

As represented by this huge stack of paper, the human genome contains more than 3 billion nucleotides or DNA letters. The first stage of the public human genome project focused on identifying marker sequences or unique tags, shown here in yellow at regular intervals throughout this book of life. Once enough sequences were tagged, various blocks of the genome were allocated to different academic centers for sequencing. To begin the sequencing process, several copies of a section of DNA, represented here as a page of text, are cleaved to produce smaller fragments. Although it looks fairly orderly, this step is small-scale shotgun, which creates numerous random fragments. Each fragment is sequenced, then computer programs align the overlap between fragments to build up an entire page. Marker sequences, shown in yellow, help establish the order of pages in the book of life. This methodical process produced huge amounts of data that have been used to virtually reassemble our genome; however, there are gaps. Repeat sequences are common in the human genome, so repeats from entirely different chromosome regions may be erroneously joined together. It will take many years to detect mismatches caused by the repeat sequences. Some regions, especially near the centromeres may never be fully finished.