Student presentations

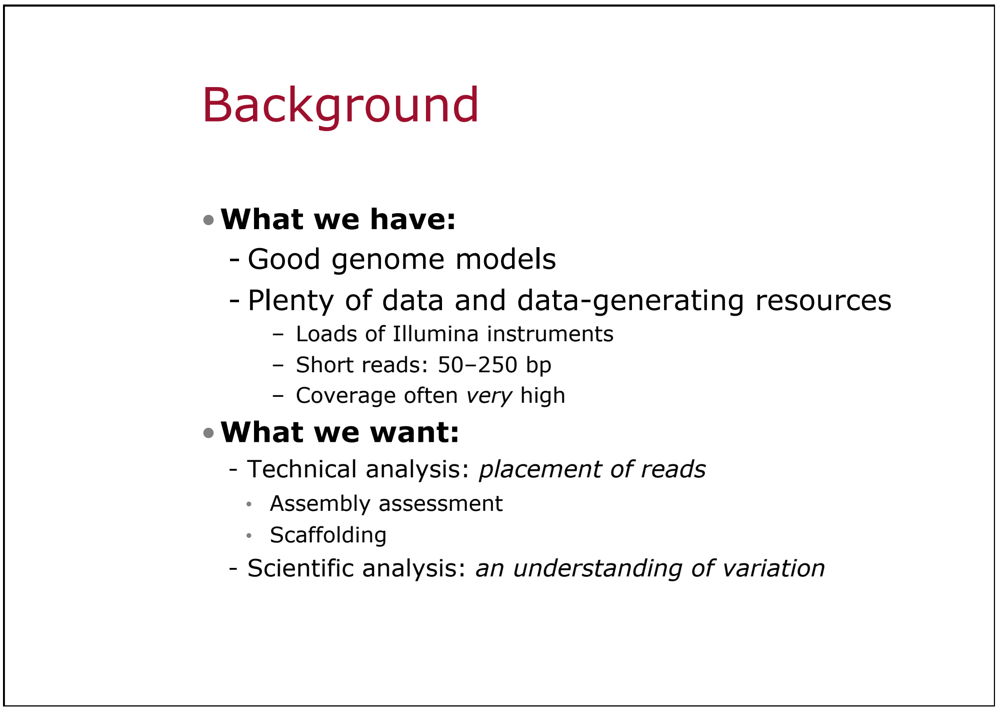
1. Assemblathon 2, by Bradnam *et al*
2. FRC for assembly comparison/evaluation, by Vezzi *et al*

ROYAL INSTITUTE OF TECHNOLOGY

Mapping short reads to a genome

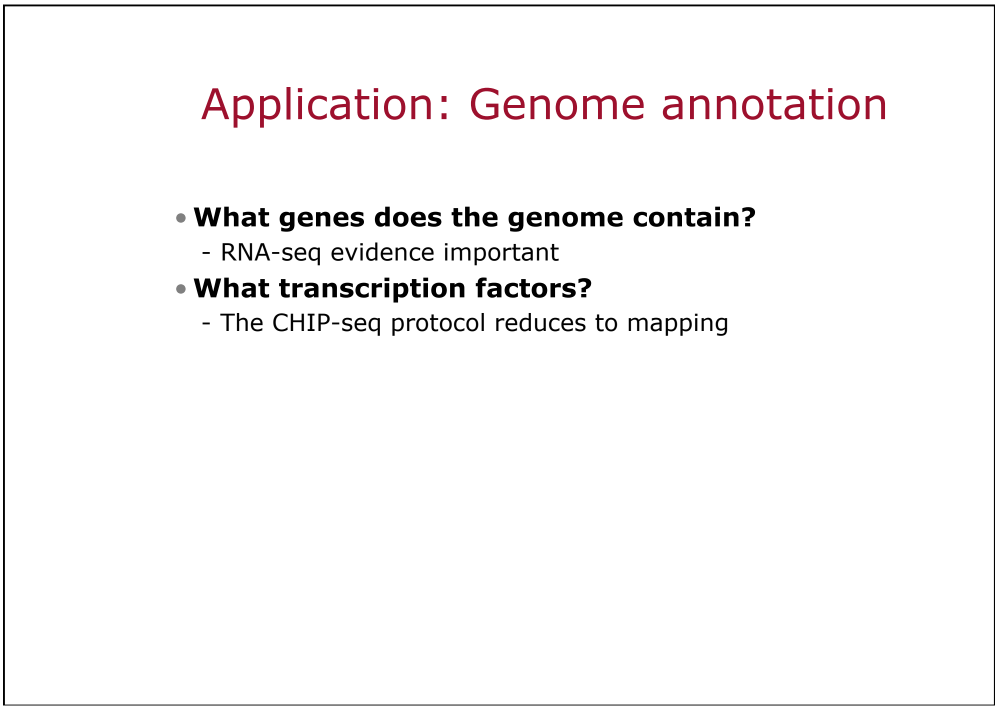
Lars Arvestad
in BB2490

Prepare for the quiz:
Trapnell and Salzberg: *How to map billions of short reads onto genomes*



Background

- **What we have:**
  - Good genome models
  - Plenty of data and data-generating resources
    - Loads of Illumina instruments
    - Short reads: 50–250 bp
    - Coverage often *very* high
- **What we want:**
  - Technical analysis: *placement of reads*
    - Assembly assessment
    - Scaffolding
  - Scientific analysis: *an understanding of variation*



Application: Genome annotation

- **What genes does the genome contain?**
  - RNA-seq evidence important
- **What transcription factors?**
  - The CHIP-seq protocol reduces to mapping

Application: Population genomics

- **What genome variation exists in the population(s)?**
  - Looking for single nucleotide variants (SNV)
    - Sometimes called ”SNPs” [snips], from Single Nucleotide Polymorphism.
      Common def: mutations with frequency > 1 %
    - In practice: all mutations
  - Structural variation (SV): inserts and deletions
  - Want to link variation to conditions and disease

Application:Differential genomics

- **Red junglefowl**
  - Wild bird
  - Healthy
  - Not fit for industrial use

- **White leghorn**
  - Domesticized bird
  - Meat and egg producer
  - Weak

Pics: Lip Kee and .brioso. at Flickr

Application: Differential genomics

LETTER

doi:10.1038/nature11837

The genomic signature of dog domestication reveals adaptation to a starch-rich diet

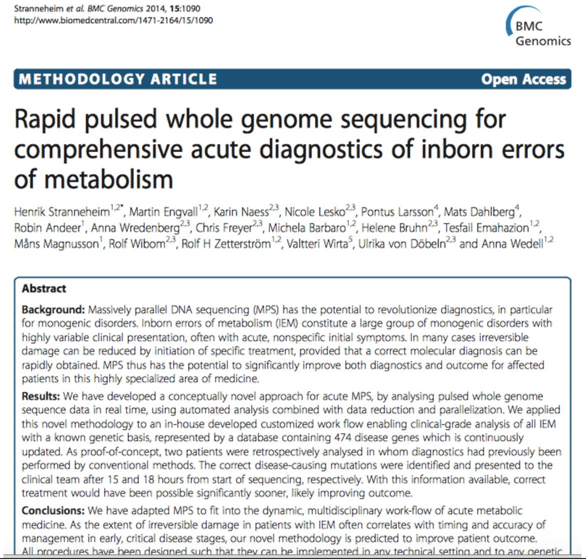
Erik Axelsson[1], Abhirami Ratnakumar[1], Maja-Louise Arendt[1], Khurram Maqbool[1], Matthew T. Webster[1], Michele Perloski[2], Olof Liberg[3], Jon M. Arnemo[4,5], Åke Hedhammar[6] & Kerstin Lindblad-Toh[1,2]

The domestication of dogs was an important episode in the development of human civilization. The precise timing and location of this event is debated[1–5] and little is known about the genetic changes that accompanied the transformation of ancient wolves into domestic dogs. Here we conduct whole-genome resequencing of dogs and wolves to identify 3.8 million genetic variants used to identify 36 genomic regions that probably represent targets for selection during dog domestication. Nineteen of these regions contain genes important in brain function, eight of which belong to nervous system development pathways and potentially underlie behavioural changes central to dog domestication[6]. Ten genes with key roles in starch digestion and fat metabolism also show signals of selection. We identify candidate mutations in key genes and provide functional support for an increased starch digestion in dogs relative to wolves. Our results indicate that novel adaptations allowing the early ancestors of modern dogs to thrive on a diet rich in starch, relative to the carnivorous diet of wolves, constituted a crucial step in the early domestication of dogs.

Domestic animals are crucial to modern human society, and it is likely colour variants in *MC1R* in pig[9] and a mutation in *TSHR* likely to affect seasonal reproduction in chicken[10], but to our knowledge in dogs no genome-wide sequence-based searches have been performed until now. To identify genomic regions under selection during dog domestication we performed pooled whole-genome resequencing of dogs and wolves followed by functional characterization of candidate genes.

Uniquely placed sequence reads from pooled DNA representing 12 wolves of worldwide distribution and 60 dogs from 14 diverse breeds (Supplementary Table 1) covered 91.6% and 94.6%, respectively, of the 2,385 megabases (Mb) of autosomal sequence in the CanFam 2.0 genome assembly[11]. The aligned coverage depth was 29.8× for all dog pools combined and 6.2× for the single wolf pool (Supplementary Table 1 and Supplementary Fig. 1). We identified 3,786,655 putative single nucleotide polymorphisms (SNPs) in the combined dog and wolf data, 1,770,909 (46.8%) of which were only segregating in the dog pools, whereas 140,818 (3.7%) were private to wolves (Supplementary Table 2). Similarly we detected 506,148 short indels and 26,619 copy-number variations (CNVs) (Supplementary Files 1 and 2). We were able to experimentally validate 113 out of 114 tested SNPs (Sup-

Application: Clinical genomics



Stranneheim *et al. BMC Genomics* 2014, **15**:1090
http://www.biomedcentral.com/1471-2164/15/1090

BMC Genomics

METHODOLOGY ARTICLE Open Access

Rapid pulsed whole genome sequencing for comprehensive acute diagnostics of inborn errors of metabolism

Henrik Stranneheim[1,2*], Martin Engvall[1,2], Karin Naess[2,3], Nicole Lesko[2,3], Pontus Larsson[4], Mats Dahlberg[4], Robin Andeer[1], Anna Wredenberg[2,3], Chris Freyer[2,3], Michela Barbaro[1,2], Helene Bruhn[2,3], Tesfail Emahazion[1,2], Måns Magnusson[1], Rolf Wibom[2,3], Rolf H Zetterström[1,2], Valtteri Wirta[5], Ulrika von Döbeln[2,3] and Anna Wedell[1,2]

Abstract

Background: Massively parallel DNA sequencing (MPS) has the potential to revolutionize diagnostics, in particular for monogenic disorders. Inborn errors of metabolism (IEM) constitute a large group of monogenic disorders with highly variable clinical presentation, often with acute, nonspecific initial symptoms. In many cases irreversible damage can be reduced by initiation of specific treatment, provided that a correct molecular diagnosis can be rapidly obtained. MPS thus has the potential to significantly improve both diagnostics and outcome for affected patients in this highly specialized area of medicine.

Results: We have developed a conceptually novel approach for acute MPS, by analysing pulsed whole genome sequence data in real time, using automated analysis combined with data reduction and parallelization. We applied this novel methodology to an in-house developed customized work flow enabling clinical-grade analysis of all IEM with a known genetic basis, represented by a database containing 474 disease genes which is continuously updated. As proof-of-concept, two patients were retrospectively analysed in whom diagnostics had previously been performed by conventional methods. The correct disease-causing mutations were identified and presented to the clinical team after 15 and 18 hours from start of sequencing, respectively. With this information available, correct treatment would have been possible significantly sooner, likely improving outcome.

Conclusions: We have adapted MPS to fit into the dynamic, multidisciplinary work-flow of acute metabolic medicine. As the extent of irreversible damage in patients with IEM often correlates with timing and accuracy of management in early, critical disease stages, our novel methodology is predicted to improve patient outcome. All procedures have been designed such that they can be implemented in any technical setting and to any genetic

Computational problem: *variant detection*

Not focused on in this course!

- **In**: A *mapping* of (paired) reads

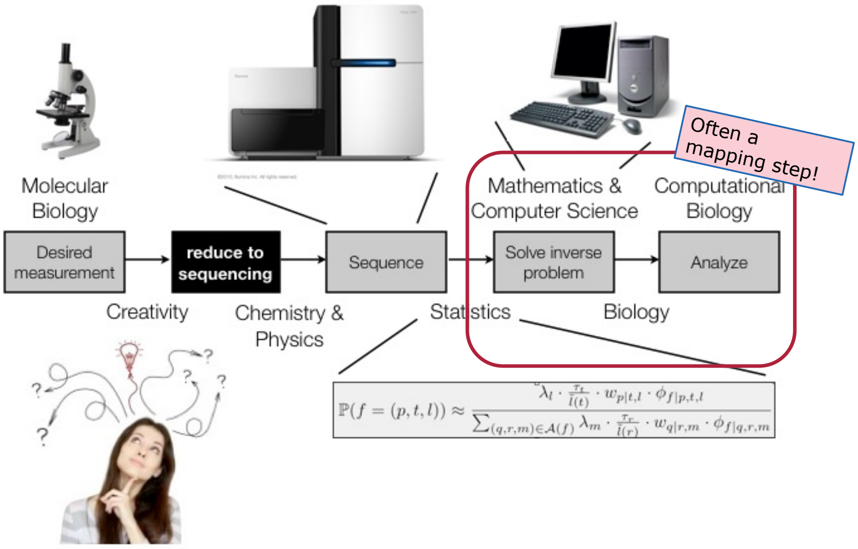
- **Out**:
  - Single nucleotide variation (SNV)

  and/or

  - Structural variation (insertion/ deletions) of various sizes

Lior Pachter's insight:

Generally: *-seq



Often a mapping step!

Molecular Biology

Desired measurement → reduce to sequencing → Sequence → Solve inverse problem → Analyze

Creativity

Chemistry & Physics

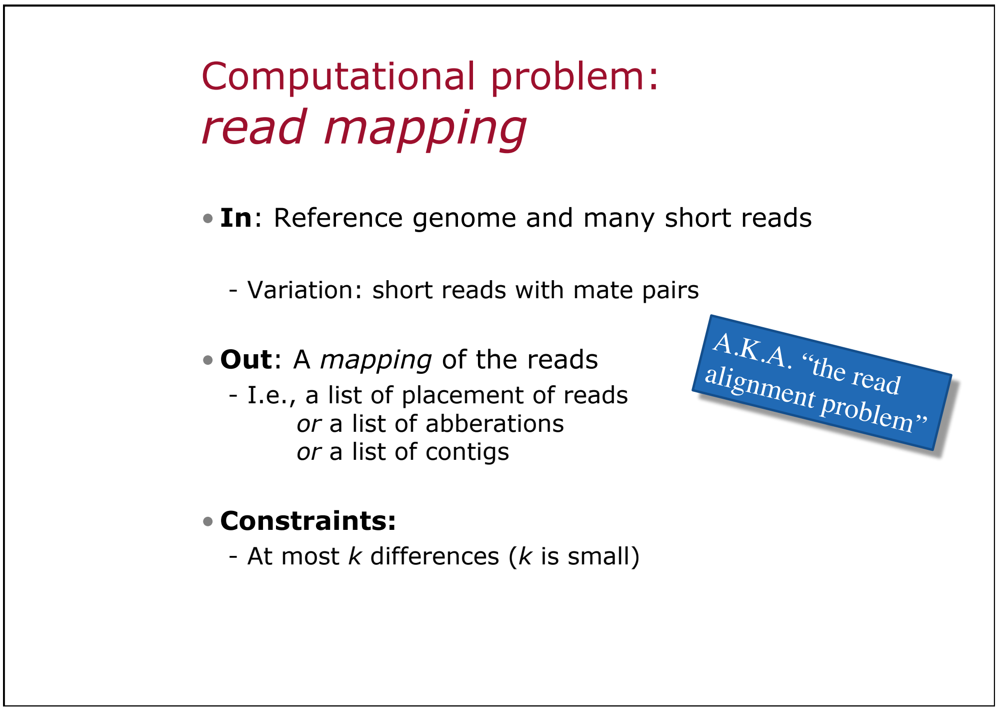
Statistics

Mathematics & Computer Science

Computational Biology

Biology

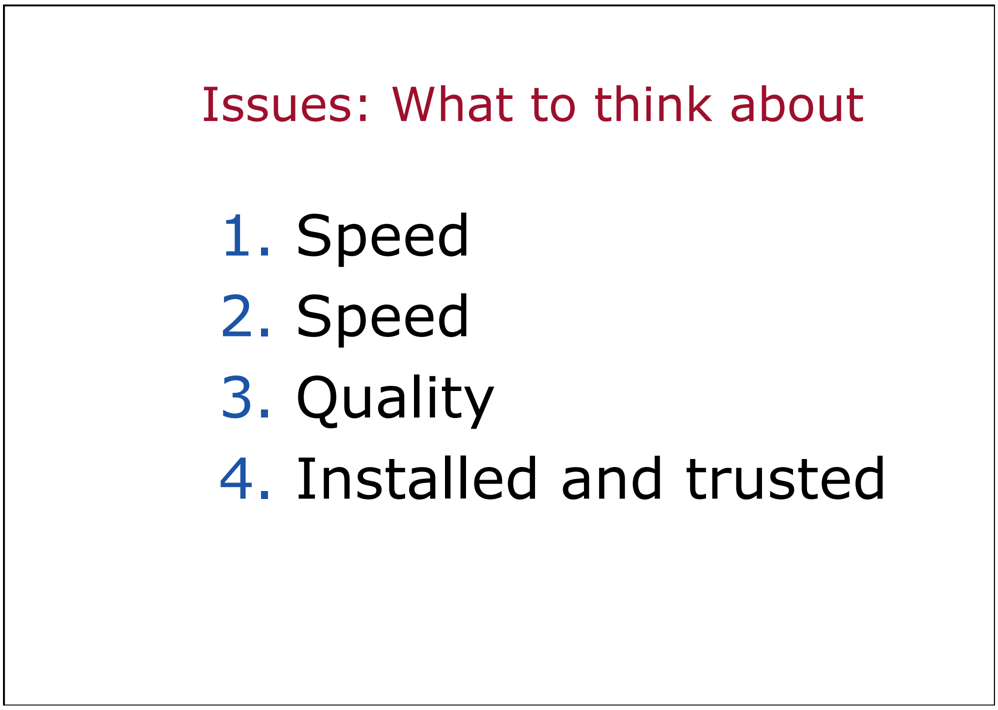
$$\mathbb{P}(f = (p,t,l)) \approx \frac{\lambda_l \cdot \frac{\tau_t}{\tilde{l}(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}}{\sum_{(q,r,m)\in\mathcal{A}(f)} \lambda_m \cdot \frac{\tau_r}{\tilde{l}(r)} \cdot w_{q|r,m} \cdot \phi_{f|q,r,m}}$$



Computational problem: *read mapping*

- **In**: Reference genome and many short reads
  - Variation: short reads with mate pairs
- **Out**: A *mapping* of the reads
  - I.e., a list of placement of reads
    or a list of abberations
    or a list of contigs
- **Constraints:**
  - At most *k* differences (*k* is small)

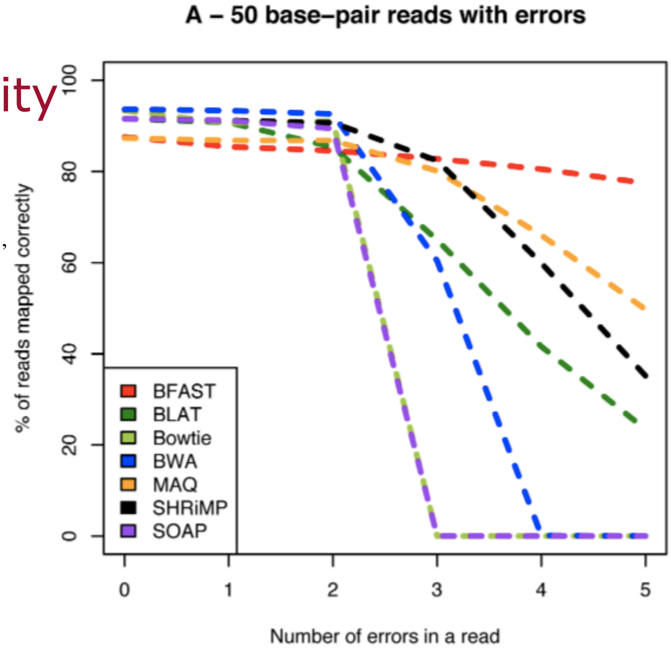
A.K.A. "the read alignment problem"



Issues: What to think about

1. Speed
2. Speed
3. Quality
4. Installed and trusted

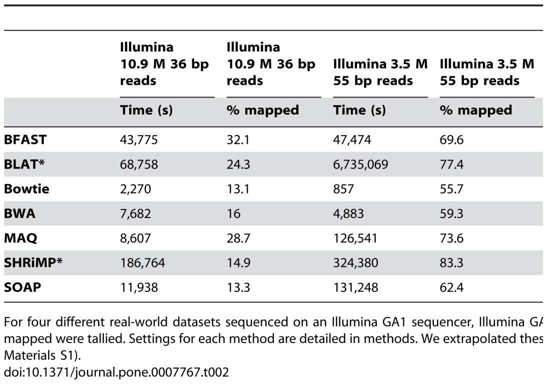
Quality



From
Homer, Merriman and Nelson,
PLoS ONE, 2009

Warning: really old data

Speed and coverage

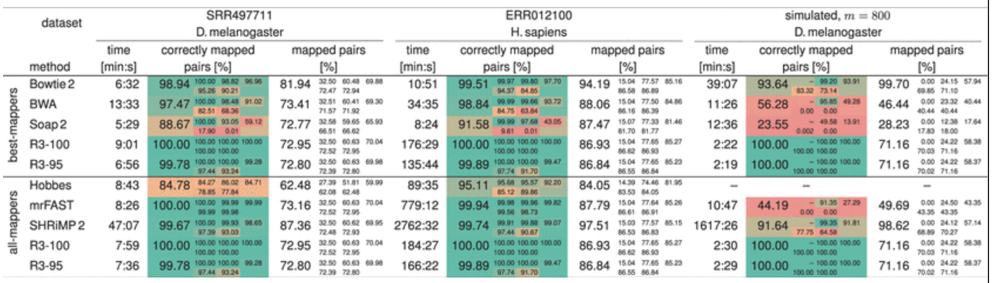


	Illumina 10.9 M 36 bp reads	Illumina 10.9 M 36 bp reads	Illumina 3.5 M 55 bp reads	Illumina 3.5 M 55 bp reads
	Time (s)	% mapped	Time (s)	% mapped
BFAST	43,775	32.1	47,474	69.6
BLAT*	68,758	24.3	6,735,069	77.4
Bowtie	2,270	13.1	857	55.7
BWA	7,682	16	4,883	59.3
MAQ	8,607	28.7	126,541	73.6
SHRiMP*	186,764	14.9	324,380	83.3
SOAP	11,938	13.3	131,248	62.4

For four different real-world datasets sequenced on an Illumina GA1 sequencer, Illumina G
mapped were tallied. Settings for each method are detailed in methods. We extrapolated thes
Materials S1).
doi:10.1371/journal.pone.0007767.t002

Homer, Merriman, and Nelson, PLoS ONE, 2009

Speed and coverage



	dataset	SRR497711 D. melanogaster					ERR012100 H. sapiens					simulated, $m = 800$ D. melanogaster				
	method	time [min:s]	correctly mapped pairs [%]		mapped pairs [%]		time [min:s]	correctly mapped pairs [%]		mapped pairs [%]		time [min:s]	correctly mapped pairs [%]		mapped pairs [%]	
best-mappers	Bowtie 2	6:32	98.94	100.00 98.82 96.96 95.26 90.21	81.94	32.50 60.48 69.88 72.47 72.94	10:51	99.51	99.97 99.80 97.70 94.37 84.85	94.19	15.04 77.57 85.16 86.58 86.89	39:07	93.64	– 99.20 93.91 83.32 73.14	99.70	0.00 24.15 57.94 69.85 71.10
	BWA	13:33	97.47	100.00 98.48 91.02 82.51 68.36	73.41	32.51 60.41 69.30 71.57 71.92	34:35	98.84	99.99 99.66 93.72 84.75 63.84	88.06	15.04 77.50 84.86 86.16 86.39	11:26	56.28	– 95.85 49.28 0.00 0.00	46.44	0.00 23.32 40.44 40.44 40.44
	Soap 2	5:29	88.67	100.00 93.05 59.12 17.90 0.01	72.77	32.58 59.65 65.93 66.51 66.62	8:24	91.58	99.99 97.68 43.05 9.61 0.01	87.47	15.07 77.33 81.46 81.70 81.77	12:36	23.55	– 49.58 13.91 0.002 0.00	28.23	0.00 12.38 17.64 17.83 18.00
	R3-100	9:01	100.00	100.00 100.00 100.00 100.00 100.00	72.95	32.50 60.63 70.04 72.52 72.95	176:29	100.00	100.00 100.00 100.00 100.00 100.00	86.93	15.04 77.65 85.27 86.62 86.93	2:22	100.00	– 100.00 100.00 100.00 100.00	71.16	0.00 24.22 58.38 70.03 71.16
	R3-95	6:56	99.78	100.00 100.00 99.28 97.44 93.24	72.80	32.50 60.63 69.98 72.39 72.80	135:44	99.89	100.00 100.00 99.47 97.74 91.70	86.84	15.04 77.65 85.23 86.55 86.84	2:19	100.00	– 100.00 100.00 100.00 100.00	71.16	0.00 24.22 58.37 70.02 71.16
all-mappers	Hobbes	8:43	84.78	84.27 86.02 84.71 78.85 77.84	62.48	27.39 51.81 59.99 62.08 62.48	89:35	95.11	95.68 95.57 92.20 85.12 89.86	84.05	14.39 74.46 81.95 83.53 84.05	–	–		–	
	mrFAST	8:26	100.00	100.00 99.99 99.99 99.99 99.98	73.16	32.50 60.63 70.04 72.52 72.95	779:12	99.94	99.98 99.96 99.82 99.56 98.73	87.79	15.04 77.64 85.26 86.61 86.91	10:47	44.19	– 91.35 27.29 0.00 0.00	49.69	0.00 24.50 43.35 43.35 43.35
	SHRiMP 2	47:07	99.67	100.00 99.93 98.65 97.39 93.03	87.36	32.50 60.62 69.95 72.48 72.93	2762:32	99.74	99.91 99.88 99.07 97.44 90.67	97.51	15.03 77.57 85.15 86.53 86.83	1617:26	91.64	– 99.35 91.81 77.75 64.58	98.62	0.00 24.12 57.14 68.89 70.27
	R3-100	7:59	100.00	100.00 100.00 100.00 100.00 100.00	72.95	32.50 60.63 70.04 72.52 72.95	184:27	100.00	100.00 100.00 100.00 100.00 100.00	86.93	15.04 77.65 85.27 86.62 86.93	2:30	100.00	– 100.00 100.00 100.00 100.00	71.16	0.00 24.22 58.38 70.03 71.16
	R3-95	7:36	99.78	100.00 100.00 99.28 97.44 93.24	72.80	32.50 60.63 69.98 72.39 72.80	166:22	99.89	100.00 100.00 99.47 97.74 91.70	86.84	15.04 77.65 85.23 86.55 86.84	2:29	100.00	– 100.00 100.00 100.00 100.00	71.16	0.00 24.22 58.37 70.02 71.16

Paired end reads: 10^7

Weese, Holtgrewe, and Reinert, Bioinformatics 2012

Popular software

All open source

- **BWA**, by Heng Li
  - BWA-SW: a Smith-Waterman step added
  - BWA-MEM: tuned for longer reads (“up to a few megabases”)
- **Bowtie2**
- Stampy
  - Good at indels, fast
- SOAP2
- ABySS-map

Monday’s student presentation

One paper:

- Langmead and Salzberg: *Fast gapped-read alignment with Bowtie 2*

- Everyone prepares at least one question!