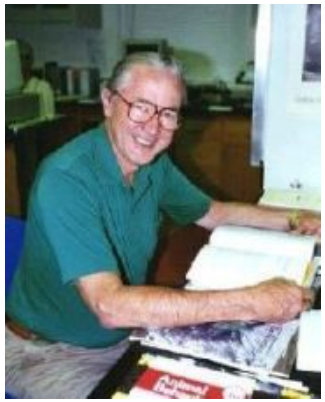
Yellow-bellied marmot genome

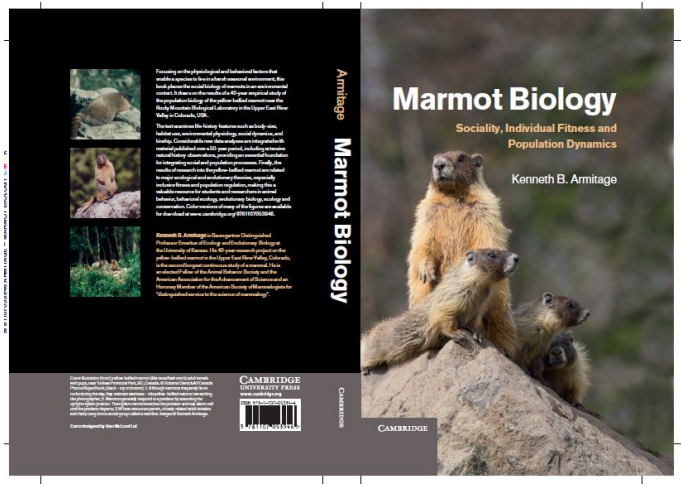
Gabriela Pinho
Graduate Student
Blumstein & Wayne Labs
EEB - UCLA

Why do we need an annotated genome?



Kenneth B. Armitage — Daniel T. Blumstein .....

1962 — 2002



Samples & measurements

Social interactions

Results include

JOURNAL OF Evolutionary Biology
eseb
European Society for Evolutionary Biology
doi: 10.1111/jeb.12700

Heritability and genetic correlations of personality traits in a wild population of yellow-bellied marmots (*Marmota flaviventris*)

M. B. PETELLE*†[1], J. G. A. MARTIN‡ & D. T. BLUMSTEIN*§

*Department of Ecology & Evolutionary Biology, University of California, Los Angeles, CA, USA
†Department of Zoology & Entomology, University of the Free State Qwaqwa, Phuthaditjhaba, South Africa
‡School of Biological Sciences, University of Aberdeen, Aberdeen, UK
§The Rocky Mountain Biological Laboratory, Crested Butte, CO, USA

PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

A test of the social cohesion hypothesis: interactive female marmots remain at home

Daniel T. Blumstein, Tina W. Wey and Karisa Tang

Proc. R. Soc. B published online 3 June 2009
doi: 10.1098/rspb.2009.0703

Tiffany Armenta
Graduate Candidate
Blumstein & Wayne Labs
EEB - UCLA

Marmot gene expression

Dispersal
- ~200 genes
- Musculature
- Metabolism
- Immune function

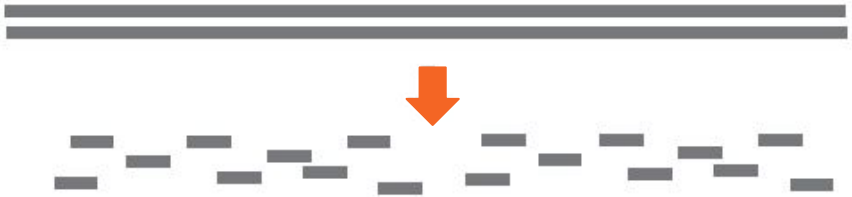
Sociality metrics
- ~600 genes

Used the thirteen-lined ground squirrel genome (8-12 MYA)
- 62% of RNA data was usable

The assembly

Sequencing

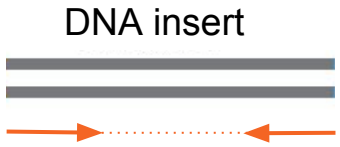
- Genomic DNA extracted from blood of a male, inbred marmot
- 4 libraries (250bp, 400bp, 700bp & 3-5kb)



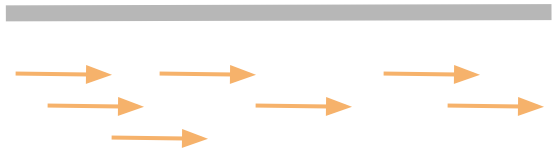
Select size

HiSeq 2500

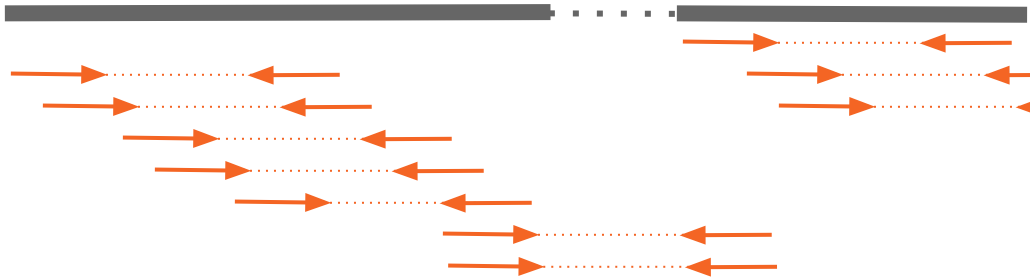
Sequencing



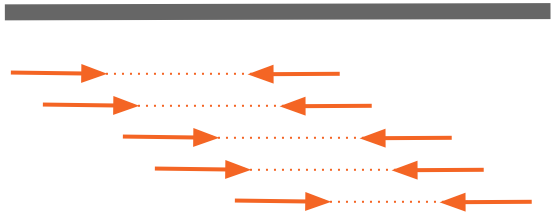
- Single-end



- Paired-end

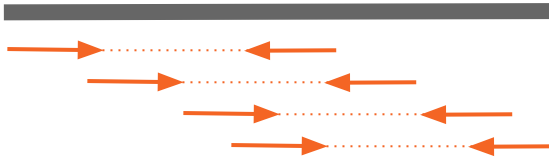


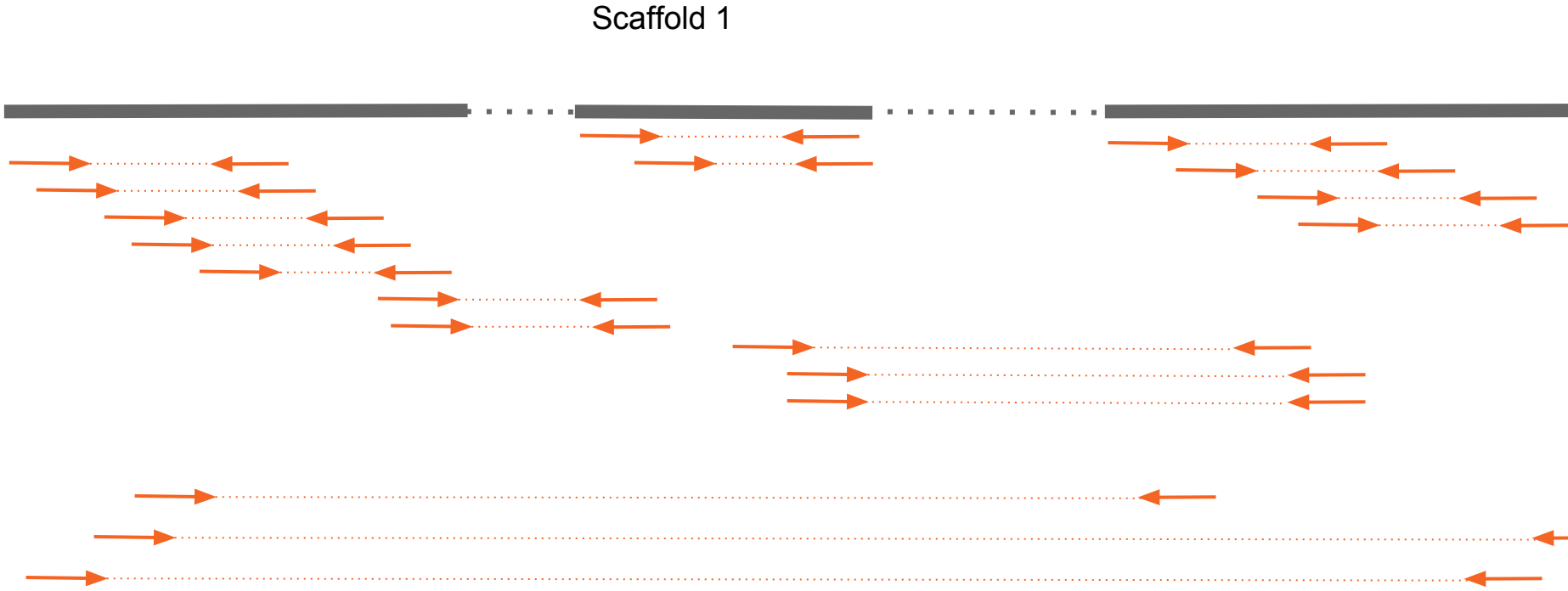
Contig 1

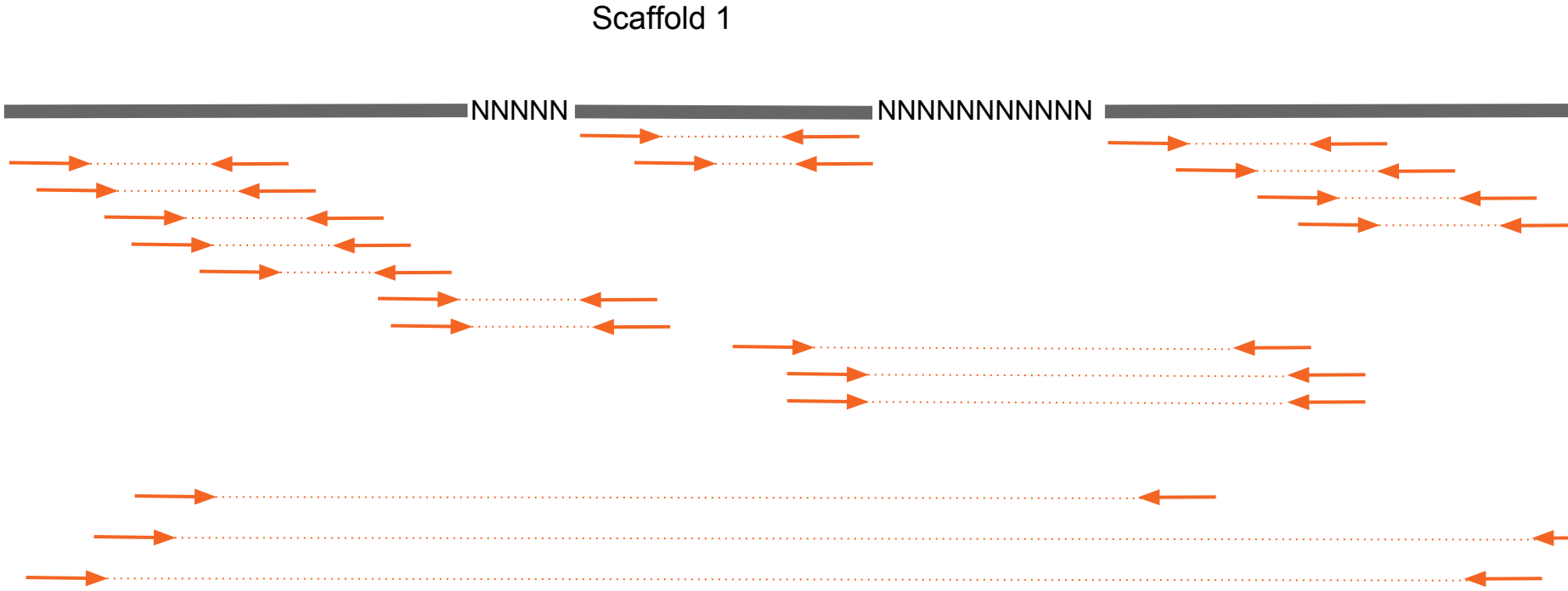


Contig 2

Contig 3




Scaffold 1


Scaffold 1
NNNNN
NNNNNNNNNNN

Reads
(single or
paired- end)

Assembler

Scaffolds
(Assembly)

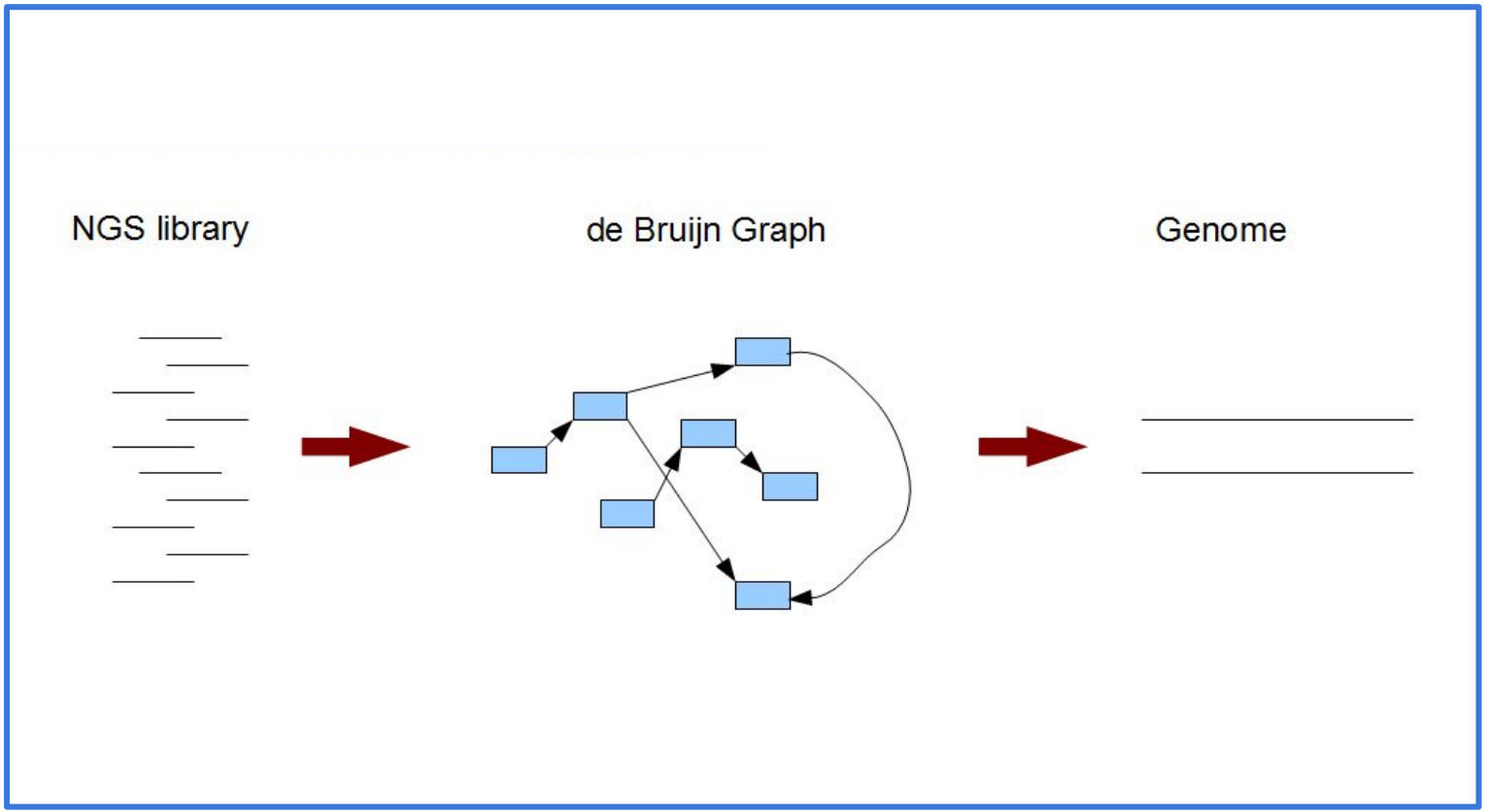
Reads (single or paired- end) → Assembler → Scaffolds (Assembly)

- **Assemblers work differently for each species**
- **ABYSS Meraculous SOAPdenovo**
- **De Bruijn graph-based assemblers**

Reads (single or paired- end)

Assembler

Scaffolds (Assembly)



Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end)

Assembler

Scaffolds (Assembly)

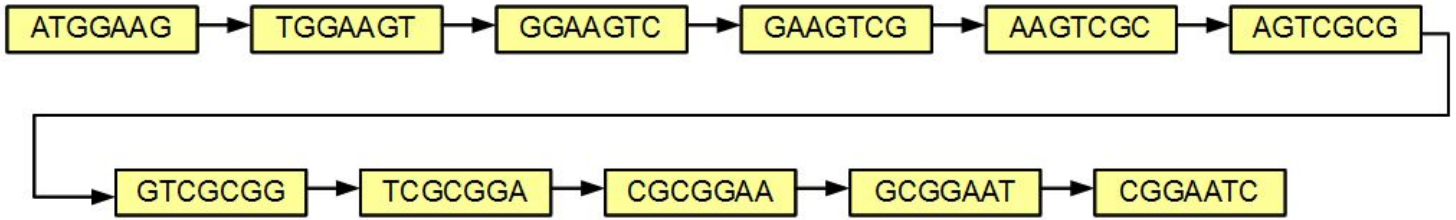
Read

ATGGAAGTCGCGGAATC

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph

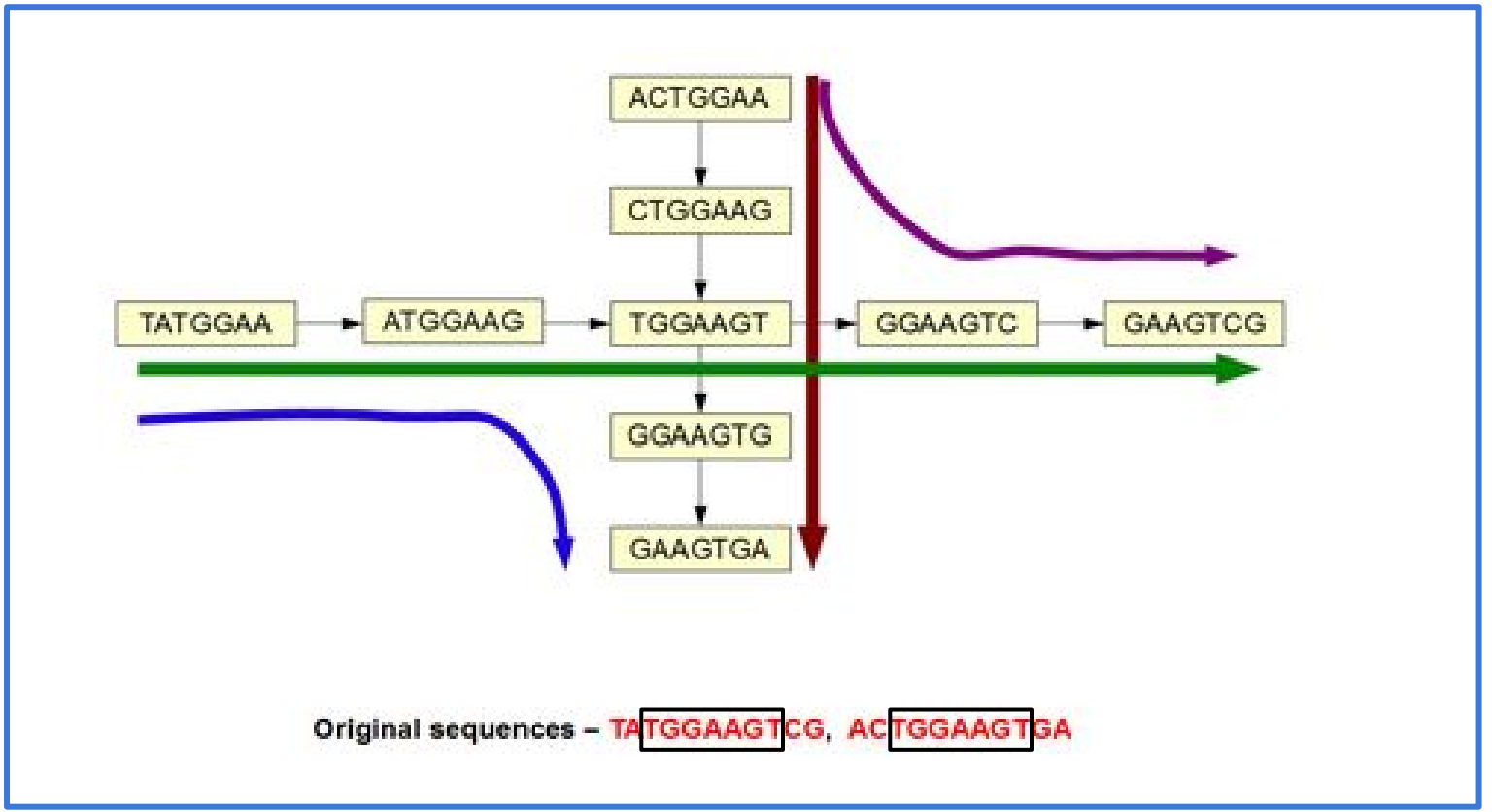


Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end)

Assembler

Scaffolds (Assembly)



Font: http://www.homolog.us/Tutorials/

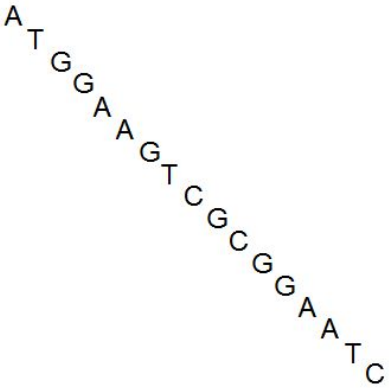
Reads (single or paired- end)

Assembler

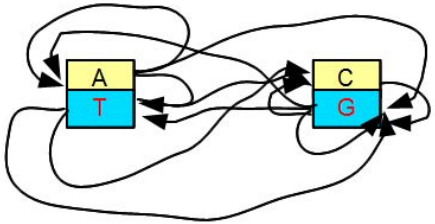
Scaffolds (Assembly)

sequence ATGGAAGTCGCGGAATC

1mers



de Bruijn graph

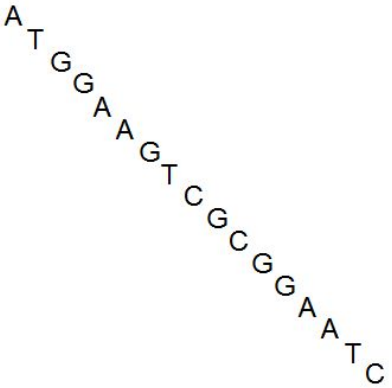


Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end) → Assembler → Scaffolds (Assembly)

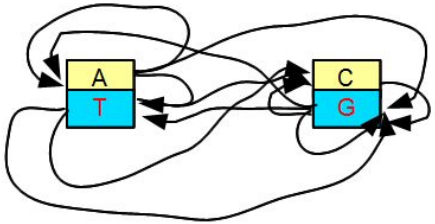
sequence ATGGAAGTCGCGGAATC

1mers



Why do you break the reads then?

de Bruijn graph



Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end) → Assembler → Scaffolds (Assembly)

- *K*-mers with errors occur fewer times than error-free *k*-mers

For each 1000 bp, Illumina mistakes 1

Genome	ACTGGGA	Sequenced 10X
K-mer	**ACTGGGA**	10
K-mer	**ATTGGGA**	1

Reads (single or paired- end) → Assembler → Scaffolds (Assembly)

Suppose there's an error

Read: GCGTACTACGCGTCTGGCCT

```
GCGTACTA: 1
 CGTACTAC: 3
  GTACTACG: 1
   TACTACGC: 1
    ACTACGCG: 2
     CTACGCGT: 1
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```

Below average

Around average

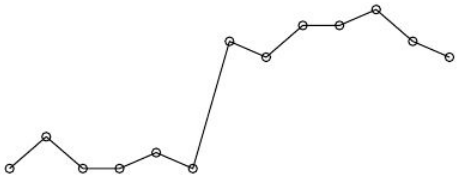
Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end)

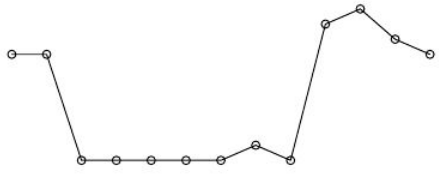
Assembler

Scaffolds (Assembly)

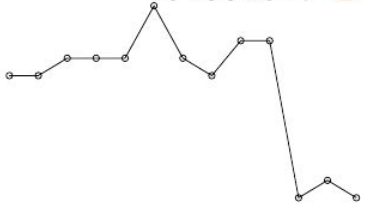
```
GCGTACTACGCGTCTGGCCT
GCGTACTA: 1
 CGTACTAC: 3
  GTACTACG: 1
   TACTACGC: 1
    ACTACGCG: 2
     CTACGCGT: 1
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```



```
GCGTATTACACGTCTGGCCT
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACA: 1
   TATTACAC: 1
    ATTACACG: 1
     TTACACGT: 1
      TACACGTC: 1
       ACACGTCT: 2
        CACGTCTG: 1
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```



```
GCGTATTACGCGTCTGGTCT
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACG: 9
   TATTACGC: 9
    ATTACGCG: 9
     TTACGCGT: 12
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGT: 1
           GTCTGGTC: 2
            TCTGGTCT: 1
```



Font: http://www.homolog.us/Tutorials/

Reads (single or paired- end) → Assembler → Scaffolds (Assembly)

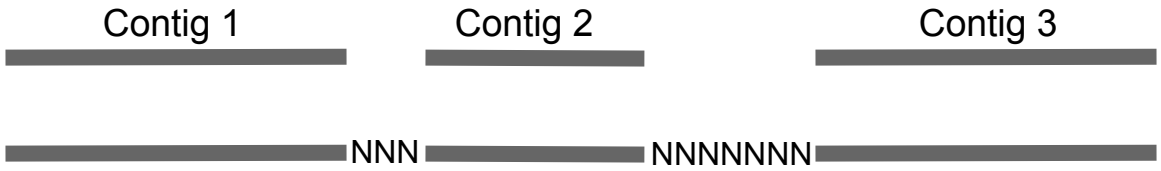
- Assemblers work differently for each species
- ABYSS Meraculous SOAPdenovo
- De Bruijn graph-based assemblers
  - Varying kmer sizes

8 assemblies

Metrics:

-N50 (Contigs)

-N50 (Scaffolds)



-Total of bases (Scaffolds)

-Gene contiguity (Busco)

-Percentage of RNA mapping (tophat)

-Mapping % to *Marmota marmota*

	N50 (scaffolds)	N50 (contigs)	Total of bases (scaffolds)
ABYSS (60kmer)	265.7kb	12.6kb	2.3 Gb
ABYSS (55kmer)	233.9kb	10.7kb	2.3 Gb
Meraculous (61kmer)	266.2Kb	6.9Kb	2.2 Gb
Meraculous (61kmer)_mode2	2.5Kb	2.2Kb	3.3 Gb
Meraculous (55kmer)	299.5Kb	7.1Kb	2.3 Gb
Meraculous (45kmer)	322.9Kb	6.6Kb	2.3 Gb
Meraculous (35kmer)	319.1kb	5.5kb	2.2 Gb
SOAPdenovo (61kmer)	82.8kb	1.3kb	3.3 Gb

	N50 (scaffolds)	N50 (contigs)	Total of bases (scaffolds)
ABYSS (60kmer)	**265.7kb**	**12.6kb**	**2.3 Gb**
ABYSS (55kmer)	233.9kb	10.7kb	2.3 Gb
Meraculous (61kmer)	266.2Kb	6.9Kb	2.2 Gb
Meraculous (61kmer)_mode2	2.5Kb	2.2Kb	3.3 Gb
Meraculous (55kmer)	299.5Kb	7.1Kb	2.3 Gb
Meraculous (45kmer)	**322.9Kb**	**6.6Kb**	**2.3 Gb**
Meraculous (35kmer)	319.1kb	5.5kb	2.2 Gb
SOAPdenovo (61kmer)	82.8kb	1.3kb	3.3 Gb

	gene contiguity (Busco)	x̄RNA mapping (%)	Mapping % to M. marmota
ABYSS (60kmer)	**C:89.0%,F:6.3%,M:4.7%,n:4104**	**83.6**	**97.2**
ABYSS (55kmer)	C:86.2%,F:8.1%,M:5.7%,n:4104	82.07	-
Meraculous (61kmer)	C:88.2%,F:6.2%,M:5.6%,n:4104	71.85	-
Meraculous (61kmer)_mode2	C:18.6%,F:29.1%,M:52.3%,n:4104	-	-
Meraculous (55kmer)	C:87.9%,F:5.8%,M:6.3%,n:4104	70.73	-
Meraculous (45kmer)	**C:87.7%,F:5.9%,M:6.4%,n:4104**	**68.32**	**88.09**
Meraculous (35kmer)	C:87.3%,F:6.1%,M:6.6%,n:4104	62.25	-
SOAPdenovo (61kmer)	C:79.5%,F:11.4%,M:9.1%,n:4104	76.76	-

ABYSS kmer60 wins!

- Checked on IGV

ABYSS kmer60 wins!

- Checked on IGV
- Average genome coverage is 95X

.. and now

- Help from the MCDB187 cohort
  - 50 “gold standard” genes

Near future applications

Comparative genomic approach

- Identify unique molecular traits
- All marmots are obligate hibernators
  - 7-8 months

    Behavioral and Physiological responses
    - Skip reproductive seasons
    - Social thermoregulation
    - Fastest growth rates in *Sciuridae*
    - Large body size

- Identify genetic signatures associated to adaptations to harsh environments

Comparative genomic approach

- “the capacity to hibernate is associated with a differential pattern of gene expression instead of with changes within gene sequences”
  - examine the level of evolutionary constraint on hibernation-related genes

- Large vs. small hibernators

Epigenetics

So many ideas!

To analyze epigenetic patterns associated with growth rates in marmots

So many ideas!

To analyze epigenetic patterns associated with growth rates in marmots

PS: Having a reference genome for this approach is a great advantage

Why this is interesting?

- One of the main predictors of marmot overwinter survival is weight
- Conditions when marmots have less time to grow:
  - Harsh years
  - Higher altitude (or slope)
- Expectation: higher growth rates in harsher conditions
- Maldonado et al. 2017

Why this is interesting?

- Costs of rapid growth rates:
  - Mature cell function
  - Somatic development
  - Immune function
- Epigenetics in blood can be associated with
  - Inflammatory, lipid and glucose metabolism genes
  - metabolic syndromes (obesity, visceral adipose tissue, coronary malfunction ...)

Thank you!

gabriela.m.pinho@gmail.com