Next-generation sequencing of pharmacogenes: a critical analysis focusing on schizophrenia treatment

Britt I. Drögemöller[a], Galen E.B. Wright[a,c], Dana J.H. Niehaus[b], Robin Emsley[b] and Louise Warnich[a]

Introduction **Because of the unmet needs of current pharmacotherapy for schizophrenia, antipsychotic pharmacogenetic research is of utmost importance. However, to date, few clinically applicable antipsychotic pharmacogenomic alleles have been identified. Nonetheless, next-generation sequencing technologies are expected to aid in the identification of clinically significant variants for this complex phenotype. The aim of this study was therefore to critically examine the ability of next-generation sequencing technologies to reliably detect variation present in pharmacogenes.**

Materials and methods **Candidate antipsychotic pharmacogenes and very important pharmacogenes were identified from the literature and the Pharmacogenomics Knowledgebase. Thereafter, the percentage sequence similarity observed between these genes and their corresponding pseudogenes and paralogues, as well as the percentage low-complexity sequence and GC content of each gene, was calculated. These sequence attributes were subsequently compared with the 'inaccessible' regions of these genes as described by the 1000 Genomes Project.**

Results **It was found that the percentage 'inaccessible genome' correlated well with GC content ($P=9.96\times10^{-5}$), low-complexity sequence ($P=0.0002$) and the presence of pseudogenes/paralogues ($P=8.02\times10^{-7}$). In addition, it was found that many of the pharmacogenes were not ideally suited to next-generation sequencing because of these genomic complexities. These included the *CYP* and *HLA* genes, both of which are of importance to many fields of pharmacogenetics.**

Conclusion **Current short read sequencing technologies are unable to comprehensively capture the variation in all pharmacogenes. Therefore, until high-throughput sequencing technologies advance further, it may be necessary to combine next-generation sequencing with other genotyping strategies.** ***Pharmacogenetics and Genomics*** **23:666–674 © 2013 Wolters Kluwer Health | Lippincott Williams & Wilkins.**

Pharmacogenetics and Genomics 2013, **23**:666–674

Keywords: antipsychotics, genome sequencing, pharmacogenes, pharmacogenomics, schizophrenia

[a]Department of Genetics, Stellenbosch University, Stellenbosch, [b]Department of Psychiatry, Stikland Hospital, Stellenbosch University and [c]South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa

Correspondence to Louise Warnich, PhD, Department of Genetics, Stellenbosch University, Private Bag XI, Matieland 7602, Stellenbosch, South Africa
Tel: + 27 21 8085888; fax: + 27 21 8085833; e-mail: lw@sun.ac.za

Received 11 April 2013 **Accepted** 8 September 2013

Introduction

Schizophrenia is one of the most debilitating mental disorders and current antipsychotic treatments have considerable limitations, with poor response rates [1], high relapse rates [2] and many severe forms of adverse drug reactions (ADRs) [3]. Antipsychotic treatment response varies markedly between individuals [4], highlighting the need for accurate genetic predictors of this phenotype. Furthermore, both schizophrenia and antipsychotic treatment responses have been shown to be highly heritable [5,6] and for this reason it seems likely that genetic variation plays an important role in antipsychotic treatment outcomes.

Although there was much initial anticipation in terms of the application of antipsychotic pharmacogenetics [5], it appears that both schizophrenia and antipsychotic treatment responses are complex phenotypes. Thus, it seems more likely that rare variants and/or several common variants in many genes will interact with one another to influence the range of phenotypes that are observed with respect to both the disorder and the treatment thereof [7]. This genomic complexity, combined with environmental influences, may explain the lack of clinically useful results that have been obtained from antipsychotic pharmacogenetic studies to date. Adding to this void, most of the past studies have focused on examining variants in single candidate genes and have been largely unable to simultaneously examine the variation present in all the genes in the genome (refer to Supplementary Table 1, Supplemental digital content 1, *http://links.lww.com/FPC/A656* for antipsychotic pharmacogenetic review articles). By analysing the entire spectrum of variation present in entire gene networks and pathways, we may be able to obtain a more comprehensive overview of the genetic factors contributing toward antipsychotic treatment phenotypes. This may be achieved through the implementation of next-generation

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (*www.pharmacogeneticsandgenomics.com*).

1744-6872 © 2013 Wolters Kluwer Health | Lippincott Williams & Wilkins DOI: 10.1097/FPC.0000000000000006

sequencing (NGS) technologies, which allow for the high-throughput analyses of all variants in all genes.

Unfortunately, although these sequencing technologies have revolutionized the field of genomics in a remarkably short time, the human genome remains complex, containing regions that are repetitive, GC rich and/or show areas of high sequence similarity. These complexities hinder NGS technologies by decreasing the accessibility of the genome or by interfering with the alignment of sequence reads [8–10]. Currently, the most affordable NGS technologies for whole-genome analyses still utilize short read sequencing. Consequently, areas of high sequence similarity are often affected by misalignments, even though there have been major advances in alignment and variant calling algorithms such as those implemented by the Burrows–Wheeler Alignment Tool [11] and the Genome Analysis Toolkit (GATK) [12]. The inability of these sequencing technologies to capture all the variation in the entire genome, without bias, is reflected by differences in sequencing coverage across the genome.

The 1000 Genomes Project, which has made excellent use of NGS to characterize the human variome across different world population groups, has drawn attention to the fact that differences in sequence coverage across the genome may act as an indication of which areas of the genome are accessible to short read sequencing [13]. Variants that occur outside of these accessible areas may not always be reliably called. In the pilot phase of the 1000 Genomes Project, the 'accessible genome' was calculated by determining which areas contained coverage that differed by a factor of 2 from the median coverage across the genome, as well as which areas had more than 10% of their reads showing mapping quality scores of less than 0 (*http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/*). All areas falling outside of these specifications were considered accessible [13]. After the completion of the sequencing of 1092 genomes from 14 populations, the 1000 Genomes Project also created a more stringent definition of the 'accessible genome'. This stated that the coverage of an area needed to be within 50% of the average coverage across the genome, only 0.1% of the reads could have mapping quality scores of 0 and the average mapping quality needed to be greater than 56 [14]. Although this definition is stringent and only 2% of sites that are called using GATK's variant quality score recalibration are likely to be false positives [14], using this strict mask may serve as a warning for problematic areas and variants called in these areas may need to be examined with caution.

The occurrence of inaccessible areas within the genome may be of particular relevance to the field of pharmacogenomics. Some of the most important pharmacogenes include the *CYP* and *HLA* gene families, both of which have been implicated in antipsychotic pharmacogenetics (refer to Supplementary Table 1, Supplemental digital content 1, *http://links.lww.com/FPC/A656*). These genes are highly polymorphic, with the *HLA* region being the most polymorphic in the human genome [15], and variations within these genes are documented and analysed using special nomenclature systems (*http://hla.alleles.org/*; *http://www.cypalleles.ki.se/*) [16]. For this reason, the variation present in these families of genes requires extensive characterization, highlighting the utility of sequencing technologies as genotyping strategies for these genes. Unfortunately, these genes also show areas of high sequence similarity to other regions of the genome because of the large number of related genes and pseudogenes (*http://hla.alleles.org/*) [17]. This draws attention to the likelihood that the *CYP* and *HLA* genes may not be well suited to NGS and it remains likely that other pharmacogenes may be affected in a similar manner.

The analyses carried out in this study aimed to assess the ability of short read NGS technologies to reliably detect the variation present in pharmacogenes related to the antipsychotic treatment of schizophrenia. Furthermore, we examined pharmacogenes, which have been shown to be most relevant to the field of pharmacogenetics in the broader sense, and compared them with the antipsychotic pharmacogenes. The assessment of all the pharmacogenes was performed by critically examining sequence coverage data in combination with the genomic complexities present in these gene regions.

Materials and methods

Identification of candidate antipsychotic pharmacogenes and very important pharmacogenes

To identify candidate genes that are of interest to schizophrenia-related antipsychotic pharmacogenetics, a literature search was performed in PubMed using the search terms 'antipsychotic pharmacogenetics' and 'antipsychotic pharmacogenomics' (*https://www.ncbi.nlm.nih.gov/pubmed/*). To ensure that the latest and most relevant candidate genes were identified, 'review' was used as an article-type filter and a publication date filter of '5 years' was incorporated into the search. References of identified articles were reviewed for additional relevant citations. Articles that were not available in English and were not related to genetic association studies examining the antipsychotic treatment of schizophrenia were excluded. The remaining articles were subsequently mined to identify genes that are annotated on the reference sequence and have been associated with antipsychotic treatment response or ADR phenotypes of relevance to the treatment of schizophrenia.

In addition to the antipsychotic pharmacogenes identified, all PharmGKB very important pharmacogenes (VIPs) (*http://www.pharmgkb.org/search/browseVip.action?browseKey=vipGenes*; accessed 15 January 2013) were included in downstream analyses. These genes were included to serve as a comparison for the antipsychotic pharmacogenes.

Furthermore, the inclusion of pharmacogenes of high relevance to other fields of pharmacogenetics allowed for the results obtained from this study to have relevance to the field of pharmacogenetics as a whole.

Critical analyses of factors potentially influencing the sequencing of pharmacogenes

Three main factors that could potentially impact the results obtained from the NGS of the identified candidate genes were considered. These were (i) high sequence similarity to paralogues or pseudogenes, (ii) GC content and (iii) repetitive or low-complexity sequences. High sequence similarity may result in misalignment of the sequence reads and GC content may affect the accessibility of gene regions for sequencing applications, whereas low-complexity sequences may affect both these aspects.

Identification of paralogues and pseudogenes

Paralogues showing greater than 70% sequence similarity to the candidate genes of interest were identified with the use of Ensembl BioMart (*http://www.ensembl.org/biomart/martview/6e75fa7cd6995c2c0cbc828188e92206*) using the Ensembl Genes 69 Database and the *Homo sapiens* genes (GTCh37.p8) dataset.

Related pseudogenes, or related functional genes in cases where the candidate genes were pseudogenes, were identified and gene sequences were obtained using NCBI's gene resource (*https://www.ncbi.nlm.nih.gov/gene/*). To determine which of the pseudogenes contained areas with more than 70% sequence similarity to the genes of interest, mVISTA was used (*http://genome.lbl.gov/vista/mvista/mvistacite.shtml*) [18].

Calculation of percentage GC content and low-complexity sequences

The GC content of each gene was calculated using Ensembl BioMart (*http://www.ensembl.org/biomart/martview/6e75fa7cd6995c2c0cbc828188e92206*). To identify the percentage of low-complexity or repetitive sequence present in each gene, the gene co-ordinates were obtained from Ensembl BioMart (*http://www.ensembl.org/biomart/martview/6e75fa7cd6995c2c0cbc828188e92206*). These co-ordinates were then used to determine the percentage of masked sequence present in each gene using the Pre-Masked Genome Search, available from RepeatMasker (*http://www.repeatmasker.org/cgi-bin/AnnotationRequest*).

Assessment of pharmacogenes using the 1000 Genomes Project mask files

Once candidate antipsychotic pharmacogenes and VIPs had been identified and examined critically, the 'accessible genome', as defined by the 1000 Genomes Project coverage data [14], was used as a proxy for the ability of these genes to be sequenced successfully. To determine which areas of the candidate genes did not fall into the 'accessible genome' (referred to as the 'inaccessible genome') as calculated by the 1000 Genomes Project 'strict mask', the bed file containing the unmasked areas was downloaded (*http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/*). Thereafter, the areas falling into the gene regions as defined by Ensembl BioMart (*http://www.ensembl.org/biomart/martview/6e75fa7cd6995c2c0cbc828188e92206*) were assessed to determine how many base pairs fell outside of the 'accessible genome'. The bed files containing the 'accessible' regions for these pharmacogenes are provided in the Supplementary material, Supplemental digital content 2 (*http://links.lww.com/FPC/A657*).

Confirmation of the 1000 Genomes Project data

To critically examine the genes with more than 50% of their gene sequence considered to be inaccessible (as defined by the 1000 Genomes Project strict masking), we utilized exome sequence data from 11 individuals, which were generated and analysed for an antipsychotic pharmacogenomics study, as described by Drögemöller *et al*. [19] (details provided in the Supplementary material, Supplemental digital content 2, *http://links.lww.com/FPC/A657*). The sequence coverage of these genes was subsequently examined using NGSrich (*http://sourceforge.net/projects/ngsrich/filesunderthe GeneralPublic*) [8], as described by the authors, using the SureSelect50Mb.bed file (*https://earray.chem.agilent.com/earray/*). If the average coverage of the gene region was less than or greater than 50% of the mean read depth for that sample, the coverage was considered to be poor for that gene.

Statistical analyses

Differences between the VIPs and antipsychotic pharmacogenes with respect to percentage inaccessibility, GC content and low-complexity sequences were assessed using the Wilcoxon–Mann–Whitney test. A Pearson's χ^2-test of independence was used to assess the differences between these two groups with respect to the number of genes with more than 70% sequence similarity to pseudogenes/paralogues. The relationship between percentage inaccessibility and the dichotomous presence of a pseudogene or paralogue (with >70% sequence similarity) was also assessed using the Wilcoxon–Mann–Whitney test. In addition, associations between percentage inaccessibility, GC content and low-complexity sequences were examined using Spearman's rank-order correlation. For group comparisons, genes belonging to both the VIP and the antipsychotic pharmacogene lists were excluded from analysis. *P* values less than 0.05 were considered significant. All statistical analyses were carried out in R (*http://www.r-project.org*).

Results

Using the search criteria for the antipsychotic pharmacogenetics literature as described above, 38 articles

remained that referred to associations that were found with variants in 152 genes (refer to Supplementary Table 1, Supplemental digital content 1, *http://links.lww.com/FPC/A656*). The antipsychotic pharmacogenetic traits that these genes were associated with were treatment response, weight gain, movement disorders, agranulocytosis, QT prolongation, hyperprolactinaemia and neuroleptic malignant syndrome. When the search was broadened to include important pharmacogenes from other fields of pharmacogenetics, it was found that PharmGKB listed 47 VIPs, of which 12 were also included in the list of antipsychotic pharmacogenes (Table 1).

Statistical analyses showed significant positive correlations between percentage inaccessibility and GC content ($\rho = 0.281$, $P = 9.96 \times 10^{-5}$), as well as percentage low-complexity sequence ($\rho = 0.269$, $P = 0.0002$). The mean percentage inaccessibility was higher for genes with paralogues/pseudogenes (i.e. 43.47%) compared with those without these homologous regions (i.e. 21.30%), and this difference was statistically significant ($P = 8.02 \times 10^{-7}$).

The results obtained from the analyses of the genomic composition (GC content, sequence similarity, low-complexity sequences and 'inaccessible genome') of the candidate antipsychotic pharmacogenes and VIPs are shown in Supplementary Table 1 (Supplemental digital content 1, *http://links.lww.com/FPC/A656*). When examining the pharmacogenes that had more than 50% of their gene regions falling into the 'inaccessible genome' (Table 2), 23 genes were identified. Utilization of the exome sequence data analysed in our laboratory to examine the sequence coverage for these 23 inaccessible genes showed that all but seven of these genes also showed poor sequence coverage for our data (Table 2). Examination of the genomic composition of the 23 genes indicated that 20 of the genes (86.96%) showed more than 70% sequence similarity to paralogues or pseudogenes. On examining the remaining three genes, it was observed that two of these three genes were affected by low-complexity sequence or high GC content. In the case of the first gene, *TGFB1*, approximately half of the sequence present in this gene was repetitive (i.e. constitutes low-complexity sequence). In the case of the second gene, *DRD4*, Fig. 1 shows how all except two of the inaccessible areas in *DRD4* are either repetitive or GC rich.

Analysis of differences between the VIPs and antipsychotic pharmacogenes with respect to percentage GC content was not statistically significant ($P = 0.915$); however, the differences between the two groups with respect to percentage inaccessibility were statistically significant ($P = 0.035$). Finally, it was shown that the VIPs were more likely to be affected by areas of high sequence similarity ($P = 0.029$), with 42.86% of the VIPs showing more than 70% sequence similarity to paralogues or pseudogenes (compared with 24.29% of the antipsychotic pharmacogenes).

Discussion

This article has identified a significant number of studies that have reported associations with various antipsychotic pharmacogenetic traits. The majority of the genes examined in these studies were associated with treatment response (52.26%), followed by weight gain (29.68%), movement disorders (23.23%) and agranulocytosis (9.03%) (refer to Supplementary Table 1, Supplemental digital content 1, *http://links.lww.com/FPC/A656*). The most likely reasons for this are as follows: (i) treatment response, although difficult to measure, is arguably the most significant hurdle in the treatment of schizophrenia and only approximately half of patients respond to treatment [1]; (ii) weight gain and related metabolic disorders are the most prominent ADRs with atypical antipsychotics

Table 1 The genomic composition of antipsychotic pharmacogenes that are also considered to be very important pharmacogenes

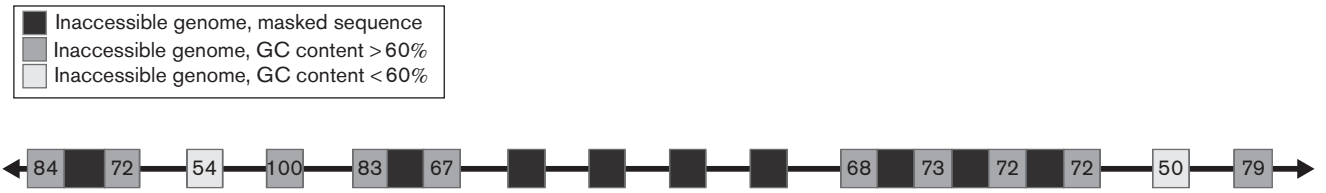
Gene	Associated antipsychotic pharmacogenetic trait	References	Paralogue/pseudogene	GC content (%)	Low-complexity sequence (%)	Inaccessible Genome (%)
ABCB1	Treatment response, weight gain	[5,20–28]	>70% sequence similarity	37.05	51.55	19.39
ADRB2	Weight gain	[29]	–	50.61	0.00	13.97
COMT	Treatment response, movement disorders	[3,5,20,21,24,26,27,30–41]	–	53.41	50.18	39.19
CYP1A2	Treatment response, movement disorders	[1,3,5,20,21,24,26,27,30,38,42,43]	>70% sequence similarity	52.03	32.79	24.48
CYP2C19	Treatment response	[24,44]	>70% sequence similarity	38.88	80.16	62.63
CYP2D6	Treatment response, movement disorders, weight gain, QT prolongation	[1,3,5,20,21–31,34,38,39,42,44–48]	>70% sequence similarity	62.68	0.00	100.00
CYP3A4	Treatment response	[20,27]	>70% sequence similarity	39.62	39.33	46.57
CYP3A5	Treatment response	[27]	>70% sequence similarity	40.47	48.48	31.64
DRD2	Treatment response, movement disorders, weight gain, hyperprolactinaemia, neuroleptic malignant syndrome	[1,3,5,20,21–39,42,44,46,47,49,50]	–	48.36	33.70	9.33
GSTP1	Movement disorders	[21,30]	>70% sequence similarity	63.05	2.84	25.74
MTHFR	Treatment response, weight gain	[20–24,26,36]	–	54.53	27.37	22.32
NQO1	Movement disorders, agranulocytosis	[5,20,21,30,38,51]	–	47.09	54.04	47.77

Table 2 The genomic composition of those genes with >50% 'inaccessible genome'

Gene	Associated antipsychotic pharmacogenetic trait	References	Paralogue/pseudogene	GC content (%)	Low-complexity sequence (%)	Inaccessible genome (%)	NGSrich reported gene sequence coverage
Antipsychotic pharmacogenes							
DRD4	Treatment response, movement disorders, weight gain	[3,5,20,21–26,30,32,33,36,38,46,47]	–	67.02	21.56	65.10	<50% of mean
GSTM1	Movement disorders	[5,21,30,38,42]	>70% sequence similarity	46.35	55.35	96.88	>50% and <50% of mean
HLA-B	Agranulocytosis	[31,51]	>70% sequence similarity	59.00	0.00	97.01	
HLA-C	Agranulocytosis	[31,51]	>70% sequence similarity	59.28	0.00	99.26	
HLA-DQA1	Agranulocytosis	[31,51]	>70% sequence similarity	39.77	44.87	77.50	
HLA-DQB1	Agranulocytosis	[1,20,21,31,39,51]	>70% sequence similarity	47.03	13.30	95.00	<50% of mean
HLA-DQB3[a]	Agranulocytosis	[51]	>70% sequence similarity	48.72	0.00	72.15	Not captured by SureSelect 50 Mb capture kit
HLA-DRB1	Agranulocytosis	[31,51]	>70% sequence similarity	40.74	33.79	99.84	>50% and <50% of mean
HLA-DRB5	Agranulocytosis	[31,51]	>70% sequence similarity	42.84	36.39	100.00	>50% and <50% of mean
HSPA1A	Agranulocytosis	[31,51]	>70% sequence similarity	59.76	0.00	97.37	>50% of mean
OXT	Treatment response	[20]	>70% sequence similarity	71.94	0.00	88.18	<50% of mean
TGFB1	Weight gain	[23,42]	–	52.41	50.01	61.39	<50% of mean
Antipsychotic pharmacogene falling into the very important pharmacogene category							
CYP2C19	Treatment response	[24,44]	>70% sequence similarity	38.88	80.16	62.63	>50% of mean
CYP2D6	Treatment response, movement disorders, weight gain, QT prolongation	[1,3,5,20,21–31,34,38,39,42,44–48]	>70% sequence similarity	62.68	0.00	100.00	–
Other very important pharmacogenes							
ADRB1	NA	NA	–	57.51	6.99	51.80	<50% of mean
BRCA1	NA	NA	>70% sequence similarity	44.09	57.07	58.64	–
CYP2A6	NA	NA	>70% sequence similarity	53.42	20.61	98.57	–
CYP2B6	NA	NA	>70% sequence similarity	44.63	60.12	73.13	>50% of mean
CYP2C9	NA	NA	>70% sequence similarity	37.78	71.96	56.82	>50% of mean
GSTT1	NA	NA	>70% sequence similarity	52.22	44.13	65.34	<50% of mean
SULT1A1	NA	NA	>70% sequence similarity	52.72	44.23	89.67	>50% of mean
TYMS	NA	NA	>70% sequence similarity	45.77	49.32	52.92	>50% of mean
VKORC1	NA	NA	>70% sequence similarity	55.01	41.11	53.17	

[a]This gene is a pseudogene.

Fig. 1



DRD4 and the corresponding inaccessible regions of the gene. The regions are not drawn to scale and are merely a representation of how genomic complexities can influence genome sequencing coverage. The blocks represent areas of the gene that are inaccessible and are: (i) masked by RepeatMasker; (ii) contain >60% GC content; and (iii) contain <60% GC content and are not masked by RepeatMasker. The numbers in the blocks indicate the GC content of these areas.

[52]; (iii) movement disorders are the most frequent ADRs observed with typical antipsychotics [52]; and (iv) agranulocytosis is a severe, and in some cases lethal, ADR associated with antipsychotic treatment [53]. These traits were associated with several different genes and, interestingly, of the 47 VIPs, 12 (25.53%) were also antipsychotic pharmacogenes. This highlights the importance of antipsychotic pharmacogenetics in the context of pharmacogenetics as a whole. Thus, NGS projects examining antipsychotic pharmacogenomics may be valuable and therefore it is necessary to critically examine the likelihood that antipsychotic pharmacogenes will be sequenced successfully.

Although the 1000 Genomes Project masking may be an overly stringent assessment of the quality of sequencing data, the findings from this study indicate that the areas of the genome that were considered to be inaccessible correlated well with all three genomic complexities that were examined (percentage GC content: $P = 9.96 \times 10^{-5}$, percentage low-complexity sequence: $P = 0.0002$, presence of paralogues/pseudogenes with >70% sequence similarity: $P = 8.02 \times 10^{-7}$). This aligns with the description of the 1000 Genomes Project masked regions, which are described as areas where reads are 'ambiguously placed' or where there are 'unexpectedly high or low numbers of aligned reads' [13]. These discrepancies in read alignment may stem from the initial capture and amplification processes during the library preparation before sequencing. However, it remains likely that a portion of these reads are successfully captured, amplified and sequenced, but alignment of these sequence reads is error prone [54]. Therefore, the library preparation and alignment algorithms associated with NGS technologies may require improvement and it is for this reason that sequencing companies are addressing these issues [10].

With reference to the specific genes that were considered inaccessible, *DRD4* may be especially important to consider for antipsychotic pharmacogenomic studies as all current antipsychotics bind to dopamine receptors [55]. This gene is affected by both low-complexity sequence and high GC content (Fig. 1), and these attributes are likely to affect the success with which this gene can be sequenced using NGS technologies. Along with this gene, a further 22 pharmacogenes contained more than 50% inaccessible sequence, with 20 of these genes showing greater than 70% sequence similarity to paralogues or pseudogenes and one gene containing approximately 50% low-complexity sequences. In total, 22/23 (95.65%) of the genes that were considered largely inaccessible (> 50% inaccessible) also contained a significant percentage of sequence complexities (> 50% of their gene regions were affected by sequence similarity, high GC content or low-complexity sequences).

Further highlighting the utility of the 1000 Genomes Project mask files was the correlation that was observed between these data and the sequence coverage data obtained for additional exomes analysed for the current study. Interestingly, when examining the seven genes that did not fulfil the requirements for poor sequence coverage in our data, the mapping quality of the reads aligning to these genes was generally poor, specifically in the case of *CYP2D6* (refer to Supplementary Fig. 1, Supplemental digital content 1, *http://links.lww.com/FPC/A656*). Taken together, these sequence coverage and genomic composition data suggest that the 1000 Genomes Project mask files may serve as a useful guide when critically analysing NGS results.

On comparing the antipsychotic pharmacogenes and VIPs, it was observed that these two groups were similar with respect to percentage GC content and low-complexity sequence. Although there was a significant difference in the percentage inaccessible sequence ($P = 0.035$), this was likely to be driven by the significant differences that were observed with respect to the number of genes with more than 70% sequence similarity to pseudogenes/paralogues ($P = 0.029$). As antipsychotic pharmacogenes are often required for vital processes such as dopamine regulation, they are more likely to be under evolutionary constraint. In contrast, the largest gene family present in the VIPs, namely the *CYP* gene family (10 of the 47 VIPs are *CYP* genes), has evolved rapidly. Furthermore,

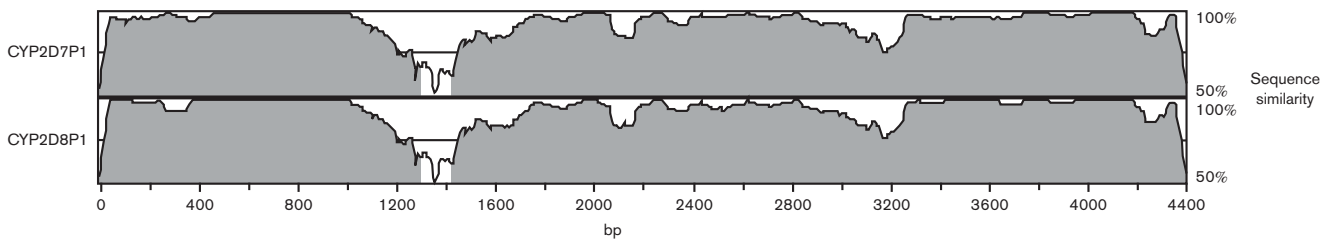
it has been hypothesized that as a result of the current nonessential nature of these genes, they have accumulated many variants [17]. The resulting large polymorphic genes families make short read sequencing of these areas very challenging and highlight the potential shortcomings with respect to NGS in the context of pharmacogenomic studies [56]. The hurdles associated with the NGS of polymorphic gene families such as the *CYP* and *HLA* genes are reflected in the large areas within these genes that are deemed inaccessible by the 1000 Genomes Project strict masking annotation. Both of these families contain genes that are 100% masked. These genes, namely the *CYP2D6* and *HLA-DRB5* (Table 2), show large regions of sequence similarity to other genes or pseudogenes within their respective families. This is clearly shown in Fig. 2, where the high sequence similarity observed between *CYP2D6* and the two corresponding pseudogenes is shown.

These findings reiterate that the presence of sequence complexities may hinder the unbiased nature of genome sequencing and it seems likely that certain areas within the genome may remain impervious to NGS variant detection, as was well demonstrated in a recent study examining kidney disease [9]. It is important to note that these sequence complexities not only complicate the identification of single nucleotide variants but may also further hamper the detection of copy number variants (CNVs). This is of particular relevance to pharmacogenes such as *CYP2D6*, where gene deletions, duplications and hybrid genes have been reported (*http://www.cypalleles.ki.se/cyp2d6.htm*). Although it is possible to examine CNVs with algorithms designed to examine NGS read depth, the genomic complexities examined in this study are likely to confound these results [57]. Thus, the examination of CNVs present in pharmacogenes that are affected by these sequence complexities is likely to yield inaccurate results.

It is well known that although NGS technologies far surpass Sanger sequencing with respect to time and cost [58], the accuracy of Sanger sequencing and CNV assays for variant detection remains unparalleled. From the results obtained in this study, it appears that current short read sequencing methods are not sufficiently reliable for variant detection for many pharmacogenes. This highlights the fact that the design of NGS pharmacogenetic assays, such as the PGRN-Seq (*http://pgrn.org/display/pgrnwebsite/Network-wide+Projects*), needs to be performed with caution. This assay warns that genes such as *CYP2D6*, *HLA-B* and *HLA-DQB3* are unlikely to be accurately sequenced because of high sequence similarities, which correlates well with our data. All three of these genes are antipsychotic pharmacogenes and *CYP2D6* is considered among the top 10 VIPs [59], possibly limiting the applicability of this assay for antipsychotic pharmacogenetic applications. The difficulties with genotyping both *CYP2D6* and the *HLA* genes have been documented previously and carefully designed genotyping strategies are required, including CNV assays [15,16,60,61]. With particular reference to using NGS for *HLA* genotyping, it appears that the use of RNAseq, which allows for greater representation of *HLA* alleles that are highly expressed, in combination with longer read lengths, to prevent misalignment, may be a better strategy [15,16]. Even so, variant detection in these genes will remain a challenge.

Although the genomic complexities associated with many of the pharmacogenes do provide unique challenges, the advancement of sequencing technologies offers the potential for the discovery of variants associated with antipsychotic response phenotypes. NGS approaches can be used to simultaneously examine known pathways as well as to discover novel targets, while allowing for the detection of both common and rare variants. However, for NGS to be truly unbiased, it is essential that all genes are represented, including well-known pharmacogenes such as the *CYP* and *HLA* families. The results from this study serve as a reference for which pharmacogenes may require careful analyses with respect to NGS data. It should, however, be noted that these results are limited as they only refer to candidate pharmacogenes and examine these genes as a whole. Future studies may need to examine additional genes and it may be necessary to focus on specific regions within these genes.

Fig. 2



Percentage sequence similarity between *CYP2D6* and the corresponding pseudogenes as calculated by mVISTA.

Conclusion

Future advances are likely to incorporate longer read lengths into sequencing technologies such as Nanopore Technology (*http://www.nanoporetech.com/*), thereby enabling comprehensive characterization of variation in all genes. Studies including well-characterized cohorts in combination with these comprehensive genotyping strategies may then obtain results that are applicable to clinical practice. In the interim, the combination of short read NGS analyses with already existing strategies, such as the use of long-range PCR or longer read NGS, may be the best strategy to examine antipsychotic pharmacogenetics.

Acknowledgements

The work reported here was supported by the following grants to the authors: Britt I. Drögemöller: National Research Foundation (NRF) research bursary and the L'Oréal-UNESCO for women in Science in Sub-Saharan Africa Fellowship; Galen E.B. Wright: National Research Foundation (NRF) research bursary; Dana J.H. Niehaus: South African Medical Council (MRC) operating research grant; Robin Emsley: New Partnership for Africa's Development (NEPAD) grant, through the Department of Science and Technology of South Africa; Louise Warnich: NRF operating research grant. This financial assistance is hereby acknowledged. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to these funding sources.

Conflicts of interest

Robin Emsley has participated in speakers/advisory boards and received honoraria from AstraZeneca, Bristol-Myers Squibb, Janssen, Lilly, Lundbeck, Organon, Pfizer, Servier, Otsuka and Wyeth. He has received research funding from Janssen, Lundbeck and AstraZeneca. For the remaining authors there are no conflicts of interest.

References

1 Lohoff FW, Ferraro TN. Pharmacogenetic considerations in the treatment of psychiatric disorders. *Expert Opin Pharmacother* 2010; **11**:423–439.
2 Robinson D, Woerner MG, Alvir JM, Bilder R, Goldman R, Geisler S, *et al.* Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. *Arch Gen Psychiatry* 1999; **56**:241–247.
3 Zhang J-P, Malhotra AK. Pharmacogenetics and antipsychotics: therapeutic efficacy and side effects prediction. *Expert Opin Drug Metab Toxicol* 2011; **7**:9–37.
4 Zhang J-P, Malhotra AK. Pharmacogenetics of antipsychotics: recent progress and methodological issues. *Expert Opin Drug Metab Toxicol* 2013; **9**:183–191.
5 Arranz MJ, de Leon J. Pharmacogenetics and pharmacogenomics of schizophrenia: a review of last decade of research. *Mol Psychiatry* 2007; **12**:707–747.
6 Van Os J, Kapur S. Schizophrenia. *Lancet* 2009; **374**:635–645.
7 Hayden E. Human genome at ten: life is complicated. *Nature* 2010; **464**:664–667.
8 Frommolt P, Abdallah AT, Altmüller J, Motameny S, Thiele H, Becker C, *et al.* Assessing the enrichment performance in targeted resequencing experiments. *Hum Mutat* 2012; **33**:635–641.
9 Kirby A, Gnirke A, Jaffe DB, Barešová V, Pochet N, Blumenstiel B, *et al.* Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat Genet* 2013; **45**:299–303.
10 Altman RB, Whirl-Carrillo M, Klein TE. Challenges in the pharmacogenomic annotation of whole genomes. *Clin Pharmacol Ther* 2013; **94**:211–213.
11 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**:1754–1760.
12 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43**:491–498.
13 Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**:1061–1073.
14 Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* An integrated map of genetic variation from 1092 human genomes. *Nature* 2012; **491**:56–65.
15 Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, *et al.* Derivation of HLA types from shotgun sequence datasets. *Genome Med* 2012; **4**:95.
16 Babik W, Taberlet P, Ejsmond MJ, Radwan J. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Mol Ecol Resour* 2009; **9**:713–719.
17 Ingelman-Sundberg M. The human genome project and novel aspects of cytochrome P450 research. *Toxicol Appl Pharmacol* 2005; **207**:52–56.
18 Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 2004; **32**:W273–W279.
19 Drögemöller B, Niehaus D, Chiliza B, Asmal L, Malhotra AK, Wright GEB, *et al.* Patterns of variation influencing antipsychotic treatment outcomes in South African first episode schizophrenia patients. *Pharmacogenomics* Under review.
20 Arranz MJ, Rivera M, Munro JC. Pharmacogenetics of response to antipsychotics in patients with schizophrenia. *CNS Drugs* 2011; **25**: 933–969.
21 Arranz MJ, Munro JC. Toward understanding genetic risk for differential antipsychotic response in individuals with schizophrenia. *Expert Rev Clin Pharmacol* 2011; **4**:389–405.
22 Risselada AJ, Mulder H, Heerdink ER, Egberts TCG. Pharmacogenetic testing to predict antipsychotic-induced weight gain: a systematic review. *Pharmacogenomics* 2011; **12**:1213–1227.
23 Lett TAP, Wallace TJM, Chowdhury NI, Tiwari AK, Kennedy JL, Müller DJ. Pharmacogenetics of antipsychotic-induced weight gain: review and clinical implications. *Mol Psychiatry* 2012; **17**:242–266.
24 Gerretsen P, Müller DJ, Tiwari A, Mamo D, Pollock BG. The intersection of pharmacology, imaging, and genetics in the development of personalized medicine. *Dialogues Clin Neurosci* 2009; **11**:363–376.
25 Balt SL, Galloway GP, Baggott MJ, Schwartz Z, Mendelson J. Mechanisms and genetics of antipsychotic-associated weight gain. *Clin Pharmacol Ther* 2011; **90**:179–183.
26 Foster A, Miller DD, Buckley P. Pharmacogenetics and schizophrenia. *Clin Lab Med* 2010; **30**:975–993.
27 Gesteira A, Barros F, Martin A, Pérez V, Cortés A, Baiget M, *et al.* Pharmacogenetic studies on the antipsychotic treatment. Current status and perspectives. *Actas Esp Psiquiatr* 2010; **38**:301–316.
28 Moons T, de Roo M, Claes S, Dom G. Relationship between P-glycoprotein and second-generation antipsychotics. *Pharmacogenomics* 2011; **12**:1193–1211.
29 Rege S. Antipsychotic induced weight gain in schizophrenia: mechanisms and management. *Aust N Z J Psychiatry* 2008; **42**:369–381.
30 Lee H-J, Kang S-G. Genetics of tardive dyskinesia. *Int Rev Neurobiol* 2011; **98**:231–264.
31 Chowdhury NI, Remington G, Kennedy JL. Genetics of antipsychotic-induced side effects and agranulocytosis. *Curr Psychiatry Rep* 2011; **13**:156–165.
32 Cacabelos R, Hashimoto R, Takeda M. Pharmacogenomics of antipsychotics efficacy for schizophrenia. *Psychiatry Clin Neurosci* 2011; **65**:3–19.
33 Di Giorgio A, Sambataro F, Bertolino A. Functional imaging as a tool to investigate the relationship between genetic variation and response to treatment with antipsychotics. *Curr Pharm Des* 2009; **15**:2560–2572.
34 Lencz T, Malhotra AK. Pharmacogenetics of antipsychotic-induced side effects. *Dialogues Clin Neurosci* 2009; **11**:405–415.
35 Reynolds GP. The pharmacogenetics of symptom response to antipsychotic drugs. *Psychiatry Investig* 2012; **9**:1–7.
36 Blanc O, Brousse G, Meary A, Leboyer M, Llorca P-M. Pharmacogenetic of response efficacy to antipsychotics in schizophrenia: pharmacodynamic

aspects. Review and implications for clinical research. *Fundam Clin Pharmacol* 2010; **24**:139–160.
37 Burdick KE, Gopin CB, Malhotra AK. Pharmacogenetic approaches to cognitive enhancement in schizophrenia. *Harv Rev Psychiatry* 2011; **19**:102–108.
38 Thelma B, Srivastava V, Tiwari AK. Genetic underpinnings of tardive dyskinesia: passing the baton to pharmacogenetics. *Pharmacogenomics* 2008; **9**:1285–1306.
39 Reynolds GP. The impact of pharmacogenetics on the development and use of antipsychotic drugs. *Drug Discov Today* 2007; **12**:953–959.
40 Vyas NS, Shamsi SA, Malhotra AK, Aitchison KJ, Kumari V. Can genetics inform the management of cognitive deficits in schizophrenia? *J Psychopharmacol* 2012; **26**:334–348.
41 Sagud M, Mück-Seler D, Mihaljević-Peles A, Vuksan-Cusa B, Zivković M, Jakovljević M, *et al*. Catechol-*O*-methyl transferase and schizophrenia. *Psychiatr Danub* 2010; **22**:270–274.
42 Plesnicar BK. Personalized antipsychotic treatment: the adverse effects perspectives. *Psychiatr Danub* 2010; **22**:329–334.
43 Tandon R, Nasrallah HA, Keshavan MS. Schizophrenia, 'just the facts' 5. Treatment and prevention. Past, present, and future. *Schizophr Res* 2010; **122**:1–23.
44 Buckley PF. Factors that influence treatment success in schizophrenia. *J Clin Psychiatry* 2008; **69 (Suppl 3)**:4–10.
45 Bishop JR, Bishop DL. Iloperidone for the treatment of schizophrenia. *Drugs Today (Barc)* 2010; **46**:567–579.
46 Maier W, Zobel A. Contribution of allelic variations to the phenotype of response to antidepressants and antipsychotics. *Eur Arch Psychiatry Clin Neurosci* 2008; **258 (Suppl 1)**:12–20.
47 Steimer W. Pharmacogenetics and psychoactive drug therapy: ready for the patient? *Ther Drug Monit* 2010; **32**:381–386.
48 Dorado P, Peñas-Lledó EM, Llerena A. CYP2D6 polymorphism: implications for antipsychotic drug response, schizophrenia and personality traits. *Pharmacogenomics* 2007; **8**:1597–1608.
49 Shiroma PR, Geda YE, Mrazek DA. Pharmacogenomic implications of variants of monoaminergic-related genes in geriatric psychiatry. *Pharmacogenomics* 2010; **11**:1305–1330.
50 Malhotra AK, Lencz T, Correll CU, Kane JM. Genomics and the future of pharmacotherapy in psychiatry. *Int Rev Psychiatry* 2007; **19**:523–530.
51 Opgen-Rhein C, Dettling M. Clozapine-induced agranulocytosis and its genetic determinants. *Pharmacogenomics* 2008; **9**:1101–1111.
52 Jafari S, Fernandez-Enright F, Huang X-F. Structural contributions of antipsychotic drugs to their therapeutic profiles and metabolic side effects. *J Neurochem* 2012; **120**:371–384.
53 Røge R, Møller BK, Andersen CR, Correll CU, Nielsen J. Immunomodulatory effects of clozapine and their clinical implications: what have we learned so far? *Schizophr Res* 2012; **140**:204–213.
54 Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, *et al*. Exome sequencing and the genetic basis of complex traits. *Nat Genet* 2012; **44**:623–630.
55 Correll CU. From receptor pharmacology to improved outcomes: individualising the selection, dosing, and switching of antipsychotics. *Eur Psychiatry* 2010; **25 (Suppl 2)**:S12–S21.
56 Drögemöller BI, Wright GEB, Niehaus DJH, Emsley RA, Warnich L. Whole-genome resequencing in pharmacogenomics: moving away from past disparities to globally representative applications. *Pharmacogenomics* 2011; **12**:1717–1728.
57 Duan J, Zhang J-G, Deng H-W, Wang Y-P. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One* 2013; **8**:e59128.
58 Lander ES. QnAs with Eric S. Lander. Interview by Prashant Nair. *Proc Natl Acad Sci USA* 2011; **108**:11319.
59 Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics* 2010; **11**:501–505.
60 Gaedigk A, Ndjountché L, Divakaran K, Dianne Bradford L, Zineh I, Oberlander TF, *et al*. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events. *Clin Pharmacol Ther* 2007; **81**:242–251.
61 Wright GEB, Niehaus DJH, Drögemöller BI, Koen L, Gaedigk A, Warnich L. Elucidation of CYP2D6 genetic diversity in a unique African population: implications for the future application of pharmacogenetics in the Xhosa population. *Ann Hum Genet* 2010; **74**:340–350.