



Stanford - South Africa

Biomedical Informatics Program



Methods for genomic variation discovery & genotyping

Caroline F. Thorn Ph.D.



Overview

- Discovery
 - Traditional
 - Computational

- Genotyping
 - high throughput, new technologies
 - low cost, low technology methods



Traditional Discovery Methods

- PCR
- Sequencing



How PCR works

- PCR = Polymerase chain reaction
- Makes copies of DNA outside of a cell - not possible before 1980's
- Uses DNA polymerase from a thermophilic bacterial that can withstand high temperatures
- Uses primers - short pieces of single stranded DNA that are complementary to the template and start the polymerase reaction



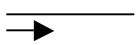
How PCR works



Tube has double stranded DNA, polymerase (aka Taq), primers and buffers (Mg to help enzyme work)



Heat tube and DNA strands come apart



At set temperature primers bind on ("annealing")
Forward primer binds on one strand, reverse primer on other strand of template



Polymerase binds on and copies the template strand ("extension")



Heat tube and DNA strands come apart



Repeat until you have lots of copies !





Analysis of PCR products

- Commonly done by gel electrophoresis
- DNA is charged and will move through a gel matrix
- Smaller pieces of DNA move more slowly
- DNA can be visualized by dye binding



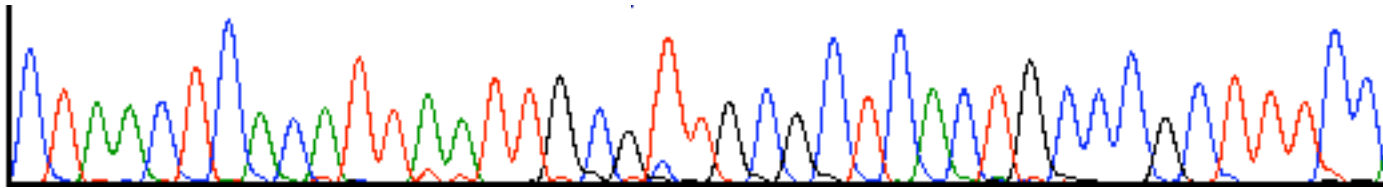
How sequencing works

- Perform PCR but include proportion of nucleotides that are modified so that elongation cannot continue 
- Each of those nucleotides has different dye - A, T, C, G
- Run products out on sequencing gel and hope to have full range of possible size products, i.e. each band has 1 extra nucleotide 



How sequencing works

- Measure the dye for each band
- Typical chromatogram looks like this:



- Software scores the bases according to the dye



Sequencing Pros/cons

- Detailed - looks at every base in gene
- Unbiased
- (If resequencing for discovery can find new SNPs depending on population)
- Expensive
- Time consuming
- Iterative
- Must sequence in both directions to be sure of SNP
- Difficult for GC-rich or repetitive regions



Computational methods for SNP discovery

- Use of databases of sequencing information and ESTs, expressed sequence tags
- Alignment of sequences
- Assess differences between them that may be caused by variation



Irizarry et al, Nature Genetics 2000

Nat Genet. 2000 Oct;26(2):233-6.

[Related Articles, Li](#)



Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences.

[Irizarry K](#), [Kustanovich V](#), [Li C](#), [Brown N](#), [Nelson S](#), [Wong W](#), [Lee C.J.](#)

- Used sequences from Unigene - 241 million nucleotides
- Kick out any sequences with low quality score and do multiple alignment with Phrap
- Used Bayesian model to infer how many predicted SNPs were true SNPs and how many from sequencing errors
- Validated by looking at the accuracy of prediction on finding coding SNPs in public data from and detailed public sequencing data for HLA gene



Computational SNPs Pros/cons

- Inexpensive
- Good first step for discovery
- Makes use of the large amount of public data
- Reliant on the quality of EST data
- And content of EST library - genes of interest may not be present, may be low copy, important SNPs may be in regions not present in ESTs
- Need good alignments
- Ultimately still have to validate by genotyping



Genotyping

- RFLP
- Taqman/primer extension methods
- Whole genome chips
- Heteroduplex method



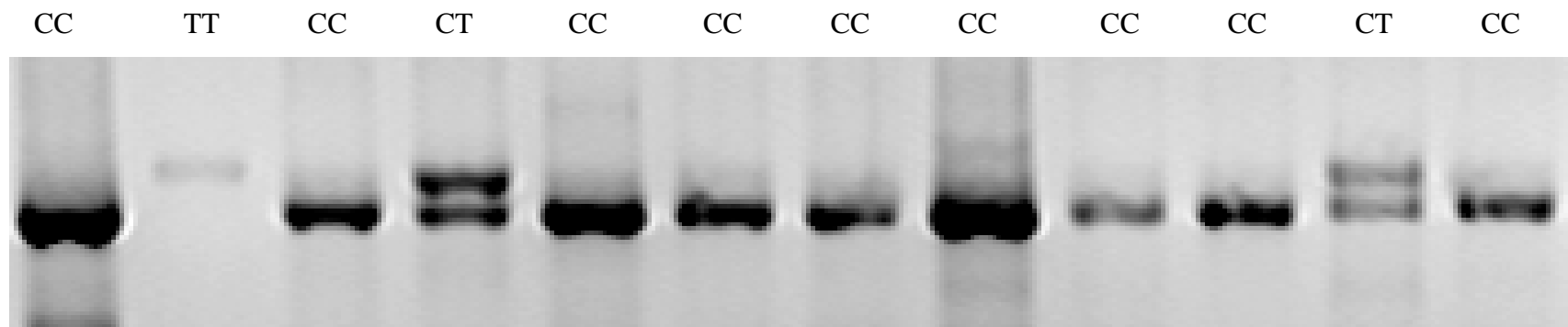
RFLP

- Restriction fragment length polymorphism
- Traditional method
- Amplify region by PCR, cut with restriction enzyme across site of SNP
- Run products out on gel
- Size difference between genotypes



Example

- CYP2C9*2



RFLP Pros/cons

- Simple
- Does not require expensive equipment (need thermal cycler, gel box, UV lamp, camera)
- Needs restriction enzymes
- Can get inconclusive results - if does not cut is it because SNP was present or reaction failed?

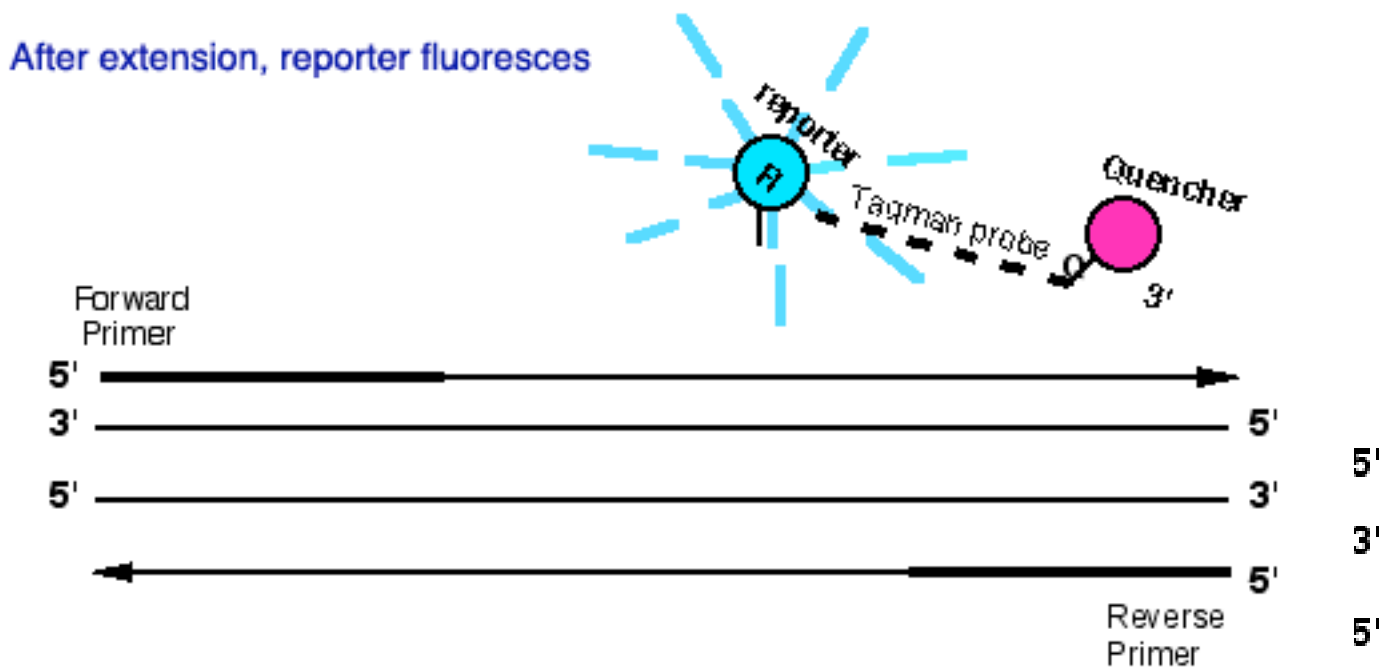


TaqMan

- TaqMan is the commercial name
- Primer extension based technique
- Fluorescent probe that binds to target DNA
- DNA polymerase extends primer, knocks off probe allowing quencher to distance from fluorophore and release of fluorescence



How TaqMan works



From <http://www.med.unc.edu/anclinic/Tm.htm>



TaqMan Pros/cons

- High throughput - can assay 96 samples in one plate, if multiplex assays can increase that
- High sensitivity/small sample volume
- Requires expensive equipment and expensive probes/primers
- Need to pick which sites to assay
- Around \$1.50/SNP/sample

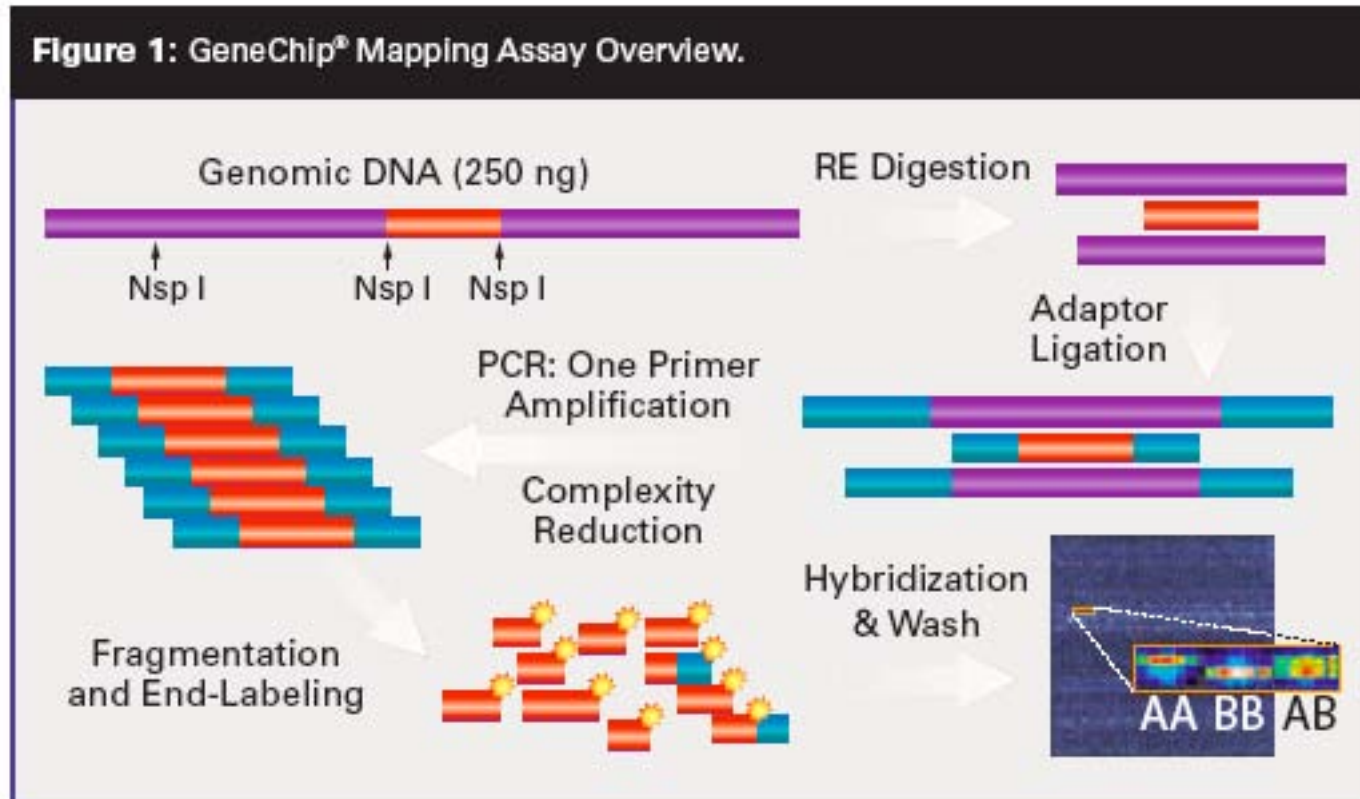


Whole genome chip

- Array technology
- Look at thousands of genotypes on one chip
- Can be gene centred or “anonymous” SNPs
- Can use for discovery or hypothesis driven



Whole genome chip



from Affymetrix



Whole genome chip Pros/cons

- Efficient
- Inexpensive after initial outlay
- High capacity
- High throughput
- Automated genotype calling
- Good way to find new targets
- Large initial investment
- Need informatics support to interpret results
- Can only genotype for biallelic SNPs
- One sample at a time
- Around \$0.15/SNP/sample

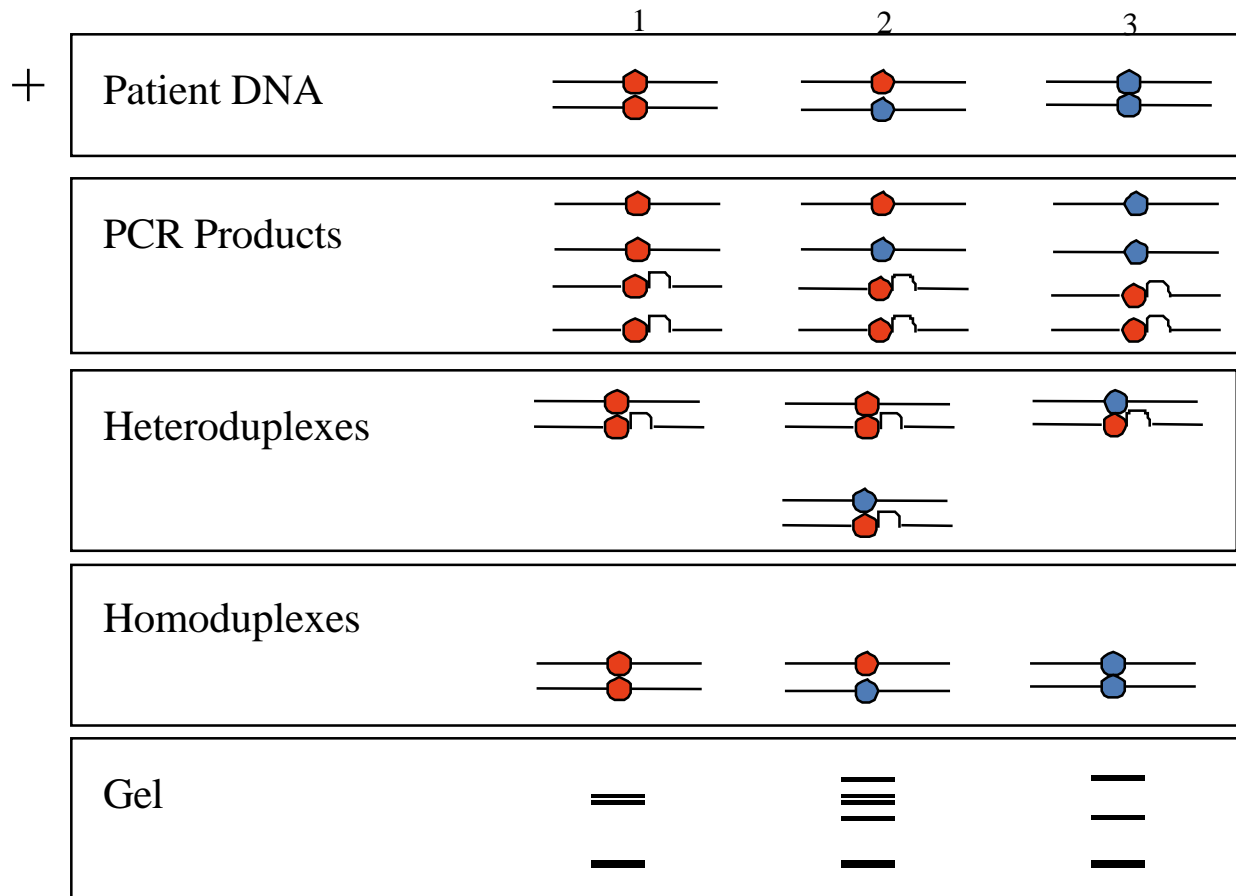
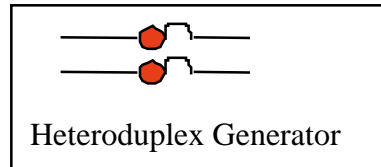


Heteroduplex method

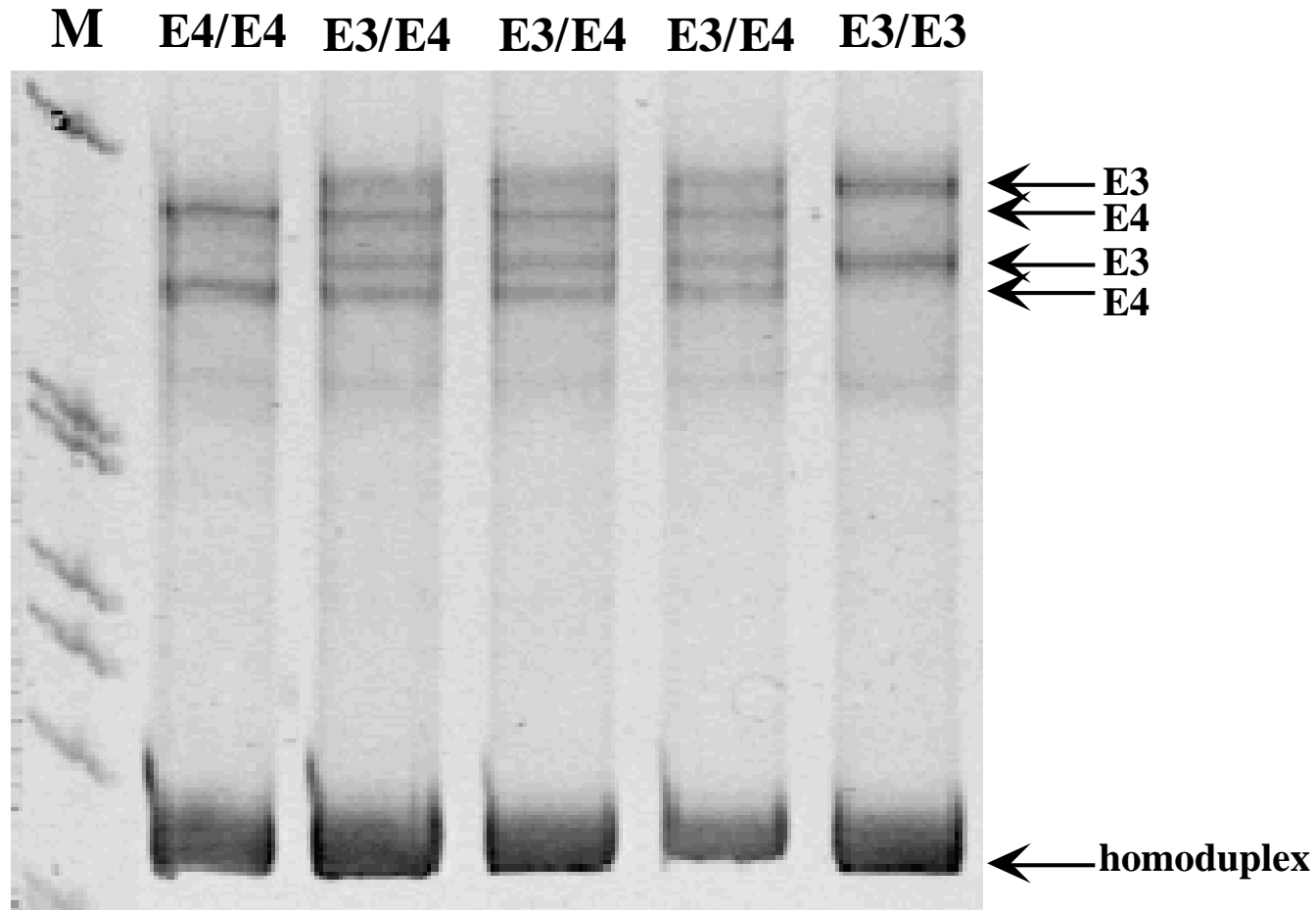
- Based on premise that 2 strands of perfectly complementary single stranded DNA when bound together (“homoduplex”) will move faster through a gel than 2 strands with mismatches bound together (“heteroduplex”)



Heteroduplex method



Heteroduplex method



Heteroduplex Pros/cons

- Simple
- Does not require expensive equipment (need thermal cycler, gel box, UV lamp, camera) AND no restriction enzymes
- Definite results
- Low throughput - slower than TaqMan, can only do ~50 at a time per SNP, unless multiplex assays
- Medium sample volume
- Relatively expensive/sample (\$1)



Summary

- New technologies have given capacity to look at many SNPs at once
- Increased need for informatics to interpret all the data
- Lower throughput methods may still be best for some situations
 - Validation
 - Discovery in new populations
- Often need combination of approaches



references

