Solving mendelian mysteries: the non-coding genome may hold the key

Enza Maria Valente, MD, PhD[1,2] and Kailash P. Bhatia, MD[3]

[1]Department of Molecular Medicine, University of Pavia, Pavia, Italy; [2]Neurogenetics Unit, IRCCS Santa Lucia Foundation, Rome, Italy; [3]Sobell Department, Institute of Neurology, University College of London, London, UK.

Summary: Despite revolutionary advances in sequencing approaches, many mendelian disorders have remained unexplained. In this issue of Cell, Aneichyk and collaborators combine genomic and cell-type specific transcriptomic data to causally link a non-coding mutation in the ubiquitous *TAF1* gene to X-linked dystonia-parkinsonism.

Over the past decade, the advent of next generation sequencing has truly revolutionized the discovery of genes causative of mendelian disorders. This previously long and painstaking process has been incredibly accelerated by this cheap, fast and easy technology, allowing to simultaneously interrogate the coding sequence of virtually all our genes. As a result, the number of genes associated to rare mendelian disorders has raised at a an exponential pace in recent years, leading to expectations that most, if not all mendelian disorders would soon be explained (Gilissen et al., 2012). Yet, after the excitement of the first times, it has become clear that the proportion of unsolved exomes was much higher than expected, and that a significant subset of monogenic disorders may relate to genetic defects lying outside the ~1.5% of our genome represented by coding sequences. Research focus is now progressively switching to the vaster, and more enigmatic non-coding regions of the genome. Variation in non-coding DNA, including single nucleotide variants, transposable elements and structural variations, hold great promises in terms of future discoveries, despite the current difficulties that relate to our largely incomplete knowledge of the functional relevance of most non-coding variants (Smedley et al., 2016). This limitation is being progressively overcome by the adoption of integrated strategies for genomic sequencing, data assembly and transcriptomic analysis. Indeed, several cancers and also a few mendelian disorders have already been linked to a dysregulated transcription process mediated at least in part by the presence of specific non-coding somatic or germinal variants (Kremer et al., 2017; Shiraishi et al., 2014). In this issue of Cell, Aneichyk and collaborators provide yet another example of a monogenic disorder being explained by a deeper understanding of the genomic sequences lying in-between exons, and of their essential role as transcriptional regulators.

The disease in question is X-linked dystonia-parkinsonism (XDP), a late-onset, progressive neurodegenerative disease endemic to the island of Panay, in the Philippines. In the past years, conventional linkage analysis allowed mapping a 449 Kb haplotype shared by all affected individuals. This common haplotype consisted of seven non-coding variants falling either within introns or in the 3' untranslated region of the *TAF1* gene, which encodes the TATA-Binding Protein (TBP)-Associated Factor-1, a protein involved in regulation of transcription by RNA polymerase II. Among the seven non-coding variants were five single nucleotide variants, a 48-bp deletion and a SINE-VNTR-Alu (SVA)-type retrotransposon insertion, none of which had any annotated function. Thus, even though a dysregulation of *TAF1* was strongly suspected in the pathogenesis of XDP, this hypothesis remained long unproven.

Now, Aneichyk and collaborators put together a very large cohort of affected male individuals, unaffected female haplotype carriers and non-carrier individuals, and employed multiple short- and

long-read sequencing strategies to perform deep sequencing and subsequent unbiased, reference-free assembly of the genomic region spanning the XDP haplotype. Intriguingly, this approach revealed sequences that were unique to the Panay population (e.g. absent in the current human reference assembly), and which included up to 47 novel variants segregating with the disease and allowed the authors to refine the minimum critical region to 203.6 Kb, which encompassed the *TAF1* gene only.

The next step to pinpoint the causative variant was to establish induced pluripotent stem cells (iPSCs) from several affected individuals, carriers and controls, and to differentiate them into neural stem cells (NSCs) and induced cortical neurons (iNs). In these models, various complementary RNA sequencing approaches led to assemble the complete transcriptional structure of *TAF1*, including 4 novel transcripts, and to evaluate expression changes (both of this specific gene and genome-wide) in patients vs controls. Of note, two transcripts (including the canonical and most abundant *TAF1* transcript) were significantly downregulated in XDP fibroblasts and NSCs compared to controls, but they were not affected in iNs or other differentiated neurons. Similarly the Novel-32i transcript, despite being rare overall, was significantly enriched in XDP lines. All these observations were directly and strongly correlated to an abnormal retention of intron 32 and decreased usage of more distal exons in XDP compared to controls. As a consequence, the levels of TAF1 protein were also reduced in these XDP cell types (Aneichyk et al., 2018).

Retrotransponsons are well known to represent a potential source of variation in the processes which rule transcription and splicing (Elbarbary et al., 2016), and have already been implicated as the culprits for other mendelian diseases and cancer (Kazazian and Moran, 2017). Thus, the presence of a SVA insertion within intron 32 of the *TAF1* gene made it an excellent candidate which could potentially interfere with correct RNA transcription, leading to aberrant transcripts and reduced *TAF1* expression. Of note, in a parallel study, authors from the same group have recently dissected the complete sequence of this SVA, and identified a polymorphic hexanucleotide repeat domain whose size showed a highly significant inverse correlation with age at onset of XDP, and which influenced the ability of the SVA to regulate transcription in a luciferase reporter assay (Bragg et al., 2017). Further corroborating the causal role of the SVA in XDP pathogenesis, removal of this retrotransposon insertion through CRISPR/Cas9 technology was able to fully rescue the normal transcriptional signature of *TAF1*, restoring physiological *TAF1* expression levels (Aneichyk et al., 2018).

This study bears several important implications. First, it provides robust evidence that combined genomic and transcriptomic strategies can succeed in disclosing the genetic defect underlying the many mendelian disorders in which whole-exome or even whole-genome sequencing alone have failed. Secondly, the observation that aberrant *TAF1* splicing and increased intron 32 retention were observed only in XDP-derived dividing NSCs but not mature neurons underlines the importance of choosing the correct cell models when performing transcriptional studies. In fact, there is mounting evidence that non-coding variants may be implicated in fine regulation of transcription and splicing, which may have a different impact on diverse cell types and in distinct temporal frames, such as embryogenesis or adult life (Andrey and Mundlos, 2017). Third, it is important to note that *TAF1* expression was only moderately impaired in patients' cells, suggesting that even subtle transcriptional alterations may eventually produce clinically appreciable deficits in the long term, especially when involving genes that are highly intolerant to variation, such as *TAF1*. Indeed, the observed inverse correlation between the polymorphic repeat sequence within the SVA and both *TAF1* expression levels and the disease onset elegantly explains at least part of the large clinical variability seen in XDP patients, suggesting that XDP may share some features of diseases associated with unstable repeat expansions. Interestingly, segregation studies in XDP pedigrees

demonstrated that the SVA repeat underwent both expansions and contractions when inherited (with a possible relation to the sex of the transmitting parent), in line with the lack of anticipation observed in XDP families. Finally, the rescue of *TAF1* splicing by CRISPR-Cas9 mediated excision of the SVA provides further proof-of-principle for potential therapeutic developments based on gene-editing techniques, although many obstacles still need to be overcome, such as the need to target specific cellular populations which, in XDP, are still to be clearly identified, and the temporary window in which such interventions may prove effective.

References

Andrey, G., and Mundlos, S. (2017). The three-dimensional genome: regulating gene expression during pluripotency and development. Development *144*, 3646-3658.

Aneichyk, T., Hendriks, W.T., Yadav, R., Shin, D., Gao, D., Vaine, C.A., Collins, R.L., Domingo, A., Currall, B., Stortchevoi, A., *et al.* (2017). Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. Cell xxxxxx

Bragg, D.C., Mangkalaphiban, K., Vaine, C.A., Kulkarni, N.J., Shin, D., Yadav, R., Dhakal, J., Ton, M.L., Cheng, A., Russo, C.T., *et al.* (2017). Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. Proc Natl Acad Sci U. S. A.*114*, E11020-E11028.

Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. Science *351*, aac7247.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur J Hum Genet *20*, 490-497.

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. New Engl J Med *377*, 361-370.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., *et al.* (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun *8*, 15824.

Shiraishi, Y., Fujimoto, A., Furuta, M., Tanaka, H., Chiba, K., Boroevich, K.A., Abe, T., Kawakami, Y., Ueno, M., Gotoh, K., *et al.* (2014). Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers. PloS one *9*, e114263.

Smedley, D., Schubach, M., Jacobsen, J.O.B., Kohler, S., Zemojtel, T., Spielmann, M., Jager, M., Hochheiser, H., Washington, N.L., McMurry, J.A., *et al.* (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet *99*, 595-606.