THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

The development and application of bioinformatics methods and software tools for computational single nucleotide polymorphism discovery

Michał Tadeusz Lorenc

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

School of Agriculture and Food Sciences

Abstract

Brassica and wheat are important crops for agriculture in Australia and world-wide. Their production is challenging because of biotic stresses such as diseases, and environmental factors including drought and soil salinity.

In comparison to the model species *Arabidopsis thaliana* and rice, the genomes of *Brassica* and wheat are both large and complex. This size and complexity makes it more difficult to determine their genome sequences..

The sequence information produced by Second Generation Sequencing (SGS) technologies allows researchers to identify for example large numbers of molecular genetic markers which can be used to study heritable traits and for applied crop improvement.

SGS technologies are speeding up genome sequencing, but they have led to vast increases in the amount of data resulting in major computational challenges. To manage this data, new computational systems have to be designed to support the SGS based research.

This thesis describes the design, implementation and validation of the SGSautoSNP pipeline, a new approach to call SNPs in large and complex crop genomes using SGS sequences. In our method the reference genome sequence is used only to assemble the reads, and SNPs are then called between these assembled reads. The pipeline includes gene prediction, SNP annotation and identifies low SNP density regions which are more conserved than high SNP density regions.

A total of 638,593 SNPs in the *Brassica napus* AA genome and 881,289 SNPs in the wheat group 7 chromosome arms were identified using the SGSautoSNP pipeline. Validation of 20 *B. napus* AA genome SNPs resulted in a SNP prediction accuracy of around 95%. Of the 28 wheat SNPs that were used for validation of the SGSautoSNP pipeline, 26 (93%) produced the expected genotype.

By combining the SGSautoSNP pipeline together with SnpEff it was possible to determine whole genome SNPs trends, transition to transversion ratios and SNP frequencies across chromosomes. Annotation of *B. napus* AA genome SNPs have revealed that 0.5% of

predicted SNPs are classified as “high effect” SNPs, and these could impact the structure of the proteins or the amino acid transcripts.

The discovered molecular markers, genes, genetic and marker annotations and gene ontology by SGSautoSNP pipeline are stored in a new developed database called SGSautoSNPdb. This information are linked to other databases in order to allow researchers to access information quick and in a biologist friendly manner.

Together, the SGSautoSNP pipeline and SGSautoSNPdb provides tools to help us to understand how natural selection has shaped the evolution of crop genomes and SNPs that can be applied to improve crops in order to secure a sufficient food-source into the future.

Declaration by author

This thesis *is composed of my original work, and contains* no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted *to qualify for the award of any* other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Papers

Lai, K., **Lorenc, M. T**., Lee, H., Berkman, P. J., Bayer, P. E., Muhindira, P. V., Ruperao, P., Fitzgerald, T. L., Zander, M, Chan, C. K., Manoli, S., Stiller, J., Batley, J. & Edwards, D 2015. Identification and characterisation of more than 4 million inter-varietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnology Journal, 13(1)*: 97-104.

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., Patel, D., **Lorenc, M. T.**, Yi, B., Long, Y., Meng, J., Raman, R., Raman, H., Lawley, C., Edwards, D. & Batley, J. 2014. Development of a high-throughput SNP array in the amphidiploid species *Brassica napus*. *Functional and Integrative Genomics, 14(1):* 643-655.

Golicz, A. A., Bayer, P. E., Martinez, P. A., Lai, K., **Lorenc, M. T**., Alamery, S., Hayward, A., Tollenaere, R., Batley, J., Edwards, D., Long, Y. & Meng, J. 2013. Characterising diversity in the *Brassica* genomes. *Acta Horticulturae*, *1005* : 33-48.

Zander, M., Patel, D. A., Van de Wouw, A., Lai, K., **Lorenc, M. T.**, Campbell, E., & Batley, J. 2013. Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional & Integrative Genomics*, *13*(3): 295-308.

Berkman, P. J., Visendi, P., Lee, H. C., Stiller, J., Manoli, S., **Lorenc, M. T**., Lai, K., Batley, J., Fleury, D., Šimková, H., Kubaláková, M., Weining, S., Doležel, J. & Edwards, D. 2013. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal*, 11:564–571.

Lorenc, M. T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P. J., Lai, K., Batley, J., Edwards, D. 2012. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology*, 1(2),370-382.

Hayashi, S., Reid, D. E., **Lorenc, M. T**., Stiller, J., Edwards, D., Gresshoff, P. M., & Ferguson, B. J. 2012. Transient Nod factor dependent gene expression in the nodulation competent zone of soybean (*Glycine max* [L.] Merr.) roots. *Plant Biotechnology Journal*, *10*(8), 995-1010.

Lai K., Duran C., Berkman P. J., **Lorenc M. T.**, Stiller J., Manoli S., Hayden, M., Forrest, K., Fleury, D., Baumann, U., Zander, M., Mason. A., Batley, J & Edwards D. 2012. Single Nucleotide Polymorphism Discovery from Wheat Next Generation Sequence Data. *Plant Biotechnology Journal*, 10 (6): 743-749

Tollenaere, R., Hayward, A., Dalton-Morgan, J., Campbell, E., McLanders, J., **Lorenc, M.**, Manoli, S., Stiller, J., Raman, R., Raman, H., Edwards, D. & Batley, J. 2012. Identification and characterisation of candidate *Rlm4* blackleg resistance genes in *Brassica napus* using next generation sequencing. *Plant Biotechnology Journal*, 10 (6): 709-715

Edwards, D., Wilcox, S., Barrero, R. A., Fleury, D., Cavanagh, C. R., Forrest, K. L., Hayden, M. J., Moolhuijzen, P., Keeble-Gagnère, G., Bellgard, M. I., **Lorenc, M. T.**, Shang, C. A., Baumann, U., Taylor, J. M., Morell, M. K., Langridge, P., Appels, R., & Fitzgerald, A. 2012. Bread matters: A national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnology Journal*, 10 (6): 703-708

Reid, D. E., Hayashi, S., **Lorenc, M.**, Stiller, J., Edwards, D., Gresshoff, P.M. & Ferguson, B. J. 2012. Identification of systemic responses in soybean nodulation by xylem sap feeding and complete transcriptome sequencing reveal a novel component of the autoregulation pathway. *Plant Biotechnology Journal*, 10 (6): 680-689

Lai, K., **Lorenc, M. T.** & Edwards, D. 2012. Genomic databases for crop improvement. *Agronomy*, 2: 67-73

Berkman, P. J., Lai, K., **Lorenc, M. T.** & Edwards, D. 2012. Next generation sequencing applications for wheat crop improvement. *American Journal of Botany*, 99 (2): 365-371

Lee, H., Lai, K., **Lorenc, M. T.**, Imelfort, M., Duran, C. & Edwards, D. 2012. Bioinformatics tools and databases for analysis of next generation sequence data. *Briefings in Functional Genomics*, 11 (1), 12-24

Lai, K., Berkman, P. J., **Lorenc, M. T.**, Duran, C., Smits, L., Manoli, S., Stiller, J. & Edwards, D. 2012. WheatGenome.info: An integrated database and portal for wheat genome information. *Plant and Cell Physiology*, 53(2): e2(1–7)

Berkman, P. J., Skarshewski, A., Manoli, S., **Lorenc, M. T.**, Stiller, J., Smits, L., Lai, K., Campbell, E., Kubaláková, M., Šimková, H., Batley, J., Doležel, J., Hernandez, P., & Edwards, D. 2012. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and Applied Genetics*, 3: 423-432

Berkman, P. J., Skarshewski, A., **Lorenc, M. T.**, Lai, K., Duran, C., Ling, E.Y., Stiller, J., Smits, L., Imelfort, M., Manoli, S., McKenzie, M., Kubaláková, M., Šimková, H., Batley, J., Fleury, D., Doležel, J. & Edwards, D. 2011. Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnology Journal*, 9 (7): 768-775

Book chapters

Lorenc, M., Boskovic, Z., Stiller, J., Duran, C. & Edwards, D. 2012. Role of Bioinformatics as a Tool for Oilseed Brassica Species. In Genetics, Genomics and Breeding of Oilseed Brassicas. Edwards, D., Parkin, I. A. P. & Batley, J. *Science Publishers*, (USA) pp 194-205

Presentations

Annotations tools at The Applied Bioinformatics Group

- Presented on 28th of September 2010
- Invited talk at Bayer CropScience, Gent, Belgium

SNP discovery in the Canola genome

- Presented on 5th of October 2010
- Invited talk at Bayer CropScience, Gent, Belgium

Genome wide SNP discovery: from *Brassica* to Wheat

- Presented on 16th of March 2011
- ACPFG Joint Research Meeting, Adelaide, Australia

Publication included in this thesis

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., Patel, D., **Lorenc, M. T.**, Yi, B., Long, Y., Meng, J., Raman, R., Raman, H., Lawley, C., Edwards, D. & Batley, J. 2014. Development of a high-throughput SNP array in the amphidiploid species *Brassica napus*. *Plant Biotechnology Journal, accepted*

Summarised and partially incorporated as paragraph in Chapter 3.

Contributor	**Statement of contribution**
Lorenc, M. T.,	SNP discovery (80%)
Dalton-Morgan, J.	Conception and design (20%) Analysis (30%) Drafting and writing (40%)
Hayward, A.	Analysis (5%) Drafting and writing (5%)
Mason, A. S.	Analysis (5%) Drafting and writing (5%)
Alamery, S.	Analysis (5%)
Tollenaere, R.	Analysis (5%)
Campbell, E.	Analysis (5%)
Patel, D.	Analysis (5%)
Yi, B.	Analysis (5%)
Long, Y.	Analysis (5%)
Meng, J.	Analysis (5%)
Raman, R.	Analysis (5%)
Raman, H.	Analysis (5%)
Lawley, C.	Conception and design (10%)
Edwards, D.	SNP discovery (20%) Conception and design (30%) Drafting and writing (10%)
Batley, J.	Conception and design (50%) Analysis (15%) Drafting and writing (40%)

Berkman, P. J., Visendi, P., Lee, H. C., Stiller, J., Manoli, S., **Lorenc, M. T**., Lai, K., Batley, J., Fleury, D., Šimková, H., Kubaláková, M., Weining, S., Doležel, J. & Edwards, D. 2013. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal*, 11:564–571.

Incorporated some results in Chapter 4.

Contributor	**Statement of contribution**
Lorenc, M. T.	Coding and analysis (15%)
Visendi, P.	Coding and analysis (15%)
Lee, H. C.	Coding and analysis (15%)
Stiller, J.	Coding and analysis (5%)
Manoli, S.	Coding and analysis (5%).
Lai, K.	Coding and analysis (5%).
Lorenc, M. T.	Drafting and writing (5%)
Batley, J.	Drafting and writing (5%)
Lai, K.	Drafting and writing (5%)
Doležel, J.	Drafting and writing (5%)
Berkman, P. J.,	Conception and design (35%) Coding and analysis (30%) Drafting and writing (40%)
Edwards, D.	Conception and design (45%) Coding and analysis (10%) Drafting and writing (40%)
Weining, S.	Conception and design (10%)
Doležel, J.	Conception and design (10%)

Lorenc, M. T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P. J., Lai, K., Batley, J., Edwards, D. 2012. Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology*, 1(2),370-382.

Partially incorporated as paragraphs in Chapter 2 and 4.

Contributor	**Statement of contribution**
Lorenc, M. T.,	Conception and design (30%) Coding (70%) Analysis (20%) Drafting and writing (30%)
Edwards, D.	Conception and design (30%) Analysis (5%) Drafting and writing (50%)
Batley, J.	Conception and design (30%) Analysis (15%) Drafting and writing (20%)
Stiller, J.	Conception and design (10%) Coding (30%) Analysis (5%)
Hayashi, S.	Analysis (15%).
Manoli, S.	Analysis (15%).
Lai, K.	Analysis (5%)
Ruperao, P.	Analysis (5%)
Visendi, P.	Analysis (5%)
Berkman, P. J.	Analysis (5%)
Lee, H.	Analysis (5%)

Contributions by others to the thesis

Professor Dave Edwards provided the idea for SGSautoSNP's algorithm. Sahana Manoli and Paul V. Muhindira provided alignments for *Brassica* and Wheat. Dr Jiri Stiller tested SGSautoSNP and provided bug reports. Kaitao Lai upgraded SGSautoSNP’s VCF implementation and Philipp Bayer performed a few improvements. Associate Professor Jacqueline Batley, Jessica Dalton-Morgan and Satomi Hayashi performed the laboratory validation of the putative SNPs discovered in SGSautoSNP.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

This thesis and the research presented within would not have been possible without the support of a number of people. Firstly, I would like to thank my supervisors Professor David Edwards, Associate Professor Jacqueline Batley and Dr Jiri Stiller for their support, encouragement and advice throughout this project. I could not have asked for better role models, each supportive, patient and inspirational. It is no easy task to review a thesis and therefore I am grateful for Jacqueline and Dave thoughtful and detailed comments.

I would also like to thank Kaye Hunt and my committee members, Dr Elizabeth Aitken and Dr Brett Ferguson for the friendly guidance and suggestions, things would be so much harder without them.

Applied Bioinformatics Group members have contributed greatly to my personal and professional time. The group has been a source of friendships as well as good advice. Thanks to SAFS/QAAFI Postgrad Association in organising international lunches and badminton which were also good source of friendships.

I am grateful to my family, my parents Danuta and Antoni and my sister Kasia for their endless support, love and believing in me. I have always been assured that there is no problem on which I cannot seek their help and advice.

This thesis is a dedicated to my sweet daughter, Adela Joyce Lorenc.

Acknowledgement of contribution

The author would like to acknowledge funding support from the University of Queensland, Prof David Edwards and funding support from the ARC

Keywords

bioinformatics, second generation sequencing, molecular markers, single nucleotide polymorphisms, *Brassica*, wheat, visualisation, genomics, genetics, databases

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102 Bioinformatics 30%

ANZSRC code: 080301 Bioinformatics Software 70%

Fields of Research (FoR) Classification

FoR code: 0604 Genetics, 20%

FoR code: 0607 Plant Biology, 20%

FoR code: 0803 Computer Software, 60%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AFLP	Amplified Fragment Length Polymorphism
API	Application Programming Interface
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BWT	Burrows Wheeler Transformation
CPU	Central Processing Unit
CSS	Cascading Style Sheets
EST	Expressed Sequence Tag
GIL	Global Interpreter Lock
GO	Gene Ontology
GOA	Gene Ontology Annotation
HPC	High-Performance Computing
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
MYA	Million Years Ago
PBS	Portable Batch System
PCR	Polymerase Chain Reaction
RAD	Restriction site Associated DNA
REST	Representational State Transfer
RFLP	Restriction Fragment Length Polymorphism
RRL	Reduced Representation Library
RT	Reference Transcript
SGS	Second Generation Sequencing
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SOAP	Short Oligonucleotide Analysis Package
SQL	Structured Query Language
SSR	Simple Sequence Repeats
STS	Sequence Tagged Site
TE	Transposable Element
Ts	Transitions
Tv	Transversions

Chapter 1: Introduction

Brassica napus L. (canola) and bread wheat (*Triticum aestivum* L.) are key crops for Australian agriculture and the export economy. Furthermore, they also have huge economic and social significance worldwide. Production of these crops is often reduced by environmental and demographic factors such as drought, disease and soil salinity. In addition, the world's population continues to grow at a rapid speed and by 2050 it is predicted that there will be more than 9 billion people on the earth, over 2 billion more than today, where there are already 1 billion people suffering from hunger, 19 March 2014). Food production around the world will need to increase by 70% in order to be able to feed this dramatically increased population (FAO, 2009). To overcome these challenging environmental and demographic factors, breeding techniques must be improved to accelerate production of new crops, and provide farmers with varieties of *Brassica* and wheat with increased yield, biotic and abiotic stress tolerance.

Both canola and wheat have genomes which are large and complex, making it difficult to apply modern molecular biological techniques. These large and complex genomes are due to genome duplication and the amplification of transposable elements. Research has shown that genome duplication occurred in almost all ancient flowering plants (Doyle et al., 2008, Soltis and Soltis, 1999) and polyploidy results in increased genome complexity (Soltis et al., 2004).

The discovery of genetic variations (polymorphisms) which can be used as molecular markers is generally easier in diploid species, because a marker often has a unique physical location, or single locus. However, in a polyploid plant species the discovery of polymorphisms is not easy because of the presence of homoeologues (Bundock et al., 2009). Despite the problems associated with genome complexity it was possible to develop a pipeline, called SGSautoSNP (Lorenc et al., 2012), for accurate SNP discovery in large complex plant genomes during this PhD project. This pipeline was used to successfully discover polymorphisms in the group 7 chromosomes (7A, 7B and 7D) of wheat, as well as across the genome of *B. napus*. The pipeline includes scripts for gene and SNP annotation which uses SNAP gene prediction software, and SNPeff, a SNP annotation and effect prediction tool. In addition, SGSautoSNP can be used to find low SNP density regions and perform gene ontology analysis using goatools to find enrichment of GO terms for genes in low or high SNP density regions.

1.1. Species of interest

1.1.1. Wheat

1.1.1.1. Wheat production

Wheat is Australia's largest crop, followed by barley and *Brassica napus* canola (http://www.abs.gov.au/, 18 March 2014). Between 2006 and 2010 Australia was the fifth largest exporter of wheat, the ninth largest producer of wheat and the sixteenth largest consumer of wheat in the world (http://www.fas.usda.gov/psdonline/, 18 March 2014). In 2009 Australia exported wheat to a value of $4.9 billion ((Australian-Bureau-of-Statistics, 2010) http://www.csiro.au/Outcomes/Food-and-Agriculture/Cereal-varieties-crop-management.aspx, 18 March 2014). From 1970 to 2012 Australia was able to increase wheat production from 7.9 to 29.9 million tonnes (380%), but production varied during this period because of environmental factors such as drought and disease (Figure 1.1).

Wheat is important worldwide because it contributes nearly 20% of the world's daily energy consumption. The consumption of wheat is expected to increase further in the coming years, because many Asian populations, which were dependent on rice as their primary food source, are eating an increasing quantity of wheat.

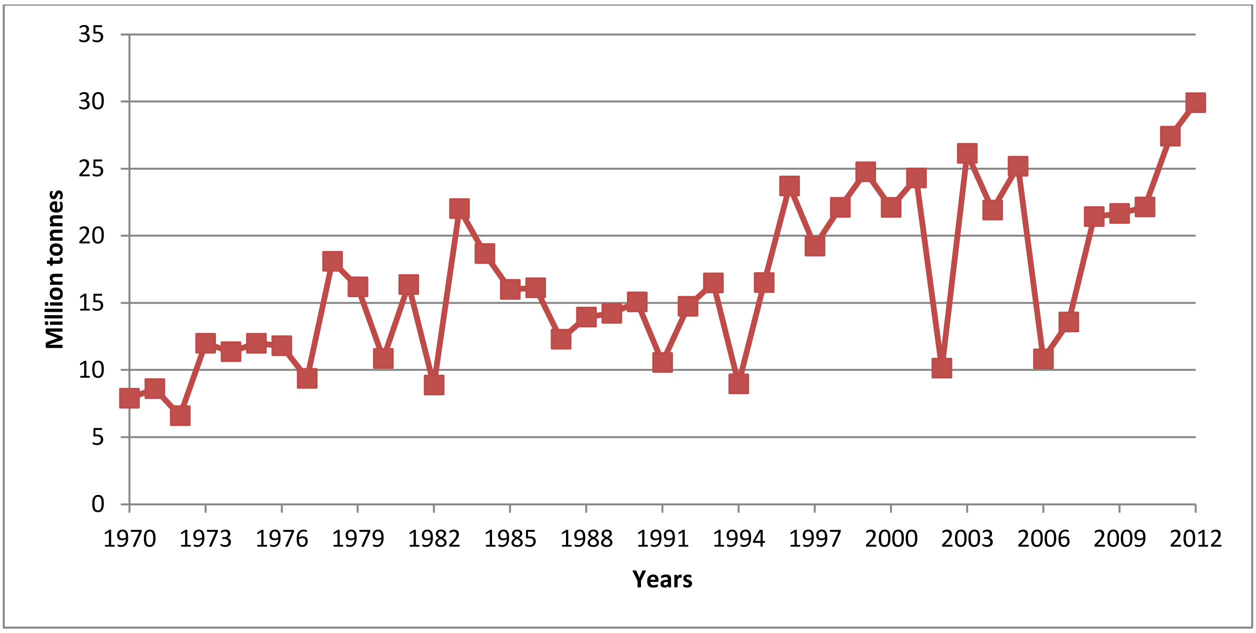


Figure 1.1: Australian wheat production has grown from 1970 to 2011.

1.1.1.2. Wheat evolution

In a divergence event, ancestral wheat split into three different diploid species between 2.5 and 6 million years ago (MYA) (Chantret et al., 2005). Between 0.5 and 3 MYA an inter-species hybridisation event occurred, which combined the genomes of *Triticum urartu* (AA) and an unknown species that provides the BB genome. The result of this was the production of the allotetraploid genome of *T. turgidum* (AABB). A second inter-species hybridisation event occurred between 7000-9500 years ago, after the domestication of wheat, and created the allohexaploid genome of *T. aestivum* (AABBDD) from domesticated *T. turgidum* (AABB) and *Aegilops tauschii* (DD). Allohexaploid means that its genome consists of six sets of chromosomes from three diploid genomes. In total *T. aestivum* has 42 chromosomes, because all three diploid donor had seven pairs of chromosomes (2n=6x=42). This evolution can be viewed in Figure 1.2.

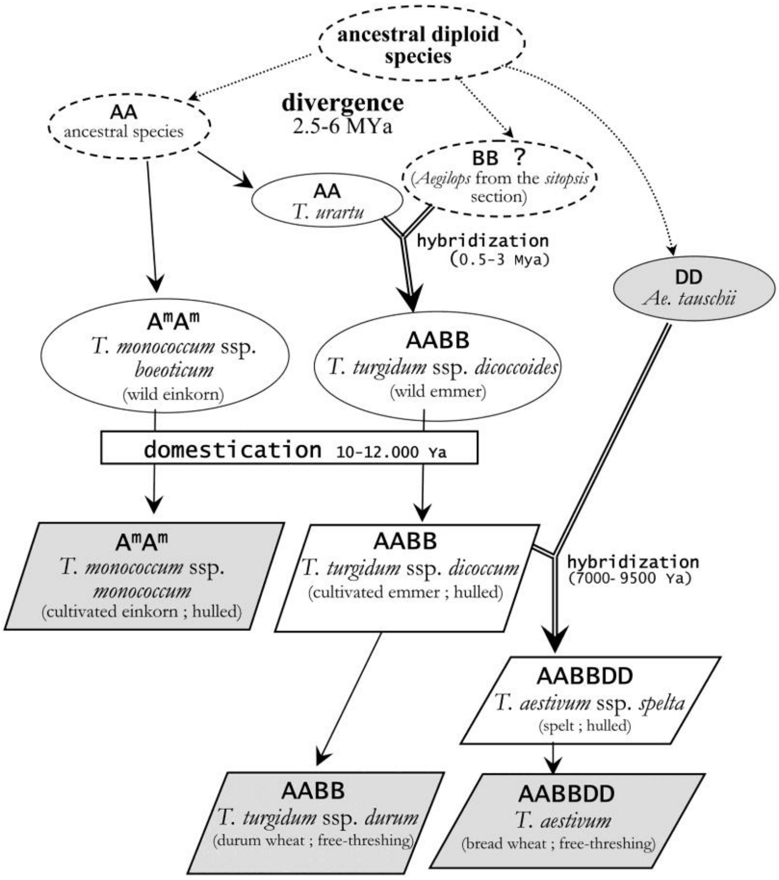

ancestral diploid species
divergence
2.5-6 MYa
AA
ancestral species
BB ?
(Aegilops from the sitopsis section)
AA
T. urartu
hybridization
(0.5-3 Mya)
DD
Ae. tauschii
A^mA^m
T. monococcum ssp. boeoticum
(wild einkorn)
AABB
T. turgidum ssp. dicoccoides
(wild emmer)
domestication 10-12.000 Ya
A^mA^m
T. monococcum ssp. monococcum
(cultivated einkorn ; hulled)
AABB
T. turgidum ssp. dicoccum
(cultivated emmer ; hulled)
hybridization
(7000- 9500 Ya)
AABBDD
T. aestivum ssp. spelta
(spelt ; hulled)
AABB
T. turgidum ssp. durum
(durum wheat ; free-threshing)
AABBDD
T. aestivum
(bread wheat ; free-threshing)

Figure 1.2: A graphical representation of the evolution of wheat species. Domesticated species are in squares and wild species are in circles. Unknown or ancestral species are surrounded by a dotted circle. Actual species are surrounded by a plain line circle (adapted from Chantret et al., 2005).

1.1.1.3. Wheat genome

The bread wheat (*Triticum aestivum*) genome is 17 Gbp in size and around 6 times larger than the human genome (Paux et al., 2010). It consists of 75% - 90% repeats (Flavell et al., 1977, Wanjugi et al., 2009) and is hexaploid, containing the A, B and D genomes, each with 7 homoeologous chromosomes. Most of the repeats are found as transposable elements (TEs) with some low-complexity repeats. This makes it more difficult to assemble this genome sequence, because it is easier to assemble shotgun DNA sequence of unique genic regions rather than long regions of repetitive DNA. A further problem is that through polylpoidy many of the genic regions, which would normally be considered unique in diploid genomes, have homoeologous copies. In a genome with homoeologous chromosomes it is more difficult to identify the exact location of genes because of the difficulty differentiating between homoeologues (Gill et al., 1991, Pedersen and Langridge, 1997).

1.1.1.4. Wheat sequence availability

Draft genome sequences of wheat were recently published; *T. aestivum* (Bread wheat) has a hexaploid AABBDD genome (Brenchley et al., 2012, International Wheat Genome Sequencing, 2014), *A. tauschii* has a diploid DD genome (Jia et al., 2013) and the *T. urartu* has a diploid AA genome (Ling et al., 2013). Sequences of individual bread wheat chromosomes arms have also been published; group 1 (1A, 1B, 1D) (Wicker et al., 2011), 4A (Hernandez et al., 2012), 5A (Vitulo et al., 2011), 5B (Sergeeva et al., 2014) and group 7 chromosomes (7A, 7B and 7D) were (Berkman et al., 2013, Berkman et al., 2012b, Berkman et al., 2011).

1.1.2. Brassica

1.1.2.1. *Brassica* importance globally

The *Brassica* genus contains many economically and agronomically important crop species with a variety of adaptation for cultivation under various environmental conditions (Batley et al., 2007). No other plant genus contains more agricultural and horticultural crops than the *Brassicas* (Hayward et al., 2012). Across many countries *Brassica* species are sources of condiments, fresh and preserved vegetables, vegetable oil, dietary fibre, vitamin C and anticancer compounds (Bohuon et al., 1998, Koo et al., 2011, Lan et al., 2000). *Brassica* species contribute to approximately 12% of the worldwide edible oil supplies and approximately 10% of the world's vegetable crop production (Mun et al., 2010). The six major cultivated *Brassica* species include *B. rapa (Chinese cabbage* and *turnip), B. oleracea (*broccoli, cabbage and cauliflower*), B. nigra* (black mustard)*, B. napus* (canola/rapeseed/oilseed rape), *B. juncea* (Indian mustard) and *B. carinata* (Ethiopian mustard). *B. rapa* (diploid AA genome) and *B. oleracea* (diploid CC genome) are grown mostly as vegetable crops. *B. nigra* (diploid BB genome) is used as a source of mustard condiment. *B. napus* (allotetraploid AACC genomes) is mainly an oil crop, *B. juncea* (allotetraploid AABB genomes) is both an oil and condiment crop, and *B. carinata* (allotetraploid BBCC genomes) is mainly a condiment crop. The *Brassica* family contributes significantly towards world food and fodder production. The relationship between the six *Brassica* species is described in the triangle of U (U, 1935) (see Figure 1.3).

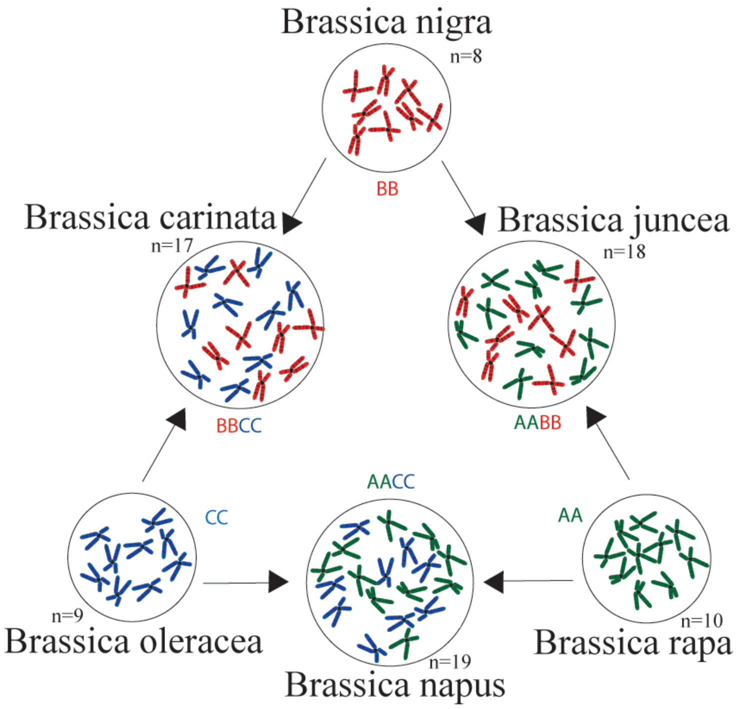


Figure 1.3: U's triangle depicting the genetic relationships between the six cultivated *Brassica* species. Chromosomes from each of the genomes A, B and C are represented by different colours. The letter n represents the number of chromosomes in each genome (adapted from http://en.wikipedia.org/wiki/Triangle_of_U).

1.1.2.2. *Brassica* evolution

Brassica diverged from the model plant *Arabidopsis thaliana* approximately 20 Mya (Yang et al., 1999). In contrast, the lineages of the species *B. rapa* (A genome) and *B. oleracea* (C genome) diverged about 3.7 MYA (Inaba and Nishio, 2002). In 2000, the genome of *A. thaliana* became the first plant genome to be sequenced (Arabidopsis-Genome-Initiative, 2000). The *A. thaliana* genome size is only 146 Mb, which is small compared to *Brassica* species (see Table 1.1) and contains few repetitive sequence regions (Bevan and Walsh, 2005). Arabidopsis and *Brassica* share approximately 85% nucleotide identity in coding regions (Cavell et al., 1998). The high level of sequence similarity between Arabidopsis and *Brassica* allows the study of the structure of *Brassica* genomes without the complete *Brassica* genome sequence being available. The *B. rap*a genome sequence was published in 2011 by the multinational Brassica Genome Sequencing Project (Wang et al., 2011). *B. rapa* was selected as the first *Brassica* species to be sequenced, because of its relatively small genome size of 529 Mb (Arumuganathan and Earle, 1991, Choi et al., 2007) in comparison to other *Brassica* species sizes (see Table 1.1), and lower complexity compared to *Brassica oleracea*. Analysis has shown that 90% of the *A. thaliana* genome and 91% of the *B. rapa* genome could be aligned in collinear blocks (Wang et al., 2011). Chromosome rearrangements resulted in chromosome number variation for the three diploid *Brassica* species, *B. nigra* (B genome; n = 8), *B. oleracea* (C genome; n = 9), and *B. rapa* (A genome; n = 10) (Lysak et al., 2005).

Hybridisation between diploid genomes, followed by chromosome doubling, produces polyploids. This lead to the creation of the amphidiploid *Brassica* species: *B. juncea*, *B. carinata* and *B. napus*. These species contain four genomes, derived from two different ancestral species. For example *B. napus* has 19 chromosomes (n = 19), 10 chromosomes from the AA genome and 9 chromosomes from the CC genome. Various *Brassica* species genome sizes currently can only be estimated with methods like Feulgen microdensity measurements and flow cytometry (see Table 1.1). .

Table 1.1: ***Brassica*****'s genome sizes (adapted from http://www.brassica.info/info/reference/genome-sizes.php).**

Species	MBp	Source	Method	Reference
B. rapa (AA)	468 - 516		flow cytometry	(Arumuganathan and Earle, 1991)
B. rapa (AA)	507	var. chinensis Pak choi	flow cytometry	(Arumuganathan and Earle, 1991)
B. rapa (AA)	511	var. rapifera turnip	flow cytometry	(Arumuganathan and Earle, 1991)
B. rapa (AA)	529		flow cytometry	(Johnston et al., 2005)
B. oleracea (CC)	599 - 618	var. italica broccoli	flow cytometry	(Arumuganathan and Earle, 1991)
B. oleracea (CC)	603	var capitata cabbage	flow cytometry	(Arumuganathan and Earle, 1991)
B. oleracea (CC)	628	var gemmifera Brussels sprout	flow cytometry	(Arumuganathan and Earle, 1991)
B. oleracea (CC)	628 - 662	var botrytis cauliflower	flow cytometry	(Arumuganathan and Earle, 1991)
B. oleracea (CC)	696		flow cytometry	(Johnston et al., 2005)
B. oleracea (CC)	752		feulgen	(Bennett and Smith, 1991)
B. oleracea (CC)	868		feulgen	(Bennett and Smith, 1976)
B. napus (AACC)	1129-1235	rapeseed	flow cytometry	(Arumuganathan and Earle, 1991)
B. napus (AACC)	1132		flow cytometry	(Johnston et al., 2005)

1.1.2.2.1. *Brassica rapa*

In India, Sweden and Finland*,* *B. rapa* is grown as an oilseed crop, but in China and Japan it is grown mostly as a leafy vegetable crop (Rakow, 2004). The leafy vegetable crops are separated into seven morphologically distinct vegetable varieties, including: var. *pekinensis* (Chinese cabbage), var. *narinosa* (Chinese savoy/taasai), var. *chinensis* (bok-choi), var. *parachinensis* (false pak choi), var. *japonica* (Mizuna/Japanese salad green), var. *campestris* (annual turnip rape) and var. *rapa* (turnip) (Dixon, 2007, Rakow, 2004). Turnip is basically a cool climate crop which is resistant to frost. Its roots are also grown for feeding livestock during autumn and winter (http://www.hort.purdue.edu/newcrop/duke_energy/brassica_rapa.html, 04 October 2013). The winter oilseed types of *B. rapa* species have the advantage that they are cold tolerant and can be grown where temperatures are too low for *B. napus*. Spring types of *B. rapa* flower earlier compared to *B. napus* and contribute to oil production in northern and western Canada (Mendham and Salisbury, 1995).

1.1.2.2.2. *Brassica oleracea*

B. oleracea represents the *Brassica* diploid CC genome and contains many vegetables including cabbage, kale, collard greens, Chinese broccoli, cauliflower, broccoli, Brussels sprouts and kohlrabi. It is rich in vitamins B and C, calcium, iron, magnesium, phosphorus, potassium and zinc. It also contains high levels of anti-oxidant and anti-cancer compounds (Weerakoon et al., 2009). *B. oleracea* are important vegetables in many countries, especially in Northern Europe and Central Asia. Cauliflower is the main vegetable in India because it can be stored without refrigeration. Cabbage and kohlrabi also have these benefits. Brussels sprouts are able to grow through a mild winter.

1.1.2.2.3. *Brassica napus*

B. napus is also known as canola, rapeseed or oilseed rape. Its genome is allopolyploid (AACC) and was produced by hybridisation between the *B. rapa* AA genome and *B. oleracea* CC genomes. Genetic mapping confirmed that the AA and CC genomes are intact in *B. napus* and have not been substantially rearranged (Parkin et al., 1995). *B. napus* is grown in Australia, East and South Asia, Europe and North and South America. It is used for bio-fuel, vegetable oil for human consumption (canola) and as a protein additive for animal stock feed.

Canola (Canadian oil, low acid) must contain less than 2% erucic acid, a known toxin, as this level causes no harm to humans. Its seeds contain about 40-43% oil. The oil of canola is used in the production of margarine and cooking oil because it has low saturated fat content (less than 7%), and is high in monounsaturated fats and omega-3 fatty acids. Canola was first grown commercially in Australia in the 1970s, and from 1970 to 2011 Australia increased production of canola from 13,000 tonnes to 299,200 tonnes per annum (see Figure 1.4). In Australia, canola is widely grown across south-east Australia and Western Australia (see Figure 1.5)

In 2012 Australia was the 14th biggest producer of Canola worldwide with 299,200 tonnes. At the same time the worldwide production was 22,254,971 tonnes (http://www.fao.org) (see Figure 1.6). Australia's export markets are Japan, China, Pakistan, Europe and Bangladesh. Canola is now Australia's third-largest crop after wheat and barley (http://www.abs.gov.au/, 18 March 2014). Furthermore, Australia exports most of its canola.

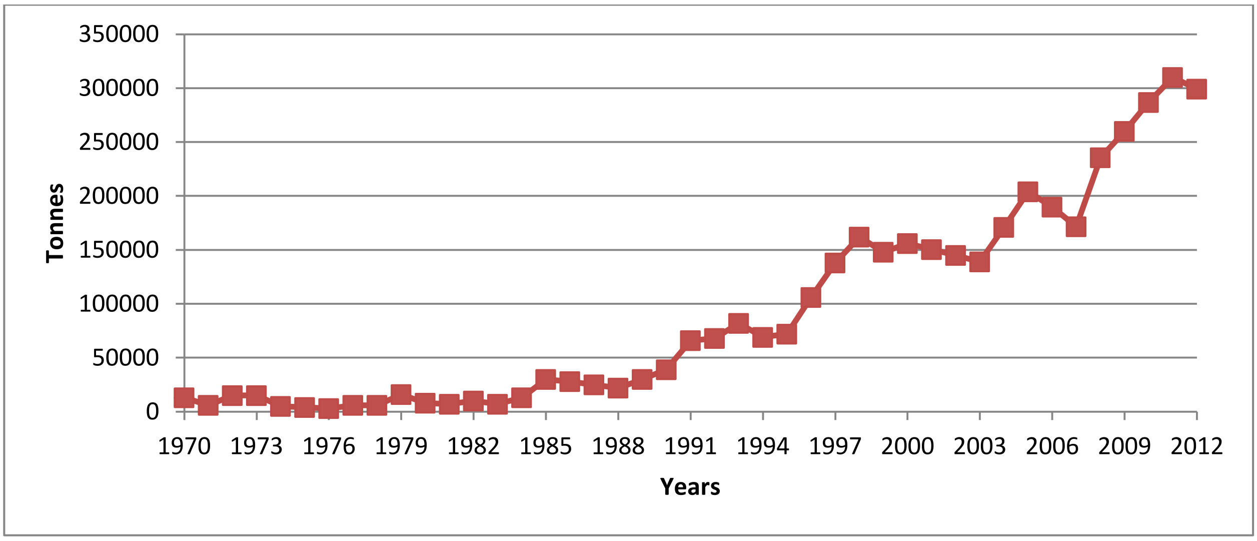


Figure 1.4: Production development of Canola in Australia between 1970 and 2011.

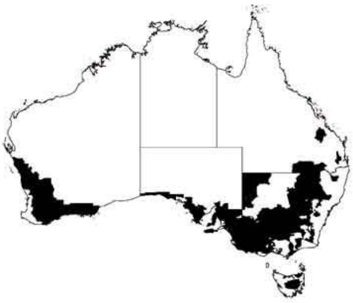


Figure 1.5: Canola production regions within Australia (adapted from http://www.rirdc.gov.au/programs/established-rural-industries/pollination/canola.cfm).

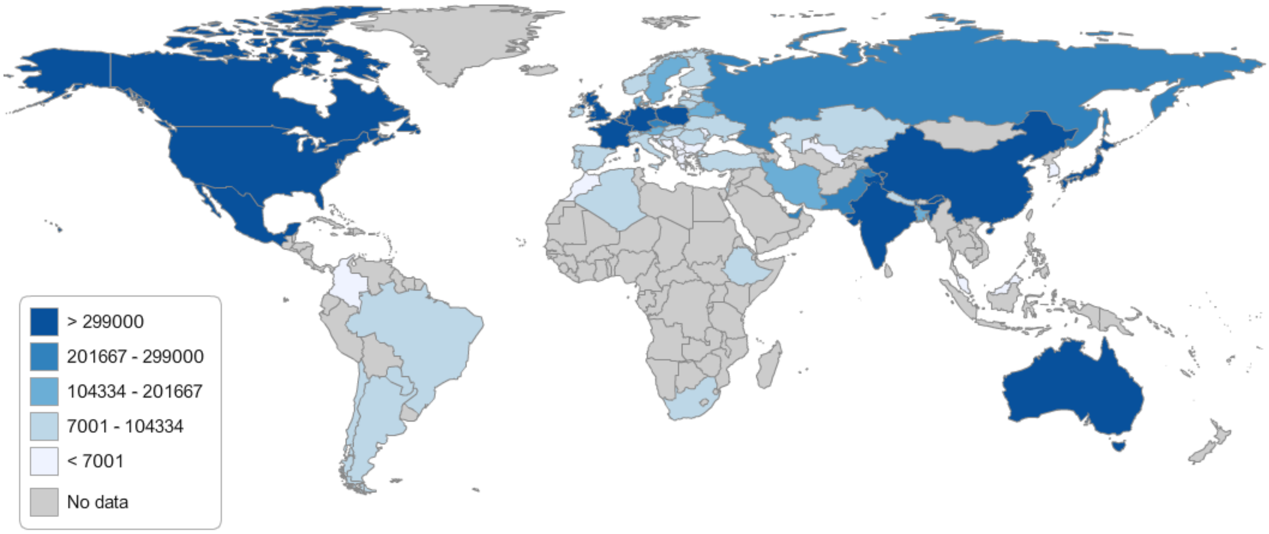


Figure 1.6: Global Oil, rapeseed production in 2012 (scale refers to tonnes) (source http://faostat3.fao.org/faostat-gateway/go/to/browse/Q/QC/E).

1.1.2.3. Brassica sequence availability

Draft genome sequences for *B. juncea* (AABB) and *B. nigra* (BB) have been produced with the possibility of publication in the near future (Golicz et al., 2012). The *B. rapa* (AA) (Wang et al., 2011) and *B. oleracea* (CC) (Liu et al., 2014, Parkin et al., 2014) genomes have been published. In future research, the published *B. oleracea* or *B. napus* genomes could replace the proprietary C genome used in chapter 3 and the whole *Brassica* analysis of chapter 3 could consider all 19 chromosomes instead just the 10 chromosomes of the A genome. These genome sequences will enhance genetic studies and provide insight into the genetic basis of important agronomic traits including nutritional seed properties and resistance to biotic and abiotic stressors (Getinet et al., 1997).

1.2. DNA sequencing technologies

1.2.1. First Generation Sequencing

The Sanger method (Sanger et al., 1977) is considered as a first-generation sequencing technology and was used from the 1970s until now (Metzker, 2010). In 2001 the first human genome was sequenced using this technology (Lander et al., 2001, Venter et al., 2001). This technology produced reasonably long sequences, up to several hundred nucleotides, with a high degree of certainty regarding the sequence accuracy. Disadvantages of the technology are the time required to generate the sequence data, as well as the limited ability to parallelise the process in order to lower the overall cost to generate data in high volumes.

1.2.2. Second Generation Sequencing

Second Generation sequencing (SGS) was introduced in 2005 by 454 Life Sciences (http://www.454.com, Margulies et al., 2005). In late 2006 another SGS platform called the Genome Analyzer was released by Solexa. Solexa has been acquired by Illumina shortly after the release of this platform (http://www.illumina.com/). SGS has accelerated DNA and RNA sequencing by producing a series of iterations continually increasing volumes of sequence data with increasing quality and read length at a lower price and increased speed (Metzker, 2010). The cost per genome dropped down from $100 million in 2001 to $10,000 in 2011 (https://www.genome.gov/sequencingcosts/ 22 October 2013). Roche’s 454 GS FLX Titanium technology (Margulies et al., 2005) is able to produce one million reads up to 1,000 nucleotides in length in one day (http://www.454.com). Illumina’s

HiSeq2000 is able to produce 600 billion nucleotides of sequence data with a read length of 150 nucleotides in around 10 days. The MiSeq can produce paired reads up to 300 bp long (http://www.illumina.com). The Illumina platforms are able to produce two different types of paired read libraries. The first library is called paired-end and can generate paired reads which have a maximum insert size under 1 Kbp. The second one is called mate-pair and is able to produce reads with maximum insert sizes of less than approximately 20 Kbp. Life Technologies' first SGS technology was called SOLiD, which is now discontinued, but was able to produce over 20 billion nucleotides per day, with a read length of up to 75 nucleotides. It was based on Polonator technology (Valouev et al., 2008). Their second platform is the Ion Torrent which produces sequence reads of 400 bp, with up to 1 Gbp of data per run (http://www.lifetechnologies.com). Life technologies' third technology is the Ion Proton which produces sequence read lengths of up to 200 bp and has a throughput of approximately 10 Gb per run (http://www.mrdnalab.com/ion_proton.html, 30 January 2014).

1.2.3. Third Generation Sequencing

Recently, new technologies were developed which promise greater volumes of sequence data and longer reads than Second Generation Sequencing. Pacific Biosciences' SMRT (Single Molecule Real Time) sequencing technology produces read lengths of around 1,000 bp with the potential to take snapshots of shorter reads over an extended fragment of over 10,000 bp (Eid et al., 2009). Oxford Nanopore introduced its USB size sequencer 'MinION', and their bench top sequencer 'GridION'. They produce extremely long reads of about 50,000 bases in length, while the 'GridION' can sequence the entire human genome in 15 minutes and the MinION' in 60 minutes. However this nanopore technology has yet to be demonstrated in a public laboratory.

1.3. Sequence analysis tools

1.3.1. Quality control of Second Generation Sequencing

SGS technologies have not reached yet the quality of sequence data compared to traditional Sanger sequencing (Robison, 2010). Each of the SGS technologies mentioned in the previous section has its own distinct error profile. Roche GS FLX technology has trouble in correctly interpreting homopolymer runs of nucleotides: it often deletes or inserts bases from the sequence output in these regions (Mardis, 2008). On the other hand, the

Illumina sequencing technology has a tendency to substitute C with A and G with T. Furthermore, the bases towards the 3' end of Illumina sequence reads have a lower base quality (Erlich et al., 2009). Sequence reads produced using ABI SOLiD sequencing technology has a similar lower quality bases towards the 3' end of the reads (Flicek and Birney, 2009). ABI SOLiD and Illumina sequencing technologies also share the trend to produce low sequence coverage of AT-rich repetitive sequences (Harismendy et al., 2009). The Ion Torrent personal genome machine (PGM) has insertion/deletion (indel) error types which are caused by incorrect flow calls. The reference genome comes in FASTA format which does not contain any quality information. However SGS data comes in FASTQ format which contains import base quality information. Both, FASTA and FASTQ, formats are described below.

1.3.1.1. FASTA format

The FASTA format was developed for and named after a biological sequence comparison algorithm (Pearson and Lipman, 1988). Until today, the FASTA format is widely used to store either DNA or amino acid sequences. Even reference genomes are stored in this format.

FASTA files are stored in plain text and start always with a header line followed by one or more sequence lines. Header lines always begin with the symbol ">" followed by text which describes the sequence lines directly below. It is recommended that all lines are shorter than 80 characters. When the length of the sequence lines is longer than 80 characters then the sequences are split across multiple lines. However, header line is never split in multiple lines. FASTA files which contains more than one header and sequence are named as multiple FASTA files**.** Figure 1.7 below gives an example of a multiple FASTA file.

```
>seq1
GTACGACGGAGTGTTATAAGATGGGAAATCGGATACCAGATGAAATTGTGGATCAGGTGCAAAAGTCGGC
AGATATCGTTGAAGTCATAGGTGATTATGTTCAATTAAAGAAGCAAGGCCGAAACTACTTTGGACTCTGT
CCTTTTCATGGAGAAAGCACACCTTCGTTTTCCGTATCGCCCGACAAACAGATTTTTCATTGCTTTGGCT
>seq2
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
```

Figure 1.7: An example of a multiple FASTA file format

1.3.1.2. FASTQ format and visualistaion of sequencing errors

The Wellcome Trust Sanger Institute invented a modified version of the standard FASTA format, FASTQ, to store sequenced reads together with the quality. Both, FASTA and FASTQ, are storing the sequence data as plain text. Each read entry contains four lines in FASTQ. The first line starts with "@"and is used as record identifier. The second line contains the read sequence. The third line contains "+" to signal the end of the read sequence. The last line contains the quality of the sequenced read. An example of this format is given in Figure 1.8 below.

```
@HWUSI-EAS762_0026:3:120:16408:21310#0/1
TAGGAGTTGGGATGAAGAAGTTATCCCAGTTTCAANNCAGGNGATTCCAGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+HWUSI-EAS762_0026:3:120:16408:21310#0/1
ffef^ggea_faRfcfddffffWffggfggcdaacBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWUSI-EAS762_0026:3:120:16469:21305#0/1
CTCGTATACTCCCACTTAGAAAAATCTTCATTATTGTAATCGAGTTTTTAGANNNNNNNNNNNNNNNNNNNNNNNNNNNN
+HWUSI-EAS762_0026:3:120:16469:21305#0/1
hhhhehghegfhhhheha[\ffff`efhahhhffh_fffffdhWeffff_BBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Figure 1.8: An example of a multiple FASTAQ file format

During sequencing process a program, called Phred, detects when a base may be wrong, and save it as a quality score. Phred quality score was developed by (Ewing et al., 1998) for Sanger sequences and later applied to SGS sequences. Phred scores are defined by the following formula:

$$q = -10 \ \log_{10} P$$

Phred quality score *q* is defined as a property which is logarithmically related to the base-calling error probabilities *p*. Thus a one percent error rate (*p* = 0.01) corresponds to a recorded quality score of 20. The value of the quality score is typically encoded as a string of single ASCII characters. One ASCII character for each base in the sequence.

Table 1.2 show the variations between these formats which exist in the relationship between ASCII characters and Phred quality scores (Cock et al., 2010). Unfortunately, FASTQ does not contain any information about which format was used for the quality score. In order to be able to distinguish between these three formats a number of tools have been developed. One such tool is FastQC which is described below.

Table 1.2: The four described FASTQ variants whereas the Illumina 1.8+ is the same as Sanger. Other columns are the range of ASCII characters permitted in the quality string and ASCII encoding offset. The last column describes possible range of scores (Cock et al., 2010).

Description	ASCII characters		Quality score
	Range	Offset	Range
Sanger	33 - 126	33	0 - 93
Solexa/Illumina	59 - 126	64	-67
Illumina 1.3+	64 - 126	64	0 - 62
Illumina 1.8+ = Sanger	33 - 126	33	0 - 93

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a quality control tool for high throughput sequence data which provides a quality control report which can find problems which originate either in the sequencer or in the starting library material. These problems could affect further analysis. Furthermore, it also provides box plot of average quality score versus read position which gives an idea of the overall quality of a sequencing run (see Figure 1.9). FastQC runs as a standalone interactive application for analysis of small numbers of FastQ files. However, it also runs in a non-interactive mode for the processing of large numbers of files.

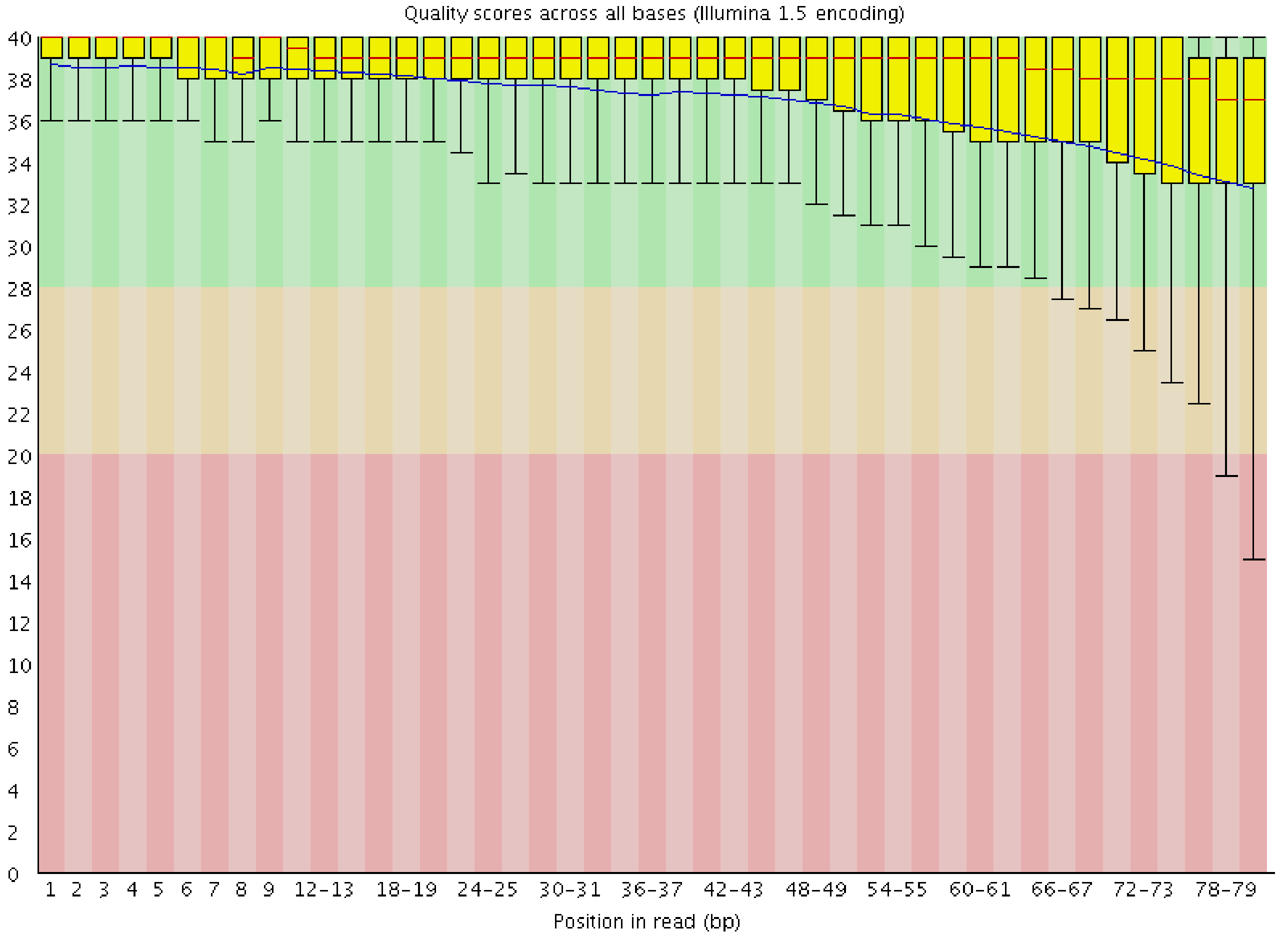


Figure 1.9 shows a box plot of read quality versus base positiona for 100bp reads. All SGS platforms show an increase in the error rate towards the ends of the reads.

1.3.2. Insert size improves accuracy of alignment

The alignment of short reads to a reference is difficult because the reference genome is often extremely large, for example the *B. napus* genome is predicted to be 1235 Mbp and contains many repetitive regions (Arumuganathan and Earle, 1991). Short reads can also have sequencing errors and may have diverged from the reference genome. During the sequencing process the DNA is randomly cut in pieces and adaptors are attached. Illumina provides three options; single-end, paired-end or mate-pair sequencing. The last two sequencing methods provide an insert size which is a distance between two reads and are described more in detail below.

Due to the short length of the reads, one read could match at many positions, but two reads separated by a gap of defined insert size provides a greater confidence of specific and accurate read mapping (Robison, 2010).

1.3.2.1. Paired-end sequencing

In single-end (SE) sequencing, only one end of a DNA fragment is sequenced, but in the paired-end (PE) process both ends of the same DNA fragment are sequenced, producing two reads called A and B. PE reads are oriented towards each other (=>.....<=) and the length of A+B is usually shorter than the DNA fragment and therefore there is a gap between A and B which is called the insert size. Unfortunately, the sequence of the DNA fragment in the gap is unknown. However, the orientation and approximate distance of A and B are known which is helpful in the downstream analysis, such as aligning the reads to a reference genome, because one of the reads is more likely outside of the repeat. Therefore a 2x100 base paired-end read with a 600 base insert size is better than a single 200 base read. Figure 1.10 shows a frequency plot of insert sizes for Illumina PE reads of *B. napus* cv. Skipton that have been aligned onto a reference sequence (chloroplast).

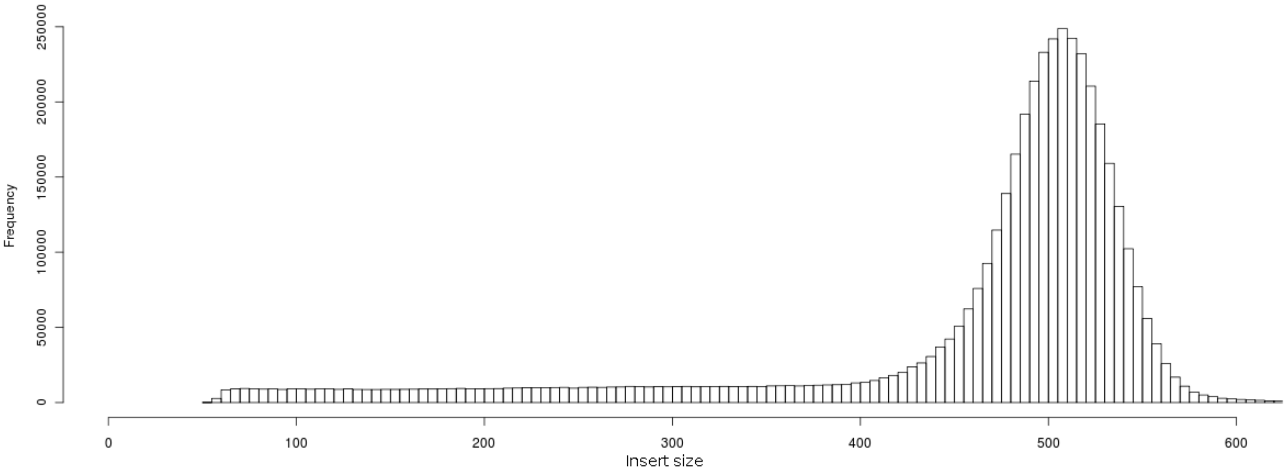


Figure 1.10: The distribution of insert sizes for an paired-end-read library of *B. napus* cv. *Skipton*

1.3.2.2. Mate-pair sequencing

Mate-pair (MP) and paired-end (PE) sequencing have two differences. The first one is that the MP reads have a larger distance between them (insert size) compare to PE reads. The second difference is that MP reads are oriented away from each other (<=.....=>) whereas PE reads are oriented towards each other (=>.....<=). However, MP libraries are frequently contaminated with PE reads (the so-called shadow library) which occur frequently in the preparation process for MP libraries. The impact on the data is that for example, a library produced with insert size of 4000 bp will mainly contain MP reads with an insert size around 4000 bp between them, but will also contain some PE reads with an insert size of 300 bp (http://sequencetagdb.info/tagdb/cgi-bin/help, 21 January 2015). Figure 1.11 shows a frequency plot of insert sizes for Illumina MP reads reads of *B. napus* cv. Skipton that have been aligned onto a reference sequence (chloroplast)..

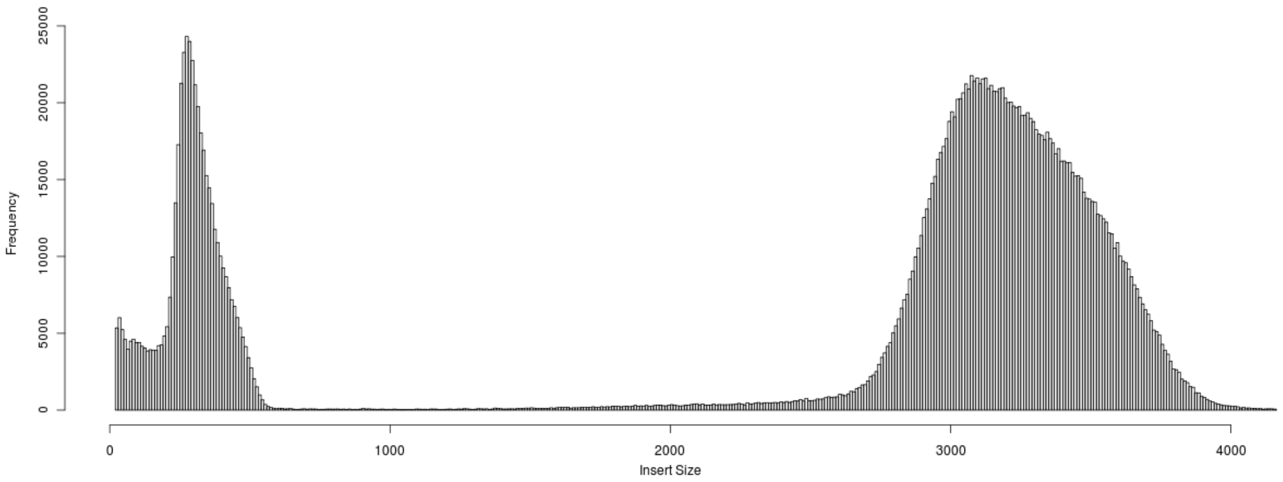


Figure 1.11: The distribution of insert sizes for a mate-pair-read library of *B. napus* cv. Skipton

1.3.3. Aligner and file formats to store the mapping

To be able to align short reads to the reference genome it is important choose an efficient and accurate aligner which can handle and analyse large-scale sequence data produced by SGS. However, to keep the aligner fast, memory-efficient and able to handle increasing read length, which happens almost every six months, aligners look for similar matches and not exact matches to the reference genome. Aligner algorithms can be roughly categorised into categories, as being based on hash tables or FM-index.

1.3.3.1. Hash table based aligner

All hash table indexing algorithms are based on the idea of Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990, Altschul et al., 1997, Li and Homer, 2010). BLAST is a fast comparison tool for biological sequences. It allows the comparison of a query sequence with a database of sequences and identifies database sequences that are similar to the query sequence. Different types of BLAST tools are available (see Table 1.3) for aligning different combinations of DNA, cDNA and protein sequence data (Altschul et al., 1990).

Table 1.3: A table of BLAST-derived programs, as featured in NCBI-BLAST.

Program	Query sequence	Subject sequence/database
blastp	Protein	Protein
blastn	Nucleotide	Nucleotide
blastx	Nucleotide	Protein
tblastn	Protein	Nucleotide
tblastx	Nucleotide	Nucleotide

The first step in the BLAST implementation is to use fast seeds detection. A hash table stores the k-mers of a specified size (word size) of the query sequence as the keys and their positions as values, and this is then searched through the database sequences. In the second step BLAST subsequently extends and joins the seeds with slower and more accurate dynamic programming Smith-Waterman (Farrar, 2007) algorithm (Shang et al., 2014). BLAST outputs statistically significant local alignments which can be can be controlled by an e-value parameter. A hit associated with an e-value of 1 means that in a database can be expected to see 1 match with a similar score simply by chance. The lower the e-value, the more significant the match is (http://blast.ncbi.nlm.nih.gov/, 24 January 2015).

BLAST compares all positions within a window, whereas SGS aligners use spaced seeds in order to improve the sensitivity of the alignment. However, spaced seeds use multiple windows in which positions can differ from the reference sequence. In order to cover all different permutations of match and mismatch positions, multiple seed masks are necessary. For example, BFAST (Homer et al., 2009) uses empirically derived optimal seed masks for given read and genome sizes (Lindner and Friedel, 2012). Other spaced seed aligner are GNUMAP (Clement et al., 2010), MAQ (Li et al., 2008), MapReduce (Schatz, 2009), PerM (Chen et al., 2009b), RMAP (Smith et al., 2009), SeqMap (Jiang and Wong, 2008), and (Lin et al., 2008). Since spaced seed approach does not allow gapped alignment, other aligners have been developed to support gapped alignments, usually after seed extension, including AGILE (Misra et al., 2011), BLAT (Kent, 2002), RazerS (Weese et al., 2009), SHRiMP (Rumble et al., 2009) and SSAHA (Ning et al., 2001).

1.3.3.2. FM-index based aligner

In bioinformatics with SGS data, and in web information retrieval, it is important to be able to index large sequences or texts for inexact pattern matching and only allow limited amount of mismatches while searching. (Policriti and Prezza, 2014). The disadvantage of using hash table index is that an alignment must be performed for each copy of the repetitive DNA sequence (Li and Homer, 2010). On the other hand, suffix array or suffix tree are the most suitable data structures for indexing DNA sequence, because only one alignment is required for repetitive sequences in the reference sequence (Li et al., 2009b). The drawback of using suffix array or suffix tree is the large memory requirement for the uncompressed data structures. In case of suffix tree it is 15-20 bytes per base of the reference (Kurtz et al., 2004) and for suffix array it is 10 bytes per base (Abouelhoda et al., 2004).

Ferragina and Manzini (Ferragina and Manzini, 2000) developed a FM-index which is compressed suffix array created from the Burrows Wheeler transformation (BWT) (Burrows and Wheeler, 1994) sequence rather than from the original reference sequence. BWT places the same bases side by side as a cluster and through this compression the FM-index uses only 0.5-2 bytes per base (Li and Homer, 2010) and is faster than their hash-based alternatives at the same sensitivity level (Flicek and Birney, 2009). Popular

FM-index based aligners for SGS such as Bowtie, BWA and SOAP2 (Yu et al., 2012) use less memory and achieve high mapping speed through some reduction in mapping sensitivity compare to hash table based methods (Nielsen et al., 2011).

The downside of FM-index is that it only provides support for exact string matching (Policriti and Prezza, 2014). For example to align SGS reads it is necessary to support inexact match search in order to be able to deal with mismatches caused by sequencing errors and differences between reference and query organisms (Langmead et al., 2009).
In order to be able to support inexact search, additional space efficient strategies such as backtracking or split-read strategy are used. The disadvantage of backtracking strategy is that query times rapidly grow exponentially and therefore it is not suitable for large patterns and numbers of errors. On the other hand the split-read strategy based technique does not suffer this exponential growth but it can be only used with a small number of errors, because split-read strategy are searched without errors (Policriti and Prezza, 2014). Bowtie and BWA use a backtracking strategy on the FM-index to search for inexact matches (Yu et al., 2012) and SOAP2 uses a split-read strategy on the FM index (Policriti and Prezza, 2014).

Bowtie and BWA use a quality-aware backtracking algorithm to search for inexact matches. Both aligners perform a depth-first search until they find alignments that satisfy a specified alignment criterion. These criterions allow a limited number of mismatches and alignments where the sum of the PHRED score at all mismatched positions are low. The higher the PHRED score, the more accurate an alignment is. Bowtie did not implement support for paired-end alignment (Langmead et al., 2009, Yu et al., 2012). However, BWA supports paired-end mapping in two steps. In the first step, it finds the positions of all the good hits, sorts them according to the chromosomal coordinates. Finally, it scans through all the potential hits to pair the two ends together (Li and Durbin, 2009b).

SOAP2 uses a split-read strategy to allow maximum two mismatches. A read is split into two fragments in order to allow only one mismatch. This mismatch can only exist in one of the two fragments. In order to allow two mismatches the read has to be split into three fragments, such that the mismatches can only exist in two of the three fragments. Paired-end reads in SOAP2 are aligning in two steps. Firstly, the two reads belonging to a pair are aligned independently. In the second step, SOAP2 searches for the pair of hits with the proper distance and correct orientation relationship. SOAP2 chooses the best hit of each read or read pair, which has small gaps or the lowest number of mismatches (Li et al., 2009b).

In order to find out which aligner performs best, a sequencing simulation and alignment evaluation software, Seal (SEquence ALignment evaluation suite, http://compbio.case.edu/seal/), has been developed. The developers compared the performance of Bowtie, BWA and SOAP2 with regard to accuracy and runtime. All three aligners have in common to build an index of a genome slowly, but to align the reads to the genome is very fast. However, the index can be reused for other reads and therefore it is not a bottleneck to build an index. Bowtie and BWA align many incorrect reads, because their algorithm tries not to miss any potential mappings. On the other hand, SOAP2 mapping accuracy is quite high even in high error reads which is useful for genotyping SNPs (Ruffalo et al., 2011).

1.3.3.3. Most popular file formats to store alignments

The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments with reference sequences, supporting single- and paired-end reads and combining reads produced by different sequencing platforms. All lines are TAB delimited. SAM format contains one header section and one alignment section. The lines in the header section start with character '@SQ' and lines in the alignment section do not. The header section represents the order of reference sequences.

Binary Alignment/Map (BAM) is the equivalent binary representation of SAM and keeps exactly the same information as SAM. BAM is compressed by the BGZF library, which is part of SAMtools, and is zlib-compatible. To achieve fast random access of alignments overlapping a specified chromosomal region a BAM file has to be sorted by coordinate and then indexed by SAMtools. Using positional sorting and indexing, applications can access a specific genomic region without loading the entire file into memory.

SAMtools is a library and software package for parsing and manipulating SAM/BAM files. It supports sort, index and merge alignments, removes PCR duplicates and generates per-position information in the pileup format and alignment viewer. SAMtools is implemented in both C and in Java, with slightly different functionality. Both are open-source and there are Perl and Python bindings to the C library. This has significantly improved the interoperability of SGS tools for alignment, visualisation and variant calling (Li et al., 2009a).

1.3.4. Duplicate removal tool

Using programs such as MagicViewer (Chapter 1.7.3) or Tablet (Chapter 1.7.2) (Milne et al., 2013) it is possible to see in the alignment that there are many exact duplicates of a read which share the same alignment position. These duplicates were created from artefacts during PCR amplification and sequencing. Keeping them would give an uneven representation of that molecule compared to other molecules and could bias the SNP calling. Therefore duplicates should be removed from the alignment using Picard tools' MarkDuplicates (http://www.cbs.dtu.dk/courses/27626/Exercises/BAM-postprocessing.php)

Picard provides Java-based command-line tools to manipulate SAM/BAM files (http://picard.sourceforge.net/). One of its tools, MarkDuplicates is able to detect duplicates in BAM files. It defines two pairs as duplicates if they align at the same position, both for their first and second reads. Only one of the duplicate read pairs with the highest average base quality is kept and the rest are deleted as duplicates using the option REMOVE_DUPLICATES=true (Pireddu et al., 2011).

1.4. Molecular genetic markers

Genetic variation in species can increase the capability of an organism to adapt to a changing environment, which helps the survival of the species. In research, the genetic variation helps in the understanding of evolution, genetic improvements and management of natural resources (Chauhan and Kumar, 2010). By introducing new and favorable traits from wild germplasm, new sources of genetic variation can be created. Molecular genetic markers can aid in understanding the genetic variation in order to improve the species. In crop research, the aim is to improve crop productivity and be able to grow crop species in more difficult climatic environments. To do these, plant breeders need to have enough diversity available to allow the production of new varieties. Genome based markers have advantages over phenotypic markers in that they are not affected by the environment, relatively easy to assay and are highly heritable. Restriction Fragment Length Polymorphism (RFLP) markers were initially used in crop plants, followed by Amplified Fragment Length Polymorphisms (AFLPs). For major crop plants many Simple Sequence Repeats (SSR) markers were used (Korzun, 2002). However, development of high-throughput genotyping with Single-Nucleotide Polymorphism (SNP) markers and linked diagnostic markers is now used for more effective molecular breeding and they are opening opportunities for genomic selection (Randhawa et al., 2013).

1.4.1. RFLP: Restriction Fragment Length Polymorphism

Restriction Fragment Length Polymorphism (RFLP) is a method that identifies variations in DNA sequences. In the first step a restriction enzyme digests the DNA sequence into fragments. Gel electrophoresis is used to separate the fragments according to their lengths. In the third step the results from gel electrophoresis are transferred to a membrane via the Southern blot method. A RFLP probe is a labelled DNA sequence that hybridises with one or more fragments of the digested DNA sample, and the hybridisation pattern reveals when a marker is polymorphic between individuals. Each fragment length is an allele and can be used in genetic analysis (Waikan and Dozy, 1978). RFLPs became obsolete because of the introduction of PCR based technologies.

1.4.2. Amplified Fragment Length Polymorphisms

Amplified Fragment Length Polymorphisms (AFLPs) are markers based on the selective amplification by PCR of fragments of genomic DNA (Vos et al., 1995). Firstly, it uses restriction enzymes to digest genomic DNA, followed by ligation of adaptors to the sticky ends of the restriction fragments. Secondly, a subset of the restriction fragments are selected to be amplified with two PCR primers that have corresponding adaptor and restriction site specific sequences. Finally, the amplified fragments are separated and visualised using gel electrophoresis techniques. AFLPs are highly sensitive for detecting polymorphisms in DNA (Mueller and Wolfenbarger, 1999), without the need for a reference sequence, and without the cost of marker discovery, such as for SNPs and SSRs. AFLPs have the disadvantage that they are anonymous and therefore have no genome information. AFLPs have been widely used for the identification of genetic variation in between varieties or closely related species of *Brassica* (Zhao et al., 2005).

1.4.3. Microsatellites/Simple Sequence Repeats

Microsatellites or Simple Sequence Repeats (SSRs) are short tandem repeat sequences of DNA (Powell et al., 1995, Turnpenny and Ellard, 2011). The repeats usually have two, three or four nucleotides (di-, tri-, and tetranucleotide repeats respectively), and can be repeated 3 to >100 times (Whittaker et al., 2003). Dinucleotide repeats are the most common SSRs, followed by tri-, tetra-, and penta-nucleotide repeats (Hamarsheh and Amro, 2011). Most polymorphic SSRs are in intergenic regions, but some of them can be found in genes, these SSRs are generally less polymorphic. As there are sometimes several alleles present at an SSR locus, genotypes within pedigrees are often fully informative, in that the progenitor of a particular allele can often be identified. Therefore SSRs can be used for determining paternity, population genetic studies and recombination mapping. Regions flanking SSRs have an increased density of SNPs (Varela and Amos, 2010). However, SSRs are frequently not identified from Second Generating Sequencing data, because the read length is to short, but this might change with Third Generation Sequencing data which provide longer reads (Edwards and Gupta, 2013).

1.4.4. Single Nucleotide Polymorphisms

Single Nucleotide Polymorphisms (SNPs) represent a single base change between two individuals at a defined position (see Figure 1.12). SNPs have advantages over SSRs in terms of cheap automated high-throughput genotyping, they are less informative due to their predominantly bi-allelic nature. SNPs appear in two different forms: transitions (purine/purine or pyrimidine/pyrimidine; C/T or G/A) or transversions (purine/pyrimidine; C/G, A/T, C/A, or T/G) (see Figure 1.13). At any position a SNP could be bi-, tri- or tetra-allelic, however in practice most SNPs are biallelic (Doveri et al., 2008). SNPs can be put in the following categories: intravarietal SNPs are differences between gene family members and homoeologues within a line, whilst varietal SNPs are differences between two varieties and can therefore be used as molecular markers (Barker and Edwards, 2009).

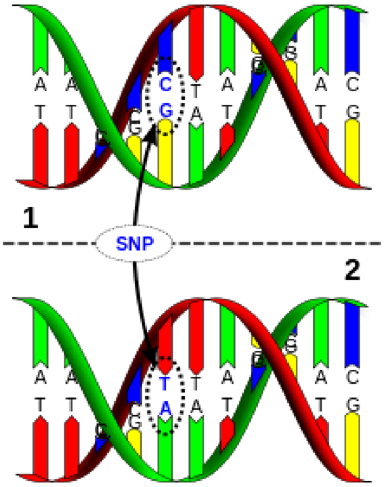


Figure 1.12: DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism) (adapted from http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism).

A synonymous SNP is a SNP that does not change the amino acid in the protein, whereas a non-synonymous SNP does. The genome-wide normalized ratio ω = N/S = non-synonymous SNPs/synonymous SNPs, is by definition normalized to 1 in most evolutionary studies (Stoletzki and Eyre-Walker, 2011). A higher N/S ratio near the telomeres and centromeres and lower N/S ratios in the middle of the chromosome arms might be observed (Begun and Aquadro, 1992, Charlesworth et al., 1987).

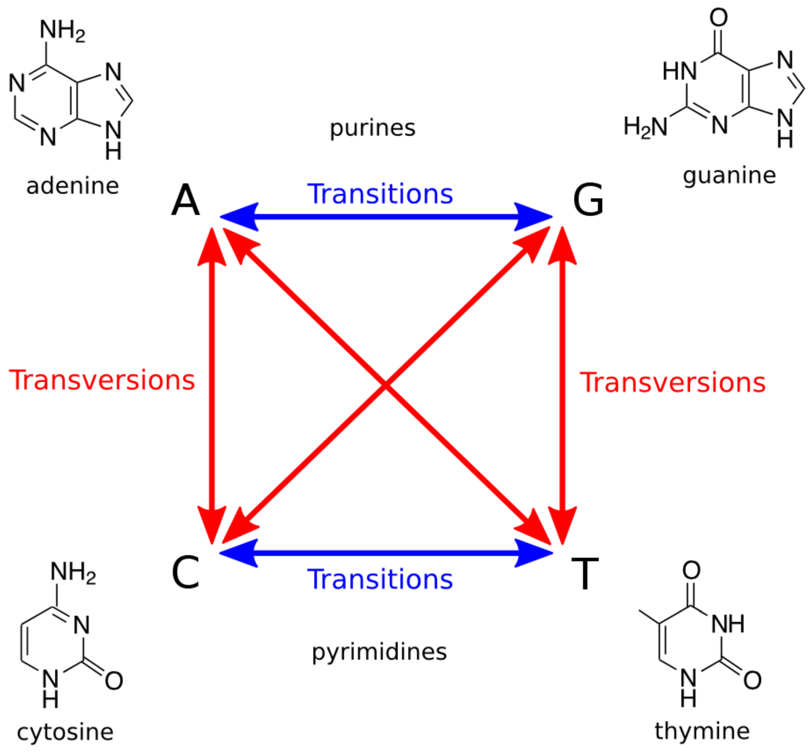


Figure 1.13: The difference between transitional and tranversional nucleotide changes (adapted from http://en.wikipedia.org/wiki/Transversion).

The low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen, 2001). SNPs represent the most frequent type of genetic polymorphism and provide a high density of markers near a locus of interest. With the introduction of Second Generation Sequencing (SGS) the cost of SNP discovery and genotyping has dropped significantly.

The development of high-throughput methods for the detection of SNPs has led to a revolution in their use as molecular markers (Duran et al., 2010b, Gupta, 2008, Rafalski, 2002, Varshney et al., 2009). SNPs may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence, are generally abundant in populations and have a low mutation rate (Edwards et al., 2007a). The principal challenge in SNP discovery remains the discrimination between true genetic polymorphisms and the often more abundant sequence or read mapping errors. SNP discovery is further confounded in polyploid species where multiple related genomes are present within each nucleus. The identification of high confidence SNPs can be based on three methods:

sequence quality score, redundancy of the polymorphism in a sequence alignment and presence of conserved haplotypes at a locus (Barker et al., 2003, Edwards et al., 2007b). SNP redundancy provides an effective means for estimating confidence in the validity of SNPs independently of sequence quality scores and has been demonstrated to be an accurate method for SNP discovery in a range of species (Batley and Edwards, 2009b, Duran et al., 2009a, Duran et al., 2009b). SNPs are used routinely in agriculture as markers in crop and livestock breeding programs, e.g. for phylogenetic analysis, cultivar identification, genetic diversity analysis, characterisation of genetic resources and association with agronomic traits (Batley and Edwards, 2009b).

1.5. In-silico Single Nucleotide Polymorphisms discovery and file formats

Sequencing data contains errors as frequent as one error every one hundred base pairs. This incorrect base could be called a SNP in some prediction software, but it does not reflect biologically relevant polymorphisms. Formats were created for different visualisation tools in order to view the newly discovered SNPs together with the alignment.

1.5.1. ACCUSA: accurate Single Nucleotide Polymorphisms calling

ACCUSA is a SNP caller which considers both the read qualities as well as the reference genome quality. Therefore it is suited for SNP discovery from genome projects in draft status. ACCUSA accepts ACE file format (http://bozeman.mbt.washington.edu/consed/distributions/README.16.0.txt), as well as the SAMtools pileup format (Li, 2011a) as input files. The problem with pileup format is that these files are huge, because this format contains all base differences at each position compared to the reference. The reference genome must be in FASTQ format, which contains the reference base quality. ACCUSA uses Bayesian analysis to compute the probability of a SNP for all aligned short reads at a given genome assembly position and for the complete alignment column including the reference base (Frohler and Dieterich, 2010).

1.5.2. AGSNP: an annotation-based, genome-wide Single Nucleotide Polymorphisms discovery pipeline

AGSNP is an annotation-based, genome-wide SNP discovery pipeline using NGS data for large and complex genomes without a reference genome sequence. Shotgun reads of one individual are annotated in order to distinguish single-copy sequences and repeat junctions with RJPrimer (You et al., 2010). Multiple genome equivalents of shotgun reads of another individual are then mapped to the annotated reads using BWA (Li and Durbin, 2009a) in order to identify putative SNPs with SAMTools (You et al., 2011). AGSNP then filters the SAMtools pileup file to increase the accuracy of putative SNPs. Furthermore, AGSNP creates validation files for Illumina's GoldenGate or Infinium assays which require a minimum of 50 bp (60 bp preferred) of sequence on either side of each SNP and a minimum of 60 bp between two contiguous SNPs. In an example of the use of AGSNP, genomic DNA and cDNA of *Ae. tauschii* accession *AS75,* as well genomic DNA of *Ae. tauschii* accession *AL8/78* were used. In a sample of 302 randomly selected putative SNPs, 84% in gene regions, 88% in repeat junctions, and 81% in uncharacterised regions were validated. The AGSNP pipeline package is available upon request (You et al., 2011).

1.5.3. NGS-SNP: Next-Generation-Sequencing - Single Nucleotide Polymorphisms

NGS-SNP (Next-Generation Sequencing SNP) is a collection of command-line Perl scripts for performing in-depth/rich annotation of SNPs using Maq (Li et al., 2008) or SAMtools (Li, 2011a) as SNP discovery programs. Both SNP callers require a reference sequence. NGS-SNP works with SNPs which were identified by the sequencing of whole genomes from any organism with a reference sequence in Ensembl and also uses NCBI Entrez Gene (Maglott et al., 2011) and UniProt (Apweiler et al., 2013) as additional information sources. SNPs are classified as synonymous, non-synonymous, 3' UTR, etc. regardless of whether or not they match existing SNP records.

NGS-SNP compares SNP positions to orthologous sequences to help to identify SNPs that affect conserved residues, or alter residues or genes linked to phenotypes in another species. This tool reports overlapping protein features or domains, provides gene ontology information, or provides flanking sequence for use in the design of validation assays. Known SNP sites in the flanking sequence and at the SNP position can be included in the output as lower case IUPAC characters, and as potentially additional alleles at the SNP site. It also maps SNP-altered residues to a protein in another species to retrieve additional information.

1.5.4. Atlas-SNP2: Atlas-Single Nucleotide Polymorphisms2

Atlas-SNP2 is based on three steps. Firstly, to reduce the computational requirements, it divides the reference genome into smaller pieces and separates NGS reads into smaller batches, each with fewer reads. Secondly, the NGS reads are anchored and aligned onto the reference sequence using BLAT (Kent, 2002) and Cross_Match (http://www.phrap.org). Reads which have multiple best hits are discarded in order to avoid mis-mapping of repeats and also to remove duplicated reads. In the last step Atlas-SNP2 predicts error probabilities of mismatches in single reads using a logistic regression model followed by a Bayesian formula to combine the likelihood estimation from multiple reads mapped to the same locus with prior SNP probabilities. The estimated posterior SNP probability is used to distinguish true SNPs from sequencing errors (Shen et al., 2010).

1.5.5. Popular file formats for Single Nucleotide Polymorphisms

1.5.5.1. GFF3: Generic Feature Format version 3

Generic Feature Format version 3 (GFF3) is based on the Generic Feature Format (GFF) and both were designed to describe genome annotation data, but GFF3, unlike GFF, is typed using a Sequence Ontology (SO). This means that the terminology being used to describe the data is standardised and organised by pre-specified relationships. This ontology was developed by the Gene Ontology Consortium (Ashburner et al., 2000) to describe the parts of genomic annotations, and how these parts relate to each other (Eilbeck et al., 2005). GFF, as well as GFF3, are tab-delimited flat file formats, which can be easily modified with a text editor and processed with shell tools such as grep (http://www.sequenceontology.org/resources/gff3.html, 24.08.2011) (Reese et al., 2010).

1.5.5.2. VCF: Variant Call Format

The variant call format (VCF) is a standardised generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations (Danecek et al., 2011). Meta-information within a VCF file provides information about the file creation, version of the reference sequence and software used. Although the generic feature format (GFF) has been extended to standardise storage of variant information in genome variant format (GVF) (Reese et al., 2010), this is not suitable for storing information across many samples. VCF files can be compressed by bgzip, a program which uses the zlib-compatible BGZF library (Li et al., 2009a). Fast data retrieval can be achieved by indexing the genomic position using tabix (Li, 2011b), a generic indexer for TAB-delimited files. Bgzip and tabix, are part of the SAMtools package (Li et al., 2009a).

VCFtools is an open-source software suite which is split into two modules. The first module provides a Perl API, and allows format validation, merging, comparing, intersecting, and making complements and basic overall statistics on VCF files. The second module is written in C++ and is used to analyse SNP data in VCF format, allowing the user to estimate allele frequencies, levels of linkage disequilibrium and various quality control metrics. An alternative tool for VCF generation and manipulation is the GATK toolkit (McKenna et al., 2010)

1.6. Single Nucleotide Polymorphisms annotation

SnpEff (SNP effect) is a platform independent open source variant effect predictor program. It annotates variants and predicts the effects of genetic variations, such as SNPs, insertions and deletions (INDELs) and multiple nucleotide polymorphisms (MNPs), based on gene annotations (Cingolani et al., 2012).

Firstly, SnpEff requires a database with gene annotation information and currently SnpEff contains over 320 databases for different reference genome versions that can be analysed. If a database build is not available then it can be built using a reference genome in FASTA format and an annotation file format such as GTF, GFF or RefSeq table. However, if annotations are not available e. g. from ENSEMBL, UCSC Genome Bioinformatics website or other specific websites, such as TAIR then gene prediction tools such as SNAP (Korf, 2004) or GlimmerHMM (Majoros et al., 2004) could be used to generate gene annotations.

Secondly, the potential effect of a SNP can be calculated with data in variant call format (VCF), which contains all SNPs and INDELs in a genome. Each variant queries the data structure to find and report all intersecting genomic regions. Whenever the regions include an exon, the coding effect of the variant is calculated. In VCF format, SNPeff stores the effect information in the information (INFO) fields using an effect (EFF) tag.

1.7. Data visualisation and data growth

In the last few years, researchers have become inundated with new exponentially growing sequencing data (see Figure 1.14), thanks to SGS technologies. Moore's Law dictates that computer technology improvements double every two years (Moore, 1998). However, DNA sequencing overtook Moore's Law, getting cheaper and faster than expected. The graph in Figure 1.14 shows data from 2001 to October 2007 representing the costs of generating DNA sequence using Sanger sequencing (First Generation Sequencing). From January 2008, the data represents the costs of sequencing using SGS.

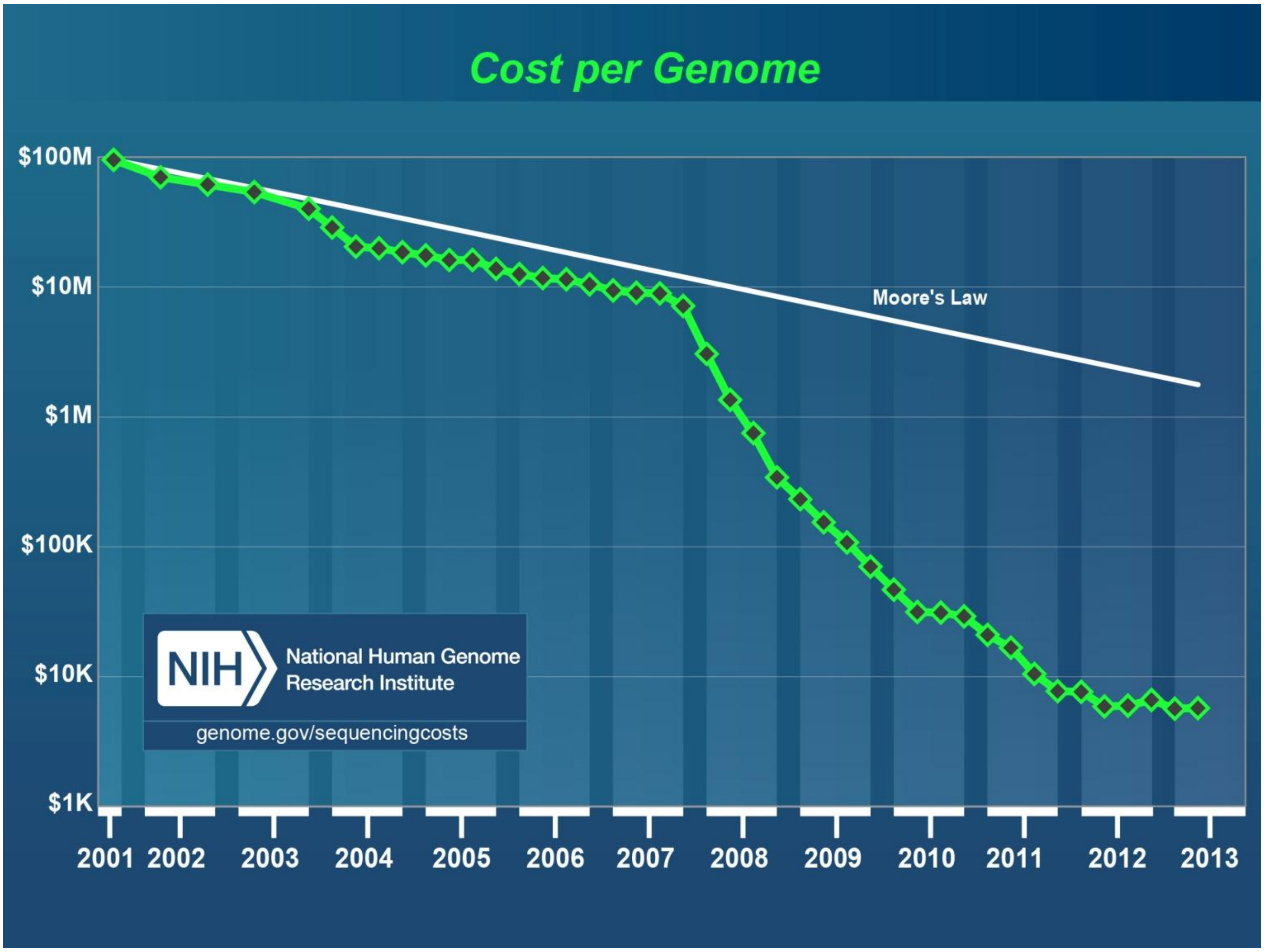


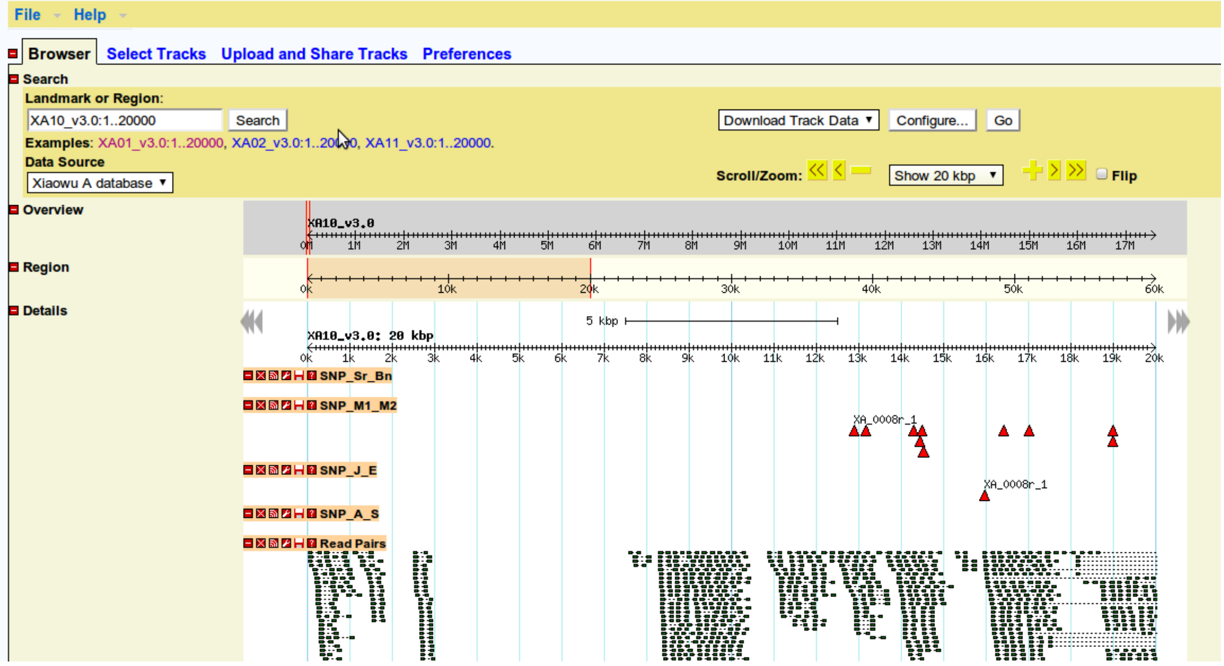
Figure 1.14: Cost of sequencing a human-sized genome. The cost of getting DNA data is dropping faster than the cost of processing data on computers (adapted from http://www.genome.gov/sequencingcosts/).

Therefore, the field of biological sequence data visualisation is a rapidly expanding field that is required to address new tasks in order to cope efficiently with the vast amounts of data produced (Pavlopoulos et al., 2013). Continued improvements in tools and adapting new hardware technologies will help researchers make sense of large volumes of data.

1.7.1. Generic Genome Browser

Generic Genome browser (GBrowse or GGB) 2 (see Figure 1.15) gives users the ability to navigate genomic sequence information and visualise various features in a series of tracks within the context of a reference sequence. It is implemented in Perl as a series of CGI scripts and designed to display genome annotations on small or large genomes. GBrowse can either use a file as a relational database through use of special functions called adaptors, or connect to a database. There are several new major features in GBrowse2 over GBrowse (Donlin, 2009) including:

- User accounts with authentication
- User interface improvements through AJAX which avoids reloading the whole page to view a new region or data track.
- Multiple processors and machines can be used to render data tracks in parallel
- It allows tracks to come from different data sources and multiple servers in the same page e.g. one track could come from a database and the other from a BAM file.


Landmark or Region: XA10_v3.0:1..20000

Examples: XA01_v3.0:1..20000, XA02_v3.0:1..20000, XA11_v3.0:1..20000.

Data Source: Xiaowu A database

Scroll/Zoom: Show 20 kbp

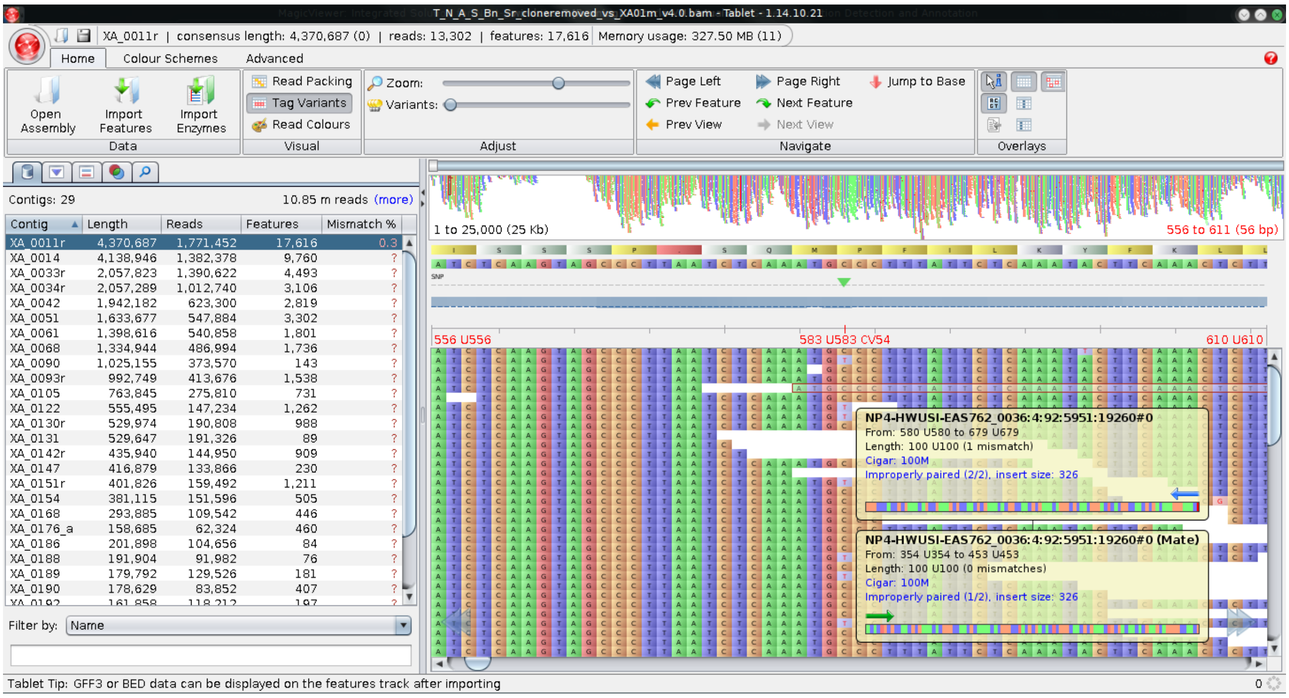
Figure 1.15: GGB2 showing *Brassica napus* AA genome SNPs called by SGSautoSNP.

GBrowse2 also contains a *Bio::DB::Sam* adaptor to visualise BAM SGS short read data alignments. Overall, Gbrowse2 allows displaying SGS data along with other annotations. Gbrowse was implemented as open-source by the project called the Generic Model Organism Database and runs on LINUX/UNIX, Mac OSX, and Windows.

1.7.2. Tablet

Tablet (see Figure 1.16) is a free graphical viewer for SGS assemblies and alignments. It provides high-quality visualisations showing data in packed (showing as many reads per line as possible without overlap) or stacked (showing one read per line) views, allowing navigation to any region of interest, whole contig overviews and data summaries. Tablet can import data from ACE, AFG, MAQ, SAM/BAM and SOAP. The latter four formats require the reference sequence to be imported separately as a FASTA file. It is also possible to import annotation features such as SNPs and indels in GFF3 format. This tool is written in Java so it can run platform independently in both 32- and 64-bit versions. It supports multi-core processor architectures to allow fast navigation of NGS data with low memory usage. Tablet has been implemented as a hybrid system that provides the advantages of memory-based (where all the data are loaded into memory) and disk cached data (only the visible segment of the dataset is loaded in the memory, and the remainder are stored on the disk).

Memory-based applications are faster for viewing and navigation, but the amount of NGS data cannot be stored in a normal desktop computer's memory. Cache-based applications can be used for NGS data because of using a minimum of memory, but data access is slower which affects the navigation (Milne et al., 2010a).



T_N_A_S_Bn_Sr_cloneremoved_vs_XA01m_v4.0.bam - Tablet - 1.14.10.21

XA_0011r | consensus length: 4,370,687 (0) | reads: 13,302 | features: 17,616 | Memory usage: 327.50 MB (11)

Contigs: 29 10.85 m reads (more)

Contig	Length	Reads	Features	Mismatch %
XA_0011r	4,370,687	1,771,452	17,616	0.3
XA_0014	4,138,946	1,382,378	9,760	?
XA_0033r	2,057,823	1,390,622	4,493	?
XA_0034r	2,057,289	1,012,740	3,106	?
XA_0042	1,942,182	623,300	2,819	?
XA_0051	1,633,677	547,884	3,302	?
XA_0061	1,398,616	540,858	1,801	?
XA_0068	1,334,944	486,994	1,736	?
XA_0090	1,025,155	373,570	143	?
XA_0093r	992,749	413,676	1,538	?
XA_0105	763,845	275,810	731	?
XA_0122	555,495	147,234	1,262	?
XA_0130r	529,974	190,808	988	?
XA_0131	529,647	191,326	89	?
XA_0142r	435,940	144,950	909	?
XA_0147	416,879	133,866	230	?
XA_0151r	401,826	159,492	1,211	?
XA_0154	381,115	151,596	505	?
XA_0168	293,885	109,542	446	?
XA_0176_a	158,685	62,324	460	?
XA_0186	201,898	104,656	84	?
XA_0188	191,904	91,982	76	?
XA_0189	179,792	129,526	181	?
XA_0190	178,629	83,852	407	?
XA_0192	161,858	118,212	197	?

Filter by: Name

NP4-HWUSI-EAS762_0036:4:92:5951:19260#0
From: 580 U580 to 679 U679
Length: 100 U100 (1 mismatch)
Cigar: 100M
Improperly paired (2/2), insert size: 326

NP4-HWUSI-EAS762_0036:4:92:5951:19260#0 (Mate)
From: 354 U354 to 453 U453
Length: 100 U100 (0 mismatches)
Cigar: 100M
Improperly paired (1/2), insert size: 326

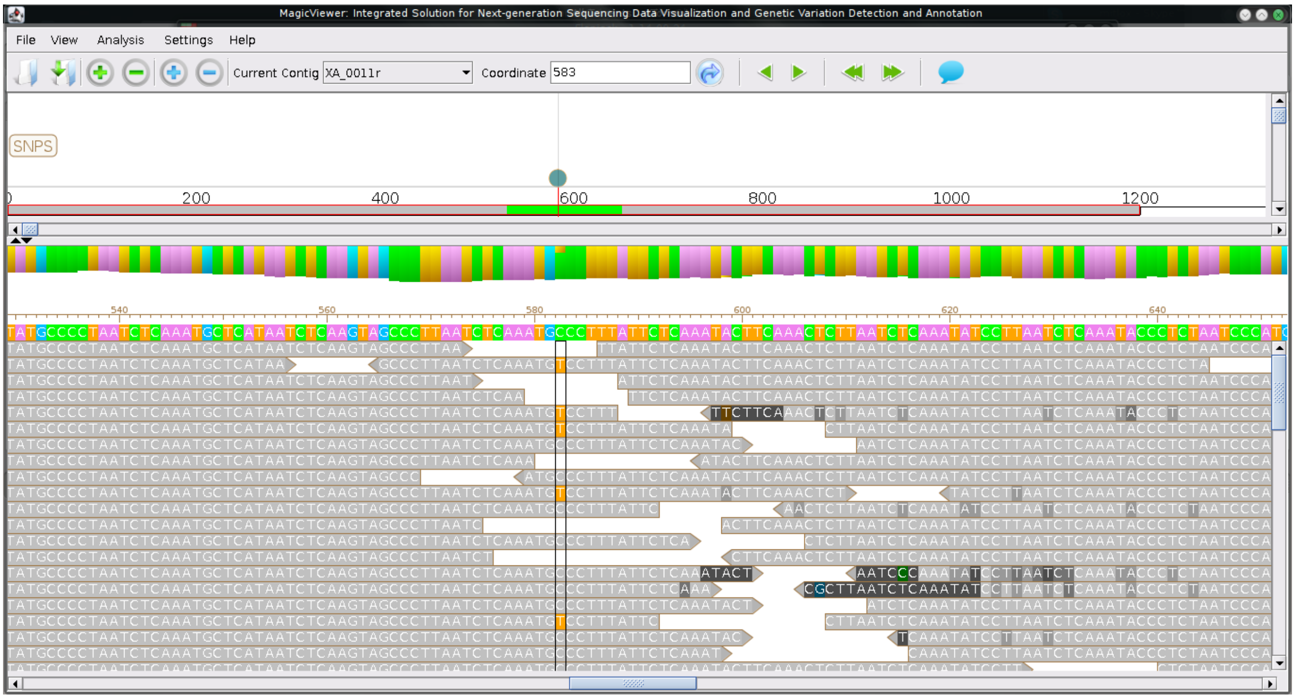
Tablet Tip: GFF3 or BED data can be displayed on the features track after importing 0

Figure 1.16: Tablet showing *Brassica napus* A genome SNPs called by SGSautoSNP.

1.7.3. MagicViewer

MagicViewer (see Figure 1.17) was developed for short read alignment visualisation and annotation. It requires a reference genome sequence in FASTA format, a sorted BAM file containing the aligned short reads and an optional reference genome annotation file in GFF format. It uses a workspace where users can save and load their most frequently used resources for quick access. Through it, users can easily load, browse, further update and modify their previous results, instead of reconstructing a new project.

MagicViewer allows the user to zoom into the image, from the whole chromosome to individual bases. When the mouse hovers on a specific read, a pop-up appears with read ID, location, base quality, read length and orientation. Users can change colours for nucleotide and background and font. The Genome Analysis Toolkit (GATK) (McKenna et al., 2010) is an open-source software framework to develop analysis tools for SGS data. It was built in MagicViewer to identify genetic variation between short reads and reference genomes. MagicViewer allows users to change parameters for heterozygosity, confidence threshold and max coverage. For candidate SNPs, it provides options (thresholds for coverage, quality, variant frequency and number of reads) for display and filtering to remove low confidence SNPs. The output of genetic variation calling is saved in a variable call format (VCF) (Danecek et al., 2011), which is the standard file format used by the 1000 Genomes Project, and it displays the SNPs above the alignment. MagicViewer allows users to adjust parameters (primer length, Tm, GC content, product Tm and the number of primers) for Primer3 in order to generate a specific genomic flanking region primers. MagicViewer is written in Java and is a cache-based viewer which uses low memory (Hou et al., 2010).

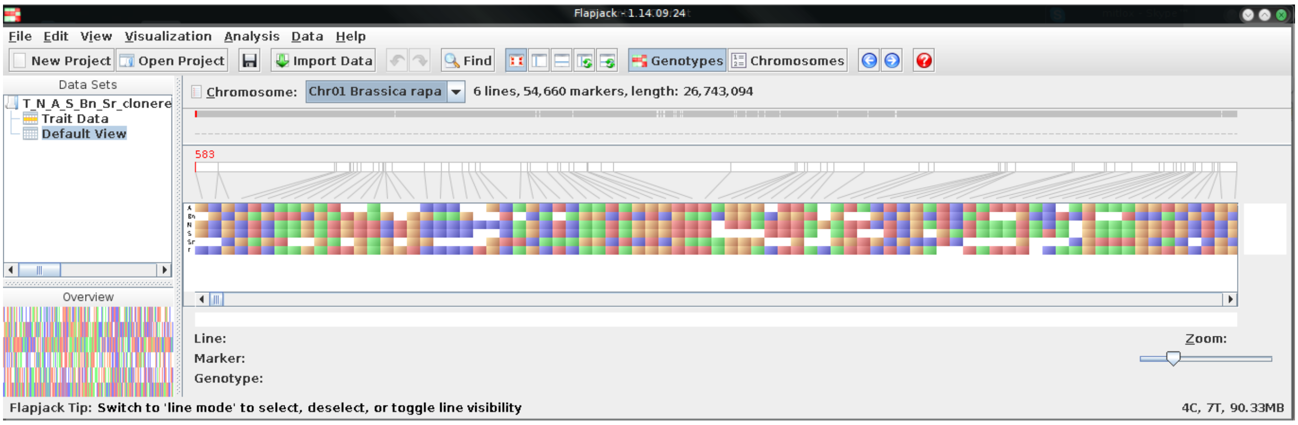

MagicViewer: Integrated Solution for Next-generation Sequencing Data Visualization and Genetic Variation Detection and Annotation
File
View
Analysis
Settings
Help
SNPS
200
400
600
800
1000
1200
540
560
580
600
620
640

Current Contig XA_0011r
Coordinate 583

Figure 1.17: MagicViewer showing *Brassica napus* A genom SNPs called by SGSautoSNP.

1.7.4. Flapjack

Flapjack (see Figure 1.18) is a visual interface for graphical genotyping applications in genetics and plant breeding. Based on the input of map, genotype and trait data Flapjack is able to provide a number of alternative graphical genotype views with individual alleles coloured by state, frequency or similarity to a given standard line. Flapjack supports a range of interactions with the data, including graphically moving lines or markers around the display, insertion or deletion of data, and sorting or clustering of lines by either genotype similarity to other lines, or by trait scores. Any map based information such as QTL positions can be aligned with graphical genotypes to identify associated haplotypes. All project results are saved in an XML-based project format and can also be exported as raw data or graphically as PNG files. Flapjack is freely available for Microsoft Windows, Mac OS X, Linux and Solaris. It is written in Java and can use multi-core processors in order to help to navigate around large or complex datasets (Milne et al., 2010b).



Chromosome: Chr01 Brassica rapa 6 lines, 54,660 markers, length: 26,743,094

Line:
Marker:
Genotype:

Flapjack Tip: Switch to 'line mode' to select, deselect, or toggle line visibility

4C, 7T, 90.33MB

Figure 1.18: Flapjack showing *Brassica napus* A genome (genome markers called by SGSautoSNP).

1.8. Summary and overview of projects presented in the following chapters

It is very important to secure food production for the future rapidly growing population in the face of global environmental change. *Brassica* and wheat are important crops species for Australia and the rest of the world. Second Generation Sequencing has accelerated genome sequencing and made it more affordable. For Second Generation Sequencing a major challenge is to store and work with the huge amount of generated sequences. Therefore new bioinformatics tools have been developed to align, visualise and assemble Second Generation Sequencing data and to analyse the genomes of crop species.

It is the purpose of this thesis is to establish and apply new bioinformatics tools for *Brassica* and wheat Second Generation Sequencing data, to provide researchers and breeders with data to assist them to develop new *Brassica* and wheat varieties that can address the global environmental challenges and feed the fast growing population.

At the start of this thesis we were not satisfied by the already available in-silico SNP discovery tools reviewed in Chapter 1.5, because either they did not provide the functionality which was required for this project or they did not share the code with us. Most of SNP discovery tools were designed for human or simple bacterial genomes. However, these tools do not work well with crop genomes which are often highly homozygous (Batley and Edwards, 2009b, Duran et al., 2009c, Imelfort et al., 2009, Lee et al., 2012). Therefore, Chapter 2 describes the novel developed pipeline for the discovery of SNP in complex genomes. The SGSautoSNP (Second-Generation Sequencing AutoSNP) (Lorenc et al., 2012) pipeline calls SNPs between different individuals using

Illumina paired read data aligned to a reference genome. SGSautoSNP does not consider the reference genome for SNP discovery. Instead, the reference is used to assemble the reads, and SNPs are then called between these assembled reads. SGSautoSNP uses BAM (Binary Alignment/Map) format in order to save memory and space. Furthermore, the pipeline can take advantage of modern multi-core CPUs in order to speed up the SNP discovery. The discovered SNPs can be viewed using a broad range of visualisation tools reviewed in section 1.7 using BAM, GFF3, VCF and Flapjack output files. There is often a requirement to generate a consensus sequence based on the reads mapped to the reference and so SGSautoSNP can generate a consensus sequence as well as marker design files for Illumina GoldenGate or Infinium assay designs. Furthermore the pipeline has been updated after it was published in 2012 and includes scripts for gene and SNP annotation which uses SNAP, a gene prediction software and SNPeff, a SNP annotation and effect prediction tool. In additional it finds SNPs in low SNP density regions and uses gene ontology analysis and goatools to find enrichment of GO terms.

Chapters 3 and 4 show the successfully predicted polymorphisms by SGSautoSNP pipeline in *Brassica* and wheat group 7 chromosomes. The *Brassica* results were stored in a novel database described in Chapter 5. The wheat group 7 chromosomes results were stored in WheatGenome.info (http://www.wheatgenome.info) which provides an integrated database and a range of web application tools to search wheat data (Lai et al., 2012a). These include links to wheat genetic maps using CMap and CMap3D (Duran et al., 2010a, Youens-Clark et al., 2009), and a wheat genome viewer based on GBrowse2 with a BLAST search portal. WheatGenome.info aims to accelerate wheat genome research and contains all data for wheat group 7 chromosomes (Berkman et al., 2013, Berkman et al., 2012a, Berkman et al., 2011). It also includes links to wheat genome data hosted at other research organisations.

Chapter 5 describes the development of a novel platform to store all SNP information discovered by SGSautoSNP. The novel platform is called SGSautoSNPdb and its database is not based on a traditional Relational Database Management System (RDMS), Instead SGSautoSNP uses a document-oriented database which has the advantage that all SNP information can be stored in one document rather than spread in multiple tables. This makes it easier for biologists to understand. Furthermore, SGSautoSNPdb is capable to manage large volumes of data produced by advances in genome technologies efficiently and fast. All information is cross linked to other databases in order to give the researcher results only a click away, instead of having to copy a particular ID and search manually in a search engine.

Chapter 2: Second Generation Sequencing Auto Single Nucleotide Polymorphisms (SGSautoSNP) computational SNP discovery and annotation pipeline

2.1. Introduction

Single nucleotide polymorphisms (SNPs) are becoming the dominant form of molecular marker for genetic and genomic analysis. The advances in second generation DNA sequencing provide opportunities to identify very large numbers of SNPs in a range of species. However, SNP identification remains a challenge for large and polyploid genomes due to their size and complexity, caused by an abundance of transposable elements (Leitch and Leitch, 2008, Meyers and Levin, 2006).

The rapidly expanding genome datasets, driven by advances in second generation DNA sequencing, present a challenge for their management and application (Batley and Edwards, 2009a). At the start of this thesis we were not satisfied by the already available in-silico SNP discovery tools reviewed in Chapter 1 such as ACCUSA (Frohler and Dieterich, 2010), AGSNP (You et al., 2011), NGS-SNP (Grant et al., 2011) and Atlas-SNP2 (Shen et al., 2010), because either they did not supported BAM files, did not use multi-core CPUs, did not calls SNPs between different individuals, instead they called SNPs between the reference genome, or they did not share the code with us. Furthermore, most SNP discovery tools were designed for human or simple bacterial genomes. However, these tools do not work well with crop genomes which are often highly homozygous (Batley and Edwards, 2009b, Duran et al., 2009c, Imelfort et al., 2009, Lee et al., 2012). Therefore, this chapter describes SGSautoSNP (Second-Generation Sequencing AutoSNP) (Lorenc et al., 2012) pipeline development which solved the above short comings from other SNP caller. For example, SGSautoSNP was developed from original concepts used in autoSNP, SNPServer and autoSNPdb (Batley and Edwards, 2009b, Duran et al., 2009a, Savage et al., 2005). Rather than attempting to identify all possible SNPs across a genome, SGSautoSNP is used to identify as many SNPs as possible with the highest confidence, with the acknowledgement that not all biologically present SNPs will be identified. SGSautoSNP method does not consider the reference genome for SNP discovery. Instead, the reference genome is used to assemble the reads, and SNPs are then called between these assembled reads. In SGSautoSNP, mismapped reads produce a heterozygous genotype call at a locus, allowing their distinction from true homozygous SNPs. Three steps were used in order to avoid calling SNPs between homeologs. Firstly, SGSautoSNP discards SNP positions where it is a base conflict within a cultivar. Secondly, only paired reads mapping to a unique location in the genome were kept for further analysis, which is guaranteed by SOAPaligner parameter (*-r 0*) (Li et al., 2009b). Lastly, an additional genome for *Brassica* project (see Chapter 3) and an additional chromosome arm for wheat project (see Chapter 4) were used to align whole genome sequenced cultivars. In the case of wheat, cultivars were mapped to the reference bread wheat chromosome arm shotgun assemblies representing homoelogous chromosomes 7A, 7B and 7D (Berkman et al., 2013), as well as 4AL (Hernandez et al., 2012). In the absence of one of the homoeologues, cultivar specific reads from the missing homoeologue would likely map to one of the other homoeologous genomes, confounding SNP discovery. The SGSautoSNP method does not consider read quality score because these scores are not very reliable, with erroneous nucleotide calls often having high quality scores caused by processes used for the generation of sequence libraries.

The SGSautoSNP pipeline produces output in GFF3, VCF, Flapjack or Illumina Infinium design, this output in particular is used as a format for further genotyping diverse populations. As well as providing an unprecedented resource for diversity analysis, the SGSautoSNP method establishes a foundation for high resolution SNP discovery in large and complex genomes.

After the SGSautoSNP pipeline was first published (Lorenc et al., 2012) new features were implemented as additional code units. It is now possible to associate SNPs with predicted genes, find SNPs in low SNP density regions and associate SNPs in genes with gene ontology classifications. Together this information from the SGSautoSNP pipeline helps us to understand how natural selection has shaped the evolution of plant genomes and provides information which can be applied for crop improvement. SGSautoSNP is freely available on request for academic use.

2.2. Methods

2.2.1. Parallel programming: with a Worker-Queues Model

To handle large and complex genomes it was necessary to develop SGSautoSNP, which is written in the Python programming language, to enable use of multi-core CPUs. However, the current and future versions of CPython, which is the default interpreter for Python, implement the Global Interpreter Lock (GIL). The GIL itself prevents more than a single native thread from running within the interpreter at any given point in time. GIL is required because CPython's memory management is not thread-safe (https://wiki.python.org/moin/GlobalInterpreterLock, 11 April 2014). Thread-safe describes a piece of code that can be called from multiple threads without causing unwanted interaction of shared data structures by multiple threads at the same time (http://en.wikipedia.org/wiki/Thread_safety, 11 April 2014).

In order to take advantage of multi-core CPU systems, Python has to start multiple interpreters and shares the data between them. Just to open and close a new interpreter would take too much time and therefore some SGSautoSNP pipeline scripts are based on a the Worker-Queue Model also know as Worker-Crew Model (Garg and Sharapov, 2001). The original published concept is based on threads. Because of the GIL it was not possible to use threads and therefore instead the SGSautoSNP pipeline uses a different Python interpreter in order to maximize concurrency, because all workers should complete their task at the same time. When using workers without a queue it is more difficult to distribute the load among workers equally. Therefore it is better to use a queue because the script can then split the task dynamically into smaller tasks and put them in a queue. Worker-Queue Model belongs to a Symmetric multiprocessing (SMP) environment, because the task queue has to be shared across all workers.

Some of the SGSautoSNP scripts first create workers on different Python interpreters which are then just there waiting for work on a different CPU core. In the next step the work is passed to each worker in the form of a share queue which contains all tasks. Each worker takes a task out of the queue and processes it. After a worker finishes a task a worker takes a new task without the need to create a new interpreter. As soon as a worker cannot get a new task it shuts down the interpreter, because all tasks have been processed or are still being processed by other workers. In order to distribute

SGSautoSNP scripts across multiple compute nodes, each chromosome was processed by the SGSautoSNP pipeline script on a single compute node on Barrine (see Appendix). With this strategy all chromosomes could be processed in parallel. More details on how some SGSautoSNP pipeline scripts work are described in Section 2.2.2.

2.2.2. SGSautoSNP workflow

The SGSautoSNP pipeline workflow is built out of multiple scripts, a graphical representation of this is shown in Figure 2.1. The user starts with *SOAPalinger.py* script and finishes the full analysis with the last script, *SGSautoSNP_summary.py*, in the SGSautoSNP pipeline. Figure 2.1 shows that after some scripts, for example *MarkDuplicates.py,* two arrows point away to other scripts. In such case the user has to make the choice whether it wants this step or not. Usually, the user can follow all arrows and include scripts. However, *MarkDuplicates.py* is a special case, because after it the user has to make a choice whether they want to discover SNPs from a pseudo chromosome or not.

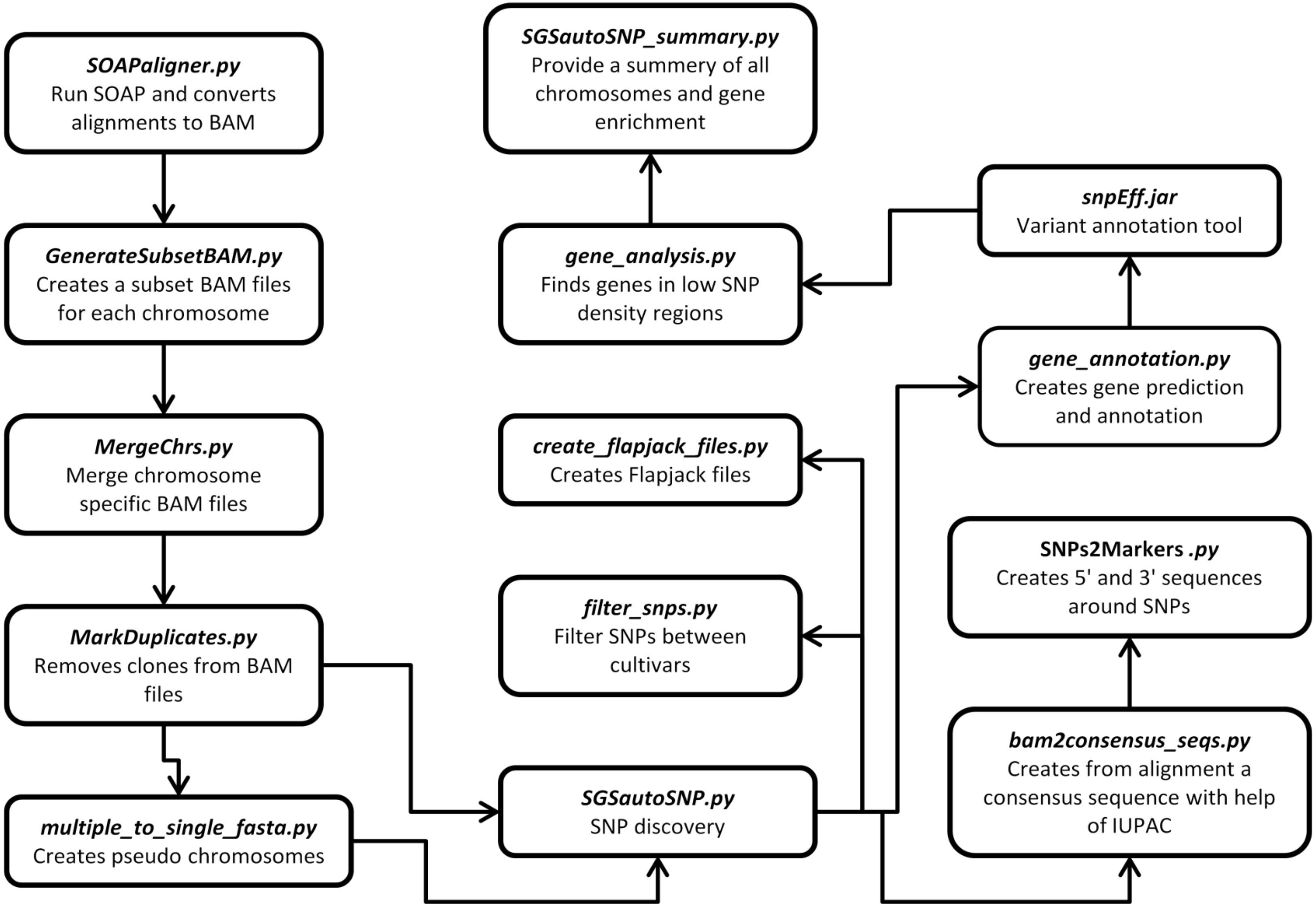
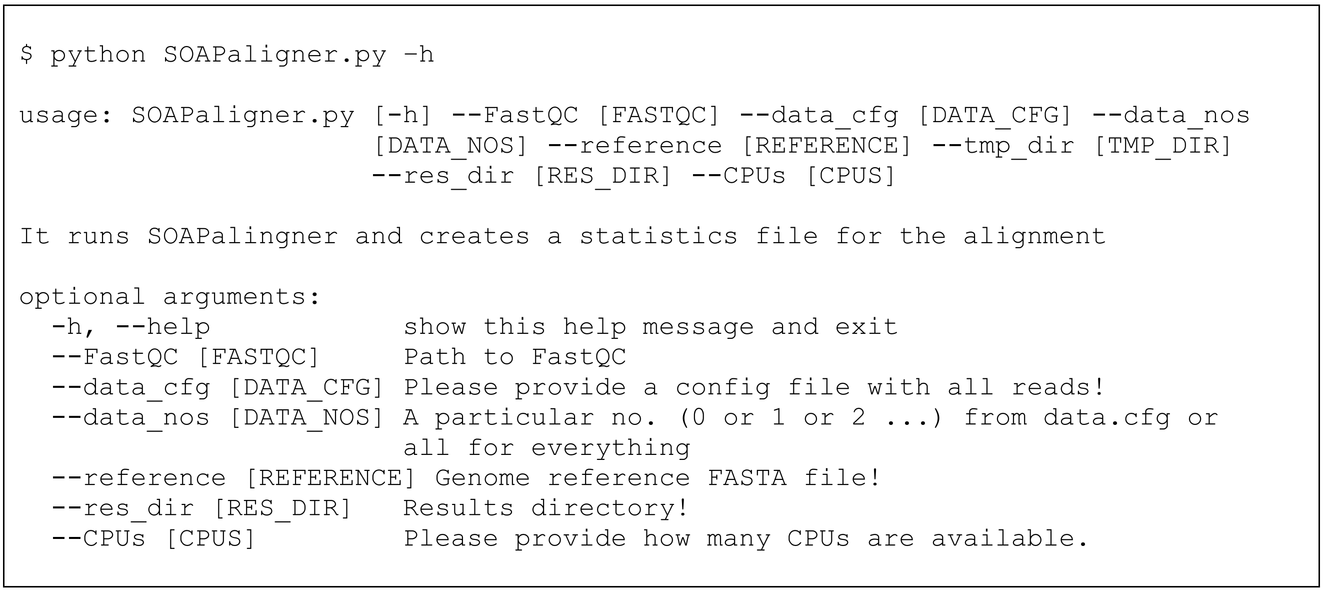

SOAPaligner.py
Run SOAP and converts alignments to BAM
GenerateSubsetBAM.py
Creates a subset BAM files for each chromosome
MergeChrs.py
Merge chromosome specific BAM files
MarkDuplicates.py
Removes clones from BAM files
multiple_to_single_fasta.py
Creates pseudo chromosomes
SGSautoSNP_summary.py
Provide a summery of all chromosomes and gene enrichment
gene_analysis.py
Finds genes in low SNP density regions
create_flapjack_files.py
Creates Flapjack files
filter_snps.py
Filter SNPs between cultivars
SGSautoSNP.py
SNP discovery
snpEff.jar
Variant annotation tool
gene_annotation.py
Creates gene prediction and annotation
SNPs2Markers .py
Creates 5' and 3' sequences around SNPs
bam2consensus_seqs.py
Creates from alignment a consensus sequence with help of IUPAC

Figure 2.1: This flowchart shows the general workflow of the SGSautoSNP pipeline and each box shows the different stages of the process of this pipeline.

2.2.2.1. Mapping reads to the reference

There were three reasons why SOAP (Li et al., 2009b) was choosen to align cultivar specific reads to the reference genome sequences for the SGSautoSNP pipeline. Firstly the SOAP algorithm is fast as described in Chapter 1. Secondly, SOAPaligner does not produce SAM or BAM files, but the developer provides a *soap2sam.pl* (http://soap.genomics.org.cn/down/soap2sam.tar.gz) script which converts SOAP results to SAM format. *SOAPaligner.py* uses this script and also uses SAMtools (Li et al., 2009a) which allows users to convert SAM to BAM format, and sort and index BAM files. Furthermore, SAMtools also helps *SOAPaligner.py* to fill in mate coordinates, ISIZE and mate related flags in alignments. Lastly, SOAP has an option, (-r 0), which removes reads where they match multiple positions equally well. This option aims to increase SNP calling accuracy by ignoring read pairs that cannot be accurately positioned on the reference. Similarly, only reads that mapped as a pair were used for SNP discovery. Due to the short length of the reads, a single read could possibly match many positions, but two reads separated by a gap of defined approximate insert size provides a greater confidence of specific and accurate read mapping. The calling of SNPs between reads aligned to a reference, while ignoring the reference allele, allows this pipeline to be applied to accurately call SNPs between individuals using a reference from a divergent species. The aim is to identify a large number of highly confident SNPs rather than all possible polymorphisms. Regions such as duplicate regions where it is not possible to accurately map sequence reads tend to lead to false SNP calls and so these regions are ignored. Regions of heterozygosity and low sequence coverage also lead to reduced SNP representation. While this pipeline does not attempt to call all biological SNPs, the very large numbers of highly accurate SNPs identified are valuable for genetic studies and the association of agronomic traits with candidate genes.



```
$ python SOAPaligner.py –h

usage: SOAPaligner.py [-h] --FastQC [FASTQC] --data_cfg [DATA_CFG] --data_nos
                      [DATA_NOS] --reference [REFERENCE] --tmp_dir [TMP_DIR]
                      --res_dir [RES_DIR] --CPUs [CPUS]

It runs SOAPalingner and creates a statistics file for the alignment

optional arguments:
  -h, --help            show this help message and exit
  --FastQC [FASTQC]     Path to FastQC
  --data_cfg [DATA_CFG] Please provide a config file with all reads!
  --data_nos [DATA_NOS] A particular no. (0 or 1 or 2 ...) from data.cfg or
                        all for everything
  --reference [REFERENCE] Genome reference FASTA file!
  --res_dir [RES_DIR]   Results directory!
  --CPUs [CPUS]         Please provide how many CPUs are available.
```

Figure 2.2: The command-line of the *SOAPaligner.py* script, showing the various usage options.

SOAP generates three results files for each cultivar: paired-end; single mapped reads; and unmapped reads. Only mapped paired reads were used for further analysis. To be able to use SOAP in an easier way and provide additional functionality, a wrapper, *SOAPaligner.py*, was written. There are a number of parameters that have to be passed to the script. The help message (-*h*) parameter outlines the parameters available for use (**Figure 2.2**). Using the config file (--*data_cfg*) parameter, the user provides all information about the reads which have to be aligned, including the insert size (minimum and maximum), cultivar abbreviation, read names and location of the files. A config file example is shown in **Figure 2.3**. Numbers in brackets represent the data set number 0..N and are used for the data set parameter (--*data_nos*) in *SOAPaligner.py.* This parameter makes it possible for each computing node to grab a particular dataset and align the reads. *SOAPaligner.py* can automatically extract reads with the file extension "*gz*" and "*bz2*". For "*bz2"* it is recommended to have lbzip2 (http://lbzip2.org/), a parallel bzip2 compression utility installed to speed up the unpacking of the reads. A further advantage of the config file is that can be used as quality control. For example, after the project has been completed it is easy to store all relevant information about the run in the related config file.

```
[0]
lane number = 2
species = Species name
cultivar = cultivar_A
library name = H45_03_001
read length = 100
read_a = <My project folder/tmp/fastq>/cultivar_A_Read_a.gz
read_b = <My project folder/tmp/fastq>/cultivar_A_Read_b.gz
min_isize = 60
max_isize = 580
cultivar abbreviation = A

[1]
lane number = 3
species = Species name
cultivar = cultivar_Bn
library name = H45_03_001
read length = 100
read_a = <My project folder/tmp/fastq>/cultivar_Bn_Read_a.bz2
read_b = <My project folder/tmp/fastq>/cultivar_Bn_Read_b.bz2
min_isize = 60
max_isize = 580
cultivar abbreviation = Bn

[N]
lane number = 4
species = Species name
cultivar = cultivar_T
library name = H45_03_001
read length = 100
read_a = <My project folder/tmp/fastq>/cultivar_T_Read_a.gz
read_b = <My project folder/tmp/fastq>/cultivar_T_Read_b.gz
min_isize = 60
max_isize = 580
cultivar abbreviation = T
```

Figure 2.3: Config file for *SOAPaligner.py* which contains all information about the reads to be aligned. Numbers in brackets represent the data set number 0..N.

2.2.2.2. Generating chromosome BAM files

The reference genome used for this analysis contains all chromosomes, which has the advantage that reads align accurately to the correct chromosome. However, for further analysis and SNP calling it is better to split the alignments by chromosome which allows reference to a particular chromosome in a genome. To allow the detection of different cultivars in the BAM files, each read ID has to be modified. All these requirements are performed using *GenerateSubsetBAM.py* (Figure 2.4).

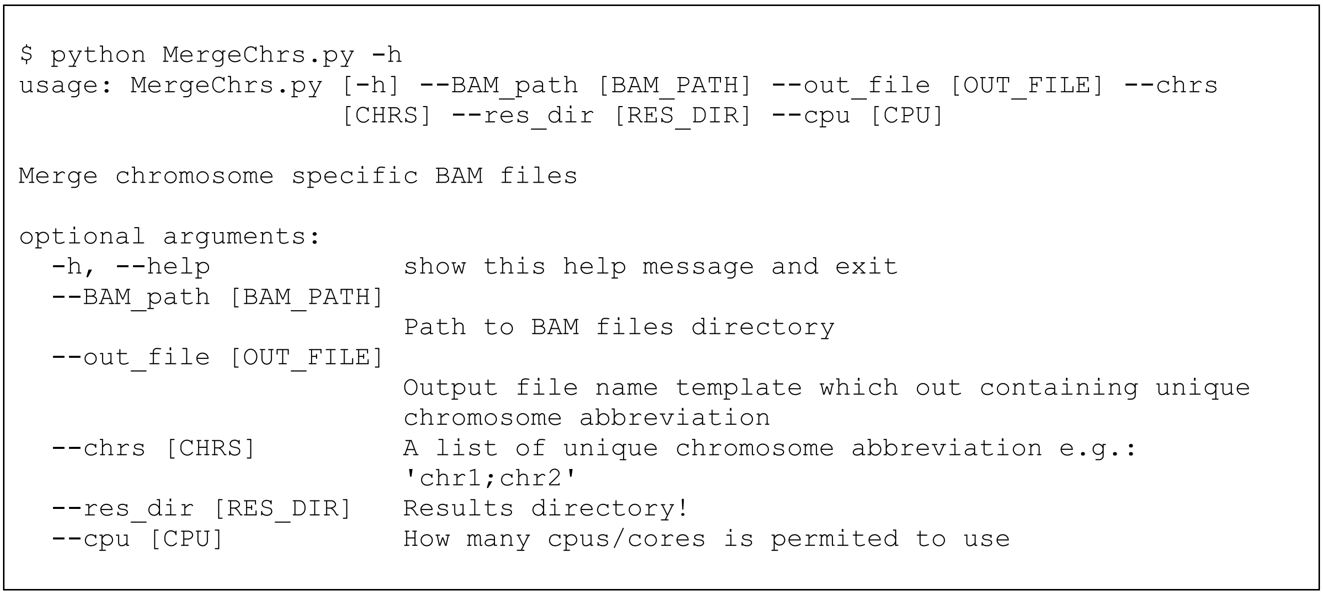
```
$ python GenerateSubsetBAM.py -h
usage: GenerateSubsetBAM.py [-h] --bam [BAM] --ref_path [REF_PATH] --chrs_refs
                            [CHRS_REFS] --cultivar [CULTIVAR] --res_dir
                            [RES_DIR] --cpu [CPU]

Creates a subset BAM files for each chromosome

optional arguments:
  -h, --help            show this help message and exit
  --bam [BAM]           BAM file name from which subset will be created!
  --ref_path [REF_PATH]
                        Path to reference files folder
  --chrs_refs [CHRS_REFS]
                        A list of unique chromosome abbreviation and reference
                        fasta file name seperated by ':' e.g.:
                        'chr1:ex1.fa;chr2:ex2.fa'
  --cultivar [CULTIVAR]
                        Cultivar abbreviation e.g. AP1 which will be inserted
                        in the BAM file in front of read ID
  --res_dir [RES_DIR]   Results directory!
  --cpu [CPU]           How many cpus/cores is permited to use
```

Figure 2.4: The command-line of the *GenerateSubsetBAM.py* script, showing the various usage options.

GenerateSubsetBAM.py uses a Workers-Queue Model where workers are created first on different Python interpreters and then wait for tasks. The jobs are passed to each worker in the form of a share queue which contains all chromosome names and the FASTA file locations. Each worker takes a task out of the queue and inserts in front of each read ID the cultivar name and at the same time creates a BAM file for the chromosome. After a task has finished the worker takes a new task without the need to create a new interpreter. As soon as a worker cannot get a new task it shuts down the interpreter, because all tasks have been processed or are still being processing by other workers. To speed this process up even more, each cultivar could be processed by a separate compute node.

2.2.2.3. Merging chromosome BAM files

During the above processing, each cultivar was split into chromosome BAM files (cultivar1-lane1-chr1, cultivar1-lane1-chr2) and were then combined together for each chromosome. The SGSautoSNP pipeline provides a script called *MergeChrs.py* (Figure 2.5). To produce one BAM file for each chromosome, which contains all cultivars, *MergeChrs.py* has to be run twice. The reason is that in the folder e.g. <My Project folder>/tmp/subset/cult1 there are more chromosome 1 BAM files of the same cultivar (cultivar1-lane1-chr1, cultivar1-lane2-chr1) from different sequencing lanes. Therefore, in the first run it is necessary to combine these lanes together (cultivar1-lane1and2-chr1) and in the next step to combine all of a cultivar's chromosomes BAM files together (cultivar1-lane1and2-chr1 + cultivar2-lane1and2-chr1). Therefore each chromosome BAM file contains all cultivars (cultivar1and2-lane1and2-chr1). *MergeChrs.py* uses a Worker-Queue-Model following the same process as the previous examples.

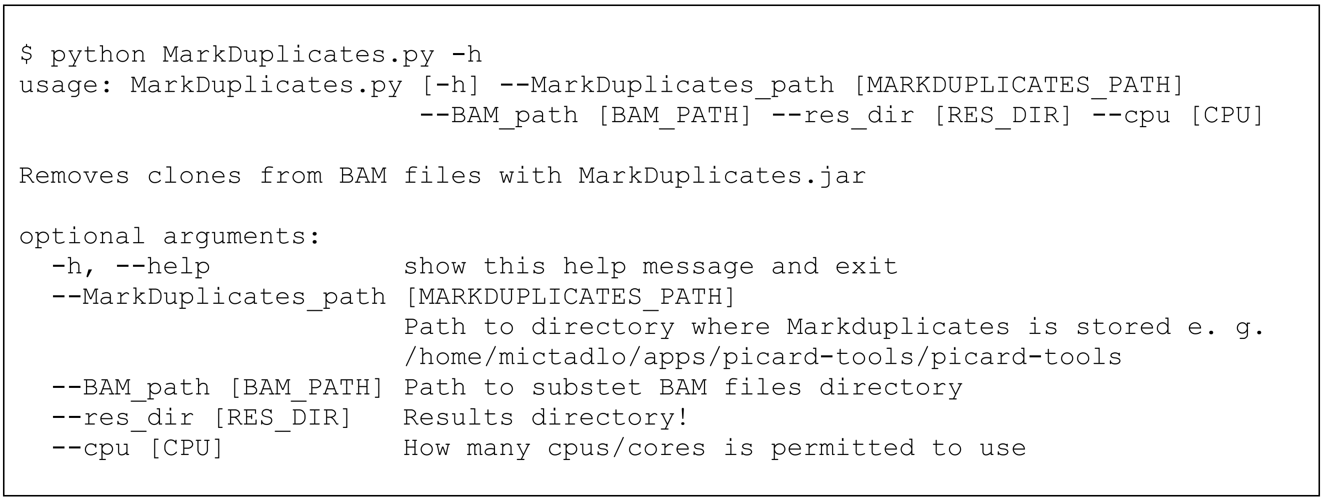


```
$ python MergeChrs.py -h
usage: MergeChrs.py [-h] --BAM_path [BAM_PATH] --out_file [OUT_FILE] --chrs
                    [CHRS] --res_dir [RES_DIR] --cpu [CPU]

Merge chromosome specific BAM files

optional arguments:
  -h, --help            show this help message and exit
  --BAM_path [BAM_PATH]
                        Path to BAM files directory
  --out_file [OUT_FILE]
                        Output file name template which out containing unique
                        chromosome abbreviation
  --chrs [CHRS]         A list of unique chromosome abbreviation e.g.:
                        'chr1;chr2'
  --res_dir [RES_DIR]   Results directory!
  --cpu [CPU]           How many cpus/cores is permited to use
```

Figure 2.5: The command-line of the *MergeChrs.py* script, showing the various usage options.

2.2.2.4. Duplicate removal

Biased representation of DNA inserts like GC content percentages and size differences can be caused by PCR amplification of DNA libraries (Dabney et al., 2013). Read sequences with the same positions on reference genome are most likely of the same insert and therefore these PCR duplicates have to be removed (Schubert et al., 2014). Picard-tools provide Java-based command-line tools to manipulate SAM/BAM files (http://picard.sourceforge.net/). One of the tools, *MarkDuplicates.jar,* is able to detect duplicate mapped reads in BAM files. It defines two pairs as duplicates if they align at the same position, both for their first and second reads. Only one of the duplicate paired reads with the highest average base quality is kept and the rest are deleted as duplicates using the option REMOVE_DUPLICATES=true (Pireddu et al., 2011). *MarkDuplicates.py* (Figure 2.6) is a wrapper for *MarkDuplicates.jar* and provides a Worker-Queue-Model to distribute the tasks where workers are created first on different Python interpreters and wait for tasks.



```
$ python MarkDuplicates.py -h
usage: MarkDuplicates.py [-h] --MarkDuplicates_path [MARKDUPLICATES_PATH]
                         --BAM_path [BAM_PATH] --res_dir [RES_DIR] --cpu [CPU]

Removes clones from BAM files with MarkDuplicates.jar

optional arguments:
  -h, --help            show this help message and exit
  --MarkDuplicates_path [MARKDUPLICATES_PATH]
                        Path to directory where Markduplicates is stored e. g.
                        /home/mictadlo/apps/picard-tools/picard-tools
  --BAM_path [BAM_PATH] Path to substet BAM files directory
  --res_dir [RES_DIR]   Results directory!
  --cpu [CPU]           How many cpus/cores is permitted to use
```

Figure 2.6: The command-line of the *MarkDuplicates.py* script, showing the various usage options.

Each worker grabs a BAM file name out of a share queue which contains all BAM file names and runs *MarkDuplicates.jar* internally. After a task has finished, the worker takes a new task without the need to create a new interpreter. As soon as a worker cannot get a new task it shuts down the interpreter, because all tasks have been processed or are still being processing by other workers. To speed this process up even more, each cultivar could be processed by a separate compute node.

2.2.2.5. Pseudo-chromosome building

For the *Brassica* work it was necessary to build pseudo chromosomes; the *multiple_to_single_fasta.py* (Figure 2.7) script creates them. A multiple FASTA file has to be provided as input to the script. During the process the sequences from each entry are concatenated, and filler sequence, e.g. 100 Ns, are inserted between each sequence. Furthermore, the script produces an additional output file in GFF3 which contains the start and end positions of the sequences (see Figure 2.8).

```
$ python multiple_to_single_fasta.py -h
Usage: multiple_to_single_fasta.py -v Chr1 -f t.m.fasta -s fasta -r r -n 100
multiple_to_single_fasta.py -v 7DS_PSMOL_0.3 -f t.m.fasta -s
ACPFG_pseudomolecule -r - -n 2000

Options:
  -h, --help            show this help message and exit
  -v PSMOLVER, --path=PSMOLVER
                        Please give PSMOL name and version eg. 7DS_PSMOL_0.3
                        or chromosome eg. Chr1
  -f MFASTA, --fasta=MFASTA
                        Please give a multiple fasta files eg. t.m.fasta.
  -s SOURCE, --source=SOURCE
                        Please give the source where the data comes from eg.
                        ACPFG_pseudomolecule or fasta.
  -r REVERSE, --reverse=REVERSE
                        Please specify what character is the orientation eg. r
                        or -.
  -n SPACERNO, --spacer=SPACERNO
                        Please give how many N do you want as spacer.
```

Figure 2.7: The command-line of the *multiple_to_single_fasta.py* script, showing the various usage options.

```
##gff-version             3
##sequence-region  XA07_v3.0       1  22305823
XA07_v3.0          fasta      contig         1    359785  .  +  .  ID=XA_0158;Name=XA_0158
XA07_v3.0          fasta      contig    359886    956461  .  +  .  ID=XA_0117;Name=XA_0117
XA07_v3.0          fasta      contig    956562   1177385  .  +  .  ID=XA_0181;Name=XA_0181
XA07_v3.0          fasta      contig   1177486   1782250  .  +  .  ID=XA_0116;Name=XA_0116
XA07_v3.0          fasta      contig   1782351   5098604  .  -  .  ID=XA_0017r;Name=XA_0017r
XA07_v3.0          fasta      contig   5098705   5484666  .  +  .  ID=XA_0153;Name=XA_0153
XA07_v3.0          fasta      contig   5484767   6809755  .  -  .  ID=XA_0069r;Name=XA_0069r
XA07_v3.0          fasta      contig   6809856  11170108  .  -  .  ID=XA_0012r;Name=XA_0012r
XA07_v3.0          fasta      contig  11170209  18223632  .  +  .  ID=XA_0003;Name=XA_0003
XA07_v3.0          fasta      contig  18223733  21429098  .  -  .  ID=XA_0019r;Name=XA_0019r
XA07_v3.0          fasta      contig  21429199  22305823  .  +  .  ID=XA_0101;Name=XA_0101
```

Figure 2.8: GFF3 file which contains the start and end positions of the sequences for Chromosome 7 of the *Brassica napus* AA genome.

2.2.2.6. SNP discovery

SGSautoSNP is different from most SNP callers in that the reference is used to assemble the reads, but SNPs are then called between these assembled reads and not between the reads and the reference. The *SGSautoSNP.py* algorithm uses two steps to call a SNP at each locus. Primary SNP calling requires a SNP redundancy score of at least 2. The SNP redundancy score is the minimum number of reads calling the SNP allele at the locus. To understand the SNP score better let's consider a random position in an alignment where cultivars contain the following bases:

- cultivar1 has 6 As
- cultivar2 has 1 G
- cultivar3 has 1 G

In the above example there are 2 Gs and 6 As, two Gs is the minimum and therefore the SNP score is 2. As at least 2 reads are required, each from at least 2 cultivars to call a SNP, the minimum read coverage at a locus to call a SNP is therefore 4. After this initial SNP call, the algorithm asks if all bases within each cultivar at a locus are the same, which would be expected for homozygous genomes. This process identifies erroneously called SNPs that are due to mis-mapping of reads.

SGSautoSNP.py uses a Worker-Queue Model which differs from the above scripts because it has also a results queue in addition to the tasks queue. Workers are created first on different Python interpreters and wait. In the next step the work is passed to each worker in the form of the share queue which contains all contigs names. Each worker takes a contig name out of the queue and processes it, and after it finishes this worker passes the result to a share results queue.

SGSautoSNP.py (Figure 2.9) produces five output types. A statistics file with the file extension '.stat' contains SNP calling statistics including: (i) scaffold name (ii) SNP number (iii) SNP types (transitions and transversions) (iv) scaffold length (see Figure 2.10). The end of this file contains a summary of all scaffolds. The first results file with the extension '.snp' contains human readable SNP information in text format which can be easily parsed to other formats. Information includes: (i) scaffold name (ii) SNP position on the scaffold (iii) SNP position on the chromosome (iv) SNP score (v) genotypes (which base and how many appear in a particular cultivar) (vi) allele (vii) SNP ID (see Figure 2.11). Three further results formats are produced. VCF (Danecek et al., 2011) files are created to allow the user to view the SNPs in MagicViewer (Hou et al., 2010) and to annotate the SNP using SnpEff (Cingolani et al., 2012) (see Figure 2.12 and Figure 2.13), whereas the chromosome VCF file only will be created by using "*--chr_output*" and "*--chr_offset*" parameters. GFF3 format (see Figure 2.14 and Figure 2.15), whereas the chromosome GFF3 file only will be created by using *"--chr_output*" and "*--chr_offset*" parameters. These results are produced for viewing in the GBrowse generic genome browser (Donlin, 2009) and Tablet (Milne et al., 2013). The help message (-*h*) parameter outlines the parameters available for *SGSautoSNP.py* (Figure 2.9).

```
$ python SGSautoSNP.py -h
usage: SGSautoSNP.py [-h] --bam [BAM] --fasta [FASTA] --snp_id_prefix
                     [SNP_ID_PREFIX] --contig_output [CONTIG_OUTPUT]
                     [--chr_offset [CHR_OFFSET]] [--chr_output [CHR_OUTPUT]]
                     --cultivars [CULTIVARS] --cpu [CPU]

SGSautoSNP a parallel SNP discovery tool for BAM files

optional arguments:
  -h, --help            show this help message and exit
  --bam [BAM]           bam file need bai file!
  --fasta [FASTA]       Input single/multiple fasta file !
  --snp_id_prefix [SNP_ID_PREFIX]
                        Please provide the prefix of each SNP ID. eg. UQ01F
  --contig_output [CONTIG_OUTPUT]
                        Please provide an output file name. SGSautoSNPwill
                        attached the following suffix to it: gff3, vcf, snp
                        and stat
  --chr_offset [CHR_OFFSET]
                        Provide an offset file and it will offset contig
                        positions on chromosome. You have to use --chr_output,
                        too.
  --chr_output [CHR_OUTPUT]
                        Please provide an GFF3 output file name for
                        chromosome. You have to use --chr_offset, too.
  --cultivars [CULTIVARS]
                        Give all cultivars which are in BAM files eg.
                        "J,E,A,S,M1,M2,Bn,Sr"
  --cpu [CPU]           How many CPUs/Cores you would like to use
```

Figure 2.9: The command-line of the *SGSautoSNP.py* script, showing the various usage options.

```
             SNPs
scaffolds    no.     mutations
XA_0158         456  A/C=49;A/G=112;A/T=67;C/G=39;C/T=133;G/T=56
XA_0117         892  A/C=85;A/G=253;A/T=150;C/G=82;C/T=223;G/T=99
XA_0181         615  A/C/T=2;A/C=63;A/G=183;A/T=91;C/G/T=1;C/G=36;C/T=178;G/T=61
XA_0116         154  A/C=15;A/G=44;A/T=20;C/G=14;C/T=42;G/T=19
XA_0017r       6984  A/C/T=2;A/C=756;A/G/T=1;A/G=2030;A/T=970;C/G=495;C/T=1969;G/T=761
XA_0153        1041  A/C/G=1;A/C=118;A/G/T=1;A/G=275;A/T=168;C/G=77;C/T=257;G/T=144
XA_0069r       4062  A/C/T=1;A/C=469;A/G/T=1;A/G=1145;A/T=588;C/G/T=1;C/G=271;C/T=1141;G/T=445
XA_0012r      15140  A/C/G=6;A/C/T=8;A/C=1673;A/G/T=9;A/G=4176;A/T=2222;C/G/T=2;C/G=1048;C/T=4247;G/T=1749
XA_0003       21644  A/C/G=4;A/C/T=9;A/C=2412;A/G/T=11;A/G=5890;A/T=3302;C/G/T=4;C/G=1682;C/T=5930;G/T=2400
XA_0019r       6652  A/C/G=2;A/C/T=4;A/C=731;A/G/T=2;A/G=1830;A/T=998;C/G/T=2;C/G=550;C/T=1813;G/T=720
XA_0101        2744  A/C/G=1;A/C=285;A/G/T=1;A/G=729;A/T=429;C/G=221;C/T=747;G/T=331
Total         60384  A/C/G=14;A/C/T=26;A/C=6656;A/G/T=26;A/G=16667;A/T=9005;C/G/T=10;C/G=4515;C/T=16680;G/T=6785
```

Figure 2.10: A statistics file contains information about SNP calling for chromosome 7 of *Brassica rapa* genome. The end of this file contains a summary of all scaffolds.

```
sca.     sca. pos  chr. pos  SNP score  genotypes                                allele  SNP id
XA_0158      1802      1802          2  A=2*G;Bn=2*G;N=12*G;S=3*G;Sr=2*C;T=24*G  C/G     UQXAH070000001
XA_0158      2136      2136          9  A=1*C;Bn=7*C;N=9*G;S=1*C;Sr=1*C;T=16*C   C/G     UQXAH070000002
                                                           …
XA_0117      6471    366356          2  A=2*G;Bn=0*X;N=4*G;S=2*T;Sr=1*G;T=0*X    G/T     UQXAH070000457
XA_0117     11659    371544          2  A=1*G;Bn=1*C;N=32*C;S=2*C;Sr=1*G;T=16*C  C/G     UQXAH070000458
                                                           …
XA_0101    876248  22305446         10  A=0*X;Bn=2*A;N=5*G;S=0*X;Sr=5*G;T=12*A   A/G     UQXAH070060383
XA_0101    876498  22305696          2  A=0*X;Bn=0*X;N=0*X;S=0*X;Sr=2*T;T=2*G    G/T     UQXAH070060384
```

Figure 2.11: A snippet of the ".snp" file which contains human readable SNP information of chromosome 7 for the *Brassica rapa* genome.

```
##fileformat=VCFv4.0
##filedate=20131031
##source=SGSautoSNP
##reference=XA07m_v3.0.fa
##phasing=allhomozygote
##INFO=<ID=DP,Number=1,Type=Integer,Description="Read depth over all samples">
##INFO=<ID=PL,Number=0,Type=String,Description="Panel">
#CHROM      POS        ID               REF  ALT  QUAL  FILTER  INFO    FORMAT    A      Bn      N       S      Sr      T
XA07_v3.0       1802  UQXAH070000001  C    G       2     .     DP=45   GT:DP   1/1:2  1/1:2  1/1:12  1/1:3  0/0:2  1/1:24
XA07_v3.0       2136  UQXAH070000002  G    C       9     .     DP=35   GT:DP   1/1:1  1/1:7  0/0:9   1/1:1  1/1:1  1/1:16
                                                              …
XA07_v3.0  22305446  UQXAH070060383  G    A      10     .     DP=24   GT:DP   ./.:0  1/1:2  0/0:5   ./.:0  0/0:5  1/1:12
XA07_v3.0  22305696  UQXAH070060384  T    G       2     .     DP=4    GT:DP   ./.:0  ./.:0  ./.:0   ./.:0  0/0:2  1/1:2
```

Figure 2.12: A snippet of the ".vcf" file which contains chromosome SNP information of chromosome 7 for the *Brassica rapa* genome.

```
##fileformat=VCFv4.0
##filedate=20131031
##source=SGSautoSNP
##reference=XA07m_v3.0.fa
##phasing=allhomozygote
##INFO=<ID=DP,Number=1,Type=Integer,Description="Read depth over all samples">
##INFO=<ID=PL,Number=0,Type=String,Description="Panel">
#CHROM    POS     ID               REF  ALT  QUAL  FILTER  INFO    FORMAT  A      Bn     N       S      Sr     T
XA_0158    1802  UQXAH070000001   C    G       2  .       DP=45   GT:DP   1/1:2  1/1:2  1/1:12  1/1:3  0/0:2  1/1:24
XA_0158    2136  UQXAH070000002   G    C       9  .       DP=35   GT:DP   1/1:1  1/1:7  0/0:9   1/1:1  1/1:1  1/1:16
                                                              …
XA_0101  876248  UQXAH070060383   G    A      10  .       DP=24   GT:DP   ./.:0  1/1:2  0/0:5   ./.:0  0/0:5  1/1:12
XA_0101  876498  UQXAH070060384   T    G       2  .       DP=4    GT:DP   ./.:0  ./.:0  ./.:0   ./.:0  0/0:2  1/1:2
```

Figure 2.13: A snippet of the ".vcf" file which contains contig SNP information of chromosome 7 for the *Brassica rapa* genome.

```
##gff-version 3
XA07_v3.0    SGSautoSNP    SNP         1802         1802   .   .   .   Name=UQXAH070000001;ID=UQXAH070000001;Contig name=XA_0158;SNP score=2;
                                                                       SNP pos. on scaffold=1802;Genotype A=2*G;Genotype Bn=2*G;Genotype N=12*G;
                                                                       Genotype S=3*G;Genotype Sr=2*C;Genotype T=24*G;Changes=C/G
                                                                                        ...
XA07_v3.0    SGSautoSNP    SNP    22305696    22305696   .   .   .   Name=UQXAH070060384;ID=UQXAH070060384;Contig name=XA_0101;SNP score=2;
                                                                       SNP pos. on scaffold=876498;Genotype A=0*X;Genotype Bn=0*X;Genotype N=0*X;
                                                                       Genotype S=0*X;Genotype Sr=2*T;Genotype T=2*G;Changes=G/T
```

Figure 2.14: A snippet of the ".gff3" file which contains chromosome SNP information of chromosome 7 for the *Brassica rapa* genome. The last column had to be split in order to fit on this side.

```
##gff-version 3
XA_0158    SGSautoSNP    SNP    1802      1802       .   .   .   Name=UQXAH070000001;ID=UQXAH070000001;Contig name=XA_0158;SNP score=2;
                                                                  SNP pos. on scaffold=1802;Genotype A=2*G;Genotype Bn=2*G;Genotype N=12*G;
                                                                  Genotype S=3*G;Genotype Sr=2*C;Genotype T=24*G;Changes=C/G
                                                                                   ...
XA_0101    SGSautoSNP    SNP    876498    876498     .   .   .   Name=UQXAH070060384;ID=UQXAH070060384;Contig name=XA_0101;SNP score=2;
                                                                  SNP pos. on scaffold=876498;Genotype A=0*X;Genotype Bn=0*X;Genotype N=0*X;
                                                                  Genotype S=0*X;Genotype Sr=2*T;Genotype T=2*G;Changes=G/T
```

Figure 2.15: A snippet of the "*.gff3*" file which contains contig SNP information of chromosome 7 for the *Brassica rapa* genome. The last column had to be split in order to fit on this side.

2.2.2.7. SNP filtering

While SNP calling may use many individuals or cultivars, SNPs that differentiate between two specific individuals or cultivars are frequently required for downstream analysis. The *filter_SNPs.py* script (Figure 2.16) parses the text '*.snp*' file which was generated by *SGSautoSNP.py* and recognises all available cultivars. In the next step it processes each SNP position and rejects cultivars which have an X as the base, which means that this cultivar was not represented at the locus position. Then it creates all cultivar combinations and generates the same format output files as *SGSautoSNP.py* for each cultivar combination, but specifically for SNPs between every pair of individuals. This script also produces a .matrix file which details the number of SNPs between all combinations of cultivars.

```
$ python filter_snps.py  -h
usage: filter_snps.py [-h] --contig_output [CONTIG_OUTPUT] --fasta [FASTA]
                      [--chr_output [CHR_OUTPUT]] [--chr_name [CHR_NAME]]
                      --snps [SNPS] --dir [DIR]

Filter SNPs between cultivars

optional arguments:
  -h, --help            show this help message and exit
  --contig_output [CONTIG_OUTPUT]
                        Please provide an output file name template.
  --fasta [FASTA]       Input fasta file !
  --chr_output [CHR_OUTPUT]
                        Please provide an GFF3 output file name for
                        chromosome. You have to use --chr_name, too.
  --chr_name [CHR_NAME]
                        For GFF3 eg. XA10_v4.0. You have to use --chr_output,
                        too.
  --snps [SNPS]         SNP file!
  --dir [DIR]           Ouput directory for results files.
```

Figure 2.16: The command-line of the *filter_snps.py* script, showing the various usage options.

2.2.2.8. Creating Flapjack files

The SGSautoSNP pipeline provides a *flapjack_files.py* script (Figure 2.17) to generate Flapjack (Milne et al., 2010b) input files. Flapjack allows the visualisation of markers, lines and their corresponding SNP calls per chromosome. It allows the selection of lines and markers from datasets. Furthermore it allows the filtering of lines by markers.

```
$ python create_flapjack_files.py –h
usage: create_flapjack_files.py [-h] --snp [SNP] --species [SPECIES]
                                [--chr_name [CHR_NAME]] --output [OUTPUT]
                                --dir [DIR]

It creates Flapjack files

optional arguments:
  -h, --help            show this help message and exit
  --snp [SNP]           SNP file
  --species [SPECIES]   eg. "Brassica napus"
  --chr_name [CHR_NAME]
                        eg. Chr1
  --output [OUTPUT]     Please provide an output file name template.
  --dir [DIR]           Ouput directory for results files.
```

Figure 2.17: The command-line of the *create_flapjack_files.py* script, showing the various usage options.

2.2.2.9. SNP density

SNP density analysis helps to identify regions of high sequence conservation and enables a greater understanding of their evolutionary history and selection. Regions with high SNP density are least conserved and regions with the lowest SNP density are the most conserved. The SGSautoSNP pipeline provides a script called *snp_density_coverage_percentage.py* (Figure 2.18) which was designed to map SNP density across each chromosome.

```
$ python snp_density_coverage_percentage.py -h
Usage: snp_density_coverage_percentage.py --chr <chromosome tag> --bam
<alignments.bam> --contigs <chr_contig.gff3> --plot <png | svg | eps> [--low
<lower_coverage_limit>] [--up <upper_coverage_limit>] [--window <window_size>]
[--snp <chromosome.snp>]

Options:
  -h, --help            show this help message and exit
  --chr=CHR             input chromosome tag
  --bam=BAM_FILE_NAME   input bam file
  -c CONTIGS_FILE_NAME, --contigs=CONTIGS_FILE_NAME
                        input contig location gff file
  --snp=SNP_FILE_NAME   input snp location gff file
  -l LLIM, --low=LLIM   input lower coverage limit [default=4]
  -u ULIM, --up=ULIM    input upper coverage limit [default=100]
  -w WINDOW_SIZE, --window=WINDOW_SIZE
                        input window size [default=5000]
  --cov_step=COV_STEP   coverage step size (positive integer 1-100)
                        [default=1]
  -p PLOT_FORMAT, --plot=PLOT_FORMAT
                        output plot format
```

Figure 2.18: The command-line of the *snp_density_coverage_percentage.py* script, showing the various usage options.

2.2.2.10. Generating consensus sequence

The *bam2consensus_seqs.py* script accepts an alignment in BAM format and generates a consensus sequence for each scaffold (Figure 2.19)**.** The script goes through all nucleotide positions and generates a consensus sequence using the following rules: (i) if no base exists at the position then an N will be inserted; (ii) if only a single read covers the locus then the algorithm uses this read sequence (iii) if more than one read covers the position, and all nucleotides are the same, this nucleotide will be inserted; (iv) if more than one read covers the position, and one single read conflicts with the others, this single read is assumed to be an error and ignored, the majority base is inserted; (iv) if more than one read covers the position, and more than one read conflicts with the others, a degenerate base is inserted using IUPAC notation (see Table 2.1).

Table 2.1: Summary of single-letter code recommendations represented by IUPAC notation (adapted from http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html).

Symbol	Bases represented	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

This script uses a Worker-Queue Model, which is the same as developed for *SGSautoSNP.py*, because it has also a results queue additional to the tasks queue. Firstly the workers are created on different Python interpreters and wait for tasks. As soon as the share task queue has been filled with all contigs names, each worker takes a contig name out of the queue and processes it in parallel with the other workers. The output file is one multiple FASTA file which include all contigs in the original BAM file.

```
$ python bam2consensus_seqs.py -h
usage: bam2consensus_seqs.py [-h] --bam [BAM] --fasta [FASTA] --cpu [CPU]
                             --output [OUTPUT]

It creates from alignment a consensus sequence with help of IUPAC

optional arguments:
  -h, --help         show this help message and exit
  --bam [BAM]        Bam file name!
  --fasta [FASTA]    Input single/multiple fasta file !
  --cpu [CPU]        How many CPUs/Cores you would like to use
  --output [OUTPUT]  Output file name template for the consensus sequence.
```

Figure 2.19: The command-line of the *bam2consensus_seqs.py* script, showing the various usage options.

2.2.2.11. Generating Illumina marker assay files

The SGSautoSNP pipeline can generate Illumina marker assay files for the design of Illumina Infinium and GoldenGate genotyping arrays. The *SNP2Markers.py* script requires as an input file the consensus sequence in FASTA format generated by *bam2consensus_seqs.py,* and the text SNP file with a '.snp' extension generated by *SGSautoSNP.py*. Additional parameters include (i) species (ii) number of cultivars (iii) SNP library name (iv) version number (v) chromosome name (vi) output directory for the results files (see Figure 2.20).

The script extracts the 5' and 3' sequence surrounding each predicted SNP in the following way: (i) the nucleotide sequence 150 bases each side of the SNP is extracted together with the SNP position in the format [A/C]; (ii) as the Illumina GoldenGate and Infinium assays designs probes up to 60 bp adjacent to the SNP, assays are discarded if this region contains any N characters within the consensus sequence.

This script uses a Worker-Queue Model together with a tasks and a results queue as previously described. Output files include a file of summary statistics '*_marker.stat' and a marker assay file for input into the Illumina SNP assay design '*_GoldenDB.csv'.

```
$ python SNPs2Markers.py -h
usage: SNPs2Markers.py [-h] --fasta [FASTA] --snp [SNP] --species [SPECIES]
                       --germplasm [GERMPLASM] --library [LIBRARY] --panel
                       [PANEL] --chr_name [CHR_NAME] --cpu [CPU] --output
                       [OUTPUT]

SNPs2Markers creates 5' and 3' sequences around SNPs

optional arguments:
  -h, --help            show this help message and exit
  --fasta [FASTA]       Consensus sequence fasta
  --snp [SNP]           SNP file
  --species [SPECIES]   eg. "Brassica napus"
  --germplasm [GERMPLASM]
                        eg. 8_canola_lines
  --library [LIBRARY]   eg. UQ_BNSNP
  --panel [PANEL]       eg. UQ_BNSNP_A_V4.0
  --chr_name [CHR_NAME]
                        eg. Chr1
  --cpu [CPU]           How many CPUs/Cores you would like to use
  --output [OUTPUT]     Please provide an output file name. SNPs2Markers will
                        attached the suffix to it:
```

Figure 2.20: The command-line of the *SNPs2Markers.py* script, showing the various usage options.

2.2.2.12. Gene annotation

Working with new genomes has the disadvantage that they may not have any annotation available. The *gene_annotation.py* script has been developed to provide basic annotation for such genomes. In the first step it uses SNAP (Korf, 2004), a gene prediction software, and then it runs BLASTp on the predicted genes to find out whether any of the genes hit a Swiss-Prot entry. Swiss-Prot is a high-quality, manually annotated, non-redundant protein sequence database maintained by the UniProt consortium. It combines information extracted from scientific literature and computational analysis. The aim of Swiss-Prot is to provide all known relevant information about a particular protein. New releases are published every 2 weeks and can be downloaded (Boutet et al., 2007). If a predicted gene does not match any Swiss-Prot entry it will be rejected. This has advantage of reducing the number of falsely predicted genes. In case a predicted gene has a hit with a Swiss-Prot entry it will save the Swiss-Prot accession number and description.

The UniProt Gene Ontology Annotation (UniProt-GOA) project at the European Bioinformatics Institute (EBI) (Barrell et al., 2009) provides a file which is in tab-delimited format that associates a Swiss-Prot accession number with one or more Gene Ontology (GO) terms (Ashburner et al., 2000). GO maintained by the Gene Ontology Consortium (http://www.geneontology.com). It is a major bioinformatics project to unify the representation of gene and gene product attributes across all genomes including plant, animal and microbial genomes. The ontology covers three domains:

- A cellular component is an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).
- Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. It does not specify where or when, or in what context, the action takes place.A biological process is series of events with a defined start and end, accomplished by sets of molecular functions.

The UniProt-GOA file was downloaded and extracted from the following server:

ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz

The *gene_annotation.py* script (Figure 2.21) will combine all results from SNAP, BLASTp and the GOA association file to create an annotation file in GFF3 format which can be used for further analysis steps.

```
$ python gene_annotation.py -h
usage: gene_annotation.py [-h] --fasta [FASTA] --out_dir [OUT_DIR] --hmm [HMM]
                          [--xml [XML]] [--blastDB [BLASTDB]] --goa [GOA]
                          [--contig_output [CONTIG_OUTPUT]]
                          [--chr_offset [CHR_OFFSET]]
                          [--chr_output [CHR_OUTPUT]] --cpus [CPUS]

Gene prediction and annotation

optional arguments:
  -h, --help            show this help message and exit
  --fasta [FASTA]       Input reference fasta file!
  --out_dir [OUT_DIR]   Output directory
  --hmm [HMM]           Hmm file for SNAP prediction
  --xml [XML]           If you have a Blast XML file
  --blastDB [BLASTDB]   If you need to run Blast than give Blast DB location
  --goa [GOA]           Please provide gene_association.goa_uniprot
  --contig_output [CONTIG_OUTPUT]
                        Please provide an output file name will attached the
                        following suffix to it: gff3, vcf, snp and stat
  --chr_offset [CHR_OFFSET]
                        Provide an offset file and it will offset contig
                        positions on chromosome. You have to use --chr_output,
                        too.
  --chr_output [CHR_OUTPUT]
                        Please provide an GFF3 output file name for
                        chromosome. You have to use --chr_offset, too.
  --cpus [CPUS]         How many CPUs/Cores you would like to use
```

Figure 2.21: The command-line of the *gene_annotation.py* script, showing the various usage options.

2.2.2.13. SNP annotation

The SnpEff variant annotation tool (Cingolani et al., 2012) was used to predict the effect of the identified SNPs from *SGSautoSNP.py* using the annotation GFF3 file of *gene_annotation.py* within different genomic DNA sequences, including putative exons, introns, and gene upstream and downstream sequences. In addition, the patterns of codon usage and the ratio of transitions/transversions resulting from SNPs were also calculated. SnpEff is a command line software tool, but before it can be used it is necessary to modify the config file called snpEff.config in the following way:

```
$ cp ~/jars/snpEff/snpEff.config <My project folder>/
$ vim <My project folder>/snpEff.config
```

The below information has to be inserted in the snpEff.config file so that SnpEff is aware of the chromosome sequences.

```
# Databases are stored here
data_dir = <My project folder>
# Databases & Genomes
# My project name
Chr1.genome : Chr1
Chr2.genome : Chr2
ChrN.genome : ChrN
```

After the configuration file has been saved it is possible to execute *SnpEff.jar* (Figure 2.22).

```
$ java  -jar snpEff.jar -h
snpEff version SnpEff 3.5h (build 2014-04-01), by Pablo Cingolani
Usage: snpEff [eff] [options] genome_version [input_file]

        variants_file                   : Default is STDIN

Options:
        -a , -around: Show N codons and amino acids around change (only in
                      coding regions). Default is 0 codons.
        -chr <string>: Prepend 'string' to chromosome name (e.g. 'chr1' instead
                       of '1'). Only on TXT output.
        -download: Download reference genome if not available. Default: false
        -i <format>: Input format [ vcf, txt, pileup, bed ]. Default: VCF.
        -fileList: Input actually contains a list of files to process.
        -o <format>: Ouput format [ txt, vcf, gatk, bed, bedAnn ]. Default: VCF.
        -s , -stats: Name of stats file (summary). Default is
                     'snpEff_summary.html'
        -noStats: Do not create stats (summary) file
        -csvStats: Create CSV summary file instead of HTML

Sequence change filter options:
        -del: Analyze deletions only
        -ins: Analyze insertions only
        -hom: Analyze homozygous variants only
        -het: Analyze heterozygous variants only
        -minQ X, -minQuality X: Filter out variants with quality lower than X
        -maxQ X, -maxQuality X: Filter out variants with quality higher than X
        -minC X, -minCoverage X: Filter out variants with coverage lower than X
        -maxC X, -maxCoverage X: Filter out variants with coverage higher than X
        -nmp: Only MNPs (multiple nucleotide polymorphisms)
        -snp: Only SNPs (single nucleotide polymorphisms)

Results filter options:
        -fi , -filterInterval  <file>: Only analyze changes that intersect with
                                       the intervals specified in this file (you
                                       may use this option many times)
        -no-downstream: Do not show DOWNSTREAM changes
        -no-intergenic: Do not show INTERGENIC changes
        -no-intron: Do not show INTRON changes
        -no-upstream: Do not show UPSTREAM changes
        -no-utr: Do not show 5_PRIME_UTR or 3_PRIME_UTR changes

Annotations options:
        -cancer: Perform 'cancer' comparisons (Somatic vs Germline).
                 Default: false
        -cancerSamples <file>: Two column TXT file defining 'oringinal \t
                               derived' samples.
        -geneId: Use gene ID instead of gene name (VCF output). Default: false
        -hgvs: Use HGVS annotations for amino acid sub-field. Default: false
        -lof: Add loss of function (LOF) and Nonsense mediated decay (NMD) tags.
        -oicr: Add OICR tag in VCF file. Default: false
        -sequenceOntolgy: Use Sequence Ontolgy terms. Default: false

Generic options:
        -c , -config: Specify config file
        -d , -debug: Debug mode (very verbose).
        -dataDir <path>: Override data_dir parameter from config file.
        -h , -help: Show this help and exit
        -if , -inOffset: Offset input by a number of bases.
                         E.g. '-inOffset 1' for one-based TXT input files
        -of , -outOffset: Offset output by a number of bases. E.g.
                          '-outOffset 1' for one-based TXT output files
        -noLog: Do not report usage statistics to server
        -t: Use multiple threads (implies '-noStats'). Default 'off'
        -q ,  -quiet: Quiet mode (do not show any messages or errors)
        -v , -verbose: Verbose mode
```

```
Database options:
        -canon: Only use canonical transcripts.
        -interval: Use a custom intervals in TXT/BED/BigBed/VCF/GFF file (you
                   may use this option many times)
        -motif: Annotate using motifs (requires Motif database).
        -nextProt: Annotate using NextProt (requires NextProt database).
        -reg <name>: Regulation track to use (this option can be used add
                     several times).
        -onlyReg: Only use regulation tracks.
        -onlyTr <file.txt>: Only use the transcripts in this file. Format: One
                            transcript ID per line.
        -ss , -spliceSiteSize <int>: Set size for splice sites (donor and
                                     acceptor) in bases. Default: 2
        -ud , -upDownStreamLen <int>: Set upstream downstream interval length
                                      (in bases)
```

Figure 2.22: The command-line of the *snpEff.jar*, showing the various usage options.

2.2.2.14. Gene analysis

The *gene_analysis.py* script (Figure 2.23) identifies genes in low SNP density regions and creates the following files:

- A file with the suffix "_SNPID_geneID_GO_LOW.tab" is a tabular separated file which contains (i) low SNP id, (ii) gene Id and (iii) GO terms
- The population and study are gene lists with one gene ID per line, which include the genes you want to compare. In our case the study file contains genes in low SNP density regions compared with the population of all genes.
- The association file contains the gene to GO term mapping which is a two-column tabular file, (i) geneID and (ii) GO terms (separated by ";" if there are multiple terms).

The population, study and association file are used for the *SGSautoSNP_summary.py* script to identify over-representation and under-representation of certain GO terms using goatools (https://github.com/tanghaibao/goatools).

```
$ python gene_analysis.py  -h
Usage: gene_analysis.py --bam <alignments.bam> --pred <snap_uniref.gff3> --loc
<contig.gff3> --snp <location.snp> --fasta <multiple.fasta> -g
<output_gene_seq.fasta> [-w <bin_size] [-s <step_size>] [-c
<coverage_cutoff_percentage] [--low <lower_coverage_limit>] [--up
<upper_coverage_limit>]

Options:
  -h, --help            show this help message and exit
  --bam=BAM_FILE_NAME   input bam file
  --pred=PRED_GENES_FILE_NAME
                        input gff file
  --loc=LOC_FILE_NAME   Optional. Input contig location gff file
  --snp=SNP_FILE_NAME   input SNP location file
  --fasta=FASTA_FILE_NAME
                        input multiple fasta file
  --dir=DIR             Please give provide output dir full path.
  --output_filename=OUTPUT_FILENAME
                        Is file name template and script extend it.
  --snp_cut=SNP_CUT     input SNP cutoff [default=2]
  --ud=UD               gene loci up and downstream amount [default=5000].
                        E.g. for --ud 5000 a gene locus will include 5000
                        bases upstream and 5000 bases downstream
  -w BIN_SIZE           input window size [default=10000]
  -s STEP_SIZE          input step size [default=1000]
  -l LLIM, --low=LLIM   input lower coverage limit [default=4]
  -u ULIM, --up=ULIM    input upper coverage limit [default=20]
  -c COV_CUT, --cov_cut=COV_CUT
                        input the percentage coverage cutoff value
                        [default=20]
```

Figure 2.23: The command-line of the *gene_analysis.py* script, showing the various usage options.

2.2.2.15. SGSautoSNP summary

The *SGSautoSNP_summary.py* script was designed to combine results from all chromosomes in order to provide a quick overview. It also uses goatools (https://github.com/tanghaibao/goatools) which is a Python library to process over- and under-representation of certain GO terms, based on Fisher's exact test. Goatools also provides several correction routines including Bonferroni, Sidak, and false discovery rate (FDR). GO is part of a larger classification effort, the Open Biomedical Ontologies (OBO), which has defined a formalised ontology for cross-species classification for genes and their protein products. GO is a structured and controlled vocabulary with a defined file-format which attempts to achieve the following goals:

- Human readability
- Ease of parsing

- Extensibility
- Minimal redundancy

Furthermore it is a formal declaration of legal relationships between terms in the ontology (Figure 2.24, http://www.geneontology.org/GO.ontology-ext.relations.shtml). The OBO file format can be downloaded from:

http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology.1_2.obo

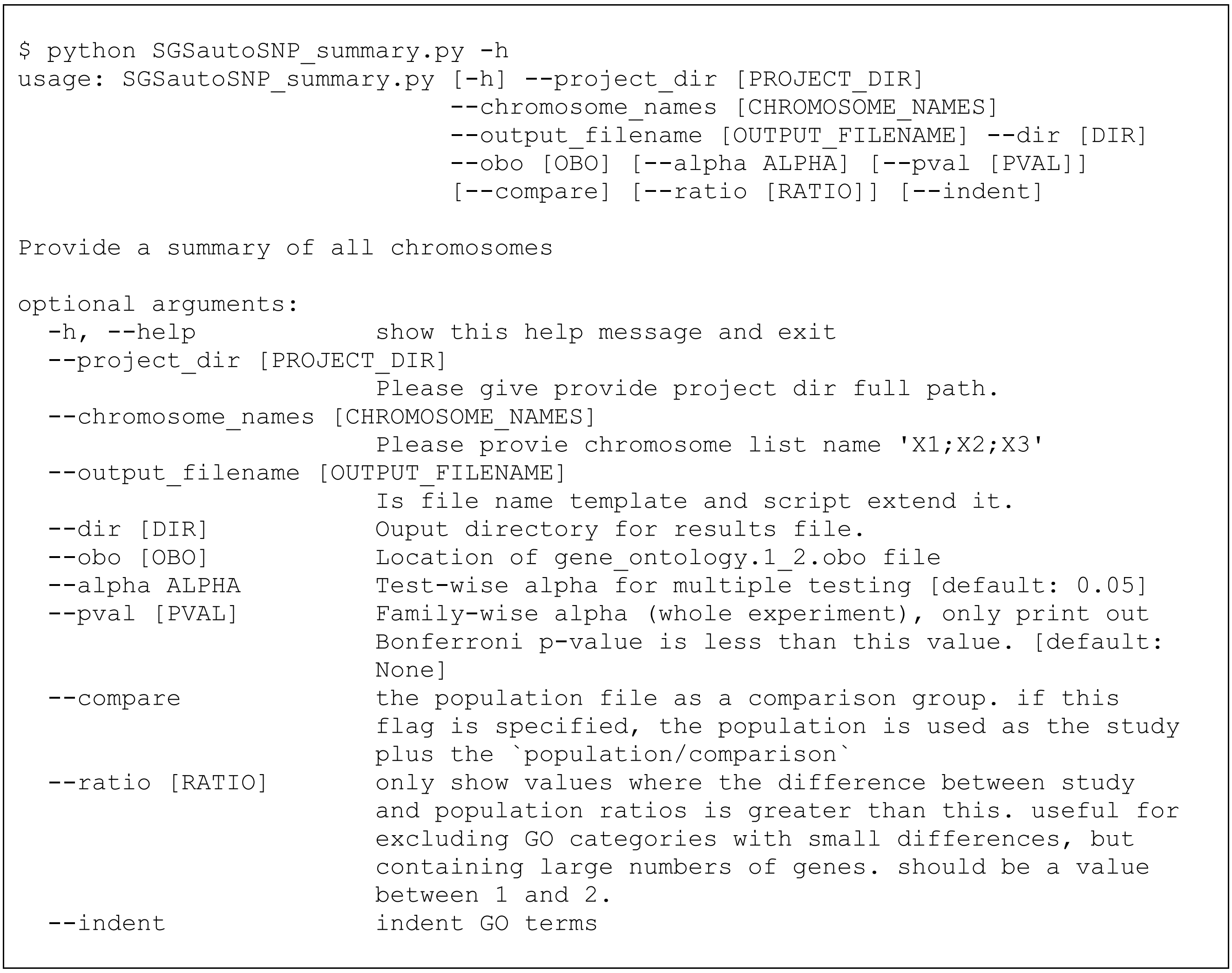
```
[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
alt_id: GO:0050876
def: "The production by an organism of new individuals that contain
some portion of their genetic material inherited from that organism."
[GOC:go_curators, GOC:isa_complete, ISBN:0198506732 "Oxford Dictionary
of Biochemistry and Molecular Biology"]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "reproductive physiological process" EXACT []
is_a: GO:0008150 ! biological_process
```

Figure 2.24 shows an example of a GO term from OBO 1.2 file format which is used by goatools.

SGSautoSNP_summary.py creates three output files:

- A file with this suffix "_snps_similarity_matrix.tab" which summarises all similarity matrices in to one.
- Another file with the suffix "_snpseffs.tab" creates a table of SnpEff chromosome results.
- The last output contains a table for all chromosomes gene enrichment results.

It is possible to run *SGSautoSNP_summary.py* on a modern PC with the following command (see Figure 2.25).



```
$ python SGSautoSNP_summary.py -h
usage: SGSautoSNP_summary.py [-h] --project_dir [PROJECT_DIR]
                             --chromosome_names [CHROMOSOME_NAMES]
                             --output_filename [OUTPUT_FILENAME] --dir [DIR]
                             --obo [OBO] [--alpha ALPHA] [--pval [PVAL]]
                             [--compare] [--ratio [RATIO]] [--indent]

Provide a summary of all chromosomes

optional arguments:
  -h, --help            show this help message and exit
  --project_dir [PROJECT_DIR]
                        Please give provide project dir full path.
  --chromosome_names [CHROMOSOME_NAMES]
                        Please provie chromosome list name 'X1;X2;X3'
  --output_filename [OUTPUT_FILENAME]
                        Is file name template and script extend it.
  --dir [DIR]           Ouput directory for results file.
  --obo [OBO]           Location of gene_ontology.1_2.obo file
  --alpha ALPHA         Test-wise alpha for multiple testing [default: 0.05]
  --pval [PVAL]         Family-wise alpha (whole experiment), only print out
                        Bonferroni p-value is less than this value. [default:
                        None]
  --compare             the population file as a comparison group. if this
                        flag is specified, the population is used as the study
                        plus the `population/comparison`
  --ratio [RATIO]       only show values where the difference between study
                        and population ratios is greater than this. useful for
                        excluding GO categories with small differences, but
                        containing large numbers of genes. should be a value
                        between 1 and 2.
  --indent              indent GO terms
```

Figure 2.25: The command-line of the SGSautoSNP_summary.py script, showing the various usage options.

2.3. Results and Discussion

SOAPaligner provides an option (-r) on how to report repeat hits (0=none; 1=random one; 2=all). This option aims to increase SNP calling accuracy by ignoring read pairs that cannot be accurately positioned on the reference. Ningyou culivar paired-end reads (80 bp) were mapped against our *B. napus* reference genome with different values for –r option in order to find out how this option effect the number of mapped reads (see Table 2.2).

Table 2.2 shows *SOAPalinger*'s -r option benchmark. Ningyou culivar paired-end read (80 bp) was mapped against *B. napus* reference genome.

	Option -r		
	0	**1**	**2**
Total Pairs	35638181	35638181	35638181
Paired	681924 (1.91%)	1522645 (4.27%)	1522645 (4.27%)
Singled	18923834 (26.55%)	40722853 (57.13%)	40722853 (57.13%)
Total Elapsed Time	5756.12 sec	5878.56 sec	9397.81 sec
Load Index Table	11.79 sec	12.75 sec	14.09 sec
Alignment	5744.33 sec	5865.81 sec	9383.72 sec

SOAPaligner.py performance for mapping six *B. napus* cultivars against our *B. napus* reference is shown in Table 2.4. *GenerateSubsetBAM.py* required 02:05:31 to extract all 10 Brassica AA chromosomes from the alignments created by *SOAPaligner.py*. *MergeChr.py* performance is presented in Table 2.4 for merging six cultivar libraries for 10 *Brassica* AA chromosomes.

SGSautoSNP.py needs, on average, 20 min to process each *Brassica* AA chromosome on a 4 CPU core computer to predict SNPs (see Figure 2.26). The higher CPU core speed up is not linear. The performance of other SGSautoSNP pipeline scripts can be found in Table 2.5. The final SGSautoSNP pipeline script *SGSautoSNP_summary.py* , just takes 00:01:19 to created the whole project summary.

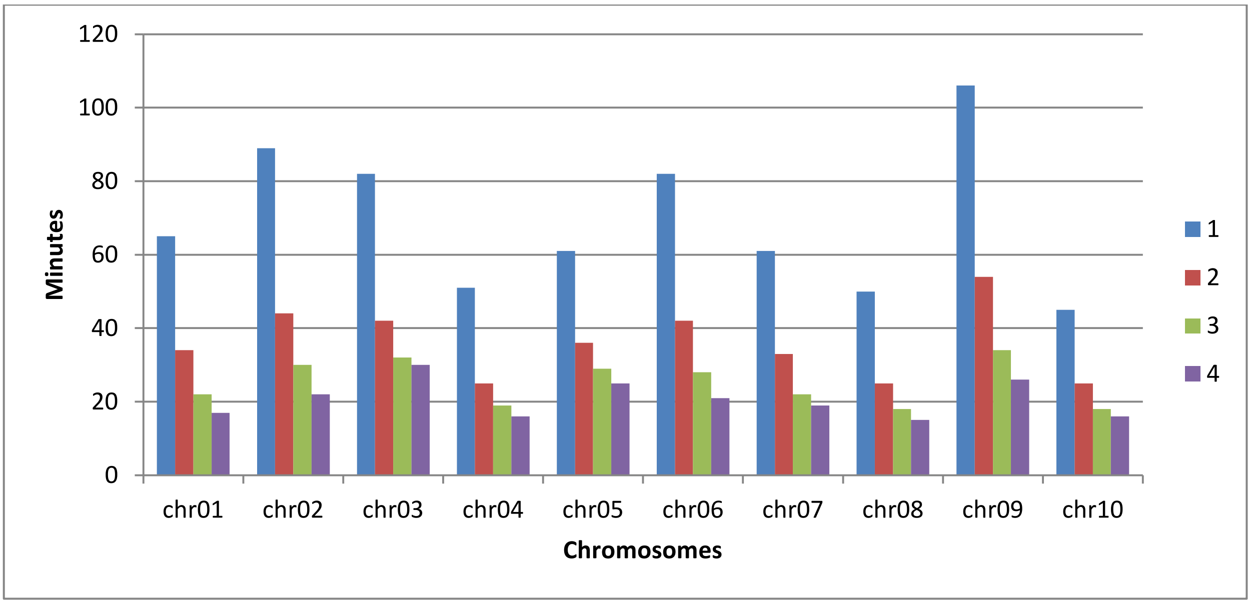


Figure 2.26: Benchmark of *SGSautoSNP.py* across all 10 *Brassica* AA chromosomes. Each chromsome was run on 1 to 4 CPUs cores.

Table 2.3: Number of contigs for all 10 *Brassica* AA chromosome.

Chromosome names	Number of contigs in each chromosome
chr01	29
chr02	19
chr03	12
chr04	20
chr05	22
chr06	16
chr07	11
chr08	19
chr09	37
chr10	7

Table 2.4 shows the performance of *SOAPaligner.py* for mapping six *B. napus* cultivars against our *B. napus* reference. Additional *MergeChr.py* performance is presented here for merging six cultivar libraries chromosomes across 10 *Brassica* AA chromosomes.

	Ag Spectrum (A)	BLN (Bn)	Ningyou (N)	Skipton (S)	Surpass (Sr)	Tapidor (T)
SOAPaligner.py	02:40:12	04:09:36	26:31:48	02:43:12	04:37:12	19:15:00
MergeChr.py	00:06:04	00:09:29	01:00:14	00:06:11	00:10:30	00:43:42

Table 2.5 shows seven SGSautoSNP pipeline script performance for 10 *Brassica* AA chromosomes.

	Chr01	Chr02	Chr03	Chr04	Chr05	Chr06	Chr07	Chr08	Chr09
bam2consensus_seqs.py	00:06:31	00:07:11	00:09:52	00:05:50	00:08:40	00:07:40	00:07:27	00:06:40	00:09:56
create_flapjack_files.py	00:00:46	00:00:55	00:01:10	00:00:59	00:00:50	00:00:70	00:00:54	00:00:38	00:00:58
filter_snps.py	00:01:29	00:01:36	00:03:06	00:02:47	00:02:59	00:02:58	00:02:57	00:03:04	00:03:27
gene_analysis.py	00:01:59	00:02:25	00:04:01	00:02:14	00:03:03	00:03:26	00:04:41	00:03:07	00:04:42
gene_annotation.py	03:26:57	03:25:25	03:50:52	02:50:17	02:55:48	03:09:17	03:02:57	02:51:40	03:26:25
multiple_to_single_fasta.py	00:01:20	00:00:59	00:00:50	00:00:40	00:00:30	00:00:30	00:00:30	00:00:20	00:00:20
snpEff.jar	00:00:59	00:01:06	00:01:31	00:02:56	00:02:08	00:01:57	00:02:08	00:02:44	00:01:22

2.4. Conclusion

SGSautoSNP is a SNP discovery and annotation pipeline and has been successfully applied to Second Generation Sequence datasets for *Brassica* and wheat group 7 chromosomes. These results are described in detail in the following two chapters. This pipeline can be used on any plant species for which Second Generation Sequence datasets exists. Furthermore, the pipeline can handle different data coming from different sequencing technologies as long they can produce BAM files. SOAP aligner could be replaced by any other aligner for short reads, such as BWA.

It turns out to ignore all repeats with –r 0 option also decreases the alignment time (5744.33 seconds) compared to report all repeats with –r 2 option (9383.72 seconds) (see Table 2.2). There is no difference in percentage of align paired and single reads (4.27% and 57.13%, respectively) by using –r 1 or 2, but there is big difference compare to the –r 0 option (1.91% and 26.55%, respectively). The other observation is that only a fraction of reads were able to align as a pair in order to provide a greater confidence of specific and accurate read mapping, because a single read could possibly match many positions, but two reads separated by a gap of defined approximate insert size are more likely to match to correct position. In other research reads have been trimmed from both the 5' and 3' ends until reaching a base with PHRED score greater than 20 and allowing at most two Ns in each read (Yu et al., 2012). As result they found that the number of reads for alignments has decreased after trimming, but more reads align.

Figure 2.26 shows that *SGSautoSNP.py* almost achieved a linear speeup on 3 CPUs cores for some chromosomes. However, using 4 CPU cores the speedup declined for all chromosomes. This is caused by a small number of contigs on each chromosomes and their length (see Table 2.3). Three workers finished, but one worker is still working on a contig which is longer and has more coverage.

Future work could include simplifying running the SGSautoSNP pipeline, because some users are unable to write PBS scripts. A solution would be to integrate the SGSautoSNP pipeline into a general bioinformatics workflow management system such as Galaxy Tool Shed (Blankenberg et al., 2014) and Yabi (Hunter et al., 2012). These systems have been designed for research scientists who do not have computer programming experience.

The SNPs identified using SGSautoSNP can be used for genotyping by sequencing, using low coverage skim sequencing of segregating populations, and calling genotypes where the low coverage sequence data aligns to a previously predicted polymorphic position. In addition, the software could be extended for applications in species which demonstrate heterozygosity.

Chapter 3: Application of SGSautoSNP in Brassica

3.1. Introduction

Brassica rapa has a diploid genome and *B. napus* has an allotetraploid genome. These genomes are both large, complex and contain many repetitive elements. These factors make it difficult to sequence and to discover SNPs in *Brassica* species for crop improvement. SNP discovery from Second-Generation Sequencing technologies is challenging due to short reads and high error rates. It is difficult to distinguish between real SNPs and sequence errors. However, we were able to reduce this issue with help of SGSautoSNP algorithm and other software included in this pipeline which is described in detail in Chapter 2.

SNP discovery has previously been performed in *B. napus*: Trick et al. discovered 41,593 putative SNPs (1 SNP/1.2 kb) between the cultivars Tapidor and Ningyou 7, through generating approximately 20 million expressed sequence tags (ESTs) from each of these two cultivars, sequenced using the Illumina Solexa platform (Trick et al., 2009). As a reference sequence they used approximately 94,000 *Brassica* species unigenes. From the detected SNPs 87.5 - 91.2% were ‘hemi-SNPs’ which are inter-varietal. In a different EST SNP discovery project a total of 604 SNPs were identified, one SNP in every 42 bp (Durstewitz et al., 2010). For this SNP analysis 100 amplicons derived from ESTs from *B. napus* varieties were compared. In the *B. rapa* genome 21,311 SNPs were discovered between 8 genotypes from re-sequencing 1,398 sequence-tagged sites (STSs). The SNP frequency was one SNP every 103 bp in exons and one SNP per 54 bp in introns (Park et al., 2010).

A total of 20,835 SNPs, one SNP every 446 bp, were discovered in eight *B. napus* inbreds, in 113,221 restriction site associated DNA markers (RAD) clusters from a *Kpn*I library (Bus et al., 2012). Huang et al. chose samples from B. napus accessions which were parents of reference mapping populations or elite cultivars (Zhongshuang11, 73290, 08-806-2, 09CB01, Tapidor, XY15, 09CB03, PY-2, Westar, PY-1). These reads were aligned against the reference *B. rapa* and *B. oleracea* sequences using SOAP2. As result they discovered 892,536 SNPs by excluding 6,331,887 SNPs that were heterozygous in at least in one individual in the genome. In 13,552 predicted genes a total of 36,458 non-synonymous

SNPs were predicted (Huang et al., 2013b). A total of 505 non-synonymous SNPs transformed amino acid codons to stop codons, whereas non-synonymous SNPs transformed stop codons to amino acid codons and their validation rate was 92% using the GoldenGate genotyping platform (Huang et al., 2013b). More than 200,000 SNPs were discovered between *B. rapa* and three oleiferous lines in RNA sequence data using the Illumina GAIIX sequencer (Paritosh et al., 2013).

The SGSautoSNP pipeline uses SnpEff for annotation and to predict the effect the SNPs on genes. This was applied here to the 10 *Brassica napus* AA genome chromosome sequences. SNPs were categorised on the basis of their structural occurrence in the exons, introns and intergenic region. Furthermore, the functional relevance of the SNPs was predicted. Proteins have a unique amino acid sequence which is specified by the DNA coding sequence. Changes to this sequence could influence protein function. Non-synonymous SNPs (nonsense or missense) change the amino acid sequence of a protein. Therefore these SNPs could have an important functional relevance to the trait studied. On the other hand synonymous SNPs do not change the codon sequence. These SNPs could modulate translation rates and protein folding and impact the protein function (Zhang et al., 2014).

3.1.1. Project aims

In this chapter the application of SGSautoSNP (Second-Generation Sequencing AutoSNP), a SNP discovery pipeline described in chapter 2, is presented for SNP discovery in the *Brassica* AA genome. These SNPs are valuable for detailed diversity analysis, marker assisted selection and genotyping by sequencing. The results in this chapter have been generated using the new SGSautoSNP pipeline, which contains more features than the orginal version published in 2012 (Lorenc et al., 2012). This include scripts for gene annotation, which uses SNAP (Korf, 2004) a gene prediction tool and SNPeff (Cingolani et al., 2012) a SNP annotation and effect prediction tool. This chapter demonstrates that the SGSautoSNP pipeline is suitable for high resolution SNP discovery and annotation and can be applied to other large and complex genomes datasets.

3.2. Material and Methods

Six *B. napus* cultivars: Ag Spectrum, BLN, Ningyou, Skipton, Surpass and Tapidor, with sequence coverage between 7.2× and 71.5×, were used for SNP discovery. Ningyou and Tapidor were sequenced on Illumina HiSeq 2000 and the others were sequenced on Illumina GAIIx.

At the start of this analysis we did not have a public *B. napus* reference genome and therefore we used the public *B. rapa* sequence (Wang et al., 2011) AA genome combined with Bayer's proprietary *B. oleracea* CC genome. Table 3.1 provides information about the chromosome and genome sizes for the *B. rapa* AA and *B. oleracea* CC genomes which were used to align the reads from the six cultivars.

The SGSautoSNP pipeline was used with default settings (see Chapter 2), except for *SOAPaligner.py* which requires the insert-size for each library. These insert-sizes can be found in Table 3.3. The pipeline was run on the Barrine computer cluster at the University of Queensland (see Appendix).

Since this work was completed two public *B. oleracea* CC genomes have been published (Liu et al., 2014, Parkin et al., 2014) and the public *B. napus AACC* genome is expected to be published in 2014. This *B. napus* genome would be more appropriate to use in future studies than the diploid progenitors.

Table 3.1: Chromosome and genome sizes for *Brassica napus* AA and CC genomes used in this study for sequence read alignment.

Chromosomes	AA (bp)	CC (bp)
chr01	26,740,857	18,290,447
chr02	27,846,329	14,513,690
chr03	32,228,999	25,073,557
chr04	20,225,473	14,202,440
chr05	23,939,834	17,430,407
chr06	26,271,742	2,917,136
chr07	22,304,823	18,811,192
chr08	21,231,227	13,194,272
chr09	37,194,012	9,244,411
chr10	17,624,101	-
Genome size	**255,607,397**	**133,677,552**

3.3. Results and Discussion

3.3.1. Reads mapping

Paired read data, between 9.38 and 67.51 Gbp, for six *B. napus* cultivars were generated (Table 3.2). These paired reads were analysed using the SGSautoSNP pipeline and mapped onto the *B. napus* reference genome which was created from the *B. rapa* and *B. olerecea* genomes. For the *B. napus* (AACC) reads it was necessary that the reference contains *B. rapa* and *B. oleracea* genomes, because if the C genome was absent from the reference, CC genome specific reads could map to the AA genome, which could confound SNP discovery. Only paired reads mapping to a unique location in the genome were kept for further analysis, which is guaranteed by SOAPaligner parameter (-*r 0*) (Li et al., 2009b). This option aims to increase SNP calling accuracy by ignoring read pairs that cannot be accurately positioned on the reference. Similarly, only reads that mapped as a pair were used for SNP discovery. Due to the short length of the reads, one read could match at many positions, but two reads separated by a gap of defined insert size provides a greater confidence of specific and accurate read mapping. Table 3.2 shows the results for all six *B. napus* cultivars paired-read mapping and Table 3.3 shows the minimum and maximum insert sizes used for SOAPalinger. Of these reads, between 4.76% and 7.72% mapped to the *B. rapa* genome and 5.64% and 9.62% mapped to the *B. oleracea* genome.

Table 3.2: Summary of *Brassica napus* cultivar data and mapping against the *B. napus* reference, which was made out of the *B. rapa* (AA) and *B. oleracea* (CC) genomes. The variety column contains in brackets the cultivar name abbreviation. The mapping information has been split into AA and CC genomes. The table also contains the growth habit and origin of the *B. napus* varieties.

***B. napus* variety**	**National origin**	**Growth habit**	**Data generated**	**Data mapped to AA**	**Read pairs mapped (AA)**	**Data mapped to CC**	**Read pairs mapped (CC)**
Ag Spectrum (A)	Australia	Spring	9.38 Gbp	0.52 Gbp	5.59%	0.62 Gbp	6.58%
BLN (Bn)	Australia	Spring	14.60 Gbp	1.13 Gbp	7.72%	1.40 Gbp	9.62%
Ningyou (N)	China	Spring	93.06 Gbp	4.57 Gbp	4.91%	6.10 Gbp	6.56%
Skipton (S)	Australia	Spring	9.55 Gbp	0.47 Gbp	4.88%	0.54 Gbp	5.64%
Surpass (Sr)	Australia	Spring	16.22 Gbp	0.77 Gbp	4.76%	1.05 Gbp	6.50%
Tapidor (T)	France	Winter	67.51 Gbp	3.98 Gbp	5.90%	6.39 Gbp	9.47%

Table 3.3: The minimum and maximum insert sizes used during the alignment with SOAPalinger for the six *B. napus* cultivars paired-reads.

***B. napus* variety**	**Minimum insert sizes (bp)**	**Maximum insert sizes (bp)**
Ag Spectrum (A)	350 - 2500	500 - 4000
BLN (Bn)	350	500
Ningyou (N)	120 - 410	350 - 820
Skipton (S)	400 - 2500	600 - 4000
Surpass (Sr)	350 - 390	510 – 550
Tapidor (T)	50 - 5600	280 - 11000

Figure 3.1 shows that for all 10 A genome chromosomes most read positions have coverage equal to or more than 4 and only a few have coverage between 1 and 3. However, the number of unmapped reads across all chromosomes is similar, except for chromosome 4, 9 and 10. Table 3.4 shows the number of bases in the genome that have coverage of at least 4 reads. SGSautoSNP requires at least two reads, each from at least two cultivars to call a SNP, the minimum coverage at a locus to call a SNP is therefore four. In tomato, (Causse et al., 2013) used a minimum coverage of eight reads to call SNPs by restricting the read coverage to eight or greater, they lost several SNPs previously detected by Sanger sequencing (Ranc et al., 2012). This confirms that it is a good approach to a minimum coverage of four to call SNPs for SGSautoSNP and allows analysis of 91.07 - 93.66% of total reads mapped to the reference (see Figure 3.1 and Table 3.4). To increase confidence in the SNP calling it is recommended to validate some of the SNPs for example with a GoldenGate Genotyping Assay (Durstewitz et al., 2010).

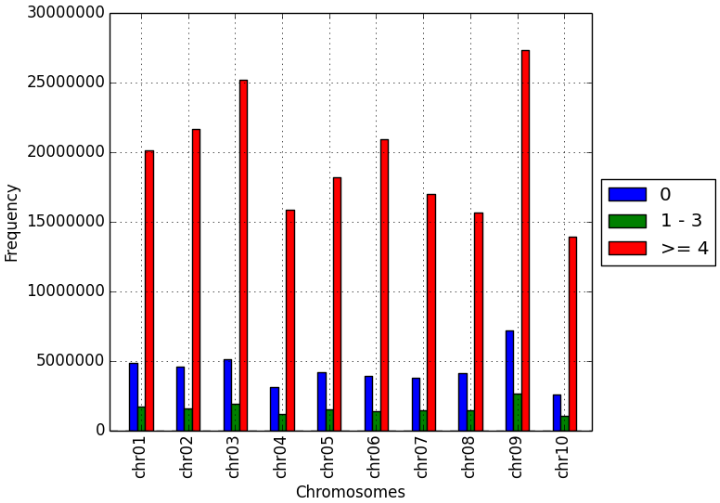


Figure 3.1: The depth of coverage of the Illumina reads at 0, 1 - 3 and >= 4 reads are shown for the assembled *Brassica* AA genome.

Table 3.4: The number of reads, with a minimum coverage of 4 that are mapped on each chromosome.

	Coverage more than 4	
Chromosome	**Reads no.**	**Reads no. in %**
A01	20145576	92.13
A02	21668216	93.17
A03	25192374	92.90
A04	15871687	92.83
A05	18216551	92.34
A06	20929709	93.66
A07	17014836	92.05
A08	15636483	91.48
A09	27312495	91.07
A10	13944935	92.83

3.3.2. Single Nucleotide Polymorphisms calling

The SNP discovery was performed only on the *B. napus* AA genome, as the CC genome was proprietary. Using the SGSautoSNP pipeline a total of 638,593 SNPs were identified across the 10 chromosomes of the AA genome, between six *B. napus* cultivars (Lorenc et al., 2012). SGSautoSNP provides a SNP score (the polymorphism must be present in a minimum of two sequence reads and is in detailed described in Chapter 2) which is a measure of confidence in SNP prediction. In this study, the SNP score ranged from 2 to 133, with an average of 7.88.

Figure 3.2 shows that most SNPs, for all 10 chromosomes in the AA genome, had coverage between 4 and 50 reads. It also shows that all 10 chromosomes have a similar curve and trend in the number of SNPs. Some SNPs had high coverage of between 118 (chromosome 10) to 381 (chromosome 7). Higher levels of coverage and the addition of more cultivars could identify more SNPs or remove previously discovered SNPs. Losing SNPs could happen because SGSautoSNP rejects a SNP if not all bases within each cultivar at a locus are the same, which is expected for homozygous genomes, and these instances are more likely to be observed with higher read coverage.

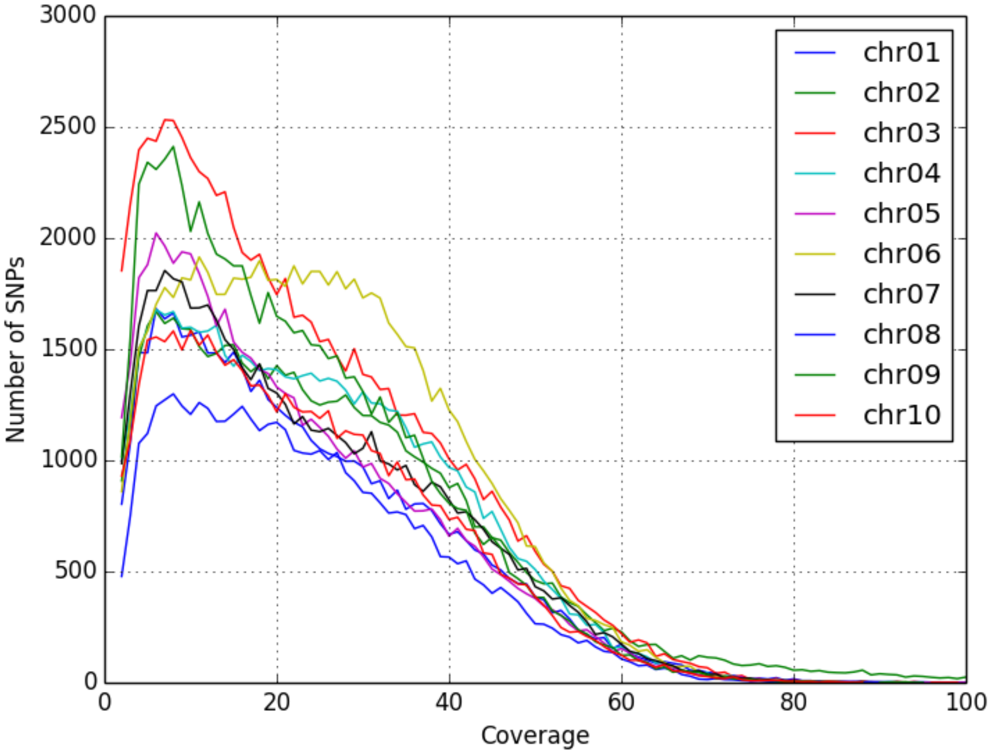


Figure 3.2: A graphical representation of the relationship between coverage and SNPs for all 10 chromosomes in the *Brassica AA* genome.

A similarity matrix was created with the number of SNPs between lines between all six cultivars in the A genome (Table 3.5). Most of the pairwise SNPs (378,652) were called between the cultivars Ningyou (N) and Tapidor (T). This may be because they are very diverse lines; Tapidor is a French winter cultivar type (i.e. it has a strong vernalisation requirement) and Ningyou 7 is a Chinese spring cultivar (i.e. it has no vernalisation requirement) (Table 3.2 and (Trick et al., 2009)). Additionally, both these cultivars had higher levels of sequence coverage. On the other hand, the lowest the number of pairwise SNPs (63,409) were called between the cultivars BLN (Bn) and Skipton (S). These Australian cultivars are both Australian spring types and therefore there may be little diversity between them (Table 3.2).

Table 3.5: The number of pairwise SNPs between the 6 cultivars in the 10 chromosomes of the *B. napus* AA genome.

	Ag Spectrum	BLNBn	Ningyou	Skipton	Surpass	Tapidor
Ag Spectrum	0	90781	228647	69498	104018	178444
BLN		0	295025	63409	115434	207930
Ningyou			0	228546	269133	378652
Skipton				0	101992	168265
Surrpass					0	177828
Tapidor						0

3.3.3. Single Nucleotide Polymorphisms validation

In a previous project (Dalton-Morgan et al., 2014), SNPs were discovered by the SGSautoSNP pipeline. These SNPs were validated by Sanger sequencing of PCR products and on a high-density, 6 K Infinium™ array for *B. napus*. This array is also able to characterise the diploid *Brassica* genomes, *B. rapa*, *B. oleracea* and *B. nigra*. Sequence libraries for *B. napus* were prepared for the Australian cultivars Ag-Spectrum, BLN2672, Skipton, Surpass 400 using the Illumina’s Genomic DNA Sample Prep Kit according to the manufacturer’s instructions. These cultivars were sequenced with the Illumina GAIIx platform to generate paired-end sequence reads between 75 and 100 bp length, with a coverage over the four varieties averaging 9.9X (Table 3.6).

Table 3.6: Sequence data used for SNP discovery (Dalton-Morgan et al., 2014)

Sample	Read number	Total read length (Gbp)	Estimated total read depth
Skipton	104,368,328	9.55	7.96×
Ag-Spectrum	103,918,490	9.38	7.81×
BLN2672	109,825,900	12.441	10.37×
Surpass 400	110,431,492	16.22	13.52×

SGSautoSNP discovered 871,806 SNPs between four cultivars with an average of one SNP per 730 bases. Of theses SNPs 498,759 were transitions (A>G or C>T) and 375,340 were transversions (A>C, A>T, C>G or G>T). However, the A genome contains 196,451 transitions and 152,956 transversions.

Initial validation of the SNPs predicted in this study was performed on 20 random selected SNPs using Sanger sequencing of PCR products and the SNP prediction accuracy was exacty 95% (Table 3.7; (Dalton-Morgan et al., 2014)). The validated SNPs all had SNP scores greater than 2, but for the one heterozygote the SNP score was 2. More extensive validation was performed using a *B. napus* 6k Infinium™ array. After the SNP prediction with the SGSautoSNP pipeline, the following filters were applied to the SNPs to build the 6k Infinium™ array:

- Within 60 bp on either side of the SNP position should be no other SNP
- For this anlaysis we only used SNP where sequence information was available for all cultivars to avoid bias from missing data
- In order to avoid the likelihood of rare alleles, SNPs were selected where the minor allele was present in more than one cultivar
- Illumina's Assay Design Tool (ADT) score has to be greater than 0.6
- All A>T and C>G tranversions were removed in order to maximise the number of positions assayed on the array, as these SNPs require two probes per locus to assay, as compared to transitions which only require one probe

The 6k Infinium™ array (*Brassica_napus_UQEvie_6k_11581453*) contains 5,306 SNPs evenly distributed over the entire A and C genomes of *B. napus*. Of these SNPs, 3,706 (69.9 %) were transitions and 1,600 (30.1 %) were tranversions. Due to poorly separated clusters, 186 (3.5 %) of the total 5,306 SNPs failed. Of these, 69 (37.1 %) were located on the A genome. After ignoring 283 (5.5 %) monomorphic SNPs out of 5,120 SNPs across the assayed samples, 4,837 (94.5 %) SNPs were successfully predicted (Dalton-Morgan et al., 2014).

Table 3.7: Summary of SNP validation (Dalton-Morgan et al., 2014).

SNP name	Forward primer	Reverse Primer	SNP score	Validation
1	GTTGGGTGGGACTAGAAAC	GCATGGAAGGCAACAC	6	True SNP
2	CTTTTAACAGTAAAGAGGGATC	GTGAGCTCCTTTCTATTTT	4	True SNP
3	CTCTTTCATTCTCCTCCATGG	AAGTATTCATAGTAAACCGAT	4	True SNP
4	CGTCATCTTCGCTTTAGGCCT	TCAAGTTTTCCTCACCAAA	4	True SNP
5	CAATGTCTTTAGCATCGTTAC	GTTAATTATTGTTCTTGTTCA	4	True SNP
6	CTCAGCCTCCTGCTCCTCAG	AGTGAGAGGGTTTTGACTCTT	4	True SNP
7	GCACCACTAATCAAACTTACCA	GTATTTCAAATGCAGAGAGATC	4	True SNP
8	CAATCCTGTAATCATAATATATGT	CAAACCCATTGATAAGTATTC	5	True SNP
9	TGCAAGCTCAGGCTCTCTTC	CAAGTTACCATCTTTAGCATC	5	True SNP
10	TCTAGTTTTGTTACTCTTGAA	AAATCACAGTACGGCGTCCC	5	True SNP
11	ACAGATCAAGCAGAACTACAGCA	CCTCATTGGTAACAAGTCTG	4	True SNP
12	AAACCATCCCTTTGTTTTCAAT	ATTATCCCAGACATTGATGAG	4	True SNP
13	TGATCGATCTATCTCTCGGT	TAACTAGACCAAAGTGAGTAG	4	True SNP
14	CACCTCGGGATAGTCCTC	GATGTGTGGGAGATGTTCAAG	22	True SNP
15	CATCCGTGTACATACTAAGAAC	GTATGGAAACTACAAACCAGC	15	True SNP
16	CTCGCTGAGGTAAGCTGAC	CGAATTATAGCTGCTCCACTC	6	True SNP
17	CTCGCTGAGGTAAGCTGAC	CGAATTATAGCTGCTCCACTC	2	Failed
18	CTCGCTGAGGTAAGCTGAC	CGAATTATAGCTGCTCCACTC	3	True SNP
19	CTCGCTGAGGTAAGCTGAC	CGAATTATAGCTGCTCCACTC	12	True SNP
20	CTCGCTGAGGTAAGCTGAC	CGAATTATAGCTGCTCCACTC	8	True SNP

3.3.4. Single Nucleotide Polymorphisms characterisation

Annotation for the 10 AA genome chromosomes was generated using the SGSautoSNP pipeline script, *gene_annotation.py* as described in Chapter 2*.* This generated gene models which are predicted based on those publically available from several organisms. For each of the 10 AA genome chromosomes a SnpEff database, including the reference genome and the genome annotation, was created and used to categorise the effects of SNPs. The output of SnpEff (Cingolani et al., 2012) provided detailed information on the number of changes and the change rate per chromosome based on the annotation. Each SNP effect was classified according to SNPeff into four classes (i) “high effect” for SNPs which modify splice sites, start or stop codons (gain or loss), (ii) “low effect” for SNPs in coding regions which do not lead to an amino acid sequence change, (iii) “moderate effect” for SNPs which led to amino acid sequence change and (iv) “modifier effect” for the SNPs located outside the genes, in introns or in non transcribed regions. Table 3.8 shows the proportion of variants in each class and that the overall impact of all variants is largely modifying (92.3% - 93.3%), followed by moderate (4.1% - 4.6%), low (2.1% - 2.5%), and high impact (0.5%). Most changes (28.3 - 30.5%) were downstream (5 kb downstream of the most distal polyA addition site) and upstream (26.3 - 29.1%) (5 kb upstream of the most distal transcription start site). The lowest frequency change was the Non Synonymous Start type and Start Lost, with 0 to 3 and 0 to 5 events between all chromosomes. The changes in the intergenic regions of the chromosomes range between 21.1 - 27.8% of the total, while the changes in introns represented between 3.7 - 4.5% of the changes. The portion of changes within the exon regions ranged between 6.6 - 7.7%. Table 3.9 shows that the non synonymous/synonymous ratio ranges from 1.8 to 2.0 across all 10 Brassica AA chromosomes. Table 3.10 shows for all 10 AA genome chromosomes the three effects per functional class, between 63.47 and 65.07% missense changes, 31.61 and 33.27% silent changes, and a small fraction of nonsense changes (3.13 to 3.59%). The Missense/Silent ratio for all chromosomes ranges between 1.91 and 2.06. These values are comparable to results observed in peach cultivars where the missense/silent ratio ranged from 1.43 - 1.53 (Fresnedo-Ramirez et al., 2013).

There were 25,323 - 45,066 transitions (Ts) and 20,300 - 37,420 transversions (Tv) identified, giving and a Ts/Tv ratio from 1.20 to 1.26 across all 10 chromosomes (Table 3.11). Table 3.11 shows that overall more transitions than transversions were predicted.

.

Table 3.8: SNPeff results for all 10 *Brassica* AA chromosomes.

	chr01		chr02		chr03		chr04		chr05		chr06		chr07		chr08		chr09		chr10	
High effect (Total)	**700**	**0.5%**	**841**	**0.5%**	**1287**	**0.5%**	**770**	**0.5%**	**741**	**0.5%**	**1036**	**0.5%**	**861**	**0.5%**	**646**	**0.5%**	**933**	**0.5%**	**866**	**0.5%**
SPLICE_SITE_ACCEPTOR	18	0.0%	32	0.0%	46	0.0%	33	0.0%	23	0.0%	46	0.0%	31	0.0%	26	0.0%	35	0.0%	32	0.0%
SPLICE_SITE_DONOR	31	0.0%	29	0.0%	49	0.0%	33	0.0%	27	0.0%	47	0.0%	30	0.0%	34	0.0%	40	0.0%	19	0.0%
START_LOST	1	0.0%	4	0.0%	3	0.0%	0	0.0%	0	0.0%	5	0.0%	4	0.0%	1	0.0%	1	0.0%	2	0.0%
STOP_GAINED	355	0.2%	406	0.2%	613	0.2%	372	0.2%	375	0.2%	476	0.2%	399	0.2%	301	0.2%	454	0.2%	428	0.3%
STOP_LOST	295	0.2%	370	0.2%	576	0.2%	332	0.2%	316	0.2%	462	0.2%	397	0.2%	284	0.2%	403	0.2%	385	0.2%
Low effect (Total)	**3,683**	**2.4%**	**3,963**	**2.3%**	**5,977**	**2.4%**	**3,540**	**2.1%**	**3,367**	**2.2%**	**5,066**	**2.4%**	**3,928**	**2.4%**	**3,091**	**2.4%**	**4,401**	**2.3%**	**4,122**	**2.5%**
NON_SYNONYMOUS_START	1	0.0%	2	0.0%	0	0.0%	1	0.0%	0	0.0%	0	0.0%	3	0.0%	0	0.0%	2	0.0%	0	0.0%
SYNONYMOUS_CODING	3,619	2.3%	3,884	2.2%	5,851	2.3%	3,464	2.1%	3,298	2.1%	4,965	2.3%	3,830	2.3%	3,023	2.3%	4,297	2.2%	4,039	2.5%
SYNONYMOUS_STOP	63	0.0%	77	0.0%	126	0.1%	75	0.0%	69	0.0%	101	0.0%	95	0.1%	68	0.1%	102	0.1%	83	0.1%
Moderate effect (Total)	**6,800**	**4.4%**	**7,560**	**4.3%**	**11,343**	**4.5%**	**6,951**	**4.1%**	**6,383**	**4.1%**	**9,217**	**4.4%**	**7,157**	**4.4%**	**5,841**	**4.5%**	**8,183**	**4.2%**	**7,520**	**4.6%**
NON_SYNONYMOUS_CODING	6,800	4.4%	7,560	4.3%	11,343	4.5%	6,951	4.1%	6,383	4.1%	9,217	4.4%	7,157	4.4%	5,841	4.5%	8,183	4.2%	7,520	4.6%
Modifier effect (Total)	**143,645**	**92.8%**	**162,138**	**92.9%**	**231,512**	**92.6%**	**157,202**	**93.3%**	**144,188**	**93.2%**	**195,971**	**92.7%**	**151,668**	**92.7%**	**119,624**	**92.6%**	**179,595**	**93.0%**	**149,604**	**92.3%**
DOWNSTREAM	45,227	29.2%	49,810	28.5%	76,277	30.5%	48,078	28.5%	43,825	28.3%	60,249	28.5%	46,692	28.5%	38,040	29.4%	55,495	28.7%	48,017	29.6%
INTERGENIC	37,384	24.1%	45,082	25.8%	52,769	21.1%	46,862	27.8%	42,998	27.8%	53,899	25.5%	41,543	25.4%	30,281	23.4%	50,297	26.0%	36,715	22.6%
INTRON	6,307	4.1%	6,895	4.0%	11,205	4.5%	6,310	3.7%	6,211	4.0%	9,567	4.5%	7,168	4.4%	5,824	4.5%	8,096	4.2%	7,022	4.3%
UPSTREAM	43,593	28.2%	48,048	27.5%	72,749	29.1%	44,757	26.6%	40,713	26.3%	57,030	27.0%	44,380	27.1%	35,961	27.8%	52,265	27.1%	45,393	28.0%
EXON	11,134	7.2%	12,303	7.1%	18,512	7.4%	11,195	6.6%	10,441	6.8%	15,226	7.2%	11,885	7.3%	9,518	7.4%	13,442	7.0%	12,457	7.7%
Total number of effects	**154,828**		**174,502**		**250,119**		**168,463**		**154,679**		**211,290**		**163,614**		**129,202**		**193,112**		**162,112**	

Table 3.9: Non synonymous, synonymous and Non synonymous/synonymous ratio for all 10 *Brassica* AA chromosomes.

	chr01	**chr02**	**chr03**	**chr04**	**chr05**	**chr06**	**chr07**	**chr08**	**chr09**	**chr10**
Synonymous	3,682	3,961	5,977	3,539	3,367	5,066	3,925	3,091	4,399	4,122
Non synonymous	6,801	7,562	11,343	6,952	6,383	9,217	7,160	5,841	8,185	7,520
Non synonymous/synonymous ratio	1.8	1.9	1.9	2.0	1.9	1.8	1.8	1.9	1.9	1.8

Table 3.10: The missense, nonsense, silent and missense/silent ratio for all 10 *Brassica* AA chromosomes.

	chr01		chr02		chr03		chr04		chr05		chr06		chr07		chr08		chr09		chr10	
MISSENSE	7,097	63.74%	7,936	64.51%	11,922	64.40%	7,284	65.07%	6,699	64.16%	9,684	63.60%	7,561	63.62%	6,126	64.36%	8,589	63.90%	7,907	63.47%
NONSENSE	355	3.19%	406	3.30%	613	3.31%	372	3.32%	375	3.59%	476	3.13%	399	3.36%	301	3.16%	454	3.38%	428	3.44%
SILENT	3,682	33.07%	3,961	32.20%	5,977	32.29%	3,539	31.61%	3,367	32.25%	5,066	33.27%	3,925	33.03%	3,091	32.48%	4,399	32.73%	4,122	33.09%
MISSENSE/ SILENT ratio	1.93		2.00		1.99		2.06		1.99		1.91		1.93		1.98		1.95		1.92	

Table 3.11: SNP information and chromosome length for the 10 *Brassica* AA genome chromosomes.

	chr01	chr02	chr03	chr04	chr05	chr06	chr07	chr08	chr09	chr10
Density (SNP/Mbp)	2,050	2,309	2,559	3,182	2,492	2,995	2,717	2,149	1,931	3,189
Chromosome length	26,743,657	27,848,129	32,230,099	20,227,373	23,941,934	26,273,242	22,305,823	21,233,027	37,197,612	17,624,701
SNPs no.	54,825	64,291	82,486	64,373	59,659	78,694	60,599	45,623	71,843	56,200
Transitions	**30,314**	**35,368**	**45,066**	**35,944**	**33,317**	**43,741**	**33,429**	**25,323**	**39,785**	**31,247**
A > G	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
C > T	50%	50%	50%	50%	50%	50%	50%	50%	50%	50%
Transversions	**24,511**	**28,923**	**37,420**	**28,429**	**26,342**	**34,953**	**27,170**	**20,300**	**32,058**	**24,953**
A > T	33%	33%	33%	34%	32%	32%	33%	32%	32%	32%
A > C	25%	25%	25%	24%	25%	25%	24%	25%	25%	25%
G > T	25%	25%	25%	25%	25%	25%	25%	25%	25%	25%
C > G	17%	17%	17%	17%	17%	17%	17%	18%	17%	17%
Ts/Tv ratio	**1.24**	**1.22**	**1.20**	**1.26**	**1.26**	**1.25**	**1.23**	**1.25**	**1.24**	**1.25**

3.3.5. Single Nucleotide Polymorphisms density

Using the SGSautoSNP pipeline a total of 638,593 SNPs were identified in the AA genome and the SNP frequency varied from 1,931 SNPs/bp to 3,189 SNPs/Mbp across the 10 chromosomes (Figure 3.3). Table 3.11 shows that the total number of SNPs varied between the different chromosomes and also shows that the highest number of SNPs were identified on chromosome 3 (82,486 SNPs) followed by chromosome 6 (78,694 SNPs) and the least numbers of SNPs were discovered on chromosome 8 (45,623 SNPs) followed by chromosome 1 (54,825 SNPs). Next the SNPs were normalised by dividing SNPs by chromosome length in Mbp. Figure 3.3 shows the normalised SNPs were the highest number of SNPs was identified on chromosome 10 followed by chromosome 4 and the least numbers of SNPs were discovered on chromosome 9 followed by chromosome 1.

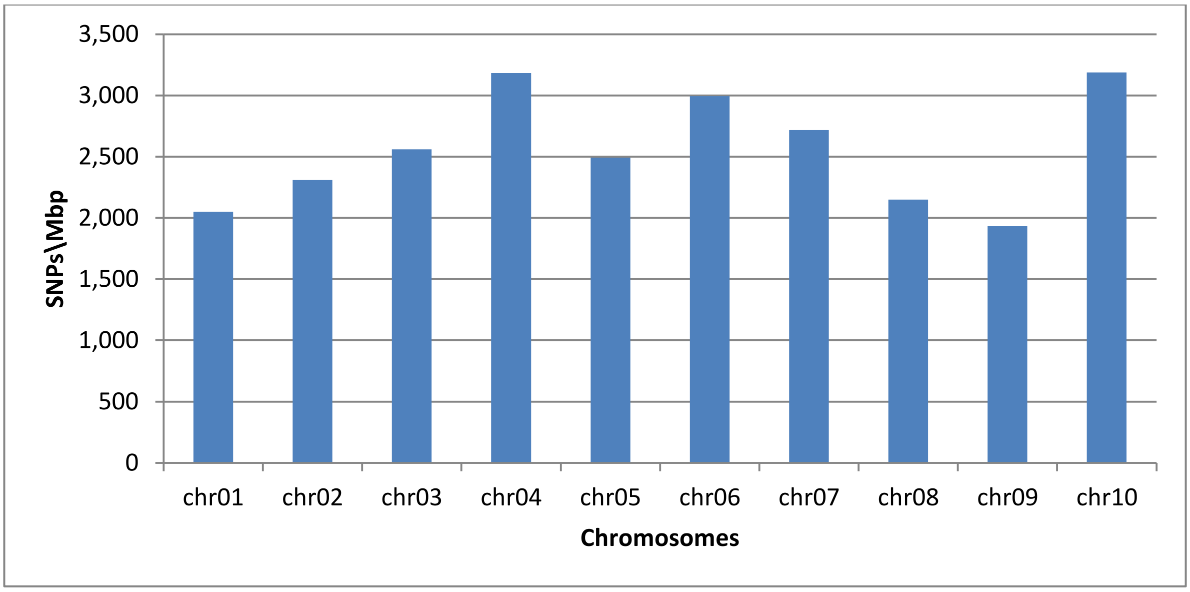


Figure 3.3: Distribution of the numbers of normalized SNPs in the *B. napus* AA genome between six *B. napus* cultivars.

The assembled sequence length of all 10 chromosomes varies from 17,624,701 bp (chromosome 9) to 37,197,612 bp (chromosome 10) (Table 3.11). In most cases the SNP numbers appear to be related to the physical size of the chromosomes. On the longest chromosome (chromosome 9, 37,197,612 bp long) 71,843 SNPs were discovered giving a density of 1,931 SNPs/Mbp, while in the shortest chromosome (chromosome 10, 17,624,701 bp long) 56,200 SNPs were discovered giving a density of 3,189 SNPs/Mbp. However, there are exceptions, for example chromosome 8 is 21,233,027 bp long and has

45,623 SNPs (2,149 SNPs/Mbp), but chromosome 4 (20,227,373 bp) is shorter and has more SNPs (64,373, 3174 SNPs/Mbp) than chromosome 8. The distribution of the SNPs along each chromosome also showed high variation as illustrated in Figure 3.4 and Figure 3.5.

Genome regions with low SNP densities could be caused by functional conservation of coding regions between otherwise diverse individuals. These regions could correspond to chromosomal regulatory or housekeeping genes blocks regions essential for an organism’s survival and/or reproduction (She et al., 2009). For most of chromosomes a few regions appeared with a low SNP density in the middle of chromosomes, except chromosome 4, 6, 7 and 10 (see Figure 3.4 and Figure 3.5). The regions of low SNP density may relate to centromere positions.

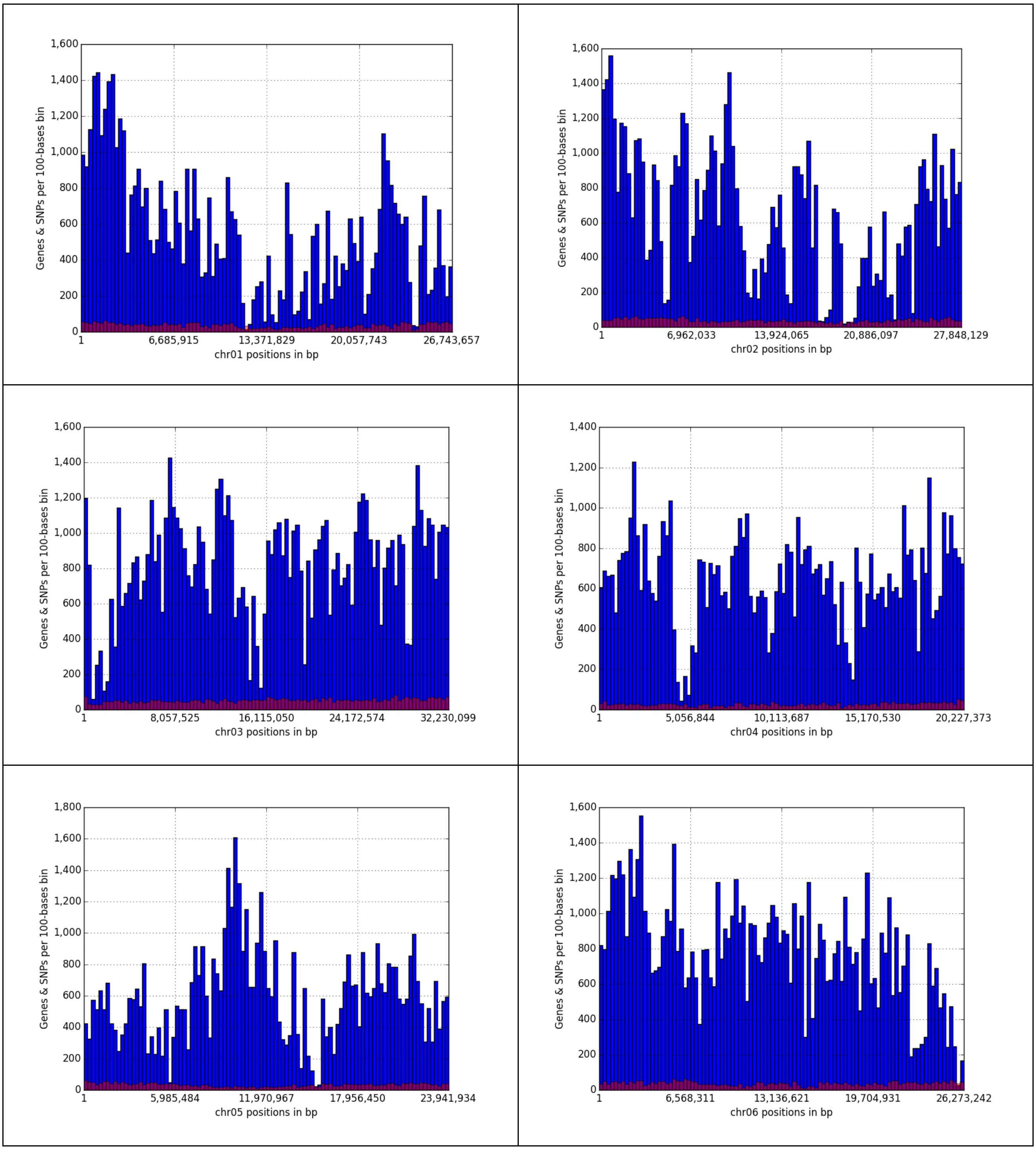


Figure 3.4: Comparison in the density distribution for *Brassica* AA 1 to 6 chromosomes. The density is given in number of genes and SNPs per 100 bases for a particular position in the chromosome. SNPs are blue, genes are red and the overlay between them is purple.

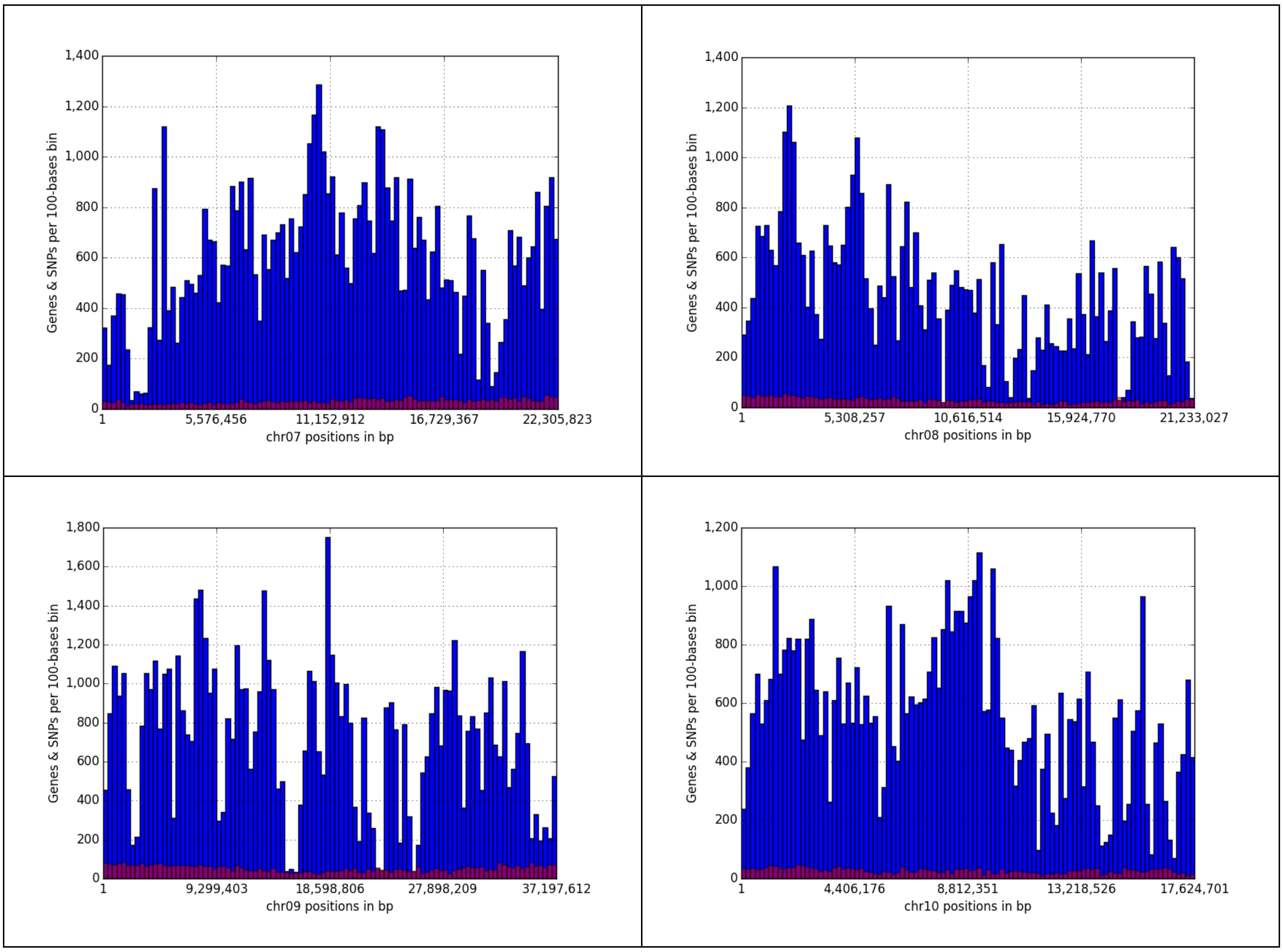


Figure 3.5: Comparison in the density distribution for *Brassica* AA 1 to 6 chromosomes. The density is given in number of genes and SNPs per 100 bases for a particular position in the chromosome. SNPs are blue, genes are red and the overlay between them is purple..

3.4. Conclusion

Second Generation Sequencing has introduced a revolution in plant research and genetics and made sequencing affordable. In this study we show that it was possible to predict thousands of SNPs in the *B. napus* AA genome using the SGSautoSNP pipeline. Paired reads from six *B. napus* cultivars were mapped between 4.76% and 7.72% mapped to the *B. rapa* genome and 5.64% and 9.62% mapped to the *B. oleracea* genome. This is most likely due to many read pairs mapping to multiple locations in this highly repetitive genomes and subsequently being ignored due to the SOAPaligner *-r 0* option, which minimises false SNP calls in repetitive regions and provides confidence in the SNP prediction. Furthermore, it helps to make sure that SNPs are only predicted in single-copy regions of the reference genome. This unmapped rate is much higher than compare for example in a tomato research from (Causse et al., 2013) who cleaned first its reads (average coverage being 11.2x) and 82 to 90% of the reads remained which were than mapped. Only 3 - 5% of the reads did not map using the stringent criteria. Other research groups could not map 15% of reads in rice (Subbaiyan et al., 2012) and 20% in tomato (*S. pimpinellifolium*) (Sato et al., 2012). The reason could be that the other genomes have lower percentage of repeated sequences.

SGSautoSNP does not consider the reference genome for SNP discovery. The calling of SNPs between reads aligned to a reference while ignoring the reference allele allows this pipeline to be applied to accurately call SNPs between individuals using a reference from a divergent species. While this pipeline does not attempt to call all biological SNPs, the very large numbers of SNPs identified are valuable for genetic studies and the association of traits with candidate agronomic genes. SGSautoSNP was successfully applied in calling SNPs in *Brassica napus*, wheat (Lorenc et al., 2012) and *Leptosphaeria maculans* (Zander et al., 2013). These three genomes vary in size and complexity (*B. napus* 1,300 Mb; *Triticum aestivum* (wheat), 16,000 Mb; *L. maculans*, 45.12 Mb), with the *B. napus* genome being allotetraploid, wheat is a hexaploid and *L. maculans* genome is haploid. This demonstrates the flexibility of the pipeline for a broad range of organisms. One limitation of the SGSautoSNP presented in this thesis is that it is designed for homozygous species and does not work efficiently when heterozygocity is present. However it has since been extended by another student for application in heterozygous species.

Firstly, we presented the application of SGSautoSNP pipeline to *Brassica napus* to identify more than 638,593 SNPs with an accuracy of greater than 95% using Sanger sequencing and 94.5 % of the successfully assayed SNPs were validated as polymorphic on the 6 K Infinium™ array. These validations rates exceed those recorded in the below *B. napus* studies and clearly demonstrating the accuracy of the SGSAutoSNP pipeline. Eight *B. napus* lines were used for Restriction Site Associated DNA (RAD) marker sequencing and validated by Sanger Sequencing with an accuracy of around 84% (Bus et al., 2012). Ninety-four genotypes from the Tapidor × Ningyou7 (TNDH) population were genotyped in order to desing a *B. napus* InfiniumTM assay which is composed of SNPs obtained from EST data and reached an accuracy of around 77 % (Delourme et al., 2013).

In this study 6 cultivars were used and 638,593 SNPs were discovered, this is more than

50% more SNPs than idenfiried with four cultivars reflecting the additional data and diversity (Dalton-Morgan et al., 2014).

In this study the Ts/Tv ratios range from 1.2 to 1.26 across all 10 chromosomes (Table 3.11) which is very similar to other *B. napus* studies such 1.29 (Dalton-Morgan et al., 2014) and 1.39 (Bus et al., 2012). These values are lower than Ts/Tv ratios observed in other plants such as 1.6 in eggplant (Barchi et al., 2011), 3.9 in maize, 1.9 in alfalfa, 1.6 in eikorn wheat (*Triticum monococcum* L.), 2.5 in barley and Lotus (Vitte and Bennetzen, 2006) and 3.29 - 3.63 peach (Fresnedo-Ramirez et al., 2013). Between 23,763 to 60,299 transitions (Ts) and 34,020 to 132,024 transversions (Tv) and a Ts/Tv ratio from 1.77 to 1.28 across all wheat group 7 chromosome arms are presented in Chapter 4. The differences in Ts/Tv ratio between different species could be caused by differential abundance of methylated cytosine in CpG dinucleotides, because methylated cytosine can be deaminated and coverted to thymine resulting in this Ts/Tv bias.

By combining the SGSautoSNP pipeline together with SnpEff it was possible to determine whole genome SNP trends, like transition to transversion ratios across chromosomes. Very low numbers of “high effect” SNPs (splice site acceptors, splice site donors, start lost codons, stop gained codons, and stop lost codons) were predicted. These SNPs could impact the structure of the proteins by changing the length of the open reading frame (ORF) or the amino acid transcripts.

A study of eight tomato cultivars identified more than 4 million SNPs (Causse et al., 2013). These SNPs caused more than 98% modifier effects which was around 5% higher than in this study. The moderate SNPs were 0.93 to 1.5%, but in this study they ranged from 4.1 – 4.6%. Low effects were 0.80 to 1.3%, but here they ranged from 2.1 – 2.5%. Finally, the high effect variants represented 0.05 to 0.1%, but in the present study they were 0.5%. Overall this study found a similar trend in the distribution of SNP effects in these 4 classes. The tomato study found 57% of SNPs were located in intergenic regions, more than double the amount found in this study. This may reflect the poor mapping of reads in the intergenic region of this polyploid species. They found 34% of SNPs in downstream or upstream regions of genes which is around 4% for downstream and 5% for upstream higher than our study.

Our non synonymous/synonymous ratio are slightly bigger (1.8 to 2.0) than those detected in cherry tomato cultivars (1.34) and cultivated tomato cultivars (1.48) (Causse et al., 2013), in wild soybean (1.36) and cultivated soybean (1.38) (Lam et al., 2010) and rice (1.2) (Subbaiyan et al., 2012). A synonymous SNP is when a DNA sequence changes, but the translated amino acid stays the same, non-synonymous SNPs mean that the translated amino acid will change. Phenotypic change can be caused by non-synonymous SNPs within transcribed genes; because an organism's interaction with the environment can be affected by alteration of the protein function or structure. Non-synonymous SNPs which can be linked to phenotypic change are the best markers (Edwards et al., 2007a). Non-synonymous SNPs are more readily tolerated in a polyploid

Chapter 4: Application of SGSautoSNP in wheat

4.1. Introduction

Bread wheat (*Triticum aestivum*) has an allohexaploid genome which is very large, complex and contains many repetitive elements. These factors make it difficult to sequence and to discover SNPs in wheat. A common way to address the above issues is to select only a portion of the genome to simplify the sequencing. One example used in wheat is to isolate and sequence individual chromosome arms, which eliminates homoeology resulting from multiple genomes and decreases the genome complexity and size (Vrana et al., 2000). SNP discovery from Second-Generation Sequencing technologies is challenging due to short reads and high error rates. It is difficult to distinguish between real SNPs and sequence or read mapping errors (Duran et al., 2009c, Imelfort et al., 2009). However, we were able to reduce this issue with help of the SGSautoSNP algorithm and other software included in this pipeline which is described in detail in Chapter 2.

In an early study there were 903 SNPs discovered, with a frequency of 1 SNP per 540 bp, among EST sequences in a collection of 12 wheat genotype from Brazil, Canada, China and Mexico (Somers et al., 2003). In another study, twenty six hexaploid wheat genotypes from diverse origins and growth habits were analysed. BAC (bacterial artificial chromosome) sequences from *T. aestivum* were used to design PCR primers and a total of 64 SNPs were discovered between the 26 genotypes (Ravel et al., 2006). An additional twenty-one SNPs were detected with a frequency of one in 76.1 bases from 56 sequences from three species of einkorn wheat (T. monococcum ssp. aegilopoides, T. monococcum ssp. monococcum and T. urartu accessions) (Chen et al., 2009a). A total of 2,659 SNPs were identified in tetraploid durum wheat (Triticum durum Desf.) between 12 cultivars. In this study, two reduced representation libraries (RRLs) were sequenced from the inbred line crosses Colosseo × Lloyd and Meridiano × Claudio using the Roche 454 GS FLX sequencer. For the SNP validation, 768 SNPs were chosen and assayed using the Illumina BeadExpress genotyping system. Only 275 (35.8%) of SNPs could be validated (Trebbi et al., 2011). Allen et al. (2011) identified 14,078 putative SNPs in 6,255 distinct reference sequences with Illumina GAIIx ESTs data from the wheat lines Avalon, Cadenza, Rialto, Savannah and Recital (Allen et al., 2011). The validation rate from a subset of

1,659 was 67%, using the KASPar genotyping platform. In a separate project, (Lai et al., 2012b) identified a total of 38,928 candidate SNPs from bread wheat Roche 454 transcriptome data, with an accuracy of 78%. These SNPs are presented in an online database (http://autosnpdb.appliedbioinformatics.com.au/). You et al. (2011) identified SNPs between two accessions of one of the diploid progenitors of bread wheat, *Aegilops tauschii* (You et al., 2011). Roche 454 sequencing of *Ae. tauschii* accession AL8/78 was combined with Applied Biosystems SOLiD sequencing of genomic DNA and cDNA from *Ae. tauschii* accession AS75 to identify a total of 497,118 candidate *Ae. tauschii* SNPs. In another project, Roche 454 sequence reads from nine wheat accessions originating from Australia, China, Mexico and USA were assembled into reference transcripts (RTs). SNP discovery was performed by mapping transcriptomes of 26 hexaploid wheat accessions, sequenced using Roche 454 and Illumina (GAIIx and HiSeq2000). A total of 25,454 SNPs were indentified with a validation rate of 85 - 90% on a 9K iSelect Beadchip Assay (Cavanagh et al., 2013).

4.1.1. Project aims

In this chapter we present the results of the application of SGSautoSNP (Second-Generation Sequencing AutoSNP), a SNP discovery pipeline described in chapter 2 to hexaploid bread wheat. Validation suggests greater than 93% of SNPs represent polymorphisms between wheat cultivars and hence are valuable for diversity analysis, marker assisted selection and genotyping by sequencing. The work in this chapter demonstrates that the SGSautoSNP pipeline is suitable for high resolution SNP discovery in very large and complex genomes (Lorenc et al., 2012).

4.2. Material and Methods

We demonstrate the potential of the SGSautoSNP pipeline by identifying SNPs between four Australian wheat cultivars; Drysdale, Gladius, Excalibur and RAC875. We used Illumina whole-genome paired read sequence data which had coverage between 8.8x and 10.8x. These sequence data were downloaded from the Bioplatforms web site (http://www.bioplatforms.com.au/datasets/wheat, 17 August 2012) (Edwards et al., 2012). The wheat group 7 (7A, 7B and 7D) chromosomes arms had been sorted by the flow cytometry method (Vrana et al., 2000). The DNA of these chromosome arms were isolated, sequenced using Illumina GAIIx and HiSeq 2000 platforms, and assembled using Velvet (Zerbino and Birney, 2008) and affterwards syntenic build were created by (Berkman et al., 2013, Berkman et al., 2012b, Berkman et al., 2011). Syntenic builds are contigs which have been ordered based on similarity to related cereal genomes. Table 4.1 shows the group 7 chromosome arm, syntenic build sizes and sequence coverage. The data for the four cultivars were mapped to the reference bread wheat chromosome arm shotgun assemblies representing homoelogous chromosomes 7A, 7B and 7D (Berkman et al., 2013), as well as 4AL (Hernandez et al., 2012). In the absence of one of the homoeologues, cultivar specific reads from the missing homoeologue would likely map to one of the other homoeologous genomes, confounding SNP discovery. An assembly from chromosome arm 4AL was included as this arm contains a reciprocal translocation with 7BS (Berkman et al., 2011). Assemblies for each of the wheat 7A, B and D chromosomes, including the syntenic builds and extra contigs were as described by (Berkman et al., 2011) and are accessible at the wheatgenome.info web site (Lai et al., 2012a) (http://www.wheatgenome.info, 17 August 2012).

The SGSautoSNP pipeline was used with default settings (see Chapter 2), except for *SOAPaligner.py* which requires insert-size for each read. These insert-sizes can be found in Table 4.3. The pipeline runs on a computer cluster (Barrine) at the University of Queensland (see Appendix).

Table 4.1: Summary of wheat group 7 chromosome data including remaining assembled contigs (extra contigs) (Berkman et al., 2013).

Chromosome arm	Chromosome arm size in Mbp	Syntenic builds length in Mbp	Extra contigs length in Mbp
7AS	407	6.85	203.95
7AL	407	7.75	248.89
7BS	360	6.62	214.51
7BL	540	5.97	248.14
7DS	381	7.47	203.38
7DL	346	13.48	224.77

A total of 40 SNPs were randomly selected from the three group 7 reference genomes for validation. The validation work described below was performed by Jacqueline Batley and Satomi Hayashi. The SNPs represented 18, 9 and 13 SNPs from the A, B and D genomes respectively and had a range of redundancy scores. Genomic DNA was isolated from the four wheat cultivars Drysdale, Gladius, Excalibur and RAC875, according to a protocol adapted from (Fulton et al., 1995). PCR amplification of the 40 loci was performed using primers designed to conserved sequence surrounding the SNPs (Table 4.5 and Table 4.6) in a 20 µL reaction volume containing 1 × iTaq PCR buffer (containing 100 mM Tris-HCl and 500 mM KCl, pH 8.3) (Bio-Rad), 200 µM each dNTP (Bio-Rad), 0.5 µM each primer, 1.5 U iTaq DNA polymerase (Bio-Rad), RNase and DNase free water (Gibco) and 60 ng DNA. Thermocycling conditions for the reaction were 94 °C for 2 min, followed by 35 cycles of: 94 °C for 30 s, annealing for 1 min at 60 °C and extension for 1 min at 72 °C. Final extension was performed at 72 °C for 10 min. Gel electrophoresis on a 1% (w/v) agarose gel in 1 × TAE buffer (Sambrook and Russell, 2001) containing ethidium bromide resolved products, which were excised and purified using a silica method based on (Boyle and Lew, 1995). The purified PCR products were Sanger sequenced using BigDye 3.1, using forward PCR primers, and analysed using an ABI3730xl. The sequences for each locus and cultivar were aligned and compared using Geneious Pro v5.4.6 (Kearse et al., 2012) with a cost matrix of 65%, a gap open penalty of 6, and a gap extension penalty of 3, and each of the SNPs assessed.

4.3. Results and Discussion

4.3.1. Reads mapping

The paired reads for each of the varieties (see Table 4.2) were processed by the SGSautoSNP pipeline and mapped onto reference genome using the *SOAPaligner.py* script, which is a wrapper around SOAPaligner (Li et al., 2009b). The reference consists of the group 7 chromosomes (7A, 7B and 7D) combined with 4AL. The *−r 0* option of SOAPaligner was applied which removes reads where they match multiple positions equally well. This option aims to increase SNP calling accuracy by ignoring read pairs that cannot be accurately positioned on the reference. Similarly, only reads that mapped as a pair were used for SNP discovery. Due to the short length of the reads, one read could match at many positions, but two reads separated by a gap of defined insert size provides a greater confidence of specific and accurate read mapping. Table 4.3 shows the minimum and maximum insert sizes used for SOAPalinger. Of the reads used for mapping, between 3.10% and 5.14% mapped to the group 7/4AL reference as read pairs (see Table 4.2). As the group 7 reference is estimated to cover approximately 14% of the complete genome, the number of mapped reads is less than predicted. This is due to of the large number of repeats in the wheat genome (Brenchley et al., 2012, Flavell et al., 1977) which prevents the reads from mapping to a unique specific location and the fact that the genome assemblies do not represent the complete chromosome arms.

Table 4.2: Summary of wheat cultivar data and mapping reference genome which is made out of the group 7 chromosomes (7A, 7B and 7D) combined with 4AL. The wheat variety column contains in brackets the cultivar name abbreviation.

Wheat variety	Data generated (Gbp)	Data mapped to reference (Gbp)	% read pairs mapped
Drysdale (D)	168	8.65	5.14
Excalibur (E)	146	5.36	3.66
Gladius (G)	180	8.47	4.70
RAC875 (R)	132	4.1	3.10

Table 4.3: The minimum and maximum insert sizes used during the alignment with SOAPalinger for four wheat cultivars paired-reads.

Wheat variety	Minimum Insert sizes	Maximum insert sizes
Drysdale (D)	80	480
Excalibur (E)	60	520
Gladius (G)	60	600
RAC875 (R)	60	370

4.3.2. Single Nucleotide Polymorphism calling

A total of 881,289 SNPs were identified across the group 7 chromosomes from the four Australian wheat varieties using the SGSautoSNP pipeline (Lorenc et al., 2012). These SNPs consisted of 63 - 70% transitions and 30 - 37% transversions (Table 4.4). The SNP frequency on the syntenic build varies across the three group 7 chromosomes, 963.3 SNPs/Mb (7A), 746.2 SNPs/Mb (7B) and 149.7 SNPs/Mb (7D). SGSautoSNP provides a SNP score which is a measure of confidence in SNP prediction. In this study, the SNP score ranged from 2 to 60, with an average of 4. All predicted SNPs have been included in a public wheat genome GBrowse database hosted at the wheatgenome.info web site (Lai et al., 2012a).

Table 4.4: Information about SNPs in 7A, 7B and 7D chromosome arms.

		Total			**Syntenic build**		
	SNP	**7A**	**7B**	**7D**	**7A**	**7B**	**7D**
Transitions	A/G	150,760	119,165	30,215	5030	3248	1014
	C/T	150,494	118,466	30,084	4724	3198	973
Transversions	A/C	37,919	33,117	9,360	1137	814	325
	A/T	24,838	22,695	8,102	911	642	284
	C/G	31,057	27,182	7,383	1149	775	247
	G/T	38,210	32,737	9,175	1107	713	294
	A/C/G	25	34	11	-	3	-
	A/C/T	41	38	15	-	1	-
	A/G/T	37	47	10	1		-
	C/G/T	29	36	7	-	2	-
Biallelic SNPs no.		433,278	353,362	94,319	14058	9390	3137
Triallelic SNPs no.		433,410	353,517	94,362	14059	9396	-
Transitions (Ts)		301,254	237,631	60,299	9754	6446	1987
Transversions (Tv)		132,024	115,731	34,020	4304	2944	1150
Ts/Tv ratio		2.28	2.05	1.77	2.26	2.19	1.73

There were between 237,631 to 60,299 transitions (Ts) and 34,020 to 132,024 transversions (Tv) and the Ts/Tv ratio ranged from 1.77 to 2.28 across all group 7 chromosome arms (Table 4.4).

4.3.3. Single Nucleotide Polymorphisms validation

Validating individual SNPs in a hexaploid species is a challenge as the amplification of loci requires the design of homoeologue specific PCR primers. Of 40 SNPs selected for validation, 12 did not produce clean PCR amplification products or Sanger sequence. This reflects inefficiency in validation rather than SNP calling errors and so these SNPs were ignored. Of the 28 SNPs that did produce clean Sanger sequence data, 26 (93%) produced the expected genotype. One SNP was homozygous across cultivars and not a true SNP, while one appeared to be heterozygous, suggesting a SNP between the homoeologous genomes rather than between cultivars. SGSautoSNP predicted correct SNPs even for the minimum SNP score of 2, although the monomorphic SNP has a score of 2 while the SNP between homeologues had a SNP score of 6 (Lorenc et al., 2012, Zander et al., 2013, Trebbi et al., 2011, Allen et al., 2011, Lai et al., 2012b, Cavanagh et al., 2013).

Table 4.5: Summary of single nucleotide polymorphism (SNP) validation in wheat chromosome 7A.

SNP primer name	Forward primer	Reverse primer	SNP score	Validation
UQ7A27	TAACATAAGCAAAGTTCTATTA	TTTGGAACACAATCGGAACTT	6	Failed
UQ7A1397	TCTATTGGATTCTTTCCGAT	TCACCCTGTGGAATGAAAGA	5	Failed
UQ7A5622	TTAGCCAAAATGGACCCAAA	CCTCTTTATTCAATCTGGAAACG	2	True SNP
UQ7A129835	TTCTTACTGTGGCTGCATCA	GCCATCCTAAACGACCTTCA	5	True SNP
UQ7A9400	GCCCATATGCAGTTCATGGT	AGAGCCAAACCTTCCCTGAT	2	Failed
UQ7A7915	CATGCCAACCCAAGTAGACC	GAAGCGTGAAAATTTCGTGA	6	True SNP
UQ7A6107	TGGTGTTTACGCTGAAGTTACC	CTGGCCTGGGCACTACATA	6	True SNP
UQ7A2603	GTCACCAACCAGCTCGAAAT	TTGTAGCTTTGCCTCTGTGAA	2	Failed
UQ7A3491	AGTCGCCGGCAGTAAAAATA	CCGAAGAAAATGTGGTGGAG	4	True SNP
UQ7A4532	TTTCCTCTAGATCTGTGCAAAATG	CATCCAGGACTGCATAAGCTC	6	True SNP
UQ7A100138	TCCCTGGTCCACGAGTTATT	AAATGGTTTGAGCCTTGTGC	7	Failed
UQ7A136305	CATCATCTTTGAAAAATCCTAGCC	TGTTCTGCAAGCTTCGTCTG	5	True SNP
UQ7A155877	AAGCTGTTGTGCCAGTGTTG	GAGCTAGCGTCGCTGACATA	4	True SNP
UQ7A180868	GACCGTCATCGAATGTAGCA	TCGTCCACCCAGACCTTATC	3	True SNP
UQ7A287189	GGCGATCATCACTTAAGAAACC	CAGTAATGAGGTTTCTGCTTGG	2	Failed
UQ7A322716	TCTGTTCGCAAACCAACG	GTGCGTTATCAGGGGAACAT	11	True SNP
UQ7A57227	ATGGGTGAAGGGAATACAGC	TGCATGCACATACAACCAAA	5	True SNP
UQ7A87191	TCAGTTCGGTAAGGATGAAGA	GAAGCAGTATGCATCTAAACTTTG	6	Heterozygous

Table 4.6: Summary of single nucleotide polymorphism (SNP) validation in wheat chromosomes 7B and 7D.

SNP primer name	Forward Primer	Reverse Primer	SNP score	Validation
UQ7B21	GCAGGGTTAATTTCTAGCAAGC	GCCTTTTATCCAAAGCCATC	8	Failed
UQ7B484	CTCAACCTCCCAAGCATGA	GCTATCCAGCTACCCTGTGC	11	Failed
UQ7B3940	GCCAGAGGCACTAGCATCAC	GGTAATTGTGGAGCAAGCAA	6	True SNP
UQ7B4960	GCATGGCATTTCAAGATCAG	GGAGGAGGACAAAGCCAGAT	5	True SNP
UQ7B5991	CCAAGCCACCACCCTTTAT	TAATCCCCGTCATCTCGAAG	4	True SNP
UQ7B120997	CTCCTCAGATGACCAATTTGC	CACCAAAATATGCTGTACAATTCTATG	7	Failed
UQ7B256895	GCAGCAGAGGTAGGCACTTC	GAAATGCTTCGAGTGTGGTG	11	True SNP
UQ7B64318	GGGTCCAGACTTCCACGTTA	CCCACATTAATTTGTACGACCTC	6	Failed
UQ7B97303	TGATTCGAGCCCATATAGGAA	AGCCATGCGGAAATATTGAG	8	True SNP
UQ7D283	TGAGTAAGACAACAATCAGAGCA	CAATGCGAGCAAAAAGATCA	5	True SNP
UQ7D429	TGTGCTGACGTGGCATCTAT	GCATGTGGAAAACGAGTGTG	3	True SNP
UQ7D689	CATCTGGCCTCAACATCAAA	TGTTGGTAGTGAGGCACTTCTT	9	Failed
UQ7D948	GGCGATACTCGATGAAAGAAA	TTGGAAACTACAATTGCACAAC	9	True SNP
UQ7D1189	GCGTGGAGTAGAGGGACAAG	TCCAAAAAGCAAAACAAATGC	4	True SNP
UQ7D1491	AGCGCAAGGAGGAGGTTAGT	GAGCCAAGTCCTTGTCAATTT	7	True SNP
UQ7D1846	AATGTGTTCCATCCAAGACG	GCCAAGGTCGACATGTGATA	10	True SNP
UQ7D2314	AAACAAGTCTGTGTTGCGTCA	TGCAGATACATGGCTCCAGA	2	Monomorphic
UQ7D20375	CTGCCACCAAACGGATTAAC	AATGCATTGGCAGTCACAAG	6	True SNP
UQ7D27168	TAATGCTATGCCGTGTCAGC	GCCACCTATTATTGAAGGCATC	2	True SNP
UQ7D38754	GAGCGAGCAATGCTAGTGTG	GAACCCATTTGATAACCGTGA	3	Failed
UQ7D59683	CGTCCACATTGTTGCAAATC	TTGACCCTGAAGGAAGGATG	6	True SNP
UQ7D68910	TTGCTTTATGCCACTGGAGA	TAGGCCGTGAAACATCAACA	3	True SNP

4.3.4. Single Nucleotide Polymorphisms characterization

Syntenic build gene annotations for the wheat group 7 (7A, 7B and 7D) chromosomes arms were used from (Berkman et al., 2013). For each of the 3 chromosomes arms a SnpEff (Cingolani et al., 2012) database was created and used to categorise the effects of SNPs. The output of SnpEff provided detailed information on the number of changes and the change rate per chromosome based on the annotation.

Each SNP effect was classified according to SNPeff into four classes (i) “high effect” for SNPs which modify splice sites, start or stop codons (gain or loss), (ii) “low effect” for SNPs in coding regions which do not lead to an amino acid sequence change, (iii) “moderate effect” for SNPs which led to amino acid sequence change and (iv) “modifier effect” for the SNPs located outside the genes, in introns or in non transcribed regions.

Table 4.7: SNPeff results for all three wheat group 7 Syntenic build chromosome arms.

	7A_SynBuild_v2.0		7B_SynBuild_v2.0		7D_SynBuild_v2.0	
High effect (Total)	**186**	**0.1%**	**162**	**0.2%**	**62**	**0.2%**
SPLICE_SITE_ACCEPTOR	21	0.0%	18	0.0%	6	0.0%
SPLICE_SITE_DONOR	6	0.0%	12	0.0%	2	0.0%
START_LOST	0	0.0%	0	0.0%	0	0.0%
STOP_GAINED	96	0.1%	101	0.1%	45	0.1%
STOP_LOST	63	0.1%	31	0.0%	9	0.0%
Low effect (Total)	**6610**	**5.3%**	**1885**	**2.7%**	**1375**	**4.5%**
NON_SYNONYMOUS_START	0	0.0%	0	0.0%	0	0.0%
SYNONYMOUS_CODING	6,610	5.3%	1,875	2.7%	1,375	4.5%
SYNONYMOUS_STOP	0	0.0%	10	0.0%	0	0.0%
Moderate effect (Total)	**4,341**	**3.5%**	**2,530**	**3.6%**	**626**	**2.0%**
NON_SYNONYMOUS_CODING	4,341	3.5%	2,530	3.6%	626	2.0%
Modifier effect (Total)	**113,341**	**91.1%**	**65,603**	**93.5%**	**28,555**	**93.3%**
DOWNSTREAM	41,899	33.7%	25,998	37.0%	10,295	33.6%
INTERGENIC	9,718	7.8%	6,298	9.0%	2,345	7.7%
INTRON	10,126	8.1%	6,708	9.6%	3,578	11.7%
UPSTREAM	40,488	32.5%	22,052	31.4%	10,282	33.6%
EXON	11,110	8.9%	4,547	6.5%	2,055	6.7%
Total number of effects	**124,478**		**70,180**		**30,618**	

Table 4.7 shows the proportion of variants in each class and that the overall impact of all variants is largely modifying (91.1 - 93.5%), followed by low (2.7 - 5.3%), moderate (2% - 3.6%), and high impact (0.1 - 0.2%). Most changes (33.6 - 37%) were downstream (5 kb downstream of the most distal polyA addition site) and upstream (31.4 – 33.6%) (5 kb upstream of the most distal transcription start site). The lowest frequency change was the Splice Site Donor, with 2 to 12 events between all 3 group 7 chromosome arms. The changes in the intergenic regions of the chromosomes range between 7.7 - 9% of the total, while the changes in introns represented between 8.1 – 11.7% of the changes. The portion of changes within the exon regions ranged between 6.5 – 8.9%.

Table 4.8 shows Non synonymous, synonymous and Non synonymous/synonymous ratio for all 3 Syntenic build wheat group 7 chromosome arms.

	7A_SynBuild_v2.0	7B_SynBuild_v2.0	7D_SynBuild_v2.0
Synonymous	6,610	1,885	1,375
Non synonymous	4,341	2,530	626
ratio	0.66	1.34	0.46

Table 4.8 shows that the non synonymous/synonymous ratio ranges from 0.46 to 1.34 across all 3 Syntenic build wheat group 7 chromosome arms.

Table 4.9 shows the missense, nonsense, silent and missense/silent ratio for all 3 Syntenic build wheat group 7 chromosome arms.

	7A_SynBuild_v2.0		7B_SynBuild_v2.0		7D_SynBuild_v2.0	
MISSENSE	4,404	63.74%	2,561	64.51%	635	64.40%
NONSENSE	96	3.19%	101	3.30%	45	3.31%
SILENT	6,610	33.07%	1,885	32.20%	1,375	32.29%
MISSENSE/ SILENT ratio	0.67		1.36		0.46	

Table 4.9 shows all 3 Syntenic build wheat group 7 chromosome arms the three effects per functional class, between 63.74 and 64.51% missense changes, 32.20 and 33.07% silent changes, and a small fraction of nonsense changes (3.19 to 3.31%). The Missense/Silent ratio for all chromosomes ranges between 0.46 and 1.36.

4.4. Conclusion

Second generation sequencing has introduced a revolution in plant research and genetics and made the generation of large quantities of data affordable. In these studies, we demonstrate that it was possible to predict hundreds of thousands of SNPs associated with wheat group 7 chromosome arms using the SGSautoSNP pipeline, offering new tools for researchers and plant breeders.

We show the application of SGSautoSNP pipeline to wheat chromosomes 7A, 7B and 7D to identify 881,289 SNPs with an accuracy of greater than 93%. These polymorphisms are available in a GBrowse genome viewer at the wheatgenome.info web site (http://www.wheatgenome.info). The successful application of the SGSautoSNP pipeline method to hexaploid wheat, diploid *Brassica* AA genome and *Leptosphaeria maculans* (Zander et al., 2013) demonstrates that this approach should work for SNP discovery in other large and complex genomes. Recently, the SGSautoSNP pipeline discovery of 4 millions SNPs across the group 7 chromosomes between 16 Australian bread wheat cultivars (Lai et al., 2015). Using more cultivars results in more SNPs, because many reference genome positions did not have enough coverage. Both studies have in common that less SNPs were found on syntenic build and significantly more SNPs are on chromosomes 7A and 7B, compared to 7D. This is consistent with previous results (Chao et al., 2009) and most likely due to early gene flow between *T. aestivum*, the tetraploid and hexaploid species resulted in greater sequence diversity within the A and B genomes than compare to D genome (Caldwell et al., 2004, Dvorak et al., 2006, Talbert et al., 1998).

Only 3.1 to 5.1% paired reads from 4 wheat cultivars mapped to the group 7 chromosome arm reference (Table 4.2). This is likely due to read pairs mapping to multiple locations in this highly repetitive genome and subsequently being ignored due to the SOAPaligner *–r 0* option. This option reduces false SNP calls through the mismapping of reads and provides confidence in the SNP prediction. Furthermore, it helps to ensure that SNPs are only predicted in low-copy regions of the genome for subsequent accurate genotyping. One caveat of this approach is that we do not find all possible SNPs but only SNPs which can be predicted and genotyped with a high accuracy.

In this chapter presented Ts/Tv ratios (1.73 to 2.28) are comparable with values seen in other plants such as 1.6 in eggplant (Barchi et al., 2011), 1.9 in alfalfa and 1.6 in einkorn wheat (*Triticum monococcum* L.). For all 10 *Brassica napus* AA chromosomes presented in chapter 3 the Ts/Tv ratio ranges from 1.20 to 1.26. This bias in transition/transversion ratio is commonly observed in SNP discovery and reflects the high degree of methlyl C to U mutation in genomes (Coulondre et al., 1978). It may be expected that the bread wheat genome is highly methylated due to the two rounds of polyploidy and high repeat content. The observed transition/transversion bias provides a level of confidence in SNP prediction accuracy since erroneously called SNPs caused by sequence read errors or mismapping would be unlikely to display such a bias.

SGSautoSNP has been applied to call SNPs in canola, wheat (4 culitvars), wheat (6 cultivars) and in the fungal genome of *L. maculans* with a prediction accuracy of 95%, 93%, 95% and 90%, respectively (Dalton-Morgan et al., 2014, Lorenc et al., 2012, Lai et al., 2015, Zander et al., 2013). This compares to an accuracy of 35.8% in tetraploid durum wheat (Triticum durum Desf.) in a study of 12 cultivars (Trebbi et al., 2011), a validation rate of 67% from wheat EST data (Allen et al., 2011) and 78% from bread wheat Roche 454 transcriptome data (Lai et al., 2012b). The previous highest validation rate for wheat SNPs was 85 - 90% in a wheat reference transcripts (RTs) project (Cavanagh et al., 2013). Hence, SGSautoSNP is a highly accurate SNP discovery pipeline and can be used for large, complex genomes.

Table 4.7 shows that the three wheat group 7 chromosomes arms (Syntenic build) missing Start Lost, Non Synonymous Start and Synonymous Stop effects, but the SNPeff results for *Brassica* described in Table 3.8 shows that (i) only chromosomes 4 and 5 missing Start Lost, (ii) chromosomes 3, 5, 6, 8 and 10 missing Non Synonymous Start and (iii) all 10 *Brassica* chromosomes have Synonymous Stop. Table 4.7 shows for Syntenic builds that 7A has the most amount of effects (124,478) follow by 43.3% less for 7B and 75.4% less for 7D. Comparing the results to *Brassica* (Table 3.8) this species has more equal number of effects distribution (129,202 - 250,119). On the other hand, the Syntenic builds' Non synonymous/Synonymous ratio for 7B is the highest (1.34), but the 7D has only 0.46 (Table 4.8). The Non synonymous/Synonymous ratio for all 10 *Brassica* chromosomes in Table 3.9 are equal distributed and range between 1.8 and 2. Table 4.9 shows for Syntenic builds that 7B has the highest Missense/Silent ratio of 1.36, but 7D has only 0.46. The

Missense/Silent ratio for *Brassica* of 1.91 – 2.06 across the 10 chromosomes is slightly higher and almost constant (Table 3.10)

Genome wide identification of hexaploid bread wheat SNPs using our pipeline is limited by the lack of publically available chromosome sequences. It is expected that draft assemblies of the remaining chromosome shotgun arms will be available to the public soon and this will enable extension of this method to whole genome SNP discovery in this species and the identification of 881,289 SNPs across the group 7 chromosomes suggests that genome wide discovery would identify a total of more than 6 million SNPs across the genome. For this project four wheat cultivars were used, however Bioplatforms Australia have now sequenced a total of 16 cultivars (Edwards et al., 2012). The detailed analysis of genetic variation in these additional cultivars is being undertaken by Kaitao Lai as part of his PhD project.

Chapter 5: SGSautoSNPdb: a database which stores all SGSautoSNP results

5.1. Introduction

It is important to develop interactive web based applications which store molecular markers, genes, genetic and marker annotations and gene ontology. Furthermore, the information from one web service should be connected to others in order to save time in finding additional information. The advantage would be that search results could be visualised in a way that a researcher can mouse over it and find more information e.g. in a chart. This would allow researchers to access information in a biologist friendly manner.

SGSautoSNP is able to discover millions of SNPs. These large numbers of SNPs require new and innovative approaches to help turn massive amounts of data into usable information. Therefore SGSautoSNPdb has been developed as a web application with a database which will aim to fulfil the above goals in order to provide plant breeders with more information about SNP markers; such as SNP annotation, primers for validation, whether the SNP is in a low SNP density region, whether the SNP has been validated, in which genes a SNP is located and GO terms. Furthermore, SGSautoSNPdb could link gene ids and GO terms to Uniprot (Apweiler et al., 2013) and QuickGO (Binns et al., 2009) respectively. This information will help plant breeders to identify SNPs and genes for important agronomic traits like drought and disease resistance (Gupta et al., 2013) which could be used for breeding new varieties in order to increase the world crops production and keep up with the growing population around the world.

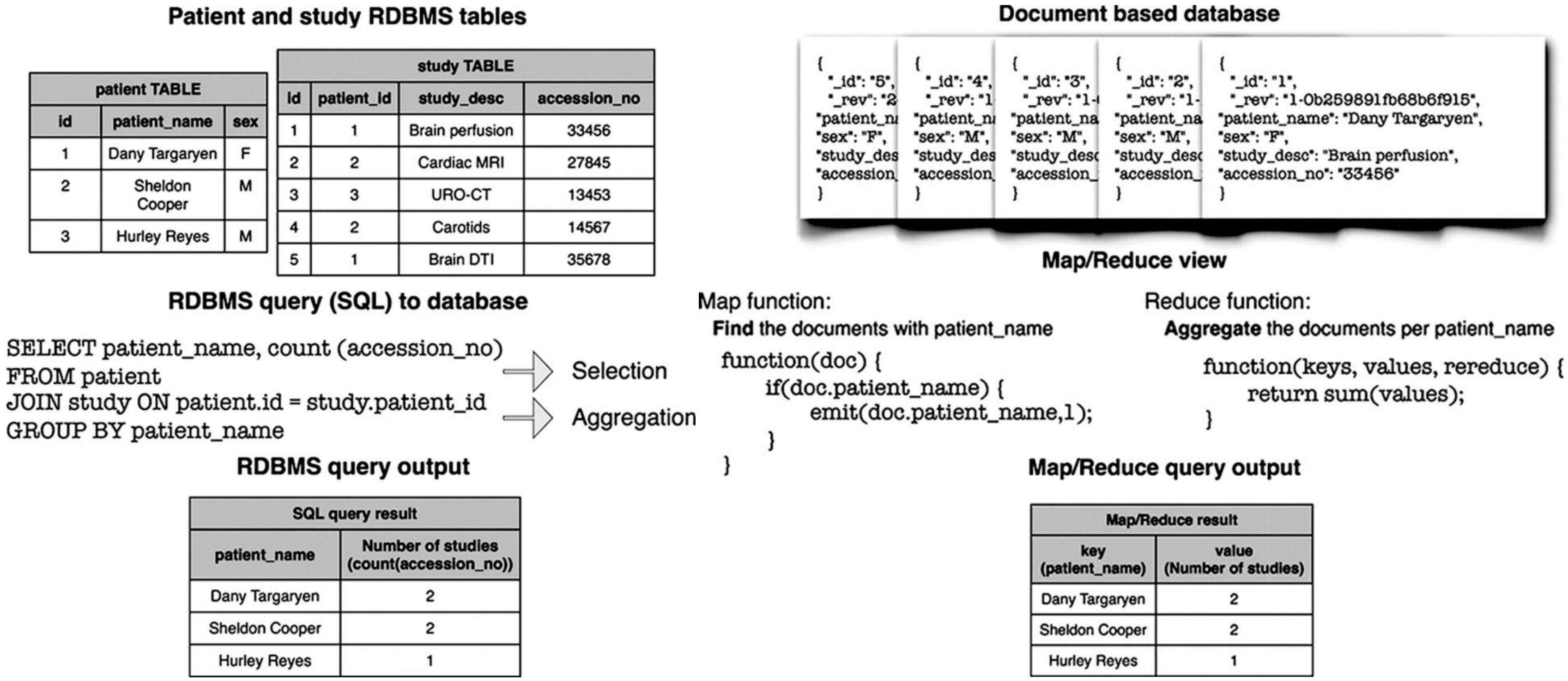
5.1.1. Choosing flexible cache and database for SGSautoSNPdb

5.1.1.1. A flexible and scalable database

The most widespread database for bioinformatics services are based on a relational model. The first step in using relational database management systems (RDBMSs) is to design a schema of tables which defines the relationship between those tables. In the next step the data has to be split into multiple tables which have to satisfy the predefined schema of the RDBMS (Rascovsky et al., 2012). Since the data is split in different tables, an issue arise for biologists how to make sense off all different tables. The solution is to use a Structured Query Language (SQL) to join the different tables together. Joins are very slow in RDSMS. On the other hand, many new databases are based on a non-relational or NoSQL model (http://nosql-database.org). The advantages of these new databases are scalability and flexibility. Avoiding designing a database schema allows making changes to the database when the requirements change while continuing access to the existing data**.** NoSQL is increasingly being used for cloud computing services like Google and Amazon (Manyam et al., 2012)**.** SGSautoSNPdb uses Apache CouchDB database (http://couchdb.apache.org/), which belongs to the group of NoSQL databases.

CouchDB is an open source NoSQL, schema-free, document-oriented database which stores data in the JavaScript Object Notation (JSON) format unlike the different tables used by relational databases. Biologists will understand all relationships in one document rather than split in different tables like in RDBMS. Each document gets a version number and, if not specified, a unique id. Furthermore, CouchDB allows storing any types of files as “attachments”. This database is written in the Erlang and JavaScript programming language. Simultaneous access by users of SGSautoSNPdb would not lead to blocking database access, because of CouchDB's non blocking concurrency implementation (Silbermann et al., 2013).

CouchDB does not support SQL queries like RDBMS, but it uses the MapReduce method introduced by Google for databases (Pike et al., 2005). This is a new method of querying large databases fast and is completely different to SQL (see Figure 5.1). MapReduce is based on two functions, Map and Reduce. On each document the Map function is executed to compute a list of key-value pairs based on the search filter criteria. The optional Reduce function must merge the list of key-value pairs from the Map function to a single value. The MapReduce method in CouchDB is written in JavaScript and stored in a “view” file together with the database. CouchDB keeps the results from the initial run of the “view” file until new documents are added or modified and only then it applies the MapReduce method on the new or updated documents. Therefore, CouchDB is able to provide fast responses when performing queries on views (Redmond and Wilson, 2012). In contrast, the SQL query must always recalculate all current data stored in the RDBMS (Rascovsky et al., 2012). In order to limit the response from querying a “view”, CouchDB API offers parameters key, startkey, endkey and limit (Manyam et al., 2012).



Patient and study RDBMS tables

patient TABLE		
id	**patient_name**	**sex**
1	Dany Targaryen	F
2	Sheldon Cooper	M
3	Hurley Reyes	M

study TABLE			
id	**patient_id**	**study_desc**	**accession_no**
1	1	Brain perfusion	33456
2	2	Cardiac MRI	27845
3	3	URO-CT	13453
4	2	Carotids	14567
5	1	Brain DTI	35678

RDBMS query (SQL) to database

```
SELECT patient_name, count (accession_no)
FROM patient
JOIN study ON patient.id = study.patient_id
GROUP BY patient_name
```

RDBMS query output

SQL query result	
patient_name	**Number of studies (count(accession_no))**
Dany Targaryen	2
Sheldon Cooper	2
Hurley Reyes	1

Document based database

```
{
  "_id": "5",
  "_rev": "2
"patient_n
"sex": "F",
"study_des
"accession_
}
```

```
{
  "_id": "4",
  "_rev": "1
"patient_n
"sex": "M",
"study_des
"accession_
}
```

```
{
  "_id": "3",
  "_rev": "1-
"patient_na
"sex": "M",
"study_desc
"accession_
}
```

```
{
  "_id": "2",
  "_rev": "1-
"patient_na
"sex": "M",
"study_desc
"accession_
}
```

```
{
  "_id": "1",
  "_rev": "1-0b259891fb68b6f915",
"patient_name": "Dany Targaryen",
"sex": "F",
"study_desc": "Brain perfusion",
"accession_no": "33456"
}
```

Map/Reduce view

Map function:

Find the documents with patient_name

```
function(doc) {
    if(doc.patient_name) {
        emit(doc.patient_name,1);
    }
}
```

Reduce function:

Aggregate the documents per patient_name

```
function(keys, values, rereduce) {
    return sum(values);
}
```

Map/Reduce query output

Map/Reduce result	
key (patient_name)	**value (Number of studies)**
Dany Targaryen	2
Sheldon Cooper	2
Hurley Reyes	1

Figure 5.1: The difference between RDBMS and CouchDB (adopted from (Rascovsky et al., 2012)).

5.1.1.2. In memory cache to store user data

Redis (http://redis.io) is an open source in memory key-value cache unlike CouchDB which is a persistent document-oriented database. SGSautoSNPdb uses Redis to store user session and query result ids on the server which are necessary for the pagination.

5.1.2. New trends in website design

In 2014 Mobile Internet users have overtaken Desktop Internet users (see Figure 5.1). Responsive Web Design uses the same front-end code for the website across devices of various sizes in order to provide the same user experience across all devices and screen sizes. SGSautoSNPdb uses Twitter's Bootstrap (http://getbootstrap.com/) which is the most popular open source HTML5, CSS and JavaScript framework for developing responsive websites in order to allow biologists to do research anywhere, anytime and on any device. Bootstrap uses CSS media queries in order to adjust the website layout for mobile or desktop devices.

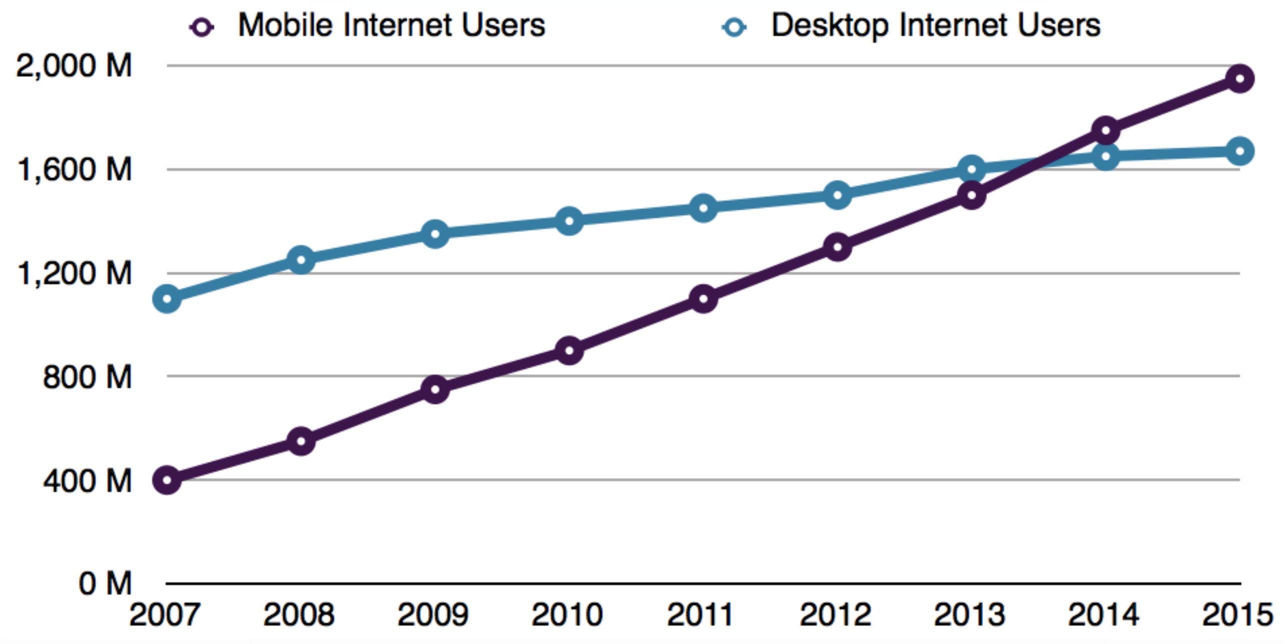


Figure 5.2: Number of global users in millions (adapted from http://smustalks.appspot.com/japan-12, 05 March 2015)

5.2. Material and Methods

5.2.1. Architecture of SGSautoSNPdb

SGSautoSNPdb uses a three-tier architecture which contains a presentation, application and database layer. The presentation layer contains the user interface which allows users to perform queries and retrieve results. This layer uses Twitter’s Bootstrap 3 in order to adjust automatically the SGSautoSNPdb’s user interface to different screen sizes. The database layer contains CouchDB which stores all the SNP data discovered by SGSautoSNP. The application layer is built with Flask (http://flask.pocoo.org/) which is a Python web framework. This layer is responsible to interact between the application and database layer. Furthermore, it uses Redis store user session and query result ids which are necessary for the fast pagination.

5.2.2. Loading *Brassica* SNPs to SGSautoSNPdb

For reading and updating (add, delete, edit) database documents CouchDB provides a RESTful (Representational state transfer) API (Application Programming Interface) which uses standard HTTP (Hypertext Transfer Protocol) methods (GET, PUT, POST, or DELETE). Any programming language which supports HTTP requests can interact with CouchDB and as response it returns JSON format.

In order to avoid using HTTP requests directly from the SGSautoSNPdb's loading script, *loadDB.py* (Figure 5.3), it uses a Python CouchDB driver (CouchDB-Python, https://code.google.com/p/couchdb-python/) that wraps REST requests into a convenient Python API. The wheat project described in Chapter 4 used SGSautoSNP version 1 and the method and results were published (Lorenc et al., 2012, Berkman et al., 2013). After SGSautoSNP version 1 was first published, new features were implemented to SGSautoSNP version 2. In SGSautoSNP v2, each output file contains a unique SNP id which makes it simple to build a SNP document out of different files and load it to SGSautoSNPdb. Furthermore, the loading script uses a SnpEff parser from the open source project (Paila et al., 2013). SGSautoSNPdb requires that the results were run with SGSautoSNP version 2. Therefore, only Brassica results described in Chapter 3 were loaded to SGSautoSNPdb. Further software dependencies can be found in Appendix.

```
$ python loadDB.py -h
usage: loadDB.py [-h] --project_dir [PROJECT_DIR]
                      --chr_name [CHR_NAMES] --specie [SPECIE]
                      --db_name [DB_NAME]

Load data to SGSautoSNPdb

optional arguments:
   -h, --help            show this help message and exit
  --project_dir [PROJECT_DIR]
                 Please give provide project dir full path.
  --chr_names [CHR_NAMES]
                        A list of unique chromosome abbreviation and
                        Chromosome folder name seperated by ':' e.g.:
                        'chr1:XA01_v3.0;chr2:XA02_v3.0'
  --specie [SPECIE]     Name of the specie e.g. "Brassica napus".
  --db_name [DB_NAME]   Name to the new database.
```

Figure 5.3: The command-line of the loadDB.py script, showing the various usage options.

Figure 5.4 shows one SNP document in JSON format taken out of SGSautoSNPdb by using CouchDB’s Futon, it is an administration tool which allows management of databases and modification of individual documents in the database (Silbermann et al., 2013). SNP id (_id), scaffold position (scaffoldPos), chromosome position (chrPos), scafffold name (scaffoldName), allele, SNP score (snpScore) and genotypes (genoTypes) are retrieved from the "chrN/SNPs/fileN_cont_out.snp" file. Whether a SNP is located in a low SNP density region (lowSNPregion) has been taken from the "chrN/gene_analysis/fileN_LOW_SNPID_geneID_GO.tab" file. Lines 6 - 18 contain all gene information, which comes from the "chrN/gene_analysis/fileN_SNPID_geneID_GO.tab" file and SNP annotation information (lines 24 - 36) is retrieved from the "chrN/snpEff/fileN_chr_out_only_genes.vcf" file. More information about the snpEff fields are described in Table 5.1. Marker specific information (lines 39-45) is retrieved from the "chrN/markers/fileN_GoldenDB.csv" file. Species and chromosome name (chrName) are given as a parameter to *loadDB.py*.

```
{
  "_id": "UQXAH010000004",
  "_rev": "1-e7157a93624cc8ef1e63c9c85c524202",
  "lowSNPregion": false,
  "specie": "Brassica napus",
  "genes": [
    {
      "geneEnd": 5754,
      "uniprotId": "Q3TPR7",
      "geneStart": 4609,
      "goIds": [
        "GO:0003674",
        "GO:0008150",
        "GO:0016020",
        "GO:0016021"
      ]
    }
  ],
  "mapping": {
    "scaffoldPos": 5442,
    "chrPos": 5442
  },
  "chrName": "chr1",
  "snpEff": {
    "effectSeverity": "LOW",
    "isCoding": false,
    "aaLength": null,
    "exon": null,
    "codonChange": null,
    "isExonic": false,
    "isLof": false,
    "gene": "Q811P0",
    "transcript": "Transcript_XA_0011r-snap.5",
    "aaChange": null,
    "biotype": null
  },
  "scaffoldName": "XA_0011r",
  "allele": "C/T",
  "marker": {
    "germplasm": "6_canola_lines",
    "fivePrimer": "ACAA...GAGC",
    "threePrimer": "GAGT...AAGG",
    "library": "UQ_BNSNP",
    "panel": "UQ_BNSNP_H_V3.0"
  },
  "snpScore": 2,
  "genotypes": [
    {
      "baseNo": 7,
      "base": "T",
      "cultivar": "A"
    },
    {
      "baseNo": 1,
      "base": "C",
      "cultivar": "Sr"
    },
    {
      "baseNo": 1,
      "base": "C",
      "cultivar": "Bn"
    },
    {
      "baseNo": 19,
      "base": "T",
      "cultivar": "N"
    },
    {
      "baseNo": 1,
      "base": "T",
      "cultivar": "S"
    },
    {
      "baseNo": 8,
      "base": "T",
      "cultivar": "T"
    }
  ]
}
```

Figure 5.4: One SNP document in JSON format taken out of SGSautoSNPdb.

Table 5.1: The effect types predicted by SnpEff and load into SGSautoSNPdb.

Effect type	**Effect description**
effectSeverity	Effect severity (LOW, MED, HIGH)
isCoding	Whether SNP is located in a coding region (except 3' & 5' UTR's of exons)
aaLength	CDS lenght in number of amino acids
exon	Exon information for SNPs that are exonic
codonChange	Codon change
isExonic	Whether the SNP affect an exon for this transcript
isLof	Whether the SNP is LOF?
gene	Gene affected by the SNP.
transcript	Transcript affected by the SNP.
aaChange	What kind of amino acid change?
biotype	Type of transcript e.g. protein-coding, rRNA etc.

5.3. Results and Discussion

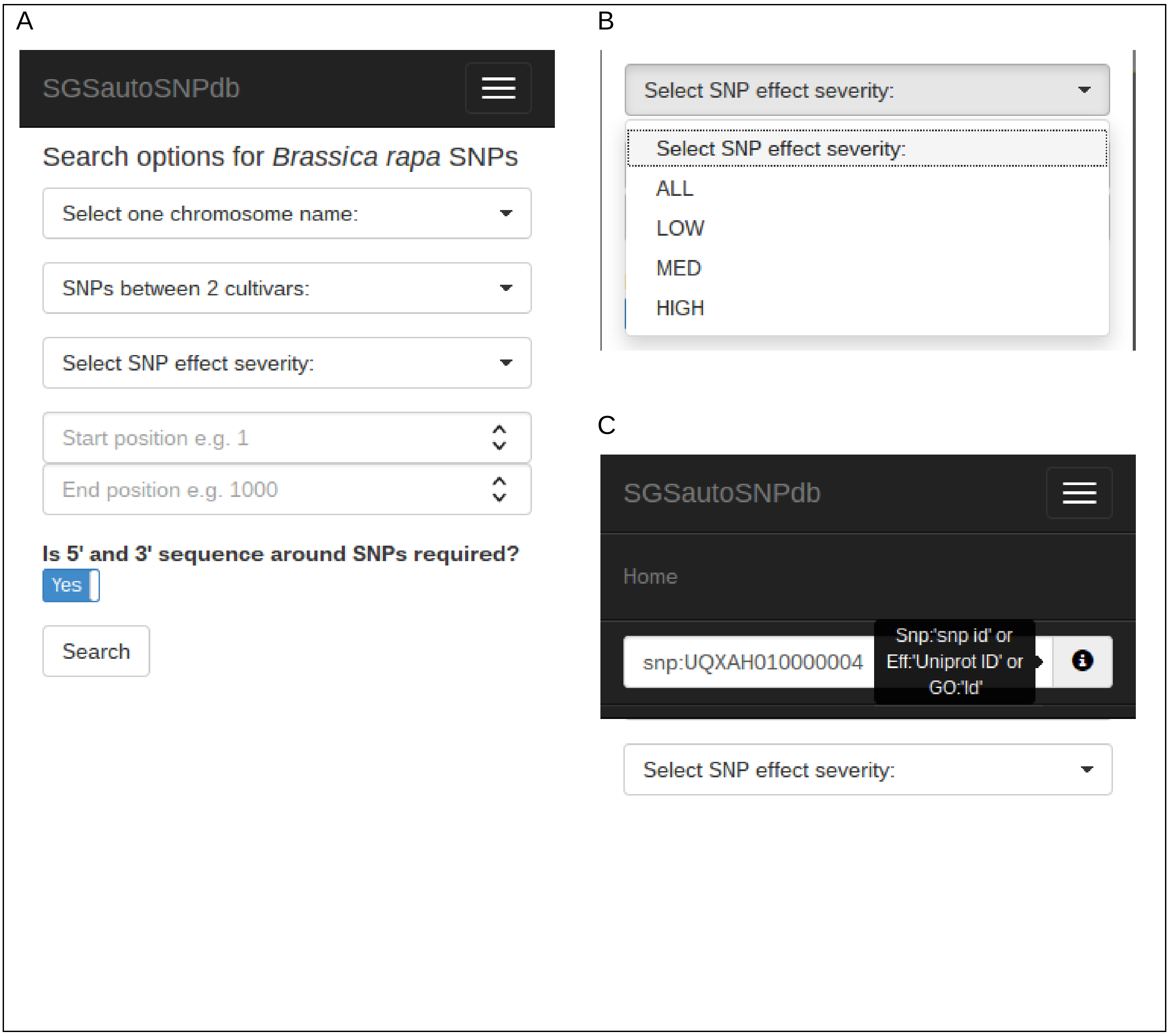
5.3.1. Two search interfaces

The aim of this project was to develop a responsive website in order to allow biologists to do research with the results produced by SGSautoSNP pipeline anywhere, anytime and on any device. SGSautoSNPdb provides two different ways to search for SNPs. The first one provides advanced search options (see Figure 5.5A) such as:

- find SNPs between two cultivars
- find SNPs in range (start to end)
- find SNPs which have a particular SNP effect severity (see Figure 5.5B)

Experienced users who are already familiar with SGSautoSNPdb can use the quick search option. This option can be found by clicking SGSautoSNPdb's menu icon (▤) on the top right side (see Figure 5.5 C). Quick search options allow searching for SNPs with known property ids such:

- for SNP id with this syntax SNP:"SNP id" for example SNP:UQXAH010000004
- for SNPs which have SNP effect gene id with this syntax EFF: "Uniprot ID" for example EFF:Q811P0
- for SNPs with a particular GO id with the syntax GO: ID for example GO:0000001



A
SGSautoSNPdb
Search options for *Brassica rapa* SNPs
Select one chromosome name:
SNPs between 2 cultivars:
Select SNP effect severity:
Start position e.g. 1
End position e.g. 1000
Is 5' and 3' sequence around SNPs required?
Yes
Search
B
Select SNP effect severity:
Select SNP effect severity:
ALL
LOW
MED
HIGH
C
SGSautoSNPdb
Home
snp:UQXAH010000004
Snp:'snp id' or Eff:'Uniprot ID' or GO:'Id'
Select SNP effect severity:

Figure 5.5: SGSautoSNPdb provides two search interfaces. Figure A shows the advanced search option and Figure B show the four possibilities for SNP effect severtity. On the other hand, Figure C shows the quick search interface for user who familiar with SGSautoSNPdb.

5.3.2. Step by step cultivar and range search

Figure 5.6 and Figure 5.7 shows screenshots of a step by step cultivar and range search in SGSautoSNPdb. Figure 5.6 A shows chromosome 1, SNPs between cultivars Ningyou (N) and Tapidor (T), all SNP effect severtity and the range from 1 to 10000 has been chosen. After clicking the search button, SGSautoSNPdb shows the first 20 SNPs (Figure 5.6 B). In this search SGSautoSNPdb has found 21 SNPs which means that pagination has been activated in order to show the additional SNP on next page. By clicking one of the SNP ids SGSautoSNPdb shows following SNP details (Figure 5.6 C and Figure 5.7 D -

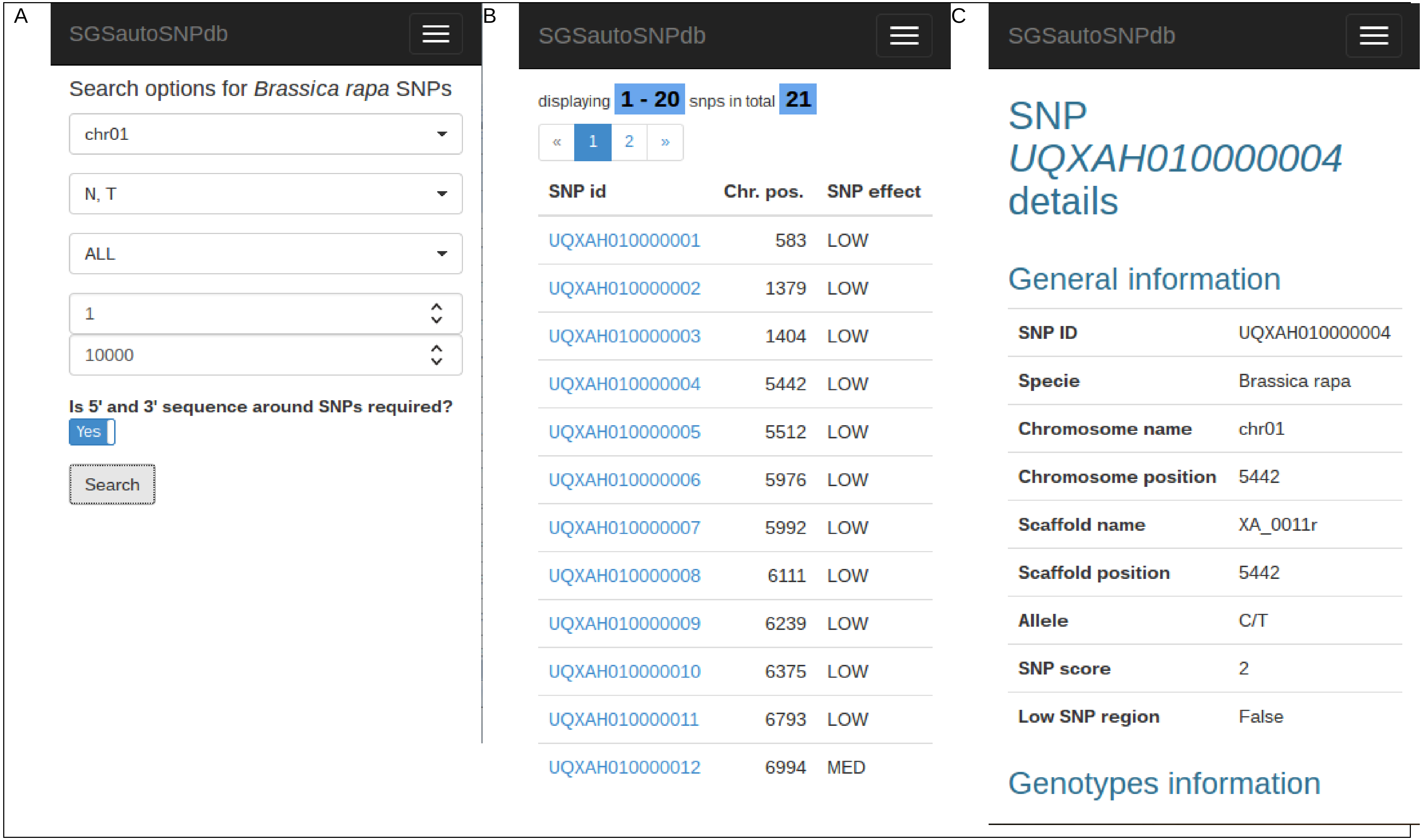
F):

- General information,
- Genotypes information,
- SNP effects and
- Marker information.

Almost all entries in SNP effects contain an information icon (ⓘ) which shows explanation from Table 5.1 when the users mouse over it. Since SGSautoSNPdb is a web application it can easily link the GO and UniProt ids to corresponding entries to third parties databases (Table 5.2).

Table 5.2 shows prefixes used in SGSautoSNPdb links for GO and UniProt Ids to their entries in the corresponding web-based repositories. The bold<ID> are place holder for real Ids.

Annotation	**Prefix URL**
Gene Ontology	http://www.uniprot.org/uniprot/**<Id>**
SwissProt	http://www.ebi.ac.uk/QuickGO/GTerm?id=**<Id>**



SGSautoSNPdb

Search options for *Brassica rapa* SNPs

chr01

N, T

ALL

1

10000

Is 5' and 3' sequence around SNPs required?

Yes

Search

SGSautoSNPdb

displaying **1 - 20** snps in total **21**

« 1 2 »

SNP id	Chr. pos.	SNP effect
UQXAH010000001	583	LOW
UQXAH010000002	1379	LOW
UQXAH010000003	1404	LOW
UQXAH010000004	5442	LOW
UQXAH010000005	5512	LOW
UQXAH010000006	5976	LOW
UQXAH010000007	5992	LOW
UQXAH010000008	6111	LOW
UQXAH010000009	6239	LOW
UQXAH010000010	6375	LOW
UQXAH010000011	6793	LOW
UQXAH010000012	6994	MED

SGSautoSNPdb

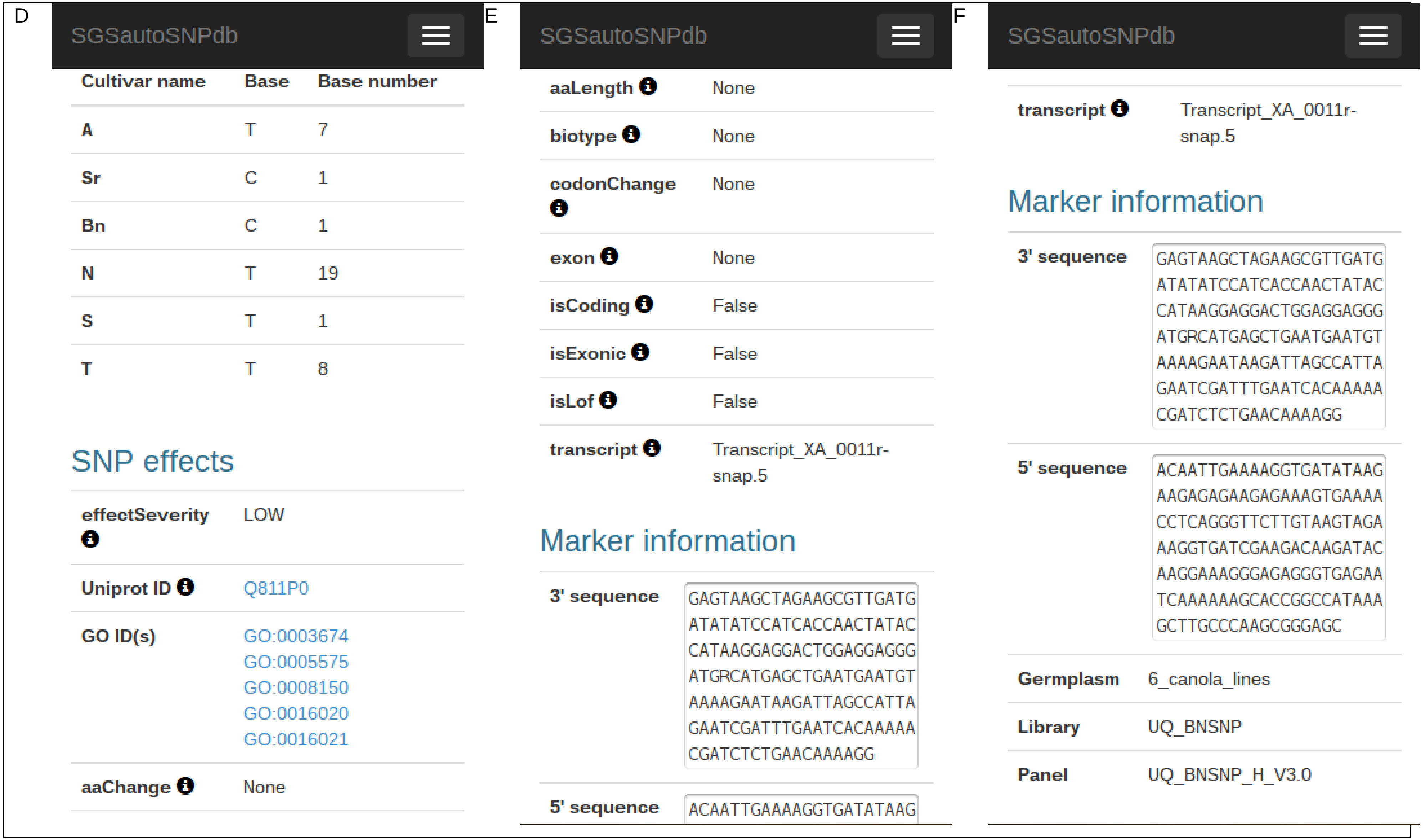
SNP *UQXAH010000004* details

General information

SNP ID	UQXAH010000004
Specie	Brassica rapa
Chromosome name	chr01
Chromosome position	5442
Scaffold name	XA_0011r
Scaffold position	5442
Allele	C/T
SNP score	2
Low SNP region	False

Genotypes information

Figure 5.6 shows an example search (A-B) and detail overview of a SNP (C-F). Screenshots of the desktop version can be found in Appendix.



SGSautoSNPdb

Cultivar name	Base	Base number
A	T	7
Sr	C	1
Bn	C	1
N	T	19
S	T	1
T	T	8

SNP effects

effectSeverity ℹ	LOW
Uniprot ID ℹ	Q811P0
GO ID(s)	GO:0003674 GO:0005575 GO:0008150 GO:0016020 GO:0016021
aaChange ℹ	None

SGSautoSNPdb

aaLength ℹ	None
biotype ℹ	None
codonChange ℹ	None
exon ℹ	None
isCoding ℹ	False
isExonic ℹ	False
isLof ℹ	False
transcript ℹ	Transcript_XA_0011r-snap.5

Marker information

3' sequence	GAGTAAGCTAGAAGCGTTGATG ATATATCCATCACCAACTATAC CATAAGGAGGACTGGAGGAGGG ATGRCATGAGCTGAATGAATGT AAAAGAATAAGATTAGCCATTA GAATCGATTTGAATCACAAAAA CGATCTCTGAACAAAAGG
5' sequence	ACAATTGAAAAGGTGATATAAG

SGSautoSNPdb

transcript ℹ	Transcript_XA_0011r-snap.5

Marker information

3' sequence	GAGTAAGCTAGAAGCGTTGATG ATATATCCATCACCAACTATAC CATAAGGAGGACTGGAGGAGGG ATGRCATGAGCTGAATGAATGT AAAAGAATAAGATTAGCCATTA GAATCGATTTGAATCACAAAAA CGATCTCTGAACAAAAGG
5' sequence	ACAATTGAAAAGGTGATATAAG AAGAGAGAAGAGAAAGTGAAAA CCTCAGGGTTCTTGTAAGTAGA AAGGTGATCGAAGACAAGATAC AAGGAAAGGGAGAGGGTGAGAA TCAAAAAAGCACCGGCCATAAA GCTTGCCCAAGCGGGAGC
Germplasm	6_canola_lines
Library	UQ_BNSNP
Panel	UQ_BNSNP_H_V3.0

Figure 5.7 shows an example search (A-B) and detail overview of a SNP (C-F) (continue).

5.3.3. Benchmarks of SGSautoSNPdb

It took 7.5 hours to load all 10 *Brassica* AA chromosome SNP data discovered in Chapter 3 into SGSautoSNPdb with help of *loadDB.py* script. For all SGSautoSNPdb possible queries it was required to generate 13 CouchDB views (see Appendix) which took on average 20 minutes to generate. Searching for SNPs in SGSautoSNPdb takes in average 2 seconds, but as soon as the results were found and stored in cache, pagination throught the results takes 1 second.

5.4. Conclusion

SGSautoSNPdb collects all SNPs and annotation data discovered by the SGSautoSNP pipeline and stores them into a flexible database. These data are accessible by any device and any display size, providing a valuable source of cultivar identification and annotated SNPs for applications such as genetic diversity analysis. Furthermore, SGSautoSNPdb provides addition links to third party resources for GO and UniProt ids.

Future work should include the expansion of loading data from all *Brassica* species, chickpea and wheat. Funding or developing a genome browser which also supports a Responsive Web Design and SGSautoSNPdb could link SNP positions to it and graphically visualize what happens at the SNP position and around it. The reason why it took 7.5 hours to load all 10 *Brassica* AA chromosome SNPs data to couchdb is because *loadDB.py* script collects all information for one SNP and then sends it to couchdb. After sending SNP data to couchdb, *loadDB.py* waits for confirmation that the SNP has been saved in couchdb. In order to avoid the waiting period, two improvements could be implemented. The first one would be to collect all SNP data for one chromosome and then send it to couchdb as a batch job. Secondly, *loadDB.py* could be extended to use more than one CPU core in order to process more than on chromosome at a time.

Chapter 6: Concluding remarks and future directions

6.1. Concluding remarks

The data for *Brassica* and wheat in this thesis were derived from Second Generation Sequencing (SGS) technology. SGS has revolutionized biological research in the 21st century. Since the introduction of this technology in 2005, by 454 Life Sciences and commercialised by Roche as the GS20 (Margulies et al., 2005), the price of sequencing “per bp” has been decreasing with the rapid increase of sequencing speed and amount of sequencing data generated per run. SGS has been successfully used for *de novo* genome sequencing, as well as re-sequencing of genomes. Researchers are now able to perform their research on complex reference crop genomes instead of only model organisms. Model organisms were previously selected for genome sequencing due to their small genome size, low complexity and fast life-cycle, however they frequently had little application in the field. In some cases they can also be related to crop species of interest, for example *Brassica*’s ancestor is shared with *Arabidopsis thaliana*. Researchers were using these model organisms to understand the fundamental structure and functions of important agricultural crop species. Using reference genomes of important crop species, rather than translation from model organisms, will allow researchers to understand evolution, functionality and genetic structure questions much better than ever before.

Many SGS sequencing technologies are able to produce paired-read sequences with different fragment and insert sizes. These paired-end sequences help to overcome repetitive sequence issues (Robison, 2010). Illumina will be producing longer reads through the newly acquired Moleculo technology. This new technology breaks DNA into large fragments that are than sequenced using Illumina’s standard sequencing technologies. Longer reads can be used to increase the haplotype resolution and will help to explore the changes in haplotype structure and composition.

The rapid increase in sequence data produced by SGS is significantly exceeding the rate of increase in disk space through the production of longer reads and increasing volumes of data per run. A solution to use less disk space and to be able to rapidly access reads in a particular position could be a reference-based compression method (Fritz et al., 2011). In this method the sequences are aligned to a reference genome and then only the differences between the reference genome and the aligned sequence are saved, unlike BAM files which store the whole alignment (Li et al., 2009a).

In chapter 2 the SGSautoSNP pipeline (Lorenc et al., 2012) is described. This was designed to call SNPs for homozygous species. This method does not consider the reference genome for SNP discovery. Instead, the reference is used to assemble the cultivar reads, and SNPs are then called between these assembled reads. In SGSautoSNP, mismapped reads produce a heterozygous genotype call at a locus, allowing their distinction from true homozygous SNPs. SGSautoSNP is able to cope with increasing input data sizes generated by SGS technologies. It was designed to run on multi-core processors and uses a workaround for Python's Global Interpreter Lock (GIL) issue. GIL prevents multiple threads to make effective use of multiple core CPUs.

In Bioinformatics many formats are not properly defined or implemented which makes it difficult to transfer them between different tools. Many new file formats were developed to solve previous file format problems or to introduce new features. However, often new formats were no longer compatible to the previous format. For example, GFF3 (General Feature Format) is not compatible to the older GFF2. Because of the incompatibility, new software libraries have to be written to enable access to the new file format information.

Usually a person or organisation who invented a new format provides a library in a particular language e.g. in C, but programmers are using different programming languages e.g. Python. Therefore a wrapper has to be written so Python can use the C library. This has problems, because each time the C library gets improved, the wrapper for Python has also to be updated which leads to a delay in using the new improvements or bug fixes. It would be ideal if all bioinformatics libraries would use a “Simplified Wrapper and Interface Generator” (SWIG, http://www.swig.org) which is a software development tool that connects programs written in C and C++ with a variety of high-level programming language such as Python.

Instead of developing new file formats it may be possible to use comma or tab delimited file formats. This would work e.g. for FASTA format which contain no relations between each entry, but it would not work where relationships between entries are important such as in GFF3. Relationship information could be stored in file formats which already exist such eXtensible Markup Language (XML, www.w3.org/TR/REC-xml), JavaScript Object Notation (JSON, http://json.org) or YAML Ain't Markup Language (YAML, http://www.yaml.org).

Many programming languages provide parsers for these three formats. Storing the same information with the smallest file size could be achieved with YAML and JSON but a larger file size is required with XML, mainly because of XML's closing tags. XML and JSON provide a binary format, Efficient XML Interchange (EXI, www.w3.org/TR/exi/) and Binary JSON (BSJON, http://bsonspec.org). EXI makes XML data up to hundreds of times smaller, increases processing speed, and increasing the transmission speed of XML across existing networks (http://www.agiledelta.com/product_efx.html, 16 November 2013). BSON provides efficient encoding/decoding compared to JSON, but the file size might be bigger than the plain JSON format.

Before we started to develop SGSautoSNP we decided to support previously available bioinformatics tools and formats to prevent duplication of work. It was a challenge to get SGSautoSNP output files to work with other tools because some of the file formats used by other tools are not well documented.

The SGSautoSNP pipeline produces a variety of output formats for visualisation and validation. The SGSautoSNP pipeline provides an unprecedented resource for diversity analysis, and establishes a foundation for high resolution SNP discovery in large and complex genomes. After the SGSautoSNP method and its application for wheat SNP discovery were published (Berkman et al., 2013, Lorenc et al., 2012), more features were implemented gaining more information than was previously possible. The SGSautoSNP pipeline now includes scripts for gene annotation, which uses SNAP (Korf, 2004), a gene prediction tool, and SNPeff (Cingolani et al., 2012), a SNP annotation and effect prediction tool. SGSautoSNP now uses a predefined directory structure to satisfy the compatibility requirements of SnpEff. In addition, the SGSautoSNP pipeline identifies SNPs in low SNP density regions and gene ontology analysis can be performed using goatools (https://github.com/tanghaibao/goatools) to identify if there is an enrichment of GO terms. The challenge in developing this pipeline was various formats which SGSautoSNP has to support to allow compatibility with other bioinformatics tools. It would be desirable that the current formats would be better documented rather than new formats invented.

Chapters 3 and 4 describe the result of using the SGSautoSNP pipeline for *Brassica* and wheat, respectively. Between 10.5 and 17.3% of paired Brassica reads could be mapped to the reference genome which was more than the wheat mapping for four wheat cultivars where only 3.1 to 5.1% of paired reads mapped to the group 7/4AL chromosomes arms. This is due to read pairs mapping to multiple locations in repetitive genomes and subsequently being ignored due to the SOAPaligner *–r 0* option and the fact that only a portion of the wheat genome is represented in the arm references.

Understanding heritable traits in crops has been accelerated worldwide by the application of molecular markers, because they allow the selection of plant characteristics without the requirement and expense of phenotyping. The rapid increase in the accessibility of genome sequence data allows the identification of genetic markers and genes underlying key traits for use in molecular breeding and crop improvement. A total of 638,593 SNPs in the *Brassica* AA genome and 881,289 SNPs in the wheat group 7 chromosome arms were identified using the SGSautoSNP pipeline. Validation of 20 *B. napus* AA genome SNPs resulted in a SNP prediction accuracy of around 95%. Of the 28 wheat SNPs that were used for validation of the SGSautoSNP pipeline, 26 (93%) produced the expected genotype. In another project Zander et al. identified 21,814 SNPs in *Leptosphaeria maculans* between two isolates. Of the 20 *L. maculans* SNPs that were used for validation, 18 (90%) produced the expected genotype (Zander et al., 2013). The *Brassica, L. maculans* and wheat validation confirms that the SGSautoSNP algorithm is accurate and works in small, as well as large and complex genomes, producing homoeologue specific markers. By combining the SGSautoSNP pipeline together with SnpEff it was possible to determine whole genome SNP trends, transition to transversion ratios and SNP frequencies across chromosomes. Annotation of *B. napus* AA genome SNPs have revealed that 0.5% of predicted SNPs are classified as “high effect” SNPs, and these could impact the structure of the proteins or the amino acid transcripts. Furthermore, the transition/transversion (Ts/Tv) ratio ranges from 1.20 to 1.26 across all 10 *Brassica napus* AA chromosomes. These values are comparable with other plants such as 1.6 in eggplant (Barchi et al., 2011), 3.9 in maize, 1.9 in alfalfa, 1.6 in eikorn wheat (*Triticum monococcum* L.), and 2.5 in barley and Lotus (Vitte and Bennetzen, 2006).

Chapter 5 describes the development of SGSautoSNPdb which is a web application with responsive web design in order to allow researchers to work on any device, any screen sizes and anywhere. As a database SGSautoSNPdb uses couchDB which is a document-oriented database. For biologists a document-oriented database is easier and faster to understand than RDBMS where the data is spread across different tables. SGSautoSNPdb at the moment contains only the SNPs and annotations discovered by SGSautoSNP from *Brassica* 10 AA chromosomes described in chapter 3.

Together the SGSautoSNP pipeline and SGSautoSNPdb provide tools to help us to understand how natural selection has shaped the evolution of crop genomes and SNPs that can be applied to improve crops.

6.2. Future direction

With the continued decrease in the "per bp" price of SGS sequencing and advent of the successor; Third Generation Sequencing technologies, an increasing amount of sequence data will be generated which allows the discovery and application of molecular markers. The use of these markers will assist breeders and researchers to speed up crop improvement in a greater diversity of species than ever before. The massive amount of data generated by SGS or 3GS will require new efficient bioinformatics tools which are able to be easily scaled up.

Draft genome sequences of wheat were recently published; *Triticum aestivum* (Bread wheat) has a hexaploid AABBDD genome (Brenchley et al., 2012), *Aegilops tauschii* has a diploid DD genome (Jia et al., 2013) and *Triticum urartu* has a diploid AA genome (Ling et al., 2013). Genome sequence of individual bread wheat chromosomes arms have also been published; group 1 (1A, 1B, 1D) (Wicker et al., 2011), 4A (Hernandez et al., 2012), 5A (Vitulo et al., 2011), 5B (Sergeeva et al., 2014) and group 7 (7A, 7B and 7D) (Berkman et al., 2013, Berkman et al., 2012b, Berkman et al., 2011).

The *B. rapa* AA (Wang et al., 2011) and *B. oleracea* CC (Liu et al., 2014, Parkin et al., 2014) genomes have been published. In addition, draft genome sequences for *B. napus*, *B. juncea* AABB and *B. nigra* BB have been produced with the possibility of publication in the near future (Golicz et al., 2012). In future research, the published *B. oleracea* genome could replace the proprietary one used in Chapter 3 and the whole *Brassica* analysis of Chapter 3 could be published with all 19 *B. napus* chromosomes instead just the 10 chromosomes. However, the public *B. napus* genome will soon be published, which would be more appropriate to use than the diploid progenitors.

These draft genome sequences will enhance genetic studies and provide insight into the genetic basis of important agronomic traits including nutritional seed properties and resistance to biotic and abiotic stressors (Getinet et al., 1997). A more complete genome might improve the mapping accuracy, because reads previously mapped could mapped better on a new genome positions. These could affect positive the SNP discovery, because miss mapped reads could cause conflict between cultivars, which could cause of losing true SNPs.

SGSautoSNP is being used for a new method called Skim Based Genotyping by Sequencing (skimGBS). SkimGBS is an alternative genotyping approach for trait mapping and can be applied to characterise recombination and for genome-wide association studies. Furthermore, it can be used to improve genome assemblies or assess structural variation. In the first step, the genomes of parents of a mapping population are re-sequenced and data aligned to the reference genome. This is followed by using SGSautoSNP to predict SNPs. Next, multiple individuals from the population are skim re-sequenced at low coverage, for example between 0.1 -1.5x and their reads are mapped to the reference genome to genotype the previously predicted SNPs. Due to the low level of sequencing, coverage is not even along the whole genome and therefore some SNPs may not be identified in certain parts of the genomes. For these missing SNPs imputation is applied by using haplotype block information to replace missing genotypes. In order to increase the genotyping resolution of a selected individual additional sequence data can be generated (Golicz et al., 2012).

By re-running the SGSautoSNP pipeline with additional cultivars we could discover more novel SNPs, however, previously validated SNPs could disappear because the new cultivar might introduce a conflict. To avoid the loss of SNPs, *loadDB.py* could be modified in the following way. The loading script could check in SGSautoSNPdb whether the new SNP position is already allocated, if yes the new SNP could replace the old SNP and keep the old SNP id. If the new SNP position is not stored in SGSautoSNPdb then this SNP will be stored in the database with a modified SNP id e.g. attached suffix which is not part of other SNP ids. A new SNP id has to be created because it very likely that the SNP id is already assigned. Another way to limit the number of new redundant SNP IDs is by only generating a new ID if the SNP has not been seen before. It could be possible to implement an option in SGSautoSNP to accept a ".snp" file from previous SNP calling. From this file SGSautoSNP could limit the number of new redundant SNP IDs by only generating a new ID if we have not seen the SNP before. These would prevent that a predicted and/or validated SNP from the past would be removed by adding more cultivars which could cause a conflict. Not only adding new cultivars could remove a SNP because of a conflict, but updating to a newer version of reference genome could also cause this. Where previously the reads could align at a particular locus position, now with the updated genome the cultivar reads might not be able to align. In order to allow mapping of the cultivar reads to the new updated genome a new aligner called BWBBLE can be introduced which allows the user to include SNP information from previous SNP calling during the alignment (Huang et al., 2013a). However, the current BWBBLE version only supports single reads, but for this project paired reads were used.

Appendix A Cloud computing environment

Cloud computing is a system to run a program on many connected computers at the same time in a cluster. A High-Performance Computing (HPC) cluster called Barrine is located at the University of Queensland's St. Lucia campus. It hosts Bioinformatics Resource Australia-EMBL (BRAEMBL, http://braembl.org.au/), provides programmatic access to various data resources and analysis tools via Web Services technologies, and is used by scientists from all over Australia. Barrine contains 384 compute nodes with over 3000 CPU cores connected via an Infiniband fabric network. The majority of nodes have 24 GB of memory and 8 CPU cores. However three nodes have 1 TB memory and 32 CPU cores. Data storage includes a 92 TB parallel network file system (Panasas) and offline storage of 2PB. In order to provide the researcher a short response time and avoid overload of any one of the login nodes, Barrine uses a load balancer to distribute researcher workloads (https://ncisf.org/barrinehpc, 11 April 2014).

Job submission and execution in Barrine is handled by the Portable Batch System (PBS) which is used to provide computing resources across the available nodes in the Linux cluster. It provides tools to submit monitor and delete user jobs. PBS has three components, a job server, a job executor and job scheduler. A job server (pbs_server) receives the user job request and protects it against system crashes. A job executor (pbs_mom) receives a copy of the job from the job server, runs the job and returns the job's output to the user. Policy control manages which, where and when a job is run.

As described in the Chapter 1, crop genomes are large and complex, and Second Generation Sequencing technologies produce a huge amount of sequence data. Although the SGSautoSNP pipeline is not hugely compute intensive, the large amounts of data available for this project required extensive computing power to process the data in a reasonable timeframe. Barrine was used to distribute SGSautoSNP scripts across multiple compute nodes, and each chromosome was processed by the SGSautoSNP pipeline script on a single compute node. With this strategy all chromosomes could be processed in parallel.

In order to be able to submit a job to Barrine, a custom Bash script, also called a PBS script, has to be prepared (see Figure 6.1), which is a request for the resources from the compute node that will be needed, including the number of CPU cores, how long it will run, information about where the data is stored and where the working directory is. Table 6.1 provides an explanation about the commands used in the PBS script. This script is submitted to the pbs_server, and jobs that use more resources than allocated by the user in the PBS script are terminated by the policy controller.

```
#!/bin/bash

# Usage: qsub -J 1-11 -o $HOME/SGSautoSNP_XA.^array_index^.out
-e $HOME/SGSautoSNP_XA.^array_index^.err -v PROJECT_DIR=<My project
folder>,CULTIVARS="T;N;A;S;Bn;Sr",SNP_ID="UQXAH" SGSautoSNP_PBS.sh

# QCIF PBS commands
#PBS -N SGSautoSNP
#PBS -l select=1:ncpus=6:mem=60G:NodeType=large
#PBS -l walltime=10:00:00
#PBS -A sf-Y82

# Generic PBS commands

source /usr/share/modules/init/bash;
module load python;

sleep $(( ($PBS_ARRAY_INDEX % 10) * 15 ))

cd $PROJECT_DIR;
COUNTER=`printf "%02d" ${PBS_ARRAY_INDEX}`;
bam=`basename BAMs/*${COUNTER}*.bam .bam`
fasta_m=genomes_m/*${COUNTER}*.fa
fasta=`basename genomes/*${COUNTER}*.fa .fa`
outputfile=$bam"_"$fasta

COMMAND="SGSautoSNP.py --bam BAMs/*${COUNTER}*.bam --fasta $fasta_m --chr_offset
genomes/*${COUNTER}*.gff3 --cultivars $CULTIVARS --snp_id_prefix $SNP_ID$COUNTER
--contig_output $PROJECT_DIR/$fasta/SNPs/${outputfile}_contig_output --
chr_output $PROJECT_DIR/$fasta/SNPs/${outputfile}_chr_output --cpu $NCPUS";

echo $COMMAND;
$COMMAND;
```

Figure 6.1: A sample of the PBS script for running *SGSautoSNP.py.* The qsub command submits the job and allocates 11 nodes with the parameter -J 1-11.

Table 6.1: PBS options in a PBS job script file.

PBS options	Description
-o	Path and file name for standard output.
-e	Path and file name for standard error.
-V	All environment variables to the job such as Path to project directory, cultivars used and SNP ID.
-N	Job name such as the name of the PBS script
-l	The number of CPU cores (ncpus), memory (mem), the type of node (NodeType) and the maximum run time (walltime)

Appendix B SGSautoSNP pipeline dependencies

The SGSautoSNP pipeline is implemented in Python 2.7 and runs from the command line on any operating system where Python is available. It is recommended not to install software dependencies, such as sudo or root user, because later it makes it easier to clean up a folder where only SGSautoSNP dependencies are installed. To do this it is necessary to create the following folder structure with the *mkdir* Linux command:

```
$ mkdir -p <My Python path>/lib/python2.7/site-packages/ <My Python path>/lib64/python2.7/site-packages
```

In order that *easy_install*, a Python package installer, can install the dependencies in the new directory structure it is necessary to update the users' *bashrc*. The best way to perform this, is to use any text editor e.g. Vim to open the *bashrc* file.

```
$ vim ~/.bashrc
```

In the bashrc file users have to insert the following line:

```
export PYTHONPATH=<My Python path>/lib/python2.7/site-packages: <My Python path>/lib64/python2.7/site-packages:$PYTHONPATH
```

After this, the file has to be saved, and to make the changes active this command has to be executed:

```
$ source ~/.bashrc
```

The following commands are to install all Python packages which SGSautoSNP requires, with the help of *easy_install* a Python package installer:

- $ easy_install --prefix=<My Python path> -UZ numpy
  - Scientific computing library for Python
- $ easy_install --prefix=<My Python path> -UZ biopython
  - To access bioimformatics files (Cock et al., 2009)
- $ easy_install --prefix=<My Python path> -UZ pysam
  - To access SAM/BAM formats
- $ git clone git://github.com/chapmanb/bcbb.git
  $ easy_install --prefix=<My Python path> -UZ bcbb/gff
  - To access GFF formats
- $ git clone https://github.com/tanghaibao/goatools.git
  $ easy_install --prefix=<My Python path> -UZ .
  $ easy_install --prefix=/home/mictadlo/apps/pymodules -UZ fisher
  - To find enrichment of GO terms
- $ easy_install --prefix=<My Python path> -UZ pandas
  - Easy-to-use data structures and data analysis tools
- $easy_install --prefix=<My Python path> -UZ lxml
  - Support for XML and HTML parsing for Pythyon
- $ easy_install --prefix=<My Python path> -UZ beautifulsoup4
  - To parse HTML files

Other non Python requirements which are necessary for the SGSautoSNP pipeline are:

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) a quality control tool for high throughput sequence data.
- SOAP (Li et al., 2009b) a tool for short read alignment.
- soap2sam.pl (http://soap.genomics.org.cn/down/soap2sam.tar.gz) is used to covert SOAP results to SAM format.
- SAMtools (Li et al., 2009a) which provide various tools for manipulating alignments in the SAM/BAM format, including sorting, merging, indexing and converting SAM to BAM.
- Picard tools (Li et al., 2009a) provides MarkDuplicates.jar a tool to remove duplicates in alignments.
- Flapjack (Milne et al., 2010b) visualisation tool for genotyping.
- Semi-HMM-based Nucleic Acid Parser (SNAP) gene prediction tool (Korf, 2004).
- SnpEff (Cingolani et al., 2012) a variant annotation and effect prediction tool.
- Blast+ (Camacho et al., 2009) an alignment search tool.

Appendix C SGSautoSNP project structure

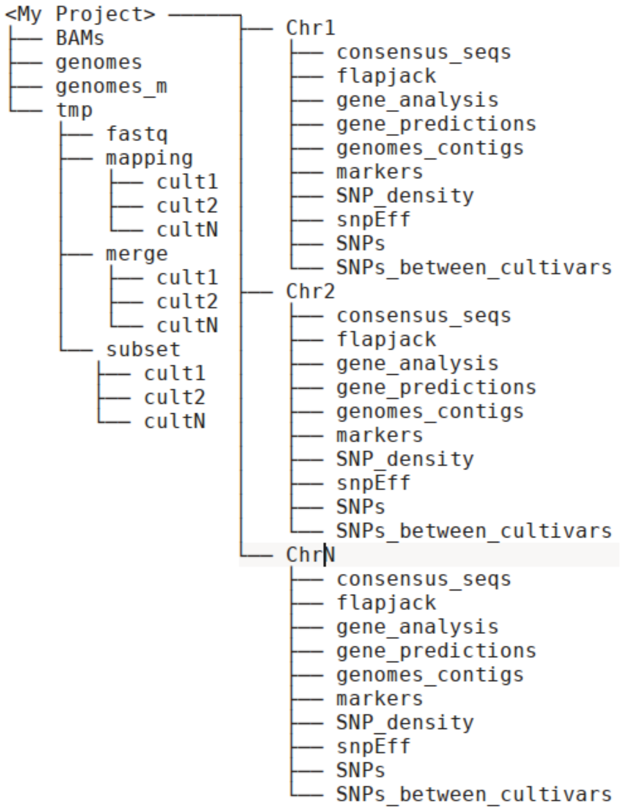
SnpEff (Cingolani et al., 2012) is a command line application, but it requires that all files have to be stored in special directories. However, SGSautoSNP allows input files which are stored in different or the same location, but it creates a lot of results files during the analysis. The SGSautoSNP pipeline extended SnpEff's directory structure in order to avoid the requirement to copy SGSautoSNP's files to the SnpEff directory structure and to provide a better overview of the SGSautoSNP's result files. To create the SGSautoSNP/SnpEff directory structure the following commands need to be executed:

- $ mkdir <My project folder>
- $ cd <My project folder>
- $ mkdir genomes genomes_m BAMs
    - *genomes folder* contains all chromosomes files in FASTA format where each of the sequences is concatenate by 100 Ns and GFF3 files which contains information about where original the sequence started.
    - *genome_m* folder contains all chromosomes files in FASTA format, but their sequences where not concatenated.
- $ mkdir -p tmp/{fastq,mapping,markDupl,merge,subset}
    - *fastq* folder contains all fastq files for all cultivars.
- $ mkdir -p tmp/mapping/{cult1,cult2,cult3}
    - *mapping* folder contains the results from *SOAPaligner.py* and sorted by cultivar names (cult1, cult2, cult3, …)
- $ mkdir -p tmp/markDupl/{cult1,cult2,cult3}
    - *markDupl* folder contains the results from *MarkDuplicates.py*
- $ mkdir -p tmp/subset/{cult1,cult2,cult3}
    - *subset* folder contains the results from *GenerateSubsetBAM.py*
- $ mkdir -p tmp/merge/{cult1,cult2,cult3}
    - *merge* folder contains the results from *MergeChrs.py*
    - *BAMs* folder contains for each chromosome a BAM file and its index file from *MergeChrs.py*
- $ for i in {1..11}; do mkdir -p `printf "XA%02d_v3.0" $i`/{consensus_seqs,gene_analysis,gene_predictions,genomes_contigs,markers, SNP_density,SNPs,SNPs_between_cultivars,snpEff}; done
    - In the above for loop the user can specify how many chromosomes are available. In the above example there were 11 chromosomes
    - *gene_analysis* folder stores the output of *gene_analysis.py* script
    - *gene_predictions* folder stores the output of *gene_annotation.py* script
    - *SNP_density* folder stores the output of *snp_density_coverage_percentage.sh script*
    - *SNPs* folder stores the output of *SGSautoSNP.py* script

- ○ *SNPs_between_cultivars* folder stores the output of *filter_snps.py* script
- ○ *snpEff* folder stores the output of SnpEff.jar application

Figure 6.2 shows a tree representation of the project directory structure. SnpEff does not recognize FASTA file with the "*.fasta*" extension and therefore the extension must be changed to "*.fa*". The best way to do this is to use the "*rename*" Linux command for all files in a directory in the following way:

- $ <My project folder>/genomes> rename .fasta .fa *.fasta
- $ <My project folder>/genomes_m> rename .fasta .fa *.fasta



```
<My Project>
├── BAMs
├── genomes
├── genomes_m
├── tmp
│   ├── fastq
│   ├── mapping
│   │   ├── cult1
│   │   ├── cult2
│   │   └── cultN
│   ├── merge
│   │   ├── cult1
│   │   ├── cult2
│   │   └── cultN
│   └── subset
│       ├── cult1
│       ├── cult2
│       └── cultN
├── Chr1
│   ├── consensus_seqs
│   ├── flapjack
│   ├── gene_analysis
│   ├── gene_predictions
│   ├── genomes_contigs
│   ├── markers
│   ├── SNP_density
│   ├── snpEff
│   ├── SNPs
│   └── SNPs_between_cultivars
├── Chr2
│   ├── consensus_seqs
│   ├── flapjack
│   ├── gene_analysis
│   ├── gene_predictions
│   ├── genomes_contigs
│   ├── markers
│   ├── SNP_density
│   ├── snpEff
│   ├── SNPs
│   └── SNPs_between_cultivars
└── ChrN
    ├── consensus_seqs
    ├── flapjack
    ├── gene_analysis
    ├── gene_predictions
    ├── genomes_contigs
    ├── markers
    ├── SNP_density
    ├── snpEff
    ├── SNPs
    └── SNPs_between_cultivars
```

Figure 6.2: Recommended SGSautoSNP pipeline project structure.

Appendix D SGSautoSNPdb software dependencies

SGSautoSNPdb uses Python in the backend and for the frontend JavaScript, HTML5 and CSS3. These are Python software dependencies:

- $ easy_install --prefix=<My Python path> -UZ Flask
  - Flask is a webframework for Python
- easy_install --prefix=<My Python path> -UZ flask-paginate
  - Flask-paginate is a paginate extension for flask
- easy_install --prefix=<My Python path> -UZ flask-wtf
  - Flask-wtf is a Flask extension for WTForms which provides forms validation.
- easy_install --prefix=<My Python path> -UZ CouchDB==0.9
  - CouchDB driver for Python.
- easy_install --prefix=<My Python path> -UZ redis
  - Redis driver for Python.
- easy_install --prefix=<My Python path> -UZ cyvcf
  - A fast Python library for VCF files.
- easy_install --prefix=<My Python path> -UZ gemini
  - This projected contains a SNPeff parser which *loadDB.py* uses.

Other non Python backend requirements:

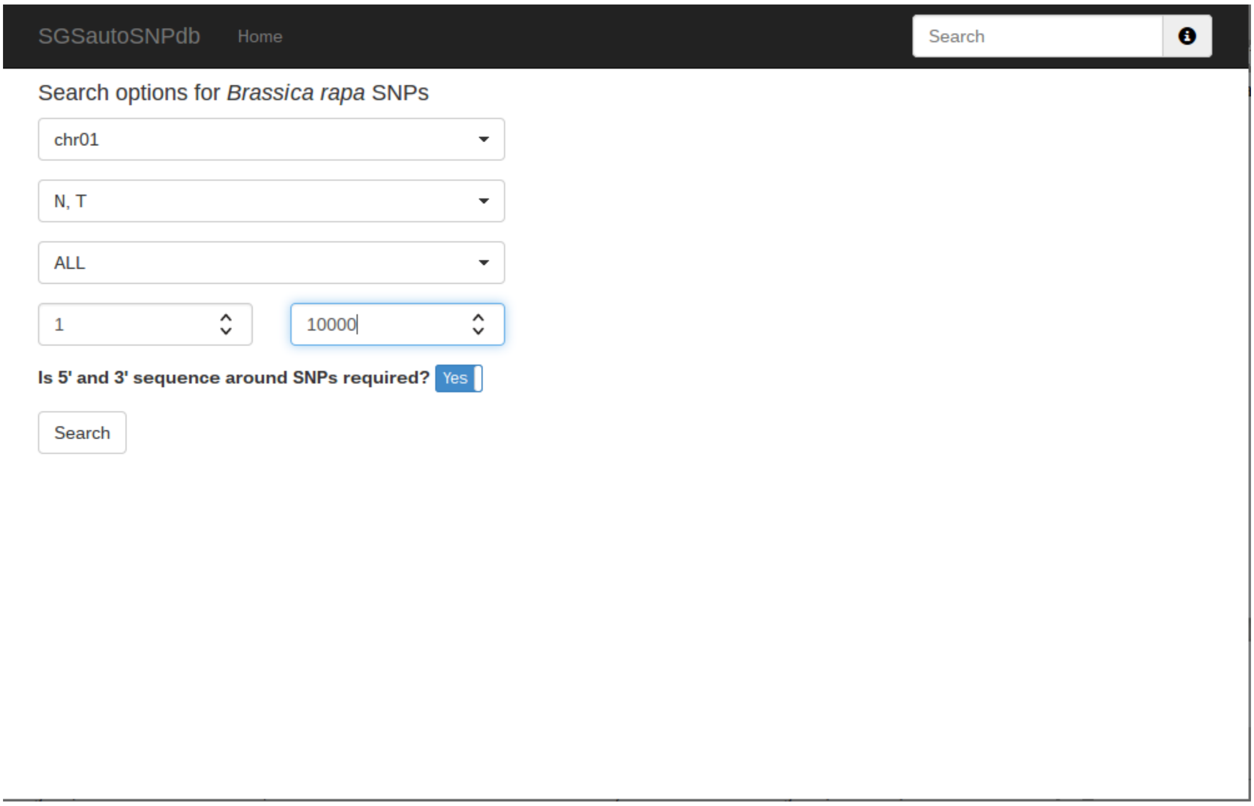
- CouchDB (http://couchdb.apache.org/)
- Redis (http://redis.io/)

The frontend dependencies are:

- Twitter's Bootsrap (http://getbootstrap.com/)
  - Allows Responsive Web Design
- Bootstrap-select (https://github.com/silviomoreto/bootstrap-select)
- An extension for Bootstrap to display nice looking select and multiselect dropdown menus.

Appendix E   Desktop version of SGSautoSNPdb

Screenshots of a step by step cultivar and range search in SGSautoSNPdb



Search options for *Brassica rapa* SNPs

chr01

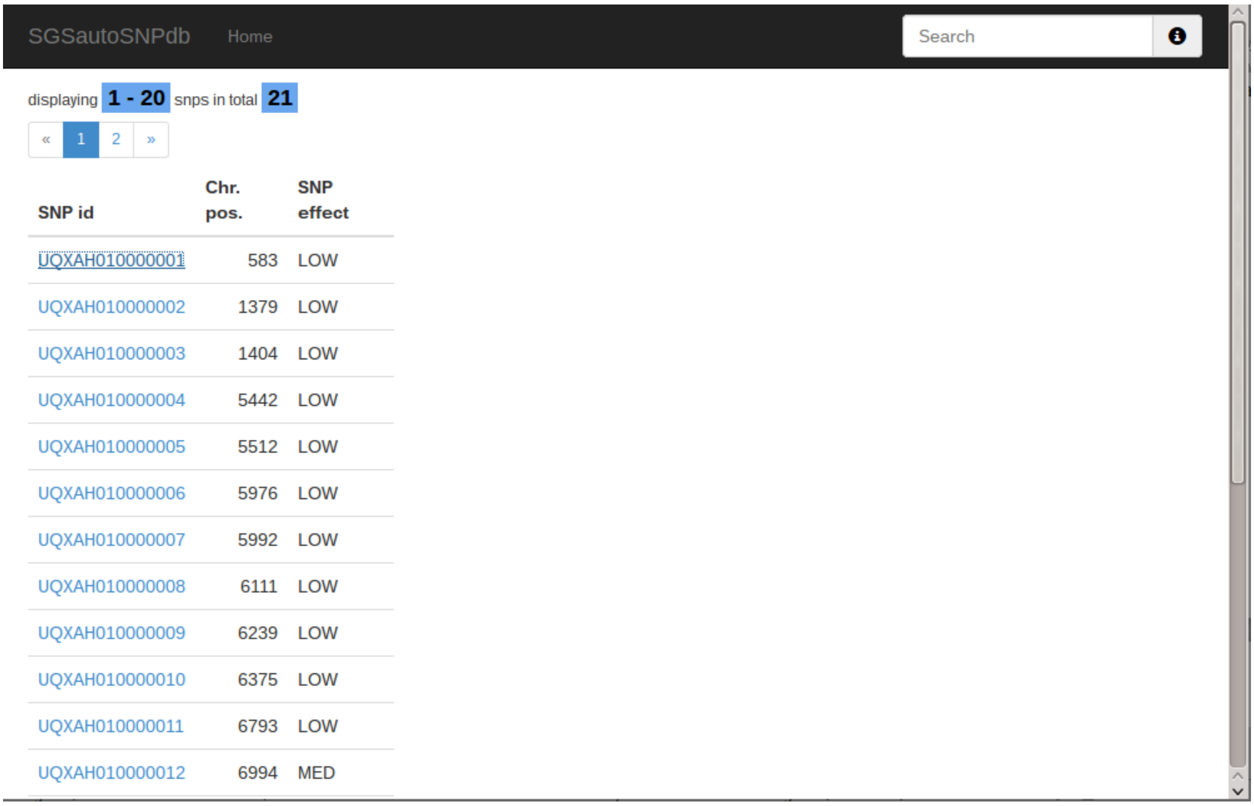
N, T

ALL

1   10000

Is 5' and 3' sequence around SNPs required? Yes

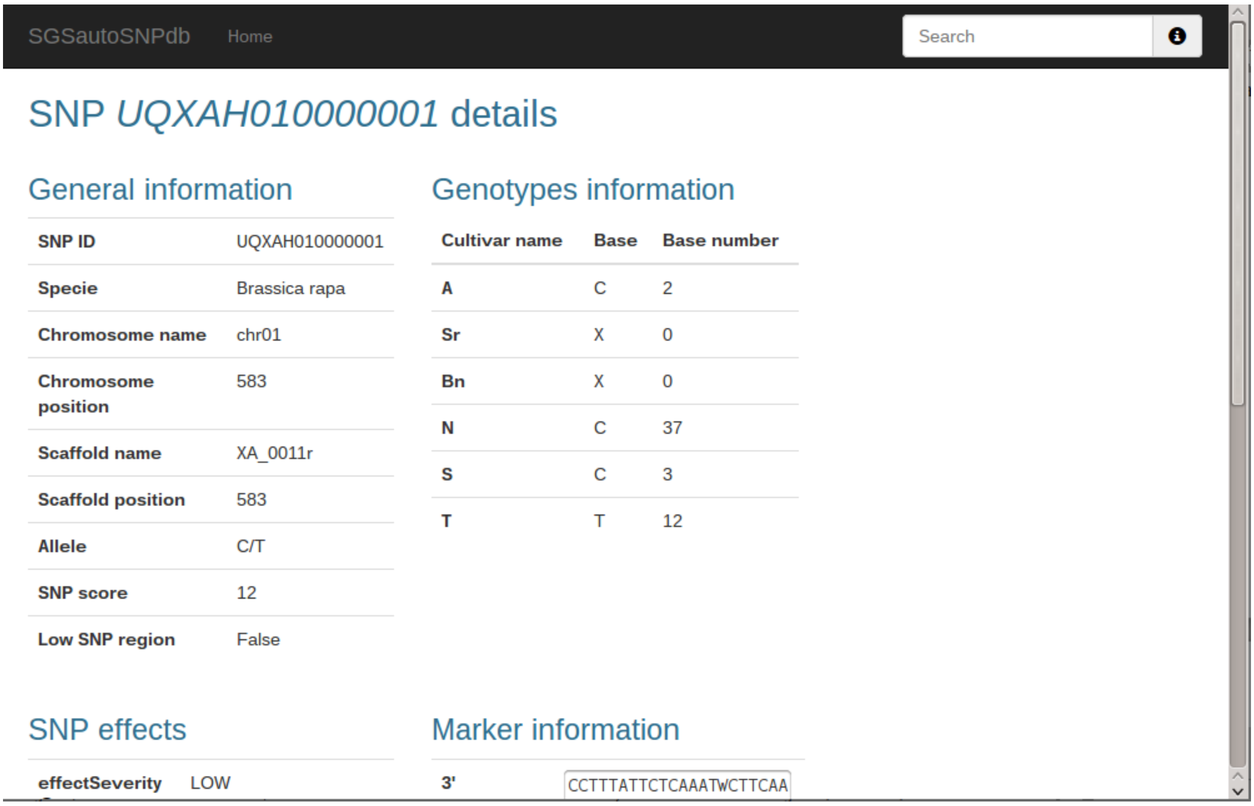
Figure 6.3 shows an example how to search between two cultivars and in range (step 1).



displaying **1 - 20** snps in total **21**

SNP id	Chr. pos.	SNP effect
UQXAH010000001	583	LOW
UQXAH010000002	1379	LOW
UQXAH010000003	1404	LOW
UQXAH010000004	5442	LOW
UQXAH010000005	5512	LOW
UQXAH010000006	5976	LOW
UQXAH010000007	5992	LOW
UQXAH010000008	6111	LOW
UQXAH010000009	6239	LOW
UQXAH010000010	6375	LOW
UQXAH010000011	6793	LOW
UQXAH010000012	6994	MED

Figure 6.4 shows all SNPs which satisfied the search criteria from Figure 6.3 (step 2).



SNP *UQXAH010000001* details

General information

SNP ID	UQXAH010000001
Specie	Brassica rapa
Chromosome name	chr01
Chromosome position	583
Scaffold name	XA_0011r
Scaffold position	583
Allele	C/T
SNP score	12
Low SNP region	False

Genotypes information

Cultivar name	Base	Base number
A	C	2
Sr	X	0
Bn	X	0
N	C	37
S	C	3
T	T	12

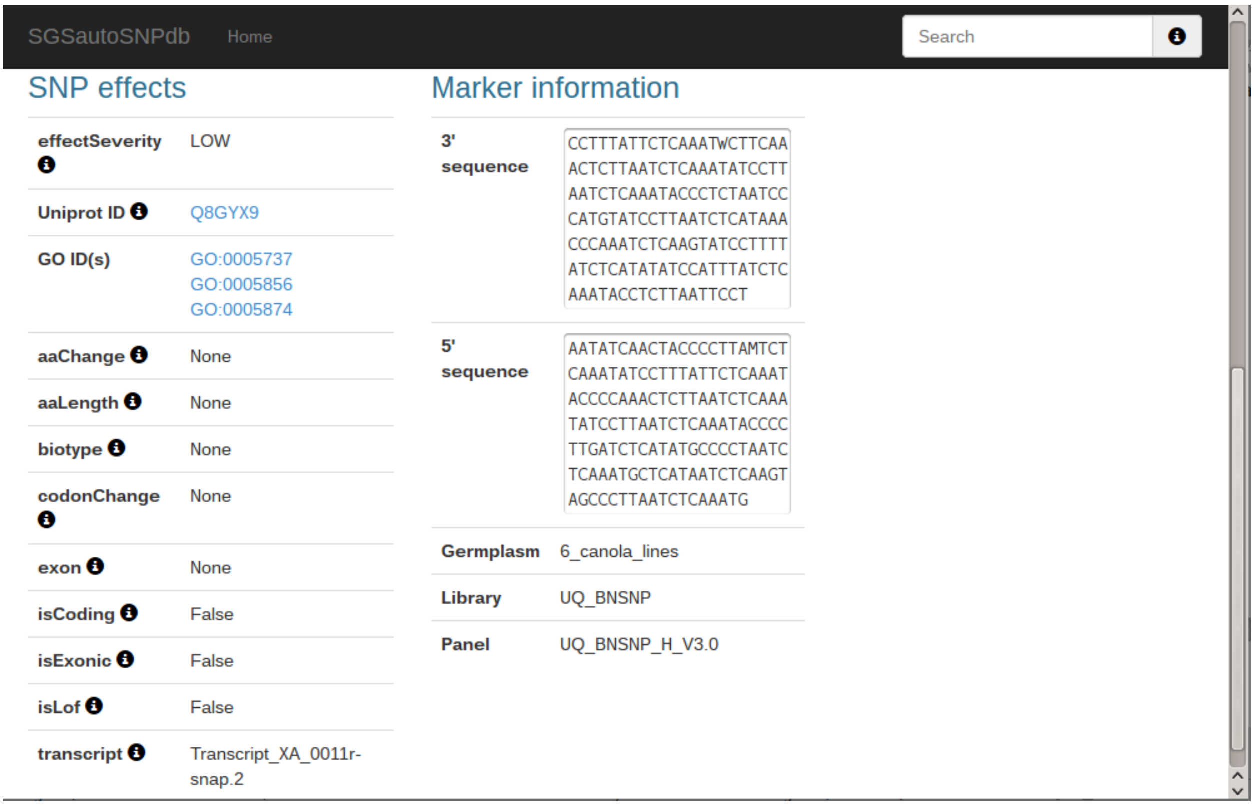
SNP effects

effectSeverity	LOW

Marker information

3'	CCTTTATTCTCAAATWCTTCAA

Figure 6.5: By clicking an SNP id in Figure 6.4 the user gets a detailed description of a SNP (step 3)



SNP effects

effectSeverity ℹ	LOW
Uniprot ID ℹ	Q8GYX9
GO ID(s)	GO:0005737 GO:0005856 GO:0005874
aaChange ℹ	None
aaLength ℹ	None
biotype ℹ	None
codonChange ℹ	None
exon ℹ	None
isCoding ℹ	False
isExonic ℹ	False
isLof ℹ	False
transcript ℹ	Transcript_XA_0011r-snap.2

Marker information

3' sequence	CCTTTATTCTCAAATWCTTCAA ACTCTTAATCTCAAATATCCTT AATCTCAAATACCCTCTAATCC CATGTATCCTTAATCTCATAAA CCCAAATCTCAAGTATCCTTTT ATCTCATATATCCATTTATCTC AAATACCTCTTAATTCCT
5' sequence	AATATCAACTACCCCTTAMTCT CAAATATCCTTTATTCTCAAAT ACCCCAAACTCTTAATCTCAAA TATCCTTAATCTCAAATACCCC TTGATCTCATATGCCCCTAATC TCAAATGCTCATAATCTCAAGT AGCCCTTAATCTCAAATG
Germplasm	6_canola_lines
Library	UQ_BNSNP
Panel	UQ_BNSNP_H_V3.0

Figure 6.6: This is a continuation from Figure 6.5.

References

ABOUELHODA, M. I., KURTZ, S. & OHLEBUSCH, E. 2004. Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms,* 2**,** 53-86.

ALLEN, A. M., BARKER, G. L., BERRY, S. T., COGHILL, J. A., GWILLIAM, R., KIRBY, S., ROBINSON, P., BRENCHLEY, R. C., D'AMORE, R., MCKENZIE, N., WAITE, D., HALL, A., BEVAN, M., HALL, N. & EDWARDS, K. J. 2011. Transcript- specific, single- nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnology Journal*, 9**,** 1086-1099.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

APWEILER, R., MARTIN, M. J., O'DONOVAN, C., MAGRANE, M., ALAM-FARUQUE, Y., ALPI, E., ANTUNES, R., ARGANISKA, J., CASANOVA, E. B., BELY, B., BINGLEY, M., BONILLA, C., BRITTO, R., BURSTEINAS, B., CHAN, W. M., CHAVALI, G., CIBRIAN-UHALTE, E., DA SILVA, A., DE GIORGI, M., DIMMER, E., FAZZINI, F., GANE, P., FEDOTOV, A., CASTRO, L. G., GARMIRI, P., HATTON-ELLIS, E., HIETA, R., HUNTLEY, R., JACOBSEN, J., JONES, R., LEGGE, D., LIU, W. D., LUO, J., MACDOUGALL, A., MUTOWO, P., NIGHTINGALE, A., ORCHARD, S., PATIENT, S., PICHLER, K., POGGIOLI, D., PUNDIR, S., PUREZA, L., QI, G. Y., ROSANOFF, S., SAWFORD, T., SEHRA, H., TURNER, E., VOLYNKIN, V., WARDELL, T., WATKINS, X., ZELLNER, H., CORBETT, M., DONNELLY, M., VAN RENSBURG, P., GOUJON, M., MCWILLIAM, H., LOPEZ, R., XENARIOS, I., BOUGUELERET, L., BRIDGE, A., POUX, S., REDASCHI, N., AUCHINCLOSS, A., AXELSEN, K., BANSAL, P., BARATIN, D., BINZ, P. A., BLATTER, M. C., BOECKMANN, B., BOLLEMAN, J., BOUTET, E., BREUZA, L., DE CASTRO, E., CERUTTI, L., COUDERT, E., CUCHE, B., DOCHE, M., DORNEVIL, D., DUVAUD, S., ESTREICHER, A., FAMIGLIETTI, L., FEUERMANN, M., GASTEIGER, E., GEHANT, S., GERRITSEN, V., GOS, A., GRUAZ-GUMOWSKI, N., HINZ, U., HULO, C., JAMES, J., JUNGO, F., KELLER, G., LARA, V., LEMERCIER, P., LEW, J., LIEBERHERR, D., MARTIN, X., MASSON, P., MORGAT, A., NETO, T., et al. 2013. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Research,* 41**,** D43-D47.

ARABIDOPSIS-GENOME-INITIATIVE 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature,* 408**,** 796-815.

ARUMUGANATHAN, K. & EARLE, E. D. 1991. Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter,* 9**,** 208-218.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-29.

AUSTRALIAN-BUREAU-OF-STATISTICS 2010. Australian farming in brief. Australia: Agriculture at a glance.

BARCHI, L., LANTERI, S., PORTIS, E., ACQUADRO, A., VALE, G., TOPPINO, L. & ROTINO, G. L. 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics,* 12**,** 304.

BARKER, G., BATLEY, J., O'SULLIVAN, H., EDWARDS, K. J. & EDWARDS, D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics,* 19**,** 421-422.

BARKER, G. L. & EDWARDS, K. J. 2009. A genome- wide analysis of single nucleotide polymorphism diversity in the world's major cereal crops. *Plant Biotechnology Journal,* 7**,** 318-325.

BARRELL, D., DIMMER, E., HUNTLEY, R. P., BINNS, D., O'DONOVAN, C. & APWEILER, R. 2009. The GOA database in 2009-an integrated gene ontology annotation resource. *Nucleic Acids Research,* 37**,** D396-D403.

BATLEY, J. & EDWARDS, D. 2009a. Genome sequence data: management, storage, and visualization. *Biotechniques,* 46**,** 333-336.

BATLEY, J. & EDWARDS, D. 2009b. Mining for single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular genetic markers. *In:* POSADA, D. (ed.) *Bioinformatics for DNA sequence analysis* New York, USA: Humana Press.

BATLEY, J., HOPKINS, C. J., COGAN, N. O. I., HAND, M., JEWELL, E., KAUR, J., KAUR, S., LI, X., LING, A. E., LOVE, C., MOUNTFORD, H., TODOROVIC, M., VARDY, M., WALKIEWICZ, M., SPANGENBERG, G. C. & EDWARDS, D. 2007. Identification and characterization of simple sequence repeat markers from *Brassica napus* expressed sequences. *Molecular Ecology Notes,* 7**,** 886-889.

BEGUN, D. J. & AQUADRO, C. F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature,* 356**,** 519 - 520.

BENNETT, M. D. & SMITH, J. B. 1976. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences,* 274**,** 227-274.

BENNETT, M. D. & SMITH, J. B. 1991. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 334**,** 309-345.

BERKMAN, P. J., LAI, K. T., LORENC, M. T. & EDWARDS, D. 2012a. Next-generation sequencing applications for wheat crop improvement. *American Journal of Botany,* 99**,** 365-371.

BERKMAN, P. J., SKARSHEWSKI, A., LORENC, M. T., LAI, K. T., DURAN, C., LING, E. Y. S., STILLER, J., SMITS, L., IMELFORT, M., MANOLI, S., MCKENZIE, M., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., FLEURY, D., DOLEZEL, J. & EDWARDS, D. 2011. Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnology Journal,* 9**,** 768-775.

BERKMAN, P. J., SKARSHEWSKI, A., MANOLI, S., LORENC, M. T., STILLER, J., SMITS, L., LAI, K. T., CAMPBELL, E., KUBALAKOVA, M., SIMKOVA, H., BATLEY, J., DOLEZEL, J., HERNANDEZ, P. & EDWARDS, D. 2012b. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical & Applied Genetics,* 124**,** 423-432.

BERKMAN, P. J., VISENDI, P., LEE, H. C., STILLER, J., MANOLI, S., LORENC, M. T., LAI, K. T., BATLEY, J., FLEURY, D., SIMKOVA, H., KUBALAKOVA, M., SONG, W. N., DOLEZEL, J. & EDWARDS, D. 2013. Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnology Journal,* 11**,** 564-571.

BEVAN, M. & WALSH, S. 2005. The Arabidopsis genome: a foundation for plant research. *Genome Research,* 15**,** 1632-1642.

BINNS, D., DIMMER, E., HUNTLEY, R., BARRELL, D., O'DONOVAN, C. & APWEILER, R. 2009. QuickGO: a web-based tool for gene ontology searching. *Bioinformatics,* 25**,** 3045-3046.

BLANKENBERG, D., VON KUSTER, G., BOUVIER, E., BAKER, D., AFGAN, E., STOLER, N., GALAXY, T., TAYLOR, J. & NEKRUTENKO, A. 2014. Dissemination of scientific software with Galaxy ToolShed. *Genome Biology,* 15**,** 403.

BOHUON, E. J. R., RAMSAY, L. D., CRAFT, J. A., ARTHUR, A. E., MARSHALL, D. F., LYDIATE, D. J. & KEARSEY, M. J. 1998. The association of flowering time quantitative trait loci with duplicated regions and candidate loci in *Brassica oleracea*. *Genetics,* 150**,** 393-401.

BOUTET, E., LIEBERHERR, D., TOGNOLLI, M., SCHNEIDER, M. & BAIROCH, A. 2007. UniProtKB/Swiss-Prot. *In:* EDWARDS, D. (ed.) *Plant bioinformatics.* Totowa, New Jersey, USA: Humana Press.

BOYLE, J. S. & LEW, A. M. 1995. An inexpensive alternative to glassmilk for DNA purification. *Trends in Genetics,* 11**,** 8.

BRENCHLEY, R., SPANNAGL, M., PFEIFER, M., BARKER, G. L. A., D'AMORE, R., ALLEN, A. M., MCKENZIE, N., KRAMER, M., KERHORNOU, A., BOLSER, D., KAY, S., WAITE, D., TRICK, M., BANCROFT, I., GU, Y., HUO, N., LUO, M. C., SEHGAL, S., GILL, B., KIANIAN, S., ANDERSON, O., KERSEY, P., DVORAK, J., MCCOMBIE, W. R., HALL, A., MAYER, K. F. X., EDWARDS, K. J., BEVAN, M. W. & HALL, N. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature,* 491**,** 705-710.

BUNDOCK, P. C., ELIOTT, F. G., ABLETT, G., BENSON, A. D., CASU, R. E., AITKEN, K. S. & HENRY, R. J. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal,* 7**,** 347-354.

BURROWS, M. & WHEELER, D. J. 1994. A block-sorting lossless data compression algorithm. *Technical Report.* Palo Alto, CA: Digital Equipment Corporation.

BUS, A., HECHT, J., HUETTEL, B., REINHARDT, R. & STICH, B. 2012. High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics,* 13**,** 281.

CALDWELL, K. S., DVORAK, J., LAGUDAH, E. S., AKHUNOV, E., LUO, M. C., WOLTERS, P. & POWELL, W. 2004. Sequence polymorphism in polyploid wheat and their d-genome diploid ancestor. *Genetics,* 167**,** 941-7.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

CAUSSE, M., DESPLAT, N., PASCUAL, L., LE PASLIER, M. C., SAUVAGE, C., BAUCHET, G., BERARD, A., BOUNON, R., TCHOUMAKOV, M., BRUNEL, D. & BOUCHET, J. P. 2013. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics,* 14**,** 791.

CAVANAGH, C. R., CHAO, S. M., WANG, S. C., HUANG, B. E., STEPHEN, S., KIANI, S., FORREST, K., SAINTENAC, C., BROWN-GUEDIRA, G. L., AKHUNOVA, A., SEE, D., BAI, G. H., PUMPHREY, M., TOMAR, L., WONG, D. B., KONG, S., REYNOLDS, M., DA SILVA, M. L., BOCKELMAN, H., TALBERT, L., ANDERSON, J. A., DREISIGACKER, S., BAENZIGER, S., CARTER, A., KORZUN, V., MORRELL, P. L., DUBCOVSKY, J., MORELL, M. K., SORRELLS, M. E., HAYDEN, M. J. & AKHUNOV, E. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 8057-8062.

CAVELL, A. C., LYDIATE, D. J., PARKIN, I. A. P., DEAN, C. & TRICK, M. 1998. Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. *Genome,* 41**,** 62-69.

CHANTRET, N., SALSE, J., SABOT, F., RAHMAN, S., BELLEC, A., LAUBIN, B., DUBOIS, I., DOSSAT, C., SOURDILLE, P., JOUDRIER, P., GAUTIER, M. F., CATTOLICO, L., BECKERT, M., AUBOURG, S., WEISSENBACH, J., CABOCHE, M., BERNARD, M., LEROY, P. & CHALHOUB, B. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *The Plant Cell Online,* 17**,** 1033-1045.

CHAO, S. M., ZHANG, W. J., AKHUNOV, E., SHERMAN, J., MA, Y. Q., LUO, M. C. & DUBCOVSKY, J. 2009. Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Molecular Breeding,* 23**,** 23-33.

CHARLESWORTH, B., COYNE, J. A. & BARTON, N. H. 1987. The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist,* 130**,** 113-146.

CHAUHAN, T. & KUMAR, R. 2010. Molecular markers and their applications in fisheries and aquaculture. *Advances in Bioscience & Biotechnology,* 1**,** 281-291.

CHEN, Q., QI, P. F., WEI, Y. M., WANG, J. R. & ZHENG, Y. L. 2009a. Molecular characterization of the pina gene in einkorn wheat. *Biochemical Genetics,* 47**,** 384-396.

CHEN, Y., SOUAIAIA, T. & CHEN, T. 2009b. PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics,* 25**,** 2514-21.

CHOI, S. R., TEAKLE, G. R., PLAHA, P., KIM, J. H., ALLENDER, C. J., BEYNON, E., PIAO, Z. Y., SOENGAS, P., HAN, T. H., KING, G. J., BARKER, G. C., HAND, P., LYDIATE, D. J., BATLEY, J., EDWARDS, D., KOO, D. H., BANG, J. W., PARK, B. S. & LIM, Y. P. 2007. The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. *Theoretical & Applied Genetics,* 115**,** 777-792.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly,* 6**,** 80-92.

CLEMENT, N. L., SNELL, Q., CLEMENT, M. J., HOLLENHORST, P. C., PURWAR, J., GRAVES, B. J., CAIRNS, B. R. & JOHNSON, W. E. 2010. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics,* 26**,** 38-45.

COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research,* 38**,** 1767-1771.

COCK, P. J. A., ANTAO, T., CHANG, J. T., CHAPMAN, B. A., COX, C. J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. & DE HOON, M. J. L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25**,** 1422-1423.

COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. & GILBERT, W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature,* 274**,** 775-780.

DABNEY, J., KNAPP, M., GLOCKE, I., GANSAUGE, M. T., WEIHMANN, A., NICKEL, B., VALDIOSERA, C., GARCIA, N., PAABO, S., ARSUAGA, J. L. & MEYER, M. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 15758-15763.

DALTON-MORGAN, J., HAYWARD, A., ALAMERY, S., TOLLENAERE, R., MASON, A. S., CAMPBELL, E., PATEL, D., LORENC, M. T., YI, B., LONG, Y., MENG, J., RAMAN, R., RAMAN, H., LAWLEY, C., EDWARDS, D. & BATLEY, J. 2014. Development of a high-throughput SNP array in the amphidiploid species *Brassica napus*. *Plant Biotechnology Journal*.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G. & DURBIN, R. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DELOURME, R., FALENTIN, C., FOMEJU, B. F., BOILLOT, M., LASSALLE, G., ANDRE, I., DUARTE, J., GAUTHIER, V., LUCANTE, N., MARTY, A., PAUCHON, M., PICHON, J. P., RIBIERE, N., TROTOUX, G., BLANCHARD, P., RIVIERE, N., MARTINANT, J. P. & PAUQUET, J. 2013. High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics,* 14**,** 120.

DIXON, G. R. 2007. Vegetable brassicas and related crucifers. *In:* ATHERTON, J. & REES, H. (eds.) *Crop production science in horticulture series.* Oxfordshire, UK: CAB International.

DONLIN, M. J. 2009. Using the generic genome browser (GBrowse). *Current Protocols in*

Bioinformatics. Saint Louis University School of Medicine, St. Louis, Missouri, USA: Wiley Online Library.

DOVERI, S., LEE, D., MAHESWARAN, M. & POWELL, W. 2008. Molecular markers-history, features and applications. *In:* KOLE, C. & ABBOTT, A. G. (eds.) *Principles and practices of plant genomics, Volume 1: genome mapping.* USA: CAB International.

DOYLE, J. J., FLAGEL, L. E., PATERSON, A. H., RAPP, R. A., SOLTIS, D. E., SOLTIS, P. S. & WENDEL, J. F. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics,* 42**,** 443-461.

DURAN, C., APPLEBY, N., CLARK, T., WOOD, D., IMELFORT, M., BATLEY, J. & EDWARDS, D. 2009a. AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Research,* 37**,** D951-D953.

DURAN, C., APPLEBY, N., VARDY, M., IMELFORT, M., EDWARDS, D. & BATLEY, J. 2009b. Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnology Journal,* 7**,** 326-333.

DURAN, C., BOSKOVIC, Z., IMELFORT, M., BATLEY, J., HAMILTON, N. A. & EDWARDS, D. 2010a. CMap3D: a 3D visualization tool for comparative genetic maps. *Bioinformatics,* 26**,** 273-274.

DURAN, C., EALES, D., MARSHALL, D., IMELFORT, M., STILLER, J., BERKMAN, P. J., CLARK, T., MCKENZIE, M., APPLEBY, N., BATLEY, J., BASFORD, K. & EDWARDS, D. 2010b. Future tools for association mapping in crop plants. *Genome,* 53**,** 1017-1023.

DURAN, C., EDWARDS, D. & BATLEY, J. 2009c. Molecular marker discovery and genetic map visualisation. *In:* EDWARDS, D., STAJICH, J. & HANSON, D. (eds.) *Bioinformatics.* New York: Springer.

DURSTEWITZ, G., POLLEY, A., PLIESKE, J., LUERSSEN, H., GRANER, E. M., WIESEKE, R. & GANAL, M. W. 2010. SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species Brassica napus. *Genome,* 53**,** 948-956.

DVORAK, J., AKHUNOV, E. D., AKHUNOV, A. R., DEAL, K. R. & LUO, M. C. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol Biol Evol,* 23**,** 1386-96.

EDWARDS, D., FORSTER, J. W., CHAGNE, D. & BATLEY, J. 2007a. What are SNPs? *In:* ORAGUZIE, N. C., RIKKERINK, E. H. A., GARDINER, S. E. & DE SILVA, H. N. (eds.) *Association mapping in plants.* New York, USA: Springer.

EDWARDS, D., FORSTER, J. W., COGAN, N. O. I., BATLEY, J. & CHAGNE, D. 2007b. Single nucleotide polymorphism discovery. *In:* ORAGUZIE, N. C., RIKKERINK, E. H. A., GARDINER, S. E. & DE SILVA, H. N. (eds.) *Association mapping in plants.* New York, USA: Springer.

EDWARDS, D. & GUPTA, P. K. 2013. Sequence based DNA markers and genotyping for cereal genomics and breeding. *In:* ., G. P. K. & VARSHNEY, R. K. (eds.) *Cereal genomics II.* New York, USA: Springer.

EDWARDS, D., WILCOX, S., BARRERO, R. A., FLEURY, D., CAVANAGH, C. R., FORREST, K. L., HAYDEN, M. J., MOOLHUIJZEN, P., KEEBLE-GAGNERE, G., BELLGARD, M. I., LORENC, M. T., SHANG, C. A., BAUMANN, U., TAYLOR, J. M., MORELL, M. K., LANGRIDGE, P., APPELS, R. & FITZGERALD, A. 2012. Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnology Journal,* 10**,** 703-708.

EID, J., FEHR, A., GRAY, J., LUONG, K., LYLE, J., OTTO, G., PELUSO, P., RANK, D., BAYBAYAN, P., BETTMAN, B., BIBILLO, A., BJORNSON, K., CHAUDHURI, B., CHRISTIANS, F., CICERO, R., CLARK, S., DALAL, R., DEWINTER, A., DIXON, J., FOQUET, M., GAERTNER, A., HARDENBOL, P., HEINER, C., HESTER, K., HOLDEN, D., KEARNS, G., KONG, X. X., KUSE, R., LACROIX, Y., LIN, S., LUNDQUIST, P., MA,

C. C., MARKS, P., MAXHAM, M., MURPHY, D., PARK, I., PHAM, T., PHILLIPS, M., ROY, J., SEBRA, R., SHEN, G., SORENSON, J., TOMANEY, A., TRAVERS, K., TRULSON, M., VIECELI, J., WEGENER, J., WU, D., YANG, A., ZACCARIN, D., ZHAO, P., ZHONG, F., KORLACH, J. & TURNER, S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323**,** 133-138.

EILBECK, K., LEWIS, S. E., MUNGALL, C. J., YANDELL, M., STEIN, L., DURBIN, R. & ASHBURNER, M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology,* 6**,** R44.

ERLICH, Y., CHANG, K., GORDON, A., RONEN, R., NAVON, O., ROOKS, M. & HANNON, G. J. 2009. DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research,* 19**,** 1243-1253.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-85.

FAO 2009. High level expert forum - how to feed the world in 2050. Rome, Italy: Agricultural development economics division.

FARRAR, M. 2007. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics,* 23**,** 156-61.

FERRAGINA, P. & MANZINI, G. Opportunistic data structures with applications. Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, 2000 2000. 390-398.

FLAVELL, R. B., RIMPAU, J. & SMITH, D. B. 1977. Repeated sequence DNA relationships in four cereal genomes. *Chromosoma,* 63**,** 205-222.

FLICEK, P. & BIRNEY, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods,* 6**,** S6-S12.

FRESNEDO-RAMIREZ, J., MARTINEZ-GARCIA, P. J., PARFITT, D. E., CRISOSTO, C. H. & GRADZIEL, T. M. 2013. Heterogeneity in the entire genome for three genotypes of peach [*Prunus persica* (L.) Batsch] as distinguished from sequence analysis of genomic variants. *BMC Genomics,* 14**,** 750.

FRITZ, M. H. Y., LEINONEN, R., COCHRANE, G. & BIRNEY, E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research,* 21**,** 734-740.

FROHLER, S. & DIETERICH, C. 2010. ACCUSA--accurate SNP calling on draft genomes. *Bioinformatics,* 26**,** 1364-1365.

FULTON, T. M., CHUNWONGSE, J. & TANKSLEY, S. D. 1995. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Molecular Biology Reporter,* 13**,** 207-209.

GARG, R. P. & SHARAPOV, I. 2001. Worker-Crew Model. *Techniques for Optimizing Applications: High Performance Computing.* 1 ed. Palo Alto, CA: Prentice Hall PTR.

GETINET, A., RAKOW, G., RANEY, J. P. & DOWNEY, R. K. 1997. Glucosinolate content in interspecific crosses of *Brassica carinata* with *B. juncea* and *B. napus*. *Plant Breeding,* 116**,** 39-46.

GILL, B. S., FRIEBE, B. & ENDO, T. R. 1991. Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum aestivum*). *Genome,* 34**,** 830-839.

GOLICZ, A. A., BAYER, P. E., MARTINEZ, P. A., LAI, K., LORENC, M. T., ALAMERY, S., HAYWARD, A., TOLLENAERE, R., BATLEY, J., EDWARDS, D., LONG, Y. & MENG, J. 2012. Characterising diversity in the *Brassica* genomes. *Acta Horticulturae,* 1**,** 33-48.

GRANT, J. R., ARANTES, A. S., LIAO, X. P. & STOTHARD, P. 2011. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics,* 27**,** 2300-2301.

GUPTA, P. K. 2008. Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology,* 26**,** 602-611.

GUPTA, P. K., RUSTGI, S. & MIR, R. R. 2013. Array-based high-throughput DNA markers and genotyping platforms for cereal genetics and genomics. *In:* ., G. P. K. & VARSHNEY, R. K. (eds.) *Cereal genomics II.* New York, USA: Springer.

HAMARSHEH, O. & AMRO, A. 2011. Characterization of simple sequence repeats (SSRs) from Phlebotomus papatasi (Diptera: Psychodidae) expressed sequence tags (ESTs). *Parasit Vectors,* 4**,** 189.

HARISMENDY, O., NG, P. C., STRAUSBERG, R. L., WANG, X., STOCKWELL, T. B., BEESON, K. Y., SCHORK, N. J., MURRAY, S. S., TOPOL, E. J., LEVY, S. & FRAZER, K. A. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology,* 10**,** R32.

HAYWARD, A., DALTON-MORGAN, J., MASON, A., ZANDER, M., EDWARDS, D. & BATLEY, J. 2012. Special Issue: Reviews; SNP discovery and applications in *Brassica napus*. *Journal of Plant Biotechnology (구 식물생명공학회지),* 39**,** 49-61.

HERNANDEZ, P., MARTIS, M., DORADO, G., PFEIFER, M., GALVEZ, S., SCHAAF, S., JOUVE, N., SIMKOVA, H., VALARIK, M., DOLEZEL, J. & MAYER, K. F. 2012. Next- generation sequencing and syntenic integration of flow- sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant Journal,* 69**,** 377-386.

HOMER, N., MERRIMAN, B. & NELSON, S. F. 2009. BFAST: an alignment tool for large scale genome resequencing. *PLoS One,* 4**,** e7767.

HOU, H. B., ZHAO, F. Q., ZHOU, L. L., ZHU, E. L., TENG, H. J., LI, X. K., BAO, Q. Y., WU, J. Y. & SUN, Z. S. 2010. MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Research,* 38**,** W732-W736.

HUANG, L., POPIC, V. & BATZOGLOU, S. 2013a. Short read alignment with populations of genomes. *Bioinformatics,* 29**,** i361-i370.

HUANG, S. M., DENG, L. B., GUAN, M., LI, J., LU, K., WANG, H. Z., FU, D. H., MASON, A. S., LIU, S. Y. & HUA, W. 2013b. Identification of genome-wide single nucleotide polymorphisms in allopolyploid crop *Brassica napus*. *BMC Genomics,* 14**,** 717.

HUNTER, A. A., MACGREGOR, A. B., SZABO, T. O., WELLINGTON, C. A. & BELLGARD, M. I. 2012. Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine,* 7**,** 1.

IMELFORT, M., DURAN, C., BATLEY, J. & EDWARDS, D. 2009. Discovering genetic polymorphisms in next- generation sequencing data. *Plant Biotechnology Journal,* 7**,** 312-317.

INABA, R. & NISHIO, T. 2002. Phylogenetic analysis of Brassiceae based on the nucleotide sequences of the S-locus related gene, SLR1. *Theoretical & Applied Genetics,* 105**,** 1159-1165.

INTERNATIONAL WHEAT GENOME SEQUENCING, C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science,* 345**,** 1251788.

JIA, J. Z., ZHAO, S. C., KONG, X. Y., LI, Y. R., ZHAO, G. Y., HE, W. M., APPELS, R., PFEIFER, M., TAO, Y., ZHANG, X. Y., JING, R. L., ZHANG, C., MA, Y. Z., GAO, L. F., GAO, C., SPANNAGL, M., MAYER, K. F. X., LI, D., PAN, S. K., ZHENG, F. Y., HU, Q., XIA, X. C., LI, J. W., LIANG, Q. S., CHEN, J., WICKER, T., GOU, C. Y., KUANG, H. H., HE, G. Y., LUO, Y. D., KELLER, B., XIA, Q. J., LU, P., WANG, J. Y., ZOU, H. F., ZHANG, R. Z., XU, J. Y., GAO, J. L., MIDDLETON, C., QUAN, Z. W., LIU, G. M., WANG, J., YANG, H. M., LIU, X., HE, Z. H., MAO, L. & WANG, J. 2013. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature,* 496**,** 91-95.

JIANG, H. & WONG, W. H. 2008. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics,* 24**,** 2395-6.

JOHNSTON, J. S., PEPPER, A. E., HALL, A. E., CHEN, Z. J., HODNETT, G., DRABEK, J., LOPEZ, R. & PRICE, H. J. 2005. Evolution of genome size in *Brassicaceae*. *Annals of Botany,* 95**,** 229-235.

KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., COOPER, A., MARKOWITZ, S., DURAN, C., THIERER, T., ASHTON, B., MEINTJES, P. & DRUMMOND, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* 28**,** 1647-1649.

KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Research,* 12**,** 656-664.

KOO, D. H., HONG, C. P., BATLEY, J., CHUNG, Y. S., EDWARDS, D., BANG, J. W., HUR, Y. & LIM, Y. P. 2011. Rapid divergence of repetitive DNAs in *Brassica* relatives. *Genomics,* 97**,** 173-185.

KORF, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics,* 5**,** 59.

KORZUN, V. 2002. Use of molecular markers in cereal breeding. *Cellular & Molecular Biology Letters,* 7**,** 811-820.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol,* 5**,** R12.

LAI, K., BERKMAN, P. J., LORENC, M. T., DURAN, C., SMITS, L., MANOLI, S., STILLER, J. & EDWARDS, D. 2012a. WheatGenome.info: an integrated database and portal for wheat genome information. *Plant & Cell Physiology,* 53**,** e2.

LAI, K., DURAN, C., BERKMAN, P. J., LORENC, M. T., STILLER, J., MANOLI, S., HAYDEN, M. J., FORREST, K. L., FLEURY, D., BAUMANN, U., ZANDER, M., MASON, A. S., BATLEY, J. & EDWARDS, D. 2012b. Single nucleotide polymorphism discovery from wheat next- generation sequence data. *Plant Biotechnology Journal,* 10**,** 743-749.

LAI, K., LORENC, M. T., LEE, H. C., BERKMAN, P. J., BAYER, P. E., VISENDI, P., RUPERAO, P., FITZGERALD, T. L., ZANDER, M., CHAN, C.-K. K., MANOLI, S., STILLER, J., BATLEY, J. & EDWARDS, D. 2015. Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnology Journal,* 13**,** 97-104.

LAM, H. M., XU, X., LIU, X., CHEN, W. B., YANG, G. H., WONG, F. L., LI, M. W., HE, W. M., QIN, N., WANG, B., LI, J., JIAN, M., WANG, J. A., SHAO, G. H., WANG, J., SUN, S. S. M. & ZHANG, G. Y. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics,* 42**,** 1053-1059.

LAN, T. H., DELMONTE, T. A., REISCHMANN, K. P., HYMAN, J., KOWALSKI, S. P., MCFERSON, J., KRESOVICH, S. & PATERSON, A. H. 2000. An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. *Genome Research,* 10**,** 776-788.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

LEE, H. C., LAI, K. T., LORENC, M. T., IMELFORT, M., DURAN, C. & EDWARDS, D. 2012. Bioinformatics tools and databases for analysis of next-generation sequence data. *Briefings in Functional Genomics,* 11**,** 12-24.

LEITCH, A. R. & LEITCH, I. J. 2008. Perspective - Genomic plasticity and the diversity of polyploid plants. *Science,* 320**,** 481-483.

LI, H. 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-2993.

LI, H. 2011b. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics,* 27**,** 718-719.

LI, H. & DURBIN, R. 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-1760.

LI, H. & DURBIN, R. 2009b. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LI, H. & HOMER, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform,* 11**,** 473-83.

LI, H., RUAN, J. & DURBIN, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research,* 18**,** 1851-1858.

LI, R. Q., YU, C., LI, Y. R., LAM, T. W., YIU, S. M., KRISTIANSEN, K. & WANG, J. 2009b. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics,* 25**,** 1966-1967.

LIN, H., ZHANG, Z., ZHANG, M. Q., MA, B. & LI, M. 2008. ZOOM! Zillions of oligos mapped. *Bioinformatics,* 24**,** 2431-7.

LINDNER, R. & FRIEDEL, C. C. 2012. A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS One,* 7**,** e52403.

LING, H. Q., ZHAO, S. C., LIU, D. C., WANG, J. Y., SUN, H., ZHANG, C., FAN, H. J., LI, D., DONG, L. L., TAO, Y., GAO, C., WU, H. L., LI, Y. W., CUI, Y., GUO, X. S., ZHENG, S. S., WANG, B., YU, K., LIANG, Q. S., YANG, W. L., LOU, X. Y., CHEN, J., FENG, M. J., JIAN, J. B., ZHANG, X. F., LUO, G. B., JIANG, Y., LIU, J. J., WANG, Z. B., SHA, Y. H., ZHANG, B. R., WU, H. J., TANG, D. Z., SHEN, Q. H., XUE, P. Y., ZOU, S. H., WANG, X. J., LIU, X., WANG, F. M., YANG, Y. P., AN, X. L., DONG, Z. Y., ZHANG, K. P., ZHANG, X. Q., LUO, M. C., DVORAK, J., TONG, Y. P., WANG, J., YANG, H. M., LI, Z. S., WANG, D. W., ZHANG, A. M. & WANG, J. 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature,* 496**,** 87-90.

LIU, S., LIU, Y., YANG, X., TONG, C., EDWARDS, D., PARKIN, I. A., ZHAO, M., MA, J., YU, J., HUANG, S., WANG, X., WANG, J., LU, K., FANG, Z., BANCROFT, I., YANG, T. J., HU, Q., WANG, X., YUE, Z., LI, H., YANG, L., WU, J., ZHOU, Q., WANG, W., KING, G. J., PIRES, J. C., LU, C., WU, Z., SAMPATH, P., WANG, Z., GUO, H., PAN, S., YANG, L., MIN, J., ZHANG, D., JIN, D., LI, W., BELCRAM, H., TU, J., GUAN, M., QI, C., DU, D., LI, J., JIANG, L., BATLEY, J., SHARPE, A. G., PARK, B. S., RUPERAO, P., CHENG, F., WAMINAL, N. E., HUANG, Y., DONG, C., WANG, L., LI, J., HU, Z., ZHUANG, M., HUANG, Y., HUANG, J., SHI, J., MEI, D., LIU, J., LEE, T. H., WANG, J., JIN, H., LI, Z., LI, X., ZHANG, J., XIAO, L., ZHOU, Y., LIU, Z., LIU, X., QIN, R., TANG, X., LIU, W., WANG, Y., ZHANG, Y., LEE, J., KIM, H. H., DENOEUD, F., XU, X., LIANG, X., HUA, W., WANG, X., WANG, J., CHALHOUB, B. & PATERSON, A. H. 2014. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications,* 5**,** 3930.

LORENC, M. T., HAYASHI, S., STILLER, J., LEE, H., MANOLI, S., RUPERAO, P., VISENDI, P., BERKMAN, P. J., LAI, K., BATLEY, J. & EDWARDS, D. 2012. Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology,* 1**,** 370-382.

LYSAK, M. A., KOCH, M. A., PECINKA, A. & SCHUBERT, I. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Research,* 15**,** 516-525.

MAGLOTT, D., OSTELL, J., PRUITT, K. D. & TATUSOVA, T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research,* 39**,** D52-D57.

MAJOROS, W. H., PERTEA, M. & SALZBERG, S. L. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics,* 20**,** 2878-2879.

MANYAM, G., PAYTON, M. A., ROTH, J. A., ABRUZZO, L. V. & COOMBES, K. R. 2012. Relax with CouchDB--Into the non-relational DBMS era of bioinformatics. *Genomics,* 100**,** 1-7.

MARDIS, E. R. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics & Human Genetics,* 9**,** 387-402.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y. J., CHEN, Z. T., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L. I., JARVIE, T. P., JIRAGE, K. B., KIM, J. B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P. G., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

MENDHAM, N. J. & SALISBURY, P. A. 1995. Physiology - crop development, growth and yield. *In:* KIMBER, D. S. & MCGREGOR, D. I. (eds.) *Brassica oilseeds: production and utilization.* Wallingford, UK: CAB International.

METZKER, M. L. 2010. Sequencing technologies—the next generation. *Nature Reviews Genetics,* 11**,** 31-46.

MEYERS, L. A. & LEVIN, D. A. 2006. On the abundance of polyploids in flowering plants. *Evolution,* 60**,** 1198-1206.

MILNE, I., BAYER, M., CARDLE, L., SHAW, P., STEPHEN, G., WRIGHT, F. & MARSHALL, D. 2010a. Tablet-next generation sequence assembly visualization. *Bioinformatics,* 26**,** 401-402.

MILNE, I., SHAW, P., STEPHEN, G., BAYER, M., CARDLE, L., THOMAS, W. T., FLAVELL, A. J. & MARSHALL, D. 2010b. Flapjack--graphical genotype visualization. *Bioinformatics,* 26**,** 3133-3134.

MILNE, I., STEPHEN, G., BAYER, M., COCK, P. J. A., PRITCHARD, L., CARDLE, L., SHAW, P. D. & MARSHALL, D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics,* 14**,** 193-202.

MISRA, S., AGRAWAL, A., LIAO, W. K. & CHOUDHARY, A. 2011. Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. *Bioinformatics,* 27**,** 189-95.

MOORE, G. E. 1998. Cramming more components onto integrated circuits. *Proceedings of the IEEE,* 86**,** 82-85.

MUELLER, U. G. & WOLFENBARGER, L. L. 1999. AFLP genotyping and fingerprinting. *Trends in Ecology & Evolution,* 14**,** 389-394.

MUN, J. H., KWON, S. J., SEOL, Y. J., KIM, J. A., JIN, M., KIM, J. S., LIM, M. H., LEE, S. I., HONG, J. K., PARK, T. H., LEE, S. C., KIM, B. J., SEO, M. S., BAEK, S., LEE, M. J., SHIN, J. Y., HAHN, J. H., HWANG, Y. J., LIM, K. B., PARK, J. Y., LEE, J., YANG, T. J., YU, H. J., CHOI, I. Y., CHOI, B. S., CHOI, S. R., RAMCHIARY, N., LIM, Y. P., FRASER, F., DROU, N., SOUMPOUROU, E., TRICK, M., BANCROFT, I., SHARPE, A. G., PARKIN, I. A., BATLEY, J., EDWARDS, D. & PARK, B. S. 2010. Sequence and structure of *Brassica rapa* chromosome A3. *Genome Biology,* 11**,** R94.

NIELSEN, R., PAUL, J. S., ALBRECHTSEN, A. & SONG, Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet,* 12**,** 443-51.

NING, Z., COX, A. J. & MULLIKIN, J. C. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res,* 11**,** 1725-9.

PAILA, U., CHAPMAN, B. A., KIRCHNER, R. & QUINLAN, A. R. 2013. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology,* 9**,** e1003153.

PARITOSH, K., YADAVA, S. K., GUPTA, V., PANJABI-MASSAND, P., SODHI, Y. S., PRADHAN, A. K. & PENTAL, D. 2013. RNA-seq based SNPs in some agronomically important oleiferous lines of *Brassica rapa* and their use for genome-wide linkage mapping and specific-region fine mapping. *BMC Genomics,* 14**,** 463.

PARK, S., YU, H. J., MUN, J. H. & LEE, S. C. 2010. Genome-wide discovery of DNA polymorphism in *Brassica rapa*. *Molecular Genetics & Genomics,* 283**,** 135-145.

PARKIN, I. A., KOH, C., TANG, H., ROBINSON, S. J., KAGALE, S., CLARKE, W. E., TOWN, C. D., NIXON, J., KRISHNAKUMAR, V., BIDWELL, S. L., DENOEUD, F., BELCRAM, H., LINKS, M. G., JUST, J., CLARKE, C., BENDER, T., HUEBERT, T., MASON, A. S., PIRES, C. J., BARKER, G., MOORE, J., WALLEY, P. G., MANOLI, S., BATLEY, J., EDWARDS, D., NELSON, M. N., WANG, X., PATERSON, A. H., KING, G., BANCROFT, I., CHALHOUB, B. & SHARPE, A. G. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology,* 15**,** R77.

PARKIN, I. A., SHARPE, A. G., KEITH, D. J. & LYDIATE, D. J. 1995. Identification of the A and C genomes of amphidiploid *Brassica napus* (oilseed rape). *Genome,* 38**,** 1122-1131.

PAUX, E., FAURE, S., CHOULET, F., ROGER, D., GAUTHIER, V., MARTINANT, J. P., SOURDILLE, P., BALFOURIER, F., LE PASLIER, M. C., CHAUVEAU, A., CAKIR, M., GANDON, B. & FEUILLET, C. 2010. Insertion site- based polymorphism markers open new perspectives for genome saturation and marker- assisted selection in wheat. *Plant Biotechnology Journal,* 8**,** 196-210.

PAVLOPOULOS, G. A., OULAS, A., IACUCCI, E., SIFRIM, A., MOREAU, Y., SCHNEIDER, R., AERTS, J. & ILIOPOULOS, I. 2013. Unraveling genomic variation from next generation sequencing data. *BioData Mining,* 6**,** 13.

PEARSON, W. R. & LIPMAN, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A,* 85**,** 2444-8.

PEDERSEN, C. & LANGRIDGE, P. 1997. Identification of the entire chromosome complement of bread wheat by two-colour FISH. *Genome,* 40**,** 589-593.

PIKE, R., DORWARD, S., GRIESEMER, R. & QUINLAN, S. 2005. Interpreting the data: Parallel analysis with Sawzall. *Scientific Programming,* 13**,** 277-298.

PIREDDU, L., LEO, S. & ZANETTI, G. 2011. SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics,* 27**,** 2159-2160.

POLICRITI, A. & PREZZA, N. 2014. Hashing and Indexing: Succinct DataStructures and Smoothed Analysis. *In:* AHN, H.-K. & SHIN, C.-S. (eds.) *Algorithms and Computation.* Springer International Publishing.

POWELL, W., MORGANTE, M., ANDRE, C., MCNICOL, J. W., MACHRAY, G. C., DOYLE, J. J., TINGEY, S. V. & RAFALSKI, J. A. 1995. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Current Biology,* 5**,** 1023-1029.

RAFALSKI, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology,* 5**,** 94-100.

RAKOW, G. 2004. Species origin and economic importance of Brassica. *In:* PUA, E. C. & DOUGLAS, C. J. (eds.) *Biotechnology in agriculture and forestry.* 54 ed.: Springer.

RANC, N., MUNOS, S., XU, J. X., LE PASLIER, M. C., CHAUVEAU, A., BOUNON, R., ROLLAND, S., BOUCHET, J. P., BRUNEL, D. & CAUSSE, M. 2012. Genome-wide association mapping in tomato (Solanum lycopersicum) is possible using genome admixture of *Solanum lycopersicum* var. cerasiforme. *G3: Genes| Genomes| Genetics,* 2**,** 853-864.

RANDHAWA, H. S., ASIF, M., POZNIAK, C., CLARKE, J. M., GRAF, R. J., FOX, S. L., HUMPHREYS, D. G., KNOX, R. E., DEPAUW, R. M., SINGH, A. K., CUTHBERT, R. D., HUCL, P. & SPANER, D. 2013. Application of molecular markers to wheat breeding in Canada. *Plant Breeding,* 132**,** 458-471.

RASCOVSKY, S. J., DELGADO, J. A., SANZ, A., CALVO, V. D. & CASTRILLON, G. 2012. Informatics in Radiology: Use of CouchDB for Document-based Storage of DICOM Objects. *Radiographics,* 32**,** 913-927.

RAVEL, C., PRAUD, S., MURIGNEUX, A., CANAGUIER, A., SAPET, F., SAMSON, D., BALFOURIER, F., DUFOUR, P., CHALHOUB, B., BRUNEL, D., BECKERT, M. & CHARMET, G. 2006. Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). *Genome,* 49**,** 1131-1139.

REDMOND, E. & WILSON, J. 2012. CouchDB. *In:* CARTER, J. (ed.) *Seven databases in seven weeks.* Texas, USA: O'Reilly.

REESE, M. G., MOORE, B., BATCHELOR, C., SALAS, F., CUNNINGHAM, F., MARTH, G. T., STEIN, L., FLICEK, P., YANDELL, M. & EILBECK, K. 2010. A standard variation file format for human genome sequences. *Genome Biology,* 11**,** R88.

ROBISON, K. 2010. Editorial: Second-generation sequencing. *Briefings in Bioinformatics,* 11**,** 455-456.

RUFFALO, M., LAFRAMBOISE, T. & KOYUTURK, M. 2011. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics,* 27**,** 2790-6.

RUMBLE, S. M., LACROUTE, P., DALCA, A. V., FIUME, M., SIDOW, A. & BRUDNO, M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol,* 5**,** e1000386.

SAMBROOK, J. & RUSSELL, D. W. 2001. *Molecular cloning: a laboratory manual,* New York, USA, Cold Spring Harbor Laboratory Press.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SATO, S., TABATA, S., HIRAKAWA, H., ASAMIZU, E., SHIRASAWA, K., ISOBE, S., KANEKO, T., NAKAMURA, Y., SHIBATA, D., AOKI, K., EGHOLM, M., KNIGHT, J., BOGDEN, R., LI, C. B., SHUANG, Y., XU, X., PAN, S. K., CHENG, S. F., LIU, X., REN, Y. Y., WANG, J., ALBIERO, A., DAL PERO, F., TODESCO, S., VAN ECK, J., BUELS, R. M., BOMBARELY, A., GOSSELIN, J. R., HUANG, M. Y., LETO, J. A., MENDA, N., STRICKLER, S., MAO, L. Y., GAO, S., TECLE, I. Y., YORK, T., ZHENG, Y., VREBALOV, J. T., LEE, J., ZHONG, S. L., MUELLER, L. A., STIEKEMA, W. J., RIBECA, P., ALIOTO, T., YANG, W. C., HUANG, S. W., DU, Y. C., ZHANG, Z. H., GAO, J. C., GUO, Y. M., WANG, X. X., LI, Y., HE, J., LI, C. Y., CHENG, Z. K., ZUO, J. R., REN, J. F., ZHAO, J. H., YAN, L. H., JIANG, H. L., WANG, B., LI, H. S., LI, Z. J., FU, F. Y., CHEN, B. T., HAN, B., FENG, Q., FAN, D. L., WANG, Y., LING, H. Q., XUE, Y. B. A., WARE, D., MCCOMBIE, W. R., LIPPMAN, Z. B., CHIA, J. M., JIANG, K., PASTERNAK, S., GELLEY, L., KRAMER, M., ANDERSON, L. K., CHANG, S. B., ROYER, S. M., SHEARER, L. A., STACK, S. M., ROSE, J. K. C., XU, Y. M., EANNETTA, N., MATAS, A. J., MCQUINN, R., TANKSLEY, S. D., CAMARA, F., GUIGO, R., ROMBAUTS, S., FAWCETT, J., VAN DE PEER, Y., ZAMIR, D., LIANG, C. B., SPANNAGL, M., GUNDLACH, H., BRUGGMANN, R., et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature,* 485**,** 635-641.

SAVAGE, D., BATLEY, J., ERWIN, T., LOGAN, E., LOVE, C. G., LIM, G. A. C., MONGIN, E., BARKER, G., SPANGENBERG, G. C. & EDWARDS, D. 2005. SNPServer: a real-time SNP discovery tool. *Nucleic Acids Research,* 33**,** W493-W495.

SCHATZ, M. C. 2009. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics,* 25**,** 1363-9.

SCHUBERT, M., ERMINI, L., SARKISSIAN, C. D., JONSSON, H., GINOLHAC, A., SCHAEFER, R., MARTIN, M. D., FERNANDEZ, R., KIRCHER, M., MCCUE, M., WILLERSLEV, E. & ORLANDO, L. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols,* 9**,** 1056-1082.

SERGEEVA, E. M., AFONNIKOV, D. A., KOLTUNOVA, M. K., GUSEV, V. D., MIROSHNICHENKO, L. A., VRÁNA, J., KUBALAKOVA, M., PONCET, C., SOURDILLE, P., FEUILLET, C., DOLEZEL, J. & SALINA, E. A. 2014. Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. *The Plant Genome,* 7.

SHANG, J., ZHU, F., VONGSANGNAK, W., TANG, Y., ZHANG, W. & SHEN, B. 2014. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int,* 2014**,** 309650.

SHE, X. W., ROHL, C. A., CASTLE, J. C., KULKARNI, A. V., JOHNSON, J. M. & CHEN, R. H. 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics,* 10**,** 269.

SHEN, Y. F., WAN, Z. Z., COARFA, C., DRABEK, R., CHEN, L., OSTROWSKI, E. A., LIU, Y., WEINSTOCK, G. M., WHEELER, D. A., GIBBS, R. A. & YU, F. L. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research,* 20**,** 273-280.

SILBERMANN, J., WERNICKE, C., POSPISIL, H. & FROHME, M. 2013. RefPrimeCouch--a reference gene primer CouchApp. *Database,* 2013**,** bat081.

SMITH, A. D., CHUNG, W. Y., HODGES, E., KENDALL, J., HANNON, G., HICKS, J., XUAN, Z. & ZHANG, M. Q. 2009. Updates to the RMAP short-read mapping software. *Bioinformatics,* 25**,** 2841-2.

SOLTIS, D. E. & SOLTIS, P. S. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in Ecology & Evolution,* 14**,** 348-352.

SOLTIS, D. E., SOLTIS, P. S. & TATE, J. A. 2004. Advances in the study of polyploidy since plant speciation. *New Phytologist,* 161**,** 173-191.

SOMERS, D. J., KIRKPATRICK, R., MONIWA, M. & WALSH, A. 2003. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome,* 46**,** 431-437.

STOLETZKI, N. & EYRE-WALKER, A. 2011. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. *Molecular Biology & Evolution,* 28**,** 1371-1380.

SUBBAIYAN, G. K., WATERS, D. L., KATIYAR, S. K., SADANANDA, A. R., VADDADI, S. &

HENRY, R. J. 2012. Genome- wide DNA polymorphisms in elite indica rice inbreds discovered by whole- genome sequencing. *Plant Biotechnology Journal,* 10**,** 623-634.

SYVANEN, A. C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics,* 2**,** 930-942.

TALBERT, L. E., SMITH, L. Y. & BLAKE, N. K. 1998. More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. *Genome,* 41**,** 402-407.

TREBBI, D., MACCAFERRI, M., DE HEER, P., SORENSEN, A., GIULIANI, S., SALVI, S., SANGUINETI, M. C., MASSI, A., VAN DER VOSSEN, E. A. G. & TUBEROSA, R. 2011. High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). *Theoretical & Applied Genetics,* 123**,** 555-569.

TRICK, M., LONG, Y., MENG, J. L. & BANCROFT, I. 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnology Journal,* 7**,** 334-346.

TURNPENNY, P. D. & ELLARD, S. 2011. *Emery's elements of medical genetics,* Philadelphia, PA, USA, Churchill Livingstone.

U, N. 1935. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany,* 7**,** 389–452.

VALOUEV, A., ICHIKAWA, J., TONTHAT, T., STUART, J., RANADE, S., PECKHAM, H., ZENG, K., MALEK, J. A., COSTA, G., MCKERNAN, K., SIDOW, A., FIRE, A. & JOHNSON, S. M. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research,* 18**,** 1051-1063.

VARELA, M. A. & AMOS, W. 2010. Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. *Genomics,* 95**,** 151-159.

VARSHNEY, R. K., NAYAK, S. N., MAY, G. D. & JACKSON, S. A. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology,* 27**,** 522-530.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. Q. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J. H., MIKLOS, G. L. G., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, C., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z. M., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W. M., GONG, F. C., GU, Z. P., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z. X., KETCHUM, K. A., LAI, Z. W., LEI, Y. D., LI, Z. Y., LI, J. Y., LIANG, Y., LIN, X. Y., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B. X., SUN, J. T., WANG, Z. Y., WANG, A. H., WANG, X., WANG, J., WEI, M. H., WIDES, R., XIAO, C. L., YAN, C. H., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-1351.

VITTE, C. & BENNETZEN, J. L. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 17638-17643.

VITULO, N., ALBIERO, A., FORCATO, C., CAMPAGNA, D., DAL PERO, F., BAGNARESI, P., COLAIACOVO, M., FACCIOLI, P., LAMONTANARA, A., SIMKOVA, H., KUBALAKOVA, M., PERROTTA, G., FACELLA, P., LOPEZ, L., PIETRELLA, M., GIANESE, G., DOLEZEL, J., GIULIANO, G., CATTIVELLI, L., VALLE, G. & STANCA, A. M. 2011. First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PloS One,* 6**,** e26421.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., VANDELEE, T., HORNES, M., FRIJTERS, A., POT, J., PELEMAN, J., KUIPER, M. & ZABEAU, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research,* 23**,** 4407-4414.

VRANA, J., KUBALAKOVA, M., SIMKOVA, H., CIHALIKOVA, J., LYSAK, M. A. & DOLEZEL, J. 2000. Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics,* 156**,** 2033-2041.

WAIKAN, Y. & DOZY, A. M. 1978. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proceedings of the National Academy of Sciences of the United States of America,* 75**,** 5631-5635.

WANG, X. W., WANG, H. Z., WANG, J., SUN, R. F., WU, J., LIU, S. Y., BAI, Y. Q., MUN, J. H., BANCROFT, I., CHENG, F., HUANG, S. W., LI, X. X., HUA, W., WANG, J. Y., WANG, X. Y., FREELING, M., PIRES, J. C., PATERSON, A. H., CHALHOUB, B., WANG, B., HAYWARD, A., SHARPE, A. G., PARK, B. S., WEISSHAAR, B., LIU, B. H., LI, B., LIU, B., TONG, C. B., SONG, C., DURAN, C., PENG, C. F., GENG, C. Y., KOH, C. S., LIN, C. Y., EDWARDS, D., MU, D. S., SHEN, D., SOUMPOUROU, E., LI, F., FRASER, F., CONANT, G., LASSALLE, G., KING, G. J., BONNEMA, G., TANG, H. B., WANG, H. P., BELCRAM, H., ZHOU, H. L., HIRAKAWA, H., ABE, H., GUO, H., WANG, H., JIN, H. Z., PARKIN, I. A. P., BATLEY, J., KIM, J. S., JUST, J., LI, J. W., XU, J. H., DENG, J., KIM, J. A., LI, J. P., YU, J. Y., MENG, J. L., WANG, J. P., MIN, J. M., POULAIN, J., WANG, J., HATAKEYAMA, K., WU, K., WANG, L., FANG, L., TRICK, M., LINKS, M. G., ZHAO, M. X., JIN, M. N., RAMCHIARY, N., DROU, N., BERKMAN, P. J., CAI, Q. L., HUANG, Q. F., LI, R. Q., TABATA, S., CHENG, S. F., ZHANG, S., ZHANG, S. J., HUANG, S. M., SATO, S., SUN, S. L., KWON, S. J., CHOI, S. R., LEE, T. H., FAN, W., ZHAO, X., TAN, X., XU, X., WANG, Y., QIU, Y., YIN, Y., LI, Y. R., et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics,* 43**,** 1035-1039.

WANJUGI, H., COLEMAN-DERR, D., HUO, N., KIANIAN, S. F., LUO, M. C., WU, J., ANDERSON, O. & GU, Y. Q. 2009. Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome,* 52**,** 576-587.

WEERAKOON, S. R., SI, P., ZILI, W., MENG, J. & YAN, G. 2009. Production and confirmation of hybrids through interspecific crossing between tetraploid *B. juncea* and diploid *B. oleracea* towards a hexaploid *Brassica* population. *16th Australian aesearch assembly on Brassicas.* Ballarat, Victoria, Australia.

WEESE, D., EMDE, A. K., RAUSCH, T., DORING, A. & REINERT, K. 2009. RazerS--fast read mapping with sensitivity control. *Genome Res,* 19**,** 1646-54.

WHITTAKER, J. C., HARBORD, R. M., BOXALL, N., MACKAY, I., DAWSON, G. & SIBLY, R. M. 2003. Likelihood-based estimation of microsatellite mutation rates. *Genetics,* 164**,** 781-787.

WICKER, T., MAYER, K. F. X., GUNDLACH, H., MARTIS, M., STEUERNAGEL, B., SCHOLZ, U., SIMKOVA, H., KUBALAKOVA, M., CHOULET, F., TAUDIEN, S., PLATZER, M., FEUILLET, C., FAHIMA, T., BUDAK, H., DOLEZEL, J., KELLER, B. & STEIN, N. 2011. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *The Plant Cell Online,* 23**,** 1706-1718.

YANG, Y. W., LAI, K. N., TAI, P. Y. & LI, W. H. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *Journal of Molecular Evolution,* 48**,** 597-604.

YOU, F. M., HUO, N. X., DEAL, K. R., GU, Y. Q., LUO, M. C., MCGUIRE, P. E., DVORAK, J. & ANDERSON, O. D. 2011. Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics,* 12**,** 59.

YOU, F. M., WANJUGI, H., HUO, N. X., LAZO, G. R., LUO, M. C., ANDERSON, O. D., DVORAK, J. & GU, Y. Q. 2010. RJPrimers: unique transposable element insertion junction discovery and PCR primer design for marker development. *Nucleic Acids Research,* 38**,** W313-W320.

YOUENS-CLARK, K., FAGA, B., YAP, I. V., STEIN, L. & WARE, D. 2009. CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics,* 25**,** 3040-3042.

YU, X., GUDA, K., WILLIS, J., VEIGL, M., WANG, Z., MARKOWITZ, S., ADAMS, M. D. & SUN, S. 2012. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining,* 5**,** 6-6.

ZANDER, M., PATEL, D. A., VAN DE WOUW, A., LAI, K. T., LORENC, M. T., CAMPBELL, E., HAYWARD, A., EDWARDS, D., RAMAN, H. & BATLEY, J. 2013. Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional & Integrative Genomics,* 13**,** 295-308.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research,* 18**,** 821-829.

ZHANG, X. Y., BAILEY, S. D. & LUPIEN, M. 2014. Laying a solid foundation for Manhattan - 'setting the functional basis for the post-GWAS era'. *Trends in Genetics,* 30**,** 140-149.

ZHAO, J. J., WANG, X. W., DENG, B., LOU, P., WU, J., SUN, R. F., XU, Z. Y., VROMANS, J., KOORNNEEF, M. & BONNEMA, G. 2005. Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. *Theoretical & Applied Genetics,* 110**,** 1301-1314.