First 12 nt in read used to calculate phasing. The algorithm in illumina that does the phase correction (identifies clutsers) etx assumes a random or even nucleotide distribution across all flow cells (assumes 25% chance of getting a A T C or G). So if you have RE at same position and those RE are all the same illumine will tell us that the reads are all crap. GBS modulates position of cut site so you don’t get uniformity across clusters in illumine. Making barcodes is nontrivial

Primer 1 reads the barcode. There is no second read.

Most significant technical issues:

-DNA quality

-DNA quantification

-accurate liquid handling

GBS vs RAD (Davey et all 2011 review in Nature Reviews Genetics)

-a lot in common

-RAD does random shearing, size selection and Ligate on Y adaptors. GBS does not use Y adaptors

-RAD can get at longer fragments

-according to Laval website, Laval does GBS.

When use a methylation sensitive cutter we only get fraction of genome that is unmethylated. Then sequencing gets an even smaller portion of genome.

We use a 6bp cutter Pst1, MsP1 is 4bp cutter

With 6bp size of library is smaller, and because the size of library is smaller you get better representation of the fragments in the library from sequencing.

If you want to increase depth of reads cut with rare 6bp cutter.

Do you need more markers at low depth or fewer markers at greater depth?

If you sequence your library again you will not recover the same fragments. You will have overlap but not identical fragments sequences. So if you use a longer cutter you’ll have higher probability of recovering the same fragments.

**Optimizing GBS in New Species**

Pst1 works well for most vertebrates. Generalizations don’t always hold. You look at distribution of fragment sizes. E.g., Pst1 doesn’t work for geese because Pst1 seems to be associated with repeat.

How do people deal with version control? Logging and audit trail. How do we make an “audit trail”?

Is everyone aware that its important to check MD5SUMS.

**Bias**

Ascertainment bias?

-we had a bad design for snp discovery.

-use all samples to discover snps, results in least amount of ascertainment bias. Basically u only recover snps so .

-Even in workshop video they report 33% of non-missing data