Genotyping & Sequencing Technologies

Jeanette Papp
Director, Genotyping & Sequencing Core
Adjunct Associate Professor
Department of Human Genetics
Core Equipment

- Roche GS FLX (454) Next Gen Sequencer
- Fluidigm Biomark
- 3730 capillary sequencers
- Taqman 7900HT Real-time PCR instrument
- CEQ 8000 sequencers
- Roche LightCycler 480
- PSQ96 Pyrosequencer
- Qiagen TissueLyser
- Agilent Bioanalyzer
- Beckman Coulter Counter
- Liquid handling robots
- PCR Machines

Human Genetics Department
- Affymetrix
- Illumina
- Solexa
- SOLiD
- Sequenom
Genetic Assays

- SNP Genotyping
- DNA Methylation Analysis
- Gene Expression
- Single Cell Gene Expression
- LOH (loss of heterozygosity)
- siRNA/RNAi, microRNA
- In/Del Analysis
- Copy Number Variation
- Microsatellite Genotyping
- Large Fragment Sizing
- AFLP, RFLP
- BAC Fingerprinting
- SAGE
- HLA Typing
- Conformation Analysis
- Allele Quantification
- Sequencing
- Resequencing
- Comparative Sequencing
- CLIA-Certified Clinical Testing
Maximize DNA Quality

All genotyping methods are sensitive to DNA...

• **Quantity** – Do you have sufficient DNA for your assay? Should you consider Whole Genome Amplification (WGA)? Advantages of WGA can be quantity and consistency. Disadvantages – amplification bias, expense. Tip: *Don’t mix WGA and non-WGA.*

• **Quality** – What is the quality of your DNA? High concentrations of poor quality DNA will not help you.

• **Consistency** – Even if the quantity and quality are adequate, if they vary widely from sample to sample, you are liable to get poor results

Is your technology more robust to DNA quality or quantity? What is the dynamic range of detection?
Sample Collection

• Where will you get your DNA?
  • Blood, Tissue, Buccal Swab, Saliva?
    • Quality
    • Consistency
    • Compliance
DNA / RNA Collection

1. Spit your saliva into the Oragene container.
2. Keep spitting until the amount of liquid saliva (not counting foam) reaches the top of the white label.
3. Tighten the cap very firmly.
4. Gently mix your saliva.

Figure 1. Agarose gel electrophoresis of DNA extracted from Oragene/saliva samples. A Lambda-Hind III digest was used as the marker in Lane 1.
Polymorphisms

• Different values at a specific genomic location (locus) between individuals in a population
• Microsatellites
  • Often variable numbers of short tandem repeats
  • *aka*: VNTR, STR
  • Fewer and younger
• SNPs
  • Single *N*ucleotide *P*olymorphism
  • Usually only two possible values at that base
  • About 10 million human SNPs have been identified
  • More and older
Genotyping Gel Autoradiograph
Fluorescent Genotyping Gel
SNPs

• Single Nucleotide Polymorphism
• Responsible for 90% of all human genetic variation
• A SNP occurs every 100-300 base pairs
• Currently almost 12 million SNPs in the NCBI SNP database
• May be within genes (coding SNP, cSNP) or outside gene (non-coding, the majority)
• May cause amino acid changes or not. If it causes an amino acid change it is called non-synonymous (nsSNP)
• Most SNPs are not responsible for a disease.
• Like microsatellites, they are used as markers for pinpointing a disease on the genome map. SNPs make particularly good markers because
  - They occur frequently throughout the genome.
  - They are older and more stable genetically.
**SNP Platforms**

**Study Size**

- **Pyrosequencing** – few SNPs, few samples, low-throughput, labor intensive
- **Taqman** – fewer SNPs, fewer samples, moderate-throughput
- **Fluidigm, SNPlex, Sequenom** – moderate numbers of SNPs and samples, high-throughput (hundreds of thousands per day)
- **Illumina** – many SNPs, many samples, ultra-high-throughput (millions per day)
- **Affymetrix** – many SNPs, many samples, ultra-high-throughput (millions per day)
- **Next Generation Sequencing** – massively parallel, ultra-ultra-high-throughput
SNP Platforms

Technology

- **Pyrosequencing** – Sequencing by synthesis. Addition of dNTPs one at a time. Luciferase generates light when nucleotides incorporated. Gives you the neighboring sequence.

- **Taqman** – Allelic discrimination. Fluorescent probe for each SNP variant. Simple, robust chemistry.

- **Fluidigm** – Microfluidics and Taqman qPCR.

- **SNPlex** – Multiplexed, fluorescently labeled oligonucleotides on capillary electrophoresis.

- **Sequenom** – Mass Spectrometry. Highly accurate and reproducible.

- **Illumina** – Microarray of tiny beads with bound oligonucleotides. Sample DNA binds to the bead oligo, and is detected by an optical fiber.

- **Affymetrix** – Microarray of probe DNA spots printed on a glass or plastic chip.
Figure 1 | The principle of Pyrosequencing and the output Pyrogram™. Double peak heights indicate incorporations of two nucleotides in a row.
Allelic Discrimination
SNPlex Chemistry Overview

Encoding
Generation of genotype (GT) specific products through multiplex oligonucleotide ligation reaction (OLA)

-Amplification
Multiplex PCR with universal primers

Decoding
Hybridization of universal ZipChute probes to amplicons and identification of eluted ZipChutes by CE

Each ZipChute probe pair is representative of one SNP.
SNPlex Raw Data
SNPlex Raw Data
Examples

Good

Bad

Good

Bad
Examples
SNP MicroArray
SNP BeadArray
SNP Platforms

Cost

- Affymetrix – $0.0006-??
- Illumina – $0.0002-??
- Sequenom – $0.05-$0.20
- SNPllex – $0.08-$0.14
- Taqman – $0.20-$0.70
- Pyrosequencing – $0.10-$0.50
When Choosing a Technology Platform, Consider:

- **Conversion rate** – How many of the SNPs you chose worked - *in silico, in vitro*?
- **Reproducibility** – Do you get the same result when you repeat the whole process?
- **Error Rate** – Compared to “true” result.
- **Concordance** – Agreement with some other method.
- **Call Rate** – How much missing data?
- **Cost** – What can you afford?
Array Data
Chip-based Genomics

- Each spot represents one genetic marker
- New generation chips hold 2.3 million SNPs
- To find genes for common, complex traits it may require DNA from 2000 individuals

4,600,000,000 SNP dataset
Sequencing Technologies
A Brief History of Sequencing...

- A adenine
- C cytosine
- T thymine
- G guanine
Sanger Sequencing
Chain Termination

Template: 3' CCGTAGCAACT 5'
Primer: 5' GG3' 3'

1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation
2. Sequence read (bottom to top) from gel autoradiogram

Figure 3-21 Fundamentals of Biochemistry, 2/e
© 2006 John Wiley & Sons
Sequencing Gel Autoradiograph
Dye Termination

TACTGGAGCAATCCGTTTCGGAAA•••A

...
Fluorescent Gel
DNA Sequencing
Capillary Sequencing

- Background
- Methods
- Applications
Human Genome Project

3 Billion Base Pairs

- **1990** - Begin: estimate **15 years, $3 billion**
- **1998** - Craig Venter starts new company *Celera*, “will sequence human genome in 3 years for $300 million”
- **2000** - Working draft
- **2003** - Complete in **13 years, $2.7 billion**
To develop novel technologies to sequence a mammalian-sized genome for approximately $1000.00.

“...the realization of the goals of this RFA is a long-range effort that is likely to require as much as ten years to achieve.”

Near-term goal - $100,000 genome.
X PRIZE Foundation
$10 million prize

For the first device to sequence 100 human genomes in 10 days or less, for less than $10,000.
Next Generation Sequencing Technologies

Strategy

• Automation
• Parallelization
• Miniaturization
Commercially Available Now

- **454 (now Roche GS FLX) – 2005**
  - 300-500 base pair reads. Good for novel genomes
  - 100 million base pairs in 8 hour run

- **Solexa - 2006**
  - Short reads, 35-50 bp. Best when there is a reference sequence (“resequencing”) or for analyzing small molecules like RNA.
  - 4 gigabases per 3 day run - the equivalent of one-third of the human genome

- **SOLiD - 2007**
  - Short reads, 35-50 bp
  - 3 GB per 3 day run

- **Helicos – 2008**
  - Short reads, 35-50 bp
  - 10 GB per 8 day run
Technology

• Roche GS-FLX (454)
  - DNA is fragmented, bound to beads, amplified till each bead has 100,000 copies of the DNA fragment
  - Pyrosequencing - When a base is added to the DNA strand pyrophosphate is released, used as a substrate for luciferase, light is emitted and detected by a camera

• Illumina (Solexa)
  - Sequencing by synthesis with fluorescently labeled nucleotides and DNA fragments bound to a slide.

• SOLiD
  - Sequencing by ligation. Bead-bound DNA molecule is interrogated with each of the 16 possible 2 base pair combinations in a fluorescently labeled oligonucleotide.

• Helicos
  - Single molecule sequencing by synthesis with fluorescently labeled bases and DNA fragments bound to a flow cell surface.
Next Gen Sequencer
Roche 454 GS FLX
Clonal Cluster Technology

- Randomly fragment DNA
- Dilute to a single DNA molecule
- Amplify single molecule to get colonies of identical DNA
- Sequence by synthesis or ligation

Roche 454
Illumina Solexa

Lots of Data!
Next-Next Sequencing

Helicos
Shipped in 2008
Single molecule sequencing

Complete Genomics
2009
Sells service, not instrument
<$5000 for whole human genome
Sequence “DNA nano-balls”

Pacific Biosciences
2010 - 2013
Single molecule
Under an hour for hundreds of $
Miniaturized - zeptoliter
# Cost to Sequence Human Genome

<table>
<thead>
<tr>
<th>Human Genome Project</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Capillary Sequencer</td>
<td>$3 million</td>
</tr>
<tr>
<td>Roche GS FLX (454)</td>
<td>$100,000</td>
</tr>
<tr>
<td>Illumina Solexa</td>
<td>&lt; $100,000</td>
</tr>
<tr>
<td>Applied Biosystems SOLiD</td>
<td>$60,000</td>
</tr>
<tr>
<td>Helicos</td>
<td>&lt;$50,000</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>$5000</td>
</tr>
<tr>
<td>BioNanomatrix???</td>
<td>$100</td>
</tr>
</tbody>
</table>

The Human Genome Project cost is $2.7 billion.
Human Genome

3 Billion Base Pairs

- **Human Genome Project** - 13 years, $2.7 billion
- **James Watson’s Genome** - two months, <$1 million, on the 454
- **Today** - weeks to months, less than $100,000
- **Tomorrow** - $5000

*The Challenge: Turning Data into Information*
Applications

As the introduction of the personal computer created new applications in computing, whole genome sequencing technology has generated new applications.
Whole Genome Sequencing

*De novo* sequence of novel genomes

- *Mycobacterium tuberculosis*
- *Vibrio cholerae*
- *Streptococcus pneumoniae*
- *Haemophilus influenzae*
- *Helicobacter pylori*
Genomic Diversity and Metagenomics

Microbial diversity in

- Human microbiome
- Microbes in honey bee colony collapse disorder
- Deep sea microenvironments
- Deep mine microbial ecology
- Environmental sampling
Comparative Genomics, Paleogenomics, Ancient DNA Analysis

- Neanderthal Genome
- Mammoth
- Ancient wolves
- Mitochondria from ancient hair shafts
Chromosome Structure

- Deletions
- Duplications
- Copy number variation
- Insertions
- Inversions
- Translocations
- Methylation Analysis
Transcriptome Analysis

• Use massively parallel sequencing to quantitatively measure gene expression across tens of thousands of samples

• Develop a *genomic signature* of
  
  - Cell differentiation state
  - Disease state
  - Therapeutic response
  - Tumor signature
Deep Resequencing

Sequencing against a reference genome

- Identify rare disease-causing variants
- Cancer, HIV mutation detection
<table>
<thead>
<tr>
<th>Research Field</th>
<th>Application</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Novo Sequencing</td>
<td>Whole Genomes of Plants, Yeasts, Fungi, Bacteria, Viruses</td>
<td>Shotgun Sequencing, Paired-End Sequencing, De Novo Assembly</td>
</tr>
<tr>
<td>Resequencing</td>
<td>Whole Genomes of Humans, Plants, Yeasts, Fungi, Bacteria, Viruses</td>
<td>Shotgun Sequencing, Mapping to Reference Sequence</td>
</tr>
<tr>
<td></td>
<td>Genomic rearrangements, Copy number variation</td>
<td>Paired-End Sequencing</td>
</tr>
<tr>
<td></td>
<td>Indels, SNPs, Somatic Mutations</td>
<td>Amplicon Sequencing</td>
</tr>
<tr>
<td>Transcriptome Analysis</td>
<td>Full-length transcripts</td>
<td>cDNA fragment sequencing</td>
</tr>
<tr>
<td></td>
<td>Multiplex Sequencing of Paired-End Ditags (Singapore MS-PET)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Serial Analysis of Gene Expression (SAGE)</td>
<td></td>
</tr>
<tr>
<td>Gene Regulation Studies</td>
<td>Small non-coding RNAs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transcription Factor Binding Sites (ChIP-Sequencing)</td>
<td></td>
</tr>
<tr>
<td>Epigenetic Changes</td>
<td>DNA Methylation Pattern</td>
<td>Amplicon Sequencing</td>
</tr>
<tr>
<td></td>
<td>Nucleosome Modifications</td>
<td>DNA fragment sequencing</td>
</tr>
<tr>
<td>Metagenomics &amp; Microbial Diversity</td>
<td>Analysis of Environmental DNA</td>
<td>Shotgun Sequencing</td>
</tr>
<tr>
<td></td>
<td>16S rRNA</td>
<td>Amplicon Sequencing</td>
</tr>
<tr>
<td>Paleogenomics</td>
<td>Whole Genome Sequencing of Ancient DNA</td>
<td>Shotgun sequencing</td>
</tr>
</tbody>
</table>
Diagnostics

- $1000 genome will allow individual sequence as diagnostic. Will it come too soon to be useful?

- In genomic research technology and data have often come before our ability to extract information and knowledge from them.
Gene Expression Technologies

- DNA microarray
- Whole Genome Expression Profiling
- ChIP-Seq - Chromatin ImmunoPrecipitation
  - protein interactions with DNA
Gene Expression Microarray

- Make cDNA reverse transcript
- Label cDNAs with fluorescent dyes

Control

Experimental

Hybridization to microarray

Laser excitation at dye-specific Hz

Laser emission

Computer calculates ratio of intensity

Red = "up-regulation"
Green = "down-regulation"
Black = constitutive expression

Principle of cDNA microarray assay for gene expression (after Gibson & Muse 2002)
Results - Raw Data
Expression Array Example

Columns = Brain tissue samples

Rows = Genes
Color bands indicate modules
Expression Array Example

85 Grade III and IV gliomas with 595 survival related genes

HC1A  HC1B  HC2A  HC2B

Neurogenesis
Synaptic transmission
Mitotic
Extra-Cellular matrix

“Gene expression profiling of gliomas strongly predicts survival.”
FDA ALERT [12/12/2007]: Dangerous or even fatal skin reactions … that can be caused by carbamazepine therapy, are significantly more common in patients with a particular … [gene form], HLA-B*1502. This [form] occurs almost exclusively in patients with ancestry across broad areas of Asia, including South Asian Indians. Genetic tests for HLA-B*1502 are already available. Patients with ancestry from areas in which HLA-B*1502 is present should be screened for the HLA-B*1502 allele before starting treatment with carbamazepine. If they test positive, carbamazepine should not be started unless the expected benefit clearly outweighs the increased risk of serious skin reactions.

http://www.fda.gov/cder/drug/InfoSheets/HCP/carbamazepineHCP.htm
2004: FDA Approves First Diagnostic Microarray

“Detects common genetic mutations that alter the body’s ability to metabolize specific types of drugs. The enzyme produced from the gene that is tested, called cytochrome P4502D6 (CYP4502D6), is active in metabolizing many types of drugs including antidepressants, antipsychotics, beta-blockers, and some chemotherapy drugs. Variations in this gene can cause a patient to metabolize these drugs abnormally fast, abnormally slow, or not at all. For example, the same dose that is safe for a patient with one variation might be too high (and therefore toxic) to a patient with a different variation who cannot metabolize the drug.”
Genome-Scale Rearrangements

Copy number variation - CNV

- Deletions
- Duplications
- Inversions
- Translocations
- Can be kilobases to megabases in size
- Common in human genome
- Can be inherited or de novo mutation
- May have more significance to human disease than previously believed

April 12, 2010
Traits Affected by CNV

- Cancer
- HIV
- Autism
- Schizophrenia
Epigenetics

• Changes in gene expression or phenotype, but not the basic structure of DNA

• Mechanisms such as methylation or protein interactions

• Can activate or silence genes

• Can be preserved when cells divide
Traits Affected by Epigenetics

• Imprinting Disorders:
  o Prader-Willi, Angelman and
  o Beckwith-Weidemann syndromes

• Cancer
Using High-Resolution Melting on the Roche LightCycler to Determine CpG-site Methylation status
Bisulfite Conversion

- Methyl-C and C are indistinguishable to DNA polymerase
- Methylation state is lost in PCR amplification
- Bisulfite treatment converts unmethylated cytosine to uracil, which becomes thymine after PCR
- Methylated cytosines are protected from bisulfite and thus unchanged
Principle of High Resolution Melting

- C to T proportion significantly changes the melting temperature of the product
- Degree of DNA methylation gives different melting profiles
- The high resolution dye from Roche intercalates and saturates evenly giving a sharp, precise melting profile
Analysis using Tm Calling

- Significantly different melting temperatures of the two species allows quantitative analysis using the Tm calling.
- By comparing the area of each peak relative to the sum of both peaks a quantitative percentage can be obtained.