Sarah Gagliano ORCID iD: 0000-0003-1306-1868

Relative impact of indels versus SNPs on complex disease

Sarah A Gagliano[1,#], Sebanti Sengupta[1], Carlo Sidore[2], Andrea Maschio[2], Francesco Cucca[2,3], David Schlessinger[4], Gonçalo R Abecasis[1,#]

1 Center for Statistical Genetics, and Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

2 Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy

3 Dipartimento di Scienze Biomediche, Università degli Studi di Sassari, Sassari, Italy

4 Laboratory of Genetics, National Institute on Aging, US National Institutes of Health, Baltimore, Maryland, USA

Corresponding author

Correspondence: 1415 Washington Heights

School of Public Health

Ann Arbor, Michigan

48109 USA sarah.gagliano@umich.edu; gonçalo@umich.edu short title/running title: indels versus SNPs

Abstract

It is unclear whether insertions and deletions (indels) are more likely to influence complex traits than abundant SNPs. We sought to understand which category of variation is more likely to impact health.

This is the author manuscript accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/gepi.22175.

Using the SardiNIA study as an exemplar, we characterized 478,876 common indels and 8,246,244 common SNPs in up to 5,949 well-phenotyped individuals from an isolated valley in Sardinia. We assessed association between 120 traits, resulting in 89 non-overlapping associated loci.

We evaluated whether indels were enriched among credible sets of potential causal variants. These credible sets included 1,319 SNPs and 88 indels. We did not find indels to be significantly enriched. Indels were the most likely causal variant in 7 loci, including one locus associated with monocyte count where an indel with causality and mechanism previously demonstrated (rs200748895:TGCTG/T) had a 0.999 posterior probability. Overall, our results show a very modest and non-significant enrichment for common indels in associated-loci.

Keywords: indels, genome-wide association, complex traits

Introduction

The relative impact of insertion and deletion variants (indels) and single nucleotide polymorphisms (SNPs) on human complex disease risk is unclear. By definition, a SNP changes a single nucleotide in the DNA sequence, whereas an indel incorporates or removes one or more nucleotides (Loewe, 2008).

SNPs in coding and non-coding regions have been implicated in both Mendelian and complex disease, and the same is true for indels. In coding regions, an insertion or deletion that is not in-frame (a multiple of three base pairs) will alter the reading frame resulting in a new set of amino acids and thus a protein product that

differs to the wild type. The presence of 40 or more CAG repeats in the first exon of the huntingtin gene (*HTT*) results in Huntington's disease (Lench et al., 2013). Even in-frame indels (insertions of deletions of three or multiples of three base pairs) in the coding sequence can also result in altered proteins. An example is a deletion in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene that removes one amino acid (phenylalanine) at position 508, which arrests protein function and leads to cystic fibrosis (Mullaney et al., 2010). With regard to non-coding regions, as with SNPs, indels may have effects on the affinity of a binding site for a regulatory factor or transcriptional machinery, or on chromatin structure. For example, insertions within the promoter region of the *FMR1* gene can cause Fragile X syndrome (Mills et al., 2006), and insertions within the promoter region of the *SNCA* gene contributes to autosomal dominant Parkinson's disease (Singleton et al., 2003).

There is a balance between mutation rates and selective constraint of indels, particularly in the coding sequence as reading frames should be maintained and thus the protein function preserved. Coding indels tend to be subject to stronger purifying selection than SNPs (Montgomery et al., 2013, de la Chaux et al., 2007).

The cumulative contribution of indels compared to SNPs to disease risk has not been thoroughly investigated. To fill this gap in our understanding, we assessed associations of indels and SNPs with 120 traits in a sample of up to 5,949 individuals from the island of Sardinia. We limited our analysis to common variants due to the nature of the study design. Acknowledging that this work lacks insight from rare variation, it nevertheless begins to move toward a better understanding of which type

of polymorphism is more likely to impact human health, and to quantify the gain by routinely including indels in GWAS.

Results

i. Imputation of Indels

A total of 928,605 out of 1,156,646 autosomal indels remained after imputation RSQR thresholds were applied, and they are distributed throughout the autosomes (**Supporting Information Figure S1**). With regard to autosomal SNPs, 17,607,889 out of 24,106,694 passed the RSQR thresholds. Applying a minor allele frequency (MAF) ≥1% cut-off to ensure the inclusion of variants with high imputation quality, there were 8,725,120 variants genome-wide (478,876 indels and 8,246,244 SNPs). All downstream analyses involve this filtered set of variants.

Imputation quality summary metrics are displayed in **Supporting Information Figure S2**.

i. Annotation of Indels

The vast majority of indels are non-coding, and only 0.2% (760) of the MAF≥1% indels fall into a coding region as defined by GENCODE v19 (Harrow et al., 2012). 58% of the indels are deletions and 42% are insertions (276,508 and 202,367, respectively). This inequality in proportions is likely due to additional challenges aligning reads containing insertions larger than the fragment size in the sequencing library (Medvedev et al., 2009).

There were fewer indels in coding regions compared to non-coding regions than expected by chance (**Table 1**). This lower relative density of indels in the coding region has been seen previously in other datasets (Mullaney et al., 2010, Lek et al., 2016).

56% of the indels are the insertion or deletion of a single base (148,355 deletions and 120,098 insertions).

We assessed the proportion of SNPs and indels within 1Mbp of associated loci (see **Table 2**) in regions of low complexity (amino acid sequences that contain repeats of single amino acids or short amino acid motifs making these regions more difficult to call) (Morgulis et al., 2006). These regions represent 8.8% of the whole genome autosomal sequences (254,665,411 base pairs). There were significantly more indels than SNPs found in these regions (chi-square= 1441.1, p= 2.5E-315), likely due to high error rate in variant calling. Of the 18,325 indels found within 1Mbp of associated loci we detected 4.3% (792) to be in regions of low complexity. Of the 308,310 SNPs found within 1Mbp of associated loci we detected 1.1% (3,354) to be in regions of low complexity.

6.3% of indels (30,124) were not found in the variant list of the NHLBI Trans-Omics for Precision Medicine (TOPMed) high-depth whole-genome sequencing effort.

ii. Association analyses

Using the MAF≥1% cutoff, 51 of the 120 traits tested had at least one variant that reached genome-wide significance (p≤5E-8), and for 33 of those 51 traits, at least one indel reached genome-wide significance. These association results allowed us to assess the relative enrichment of indels and SNPs among trait-associated variants. There were 9,474 variants that reached genome-wide significance, of which 494 are indels, in at least one of the traits tested for association.

Of the significant indels, 19 are not found in the TOPMed variant list).

i. Impact of indels versus SNPs

We investigated whether indels are more likely than SNPs to be potentially causal. To obtain an estimate of indel enrichment among potentially causal variants we assessed the proportion of indels to SNPs within 1Mbp of associated loci (N=89) compared to the rest of the genome for variants MAF≥1%. We set a wide base pair range to ensure that all possible causal variants would be included in the computation of credible sets (see below) regardless of the linkage disequilibrium structure at the loci. Indels were not significantly enriched (estimate= e^(0.09)= 1.09; p= 0.88).

In order to address the inherent genomic alignment and calling challenges in regions of low complexity, we removed SNPs and indels that fall into those regions and then repeated the analysis to estimate the enrichment parameter. 4% of indels (N= 792) in the associated loci were removed, and 1% of SNPs (N= 3,354) were removed.

The estimated enrichment parameter remained non-significant (estimate= 1.41, p= 0.64).

We obtained 95% credible sets of potentially causal variants. Of the variants in the credible sets, the distribution of effect sizes did not significantly differ between indels and SNPs (Mann-Whitney U test p= 0.91). Indels were the most likely causal variant in in 7 out of the 89 associated loci assessed. One of those sets contained only one variant, solely an indel in the 3-prime UTR of the *TNFSF13B* gene (rs200748895:TGCTG/T, chi-square = 24.9, posterior probability= 0.999) for association with monocyte count. This complex polymorphism has been identified as the causal variant at this locus in previous work (Steri et al., 2017). Thus, the identification of a known causal variant provided us with reassurance of the utility of our method. Of variants with a posterior probability ≥0.1, 8% (14/182) were indels. Although 8% of indels with a posterior probability ≥0.1 within the credible sets is not more than expected by chance (chi-square= 0.74, p= 0.39), this percentage nevertheless is higher than the proportion of indels in the total number of variants.

As a positive control for the enrichment parameter, we assessed whether missense SNPs are enriched in trait-associated loci given their direct consequence on the amino acid chain and thus the resulting protein. The estimated enrichment parameter showed that missense SNPs are more likely to be potentially causal than other variants (estimate= e^(3.92)= 50.4, p= 1.6E-10) (**Table 2**). Relatively there are more indels than missense SNPs in the genome, and thus the non-significant enrichment results for the indel versus SNP analysis is unlikely due to a lack of power possibly a result of the lower imputation quality in indels compared to SNPs. Genome-wide there were

31,112 missense SNPs with MAF ≥1% in the dataset, of which 7.6% (2,356) fell into trait-associated loci.

As a complementary analysis to the missense analysis we also performed a coding indel enrichment analysis. Of the 760 indels in the dataset with MAF ≥1% that fall into coding sequences, 8.0% (61) were in trait-associated loci. The estimated enrichment parameter showed that coding indels are not more likely to be potentially causal than other variants (estimate= e^(-16.3) = 8.3E-8, p= 0.90) (**Table 2**). However, we acknowledge the lack of power in this particular sub-analysis of a small subset of variation.

Discussion

Using association results from the SardiNIA cohort of up to 5,949 individuals for 120 traits, we did not find evidence of common indels more likely to be potentially causal than SNPs with regard to associations to complex traits. On a similar note but looking at only the coding sequence, Montgomery et al. (2013) did not find direct evidence that potentially causal classes of coding indels are enriched for associations compared to known disease-associated SNPs present in the GWAS Catalog.

The modest sample size in our study limits the capacity to identify causal variants. However, our analysis strategy allowed us to evaluate enrichment of indels at loci even in situations where we could not pinpoint an individual causal variant, which may require studies with larger sample sizes or multiple ancestries. We also acknowledge that a subset of variants achieving the widely accepted genome-wide

significant p-value threshold ($p \leq 5E\text{-}8$) could be false positive signals. Additionally, we applied a MAF threshold to ensure the integrity of the imputed genotypes, but in doing so we removed potentially causal rare variants, possibly biasing our analysis. Future studies with larger sample size will help in addressing these limitations by increasing the statistical power. Finally, *in vitro* and *in vivo* experimental designs are required to verify the functionality of the variants in question. We employ an *in silico* method to address potential "causality", which can guide the choice of variants to carry forward to these subsequent experiments.

Investigation into the relative impact of common and also lower frequency indels compared to SNPs in the context of larger more diverse samples and more phenotypes is warranted.

Methods

ii. SardiNIA study dataset

In brief, we genotyped 6,602 individuals from four villages in the Lanusei valley on Sardinia (>60% of the adult population). Each sample was genotyped on four different Illumina Infinium arrays: OmniExpress, Cardio-Metabochip (Voight et al., 2012), Immunochip (Parkes et al., 2013), and Exome Chip. We also performed low-depth (~4x coverage) whole-genome sequencing on 3,839 individuals, 2,340 of whom we also genotyped. Study samples, genotyping, sequencing and variant calling have been previously described (Sidore et al., 2015).

Over 100 traits (e.g. blood lab measurements, anthropometric values) have been measured at 4 to 5 time-points. We looked at 120 traits from the first visit, for which the majority of the individuals have measurements (median number of samples with at least one measurement per trait= 5814, first quartile= 5473, third quartile= 5923). The traits have been previously summarized (Pilia et al., 2006).

iii. Imputation

We imputed autosomal SNPs and indels for the individuals who were successfully genotyped on all four arrays (N=6,602) using Minimac3 (Das et al., 2016) using the Sardinia sequencing data as the reference panel with genomic locations corresponding to GRCh37. This sequencing panel has 1,156,646 biallelic indels and 24,106,694 biallelic SNPs. We imputed SNPs based on only the phased SNPs in the reference panel (i.e. indels removed), and then we imputed indels based on both the SNPs and indels in the reference panel.

For very rare SNPs (i.e. MAF <0.5%) we have shown (data not published) that imputation with a Sardinian reference panel outperforms imputation with the reference panel of the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016) samples, which includes 3,445 Sardinian individuals.

After imputation, we retained only markers with an imputation quality (RSQR) >0.3 or >0.6 if the estimated minor allele frequency (MAF) was ≥1% or <1%, respectively (Pistis et al., 2015).

iv. Annotation of Indels

The likely functional impact of variants was annotated using the Ensembl Variant Effect Predictor (McLaren et al., 2010) to annotate consequences. VT (Tan et al., 2015) was used to annotate DUST low complexity regions (Morgulis et al., 2006). We also looked at the MAF distributions, and the lengths of the indels.

We identified which indels were not found in any of the populations in the NHLBI TOPMed release 3a variant list (NHLBI TOPMed Project Freeze 3a. https://www.nhlbiwgs.org). The 3a release contains 170 million variants on 14,559 individuals, and was accessed through the BRAVO browser (https://bravo.sph.umich.edu). Indels were considered present in the TOPMed Project if there was an exact match by chromosome, start position, reference allele and alternate allele.

v. Association analyses

Association analyses were performed for 120 quantitative traits measured in 1,460 – 5,949 individuals (median= 5,814) from Visit 1 of the SardiNIA cohort. Associations were run in EPACTS (Kang, H.M., Zhan, X., Sim, X., Ma, C. EPACTS: A flexible and efficient sequence-based genetic analysis software package; Presented at the 62nd Annual Meeting of The American Society of Human Genetics, November 2012, San Francisco) using the age, age^2 and sex-adjusted inverse-normalized residuals of the outcomes as input to the Efficient Mixed Model Association eXpedited (EMMAX) (Kang et al., 2010) single variant test (i.e. a linear model with a kinship matrix).

vi. Impact of indels versus SNPs

To obtain an estimate of indel enrichment among potentially causal variants we assessed the proportion of indels to SNPs within 1Mbp of associated loci compared to the rest of the genome (Sengupta et al. *In prep*). We used a filter of MAF≥1%, and estimated an enrichment parameter, which denotes how much more likely indels are potentially causal compared to SNPs. This iterative procedure of essentially maximizing a log odds ratio of a two-by-two table to obtain the enrichment parameter is summarized in **Figure 1**. If this estimate was statistically significant, we would use it to calculate the posterior probability of each variant being causal. To illustrate, say a SNP and an indel have similar p-values, but the enrichment parameter suggests the enrichment of indels at associated loci. The indel would thus receive a higher posterior probability.

For the traits, we identified all of the variants (MAF≥1%) with a significant GWAS p-value (p≤5E-8). We took 500K base pairs downstream and 500K base pairs upstream the most significant variant to obtain the locus. For any overlapping loci, within a trait or among traits, we retained the locus with the most significant p-value and dropped the other loci. Then we obtained all of the variants (MAF≥1%) within the non-overlapping 1Mbp loci, and annotated them with regard to being a SNP or an indel.

As a positive control, for the same non-overlapping 1Mbp associated loci used for the indel versus SNP analysis, we re-annotated the variants, this time with regard to

being a either a missense SNP or not a missense SNP or an indel. As a complementary analysis, we also re-annotated the variants with regard to being a coding indel or not, where coding regions were defined by GENCODE v19.

Conflicts of Interests

The authors declare no conflicts of interests.

Acknowledgements

We thank all the volunteers who generously participated in this study and made this research possible. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

References

Das, S. *et al.* (2016). Next-generation genotype imputation service and methods. *Nat. Genet.*, 48, 1284–1287. *doi*: 10.1038/ng.3656

de la Chaux, N., Messer, P.W., & Arndt, P.F. (2007). DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol.*, 7, 191. *doi:*10.1186/1471-2148-7-191

Harrow, J. *et al.* (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), 1760-1774. *doi*:10.1101/gr.135350.111

Kang, H.M. *et al.* (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.*, 42, 348-54. *doi*:10.1038/ng.548

Lek, M. *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536, 285-291. *doi*:10.1038/nature19057

Lench, N. *et al*. (2013). The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made. *Prenat. Diagn.*, 33, 555–562. *doi*:10.1002/pd.4124

Loewe, L. (2008) Genetic mutation. *Nature Education* 1(1):113

McCarthy S. *et al*. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*., 48, 1279–1283. *doi*:10.1038/ng.3643

McLaren, W. *et al*. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26, 2069–2070. *doi*:10.1093/bioinformatics/btq330

Medvedev P., Stanciu M., & Brudno M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6, S13–S20. *doi*:10.1038/nmeth.1374

Mills, R.E. *et al.* (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16(9), 1182-1190. *doi*:10.1101/gr.4565806

Montgomery, S. *et al.* (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.*, 23(5), 749-761. *doi*:10.1101/gr.148718.112

Morgulis, A., Gertz, E.M. Schäffer, A.A. & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, 13(5), 1028-1040. *doi*:10.1089/cmb.2006.13.1028

Mullaney, J.M., Mills, R.E., Pittard, W.S., & Devine, S. (2010). Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet.*, 19(R2), R131-R136. *doi*:10.1093/hmg/ddq400

Parkes, M., Cortes, A., van Heel, D.A., & Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.*, 14, 661–673. *doi*:10.1038/nrg3502

Pilia, G. *et al.* (2006). Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genetics*, 2(8):e132. *doi*:10.1371/journal.pgen.0020132

Pistis, G. *et al*. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.*, 23, 975–983. *doi*:10.1038/ejhg.2014.216

Sidore, C. *et al.* (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.*, 47, 1272-1281. *doi*:10.1038/ng.3368

Singleton, A.B. *et al.* (2013). α-Synuclein Locus Triplication Causes Parkinson's Disease. *Science*, 302(5646), 841. *doi*:10.1126/science.1090278

Steri, M. *et al.* (2017). Overexpression of the Cytokine BAFF and Autoimmunity Risk. *NEJM*, 376, 1615-1626. *doi*:10.1056/NEJMoa1610528

Tan, A., Abecasis, G.R., & Kang, H.M. (2015). Unified representation of genetic variants. *Bioinformatics*, 31(13), 2202-2204. *doi*:10.1093/bioinformatics/btv112

Voight, B.F. *et al*. (2012). The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* 8, e1002793. *doi*:10.1371/journal.pgen.1002793

Tables

Table 1. Numbers of coding and non-coding variation (MAF≥1%) in SardiNIA

	Coding	Non-coding	Total
SNP	60,844 (0.7% of SNPs)	8,185,400 (99.3% of SNPs)	8,246,244 (95% of variants)
Indel	**760** (0.2% of Indels)	478,116 (99.8% of indels)	478,876 (5% of variants)
Total	83,223	8,641,897	8,725,120

Table 2. Enrichment results for indels and SNPs in associated loci and for controls

Category	% within 1Mbp associated loci	Enrichment parameter, λ (P)
INDEL vs. SNP	3.8% (18,325) vs. 3.7% (308,310)	0.09 (0.88)
Missense vs. not	7.6% (2,356) vs. 3.7% (324,279)	3.92 (1.6E-10)
Coding INDEL vs. not	8.0% (61) vs. 3.7% (326,574)	(-16.3) (0.90)

Figures

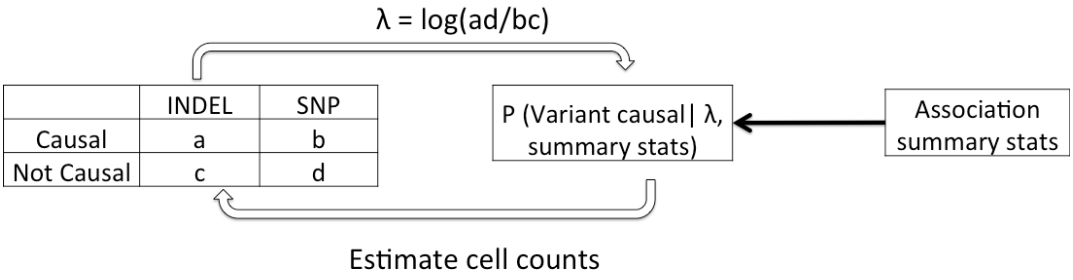

Figure 1. Iterative procedure to estimate enrichment parameter, λ.