RESEARCH NOTE

Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes [v1; ref status: indexed, http://f1000r.es/4zj]

Ron Ammar¹, Tara A. Paton², Dax Torti¹, Adam Shlien³, Gary D. Bader¹,⁴,⁵

¹The Donnelly Centre, University of Toronto, Toronto, ON, M5S3E1, Canada
²The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, M5G0A4, Canada
³Department of Laboratory Medicine and Pathobiology, University of Toronto; Program in Genetics and Genome Biology & Department of Paediatric Laboratory Medicine The Hospital for Sick Children, Toronto, ON, M5G1X8, Canada
⁴Department of Computer Science, University of Toronto, Toronto, ON, M5S3G4, Canada
⁵Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S1A8, Canada

Abstract

Haplotypes are often critical for the interpretation of genetic laboratory observations into medically actionable findings. Current massively parallel DNA sequencing technologies produce short sequence reads that are often unable to resolve haplotype information. Phasing short read data typically requires supplemental statistical phasing based on known haplotype structure in the population or parental genotypic data. Here we demonstrate that the MinION nanopore sequencer is capable of producing very long reads to resolve both variants and haplotypes of HLA-A, HLA-B and CYP2D6 genes important in determining patient drug response in sample NA12878 of CEPH/UTAH pedigree 1463, without the need for statistical phasing. Long read data from a single 24-hour nanopore sequencing run was used to reconstruct haplotypes, which were confirmed by HapMap data and statistically phased Complete Genomics and Sequenom genotypes. Our results demonstrate that nanopore sequencing is an emerging standalone technology with potential utility in a clinical environment to aid in medical decision-making.
Introduction
An important aspect of precision medicine is the study of how genes influence individual response to drug therapies, known as pharmacogenomics (PGx). PGx genotyping impacts the choice of drug dosing in many medical contexts. As an example, in acute lymphoblastic leukemia patients the metabolizer status of thiopurine methyltransferase (TPMT) must be considered when calculating the initial drug dose of mercaptopurine (6-MP) to ensure proper treatment and avoid fatal toxicity\(^1\).\(^2\).

PGx data are typically collected by sequencing a small panel of known PGx genes via traditional Sanger sequencing, or targeted genotyping technologies\(^3\). Diagnostic labs are also exploring the use of whole genome or exome sequencing (WGS, WES) for PGx. However, existing methods have various limitations, which may lead to adverse drug responses. WGS and WES methods may fail to capture or provide adequate sequence coverage for certain PGx loci. Targeted genotyping approaches, such as Taqman (Life Technologies), Luminex (Luminex Corp.) or Sequenom (Agena Bioscience), can fail to detect novel loss-of-function mutations due to their selective interrogation of predefined genomic loci. In a diagnostic clinic, where results are often required within days of administering diagnostic tests, some existing technologies can delay the return of clinical results. Current massively parallel technologies have high capital costs (ranging from $100,000–$1,000,000) requiring high maintenance fees and high electricity costs. These factors hinder their adoption by smaller clinical laboratories to purchase and maintain large instruments to perform in-house genotyping. Alternatively, a laboratory can send these PGx samples to a third party service for a fee, but may wait up to several months for a clinical report.

Clinical haplotypes, tightly-linked collections of inherited alleles, that are responsible for a plethora of medical phenotypes including patient drug response are important in many medical sequencing applications. Information about haplotypes, or genotype phase, can be inferred from parental genotypes or genetic pattern frequency in the human population, however, these predictions can be inaccurate if de novo or rare haplotypes are encountered in a patient\(^5\).\(^6\). Due to the chromosomal distance between alleles, current short read technologies in use for PGx are often unable to resolve haplotype information without supplemental statistical phasing or parental genotypic data.

Recently, Oxford Nanopore Technologies Inc. has developed the MinION, a real-time nanopore-based DNA sequencing instrument which is compact, inexpensive and faster than most established DNA sequencing technologies. The time it takes from initiation of library preparation to basecalling the first sequence read is approximately 3 hours and the instrument is capable of detecting long sequence reads in excess of 50kb (according to the manufacturer, ONT). Nanopore-based technology promises major advances in DNA sequencing by offering an inexpensive (e.g., on the order of $1000) pocket-sized device for clinical diagnostics or field experiments.

Here we report nanopore-based sequencing of three clinically relevant PGx genes to identify medically actionable variants and haplotypes without statistical phasing.

Methods

**PCR amplification**

Primer sequences for HLA-A, HLA-B and CYP2D6 are available in Table S1. For CYP2D6, we designed primers to specifically amplify CYP2D6 while not amplifying the 94% identical CYP2D7. PCR primer specificity was verified using UCSC in silico PCR. We used the standard protocol (for fragments up to 8 kb) of the KAPA LongRange HotStart PCR system: 5x KAPALongRange Buffer (without Mg\(^{++}\)) 1x, MgCl\(_2\) (25 mM) 1.75 mM, dNTPs (10 mM each dNTP) 0.3 mM, Fwd primer (10 μM) 0.5 μM, Rev primer (10 μM) 0.5 μM 50ng of genomic DNA (1ul of a 50ng/ul preparation), KAPA LongRange HotStart DNA Polymerase (2.5 U/ul) 1.25 U/50 μl, PCR grade water up to 50μl. For HLA-A and HLA-B, genomic sequence was downloaded from UCSC Browser with common SNPs masked. Primers were designed using Primer3 using parameters of 68°C for optimal annealing temp and 26bp minimum primer length\(^6\).

**Oxford Nanopore genomic DNA library preparation**

The DNA libraries were prepared using the Oxford Nanopore Genomic DNA Sequencing protocol (SQK-MAP003). 1.5μg of PCR product was used (instead of the suggested 1μg based on improved yield in earlier testing) with equimolar amounts of CYP2D6, HLA-A and HLA-B amplicons in solution. DNA was not fragmented because the PCR amplicons were already at the desired size for sequencing and downstream haplotyping (4–5Kbp; Table S1). In accordance with the protocol, we end-repaired the DNA with the NEBNext end repair module (New England Biolabs, cat. no. E6050) and subsequently dA-tailed the sample using the NEBNext dA-tailing module (New England Biolabs cat. no. E6053), prior to ligation of nanopore-specific adapters. All purifications were accomplished with Agencourt AMPure XP beads (Beckman Coulter Inc., cat. no. A63880). Throughout the library preparation, care was taken not to vortex or vigorously pipette/mix the library to avoid shearing the DNA into smaller fragments.

**Oxford Nanopore MinION sequencing and basecalling**

The MinION flowcell (R7.3 flowcell chemistry) was run for 24 hours using the MinKNOW software (v47.3) producing 24,859 fast5 files, corresponding to individual reads from base detection events at specific nanopore channels. Online basecalling was performed using the Metrichor software (v2.23). The MinION outputs 3 reads for each dsDNA molecule that passes through a pore. The leading ssDNA is referred to as “1D template” and its complementary ssDNA strand is the “1D complement”. When both 1D template and 1D complement reads are basecalled, a 2D consensus sequence is determined based on complementarity. We observed 19,655 1D template reads, 9,584 1D complement reads and 7,540 2D reads. The mean lengths were 2,693bp for 1D template, 2,706bp for 1D complement and 3,486bp for 2D consensus.

**Read alignment**

Existing massively parallel sequencing instruments, such as the Illumina HiSeq 2500, produce accurate short reads typically up to ~250bp in length. These sequencers can produce hundreds of millions of reads which need to be rapidly aligned to a reference genome. Current computational methods for accurate alignment...
of these reads, including BWA\(^1\) and Bowtie2\(^2\) are based on the Burrows-Wheeler Transform FM index, and they are designed to align short reads with minimal variation to a reference assembly. BWT-FM methods are insufficiently sensitive to align much longer reads with higher error rates\(^3\). These long reads, generated by single molecule sequencers such as the Oxford Nanopore MinION or the Pacific Biosciences RS II, have a significantly higher error rate, enriched for insertions or deletions (indels) rather than substitutions\(^4\). Mapping of these reads is best suited to aligners that were originally designed for whole genome alignments, such as LAST\(^5\). We chose to use BLASR, originally developed for the Pacific Biosciences system, to align our data, because it was designed to align long error-prone reads rather than genomes\(^6\).

All reads were aligned to the human genome reference assembly GRCh37.p13 (hg19) using default BLASR parameters (gap open penalty = ten, gap extension penalty = zero, minimum seed length = 12). The use of other parameters, such as a gap-open penalty of zero (with default gap extension penalty = zero) did not alter the results, even though it may be expected to do so given the prevalence of indels expects in single molecule nanopore sequencing. The majority of successfully aligned long read fragments were obtained from 2D basecalls (Table 1), and these were of higher quality because they are consensus reads constructed from corresponding 1D template and complement. For the final alignment step, for each separate read event (1D template, 1D complement and 2D consensus), we selected the 2D if it was available. Since the 1D reads typically had lower mapping accuracy and significantly shorter aligned fragment lengths (Table 1), these were not included in our variant or haplotype calling analysis.

Finally, we performed two separate alignments depending on our desired sequencing application: a) gene targets with highly similar nearby genes; and b) highly polymorphic gene targets. The first analysis only selected the single best alignment for each long read. This was critical for the gene CYP2D6 because the CYP2D locus on chromosome 22 harbors two paralogous pseudogenes CYP2D7 and CYP2D8\(^P\). In particular, CYP2D6 and CYP2D7 are highly similar (94% identity, BLAST E-value = 0.0) and are positioned in tandem on the chromosome. By allowing reads to only map to a single best hit, we were able to verify that the PCR selectively amplified CYP2D6 and not nearby related genes (see coverage in Figure 1). Our second analysis was performed due to the high degree of polymorphism in the MHC locus on chromosome 6. As part of the MHC haplotype project\(^7\), multiple reference contigs for this highly variable region are included in the GRCh37 reference assembly as indicated in the release notes (http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html). Since long reads from the NA12878 HLA-A and HLA-B genes mapped to different chromosome 6 reference haplotype contigs, by allowing multiple alignments to the reference (up to 10) for each read, we could gather all reads for a single gene in a single pileup to any of the eight HLA-A/B loci to generate a consensus sequence. For this study, we used the reference chromosome 6 contig NC_000006.11 (not the MHC haplotype project contigs).

### Variant detection and haplotype identification

Due to the long reads, high error rates and continuously evolving error profile of the MinION basecalls at this early stage of technology roll out, variant callers such as the Genome Analysis Toolkit’s UnifiedGenotyper or HaplotypeCaller\(^8\) were unable to identify variants or haplotypes in the MinION sequence data during our trials. Variant and haplotype level information, however, was readily accessible based on coverage of aligned reads, which we extracted using SAMTools via the PySam wrapper (http://github.com/pysam-developers/pysam)\(^9\).

Mean coverage was 1236.4× for CYP2D6 (single best hit alignment), 785.5× for HLA-A (multi-hit alignment) and 1416.3× for HLA-B (multi-hit alignment).

Variants were detected using a naïve threshold requiring 1/3 of reads to contain the variant genotype at that position. While this was effective for substitution detection, we are likely to detect many false positive deletions due to the high deletion error rate (Table 1). Haplotype proportions were identified by interrogating clinical marker positions (Table S2) across all reads aligned to a particular gene to establish the proportion of reads corresponding to each haplotype. Pharmacogenomic haplotypes were verified by comparison to diagnostic data sets (see below) using the MedSavant software (www.medsavant.com; manuscript in review) with the pharmacogenomics app that we developed. The PGx app interprets human pharmacogenomic variants with medically actionable output based on published guidelines established by the Clinical Pharmacogenetics Implementation Consortium (CPIC) and the Pharmacogenomics Knowledgebase (www.pharmgkb.org).

### Validation of genotypes by Complete Genomics, and clinical diagnostic Sequenom MassARRAY and qPCR

Complete Genomics WGS data for NA12878 were obtained from the public 69 genomes project (CG analysis pipeline version 2.0.0; http://www.completegenomics.com/public-data/69-Genomes/)\(^10\).

| Table 1. Basecall and read mapping statistics. |
|------------------------------------------------|
| **Read Type** | **Number of reads** | **Mean Length, unaligned (bp)** | **Number of reads aligned** | **% of reads aligned** | **Mean Length, aligned fragment (bp)** | **Mean substitution frequency** | **Mean deletion frequency** | **Mean insertion frequency** | **Mean mapping accuracy (for each read [match bp/total bp])** |
|------------------------------------------------|
| 1D template | 19655 | 2693.7 | 3793 | 19.3% | 872.8 | 8.9% | 13.9% | 5.7% | 71.5% |
| 1D complement | 9584 | 2705.7 | 2717 | 28.3% | 292.7 | 7.3% | 15.4% | 4.1% | 73.2% |
| 2D consensus | 7540 | 3486.3 | 4761 | 63.1% | 2952.3 | 7.0% | 13.3% | 5.3% | 74.3% |
10ng of genomic DNA from NA12878 was genotyped for 36 SNP, indel and copy number variants for CYP2D6 using the iPLEX® ADME CYP2D6 Panel v1.0, developed by Assays by Agena (formerly Sequenom) on the MassARRAY4 System. Haplotype assignment and copy number determination was done using Typer software version 4.0 (Agena Biosciences).

In parallel, copy number estimation of CYP2D6 was performed using the Taqman copy number assays Hs04502391_cn and Hs04083572_cn (Life Technologies) using the manufacturer’s recommended protocol (Figure S1). The assay was performed in quadruplicate on 10ng genomic DNA for each sample in a 96-well plate. The 10μL reaction mix consisted of 5μL 2× Taqman Genotyping Master Mix (Life Technologies), 0.5μL of 20X copy number assay (described above), 0.5μL TaqMan RNase P Copy Number Reference Assay (Life Technologies cat. no. 4403326), 2μL water and 2μL of 5ng/μL genomic DNA. Cycling conditions for the reaction were 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. Samples were analyzed using the ViiA™ 7 Real-Time PCR System (Life Technologies) and analyzed using Copy-Caller Software (Life Technologies). The HuRef sample was used as a 2-copy calibrator sample.

Validation of haplotypes by statistical phasing and HapMap

Complete Genomics WGS and Sequenom MassARRAY genotypes were statistically phased using the BEAGLE software (v4.0) and the 1000 Genomes Project phase 3 reference panel (http://faculty.washington.edu/browning/beagle/beagle.html). For the HLA-A/B genes, phase information was obtained from the HapMap phase 2 data (http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en). HLA-A/B alleles were determined using the GATK HLACaller software package (http://gatkforums.broadinstitute.org/discussion/65/hla-caller).

Results

To evaluate the MiniION for diagnostic PGx sequencing, we selectively amplified and sequenced the genes CYP2D6, HLA-A, and HLA-B from the CEPH/UTAH pedigree 1463 sample NA12878. CYP2D6 is a pharmacogenetically vital cytochrome P450 gene because it encodes a protein responsible for metabolism of 20% of clinically used drugs11. The diagnostic relevance of 2D6 is derived from its significant polymorphism which contributes to dramatic inter-individual variability in enzyme activity11. Also important are the HLA genes which are among the most polymorphic loci in the human genome, making their sequencing and confident typing difficult with current short read DNA sequencing methods. These three genes were also chosen for sequencing due to their length (4–5Kbp), which did not require long range PCR amplification methods.
PCR amplicons of these three genes from NA12878 (CEPH/Utah Pedigree 1463) were sequenced on the MinION instrument yielding 19655 read events. Each read event could be basecalled in multiple forms, as a template or complement strand (1D) or as a consensus of the two (2D), and we obtained 36779 1D and 2D reads in total. For the purpose of diagnostic evaluation, we chose to align only the consensus 2D reads due to their lower error rate and extended length (Table 1; see Methods). Reads were aligned to the human genome (GRCh37), with an abundance of aligned reads 4–5Kb in length representing full-length PCR amplicons. As well, smaller aligned read fragments were observed, some of which are speculated to be byproducts of shearing during experimental DNA handling (Figure 2A, Table S1).

With depth of coverage of ~1000x for each of the genes, many chromosomal positions were called with 70–90% consensus, demonstrating that as coverage of loci increases on the MinION, confidence improves with regard to specific base calls (Figure 2B). While the MinION basecalls are emitted with a comparatively high error rate (Table 1), the majority of errors appear to be randomly distributed across the length of the reads, which is why increasing coverage can yield a consensus that matches variant calls from existing sequencing and genotyping platforms such as Illumina, Complete Genomes and Sequenom.

MinION-called variants and haplotypes were validated against statistically phased genotypes from multiple platforms including Complete Genomes and Sequenom MassARRAY (see Methods). Based on the statistically phased genotypes, we determined that NA12878 possesses both the *3 and *4 loss-of-function alleles for CYP2D6, and this *3/*4 diplotype is interpreted as reduced metabolism of drugs such as codeine (an opiate) and olanzapine (an atypical antipsychotic)\textsuperscript{20–22}.

CYP2D6 haplotype proportions were identified by interrogating clinical marker positions across all aligned reads to establish the proportion of reads corresponding to each PGx haplotype (Table S2). Only reads spanning all clinical markers were included (n = 404), so that haplotypes could be measured by linkage of markers on a single DNA molecule. The MinION data confirmed the statistically-phased haplotypes by direct interrogation of markers from individual reads (Figure 2C). However, we also observed a prominent *2 haplotype, which we could not account for given that our Sequenom MassARRAY and qPCR results indicated that the CYP2D6 locus was diploid (no copy number variation) and could only correspond to a *3/*4 diplotype. To determine whether the *2 haplotype could arise from mismatched CYP2D7 DNA, which was not supposed to be PCR amplified (see Methods), we interrogated four positions with different bases between CYP2D6 and CYP2D7.

**Figure 2.** A. Length distribution of aligned reads. 4–5Kb reads represent full-length PCR amplicons. Slightly smaller fragments were likely byproducts of shearing during DNA handling in the experimental protocol. B. With depth of coverage of ~1000x for each of the genes, many chromosomal positions were called with 70–90% consensus. This is a short window of aligned reads for the *4 locus of CYP2D6 with over 1200x coverage. The heterozygous *4 allele rs1065852 is indicated with the arrow. C. Proportions of haplotypes of CYP2D6, HLA-A and HLA-B when directly measured from individual reads spanning all haplotype markers.
and CYP2D7 reference sequences and found that all reads corresponded to CYP2D6 (Supplementary File S1). Finally, we hypothesized that the *2 haplotype might arise due to *3 and *4 duplexes forming during PCR (effectively outcompeting the primer binding during the annealing step), but this was ruled out by identifying the *2 haplotype using only 1D reads. It is possible that this *2 haplotype arose either due to early cycle template switching during PCR or sample contamination. Also, the relative proportion of haplotypes was likely skewed by potential biases during PCR amplification.

The HLA-A and HLA-B haplotypes were determined in the same way as the CYP2D6 haplotypes, using predefined markers from the HapMap project. In HLA-A (only spanning reads, n = 203), the most abundant haplotype matched the transmitted haplotypes of the parents NA12891 and NA12892, which both transmitted an identical haplotype (Figure 2C). Accounting for the errors in MinION sequencing, when allowing for a single mismatch in the haplotype, ~85% of reads confirm the NA12878 diploid. In HLA-B (only spanning reads, n = 202), the majority of reads corresponded to the transmitted and untransmitted haplotypes of the parent NA12891, with only 8.4% of reads corresponding to the transmitted haplotype of parent NA12892 (Figure 2C). This could be a result of potential contamination suspected earlier in described above with CYP2D6. As suggested for CYP2D6, the relative proportion of HLA haplotypes was likely also affected by PCR bias during amplification. HLA alleles were called with 4-digit resolution using the GATK HLAcaller, but due to the high error rates of nanopore reads, HLA alleles did not match with alleles called using HapMap data (Table S3).

Long sequence reads aid haplotype identification by determining which genetic variants are in phase (i.e. on the same DNA strand). If the TPMT genotypes from the example above were sequenced using nanopore-based long read technology, the *1/*3A diplotype would likely be called correctly (note that rs1142345 and rs1800460 are only 8,310bp apart).

While nanopore sequencing with the MinION is demonstrably error-prone in its current stage of development, we assert that this technology holds promise for clinical applications because accurate consensus sequences can be built with sufficient coverage given the high number of reads generated. As well, we have been able to successfully call haplotypes from long reads de novo in the absence of parental haplotypes or statistical phasing. The MinION device produced sufficiently long mappable reads to phase all variants in the loci examined. As error rates on the MinION decrease, we can expect to deconvolute these data into more accurate diplotypes with less noise and will be able to measure how much multi-sample multiplexing can be supported by a single run.

According to the CPIC, 63 genes and 132 drugs have guidelines for pharmacogenomic status (http://www.pharmgkb.org/cpic/pairs), and this list is constantly expanding. With increasing guidelines and demands for PGx in the clinic, affordable and rapid nanopore sequencing may hold great utility.

Data availability
figshare: Nanopore reads and alignments, doi: http://dx.doi.org/10.6084/m9.figshare.1289717

Raw nanopore reads and alignment files are available at the NCBI Sequence Read Archive, accession SRP051851 (http://www.ncbi.nlm.nih.gov/sra/?term=SRP051851).

Conclusions and discussion
Phasing of genotypes is critical to prevent misinterpretation of PGx variants. The importance of correct phasing of PGx genotypes is illustrated with the gene TPMT, which plays a critical role in the metabolism of thiopurine, a drug used to treat acute lymphoblastic leukemia. In a recent study, an individual was reported to have a TPMT *3B/*3C diplotype, based on observed heterozygous genotypes for the rs1142345 and rs1800460 variants, but this was a misinterpretation due to faulty haplotyping and *1/*3A is the correct diplotype. The rs1800460 variant is present in both *3A and *3B haplotypes while the rs1142345 variant is present in both *3A and *3C haplotypes. As a result, it was possible for the individual to have a *1/*3A diplotype or a *3B/*3C diplotype. Clinically, a *1/*3A diplotype corresponds to an intermediate metabolizer, requiring a 30–70% reduction in thiopurine dose, while a *3B/*3C diplotype corresponds to a poor metabolizer with a 90% reduction in dose. An individual who receives a standard dose and is a poor metabolizer can experience fatal toxicity, while a low dose for a normal metabolizer can lead to disease progression. Clinical trials have demonstrated the medical importance of TPMT haplotyping in treatment of myeloid leukemias and non-malignant immunologic disorders.

Author contributions
RA, AS and GDB conceived the study. TAP designed and performed the PCR amplification. RA and DT performed the library preparation. RA performed the sequencing and downstream analysis. RA, TAP and GDB drafted the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests
R.A. is a member of the Oxford Nanopore Technologies Inc. MinION Access Programme and the MinION instrument and R7.3 flowcells were received free of charge.

Grant information
This study was funded by a Large-scale Applied Project grant from Genome Canada and the Ontario Genomics Institute (grant ID OGI-068).

We confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preperation of the manuscript.

Acknowledgments
The authors thank Stephen Scherer and Peter Ray of the Hospital for Sick Children for contributing diagnostic validation data.
**Supplementary material**

**Figure S1.** CopyCaller software analysis of data from Taqman copy number assays Hs04502391_cn and Hs04083572_cn. CYP2D6 from sample NA12878 was observed to be diploid.

**Table S1.** Primer sequences and amplicon lengths for HLA-A, HLA-B and CYP2D6.

| Gene   | Primer Name | Primer seq                             | Chromosomal Coordinates       | Amplicon Size |
|--------|-------------|----------------------------------------|-------------------------------|---------------|
| CYP2D6 | CYP2D6-2F   | TAGCTCCCTGAGCCCATGATTTGCTTT            | chr22:42,522,077-42,527,144   | 5,067 bp      |
| CYP2D6 | CYP2D6-2R   | CCGTGTTAATCCGAGGCAAGCTTGGAG           |                               |               |
| HLA-A  | HLAA-2F     | AGAAGAGTCCAGGACAGTAGTAAAGTG           | chr6:29,909,854-29,913,805    | 3,951 bp      |
| HLA-A  | HLAA-2R     | TTCTACTGAGGCGCAAGGCAATGGAG            |                               |               |
| HLA-B  | HLAB-2F     | TGGATTCAGCAACAGACGCTAAGACACG          | chr6:31,321,279-31,325,303    | 4,024 bp      |
| HLA-B  | HLAB-2R     | GCCTCTCCCTGCTTCAAGACAGATCTCT          |                               |               |

**Table S2.** Haplotype translation table for CYP2D6.

| Haplotype Id | CYP2D6 | rs1065852 | rs28371706 | rs5030655 | rs3892097 | rs35742686 | rs5030656 | rs16947 | rs28371725 | rs1135840 |
|--------------|--------|-----------|------------|-----------|-----------|------------|-----------|---------|------------|-----------|
| PA165816576  | *1     | G         | G          | A         | C         | T          | CTT       | G       | C          | C         |
| PA165816577  | *2     | G         | G          | A         | C         | T          | CTT       | A       | C          | G         |
| PA165816578  | *3     | G         | G          | A         | C         | -          | CTT       | G       | C          | C         |
| PA165816579  | *4     | A         | G          | A         | T         | T          | CTT       | G       | C          | G         |
| PA165948092  | *5     | -         | -          | -         | -         | -          | -         | -       | -          | -         |
| PA165816581  | *6     | G         | G          | C         | T         | CTT        | G         | C       | C          | C         |
| PA165948317  | *9     | G         | G          | A         | C         | T          | -         | G       | C          | C         |
| PA165816582  | *10    | A         | G          | C         | T         | CTT        | G         | C       | G          | C         |
| PA165816583  | *17    | G         | A          | C         | T         | CTT        | A         | C       | G          | G         |
| PA165816584  | *41    | G         | G          | A         | C         | T          | CTT       | A       | T          | G         |

**Table S3.** HLA alleles called with 4-digit resolution using the GATK HLACaller.

| Locus  | HapMap A1 | HapMap A2 | Nanopore A1 | Nanopore A2 |
|--------|-----------|-----------|-------------|-------------|
| HLA-A  | 0101      | 1101      | 0132        | 0312        |
| HLA-B  | 0801      | 5601      | 0765        | 5510        |
Supplementary File S1
BLAST alignment of CYP2D6 and CYP2D7. These genes have 94% identity (E-value = 0.0). To determine if our reads were generated from CYP2D7 reads, we interrogated reads at the paralogous positions circled in red.
Dot Matrix View

Descriptions
Sequences producing significant alignments:

| Description                                                                 | Max score | Total score | Query cover | E value | Ident | Accession     |
|----------------------------------------------------------------------------|-----------|-------------|-------------|---------|-------|---------------|
| Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly                    | 6582      | 6582        | 99%         | 0.0     | 94%   | NC_000022.10  |

Alignments

Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly
Sequence ID: refNC_000022.10 Length: 51304566 Number of Matches: 1
Range 1: 42522501 to 42526883

| Score       | Expect | Identities | Gaps | Strand | Frame |
|-------------|--------|------------|------|--------|-------|
| 6582 bits(3564) | 0.0    | 4142/4408(94%) | 91/4408(2%) | Plus/Plus |       |

Features:

Query 42536214
Subject 42522501

Query 42536274
Subject 42522561

Query 42536334
Subject 42522621

Query 42536393
Subject 42522680

Query 42536453
Subject 42522740

Query 42536513
Subject 42522800

Query 42536573
Subject 42522860

Query 42536633
Subject 42522920

Query 42536693
Subject 42522980

Query 42536753
Subject 42523040

Query 42536813
Subject 42523100

Query 42536873
Subject 42523160

Query 42536933
Subject 42523220

Query 42536993
Subject 42523279

http://blast.ncbi.nlm.nih.gov/Blast.cgi
References

1. Evans WE, Hon YY, Borngaars L, et al: Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine. J Clin Oncol. 2001; 19(8): 2293–2301. PubMed Abstract

2. Relling MV, Gardner EE, Sandborn WJ, et al: Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin Pharmacol Ther. 2011; 89(3): 387–391. PubMed Abstract

3. Mizzi C, Peters B, Mitropoulou C, et al: Personalized pharmacogenomics profiling using whole-genome sequencing. Pharmacogenomics. 2014; 15(9): 1223–1234. PubMed Abstract

4. Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81(5): 1084–1097. PubMed Abstract

5. Browning SR, Browning, BL: Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011; 12(10): 703–714. PubMed Abstract

6. Untergasser A, Cutcutache I, Koressaar T, et al: Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012; 40(15): e115. PubMed Abstract

7. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14): 1754–1760. PubMed Abstract

8. Langmead B, Saltzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4): 357–359. PubMed Abstract

9. Chaisson MJ, Tesler G: Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012; 13: 238. PubMed Abstract

10. Kielbasa SM, Wan R, Sato K, et al: Adaptive seeds tame genomic sequence comparison. Genome Res. 2011; 21(3): 487–493. PubMed Abstract

11. Zanger UM, Schwab M: Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther. 2013; 138(1): 103–41. PubMed Abstract

12. Horton R, Gibson R, Coggil P, et al: Variation analysis and gene annotation of eight MHC haplotypes: The MHC HaploType Project. Immunogenetics. 2008; 60(1): 1–18. PubMed Abstract

13. van der Auwera GA, Carneiro MO, Hartl C, et al: From fastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinform. 2013; 11(110): 11.10.1–11.10.33. PubMed Abstract

14. Li H, Handsaker B, Wysoker A, et al: The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–2079. PubMed Abstract

15. Dmancic R, Sparks AB, Callow MJ, et al: Human genome sequencing using unchained base reads on self-assembling DNA nanocrystals. Science. 2010; 327(5961): 78–81. PubMed Abstract

16. Frazier KA, Ballinger DG, Cox DR, et al: A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164): 851–861. PubMed Abstract

17. Leckband SG, Kelsoe JR, Dunnenberger HM, et al: Clinical Pharmacogenetics Implementation Consortium guidelines for HLA-B genotype and carbamazepine dosing. Clin Pharmacol Ther. 2013; 94(3): 324–8. PubMed Abstract

18. Martin MA, Hoffman JM, Freimuth RR, et al: Clinical Pharmacogenetics Implementation Consortium Guidelines for HLA-B Genotype and Abacavir Dosing: 2014 update. Clin Pharmacol Ther. 2014; 95(5): 499–500. PubMed Abstract

19. Hershel MS, Callaghan JT, Tassaneeyakul W, et al: Clinical Pharmacogenetics Implementation Consortium guidelines for human leukocyte antigen-B genotype and allopurinol dosing. Clin Pharmacol Ther. 2013; 93(2): 153–8. PubMed Abstract

20. Zhou, SF: Polymorphism of human cytochrome P450 2D6 and its clinical significance; part II. Clin Pharmacokinet. 2009; 48(12): 761–804. PubMed Abstract

21. Creers KR, Gaagdijk A, Dunnenberger HM, et al: Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. Clin Pharmacol Ther. 2014; 95(4): 376–82. PubMed Abstract

22. Hicks JK, Sven JJ, Thom CF, et al: Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. Clin Pharmacol Ther. 2013; 93(5): 402–8. PubMed Abstract

23. Odellberg SJ, Weiss RB, Hata A, et al: Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res. 1995; 23(11): 2049–2057. PubMed Abstract

24. Listgarten J, Brumme Z, Kadie C, et al: Statistical resolution of ambiguous HLA typing data. PLoS Comput Biol. 2008; 4(2): e1000016. PubMed Abstract

25. Brownstein CA, Margulies DM, Manzi SF: Misinterpretation of TPMT by a DTC Genetic Testing Company. Clin Pharmacol Ther. 2014; 95(6): 598–600. PubMed Abstract

26. Stanulla M, Schoefler E, Fohr T, et al: Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia. JAMA. 2005; 293(12): 1485–1489. PubMed Abstract

27. Ammar R, Paton TA, Torti D, et al: Nanopore reads and alignments, figshare. 2015. Data Source
Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 30 March 2015
doi:10.5256/f1000research.6463.r7919

Thomas Hoenen
Laboratory of Virology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA

This manuscript describes the analysis of three PCR-fragments generated from a human reference sample using a MinION sequencing device. Despite the preliminary nature of the data, and the fact that the paper is based on a single run of a single sample, it will nevertheless be of significant interest, and publication as a research note is in my view justified. In fact, given the broad interest in this emerging new technology, and its applicability to all life sciences (not just human (pharmaco-)genetics), the readership might be much wider than anticipated by the authors. This constitutes at the same time the biggest weakness of the paper, since it is in parts rather inaccessible for readers from fields other than human genetics (e.g. virology or microbiology), for whom this paper might nevertheless be highly relevant.

In order to address the concern that the manuscript is based on a single MinION run, and given the nature of the MAP program, it should be no problem for the authors to repeat the experiment with another sample, and while doing so specifically address the concerns of Dr. Kennedy. In addition, the authors should strive to improve accessibility to a wider audience wherever possible. Finally, it would be helpful to include additional experimental details that are currently missing.

Specific comments:

- PCR cycling conditions should be provided.
- The authors refer to the SQK-MAP003 sequencing protocol. As far as I am aware, Oxford Nanotechnologies does not make the detailed protocol available to people outside the MAP program, although there have been indications that a non-technical version of the protocol will be made available for publication purposes. The authors should reference such a protocol (including a link) as soon as possible, and approach ONT about making it available to the general public, if this hasn’t happened already.
- Did the authors perform a PCR purification prior to the library preparation? If so, what was the volume/ratio of Agencourt beads to sample?
- How did the authors extract reads from the fast5 files? Did they use poretools, or another tool (which should be referenced)?
- In general, providing more details about the exact bioinformatics workflow would be helpful.
- What was the rationale for the cut-off of 1/3 of reads for variant calling?
- It seems odd that the length of the aligned fragments from the 1D reads is so much shorter than the read length. In our hands (using a similar approach on ~2 kb PCR products amplified from virus genomes, albeit with a later chemistry/protocol version (SQK-MAP004) and using LAST for the alignment) we get much longer average alignments (92% for 2D reads, 82% for template reads,
and 85% for complement reads, vs. 85%, 23% and 11% reported by the authors). This could either indicate significant advances in base-calling accuracy since the authors performed their experiments, or that their alignment is suboptimal. It might be very interesting to see whether the authors can get longer alignments using LAST or other alignment softwares in their workflow.

- It would be very helpful if the authors could repeat the experiment using the newest chemistry/protocol/software versions, which have changed considerably over the last months. At the same time this would allow them to address many of the concerns of Dr. Kennedy.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Competing Interests:** I am a participant in the MinION access program. Other than that, I have no connections to the authors, and I do not know them. I do not have any competing interests.
a haploid (CYP2D6*5) case or multicopy case, to provide confidence that the assay was working as expected. This is relevant because of the question raised by the *2 haplotype which shows up in the MinIon analysis of NA12878.

Results

Page 6, first para: This description of the possible origins of the mystery CYP2D6*2 haplotype needs some editing for improved clarity. Not clear what is meant by “*3 and *4 duplexes forming”. Also not clear what would cause PCR biases alluded to in the last sentence of this paragraph (and of the following paragraph).

Typos/suggested edits:

Abstract

MinIOn > MinIon
Should refer to NA12878 as “reference sample” rather than just “sample”
Suggest sentence be modified thus for clarity: “…statistically phased genotype data from Complete Genomics and Sequenom.”
Suggest delete “Standalone” in penultimate line.

Introduction

Suggest this sentence be changed: However, existing methods have various limitations, which may lead to adverse drug responses. > However, existing methods have various limitations, which may lead to failure to detect variants of pharmacogenetic significance.

Methods

Page 3, 2nd para: indels expects > indels expected
Page 4 para 2 – clarify “The HuRef sample…”
Figure 1 title: Integrate > Integrative

Results

Page 5, para 4. First sentence should read thus, for improved clarity: “CYP2D6 haplotype proportions in MinIon data were identified by interrogating clinical marker positions…”

Supp File S1
It would be helpful to indicate which of the sequences is CYP2D6 and which is CYP2D7.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I am also a participant in the MinIon Access Programme, and my laboratory is working with the device. I met Dr Ammar at a meeting organised by ONT, and we have a shared interest in use of this technology for pharmacogenetic analysis. However, I have not been involved in the work described here.