SeqSNP tGBS as alternative for array genotyping in routine breeding programs

Executive summary

Refinement of targeted genotyping by sequencing (tGBS) technology has led to the development of the LGC Genomics SeqSNP genotyping platform. The development of SeqSNP provides a cost-efficient, flexible and scalable mid-plex genotyping platform as a service or as bespoke kits for in-house targeted sequence based genotyping.

SeqSNP allows for the assessment of complex traits in all modern breeding programs. It provides an alternative to fixed arrays and is ideally placed for the application of genomic selection (GS). This particular study highlighted an application in a plant breeding program and clearly shows that SeqSNP results not only correlate with existing array genotyping platforms, but also surpasses other sequence based genotyping options in *de novo* SNP discovery and the analysis of multi-allelic target SNP sequences.

SeqSNP service 'all inclusive' options include:

- Plant sampling kits.
- DNA extraction.
- Probe library design.
- Probe library synthesis.
- Sequencing.
- Data analysis

The sequence coverage obtained in SeqSNP combined with imputation can truly be applied for a range of crossing strategies (bi-parental crosses, landrace trait introgression, and hybrid production). The incorporation of SeqSNP in the analysis of training populations will contribute to more accurate genomic estimated breeding values (GEBVs).

The success of GS depends on the ability of genotypic data to capture genetic variation among the training populations and prediction individuals at low cost (10). Highly heterotic species or segregating populations in crossing programs such as potato and maize will benefit from probe design (as in SeqSNP) that has flexibility for target sequences, avoiding variability in DNA due to heterosis. The capability of SeqSNP to become an established tool in the breeder's toolbox by turnaround times for data generation fitting into breeding cycles, is expected to have an impact on the development of novel varieties for all species.

The contribution of the selection of molecular markers, together with other farm management systems, to breeding strategies from first implementation in 1974 (11) has catapulted global agricultural productivity. The ability for improvements in complex traits such as yield, drought tolerance and nitrogen usage efficiency can only truly be assessed by the application of high density markers. Accessibility to applications such as tGBS must be possible to all breeders to enable the deficits in agricultural production to be overcome. With the expected global population to reach 9 billion by 2050 agricultural sustainability is in question. The flexibility, scalability and cost efficiency of technologies such as SeqSNP will provide a solution in part by the ability of the breeding community to have access to cutting edge technologies through service options and kits.

Introduction

A major gap in mid-plex genotyping exists in the process flow for all molecular breeding programs and has recently been described as the dead space for genotyping (1). This has been addressed by the development of SeqSNP service by LGC Genomics; a targeted genotyping by sequencing method. SeqSNP not only provides flexibility in single nucleotide polymorphism (SNP) sequence selection, but also scalability in sample numbers which can be restrictive on other genotyping platforms. Independently analysed data is presented here, substantiating that the SeqSNP service delivers genotyping data with high concordance to array genotyping and provides a more flexible and scalable alternative. SeqSNP is the next stage in genotype-based sequencing for all breeding communities.

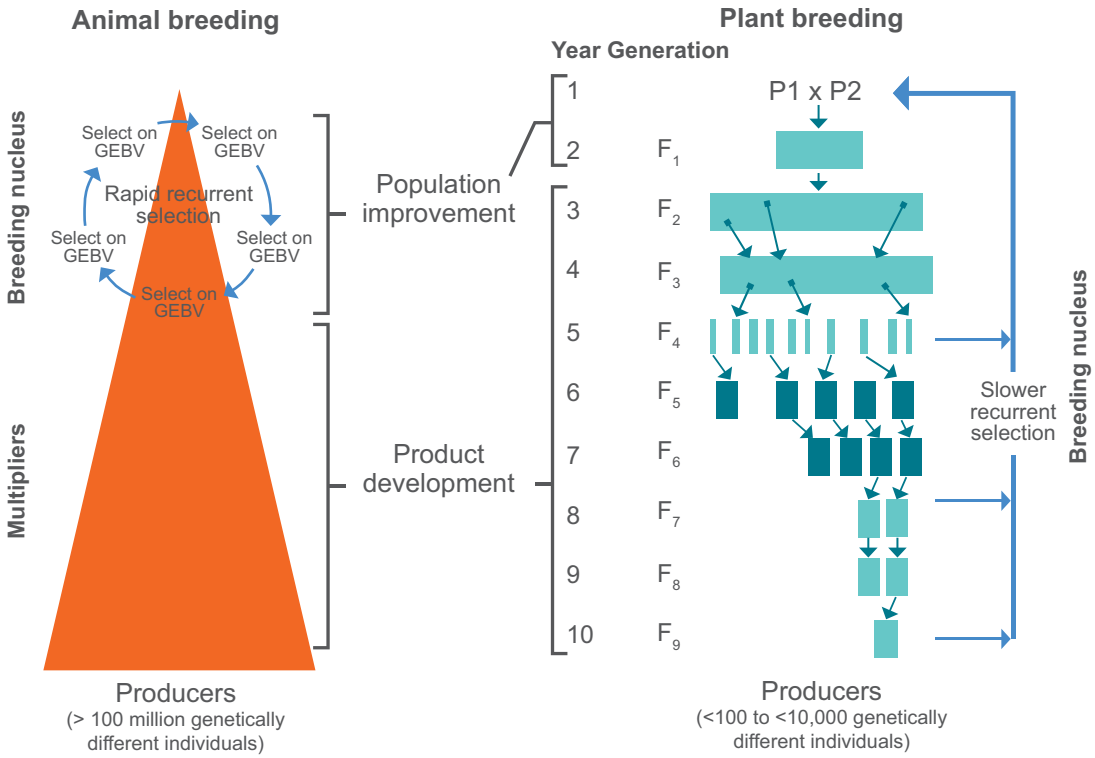

Fig 1: Comparison of plant and animal breeding strategies. Courtesy of Hickey et al (2017)

Background

High-resolution sequence-based genotyping has been extensively applied to livestock breeding programs as a means to accurately predict and select livestock where investment needs to deliver maximum genetic gains. This investment has improved the productivity of livestock breeding programs for complex traits, such as milk, leading to an annual genetic gain in milk yield for genotyped cows (with a record and born from 2009 through 2012) by 24 kg (2). The use of genomic selection in early measured traits for sheep and beef cattle produced 20-40% more genetic gain when combined with reproductive technologies (3).

The unification in breeding strategies through applications such as GS in animal and plant breeding has converged (Fig 1). Sequencing cost efficiencies are now within budgeting constraints in plant breeding programs, which are restrictive when compared to livestock programs. Yet the challenges in agricultural sustainability in the future are no less inconsequential.

The available technologies to perform genome wide association studies (GWAS) such as arrays have until now required expensive investment in terms of setup costs, which has required the amalgamation of groups into consortia. This would allow inclusion of limited germplasm of individual breeding programs and long term commitment of sample number to make the service cost efficient. This has led to a redundancy of data generated, which is a false economy in the investment into array based genotyping platforms. This is compounded with issues such as rigidity in sequence selection, lack of scalability and a lack of flexibility. For realistic and pragmatic applications of GEBVs, incorporating SNP marker information with cost efficient targeted sequencing options are needed.

SeqSNP provides a new alternative approach to arrays. It is a complete pipeline solution for development of varieties with improved genetic gains that is needed for all agricultural sustainability challenges faced as a global community and which need to be addressed immediately.

Target enriched sequencing

In most crop genomes, the exome corresponds to only 1-2% of the entire genome; therefore, the targeting and sequencing of only these regions significantly reduces sequencing and computing costs (4). Costs and assembly difficulties associated with whole genome sequencing (WGS) make approaches like target enriched sequencing feasible and appropriate. It enables the development of probe libraries not only for GWAS, but can encompass exome capture data to enable screening of populations intended for GS strategies. Consequently, considerable effort has been devoted to develop ‘target-enrichment’ methods, in which genomic regions are selectively captured from a DNA sample before sequencing. Resequencing the genomic regions that are retained is more time and cost-effective, and the resulting data are considerably less cumbersome to analyse (5).

There is a number of commercially available targeted enriched genotyping by sequencing technologies

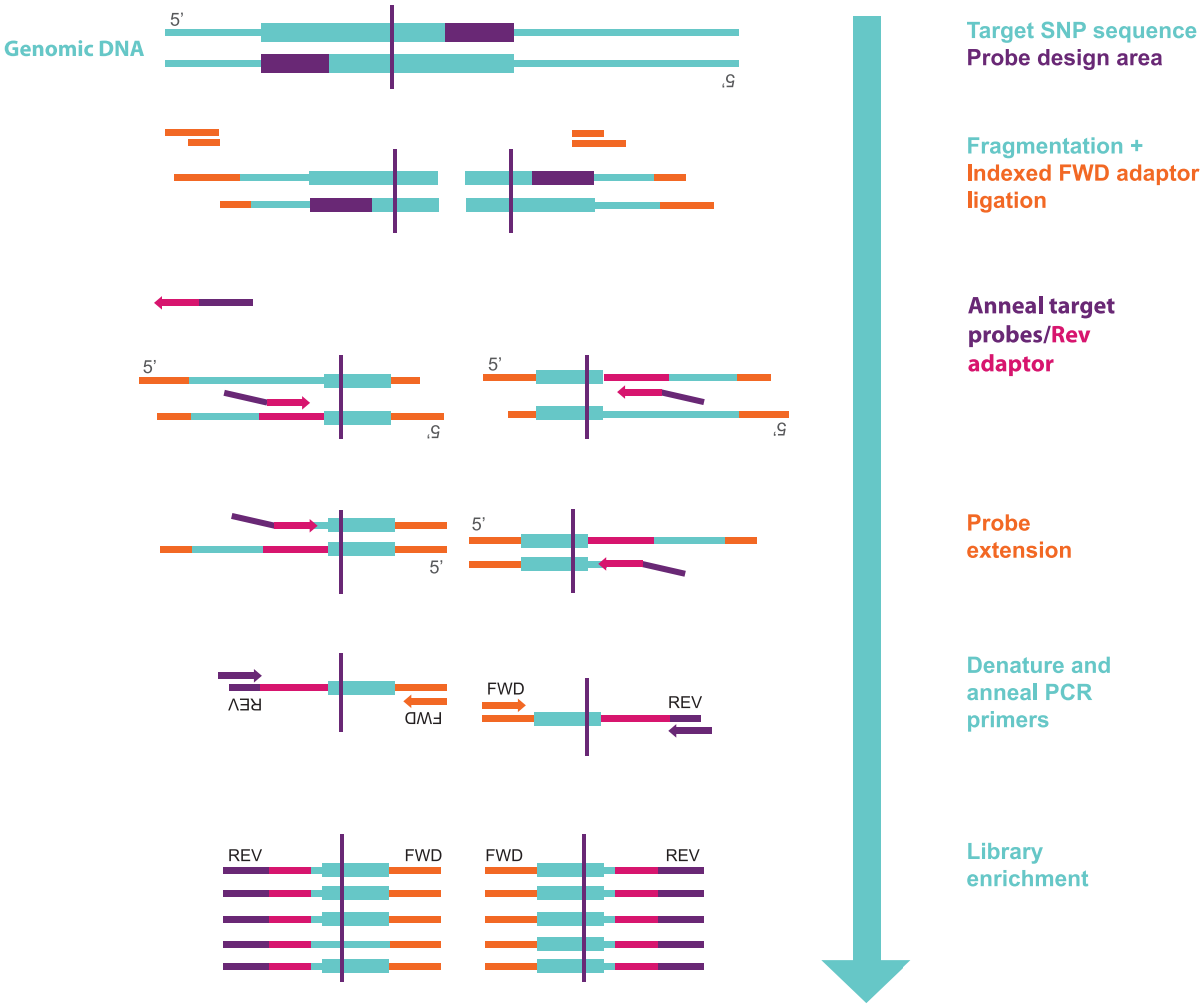

Fig 2: Single primer enrichment technology (SPET) methodology for the production of probe libraries for tGBS

1) Biotinylated capture: Requires *de novo* protocol set up, limited to only specific probe libraries and no service option available.
2) Solid state capture: Array-based genotyping lacks flexibility as the markers on a designed array are fixed. Arrays are also subject to an ascertainment bias related to the number of samples and criteria used in SNP detection (6). In addition, if additional SNPs are later required the array must be redesigned, a process that can be expensive (7).
3) Alternative in-solution targeted genotyping by sequencing: Limited to multiplex to 2000 markers and direct SNP discrimination.
4) Amplicon based sequencing: Discriminates targeted SNPs, reduces flexibility of probe design and prevents accommodation of minor allele frequency (MAF) in flanking SNP sequence.

In comparison to array-based sequencing, probe-based SeqSNP enables the flexibility of developing core panels to which future marker discoveries can be added, without the need of re-design or re-synthesis of probe libraries.

Probe library design for SeqSNP target regions surrounding a targeted SNP sequence, thus enabling probe design flexibility which can accommodate surrounding MAF, providing higher conversion rates and higher accuracy in data generated (Fig 2). Another advantage of SeqSNP's flexibility in probe design is the identification of *de novo* SNP identification in the surrounding region of the target SNP marker. The study identifies this unique ability with tGBS as part of the SeqSNP service. In addition, the utilisation of two oligonucleotide probes for the sequencing of targets contributes to the cost efficiency of the service offering.

SeqSNP has answered the need for capacity to produce bespoke probe libraries that are applicable for diverse long term breeding objectives for individual breeders. The main advantages of SeqSNP over existing tGBS technologies include:

- Lower set up costs, as costs are dependent on SNP and sample number, with no solid array set-up.
- Flexible marker selection: Up to 100K SNPs per sample in a single run.

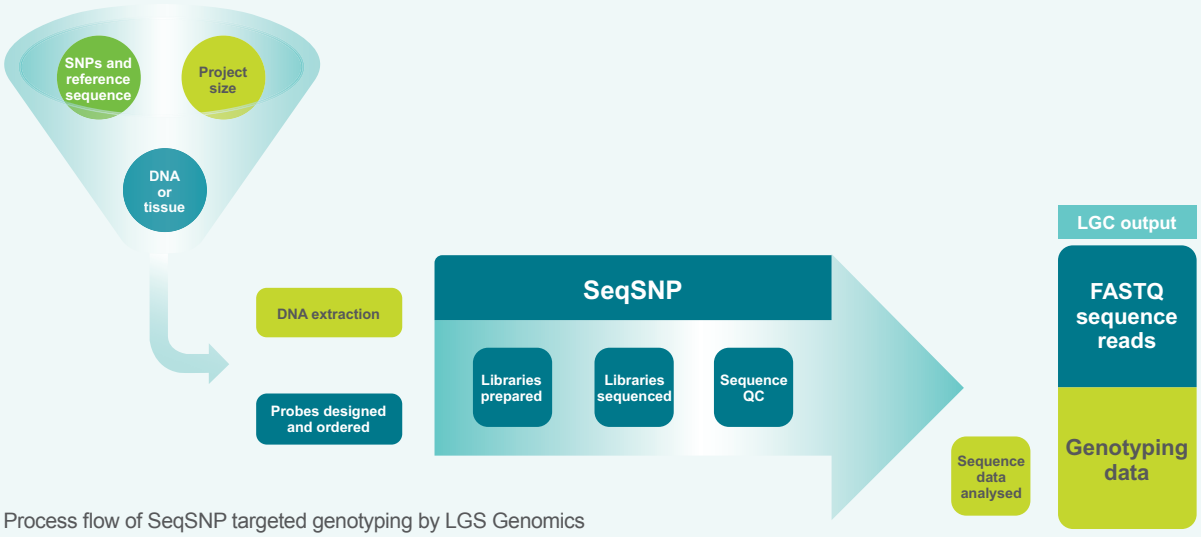

Fig 3: Process flow of SeqSNP targeted genotyping by LGS Genomics sequencing service. Turnaround times with prior knowledge of reference sequence and SNP information can be 8- 12 weeks to enable selection of lines for continuation within a breeding cycle.

- *De novo* variants (including structural variants) detected in target SNP region.
- Cost effective: Highly efficient enrichment methods reduce day-to-day operation costs.
- Shorter turnaround times for probe library production when compared to manufacture of arrays.
- Accessibility to sequencers and high throughput extraction instrumentation, without capital investment.
- Fragmentation of the gDNA replaces mechanical shearing: Simultaneous digestion and labelling of DNA fragments simplify the workflow.
- Single primer target enrichment technology (SPET) enables highly flexible and scalable custom panel design (Fig 2).
- Dual-index sample barcoding enables multiplex sequencing of over 3000 samples in a single sequencing lane, which allows for further scalability without limitations.

The importance of establishing a pipeline for the breeding community for the incorporation of tGBS into breeding program timelines cannot be underestimated. One of the major limitations of arrays for genotyping is the length of time for array production. It can take from 3 to 6 months solely to design and manufacture a fixed set of markers. Compounded by the lack of scalability in array technology, SeqSNP provides a solution for increased sample numbers associated with applications such as GS. With planning, SeqSNPs' process flow enables the design and manufacture of probe libraries, DNA extraction, sequencing and data analysis to fit into plant breeding cycles (Fig 3).

Plant sampling and DNA extraction

The current requirement of high molecular weight and high quality DNA for all plant-based sequencing for genotyping begins with either leaf or seed samples, and with GS, the sample numbers are expected to be in the thousands. The methods undertaken for plant sampling and extraction using standard sampling protocols include expensive and time consuming methods, including freeze-drying leaf material, the requirement of dry ice or cool boxes for collection and laborious extraction protocols. Compromising starting material or DNA quality at the commencement of any genotyping approach can lead to up to 30% loss of subsequent data (Fig 4) and using crude extraction methods can also lead to a reduction in data quality as is seen in this current study. Ultimately the ramifications of loss of data is translated to loss of data point per $ spent.

The SeqSNP service option includes an LGC plant sample collection kit, which can stabilise leaf material at source without the need of cumbersome equipment. The kit utilises a molecular desiccant which, when used correctly, will

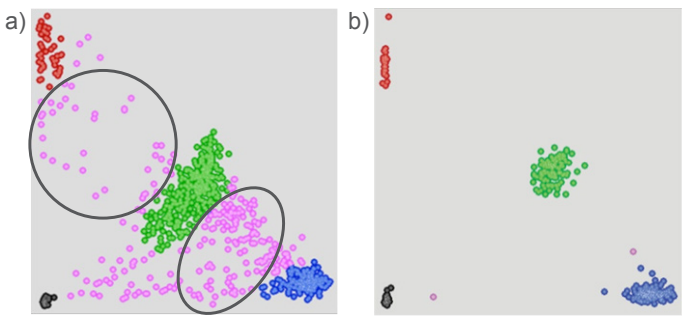

Fig 4: Impact of plant sample collection using LGC plant sample collection kits on data quality using endpoint Kompetitive Allele Specific PCR (KASP®) genotyping. The same KASP SNP marker was interrogated in both data plots. a) Leaf material sampled in tubes. Pink dots highlighted in clusterplot indicate datapoints which could not be assigned a genotype. b) Leaf samples taken in LGC plant sample collection kits.

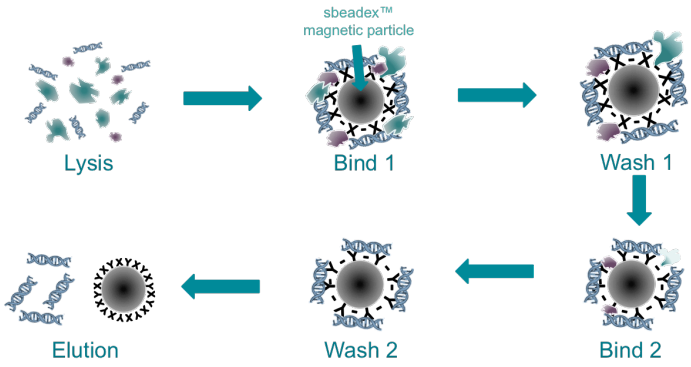

Fig 5: sbeadex extraction protocols follow a unique technology which includes a two-step binding mechanism that enables a second wash step using pure water.

maintain integrity of leaf material for any DNA extraction chemistry, leading to production of high molecular weight DNA in quantities required for tGBS.

The SeqSNP pipeline uses LGC sbeadex® proprietary extraction chemistries. A number of standard sbeadex extraction protocols exist for manual and automated extractions of plant material for a wide range of species (including difficult species such as rubber, cocoa and sunflower).

sbeadex has been developed to accommodate for the following variables, allowing for the generation of high molecular weight nucleic acids, suitable for sensitive downstream processes:

- DNA yield
- DNA concentration
- Process time
- Throughput
- Instrumentation
- Unusual properties of samples e.g. presence of secondary metabolites

sbeadex chemistries use magnetic microparticles and a novel two-step binding mechanism to bind and purify nucleic acids. Combined with the washing steps, this unique process effectively removes impurities and potential inhibitors of enzymatic reactions (Fig 5). The absence of any organic solvents in the final wash buffers prevents nucleic acid preparation from being contaminated with inhibitory remnants of these solvents, and this shortens the overall extraction time due to the elimination of unnecessary drying and heating steps. Finally, the nucleic acids are eluted and ready for use in a wide range of downstream processes

Study design

An essential output from genotype based sequencing is the ability to reconstruct the resulting data to previously identified mapped sequences which can relate to phenotypic traits exhibited in selected populations. One of the objectives in this study is to assess the optimisation of DNA quantity, validation and comparison of previously mapped array data to the tGBS method through the SeqSNP service at LGC Genomics. The results of the comparison were generated independently by a third party.

500 sugar beet markers were selected with

- 471 identical with array chip data (varying quality)
- 29 alternative markers not included on array
- 6 multi-allelic markers

192 samples

- 105 common with array chip samples
- 44 duplicates, comprised of leaf punches of various numbers (6, 4, 2 and leaf fragment) and previously extracted DNA samples
- 43 samples not previously tested

Read mapping: Bowtie2

- SNP calling: Varscan (min-cov=3)
- Use of coverage information to fill monomorphic markers (not detected with Varscan)

Leaf samples taken for the study ranged from 1 - 10 leaf punches from seedlings (3 weeks post germination). Plant material was sampled using the LGC plant sample collection kit, the plate was sealed with perforated (gas-permeable) strip caps and placed in a heavy-duty, sealed bag with desiccant to dehydrate and preserve the leaf tissue during transit to LGC Genomics in Berlin, Germany, for DNA extraction and genotyping. The study also included previously extracted DNA and leaf fragment extracts using proprietary chemistry and protocols.

In brief, total genomic DNA was isolated from 149 plant tissue samples using LGC's sbeadex DNA extraction, performed at LGC Genomics. Isolated DNA was analysed using UV spectrophotometry to estimate both the quality and quantity of the DNA. The tGBS probe design and application was carried out at LGC Genomics in Berlin and sequencing was carried out on an Illumina platform.

Results and discussion

The array data generated in these results were obtained by an independent third party. The results compare sequencing data generated by tGBS and existing array data from a sub-selection of samples and SNP sequences.

Mapping data

Array data was generated from leaf fragment analysis. SeqSNP tGBS involved varying quantities of starting material to ensure sufficient quantity and quality DNA to produce comparable results (Fig 6).

Percentage call differences between array data and tGBS were on average 4.6% with a median of 2%.

44 markers >5% differences vs 369 <=5% differences.

The variability in standard deviations in leaf fragment and 10 leaf punches was the most significant and can be explained by the degree of disruption during homogenisation steps in the extraction protocols used. Similar standard deviations can be observed in the number of reads generated from leaf fragment data.

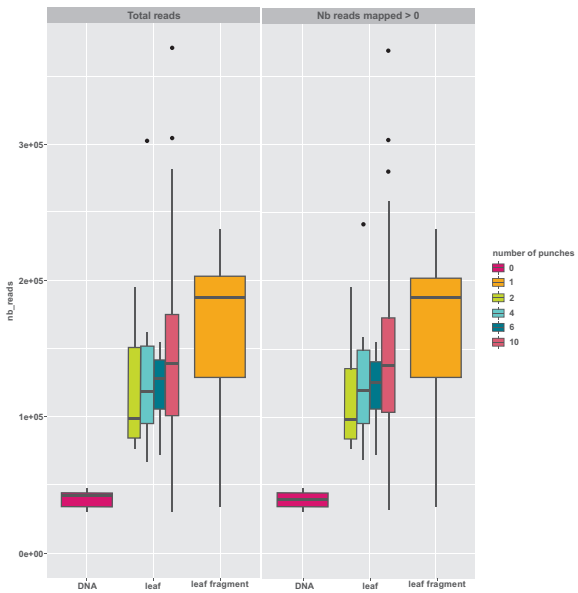

Figure 6: The number of mapped reads is based on bowtie output. The data indicates that there is high concordance of results produced by SeqSNP and array data. The result concludes that there are similar numbers of mapped reads between the two technologies for leaf material. Mapped reads from DNA extracted using crude extraction methods were significantly lower.

The impact of DNA quality can be seen in the number of reads and mapped reads using DNA extracted using crude extraction protocols. There was a 73% reduction in the mappable reads when compared to high quality sbeadex DNA extraction chemistry, coroborating the fact that high quality DNA is a necessity at the moment for all sequence based genotyping.

DNA optimisation

Applying GS in breeding programs requires streamlining of existing process flows as has been applied in marker assisted selection breeding strategies. A particular bottle neck can be the sampling of leaf material intended for sequence based genotyping and GS. The high sample numbers needed for accurate estimation of GEBV are inhibited by the quantity of leaf material required for high quality of DNA extractions.

The SeqSNP pipeline utilises LGC proprietary extraction chemistry (sbeadex) which has been shown to produce on average 4.5 µg total DNA for 5X 5 mm leaf punches (Fig 7). Consistency in DNA quantities extracted and number of reads is achieved with 2 - 6 leaf discs (Fig 7 and 8). Greater variation in the 3rd quartile of data produced in the total number of reads is seen with 10 leaf discs and

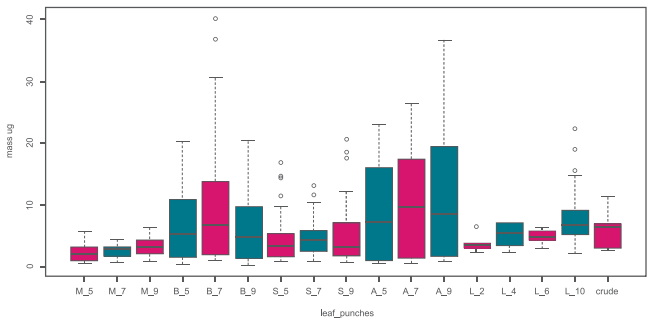

Fig 7: Total mass produced for maize (M), broccoli (B), sunflower (S) and apple (A) from 5, 7 and 9 leaf punches respectively extracted using LGC sbeadex extraction protocols. Total DNA extracted from 2, 4, 6 and 10 5mm leaf discs in anonymised species (L_2, L_4, L_6 and L_10), and crude DNA extraction using proprietary DNA extraction protocol.

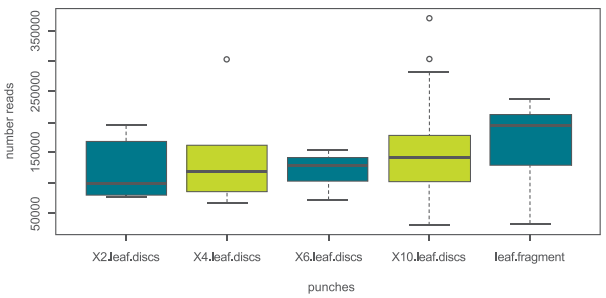

Fig 8: Sequencing read generated for varying quantities of starting leaf material extracted using LGC sbeadex DNA extraction chemistry.

crude extractions. This can be explained by reduced uniformity in homogenisation of starting material and verifies that increasing the number of leaf discs to 10 would lead to variability in total mass of DNA extracted without a substantial net gain in total mass of DNA obtained. In practical terms, the additional time and effort sampling excessive leaf material would be minimised, which again would enable the application of tGBS in existing breeding program workflows.

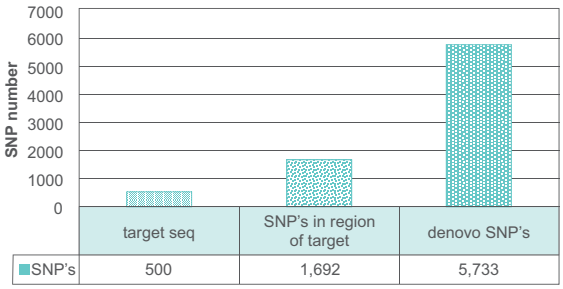

Fig 9: Summary of SNP sequence targeted, SNP detection in surrounding sequence of target SNPs and *de novo* SNP's identified in other regions.

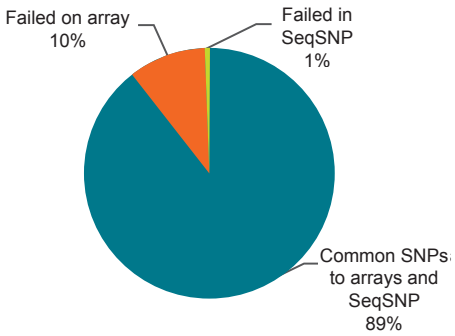

Fig 10: Chart illustrating the percentage of SNPs common for both array and SeqSNP, with percentages of failed SNPs in both technologies.

SNP detection

The quality of data generated by high-resolution sequence-based technology is essential for the breeding community to consider alternative platforms. *De novo* SNP information gained from screening should be expected from technological advances in sequencing protocols. The design of probes to surrounding sequence, and not directly to the SNP in question, makes SeqSNP the next generation method for molecular marker breeding. Specifically, this approach allows variation in germplasm to be accommodated without the re-sequencing of target regions and maintains greater than 95% confidence in data generated. The additional benefits were detection of *de novo* SNP markers using SeqSNP tGBS approach which resulted in the identification of 5,733, previously uncharacterised, additional SNPs (Fig 9).

The data produced in this study (Fig 10) shows directly comparable results for LGC Genomics SeqSNP service with array genotyping. Optimisations of sampling procedures and efficient DNA extraction protocols and flexible probe library design have shown that high quality tGBS can be generated effectively and cost efficiently with a substantial reduction in SNP failure rates. The impact of SNP detection failure rate for estimating GEBV in breeding strategies not only leads to loss of data and increase in costs per sample, but could reduce the value of association data for target traits.

Imputation/coverage

To date, studies have shown that by reducing the sequencing coverage depth, a higher proportion of missing or inaccurate data is obtained. This impacts the overall results by reducing accuracy when identifying the allele-frequency at each locus (7). The minimum depth/coverage for plant breeding strategies and selection of training populations (TP) can be impacted by the read depth/coverage and need careful consideration. It has been shown previously that coverage of X25 can be considered to be sufficient depth for representation commonly used bi-parental crossing strategies for SNP array data (8). The composition of the TP, its size, and its relatedness to the parents are key elements in determining the prediction accuracy of GS (9). Diverse association panels require substantially more markers than recombinant inbred lines for effective mapping and trait association. The target read depth/coverage for the study was X8 coverage. The results for the study generated sequence coverage from leaf material was on average X50, and from previous fragmented leaf DNA extraction methods X70 (Fig 11). From the impartial data produced, LGC's SeqSNP tGBS service is expected to increase in accuracy and reduce the proportion of missing data for the application of tGBS. Using a combination of the SeqSNP service together with imputation could be sufficient for diverse breeding strategies implemented in GS. The pricing structure offered by LGC's service encompasses the variability and scalability in application of tGBS proposition making it a cost-efficient option as an alternative to arrays and for new GS applications.

	min.	max.	avg.	median
DNA	1	170	21.5	17.0
Leaf_fragment	1	383	70.7	42.0
Leaf_2punches	1	355	50.4	33.0
Leaf_4punches	1	279	52.8	35.0
Leaf_6punches	1	268	55.9	37.0
Leaf_10punches	1	515	59.2	37.0

Fig 11: Sequence coverage for 499 SNP markers selected in study.

Conclusion

Comparison of tGBS and array genotyping has been shown in this study to be comparable for data quality and quantity. The study also highlights the advantages of SeqSNP, a tGBS technology, to be a superior for flexibility and scalability over array genotyping, providing the breeding community with a new alternative cost efficient mid-plex genotyping option.

1) A Burrage, Bristol University, “Rapid and affordable genotyping by sequencing optimised for hexploid wheat”, Monogram 2018.

2) Y.Masuda, P.M.VanRaden, I.Misztal, T.J.Lawlor (2018). Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. *Journal of Dairy Science research, Volume 101, Issue 6,* 5194–5206 https://doi.org/10.3168/jds.2017-13310.

3) Granleese, T., Clark, S. A., Andrew A. Swan A, A., Julius, H. J., Werf, V. (2015). Increased genetic gains in sheep, beef and dairy breeding programs from using female reproductive technologies combined with optimal contribution selection and genomic breeding values. *Genet Sel Evol.*; 47(1): 70. https://doi:10.1186/s12711-015-0151-3.

4) Torkamaneh, D., Boyle, B., Belzile, F., (2018). Effcient genome-wide genotyping strategies and data integration in crop plants. *Theoretical and Applied Genetics, Volume 131, Issue 3*, pp 499–511.https://doi.org/10.1007/s00122-018-3056-z.

5) Burridge, A. J., Wilkinson, P. A., Winfield, M. O., Barker, G. L. A., Allen, A. M., Coghill, J. A., Waterfall, C. and Edwards, K. J. (2017). Conversion of array-based single nucleotide polymorphic markers for use in targeted genotyping by sequencing in hexaploid wheat (Triticum aestivum). *Plant Biotechnol*. J., https://doi.org/10.1111/pbi.12834.

6) Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Albrechtsen, A., Nielsen, F, C., Nielsen, R. (2010); *Mol. Biol. Evol.* 7(11):2534–2547. https://doi:10.1093/molbev/msq148.

7) Michael J. Thomson. (2014). High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breed. Biotech.* 2(3):195-212 https://doi.org/10.9787/PBB.2014.2.3.195.

8) Cericola F., Lenk I., Fè D., Byrne S., Jensen C.S., Pedersen M. G., Asp T., Jensen J., Janss L. (2018). Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (Lolium perenne L.); *Plant Sci.*, https://doi.org/10.3389/fpls.2018.00369.

9) Bassi, F.M., et al., Bentley A. R., Charmet. G., Rodomir O., Crossa. J.,(2015). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.), *Plant Sci*. http://dx.doi.org/10.1016/j.plantsci.2015.08.021.

10) Duhnenab, A., Grasb, A., Teyssèdreb, S., Romestantb, M., Claustresb, B., Jean Daydéc, J., Mangin, B. (2017). Genomic selection for yield and seed protein content in soybean: A study of breeding program data and assessment of prediction accuracy, *Crop Science*, vol 57, May - June 2017. doi: 10.2135/cropsci2016.06.0496.

11) Grodzicker T, Williams J, Sharp P, Sambrook J (1974) Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harbor Symp Quant Biol* 39:439–446. doi:10.1101/SQB.1974.039.01.056

www.lgcgroup.com/genomics • genomics@lgcgroup.com

@LGCGenomics LGC.Genomics lgc-genomics

Science for a safer world