ABSTRACT: 163

AN OPTIMIZED GENOTYPING BY SEQUENCING (GBS) APPROACH FOR RAPID AND COST-EFFECTIVE SNP DISCOVERY AND GENOTYPING IN SOYBEAN

SONAH H, BASTIEN M, IQUIRA E, TARDIVEL A, BOYLE B, LAROCHE J AND BELZILE F*

Département de Phytologie, Faculté des Sciences de l'agriculture et de l'alimentation and Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Pavillon Charles-Eugène-Marchand, Québec, QC G1V 0A6, Canada

**E-mail: francois.belzile@fsaa.ulaval.ca*

Highly parallel SNP genotyping platforms have been developed for soybean such as a 1,536-SNP GoldenGate assay and, more recently, a 50k SNP Infinium assay. Unfortunately, for most individual researchers, these platforms carry a very high cost per sample. In contrast, recently developed genotyping by sequencing approaches offer a highly flexible and cost effective alternative for simultaneous SNP discovery and genotyping. In the present investigation, we have developed an optimized GBS library preparation protocol for soybean and an analysis pipeline to call SNPs and indels from the resulting sequence reads. We used a set of eight diverse soybean genotypes to conduct a pilot scale test of the protocol and pipeline. A total of 5.5 M reads (660k reads/sample) were obtained and processed by our pipeline which involves editing the reads to retain only soybean sequence, mapping the reads to the soybean genome, calling SNPs on the resulting alignments and imputing missing data. A total of 10,120 high quality SNPs were obtained and the distribution of these SNPs mirrored closely the distribution of gene-rich regions in the soybean genome. A total of 39.5% of the SNPs were present in genic regions and 52.5% of these were located in the coding sequence. Moreover, validation of over 400 genotypes at a set of randomly selected SNPs and InDels using Sanger sequencing showed a 98% success rate. Wider scale confirmation of the quality of the SNP data was obtained by applying this GBS approach to highly structured populations such as NILs or RILs, and the expected clustering and segregation patterns were observed, examples of which will be provided. In its current implementation (96-plex libraries sequenced on a single Illumina HiSeq lane), between 20,000 and 40,000 SNPs are typically obtained among sets of unrelated lines whereas over 1,000 SNPs can be expected to be genotyped in segregating progeny of a biparental cross. The approach to obtain high quality SNPs developed here will be helpful for marker assisted genomics as well as assessment of available genetic resources for effective utilisation in soybean breeding.

An optimized genotyping by sequencing (GBS) approach for rapid and cost-effective SNP discovery and genotyping in soybean

François Belzile
Dept. de phytologie/IBIS
Université Laval

WSRC
Durban
February 2013

UNIVERSITÉ LAVAL

Common complaints about markers

Could SNP genotyping help us overcome this?

Typical SNP platform development

1. Identify polymorphic regions
   - Sequence the same region in many individuals to identify specific positions that are polymorphic
2. Develop a multilocus genotyping assay
   - DNA “chips” for 96, 1536, 60K, 1M SNPs
     - Golden Gate assay in soybean (1,536 SNPs)
     - Infinium assay in soybean (~60K SNPs)
3. Perform the actual genotyping
   - Not all interrogated loci will prove polymorphic in any given collection of lines

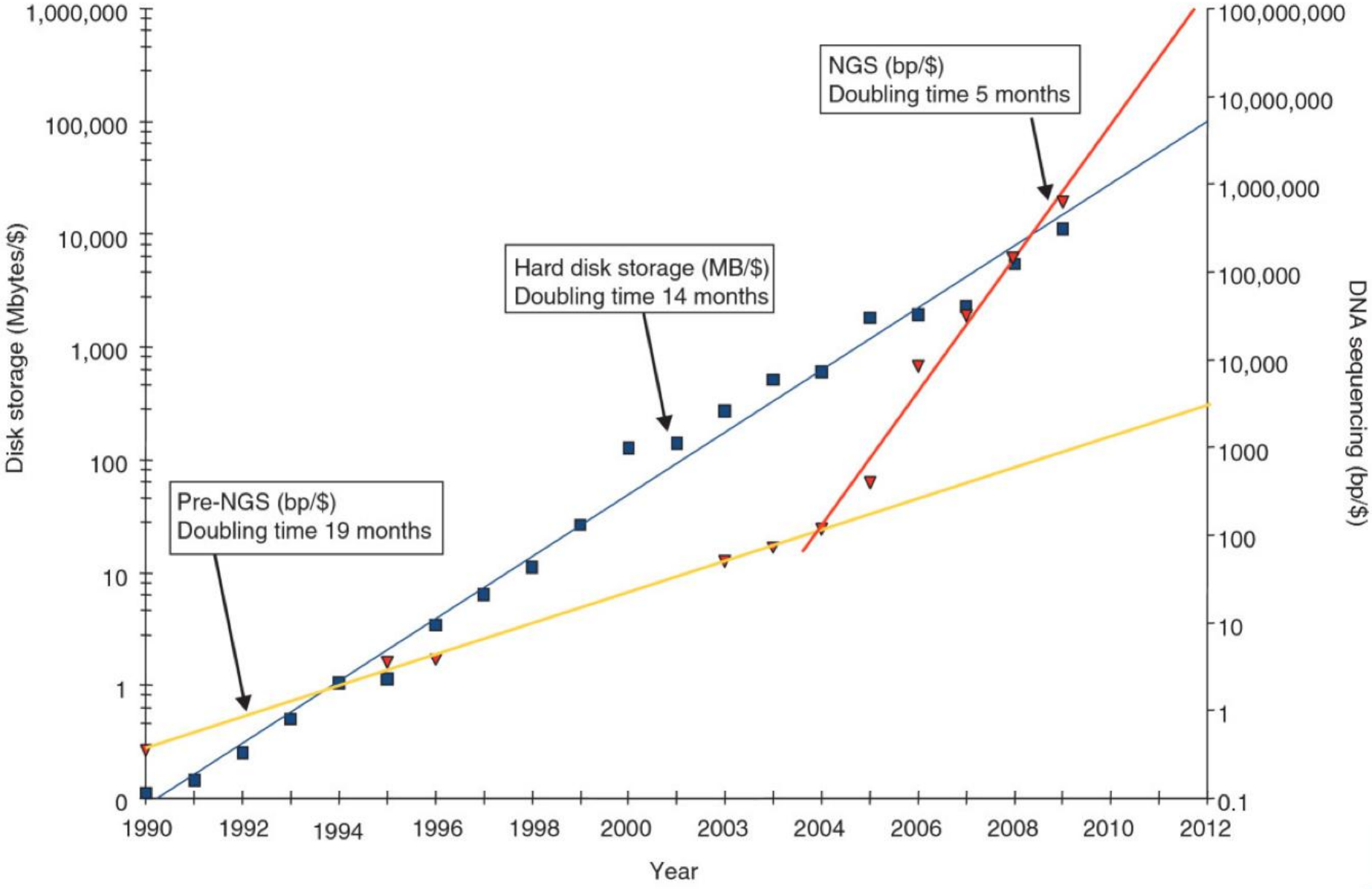
Number of Informative SNPs
Soybean GoldenGate Assay

SNPs polymorphic between pairs of genotypes (out of 1,536)

Population	Genotypes	Mean	95% CI
Elite	96	458	324 - 591
Landrace	96	544	355 - 732
Elite *vs.* Landrace	96 *vs.* 96	590	454 - 724
Northern Elite	75	443	320 - 566
Southern Elite	21	391	262 - 518
N. Elite *vs.* S. Elite	75 *vs.* 21	493	372 – 612

(D. Hyten, USDA; pers. comm.)

NextGen Sequencing a Game-Changer



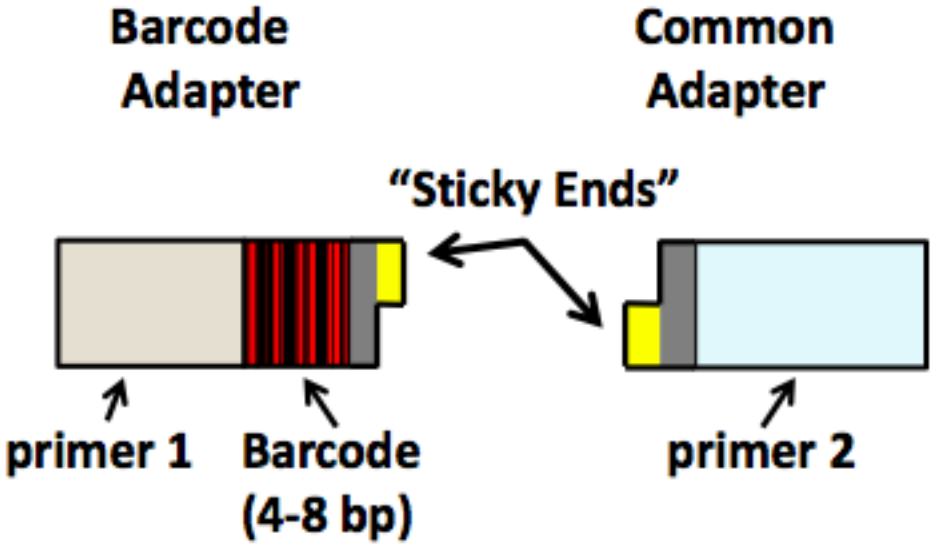
Genotyping by sequencing

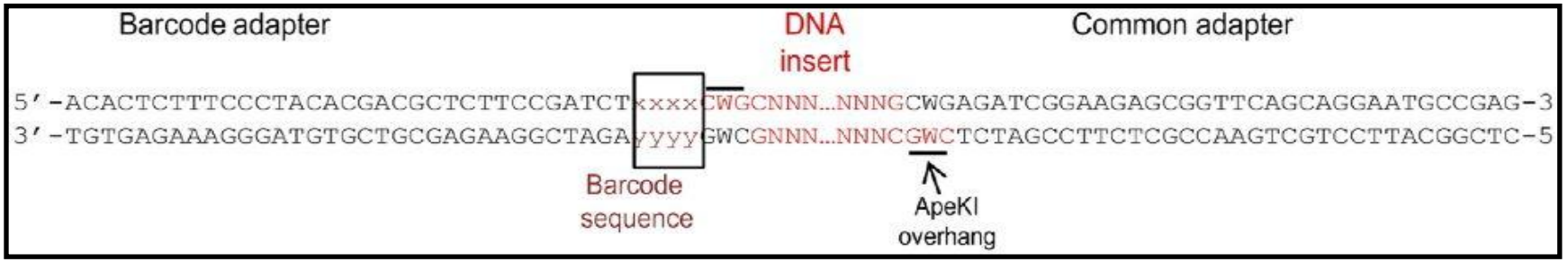
- Leverage the power of next-gen sequencing to simultaneously identify and genotype SNPs
  - RAD-seq ("Restriction site associated DNA sequencing")
    - Baird et al., 2008
  - GBS ("Genotyping by sequencing")
    - Elshire et al., 2011

Overview of GBS

A) Restriction digestion (*Ape*KI)

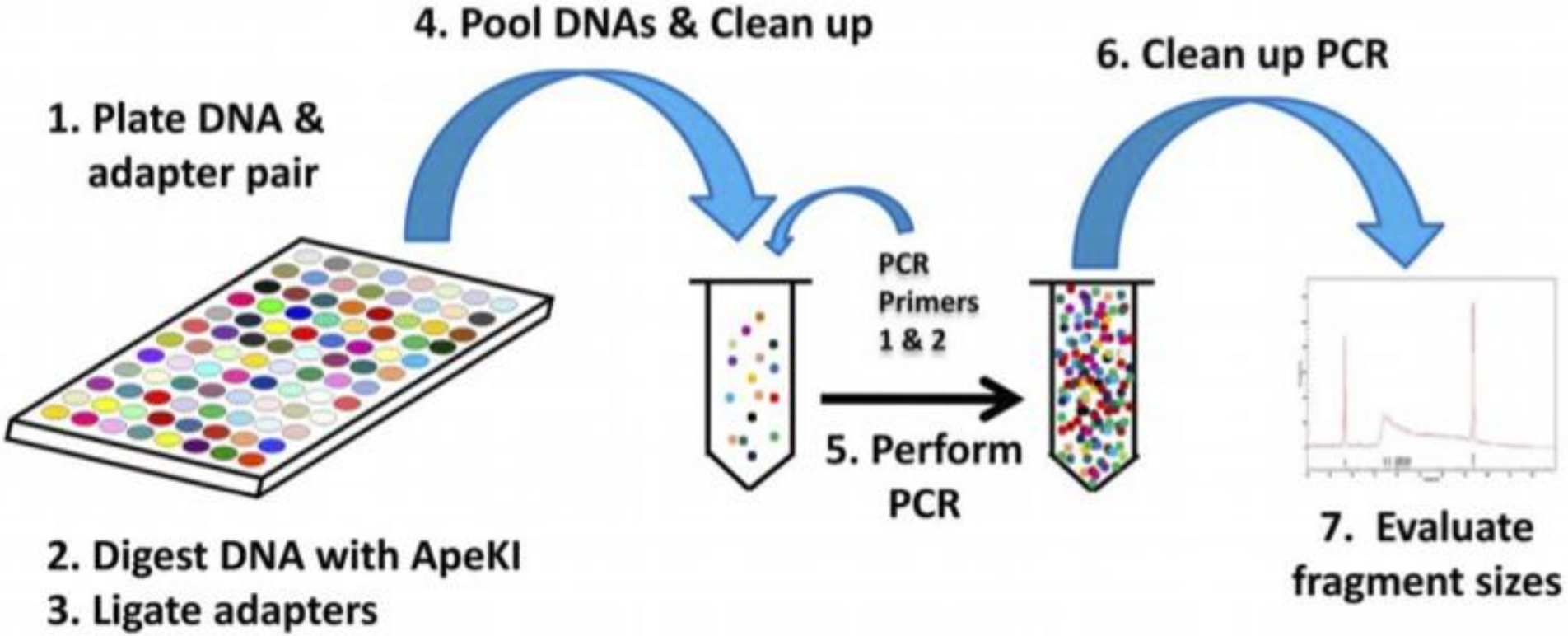
B) Ligation of two flanking adapters





Structure of a GBS amplicon

GBS: step by step



8. Sequence on Illumina machine
9. Bioinformatic analysis

Analysis pipeline

FASTQ file from sequencer

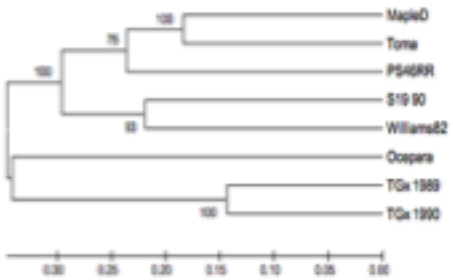
BC1 BC2 BC3 .....BC46 BC47 BC48

Sequence editing

Alignment onto reference genome

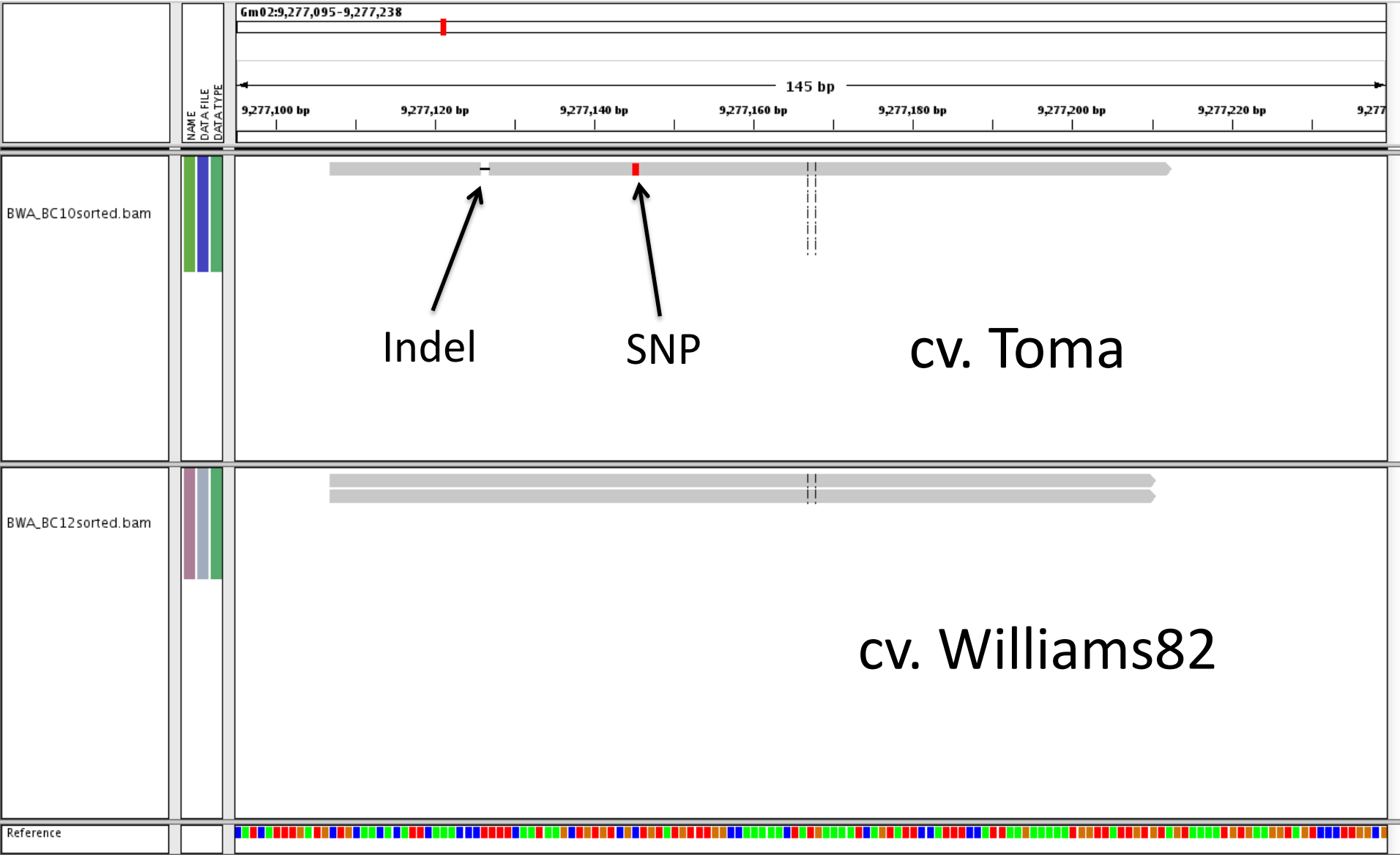
Identification and genotyping of SNPs

Visualisation of SNP markers



Validation of SNP markers

Types of polymorphisms detected



Results of pilot test

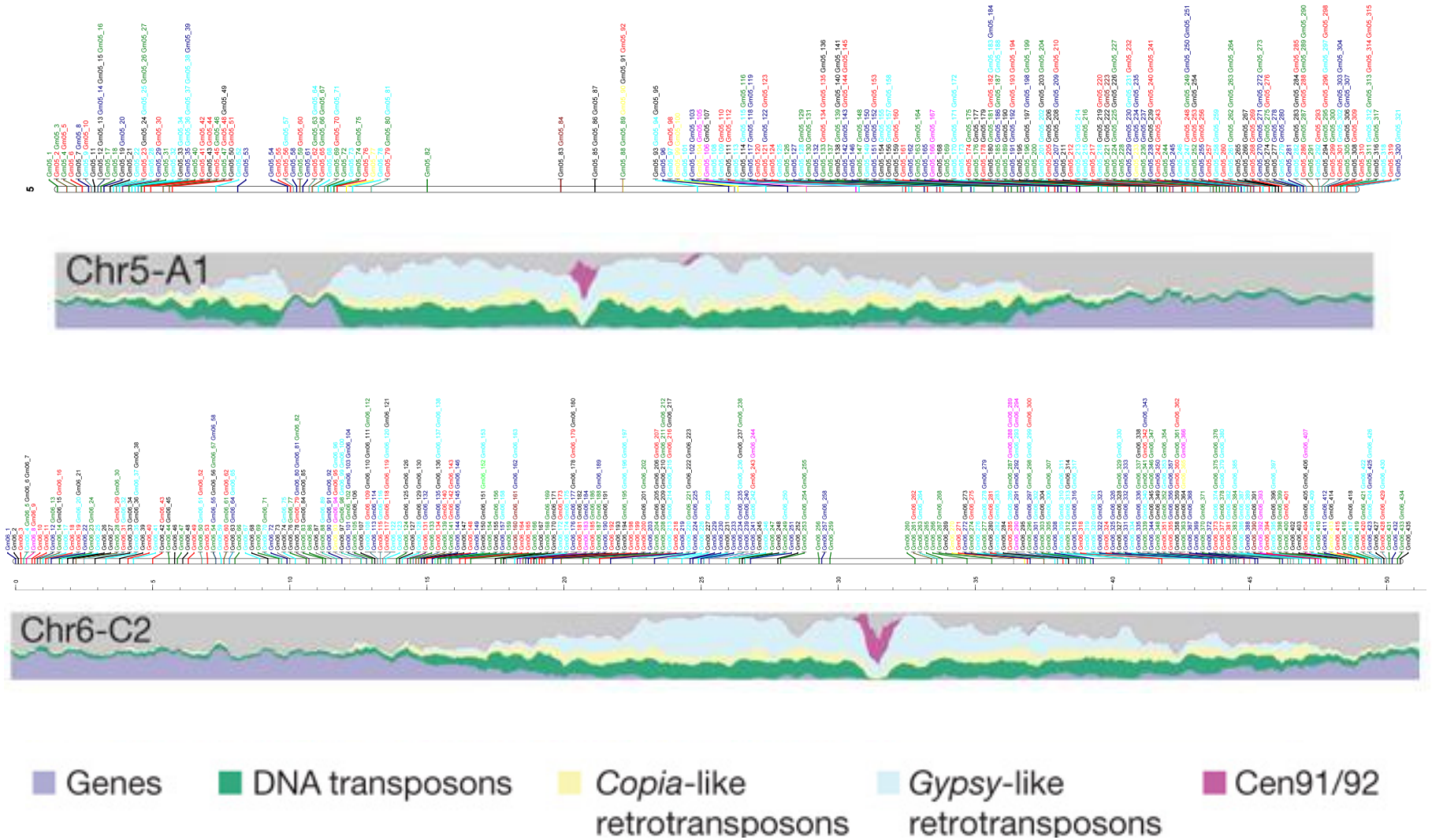
- 8 soybean lines on Illumina GAIIx (« old »)
- Number of reads
  - In total, >5,000,000 reads
  - Mean = 660,000 (440,000 to 1,005,000) reads
    - 50% of reads are unique
    - 50% of reads are seen at least twice

  ] in 1 DNA sample
- Excellent quality of the reads

Catalog of SNPs and indels

- Loci called polymorphic for the 8 test lines
  - 9,041 SNPs
  - ~ 500 indels
- Distribution on the chromosomes
  - ~500 SNPs/indels per chromosome
  - On average, 1 marker every 0.3 cM
  - Only a few hundred markers are not assigned to chromosomes (unmapped contigs/scaffolds)

Chromosomal distribution of SNPs

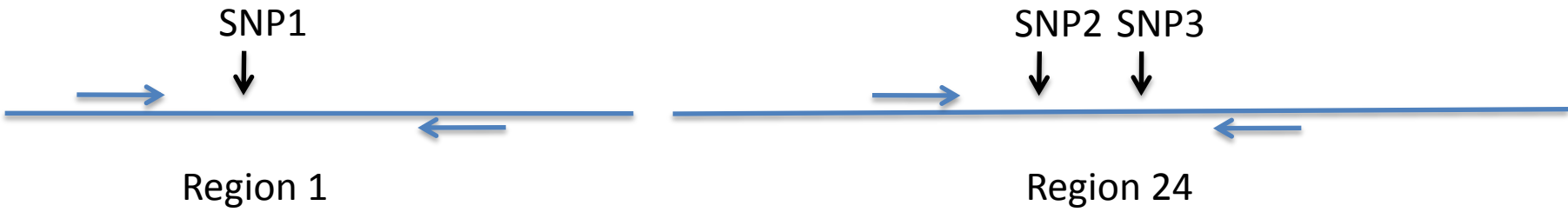


Validation of markers

- Sanger resequencing of SNPs ✔
- Phylogenetic analyses ✔
- Linkage analysis/mapping
  - Near isogenic lines ✔
  - Recombinant inbred lines ✔
- LD analysis and association mapping ✔

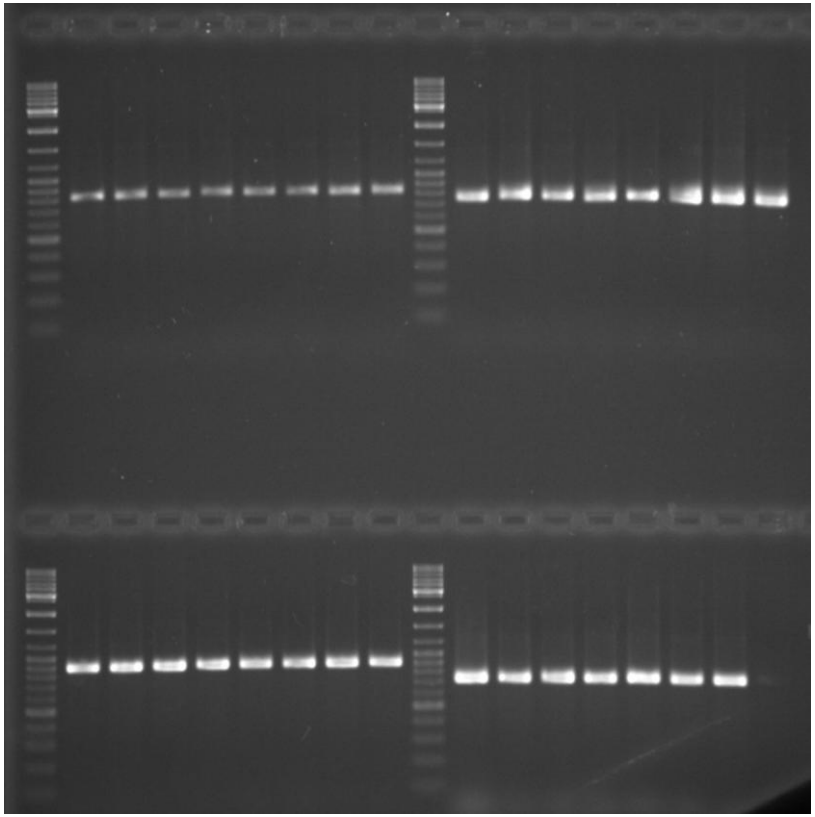
How good are the SNPs?

- Genomic region supposed to contain a SNP
  - Is there really a polymorphic site at this position?
  - Is the genotype called the right one?



Validation by resequencing

- 24 regions selected at random and PCR amplified on 2 sets of 8 soybean lines



- All regions were successfully amplified on practically all lines
- Sequencing showed that the predicted polymorphic site was indeed present in all cases

Validation by resequencing

❖1 SNP on chromosome 04 including 1 heterozygote

CHR	PLATFORM	SNP	BC01	BC02	BC03	BC04	BC05	BC06	BC07	BC08
Gm04	GBS	G/A	G	A	A	G	HETERO	A	G	G

Summary of validation tests

- Existence of a SNP in amplified region
  - Confirmed in all cases studied to date (24/24)
- Accuracy of the genotype calls
  - Over 400 genotype calls tested (24 PCR x 16 lines)
  - Accurate at >98%
  - Most often, the discordant genotype calls involve predicted heterozygotes

How many SNPs do you want?

- What are your goals?
  - Association analyses
    - maximize SNP number to achieve exhaustive genome coverage
    - How many lines?
    - How much linkage disequilibrium?
  - Genetic mapping in biparental populations
    - Hundreds of markers usually sufficient
    - Possibly increase multiplexing (number of individuals per library)

- In GBS, this can be achieved by varying the degree of complexity reduction

Complexity reduction: how much and how?

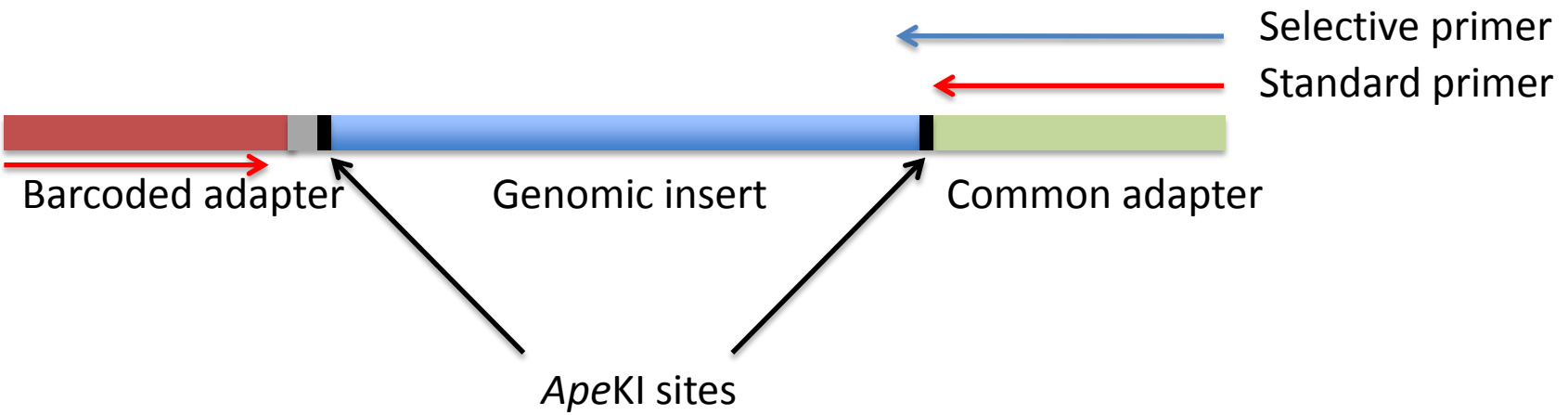
- Option 1: reducing complexity through the choice of an appropriate restriction enzyme

Enzyme	Site	# of sites	Coverage[1]
ApeKI	G*CWGC	807,990	2
PstI	CTGCA*G	125,238	12
SbfI	CCTGCA*GG	4,521	346

- Option 2: amplify and sequence only a subset of the amplicons using primers with selective bases

Tailoring GBS with selective amplification

- The idea: to amplify only a subset of fragments to decrease SNP number and increase depth



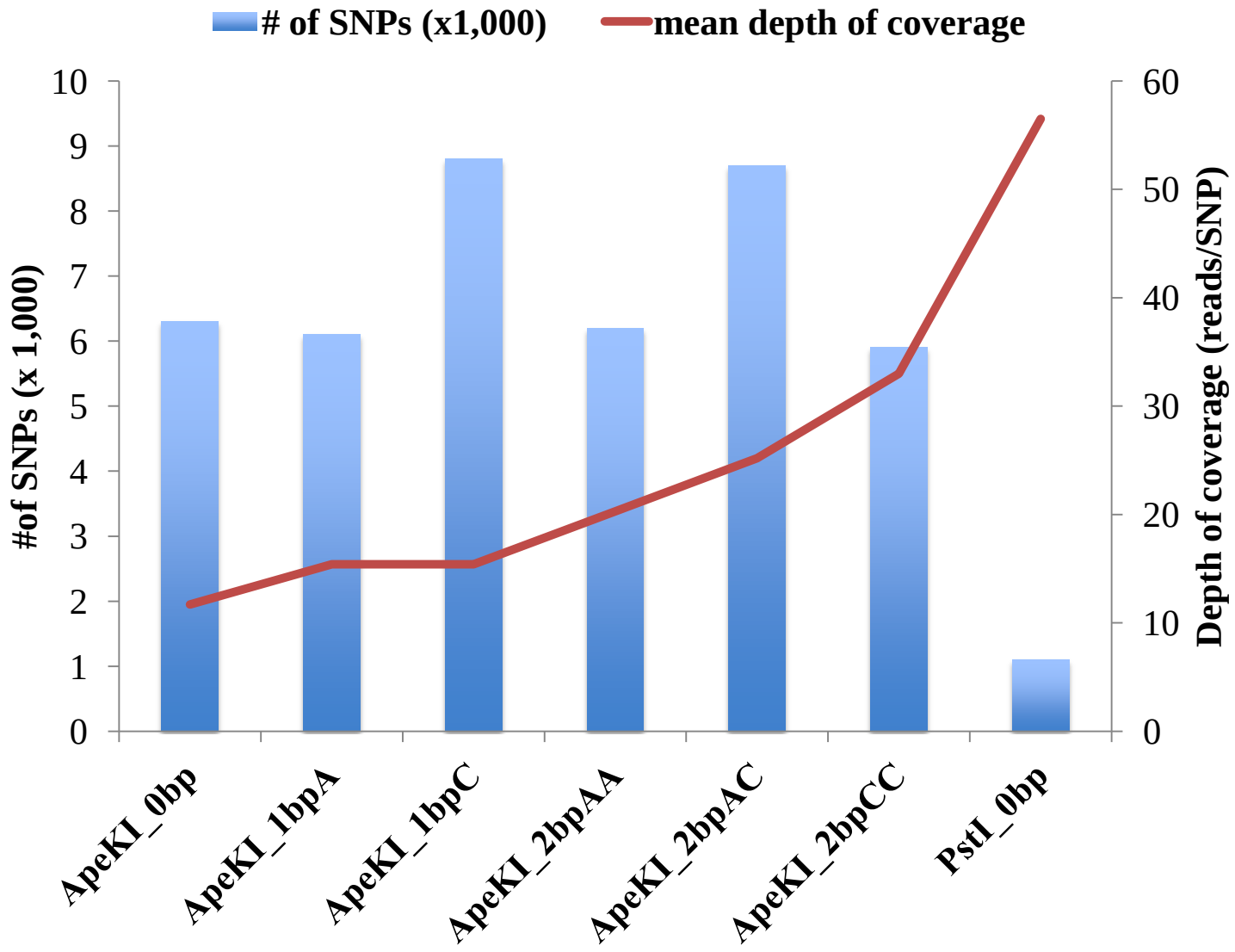
3’ AACGWCTCTAGCCTTCTCGCCAAGTCGTTTACGGCTCTGGCTAGAGCATACGGCAGAAGAC

3'
ACGWCTCTAGCCTTCTCGCCAAGTCGTTTACGGCTCTGGCTAGAGCATACGGCAGAAGACGA
AC 5’ 3’

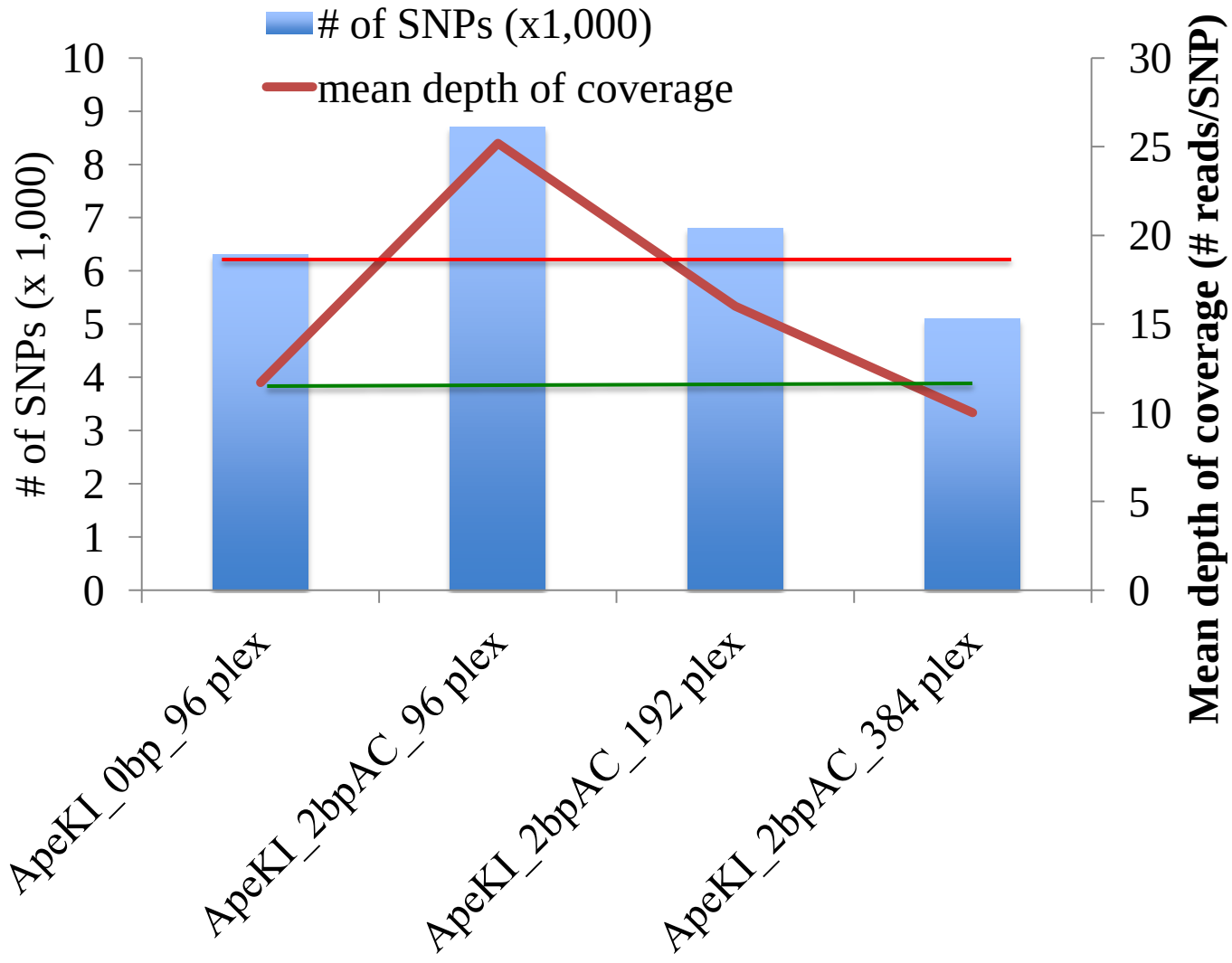
5’ nnnnGCWGAGATCGGAAGAGCGGTTCAGCAAATGCCGAGACCGATCTCGTATGCCGTCTTCTG
3’ nnnnCGWCTCTAGCCTTCTCGCCAAGTCGTTTACGGCTCTGGCTAGAGCATACGGCAGAAGAC

*Ape*KI

Impact of enzyme and selective primers in soybean (2 cultivars)



Increased multiplexing potential



OPEN ACCESS Freely available online

PLOS ONE

An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping

Humira Sonah[1], Maxime Bastien[1], Elmer Iquira[1], Aurélie Tardivel[1], Gaétan Légaré[2], Brian Boyle[2], Éric Normandeau[3], Jérôme Laroche[4], Stéphane Larose[4], Martine Jean[1], François Belzile[1]*

1 Département de Phytologie and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, **2** Plateforme d'analyses génomiques and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, **3** Département de Biologie, and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada, **4** Plate-forme de bio-informatique and Institut de biologie intégrative et des systèmes, Université Laval, Quebec City, Quebec, Canada

January 2013

Using GBS to characterize various germplasm collections

- Public breeders' germplasm from Canada (305 lines)
  - Current and past registered varieties
  - Advanced breeding lines/a few foreign accessions
- Maturity panel (144 lines) (#161)
  - NILs for various maturity loci/genes
  - Various very early maturing lines
- IITA soybean breeding program (300 lines) (#352)
  - Nodulation and ASR resistance
- Adapted lines (130 lines) and PIs (110 lines) (#162)
  - Association mapping of sclerotinia resistance

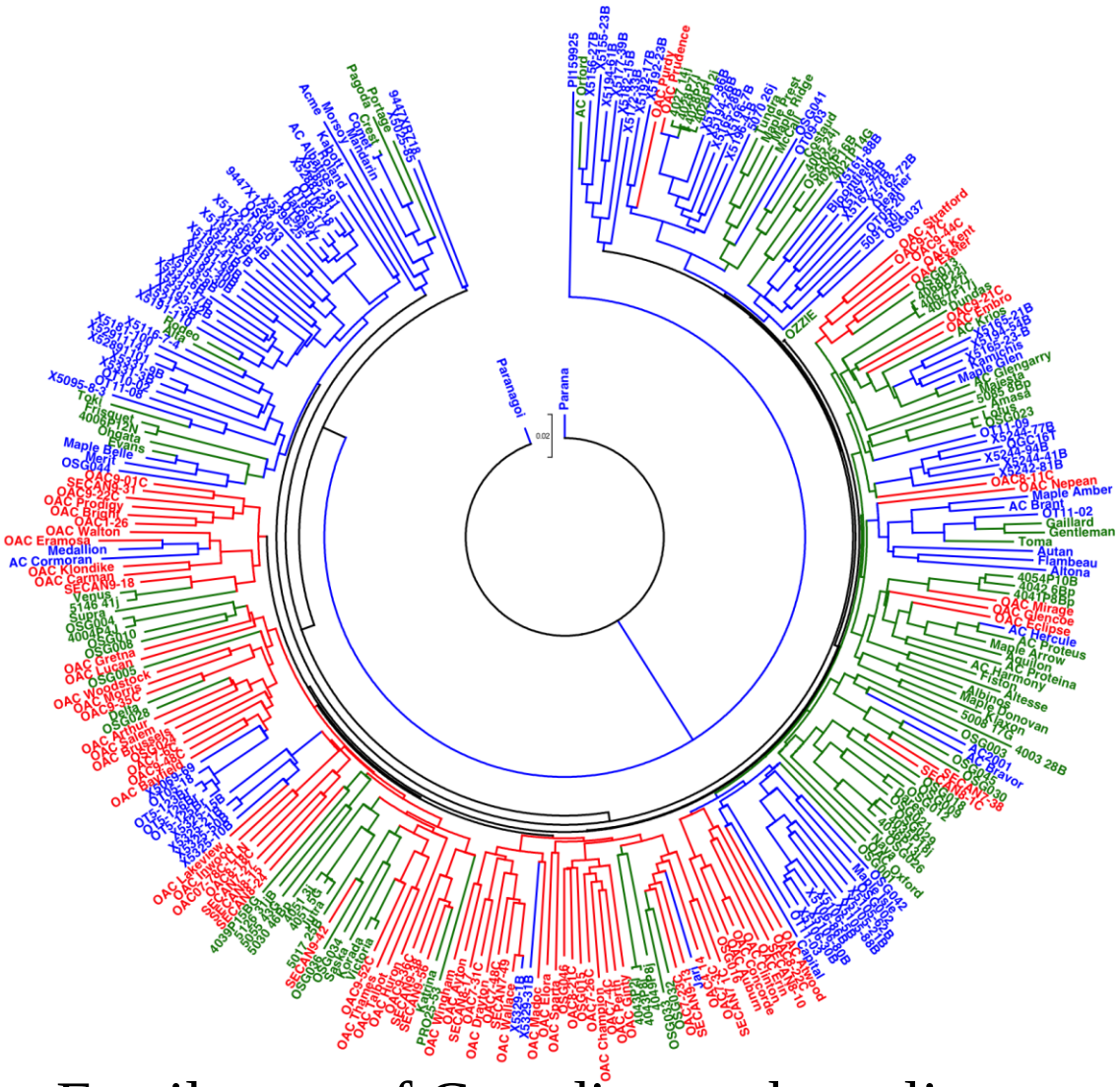
SNP genotyping of Canadian lines

- Originally to be done using the Universal Soybean Linkage Panel 1.0
  - Commercially available platform for genotyping up to 1,536 SNP loci in soybean
  - Based on previous work done at USDA, we expected between 700 and 1,000 SNPs to be informative in a population of this size
- Conducted using a genotyping by sequencing (GBS) as per Sonah et al. 2013

Results of genotyping by GBS (Canadian public breeding-305 lines)

- Obtained >40,000 informative markers
  - 50-80x more than anticipated using USLP1.0
  - Similar to what we could have obtained with the Infinium assay (but cheaper)
- >23,000 are “high-quality”
  - Mean number of reads/SNP/line ≥ 2
  - <20% missing data
  - Imputation can accurately “fill in the blanks”

Population structure in Canadian lines



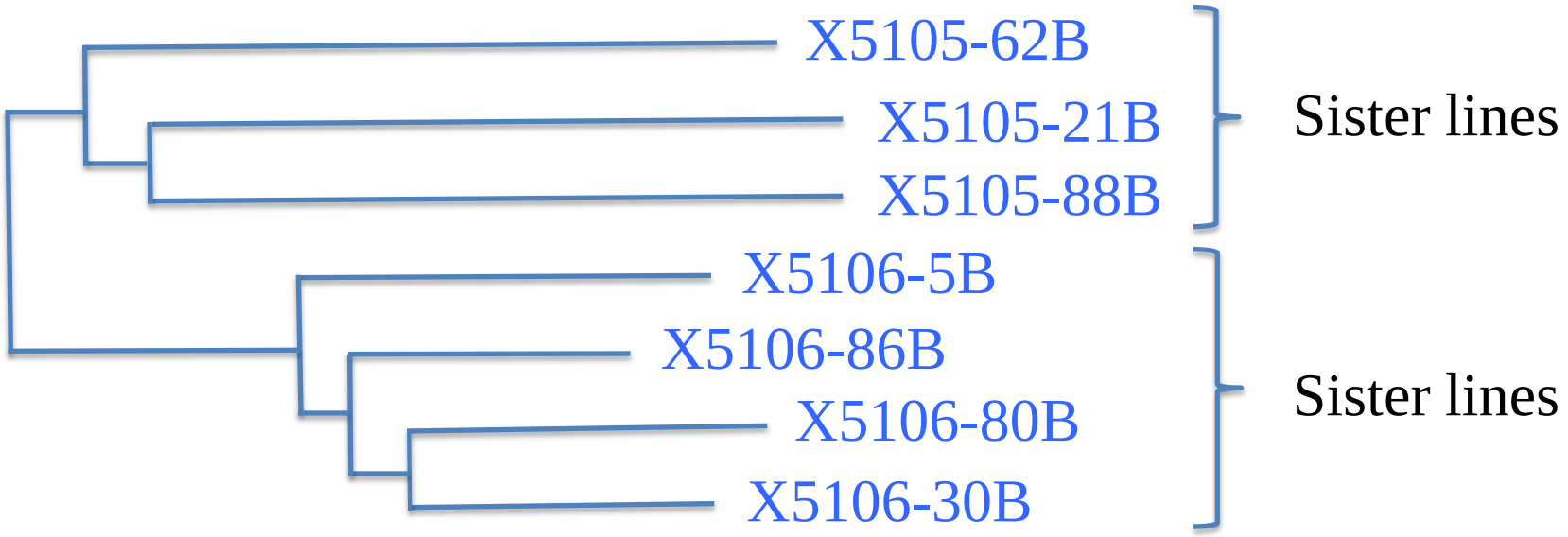
Family tree of Canadian soybean lines

Red = Guelph

Blue = Ottawa

Green = Beloeil

Population structure in Canadian lines
Let's zoom in!

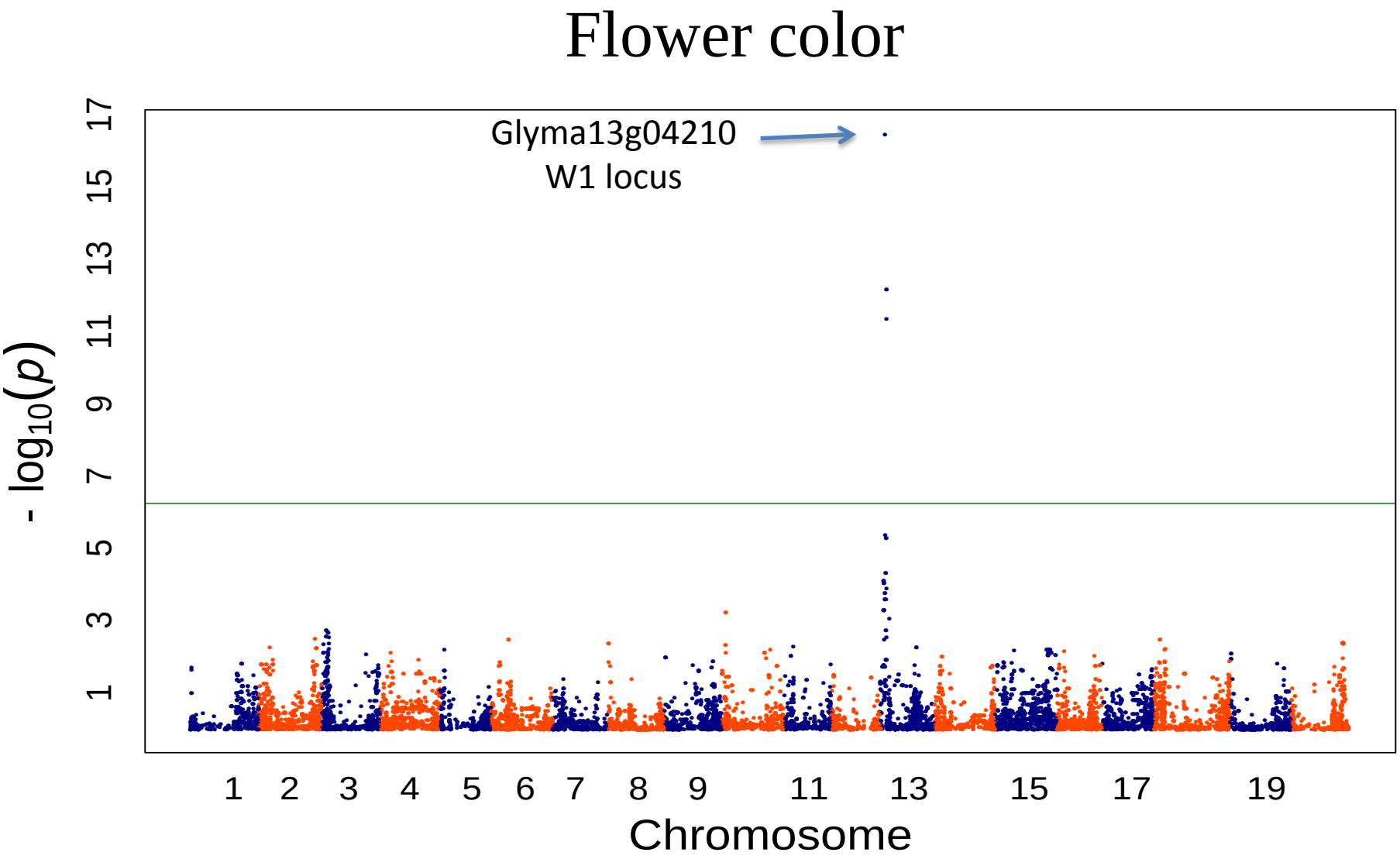


The architecture of the tree makes great sense in that lines derived from the same cross (i.e. same parents) group together

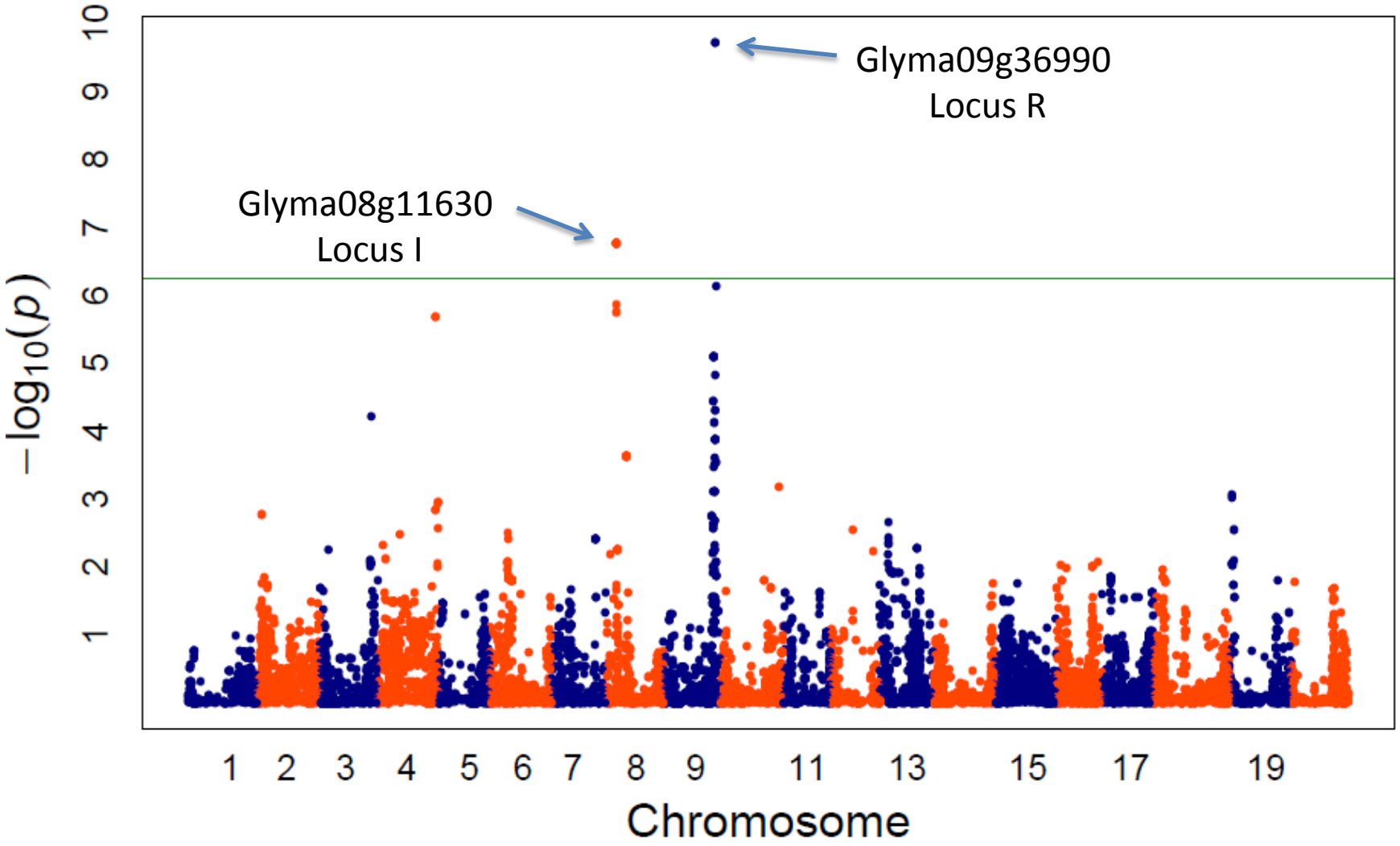
Association mapping

- Perform association tests on simple traits as a trial run
- Obtain field data to characterize the phenotype of a common set of lines (2012 and following)
  - Height
  - Maturity
  - Seed traits
- Perform association tests on agronomic traits

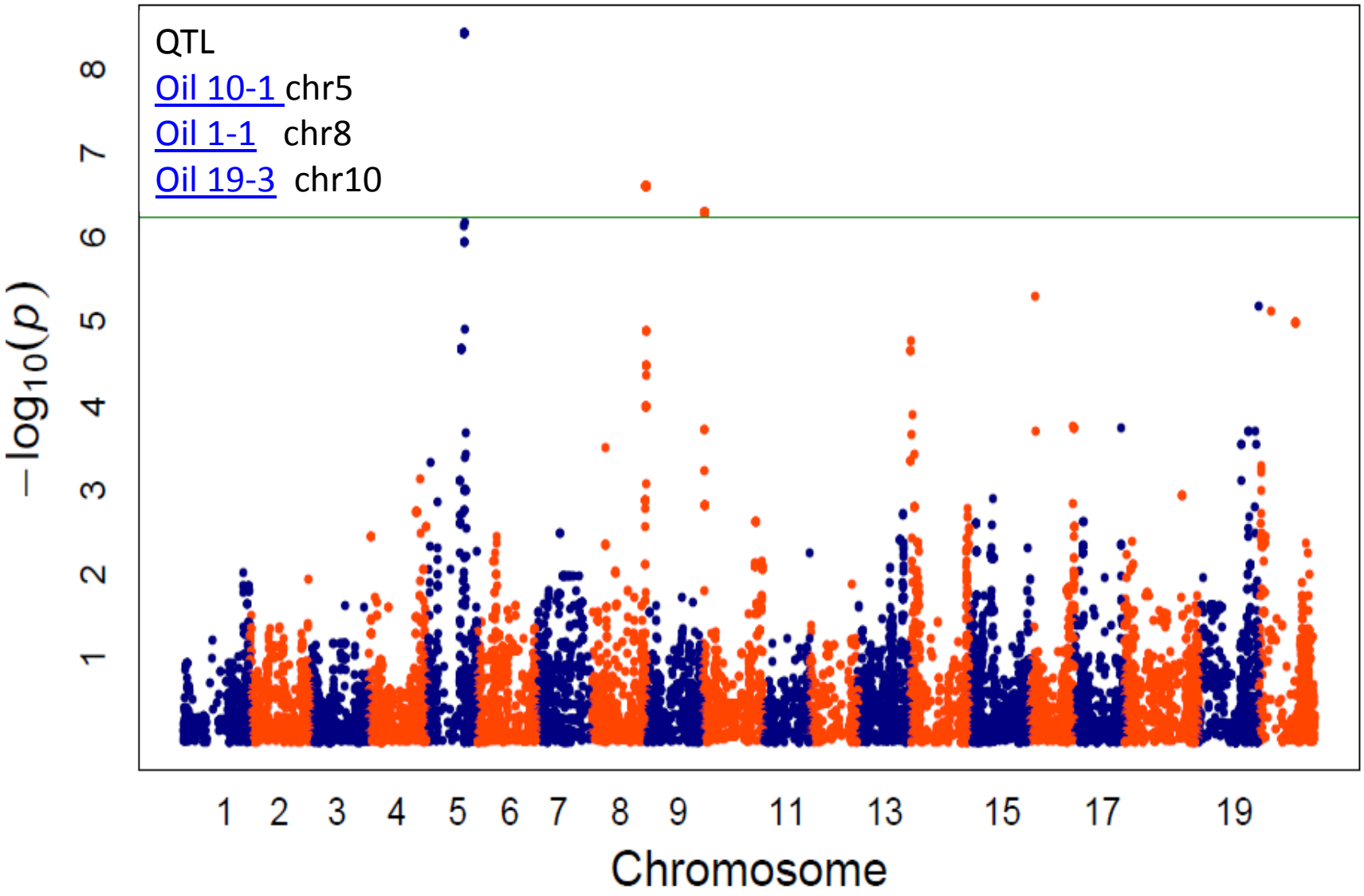
Association tests-simple trait



Hilum color


Glyma09g36990
Locus R
Glyma08g11630
Locus I
−log10(p)
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 11 13 15 17 19
Chromosome

Oil content


QTL
Oil 10-1 chr5
Oil 1-1 chr8
Oil 19-3 chr10
−log10(p)
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
11
13
15
17
19
Chromosome

On-going work

- Continued efforts to improve the genotyping capacity of the GBS pipeline
- Continued efforts to identify and validate candidate QTLs in the context of marker-assisted breeding
- Exploration of genomic selection as an alternative selection strategy

Acknowledgements

UNIVERSITÉ LAVAL

Belzile Lab

Humira Sonah
Maxime Bastien
Elmer Iquira
Aurélie Tardivel
Martine Jean

Bernatchez Lab

Éric Normandeau

Sequencing lab

Gaétan Légaré
Brian Boyle

Bioinformatics

Jérôme Laroche
Stéphane Larose

Cornell University

Ed Buckler's group

UNIVERSITY of GUELPH

Istvan Rajcan

CEROM
Centre de recherche sur les grains inc.

Louise O'Donoughue

NSERC CRSNG

Canadian International Development Agency

Agriculture and Agri-Food Canada

Elroy Cober

GBS as a service

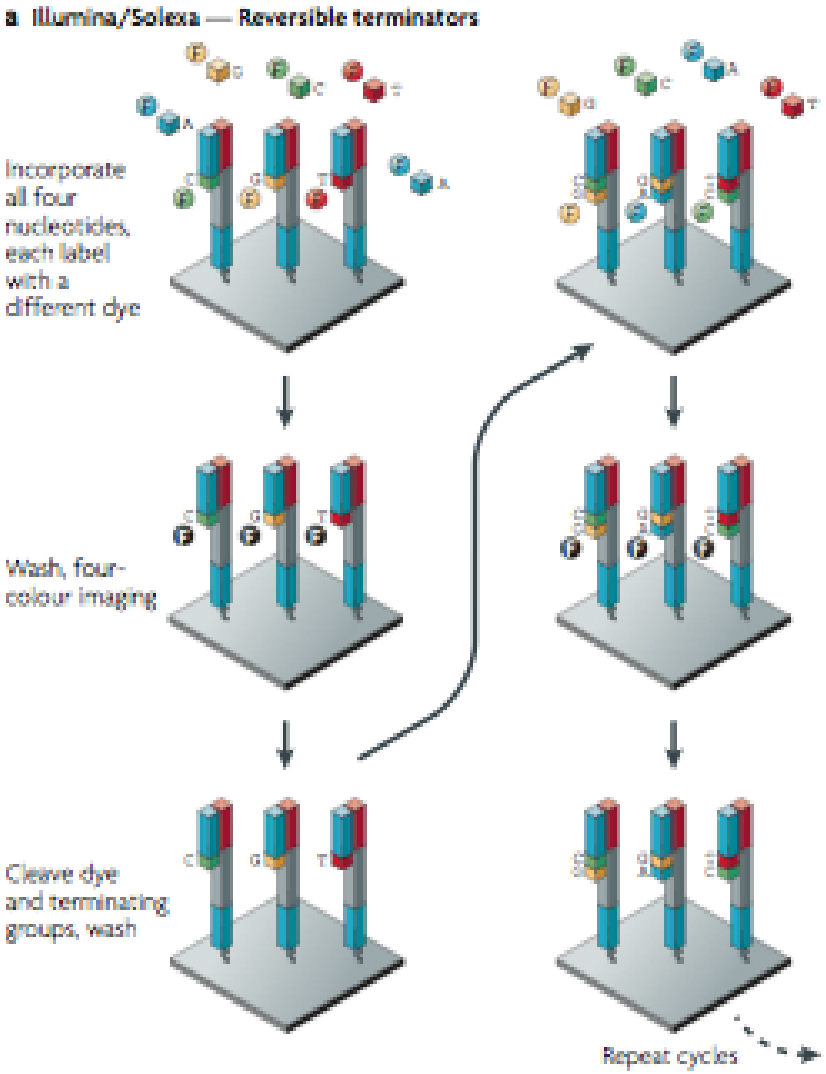
Offered by the Sequencing Service Lab at the Institute for Integrative and Systems Biology (IBIS) at Université Laval (Canada)

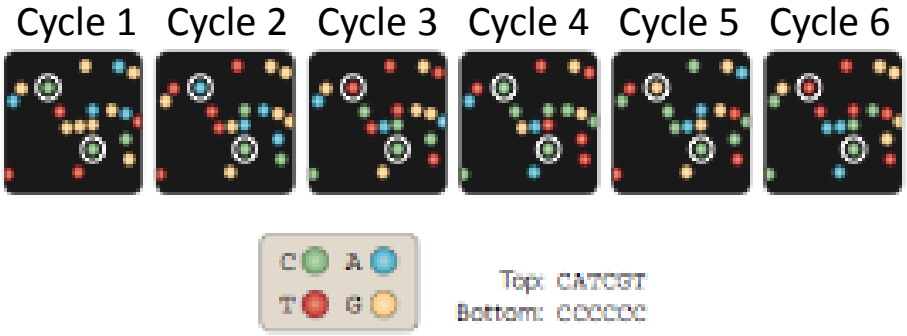
Contact Dr. Brian Boyle at: Brian.Boyle@ibis.ulaval.ca

Thank you for your attention

QUESTIONS?

Illumina sequencing by synthesis





- At each cycle, one and only one labeled base is added
- Imaging at each cycle determines which base had been added to each cluster
- After 100 cycles, the sequence of exactly 100 bases has been determined for each cluster and this produces a "read"
- For each base in the read, a quality score is assigned to indicate how good the base call was

(Metzker, 2010)

How much does it cost?

- DNA extraction ~5$
- GBS library preparation (96-plex) ~7-10$
- Sequencing (96-plex) ~12-15$

For less than 30$ per sample, we can obtain many thousand SNPs per individual

How long does it take?

- DNA extraction and quantification/QC — 1 wk.
- GBS library preparation (UL-IBIS) — 1 wk.
- Sequencing — ~4 wk.
- Analytical pipeline — ~2 wk.

In about 2 months, you can obtain a lot of data with very little hands-on time