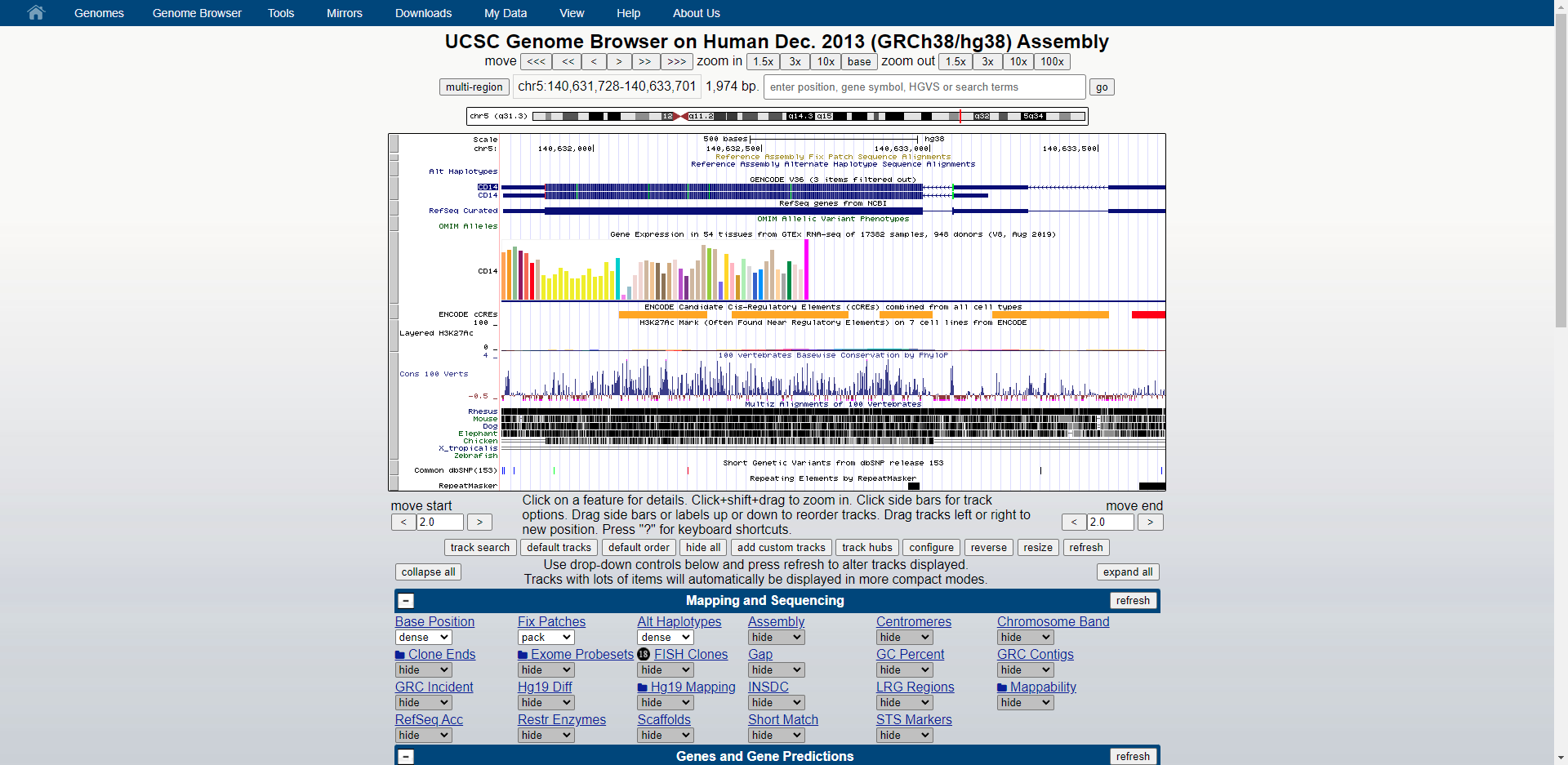
**第1次作业**

自硕21 崔晏菲 2021210976

注：单人完成作业，故无法下载所有的染色体，只挑了有代表性的进行分析

**1. UCSC Genome Browser**



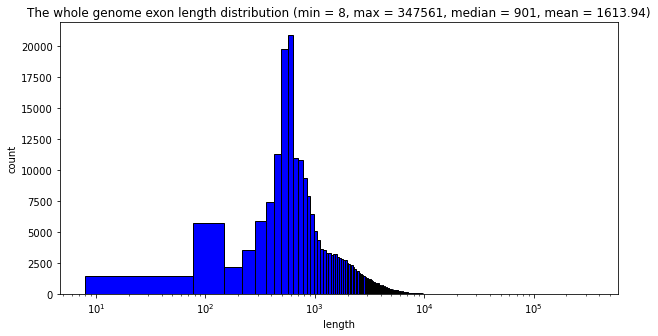
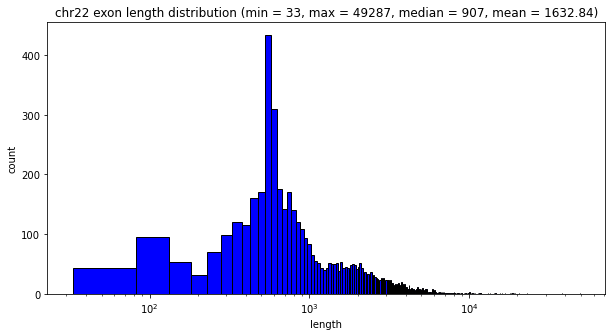
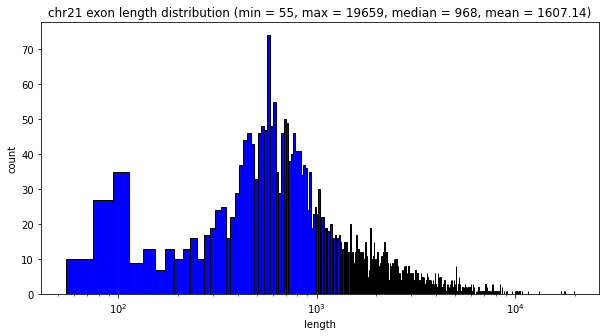
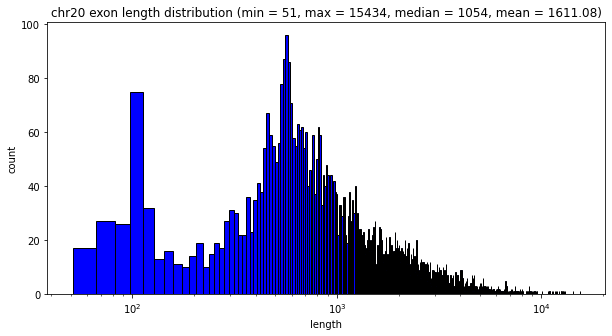
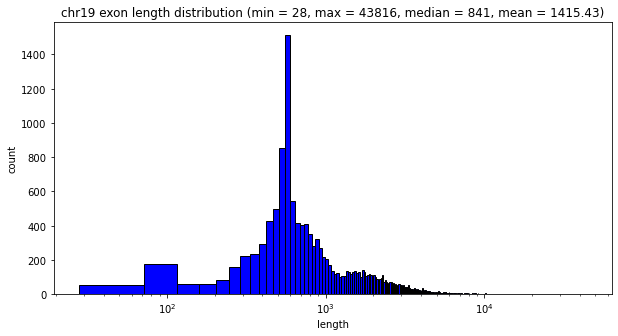
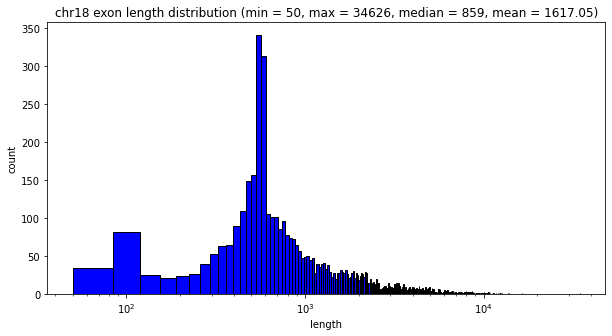
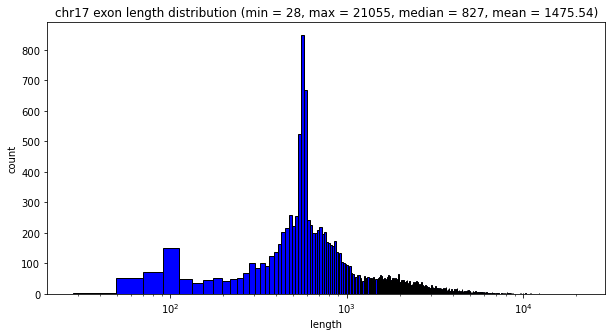
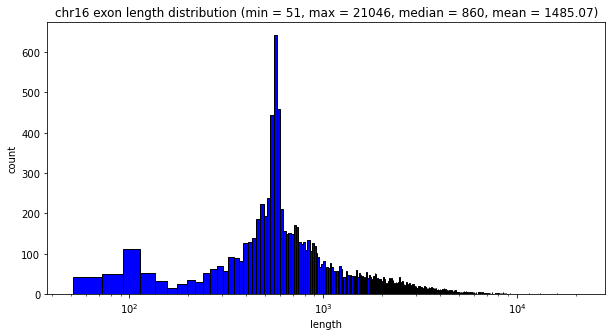
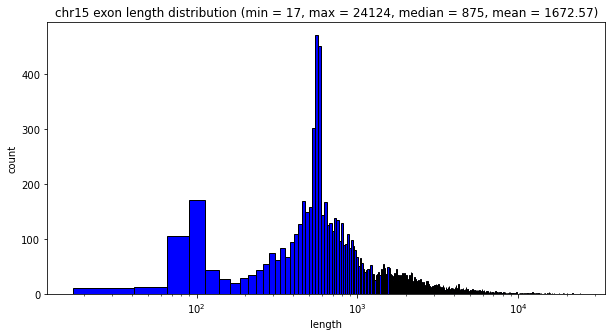
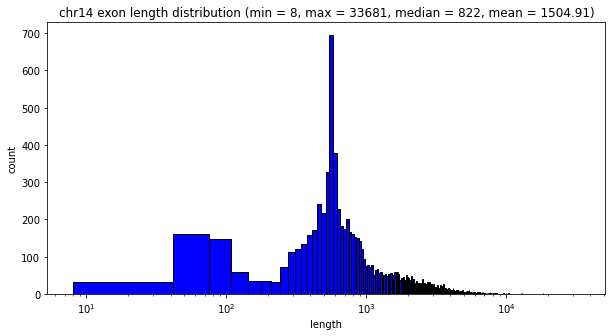
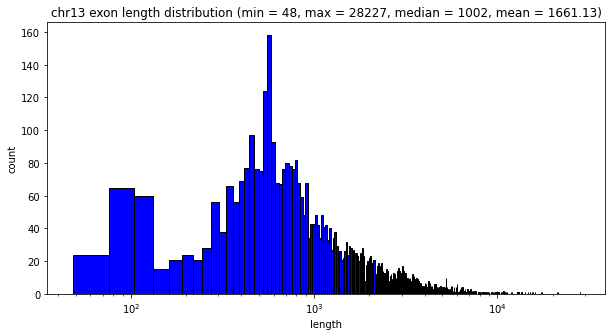
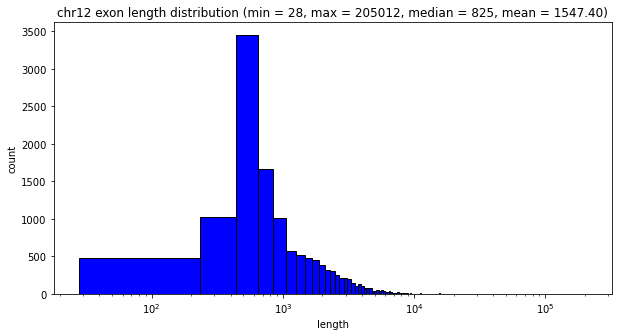
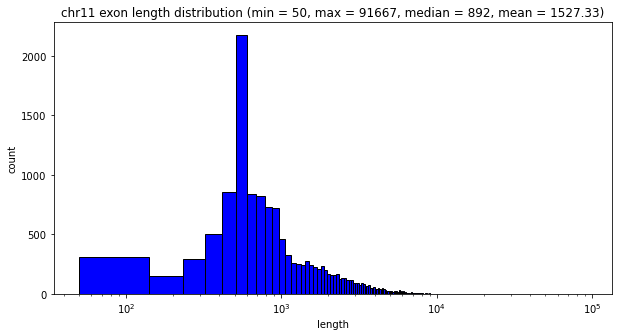
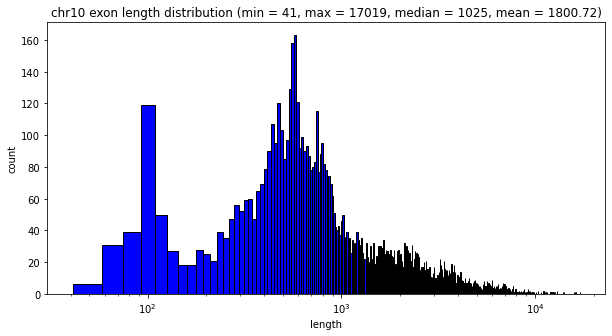
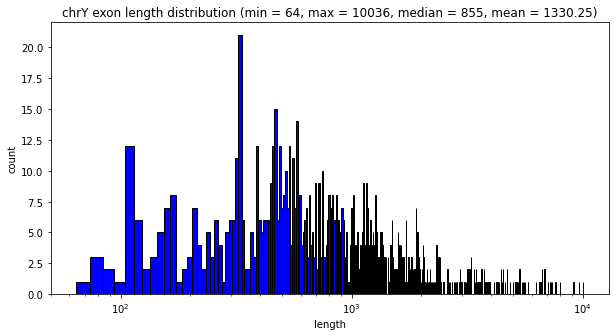
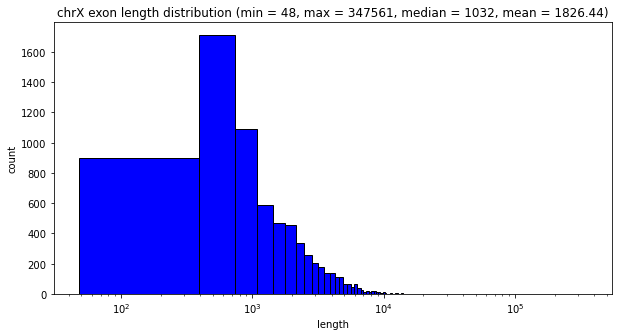
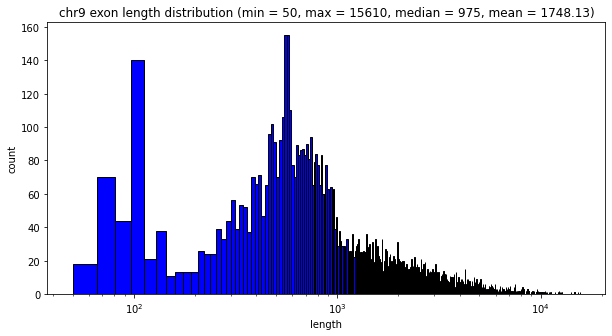
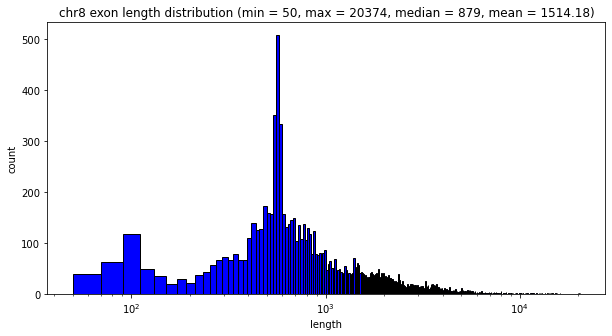
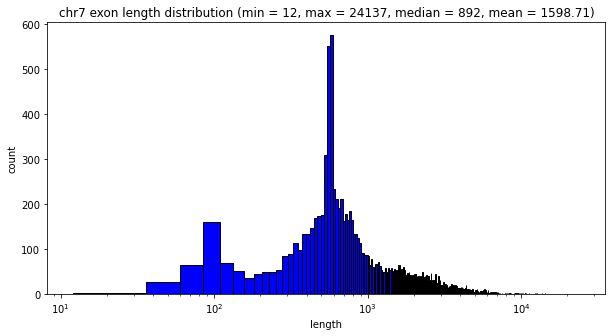
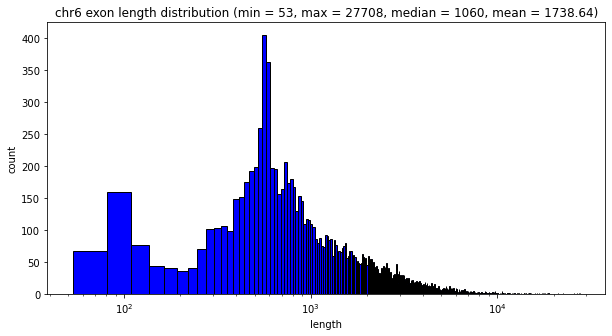
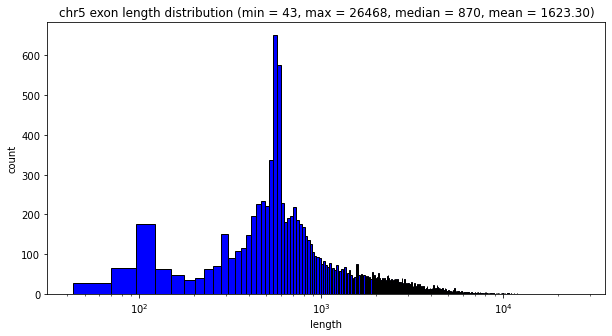
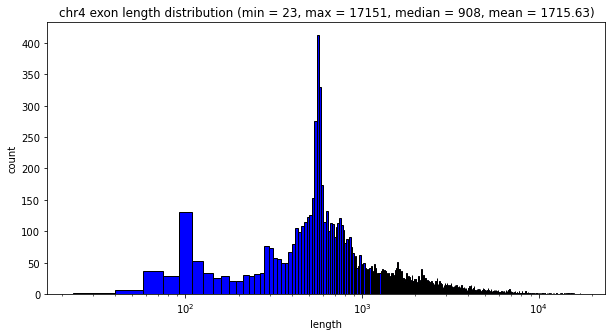
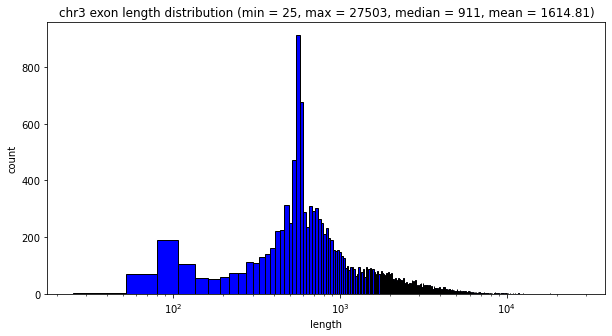
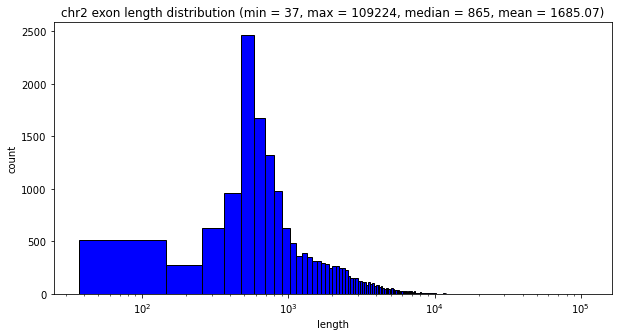
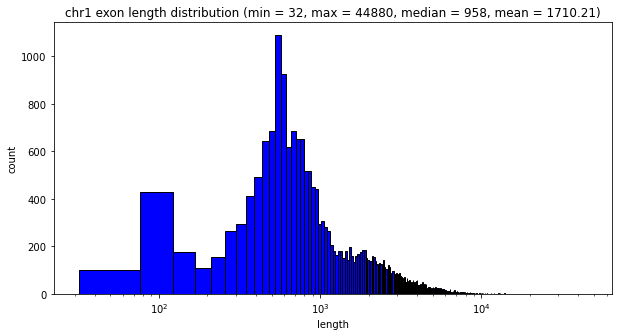
这本质上是一个关系型数据库。

**2. Count the number of protein-coding genes on each chromosome.**

|  |  |
| --- | --- |
| 染色体 | 蛋白质编码基因数 |
| chr1 | 20,376 |
| chr2 | 16,951 |
| chr3 | 14,438 |
| chr4 | 9,533 |
| chr5 | 11,020 |
| chr6 | 10,604 |
| chr7 | 11,016 |
| chr8 | 9,693 |
| chr9 | 7,768 |
| chr10 | 8,124 |
| chr11 | 13,631 |
| chr12 | 12,939 |
| chr13 | 4,244 |
| chr14 | 8,517 |
| chr15 | 8,603 |
| chr16 | 11,048 |
| chr17 | 13,708 |
| chr18 | 4,497 |
| chr19 | 13,790 |
| chr20 | 5,456 |
| chr21 | 2,983 |
| chr22 | 4,996 |
| chrX | 7,169 |
| chrY | 976 |

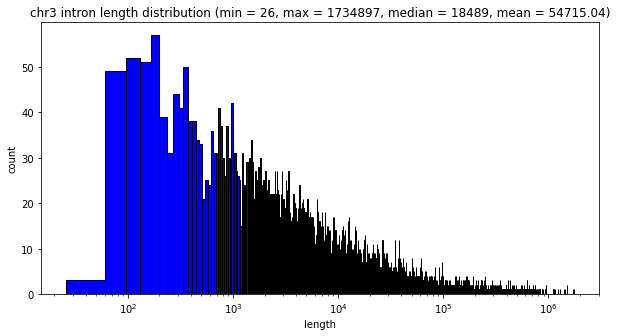
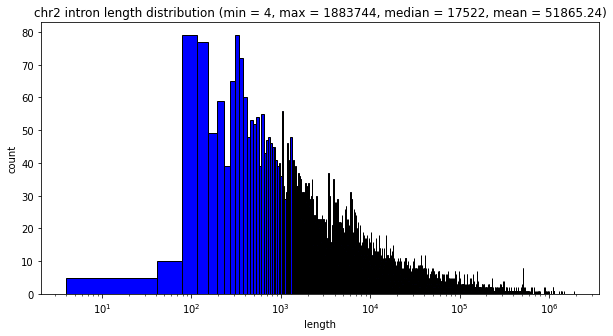
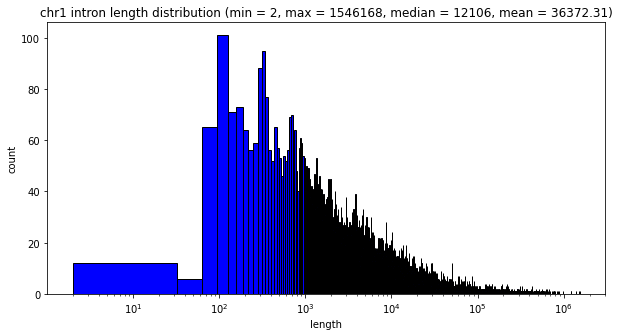
**3. Summarize length distributions of exons, introns and genes on each chromosome and in the whole genome using histograms, and also provide statistics like median, mean, max/min, etc..**

**3.1 外显子长度分布**

****

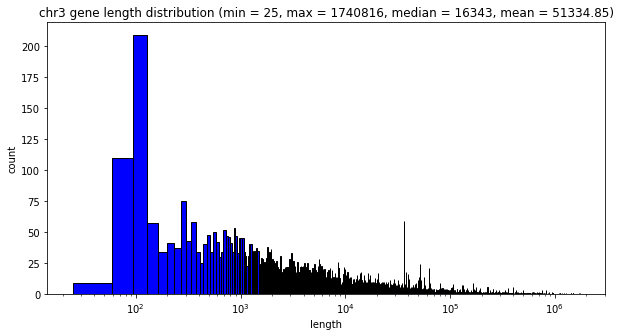
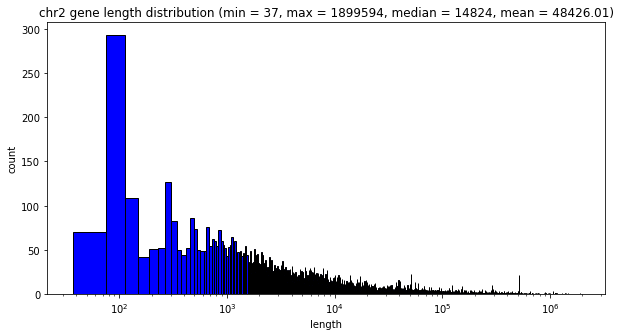
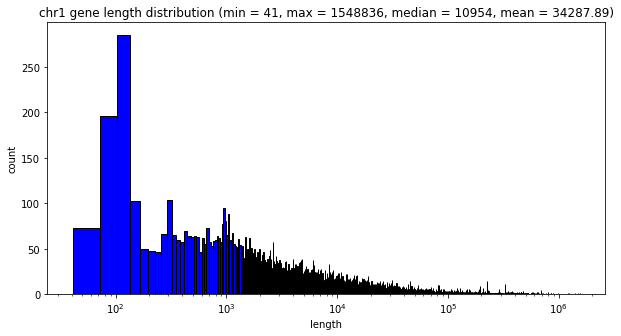
**3.2 内含子长度分布**

由于计算长度的分布需要下载整个内含子序列，而内含子序列文件过大（例如:chr1的内含子序列文件超过600MB！），因此，这里只计算了chr1~chr3的内含子长度分布。



**3.3 基因长度分布**

由于计算长度的分布需要下载整个基因序列，而基因序列文件过大（例如:chr1的基因序列文件超过600MB！），因此，这里只计算了chr1~chr3的内含子长度分布。

****

**4. Reference**

[1] Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. [**Centromere reference models for human chromosomes X and Y satellite arrays**](http://genome.cshlp.org/content/24/4/697.abstract). Genome Res. 2014 Apr;24(4):697-707. Epub 2014 Feb 5.

[2] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep;22(9):1760-74. PMID: 22955987; PMC: PMC3431492

[3] Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006;7 Suppl 1:S4.1-9. PMID: 16925838; PMC: PMC1810553

[4] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, Paul Flicek, GENCODE reference annotation for the human and mouse genomes, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D766–D773, https://doi.org/10.1093/nar/gky955