Title: Variation in novel exons (RACEfrags) of the MECP2 gene in Rett syndrome patients and controls.

Authors: Makrythanasis P, Kapranov P, Bartoloni L, Reymond A, Deutsch S, Guigó R, Denoeud F, Drenkow J, Rossier C, Ariani F, Capra V, Excoffier L, Renieri A, Gingeras TR, Antonarakis SE

Journal: Human mutation

Year: 2009 Sep

Volume: 30

Issue: 9

Pages: E866-79

DOI: 10.1002/humu.21073

Published in final edited form as:
Hum Mutat. 2009 September ; 30(9): E866–E879. doi:10.1002/humu.21073.

Variation in Novel Exons (RACEfrags) of the *MECP2* Gene in Rett Syndrome Patients and Controls

Periklis Makrythanasis[1], **Philipp Kapranov**[2,10], **Lucia Bartoloni**[1], **Alexandre Reymond**[3], **Samuel Deutsch**[1], **Roderic Guigó**[4,5], **France Denoeud**[4], **Jorg Drenkow**[2], **Colette Rossier**[1], **Francesca Ariani**[6], **Valeria Capra**[7], **Laurent Excoffier**[8], **Alessandra Renieri**[6], **Thomas R Gingeras**[2,9], and **Stylianos E Antonarakis**[1,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel Servet, CH-1211 Geneva, Switzerland [2]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA [3]Center for Integrative Genomics, University of Lausanne, Genopode Building, 1015 Lausanne, Switzerland [4]Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra, Dr. Aiguader 88, 08003 Barcelona, Spain [5]Center for Genomic Regulation, Dr. Aiguader 88, 08003 Barcelona, Spain [6]Medical Genetics, Molecular Biology Department, University of Siena, Viale Bracci 2, 53100 Siena, Italy [7]Unitá Operativa Neurochirurgia, Istituto G. Gaslini, 16148 Genova, Italy [8]CMPG, Zoological Institute, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland

Abstract

The study of transcription using genomic tiling arrays has lead to the identification of numerous additional exons. One example is the *MECP2* gene on the X chromosome; using 5'RACE and RT-PCR in human tissues and cell lines, we have found more than 70 novel exons (RACEfrags) connecting to at least one annotated exon.. We sequenced all *MECP2*-connected exons and flanking sequences in 3 groups: 46 patients with the Rett syndrome and without mutations in the currently annotated exons of the *MECP2* and *CDKL5* genes; 32 patients with the Rett syndrome and identified mutations in the *MECP2* gene; 100 control individuals from the same geoethnic group. Approximately 13kb were sequenced per sample, (2.4Mb of DNA resequencing). A total of 75 individuals had novel rare variants (mostly private variants) but no statistically significant difference was found among the 3 groups. These results suggest that variants in the newly discovered exons may not contribute to Rett syndrome. Interestingly however, there are about twice more variants in the novel exons than in the flanking sequences (44 vs. 21 for approximately 1.3 Mb sequenced for each class of sequences, p = 0.0025). Thus the evolutionary forces that shape these novel exons may be different than those of neighboring sequences.

Keywords

MECP2; Rett syndrome; RACEfrags; SNP; rare variants; positive selection

*Correspondence to Stylianos E Antonarakis, Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel Servet, CH-1211 Geneva, Switzerland. Stylianos.Antonarakis@unige.ch.
[9]Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 11724, USA;
[10]Present address: Helicos BioSciences, One Kendall Sq., Cambridge, MA, 02139, USA
Communicated by Garry R. Cutting

INTRODUCTION

The elucidation of the genome function is one of the most important challenges in biomedical research. The ENCODE pilot project (Birney, et al., 2007) provided initial data for the functional analysis of a selected 1% of the human genome. Surprisingly, the ENCODE study has revealed the existence of many additional exons connected to the currently annotated genes (Denoeud, et al., 2007). 5' RACE (Rapid Amplification of cDNA Ends) experiments in many different tissues, followed by hybridization on tiling repeat-masked arrays, have shown numerous "exons" (referred to as RACEfrags), not previously described, of unknown function, extending further beyond the known 5' boundaries of genes. Subsequent RT-PCR and sequencing experiments of the amplicons have verified the RACEfrags (Djebali, et al., 2008).

RACEfrags have been reported to be expressed at a similar level to the annotated coding exons, and many are found in a single tissue or cell line (Denoeud, et al., 2007). Remarkably they are not very well conserved throughout mammalian evolution (Birney, et al., 2007; Denoeud, et al., 2007; Margulies, et al., 2007). The biological importance of these new "exons" is unclear. In order to provide functional evidence we sought to search for pathogenic mutations in RACEfrags of genes that are associated with known genetic disorders.

We selected the MECP2 (MIM# 300005) gene on the Xq chromosome as the case study. MECP2 has two isoforms (Kriaucionis and Bird, 2004; Mnatzakanian, et al., 2004) and pathogenic mutations are responsible for Rett syndrome (MIM# 312750) (Amir, et al., 1999; Wan, et al., 1999), a frequent monogenic disorder with severe mental retardation in girls (Bienvenu, et al., 2006). The patients are typically born normal but their development is abruptly stopped before the second year of life, resulting in a loss of the acquired abilities and characteristic physical and behavioral findings (Percy and Lane, 2004; Segawa and Nomura, 2005; Weaving, et al., 2005). In the classic cases of Rett syndrome, mutations in the MECP2 gene are found in 90–95% of patients; however in the atypical forms (Congenital, Forme Fruste: milder form, PSV: Preserved Speech Variant, Rett-like: patients not fulfilling the diagnostic criteria but with clinical similarities, and early seizures) this rate drops to less than 45% (Mari, et al., 2005; Scala, et al., 2005; Weaving, et al., 2005). Since a considerable number of novel RACEfrags have been linked with the annotated exons of the MECP2 gene, we hypothesized that variants within the newly described MECP2 RACEfrags may be causally related to cases of Rett syndrome without known mutation in the MECP2, CDKL5 (MIM# 300203) or FOXG1 (MIM# 164874) genes (the latter genes have been associated with a minority of Rett syndrome cases (Mencarelli, et al., 2009; Tao, et al., 2004; Weaving, et al., 2004)). Patients' samples were provided from the Italian Rett Bank (Sampieri, et al., 2007).

Our results show that RACEfrag variants are found in similar numbers in the samples from Rett syndrome patients and the matched controls suggesting that nucleotide variants in the newly discovered exons studied do not contribute to the pathogenesis of Rett syndrome. Interestingly however, the variants in the novel exons are twice as frequent as those found in flanking sequences. The significance of this result remains to be elucidated, implying nonetheless that the evolutionary forces that shape these novel exons may be different than those of neighboring sequences.

MATERIALS AND METHODS

Samples

DNA samples from a total of 46 female patients with the clinical diagnosis of Rett syndrome or one of its variants with no known mutations in the MECP2 (RefSeq NM_004992.3), CDLK5 or FOXG1 genes, (referred to as group "PNM") were collected. In addition, samples from 32 female Rett syndrome patients with known mutations in the MECP2 gene were studied (referred to as group "PM"). The patients of groups "PNM" and "PM" were provided by the Italian Rett Bank (Sampieri, et al., 2007) and the frequencies of the clinical categories of the Rett syndrome for each group can be found in Figures 1A and 1B. For 28 of the patients from group "PM" detailed clinical scoring exists based on 10 or 24 clinical parameters (Supp. Table S1). As expected there is a statistical difference in the phenotypic scoring of patients with classic Rett and those with the PSV variant, ($p= 0.0018$ for 10 phenotypic variables or $p= 0.0008$ for 24 phenotypic variables). There was no difference between the patients with classic Rett and those with FF or Rett-like variants (likely because of the small sample size).

The control group, (referred to as "CNT"), consisted of 100 unrelated, randomly selected females from the Gaslini Children's Hospital blood bank in Genova, Italy. The control subjects were Italian with ancestors from all parts of the country, and none had a first-degree relative with Rett syndrome. The samples from all control subjects were anonymous.

Identification of RACEfrags

The coordinates of the RACEfrags connected with the MECP2 gene that are studied here were taken from our recently published studies (Denoeud, et al., 2007; Djebali, et al., 2008). Briefly, using an oligonucleotide primer on the 3rd exon of MECP2, 5'RACE reactions were performed in cDNA samples and subsequently the products were hybridized against repeat-masked tiling arrays of the ENCODE (Birney, et al., 2007) regions consisting of 25-mers with 4bp overlap. Data from 13 tissues (brain, placenta, colon, small intestine, heart, spleen, kidney, stomach, lung, testis, muscle, liver, fetal and adult brain) were used (Denoeud, et al., 2007; Djebali, et al., 2008).

Verification of RACEfrags and connectivity

Ten MECP2-related RACEfrags out of the 71 identified were randomly selected for validation and identification of the exact splice sites. cDNA from the original samples used for identification of the RACEfrags (Denoeud, et al., 2007) were amplified by using the primer of the index exon as forward and a primer on the RACEfrag as reverse. The reverse primers were chosen not to span possible splice junctions. Amplification was done using JumpStart (Sigma) and subsequently the fragments were cloned using the TopoTA kit (Invitrogen), Three clones were picked per plate and sequenced (ABI 3130XL). All the RT-PCRs resulted in at least one identifiable amplicon by sequencing.

In order to estimate the abundance of the RACEfrags in relation to the annotated exons Real-Time PCR (RTPCR) assays were performed using 8ng of fetal brain polyA+ RNA (Ambion) per reaction using one-step qRTPCR kit (ABI). Given variations in primer efficiencies, two primer pairs were selected per variant: one primer pair exactly overlapping the variant, and another within 200 bp of the variant; additionally 3 primer pairs to 3 different exons of MeCP2 were also selected for comparison (Supp. Table S2). A total of 8 RACEfrag were tested.

Detection of nucleotide variants

All RACEfrags studied were sequenced in all 178 individuals by Sanger sequencing. Amplification was done using JumpStart (Sigma). In all cases touchdown PCR programs were used. The total number of PCR cycles was forty, during the first ten the annealing temperature was gradually lowered from 65 to 55°C. A final round of amplification lasting 5 minutes was added at the end of the program. The sequence fragments included the RACEfrag with at least 50bp upstream and downstream and when two adjacent RACEfrags were separated with a maximum distance of 500bp all the intermediate fragment was sequenced as well. Twenty-one PCR amplicons were amplified per sample (Figure 2). Primers were designed using mostly Primer3 and are shown in Supp. Table S3.

The total amount of readable sequence for each DNA sample was 13118 bp, of which 7187 nucleotides (54.8%) were from RACEfrags and 5931 nucleotides (45.2%) from the flanking areas. Variant identification was performed by manually reading the sequences using the Staden Package. In order to define whether an identified variant should be considered inside or outside a RACEfrag the hybridization data were used along with the RT-PCR data when these were available. (Figure 3)

Statistical analysis

Fisher exact and Mann-Whitney-Wilcoxon tests were performed using the free statistical software R. For the detection of difference in the number of variants within and outside of RACEfrags, empirical p-value was calculated by performing 30.000 permutations. More specifically the position of the variants was randomly permuted along the whole sequenced area.

We used the Variscan Software (Hutter, et al., 2006; Vilella, et al., 2005) to perform a population genetics analysis to calculate the nucleotide diversity (π or “pi”), the level of nucleotide polymorphism (θ or “theta”) and the Tajima’s D value. For this analysis we concatenated the sequenced DNA fragments to generate 3 sequence sets (i) All sequenced regions (ii) Only regions corresponding to RACEfrags, (iii) Only regions flanking RACEfrags.

RESULTS

A total of 13118 bp of DNA per sample (21 amplicons, 48.1% GC content, total amount of sequence 2.33Mb) were sequenced. Of these, 7187 bp (54,8%) corresponded to the RACEfrags as determined after the RT-PCRs (48.9% GC content). All the novel RACEfrags linked to the annotated MECP2 gene irrespective of the tissue of origin or the detection method are shown in Figure 2. The remaining 5931 bp (45.2%) per sample correspond to areas flanking the RACEfrags.

RT-PCR experiments to verify the connectivity of the RACEfrags showed that in 15 out of 26 cases canonical splice sites (GT/AG) were used. For the remaining splicing events, the donor site was non canonical in 10 cases (sequences identified: TA, GG, CC, CT), and the acceptor was non canonical in 3 cases (sequences identified: GA, CC) (Djebali, et al., 2008). We examined all 6 reading frames for a possible ORF but we noted that the RACEfrags (with the exception of one exonified Alu sequence) do not show evidence for protein-coding capacity (data not shown).

The median abundance of these RNAs was 2.7 fold lower then that of the MeCP2 annotated exons, based on the difference in the median Ct of the MeCP2 exons vs median Ct of all 16 primer pairs in the SYBRgreen based assay (Supp. Table S4).

Analysis of the re-sequencing data allowed the detection of a total of 65 novel sequence variants were identified; Table 1 lists all the sequence variants and includes their genomic position (hg17), the nucleotide change, the number of occurrences, the group in which they were found (patients, controls), and whether they were found within or outside the RACEfrags. All, but one (deletion of 1bp) were nucleotide substitutions; of these changes 50 (78%) were transitions, and 14 were transversions. For the first 5 variants identified in patients without known MECP2/CDKL5 mutations, samples from the parents were sequenced as well. In all cases the variants were either paternally or maternally inherited.

In order to assess whether variants within MECP2 RACEfrags could be involved in Rett syndrome we compared the frequency of variants in the RACEfrags among the 3 groups of the study. Table 2 provides the number of samples with variants (total or within and outside the RACEfrags) per group along with the respective p-values calculated after comparing the two groups of patients (“PNM” and “PM”) with the control group “CNT”. No p-value reached statistical significance. In addition, no statistical difference was noted when comparing the number of variants in patients with classic Rett versus the other phenotypic groups. Thus, no evidence was found for pathogenic mutations causing Rett syndrome in the MECP2 RACEfrags.

Interestingly, however, we observed a difference in the total number of variants in and outside the RACEfrags (Figure 3 shows the definition used in this study concerning the position of a SNP inside or outside a RACEfrag): 44 variants were found inside the RACEfrags, and only 21 were detected outside. The resulting empirical p-value after 30.000 permutations is 0.025 showing that there is a statistically significant difference in the frequency of variants inside versus outside the RACEfrags.

The observation that RACEfrags contained twice as many rare variants than the neighbouring sequences is remarkable. The θ value of all the target DNA segments studied is 8.86×10^{-4}; however the θ value for RACEfrags was 10.56×10^{-4} while that of the flanking regions was only 6.79×10^{-4}.

The mean π value of our data set is 1.64×10^{-4}; this value is 4 times lower than the average value of 6.82×10^{-4} reported from the study of 322 genes resequenced in 23 CEPH individuals (22Kb per sample, 7108kb total) (Bhangale, et al., 2005) (http://pga.gs.washington.edu/summary_stats.html) although significant variation is noted with the values of that study ranging from 0.21 to 27.16×10^{-4}.

By performing an analysis on the whole concatenated sequenced area, we find a value of Tajima’s D statistic, which measures deviations form neutrality of population sationarity equal to −2.38. The values of Tajima’s D computed separately on the concatenated RACEfrag and on non RACEfrag areas, the values are −2.51 and −1.78 respectively showing that the RACEfrags are the main contributors to the large negative value observed in that region.

DISCUSSION

The recent discovery that a substantial part of our genome is likely to be transcribed is challenging established views on the functional fraction of the genome (Cheng, et al., 2005). The availability of genomic tiling arrays provide novel transcription maps that include not only the already annotated genic regions, but also a substantial number of additional regions either inter- or intragenic. RACE and RT-PCR reactions utilizing RNAs from a large number of tissues and cell lines provide a connectivity of the novel transcribed sites with the annotated transcripts (Djebali, et al., 2008). The function of this rich transcription network is unknown and currently under investigation in numerous laboratories.

The function of a genomic element is often identified or inferred by studying the phenotypic consequences of either natural or induced pathogenic mutations (Antonarakis and Beckmann, 2006). We thus hypothesized that if RACEfrags were functionally linked to the annotated connected gene, then mutations in these RACEfrags may cause the same phenotype as their connected gene. Following our work on novel RACEfrags in the pilot ENCODE regions of the human genome (Birney, et al., 2007) we selected the MECP2 gene on Enm006 on the X chromosome as a test case. The annotated version of the MECP2 gene has been linked to several additional RACEfrags, and pathogenic mutations in the annotated exons cause Rett syndrome (MIM# 312750), an X-linked dominant disease (Amir, et al., 1999; Wan, et al., 1999). Since not all patients with the clinical diagnosis of Rett syndrome have pathogenic MECP2 mutations, we reasoned that pathogenic or predisposing mutations could be found in the DNAs of such Rett syndrome patients.

The results of this study suggest that rare variants in the MECP2-linked RACEfrags may not contribute to the Rett syndrome. However, since it is unknown which of the RACEfrags are likely to be important in the expression of the gene in the brain, and which of the variants detected are likely to affect the function of the transcript, a potential effect could be missed if all identified variants are considered.

We observed that RACEfrags harbor approximately twice as many variants than the flanking areas (44 vs 21; p= 0.025) for roughly the same amount of sequence data. The θ value of the whole DNA region sequenced was 8.86×10^{-4}, but RACEfrags alone had $\theta = 10.56 \times 10^{-4}$ versus 6.79×10^{-4} in the flanking regions. Data from literature are shown in Table 3 and indicate that θ varies from 3.6 and 4×10^{-4} in the coding regions to $5–8 \times 10^{-4}$ in the introns. This suggests that the theta of RACEfrags found in this study is an outlier while the flanking intronic regions have a θ value similar to that found in other studies.

Furthermore the θ value of the RACEfrags is similar to that calculated in an intergenic region in chromosome 8q24 (10.2×10^{-4}) (Yeager, et al., 2008), a gene desert linked to various types of cancer (Haiman, et al., 2007; Kiemeney, et al., 2008; Zanke, et al., 2007). The accumulation of variants in this region could signal the presence of a functional element that evolves rapidly.

Tajima's D value measuring deviation from neutral evolution has a value of −2.58 in the RACEfrag regions and of −1.78 outside the RACEfrags. These negative values are compatible with either positive selection or population expansion (Carlson, et al., 2005). The distinction between the two alternatives is difficult given the fact that the human population has expanded rapidly during the last millennia. It is apparent however that the RACEfrags and their flanking regions are under different evolutionary constraints.

This study has failed to show that nucleotide variation within the RACEfrags is responsible for the Rett syndrome or its phenotypic variants. This does not however mean that the RACEfrags examined are non-functional. Functional analyses in cell culture systems or whole animals may contribute to the functional annotation of each MECP2-linked RACEfrag. The difficulty, however, in using experimental animals is that the majority of RACEfrags are not well conserved among different species (Denoeud, et al., 2007). Thus the study of natural pathogenic mutations is often the main way to verify functional genomic regions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the ENCODE project grant from the NIH, the Italian Telethon grant (GTB07001C) to ARen, and the MHV fellowship to LB.

REFERENCES

Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet. 1999; 23(2): 185–188. [PubMed: 10508514]

Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nat Rev Genet. 2006; 7(4):277–282. [PubMed: 16534515]

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. Hum Mol Genet. 2005; 14(1):59–69. [PubMed: 15525656]

Bienvenu T, Philippe C, De Roux N, Raynaud M, Bonnefond JP, Pasquier L, Lesca G, Mancini J, Jonveaux P, Moncla A, et al. The incidence of Rett syndrome in France. Pediatr Neurol. 2006; 34(5):372–375. [PubMed: 16647997]

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146):799–816. [PubMed: 17571346]

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999; 22(3):231–238. [PubMed: 10391209]

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. 2005; 15(11):1553–1565. [PubMed: 16251465]

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005; 308(5725):1149–1154. [PubMed: 15790807]

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 2007; 17(6):746–759. [PubMed: 17567994]

Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, et al. Efficient targeted transcript discovery via array-based normalization of RACE libraries. Nat Methods. 2008

Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet. 2007; 39(5):638–644. [PubMed: 17401364]

Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet. 1999; 22(3):239–247. [PubMed: 10391210]

Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics. 2006; 7:409. [PubMed: 16968531]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002; 12(6):996–1006. [PubMed: 12045153]

Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet. 2008

Kriaucionis S, Bird A. The major form of MeCP2 has a novel N-terminus generated by alternative splicing. Nucleic Acids Res. 2004; 32(5):1818–1823. [PubMed: 15034150]

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5(10):e254. [PubMed: 17803354]

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. Pattern of sequence variation across 213 environmental response genes. Genome Res. 2004; 14(10A):1821–1831. [PubMed: 15364900]

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 2007; 17(6):760–774. [PubMed: 17567995]

Mari F, Azimonti S, Bertani I, Bolognese F, Colombo E, Caselli R, Scala E, Longo I, Grosso S, Pescucci C, et al. CDKL5 belongs to the same molecular pathway of MeCP2 and it is responsible for the early-onset seizure variant of Rett syndrome. Hum Mol Genet. 2005; 14(14):1935–1946. [PubMed: 15917271]

Mencarelli MA, Kleefstra T, Katzaki E, Papa FT, Cohen M, Pfundt R, Ariani F, Meloni I, Mari F, Renieri A. 14q12 Microdeletion syndrome and congenital variant of Rett syndrome. Eur J Med Genet. 2009

Mnatzakanian GN, Lohi H, Munteanu I, Alfred SE, Yamada T, MacLeod PJ, Jones JR, Scherer SW, Schanen NC, Friez MJ, et al. A previously unidentified MECP2 open reading frame defines a new protein isoform relevant to Rett syndrome. Nat Genet. 2004; 36(4):339–341. [PubMed: 15034579]

Percy AK, Lane JB. Rett syndrome: clinical and molecular update. Curr Opin Pediatr. 2004; 16(6): 670–677. [PubMed: 15548931]

Renieri A, Mari F, Mencarelli MA, Scala E, Ariani F, Longo I, Meloni I, Cevenini G, Pini G, Hayek G, et al. Diagnostic criteria for the Zappella variant of Rett syndrome (the preserved speech variant). Brain Dev. 2009; 31(3):208–216. [PubMed: 18562141]

Sampieri K, Meloni I, Scala E, Ariani F, Caselli R, Pescucci C, Longo I, Artuso R, Bruttini M, Mencarelli MA, et al. Italian Rett database and biobank. Hum Mutat. 2007; 28(4):329–335. [PubMed: 17186495]

Scala E, Ariani F, Mari F, Caselli R, Pescucci C, Longo I, Meloni I, Giachino D, Bruttini M, Hayek G, et al. CDKL5/STK9 is mutated in Rett syndrome variant with infantile spasms. J Med Genet. 2005; 42(2):103–107. [PubMed: 15689447]

Segawa M, Nomura Y. Rett syndrome. Curr Opin Neurol. 2005; 18(2):97–104. [PubMed: 15791137]

Tao J, Van Esch H, Hagedorn-Greiwe M, Hoffmann K, Moser B, Raynaud M, Sperner J, Fryns JP, Schwinger E, Gecz J, et al. Mutations in the X-linked cyclin-dependent kinase-like 5 (CDKL5/STK9) gene are associated with severe neurodevelopmental retardation. Am J Hum Genet. 2004; 75(6):1149–1154. [PubMed: 15499549]

Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics. 2005; 21(11):2791–2793. [PubMed: 15814564]

Wan M, Lee SS, Zhang X, Houwink-Manville I, Song HR, Amir RE, Budden S, Naidu S, Pereira JL, Lo IF, et al. Rett syndrome and beyond: recurrent spontaneous and familial MECP2 mutations at CpG hotspots. Am J Hum Genet. 1999; 65(6):1520–1529. [PubMed: 10577905]

Weaving LS, Christodoulou J, Williamson SL, Friend KL, McKenzie OL, Archer H, Evans J, Clarke A, Pelka GJ, Tam PP, et al. Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation. Am J Hum Genet. 2004; 75(6):1079–1093. [PubMed: 15492925]

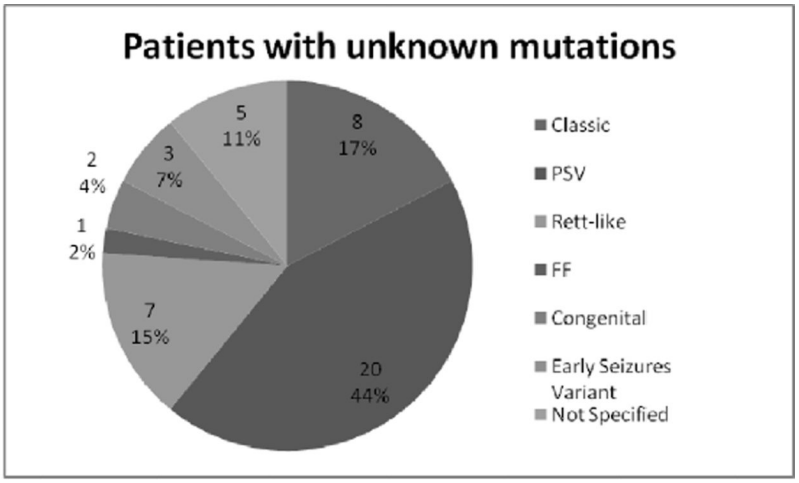
Weaving LS, Ellaway CJ, Gecz J, Christodoulou J. Rett syndrome: clinical review and genetic update. J Med Genet. 2005; 42(1):1–7. [PubMed: 15635068]

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452(7189):872–876. [PubMed: 18421352]

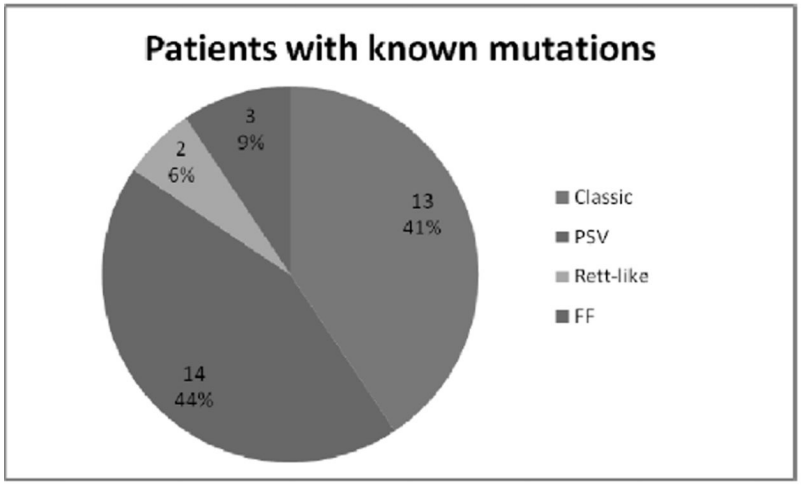
Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, Orr N, Matthews C, Qi L, Crenshaw A, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24

associated with prostate and colon cancers. Hum Genet. 2008; 124(2):161–170. [PubMed: 18704501]

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007; 39(8):989–994. [PubMed: 17618283]



A. Clinical diagnostic spectrum among the 46 patients with Rett syndrome without identifiable mutation in the MECP2 or CDKL5 genes, group "PNM". (PSV = Preserved Speech Variant, FF = Forme Fruste)



B. Clinical diagnostic spectrum among the 32 patients with Rett syndrome and identified mutations in the MECP2 gene, group "PM". (PSV = Preserved Speech Variant, FF = Forme Fruste)

Figure 1.

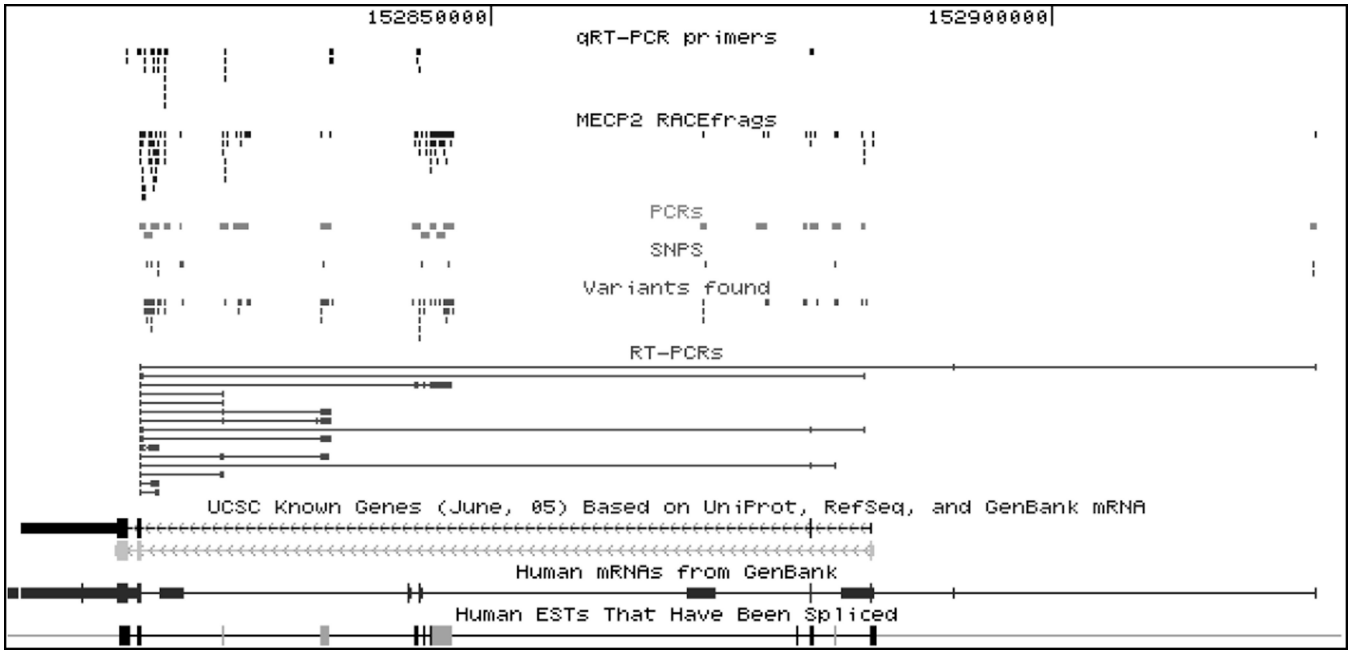


Figure 2.

Custom image from the UCSC genome browser (http://genome.ucsc.edu) (Kent, et al., 2002) showing from top to bottom: (i) coordinates of the X chromosome, NCBI Build 35 (ii) the exact positions of the primers used in the qRT-PCR to estimate the abundance of the RACEfrags, (iii) the identified MECP2 RACEfrags, (iv) the PCR fragments used for sequencing (primers are found in Supp. Table S2), (v) the known SNPs that were identified, (vi) the novel variants identified during this study (vii) the RT-PCRs. The annotated exons of MECP2, known mRNAs, and EST tracks are shown. Complementary information can be found in Fig 2 in Djebali et al, 2008(Djebali, et al., 2008)

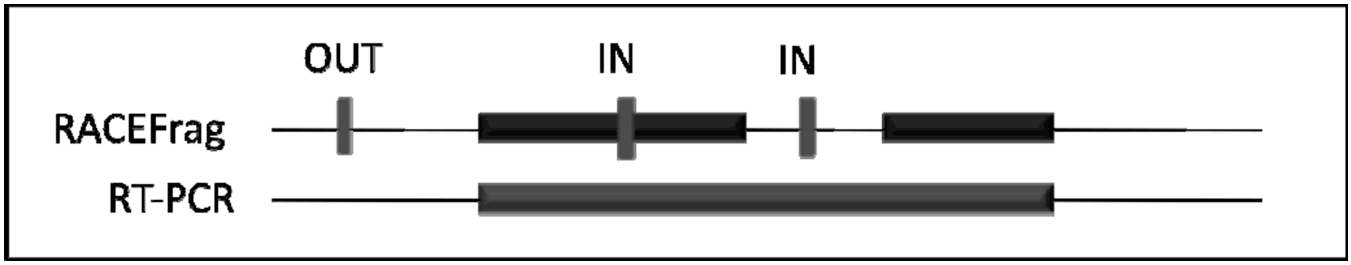


Figure 3.
Localization of the different variants within and outside of RACEfrags. The distinction is made after RACE and RT-PCR experiments as described in the text.

Table 1

Catalogue of the variants discovered during the sequencing of the RACEfrags and the adjacent regions.

Variant	Group	no of individuals	inside in original Rxfrags	inside after after RT-PCR
g.152819070 T>C	CNT	1	þ	þ
g.152819225 G>A	PNM	1	þ	þ
g.152819301 T>C	CNT	1	þ	þ
g.152819367 A>G	PNM	1		þ
g.152819418 T>A	CNT	1		þ
g.152819538delA	CNT	1		þ
g.152819716 C>T	CNT	1		þ
g.152819769 T>C	CNT	1		þ
g.152819814 G>A	CNT	1		þ
g.152819824 G>A	CNT	1	þ	þ
g.152819969 G>A	PNM	2	þ	þ
g.152819995 G>A	CNT	1	þ	þ
g.152820283 G>A	CNT	2	þ	þ
g.152820398 C>A	PM	1	þ	þ
g.152820464 C>T	PM	1	þ	þ
g.152820880 G>A	PM, CNT	3	þ	þ
g.152820972 G>T	PNM	1		
g.152822513 C>G	CNT	1		
g.152826382 C>T	PNM	1	þ	þ
g.152827405 T>G	PNM, PM, CNT	4	þ	þ
g.152827438 C>T	CNT	1	þ	þ
g.152827764 G>T	CNT	1	þ	þ
g.152828230 C>T	PM	1		
g.152828397 G>A	PM	1		
g.152834889 C>T	CNT	1		þ
g.152834899 C>T	CNT	1	þ	þ
g.152834952 C>A	CNT	1	þ	þ
g.152835019 C>T	PNM, PM, CNT	7		þ
g.152835204 T>C	CNT	1		þ
g.152835546 T>C	PNM	1		þ
g.152835857 C>T	PNM	1		
g.152843112 A>G	CNT	1		
g.152843604 C>T	PNM	1	þ	þ
g.152843622 T>C	CNT	1	þ	þ
g.152843625 T>C	CNT	1	þ	þ
g.152843695 A>C	PNM	1	þ	þ

Variant	Group	no of individuals	inside in original Rxfrags	inside after after RT-PCR
g.152843701 T>C	PNM	1	þ	þ
g.152843955 T>A	CNT	1		
g.152844044 G>A	CNT	1		þ
g.152844597 C>T	CNT	1	þ	þ
g.152844796 G>A	CNT	1	þ	þ
g.152845376 A>G	PM	1	þ	þ
g.152845794 T>C	PM	1	þ	þ
g.152845924 A>G	CNT	1		
g.152845948 G>A	CNT	1		
g.152846038 G>A	PNM, PM, CNT	3	þ	þ
g.152846239 T>C	CNT	1	þ	þ
g.152846244 G>A	CNT	1	þ	þ
g.152846387 C>T	CNT	1		
g.152846510 A>G	CNT	1		þ
g.152846581 T>A	CNT	1		þ
g.152845879 C>T	PNM	1	þ	þ
g.152868815 T>A	PNM	1		
g.152868846 C>T	CNT	1		
g.152868980 A>G	PM	1		
g.152874498 T>C	CNT	1		
g.152874573 T>G	CNT	2		
g.152877868 A>T	CNT	2		
g.152878128 G>A	CNT	1	þ	þ
g.152878962 C>T	PNM, PM, CNT	11		
g.152880633 C>G	PNM	2		
g.152880915 G>C	PM	1		
g.152883115 C>A	CNT	1		
g.152883356 G>A	PNM	1		
g.152819564 A>G	CNT	1		þ

In the group column “PNM” stands for the patients where no mutation is found, “PM” for the group of patients where mutations have been found and “CNT” for the healthy individuals control group.

Table 2

A. Total number of variants											
Variants found only once (het freq = 0.56%)						**Variants found 2–4 times (het freq = 1.12–2.25%)**					
	Group CNT	**Group PNM**	**p-value**	**Group PM**	**p-value**		**Group CNT**	**Group PNM**	**p-value**	**Group PM**	**p-value**
No of individuals with variants	35	12	0.342	8	0.387	No of individuals with variants	10	7	0.409	3	1
Total no of patients	100	46		32		Total no of patients	100	46		32	
						p-values are calculated by Fisher's exact test (sum of small p-values)					

B. Variant's frequencies based on their location within or outside the RACEFrag											
Variants found only once (het freq = 0.56%)						**Variants found 2–4 times (het freq = 1.12–2.25 %)**					
	Group CNT	**Group PNM**	**p-value**	**Group PM**	**p-value**		**Group CNT**	**Group PNM**	**p-value**	**Group PM**	**p-value**
No of individuals with variants IN	25	8	0.396	4	0.219	No of individuals with variants IN	6	5	0.323	3	0.423
No of individuals with variants OUT	10	4	1	4	0.744	No of individuals with variants OUT	4	2	1	0	0.572
Total no of patients	100	46		32		Total no of patients	100	46		32	
						IN and OUT are based on the hybridization data, taking into account the RT-PCRs					

A: Number of variants in each studied group; p-values are from the comparison of the relative frequencies between groups CNT and PNM and groups CNT and PM B: Number of variants within and outside the RACEfrags in each group studied p-values are calculated as above.

Table 3

The theta values (population mutation parameter or level of nucleotide polymorphism) calculated from different studies are shown.

Genomic area	Theta value (10−4)	Reference
CDS, Venter genome	3.6	(Levy, et al., 2007)
CDS, Watson genome	4	(Wheeler, et al., 2008)
genes (CDS + introns)	5.4	(Cargill, et al., 1999)
genes (CDS + introns)	5.6	(Bhangale, et al., 2005)
introns, Venter genome	5.6	(Levy, et al., 2007)
introns, Watson genome	6.2	(Wheeler, et al., 2008)
Venter genome, total	6.2	(Levy, et al., 2007)
Watson genome, total	6.5	(Wheeler, et al., 2008)
genes (CDS + introns)	6.7	(Livingston, et al., 2004)
CDS	8	(Halushka, et al., 1999)
8q24 area	10.2	(Yeager, et al., 2008)
non RACEfrags(introns)	6.8	present study
RACEfrags	10.5	present study

The values from the present study are listed on the bottom. All of them are represented as 10^{-4}. CDS : Coding Sequence