PACIFIC BIOSCIENCES®

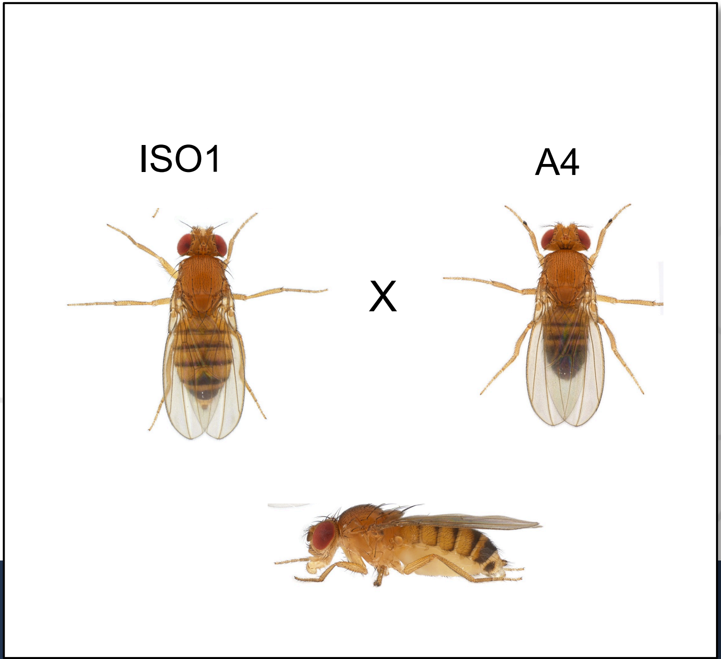
Assembling the drosophila genome with IPA and HiFi data

Zev Kronenberg

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2020 by Pacific Biosciences of California, Inc. All rights reserved.

AGENDA

- Introduction
- Setup PacBio software
- Learn about command line interface
- Run an IPA assembly



Today’s dataset

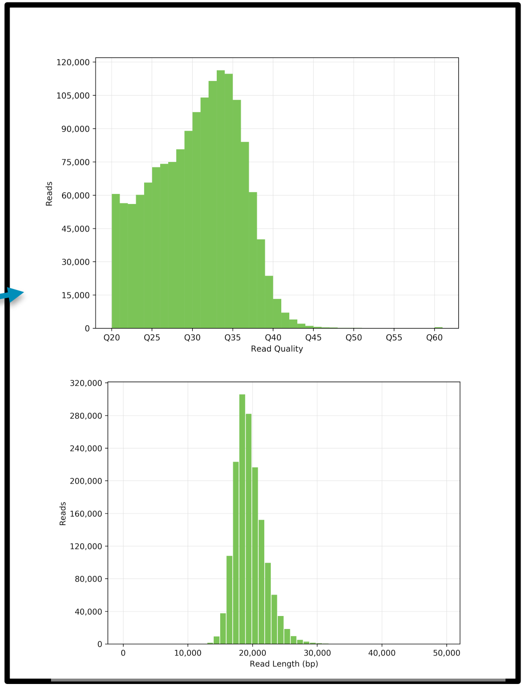
Drosophila

DROSOPHILA DATASET INFO

19 Kb dataset

- Processing of PacBio data
  - Circular consensus algorithm was done with SMRT Link
  - Data was subsampled down to 38x depth of coverage (DOC)

- Short read data
  - Standard procedures
  - Both parental strains were sequenced (70-90x DOC)
  - Utility
    - Trio binning
    - Phasing evaluation



SETUP CONDA – A SOFTWARE MANAGEMENT SYSTEM

- Google: `conda install`
  - https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html
- Follow link for download (linux x86)
- Follow install directions, setup BASH shell!

- You will likely need to source your .bashrc if conda isn't in your path after setup

- Why does PacBio use Conda?
  - Central repository of many bioinformatic tools
  - We can distribute binary code
  - We can post updates quickly
  - Many more…

SETUP IPA

PACBIO® & BIOCONDA®

- Google: `pacbio bioconda`
  - https://github.com/PacificBiosciences/pbbioconda
  - Familiarize yourself with available command line tools available for download.
  - Go to IPA bioconda wiki:
  - https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler

While efforts have been made to ensure that releases on Bioconda live up to the quality that PacBio strives for, we make no warranty regarding any Bioconda release.

SETUP IPA

PACBIO® & BIOCONDA®

```
conda create -n ipa -c
bioconda -c conda-forge -c
defaults conda activate ipa
conda install pbipa
```

UC DAVIS SETUP

PACBIO & BIOCONDA

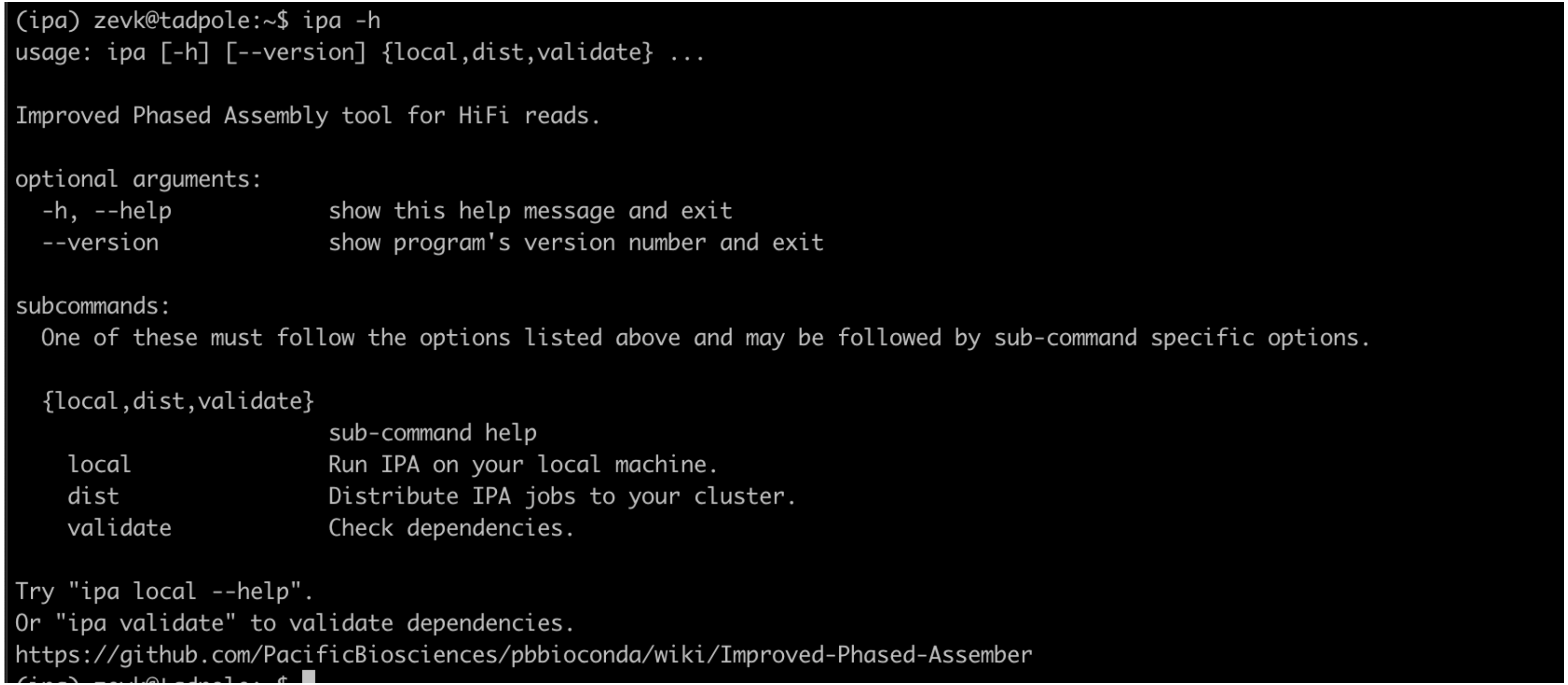
```
eval "$(/share/biocore/shunter/2020-07-15-IPA-tests/conda/bin/conda shell.bash hook)"
```

IPA

IMPROVED PHASED ASSEMBLY

COMMAND LINE

PACBIO® & BIOCONDA®

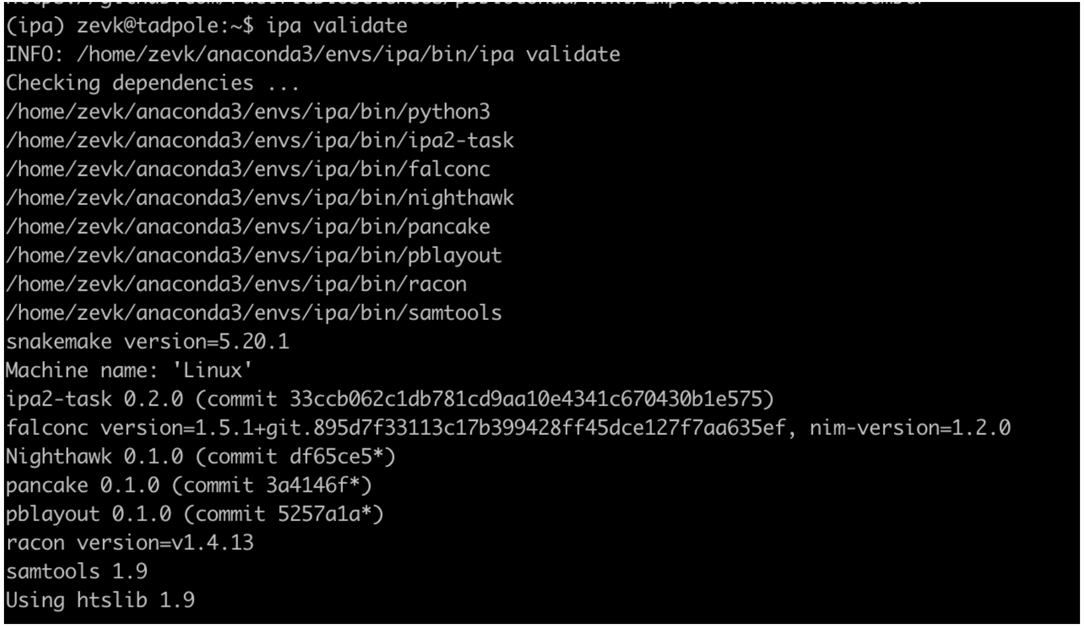


```
(ipa) zevk@tadpole:~$ ipa -h
usage: ipa [-h] [--version] {local,dist,validate} ...

Improved Phased Assembly tool for HiFi reads.

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit

subcommands:
  One of these must follow the options listed above and may be followed by sub-command specific options.

  {local,dist,validate}
                        sub-command help
    local               Run IPA on your local machine.
    dist                Distribute IPA jobs to your cluster.
    validate            Check dependencies.

Try "ipa local --help".
Or "ipa validate" to validate dependencies.
https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assember
(ipa) zevk@tadpole:~$ 
```

KNOW YOUR VERSIONS

PACBIO® & BIOCONDA®

- We commonly update IPA
- Before running assembly update IPA
- Keep track of your versions
- Avoid updating mid-assembly



```
https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler
(ipa) zevk@tadpole:~$ ipa validate
INFO: /home/zevk/anaconda3/envs/ipa/bin/ipa validate
Checking dependencies ...
/home/zevk/anaconda3/envs/ipa/bin/python3
/home/zevk/anaconda3/envs/ipa/bin/ipa2-task
/home/zevk/anaconda3/envs/ipa/bin/falconc
/home/zevk/anaconda3/envs/ipa/bin/nighthawk
/home/zevk/anaconda3/envs/ipa/bin/pancake
/home/zevk/anaconda3/envs/ipa/bin/pblayout
/home/zevk/anaconda3/envs/ipa/bin/racon
/home/zevk/anaconda3/envs/ipa/bin/samtools
snakemake version=5.20.1
Machine name: 'Linux'
ipa2-task 0.2.0 (commit 33ccb062c1db781cd9aa10e4341c670430b1e575)
falconc version=1.5.1+git.895d7f33113c17b399428ff45dce127f7aa635ef, nim-version=1.2.0
Nighthawk 0.1.0 (commit df65ce5*)
pancake 0.1.0 (commit 3a4146f*)
pblayout 0.1.0 (commit 5257a1a*)
racon version=v1.4.13
samtools 1.9
Using htslib 1.9
```

RUN IPA ON THE CLUSTER

PACBIO® & BIOCONDA®

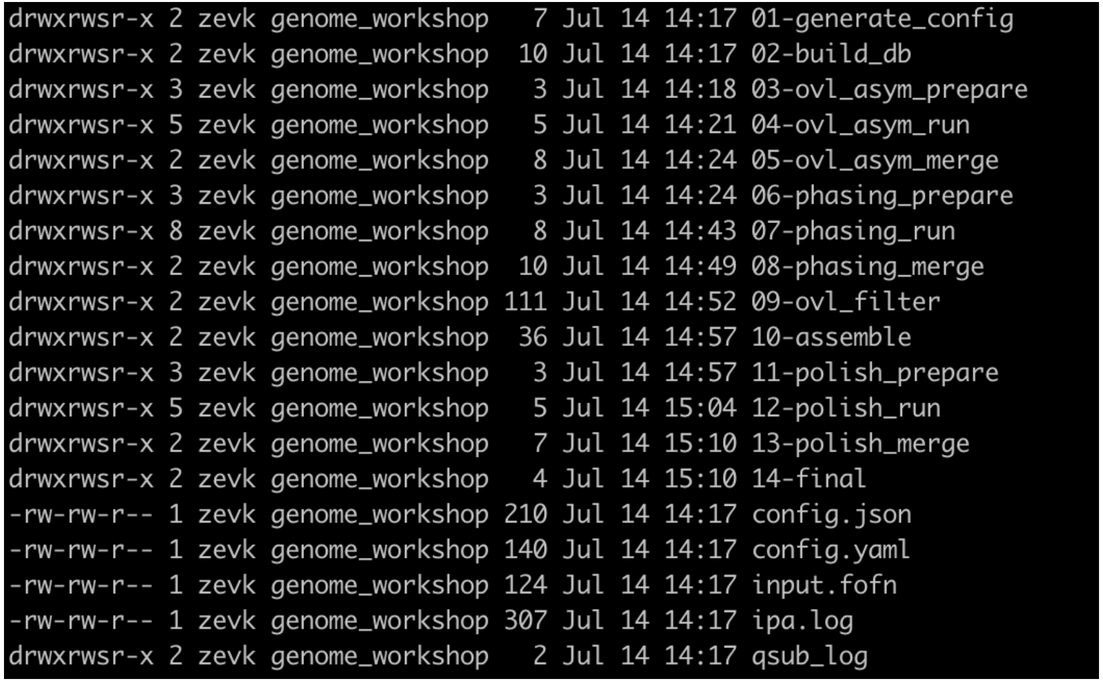
```
ipa dist -i ../hifi_long_read_data/ELF_19kb.m64001_190914_015449.Q20.38X.fasta \
--nthreads 24 --njobs 30 --cluster-args 'sbatch -J zev-ipa.{rule} -t 45  \
-c {params.num_threads} -e stderr -o stdout --get-user-env \
--chdir pacbio_2020_data_drosophila/hifi_long_read_diploid_ipa_assembly_cluster '
```

RUN IPA LOCALLY

PACBIO® & BIOCONDA®

```
ipa local --nthreads 48 --njobs 2 -i ELF_19kb.m64001_190914_015449.Q20.38X.fasta
```

STAGES OF IPA



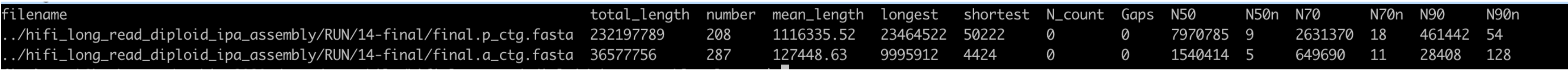
```
drwxrwsr-x 2 zevk genome_workshop   7 Jul 14 14:17 01-generate_config
drwxrwsr-x 2 zevk genome_workshop  10 Jul 14 14:17 02-build_db
drwxrwsr-x 3 zevk genome_workshop   3 Jul 14 14:18 03-ovl_asym_prepare
drwxrwsr-x 5 zevk genome_workshop   5 Jul 14 14:21 04-ovl_asym_run
drwxrwsr-x 2 zevk genome_workshop   8 Jul 14 14:24 05-ovl_asym_merge
drwxrwsr-x 3 zevk genome_workshop   3 Jul 14 14:24 06-phasing_prepare
drwxrwsr-x 8 zevk genome_workshop   8 Jul 14 14:43 07-phasing_run
drwxrwsr-x 2 zevk genome_workshop  10 Jul 14 14:49 08-phasing_merge
drwxrwsr-x 2 zevk genome_workshop 111 Jul 14 14:52 09-ovl_filter
drwxrwsr-x 2 zevk genome_workshop  36 Jul 14 14:57 10-assemble
drwxrwsr-x 3 zevk genome_workshop   3 Jul 14 14:57 11-polish_prepare
drwxrwsr-x 5 zevk genome_workshop   5 Jul 14 15:04 12-polish_run
drwxrwsr-x 2 zevk genome_workshop   7 Jul 14 15:10 13-polish_merge
drwxrwsr-x 2 zevk genome_workshop   4 Jul 14 15:10 14-final
-rw-rw-r-- 1 zevk genome_workshop 210 Jul 14 14:17 config.json
-rw-rw-r-- 1 zevk genome_workshop 140 Jul 14 14:17 config.yaml
-rw-rw-r-- 1 zevk genome_workshop 124 Jul 14 14:17 input.fofn
-rw-rw-r-- 1 zevk genome_workshop 307 Jul 14 14:17 ipa.log
drwxrwsr-x 2 zevk genome_workshop   2 Jul 14 14:17 qsub_log
```

GETTING ASM STATS

PACBIO® & BIOCONDA®

module load assembly_stats/1.0.1

assembly-stats -t final.p_ctg.fasta final.a_ctg.fasta | column -t



```
filename                                                   total_length  number  mean_length  longest   shortest  N_count  Gaps  N50      N50n  N70      N70n  N90     N90n
../hifi_long_read_diploid_ipa_assembly/RUN/14-final/final.p_ctg.fasta  232197789     208     1116335.52   23464522  50222     0        0     7970785  9     2631370  18    461442  54
../hifi_long_read_diploid_ipa_assembly/RUN/14-final/final.a_ctg.fasta  36577756      287     127448.63    9995912   4424      0        0     1540414  5     649690   11    28408   128
```

LOOKING AT YOUR ASM

PACBIO® & BIOCONDA®

Make GFA

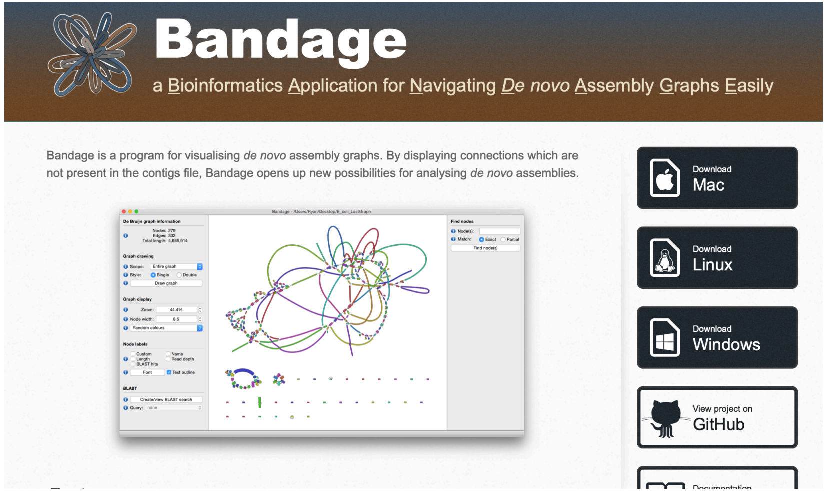
```
pblayout subgraph --tp all_ctg.tp sg_edges_list tp_graph
```

Remove Double Edges

```
perl -lane '$_ =~ s/:B|:E//g; print' tp_graph.csv  >
tp_graph_single.csv ; perl -lane '$_ =~ s/:B|:E//g;
print' tp_graph.gfa  > tp_graph_single.gfa
```

LOOKING AT YOUR ASM

PACBIO® & BIOCONDA®



Bandage

a Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily

Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.

Bandage - /Users/Ryan/Desktop/E_coli_LastGraph

De Bruijn graph information

Nodes: 279
Edges: 332
Total length: 4,685,914

Graph drawing

Scope: Entire graph
Style: ● Single ○ Double

Graph display

Zoom: 44.4%
Node width: 8.5
Random colours

Node labels

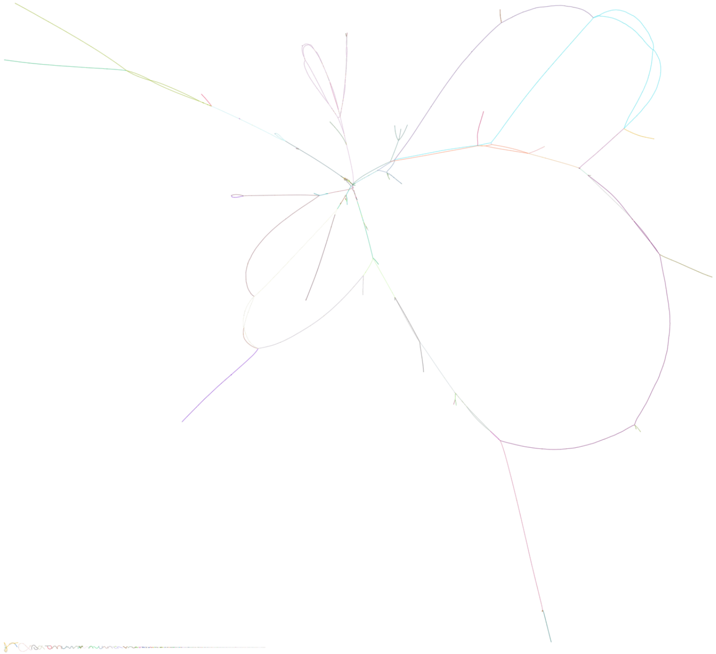
□ Custom □ Name
□ Length □ Read depth
□ BLAST hits
☑ Text outline

BLAST

Query: none

Find nodes

Node(s):
Match: ● Exact ○ Partial



PACBIO®

www.pacb.com

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2020 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. Pacific Biosciences does not sell a kit for carrying out the overall No-Amp Targeted Sequencing method. Use of these No-Amp methods may require rights to third-party owned intellectual property. FEMTO Pulse and Fragment Analyzer are trademarks of Agilent Technologies Inc.

All other trademarks are the sole property of their respective owners.