Long-read sequencing offers path to more accurate drug metabolism profiles

In the complex drug discovery process, one of the looming questions for any new compound is how it will be metabolised in a human body. While there are several methods for evaluating this, one of the most common involves CYP2D6, the enzyme encoded by the cytochrome P450-2D6 gene. This enzyme is involved in metabolising a quarter of all commonly used medications, making it an important target for ADME and pharmacogenomics studies. It is known to activate some drugs and to play a role in the deactivation or excretion of others.

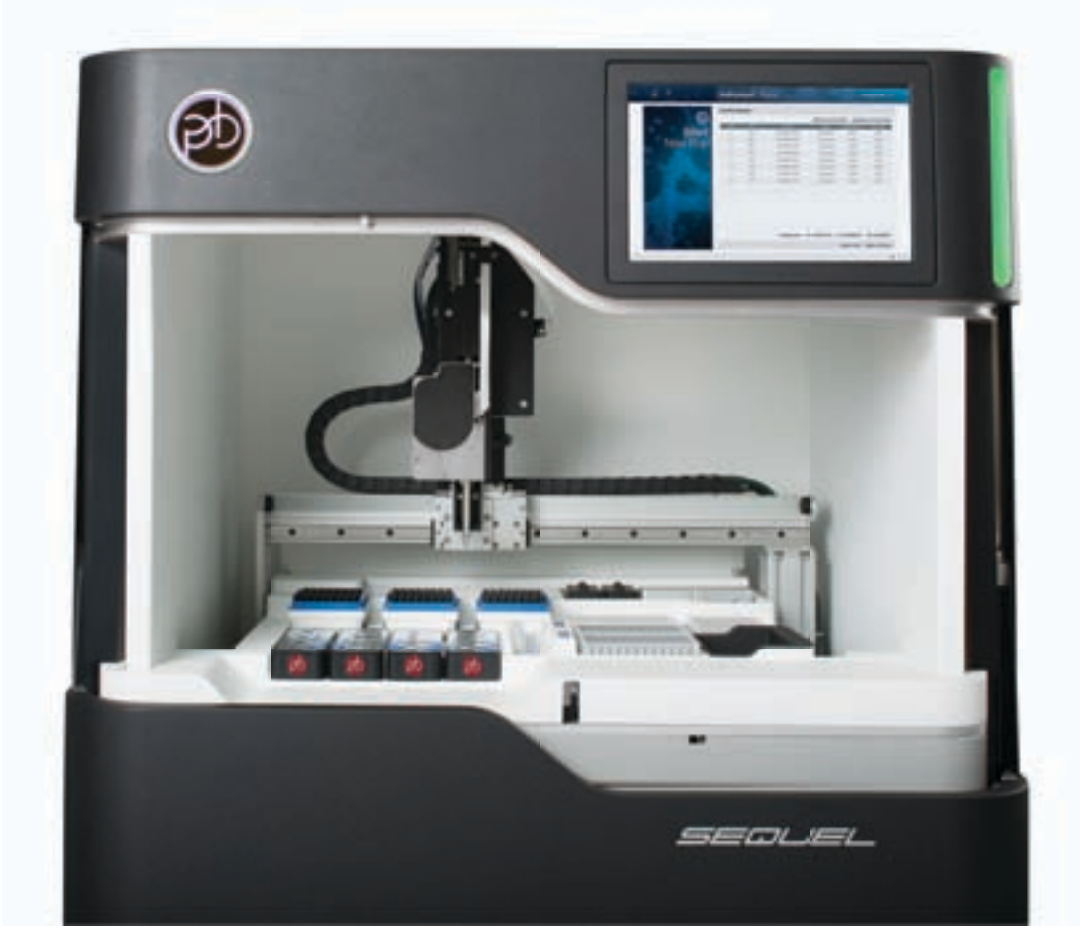
By Dr Jenny Ekholm

The CYP2D6 gene harbours many variants that contribute to the differences in how people metabolise everything from antipsychotic therapies to painkillers. From a pharmacology perspective, it is critical to understand the function of this gene in order to develop the best possible compounds, determine optimal dosing protocols, and target drugs to the right patient population.

However, elucidating CYP2D6 function through variant detection has been a remarkably difficult task. The gene is highly polymorphic, with more than 150 different alleles classified for it – a number that is expected to keep growing. CYP2D6 is known for complex structural variations, such as rearrangements, duplications and deletions. To make matters worse, it has a nearby pseudogene with incredibly high sequence identity; efforts to query CYP2D6 are frequently confounded by misleading results from this pseudogene.

Interrogating this important gene is typically done through genotyping or sequencing. FDA-approved molecular diagnostic platforms scan the gene, looking for only the most common alleles. These platforms cannot identify novel alleles, nor can they distinguish clinically meaningful variants reflective of a less common allele. In drug discovery research, scientists often use TaqMan assays or Luminex for genotyping, while some use Sanger sequencing or short-read sequencing to characterise CYP2D6. Unfortunately, these sequencing technologies lack the resolution needed to provide a clear and accurate picture of the gene and its pseudogene.

Recently, long-read sequencing has become an alternative approach for resolving this region.



The Sequel System from PacBio

Single molecule, real-time (SMRT) sequencing produces extremely long reads that fully span the CYP2D6 gene. Scientists have now demonstrated that SMRT sequencing offers accurate, base-level resolution even for the most complex elements within this gene. With a complete view of the CYP2D6 gene, it is now possible for scientists to achieve a more thorough understanding of drug metabolism for each new therapy considered. Meanwhile, the rise of population-based genome studies has shown that some clinically-relevant CYP2D6 variants are specific to certain populations or ethnicities, offering better guidance for pharmacogenomic studies. By combining SMRT sequencing with those study results, it should be possible to assemble a complete database of CYP2D6 alleles and to better predict metaboliser phenotypes for a more straightforward path to selecting a target population that will respond most effectively to a drug.

CYP2D6 complexity

Located on chromosome 22, the CYP2D6 gene shares 97% of its sequence with the pseudogene CYP2D7P1, found only 10 kilobases away. The situation already presents a real challenge for genotyping or sequencing platforms, and is made even more complex by the presence of copy number variation, structural variation and more. In this genetic morass, it is not enough to detect variants; proper classification of a CYP2D6 metaboliser profile requires that variants be phased into distinct haplotypes to resolve the precise number of functional or dysfunctional gene copies in an individual. When a gene duplication is detected, for instance, technologies that cannot phase alleles usually have no way of distinguishing between two copies of an allele with increased function or two copies of an allele with decreased function.

Traditionally, scientists have used Sanger sequencing or genotyping platforms to profile the CYP2D6 gene. More recently, next-generation sequencing (NGS) has been incorporated as well. None of these approaches can provide a complete view of the gene and its breadth of variation, though. Genotyping tools are restricted to seeing known variants, and because there are so many catalogued alleles for this gene, most platforms must limit their queries to the top dozen or so vari-

References

1 Qiao, W, Yang, Y, Sebra, R, Mendiratta, G, Gaedigk, A, Desnick, RJ and Scott, SA (2016). Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. Hum. Mutat., 37:315-323. doi: 10.1002/humu.22936. http://onlinelibrary.wiley.com/doi/10.1002/humu.22936/abstract.

2 Buermans, HPJ, Vossen, RHAM, Anvar, SY, Allard, WG, Guchelaar, H-J, White, SJ, den Dunnen, JT, Swen, JJ and van der Straaten, T (2017). Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing. Hum. Mutat., 38: 310-316. doi: 10.1002/humu.23166. http://onlinelibrary.wiley.com/doi/10.1002/humu.23166/full.

3 Seo, J, Rhie, A, Kim, J, Lee, S et al (13 October 2016). De novo assembly and phasing of a Korean human genome. Nature 538, 243–247, doi:10.1038/nature20098. http://www.nature.com/nature/journal/v538/n7624/full/nature20098.html.

4 Moya, G, Dorado, P, Ferreiro, V, Naranjo, MEG, Peñas-Lledó, EM and LLerena, A (12 April 2016). High frequency of CYP2D6 ultrarapid metabolizer genotypes in an Ashkenazi Jewish population from Argentina. The Pharmacogenomics Journal. doi:10.1038/tpj.2016.27. http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201627a.html.

5 Jittikoon, J, Mahasirimongkol, S, Charoenyingwattana, A, Chaikledkaew, U et al (2016). Comparison of genetic variation in drug ADME-related genes in Thais with Caucasian, African and Asian HapMap populations. Journal of Human Genetics 61, 119-127; doi:10.1038/jhg.2015.115. http://www.nature.com/jhg/journal/v61/n2/abs/jhg2015115a.html.

Continued on page 28

ations. Sanger sequencing is laborious and costly, making it an unrealistic option for the high-throughput demands of drug discovery. NGS offers a higher-capacity means of interrogating the gene and requires no *a priori* knowledge, but most tools generate reads only a few hundred bases long. These snippets get stitched together, but with the complexity and repetition of CYP2D6 sequence and its pseudogene, many reads get conflated or misaligned. The end result is an error-prone alignment that makes it all but impossible to call a CYP2D6 profile with accuracy or to identify novel CYP2D6 alleles.

These limitations have prevented scientists from discovering the full universe of natural CYP2D6 variation, and therefore from identifying drug metabolism profiles for all individuals. Having this information is essential for removing confounding variables from pharmacogenomic studies and other projects designed to understand a drug's activity in each person.

A new approach

Recently, scientists have reported results from evaluating SMRT sequencing to call CYP2D6 profiles. With its multi-kilobase-long reads, this technology can fully span the gene and its copies in a single amplicon. Read length is important for phasing, too, making it possible to link distant SNPs or other types of variation.

At the Icahn School of Medicine at Mount Sinai, a team led by Stuart Scott, Wanqiong Qiao and Yao Yang conducted the first evaluation of SMRT sequencing for the CYP2D6 gene[1]. They used long-range PCR to build 5kb amplicons covering the entire gene as well as any upstream or downstream copies.

The approach was first validated on a set of Coriell DNA samples that had been previously classified with other CYP2D6 profiling tools. Results from the SMRT sequencing pipeline were concordant with known profiles for all 10 samples, and in some cases showed variation that had not been found with the other technologies. With that information, scientists were able to spot novel alleles and allele-specific duplication while refining genotype calls for some samples.

Next, the team moved on to samples that had been profiled before but had yielded ambiguous results. With SMRT sequencing, they were able to resolve the full amplicon sequence without any need for assembly, eliminating the troublesome step that had caused some of the previously inconclusive findings for those 14 samples. The new long-read data offered clear explanations for the discrepancies seen earlier, providing improved resolution and structural variant detection.

Perhaps the most important result of this investigation was the discovery that existing CYP2D6 profiling methods misclassify an individual's genotype more often than expected. While the sample numbers in the study were small, each testing series led to the revision of genotype calls for some 20% of samples. Many of these changes involved reporting novel alleles, while others reclassified samples to rare alleles not covered by common CYP2D6 genotyping platforms. The scientists were not looking for novel alleles, but found three of them anyway. For a region as well-characterised as CYP2D6, this was an unexpected finding that suggests there is far more variation in this gene that has been missed due to technical limitations.

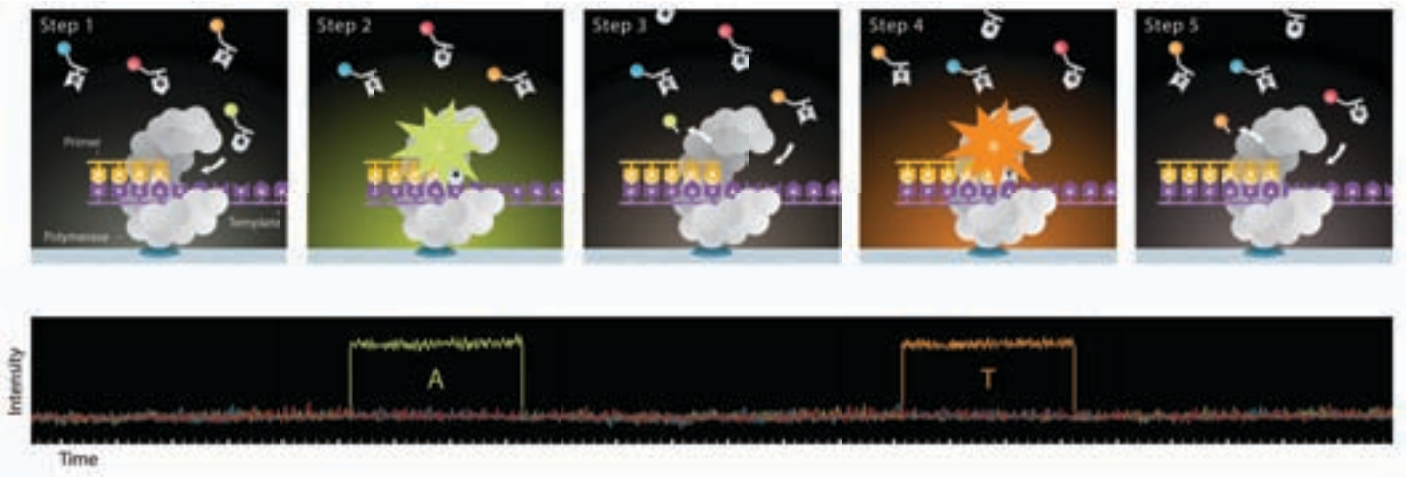
Since that original study, the Mount Sinai team has improved on the method to make it robust for routine lab use. They incorporated barcoding and a higher-throughput platform for SMRT sequencing to allow for multiplexing as many as 384 samples in a run, generating better than 100-fold coverage of CYP2D6 for each sample.

In a separate project, Henk Buermans and Tahar van der Straaten led a team at Leiden University Medical Center to assess the performance of SMRT sequencing for CYP2D6 profiling[2]. Like the Mount Sinai approach, these scientists used long-range PCR and barcoded amplicons to cover the gene and its copies. However, they used a two-step barcoding system and a longer amplicon (6.6kb) to generate more information in a streamlined protocol.

The Leiden scientists sequenced 24 samples, comparing long-read results to data generated from a standard CYP2D6 genotyping platform. The SMRT sequencing approach produced full-length CYP2D6 sequences for all individuals, and results were in agreement with the genotyping calls. Of the variants detected, nearly 10% were unique and many were novel, including SNPs, insertions and deletions. The team was able to refine genotype calls based on the long-read data and add to the community's understanding of natural variation in the gene.

Population studies

As the catalogue of possible CYP2D6 alleles has been fleshed out, it has become increasingly possible to associate some specific metaboliser profiles with certain populations where those genotypes are most common. The rise of population-focused genome projects has also contributed insight by finding variants that appear to exist only in specific groups.



The SMRT Sequencing Process
Step 1: Fluorescent phospholinked labelled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats

For example, an effort to produce a reference-grade genome assembly for a Korean individual led to the identification of a clinically relevant CYP2D6 duplication. Scientists at Seoul National University and other institutions combined many technologies, including SMRT sequencing, to generate a high-quality, *de novo* assembly from a Korean genome[3]. With long-read sequence data, they were able to build a highly-contiguous assembly and identify thousands of structural variants that had never been seen before. They also phased the genome, creating separate assemblies for each haplotype.

With that foundation, the team examined certain areas of interest in greater depth, including the CYP2D6 gene. The individual sequence harboured a duplication of the gene in one allele, a finding that likely has clinical relevance for that person's ability to metabolise drugs.

Several recent population studies have highlighted the difference in allele frequencies among various groups. For instance, scientists compared the results of CYP2D6 profiles among people of Ashkenazi Jewish descent and North American Caucasians or Argentines, finding in both cases that ultra-rapid metaboliser profiles were significantly more common in the Ashkenazi population[4]. That echoed previous findings that Ashkenazi Jews more frequently had increased CYP2D6 metabolic activity but expanded the comparison to populations outside the US This information could have important implications for therapeutic selection and dosing for people of Ashkenazi descent.

In an analysis of the Thai population, a team assessed the activity of CYP2D6 and many other genes related to drug metabolism, absorption and more[5]. Results of the genetic variants found in nearly 200 Thai individuals were compared to publicly-available HapMap data for Caucasian, African and Asian populations. The team uncovered variants in CYP2D6 and other genes that showed statistically significant differences in allele frequency between Thais and the other groups. "The results could explain clinical variability in pharmacokinetics and pharmacodynamics of drugs in Thais based on genetic variations," the scientists concluded in their publication.

In a final example, scientists found greater natural genetic diversity in CYP2D6 alleles in African populations compared to other groups[6]. While increased genetic diversity among Africans is widely accepted in general, these particular findings suggest that differences in drug metabolism-related genes could explain why more adverse drug reactions are reported among people in Africa. The team's analysis highlights "a need for optimisation of drug therapy and drug development there," the scientists reported in their paper, calling for new efforts to "discover uniquely African alleles and to identify populations at a potentially increased risk of drug-induced adverse events or drug inefficacy."

Moving forward

The ability to fully resolve CYP2D6 with SMRT sequencing affords us new opportunities to improve the success rates in drug discovery and development. One obvious benefit is that it is now possible to conduct enough sequencing to characterise the breadth of natural genetic variation for CYP2D6 – and eventually for all genes related to drug metabolism. It is remarkable to realise that our understanding of CYP2D6 activity is no longer limited by technology, but by the availability of samples. That is a much more straightforward

Continued from page 26

problem to solve, particularly as the results of more public genome projects and population studies become available and as institutions find new ways to share samples. Future efforts will focus on characterising CYP2D6 variation among populations that have been under-represented in studies so far.

Of course, once the catalogue of CYP2D6 alleles is complete, it will take many studies to detail the phenotypic effect of each possible variant. This will not be easy, but having associated enzyme activity phenotypes with known alleles in the past, at least we are familiar with the methods required for this task.

Another important step will be determining how closely the natural CYP2D6 variation in humans is mirrored by animal models used for evaluating new compounds. Several investigations have already compared the human gene to its ortholog in other species. Last year, a team of scientists reviewed analyses of CYP2D genes in non-human primates[7]. They found material differences in how various species metabolised certain drugs, which may be reflective of divergent gene sequences. "Further studies are warranted to elucidate the structural and functional features of CYP2D members in non-human primates and thus offer a solid base for the application of these animals in drug discovery," the scientists recommended in the paper. In a separate publication, scientists used computational analysis to identify and interrogate genes related to the cytochrome P450 family in organisms ranging from plants to fish to mammals [8]. They found a high level of conservation for most of the genes, suggesting in their report that an improved understanding "of the evolutionary relationships and functional implications of the human CYP superfamily" would be helpful for drug discovery.

A longstanding challenge in pharmacogenomics has been the lack of consistent results across studies, leading some in the field to question the clinical utility of trying to associate traits such as CYP2D6 genotype with drug selection or dosing. However, it is quite possible that these studies have not been as robust as expected; after all, if 20% of participants had their CYP2D6 genotype misclassified – as the Mount Sinai work suggests is likely – that could explain why results were inconsistent. Future studies and clinical trials might produce more reliable and clinically meaningful results by incorporating SMRT sequencing for a high-resolution view of the CYP2D6 locus in participants. This has major implications for success rates in matching the safest and most effective drug, at the right dose, to the right person. DDW

Dr Jenny Ekholm *is a Senior Scientist specialising in human biomedical applications at PacBio. She has more than 10 years' experience in human genetics, applied techniques, statistical analysis and laboratory management. She earned her PhD in human/medical genetics at the University of Helsinki and completed a postdoctoral fellowship at UCLA.*

6 Rajman, I, Knapp, L, Morgan, T and Masimirembwa, C (2017). Implications of African allele frequency variation in CYP450 genes for drug safety, efficacy and the future of drug development. EBioMedicine. ISSN 2352-3964 (Volume 17, March 2017, Pages 67-74). http://www.sciencedirect.com/science/article/pii/S2352396417300762.

7 He, Z, Chen, X, Yang, Y and Zhou, S (July 2016). A Comparison of Non-Human Primate Cytochrome P450 2D Members and the Implication in Drug Discovery. Current Drug Metabolism. Volume 17, Number 6, July 2016, pp. 520-527(8). http://www.ingentaconnect.com/contentone/ben/cdm/2016/00000017/00000006/art00003.

8 Pan, S, Xue, D, Li, Z, Zhou, Z et al (2016). Computational Identification of the Paralogs and Orthologs of Human Cytochrome P450 Superfamily and the Implication in Drug Discovery. Int. J. Mol. Sci. 2016, 17(7), 1020; doi:10.3390/ijms17071020. http://www.mdpi.com/1422-0067/17/7/1020/htm.