高倍率基因體組裝的加速與改進
--運用讀序選取新流程
Speeding & improving genome assembly of high-depth NGS data by read subset-driven workflows

Yu-Jung Chang
Institute of Information Science, Academia Sinica
2017/10/25

Outline

* Introduction
* Read subset selection

* Subset selection for paired-end (PE) reads
  * Experiments on the grouper dataset
* Discussion and conclusions

The genomic NGS data

* Recent progress in NGS technology has afforded several improvements
  * ultra-high throughput at **low cost,**
  * **very high read quality**, and
  * **substantially increased sequencing depth**

State-of-art NGS sequencers

* Illumina MiSeq system can generate ~15 Gbp per run,
  * with >80% bases above Q30 and
  * a sequencing depth of **up to several 1000x for small genomes.**
* Illumina HiSeq 2500 can generates up to 1 Tbp per run,
  * with >80% bases above Q30 and
  * **often >100x sequencing depth for large genomes**

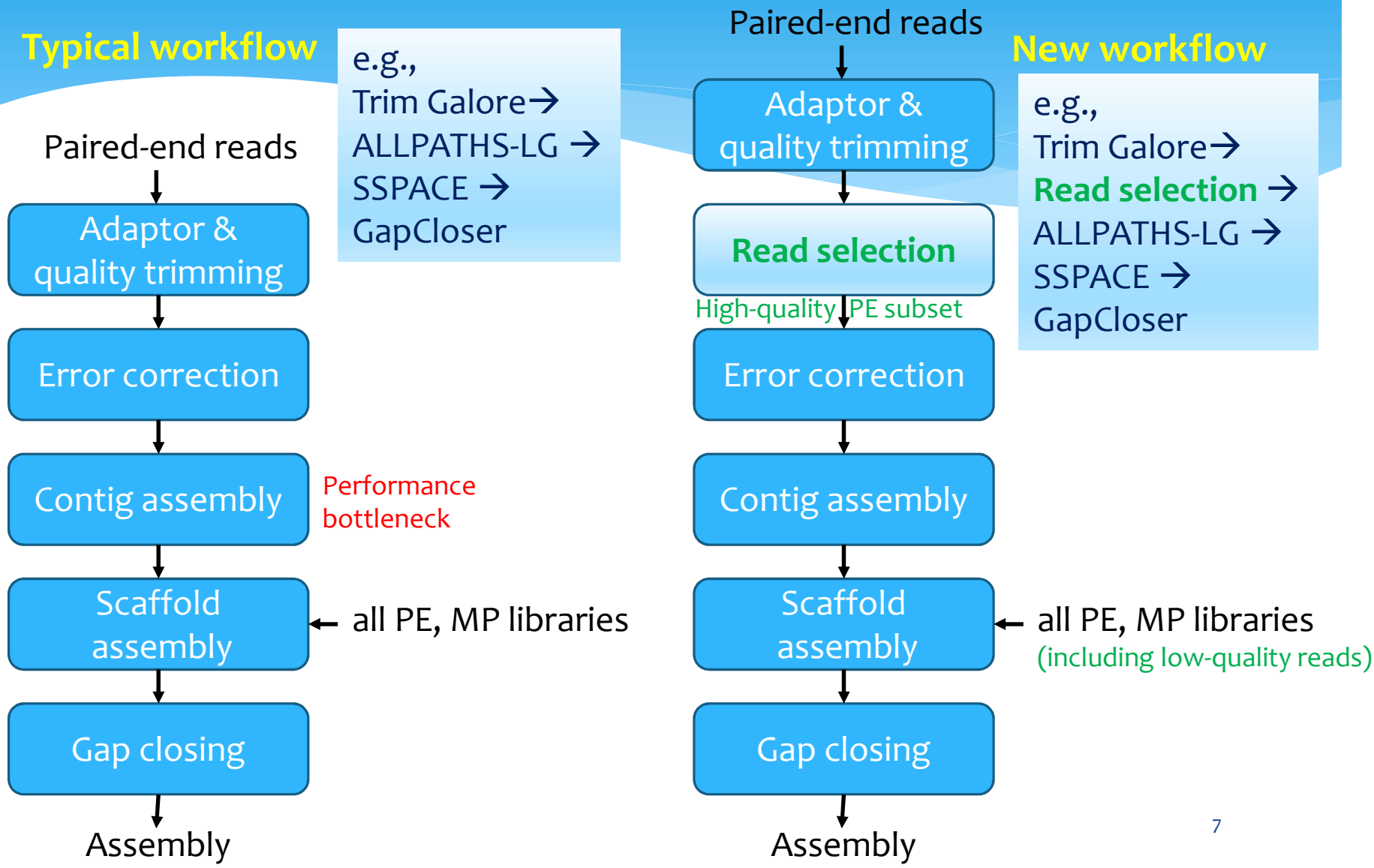
To assemble the high-depth NGS data

* Time-consuming and resource-demanding assembly
  * The ALLPATHS-LG assembly took 50+ days to assemble the 125G-bp grouper data (~110x depth) on a machine of 32 cores and 1TB RAM
  * The long analysis is a nightmare
    * When you want to try multiple sets of parameters
    * When power failure happens
* The more depth of NGS data, the better assembly results?
  * More redundancy of sequences
    * **Adv:** will help fixing errors & have more chance to close gaps
    * **Disadv:** will also have larger number errors in wider regions
      * More chance to form complex error patterns that cannot be solved by assemblers
      * Of course more time-consuming & resource-demanding
  * *How to reduce the disadvantages while retaining the advantages?*

Read subset selection for high-depth NGS data

* Idea:
  * Select high-quality reads to assemble into skeleton for further processing
    * Lower redundancy to speed up genome assembly
      * E.g., 110x → 50x
    * Can achieve almost equal or better assembly results
  * Low-quality reads are used later in the workflow

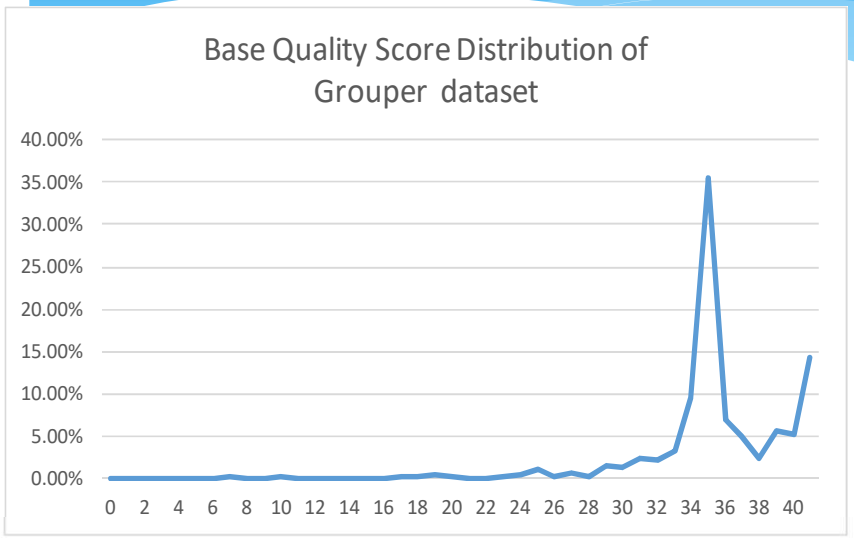
Genome assembly workflows



The MinimalQ strategy

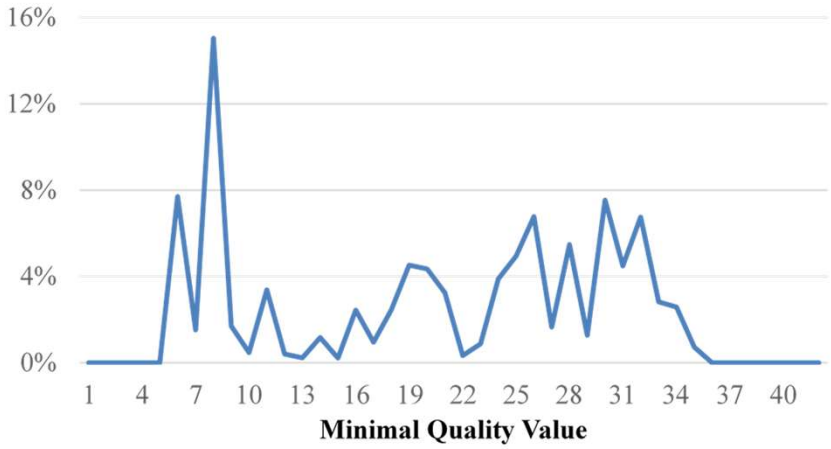
* *MinimalQ* strategy identifies the minimal quality value of each read, and
  * sets a threshold of **selecting reads with minimal quality value no smaller than the threshold**
    * Keep reads with MinimalQ $\geq$ t as the selected subset
  * Assumption: reads with very low MinimalQ values are likely to cause mis-assemblies
* MinimalQ of paired-end reads
  * MinimalQ of PE = min(MinimalQ of read1, MinimalQ of read2)
    * Both reads with MinimalQ $\geq$ t

Statistics of minimal quality value for the PEs in the grouper dataset

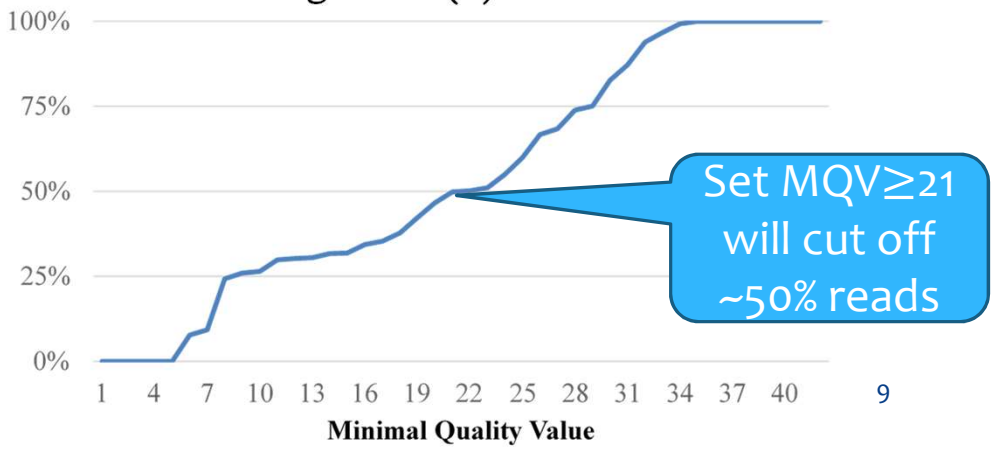


* Bases' quality
  * Peak: 35 (y: 36%)
* PEs' Min. quality value
  * Peak: 8 (y: 15%)

(a) MinmalQ distribution



(b) Cumulative figure of (a)



Experiments of PE subset selection

* Dataset: the giant grouper dataset of HiSeq
* Selection strategy: MinimalQ
* Assembler: ALLPATHS-LG
* Evaluation: QUAST quality assessment tool

Experimental steps for PEs

* Step 1:
  * A 50% subset of PEs was selected using MinimalQ strategy
    * Use MQV-21
* Step 2:
  * ALLPATHS-LG assembler was used to obtain contigs and scaffolds
  * 5 mate-pair libraries, with insert lengths ~2K, ~4K, ~6K, ~8K, ~10K bp, were used for both the original and selected datasets
    * The size of each mate-pair library is ~4.4G bp
* Step 3:
  * The results were evaluated using QUAST assessment tool

Comparing the assembly results of PE subset selection for the grouper dataset

		Original dataset	Selected subset
Dataset characteristics			
	Dataset size (G bp)	125	63
	# read pairs	319,878,932	158,651,599
	Mean length of reads	195.3	198.6
	%GC content of reads	41.0%	39.7%
Results of contigs			
	# contigs	39,911	53,488
	Total contig length	996,203,993	991,109,739
	N50 contig size (K bp)	82.2	43.5

Results of scaffolds

		Original dataset	Selected subset
	# scaffolds	3,917	4,043
	Total scaffold length	1,076,396,971	1,062,462,514
	Largest scaffold length	12,701,604	21,777,629
	N50 scaffold size (K bp) (L50 number)	3,354 (97 scaffolds)	5,443 (61 scaffolds)
	N75 scaffold size (K bp) (L75 number)	1,429 (218 scaffolds)	2,493 (131 scaffolds)
	%GC of scaffolds	41.23%	41.17%
	# 'N's	79,902,759	71,510,549
	# 'N's per 100K bp	7,423.10	6,730.57
	# scaffolds for 1G bp [3]	482	304

QUAST results: cumulative length, GC%

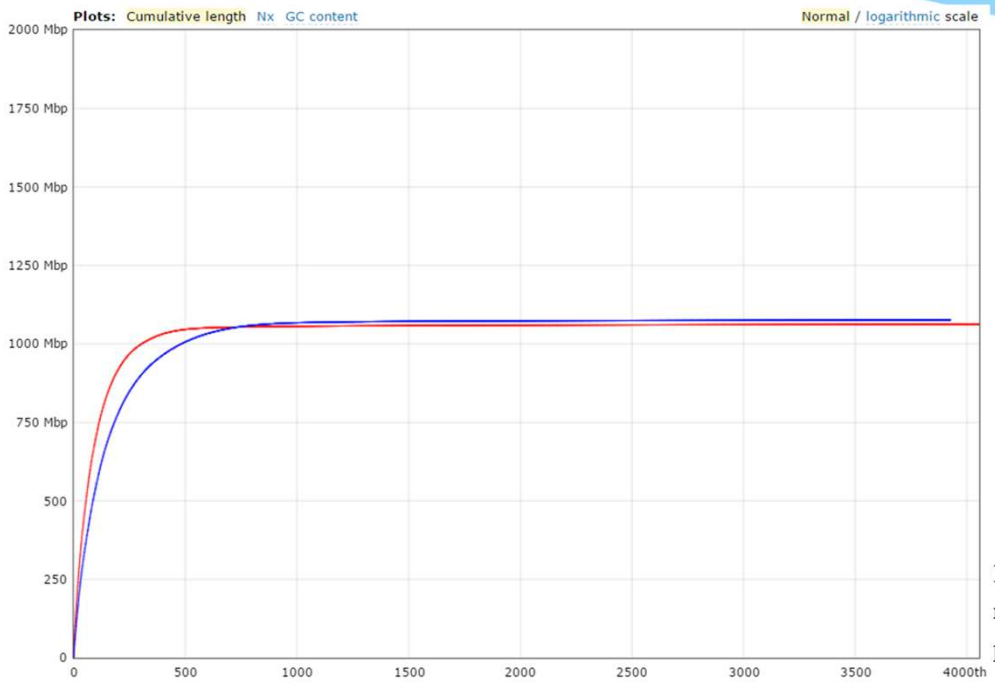


Figure S6. Comparison of the cumulative length of scaffolds. The *x*-axis denotes the top *x* long scaffolds (ordered from largest (scaffold #1) to smallest). The *y*-axis denotes their cumulative length. The original dataset uses blue curve; the selected subset uses red curve

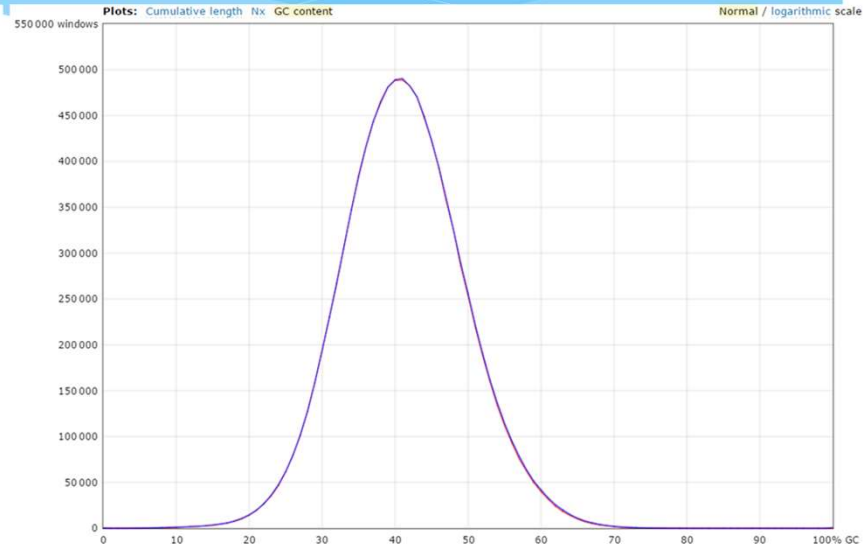


Figure S8. Comparison of the GC content of scaffolds. Scaffolds are broken into nonoverlapping 100 bp windows. The figure shows numbers of windows for each GC percentage. The original dataset uses blue curve; the selected subset uses red curve.

QUAST results: N*x* plot

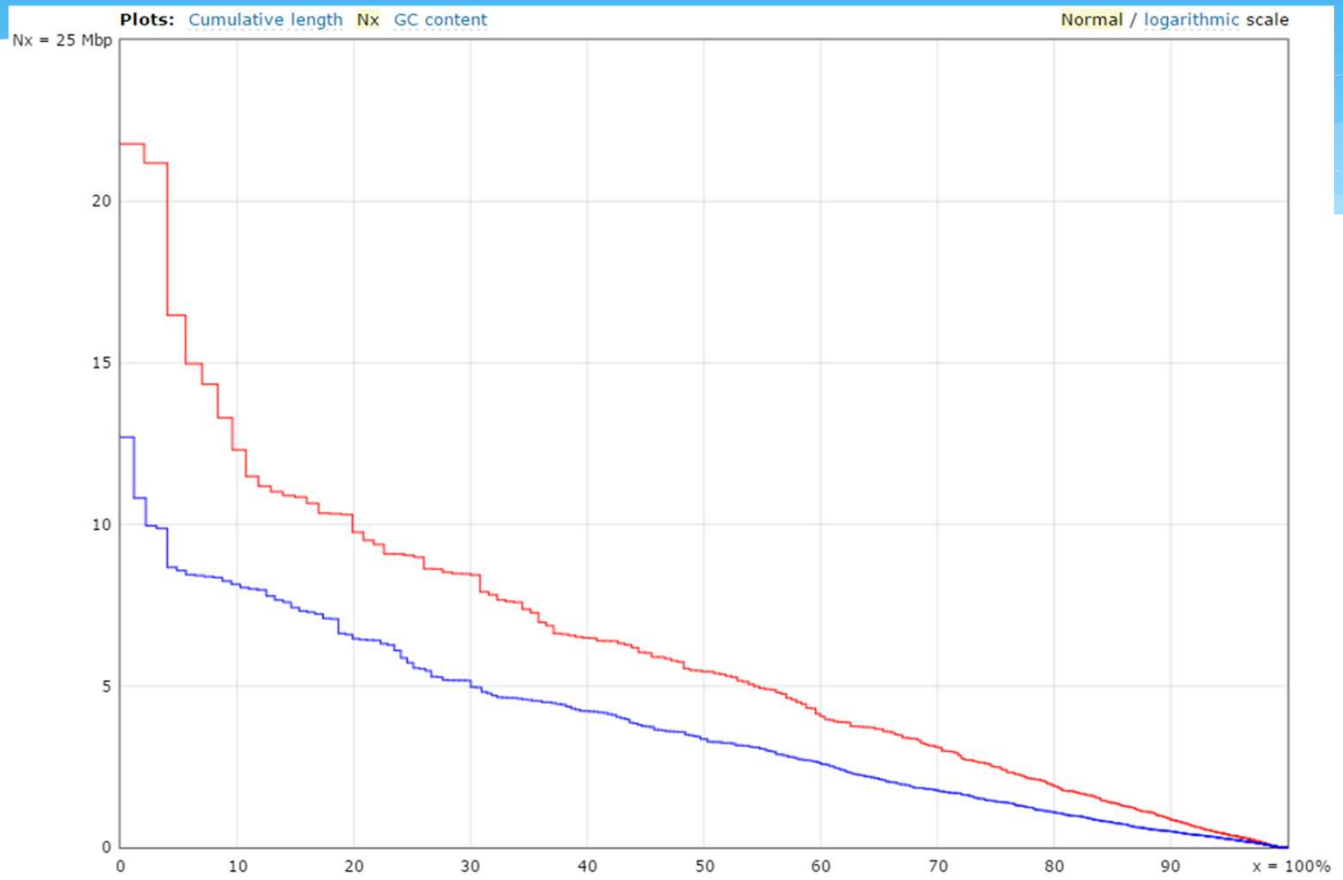


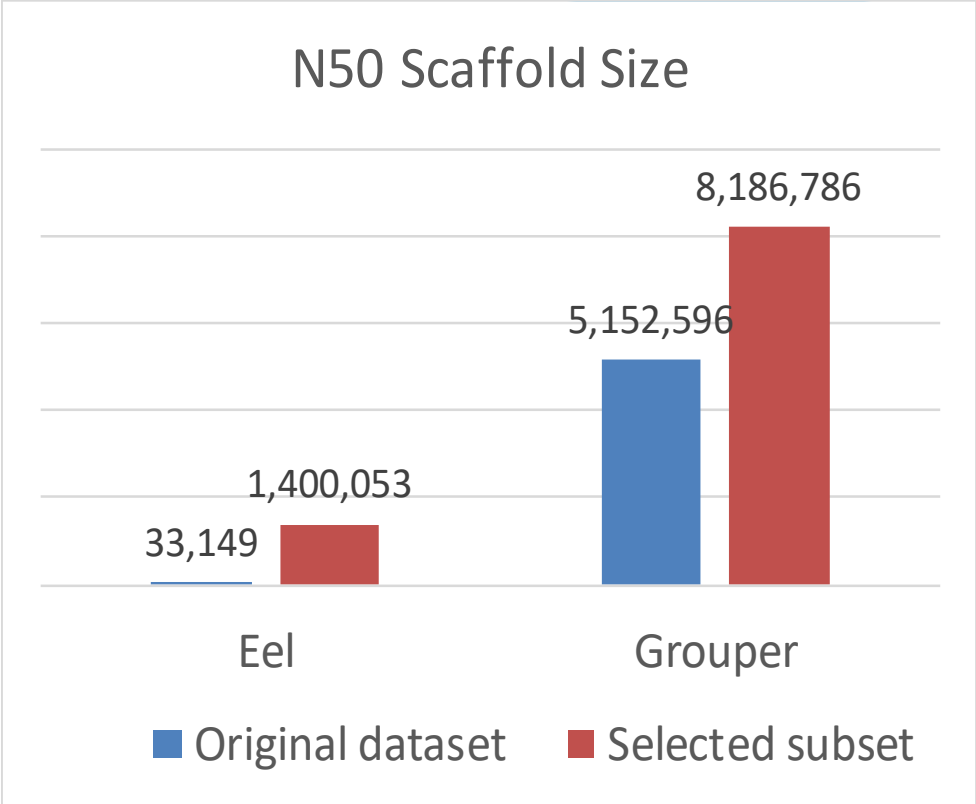
Figure S7. Comparison of N*x* of scaffolds. N*x* (where $0 \leq x \leq 100$) is the largest scaffold length, L, such that using contigs of length ≥ L accounts for at least *x*% of the bases of the assembly. The original dataset uses blue curve; the selected subset uses red curve.

Runtimes of the two ALLPATHS-LG assemblies

* The original dataset
  * ALLPATHS-LG took 50+ days
    * on a machine of 32 cores and 1TB RAM
      * Peak RAM usage: ~600GB
* The selected subset
  * ALLPATHS-LG took 7.2 days
    * on a machine of 40 cores and 1.5TB RAM
      * Peak RAM usage: ~390GB
  * PE subset selection took ~2 hr
    * on a 10-node Hadoop cluster
    * Open source software
      * https://github.com/moneycat/QReadSelector

Eel genome assemblies by Trim+[Select]+ALP+SS+GC workflow

Eel	Original dataset	Selected subset
Dataset size	135 G bp	49 G bp
Total scaffold length	333 M bp	1028 M bp
Largest scaffold length	1,762,478	9,640,288
N50 scaffold size	33,149	1,400,053



Guidelines for applying MinimalQ strategy to your datasets

* **Suitable for genome assembly of high-depth NGS data**
  * Suggest to select reads with at least 50x coverage depth
    * we suggest determining the initial subset sizes by sufficient coverage depths.
    * Desai et al [1] suggest that 50x data is enough to get good genome coverage for assemblies of small and moderate sized genomes
    * For large genomes, we suggest to initial at 50x-60x
* **Suitable for genome assembly with multiple PE & MP libraries**
  * If the dataset has only one or two PE, the gain of scaffolding would become inferior.
* **Suitable for datasets with good quality**
  * poor-quality datasets → right-skewed MinimalQ distribution
    * Most reads have low MinimaQ (say MinimalQ < Q15)
  * The read length also affects MinimalQ values
  * ➔ We provide a new read subset selection method %HighQ(*x*)

[1] Desai et al. **Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data.** *PLoS One*. 2013.

Read subset selection by high-quality percentage

* $\%HighQ(x) = \frac{\#bases\ with\ quality\ value \geq x}{\#bases\ of\ a\ read}$
  * $MinimalQ \geq x$ can be specified as $\%HighQ(x) \geq 100$
    * i.e., 100% of bases with quality values $\geq x$
  * You can also specify $\%HighQ(20) \geq 95$
    * i.e., 95% of bases with quality values ≥ 20
  * Guideline:
    * Make sure the dataset is high-depth and has multiple PE & MP

Software demonstration (1/2)

* Step0. Put read1, read2 fastq files in a folder, e.g., C:\data
* Step1. Probe the quality distribution of PE reads
  * Usage: **peQdist** read1.fq read2.fq out.csv
  * Windows 8+ edu/pro/… (needs DOCKER for windows)
    * **> docker run -v C:\data:/data abnerchang/subset /ngs/peQdist /data/read1.fq /data/read2.fq /data/out.csv**
      * PS. The system auto-downloads the docker image for the 1st time.
  * Linux
    * $ sudo docker run -v /data:/data abnerchang/subset /ngs/peQdist /data/read1.fq /data/read2.fq /data/out.csv
* Step2. Check the quality distribution and determine the thresholds
  * Say MinmalQ ≥ 21, %HighQ(20) ≥ 98.5

Software demonstration (2/2)

* Step2. Select the high-quality subset of PE reads
  * Usage: **peQselect** r1.fq r2.fq outPrjName %HighQ [QTh]
    * QTh: threshold x of quality value; default 20; range [0..41]
    * Output: <outPrjName>-r1.fq, <outPrjName>-r2.fq
  * Windows
    * MinimalQ $\geq$ 21
      * **> docker run -v C:\data:/data abnerchang/subset /ngs/peQselect /data/read1.fq /data/read2.fq /data/subsetMinQ-21 100 21**
    * %HighQ(20) $\geq$ 98.5
      * **> docker run -v C:\data:/data abnerchang/subset /ngs/peQselect /data/read1.fq /data/read2.fq /data/subsetHiQ-98.5 98.5**

Thank You!

* References
  * Chih-Hao Fang, Yu-Jung Chang, Wei-Chun Chung, Ping-Heng Hsieh, Chung-Yen Lin and Jan-Ming Ho, **Subset selection of high-depth next generation sequencing reads for de novo genome assembly using MapReduce framework**. *BMC Genomics*, volume 16, number Suppl 12, pages S9, Dec. 2015.
* Software
  * https://hub.docker.com/r/abnerchang/
  * Welcome to give comments/suggestions to yjchang@iis.sinica.edu.tw