ROYAL INSTITUTE
OF TECHNOLOGY

Mapping short reads to a genome

ROYAL INSTITUTE OF TECHNOLOGY

Background

- **What we have:**
  - Good genome models
  - Plenty of data and data generating resources
    - Illumina + Solid instruments
    - Short reads: 30 – 100 bp
    - Coverage often *very* high
- **What we want:**
  - A better understanding of variation

ROYAL INSTITUTE OF TECHNOLOGY

Application: Population genomics

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

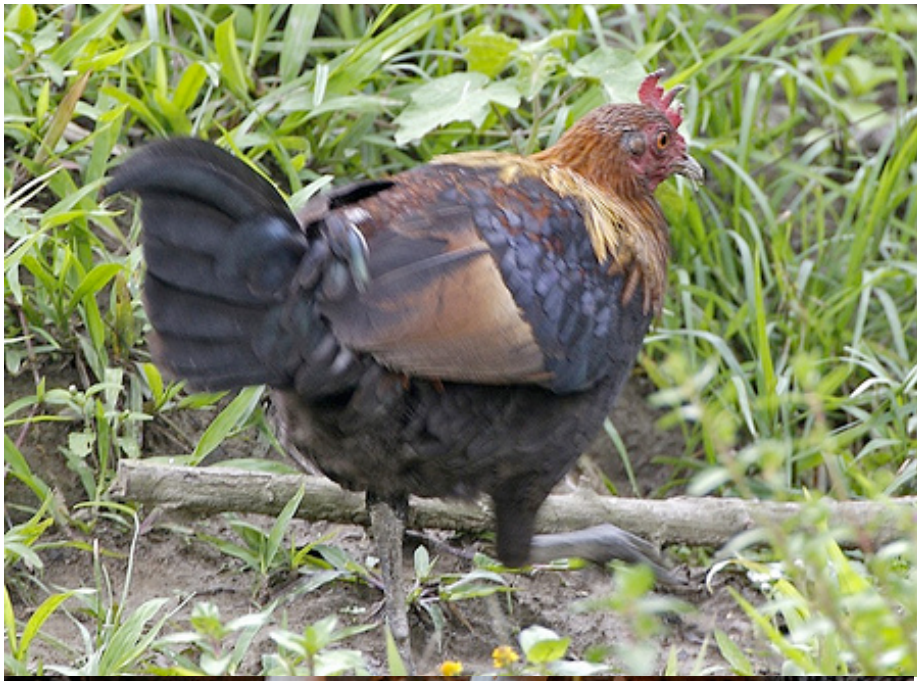
ROYAL INSTITUTE OF TECHNOLOGY

Systems biology?

- Professor 1: ”I am doing Systems Biology”
- Professor 2: ”No, you don’t, I do.”

- *Systematic* biology is not *Systems* biology

Application: Differential genomics



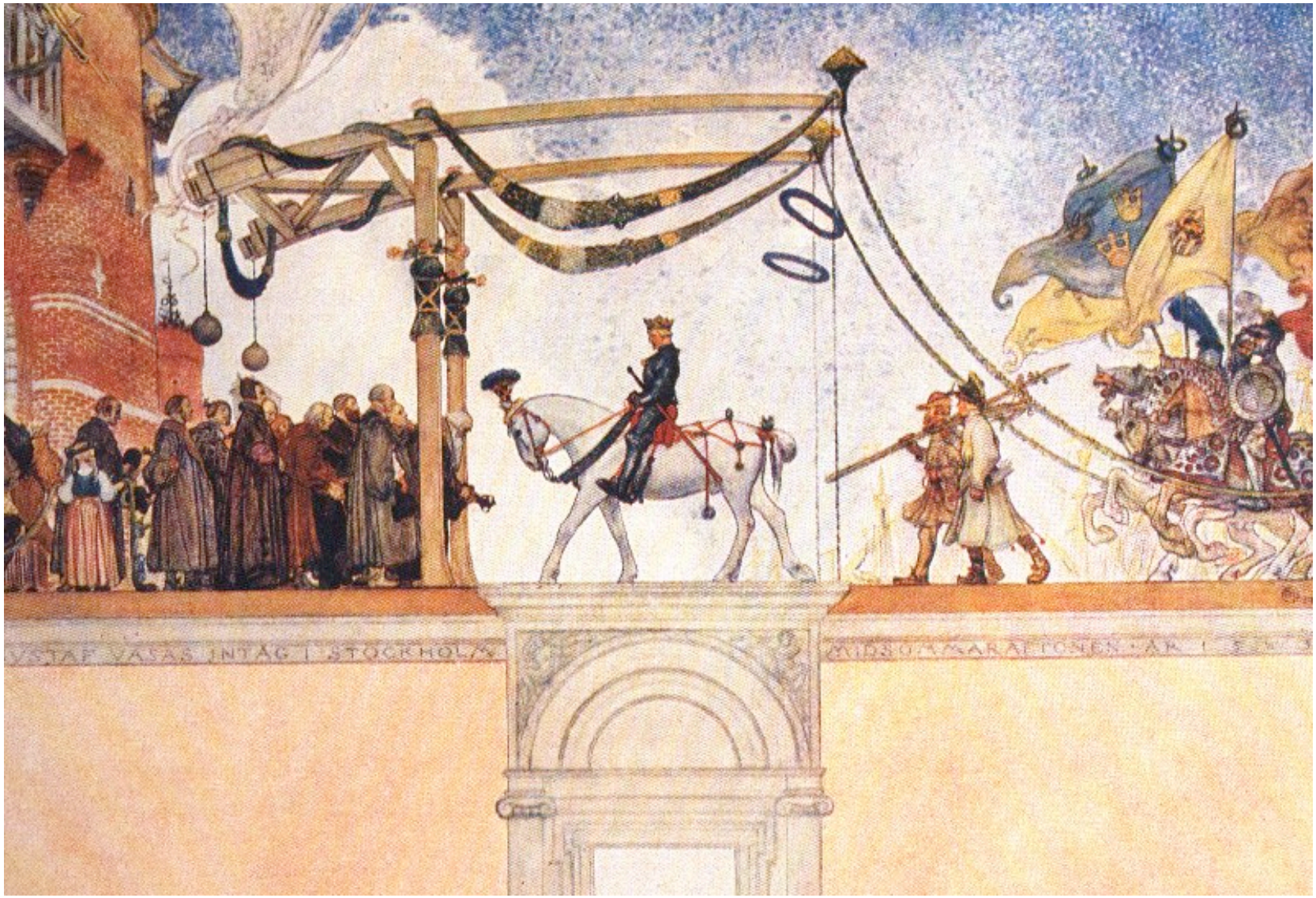
- **Red junglefowl**
  - Wild bird
  - Healthy
  - Not fit for industrial use

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

Pics: Lip Kee and .brioso. at Flickr

ROYAL INSTITUTE OF TECHNOLOGY

Application: Differential genomics



”Gustav Vasa enters Stockholm” by Carl Larsson— at Nationalmuseum

ROYAL INSTITUTE OF TECHNOLOGY

Computational problem

- **In**: Reference genome and many short reads

  - Variation: short reads with mate pairs
  - Variation: Solid reads in colorspace

- **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

ROYAL INSTITUTE OF TECHNOLOGY

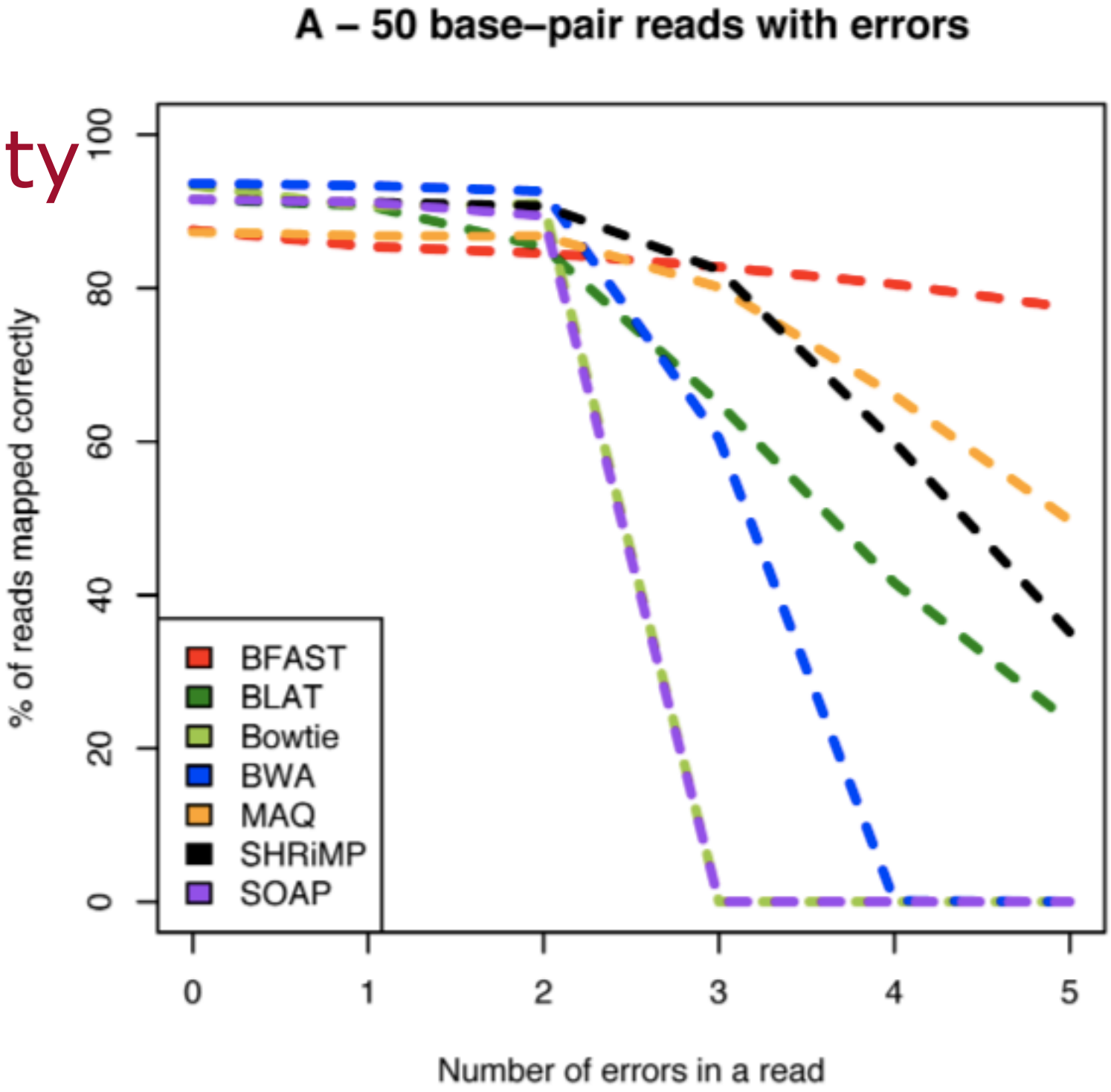
Issues: What to think about

1. Speed
2. Speed
3. Speed
4. Quality

ROYAL INSTITUTE OF TECHNOLOGY

Quality

From
Homer, Merriman and Nelson,
PLoS ONE, 2009



ROYAL INSTITUTE OF TECHNOLOGY

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	ABI SOLiD 1 M 25 bp read	ABI SOLiD 1 M 25 bp read	ABI SOLiD 1 M 50 bp read	ABI SOLiD 1 M 50 bp read
	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped
BFAST	43,775	32.1	47,474	69.6	9,590	66	42,856	72.5
BLAT*	68,758	24.3	6,735,069	77.4	NA	NA	NA	NA
Bowtie	2,270	13.1	857	55.7	NA	NA	NA	NA
BWA	7,682	16	4,883	59.3	21,179	74.7	845	47.8
MAQ	8,607	28.7	126,541	73.6	7,602	63.6	6,680	68.1
SHRiMP*	186,764	14.9	324,380	83.3	2,977	2.4	32,644	70.4
SOAP	11,938	13.3	131,248	62.4	NA	NA	NA	NA

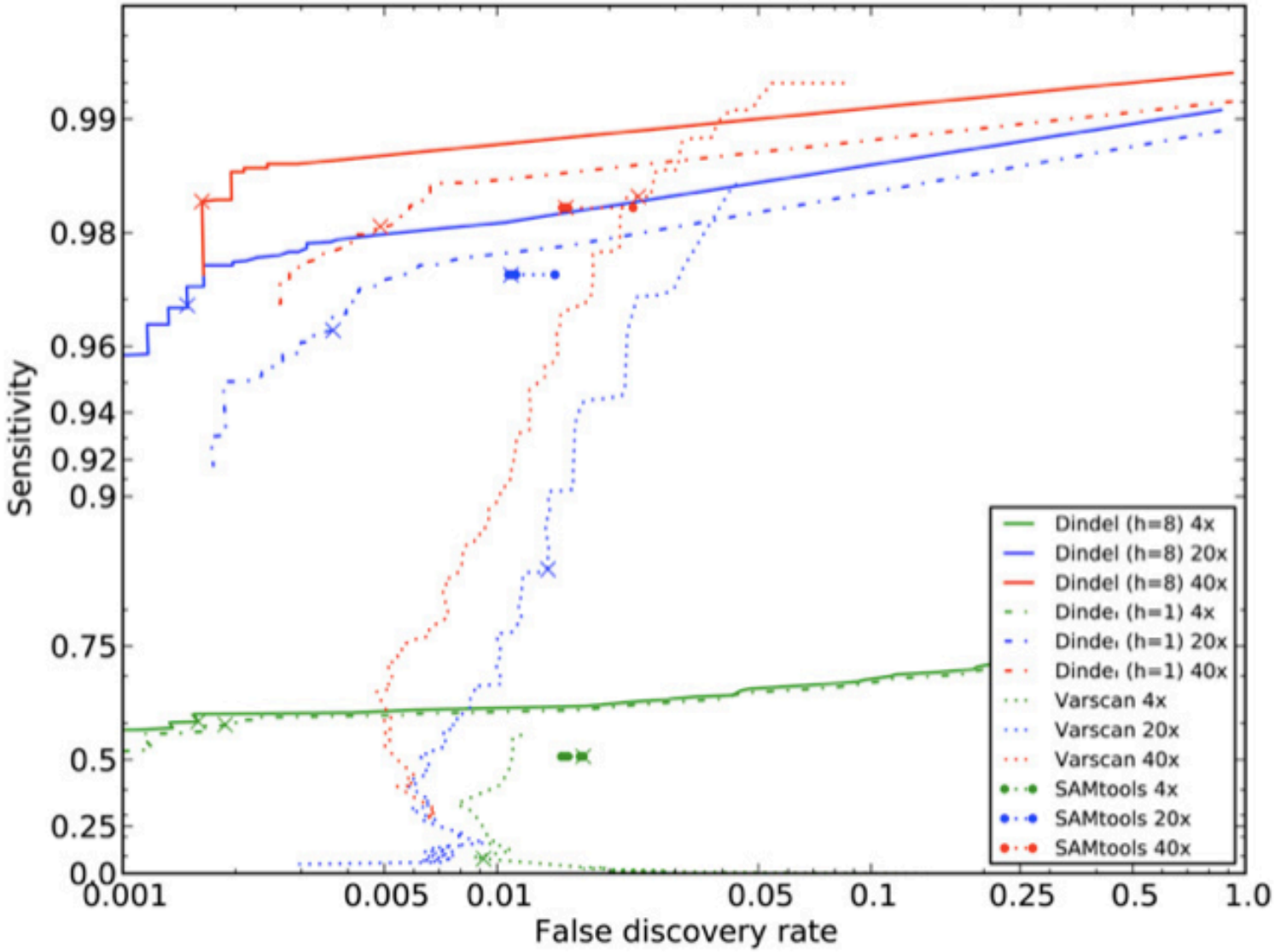
For four different real-world datasets sequenced on an Illumina GA1 sequencer, Illumina GAII and an ABI SOLiD sequencer, the run time and the fraction of reads mapped were tallied. Settings for each method are detailed in methods. We extrapolated these values for those methods denoted with an asterisk (*) (see Supplemental Materials S1).

doi:10.1371/journal.pone.0007767.t002

Homer, Merriman, and Nelson, PLoS ONE, 2009

ROYAL INSTITUTE OF TECHNOLOGY

Indel sensitivity



From Albers *et al*, Genome Research, 2011

ROYAL INSTITUTE OF TECHNOLOGY

Popular software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Other software

- **SHRiMP**

Table from Trapnell and Salzberg, Nat Biotech, 2009