Amplicon #1

3700 Loading

96 of the 450 DNA samples in SNP Discovery Panel

Amplicon #1

Forward Reaction

NIEHS SNP Tutorial
Department of Genome Sciences
University of Washington

January 30-31, 2006

Sequencing production and data analysis pipeline

- Amplify DNA
- Sequence each end of the fragment
- PolyPhred
- Consed
- Sequencing production and data analysis pipeline
- Customized software tools
  - Primer design algorithm
  - Custom LIMS and database to track all aspects of data production and quality
  - Robotics used to automate sample handling

Robotics
Automated sample handling and plate setup

- Packard MultiProbe
- PerkinElmer EvolutionP3

Data tracking using customized LIMS

- Tracks all aspects of sequencing production
  - Primer sequences
  - DNA samples
  - PCR quality
  - Inventories sequence chromatograms
  - Records read lengths, quality scores
  - Make sample tracking sheets for sequencers
  - All genotypes can be traced back to DNA sample

EGP Bench: Custom LIMS
Organized by gene

EGP Bench
Generates sample tracking form
Re-sequencing pipeline

- Gene design: automated primer picking software
  - Exons, 2 kb upstream of first exon, 2 kb downstream of last exon
  - Genes larger than 30 kb have 10% of introns scanned
- Prior to amplification and re-sequencing, problematic GC-rich regions, alu repeats, polynucleotide tracts, and pseudogenes identified

Re-sequencing pipeline

- PCR conditions optimized for each amplicon
- Failed optimization reactions repeated, primers redesigned upon second failure

Re-sequencing pipeline

- PCR quality spot checked prior to sequencing
- Failed PCR reactions repeated
- Refractory amplicons subjected to:
  - 10% DMSO
  - Alternate thermocycling parameters
  - Primer redesign

Re-sequencing pipeline

- Standard ABI BDT chemistry
  - Optimized for reaction volume and dilution
- Universal primer sequences standardize sequencing reaction conditions
- ABI 3730 capillary electrophoresis automated sequencers

Data analysis

- Amplify DNA
- Sequence each end of the fragment
- Customized software tools
  - Determine quality of base calls
  - Align traces
  - Identify potential polymorphisms
  - Identify potential functional SNPs
  - Infer haplotypes, identify tagSNPs
  - Prepare summary stats (allele freqs, HW)
  - Publish data to website and NCBI databases

Sequence Analysis

- Polyphred is the engine of our sequence analysis
- Consed provides the framework to
  - Add reads to assembly
  - Annotate reference sequence
  - Review genotypes
  - Review sequence quality
  - Resolve ambiguous genotypes using VQG to view LD
  - Identify potential allele-specific or non-unique amplification (pseudogenes) using Hardy-Weinberg equilibrium
Hardy-Weinberg Disequilibrium

Extensive Quality Control Protocols and Checkpoints Built into the System

1. Preparation of the reference sequence
   - BLAT analysis of sequence to identify closely homologous regions or pseudogenes
   - Verification of candidate gene mapping and exonic locations
   - Automated entry of baseline sequence and candidate gene information into LIMS

2. PCR primer design
   - BLAST analysis of all primer sequence to ensure specificity
   - Identification of sequence context elements and repetitive sequence which reduce sequencing read lengths and quality
   - Automated entry of all PCR primer sequences and mapping into LIMS
   - Tracking of all PCR primers to candidate gene and ordering information
Extensive Quality Control Protocols and Checkpoints Built into the System

3. PCR amplification
- Verification of PCR amplification and sizing
- Entry of PCR conditions and PCR results LIMS – linked to specific primers
- Robotic transfer of all DNA samples into pre-made, quality controlled PCR plates

4. DNA sequencing
- Robotic transfer of all diluted PCR amplicons into pre-made, quality controlled sequencing plates
- Entry of sequencing reaction data into LIMS – linked to specific PCR amplicons and PCR events
- Generation of virtual barcode for each sequencing sample
- Automated generation of sequencing sample sheet (with virtual barcode)
- Daily sequencing reports automatically generated and emailed to laboratory technicians

5. Gene assembly and polymorphism analysis
- Automated entry of sample chromatogram data in LIMS – linked to virtual barcode
- Automated entry of sample chromatogram QC data – Phred quality and read lengths
- Confirmation of orientation and location of sequence data on reference sequence during assembly
- Review of all tagged SNPs by data analyst to confirm quality
- Confirmation of all genotypes using double-stranded data
- Automated entry of polymorphism location and sample genotypes into LIMS

6. Final data processing
- Confirmation of Hardy-Weinberg equilibrium for all sites (proportion of expected genotypes per site which can reveal problems stemming from allele-specific PCR amplification).

Data publishing

- Text files published to NIEHS SNPs web site and NCBI databases
  - SNP summary data
  - Genotypes
  - Final reference sequence
- Graphical data summaries with GeneSNPs and Visual Genotype images

Data formats published to web facilitate association studies

Summary

- Amplicons designed to tile across gene region using Tm-matched PCR primers
- Amplicons sequenced using standard ABI BDT chemistry
- Amplicon sequences assembled into contigs, annotated and reviewed using Consed
- Polyphred 5.0 identifies potential SNPs, annotated and reviewed using Consed
- Custom LIMS tracks all aspects of data production and analysis
- Rapid publishing of data files to web and national databases