

[\[TOC\]](#)

数据分析

1. 野生稻

1.1 类型

正如图1所示，野生稻中共有2748个SV，其中DEL的数量最多（1428），INS的数量次之（1206），INV的数量是最少的，仅有6个。

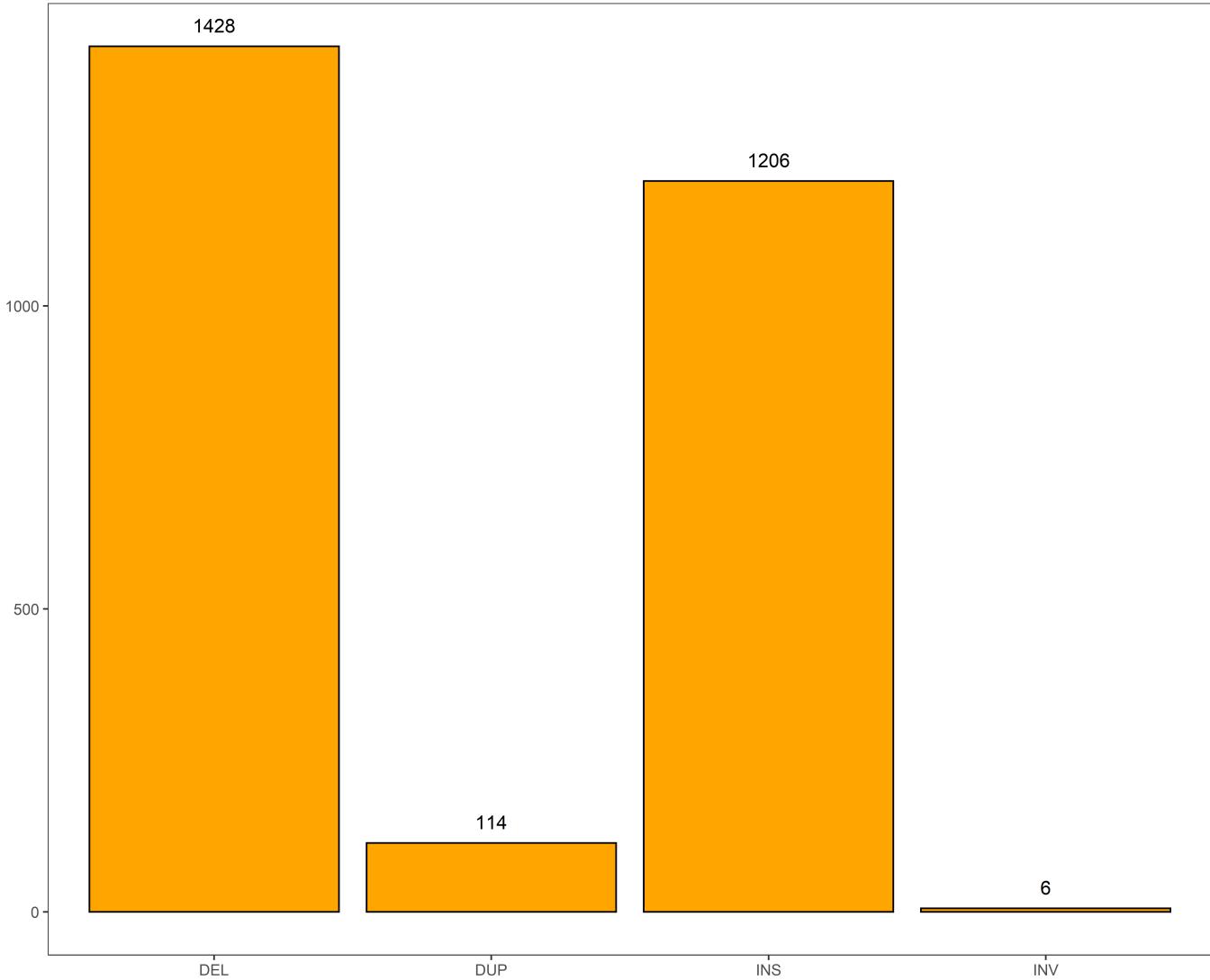


图1

1.2 长度

图2是野生稻中SV长度分布的频率统计图，总体来看，SV的长度分布于31bp到60000bp之间。此外SV的长度主要分布在100bp~1000bp之间，一共有1338个占总体变异的48.69%。而在这其中近乎一半片段都聚集在200bp到500bp之间（50.67%）。同时SV长度在31bp-100bp之间的SV数目也不在少数，共有840份，占总体变异的

30.57%。只有极少部分的SV长度大于5000，仅占总体变异的4.37%。

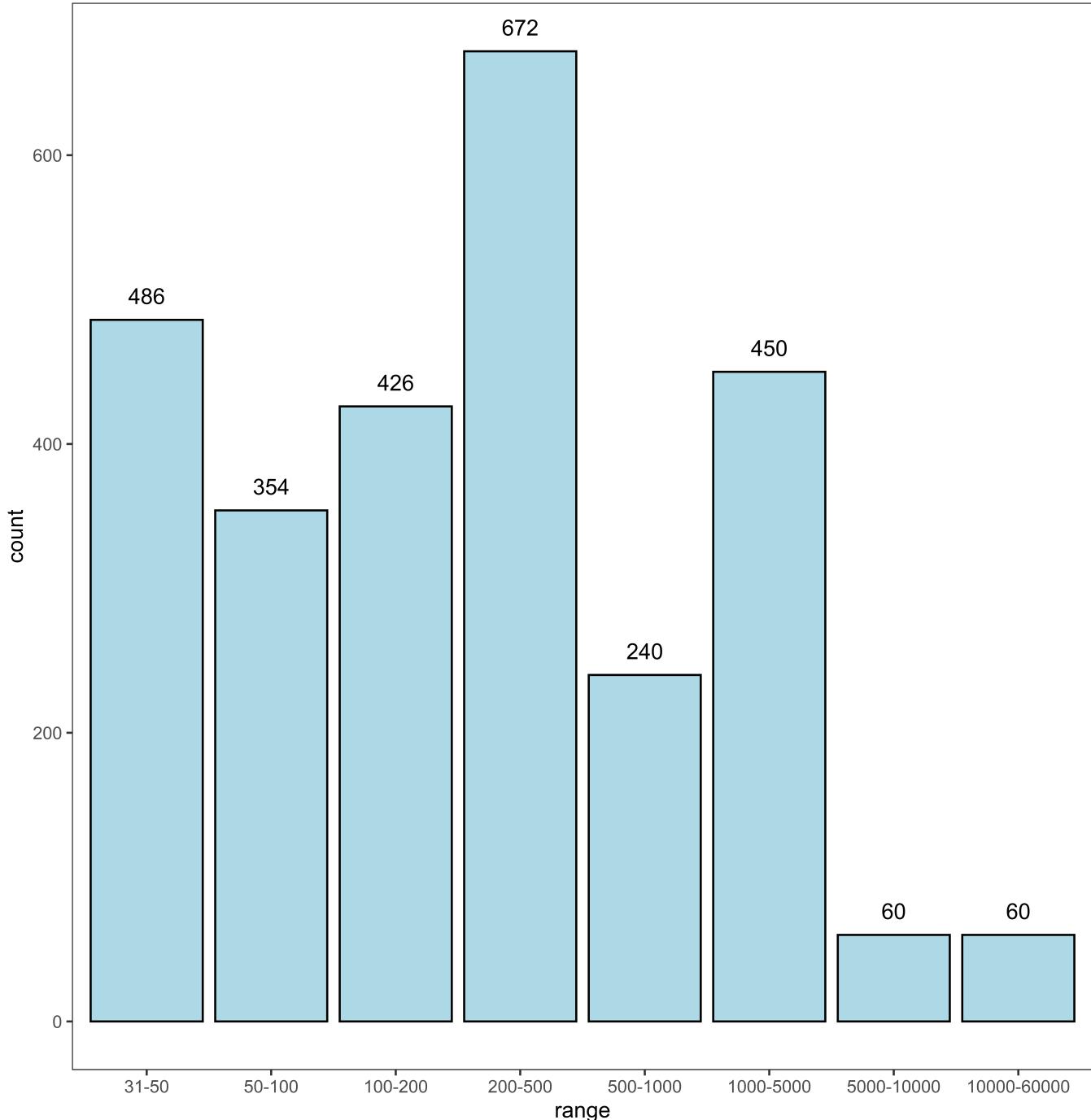


图2

1.3 SV在染色体上的分布情况

图3描述了102份参考基因组中野生稻和栽培稻SV在染色体上的分布，图3-a和图3-b分别是栽培稻和野生稻是以1Mb滑窗大小进行SV数目统计的结果，图3-c和图3-d分别是栽培稻和野生稻是以100kb滑窗大小进行SV数目统计的结果。就栽培稻而言，当滑窗大小为1Mb时结果并不显著，没有较大的峰值出现，但是当滑窗大小调整为100kb后，可以看到，在1号染色体的前端和中部位置出现了峰值，2号染色体、4号染色体、10号染色体、11号染色体和11号染色体都有明显峰值出现，尤其是11号染色体的末端SV的数目最高甚至达到了300。就野生稻而言，当滑窗大小为1Mb时，我们可以看出在1号染色体和10号染色体上出现了一共较大的峰值，且都定位于染色体的前端。除2号、6号和8号染色体是SV的分布较为均匀外，其余染色体是均出现了或大或小的峰值。而当我们将滑窗大小调整为100kb以后，可以看到1号染色体和10号染色体的峰值仍然出现在染色体的前半部分，这一点与图3-b一致是一致的。

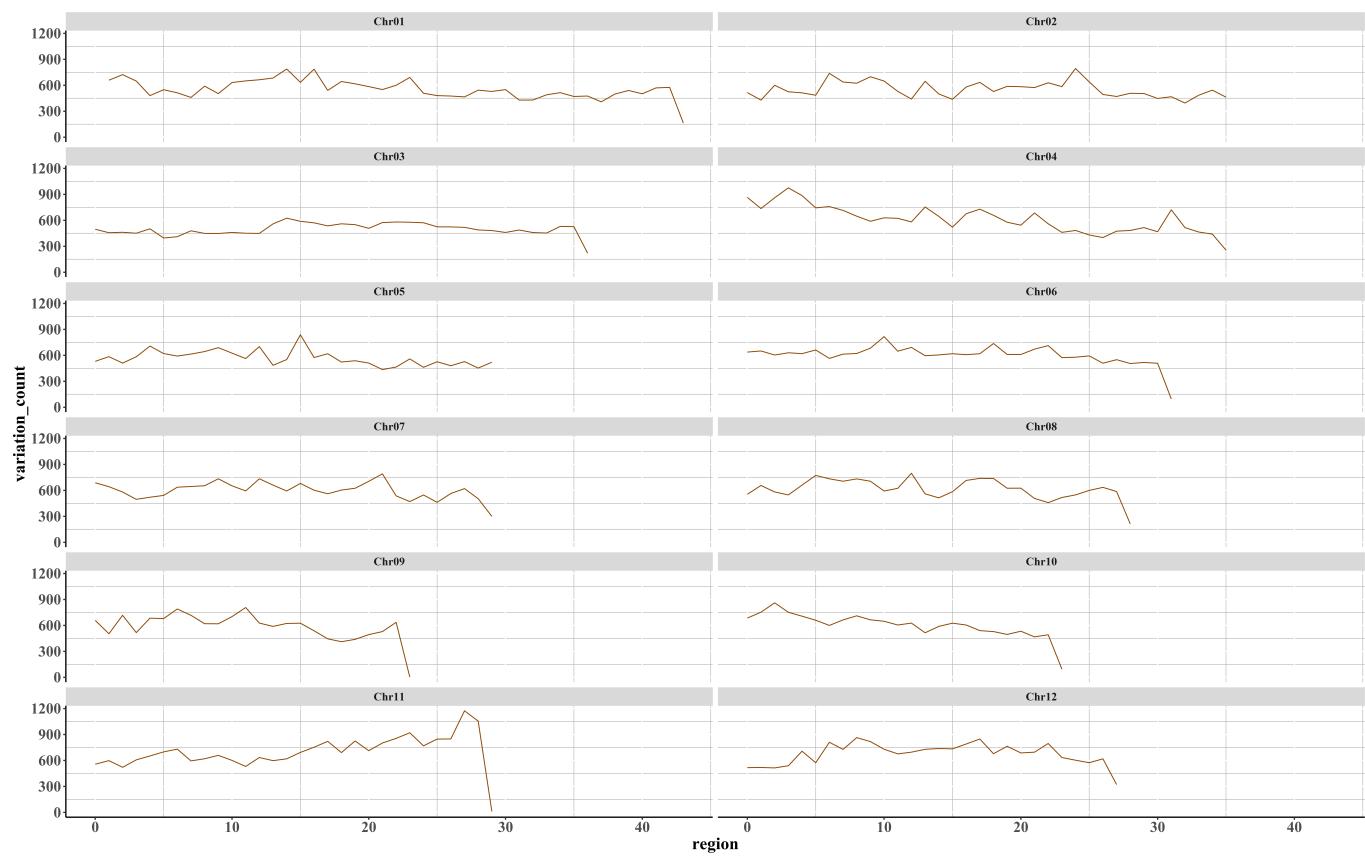


图3-a

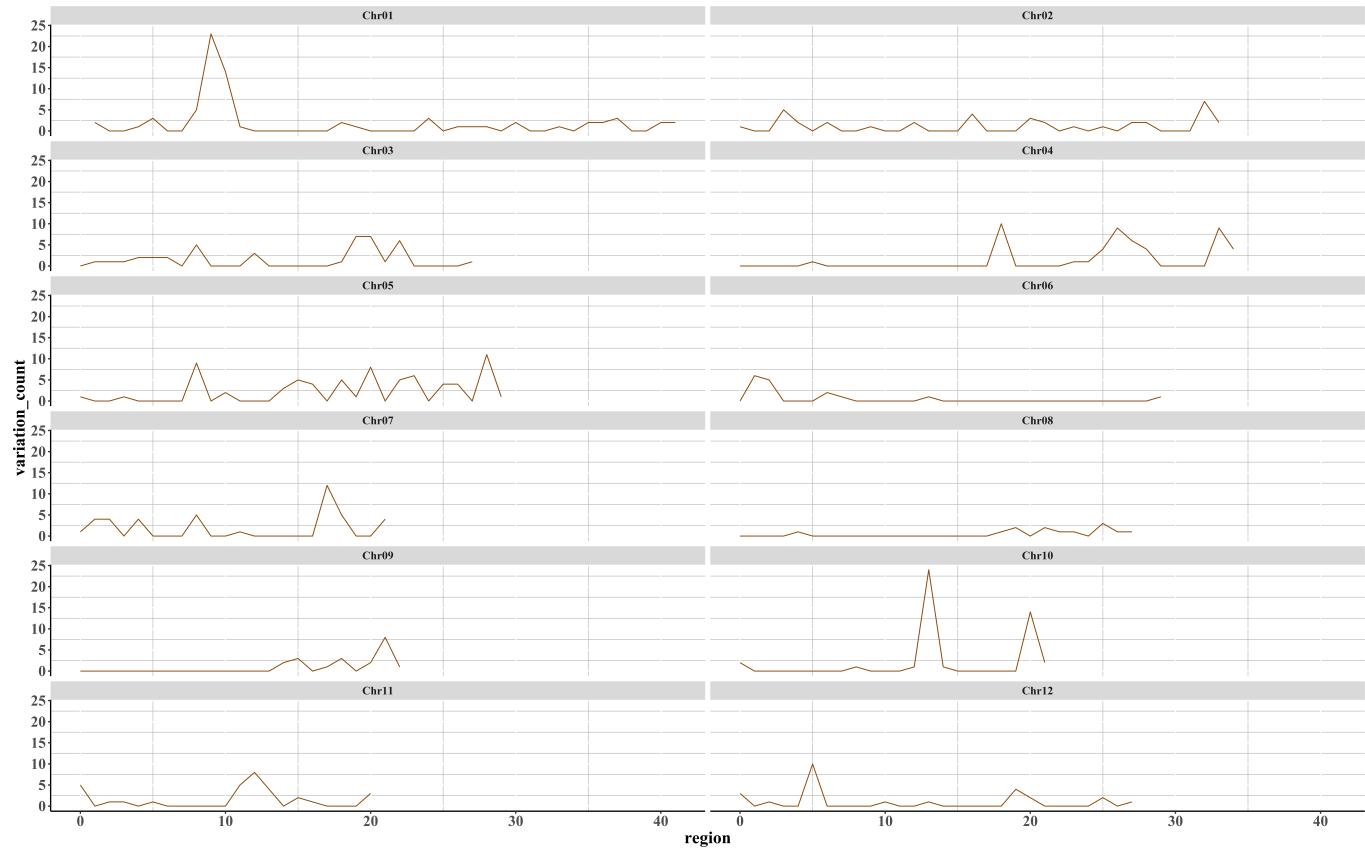


图3-b

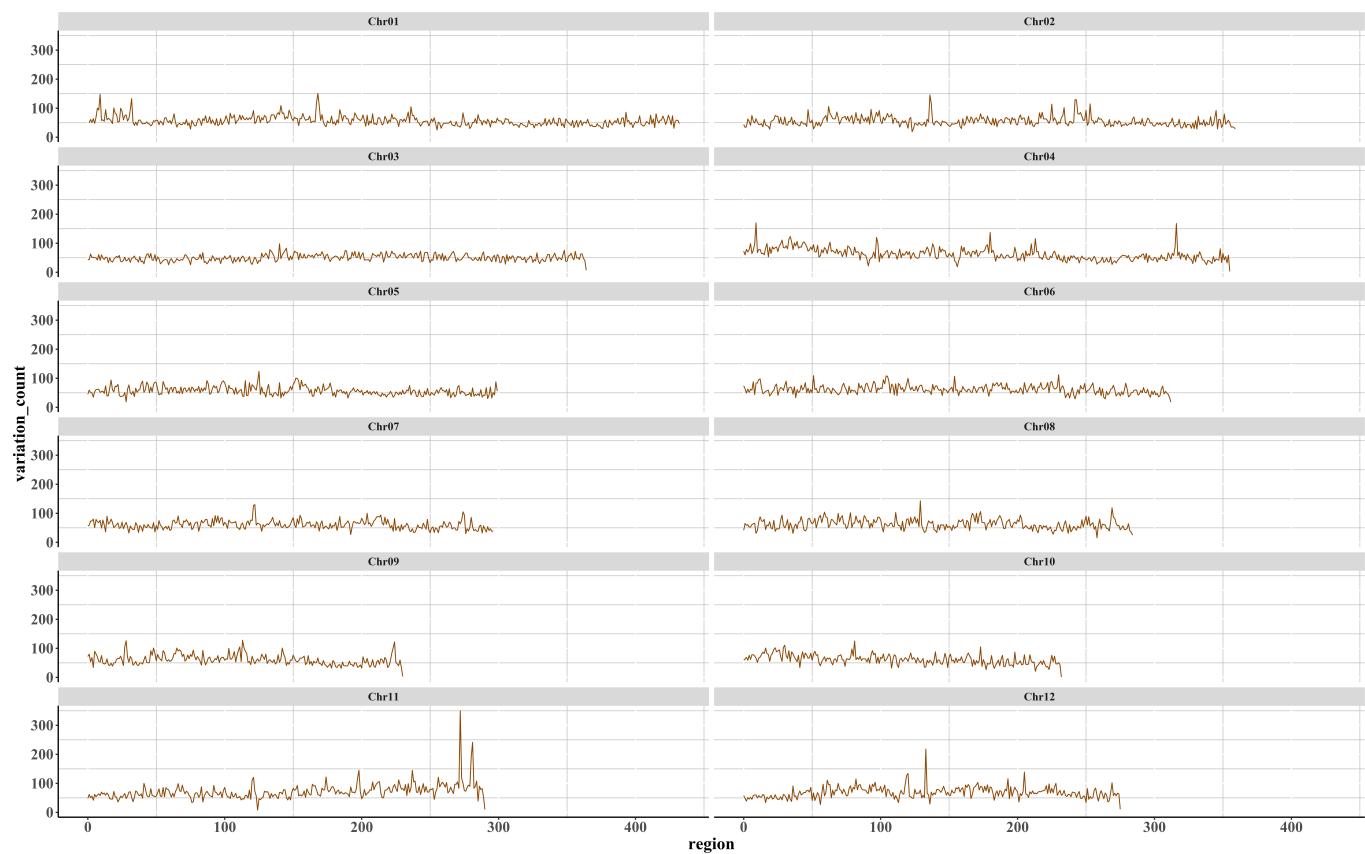


图3-c

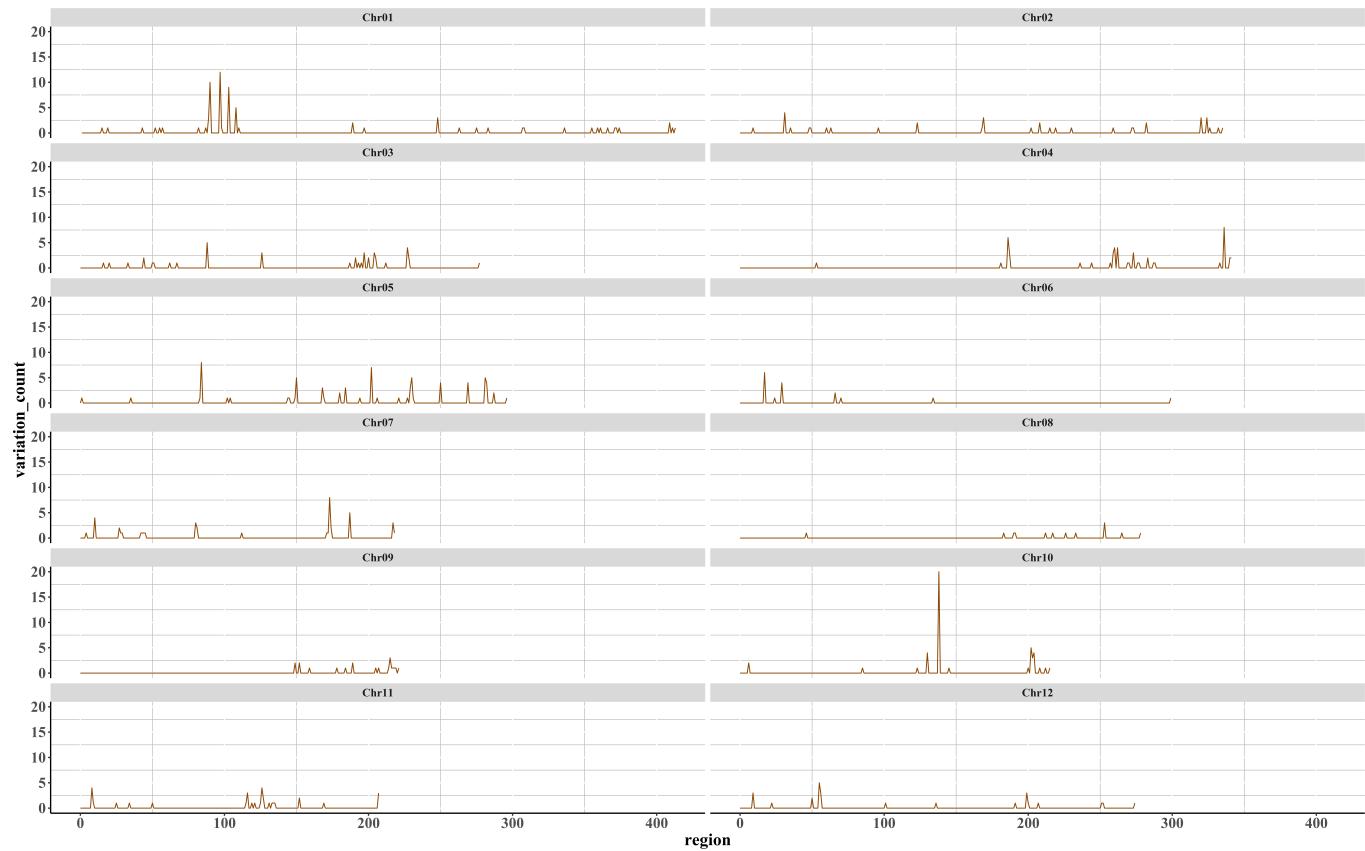


图3-d

102reference

表一的上半部分统计了102份参考基因组在453份3k数据库中高测序深度水稻中SV覆盖情况，而图4则将这一情况进行可视化。102份参考基因组中的SV数目为127096，其中108352份SV在453份材料中至少匹配到了一次，整体覆盖度在85.25%；此外有18744份SV并没有在453份材料中成功匹配，而在这其中大部分SV长度的长度集中在30-50bp，总体占比达32.62%；紧随其后的是长度位于101-1000bp之间的SV，总体占比28.36；随后便是长度位于51-100bp之间SV，共有3994份，占比21.31%；接下来是长度在1000bp-10000bp之间的SV，共有2950份，占比15.74%。此外在108352份SV中还有11209份SV是稀有SV，这些SV均只在一份材料中被成功检测到。稀有SV中，数量最多SV长度区间为100kb-1000kb，一共有4037份，占比36.02%，紧随其后的长度区间为1000kb到10000kb，一共有2706份，占比24.14%。此外就SV类型而言，在未匹配到的18744份SV中，大部分都是DEL，而在成功匹配到的这些SV中，DEL和INS的数量基本相当，但当匹配数量超过400时，INS又稍占上风。

表1

The SVs mapping rate of 453 genomes with 102 reference ↵

	Frequency ↵	DEL ↵	INS ↵
SV distribution ↵ among the 3K ↵ (Number) ↵	0 ↵ 1 ↵ 2-100 ↵ 101~200 ↵ 201~300 ↵ 301~400 ↵ >400 ↵	12494 ↵ 5475 ↵ 31994 ↵ 4862 ↵ 2931 ↵ 1689 ↵ 141 ↵	6250 ↵ 5734 ↵ 44534 ↵ 5350 ↵ 3415 ↵ 1959 ↵ 268 ↵
SV length ↵ (bp) ↵	31~50 ↵ 51~100 ↵ 101~1000 ↵ 1001~10000 ↵ >10000 ↵	14318 ↵ 9267 ↵ 13797 ↵ 6655 ↵ 3055 ↵	5905 ↵ 5694 ↵ 23545 ↵ 26107 ↵ 9 ↵
Total ↵	↪	47092 ↵	61260 ↵

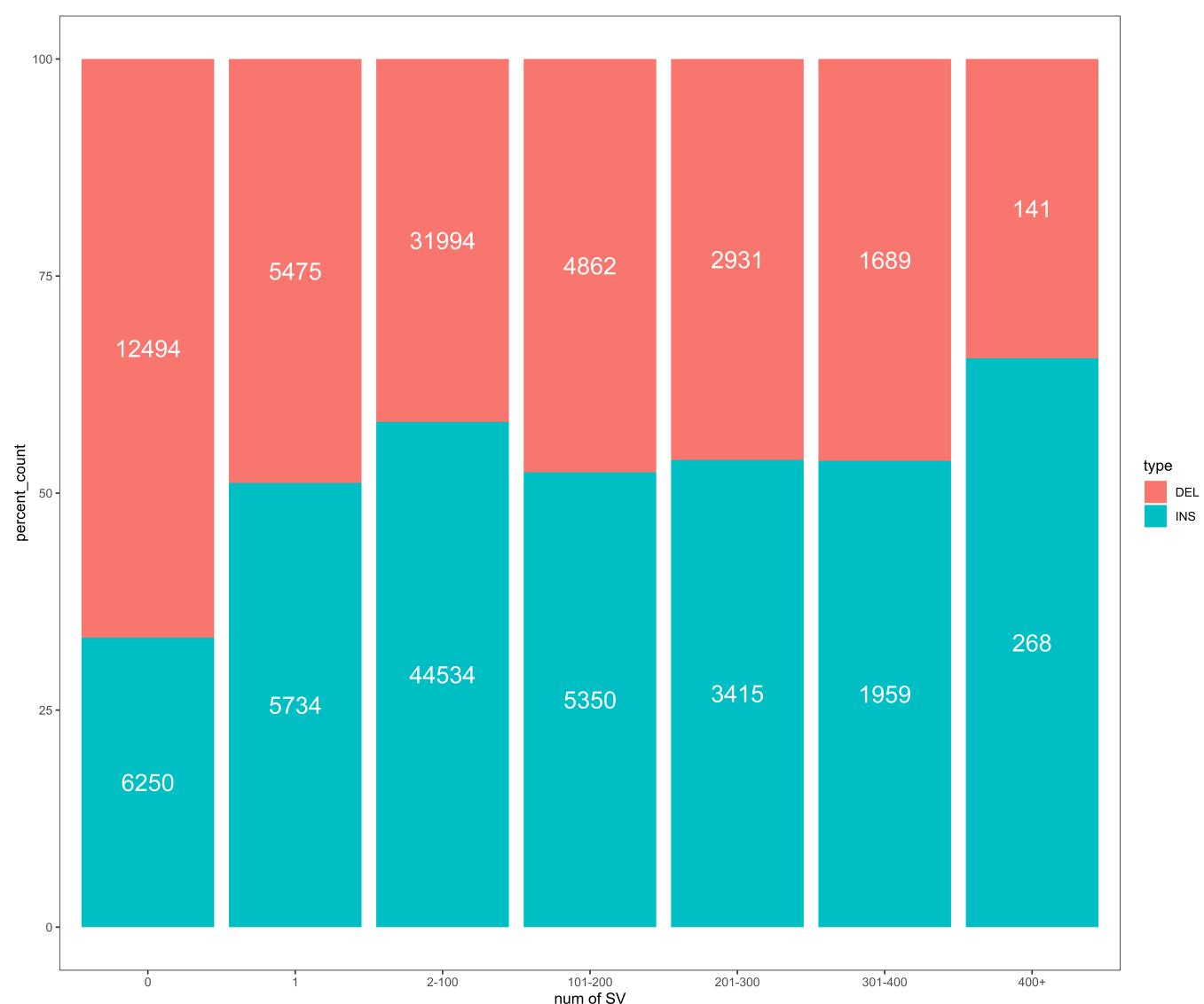


图4

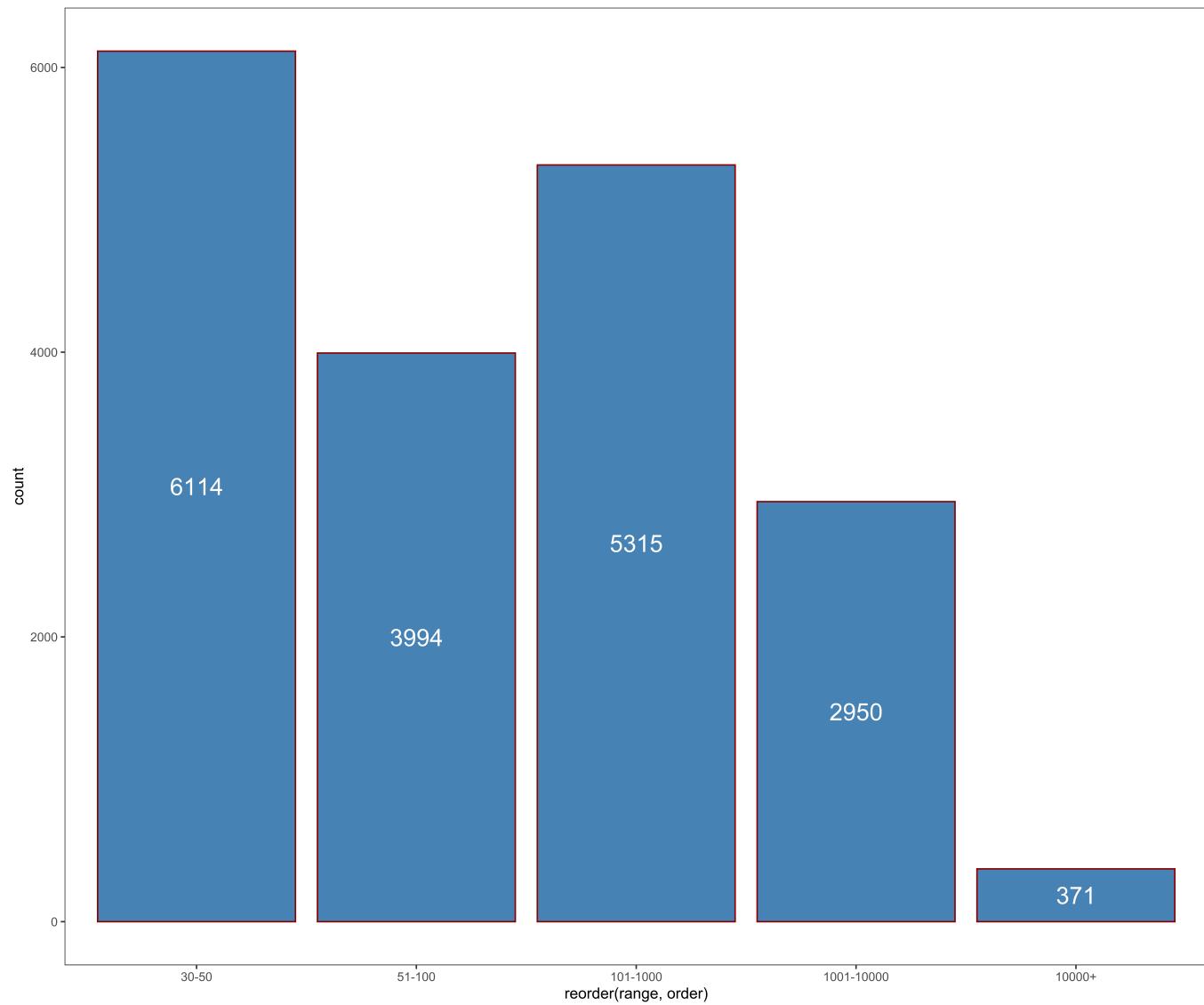


图5-1 102份参考基因组中未被453份材料中SV匹配到的SV分布

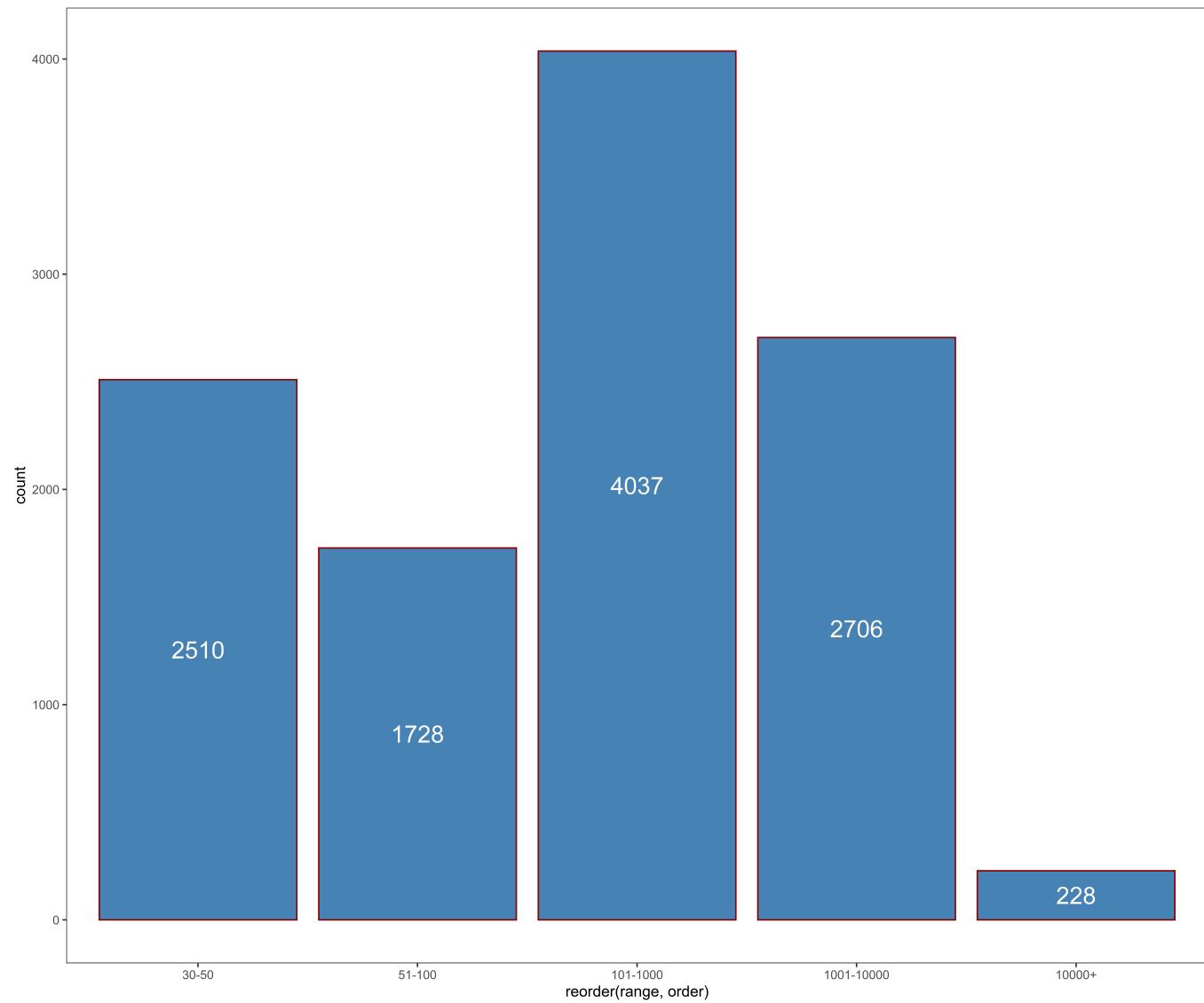


图5-2 102份参考基因组中稀有SV的分布

图6是102份参考基因组中SV在染色体上的分布情况。12条染色体上分布着不同类型的SV，其中DEL和INS数量最多，在12条染色体上都有分布，且DEL和INS往往会在同一位置。此外DUP数量相对较少，只分布在部分染色体上，7号和11号染色体上并没有DUP变异，4号染色体上分布的DUP变异数量最多。此外INV变异的数量是最少的，只有六个，分布于1号染色体的末端。

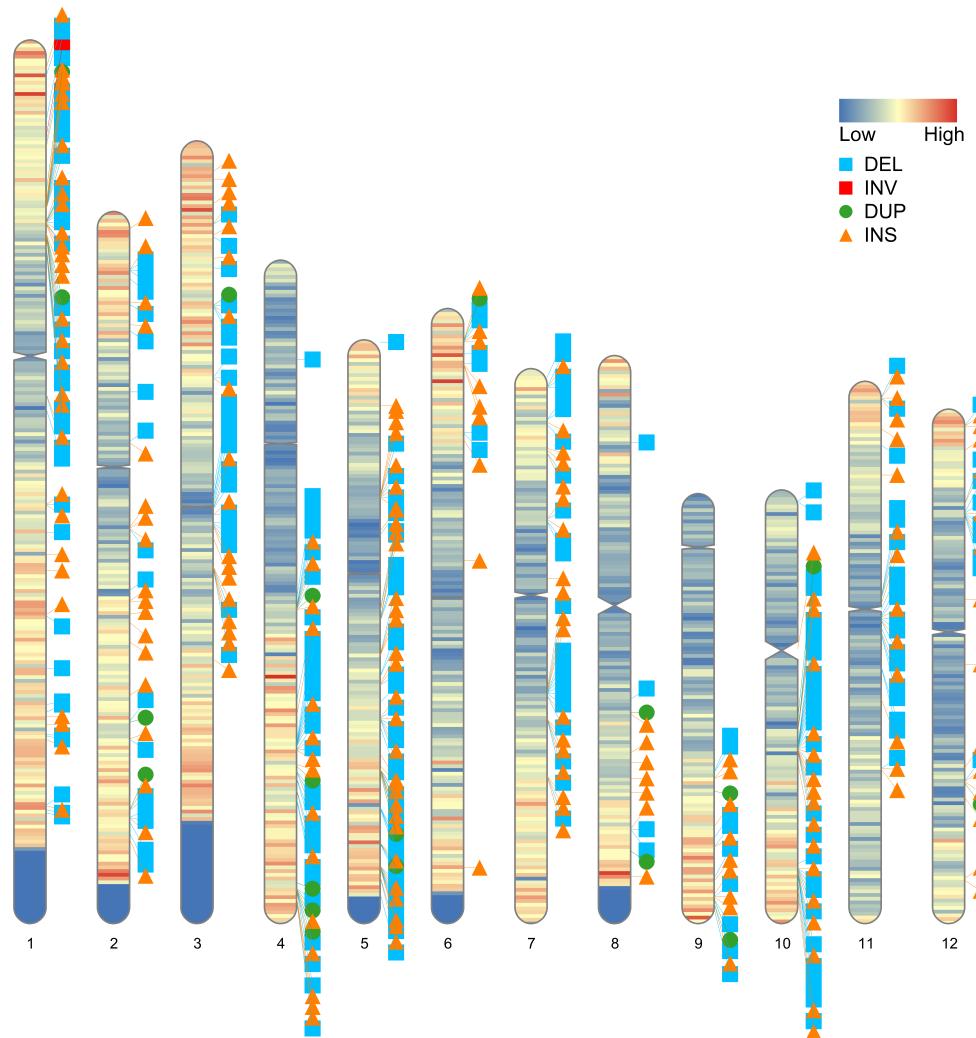


图6