Exploring Structural Variants in Environmentally-Sensitive Gene Families

Nevin Dale Young^{1,2,3}, Peng Zhou^{1,4}, Kevin A. T. Silverstein⁴

Addresses

¹ Department of Plant Pathology

495 Borlaug Hall

² Department of Plant Biology

220 BioScience Building

University of Minnesota

St. Paul, MN 55108 USA

Email: neviny@umn.edu

³ Corresponding Author

⁴ Minnesota Supercomputer Institute

University of Minnesota

Minneapolis, MN, 55455, USA

Abstract

Environmentally sensitive plant gene families like NBS-LRRs, receptor kinases, defensins and others, are known to be highly variable. However, most existing strategies for discovering and describing structural variation in complex gene families provide incomplete and imperfect results. The move to *de novo* genome assemblies for multiple accessions or individuals within a species is enabling more comprehensive and accurate insights about gene family variation. Earlier array-based genome hybridization and sequence-based read mapping methods were limited by their reliance on a reference genome and by misplacement of paralogous sequences. Variant discovery based on *de novo* genome assemblies overcome the problems arising from a reference genome and reduce sequence misplacement. As *de novo* genome sequencing moves to the use of longer reads, artifacts will be minimized, intact tandem gene clusters will be constructed accurately, and insights into rapid evolution will become feasible.

Abbreviations

CGH	Comparative genome hybridization
CNV	Copy number variation
CRP	Cysteine rich peptide
IRLC	Inverted repeat lacking clade
kbp	Kilobase pairs
Mbp	Megabase pairs
NB-ARC	Nucleotide-binding, APAF-1, R proteins, CED-4
NBS-LRR	Nucleotide binding site, leucine-rich repeat
NCR	Nodule-specific, cysteine-rich peptide
PAV	Presence-absence variation
RLK	Receptor-like kinase
RLP	Receptor-like protein
RD	Read-depth
RP	Read-pair
SR	Split-read
SV	Structural variation

Introduction

Recent advances in DNA sequencing, genome assembly, and structural variant (SV) discovery are having a major impact on our understanding of large plant gene families. This is especially relevant for environmentally-sensitive gene families like nucleotide binding site-leucine rich repeats (NBS-LRRs), receptor-like kinases (RLKs), F-box proteins, and cysteine-rich peptides (CRPs). These gene families are broadly implicated in biotic and abiotic interactions, and like transposon sequences, they have long been known to exhibit high levels of variation within genomes [1–5]. Nonetheless, research has generally been constrained by relatively weak methods for SV discovery [6–9] and reliance on reference genome sequences until very recently.

Existing discovery methods are limited in power for exploring variants

Early attempts to interrogate sequence variation were generally based on resequencing microarrays [1] and comparative genome hybridization (CGH) [3,10,11], providing the first large-scale surveys of plant genome sequence variation. With the advent of next generation sequencing, SV discovery quickly transitioned to sequenced-based methods, reviewed in a recent article by Pirooznia *et al* [9]. These methods broadly fit into three categories: Read-Pair (RP), Split-Read (SR), and Read-Depth (RD) – plus a fourth, SV discovery based on genome assemblies – discussed in detail below (Table 1). RP methods begin by mapping short reads to a reference genome followed by comparing the distance between read pairs with the expected distance (insert size). Deviations from the expected size distribution can then be used to detect unbalanced SVs like insertions and deletions. SR approaches, by contrast, make use of read pairs where only one read in a pair has a reliable mapping while the other is either unmapped or only partially mapped to the genome (soft-clipped). The unmapped and soft-clipped reads are then used to predict the breakpoint of a SV at base pair resolution. Finally, RD

approaches assume a correlation between depth of coverage for a genomic region and the copy number of that region, enabling prediction of copy number variants (CNVs) [12].

Each approach has its own strengths and weaknesses. While RP methods are good at detecting medium-sized SVs, they are insensitive to smaller events due to problems separating small deviations (true signal) from normal background variability, a feature highly dependent on library construction [6]. The RD method is the only one to accurately predict absolute copy number, but its breakpoint resolution is poor and it often misses smaller events [8]. Similarly, SR algorithms can detect a wide range of SVs at base pair resolution, but are limited by read length while also tending to ignore large-scale SVs [9]. All three methods have poor performances in repetitive regions since they rely on confident placement of reads in the first place [7]. This is especially important in the case of repetitive genome regions, including large gene families where tandem duplication frequently occurs [13,14].

Read-mapping SV discovery also relies on a reference genome sequence and therefore necessarily overlooks regions that are missing from the reference. RD methods consequently uncover more deletions than insertions relative to the reference [15]. Read mapping will also generally attempt to place reads against the best observed hit, even if that sequence is a paralog. In a recent human genome study, standard variant call sets missed up to 25% of true variants when read mapping approaches were critically compared to long read approaches, with the greatest impact in repetitive regions of the genome [16]. Given that environmentally-sensitive plant gene families are repetitive and frequently organized into clusters [13,17], reliance on a single reference genome sequence clearly hinders efforts to describe the landscape of SV in plant disease resistance genes and other large gene families.

Environmentally-sensitive gene families have long been known to be highly variable

Early *Arabidopsis* re-sequencing highlighted the extensive level of variation in NBS-LRRs and other environmentally-sensitive gene families. Array-based re-sequencing of 20 *Arabidopsis* accessions revealed that 60% of NBS-LRRs contained some type of major-effect change (variants that lead to disruption or significant modification in protein product) compared with less than 10% in all known expressed sequences at the time [1]. In the same study, 25% of F-box proteins and 16% of RLKs also contained major-effect changes. Similar estimates of variation among environmentally-sensitive gene families were observed in a later sequence-based study of 80 *Arabidopsis* accessions [18]. Not only was DNA sequence variation highest in environmentally-sensitive gene families, so was expression variation among accessions [19]. An interesting corollary to these results, however, was the observation that compensating mutations frequently restore the coding potential of many major impact SNPs [19]. This does not necessarily extend to the impact of SVs, which are often more consequential and/or not so easily compensated. In soybean, CNVs uncovered by CGH were likewise over-represented among NBS-LRR and receptor-like proteins (RLP) genes, where 29% of clustered NBS-LRRs and 68% of clustered RLP co-localized with CNVs compared with just 1.5% in the reference gene set [5]. Disproportionately high levels of structural and major-effect variation among environmentally-sensitive gene families were also reported in rice [2], maize [3,11], soybean [5], and *Medicago truncatula* [4]. Clearly, early SV discovery based on CGH and short read alignment to a reference could be successful at revealing the very high level of variability in gene families associated with biotic and abiotic interactions. Today, even more effective methods for SV exploration are extending these results.

Recent advances in creating high quality assemblies

The foundation for better discovery and analysis of plant SVs is direct comparison based on *de novo* genome assemblies across accessions from a single taxon. Theoretically, all forms of SVs can be accurately detected by comparing *de novo* assemblies – provided the underlying sequence reads are long and accurate enough. Such direct genome sequence comparisons have been the mainstay of synteny analysis for years (for example, [20]). Exploring SV in plant genomes generally, and large gene families in particular, is transformed by such side-by-comparisons. Reasonably high quality assemblies can be created with short Illumina reads, especially when using an algorithm optimized for large, complex genomes, such as ALLPATHS-LG [21]. These assemblies typically have contig N50s (the length of the shortest contig in the size-ordered set representing 50% of the assembly) exceeding 25 kbp and scaffold N50s of 200 kbp or more. Nonetheless, early *de novo* assemblies also tended to be biased and error-prone in regions containing duplications and repeats [22]. Obviously, this impacts the effectiveness of SV detection in large gene families and highlights the need for even better sequencing, assembly and SV discovery technology (something on the way with better and more economical long read methodology; see below).

Better genome assembly has already had a major impact on SV discovery and pan-genomic analysis in plants. Focusing on six wild *Orzya* genomes descended from a common AA-genome ancestor, Zhang et al uncovered massive levels of SV – between 230,000 and 540,000 instances for both insertions and deletions – and examined the role of gene family expansions and contractions [23]. Altogether, the genomes contained more than 39,000 gene families with a subset of over 10,000 common (core) to all the examined species (both wild and cultivated). Many of these lineage-specific expansions and losses were associated with specific phenotypic adaptations. Estimates of turnover and SV levels were especially high in families related to floral development and reproduction as well as disease resistance. In another rice study, *de novo* assemblies for two *O. sativa* accessions uncovered 4.8 and 8.2 Mbp of accession-specific sequence compared with the Nipponbare (IRGSP-1.0) assembly [24]. Of this accession-specific genome space, only 555-760 kbp were

classified as non-repetitive and genome-specific. In wild soybean (*Glycine soja*), *de novo* assemblies of seven accessions revealed 1,987 genes affected by CNVs and 338 genes with PAVs that could only be detected using the assembled genomes [25]. Comparing the genomes in terms of shared versus accession-specific sequences indicated a dispensable genome size of 20% (51.4% of gene families). This value differs from a much higher estimate of 82.7% based on a pan-genomic analysis of transcript assemblies in maize [26]. The authors of the soybean study suggest that an autogamous versus outbred reproductive strategy might explain the difference [25], but so could differences in technology (genome sequencing in soybean versus transcriptome sequencing in maize; seven accessions sequenced in soybean versus 503 in maize).

***De novo* genome assemblies provide a better picture of gene family variability**

Just as high quality *de novo* assemblies are expanding our capacity to examine plant genome variation broadly, they also make it possible to ask more powerful questions about complex gene families. Two areas where this emerging capacity is already impacting research are better estimates of genome-wide inventories and opportunities to explore gene families in the context of a pan-genome.

An early study of NBS-LRRs and RLKs based on comparative filter hybridization, a classic technique involving solution-based hybridization to nylon membranes with spotted genomic DNA, indicated that NBS-LRR numbers varied by a factor of more than three-fold in *Glycine* and *Oryza* and as much as five-fold among *Gossypium* accessions [27]. RLKs also exhibited a comparable range, varying two-fold in *Oryza* and three-fold in *Glycine*. If true, this level of variation in gene family size within individual species would be truly remarkable. With the advent of high quality re-sequencing and *de novo* assemblies, more precise estimates of gene family numbers have become possible. Guo *et al* compared a *de novo* assembly for *A. lyrata* with the

A. thaliana reference (Columbia) and found 159 NBS-LRRs in *A. thaliana* compared with 185 in *A. lyrata* (a difference of 1.16x) using a stringent hidden Markov model profile [28]. Surveying high quality genome assemblies for AA genome *Oryza* species, Zhang [23] found 450 and 631 NBS-LRRs in the genomes of two rice cultivars (a range of 1.40x) and between 307 and 489 within the wild AA species genomes (a range of 1.59x) [23]. In a survey of seven *Glycine soja de novo* genome assemblies [25], NBS-LRRs varied between 334 and 382, a range of only 1.14x, while in *M. truncatula*, we have discovered between 747 and 860 NBS-LRRs in a survey of 16 different accessions (a range of 1.15x) (unpublished). In other words, NBS-LRR numbers do vary among accessions or closely related species, but significantly less than twofold.

In our lab, we are particularly interested in nodule-specific cysteine-rich peptides (NCRs). These genes are part of the larger cysteine-rich peptide family with remote homology to plant defensins [29,30]. They are small, secreted peptides encoded by one- and two-exon genes spread throughout the genome, frequently in clusters [30–32]. Surprisingly, they seem to be found exclusively in the nitrogen-fixing nodules of Inverted-Repeat Lacking Clade (IRLC) [29] and *Aeschynomene* [33] legumes throughout the process of nodulation [34]. Although the full set of functions for NCRs is still unknown, a few have been shown to act as symbiotic plant effectors directing bacteroid differentiation, leading to speculation that NCRs are antimicrobial peptides recruited by certain legumes to dominate endo-symbionts [35,36]. To learn more about variation among NCRs – and to compare their genome architecture with NBS-LRRs – we created 16 *de novo* genome sequences for *M. truncatula*. One of the first observations from this work has been that genome-wide, NCR numbers among the *Medicago* accessions range from 606 to 712 (1.17x) an estimate remarkably similar to the range observed for *Medicago* NBS-LRRs (unpublished results).

Environmentally-sensitive gene families like NBS-LRRs and NCRs are expected to be major components of the genome regions distinct from what was originally

discovered in earlier “reference” genome sequencing. This is hardly surprising given the high level of documented variability among these gene families and their comparably high level of presence-absence variation. In *Arabidopsis*, for example, 33.3% of NBS-LRR genes in the reference genome appeared to be deleted in at least one of 80 accessions compared with only 12.5% of genes in the genome as a whole [28]. Similar observations have been made for the fraction of gene families in the novel or accession-specific gene space. As a point of reference, NBS-LRRs vary from 0.6% (*Arabidopsis*) to 1% (rice, poplar and *Medicago*) to 2% (potato) of all gene models in a sampling of published plant reference genomes [37–41]. The picture is very different when looking at accession-specific genome regions. In *Oryza*, Schatz et al [24] re-sequenced three divergent rice genomes and found that genes containing the NB-ARC domain constituted 12% of lineage-specific genes in *O. sativa indica* (IR64) compared with just 0.35% of the genes shared among all three genomes [24]. In our analysis of *M. truncatula* genomes, we found that taken together, NBS-LRR, F-box, LRR, and NCR genes together account for 33% of all Pfam-defined genes in the non-transposon, accession-specific gene space.

A very different approach to exploring the disposable genome is to treat it as a metagenomic project. Yao et al [42] examined the pan-genome of 1483 cultivated rice accessions, sequencing them all to just 1-3X genome coverage [42]. After aligning to the Nipponbare reference and then examining sequences that could not be aligned, they uncovered thousands of protein genes including many agronomically important ones missing from the reference. Using association and linkage disequilibrium mapping, more than 78% of the disposable sequence could be positioned relative to the reference. Ultimately, the study focused on 1913 predicted genes in the dispensable genome, 1190 of which were associated with Pfam domains. Of these, a remarkable 166 were NB-ARCs (14%) and 138 were protein kinases (11.6%). Clearly, this metagenomic approach to pan-genome analysis successfully discovers environmentally-sensitive (and presumably all types of) gene families compared with high quality assemblies. It would be challenging, however, to extend the strategy to detailed examination of specific SVs or the impact

of different SVs on extended haplotypes in the genomes of different plant accessions.

Better SV understanding through better sequencing

Combining these observations about the SV architecture of environmentally-sensitive gene families, one is left with a curious paradox. While the total number of NBS-LRR, NCRs, and other environmentally-sensitive gene family members remains approximately constant across accessions within a species, the actual membership of families is quite dynamic. Given their role in biotic and abiotic interactions, this is hardly surprising. Different protein specificities are needed for different pathogens, symbionts, or abiotic challenges. From a genome evolution perspective, however, families with relatively constant gene numbers but highly variable membership raise intriguing questions.

It remains crucial, therefore, that DNA sequencing and genome assembly improve still further as a basis for SV discovery. As sequencing reads become longer with better error-correction algorithms, the quality and contiguity of *de novo* assemblies also increase [24,43]. Pacific Biosciences (PacBio) technology continues to make strides in generating longer reads and sequencing larger genomes, progressing all the way from bacteria to humans in just the past few years. With the recent full assembly of a human genome using PacBio sequencing and BioNano technology [44], researchers have shown that all types of SVs, including the most difficult tandem and segmental duplications, can be accurately recovered [43]. In this study, most (~95%) ALLPATH assembly SV predictions could be validated by comparison with the long-read-based PacBio assembly, but the PacBio assemblies uncovered far more total SV events [43]. Undoubtedly, many of the missed variants involved CNVs within dynamic gene families. High quality assemblies can now be enhanced even further using *in vitro* reconstituted chromatin techniques such as Hi-C [45] and Chicago [46]. For projects that cannot afford the 70X - 100X coverage required for a full PacBio assembly, success is being achieved by researchers using hybrid

approaches with lower PacBio coverage [47], such as PBJelly2 [48] or ALPACA (J. Miller, J. Mudge, K. Silverstein, personal communication). These hybrid approaches can save considerable cost at the potential expense of some accuracy and contiguity. As costs come down, ever more accurate examination of complex SV among environmentally-sensitive gene families in populations becomes feasible.

Conclusions

Large plant gene families that play a role in environmental interactions are abundant, important, and complex. Understanding their architecture, diversity and evolution is key to understanding their role in plant biology. Advances in sequencing technology and the move to *de novo* genome sequencing as a basis for describing SV are now making it possible to explore these areas in more depth and with better accuracy. Already, the results demonstrate that environmentally-sensitive gene families play an especially important role in the pan-genome.

Acknowledgements

We thank Jason Miller, Joann Mudge, and Peter Tiffin for useful discussions. We regret that it was not possible to cite the works of many additional scientists who contributed to this field as a result of space limitations. This work was supported, in part, by NSF grant DBI-1237993.

References

1. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds D a., et al.: **Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*.** *Science* 2007, **317**:338–42.
2. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al.: **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nat. Biotechnol.* 2011, **30**:105–111.
3. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al.: **Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.** *PLoS Genet.* 2009, **5**:e1000734.
4. Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti A K, Woodward JE, May GD, Gentzbittel L, et al.: **Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*.** *Proc. Natl. Acad. Sci.* 2011, **108**:E864–E870.
5. McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh J a., Stupar RM: **Structural variants in the soybean genome localize to clusters of biotic stress-response genes.** *Plant Physiol.* 2012, **159**:1295–308.
6. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing.** *Nat. Methods* 2009, **6**:S13–S20.
7. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: **Sensitive and accurate detection of copy number variants using read depth of coverage.** *Genome Res.* 2009, **19**:1586–92.
8. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping.** *Nat. Rev. Genet.* 2011, **12**:363–376.

- *9. Pirooznia M, Goes FS, Zandi PP: **Whole-genome CNV analysis: advances in computational approaches.** *Front. Genet.* 2015, **6**:138.

A substantive review about the informatic approaches to structural variant and copy number variant discovery based on short read sequencing technology.

- 10. DeBolt S: **Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales.** *Genome Biol. Evol.* 2010, **2**:441–53.
- 11. Swanson-Wagner R a., Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM: **Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor.** *Genome Res.* 2010, **20**:1689–99.
- 12. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A.: **Statistical challenges associated with detecting copy number variations with next-generation sequencing.** *Bioinformatics* 2012, **28**:2711–2718.
- 13. Silverstein KAT, Moskal WA, Wu HC, Underwood BA, Graham MA, Town CD, VandenBosch KA: **Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants.** *Plant J.* 2007, **51**:262–280.
- 14. McHale L, Tan X, Koehl P, Michelmore RW: **Plant NBS-LRR proteins: adaptable guards.** *Genome Biol.* 2006, **7**:212.
- 15. Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD, Mackay TFC, et al.: **Genomic variation and its impact on gene expression in *Drosophila melanogaster*.** *PLoS Genet.* 2012, **8**:e1003055.
- 16. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al.: **Comprehensive variation discovery in single human genomes.** *Nat. Genet.* 2014, **46**:1350–5.
- 17. Meyers BC: **Genome-Wide Analysis of NBS-LRR-Encoding Genes in *Arabidopsis*.** *Plant Cell Online* 2003, **15**:809–834.

18. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al.: **Whole-genome sequencing of multiple *Arabidopsis thaliana* populations.** *Nat. Genet.* 2011, **43**:956–963.
19. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al.: **Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*.** *Nature* 2011, **477**:419–423.
20. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al.: **Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids.** *Plant Physiol.* 2008, **148**:1772–81.
21. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, et al.: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc. Natl. Acad. Sci.* 2011, **108**:1513–1518.
22. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al.: **Personalized copy number and segmental duplication maps using next-generation sequencing.** *Nat. Genet.* 2009, **41**:1061–7.
- **23. Zhang Q-J, Zhu T, Xia E-H, Shi C, Liu Y-L, Zhang Y, Liu Y, Jiang W-K, Zhao Y-J, Mao S-Y, et al.: **Rapid diversification of five *Oryza* AA genomes associated with rice adaptation.** *Proc. Natl. Acad. Sci. U. S. A.* 2014, **111**:E4954–62.

A comparative genome analysis of closely related *Oryza* species based on high quality *de novo* genome assemblies, highlighting the role of structural variation and lineage-specific gene families and noting the substantial impact of resistance-related processes.

- **24. Schatz MC, Maron LG, Stein JC, Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, et al.: **Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica.** *Genome Biol.* 2014, **15**:506.

In order to explore the pan-genome of rice, the authors constructed high quality

de novo assemblies for three divergent rice varieties, exploring the lineage-specific gene spaces, which contained a notably high percentage of NB-ARC gene sequences.

- **25. Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, et al.: **De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits.** *Nat. Biotechnol.* 2014, **32**:1045–1052.

By sequencing and assembling seven diverse accessions of the wild soybean, *Glycine soja*, the authors created a pan-genome, described the core genome, which consisted of 80% of the pan-genome present in all seven accessions, and discovered lineage-specific genes.

- *26. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al.: **Insights into the maize pan-genome and pan-transcriptome.** *Plant Cell* 2014, **26**:121–35.

To explore the pan-genome of maize, the authors used transcriptome data from more than 500 inbred lines, identifying transcript assemblies expressed in all or just a sub-set of lines, going on to conclude that the maize genome is restricted in content.

27. Zhang M, Wu YH, Lee MK, Liu YH, Rong Y, Santos TS, Wu C, Xie F, Nelson RL, Zhang H Bin: **Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors.** *Nucleic Acids Res.* 2010, **38**:6513–6525.
28. Guo Y-L, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D: **Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*.** *Plant Physiol.* 2011, **157**:757–69.
29. Mergaert P, Nikovics K, Kelemen Z, Maunoury N, Vaubert D, Kondorosi A, Kondorosi E: **A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs.** *Plant Physiol.* 2003, **132**:161–73.

30. Graham M a., Silverstein KAT, Cannon SB, VandenBosch KA: **Computational identification and characterization of novel genes from legumes.** *Plant Physiol.* 2004, **135**:1179–97.
31. Alunni B, Kevei Z, Redondo-Nieto M, Kondorosi A, Mergaert P, Kondorosi E: **Genomic organization and evolutionary insights on GRP and NCR genes, two large nodule-specific gene families in *Medicago truncatula*.** *Mol. Plant. Microbe. Interact.* 2007, **20**:1138–48.
32. Maróti G, Downie JA, Kondorosi É: **Plant cysteine-rich peptides that inhibit pathogen growth and control rhizobial differentiation in legume nodules.** *Curr. Opin. Plant Biol.* 2015, **26**:57–63.
33. Guefrachi I, Pierre O, Timchenko T, Alunni B, Barrière Q, Czernic P, Villaécija-Aguilar J-A, Verly C, Bourge M, Fardoux J, et al.: ***Bradyrhizobium BclA* is a peptide transporter required for bacterial differentiation in symbiosis with *Aeschynomene* legumes.** *Mol. Plant-Microbe Interact.* 2015, doi:10.1094/MPMI-04-15-0094-R.
34. Nallu S, Silverstein K a T, Samac D a., Bucciarelli B, Vance CP, VandenBosch K a: **Regulatory patterns of a large family of defensin-like genes expressed in nodules of *Medicago truncatula*.** *PLoS One* 2013, **8**:e60355.
35. Van de Velde W, Zehirov G, Szatmari A, Debreczeny M, Ishihara H, Kevei Z, Farkas A, Mikulass K, Nagy A, Tiricz H, et al.: **Plant peptides govern terminal differentiation of bacteria in symbiosis.** *Science* 2010, **327**:1122–6.
36. Haag AF, Baloban M, Sani M, Kerscher B, Pierre O, Farkas A, Longhi R, Boncompagni E, Hérouart D, Dall'angelo S, et al.: **Protection of *Sinorhizobium* against host cysteine-rich antimicrobial peptides is critical for symbiosis.** *PLoS Biol.* 2011, **9**:e1001169.
37. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796–815.
38. Kohler A, Rinaldi C, Duplessis S, Bauchner M, Geelen D, Duchaussoy F, Meyers BC, Boerjan W, Martin F: **Genome-wide identification of NBS resistance genes in *Populus trichocarpa*.** *Plant Mol. Biol.* 2008, **66**:619–36.

39. Zhou T, Wang Y, Chen JQ, Araki H, Jing Z, Jiang K, Shen J, Tian D: **Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes.** *Mol. Genet. Genomics* 2004, **271**:402–415.
40. Sequencing Consortium TPG: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475**:189–195.
41. Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito V a., Mayer KFX, Gouzy J, Schoof H, et al.: **The *Medicago* genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480**:520–4.
- **42. Yao W, Li G, Zhao H, Wang G, Lian X, Xie W: **Exploring the rice dispensable genome using a metagenome-like assembly strategy.** *Genome Biol.* 2015, **16**:187.

In order to explore the rice pan-genome, the authors sequenced nearly 1500 cultivated rice accessions to low coverage (1-3x) and then treated the data from a metagenomic perspective. They went onto assign genomic positions for 78% of the novel sequences and perform association mapping, demonstrating that a significant percent of phenotypes are associated with loci that reside in the dispensable genome space.

- **43. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A, et al.: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nat. Methods* 2015, **12**:780–786.

The first *de novo* genome sequence (in human or anywhere) based entirely on long-read, single-molecule sequencing, leading to an assembly with a remarkably long N50 (30 Mb) and discovery of SVs missed by other methods.

44. Lam ET1, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY: **Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly.** *Nat Biotechnol.* 2012, **30**:771-6.

45. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions.** *Nat. Biotechnol.* 2013, **31**:1119–25.
46. Putnam NH, O'Connell B, Stites JC, Rice BJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE: **Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage.** *arXiv:1502.05331 [q-bio]* 2015, [no volume].
47. Byrne SL, Nagy I, Pfeifer M, Armstead I, Swain S, Studer B, Mayer K, Campbell JD, Czaban A, Hentrup S, et al.: **A syntenic-based draft genome sequence of the forage grass *Lolium perenne*.** *Plant J.* 2015, doi:10.1111/tpj.13037.
48. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al.: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One* 2012, **7**:e47768.

Table 1. Sequence-Based Structural Variant Discovery

Current Methods	Recipe	Constraints	Implications	Large Gene Families
Read-Mapping	Read Pair; Split-Read; Read Depth	Reliance on reference; Misplacement; Paralog confusion; Signal-to-Noise	CNV miscounts; SVs inconsistent between methods; Pan-genomics impossible	Collapsed, missing and unmapped members; Confounded cluster organization
<i>De novo</i> Short Read	Illumina + ALLPATHS + Genome comparison	Misassembly; Sequence gaps; Collapsed repeats	CNV miscounts; SVs uncertain at gaps; Incomplete gene models	Collapsed members; Short-range architecture; Confounded cluster organization
Emerging Methods	Recipe	Remaining Challenges	Opportunities	Large Gene Families
<i>De novo</i> Long Read	PacBio + Hybrid assembly + Genome comparison	Defining "novel" genes; Constructing valid pan-genome	Minimizes artifacts; Insights into rapid evolution; More reliable SV prediction	Intact tandem clusters; Inter-gene rearrangement and conversion events