THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Analysis of the Chickpea genome using next generation sequencing data

PRADEEP RUPERAO

M.Tech (Bioinformatics)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

School of Agriculture and Food Sciences

Abstract

Chickpea is an annual, self-pollinating, diploid (2n=2x=16) and the second most important food legume crop with major production areas in the Indian sub-continent, West Asia and North Africa. Australia is the largest exporter of chickpea and yield for 2015 has been forecasted at 990,000 tonnes (http://www.agriculture.gov.au/abares/media-releases/2015). The main constraint for increasing yield is the susceptibility of the plant to foliar disease, ascochyta blight, fusarium wilt. Chickpea breeding aims at high yielding cultivars that combine long-lasting resistance against both biotic and abiotic stress. Emerging sequencing technologies have enhanced the availability of crop genomic resources. Analysis of huge amounts of genomic data comes with major computational challenges.

This thesis describes a new method for assessing the quality of reference genome assembly by flow sorting and isolating chromosomes based on size. This chromosomal DNA was sequenced and mapped to the reference genome assembly. This approach will reduce the genome DNA complexity to a chromosome level and is expect to correspond to only one chromosome assembly. Appling this approach on both released *desi* and *kabuli* reference genomes has shown miss-assembly. The *desi* genome was of poor quality compared to the *kabuli* genome. To fix these mis-assemblies, we developed a novel method called skimGBS for rearranging the fragmented sequences. In this approach, genotypes were called from population individuals to construct haplotype blocks. Based on the haplotype block signature, contigs/fragmented sequences were reordered as new assemblies. Using this approach, both *desi* and *kabuli* genomes were improved by placing unplaced contig sequences into chromosomes. Furthermore, these improved reference assemblies were assessed and annotated.

This thesis also reports identification of more than 800,000 high quality SNPs by sequencing 69 diverse Australian chickpea accessions. Gene loss, genetic relatedness, population structure and diversity analysis was also performed. The public accessibility of the data and above results provides a valuable resource to support chickpea research.

The developed methodology can also apply to other genomics studies, and will therefore be a valuable approach to assist crop improvement and further breeding approaches.

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

Pradeep Ruperao, Chon-Kit Kenneth Chan, Sarwar Azam, Miroslava Karafiátová, Satomi Hayashi, Jana Čížková, Rachit K. Saxena, Hana Šimková, Chi Song, Jan Vrána, Annapurna Chitikineni, Paul Visendi, Pooran M. Gaur, Teresa Millán, Karam B. Singh, Bunyamin Taran, Jun Wang, Jacqueline Batley, Jaroslav Doležel, Rajeev K. Varshney, David Edwards. "A chromosomal genomics approach to assess and validate the *desi* and *kabuli* draft chickpea genome assemblies." Plant Biotechnology Journal, 12 (2014), 778-86.

Bayer Philipp E, **Pradeep Ruperao**, Annaliese S. Mason, Jiri Stiller, Chon-Kit Kenneth Chan, Satomi Hayashi, Yan Long, Jinling Meng, Tim Sutton, Paul Visendi, Rajeev K. Varshney, Jacqueline Batley, David Edwards. "High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*." Theoretical and Applied Genetics 128, no. 6 (2015): 1039-1047.

Sandip M Kale, Deepa Jaganathan, **Pradeep Ruperao**, Charles Chen, Ramu Punna, Himabindu Kudapa, Mahendar Thudi, Manish Roorkiwal, Mohan AVSK Katta, Dadakhalandar Doddamani, Vanika Garg, P B Kavi Kishor, Pooran M Gaur, Henry T Nguyen, Jacqueline Batley, David Edwards, Tim Sutton and Rajeev K Varshney. (2015) “Prioritization of candidate genes in "QTL-hotspot" region for drought tolerance in chickpea (*Cicer arietinum L*.)” Scientific Reports 5, no. 15296 (2015).

Mohd-Yusoff Nur Fatihah, **Pradeep Ruperao**, Nurain Emylia Tomoyoshi, David Edwards, Peter M. Gresshoff, Bandana Biswas, and Jacqueline Batley. "Scanning the Effects of Ethyl Methanesulfonate on the Whole Genome of *Lotus japonicus* Using Second-Generation Sequencing Analysis." G3: Genes| Genomes| Genetics 5, no. 4 (2015): 559-567.

Doddamani Dadakhalandar, Aamir W. Khan, Mohan AVS K. Katta, Gaurav Agarwal, Mahendar Thudi, **Pradeep Ruperao**, David Edwards, and Rajeev K. Varshney. "CicArVarDB: SNP and InDel database for advancing genetics research and breeding applications in chickpea." Database (2015): bav078.

Azam Sarwar, Abhishek Rathore, Trushar M. Shah, Mohan Telluri, Bhanu Prakash Amindala, **Pradeep Ruperao**, Mohan AVSK Katta, and Rajeev K. Varshney. "An Integrated SNP Mining and Utilization (ISMU) Pipeline for Next Generation Sequencing Data." PLoS One (2014): e101754.

Shengyi Liu, Yumei Liu, Xinhua Yang, Chaobo Tong, David Edwards, Isobel A. P. Parkin, Meixia Zhao, Jianxin Ma, Jingyin Yu, Shunmou Huang, Xiyin Wang, Junyi Wang, Kun Lu, Zhiyuan Fang, Ian Bancroft, Tae-Jin Yang, Qiong Hu, Xinfa Wang, Zhen Yue, Haojie Li, Linfeng Yang, Jian Wu, Qing Zhou, Wanxin Wang, Graham J King, J. Chris Pires, Changxin Lu, Zhangyan Wu, Perumal Sampath, Zhuo Wang, Hui Guo, Shengkai Pan, Limei Yang, Jiumeng Min, Dong Zhang, Dianchuan Jin, Wanshun Li, Harry Belcram, Jinxing Tu, Mei Guan, Cunkou Qi, Dezhi Du, Jiana Li, Liangcai Jiang, Jacqueline Batley, Andrew G Sharpe, Beom-Seok Park, **Pradeep Ruperao**, Feng Cheng, Nomar Espinosa Waminal, Yin Huang, Caihua Dong, Li Wang, Jingping Li, Zhiyong Hu, Mu Zhuang, Yi Huang, Junyan Huang, Jiaqin Shi, Desheng Mei, Jing Liu, Tae-Ho Lee, Jinpeng Wang, Huizhe Jin, Zaiyun Li, Xun Li, Jiefu Zhang, Lu Xiao, Yongming Zhou, Zhongsong Liu, Xuequn Liu, Rui Qin, Xu Tang, Wenbin Liu, Yupeng Wang, Yangyong Zhang, Jonghoon Lee, Hyun Hee Kim, France Denoeud, Xun Xu, Xinming Liang, Wei Hua, Xiaowu Wang, Jun Wang, Boulos Chalhoub & Andrew H Paterson. "The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes." Nature Communications 5, 3930 (2014).

Neil Wratten Harsh Raman, Rosy Raman, Andrzej Kilian, Frank Detering, Jason Carling, Neil Coombes, Simon Diffey, Gururaj Kadkol, David Edwards, Margaret McCully, **Pradeep Ruperao**, Isobel A. P. Parkin, Jacqueline Batley, David J. Luckett. "Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*." PLoS One (2014): e101673.

Kaitao Lai, Michał T. Lorenc, Hong Lee, Paul J. Berkman, Philipp Emanuel Bayer, Paul Visendi, **Pradeep Ruperao**, Timothy L. Fitzgerald, Manuel Zander, Chon-Kit Kenneth Chan, Sahana Manoli, Jiri Stiller, Jacqueline Batley, David Edwards. "Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat." Plant Biotechnology Journal 13, no. 1 (2015): 97-104.

Book chapters

Pradeep Ruperao, David Edwards (2015) Bioinformatics: identification of markers from next generation sequence data Methods and Protocols, Methods in Molecular Biology, Vol. 1245. Ed. Batley, J. Humana Press (Springer) (USA).

Pradeep Ruperao, David Edwards (2015) TAG sequence identification of genomic regions: TAGdb. Methods and Protocols, Methods in Molecular Biology. Ed. Edwards, D. Humana Press (Springer) (USA).

Conference Posters

Pradeep Ruperao, Philipp Bayer, Chon-Kit Kenneth Chan, Satomi Hayashi, Manish Roorkiwal, Anu Chitikineni, Tim Sutton, Jacqueline Batley, Rajeev K. Varshney, and David Edwards. Constructing improved chickpea genome assemblies using skimGBS. The Plant and Animal Genome: Asia conference, 13th-15th July, 2015, Singapore.

Pradeep Ruperao, Philipp E. Bayer, Chon-Kit Kenneth Chan, Satomi Hayashi, Manish Roorkiwal, Anu Chitikineni, Tim Sutton, Jacqueline Batley, Rajeev K. Varshney, David Edwards. Constructing improved chickpea reference genome assemblies using skimGBS, The Plant Genome Congress: Asia conference, 19th-20th March, 2015, Kuala Lumpur, Malaysia.

Philipp Emanuel Bayer, **Pradeep Ruperao** , Annaliese S. Mason, Jiri Stiller ,Kenneth Chan, Satomi Hayashi ,Salman Alamery, Reece Tollenaere ,Paul Visendi ,Yan Long, Jinling Meng, Tim Sutton, Rajeev K Varshney, Jacqueline Batley ,David Edwards. Using skim-based genotyping by sequencing for trait association and QTL cloning in *Brassica napus*. The Plant and Animal Genome: San Diego conference, 10th-14th January, 2015, California, USA.

Sandip M Kale, Deepa Jaganathan, **Pradeep Ruperao**, Charles Chen, Ramu Punna, Himabindu Kudapa, Mahendar Thudi, Manish Roorkiwal ,Krishnamohan Katta, P B Kavi Kishor, Pooran M Gaur, Jacqueline Batley , David Edwards, Tim Sutton, Rajeev K Varshney. Identification of Candidate Genes from a "QTL-Hotspot" Region for Drought Tolerance in Chickpea Using Whole Genome Resequencing

Based Bin Mapping. The Plant and Animal Genome: San Diego conference, 10th-14th January, 2015, California, USA.

Dadakhalandar Doddamani, Aamir W Khan, Krishnamohan Katta, Gaurav Agarwal, **Pradeep Ruperao**, David Edwards, Rajeev K Varshney. CicArVarDB: Chickpea Variation Database. The Plant and Animal Genome: San Diego conference, 10th-14th January, 2015, California, USA.

Pradeep Ruperao, Philipp E. Bayer, Chon-Kit Kenneth Chan, Sarwar Azam, Miroslava Karafiátová, Satomi Hayashi, Jana Čížková, Rachit K Saxena, Hana Šimková, Chi Song, Jan Vrána, Annapurna Chitikineni, Paul Visendi, Pooran M Gaur, Teresa Millán, Karam B. Singh, Bunyamin Taran, Jun Wang, Jacqueline Batley, Jaroslav Doležel, Rajeev K. Varshney, David Edwards. A chromosomal genomics approach to assess and validate the *desi* and *kabuli* draft chickpea genome assemblies, The Plant and Animal Genome: San Diego conference, 11th-15th January, 2014, California, USA.

Kaitao Lai, Michal Lorenc, Hong Lee, Paul Berkman, Paul Visendi Muhindira, Philipp Bayer, **Pradeep Ruperao**, Kenneth Chan, Sahana Manoli, Jiri Stiller, Jacqueline Batley, Dave Edwards. Genetic variation across Australian bread wheat varieties, The Plant and Animal Genome: San Diego conference, 14th January, 2013, California, USA.

Presentations

Pradeep Ruperao and Dave Edwards. Gene variability in chickpea genotypes. Pawsey supercomputing workshop, 17-18 August, 2015, Perth, Australia.

Pradeep Ruperao, Philipp Bayer, Chon-Kit Kenneth Chan, Satomi Hayashi, Manish Roorkiwal, Anu Chitikineni, Tim Sutton, Jacqueline Batley, Rajeev K. Varshney, and David Edwards. High resolution skim genotyping assists in Chickpea (*Cicer arietinum*) reference genome assessment and improvement. The Plant and Animal Genome: Asia conference, 13-15 July, 2015, Singapore.

Pradeep Ruperao, Philipp Bayer, Chon-Kit Kenneth Chan, Satomi Hayashi, Manish Roorkiwal, Anu Chitikineni, Tim Sutton, Jacqueline Batley, Rajeev K. Varshney, and David Edwards. Analysis of Chickpea genome using next generation sequencing data. Australia India Strategy Research Fund Project meeting. 22 Feburay 2015. ICRISAT, Hyderabad, India.

Pradeep Ruperao, Philipp Bayer, Chon-Kit Kenneth Chan, Satomi Hayashi, Manish Roorkiwal, Anu Chitikineni, Tim Sutton, Jacqueline Batley, Rajeev K. Varshney, and David Edwards. Analysis of Chickpea genome using next generation sequencing data. B3 Symposium, 24 – 25 November 2014, Queensland University of Technology, Brisbane, Australia.

Publications included in this thesis

1. **Pradeep Ruperao**, Chon-Kit Kenneth Chan, Sarwar Azam, Miroslava Karafiátová, Satomi Hayashi, Jana Čížková, Rachit K. Saxena, Hana Šimková, Chi Song, Jan Vrána, Annapurna Chitikineni, Paul Visendi, Pooran M. Gaur, Teresa Millán, Karam B. Singh, Bunyamin Taran, Jun Wang, Jacqueline Batley, Jaroslav Doležel, Rajeev K. Varshney, David Edwards. "A chromosomal genomics approach to assess and validate the *desi* and *kabuli* draft chickpea genome assemblies." Plant Biotechnology Journal 12, no. 6 (2014): 778-786. Incorporated into Chapter 2.

Contributor	Statement of contribution
Pradeep Ruperao (85%)**,** Chon-Kit Kenneth Chan (5%), Sarwar Azam (5%), Paul Visendi (5%)	Analysis and interpretation of data and coding
Pradeep Ruperao (60%), Jacqueline Batley (10%), Chon-Kit Kenneth Chan (10%), David Edwards (20%)	Writing publication
David Edwards (60%), Jacqueline Batley (20%), Rajeev K Varshney (20%)	Project supervision
Satomi Hayashi (100%)	DNA Sequencing and data production
Rachit K Saxena (100%)	Provided material for sequencing
Hana Šimková (10%), Chi Song (10%), Jan Vrána (10%), Annapurna Chitikineni (10%), Pooran M Gaur (10%), Teresa Millán (5%), Karam B Singh (10%), Bunyamin Taran (10%), Jun Wang (10%), Jana Čížková (5%), Miroslava Karafiátová (10%).	Provided comments on manuscript
Jaroslav Doležel (100%)	Chromosome isolation
David Edwards (80%), Rajeev K Varshney (20%)	Experiment designed and conception

2. Bayer Philipp E, **Pradeep Ruperao**, Annaliese S. Mason, Jiri Stiller, Chon-Kit Kenneth Chan, Satomi Hayashi, Yan Long, Jinling Meng, Tim Sutton, Paul Visendi, Rajeev K. Varshney, Jacqueline Batley, David Edwards. "High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*."

Theoretical and Applied Genetics 128, no. 6 (2015): 1039-1047. Incorporated into Chapter 3.

Contributor	**Statement of contribution**
Philipp E Bayer (70%), **Pradeep Ruperao** (30%)	Designed experiments and improved SkimGBS pipeline
Pradeep Ruperao (100%)	Analysis and interpretation of chickpea data
Philipp E Bayer (100%)	Analysis and interpretation of brassica napus data
Annaliese S Mason (100%)	Help with centromere location and recombination mappings
Jiri Stiller (30%), Chon-Kit Kenneth Chan (70%)	Coding
Satomi Hayashi (100%)	DNA-sequencing
Yan Long (40%), Jinling Meng (60%)	Provided material for sequencing of the Brassica DH population
Paul Visendi (100%)	Helped with debugging and comments on paper
Philipp E Bayer (70%), Pradeep Ruperao (10%), David Edwards (20%)	Writing
Rajeev K Varshney (100%)	Grew the *Cicer* RIL population
David Edwards (60%) and Jacqueline Batley (20%), Rajeev K Varshney (20%)	Supervision

3. Sandip M Kale, Deepa Jaganathan, **Pradeep Ruperao**, Charles Chen, Ramu Punna, Himabindu Kudapa, Mahendar Thudi, Manish Roorkiwal, Mohan

AVSK Katta, Dadakhalandar Doddamani, Vanika Garg, P B Kavi Kishor, Pooran M Gaur, Henry T Nguyen, Jacqueline Batley, David Edwards, Tim Sutton and Rajeev K Varshney. (2015) “Prioritization of candidate genes in "QTL-hotspot" region for drought tolerance in chickpea (*Cicer arietinum L*.)” Scientific Reports 5, no. 15296 (2015). Partially incorporated in Chapter 3.

Contributor	Statement of contribution
Pradeep Ruperao (100%)	SNP and GBS analysis
Sandip M Kale (50%), Deepa Jaganathan (50%)	QTL analysis and gene enrichment analysis
Himabindu Kudapa (50%), Deepa Jaganathan (50%)	Gene expression analysis
Sandip M Kale (40%), Mohan AVSK Katta (20%), Dadakhalandar Doddamani (20%), **Pradeep Ruperao** (20%)	Coding
Pooran M Gaur (100%)	Provided genetic material
Sandip M Kale (20%), Deepa Jaganathan (20%), Mahendar Thudi (10%), Jacqueline Batley (10%), David Edwards (10%), Tim Sutton (10%) & Rajeev K Varshney (20%)	Wrote the manuscript
Charles Chen (5%), Ramu Punna (5%), Himabindu Kudapa (5%), Mahendar Thudi (5%), Manish Roorkiwal (5%), Mohan AVSK Katta (5%), Dadakhalandar Doddamani (5%), **Pradeep Ruperao** (5%), Vanika Garg (5%), P B Kavi Kishor (5%), Pooran M Gaur (10%), Henry T Nguyen (10%), Jacqueline Batley (10%), David Edwards (10%), Tim Sutton (10%)	Provided comments on manuscript
David Edwards (10%), Jacqueline Batley (10%), Tim Sutton (20%), Rajeev K Varshney (60%)	Project supervision

David Edwards (20%), Tim Sutton (20%), Rajeev K Varshney (60%)	Experiment designed and conception

Contributions by others to the thesis

Principal supervisor Prof. David Edwards and co-supervisor Dr. Rajeev K Varshney contributed to the conception and design of the project. Prof. David Edwards and co-supervisor Prof. Jacqueline Batley contributed to the editing of the sections and critically revising the sections in the thesis. Dr. Rajeev K Varshney, Prof. David Edwards, Prof. Jacqueline Batley and Dr. Tim Sutton selected chickpea germplasms for whole genome sequencing. Michal T. Lorenc implemented SGSautoSNP and Philipp E Bayer and CCK Chan further modified for additional features. Philipp E Bayer also implemented the contigPlacer program. Agnieszka A Golicz implemented gene loss program SGSgeneloss. Annaliese S. Mason and Satomi Hash produced sequence data.

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

None

Acknowledgements

This thesis and the research presented would not have been possible without the support of a number of people. Firstly, i would like to thank my parents for their support to move on with my research towards PhD in Australia. For my PhD, it was a sign of encouragement to have a positive response from my wife in the early days of my married life. I thank my wife for understanding the importance of my dedication towards this committed work and being supportive in all aspects of this work.

I personally thank to my co-supervisor (and also friend and well-wisher) Dr. Rajeev K Varshney who encouraged me for this project to work with Prof. Dave Edwards and Prof. Jacqueline Batley. I sincerely thank my principal supervisor, Prof. Dave Edwards, for his guidance, understanding, patience, and most importantly, focus on strengthening my weakness in (PhD) research work. His positive attitude, pleasing and friendly nature towards students made me more responsible to reach his expectations. I am also thankful to my co-supervisor Prof. Jacqueline Batley for her assistance and guidance in research, manuscripts and thesis writings. I thank all supervisors for being patience to expect research progress while I got busy with other personal and official issues.

Additionally, I would like to thank my un-official co-supervisor Dr. Jiri Stiller, who thought me research working style and coding behaviour. I also thank my colleagues from CEG-ICRISAT, The University of Queensland and University of Western Australia (Prof. Dave and Prof. Jacqueline group). They have helped me in generating data, understanding the work flow, scripts and computational resource. Some of them have added their ideas to do the best in understanding the analysis and polished my thinking about my PhD research.

I acknowledge the funding provided by Australia and India government towards Australia India Strategic Research Fund (AISRF) for chickpea crop improvement. I also acknowledge financial support from GRDC and School of Agriculture and Food Science (SAFS) towards the travel grant. I also appreciate the support from Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF), Queensland Facility for Advanced Bioinformatics (QFAB),

PAWSEY supercomputing centre, University of Western Australia and University of Queensland.

Keywords

Chickpea, desi, kabuli, next generation sequencing, genome assembly, genome assessment, SGSautoSNP, SGSGeneloss, SkimGBS and GBrowse.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102 Bioinformatics, 50%

ANZSRC code: 060408 Genomics, 50%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 50%

FoR code: 0604, Genetics, 50%

Table of Contents

List of Figures

List of Tables

List of Abbreviations used in the thesis

2GS	Second-generation sequencing
bp	Base pairs
Gbp	Gbp gigabase-pairs (billion base-pairs)
GO	GO gene ontology
indel	Indel insertion or deletion mutation
Mbp	Mbp megabase-pairs (million base-pairs)
NGS	NGS next-generation sequencing
QTL	QTL quantitative trait loci
SRA	SRA short read archive
WGS	Whole-genome shotgun/sequence
autoSNPdb	Automatic annotated Single Nucleotide Polymorphism database
BLAST	Basic Local Alignment Search Tool
EST	Expressed Sequence Tag
GBrowse2	Generic Genome Browser version 2
NGS	Next-generation Sequencing
PCR	Polymerase Chain Reaction
SNP	Single-Nucleotide Polymorphism
GBS	Genotyping by Sequencing
TAGdb	Tag database
CNV	Copy Number Variations
PAV	Presence and Absence Variations
FAOSTAT	Food and Agriculture Organization Corporate Statistical Database

1 Literature review

1.1 Introduction and classification

Chickpea (*Cicer arietinum L.*) is a post-rainy season grain crop legume and is the second most widely grown food legume crops after common bean, with annual production of 13.1 M tons in 2013 (FAOSTAT). It is an annual, self-pollinated, diploid (2n=16) plant grown in a wide range of environments including the Mediterranean, South and West Asia, North America, and East Africa. Turkey and India are two major centres of diversity.

Globally there are around 98,000 chickpea accessions held at various germplasm collections, mainly at ICRISAT, ICARDA and national gene banks of Russia and USA. To better target and utilize the germplasm by plant breeders, the trait characteristics of accessions are made available. These can be retrieved with ICIS (International Crop Information System) platform for search and retrieved across all combined relational databases from ICRISAT, ICARDA, USDA and ATFCC. Chickpea is a widely cultivated food legume, domesticated in the fertile Crescent (Zohary D, 2012). Both wild and cultivated chickpeas were predominantly self-crossed (Singh F, 1995). Cultivated chickpea varieties differ from its wild relatives in its plant growth habit, altered phenology and seed coat texture. Further variations exist within cultivated chickpea, such as seed features like size, shape, coat thickness, and colour. This again categorized as *desi* and *kabuli* types.
A major form of chickpea as *kabuli* type has unique morphological characteristics like a white flower and light-coloured seed coats that do not exist in the wild progenitor. The origin of *kabuli* types has been enigmatic and polyphyletic. The *kabuli* form has developed from the phase of phenotypic diversification after the initial domestication of cultivated chickpea (Varma Penmetsa *et al.*, 2016). The greater similarity of *desi* seed with wild type (*C.reticulatum*) seed represents the early domesticated form (Ladizinsky G, 1976) and *kabuli* as subsequently derived type.

Two main chickpea cultivar groups, *kabuli* and *desi*, are classified on the basis of morphological characters and geographic distribution. The *desi* type are characterized by pink flowers small usually angular, yellow-brown coloured seeds

and grown mainly in Indian subcontinent and east Africa. The *kabuli* type chickpea are native to the Mediterranean and Near-East region, having white flowers and large smooth or wrinkled light-coloured seeds. Grains of *kabuli* are bigger in size, weighing in excess of 26 grams per 100 seeds, round white or cream coloured and thinner in seed coat than *desi*. Plants of *kabuli* types are tall and are devoid of purple or violet pigmentation of the leaves, stems, flowers and pods. *Desi* types generally have seeds that are less than 26 grams per 100 seeds and have a rough, angular appearance with various colourations from light tan to black with all gradations in between(Muehlbauer and Rajesh, 2008) . *Desi* type is consumed mostly in the Middle East and South-East Asia whereas *kabuli* is a valuable global commodity. The mean value of protein (25.4 vs 24.4%), fat (3.7 vs 5.1%), carbohydrate (47.4 vs 55%), crude fibre (11.2 vs 3.9%), ash (3.2 vs 2.8%) and caloric value (327 vs 365 kcal/100 g) were for *desi* vs *kabuli* chickpeas respectively (Khan *et al.*, 1995) . The chickpeas products contained 8.9-21.1% protein (N x 6.25), 3.1-21.8% fat, 53.4-75.9% carbohydrate, 1.6-11.1% crude fibre, 1.2-5.9% ash, 226-360mg Ca, 126-315 mg P, 3.8-8.2 mg Fe, 1.8-5.4 mg Zn, 1.5-5.4 mg Mn, 0.6-1.1 mg Cu and 370-490 kcal per 100 g. *Kabuli* has greater tolerance to cold, and iron deficiency than the *desi* (G.C. Hawtin, 1979) . The mean seed coat percentage, which plays an important role in nutritive value and cooking time, reported being 14.2 for *desi* and 4.9 for the *kabuli* (Singh.U, 1980) . Candidate genes responsible for disease resistance, agronomic traits and traits that distinguish the *desi* and *kabuli* were identified (Varshney *et al.*, 2013) and molecular level diversity was documented using a *desi* x *kabuli* cross (Bharadwaj *et al.*, 2011). Chickpea is a good source of many essential mineral nutrients, protein, and dietary fibre and are low in saturated fat, rich in unsaturated fatty acids and β- carotene and also cheap source of protein for people in developing countries. Additionally, it improves soil fertility fixing atmospheric nitrogen, meeting up to 80% of its nitrogen requirement from symbiotic nitrogen fixation (Saraf and M.A.S., 1998). In India the crop is consumed primarily as dhal, a preparation produced by decortication of the seed and separating the cotyledons. The decorticated and split cotyledons are then used to produce a thick soup that is generally served with rice. Chickpea is also used as a whole pulse and is soaked and boiled. “Chole” is a traditional dish made from whole chickpeas in India. Dry roasted chickpea seasoned with various spices are a popular snack in most countries of the Middle East and North Africa. Similar to spinach dish, the young

leaves of chickpea are often harvested green and cooked. In the semi-arid regions, the crop is grown as a cool weather crop while in many production regions it is grown as winter crop, particularly in South Asia, the Mediterranean, Australia and Mexico. Maturation of cultivars varies and cultivars will complete their life cycle between 65 and 180 days (Muehlbauer and Rajesh, 2008). Chickpea is considered as a low input crop and is mainly grown on residual soil moisture. The major constrains of chickpea are biotic stress like fusarium wilt, pod borer and ascochyta blight and abiotic stress like temperatures, drought and salinity (Clarke and Siddique, 2004; Pande *et al.*, 2005). Due to recent advancements in genomics including sequencing technology, considerable progress has been made in chickpea improvement using conventional breeding approaches. Before 2005, the global production has remained stagnant (Millan *et al.*, 2006) but lately the total production of chickpea has increased steadily to 13 million tonnes (Figure 1-1).

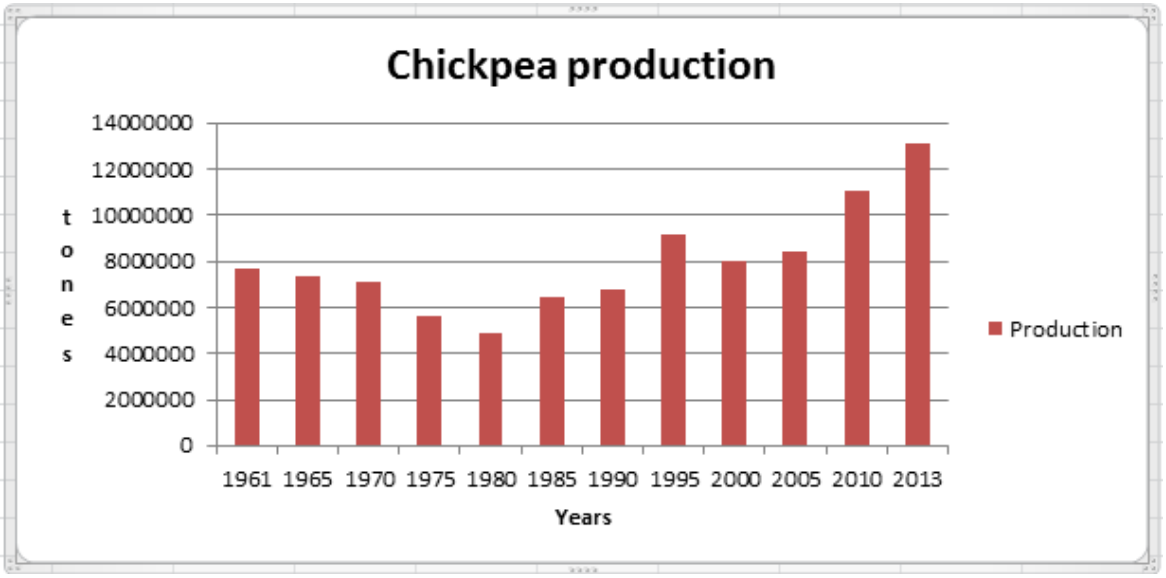


Figure 1-1: Global chickpea annual production from 1961 to 2013.

The annual production of chickpea is 13.1 million tons (CGIAR, 2013), which has expanded from 5 million metric tons (mmt) (in 1980) over the past 3 decades. During 2013, the global chickpea harvested area was 13.5 million ha, production was 13.1 mmt and yield was 0.96 tonne/hectare (FAOSTAT, 2013). Chickpea is grown in over 50 countries of which India (8.8 mmt) is the largest chickpea producing country followed by Australia (0.81 mmt), Turkey (0.56 mmt), Myanmar (0.49 mmt), Ethiopia

(0.24 mmt), Iran (0.29mmt), Pakistan (0.75mmt) and Canada (0.16mmt) (based on 2013 annual production) (FAOSTAT, 2013). Figure 1-2 illustrates the global chickpea production in metric tonnes.

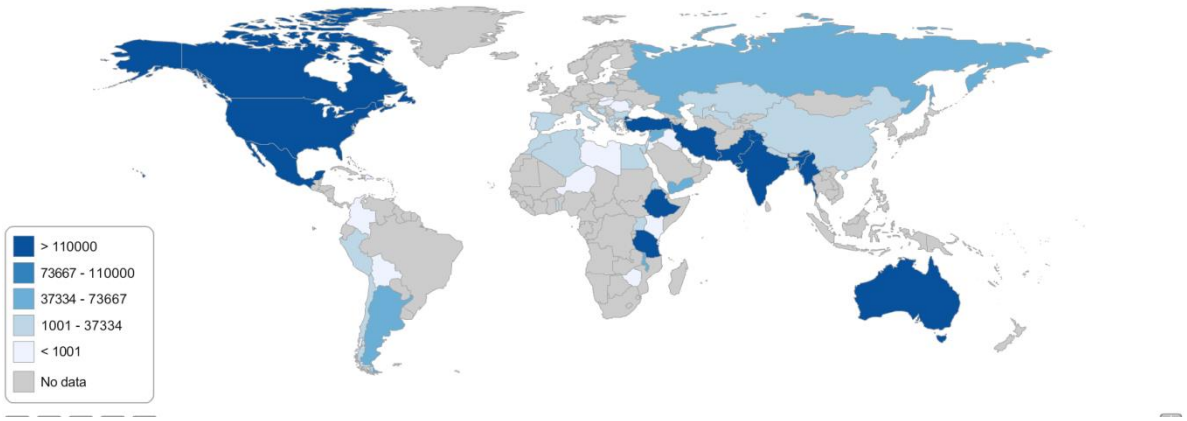


Figure 1-2 : Illustrating global chickpea production in the year 2013 (FAOSTAT).

Population growth is expected to reach 9 billion by 2050. The availability of the reference genome assembly of chickpea *desi* and *kabuli* types, identifying genetic variation, markers and genes assist breeders in designing high-yield varieties.

1.2 Archaeological sites

The genus *Cicer* holds more than 40 species (van der Maesen, 1987) which includes nine annual species. The closest species is *Cicer reticulatum* Ladiz, found in oak shrub formations on hilly limestone bedrock (Ladizinsky, 1975) which is the immediate wild progenitor of domesticated chickpea (Patil *et al.*, 1995). Early chickpea domestication has been found at several ancient sites, including the Pre-pottery Neolithic B period of a number of Near-East sites (Zohary, 1993) and Dja'de (9,000 BC) in Syria, Cayonu (7250-6750 BC), Hacilar (6700 BC), and Akarcay tepe (7280-8700 BP) in Turkey and Jericho (8350 BC to 7370 BC) in West Bank (Kris, 2014). The earliest occurrence of chickpea in India dates back to 2000 BC at Atranjikhera in Uttar Pradesh (Chowdhury, 1971).

1.3 Environmental and growth challenges in chickpea production

With increased global warming, the amount of crop yield has drastically declined due to various stresses. The amount of atmosphere carbon dioxide, temperature and rainfall affect crop yield dramatically (Mendelsohn *et al.*, 1994). The impact of climate change on crops effect not only by period of cultivable season but also affecting soil nutrition, dominance of pest and disease. The increasing temperature may promote increase crop production in the cooler temperature regions but decrease production in warmer tropical regions. Crop production not only depends on climate change but also on the agricultural systems and crop varieties, including their ability to resist stress. To cope with these challenges, adaptations strategies are applied including cultivating better germplasm.

1.3.1 Abiotic stress

1.3.1.1 Drought

Drought tolerance is defined as the ability of a genotype/variety within a species to be more productive than others under water deficit conditions (Quisenberry, 1982). Chickpea is one of the important cool season food legume crops and is generally subjected to drought. Breeding for drought tolerance is complex because of the interaction of drought with other stresses. However, traits such as dehydration avoidance, early flowering and maturity are components of drought tolerance.

1.3.1.2 Heat

Heat is another component affecting seed yield of grain legume crops during flowering and pod-filling stages. The high temperature leads to accelerated rate of plant development (Gan *et al.*, 2004) and shortened period of growth of reproductive organs (Angadi *et al.*, 2000). High temperature during reproductive development also has negative impact on pollen viability, fertility (Hall, 2004), floral bud development (Russell-Stevens *et al.*, 2002) and also seed composition (Thomas *et al.*, 2003). Chickpea is sensitive to high temperatures mainly at the reproductive stage (Singh *et al.*, 1994).

1.3.1.3 Cold

In contrast to high temperature, low temperature (either chilling or freezing) affects crop production through low growth rate and increased chlorosis and necrosis. Cold also affects symbiotic maintenance and causes damage during the reproductive

phase (Maqbool *et al.*, 2010). Freezing stress is common during vegetative growth in Europe and Central Asia, especially during early spring (Singh *et al.*, 1994).

1.3.1.4 Higher level of CO_2

Increasing CO_2 levels could provide a basis to offset losses in agricultural production caused by other factors like drought and temperature stress. Recent studies reported that cereals have developed sink capacity to perform better under high CO_2 condition as they can produce more photo-assimilates. In legumes N_2 fixation levels increases at elevated CO_2, which is critical in maintaining C and N ratio in the ecosystem (Rogers *et al.*, 2009).

1.3.2 Biotic stress

The impact of increasing temperature and CO_2 on pest population and pest – crop interaction has both pros and cons (Deka S, 2009). The productivity and quality of chickpea are affected by soil borne diseases (wilt/root and nematodes) and insect pests (pod borer, leaf miner). The extent of damage caused by these pests depends on factors like temperature, moisture, pathogen and insects. Sometimes climate change can also favour the chickpea crop (from insects affecting). For example, in North Africa pod borers attack chickpea at very late stage without affecting the crop production. Due to climate change, the level of infestation has changed which could be a major threat to chickpea production in West Asia and North Africa. Late rains can cause heavy pod infection by ascochyta blight and rainfall above critical threshold leading to blight epidemics. Thus, in addition to existing biotic constraints to chickpea productivity, new pests are likely to become important under climate change like increased dry conditions can favour dry root rot in tropical countries.

1.4 Chickpea genomics

Chickpea germplasm is maintained by different institutes across the world, including the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), India; the National Bureau of Plant Genetics Resource (NBPGR), India and the International Centre for Agricultural Research in Dry land Areas (ICARDA), Syria. The availability of large number of accessions will enhance the breeding programmes (Upadhyaya *et al.*, 2011).

Advanced bioinformatics tools will enhance the understanding of plant genomics at various biological and environmental levels, and boost traditional plant breeding

techniques. Recent emerging technologies from whole genome sequencing (WGS) to discovery of novel and high throughput genetic and molecular techniques are now enabling researchers to study the genes and associated functions contributing to the improved productivity and quality of crop varieties. Recently national and international collaborations have come forward to draft the reference genome of both *desi* and *kabuli* types. The International Chickpea Genetics and Genomics Consortium (ICGGC) produced a genome assembly of ~532 Mb (71.89%) for the CDC Frontier *kabuli* type, along with 28,269 gene predictions (Varshney *et al.*, 2013). The genome was analysed in detail, including gene prediction, functional analysis, synteny, phylogenetic analysis, SNP identification from five genotypes (ICC506, ICC1882, ICC4958, ICCC37, PI489777), SSR and genetic diversity analysis of 90 genotypes. The Indian National Institute of Plant Genome Research, India group also released a *desi* (ICC4958) genome assembly of 520 Mb, covering 70% of the genome with 27,571 predicted genes (Jain *et al.*, 2013). Analysis of this genome included repetitive element prediction, gene prediction/duplication and functional annotation. Gene expression analysis was also performed.

Application of biotechnology approaches to chickpea crops can overcome production constraints such as biotic and abiotic stress to develop varieties with improved resistance/tolerance to these stresses.

1.4.1 Gene discovery and QTL analysis

In the progress of crop improvement through phenotypic selection, breeders experience the difficulties such as environment interaction, expense and time consuming screening methodologies for target traits. Molecular markers as a selection tool provide the efficiency of breeding program by reducing environmental variability, facilitating earlier selection, reduced cultivation expenses and labour. Molecular markers can be linked to agronomic traits, and targeting these regions can facilitate breeding strategies for crop improvement. Useful characteristics had been introgressed from one chickpea type to the other (Bahl.P.N, 1988) . For example resistance to Fusarium wilt or Helicoverpa had been transferred to *kabuli* types and the resistance to Ascochyta blight from *kabuli* to *desi* (Yadav, 2006) . With available limited genetic resources till the early eighties, *kabuli* were thought not to be adapted for warmer environments. However, the knowledge generated by the gene action and genetic variation for crop improvement efforts has enabled to transfer useful

alleles from the *desi* to *kabuli* varieties such as ICCV2 and KAK 2. There are numerous known molecular markers linked to stress tolerance (Dita *et al.*, 2006; Millan *et al.*, 2006), salt tolerance (Samineni, 2010), drought tolerance (Azam *et al.*, 2014; Kale *et al.*, 2015), heat tolerance (Thudi *et al.*, 2014), yield (Rehman *et al.*, 2011) and biotic stress tolerance (Flandez-Galvez *et al.*, 2003) in chickpea. With availability of a reference genome, the potential to map different QTL and identify the linked molecular markers facilitates the transfer of several QTL in one improved cultivar simultaneously.

1.4.2 Genetic variation

The occurrence of genetic variation provides the genetic material for natural selection. These variations brought by mutation result in permanent changes in DNA molecules and this genetic material are transferred to the next generation. In chickpea, many genomic studies focus on biotic fungal, bacterial, viral, nematode and insect stress affecting the crop. Recently QTLs related to resistance to Ascochyta blight, fusarium wilt, drought tolerance and pod borer diseases were studied (Kale *et al.*, 2015; Millan *et al.*, 2006; Rajesh *et al.*, 2007). In recent years, the incidence of disease is getting less like, Fusarium wilt has been reported in the US, Spain and India (Jimenez-Gasco *et al.*, 2004) and genes linked to fusarium wilt resistance have also been reported for various races (Millan *et al.*, 2006; Sharma *et al.*, 2004). Recently, molecular markers of these genes were validated in commercial cultivars (Gowda *et al.*, 2009).

1.5 Sequencing technologies

With new sequencing technologies it is possible to re-sequence the entire plant genome or transcriptome more efficiently (at low cost). Sequencing multiple (hundreds or even thousands) of related genomes will help predict the genetic diversity within and between germplasm pools. Availability of sequencing information enhances the detailed study and development of large scale molecular markers for linkage mapping, association mapping, epigenetic modifications, population genetics and building genome assemblies (Close *et al.*, 2009). The current sequencing technologies (second generation sequencing - SGS) produce more coverage data compared to the traditional Sanger method (Pareek *et al.*, 2011). Second Generation Sequencing (SGS) technologies are mainly based on immobilization of DNA samples onto a solid support, cyclic automated sequencing reactions and the

detection of the molecular events by imaging. Technologies include GS FLX by 454 Life Technologies/Roche; the CGA Platform by Complete Genomics; Genome Analyzer, HiSeq and MiSeq by Illumina, PacBio RS by Pacific Biosciences, Ion Proton from Life Technologies and SOLiD by Applied Biosystems. Each platform can generate data required for a draft genome with its own potential range of quality and accuracy (Table 1-1). Among Roche, Illumina and Life Technologies sequencing platforms, performance wise, HiSeq has the highest throughput per run and lowest error rates.

Table 1-1: Sequencing platform and properties.

Platform	Reads	Max read length	Run time	Application reference
Sequence by ligation				
SOLiD 5500 Wildfire	~700 M	50 (SE)	6 d	Tomato (Sato *et al.*, 2012)
SOLiD 550 xl	~1.4 B	50 (SE)	10 d	Rubber tree (Rahman *et al.*, 2013)
BGISEQ-500 FCS	!!	50 – 100 (SE/PE)	24 h	(Yuzuki, 2015)*
Sequence by synthesis				
Illumina MiSeq v3	50 M	300 (PE)	21-56 h	Chloroplast genome, Venturia inaequalis (Mariac *et al.*, 2014; Villani *et al.*, 2016)
Illumina NextSeq 500/550 High output	800 M	150 (PE)	29 h	Ginger (Gaur *et al.*, 2016)
Illumina HiSeq2500 v4	4 B	125 (PE)	6 d	Wheat (Saint Pierre *et al.*, 2016)
Illumina HiSeq3000/4000	2.5 B	150 (PE)	3 d	Wheat (Holtz *et al.*, 2016)
Illumina HiSeq X	3 B	150 (PE)	<3 d	(Fulton, 2016)*
454 GS FLX Titanium XL+	~1 M	Up to 1Kb (SE,PE)	23 h	Chickpea (Parween *et al.*, 2015; Varshney *et al.*, 2013)
Ion Proton	60-80M	200 (SE)	2 – 4 h	(Rife *et al.*, 2015)*
Single-molecule real-time long reads				

Pacific BioSciences RSII	~55,000	~20Kb	4 h	Rice (Wang *et al.*, 2014)
Pacific BioSciences Sequel	~350,000	8-12Kb	6 h	(Heger, 01 Oct 2015)*
Oxford Nanopore MK 1 MinION	>100,000	Up to 200Kb	48 h	!!

B = Billion, M = Million, h = hours, d = days, Kb = kilobase pairs, PE = paired end sequencing, SE = single end sequencing, !! = recent product (information not available), * = Product reference

1.5.1 First generation sequencing

In 1975, Sanger introduced the concept of DNA sequencing by primed synthesis with DNA polymerase (Sanger, 1975). Later in 1977, Sanger introduced dideoxy DNA sequencing technique based on the chain-terminating dideoxynucleotide analogues (Sanger *et al.*, 1977) and Allan Maxam introduced the chemical degradation DNA sequencing technique (Maxam and Gilbert, 1977). Sanger sequencing was commercialized by Applied Biosystems (ABI), EMBL and Pharmacia Amersham, later General Electric (GE) healthcare.

1.5.2 Second generation sequencing

In 2005, 454 Life Sciences introduced the first commercially available NGS platform, named GS20. Later Roche applied science acquired 454 Life sciences and extended the improvements to introduce GSFLX titanium. Further developments in sequencing were later commercialized by Roche, Illumina-Solexa, ABI, Helicos and other companies. Principally the DNA molecules are sequenced either in a stepwise iterative process or in continuous real-time manner in a massively parallel fashion. Second generation sequencing produces massive qualitative and quantitative DNA information at low cost in short time frame.

1.5.2.1 Genome Analyzer, HiSeq, MiSeq

The Genome Analyzer and more recent HiSeq and MiSeq systems from Illumina are based on immobilizing linear sequencing library fragments using solid support amplification. This technology use fluorescent labelled reversible terminator nucleotides. Sequencing by synthesis employs four distinct fluorophores and

reversibly terminated nucleotides producing reads up to 300 bp. The Illumina sequencing technology dominates the DNA sequencing market (Bentley *et al.*, 2008).

1.5.2.2 CGA Platform

The CGA technology is based on preparation of circular DNA libraries and rolling circle amplification (RCA) to generate DNA nanoballs that are arrayed on a solid support (Drmanac *et al.*, 2010). Currently this system has only been optimised for human genome sequencing due to the complexity of data analysis.

1.5.3 Third generation sequencing

1.5.3.1 PacBio RS

PacBio RS is a single molecule real time (SMRT) sequencing system developed by Pacific Biosciences (Eid *et al.*, 2009). The high error rate of Pacific Biosciences reads has limited their direct use in SNP discovery, but their long reads offer an advantage to scaffold and validate genomic sequences.

1.5.3.2 Semiconductor sequencing

Life Technology and Ion Torrent have developed the Ion Personal Genome Machine. It is embedded with an array of semiconductor chips capable of detecting nucleotide incorporation (by the release of a hydrogen ion from natural nucleotides) and sensing pH changes. DNA sequencing is done based on detection of hydrogen ions that are released during the polymerization of DNA.

1.5.3.3 Nanopore sequencing

Nanopore sequencing is based on a theory that recording the current nucleic acids passing through a pore could be used to distinguish the sequence of individual bases within the DNA chain. Oxford Nanopore included the exonuclease and standard DNA sequencing would represents a possible solution to some of the technical hurdles found in nanopore sequencing (Clarke *et al.*, 2009).

1.6 Genome assemblies

With the recent advancements in high through put sequencing technologies, researchers are able to sequence genomes more accurately, faster and cheaper than using Sanger sequencing. The data quality also increased along with the length of sequence reads compared to short reads produced at the initial stage of NGS

development. Converting the fragmented genome into a high-quality, finished genome sequence is a bottleneck of these projects.

Assembly is a procedure to arrange all fragments of overlapping sequence reads to construct a consensus long sequence. At the initial stage of assembly, reads are grouped into contigs and then contigs into scaffolds. Based on the data coverage, genome size and repetitiveness of genome, most assemblers generate partially assembled output as contig sequences. Scaffolding is a procedure of assembling to order and orient the contigs to construct long sequences generally called as supercontigs or scaffolds. The quality of assemblies is measured by comparing the number of bases incorporated into contig and scaffolds to the estimated size of the genome. Generally the assembly size is measured by statistics including maximum length of contigs, combined total length and N50. The genome N50 represents the smallest contig in the set whose combined length represents at least 50% of the assembly.

Even though advance sequencing technologies generate high throughput reads for assembly, the read lengths are short. A challenge for assembly software is handling repeat sequences, especially repeats longer than the reads. Sequencing errors are an additional challenge in assembly. Software must be able to tolerate the imperfect sequence alignments to avoid missing true joins. False positive joins induce the formation of chimeric assemblies. Coverage variations by copy number variations or by chance in WGS may also induce gaps in assemblies.

Assemblers such as EULER (Pevzner *et al.*, 2001), ALLPATHS (Butler *et al.*, 2008), Velvet (Zerbino and Birney, 2008) and ABySS (Simpson *et al.*, 2009) utilize the *de Brujn* graph approach. The Euler software was developed for Sanger reads (Pevzner *et al.*, 2001) and later modified for short 454 GS20 reads (Chaisson *et al.*, 2004) and Illumina/Solexa reads (Chaisson and Pevzner, 2008). The *de Brujn* graph based SOAPdenovo (Li *et al.*, 2010b) assembler was used for large eukaryotic genome sequences like giant panda (Li *et al.*, 2010a). To make use of increasing read length, SGA (Simpson and Durbin, 2012) and femi (Li, 2012) have been developed based on the string graph approach. Similarly, Forge (DiGuistini *et al.*, 2009), MIRA and PCAP (Huang and Yang, 2005) are long-read assemblers (Forge and MIRA can also handle short reads).

With the availability of genome, transcriptome, molecular markers and bioinformatics tools, genome assisted breeding approaches have advanced for several plants. Draft genome sequences have become available for crop species (Table 1-2) such as rice (Barry, 2001; Goff *et al.*, 2002; Yu *et al.*, 2002), sorghum (Bedell *et al.*, 2005; Paterson *et al.*, 2009), poplar (Tuskan *et al.*, 2006), grape (Ming *et al.*, 2008), medicago (Young *et al.*, 2011), wheat (Ling *et al.*, 2013), maize (Martienssen *et al.*, 2004), barley (Mayer *et al.*, 2012), chickpea (Varshney *et al.*, 2013), tomato (Sato *et al.*, 2012) and soybean (Schmutz *et al.*, 2010).

Table 1-2: Crop reference genome availability.

Crop	Reference	Web/Availability
Rice	Barry, 2001; Goff et al., 2002; Yu et al., 2002	http://www.plantgdb.org/OsGDB/
Sorghum	Bedell et al., 2005; Paterson et al., 2009	http://www.phytozome.net/sorghum
Poplar	Tuskan et al., 2006	http://www.phytozome.net/poplar.php
Grape	Ming et al., 2008	http://www.phytozome.net/grape.php
Medicago	Young et al., 2011	http://www.medicagohapmap.org
Wheat	Ling et al., 2013	http://www.wheatgenome.org
Maize	(Martienssen et al., 2004	http://www.maizegenome.org
Barley	Mayer et al., 2012	http://www.public.iastate.edu/imagefpc/IBSC%20Web page/IBSC%20Template-home.html
Chickpea	(Jain et al., 2013; Varshney et al., 2013	http://www.cicer.info
Soybean	Schmutz et al., 2010	http://www.phytozome.net/soybean
Tomato	Sato et al., 2012	http://solgenomics.net/organism/Solanum_ lycopersicum/genome
Foxtail millet	Zhang et al. 2012	http://www.jgi.doe.gov/sequencing/why/99178.html

1.7 Next-Generation sequence analysis tools

Compared to traditional Sanger sequences, sequences generated from NGS are shorter making downstream analysis more challenging. In addition to short sequence read size, NGS platforms generate terabytes of data increasing the computational

resource requirement for sequencing laboratories. Next generation sequence data can be produced either for genome assembly or downstream analysis such as variant (SNP) and genotype calling, detection of structural variations (SVs), copy number variations (CNVs), genome wide association studies (GWAS) and phylogenic studies.

1.7.1 Quality control (QC)

This is the initial (pre-processing) step to evaluate the quality of raw reads and process them to remove, trim or correct reads that do not meet standards for downstream analysis. Pre-processing includes visualising the data quality, trimming reads and reporting overall sequence properties (Table 1-3). FastQC is a widely used user friendly tool compatible with all major sequencing platforms. In addition to quality assessment and processing, tools like htSeqTools (Planet *et al.*, 2012b) and SolexaQA (Cox *et al.*, 2010a) also has visualisation functionality. A few tools are specific to sequencing platform like FASTX-Toolkit (hannonlab.cshl.edu/fastx_toolkit), PIQA (Martinez-Alcantara *et al.*, 2009) and TileQC (Dolan and Denver, 2008a) specific to the Illumina platform.

Table 1-3: NGS data pre-processing software.

Tool	Reference	Availability
FastQC	-	http://www.bioinformatics.babraham.ac.uk
ContEst	(Cibulskis *et al.*, 2011)	http://www.broadinstitute.org/cancer/cga/contest
SolexaQA	(Cox *et al.*, 2010b)	http://solexaqa.sourceforge.net/
FASTX Toolkit	-	http://hannonlab.cshl.edu/fastx_toolkit/
PIQA	(Martinez-Alcantara *et al.*, 2009)	http://bioinfo.uh.edu/PIQA
htSeqTools	(Planet *et al.*, 2012a)	www.bioconductor.org
TileQC	(Dolan and Denver, 2008b)	http://science.oregonstate.edu/~dolanp/tileqc
Trimommatic	(Bolger *et al.*, 2014)	http://www.usadellab.org/cms/?page=trimmomatic
NGS QC Toolkit	(Patel and Jain, 2012)	http://www.ngs-qc.org/

1.7.2 Alignment tools

Alignment tools are the software that aligns sequence reads to a reference genome based on the base similarity between sequences read and reference bases. Some of the publicly available short-read mapping software (listed in Table 1-4) include SOAP (Li *et al.*, 2008b), MAQ (Li *et al.*, 2008a), Bowtie (Langmead *et al.*, 2009a), BWA (Li and Durbin, 2009; Li and Durbin, 2010), RMAP (Smith *et al.*, 2008), CloudBurst (Schatz, 2009), SHRiMP (Rumble *et al.*, 2009), RazerS (Weese *et al.*, 2012), PerM (Chen *et al.*, 2009b), degemehl (Hoffmann *et al.*, 2009), GenomeMapper (Schneeberger *et al.*, 2009), and BOAT (Zhao *et al.*, 2009). Mapping software designed specifically for sequencing platforms are Eland from Illumina (www.illumina.com), Corona from ABI (McKernan *et al.*, 2009), Reference mapper from 454 (www.454.com). Other software includes ZOOM (Lin *et al.*, 2008), CLC Workbench (www.clcbio.com), Novoalign (www.novocraft.com), Myrialign

(www.savannah.nongnu.org/projects/myrialign) and Mr. Fast (www.mrfast.sf.net). Mapping based variant discovery algorithms include MoDIL (Lee *et al.*, 2009), VariationHunter (Hormozdiari *et al.*, 2009) and BreakDancer (Chen *et al.*, 2009a).

Table 1-4: NGS data alignment software.

Name	Reference	Available
Bowtie	(Langmead *et al.*, 2009b)	http://bowtie-bio.sourceforge.net
BWA	(Li and Durbin, 2009)	http://bio-bwa.sourceforge.net
BLAT	(Kent, 2002)	www.soe.ucsc.edu/~kent/
Geneious	www.geneiousserver.com	www.geneiousserver.com
MAQ	(Li *et al.*, 2008a)	www.maq.sourceforge.net
MOSAIK	(Lee *et al.*, 2014)	http://bioinformatics.bc.edu/marthlab/Mosaik
NovoalignCS	www.novocraft.com	www.novocraft.com
NextGENe	-	http://softgenetics.com/NextGENe.html
NextGenMap	(Sedlazeck *et al.*, 2013)	http://github.com/cibiv/NextGenMap/wiki
SHRiMP	(Rumble *et al.*, 2009)	http://combio.cs.toronto.edu/shrimp
SOAP2	(Li *et al.*, 2009b)	http://sourceforge.net/projects/soap3dp/
SToRM	-	http://bioinfo.lifl.fr/yasss/iedera_solid/storm

1.7.3 Variant discovery

Sequence coverage is an important factor in variant discovery as called mutations need to be supported by several reads. Variant calling aims to determine in which positions there is polymorphism. Some analysis calls polymorphisms between the varieties (i.e variant bases between the multiple individuals) without considering the reference base (Lorenc *et al.*, 2012). SNPs are the most commonly used molecular markers for downstream usefulness in linkage map construction, genetic diversity analysis, association mapping and marker assisted selection (MAS).

1.7.4 Reference genome (features) visualisation tools

Different categories of visualisation tools are available (Table 1-5) including genome browsers that allow users to browse mapped data with additional feature files.

Table 1-5: NGS data visualisation tools.

Tools	User interface	OS	Zoom	Paired end	Long reads	Input format
BamView	Java	Windows/Unix	Partial	Yes	Yes	SAM
Gap5	X11/tk	Unix	Yes	Yes	Yes	Gap5
IGV	Java	Windows/Unix	Yes	Partial	Yes	SAM
MapView	.NET	Windows/Unix	Yes	Yes	No	MVF
Maqview	OpenGL	Unix	Yes	No	No	MAQ
SAMtools	Curses	Windows/Unix		Orphan only	Yes	SAM
Tablet	Java	Windows/Unix	Yes	No	Yes	SAM
Gbrowse	Perl	Web	Yes	Yes	Yes	SAM
Jbrowse	JavaScript/Perl	Web	Yes	Yes	Yes	SAM

1.8 Molecular markers

DNA based markers are more favourable than phenotypic markers as they are highly heritable and not influenced by external factors such as biotic or abiotic stress. Several molecular marker technologies have been developed for genome analysis, assessing the difference between individuals within a species and for association studies. Traditional markers like restriction fragment length polymorphis (RFLP), amplified fragment length polymorphism (AFLP) and cleaved amplified polymorphic sequence (CAPS) and modern markers like single nucleotide polymorphism (SNP) are used for genetic maps (Gaur *et al.*, 2012; Millan *et al.*, 2010; Tar'an *et al.*, 2007). Single Nucleotide Polymorphisms (SNPs) can be rapidly and cheaply identified through bioinformatics and have many uses in genetics, such as the detection of alleles associated with disease, genome mapping, diversity and evolutionary studies (Kale *et al.*, 2015; Lai *et al.*, 2015).

SNPs are the most abundant class of markers present in both plants and animals. The high density of SNP markers makes SNPs an ideal marker of choice. In chickpea, the range of SNP density is between one in 36 bp (Gujaria *et al.*, 2011) to 973 bp (Jhanwar *et al.*, 2012). However, the SNP density is highly dependent on the diversity and number of accessions used in the analysis.

1.8.1.1 SNP prediction tools

SNPs have become the marker of choice in genetic studies. Some software are available with sequencing packages such as CASAVA (consensus assessment of sequence and variation). Some commercial software such as NextGENe, CLC Genomics workbench and free-ware program such as SNPdetector (Zhang *et al.*, 2005), ACCUSA (Frohler and Dieterich, 2010), AGSNP (You *et al.*, 2012), PolyScan (Chen *et al.*, 2007), samtools (http://samtools.sourceforge.net/), GATK (Zhu *et al.*, 2015) and SGSautoSNP (Lorenc *et al.*, 2012). AutoSNP (Batley *et al.*, 2003b; Duran *et al.*, 2009a) was extended as AutoSNPdb (Duran *et al.*, 2009c; Savage *et al.*, 2005) and can now integrate both Sanger and Roche 454 data for SNP discovery. Improved version of autoSNP, SGSautoSNP calls SNPs from Illumina sequence data (Lorenc *et al.*, 2012).

1.9 SNP genotyping

With the recent developments in sequence technologies, several new technologies for large scale SNP genotyping have been developed. Illumina GoldenGate and Infinium genotyping platforms have been widely used in many crops including wheat (Akhunov *et al.*, 2009), rice (Parida *et al.*, 2012), lentil (Sharpe *et al.*, 2013) and chickpea (Deokar *et al.*, 2014). Developments in DNA sequencing technologies have led to several cost-effective SNP discovery and genotyping approaches such as genotyping by sequencing (GBS) (Elshire *et al.*, 2011) skim-based genotyping, restriction site associated DNA (RAD) (Baird *et al.*, 2008) and complexity reduction of polymorphic sequences (CRoPS) (Table 1-6) (van Orsouw *et al.*, 2007)

.

Table 1-6: SNP genotyping platforms.

Genotyping platform	Technology	SNP X sample	Advantages
Illumina Infinium iSelect HD	Fixed array	3,072 - 700K SNPs x 24 samples	Highly multiplexed
Affymetrix Axiom	Fixed array	50K SNPs x 384 samples; 650K SNPs x 96 samples	Highly multiplexed
Douglas Array Tape	Flexible, PCR-based	1 SNP/sample x 76,800 reactions/reel	Ultra high-throughput
Fluidigm Dynamic Arrays	Flexible, PCR-based	96 SNPs x 96 samples; 24 SNPs x 192 samples	High-throughput
RE-based GBS	Genotyping by sequencing	~10K-100K SNPs x 96 or 384 samples	Lots of data relative to cost
Amplicon sequencing	Genotyping by sequencing	Variable (e.g. 20-500 SNPs x 48-384 samples)	Multiple targeted loci at once
skimGBS	Genotyping by sequencing	Resequencing at low coverage	Multiple targeted loci at once

1.10 Summary

Chickpea is an important crop plant feeding a significant fraction of the world population. With the growing population chickpea improvement is required to meet demand. Sequencing technologies have been improved through emerging NGS sequencing platforms. This has laid path to support the completion of reference genome sequences for many important corps, including chickpea. NGS has further expanded to offer approaches for profiling of mRNAs, small RNAs, and variation studies. NGS can accelerate the development of molecular markers. SNPs are regarded as the highest resolution of genetic variations in comparison with other kind of markers. Additionally, for breeders, these genetic molecular markers have become the optimal choice for breeding through marker assisted selection (MAS).

The purpose of this thesis is to establish a new approach to utilise NGS data for reference genome assessment and improvement, identify inter-varietal SNPs, validation and diversity assessment between chickpea cultivars.

As an outcome from this study, an online integrated database and portal was established to provide the reference genome assembly of chickpea *desi* and *kabuli* types, identified genetic variations, markers and genes to assist breeders in designing high-yielding varieties in the face of climate change.

2 Chickpea genome assessment

2.1 Introduction

The sequencing of plant genomes such as *Arabidopsis thaliana* (Kaul *et al.*, 2000), *Oryza sativa* (Matsumoto *et al.*, 2005), *Carica papaya* (Ming *et al.*, 2008), and *Zea mays* (Schnable, 2012) was initiated using first generation capillary sequencing during the last 20 years. Since then many other plant genomes were sequenced with second generation sequencing, including *Fragaria vesca* (Shulaev *et al.*, 2011), *Solanum lycopersicum* (Tomato), *Cicer aritenium*, *Cajanus cajan* (Varshney *et al.*, 2012) and Wheat (Mayer *et al.*, 2014). Sequencing technology improvements have boosted the availability of the number of plant reference genomes. Genome complexity heterozygosity and repeats (Gore *et al.*, 2009) remain challenges for *de novo* assembly of plant genomes. *De novo* assembly can create a highly fragmented assembly rather than contigs/scaffolds covering large chromosomal regions.

It is essential to assess when an assembly is good enough to be useful. Assembling a reference genome requires bioinformatics skills to assess the quality of results at various stages of assembly. There has been a lot of work developing tools to assess the quality of reference genomes.

Chickpea is the second most important grain legume crop in the world. Based on the morphology and nutrition value, chickpea is of two types: small and dark seed coated *desi* and large and creamy coloured *kabuli* types (discussed in Chapter 1). Genome assemblies have recently become available for both *kabuli* (CDC Frontier) (Varshney *et al.*, 2013) and *desi* (Jain *et al.*, 2013) types. These genome assemblies appear to be different, so we studied the differences to assess the quality of both reference genomes.

2.2 Methods

2.2.1 Experimental design

Initially both *desi* and *kabuli* reference genomes were compared to find similarity at the base pair level, followed by assessing the quality of each reference genome using isolated chromosomes sequence reads.

2.2.2 *Desi* and *kabuli* genome comparison

The whole genome statistics were produced with in-house developed scripts '*Congit_stats.pl*', '*fasta_size.pl*' and '*gc_distr.pl*' (developed by Jiri Stiller). SyMAP v4.2 (Soderlund *et al.*, 2011) was employed for pairwise comparison to produce whole genome dot plots comparing of *desi* and *kabuli* chromosomes.

2.2.3 Chromosome isolation and sequencing

Dr. Jaroslav Dolezel from Institute of Experimental Botany, Czech Republic isolated the chromosomes from both chickpea types. Actively growing roots of both chickpea *desi* and *kabuli* types were used for chromosome isolation using flow cytometric analysis and sorting of mitotic chromosomes (Vlacilova *et al.*, 2002). Dr. Jacqueline Batley amplified the isolated chromosomal DNA for further sequencing with Illumina TrueSeq platform to sequence each chromosome individually to produce pair data ranging from 2.4x to 109.4x coverage (Table 2-3).

2.2.4 Genome assessment

SOAPaligner-v2.21 (Li *et al.*, 2009b) was used for mapping individual chromosome sequence reads to *desi* and *kabuli* reference genomes. For high confidence, reads were mapped uniquely to the reference avoiding the reads having multiple hits and only pair mapped reads were considered. A perl script '*soap2heatmap.pl*' was used to convert soap mapped data to numerical coverage data representing the read count at each reference base position of a chromosome/sequence.

perl soap2heatmap.pl –file AchrMapPair.soap –ref Kabuliv1.0.fasta

Above numerical coverage file was converted to Circos compatible format using a *perl script nc2circos.pl*

perl nc2circos.pl –I AchrMapPair.nc –w 5000 –c Ca1

Circos v0.56 (Krzywinski *et al.*, 2009a) was used to produce circular heat-maps to demonstrate the read mapping on all chromosomes.

2.2.5 Contig sequence assessment

Circular heat-maps were produced for assessing the pseudomolecules but it is difficult to produce heat-maps for thousands of contiguous sequences. The mapped

pair read information from session 2.2.4 was used to find the contig sequence corresponding to respective chromosome.

The soap files for each isolated chromosome sequence read mapped data were converted to bam format using a custom python script developed by Michal T Lorence.

Soap2bam.py –s AchrMapPair.soap –f Kabuliv1.0.fasta –r /output/dir/path/

A horizontal coverage of mapped isolated chromosome sequence reads on each contig was recorded using “mpileup” feature of “Samtools” package (Li *et al.*, 2009a). Results from each bam file were collected separately and parsed to compare the horizontal reference sequence captured with each genetic chromosome using a perl script ‘*GroupContigsFromMultipleBam.pl*’.

GroupContigsFromMultipleBam.pl –ref Kabuliv1.0.fasta –out parseContigMapFreq.txt

Similarly, a vertical coverage of mapped genetic chromosome reads on each contig was recorded with Samtools and GroupContigsFromMultipleBam.pl to report and compare with horizontal coverage.

2.3 Results and discussion

2.3.1 *Desi* and *kabuli* reference genome comparison

An initial genome assembly statistic of both *desi* and *kabuli* reference genomes revealed the difference between the assemblies in terms of quality (Table 2-1). Second generation sequence data was used to build both *desi* and *kabuli* reference genome assemblies. The *kabuli* assembly was constructed from Illumina data, while the *desi* assembly was based on a hybrid approach by combining Roche/454 and Illumina data. The *kabuli* assembly captured 532 Mbp (with scaffolds greater than 1000 bp) of sequence while *desi* captured 456 Mbp having contigs with a minimum size of 200bp. The *desi* genome was more fragmented, with a total of 32,935 scaffolds greater than 1Kb and an N50 of 106 Kbp, compared to 7,163 scaffolds and an N50 of 39,989 Kbp for *kabuli*.

Table 2-1: Released draft genome stats.

Features	***Desi* draft**	***Kabuli* draft**
Total assembly size (Mb)	456 (52.5%)^	532 (60.3%)*
Number of scaffolds	32,935	7,163
Minimum reported scaffold length (bp)	1,000	1,000
Maximum scaffold length (Kbp)	23,376	59,460
Average scaffold length (bp)	13,857	74,311
N50 length (Kb)	106	39,989
GC content (%)	25.63	30.80%
Genome captured in pseudomolecule (Kbp)	124,386 (14.33 %)^	347,247 (39.37%)*
Protein coding genes	27,571	28,269
Average gene length (bp)	3,122	3,055
Average coding sequence length (bp)	962	1,166

^ Considering 1C = 868 Mbp (Table 2-1)

***** Considering 1C = 882 Mbp (Table 2-1)

Pairwise comparison of each of the pseudomolecules from the two assemblies revealed numerous structural variations (Figure 2-1). These differences include both long and short regions where the orientations of the sequence differed. There were differences in the position of regions within pseudomolecules, for example, first half of *desi* pseudomolecule Ca5 is inverted and matches the centre of *kabuli* pseudomolecule Ca5. Interestingly, we observed several large regions of similarity between unrelated pseudomolecules. These include *desi* pseudomolecule Ca8 matching a region at the start of *kabuli* pseudomolecule Ca7, while *kabuli* pseudomolecule Ca8 matches the last third of *desi* pseudomolecule Ca3. A large portion of *kabuli* pseudomolecule Ca6 matched the second half of *desi* pseudomolecule Ca2. These differences suggest mis-assembly of one or both draft genome assemblies.

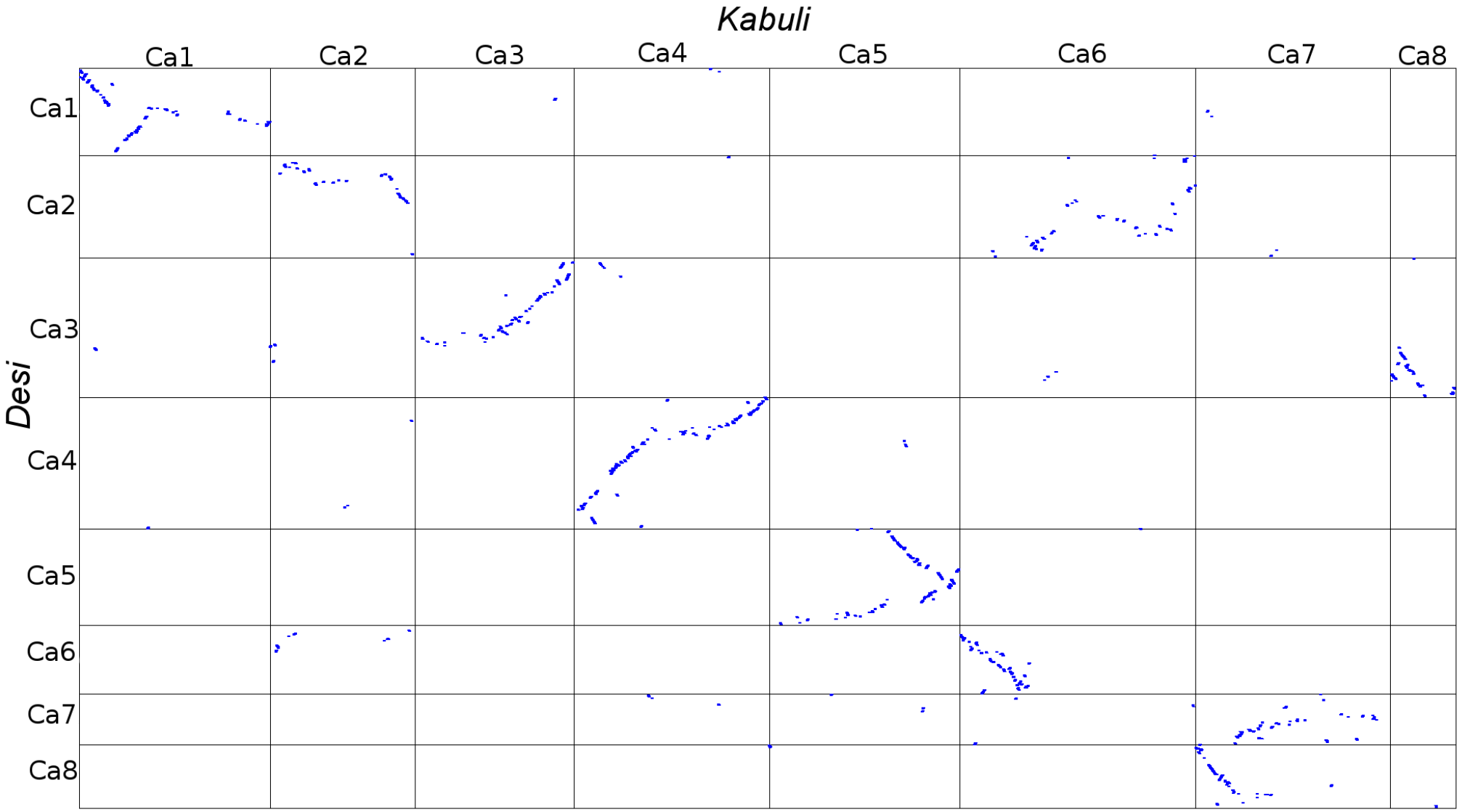


Figure 2-1: Dot plot matrix of a comparison of the *kabuli* and *desi* draft chromosome assemblies.

2.3.2 Isolation and sequencing of chickpea chromosomes

To assess and validate the assembled pseudomolecules from the two genome assemblies, our collaborator, Dr. Jaroslav Dolezel from the Institute of Experimental Botany, Czech Republic isolated and individual chromosomes from both *kabuli* and *desi* varieties of chickpea. All the chromosomes were sorted using flow cytometry method from the sequenced reference *kabuli* 'CDC Frontier', with chromosomes D and E (Ca4 and Ca7) sorted together as a group, while chromosomes A, B and H were sorted from the sequenced reference *desi* 'ICC 4958'. Dr. Jacqueline Batley's group amplified the DNA from these isolated chromosomes to produce samples suitable for sequencing using Illumina technology (Table 2-2). Based on the microscopic observations, all chromosome isolates could be sorted at high purity from both genotypes.

Table 2-2: Isolated chromosome sequence data.

Chromosome	Pair reads	Coverage
Desi		
A (Ca5)	69,857,714	84.7x
B (Ca3)	75,054,367	109.4x
H (Ca8)	14,248,627	45.5x
Kabuli		
A (Ca5)	2,067,921	5.9x
B (Ca3)	1,604,500	5.4x
C (Ca6)	3,906,516	17.3x
D/E (Ca7/Ca4)	6,641,561	16.5x
F (Ca2)	3,263,085	17.6x
G (Ca1)	4,030,401	24.4x
H (Ca8)	851,604	2.4x

2.3.3 Estimation of molecular sizes of chickpea chromosomes

Dr. Jaroslav Dolezel also helped in estimating the molecular size of individual chromosomes based on relative chromosome lengths at mitotic metaphase. Dr. Jacqueline Batley and Dr. Satomi Hayashi sequenced the isolated chromosome DNA and Dr. Rajeev K Varshney and Dr. Rachit K Saxena provided the linkage maps to associate with isolated chromosomes. The differences in chromosome size support to estimate the chromosome size with in genotype and also between *desi* and *kabuli* chromosomes as large as 10 Mbp for chromosomes A and B and as small as several hundred Kbp for chromosome F (Table 2-3). Although the differences between the two types of chickpea may be recognized in part to differences in chromatin condensation, they correspond well to differences between flow karyotypes of *desi* and *kabuli* and differences in chromosome peak (chromosome isolation peak) positions (Ruperao *et al.*, 2014b). For example, chromosomes F and G of *desi* 'ICC 4958' differ by about 7 Mbp (7%), and their peaks cannot be discriminated based on flow karyotype. In *kabuli* 'CDC Frontier', the two chromosomes differ by about 10 Mbp (11%) and can be discriminated.

Table 2-3: Chromosome size estimation.

Cicer arietinum				**Linkage group (Thudi *et al.*, 2011)**	**Pseudomolecule**	**Relative chromosome length [%]**		**Molecular chromosome size [Mbp]**	
Desi		***Kabuli***							
Peak	**Chromosome**	**Peak**	**Chromosome**			***Desi***	***Kabuli***	***Desi***	***Kabuli***
I	H	I	H	LG 8	Ca8	7.2	7.8	62.5	68.80
II	G	II	G	LG 2	Ca2	9.9	9.4	85.93	82.91
	F	III	F	LG 1	Ca1	10.7	10.5	92.88	92.61
III	E	IV	E	LG 4	Ca4	11.5	11.1	99.82	97.90
IV	D		D	LG 7	Ca7	12.6	11.8	109.37	104.01
	C	V	C	LG 6	Ca6	13.2	12.8	114.58	112.90
V	B	VI	B	LG 3	Ca3	15.8	16.7	137.14	147.29
VI	A	VII	A	LG 5	Ca5	19.0	19.8	164.92	174.64

2.3.4 Comparison of pseudomolecule assemblies

A major portion of the *kabuli* assembly sequence was integrated into pseudomolecules (347 247 Kbp) compared to *desi* (124 386 Kbp). The size of the *kabuli* pseudomolecule (39.37%) is much larger than *desi* pseudomolecule (14.33%) (Table 2-4). Individual pseudomolecules differed in size and their predicted chromosome size is shown in Table 2-4. Initial chromosomes sequence mapping results demonstrate that, chromosome A, B and H pseudomolecules of *kabuli* type were assembled to an average of only about 26% of their predicted size. The smaller than expected pseudomolecule size of these three chromosomes could be explained by the presence of satellite CaRep2 on chromosomes A and B, satellite CaSat2 on chromosomes A and H, and the 45S rDNA locus on chromosome A (Zatloukalová et al., 2011). These differences could be due to highly repetitive regions, which are likely to collapse into shorter regions during de Bruijn graph-based whole-genome assembly.

Table 2-4: Pseudomolecule size and percentage of predicted chromosome size.

Pseudomolecule (Chromosome)	***Desi***	***Kabuli***
Ca1 (F)	14,791,696 (15.9%)	48,359,943 (52.2%)
Ca2 (G)	17,304,114 (20.1%)	36,634,854 (44.1%)
Ca3 (B)	23,376,002 (17.0%)	39,989,001 (27.1%)
Ca4 (E)	22,093,647 (22.1%)	49,191,682 (50.2%)
Ca5 (A)	16,301,343 (9.8%)	48,169,137 (27.5%)
Ca6 (C)	11,482,212 (10%)	59,463,898 (52.6%)
Ca7 (D)	8,461,617 (7.7%)	48,961,560 (47.0%)
Ca8 (H)	10,574,966 (16.9%)	16,477,302 (23.9%)

Mapping *kabuli* isolated chromosome sequence reads to *kabuli* reference genome assembly demonstrated the correspondence with pseudomolecule. Chromosome F and G reads mapped to Ca2 and Ca1, respectively, which is opposite to earlier

assignments (Millan et al., 2010; Thudi et al., 2011; Zatloukalová et al., 2011). Inspection of the read mapping density (Figure 2-2) suggested that chromosome F data included sequences specific for pseudomolecule G and vice versa. Chromosome C and the chromosome D/E group also shared contamination (with other chromosomes), while chromosomes A, B and H demonstrated a greater purity.

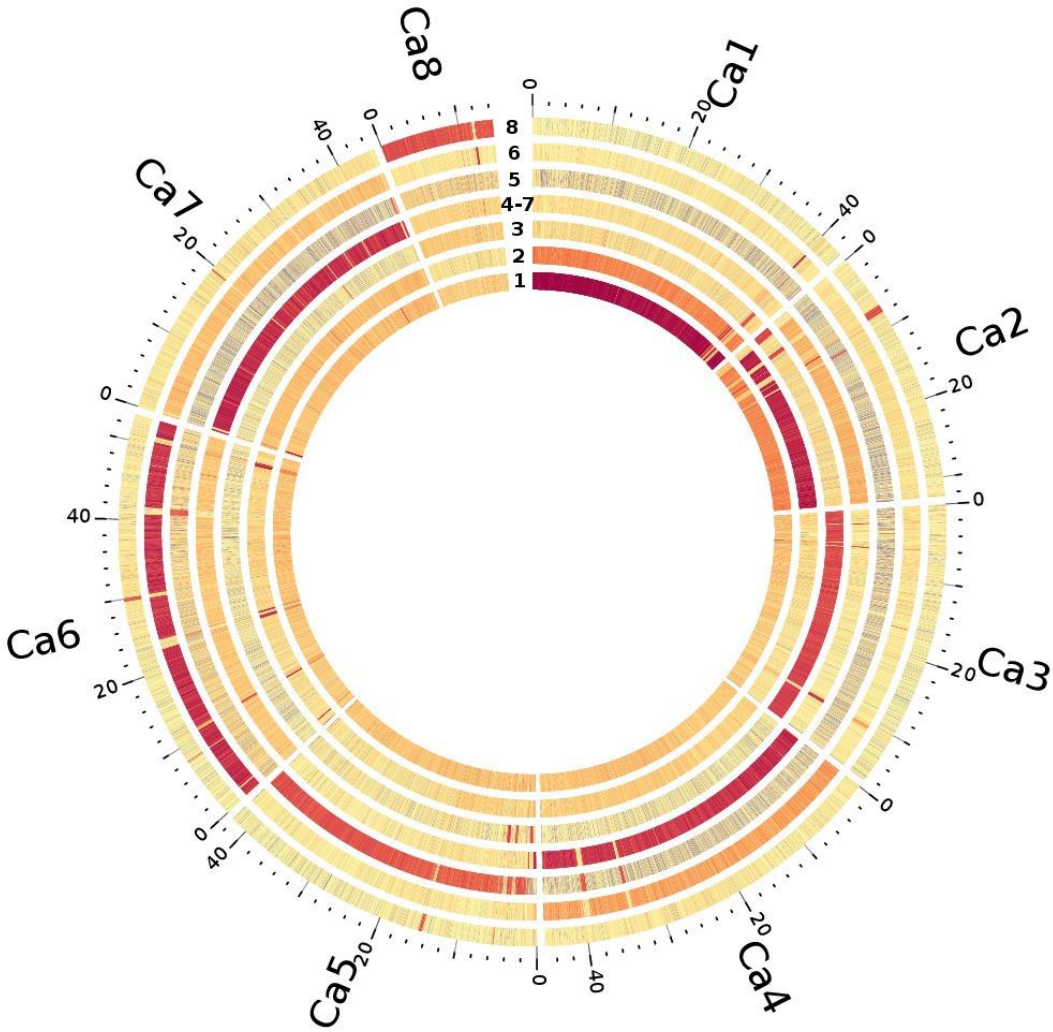


Figure 2-2: *Kabuli* reference genome assessment. Each circular heatmap represents *kabuli* reference genome assessment with each isolated chromosome sequence reads as below

1) 1 = G chromosome (Ca1)
2) 2 = F chromosome (Ca2)

3) 3 = B chromosome (Ca3)
4) 4-7 = DE chromosome (Ca4,7)
5) 5 = A chromosome (Ca5)
6) 6 = C chromosome (Ca6)
7) 8 = H chromosome (Ca8)

In addition to the chromosome sequence paired reads mapping to corresponding chromosomes, reads mapped on non-corresponding chromosomes forming high density mapped reads (showing mis-assemblies) (Figure 2-2). For example, a region from 40,141,642 to 40,436,753 bp on pseudomolecule Ca1 had very few reads mapping from the corresponding isolated chromosome G. Interestingly, the same region had high density of chromosome C reads indicating the mis-assembly region belongs to Ca6. Numerous such small non-corresponding chromosome sequence read mapping patterns were observed and interpreted to be misplaced regions. In total, we observed 46 regions ranging in size from 57 to 1371 Kbp and representing 16,164 Kbp (3.0%) of the pseudomolecule assemblies that were placed into the wrong pseudomolecule (Table 2-5). Ca8 has only a single region of 341 Kbp which should be located on pseudomolecule Ca6 (Figure 2-3). In contrast, pseudomolecule Ca6 contains 11 sequence regions which should be relocated onto other pseudomolecules.

Table 2-5: *Kabuli* reference genome miss-assemblies origins.

Pseudomolecule	Start	End	Length	Chromosome
Ca1	17,709,355	17,768,444	59,089	H (Ca8)
Ca1	39,419,875	39,639,265	219,390	F (Ca2)
Ca1	39,875,495	40,137,184	261,689	F (Ca2)
Ca1	40,141,642	40,436,753	295,111	C (Ca6)
Ca1	40,737,888	41,342,099	604,211	B (Ca3)
Ca2	1	1,370,632	1,370,631	B (Ca3)
Ca2	4,000,604	4,701,466	700,862	B (Ca3)
Ca2	5,875,970	5,981,305	105,335	G (Ca1)
Ca2	6,978,984	7,977,546	998,562	H (Ca8)
Ca2	8,329,465	8,839,803	510,338	D/E (Ca4/Ca7)
Ca2	9,713,165	10,056,176	343,011	A (Ca5)
Ca3	2,147,981	2,291,454	143,473	D/E (Ca4/Ca7)
Ca3	5,222,428	5,315,148	92,720	G (Ca1)
Ca3	5,817,530	5,985,956	168,426	D/E (Ca4/Ca7)
Ca3	16,652,391	16,736,790	84,399	C (Ca6)
Ca4	33,548,471	34,015,816	467,345	A (Ca5)
Ca4	39,258,943	40,092,009	833,066	A (Ca5)
Ca5	1	475,756	475,755	D/E (Ca4/Ca7)
Ca5	479,469	967,381	487,912	B (Ca3)
Ca5	1,056,592	1,302,990	246,398	D/E (Ca4/Ca7)
Ca5	1,443,073	1,575,871	132,798	B (Ca3)
Ca5	3,057,844	3,436,790	378,946	B (Ca3)
Ca5	4,300,510	4,897,203	596,693	B (Ca3)
Ca5	5,459,276	5,583,037	123,761	H (Ca8)
Ca5	14,570,063	14,984,575	414,512	H (Ca8)
Ca6	1	95,085	95,084	F (Ca2)
Ca6	1,273,532	1,471,783	198,251	F (Ca2)
Ca6	8,091,437	8,171,394	79,957	H (Ca8)
Ca6	10,834,481	11,048,871	214,390	F (Ca2)
Ca6	11,049,103	11,286,027	236,924	D/E (Ca4/Ca7)
Ca6	22,607,989	23,187,652	579,663	F (Ca2)
Ca6	23,191,730	23,605,994	414,264	A (Ca5)
Ca6	23,607,585	23,929,615	322,030	F (Ca2)
Ca6	29,686,973	30,164,801	477,828	H (Ca8)
Ca6	40,809,786	41,720,980	911,194	A (Ca5)
Ca6	50,526,024	51,189,107	663,083	F (Ca2)
Ca7	259,820	579,136	319,316	G (Ca1)
Ca7	10,581,321	10,638,294	56,973	G (Ca1)
Ca7	19,431,822	19,639,120	207,298	H (Ca8)
Ca7	24,424,554	24,528,692	104,138	F (Ca2)
Ca7	31,258,312	31,438,623	180,311	B (Ca3)
Ca7	34,418,700	34,570,247	151,547	A (Ca5)
Ca7	37,738,053	37,860,082	122,029	G (Ca1)
Ca7	37,864,185	37,939,907	75,722	A (Ca5)
Ca7	44,316,766	44,615,144	298,378	A (Ca5)
Ca8	11,734,308	12,075,396	341,088	C (Ca6)

Some mis-assemblies appeared during scaffolding and others appeared within contigs as chimeric contig assembly. Many of the mis-assembled regions were also flanked by highly repetitive retrotransposon sequences, although there was no clear correlation between the presence of these sequences and the type of mis-assembly.

This chromosomal genomics approach connected with next generation sequencing has limited the genome complexity to chromosome level which supports the identification of genome mis-assembles at high resolution. This resolution will also facilitate the relocation of these regions into their respective pseudomolecule. However, this approach cannot identify intra-chromosomal mis-assembly or mis-assemblies between chromosomes which cannot be separated physically by flow sorting. In case of pool isolation of chromosome D and E, it is difficult to identify the mis-assembled regions specifically belonging to chromosome E (Ca4) or chromosome D (Ca7).

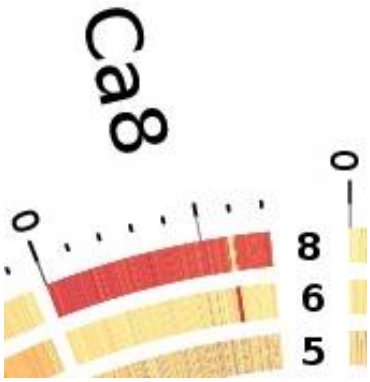


Figure 2-3 : Circos heat map demonstrating miss-assembly in *kabuli* reference.

2.3.5 *Kabuli* contig sequence assessment

The 7,163 *kabuli* unplaced contigs were assessed to identify the correspondence with respective isolated chromosome. On mapping individual (isolated) chromosome sequence data to the *kabuli* reference genome (including all extra contigs), the paired reads also mapped to same contigs/scaffold indicating that these contigs/scaffolds correspond to the respective chromosomes. For each contig/scaffold, the horizontal coverage of each mapped isolated chromosome sequence was assessed to compare and assign to the respective chromosome. To add more confidence, a vertical coverage (read count) of each isolated chromosome

sequence reads mapping to each contig/scaffold were compared and validated with above assignment.

A total of 4,852 contigs/scaffolds spanning 177.47 Mb of sequence were assigned to respective chromosomes. A total of 1,206 contigs with sequence length of 39 Mbp assigned to A(Ca5) chromosome and only 488 contigs (but more contigs than H(Ca8) chromosome assigning 457 contigs) spanning 9.19 Mb sequence assigned to G(Ca1) chromosome (Table 2-6). This additional information of extra contigs can be used to improve and increase the chromosome sequence content.

Table 2-6: *Kabuli* contig sequence analysis.

Chromosome	Number of sequences	Sequence length
A(Ca5)	1,206	39,091,098
B(Ca3)	738	38,588,828
C(Ca6)	588	20,017,227
DE(Ca4,7)	877	36,850,083
F(Ca2)	498	20,460,868
G(Ca1)	488	9,191,325
H(Ca8)	457	13,274,427
Total	4,852	177,473,856

2.3.6 *Desi* reference genome assessment

To examine the differences between *desi* and *kabuli* draft genome sequences that reflect true structural genome variation or pseudomolecule mis-assembly, isolated and sequenced chromosomes A, B and H reads from *desi* type were mapped together with the related *kabuli* chromosome-specific reads to the *desi* reference genome (Figure 2-4). Both *desi* and *kabuli* isolated chromosomes demonstrated almost identical mapping patterns on the pseudomolecules suggesting that both genomes have similar genome sequence. Interestingly, the chromosome B (Ca3) reads from *kabuli* and *desi* only matched the first half of *desi* pseudomolecule Ca3. The remaining half of Ca3 mapped with sequence reads from isolated chromosome H (Ca8), suggesting that the observed differences between the *desi* and *kabuli* reference genome assemblies are not due to structural genome differences but due to mis-assembly of the *desi* reference genome.

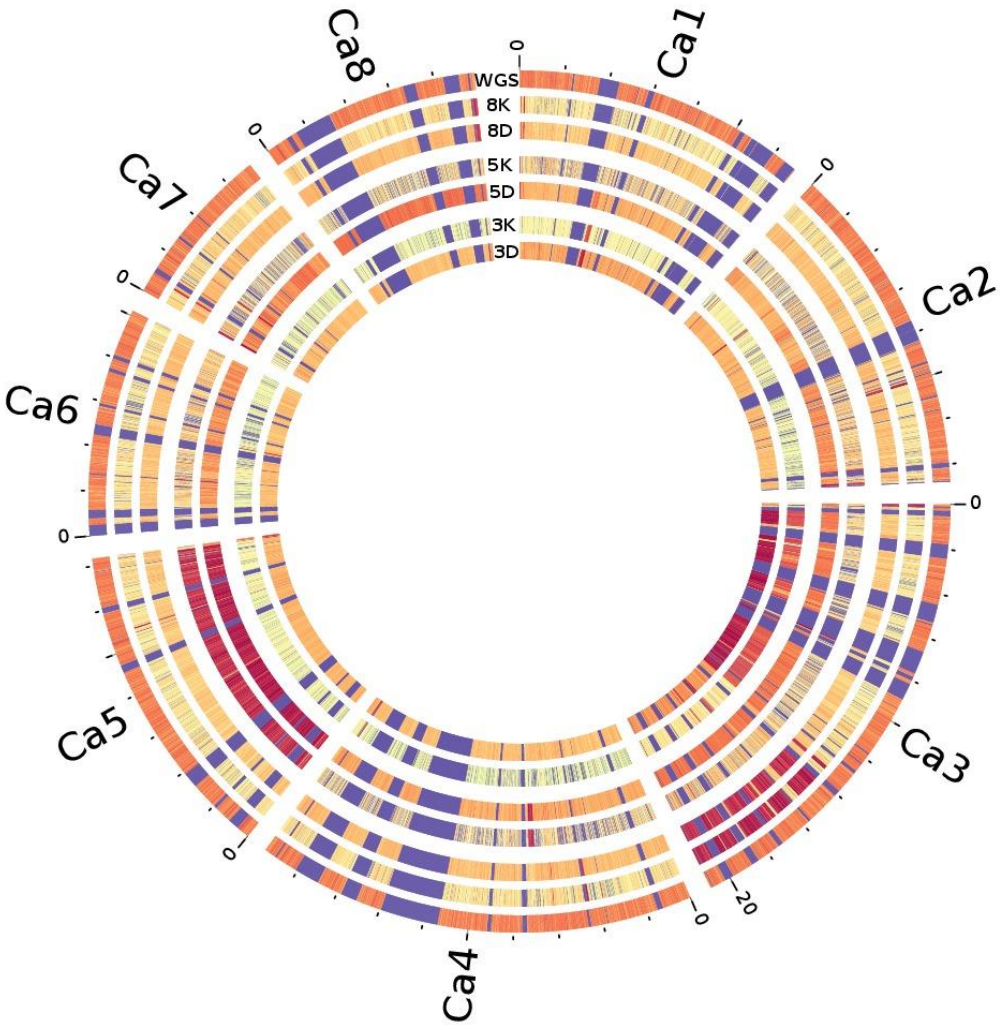


Figure 2-4: Circos heat map illustrating *desi* genome assessment. Assessment with both *desi* (D) and *kabuli* (K) chromosomes 3, 5, 8 and whole genome sequence data (WGS).

There were some regions of the *desi* pseudomolecules where no reads mapped. We investigated these regions by mapping *desi* whole-genome sequence data to the *desi* pseudomolecules (Figure 2-4) and found no reads mapped to these regions. This concluded that these regions of the *desi* reference pseudomolecules do not reflect the physical content of the *desi* genome. Extraction of the sequence for these regions and (BLAST) comparison with the Swiss-Prot gene database failed to identify significant genes, again suggesting that these regions are not true genome sequences.

2.3.7 *Desi* unplaced contig sequence assessment

Desi contiguous sequence assessment was performed with *kabuli* isolated chromosome sequence data. Both vertical and horizontal coverage of mapped reads on each extra contig were recorded.

Based on the above analysis, a total of 106,816 contigs spanning 370 Mb of sequence has revealed the correspondence with respective isolated chromosome data. The chromosome DE(Ca4,7) has the biggest number of unplaced contigs spanning 90.6Mb of sequence (Table 2-7).

Table 2-7: *Desi* contiguous sequence analysis.

Chromosome	Number of sequences	Sequence length
A(Ca5)	19,777	63,381,859
B(Ca3)	15,950	57,764,518
C(Ca6)	14,375	54,542,218
DE(Ca4,7)	20,972	90,635,403
F(Ca2)	12,497	43,207,629
G(Ca1)	12,261	39,146,091
H(Ca8)	10,984	21,623,569
Total	106,816	370,301,287

2.4 Conclusion

Assembly validation with physical anchoring of genetically mapped markers may be prone to errors, so a new approach (chromosomal genomics) was introduced to validate the reference genome assemblies. The possibility of isolating individual chromosomes combined with next generation sequencing permits the validation of genome assemblies at the chromosome level. Using this approach, both *desi* and *kabuli* reference genome quality was assessed and mis-assemblies were identified.

3 Constructing improved Chickpea Genome Assemblies using skimGBS

3.1 Introduction

The chickpea genome project began in 2012 with two scientific groups working on each chickpea type (*desi* and *kabuli*). International Chickpea Genetics and Genomics Consortium (ICGGC) first published the chickpea *kabuli* (CDC Frontier) genome in early 2013 and in the same year, National Institute of Plant Genomics Research (NIPGR), India published the *desi* (ICC4958) draft reference genome.

Chickpea genome assessment with isolated chromosomes sequence data enabled to assess both *kabuli* (Varshney *et al.*, 2013) and *desi* (Jain *et al.*, 2013) genome qualities and also able to locate the wrongly placed sequence regions (Ruperao *et al.*, 2014a), indicating there is a need to improve both genomes (Chapter 2).

The inaccurate genetic information of cultivated chickpea varieties weakens the efforts of marker-assisted crop improvement and production of elite cultivars with durable stress-resistance by modern breeding methods. The better understanding of proper location, variation and function of individual genes will enable to study the evolutionary relationships revealing previously unknown regulatory mechanisms.

Following the release of both CDC Frontier (*kabuli* type) chickpea reference genome by the (ICGGC) and ICC 4958 (drought tolerant *desi* type) from NIPGR (India), we assembled an improved version using skim-based genotyping using a RIL population with the wildtype and a domesticated cultivar as parents (ICC 4958 x PI489777). The released *kabuli* v1.0 spans 532.2Mb of which chromosomes cover 347.2 Mb of sequence and 7,163 unplaced scaffolds. Similarly, the *desi* genome spans 519.8Mb of assembly size out of which 124Mb of sequences was integrated into pseudomolecules and remaining 395Mb of unplaced sequence.

This study has developed a novel method combining the chromosomal genomics method with skimGBS method to build a high quality of draft reference (*desi* and *kabuli*) genome. SkimGBS method calls genotypes from each population individual

using parental SNPs. One of the applications of calling genotypes is mapping the frequency and location of recombination events. Genetic recombination is one of the key mechanisms underlying the evolution of plants (Schnable *et al.*, 1998). It is one of the major sources of genetic variation by forming novel allelic combinations. Both reciprocal (crossover) and non-reciprocal (non-crossover) exchange of genetic material between chromosomes are initiated during meiosis. Recombination events in plants have been genetically mapped using molecular markers (Farkhari 2011, Yao 2002). Based on recombination patterns in population individuals, a novel method was implemented to improve genome assembly.

With the improved assembly, the gene models were predicated using a hybrid pipeline which integrates both *ab initio* and homology based gene modelling.

3.2 Methods

3.2.1 Population sequence data

A total of 5.9 and 7.2 Gbp of whole genome sequence data were generated from *C. arietinum* cultivar ICC4958 (*desi*) and *C. reticulatum* PI489777 (wild-type) which represent an estimated coverage of 7.9× and 9.7× respectively. Chickpea population ICC4958 x PI489777 was chosen to improve the *kabuli* draft reference genome. A total of 20.9 Gbp of Illumina paired read sequence data were generated for 46 progeny individuals of ICC4958 x PI489777 population with coverage range between 0.13× and 1.54×, with an average of 0.58× (Appendix 1).

Prof. Jacqueline Batley's group provided the parental WGS and skim sequenced population data.

3.2.2 SkimGBS analysis

Initially, parental reads (of ICC4958 and PI489777) were mapped to the reference genome uniquely (*kabuli* reference genome) using SOAP2aligner (Li *et al.*, 2009b) (-r 0, reads map non-repetitively). Only reads aligned in pairs were used in downstream analysis. Parental SNPs were called using SGSautoSNP (Lorenc *et al.*, 2012) and genotypes were called by mapping the progeny reads (ICC4958xPI489777) back to same reference.

An in-house developed perl script '*snp_genotyping_all.pl*' (appliedbioinformatics.com.au/index.php/SkimGBS) (Bayer *et al.*, 2015b) compares the progeny read alignments with parental genotypes to assign genotypes.

perl snp_genotyping_all.pl –b bamfile.bam –gff SNP.gff –off Refoffset.gff –f Ref.fasta –o outdir –c1 A –c2 B –h AB –s 1 >outdir/Ca1.map

SNPs that had similar called genotype from all individuals are removed with a python script '*removeMonomorphicSNPs.py*' (developed by Philipp E Bayer).

removeMonomorphicSNPs.py Ca1.map Ca1.dat

With low coverage sequence data from population individuals may not be possible to call genotype allele for all SNPs. To overcome this challenge, the genotype gaps with both adjacent alleles representing the same parental genotype were imputed. Imputation was performed based on adjacent genotypes using '*imputeFlapjackSNPs.py*' (developed by Philipp E Bayer).

imputeFlapjackSNPs.py Ca1.dat

3.2.3 Recombination events

Using the above called genotypes (imputed) data for ICC4958 X PI489777 population, all recombination were recorded. The recombinations were categorised based on their lengths, ranging between 20bp-10kb defined as gene conversions and longer than 10kb defined as crossovers. The overlapping crossovers and gene conversions are filtered using *fuzzyRecombinationfilter.py*. The distribution of recombination events was plotted using Python v2.7.

3.2.4 Genome improvement

Initially, both genome assemby quality were assessed with isolated chromosome data to identify the potential mis-assembly co-ordinates (Chapter 2). Since *desi* reference genome is of poor quality, a synteny based assembly using the *kabuli* draft genome as reference was assembled as shown in Figure 3-1.

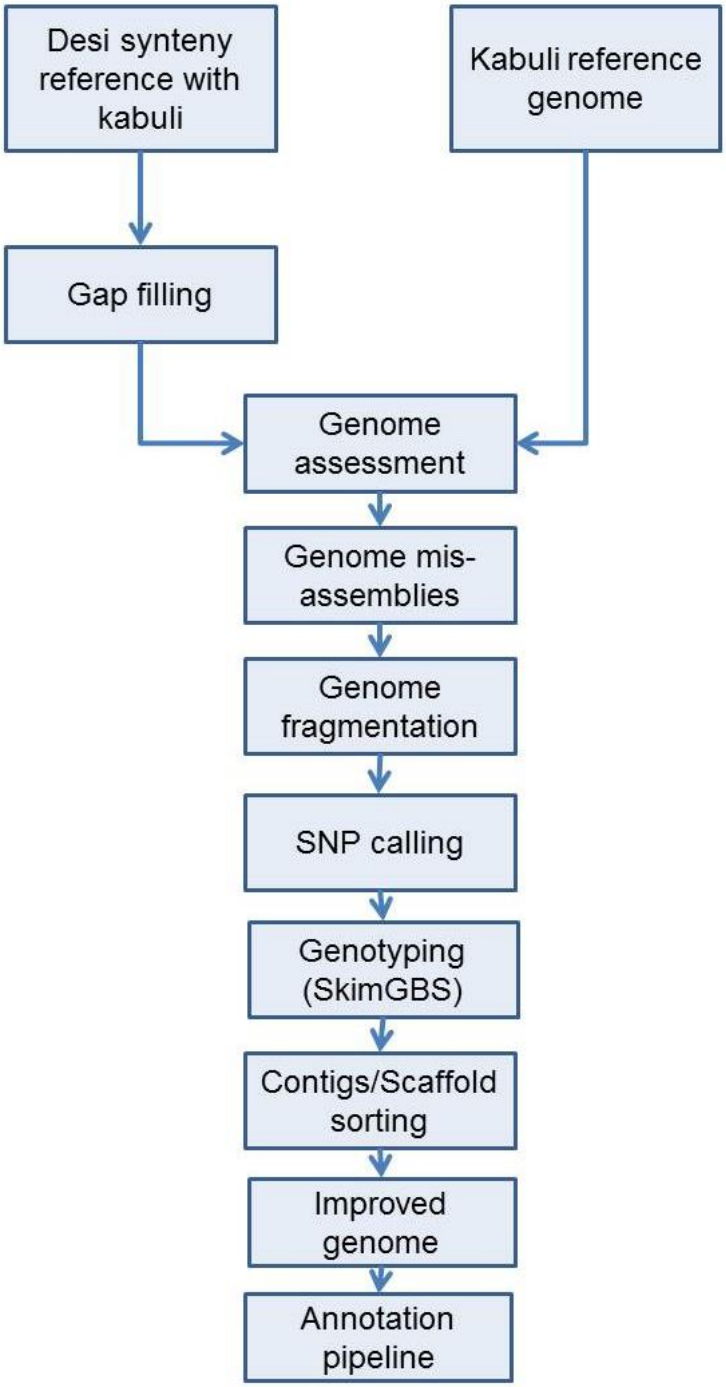


Figure 3-1: Schematic work flow of genome improvement.

Kabuli genome was fragmented at mis-assembly boundaries using *retrieve_subseq4mfasta.pl* and parental reads (ICC4958 and PI489777) were mapped with SOAP2 (Chen *et al.*, 2009a) to call SNPs using SGSautoSNP (Lorenc *et al.*, 2012). All 46 population individuals reads were mapped to fragmented reference with SOAP2aligner (Li *et al.*, 2009b) as described above. Genotypes were

called from each individual using the skimGBS pipeline (appliedbioinformatics.com.au/index.php/SkimGBS) (Bayer *et al.*, 2015a). Based on recombination patterns all unplaced contigs were sorted to construct new pseudomolecules with contigPlacer.

ContigPlacer (https://github.com/philippbayer/contigplacer) is a program that places unplaced contigs around placed contigs by comparing SNPs located on these contigs with each other. It only works with unplaced contigs on which SNPs are located.

First, contigPlacer constructs a representative block of SNPs using the SNPs on each contig (Figure 3-2). These represent the majority of alleles of these SNPs for each individual. For example, three SNPs are present on contig 1, with the first individual always has allele A, the second individual has mostly A alleles with a few B alelles, and the third individual having only B alleles. The representative block is made by considering a major allele for each of these individuals, so it would be A for the first individual and second individual, and B for the third individual (Figure 3-2).

Then, the representative blocks of all unplaced contigs are compared with the placed contigs representative blocks using a penalized Hamming distance (score). Based on the similarity or difference between representative block comparisons, a set of rules were framed to score. For each comparison of two representative blocks each individual's alleles are compared and following rules applied.

a) If an individual's alleles are different, 1 is added to the score.
b) If one of the individual allele is missing, 0.5 is added to the score.
c) If both are missing, 0.75 is added to the score.
d) If both alleles are not missing and identical, 0 is added to the score.

For each representative block (of an unplaced contig), the two placed contigs with the lowest score ("partner-contigs") are retained. If two placed contigs corresponds to different pseudomolecule, then unplaced contig is not placed.

Next, each unplaced contig is placed after the contig with the lowest Hamming distance. If several former unplaced contigs are placed after the contig, they are sorted by increasing score.

Lastly, contigPlacer tries to find the orientation of the contig by merging representative blocks of the last and the first five SNPs located on the unplaced contig. This step is performed only with contigs on which ten or more SNPs are located. These two representative blocks are compared to the representative block of the partner-contig (Figure 3-3). If the representative block of the last five SNPs has a lower Hamming distance to the partner contig than the representative block of the first five SNPs, then the contig is reverse-complemented. In all other cases the original orientation is kept.

contigPlacer parse gff3 files detailing the start- and end coordinates of each contig, and a Flapjack dat and map file detailing the positions and the alleles of all SNPs. It produces a list of unplaced contigs and, for each pseudo-chromosome, a new gff3 file detailing the placement of newly placed contig along with the placement of all already placed contigs.

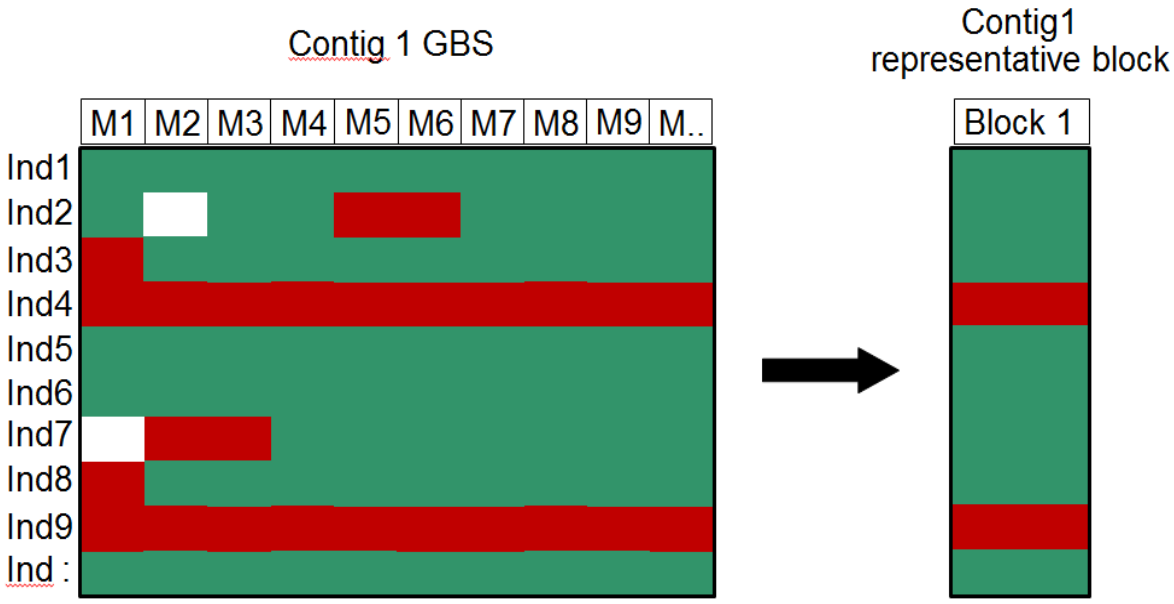


Figure 3-2: Creating a representative block for each contig sequence.

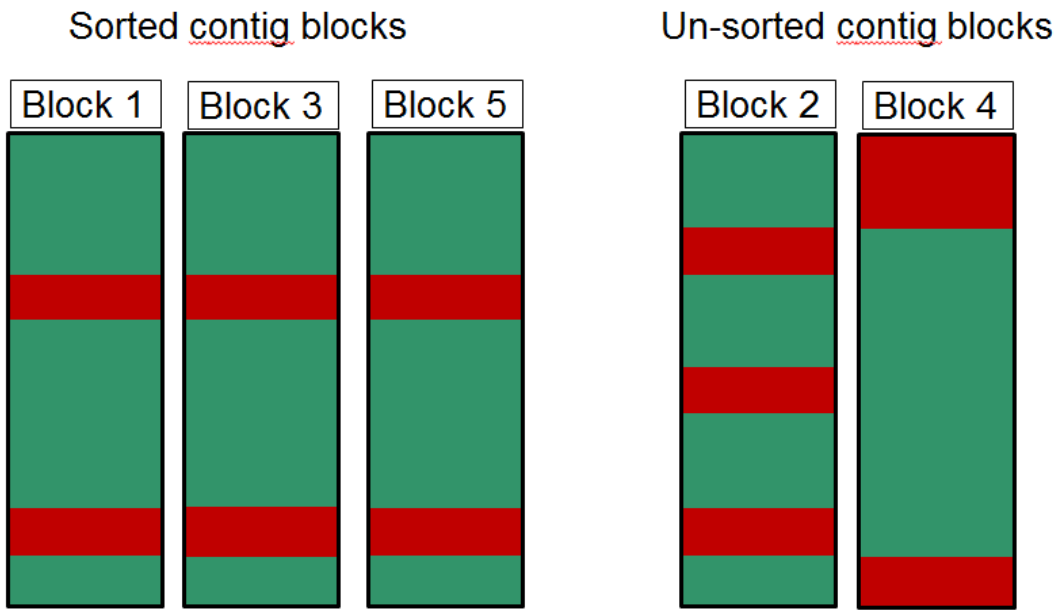


Figure 3-3: Comparison of representative blocks and sorting respective contig sequences.

Based on the list of all sorted contig ids, a multi-fasta file was created using '*retrieve_fastas.pl*' and pseudomolecules were created with '*multiple_to_single_fasta_v2.5.py*' bridging contigs with 100Ns. Finally sequence flanking with long Ns (from released version) were replaced with 100Ns to make a clean improved assembly.

For *desi* reference genome, clones were removed from the whole genome Illumina sequence data of *desi* (ICC4958) from SRA SRR500634 using '*remove_possible_clones.pl*'. Reads were cleaned using '*trimConverter.py*' (base-pairs with Phred quality scores below 20 were removed and reads shorter than 75 bp were excluded). The remaining filtered reads were mapped to *kabuli* (Varshney *et al.*, 2013) genome using SOAP2 (Chen *et al.*, 2009a) and conserved chromosomes sequence was build using mpileup program in Samtools (Li *et al.*, 2009a). GapFiller (Boetzer and Pirovano, 2012) program was used to fill gaps and unused reads were assembled with Velvet (Zerbino and Birney, 2008) assembler. The newly assembled contigs were added to reference genome as extracontigs. This intermediate reference genome was assessed with isolated chromosomes sequence data to identify and fragment the sequence at mis-assembly regions. The fragmented

sequences together with extra contigs were BLAST search against (improved) *kabuli* reference genome to find the synteny contigs corresponding to each chromosome. The synteny contigs were mapped to corresponding *kabuli* chromosome with LASTZ (Harris *et al.*, 2010) and the output was parsed with '*LASTZSorter.py*' to sort and orient *desi* contig sequences to create pseudomolecules.

3.2.5 *Desi* and *kabuli* genome comparison

Whole genome statistics were produced with in-house developed scripts '*Contig_stats.pl', 'fasta_size.pl*' and '*gc_distr.pl*'. SyMAP v4.2 (Soderlund *et al.*, 2011) was employed for pairwise comparison of *desi* and *kabuli* reference genomes, producing whole genome dot plots pairwise comparison of *desi* and *kabuli* chromosomes.

3.2.6 Gene predictions

The predictions of protein coding genes from the *desi* and *kabuli* genomes involved *ab initio* and homology based approaches using Maker (Cantarel *et al.*, 2008) software. Initially, the conserved genes were identified with Cegma (Parra *et al.*, 2007), and for *ab initio* prediction, Augustus (Stanke *et al.*, 2006) trained Arabidopsis genes and Hidden Markov Model based gene finders such as SNAP (Korf, 2004) and GeneMark (Lukashin and Borodovsky, 1998) were used. For protein homology evidence, 32,097 chickpea proteins downloaded from GenBank were used. In addition to 45,038 chickpea ESTs, 269,501 Medicago ESTs and mRNA-seq (SRR491157) from chickpea as supporting EST evidence (from GenBank) were used. Using this data as input Maker predicted 30,908 genes from *desi* and 33,473 from *kabuli*.

3.3 Results and discussion

3.3.1 SkimGBS and genetic recombination

WGS data of ICC4958 and PI489777 were mapped to the *kabuli* (Varshney *et al.*, 2013) reference to call SNPs. To minimise the number of SNPs from homologous regions, only non-repetitively aligned reads were used for SNP calling. A total of 555,346 SNP were predicted with SGSautoSNP (Table 3-1). Illumina paired reads were generated for 46 progeny individuals (total of 20.9 Gbp) with approximately 1x coverage for each individual. On mapping these population individuals sequence reads to the reference, the skimGBS pipeline called genotypes with an average of

147,363 per individual. Due to low coverage of the sequence based genotyping, some alleles were not called. The 43,722 SNPs having missing alleles were discarded and a side-way imputation added 15,221,101 alleles. Out of total of 6,778,700 called alleles, 6,440 SNPs exhibiting heterozygosity were removed from subsequent analysis.

Table 3-1: Parental (ICC4958 and PI489777) SNP prediction.

Name	Length	SNP count	SNPs/Kbp	Transitions/Transversions
Ca1	48,359,943	62,154	1.28	1.681
Ca2	36,634,854	45,836	1.25	1.732
Ca3	39,989,001	59,818	1.49	1.687
Ca4	49,191,682	69,229	1.4	1.645
Ca5	48,169,137	63,406	1.31	1.7
Ca6	59,463,898	75,323	1.26	1.698
Ca7	48,961,560	52,550	1.07	1.664
Ca8	16,477,302	20,303	1.23	1.561
Extra contigs	186,473,055	106,727	0.57	1.726

A total of 4,675 recombinations ranging 20bp to 10kb in length were categorised as gene conversions (Yang et al. 2012) and 3,960 longer than this range were reported as crossovers events. One individual in the population had higher gene conversions than other individuals, affecting the average gene conversion. On close examination of genotype patterns, some recombinations were commonly seen in most of the individuals and this may due to the mis-assemblies in reference genome. To avoid these false positive recombinations, filtering was performed to ignore such common recombination between the individuals. After filtering, the gene conversions ranged from 5 (RIL4) to 22 (RIL29), and crossovers ranged from 0 (RIL7) to 60 (RIL29). There were 219 crossovers and 256 gene conversions identified (Table 3-2). A similar number of crossovers and gene conversions were also seen in *B. napus* (Bayer et al., 2015b). Chromosome wide recombinations on Ca1 before and after filtering were shown in Figure 3-4 and Figure 3-5. After filtering, Ca6 had the highest

average number of crossovers with 1.44 and Ca3 had the lowest average number of crossovers with 0.19, whereas, Ca5 had the highest (1.0) and Ca7 had the lowest number of (0.69) gene conversions (Table 3-3). On comparison of chickpea erroneous rate of recombination with *Brassica* A and C genome demonstrates that *C.arietinum* has low error rate compared to *Brassica* species which depends on genome assembly quality and likelihood of mis-assembly regions. This indicates that *Brassica* A and C genomes may have high mis-assemblies than release *kabuli* reference genome (Bayer *et al.*, 2015b). Compared to *kabuli*, *desi* reference genome is of poor quality (Chapter 2) and can be expect to have higher error rate than *Brassica* species.

Table 3-2: Recombination crossovers and gene conversions in RIL individuals (after filtering of heterozygous individuals), before and after filtering of overlapping recombination.

Name of element	Crossovers before filtering	Gene conversions before filtering	Crossovers after filtering	Gene conversions after filtering
RIL1	85	120	2	8
RIL2	121	132	4	7
RIL3	146	98	8	7
RIL4	117	108	1	5
RIL5	134	105	3	7
RIL6	46	84	22	8
RIL7	68	34	0	7
RIL8	72	30	1	7
RIL10	65	68	3	9
RIL11	124	78	1	6
RIL12	102	34	2	6
RIL13	108	89	12	7
RIL14	98	42	4	5
RIL15	101	41	1	6
RIL16	87	80	2	5
RIL17	108	72	2	7
RIL18	116	98	3	7
RIL20	177	244	5	7
RIL21	134	222	8	7
RIL22	114	98	2	6
RIL24	133	132	1	7
RIL25	83	86	2	7
RIL26	97	121	1	8
RIL27	127	70	9	6
RIL29	212	1,174	60	22
RIL30	121	154	5	7
RIL31	119	162	5	5
RIL32	109	173	6	7
RIL33	96	52	19	7
RIL34	114	109	1	8
RIL35	134	89	7	5
RIL36	103	150	3	7
RIL37	37	74	1	8
RIL40	140	101	7	7
RIL41	109	92	1	6
RIL42	103	59	5	5

Table 3-3: Count of crossovers, gene conversions, after filtering of overlapping recombination events for individuals of the Chickpea DH population after filtering of non-double haploid individuals, counted per chromosome.

Chromosome	Minimum crossovers	Maximum crossovers	Average crossovers	Minimum gene conversion	Maximum gene conversion	Average gene conversion
Ca1	0	2	0.47	0	2	0.89
Ca2	0	16	0.69	0	1	0.89
Ca3	0	3	0.19	0	3	0.97
Ca4	0	14	0.97	0	1	0.94
Ca5	0	11	1.14	0	7	1
Ca6	0	10	1.44	0	2	0.92
Ca7	0	7	0.58	0	1	0.69
Ca8	0	9	0.58	0	5	0.81

Some chromosomes show a greater frequency of recombination towards the telomeres, but others exhibit even distribution. After filtering, RIL29 exhibited a much higher non-crossover rate than other individuals. This could be because of reads for these individuals were mixed with several different individuals, leading to a large number of recombinations. Previous studies suggest that lower numbers of recombination events occur around centromeres and a greater number of crossover events occur towards telomeres (Farkhari *et al.*, 2011; Helms *et al.*, 1992).

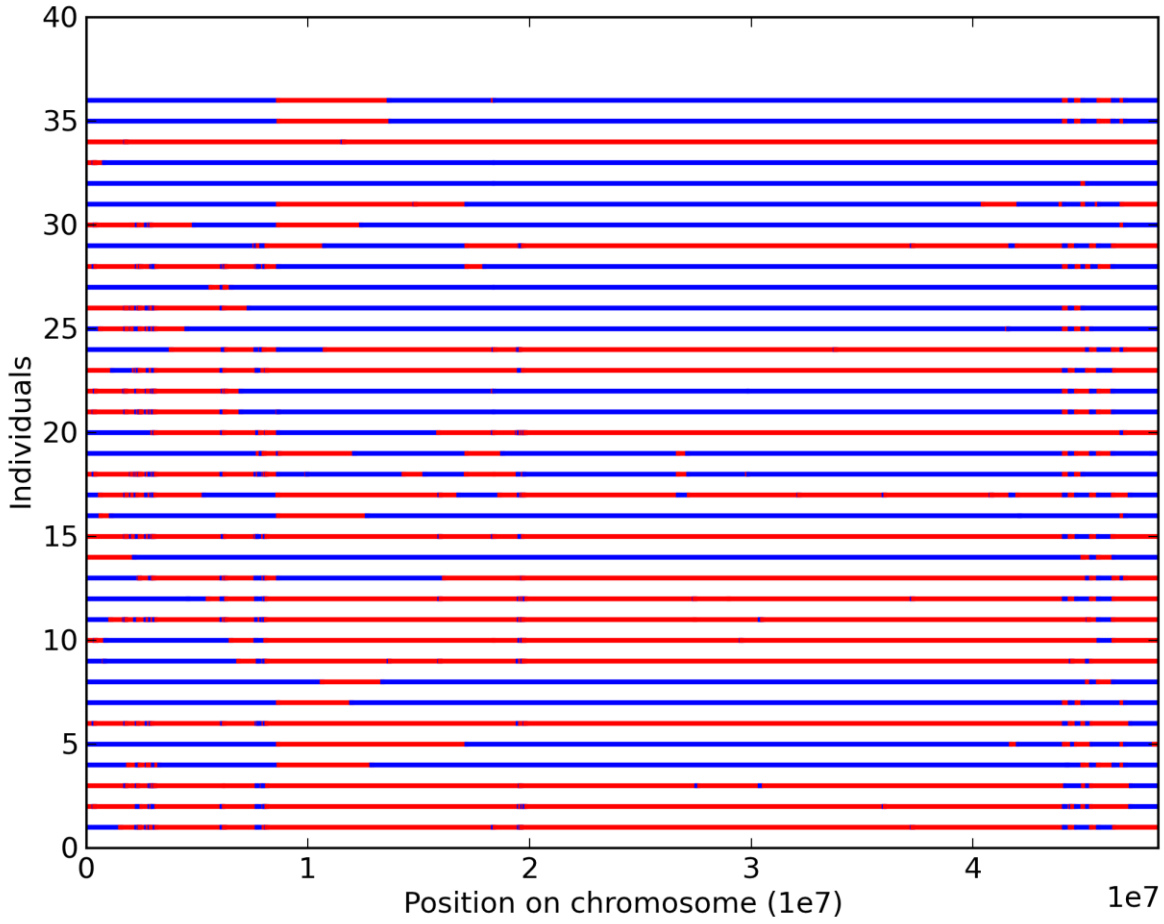


Figure 3-4: Recombination map of Chromosome 1 before filtering of overlapping recombination. (Red:ICC4958, Blue: PI489777, white: missing).

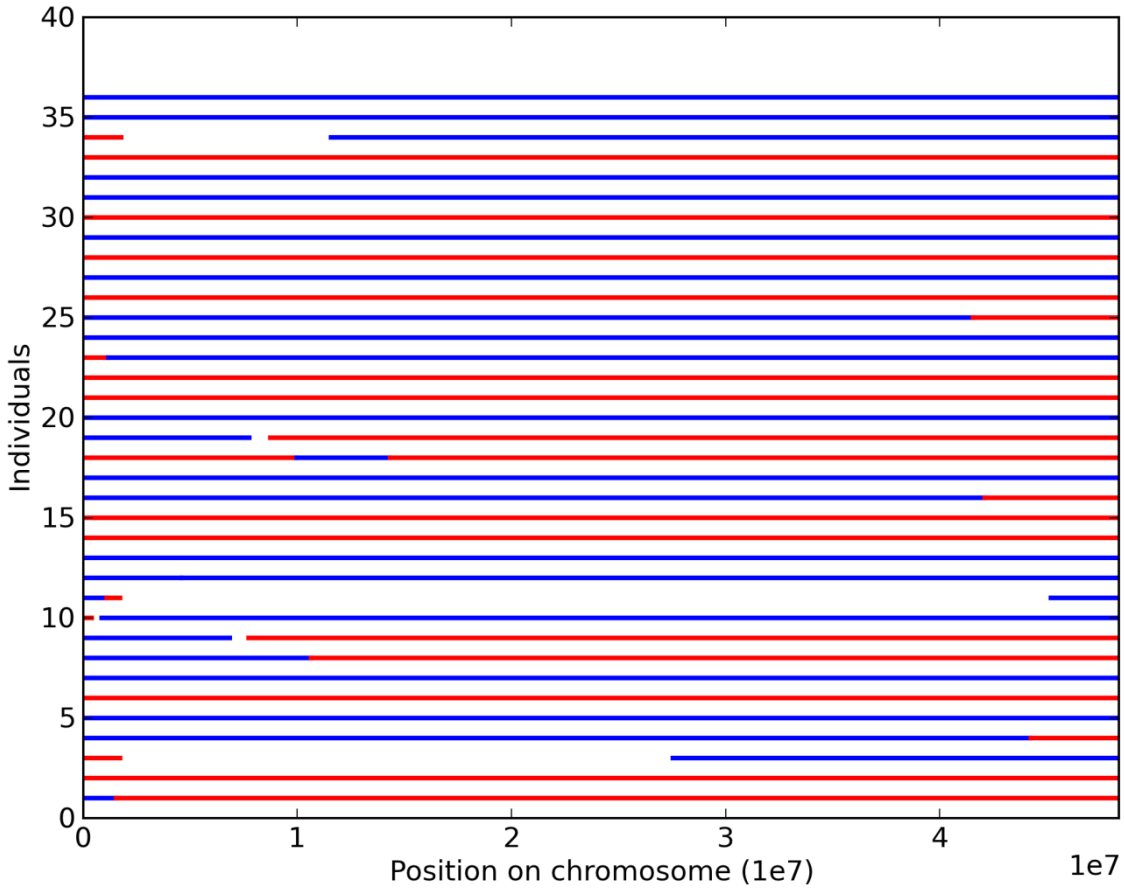


Figure 3-5: Recombination map of Chromosome 1 after filtering of overlapping recombination. (Red:ICC4958, Blue: PI489777, white: missing).

3.3.2 *Desi* and *kabuli* genome improvement

Following the *kabuli* chickpea reference genome assessment with isolated chromosomes (Chapter 2), the genome was fragmented based on earlier identified mis-assembly boundaries (Ruperao *et al.*, 2014b). The split genome consisted of 7,256 contig sequences (including extra contigs). Whole genome sequence paired data of ICC4958 (7.9x) and PI489777 (9.7x) were individually mapped to the fragmented reference genome. Using these mappings SGSautoSNP (Lorenc *et al.*, 2012) identified 555,276 SNPs in 2,660 contigs. Using above used 46 individuals of the ICC 4958 x PI 489777 population, genotypes were called at each SNP position. All wrongly placed sequence regions and extra contigs were manually grouped as unplaced contigs and the remaining sequences were grouped as placed contigs. Based on the pattern of called genotypes and recombinations, the unplaced contigs were placed in appropriate position of chromosomes. ContigPlacer was able to place 307 (76.04Mb size) contigs in their respective chromosomes, resulting in an

improved *kabuli* reference genome that has 423.28 Mbp of chromosomal sequences (Appendix 2). This improved assembly, spans 472.09Mb and contains 5,176 scaffolds including 76.04Mb of extra contig sequences into chromosomes.

Based on the chromosomal genomic assessment (in Chapter 2), the *desi* (ICC 4958) reference genome was observed to be of poor quality and also included non-chickpea genome sequences (Ruperao *et al.*, 2014b). SRA data from NCBI with id SRR500634 was QC filtered and the remaining 110,501,088 paired-end reads were mapped to the published *kabuli* reference genome. Out of the total reads, 61,352,802 mapped reads represented conserved sequence regions between *desi* and *kabuli* reference genomes (including 6,059 contigs). The unconserved 64,983 (out of 237,092) gaps in the *desi* (synteny) reference were filled with above unmapped paired reads to produce 12,144,308 bp (out of 65,227,235) sequences. The remaining unmapped reads were assembled to produced 113,607 contigs and contigs greater than 500bp (3,857 contigs) were included in the draft genome assembly as extra contigs. Since the released *kabuli* genome has mis-assemblies (Chapter 2), the new *desi* reference genome may also carry the same mis-assemblies. On assessing desi genome with isolated chromosome, 43 mis-assembly sequence regions were identified, isolated (based on mapping isolated chromosome sequence data) and fragmented into 89 contiguous sequences. To sort the order and orientation of fragmented *desi* reference genome, initially contigs similar to *kabuli* pseudomolecules sequence were filtered. Among 10,005 fragmented *desi* genome contigs, BLAST identified 4,071 contigs having significant similarity with *kabuli* pseudomolecules. Each set of synteny contigs corresponding to each *kabuli* single pseudomolecule were LASTZ mapped to (improved) *kabuli* genome produced synteny co-ordinates (in ".lav" map file). On parsing the map information the contig corresponding to chromosome were sorted and re-oriented to make final *desi* pseudomolecules. Among the remaining 5,939 contigs (did not show synteny with *kabuli* pseudomolecules) only contigs more than 1 Kb in length (3,313) were used as extra contigs. Total span of final *desi* genome sequence is 427.16 Mb with 30.12% of GC count, of which eight-nine percent (381.75 Mb) of total assembly was assigned to eight pseudomolecules. Ca3 is the longest sequence with 61.12 Mb and shortest pseudomolecules Ca8 have 18.29 Mb of sequence.

3.3.3 *Desi* and *kabuli* reference genome assessment

Both *desi* and *kabuli* improved reference genomes were assessed with their respective isolated chromosome sequence data. All isolated *kabuli* chromosome sequence reads (Ruperao *et al.*, 2014b) were mapped to both reference genomes and heatmaps were produced to assess the genome quality. Both genomes show clear correspondence of all physical chromosomes with their respective isolated chromosomes (Figure 3-6, Figure 3-7). There are a few gaps as shown on the *desi* heatmaps indicating that the *kabuli* chromosome reads did not map, possibly due to the sequence uniqueness in *desi* genome. To verify this, WGS of ICC4958 (*desi*) Illumina paired-end reads were mapped on the improved genome and found sequence reads covering the sequence gaps (Figure 3-7).

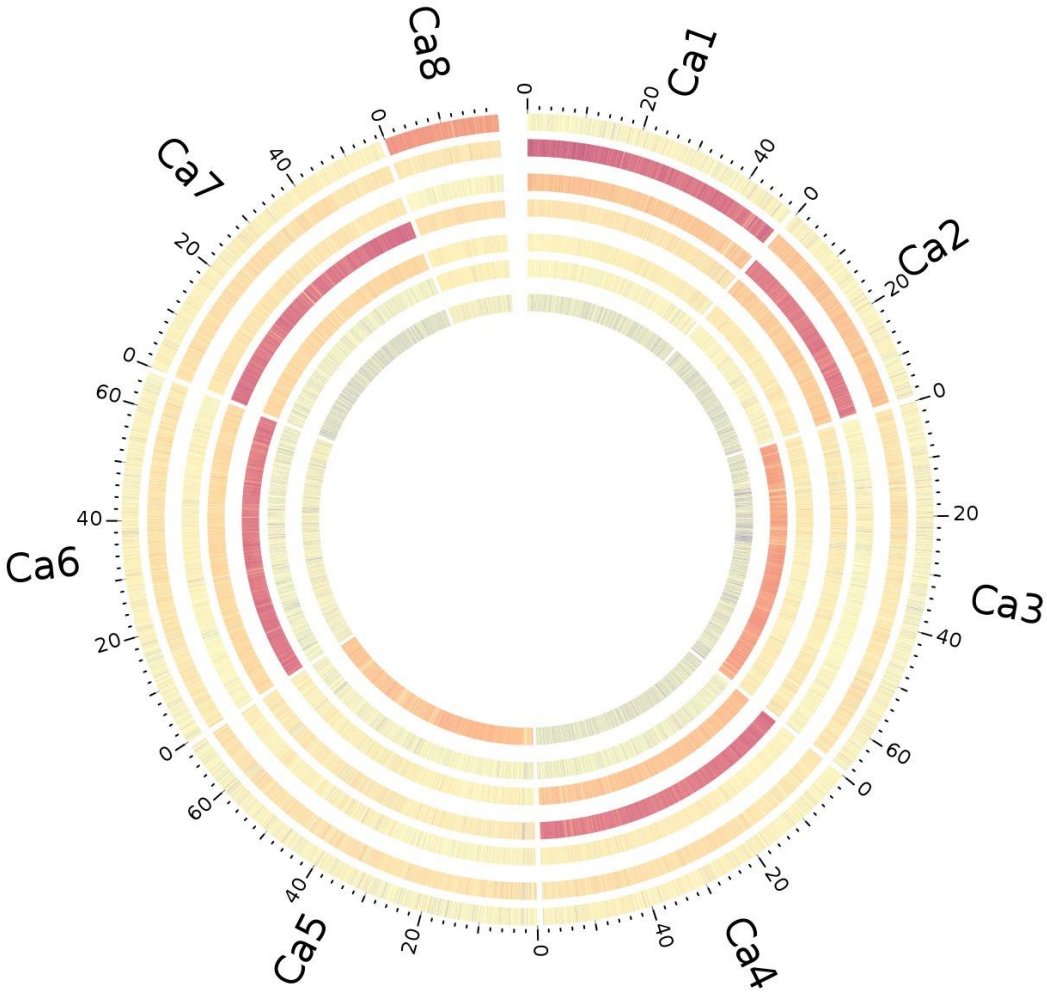


Figure 3-6: Circos heat map demonstrates the quality of *kabuli* reference genome. Assessment with *kabuli* chromosomes sequence data with A being on the inner-most lane, followed by B, C, DE, F, G and H on the outer-most lane.

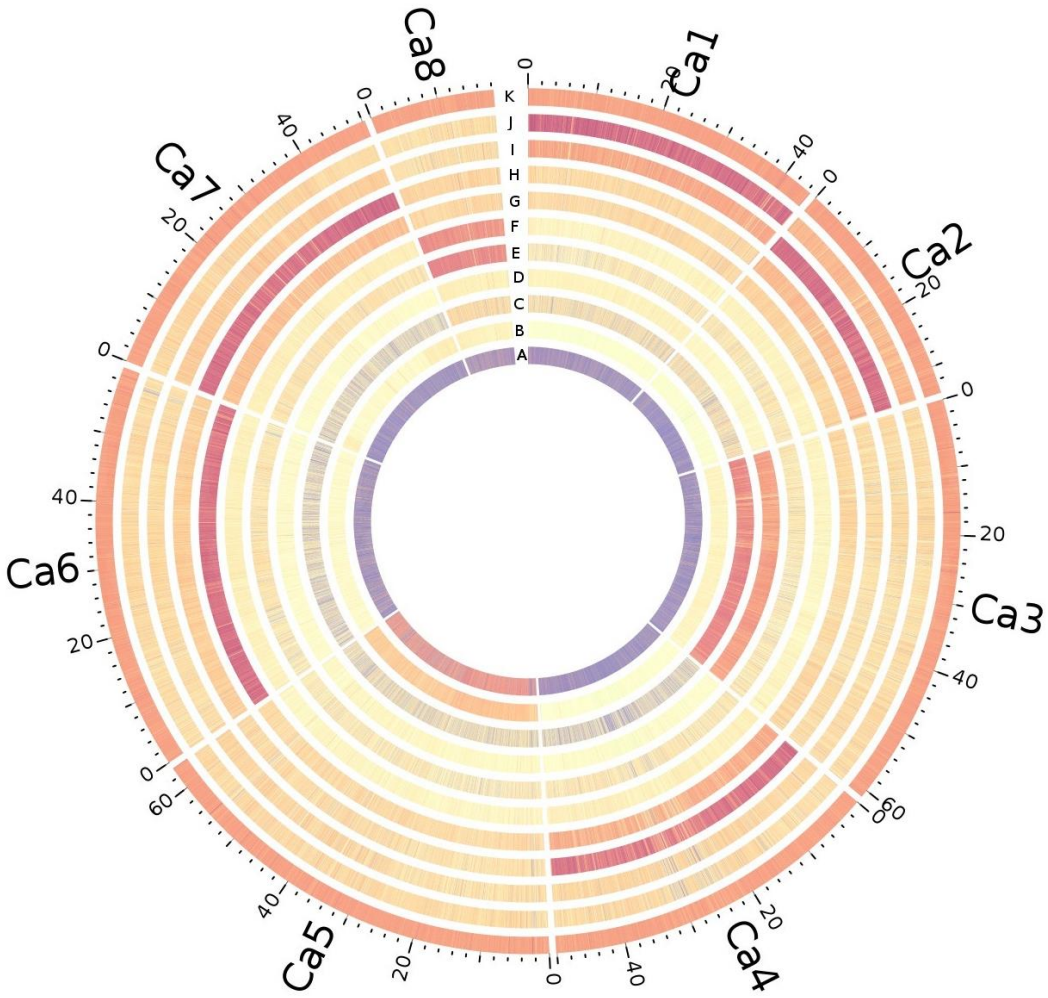


Figure 3-7: Circos heat map demonstrating quality of improved *desi* reference genome. Each circular heat-maps represents as below.

***Kabuli* chromosomes**

A = A chromosome
C = B chromosome
E = H chromosome
G = C chromosome
H = DE chromosome
I = F chromosome
J = G chromosome

***Desi* chromosomes**

B = A chromosome
D = B chromosome

F = H chromosome
K = ICC4958 (*desi*) WGS data

3.3.4 *Desi* and *kabuli* reference genome comparison

The statistics shown in Appendix 2 confirm that the two draft chickpea genomes are different in assembly quality and size. The *kabuli* (Varshney *et al.*, 2013) reference genome assembly was assembled from Illumina data while the *desi* (Jain *et al.*, 2013) reference genome used a hybrid approach combining Roche/454 and Illumina data. In this study, only Illumina data was used for genome improvement. The improved *kabuli* assembly captures 423.28 Mb of chromosomal sequence (57.20% of estimated genome size) and the *desi* assembly captures 413.30 Mb of chromosomal sequence (55.85% of estimated genome size), representing a similar proportion of chromosomal sequence. However, the *kabuli* reference based *desi* genome assembly is 60 Mb smaller having total of 12,686 extra contigs and a N50 of 50,017,301bp compared to the *kabuli* genome with 5,176 extra contigs and a N50 of 58,848,791bp (Appendix 2). Construction of (*desi*) genomic assembly based on the *kabuli* draft genome has restricted the *desi* genome size; however, *desi* unique sequence in form of 12,678 contigs represents the distinctness.

Pairwise comparison of each chromosome from the two improved genome assemblies shows major sequence similarity between *desi* and *kabuli* assemblies (Figure 3-8). Since *desi* contigs were ordered and oriented based on synteny with *kabuli* genome, the sequence regions in *desi* genome match with those of corresponding pseudomolecules of the *kabuli* genome assembly.

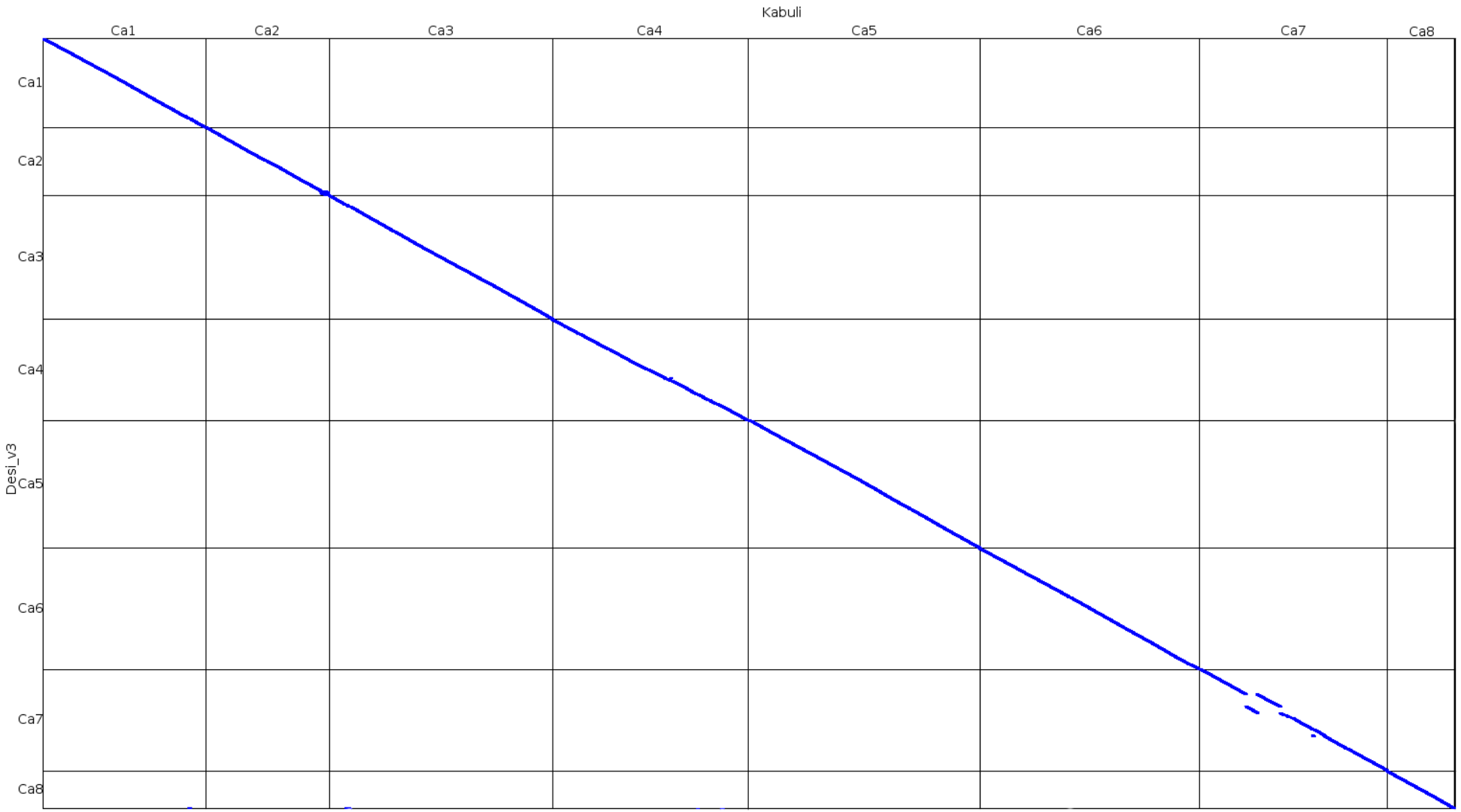


Figure 3-8: Dotplot comparison of *desi* and *kabuli* pseudomolecules.

3.3.5 Unplaced contig sequence assessment

A total 1,874 and 1,998 unplaced contigs (spanning 40.64 Mb and 37.74 Mb) from both *desi* and *kabuli* reference genomes showed correspondence with isolated chromosomes (Table 3-4). Compared to contigs in *desi* and *kabuli* released reference genome, improved version has placed 329.65 Mb and 139.73 Mb of chromosomal sequence in respective chromosomes. Even though, unplaced contigs from both improved genomes show correspondence with isolated chromosomes, the skimGBS method did not place them in pseudomolecules. This may be due to lack of SNPs on contigs and for contigs having SNPs, the genotype based representative block of these contigs did not show similarity with any of the other representative blocks of contig/sequence in pseudomolecules. Out of 7,256 split (at mis-assembled co-ordinates) *kabuli* genome scaffolds, only 2,655 scaffolds contain SNPs between PI489777 and ICC4958. The remaining 4,601 scaffolds do not carry SNPs. Out of the 10,005 fragmented *desi* scaffolds showing synteny to the *kabuli* reference genome, only 1,784 contigs/scaffolds exhibit SNPs between ICC1882 and ICC4958. Sequences without SNPs cannot be placed by contigPlacer.

Table 3-4: Contiguous sequence assessment.

	***Desi* (Jain *et al.*, 2013)**		***Desi* improved**		***Kabuli* (Varshney *et al.*, 2013)**		***Kabuli* improved**	
Chromosome	**Number of sequences**	**Sequence length**	**Number of sequence**	**Sequence length**	**Number of sequences**	**Sequence length**	**Number of sequence**	**Sequence length**
A(Ca5)	19,777	63,381,859	10	19,742	1,206	39,091,098	12	46,495
B(Ca3)	15,950	57,764,518	198	3,794,568	738	38,588,828	232	3,223,680
C(Ca6)	14,375	54,542,218	303	2,005,961	588	20,017,227	348	2,627,537
DE(Ca4,7)	20,972	90,635,403	357	8,288,259	877	36,850,083	409	6,681,564
F(Ca2)	12,497	43,207,629	449	15,209,439	498	20,460,868	464	14,696,089
G(Ca1)	12,261	39,146,091	237	2,217,343	488	9,191,325	244	1,429,262
H(Ca8)	10,984	21,623,569	320	9,108,552	457	13,274,427	289	9,037,363
Total	**106,816**	**370,301,287**	**1,874**	**40,643,864**	**4,852**	**177,473,856**	**1,998**	**37,741,990**

3.3.6 Genome annotation

The protein-coding genes were analysed based on various gene prediction approaches including *ab initio* and homology-based gene prediction, followed by merging the predicted results. In *desi*, a total of 30,908 genes (a total of 123.8 Mbp) were predicted, with an average gene length of 4,008 bp, an average coding sequence length of 198.38 bp (total of 35.53 Mbp) and an average of 5 exons per gene. The overall GC content with in the coding sequence was higher (40.54%) than that of whole genome (30.2%). In *kabuli*, 33,473 genes with a total size of 127.7 Mbp were identified, with an average gene length of 3,816.14 bp, coding sequence length of 204.42 bp (total of 38.5 Mbp) and an average of 5 exons per gene Similar to the *desi* genome, the overall GC content for the *kabuli* coding sequence is higher (40.45%) than that of the whole genome (29.34%).

3.4 Conclusion

We describe a new and high quality chickpea *desi* and *kabuli* reference genome, representing substantial improvements over the previous versions (Jain *et al.*, 2013; Varshney *et al.*, 2013) in 2013. The improved *kabuli* reference genome consists of 423.28 Mb of the respective chromosomal sequence (compared to 347.24 Mb in the previous release) and *desi* assembly has 381 Mb of chromosome sequence. Both improved genomes have been validated with isolated chromosome and no mis-assemblies were identified. The high quality of curation on chromosomal sequence and associated gene models will serve as much better reference for chickpea researchers. A novel method of assessing the genome quality with isolated chromosomes was implemented to improve genome assemblies. This method can be used in other projects to increase genome assembly qualities.

4 Chickpea variation study, SNP density and gene loss

4.1 Introduction

With improved sequencing technologies (NGS) it is possible to re-sequence entire plant genomes at high coverage more efficiently and economically compared to traditional sequencing methods. Using this sequencing technology DNA from whole populations can be sequenced to gain insights into population genetics (Varshney *et al.*, 2009). NGS technologies with available reference genome sequences have revolutionized the study of population level plant diversity for chickpea. The ability of producing whole genome sequencing (WGS) using the NGS approach has significant impact on crop genetics and breeding.

Molecular breeding strategies have adapted to advance improvement programmes in several crops, including legumes such as soybean and common bean (Chamarthi *et al.*, 2011). Implementation of a similar approach in chickpea has been relatively slow due to the limited availability of molecular markers. Among different marker systems, SSR and SNPs are the markers of choice for genetic and plant breeding applications (Close *et al.*, 2009). Within these types of markers, the high abundance of SNP markers has become the optimal choice. NGS technology has enhanced the detection of large number of SNPs in a relative short time frame (Shendure and Ji, 2008). SNP density between species is significant, ranging from one SNP per 16bp in Eucalyptus species (Kulheim *et al.*, 2009) to one SNP per 7000bp in tomato (Labate and Baldo, 2005). The range of SNP density in chickpea varies between one SNP per 36bp (Gujaria *et al.*, 2011) to one SNP per 973bp (Jhanwar *et al.*, 2012). The high frequency of SNPs compared to SSR markers makes SNPs the marker of choice (Gaur *et al.*, 2012).

In addition to the advances in sequencing technologies, several new technologies for large scale SNP genotyping have been developed. Depending on these technologies, some platforms can integrate from a few thousand SNPs to million SNPs. Illumina GoldenGate and Infinium genotyping platforms have been widely used in many crops. SNP genotyping of large scale segregating populations

becomes cost effective for developing high density genetic maps, genome wide association mapping, marker assisted selection (MAS) and genomic selection (GS) studies. The genome wide marker trait association depends on the marker density and distribution, and hence there is a need to develop SNP based genotyping allowing association studies and recombinations in a given populations (Chapter 3).

Genetic variations leading to duplication, polyploidy and gene loss (canola, blackleg fungus (Golicz *et al.*, 2015), wheat (Gornicki *et al.*, 1997)) are very common in the evolution of plants. Legumes of the large Papilionoideae subfamily have undergone whole genome duplication (Cannon *et al.*, 2006). Over the time of evolution, plant genomes become diploid, accompanied by major variation in the genome structure, like gene loss and chromosomal rearrangements. The gene loss, silencing, and activation occur in the early stages of allopolyploid formation. Recent studies demonstrate that genes can be lost from one parental genome as early as the F1 hybrid forms or even after chromosome doubling (Kashkush *et al.*, 2002). The genes that are present in one genome and absent in another genome have been defined as presence/absence variation (PAV) or CNV (copy number variations). Illegitimate recombination events can also be one of the reason of large scale of gene loss (Devos *et al.*, 2002) or PAV. These gene losses also lead to loss of function and result in ongoing genetic deterioration. Analysis of these variations highlights the association with ancestral evolution events and domestication (Springer *et al.*, 2009).

In addition to trait association studies, genetic diversity can be estimated using genetic molecular markers. A number of different molecular markers are in use to determine genetic relationships between the chickpea accessions.

This chapter present results of genome-wide SNP discovery, gene loss and genetic relatedness in 47 chickpea cultivars. Among the 47 chickpea varieties, 29 varieties were sequenced and used for assessing a *kabuli* reference genome (Varshney *et al.*, 2013) and the remaining 18 varieties were sequenced as part of an Australia India Strategic Research Fund (AISRF) funded project.

4.2 Methods

4.2.1 Experimental workflow

This chapter focus mainly on variant (SNP) discovery, diversity and gene variability between 47 chickpea varieties. The schematic outline of the experimental design is shown in Figure 4-1. For SNP discovery, WGS pair reads were mapped with high specificity to the released *kabuli* (Varshney *et al.*, 2013) reference genome. Only paired mapped data was used for SNP discovery and density plots were produced for genome wide variation study. For downstream analysis, SNPs were further studied with structural and functional annotation and validated with skimGBS based population data. Genome wide variation, evolutionary relatedness and admixture analysis was also performed. Using the same reads mapped with reduced specificity, a gene variability study followed by gene based relatedness between all chickpea cultivars was performed. In this study a significant percent of unmapped reads was also used for novel sequence study.

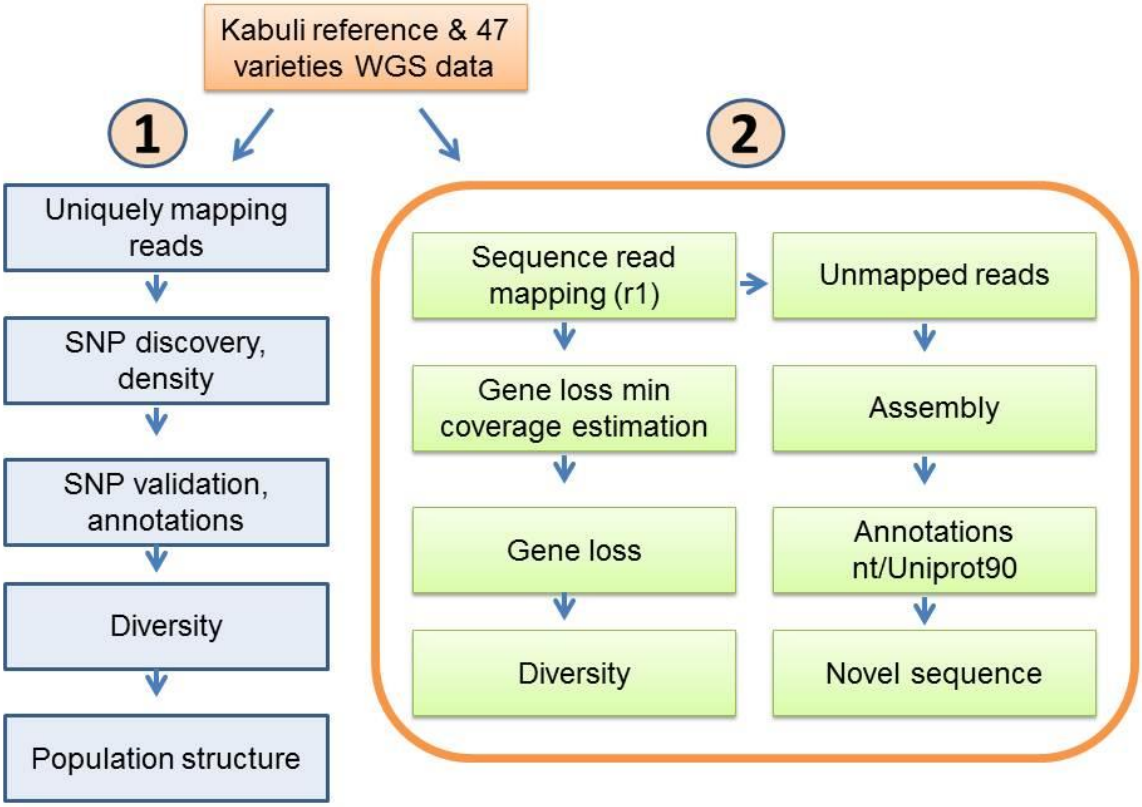


Figure 4-1: Schematic diagram describing the experimental approach.1: SNP discovery and downstream analysis and 2) Gene variability and novel sequence study.

4.2.2 Whole genome sequence data production

Dr. Jacqueline Batley group performed whole genome shotgun sequencing using HiSeq 2500 sequencing platform. A total of 18 chickpea variety paired-end sequencing libraries were constructed with insert size of ~300bp. In total, 65.63 GB of paired read data with 101 bp read length were generated with a coverage ranging from 5.5x to 11.4 x (average 9.7x). Furthermore, 29 chickpea accessions (average 9.7x coverage) WGS data were also downloaded from ICRISAT, which are used for building *kabuli* draft genome assembly (Varshney *et al.*, 2013) (Appendix 3).

4.2.3 Read mapping and SNP identification

Illumina paired reads of all 47 chickpea accessions were mapped to the *kabuli* draft genome using SOAP2 v2.21 (Li *et al.*, 2009b) using –r 0 as parameter (which ignores reads having multiple hits). The SOAP output format was converted to BAM

format using Samtools v0.1.17 (Li *et al.*, 2009a). Duplicated reads were removed with MarkDuplicates from the Picard tools (http://broadinstitute.github.io/picard/) and remaining reads in BAM format were merged using samtools. The resulting BAM file was used for SNP discovery using the SGSautoSNP pipeline (Lorenc *et al.*, 2012). Originally, SGSautoSNP removes all SNPs with more than one allele in an individual (ignoring heterozygous SNPs). SGSautoSNP was improved to retain such heterozygous SNPs on the condition that each allele's Phred quality score is at least 20. Since this still introduces SNPs from single mis-mapped reads, only heterozygous SNPs showing heterozygosity in at least eight individuals were considered. All homozygous genotypes that differ between the varieties are considered as homozygous SNPs. The SGSautoSNP pipeline was also used to call heterozygous alleles from each variety and later filtered based on number of cultivars exhibiting heterozygous alleles and allele frequency. To visualise the homozygosity and heterozygosity at chromosomal level, SNP density plots were produced using Circos v0.64 (Krzywinski *et al.*, 2009b).

4.2.4 Validation of SNPs using skim genotyping by sequencing

A set of 46 population individuals (RILs developed by crossing parents) ICC 4958 x PI 489777 sequence reads were mapped to the *kabuli* reference genome using SOAP2 v2.21. A custom script '*snp_genotyping_all.pl* (Bayer *et al.*, 2015a) was used to call genotypes from each individual of population as explained in chapter 3.

4.2.5 Functional annotation of SNPs

SNPeff (Cingolani *et al.*, 2012) software was used to functionally annotate predicted SNPs using gene models published with the chickpea reference genome (Varshney *et al.*, 2013). For each SNP, the location and the functional annotation (non-synonymous, synonymous, stop codon gain or loss, and amino acid changes) were determined.

4.2.6 Gene loss estimation

Gene loss detection is based on the quantification of reads aligned to exons in a reference. This method counts the number of exons present in the reference and calculates the percentage of their length aligned with matching NGS data. Exons in the reference with no reads aligned might be lost or diverged in the sequence.

In this method WGS data of all 47 chickpea cultivars were mapped to the *kabuli* reference genome using SOAP2, allowing reads to align to their best position within available multiple hits. An in-house developed program called SGSGeneLoss (Golicz *et al.*, 2015) assesses mapped reads associated with exons in genes and decides whether a gene is present or lost. An R script filters the non-covered read regions on the reference and plots a graph showing exons covered, coverage depth, non-covered regions and coverage category (Figure 4-2).

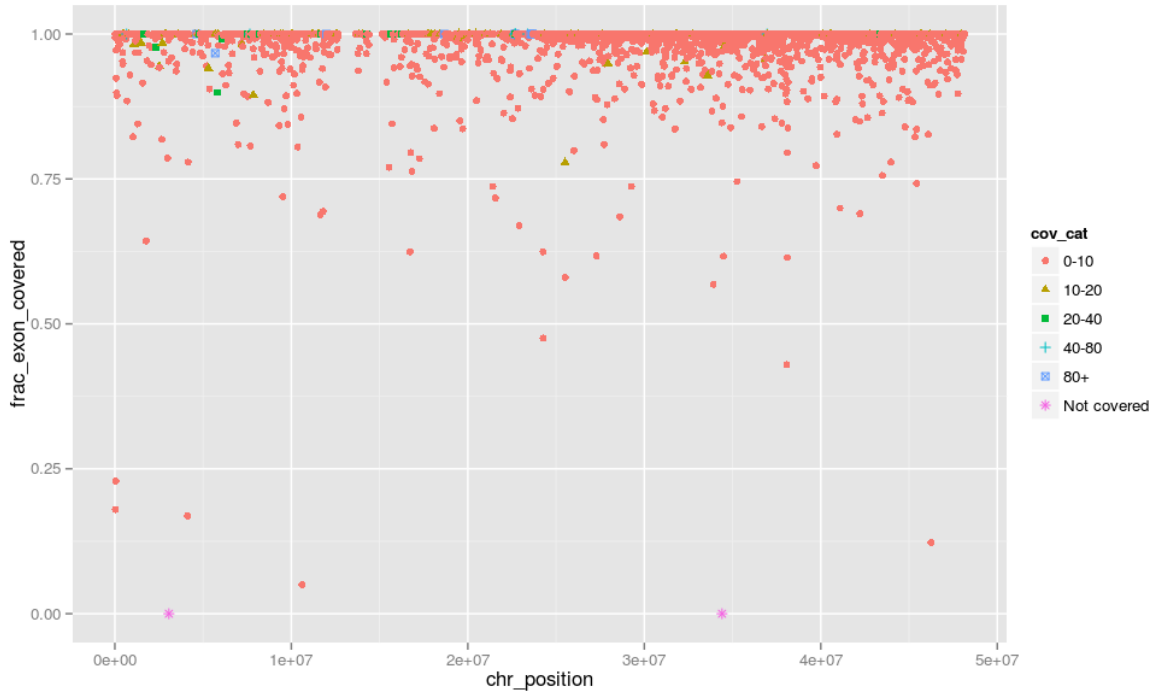


Figure 4-2: An example of Ca5 of ICC1882 exon coverage plots, a dot represents a gene. Each colour exemplifies different coverage categories. The pink star shape represents genes with no aligned reads.

For minimum coverage estimation, Illumina *desi* ICC4958 reads (Jain *et al.*, 2013) were used and split into different subsets having different levels of coverage. These subsets of paired reads were mapped to the *desi* reference using SOAP2 (Li *et al.*, 2009b) and gene loss estimated at each level.

4.2.7 Identification of gene loss events

Based on above estimation, for this study, SOAP2 (Li *et al.*, 2009b) (-r1 parameters) was used to map WGS paired read data of each chickpea variety to chickpea *kabuli* reference genome (including chloroplast sequence NC_011163) allowing reads

having multiple hits to map one best hit. To identify the candidates of lost genes, available gene models for the *kabuli* reference genome were used to assess the exon coverage. In this study, a total of 2,253 gene loss events were identified across 47 chickpea accessions (summed up from all cultivars).

4.2.8 Identification of novel genes

After mapping WGS paired reads to CDCFrontier *kabuli* genome, we pooled and assembled the unmapped reads from each accession using Velvet optimiser (Zerbino and Birney, 2008). Contigs shorter than 500bp were excluded and the remaining contig sequences (with an average length of 1.4Kb) were used as novel sequence. We used BLAST (Camacho *et al.*, 2009) to find local alignments to all these candidate novel contigs against NCBI nt database (nucleotide sequence database) to identify diverse homologs. All (326) contig sequences corresponding to mitochondria were excluded. The remaining contig sequences (46,262) indeed have more or less similar homologous sequences with identity >70%.

SNAP (Korf, 2004) and AUGUSTUS (Stanke *et al.*, 2006) were used for *ab initio* gene prediction. mRNA (SRR491157), 45,038 ESTs and 32,097 chickpea proteins (all downloaded from NCBI) were used for homology based predictions. Finally, predictions from both approaches were combined using MAKER (Cantarel *et al.*, 2008) software. We annotated these sequences through BLAST comparison with NCBI nt database.

Both mitochondrial and novel sequences were appended to the *kabuli* reference genome (also having chloroplast sequence) to identify the genome wide gene variability. SGSGeneloss (Golicz *et al.*, 2015) was re-run to predict presence or absence of genes in each accession.

4.2.9 Genetic diversity and phylogeny

MEGA 5.2 (Tamura *et al.*, 2011) was used to calculate pairwise Nei's genetic distance between all chickpea varieties. The wild type (PI489777) was used to root the phylogenetic tree. Population structure was inferred using fastSTRUCTURE (Raj *et al.*, 2014) and Distruct1.1 was used for displaying the population structure.

A combination of *kabuli* variable genes and novel sequence genes was used to construct a whole-genome phylogeny tree based on the gene content. We generated

a matrix of binary discrete characters ('P' and 'A' for presence and absence). Using this matrix, we constructed a phylogeny tree implementing the distance method form MAGE5.2 (Tamura *et al.*, 2011).

4.3 Results and discussion

4.3.1 SNP discovery

A set of 47 chickpea varieties (Appendix 3) WGS data was mapped to *kabuli* reference genome (Varshney *et al.*, 2013). Using SOAP (Li *et al.*, 2009b) with –r 0 an average of 39 million pair reads were mapped (maximum 70 million and minimum 11 million paired reads) to reference genome (Appendix 4). Only uniquely mapped pair reads were considered for SNP calling.

SNP discovery resulted in the identification of 611,365 homozygous and 1,367,577 heterozygous SNPs. Homozygous SNPs were identified as variant alleles which are homozygous across all varieties, while SNPs which showed more than two alleles in at least one individual are reported as heterozygous SNPs.

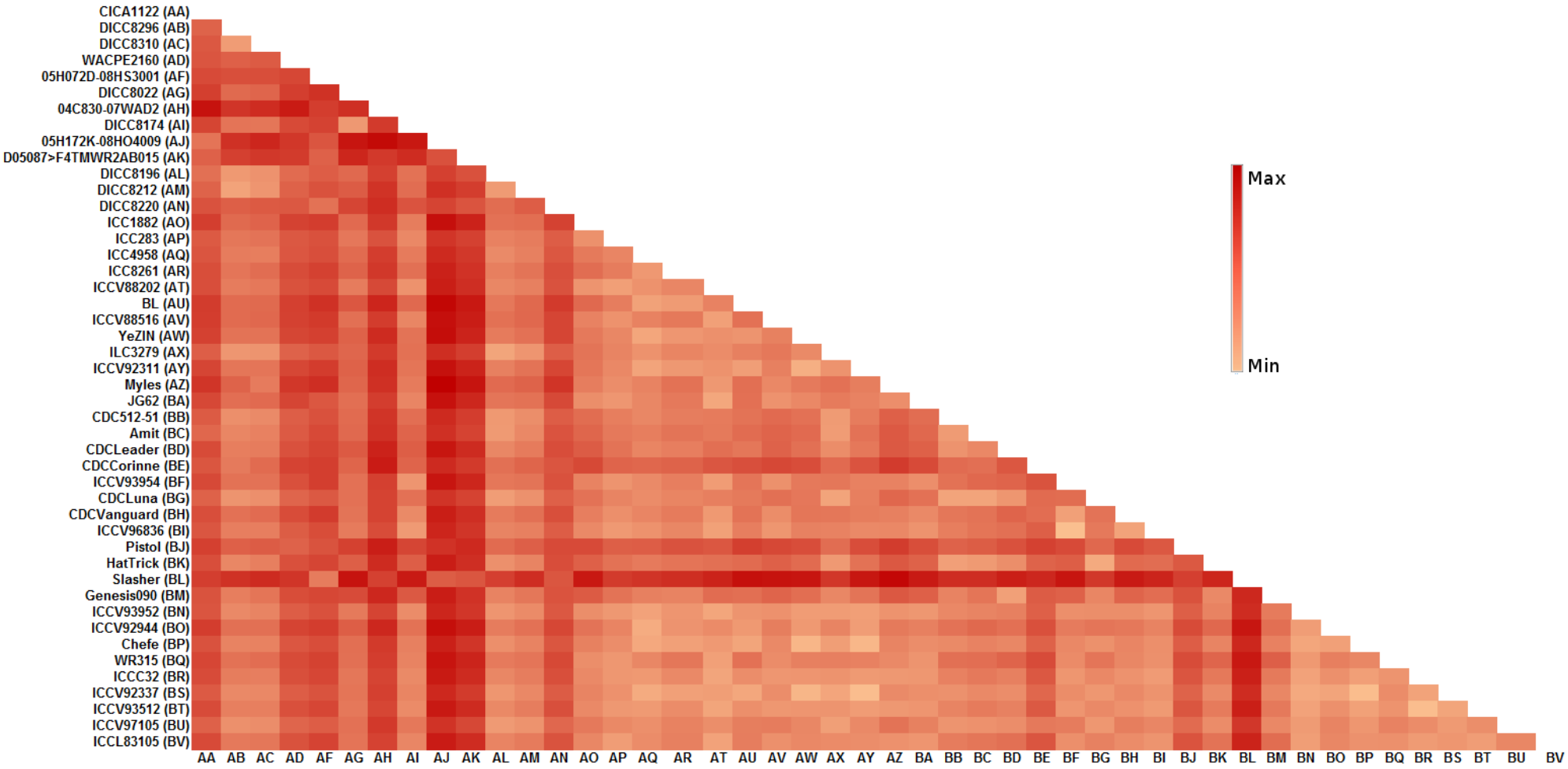


Figure 4-3: Heatmaps demonstrating the level of variation and diversity between chickpea cultivars (wild type excluded).

An inter-varietal SNP matrix was constructed for homozygous SNPs (Figure 4-3). Pairwise SNP numbers between varieties varied from 1,874 (ICCV93954 - *desi* and ICCV96836 - *desi*) and 392,672 (05H172K-08HO4009 - *desi* and PI489777 - wild). The number of homozygous SNPs between the PI489777 wild type and the domesticated varieties is much greater than between domesticated varieties, highlighting the genetic bottleneck due to domestication and breeding. *Desi* varieties demonstrated the greatest diversity, with homozygous SNP density ranging from 3.5 SNPs/Mbp (between ICCV93954 and ICCV96836) to 78 SNPs/Mbp (between Myles and FLIP94-079C-01-481*03HS003). In contrast, SNP density in *kabuli* varied from 4.3 SNPs/Kbp between ICCV92337 and ICCV92311 and 36 SNPs/Mbp in between Amit and YeZIN.

The number of heterozygous SNPs varied between 6,046 (CDC512-51) and 94,340 (ILC3279) (Appendix 5). Heterozygous SNP density ranged from 0.01 - 0.17 SNPs/Kbp. In inbreeding species, this relatively low level of heterozygosity in chickpea results from a low level of outcrossing (<1%)(Tayyar *et al.*, 1996). Among chickpea *desi* types, CDC512-51 (6,046) was the most homozygous and 05H172K-08HO4009 (77,177) the most heterozygous. A similar number of heterozygous SNPs was identified in *kabuli* types, varying between 12,997 (ICCV92337) and 94,340 (ILC3279).

Between the different chickpea chromosomes, Ca4 has the highest density of homozygous SNPs (1.83 SNPs/Kbp) followed by Ca1 (1.45 SNPs/Kbp), while Ca7 is the least polymorphic (1.10 SNPs/Kbp), a similar distribution of marker density was observed in earlier studies (Deokar *et al.*, 2014; Kujur *et al.*, 2013). For heterozygous SNPs across all varieties, Ca 4 exhibited the highest heterozygosity (2.6 SNPs/Kbp) and chromosome 8 the least heterozygosity (2.0 SNPs/Kbp) (Table 4-1).

Table 4-1: Genome wide heterozygous SNPs identified in all chickpea varieties.

Chromosome	Heterozygous SNPs	Kbp^{-1}
Ca1	120,997	2.50
Ca2	84,735	2.31
Ca3	88,657	2.21
Ca4	129,204	2.62
Ca5	102,331	2.12
Ca6	134,244	2.25
Ca7	117,071	2.39
Ca8	34,137	2.07
Extra contigs	556,201	2.98
Total	1,367,577	2.56

Since Ca4 shows great divergence between *desi* and *kabuli* types (Millan *et al.*, 2010; Varshney *et al.*, 2013), the high SNP density in this chromosome is expected. Homozygous SNP density varied along the lengths of the pseudomolecules and between pseudomolecules (Figure 4-4). For example, the first half of chickpea chromosome Ca1 has a higher density of homozygous SNPs compared to the remaining half of chromosome. A similar pattern of SNP density is observed in chromosomes Ca7 and Ca8. This non-uniform SNP density pattern is observed in other plants such as wheat (Berkman *et al.*, 2013; Lai *et al.*, 2015) and reflects rounds of selection and local genetic bottlenecks. The heterozygous SNP density varied between the cultivars and also within the cultivars (between and across the chromosomes). For example, CDCLuna exhibits high density on the second half of chromosome 4 (Figure 4-4).

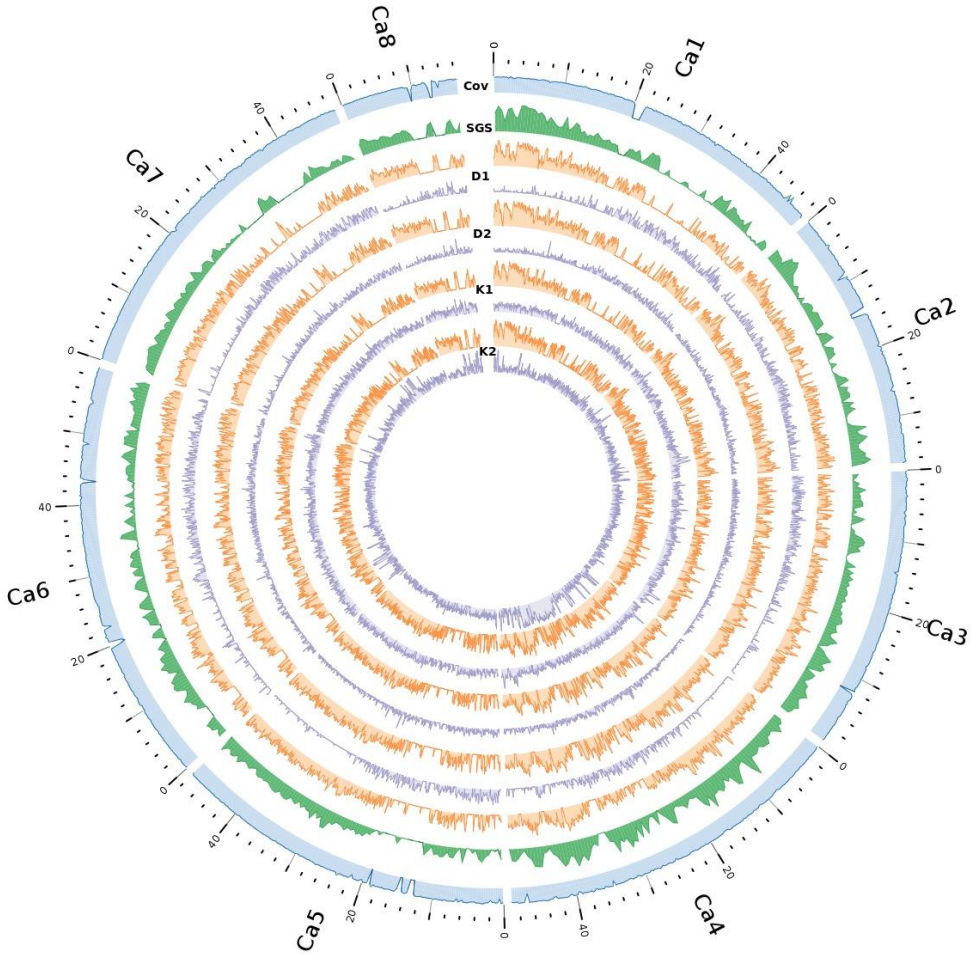


Figure 4-4: Homozygous and heterozygous SNP density. Image showing homozygous SNPs in orange colour plots and heterozygous SNPs in purple colour plots for chickpea *kabuli* types ILC3279 (K1), CDCLuna (K2) and *desi* types HatTrick (D1), ICCV96836 (D2), SGSautoSNP-SNP density plot, chickpea 47 varieties coverage plot (Cov).

4.3.1.1 Functional SNP annotation

To assess the effect of predicted polymorphisms between the 47 chickpea varieties, we functionally annotated all homozygous SNPs (Table 4-2). Among the reported homozygous 611,365 SNPs, transitions (Tr) were more abundant (368,805) than transversions (Tv 243,518) with a Tr/Tv ratio of 1.5. This bias was also observed in

previous chickpea analyses (Deokar *et al.*, 2014; Roorkiwal *et al.*, 2014a) and is frequently observed in other crop species such as wheat (Lai *et al.*, 2012b), barley (Duran *et al.*, 2009b) and maize (Barker *et al.*, 2003; Batley *et al.*, 2003a). Out of a total of 611,365 homozygous SNPs, the majority 37.0% (284,832) were found in intergenic regions. The remaining SNPs were located in intronic 11.8% (90,973), exonic 5.6% (43,057), downstream 23.9% (183,735) and upstream 21.6% (166,261) regions. Of the homozygous SNPs predicted to be in coding regions 56.9% (24,285) were synonymous while 43.1% (18,365) were predicted to change the structure of 9,807 encoded proteins. These may be considered candidate causal SNPs for heritable differences between these chickpea varieties.

Table 4-2: SNP annotations.

	Ca1	Ca2	Ca3	Ca4	Ca5	Ca6	Ca7	Ca8
SNPs	70,196	48,483	54,640	90,070	60,561	76,240	54,087	19,721
Mis-sense	2,676	1,747	2,248	3,209	2,677	2,893	1,869	1,170
Non-sense	38	34	30	42	34	34	28	8
Silent	3,745	2,209	2,914	4,212	3,443	3,654	2,455	1,688
Mis/Silent ratio	0.7146	0.7909	0.7714	0.7619	0.7775	0.7917	0.7613	0.6931
Transitions	41,919	29,631	33,434	53,354	37,376	47,082	32,592	11,647
Transversions	28,393	18,913	21,270	36,867	23,238	29,248	21,542	8,091
Ts/Tv	1.4764	1.5667	1.5719	1.4472	1.6084	1.6098	1.513	1.4395
Intron	13,635	8,876	10,388	16,428	12,785	13,695	9,346	5,748
Non-Synonymous-coding	2,657	1,738	2,233	3,182	2,664	2,868	1,860	1,163
Non-Synonymous-start	1	1	0	4	0	0	1	0
Start-lost	8	2	1	2	6	7	2	1
Stop-gained	38	34	30	42	34	34	28	8
Stop-lost	7	6	11	17	6	15	5	5
Synonymous-coding	3,733	2,207	2,908	4,205	3,436	3,646	2,451	1,684
Synonymous-stop	6	2	3	5	6	6	3	4

4.3.1.2 SNP Validation

There are several high throughput methods for the validation of SNPs. We used skim-based genotyping by sequencing to assess the segregation of homozygous and heterozygous SNPs in a population of 46 individuals from a cross between ICC4958 and PI489777. We identified 371,461 homozygous SNPs between these two parents as well as 33,804 and 7,849 heterozygous SNPs within ICC4958 and

PI489777 respectively. After mapping reads from the segregating population individuals, the genotype was called and the SNPs assessed for its segregation in the population. A total of 342,644 out of 371,461 homozygous SNPs (92.3%) were validated as segregating in the population (Table 4-3).

Table 4-3: Homozygous SNPs validation with genotypes called from ICC4958 x PI489777 population individuals.

Chromosomes	SNPs predicted	Missing data	True	True SNP percent
Ca1	41,432	3,094	38,338	92.5
Ca2	31,253	2,453	28,800	92.1
Ca3	41,693	2,584	39,109	93.8
Ca4	43,355	5,087	38,268	88.2
Ca5	44,645	2,901	41,744	93.5
Ca6	51,946	3,223	48,723	93.7
Ca7	35,743	2,963	32,780	91.7
Ca8	13,579	2,056	11,523	84.8
Extra contigs	67,815	4,456	63,359	93.4
Total	**371,461**	**28,817**	**342,644**	**92.2**

The heterozygous SNPs were filtered to 5,299 and 1,930 for ICC4958 and PI489777 respectively by removing SNPs which have insufficient reads in the population to accurately determine a segregating heterozygous SNP. A total of 3,363 (63.4%) and 1,559 (80.7%) of ICC4958 and PI489777 heterozygous SNPs were predicted to be valid using this approach (Table 4-4).

Table 4-4: Heterozygous SNP validation with genotypes called from ICC4958 x PI489777 population individuals.

	Heterozygous SNPs		Monomorphs		True SNPs		True SNP percent	
	ICC4958	**PI489777**	**ICC4958**	**PI489777**	**ICC4958**	**PI489777**	**ICC4958**	**PI489777**
Ca1	281	102	98	16	183	86	65.1	84.3
Ca2	178	67	65	15	113	52	63.4	77.6
Ca3	138	59	48	10	90	49	65.2	83.0
Ca4	165	49	64	11	101	38	61.2	77.5
Ca5	222	81	96	20	126	61	56.7	75.3
Ca6	268	100	107	22	161	78	60.0	78
Ca7	265	112	101	21	164	91	61.8	81.2
Ca8	55	19	15	1	40	18	72.7	94.7
ExConti	3727	1341	1342	255	2385	1086	63.9	80.9
Total	5299	1930	1936	371	3363	1559	63.4	80.7

4.3.2 Gene loss study

4.3.2.1 Estimating minimal coverage requirements

The accuracy of gene loss based on sequence read mapping depends on the sequencing depth and sequencing coverage. Insufficient coverage of sequence reads data may lead to false positive results due to missing genomic segments. To estimate the minimum coverage for gene loss study, chickpea WGS data of ICC4958 (SRX149431) was split into 1x, 2x, 5x, 8x, 10x, 12x, 15x, 20x, 30x, 40x, 50x and 60x coverage. Using each subset, lost genes are predicted and the number of genes lost plotted in Figure 4-5 against the coverage. The curve resting at the constant level is between 10x and 15x coverage, and can be considered as minimum coverage required achieving accurate results.

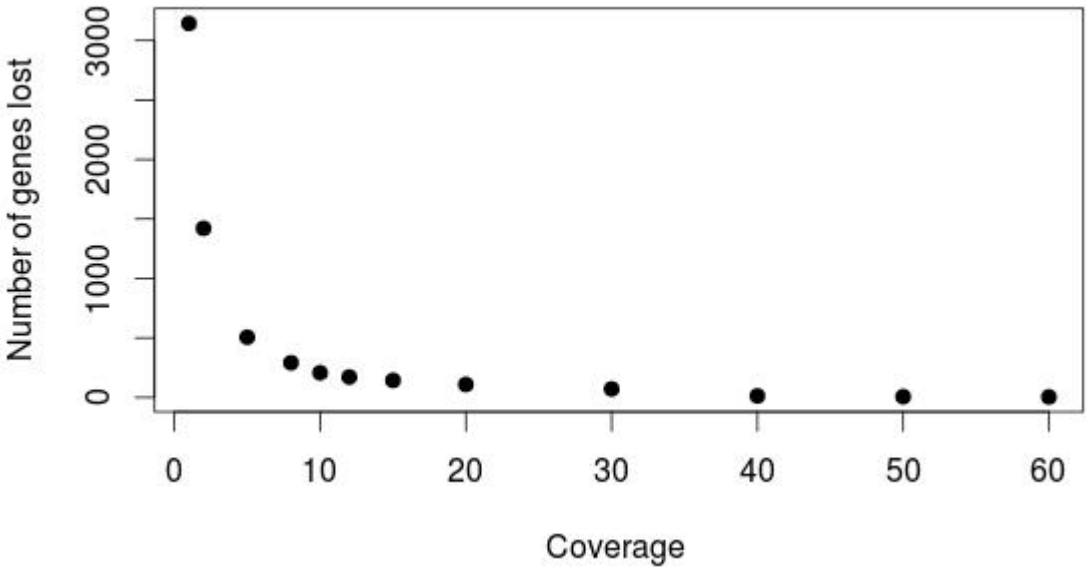


Figure 4-5: Minimum coverage estimation with number of lost genes at different coverage levels.

4.3.2.2 Mapping whole genome sequence data

Using all 47 chickpea varieties/accessions WGS data from the above SNP analysis (Appendix 3), the paired reads were mapped to released chickpea *kabuli* reference with SOAP2 v2.21 (Li *et al.*, 2009b). The mapping rate in different accessions varied from 40% to 78% (pair mapped reads), and the average mapping rate was >62% for

most accessions (Appendix 4). The difference in mapping rates is due to divergence between the accessions and reference.

4.3.2.3 Gene loss events

Based on the difference in mapping rates, we identified genes absent in accessions and found the genome coverage varied from 2.4x to 14.3x (only pair mapped reads). In all accessions, there were some genes with very low coverage, whereas in the wild type (PI48977), we observed many genes having no coverage. Using an in-house developed program SGSGeneLoss (Golicz et al., 2015), we identified 2,253 possible gene loss events from 820 genes based on read coverage from all 47 chickpea varieties (Appendix 6 and Appendix 7).

The wild variety of *Cicer* species has 225 gene loss events compared to the domesticated accessions. The difference in gene content between wild and domesticated is likely due to selection. Based on gene loss and SNP study, major genetic variation was observed on Ca4, which indicates the chromosome has significant genetic diversity. A similar diversity was also seen earlier in (Varshney *et al.*, 2013). Among the remaining accessions, CDCLuna having 210, Slasher having 181, ICCV 9683 having 160 and Pistol having 127 lost gene events, indicated that these accessions are more divergent than other accessions. Interestingly, the ILC 3279 (*kabuli* type) accession has a single lost gene representing the evolutionary closely relatedness to the reference genome. Among *desi* varieties, ICCV 3996 (10) and Slasher (181) have the minimum and maximum gene loss events and CDCLuna has the maximum lost genes in the *kabuli* type. Supporting our previous SNP analysis, the highest number of gene loss events were reported on chromosome 4 (673, 29.8%), followed by chromosome 1 (432, 19.3%) and chromosome 2 (295, 13%). The lowest number of gene loss events was found on chromosome 8 (65, 2.88%). The average gene loss density was highest in the chromosome 4 (14.3), followed by the chromosome 1 (9.2) and 7 (5.1) and lowest in chromosome 8 (1.3)

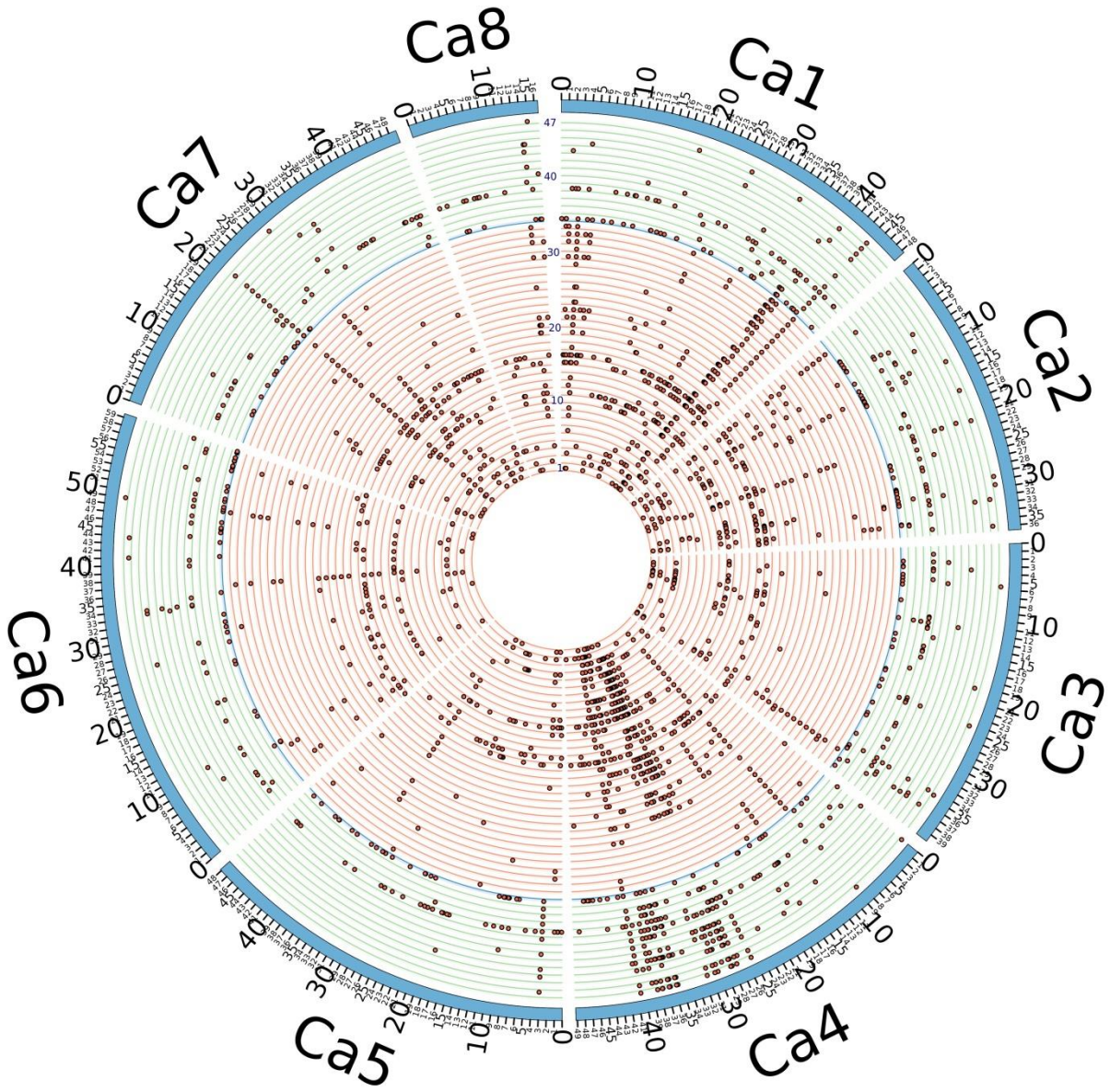


Figure 4-6: Chickpea genome wide gene loss events (red doted circles) found in 47 chickpea varieties (each circular line). Inner red coloured circles represent *desi* chickpea varieties, the blue coloured line is for the wild type, and the green coloured lines for *kabuli* varieties.

Among all the varieties in the *desi* group with 33 cultivars, 1,491 gene loss events were reported for 517 genes. From 13 *kabuli* cultivars only 537 gene loss events were found for 316 genes (Figure 4-7). Overall 820 genes showed genome wide gene variability from 47 chickpea varieties. The *kabuli* gene models were used for BLAST searches against the nt database to assign putative functions. Among 2,253

gene models, 1,983 showed function annotations and 270 did not have any functional annotation.

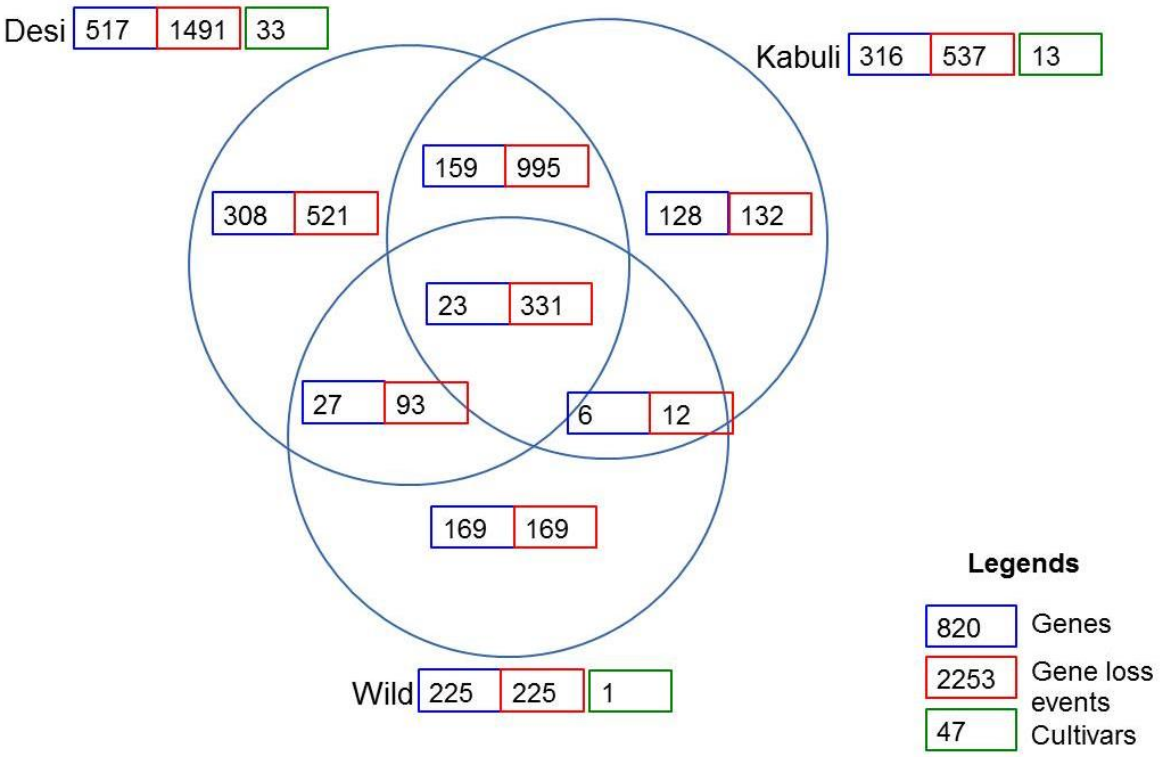


Figure 4-7: Gene loss events reported from respective genes of *desi*, *kabuli* and wild type chickpea cultivars.

4.3.2.4 Novel gene identification

Using the unmapped reads (Appendix 8), we pooled and assembled reads to produce 46,588 contigs having 66.9Mbp (55.2Mb without Ns) of novel sequence. On comparing these sequence with nt database, 326 contigs spanning 529.7Kb (excluding N's 515.7Kb) correspond to mitochondria and the remaining 46,263 contigs spanning 66.4Mb (54.7Mb excluding Ns) were identified as novel sequences. The GC content of these 66.4Mb sequences is 27.8%, which is comparably more than the GC ratio of the genome (26.0%). Annotating these novel sequences using both *ab initio* and homology based gene prediction, 611 genes were predicted with an average length of 4.5 Kb and 4 exons each. These genes were annotated with nt database and gene variability in all chickpea varieties was reported using the method for identifying *kabuli* variable genes. Among all accessions, DICC8220 has the

lowest number of lost genes (28) followed by ICCV92311 and ICC283 (31). The highest number of lost genes (84) was reported in CDCLuna (Appendix 7).

4.3.3 Genetic diversity

4.3.3.1 SNP based genetic diversity

The identification of large number of SNPs permits the characterisation of genetic diversity across chickpea germplasm. Previous studies suggest a low level of genetic diversity in cultivated germplasm compared to wild species (Ahmad *et al.*, 2010; Sudupak *et al.*, 2002; Sudupak *et al.*, 2004; Udupa *et al.*, 1999). The 611,365 homozygous SNPs between the 47 chickpea varieties were used to calculate Nei's pairwise genetic distance (Figure 4-8). This also supports our analysis where the majority of identified SNPs are between the wild type PI489777 and the cultivated varieties, with relatively little diversity within the *desi* and *kabuli* cultivars

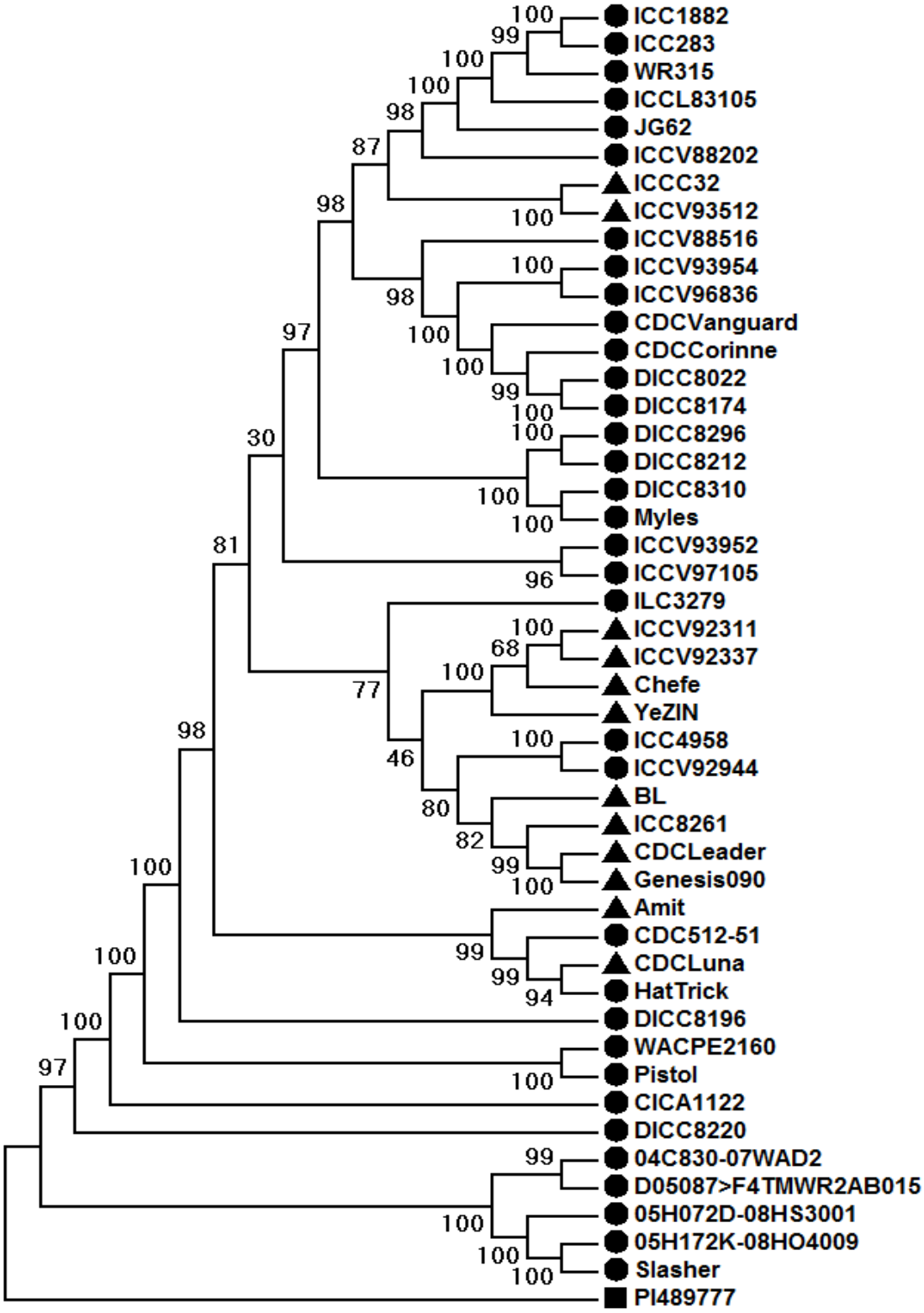


Figure 4-8: Distance based genetic diversity analysis showing the genetic relationship between 47 chickpea varieties (circles=*Desi*, triangles=*Kabuli*, square=Wild type).

We investigated population structure using homozygous SNPs between *desi* and *kabuli* types without introducing any *a* priori classification. We used a variational algorithm implemented in the software fastStructure (Raj *et al.*, 2014). Using the log-likelihood method implemented in fastStructure resolved four clusters (k=4), suggesting that there are four groups of cultivars (Figure 4-9).

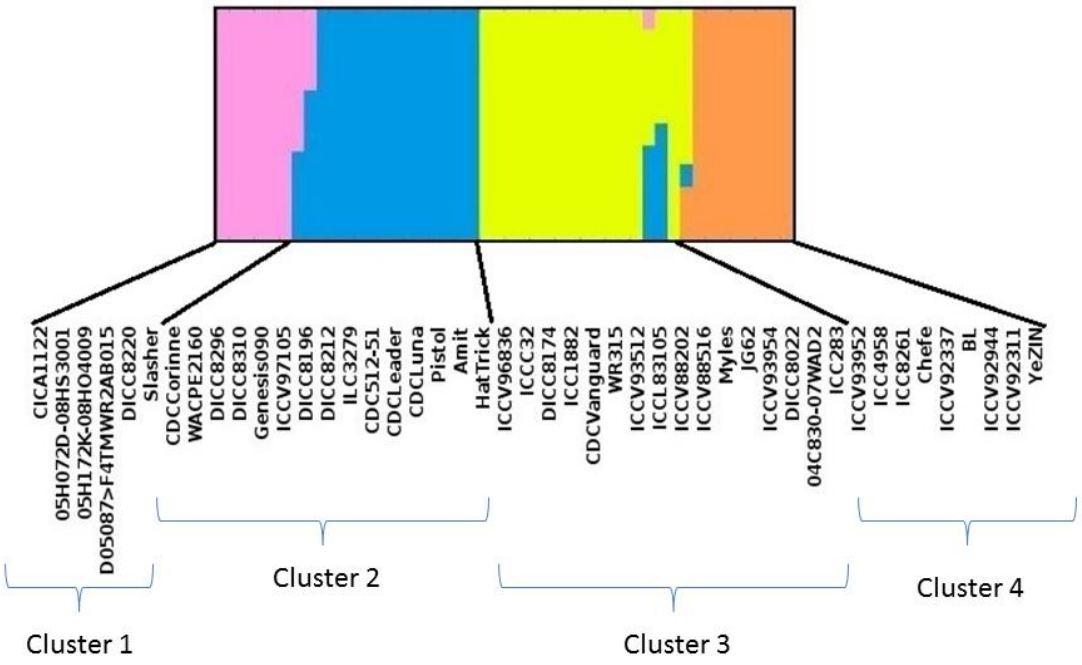


Figure 4-9: Population structure inferred using *desi* and *kabuli* whole genome SNP dataset for K=4. Each genotype is represented by a horizontal line, which is partitioned into coloured segments that represent the estimated membership fractions in the K clusters.

The results indicate that CDCCorinne and WACPE2160 belonging to cluster 2 show partial membership with cluster 1. Similarly, DICC8022 and 04C830-07WAD2 from cluster 3 and ICCV93952 of cluster 4 have partial cluster membership, indicating hybrids or intermediate forms. Among the chickpea varieties used for population structure analysis, the above 5 chickpea accessions have higher genetic variability than other varieties. The population structure, diversity and genetic distance analysis

support the mixed use of both *kabuli* and *desi* in breeding varieties. Based on earlier studies with different markers, cultivated chickpea was found to be more closely related to *C. reticulatum* than *C. echinospermum* (Croser *et al.*, 2003; Iruela *et al.*, 2002; Nguyen *et al.*, 2004; Roorkiwal *et al.*, 2014b).

4.3.3.2 Gene loss phylogeny

On combining 820 variant genes from CDCFrontier *kabuli* reference genome with 611 genes from novel sequences, we investigated the evolution of gene flow between 47 chickpea varieties. On comparing genes between any two varieties, a gene that is absent in one variety could have either been lost in the first variety or gained in the second variety. To detect the genic relationship, a phylogenic tree was constructed based on gene content (i.e., presence or absence of gene) from all 47 varieties. The phylogenetic tree (Figure 4-10) shows most of the *desi* type varieties clustered on the first half of the tree, and the second half presents a mixture of *desi* and *kabuli* types. Few varieties like PI489777 (Turkey), CDC Luna (Canada), Slasher (Australia), Pistol (Australia), ICCV96836 (ICRISAT) and Hat Trick (Australia) are too distinct to cluster with other domesticated accessions (Appendix 8). The majority of gene variation observed in these accessions may be due to their different evolutionary histories. We also found that wild chickpea have retained diversity that seems to have been lost in the cultivated chickpea varieties. The phylogenetic tree suggests that differential gene loss/gain may have played important roles in the evolution of some chickpea accessions. This gene loss study provides a valuable resource for recovering useful genes from diverse chickpea types.

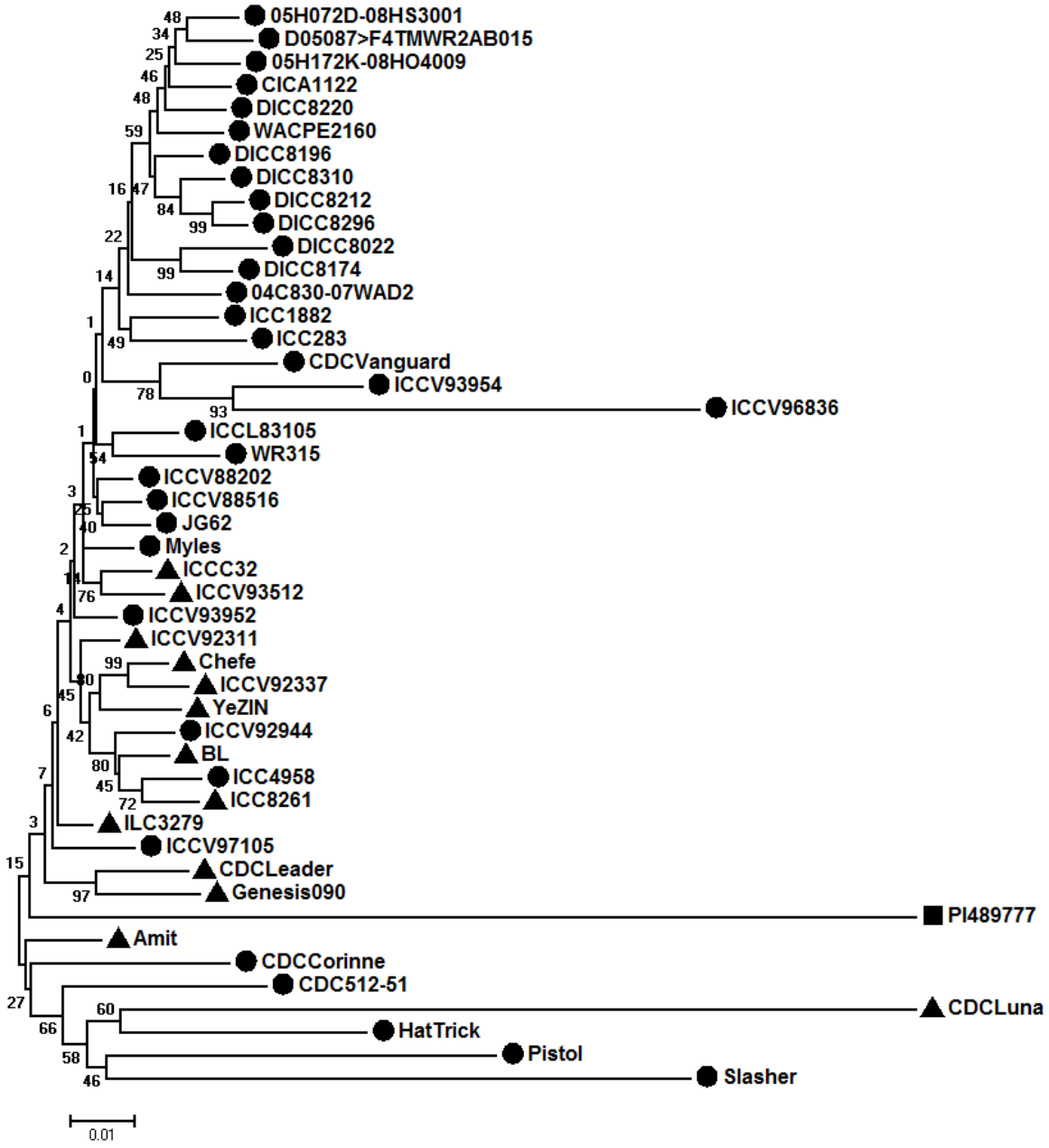


Figure 4-10: Phylogenetic analysis using chickpea variable genes. Distance based phylogeny constructed based on the presence or absence of variable gene gain/loss in each chickpea variety.

4.4 Conclusions

This study reports a large number of SNPs. These SNP markers can be used to tackle the issue of narrow genetic diversity in the domesticated gene pool by facilitating genetic enhancement and breeding. Nucleotide diversity analysis provides an assessment of genetic based relatedness between a wide range of chickpea

cultivars. The gene loss study revealed that variation in gene content is considerable. These single nucleotide variations or gene variations are likely to have a pronounced effect on phenotypic diversity and represent a crucial target for the assessment of genomic diversity. The known genetic structure and relatedness between 47 chickpea accessions may be useful in finding suitable parental combinations for developing new segregating mapping populations for different traits of interest to chickpea breeders. Additionally, the SNPs can be chosen as polymorphic markers from existing mapping populations which can be then used for constructing genetic maps and trait association studies. The ability to sequence many chickpea lines will lead to the increase in sequence based genotyping for diversity assessment, genome structure validation and gene trait association.

These results provide a valuable resource for recovering useful alleles and genes from wild chickpea and also aid in carrying out future in depth studies of population genetics, marker-assisted breeding and gene identification studies in chickpea.

5 Chickpea genomic resources (Cicer.info)

5.1 Introduction

Cicer.info (www.cicer.info) database project provides a bioinformatics framework to organise biological information. It is a comprehensive source of available genomic and associated tools developed for *Cicer* research. The availability of the draft genomes sequence enables analysis of huge amounts of data and to record variation between the *Cicer* cultivars. The availability of the two chickpea reference genomes allows the discovery of large numbers of SNPs with genomic locations which can be used for trait mapping and candidate gene identification. However managing, visualising and extracting information to support trait discovery from these datasets is a challenge (Batley and Edwards, 2009). To address this, we have developed an online chickpea genomics resource: called cicer.info, which provides a variety of web-based system for assessing both *desi* and *kabuli* chickpea reference genomes together with their associated annotations. The database was built on a very general and compatible platform to easily customise and integrate data. Cicer.info database provides a variety of sequence similarity search services, including BLAST and TAGdb.

5.1.1 Genome viewers

There are a number of viewers available for visualisation of genome at nucleotide level. The primary function of genome viewers, mainly focusing on molecular markers and other features, is the ability to compare multiple sequences (sequence reads) in the form of alignment (Chapter 1). Genome browsers are mainly graphical interface for display of information from alignment files against a reference. It enables researchers to visualise and browse entire genome with associated annotation, gene predictions, structural variations, proteins and regulatory elements. The associated data for reference genome is usually from multiple diverse sources linked with genome coordinates. The visual analysis shows the nucleotide context of variant occurrence and other factors of breeding interest. Many software packages are available with unique visualisation features as discussed in Chapter 1. Artemis is a genome viewer available from Sanger Institute, developed in Java and compatible to all OS platforms and it allows editing the sequence annotations and features

(Rutherford *et al.*, 2000). It also supports EMBL, GENBANK and FASTA format. Apollo is similar to Artemis developed on same platform and is a part of the GMOD project accessible online (Lee *et al.*, 2013). NCBI Genome workbench is a workbench of tools that allow user to organize and retrieve sequence data from NCBI database. This software allows viewing sequences as flat sequence files, phylogenetic trees and alignments. UCSC genome browser is an online genome browser providing access to genome sequence data from a variety of major model organisms (Rosenbloom *et al.*, 2015). It is an open-source, web-based tool suite built on top of MySQL for rapid querying. UCSC site also hosts a set of genome analysis tools including mining the information from database. JBrowse is another genome browser developed with Java and HTML5 platform and comparatively fast and very light server resource requirements (Skinner *et al.*, 2009). It supports a wide range of input formats. Integrative Genomics Viewer (IGV) is another browser handles large data sets while providing a smooth user experience at all levels of genome resolution (Thorvaldsdottir *et al.*, 2013). It supports the iteration of array-based, NGS, and phenotypic data.

Generic genome browser (GBrowse) is a web-server application implemented in Perl and capable of handling NGS data in binary map format (BAM). GBrowse allows easily upload or download the genomic features and also able to handling huge data either in MySQL or flat file format to visualise features from chromosome level to base level is an additional feature. Among the wide range of availability of genome browsers, in this study we have used GBrowse (Stein, 2013) as online genome browser.

5.1.1.1 GBrowse

GBrowse is a combination of database and interactive web pages for manipulating and displaying genomic features (Stein, 2013). It is a Perl-based server and can be run on Linux, Mac OSX and Windows PC platforms. It can be easily integrated with other datasets at the database level. Since it is not dependent on any specific data model or pipeline, it is suitable for any research group that needs to maintain a set of sequence annotations. It can run locally on a less powerful computer or even on a high-end server to make it available online for research community. GBrowse can be used for displaying genomic features on small genomes, as easy as on large genome such as wheat. The main advantage of GBrowse is the usage of different

adaptors connecting to various types of databases. Usually GBrowse is set up with a flat file adaptor, suitable for genomes having no more than 20,000 features, or using a relational database adaptor for large genomes. GBrowse provides both simultaneous bird eye and detailed views of the genome (Figure 5-1). The zoom in/out option allows users to view genomics features at different resolution. The user friendly interface allows navigating across the length of chromosome or contig sequence.

The browser window can jump to another specific region of the genome by entering the co-ordinates in the provided search box (Landmark or Regions box). Features are categorised as customizable tracks that can be ordered by various attributes. In GBrowse “Select Tracks” tab lists all available tracks that are available to visualise in browse and can either be selected or removed. Details of features can be made available either in popup balloons or in detail pages. The search option in GBrowse allows searching by specific chromosome region, gene, annotation id, name or comments across the database. Third party annotations can easily uploaded to GBrowse in GFF formats. Data can be downloaded from a track by selecting download track data from the drop-down list and any data associated with the current view will be downloaded in GFF3 format. There are more options provided to download as per requirement of the user. GBrowse also enables a user to share screenshot with others by bookmarking or exporting the current link. GBrowse is capable of displaying density plots and individual reads based on the standard SAM/BAM file format.

Figure 5-1: GBrowse viewer web interface.

5.1.2 Genomics features

Genomic features are (structural/functional) information of a genomic location in form of gene co-ordinates, annotations, sequence read alignments, ontology association, database cross reference, intron/exon regions, markers and restriction sites.

5.1.2.1 Synteny genes/markers

The term synteny has been expanded for homology studies. Synteny describes the physical location of genetic loci on the chromosome relating to the order in closely related species/individual. In other words, synteny can be defined as the conservation of blocks of order within two sets of chromosomes that are compared with each other. Where a reference genome is not available, a genetic linkage map represents the order of the known molecular genetic markers along a given chromosome of a species. This provides an overview of the genome's organisation. Comparative genetic study is a valuable technique to identify similarity and difference between species. Synteny studies also help in identifying candidate genes underlying a QTL, genome comparisons, identify syntenic regions and view comparative genetic maps in selected species (Duran *et al.*, 2009d).

5.1.2.2 Gene models

Annotations include structural annotation (consists of the identification of genomic elements such as ORF, coding regions, regulatory motifs) and functional annotation (consists of attaching biological information to genomic elements such as biochemical, biological and expression details). Computational gene prediction has been an active area and commonly used for wide range of genomics studies. Gene predictions were mainly performed using *ab initio* methods, such as Genscan (Burge and Karlin, 1997) and Gene(Reese *et al.*, 2000), or homology based (or alignment-based) methods such as Procrustes (Gelfand *et al.*, 1996) and Genewise. To attain the advantages of both methods, hybrid approaches were introduced to increase the accuracy of gene predictions. Recently, machine learning techniques have been successfully applied to solve various important biological problems such as genome annotations (Yip *et al.*, 2013), pattern recognition (de Ridder *et al.*, 2013), disease associated genes (Lopez-Bigas and Ouzounis, 2004; Sun *et al.*, 2009) and discovery of gene-gene interaction in disease data (Okser *et al.*, 2013; Upstill-Goddard *et al.*, 2013). The process of attaching biological information to the gene models is the genome functional annotation and it is done using annotation tools or annotation pipeline. The basic method of annotating is by using BLAST for finding similarities of predicted genes with well-studied and annotated genes (or genomes).

5.1.2.3 RNA (miRNA, snRNA, rRNA and tRNA)

A functional RNA molecule that is not translated into a protein is non-coding RNA and the DNA from which these are transcribed is often called an RNA gene. RNA is an important biomolecule involved in catalysation, post-transcriptional modification and gene regulation. RNA prediction tools FRUUT (Milo *et al.*, 2013) is a pairwise RNA structural alignment tool based on the comparison of RNA trees. GraphClust (Heyne *et al.*, 2012) is a fast RNA structural clustering method of local RNA secondary structures. These clusters are further refined with LocARNA and CMsearch. INFERNAL(Nawrocki *et al.*, 2009b) is a software usually used for miRNA and snRNA gene prediction using Rfam database.

5.1.2.4 Markers

Genetic markers are fragments of DNA that are associated with known location within a genome. These markers can be as short as SNP or as long as minisatellites and they are mainly applied for species identification, genetic variation, population

structure and comparative studies. Even though there are many genetic markers (discussed in Chapter 1), because of their abundance, high throughput markers like SNPs are mainly used.

5.1.2.5 Next generation sequence data

GBrowse version 2.0 supports NGS data (DNA-seq and RNA-seq data) by reading data directly from BAM sequence alignment files. The read alignment can be displayed at multiple resolutions from a whole-chromosome coverage histogram to individual base pairs. Alignment files can be uploaded to the browser.

5.1.3 Sequence similarity search

Sequence similarity search is a method of searching a query sequence. A sequence alignment may be a consequence of functional, structural, or evolutionary relationships between the sequences. Based on the algorithm and type of sequence, there are different tools for similarity search like FASTA, BLAST, LALIGN and SSEARCH are local alignment tools and clustalx/w, Needle, Stretcher are global alignment tools. BLAST has become the commonly used sequence similarity search tool. The BLAST program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

5.1.4 TAGdb search

TAGdb (Lai *et al.*, 2012a) is an online database that enables researchers to identify similarity between paired read sequences and submitted query sequence. Identified tags can be used design oligonucleotide primers for the PCR amplification of the region in the genome.

5.2 Methods

We have developed a web-based application for displaying genomic annotations and other features associated with chickpea genome. For the chickpea research community, features of GBrowse include the ability to scroll and zoom throughout the genome, and also to look for region of genome by searching for landmarks or performing a search based on feature keywords. This genomic resource also enables users to compare the feature at a region of interest and also turn on or off the associated track. The GBrowse also plugin to web-based BLAST tools to visualise the similarity hits.

5.2.1 GBrowse installation

GBrowse was chosen to display the chickpea genome. GBrowse 2.0 was installed on Ubuntu 12.04.2 LTS. The GBrowse was installed by following the instruction provided at http://gmod.org/wiki/GBrowse_Ubuntu_HOWTO.

5.2.1.1 GBrowse configuration file

A GBrowse configuration file is an important component in GBrowse for configuring what, where and how to display the data. This file specifies the location of database, flat files in form of GFF3 files and also other display settings including feature tracks. GBrowse can automatically create configuration file for user uploaded GFF3 files (or Feature File Format - FFF), and place file at proper location. Such automatically created configuration files are saved in the configuration directory specified at the time of installation. Various feature formats like BED, GFF, GFF3 and BAM format with different display styles are accepted for uploading into GBrowse as mentioned at http://www.cicer.info/gbrowse2/annotation_help.html. GBrowse.conf is located at /etc/gbrowse/ is used to store the common configuration settings.

All default settings are used for GBrowse interface and both databases “Chickpea *Kabuli*” and “Chickpea *Desi*” are used in one interface. Both publicly available chickpea type genomes and its associated genome feature files were loaded into MySQL database.

To creat a database

bp_seqfeature_load -c -u <username> -p <password> -d dbi:mysql:<databaseName>:cicer *.fasta *.gff3

To loading features into database

bp_seqfeature_load -u <username> -p <password> -d dbi:mysql:<databaseName>:cicer *.fasta *.gff3

5.2.1.2 Glyph library

A pictorial presentation of genomic feature is defined as “glyph” in the configuration file. GBrowse gives the control to the user to choose any glyph styles by selecting any shape from settings option of any track. It is possible to add more glyphs by editing the Glyph library file. The GBrowse package includes a perl script

getNewGlyphYAML.pl that can generate the YAML content to be inserted into the Glyph library.

5.2.2 Genomic features

GBrowse can handle data as file-based database or relational database such as MySQL. GBrowse requires genome annotation files in GFF3 format (Generic Feature Format Version 3), which is a standard and commonly used format for representing genomic feature. GFF3 file represents each feature in tab-delimited plain text file consisting nine columns specifying features and corresponding genomic co-ordinates. GBrowse allows users to directly supply their own GFF3 files to the GBrowse sever. On successful upload, GBrowse automatically validates and move the genomic features in the GFF3 file into GBrowse database.

5.2.2.1 Synteny between crops

All genome annotations for *A. thaliana*, *O. sativa*, *M. truncatula* were downloaded from Phytozome database and sequence similarity search was performed against Chickpea *desi* and *kabuli* gene models using BLAST tool. The tabular output format was chosen for further filtering.

blastx -db uniref90 -query Cicer_arietinum_pseudo-chr_v1.0.gene.cds.fa.5 -evalue 1e-10 -outfmt 6 -out Cicer_arietinum_pseudo-chr_v1.0.gene.cds.fa.5.blastx -num_threads 4

Few chickpea genes may have multiple hits and they are filtered with following command to have a top best hit.

sort -k1,1 -k12,12rn AllblastTabular.blastx |awk 'BEGIN [last=""] [if ($1!=last)[print $0; last=$1]]' >filterblastTabular.blastx

This tabular filtered BLAST output was converted to gff3 format using a perl script “BLAST_to_gff.pl”.

perl BLAST*_to_gff.pl -i filterblastTabular.blastx > FilterblastTabular.blastx.gff*

The uniprot descriptions are assigned to similar chickpea genes by using addFastaDescriptToGff3.pl script.

perl addFastaDescriptionToGff3.pl -g FilterblastTabular.blastx.gff -i All_uniprotIds.txt

The output gff3 file with chickpea gene coordinates and associated annotations were uploaded into GBrowse. Above pipeline was also run on nr database, to assign synteny genes from wide range of genomes.

5.2.2.2 Synteny within Cicer species

The synteny between *desi* and *kabuli* genomes was performed with SyMap v4.2 (Soderlund *et al.*, 2011). The similar sequence regions showing the synteny between two genomes were extracted with MySQL query system provided within the software.

5.2.2.3 Gene annotation feature

Both *desi* and *kabuli* genome genes were annotated using the UniRef90 database. Complete genes of individual genomes were searched against UniRef90 database using BLAST and genome annotations were assigned to each gene by using the same pipeline mentioned in “synteny between crops”.

5.2.2.4 Repeats markers

Repeat sequences were identified with MISA (Thiel *et al.*, 2003) and TRF finder (Benson, 1999) and transposable elements were predicted with Repeatmasker (Bergman and Quesneville, 2007).

5.2.2.5 SNP markers

Whole genome sequence reads of 47 chickpea varieties were mapped on *kabuli* reference individually using SOAP2 (Li *et al.*, 2009b) software and SNPs were identified with SGSautoSNP (Lorenc *et al.*, 2012) software. More details are discussed in Chapter 2 & 4.

5.2.2.6 Non-coding RNA genes

The tRNA genes were predicted by tRNAscan-se (Schattner *et al.*, 2005) and remaining RNA genes predicted with cmsearch (Nawrocki *et al.*, 2009a) software against the Rfam database (Burge *et al.*, 2013).

5.2.2.7 Isolated Chromosome Sequence (ICS) mapped reads

The chickpea chromosomes were isolated using flow sorting method and sequenced them individually (Ruperao *et al.*, 2014b; Vlacilova *et al.*, 2002). For *kabuli* genome all chromosomes were isolated, exceptionally chromosome 4 and 7 were group isolated and sequenced and in case of *desi* genome, chromosome 3, 5 and 8 were

sequenced. More details of this data analysis were discussed in Chapter 2 (Ruperao *et al.*, 2014b).

5.2.3 BLAST search tool

BLAST portal developed in Perl-CGI providing the web-interface at front end and plugin to BLAST tool at back end. Based on the BLAST output the links on summary page were directed to show corresponding position on GBrowse (Stein, 2013).

5.2.4 TAGdb tool

TAGdb is another web-based application to visualise NGS paired sequence tags that share identity with submitted query sequence (Marshall *et al.*, 2010). A total of eight accession WGS data, 4 for each *desi* and *kabuli* types, and ICS data of both *desi* and *kabuli* types were made publically available.

5.3 Results and discussion

5.3.1 Database contents

It is a challenge to visualise the genome and its associated genomic features in raw data format, and especially for non-bioinformatics researchers. “Cicer.info” provides access to a variety of web-based systems hosting chickpea genome and genomic data to support chickpea research and crop improvement. It is integrated database resource which includes multiple web-based applications. Both *desi* and *kabuli* genome assemblies were made available in GBrowse interface with a BLAST search portal. Chickpea next generation genome sequence data can be searched using TAGdb. Users can also easily download all the raw data via this portal that includes both chickpea genomes and their associated genome feature files in GFF format, and isolated chromosome sequence (ICS) reads mapped in bam format files.

5.3.2 A *Cicer* genome viewer

The application of NGS technology and advanced bioinformatics tools has enabled the assembly and annotations of the genes and genomic variations study of chickpea varieties. The chickpea draft reference genome of *desi* (Jain *et al.*, 2013) and *kabuli* (Varshney *et al.*, 2013) from NIPGR (India) and ICGGC are hosted in a GBrowse2 database at cicer.info for public access prior to publication. GBrowse2 is a user-friendly search and visualisation tools for genome sequence data and

associated annotation. The statistics of both *desi* and *kabuli* reference genome assemblies were also made available on web portal (Table 5-1).

Table 5-1: *Desi* and *kabuli* reference genome statistics.

Features	***Desi* draft genome**	***Kabuli* draft genome**
Total assembly size (Mb)	456	532
Number of scaffolds	32,935	7,163
Minimum reported scaffold length (bp)	1,000	1,000
Maximum scaffold length (Kbp)	23,376	59,460
Average scaffold length (bp)	13,857	74,311
N50 length (Kb)	106	39,989
GC content (%)	25.6	30.8
Genome captured in pseudomolecule (Kbp)	124,386	347,247
Protein coding genes	27,571	28,269
Average gene length (bp)	3,122	3,055
Average coding sequence length (bp)	962	1,166

In addition to draft genome we have provided annotated gene models, repeat sequences, structural RNAs and restriction sites. More than 600,000 predicted SNPs between 47 *kabuli* varieties have been identified and added to the *kabuli* GBrowse database as a specific track facilitating the genetic association of agronomic traits with candidate genes (Chapter 4). The genome viewer consists of three sliding windows, the chromosome/scaffold/contig overview window, the region window and a detail window. The detail window contains tracks showing different genomic features. The powerful function of GBrowse enables user to close, re-arrange track ordering and saving tracks.

This GBrowse genome viewer enables researchers to interact with the chickpea genome from the chromosome to nucleotide level in a user friendly manner. The annotated gene and SNP data supports chickpea researchers and breeders in their

identification of the genomic basis of important traits and provides tools for the selective breeding of allelic variants.

5.3.3 Search tool

The GBrowse search tool provides the ability to access information about nomenclature, genome location and function of gene/markers. This functionality provides user to jump to specific landmarks or regions in the genome. The search function also enables users to find genes using gene ID or name, or even keywords in the functional annotation of a gene. When keywords are searched, the result is returned as a list of keyword occurrence with its corresponding position on pseudomolecule/scaffold as shown in Figure 5-2.

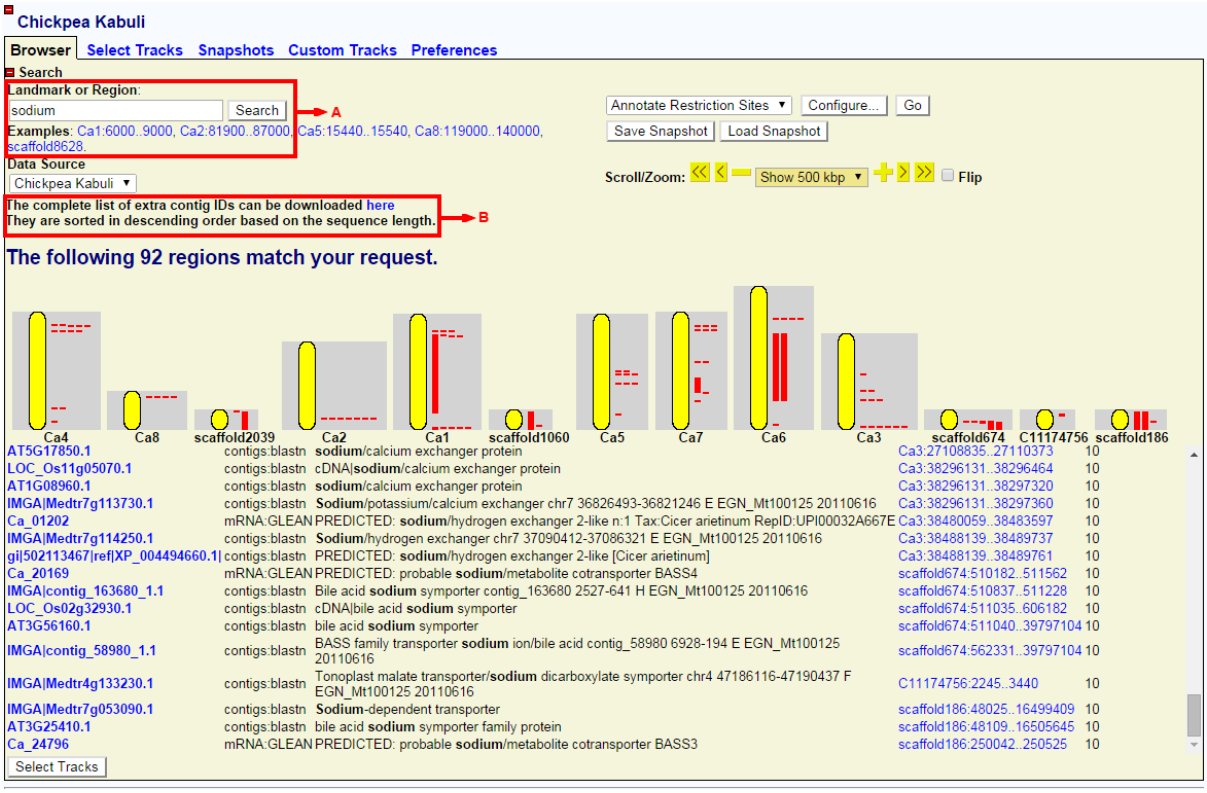


Figure 5-2: GBrowse search functionality. GBrowse (example) search sodium keyword on entire genome and can visualise the keyword occurrence at respective sequence positions. Box A shows the search text box and box B show the link to list available contigs.

5.3.4 Genomic features

5.3.4.1 Gene models

The desi genome has 27,571 gene models and only 8,745 are corresponding to pseudomolecules, whereas, among total 28,269 genes from *kabuli* genome, 23,550 genes are from chromosomes. These released gene models were annotated with the UniRef90 database using BLAST and an in-house annotation pipeline. The detail annotation of each gene is made available on GBrowse associated with gene models. The gene link is provided with a popup balloon, allowing user to choose to view either UniRef90 annotations or default GBrowse detail page option (Figure 5-4). These details help researchers to study genes in detail about the structural or functional properties.

5.3.4.2 Non-coding RNA genes

This group shows the different type of RNA such as miRNA, snRNA, rRNA and tRNA sequences predicted. In *desi* genome, a total of 627 tRNA, 249 rRNAs and in *kabuli*, 420 miRNA, 478 rRNA, 647 snRNA, 700 tRNAs were made available with reference genome under RNA group track.

5.3.4.3 SNPs

Based on earlier genome assessment (Chapter 2), *kabuli* genome carries few mis-assemblies compared to *desi* genome. So, using *kabuli* genome as reference, 47 chickpea varieties were mapped and SNPs were predicted. Among these 47 chickpea varieties, 29 published varieties are used with additionally 18 Australian chickpea varieties WGS data. This complete dataset consisted of *desi*, *kabuli* and wild type varieties (Appendix 3). On *kabuli* genome, SNP track host more than 600,000 candidate SNPs identified from 47 chickpea varieties of *kabuli* type using SGSautoSNP 20 (Chapter 4). These SNPs were categorised into synonymous (in blue colour), non-synonymous (in red colour) and non-coding region (in green colour).

5.3.4.4 Repeats

Desi genome has 210Mb of the draft genome sequence was identified as interspersed repeat sequences, constituting 40% (Jain *et al.*, 2013) and the *kabuli* genome is composed of 49% of transposable elements and unclassified repeats (Varshney *et al.*, 2013). The most abundant tandem repeats in the *kabuli* genome

assembly are 163bp, 100bp and 74bp unit repeats. Based on the type of repeats, this group has four tracks named as transposons, TRF, SSR (by MISA) and TE proteins. More details of individual features are made available on GBrowse details pages.

5.3.4.5 Synteny

There is little diversity between *desi* and *kabuli* types with the wild-type showing many differences to both (Chapter 2 & 4) (Varshney *et al.*, 2013). A synteny study between *desi* and *kaubli* was performed and made available as the synteny track in both databases. Clicking on a synteny feature in the *kabuli* genome viewer can directly link to the synteny location in the syntenic region in the *desi* genome viewer and *vice versa*. Chickpea genes showing synteny with other major model plants can be visualised in the O.*sativa*, A. *thaliana* and *M. truncatula* tracks. The links associated with these feature allows user to be redirected to the gene location in their corresponding genome browser in Phytozome (Goodstein *et al.*, 2012).

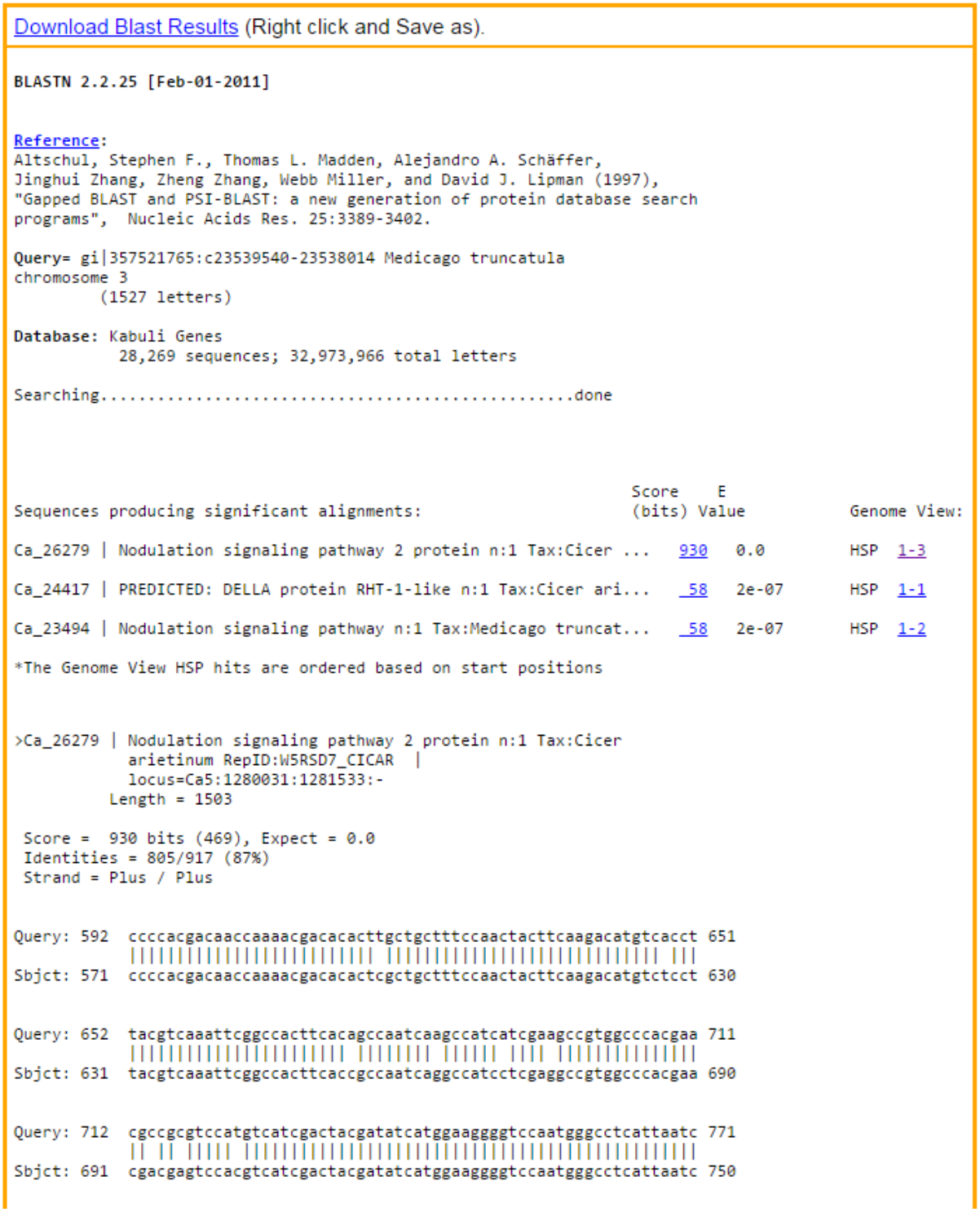
5.3.4.6 Isolated chromosome sequence (ICS) read alignments

The ability of isolating individual chromosomes and sequencing through high-throughput has enabled to assess *desi* and *kabuli* reference genomes at base pair levels. In the earlier assessment *desi* reference genome was poor in quality compared to the *kabuli* reference genome. However, *kabuli* genome has more than 60 mis-assemblies regions that were wrongly placed in non-corresponding chromosomes (Chapter 2). For both *desi* and *kabuli* reference genome, the mapped data is made available in GBrowse.

5.3.5 Chickpea BLAST portal

A BLAST portal enables sequence similarity searches of both *desi* and *kabuli* genomes, with results displayed in the GBrowse2. For both *desi* and *kabuli* genomes, not only draft genome database is made available but also the annotated genes are made accessible to search via a user-supplied query sequence. The BLAST tool was also made flexible to perform a search against either chickpea *desi or kabuli* genes alone. A BLAST search against gene models produces output as a summary page listing predicted gene annotation. DNA or protein query sequence can be uploaded or pasted in the web-based input text box in FASTA format. The results are displayed in BLAST summary page showing list of hits ordered by the

alignment score and followed by alignments showing the sequence similarity between the query and subject sequence. This BLAST summary page also display two different links, one link direct to the respective alignment and other one to the BLAST hit showing corresponding alignment to reference genome. To further analysis or to preserve the results, the BLAST summary page was made available to download through the link provided on top of the BLAST summary page (Figure 5-3).



Download Blast Results (Right click and Save as).

```
BLASTN 2.2.25 [Feb-01-2011]


Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|357521765:c23539540-23538014 Medicago truncatula
chromosome 3
         (1527 letters)

Database: Kabuli Genes
           28,269 sequences; 32,973,966 total letters

Searching.................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value           Genome View:

Ca_26279 | Nodulation signaling pathway 2 protein n:1 Tax:Cicer ...   930   0.0         HSP  1-3

Ca_24417 | PREDICTED: DELLA protein RHT-1-like n:1 Tax:Cicer ari...    58   2e-07       HSP  1-1

Ca_23494 | Nodulation signaling pathway n:1 Tax:Medicago truncat...    58   2e-07       HSP  1-2

*The Genome View HSP hits are ordered based on start positions


>Ca_26279 | Nodulation signaling pathway 2 protein n:1 Tax:Cicer
            arietinum RepID:W5RSD7_CICAR  |
            locus=Ca5:1280031:1281533:-
          Length = 1503

 Score =  930 bits (469), Expect = 0.0
 Identities = 805/917 (87%)
 Strand = Plus / Plus


Query: 592  ccccacgacaaccaaaacgacacacttgctgctttccaactacttcaagacatgtcacct 651
            ||||||||||||||||||||||||| |||||||||||||||||||||||||||||| |||
Sbjct: 571  ccccacgacaaccaaaacgacacactcgctgctttccaactacttcaagacatgtctcct 630


Query: 652  tacgtcaaattcggccacttcacagccaatcaagccatcatcgaagccgtggcccacgaa 711
            ||||||||||||||||||||||| |||||||| |||||| |||| ||||||||||||||
Sbjct: 631  tacgtcaaattcggccacttcaccgccaatcaggccatcctcgaggccgtggcccacgaa 690


Query: 712  cgccgcgtccatgtcatcgactacgatatcatggaaggggtccaatgggcctcattaatc 771
            || || ||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 691  cgacgagtccacgtcatcgactacgatatcatggaaggggtccaatgggcctcattaatc 750
```

Figure 5-3: BLAST **summary page. Image illustrating the** BLAST **summary page showing** BLAST **hits, links and alignment view.**

The hyperlink in BLAST summary page directs to the GBrowse view to show the BLAST hit on corresponding position of subject sequence (Figure 5-4). The results are displayed in three windows: the overview window, region window and details

window. The reference view can be navigated by zooming or scrolling through the sequences. Several tracks of annotation are available, including SNP, SSR, CDS, 6 frame translations, genes with UniRef90 annotations, synteny between *desi* and *kabuli*, non-coding RNA genes and ICS mapped reads. All of these features tracks can be expanded by clicking the associated plus button, and each feature provide a link to show the feature details. This genomic resource allows the investigation of genes surrounding a locus of interest than was previously possible in chickpea.

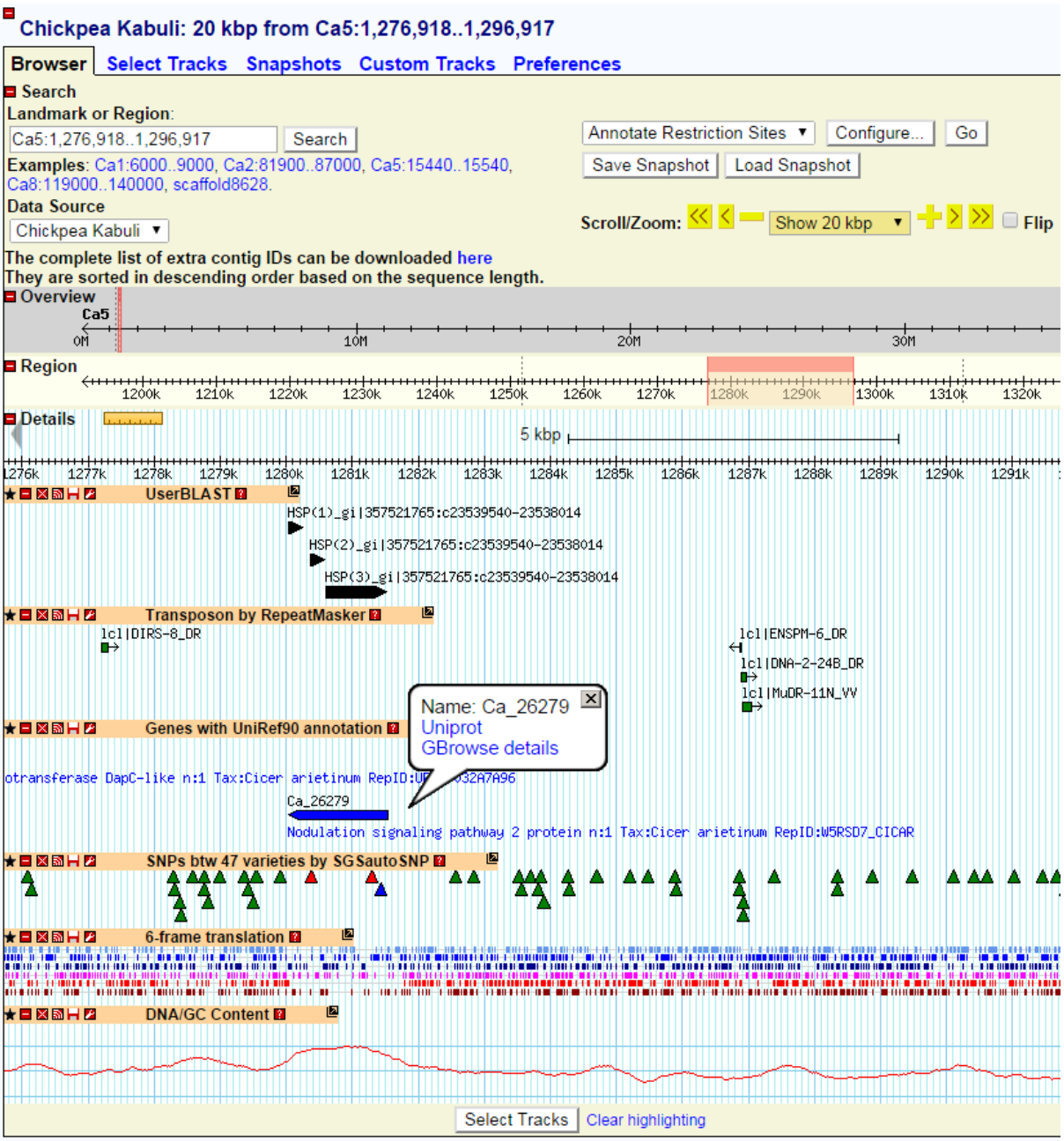


Figure 5-4: BLAST **similarity hits in GBrowse. Image illustrating the available features on *kabuli* reference genome and also showing** BLAST **hists.**

5.3.6 Chickpea data on TAGdb

TAGdb is an online database system designed to identify and visualise the NGS paired sequence tags that share identity with a query sequence (Marshall *et al.*, 2010). The TAGdb interface request a FASTA query sequence and an email to notify the results. The results are stored, so users can retrieve previous query searches. User can select the species of interest from the species list; this will populate the list of available libraries in short paired-read library list. Currently, for each of *desi* and *kabuli* types, four accession WGS data and ICS were made available. On starting the search, TAGdb sends an E-mail to the user stating that the job has started successfully and provides a link to the results web page. On completing the search, TAGdb sends another E-mail to confirm completion, together with a link to the results.

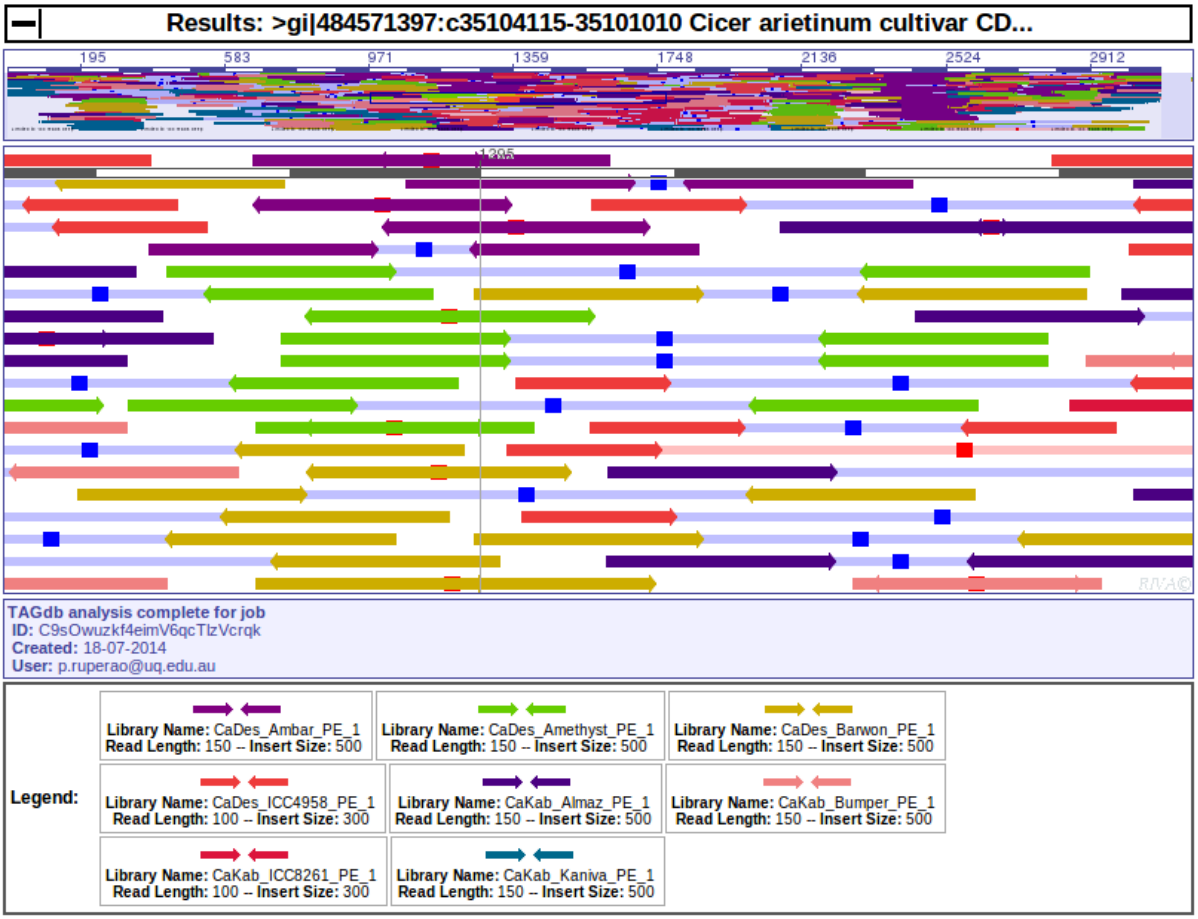


Figure 5-5: TAGdb query sequence read alignment view. TAGdb result showing sequenced reads from different cultivars mapped to a query sequence.

TAGdb result page has two windows an overview and zoomed region of the read alignments (Figure 5-5). Different coloured arrows represent the different datasets, and paired reads are connected by a line, with a blue rectangle spot in middle represents the expected orientation and paired read distance. All aligned reads are viewed as a table or can be downloaded as a multi-FASTA format file (from provided link) for further analysis. This web-based tool provides access to chickpea WGS data being produced by new sequencing technologies. The identification of a large number of reads alignment may concludes the existence of a given query sequence in the chosen cultivar/variety. Case study 1 demonstrates the application of this tool for gene discovery in various accessions without the requirement of a reference genome.

5.4 Case study

5.4.1 Study 1

Aim: To find the presence or absence of genes/sequence (Polycomb group protein embryonic flower MTR_5g013150) in various chickpea accession using TAGdb.

Method: In TAGdb, after providing the optional e-mail, a query sequence of interest (MTR_5g013150) was uploaded and from list of species, chickpea was selected to populate the list of available short pair-read library sequence data. A total of 8 accessions listing whole genome sequence datasets were selected (4 *desi* (Ambar, Amethyst, Barwon and ICC4958) and 4 *kabuli* (Almaz, Bumper, ICC8261 and Kaniva) and the job was submitted.

Results: After the job finished, the result page was accessed via a hyperlink on the submission page or in a notification email. The results page shows the output in two panels (Figure 5-5), the top panel shows a global view of all aligned reads. The bottom panel shows the sequence information of all aligned reads displayed in first window. Whole genome sequence reads from all 8 chickpea accessions had significant similarity with the query sequence suggesting the presence of query gene in all chickpea accessions. These TAGdb results are also downloadable in FASTA format through the link provided in result page.

5.4.2 Study 2

Aim: To find the gene/sequence locus “Nodulation-signalling pathway 2 protein MTR_3g072710” in the *kabuli* genome and any associated SNPs

Method: Using the chickpea BLAST GBrowse tool, a FASTA query sequence representing the gene MTR_3g072710 from NCBI was pasted in the sequence input box. Chickpea *kabuli* annotated genes were selected and BLAST search started by clicking the start button.

Results: The BLAST results summary page confirms that *kabuli* genome contains the nodulation-signalling pathway 2 protein as gene id Ca_26279. The link provided in the BLAST hit score corresponds to respective the alignment and the genome view link directs the user to the GBrowse view showing the BLAST hits on chromosome 5 of the *kabuli* genome. In the BLAST summary page (Figure 5-3), the results of the sequence similarity search are also available to download through the respective link. The GBrowse database also provides additional features and after opening all tracks, the user can see that the identified gene has a higher GC content than flanking regions and also contains two non-synonymous SNPs (red coloured) and one synonymous SNP (blue colour) (Figure 5-3).

5.5 Conclusion

The cicer.info database hosts a range of chickpea genome information with unrestricted public access. The available genome information related to different chickpea cultivars and wild relatives are made available. The cicer.info resource provides researchers access to genomic information and associated genomic features. The portal also allows the users to compare query sequences with genomic data, identify genes at loci of interest and related structural and functional information. The cicer.info web-accessible resources host not only databases, genome browsers, searching utilities but also a platform to host the advance version of genome assemblies, associated genomic features and downstream results. The collective source of genomic information with associated features and web-based tools will greatly help researchers in improving this valuable crop.

6 Concluding remarks and further directions

6.1 Conclusion

High throughput sequencing technologies has enabled to sequence orphan crops. NGS technologies cannot produce complete single sequence length of a chromosome, but are able to generate large number of reads. Genome assembly software combines these fragmented sequence reads into large region called contigs. Like different sequence technologies produce sequence data of different quality, even assembly programs also use different approaches or algorithms to handle genome assembly challenges.

The National Institute of Plant Genome Research (NIPGR) has produced the *desi* (ICC4958) reference genome build using both 454/Roche GS FLX and Illumina data and International Chickpea Genetics and Genomics Consortium has produced the *kabuli* (CDC Frontier *kabuli* type) reference genome build with only Illumina data (Chapter 5). Genome assembly is difficult computational task, and new approaches to assembly are still under development (Narzisi and Mishra, 2011; Tang *et al.*, 2014).

On initial genome comparison at base level (pairwise chromosome comparison), *desi* chromosomes are small compared to *kabuli* chromosomes. This is in contrast to cytogenetic genome size estimation made in both *desi* and *kabuli* genome (Chapter 2). A new method has been developed to assess the draft genome assembly. Taking the advantage of variable length of cytogenetic chromosome size, the chromosomes were flow sorted and isolated to sequence them individually. Using three chromosomes from *desi* genome and all chromosomes from *kabuli* type, the respective reference genomes were assessed with individual chromosome sequence data (Chapter 2). By isolating and sequencing the chromosomes individually, the genome complexity was reduced to chromosome level and so it was possible to evaluate the genome assembly chromosome by chromosome (Chapter 2). This method can also be applied to other crops where isolation of chromosomes is possible.

In parallel to recent advancements in sequencing techniques, new technologies are emerging to genotype large numbers of SNPs (Chapter 1). Modern genotyping methods can integrate up to one million SNPs. Using sequencing technology, a novel genotyping method skimGBS was used in this study. SkimGBS is an alternative genotyping approach for trait mapping and can be applied to characterise recombination and to mine SNPs for association studies (Chapter 4). Furthermore, data gained from skimGBS can be used to improve genome assemblies and to assess structural variation (Chapter 4). This study applies a new method to improve the draft genome by comparing representative haplotype block of each contig sequence. Using this approach both *desi* and *kabuli* draft genomes were improved by integrating unassigned contig sequence into respective chromosome sequences (Chapter 3). This project has successfully generated high-quality reference genomes that may be subsequently used as reference genomes to facilitate the re-sequencing and annotation of related species through comparative methods.

In chapter 4, 611,365 SNPs located on the released *kabuli* assembly were identified between 47 chickpea varieties having different geographical origins. Some of the chickpea accessions were identified as admixtures of *desi* and *kabuli* types. In addition, this study also focused on gene variability in 47 chickpea cultivars using SGSgeneloss. This study enabled to report variable genes between chickpea accessions.

In Chapter 5, a web based system (Cicer.info portal) was made available hosting *desi* and *kabuli* reference genome along with associated annotation, synteny with related species, miss-assemblies, SSRs and SNP (including annotations) to support chickpea research and crop improvement. Cicer.info portal was established to host other web applications like BLAST sequence search portal and TAGdb for searching chickpea NGS sequence data. Additionally, raw WGS data of chickpea varieties and genome assemblies were made freely accessible via web interface at http://www.cicer.info.

The present study constitutes the first report comparing *desi* and *kabuli* genomes and identifying SNP markers for population genetics analysis in cultivated chickpea varieties. This study shows that cultivars are showing high genetic variability with lost genes, genetic diversity and population structure. The generated knowledge about

the levels of diversity and population structure of chickpea germplasm is an important contribution to crop breeding and conservation.

6.2 Future directions

While concluding this thesis, recently, NIPGR, India has released an advanced version of the *desi* reference genome (Parween *et al.*, 2015). This version can be assessed using the same chromosomal genomics method used in Chapter 2. *Kabuli* and *desi* genomes were improved substantially with combined chromosomal genomics and the skimGBS method. This method has increased the quality and length of chromosomes in both chickpea reference assemblies. However, the sequence within each chromosome can further be improved with clustering, and comparison of meta-snps within respective chromosomes would lead to better quality of reference genome.

Emerging NGS technologies and bioinformatics tools have significant implications for crop genetics and breeding. The development of many genetic resources like WGS, transcript and gene variability studies allows for many approaches to crop improvement. Genome wide sequence data produced for 47 chickpea varieties will facilitate the understanding of role of genome complexity and recombination hotspots in crop genetics and breeding. With low coverage of data, skimGBS approach is flexible for trait association, SNP validation, while higher coverage enables fine mapping of recombinations events, validating genome assemblies and identifying the structural variations.

Identification of variable genes in chickpea varieties shows that genetic alterations in chickpea genome have occurred at gene level. The notable differences such as gene variability in 47 chickpea accessions suggest interspecific differences and can further extend to study in wide range of chickpea population. Gene loss method can also be applied to recently WGS sequenced 150 chickpea accessions (AISRF funded chickpea project). On further analysis of gene loss study can define the core and pan genome to include not only annotated genes but also non-coding regions.

The availability of a large number (611,365) of SNPs developed through NGS will facilitate trait mapping and assist in marker assisted breeding (Chapter 4). Large scale development of molecular markers using NGS also assists linkage mapping and association studies. With the increasing availability of genomic data, NGS

technologies accelerate the genome assisted breeding and transformation technologies for crops improvement.

Among the advanced genotyping methods, SkimGBS is efficient in genotyping the huge number of SNPs. This genotyping can be used for genome assessment, genome improvement, SNP validation and association studies.

The outcome of this thesis is introducing a method to assess chickpea reference genome and followed by improvement through a combined chromosomal genomics approach with skimGBS method. Using this approach, *desi* and *kabuli* draft genome quality has increased, additionally, high confident SNP information between 47 chickpea varieties also made available. The above results associated with improved reference genome version could assist to support further studies for chickpea crop improvement.

Reference

Ahmad, F., Khan, A. I., Awan, F. S., Sadia, B., Sadaqat, H. A. and Bahadur, S. (2010) Genetic diversity of chickpea (Cicer arietinum L.) germplasm in Pakistan as revealed by RAPD analysis. *Genet Mol Res* **9**, 1414-1420.

Akhunov, E., Nicolet, C. and Dvorak, J. (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor Appl Genet* **119**, 507-517.

Angadi, S. V., Cutforth, H. W., Miller, P. R., McConkey, B. G., Entz, M. H., Brandt, S. A. and Volkmar, K. M. (2000) Response of three Brassica species to high temperature stress during reproductive growth. *Can J Plant Sci* **80**, 693-701.

Azam, S., Rathore, A., Shah, T. M., Telluri, M., Amindala, B., Ruperao, P., Katta, M. A. V. S. K. and Varshney, R. K. (2014) An Integrated SNP Mining and Utilization (ISMU) Pipeline for Next Generation Sequencing Data. *Plos One* **9**.

Bahl.P.N (1988) *Chickpea*. New Delhi, India:Oxford and IBH publishing Co. Pvt. Ltd.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. and Johnson, E. A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K. J. and Edwards, D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* **19**, 421-422.

Barry, G. F. (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol* **125**, 1164-1165.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J. and Edwards, D. (2003a) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol* **132**, 84-91.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J. and Edwards, D. (2003b) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol* **132**, 84-91.

Batley, J. and Edwards, D. (2009) Genome sequence data: management, storage, and visualization. *Biotechniques* **46**, 333-+.

Bayer, P. E., Ruperao, P., Mason, A. S., Stiller, J., Chan, C. K. K., Hayashi, S., Long, Y., Meng, J. L., Sutton, T., Visendi, P., Varshney, R. K., Batley, J. and Edwards, D. (2015a) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. *Theor Appl Genet* **128**, 1039-1047.

Bayer, PhilippE, Ruperao, Pradeep, Mason, AnnalieseS, Stiller, Jiri, Chan, Chon-KitKenneth, Hayashi, Satomi, Long, Yan, Meng, Jinling, Sutton, Tim, Visendi, Paul, Varshney, RajeevK, Batley, Jacqueline and Edwards, David (2015b) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. *Theor Appl Genet* **128**, 1039-1047.

Bedell, J. A., Budiman, M. A., Nunberg, A., Citek, R. W., Robbins, D., Jones, J., Flick, E., Rohlfing, T., Fries, J., Bradford, K., McMenamy, J., Smith, M., Holeman, H., Roe, B. A., Wiley, G., Korf, I. F., Rabinowicz, P. D., Lakey, N., McCombie, W. R., Jeddeloh, J. A. and Martienssen, R. A. (2005) Sorghum genome sequencing by methylation filtration. *Plos Biol* **3**, 103-115.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Cheetham, R. K., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X. H., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X. L., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Cooley, R. N., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fajardo, K. V. F., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Jones, T. A. H., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S. J., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z. M., Ng, B. L., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Pinkard, D. C., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Rodriguez, A. C., Roe, P. M., Rogers, J., Bacigalupo, M. C. R., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Sohna, J. E. S., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J. W., Worsley, G. J., Yan, J. Y., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Bergman, C. M. and Quesneville, H. (2007) Discovering and detecting transposable elements in genome sequences. *Brief Bioinform* **8**, 382-392.

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. and Edwards, D. (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol J* **11**, 564-571.

Bharadwaj, C., Chauhan, S. K., Yadav, S., Satyavathi, C. T., Singh, R., Kumar, J., Srivastava, R. and Rajguru, G. (2011) Molecular marker-based linkage map of chickpea (Cicer arietinum) developed from desi x kabuli cross. *Indian J Agr Sci* **81**, 116-118.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biol* **13**.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P. and Bateman, A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41**, D226-D232.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C. and Jaffe, D. B. (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res* **18**, 810-820.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009) BLAST plus : architecture and applications. *Bmc Bioinformatics* **10**.

Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Scheix, T., Spannagl, M., Monaghan, E., Nicholson, C., Humphray, S. J., Schoof, H., Mayer, K. F. X., Rogers, J., Quetier, F., Oldroyd, G. E., Debelle, F., Cook, D. R., Retzel, E. F., Roe, B. A., Town, C. D., Tabata, S., Van de Peer, Y. and Young, N. D. (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes (vol 103, pg 14959, 2006). *P Natl Acad Sci USA* **103**, 18026-18026.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S. and Yandell, M. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-196.

Chaisson, M. J. and Pevzner, P. A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res* **18**, 324-330.

Chaisson, M., Pevzner, P. and Tang, H. X. (2004) Fragment assembly with short reads. *Bioinformatics* **20**, 2067-2074.

Chamarthi, S. K., Kumar, A., Vuong, T. D., Blair, M. W., Gaur, P. M., Nguyen, H. T. and Varshney, R. K. (2011) Trait Mapping and Molecular Breeding. *Biology and Breeding of Food Legumes*, 296-313.

Chen, K., McLellan, M. D., Ding, L., Wendl, M. C., Kasai, Y., Wilson, R. K. and Mardis, E. R. (2007) PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res* **17**, 659-666.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q. Y., Locke, D. P., Shi, X. Q., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L. and Mardis, E. R. (2009a) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-U676.

Chen, Y. H., Souaiaia, T. and Chen, T. (2009b) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics* **25**, 2514-2521.

Chowdhury, K. A., Saraswat, K. S., Hasan, S. N., and Gaur, R. C (1971) *4,000–3,500 year old barley, rice, and pulses from Atranjikhera*.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M. and Getz, G. (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2602.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. Y. and 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 w(1118); iso-2; iso-3. *Fly* **6**, 80-92.

Clarke, H. J. and Siddique, K. H. M. (2004) Response of chickpea genotypes to low temperature stress during reproductive development. *Field Crop Res* **90**, 323-334.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265-270.

Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y. H., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J. T., Wanamaker, S., Bozdag, S., Roose, M. L., Moscou, M. J., Chao, S. A. M., Varshney, R. K., Szucs, P., Sato, K., Hayes, P. M., Matthews, D. E., Kleinhofs, A., Muehlbauer, G. J., DeYoung, J., Marshall, D. F., Madishetty, K., Fenton, R. D., Condamine, P., Graner, A. and Waugh, R. (2009) Development and implementation of high-throughput SNP genotyping in barley. *Bmc Genomics* **10**.

Cox, M. P., Peterson, D. A. and Biggs, P. J. (2010a) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics* **11**, 485.

Cox, M. P., Peterson, D. A. and Biggs, P. J. (2010b) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics* **11**.

Croser, J. S., Ahmad, F., Clarke, H. J. and Siddique, K. H. M. (2003) Utilisation of wild Cicer in chickpea improvement - progress, constraints, and prospects. *Aust J Agr Res* **54**, 429-444.

de Ridder, D., de Ridder, J. and Reinders, M. J. T. (2013) Pattern recognition in bioinformatics. *Brief Bioinform* **14**, 633-647.

Deka S, Byjesh K, Kumar U, Choudhary R (2009) Climate change and impacts on crop pests—A critique. Thermotolerance in the crop under climate change. In: *Challenges and Opportunities in Agrometeorology* pp. 147–149. New Delhi.

Deokar, A. A., Ramsay, L., Sharpe, A. G., Diapari, M., Sindhu, A., Bett, K., Warkentin, T. D. and Tar'an, B. (2014) Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. *Bmc Genomics* **15**.

Devos, K. M., Brown, J. K. M. and Bennetzen, J. L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12**, 1075-1079.

DiGuistini, S., Liao, N. Y., Platt, D., Robertson, G., Seidel, M., Chan, S. K., Docking, T. R., Birol, I., Holt, R. A., Hirst, M., Mardis, E., Marra, M. A., Hamelin, R. C., Bohlmann, J., Breuil, C. and Jones, S. J. M. (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol* **10**.

Dita, M. A., Rispail, N., Prats, E., Rubiales, D. and Singh, K. B. (2006) Biotechnology approaches to overcome biotic and abiotic stress constraints in legumes. *Euphytica* **147**, 1-24.

Dolan, P. C. and Denver, D. R. (2008a) TileQC: a system for tile-based quality control of Solexa data. *BMC bioinformatics* **9**, 250.

Dolan, P. C. and Denver, D. R. (2008b) TileQC: A system for tile-based quality control of Solexa data. *BMC bioinformatics* **9**.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L. S., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X. D., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. and Reid, C. A. (2010) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78-81.

Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., Batley, J. and Edwards, D. (2009a) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res* **37**, D951-953.

Duran, C., Appleby, N., Edwards, D. and Batley, J. (2009b) Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation. *Curr Bioinform* **4**, 16-27.

Duran, C., Appleby, N., Vardy, M., Imelfort, M., Edwards, D. and Batley, J. (2009c) Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnol J* **7**, 326-333.

Duran, Chris, Edwards, David and Batley, Jacqueline (2009d) Genetic Maps and the Use of Synteny. In: *Plant Genomics* (Gustafson, J.P., Langridge, P. and Somers, D.J. eds), pp. 41-55. Humana Press.

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. and Turner, S. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133-138.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. and Mitchell, S. E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One* **6**.

FAOSTAT FAO Statistical databases. In: *Food and Agriculture Organization (FAO)*.

Farkhari, M., Lu, Y. L., Shah, T., Zhang, S. H., Naghavi, M. R., Rong, T. Z. and Xu, Y. B. (2011) Recombination frequency variation in maize as revealed by genomewide single-nucleotide polymorphisms. *Plant Breeding* **130**, 533-539.

Flandez-Galvez, H., Ades, P. K., Ford, R., Pang, E. C. K. and Taylor, P. W. J. (2003) QTL analysis for ascochyta blight resistance in an intraspecific population of chickpea (Cicer arietinum L.). *Theor Appl Genet* **107**, 1257-1265.

Frohler, S. and Dieterich, C. (2010) ACCUSA-accurate SNP calling on draft genomes. *Bioinformatics* **26**, 1364-1365.

Fulton, Robert (2016) Comprehensive, Cost Competitive, Genomics Capabilities at The McDonnell Genome Institute at Washington University. San Diego, CA: Plant & Animal Genome Conference XXIV.

G.C. Hawtin, K.B. Singh (1979) Kabuli-desi introgression: problems and prospects. In: *Proceedings of the International Workshop on Chickpea Improvement* pp. 51-60.

Gan, Y., Angadi, S. V., Cutforth, H., Potts, D., Angadi, V. V. and McDonald, C. L. (2004) Canola and mustard response to short periods of temperature and water stress at different developmental stages. *Can J Plant Sci* **84**, 697-704.

Gaur, Mahendra, Das, Aradhana, Sahoo, Rajesh Kumar, Mohanty, Sujata, Joshi, Raj Kumar and Subudhi, Enketeswara (2016) Comparative transcriptome analysis of ginger variety Suprabha from two different agro-climatic zones of Odisha. *Genomics Data* **9**, 42-43.

Gaur, R., Azam, S., Jeena, G., Khan, A. W., Choudhary, S., Jain, M., Yadav, G., Tyagi, A. K., Chattopadhyay, D. and Bhatia, S. (2012) High-Throughput SNP Discovery and Genotyping for Constructing a Saturated Linkage Map of Chickpea (Cicer arietinum L.). *DNA Res* **19**, 357-373.

Gelfand, M. S., Mironov, A. A. and Pevzner, P. A. (1996) Gene recognition via spliced sequence alignment. *P Natl Acad Sci USA* **93**, 9061-9066.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchinson, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J. P., Miguel, T., Paszkowski, U., Zhang, S. P., Colbert, M., Sun, W. L., Chen, L. L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y. S., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). *Science* **296**, 92-100.

Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P., Fitzgerald, T. L., Edwards, D. and Batley, J. (2015) Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Funct Integr Genomic* **15**, 189-196.

Goodstein, D. M., Shu, S. Q., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D. S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**, D1178-D1186.

Gore, M. A., Chia, J. M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., Peiffer, J. A., McMullen, M. D., Grills, G. S., Ross-Ibarra, J., Ware, D. H. and Buckler, E. S. (2009) A First-Generation Haplotype Map of Maize. *Science* **326**, 1115-1117.

Gornicki, P., Faris, J., King, I., Podkowinski, J., Gill, B. and Haselkorn, R. (1997) Plastid-localized acetyl-CoA carboxylase of bread wheat is encoded by a single gene on each of the three ancestral chromosome sets. *P Natl Acad Sci USA* **94**, 14179-14184.

Gowda, S. J. M., Radhika, P., Kadoo, N. Y., Mhase, L. B. and Gupta, V. S. (2009) Molecular mapping of wilt resistance genes in chickpea. *Mol Breeding* **24**, 177-183.

Gujaria, N., Kumar, A., Dauthal, P., Dubey, A., Hiremath, P., Prakash, A. B., Farmer, A., Bhide, M., Shah, T., Gaur, P. M., Upadhyaya, H. D., Bhatia, S., Cook, D. R., May, G. D. and Varshney, R. K. (2011) Development and use of genic molecular markers (GMMs) for construction of a transcript map of chickpea (Cicer arietinum L.). *Theor Appl Genet* **122**, 1577-1589.

Hall, A. E. (2004) Breeding for adaptation to drought and heat in cowpea. *Eur J Agron* **21**, 447-454.

Harris, B., Riemer, C. and Miller, W. (2010) LastZ.

Heger, M (01 Oct 2015) PacBio launches higher-throughput, lower-cost single-molecule sequencing system.

Helms, C., Mishra, S. K., Riethman, H., Burgess, A. K., Ramachandra, S., Tierney, C., Dorsey, D. and Doniskeller, H. (1992) Closure of a Genetic-Linkage Map of Human Chromosome-7q with Centromere and Telomere Polymorphisms. *Genomics* **14**, 1041-1054.

Heyne, S., Costa, F., Rose, D. and Backofen, R. (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* **28**, i224-232.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. and Hackermuller, J. (2009) Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. *PLoS computational biology* **5**.

Holtz, Y., Ardisson, M., Ranwez, V., Besnard, A., Leroy, P., Poux, G., Roumet, P., Viader, V., Santoni, S. and David, J. (2016) Genotyping by Sequencing Using Specific Allelic Capture to Build a High-Density Genetic Map of Durum Wheat. *Plos One* **11**.

Hormozdiari, F., Alkan, C., Eichler, E. E. and Sahinalp, S. C. (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**, 1270-1278.

Huang, X. and Yang, S. P. (2005) Generating a genome assembly with PCAP. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 11**, Unit11 13.

Iruela, M., Rubio, J., Cubero, J. I., Gil, J. and Millan, T. (2002) Phylogenetic analysis in the genus Cicer and cultivated chickpea using RAPD and ISSR markers. *Theor Appl Genet* **104**, 643-651.

Jain, M., Misra, G., Patel, R. K., Priya, P., Jhanwar, S., Khan, A. W., Shah, N., Singh, V. K., Garg, R., Jeena, G., Yadav, M., Kant, C., Sharma, P., Yadav, G., Bhatia, S., Tyagi, A. K. and Chattopadhyay, D. (2013) A draft genome sequence of the pulse crop chickpea (Cicer arietinumL.). *Plant J* **74**, 715-729.

Jhanwar, S., Priya, P., Garg, R., Parida, S. K., Tyagi, A. K. and Jain, M. (2012) Transcriptome sequencing of wild chickpea as a rich resource for marker development. *Plant Biotechnol J* **10**, 690-702.

Jimenez-Gasco, M. D., Milgroom, M. G. and Jimenez-Diaz, R. M. (2004) Stepwise evolution of races in Fusarium oxysporum f. sp ciceris inferred from fingerprinting with repetitive DNA sequences. *Phytopathology* **94**, 228-235.

Kale, Sandip M., Jaganathan, Deepa, Ruperao, Pradeep, Chen, Charles, Punna, Ramu, Kudapa, Himabindu, Thudi, Mahendar, Roorkiwal, Manish, Katta, Mohan Avsk, Doddamani, Dadakhalandar, Garg, Vanika, Kishor, P. B. Kavi, Gaur, Pooran M., Nguyen, Henry T., Batley, Jacqueline, Edwards, David, Sutton, Tim and Varshney, Rajeev K. (2015) Prioritization of candidate genes in “QTL-hotspot” region for drought tolerance in chickpea (Cicer arietinum L.). *Scientific Reports* **5**, 15296.

Kashkush, K., Feldman, M. and Levy, A. A. (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* **160**, 1651-1659.

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X. Y., Town, C. D., Venter, J. C., Fraser, C. M., Tabata, S., Nakamura, Y., Kaneko, T., Sato, S., Asamizu, E., Kato, T., Kotani, H., Sasamoto, S., Ecker, J. R., Theologis, A., Federspiel, N. A., Palm, C. J., Osborne, B. I., Shinn, P., Conway, A. B., Vysotskaia, V. S., Dewar, K., Conn, L., Lenz, C. A., Kim, C. J., Hansen, N. F., Liu, S. X., Buehler, E., Altafi, H., Sakano, H., Dunn, P.,

Lam, B., Pham, P. K., Chao, Q., Nguyen, M., Yu, G. X., Chen, H. M., Southwick, A., Lee, J. M., Miranda, M., Toriumi, M. J., Davis, R. W., Wambutt, R., Murphy, G., Dusterhoft, A., Stiekema, W., Pohl, T., Entian, K. D., Terryn, N., Volckaert, G., Salanoubat, M., Choisne, N., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Artiguenave, F., Weissenbach, J., Quetier, F., Wilson, R. K., de la Bastide, M., Sekhon, M., Huang, E., Spiegel, L., Gnoj, L., Pepin, K., Murray, J., Johnson, D., Habermann, K., Dedhia, N., Parnell, L., Preston, R., Hillier, L., Chen, E., Marra, M., Martienssen, R., McCombie, W. R., Mayer, K., White, O., Bevan, M., Lemcke, K., Creasy, T. H., Bielke, C., Haas, B., Haase, D., Maiti, R., Rudd, S., Peterson, J., Schoof, H., Frishman, D., Morgenstern, B., Zaccaria, P., Ermolaeva, M., Pertea, M., Quackenbush, J., Volfovsky, N., Wu, D. Y., Lowe, T. M., Salzberg, S. L., Mewes, H. W., Rounsley, S., Bush, D., Subramaniam, S., Levin, I., Norris, S., Schmidt, R., Acarkan, A., Bancroft, I., Quetier, F., Brennicke, A., Eisen, J. A., Bureau, T., Legault, B. A., Le, Q. H., Agrawal, N., Yu, Z., Martienssen, R., Copenhaver, G. P., Luo, S., Pikaard, C. S., Preuss, D., Paulsen, I. T., Sussman, M., Britt, A. B., Selinger, D. A., Pandey, R., Mount, D. W., Chandler, V. L., Jorgensen, R. A., Pikaard, C., Juergens, G., Meyerowitz, E. M., Theologis, A., Dangl, J., Jones, J. D. G., Chen, M., Chory, J., Somerville, M. C. and In, Ar Gen (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815.

Kent, W. J. (2002) BLAT - The BLAST-like alignment tool. *Genome Res* **12**, 656-664.

Khan, M. A., Akhtar, N., Ullah, I. and Jaffery, S. (1995) Nutritional-Evaluation of Desi and Kabuli Chickpeas and Their Products Commonly Consumed in Pakistan. *Int J Food Sci Nutr* **46**, 215-223.

Korf, I. (2004) Gene finding in novel genomes. *BMC bioinformatics* **5**.

Kris, Hirst (2014) The Domestication History of Chickpeas. About education.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009a) Circos: An information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009b) Circos: An information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645.

Kujur, A., Bajaj, D., Saxena, M. S., Tripathi, S., Upadhyaya, H. D., Gowda, C. L. L., Singh, S., Jain, M., Tyagi, A. K. and Parida, S. K. (2013) Functionally Relevant Microsatellite Markers From Chickpea Transcription Factor Genes for Efficient Genotyping Applications and Trait Association Mapping. *DNA Res* **20**, 355-373.

Kulheim, C., Yeoh, S. H., Maintz, J., Foley, W. J. and Moran, G. F. (2009) Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. *Bmc Genomics* **10**.

Labate, J. A. and Baldo, A. M. (2005) Tomato SNP discovery by EST mining and resequencing. *Mol Breeding* **16**, 343-349.

Ladizinsky, G (1975) *A new Cicer from Turkey*:Notes Roy. Bot. Gard. Edinb.

Ladizinsky G, Adler A (1976) The origin of chickpea of Cicer arietinum. *L. Euphytica* **25**.

Lai, K. T., Berkman, P. J., Lorenc, M. T., Duran, C., Smits, L., Manoli, S., Stiller, J. and Edwards, D. (2012a) WheatGenome.info: An Integrated Database and Portal for Wheat Genome Information. *Plant Cell Physiol* **53**.

Lai, K. T., 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. and Edwards, D. (2012b) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol J* **10**, 743-749.

Lai, K. T., 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. and Edwards, D. (2015) Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnol J* **13**, 97-104.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009a) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009b) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**.

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., Stein, L., Holmes, I. H., Elsik, C. G. and Lewis, S. E. (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**.

Lee, S., Hormozdiari, F., Alkan, C. and Brudno, M. (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* **6**, 473-474.

Lee, W. P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P. and Marth, G. T. (2014) MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. *Plos One* **9**.

Li, H. (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* **28**, 1838-1844.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Proc, 1000 Genome Project Data (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Li, H., Ruan, J. and Durbin, R. (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b) SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-714.

Li, R. Q., Fan, W., Tian, G., Zhu, H. M., He, L., Cai, J., Huang, Q. F., Cai, Q. L., Li, B., Bai, Y. Q., Zhang, Z. H., Zhang, Y. P., Wang, W., Li, J., Wei, F. W., Li, H., Jian, M., Li, J. W., Zhang, Z. L., Nielsen, R., Li, D. W., Gu, W. J., Yang, Z. T., Xuan, Z. L., Ryder, O. A., Leung, F. C. C., Zhou, Y., Cao, J. J., Sun, X., Fu, Y. G., Fang, X. D., Guo, X. S., Wang, B., Hou, R., Shen, F. J., Mu, B., Ni, P. X., Lin, R. M., Qian, W. B., Wang, G. D., Yu, C., Nie, W. H., Wang, J. H., Wu, Z. G., Liang, H. Q., Min, J. M., Wu, Q., Cheng, S. F., Ruan, J., Wang, M. W., Shi, Z. B., Wen, M., Liu, B. H., Ren, X. L., Zheng, H. S., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X. Y., Lu, Z. H., Zheng, H. C., Li, Y. R., Steiner, C. C., Lam, T. T. Y., Lin, S. Y., Zhang, Q. H., Li, G. Q., Tian, J., Gong, T. M., Liu, H. D., Zhang, D. J., Fang, L., Ye, C., Zhang, J. B., Hu, W. B., Xu, A. L., Ren, Y. Y., Zhang, G. J., Bruford, M. W., Li, Q. B., Ma, L. J., Guo, Y. R., An, N., Hu, Y. J., Zheng, Y., Shi, Y. Y., Li, Z. Q., Liu, Q., Chen, Y. L., Zhao, J., Qu, N., Zhao, S. C., Tian, F., Wang, X. L., Wang, H. Y., Xu, L. Z., Liu, X., Vinar, T., Wang, Y. J., Lam, T. W., Yiu, S. M., Liu, S. P., Zhang, H. M., Li, D. S., Huang, Y., Wang, X., Yang, G. H., Jiang, Z., Wang, J. Y., Qin, N., Li, L., Li, J. X., Bolund, L., Kristiansen, K., Wong, G. K. S., Olson, M., Zhang, X. Q., Li, S. G., Yang, H. M., Wang, J. and Wang, J. (2010a) The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317.

Li, R. Q., Yu, C., Li, Y. R., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009b) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967.

Li, R. Q., Zhu, H. M., Ruan, J., Qian, W. B., Fang, X. D., Shi, Z. B., Li, Y. R., Li, S. T., Shan, G., Kristiansen, K., Li, S. G., Yang, H. M., Wang, J. and Wang, J. (2010b) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272.

Lin, H., Zhang, Z. F., Zhang, M. Q., Ma, B. and Li, M. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**, 2431-2437.

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. and Wang, J. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* **496**, 87-90.

Lopez-Bigas, N. and Ouzounis, C. A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108-3114.

Lorenc, Michał T., Hayashi, Satomi, Stiller, Jiri, Lee, Hong, Manoli, Sahana, Ruperao, Pradeep, Visendi, Paul, Berkman, Paul J., Lai, Kaitao, Batley, Jacqueline and Edwards, David (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* **1**, 370-382.

Lukashin, A. V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115.

Maqbool, A., Shafiq, S. and Lake, L. (2010) Radiant frost tolerance in pulse crops-a review. *Euphytica* **172**, 1-12.

Mariac, C., Scarcelli, N., Pouzadou, J., Barnaud, A., Billot, C., Faye, A., Kougbeadjo, A., Maillol, V., Martin, G., Sabot, F., Santoni, S., Vigouroux, Y. and Couvreur, T. L. P. (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Mol Ecol Resour* **14**, 1103-1113.

Marshall, D. J., Hayward, A., Eales, D., Imelfort, M., Stiller, J., Berkman, P. J., Clark, T., McKenzie, M., Lai, K. T., Duran, C., Batley, J. and Edwards, D. (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods* **6**.

Martienssen, R. A., Rabinowicz, P. D., O'Shaughnessy, A. and McCombie, W. R. (2004) Sequencing the maize genome. *Curr Opin Plant Biol* **7**, 102-107.

Martinez-Alcantara, A., Ballesteros, E., Feng, C., Rojas, M., Koshinsky, H., Fofanov, V. Y., Havlak, P. and Fofanov, Y. (2009) PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* **25**, 2438-2439.

Matsumoto, T., Wu, J. Z., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B. A., Baba, T., Sakata, K., Nagamura, Y., Aoki, H., Arikawa, K., Arita, K., Bito, T., Chiden, Y., Fujitsuka, N., Fukunaka, R., Hamada, M., Harada, C., Hayashi, A., Hijishita, S., Honda, M., Hosokawa, S., Ichikawa, Y., Idonuma, A., Iijima, M., Ikeda, M., Ikeno, M., Ito, K., Ito, S., Ito, T., Ito, Y., Ito, Y., Iwabuchi, A., Kamiya, K., Karasawa, W., Kurita, K., Katagiri, S., Kikuta, A., Kobayashi, H., Kobayashi, N., Machita, K., Maehara, T., Masukawa, M., Mizubayashi, T., Mukai, Y., Nagasaki, H., Nagata, Y., Naito, S., Nakashima, M., Nakama, Y., Nakamichi, Y., Nakamura, M., Meguro, A., Negishi, M., Ohta, I., Ohta, T., Okamoto, M., Ono, N., Saji, S., Sakaguchi, M., Sakai, K., Shibata, M., Shimokawa, T., Song, J. Y., Takazaki, Y., Terasawa, K., Tsugane, M., Tsuji, K., Ueda, S., Waki, K., Yamagata, H., Yamamoto, M., Yamamoto, S., Yamane, H., Yoshiki, S., Yoshihara, R., Yukawa, K., Zhong, H. S., Yano, M., Sasaki, T., Yuan, Q. P., Shu, O. T., Liu, J., Jones, K. M., Gansberger, K., Moffat, K., Hill, J., Bera, J., Fadrosh, D., Jin, S. H., Johri, S., Kim, M., Overton, L., Reardon, M., Tsitrin, T., Vuong, H., Weaver, B., Ciecko, A., Tallon, L., Jackson, J., Pai, G., Van Aken, S., Utterback, T., Reidmuller, S., Feldblyum, T., Hsiao, J., Zismann, V., Iobst, S., de Vazeille, A. R., Buell, C. R., Ying, K., Li, Y., Lu, T. T., Huang, Y. C., Zhao, Q., Feng, Q., Zhang, L., Zhu, J. J., Weng, Q. J., Mu, J., Lu, Y. Q., Fan, D. L., Liu, Y. L., Guan, J. P., Zhang, Y. J., Yu, S. L., Liu, X. H., Zhang, Y., Hong, G. F., Han, B., Choisne, N., Demange, N., Orjeda, G., Samain, S., Cattolico, L., Pelletier, E., Couloux, A., Segurens, B., Wincker, P., D'Hont, A., Scarpelli, C., Weissenbach, J., Salanoubat, M., Quetier, F., Yu, Y., Kim, H. R., Rambo, T., Currie, J., Collura, K., Luo, M. Z., Yang, T. J., Ammiraju, J. S. S., Engler, F., Soderlund, C., Wing, R. A., Palmer, L. E., de la Bastide, M., Spiegel, L., Nascimento, L., Zutavern, T., O'Shaughnessy, A., Dike, S., Dedhia, N., Preston, R., Balija, V., McCombie, W. R., Chow, T. Y., Chen, H. H., Chung, M. C., Chen, C. S., Shaw, J. F., Wu, H. P., Hsiao, K. J., Chao, Y. T., Chu, M. K., Cheng, C. H., Hour, A. L., Lee, P. F., Lin, S. J., Lin, Y. C., Liou, J. Y., Liu, S. M., Hsing, Y. I., Raghuvanshi, S., Mohanty, A., Bharti, A. K., Gaur, A., Gupta, V., Kumar, D., Ravi,

V., Vij, S., Kapur, A., Khurana, P., Khurana, P., Khurana, J. P., Tyagi, A. K., Gaikwad, K., Singh, A., Dalal, V., Srivastava, S., Dixit, A., Pal, A. K., Ghazi, I. A., Yadav, M., Pandit, A., Bhargava, A., Sureshbabu, K., Batra, K., Sharma, T. R., Mohapatra, T., Singh, N. K., Messing, J., Nelson, A. B., Fuks, G., Kavchok, S., Keizer, G., Llaca, E. L. V., Song, R. T., Tanyolac, B., Young, S., Il, K. H., Hahn, J. H., Sangsakoo, G., Vanavichit, A., de Mattos, L. A. T., Zimmer, P. D., Malone, G., Dellagostin, O., de Oliveira, A. C., Bevan, M., Bancroft, I., Minx, P., Cordum, H., Wilson, R., Cheng, Z. K., Jin, W. W., Jiang, J. M., Leong, S. A., Iwama, H., Gojobori, T., Itoh, T., Niimura, Y., Fujii, Y., Habara, T., Sakai, H., Sato, Y., Wilson, G., Kumar, K., McCouch, S., Juretic, N., Hoen, D., Wright, S., Bruskiewich, R., Bureau, T., Miyao, A., Hirochika, H., Nishikawa, T., Kadowaki, K., Sugiura, M. and Project, Int Rice Genome Sequencing (2005) The map-based sequence of the rice genome. *Nature* **436**, 793-800.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *P Natl Acad Sci USA* **74**, 560-564.

Mayer, K. F. X., Rogers, J., Dolezel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A. J., Sourdille, P., Endo, T. R., Dolezel, J., Kubalakova, M., Cihalikova, J., Dubska, Z., Vrana, J., Sperkova, R., Simkova, H., Rogers, J., Febrer, M., Clissold, L., McLay, K., Singh, K., Chhuneja, P., Singh, N. K., Khurana, J., Akhunov, E., Choulet, F., Sourdille, P., Feuillet, C., Alberti, A., Barbe, V., Wincker, P., Kanamori, H., Kobayashi, F., Itoh, T., Matsumoto, T., Sakai, H., Tanaka, T., Wu, J. Z., Ogihara, Y., Handa, H., Pozniak, C., Maclachlan, P. R., Sharpe, A., Klassen, D., Edwards, D., Batley, J., Olsen, O. A., Sandve, S. R., Lien, S., Steuernagel, B., Wulff, B., Caccamo, M., Ayling, S., Ramirez-Gonzalez, R. H., Clavijo, B. J., Steuernagel, B., Wright, J., Pfeifer, M., Spannagl, M., Mayer, K. F. X., Martis, M. M., Akhunov, E., Choulet, F., Mayer, K. F. X., Mascher, M., Chapman, J., Poland, J. A., Scholz, U., Barry, K., Waugh, R., Rokhsar, D. S., Muehlbauer, G. J., Stein, N., Gundlach, H., Zytnicki, M., Jamilloux, V., Quesneville, H., Wicker, T., Mayer, K. F. X., Faccioli, P., Colaiacovo, M., Pfeifer, M., Stanca, A. M., Budak, H., Cattivelli, L., Glover, N., Martis, M. M., Choulet, F., Feuillet, C., Mayer, K. F. X., Pfeifer, M., Pingault, L., Mayer, K. F. X., Paux, E., Spannagl, M., Sharma, S., Mayer, K. F. X., Pozniak, C., Appels, R., Bellgard, M., Chapman, B., Pfeifer, M., Pfeifer, M., Sandve, S. R., Nussbaumer, T., Bader, K. C., Choulet, F., Feuillet, C., Mayer, K. F. X., Akhunov, E., Paux, E., Rimbert, H., Wang, S. C., Poland, J. A., Knox, R., Kilian, A., Pozniak, C., Alaux, M., Alfama, F., Couderc, L., Jamilloux, V., Guilhot, N., Viseux, C., Loaec, M., Quesneville, H., Rogers, J., Dolezel, J., Eversole, K., Feuillet, C., Keller, B., Mayer, K. F. X., Olsen, O. A., Praud, S. and Iwgsc (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* **345**.

Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A., Matsumoto, T., Sato, K., Schulman, A., Muehlbauer, G. J., Stein, N., Ariyadasa, R., Schulte, D., Poursarebani, N., Zhou, R. N., Steuernagel, B., Mascher, M., Scholz, U., Shi, B. J., Langridge, P., Madishetty, K., Svensson, J. T., Bhat, P., Moscou, M., Resnik, J., Close, T. J., Muehlbauer, G. J., Hedley, P., Liu, H., Morris, J., Waugh, R., Frenkel, Z., Korol, A., Berges, H., Graner, A., Stein, N., Steuernagel, B., Taudien, S., Groth, M., Felder, M., Platzer, M., Brown, J. W. S., Schulman, A., Platzer, M., Fincher, G. B., Muehlbauer, G. J., Sato, K., Taudien, S., Sampath, D., Swarbreck, D., Scalabrin, S., Zuccolo, A., Vendramin, V., Morgante, M., Mayer, K. F. X., Schulman, A. and Conso, Int Barley Genome Sequencing (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-+.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y. T., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H. N., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De la Vega, F. M. and Blanchard, A. P. (2009) Sequence and

structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527-1541.

Mendelsohn, R., Nordhaus, W. D. and Shaw, D. (1994) The Impact of Global Warming on Agriculture - a Ricardian Analysis. *Am Econ Rev* **84**, 753-771.

Millan, T., Clarke, H. J., Siddique, K. H. M., Buhariwalla, H. K., Gaur, P. M., Kumar, J., Gil, J., Kahl, G. and Winter, P. (2006) Chickpea molecular breeding: New tools and concepts. *Euphytica* **147**, 81-103.

Millan, T., Winter, P., Jungling, R., Gil, J., Rubio, J., Cho, S., Cobos, M. J., Iruela, M., Rajesh, P. N., Tekeoglu, M., Kahl, G. and Muehlbauer, F. J. (2010) A consensus genetic map of chickpea (Cicer arietinum L.) based on 10 mapping populations. *Euphytica* **175**, 175-189.

Milo, N., Zakov, S., Katzenelson, E., Bachmat, E., Dinitz, Y. and Ziv-Ukelson, M. (2013) Unrooted unordered homeomorphic subtree alignment of RNA trees. *Algorithms for molecular biology : AMB* **8**, 13.

Ming, R., Hou, S. B., Feng, Y., Yu, Q. Y., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C. X., Qian, W. B., Shen, J. G., Du, P., Eustice, M., Tong, E., Tang, H. B., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M. L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P. Z., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J. M., Wang, J. P., Na, J. K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X. Y., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H. R., Irikura, B., Paidi, M., Jiang, N., Zhang, W. L., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y. J., Burroughs, A. M., Luo, M. C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J. M., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L. and Alam, M. (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**, 991-U997.

Muehlbauer, FredJ and Rajesh, P. N. (2008) Chickpea, a Common Source of Protein and Starch in the Semi-Arid ropics. In: *Genomics of Tropical Crop Plants* (Moore, P. and Ming, R. eds), pp. 171-186. Springer New York.

Narzisi, G. and Mishra, B. (2011) Comparing De Novo Genome Assembly: The Long and Short of It. *Plos One* **6**.

Nawrocki, E. P., Kolbe, D. L. and Eddy, S. R. (2009a) Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337.

Nawrocki, E. P., Kolbe, D. L. and Eddy, S. R. (2009b) Infernal 1.0: inference of RNA alignments (vol 25, pg 1335, 2009). *Bioinformatics* **25**, 1713-1713.

Nguyen, T. T., Taylor, P. W. J., Redden, R. J. and Ford, R. (2004) Genetic diversity estimates in Cicer using AFLP analysis. *Plant Breeding* **123**, 173-179.

Okser, S., Pahikkala, T. and Aittokallio, T. (2013) Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. *Biodata Min* **6**.

Pande, S., Siddique, K. H. M., Kishore, G. K., Bayaa, B., Gaur, P. M., Gowda, C. L. L., Bretag, T. W. and Crouch, J. H. (2005) Ascochyta blight of chickpea (Cicer arietinum L.): a review of biology, pathogenicity, and disease management. *Aust J Agr Res* **56**, 317-332.

Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011) Sequencing technologies and genome sequencing. *Journal of applied genetics* **52**, 413-435.

Parida, S. K., Mukerji, M., Singh, A. K., Singh, N. K. and Mohapatra, T. (2012) SNPs in stress-responsive rice genes: validation, genotyping, functional relevance and population structure. *Bmc Genomics* **13**.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* **23**, 1061-1067.

Parween, S., Nawaz, K., Roy, R., Pole, A. K., Suresh, B. V., Misra, G., Jain, M., Yadav, G., Parida, S. K., Tyagi, A. K., Bhatia, S. and Chattopadhyay, D. (2015) An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). *Scientific Reports* **5**.

Patel, R. K. and Jain, M. (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *Plos One* **7**.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H. B., Wang, X. Y., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. A., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. A., Wang, Y., Zhang, L. F., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., Mehboob-ur-Rahman, Ware, D., Westhoff, P., Mayer, K. F. X., Messing, J. and Rokhsar, D. S. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556.

Patil, P. B., Vrinten, P. L., Scoles, G. J. and Slinkard, A. E. (1995) Variation in the Ribosomal-Rna Units of the Genera Lens and Cicer. *Euphytica* **83**, 33-42.

Pevzner, P. A., Tang, H. X. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. *P Natl Acad Sci USA* **98**, 9748-9753.

Planet, E., Attolini, C. S. O., Reina, O., Flores, O. and Rossell, D. (2012a) htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**, 589-590.

Planet, E., Attolini, C. S., Reina, O., Flores, O. and Rossell, D. (2012b) htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**, 589-590.

Quisenberry (1982) Breeding for drought resistance and plant water use efficiency. In: *Breeding Plants for Less Favorable Environments* (Christiansen, M.N.L., C.F ed) pp. 289–327. New York, NY Wiley.

Rahman, A. Y. A., Usharraj, A. O., Misra, B. B., Thottathil, G. P., Jayasekaran, K., Feng, Y., Hou, S. B., Ong, S. Y., Ng, F. L., Lee, L. S., Tan, H. S., Sakaff, M. K. L. M., Teh, B. S., Khoo, B. F., Badai, S. S., Ab Aziz, N., Yuryev, A., Knudsen, B., Dionne-Laporte, A., Mchunu, N. P., Yu, Q. Y., Langston, B. J., Freitas, T. A. K., Young, A. G., Chen, R., Wang, L., Najimudin, N., Saito, J. A. and Alam, M. (2013) Draft genome sequence of the rubber tree Hevea brasiliensis. *Bmc Genomics* **14**.

Raj, A., Stephens, M. and Pritchard, J. K. (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* **197**, 573-U207.

Rajesh, P. N., McPhee, K. E., Ford, R., Pittock, C., Kumar, J. and Muehlbauer, J. (2007) Ciceromics: Advancement in Genomics and Recent Molecular Techniques. *Chickpea Breeding and Management*, 445-457.

Reese, M. G., Kulp, D., Tammana, H. and Haussler, D. (2000) Genie - Gene finding in Drosophila melanogaster. *Genome Res* **10**, 529-538.

Rehman, A. U., Malhotra, R. S., Bett, K., Tar'an, B., Bueckert, R. and Warkentin, T. D. (2011) Mapping QTL Associated with Traits Affecting Grain Yield in Chickpea (Cicer arietinum L.) under Terminal Drought Stress. *Crop Sci* **51**, 450-463.

Rife, T. W., Wu, S. Y., Bowden, R. L. and Poland, J. A. (2015) Spiked GBS: a unified, open platform for single marker genotyping and whole-genome profiling. *Bmc Genomics* **16**.

Rogers, A., Ainsworth, E. A. and Leakey, A. D. (2009) Will elevated carbon dioxide concentration amplify the benefits of nitrogen fixation in legumes? *Plant Physiol* **151**, 1009-1016.

Roorkiwal, M., Nayak, S. N., Thudi, M., Upadhyaya, H. D., Brunel, D., Mournet, P., This, D., Sharma, P. C. and Varshney, R. K. (2014a) Allele diversity for abiotic stress responsive candidate genes in chickpea reference set using gene based SNP markers. *Front Plant Sci* **5**.

Roorkiwal, M., von Wettberg, E. J., Upadhyaya, H. D., Warschefsky, E., Rathore, A. and Varshney, R. K. (2014b) Exploring Germplasm Diversity to Understand the Domestication Process in Cicer spp. Using SNP and DArT Markers. *Plos One* **9**.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S.,

Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M. and Kent, W. J. (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**, D670-D681.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS computational biology* **5**.

Ruperao, P., Chan, C. K., Azam, S., Karafiatova, M., Hayashi, S., Cizkova, J., Saxena, R. K., Simkova, H., Song, C., Vrana, J., Chitikineni, A., Visendi, P., Gaur, P. M., Millan, T., Singh, K. B., Taran, B., Wang, J., Batley, J., Dolezel, J., Varshney, R. K. and Edwards, D. (2014a) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant biotechnology journal*.

Ruperao, P., Chan, C. K. K., Azam, S., Karafiatova, M., Hayashi, S., Cizkova, J., Saxena, R. K., Simkova, H., Song, C., Vrana, J., Chitikineni, A., Visendi, P., Gaur, P. M., Millan, T., Singh, K. B., Taran, B., Wang, J., Batley, J., Dolezel, J., Varshney, R. K. and Edwards, D. (2014b) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant Biotechnol J* **12**, 778-786.

Russell-Stevens, M., Plane, D. C., Summerscales, J., Schulz, P. and Papakyriacou, M. (2002) Effect of elevated temperature on ultimate tensile strength and failure modes of short carbon fibre reinforced magnesium composite. *Mater Sci Tech-Lond* **18**, 501-506.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Saint Pierre, C., Burgueño, J., Crossa, J., Fuentes Dávila, G., Figueroa López, P., Solís Moya, E., Ireta Moreno, J., Hernández Muela, V. M., Zamora Villa, V. M., Vikram, P., Mathews, K., Sansaloni, C., Sehgal, D., Jarquin, D., Wenzl, P. and Singh, Sukhwinder (2016) Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. *Scientific Reports* **6**, 27312.

Samineni, Srinivasan (2010) Physiology, genetics and QTL mapping of salt tolerance in chickpea (Cicer arietinum L.). pp. xvi, 122 p.: University of Western Australia.

Sanger, F. (1975) The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society* **191**, 317-333.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA* **74**, 5463-5467.

Saraf, C.S.R., O.P.; Hegde, D.M.; Yadav, R.L.; Shivkumar, B.G.; Bhattarai, S.; and Razzaque, and M.A.S., M.A (1998) Biological Nitrogen Fixation and Residual Effects of Winter Grain Legumes in Rice and Wheat Cropping Systems of the Indo-Gangetic Plain. New Delhi: India:Oxford and IBH Publishing Co. Pvt. Ltd.

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., Mayer, K., Jia, Z. Q., Zhang, J. H., Ye, Z. B. A., Bishop, G. J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, R., Singh, M., Singh, A., Pal, J. K., Pandit, A., Singh, P. K., Mahato, A. K.,

Dogra, V., Gaikwad, K., Sharma, T. R., Mohapatra, T., Singh, N. K., Causse, M., Rothan, C., Schiex, T., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P., Vautrin, S., Zouine, M., Latche, A., Rousseau, C., Regad, F., Pech, J. C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., del Carmen, A. F., Monforte, A., Granell, A., Fernandez-Munoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S., Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perrotta, G., Daddiego, L., Bryan, G., Orozco, M., Pastor, X., Torrents, D., van Schriek, K. N. V. M. G. M., Feron, R. M. C., van Oeveren, J., de Heer, P., daPonte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., van Eijk, M. J. T., Janssen, A., van Haaren, M. J. J., Jo, S. H., Kim, J., Kwon, S. Y., Kim, S., Koo, D. H., Lee, S., Hur, C. G., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jocker, A., Warfsmann, J., Gobel, U., Kawamura, S., Yano, K., Sherman, J. D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E. W. M., van de Belt, J., van Haarst, J. C., Peters, S. A., van Staveren, M. J., Henkens, M. H. C., Mooyman, P. J. W., Hesselink, T., van Ham, R. C. H. J., Jiang, G. Y., Droege, M., Choi, D., Kang, B. C., Kim, B. D., Park, M., Kim, S., Yeom, S. I., Lee, Y. H., Choi, Y. D., Li, G. C., Gao, J. W., Liu, Y. S., Huang, S. X., Fernandez-Pedrosa, V., Collado, C., Zuniga, S., Wang, G. P., Cade, R., Dietrich, R. A., Rogers, J., Knapp, S., Fei, Z. J., White, R. A., Thannhauser, T. W., Giovannoni, J. J., Botella, M. A., Gilbert, L., Gonzalez, R., Goicoechea, J. L., Yu, Y., Kudrna, D., Collura, K., Wissotski, M., Wing, R., Schoof, H., Meyers, B. C., Gurazada, A. B., Green, P. J., Mathur, S., Vyas, S., Solanke, A. U., Kumar, R., Gupta, V., Sharma, A. K., Khurana, P., Khurana, J. P., Tyagi, A. K., Dalmay, T., Mohorianu, I., Walts, B., Chamala, S., Barbazuk, W. B., Li, J. P., Guo, H., Lee, T. H., Wang, Y. P., Zhang, D., Paterson, A. H., Wang, X. Y., Tang, H. B., Barone, A., Chiusano, M. L., Ercolano, M. R., D'Agostino, N., Di Filippo, M., Traini, A., Sanseverino, W., Frusciante, L., Seymour, G. B., Elharam, M., Fu, Y., Hua, A., Kenton, S., Lewis, J., Lin, S. P., Najar, F., Lai, H. S., Qin, B. F., Qu, C. M., Shi, R. H., White, D., White, J., Xing, Y. B., Yang, K. Q., Yi, J., Yao, Z. Y., Zhou, L. P., Roe, B. A., Vezzi, A., D'Angelo, M., Zimbello, R., Schiavon, R., Caniato, E., Rigobello, C., Campagna, D., Vitulo, N., Valle, G., Nelson, D. R., De Paoli, E., Szinay, D., de Jong, H. H., Bai, Y. L., Visser, R. G. F., Lankhorst, R. M. K., Beasley, H., McLaren, K., Nicholson, C., Riddle, C., Gianese, G. and Consortium, Tomato Genome (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., Mongin, E., Barker, G., Spangenberg, G. C. and Edwards, D. (2005) SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res* **33**, W493-495.

Schattner, P., Brooks, A. N. and Lowe, T. M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686-W689.

Schatz, M. C. (2009) CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25**, 1363-1369.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J. X., Mitros, T., Nelson, W., Hyten, D. L., Song, Q. J., Thelen, J. J., Cheng, J. L., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S. Q., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J. C., Tian, Z. X., Zhu, L. C., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X. C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C. and Jackson, S. A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183.

Schnable, P. S. (2012) The B73 maize genome: Complexity, diversity, and dynamics (November, pg 1112, 2009). *Science* **337**, 1040-1040.

Schnable, P. S., Hsia, A. P. and Nikolau, B. J. (1998) Genetic recombination in plants. *Curr Opin Plant Biol* **1**, 123-129.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**.

Sedlazeck, F. J., Rescheneder, P. and von Haeseler, A. (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791.

Sharma, K. D., Winter, P., Kahl, G. and Muehlbauer, F. J. (2004) Molecular mapping of Fusarium oxysporum f. sp ciceris race 3 resistance gene in chickpea. *Theor Appl Genet* **108**, 1243-1248.

Sharpe, A. G., Ramsay, L., Sanderson, L. A., Fedoruk, M. J., Clarke, W. E., Li, R., Kagale, S., Vijayan, P., Vandenberg, A. and Bett, K. E. (2013) Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. *Bmc Genomics* **14**.

Shendure, J. and Ji, H. L. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J. M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T. L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W. Q., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E. and Folta, K. M. (2011) The genome of woodland strawberry (Fragaria vesca). *Nat Genet* **43**, 109-116.

Simpson, J. T. and Durbin, R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**, 549-556.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, I. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123.

Singh F, Diwakar B (1995) *Chickpea botany and production practices*. Hyderabad:ICRISAT.

Singh, K. B., Malhotra, R. S., Halila, M. H., Knights, E. J. and Verma, M. M. (1994) Current Status and Future Strategy in Breeding Chickpea for Resistance to Biotic and Abiotic Stresses. *Euphytica* **73**, 137-149.

Singh.U, Kumar.J, Jambunathan.R, Smithson.J.B (1980) Variability in the seed coat content of desi and kabuli chickpea cultivars. In: *Chickpea Newsl* p. 18.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. and Holmes, I. H. (2009) JBrowse: A next-generation genome browser. *Genome Res* **19**, 1630-1638.

Smith, A. D., Xuan, Z. Y. and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics* **9**.

Soderlund, C., Bomhoff, M. and Nelson, W. M. (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39**.

Springer, N. M., Ying, K., Fu, Y., Ji, T. M., Yeh, C. T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. A., Nettleton, D. and Schnable, P. S. (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *Plos Genet* **5**.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-W439.

Stein, L. D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform* **14**, 162-171.

Sudupak, M. A., Akkaya, M. S. and Kence, A. (2002) Analysis of genetic relationships among perennial and annual Cicer species growing in Turkey using RAPD markers. *Theor Appl Genet* **105**, 1220-1228.

Sudupak, M. A., Akkaya, M. S. and Kence, A. (2004) Genetic relationships among perennial and annual Cicer species growing in Turkey assessed by AFLP fingerprinting. *Theor Appl Genet* **108**, 937-944.

Sun, J. B., Patra, J. C. and Li, Y. J. (2009) Functional Link Artificial Neural Network-based Disease Gene Prediction. *Ieee Ijcnn*, 425-432.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28**, 2731-2739.

Tang, H. B., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A. N., Zhou, S. G., Gentzbittel, L., Childs, K. L., Yandell, M., Gundlach, H., Mayer, K. F. X., Schwartz, D. C. and Town, C. D. (2014) An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *Bmc Genomics* **15**.

Tar'an, B., Warkentin, T. D., Tullu, A. and Vandenberg, A. (2007) Genetic mapping of ascochyta blight resistance in chickpea (Cicer arietinum L.) using a simple sequence repeat linkage map. *Genome* **50**, 26-34.

Tayyar, R., Federici, C. V. and Waines, J. G. (1996) Natural outcrossing in chickpea (Cicer arietinum L). *Crop Sci* **36**, 203-205.

Thiel, T., Michalek, W., Varshney, R. K. and Graner, A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* **106**, 411-422.

Thomas, J. M. G., Boote, K. J., Allen, L. H., Gallo-Meagher, M. and Davis, J. M. (2003) Elevated temperature and carbon dioxide effects on soybean seed composition and transcript abundance. *Crop Sci* **43**, 1548-1557.

Thorvaldsdottir, H., Robinson, J. T. and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192.

Thudi, M., Bohra, A., Nayak, S. N., Varghese, N., Shah, T. M., Penmetsa, R. V., Thirunavukkarasu, N., Gudipati, S., Gaur, P. M., Kulwal, P. L., Upadhyaya, H. D., KaviKishor, P. B., Winter, P., Kahl, G., Town, C. D., Kilian, A., Cook, D. R. and Varshney, R. K. (2011) Novel SSR Markers from BAC-End Sequences, DArT Arrays and a Comprehensive Genetic Map with 1,291 Marker Loci for Chickpea (Cicer arietinum L.). *Plos One* **6**.

Thudi, M., Upadhyaya, H. D., Rathore, A., Gaur, P. M., Krishnamurthy, L., Roorkiwal, M., Nayak, S. N., Chaturvedi, S. K., Basu, P. S., Gangarao, N. V. P. R., Fikre, A., Kimurto, P., Sharma, P. C., Sheshashayee, M. S., Tobita, S., Kashiwagi, J., Ito, O., Killian, A. and Varshney, R. K. (2014) Genetic Dissection of Drought and Heat Tolerance in Chickpea through Genome-Wide and Candidate Gene-Based Association Mapping Approaches. *Plos One* **9**.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. and Rokhsar, D. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-1604.

Udupa, S. M., Robertson, L. D., Weigand, F., Baum, M. and Kahl, G. (1999) Allelic variation at (TAA)(n) microsatellite loci in a world collection of chickpea (Cicer arletinum L.) germplasm. *Mol Gen Genet* **261**, 354-363.

Upadhyaya, H. D., Thudi, M., Dronavalli, N., Gujaria, N., Singh, S., Sharma, S. and Varshney, R. K. (2011) Genomic tools and germplasm diversity for chickpea improvement. *Plant Genet Resour-C* **9**, 45-58.

Upstill-Goddard, R., Eccles, D., Fliege, J. and Collins, A. (2013) Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief Bioinform* **14**, 251-260.

van der Maesen, L. J. G (1987) Origin, history and taxonomy of chickpea. In: *The Chickpea* (M. C. Saxena and Singh, K.B. eds), p. 34. Wallingford, UK: CAB International.

van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H. and van Eijk, M. J. T. (2007) Complexity Reduction of Polymorphic Sequences (CRoPS (TM)): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *Plos One* **2**.

Varma Penmetsa, R., Carrasquilla-Garcia, Noelia, Bergmann, Emily M., Vance, Lisa, Castro, Brenna, Kassa, Mulualem T., Sarma, Birinchi K., Datta, Subhojit, Farmer, Andrew D., Baek, Jong-Min, Coyne, Clarice J., Varshney, Rajeev K., von Wettberg, Eric J. B. and Cook, Douglas R. (2016) Multiple post-domestication origins of kabuli chickpea through allelic variation in a diversification-associated transcription factor. *New Phytologist*, n/a-n/a.

Varshney, R. K., Chen, W. B., Li, Y. P., Bharti, A. K., Saxena, R. K., Schlueter, J. A., Donoghue, M. T. A., Azam, S., Fan, G. Y., Whaley, A. M., Farmer, A. D., Sheridan, J., Iwata, A., Tuteja, R., Penmetsa, R. V., Wu, W., Upadhyaya, H. D., Yang, S. P., Shah, T., Saxena, K. B., Michael, T., McCombie, W. R., Yang, B. C., Zhang, G. Y., Yang, H. M., Wang, J., Spillane, C., Cook, D. R., May, G. D., Xu, X. and Jackson, S. A. (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* **30**, 83-U128.

Varshney, R. K., Nayak, S. N., May, G. D. and Jackson, S. A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* **27**, 522-530.

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., Cannon, S., Baek, J., Rosen, B. D., Tar'an, B., Millan, T., Zhang, X. D., Ramsay, L. D., Iwata, A., Wang, Y., Nelson, W., Farmer, A. D., Gaur, P. M., Soderlund, C., Penmetsa, R. V., Xu, C. Y., Bharti, A. K., He, W. M., Winter, P., Zhao, S. C., Hane, J. K., Carrasquilla-Garcia, N., Condie, J. A., Upadhyaya, H. D., Luo, M. C., Thudi, M., Gowda, C. L. L., Singh, N. P., Lichtenzveig, J., Gali, K. K., Rubio, J., Nadarajan, N., Dolezel, J., Bansal, K. C., Xu, X., Edwards, D., Zhang, G. Y., Kahl, G., Gil, J., Singh, K. B., Datta, S. K., Jackson, S. A., Wang, J. and Cook, D. R. (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat Biotechnol* **31**, 240-246.

Villani, Sara M., Ayer, Katrin and Cox, Kerik D. (2016) Molecular Characterization of the sdhB Gene and Baseline Sensitivity to Penthiopyrad, Fluopyram, and Benzovindiflupyr in Venturia inaequalis. *Plant Disease*, PDIS-12-15-1512-RE.

Vlacilova, K., Ohri, D., Vrana, J., Cihalikova, J., Kubalakova, M., Kahl, G. and Dolezel, J. (2002) Development of flow cytogenetics and physical genome mapping in chickpea (Cicer arietinum L.). *Chromosome Res* **10**, 695-706.

Wang, M. H., Yu, Y., Haberer, G., Marri, P. R., Fan, C. Z., Goicoechea, J. L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J. S. S., Cossu, R. M., Maldonado, C., Chen, J., Lee, S., Sisneros, N., de Baynast, K., Golser, W., Wissotski, M., Kim, W., Sanchez, P., Ndjiondjop, M. N., Sanni, K., Long, M. Y., Carney, J., Panaud, O., Wicker, T., Machado, C. A., Chen, M. S., Mayer, K. F. X., Rounsley, S. and Wing, R. A. (2014) The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat Genet* **46**, 982-+.

Weese, D., Holtgrewe, M. and Reinert, K. (2012) RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics* **28**, 2592-2599.

Yadav, S.S., Kumar, J., Yadav, S.K., Singh, S., Yadav, V.S., Turner, N.C. & Redden, R (2006) Evaluation of Helicoverpa and drought resistance in desi and kabuli chickpea. *Plant Genetic Resources* **4(3)**, 198-203.

Yip, K. Y., Cheng, C. and Gerstein, M. (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol* **14**.

You, F. M., Deal, K. R., Wang, J. R., Britton, M. T., Fass, J. N., Lin, D. W., Dandekar, A. M., Leslie, C. A., Aradhya, M., Luo, M. C. and Dvorak, J. (2012) Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. *Bmc Genomics* **13**.

Young, N. D., Debelle, F., Oldroyd, G. E. D., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., Mayer, K. F. X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S. G., Mudge, J., Bharti, A. K., Murray, J. D., Naoumkina, M. A., Rosen, B., Silverstein, K. A. T., Tang, H. B., Rombauts, S., Zhao, P. X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X. B., Doyle, J. J., Dudez, A. M., Farmer, A. D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A. J., Green, P. J., Hallab, A., Hartog, M., Hua, A., Humphray, S. J., Jeong, D. H., Jing, Y., Jocker, A., Kenton, S. M., Kim, D. J., Klee, K., Lai, H. S., Lang, C. T., Lin, S. P., Macmil, S. L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E. L., Mun, J. H., Najar, F. Z., Nicholson, C., Noirot, C., O'Bleness, M., Paule, C. R., Poulain, J., Prion, F., Qin, B. F., Qu, C. M., Retzel, E. F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A. J., Sherrier, D. J., Shi, R. H., Sims, S., Singer, S. R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B. B., Wang, K. Q., Wang, M. Y., Wang, X. H., Warfsmann, J., Weissenbach, J., White, D. D., White, J. D., Wiley, G. B., Wincker, P., Xing, Y. B., Yang, L. M., Yao, Z. Y., Ying, F., Zhai, J. X., Zhou, L. P., Zuber, A., Denarie, J., Dixon, R. A., May, G. D., Schwartz, D. C., Rogers, J., Quetier, F., Town, C. D. and Roe, B. A. (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524.

Yu, J., Hu, S. N., Wang, J., Wong, G. K. S., Li, S. G., Liu, B., Deng, Y. J., Dai, L., Zhou, Y., Zhang, X. Q., Cao, M. L., Liu, J., Sun, J. D., Tang, J. B., Chen, Y. J., Huang, X. B., Lin, W., Ye, C., Tong, W., Cong, L. J., Geng, J. N., Han, Y. J., Li, L., Li, W., Hu, G. Q., Huang, X. G., Li, W. J., Li, J., Liu, Z. W., Li, L., Liu, J. P., Qi, Q. H., Liu, J. S., Li, L., Li, T., Wang, X. G., Lu, H., Wu, T. T., Zhu, M., Ni, P. X., Han, H., Dong, W., Ren, X. Y., Feng, X. L., Cui, P., Li, X. R., Wang, H., Xu, X., Zhai, W. X., Xu, Z., Zhang, J. S., He, S. J., Zhang, J. G., Xu, J. C., Zhang, K. L., Zheng, X. W., Dong, J. H., Zeng, W. Y., Tao, L., Ye, J., Tan, J., Ren, X. D., Chen, X. W., He, J., Liu, D. F., Tian, W., Tian, C. G., Xia, H. G., Bao, Q. Y., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W. M., Li, P., Chen, W., Wang, X. D., Zhang, Y., Hu, J. F., Wang, J., Liu, S., Yang, J., Zhang, G. Y., Xiong, Y. Q., Li, Z. J., Mao, L., Zhou, C. S., Zhu, Z., Chen, R. S., Hao, B. L., Zheng, W. M., Chen, S. Y., Guo, W., Li, G. J., Liu, S. Q., Tao, M., Wang, J., Zhu, L. H., Yuan, L. P. and Yang, H. M. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp indica). *Science* **296**, 79-92.

Yuzuki, D (2015) BGISEQ-500 debuts at the international congress of genomics 10.: Next Generation Technologist.

Zerbino, D. R. and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829.

Zhang, J., Wheeler, D. A., Yakub, I., Wei, S., Sood, R., Rowe, W., Liu, P. P., Gibbs, R. A. and Buetow, K. H. (2005) SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS computational biology* **1**, e53.

Zhao, S. Q., Wang, J., Zhang, L., Li, J. T., Gu, X. C., Gao, G. and Wei, L. P. (2009) BOAT: Basic Oligonucleotide Alignment Tool. *Bmc Genomics* **10**.

Zhu, P. Y., He, L. Y., Li, Y. Q., Huang, W. P., Xi, F., Lin, L., Zhi, Q. H., Zhang, W. W., Tang, Y. T., Geng, C. Y., Lu, Z. Y. and Xu, X. (2015) OTG-snpcaller: An Optimized Pipeline Based on TMAP and GATK for SNP Calling from Ion Torrent Data (vol 9, e97507, 2014). *Plos One* **10**.

Zohary D, Hopf M, Weiss E. (2012) *Domestication of plants in the Old World: the origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford:Oxford University Press.

Zohary, D., Hopf, M (1993) *The Origin and Spread of Cultivated Plants in West Asia, Europe, and the Nile Valley*. Oxford, UK:Oxford Univ. Press.

Appendix 1: ICC4958xPI489777 population data

Library Name	Number of Single Reads	Number of Paried Reads	Total # bp	Estimated Coverage
PI489777XICC4958_RIL43_PE_1	6,245,346	3,122,673	624,534,600	0.84
PI489777XICC4958_RIL44_PE_1	4,738,574	2,369,287	473,857,400	0.64
PI489777XICC4958_RIL45_PE_1	4,846,926	2,423,463	484,692,600	0.65
PI489777XICC4958_RIL46_PE_1	3,900,494	1,950,247	390,049,400	0.52
PI489777XICC4958_RIL47_PE_1	4,845,098	2,422,549	484,509,800	0.65
PI489777XICC4958_RIL48_PE_1	4,642,482	2,321,241	464,248,200	0.62
PI489777XICC4958_RIL49_PE_1	4,369,786	2,184,893	436,978,600	0.59
PI489777XICC4958_RIL50_PE_1	5,058,830	2,529,415	505,883,000	0.68
PI489777XICC4958_RIL1_PE_2	3,684,872	1,842,436	368,487,200	0.49
PI489777XICC4958_RIL2_PE_2	5,205,294	2,602,647	520,529,400	0.70
PI489777XICC4958_RIL3_PE_2	2,747,662	1,373,831	274,766,200	0.37
PI489777XICC4958_RIL4_PE_2	5,444,572	2,722,286	544,457,200	0.73
PI489777XICC4958_RIL5_PE_2	4,591,078	2,295,539	459,107,800	0.62
PI489777XICC4958_RIL6_PE_2	3,629,626	1,814,813	362,962,600	0.49
PI489777XICC4958_RIL7_PE_2	2,320,068	1,160,034	232,006,800	0.31
PI489777XICC4958_RIL8_PE_2	4,361,032	2,180,516	436,103,200	0.58
PI489777XICC4958_RIL10_PE_2	2,796,146	1,398,073	279,614,600	0.37
PI489777XICC4958_RIL11_PE_2	2,117,620	1,058,810	211,762,000	0.28
PI489777XICC4958_RIL12_PE_2	962,922	481,461	96,292,200	0.13
PI489777XICC4958_RIL13_PE_2	1,657,602	828,801	165,760,200	0.22
PI489777XICC4958_RIL14_PE_2	1,595,262	797,631	159,526,200	0.21
PI489777XICC4958_RIL15_PE_2	1,226,102	613,051	122,610,200	0.16
PI489777XICC4958_RIL16_PE_2	3,740,960	1,870,480	374,096,000	0.50
PI489777XICC4958_RIL17_PE_2	2,454,692	1,227,346	245,469,200	0.33
PI489777XICC4958_RIL18_PE_2	2,820,084	1,410,042	282,008,400	0.38
PI489777XICC4958_RIL19_PE_2	4,250,980	2,125,490	425,098,000	0.57
PI489777XICC4958_RIL20_PE_2	6,495,916	3,247,958	649,591,600	0.87
PI489777XICC4958_RIL21_PE_2	4,303,956	2,151,978	430,395,600	0.58
PI489777XICC4958_RIL22_PE_2	3,205,260	1,602,630	320,526,000	0.43
PI489777XICC4958_RIL23_PE_2	5,937,812	2,968,906	593,781,200	0.80
PI489777XICC4958_RIL24_PE_2	4,271,118	2,135,559	427,111,800	0.58
PI489777XICC4958_RIL25_PE_2	3,084,224	1,542,112	308,422,400	0.42
PI489777XICC4958_RIL26_PE_2	4,431,488	2,215,744	443,148,800	0.60
PI489777XICC4958_RIL27_PE_2	2,772,592	1,386,296	277,259,200	0.37
PI489777XICC4958_RIL28_PE_2	3,471,898	1,735,949	347,189,800	0.47
PI489777XICC4958_RIL29_PE_2	5,599,984	2,799,992	559,998,400	0.76
PI489777XICC4958_RIL30_PE_2	3,599,042	1,799,521	359,904,200	0.49
PI489777XICC4958_RIL31_PE_2	5,424,262	2,712,131	542,426,200	0.73
PI489777XICC4958_RIL32_PE_2	3,485,042	1,742,521	348,504,200	0.47

PI489777XICC4958_RIL33_PE_2	1,160,832	580,416	116,083,200	0.16
PI489777XICC4958_RIL34_PE_2	4,481,104	2,240,552	448,110,400	0.61
PI489777XICC4958_RIL35_PE_2	2,737,340	1,368,670	273,734,000	0.37
PI489777XICC4958_RIL36_PE_2	4,967,114	2,483,557	496,711,400	0.67
PI489777XICC4958_RIL37_PE_2	5,896,262	2,948,131	589,626,200	0.80
PI489777XICC4958_RIL39_PE_2	3,055,534	1,527,767	305,553,400	0.41
PI489777XICC4958_RIL40_PE_2	3,452,314	1,726,157	345,231,400	0.47
PI489777XICC4958_RIL41_PE_2	3,386,416	1,693,208	338,641,600	0.46
PI489777XICC4958_RIL42_PE_2	2,585,090	1,292,545	258,509,000	0.35
PI489777XICC4958_RIL43_PE_2	2,403,796	1,201,898	240,379,600	0.32
PI489777XICC4958_RIL44_PE_2	3,286,314	1,643,157	328,631,400	0.44
PI489777XICC4958_RIL45_PE_2	4,609,384	2,304,692	460,938,400	0.62
PI489777XICC4958_RIL46_PE_2	4,071,794	2,035,897	407,179,400	0.55
PI489777XICC4958_RIL47_PE_2	2,637,320	1,318,660	263,732,000	0.36
PI489777XICC4958_RIL48_PE_2	3,870,340	1,935,170	387,034,000	0.52
PI489777XICC4958_RIL49_PE_2	4,522,322	2,261,161	452,232,200	0.61

Appendix 2: *Desi* and *kabuli* reference genome statistics

	***Kabuli* Reference**		***Desi* Reference**	
Chromosome length in bp	**Varshney et al**	**Version2.6.2**	**Jain et al**	**Version3.0**
Ca1	48,359,943	49,034,091	14,791,696	44,557,692
Ca2	36,634,854	37,205,496	17,304,114	33,970,409
Ca3	39,989,001	66,825,002	23,376,002	61,121,219
Ca4	49,191,682	58,848,791	22,093,647	50,120,497
Ca5	48,169,137	69,325,083	16,301,343	63,456,816
Ca6	59,463,898	65,873,034	11,482,212	59,874,348
Ca7	48,961,560	56,246,106	8,461,617	50,353,284
Ca8	16,477,302	19,930,104	10,574,966	18,298,786
Total chromosome sequence length	347,247,377	423,287,707	124,385,597	381,753,051
Total GC count:	94,556,178 bp	138,503,164 bp	139,977,726 bp	128,643,194 bp
GC %:	27.23%	29.34%	26.93%	30.12 %
Genome stats				
Total Number of Contigs:	7,163	5,176	181,462	3,321

Total Length of Contigs:	532,289,632 bp	472,095,590 bp	519,846,222 bp	427,165,656 bp
Total Length of Contigs (excluding N's):	450,007,017 bp (15.46% N's)	449,962,214 bp (4.69% N's)	448,120,765 bp(13.80% N's)	403,568,312 bp (5.52% N's)
Largest Contig:	59,463,898 bp	69,325,083 bp	23,376,002 bp	63,456,816 bp
Smallest Contig:	1,000 bp	1,000 bp	200 bp	1,000 bp
Contigs 5000bp or larger:	2,394	1,122	8,990	832
Total bases	523,805,173 bp	465,053,891 bp	410,261,893 bp	422,849,374 bp
Percentage	98.41%	98.51%	78.92%	98.99%
Contigs 1000bp or larger:	7,163	5,176	32,968	3,321
Total bases	532,289,632 bp	472,095,590 bp	456,440,147 bp	427,165,656 bp
Percentage	100.00%	100.00%	87.80%	100.00%

Appendix 3: Whole genome sequence data for 47 chickpea varieties (data produced for 18 varieties and 29 chickpea varieties data downloaded from ICRISAT)

Variety/Accession	**Type**	**Paired reads**	**Coverage**
FLIP94-509C--97C016-20	*Desi*	40,982,925	11.0x
ICCV96970--ICCV04516-9	*Desi*	33,003,314	8.9x
ICCV92809--ICCV04516	*Desi*	35,553,313	9.6x
8627P-2--ICC13729	*Desi*	46,258,419	12.5x
97020-1147--ICC8294	*Desi*	41,874,591	11.3x
ICCV98503--Moti	*Desi*	27,265,658	7.3x
Rupali--97037-1178-1001	*Desi*	30,232,199	8.17x
ICCV96836--PBG5	*Desi*	29,819,898	8.05x
FLIP94-079C--01-481*03HS003	*Desi*	56,933,653	15.3x
GENESIS510--KYABRA	*Desi*	38,835,306	10.4x
ICCV3996--ICCV93954	*Desi*	32,862,755	8.88x
ICCV96836--ICC12004	*Desi*	38,860,925	10.5x
Sonali--ICC12004	*Desi*	40,711,212	11.0x
ICC1882	*Desi*	29,731,490	8.3x

ICC283	*Desi*	27,661,086	7.47x
ICC4958	*Desi*	29,523,997	7.97x
ICCV88202	*Desi*	36,726,403	10.0x
ICCV88516	*Desi*	43,699,669	11.9x
Myles	*Desi*	49,026,648	13.3x
JG62	*Desi*	36,748,200	10.0x
CDC512-51	*Desi*	18,372,733	5.01x
CDCCorinne	*Desi*	39,452,181	10.7x
ICCV93954	*Desi*	38,244,146	10.4x
CDCVanguard	*Desi*	36,552,733	9.97x
ICCV96836	*Desi*	29,335,784	8.00x
Pistol	*Desi*	33,467,106	9.13x
HatTrick	*Desi*	32,021,614	8.74x
Slasher	*Desi*	31,093,427	8.48x
ICCV93952	*Desi*	32,690,755	8.92x
ICCV92944	*Desi*	48,813,407	13.3x
WR315	*Desi*	37,181,692	10.2x
ICCV97105	*Desi*	27,953,682	14.8x
ICCL83105	*Desi*	25,566,405	6.97x
ICC8261	*Kabuli*	33,652,270	9.09x
BL	*Kabuli*	38,691,064	10.5x
YeZIN	*Kabuli*	43,773,290	11.9x
ILC3279	*Kabuli*	45,607,807	12.4x
ICCV92311	*Kabuli*	43,347,618	11.8x
Amit	*Kabuli*	32,947,889	8.9x
CDCLeader	*Kabuli*	30,934,656	8.4x
CDCLuna	*Kabuli*	34,936,166	9.5x
Genesis090	*Kabuli*	32,210,496	8.79x
Chefe	*Kabuli*	28,627,727	7.8x
ICCC32	*Kabuli*	30,618,033	8.35x
ICCV92337	*Kabuli*	24,438,550	6.67x
ICCV93512	*Kabuli*	29,039,467	7.9x
PI489777	Wild	36,062,972	9.7x

Appendix 4: Whole genome sequence read mapping distribution

				Unmapped Reads		
Accessions	**Total Reads**	**Paired (%)**	**Single (%)**	**Unmapped Pair**	**Unmapped Single**	**Unmapped (%)**
CICA1122	95,744,680	53,201,150(66.6%)	10,733,690(13.4%)	9,064,762	6,840,158	15,904,920(19.9%)
DICC8296	82,899,161	35,732,004(55.1%)	11,153,039(17.2%)	10,873,576	7,133,483	18,007,059(27.7%)
DICC8310	88,101,338	40,096,500(57.4%)	11,412,998(16.3%)	11,028,322	7,267,598	18,295,920(26.2%)
WACPE2160	111,855,695	54,241,164(60.3%)	13,790,677(15.3%)	13,101,246	8,810,681	21,911,927(24.4%)
05H072D-08HS3001	98,252,023	52,616,020(64.8%)	11,442,277(14.1%)	9,877,206	7,219,657	17,096,863(21.1%)
DICC8022	64,906,917	34,439,562(64.3%)	7,738,901(14.5%)	6,291,868	5,072,359	11,364,227(21.2%)
04C830-07WAD2	72,528,815	37,616,328(63.4%)	8,601,861(14.5%)	7,730,282	5,425,031	13,155,313(22.2%)
DICC8174	72,249,171	36,203,234(61.7%)	8,834,995(15.1%)	7,933,922	5,671,549	13,605,471(23.2%)
05H172K-08HO4009	132,254,225	70,664,532(64.7%)	15,351,559(14.1%)	13,559,368	9,559,699	23,119,067(21.2%)
D05087>F4TMWR2AB015	91,949,381	48,414,622(64.0%)	10,819,189(14.3%)	9,680,550	6,677,235	16,357,785(21.6%)
DICC8196	80,661,863	38,453,726(59.5%)	10,195,603(15.8%)	9,531,300	6,474,967	16,006,267(24.8%)
DICC8212	95,608,375	44,961,084(58.8%)	12,254,009(16.0%)	11,197,722	7,998,919	19,196,641(25.1%)
DICC8220	98,890,005	48,927,214(61.1%)	12,232,481(15.3%)	11,080,300	7,784,855	18,865,155(23.6%)
ICC1882	72,762,214	29,933,782(53.4%)	9,490,956(16.9%)	10,372,790	6,295,948	16,668,738(29.7%)
ICC283	69,354,436	24,934,824(48.2%)	9,228,832(17.8%)	11,535,434	6,059,956	17,595,390(34.0%)
ICC4958	73,256,616	28,100,672(50.9%)	9,024,240(16.4%)	12,092,118	5,973,734	18,065,852(32.7%)
ICC8261	83,627,913	31,359,780(50.0%)	10,416,663(16.6%)	14,022,080	6,903,655	20,925,735(33.4%)

PI489777	91,434,199	26,827,586(40.7%)	13,509,011(20.5%)	16,155,316	9,393,485	25,548,801(38.8%)
ICCV88202	85,825,561	45,081,384(64.4%)	9,133,723(13.0%)	9,643,710	6,161,517	15,805,227(22.6%)
BL	92,344,236	47,742,022(63.7%)	9,840,566(13.1%)	10,734,974	6,645,850	17,380,824(23.2%)
ICCV88516	106,172,703	51,542,362(60.8%)	11,936,847(14.1%)	13,292,638	8,054,109	21,346,747(25.2%)
YeZIN	98,322,645	60,629,532(72.2%)	9,078,599(10.8%)	8,044,466	6,262,791	14,307,257(17.0%)
ILC3279	102,436,490	57,624,778(67.7%)	10,129,748(11.9%)	10,403,056	6,937,926	17,340,982(20.4%)
ICCV92311	101,090,337	52,750,564(64.1%)	10,714,807(13.0%)	11,654,896	7,157,587	18,812,483(22.9%)
Myles	108,479,213	64,300,348(70.3%)	10,079,863(11.0%)	10,038,208	7,011,293	17,049,501(18.6%)
JG62	86,997,517	45,211,752(63.9%)	9,350,119(13.2%)	9,861,828	6,355,995	16,217,823(22.9%)
CDC512-51	39,666,280	27,871,654(78.9%)	3,092,278(8.8%)	2,316,198	2,034,976	4,351,174(12.3%)
Amit	77,048,879	42,732,606(66.8%)	8,188,015(12.8%)	7,907,110	5,157,019	13,064,129(20.4%)
CDCLeader	64,750,855	44,423,248(77.7%)	5,147,613(9.0%)	4,178,670	3,411,327	7,589,997(13.3%)
CDCCorinne	77,810,174	48,575,730(72.7%)	7,296,052(10.9%)	6,180,672	4,788,524	10,969,196(16.4%)
ICCV93954	84,379,830	49,434,280(69.2%)	8,970,954(12.6%)	5,628,914	7,358,384	12,987,298(18.2%)
CDCLuna	70,889,571	42,382,694(70.5%)	7,010,367(11.7%)	4,911,464	5,836,791	10,748,255(17.9%)
CDCVanguard	83,524,831	45,128,486(65.5%)	9,046,307(13.1%)	7,270,056	7,404,963	14,675,019(21.3%)
ICCV96836	59,919,706	34,300,868(68.3%)	6,246,466(12.4%)	4,535,036	5,151,150	9,686,186(19.3%)
Pistol	26,003,682	15,162,000(69.4%)	2,504,646(11.5%)	2,382,372	1,786,146	4,168,518(19.1%)
HatTrick	25,648,158	14,814,622(69.0%)	2,452,712(11.4%)	2,506,096	1,684,316	4,190,412(19.5%)
Slasher	19,937,205	11,697,496(69.8%)	1,902,407(11.3%)	1,830,802	1,337,849	3,168,651(18.9%)
Genesis090	73,381,656	40,329,952(66.3%)	7,856,272(12.9%)	6,544,756	6,052,960	1,2597,716(20.7%)
ICCV93952	66,686,398	33,017,014(61.3%)	7,978,852(14.8%)	6,624,676	6,220,590	12,845,266(23.9%)
ICCV92944	119,255,514	51,719,770(55.7%)	14,647,632(15.8%)	14,891,744	11,552,312	26,444,056(28.5%)

Chefe	70,258,796	32,447,392(58.4%)	8,502,172(15.3%)	8,029,368	6,625,248	14,654,616(26.4%)
WR315	70,263,691	33,604,648(59.9%)	8,329,719(14.8%)	7,709,242	6,455,420	14,164,662(25.2%)
ICCC32	60,287,869	27,294,618(57.6%)	7,197,809(15.2%)	7,318,756	5,578,965	12,897,721(27.2%)
ICCV92337	59,866,791	27,844,050(58.8%)	7,068,635(14.9%)	6,978,048	5,499,005	12,477,053(26.3%)
ICCV93512	69,387,078	33,226,336(60.1%)	8,023,542(14.5%)	7,784,432	6,284,168	14,068,600(25.4%)
ICCV97105	67,813,255	31,697,324(59.0%)	7,966,449(14.8%)	7,858,616	6,216,125	14,074,741(26.2%)
ICCL83105	63,834,379	27,830,020(56.0%)	7,792,485(15.7%)	8,106,720	5,999,217	14,105,937(28.4%)

Appendix 5: Heterozygosity and homozygous SNPs observed in each chickpea variety

		Heterozygous		Homozygous	
Chickpea variety	**Type**	**SNP**	**Density Kb-1**	**SNPs**	**Density Kb-1**
CDC512-51	*Desi*	6,04	0.01	516,09	0.69
ICCL83105	*Desi*	13,7	0.02	517,47	0.69
WR315	*Desi*	17,7	0.03	520,31	0.70
CDCCorinne	*Desi*	19,1	0.03	524,84	0.70
JG62	*Desi*	20,7	0.03	533,91	0.72
DICC8022	*Desi*	22,5	0.04	526,38	0.71
Slasher	*Desi*	22,7	0.04	531,54	0.71
ICCV97105	*Desi*	23,3	0.04	532,11	0.71
ICC1882	*Desi*	24,0	0.04	524,34	0.70
HatTrick	*Desi*	24,5	0.04	555,12	0.75
ICC283	*Desi*	25,6	0.04	480,32	0.64
04C830-07WAD2	*Desi*	27,2	0.05	532,66	0.71
DICC8174	*Desi*	27,5	0.05	528,13	0.71
Pistol	*Desi*	30,2	0.05	538,97	0.72
ICC4958	*Desi*	30,2	0.05	501,74	0.67
ICCV96836	*Desi*	34,4	0.06	473,56	0.63
ICCV88516	*Desi*	34,6	0.06	540,80	0.73
DICC8296	*Desi*	35,4	0.06	539,12	0.72
DICC8310	*Desi*	39,4	0.07	544,43	0.73
D05087>F4TMWR2AB01	*Desi*	39,7	0.07	541,52	0.73
ICCV93952	*Desi*	42,4	0.07	529,84	0.71
Myles	*Desi*	45,8	0.08	546,14	0.73
DICC8212	*Desi*	45,9	0.08	552,76	0.74
ICCV92944	*Desi*	46,1	0.08	535,19	0.72
05H072D-08HS3001	*Desi*	49,0	0.09	540,87	0.73
DICC8220	*Desi*	49,3	0.09	545,72	0.73
ICCV88202	*Desi*	50,7	0.09	539,08	0.72
ICCV93954	*Desi*	52,3	0.09	515,39	0.69
CDCVanguard	*Desi*	52,7	0.09	517,67	0.69
CICA1122	*Desi*	53,2	0.09	541,12	0.73
DICC8196	*Desi*	67,9	0.12	550,68	0.74
WACPE2160	*Desi*	74,6	0.13	545,14	0.73
05H172K-08HO4009	*Desi*	77,1	0.14	552,78	0.74
ICCV92337	*Kabuli*	12,9	0.02	519,22	0.70
CDCLeader	*Kabuli*	16,7	0.03	538,87	0.72
Amit	*Kabuli*	17,7	0.03	554,90	0.74
Chefe	*Kabuli*	18,4	0.03	526,04	0.71
ICCC32	*Kabuli*	18,7	0.03	529,56	0.71
ICCV93512	*Kabuli*	20,8	0.03	526,34	0.71
BL	*Kabuli*	22,2	0.04	540,62	0.73
Genesis090	*Kabuli*	24,2	0.04	534,00	0.72
ICCV92311	*Kabuli*	31,5	0.05	543,35	0.73

ICC8261	*Kabuli*	39,1	0.07	510,80	0.69
YeZIN	*Kabuli*	52,9	0.09	537,07	0.72
CDCLuna	*Kabuli*	71,6	0.13	472,36	0.63
ILC3279	*Kabuli*	94,3	0.17	562,59	0.76
PI489777	Wild	75,0	0.14	491,89	0.66

Appendix 6: Gene loss events identified in all 47 chickpea varieties

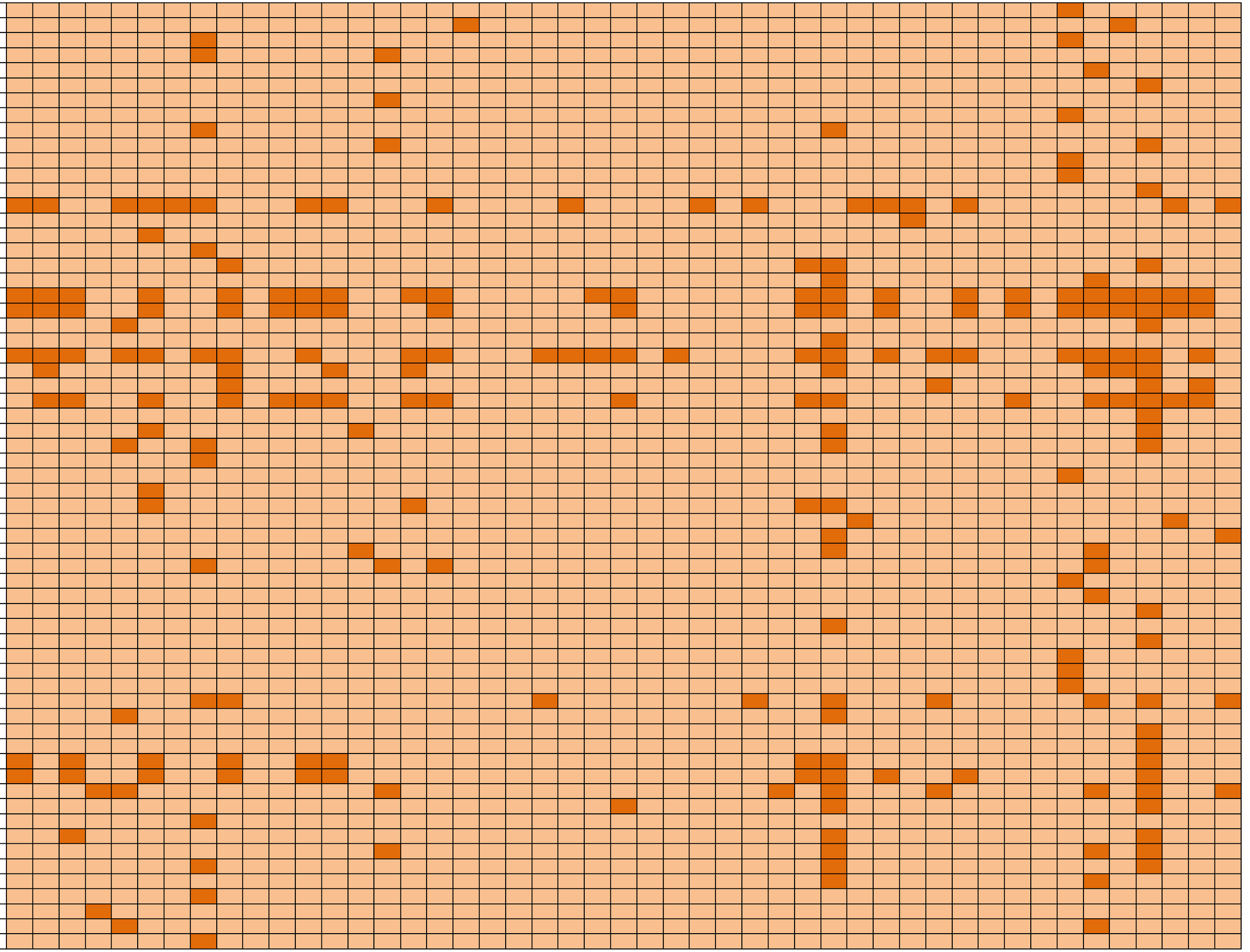
Accession	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Total
FLIP94-509C/97C016-20	9	2	3	2	1	4	2	0	23
ICCV96970/ICCV04516-9	3	2	2	3	1	2	2	0	15
ICCV 92809 / ICCV 04516	2	2	1	4	2	1	3	0	15
8627P-2 / ICC13729	10	1	1	3	2	3	0	1	21
97020-1147/ICC8294	12	5	0	5	0	4	2	1	29
ICCV 98503 / Moti	13	1	1	13	1	1	2	0	32
Rupali / 97037-1178-1001	8	6	4	17	1	2	1	0	39
ICCV 96836 / PBG 5	8	5	1	15	1	1	2	1	34
FLIP94-079C/01-481*03HS003	8	4	1	3	2	1	2	2	23
GENESIS510/KYABRA	12	4	1	2	2	2	2	2	27
ICCV 3996 / ICCV 93954	4	2	1	1	0	1	1	0	10
ICCV 96836 / ICC 12004	3	6	1	1	1	2	1	0	15
Sonali / ICC 12004	10	2	2	4	0	2	3	2	25
ICC1882	8	3	0	18	2	4	4	2	41
ICC283	10	6	2	19	1	5	5	3	51
ICC4958	3	2	2	18	1	1	2	0	29
ICC8261	1	2	3	20	0	1	1	1	29
PI489777	33	44	27	29	29	36	21	6	225
ICCV 88202	5	3	0	15	0	0	3	0	26
BL	3	1	2	19	1	0	1	0	27
ICCV 88516	7	2	0	15	0	0	2	0	26
YeZIN	5	5	3	20	1	2	1	0	37
ILC 3279	1	0	0	0	0	0	0	0	1
ICCV 92311	0	2	2	14	0	2	1	0	21
Myles	1	1	0	13	4	1	4	2	26
JG62	5	4	1	13	2	3	1	0	29
CDC 512-51	9	6	14	7	4	3	9	2	54
Amit	8	2	2	0	1	0	2	0	15
CDC Leader	4	1	3	9	2	2	6	0	27
CDC Corinne	13	9	5	13	7	3	8	2	60
ICCV 93954	13	14	5	30	2	2	11	2	79

CDC Luna	33	24	27	35	28	26	26	11	210
CDC Vanguard	14	7	2	19	0	1	8	0	51
ICCV 96836	35	19	15	35	14	17	21	4	160
Pistol	30	24	8	20	14	14	14	3	127
Hat Trick	10	7	11	14	10	9	8	4	73
Slasher	47	27	15	30	9	20	28	5	181
Genesis 090	3	7	4	12	0	3	3	1	33
Chefe	4	1	0	12	1	0	1	1	20
WR 315	3	2	1	23	3	2	2	1	37
ICCC 32	1	1	4	25	1	0	1	1	34
ICCV 92337	8	9	0	16	1	5	5	0	44
ICCV 93512	1	3	0	18	2	1	2	1	28
ICCV 97105	2	1	3	26	2	1	2	1	38
ICCL 83105	4	4	2	18	1	3	3	2	37
ICCV 93952	7	3	3	6	2	0	7	1	29
ICCV 92944	3	7	1	19	2	3	5	0	40

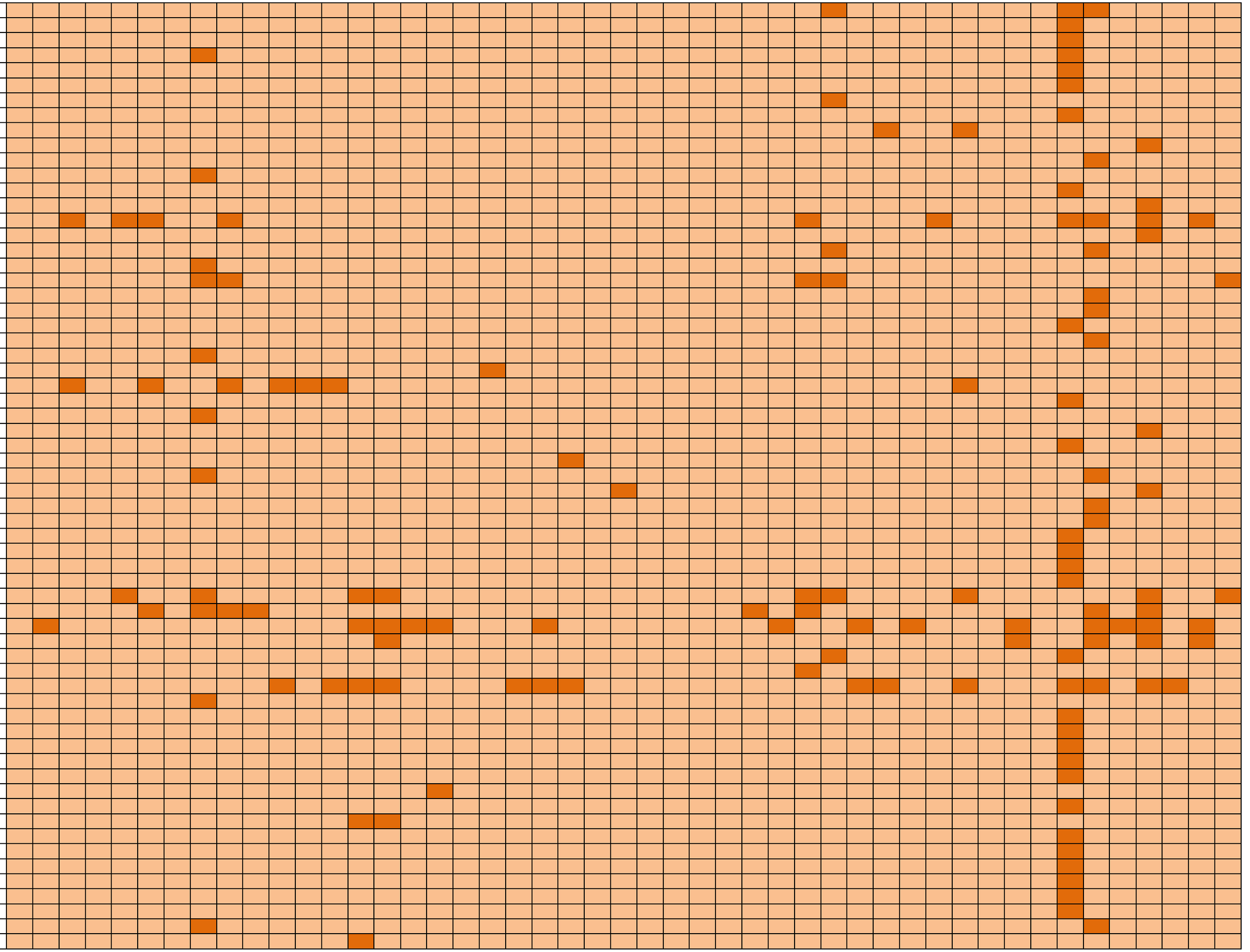
Appendix 7: Gene loss events in 47 chickpea varieties

Gene present Gene loss

Geneid	Gene start	Gene end	Gene length	No: of Exons	8627P-2--ICC13729	97020-1147--ICC8294	Amit	BL	CDC512-51	CDCCorinne	CDCLeader	CDCLuna	CDCVanguard	Chefe	FLIP94-079C--01-481*03HS003	FLIP94-509C--97C016-20	GENESIS510--KYABRA	Genesis090	HatTrick	ICC1882	ICC283	ICC4958	ICC8261	ICCC32	ICCL83105	ICCV3996--ICCV93954	ICCV88202	ICCV88516	ICCV92311	ICCV92337	ICCV92809--ICCV04516	ICCV92944	ICCV93512	ICCV93952	ICCV93954	ICCV96836	ICCV96836--ICC12004	ICCV96836--PBG5	ICCV96970--ICCV04516-9	ICCV97105	ICCV98503--Moti	ILC3279	JG62	Myles	PI489777	Pistol	Rupali--97037-1178-1001	Slasher	Sonali--ICC12004	WR315	YeZIN
Chromosome 1																																																		
Ca_00004	45719	46399	680	2																																														
Ca_00052	455636	455824	188	1																																														
Ca_00097	785086	787209	2123	3																																														
Ca_00101	831963	832705	742	2																																														
Ca_00162	1308237	1308608	371	1																																														
Ca_00208	1675656	1675943	287	1																																														
Ca_00209	1677366	1677653	287	1																																														
Ca_00283	2266640	2267107	467	1																																														
Ca_00317	2542242	2543014	772	2																																														
Ca_00326	2649601	2650122	521	1																																														
Ca_00339	2748677	2749243	566	1																																														
Ca_00345	2806713	2808359	1646	1																																														
Ca_00346	2817540	2819558	2018	2																																														
Ca_00355	2904878	2905240	362	1																																														
Ca_00387	3166368	3166667	299	1																																														
Ca_00511	4292038	4292199	161	1																																														
Ca_00513	4306017	4306538	521	1																																														
Ca_00521	4362062	4362370	308	1																																														
Ca_00526	4402683	4402865	182	1																																														
Ca_00538	4542693	4544758	2065	2																																														
Ca_00603	5129020	5130739	1719	4																																														
Ca_00613	5249811	5250446	635	2																																														
Ca_00622	5360348	5361118	770	1																																														
Ca_00625	5384708	5385453	745	2																																														
Ca_00631	5462501	5464601	2100	3																																														
Ca_02544	12105164	12106126	962	4																																														
Ca_02581	11692195	11692506	311	1																																														
Ca_02600	11501842	11502423	581	2																																														
Ca_02718	10530991	10531230	239	1																																														
Ca_02719	10529547	10530113	566	1																																														
Ca_02752	10289347	10290080	733	2																																														
Ca_02886	9099706	9100324	618	3																																														
Ca_06870	16985686	16985913	227	1																																														
Ca_06932	16273872	16275015	1143	2																																														
Ca_06939	16137809	16138144	335	1																																														
Ca_07005	15286584	15286949	365	1																																														
Ca_07014	15206045	15206257	212	1																																														
Ca_07050	14801284	14801733	449	1																																														
Ca_07898	6361883	6362568	685	2																																														
Ca_07966	7061995	7062504	509	2																																														
Ca_07991	7370335	7370670	335	1																																														
Ca_08083	8480091	8480474	383	1																																														
Ca_08840	20031603	20031869	266	1																																														
Ca_08852	19826052	19826558	506	1																																														

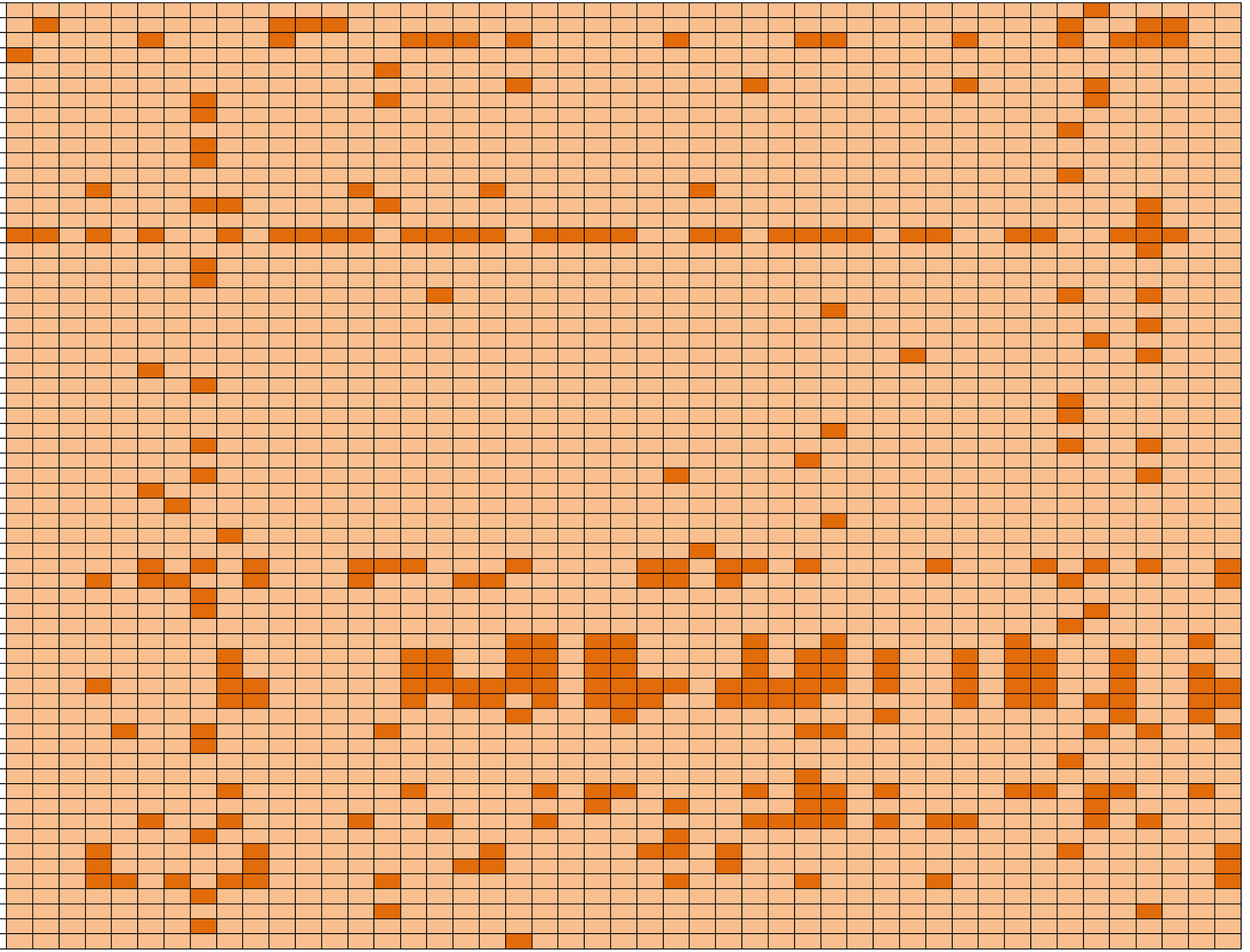
Ca_08853	19817520	19817822	302	1	
Ca_08857	19791244	19792226	982	2	
Ca_12920	47332133	47332573	440	1	
Ca_12927	47258522	47258779	257	1	
Ca_12944	46952359	46952541	182	1	
Ca_13534	43997550	43997768	218	1	
Ca_13545	43824166	43824525	359	1	
Ca_13555	43712100	43712836	736	2	
Ca_13581	43233263	43233466	203	1	
Ca_13584	43139016	43139207	191	1	
Ca_13589	43034211	43034390	179	1	
Ca_13603	42698780	42700459	1679	6	
Ca_13614	42551474	42552013	539	1	
Ca_13897	40212830	40213207	377	1	
Ca_13898	40210948	40211256	308	1	
Ca_13903	40062316	40064339	2023	2	
Ca_13911	39789691	39789933	242	1	
Ca_13925	39578121	39578300	179	1	
Ca_13932	39413032	39413289	257	1	
Ca_13934	39361528	39362702	1174	2	
Ca_13935	39353260	39353544	284	1	
Ca_13937	39275837	39276010	173	1	
Ca_13939	39220786	39222606	1820	1	
Ca_13945	39092588	39092779	191	1	
Ca_13947	38990169	38991238	1069	2	
Ca_13950	38859796	38860200	404	1	
Ca_13951	38838574	38839035	461	1	
Ca_13965	38361334	38361525	191	1	
Ca_14067	13034551	13034802	251	1	
Ca_14109	13441742	13442011	269	1	
Ca_14752	17883227	17884054	827	1	
Ca_14767	17631035	17631600	565	2	
Ca_14768	17628759	17629170	411	2	
Ca_14769	17627660	17628407	747	2	
Ca_14770	17621639	17622215	576	2	
Ca_17382	18738046	18738333	287	1	
Ca_17384	18704090	18704287	197	1	
Ca_18314	25074563	25074901	338	1	
Ca_18332	24741896	24742147	251	1	
Ca_18476	41097890	41098606	716	3	
Ca_18512	28424221	28424691	470	1	
Ca_18576	25632815	25633962	1147	2	
Ca_18577	25641842	25642165	323	1	
Ca_18592	26072625	26073411	786	2	
Ca_18593	26081028	26081570	542	1	
Ca_18594	26093183	26094166	983	1	
Ca_18601	26309122	26309358	236	1	
Ca_19285	24193925	24194359	434	1	
Ca_19296	23857314	23857913	599	1	
Ca_19306	23629039	23629263	224	1	
Ca_19316	23437045	23437470	425	2	
Ca_19317	23427872	23428316	444	2	
Ca_19451	33644599	33645093	494	1	
Ca_19461	33784397	33784858	461	1	
Ca_20640	22087432	22087620	188	1	
Ca_21403	35173617	35173937	320	1	
Ca_21420	34669958	34670455	497	1	
Ca_21423	34623795	34623989	194	1	
Ca_21552	46534408	46534653	245	2	
Ca_21677	36118811	36119374	563	1	
Ca_21688	36430152	36430628	476	2	
Ca_21693	36499526	36500713	1187	1	
Ca_21823	36956934	36957326	392	1	

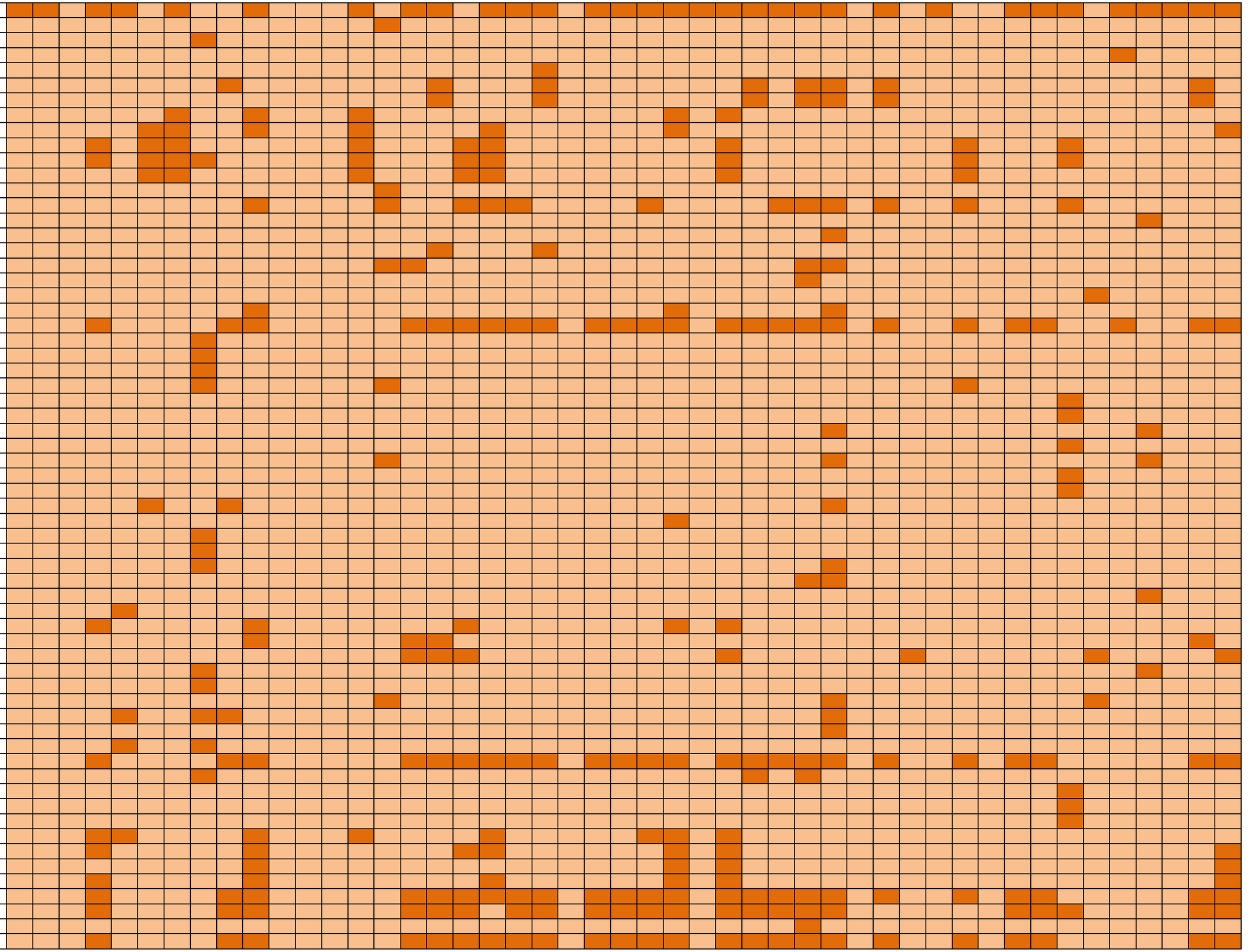
Ca_21832	37196486	37196923	437	1	
Ca_21845	31186694	31186918	224	1	
Ca_21847	31220339	31221158	819	2	
Ca_21850	31270079	31270365	286	2	
Ca_22100	22018370	22018702	332	1	
Ca_22348	22971891	22972190	299	1	
Ca_22353	23068754	23069245	491	1	
Ca_22463	30322188	30322463	275	1	
Ca_22464	30320191	30320466	275	1	
Ca_22531	41936880	41937169	289	2	
Ca_22534	41981044	41981649	605	1	
Ca_22574	29104969	29105154	185	1	
Ca_23116	41425352	41425510	158	1	
Ca_23117	41433596	41434006	410	1	
Ca_23491	44180545	44180802	257	1	
Ca_23495	44269155	44269634	479	1	
Ca_24108	35595323	35595652	329	1	
Ca_24431	45337248	45337565	317	1	
Ca_25720	45129427	45129945	518	1	
Ca_26024	45296535	45296969	434	1	
Ca_26102	35398517	35398924	407	1	
Ca_26403	31861134	31861661	527	1	
Ca_27004	31688303	31688638	335	1	
Ca_27006	31711655	31711951	296	1	
Ca_27320	34484335	34484505	170	1	
Ca_27634	34002761	34004719	1958	7	
Ca_27635	34033384	34033826	442	2	
Chromosome 2					
Ca_09721	35556775	35557359	584	1	
Ca_09723	35563905	35564348	443	1	
Ca_09744	35726987	35727285	298	2	
Ca_09745	35729836	35730723	887	3	
Ca_09748	35754212	35754652	440	1	
Ca_09760	35844818	35845680	862	3	
Ca_09770	35959011	35959460	449	1	
Ca_10195	33326287	33326592	305	1	
Ca_10274	32439778	32441846	2068	2	
Ca_10282	32372530	32372940	410	1	
Ca_10517	3365868	3366137	269	1	
Ca_10552	2921891	2922519	628	2	
Ca_11621	14135858	14136250	392	1	
Ca_11650	13573181	13574728	1547	3	
Ca_11657	13523925	13524785	860	2	
Ca_11673	13363046	13364455	1409	2	
Ca_11674	13362462	13362897	435	2	
Ca_11679	13298865	13299134	269	1	
Ca_12424	31376828	31377830	1002	3	
Ca_12426	31356362	31357440	1078	2	
Ca_12442	31213230	31213883	653	1	
Ca_12455	31083508	31083897	389	2	
Ca_12477	30815811	30816173	362	1	
Ca_12496	30541822	30542406	584	1	
Ca_12531	2490150	2490605	455	1	
Ca_12548	2296403	2297272	869	2	
Ca_12549	2295027	2295904	877	2	
Ca_14251	25185971	25187793	1822	4	
Ca_14310	26035534	26035971	437	1	
Ca_14317	29239474	29239692	218	1	
Ca_14322	29166095	29166534	439	3	
Ca_14348	28899881	28900207	326	1	
Ca_14349	28878855	28879181	326	1	
Ca_14350	28871938	28872264	326	1	
Ca_14354	28809270	28809747	477	2	

Ca_14355	28807945	28808422	477	2	
Ca_14356	28756626	28756964	338	1	
Ca_14357	28746997	28747329	332	1	
Ca_14359	28724104	28724364	260	1	
Ca_14362	28620965	28621525	560	1	
Ca_14363	28584494	28584868	374	1	
Ca_14367	28513344	28513544	200	1	
Ca_14666	5406540	5406983	443	2	
Ca_14668	5388858	5394555	5697	7	
Ca_14694	5175294	5175794	500	2	
Ca_14939	23907011	23907394	383	1	
Ca_14973	23147873	23148166	293	1	
Ca_15279	34529978	34530670	692	1	
Ca_15281	34551351	34552196	845	1	
Ca_15293	34651699	34652091	392	1	
Ca_15662	27448840	27448998	158	1	
Ca_15665	27578412	27578879	467	1	
Ca_16007	17402067	17404092	2025	2	
Ca_16011	17242480	17242725	245	1	
Ca_16079	30130380	30133959	3579	9	
Ca_16080	30100100	30101196	1096	2	
Ca_16104	29893471	29893716	245	1	
Ca_16105	29874190	29875489	1299	2	
Ca_16116	29742669	29742863	194	1	
Ca_16139	29454023	29455062	1039	2	
Ca_16886	34823740	34824592	852	2	
Ca_17246	26451094	26451919	825	2	
Ca_17250	26510821	26514345	3524	2	
Ca_17258	26648616	26649449	833	1	
Ca_17564	10760821	10761186	365	1	
Ca_17817	32049395	32049631	236	1	
Ca_17824	31927905	31928537	632	1	
Ca_17833	31848007	31848504	497	1	
Ca_17840	31792794	31794311	1517	2	
Ca_17847	31702917	31704113	1196	1	
Ca_18081	12581745	12582071	326	1	
Ca_18083	12610634	12610864	230	1	
Ca_18091	12778978	12779771	793	2	
Ca_18092	12781058	12781997	939	2	
Ca_18102	12979967	12980464	497	1	
Ca_18114	7549317	7549496	179	1	
Ca_18138	7897134	7898672	1538	3	
Ca_18143	7930181	7930575	394	2	
Ca_18754	12117811	12118422	611	1	
Ca_18755	12116739	12116969	230	1	
Ca_18758	12086428	12086712	284	1	
Ca_18765	11893567	11893923	356	1	
Ca_18783	11525916	11526310	394	2	
Ca_19149	6831106	6831273	167	1	
Ca_19160	6718329	6718508	179	1	
Ca_19161	6709567	6709746	179	1	
Ca_19702	5987406	5989227	1821	2	
Ca_19703	5996645	5996998	353	1	
Ca_20034	9282842	9283150	308	1	
Ca_20394	9919707	9920069	362	1	
Ca_21095	4767089	4767805	716	1	
Ca_21653	24463951	24464781	830	2	
Ca_21655	24442764	24443033	269	1	
Ca_21656	24423541	24423831	290	1	
Ca_21657	24375110	24376666	1556	5	
Ca_21658	24373841	24374561	720	3	
Ca_22900	24600553	24600768	215	1	
Ca_22904	24554630	24555100	470	1	

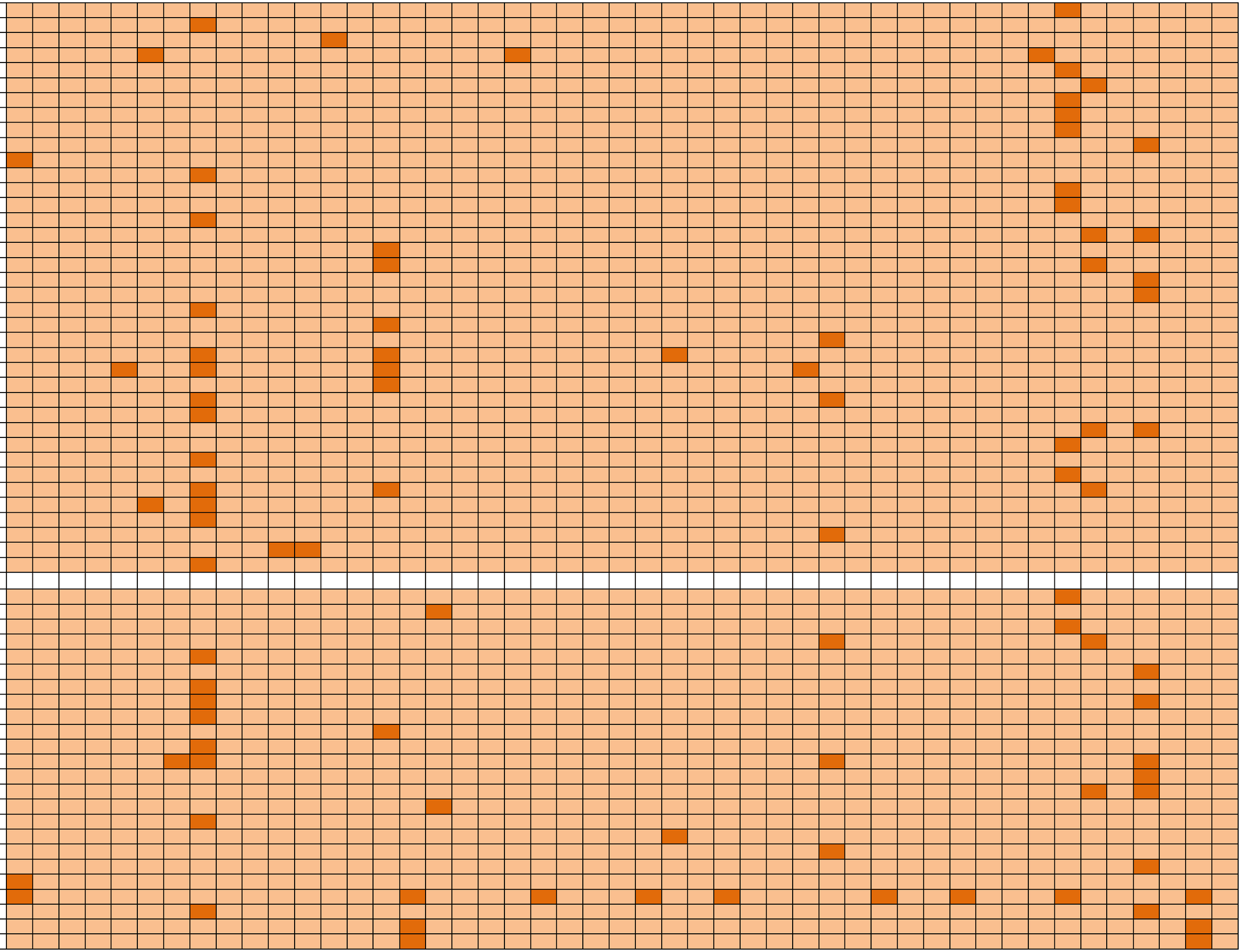
Ca_23048	21620722	21621060	338	1	
Ca_23057	21783587	21783862	275	1	
Ca_23768	20934495	20934650	155	1	
Ca_24540	20008455	20008637	182	1	
Ca_24623	21222370	21222804	434	1	
Ca_24624	21238167	21238367	200	1	
Ca_24633	21554091	21554627	536	1	
Ca_24646	15914672	15914863	191	1	
Ca_25092	9670579	9670758	179	1	
Ca_25180	29391257	29391616	359	1	
Ca_25219	19136531	19136821	290	1	
Ca_25225	19347506	19347790	284	1	
Ca_26872	7200733	7201221	488	1	
Ca_28004	4919082	4919822	740	1	
Chromosome 3					
Ca_00717	34403381	34403791	410	1	
Ca_00718	34408395	34409038	643	2	
Ca_00733	34612262	34612948	686	2	
Ca_00775	35010884	35013636	2752	7	
Ca_00784	35112749	35113243	494	1	
Ca_00833	35508163	35508471	308	1	
Ca_00857	35689583	35689843	260	1	
Ca_00858	35689954	35690721	767	1	
Ca_00879	35881177	35881557	380	1	
Ca_00929	36228096	36229851	1755	4	
Ca_00989	36678951	36679361	410	1	
Ca_01068	37349723	37349878	155	1	
Ca_01073	37416630	37416899	269	2	
Ca_01074	37422395	37422647	252	2	
Ca_01135	37965518	37969220	3702	12	
Ca_01194	38421456	38422030	574	2	
Ca_01284	39244446	39245405	959	3	
Ca_01325	39546460	39546972	512	1	
Ca_06048	24729500	24729691	191	1	
Ca_06068	24481047	24481501	454	2	
Ca_06078	24394415	24394633	218	1	
Ca_06125	23960897	23961211	314	1	
Ca_06136	23882060	23883467	1407	4	
Ca_06247	22642440	22642877	437	1	
Ca_07144	28288013	28288921	908	1	
Ca_07182	28759992	28760219	227	1	
Ca_07226	29197076	29199700	2624	2	
Ca_07227	29206781	29207212	431	1	
Ca_07368	30480214	30480393	179	1	
Ca_08226	26818524	26819030	506	1	
Ca_08227	26803718	26804353	635	1	
Ca_09397	19950586	19951356	770	2	
Ca_09453	20703959	20704147	188	1	
Ca_09454	20749730	20749987	257	1	
Ca_09466	20958403	20958867	464	2	
Ca_11957	32342931	32343236	305	1	
Ca_11976	32497175	32497594	419	1	
Ca_12011	32771616	32771927	311	1	
Ca_12205	31063703	31064134	431	1	
Ca_12206	31091019	31092051	1032	2	
Ca_16526	7428286	7428489	203	1	
Ca_17037	15567017	15567277	260	1	
Ca_17044	15670925	15675622	4697	17	
Ca_18252	16218361	16218585	224	1	
Ca_18258	16353665	16354207	542	1	
Ca_18260	16382686	16382988	302	1	
Ca_18289	16916353	16917231	878	2	
Ca_19343	19454041	19454484	443	1	

Ca_19363	10947744	10948361	617	1	
Ca_19365	11109691	11110023	332	1	
Ca_19366	11130990	11131202	212	1	
Ca_19367	11132249	11132479	230	1	
Ca_19375	11371672	11372266	594	2	
Ca_19394	2276927	2277160	233	1	
Ca_19424	1201736	1202113	377	1	
Ca_19645	18778000	18778230	230	1	
Ca_20651	5623027	5623386	359	1	
Ca_20652	5602059	5602271	212	1	
Ca_20963	9321853	9322014	161	1	
Ca_20964	9333037	9333279	242	1	
Ca_20967	9383619	9384118	499	3	
Ca_20974	9503572	9504039	467	2	
Ca_21016	12736531	12736830	299	1	
Ca_21018	12645505	12645896	391	2	
Ca_21022	12558578	12558793	215	1	
Ca_21023	12505848	12506456	608	1	
Ca_21128	17914561	17914749	188	1	
Ca_21336	3778617	3778841	224	1	
Ca_21337	3744685	3744960	275	1	
Ca_22077	4497960	4498502	542	1	
Ca_22167	3117711	3117920	209	1	
Ca_22249	11978109	11978789	680	3	
Ca_22256	11841864	11842106	242	1	
Ca_22493	15050452	15050607	155	1	
Ca_22508	14556395	14557012	617	1	
Ca_23293	13987743	13988042	299	1	
Ca_23342	17336226	17336504	278	1	
Ca_23725	8811129	8811317	188	1	
Ca_23870	5200365	5200568	203	1	
Ca_23874	5143078	5143383	305	1	
Ca_23877	5081901	5082179	278	1	
Ca_23879	5056654	5057175	521	2	
Ca_24548	3497505	3497720	215	1	
Ca_24549	3463826	3464104	278	1	
Ca_24559	3217777	3218661	884	1	
Ca_24826	6167394	6167585	191	1	
Ca_25284	9027217	9027447	230	1	
Ca_25320	13629102	13629371	269	1	
Ca_26371	2386776	2386991	215	1	
Ca_26374	2473286	2473516	230	1	
Ca_26376	2502956	2503345	389	1	
Ca_26570	8404761	8405138	377	1	
Ca_26572	8448905	8449192	287	1	
Chromosome 4					
Ca_03452	7843353	7844752	1399	2	
Ca_03619	6250675	6251265	590	1	
Ca_03634	6112656	6113012	356	1	
Ca_03772	4860719	4861860	1141	2	
Ca_03804	4543392	4543694	302	1	
Ca_03864	3886923	3887255	332	1	
Ca_03865	3882989	3883363	374	2	
Ca_03866	3878262	3879440	1178	1	
Ca_04324	10910117	10910764	647	1	
Ca_04354	11223694	11224347	653	1	
Ca_04470	12476815	12476997	182	1	
Ca_04485	12617727	12618350	623	1	
Ca_04548	13252578	13253678	1100	1	
Ca_04583	13645639	13646844	1205	2	
Ca_04593	13711171	13711644	473	1	
Ca_04597	13733160	13734765	1605	2	
Ca_04637	14099242	14099856	614	1	

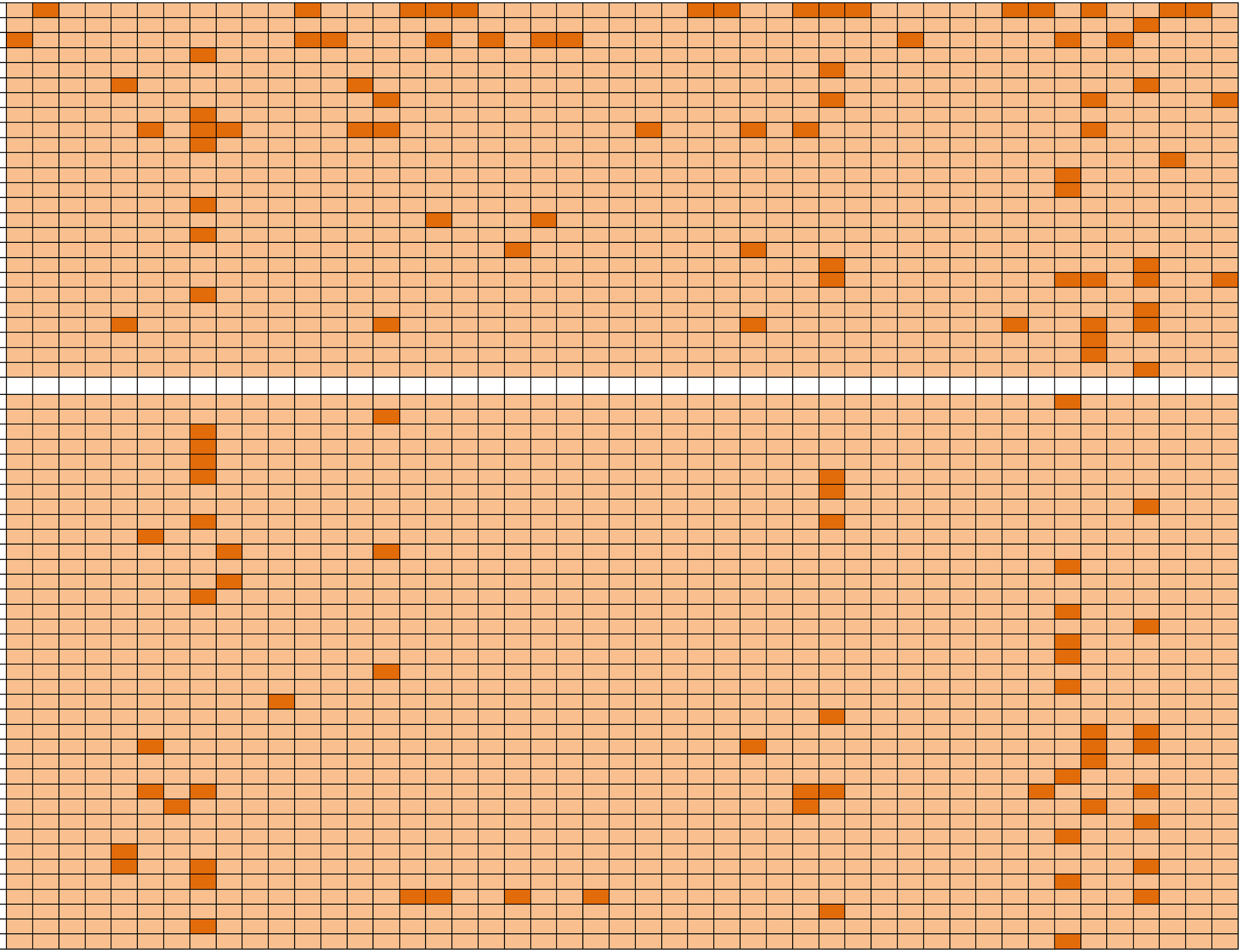
Ca_04644	14140900	14142225	1325	3	
Ca_04650	14179971	14180792	821	1	
Ca_05385	17413769	17415327	1558	2	
Ca_05412	17116534	17127194	10660	6	
Ca_05428	16961052	16961843	791	1	
Ca_05508	16004115	16005859	1744	2	
Ca_05525	15806447	15806749	302	1	
Ca_05548	15493224	15493718	494	1	
Ca_05568	15308474	15308632	158	1	
Ca_05570	15288852	15289235	383	1	
Ca_05587	15133469	15133864	395	1	
Ca_05620	14823751	14824308	557	1	
Ca_07647	77019	77387	368	1	
Ca_07824	2041269	2041547	278	1	
Ca_08336	8451347	8451631	284	1	
Ca_08362	8673418	8675706	2288	2	
Ca_08380	8825639	8826141	502	2	
Ca_08456	9787833	9788723	890	1	
Ca_09064	43840805	43841422	617	1	
Ca_09123	44518390	44520552	2162	3	
Ca_09166	45033444	45034499	1055	2	
Ca_09191	45274952	45275230	278	1	
Ca_09192	45292718	45292906	188	1	
Ca_10753	48329992	48331106	1114	4	
Ca_10756	48313227	48314009	782	1	
Ca_10763	48246915	48248708	1793	3	
Ca_10851	47302576	47302770	194	1	
Ca_10863	47181320	47181628	308	1	
Ca_10896	42569731	42569934	203	1	
Ca_10958	41472705	41472953	248	1	
Ca_10985	41243013	41243572	559	2	
Ca_11010	40995269	40995508	239	1	
Ca_11011	40984692	40984988	296	1	
Ca_12135	3329602	3330069	467	1	
Ca_13051	39403291	39404805	1514	2	
Ca_13052	39389951	39394499	4548	2	
Ca_13089	38864426	38864767	341	1	
Ca_13090	38839044	38839400	356	1	
Ca_13127	38295199	38295411	212	1	
Ca_13995	18956991	18957693	702	2	
Ca_14050	19617999	19618229	230	1	
Ca_14054	19679678	19680342	664	2	
Ca_14168	29727470	29727721	251	1	
Ca_14178	29953433	29953964	531	2	
Ca_14179	29955360	29955542	182	1	
Ca_14180	30023910	30024247	337	2	
Ca_14200	30470940	30471131	191	1	
Ca_14217	31029436	31029648	212	1	
Ca_14234	31415397	31415708	311	1	
Ca_14417	23834911	23835657	746	1	
Ca_14422	23774312	23774761	449	1	
Ca_14464	22874123	22874332	209	1	
Ca_14473	22778929	22779384	455	1	
Ca_14790	37112859	37113274	415	2	
Ca_14794	37022819	37023034	215	1	
Ca_14798	36926307	36926537	230	1	
Ca_14809	36741708	36742169	461	1	
Ca_14820	36578620	36581125	2505	5	
Ca_14823	36556704	36557060	356	1	
Ca_14844	36271939	36272196	257	1	
Ca_14860	35933820	35934359	539	1	
Ca_14862	35899569	35899760	191	1	
Ca_14869	40566999	40569102	2103	7	

Ca_14871	40544646	40544834	188	1	
Ca_14891	40338227	40338688	461	1	
Ca_14912	39985011	39985328	317	1	
Ca_14915	39954258	39957667	3409	3	
Ca_14932	39693458	39694969	1511	6	
Ca_14933	39666505	39667476	971	1	
Ca_14934	39663118	39664140	1022	2	
Ca_15103	37765200	37765622	422	1	
Ca_15115	37652563	37654116	1553	2	
Ca_15122	37600338	37600526	188	1	
Ca_15123	37599202	37599633	431	1	
Ca_15124	37590074	37592294	2220	2	
Ca_15434	31476958	31477272	314	1	
Ca_15436	31542132	31542557	425	1	
Ca_15458	31994691	31995410	719	1	
Ca_15461	32068837	32069301	464	1	
Ca_15462	32077738	32113952	36214	5	
Ca_15477	32448225	32448434	209	1	
Ca_15481	32527577	32527801	224	1	
Ca_16568	25263395	25263598	203	1	
Ca_16570	25299768	25300115	347	1	
Ca_16571	25321094	25321954	860	1	
Ca_16584	25558053	25558334	281	1	
Ca_16589	25633232	25633588	356	1	
Ca_16592	25681361	25681663	302	1	
Ca_16617	26249817	26250176	359	1	
Ca_17120	19829357	19829806	449	2	
Ca_17122	19860402	19860575	173	1	
Ca_18379	45817247	45817564	317	1	
Ca_18390	45718234	45719151	917	1	
Ca_18662	18296899	18297090	191	1	
Ca_19255	46625897	46626154	257	1	
Ca_19279	46334366	46335276	910	2	
Ca_19991	34792478	34792807	329	1	
Ca_20012	35181388	35182374	986	1	
Ca_20022	35416159	35416851	692	1	
Ca_20144	21947443	21947826	383	1	
Ca_20346	34661191	34661466	275	1	
Ca_20347	34533558	34533731	173	1	
Ca_20357	34216825	34217049	224	1	
Ca_20358	34209076	34209807	731	2	
Ca_20371	33818468	33820750	2282	3	
Ca_20435	26720613	26724115	3502	4	
Ca_20446	26924212	26924550	338	1	
Ca_20447	26926943	26927269	326	1	
Ca_20454	27048752	27049135	383	1	
Ca_20860	24563635	24563931	296	1	
Ca_21306	27716728	27716913	185	1	
Ca_21311	27628788	27629042	254	1	
Ca_21315	27529697	27530233	536	1	
Ca_21324	27371524	27372135	611	1	
Ca_21609	18587427	18587705	278	1	
Ca_22431	35467413	35469478	2065	3	
Ca_22432	35483613	35484749	1136	2	
Ca_22433	35487017	35489174	2157	5	
Ca_22436	35565799	35566029	230	1	
Ca_22437	35567441	35567644	203	1	
Ca_22439	35590420	35590917	497	1	
Ca_22450	35757535	35757951	416	1	
Ca_23168	28233722	28234219	497	2	
Ca_23169	28222818	28223327	509	1	
Ca_23172	28164613	28165536	923	2	
Ca_23179	28039250	28039669	419	1	

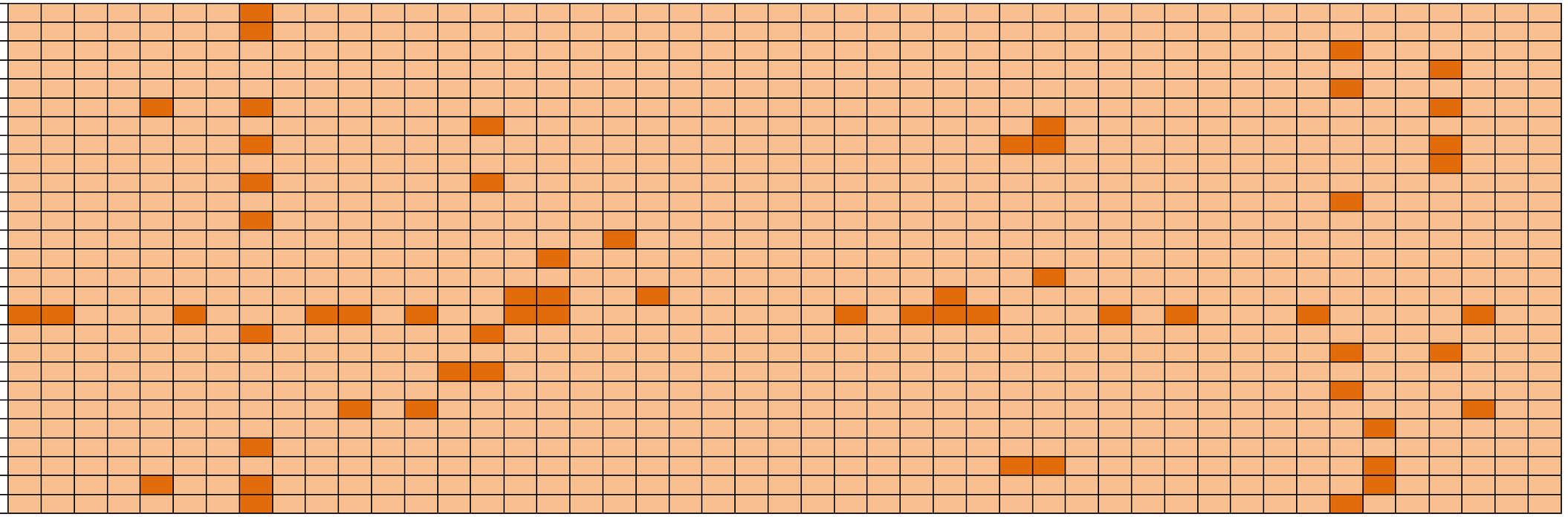
Ca_23679	43237360	43237812	452	1	
Ca_23699	28840387	28841136	749	1	
Ca_23709	29137645	29138154	509	1	
Ca_23710	29146748	29147392	644	1	
Chromosome 5					
Ca_01410	38085102	38085560	458	1	
Ca_01448	37739352	37739837	485	1	
Ca_01513	37027359	37027664	305	1	
Ca_01516	37005769	37006044	275	1	
Ca_01635	35903693	35904235	542	1	
Ca_01639	35868652	35868999	347	1	
Ca_01736	34988539	34988817	278	1	
Ca_01805	34403121	34403597	476	1	
Ca_01840	34075765	34075974	209	1	
Ca_01842	34066949	34067257	308	1	
Ca_01894	33626144	33627070	926	1	
Ca_03928	44847705	44847950	245	1	
Ca_03942	44955319	44955483	164	1	
Ca_03988	45433897	45435003	1106	1	
Ca_04023	45735705	45735881	176	1	
Ca_04055	46060966	46062552	1586	5	
Ca_04138	46783200	46783766	566	1	
Ca_04682	30053366	30053611	245	1	
Ca_04731	30487200	30487478	278	1	
Ca_04741	30531521	30531853	332	1	
Ca_04827	31213397	31217778	4381	14	
Ca_04828	31224345	31225730	1385	4	
Ca_04879	31753322	31754094	772	2	
Ca_04888	31823979	31824191	212	1	
Ca_05003	33064119	33064289	170	1	
Ca_05005	33086516	33087121	605	2	
Ca_07606	40851987	40853474	1487	5	
Ca_07616	40946648	40946926	278	1	
Ca_08911	27300968	27301774	806	1	
Ca_08968	26518174	26519421	1247	2	
Ca_09035	25526859	25527020	161	1	
Ca_09037	25496241	25496633	392	1	
Ca_12637	44190357	44191430	1073	2	
Ca_13366	28612099	28612434	335	1	
Ca_13382	28412670	28413259	589	2	
Ca_15569	24269461	24269643	182	1	
Ca_15591	24565148	24565864	716	1	
Ca_16621	29266949	29267591	642	2	
Ca_16672	29922647	29923691	1044	2	
Ca_17075	12551519	12551866	347	1	
Ca_17109	11917286	11917609	323	1	
Ca_17674	23743874	23744244	370	2	
Ca_17687	23626190	23633608	7418	14	
Ca_17913	16182485	16183303	818	2	
Ca_17936	16719330	16719683	353	1	
Ca_17941	16763175	16763996	821	1	
Ca_17944	16830942	16832486	1544	2	
Ca_17946	16849554	16849802	248	1	
Ca_18159	50651	54391	3740	1	
Ca_18193	376819	377076	257	1	
Ca_18210	20482631	20482810	179	1	
Ca_18239	21271812	21272072	260	1	
Ca_18687	3063254	3063739	485	1	
Ca_18706	8330334	8332409	2075	3	
Ca_18718	8006691	8007086	395	1	
Ca_18735	7654891	7655298	407	1	
Ca_18737	7645814	7646116	302	1	
Ca_19205	41768294	41769127	833	4	

Ca_19961	9068424	9070229	1805	2	
Ca_19978	9363904	9364278	374	1	
Ca_19979	9365693	9366085	392	1	
Ca_19983	9378066	9378239	173	1	
Ca_20075	22665019	22665246	227	1	
Ca_20076	22656902	22657081	179	1	
Ca_20340	8893317	8893934	617	1	
Ca_20499	9797941	9798162	221	1	
Ca_20500	9799386	9799757	371	1	
Ca_20730	5883343	5883831	488	1	
Ca_20832	14254398	14257030	2632	4	
Ca_21222	10649460	10649825	365	1	
Ca_21224	10607363	10608847	1484	2	
Ca_21566	7196991	7197230	239	1	
Ca_22474	17756631	17757140	509	1	
Ca_22480	17567690	17567908	218	1	
Ca_22481	17558490	17558708	218	1	
Ca_22482	17544860	17545135	275	1	
Ca_22487	17213407	17213826	419	1	
Ca_22620	5204628	5205887	1259	1	
Ca_22635	5687631	5687867	236	1	
Ca_22852	18088303	18088665	362	1	
Ca_22854	18123548	18123889	341	1	
Ca_22855	18126997	18127314	317	1	
Ca_23222	618926	619384	458	1	
Ca_23224	692055	693206	1151	2	
Ca_23225	700353	700670	317	1	
Ca_23236	1166233	1166730	497	1	
Ca_23756	10780431	10780604	173	1	
Ca_23788	9608780	9614892	6112	15	
Ca_24462	21801186	21801665	479	1	
Ca_24463	21806212	21807289	1077	2	
Ca_24486	18745074	18745307	233	1	
Ca_24491	18957920	18958177	257	1	
Ca_26197	4152732	4153049	317	1	
Ca_26278	1300123	1300896	773	1	
Ca_26281	1258935	1259448	513	2	
Ca_27587	22487751	22488140	389	1	
Chromosome 6					
Ca_05115	13431032	13431661	629	1	
Ca_05205	14348550	14351373	2823	5	
Ca_05209	14383923	14385088	1165	3	
Ca_05262	14885226	14886134	908	1	
Ca_05276	15005977	15006216	239	1	
Ca_05344	15662691	15663047	356	2	
Ca_05713	5787371	5787799	428	1	
Ca_05771	5281697	5281954	257	1	
Ca_05925	3783770	3784835	1065	2	
Ca_06353	17388057	17388710	653	1	
Ca_06424	18058474	18058863	389	1	
Ca_06428	18096314	18096466	152	1	
Ca_06433	18139113	18140033	920	1	
Ca_06469	18585266	18585565	299	1	
Ca_06493	18836478	18837512	1034	4	
Ca_08565	10248574	10248969	395	1	
Ca_08572	10180965	10181613	648	2	
Ca_08595	9953041	9953598	557	1	
Ca_08689	9077171	9078731	1560	3	
Ca_09598	7792847	7793062	215	1	
Ca_09617	7653829	7655158	1329	4	
Ca_10354	2258182	2258619	437	1	
Ca_10423	2862654	2865633	2979	5	
Ca_10424	2867734	2867940	206	1	

Ca_10428	2904769	2905917	1148	2
Ca_10442	3010495	3010767	272	1
Ca_11046	24127127	24127429	302	1
Ca_11099	23551590	23551914	324	2
Ca_11145	23025163	23025345	182	1
Ca_11149	22992718	22993023	305	1
Ca_13436	50752208	50752480	272	1
Ca_13477	50158098	50158663	565	3
Ca_13493	49755206	49755457	251	1
Ca_13526	49352964	49354318	1354	2
Ca_13660	57810771	57811106	335	1
Ca_13667	57737137	57737397	260	1
Ca_13669	57724934	57725383	449	1
Ca_13707	57340370	57340792	422	1
Ca_13838	45413290	45413601	311	1
Ca_13842	45514784	45515197	413	1
Ca_13866	45940794	45941487	693	2
Ca_13870	46005226	46005729	503	2
Ca_13883	46207910	46208128	218	1
Ca_13889	46302677	46304182	1505	3
Ca_14580	27811927	27812649	722	1
Ca_14588	27710409	27710732	323	1
Ca_14628	27185060	27185521	461	1
Ca_15171	31400790	31401750	960	2
Ca_15188	31618582	31618815	233	1
Ca_15371	58272313	58273060	747	2
Ca_15373	58311090	58311323	233	1
Ca_15375	58342187	58342648	461	1
Ca_15782	35320334	35322221	1887	4
Ca_15801	34914097	34914297	200	1
Ca_15929	39875363	39875536	173	1
Ca_16269	54150388	54150954	566	1
Ca_16445	30348906	30349172	266	1
Ca_16722	26909087	26909455	368	1
Ca_16734	9053846	9054292	446	1
Ca_16902	6314311	6315127	816	4
Ca_16931	6724333	6726363	2030	6
Ca_16955	6921559	6921747	188	1
Ca_17406	56687993	56688166	173	1
Ca_17425	56894964	56895707	743	1
Ca_17443	29495778	29496014	236	1
Ca_17469	28990387	28991953	1566	7
Ca_17470	28962938	28964584	1646	3
Ca_17495	28539074	28542821	3747	2
Ca_17538	52480820	52481202	382	2
Ca_17721	47144574	47144907	333	2
Ca_17744	46559013	46559198	185	1
Ca_17745	46543625	46544125	500	1
Ca_18835	59456742	59456924	182	1
Ca_18836	59456231	59456464	233	1
Ca_18842	59401205	59403031	1826	3
Ca_18850	59244984	59246904	1920	3
Ca_18854	59202619	59202804	185	1
Ca_18857	59172434	59174266	1832	1
Ca_18881	53388904	53389215	311	1
Ca_19031	56039397	56039690	293	1
Ca_19036	56077922	56078310	388	2
Ca_19183	51157483	51158079	596	1
Ca_19187	51085942	51086217	275	1
Ca_19188	51078450	51078692	242	1
Ca_19542	37833385	37833618	233	1
Ca_19551	37987932	37988249	317	1
Ca_19567	38334190	38334456	266	1

Ca_19821	36467844	36468476	632	3	
Ca_19874	21304103	21304387	284	1	
Ca_20714	49002147	49002371	224	1	
Ca_20721	49162782	49163459	677	1	
Ca_20727	49289767	49290165	398	1	
Ca_21201	54783030	54783566	536	2	
Ca_21578	41079357	41079740	383	1	
Ca_21588	40774226	40774843	617	1	
Ca_22059	33792285	33792554	269	1	
Ca_22067	33658312	33658905	593	1	
Ca_22068	33616577	33616804	227	1	
Ca_22926	53319539	53321045	1506	2	
Ca_23153	41408818	41409000	182	1	
Ca_23454	47675164	47675460	296	1	
Ca_24013	25881568	25882320	752	2	
Ca_24239	44460361	44460681	320	1	
Ca_24456	33346700	33347680	980	2	
Ca_24847	12233085	12233288	203	1	
Ca_24866	43769302	43769493	191	1	
Ca_25120	24723134	24723720	586	2	
Ca_25432	56323600	56324643	1043	1	
Ca_25505	26158794	26159069	275	1	
Ca_25837	32437033	32437287	254	1	
Ca_25855	51879074	51880356	1282	2	
Ca_26378	48273	48527	254	1	
Chromosome 7					
Ca_02951	4914653	4915445	792	2	
Ca_03007	4480792	4481040	248	1	
Ca_03023	4351284	4351775	491	1	
Ca_03213	2650196	2650759	563	1	
Ca_03246	2248461	2248772	311	1	
Ca_06636	7234321	7234689	368	1	
Ca_06716	6368673	6369112	439	2	
Ca_06747	6105778	6105969	191	1	
Ca_06751	6079807	6080217	410	1	
Ca_06791	5664239	5665690	1451	1	
Ca_06793	5653156	5653536	380	1	
Ca_06797	5627322	5627591	269	2	
Ca_06805	5577279	5577734	455	1	
Ca_09233	12874249	12874602	353	1	
Ca_09246	12738054	12738371	317	1	
Ca_09312	11855221	11855385	164	1	
Ca_09871	16075918	16076250	332	1	
Ca_09872	16075075	16075869	794	1	
Ca_09891	15811178	15811750	572	3	
Ca_09927	15253398	15253676	278	1	
Ca_10053	32207878	32208231	353	1	
Ca_10107	31534514	31534993	479	1	
Ca_10133	31246434	31246613	179	1	
Ca_10145	31109627	31109917	290	1	
Ca_12317	19074134	19074439	305	1	
Ca_12336	19365547	19365732	185	1	
Ca_12365	19982481	19982843	362	1	
Ca_12371	20031951	20032325	374	1	
Ca_12775	10137354	10138151	797	1	
Ca_12812	10548345	10548545	200	1	
Ca_12819	10636353	10637036	683	1	
Ca_13149	8535090	8535344	254	1	
Ca_13178	8326057	8326314	257	1	
Ca_13233	7827780	7827938	158	1	
Ca_13739	36423660	36424040	380	1	
Ca_13793	37112934	37113500	566	2	
Ca_14485	20930351	20930599	248	1	

Ca_14486	20938428	20942027	3599	1	
Ca_14487	20942285	20944621	2336	1	
Ca_14551	22358349	22360160	1811	2	
Ca_15295	9188283	9188821	538	2	
Ca_15296	9179554	9180092	538	2	
Ca_15297	9169141	9169475	334	2	
Ca_15363	8553015	8553269	254	1	
Ca_15716	45069190	45069582	392	1	
Ca_15726	45250992	45253992	3000	2	
Ca_15730	45305212	45305466	254	1	
Ca_15883	16960656	16960916	260	1	
Ca_16021	13190969	13191487	518	1	
Ca_16031	13386914	13387438	524	2	
Ca_16174	33545334	33545537	203	1	
Ca_16217	27633097	27633591	494	1	
Ca_16225	27344680	27345063	383	1	
Ca_16226	27236854	27237111	257	1	
Ca_16252	26679441	26679710	269	1	
Ca_16406	38300541	38300837	296	1	
Ca_16425	38676032	38676298	266	1	
Ca_16428	38777742	38777960	218	1	
Ca_17327	9326521	9327784	1263	4	
Ca_17331	9283151	9283915	764	2	
Ca_17645	35265559	35265966	407	1	
Ca_17991	41603123	41603380	257	1	
Ca_18919	28292110	28293198	1088	1	
Ca_18960	28715507	28718395	2888	8	
Ca_19483	47111303	47111764	461	1	
Ca_19485	47119328	47119789	461	1	
Ca_19486	47129235	47129426	191	1	
Ca_19487	47147571	47147804	233	1	
Ca_19539	22544731	22545090	359	1	
Ca_19585	39315374	39315664	290	1	
Ca_19589	39347069	39347481	412	2	
Ca_19600	39554788	39555333	545	1	
Ca_19603	39589041	39589463	422	1	
Ca_19604	39592972	39593468	496	2	
Ca_19605	39600834	39601109	275	1	
Ca_19920	23969102	23970048	946	4	
Ca_20189	43054829	43055215	386	2	
Ca_20190	43053696	43054013	317	1	
Ca_20275	40750153	40750746	593	1	
Ca_21356	43840642	43840890	248	1	
Ca_21362	44003743	44004021	278	1	
Ca_21363	44015565	44015861	296	1	
Ca_22123	18510318	18510682	364	2	
Ca_22124	18509104	18509451	347	1	
Ca_22130	18416044	18416345	301	2	
Ca_22136	18290960	18291190	230	1	
Ca_22142	18094328	18094633	305	1	
Ca_22540	47709452	47709622	170	1	
Ca_22542	47785743	47786060	317	1	
Ca_22751	46306825	46307094	269	1	
Ca_22790	37826035	37826217	182	1	
Ca_23316	30898205	30898534	329	1	
Ca_23317	30874338	30874721	383	1	
Ca_23597	25078798	25079082	284	1	
Ca_23598	25080903	25081238	335	1	
Ca_26607	28164480	28164866	386	1	
Chromosome 8					
Ca_01966	5850644	5851120	476	1	
Ca_02041	5114267	5114560	293	1	
Ca_02082	4755748	4756020	272	1	

Ca_02104	4561596	4562060	464	1	
Ca_02155	4131908	4132273	365	1	
Ca_02198	3778814	3779655	841	2	
Ca_02211	3658952	3659518	566	2	
Ca_02244	3371755	3374346	2591	6	
Ca_10613	7729646	7729861	215	1	
Ca_10692	6999680	7000327	647	1	
Ca_10723	6641371	6641706	335	1	
Ca_11538	9148677	9149009	332	2	
Ca_11914	253534	253872	338	2	
Ca_12962	15152766	15153023	257	1	
Ca_12970	15027990	15028829	839	1	
Ca_12973	14990323	14992169	1846	3	
Ca_13003	14650679	14651059	380	1	
Ca_13007	14600941	14602911	1970	2	
Ca_13026	14389743	14393719	3976	4	
Ca_13042	14192246	14194082	1836	3	
Ca_15028	1458599	1458877	278	1	
Ca_15501	15941863	15942093	230	1	
Ca_15508	16005725	16007003	1278	3	
Ca_15530	16158476	16159180	704	3	
Ca_15540	16254182	16254634	452	2	
Ca_16836	11218231	11218650	419	1	
Ca_19722	10656927	10657226	299	2	
Ca_20110	13326807	13327067	260	1	
Ca_23583	6346774	6346938	164	1	
Ca_23584	6350350	6350553	203	1	

Appendix 8: Novel gene variability in all 47 chickpea cultivars

Accession	Genes lost
DICC8220	28
ICC283	31
ICCV92311	31
ICCV88202	32
ILC3279	32
Myles	32
WACPE2160	33
05H072D-08HS3001	33
ICCV93952	33
ICC1882	34
JG62	34
04C830-07WAD2	35
ICC8261	35
ICCV88516	36
YeZIN	37
ICCC32	37
ICCV93512	37
ICCV97105	37
CICA1122	38
ICCV92944	38
ICCL83105	38
DICC8196	39
ICC4958	39
DICC8310	40
DICC8174	40
05H172K-08HO4009	40
D05087>F4TMWR2AB015	41
Chefe	41
ICCV92337	42
DICC8022	43
BL	43
WR315	46

ICCV93954	48
CDCVanguard	48
DICC8212	50
DICC8296	51
CDCLeader	52
Amit	53
CDCCorinne	53
Genesis090	53
PI489777	55
Pistol	63
CDC512-51	65
ICCV96836	67
Slasher	68
HatTrick	73
CDCLuna	84