VARIANT PRIORIZATION AND ANALYSIS INCORPORATING PROBLEMATIC REGIONS OF THE GENOME

ANIL PATWARDHAN

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: apatwardhan@personalis.com

MICHAEL CLARK

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: michael.clark@personalis.com

ALEX MORGAN

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: alex.morgan@personalis.com

STEPHEN CHERVITZ

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: schervitz@personalis.com

MARK PRATT

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: mark.pratt@personalis.com

GABOR BARTHA

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: gabor.bartha@personalis.com

GEMMA CHANDRATILLAKE

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: gemma.chandratillake@personalis.com

SARAH GARCIA

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: sarah.garcia@personalis.com

NAN LENG

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: nan.leng@personalis.com

RICHARD CHEN

Personalis Inc., 1350 Willow Road, Suite 202
Menlo Park, CA, 94025, USA
Email: richard.chen@personalis.com

In case-control studies of rare Mendelian disorders and complex diseases, the power to detect variant and gene-level associations of a given effect size is limited by the size of the study sample. Paradoxically, low statistical power may increase the likelihood that a statistically significant finding is also a false positive. The prioritization of variants based on call quality, putative effects on protein function, the predicted degree of deleteriousness, and allele frequency is often used as a mechanism for reducing the occurrence of false positives, while preserving the set of variants most likely to contain true disease associations. We propose that specificity can be further improved by considering errors that are specific to the regions of the genome being sequenced. These problematic regions (PRs) are identified a-priori and are used to down-weight constitutive variants in a case-control analysis. Using samples drawn from 1000-Genomes, we illustrate the utility of PRs in identifying true variant and gene associations using a case-control study on a known Mendelian disease, cystic fibrosis(CF).

1. Introduction

Exome sequencing is a potentially powerful tool in detecting variants and genes responsible for both simple and complex diseases. Recent successes in identifying the causal variants of several Mendelian or monogenic disorders[1-4] have highlighted the utility of heuristic methods of variant filtering and prioritization in the discovery process. These methods often preferentially retain or prioritize variants based on novelty, functional impact, putative effects in the protein coding regions (i.e. missense/nonsense substitutions, coding indels, and splice site-acceptor and donor sites), population frequency, and/or concordance with a subjective assessment of phenotypic features[5]. This biologically informed reduction in the number of variants helps maintain statistical power by reducing the number of formally tested hypotheses and the subsequent impact of multiple testing correction procedures required in high-throughput experiments[6].

While these strategies may enrich the set of disease-associated variants based on variant/functional-level information and disease phenotype, they do not directly address the occurrence of false positives stemming from sequencing inaccuracies. Exome sequencing coverage varies greatly across the genome[7-8] with some regions under-covered due to areas of low-complexity, areas of high GC content, and the occurrence of segmental duplications and homopolymers[9-10]. In case-control studies investigating variant-disease associations, alignment and mapping errors in these problematic regions (PRs) reduces the sensitivity to detect true associations in these regions and may introduce false positive associations in instances where cases and controls have differential coverage depths[11]. The integration of PR information in a case-control analysis may help identify false discoveries not readily identified by other commonly used methods of variant prioritization.

Using a well-characterized set of samples drawn from 1000-Genomes[12] we illustrate the utility of PRs in resolving known causal variants in cystic fibrosis (CF). Combined with other variant prioritization methods, the use of PRs improves the specificity of both standard variant association tests and gene-level collapsing methods in identifying true associations despite limited sample sizes.

2. Methods

2.1. *Subject Samples*

All DNA samples were drawn from the 1000Genomes project. Samples were drawn from a pool of subjects broadly identified as Caucasian and known to be affected (cases) or unaffected (controls) with CF. Information regarding ethnicity, sex, and known mutations in this group of samples, including those samples harboring the ΔF508 common founder mutation, were obtained from the *CFTR* Human Gene Mutation Panel records at the Center for Disease Control[13] and Coriell Institute for Medical Research[14] websites. Cases and controls were sequenced separately using identical platforms and technologies. Raw sequencing data were aligned and variants were called simultaneously for all case and control samples.

2.2. *Genomic Library Construction, Exome Sequencing, Alignment and Variant Calling*

DNA libraries were prepared using Illumina TruSeq Genomic DNA High throughput Sample Prep Kits (Illumina, San Diego, CA) and exome enrichment (targeting 62Mb) was accomplished using the TruSeq Exome Target Enrichment kit (Illumina, San Diego, CA) according to manufacturer's protocols. Sequencing was performed using Illumina Hiseq2000 or HiSeq2500 sequencers with single lane, paired-end 2X100bp reads. DNA fragments were generated and amplified using Clonal Single Molecule Array technology (Illumina, San Diego, CA). The sequences were determined using the Clonal Single Molecule Array and Sequencing-by-Synthesis using Illumina's proprietary instrumentation and Reversible Terminator Chemistry. Sequencing reads of at least 2x100bp in length for a total of at least 8Gb of sequence data per sample were generated for each sequenced sample.

Raw sequence data were in FASTQ format and were analyzed in multisample mode with standard (Sanger) Phred-scale quality scores. The Pipeline then uses an integrated set of proprietary and public analysis tools to align and variant call genomic sequencing data. Gapped alignment is performed using the popular Burrows-Wheeler Aligner (BWA) combined with Picard and the Genome Analysis Toolkit (GATK) to improve sequence alignment and to correct base quality scores. Data was aligned to the hg19 genome, producing standard, compressed Binary Alignment Map (BAM) format files.

GATK's Unified Genotyper module provides the Pipeline's core set of SNV calls and their accompanying quality metrics. Calls are enhanced by proprietary SNV accuracy software which incorporates both genomic context and sequence alignment information into a model that corrects miscalled loci. All calls are made on BAM files that have been recalibrated by GATK's base quality score recalibration (BQSR). SNV and small indels are reported in VCF format. Reference calls and no-call information is returned in BED files.

Variants were annotated using the Personalis Annotation Engine, which applied population frequencies, genetic region information, effect on genes, protein impact, protein-protein interactions and additional structural and functional features to the variants.

2.3. *Problematic Regions of the Genome*

Based on a previous study of discordant variant calls among multiple sequencing platforms[8,15] and further work in elucidating the mechanisms underlying these errors[16], a database of PRs was constructed. PRs are comprised of regions having >3X the average error rate seen among variant calls deemed high-quality by VQSR (i.e. largely PASS calls). PRs included those regions of the genome with high GC content, low coverage, degeneracy due to redundant paralogous sequences, low complexity repetitive elements, segmental duplications, and compression regions[17] for which large amounts of discordance in variant calls were previously observed. It also includes HLA regions and breakpoint library regions for structural variants (BreakSeq[18]). While PR regions are not always mutually exclusive in terms of their categorization, the bulk of PRs (~70%) are due to 100bp regions having >70%

GC content, degenerate 100bp single reads, and simple repeats > 100bp long. Variants called in the case-control analyses were mapped onto the PR database and were flagged as potentially problematic variant calls if they fell into a PR region.

2.4. *Case Control Analysis*

Eighteen unrelated subjects with CF were matched to 54 unrelated and unaffected subjects based on sex and broad ethnic category (i.e. Caucasians) to form a 1:3 case-control study design. In a second case-control analysis, the case-group was redefined to only include the subset of CF-affected individuals without the ΔF508 founder mutation. These 8 case-subjects were again compared to the same 54 unaffected control subjects to form a ~1:7 case-control match. Analysis was performed independently in each of these case-control studies to investigate variant and gene associations with CF.

In each study, variants were removed from analysis if they failed our internal QC requirements. These QC standards required that 1) no more than 10% of the data was missing across case samples and/or control samples and 2) the multi-sample variant call from GATK's Variant Quality Score Recalibration (VQSR) was "PASS"- indicating that there was sufficient evidence that the site was really variant in one or more samples. In order to reduce the likelihood of false discoveries when reporting CF-associated variants and genes, variants were also filtered to retain only those that were protein-coding. These filtering criteria were used when reporting variant-level associations with CF and as input criteria when testing for gene-level associations.

Remaining variants were assessed for association with disease-status using Fisher's Exact Test. Effect size was summarized as the Odds Ratio (OR) calculated from the conditional maximum likelihood estimate of a 2x2 contingency table containing alternative and reference allele counts in cases and controls assuming an additive model. Significance testing of the null of conditional independence (OR=1) used a two-tailed test.

Analysis of the second case-control study, in which all cases with the ΔF508 founder mutation were removed, was done to investigate the occurrence of PRs in studies where smaller effect sizes among causal variants could be expected. This required detection strategies that could accommodate the genetic heterogeneity of the remaining affected individuals- since known causal variants were interspersed throughout the *CFTR* gene[13-14]. Given the challenges in detecting rare variant enrichment with a limited number of heterogeneous case samples, we collapsed the variant-level associations based on gene-membership. An implementation of the Combined Multivariate and Collapsing (CMC) method[19] was used to assess the combined association of variants within the same gene to CF. Variants were binned into groups based on their respective gene membership and further binned (rare vs. common) based on a 1000-Genome derived minor allele frequency (MAF) cutoff of 5%. A multivariate test, Hotelling T-squared, was performed on the counts within all bins to determine differences among the cases and controls with asymptotic p-

values calculated based on the F-distribution. The method of Storey[20] was used to calculate FDR-adjusted p-values (i.e. q-values).

3. Results

The application of filtering criteria related only to QC-criteria (i.e. variant-call quality and missing data) among the 18 cases and matched controls, resulted in 541,119 variants for which association with CF was tested. Distribution of observed -$\log_{10}$(p-value) revealed departure from the expected distribution and severe inflation of type-1 error (Figure 1). Filtering of variants to include only those that were protein-coding reduced the number of variants 10-fold (54,178) and improved data characteristics.

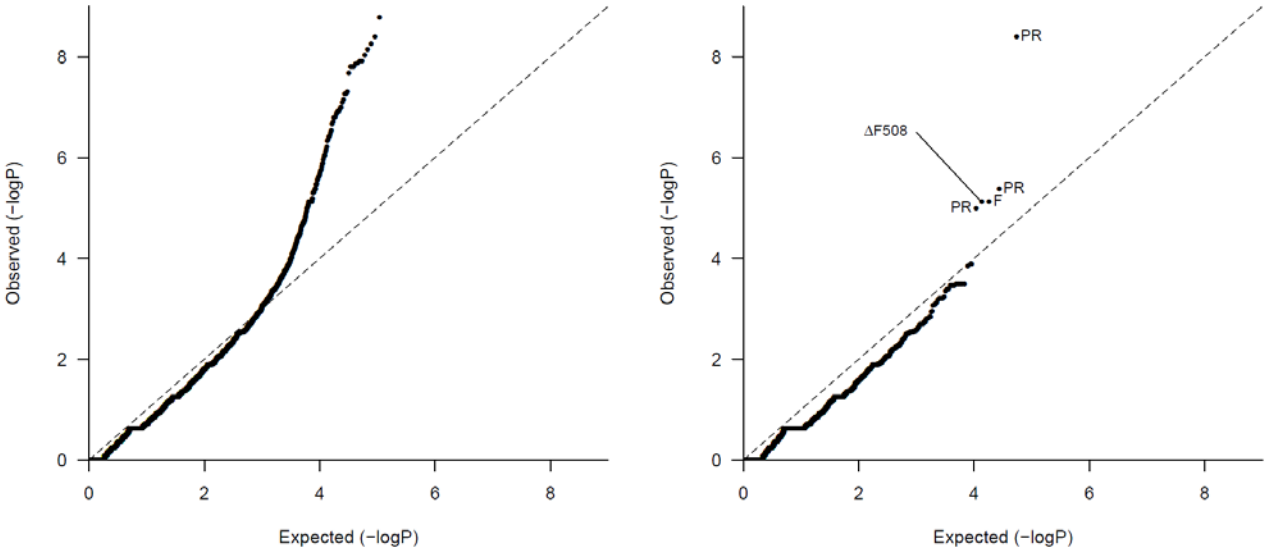

Figure 1. Q-Q Plot comparing the expected normal distribution of –log(p-values) to the observed distribution revealed inflated Type-I error when only data quality filters are applied (left). Filtering variants to include only those that are protein-coding (right) improves the data characteristics and revealed significant ($p<10^{-5}$) true associations (ΔF508), false positives occurring in problematic regions (PR), and false positives that would have been removed based on low allele frequency requirements (F).

Given the number of variants available after QC-criteria and protein-coding filters were applied, an exome-wide significance threshold was set at a p-value of 10^{-5}. At this level, five variants were significantly associated with CF-status, including the known causal variant ΔF508 (rs199826652) that was present in eight affected individuals. Three variants, all indels, occurred in PR regions (Figure 1, "PR"), and one SNP was a missense mutation in the gene DOK3 (Table 1). For variants occurring in PRs of the genome, the underlying presence of simple repeats (*POU4F2, KIAA0664*) or interspersed repeats (*COPB1*) caused sequencing

errors. The SNP, rs3749728, had an allele frequency of 14% according to 1000-Genomes, and would have been identified as a likely false positive based on low frequency assumptions often used for rare, Mendelian disorders (Figure 1, "F").

Table 1. Five variants significantly ($p<10^{-5}$) associated with CF-status after applying QC criteria and and protein-coding filters. Also show are the associated frequencies (MAF) and occurrences in PRs

dbSNP/Gene	Chromosome Position	Ref/Alt Allele	MAF	PR	p-value
POU4F2	Chr4: 147560457	TGGCGGCGGCGGC/ TGGC,TGGCGGCGGCGGCGGC,TGGCGGC,T		Yes	$4.0x10^{-9}$
COPB1	Chr11: 14521144	CGTA/C		Yes	$4.2x10^{-6}$
rs3749728/ *DOK3*	Chr5: 176936819	C/G	14%	No	$7.7x10^{-6}$
rs199826652/ *CFTR*	Chr7: 117199644	ATCT/A	1%	No	$7.7x10^{-6}$
KIAA0664	Chr17: 2595272	GCCCCCGCCACGCCCCCGCCGCGCACCTG/ G,GCCCCGCCGCGCACCTG		Yes	$1.0x10^{-5}$

Aside from the ΔF508 mutation, no other variants in the *CFTR* gene occurred in more than 3 case samples, reflecting the genetic heterogeneity of CF. Since an analysis of only case-samples not harboring the ΔF508 founder mutation would be severely underpowered to detect the smaller effect sizes of the remaining *CFTR* variants, we aggregated variant effects based on gene-membership (i.e. collapsing). Subsequent association testing of 10522 genes with CF-status revealed 15 genes with FDR-controlled p-values (q-values) < .05. Of these, *CFTR* was ranked the 4th gene by p-value. Table 2 summarizes these 15 genes, the nominal p-values derived from the CMC test-statistic, the number of variants contributing to the test statistic, the percentage of those variants found in PRs and the predominant PR type. Collectively, out of the 27 variants occurring in PRs and contributing to these collapsing results, the majority (14) occurred in areas of high-GC content, 10 occurred among segmental duplications, and the remaining occurring among areas of low complexity/simple repeats. Notably one gene association listed in Table 2, *ATF7IP2*, had no constitutive variants in PRs, yet was ranked higher than the known causal gene (i.e. *CFTR*). Further examination of this result revealed good coverage in this area across samples indicating that this was likely reflecting a true difference between cases and controls. However, 3 out of 4 constitutive variants had MAFs > 5%, indicating that these differences are unlikely to be causally related to CF and would be typically excluded using MAF threshold filters.

Table 2. Collapsing results using only CF-affected samples without the ΔF508 mutation revealed 15 genes with q-values<0.05. The nominal p-values, the percentage of those variants in PRs, and the predominant PR type is shown.

Gene	p-value	Number of variants	Percentage of variants in PR	PR types
POU4F2	$1.5x10^{-9}$	2	100%	Repetitive sequence, High GC
MSX1	$4.9x10^{-8}$	5	40%	High GC
ATF7IP2	$2.7x10^{-7}$	4	0%	--
CFTR	**$4.5x10^{-7}$**	**29**	**0%**	--
FUZ	$1.2x10^{-6}$	3	33%	High GC
C8orf74	$1.2x10^{-6}$	5	40%	High GC
TRIM10	$1.2x10^{-6}$	7	0%	--
COL6A1	$3.4x10^{-6}$	6	33%	High GC
PTK2B	$6.1x10^{-6}$	8	0%	--
FAM108A1	$1.2x10^{-5}$	2	100%	Segmental Duplication
MAP7D1	$1.6x10^{-5}$	7	43%	High GC
SCN10A	$1.8x10^{-5}$	13	0%	--
DIDO1	$3.5x10^{-5}$	4	24%	High GC
FLG	$4.0x10^{-5}$	9	89%	Segmental Duplication
COPB1	$8.7x10^{-5}$	2	50%	Repetitive sequence

4. Discussion

In retrospective observational studies of disease association, where disease-affected samples (cases) may be compared to previously sequenced shared controls, alignment and mapping errors can create false evidence for polymorphisms when there are differences in coverage and read depth between groups. Recent evidence has shown that these types of errors can persist when the same genome is sequenced twice under identical analytical environments[16]. Even in carefully designed case-control studies, where samples are matched appropriately and are collected, sequenced, and analyzed together to avoid experimental bias, these errors reduce statistical power for detecting true disease associations.[20] Reduction in these errors are essential for many diseases in which it is a challenge to sufficiently power a case-control study, and is particularly important for complex diseases in which filtering based on frequency thresholds and functional impact may not be appropriate, and where expected effect sizes for a single variant/gene are small or moderate.

CF and the associated study samples used here provide a dataset well-suited to testing the effects of PRs on detection specificity, given that the underlying causal gene and mutations are well-known. Even with a limited number of case samples, we are sufficiently powered to detect variants or genes known to be associated with CF, but suffer from an inflated Type-I error rate. While the effects of these errors can be mitigated through the use of commonly used filtering criteria using a-priori knowledge of the disease (e.g. rare, Mendelian, monogenic), their presence indicates a likely underlying source of bias occurring in the study. No evidence of population stratification was observed when the variance across samples was summarized using principal components- largely discounting biases that might have arisen during the case-control matching process. A potential source of this high error rate may be due to the use of a multi-sample VQSR variant-quality call. In multi-sample mode, a VQSR filter call of "PASS" denotes that the variant call is likely correct in at least one sample- but does not insure it is of sufficient quality across all samples. Variants in which a subset of samples contain low quality calls may introduce false positives associations when those calls occur disproportionality in either the case or control groups. The use of sample-specific (rather than multi-sample) variant-quality calls may help target only those variants of sufficient quality across all samples, providing a higher quality set of variants for association testing in downstream analysis.

Even with the use of filtering criteria, sequencing errors that occur in PRs of the genome cause several false-discoveries to persist. While the variants in Table 1 included those related to errors in covering repeat sequences, examination of PRs in Table 2 revealed that the majority of errors were related to areas of high-GC content and the occurrences of segmental duplications. A comprehensive database integrating these regions provides a mechanism to identify and experimentally or statistically address these potential sources of error.

While the rational use of variant prioritization and/or filtering can enrich the pool of variants likely to be associated with disease, the concomitant reduction in detection sensitivity often increases the Type-II error rate. Filtering variants based on PRs would be particularly problematic in this regard, given that these occur throughout the genome and are not directly related to disease characteristics. Alternative strategies have used probabilistic models incorporating read-specific quality scores and/or sequencing training data in an effort to distinguish true variants from sequencing errors[22-23]. The outcome is typically a decision rule designed to improve false-positive or false-negative error rates in variant detection, or a scoring system in which variants can be differentially weighted in subsequent analysis. While these approaches are certainly improvements over simple filtering of variants, they do not explicitly model all sources of errors inherent in the sequence data itself, including areas of degeneracy, high GC content or areas of low-complexity.

Regardless of the strategy used to distinguish sequencing errors from true discoveries, the errors in the sequence data still exist. The greatest potential impact of a database of PRs is in the identification of areas in the genome that should be targeted for improved coverage- the result being reductions in sequencing error rates[16] regardless of the underlying cause. Improvements in coverage can have beneficial effects on sensitivity; and will improve specificity in large-scale studies where the error rates can differ across samples.

References

1. Ng, S. B. et al. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42, 30–35 (2010).
2. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–793 (2010).
3. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
4. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42,** 483–485 (2010).
5. Ku, C.-S., Naidoo, N. & Pawitan, Y. Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* **129,** 351–370 (2011).
6. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 9546–9551 (2010).
7. Hedges, D. J. *et al.* Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* **6,** e18595 (2011).
8. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).
9. Wang, W., Wei, Z., Lam, T.-W. & Wang, J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep* **1,** 55 (2011).

10. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* **8,** e62856 (2013).
11. Garner, C. Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol.* (2011).
12. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
13. http://wwwn.cdc.gov/clia/Resources/GetRM/pdf/ CF_Characterized_Other.pdf
14. http://ccr.coriell.org/Sections/BrowseCatalog/Diseases.aspx?a=C&coll=&PgId=3
15. Lam, H. Y. K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30,** 78–82 (2012).
16. West, J. *et al.*, Analytical Validity of Genome Sequencing Platforms for Medical Interpretation. Poster presentation at AGBT (2013).
17. Glusman, G. *et al*. Compressions in Human Reference Sequences Identified Using Genome Sequences From Multiple Pedigrees (8 families, 45 complete genomes). Poster presentation at AGBT (2011).
18. Lam, H. Y. et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library
19. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).
20. Storey, J. D. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64: 479–498 (2002).
21. Zhi, D. & Chen, R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. *PLoS ONE* **7,** e31358(2012).
22. Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* **20,** 273–280 (2010).
23. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).