
CrossMark

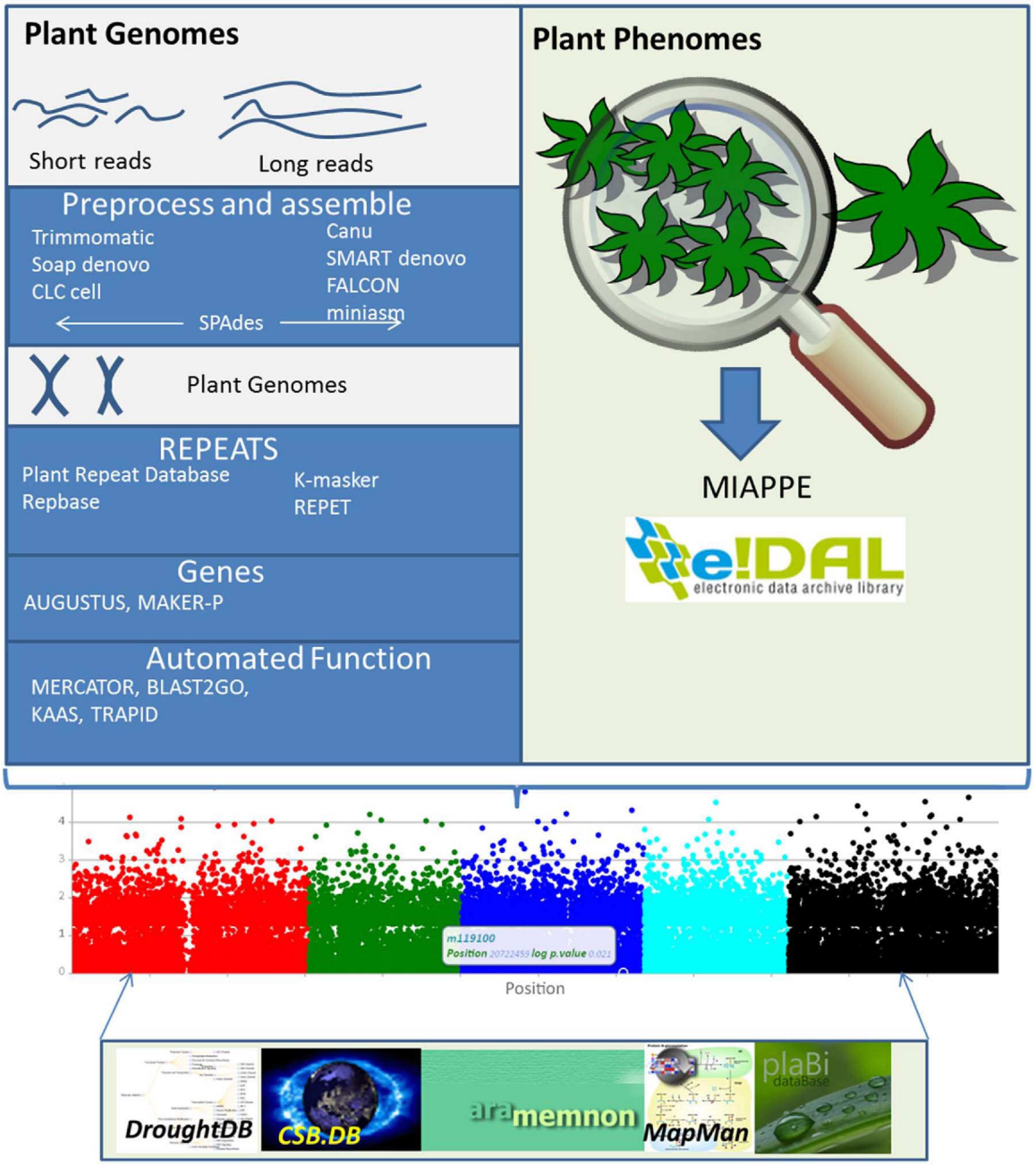

Fig. 1. From genomes and phenomes to candidate genes: Fig. 1 shows which tools one can typically use to assemble and annotate plant genomes and what standards are being developed for plant phenotyping.
The analysis of genomes and phenotypes leads to candidate regions which can be further delineated.

Furthermore, plant sciences feature several communities with specific needs and interests. As an example, a researcher working on sugar beet (Dohm et al., 2014) or carrots (Iorizzo et al., 2016) will be more interested in below ground organs and their development than a researcher working on tomatoes or barley. However, a common ground is that the plants are (also) looked at from a breeding perspective in order to improve plant yield and/or resilience. This process is greatly expedited by the development of statistical analysis and model based analysis of plant genetic (genomic) and phenotypic data (Hammer et al., 2006) as well as the maturation and development of plant phenotyping technology (Fiorani and Schurr, 2013).

Within this review we attempt to shed light on the analysis of plant genomes, describe current problems as well as how plant genomes can be best leveraged in conjunction with high throughput phenotyping to accelerate selective plant breeding.

We provide a detailed list of tools which can be used in the process of genome assembly, annotation and linking it to phenotypic plant data.

2. *De novo* genome assembly

Plant *de-novo* genome assembly is notoriously difficult (Claros et al., 2012), mainly due to the problems mentioned above. This has prompted the development of tools which can cope with these difficulties, some of which also serve the wider scientific community. A notable example of this is the raw data preprocessing tools 'Trimmomatic' http://www.usadellab.org/cms/?page=trimmomatic (Bolger et al., 2014b) which was developed due to the necessity for a highly efficient adapter trimmer during a plant genome sequencing project and has since been widely adopted by the whole scientific community due to its flexibility, speed and efficiency.

After read data preprocessing, error correction of reads is frequently carried out which can either be done by the assembler tool itself, as in the case of Canu https://github.com/marbl/canu (Koren et al., 2017) or SOAP denovo http://soap.genomics.org.cn/soapdenovo.html (Luo et al., 2012), or by using separate tools such as those reviewed in (Yang et al., 2013) or by using graph based analysis (Joppich et al., 2015). The later graph based approach demonstrated that such an analysis would correct reads which are missed by other typical frameworks which employ kmer counting methods.

The assembling of sequencing data is often undertaken as a 'trial and error' approach. The available tools each have their own strengths and weakness and multiple runs whilst optimizing parameters is often necessary to produce the best assembly. The source of the input data can also determine which assembler is used. There are three main aspects which differentiate the different assemblers; 1) the underlying algorithm used 2) how aggressive the tool is at dealing with false positives/negatives 3) the heuristics which are employed by the assembler. Each of these aspects combine to produce vastly different results which are not guaranteed to be correct. Determining how to assess the correctness of the output from an assembler is also challenging and has been covered by the Assemblathon challenge (DOI: 10.1186/2047-217X-2-10).

Assemblers which have so far been used in projects associated with plant genomes include commonly used short read assemblers such as ABySS http://www.bcgsc.ca/platform/bioinfo/software/abyss (Simpson et al., 2009), DISCOVAR (*de novo*) https://software.broadinstitute.org/software/discovar/blog/(Weisenfeld et al., 2014) and velvet https://www.ebi.ac.uk/~zerbino/velvet/ (Zerbino and Birney, 2008). Long reads have been assembled using assemblers such as Canu, SMART denovo https://github.com/ruanjue/smartdenovo and Falcon https://github.com/PacificBiosciences/FALCON as well as Hybrid assemblers such as SPAdes http://bioinf.spbau.ru/spades (Bankevich et al., 2012). Additionally, a modified version of SOAPdenovo was used to assemble the genome of a wild tomato (Bolger et al., 2014a). Even commercial software, such as the CLC assembly cell or NRGene can sometimes be helpful for plant genome assemblies (Bauer et al., 2017; International Barley Genome Sequencing et al., 2012). The wide variety of assembler software used reflects the diversity encountered when assembling plant genomes. This is aided by assembly strategies relying on a very large set of genetic markers guiding the assembly process (Hirsch et al., 2016) and experimental techniques which provide large scale proximity information (Kalhor et al., 2011; Mascher et al., 2017). As assemblers and sequencing technologies improve further, we can expect a greater number of high quality genome assemblies from large and complex plant genomes. These are expected to pose new challenges for downstream bioinformatic processing in regard to management, annotation, comparative analyses and visualization.

3. Genome annotation

Assembling a high quality genome is indeed an arduous task, but is still only the first step to providing meaningful biological data. Gigabases of nucleotide data without any form of interpretation is useful only in very niche scientific pursuits. To fully realize the value of a genome assembly, it must undergo a process known as annotation. This process can be further characterized into structural annotation, whereby the structures of genomic features are delineated, and functional annotation, whereby functions are ascribed to these structures.

The main structures of interest in genomes are typically genes which provide the core instructions which all organisms depend on. Typical gene finding tools such as AUGUSTUS http://augustus.gobics.de/ (Stanke and Waack, 2003) and/or the plant adapted MAKER-P http://www.yandell-lab.org/software/maker-p.html (Campbell et al., 2014) will annotate a genome with or (sometimes) without external evidence. Despite recent efforts to make these tools more automated and user friendly http://exon.gatech.edu/braker1.html (Hoff et al., 2016), obtaining optimal results still requires a certain level of expertise. Incorporating extrinsic RNAseq evidence generally results in superior results, but care needs to be taken not to overlook real genes (false negatives) or mispredict genes (false positives).

Indeed, due to the ever increasing demand for gene finding, GCBN (the Germany BioGreenformatics Network) at HMGU is developing a fully automated pipeline for plant gene finding.

Another annotation strategy exploits the fact that closely related plant species often retain gene order, i.e. synteny, which can be leveraged for gene finding and prediction purposes. This strategy relies on structured databases or tools such as those offered by CoGe (https://genomevolution.org/coge/) (Tang et al., 2015). For example, CoGe offers "SynMap" to compare syntenic regions and *e.g.* SynFind to find homologs. Another particular successful example is the GenomeZipper developed originally for barley (Mayer et al., 2009) and has not only been used for other Triticae like rye (Martis et al., 2013) but the approach was even adopted by other groups for more exotic species such as the mulberry tree (He et al., 2013).

Another major task in genome structural annotation involves the identification of transposons and other repetitive elements. This is particular relevant for plants as many (large) genomes are extensively composed of transposable elements and/or their relics. These have been calculated at 3.7 Gbp (81%) of mainly retrotransposons in the most recent 4.6 Gbp assembly of barley (Mascher et al., 2017). In order to address this challenge, one can detect some of them *de-novo* using *e.g.* the GCBN-IPK developed K-masker (Schmutzer et al., 2014) or the REPET package https://urgi.versailles.inra.fr/Tools/REPET (Flutre et al., 2011), but this needs to be complemented by tapping into existing plant data projects. For that purpose GCBN-HMGU collects, curates and classifies plant transposable elements in a partially automated large scale approach to complement the more animal focused REPBASE database http://www.girinst.org/repbase/ (Bao et al., 2015). As GCBN is continuously annotating new genomes, more and more data is fed to this database (Spannagl et al., 2017). This now leaves GCBN-HMGU as the last plant specific provider, as the TIGR plant repeat database (Ouyang and Buell, 2004) was discontinued in 2017 due to lack of funding. Transposable elements, which were once dismissed as merely invasive items, are experiencing a resurgence in interest as evidence accumulates to show that plants sometimes domesticate transposons or transposon promoters (Bolger et al., 2014a). This has been shown to cause divergence in regulation of flowering time (Lutz et al., 2015) and also that their role in genome and epigenome modification by providing a means of variability is important (Maumus and Quesneville, 2014).

Once the structure of the genes have been identified in a genome, it is then necessary to ascribe function to these genes. Whilst it may appear than the function for many genes has already been discerned, one has to keep in mind that many plant genes currently have no function that can be attributed simply due to the lack of knowledge. This holds true even for the highly researched model plant *Arabidopsis thaliana* (Bolger et al., 2017). A first approach for functional annotation can be to ascribe function based on sequence similarity. This can be as straightforward as performing a blast search against as similar species. In the case of *Triticae*, one could use the GCBN-IPK Blast server http://webblast.ipk-gatersleben.de/, a well annotated resource for barley and rye which integrates many different datasets such as exome capture specific sets which are commonly not accessible.

Generally, a comprehensive annotation pipeline will integrate similarity searches, domain architecture analysis along with other available data as discussed in detail in Bolger et al. (2017). GCBN-FZJ currently provides the Mercator annotation pipeline http://www.plabipd.de/portal/mercator-sequence-annotation, which was developed specifically for plants (Lohse et al., 2014). This pipeline is currently being reworked to enhance performance both in terms of speed and accuracy and incorporates sensitive detection techniques using a manually curated knowledge base (http://www.plabipd.de/portal/mercator-ii-alpha-version-). This online pipeline additionally classifies genes according to function and provides a simple, easy to interpret

human readable annotation. Thus GCBN is complementing automated resources for gene functional annotation such as BLAST2GO (Conesa and Gotz, 2008) which specifically uses GO terms and is a generalist tool, KAAS (Moriya et al., 2007) which focusing on KEGG pathways and the generalist plant focus TRAPID (Van Bel et al., 2013). Here, BLAST2GO performs well for plants, however to unlock its full potential, a license is required for faster analysis. Differences and particular advantages of the different tools have recently been reviewed (Bolger et al., 2017).

For specialized annotation needs, there are many databases dedicated to particular protein families or functions such as the ARAMEMNON database http://aramemnon.uni-koeln.de/ (Schwacke et al., 2003), an extensive data resource which focuses on data pertaining to plant membrane proteins. This multi-species database provides a comprehensive resource including sequences, topological predictions and subcellular localization predictions. The database also maintains and frequently updates functional descriptions of proteins based on publications.

4. Genome re-sequencing

Scientists or breeders often work on species which already have their genomes sequenced. Whilst this resource is sufficient for many endeavors, there are many situations which warrant re-sequencing of a species. Assemblies which are created from re-sequencing a plant are typically performed using the existing assembly as a guide or reference. This form of reference-based assembly is considerably easier than a *de-novo* genome assembly but can still present numerous difficulties.

The challenges presented by plant genomes, such as large size, heterozygosity and ploidy levels remain relevant for reference-based assemblies, albeit to a lesser extent than *de novo* assembly. Ploidy however remains the greatest challenge, since it results in multiple copies of the same genes which may need to be distinguished during data analysis. In the case of allopolyploids species *e.g.* rape seed; this is of particular importance as the genes on the corresponding loci may have functionally diverged. Distinguishing between the genes at corresponding loci in autopolyploid is usually of lower importance given that the genome duplication event resulted in (nearly) identical copies of the genes, and thus it is seldom the case that different alleles perform very diverged functions. The typical alignment and variant calling pipelines depend on (i) mapping a read to the correct genomic region, (ii) determine if there is a variant present in this region and (iii) in the case of autopolyploid species such as potato, one needs to be able to call beyond simple two-allelic combinations when considering single nucleotide polymorphisms (SNPs). In case of allopolyploid species, the genes from the corresponding loci are sufficiently divergent to require analysis as separate genes. It should also be noted that plant genomes do not typically have the high number of validated SNPs which are available to scientists working on the human genome.

The use of multiple pipelines to overcome these problems has been evaluated (Schmutzer et al., 2015) where the consensus results from three different variant callers on rape seed were compiled. In another pre-study, 8 variant calling tools on an even larger collection of pipelines to ascertain which performs best on the maize crop (Muraya et al., 2015). Training courses are available on this topic with the course material available online (http://www.plabipd.de/portal/workshop). Furthermore, GCBN-FZJ is currently comparing complete pipelines commencing with adapter trimming using Trimmomatic (Bolger et al., 2014b) to alignment using bowtie http://bowtie-bio.sourceforge.net/ and BWA http://bio-bwa.sourceforge.net/followed by SNP calling programs. This data is then compared to the data stemming from independent sources.

5. Phenotypes

A major goal behind plant genomics is to understand and predict the effect, changes to the genome have on phenotypes. While plant selection and crossing to manipulate phenotypes has been practiced well before the days of Mendel, revealing the underlying genetic code has provided scientists with a new toolbox to further refine and expedited this process. Given the recent surge of genomic data, there is currently concern among the scientific community that this genomic revolution is outpacing the availability of phenotyping data especially if the latter is not shared in a useful manner (Zamir, 2013).

Genomic data has a major role to play in crop genetic improvements and breeding programs. However, considerable gain can only be achieved by tightly coupling genomic discovery to plant phenomics (Cobb et al., 2013). While many applications for high-throughput and minimally-invasive phenotyping methods are being developed for both controlled environment and field experiments (Fahlgren et al., 2015; Fiorani and Schurr, 2013), data analysis remains a challenge. This is due to many factors including differences based on plant growth environment (Poorter et al., 2016) and even based on data storage, common ontologies and standards which had been lacking. These factors combine to limit the potential of many tools due to the low interoperability of the data and tools. To overcome these shortcoming, the minimal information standard on plant phenotype data (MIAPPE − Minimum Information About Phenotyping Experiment) was proposed (Krajewski et al., 2015) and a first implementations developed (Cwiek-Kupczynska et al., 2016) using ISA-Tab (Investigation/Study/Assay tab-delimited), a framework used to collect and communicate complex metadata. The MIAPPE checklist consists of attributes that can be classified within the following sections: general metadata, timing and location of experiments, biosource, environment (aerial, soil), treatments, experimental design, sample collection, processing, and management, and observed variables. When appropriate, publicly available ontologies are indicated for each section as recommended terminology.

Plant phenotyping is the quantitative appraisal of traits from a given genotype in a given environment and experiment, which range from scalar (plant height), multi-value (chemical or transcriptional) to image-based (pictures) and include both directly measured attributes and those derived from analysis, *e.g.* leaf area from shoot images. These heterogeneous data presents problems not only for analysis, but additionally for long-term access in a useful manner once the results have been published as standards are only emerging slowly (Cwiek-Kupczynska et al., 2016; Krajewski et al., 2015). To tackle this issue, GCBN-IPK has developed the Plant Genomics and Phenomics Research Data Repository (PGP) https://edal.ipk-gatersleben.de/repos/pgp/ (Arend et al., 2016a) as an infrastructure to comprehensively publish plant research data. This covers cross-domain datasets which are not being published in hitherto developed central repositories for reasons of data volume and/or data domain. This includes data such as plant phenotyping and microscopy images, incomplete genomic data, genotyping data, visualizations of morphological plant models, mass spectrometry data as well as software code and related documents. PGP is based on the e!DAL data publication infrastructure (Arend et al., 2014). Using this infrastructure, a reference experiment comprising multiple data domains is described using ISATab and published in PGP as a part of a research article (Arend et al., 2016b; Junker et al., 2014). All semantic and technical documentations, measured parameters, protocols and references to ontologies are manually described using ISATab format. The dataset is published as DOI:10.5447/IPK/2016/7 All raw files of such ISATab formatted data publications are stored in the PGP repository. PGP has been used to publish 115 DOIs which refer to more than 156,000 files. The PGP repository is accepted as a data repository for the Nature Publishing Group and is registered in re3data.org, OpenAIRE and DataCite, three of the main meta repositories for research data.

6. Data analysis

Data analysis is often the most overlooked task during the planning

of experiments, but has the potential to provide the highest returns on effort investments. It is indeed true that the quality of data gathered will massively influence the quality of the outcome, but it is equally true that even the best quality data is unlikely to surrender insights without appropriate data analysis.

Expression data for plants is ubiquitous in almost all public resources, ranging from microarray to RNASeq data. Indeed the plant community is well served with resources which have collated this data such as GENEVESTIGATOR https://genevestigator.com/ (Zimmermann et al., 2004, 2008), which is the one of the highest cited gene expression resource in plant biology, albeit at a monetary cost for full use. Other resources which are offered for free include BAR http://bar.utoronto.ca/ (Winter et al., 2007), which mostly focuses on model organisms and by GCBN in the form of the RNASeqExpressionBrowser (http://pgsb.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/index.jsp) resources (Nussbaumer et al., 2014). The latter is an open source web interface featuring expression data for barley and bread wheat and is used mainly by wet lab scientists to facilitate data interpretation for their genes of interest. These resources allow users to query for potentially candidate genes and provides immediate answer to questions such as: Is my favorite gene up-regulated under drought conditions? Or does this gene react to pathogens?

In cases where users generate their own expression data, analysis and interpretation of this data requires a more hands-on approach. For these analyses, GCBN hosts the RobiNA http://mapman.gabipd.org/robin (Lohse et al., 2012, 2010) and MapMan downloadable tools http://mapman.gabipd.org/ (Jaiswal and Usadel, 2016; Urbanczyk-Wochniak et al., 2006), which provide complete solutions for expression analysis. RobiNA allows users to analyze microarray and RNAseq data by providing a graphical user interface to model the experimental design and thus perform appropriate analysis. MapMan is a platform-independent application which allows the analysis and biological interpretation of plant omics data by mapping genes, metabolites and proteins onto metabolic, regulatory and developmental pathways. Unlike similar available tools for the GO ontology, MapMan attempts to (i) minimize the redundancy in overviews, i.e. in large diagrams, genes are usually shown only once and (ii) to accurately access and annotate the pathways based on the same framework that Mercator uses.

On the next level, these expression compendia can be used to make new inferences about gene–gene interaction. Indeed, the collation of expression datasets to derive co-expression networks, thus allowing simple gene based queries to find genes behaving similarly is a tried and tested service used by the plant community (Usadel et al., 2009). Moreover, GCBN is maintaining CSBDB.de which is the first published plant co-expression database (Steinhauser et al., 2004) and also the use of novel algorithms to glean the best possible data out of expression data, developed at the Max Planck Institute of Molecular Plant Physiology (Mutwil et al., 2010). Further data analysis tools offered by GCBN include a tool to compute measures of association and functional inference in the form of Corto http://www.usadellab.org/cms/index.php?page=corto. This approach of 'guilt by association' has been further developed within GCBN and used to successfully predict seed attributes in the model species *Arabidopsis thaliana* (Vasilevski et al., 2012; Voiniciuc et al., 2015a, 2015b). Currently, efforts are underway by GCBN to apply these techniques to rape seed and other *Brasicacae*, leveraging the model Arabidopsis.

From the whole plant transcriptomics perspective, users are sometimes confronted with the problem of identifying the transcriptional response of a mutant isolated from an informed candidate gene screen looking e.g. for drought specific mutants. However, it rarely happened that the transcriptomic response of a mutant exactly matches the intended profile. Often, one observes either a mixed response where the response actually consists of the primary response to e.g. drought and maybe a secondary one to other environmental factors. Sometimes one might even encounter an unexpected transcriptomic response. Here the research community working on human data has developed novel algorithms to compare such data sets to large gene expression compendia in what is called "physiospace" (Lenz et al., 2013). Here an individual data set is compared to a large compendium of data after transformation; GCBN has adapted this approach for plant data to unravel transcriptomic responses.

In order to facilitate plant genome comparison, GCBN-HMGU has developed a tool called CrowsNest which leverages conserved gene order between species. CrowsNest http://pgsb.helmholtz-muenchen.de/plant/crowsNest/index.jsp allows the user to explore syntenic relationships between species at different levels of detail, ranging from whole genome comparisons over chromososmes down to single genes. These connections are especially valuable for knowledge transfer from well characterized genes of reference species to as yet uncharacterized genes of newly sequenced species.

The aforementioned generic plant resources are complemented by specialized databases and services. GCBN-FZJ has developed a specialized database for large plant enzyme gene families, and as another example DroughtDB http://pgsb.helmholtz-muenchen.de/droughtdb/ (Alter et al., 2015) at GCBN-HMGU. It contains a manually curated set of drought stress genes from model species (Arabidopsis and rice) that have an experimentally verified function in drought tolerance. It interconnects them between nine species, including maize and barley, *via* computed orthology. This resource allows breeders to easily check candidate genes or explore candidate genetic regions for the occurrence of these curated genes.

Finally, it is of utmost importance, especially for crops, to be able to link genotypes to phenotypes. Whilst GCBN is not involved in the development of the underlying statistics, it employs state of the art tools such as fastLMM https://www.microsoft.com/en-us/research/project/fastlmm/ (Lippert et al., 2011) and helps in interpreting such datasets by defining genes underlying QTL (Millet et al., 2016).

7. Discussion

As high throughput sequencing and phenotyping technologies mature, it might be expected that the development of new bioinformatics services and pipelines becomes redundant due to achieving optimal solutions. The reality is however the opposite as the increasing use of these technologies warrants new techniques/algorithms to deal with this ever expanding mass of data. Even from a simple storage point of view, providing sustainable fast access to large data requires considerable expertise. Integration of data, especially heterogeneous phenotyping data, into existing or new data structures frequently needs to be performed before any data analysis can be carried out. This step additionally depends on the definition of standards, without which, phenotyping data is impossible to compare and integrate. As the distinction between wet-lab scientist and data scientist becomes increasing blurred, given that all aspects of research require some degree of data handling and analysis, training in fundamental bioinformatics techniques is imperative.

Within Germany GCBN is uniquely situated to undertake these services. Internationally, this ties into the ELIXIR project of Europe, where the GCBN plant side will likely be strengthening ELIXIR and complement its activities. Internationally, the US CyVerse infrastructure project which started out as iPLANT (Merchant et al., 2016) is a similar endeavor. Certain aspects of the setup are however different, as CyVerse mainly re-uses tools developed by the community such as Trimmomatic from GCBN in the CyVerse pipelines and allows them to be run on large infrastructures. GCBN actively maintains and develops their own services for the need of wet lab biologists. Therefore, these two initiatives are complementing each other perfectly.

8. Conclusions

The impact which the genomics revolution has made on plant science is undeniable. Within a decade of the first published plant genome,

sequencing has become a staple of most plant laboratories. As these technologies further improve and their cost decreases, the need for bioinformatics analysis and tools will clearly follow this trend. GCBN is already providing bioinformatics solutions not only to genomics data, but also to phenotyping data. It is after all the overarching goal of many scientists to use genomics data to predict and manipulate plant phenotypes. This goal requires extensive bioinformatics analysis on both genomics as well as the phenotyping data, a target GCBN is working on achieving.

Acknowledgements

GCBN wants to acknowledge funding be the German Ministry of Education and Research FKZ031A536A-C. In addition we acknowledge partial funding by the German Ministry of Education and Research for the German Plant Phenotyping network031A053 and the Plant Primary database FKZ0315961 projects and the NRW Strategieprojekt BioSC (no. 313/323-400-002 13).

References

Alter, S., Bader, K.C., Spannagl, M., Wang, Y., Bauer, E., Schon, C.C., Mayer, K.F., 2015. DroughtDB: an expert-curated compilation of plant drought stress genes and their homologs in nine species. Database: J. Biol. Databases Curation 2015, bav046.

Arabidopsis Genome, I., 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D., Scholz, U., 2014. e!DAL—a framework to store, share and publish research data. BMC Bioinf. 15, 214.

Arend, D., Junker, A., Scholz, U., Schuler, D., Wylie, J., Lange, M., 2016a. PGP repository: a plant phenomics and genomics data publication infrastructure. Database: J. Biol. Databases Curation 2016.

Arend, D., Lange, M., Pape, J.-M., Weigelt-Fischer, K., Arana-Ceballos, F., Mücke, I., Klukas, C., Altmann, T., Scholz, U., Junker, A., 2016b. Quantitative monitoring of Arabidopsis thaliana growth and development using high-throughput plant phenotyping. Sci. Data 3, 160055.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11.

Bauer, E., Schmutzer, T., Barilar, I., Mascher, M., Gundlach, H., Martis, M.M., Twardziok, S.O., Hackauf, B., Gordillo, A., Wilde, P., Schmidt, M., Korzun, V., Mayer, K.F.X., Schmid, K., Schön, C.-C., Scholz, U., 2017. Towards a whole-genome sequence for rye (Secale cereale L.). Plant J. 89, 853–869.

Bolger, A., Scossa, F., Bolger, M.E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sorensen, I., Lichtenstein, G., Fich, E.A., Conte, M., Keller, H., Schneeberger, K., Schwacke, R., Ofner, I., Vrebalov, J., Xu, Y., Osorio, S., Aflitos, S.A., Schijlen, E., Jimenez-Gomez, J.M., Ryngajllo, M., Kimura, S., Kumar, R., Koenig, D., Headland, L.R., Maloof, J.N., Sinha, N., van Ham, R.C., Lankhorst, R.K., Mao, L., Vogel, A., Arsova, B., Panstruga, R., Fei, Z., Rose, J.K., Zamir, D., Carrari, F., Giovannoni, J.J., Weigel, D., Usadel, B., Fernie, A.R., 2014a. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat. Genet. 46, 1034–1038.

Bolger, A.M., Lohse, M., Usadel, B., 2014b. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., Mayer, K.F., 2014c. Plant genome sequencing – applications for crop improvement. Curr. Opin. Biotechnol. 26, 31–37.

Bolger, M.E., Arsova, B., Usadel, B., 2017. Plant genome and transcriptome annotations: from misconceptions to simple solutions. Brief. Bioinform https://doi.org/10.1093/bib/bbw135.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.H., Childs, K.L., Sun, Y., Jiang, N., Yandell, M., 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164, 513–524.

Claros, M.G., Bautista, R., Guerrero-Fernandez, D., Benzerki, H., Seoane, P., Fernandez-Pozo, N., 2012. Why assembling plant genome sequences is so challenging. Biology 1, 439–459.

Cobb, J.N., Declerck, G., Greenberg, A., Clark, R., McCouch, S., 2013. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor. Appl. Genet. 126, 867–887.

Conesa, A., Gotz, S., 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genom. 2008, 619832.

Cwiek-Kupczynska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., van Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.A., Scholz, U., van Schriek, M., Seren, U., Usadel, B., Weise, S., Kersey, P., Krajewski, P., 2016. Measures for interoperability of phenotypic data: minimum information requirements and formatting. Plant Methods 12, 44.

De Luca, V., St Pierre, B., 2000. The cell and developmental biology of alkaloid biosynthesis. Trends Plant Sci. 5, 168–173.

Dohm, J.C., Minoche, A.E., Holtgrawe, D., Capella-Gutierrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sorensen, T.R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, P.F., Schmidt, T., Gabaldon, T., Lehrach, H., Weisshaar, B., Himmelbauer, H., 2014. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505, 546–549.

Fahlgren, N., Gehan, M.A., Baxter, I., 2015. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr. Opin. Plant Biol. 24, 93–99.

Fernie, A.R., 2007. The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. Phytochemistry 68, 2861–2880.

Fiorani, F., Schurr, U., 2013. Future scenarios for plant phenotyping. Annu. Rev. Plant Biol. 64, 267–291.

Flutre, T., Duprat, E., Feuillet, C., Quesneville, H., 2011. Considering transposable element diversification in de novo annotation approaches. PLoS One 6, e16526.

Fujii, S., Kubo, K., Takayama, S., 2016. Non-self- and self-recognition models in plant self-incompatibility. Nat. Plant 2, 16130.

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., Podlich, D., 2006. Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci. 11, 587–593.

He, N., Zhang, C., Qi, X., Zhao, S., Tao, Y., Yang, G., Lee, T.H., Wang, X., Cai, Q., Li, D., Lu, M., Liao, S., Luo, G., He, R., Tan, X., Xu, Y., Li, T., Zhao, A., Jia, L., Fu, Q., Zeng, Q., Gao, C., Ma, B., Liang, J., Wang, X., Shang, J., Song, P., Wu, H., Fan, L., Wang, Q., Shuai, Q., Zhu, J., Wei, C., Zhu-Salzman, K., Jin, D., Wang, J., Liu, T., Yu, M., Tang, C., Wang, Z., Dai, F., Chen, J., Liu, Y., Zhao, S., Lin, T., Zhang, S., Wang, J., Wang, J., Yang, H., Yang, G., Wang, J., Paterson, A.H., Xia, Q., Ji, D., Xiang, Z., 2013. Draft genome sequence of the mulberry tree Morus notabilis. Nat. Commun. 4, 2445.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., Fields, C.J., Wright, C.L., Koehler, K., Springer, N.M., Buckler, E., Buell, C.R., de Leon, N., Kaeppler, S.M., Childs, K.L., Mikel, M.A., 2016. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28, 2700–2714.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., Stanke, M., 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with geneMark-ET and AUGUSTUS. Bioinformatics 32, 767–769.

International Barley Genome Sequencing, Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., Wise, R.P., Stein, N., 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491, 711–716.

International Wheat Genome Sequencing, C., 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., Cavagnaro, P., Yildiz, M., Macko-Podgorni, A., Moranska, E., Grzebelus, E., Grzebelus, D., Ashrafi, H., Zheng, Z., Cheng, S., Spooner, D., Van Deynze, A., Simon, P., 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. 48, 657–666.

Jaiswal, P., Usadel, B., 2016. Plant pathway databases. Methods Mol. Biol. 1374, 71–87.

Jiao, W.-B., Schneeberger, K., 2017. The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. 36, 64–70.

Jiao, W.B., Garcia Accinelli, G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Humann, U., Reinhard, R., Koch, M.A., Swan, D., Clavijo, B., Coupland, G., Schneeberger, K., 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. http://dx.doi.org/10.1101/gr.213652.116.

Joppich, M., Schmidl, D., Bolger, A.M., Kuhlen, T., Usadel, B., 2015. PAGANtec: openMP parallel error correction for next-generation sequencing data. In: Terboven, C., de Supinski, B.R., Reble, P., Chapman, B.M., Müller, M.S. (Eds.), OpenMP: Heterogenous Execution and Data Movements. 11th International Workshop on OpenMP, IWOMP 2015, Aachen, Germany, October 1–2, 2015, Proceedings. Springer International Publishing, Cham. pp. 3–17.

Junker, A., Muraya, M.M., Weigelt-Fischer, K., Arana-Ceballos, F., Klukas, C., Melchinger, A.E., Meyer, R.C., Riewe, D., Altmann, T., 2014. Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. Front. Plant Sci. 5, 770.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., Chen, L., 2011. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. 30, 90–98.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736.

Krajewski, P., Chen, D., Cwiek, H., van Dijk, A.D., Fiorani, F., Kersey, P., Klukas, C., Lange, M., Markiewicz, A., Nap, J.P., van Oeveren, J., Pommier, C., Scholz, U., van Schriek, M., Usadel, B., Weise, S., 2015. Towards recommendations for metadata and data handling in plant phenotyping. J. Exp. Bot. 66, 5417–5427.

Lenz, M., Schuldt, B.M., Muller, F.J., Schuppert, A., 2013. PhysioSpace: relating gene expression experiments from heterogeneous sources using shared physiological processes. PLoS One 8, e77627.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D., 2011. FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833–835.

Lohse, M., Nunes-Nesi, A., Kruger, P., Nagel, A., Hannemann, J., Giorgi, F.M., Childs, L., Osorio, S., Walther, D., Selbig, J., Sreenivasulu, N., Stitt, M., Fernie, A.R., Usadel, B.,

2010. Robin: an intuitive wizard application for R-based expression microarray quality assessment and analysis. Plant Physiol. 153, 642–651.
Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M., Usadel, B., 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 40, W622–627.
Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A.R., Stitt, M., Usadel, B., 2014. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ. 37, 1250–1258.
Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.W., Wang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18.
Lutz, U., Pose, D., Pfeifer, M., Gundlach, H., Hagmann, J., Wang, C., Weigel, D., Mayer, K.F., Schmid, M., Schwechheimer, C., 2015. Modulation of ambient temperature-dependent flowering in Arabidopsis thaliana by natural variation of FLOWERING LOCUS M. PLoS Genet. 11, e1005588.
Martis, M.M., Zhou, R., Haseneyer, G., Schmutzer, T., Vrana, J., Kubalakova, M., Konig, S., Kugler, K.G., Scholz, U., Hackauf, B., Korzun, V., Schon, C.C., Dolezel, J., Bauer, E., Mayer, K.F., Stein, N., 2013. Reticulate evolution of the rye genome. Plant Cell 25, 3685–3698.
Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., Bayer, M., Ramsay, L., Liu, H., Haberer, G., Zhang, X.Q., Zhang, Q., Barrero, R.A., Li, L., Taudien, S., Groth, M., Felder, M., Hastie, A., Simkova, H., Stankova, H., Vrana, J., Chan, S., Munoz-Amatriain, M., Ounit, R., Wanamaker, S., Bolser, D., Colmsee, C., Schmutzer, T., Aliyeva-Schnorr, L., Grasso, S., Tanskanen, J., Chailyan, A., Sampath, D., Heavens, D., Clissold, L., Cao, S., Chapman, B., Dai, F., Han, Y., Li, H., Li, X., Lin, C., McCooke, J.K., Tan, C., Wang, P., Wang, S., Yin, S., Zhou, G., Poland, J.A., Bellgard, M.I., Borisjuk, L., Houben, A., Dolezel, J., Ayling, S., Lonardi, S., Kersey, P., Langridge, P., Muehlbauer, G.J., Clark, M.D., Caccamo, M., Schulman, A.H., Mayer, K.F.X., Platzer, M., Close, T.J., Scholz, U., Hansson, M., Zhang, G., Braumann, I., Spannagl, M., Li, C., Waugh, R., Stein, N., 2017. A chromosome conformation capture ordered sequence of the barley genome. Nature 544, 427–433.
Maumus, F., Quesneville, H., 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nat. Commun. 5, 4104.
Mayer, K.F., Taudien, S., Martis, M., Simkova, H., Suchankova, P., Gundlach, H., Wicker, T., Petzold, A., Felder, M., Steuernagel, B., Scholz, U., Graner, A., Platzer, M., Dolezel, J., Stein, N., 2009. Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 151, 496–505.
Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., Antin, P., 2016. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. PLoS Biol. 14, e1002342.
Millet, E.J., Welcker, C., Kruijer, W., Negro, S., Coupel-Ledru, A., Nicolas, S.D., Laborde, J., Bauland, C., Praud, S., Ranc, N., Presterl, T., Tuberosa, R., Bedo, Z., Draye, X., Usadel, B., Charcosset, A., Van Eeuwijk, F., Tardieu, F., 2016. Genome-wide analysis of yield in Europe: allelic effects vary with drought and heat scenarios. Plant Physiol. 172, 749–764.
Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182–185.
Muraya, M.M., Schmutzer, T., Ulpinnis, C., Scholz, U., Altmann, T., 2015. Targeted sequencing reveals large-scale sequence polymorphism in maize candidate genes for biomass production and composition. PLoS One 10, e0132120.
Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O., Persson, S., 2010. Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol. 152, 29–43.
Nussbaumer, T., Kugler, K.G., Bader, K.C., Sharma, S., Seidel, M., Mayer, K.F., 2014. RNASeqExpressionBrowser–a web interface to browse and visualize high-throughput expression data. Bioinformatics 30, 2519–2520.
Ouyang, S., Buell, C.R., 2004. The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 32, D360–363.
Poorter, H., Fiorani, F., Pieruschka, R., Wojciechowski, T., van der Putten, W.H., Kleyer, M., Schurr, U., Postma, J., 2016. Pampered inside, pestered outside? Differences and similarities between plants growing in controlled conditions and in the field. New Phytol. 212, 838–855.
Pucker, B., Holtgrawe, D., Rosleff Sorensen, T., Stracke, R., Viehover, P., Weisshaar, B., 2016. A de novo genome sequence assembly of the Arabidopsis thaliana accession Niederzenz-1 displays presence/absence variation and strong synteny. PLoS One 11, e0164321.
Schmutzer, T., Ma, L., Pousarebani, N., Bull, F., Stein, N., Houben, A., Scholz, U., 2014. Kmasker–a tool for in silico prediction of single-copy FISH probes for the large-genome species Hordeum vulgare. Cytogenet. Genome Res. 142, 66–78.
Schmutzer, T., Samans, B., Dyrszka, E., Ulpinnis, C., Weise, S., Stengel, D., Colmsee, C., Lespinasse, D., Micic, Z., Abel, S., Duchscherer, P., Breuer, F., Abbadi, A., Leckband, G., Snowdon, R., Scholz, U., 2015. Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant Brassica napus. Sci. Data 2, 150072.
Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W.B., Flugge, U.I., Kunze, R., 2003. ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol. 131, 16–26.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123.
Spannagl, M., Nussbaumer, T., Bader, K., Gundlach, H., Mayer, K.F., 2017. PGSB/MIPS PlantsDB database framework for the integration and analysis of plant genome data. Methods Mol. Biol. 1533, 33–44.
Stanke, M., Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 (Suppl. 2), ii215–ii225.
Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O., Kopka, J., 2004. CSB.DB: a comprehensive systems-biology database. Bioinformatics 20, 3647–3651.
Tang, H., Bomhoff, M.D., Briones, E., Zhang, L., Schnable, J.C., Lyons, E., 2015. SynFind: compiling syntenic regions across any set of genomes on demand. Genome Biol. Evol. 7, 3286–3298.
Urbanczyk-Wochniak, E., Usadel, B., Thimm, O., Nunes-Nesi, A., Carrari, F., Davy, M., Blasing, O., Kowalczyk, M., Weicht, D., Polinceusz, A., Meyer, S., Stitt, M., Fernie, A.R., 2006. Conversion of MapMan to allow the analysis of transcript data from Solanaceous species: effects of genetic and environmental alterations in energy metabolism in the leaf. Plant Mol. Biol. 60, 773–792.
Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., Provart, N.J., 2009. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. 32, 1633–1651.
Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., Vandepoele, K., 2013. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol. 14, R134.
Vasilevski, A., Giorgi, F.M., Bertinetti, L., Usadel, B., 2012. LASSO modeling of the Arabidopsis thaliana seed/seedling transcriptome: a model case for detection of novel mucilage and pectin metabolism genes. Mol. Biosyst. 8, 2566–2574.
Voiniciuc, C., Gunl, M., Schmidt, M.H., Usadel, B., 2015a. Highly branched xylan made by IRREGULAR XYLEM14 and MUCILAGE-RELATED21 links mucilage to arabidopsis seeds. Plant Physiol. 169, 2481–2495.
Voiniciuc, C., Schmidt, M.H., Berger, A., Yang, B., Ebert, B., Scheller, H.V., North, H.M., Usadel, B., Gunl, M., 2015b. MUCILAGE-RELATED10 produces galactoglucomannan that maintains pectin and cellulose architecture in Arabidopsis seed mucilage. Plant Physiol. 169, 403–420.
Weisenfeld, N.I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., Nusbaum, C., Lander, E.S., MacCallum, I., Jaffe, D.B., 2014. Comprehensive variation discovery in single human genomes. Nat. Genet. 46, 1350–1355.
Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., Provart, N.J., 2007. An Electronic Fluorescent Pictograph browser for exploring and analyzing large-scale biological data sets. PLoS One 2, e718.
Yang, X., Chockalingam, S.P., Aluru, S., 2013. A survey of error-correction methods for next-generation sequencing. Brief. Bioinform. 14, 56–66.
Zamir, D., 2013. Where have all the crop phenotypes gone? PLoS Biol. 11, e1001595.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.
Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., Gruissem, W., 2004. GENEVESTIGATOR: Arabidopsis microarray database and analysis toolbox. Plant Physiol. 136, 2621–2632.
Zimmermann, P., Laule, O., Schmitz, J., Hruz, T., Bleuler, S., Gruissem, W., 2008. Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. Mol. Plant 1, 851–857.