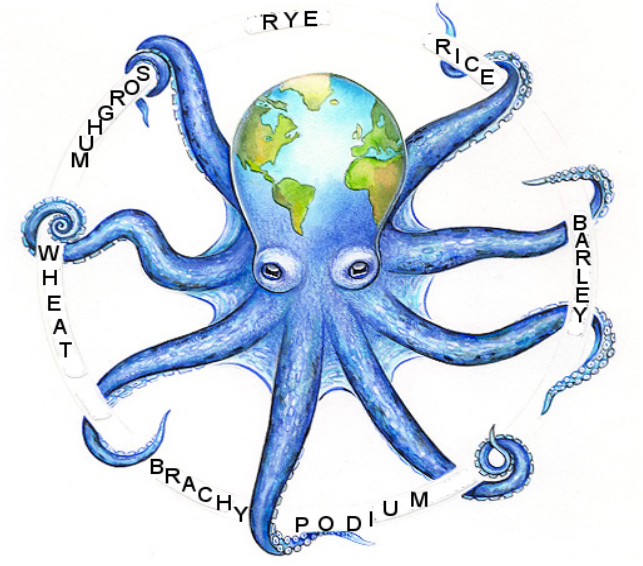
Application of the GenomeZipper-Approach on Wheat



Mihaela Martis
MIPS, Munich

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Overview

- From barley to wheat
- Introduction to the GenomeZipper Pipeline
  - Repeat masking
  - Homology search
  - Synteny detection
  - GenomeZipper
- Live Demo of the GenomeZipper on the short or long arm of 1A or 1B
- Discussion of the results

From Barley to Wheat

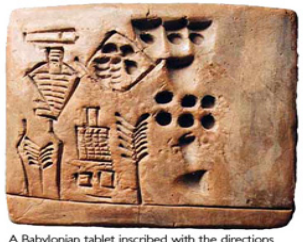
10500 BC

9000 BC

early Neolithic: first dating of beer

7800 BC

4500 BC



A Babylonian tablet inscribed with the directions for brewing beer (c. 3100BCE)

Ancient Babylonians drinking beer through straws

present

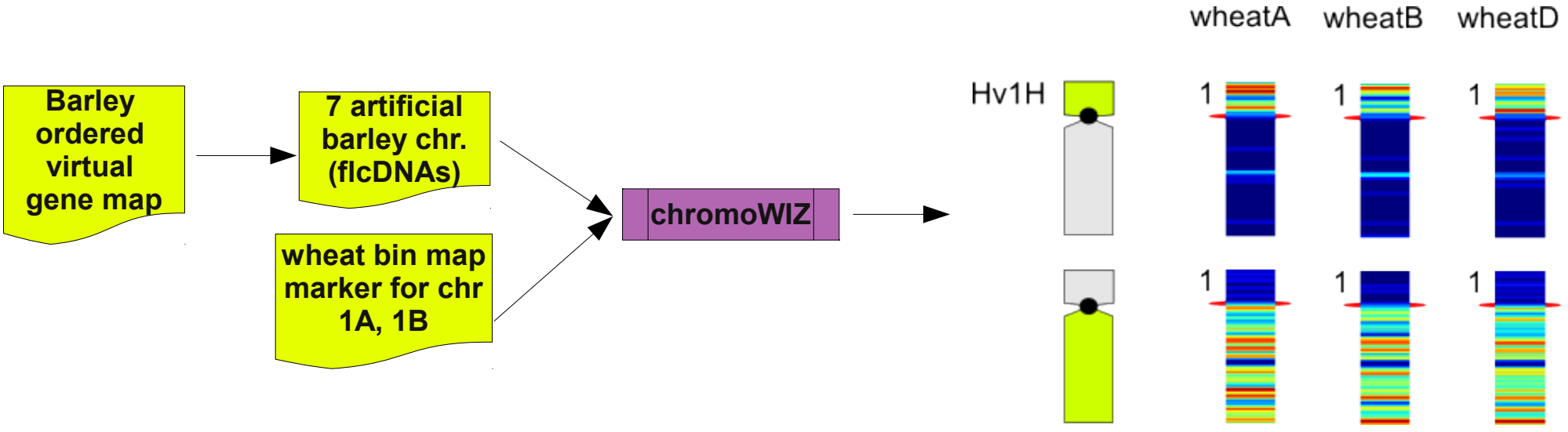
start 2008

2011

2012

. . .

From Barley 1H to Wheat 1A, 1B



Raw lab data:

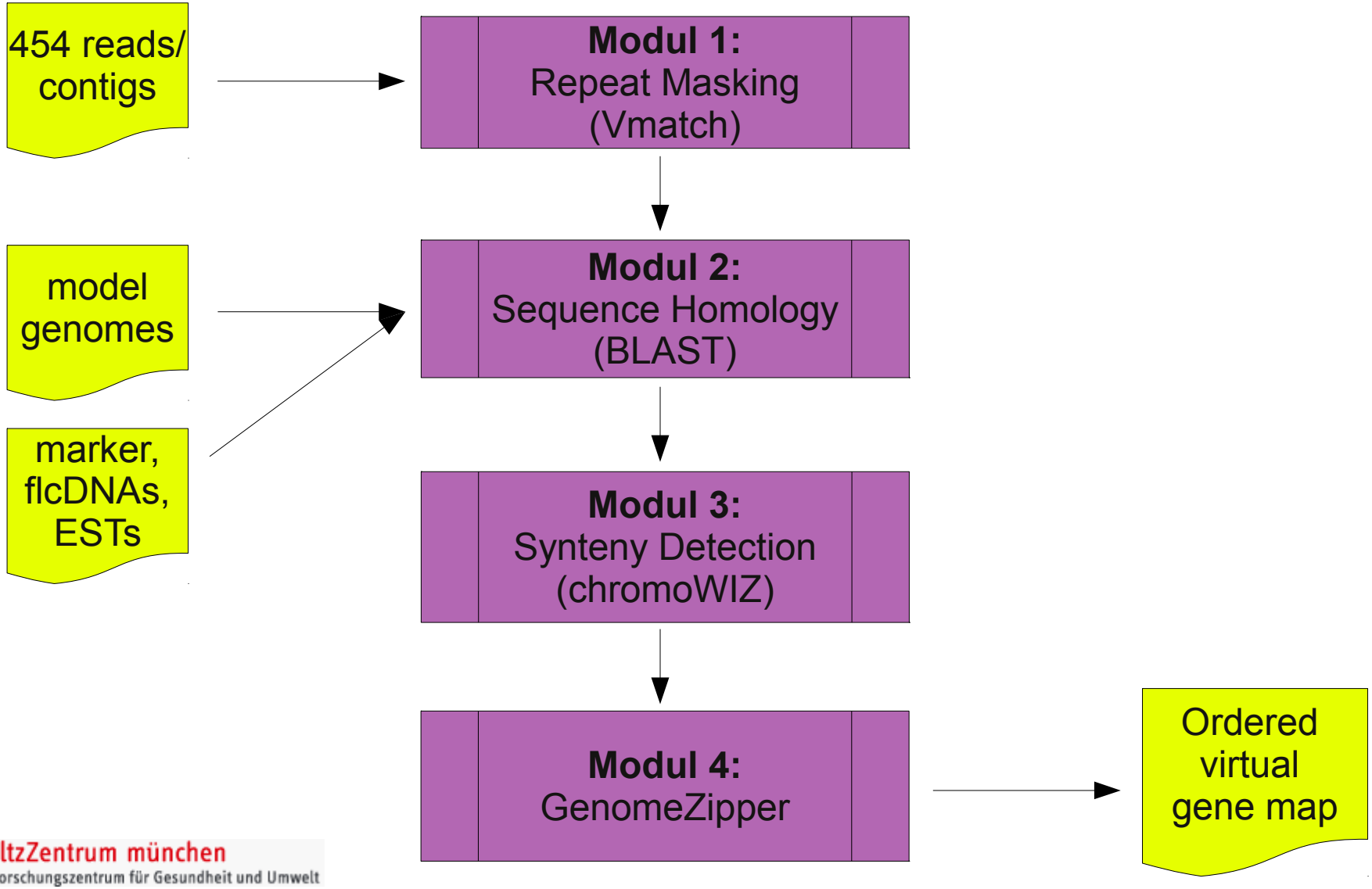
wheat	Mb	# reads	Min bp	Max bp
1AS	400.5	1062650	12	656
1AL	782.9	2058861	2	2023
1BS	728	1939403	11	2035
1BL	834.5	2177887	2	744

longestRead.fa

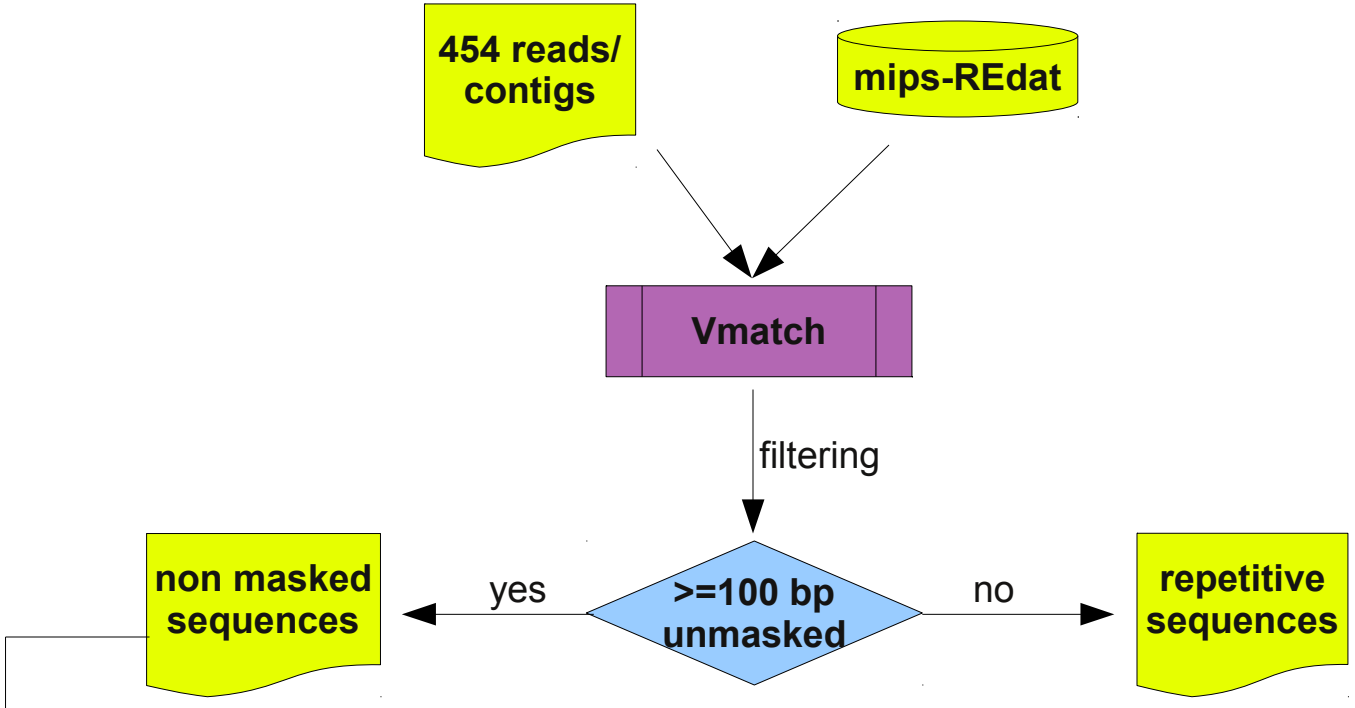
What is the GenomeZipper?

- Is an approach developed to create an ordered virtual gene map for a chromosome
- It smartly combines chromosom sorting, next generation sequencing, genetic maps, flcDNAs and systematic exploitation of conserved synteny with model grasses
- It provides a valuable surrogate for the gene space of the analyzed chromosome/genome
- Requirements:
  - Masked 454 reads/contigs
  - Orthologs from syntenic regions

GenomeZipper Pipeline



Modul 1: Repeat Masking

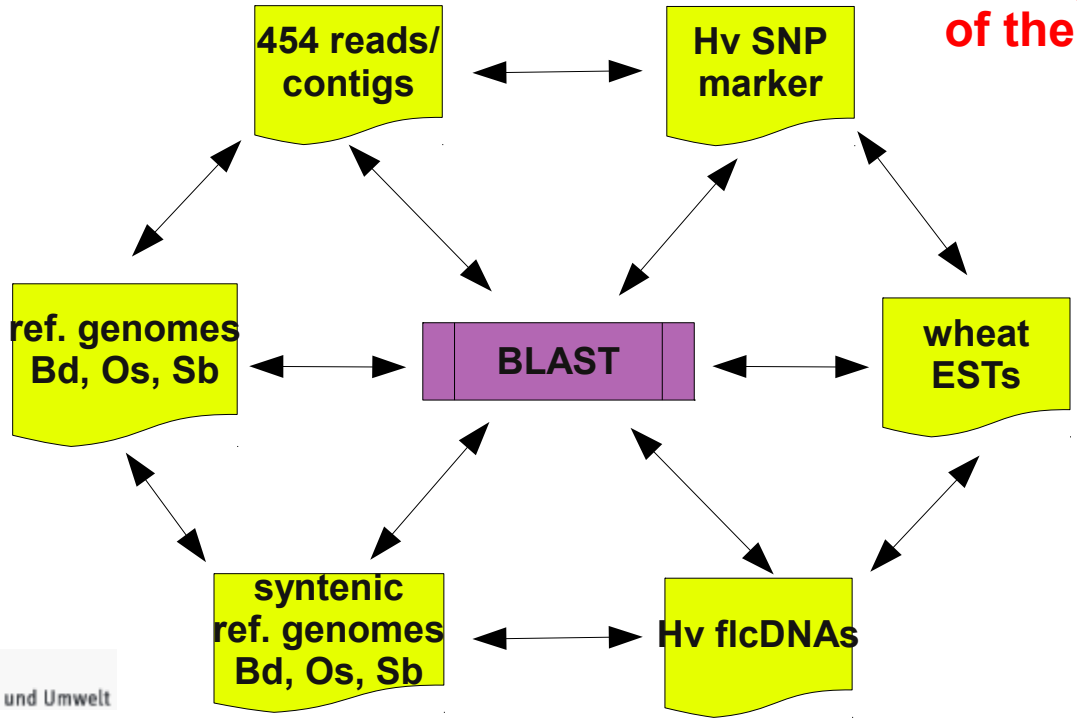


wheat	N's %	# non masked reads	Min bp	Max bp
1AS	75.8	259212	100	656
1AL	71.6	576845	100	2023
1BS	69.9	582714	100	2035
1BL	71.2	624535	100	744

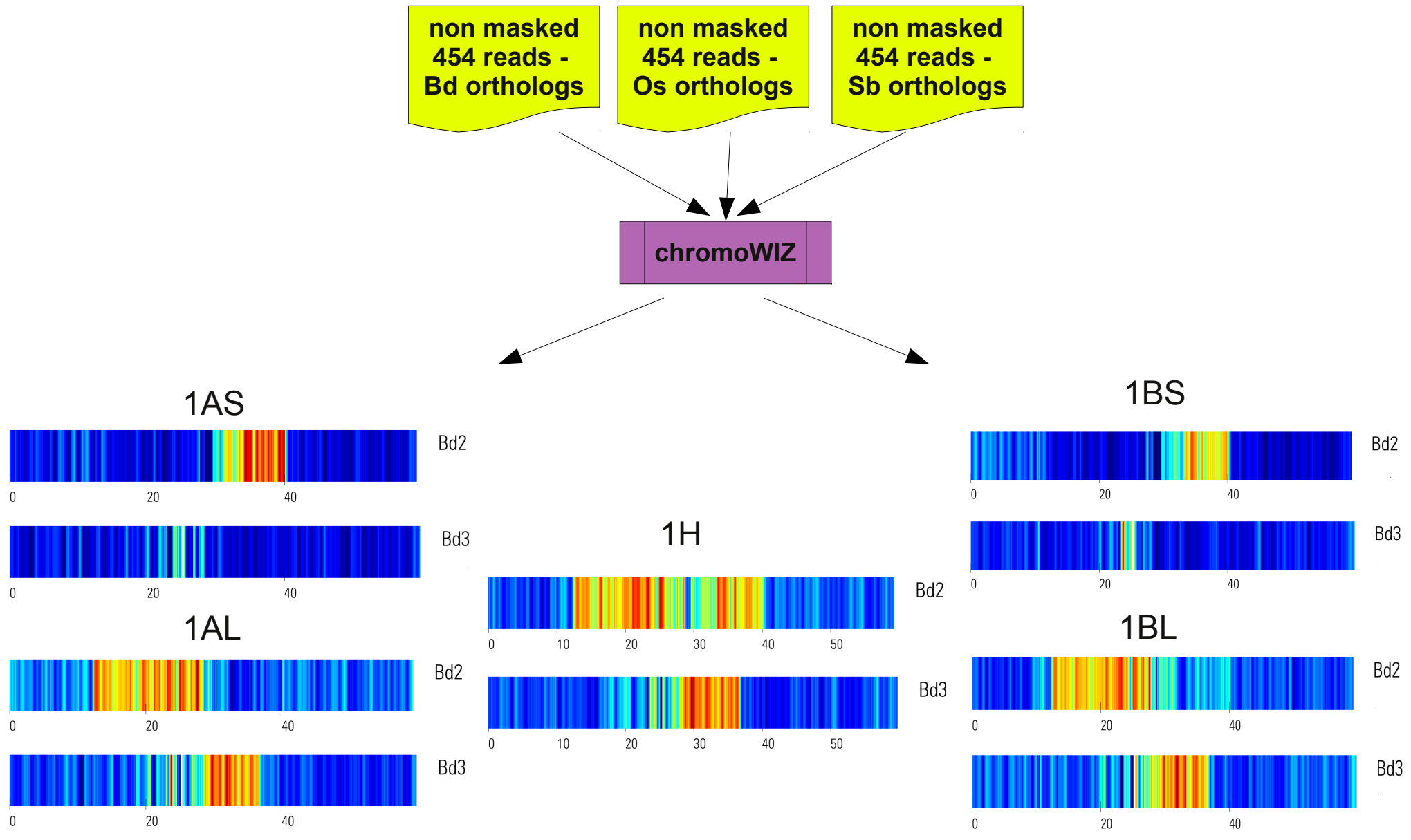
Modul 2: Sequence Homology

- normal & bidirectional: blastp, blastx, blastn
- 54 Blasts per chromosome arm
- 80 Blasts per chromosome
- 1680 per A, B & D genome

The runtime of the blasts is the biggest limitation factor of the pipeline!!!



Modul 3: Synteny Detection



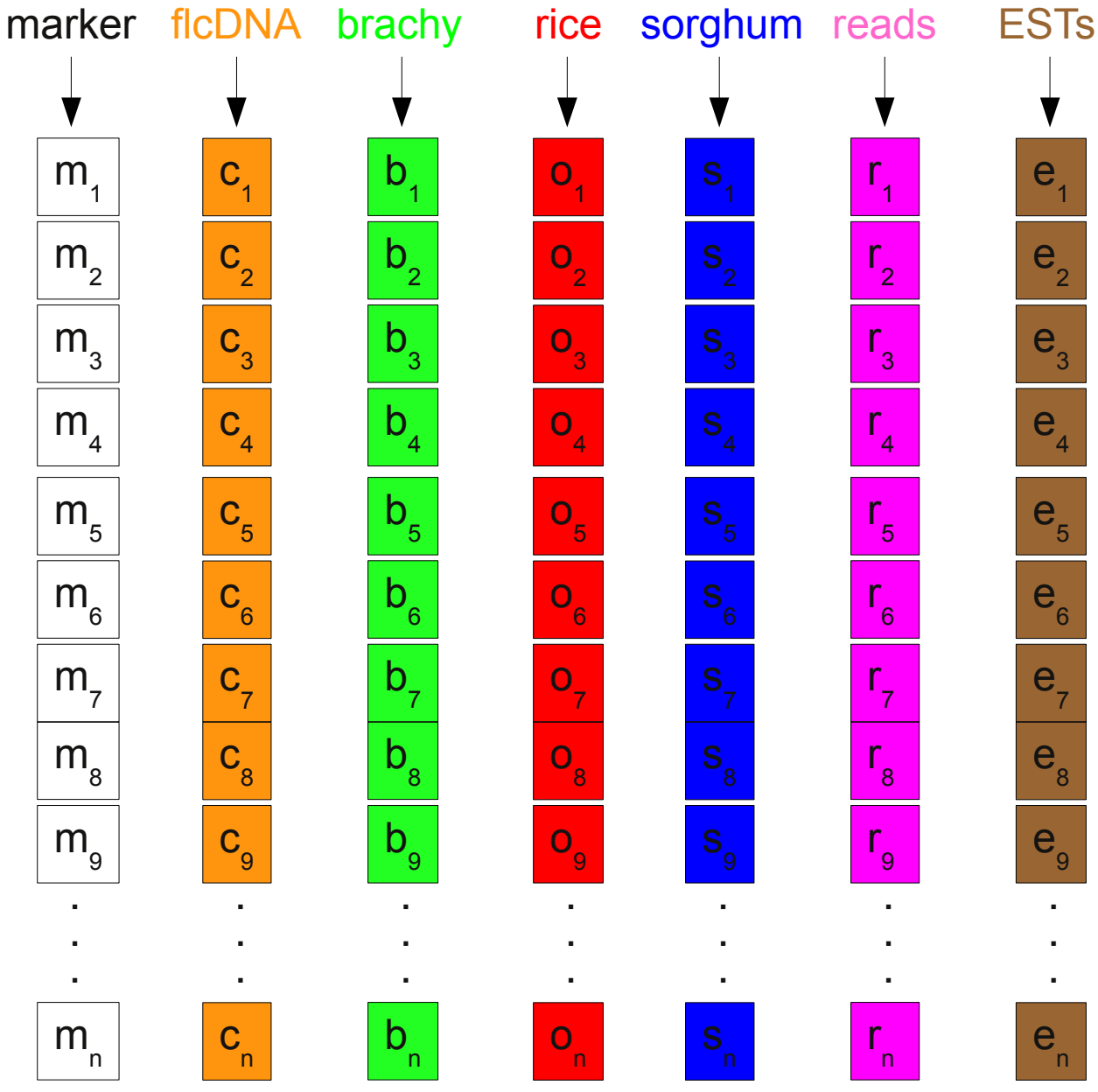
Modul 4: GenomeZipper

- Configuration File
- Program fetch:

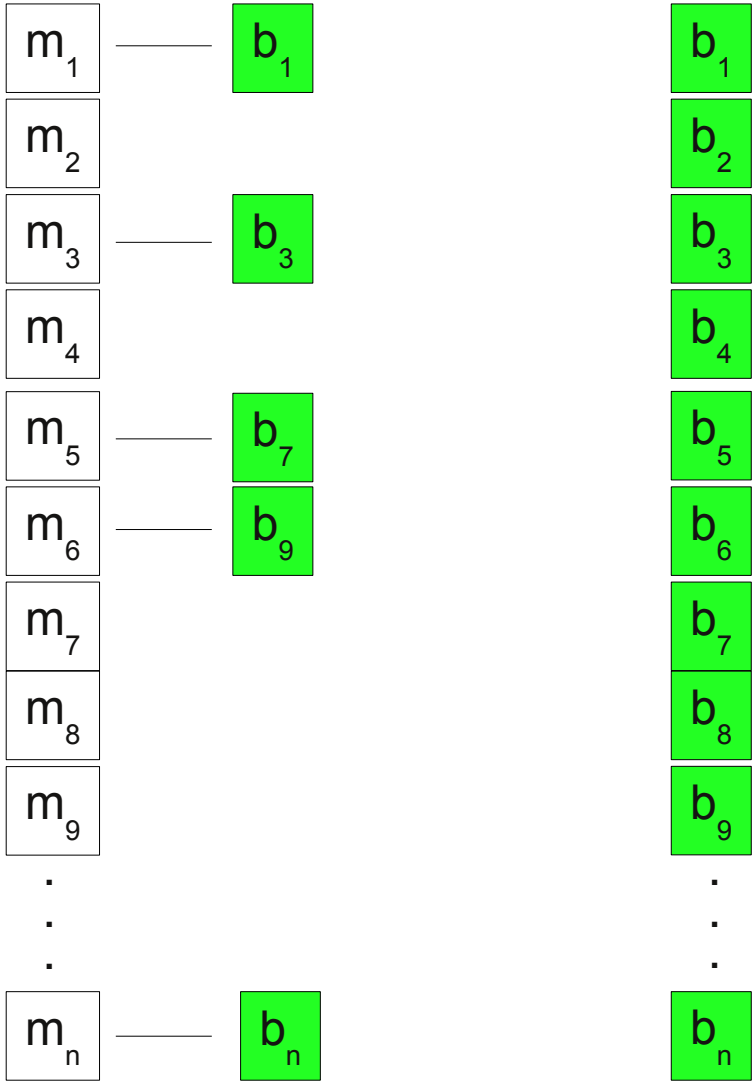
  perl genomeZipper.pl configFile

- Runtime: ~ 2 minutes (short arm) and ~ 9 minutes (long arm)

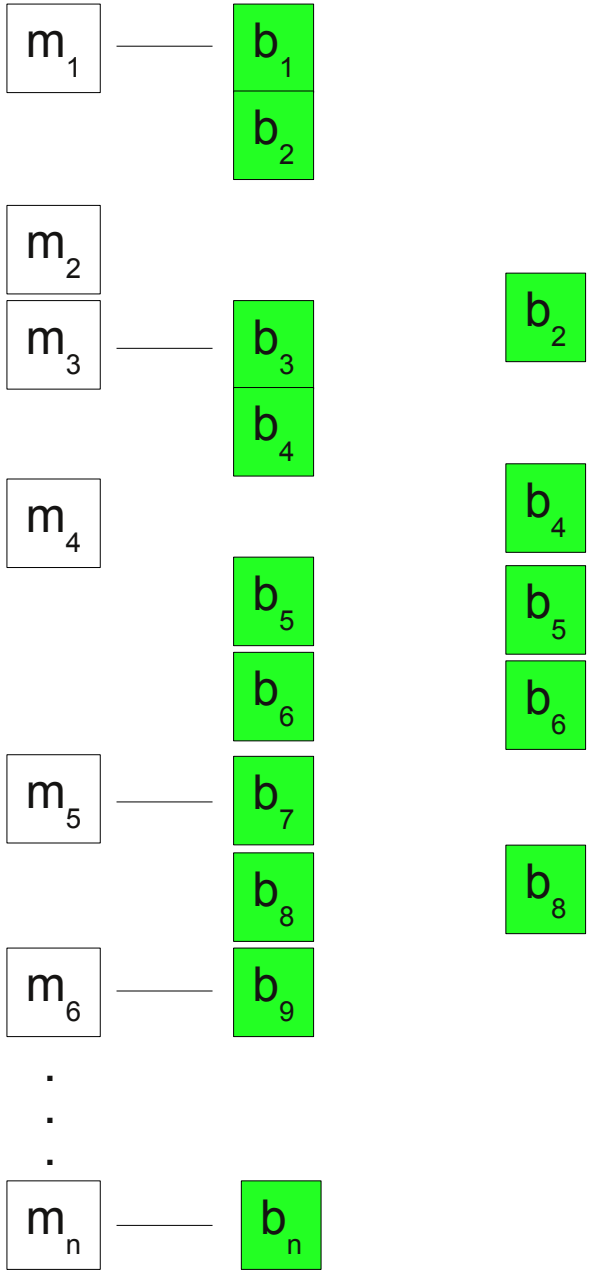
Input data:


marker
flcDNA
brachy
rice
sorghum
reads
ESTs
m1 m2 m3 m4 m5 m6 m7 m8 m9 . . . mn
c1 c2 c3 c4 c5 c6 c7 c8 c9 . . . cn
b1 b2 b3 b4 b5 b6 b7 b8 b9 . . . bn
o1 o2 o3 o4 o5 o6 o7 o8 o9 . . . on
s1 s2 s3 s4 s5 s6 s7 s8 s9 . . . sn
r1 r2 r3 r4 r5 r6 r7 r8 r9 . . . rn
e1 e2 e3 e4 e5 e6 e7 e8 e9 . . . en

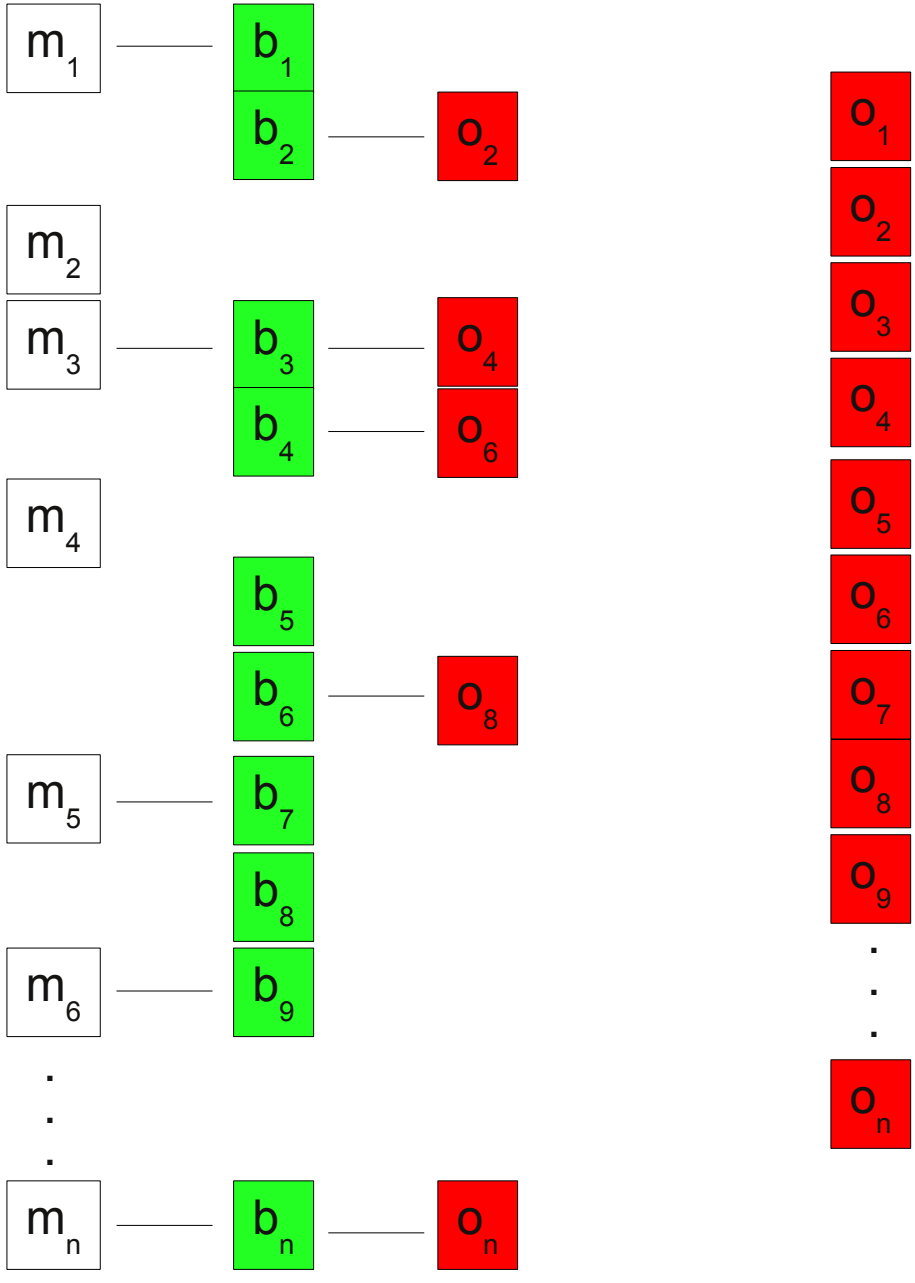
Step 1: Combine Marker With Brachypodium Genes



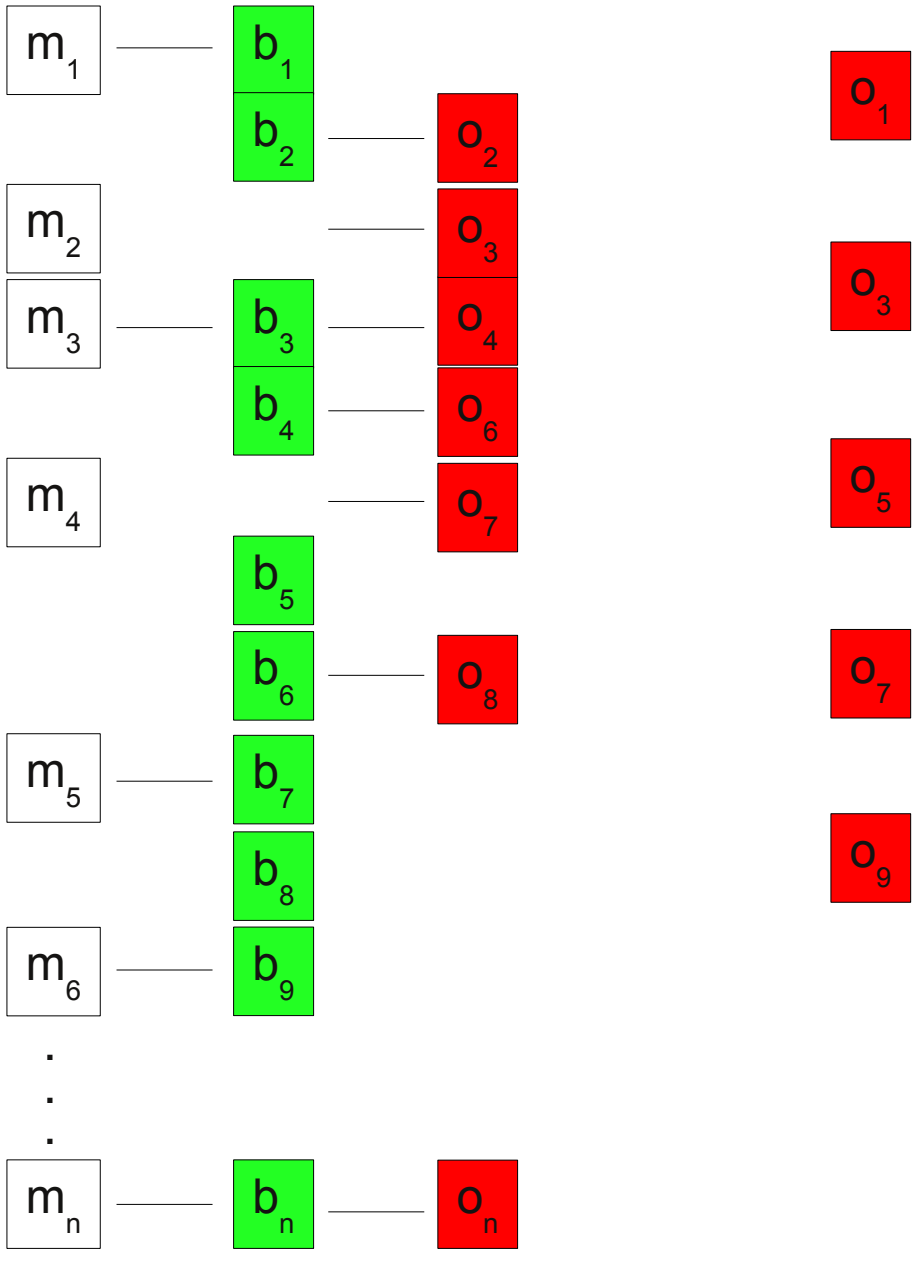
Step 2: Insert Remaining Brachypodium Genes



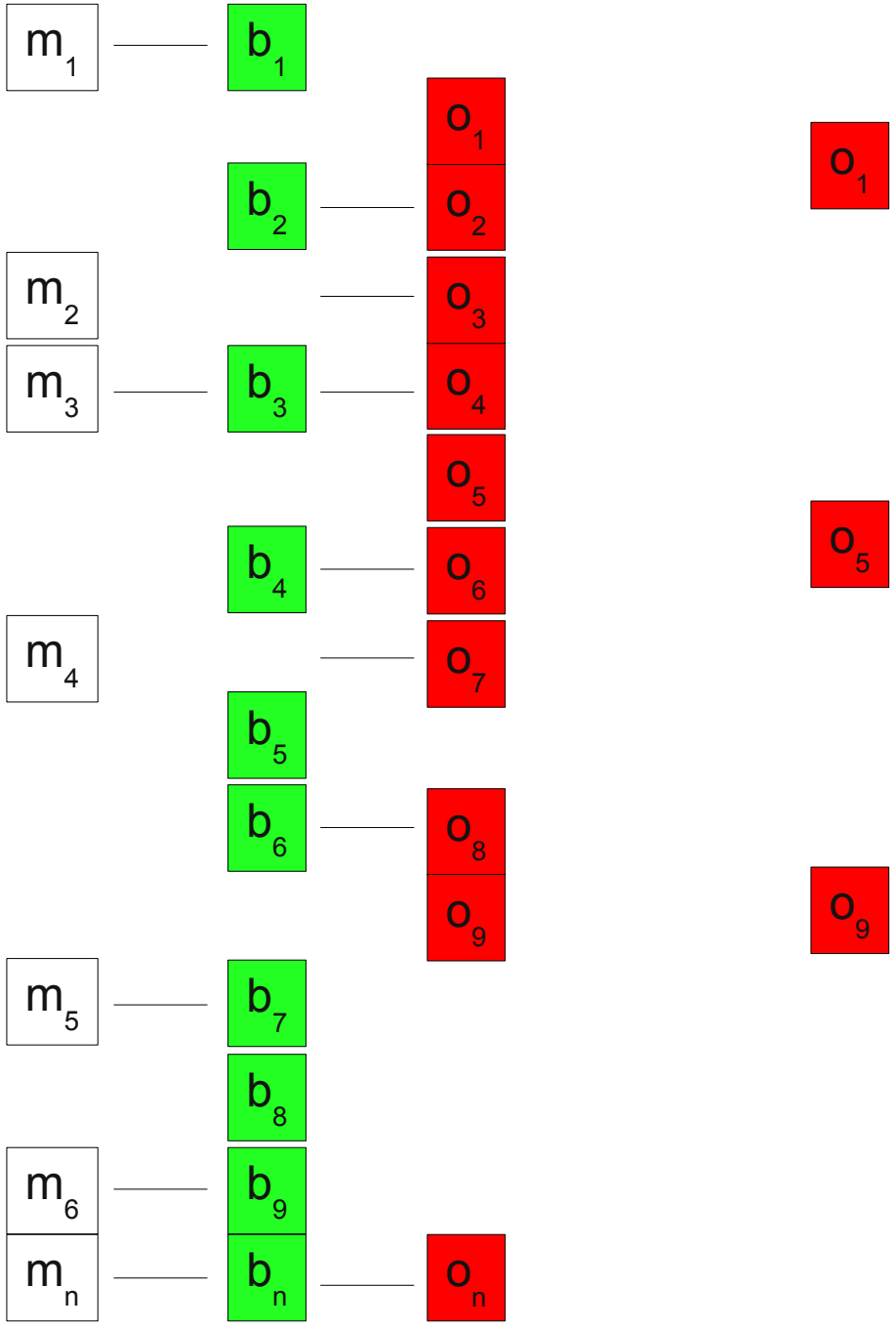
Step 3: Insert Brachypodium – Rice BBHs



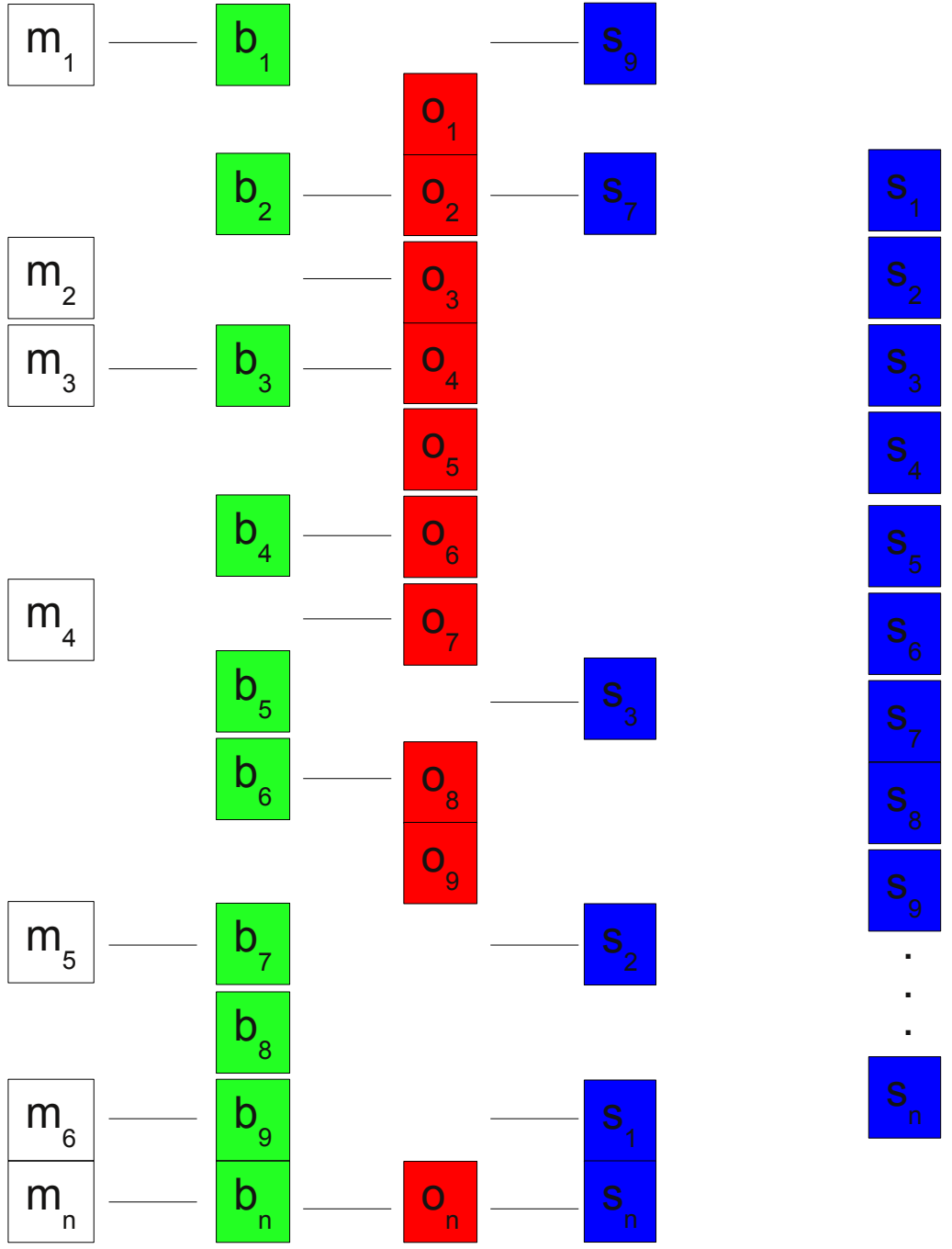
Step 4: Combine Remaining Rice Genes With Marker



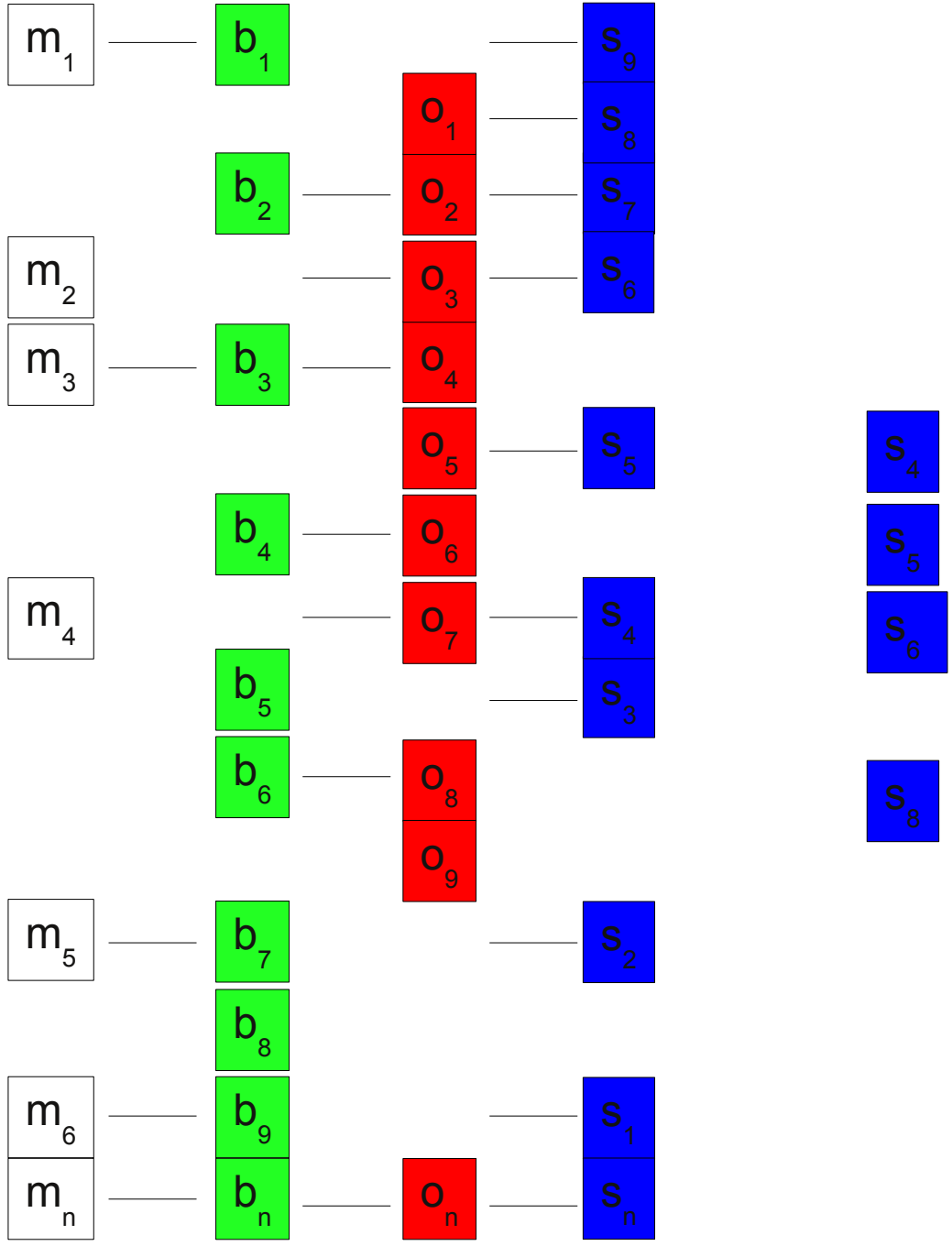
Step 5: Insert Remaining Rice Genes



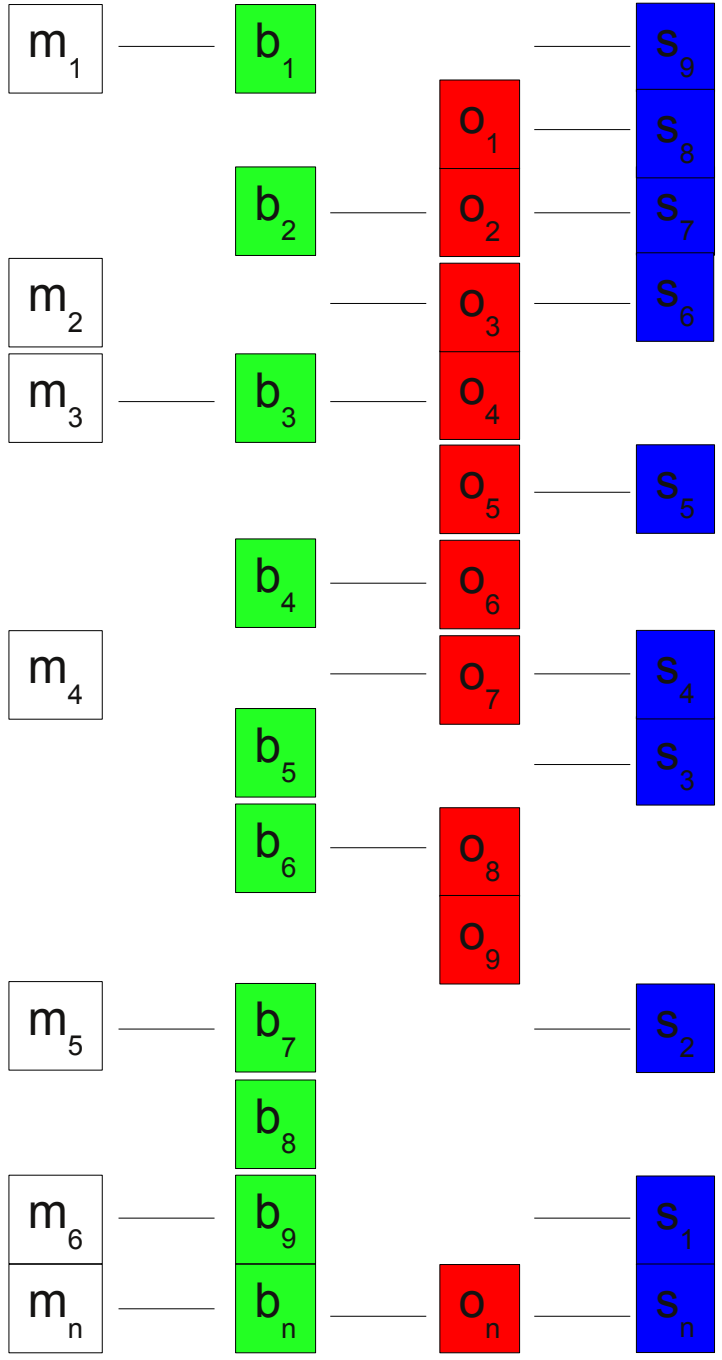
Step 6: Insert Brachypodium – Sorghum BBHs



Step 7: Insert Rice – Sorghum BBHs



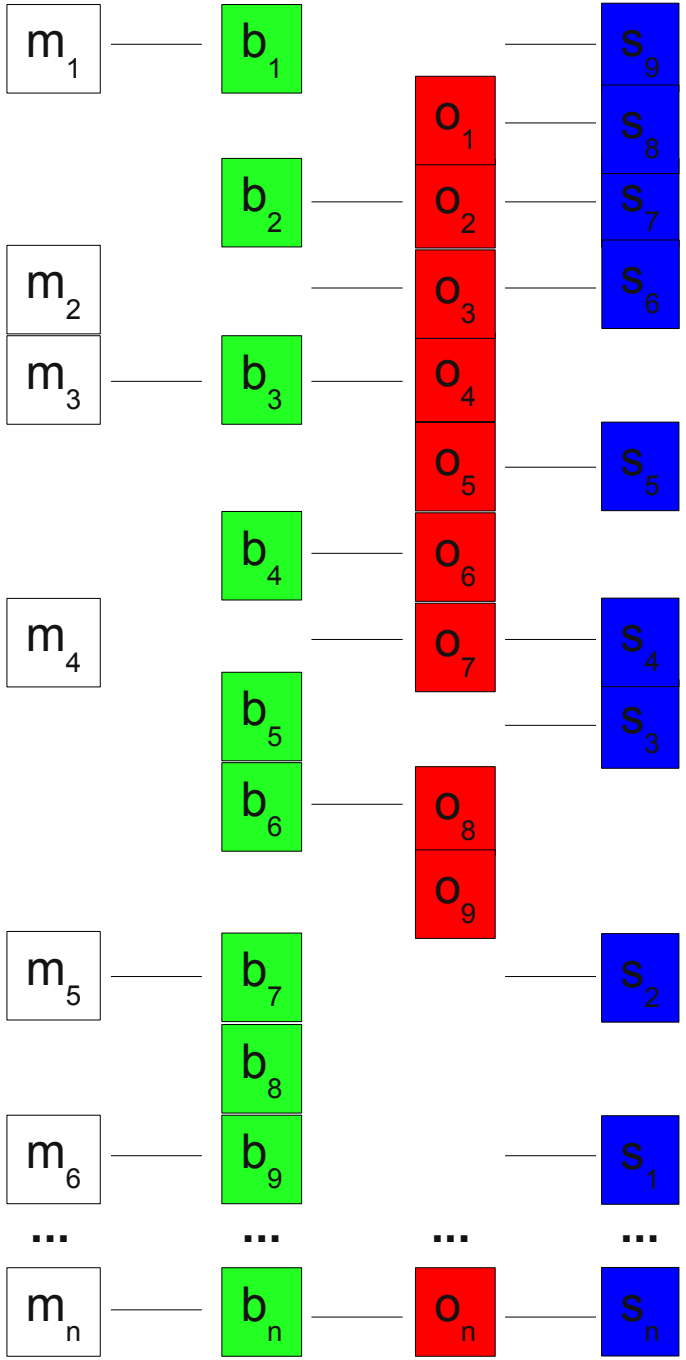
Step 8: Combine Remaining Sorghum Genes With Marker



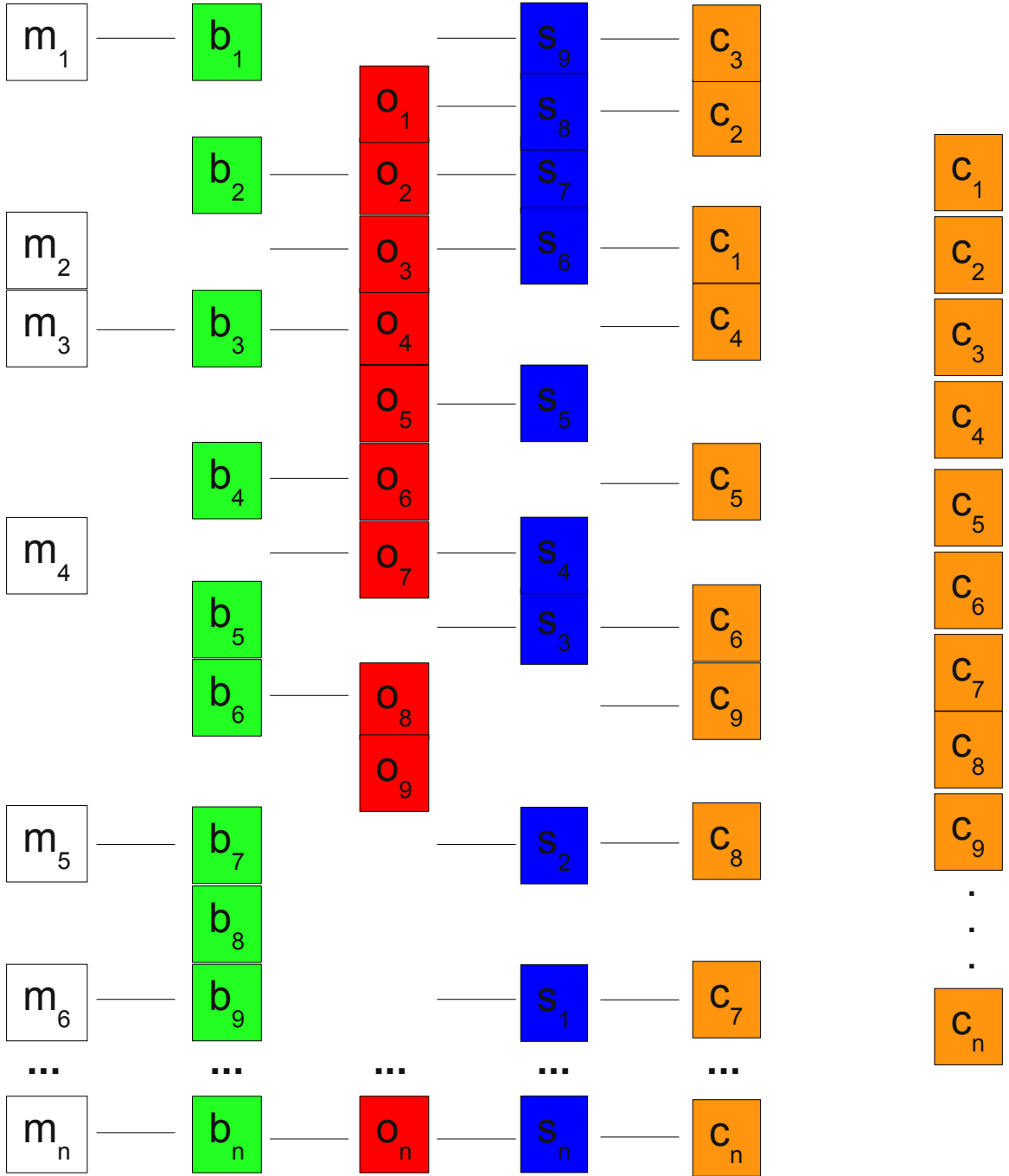
Nothing to do, no sorghum genes left over in this example!

Step 9: Insert Remaining Sorghum Genes

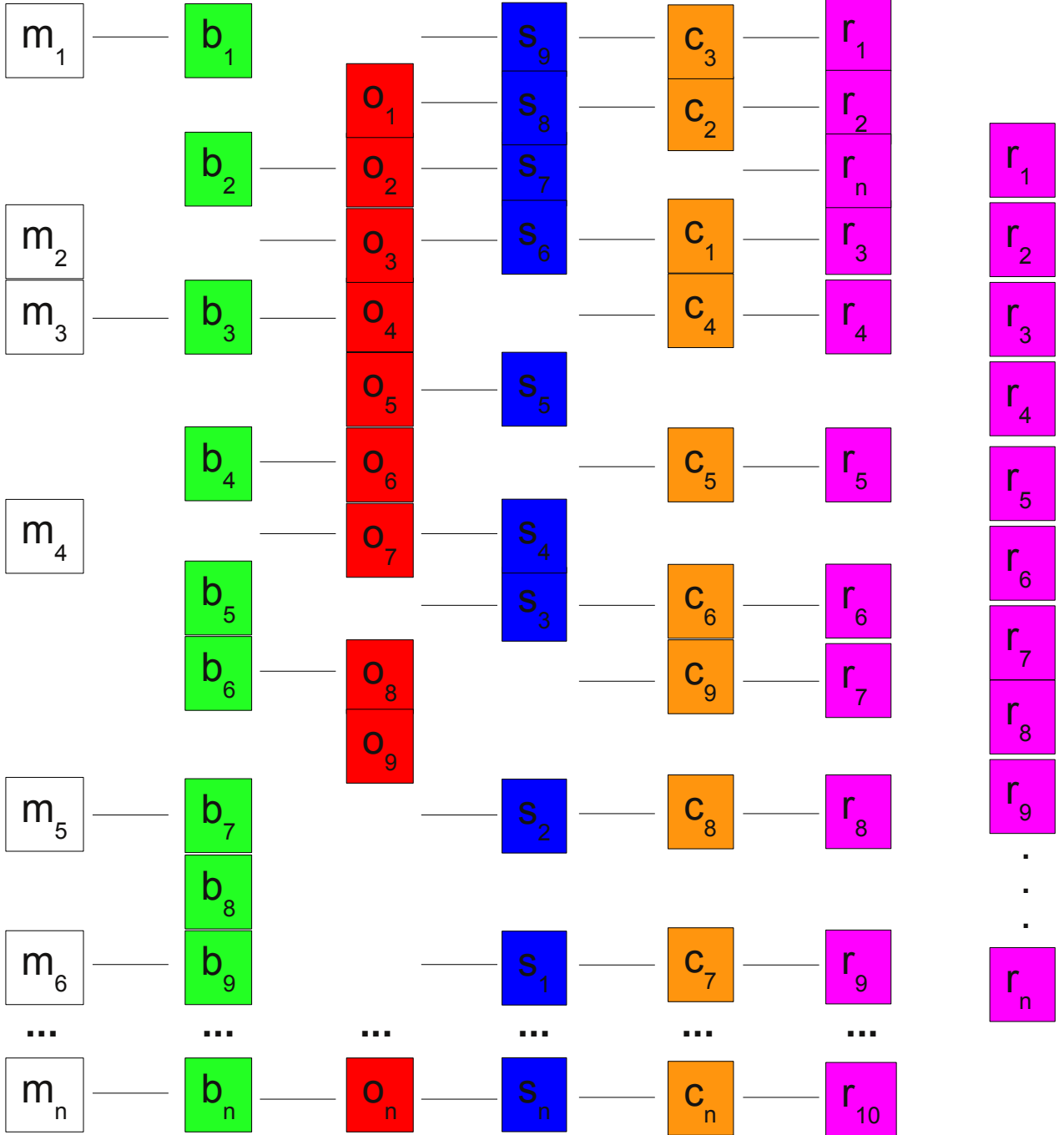
Nothing to do, no sorghum genes left over in this example!



Step 10: Combine FlcDNAs with marker/Bd/Os/Sb

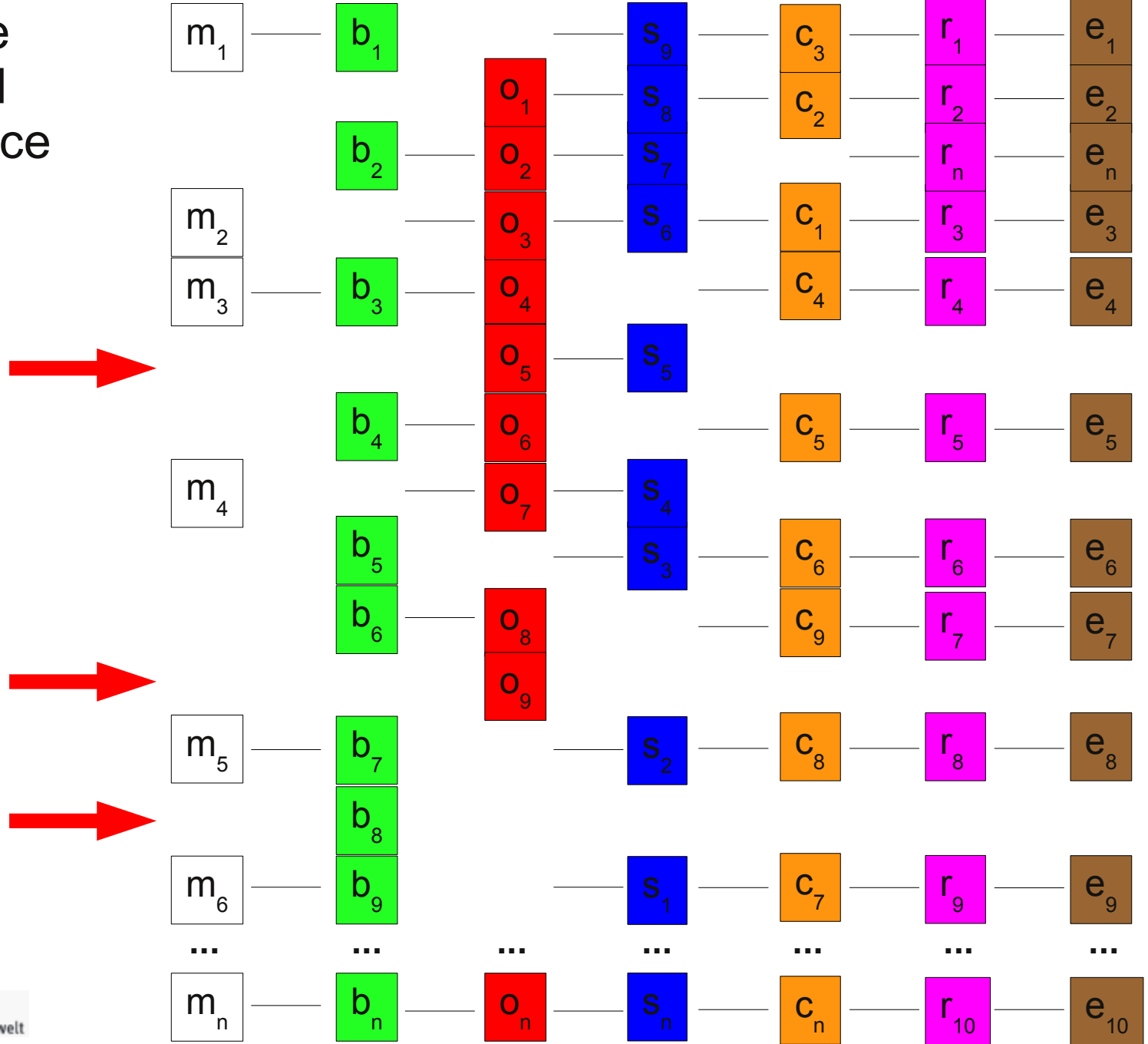


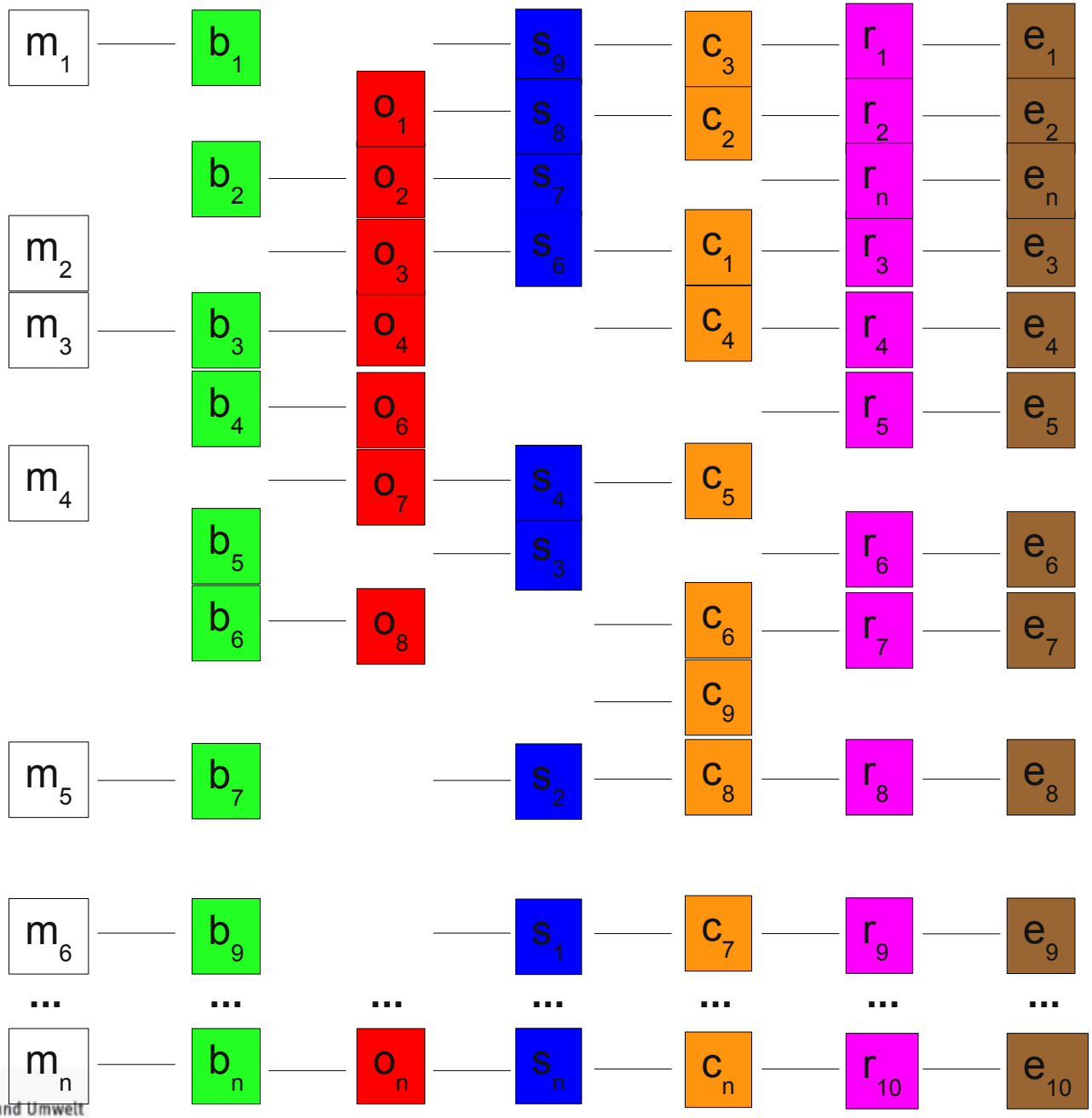
Step 11: Combine reads with marker/Bd/Os/Sb



Step 12: Combine ESTs with reads

Step 13: Remove data without read or marker evidence

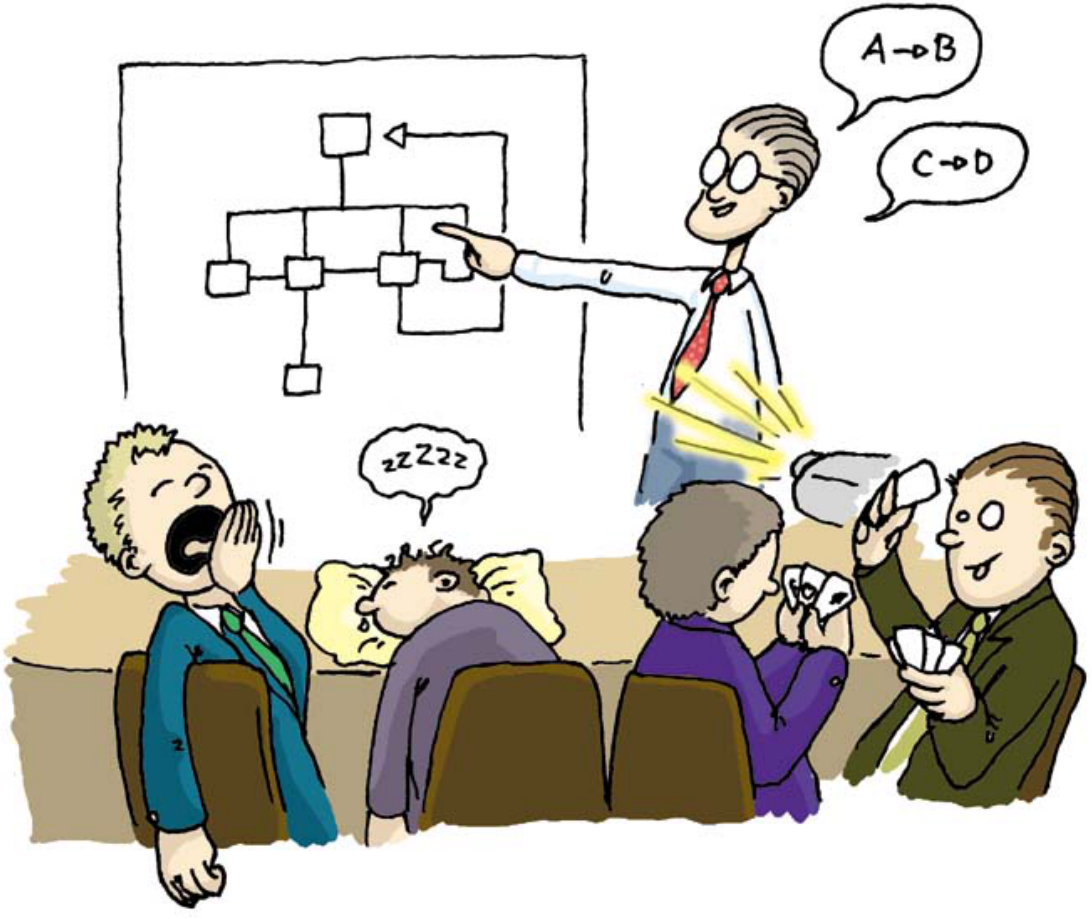




GenomeZipper Output

- Output consists of two files:
  - Tab-delimited file with the ordered gene map
  - gz_1AS.txt gz_1AL.txt gz_1BS.txt gz_1BL.txt
  - gz_1AS.xls gz_1AL.xls gz_1BS.xls gz_1BL.xls
  - A file with statistics:
    - gz_stat_1AS.txt, gz_stat_1AL.txt
      gz_stat_1BS.txt gz_stat_1BL.txt

	# loci	# marker with Bd/Os/Sb hit	# reads	# ESTs	# hv FlcDNAs	# Bd	# Os	# Sb
1A	2763	177	27596	1770	1420	1798	1386	1492
1B	2884	165	35151	1820	1450	1799	1485	1564



Thank you for your attention!