High-fidelity sequencing of complex genomic regions and full-length transcripts

Introduction

In this technical application note we demonstrate how long reads can be used to study hard-to-sequence regions. We will give an overview of the methods and considerations when targeting genomic regions or transcript sequencing and present various results and examples of applications.

The human genome is comprised of complex genetic variants spanning thousands of bases with important impacts in development and disease. The transcriptome also carries its share of complexity such as splicing of RNA isoforms and gene fusions. The read lengths obtained by most next-generation sequencing platforms does not allow for an unambiguous, confident analysis in these complex regions. In the paper: "Performance of High-Throughput Sequencing for the Discovery of Genetic Variation Across the Complete Size Spectrum" [1] the authors describe how using second generation high throughput sequencers only captures a fraction of known variations. They show how difficult it is to detect deletions, insertions and copy number variations (CNVs) in the 100bp to 10kb size range and repetitive DNA sequences. Furthermore, in the paper "Resolving the complexity of the human genome using single-molecule sequencing" [2] the authors sequenced the hydatidiform mole cell line (CHM1) using PacBio, a third generation long-read sequencing technology, and compared it to the human GRCh37 reference.

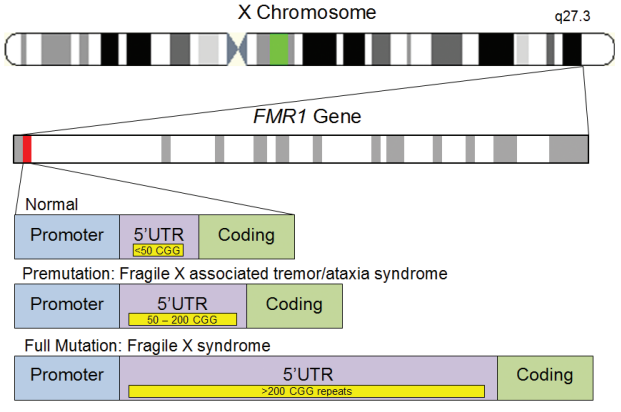

Figure 1. ***Structure of the FMR1 gene.***
For normal individuals the CGG repeats in the Fragile X mental retardation gene 1 (FMR1) are between 5 and 50. In the premutation range the repeats are between 50 and 200, in this range the mRNA expression levels are elevated but the Fragile X mental retardation protein (FMRP) levels decrease. There is an increased risk Fragile X-associated tremor/ataxia syndrome (FXTAS). Above 200 repeats transcription is silenced due to hypermethylation. The absence of FMRP results in Fragile X syndrome.

Using multi-kilobase reads, they were able to close or shrink half of the remaining euchromatic gaps, most of which contained GC-rich sequences composed of short tandem repeats (STR) in the kilobase range. Furthermore, 85% of all CNVs, 92% of all insertions and 69% of all deletions found in this study were novel and only a fraction could be detected with short-read technologies. Finally, the authors found a significant number of highly complex structural variants containing repeats of which only 1% are included in current genome assemblies generated using other sequencing technologies.

The PacBio technology can be applied towards the study of genomic structural rearrangements. For example, in the paper "Sequencing the unsequenceable: Expanded CGG-repeat alleles of the Fragile X gene" [3] the authors demonstrate how PacBio can be used to sequence expanded CGG repeats in the 5' untranslated region (5'UTR) of the FMR1 gene while still covering the full mutation repeat range. The ability to sequence different expansion lengths is important since differing lengths may appear in clinically relevant disorders such as the Fragile X syndrome (Figure 1). Beyond the sequencing of the short repeats (34-36 CGG repeats) and mid-size repeats (95 CGG repeats) the authors successfully amplify and sequence repeats within the full-mutation allelic distributions (~750 CGG repeats).

Targeting methods suited for long-read sequencing

The PacBio technology is a long-read, single molecule, real-time sequencing method. Each molecule loaded into the system is sequenced independently using an immobilized polymerase incorporating fluorescent bases in real-time. Each time a single base is incorporated the signal is recorded by one of the four cameras corresponding to each base and color. First and second generation sequencing technologies are hindered by phasing error where accurate base calls always require that a multitude of identical (amplified) DNA fragments be sequenced in synchrony. This limits the read length at which a base can be confidently called. Among other next-generation sequencers, PacBio stands alone with an average raw read length of 10kb (Figure 2).

To target specific regions in a large genome, one could use a capture approach [5] or a PCR amplification. When PCR is used, special care must be taken in the choice of the polymerase and conditions to obtain full-length multi-kilobase products and avoid artifacts (high fidelity polymerases and elongation time). Repeated structures can also be difficult to amplify due to the formation of secondary structures which can hinder amplification or even result in fragments

McGill
gqinnovationcenter.com

McGill University and Génome Québec Innovation Centre
740 Docteur-Penfield Ave., Montréal, QC Canada H3A 0G1
T 514 398-7211 • infoservices@genomequebec.com

GenomeQuébec
GenomeCanada

High-fidelity sequencing of complex genomic regions and full-length transcripts

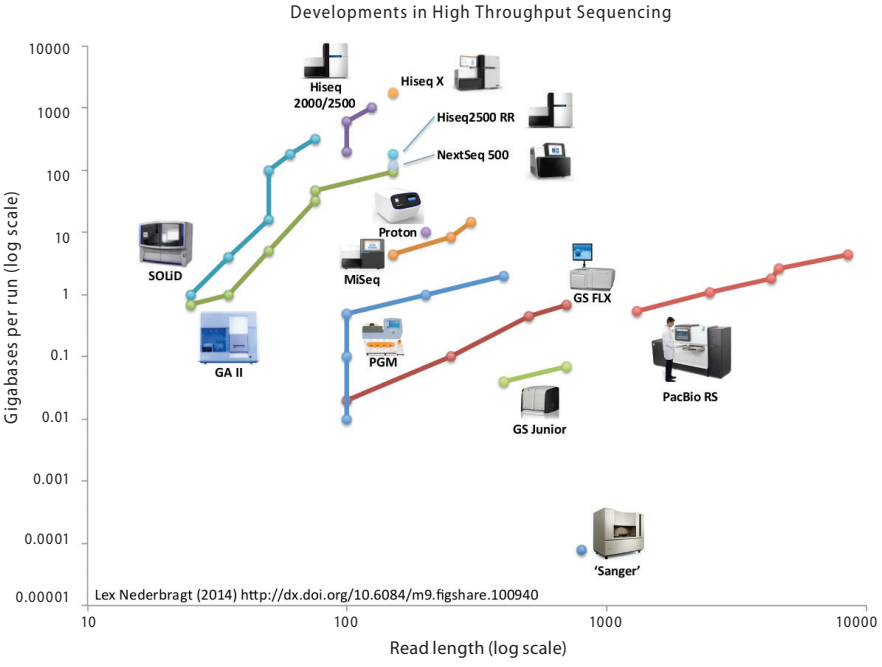

Figure 2. ***Comparison of average read length produced by next-generation sequencers over time.***
On the graph, the rightmost point for PacBio corresponds to the average read length versus Gigabase for a SMRTcell using P5-C3 chemistry run on the PacBio RSII system. We are currently using the PacBio P6-C4 chemistry which has a comparable read length distribution to the P5-C3 chemistry with added accuracy [4].

not representative of the genome [6,7]. In order to sequence full-length RNAs, a cDNA protocol that retains the integrity of the transcript should be used (example: Clontech - *SMARTerTM PCR cDNA Synthesis Kit or Invitrogen - GeneRacerTMRLM RACE kit).* Protocols that generate cDNA using random priming are not appropriate since the resulting cDNA is not full length. In the event that multiplexing is necessary, barcoding of the different samples can be performed independently and prior to the PacBio library stage. Typically, we introduce a 16-base barcode on both ends of the fragment to be amplified.

Application and Results

RNA Sequencing

Using short-read sequencing technologies to identify and quantify the different transcripts of a gene is difficult since the read length is typically shorter than the mRNA molecule and cannot sequence across complex splicing events. Using multi-kilobase reads allows us to sequence full-length mRNA molecules and directly capture different patterns of alternative tanscriptional events. Likewise, patterns of alternative tracriptional start sites and/or polyadenylation sites may also be layered onto the alternative splicing patterns, thereby providing a much more complete inventory of the transcript isoform content. The subsequent analysis is not only greatly simplified but does not rely on averages and removes the requirement for assumptions about isoforms.

In a collaborative study, we amplified and sequenced DSCAM transcripts in *Drosophila* (unpublished). The DSCAM gene very precisely controls neuronal branching patterns in the developing *Drosophila* by expressing different cell-surface receptors. DSCAM is a member of the immunoglobulin (Ig) superfamily and contains ten Ig, six fibronectin domains and a single transmembrane and cytoplasmic domain. The DSCAM gene produces a mRNA of 7.8kb containing 24 exons including 4 with variant exons that are mutually exclusive which can give rise to a total of 38,016 possible isoforms (Figure 3). Which isoforms are expressed depends on tissue type and larval stage. *Wei Sun* et al. [8] present a method for resolving the pairing of the different variants with a short-read sequencing strategy (Illumina). To achieve this, the cDNA must be amplified, circularized and amplified again to regroup exons 4, 6 and 9 in proximity of each other within a 1kb window. The fragments are then formed into clusters on the flowcell and multiple sequencing primers are used to sequence into exons 4, 6 and 9 as well as the barcodes. PacBio allowed us to do the same but with only the PCR step and the ligation of the SMRTbells. After cDNA generation and amplification we were able to sequence 2kb portions of the transcripts covering exons 4, 6 and 9 and found 1,649 unique isoforms. We also confirmed the three exon distributions with a control experiment where the exons were sequenced independently (data not shown).

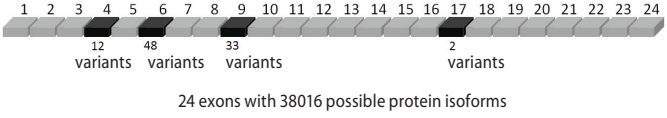

Figure 3. ***Structure of the DSCAM gene.***
In the genome, 115 exons are present but only 24 are spliced and expressed. Exons 4, 6, 9 and 17 have 12, 48, 33, and 2 alternate forms, respectively.

High-fidelity sequencing of complex genomic regions and full-length transcripts

Repeat expansion and fusion genes

In another project we used our long reads to validate a gene fusion involved in a deadly pediatric brain tumour (Embryonal tumour with multilayered rosettes – ETMR [9]). In this cancer an overamplification of the C19MC microRNA cluster is driven by the fusion to TTYH1. We validated the gene fusion by PacBio RNA-sequencing of the transcripts after targeted amplification. We performed long-range PCR-targeted amplification after reverse transcription of total RNA. The forward primers targeted the exons adjacent to the fusion point and the reverse primers targeted the fused sequence including microRNA clusters.

The different primer combinations resulted in multiple PCR products in the 2-6kb range which is in agreement with the PCR design. Using individual raw reads, we constructed a consensus sequence with 98.4% identity with the reference using PacBio's ConsensusTool.sh command with the AmpliconAnalysis option (smrtpipes v2.3 patch1). The consensus sequence generated from 20X PacBio reads shows a perfect match to the reference genome revealing the fusion gene structure in which the first 4 exons and part of the intron of TTYH1 are fused to the X microRNA cluster (Figure 4).

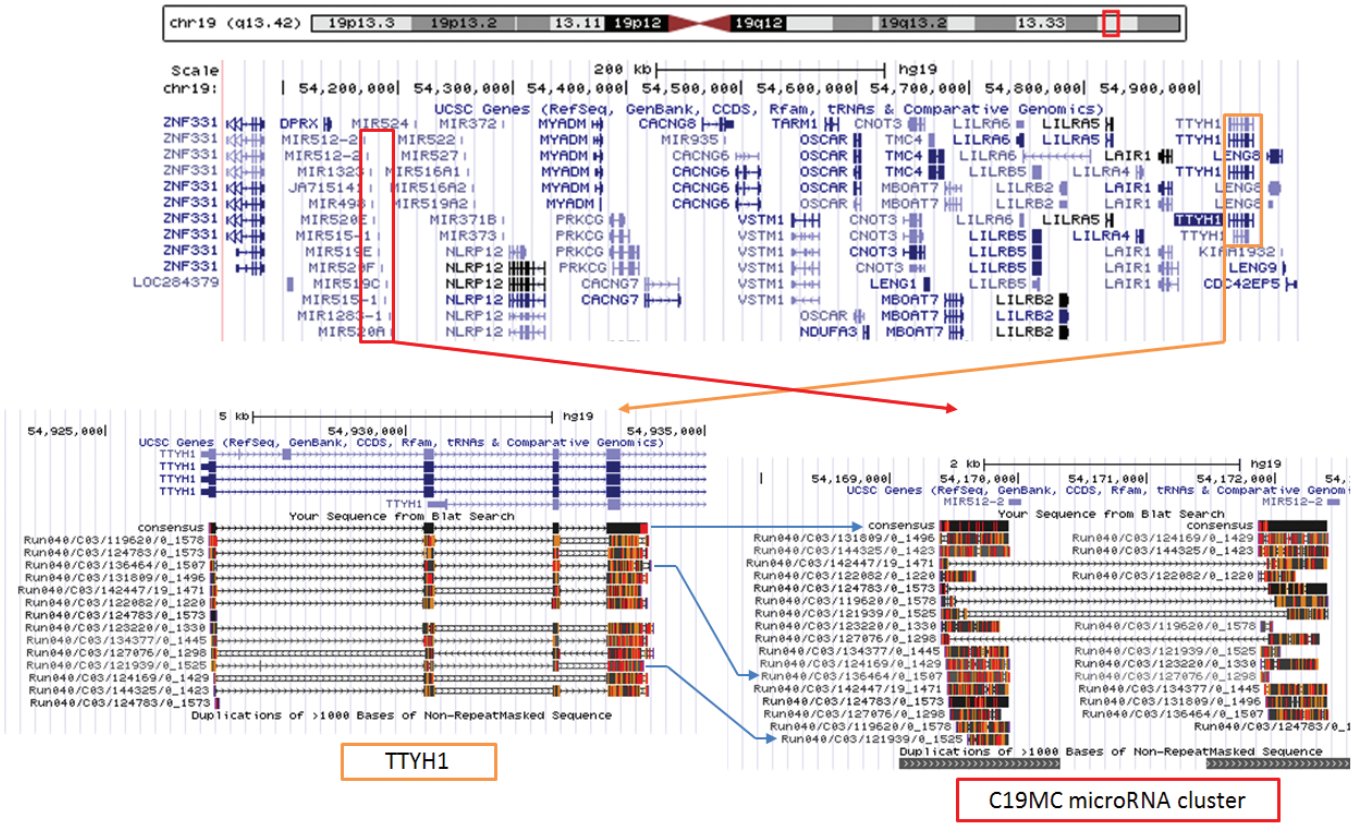

Figure 4. Structure of the gene fusion between the TTYH1 gene (chr19:54,926,605-54,947,899) and the C19MC microRNA cluster chr19:54,167,483-54,267,895).

These two loci are normally spaced by 700kb in the genome. We aligned both raw reads and a reconstructed consensus sequence from our full-length sequenced transcripts to a human genome reference using blat and the UCSC browser Human Feb. 2009 (GRCh37/hg19). Note that when the alignment is colored in black it indicates that the mapping is perfect between the reads and the reference. The amplified fusion transcripts first map to four exons from the TTYH1 gene as well as the intronic sequence following the fourth exon (around 7,000 bases). The transcript then maps on the C19MC microRNA cluster over two regions upstream of the microRNA, this is expected since the microRNAs are nested inside duplications. Example alignments from the same read on each locus are shown by the blue arrows. We observe that the consensus sequence has a better alignment to the rightmost repeat since shown by an alignement in black with the absence of red lines. The consensus sequence here was constructed using 20X of raw read coverage. Because PacBio errors occur randomly along the reads we are able to filter them out using multiple reads from the same amplification. Typically the accuracy jumps from 85% to high 90s by merging a few raw reads. Starting at 20X we no longer see any insertions or deletions and we have perfect consensus with the reference with only a few differences near the transcribed fusion point. We also observed a single base mismatch in the third exon of TTYH1 but this call is well supported by all the raw reads at that genomic coordinate. This could be a real variant but we must rule out the possibility of PCR artifacts with more replicates.

High-fidelity sequencing of complex genomic regions and full-length transcripts

Conclusion

We showed in this technical application note that PacBio technology data can be used to validate and discover the precise composition and distribution of full-length RNA transcripts and an example of a complex fusion event causing a rare tumor in humans. These studies would have been very difficult if not impossible using any second-generation short-read sequencing approach. Other potential applications of PacBio include phasing of single nucleotide polymorphisms (SNPs) and of small insertions and deletions (indel), short tandem repeat expansions, variable number tandem repeats (VNTR), retro-element insertions such as Line 1 elements, alternative splicing with full-length RNA (cDNA) sequencing, duplications, inversions and fusions. Larger variations such as chromosomal rearrangements can also be detected by performing local re-assemblies with PacBio reads [2]. Longer reads have also been shown to help HLA typing (human leukocyte antigen) genes in the MHC (major histocompatibility complex) region of the human genome. Longer reads bring critical SNP phasing information which is difficult to obtain with other technologies [10,11]. Both targeted design and the generation of PacBio long reads are offered as services at the Innovation Centre.

Acknowledgements

Dr. Jacek Majewski, McGill University
Dr. Nada Jabado, Montreal Children's Hospital
Dr. Brian Chen, Research Institute of the McGill University Health Centre

References

[1] Performance of High-Throughput Sequencing for the Discovery of Genetic Variation Across the Complete Size Spectrum. *Andy Wing Chun Pang et al.* Genes Genomes Genetics (2014)

[2] Resolving the complexity of the human genome using single-molecule sequencing. *Mark J. P. Chaisson et al.* Nature Letter (2014)

[3] Sequencing the unsequenceable: Expanded CGG-repeat alleles of the Fragile X gene *Erick W. Loomis et al.* Genome Research (2013)

[4] Nederbragt, Lex (2012): developments in NGS. figshare. http://dx.doi.org/10.6084/m9.figshare.100940 Retrieved 21:15, Dec 08, 2014 (GMT)

[5] Targeted Sequencing on the PacBioRS Using Agilent Technologies Sure Select Target Enrichment http://pacificbiosciences.com/pdf/Technical_Note_Targeted_Sequencing_PacBio_RS_AgilentTechnologiesSureSelectTarget.pdf Pacific Biosciences

[6] A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR X.Y. Hauge et al. Human Molecular Genetics (1993)

[7] Improving sequencing quality from PCR products containing long mononucleotide repeats *Aron J. Fazekas et al.* Biotechniques (2010)

[8] Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing *Wei Sun et al.* EMBO (2013)

[9] Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR *Claudia L Kleinman et al.* Nature Genetics (2014)

[10] A fault-tolerant method for HLA typing with PacBio data. Chia-Jung Chang et al. BMC Bioinformatics (2014)[11] Allele Level Sequencing and Phasing of Full length HLA Class I and II Genes Using SMRT Sequencing Technology http://www.pacificbiosciences.com/pdf/Poster_AlleleLevelSequencing_PhasingFullllengthHLAClassIandIIGenes.pdf Pacific Biosciences

Client Management Office: 514 398-7211
infoservices@genomequebec.com

Assistant Scientific Director:
Alexandre Montpetit, PhD 514 398-3311
alexandre.montpetit@mail.mcgill.ca ext. 00913

Haïg Djambazian, Claudia Kleinman, Albena Pramatarova, Brian Chen, Tsung-Jung Lin, Geneviève Geneau, Patrick Willett, Pierre Bérubé, Alfredo Staffa, Rob Sladek, Alexandre Montpetit

McGill University and Génome Québec Innovation Centre; Montréal, Québec, Canada.

492-27_GQ17 (03-15)