HUMAN GENOME – FROM PIECES TO PATTERNS

Meena Kishore Sakharkar 1, Pandjassarame Kangueane 1, Bagavathi S. Perumal 1, Vincent Tak Kwong Chow 2, Eric Sorscher 3, Kishore R. Sakharkar 4 and Aubrey Hill 3

1 NCSV, MAE, Nanyang Technological University, Singapore 639798, 2 Department of Microbiology, National University of Singapore, Singapore 119260, 3Department of Medicine, University of Alabama at Birmingham, 796 Birmingham, AL 35294-0005, USA, 4BioInformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Materials and methods
4. Results & Discussion
   4.1. From pieces to patterns by retroposition
   4.2. From pieces to patterns and gene assemblage
5. Conclusion
6. Caveats
7. Acknowledgements
8. References

1. ABSTRACT

A profile of exon-intron lengths in genes shows a normal distribution. This observation suggests that different genes may have portions of their total exon and/or intron lengths in common. In order to explore the common exon-intron structural patterns that may arise due to common lengths across genes, we compared the exon-intron length patterns of annotated human genes. We discovered 1762278 conserved arrangements of exon-intron length across the otherwise unrelated and diverse genomic landscape. The existence of common exon-intron length patterns across unrelated genes suggests for their role of in gene assemblage and human genome design and architecture.

2. INTRODUCTION

The boundaries between exons and introns are blurring (1). Researchers have shown how dynamic the genome is and how the repetitive elements of so-called junk DNA are actually, as Makalowski puts it, "a genomic treasure" and "a source of 'ready-to-use-motifs' " increasing an organism's evolutionary flexibility (1). The structural and functional properties of DNA change as a function of its nucleotide composition. The human genome has been described as exon islands in a vast sea of introns (2). The size and prevalence of introns (about 25%) in more complex organisms suggests that introns could be important functional elements in large genomes (3). In order to understand the structure and evolution of genes
Patterns in human genome

and genomes, it is important to know the general statistical characteristics of the exon-intron structures. The first compilation of exon-intron structures in eukaryotic genes was published by Hawkins in 1988 (4). Since then many disparate reports have been presented and the use of patterns in exons and introns to understand gene structure is becoming increasingly ubiquitous. Also, the sequential arrangement of coding (exons) and non-coding (introns) regions is of particular interest from a biological viewpoint in revealing essential details necessary for understanding the assembly of the spliceosome and the splicing process in general. Recently, it was reported that different genes have portions of their total exon-intron sequential structure in common. The analysis reported more than 200 patterns of length 2 (length 2 implies a block of exon and an intron) or greater among the 72 human genes (5). The data is novel and needs further analysis at genome level. Exploration of such patterns at the genome level will confirm their factual nature and provide clues to their role in genome design and gene architecture. Due to the size and complexity of human genes, it has not been possible in the past to determine whether common exon-intron size patterns exist among different genes by visual inspection. Our goal in this exploration is to get a list of all such conserved patterns of length two or more and to compare the exon-intron patterns across the human genome landscape representing the length of the exons and introns as integers while ignoring the underlying nucleotide sequence (5). This allows us to study the distribution of exon-intron blocks in human genes to determine whether one can detect the reuse of exon-intron sequences and to use the frequency of such reuse to estimate how many ancestral exon-intron block sequences there might have been. This study is imperative for the theoretical study of the origin and evolution of genes and genomes. These findings could help improve gene structure prediction by computational methods by providing better understanding of factors that govern genome assembly and design.

3. MATERIALS AND METHODS

The exon-intron structure of human genes was obtained from human genome data downloaded from NCBI build 34 and AceView (http://www.ncbi.nlm.nih.gov/AceView/) database (6). The length, in nucleotides, for each exon or intron was alternately entered as an element of an array representing the structure of that gene. The arrays were indexed such that the first element was designated exon (starting at '0'), the second element as intron, and so on. This scheme resulted in the exons residing in the even numbered elements of the array and the introns residing in the odd-numbered elements of the array. The array representing the exon-intron structure of the human FUCA1 and ADH5 gene is presented in Table 1. Pair wise comparisons of 27115 human genes (246632 exons and 218575 introns) were performed. This involved $7.35 \times 10^9$ gene pair comparisons. Each of the ‘exon and the subsequent intron’ or ‘intron and subsequent exon’ which we call a “block of length 2” from each gene was compared to other “blocks of length 2” from all annotated genes throughout the database, to identify pairs of genes that have at least one block of length 2 matched in length (99% similar in length). The lengths were extracted based on CDS feature as described earlier (7). Extensions of block lengths from 3 to >25 were computed and tabulated (Table 2). The identified pairs were further categorized into subgroups based on the start of the match with exon or intron, and whether the genes are on the sense or anti-sense strand (Table 3). Redundancy in blocks of length 2 to >25 was computed (at 99% of components length) to identify pairs of unique length patterns (Table 4). Each of the entries from genome data were matched with corresponding Aceview entry by using Gene name and the exon/intron lengths were identified in any of the splice variants in Aceview and the lengths were categorized as confirmed components. AceView is a database that offers an integrated view of the human genes as reconstructed by alignment of all publicly available mRNAs and ESTs on the genome sequence. These confirmed components were mapped back onto the initial pool of blocks of length 2 to >25 so as to identify the pairs that have Aceview validated block lengths (Table 4). A dataset on new patterns of blocks of length 2 from higher block lengths were generated and categorized (Table 5). Similar analysis was performed to identify unique blocks of length 3 and tabulated (Table 6).

4. RESULTS & DISCUSSION

4.1. From pieces to patterns by retroposition

The Human genome contains 246632 exons and 218575 introns from 27115 gene sequences (build 34 version 3.0). 1762278 patterns of “blocks of length 2” were identified across the genomic landscape that had at least 1 exon and 1 intron length in common (99% similar in length ). Distributions of the total number of pairs based on the number of matched exon-intron lengths they contain show a wavelet distribution (Figure 1). The number of genes that share blocks of length 3 is 16380, this number decreases to 637 when the block size is 4 and then increases again to 1335 for 5 (Table 2). This pattern is observed in both the sense and the anti-sense strand (Table 3). Such alternate patterns suggest that odd number pairs are more frequent than even number pairs. This suggests that similar components are more preferred at the boundaries. For example, it is more frequent to find a block of length 5 with exon-intron-exon-intron-exon than to find a block of length 4 e.g. exon-intron-exon-intron. Also, pairs beginning with an exon and ending in exon were found to be more frequent than the ones beginning with introns and ending with introns (Table 3). Though, the dynamics of coding regions are expected to be more correlated with functional complexity and diversity than are the dynamics of non-coding regions, it suggests that the dynamics of coding regions are not independent of non-coding regions. However, the possible reasons for such occurrences are obscure.

Interestingly, genes that share blocks of length 4 or more lie on the same chromosome 90% of the time (Table3; Figure 2). The prevalence of similar blocks on same chromosome suggests for their evolution by duplication and divergence. These results are in accordance with previous findings and suggest that the dynamics of
Table 1. Gene comparison matrix for FUCA1 and ADH5

<table>
<thead>
<tr>
<th>Position</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (bp)</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>12</td>
<td>0.032</td>
<td>0.089</td>
<td>0.087</td>
<td>0.113</td>
<td>0.060</td>
<td>0.063</td>
<td>0.713</td>
<td>0.063</td>
<td>0.120</td>
<td>0.085</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>102</td>
<td>0.273</td>
<td>0.756</td>
<td>0.739</td>
<td>0.930</td>
<td>0.663</td>
<td>0.629</td>
<td>0.057</td>
<td>0.534</td>
<td>0.980</td>
<td>0.723</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2998</td>
<td>0.758</td>
<td>0.740</td>
<td>0.928</td>
<td>0.572</td>
<td>0.543</td>
<td>0.826</td>
<td>0.073</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>142</td>
<td>0.380</td>
<td>0.951</td>
<td>0.972</td>
<td>0.746</td>
<td>0.706</td>
<td>0.743</td>
<td>0.704</td>
<td>0.993</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>120</td>
<td>0.823</td>
<td>0.236</td>
<td>0.005</td>
<td>0.974</td>
<td>0.798</td>
<td>0.302</td>
<td>0.811</td>
<td>0.894</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>102</td>
<td>0.341</td>
<td>0.567</td>
<td>0.563</td>
<td>0.387</td>
<td>0.270</td>
<td>0.992</td>
<td>0.671</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>444</td>
<td>0.512</td>
<td>0.500</td>
<td>0.727</td>
<td>0.848</td>
<td>0.805</td>
<td>0.557</td>
<td>0.049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>220</td>
<td>0.588</td>
<td>0.614</td>
<td>0.627</td>
<td>0.482</td>
<td>0.914</td>
<td>0.868</td>
<td>0.455</td>
<td>0.641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>124</td>
<td>0.547</td>
<td>0.560</td>
<td>0.385</td>
<td>0.277</td>
<td>0.225</td>
<td>0.502</td>
<td>0.176</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>86</td>
<td>0.364</td>
<td>0.993</td>
<td>0.966</td>
<td>0.797</td>
<td>0.712</td>
<td>0.715</td>
<td>0.965</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2202</td>
<td>0.969</td>
<td>0.992</td>
<td>0.682</td>
<td>0.420</td>
<td>0.399</td>
<td>0.890</td>
<td>0.099</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>139</td>
<td>0.372</td>
<td>0.971</td>
<td>0.993</td>
<td>0.763</td>
<td>0.692</td>
<td>0.728</td>
<td>0.986</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>131</td>
<td>0.058</td>
<td>0.059</td>
<td>0.044</td>
<td>0.025</td>
<td>0.024</td>
<td>0.033</td>
<td>0.060</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>25</td>
<td>0.067</td>
<td>0.185</td>
<td>0.181</td>
<td>0.041</td>
<td>0.693</td>
<td>0.641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Row: FUCA1 (X-axis), Column: ADH5 (Y-Axis), Length is in nucleotides, E = Exon; I = intron

Table 2. Distribution of block lengths (2 to >25) showing # of matched pairs identified; # of pairs showing same gene names and # of pairs showing different gene names

<table>
<thead>
<tr>
<th>Lengths of Blocks</th>
<th>Matches with different gene names</th>
<th>Matches with same gene names</th>
<th>Total # of matches</th>
<th>Matches with same gene name (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1737702</td>
<td>334</td>
<td>1738036</td>
<td>9.02e-2</td>
</tr>
<tr>
<td>3</td>
<td>175755</td>
<td>805</td>
<td>16380</td>
<td>4.9</td>
</tr>
<tr>
<td>4</td>
<td>358</td>
<td>279</td>
<td>637</td>
<td>44.0</td>
</tr>
<tr>
<td>5</td>
<td>845</td>
<td>490</td>
<td>1335</td>
<td>36.7</td>
</tr>
<tr>
<td>6</td>
<td>142</td>
<td>265</td>
<td>407</td>
<td>65.1</td>
</tr>
<tr>
<td>7</td>
<td>159</td>
<td>508</td>
<td>667</td>
<td>76.2</td>
</tr>
<tr>
<td>8</td>
<td>51</td>
<td>169</td>
<td>230</td>
<td>76.8</td>
</tr>
<tr>
<td>9</td>
<td>95</td>
<td>307</td>
<td>402</td>
<td>76.4</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>120</td>
<td>150</td>
<td>86.7</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>401</td>
<td>451</td>
<td>88.9</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>127</td>
<td>148</td>
<td>85.8</td>
</tr>
<tr>
<td>13</td>
<td>33</td>
<td>232</td>
<td>289</td>
<td>87.2</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>124</td>
<td>134</td>
<td>94.1</td>
</tr>
<tr>
<td>15</td>
<td>30</td>
<td>256</td>
<td>286</td>
<td>89.5</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>130</td>
<td>136</td>
<td>95.6</td>
</tr>
<tr>
<td>17</td>
<td>21</td>
<td>236</td>
<td>257</td>
<td>91.8</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>121</td>
<td>129</td>
<td>93.8</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>172</td>
<td>189</td>
<td>91.0</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>115</td>
<td>130</td>
<td>97.4</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>81</td>
<td>84</td>
<td>96.4</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>153</td>
<td>160</td>
<td>95.6</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>70</td>
<td>71</td>
<td>98.6</td>
</tr>
<tr>
<td>&gt;=25</td>
<td>36</td>
<td>1464</td>
<td>1500</td>
<td>97.6</td>
</tr>
</tbody>
</table>

Note: Block lengths >25 are clustered together.

Table 3. Distribution of block lengths in the sense and anti-sense strand

<table>
<thead>
<tr>
<th>Lengths of blocks</th>
<th>Blocks starting with exon</th>
<th>Blocks starting with intron</th>
<th>Total</th>
<th>% of matches on same Chr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2161729</td>
<td>219236</td>
<td>434200</td>
<td>869668</td>
</tr>
<tr>
<td>3</td>
<td>3442</td>
<td>3682</td>
<td>5855</td>
<td>12979</td>
</tr>
<tr>
<td>4</td>
<td>108</td>
<td>112</td>
<td>299</td>
<td>128</td>
</tr>
<tr>
<td>5</td>
<td>327</td>
<td>304</td>
<td>156</td>
<td>1287</td>
</tr>
<tr>
<td>6</td>
<td>106</td>
<td>78</td>
<td>212</td>
<td>69</td>
</tr>
<tr>
<td>7</td>
<td>225</td>
<td>361</td>
<td>50</td>
<td>636</td>
</tr>
<tr>
<td>8</td>
<td>54</td>
<td>40</td>
<td>8</td>
<td>102</td>
</tr>
<tr>
<td>9</td>
<td>197</td>
<td>152</td>
<td>38</td>
<td>387</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>41</td>
<td>9</td>
<td>79</td>
</tr>
<tr>
<td>11</td>
<td>202</td>
<td>219</td>
<td>18</td>
<td>459</td>
</tr>
<tr>
<td>12</td>
<td>33</td>
<td>26</td>
<td>76</td>
<td>32</td>
</tr>
<tr>
<td>13</td>
<td>138</td>
<td>129</td>
<td>11</td>
<td>278</td>
</tr>
<tr>
<td>14</td>
<td>44</td>
<td>38</td>
<td>2</td>
<td>84</td>
</tr>
<tr>
<td>15</td>
<td>141</td>
<td>120</td>
<td>16</td>
<td>277</td>
</tr>
<tr>
<td>16</td>
<td>28</td>
<td>29</td>
<td>4</td>
<td>61</td>
</tr>
<tr>
<td>17</td>
<td>161</td>
<td>83</td>
<td>11</td>
<td>255</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>31</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>19</td>
<td>39</td>
<td>87</td>
<td>9</td>
<td>183</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
<td>16</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>21</td>
<td>69</td>
<td>53</td>
<td>6</td>
<td>128</td>
</tr>
<tr>
<td>22</td>
<td>18</td>
<td>25</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>23</td>
<td>84</td>
<td>70</td>
<td>1</td>
<td>155</td>
</tr>
<tr>
<td>24</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>&gt;=25</td>
<td>578</td>
<td>663</td>
<td>1249</td>
<td>1349</td>
</tr>
</tbody>
</table>

Note: Block lengths >25 are clustered together. Matched pairs on the same chromosome are also shown.
Patterns in human genome

<table>
<thead>
<tr>
<th>Lengths of blocks</th>
<th># of unique length patterns</th>
<th>Total # of matched pairs</th>
<th># of verified pairs</th>
<th># of verified pairs/total # of matched pairs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>286756</td>
<td>17358636</td>
<td>339574</td>
<td>19.5</td>
</tr>
<tr>
<td>3</td>
<td>10417</td>
<td>16380</td>
<td>2958</td>
<td>18.1</td>
</tr>
<tr>
<td>4</td>
<td>335</td>
<td>637</td>
<td>34</td>
<td>5.3</td>
</tr>
<tr>
<td>5</td>
<td>548</td>
<td>1335</td>
<td>57</td>
<td>4.3</td>
</tr>
<tr>
<td>6</td>
<td>232</td>
<td>497</td>
<td>14</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>361</td>
<td>667</td>
<td>1</td>
<td>7.6</td>
</tr>
<tr>
<td>8</td>
<td>155</td>
<td>220</td>
<td>18</td>
<td>8.2</td>
</tr>
<tr>
<td>9</td>
<td>255</td>
<td>402</td>
<td>18</td>
<td>9.5</td>
</tr>
<tr>
<td>10</td>
<td>109</td>
<td>150</td>
<td>8</td>
<td>5.3</td>
</tr>
<tr>
<td>11</td>
<td>213</td>
<td>451</td>
<td>92</td>
<td>20.4</td>
</tr>
<tr>
<td>12</td>
<td>98</td>
<td>148</td>
<td>10</td>
<td>6.8</td>
</tr>
<tr>
<td>13</td>
<td>169</td>
<td>289</td>
<td>23</td>
<td>8.0</td>
</tr>
<tr>
<td>14</td>
<td>91</td>
<td>135</td>
<td>11</td>
<td>8.1</td>
</tr>
<tr>
<td>15</td>
<td>156</td>
<td>286</td>
<td>17</td>
<td>5.9</td>
</tr>
<tr>
<td>16</td>
<td>87</td>
<td>156</td>
<td>13</td>
<td>9.6</td>
</tr>
<tr>
<td>17</td>
<td>131</td>
<td>257</td>
<td>25</td>
<td>9.7</td>
</tr>
<tr>
<td>18</td>
<td>81</td>
<td>129</td>
<td>12</td>
<td>9.3</td>
</tr>
<tr>
<td>19</td>
<td>127</td>
<td>189</td>
<td>45</td>
<td>23.8</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>77</td>
<td>15</td>
<td>19.5</td>
</tr>
<tr>
<td>21</td>
<td>98</td>
<td>132</td>
<td>29</td>
<td>22.0</td>
</tr>
<tr>
<td>22</td>
<td>50</td>
<td>84</td>
<td>13</td>
<td>15.5</td>
</tr>
<tr>
<td>23</td>
<td>78</td>
<td>160</td>
<td>41</td>
<td>25.5</td>
</tr>
<tr>
<td>24</td>
<td>40</td>
<td>71</td>
<td>11</td>
<td>15.5</td>
</tr>
<tr>
<td>&gt;25</td>
<td>682</td>
<td>1500</td>
<td>354</td>
<td>23.6</td>
</tr>
</tbody>
</table>

Percentage of blocks verified by Aceview; Note: Block lengths >25 are clustered together.

Table 5. Distribution of block lengths, number of block lengths covered in blocks of length 2 and number of new blocks of length 2 identified in blocks of length 3 to >25

<table>
<thead>
<tr>
<th>Lengths of blocks</th>
<th># of unique patterns with Blocks of length 2 (A)</th>
<th># of patterns that are covered in blocks of length 2 (B)</th>
<th># of patterns that are not covered in blocks of length 2 (C)</th>
<th>B/A (%)</th>
<th>C/B (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16373</td>
<td>16089</td>
<td>284</td>
<td>98.3</td>
<td>98.3</td>
</tr>
<tr>
<td>4</td>
<td>851</td>
<td>790</td>
<td>65</td>
<td>92.5</td>
<td>92.5</td>
</tr>
<tr>
<td>5</td>
<td>1849</td>
<td>1619</td>
<td>230</td>
<td>87.6</td>
<td>87.6</td>
</tr>
<tr>
<td>6</td>
<td>998</td>
<td>902</td>
<td>96</td>
<td>90.4</td>
<td>90.4</td>
</tr>
<tr>
<td>7</td>
<td>1977</td>
<td>1761</td>
<td>216</td>
<td>89.1</td>
<td>89.1</td>
</tr>
<tr>
<td>8</td>
<td>998</td>
<td>912</td>
<td>76</td>
<td>92.1</td>
<td>92.1</td>
</tr>
<tr>
<td>9</td>
<td>1928</td>
<td>1733</td>
<td>195</td>
<td>89.9</td>
<td>89.9</td>
</tr>
<tr>
<td>10</td>
<td>937</td>
<td>851</td>
<td>86</td>
<td>90.8</td>
<td>90.8</td>
</tr>
<tr>
<td>11</td>
<td>1989</td>
<td>1814</td>
<td>177</td>
<td>91.2</td>
<td>91.2</td>
</tr>
<tr>
<td>12</td>
<td>975</td>
<td>879</td>
<td>92</td>
<td>90.5</td>
<td>90.5</td>
</tr>
<tr>
<td>13</td>
<td>1924</td>
<td>1761</td>
<td>163</td>
<td>91.5</td>
<td>91.5</td>
</tr>
<tr>
<td>14</td>
<td>1144</td>
<td>1071</td>
<td>73</td>
<td>93.6</td>
<td>93.6</td>
</tr>
<tr>
<td>15</td>
<td>2103</td>
<td>1903</td>
<td>200</td>
<td>90.5</td>
<td>90.5</td>
</tr>
<tr>
<td>16</td>
<td>1262</td>
<td>1165</td>
<td>97</td>
<td>92.3</td>
<td>92.3</td>
</tr>
<tr>
<td>17</td>
<td>2002</td>
<td>1800</td>
<td>202</td>
<td>89.9</td>
<td>89.9</td>
</tr>
<tr>
<td>18</td>
<td>1312</td>
<td>1212</td>
<td>100</td>
<td>92.4</td>
<td>92.4</td>
</tr>
<tr>
<td>19</td>
<td>2105</td>
<td>1927</td>
<td>178</td>
<td>91.3</td>
<td>91.3</td>
</tr>
<tr>
<td>20</td>
<td>886</td>
<td>824</td>
<td>62</td>
<td>93.0</td>
<td>93.0</td>
</tr>
<tr>
<td>21</td>
<td>1785</td>
<td>1640</td>
<td>145</td>
<td>91.9</td>
<td>91.9</td>
</tr>
<tr>
<td>22</td>
<td>1032</td>
<td>969</td>
<td>63</td>
<td>93.9</td>
<td>93.9</td>
</tr>
<tr>
<td>23</td>
<td>1636</td>
<td>1506</td>
<td>130</td>
<td>92.1</td>
<td>92.1</td>
</tr>
<tr>
<td>24</td>
<td>918</td>
<td>871</td>
<td>47</td>
<td>94.9</td>
<td>94.9</td>
</tr>
<tr>
<td>25</td>
<td>19720</td>
<td>18285</td>
<td>1455</td>
<td>92.7</td>
<td>92.7</td>
</tr>
<tr>
<td>Total</td>
<td>66090</td>
<td>62284</td>
<td>4406</td>
<td>93.4</td>
<td>93.4</td>
</tr>
</tbody>
</table>

Percentage of blocks verified by Aceview are also shown; Note: Block lengths >25 are clustered together.

The view is further validated by the observations on function of introns (8). The data implies on the role of segmental gene duplications in genome evolution (9) and also confirms that gene duplication on the same chromosome is prevalent than inter-chromosomal gene duplication (10).

Because of several confounding aspects such as domain shuffling, the existence of isoforms derived from alternative splicing, and annotation errors in the databases detecting duplicate genes in a genome is not a simple task. To test if length patterns of blocks could help identify duplicate genes we explored the minimum number of blocks required for two genes to have the same gene name. It is interesting to see that genes that share blocks of length 10 or more have same gene name >85% of the time (Table 2; Figure 3). These results suggest that exon lengths and intron lengths play a significant role in gene assembly and function. A distribution profile on the number of matched pairs (total # of blocks identified) on each chromosome identified chromosome 16 and chromosome 19 with maximum number of hits (Figure 4). Both of these chromosomes are reported as duplication rich with highest number of segmental duplications (11-12). Chromosome 19 is also notable for the prevalence of duplication structures such as tandemly clustered gene structures. These findings are thus complementary to previous reports and present a new method for identification and detection of duplicate genes. On average a human gene usually has 10 exons (13). The presence of exon–intron pairs (blocks of length 2) of almost identical length but of dissimilar base sequence in numerous locations throughout the human genome supports the idea that introns are the result of propagation by replicative transposons or retrotransposons as suggested by Roger et al. (14). If this is the case they should be expected
Patterns in human genome

Table 6. Distribution of block lengths, number of block lengths covered in blocks of length 3 and number of new blocks of length 3 identified in blocks of length 4 to >25

<table>
<thead>
<tr>
<th>Lengths of blocks</th>
<th># of unique patterns with lengths of block 3 (A)</th>
<th># of patterns that are not covered in blocks of length 3 (B)</th>
<th># of patterns that are not covered in blocks of length 3</th>
<th>B/A (%)</th>
<th># of verified patterns (C)</th>
<th>C/B (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>578</td>
<td>105</td>
<td>473</td>
<td>18.2</td>
<td>33</td>
<td>51.4</td>
</tr>
<tr>
<td>5</td>
<td>1414</td>
<td>192</td>
<td>1222</td>
<td>13.6</td>
<td>62</td>
<td>32.3</td>
</tr>
<tr>
<td>6</td>
<td>808</td>
<td>81</td>
<td>727</td>
<td>10.0</td>
<td>24</td>
<td>29.6</td>
</tr>
<tr>
<td>7</td>
<td>1665</td>
<td>176</td>
<td>1489</td>
<td>10.6</td>
<td>72</td>
<td>48.9</td>
</tr>
<tr>
<td>8</td>
<td>854</td>
<td>60</td>
<td>794</td>
<td>9.2</td>
<td>20</td>
<td>33.3</td>
</tr>
<tr>
<td>9</td>
<td>1702</td>
<td>144</td>
<td>1558</td>
<td>8.5</td>
<td>61</td>
<td>42.4</td>
</tr>
<tr>
<td>10</td>
<td>835</td>
<td>47</td>
<td>788</td>
<td>5.6</td>
<td>21</td>
<td>44.7</td>
</tr>
<tr>
<td>11</td>
<td>1800</td>
<td>130</td>
<td>1670</td>
<td>7.2</td>
<td>69</td>
<td>53.1</td>
</tr>
<tr>
<td>12</td>
<td>890</td>
<td>66</td>
<td>824</td>
<td>7.4</td>
<td>24</td>
<td>36.4</td>
</tr>
<tr>
<td>13</td>
<td>1774</td>
<td>123</td>
<td>1651</td>
<td>6.9</td>
<td>47</td>
<td>38.2</td>
</tr>
<tr>
<td>14</td>
<td>1061</td>
<td>75</td>
<td>986</td>
<td>7.1</td>
<td>35</td>
<td>46.7</td>
</tr>
<tr>
<td>15</td>
<td>1964</td>
<td>123</td>
<td>1941</td>
<td>6.3</td>
<td>46</td>
<td>37.4</td>
</tr>
<tr>
<td>16</td>
<td>1181</td>
<td>68</td>
<td>1113</td>
<td>5.8</td>
<td>36</td>
<td>52.9</td>
</tr>
<tr>
<td>17</td>
<td>1883</td>
<td>105</td>
<td>1778</td>
<td>5.6</td>
<td>55</td>
<td>52.4</td>
</tr>
<tr>
<td>18</td>
<td>1237</td>
<td>77</td>
<td>1160</td>
<td>6.2</td>
<td>43</td>
<td>55.8</td>
</tr>
<tr>
<td>19</td>
<td>1994</td>
<td>133</td>
<td>1861</td>
<td>6.7</td>
<td>57</td>
<td>42.9</td>
</tr>
<tr>
<td>20</td>
<td>846</td>
<td>44</td>
<td>802</td>
<td>5.2</td>
<td>24</td>
<td>54.5</td>
</tr>
<tr>
<td>21</td>
<td>1711</td>
<td>111</td>
<td>1600</td>
<td>6.5</td>
<td>60</td>
<td>54.1</td>
</tr>
<tr>
<td>22</td>
<td>906</td>
<td>51</td>
<td>915</td>
<td>5.2</td>
<td>31</td>
<td>60.8</td>
</tr>
<tr>
<td>23</td>
<td>1572</td>
<td>86</td>
<td>1486</td>
<td>5.5</td>
<td>40</td>
<td>46.5</td>
</tr>
<tr>
<td>24</td>
<td>880</td>
<td>49</td>
<td>831</td>
<td>5.6</td>
<td>34</td>
<td>69.4</td>
</tr>
<tr>
<td>25</td>
<td>19911</td>
<td>1197</td>
<td>18714</td>
<td>6.0</td>
<td>658</td>
<td>55.0</td>
</tr>
<tr>
<td>Total</td>
<td>47346</td>
<td>3253</td>
<td>44303</td>
<td>6.8</td>
<td>1227</td>
<td>49.3</td>
</tr>
</tbody>
</table>

Percentage of blocks verified by Aceview are also shown. Note: Block lengths >25 are clustered together.

4.2. From pieces to patterns and gene assemblage

The modular view of evolution suggests for the assembly of new genes from copies of pieces of various older genes, rapidly building new functions from a novel collection of already reliable parts. By analyzing the structure of genes at the genome level, it should be possible to discern the ancient blocks within. To test this hypothesis we derived a unique set of blocks of length 2 present in different blocks lengths (length 3 to length >25). The number of unique patterns identified for a block of length 2 is 206756 and for a block of length 3 is 10417 (Table 4). This indicates that blocks of length 2 provide maximum number of pieces for gene assemblage and probable functional diversity. We further proceeded with identification of unique length patterns of blocks of length 2 and 3 and among the higher block lengths 3 to >25 and 4 to >25, respectively (Table 5 and Table 6). 16373 blocks of length 2 were identified in blocks of length 3, however, 16089 of them (~98%) are previously covered in blocks of length 2. So the new patterns of blocks of length 2 that re-recruits pieces at its disposal rather than repeatedly starting from scratch.

5. CONCLUSION

The exon-intron length patterns of annotated human genes were examined and they revealed common structural patterns (1762278 conserved arrangements of exon-intron length) within the human genome. Exploration of the blocks of variant length shows interesting data highlighting their occurrence by retroposition and suggesting their possible role in gene assemblage. The results support the process of creating new combinations of exons by recruitment of pieces of exons/introns and creation of present-day gene architecture in the human genome. We conclude that the modular protein evolution by exon-shuffling and retroposition are the probable processes that have contributed significantly to human genome design.

6. CAVEATS

As this analysis is strictly based on CDS feature in genome data, it does not take into account the first exon and is biased towards internal coding exons of the gene. Nonetheless, this analysis hints at the possible role of non-coding DNA in genome architecture and design and provides a platform for understanding the human genome and issues in gene evolution. This stringent criterion of
Patterns in human genome

Figure 1. Distribution of block lengths (exon-intron and intron-exon). It is clear that odd length blocks are more than even length blocks and matched pairs with exons as start are more prevalent.

Figure 2. Percentage of block length matches on same chromosome.
Patterns in human genome

Figure 3. Percentage of block length matches with same gene name.

Figure 4. Percentage of block length matches on each chromosome.

99% length match excluded many interesting patterns but also ensured that the remaining ones were genuine and potentially biologically significant.

7. ACKNOWLEDGEMENTS

This manuscript is based on the project supported by A*STAR - BMRC grant #03/1/22/19/242.

8. REFERENCES

Patterns in human genome


Patterns in human genome


**Key Words**: Exon-Intron Length Patterns, Human Genome, Block Length, Duplication, Genome Design, Gene Architecture

**Send correspondence to**: Meena Kishore Sakharkar Ph.D., Nanyang Technological University, Singapore 639798; Tel: 65 6 790 5836; Fax: 65 6 774 4340; E-mail: mmeena@ntu.edu.sg

http://www.bioscience.org/current/vol10.htm