Reduced Representation Genotyping Methods

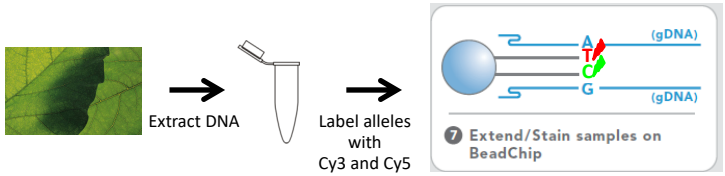
What reduced representation methods are available to researchers?

- Array based Genotyping
- Sequence Based Genotyping
- (Exome) Capture
- Transcriptome as a proxy to the genome

Reduced Representation Approaches

- Array based Genotyping
- Sequence Based Genotyping
- Exome Capture (solid and liquid)
- Transcriptome as a proxy to the genome

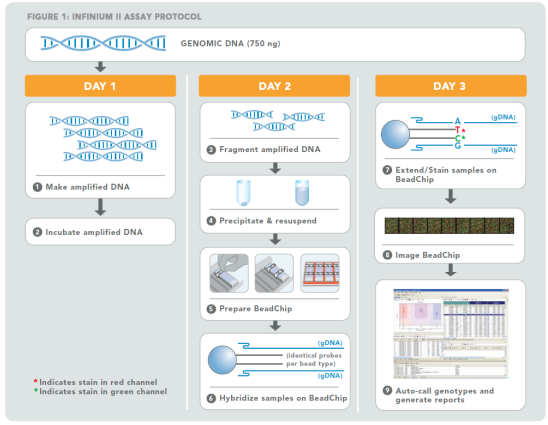
Infinium SNP Chips



- Uses single base extension and staining with Cy3 and Cy5
- Scan and process using GenomeStudio
- Assay reports only bi-allelic SNPs because it uses a two channel scanner

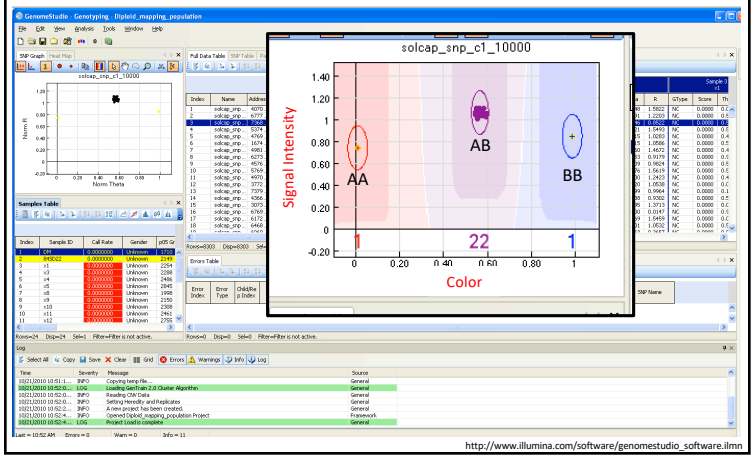
http://www.illumina.com/technology/infinium_hd_assay.ilmn

Infinium SNP Chip



http://www.illumina.com/technology/infinium_hd_assay.ilmn

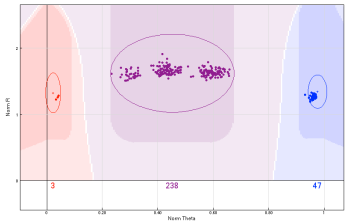
Genotype Cluster Calling



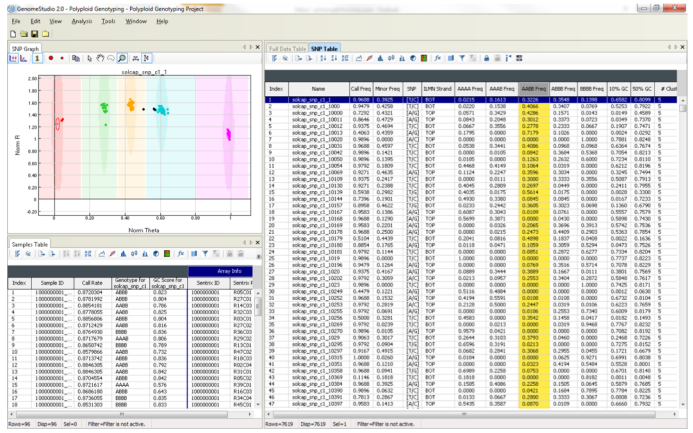
http://www.illumina.com/software/genomestudio_software.ilmn

GenomeStudio Software

- Designed for use with diploid species
  - Clusters are called as AA, AB, BB
- Autotetraploids can have at least 5 marker classes
  - AAAA, AAAB, AABB, ABBB, BBBB, also nulls (i.e. AAA)



Polyploid Genotyping Module



Genotyping Arrays

Strengths??

- High quality data
- Minimal missing data

Weaknesses??

- Ascertainments bias
- Smaller number of markers

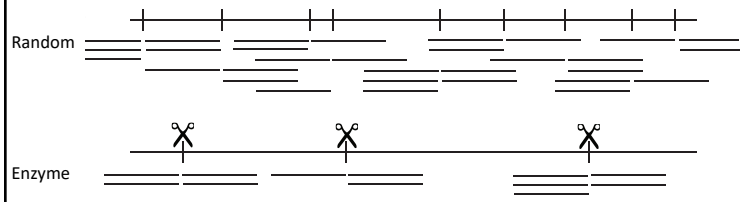
Reduced Representation Approaches

- Array based Genotyping
- Sequence Based Genotyping
- Exome Capture (solid and liquid)
- Transcriptome as a proxy to the genome

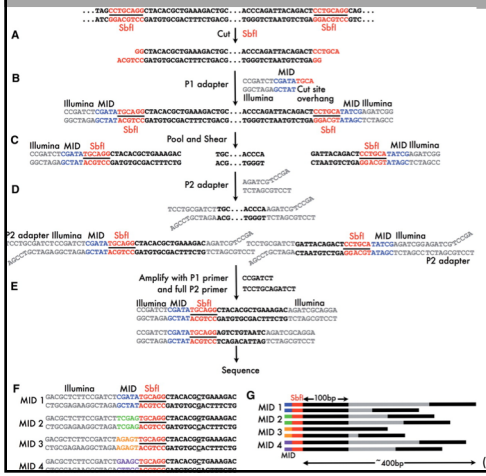
Sequence Based Genotyping

- Many different flavors of Sequence Based Genotyping
  - Skim sequencing (e.g. sequence to 0.1x coverage)
  - Reduced representation sequencing
    - RADSeq
    - Genotyping-by-Sequencing (GBS)
    - KeyGene Sequence Based Genotyping (SBG)
    - RAPiD Seq

Enzyme Based Approaches



RADSeq



Shear genomic DNA with RE

Ligate modified Illumina adapter with barcode and cut site overhang

Pool fragments and randomly shear

Ligate second adapter with divergent end

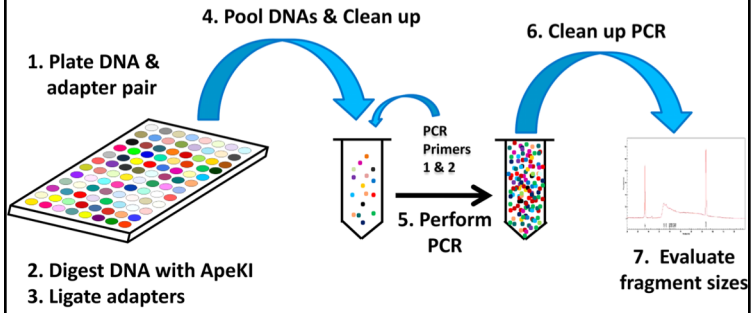
PCR Amplify

Sequence and split based on barcode sequence

(Davey et al., Briefings in Functional Genomics 2010)

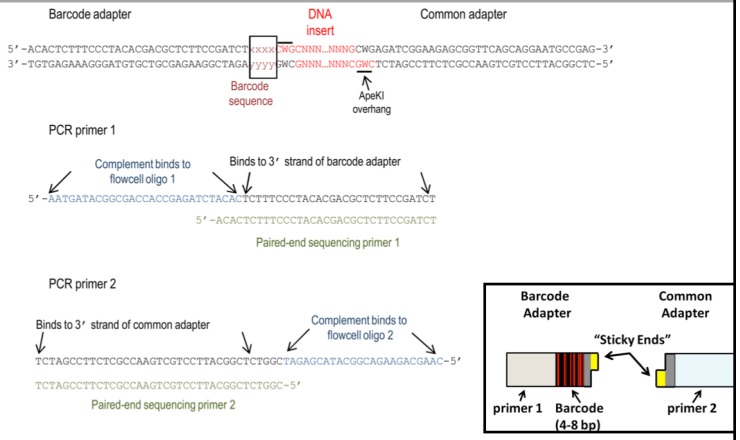
GBS Library Construction

Modified version of RADSeq that omits the shearing step
Popular variant in plant research



(Elshire et al., PLoS ONE 2011)

GBS Adapters, PCR, and Sequencing Primers



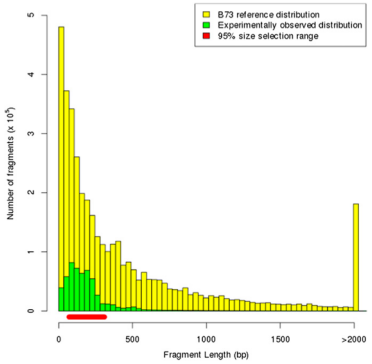
(Elshire et al., PLoS ONE 2011)

Optimizing GBS

- Variable barcode lengths to help with phasing
- Select RE that leaves 2-3 bp overhang to promote efficient adapter ligation to insert DNA
- Use RE with relatively few recognition sites in repetitive sequences
- Ideally, select RE that is methylation sensitive to increase fragments in low copy regions of genome
- Best with homozygous individual with reference genome, but not required
- Increased ploidy and heterozygosity require more sequence depth

Reality of GBS

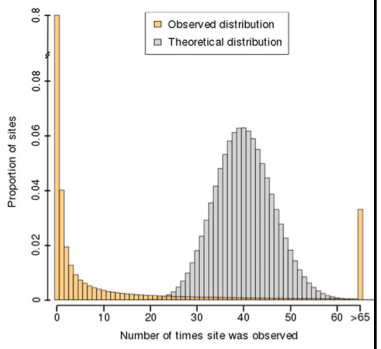
- Enzymes don't cut perfectly resulting in extra cut sites and untouched cut sites
- Maize B73 GBS Study
  - 40x theoretical coverage over expected cut sites
  - Touched 27.4% of these with at least one read
  - 12% of touched sites were not perfect matches to a G[T/A]GC cut site



(Beissinger et al., Genetics 2013)

Reality of GBS

- Can observe barcode bias
- Uneven number of reads per fragment
  - Up to 95,014 reads from a single fragment
  - Many fragments with 0 reads
- Overrepresented sequences
  - Organellar DNA found in nuclear DNA
  - Collapsed repeats in the reference genome sequence
  - GC bias



Result -> High Frequency of missing data

(Beissinger et al., Genetics 2013)

GBS-type Methods

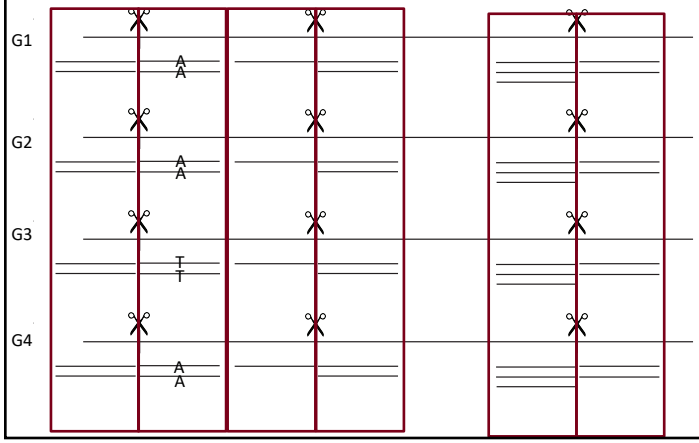
Strengths??

- Inexpensive per data point
- Small amount of sequencing to "cover" the genome

Weaknesses??

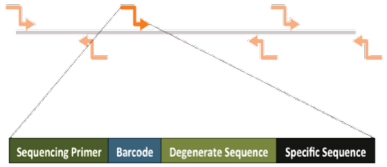
- A lot of missing data
- Uneven coverage

Restriction Enzyme Based Methods

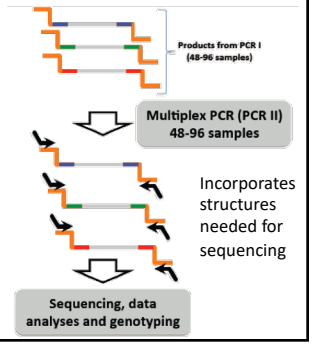


RAPiD Seq Technology by RAPiD Genomics

- Reduced representation genotyping by sequencing using PCR amplification
  - PCR 1 performed on individual samples
  - PCR 2 performed on pooled products from PCR 1



Primers designed for specific number and of desired loci



Incorporates structures needed for sequencing

http://www.rapid-genomics.com

RAPiD Seq

- "Fast, scalable, flexible number of markers, low amounts of DNA, and affordable"
- Repeatability over 99%
- Low missing data without imputation

Simulated multiplexing	Subset percentage	# SNPs*	Missing data (%)	Average Depth (x)
24	1	68,352	0.14	26
48	0.5	40,279	0.18	29
96	0.25	21,630	0.17	27
144	0.17	14,853	0.16	26
192	0.125	10,540	0.16	25
384	0.0625	4,808	0.15	24

>97%

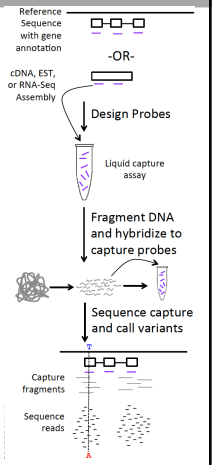
http://www.rapid-genomics.com

Reduced Representation Approaches

- Array based Genotyping
- Genotyping-by-Sequencing (GBS)
- Exome Capture (solid and liquid)
- Transcriptome as a proxy to the genome

(Exome) Capture Overview

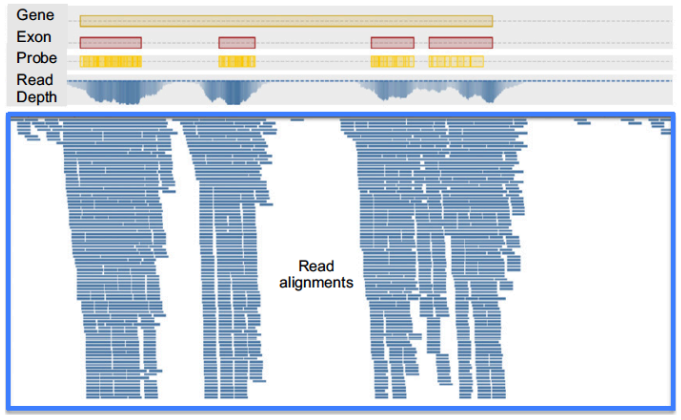
- Methods
  - Solid phase capture (not used much anymore)
    - Probes fixed to solid surface
    - DNA is hybridized and non-specific hybridization is removed through washes
    - Capture is eluted and sequenced
  - Liquid capture (far more common)
    - Biotinylated DNA or RNA probes are suspended in solution
    - Genomic DNA is captured through annealing
    - Captured sequence is eluted and sequenced
- Issues
  - Sensitivity
  - Specificity
  - How much to sequence to get all new alleles?



Exome Capture

- What to design probes from?
  - Sequenced genomic DNA, *de novo* transcript assembly, cDNA libraries, whole genome sequencing reads aligned to a closely related species
- Bias based on what was used for probe design
  - Lowly expressed genes/alleles not included
  - Genes not in reference assembly or not annotated as genes not included
- Need to take exon boundaries into consideration during probe design
  - Without intron sequence, poor capture efficiency

Visualizing Exome Capture Read Alignment



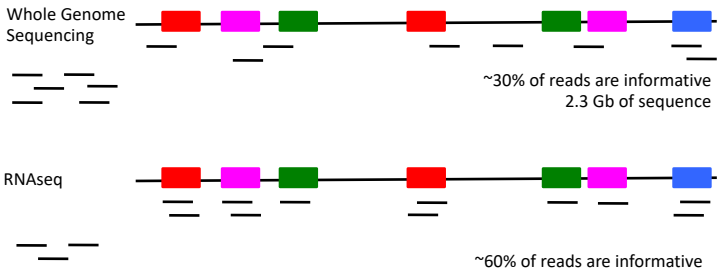
(Evans et al., The Plant Journal 2014)

Reduced Representation Approaches

- Array based Genotyping
- Genotyping-by-Sequencing (GBS)
- Exome Capture (solid and liquid)
- Transcriptome as a proxy to the genome

Transcriptome as a Proxy to the Genome

In maize multiple copies of genes
Extensive repetitive intergenic sequence (~85% transposable elements)



Limitation of RNAseq for Variant Detection??

- Genes/alleles must be expressed in the tissue used to detect variants
- Select a tissue that has a high percentage of genes expressed in it
  - Example: seedling tissue in maize has at least 66% of the annotated genes expressed (Sekhon and Lin et al., 2011)
- Ideal with highly homozygous inbred lines
  - Removes concerns of allele specific expression

Summary of Sequence Based Reduced Representation Genotyping Methods

Table 1: Advantages and disadvantages of approaches to interrogate genome diversity

	Advantages	Disadvantages
Whole Genome		
Whole-genome re-sequencing	-SNPs, InDels, CNV and PAV genome variation can be assayed -The entire genome is assayed	-Reference sequence needed -Can be expensive and not cost effective
Reduced representation		
Exome capture	-No reference sequence needed -SNPs, InDels, CNV and PAV[a] genome variations can be assayed	-*A priori* knowledge required -Capture bias -Initial capture design effort
Genotyping-by-sequencing (RAD/GBS)	-No reference sequence needed -Large number of polymorphic loci assayed -No *a priori* knowledge required	-Large amount of missing data -InDels and structural variants are less readily detected
Transcriptome sequencing	-No reference sequence needed -Can be assembled and used as a proxy for genic regions of the genome -Assay genome and transcriptome variation -Large number of polymorphic loci assayed	-Only assay-expressed genes -Allele-specific expression can result in mis-genotyped individuals -Cannot assay structural variation

Note: [a]Relative to the sequence space from which the probe set was designed

(Hirsch et al., Briefings in Functional Genomics 2014)

Reduced Representation Takeaways

- Using different methods to shrink down the effective size of the genome
- Decreases the cost of genotyping an individual
- Important to understand ascertainment bias that comes from different approaches
- Depending on the downstream application, the high missing data for some approaches can be problematic