Supplemental information

Datasets

The datasets used in the paper are publicly available at:

I: http://1001genomes.org/data/software/shoremap/shoremap_2.0\/data/reads/Schneeberger.2009/Schneeberger.2009.single_end.gz

II: http://1001genomes.org/data/software/shoremap/shoremap_2.0/data/reads/Galvao.2012/Galvao.2012.reads1.fq.gz,http://1001genomes.org/data/software/shoremap/shoremap_2.0/data/reads/Galvao.2012/Galvao.2012.reads2.fq.gz

III: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR611/SRR611084//SRR611084.sra,ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR611/SRR611085//SRR611085.sra

IV: artificial pair-ended dataset for *A.thaliana* created with the `wgsim` program from the Samtools package.

V: http://www.ncbi.nlm.nih.gov/sra/SRX148888

Reference genomes

- TAIR10 for datasets II-IV ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/*.fas
- TAIR8 for dataset I ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/
- H.sapiens, NCBI v37 ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/h_sapiens_37_asm.ebwt.zip

Description of computational facilities

1. HPC: Multinode short-read alignment was performed on the Milou cluster (http://www.uppmax.uu.se/the-milou-cluster), equipped with dual 8-core Intel Xeon E5-2660, (2.2 GHz, 2 MB L2 cache, 20 MB

L3 cache), 128 GB of RAM, Infiniband node-to-node network connection, and 10Gbit/s uplink.

2. Storage: Gulo (http://www.uppmax.uu.se/gulo) is a custom built Lustre 2.4 system using 8 HP nodes with MDS600 storage boxes and an additional node for metadata handling. In total, it provides roughly 1 PB of storage and is accessed with Lustre's own protocol. It supports data striping over multiple nodes and disk targets and can give a theoretical single file read performance of up to 80 Gbits per second.

3. Our Hadoop test platform was deployed on a private cloud at UPPMAX using the OpenNebula (http://opennebula.org) cloud management system. Each physical node was equipped with two 4-core Intel Xeon 5420 (2.50 GHz; 12 MB L2 cache), 16 GB RAM, one 1 TB SATA disk and Gigabit Ethernet. The cluster was set up with Cloudera Hadoop Distribution version 2.0.0-cdh4.5.0 (http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html). Note that the physical hardware provided less RAM than desired. Each VM node has 7 cores and 14 GB of RAM (2 GB/core), which is less than recommended for Hadoop. We would expect to see much better performance with twice as much memory.

Crossbow preprocessing stage as a Bash script

The following Bash script mimics functionality of the Crossbow's preprocessing stage, and uses multi-core parallelism. The script reformats the `bzip2`-compressed short-reads in FASTQ format into a single text file where each line contains following tab-separated fields:

- read header
- forward read
- forward read qualities
- reverse read
- reverse read qualities

```
   pbzip2 -dc $file1 | paste - - - - -d'\t' | cut -f1,2,4 | paste
- -d' ' <(pbzip2 -dc $file2) | paste - - - - -d'\t' | cut -f2,4)
| pbzip2 -cz > $fileOut
```