基于结构冗余序列去重优化 DNA 测序数据压缩：补充材料

Optimizing DNA Sequencing Data Compression via Structural Redundant Reads Deduplication：Supplementary Data

S1 实验测试数据集

实验采用 NCBI 开源数据库(https://www.ncbi.nlm.nih.gov) 中 8 组真实开数据集用于实验评估。数据集下载使用 sra-tools 工具，其脚本配置参见：https://github.com/ncbi/sra-tools.

数据集 1： C.arietinum 鹰嘴豆

URL: https://www.ebi.ac.uk/ena/browser/view/SRR13556216

下载脚本：

```
cd SR2C/data
prefetch SRR13556216
fastq-dump SRR13556216
rm -rf SRR13556216 SRR13556216_2.fastq
```

数据集 2： Human 人类宏基因组

URL: https://www.ebi.ac.uk/ena/browser/view/SRR16553126

下载脚本：

```
prefetch SRR16553126
fastq-dump SRR16553126
rm -rf SRR16553126 SRR16553126_2.fastq
```

数据集 3&4：M.fascicularis 食蟹猕猴

URL: https://www.ebi.ac.uk/ena/browser/view/SRR8386204

下载脚本：

```
prefetch SRR8386204
fastq-dump SRR8386204
rm -rf SRR8386204
```

数据集 5&6：Mouse.tumor 小鼠肿瘤

URL: https://www.ebi.ac.uk/ena/browser/view/SRR17794741

下载脚本：

```
prefetch SRR17794741
fastq-dump --split-files SRR17794741
rm -rf  SRR17794741
```

数据集 7：S.fontinalis-1 美洲红点鲑

URL: https://www.ebi.ac.uk/ena/browser/view/SRR11995278

下载脚本：

```
prefetch SRR11995278
fastq-dump SRR11995278
```

```
rm -rf SRR11995278
```

数据集 8：S.fontinalis-2 美洲红点鲑

URL: https://www.ebi.ac.uk/ena/browser/view/SRR11994956

下载脚本：

```
prefetch SRR11994956
fastq-dump SRR11994956
rm -rf SRR11994956
```

S2 补充实验结果

表 S1. 算法 Minirmd 和 SR2C 在 8 组数据集运行得到的剩余序列数目(单位：条)

数据集 (Datasets)	*NMis* = 0		*NMis* = 1		*NMis* = 2		*NMis* = 3	
	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C
C.arietinum	27674276	**25538394**	25122345	**23712805**	23146415	**22014323**	21613589	**20624364**
Human	1134534	**1102078**	1020979	**1017296**	982487	**980033**	961736	**959526**
M.fascicularis-1	2330651	**2321698**	1213087	**1205610**	970240	**962876**	802288	**794989**
M.fascicularis-2	2270237	**2261396**	1260967	**1253703**	1020431	**1013267**	889445	**882339**
Mouse.tumor-1	416097	**411960**	204643	**204506**	106609	**106516**	66036	**65947**
Mouse.tumor-2	459250	**455642**	231544	**231292**	121670	**121500**	75559	**75404**
S.fontinalis-1	679936	**679731**	630201	**630072**	590035	**589932**	556059	**555965**
S.fontinalis-2	3392223	**3344524**	2990103	**2989442**	2728974	**2728446**	2515730	**2515288**
总计	38357204	**36115423**	32673869	**31244726**	29666861	**28516893**	27480442	**26473822**

表 S2. 算法 Minirmd 和 SR2C 在 8 组数据集运行得到的内存开销(单位：KB)

数据集 (Datasets)	*NMis* = 0		*NMis* = 1		*NMis* = 2		*NMis* = 3	
	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C
C.arietinum	13330652	**8540788**	18001044	**7850524**	18578944	**7278204**	19739192	**6824548**
Human	1324260	**343900**	2108904	**307072**	2206368	**296568**	2401484	**290532**
M.fascicularis-1	4953592	**716720**	7560836	**372956**	7887288	**292516**	8538700	**244796**
M.fascicularis-2	4953300	**697284**	7560860	**385708**	7886788	**306784**	8538580	**268436**
Mouse.tumor-1	888316	**143540**	1325736	**71888**	1381732	**38276**	1494224	**24796**
Mouse.tumor-2	888876	**157748**	1331868	**80812**	1385460	**43268**	1494312	**27952**
S.fontinalis-1	735724	**217376**	1123732	**201684**	1172236	**189896**	1269204	**179628**
S.fontinalis-2	5592384	**1064112**	7882296	**947280**	8165280	**867704**	8731848	**802712**
平均值	4083388	**1485184**	5861910	**1277241**	6083012	**1164152**	6525943	**802712**

表 S3. 算法 Minirmd 和 SR2C 在 8 组数据集运行得到的时间开销(单位：HH:MM:SS)

数据集 (Datasets)	*NMis* = 0		*NMis* = 1		*NMis* = 2		*NMis* = 3	
	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C	Minirmd	SR2C
C.arietinum	00:06:40	**00:57:33**	**00:29:21**	**00:52:03**	**00:17:14**	00:47:56	**00:13:06**	00:44:51
Human	**00:00:13**	00:00:54	**00:00:18**	00:00:47	**00:00:25**	00:00:44	00:00:29	**00:00:46**
M.fascicularis-1	**00:00:44**	00:02:11	00:01:47	**00:01:00**	00:01:50	**00:00:49**	00:02:04	**00:00:38**
M.fascicularis-2	**00:00:43**	00:02:07	00:01:49	**00:01:00**	00:01:53	**00:00:51**	00:02:10	**00:00:43**
Mouse.tumor-1	**00:00:24**	00:00:32	00:00:24	**00:00:14**	00:00:23	00:00:01	**00:00:25**	00:00:05
Mouse.tumor-2	**00:00:27**	00:00:51	00:00:35	**00:00:18**	00:00:27	**00:00:08**	00:00:22	**00:00:06**
S.fontinalis-1	**00:00:11**	00:00:43	**00:00:28**	00:00:41	**00:00:24**	00:00:34	**00:00:26**	00:00:31
S.fontinalis-2	**00:01:44**	00:05:05	00:04:56	**00:04:28**	00:04:25	**00:04:00**	00:04:20	**00:03:45**
平均值	**00:01:23**	00:08:45	**00:04:57**	00:07:34	**00:03:23**	00:06:53	**00:02:55**	00:07:20

表 S4. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的文件规模(单位：KB)

数据集 (Datasets)	Reads (KB)	Without SR2C				With SR2C			
		Pigz	PBzip2	XZ	7Z	Pigz	PBzip2	XZ	7Z
C.arietinum	3290620	861399	806193	**535604**	**533003**	**758942**	**713971**	555339	554035
Human	264860	**31592**	**24917**	**24903**	**24695**	33679	32998	24922	24963
M.fascicularis-1	1010380	298906	284658	97794	94049	**119432**	**107110**	**86370**	**89656**
M.fascicularis-2	1010380	298915	284688	98547	94854	**118534**	**106243**	**86652**	**89881**
Mouse.tumor-1	247098	35329	20419	10235	10195	**16963**	**13557**	**8575**	**8639**
Mouse.tumor-2	247098	35616	20835	10473	10432	**17766**	**14157**	**8974**	**9041**
S.fontinalis-1	166494	39403	34787	**13857**	**13871**	**21502**	**19407**	13899	13927
S.fontinalis-2	1384369	330651	292059	114968	113232	**139508**	**125452**	**94255**	**93647**
平均值	952662	241476	221070	113298	111791	**153291**	**141612**	**109873**	**110474**

表 S5. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的压缩比提升(单位：Ratio)

数据集(Datasets)	Reads (KB)	Pigz	PBzip2	XZ	7Z	平均值
C.arietinum	3290620	13.500%	12.917%	-3.554%	-3.796%	**4.767%**
Human	264860	-6.197%	-24.489%	-0.076%	-1.074%	**-7.959%**
M.fascicularis-1	1010380	150.273%	165.762%	13.227%	4.900%	**83.540%**
M.fascicularis-2	1010380	100.473%	47.171%	16.704%	15.385%	**44.933%**
Mouse.tumor-1	247098	108.271%	50.616%	19.359%	18.011%	**49.064%**
Mouse.tumor-2	247098	100.473%	47.171%	16.704%	15.385%	**44.933%**
S.fontinalis-1	166494	83.253%	79.250%	-0.302%	-0.402%	**40.450%**
S.fontinalis-2	1384369	137.012%	132.805%	21.975%	20.914%	**78.177%**
平均值	952662	**92.345%**	**78.999%**	**10.132%**	**7.434%**	**47.228%**

表 S6. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的压缩时间开销(单位：HH:MM:SS)

数据集 (Datasets)	Reads (KB)	Without SR2C				With SR2C			
		Pigz	PBzip2	XZ	7Z	Pigz	PBzip2	XZ	7Z
C.arietinum	3290620	00:09:02	00:00:46	**00:10:35**	00:14:53	**00:06:59**	**00:00:40**	00:10:39	**00:14:32**
Human	264860	00:00:14	00:00:07	00:02:07	00:02:25	**00:00:12**	**00:00:02**	**00:02:02**	**00:01:15**
M.fascicularis-1	1010380	00:02:40	00:00:16	00:05:08	00:05:32	**00:00:27**	**00:00:05**	**00:04:27**	**00:04:13**
M.fascicularis-2	1010380	00:02:33	00:00:16	00:05:39	00:05:18	**00:00:32**	**00:00:05**	**00:04:54**	**00:04:22**
Mouse.tumor-1	247098	00:00:13	00:00:05	00:01:27	00:01:31	**00:00:06**	**00:00:01**	**00:00:49**	**00:00:37**
Mouse.tumor-2	247098	00:00:19	00:00:05	00:01:25	00:01:33	**00:00:07**	**00:00:01**	**00:00:48**	**00:00:40**
S.fontinalis-1	166494	00:00:16	00:00:03	00:02:20	00:02:17	**00:00:10**	**00:00:01**	**00:01:15**	**00:00:54**
S.fontinalis-2	1384369	00:02:45	00:00:22	**00:03:50**	00:08:11	**00:00:45**	**00:00:07**	00:04:56	**00:03:57**
平均值	952662	00:02:15	00:00:15	00:04:04	00:05:12	**00:01:10**	**00:00:08**	**00:03:44**	**00:03:49**

表 S7. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的压缩内存开销(单位：KB)

数据集 (Datasets)	Reads (KB)	Without SR2C				With SR2C			
		Pigz	PBzip2	XZ	7Z	Pigz	PBzip2	XZ	7Z
C.arietinum	3290620	**6172**	**72116**	7599220	**3601116**	6688	75328	**7500016**	3672028
Human	264860	**5768**	**69628**	1467600	**697100**	6636	79500	**757996**	697172
M.fascicularis-1	1010380	**6340**	**73648**	4238156	3461328	6544	73812	**1656040**	**1523060**
M.fascicularis-2	1010380	**6224**	**73508**	4238928	3461552	6524	74052	**1653292**	**1495948**
Mouse.tumor-1	247098	**5896**	**69384**	1279856	697220	6500	77576	**647592**	**643532**
Mouse.tumor-2	247098	**5888**	**69524**	1280128	697156	6512	79336	**673248**	**664456**
S.fontinalis-1	166494	**6024**	**71408**	800936	697072	6560	75448	**687652**	**669464**
S.fontinalis-2	1384369	**6040**	**73136**	5831548	3520396	6396	74600	**2180468**	**1720896**
平均值	952662	**6044**	**71544**	3342047	2104118	6545	76207	**1969538**	**1385820**

表 S8. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的解压缩时间开销(单位：HH:MM:SS)

数据集 (Datasets)	Reads (KB)	Without SR2C				With SR2C			
		Pigz	PBzip2	XZ	7Z	Pigz	PBzip2	XZ	7Z
C.arietinum	3290620	**00:00:13**	**00:00:16**	**00:00:49**	**00:00:08**	00:00:15	00:00:19	**00:00:49**	00:00:10
Human	264860	**00:00:01**	**00:00:01**	**00:00:03**	**00:00:02**	**00:00:01**	**00:00:01**	**00:00:03**	**00:00:02**
M.fascicularis-1	1010380	00:00:05	00:00:07	00:00:11	**00:00:04**	**00:00:02**	**00:00:03**	**00:00:10**	00:00:06
M.fascicularis-2	1010380	00:00:05	00:00:07	00:00:11	**00:00:06**	**00:00:02**	**00:00:03**	**00:00:10**	**00:00:06**
Mouse.tumor-1	247098	**00:00:01**	**00:00:01**	00:00:02	**00:00:01**	**00:00:01**	**00:00:01**	**00:00:01**	**00:00:01**
Mouse.tumor-2	247098	**00:00:01**	**00:00:01**	00:00:02	**00:00:01**	**00:00:01**	**00:00:01**	**00:00:01**	**00:00:01**
S.fontinalis-1	166494	**00:00:01**	**00:00:01**	**00:00:02**	**00:00:01**	**00:00:01**	**00:00:01**	**00:00:02**	**00:00:01**
S.fontinalis-2	1384369	00:00:07	00:00:08	00:00:14	00:00:04	**00:00:03**	**00:00:03**	**00:00:11**	**00:00:06**
平均值	952662	00:00:04	00:00:05	00:00:12	**00:00:03**	**00:00:03**	**00:00:04**	**00:00:11**	00:00:04

表 S9. 算法 SR2C 级联通用压缩算法 Pigz、Pbzip2、XZ 和 7Z 在 8 组数据集运行得到的解压缩内存开销(单位：KB)

数据集 (Datasets)	Reads (KB)	Without SR2C				With SR2C			
		Pigz	PBzip2	XZ	7Z	Pigz	PBzip2	XZ	7Z
C.arietinum	3290620	992	**54896**	66564	**2445572**	**988**	59492	**66552**	2541752
Human	264860	1012	**48068**	**66560**	284252	**1008**	53212	**66560**	**131920**
M.fascicularis-1	1010380	1012	55232	66564	1081688	**1008**	**51992**	**66560**	**408480**
M.fascicularis-2	1010380	**1000**	51280	66564	1082336	1012	**50924**	**66560**	**405908**
Mouse.tumor-1	247098	**1008**	48856	66560	252568	**1008**	**48180**	**64492**	**73048**
Mouse.tumor-2	247098	**1008**	49912	**66564**	252836	**1008**	**48388**	**66564**	**77800**
S.fontinalis-1	166494	1012	**47676**	**66564**	177528	**1008**	48532	**66564**	**87836**
S.fontinalis-2	1384369	**1008**	55588	**66564**	1466436	**1008**	**53136**	**66564**	**532024**
平均值	952662	1007	**51439**	66563	880402	**1006**	51732	**66302**	**532346**