Position Kernels as a Key to Making Sense of Very Rare and Private Single-Nucleotide Variants

Ulrich Bodenhofer
LIT AI Lab & Institute of Bioinformatics
Johannes Kepler University Linz, Austria
bodenhofer@bioinf.jku.at

Sepp Hochreiter
LIT AI Lab & Institute of Bioinformatics
Johannes Kepler University Linz, Austria
hochreit@bioinf.jku.at

Abstract

We present an approach for convolving single-nucleotide variants (SNVs) with a position kernel in order to augment SNVs with information about close-by SNVs. By means of the Position-Dependent Kernel Association Test (PODKAT), we demonstrate the potential of this approach to leverage the analysis of rare and private SNVs. Finally, we also provide some ideas how machine-learning based predictions from genomic data can benefit from this augmentation.

1 Introduction

High-throughput sequencing technologies have facilitated the identification of large numbers of single-nucleotide variants (SNVs). Genome-wide association studies [7, 15] have become standard in statistical genetics and helped to find many associations of SNVs with diseases or other traits. While it is common to test for statistical associations between single SNVs and the traits, these single-marker tests are generally underpowered for rare SNVs and for complex traits that involve interactions of SNVs on multiple loci. So, many genetic influences remain elusive, where rare SNVs are supposed to play one of the crucial roles in this "missing heritability" [16]. In order to cope with the statistical challenges of analyzing rare SNVs, different collapsing strategies have been proposed with the aim to improve statistical power by considering multiple SNVs occurring in a region simultaneously. Such strategies can be classified into burden tests and non-burden tests [19], where the acclaimed SNP-set Kernel Association Test[1] (SKAT) [23] is an important representative of the latter.

Several large sequencing studies, such as, the 1000 Genomes Project [21], the UK10K project [22], or the NHLBI-Exome Sequencing Project [20], have consistently reported a large proportion of private SNVs, that is, SNVs that are unique to a family or even a single individual. Non-burden tests like SKAT are typically utilizing correlations between SNVs to increase statistical power — a strategy that is not applicable to private SNVs, since singular events are generally uncorrelated. Burden tests are potentially able to deal with private SNVs, but only if the number of private SNVs occurring in a region is correlated with the trait under consideration. Moreover, burden tests have a disadvantage if deleterious and protective SNVs occur together in the same region.

Association tests try to find only statistical correlations between SNVs and traits. With recent advances in machine learning, the prediction of phenotypes (traits, diseases, etc.) from genotypes has become increasingly feasible. Among other methods, deep networks have a great potential in this direction, but the large number of inputs in conjunction with the relatively low number of samples poses a serious problem, since very large numbers of parameters need to be learned from relatively few samples. Diet networks [18], for instance, address this challenge by a parametrization technique that reduces the number of free parameters. This is achieved by learning the weights between inputs and the first hidden layer using an auxiliary network that utilizes the similarities between SNVs

[1] formerly known as Sequence Kernel Association Test

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

across samples. While this is generally a very attractive idea that works well for common SNVs, rare and private SNVs will generally expose little similarity to other SNVs, so the auxiliary network will not be able to meaningfully generalize weights across SNVs.

In this contribution, we present an approach for convolving SNVs with a position kernel. This approach follows the assumption that, the closer two SNVs are on the genome, the more likely they have similar effects on the trait under consideration. This assumption is fulfilled as long as deleterious, neutral, and protective variants are grouped sufficiently well along the genome. Convolving the genotype matrix with a position kernel then allows for uncovering positional similarities even of very rare and private SNVs in a way that they can be used directly as inputs for further processing, be it an association test or a predictive model. We motivate this approach by introducing the Position-Dependent Kernel Association Test (PODKAT) [2, 1], a generalization of SKAT which uses this idea to better deal with very rare and private SNVs. Finally, we will expose possibilities to use this approach in other settings as well.

2 The Position-Dependent Kernel Association Test (PODKAT)

In line with SKAT [23], PODKAT uses a variance component score test to test for associations between genotypes and traits. PODKAT and SKAT assume that traits are distributed according to the following semi-parametric mixed models:

$$\operatorname{logit}\big(p(y=1)\big) = \alpha_0 + \boldsymbol{\alpha}^T \cdot \boldsymbol{x} + h(\boldsymbol{z}) \qquad \text{(if trait is binary)}$$

$$y = \alpha_0 + \boldsymbol{\alpha}^T \cdot \boldsymbol{x} + h(\boldsymbol{z}) + \varepsilon \qquad \text{(if trait is continuous)}$$

In the above formulas, y is the trait, $\boldsymbol{x}$ is the covariate vector (if any), $\boldsymbol{z}$ is the genotype vector, α_0 is the intercept, $\boldsymbol{\alpha}$ are the fixed effect coefficients, $h(.)$ is an unknown centered smooth function and ε is the error term. SKAT and PODKAT both assume that the function $h(.)$ is from a function space that is generated by a given positive semi-definite kernel function $K(.,.)$ [12].

The null hypothesis is that y is not influenced by the genotype:

$$p(y=1) = \operatorname{logit}^{-1}\big(\alpha_0 + \boldsymbol{\alpha}^T \cdot \boldsymbol{x}\big) \qquad \text{(if trait is binary)}$$

$$y = \alpha_0 + \boldsymbol{\alpha}^T \cdot \boldsymbol{x} + \varepsilon \qquad \text{(if trait is continuous)}$$

As mentioned above, we use a *variance component score test* [11, 12, 23] to test against the null hypothesis. The test statistics is given as

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T \cdot \mathbf{K} \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}}),$$

where $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is the vector of residuals of the (logistic) linear model has been trained to fit traits to the covariates only (the so-called *null model*) and $\mathbf{K}$ is a positive semi-definite kernel matrix defined as $K_{i,j} = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$, where $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are the genotypes of the i-th and j-th sample. The kernel matrix $\mathbf{K}$ measures the pairwise similarity of genotypes of samples. Since $\mathbf{K}$ is positive semi-definite, Q is non-negative. The more structure the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ and the matrix $\mathbf{K}$ share, the larger Q. If the residuals and the genotypes are independent, i.e. if the test's null hypothesis holds true, large values can only occur by pure chance with a low probability. Hence, we test whether the actually observed Q value is higher than expected by pure chance. More specifically, the test's p-value is computed as the (estimated) probability of observing a value under the null hypothesis that is at least as large as the observed Q. For continuous traits and normally distributed noise ε, the residuals are normally distributed and the distribution of Q is obviously a *mixture of* χ^2 *distributions*. For binary traits, Q approximately follows a *mixture of* χ^2 *distributions*, too [11, 23]. This null distribution can be estimated efficiently to compute p-values without permutation testing [5, 13].

As mentioned above, the choice of a kernel function that computes the pairwise similarities of the samples' genotypes is essential. The simplest kernel is the *linear kernel* that computes the similarities as the outer product of the genotype matrices: $\mathbf{K} = \mathbf{Z} \cdot \mathbf{Z}^T$. This kernel is a standard choice in SKAT as well as the *weighted linear kernel* which weighs the genotype matrix before computing the outer product, i.e. $\mathbf{K} = \mathbf{Z} \cdot \mathbf{W} \cdot \mathbf{W}^T \cdot \mathbf{Z}^T$, with $\mathbf{W}$ being a diagonal weight matrix that assigns weights to SNVs. By these weights, it becomes possible to put more emphasis on rare emphasis than on common SNVs. The weights are usually functions of the minor allele frequency of the SNVs.

PODKAT convolves the genotype matrix with a position kernel before computing the linear kernel:

$$\mathbf{K} = \mathbf{Z} \cdot \mathbf{W} \cdot \mathbf{P} \cdot \mathbf{P}^T \cdot \mathbf{W}^T \cdot \mathbf{Z}^T,$$

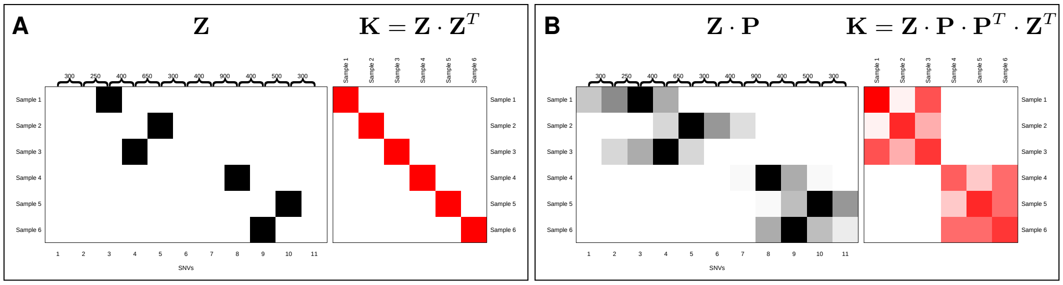

Figure 1: Toy example demonstrating how the position kernel takes private SNVs into account. **A:** genotype matrix $\mathbf{Z}$ (left) and kernel matrix of the linear kernel (right); **B:** convolution of genotype matrix $\mathbf{Z}$ with position kernel (left) and resulting kernel matrix (right).

The matrix $\mathbf{P}$ is a positive semi-definite kernel matrix that measures the similarities/closeness of positions of SNVs [3], i.e.

$$P_{i,j} = \max\left(1 - \tfrac{1}{r}|\mathrm{pos}_i - \mathrm{pos}_j|, 0\right),$$

where pos_i and pos_j are the genomic positions of the i-th and the j-th SNV, respectively. The hyperparameter r determines the *maximal radius of tolerance* beyond which two SNVs are considered completely dissimilar. Note that, by this approach, PODKAT allows for a smooth interpolation between SKAT (for $r < 1$) and a weighted burden test [14] (for $r \to \infty$). In that respect, PODKAT resembles SKAT-O [10], but without inheriting the burden test's disadvantage that the presence of deleterious and protective SNVs in the same window dilutes detection power.

Fig. 1 shows a toy example that demonstrates how the position kernel takes private SNVs into account while the linear kernel is unable to do that. On the left, Panel A, shows a genotype matrix $\mathbf{Z}$ of 6 samples each of which has one private SNV (minor alleles are 1's and major alleles are 0's in the matrix). The right-hand side of Panel A shows the kernel matrix that would be obtained for the linear kernel. Since all SNVs are private, the kernel matrix is diagonal, which does not allow for any meaningful association testing. The left side of Panel B shows the convolution of $\mathbf{Z}$ with the position kernel matrix $\mathbf{P}$, and the right side shows the resulting kernel matrix. Suddenly, two blocks of samples become visible. If, as an example samples 1–3 are cases and samples 4–6 are controls, PODKAT would be able to detect an association, while SKAT with the linear kernel would fail.

The test statistic of PODKAT can be decomposed into individual contributions of single variants:

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T \cdot \mathbf{Z} \cdot \mathbf{W} \cdot \mathbf{P} \cdot \mathbf{P}^T \cdot \mathbf{W}^T \cdot \mathbf{Z}^T \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \|\mathbf{P}^T \cdot \mathbf{W}^T \cdot \mathbf{Z}^T \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}})\|^2.$$

So, Q is the squared norm of a vector $\boldsymbol{p} = \mathbf{P}^T \cdot \mathbf{W}^T \cdot \mathbf{Z}^T \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}})$ with one entry per SNV that can be interpreted as the SNV's contribution to the test statistic Q. The signs of the entries of $\boldsymbol{p}$ indicate the direction of the association (deleterious or protective) and normalizing the squares of contributions allows for quantifying relative contributions.

3 Results

To validate PODKAT's actual ability to better deal with rare and private SNVs, we performed an extensive set of simulation experiments. Fig. 2 shows some results obtained for different numbers of simulated genomes with simulated traits. The results have been obtained for a window size of 5kb and with a genome-wide significance threshold of $\alpha = 10^{-6}$. The plots show that, for given sample size, PODKAT achieves better detection power. For a given detection power, PODKAT potentially suffices with fewer samples. For different significance thresholds and window sizes, the insights are similar (results not shown). If unweighted variants of SKAT and PODKAT are used, the results are much worse (not shown) because the traits have intensionally be simulated such that rare and private SNVs have a higher effect size.

Similar results were obtained for whole-genome and whole-exome data from the UK10K project [22] with simulated traits. However, type I error simulations have shown that SKAT with the weighted IBS kernel [23] has strongly inflated p-values for binary traits, so the results shown by the orange curve in the left panel of Fig. 2 may be overly optimistic.

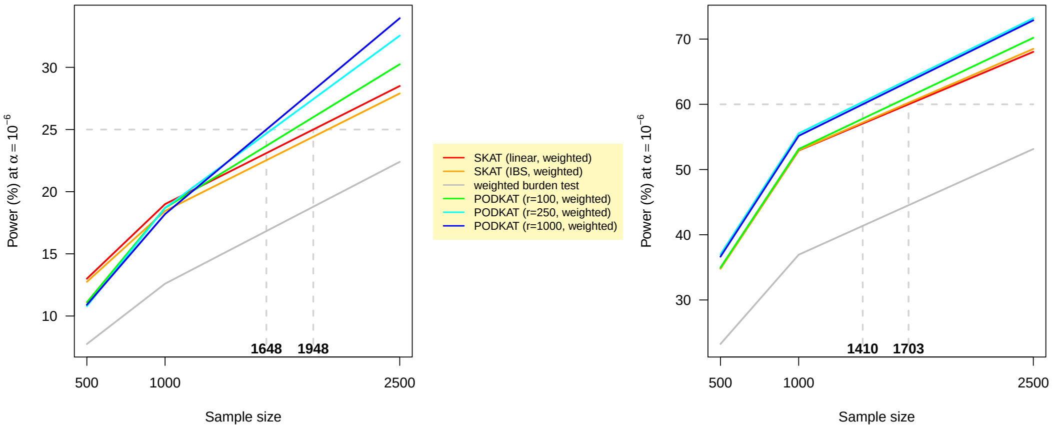

Figure 2: Power simulations for binary (left) and continuous traits (right) using different weighted variants of SKAT and PODKAT.

Results with real-world data were also highly encouraging. For a study investigating genetic associations with intolerances to nonsteroidal anti-imflammatory drugs (NSAIDs) [4], a new association with a rare SNV has been found. We also investigated all phenotypes available for the two whole-genome data sets within the UK10K project which are subsets of the Avon Longitudinal Study of Parents and Children (ALSPAC) [6] and the TwinsUK study [17]. While many of the found associations were already known and also found with other methods, there were indeed some interesting new associations in previously unknown, also non-coding, regions. Though being promising, the results are preliminary and require more detailed interpretations and follow-up.

PODKAT's ability to quantify the contributions of individual SNVs can also be used for feature selection. A recent paper uses machine learning to predict antibiotics resistances of various *Pseudomonas aeruginosa* strains from their genomes [9]. In a first step, PODKAT is applied window-wise and candidate SNVs are selected on the basis of their individual contributions to the associations. Then the Potential Support Vector Machine (PSVM) [8] is used to actually predict antibiotics resistances from the set of selected SNVs. This reduction of input dimension is indispensable for tasks in which the number of SNVs is too large for feeding them into a predictor directly. However, even in cases where a machine learning model can be applied to the genotype matrix directly, the PODKAT-based feature selection has often turned out to be advantageous.

4 Conclusion and Future Prospects

The results above demonstrate that convolving SNV data with a position kernel can leverage the analysis of very rare and private SNVs. This is true for the association test PODKAT which is still in line with the traditional approach of association testing. The results shown in [9], however, indicate that this idea of putting rare and private SNVs in context with their genomic proximity has great potential also when actually making predictions from genomic data using machine learning methods. Though this is only a vague idea at the moment, convolutions with a position kernel may also help machine learning methods to learn from SNVs without using any association test like PODKAT. Reconsider Fig. 1: convolving the genotype matrix with a position kernel (see left graphics in panel **B**) augments each SNV with information about other SNVs in its proximity. That may ease predictions if this convolved genotype matrix is fed into a machine learning method directly, be it a deep neural network or any other suitable method. Furthermore, an approach like Diet Networks [18] would benefit from this augmentation too, since similarities between close-by rare/private SNVs would suddently become apparent to the auxiliary network. More specifically, similar input weights will be assigned to rare/private SNVs that are close to each other. Without the convolution with a position kernel, the SNVs would be dissimilar and the auxiliary network would not be able to learn any meaningful representation of input weights for these SNVs.

References

[1] U. Bodenhofer. *PODKAT: An R Package for Association Testing Involving Rare and Private Variants*, 2015. Software Manual.

[2] U. Bodenhofer and S. Hochreiter. PODKAT: a non-burden test for associating complex traits with rare and private variants. In *63rd Annual Meeting of the American Society of Human Genetics*, Boston, October 2013. (poster).

[3] U. Bodenhofer, K. Schwarzbauer, M. Ionescu, and S. Hochreiter. Modeling position specificity in sequence kernels by fuzzy equivalence relations. In J. P. Carvalho, D. Dubois, U. Kaymak, and J. M. C. Sousa, editors, *Proc. Joint 13th IFSA World Congress and 6th EUSFLAT Conference*, pages 1376–1381, Lisbon, July 2009.

[4] J. A. Cornejo-García, L.-B. Liou, N. Blanca-López, I. Doña, C.-H. Chen, Y.-C. Chou, H.-P. Chuang, J.-Y. Wu, Y.-T. Chen, M. del Carmen Plaza-Serón, C. Mayorga, R. M. Guéant-Rodríguez, S.-C. Lin, M. J. Torres, P. Campo, C. Rondón, J. J. Laguna, J. Fernández, J.-L. Guéant, G. Canto, M. Blanca, and M. T. M. Lee. Genome-wide association study in NSAID-induced acute urticaria/angioedema in spanish and han chinese populations. *Pharmacogenomics*, 14:1857–1869, 2013.

[5] R. B. Davies. The distribution of a linear combination of χ^2 random variables. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, 29:323–333, 1980.

[6] J. Golding and ALSPAC Study Team. The Avon longitudinal study of parents and children (ALSPAC)—study design and collaborative opportunities. *Eur. J. Endocrinol.*, 151(Suppl.):U119–U123, 2004.

[7] J. Hardy and A. Singleon. Genomewide association studies and human disease. *N. Engl. J. Med.*, 360(2):1759–1768, 2009.

[8] S. Hochreiter and K. Obermayer. Support vector machines for dyadic data. *Neural Comput.*, 18:1472–1510, 2006.

[9] A. Khaledi, M. Schniederjans, S. Pohl, R. Rainer, U. Bodenhofer, B. Xia, F. Klawonn, S. Bruchmann, M. Preusse, D. Eckweiler, A. Dötsch, and S. Häussler. Transcriptome profiling of antimicrobial resistance in *pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.*, 60(8):4722–4733, 2016.

[10] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, 91(2):224–237, 2012.

[11] X. Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

[12] D. Liu, D. Ghosh, and X. Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9:292, 2008.

[13] H. Liu, Y. Tang, and H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Stat. Data Anal.*, 53:853–856, 2009.

[14] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

[15] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, 363(2):166–176, 2010.

[16] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, October 2009.

[17] A. Moayyeri, C. J. Hammond, D. J. Hart, and T. D. Spector. The UK Adult Twin Registry (TwinsUK resource). *Twin Res. Hum. Genet.*, 16(1):144–149, 2013.

[18] A. Romero, P. L. Carrier, A. Erraqabi, T. Sylvain, A. Auvolat, E. Dejoie, M.-A. Legault, M.-P. Dubé, J. G. Hussin, and Y. Bengio. Diet networks: thin parameters for fat genomics. *CoRR*, arXiv/1611.09340, 2017. (published at ICLR 2017).

[19] S. Lee S, G. R. Abecasis, M. Boehnke, and X. Lin. are-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, 95(1):5–23, 2014.

[20] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337:64–69, July 2012.

[21] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, November 2012.

[22] The UK10K Proejct Consortium. The UK10K project identifies rare variants in health and disease. *Nature*, 526:82–90, October 2015.

[23] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89(1):82–93, 2011.