Ryan Lorig-Roach
6/12/17

NOTCH2NL Independent Gene Assembly Verification in H9 ESCs with Nanopore cDNA

Introduction

NOTCH2NL is a gene family found exclusively in humans, and implicated in several conditions such as autism, schizophrenia, microcephaly, and microcephaly [2][3]. Its expression leads to prolonged proliferation of neural progenitors, and thus is thought to be directly related to brain size [2]. The gene exists in 4 paralogs across the 1p12 and 1q21.1 region, referred to as NOTCH2NL-A, NOTCH2NL-B, NOTCH2NL-C, and NOTCH2NL-D. The paralogs are >=99.2% similar in exon content, in some cases with only several bases of difference between them. For this reason, assembling the NOTCH2NL gene family and phasing their haplotypes has only recently become possible with the advent of 10X Genomics [1]. In order to verify assemblies generated with the 10X Genomics platform, cDNA libraries were generated for the Oxford Nanopore with the same cell line (H9 ESC). Using Nanopore transcript data, the proposed assembly of each paralog variant has been assessed independently with HMM-based sequence alignment and unsupervised clustering.

Methods

cDNA libraries were prepared from H9 ESCs and enriched with an array of 60,000 probes designed by Ian Fiddes to tile the 1p12 and 1q21.1 region in its entirety. The sequences were then basecalled with Metrichor and aligned to a NOTCH2 consensus using a Nanopore optimized algorithm called MarginAlign (developed by Benedict Paten). 10X whole cell genomics data was enriched with the same probe array, and was assembled into paralog variants using a custom assembler developed by Alex Bishara (see Table 1).

Variant Name	Consensus Position (bp)										
	49	53	91	124	150	151	155	164	271	…	1096
NOTCH2-1	T	T	CGGC	G	G	CC	G	G	GCGGCGGAGGA	…	G
NOTCH2NL-A-1	T	C	CGGC	G	G	C	G	G	GCGGCGGCGGA	…	A
NOTCH2NL-A-2	T	C	CGGC	G	G	C	G	G	GCGGCGGCGGA	…	A
NOTCH2NL-B-1	T	C	C	C	G	CC	G	G	GA	…	A
NOTCH2NL-B-2	T	C	C	C	G	CC	G	G	GA	…	A
NOTCH2NL-C-1	T	C	CGGC	G	G	CC	T	C	GCGGCGGCGGCGGCGGAGGA	…	A
NOTCH2NL-C-2	T	C	CGGC	G	G	CC	T	C	GCGGCGGCGGC	…	A
NOTCH2NL-D-1	C	C	CGGC	G	A	CC	G	G	GCGGCGGCGGCGGAGGAGGA	…	A

Table 1: Abridged feature table for paralog assemblies. Each NOTCH2NL paralog and its observed variants were generated with 10X Genomics data, with the ancestral NOTCH2 paralog included for completeness' sake. At each variable position, the observed feature for each paralog is shown. Note that the ORF begins at consensus position 307bp. NOTCH2 and NOTCH-D were determined to be homozygous in H9, while all other paralogs were heterozygous.

The Nanopore reads were then reduced into feature vectors containing all 33 variant sites along the first 1100bp of all transcripts. These transcripts were aligned using a Markov model with one path for each of the unverified paralog assemblies (see Figure 1). Since the

transcripts are already aligned to a consensus, there is no need for reverse transitions in the model, and since variation or recombination between paralogs is already accounted for in the assemblies, no transitions between paths are allowed. This vastly simplifies the Forward algorithm, and the maximum probability path (usually determined with the Viterbi Algorithm) is trivial to calculate under these conditions. All mismatches were assumed to be errors, and were given an emission probability of 0.1 to approximate the error rate of the Nanopore. The success of the assembly is evaluated by this aligner depending on whether there are non-erroneous transcripts that do not align well to any path.

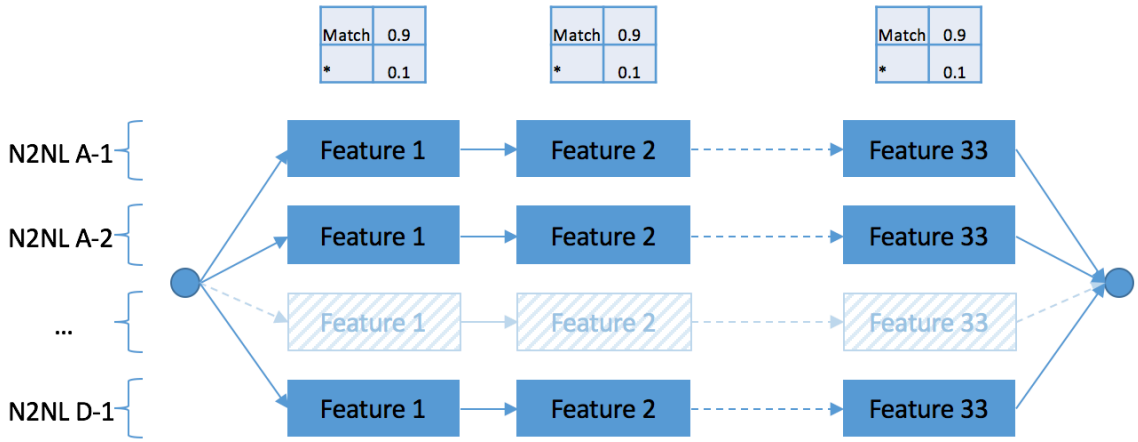

Figure 1: Model diagram for the simplified HMM aligner. Each state consists of a feature obtained from the preliminary assembly, and each path corresponds to the set of features obtained from a single paralog variant. Failed reads that contain no aligned features (missing data) will match equally well to all 8 paths, resulting in a match probability of $1/8^{th}$ or 0.125, which is the lowest possible score.

Since this alignment method is not entirely independent from the assembly method, KMeans clustering was also performed on the Nanopore transcripts. The clusters obtained from this step were then aligned through the Markov aligner to see whether their transcripts tended to fall into a single paralog or not. In this case, the proposed assembly would be validated if each variant had its own non-overlapping cluster.

Results and Discussion

Initial results from the alignment of transcripts show notable variation in the expression of each paralog (Table 2), and in particular the B-2 paralog has a surprising lack of transcripts. However, upon further inspection this was determined to be the result of an in-frame stop which leads to nonsense-mediated decay. The ancestral NOTCH2 and NOTCH2NL-D paralogs were observed in very low frequency, which is partially explained because D is thought to be a pseudogene. Finally, the C paralog transcripts lack unique alignments, but this is most likely because the only differing feature between C-1 and C-2 was a CGG repetitive region (see Table 1: 271bp), which is poorly characterized by the Nanopore. The deficiency in total alignments is resolved in Figure 1B and 1C, where all alignments are shown regardless of whether they are

Variant	Matches (unique)
NOTCH2-1	7
NOTCHNL-A-1	197
NOTCHNL-A-2	201
NOTCHNL-B-1	529
NOTCHNL-B-2	40
NOTCHNL-C-1	18
NOTCHNL-C-2	0
NOTCHNL-D-1	46
Total	1038

Table 2: Raw read counts for NOTCH2NL variants. The number of unique alignments for each path in the model are shown. 1038 of the original 1483 reads were aligned uniquely to one variant.

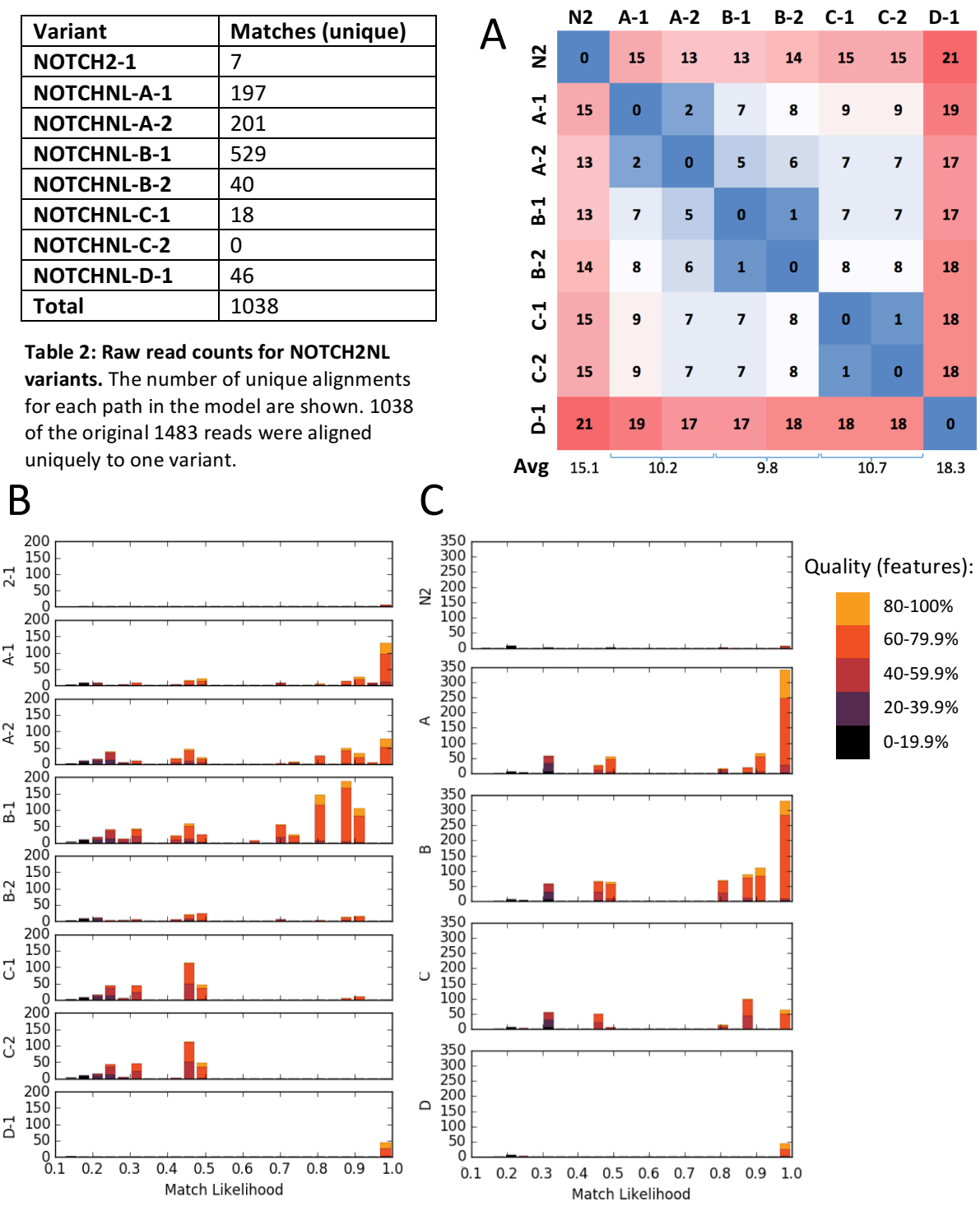

Figure 1: (A) Difference matrix for assembled paralog features and (B,C) Score distributions for aligned H9 transcripts. Shown in **Panel A** is the number of differing features between variants. Binned histograms are shown in **Panels B and C** with the probability of the most likely path through the Markov model determining the bin for each transcript. Transcripts with multiple equal alignments were added to all bins with the maximum probability score. "Quality" refers to the number of non-missing features present in each transcript's feature vector. **Panel B** shows the alignments for all variants, while **Panel C** shows alignments for paralogs after collapsing paralogous paths together in the model. The ancestral NOTCH2 paralog is referred to as "2-1" or "N2".

unique or not. The enrichment array may also bias transcript quantities, so these results should not necessarily be taken as a direct measurement of expression.

Collapsing the paralogous paths in the Markov model resulted in a shift of match probabilities within each paralog towards 1.0 (Figure 1C), indicating that most of the poorly aligned reads were ambiguous within their paralog, and not between different paralogs. This is likely because of the minimal difference that exists between sequences of the same paralog, sometimes amounting to only 1 SNP. However, about 50 feature vectors still failed to align uniquely, despite having at least 70% of their features present (not lost to sequencing error). These strange transcripts mapped to both N2NL-A and B with roughly 50% likelihood, as seen in the score distributions in Figure 1C.

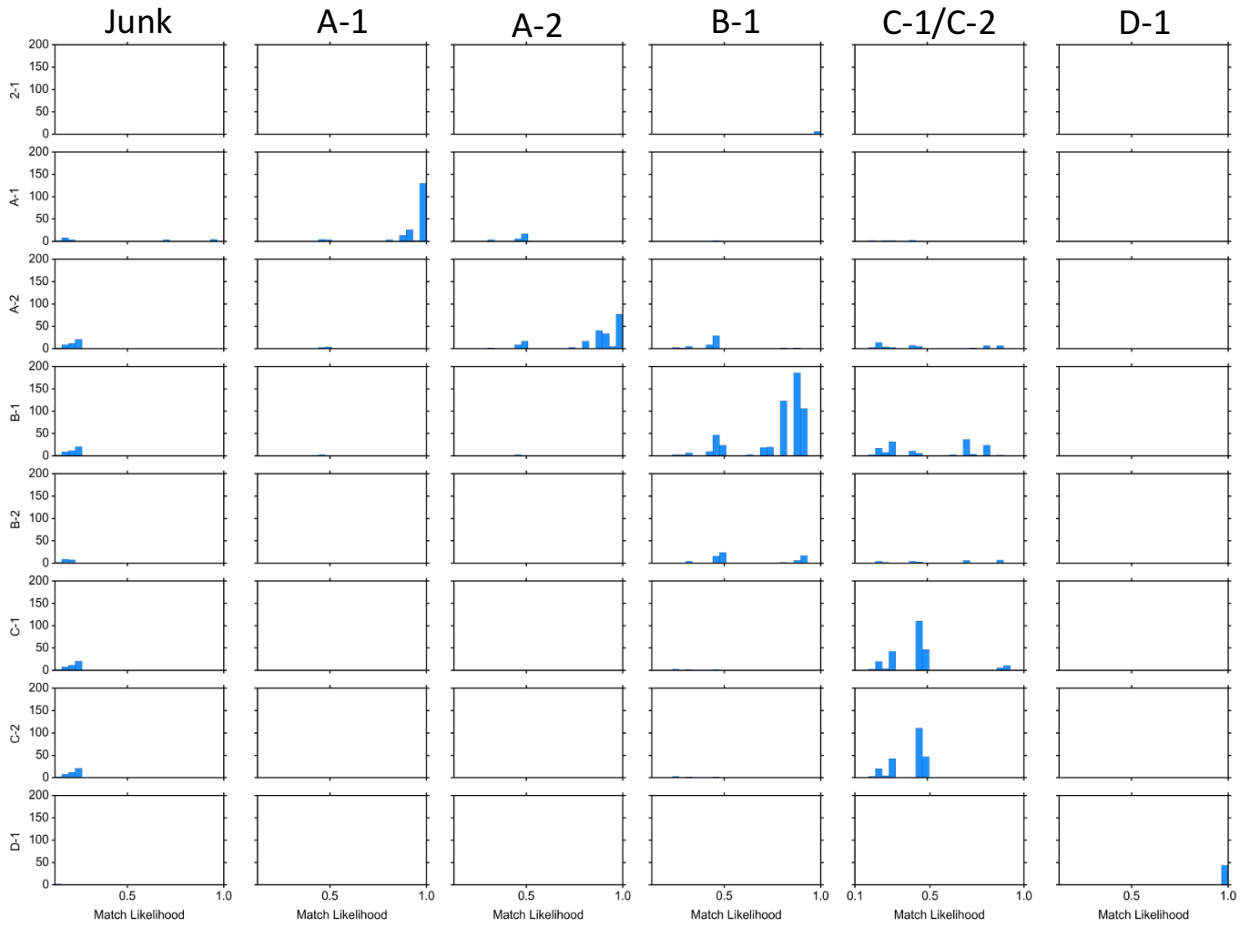

Figure 2: KMeans clustering of transcript feature vectors. Feature vectors for all Nanopore reads were binarized and clustered using the Python SKLearn KMeans algorithm. The transcripts of each cluster were then aligned through the Markov model to determine whether paralog variants were organized into separate clusters by the unsupervised algorithm. Each **column** represents a cluster and each **row** shows the score distribution for a given paralog variant. Initial clustering with K=8,7 yielded duplicate clusters, prompting reduction of the K parameter to 6. Note the overall lack of overlap between paralogs within each column.

To assess whether the poorly mapping sequences belonged to a true unknown transcript, their original fastq sequences were collected and assembled with two assemblers:

Canu and Geneious. Both assemblies failed, so it is presumed that while these reads retained most of their features during basecalling and consensus alignment, these features are erroneous. The reads were also aligned to hg38 with BLAT, which yielded poor quality alignments to all annotated NOTCH2NL paralogs, with many intronic partial alignments.

Despite minimal differing features, unsupervised clustering was still able to generate groups that were unique for most of the variants, with the exception of the poorly transcribed NOTCH2 and NOTCH2NL-B-2 genes (Figure 2). Again, the NOTCH2NL-C variants were indistinguishable because of their lack of sequenceable features. The cluster configuration shown in Figure 2 was not the only result for K=6, but it appeared to be the most stable. Other configurations showed duplicate clusters for N2NL-A variants.

Conclusion

With at least 50 reads aligning with >90% probability to most paralog variants, and hundreds of reads aligning to each paralog after collapsing their variants, we can be reasonably confident that no errors exist in the assembly. Unfortunately, it appears that individual C variants will not be discernable with Nanopore technology unless they experience more divergence, or if homopolymer characterization improves. Confirmation by unsupervised clustering improves the independence of this assessment.

Acknowledgements

Ian Fiddes
Benedict Paten
Colleen Bosworth
Alex Bishara

SciKit Learn Cluster library
PySam library
VCF python library
Canu assembler
Geneious assembler

References

1. Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January 16, 2017): 14049. doi:10.1038/ncomms14049.
2. Jacobs et al (unpublished)
3. Stefansson, Hreinn, Dan Rujescu, Sven Cichon, Olli P. H. Pietiläinen, Andres Ingason, Stacy Steinberg, Ragnheidur Fossdal, et al. “Large Recurrent Microdeletions Associated with Schizophrenia.” *Nature* 455, no. 7210 (September 11, 2008): 232–36. doi:10.1038/nature07229.