A new approach to genotyping: Single Primer Enrichment Technology (SPET), an integrated system for both targeted and *de novo* genotyping

Fornasiero Alice[1], Pinosio Sara[1], Don Heath Joe[2], Dell'Acqua Matteo[3], Pè Mario Enrico[3],
Cattonaro Federica[4], Morgante Michele[1], Scaglione Davide[4]

[1] Istituto di Genomica Applicata – Udine (Italy); [2] NuGEN Technologies Inc. – San Carlos (CA - USA);
[3] Institute of Life Sciences, Scuola Superiore Sant'Anna – Pisa (Italy); [4] IGA Technology Services – Udine (Italy)

e-mail:
afornasiero@appliedgenomics.org

Introduction

DNA array-based technology is a powerful technique for the detection of targeted SNPs, but it is limited in throughput scalability and prone to ascertainment bias.

Genotyping-by-sequencing (GBS) is a robust approach for enabling large scale, whole-genome studies of genetic variation by the random genotyping of a reduced fraction of the genome.

Single Primer Enrichment Technology (SPET) is an innovative approach that integrates both the **targeting of known polymorphisms**, allowing to perform targeted genotyping, and a ***de novo* genotyping**, allowing random SNP discovery. The Allegro Targeted Genotyping (NuGEN Inc.) for SPET analysis relies on a panel of probes targeting selected SNPs and leverages sequencing information for novel allele discovery.

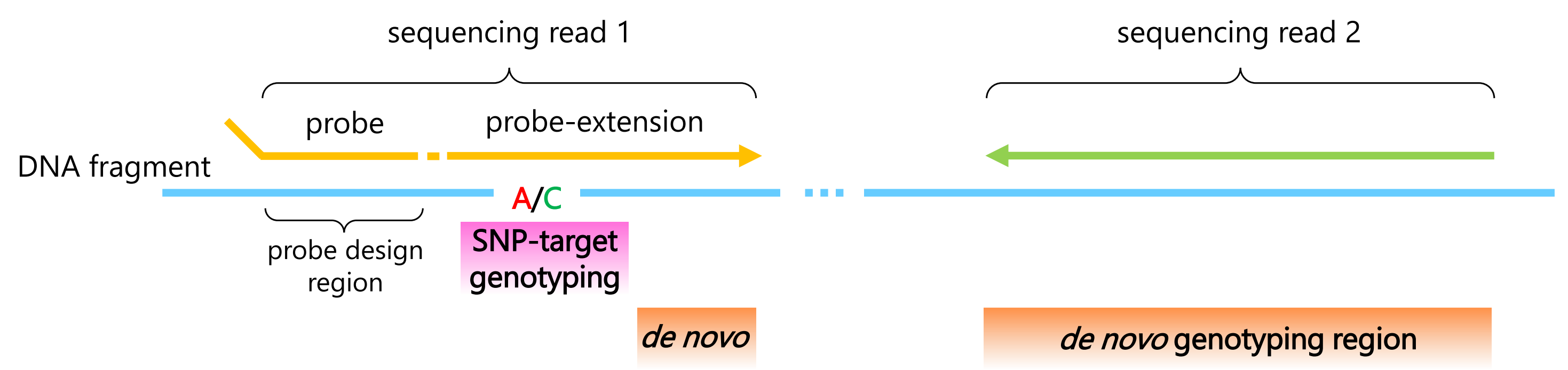
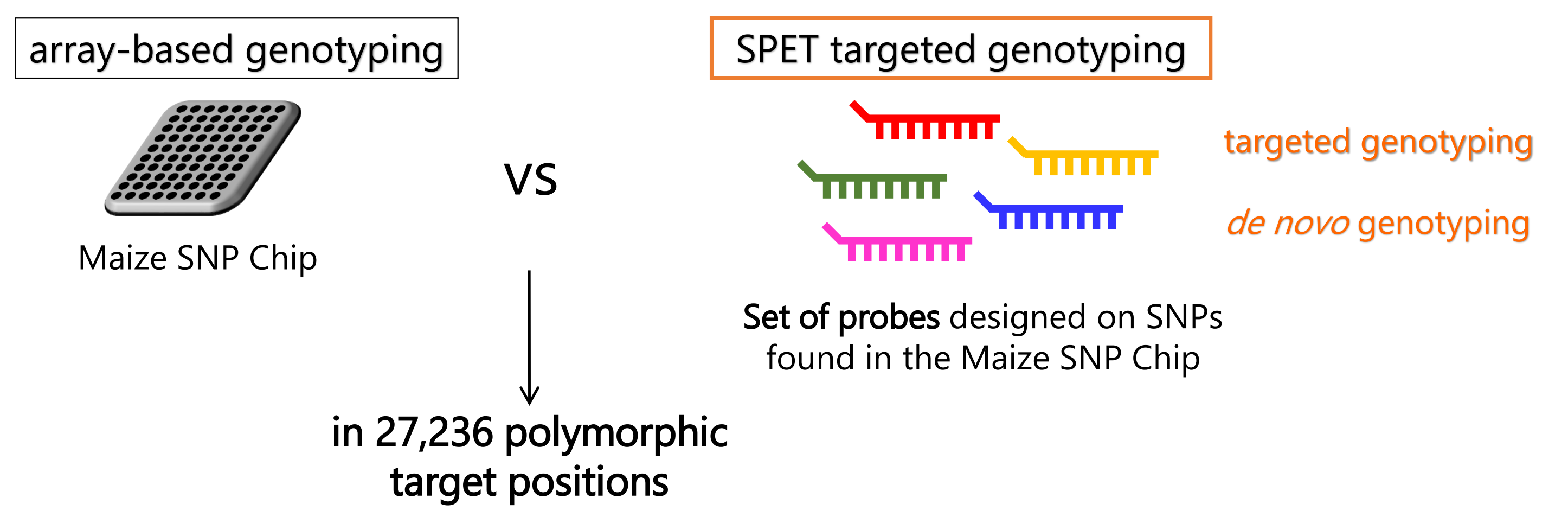


Figure 1. **The Allegro Targeted Genotyping** (NuGEN Inc.) relies on probe hybridization and extension to target a panel of selected SNPs and allows SNP discovery by exploiting read information from single-read sequencing or from paired-end sequencing.

Aim of the study

Two replicates of five *Zea mays* **inbred lines** (F7, H99, HP301, Mo17 and W153R) and five **F1-crosses** (A632 x B73, B73 x B96, B73 x F7, B73 x Mo17 and W153 x HP301) were assayed using the targeted genotyping method.

The aim was to measure the **performance of the SPET technology** comparing it to the array-based technology, and to validate the method as a genotyping solution suitable for both targeted and *de novo* genotyping.



Set of probes designed on SNPs found in the Maize SNP Chip

in 27,236 polymorphic target positions

Figure 2. **SPET evaluation analysis in maize.** The accuracy of SPET genotyping was evaluated on the three inbred lines (F7, HP301 and Mo17) and on all the five F1-crosses for which genotypes called using the array technology were already available.

Results

- Samples were sequenced on either the Illumina HiSeq2500 or NextSeq500 to produce single-end 125 bp reads. Sequencing, alignment and coverage statistics are shown in Table 1.
- The target site coverage distribution (Figure 3) showed an homogeneous performance of the probe panel at a coverage of about 50x.
- The **accuracy** of the targeted genotyping was about 98.2% considering **all positions** at a coverage of 20x, and reached a plateau at about 50x. Positions targeted by **two probes** had a higher percentage of accordance (98.5% at 20x coverage) with SNP Chip (Figure 4A). The accuracy at **homozygous sites** was higher with respect to heterozygous ones (Figure 4B).
- The **reproducibility** of the method, obtained comparing the genotype calls for the two replicates, ranged from 97.87% to 99.73% (Figure 5). Increasing the minimum required coverage led to an increase of reproducibility, with a plateau at about 50x.
- Our results **validated** 98.2% of polymorphic sites compared with the SNP Chip, with 381,575 true calls over 388,551 total calls at a coverage of 20x.
- The ***de novo* genotyping** allowed the discovery of additional polymorphic sites in the regions targeted by the probes. The frequency of the alternative allele in the discovered polymorphic positions was lower than the one in the target site positions (Figure 6.)

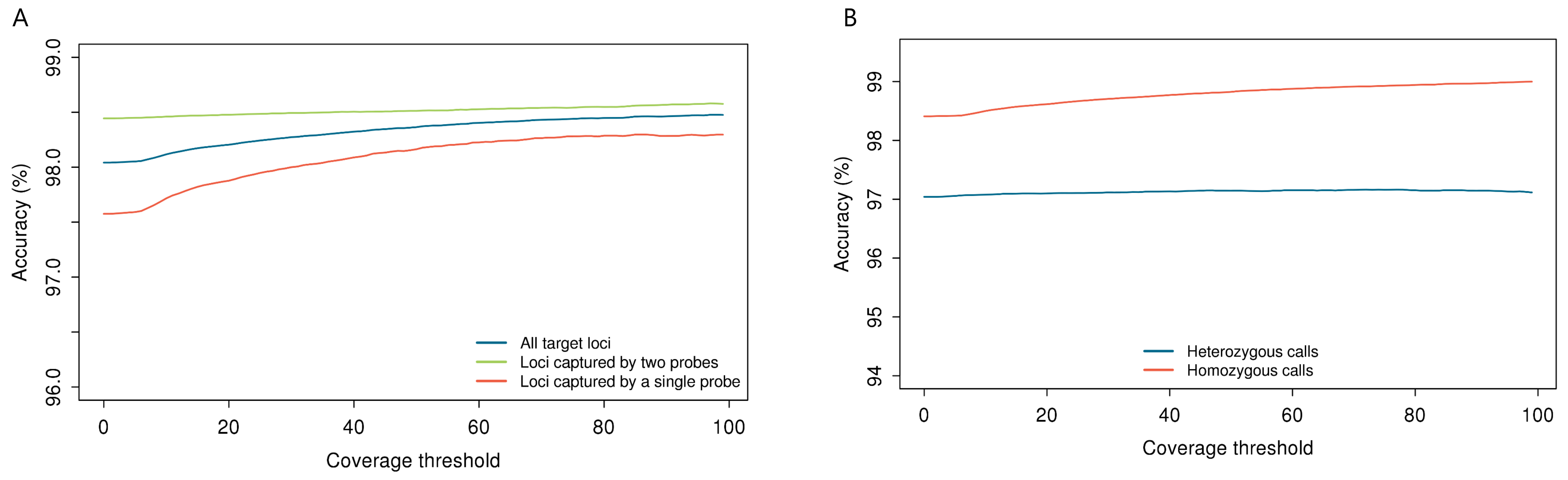


Figure 4. **Accuracy of the targeted genotyping method**. (A) Percentage of genotype calls in accordance between SNP Chip and SPET at increasing coverage thresholds, considering one probe, two probes and all positions; (B) percentage of accuracy considering separately homozygous and heterozygous calls at increasing coverage thresholds.

Sequencing and Alignment Statistics	
Raw read number	25,701,144
Aligned read number	22,975,658
Mean Coverage	90

Table 1. **Sequencing and alignment statistics.** The table reports the mean total number of sequenced reads, the mean number of high quality aligned reads and the mean coverage obtained on the target regions across samples.

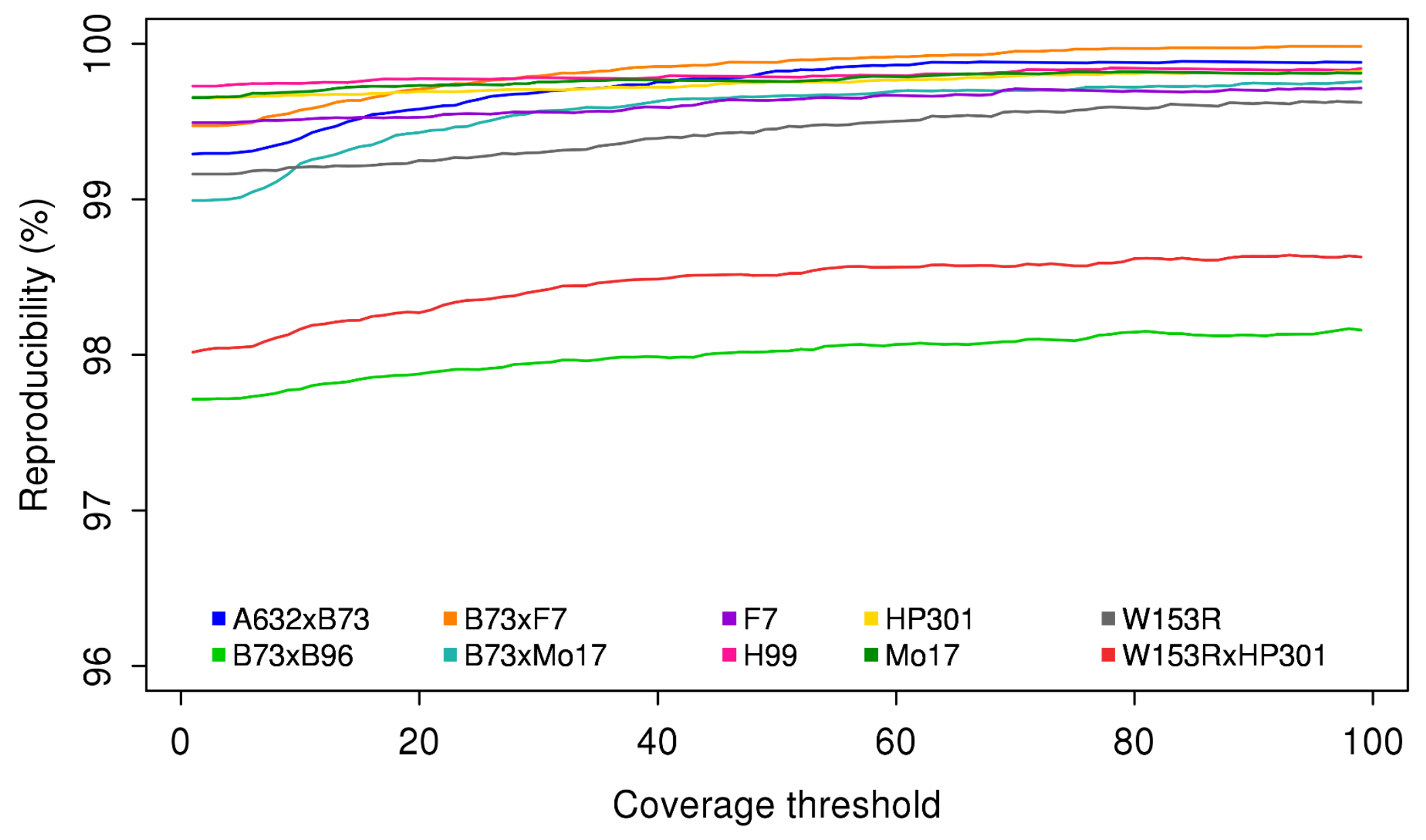


Figure 5. **Reproducibility of the targeted genotyping method.** Reproducibility was measured by comparing the genotype calls obtained for the two replicates of each sample, at increasing coverage.

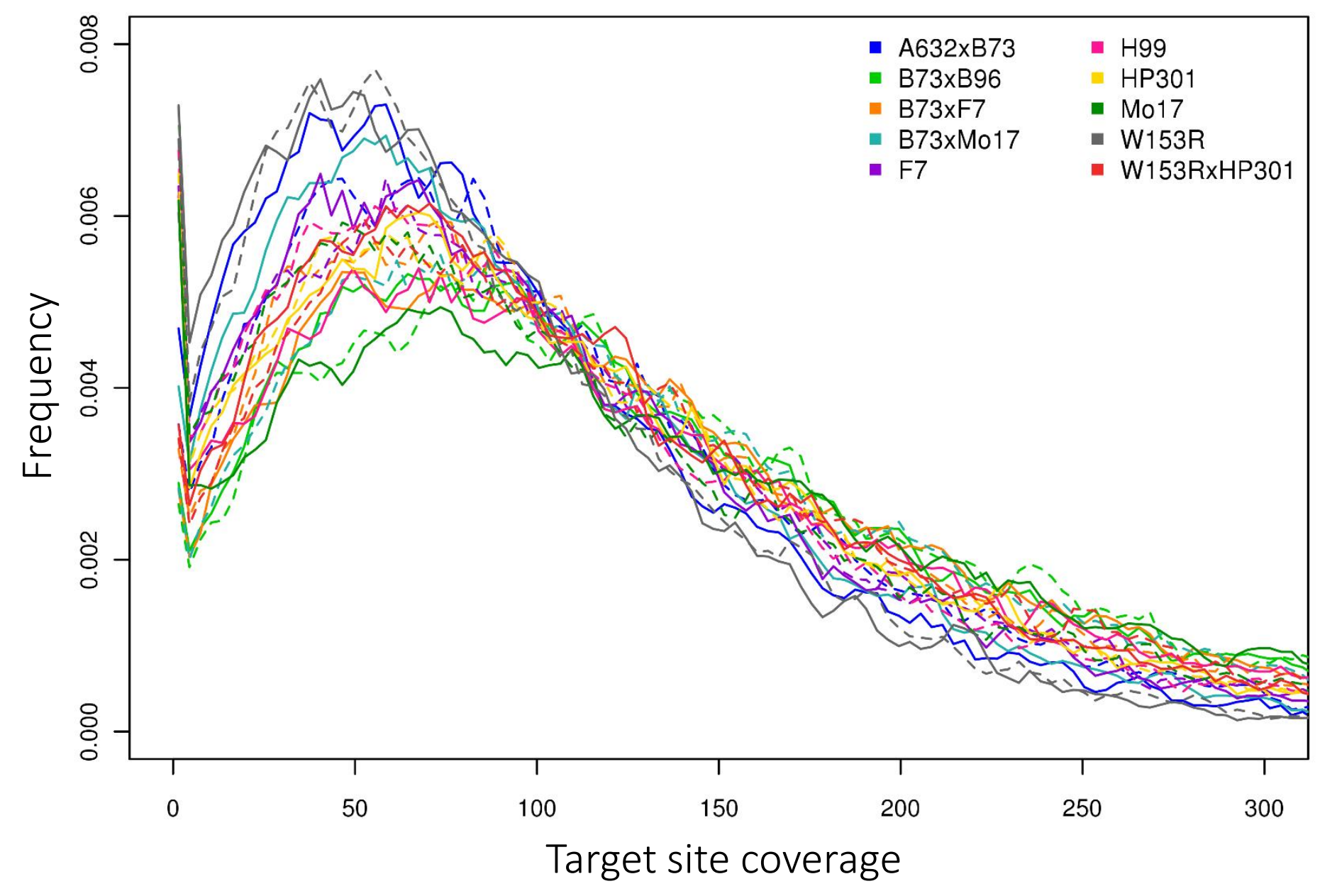


Figure 3. **Target site coverage distribution**. Distribution of the coverage obtained at each target site. For each sample, replicates are represented by dashed and solid lines.

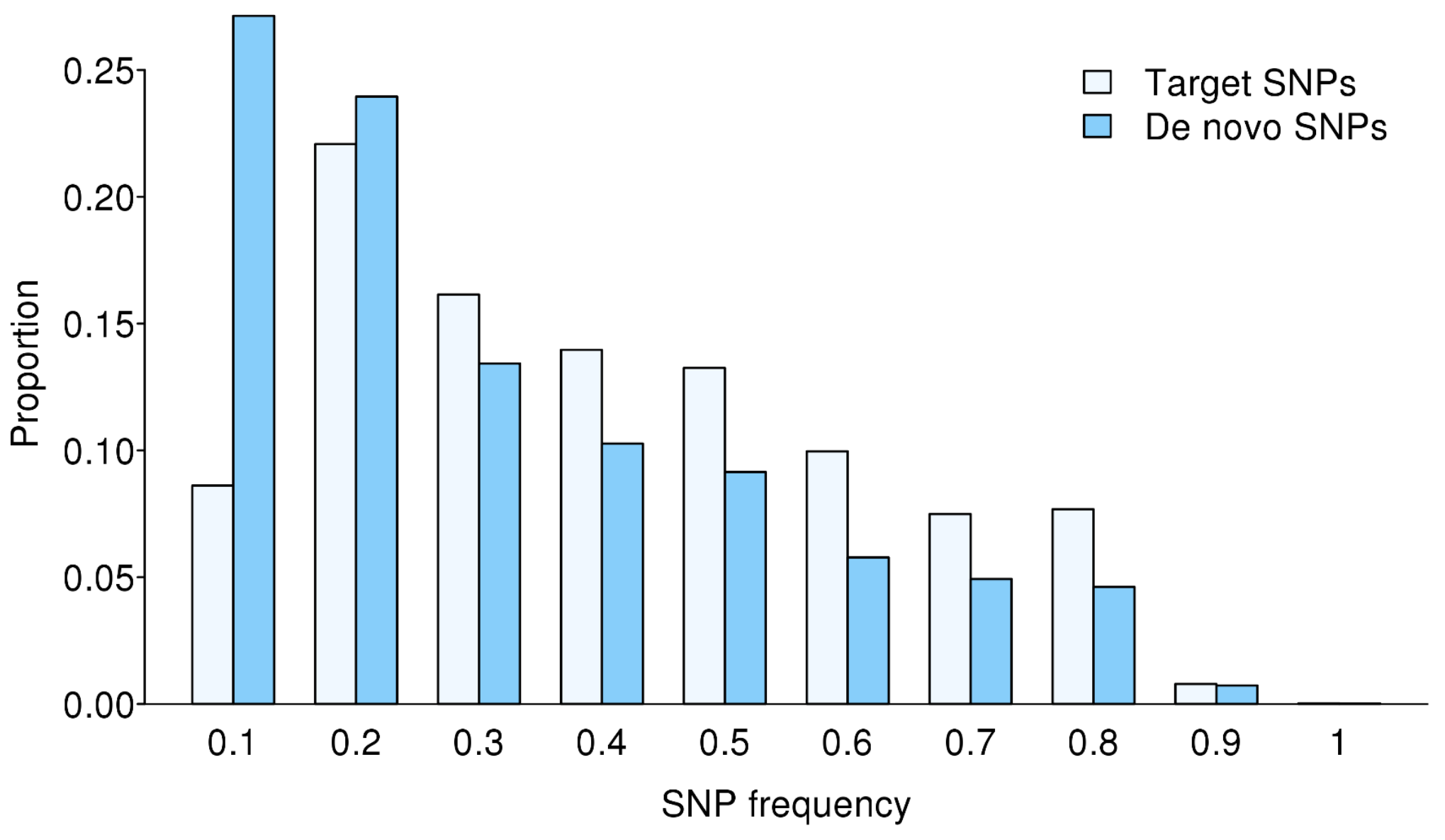


Figure 6. **Alternative allele frequency distribution.** The histogram shows the frequency of the alternative alleles in the *de novo* discovered polymorphic sites, compared to the target sites.

Conclusions

The analysis of the whole region captured by the probes designed to target the 27,236 sites allowed the detection of **49,443 additional polymorphic sites**, present in at least one of the analysed sample, **thanks to the *de novo* genotyping approach**.

Our experiment showed that SPET technology is a **powerful tool for high-throughput, cost-effective genotyping**. Furthermore, the possibility to provide additional data thanks to the *de novo* SNP genotyping allows to **circumvent ascertainment bias**, which is one limitation of the array technology.

The results showed that the SPET approach is a promising solution for a wide-range spectrum of analysis, from **fine mapping** to **GWAS**, thanks to its combination of targeted genotyping and *de novo* SNP discovery.

Methods

Cutadapt (Martin, 2011) and ERNE filter (erne.sourceforge.net) were used to remove adaptor sequences and low quality 3' ends from short reads, respectively.

Reads were aligned to the *Zea mays* v4 reference genome using the short read aligner BWA-MEM (Li et Durbin, 2009).

SNP calling was performed on uniquely aligned reads using the GATK software (McKenna et al., 2010).

References

Dell'Acqua, M. et al. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biology*. 16(1), 167.
Li, H. et Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 25(14), 1754-60.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*.
McKenna A. et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297-303.

Acknowledgements

Novel variation in plant breeding and the plant pan-genomes (NOVABREED). Grant agreement n. 294780 - European Research Council.