SUPPLEMENTARY MATERIAL 1

Chimeric transcripts resulting from complex duplications in chromosome Xq28

Human Genetics

Luciana W. Zuccherato[1], Benjamin Alleva[2], Marjorie A. Whiters[1], Claudia M. B. Carvalho[1,3], James R. Lupski[1,4,5]*

[1]Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX, USA.

[2]Biology Department, University of Iowa, Iowa City, IA, USA.

[3]Centro de Pesquisas René Rachou - FIOCRUZ, Belo Horizonte, MG, Brazil.

[4]Pediatrics, Baylor College of Medicine, Houston, TX, USA.

[5]Texas Children's Hospital, Houston, TX, USA.

*Corresponding author:

jlupski@bcm.edu

SUPPLEMENTARY METHODS

Subjects

We analyzed a cohort of 38 unrelated male subjects with *MECP2* duplication syndrome, in which the genomic rearrangements were previously identified by microarray-based comparative genomic hybridization (array CGH) of the *MECP2* region on Xq28 and breakpoint junction mapping at nucleotide resolution was achieved (Carvalho et al. 2009, 2011, 2013). This enabled *in silico* predictions for evidence of potential formation of fusion genes based on the complex genomic content observed in the personal genomes of those patients. The cohort allows for the study of the formation and expression of novel genes generated by alternative repair mechanisms and whether one-generation acquired genomic complexity can contribute to increase the transcript repertoire in humans. The study was performed with the approval of the Baylor College of Medicine Institutional Review Board.

Detection of the fusion gene transcripts

Total RNA was extracted from immortalized lymphoblast cell lines from the patients using the RNeasy Plus Mini kit according to the manufacturer's protocol (Qiagen, Valencia, CA). The cDNAs were synthesized from 1 μg of RNA using ProtoScript First Strand cDNA Synthesis Kit (New England Biolabs, Ipswich, MA). The detection of the fusion gene transcripts from the junctions were performed using the primers 3204_TEX_F: GTTCTTCCATCAATGCTTGC and 3204_BCA_R: GTGACGGAAAAGGTGAACCT for the patient BAB3204; 3161_F8_short_F: TGCCCTGATGAGGTGCAAAG, 3161_F8_long_F: AGGGTGCCCGTCAGAAGTTC and 3161_CSAG_R: CTTGCCCTTGTGGTCCTGCT for the patient BAB3161. The following primers were designed to amplify the complete genomic regions of the four variants of gene *BCAP31/TEX28* (3204_BCAP_exon1_1F: ACCCTGTTCTCGCCCCTC; 3204_BCAP_exon1_2F: GAGAGTTCTGTTGCTGCGGC; 3204_BCAP_exon1_3F: GAAGCCCCACCTGGAGGA and 3204_BCAP_exon1_4F: GGCCTCCGGGACGGTGTG; 3204_TEX28_exon5R:

TGTGAGGCAGGTGGGCGTCT) and the variants of the transcript *F8/CSAG1* (3161_F8_exon1_long_F: GCTTAGTGCTGAGCACATCC; 3161_F8_exon1_short_F: GCGTCCCCCTCGGCGGG and 3161_CSA_exon5_R: TCATTTTACAACATGTTCATT). The RT-PCR positive control (*KDMC4*) was amplified using the primers KDM4CF1: CTGCATCCAGTGTTCCTACG and KDM4CR: CCTCAGGAAATGTGTCTCTGC. All the PCRs were performed using HotStar Taq DNA polymerase (Qiagen, Valencia, CA), except for the reaction using the primers 3161_F8_exon1_long_F and 3161_CSA_exon5_R, in which long-range PCR was performed using TaKaRa LA *Taq* (Mountain View, CA). The PCR products were further sequenced by the Sanger method. For the long variant of the *F8/CSAG1*, a primer walking strategy was adopted to cover the full sequence of the transcript using the primers 3161_F8_exon1_ls: TGTGCCTTTTGCGATTCTGC; 3161_F8_ex1_1Fs: GGGAAAAGTTGGCACTCAGA; 3161_F8_ex1_1Rs: CAGCATCCCTATCCTGCATC; 3161_F8_ex9_1Fs: TTCAGCATGAATCAGGAATCT; 3161_F8_ex10_1Rs: TCAGTGATTCCGTGAGGGTA; 3161_F8_ex14_Fs: TGCCACAACTCAGACTTTCG; 3161_F8_ex14_Fs_2: GGACAACTGCAGCAACAGAG; 3161_F8_ex14_Fs_3: CCAGATGCACAAAATCCAGA; 3161_F8_ex14_Rs: TGCTGCTGGAAGATGAGAAG; 3161_F8_ex14_Fs_4: AGCAGCAATAAATGAGGGACA; 3161_F8_ex20_Fs: GGAATGGCTTCTGGACACAT.

***In Silico* Analysis**

The cDNA sequences were analyzed using the UCSC Genome browser (assembly hg19), dbSNP (build 137) and Refseq databases. Protein translation and prediction of the protein size were performed in the Sequencher software (Ann Arbor, MI), and the *InterPro* protein sequence analysis tool (http://www.ebi.ac.uk/interpro) was used to predict the domains of the observed transcripts. Additionally, the putative conserved domains and protein similarities for the inserted LTR sequence were investigated using the NCBI blastp tool (http:// blast.ncbi.nlm.nih.gov/Blast.cgi) and the protein query on UCSC Blat Search Genome tool (assembly hg19; https://genome.ucsc.edu/cgi-bin/hgBlat?command=start).

REFERENCES

Carvalho CM, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ, Ramocki MB, Nawara M, Obersztyn E, Vianna-Morgante AM, Stankiewicz P, Zoghbi HY, Cheung SW, Lupski JR (2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. Hum Mol Genet 18: 2188-203

Carvalho CM, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, Belmont JW, Hastings PJ, Lupski JR (2013) Replicative mechanisms for CNV formation are error prone. Nat Genet 45: 1319-26

Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH, Friehling L, Lee S, Smith R, Del Gaudio D, Withers M, Liu P, Cheung SW, Belmont JW, Zoghbi HY, Hastings PJ, Lupski JR (2011) Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet 43: 1074-81