


Focus	Sequence and structure determination and open to public of reference mouse genome based on long one-molecule sequencing																						
PI	Yoichi Gondo RIKEN																						
Period	FY2015																						
Overview	<p>Just after the completion of human genome sequencing, the mouse genome was determined prior to any other higher organisms. Many standard mouse inbred strains have been sequenced but quite a few gaps still remain as shown in Figure. Human genome reference sequences have already been revised in 2015 in order to fill the gaps as well as to finalize the structural variations like inversions, translocations, duplications and so forth. Accordingly, it becomes necessary to revise the mouse genome reference sequences as one of the most widely used model systems for, e.g., human diseases. Thus, RIKEN conducts the revision of reference mouse genome based on two big data of mouse genome sequences. One is a set of ultra-deep short-reads that was produced by the large-scale analysis of mouse genome and mutations supported by Genome Sciences. The other will be a newly produced set of one-molecule long-reads produced by a latest sequencing technology. Combining together, not only the gaps but also all the structural variations will be revised with the highest precision. RIKEN will make it open to public the assembled as well as original data. Thus, all researchers may further revise and/or add various annotations to the mouse genome reference.</p> <table border="1" data-bbox="454 824 1225 1332"> <thead> <tr> <th>Mouse Genome Reference Data</th> <th><i>Mus musculus</i></th> </tr> </thead> <tbody> <tr> <td>NCBI/Genome Reference Consortium</td> <td>GRCm38.p4</td> </tr> <tr> <td>Total Bases in Assembly</td> <td>2,803,568,840</td> </tr> <tr> <td>Total Non-N Bases in Assembly</td> <td>2,724,212,084</td> </tr> <tr> <td>Total assembly gap length</td> <td>79,356,756</td> </tr> <tr> <td>Gaps between scaffolds</td> <td>191</td> </tr> <tr> <td>Number of scaffolds</td> <td>293</td> </tr> <tr> <td>Scaffold N50</td> <td>52,589,046</td> </tr> <tr> <td>Number of contigs</td> <td>796</td> </tr> <tr> <td>Contig N50</td> <td>32,273,079</td> </tr> <tr> <td>Total number of chromosomes and plasmids</td> <td>22</td> </tr> </tbody> </table> <p>As of Oct. 26, 2015, 191 unspanned gaps and a total of ~80Mb spanned gaps still remain.</p> 	Mouse Genome Reference Data	<i>Mus musculus</i>	NCBI/Genome Reference Consortium	GRCm38.p4	Total Bases in Assembly	2,803,568,840	Total Non-N Bases in Assembly	2,724,212,084	Total assembly gap length	79,356,756	Gaps between scaffolds	191	Number of scaffolds	293	Scaffold N50	52,589,046	Number of contigs	796	Contig N50	32,273,079	Total number of chromosomes and plasmids	22
Mouse Genome Reference Data	<i>Mus musculus</i>																						
NCBI/Genome Reference Consortium	GRCm38.p4																						
Total Bases in Assembly	2,803,568,840																						
Total Non-N Bases in Assembly	2,724,212,084																						
Total assembly gap length	79,356,756																						
Gaps between scaffolds	191																						
Number of scaffolds	293																						
Scaffold N50	52,589,046																						
Number of contigs	796																						
Contig N50	32,273,079																						
Total number of chromosomes and plasmids	22																						
Progress	<p>(written in Japanese)</p> <p>(1) ショートリードビッグデータに基づく高精度塩基配列決定公開されている C57BL/6J マウスゲノム由来の標準参照配列 GRCm38p4 (mm10) と比較検討した結果、本邦で広く利用されている C57BL/6Jcl 標準近交系マウスのゲノム配列では、少なくとも 4601 箇所の塩基配列が異なっていることを確認した。また、この解析の過程で、広く遺伝的解析などに用いられる公開マウス dbSNP 情報のうち少なくとも 1 万個の SNP 情報は間違っていることを強く示す解析結果を得た。</p> <p>(2) 1 分子長鎖リードのための諸条件の検討さまざまな臓器から抽出したゲノム DNA をパルスフィールド電気泳動法を用いて比較検討し、超高分子ゲノム DNA の調製には、肺ゲノムが最も適することを確認した。また、平均 10kb 長で 10×カバレッジの解読するためには、PacBioRSII の 6 時間モードが最も適していると総合的に判断し、そのためのライブラリー作製およびシーケンシングを行った。</p> <p>(3) マウスゲノムの 1 分子長鎖解読 最終的に 7,519,270 の総リード数を得た。解読塩基数はのべ 96,522,590,619bp と約 30×カバレッジのビッグデータを得ることができ当初目的を大きく越えた。また平均リード長も 12,837bp と当初目的の平均 10kb を上回った値が得られた。</p> <p>(4) 成果公開のための整備 DDBJ からオリジナルデータを公開するための登録 ID を取得した。また総合的にアセンブリーしたデータを公開するため、理研情報基盤センターが整備している理研メタデータベース上に公開用プラットフォームを構築した。本研究による成果と公開により、疾患モデル開発における責任遺伝子や原因変異同定の精度が飛躍的にあがる。さ</p>																						

らに、ゲノム参照配列が基盤となる一般的な PCR 解析やゲノム編集の精度向上にも大きく寄与し、モデル動物としてのマウスリソースの価値が高まる。

【成果の発信】

1. "Detection of spontaneous mutations in mammalian genomes." Y. Gondo in "Biological & Medical Science based on Physics: Radiation and Physics, Physics on Medical Science, Modeling for Biological System" Kyoto, 2015/11/5, 国際会議
2. "Needs of fundamental revision of mouse genome reference sequences" Y. Gondo et al. in "29th International Mammalian Genome Conference (IMGC2015)" Yokohama, 2015/11/10, 国際会議
3. 「標準マウス系統 C57BL/6J における自然発生突然変異の解析法の確立」権藤洋一ほか、第 38 回日本分子生物学会/第 88 回日本生化学会合同大会、神戸市、2015/12/1, 国内会議
4. "Genetics, mutagenesis and genomics with standardized mouse inbred strains" Y. Gondo at School of Pharmacy and School of Life Sciences, Fudan University, Shanghai, China, 2015/12/17, 海外招待講演
5. "Accumulation and detection of germline spontaneous mutations in C57BL/6Jcl inbred mouse strain" Y. Gondo et al. in "30th International Mammalian Genome Conference (IMGC2015)" Orland, USA, 2016/7/14, 国際会議