Abstract

In comparison to single nucleotide polymorphisms (SNPs), there is currently less knowledge regarding insertions or deletions (INDELs). In this study, we performed whole-genome variant calling and observed that 63% of INDELs in *A. thaliana* natural accessions and 64% in *D. melanogaster* inbred populations were misclassified as multi-allelic, where the same underlying clustered variants were presented as alternative alignments. Here we took multiple sequence alignment strategies to synchronise genetic variants. The improved variant calling was used to perform genome-wide association studies (GWAS). And then the distribution pattern of INDELs in protein coding regions was analysed.

Short INDELs and structural variants have been increasingly used for GWAS, however, only very limited cases were reported where they could improve the traditional SNP-based GWAS. Here, we developed the software Irisas to reclassify and re-annotate genome-wide variants and proposed a burden test as a robust measure that integrates the predicted functional impact of SNPs, INDELs, and structural variants. We re-analysed two publicly available datasets with multiple traits in *A. thaliana* and *D. melanogaster* using our method. When performing GWAS using SNPs, significant variants explained on average 37% and 18% of variance for multiple traits. INDELs and burden tests however explained an additional 10% and 3% of the phenotypic variance in *A. thaliana* and *D. melanogaster*, respectively. The novel loci that previous GWAS have failed to associate with contain established candidate genes, some of which were confirmed by us using transfer DNA (T-DNA) mutant lines. Our study highlights the value of integrated analysis of multiple types of variants for GWAS in plants and animals.

We found that INDELs without causing open reading frame (ORF) -shift are overrepresented in protein coding regions, in agreement with previous studies. In this study, we showed that the combination of double ORF-shift INDELs with overall lengths divisible by three appears at significantly higher frequency than other combinations. With the *A. thaliana* accessions from 1001 Genomes Project, we identified 195 ORF state conserved genes that had double ORF shifting INDELs. In 164 of these 195 genes, we were also able to identify alleles containing only one of the double INDELs, which suggested that the double-INDEL appeared sequentially. Focusing on these 195 genes, the analysis of protein domain indicated that a majority of affected domains were rescued by the second INDEL. All reference alleles, single-INDEL alleles and double-INDEL alleles were widely distributed in terms of both

geographical position and population structure. Balancing selection might be the driving force of this two-step evolution.

Zusammenfassung

Im Vergleich zu Single nucleotide polymorphisms (SNPs) sind Insertionen oder Deletionen (INDELs) wenig erforscht. In dieser Studie haben wir eine Suche nach Sequenzvarianten in vollständig sequenzierten Genomen durchgeführt und haben herausgefunden, dass 63% der INDELs in *A. thaliana* und 64% in *D. melanogaster* falsch als multi-allelisch klassifiziert sind, wo dieselben Cluster von Sequenzvarianten als unterschiedliche Alignments dargestellt werden. Wir haben hier Mehrfach-Alignment-Strategien verwendet, um die genetischen Sequenzvarianten zu synchronisieren. Die Verbesserte Suche nach Sequenzvarianten wurde dann für genomweite Assoziationsstudien (GWAS) genutzt, und das Verteilungsmuster von INDELs in Protein kodierenden Regionen wurde auch analysiert.

Kurze INDELs und strukturelle Varianten werden immer häufiger für GWAS genutzt, jedoch in nur wenigen Fällen wurde eine Verbesserung gegenüber herkömmlicher SNP-basierenden GWAS erzielt. Hier haben wir die Software Irisas entwickelt, die genomweit Sequenzvarianten re-klassifiziert und re-annotiert, und schlagen einen Belastungstest als robustes Maß vor, der den funktionellen Einfluss von SNPs, INDELs und strukturellen Varianten vorausberechnet und integriert. Wir haben zwei öffentlich zugängliche Datensätze mit multiplen Phänotypen in *A. thaliana* und *D. melanogaster* mit unserer Methode re-analysiert. In GWAS mit herkömmlichen SNPs erklären signifikante Sequenzvarianten 37% und 18% der multiplen Phänotypen. INDEL und Belastungstest erklärten jedoch zusätzlich 10% und 3% der phänotypischen Variation in *A. thaliana* beziehungsweise *D. melanogaster.* Die neuen Loci, die herkömmliche GWAS nicht identifizieren konnten enthalten gut etablierte Kandidatengene, von denen einige durch T-DNA Insertionslinien bestätigt werden konnten. Unsere Studie unterstreicht den Wert von integrierten Analysen, die multiple Sequenzvarianten für GWAS in Pflanzen und Tieren nutzen.

Wir konnten zeigen, dass INDELs, die keine Verschiebung des Leserasters verursachen in Protein-codierenden Regionen überrepräsentiert sind, was im Einklang mit vorherigen Studien ist. Des weiteren konnten wir zeigen, dass die Kombination von doppelten INDELs mit Verschiebungen des Leserasters zusammen mit der Teilbarkeit durch drei signifikant häufiger vorkommt als andere Kombinationen. In den *A. thaliana* Linien des 1001 Genom Projekts identifizierten wir 195 doppelt-INDELs mit konserviertem Leseraster. Für 164 von diesen 195 Genen waren wir in der Lage eine andere Allele zu identifizieren, die nur eines

der beiden INDELs hat, was nahe legt, dass die doppelt-INDELs sequenziell aufgetreten sind. Die fokussierte Analyse der Proteindomäne dieser 195 Gene weist darauf hin, dass die Mehrheit dieser Proteindomäne durch ein zweites INDEL gerettet wurden. Alle Referenz-Allele, einfach-INDELs und doppelt-INDELs waren weit verbreitet, sowohl geographisch als auch populationsgenetisch. Unsere Ergebnisse suggerieren, dass ausgleichende Selektion die treibende Kraft dieser zweistufigen Evolution sein könnte.