Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome

Stefan Kirsch  
*University of Heidelberg*

Birgit Weiß  
*University of Heidelberg*

Tracie L. Miner  
*Washington University School of Medicine in St. Louis*

Robert H. Waterston  
*Washington University School of Medicine in St. Louis*

Royden A. Clark  
*University of Washington - Seattle Campus*

*See next page for additional authors*

Follow this and additional works at: [http://digitalcommons.wustl.edu/open_access_pubs](http://digitalcommons.wustl.edu/open_access_pubs)

**Recommended Citation**

Kirsch, Stefan; Weiβ, Birgit; Miner, Tracie L.; Waterston, Robert H.; Clark, Royden A.; Eichler, Evan E.; Münch, Claudia; Schempp, Werner; and Rappold, Gudrun, "Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome." *Genome Research*. 15., 195-204. (2005).  
[http://digitalcommons.wustl.edu/open_access_pubs/2084](http://digitalcommons.wustl.edu/open_access_pubs/2084)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Authors
Stefan Kirsch, Birgit Weiß, Tracie L. Miner, Robert H. Waterston, Royden A. Clark, Evan E. Eichler, Claudia Münch, Werner Schempp, and Gudrun Rappold

This open access publication is available at Digital Commons@Becker: http://digitalcommons.wustl.edu/open_access_pubs/2084
Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome

Stefan Kirsch, Birgit Weiß, Tracie L. Miner, et al.

*Genome Res.* 2005 15: 195-204
Access the most recent version at doi:10.1101/gr.3302705

<table>
<thead>
<tr>
<th>References</th>
<th>This article cites 40 articles, 15 of which can be accessed free at: <a href="http://genome.cshlp.org/content/15/2/195.full.html#ref-list-1">http://genome.cshlp.org/content/15/2/195.full.html#ref-list-1</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Creative Commons License</td>
<td>This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <a href="http://genome.cshlp.org/site/misc/terms.xhtml">http://genome.cshlp.org/site/misc/terms.xhtml</a>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at <a href="http://creativecommons.org/licenses/by-nc/3.0/">http://creativecommons.org/licenses/by-nc/3.0/</a>.</td>
</tr>
<tr>
<td>Email Alerting Service</td>
<td>Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="http://genome.cshlp.org/subscriptions">click here</a>.</td>
</tr>
</tbody>
</table>

To subscribe to *Genome Research* go to: [http://genome.cshlp.org/subscriptions](http://genome.cshlp.org/subscriptions)
Interchromosomal segmental duplications of the pericentromeric region on the human Y chromosome

Stefan Kirsch,1 Birgit Weiβ,1 Tracie L. Miner,2 Robert H. Waterston,2,5 Royden A. Clark,3 Evan E. Eichler,3 Claudia Münch,4 Werner Schempp,4 and Gudrun Rappold1,6

1Institute of Human Genetics, University of Heidelberg, INF 366, 69120 Heidelberg, Germany; 2Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63108, USA; 3University of Washington, Genome Sciences, HSB K336B, Seattle, Washington 98195, USA; 4Institute of Human Genetics, University of Freiburg, 79106 Freiburg, Germany; 5University of Washington, Genome Sciences, HSB K357, Seattle, Washington 98195, USA

Basic medical research critically depends on the finished human genome sequence. Two types of gaps are known to exist in the human genome: those associated with heterochromatic sequences and those embedded within euchromatin. We identified and analyzed a euchromatic island within the pericentromeric repeats of the human Y chromosome. This 450-kb island, although not recalcitrant to subcloning and present in 100 tested males from different ethnic origins, was not detected and is not contained within the published Y chromosomal sequence. The entire 450-kb interval is almost completely duplicated and consists predominantly of interchromosomal rather than intrachromosomal duplication events that are usually prevalent on the Y chromosome. We defined the modular structure of this interval and detected a total of 128 underlying pairwise alignments (≥90% and ≥1 kb in length) to various autosomal pericentromeric and ancestral pericentromeric regions. We also analyzed the putative gene content of this region by a combination of in silico gene prediction and paralogy analysis. We can show that even in this exceptionally duplicated region of the Y chromosome, eight putative genes with open reading frames reside, including fusion transcripts formed by the splicing of exons from two different duplication modules as well as members of the homeobox gene family DUX.

Last year the male-specific sequence of the human Y chromosome was announced (Skaletsky et al. 2003). Determining the sequence of the human Y was an enormous task due to the rather highly repetitive genomic landscape of this chromosome shaped by an extraordinary evolutionary history. The majority (41 Mb) of the entire chromosome (63 Mb) is comprised of three blocks of highly reiterated satellites as well as other repeat sequences. Even the 23-Mb male-specific euchromatic region appears to have an unusual genetic structure with very large gene-rich palindromes. Near-perfect sequence duplications appear to preserve their structural integrity due to gene conversion events. Within this complex genomic environment, the chromosome generated its present gene repertoire by gene acquisition from the autosomes and the X chromosome, followed by selective gene amplification.

A surprising finding from the analysis of the human genome sequence was that 5% of our genome consists of segmental duplications. The size, fraction, and degree of sequence identity of these segmental duplications are unique to the human genome. Analysis of these regions has also revealed that they are composed of DNA containing partial copies or complete genomic sequences composed of exons and introns. Indeed, these highly active regions have also been demonstrated to be the birthplace of novel genes (Samonte and Eichler 2002). We have isolated and sequenced a 554-kb genomic segment from the human Y chromosome that contains a 450-kb euchromatic island hidden between satellite 3 sequence, adjacent to the centromere on the long arm of the chromosome. Analysis of the Y-specific pericentromeric sequence revealed that it is almost entirely composed of blocks of duplicated genomic segments sharing 95%–99% homology to multiple human autosomes. Segmental duplications with nearly identical sequence of this range have been detected throughout the genome within exceptional, mostly pericentromeric regions (Horvath et al. 2000b; Bailey et al. 2002; Eichler et al. 2004; Rudd and Willard 2004; She et al. 2004). Here, we provide a detailed analysis of the structural composition of this euchromatic island in pericentromeric Yq11. The extent and degree of homology between this region and paralogous segments elsewhere in the genome were evaluated and the mosaic modular structure defined. Eight putative genes with open reading frames have been identified.

Results

The pericentromeric region in Yq11

The Y chromosomal centromere is essentially composed of aliphid DNA surrounded by a range of other satellite and nonsatellite repeated sequences in a complex region spanning several hundred kilobases. To analyze the region next to the centromere unambiguously, we generated several Y-specific sequence-tagged sites (STSs) (Kirsch et al. 2002). STS SKY1 (AJ487121) is contained...
within the P1-derived artificial chromosome clone (PAC) RP1-85D24 (AC134013). The Y-specific STS SKY2 (AJ487122) was identified within bacterial artificial chromosome clone (BAC) RP11-295P22 (AC134879). Additional Y-specific STTs, SKY5–7, were generated by sequencing either BAC- or PAC-end fragments as well as internal sequences from YAC 17C12C (Kirsch et al. 2002). Using RP1-85D24 and RP11-295P22 as seed clones, we set out to close a gap of unknown size by a combination of PCR-based screening and hybridization to a single PAC library generated from one male individual, to avoid allelic variation (Fig. 1C). PAC clones RP1-131M06 (AC134878) and RP11-886I11 (AC134882) overlap with RP1-85D24 and RP11-295P22 and completely cover the genomic region between RP1-85D24 and RP11-295P22. The entire region of overlap between RP1-85D24 and RP11-322K23 (the most proximal clone in the Y-chromosomal clone-based map; Tilford et al. 2001) is restricted to satellite sequences, and proof of overlap was confirmed by sequence-family variant (SFV) typing (for details see Kirsch et al. 2002). This method relies on subtle variations as characteristic features of closely related but nonallelic sequences. Satellite probes from a previous pulsed-field-derived map (Cooper et al. 1992) were subsequently used to verify the integrity of the contig. Attempts to extend the contig towards the centromere resulted in the identification of heavily unstable genomic clones which consisted almost exclusively of satellite sequences. A minimum tiling path for sequencing (Fig. 1C) consisting of three BAC clones and one PAC clone was determined. Sequencing was carried out at the Washington University Genome Sequencing Center (WUGSc, St. Louis, Missouri). The four genomic clones include Y-specific STSs (Kirsch et al. 2002). Finished BAC clones share a 100% identity over their entire regions of overlap, whereas PAC clone RP1-85D24, which was derived from a different library (individual), shares 99.8% sequence identity (RP1-85D24 ↔ RP11-131M06 71279 bp; RP1-131M06 ↔ RP11-886I11 33248 bp; RP11-886I11 ↔ RP11-295P22 10705 bp). The entire overlap of 10417 bp between RP11-295P22 and RP11-322K23 consists of satellite 3 repeats. Sequences were assembled to form a contiguous sequence of 554,625 bp. Comparison of the complete sequence with the current human genome assembly of the National Center for Biotechnology Information (NCBI) indicated that this pericentromeric Yq11 contig is not part of the publicly available Y chromosome sequence. To determine whether this region reflects a low-frequency polymorphism or a constant part of the human Y chromosome, we studied 100 male individuals from different ethnic backgrounds for the presence of the Y-derived STSs SKY1, SKY2, and DUXY. All tested male individuals were scored positive for all three markers, whereas none of the markers was found in female controls.

**Segmental duplications in the pericentromeric Yq11 region**

To investigate the molecular and chromosomal structure of this region, we examined the duplication content of the 554-kb segment by comparing it to the human genome (Build 34). To detect extensive internal and external pairwise chromosomal similarities, whole genome assembly comparison (WGAC) was used whereby simultaneously large gaps or insertions/deletions within the DNA were bridged (Bailey et al. 2001; see Methods). This analysis revealed that 80.2% (444,601/554,625 bp) of the pericentromeric sequence was composed of segmental duplications. In addition, 73.8% (409,187 bp) and 5.3% (29,289

---

**Figure 1.** Location of a euchromatic island flanked by centromeric satellite 3 repeats on the long arm of the Y chromosome. (A) Minimum tiling path for sequencing the Y chromosome as published by Tilford et al. (2001). (B) Enlarged view of the genomic region encompassing the centromere and satellite 3 repeats (Tilford et al. 2001). (C) Illustration of the P1-derived artificial chromosome (PAC) and bacterial artificial chromosome (BAC) clones assembled into the pericentromeric Yq11 contig. Blue lines indicate name and sequence length of respective clones. Clone names include library origin; accession nos. are in parentheses. PAC RP1-85D24 extends 2 kb into the satellite 3 sequence block forming a constant part of the human Y chromosome centromere (Tyler-Smith 1987). The overlap sizes of the clones are as follows: RP1-85D24 ↔ RP11-131M06 71,279 bp; RP11-131M06 ↔ RP11-886I11 33,248 bp; RP11-886I11 ↔ RP11-295P22 10,705 bp; RP11-295P22 overlaps by 10,417 bp with RP11-322K23, the most centromeric clone presented in Tilford et al. (2001). The distal half of RP11-295P22 consists exclusively of satellite 3 repeat sequence. Subtracting satellite 3 segments from the entire 554-kb sequence discloses previously unknown 450 kb of euchromatic DNA sequence.
The Yq11 pericentromeric region

Modular structure and gene content of the pericentromeric region in Yq11

Successive rearrangements account for the complex structure of genomic segments consisting of segmental duplications. To define the modular structure of pericentromeric Yq11, we set out to trace back each distinct sequence block of the 554-kb sequence and its paralogous counterparts on other chromosomes to a common ancestral progenitor. Due to the mosaic structure of this region of the human Y chromosome, the definition of modules based on the identification of junctional boundaries shared with other chromosomes alone is not sufficient to decode its patchwork structure completely. Because gene sequences were used successfully in previous analyses to delineate modules in more complex duplications (Eichler et al. 1996; Shaikh et al. 2000), we used the same strategy for the pericentromeric Yq11 region. In total, we defined 37 modules that were distributed inter- (36 modules) and intrachromosomally (one). All modules were present only once in the Yq11 region. Twenty-nine modules were identified solely on the basis of well demarcated junctional boundaries. Ten of them presented one genic sequence, whereas four showed two genic features derived from different ancestral loci. Two further genic features were shown to spread across junctional boundaries. No transcriptional activity was documented among these 20 duplication modules. Several degenerative processed pseudogenes have been detected in Yq11 with multiple copies in other pericentromeric regions.

Of the 20 identified genic sequences sharing significant homology to the 554-kb sequence, all are present in at least one other location in the human genome. The principle gene-related features of each genic sequence are summarized in Table 2. Direct comparison with the NCBI dbEST database by best expressed sequence tag (EST) placement showed significant nucleotide identity to the Y chromosome.

Degenerated processed pseudogenes and genes with partial exon–intron structure

Of the 20 gene segments found on Yq11, 13 are unlikely to be functional genes. Eight have features of degenerated processed pseudogenes, and five genes show only partial exon–intron structure (Fig. 4B,C). All of them were derived from more complete genes elsewhere in the genome and propagated as part of segmental duplications to the human Y chromosome. They present an overall nucleotide sequence identity of 82%–97%.

A striking example is the processed pseudogene of asparagine synthetase (ASNS). This processed pseudogene consists of a proximal part encompassing exons 1 and 2, and a distal part comprising exons 5 and 6 of the functional ASNS gene. Whereas the distal part has a nucleotide identity of 96%, the proximal part has merely 85%. Thus, the fragments of the ASNS processed pseudogene result from temporally different retrotranspositions subsequently juxtaposed through the process of paralogous recombination.

A second interesting case involves a...
Y-chromosomal member of the recently identified ARP3\/*H9252 pseudogenes-derived gene family locating to different chromosomes (FKSG72–74) (Fig. 4C). Genic sequence 17a, comprising only a portion of the degenerated processed pseudogene 17, may therefore also represent a functional gene. It has experienced a one-base pair deletion resulting in an altered carboxy-terminal portion after amino acid 88. This frameshift leads to a premature termination of the open reading frame (ORF), resulting in a protein of only 107 amino acids. All paralogs are predicted to encode single-exon ORFs of 193–199 amino acids with 98%–100% of identity to the well defined ARP3\/*ORF (transcripts 17 and 17a; Fig. 4C).

ESTs and candidate genes

The other genic segments are likely to be functional. Besides two EST clusters (UniGene EST Al6758041 and TIGR THC1666755) with unknown function, several interesting candidate genes were analyzed in more detail.

The four paralogous mRNAs FLJ39633, FLJ00219, FLJ35473, and pp5644 with a nucleotide identity between 94% and 100% are weakly homologous to tektin A1 (A46170), a cytoskeletal protein from Strongylocentrotus purpuratus (Norrander et al. 1992). SIM4 analyses revealed that FLJ39633 (AK096952) is composed of eight exons (Fig. 4B,C), whereas the remaining three mRNAs only contain parts of paralogous copies on other chromosomes. The genomic structure discloses that exons 1 and 2 and exons 3 to 5 reside within a few kilobases, whereas the distance between exon 2 and 3 is more than 50 kb. Remarkably, this intron extends over five different defined modules, of which three were identified by the presence of genic sequences. All three internal genic sequence paralogs show an inverted transcriptional orientation relative to the tektin A1-homologous mRNAs, thereby excluding the incidental incorporation into a growing transcript.

We also found four copies of the DUX (double homeobox) gene family (Fig. 4). DUX genes encode single-exon ORFs (Beckers et al. 2001) with highest homologies to the paired-type homeobox genes PAX3 and PAX7. The Y-derived copies (DUXY1–
The Yq11 pericentromeric region

DUXY4) are organized as a tandem repeat (Fig. 5). The DUXY gene family members are predicted to encode ORFs of 110 amino acids with conserved amino termini, including the first homeodomain (Fig. 6). None of the Y-derived copies encodes a complete second homeodomain. A nuclear localization signal was defined in the amino terminal part of the paired-type homeodomain in several proteins. A similar stretch of basic amino acids is still present in all DUXY genes. A TATA box and a transcription start site were found upstream of the predicted translation start site of DUXY1–4. There is no poly(A) signal in DUXY genes, which might explain the absence of DUX mRNA sequences in EST databases.

Table 1. Sequence and comparative FISH results for interchromosomal duplications

<table>
<thead>
<tr>
<th>Clone Library</th>
<th>Beginning (bp)</th>
<th>Ending (bp)</th>
<th>WGAC</th>
<th>FISH</th>
</tr>
</thead>
<tbody>
<tr>
<td>8SD24 RPCI-1</td>
<td>1</td>
<td>155655</td>
<td>1q21, 2q11, 2q21, 3p13, 4p12, 9p11, 9q13, 9q22, 11q13, 14cen, 15q11, 16p11, 21q11, 22q11</td>
<td>1p11/p12, 1q12/q21, 1q44, 2cen, 2q14.3/q21, 3cen, 4cen, 9p11/p12, 9q12/q13, 10cen, 11q13, 13cen, 13p11.2, 13p13, 14cen, 14p11.2, 14p13, 15cen, 15p11.2, 15p13, 20cen, 21cen, 21p11.2, 21p12, 22p11.2, 22p13</td>
</tr>
<tr>
<td>131M06 RPCI-11</td>
<td>84376</td>
<td>240189</td>
<td>1q21, 2q11, 2q21, 3p13, 4p12, 9q13, 14cen, 15q11</td>
<td>1q12/q21, 1cen, 2q14.3/q21, 3cen, 3p12, 14cen, 14p11.2, 14p13, 15cen, 15p11.2, 15p13, 20cen, 21cen, 21p11.2, 21p12, 22p11.2, 22p13</td>
</tr>
<tr>
<td>886I11 RPCI-11</td>
<td>206941</td>
<td>359595</td>
<td>1q21, 2q11, 3p13, 4q23, 10p11, 16p11, 17q11, 18cen, 21q11, 22q11</td>
<td>1q12/q21, 1cen, 3cen, 4cen, 4q22–24, 9cen, 9q12/q13, 10cen, 13cen, 13p11.2, 13p13, 14cen, 14p11.2, 14p13, 15cen, 15p11.2, 15p13, 15cen, 16cen, 21cen, 21p11.2, 21p12, 22p11.2, 22p13</td>
</tr>
<tr>
<td>295P22 RPCI-22</td>
<td>348890</td>
<td>554625</td>
<td>1q21, 2p11, 2q21, 3p26, 4q23, 7p11, 7q11, 9p11, 9q13, 10p11, 11cen, 15q11, 16p11, 17cen, 18p11, 21q11, 22q11, 22q12, Yq11</td>
<td>1q12/q21, 3cen, 9p11/p12, 9q12/q13, 10cen, 13cen, 13p11.2, 13p13, 14cen, 14p11.2, 14p13, 15cen, 15p11.2, 15p13, 15cen, 16cen, 21cen, 21p11.2, 21p12, 22p11.2, 22p13</td>
</tr>
</tbody>
</table>

Table 2. Genic paralogies in pericentromeric Yq11

<table>
<thead>
<tr>
<th>Region of homology</th>
<th>Gene</th>
<th>Name</th>
<th>GenBank acc. no.</th>
<th>Ancestral locus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degenerated processed pseudogenes</td>
<td>55219-58321 FLJ42128</td>
<td>Testis cDNA clone</td>
<td>AK124122</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>68740-70354 LOC339742</td>
<td>Image cDNA clone</td>
<td>BC045732</td>
<td>Chr 2</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>100202-106214 ASNS</td>
<td>Asparagine synthetase</td>
<td>NM_133436</td>
<td>Chr 7q21</td>
<td>Arfin et al. 1983</td>
</tr>
<tr>
<td></td>
<td>142459-148130 FLJ35140</td>
<td>Kazusa cDNA clone</td>
<td>AK092459</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>164406-174880 FLJ00310</td>
<td>Kazusa cDNA clone</td>
<td>AK090412</td>
<td>Chr 1</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>296713-299847 PABPC1</td>
<td>Poly(A) binding protein, cytoplasmic 1</td>
<td>NM_002568</td>
<td>Chr 8q23</td>
<td>Grange et al. 1987</td>
</tr>
<tr>
<td></td>
<td>357596-359239 ARP3</td>
<td>Actin-related protein 3-beta</td>
<td>NM_020445</td>
<td>Chr 7q32-36</td>
<td>Machesky and Goud 1999</td>
</tr>
<tr>
<td></td>
<td>407551-409679 TRIM43</td>
<td>Tripartite motif-containing 43</td>
<td>NM_013800</td>
<td>Chr 2</td>
<td>Unpublished</td>
</tr>
<tr>
<td>ESTs</td>
<td>32480-33082 Hs.252460</td>
<td>UNIGENE EST cluster</td>
<td>/</td>
<td>Chr 11</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>236680-242264 THC166675</td>
<td>TIGR EST cluster</td>
<td>/</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Genes with partial exon-intron structure</td>
<td>46985-47778 AF038169</td>
<td>IMAGE cDNA clone</td>
<td>BC043584</td>
<td>Chr 1</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>134919-135544 C21orf81</td>
<td>Chromosome 21 unknown ORF 81</td>
<td>AF426257</td>
<td>Chr 21</td>
<td>Reymond et al. 2002</td>
</tr>
<tr>
<td></td>
<td>302288-347418 LOC1S0159</td>
<td>CG10806-like IMAGE cDNA clone</td>
<td>NM_139173</td>
<td>Chr 4</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>374270-380489 CHEK2</td>
<td>CHK2 checkpoint homolog</td>
<td>NM_007194</td>
<td>Chr 22q12</td>
<td>Matsuoka et al. 1998</td>
</tr>
<tr>
<td></td>
<td>435493-435986 MGCl32713</td>
<td>IMAGE cDNA clone</td>
<td>BC034141</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Potential coding gene</td>
<td>115523-120823 FLJ00219</td>
<td>Kazusa cDNA clone</td>
<td>AK074146</td>
<td>Chr 13</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>115523-176727 FLJ35473</td>
<td>Kazusa cDNA clone</td>
<td>AK092792</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>115523-176733 FLJ39663</td>
<td>Kazusa cDNA clone</td>
<td>AK096952</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>115520-176692 pp5644</td>
<td>cDNA clone</td>
<td>AF289559</td>
<td>Unknown</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td>268641-269264 DXU1</td>
<td>Double homeobox, 1</td>
<td>NM_012146</td>
<td>4q35</td>
<td>Ding et al. 1998</td>
</tr>
<tr>
<td></td>
<td>276644-276976 DXU1</td>
<td>Double homeobox, 1</td>
<td>NM_012146</td>
<td>4q35</td>
<td>Ding et al. 1998</td>
</tr>
<tr>
<td></td>
<td>283281-283613 DXU1</td>
<td>Double homeobox, 1</td>
<td>NM_012146</td>
<td>4q35</td>
<td>Ding et al. 1998</td>
</tr>
<tr>
<td></td>
<td>294054-294476 DXU1</td>
<td>Double homeobox, 1</td>
<td>NM_012146</td>
<td>4q35</td>
<td>Ding et al. 1998</td>
</tr>
<tr>
<td></td>
<td>357403-359228 FKSG74</td>
<td>FKSG74</td>
<td>AF026352</td>
<td>Chr 16</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

The position of each feature within the 554kb contiguous sequence is given, together with the accession number and EST cluster information. Unknown ancestral locus indicates either that BAC clones corresponding to that sequence are not chromosomally assigned or a corresponding genomic sequence is not available. References and ancestral locus of each genic paralogy is shown.
Fundamentally, the pericentromeric region in Yq11 can be subdivided into satellite 3 sequences and a euchromatic island flanked by these repeats. The presence of the satellite 3 sequence was interpreted by Skaletsky and colleagues (2003) as the end of the euchromatin, but new additional Y-specific markers revealed this euchromatic island. Investigating this euchromatic island encompassed by satellite sequences has illuminated its complex structure and the dynamic history of sequences located in this region. We have characterized all the segmental duplications and provide a genome-wide view of the results. The comparison of FISH and sequence homology analysis of this region strongly suggests an underrepresentation of pericentromeric regions of the acrocentric chromosomes in the current human genome sequence.

Ninety-three percent of the sequenced euchromatic island was shown to be involved in segmental duplications (90% identity and 1 kb in length). Its highly duplicative nature was used to organize the genomic segment into minimal evolutionary shared segments (modules) and to assess the transcriptional potential of the duplications. Our analysis shows quite a striking correspondence to the whole-chromosome study by Bailey et al. (2002). The balanced distribution of pairwise genetic-distance estimates (K values) supports the observation of frequently occurring duplications over the past 30 million years of evolution. The
The Yq11 pericentromeric region

There are only quantitative differences between different chromosomes, including the Y. Our work shows in detail that part of the pericentromeric region of the Y chromosome consists of a mosaic of interchromosomal segmental duplications, indicating that this region is typical of autosomal pericentromeric regions.

With the exception of the DUXY1–4 genes, segmental duplications of Yq11-paralogous sequences have been involved in the multiplication of all genes sequences in this region. Among them, we detected two genes with intact ORFs (transcripts 7 and 17a), five genes with partial exon–intron structure, eight degenerated processed pseudogenes, and two EST clusters with an unidentified ORF. Members of one gene family could only be detected by direct homology comparison using the nr database of NCBI, as their genomic environment does not reflect the characteristics of a segmental duplication. The fate of the duplicated genes sequences is consistent with the birth-and-death model of gene evolution. Beyond that, the existence of genes possessing an intact ORF in Yq11-pericentromeric-paralogous segments supports the hypothesis that these duplications have an important evolutionary impact on functional change (Nei et al. 1997).

The homeobox-containing DUX gene family cluster in pericentromeric Yq11 is the first completely sequenced cluster of DUX genes in the human genome. The first members of the DUX gene family were identified on distal 4q (Ding et al. 1998), and paralogous gene clusters have been mapped to chromosomes 10, 13, 14, 15, 21, and 22 in the meantime. Proof of active transcription of the DUX genes has been provided for six family members (DUX1–5, -10). Similar to DUX2, none of the four Y-chromosomal DUX copies preserves a complete second homeodomain. Nevertheless, the high nucleotide sequence identities of these gene candidates warrants expression profiling to determine whether the Y-specific DUX copies adopted a distinct tissue-specific expression pattern. This is especially intriguing for DUXY1, as its predicted carboxy terminus presents striking differences from all other members, although the homeodomain I is extremely highly conserved. It will be interesting to find out if this carboxy terminus is of particular importance to the putative protein. One of the DUX gene family members, DUX4, has been suggested to play a role in falciparum hemoparasite pathogenesis (FISH) (Gabriëls et al. 1999).

By investigating the FKS72–74 gene family and the tektin A1 homologous genes, two interesting features contributing to the continuing process of gene duplication and divergence were observed: First, the development of a possible functional FKS72–74 gene family member (transcript 17a) from a degenerated retrotransposed pseudogene of ARPS (transcript 17), and second, the integration of part of the distal end of a degenerated processed pseudogene FLJ00310 (transcript 8) into exon 1 of the tektin A1 homologous genes. A similar observation of exon usage in reverse transcription orientation was presented by Bailey et al. (2002). Actively transcribed retrotransposed pseudogenes might furthermore be involved in the regulation of their functional
progenitors (Hirotsune et al. 2003). As the human genome sequences of pericentromeric regions will eventually approach completion, it will be interesting to find out whether these genomic alterations are intrinsic evolutionary features of paralogous sequence blocks.

All potential genes described here are embedded in regions with extensive homologies to each other. Such highly duplicated regions with a locally increased number of recombination-promoting CAGGG repeats are prone to chromosomal rearrangements (Ji et al. 2000; Samonte and Eichler 2002; Stankiewicz and Lupski 2002). These rearrangements might lead to an abnormal phenotype by disruption or dosage alteration of the involved genes. Therefore, investigation of the eight potential genes as candidates for human disease mapped to paralogous sequences of nonhomologous chromosomes might shed new light on human disorders. Especially, the search for the yet unknown stature (GCY) (Smith et al. 1985; Ogata and Matsuo 1992) and gonadoblastoma locus on the Y chromosome (GBY) (Page 1987; Salo et al. 1995; Tsuchiya et al. 1995) might profit from the putative gene sequences provided in this study.

Methods

Generation and analysis of pericentromeric Y-specific STSs

Y-specific STSs termed SKY1, 2, 5, 6, 7 were derived from YAC, BAC, and PAC end sequences or from clone-internal sequences amplified by various combinations of Alu primers (Kirsch et al. 2002).

Subjects and PCR analysis

Subjects were unrelated healthy male individuals belonging to various Central European and Mediterranean populations. Additional samples from the Middle East and Far East were also included. Genomic DNA samples were extracted from peripheral blood leukocytes using standard protocols. PCR cycling conditions and analysis were as described by Kirsch et al. (2002). Primer sequences (5′-H11032–3′/H11032) for DUXY are as follows: DUXYfor, CCGACACCTTCGGACAGCAC; DUXYrev, GTGGTCTGGGATCCGGTGAC.

Isolation and sequencing of chromosome Y clones

Large-insert genomic PAC and BAC clones were identified through screening of PCR pools from the RPCI1,3–5 and RPCI11 male human genomic libraries provided by the German Resource Center (RZPD). From a total of 20 positive clones, four were selected to form the minimum tiling path for sequencing, namely RP1-85D24 (AC140113), RP11-131M06 (AC134878), RP11-886111 (AC134882), and RP11-295P22 (AC134879). Genomic sequencing was carried out by conventional high-throughput sequencing techniques. Finished sequences from overlapping clones were assembled into one contiguous sequence of 554,625 bp.

Detection of segmental duplications

We used whole-genome assembly comparison (WGAC; Bailey et al. 2001) to detect segmental duplications within the contiguous

Figure 6. Comparison of predicted amino-acid sequence of cDNAs and paralogous genomic copies of double homeobox DUX-like genes. DUX1, DUX2, DUX3, DUX4, DUX5, and DUX10 represent human double homeobox-containing genes from 4q35. All genes consist only of a single exon. The color code corresponds to the CLUSTALW default for amino-acid sequence comparison. The boxes indicate the 60-amino-acid conserved homeodomain. Analogous to the DUX gene family member DUX2, none of the Y-derived DUX family members contain a complete copy of the second homeodomain (Homeobox II). The location of a 1-bp deletion in DUXY1 relative to all other family members is indicated, resulting in a frameshift and a C-terminally altered amino acid sequence (purple). We resequenced the DUXY1 copy from a PCR product amplified from BAC RP11-886111 (AC134882) and a normal male individual and confirmed the accuracy of the 1-bp deletion. Black stars indicate stop codons.
Fluorescence in situ hybridization (FISH)

FISH analysis of chromosomal metaphase spreads derived from lymphocytes was performed from two different human males. Prior to FISH, the slides were treated with RNase followed by pepsin digestion as described (Ried et al. 1992). FISH followed the method described by Scheimpf (1995). Chromosome in situ suppression was applied to the clones from the Human Male Genome PAC library (RPCI1) and from the BAC library (RPCI11) (RPCI1-85D24, RPCI11-131M06, RPCI11-886111, RPCI11-295P22). The probes D4Z1 and D9Z1 (Appligene Oncor) were used as a marker for the centromere of chromosome 4 and for the centromere and heterochromatic region of chromosome 9. After FISH the slides were counterstained with DAPI (0.14 µg/mL) and mounted in Vectashield (Vector Laboratories). Preparations were evaluated using a Zeiss Axioshot epifluorescence microscope equipped with single-bandpass filters for excitation of red, green, and blue (Chroma Technologies). During exposures, only excitation filters were changed, allowing for pixel-shift-free image recording. Images of high magnification and resolution were obtained using a black-and-white CCD camera (Photometrics Kopak KAF 1400; Kodak) connected to the Axiophot. Camera control and digital image acquisition involved the use of an Apple Macintosh Quadra 950 computer.

Potential gene content and module definition

After in silico detection and manual correction, a BLAST analysis of the nonredundant part of the 554-kb segment versus the full-length NCBI Locus Link/reference sequence (RefSeq) plus UniGene and TIGR human transcripts was carried out. Transcripts with high homologies to the Y-chromosomal sequence were selected for further analysis. Transcripts derived from specific autosomal regions that differed in exon composition were analyzed in detail. A total of 24 genic sequences were extracted from the Y-specific sequence, and the most likely allelic loci of the transcript in the human genome were identified using BLASTN. Exon-intron structures were delineated by Sim4. The underlying genomic sequence of the allelic loci was RepeatMasked and reblasted against the 554-kb segment to characterize its modular structure.

Acknowledgments

The human genomic PAC and BAC libraries used in this work were constructed at the RPCI in Buffalo, NY. Clones isolated from these libraries were purchased from the same institution. Satellite probes p22hom48.4, pHY10, pkFC68, pkFC52, pkFC11, pkFC37, and pkFC43 were kindly provided by Chris Tyler-Smith (Oxford). This work was supported by a grant from Pharmacia AB, Stockholm and the Deutsche Forschungsgesellschaft (Ra 380/10-1).

Note added in proof

According to the current human genome assembly (NCBI Build 35, May 2004), the sequenced portion of the Y chromosome must be inserted at position 12.15 Mb of the human Y chromosome (in between the proximal block of satellite 3 sequences and the assembled contig NT_011875). The genomic clone RP11-322K23 connects our contig to NT_011875.

References


