De Novo Genome Metassembly

Alejandro H. Wences*, Paul Baranay, Michael Schatz
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory

*alhernan@cshl.edu

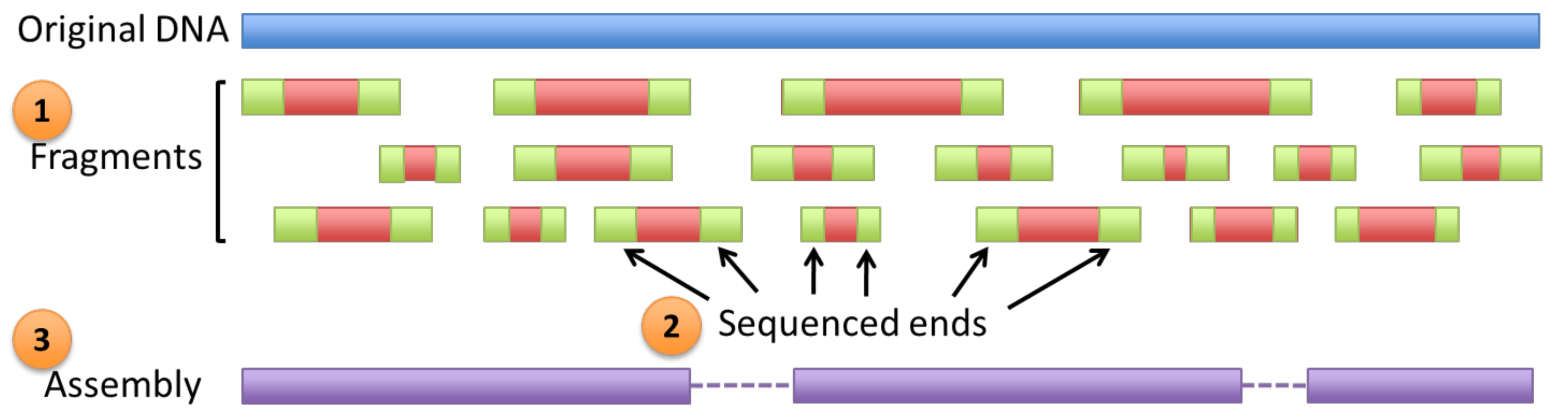
Summary

Sequencing projects typically create several draft assemblies of the genome either by employing several different assemblers or by incrementally adjusting the input parameters of a single assembler. The latter is of special importance for de Bruijn graph based methods where the choice of K-mer size can have a dramatic influence on the assembly quality and contiguity. In fact, there may even be different optimal values for different parts of the genome depending on their repeat or sequence composition. Today, genome sequencing projects usually select a single draft assembly, with a single set of parameters, as the candidate for publication. Instead of discarding the extra assemblies, we propose using them in the process of "metassembly" which combines information from several input assemblies into a single output assembly. The final output will be superior to any of its constituents, and allows us to merge together the locally best algorithms and parameters for the genome.

Genome assembly

Genome assembly is the process of determining an organism's DNA sequence from a library of sequenced reads. Frequently, next-generation sequence methods are employed to create libraries of millions of very short reads, about 100 base pairs in length.

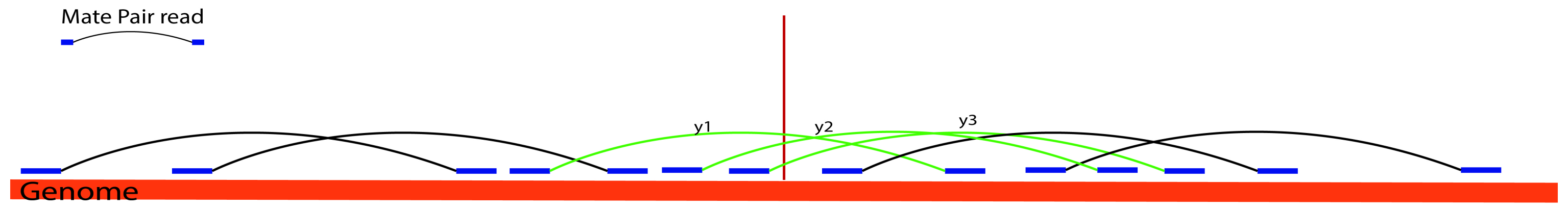
Schematic representation of a next-generation assembly project:



The accuracy and contiguity of the resulting assembly depend on several factors such as: sequencing errors, the presence of repetitive sequences within the genome, and polymorphism. Typical assembly errors are: insertion/deletion events and chromosome breaking.

Mathematical justification

Insertion/deletion errors in the assembly sequence can be pinpointed by realigning a mate pair library of a particular insert length to the assembly sequence itself. Insertion/deletion errors are flagged by deviations of the resulting insert lengths from the expected insert length distribution. The CE statistic or compression-expansion statistic compares the mean insert length of mate pairs that span a particular base pair in the assembly sequence against the expected null hypothesis mean insert length. We infer the null hypothesis mean insert length by averaging the insert lengths of the entire set of mate pairs. Large positive *CE* values (higher separation) indicate expansion errors caused by insertions, whereas negative *Z* values indicate compression errors caused deletions.



$n_i = number\ of\ mate\ pairs\ spanning\ position\ 'i'$

$y_m = insert\ length\ of\ mate\ pair\ 'm'$

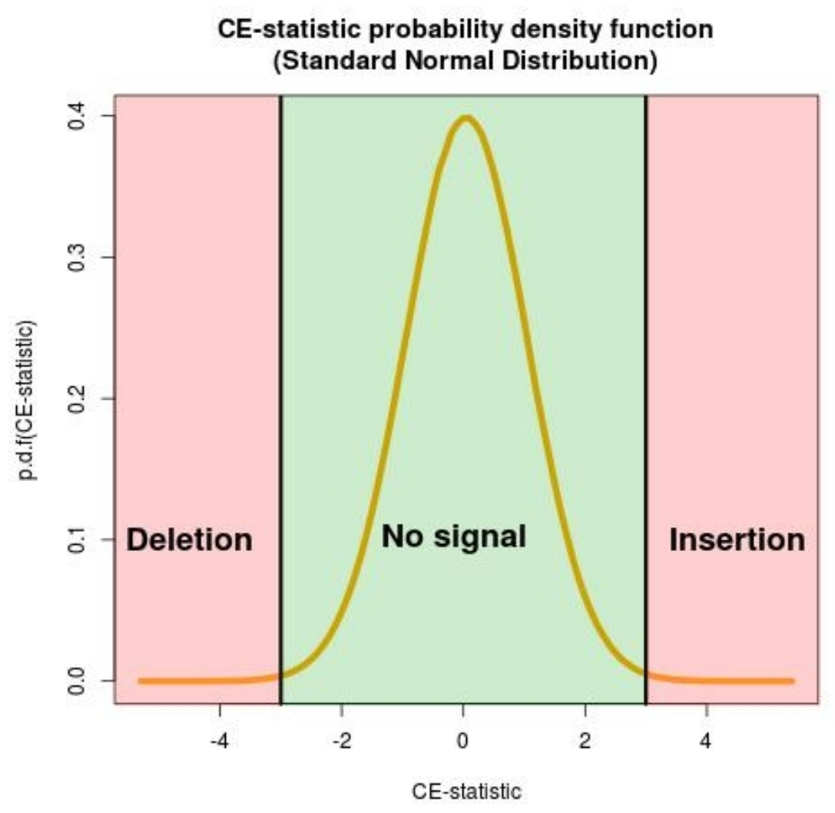
$\overline{Y}_i = mean\ insert\ length\ at\ position\ 'i'$

$\mu_0 = null\ hypothesis\ mean\ insert\ length$

$\sigma_0 = null\ hypothesis\ insert\ length\ standard\ deviation$

$$\overline{Y}_i = 1/n_i \sum_{m=1}^{n_i} y_m$$

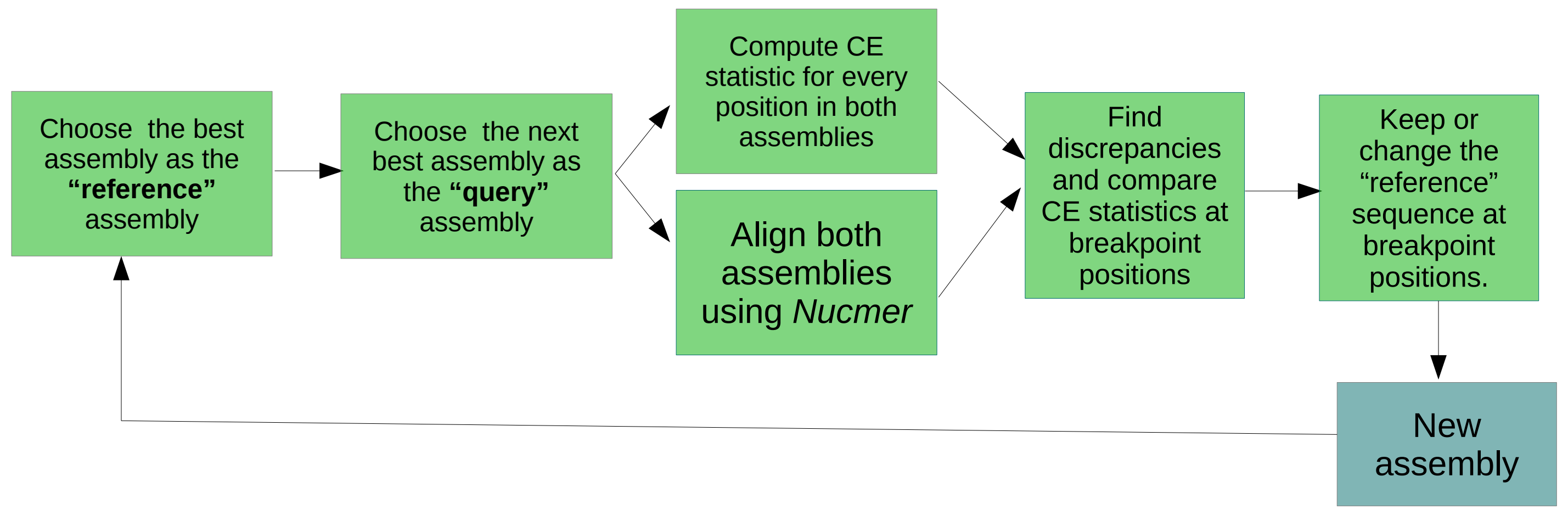
$$CE_i = \frac{\overline{Y}_i - \mu_0}{\frac{\sigma_0}{\sqrt{n_i}}}$$



Metassembly

In order to merge multiple assemblies into one superior assembly the Metassembler merges assemblies in a pairwise, cumulative fashion.

Pairwise, cumulative metassemby:


Choose the best assembly as the **"reference"** assembly

Choose the next best assembly as the **"query"** assembly

Compute CE statistic for every position in both assemblies

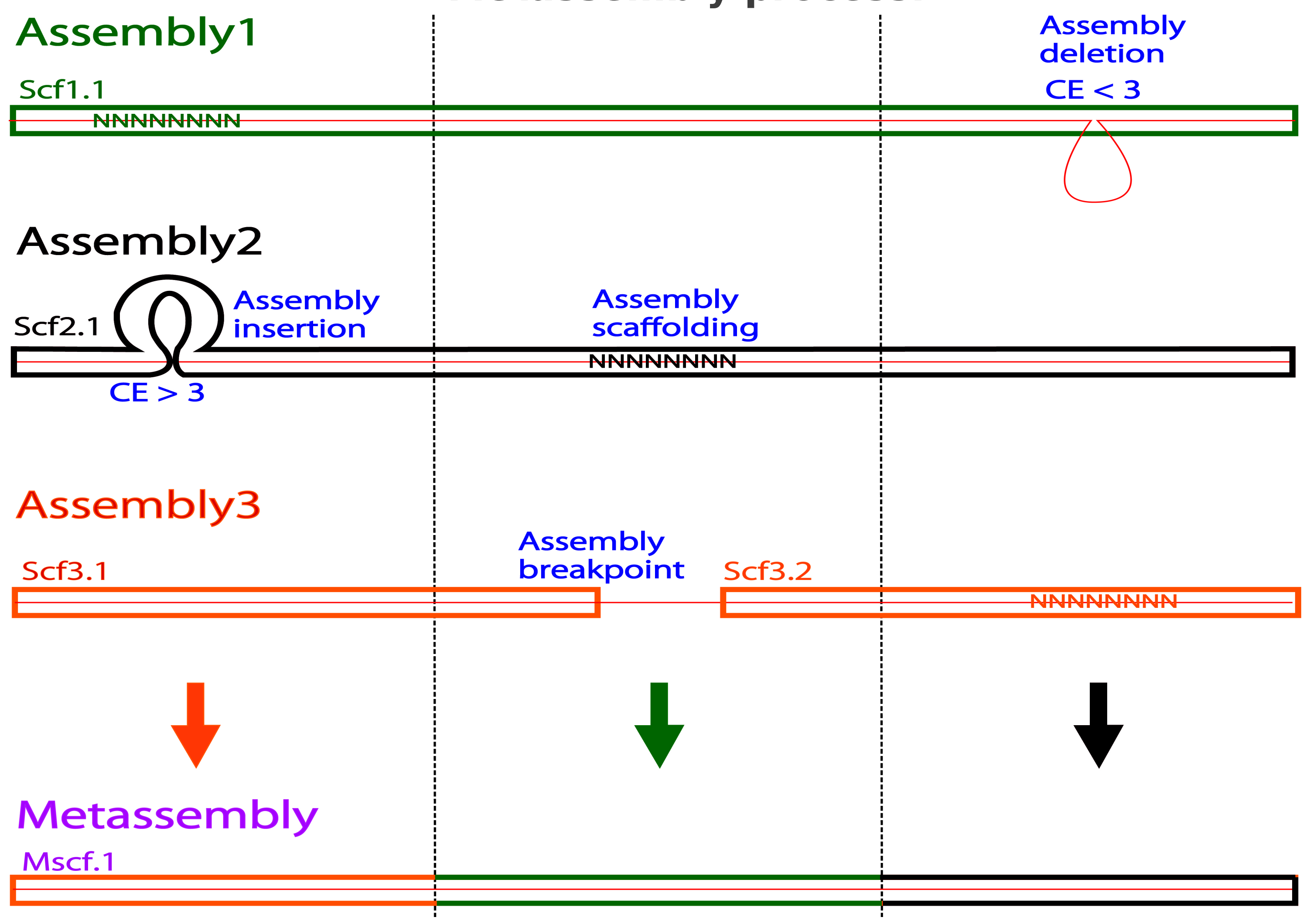
Align both assemblies using *Nucmer*

Find discrepancies and compare CE statistics at breakpoint positions

Keep or change the "reference" sequence at breakpoint positions.

New assembly

Schematic representation of the Metassembly process:



References

- Dent Earl, et al. "Assemblathon 1: A competitive assessment of de novo short read assembly methods". Genome Research 2011
- Keith R. Bradnam, et al. "Assemblathon2: evaluating de novo methods of genome assembly in three vertebrate species". 2013 GigaScience. Under review. arXiv:1301.5406
- Steven L., et al. "GAGE: A critical evaluation of genome assemblies and assembly algorithms"
- Kurtz, S., et al. "Versatile and open software for comparing large genomes." Genome Biology 2004.
- Zimin A., et al. "Assembly reconciliation." Bioinforamtics 2008.

Results: Assemblathon1. Merge top 5 assemblies

GAGE assembly evaluation tool and Metassembly report:

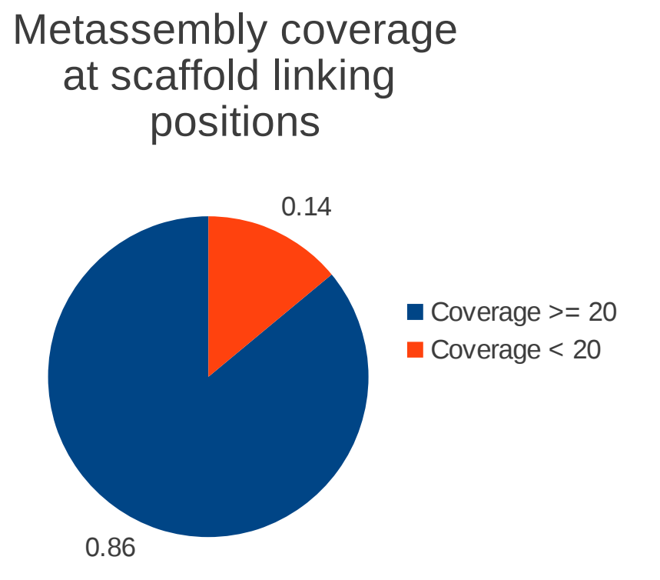
Data/Assembly	1	1-2	1-2-3	1-2-3-4	1-2-3-4-5
Scf mean	**0.122 Mb**	**0.472 Mb**	**1.43 Mb**	**1.913 Mb**	**2.45 Mb**
Scf N50	**8.28 Mb**	**10.62 Mb**	**10.62 Mb**	**11.58 Mb**	**11.58 Mb**
Scf Max	**17.10 Mb**	**30.29 Mb**	**30.28 Mb**	**38.65 Mb**	**38.66 Mb**
Ctg N50	**207 Kb**	**447 Kb**	**219 Kb**	**174 Kb**	**181 Kb**
Links	NA	40	0	2	1
Gap closures	NA	504	78	207	260
Insertion/deletion changes	NA	180	519	326	118
Missing reference bases	114 Kb (0.10%)	151 Kb (0.13%)	245 Kb (0.22%)	231 Kb (0.21%)	190 Kb (0.17%)
Dup Ref bases	1.233 Mb	1.139 Mb	0.404 Mb	0.281 Mb	0.255 Mb
Compressed Ref Bases	514 Kb	468 Kb	521 Kb	555 Kb	482 Kb
SNPS	224 K	50 K	50 K	50 K	51 K
Indels <5	35 K	37 K	37 K	37 K	37 K
Indels >=5	7,655	7,571	7,402	7,367	7,414
Inversions	95	84	84	79	96
Relocations	55	69	59	51	55
Translocations	2	4	4	4	2

Results: Assemblathon2. Merge top 3 assemblies

Metassembly report for Fish dataset:

Data/Assembly	1	1-2	1-2-3
Scf mean	**302 Kb**	**400 Kb**	**422 Kb**
Scf N50	**3.71 Mb**	**3.94 Mb**	**3.96 Mb**
Scf Max	**25.47 Mb**	**25.55 Mb**	**25.55 MB**
Ctg N50	**24581**	**29122**	**29963**
Links	NA	152	9
Gap closures	NA	7793	1288
Insertion/deletion changes	NA	2076	13457

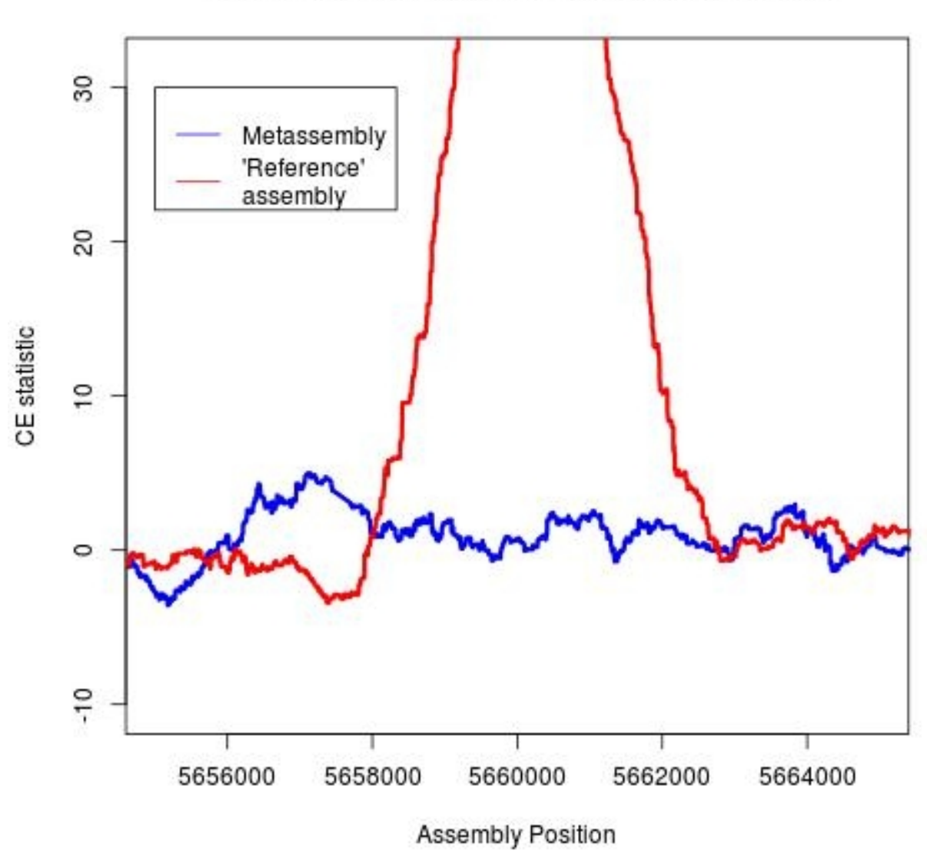
Metassembly coverage at scaffold linking positions



Metassembly report for Bird dataset:

Data/Assembly	1	1-2	1-2-3
Scf mean	**264 Kb**	**422 Kb**	**683 Kb**
Scf N50	**14.94 Mb**	**17.34 Mb**	**17.70 Mb**
Scf Max	**65.89 Mb**	**74.17 Mb**	**74.18 Mb**
Ctg N50	**56837**	**93395**	**107201**
Links	NA	166	39
Gap closures	NA	16,148	5,088
Insertion/deletion changes	NA	922	1,649

CE statistic comparison: BEFORE and AFTER



Future Work

Metassembler is broadly applicable to any sequencing project where multiple draft assemblies are created. In particular, we hope to apply Metassembler to :

- Rice and other plant genomes
- Snake genome (Assemblathon2 dataset)