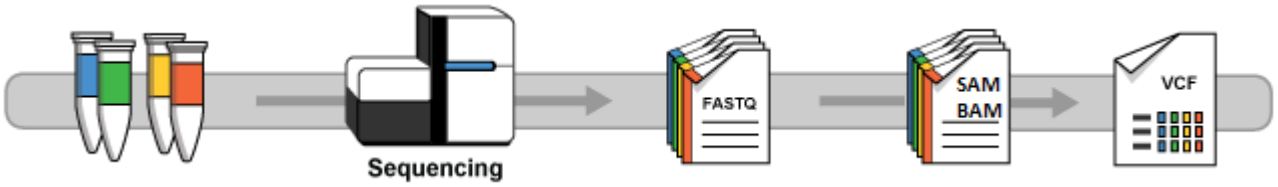
Lossy compression of genomics datasets

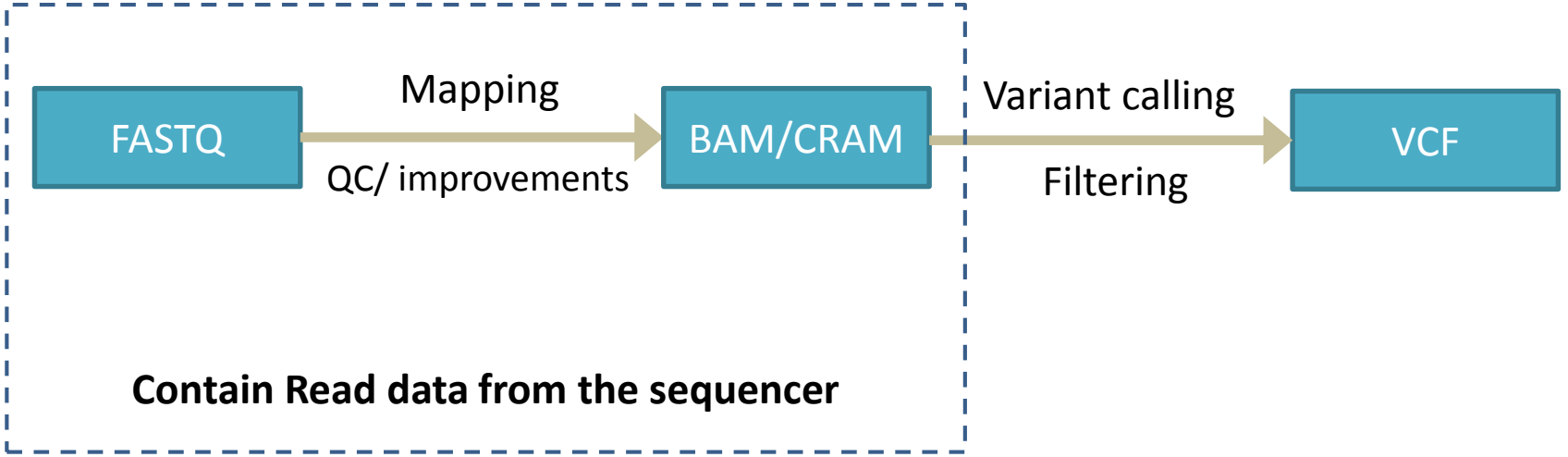
October, 2015

Dr Dan Greenfield

PetaGene

Overview - NGS





Overview - NGS

- Example Read:

Sequence bases:	GCAGTATGCCTGGTGTATTTCAGAAACAACCA
Quality scores (QS):	@CCDFDEDFIHHDGGI@GI@FGH?<@A<I?>@

- For Illumina reads, QS takes 2.3-2.8x more space than sequence reads (compressed)
- Lossy compression is reaching its limits and these limits are not good enough!
- *Lossless compression*

Lossy compression - overview

- Lossy compression for quality scores:
  - Binning/quantization:
    Illumina 8bin, UniBinning, Truncating, LogBinning
  - Smoothing: P-Block, R-Block
  - Rate distortion: QualComp, QVZ
  - Corpus based: RQS/Quartz, GeneCodeq

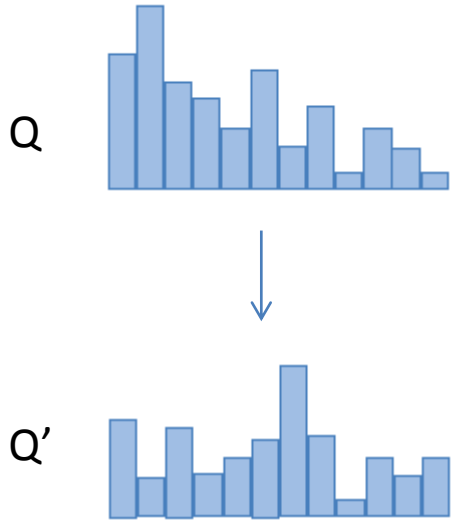
- Impact of loss?!

Lossy compression: evaluation

- Two approaches to evaluation
  - Rate-distortion theory metrics
  - Impact in downstream applications

Lossy compression: evaluation

- Some reported work uses metrics common to rate-distortion theory, e.g.:
  - Mean squared error
  - L1
  - Log(1+L1)
  - Max:min distance



Lossy compression: evaluation

- Measure the impact of compression in downstream applications
  - Genotyping accuracy
- Genotyping accuracy metrics:
  - ROC curves
  - Precision
  - Recall
  - F-score
- Many lossy compression algorithms claim to improve genotyping accuracy!

Lossy compression: evaluation

- Naïve compression algorithms do not utilise valuable information that can guide lossy compression
  - E.g. similarities between the sequencing sample and the reference genome of the sample species
- Quartz

- Quality scores are not equally important!

Lossy compression: evaluation

- Two approaches …
  - Rate-distortion theory metrics
  - Impact in genotyping accuracy
- GeneCodeq and Quartz look very bad in rate-distortion theory metrics
- … but are much better in genotyping accuracy!

- Rate distortion metrics are not really suitable.

Measuring genotyping accuracy

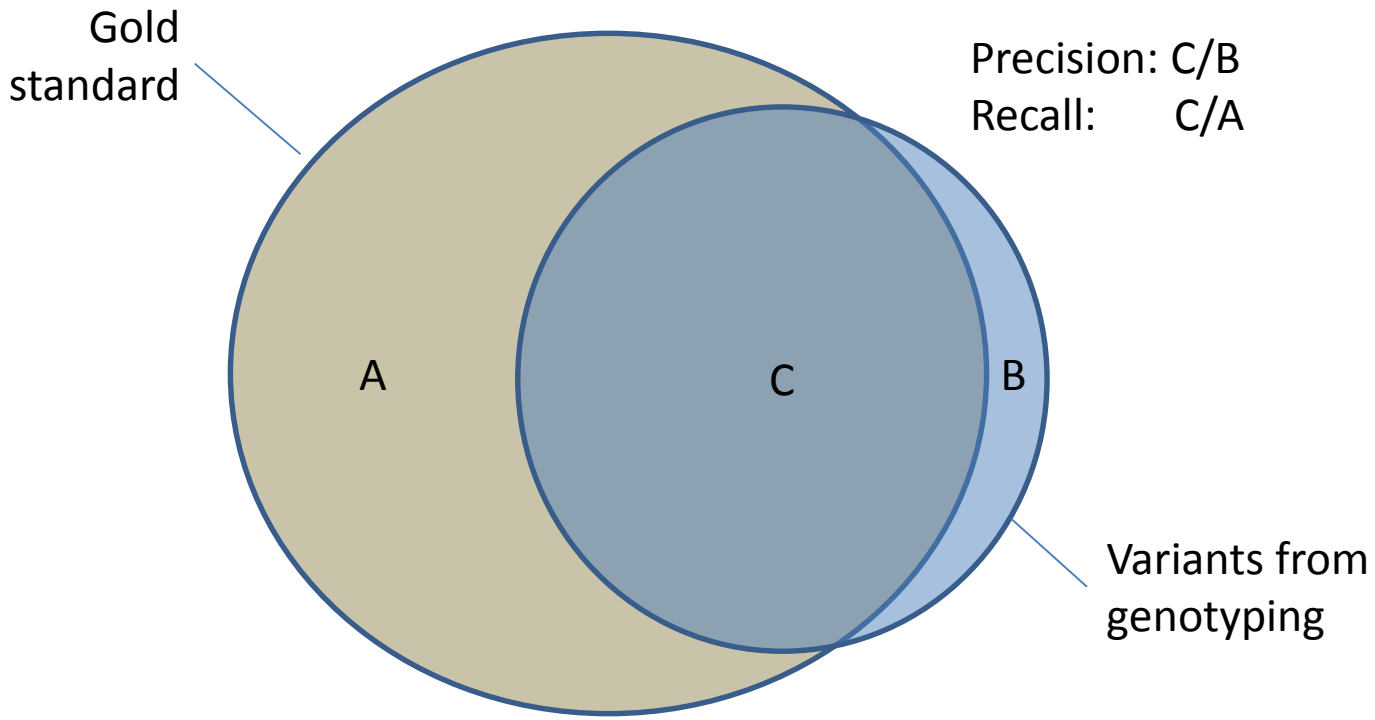
- Other lossy compression algorithms also claim improvements in genotyping accuracy
- How can accuracy be improved when we are reducing information provided to the tools?

Measuring genotyping accuracy

- The output of genotyping is a vcf file (listing identified variants)

```
...
1       858801  . A           G       87      .   DP=5;VDB=0.913383  GT:PL   1/1:114,12,0
1       861808  . TG          T       70      .   DP=5;VDB=0.522837  GT:PL   1/1:97,12,0
1       861630  . GTTTCTTTC   GTTTC   107     .   DP=5;VDB=0.807392  GT:PL   1/1:134,15,0
...
```

Measuring genotyping accuracy



- Compare lossless and lossy versions
- These measurements do not capture filtering of variants based on quality, hence the use of ROC curves…

Choosing a gold standard

- What's commonly used is far from 'golden'
- E.g. NA12878
  - Illumina Platinum: 4,495,450 variants
  - Genome in a bottle: 3,163,064 variants
- Choose carefully, otherwise it introduces biases!

Choosing a gold standard

- Example:
- Two approaches producing variants sets X, Y

Approach	True precision	True recall
X	99%	99%
Y	90%	80%

- Compare them against ‘gold’ standards A, B:
  - A: 30% of true variants are missing, 5% are false
  - B: 5% of true variants are missing, 30% are false

Choosing a gold standard

- Example (…continued)
- Comparing against A

Approach	Observed precision (A)	Observed recall (A)
X	~70%	~95%
Y	~80%	~95%

Comparing against B

Approach	Observed precision (B)	Observed recall (B)
X	~95%	~70%
Y	~88%	~71%

- X seems worse than Y when using A!

Approach	True precision	True recall
X	99%	99%
Y	90%	80%

Lossy compression and genotyping accuracy

- In the absence of new information, genotyping accuracy should not be expected to improve
- If seen, improvements could be due to:
  - Flaws in the variant calling pipeline
  - Variant callers not leveraging all available information
  - By misleading measurements
- Corpus based approaches (GeneCodeq, Quartz) can improve accuracy by utilising the information available in the corpus
- The tradeoff between loss of genotyping accuracy and entropy reduction in quality scores remains!

GeneCodeq

- Utilising Coding Theory and Bayesian probability to adjust quality scores

- Reference genome

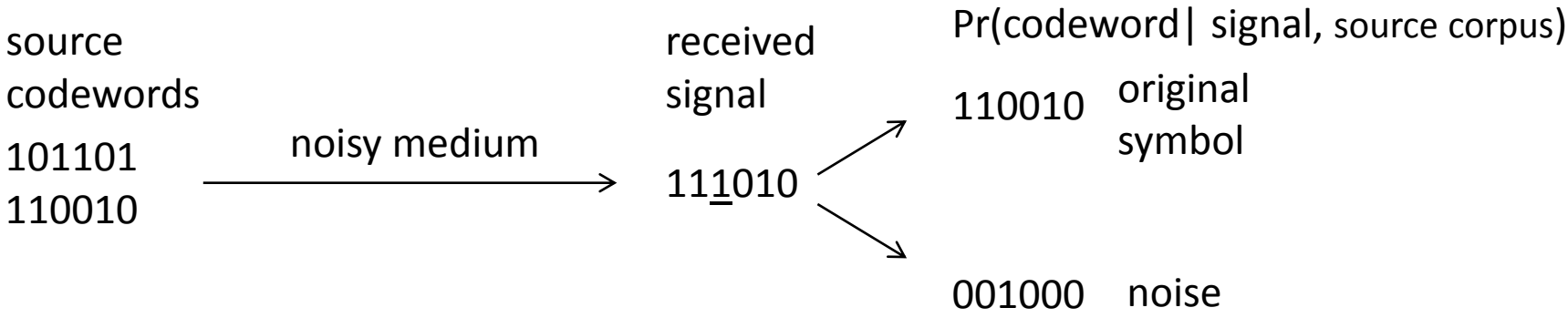
Mutation

Sample genome

… or transmission via a noisy medium

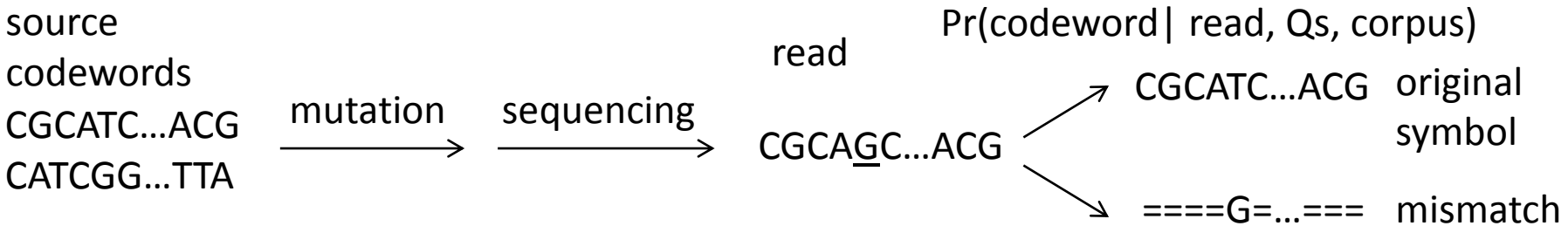
GeneCodeq

- Transmission through a noisy medium



GeneCodeq

- Sequencing as a coding theory problem



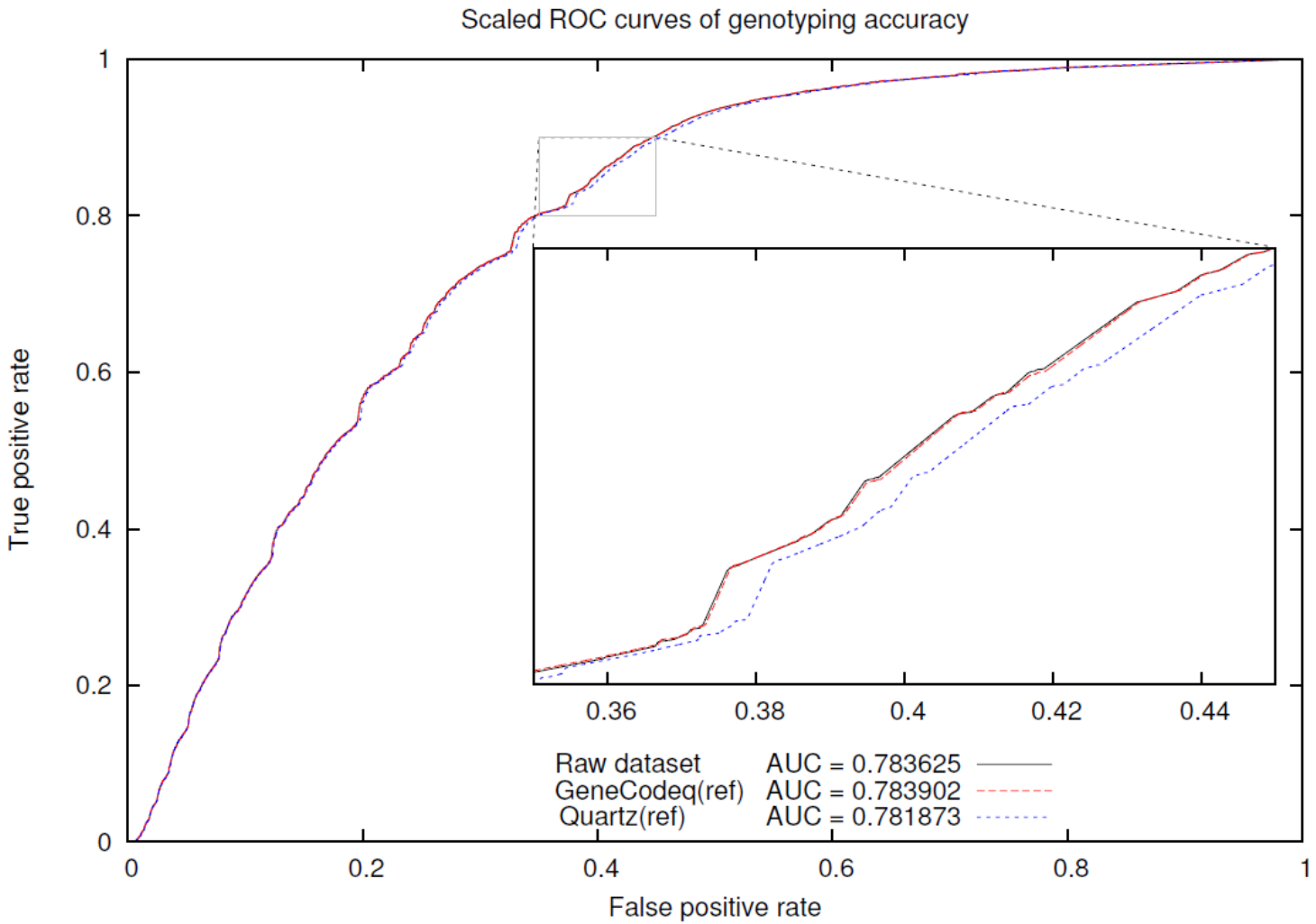
Q = Pr(seq err) ⟶ Q' = Pr(seq err | read, Qs, corpus)

- Example corpus: reference genome

GeneCodeq

- Good compression, but what about genotyping accuracy?

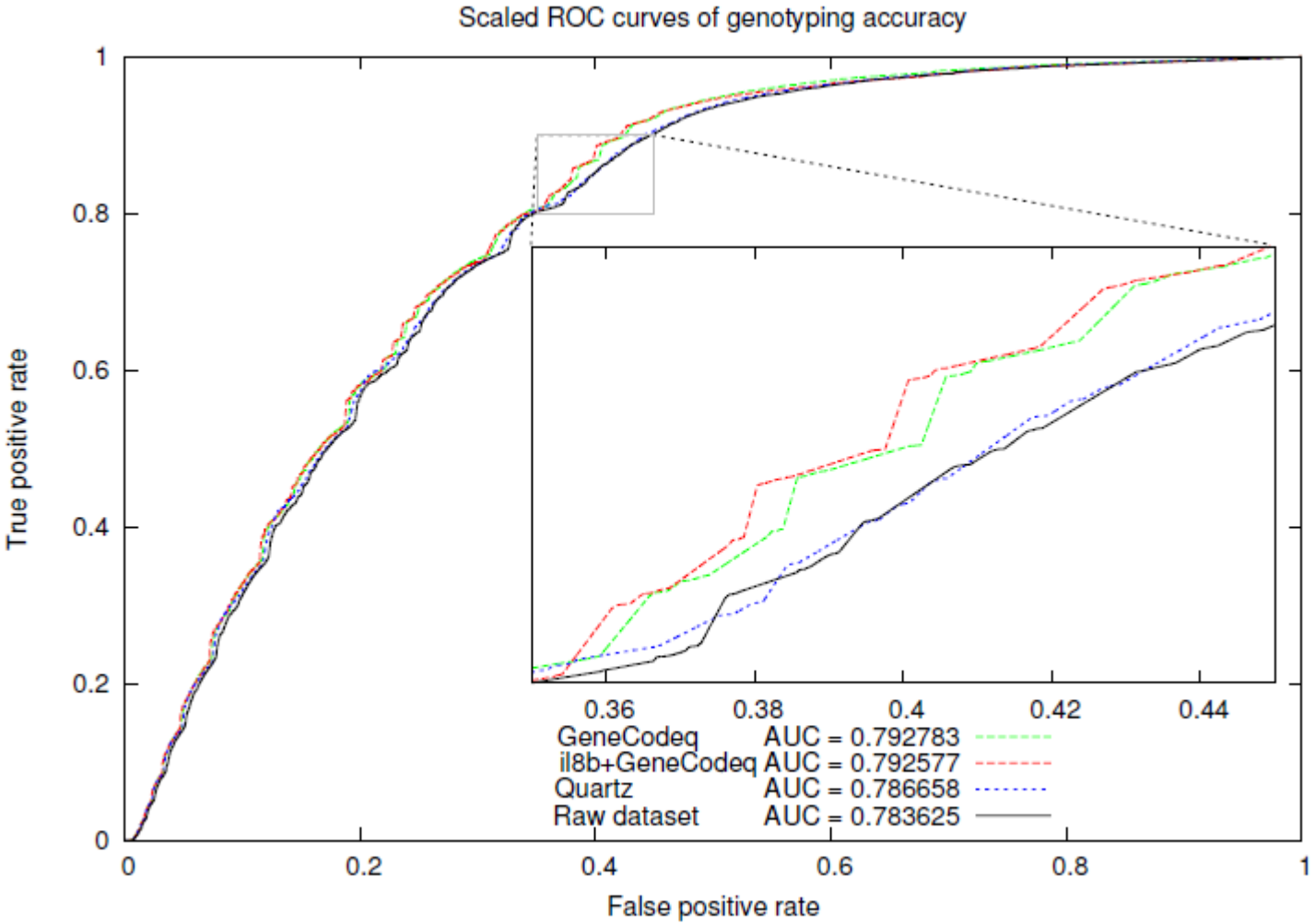
GeneCodeq


Scaled ROC curves of genotyping accuracy
True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
0.36
0.38
0.4
0.42
0.44
Raw dataset AUC = 0.783625
GeneCodeq(ref) AUC = 0.783902
Quartz(ref) AUC = 0.781873

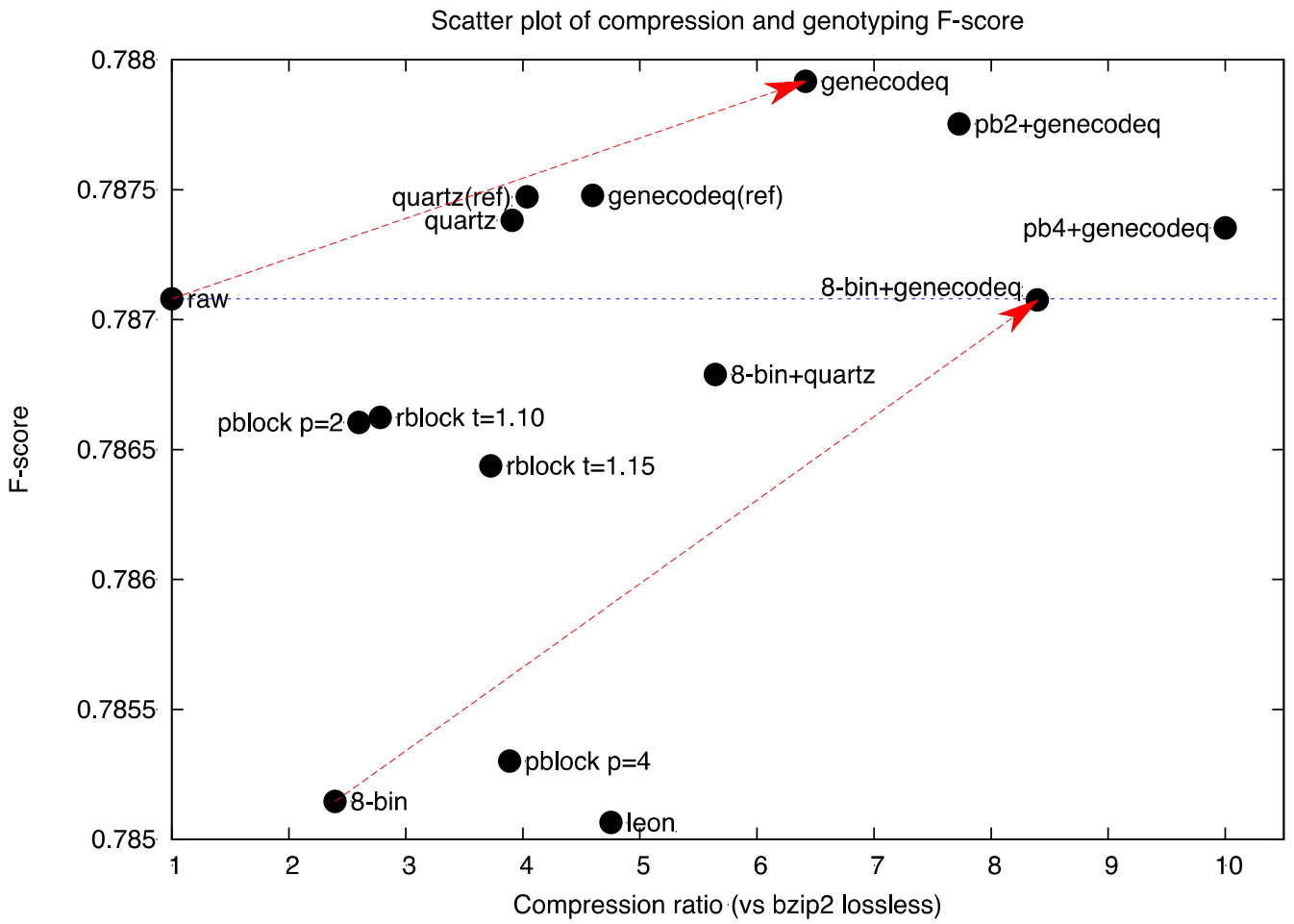
GeneCodeq

- Good compression, but what about genotyping accuracy?
- Genotyping accuracy is not reduced!
- … Genotyping accuracy can be improved using a richer corpus
  - E.g. add information about common variants from the 1000 Genomes Project

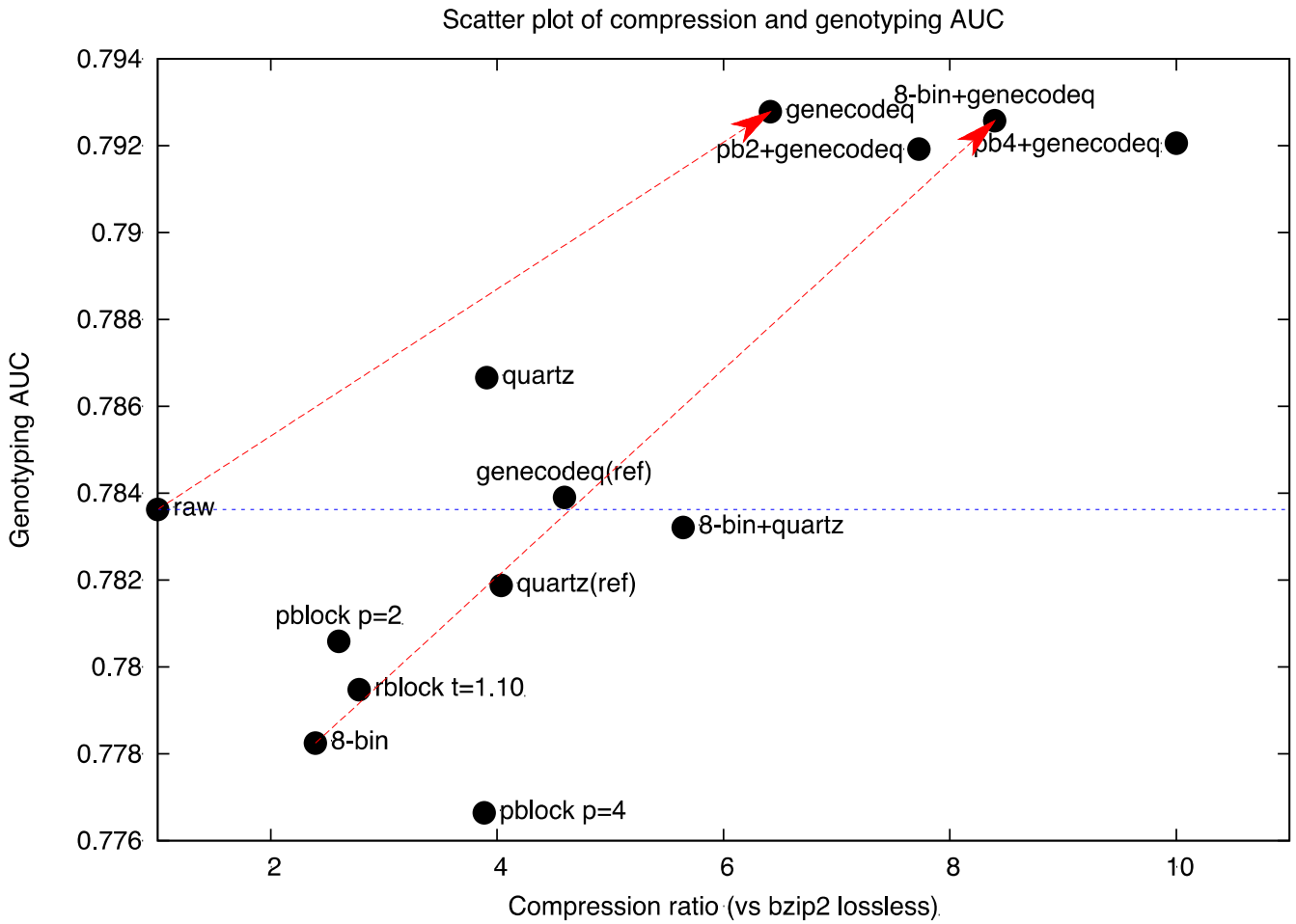
GeneCodeq


Scaled ROC curves of genotyping accuracy
True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1
0.36
0.38
0.4
0.42
0.44
GeneCodeq AUC = 0.792783
il8b+GeneCodeq AUC = 0.792577
Quartz AUC = 0.786658
Raw dataset AUC = 0.783625

F-Score vs Compression Ratio



AUC vs Compression Ratio



Acknowledgements

- Thanks to:
  - Alban Rrustemi (PetaGene)
  - Oliver Stegle (EMBL-EBI)

- Download the paper and evaluation of GeneCodeq from: www.petagene.com/eval/genecodeq/