Exploring the genomes of elite short-season soybeans in Canada

Davoud Torkamaneh and François Belzile

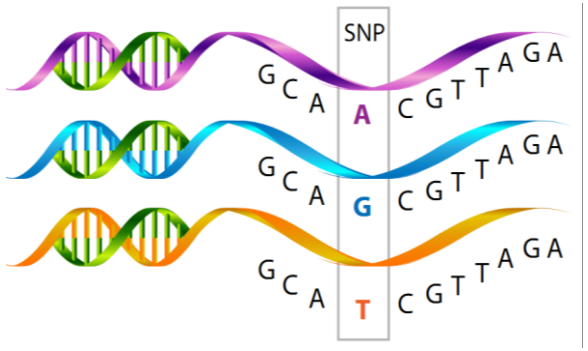
Outline

1. Definitions
2. Can we extensively characterize the nucleotide variation in short-season soybean?
    2.1. How many samples need to be sequenced?
    2.2. Can we create an accurate catalogue of SNPs?
3. Can new bioinformatics tools allow us to discover and genotype large structural variants (SVs) in soybean?
    3.1 Which types of SVs can we detect?
    3.2 At what level of accuracy?

Types of genetic variation

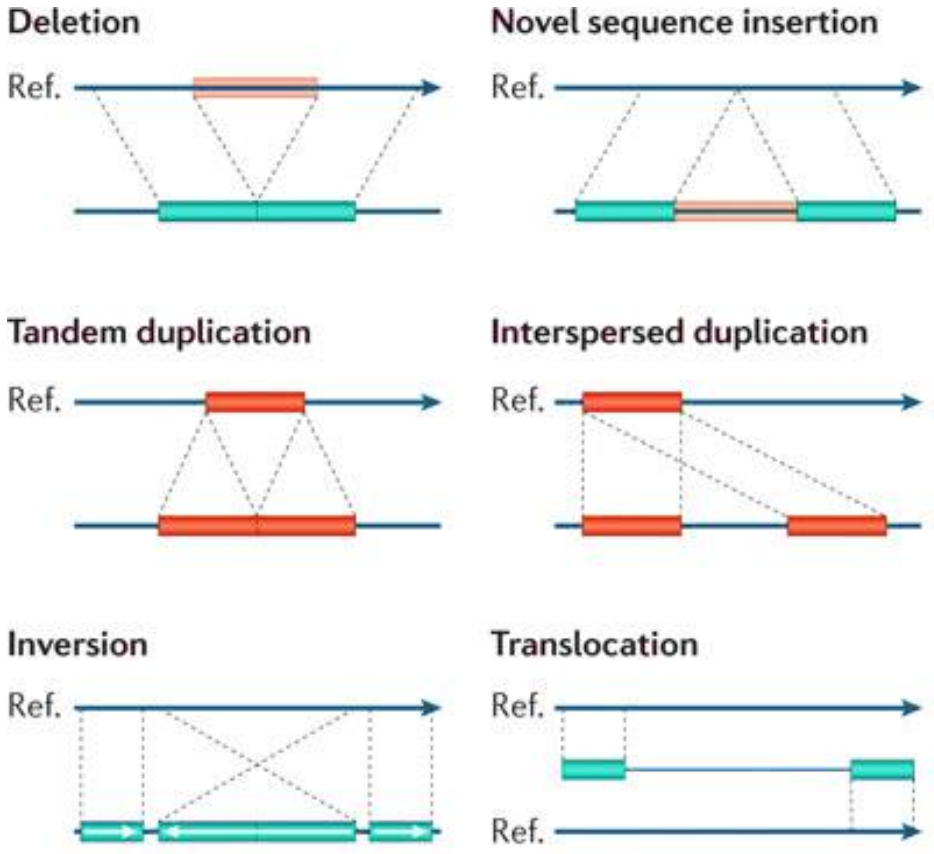
At the nucleotide level

- Single Nucleotide Polymorphism (SNPs)
- Multiple Nucleotide Polymorphism (MNPs)
- Small insertions and deletions (Indels)



At the structural level

- Deletions
- Insertions
- Duplications
- Inversions
- Translocations
- CNVs



Discovery of genetic variants

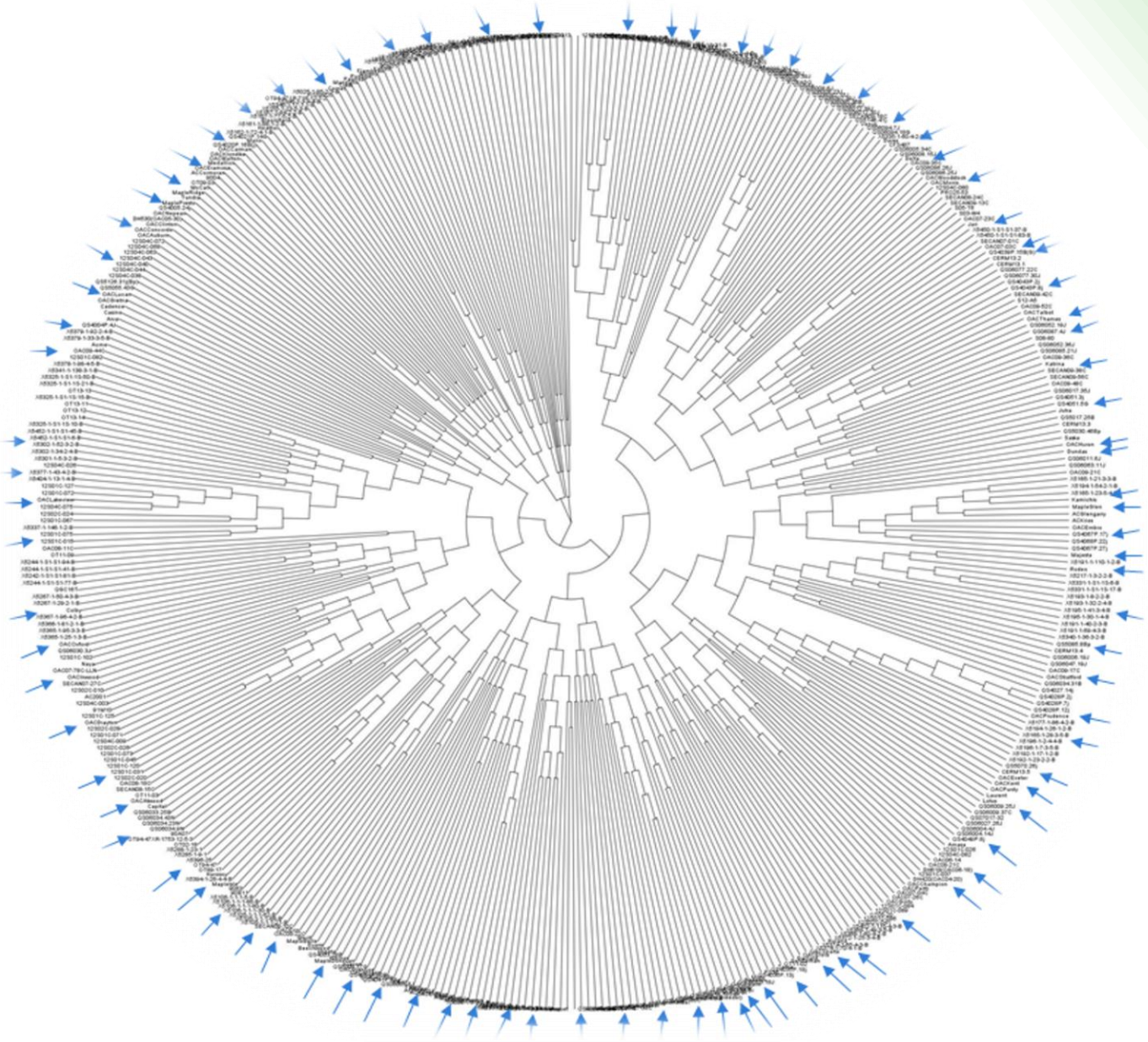
- **Can be achieved most extensively through whole-genome sequencing (WGS) of lines that are of interest**
  - Selection of lines
  - Library preparation
  - Sequencing
  - Bioinformatics to extract useful information about variants from mountains of sequence data

How do we select the samples for WGS?

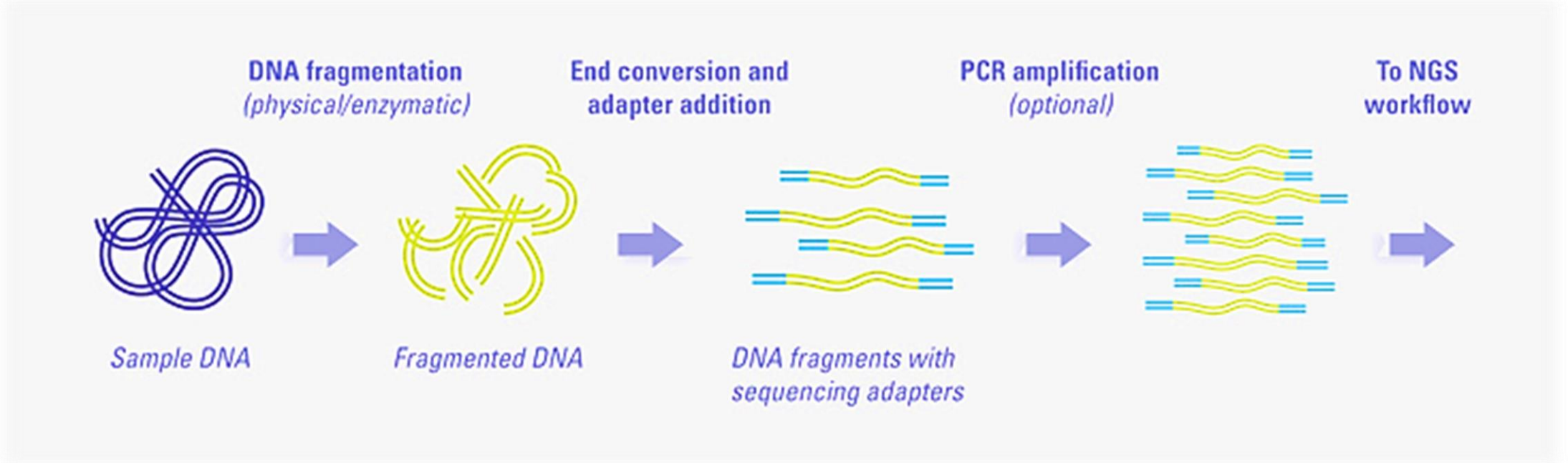
I. 441 Canadian soybean accessions were genotyped using GBS

II. A cladogram was constructed using GBS data for 441 accessions

III. Among these, 102 accessions were selected (arrows) for WGS by selecting accessions residing in each of the major branches of the tree

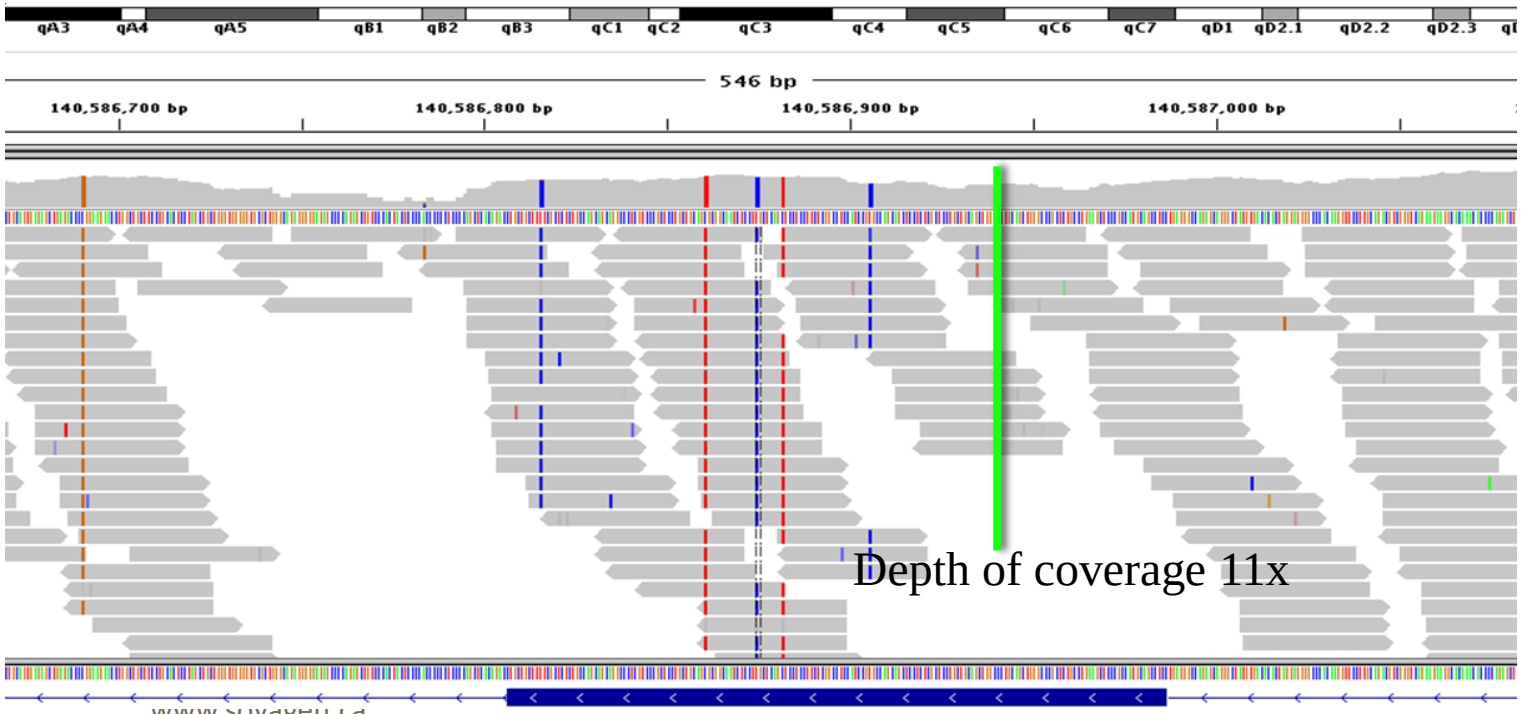


NGS library preparation



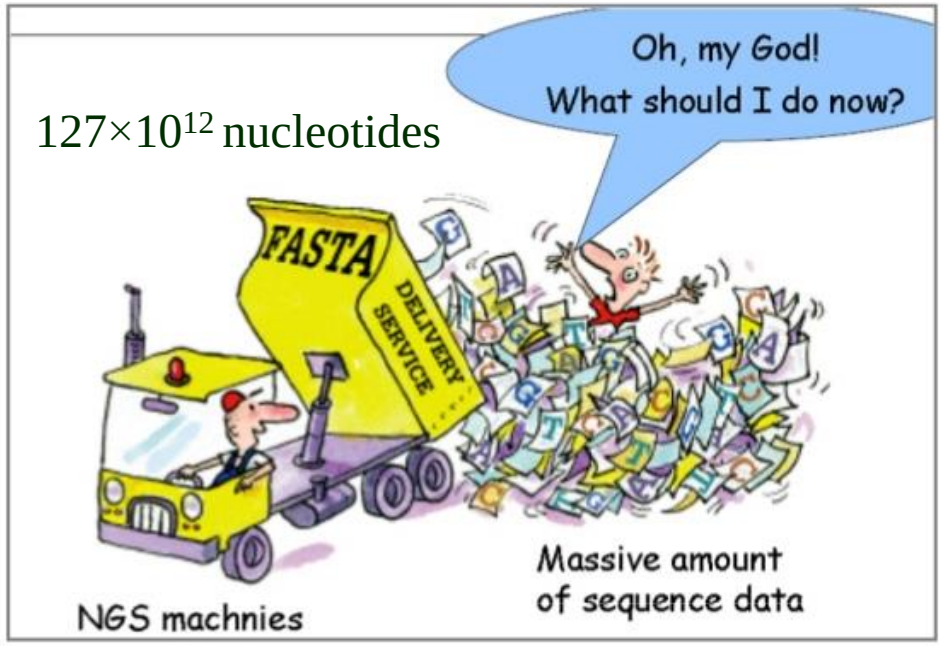
Sequencing

- ✓ Illumina HiSeq
- ✓ 127 trillion nucleotides
- ✓ Median depth of coverage 11x
- ✓ Covering 97.6% of the *G. max* genome sequence (at least one read)



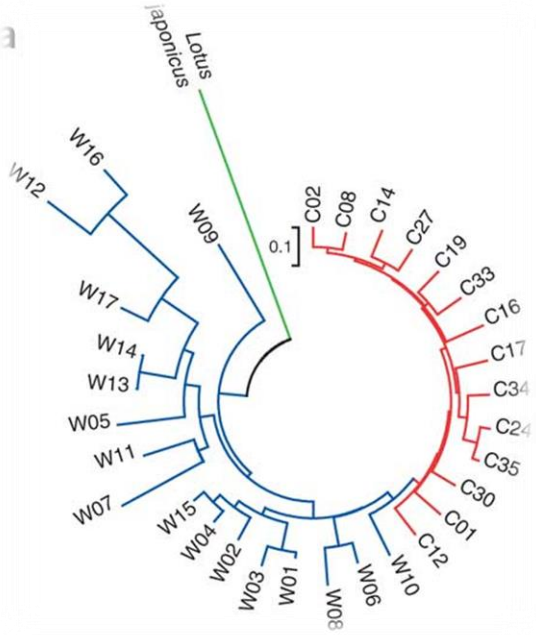
- ✓ Suitable depth of coverage
- ✓ Excellent genome coverage

The challenge of NGS data analysis
Analytical pipeline



Bioinformatics pipeline

SOAPsnp
Fast-WGS



Comparison of two WGS pipelines

- ✓ Comparison of our new analytical pipeline (Fast-WGS) with SOAPsnp, the pipeline used in most/all prior WGS in soybean

Pipeline/Variants	SNPs	MNPs	Indels	Computing time*
Fast-WGS	4,071,378	284,836	642,015	81 hours
SOAPsnp	4,124,216	ND	512,418	261 hours

- ✓ 7% more variants called by Fast-WGS
- ✓ Detection of MNPs by Fast-WGS
- ✓ Fast-WGS runs 3 times faster than SOAPsnp

Genotype accuracy

- Shared data:

19 shared accessions x 35,481 SNP loci =

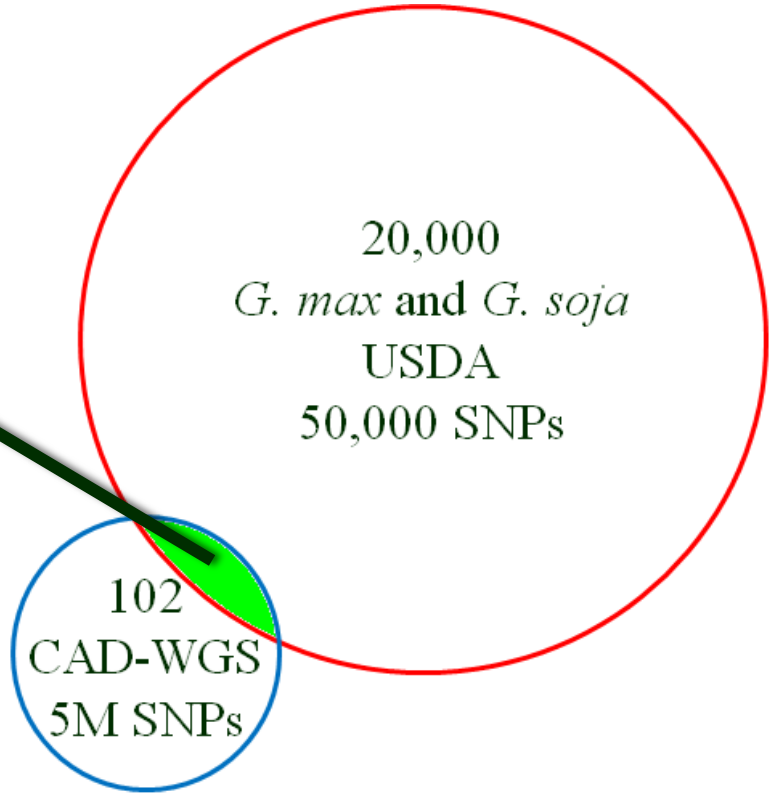
674,139 genotypes (data points)

OPEN ACCESS Freely available online

PLOS ONE

Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean

Qijian Song[1], David L. Hyten[1¤], Gaofeng Jia[1], Charles V. Quigley[1], Edward W. Fickus[1], Randall L. Nelson[2], Perry B. Cregan[1*]



Plant Biotechnology Journal

Plant Biotechnology Journal (2017), pp. 1–11

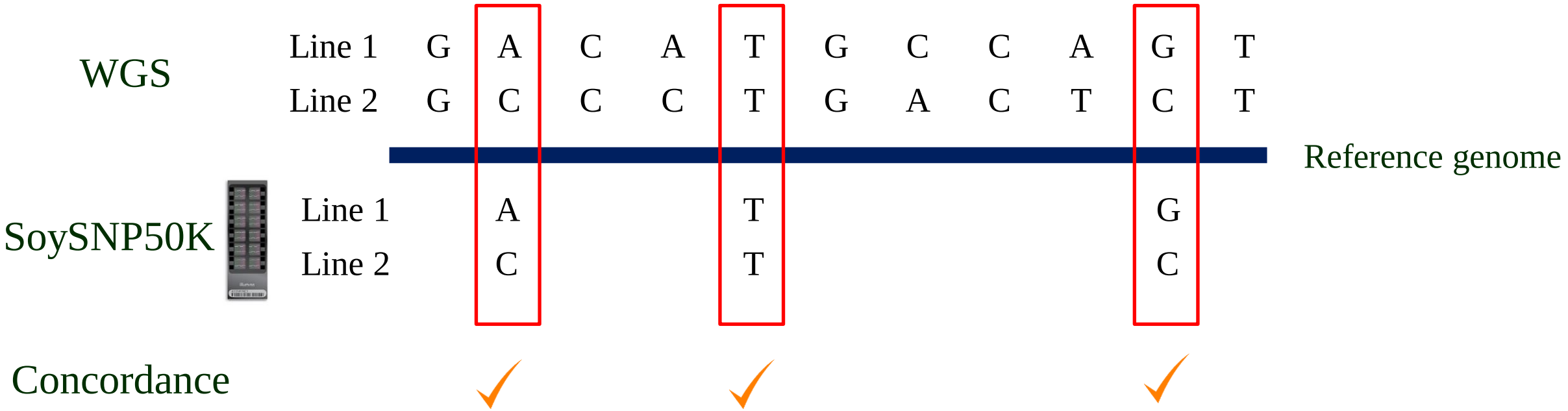
doi: 10.1111/pbi.12825

Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean

Davoud Torkamaneh[1,2], Jérôme Laroche[2], Aurélie Tardivel[1,2,3], Louise O'Donoughue[3], Elroy Cober[4], Istvan Rajcan[5] and François Belzile[1,2,*]

Genotype accuracy

Are the genotype calls made through WGS in agreement with the calls made using the SoySNP50K chip at the same nucleotide position in the same accession?



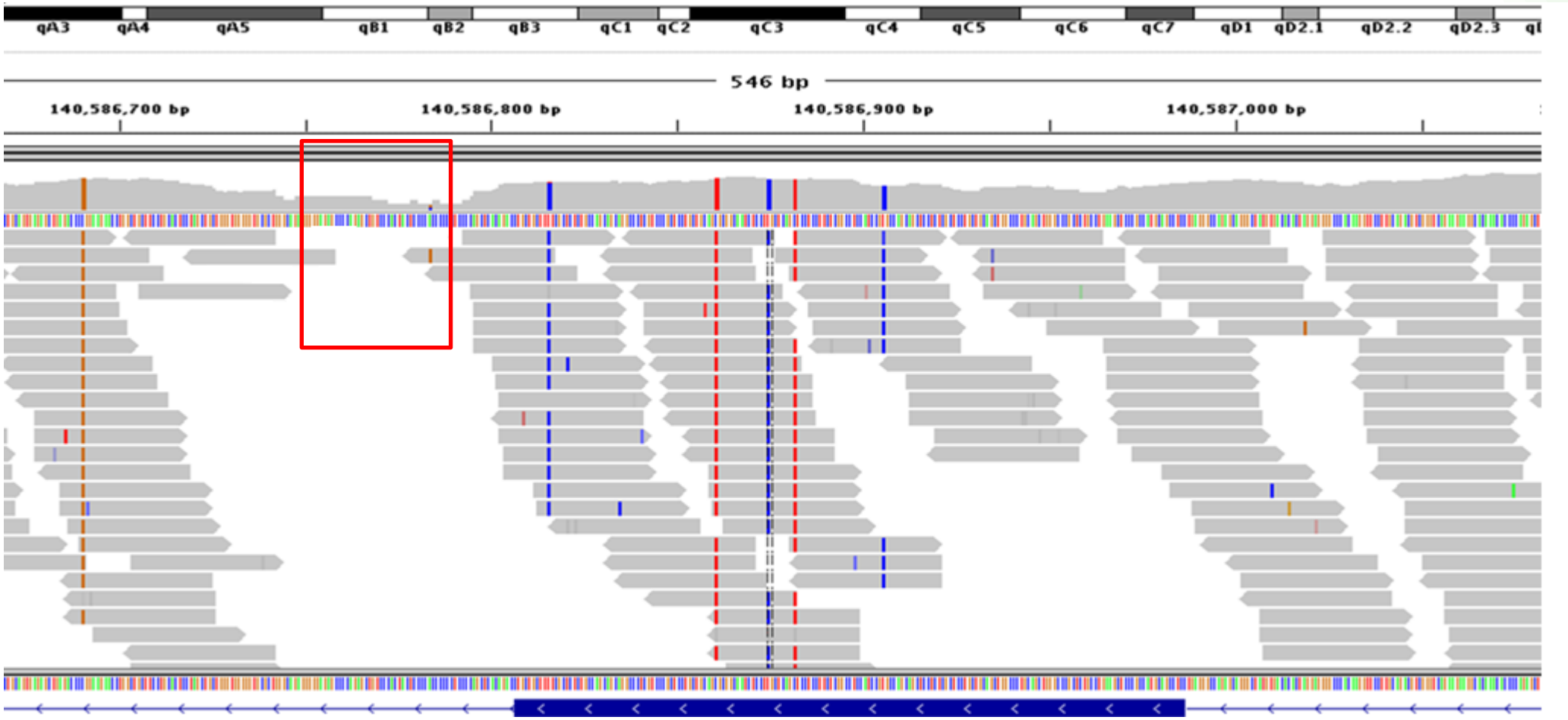
Genotype accuracy

Variants/Pipeline	Fast-WGS	Concordance (%)	SOAPsnp	Concordance (%)
Shared genotypes*	674,139		645,070	
Homozygous	668,672	99.7	641,215	97.1
Heterozygous	3,842	98.6	2,152	91.8
Indels	1,625	96.1	1,703	89.5

*Shared genotypes with SoySNP50K dataset

- ✓ Fast-WGS calls highly accurate variants
- ✓ Overall high level of accuracy of dataset : ~99.5%

The problem of missing data!



Missing data imputation

9% missing data

Missing data imputation

A	T	T	G	C	A	T
A	T	T	G	C	G	T
A	T	T	N	C	G	T
A	C	C	A	C	G	A

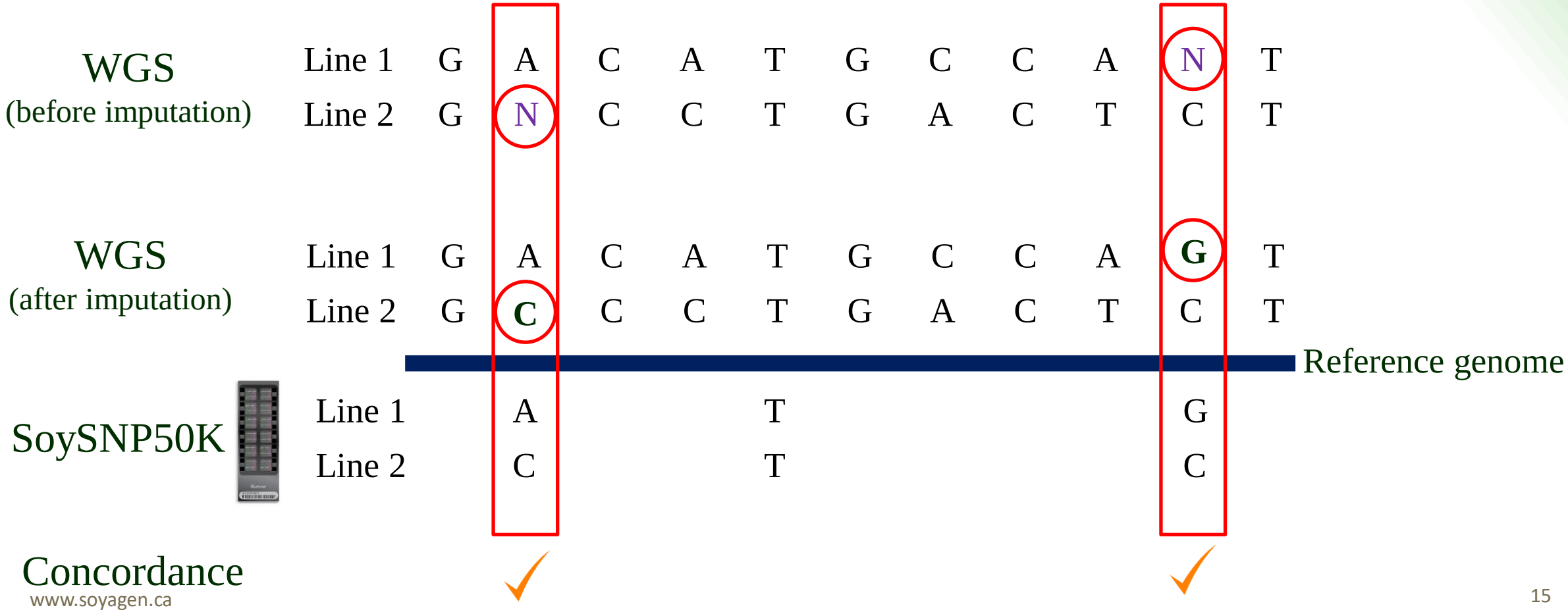
A	T	T	G	C	A	T
A	T	T	G	C	G	T
A	T	T	G	C	G	T
A	C	C	A	C	G	A

Is this accurate?

Torkamaneh and Belzile 2015

Missing data imputation accuracy

Are the imputed genotypes (initially missing data) in agreement with the calls made using the SoySNP50K chip at the same nucleotide position in the same accession?

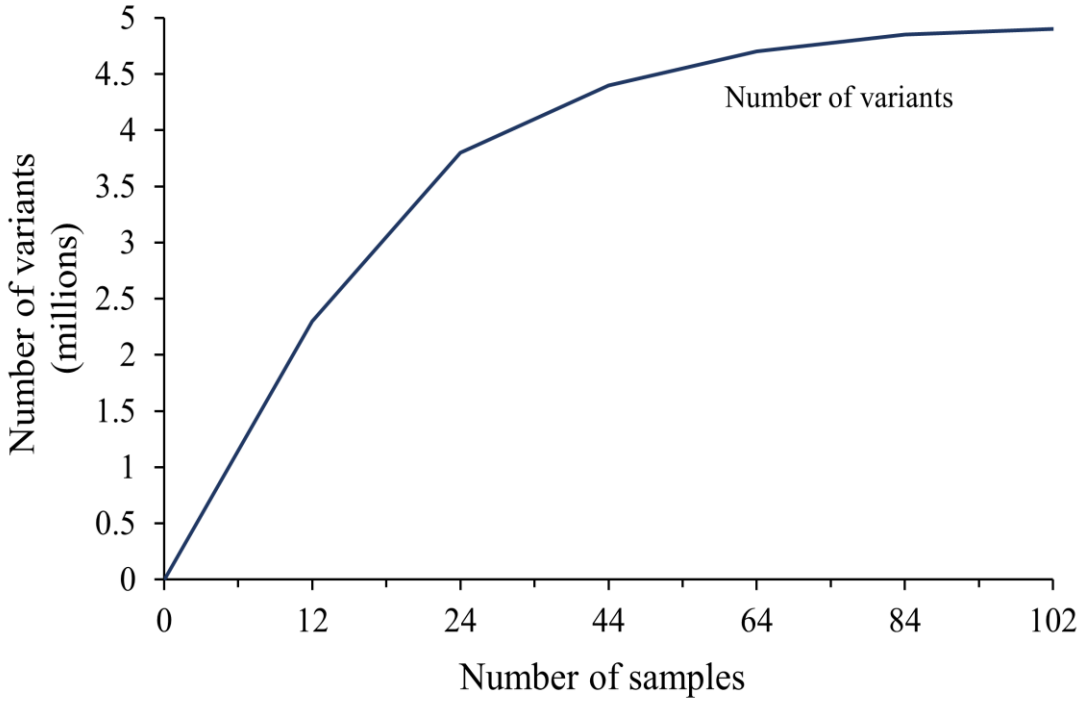


Missing data imputation accuracy

Variants	WGS dataset	Imputation accuracy (%)
Number of homozygous genotypes	594	98.8
Number of heterozygous genotypes	41	92.7
Total	635	98.6

✓High level of imputation accuracy

Is WGS of 102 accessions enough?



- To determine the level of saturation among Canadian soybean, subsets of samples of increasing size were randomly selected and analyzed (N=12, 24, 44, 64, 84, and 102)
- The number of variants discovered did not increase much beyond 80 accessions

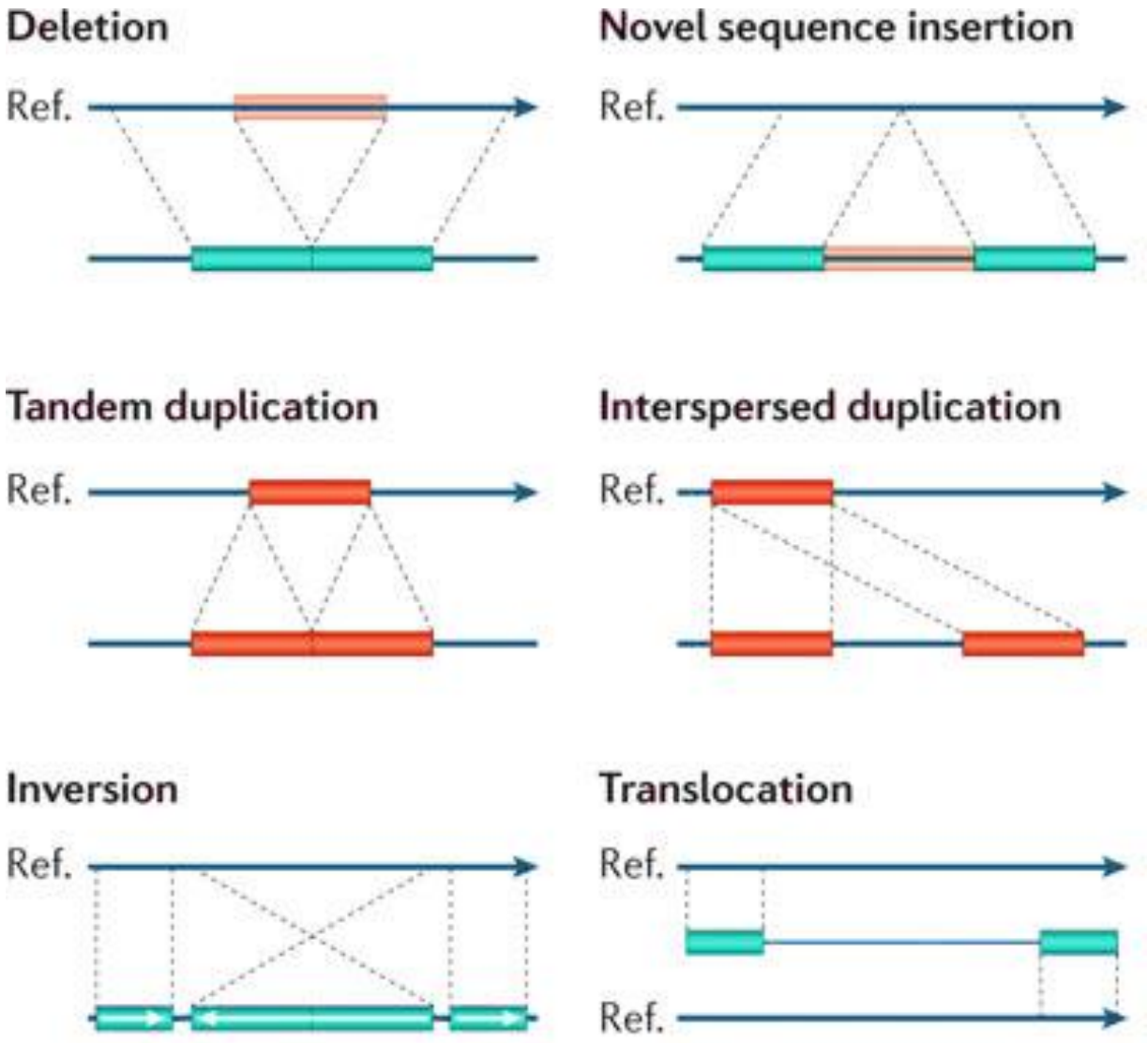
✓SNP catalogue is highly extensive

Conclusions for nucleotide variation

- ✓ Complete genome sequencing of 102 Canadian short-season soybean
- ✓ ~5 M nucleotide variants
- ✓ High level of accuracy
- ✓ Extensive capture of SNP and haplotype diversity

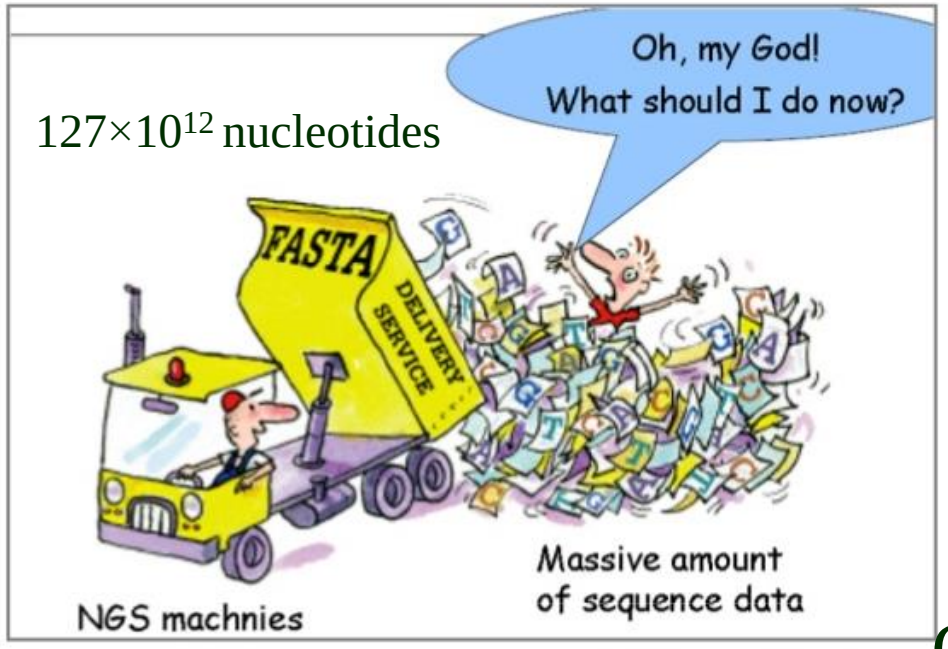
Structural variation

Exploration and characterization of all types of SVs

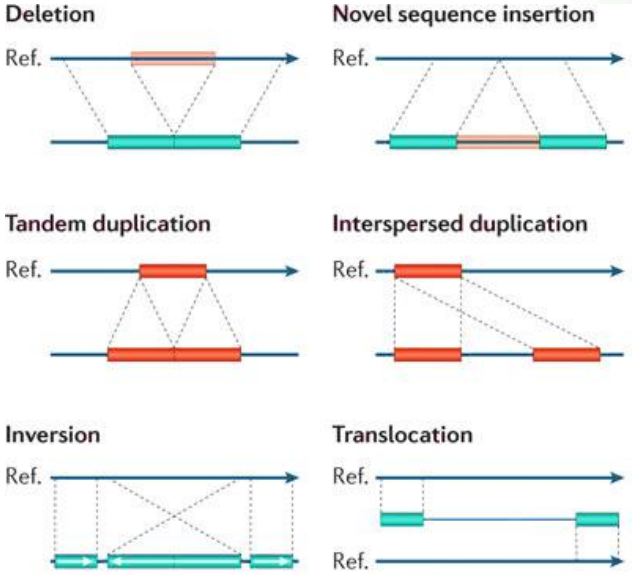


Structural variation

Exploration and characterization of all types of SVs



Bioinformatics tools



Combination of several tools:

BreakDancer, CNVnator and **LUMPY.**

Types of SVs and their characteristics

SV type	Number of SV sites	SV size	Median size of SV (bp)	SV site breakpoint precision (bp)
Deletion	63,556	10bp-3Mb	118	±3*
Insertion	16,442	32bp-3Mb	149	±4*
Duplication (disperse duplication)	2,865	66bp-3Mb	2,546	±15†
Inversion	3,965	33bp-2.8Mb	138	±12‡
CNV (tandem duplication)	1,435	500bp-1.5Mb	5,714	-
Translocation (intrachromosomal)	3,011	30bp-2Mb	124	±6
Translocation (interchromosomal)	78	11kp-3Mb	245,312	±35

- ✓ Complete collection of SVs in soybean (all classes)
- ✓ Deletions and insertions constitute largest class of SVs (~86%)
- ✓ Size of SVs ranges from 10bp to 3Mb
- ✓ ~90% of SVs are small (50-300 bp)

SVs accuracy estimation

- Lack of external dataset for comparison (such as CGH array)

- **Alternative approaches:**

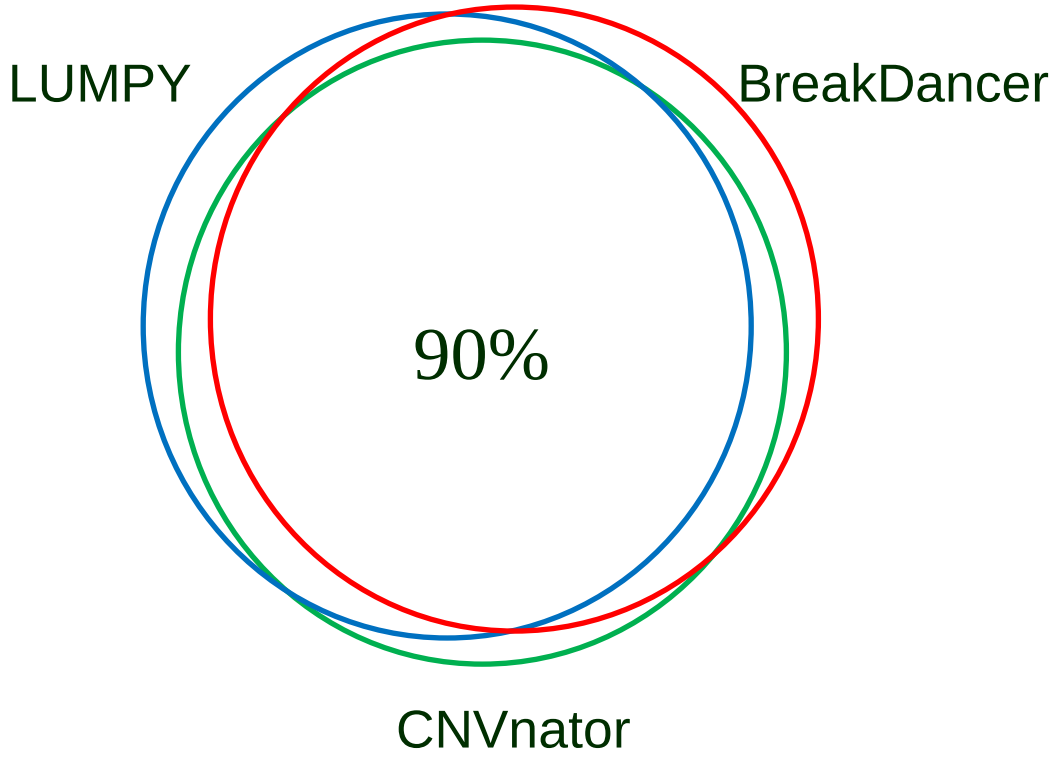
  I. Overlap among different tools

  II. Validation using SNP dataset

  III. Experimental validation (PCR and sequencing)

How accurate are these SVs?

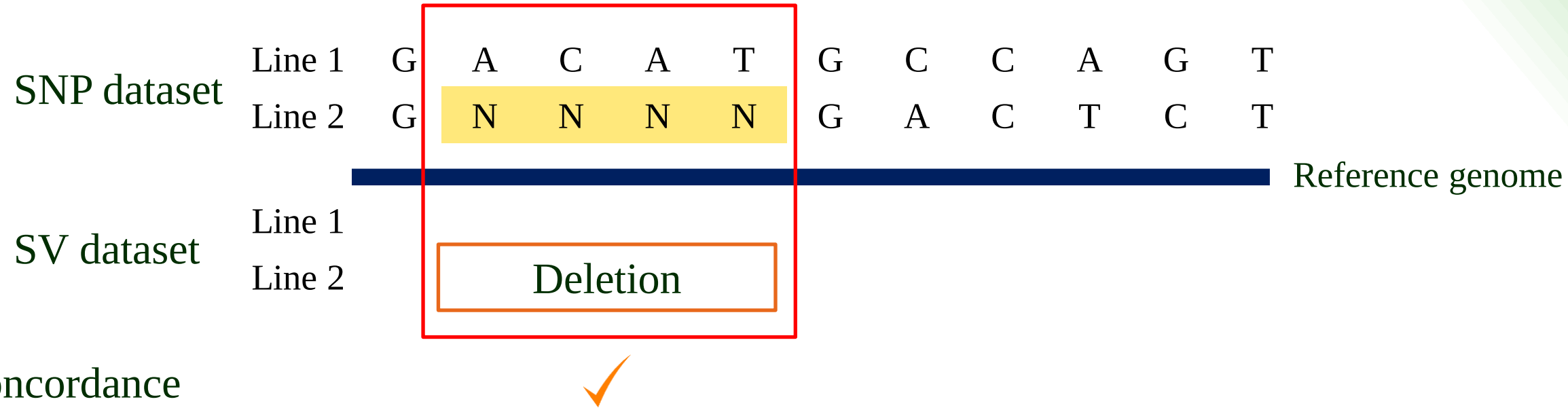
I. Overlap among different tools



✓ ~90% of concordance for all types of SVs

How accurate are these SVs?

II. Validation using the WGS SNP dataset



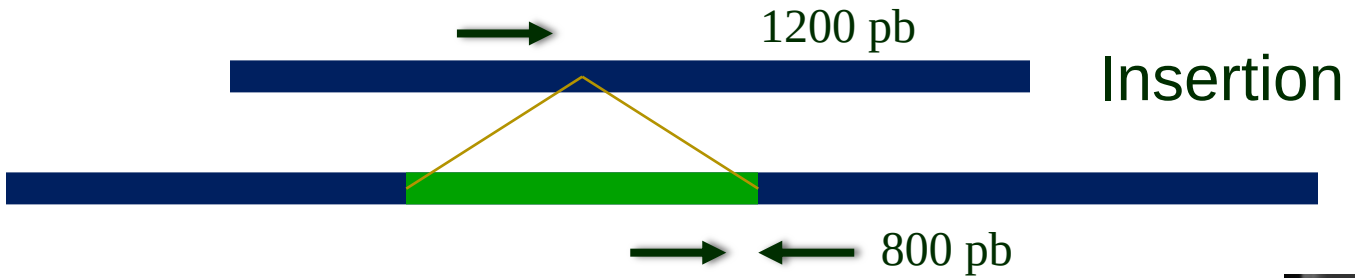
Concordance

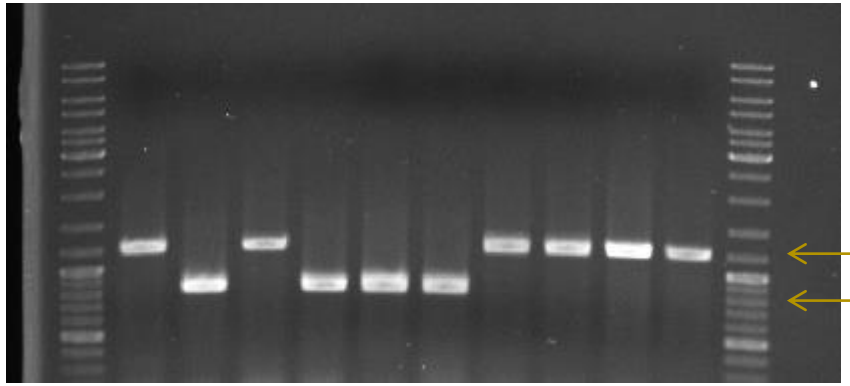
✓ High level of concordance for called deletions: ~94%

How accurate are these SVs?

III. Experimental validation (PCR and sequencing)

- 40 SVs representing different types of SV of different sizes





- ✓ 80% (32/40) of concordance between WGS data and PCR results
- ✓ High level of validation using PCR

Conclusions for structural variation

- ✓ Extensive catalogue of SVs in soybean (all classes)
- ✓ ~92K structural variants
- ✓ High level of accuracy
- ✓ Extensive capture of SV diversity

Overall conclusions

We sequenced 102 elite soybean lines from Canada, resulting in the most extensive capture of genetic diversity among cultivated accessions from a single country to date.

This collection is very important for several reasons:

i) Representative of short-season soybean germplasm;
ii) Extensive capture of SNP and haplotype diversity
iii) First complete collection of structural variants
iv) Highly accurate collection of nucleotide and structural variants