysis. We use this approach to map the repertoire of endogenous transposons in different laboratory strains of Saccharomyces cerevisiae and demonstrate that transposons are a source of extensive genomic variation. We also apply this method to mapping bacterial transposon insertion sites in a yeast genomic library. This unique whole genome view of transposon location will facilitate our exploration of transposon dynamics, as well as defining bases for individual differences and adaptive potential.

Comparison of Ty1 & Ty2 locations in two unrelated yeast strains

Chr:
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI

S288c - red
RM11 - green

gene 1 gene 2
gene 3 gene 4

transposon capture point

gene 3
RSE / HSE Value

- RSE / HSE can produce more comprehensive and accurate information for NGS and conventional sequencing
- Quick adaptation to any target region or pathway of interest
  - Primer sets easily modified as needed based on new information
  - Permits efficient iterative “walking” through difficult regions
  - Discovers de novo sequence, unknown linkage and structural variants
  - Determines unambiguous molecular haplotypes
- Direct follow-up on SNP-based GWA studies possible
  
  *(directly builds on existing infrastructure, samples and knowledge base)*
Summary – Application areas:

RSE / HSE validated for:

- Resolution of ambiguities and complex regions
- Isolation of candidate regions for Next-Generation Sequencing
- Identification of insertion points, rearrangements & copy number
- Identification of new alleles
- Identification of unknown sequence & new SNPs
- Identification of novel structural variants