6660–6672 Nucleic Acids Research, 2012, Vol. 40, No. 14 Published online 17 April 2012
doi:10.1093/nar/gks302

Sequence and expression analysis of gaps in human chromosome 20

Sheroy Minocherhomji[1], Stefan Seemann[2,3], Yuan Mang[1,2], Zahra El-schich[1], Mads Bak[1,2], Claus Hansen[1,2], Nickolas Papadopoulos[4], Knud Josefsen[5], Henrik Nielsen[2,6], Jan Gorodkin[2,3], Niels Tommerup[1,2] and Asli Silahtaroglu[1,*]

[1]Wilhelm Johannsen Centre for Functional Genome Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen N, Denmark, [2]Center for Non-coding RNA in Technology and Health, [3]Division of Genetics and Bioinformatics, IBHV, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark, [4]Ludwig Center for Cancer Genetics, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA, [5]Bartholin Institute, Copenhagen University Hospital, Section 3733, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark and [6]Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen N, Denmark

Received February 6, 2012; Revised and Accepted March 21, 2012

ABSTRACT

The finished human genome-assemblies comprise several hundred un-sequenced euchromatic gaps, which may be rich in long polypurine/polypyrimidine stretches. Human chromosome 20 (chr 20) currently has three unfinished gaps remaining on its q-arm. All three gaps are within gene-dense regions and/or overlap disease-associated loci, including the *DLGAP4* locus. In this study, we sequenced ~99% of all three unfinished gaps on human chr 20, determined their complete genomic sizes and assessed epigenetic profiles using a combination of Sanger sequencing, mate pair paired-end high-throughput sequencing and chromatin, methylation and expression analyses. We found histone 3 trimethylated at Lysine 27 to be distributed across all three gaps in immortalized B-lymphocytes. In one gap, five novel CpG islands were predominantly hypermethylated in genomic DNA from peripheral blood lymphocytes and human cerebellum. One of these CpG islands was differentially methylated and paternally hypermethylated. We found all chr 20 gaps to comprise structured non-coding RNAs (ncRNAs) and to be conserved in primates. We verified expression for 13 candidate ncRNAs, some of which showed tissue specificity. Four ncRNAs expressed within the gap at *DLGAP4* show elevated expression in the human brain. Our data suggest that unfinished human genome gaps are likely to comprise numerous functional elements.

INTRODUCTION

Completion of the sequence analysis of the human genome provided in-depth characterization of the physical sequence of genomic DNA (gDNA), albeit with numerous unfinished gap regions (1,2). After almost 11 years of post-analysis, these gap regions remain present and uncharacterized in both euchromatic and heterochromatic regions of the human nuclear genome (3). Some of the euchromatic gaps are flanked by segmental duplications and remain technically challenging to sequence. Alternative approaches previously used for closing some of these gaps elsewhere in the genome, include fosmid resources (4), transcript-based approaches between paired expressed sequence tags (5), chromosome walking and shotgun sequencing (6) and more recently sequencing by synthesis using next-generation high-throughput 454 sequencing technology (7). None of these single approaches has been successful in closing all the remaining gaps in the human genome, possibly due to their unique structural or sequence identity.

The sequence assembly of human chromosome 20 (chr 20) currently has three remaining unfinished gaps, all located on the long arm (q-arm) of the chromosome (8). All three gaps are within gene-dense euchromatic areas. There are no segmental duplications up to a distance of >100 kb on either side of these gaps. Since all three human chr 20 gaps lay in gene-dense areas or overlapped loci associated with human disorders (9,10), we hypothesized that their molecular and transcriptional characterizations will provide valuable insight into the regulation of surrounding regions. In this study, we have successfully sequenced ~99% of the unfinished sequences and determined the correct sizes of the three remaining gaps

*To whom correspondence should be addressed. Tel: +45 35 32 76 21; Fax: +45 35 32 78 45; Email: asli@sund.ku.dk

on human chr 20. Furthermore, we characterized their chromatin and methylation compositions, assessed transcription (predicted and actual transcripts) and measured sequence and RNA structure conservation (Figure 1A).

MATERIALS AND METHODS

Cell culture

Epstein–Barr virus (EBV) transformation of peripheral blood lymphocytes (PBLs) was performed using established protocols (11). Mouse–human cell hybrid cell lines containing only one single paternally or maternally derived human chr 20 were generated using mouse E2 cells as published elsewhere (12). The maternally inherited chr 20 homolog was disrupted by a chromosomal translocation t(8:20)(p12;q11.23) that we previously reported (13).

Sanger sequencing

Overlapping polymerase chain reaction (PCR) primers (Supplementary Table S1) were designed using Primer 3 software (14). Touchdown PCR followed by Sanger sequencing (15), was adopted to sequence and close chr 20 gaps using gDNA from human PBLs of multiple controls.

Mate pair paired-end Illumina sequencing of gDNA

Mate pair paired-end sequencing of multiple gDNA libraries was performed as previously published (16), using an Illumina-recommended protocol and sequenced on a Genome Analyzer II (Illumina). Genomic DNA was

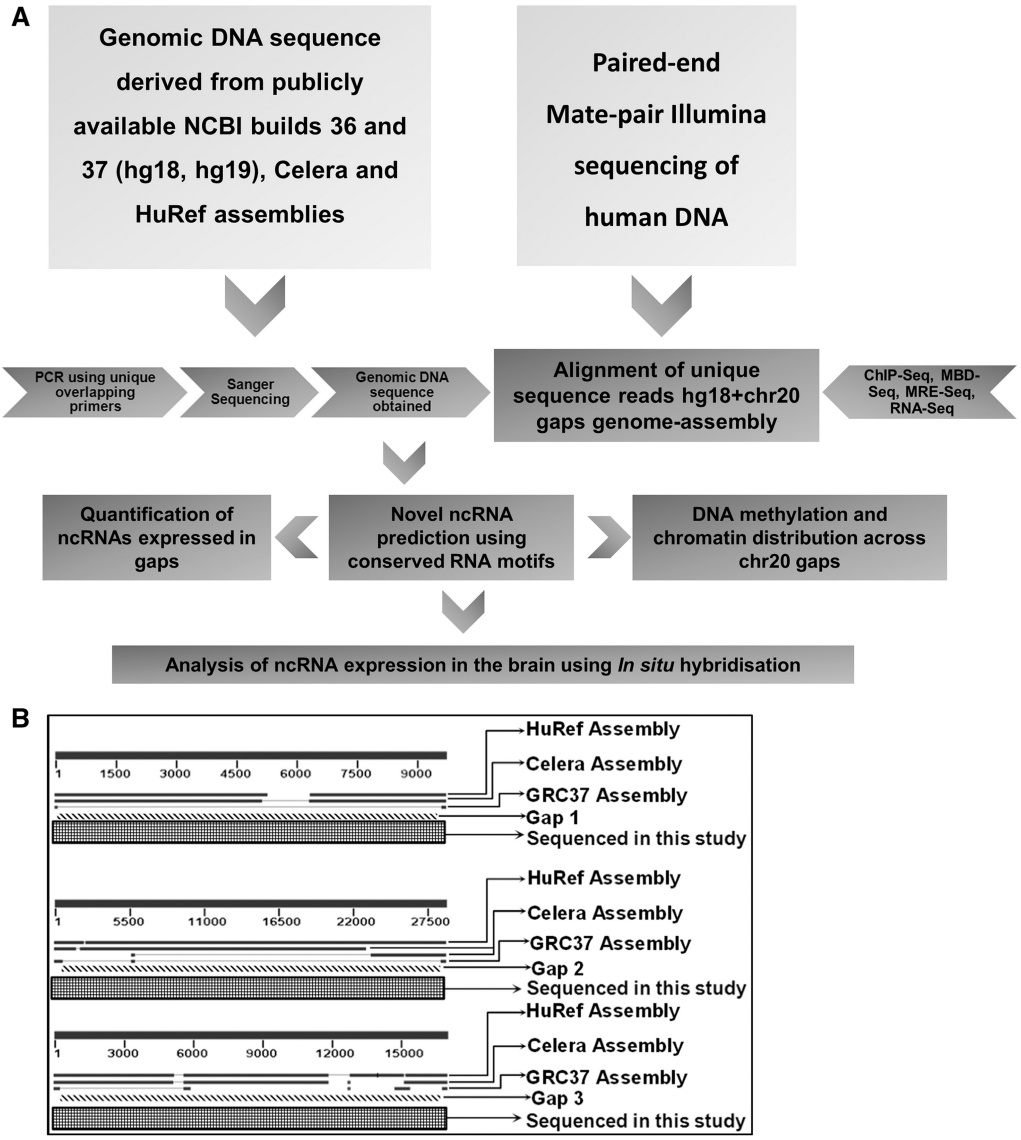

Figure 1. Strategy followed to sequence and characterize previously unfinished human chr 20 gaps. (**A**) Schematic representation of experimental workflow and study design. (**B**) BLASTN analysis of three remaining human chr 20 gaps (1, 2 and 3) in the HuRef, Celera and GRC37 (hg19) assemblies. Stripped rectangles denote gap regions sequenced here.

obtained from PBLs of three balanced chromosomal translocation carriers (not used for PCR and Sanger sequencing) with intact chr 20 homologs. Fragmented gDNA was sonicated to sizes of 3-kb fragments.

Immunoprecipitation of hypermethylated gDNA using his-MBD2b (MBD-Seq)

Approximately 2–4 μg of gDNA was obtained from PBLs of either a translocation carrier (13), phenotypically normal controls or purchased normal human cerebellum (41-year-old male donor, Cat # TCB-4098, Capital Biosciences) and sonicated/nebulized, followed by purification using QIAquick columns and purification kits (Qiagen). Immunoprecipitation (IP) was performed using a his-tagged MBD2b (termed MBD) antibody (0.1 μg/μl) and the MethylCollector kit (Cat # 5502, Active Motif) according to manufacturer's instructions followed by high-throughput sequencing.

Methylation-sensitive restriction enzyme digestion sequencing of gDNA

Restriction enzyme digests were performed using the methylation-sensitive *H*paII enzyme according to the manufacturer's instructions (New England Biolabs). Digested gDNA was size selected using 2% agarose gel electrophoresis. A region corresponding to 150–300 bp was excised and purified using a gel purification kit (Qiagen) followed by high-throughput sequencing.

Chromatin IP sequencing

Chromatin IP (ChIP) was performed using sonicated chromatin obtained from EBV-transformed PBLs of a translocation carrier (13) and antibodies for histone 3 trimethylated at Lysine 27 (H3K27me3) (Active Motif, Cat. #39155), H3+ (Abcam, Cat #ab1791) or IgG (Cell Signaling Technology Cat #2729S) as previously published (17) or using the ChIP-IT express kit (Cat # 53008, Active Motif) followed by high-throughput sequencing.

High-throughput Illumina sequencing of RNA sequencing

High-throughput RNA sequencing (RNA-Seq) was performed using poly-A RNA from human hippocampus and temporal lobe (Ambion) tissues together with paired-end sequencing adapters, an Illumina-recommended sequencing protocol and run on a Genome Analyzer II (Illumina). In addition, the Body Map 2.0 sequence data archived at the European Nucleotide Archive (ENA) under accession ERP000546 was used for RNA-Seq analysis of other human tissues as described elsewhere (18).

High-throughput Illumina sequencing

Library preparation of gDNA obtained from MBD-Seq, methylation-sensitive restriction enzyme sequencing (MRE-Seq) and ChIP-Seq experiments was performed using single read or paired-end adapters and sequenced on a Genome Analyzer II according to manufacturer's instructions (Illumina). ENCODE raw sequenced reads for H3K27me3 ChIP on GM12878, EBV–PBLs were obtained from UCSC genome browser (genome.ucsc.edu) (19). Unique sequenced reads from all high-throughput experiments including ENCODE H3K27me3 ChIP-Seq experiments were re-aligned to an altered hg18 assembly that included sequence newly identified in this study for all three chr 20 gaps (termed hg18 + chr 20 gaps genome assembly) using Bowtie (20). The 709 bp starting from the 5′-end of gap 2 was not included in the re-alignment assembly (hg18 + chr 20 gaps) file so as to not change the genome base positions (chromosome offsets) on either side of the gap 2 during re-alignment. Importantly, no significant reads were shown to align to this region in any ChIP-Seq, MBD-Seq or MRE-Seq experiment. The required numbers of unsequenced nucleotides (NNNs) were maintained in gaps 1 and 3 as they were estimated to be 20 and 110 kb, respectively, in the hg18 human-genome assembly. MBD-Seq, MRE-Seq and ChIP-Seq reads were analyzed using Cisgenome software as published elsewhere using a false discovery rate or <0.01 (21). ChIP-Seq data were analyzed using H3+ or IgG as negative controls. Paired-end RNA-Seq reads had a maximum distance of ~5 kb between each pair.

In situ hybridizations

In situ hybridizations were performed using purchased control human cerebellum brain sections (70-year-old female donor, Cat # TCB-4089 Capital Biosciences) as previously published (22) using *Fluorescein*-labeled oligos (Supplementary Table S7). Signals were analyzed using an MVX10 microscope, F-View black/white and Cell^P software (Olympus).

Bisulfite allelic sequencing

Bisulfite conversion of gDNA was performed using the Methylation Gold kit (Zymo Research) according to manufacturer's instructions. Bisulfite allelic sequencing was performed using primers listed in Supplementary Table S1, Methyl-specific PCR, touchdown PCR conditions (15,23) and Sanger sequenced on an ABI 3730×1 DNA analyzer (Applied Biosystems). Sequence output was analyzed using Chromas software (Technelysium) and BLASTN (24).

Expression analysis

Complementary DNA (cDNA) was prepared from total RNA obtained from a human multi-tissue panel (NSGene) and human brain panel (BioCat), using Superscript II Reverse Transcriptase (Invitrogen). Real time PCR (RT-PCR) was performed with multiple batches of cDNA samples, primers listed in Supplementary Table S1 and run on the Opticon 2 thermocycler (Bio-Rad) using FastStart DNA Masterplus SYBR green I (Roche). RT-PCR using cDNA prepared from RNA of a human multi-region brain panel (BioCat) was performed in duplicates or triplicates. Data were quantified and normalized using the geometric mean normalization factor as published elsewhere (25) and housekeeping genes: *GAPDH, ATP5A1,*

COX4A, ATP6A, ALAS1, B2M, ATP8A, PBGD, G6PD and *HPRT* (Supplementary Table S1).

Northern blot analysis

Northern blot analysis was performed using purchased total human cerebellum RNA of a 26-year-old normal human male (Cat # 4099, Capital Biosciences). Amplified PCR products were purified using the QIAquick purification kit and columns (Qiagen). Northern blotting was performed using methods published previously (26,27) using purified PCR products end labeled with T4 PNK and [γ-^{32}P] ATP or body labeled by the random primer method and [α-^{32}P] ATP (specific activity >3000 Ci/mmol, PerkinElmer).

CpG island prediction

CpG islands (Supplementary Table S3) were predicted using previously established criteria (28,29) and the CpG island Searcher tool (30).

Visualization of aligned high-throughput sequenced reads in chr 20 gaps

The graphical user interface MapView (31), Cisgenome (21) and/or the genome browser at UCSC (19) were used to visualize short-read data for each of the gap regions at nucleotide level, using high-throughput MRE-Seq MBD-Seq, ChIP-Seq, RNA-Seq and mate pair paired-end sequenced data sets.

Conservation analysis of human chr 20 gaps

The gap regions and 23 other genome assemblies were aligned using the lastz alignment (version 1.02, gap open penalty of O = 400 and gap extension penalty of E = 3) (32). Lastz output was chained by the axtChain program (minimum chain score of 3000 and 'medium' linear gap matrix) and chained alignments were processed into nets by chainNet and netSyntenic, which are all available from Webb Miller's lab at Penn State University (http://www.bx.psu.edu/miller_lab/). Multiple alignments were created with roast from the tba/multiz alignment program (33) version 012909 also available from Webb Miller's lab. A dynamic programming range of 30 and a minimum block size of size 1 were employed. Phylogenetic trees were first estimated by the neighbor joining algorithm under the K80 distance model. The branch lengths were maximum-likelihood estimated using the HKY85 substitution model. The evolutionarily conserved elements in the multiple alignments of the three gaps are measured by phastCons scores (34). We let phastCons (35) itself estimate the phylogenetic models (conserved and non-conserved) directly from the 24-way multiple alignments and the previously calculated neighbor joining trees (phastCons' option 1).

Prediction of structured RNAs

We searched the gap regions that were not repeat masked for known non-coding RNAs (Rfam) (36) by using INFERNAL ($E < 1E^{-6}$) (37) and BLAST ($E < 1E^{-10}$). We applied CMfinder (38) to simultaneously predict alignment and consensus secondary structures of syntenic sequences. The resulting RNA structural alignments are described by covariance models (CMs) (39) that can be searched by INFERNAL. First, we screened the gap regions for structured RNA motifs previously predicted by CMfinder (38) using INFERNAL ($E < 1E^{-6}$). These motifs were found by a human genome-wide screen similar to the screen of the ENCODE regions (40) and ran conserved in a pool of 17 vertebrates. Second, we run CMfinder on the 24-way multiple alignments of the gap regions ($P > 60$).

RESULTS

Location and Sanger sequencing of previously unfinished gaps on the long arm of human chr 20

Starting from the centromere on the long arm of human chr 20, the first unfinished gap is located within the *DLGAP4*/*SAPAP4* [discs, large (*Drosophila*) homolog-associated protein 4 or PSD-95/SAP90-associated protein 4] at position chr 20:34360500-34380499 (NCBI build 36: hg18) (Supplementary Figure S1A). The second gap (chr 20:60524833-60551882, hg18) is distal to the *C20orf200*/NCRNA00335 (Chr20 Open Reading Frame (ORF) 200/Non-coding RNA 335) (Supplementary Figure S1B). Whereas the third gap (chr 20:60623815-60733814, hg18) is distal to the *SLCO4A1* (Solute Carrier Organic anion transporter family, member 4 A1) (Supplementary Figure S1C). From here on in this study each of these unfinished gaps on human chr 20 is annotated as gap 1, gap 2 and gap 3, respectively. According to the hg18 genome assembly the three unfinished gaps are estimated to be 20 kb (gap 1), 27 050 bp (gap 2) and 110 kb (gap 3), respectively. The recently released GRC37 (hg19) genome assembly did not contain any new sequence representation for any of these three gaps on chr 20. The sizes of these chr 20 gaps according to the GRC37/hg19 genome assembly were estimated to be 50 kb each and located at positions chr 20:34897086-34947085 (gap 1), chr 20:61091438-61141437 (gap 2), chr 20:61213370-61263369 (gap 3). Neither the hg18 nor the hg19 assemblies contain the complete gDNA sequence for any of the three remaining human chr 20 gaps.

Here, we successfully obtained ~99% nucleotide sequence and identified the correct sizes for each of the three remaining gaps on human chr 20. We initially combined sequences available in the publicly released NCBI build 36/hg18 and GRC37/hg19 or the Celera and HuRef genome-assemblies using BLASTN (24) analysis of their flanking regions (Figure 1B). Following this, we Sanger sequenced all three-gap regions on human chr 20 in gDNA obtained from PBLs. We identified the complete and correct genomic sizes of all three gaps to be 9485 bp (gap 1), 27 759 bp (gap 2) and 16 383 bp (gap 3) (Supplementary Figure S1A–S1C), respectively. Starting from the known sequence outside each gap we designed overlapping PCR primers to amplify and sequence these gaps from the outside, going inwards. BLASTN analysis confirmed that most of the sequences obtained here, have matched with sequence already available in either the

Celera or HuRef assemblies or both, albeit for some differences in single nucleotides, possibly indicative of single nucleotide polymorphisms (Figure 1B). None of the sequence we obtained is represented in neither hg18 nor hg19 genome assemblies. We also obtained novel sequence for some regions, previously not represented in either Celera or HuRef assemblies. Some regions that were indeed identified as unfinished in the Celera and HuRef assemblies were found to be artificial, suggesting that their flanking regions at either end were in fact joined. We found no evidence for segmental duplications on either side of these gaps up to a distance of <100 kb.

Mate pair paired-end Illumina sequencing of human chr 20 gaps

Next, using paired-end sequencing of multiple mate pair human PBL gDNA libraries we confirmed sequence identity of all three human chr 20 gaps (Figure 2A–C). We used multiple gDNA samples, which were in addition to the ones used for Sanger sequencing. We modified the hg18 genome assembly in a way that the gaps having the unsequenced nucleotides (NNN) on the long arm of chr 20 were replaced with sequence we obtained previously using Sanger sequencing, taking care not to change the genome base position (chromosome offsets) on either side of these gaps. We termed this altered human genome assembly: hg18 + chr 20 gaps. We successfully obtained the complete sequence for gap 2 but were unable to generate sequence for ~348 bp in gap 1 and ~413 bp in gap 3 due to the highly repetitive nature of these regions as established using Sanger sequencing. Since we obtained an individual PCR product for these repetitive regions in gap 1 and gap 3, we were able to estimate their sizes but unable to generate complete nucleotide sequence (Supplementary Figure S1A and S1C). Using RepeatMasker (v3.3.0) (41) and the default human repeat database we found all the three gap regions on chr 20 to have long interspersed nuclear elements, small interspersed nuclear elements, long transposon repeats and simple and low complexity repeats, yet at low copy numbers (Supplementary Table S2). Within gap 1 we obtained novel sequence for 1052 bp within a ~1400-bp region that was not present in any genome assembly. The sequence for the remaining ~348 bp contained within this region could not be generated due to a continuous TG repeat.

H3K27me3 is distributed across all three human chr 20 gaps

Following sequence analysis and identification of the actual sizes of the three chr 20 gaps we characterized their chromatin distribution. H3K27me3 is associated with repressed chromatin and is distributed on both the flanking sides of all three chr 20 gaps in most lineage-committed and cultured immortalized human somatic cells (42). Using ChIP-Seq we found H3K27me3 to be distributed across all three-gap regions in EBV–PBLs (Figure 3A), suggesting them to be transcriptionally silenced. H3K27me3 distribution across all the three gaps was identified at the higher end of H3K27me3 distribution for the genome.

Differentially methylated CpG island in human chr 20 gap 2

Next, we assessed gDNA methylation (Ch_3^-) at the carbon-5 residue of the cytosine base (termed 5mC) of CpG dinucleotides across all three human chr 20 gaps. Using MBD-Seq and MRE-Seq we found no widespread enrichment of gDNA hypermethylation at CpG dinucleotides in either gap 1 or gap 3 in PBLs or human cerebellum (Figure 3B). However, we found hypomethylation and hypermethylation peaks to be widespread across gap 2 in gDNA from both PBLs and human cerebellum (Figure 3C). We identified five new CpG islands within gap 2, including one lying at its 3′-terminal end that extended outside the gap (Figure 3C). We also used a more stringent classification that predominantly identifies CpG islands at the 5′-end of genes (28) and identified three CpG islands. These overlapped with the CpG islands predicted earlier in gap 2 using the less stringent criteria (Figure 3C and Supplementary Table S3). Using MBD-Seq we found all CpG islands in gap 2 to be predominantly enriched for hypermethylated CpGs (Figure 3C). However, we found CpG island 4/1a within gap 2 to have both hypomethylated and hypermethylated peaks (Figure 3C), demonstrating it to be a differentially methylated region (43). In order to assess whether CpG island 4/1a was maternally or paternally methylated we took advantage of our previously established mouse–human E2 cell hybrid cell lines. Each cell hybrid contained only a single paternally or maternally derived human chr 20. Bisulfite allelic sequencing of gDNA isolated from these mouse–human cell hybrids showed CpG island 4/1a to be paternally hypermethylated (Figure 3D). As a control we tested the methylation status of the paternally expressed *MCTS1* (*PSIMCT-1)* pseudogene and *NNAT* (Neuronatin) gene located on the q-arm of human chr 20. We confirmed them to be paternally hypomethylated and maternally hypermethylated in these mouse–human E2 cells (Supplementary Figure S2, data not shown) similar to data reported elsewhere (44,45).

Conservation and expression analysis of human chr 20 gaps

Since the analyzed regions are not annotated as gaps in other organisms including primates we performed pairwise alignments (lastz) (32) of the three gap regions including 500 bp up and downstream with several other species. We found all three gaps on human chr 20 to be well covered and conserved in primates but not in more distant mammals (Figure 4A–F). Furthermore, we found evolutionary conserved elements to be present in each human chr 20-gap region (Supplementary Figure S3A–S3C). The multiple alignments of the gap regions are slightly more conserved than the 17-way UCSC multiple alignments that cover around 50% of the human (hg18) genome. This is mostly due to the primate specificity of gap regions and presence of many regions of low complexity within these gaps. In order to assess transcriptional

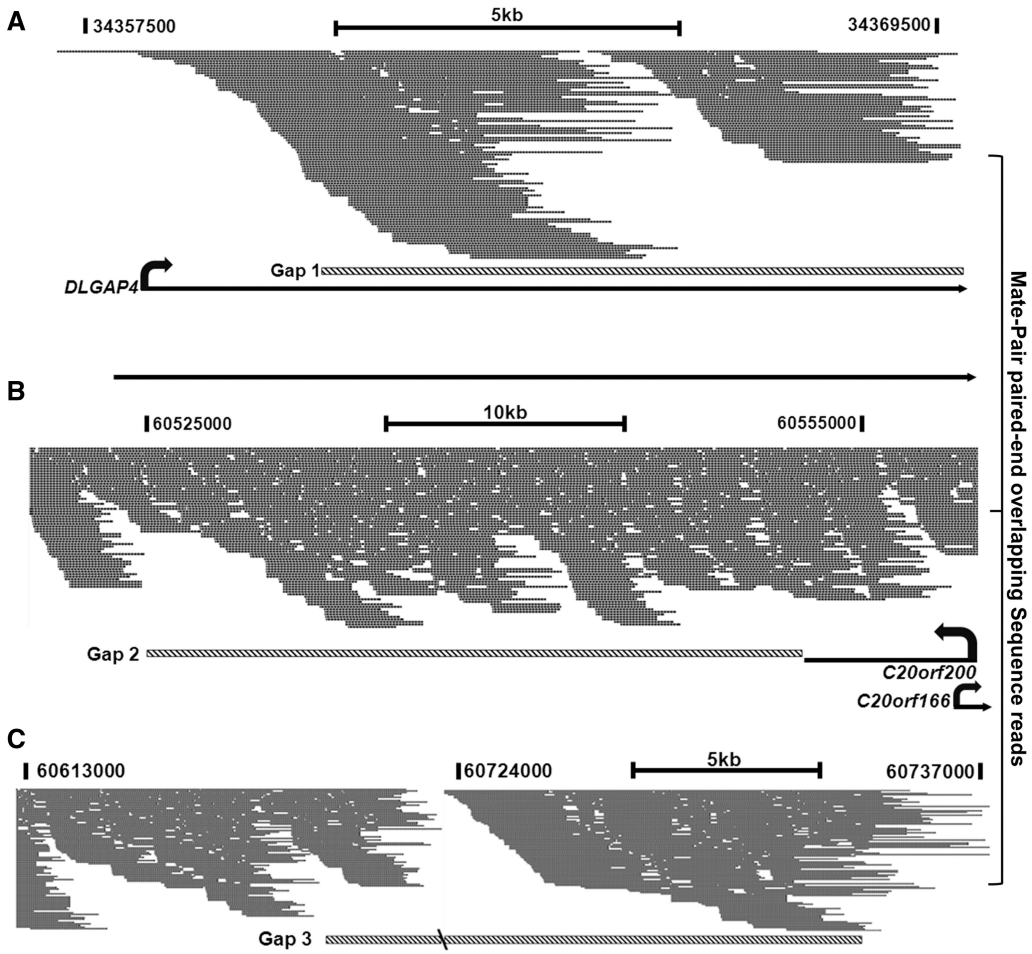

Figure 2. Mate pair paired–end Illumina sequencing of human chr 20 gap regions. (**A-C**) Representative UCSC genome browser views of mate pair paired–end Illumina sequenced unique reads aligned to the three human chr 20 gap regions using the edited hg18 + chr 20 gaps human genome assembly. Overlapping mate pair paired-end reads align across all three chr 20 gap. Regions are drawn to scale without disturbing chromosome offsets.

activity within the chr 20 gaps, we initially screened for ORFs using the ORF finder at NCBI (46) and a minimum gDNA region of >300 bp. We predicted 6 ORFs in gap 1; 38 ORFs in gap 2 and 24 ORFs in gap 3 (Supplementary Table S4). Most ORFs in gap 2 overlapped with hypermethylation peaks identified (Figure 3C and Supplementary Tables S4).

Using an in house non-coding RNA (ncRNA) annotation pipeline (47) we searched for known as well as *de novo* predicted RNA structures. Structural homology to known ncRNAs from Rfam v10.0 (48) was searched using tools in the INFERNAL package. Structural similarity to known ncRNAs from Rfam v10.0 was searched using tools in the INFERNAL package and BLAST. INFERNAL tools were also used to search for structural similarity to CMfinder-generated models of *de novo* ncRNA structure predictions elsewhere in the genome. These CMfinder models comprise structural features conserved in several vertebrate genomes. In addition, the generated 24-way multiple alignments of the gap regions were screened by CMfinder to find models of novel structured RNA motifs. Expression from these gaps was subsequently assessed using the multi-tissue RNA-Seq Body map 2.0 data. Though there are no matches to known ncRNAs in these gap regions, we predicted 16 structured RNA motifs in the multiple alignments of the gap regions (Supplementary Tables S5) and 19 additional structured RNAs in fourteen regions similar to RNA structures elsewhere in the genome (Supplementary Tables S6) with variable sizes of <250 nt. The amount of RNA structures is slightly lower in the gap regions compared to the conserved RNA structures in the human genome (predicted by CMfinder with pscore ≥60; data not shown). Sixteen of the structured RNAs are <50 nt (12 from the first and 4 from the second set of predictions) and 4 others originate from the same genomic locus. We chose to validate expression for the remaining 11 candidates with similar conserved structure elsewhere in the human genome (Supplementary Tables S7 and S8). These ncRNAs are mostly primate specific in the gap regions and thus predicted with low scores from the 24-way multiple alignments (due to short evolutionary distances).Three of these ncRNA candidates (numbers 5, 6 and 9) are present in gap 1 and a set of four ncRNA

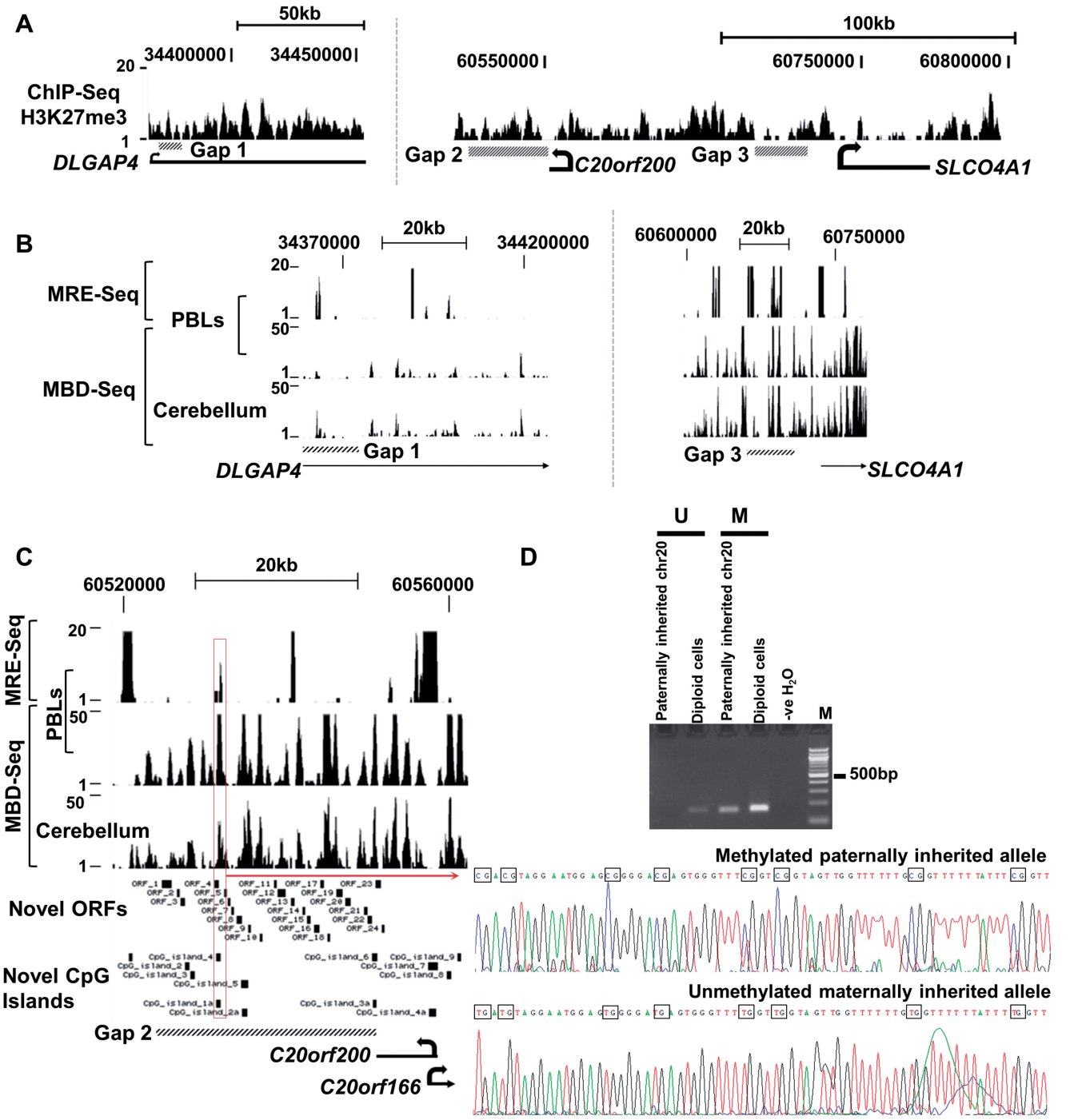

Figure 3. Chromatin distribution and DNA methylation profile across human chr 20-gap regions. (**A**) H3K27me3 is distributed across all three human chr-20 gap regions in EBV-PBLs. (**B**) MRE-Seq (hypomethylated) and MBD-Seq (hypermethylated) aligned peak reads within gaps 1 and 3. Methylation distribution is similar in PBL and human cerebellum for both gaps 1 and 3. (**C**) Methylation profile of gap 2 shows it to be enriched for both hypomethylated and hypermethylated CpGs, respectively. CpG island 4/1a (red rectangle) is enriched for both MBD-Seq and MRE-Seq reads suggesting it to be a differentially methylated locus. CpG islands 2–6 were annotated using the classification: 500-bp region of genomic DNA with ≥50% CG content and an observed CpG to expected CpG ratio of 0.6. CpG islands 1a–3a were classified based on a CpG island being a 500-bp region of genomic DNA with ≥55% CG content and an observed CpG to expected CpG ratio of 0.65. Images are drawn to scale. (**D**) Differentially methylated and paternally hypermethylated CpGs island 4/1a within gap 2 using bisulfite allelic sequencing and gDNA from mouse–human cell hybrid cell lines or diploid human cells.

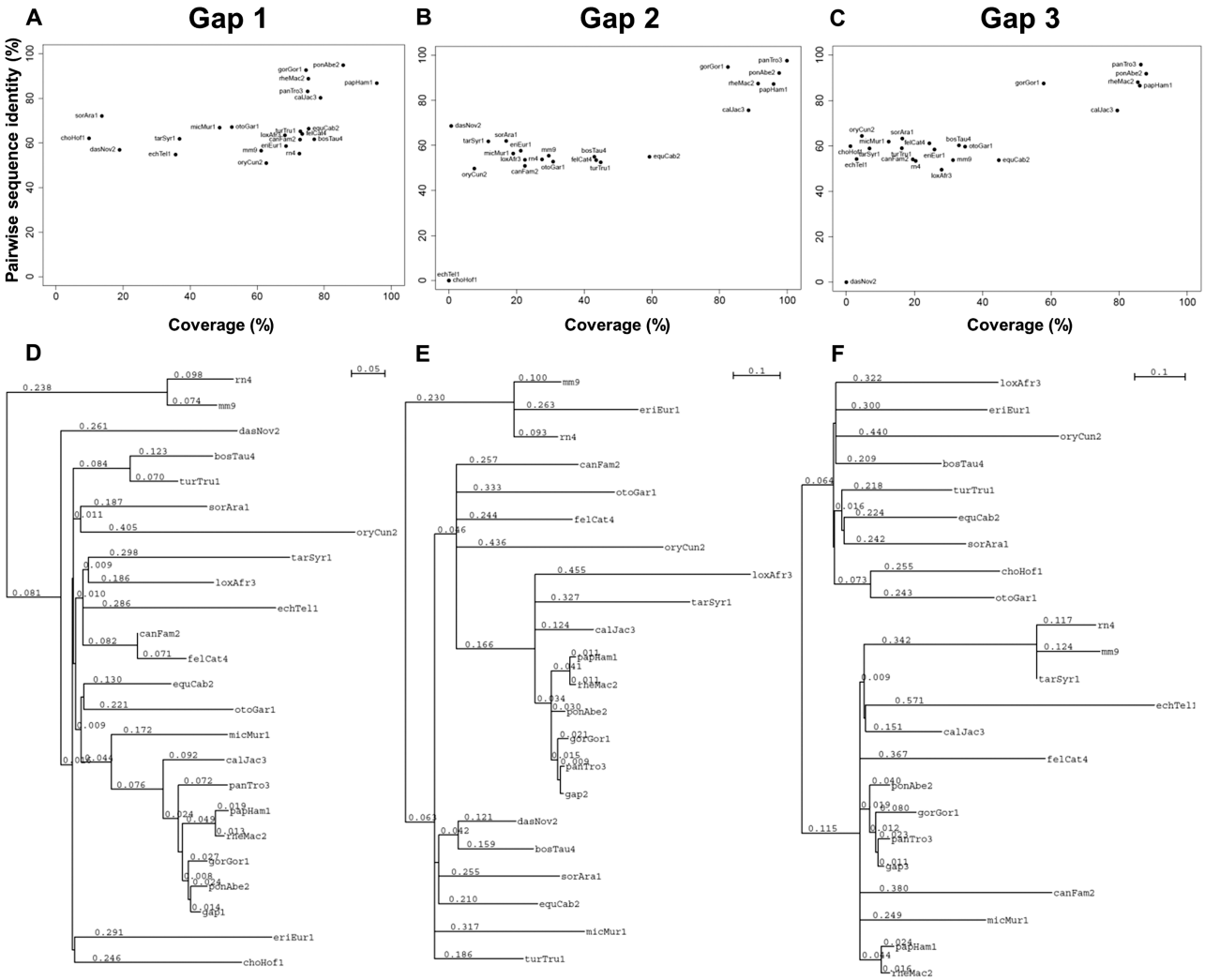

Figure 4. Conservation analysis of human chr 20 gaps. The coverage versus pairwise sequence identity of 23 organisms are shown as scatterplots for gap 1 (**A**), gap 2 (**B**) and gap 3 (**C**). Primates are clearly separated from more distant organisms. In the phylogenetic trees, gap 1 (**D**), gap 2 (**E**) and gap 3 (**F**) are all seen to be clustered together with primate species. Phylogenetic trees (generated from the 24-way multiple alignments) show all three human gaps regions tightly clustered with primates.

candidates each are in gap 2 (numbers 1, 4, 7 and 8) and gap 3 (numbers 2, 3, 10 and 11).

Candidate ncRNA numbers 10 and 11 originated from the same genomic region in gap 3. Using BLASTN (E: $<1e^{-157}$) we found no significant overlap in sequence between the ncRNA candidates and coding sequences anywhere else in the human genome (hg18 genome assembly and data not shown). We initially validated expression of all 11 CMfinder-predicted structured ncRNAs using universal human reference total RNA. Using qPCR and gel electrophoresis we identified the sizes of all products amplified to be of the predicted size (Figure 5A). To assess tissue specificity of four randomly chosen candidate ncRNAs we selected two candidate ncRNAs each from gap 1 (numbers 5 and 9) and gap 2 (numbers 7 and 8) and assessed their tissue-wide expression using total RNA from a multi-tissue human panel. Both ncRNA candidate numbers 5 and 9 are expressed in a tissue-specific manner and predominantly in human brain regions (Figure 5B). Candidate numbers 7 and 8 are similarly expressed in a tissue-specific manner in tissues including the colon and kidney but at low levels in the brain (Figure 5B). Since ncRNA candidate numbers 5, 6 and 9 are expressed from the *DLGAP4* locus known to be highly expressed in the rat brain (49), we quantified their expression in the human brain. We found all three candidate ncRNAs to be widely expressed in the human brain with elevated expression in the cerebellum, occipital and temporal lobes (Figure 5C). Their covariant CMfinder predicted ncRNA structures are conserved in the same organisms including macaque, mouse, rat, cow and dog (Figure 5D). The predicted structures and conservation of all putative ncRNAs are listed in Supplementary Tables S4 and S5 excluding candidate ncRNA number 8 as it had a weakly conserved predicted secondary structure despite of its statistical significance calculated by INFERNAL. We determined that the *DLGAP4* mRNA isoform is expressed specifically in human brain samples

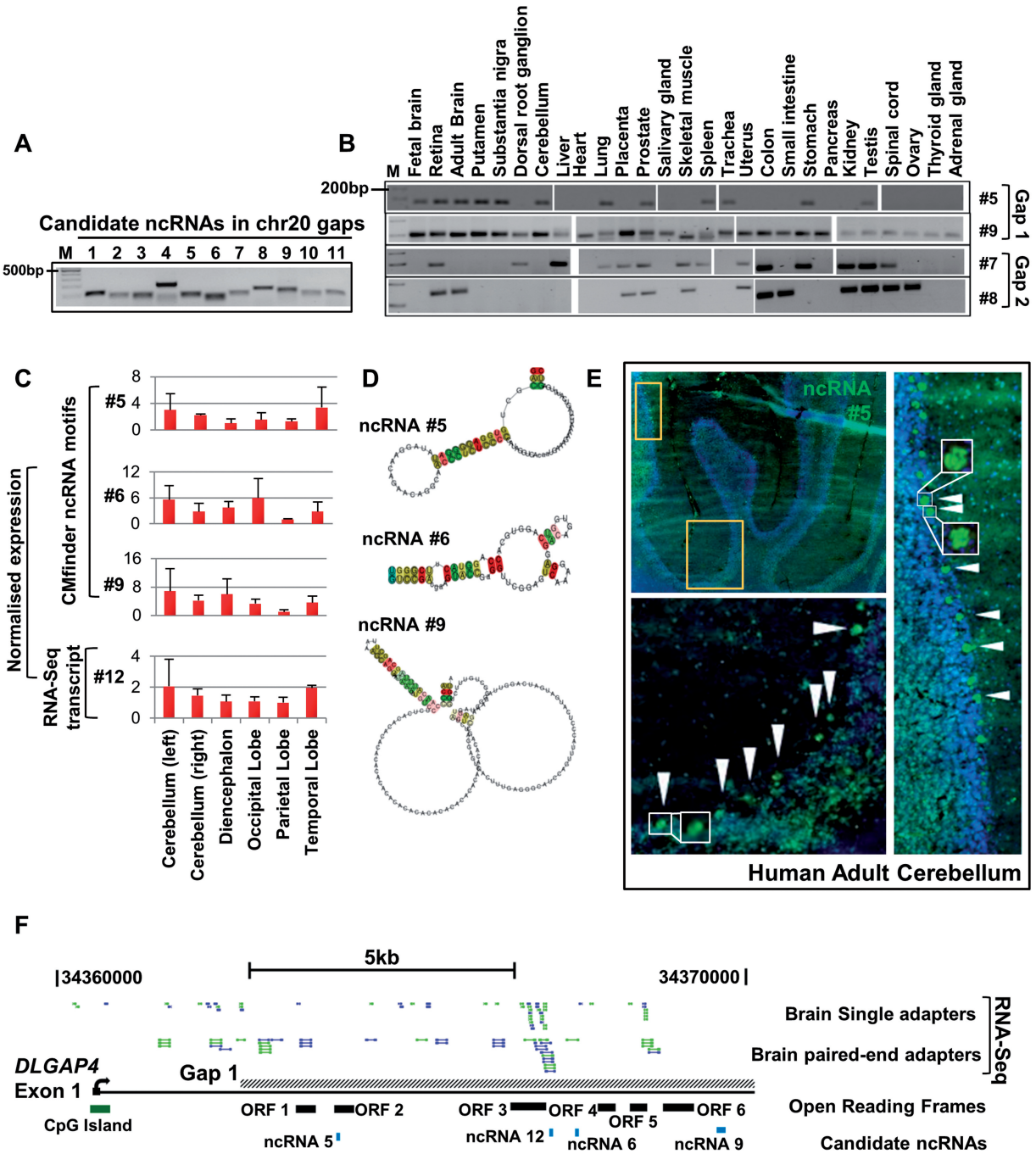

Figure 5. *De novo* prediction and experimental verification of structured ncRNAs within human chr 20 gaps. (**A**) Expression confirmation of ncRNAs predicted by CMfinder using total universal reference RNA. (**B**) Expression of candidate ncRNAs in a human multi-tissue RNA panel. (**C**) Elevated expression of candidate ncRNAs: 5, 6, 9 and 12 in a human multi-region brain panel. (**D**) RNA structure of ncRNA numbers: 5, 6, 9 predicted using INFERNAL. The Vienna RNA conservation coloring scheme highlights the mutational pattern with respect to the structure and a circle around a variable base(s) marks consistent and compensatory mutations. The color coding is explained in a color scheme legend. There are six different canonical base pairs (G-C, C-G, A-U, U-A, G-U, U-G) described by the different colors (*x*-axis), and if the base pair is not supported in a sequence of the alignment the color shading gets lighter (*y*-axis, incompatible pairs). (**E**) Candidate ncRNA number 5 is specifically expressed in the nucleus of Purkinje cells in the human cerebellum. (**F**) Localized view of the DLGAP4 promoter (chr 20:34,355,500–34,385,499, hg18 + chr 20 gaps) showing candidate ncRNAs in gap 1. RNA-Seq tracks show forward and reverse aligned reads in blue and green, respectively.

(Supplementary Figure S4A). Using qPCR we found no inclusion of sequences from gap 1 transcribed into the mRNA sequence of *DLGAP4* isoform a (data not shown). Subsequently, we characterized the spatial and temporal expressions of ncRNA candidate number 5 in gap 1, expressed from within this brain-specific *DLGAP4* locus in the human brain. This candidate ncRNA number 5 expressed from the sense strand was highly expressed in

Purkinje neurons in the cerebellum similar to the *DLGAP4* mRNA isoform a (Figure 5E, data not shown). Importantly, we found candidate ncRNA number 5 to be expressed in a specifically localized manner in the nucleus of Purkinje neurons (Figure 5E) attributing it a possible regulatory role. We ruled out the possibility of these candidate ncRNAs as being additional non-coding exons of *DLGAP4* mRNA isoform as no sequence from the gap was transcribed to mature *DLGAP4* mRNA isoform a transcript (data not shown).

We next asked if any other transcribed elements could be identified through an independent high-throughput RNA-Seq approach. We found some overlap of predicted structured ncRNAs within gap 1 in RNA samples from different regions of the human brain (Figure 5F, Supplementary Figure S4A and Supplementary Tables S5 and S6). However, most regions predicted using CMfinder as candidate ncRNAs were not identified in the RNA-Seq data sets. Furthermore, we found one additional transcript (ncRNA candidate number 12, Figure 5F and Supplementary Table S6) expressed from the sense strand in gap 1 and in different regions of the brain. We also found one additional transcript in gap 2 specifically expressed in RNA from PBLs (termed candidate ncRNA number 13, Supplementary Figure S4B and Supplementary Table S6) using the RNA-Seq BodyMap 2.0 data set. Candidate ncRNA numbers 12 and 13 were identified using RNA-Seq and not by CMfinder/INFERNAL predictions. Quantification of ncRNA number 12 in our multi-region human brain panel identified it to be elevated in expression in brain regions that was similar to the expression pattern of the other ncRNA candidate's numbers 5, 6 and 9 identified previously (Figure 5C). RNA-Seq data also identified ncRNA candidate number 13 expressed in PBLs to have two potential exons with total size of ~680 bp (Supplementary Figure S4C). None of the candidate ncRNAs identified here was present in any repetitive elements (Supplementary Table S2).

DISCUSSION

In this study, we have successfully sequenced three previously unfinished gaps on the q-arm of human chr 20. With the addition of these 52 866 sequenced nucleotides, the full size and sequence of chr 20 is now complete with the exception of short repeat stretches in gap 1 (~348 bp) and gap 3 (~413 bp). We determined the complete sizes of all three remaining gaps on human chr 20 as 9485 bp (gap 1), 27 759 bp (gap 2) and 16 383 bp (gap 3) using gDNA from multiple normal individuals using Sanger sequencing and high-throughput mate pair sequencing. This confirms the robustness of the methodologies used here for sequencing of previously unfinished human genome gaps. Analysis at the outset in BLASTN of these human chr 20 gaps while including the data from the Celera and HuRef assemblies revealed that sequence present in one assembly (Celera or HuRef) was sometimes identified as a gap in the other assembly or vice versa. Most of the remaining unfinished regions in the Celera and HuRef assemblies were complemented each other and were similarly sized to what we obtained in this study but were not present at all in the hg18 or hg19 genome assemblies. Although the hg19/GRC37 genome assembly was released to include further sequencing coverage of the human genome, no sequence representation for these chr 20 gaps was made (50). A similar approach may be applied to re-sequence and close other remaining unfinished gaps of the human genome. However, for this approach to be successful there should be no segmental duplications flanking the immediate gap region on either side. In addition, some sequence should be represented in either the publicly released Celera or HuRef genome assemblies. Although sequencing of these gaps is not trivial, a systematic approach based on a combination of techniques used here and in other studies (4,7) applied together, may help in closing the remaining unfinished human genome euchromatin gaps.

Evidence from previous studies has shown gap regions to be rich in purine and/or pyrimidine stretches that are known to form Z-DNA structures (51). However, we found long stretches of TG and CT repeats in two of the three remaining human chr 20-gap regions. The repetitive nature of these sequences render them problematic to cloning techniques including those used in the human genome project (1). Gap 1 is comprised of >500 bp of GT and >300 bp of CT continuous nucleotide repeats. Gap 2 contains long repeats, however, gap 3 contains a >400 bp CCATC stretch. Some of these sequences are present in the publicly released Celera genome assembly, suggesting the shotgun approach may have assisted in their sequencing and assembly (52). The latter is supported by our ability to align human mate pair paired-end genomic reads from multiple and independent experiments to these gap sequences. We did try to sequence the remaining >500 bp regions in gaps 1 and 3 using unmapped mates from our Illumina sequence assemblies and one other independent source (SRA accession ERX009608), however, we were unable to obtain enough coverage to close them.

We characterized five novel CpG islands within gap 2 in gDNA from PBLs and human cerebellum tissue, three of which are GC rich. We also identified CpG island 4/1a to be differentially methylated and paternally hypermethylated. Human chr 20 has several maternally imprinted genes most of which are paternally expressed (44,45). Since we found CpG island 4/1a in gap 2 to be paternally methylated, we speculate that expression at this locus may occur from its maternally inherited hypomethylated allele in a tissue-specific manner. Evidence from recent studies has shown tissue-specific expression of certain loci in the nuclear genome may be regulated by *cis*/*trans* acting ncRNAs. Emerging functional roles for ncRNAs in regulating the expression of loci highly expressed in the brain and involved in regulating pluripotency and neurogenesis is also becoming increasingly evident (53,54). Taking these observations into consideration, we searched human chr 20 gaps for ncRNAs using a combination of computational analyses and high-throughput RNA-Seq. We confirmed expression of a total of 13 (11 CMfinder/INFERNAL predicted structured ncRNAs and 2 identified using

RNA-Seq) candidate ncRNAs in these 3 human chr 20 gaps. CMfinder and INFERNAL have been used to predict conserved RNA secondary motif structures elsewhere in the human genome (38,40). We found most of these predicted ncRNAs to have structural similarity to other RNA structures transcribed elsewhere in the human genome, most of which were also found in other vertebrates.

Almost any genomic locus can theoretically fold into thermodynamically predicted RNA secondary structures, which is what RNAfold calculates. It has also been shown that secondary structure alone is statistically insignificant for the detection of putative ncRNAs (55). On the other hand, agreement in the literature exists for comparative sequence analysis and comparison to structural background as a means to discover real and/or functional ncRNAs (38,56,57). Therefore, the applied prediction method using CMfinder accounts for thermodynamics and compensatory base pair changes (A-U in human and G–C in mouse) in structurally re-aligned sequences from 18 different vertebrate genomes and compares the signal to the structural background for finding statistically significant ncRNA structures.

Quantification of ncRNA numbers 5, 6, 9 and 12 in the human brain panel identified all four to be elevated in the human cerebellum. Furthermore, we found ncRNA number 5 to be expressed in specific regions of Purkinje neuron cells in adult human cerebellum. These candidate ncRNAs are expressed at elevated levels in the brain and from within the *DLGAP4* locus coding for the brain-specific *DLGAP4* mRNA isoform a. Their elevated expression in the human brain and high clustering in certain regions of the Purkinje cells including the nucleus may in fact drive expression of *DLGAP4* mRNAs or other mRNAs in the brain. We speculate that putative ncRNAs characterized here may have roles in the regulation of disease-linked loci at 20q (9,10). Since we found candidate ncRNA numbers 12 and 13 using high-throughput RNA-Seq and not by prediction-based methods, we suggest a combination of approaches should be used to identify new putative ncRNAs in other regions of the human genome as well. Northern blot analysis using total RNA from whole-human brain and human cerebellum tissues together with randomly or end-labeled PCR amplified probes corresponding to three putative structured ncRNA numbers 5, 6 and 9 failed to reveal their size (data not shown). This may be due to the small length of each probe (<250 bp) used and the fact that this was based solely on the size of the predicted structural part. Furthermore, these ncRNAs are expressed at relatively low levels compared to mRNAs as evidenced by qPCR and may exist as multiple transcripts of varying length or spliced variants making them more difficult to detect by northern blot analysis. Difficulties associated with the confirmation of ncRNAs by northern blot analysis have previously been reported (58,59). Analyzing flanking regions on either side of these three human chr 20 gaps identified them as being conserved in primates and having evolutionary conserved elements within them. From an evolutionary perspective this area of chr 20 is dense in genes associated with development and its rather low conservation suggests functions unique to primates and especially *Homo sapiens*. Further work would be necessary to understand the specific nature and functional significance of these expressed transcripts, if any. In conclusion, our study shows that unfinished euchromatic gaps in the human genome may harbor specific epigenetic domains and contain conserved genomic elements with regulatory potential, including differentially methylated CpG-islands and other functional elements, as well as ncRNA genes, the function of which is now open for further study.

ACCESSION NUMBERS

All novel sequence obtained here has been deposited at GenBank under accessions: JN854134, JN854135 and JN854136. High-throughput sequencing data have been archived at GEO under accession GSE35405 at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=bxkrzmmaceyyqjm&acc=GSE35405

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–9 and Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

The authors thank the technical staff at Wilhelm Johannsen Centre for Functional Genome Research and Zizhen Yao and Larry Ruzzo for the development of CMfinder. S.M. is a Marie Curie Early Stage Research Fellow (*EU FP6*).

FUNDING

EU FP6 Marie Curie Research Trainning Network ''Chromatin Plasticity'' (to A.S.); University of Copenhagen, Faculty of Health Sciences partial PhD Scholarship (to S.M.); Danish Ministry of Science, Technology and Innovation and the Danish National Research Council (to A.S. and N.T.); Danish Council for Independent Research (Technology and Production Sciences); Danish Council for Strategic Research (Program Commission on Strategic Growth Technologies); Danish Centre for Scientific Computation (to J.G.); Lundbeck Foundation (to J.G., N.T. and A.S., in part). Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation. Funding for open access charge: University of Copenhagen.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

2. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**. 931–945.
3. Eichler,E.E., Clark,R.A. and She,X. (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.*, **5**, 345–354.
4. Bovee,D., Zhou,Y., Haugen,E., Wu,Z., Hayden,H.S., Gillett,W., Tuzun,E., Cooper,G.M., Sampas,N., Phelps,K. *et al.* (2008) Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.*, **40**, 96–101.
5. Sogayar,M.C. and Camargo,A.A. (2004) A Transcript Finishing Initiative for Closing Gaps in the Human Transcriptome. *Genome Res.*, **14**, 1413–1423.
6. Cole,C.G., McCann,O.T., Collins,J.E., Oliver,K., Willey,D., Gribble,S.M., Yang,F., McLaren,K., Rogers,J., Ning,Z. *et al.* (2008) Finishing the finished human chromosome 22 sequence. *Genome Biol.*, **9**, R78.
7. Garber,M., Zody,M., Arachchi,H., Berlin,A., Gnerre,S., Green,L., Lennon,N. and Nusbaum,C. (2009) Closing gaps in the human genome using sequencing by synthesis. *Genome Biol.*, **10**, R60.
8. Deloukas,P., Matthews,L.H., Ashurst,J., Burton,J., Gilbert,J.G., Jones,M., Stavrides,G., Almeida,J.P., Babbage,A.K., Bagguley,C.L. *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.
9. Tranebjaerg,L., Teslovich,T.M., Jones,M., Barmada,M.M., Fagerheim,T., Dahl,A., Escolar,D.M., Trent,J.M., Gillanders,E.M. and Stephan,D.A. (2003) Genome-wide homozygosity mapping localizes a gene for autosomal recessive non-progressive infantile ataxia to 20q11-q13. *Hum. Genet.*, **113**, 293–295.
10. Allen-Brady,K., Miller,J., Matsunami,N., Stevens,J., Block,H., Farley,M., Krasny,L., Pingree,C., Lainhart,J., Leppert,M. *et al.* (2008) A high-density SNP genome-wide linkage scan in a large autism extended pedigree. *Mol Psychiatry*, **14**, 590–600.
11. Pattengale,P., Smith,R. and Gerber,P. (1973) Selective transformation of B lymphocytes by E.B. virus. *Lancet*, **302**, 93–94.
12. Yan,H., Papadopoulos,N., Marra,G., Perrera,C., Jiricny,J., Boland,C.R., Lynch,H.T., Chadwick,R.B., de la Chapelle,A., Berg,K. *et al.* (2000) Conversion of diploidy to haploidy. *Nature*, **403**, 723–724.
13. Hertz,J.M., Sivertsen,B., Silahtaroglu,A., Bugge,M., Kalscheuer,V., Weber,A., Wirth,J., Ropers,H.H., Tommerup,N. and Tumer,Z. (2004) Early onset, non-progressive, mild cerebellar ataxia co-segregating with a familial balanced translocation t(8;20)(p22;q13). *J. Med. Genet.*, **41**, e25.
14. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
15. Korbie,D.J. and Mattick,J.S. (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat. Protoc.*, **3**, 1452–1456.
16. Halgren,C., Kjaergaard,S., Bak,M., Hansen,C., El-Schich,Z., Anderson,C.M., Henriksen,K.F., Hjalgrim,H., Kirchhoff,M.S., Bijlsma,E.K. *et al.* (2011) Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. *Clin. Genet.*, July 29 (doi: 10.1111/j.1399-0004.2011.01755.x; epub ahead of print).
17. Pasini,D., Cloos,P.A., Walfridsson,J., Olsson,L., Bukowski,J.-P., Johansen,J.V., Bak,M., Tommerup,N., Rappsilber,J. and Helin,K. (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature*, **464**, 306–310.
18. Bradley,R.K., Merkin,J., Lambert,N.J. and Burge,C.B. (2012) Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.*, **10**, e1001229.
19. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
20. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
21. Ji,H., Jiang,H., Ma,W., Johnson,D.S., Myers,R.M. and Wong,W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotech.*, **26**, 1293–1300.
22. Silahtaroglu,A.N., Nolting,D., Dyrskjot,L., Berezikov,E., Moller,M., Tommerup,N. and Kauppinen,S. (2007) Detection of microRNAs in frozen tissue sections by fluorescence in situ hybridization using locked nucleic acid probes and tyramide signal amplification. *Nat. Protoc.*, **2**, 2520–2528.
23. Don,R.H., Cox,P.T., Wainwright,B.J., Baker,K. and Mattick,J.S. (1991) Touchdown PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.*, **19**, 4008–4008.
24. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
25. Vandesompele,J., De Preter,K., Pattyn,F., Poppe,B., Van Roy,N., De Paepe,A. and Speleman,F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, **3**, RESEARCH0034.
26. Josefsen,K. and Nielsen,H. (2011) Northern blotting analysis. *Methods Mol. Biol.*, **703**, 87–105.
27. Andersen,K.L. and Nielsen,H. (2012) Experimental identification and analysis of macronuclear non-coding RNAs from the ciliate Tetrahymena thermophila. *Nucleic Acids Res.*, **40**, 1267–1281.
28. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
29. Gardiner-Garden,M. and Frommer,M. (1987) CpG Islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
30. Takai,D. and Jones,P.A. (2003) The CpG island searcher: a new WWW resource. *In Silico Biol.*, **3**, 235–240.
31. Bao,H., Guo,H., Wang,J., Zhou,R., Lu,X. and Shi,S. (2009) MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics*, **25**, 1554–1555.
32. Hudek,A.K. and Brown,D.G. (2011) FEAST: sensitive local alignment with multiple rates of evolution. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 698–709.
33. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.*, **14**, 708–715.
34. Felsenstein,J. and Churchill,G.A. (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93–104.
35. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
36. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
37. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
38. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder-a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
39. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
40. Torarinsson,E., Yao,Z., Wiklund,E.D., Bramsen,J.B., Hansen,C., Kjems,J., Tommerup,N., Ruzzo,W.L. and Gorodkin,J. (2008) Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.*, **18**, 242–251.
41. Chen,N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*, Chapter 4, Unit 4 10.
42. Mikkelsen,T., Ku,M., Jaffe,D., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T. and Koche,R. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
43. Yagi,S., Hirabayashi,K., Sato,S., Li,W., Takahashi,Y., Hirakawa,T., Wu,G., Hattori,N., Ohgane,J., Tanaka,S. *et al.* (2008) DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Res.*, **18**, 1969–1978.
44. Evans,H.K., Wylie,A.A., Murphy,S.K. and Jirtle,R.L. (2001) The neuronatin gene resides in a 'Micro-imprinted' domain on human chromosome 20q11.2. *Genomics*, **77**, 99–104.

45. Monk,D., Arnaud,P., Frost,J.M., Wood,A.J., Cowley,M., Martin-Trujillo,A., Guillaumet-Adkins,A., Iglesias Platas,I., Camprubi,C., Bourc'his,D. *et al*. (2011) Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes. *Nucl. Acids Res*., **39**, 4577–4586.
46. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res*., **31**, 28–33.
47. Gorodkin,J., Hofacker,I.L., Torarinsson,E., Yao,Z., Havgaard,J.H. and Ruzzo,W.L. (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol*., **28**, 9–19.
48. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al*. (2011) Rfam: wikipedia, clans and the "decimal" release. *Nucleic Acids Res*., **39**, D141–D145.
49. Kindler,S., Rehbein,M., Classen,B., Richter,D. and Böckers,T.M. (2004) Distinct spatiotemporal expression of SAPAP transcripts in the developing rat brain: a novel dendritically localized mRNA. *Brain Res. Mol. Brain Res*., **126**, 14–21.
50. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.-C., Agarwala,R., McLaren,W.M., Ritchie,G.R.S. *et al*. (2011) Modernizing Reference Genome Assemblies. *PLoS Biol*., **9**, e1001091.
51. Konopka,A.K., Reiter,J., Jung,M., Zarling,D.A. and Jovin,T.M. (1985) Concordance of experimentally mapped or predicted Z-DNA sites with positions of selected alternating purine-pyrimidine tracts. *Nucleic Acids Res*., **13**, 1683–1701.
52. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
53. Orom,U.A., Derrien,T., Beringer,M., Gumireddy,K., Gardini,A., Bussotti,G., Lai,F., Zytnicki,M., Notredame,C., Huang,Q. *et al*. (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46–58.
54. Ng,S.-Y., Johnson,R. and Stanton,L.W. (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J*, **31**, 522–533.
55. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
56. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
57. Gorodkin,J. and Hofacker,I.L. (2011) From Structure Prediction to Genomic Screens for Novel Non-Coding RNAs. *PLoS Comput. Biol*., **7**, e1002100.
58. Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*., **14**, 331–342.
59. Hüttenhofer,A. and Vogel,J. (2006) Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*., **34**, 635–646.