Comparative genomic analysis of two maize lines differing in herbicide resistance

Medhat Helmy Shehata Mahmoud

Prof. dr hab. Wojciech M. Karłowski, Supervisor

Dr Marek Żywicki, Auxiliary supervisor

Institute of Bioorganic Chemistry Polish Academy of Sciences
Protein Biosynthesis Department

Poznań, Poland 2017

I want to express my honest gratitude to prof. dr hab. Tomasz Twardowski,
the Principal Investigator of Maestro grant,
for giving me the opportunity to take part in this interesting project.

*Sincere thanks to my supervisor, prof. dr hab. Wojciech M. Karłowski
for his academic supervision and guidance while doing this thesis
as well as for allowing me to conduct this work in his facility.*

*Special thanks to my auxiliary supervisor, dr Marek Żywicki
for his patience, indulgence and most appreciated and needed guidance and aid.*

*Warm thanks to the whole crew in the Department of Computational Biology
at the Institute of Molecular Biology and Biotechnology,
Faculty of Biology A. Mickiewicz University in Poznan.*

I would like to thank Dr.Kamilla Bąkowska-Żywicka,
who as a good friend was always willing to help and give her best suggestions.

I would also like to thank my parents and my brothers.
They were always supporting me and encouraging me with their best wishes.

Table of content

Abstract

Zea mays, more commonly referred to as maize, is widely cultivated all over the world and constitutes one of the most important crops that used as feed and food resource as well as bio-fuels. Millions of tons of maize are produced every year and there is still a growing demand of production. Food and Agriculture Organization (FAO) estimated that by 2050 food production must increase by 70 percent in order to satisfy the needs of world's population. To achieve that goal, high-production rate pesticides are used to eliminate and reduce natural competitors of maize such as pests, fungi and weeds. Half of currently used pesticides constitute herbicides. The best known is a glyphosate - selective, systemic herbicide with environmental- and human-safe profile. Glyphosate was firstly used as a pre-emergent, preventing the weed seeds from germinating. However, in order to achieve better yields protection, there was a need to apply it as a post-emergent (on grown weed) and for that reason scientists aimed at production of transgenic plants which would be tolerant to the glyphosate. There is a great deliberation until now about the use of genetically modified organisms. In this dissertation, I took the advantage of the existence of maize inbred line, which is naturally resistant to the glyphosate, to study the genetic variations between the tolerant and sensitive maize lines. This goal was achieved by using high-throughput sequencing data. To overcome the complexity of the maize genome (more than 85% of repeated sequences) I made use of two sequencing technologies: Illumina (generating short but highly accurate sequencing reads) and SMRT PacBio (longer but less precise reads). Since PacBio sequencing errors might reach up to 15%, there was a need for establishing the best computational tools for the reads correction. I have compared 5 different tools and found out that HALC performance excelled the rest. Correction results were divided into two criteria: when split or full reads were taken into account. HALC performed best in both cases, Proovread performed better with full reads and LoRDEC with split reads. The reads corrected by HALC provided significantly higher sensitivity in detection of structural variants, compared to raw, uncorrected reads.

I have revealed the existence of more than 11 thousand structural variants, 4 million of single nucleotide polymorphisms (SNPs) and around 800 thousands of insertions or deletions (indels) differentiating between the two studies maize lines. Some of them were located within the gene encoding for the 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), which is a biological target of the glyphosate, however their predicted impact on protein sequence or expression was low. At the same time I was able to identify multiple high impact variants located within the genes encoding other enzymes of shikimate pathway (bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase and chorismate synthase), as well as enzymes and transporter proteins influencing the availability of the phosphoenolpuryvate which is the substrate of shikimate pathway.

1. Introduction

Maize (*Zea mays* L.) also known as corn, originated in the Balsas river valley of present-day Mexico 10,000 years ago. It is a diploid plant, phylogenetically close to *Sorghum bicolor* and belongs to the *Panicoideae* subfamily.

Maize is considered as one of the most important crops all over the world and its production is continuously increasing. Besides maize utilization as food, it also used as livestock feed, raw material for industry production, bio-fuel and a model organism in scientific research. Maize tremendous genetic diversity was useful in elucidation of plant transcriptional networks.

According to the Food and Agriculture Organization of United Nations (FAO) 2014 estimations, there are 18 million hectares of maize in Europe, yielding to the production of around 68 903 hectogram/hectare. Maize is also considered a leading crop in the world, with 825 million metric tons produced in 2010 (Awika, 2011). Moreover, 50% of the annual calories for humans and 34% of the production for animal feed comes from cereals (FAO, 2014).

FAO estimated that about 795 million people of the 7.3 billion people in the world (one in nine) were suffering from chronic undernourishment in 2014-2016. Almost all malnourished people, 780 million, live in the developing countries. Therefore, the estimations might be presented as follows: 12.9 percent (one in eight) of the population in developing counties is chronically malnourished (Worldhunger, 2016). It is expected that the world population will reach 9.1 billion by 2050. This means that the food production would have to increase by 70% to provide an adequate amount of food. However, the world's total arable land has almost reached its maximum. In addition, climate change and environmental degradation reduced the available agricultural lands, resulting in more challenging situation to feed the world population (Alexandratos & Bruinsma, 2012; Hanjra & Qureshi, 2010).

Zea mays has a large genome of about 2.4 gigabases and around 40 000 of genes, spread across 10 chromosomes. Some of the chromosomes contain highly repetitive heterochromatic domains called "chromosomal knobs". The genome is transposon-rich and its organization is complex. It is composed of approximately 85% repetitive sequences, like transposable elements (TE). Moreover, maize is one of the crop species with high genetic variation between different lines of the same species. These differences might reach from 25 % up to 84 %, largely due to the differences in TE content, since the activation and/or loss of TEs have a great impact on the genome structure. Moreover, single nucleotide polymorphisms (SNPs), small insertion and deletions (indels), structural variants (SVs) such as copy number variations (CNV), inversion and translocations, also take a great part in provoking substantial differences between maize genomes. It was shown in numerous studies that especially genome structure variations are associated with a wide range of plant phenotypic traits, involved in metabolic fluctuation and regulation of gene expression (Huang & Han, 2012, Sebat, 2007). This suggests that SVs play important role in plant phenotypic variations.

Number of grasses as well as annual broadleaf weeds represents active competitors of maize in the field. This results in both, yield and economic losses and therefore, the need of specific and efficient elimination of weeds, including application of herbicides such as glyphosate. Glyphosate is a highly specific herbicide that targets enolpyruvylshikimate-3-phosphate synthase (EPSPS), an enzyme which is involved in synthesis of three aromatic amino acids: tyrosine, tryptophan, and phenylalanine.

Glyphosate was primarily used as a pre-emergent herbicide, preventing germination of seeds but in order to simplify its use it is now commonly applied post-emergent, on existing weed. This raised the need for generation of genetically-engineered herbicide-resistant (HR) crops, which can be commercialized. Introducing new genes to produce genetically modified crops (GMC) with beneficial traits (like herbicide-resistance) is of a high demand in order to overcome the continuous need of highly productive crops. However, many studies have shown possible negative effect on soil, water, wild species and humans as a result of using GM plants. This issue raised a special interest in using non-genetically modified organisms (non-GMO) which would be beneficial in specific traits.

In this study, I have analyzed genome structures of two naturally inbred maize lines, which substantially differ in resistance to the herbicide glyphosate. This was achieved by using different sequencing technologies (Illumina and SMRT PacBio) to identify the genome structure variations, single nucleotide polymorphism and insertions-deletions (indels), which could be involved in gaining herbicide resistance without introduction of foreign genes into plant system in non-GMO maize.

2. Background

2.1. Maize importance

Maize (*Zea mays*) is one of the world's leading grain cereals besides rice and wheat. The world production was about 844,4 million tons of grain in 2010 (Nuss & Tanumihardjo, 2010). The leading world suppliers are USA, China, Brazil and Mexico. Humans consume four types of cultivated maize; everta Sturt (popcorn), *Z. mays* var. indentata Sturt (dent), indurate Sturt (flint), saccharata Sturt (sweet) and amylacea Sturt (flour) (Nuss & Tanumihardjo, 2010).

Maize serves as a food and feed source and the three major constituents of maize kernel are lipids, proteins and polysaccharides. A typical maize kernel is composed of 1–3 % of sugar, 4–5 % of lipids, 8–10 % of proteins, 70–75 % of starch and 1–4 % of ash. The kernel contains phytosterols in high levels, which have a role in reduction of cholesterol. The maize kernel serves as a raw material to produce flour, bread, porridge, gruel, steamed products, snacks and other foods (Arendt & Zannini, 2013). In addition, it is used in ethnomedicine as anti-diabetic and anti-inflammatory agent, in urinary diseases, gallstones and malaria (Abo, Fred-Jaiyesimi, & Jaiyesimi, 2008, Owoyele et al., 2010).

Maize natural history began already nine thousand years ago, when farmers from Mexico started to collect wild grass seeds. Teosintes are the closest relatives of maize and archaeological findings suggest that domestication occurred between 6 250 to 10 000 years ago in southern Mexico (Piperno et al., 2007). Female and male flowers in unisexual maize are born on separate stems. Maize is fertilized through natural cross-pollination and wind pollination (Strable & Scanlon, 2009), which makes it amenable to genetic analysis. Maize geographical cultivation range is the widest among all other crops, ranging between Chile (40 °S) and Canada (50 °N) and 3 400 m above sea level in Andean mountains and Caribbean islands, making it to be grown on more areas and continents than any other crop (Tenaillon & Charcosset, 2011).

In contrast to other cereals with high economic value like wheat, rice and barley which utilize C3 carbon metabolism, maize adapted to C4 metabolism, making it highly efficient in carbon fixation. Furthermore, maize adapted to low water availability, high light intensities and temperature (Gowik & Westhoff, 2011). It has been also adopted for phytomanagement of cadmium-contaminated soils (Rizwan et al., 2016).

2.2. Maize genome structure

Maize is a tall annual diploid plant that belongs to *Poaceae* family (true grass). Similar to other plant genomes, 98% of its genome constitutes non-coding sequences. The rest are wide desert of repeats that remain repressed through the cell cycle (Rodgers-Melnick, Vera, Bass, & Buckler, 2016). The maize genome ~70 million years ago went through several rounds of genome duplication from a paleopolyploid ancestor, *Z.mays* ssp. *parviglumis,* (Paterson, Bowers, & Chapman, 2004, Matsuoka et al., 2002; Piperno & Flannery, 2001). Expansive amplification of transposable elements as well as genome duplication attributed much to the size variation of maize and other cereal genomes (Bennetzen, 2000). Maize genome reveals tremendously high levels of genetic diversity represented with SNPs, indels polymorphisms and SVs (Messing & Dooner, 2006). The maize genome is composed of ~ 2.4 billion base pairs (bp) containing ~ 40 000 genes (Law et al., 2015). Genes typically have larger introns than orthologous genes in rice and sorghum, because of insertion of repetitive elements in cyclic way inside each other (Wei et al., 2009; Yang & Bennetzen, 2009). The genes are spread across 10 structurally diverse chromosomes, that undergone dynamic changes in chromatin composition. The complexity of maize genome is also increased by variable amounts of centromeric satellite repeat (CentC).

The reference maize genome was obtained by sequencing B73 maize line, since it is the most utilized inbred line and it gives the highest yield (Strable & Scanlon, 2009). The first assembly of maize genome was released in 2009. The authors used minimum tiling path bacterial artificial chromosome and fosmids, guided by physical mapping. The whole assembly consisted of more than 100 000 small contigs, however poorly ordered and oriented. At the time of 2009, the genome size was estimated to be 2.3 gigabases and about 32 000 of genes, which were annotated using evidence-based approaches, such as expressed sequence tags (EST) and RNA-sequencing (RNA-seq) data as well as *ab initio* approaches. It was revealed that the highly abundant transposable elements are dispersed non-uniformly across the whole genome (P. S. Schnable et al., 2009; Wei et al., 2009). These TEs were responsible for the amplification of numerous gene fragments and as a result affected the genome composition, size and positions of centromeres. Well-annotated protein coding genes as well as extensive RNA-seq resources enabled high accuracy gene model prediction (Campbell et al., 2014, Law et al., 2015).

The genome has been updated since then and the current version 4 was published in 2017 (http://www.maizegdb.org). The improvement in the genome assemble was performed using single-molecule sequencing technology. This allowed for the genome assembly in 2 958 contigs (that is 33 times less than in previous assembly), half of the total assembly was performed with the contigs larger than 1.2 Mb (Jiao et al., 2017).

Maize genome size is 6-fold larger than in rice and roughly the size of a human genome. Moreover, it is organized in a complex way. Genetic analysis of duplicated genes revealed the existence of homologous regions (Wendel, Stuber, Edwards, & Goodman, 1986; Bennetzen, 2000). Transposon-rich maize genome contains about ~855 transposable elements families, where DNA transposons, ale much less frequent than retrotransposons (Schnable et al., 2009). The most common type of transposons in maize are retro transposons, inserted inside each other and repeated in a cyclic way which makes it very hard to assemble the genome. The most complex of these families are *mutator,* massively different in both, sequence and size and 262 Pack-MULEs (Mutator-like elements that contain gene fragments) carrying fragments of 226 nuclear genes (Schnable et al., 2009).

Another type of DNA transposons that exist in maize and cause movement of large pieces of DNA are helitrons. They exist in fungi, plants and animals but they are significantly variable, active and abundant in plants. While in maize helitrons are enriched in poor-gene regions, in other organisms they are located mainly within the gene-rich regions (Schnable et al., 2009; Bennetzen & Hake, 2009; Morgante et al., 2005). More than 75 % of the reference maize genome constitutes long terminal repeats (LTR) which exhibit non-uniform, family specific distribution in maize genome. Due to the differences in TEs content, maize inbred lines genomic sequences differ from 25 to 84 % (Q. Wang & Dooner, 2006). Comparative genomic hybridization assay between maize inbred lines B73 and Mo17 revealed remarkable amount of structure variations (Springer et al., 2009). It was found that TEs play important role in adaption and in resistance, such as in *Drosophila melanogaster* where phenotypic variations have been associated with TE polymorphisms near genes. As an example, repeated adaptive insertion of TEs in the 5′ end of the *Cyp6g1* gene lead to its over-transcription and increasing pesticide resistance (Schmidt et al., 2010).

Single nucleotide polymorphisms (SNPs) were also at focus of many genetic analysis, however it was found out that the major phenotypic consequences are caused by structural variations (Sebat, 2007). Thus, there is a growing appreciation for structure variations role in creation of phenotypic variations (Hurles, Dermitzakis, & Tyler-Smith, 2008).

Considering single nucleotide polymorphisms, it was found that the frequency of SNPs between humans and chimpanzees is smaller than between different maize inbreds lines (Messing & Dooner, 2006; Buckler, Gaut, & McMullen, 2006). It has been estimated that SNPs in maize appear on average at every ~80 bps and insertion or deletion - at every ~300 bps (Fu et al., 2006). As a consequence, when comparing two random inbred maize lines, on average one polymorphism will be found every ~100 bps (Ada Ching Mark Jung, Maurine Dolan, Oscar S (Howie) Smith, Scott Tingey, Michele Morgante and Antoni J Rafalski et al., 2002). Besides, genetic variations within active chromatin regions account for ~40 % of phenotypic variations in agronomic traits (Rodgers-Melnick, Vera, Bass, & Buckler, 2016). Understanding intraspecific variations has crucial implications for plant breeding and improvement of crops.

2.3. Herbicides

It has been estimated that the occurrence of pests, diseases and weeds reduces the productivity of the agricultural crops up to 90 % (OERKE, 2006). Weeds alone cause a great loss in economically important crops - for example a wild grass *Echinochloa crus-galli* reduces maize biomass by 50-57 % per m^2 (M. J. Kropff, 1984, Bosnic & Swanton, 1997). Chemical weed management using herbicides is the most economical and widely used weed management technique in the world (Fernando, Manalil, Florentine, Chauhan, & Seneweera, 2016).

Herbicides are phytotoxic chemical products used to eliminate weeds by inhibition of their germination (pre-emergent herbicides) or growth (post-emergent herbicides) (de Souza, Guedes, & Fontanetti, 2016). Herbicides can be classified as selective ones that targets special weed species, leaving the desired crop unharmed and nonselective ones which eliminate all plant material in the targeted area.

The first use of herbicides was after Second World War by introducing 2,4D (Duke, Stephen; Powles, 2008). Since then, the usage of herbicides increased dramatically in order to satisfy the needs of growing population, and to provide both, economic and labor benefits, thereby reducing the cost of farming and save energy.

Between 1974 and 2014 the usage of glyphosate-based herbicides increased ~100 folds and it is expected to increase even more (Vandenberg et al., 2017). At present, half of the pesticides used constitute herbicides (de Souza, Guedes, & Fontanetti, 2016).

The number-one selling herbicide in the world is glyphosate (N-phosphonomethyl-glycine), developed, commercialized under the trade name of Roundup and patented by Monsanto. Glyphosate is an environmentally safe nonselective herbicide with a broad spectrum of effects. It inhibits action of enolpyruvylshikimate-3-phosphate synthase (EPSPS), an enzyme which is absolutely required for the survival of plants (Dill et al., 2008).

EPSPS is not present in vertebrates therefore, the glyphosate is safe to humans. EPSPS catalyzes the transfer of the enolpyruvyl moiety of phosphoenolpyruvate (PEP) to the 5-hydroxyl of shikimate-3-phosphate (S3P) to produce enolpyruvyl shikimate-3-phosphate (EPSP) and inorganic phosphate in a sixth step of the shikimate pathway (R. Bentley & Haslam, 1990).

EPSPS enzyme is targeted in the plant chloroplast-localized pathway that leads to the biosynthesis of aromatic amino acids (Pollegioni, Schonbrunn, & Siehl, 2011). Thus, targeting the EPSPS enzyme prevents the synthesis of chorismate-derived aromatic amino acids and secondary metabolites in plants; pigments, flavonoids, auxins, phytoalexins, lignin, and tannins (Howe et al., 2002). As a result, various processes are affected, which lead to plant death, including a failure to produce compounds that depends on shikmate pathway, disruption of the carbon flow, and decrease in protein synthesis due to concentrations of aromatic amino acids reduction (Becerril, Duke, & Lydon, 1989). In weeds translocation of glyphosate takes place through the phloem, causing the death of the root system and the reproductive structures of perennial plants such as rhizomes, bulbs, and tubers.

EPSPS enzymes are divided into two classes, based on their intrinsic glyphosate sensitivity. Class I, found in all plants and a number of gram-negative bacteria such as *Escherichia coli*, is inhibited by low micromolar concentrations of glyphosate. Class II exists in microbes, including *Agrobacterium sp.* strain CP4, *Achromobacter sp.* strain LBAA and is naturally glyphosate-tolerant (Barry, Kishore, Padgette, & Stallings, 1997).

Glyphosate was initially used as a pre-emergent, before crop seeding, but for more effective results, it now commonly used as a post-emergent. The latter raises the possibility of elimination of all plants in the treatment area since a glyphosate is a nonselective herbicide. In 1996 a first transgenic glyphosate-tolerant soybean was introduced to the market allowing for safe and economic post-emergent application of a glyphosate, to remove weeds without causing any damage in the crop. In 2012 59 % of 170.3 million hectares was occupied by transgenic crops.

The idea of producing genetically modified plants resistant to glyphosate begins already in 1980 when scientists tried to identify glyphosate-insensitive EPSP synthases that could be introduced to crops in order to gain resistance. But an increase of tolerance to the glyphosate is connected with the decrease in phosphoenolpyruvate (PEP) enzyme affinity and results in decrease of its catalytic efficiency. After identifying naturally occurring glyphosate-tolerant microbes like *Agrobacterium sp.* strain CP4, they were used to produce transgenic glyphosate-tolerant crops. P101S was the first single-site mutation in EPSP gene, reported to conferee resistance to glyphosate in *Salmonella typhimurium* (Comai, Sen, & Stalker, 1983), followed by the discovery of *Klebsiella pneumoniae* G96A single-site mutation (Sost & Amrhein, 1990) and T102I/P106S in *Z. mays* (Pollegioni et al., 2011).

Two basic strategies are followed to produce glyphosate-resistant crops: the first is to intensively express the targeted enzyme (used in the commercial glyphosate-tolerant crops). The second one is detoxification of the glyphosate molecule.

Genetically modified plants resistant to the glyphosate, called Roundup Ready® plants, carry the gene coding for a glyphosate-insensitive form of EPSP enzyme obtained from *Agrobacterium* sp. strain CP4. Once introduced into the plant genome, the gene product, CP4 EPSP synthase, confers the resistance of a transgenic crop to the glyphosate (Todd Funke, Han, Healy-Fried, Fischer, & Schönbrunn, 2006). Another strategy to produce glyphosate-resistant plants is to introduce a mutated version of EPSPS, which activity is not inhibited by the glyphosate (TIPS). As a result, in both cases, the glyphosate accumulates in plant meristems which may interfere with reproductive development and decrease the crop yield (Pline, Wilcut, Duke, Edmisten, & Wells, 2002). Therefore, better results might be achieved by metabolic detoxification of a glyphosate, using native plant gene-encoded or transgene-encoded enzymes (Pollegioni et al., 2011).

There are lingering concerns about the possible effects of genetically modified crops (GMC) on both, health and environment. These issues the acceptance of GMC especially in Europe and Japan (Todd Funke, Han, Healy-Fried, Fischer, & Schönbrunn, 2006; Hellsten, 2006), beside that the current safety assessments rely heavily on studies conducted over 30 years ago (Vandenberg et al., 2017). Dramatic expansion of genetically modified crops caused by the powerful scientific techniques development poses direct and indirect environmental implications. The current state of knowledge shows that GMC convey damaging impacts on the environment such as modification in crop pervasiveness or invasiveness, the emergence of herbicide and insecticide tolerance, transgene stacking and disturbed biodiversity.

The first report presenting the possible risks associated with genetically modified organisms (GMOs) was published in 1982 (Sharples, 1983) and caused more scientific interest on that aspects. The environmental risk of GMOs could be summarized as follows:

I. Risk associated with a biodiversity, such as effects on soil and non-targeted species.

II. Risk associated with a gene flow and genetic recombination.

III. Risk associated with the evolution, such as development of resistant weeds and/or insects (Tsatsakis et al., 2017).

Besides that, the use of GMOs might be associated with a relevant unintended increase of allergens compared with conventionally produced crops (Selb et al., 2017). Also, an introduction of non-native genetically modified plants (GMPs) could lead to potential environmental risk which consequences cannot be predicted. These are: transmission of transgenic sequence to the related wild species, including weeds, by the horizontal gene transfer. Implication of this could lead to evolution of pests and pathogens with high resistance to new pesticides or the emergence of new viral pathogens (Beckie, Warwick, Hall, & Neil Harker, 2012; H.-L. Yu, Li, & Wu, 2011). Another possibility is a spread of transgenic contamination or hybridization between GMCs and its compatible wild type crops (Cruz-Reyes, Avila-Sakar, Sanchez-Montoya, & Quesada, 2015). In addition, there are indirect effects of GMPs on wildlife biodiversity, water, soil, reduction of insect, weeds and pest control. The most important there is reasonable hypothesis that regular use of glyphosate on genetically modified crop field could lead to the development of glyphosate resistance (Tsatsakis et al., 2017).

2.4. Sequencing technologies

The first DNA sequencing method was developed by Sanger and colleagues and Maxam and Gilbert in 1970 using chain termination and fragmentation techniques, respectively. This revolutionary step in biology helped scientists to decipher complete genes and later on the entire genomes. Sanger sequencing enabled completion of the first human genome sequence in 2004. But there was a continuous need for more cost-effective techniques. In the same year, National Human Genome Research Institute (NHGRI) started an initiative to fund a project aiming at reduction of the costs of human genome sequencing to just 1 000 dollars. This activity paves the way to the next-generation sequencing (NGS) technologies development.

NGS techniques sequence nucleotides cheaper and faster and provide higher throughput than the Sanger method. These new methods opened a new era of molecular biology and genomics. NGS was named as the second generation sequencing technology. It has three major improvements over Sanger method: (i) it does not require cloning into bacteria, (ii) NGS processes millions of sequencing reactions in parallel and (iii) base detection is performed cyclically (Van Dijk, Auger, Jaszczyszyn, & Thermes, 2014).

Three major platforms of NGS technologies have been invented: Roche (formerly 454 Life Sciences, Branford, CT, USA, discontinued since 2016), Ion Torrent by Life Technologies - as a part of Thermo Fisher (Waltham, MA, USA) and Illumina (San Diego, CA, USA) (S. T. Park & Kim, 2016).

Illumina sequencing is the most widely used, and supplies most NGS platforms in the world. It currently offers the highest throughput per run and the lowest cost per-base (L. Liu et al., 2012). The recent Illumina platform delivers 1.8 Tb of sequence per run in three days from ~ 6 billion reads with 150 bp in length, it was designed for whole genome sequencing (WGS) (S. T. Park & Kim, 2016).

Genomic research has been revolutionized since the NGS technology has been released because it brings the power of whole-genome sequencing to small laboratories. Besides that, gene expression studies changed from depending on microarrays to NGS, enabling scientists to quantify and identify gene expression without previous knowledge of a particular gene (P. J. Park, 2009). Since the first large-scale project that concerned genetic variation within 1 000 human genomes, several projects have been launched (Genome 10K Community of Scientists, 2009) which significantly increased our understanding of relations between genomic variation and phenotype (Kilpinen & Barrett, 2013). NGS became a crucial technology in basic science - it is used to measure genetic variants between organism and the reference since it enabled whole-exome sequencing (WES), targeted sequencing or WGS. All of these approaches increased our knowledge about SNPs, indels and SVs within genomes, which can now be easily identified using different tools for NGS data analysis (Ng & Kirkness, 2010).

However, there are several disadvantages of NGS technology, such as relatively short reads or biases in regions with high/low GC content. Because both, NGS and

Sanger platform produce reads which are sometimes shorter than length of the repeats, these regions are hard to sequence and as a consequence, the genome assembly becomes more challenging (Van Dijk et al., 2014).

To overcome such hindrances, a single molecule detection system sequencing was developed (Helicos BioSciences) (Pushkarev, Neff, & Quake, 2009). In this system, DNA is not amplified before sequencing. Helicos technology was an intermediate between second and a third-generation sequencing. Pacific Biosciences (PacBio) is the first commercially available third generation sequencing (TGS) technology and allows for unique single molecule real-time (SMRT) sequencing (Schadt et al., 2010).

PacBio presents several advantages over second generation sequencing: (i) average read length of > 20 kbp and maximum read length of > 60 kbp (best record is 92.7 kbp as of Nov. 2016), (ii) low degree of sequence composition bias, (iii) simultaneous epigenetic characterization and (iv) high accuracy of consensus sequence with coverage > 30X.

These advantages enable high resolution and analysis of hard-to-sequence regions in complex genomes (Koren et al., 2013) as well as sequencing of full transcriptomes, which was not possible before (Wang et al., 2016). All of these features make PacBio ideal for *de novo* assembly, finishing genome assemblies, improving draft genomes and finding new genes isoforms.

The SMRT PacBio technology works by detecting fluorescence in a real time by incorporating phosphate-labeled nucleotides with single DNA polymerase during DNA synthesis process (Schadt et al., 2010). PacBio template is a circular, double stranded DNA which is generated by ligation of hairpin adapters at both ends. The sequencing library is loaded in 150 000 wells array with so called zero-mode waveguides (ZMW). The well size is 50 nm in diameter and 100 nm in depth and is presented in nanofabricated consumable chip, 1 cm squared in diameter (Rhoads & Au, 2015).

2.5. Single Nucleotide Polymorphisms and structure variation detection

Various analysis pipelines that combine both, short read sequences aligners and variant callers, are used to detect genomic variations, such as SNPs and indels. For example, a combination of variant callers: Genome Analysis Tool Kit HaplotypeCaller (GATK-HC) (McKenna et al., 2010), Samtools mpileup (H. Li et al., 2009) or Freebayes (Garrison & Marth, 2012) with sequence aligners: BWA-MEM (Heng Li, 2013), Bowtie2 (Ben Langmead & Salzberg, 2012) or Novoalign (http://novocraft.com/) are widely used.

Other methods use heuristic approaches to detect SNPs, for example: Atlas-SNP2 (Brockman et al., 2008) or VarScan (Koboldt et al., 2009). VarScan takes into consideration a combination of features from the sequencing platform (like Illumina and Roche 454) and different alignment methods. Followed by series of filters (e.g. read depth, strand specific depth, per-base quality and number or reads) carrying certain alleles to detect SNPs.

Hoberman et al., 2009 proposed ProbHD pipeline, which uses a machine-learning approach (random forest method), producing heterozygosity score for each base considering mutable features. It was designed especially for Roche 454 and its main features are as follows: per-base quality scores, read cycle (within-read relative position), homopolymer length, strand-specific depths, read alignment quality and total read depth.

Logistic regression used to recalibrate per-base quality scores for every base carrying the non-reference allele as first step in two steps process is done by Atlas-SNP2 (Shen et al., 2010). The first step is to detect SNPs and is followed by accumulation of information across all reads that carry non-reference allele. It also includes read depth and prior knowledge of overall sequencing, adopted by Bayesian approach. In real data, the training set is independent of pre-existing data set generated by the same Roche machine and base-calling technology.

GATK include several algorithms, like UnifiedGenotyper and HaplotypeCaller to call variant from realigned and recalibrated reads. The HaplotypeCaller is able to detect both, SNPs and indels simultaneously, providing more accurate calls than the UnifiedGenotyper algorithm. In the regions of high variability, it ignores the existing mapping information and *de novo* reassembles the reads in that region. A comparison between GATK and SamTools found out that GATK provides more accurate results – the positive predicted value was 92.55 % in case of GATK *vs* 80,35 % in SamTools case. It was also revealed that in difficult to call regions, GATK HaplotypeCaller gives more accurate results (Pirooznia et al., 2014).

Genotype variants could be classified according to their size, where variants of size ≥ 50 bp in length are considered as structure variations (including insertions, deletions, duplications, inversions and large-scale structural rearrangements) (MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014). Structure variations could be considered as drivers of evolution, resulting in phenotypic variations of a trait, ecological adaptation and speciation (Kirkpatrick, 2010; Long et al., 2013; Vlad, Rappaport, Simon, & Loudet, 2010).

SVs are quite abundant and have phenotypic consequences (P. Lu et al., 2012). They are divided into two categories. The first one is balanced SVs, which involve no net loss or gain of genetic material like inversions (chromosome part reversed) and translocations, where segment of chromosome is transferred to the same chromosome (intrachromosomal) or to another chromosome (interchromosomal). The second category is imbalanced, which includes insertion, deletion and duplication (Lin, Bonnema, Sanchez-Perez, & De Ridder, 2014).

Structural variations detection is severely limited by the usage of short reads. Even paired-end reads cannot resolve accurately large-scale structural mutations (Schatz, Delcher, & Salzberg, 2010). However, efficient identification of SVs could be substantially enhanced using long read sequencing platforms such as PacBio, 10X genomics or Oxford Nanopore (Koren & Phillippy, 2015, H. Lu, Giordano, & Ning, 2016). Different tools were developed to identify structure variations. PBhoney (English, Salerno, & Reid, 2014) uses two alternative algorithms, long-read discordance (PBhoney Spots), and interrupted mapping (PBhoney Tails). Sniffles (Sedlazeck et al., 2017) uses evidence from split-read alignments, high-mismatch regions, and coverage analysis to identify SVs. MultiBreak-SV (Ritz et al., 2014) identifies structural variants from next-generation paired end data, third-generation long read data, or data from a combination of sequencing platforms. Parliament (English et al., 2015) is a publicly available consensus SV-calling infrastructure that merges multiple data types and SV detection methods.

2.6. SMRT PacBio correction

It has been estimated that the error rate in long sequencing reads is ~ 15% of read length and is randomly distributed. For that reason, in order to analyze the data produced by SMRT PacBio instrument, different algorithm approaches have been developed to overcome these high error rates. These algorithms either depend only on long reads or combine them with other, short high fidelity reads produced by PacBio machines (e.g. circular consensus sequences - CCS, which accuracy is ~ 99.99 %) or any other short, accurate read technologies (such as Illumina). These strategies were first designed for SMRT PacBio, as it was the first commercially available long-read sequencing technology (Koren & Phillippy, 2015).

Overlap-Layout-Consensus (OLC) methods

It was the first method used together with the whole genome sequencing like Celera Assembler (E. W. Myers et al., 2000) and Allora (Rasko et al., 2011). The algorithm assembles the genome, first by detecting overlap between reads, then by assembly using the Overlap-Layout-Consensus approach.

OLC performs perfectly with high fidelity sequencing reads which contain few errors. But when the error rate is reaching ~ 15 %, this approach is more computationally demanding to detect overlaps between the reads and produces high rates of false-positive and false-negative overlaps. Such overlaps could lead to the production of misassembled genomes in its worst case, or at least complicate the assembly graph. Therefore, to facilitate OLC assembly of long erroneous reads, they are corrected using the hierarchical methods described below.

Hierarchical methods of reads correction

Hierarchical non-hybrid method

In this approach, long reads are corrected without usage of short reads. The algorithm begins with aligning of long reads against each other. In this way, the most obvious overlaps are identified first, usually consisting of one read entirely contained within another.

Longest sequences are then corrected using a consensus of the data and assembled with an OLC method. Around 24 hours are needed in this approach to assemble bacterial genome and coverage around 100 x (Koren & Phillippy, 2015). The available correction tools that are based on this approach are: PBcR (Koren et al., 2012), HGAP (Chin et al., 2013), Dazzler (G. Myers, 2014) and Sprai (Miyamoto et al., 2014). Different comparisons of assemblies have demonstrated that the hierarchical non-hybrid approach outperforms others, when sufficient coverage is available (Harhay et al., 2013).

Hierarchical hybrid method

It begins with a process called pre-assemble, correction or scrubbing, to improve the quality of long reads. It involves mapping of multiple short reads to a single long read. Errors are then identified and corrected using consensus alignment.

In a hybrid mode, highly accurate complementary reads need to be used, like Illumina short reads, CCS or any other short reads. The corrected sequences are highly accurate and can be assembled using a traditional OLC approach afterwards.

The following tools use this approach: PBcR (Koren et al., 2012), LSC, ECTools (Lee, Gurtowski, & Yoo, 2014), LoRDEC (Salmela & Rivals, 2014), Proovread (Hackl, Hedrich, Schultz, & Forster, 2014) and DBG2OLC (Ye et al., 2015).

Other approaches focused on the performance or called "assembly boosters" and rely on improvements from pre-assembling (contigs) the secondary technology prior to the correction (Lee et al., 2014) or aligning the long reads to a *de Bruijn* graph (read threading) in which the tools try to solve short read assembly graph using long reads. These methods perform well even when there is a low coverage (20x-50x).

Hierarchical hybrid tools used for SMRT PacBio correction

Proovread

Proovread is a hybrid correction pipeline for SMRT PacBio reads, that is flexible and can be adapted to the different hardware, from laptop to the high-performance computing cluster. Proovread is based on the following hybrid alignment scoring scheme:

I. The costs of gaps in long reads (LRs) which represent deletions are about twice the gaps in short reads (SRs), which represent insertions.

II. Penalty of mismatches are estimated as at least 10 times the cost of SR, because substitutions compared to SRs are rare (about 1 %).

III. Contrasting to biological scenarios, continuous insertions or deletions are less likely in PacBio reads, because the errors distribution is random so they assign higher cost for gap extension over gap opening.

Proovread by default uses BWA-MEM as a mapper. It accepts either FASTA or FASTQ format. Beside short reads, Proovread also can use unitigs or SAM file, leaving the mapping to the user. The software workflow is as follows: short reads are mapped onto erroneous and chimeric long reads, the resulting mappings are refined to distinguish between valid and non-valid mapping, using algorithm to assess length normalization score. Dividing of LRs to the consecutive bins and each SRs mapping assigned to a bin by its center is then achieved. For each bin, the highest alignment score is considered for the next step, which is a calculation of the consensus sequence.

In order to compute the consensus sequence, the Proovread uses matrix, where each nucleotide is represented by a column for each LR. Then, the consensus sequence is filled with alignment information for short reads, where empty cell represent insertion in LR and multiple nucleotides in cell represent deletion. Alignment with gaps near the end is trimmed. If there are no SR bases in the alignment, then the LR sequence is kept. During consensus sequence generation, the majority of errors are removed and possible chimeric break points are identified. New quality scores are concluded from the coverage and the composition of the consensus sequence at each position. Afterwards, the processed reads and chimera annotations are written to files.

Reads obtained from the consensus step constitute untrimmed corrected reads. In the next step, they are returned in FASTQ format with ASCII-encoded consensus phred quality scores. The reads contain both, high-accuracy regions as well as uncorrected regions and unprocessed chimeras. Such reads can be trimmed using a quality cutoff and the chimera annotations, resulting in high-accuracy long reads (Hackl et al., 2014).

LoRDEC

Long Read DBG (*de Bruijn* graph) Error Correction (LoRDEC), unlike other tools which uses alignment to call consensus, which require long running time and it is also parameter dependent, LoRDEC depends on *de Bruijn* graph.

Instead of aligning the short sequences directly to the long reads, LoRDEC first builds a concise DBG representing the short reads. For each erroneous region of LRs, the tool is searching for an optimal path within the DBG by traversing the LR through the appropriate paths to find an alternative correct sequence. To avoid introduction of erroneous bases during the correction process, the software filters out any *k*-long substring (*k*-mer) which occurs less than *s* times within the SRs (*s* is defined by the user). LoRDEC performs two passes over the long read, one in each direction. This program requires 93% less memory than any other software and performs six times faster (Salmela & Rivals, 2014). The program takes long and short read and an odd integer *k* that will be used as *k*-mer size. LoRDEC estimates a threshold - number of times *k*-mer appears in the read set. Each solid *k*-mer makes a node in the DBG graph and arcs links two nodes if they are overlaying by *k-1* to form the DBG graph. LoRDEC uses GATB libraries (searching for an optimal path within the DBG), which uses bloom filter to store the DBG graph and also store false-positive *k*-mers, which helps to traverse only solid *k*-mers. For each PacBio read, LoRDEC tries to find equivalent solid *k*-mer in the DBG graph, for weak *k*-mers in LR the software corrects it by finding the best path between two solid *k*-mers within the DBG graph boarding the weak *k*-mer. For each weak *k*-mer existing between two solid *k*-mers, LoRDEC finds an alternative path from the DBG and modifies the sequence on the fly. As result, the weak *k*-mers turn into solid ones and finally, the shortest path between the first and last solid *k*-mer is found.

In the last step, FASTA file is created with bases classified as solid or weak. This classification depends on the base belongings to the solid or weak *k*-mer. Solid bases are presented in upper cases and the weak ones in the lower cases. Two options exist, either trimming reads only (the tools will trim only weak bases in each end) or trimming and spitting (the tool will remove all weak bases).

LSC

LSC is a computational method to correct LRs using SRs, aiming at reducing the error rate of sequencing reads. LSC applies a homopolymer compression (HC) transformation strategy to increase the sensitivity of SR-LR alignment, without loss of the alignment accuracy. This reduces the error rate by ~ 3 folds.

The idea of HC is that any two consecutive nucleotides of compressed sequence must be different, for example “GGGCCCAAA” will be transformed to “GCA”. This reduces the information content and makes it useful for alignment, so instead of having four degrees of freedom in each new position, only three exist. This reduces probability of finding repeat alignment hit by chance, because compressed reads have equivalent length of regular reads by a factor $\log_4(3)$.

The program works in five steps: SRs quality control, HC transformation, SRs-LRs alignment, error correction and decompression transformation (Au et al., 2012). It uses Novoalign for aligning because of its sensitivity, but for better computational efficiency it also can be modified with faster aligners like BWA (Heng Li & Durbin, 2009) or Seqalto (Mu et al., 2012). After the alignment is performed, the LRs are modified according to consensus information from the aligned SRs and four types of correction points are performed: HC points, point mismatches, deletions and insertions.

PBcR

PacBio corrected reads (PBcR) is the first method that was introduced as a correction algorithm and assembly strategy that uses short high-fidelity sequences to correct the error in single-molecule sequences, implemented as part of the Celera Assembler, it trims and corrects reads.

First, high-identity short-read sequences are simultaneously mapped to all long-read sequences, then repeats are resolved by placing each SR sequence in its highest identity repeat copy, chimera and trimming problems are detected and corrected within the long-read sequences, and lastly, a consensus sequence is computed for each LR sequence, based on a multiple alignment of the short-read sequences.

The corrected, 'hybrid' PBcR reads might be then *de novo* assembled alone, in combination with other data or exported for other applications. The algorithm to correct and assemble PacBio RS sequences is using an OLC (overlap-layout-consensus) approach (Koren et al., 2013).

HALC

A lot of tools make use of the continuity of contigs and de Bruijn graph, because that approach enables more error rich regions in the long reads to be aligned and corrected, and also the alignment of long reads to the contigs or de Bruijn graph is much faster than aligning short reads to long reads.

Examples on that are LoRDEC (Salmela & Rivals, 2014) and Jabba (Miclotte, Heydari, Demeester, Audenaert, & Fostier, 2015), but most of the correction tools lose a great amount of data (Bao & Lan, 2017), that is because:

- Lack of reference read, because some long reads do not have enough short reads coverage.
- High error rate regions in the long reads, which makes it difficult to align reference reads to it.

Some of the existing tools address the high error rate regions in the long reads, by using contigs or de Bruijn graphe, but none can address the lack of reference issue.

HALC: High throughput Algorithm for Long read error Correction (Bao & Lan, 2017); further address the error richness problem and the lack of reference data. It uses contigs assembled from short reads; First HALC aligns the long reads with relatively low identity requirement to the assembled contigs, so the long reads will not only align to the true genome region but also to genome region's repeats in the contigs. This allows overcoming the lack of reference reads issue. After that it validates each aligned region using referencing other long read regions' alignments and adjacent alignment based validation approach which is the alignments of a long read region and its adjacent regions in the same long read are validated together, and the ones aligned adjacent to each other in the contigs are accepted.

3. Goals

The main goal of the dissertation was identification of genomic variations between two *Zea mays* lines, obtained with classical breeding methods and characterized with substantial differences in sensitivity to the herbicide glyphosate. A putative role of genetic variants in gaining herbicide resistance without introduction foreign genes into plant system was assessed. The scope of the study was the computational analysis of genome sequencing data obtained from both lines.

In order to achieve this goal, first the maize genome complexity had to be overcome. Therefore, two different high-throughput sequencing technologies were used: different insert size Illumina short read libraries, for the detection of small genetic variations such as single nucleotide polymorphism (SNPs), insertions and deletions (indels), and SMRT PacBio long reads to detect larger structural variations.

SMRT PacBio continuous long reads (CLR) have a high error rate, reaching ~ 15% of randomly distributed errors, with high tendency towards deletions and insertion. Prior to use of PacBio sequencing data for structure variant analysis, reads correction was required. Since there are number of tools available, designed specifically for PacBio read correction, my first objective was to assess the performance and accuracy of hybrid correction tools.

Secondly, Illumina reads were used to detect SNPs and indels between the two *Z.mays* lines and to find the specific genotype variants associated with glyphosate resistance.

The final step was to use PacBio to identify SVs in both lines, and to functionally annotate both variants from Illumina and PacBio.

4. Materials and methods

4.1. NGS and PacBio data

The reference genome that was used to align and predict genetic variants was maize AGPv4. It was obtained from plant Ensembl genome browser (ftp://ftp.ensemblgenomes.org/pub/plants/release-33/fasta/zea_mays). Annotation of SNPs, indels and structure variations were performed using variant effect predictor (VEP) version 88 (https://github.com/Ensembl/ensembl-tools/archive/release/88.zip). The standalone offline version of VEP was used, the case files used to annotate data were downloaded from (ftp://ftp.ensemblgenomes.org/pub/plants/release-33/vep/zea_mays_vep_33_AGPv4.tar.gz).

Plant sequence data source

The genomes of two non-GMO inbred *Zea mays* lines differing in herbicide tolerance were sequenced by Dr Agata Tyczewska and Dr Joanna Gracz at the Institute of Bioorganic Chemistry Polish Academy of Sciences in Poznań, Protein Biosynthesis Department. Those lines were acquired from Plant Breeding and Acclimatization Institute - National Research Institute.

The herbicide glyphosate was applied post-emergent to the plants and the response of the two lines was studied (Figure (4.1)). Three weeks after application of the glyphosate, the sensitive line was dead while the tolerant line was still growing.

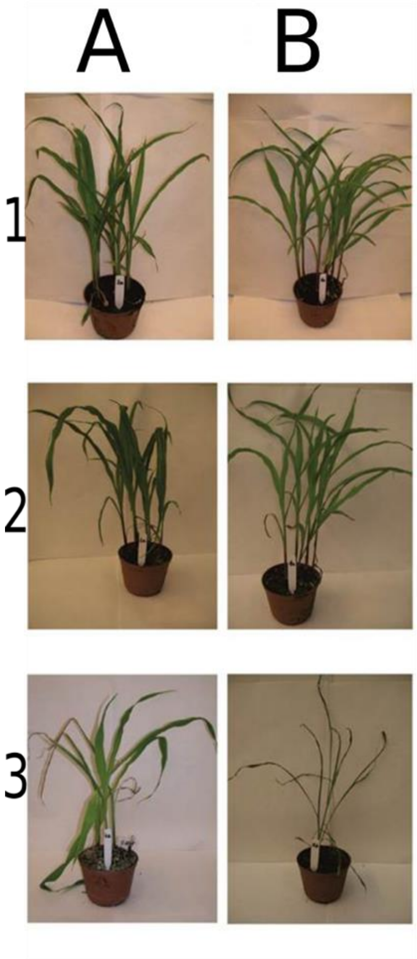

Fig 4.1 Differential response of investigated maize lines to glyphosate. Tolerant plant(A) and sensitive plant (B) one week (1), 2 weeks (2) and three weeks (3) after applying the herbicide (provided by A. Tyczewska, J. Gracz, T. Twardowski, IBCh PAS).

Two sequencing technologies were used: Illumina short reads and SMRT PacBio long reads. Illumina libraries were as follows: two paired-end libraries with insert length of 400 bp and 500 bp and two mate-pair libraries with insert length of 8 kb and 11 kb. For long read SMRT PacBio, RS II technology was used.

Preparation and assessment of sequencing data

Illumina sequence files were transformed first to ubam files, using Picard FastqToSam tools version 2.2.2 (http://broadinstitute.github.io/picard), adding library and group name for each newly created file. Illumina reads were sorted and adaptors were marked using SortSam and MarkIlluminaAdapters respectively from picard package. Reads were aligned using BWA-mem with default parameters, and seting –M flag to mark split reads as secondary aligned. Coverage was calculated using plotCoverage script from deepTools pacakge (https://github.com/fidelram/deepTools).

For SMRT PacBio, 38 SMRT cells for tolerant line and 40 SMRT cells for sensitive line were sequenced. The raw data encoded in H5 format, represent the polymerase reads which is formed of both the DNA that was sequenced plus the adaptors. To extract sub-reads as shown in figure (4.2), adaptors need to be removed. For this purpose, RS_Subreads.1 protocol was used from SMRT portal version 2.2.0 (https://github.com/PacificBiosciences/SMRT-Analysis/wiki/SMRT-Analysis-Software-Installation-v2.2.0).

Reads were aligned using BWA (Heng Li, 2013) version 0.7.10 using -x pacbio which set these parameters: minimum seed length (-k17), where matches shorter than 17 will be removed, max gap set to 40 where gaps longer than 40 will not be found (-W40), (-r10) a key heuristic parameter for tuning the performance it re-seeding for a MEM longer than minimum seed length, matching score 2 (-A2), mismatch penalty 5 (-B5), gap opening penalty 2 (-O2), (-E1) setting gap penalty to gap opening + minimum seed length * E), clipping penalty 0 (-L0). And –M flags which mark shorter split hits as secondary.

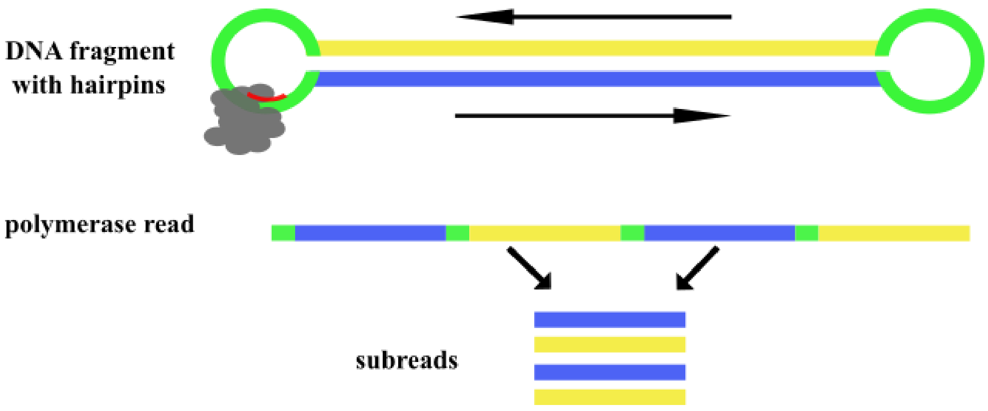

Figure 4.2 Illustration of PacBio sub-reads extraction from the polymerase read. At the top of the graph is the SMRT bell template, formed of green hairpin adaptor ligated to double stranded DNA in blue and yellow, the enzyme (DNA polymerase) in gray will replicate giving the polymerase read, when adapter sequences are removed from the polymerase read, the read is split into multiple sub-reads.

4.2. Assessment of the correction tools performance

Organism selection

Organisms were selected from different kingdoms showing diverse genome structure and complexity. The organisms were as follows: *Homo sapiens, Oryza sativa L., Trypanosoma brucei, Saccharomyces cerevisiae* and *Escherichia coli K-12*. They were all downloaded from Ensembl (http://www.ensembl.org/index.html) database. From each organism, I chose one chromosome.

Human *(Homo sapiens)* chromosome 21, which contains 46,709,983 bp, including 234 coding genes, 400 non-coding genes, it consists of different classes of interspersed repeats, and it is the smallest human chromosome, comprising about 1.2% of the human genome.

Rice (*Oryza sativa L.*) recognized as the leading experimental model for functional and evolutionary genomics of cereals. Chromosome 3 was chosen, it is the second largest rice chromosome and one of the most euchromatic chromosomes. The chromosome contains 36,413,819 bp including 4,271 coding and 5,456 non-coding genes.

Trypanosoma brucei, is the etiological agent of human sleeping sickness and Nagana (Animal trypanosomiasis or sleeping sickness) in animals. Here, chromosome 11 was chosen, it contains 5,261,801 bp and includes 1,693 coding and 82 non-coding genes.

Baking yeast *(Saccharomyces cerevisiae)* was the first Eukaryotic genome to be completely sequenced. Species was diverged approximately 600 to 300 million years ago and it is a significant tool in the study of DNA damage and repair mechanisms. It contributed to the identification of arguably more mammalian genes that affect aging than any other model organism; chromosome 4 was chosen which contains 1,531,933 bp, 853 coding and 32 non-coding genes.

Escherichia coli K-12 genome was observed to contain a significant number of transposable genetic elements, repeat elements, cryptic prophages and bacteriophage remnants. Genome consists of 4,558,660 bp 4,051 coding and 174 non-coding genes.

Data for correction assessment

To simulate short pair-end Illumina reads ArtificialFastqGenerator version 1.0.0 (Frampton, Houlston, Gardet, Stevens, & Sharma, 2012) was used. It takes a reference genome sequence as input and outputs artificial paired-end FASTQ files containing Phred quality scores, the default parameters were used except for the peak coverage mean for a region (-CMP), it was set to 50 and the length of each read (-RL) was set to 100.

Pbsim version 1.0.3 (Ono, Asai, & Hamada, 2013), was used for simulating CLR PacBio reads with depth equals to 20. The model provided by the software was used and error distribution set to 60% insertions, 30% deletions and 10% substitution.

Artificial read simulator Pbsim generates a MAF file besides the artificial reads file, this file contains both the origin sequence of the read and the artificial erroneous generated reads in format of the MAF alignment. The erroneous reads will be used by the correction software, and then the result of the correction will be compared with the read origin (error free read) in the MAF file to assess the correction efficiency of the used software

Correction of long PacBio reads

LoRDEC (Salmela & Rivals, 2014), version 0.5 was used with k-mer size of 21 and solid k-mer 3. Proovread version 2.12 (Hackl et al., 2014) was used with the default parameters, except of the minimum correction read length, which was set to 200 bp. PBcR version 8.3 (Koren et al., 2013) and LSC version 2.0 (Au et al., 2012) were used with default parameters, and HALC (Bao & Lan, 2017) version 1.1 was used to corrected long reads using contigs from short reads assembled by using SOAPdenovo2 (Luo et al., 2012) version 2.04 with default parameters.

Correction tools are classified as follow: class I that produces both trimmed and untrimmed reads, like LoRDEC, Proovread and HALC. Class II produces only one kind of reads either trimmed or untrimmed, for example PBcR, produces only trimmed reads. LSC produces only untrimmed reads.

The difference between each type of reads that are produced is as follow: the untrimmed reads where there are parts of reads that were not corrected, and still included in the corrected reads; the trimmed reads, in which parts of long read containing regions that were not corrected are removed. Figure (4.3) show the difference between outputs in each case.

To evaluate the correction process, first the untrimmed corrected reads were aligned to the original sequence from MAF file to measure similarity after correction. The same process was done to the trimmed corrected reads. Similarity, reads length, number of reads and nucleotides, and number of lost bases and reads were put in consideration when evaluation the efficiency of correction, for both trimmed and untrimmed reads.

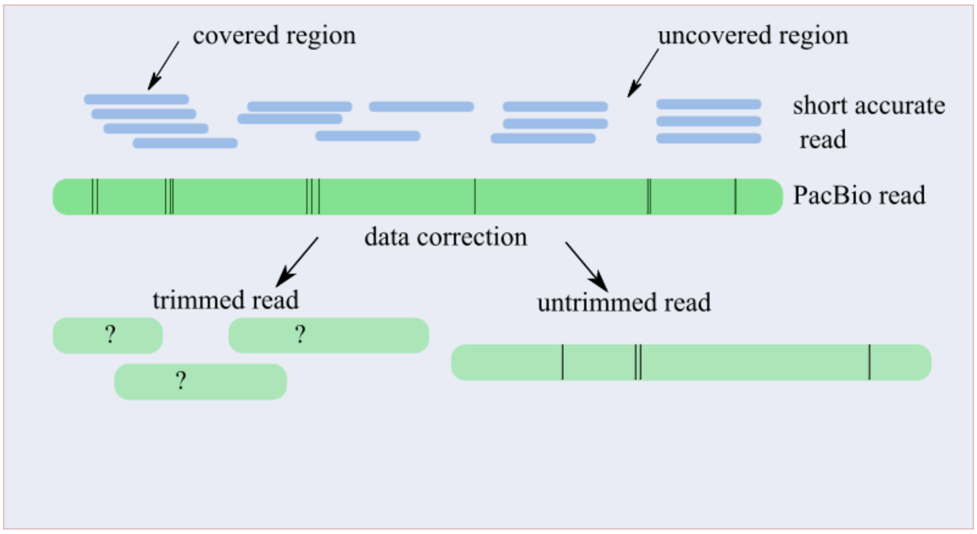

Figure 4.3 **Differences between trimmed and untrimmed reads**. Short reads are presented in blue color, PacBio uncorrected read showed in green color contain errors (black bars). Corrected PacBio reads (light green) contain uncorrected bases (black bars) or bases that were not corrected or wrongly corrected (question mark).

Trimmed reads were aligned locally using Smith-Waterman algorithm, while untrimmed reads were globally aligned using Needleman-Wunsch algorithm. The aligning parameters for both local and global were set to: gap opening 10 and gap extension 0.5. The remaining parameters were default. Similarity result between the corrected reads and the original reads used as indicator of correction efficiency.

GC content, complexity and repeat percentage for corrected and uncorrected PacBio reads were measured using in house Python and R scripts. For complexity calculation, Local Composition Complexity (LCC) was used. RepeatMasker (http://www.repeatmasker.org/) version open-4.0.6 was used to identify repeats region, and bedtools intersect to identify reads falling in these regions. From those results, correction status and number of corrected bases were calculated.

4.3. Calling variance

GATK haplotypecaller (GATK HC) version 3.5 (DePristo et al., 2011) was used to call variance between maize lines. I followed the GATK best practice workflow as shown in Figure (4.4). The first step was to convert FASTQ reads into unmapped BAM (uBAM) file, using Picard FastqToSam tools version 2.2.2 (http://broadinstitute.github.io/picard). Second, MarkIlluminaAdapters from Picard tools was used to mark adapters in uBAM file.

Reads were aligned using BWA-MEM version 0.7.10 (Heng Li & Durbin, 2009), duplicates were marked using Picard MarkDuplicates and then masked. Reads were realigned around identified indels as suggested by the GATK best practice using RealignerTargetCreator.

Variants were called using GATK HaplotypeCaller with --emitRefConfidence flag (mode for emitting reference confidence scores), for cohorts' variant calling. Lastly GenotypeGVCFs from GATK tools used to call the raw SNPs and indels from gVCF from the previous step.

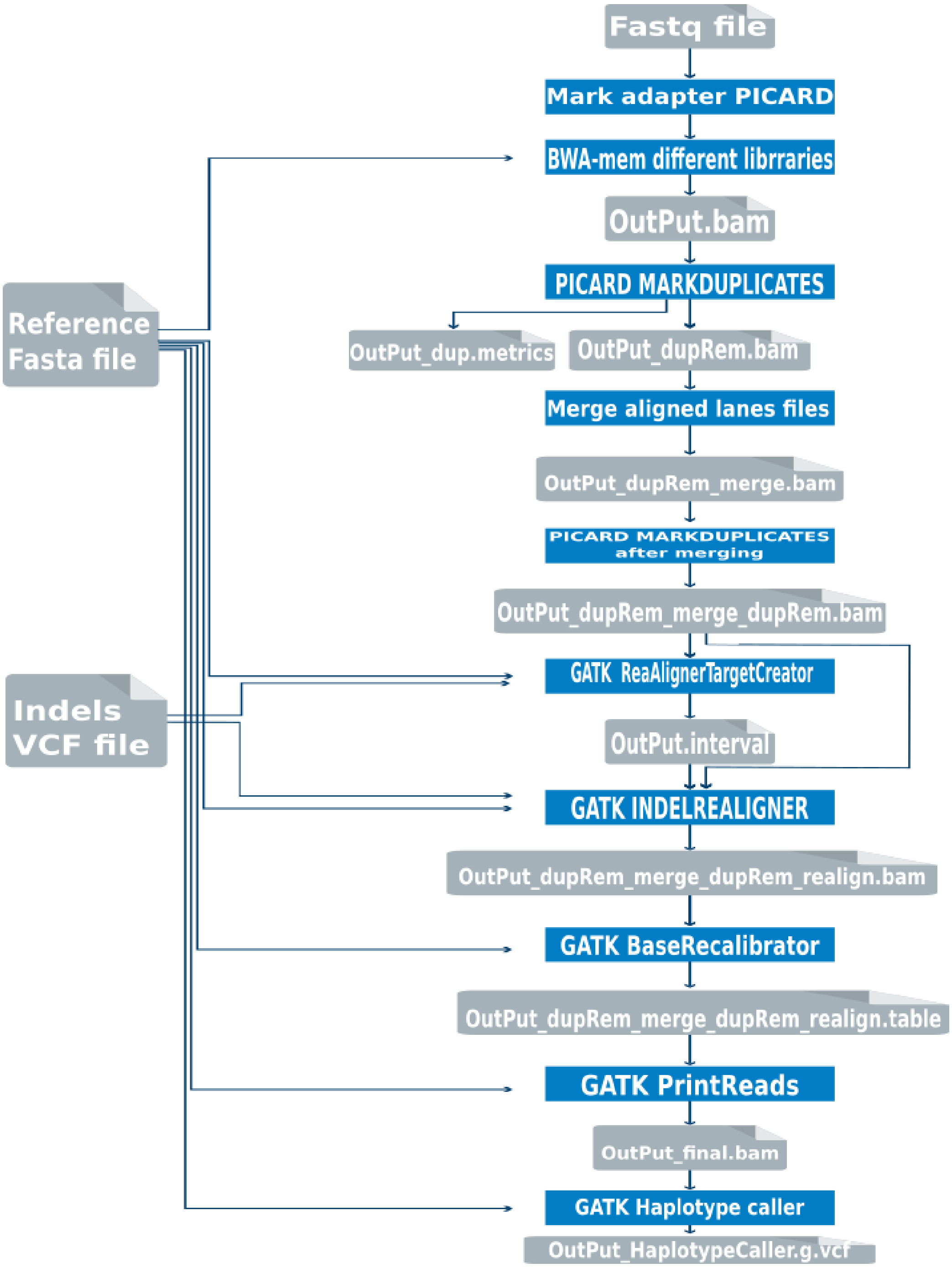

Figure 4.4 GATK haplotype caller workflow for SNPs and indels.

Variants were filtered using hard filtering for SNPs, using best practice suggested parameters, where SNPs matching any of these conditions will be marked FILTER, and will not be considered in further analysis, the remaining SNPs will be annotated as PASS (QD < 2.0). The employed parameters include: QualByDepth is the variant confidence (from the QUAL field) divided by the unfiltered depth of non-reference samples reads – variants with value less than 2 were omitted; Fisher's Exact Test, which detect strand bias, more bias is indicative of false positive calls SNPs with Fisher value greater than 60 (FS > 60.0); SNPs that have Root Mean Square of the mapping quality of the reads across all samples less than 40 were ommited (MQ < 40.0); the u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with ref bases *vs.* those with the alternate allele less than -12.5 (MQRankSum < -12.5); the u-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele, so if the alternative allele only seen near the end of read this indicates error (ReadPosRankSum < -8.0); and for indels I used "--filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0".

Structure variations detection

Two methods of structural variants identification were employed, Sniffles (Sedlazeck et al., 2017) in version 1.0.3 and PBSuite (English, Salerno, & Reid, 2014) in version 15.8.24. In case of Sniffles, after aliging raw uncorrected PacBio reads with BWA-mem using the parmeters mentioned before, structural variants were detected with minimum number of supporting reads set to 2. PBSuite was used with both corrected and uncorrected subreads. I used tails and spots algorithm to identify SVs based on interrupted mapping and read discordance respectively. Minimum read supporting SVs were set to 2. Next, I filtered SVs to only SVs that supported with 5 reads. Blasr (Chaisson & Tesler, 2012) version 2.0.0 was used to align the reads (corrected and uncorrected) default parameters were used except for -bestn 1 to report the top aligned read, –sam to get output in sam format, and -clipping subread to report clipped reads in the sam output file, the resulted sam file was converted to bam then using samtools. The bam file used as input for PBSuite pie algorithm which extracts unmapped, soft-clipped read tails and consolidate them with the aligned results. After this step both Tails and Spots algorithms from PBSuite were used to identify SVs.

The comparison between variants detected with multiple approaches was performed using bedtools suite. The number of exact overlap between corrected and uncorrected subreads identified SVs were identified using bedtools intersect by setting -f 1.0 for 100% overlap and set –r flag which requires fraction of overlap to be reciprocal for both samples.

Functional analysis

The identified SNPs and indels were filtered, selecting only variants that were specific to the glyphosate-tolerant *Z. mays* line. SnpSift (http://snpeff.sourceforge.net/SnpSift.html) was used, by using filter option with genotype (GEN[1].GT == '0/0'), where “GEN[1]” refers to the sensitive line genotype, thus any identified SNPs or indels that does not meet this criteria will be filtered out.

Variant Effect Predictor (VEP) was used to annotate SNPs. In case of SVs the gene ontology term accession associated with the SV affected genes were extracted using plant BioMart from Plants genome browser (http://plants.ensembl.org/index.html), then was summarized using REViGO (Supek, Bošnjak, Škunca, & Šmuc, 2011).

5. Result and discussions

5.1. Characterization of the genome sequencing data

Illumina reads

BWA-mem was used to align Illumina reads to the reference genome. Next, alignment rate for each library and the genome coverage (percentage and coverage) were calculated (Table 5.1). Observed high alignment rates indicate that both lines are relatively close to the reference maize B73 line. Also, the overall average coverage with Illumina reads is relatively high (>67x). The detailed examination of coverage distribution revealed that in both lines one can observe the enrichment of bases with coverage ~30x (Figure 5.1, left panel), suggesting the saturation of the genome with sequencing data. Moreover, based on cumulative distribution of genome coverage, one can observe that around 45% of the genome was covered with at least 50 reads. Based on those observations, I concluded that the coverage with high accuracy Illumina reads is sufficient for variant identification.

Table 5.1 The alignment results for Illumina libraries in glyphosate-tolerant and glyphosate-sensitive *Zea mays* lines.

	Tolerant line			Sensitive line		
Library	**Millions of reads**	**Genome coverage**	**Alignment rate**	**Millions of reads**	**Genome coverage**	**Alignment Rate**
400 bp	816	37.5	97.1%	821	37.7	96.6%
500 bp	664	27.8	85.3%	615	29.1	97.0%
8 kb	129	3.7	86.4%	74	3.0	87.7%
11 kb	59	2.2	86.0%	116	2.6	70.0%

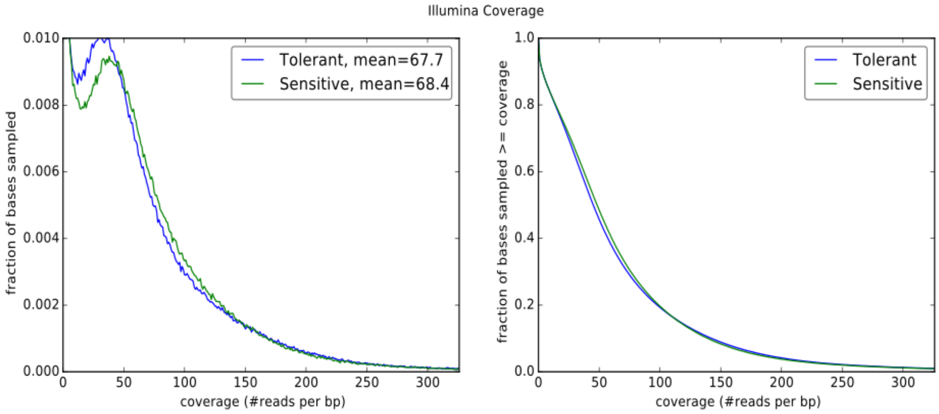

Figure 5.1 Distribution of genome coverage with Illumina reads. The results for the glyphosate-tolerant *Z. mays* line are presented in blue and the results for the sensitive *Z. mays* line – in green. A graph on the left shows the distribution of base coverage while the graph on the right presents the cumulative distribution of genome coverage.

PacBio reads

In the first step, I have performed filtering of PacBio reads using SMRT portal. The results are shown in Table (5.2). I have observed that the qulaity of reads after filtering reached ~84%. Filtering removed mostly the very short, erroneous reads, resulting in high quality dataset for downstream analyses. In the next steps, subreads representing the actual inserts were extracted and corrected with HALC. To assess the genome coverage with obtained reads, I have aligned them to reference maize genome using blasr. A summary of SMRT PacBio reads can be found in Table (5.2).

The average coverage for both lines was > 3x, with a slightly higher value observed in sensitive line (3.59x vs 3.06x in tolerant line). Around 75% of the genome was covered (Figure 5.2, right panel), suggesting rather uniform distribution of the reads across the genome. Due to low overall coverage of the genome with PacBio data, I have investigated in more details the coverage for genic and intergenic regions separately, to assess the ability of structural variant identification (Figure 5.3). In genic

Table 5.2 PacBio reads in glyphosate-tolerant and sensitive *Z. mays* lines.

	Tolerant				Sensitive			
	raw	quality filtered	subreads	corrected	raw	quality filtered	subreads	corrected
Number of bases (millions)	8 402	7 514	7 489	7 309	10 166	9 136	9 107	8 898
Number of reads (millions)	5.711	1.711	2.226	2.226	6.011	1.891	2.497	2.497
N50	6 401	6 476	4 594	4 509	7 142	7 211	4 931	4 837
Mean read length	1 471	4 389	3 363	3 283	1 691	4 830	3 646	3 562
Mean read quality	0.27	0.836	-	-	0.281	0.84	-	-
Coverage	-	-	3.06	-	-	-	3.59	-

regions, the difference in coverage between sensitive and tolerant line turned out to be increased (accordingly, 78% vs 62% of genic regions covered). The minimum coverage of 2 necessary to identify the structural variants was observed for 55% or 70% of genic regions and 60%or 65% of intergenic regions in tolerant or sensitive line

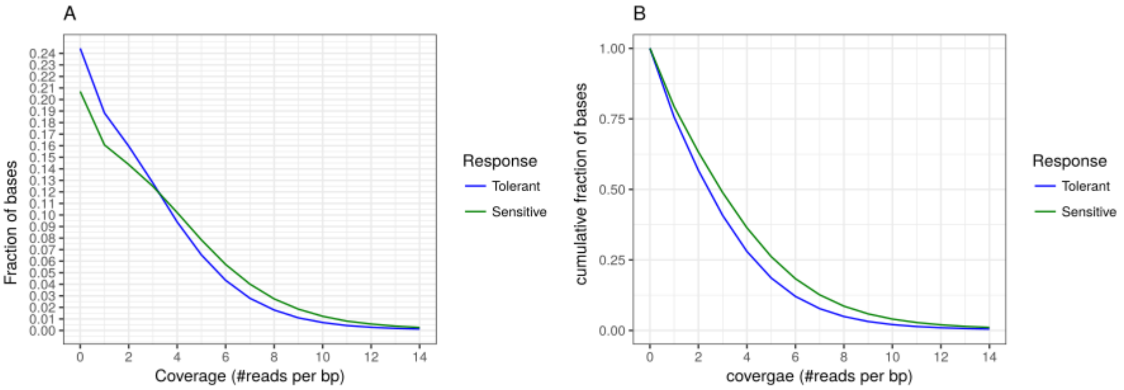

Figure 5.2 Distribution of genome coverage with PacBio reads. (A) Distribution of genome coverage, (B) cumulative genome coverage showing that 75% of genome has been covered with data.

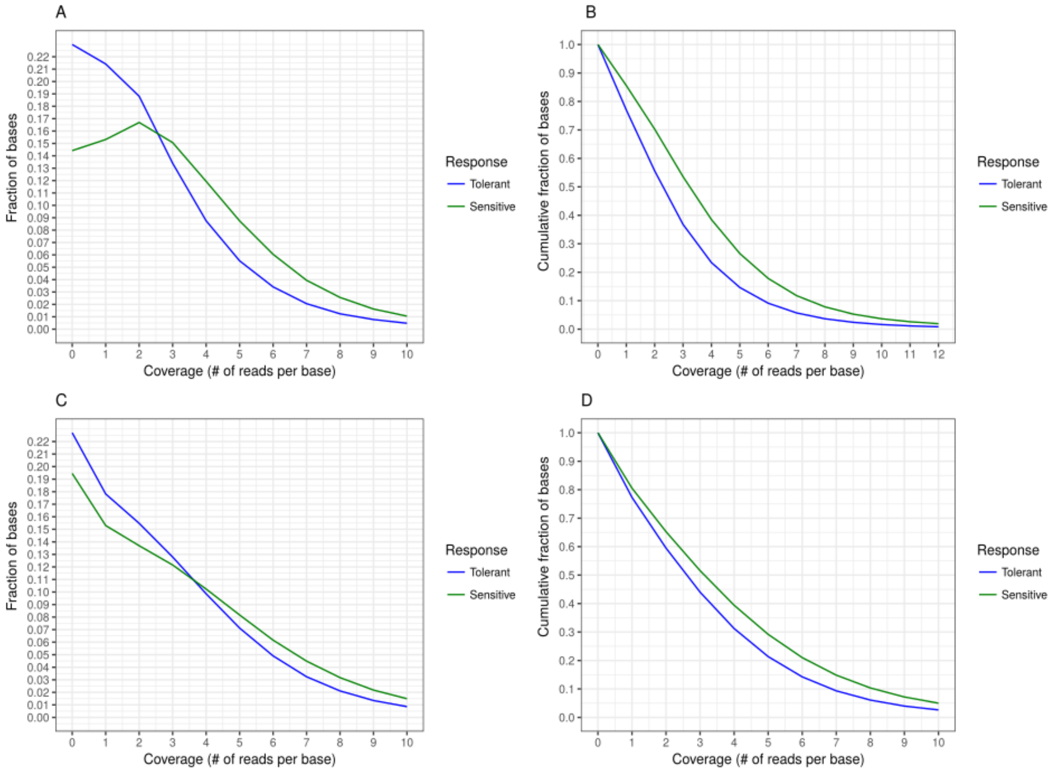

Figure 5.3 Distribution of genic and intergenic regions coverage with PacBio reads. (A) distribution of genic regions coverage, (B) cumulative distribution of genic regions coverage, (C) distribution of genic intergenic regions coverage and (D) cumulative distribution of intergenic regions coverage.

accordingly. Thus, the obtained PacBio data will allow identification of over a half of all potential structural variants between the lines.

5.2. Reads correction efficiency

To evaluate the performance of the available correction tools, I have followed the workflow presented in Figure (5.4). After selecting the organisms based on the genome complexity, I have generated both, Illumina pair-end and long PacBio reads using artificial read generator tools as described in the methods section. The MAF alignment file produced by PacBio artificial read generator software contained both, artificial PacBio reads and the original sequences of the corresponding regions. Correction of the erroneous artificial reads was done using different dedicated tools (LoRDEC, PBcR, Proovread, LSC and HALC) on different datasets derived from model organisms (human, rice, *Trypanosoma*, yeast and *E. coli*).

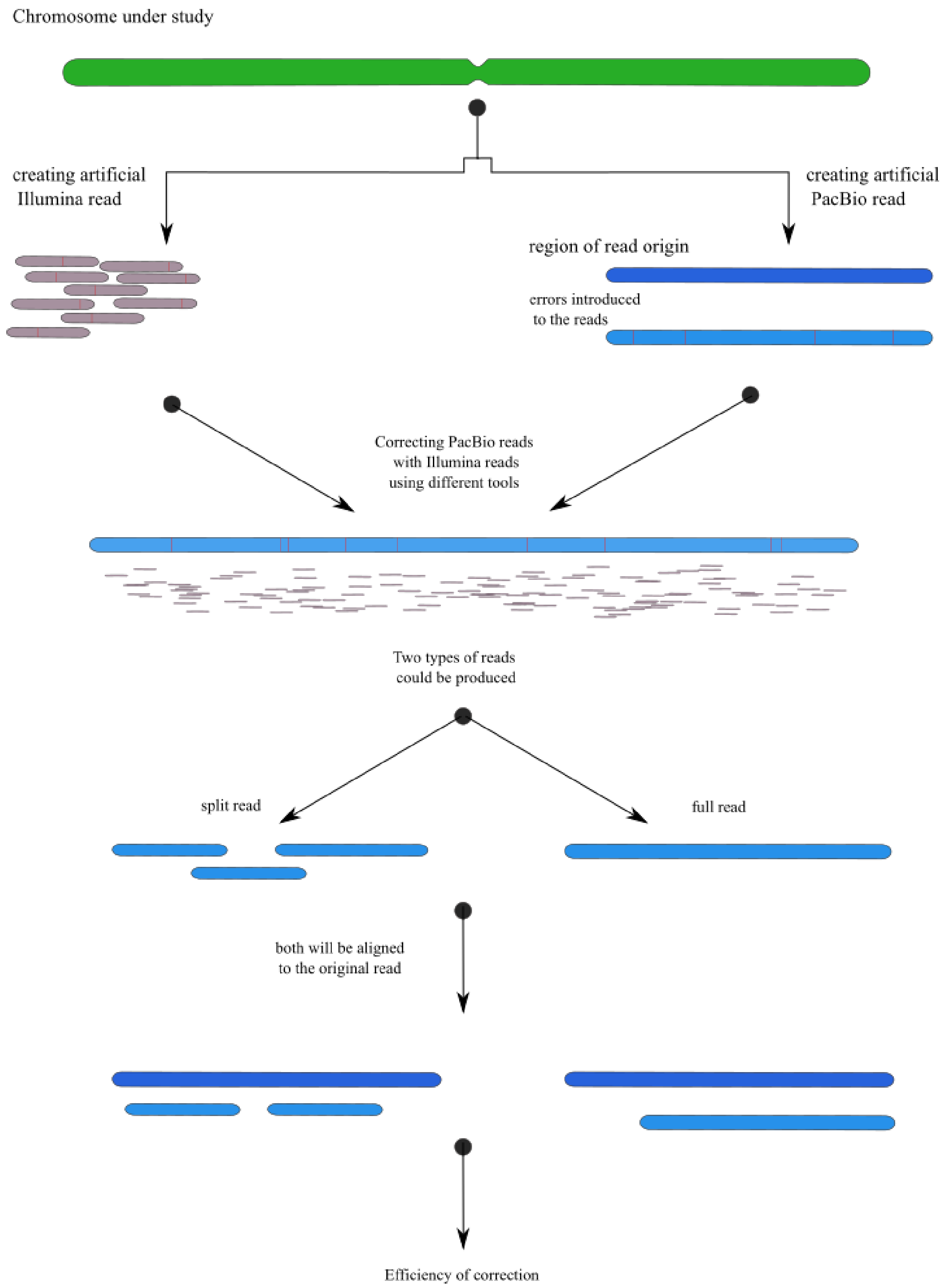
Chromosome under study
creating artificial Illumina read
creating artificial PacBio read
region of read origin
errors introduced to the reads
Correcting PacBio reads with Illumina reads using different tools
Two types of reads could be produced
split read
full read
both will be aligned to the original read
Efficiency of correction

Figure 5.4 Read correction assessment workflow. Erroneous artificial reads are generated, corrected and back-aligned against the original reads**.**

As the result of correction process, programs either return full-length corrected reads that contained some uncorrected regions, or trimmed and split reads where uncorrected regions has been removed. Proovread, LoRDEC and HALC are able to produce both types of corrected reads (full-length and trimmed reads). PBcR produces only trimmed reads and they were locally aligned. LSC produces full reads only. Uncorrected and corrected reads were compared with the original sequence of the region of origin (error-free read) extracted from the MAF file created by PacBio artificial read generation software. Trimmed reads were aligned locally using Smith-Waterman algorithm and untrimmed reads were aligned globally using Needleman-Wunsch algorithm. The increase in similarity between the uncorrected and corrected reads to the sequence of origin has been used as indicator of correction efficiency. Custom Python scripts were written to evaluate the correction efficiency of all correction tools (https://github.com/MeHelmy/AssessPy). Both types of corrected reads (full-length and trimmed) were evaluated separately. The major metrics of read quality that were investigated include read similarity to the sequence of origin, read length, number of corrected reads and bases

High accuracy trimmed reads

The first analyzed type of output from PacBio reads correction are trimmed high accuracy reads. The main aim here is to obtain dataset composed of reads with highest possible accuracy, which could be employed in applications requiring high confidence sequence information, as analysis of SNPs and haplotype identification. The major drawback is the substantial fragmentation of the PacBio reads caused by removal of low accuracy regions of PacBio reads, which usually reveal very low coverage with high accuracy Illumina reads. Thus, in the assessment of tools producing this kind of output I have focused mostly on examination of accuracy of corrected reads, their fragmentation and loss of data caused by removal of uncovered regions or highly error regions.

All tested tools, LoRDEC, Proovread, PBcR and HALC were found to produce highly accurate corrected reads with vast majority of them revealing accuracy > 99% (Table 5.3). From this point of view the best performing software was Proovread, which especially in complex genomes of rice and human was able to produce more accurate reads than other tools.

As expected, in case of all tools I have observed substantial fragmentation of the reads. It was revealed by increase of read number and decrease of read lengths after correction (Table 5.4, Figure 5.5) except in case of HALC, which have the highest average read length and also lowest read fragmentation compared to other tools. As mentioned previously, this is achieved by the price of lower accuracy of reads than in

Table 5.3 Summary of correction efficiency for reads reaching more than 99% or 100% identity to the source sequences in high accuracy output mode.

	LoRDEC		Proovread		PBcR		HALC	
Read identity	**99%**	**100%**	**99%**	**100%**	**99%**	**100%**	**99%**	**100%**
E.coli	98.42	95.26	98.99	88.38	97.66	74.31	96.88	90.13
Trypanosoma	93.93	87.21	96.85	86.11	95.58	69.64	91.30	77.00
Yeast	97.8	94.4	97.99	86.68	97.09	70.17	95.58	86.96
Rice	70.86	61.26	89.48	77.23	88.54	64.4	59.46	37.77
Human	73.07	62.14	90.2	78.27	86.41	62.54	58.32	31.17

Table 5.4 Summary of high accuracy trimmed read.

		E.coli	Trypanosoma	Yeast	Rice	Human
Input read count		30 364	34 620	10 198	24 3426	26 6254
Median read length		2 393	2 374	2 394	2 379	2 381
Average read length		3 003	2 986	3 004	2 987	2 995
PBcR	Output read count	79 633	94 887	28 818	46 7204	61 9752
	Median read length	377	398	410	327	350
	Mean read length	460	491	511	385	422
Proovread	Output read count	59 195	65 055	19 295	46 3115	51 1532
	Median read length	506	517	551	378	374
	Mean read length	692	721	763	499	493
LoRDEC	Output read count	86 537	96 682	29 006	66 4961	72 6469
	Median read length	515	503	524	447	409
	Mean read length	664	673	691	593	529
HALC	Output read count	30 981	37 594	10 393	35 9908	51 2152
	Median read length	2 179	2 000	2 179	1 344	897
	Mean read length	2 729	2 519	2 721	1 752	1 239

case of Proovread (Table 5.3) beside that HALC performed the best in case of larger genomes with around two-fold higher read lengths. In case of simple genomes (*E. coli*, *Trypanosoma*, yeast), all tested tools performed relatively similar with exception of PBcR which produced higher amounts of shorter reads when compared to Proovread, LoRDEC and HALC. The highest average and median read length, associated by lowest read number, was reached in case of HALC followed by Proovread (Table 5.4). In complex genomes (human, rice), the performance of HALC and LoRDEC excels Proovread. Both mean and average read lengths are higher (Table 5.4). Also, HALC exceed both Proovread and LoRDEC in read length distribution, when we take a closer look to read length returned by Proovread and LoRDEC, results shows relatively similar number of very short reads for both tools, but LoRDEC produce longer reads (length ~600-2000 nt) than Proovread (Figure 5.5).

Next, I have examined the number of low accuracy uncorrected bases removed by each program. Looking at the lengths of clipped regions from each corrected PacBio read it was found that in all cases HALC had clipped fewer bp than Proovread, LoRDEC and PBcR (Figure 5.6). PBcR was observed to remove much larger portions of the reads (>5000 nt), than other tools. This feature is also reflected in shorter output read lengths observed for this tool (Table 5.4).

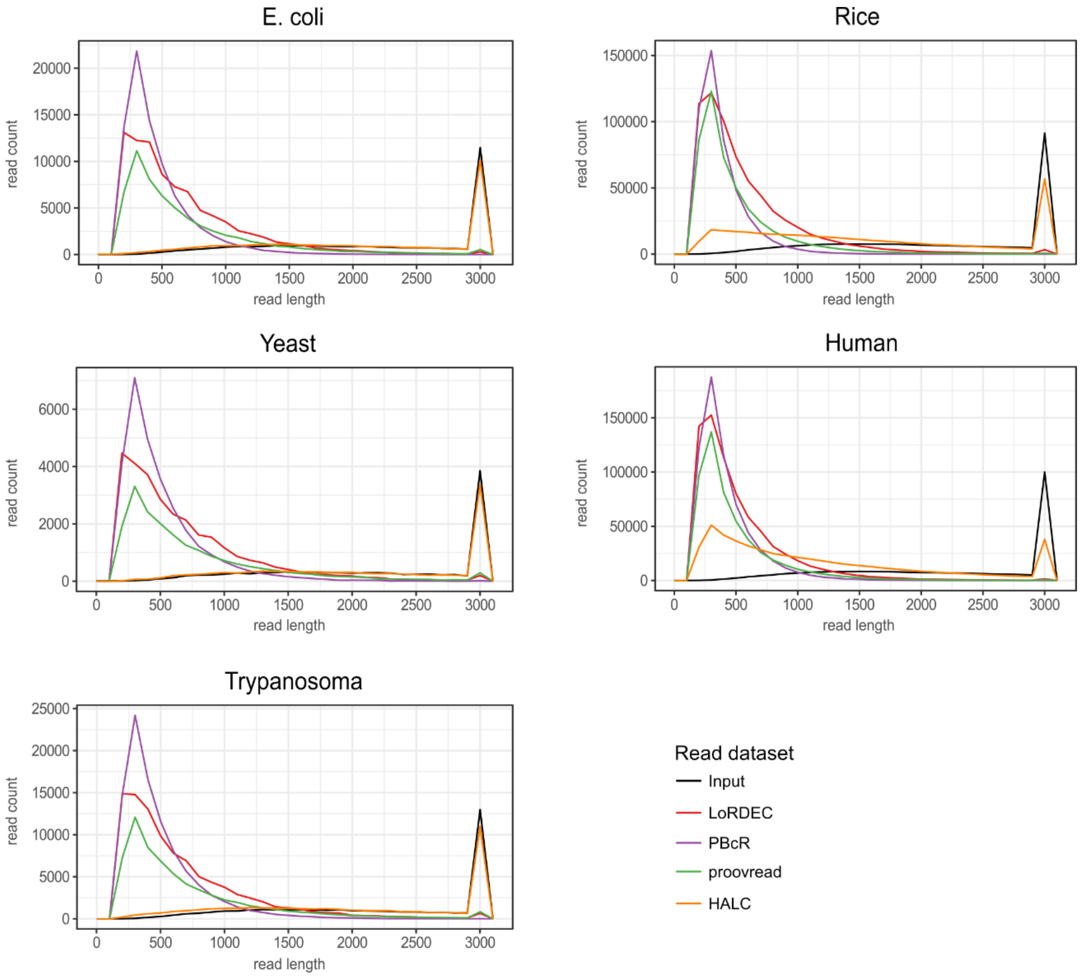

Figure 5.5 Read length distribution of high accuracy trimmed reads. The peak at 3000 represent reads longer than 3000 nt**.**

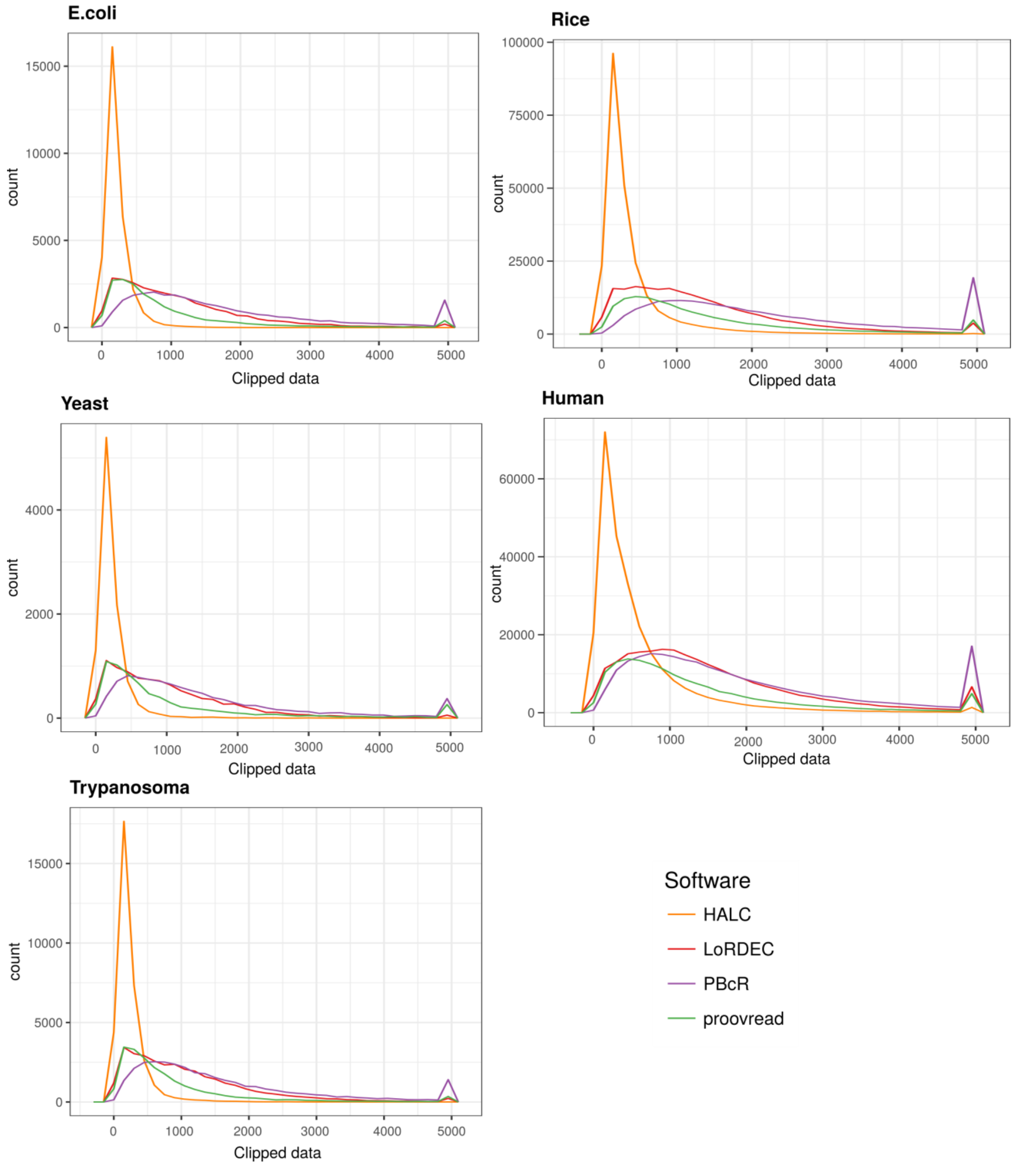

Figure 5.6 Distribution of amounts of low quality nucleotides removed from individual PacBio reads.

Since in the above considerations, only corrected reads were taken into account, they do not completely reflect the total amount of data lost during the correction process. To estimate this, I have calculated the number of corrected bases present in the reads produced by all tested algorithms. The results clearly indicate that in all organisms, HALC return significantly higher amounts of data than other tools (Table 5.5). In case of rice it is two folds more than Proovread and around one and half fold more than LoRDEC and around two folds more in human than both, the worst performance was in PBcR.

From above result, one could see that taking into consideration high accuracy trimmed reads output, PBcR performed worst in all considered measures and HALC is superior to all the used tools. Both remaining tools, Proovread and LoRDEC have relatively similar performance, depending on the downstream application of corrected reads and organism employed in the project. LoRDEC provide significantly higher amounts of corrected bases with similar accuracy as Proovread. In less complex genomes of *E.coli*, *Trypanosoma* and yeast LoRDEC provide reads of similar length distribution as Proovread, but in higher amounts. However, in complex genomes of rice and human, it outperforms Proovread tool in all aspects: number of bases, length and number of reads.

Proovread performed slightly better in terms of accuracy of corrected reads than rest of software excluding HALC. The difference was clearly visible for human dataset, where it outperformed other tools. On the other hand, in simple organisms it was able to produce slightly longer reads than LoRDEC, but in lower amounts. The

Table 5.5 The percentage of corrected bases output in high accuracy trimmed reads mode.

Organism	PBcR	Proovread	LoRDEC	HALC
E.coli	40.18	44.92	62.99	92.64
Trypanosma	45.00	45.36	62.75	91.21
Yeast	48.02	48.05	65.35	92.15
Rice	24.70	31.69	53.08	85.10
Human	32.65	31.54	47.11	76.98

number of obtained bases was in each case significantly lower than with LoRDEC.

To conclude, the best tool for correction of PacBio reads, when the aim is to obtain high accuracy trimmed reads, is HALC.

Full length reads

The high lengths of the reads are the major advantage of the PacBio sequencing. Full length corrected reads, even if they contain some of low accuracy regions, are very useful in scaffolding during the genome assembly in order to resolve repeat regions and properly place contigs. Form tested correction tools, 4 had a possibility to produce this kind of output: Proovread, LoRDEC, LSC and HALC. The major parameters which reflect the correction efficiency of full-length reads is their similarity to sequence of origin (Table 5.6).

In all cases, the highest accuracy of corrected reads was reached by HALC. In simple genomes, it was able to produce more than 75% of reads with accuracy >99%

Table 5.6 Summary of correction efficiency for reads reaching more than 99% or 100% identity to the source sequences in high accuracy output mode.

	LoRDEC		**Proovread**		**LSC**		**HALC**	
Read identity	**99%**	**100%**	**99%**	**100%**	**99%**	**100%**	**99%**	**100%**
E.coli	8.66	1.36	21.97	1.04	0.01	0	84.38	20.25
Trypanosoma	8.42	1.24	24.47	1.46	0.01	0	75.95	16.41
Yeast	10.51	1.44	31.37	2.19	0.01	0	82.54	19.75
Rice	2.52	0.38	5.06	0.23	0.01	0	40.31	7.23
Human	2.62	0.48	4.24	0.18	0.01	0	24.91	4.31

(Table 5.6). In complex ones, the value drooped to reach up to 40, but still significant up to 8 times more than the nearest achieved results by Proovread. The superiority of HALC was even clearer when analyzing the distribution of corrected reads accuracy (Figure 5.7). In all cases, the peak density of read accuracy was significantly better than of other tools reaching up to 95% in complex genome, and in simple organism, it was ~98%, whereas the nearest achieved results were by Proovread in complex ones, ~90% of accuracy and 95% in simple genomes.

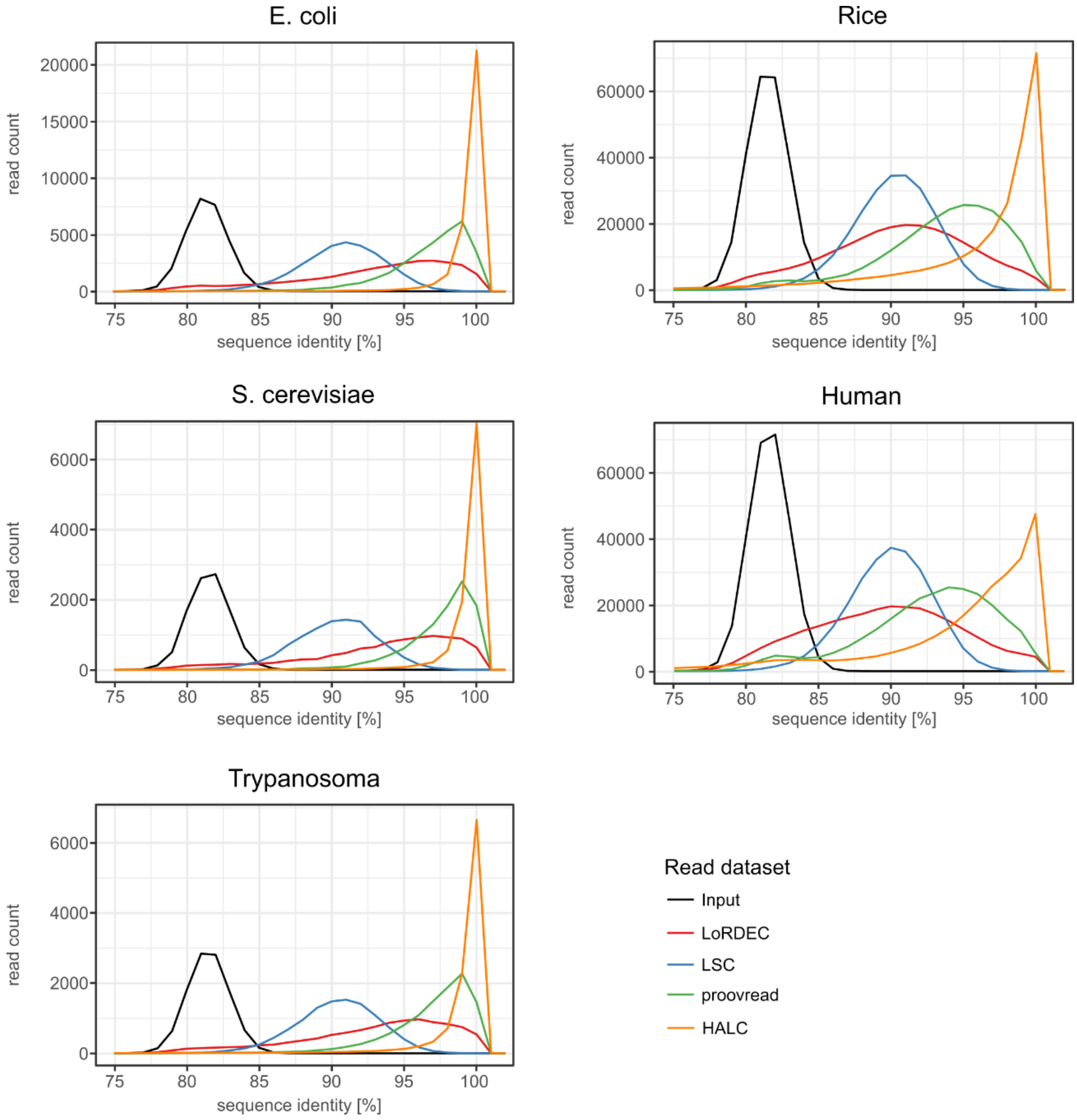

Figure 5.7 Distribution of read accuracy for full-length reads output compared to the accuracy of reads before correction.

Although in case of full-length read output, one would expect to obtain all of the input reads in the original full-length, I have observed the minor loss of data (Table 5.7). Only HALC and LoRDEC were able to return 100% of the input reads. Other tools seem to remove the minor fraction of PacBio reads which were not corrected at all. All tools remove also some of the bases from the ends of the reads, resulting in minor shortening of the corrected reads, when compared with input. The most aggressive in this behavior was Proovread, with shortening of median read length within the range of 30-60nt.

Table 5.7 Summary of full-length corrected reads.

		E.coli	Trypanosoma	Yeast	Rice	Human
Input read count		30 364	34 620	10 198	243 426	266 254
Median read length		2 393	2 374	2 394	2 379	2 381
Average read length		3 003	2 986	3 004	2 987	2 995
LSC	Output read count	30 317	34 560	10 183	242 969	265 723
	Median read length	2 302	2 292	2 312	2 294	2 299
	Mean read length	2 889	2 878	2 901	2 880	2 892
Proovread	Output read count	30 360	34 616	10 196	243 397	266 220
	Median read length	2 258	2 244	2 255	2 262	2 270
	Mean read length	2 834	2 819	2 831	2 845	2 858
LoRDEC	Output read count	30 364	34 620	10 198	243 426	266 254
	Median read length	2 295	2 279	2 292	2 293	2 302
	Mean read length	2 874	2 858	2 871	2 876	2 894
HALC	Output read count	30 364	34 620	10 198	243 426	266 254
	Median read length	2 240	2 230	2 245	2 237	2 252
	Mean read length	2 811	2 799	2 813	2 811	2 835

Factors influencing the correction efficiency

Obtained results suggested, that complexity of the genome influence the correction efficiency. Thus, in the next step, I have assessed whether any of genomic sequence characteristics could differentiate the efficiently corrected and uncorrected regions of the PacBio reads. I have examined the GC and repeat content, as well as complexity. Repeats were identified with RepeatMasker and GC content was calculated using in house Python script. As the measure of complexity, Local Composition Complexity (LCC) has been used. Identified features were overlaid with the positions of corrected read fragments and analyzed. Results indicate that there is no significant correlation between analyzed features and ability of software to correct given regions of the read (Figure 5.8). The highest difference between corrected and

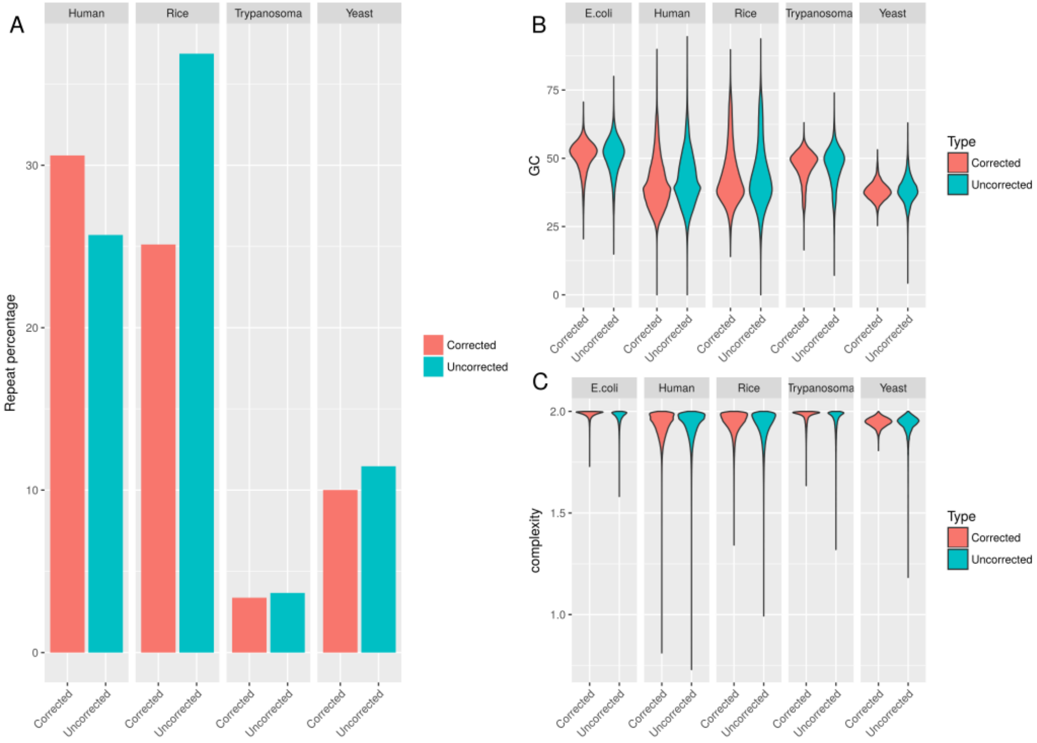

Figure 5.8 Dependence of correction efficiency on genome features. The differences in (A) repeat content, (B) GC content and (C) sequence complexity between corrected and uncorrected regions of PacBio reads has been analyzed.

uncorrected read regions was observed for repeat content in rice, however it reached only 10% of difference, thus it does not explain why the regions were not corrected.

On the other hand, I have also tested the performance of the read correction when whole PacBio dataset is divided into smaller chunks. For each subset, all Illumina reads has been employed. By this approach I was able to slightly boost the correction efficiency, however by cost of introduction of minor amount of additional errors. This observation suggests that the major reason for drop of correction efficiency in complex genomes is the mapping of the Illumina reads into multiple PacBio reads. When the PacBio dataset is reduced, more Illumina reads can be mapped fulfilling criteria of maximal number of mapping sites, thus the coverage of the PacBio reads increase. The drawback of this approach is introduction of additional errors, which are related to mapping of Illumina reads which in fact reflect other PacBio reads, not included in given chunk.

Computational performance

Computational performance of the tools was measured on workstation equipped with Ubuntu 16.04.1 operating system, Intel(R) Xeon(R) CPU E5-1660, 3.30GHz, x86_64, 12 CPUs and 32GB RAM. *E.coli* was used as a model to measure software performance (speed, CPU and RAM usage). The dependence of performance on dataset size was measured by dividing *E.coli* genome into four chunks of different size and performing correction on each of them. These chunks are: 100%, 75%, 50% and 25% of the genome. Computationally, LoRDEC exceeded the performance of the other software followed by Proovread. LSC comes at the last place, it the most consuming tool regarding CPU and memory usage (Table 5.8) beside that in case of HALC it demands to run another tool (SOAPdenovo2) because it depends on its output contigs to run.

Table 5.8 A comparison of the computational performance of the correction tools.

Software	User time [s]	System time [s]	CPU load	Total time	Memory [MB]
Dataset size: 100%					
Proovread	49302.87	6811.14	676%	2:18:11.62	4803.53
LoRDEC	2943.50	532.17	790%	7:19.84	450.98
LSC	58944.97	195.09	934%	1:45:27.54	1156.42
PBcR	20538.56	1140.17	383%	1:34:11.75	1995.64
HALC	4665.50	12.20	857%	9:05.74	2774.58
Dataset size: 75%					
Proovread	36491.06	4800.70	634%	1:48:29.04	4153.96
LoRDEC	1892.43	291.01	619%	5:52.18	428.29
LSC	44517.65	119.67	881%	1:24:22.33	1193.95
PBcR	13176.43	709.95	374%	1:01:46.37	1570.87
HALC	3746.86	9.51	845%	7:24.05	2223.95
Dataset size: 50%					
Proovread	23095.22	3130.62	621%	1:10:19.56	2823.46
LoRDEC	1159.71	258.07	658%	3:35.17	407.08
LSC	25472.81	85.09	902%	47:13.07	1079.21
PBcR	9525.22	625.45	342%	49:24.63	1321.88
HALC	2815.30	6.67	841%	5:35.34	1262.47
Dataset size: 25%					
Proovread	10920.95	1588.77	615%	33:51.72	1396.52
LoRDEC	576.78	84.85	630%	1:44.87	377.20
LSC	13327.00	52.06	891%	25:01.43	1059.08
PBcR	4419.74	302.41	335%	23:28.5	1159.16
HALC	1695.24	3.34	803%	3:31.48	651.01

Conclusions on PacBio read correction tools

Presented result show that decision of the tool employed for correction of PacBio reads is crucial for quality of the resulting dataset. In order to obtain the high accuracy corrected reads, one have to consider the substantial loss of data, accompanied by severe fragmentation of the reads. Since most of the obtained reads are within the range of length typical for Illumina reads, thus the employment of such approach is questionable in terms of cost-efficiency. On the other hand, the correction aimed to obtain the full-length corrected reads result in substantial increase of read accuracy, especially in simple genomes, without drawbacks related to read fragmentation. Thus, this kind of output is well suited for employment in genome assembly.

For trimmed reads, the best overall performing tool was HALC, Followed by LoRDEC as confirmed also by (Abdel-Ghany et al., 2016). The highest possible accuracy of reads can be reached by using Proovread, however with substantially higher data loss. Therefore, it is a tradeoff between more accurate corrected reads using Proovread or longer average read length and high throughput by HALC.

When aiming for full-length corrected reads, HALC clearly outperforms other tools. In simple organism it provides very high accuracy of corrected reads and in complex genomes also is the best one.

HALC, LoRDEC and Proovread could work on any machine, but LoRDEC is much more efficient computationally. This is achieved by the usage of De Bruijn Graphs (DBG) during the correction, which could be stored and used with other tools to assembly the short reads alone. No other tested software provides similar functionality. On the other hand, Proovread provide the functionality of identification of chimeric reads. All the functionalities of tested tools are compared in (Table 5.9).

Based on above, there is a space for more tools to be developed, which would be less sensitive to multiple mapping of the short reads. Due to this limitation, all the existing tools lose a tremendous amount of data as shown from the results, beside that either they lack speed or accuracy, sometime both.

Table 5.9 A summary of the correction software features,

	Proovread	**LoRDEC**	**PBcR**	**LSC**	**HALC**
Input FASTA or FASTQ	✓	✓	✓	✓	✓ (contig)
output FASTQ	✓	○	✓	✓	○
produce full reads	✓	✓	○	✓	✓
produce trimmed reads	✓	✓	✓	○	✓
accept other inputs	✓	○	○	○	○
output DBG from SR	○	✓	○	○	○
CPU and memory efficiency	✓✓✓	✓✓✓✓	✓✓	✓	✓✓✓
detection of chimeric reads	✓	○	○	○	○
location of corrected read part	✓	○	○	○	○

5.3. Identification of genetic variation between maize lines

The major aim of the work was to identify the genetic variants between maize lines, which substantially differ in tolerance to glyphosate. In my work, I have employed two strategies: first, I have used the long corrected PacBio reads to find large structural variations, and then I have employed high accuracy Illumina reads to detect SNPs and small indels.

Identification of structural variants

For detection of large structural variants (SVs), I took the advantage of long PacBio reads, which are able to cover substantial chromosomal regions and directly reveal large deletions, insertion, duplications and inversions. In the first step, I have compared the sensitivity of variant detection, when using corrected and uncorrected reads. For this purpose, PBSuite tools were used. Both corrected and uncorrected subreads were aligned to the reference genome using blasr as mentioned previously in method section. Tails and Spots algorithms from PBSuite were used to identify SVs based on interrupted mapping and read discordance respectively, but first PBSuite pie algorithm was used to identify unaligned tails for reads then the result was passed to both Tails and Spots to identify SVs.

Next, I investigated the overlap between corrected and uncorrected subreads identified by PBSuite, by comparing overlap between the SVs resulting from employment of corrected and uncorrected reads. There were a very few variants reported by both approaches with exactly the same position boundaries (81 in tolerant and 102 in sensitive line), but when analyzing any size of overlap, there was a significant number of variants identified by both approaches (Figure 5.9). In general, in each case the number of reported deletions was significantly higher than of insertions. The employment of corrected reads doubled the number of identified variants. Since the number of variants detected with uncorrected reads was usually much smaller compared to those detected with corrected reads, I concluded that employment of read correction increase sensitivity of variant detection.

To have a closer insight into differences between the approaches, I have analyzed the distributions of overlap size among the variants detected by both (Figure 5.11). The overlap has been expressed as a percent of variant length. Vast majority of deletions were detected by both approaches with very similar boundary positions, in

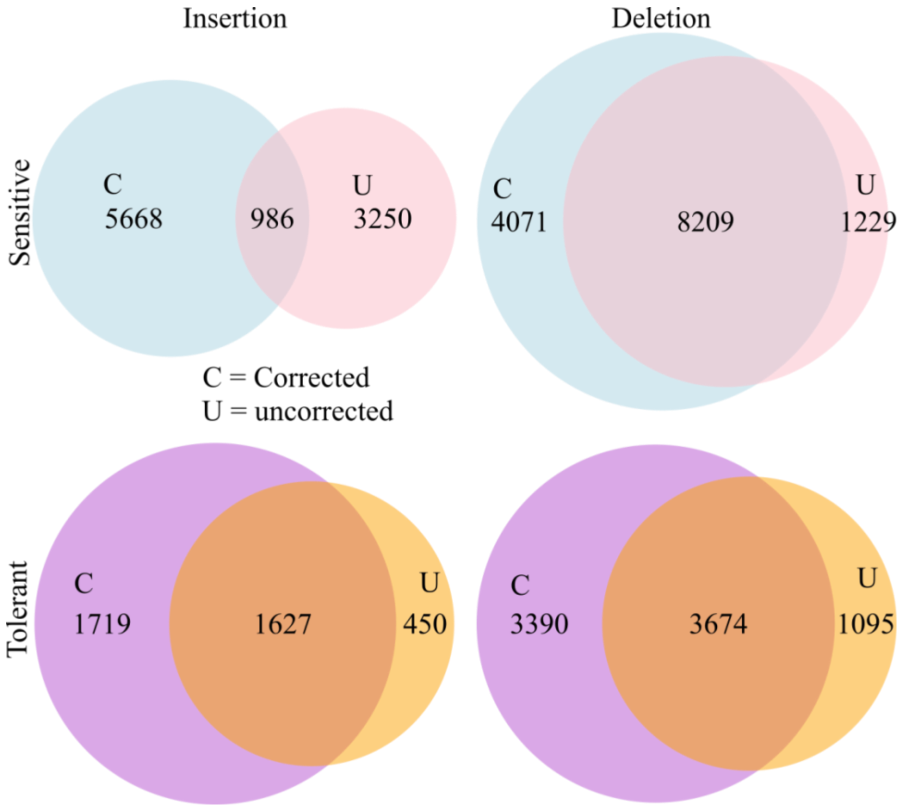

Figure 5.9 Comparison of amounts of structural variants identified by PBSuite using either corrected or uncorrected PacBio reads.

contrast to insertions, where usually concordance of variant position was below 50%. This observation could be explained by methodology of insertion detection, where inserted fragment is determined by its disagreement with reference sequence. In case of uncorrected reads, this process is hampered by local accumulation of sequencing errors.

In order to prove the above observations, I have performed the variant detection using uncorrected reads with another tool, Sniffles, which was especially designed to work on uncorrected PacBio data (Figure 5.10). Although the number of detected variants was much higher than in case of PBSuite, the concordance with variants detected with corrected reads was similar or weaker (case of insertions in

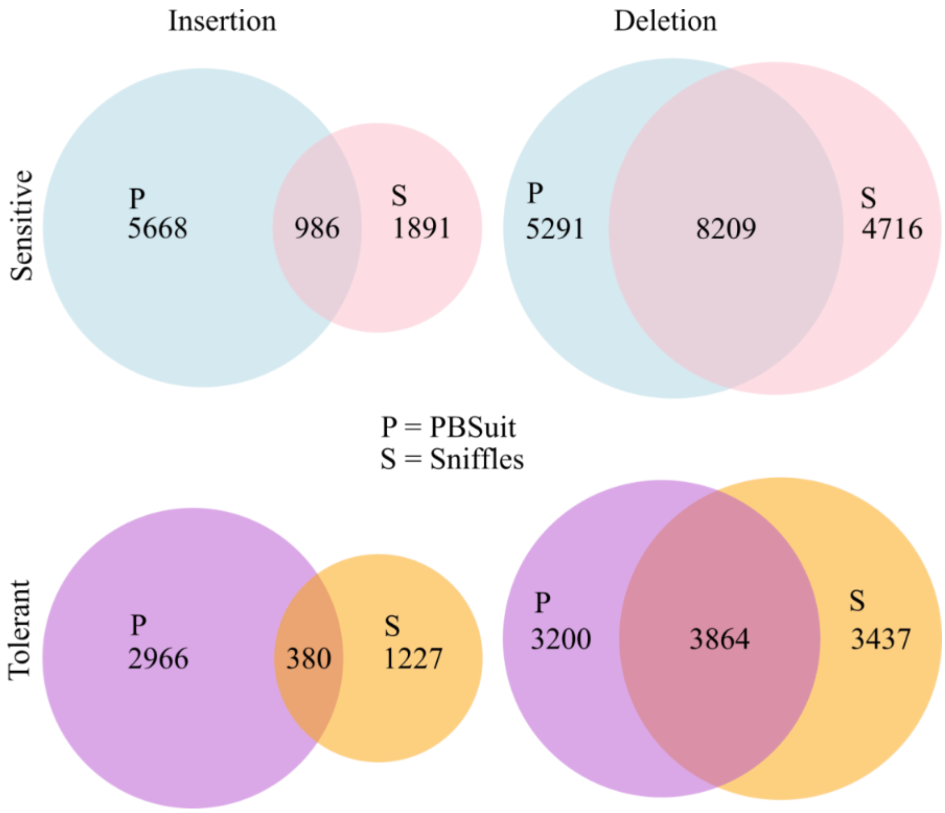

Fig 5.10 Comparison of amounts of structural variants identified using either PBSuite with corrected or Sniffles with uncorrected PacBio reads.

tolerant line). Also, the overlap of variants detected by both methods was smaller than before (Figure 5.11). The increased number of discordant variants identified with Sniffles was probably the result of employment of poorly aligned regions of PacBio reads by Sniffles algorithm. However, the lack of their confirmation by corrected reads suggest that most of them are rather derived from misaligned reads, which after correction can be efficiently aligned to other genomic loci.

Thus, based on presented results, I have decided to use in further analysis the variants detected using PBSuite with reads corrected by HALC.

After identifying the SVs in both lines, I have analyzed the distribution of read coverage supporting individual variants (Figure 5.12). Almost all insertions (~98%) were supported by at least two reads, whereas in case of deletions it was only 76-78% (Figure 5.12A). The coverage of all types of variants on both lines revealed good saturation by reads with peak coverage of 3 for deletions and 4 for insertions (Figure 5.12B).

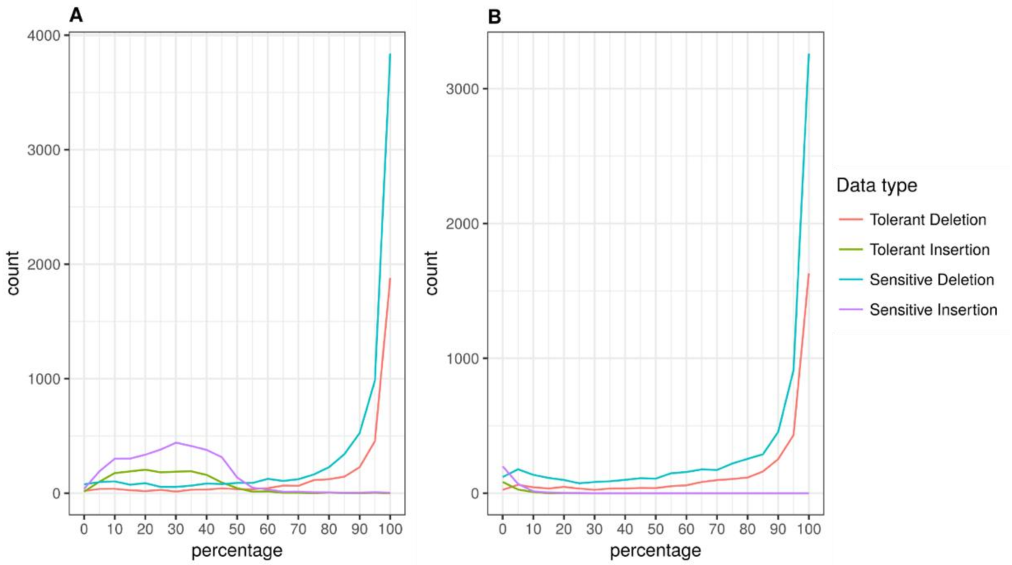

Figure 5.11 Distribution of overlap between positions of identified variants. (A) Results from PBSuite using corrected and uncorrected reads. (B) Results from PBSuite using corrected and Sniffles using uncorrected reads.

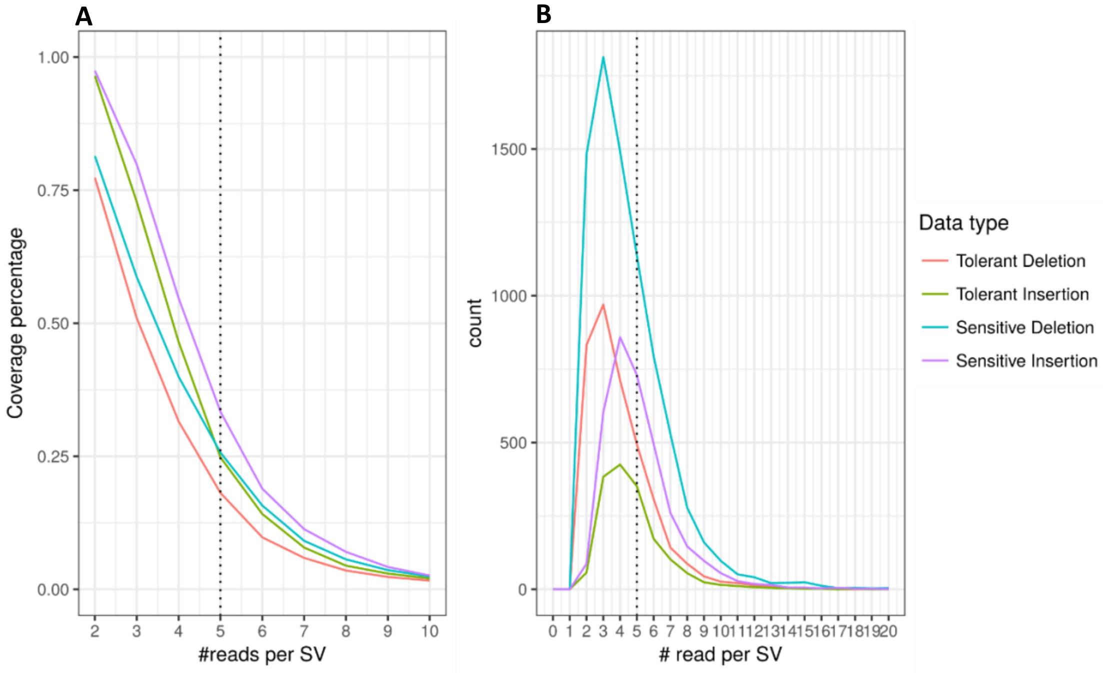

Figure 5.12 Coverage of structural variants by PacBio reads. (A) Cumulative distribution of read coverage. (B) Distribution of supporting read count for structural variants. Dotted line represents coverage cutoff applied for downstream analysis.

In the next step, deletions and insertions from both lines were separately intersected using mergeSVcallers (https://github.com/zeeev/mergeSVcallers). For further analysis, all variants identified in both lines were removed, resulting in a set of line-specific SVs. In comparison of two genomes, the deletion in one of them is equivalent with the same insertion in the other. Since SVs has been detected in both analyzed maize lines separately based on alignment to B73 reference maize genome, in the next step I have combined the results to reveal the direct differences between analyzed lines. Thus, deletions from tolerant line and insertions from sensitive line were pulled together, resulting in a set of deletions observed in tolerant line in comparison to sensitive line. Similarly, insertions from tolerant line have been pulled with deletions from sensitive line resulting in set on insertions observed in tolerant line. Based on (Fang, Hu, Wang, & Wang, 2016), to achive optimium accuracy for SVs detection, in further analysis we have used only variants supported by 5 or more reads.

As the result, total of 11 172 structural variants were identified representing 6062 insertions and 5110 deletions. The sets of identified SVs were functionally annotated using Ensembl variant effect predictor (VEP) to predict the result of the variant on gene expression (Figure 5.13). Vast majority of the variants were found to influence noncoding parts of the genome (upstream/downstream regions of the genes, intronic and intergenic regions). Thus, to focus on variants with high possibility of phenotype manifestation, for further analysis I have selected only those predicted to have high impact, whereas variants with low, moderate or modifier impact has been rejected. When analyzing such narrowed set of variants, the most frequent effect of

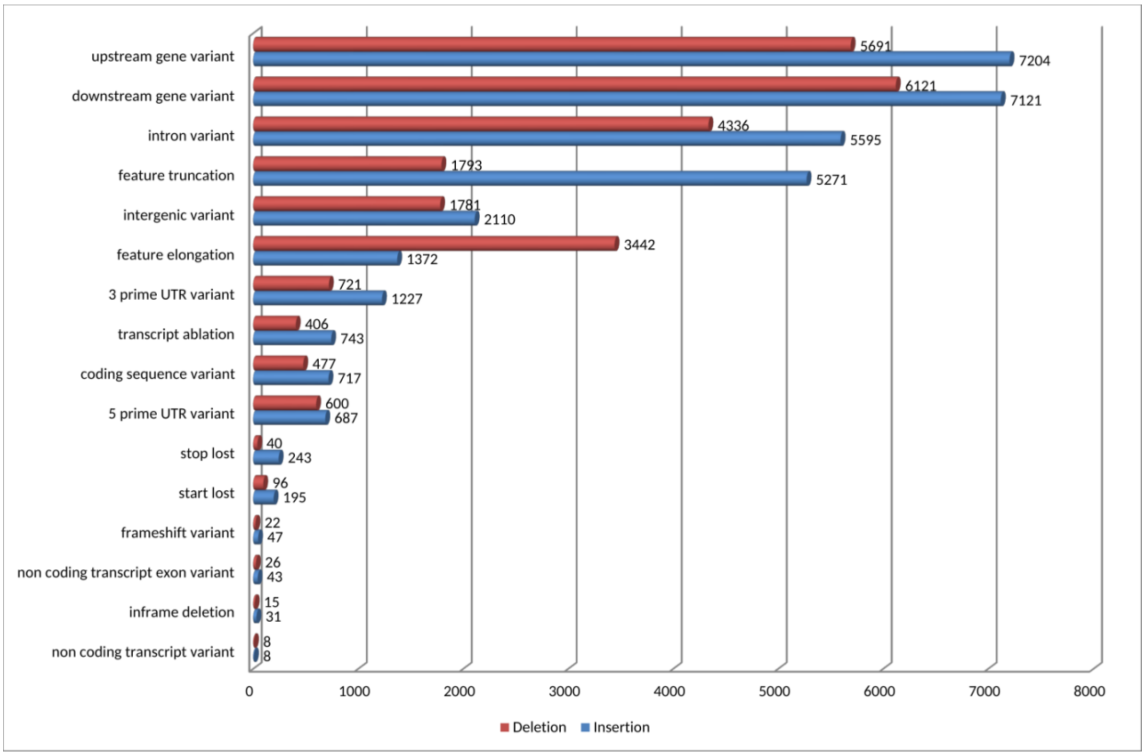

Figure 5.13 Predicted consequences of identified structural variants.

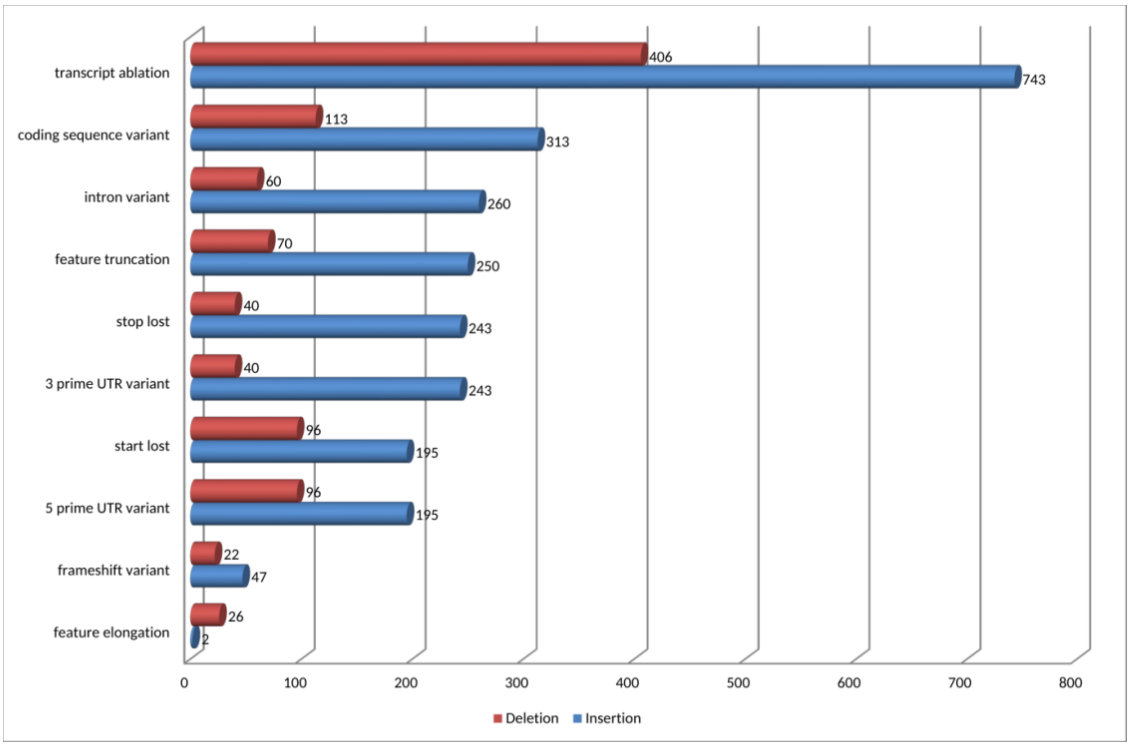

Figure 5.14 Predicted high impact consequences of identified structural variants.

the variants was the transcript ablation (Figure 5.14). Other predicted effects, like coding sequence variant, feature truncation or start/stop lost were less frequent. At the same time only few cases of feature elongation or frameshift were identified.

Next, I have analyzed the distribution of gene ontology terms associated with genes influenced by high impact structural variants. The obtained list of GO terms was summarized using REViGO, as explained in the methods section (Figure 5.15, 5.16). Among GO terms associated with genes affected by deletions, the most abundant in molecular function division were iron ion binding, methyltransferase activity, transferase activity and RNA binding. In biological process annotation, the most abundant GO terms were glycogen biosynthesis, protein K48-linked deubiquitination, carbohydrate metabolism and metal ion transport.

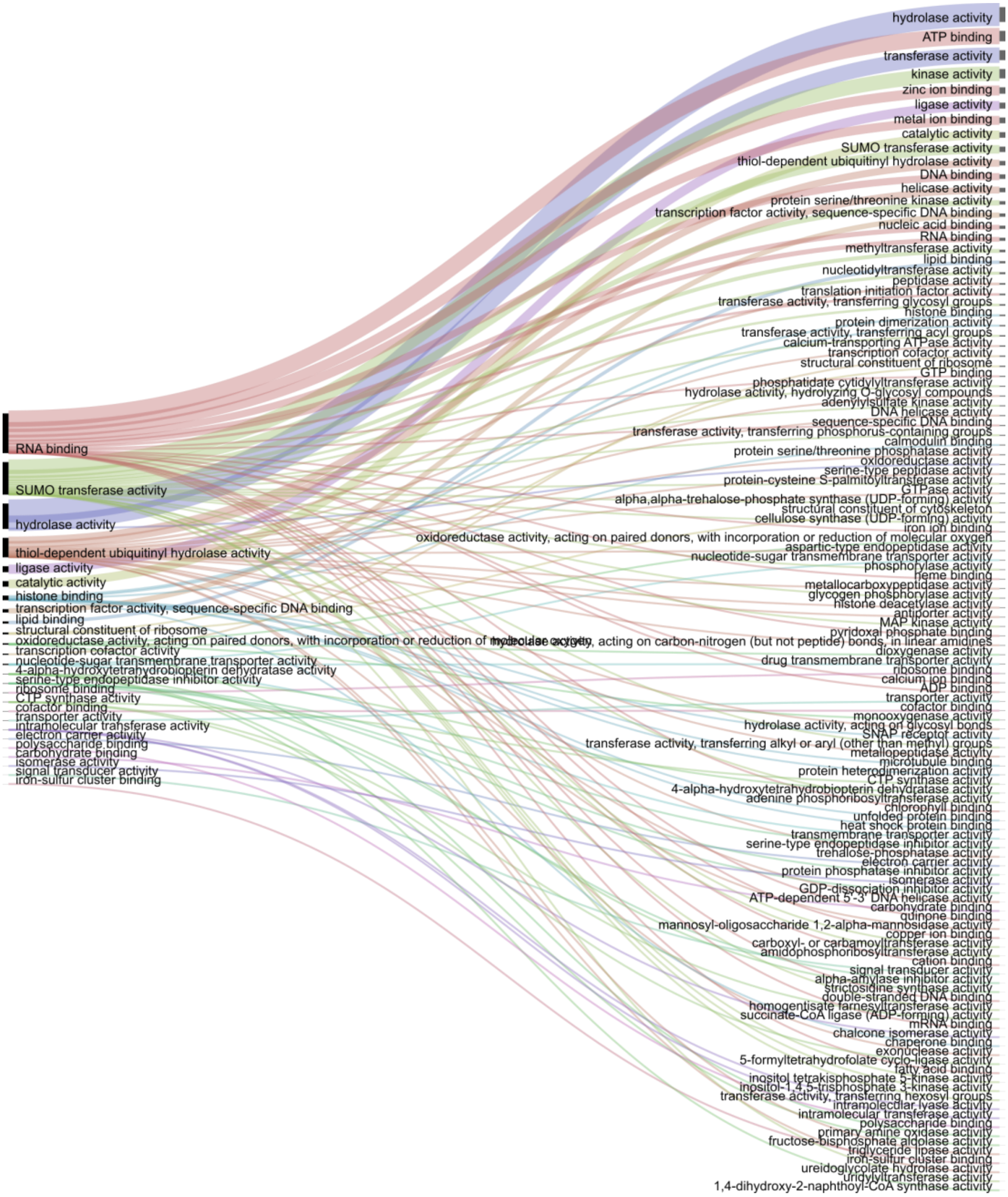
RNA binding
SUMO transferase activity
hydrolase activity
thiol-dependent ubiquitinyl hydrolase activity
ligase activity
catalytic activity
histone binding
transcription factor activity, sequence-specific DNA binding
lipid binding
structural constituent of ribosome
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
transcription cofactor activity
nucleotide-sugar transmembrane transporter activity
4-alpha-hydroxytetrahydrobiopterin dehydratase activity
serine-type endopeptidase inhibitor activity
ribosome binding
CTP synthase activity
cofactor binding
transporter activity
intramolecular transferase activity
electron carrier activity
polysaccharide binding
carbohydrate binding
isomerase activity
signal transducer activity
iron-sulfur cluster binding
hydrolase activity
ATP binding
transferase activity
kinase activity
zinc ion binding
ligase activity
metal ion binding
catalytic activity
SUMO transferase activity
thiol-dependent ubiquitinyl hydrolase activity
DNA binding
helicase activity
protein serine/threonine kinase activity
transcription factor activity, sequence-specific DNA binding
nucleic acid binding
RNA binding
methyltransferase activity
lipid binding
nucleotidyltransferase activity
peptidase activity
translation initiation factor activity
transferase activity, transferring glycosyl groups
histone binding
protein dimerization activity
transferase activity, transferring acyl groups
calcium-transporting ATPase activity
transcription cofactor activity
structural constituent of ribosome
GTP binding
phosphatidate cytidylyltransferase activity
hydrolase activity, hydrolyzing O-glycosyl compounds
adenylylsulfate kinase activity
DNA helicase activity
sequence-specific DNA binding
transferase activity, transferring phosphorus-containing groups
calmodulin binding
protein serine/threonine phosphatase activity
oxidoreductase activity
serine-type peptidase activity
protein-cysteine S-palmitoyltransferase activity
GTPase activity
alpha,alpha-trehalose-phosphate synthase (UDP-forming) activity
structural constituent of cytoskeleton
cellulose synthase (UDP-forming) activity
iron ion binding
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
aspartic-type endopeptidase activity
nucleotide-sugar transmembrane transporter activity
phosphorylase activity
heme binding
metallocarboxypeptidase activity
glycogen phosphorylase activity
histone deacetylase activity
antiporter activity
MAP kinase activity
pyridoxal phosphate binding
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
dioxygenase activity
drug transmembrane transporter activity
ribosome binding
calcium ion binding
ADP binding
transporter activity
cofactor binding
monooxygenase activity
hydrolase activity, acting on glycosyl bonds
SNAP receptor activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
metallopeptidase activity
microtubule binding
protein heterodimerization activity
CTP synthase activity
4-alpha-hydroxytetrahydrobiopterin dehydratase activity
adenine phosphoribosyltransferase activity
chlorophyll binding
unfolded protein binding
heat shock protein binding
transmembrane transporter activity
serine-type endopeptidase inhibitor activity
trehalose-phosphatase activity
electron carrier activity
protein phosphatase inhibitor activity
isomerase activity
GDP-dissociation inhibitor activity
ATP-dependent 5'-3' DNA helicase activity
carbohydrate binding
quinone binding
mannosyl-oligosaccharide 1,2-alpha-mannosidase activity
copper ion binding
carboxyl- or carbamoyltransferase activity
amidophosphoribosyltransferase activity
cation binding
signal transducer activity
alpha-amylase inhibitor activity
strictosidine synthase activity
double-stranded DNA binding
homogentisate farnesyltransferase activity
succinate-CoA ligase (ADP-forming) activity
mRNA binding
chalcone isomerase activity
chaperone binding
exonuclease activity
5-formyltetrahydrofolate cyclo-ligase activity
fatty acid binding
inositol tetrakisphosphate 5-kinase activity
inositol-1,4,5-trisphosphate 3-kinase activity
transferase activity, transferring hexosyl groups
intramolecular lyase activity
intramolecular transferase activity
polysaccharide binding
primary amine oxidase activity
fructose-bisphosphate aldolase activity
triglyceride lipase activity
iron-sulfur cluster binding
ureidoglycolate hydrolase activity
uridylyltransferase activity
1,4-dihydroxy-2-naphthoyl-CoA synthase activity

Figure 5.15 Tree view of molecular function gene ontology terms associated with genes affected by deletion structural variants.

Iron ion binding GO term in maize is associated with a response to any process that results in a change in state or activity of a cell or an organism (in terms of secretion, enzyme production, gene expression). That includes also stress response to

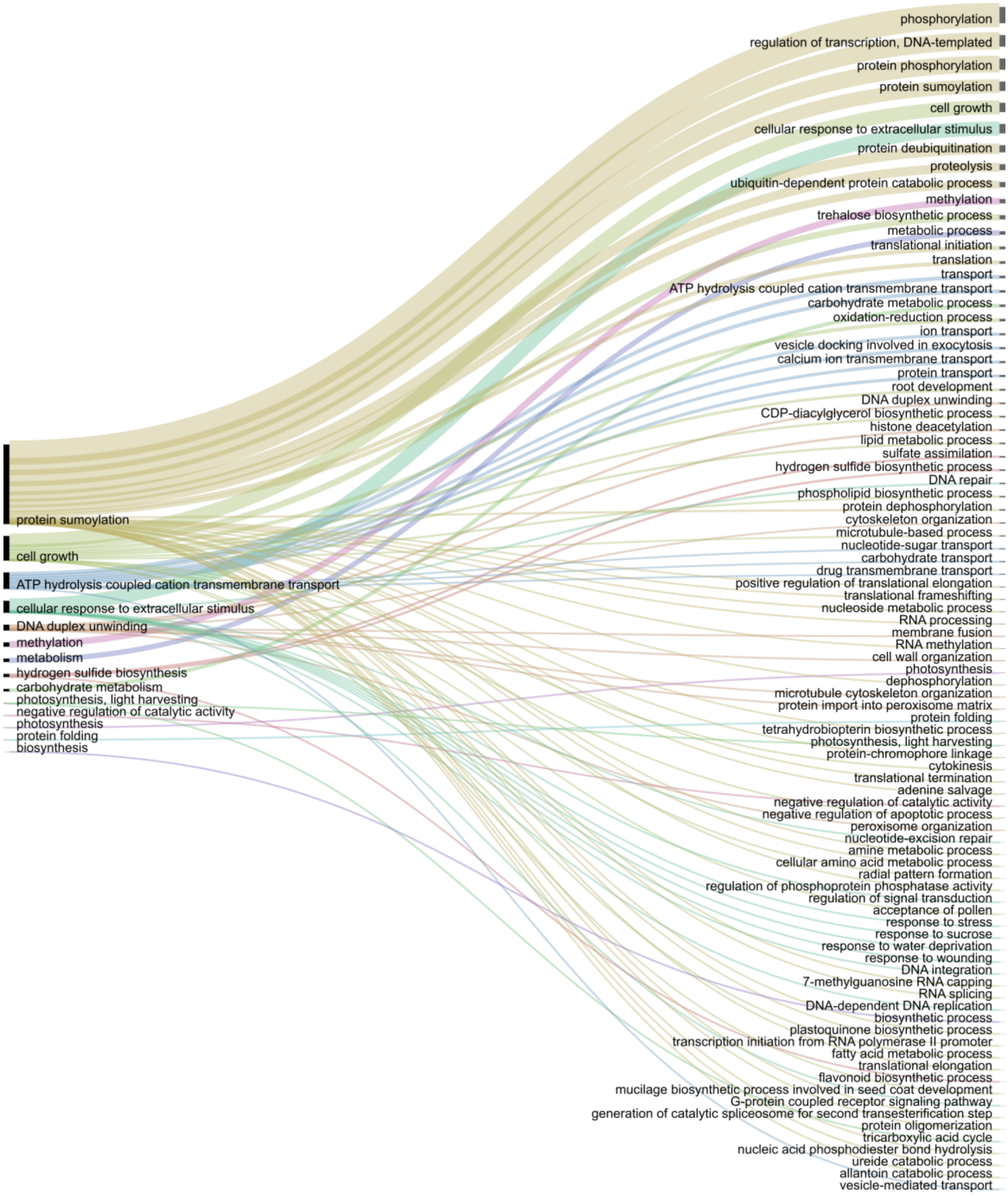

Figure 5.16 Tree view of biological process gene ontology terms associated with genes affected by deletion structural variants.

cold, fungus, wound and others. From the genes identified to be affected by deletions, to this category belongs Zm00001d037385, which encodes Adenine nucleotide alpha hydrolase-like superfamily protein, known to be associated with response to stress.

There were also other genes that were annotated to have response to different kinds of stresses like heat, nematode and oxidative stress. Those genes are Zm00001d052001, Zm00001d010838 and Zm00001d002426.

Among the GO terms of genes affected by identified insertions, the most abundant GO terms in molecular function were RNA binding, SUMO transferase

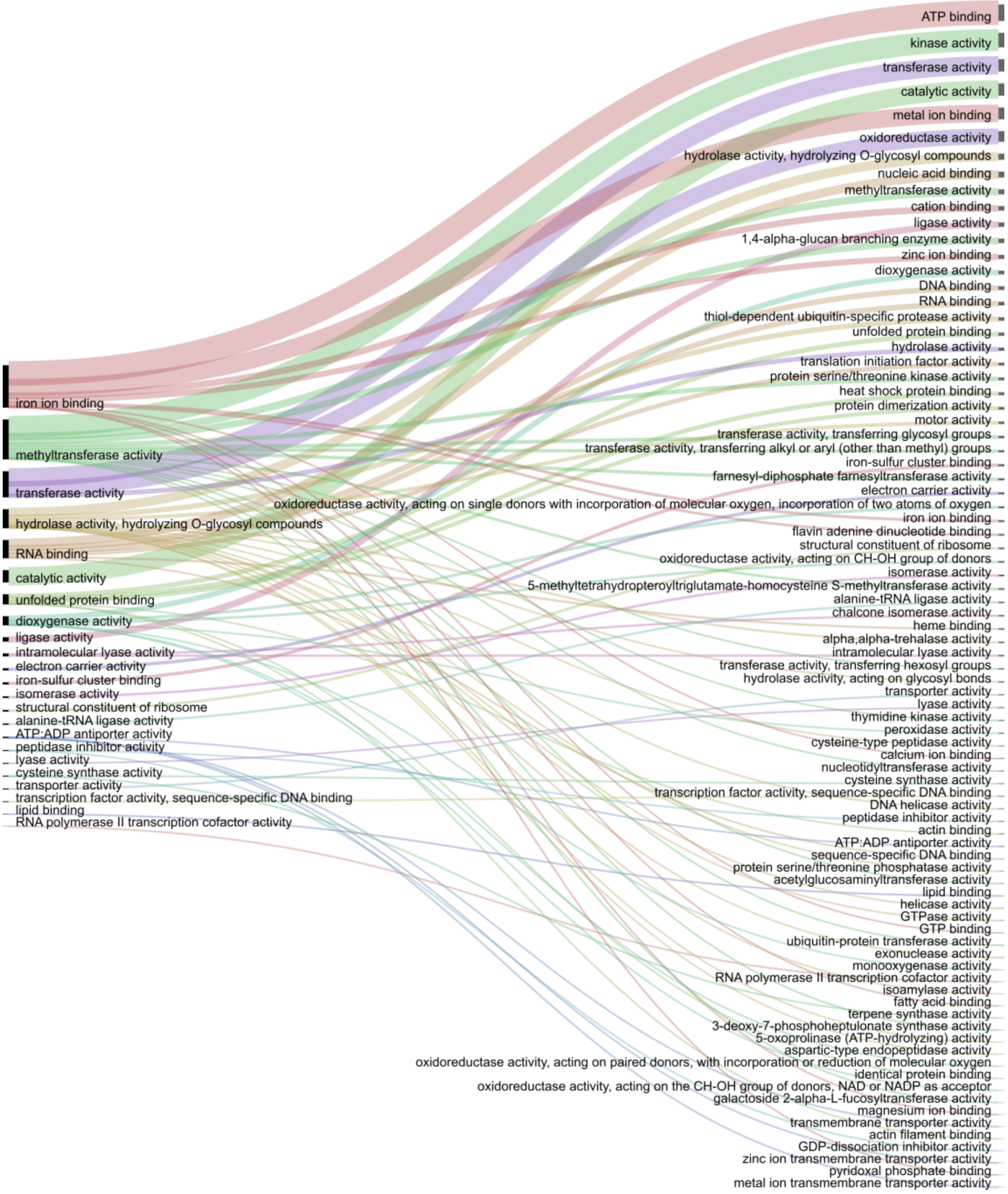

Figure 5.17 Tree view of biological process gene ontology terms associated with genes affected by insertion structural variants.

activity and hydrolase activity, whereas in biological process - protein sumoylation, cell growth, ATP hydrolysis and cellular response to extracellular stimulus (Figure 5.17, 5.18).

I have identified also a number of genes that were involved in response to external stimulus like Zm00001d010974, Zm00001d010974, Zm00001d042842 and

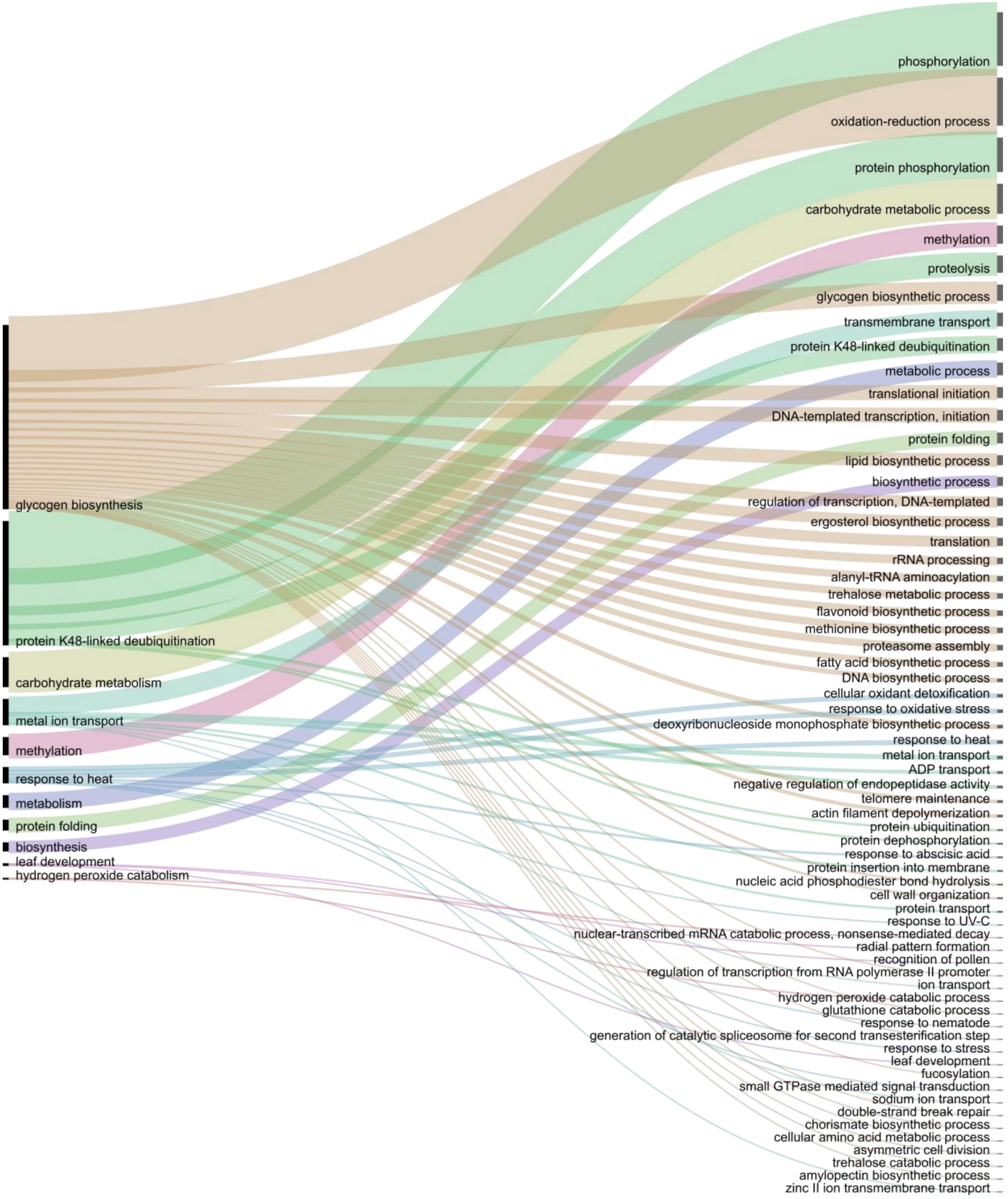
glycogen biosynthesis
protein K48-linked deubiquitination
carbohydrate metabolism
metal ion transport
methylation
response to heat
metabolism
protein folding
biosynthesis
leaf development
hydrogen peroxide catabolism
phosphorylation
oxidation-reduction process
protein phosphorylation
carbohydrate metabolic process
methylation
proteolysis
glycogen biosynthetic process
transmembrane transport
protein K48-linked deubiquitination
metabolic process
translational initiation
DNA-templated transcription, initiation
protein folding
lipid biosynthetic process
biosynthetic process
regulation of transcription, DNA-templated
ergosterol biosynthetic process
translation
rRNA processing
alanyl-tRNA aminoacylation
trehalose metabolic process
flavonoid biosynthetic process
methionine biosynthetic process
proteasome assembly
fatty acid biosynthetic process
DNA biosynthetic process
cellular oxidant detoxification
response to oxidative stress
deoxyribonucleoside monophosphate biosynthetic process
response to heat
metal ion transport
ADP transport
negative regulation of endopeptidase activity
telomere maintenance
actin filament depolymerization
protein ubiquitination
protein dephosphorylation
response to abscisic acid
protein insertion into membrane
nucleic acid phosphodiester bond hydrolysis
cell wall organization
protein transport
response to UV-C
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
radial pattern formation
recognition of pollen
regulation of transcription from RNA polymerase II promoter
ion transport
hydrogen peroxide catabolic process
glutathione catabolic process
response to nematode
generation of catalytic spliceosome for second transesterification step
response to stress
leaf development
fucosylation
small GTPase mediated signal transduction
sodium ion transport
double-strand break repair
chorismate biosynthetic process
cellular amino acid metabolic process
asymmetric cell division
trehalose catabolic process
amylopectin biosynthetic process
zinc II ion transmembrane transport

Figure 5.18 Tree view of biological process gene ontology terms associated with genes affected by insertion structural variants.

two MYB transcription factors Zm00001d029963 and Zm00001d037836. MYB is large family of regulatory proteins involved in controlling various processes like responses to biotic and abiotic stresses, development, differentiation, metabolism and defense.

SNPs and indels

To identify small changes between the analyzed maize lines I have employed the high accuracy data from illumina sequencing. The workflow was based on alignment of each read dataset to reference genome of maize B73 line obtained from Encode database followed by identification of variations between each line and reference. Since I have called SNPs and indels with GATK in cohorts' mode, only SNPs differentiating both lines were reported (Figure 5.19). This resulted in a huge number of identified variants (Table 5.10). In order to reduce it, based on the knowledge that the reference B73 line is glyphosate-sensitive, I have extracted only variants which were specific to tolerant line (Figure 5.19). Such approach resulted in a comprehensive list of genetic variations observed exclusively in glyphosate-tolerant line.

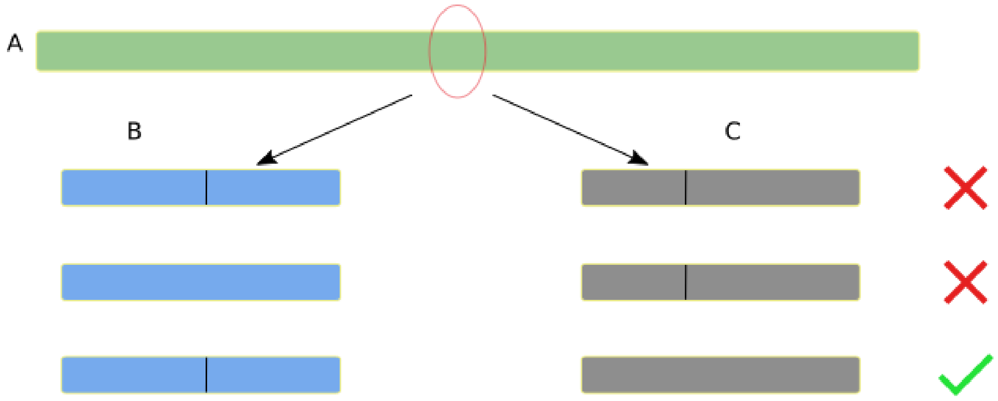

Fig 5.19 Identification schema for genetic variations associated with glyphosate-resistance phenotype. (A) the green bar represent the reference genome, red circle shows the variation hot spot, (B) genetic variants from tolerant line, (C) genetic variants from the sensitive line. Only variants that exists in the tolerant line with no equivalent variant at the same location in sensitive line are taken into consideration.

Table 5.10 Summary of identified small variants.

Variant group	SNPs	Indels
All	13,778,463	2,443,262
Tolerant line-specific	4,068,829	729,866
Located in coding sequence	113,775	15,277

Above procedure resulted in identification of 4,068,829 SNPs and 729,866 indels (Table 5.10). Among them, 113,775 SNPs and 15,277 indels were located within the protein coding regions. To gain the insight into functional consequences of the identified variants, I have performed the annotation with Ensembl variant effect predictor (VEP). The consequences of variations are shown in (Figure 5.20, 5.21) for all and high impact consequences respectively.

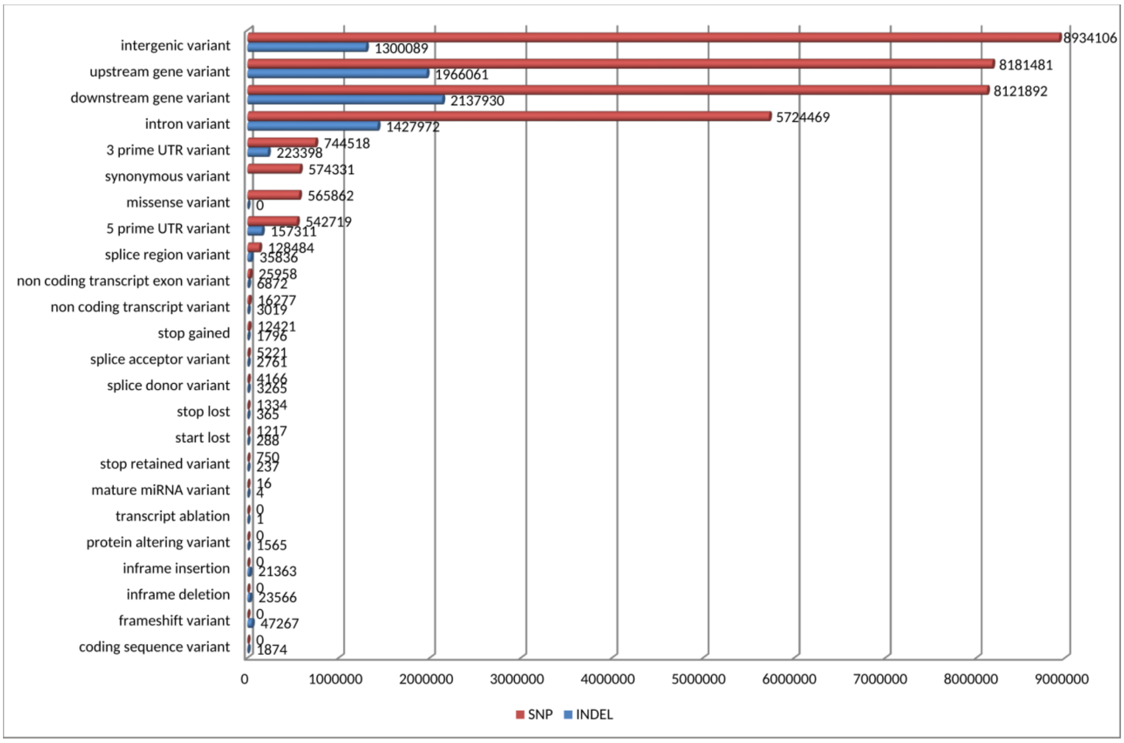

Figure 5.20 Predicted consequences of identified indels and SNPs.

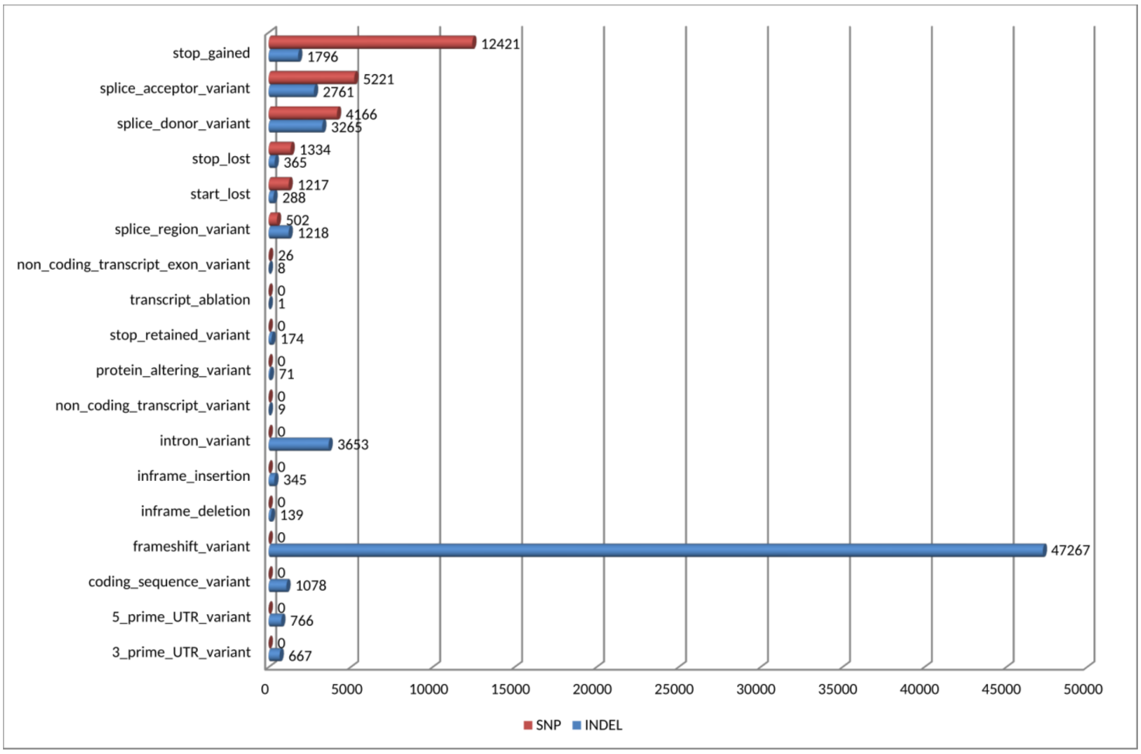

Figure 5.21 Predicted high impact consequences of identified indels and SNPs.

The variants predicted to have high impact were further annotated with the GO terms assigned to genes effected by those variants and summarized using REViGO (Figures 5.22, 5.23, 5.24, 5.25). Analysis of the distribution of the identified molecular function GO terms revealed that in case of both, SNPs and indels, the most affected group of genes belong to category of RNA binding, similarly to the results from annotation of large insertions identified by PBSuite. The second most abundant GO term, methyltransferase activity, was the most abundant GO term in revealed within genes affected by large deletions. Interestingly, a largest group of genes affected with SNPs and indels belong to biological process GO category of DNA repair, suggesting potential difference between the analyzed maize lines in ability to maintain the genome homeostasis.

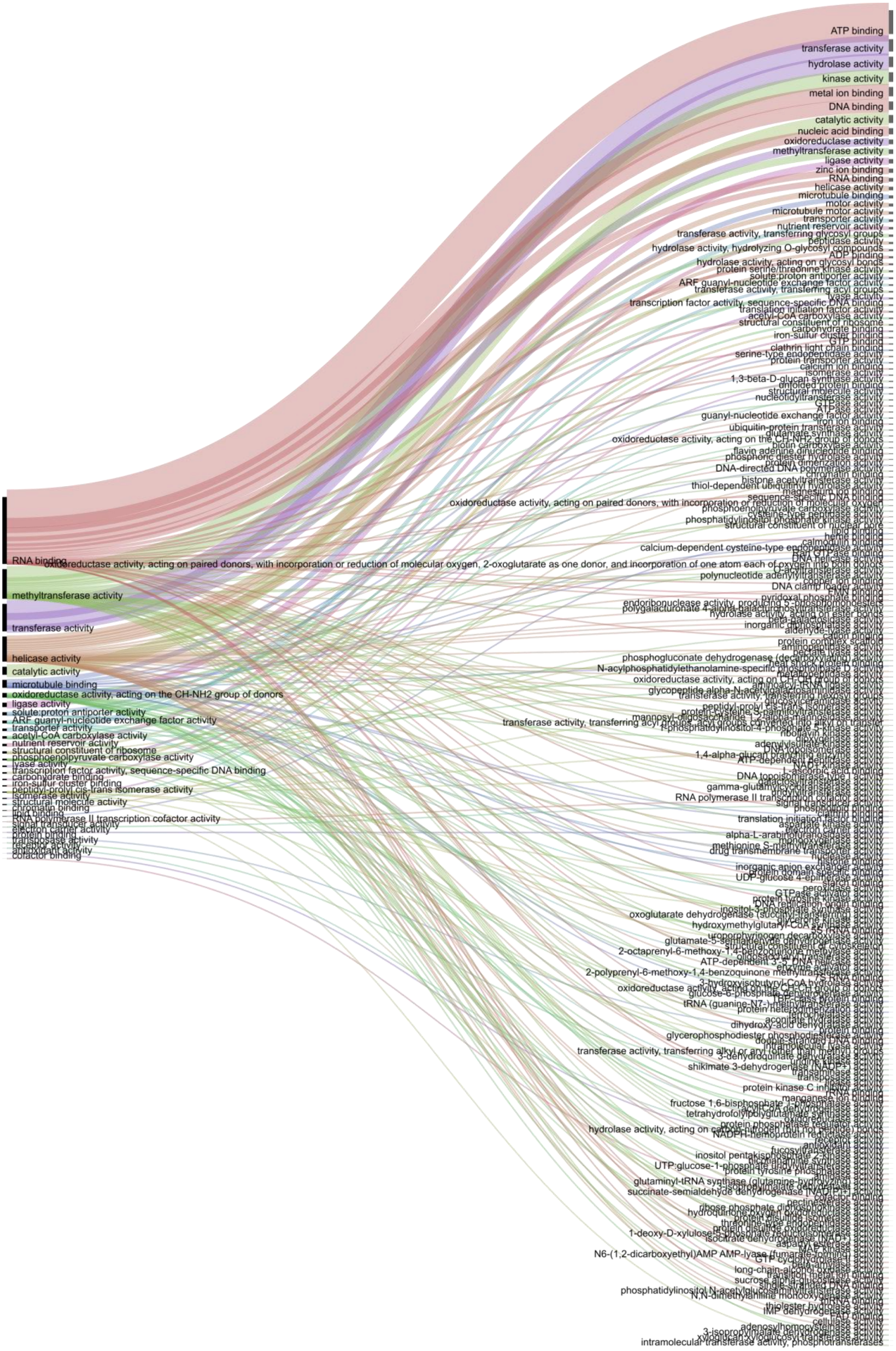
ATP binding
transferase activity
hydrolase activity
kinase activity
metal ion binding
DNA binding
catalytic activity
nucleic acid binding
oxidoreductase activity
methyltransferase activity
ligase activity
zinc ion binding
RNA binding
helicase activity
microtubule binding
motor activity
microtubule motor activity
transporter activity
nutrient reservoir activity
transferase activity, transferring glycosyl groups
peptidase activity
hydrolase activity, hydrolyzing O-glycosyl compounds
ADP binding
hydrolase activity, acting on glycosyl bonds
protein serine/threonine kinase activity
solute:proton antiporter activity
ARF guanyl-nucleotide exchange factor activity
transferase activity, transferring acyl groups
lyase activity
transcription factor activity, sequence-specific DNA binding
translation initiation factor activity
acetyl-CoA carboxylase activity
structural constituent of ribosome
carbohydrate binding
iron-sulfur cluster binding
GTP binding
clathrin light chain binding
serine-type endopeptidase activity
protein transporter activity
calcium ion binding
isomerase activity
1,3-beta-D-glucan synthase activity
unfolded protein binding
structural molecule activity
nucleotidyltransferase activity
GTPase activity
ATPase activity
guanyl-nucleotide exchange factor activity
iron ion binding
ubiquitin-protein transferase activity
glutamate synthase activity
oxidoreductase activity, acting on the CH-NH2 group of donors
biotin carboxylase activity
flavin adenine dinucleotide binding
phosphoric diester hydrolase activity
protein dimerization activity
DNA-directed DNA polymerase activity
chromatin binding
histone acetyltransferase activity
thiol-dependent ubiquitinyl hydrolase activity
magnesium ion binding
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
sequence-specific DNA binding
phosphoenolpyruvate carboxylase activity
cysteine-type peptidase activity
phosphatidylinositol phosphate kinase activity
structural constituent of nuclear pore
lipid binding
heme binding
calmodulin binding
calcium-dependent cysteine-type endopeptidase activity
Ran GTPase binding
DNA helicase activity
polynucleotide adenylyltransferase activity
DNA clamp loader activity
copper ion binding
pyridoxal phosphate binding
FMN binding
endoribonuclease activity, producing 5'-phosphomonoesters
polygalacturonate 4-alpha-galacturonosyltransferase activity
hydrolase activity, acting on ester bonds
beta-galactosidase activity
inorganic diphosphatase activity
cation binding
aminopeptidase activity
protein complex binding
phosphogluconate dehydrogenase (decarboxylating) activity
heat shock protein binding
N-acylphosphatidylethanolamine-specific phospholipase D activity
oxidoreductase activity, acting on CH-OH group of donors
glycopeptide alpha-N-acetylgalactosaminidase activity
transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer
mannosyl-oligosaccharide glucosidase activity
peptidyl-prolyl cis-trans isomerase activity
protein-cysteine S-palmitoyltransferase activity
1-phosphatidylinositol-4-phosphate 5-kinase activity
riboflavin kinase activity
adenylylsulfate kinase activity
1,4-alpha-glucan branching enzyme activity
DNA topoisomerase activity
ATP-dependent peptidase activity
NAD+ kinase activity
L-ascorbic acid binding
DNA topoisomerase type I activity
gamma-glutamyltransferase activity
RNA polymerase II transcription cofactor activity
signal transducer activity
phospholipid binding
translation initiation factor binding
aspartic-type endopeptidase activity
electron carrier activity
alpha-L-arabinofuranosidase activity
methionine S-methyltransferase activity
drug transmembrane transporter activity
histone binding
inorganic anion exchanger activity
protein domain specific binding
UDP-glucose 4-epimerase activity
starch binding
peroxidase activity
GTPase activator activity
protein tyrosine kinase activity
DNA replication origin binding
inositol-3-phosphate synthase activity
oxoglutarate dehydrogenase (succinyl-transferring) activity
hydroxymethylglutaryl-CoA synthase activity
5S rRNA binding
uroporphyrinogen decarboxylase activity
glutamate-5-semialdehyde dehydrogenase activity
2-octaprenyl-6-methoxy-1,4-benzoquinone methylase activity
ATP-dependent 3'-5' DNA helicase activity
2-polyprenyl-6-methoxy-1,4-benzoquinone methyltransferase activity
3-hydroxyisobutyryl-CoA hydrolase activity
oxidoreductase activity, acting on the CH-CH group of donors
glucose-6-phosphate dehydrogenase activity
tRNA (guanine-N7-)-methyltransferase activity
protein heterodimerization activity
aconitate hydratase activity
dihydroxy-acid dehydratase activity
glycerophosphodiester phosphodiesterase activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
3-dehydroquinate dehydratase activity
shikimate 3-dehydrogenase (NADP+) activity
protein kinase C inhibitor activity
manganese ion binding
fructose 1,6-bisphosphate 1-phosphatase activity
tetrahydrofolylpolyglutamate synthase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
protein phosphatase regulator activity
NADPH-hemoprotein reductase activity
inositol pentakisphosphate 2-kinase activity
fucosyltransferase activity
UTP:glucose-1-phosphate uridylyltransferase activity
protein tyrosine phosphatase activity
glutaminyl-tRNA synthase (glutamine-hydrolyzing) activity
succinate-semialdehyde dehydrogenase [NAD(P)+] activity
pectinesterase activity
ribose phosphate diphosphokinase activity
hydroquinone:oxygen oxidoreductase activity
threonine-tRNA ligase activity
1-deoxy-D-xylulose-5-phosphate synthase activity
isocitrate dehydrogenase (NAD+) activity
aspartate kinase activity
N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity
GTP cyclohydrolase I activity
long-chain-acyl-CoA dehydrogenase activity
transition metal ion binding
sucrose alpha-glucosidase activity
single-stranded DNA binding
phosphatidylinositol N-acetylglucosaminyltransferase activity
N,N-dimethylaniline monooxygenase activity
IMP dehydrogenase activity
thiolester hydrolase activity
FAD binding
adenosylhomocysteinase activity
cellulase activity
3-isopropylmalate dehydratase activity
xyloglucan:xyloglucosyl transferase activity
intramolecular transferase activity, phosphotransferases
RNA binding
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
methyltransferase activity
transferase activity
helicase activity
catalytic activity
microtubule binding
oxidoreductase activity, acting on the CH-NH2 group of donors
ligase activity
solute:proton antiporter activity
ARF guanyl-nucleotide exchange factor activity
transporter activity
acetyl-CoA carboxylase activity
nutrient reservoir activity
structural constituent of ribosome
phosphoenolpyruvate carboxylase activity
lyase activity
transcription factor activity, sequence-specific DNA binding
carbohydrate binding
iron-sulfur cluster binding
peptidyl-prolyl cis-trans isomerase activity
isomerase activity
structural molecule activity
chromatin binding
lipid binding
RNA polymerase II transcription cofactor activity
signal transducer activity
electron carrier activity
protein binding
translocase activity
receptor activity
antioxidant activity
cofactor binding

Figure 5.22 Tree view of molecular function gene ontology terms associated with genes affected by SNPs.

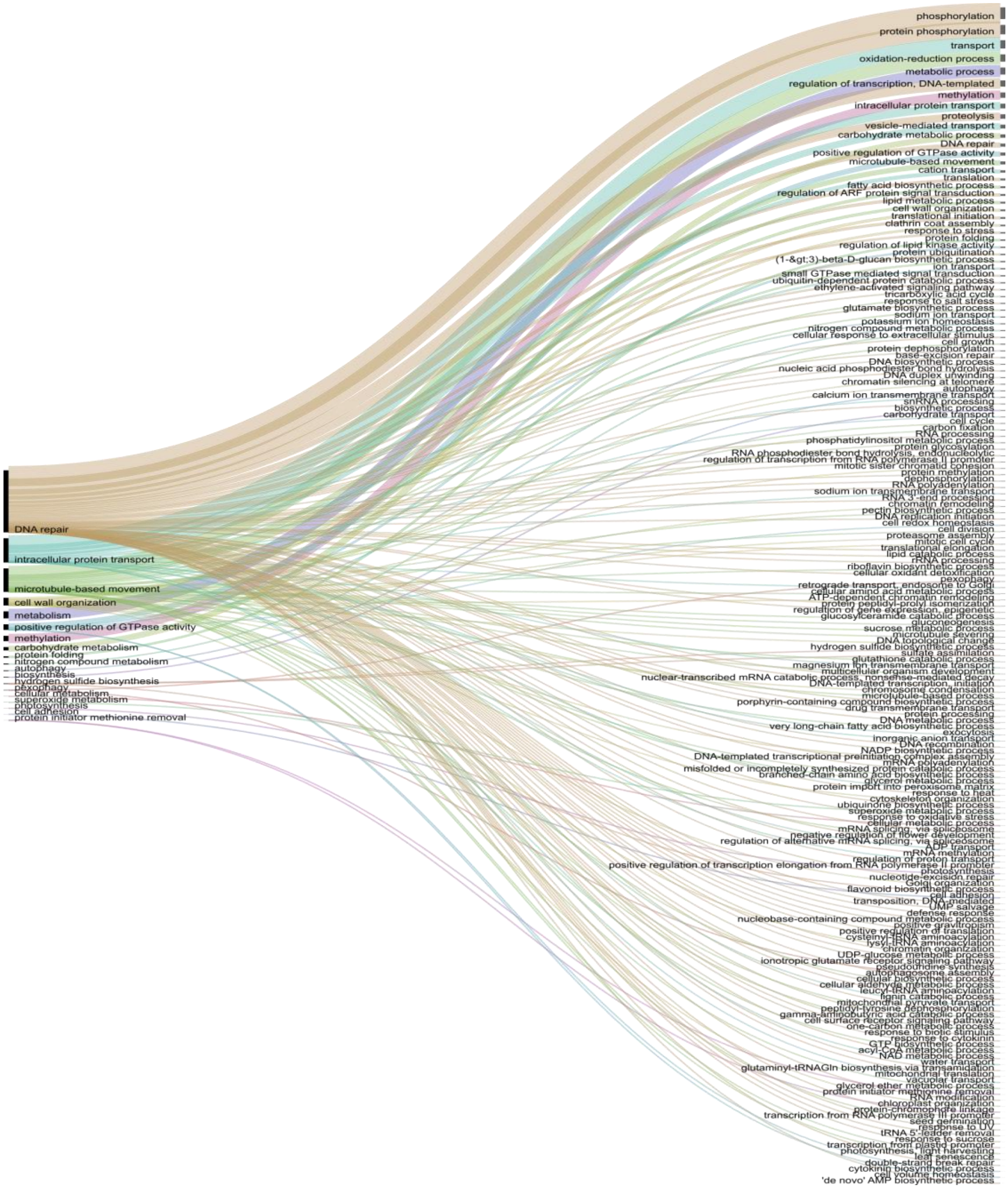
DNA repair
intracellular protein transport
microtubule-based movement
cell wall organization
metabolism
positive regulation of GTPase activity
methylation
carbohydrate metabolism
protein folding
nitrogen compound metabolism
autophagy
biosynthesis
hydrogen sulfide biosynthesis
pexophagy
cellular metabolism
superoxide metabolism
photosynthesis
cell adhesion
protein initiator methionine removal
phosphorylation
protein phosphorylation
transport
oxidation-reduction process
metabolic process
regulation of transcription, DNA-templated
methylation
intracellular protein transport
proteolysis
vesicle-mediated transport
carbohydrate metabolic process
DNA repair
positive regulation of GTPase activity
microtubule-based movement
cation transport
translation
fatty acid biosynthetic process
regulation of ARF protein signal transduction
lipid metabolic process
cell wall organization
translational initiation
clathrin coat assembly
response to stress
protein folding
regulation of lipid kinase activity
protein ubiquitination
(1-&gt;3)-beta-D-glucan biosynthetic process
ion transport
small GTPase mediated signal transduction
ubiquitin-dependent protein catabolic process
ethylene-activated signaling pathway
tricarboxylic acid cycle
response to salt stress
glutamate biosynthetic process
sodium ion transport
potassium ion homeostasis
nitrogen compound metabolic process
cellular response to extracellular stimulus
cell growth
protein dephosphorylation
base-excision repair
DNA biosynthetic process
nucleic acid phosphodiester bond hydrolysis
DNA duplex unwinding
chromatin silencing at telomere
autophagy
calcium ion transmembrane transport
snRNA processing
biosynthetic process
carbohydrate transport
cell cycle
carbon fixation
RNA processing
phosphatidylinositol metabolic process
protein glycosylation
RNA phosphodiester bond hydrolysis, endonucleolytic
regulation of transcription from RNA polymerase II promoter
mitotic sister chromatid cohesion
protein methylation
dephosphorylation
RNA polyadenylation
sodium ion transmembrane transport
RNA 3'-end processing
chromatin remodeling
pectin biosynthetic process
DNA replication initiation
cell redox homeostasis
cell division
proteasome assembly
mitotic cell cycle
translational elongation
lipid catabolic process
rRNA processing
riboflavin biosynthetic process
cellular oxidant detoxification
pexophagy
retrograde transport, endosome to Golgi
cellular amino acid metabolic process
ATP-dependent chromatin remodeling
protein peptidyl-prolyl isomerization
regulation of gene expression, epigenetic
glucosylceramide catabolic process
gluconeogenesis
sucrose metabolic process
microtubule severing
DNA topological change
hydrogen sulfide biosynthetic process
sulfate assimilation
glutathione catabolic process
magnesium ion transmembrane transport
multicellular organism development
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
DNA-templated transcription, initiation
chromosome condensation
microtubule-based process
porphyrin-containing compound biosynthetic process
drug transmembrane transport
protein processing
DNA metabolic process
very long-chain fatty acid biosynthetic process
exocytosis
inorganic anion transport
DNA recombination
NADP biosynthetic process
DNA-templated transcriptional preinitiation complex assembly
mRNA polyadenylation
misfolded or incompletely synthesized protein catabolic process
branched-chain amino acid biosynthetic process
glycerol metabolic process
protein import into peroxisome matrix
response to heat
cytoskeleton organization
ubiquinone biosynthetic process
superoxide metabolic process
response to oxidative stress
cellular metabolic process
mRNA splicing, via spliceosome
negative regulation of flower development
regulation of alternative mRNA splicing, via spliceosome
ADP transport
mRNA methylation
regulation of proton transport
positive regulation of transcription elongation from RNA polymerase II promoter
photosynthesis
nucleotide-excision repair
Golgi organization
flavonoid biosynthetic process
cell adhesion
transposition, DNA-mediated
UMP salvage
defense response
nucleobase-containing compound metabolic process
positive gravitropism
positive regulation of translation
cysteinyl-tRNA aminoacylation
lysyl-tRNA aminoacylation
chromatin organization
UDP-glucose metabolic process
ionotropic glutamate receptor signaling pathway
pseudouridine synthesis
autophagosome assembly
cellular biosynthetic process
cellular aldehyde metabolic process
leucyl-tRNA aminoacylation
lignin catabolic process
mitochondrial pyruvate transport
peptidyl-tyrosine dephosphorylation
gamma-aminobutyric acid catabolic process
cell surface receptor signaling pathway
one-carbon metabolic process
response to biotic stimulus
response to cytokinin
GTP biosynthetic process
acyl-CoA metabolic process
NAD metabolic process
water transport
glutaminyl-tRNAGln biosynthesis via transamidation
mitochondrial translation
vacuolar transport
glycerol ether metabolic process
protein initiator methionine removal
RNA modification
chloroplast organization
protein-chromophore linkage
transcription from RNA polymerase III promoter
seed germination
response to UV
tRNA 5'-leader removal
response to sucrose
transcription from plastid promoter
photosynthesis, light harvesting
leaf senescence
double-strand break repair
cytokinin biosynthetic process
cell volume homeostasis
'de novo' AMP biosynthetic process

Figure 5.23 Tree view of biological process gene ontology terms associated with genes affected by SNPs.

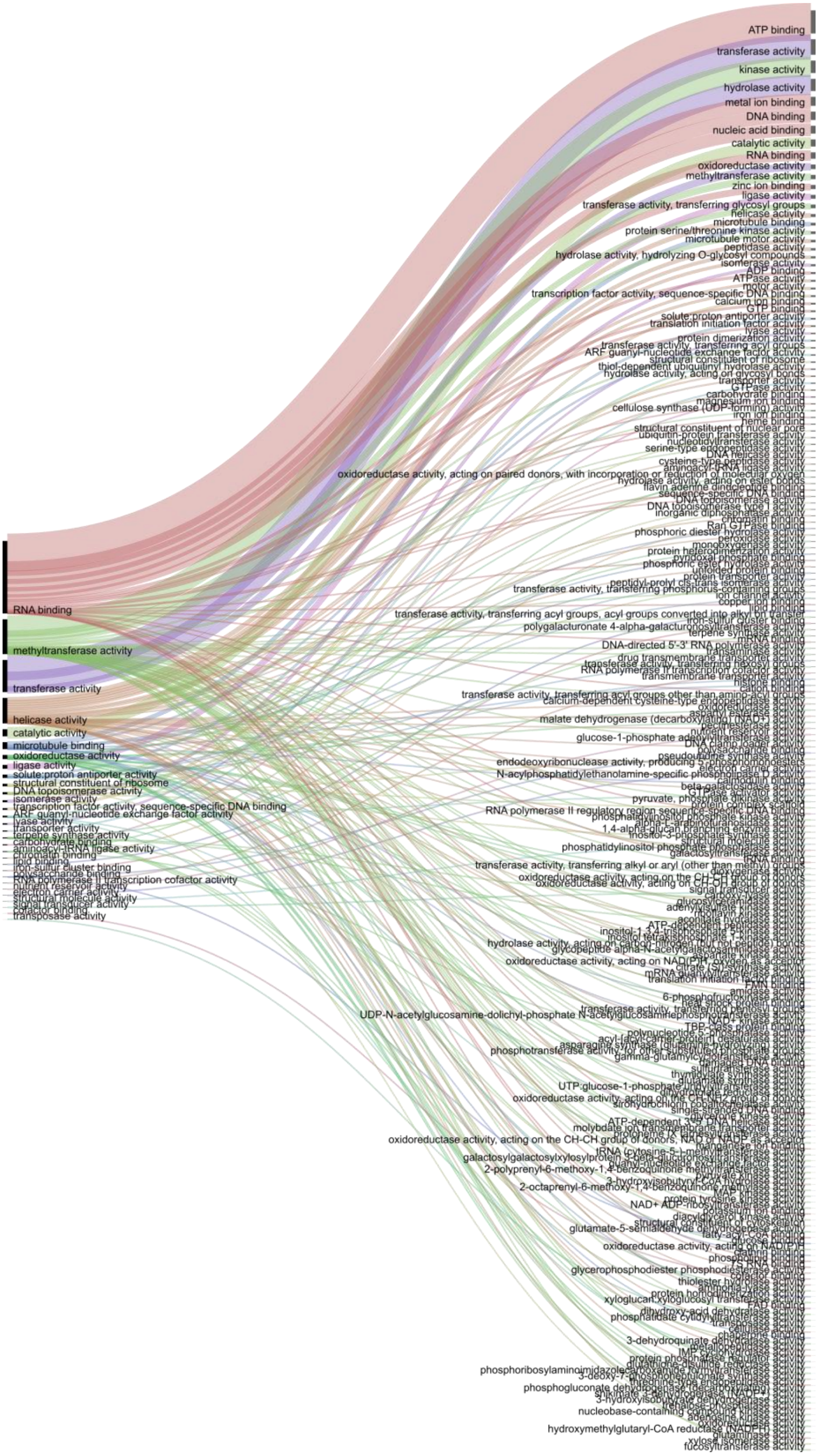
ATP binding
transferase activity
kinase activity
hydrolase activity
metal ion binding
DNA binding
nucleic acid binding
catalytic activity
RNA binding
oxidoreductase activity
methyltransferase activity
zinc ion binding
ligase activity
transferase activity, transferring glycosyl groups
helicase activity
microtubule binding
protein serine/threonine kinase activity
microtubule motor activity
peptidase activity
hydrolase activity, hydrolyzing O-glycosyl compounds
isomerase activity
ADP binding
ATPase activity
motor activity
transcription factor activity, sequence-specific DNA binding
calcium ion binding
GTP binding
solute:proton antiporter activity
translation initiation factor activity
lyase activity
protein dimerization activity
transferase activity, transferring acyl groups
ARF guanyl-nucleotide exchange factor activity
structural constituent of ribosome
thiol-dependent ubiquitinyl hydrolase activity
hydrolase activity, acting on glycosyl bonds
transporter activity
GTPase activity
carbohydrate binding
magnesium ion binding
cellulose synthase (UDP-forming) activity
iron ion binding
structural constituent of nuclear pore
heme binding
ubiquitin-protein transferase activity
nucleotidyltransferase activity
serine-type endopeptidase activity
DNA helicase activity
cysteine-type peptidase activity
aminoacyl-tRNA ligase activity
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
hydrolase activity, acting on ester bonds
flavin adenine dinucleotide binding
sequence-specific DNA binding
DNA topoisomerase activity
DNA topoisomerase type I activity
inorganic diphosphatase activity
chromatin binding
Ran GTPase binding
phosphoric diester hydrolase activity
peroxidase activity
monooxygenase activity
protein heterodimerization activity
pyridoxal phosphate binding
phosphoric ester hydrolase activity
unfolded protein binding
protein transporter activity
peptidyl-prolyl cis-trans isomerase activity
transferase activity, transferring phosphorus-containing groups
ion channel activity
copper ion binding
lipid binding
transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer
iron-sulfur cluster binding
polygalacturonate 4-alpha-galacturonosyltransferase activity
terpene synthase activity
DNA-directed 5'-3' RNA polymerase activity
transaminase activity
drug transmembrane transporter activity
transferase activity, transferring hexosyl groups
RNA polymerase II transcription cofactor activity
transmembrane transporter activity
histone binding
cation binding
transferase activity, transferring acyl groups other than amino-acyl groups
calcium-dependent cysteine-type endopeptidase activity
oxidoreductase activity
malate dehydrogenase (decarboxylating) (NAD+)
pectinesterase activity
nutrient reservoir activity
glucose-1-phosphate adenylyltransferase activity
DNA clamp loader activity
polysaccharide binding
pseudouridine synthase activity
endodeoxyribonuclease activity, producing 5'-phosphomonoesters
electron carrier activity
N-acylphosphatidylethanolamine-specific phospholipase D activity
beta-galactosidase activity
GTPase activator activity
pyruvate, phosphate dikinase activity
RNA polymerase II regulatory region sequence-specific DNA binding
phosphatidylinositol phosphate kinase activity
alpha-L-arabinofuranosidase activity
1,4-alpha-glucan branching enzyme activity
inositol-3-phosphate synthase activity
phosphatidylinositol phosphate phosphatase activity
structural molecule activity
galactosyltransferase activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
dioxygenase activity
oxidoreductase activity, acting on the CH-CH group of donors
oxidoreductase activity, acting on CH-OH group of donors
signal transducer activity
glucosylceramidase activity
adenylylsulfate kinase activity
riboflavin kinase activity
ATP-dependent peptidase activity
inositol-1,3,4-trisphosphate 5/6-kinase activity
inositol tetrakisphosphate 1-kinase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
glycopeptide alpha-N-acetylgalactosaminidase activity
oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor
mRNA guanylyltransferase activity
translation initiation factor activity
FMN binding
6-phosphofructokinase activity
UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosaminephosphotransferase activity
transferase activity, transferring pentosyl groups
TPP-class protein binding
polynucleotide 5'-phosphatase activity
asparagine synthase (glutamine-hydrolyzing) activity
phosphotransferase activity, alcohol group as acceptor
gamma-glutamyltransferase activity
damaged DNA binding
sulfotransferase activity
thymidylate synthase activity
UTP:glucose-1-phosphate uridylyltransferase activity
oxidoreductase activity, acting on the CH-NH2 group of donors
single-stranded DNA binding
ATP-dependent 3'-5' DNA helicase activity
molybdate ion transmembrane transporter activity
oxidoreductase activity, acting on the CH-CH group of donors, NAD or NADP as acceptor
tRNA (cytosine-5-)-methyltransferase activity
manganese ion binding
galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase activity
2-polyprenyl-6-methoxy-1,4-benzoquinone methyltransferase activity
3-hydroxyisobutyryl-CoA hydrolase activity
2-octaprenyl-6-methoxy-1,4-benzoquinone methylase activity
protein tyrosine kinase activity
NAD+ ADP-ribosyltransferase activity
structural constituent of cytoskeleton
diacylglycerol kinase activity
glutamate-5-semialdehyde dehydrogenase activity
fatty-acyl-CoA binding
oxidoreductase activity, acting on NAD(P)H
phosphatase binding
glycerophosphodiester phosphodiesterase activity
thiolester hydrolase activity
ammonia-lyase activity
protein homodimerization activity
xyloglucan:xyloglucosyl transferase activity
dihydroxy-acid dehydratase activity
phosphatidate cytidylyltransferase activity
3-dehydroquinate dehydratase activity
chaperone binding
IMP cyclohydrolase activity
protein phosphatase regulator activity
phosphoribosylaminoimidazolecarboxamide formyltransferase activity
phosphogluconate dehydrogenase (decarboxylating) activity
shikimate 3-dehydrogenase (NADP+) activity
3-hydroxyisobutyrate dehydrogenase activity
nucleobase-containing compound kinase activity
trehalose-phosphatase activity
adenosine kinase activity
hydroxymethylglutaryl-CoA reductase (NADPH) activity
glutaminase activity
xylose isomerase activity
fucosyltransferase activity
RNA binding
methyltransferase activity
transferase activity
helicase activity
catalytic activity
microtubule binding
oxidoreductase activity
ligase activity
solute:proton antiporter activity
structural constituent of ribosome
DNA topoisomerase activity
isomerase activity
transcription factor activity, sequence-specific DNA binding
ARF guanyl-nucleotide exchange factor activity
lyase activity
transporter activity
terpene synthase activity
carbohydrate binding
aminoacyl-tRNA ligase activity
chromatin binding
lipid binding
iron-sulfur cluster binding
polysaccharide binding
RNA polymerase II transcription cofactor activity
nutrient reservoir activity
electron carrier activity
structural molecule activity
signal transducer activity
cofactor binding
transposase activity

Figure 5.24 Tree view of molecular function gene ontology terms associated with genes affected by indels.

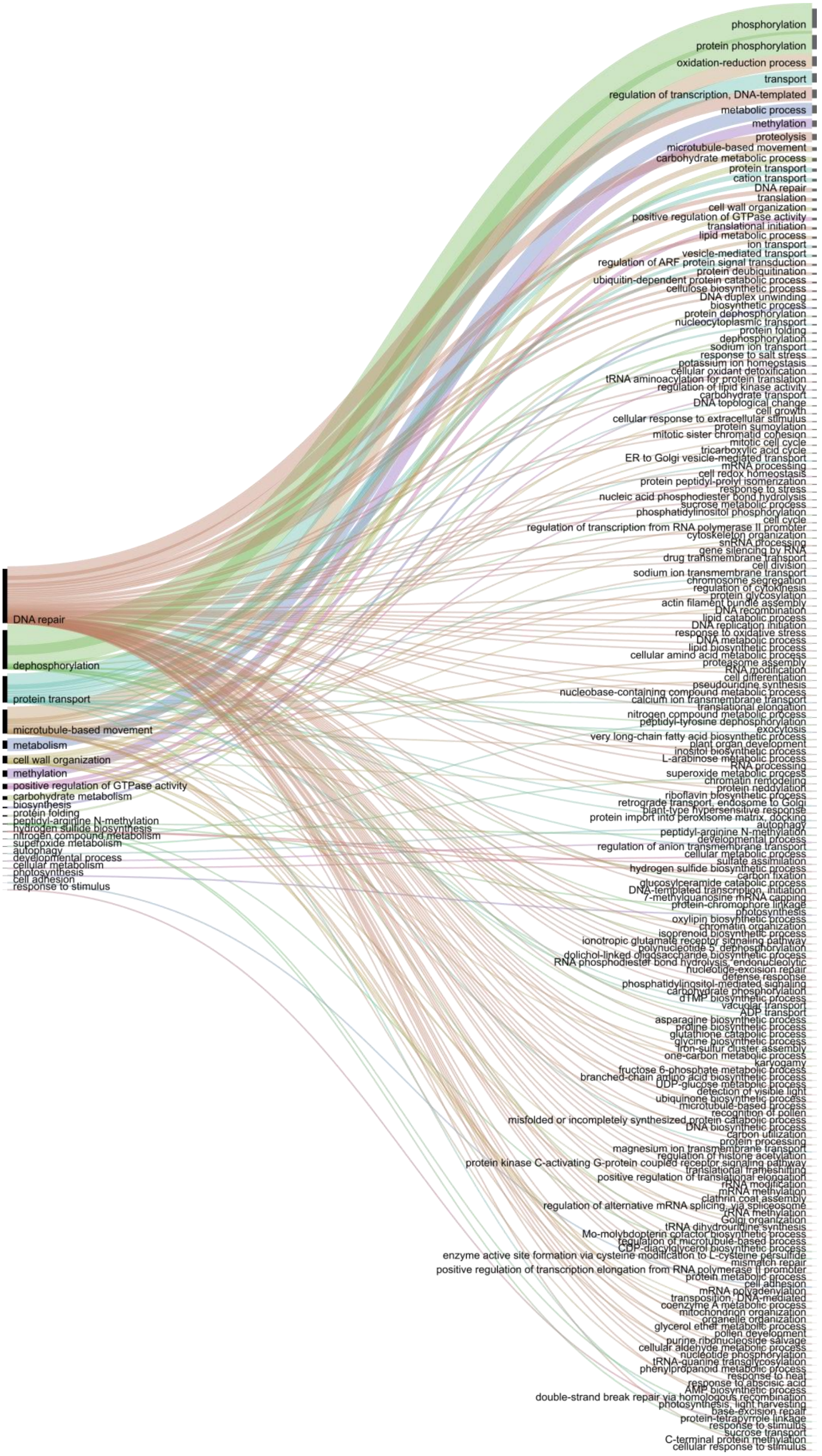
DNA repair
dephosphorylation
protein transport
microtubule-based movement
metabolism
cell wall organization
methylation
positive regulation of GTPase activity
carbohydrate metabolism
biosynthesis
protein folding
peptidyl-arginine N-methylation
hydrogen sulfide biosynthesis
nitrogen compound metabolism
superoxide metabolism
autophagy
developmental process
cellular metabolism
photosynthesis
cell adhesion
response to stimulus
phosphorylation
protein phosphorylation
oxidation-reduction process
transport
regulation of transcription, DNA-templated
metabolic process
methylation
proteolysis
microtubule-based movement
carbohydrate metabolic process
protein transport
cation transport
DNA repair
translation
cell wall organization
positive regulation of GTPase activity
translational initiation
lipid metabolic process
ion transport
vesicle-mediated transport
regulation of ARF protein signal transduction
protein deubiquitination
ubiquitin-dependent protein catabolic process
cellulose biosynthetic process
DNA duplex unwinding
biosynthetic process
protein dephosphorylation
nucleocytoplasmic transport
protein folding
dephosphorylation
sodium ion transport
response to salt stress
potassium ion homeostasis
cellular oxidant detoxification
tRNA aminoacylation for protein translation
regulation of lipid kinase activity
carbohydrate transport
DNA topological change
cell growth
cellular response to extracellular stimulus
protein sumoylation
mitotic sister chromatid cohesion
mitotic cell cycle
tricarboxylic acid cycle
ER to Golgi vesicle-mediated transport
mRNA processing
cell redox homeostasis
protein peptidyl-prolyl isomerization
response to stress
nucleic acid phosphodiester bond hydrolysis
sucrose metabolic process
phosphatidylinositol phosphorylation
cell cycle
regulation of transcription from RNA polymerase II promoter
cytoskeleton organization
snRNA processing
gene silencing by RNA
drug transmembrane transport
cell division
sodium ion transmembrane transport
chromosome segregation
regulation of cytokinesis
protein glycosylation
actin filament bundle assembly
DNA recombination
lipid catabolic process
DNA replication initiation
response to oxidative stress
DNA metabolic process
lipid biosynthetic process
cellular amino acid metabolic process
proteasome assembly
RNA modification
cell differentiation
pseudouridine synthesis
nucleobase-containing compound metabolic process
calcium ion transmembrane transport
translational elongation
nitrogen compound metabolic process
peptidyl-tyrosine dephosphorylation
exocytosis
very long-chain fatty acid biosynthetic process
plant organ development
inositol biosynthetic process
L-arabinose metabolic process
RNA processing
superoxide metabolic process
chromatin remodeling
protein neddylation
riboflavin biosynthetic process
retrograde transport, endosome to Golgi
plant-type hypersensitive response
protein import into peroxisome matrix, docking
autophagy
peptidyl-arginine N-methylation
developmental process
regulation of anion transmembrane transport
cellular metabolic process
sulfate assimilation
hydrogen sulfide biosynthetic process
carbon fixation
glucosylceramide catabolic process
DNA-templated transcription, initiation
7-methylguanosine mRNA capping
protein-chromophore linkage
photosynthesis
oxylipin biosynthetic process
chromatin organization
isoprenoid biosynthetic process
ionotropic glutamate receptor signaling pathway
polynucleotide 5' dephosphorylation
dolichol-linked oligosaccharide biosynthetic process
RNA phosphodiester bond hydrolysis, endonucleolytic
nucleotide-excision repair
defense response
phosphatidylinositol-mediated signaling
carbohydrate phosphorylation
dTMP biosynthetic process
vacuolar transport
ADP transport
asparagine biosynthetic process
proline biosynthetic process
glutathione catabolic process
glycine biosynthetic process
iron-sulfur cluster assembly
one-carbon metabolic process
karyogamy
fructose 6-phosphate metabolic process
branched-chain amino acid biosynthetic process
UDP-glucose metabolic process
detection of visible light
ubiquinone biosynthetic process
microtubule-based process
recognition of pollen
misfolded or incompletely synthesized protein catabolic process
DNA biosynthetic process
carbon utilization
protein processing
magnesium ion transmembrane transport
regulation of histone acetylation
protein kinase C-activating G-protein coupled receptor signaling pathway
translational frameshifting
positive regulation of translational elongation
rRNA modification
mRNA methylation
clathrin coat assembly
regulation of alternative mRNA splicing, via spliceosome
rRNA methylation
Golgi organization
tRNA dihydrouridine synthesis
Mo-molybdopterin cofactor biosynthetic process
regulation of microtubule-based process
CDP-diacylglycerol biosynthetic process
enzyme active site formation via cysteine modification to L-cysteine persulfide
mismatch repair
positive regulation of transcription elongation from RNA polymerase II promoter
protein metabolic process
cell adhesion
mRNA polyadenylation
transposition, DNA-mediated
coenzyme A metabolic process
mitochondrion organization
organelle organization
glycerol ether metabolic process
pollen development
purine ribonucleoside salvage
cellular aldehyde metabolic process
nucleotide phosphorylation
tRNA-guanine transglycosylation
phenylpropanoid metabolic process
response to heat
response to abscisic acid
AMP biosynthetic process
double-strand break repair via homologous recombination
photosynthesis, light harvesting
base-excision repair
protein-tetrapyrrole linkage
response to stimulus
sucrose transport
C-terminal protein methylation
cellular response to stimulus

Figure 5.25 Tree view of biological process gene ontology terms associated with genes affected by indels.

5.4. Variants potentially associated with glyphosate resistance

The presented results suggest high genetic distance between analyzed glyphosate tolerant and sensitive line. The functional diversity of genes affected by identified changes suggest that vast majority of variants are related rather to the line-specific variability, which is not directly related to the phenotype of glyphosate resistance. Thus, to spot the variants which could be potentially related to the phenotype of interest, I have analyzed in more details the genes known to be potentially involved in glyphosate response.

First, I have analyzed the variants located within the EPSPS gene, which is the direct target of glyphosate. No large structural variants have been found around EPSPS gene in range of ±10kb. I have found however several SNPs and indels within, or in vicinity of the EPSPS gene (Figure 5.26). All variants were predicted with VEP to have only moderate or modifier impact on gene expression and none of them was located within the coding region. Four of SNPs were found to affect the 3'UTR. Few of the

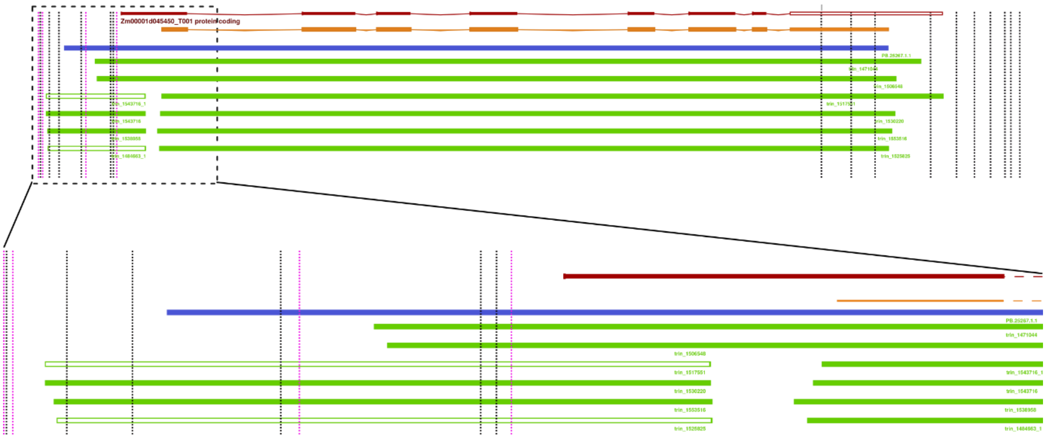

Figure 5.26 Location of identified SNPs (black) and indels (violet) on EPSPS gene. In red, the EPSP gene structure according to version 4 annotation of maize genome, in orange the annotation in version 3. Blue bars represent the publicly available full length EPSPS transcript isoform identified with iso-seq approach, green bars represent publicly available Trinity-assembled transcripts from RNA-seq data.

variants were also located upstream from the annotated CDS start, but no 5’UTR annotation was available. Thus, by analyzing the annotation of the EPSPS loci in Ensembl genome browser, I have noticed that there are known ESTs which extend the annotation of the EPSPS gene toward the 5’UTR region. Especially, the presence of the EST from full-length transcript sequencing using Iso-seq approach suggest that at least some of the identified variants fall within the 5’UTR region of the EPSPS. It is however impossible to predict the exact functional consequence of those changes.

I was able to identify other genes encoding proteins involved in shikimate pathway that were affected by detected variations. Those include bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase and chorismate synthase

Table 5.11 Shikimate pathway genes affected by SNPs and indels.

Gene	Protein	Variation	Consequence
100272333	Bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase chloroplastic	SNPs, indels	Splice donor/ frame shift
100381407	Chorismate synthase chloroplastic	SNPs, indels	Splice donor/ frame shift

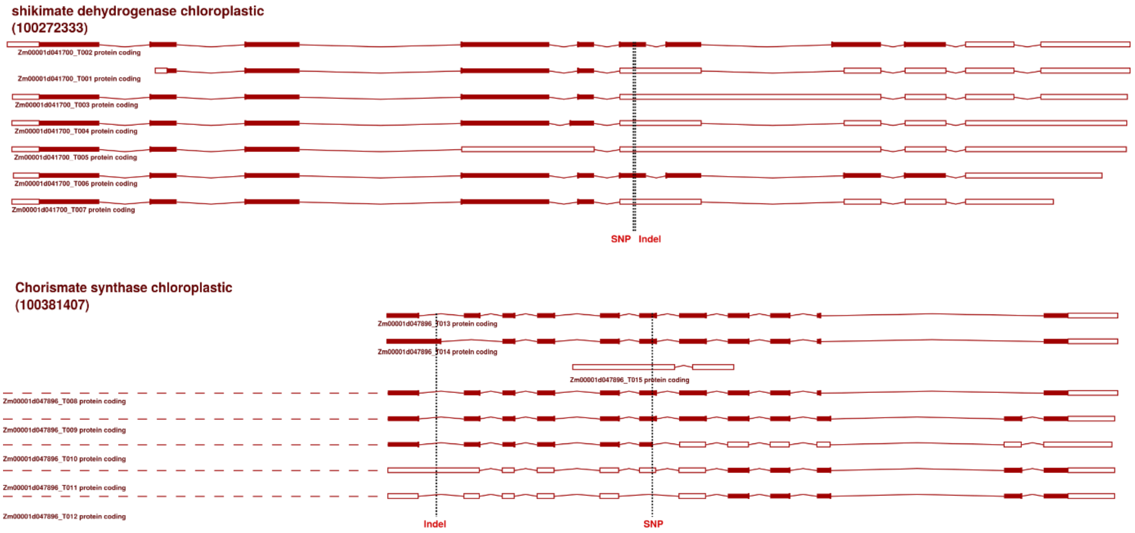
shikimate dehydrogenase chloroplastic
(100272333)
Zm00001d041700_T002 protein coding
Zm00001d041700_T001 protein coding
Zm00001d041700_T003 protein coding
Zm00001d041700_T004 protein coding
Zm00001d041700_T005 protein coding
Zm00001d041700_T006 protein coding
Zm00001d041700_T007 protein coding
SNP
Indel
Chorismate synthase chloroplastic
(100381407)
Zm00001d047896_T013 protein coding
Zm00001d047896_T014 protein coding
Zm00001d047896_T015 protein coding
Zm00001d047896_T008 protein coding
Zm00001d047896_T009 protein coding
Zm00001d047896_T010 protein coding
Zm00001d047896_T011 protein coding
Zm00001d047896_T012 protein coding
Indel
SNP

Figure 5.27 Genetic variations in genes encoding proteins involved in shikimate pathway.

(Table 5.11, Figure 5.27). In both cases, predicted impact on gene expression was high, affecting splice sites and causing potential frame shifts. Such changes could lead to changes in protein expression (e.g. via directing transcripts into nonsense-mediated decay) and in protein activity. Although, I cannot predict the actual functional consequences, it is potentially possible that both enzymes could gain higher activity. In such case it would be possible to at least partially compensate for inhibitory effect of glyphosate on EPSPS by increase of EPSPS substrate synthesis and elevated intake of the EPSPS product.

Another group of genes, which have been reported to be potentially involved in glyphosate resistance, are phosphate transporters, which could participate in active transport of glyphosate into plant cell. I have identified a number of genes encoding phosphate transporters that were affected by SNPs and indels predicted to have high impact on gene expression (Table 5.12). Two of them, phosphate transporters 1 and 2 have been recently shown to be differentially regulated by miRNAs involved in

Table 5.12 Phosphate transporters genes affected by variations.

Gene	Protein	Variation	Consequence
Zm00001d051945	Phosphate transporter 2	SNP	Splice acceptor
Zm00001d018445	Phosphate transporter 3	indel	Splice donor
Zm00001d012747	Putative sugar phosphate/phosphate translocator	SNP, indel	Splice donor/ Splice acceptor
Zm00001d021653	Glucose-6-phosphate/phosphate translocator 2	indel	Frame shift
100191756	Probable sugar phosphate/phosphate translocator	indel	Frame shift
Zm00001d011388	Putative sugar phosphate/phosphate translocator	indel	Frame shift

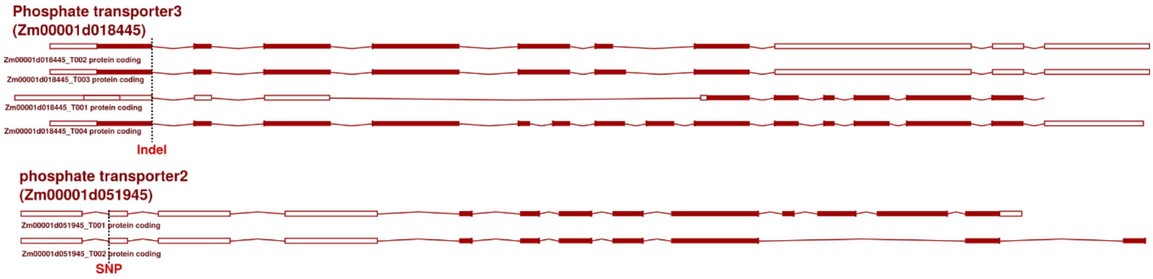
Phosphate transporter3
(Zm00001d018445)
Zm00001d018445_T002 protein coding
Zm00001d018445_T003 protein coding
Zm00001d018445_T001 protein coding
Zm00001d018445_T004 protein coding
Indel
phosphate transporter2
(Zm00001d051945)
Zm00001d051945_T001 protein coding
Zm00001d051945_T002 protein coding
SNP

Figure 5.28 Genetic variations in phosphate transporter genes.

response to glyphosate (Żywicki, Gracz, Karłowski, Twardowski, & Tyczewska, 2015). In both of them, identified variations affect the splicing of the first intron, which is crucial in regulation of the gene expression (Figure 5.28). Four remaining genes encode sugar phosphate/phosphate translocators which are responsible for transport of Glc6P, but also inorganic phosphate, 3-phosphoglycerate, triose phosphates and, to a lesser extent, phosphoenolpyruvate (PEP) which is the substrate for shikimate pathway directly affected by glyphosate.

In fact, I was able to identify more genes affected by high impact genetic changes which are related to cellular availability of the phosphoenolpyruvate (PEP) (Table 5.13). Four of identified genes encode chloroplastic phosphoenolpyruvate/phosphate translocators which are important for maintaining appropriate concentration of PEP in chloroplasts, where shikimate pathway take place. Phosphoenolpyruvate carboxylase (PEPC) is responsible for catalysis of the addition of bicarbonate to PEP, forming oxaloacetate and inorganic phosphate. This reaction is directing PEP into the carbon fixation pathway, thus, lowering its availability for shikimate pathway. Last of identified genes encode phosphoenolpyruvate carboxylase kinase, which regulate the activity of the PEPC by its phosphorylation.

Table 5.13 Genes associated with phosphoenolpyruvate availability affected by variations.

Gene	Protein	Variation	Consequence
100283648	Phosphoenolpyruvate/phosphate translocator 1 chloroplastic	SNP	Splice acceptor
103649694	Phosphoenolpyruvate/phosphate translocator 2 chloroplastic	indels	Frame shift
Zm00001d044715	Phosphoenolpyruvate/phosphate translocator 2 chloroplastic	indel	Frame shift
Zm00001d037659	Phosphoenolpyruvate/phosphate translocator 2 chloroplastic	SV insertion	Stop lost
542372	Phosphoenolpyruvate carboxylase	SNP	Stop gained
Zm00001d053453	Phosphoenolpyruvate carboxylase isoform 1	SNPs, indels	Splice acceptor/Splice donor
542479	Phosphoenolpyruvate carboxylase2	SNP	Splice donor
Zm00001d024980	Phosphoenolpyruvate carboxylase 3	SNPs, indels	Stop gained/ frame shift
103649899	Phosphoenolpyruvate carboxylase 4	SNP	Splice acceptor
Zm00001d051156	Putative phosphoenolpyruvate carboxylase kinase family protein	SNPs, indel	Stop gained/frame shift

It has been shown in rice, that silencing of the PEPC gene lead to 50-60% increase of the activity of the shikimate pathway, due to higher PEP availability (Masumoto et al., 2010). Since all mentioned above genes are affected by high impact variants (Table 5.13, Figure 5.29), one could expect the differences in PEP availability between analyzed maize lines. Potentially, if the cumulated effect of altered PEP-related genes would lead to increase of chloroplastic PEP availability, it could lead to stimulation of shikimate pathway and at least partially compensate inhibition of EPSPS by glyphosate. This however has to be addressed by in-depth experimental study.

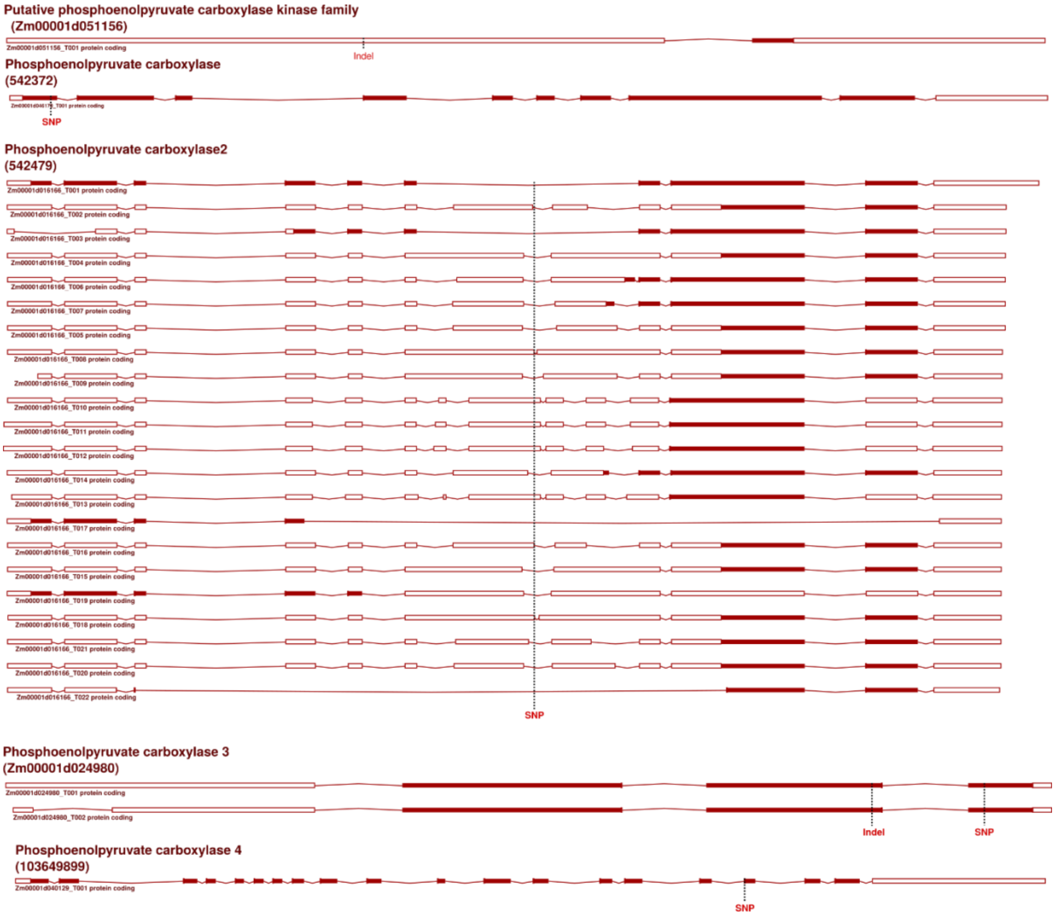
Putative phosphoenolpyruvate carboxylase kinase family
(Zm00001d051156)
Zm00001d051156_T001 protein coding
Indel
Phosphoenolpyruvate carboxylase
(542372)
Zm00001d046170_T001 protein coding
SNP
Phosphoenolpyruvate carboxylase2
(542479)
Zm00001d016166_T001 protein coding
Zm00001d016166_T002 protein coding
Zm00001d016166_T003 protein coding
Zm00001d016166_T004 protein coding
Zm00001d016166_T006 protein coding
Zm00001d016166_T007 protein coding
Zm00001d016166_T005 protein coding
Zm00001d016166_T008 protein coding
Zm00001d016166_T009 protein coding
Zm00001d016166_T010 protein coding
Zm00001d016166_T011 protein coding
Zm00001d016166_T012 protein coding
Zm00001d016166_T014 protein coding
Zm00001d016166_T013 protein coding
Zm00001d016166_T017 protein coding
Zm00001d016166_T016 protein coding
Zm00001d016166_T015 protein coding
Zm00001d016166_T019 protein coding
Zm00001d016166_T018 protein coding
Zm00001d016166_T021 protein coding
Zm00001d016166_T020 protein coding
Zm00001d016166_T022 protein coding
SNP
Phosphoenolpyruvate carboxylase 3
(Zm00001d024980)
Zm00001d024980_T001 protein coding
Zm00001d024980_T002 protein coding
Indel
SNP
Phosphoenolpyruvate carboxylase 4
(103649899)
Zm00001d040129_T001 protein coding
SNP

Figure 5.29 Genetic variations in genes encoding proteins related to lowering the availability of phosphoenolpyruvate for shikimate pathway.

The last group of genes that were affected by variations includes genes encoding proteins involved in multidrug and toxic compounds extrusion (Table 5.14, Figure 5.30). This family of proteins constitutes one of the largest transporter families in plants which are conserved in living organisms. They function as multidrug resistance proteins by transporting drugs and synthetic substances through membrane. They also have been shown to be involved in a wide variety of physiological functions throughout plant development via transporting a broad range of substrates such as organic acids, plant hormones and secondary metabolites.

Changes in their function caused by identified variations could lead to glyphosate resistance by active lowering of the intracellular amount of herbicide in maize cells.

Table 5.14 Multidrug and toxic compounds extrusion genes affected by variations.

Gene	Protein	Variation	Consequence
Zm00001d005080	Multidrug and toxic compound extrusion5	SNPs, indels	Splice acceptor/frame shift
Zm00001d013810	Multidrug and toxic compound extrusion2	indels, SV insertion	Frame shift/start lost
100383875	Multidrug and toxic compound extrusion3	indel	Frame shift
100193278	Multidrug and toxic compound extrusion4	indel	Frame shift
Zm00001d035115	Multidrug and toxic compound extrusion1	indel	Stop gained

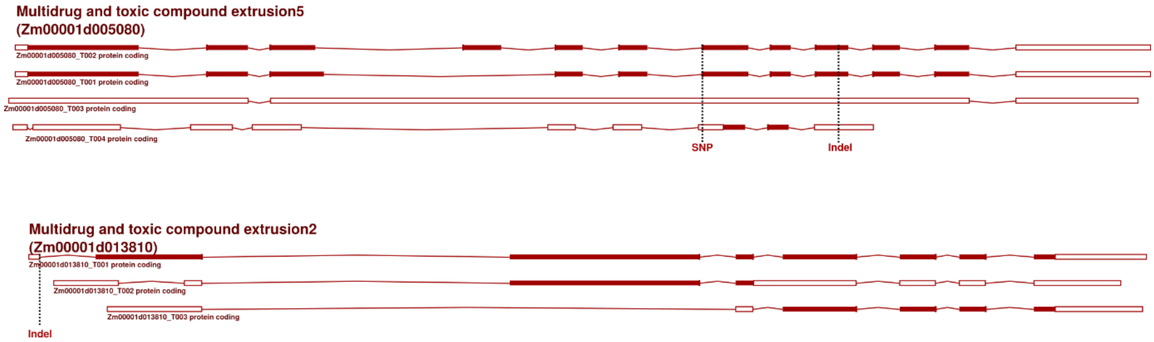
Multidrug and toxic compound extrusion5
(Zm00001d005080)
Zm00001d005080_T002 protein coding
Zm00001d005080_T001 protein coding
Zm00001d005080_T003 protein coding
Zm00001d005080_T004 protein coding
SNP
Indel
Multidrug and toxic compound extrusion2
(Zm00001d013810)
Zm00001d013810_T001 protein coding
Zm00001d013810_T002 protein coding
Zm00001d013810_T003 protein coding
Indel

Figure 5.30 Genetic variations in selected genes encoding proteins involved in multidrug and toxic compound extrusion.

6. Conclusions

The continuing reports about weeds gaining the glyphosate resistance and social defiance against genetically modified organisms are significant factors contributing to the urge of revealing the molecular mechanisms of natural glyphosate resistance, which could be gained by plants without use of genetic modifications or gene transfer. In my thesis, I have approached this problem by in-depth analysis of genetic differences between two natural (not genetically modified) maize lines that significantly differ in resistance to glyphosate. For this purpose, I have employed the genomic data obtained by two leading sequencing technologies – Illumina and PacBio.

Since PacBio sequencing reads are suffering from high error rate, the first step of my thesis was to find the best method for PacBio error correction using short, high accuracy illumina reads. Based on the obtained results, I have found out that the best correction performance is achieved by HALC program, whereas LoRDEC was the fastest tool and Proovread was producing most accurate short reads, though with low throughput. During examination of data correction tools, I have divided the output data into two types: split reads (where the uncorrected parts are removed) and full reads (which contain both corrected and uncorrected portions of the reads). In both, the split read and full length read, HALC was performing significantly better than other tools, regarding speed, accuracy and throughput. It was directly followed by LoRDEC.

Most of the tested tools either lack speed or efficiency. One has to take in consideration the continuous development in the real time sequencing technology, which will lead to higher throughput. New tools should be developed to achieve better speed, output and accuracy. Correction process could be enhanced by incorporation of self-correction process as a first step, especially if there is sufficient coverage, since the errors are randomly distributed with low biases towards GC rich regions.

Based on those observations, I have employed HALC to correct PacBio sequencing data from maize. To detect full range of differences between herbicide-tolerant and sensitive *Zea mays* lines, genetic variants were identified at two levels. First, large structure variations (SVs), using corrected PacBio reads were detected, next, single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) were revealed using Illumina short reads.

Before identifying structure variations, first I compared the influence of PacBio reads correction on sensitivity of SVs detection. As shown from the results, the employment of corrected PacBio reads provided significantly higher sensitivity than using raw uncorrected reads. Thus, I was able to detect 6062 large insertions 5110 large deletions, which differentiate the tolerant and the sensitive lines. Also, by analysis of high accuracy Illumina sequences, I was able to detect 4,068,829 SNPs and 729,866 indels that were specific to glyphosate-tolerant maize line. Such high number of variants between analyzed lines reflect their high genetic diversity. Additionally, one could expect that observed difference in sensitivity to glyphosate could be the effect of multiple changes in protein expression or function. Therefore, I have performed in-depth analysis of variations located within the genes which could potentially be involved in glyphosate resistance (candidate-genes).

The first gene, I have examined was gene encoding EPSPS protein, which is the direct target of glyphosate. Its inhibition lead to alteration of the shikimate pathway, which is crucial for plant survival. Although I have found 20 variations within this gene, all of them were predicted to have moderate or unknown (modifier) effect on gene expression. They were located in UTR regions and in close vicinity of the gene. Although, at this point, it is very hard to judge whether expression of EPSPS could be altered by those variations, I might conclude that the genetic bases of the tolerance in analyzed maize line is not directly connected with EPSPS protein.

I have found however, other genes of shikimate pathway, which were altered by variants predicted to have high impact on gene expression, encoding bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase and chorismate synthase (Figure 6.1). Those enzymes catalyze 3 steps of the shikimate pathway. Their alteration could have severe consequences, either limiting or inducing the efficiency of the pathway. If the latter case would be true, it could potentially compensate the decreased activity of the EPSPS caused by glyphosate. Similar effect could be gained by increased availability of the shikimate pathway substrate, phosphoenolpuryvate (PEP). As has been shown in rice, deactivation of phosphoenolpuryvate carboxylase, which is directing PEP into carbon fixation pathway, lead to substantial increase of shikimate pathway activity resulting in 50-60% increase of chorismate synthesis. In my study, I have found genes encoding phosphoenolpuryvate carboxylases and its regulator, phosphoenolpuryvate carboxylase kinase, to be affected by identified variants of high impact on gene expression (Figure 6.1). Although it is not possible to predict the effect of those variants on protein levels and activity, potential decrease of phosphoenolpuryvate carboxylase activity could lead to increased availability of PEP, providing another level of EPSP inhibition compensation.

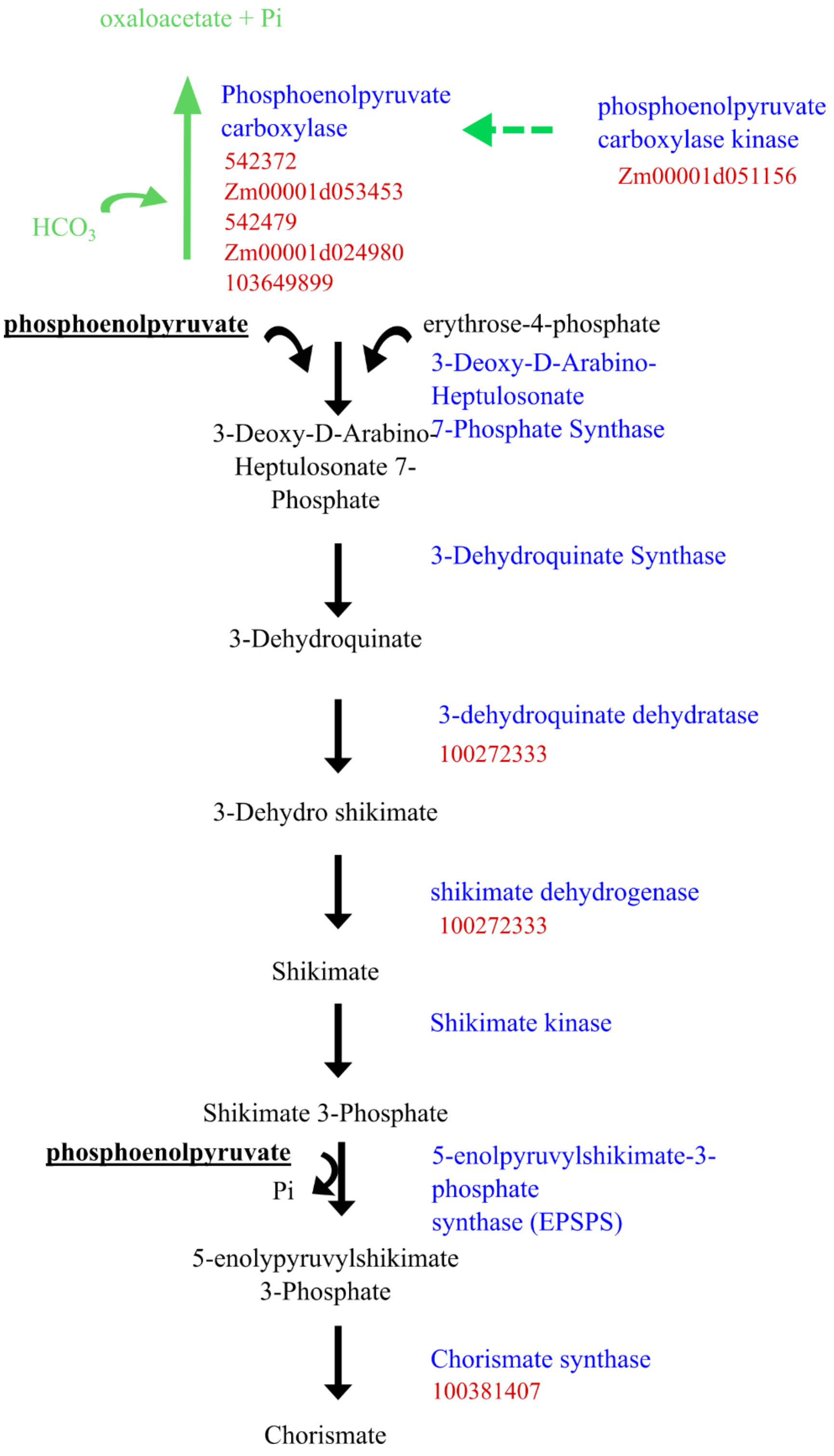

Figure 6.1 Shikimate pathway in plants. Enzymes catalyzing individual steps are shown in blue. IDs of maize genes affected by high impact variants revealed in my study are shown in red. Green color represent steps directing PEP into carbon fixation pathway.

Other possible scenario for observed glyphosate resistance is related to genes potentially involved in glyphosate transport. I have observed the high impact variants located within the genes encoding phosphate transporters and multidrug and toxic compound extrusion. Those proteins are suspected to be the major factors involved in active transport of glyphosate through the cell membrane. Thus, changes in their activity or expression could alter the effective intracellular concentration of glyphosate.

All the identified genes altered by identified variants should be however treated as potentially involved in glyphosate resistance phenotype. The existence of the variants differentiating glyphosate-tolerant and sensitive line within the described genes is however a strong indication for further biochemical studies, which in future will reveal the exact molecular mechanisms of this phenomenon.

Tables

Figures

References:

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., … Reddy, A. S. N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, *7*, 11706. http://doi.org/10.1038/ncomms11706

Abo, K. A., Fred-Jaiyesimi, A. A., & Jaiyesimi, A. E. A. (2008). Ethnobotanical studies of medicinal plants used in the management of diabetes mellitus in South Western Nigeria. *Journal of Ethnopharmacology*, *115*(1), 67–71. http://doi.org/10.1016/j.jep.2007.09.005

Ada Ching Mark Jung, Maurine Dolan, Oscar S (Howie) Smith, Scott Tingey, Michele Morgante and Antoni J Rafalski, K. S. C., Ching, A., Caldwell, K. S., Jung, M., Dolan, M., Smith, O. S., … Rafalski, A. J. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics*, *3*(19), 19. http://doi.org/http://www.biomedcentral.com/1471-2156/3/19

Adawy, S. S. M., Stupar, R. M., & Jiang, J. (2004). Fluorescence In Situ Hybridization Analysis Reveals Multiple Loci of Knob-associated DNA Elements in One-knob and Knobless Maize Lines. *Journal of Histochemistry and Cytochemistry*, *52*(8), 1113–1116. http://doi.org/10.1369/jhc.4B6335.2004

Alexandratos, N., & Bruinsma, J. (2012). WORLD AGRICULTURE TOWARDS 2030 / 2050 The 2012 Revision. *Food and Agriculture Organization of the United Nations*, (12), 146. http://doi.org/10.1016/S0264-8377(03)00047-4

Altshuler, D., Lander, E., & Ambrogio, L. (2010). A map of human genome variation from population scale sequencing. *Nature*, *476*(7319), 1061–1073. http://doi.org/10.1038/nature09534.A

Ambawat, S., Sharma, P., Yadav, N. R., & Yadav, R. C. (2013). MYB transcription factor genes as regulators for plant responses: an overview. *Physiology and Molecular Biology of Plants : An International Journal of Functional Plant Biology*, *19*(3), 307–21. http://doi.org/10.1007/s12298-013-0179-1

Arendt, E. K., & Zannini, E. (2013). *Maizes*. *Cereal Grains for the Food and Beverage Industries*. http://doi.org/10.1533/9780857098924.67

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., … Wong, W. H. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*, *110*(50), E4821–E4830. http://doi.org/10.1073/pnas.1320101110

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio Long Read Accuracy by Short Read Alignment. *PLoS ONE*, *7*(10), 1–8. http://doi.org/10.1371/journal.pone.0046679

Bai, L., Singh, M., Pitt, L., Sweeney, M., & Brutnel, T. P. (2007). Generating novel allelic variation through Activator insertional mutagenesis in maize. *Genetics*, *175*(3), 981–992. http://doi.org/10.1534/genetics.106.066837

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4), 333–337. http://doi.org/10.1038/nmeth.1935

Bansal, V., Harismendy, O., Tewhey, R., Murray, S. S., Schork, N. J., Topol, E. J., & Frazer, K. A. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Research*, *20*(4), 537–545. http://doi.org/10.1101/gr.100040.109

Bao, E., & Lan, L. (2017). HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics*, *18*(1), 204. http://doi.org/10.1186/s12859-017-1610-3

Barry, G., Kishore, G., Padgette, S. R., & Stallings, W. C. (1997). Glyphosate-Tolerant 5-enolpyruvylshikimate-3-phosphate synthases. *US Patent …*. Retrieved from http://www.google.com/patents?hl=en&lr=&vid=USPAT5633435&id=2EIaAAAAEBAJ&oi=fnd&dq=Glyphosate-Tolerant+5-enolpyruvylshikimate-3-phosphate+synthases&printsec=abstract

Becerril, J. M., Duke, S. O., & Lydon, J. (1989). Glyphosate effects on shikimate pathway products in leaves and flowers of velvetleaf. *Phytochemistry*, *28*(3), 695–699. http://doi.org/10.1016/0031-9422(89)80095-0

Beckie, H. J., Warwick, S. I., Hall, L. M., & Neil Harker, K. (2012). Pollen-mediated gene flow in wheat fields in Western Canada. *AgBioForum*, *15*(1), 36–43. http://doi.org/10.2135/cropsci2010.03.0176

Bennetzen, J. L. (2000). Comparative Sequence Analysis of Plant Nuclear Genomes: Microcolinearity and Its Many Exceptions. *The Plant Cell*, *12*(July), 1021–1030. http://doi.org/10.1105/tpc.12.7.1021

Bennetzen, J. L., & Hake, S. (2009). *Handbook of maize: Genetics and genomics*. *Handbook of Maize: Genetics and Genomics*. http://doi.org/10.1007/978-0-387-77863-1

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. http://doi.org/10.1038/nature07517

Bentley, R., & Haslam, E. (1990). The Shikimate Pathway — A Metabolic Tree with Many Branche. *Critical Reviews in Biochemistry and Molecular Biology*, *25*(5), 307–384. http://doi.org/10.3109/10409239009090615

Bhatt, A. M., Zhang, Q., Harris, S. A., White-Cooper, H., & Dickinson, H. (2004). Gene structure and molecular analysis of Arabidopsis thaliana ALWAYS EARLY homologs. *Gene*, *336*(2), 219–229. http://doi.org/10.1016/j.gene.2004.03.033

Bosnic, A., & Swanton, C. (1997). Influence of barnyardgrass (Echinochloa crus-galli) time of emergence and density on corn (Zea mays). *Weed Science*, *45*(2), 276–282.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., … Jaffe, D. B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, *18*(5), 763–70. http://doi.org/10.1101/gr.070227.107

Browning, B. L., & Yu, Z. (2009). Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *American Journal of Human Genetics*, *85*(6), 847–861. http://doi.org/10.1016/j.ajhg.2009.11.004

Buckler, E. S., Gaut, B. S., & McMullen, M. D. (2006). Molecular and functional diversity of maize. *Current Opinion in Plant Biology*. http://doi.org/10.1016/j.pbi.2006.01.013

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., … Yandell, M. (2014). MAKER-P : A Tool Kit for the Rapid Creation , Management , and Quality Control of Plant. *Plant Physiology*, *164*(February), 513–524. http://doi.org/10.1104/pp.113.230144

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, *13*(1), 238. http://doi.org/10.1186/1471-2105-13-238

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. http://doi.org/10.1038/nmeth.2474

Coe, E. H. (2001). The origins of maize genetics. *Nature Reviews. Genetics*, *2*(11), 898–905. http://doi.org/10.1038/35098524

Comai, L., Sen, L. C., & Stalker, D. M. (1983). An Altered aroA Gene Product Confers Resistance to the Herbicide Glyphosate. *Science (New York, N.Y.)*, *221*(4608), 370–371. http://doi.org/10.1126/science.221.4608.370

Cruz-Reyes, R., ?vila-Sakar, G., S?nchez-Montoya, G., & Quesada, M. (2015). Experimental assessment of gene flow between transgenic squash and a wild relative in the center of origin of cucurbits. *Ecosphere*, *6*(12), art248. http://doi.org/10.1890/ES15-00304.1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. http://doi.org/10.1093/bioinformatics/btr330

de Souza, C. P., Guedes, T. de A., & Fontanetti, C. S. (2016). Evaluation of herbicides action on plant bioindicators by genetic biomarkers: a review. *Environmental Monitoring and Assessment*, *188*(12), 694. http://doi.org/10.1007/s10661-016-5702-8

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Dill, G. M., CaJacob, C. A., & Padgette, S. R. (2008). Glyphosate-resistant crops: Adoption, use and future considerations. In *Pest Management Science* (Vol. 64, pp. 326–331). http://doi.org/10.1002/ps.1501

Duke, Stephen; Powles, S. (2008). Glyphosate: a once in a century herbicide. *Pest Management Science*, *64*, 319–325. http://doi.org/10.1002/ps.1218

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133–8. http://doi.org/10.1126/science.1162986

English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., … Gibbs, R. A. (2015). Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics*, *16*(1), 286. http://doi.org/10.1186/s12864-015-1479-3

English, A. C., Salerno, W. J., & Reid, J. G. (2014). PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, *15*(1), 180. http://doi.org/10.1186/1471-2105-15-180

Evans, M. M. S. (2007). The indeterminate gametophyte1 Gene of Maize Encodes a LOB Domain Protein Required for Embryo Sac and Leaf Development. *THE PLANT CELL ONLINE*, *19*(1), 46–62. http://doi.org/10.1105/tpc.106.047506

FAO. (2016). Food and Agriculture Organization of United Nations. Retrieved from http://faostat.fao.org

FAOSTAT. (2014). FAOSTAT. *Food and Agricultural Organization of the United Nations*. Retrieved from http://data.fao.org/ref/262b79ca-279c-4517-93de-ee3b7c7cb553.html?version=1.0

Fang, L., Hu, J., Wang, D., & Wang, K. (2016). Evaluation on Detection of Structural Variants by Low-Coverage Long-Read Sequencing. *bioRxiv*, 92544. http://doi.org/10.1101/092544

Fernando, N., Manalil, S., Florentine, S. K., Chauhan, B. S., & Seneweera, S. (2016). Glyphosate Resistance of C3 and C4 Weeds under Rising Atmospheric CO2. *Frontiers in Plant Science*, *7*. http://doi.org/10.3389/fpls.2016.00910

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant Transposable Elements: Where Genetics Meets Genomics. *Nature Reviews Genetics*, *3*(5), 329–341. http://doi.org/10.1038/nrg793

Frampton, M., Houlston, R., Gardet, A., Stevens, C., & Sharma, Y. (2012). Generation of Artificial FASTQ Files to Evaluate the Performance of Next-Generation Sequencing Pipelines. *PLoS ONE*, *7*(11), e49110. http://doi.org/10.1371/journal.pone.0049110

Fu, Y., Wen, T.-J., Ronin, Y. I., Chen, H. D., Guo, L., Mester, D. I., … Schnable, P. S. (2006). Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. *Genetics*, *174*(3), 1671–83. http://doi.org/10.1534/genetics.106.060376

Funke, T., Han, H., Healy-Fried, M. L., Fischer, M., & Schönbrunn, E. (2006). Molecular basis for the herbicide resistance of Roundup Ready crops. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(35), 13010–5. http://doi.org/10.1073/pnas.0603638103

Funke, T., Yang, Y., Han, H., Healy-Fried, M., Olesen, S., Becker, A., & Schonbrunn, E. (2009). Structural Basis of Glyphosate Resistance Resulting from the Double Mutation Thr97 -&gt; Ile and Pro101 -&gt; Ser in 5-Enolpyruvylshikimate-3-phosphate Synthase from Escherichia coli. *Journal of Biological Chemistry*, *284*(15), 9854–9860. http://doi.org/10.1074/jbc.M809771200

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Retrieved from http://arxiv.org/abs/1207.3907

Genome 10K Community of Scientists. (2009). Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity*, *100*(6), 659–674. http://doi.org/10.1093/jhered/esp086

Gordon, D., Huddleston, J., Chaisson, M. J., Hill, C. M., Kronenberg, Z. N., Munson, K. M., … Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science (New York, N.Y.)*, *352*, aae0344. http://doi.org/10.1126/science.aae0344

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., … Wang, Z. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS ONE*, *10*(7). http://doi.org/10.1371/journal.pone.0132628

Gowik, U., & Westhoff, P. (2011). The path from C3 to C4 photosynthesis. *Plant Physiology*, *155*(1), 56–63. http://doi.org/10.1104/pp.110.165308

Hackl, T., Hedrich, R., Schultz, J., & F??rster, F. (2014). Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, *30*(21), 3004–3011. http://doi.org/10.1093/bioinformatics/btu392

Han, Y., Qin, S., & Wessler, S. R. (2013). Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics*, *14*(1), 71. http://doi.org/10.1186/1471-2164-14-71

Hanjra, M. A., & Qureshi, M. E. (2010). Global water crisis and future food security in an era of climate change. *Food Policy*, *35*(5), 365–377. http://doi.org/10.1016/j.foodpol.2010.05.006

Harhay, G. P., Koren, S., Phillippy, A. M., Mcvey, D. S., Kuszak, J., Clawson, M. L., … Smith, T. P. L. (2013). Complete Closed Genome Sequences of Mannheimia haemolytica Serotypes A1 and A6, Isolated from Cattle. *Genome Announc*, *1*(3), 188–13. http://doi.org/10.1128/genomeA.00188-13

Healy-Fried, M. L., Funke, T., Priestman, M. A., Han, H., & Schonbrunn, E. (2007). Structural Basis of Glyphosate Tolerance Resulting from Mutations of Pro101 in Escherichia coli 5-Enolpyruvylshikimate-3-phosphate Synthase. *Journal of Biological Chemistry*, *282*(45), 32949–32955. http://doi.org/10.1074/jbc.M705624200

Hellsten, I. (2006). Focus On Metaphors: The Case Of “Frankenfood” On The Web. *Journal of Computer-Mediated Communication*, *8*(4), 0–0. http://doi.org/10.1111/j.1083-6101.2003.tb00218.x

Hoberman, R., Dias, J., Ge, B., Harmsen, E., Mayhew, M., Verlaan, D. J., … Pastinen, T. (2009). A probabilistic approach for SNP discovery in high-throughput human resequencing data. *Genome Research*, *19*(9), 1542–1552. http://doi.org/10.1101/gr.092072.109

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends in Genetics : TIG*, *24*(5), 238–45. http://doi.org/10.1016/j.tig.2008.03.001

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*(December), 17875. http://doi.org/10.1038/srep17875

International Rice Genome Sequencing Project, Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., … Wilson, R. K. (2005). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Nature*, *326*(5956), 1112–1115. http://doi.org/10.1038/nature03895

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., … Ware, D. (2017). Improved maize reference genome with. *Nature*. http://doi.org/10.1038/nature22971

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., … Turro, N. J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(52), 19635–19640. http://doi.org/10.1073/pnas.0609513103

Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. http://doi.org/10.1101/gr.229202. Article published online before March 2002

Kilpinen, H., & Barrett, J. C. (2013). How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics*. http://doi.org/10.1016/j.tig.2012.10.001

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, *8*(9). http://doi.org/10.1371/journal.pbio.1000501

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. http://doi.org/10.1093/bioinformatics/btp373

Kolkman, J. M., Conrad, L. J., Farmer, P. R., Hardeman, K., Ahern, K. R., Lewis, P. E., … Brutnell, T. P. (2005). Distribution of Activator (Ac) throughout the maize genome for use in regional mutagenesis. *Genetics*, *169*(2), 981–995. http://doi.org/10.1534/genetics.104.033738

Koren, S., Harhay, G. P., Smith, T. P. L., Bono, J. L., Harhay, D. M., Mcvey, S. D., … Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, *14*(9), R101. http://doi.org/10.1186/gb-2013-14-9-r101

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, *23*, 110–120. http://doi.org/10.1016/j.mib.2014.11.014

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., … Adam M Phillippy. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. http://doi.org/10.1038/nbt.2280

Lander, E. S., Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., & Linton, L. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*(6803), 513–516. http://doi.org/10.1038/35035083

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. http://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, *10*(3), R25. http://doi.org/10.1186/gb-2009-10-3-r25

Law, M., Childs, K. L., Campbell, M. S., Stein, J. C., Olson, A. J., Holt, C., … Yandell, M. (2015). Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. *Plant Physiology*, *167*(1), 25–39. http://doi.org/10.1104/pp.114.245027

Le, S. Q., & Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, *21*(6), 952–960. http://doi.org/10.1101/gr.113084.110

Lee, H., Gurtowski, J., & Yoo, S. (2014). Error correction and assembly complexity of single molecule sequencing reads. *bioRxiv*, 1–17. http://doi.org/10.1101/006395

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*, *299*(5607), 682–686. http://doi.org/10.1126/science.1079700

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv*, *0*(0), 3. http://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. http://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, *21*(6), 940–951. http://doi.org/10.1101/gr.117259.110

Lin, K., Bonnema, G., Sanchez-Perez, G., & De Ridder, D. (2014). Making the difference: Integrating structural variation detection tools. *Briefings in Bioinformatics*, *16*(5), 852–864. http://doi.org/10.1093/bib/bbu047

Liu, J., He, Y., Amasino, R., & Chen, X. (2004). siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes & Development*, *18*(23), 2873. http://doi.org/10.1101/gad.1217304

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*. http://doi.org/10.1155/2012/251364

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., … Nordborg, M. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature Genetics*, *45*(8), 884–890. http://doi.org/10.1038/ng.2678

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. http://doi.org/10.1016/j.gpb.2016.05.004

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A. J., Li, T., & Ma, H. (2012). Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Research*, *22*(3), 508–18. http://doi.org/10.1101/gr.127522.111

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. http://doi.org/10.1186/2047-217X-1-18

M. J. Kropff, F. J. H. V. and C. J. T. S. (. (1984). Competition between a maize crop and a natural population of Echinochloa crus-gali. *Netherlands Journal of Agricultural Science*, *32*(November), 324–327.

Masumoto, C., Miyazawa, S.-I., Ohkawa, H., Fukuda, T., Taniguchi, Y., Murayama, S., Miyao, M. (2010). Phosphoenolpyruvate carboxylase intrinsically located in the chloroplast of rice plays a crucial role in ammonium assimilation. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(11), 5226–31. http://doi.org/10.1073/pnas.0913127107

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G., J., Buckler, E., & Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences*, *99*(9), 6080–6084. http://doi.org/10.1073/pnas.052125199

Maxam, a M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–4. http://doi.org/10.1073/pnas.74.2.560

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. http://doi.org/10.1101/gr.107524.110

Meisel, L., & Lam, E. (1996). The conserved ELK-homeodomain of KNOTTED-1 contains two regions that signal nuclear localization. *Plant Molecular Biology*, *30*(1), 1–14. http://doi.org/10.1007/BF00017799

Messing, J., & Dooner, H. K. (2006). Organization and variability of the maize genome. *Current Opinion in Plant Biology*. http://doi.org/10.1016/j.pbi.2006.01.009

Minichiello, M. J., & Durbin, R. (2006). Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs. *The American Journal of Human Genetics*, *79*(5), 910–922. http://doi.org/10.1086/508901

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., … Nakamura, S. (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, *15*(1), 688–675. http://doi.org/10.1186/1471-2164-15-699

Moose, S. P., Dudley, J. W., & Rocheford, T. R. (2004). Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends in Plant Science*, *9*(7), 358–364. http://doi.org/10.1016/j.tplants.2004.05.005

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., & Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*, *37*(9), 997–1002. http://doi.org/10.1038/ng1615

Mu, J. C., Jiang, H., Kiani, A., Mohiyuddin, M., Asadi, N. B., & Wong, W. H. (2012). Fast and accurate read alignment for resequencing. *Bioinformatics*, *28*(18), 2366–2373. http://doi.org/10.1093/bioinformatics/bts450

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., … Venter, J. C. (2000). A Whole-Genome Assembly of Drosophila. *Science*, *287*(5461). Retrieved from http://science.sciencemag.org/content/287/5461/2196

Myers, G. (2014). Efficient local alignment discovery amongst noisy long reads. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8701 LNBI, pp. 52–67). http://doi.org/10.1007/978-3-662-44753-6_5

Nagasaki, H., Sakamoto, T., Sato, Y., & Matsuoka, M. (2001). Functional analysis of the conserved domains of a rice KNOX homeodomain protein, OSH15. *The Plant Cell*, *13*(9), 2085–98. http://doi.org/10.1105/TPC.13.9.2085

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Methods Mol Biol*, *628*, 215–226. http://doi.org/10.1007/978-1-60327-367-1_12

Nuss, E. T., & Tanumihardjo, S. A. (2010). Maize: A paramount staple crop in the context of global nutrition. *Comprehensive Reviews in Food Science and Food Safety*, *9*(4), 417–436. http://doi.org/10.1111/j.1541-4337.2010.00117.x

OERKE, E.-C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, *144*(1), 31. http://doi.org/10.1017/S0021859605005708

Ono, Y., Asai, K., & Hamada, M. (2013). PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics*, *29*(1), 119–121. http://doi.org/10.1093/bioinformatics/bts649

Owoyele, B. V, Negedu, M. N., Olaniran, S. O., Onasanwo, S. a, Oguntoye, S. O., Sanya, J. O., … Soladoye, A. O. (2010). Analgesic and anti-inflammatory effects of aqueous extract of Zea mays husk in male Wistar rats. *Journal of Medicinal Food*, *13*(2), 343–347. http://doi.org/10.1089/jmf.2008.0311

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics*, *10*(10), 669–80. http://doi.org/10.1038/nrg2641

Park, S. T., & Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *International Neurourology Journal*, *20*(Suppl 2), S76-83. http://doi.org/10.5213/inj.1632742.371

Paterson, A. H., Bowers, J. E., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences*, *101*(26), 9903–9908. http://doi.org/10.1073/pnas.0307901101

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*(8), 780–786. http://doi.org/10.1038/nmeth.3454

Piperno, D. R., & Flannery, K. V. (2001). The earliest archaeological maize (Zea mays L.) from highland Mexico : New accelerator mass spectrometry. *Archaeology*, *98*(4), 2101–2103.

Piperno, D. R., Moreno, J. E., Iriarte, J., Holst, I., Lachniet, M., Jones, J. G., … Castanzo, R. (2007). Late Pleistocene and Holocene environmental history of the Iguala Valley, Central Balsas Watershed of Mexico. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11874–11881. http://doi.org/10.1073/pnas.0703442104

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14. http://doi.org/10.1186/1479-7364-8-14

Pline, W. A., Wilcut, J. W., Duke, S. O., Edmisten, K. L., & Wells, R. (2002). Tolerance and accumulation of shikimic acid in response to glyphosate applications in glyphosate-resistant and nonglyphosate-resistant cotton (Gossypium hirsutum L.). *Journal of Agricultural and Food Chemistry*, *50*(3), 506–12. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11804521

Pollegioni, L., Schonbrunn, E., & Siehl, D. (2011). Molecular basis of glyphosate resistance-different approaches through protein engineering. *The FEBS Journal*, *278*(16), 2753–66. http://doi.org/10.1111/j.1742-4658.2011.08214.x

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, *27*(9), 847–50. http://doi.org/10.1038/nbt.1561

Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., … Waldor, M. K. (2011). Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany. *New England Journal of Medicine*, *365*(8), 709–717. http://doi.org/10.1056/NEJMoa1106920

Rayburn, A. L., Biradar, D. P., Bullock, D. G., & McMurphy, L. M. (1993). Nuclear DNA content in F1 hybrids of maize. *Heredity*, *70*(3), 294–300. http://doi.org/10.1038/hdy.1993.42

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. http://doi.org/10.1016/j.gpb.2015.08.002

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., & Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, *30*(24), 3458–3466. http://doi.org/10.1093/bioinformatics/btu714

Rizwan, M., Ali, S., Qayyum, M. F., Ok, Y. S., Zia-ur-Rehman, M., Abbas, Z., & Hannan, F. (2016). Use of Maize (Zea mays L.) for phytomanagement of Cd-contaminated soils: a critical review. *Environmental Geochemistry and Health*, pp. 1–19. http://doi.org/10.1007/s10653-016-9826-0

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., & Buckler, E. S. (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(22), E3177-84. http://doi.org/10.1073/pnas.1525244113

Roy, S. (2016). Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. *Plant Signaling & Behavior*, *11*(1), e1117723. http://doi.org/10.1080/15592324.2015.1117723

Salmela, L., & Rivals, E. (2014). LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, *30*(24), 3506–3514. http://doi.org/10.1093/bioinformatics/btu538

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2). http://doi.org/10.1093/hmg/ddq416

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*. http://doi.org/10.1101/gr.101360.109

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., … Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, *39*(7 Suppl), S7–S15. http://doi.org/10.1038/ng2093

Schmidt, J. M., Good, R. T., Appleton, B., Sherrard, J., Raymant, G. C., Bogwitz, M. R., … Robin, C. (2010). Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genetics*, *6*(6), 1–11. http://doi.org/10.1371/journal.pgen.1000998

Schnable, J. C., & Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS ONE*, *6*(3). http://doi.org/10.1371/journal.pone.0017855

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112–1115. http://doi.org/10.1126/science.1178534

Schönbrunn, E., Eschenburg, S., Shuttleworth, W. A., Schloss, J. V, Amrhein, N., Evans, J. N., & Kabsch, W. (2001). Interaction of the herbicide glyphosate with its target enzyme 5-enolpyruvylshikimate 3-phosphate synthase in atomic detail. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1376–80. http://doi.org/10.1073/pnas.98.4.1376

Sebat, J. (2007). Major changes in our DNA lead to major changes in our thinking. *Nature Genetics*, *39*(7 Suppl), S3-5. http://doi.org/10.1038/ng2095

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. (2017). Accurate detection of complex structural variations using single molecule sequencing. *Doi.org*, 169557. http://doi.org/10.1101/169557

Selb, R., Wal, J. M., Moreno, F. J., Lovik, M., Mills, C., Hoffmann-Sommergruber, K., & Fernandez, A. (2017). Assessment of endogenous allergenicity of genetically modified plants exemplified by soybean – Where do we stand? *Food and Chemical Toxicology*, *101*, 139–148. http://doi.org/10.1016/j.fct.2017.01.014

Sharma, M., & Pandey, G. K. (2015). Expansion and Function of Repeat Domain Proteins During Stress and Development in Plants. *Frontiers in Plant Science*, *6*, 1218. http://doi.org/10.3389/fpls.2015.01218

Sharples, F. E. (1983). Spread of organisms with novel genotypes: thoughts from an ecological perspective. *Recomb DNA Tech Bull*, *6*(2), 43–56.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., … Yu, F. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research*, *20*(2), 273–280. http://doi.org/10.1101/gr.096388.109

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., … Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, *309*(5741), 1728–32. http://doi.org/10.1126/science.1117389

Smith, L. G., Greene, B., Veit, B., & Hake, S. (1992). A dominant mutation in the maize homeobox gene, Knotted-1, causes its ectopic expression in leaf cells with altered fates. *Development (Cambridge, England)*, *116*(1), 21–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1362381

Sost, D., & Amrhein, N. (1990). Substitution of Gly-96 to Ala in the 5-enolpyruvylshikimate-3-phosphate synthase of Klebsiella pneumoniae results in a greatly reduced affinity for the herbicide glyphosate. *Archives of Biochemistry and Biophysics*, *282*(2), 433–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2241161

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., … Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, *5*(11), e1000734. http://doi.org/10.1371/journal.pgen.1000734

Stallings, W. C., Abdel-Meguid, S. S., Lim, L. W., Shieh, H. S., Dayringer, H. E., Leimgruber, N. K., … Kishore, G. M. (1991). Structure and topological symmetry of the glyphosate target 5-enolpyruvylshikimate-3-phosphate synthase: a distinctive protein fold. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 5046–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11607190

Strable, J., & Scanlon, M. J. (2009). Maize (Zea mays): A model organism for basic and applied research in plant biology. *Cold Spring Harbor Protocols*, *4*(10), 1–10. http://doi.org/10.1101/pdb.emo132

Stupar, R. M., & Springer, N. M. (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, *173*(4), 2199–2210. http://doi.org/10.1534/genetics.106.060699

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, *6*(7), e21800. http://doi.org/10.1371/journal.pone.0021800

Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., & Messing, J. (2004). Close split of sorghum and maize genome progenitors. *Genome Research*, *14*(10 A), 1916–1923. http://doi.org/10.1101/gr.2332504

Tenaillon, M. I., & Charcosset, A. (2011). A European perspective on maize history. *Comptes Rendus - Biologies*. http://doi.org/10.1016/j.crvi.2010.12.015

Tilgner, H., Grubert, F., Sharon, D., & Snyder, M. P. (2014). Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proceedings of the National Academy of Sciences*, *111*(27), 9869–9874. http://doi.org/10.1073/pnas.1400447111

Treutlein, B., Gokce, O., Quake, S. R., & Südhof, T. C. (2014). Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proceedings of the National Academy of Sciences*, *111*(13), E1291–E1299. http://doi.org/10.1073/pnas.1403244111

Tsatsakis, A. M., Nawaz, M. A., Kouretas, D., Balias, G., Savolainen, K., Tutelyan, V. A., … Chung, G. (2017). Environmental impacts of genetically modified plants: A review. *Environmental Research*, *156*, 818–833. http://doi.org/10.1016/j.envres.2017.03.011

Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A., & Brown, S. D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, *30*(19), 2709–2716. http://doi.org/10.1093/bioinformatics/btu391

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*. http://doi.org/10.1016/j.tig.2014.07.001

Vandenberg, L. N., Blumberg, B., Antoniou, M. N., Benbrook, C. M., Carroll, L., Colborn, T., … Myers, J. P. (2017). Is it time to reassess current safety standards for glyphosate-based herbicides? *Journal of Epidemiology and Community Health*, jech-2016-208463. http://doi.org/10.1136/jech-2016-208463

Vlad, D., Rappaport, F., Simon, M., & Loudet, O. (2010). Gene transposition causing natural variation for growth in Arabidopsis thaliana. *PLoS Genetics*, *6*(5), 21. http://doi.org/10.1371/journal.pgen.1000945

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., … Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, *7*, 11708. http://doi.org/10.1038/ncomms11708

Wang, H., & Bennetzen, J. L. (2012). Centromere retention and loss during the descent of maize from a tetraploid ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(51), 21004–9. http://doi.org/10.1073/pnas.1218668109

Wang, J. J., Wang, W., Li, R., Li, Y., Tian, G., Fan, W., … Wang, J. J. (2008). The diploid genome sequence of an Asian individual. *Nature*, *456*(7218), 60–5. http://doi.org/10.1038/nature07484

Wang, Q., & Dooner, H. K. (2006). Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17644–9. http://doi.org/10.1073/pnas.0603080103

Wei, F., Zhang, J., Zhou, S., He, R., Schaeffer, M., Collura, K., … Wing, R. A. (2009). The Physical and Genetic Framework of the Maize B73 Genome. *PLoS Genetics*, *5*(11), e1000715. http://doi.org/10.1371/journal.pgen.1000715

Wendel, J. F., Stuber, C. W., Edwards, M. D., & Goodman, M. M. (1986). Duplicated chromosome segments in maize (Zea mays L.): further evidence from hexokinase isozymes. *Theoretical and Applied Genetics*, *72*(2), 178–185. http://doi.org/10.1007/BF00266990

Worldhunger. (2015). 2015 World Hunger and Poverty Facts and Statistics by WHES. *World Hunger*, 1–7. Retrieved from http://www.worldhunger.org/articles/Learn/world hunger facts 2002.htm

Yang, L., & Bennetzen, J. L. (2009). Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences*, *106*(31), 12832–12837. http://doi.org/10.1073/pnas.0905563106

Ye, C., Hill, C., Koren, S., Ruan, J., Zhanshan, Ma, … Zimin, A. (2015). DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph. *arXiv*, 1410.2801v2. Retrieved from http://arxiv.org/abs/1410.2801

Yu, H.-L., Li, Y.-H., & Wu, K.-M. (2011). Risk Assessment and Ecological Effects of Transgenic Bacillus thuringiensis Crops on Non-Target Organisms(F). *Journal of Integrative Plant Biology*, *53*(7), 520–38. http://doi.org/10.1111/j.1744-7909.2011.01047.x

Yu, Q., Cairns, A., & Powles, S. (2007). Glyphosate, paraquat and ACCase multiple herbicide resistance evolved in a Lolium rigidum biotype. Planta, 225(2), 499–513. http://doi.org/10.1007/s00425-006-0364-3

Żywicki, M., Gracz, J., Karłowski, W., Twardowski, T., & Tyczewska, A. (2015). Expression of miRNAs involved in phosphate homeostasis and senescence is altered in glyphosate-treated maize. Acta Physiologiae Plantarum, 37(12), 265. http://doi.org/10.1007/s11738-015-2022-5