

Error-Correction Methods for High-Throughput Sequencing Data

Some slides contributed by D.Weese, FU Berlin

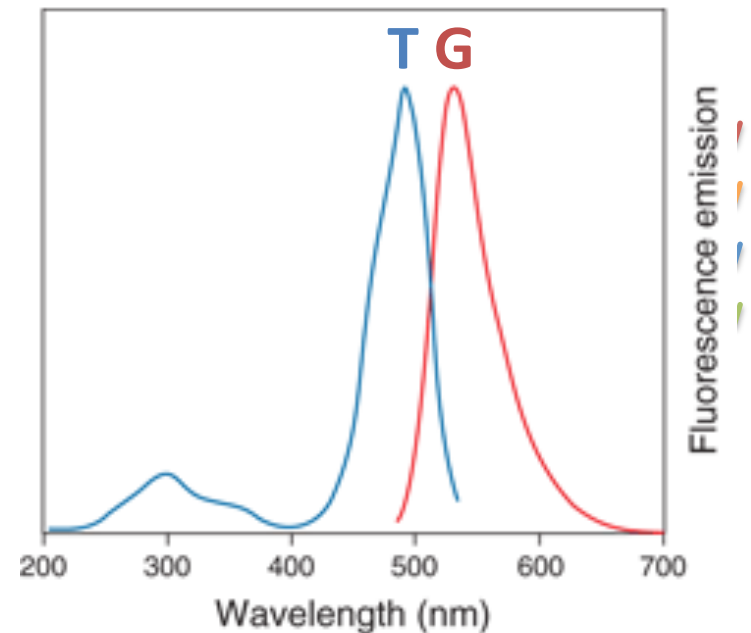
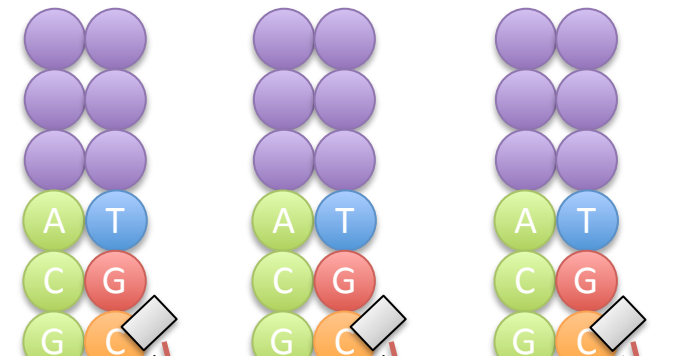
Contents

- Introduction
- Three Different Approaches
 - MSA
 - k-spectrum
 - Suffix tree
- Results and Conclusion

INTRODUCTION

Error Sources - Illumina

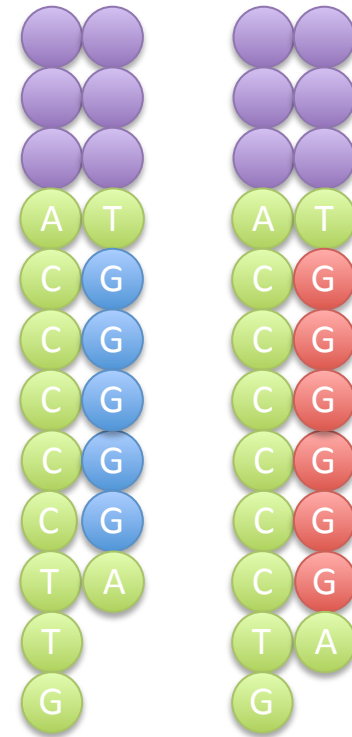
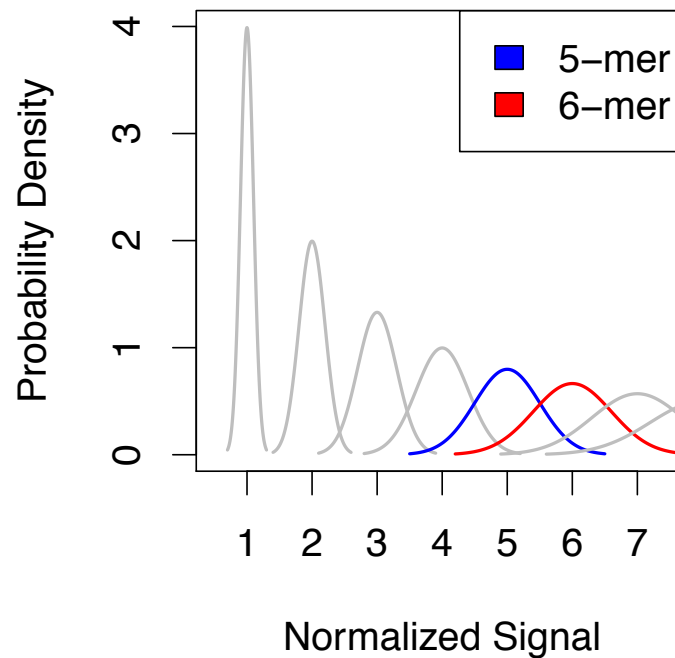
- Dephasing (Dohm *et al.*, 2008)
 - Some templates lag behind
 - Mixed signals towards the 3'-end
- Overlapping emission spectra
 - T can be mistaken for G
 - 4 times more G→T errors compared to G→A or G→C errors
- Dominant errors:
 - **Substitution errors**



Error Sources – Roche/454

- Signal Resolution and Noise

- Light intensity corresponds to homopolymer length
- Difficult to distinguish signals for homopolymers >6bp (Mardis, 2008)



- Dominant errors:

- Indels

The need to correct errors

- Sequencing errors affect almost every HTS application:
 - Fragmented contigs in assembly
 - Error correction reduces mismatches in assembly by 50% (MacManes *et al.*, 2013)
 - Misaligned or unaligned reads → inaccurate SNP detection
 - After correction more reads align to SNP locations, more SNPs can be discovered (Kelley *et al.*, 2010)

BUT!

- Sequencing errors must be distinguished from:
 - Heterozygous SNPs in polyploid organisms
 - Differences in repeat copies

How to correct errors in theory

- **Scenario 1:** Suppose **sequenced genome and read layout** are known
 - Errors are mismatches/indels between reads and genome
 - Use genome to correct reads
 - Impossible, often there is not even a *reference genome*

```
ATACAATTATATCTTATTTCCATTCCCATATGTGGTACGCAATATCCTAAA
  ACAATTA    CTTATTT  ATTCCCA    GTGGTAC  CAATAT  CTAAA
ATACAAT  TATCTTA    CCATTCCC    TGTGGTA  GCAATAT  TAAA
TACAATT  ATCTTAT    CATTCCCA    TGGTACG
  ACAATTA          ATTTCCA  CCCATAT  GGTACGC    TCCTAAA
ACAATTA  TCTTATT  CCATTCC    TGTGGTA  AATATCC
```

How to correct errors in theory

- **Scenario 2:** Suppose **only read layout** is known
 - We could infer the genome from the consensus
 - Use consensus to correct reads
 - Requires that correct bases are dominant (data redundancy)
 - Computing the complete MSA (multiple sequence alignment) is as hard as de-novo assembly

```
ATACAATTATATCTTATTTCCATTCCCATATGTGGTACGCAATATCCTAAA
ACAATTA    CTTATTT    ATTCCCA    GTGGAAC  CAATAT  CTAAA
ATACAAT    TATCTTA    CCATTCC    TGTGGAA  GCAATAT  TAAA
TACAATT    ATCTTAT    CATTCCCA    TGGAACG
ACAATTA            ATTTCCA    CCCATAT    GGAACGC    TCCTAAA
ACAATTA    TCTTATT    CCATTCC    TGTGGAA    AATATCC
```


How to correct errors in theory

- **Scenario 3:** Suppose **only reads** are given
 - We could locally approximate the MSA
 - Requires to detect overlapping reads
 - Could be feasible if we use seeding and some heuristics

```
ATTTCATTCCCATAT
      ATTCCCA
      CCATTCC
      CATTCCCA
ATTTCCA  CCCATAT
      CCATTCC
```

Pitfalls: SNPs vs. Errors

- Heterozygous SNPs could easily be mistaken for sequencing errors
- Solution:
 - Assume uniform sampling from both haplotypes
 - Use distribution of base frequencies in the MSA column

```
haplotype1 ****GATTTCATAT*****
haplotype2 ****GATTACATAT*****ATTCCCCATT*****
      GATTACA                      ATTCCCC
      GATTTCA                      TTCCCCA
      ATTTCAT                      TTCCCCA
      TTTCAATT                     TCCCTAT
      TTACATT                      CCCCAATT
      TTACATT                      CCCCAATT
      TACATAT
      TTCATAT
      ↑
heterozygous SNP
```

error

Pitfalls: Repeats vs. Errors


- Differences between repeat copies

*****GATTACATAT*****GATTACATAT*****TATTCCATAT*****
GATTACA ATTACAT TATTCCA
TTACATA TACATAT TATTCCA
TCCATAT

- Would be compressed by MSA

GATTACA
TATTCCA
TATTCCA
ATTACAT
TTACATA
TACATAT
TCCATAT

coincident
differences



Related Work

- MSA based:
 - Coral (Salmela *et al.*, 2011)
 - ECHO (Kao *et al.*, 2011)
- k-spectrum based:
 - EULER-SR (Chaisson *et al.*, 2009)
 - CUDA-EC (Shi *et al.*, 2010)
 - Reptile (Yang *et al.*, 2010)
 - Quake (Kelley *et al.*, 2010)
 - Allpaths-LG (Gnerre *et al.*, 2011)
- Suffix tree/array based:
 - SHREC (Schröder *et al.*, 2009)
 - Hybrid SHREC (Salmela *et al.*, 2010)
 - HiTEC (Ilie *et al.*, 2011)

MSA APPROACH

MSA Approach

- Rationale:
 - Assume that reads overlap in the genome if they share a k-mer (substring of length k)
 - Use k-mer as anchor to incrementally form a multiple alignment
 - Sequencing errors appear as alignment errors

TTTCAATTCCCATATGTGGTAC

TCAATTCCCATATA

CAATTCC**CATAT**

ATTC**GCATAT**GT

TTCC**CATAT**GTG

TTCC**CATAT**GTG

TCC**CATAT**GTGG

CC**CATAT**GTGGT

Coral – Alignment Step

1. Build k-mer index of all reads and their reverse-complements
2. For each read r compute MSA:
 - Start with only r
 - Incrementally anchor reads to the consensus that share a k-mer with r
 - Extend alignment using banded Needleman-Wunsch alignment

```
GATTACA-TAT
AGATT-CA
      TACACTAT
      ATTACACT
      GATTACAC
```

Coral – Correction Step

- Evaluate rows and columns of MSA
 - *quality*: fraction of read alignment agreeing with consensus
 - *support*: fraction of reads agreeing with a consensus base
 - *weighted support*: quality sum of agreeing bases
- Correct deviating bases with consensus if ...
 - minimal *quality* exceeds threshold
 - (*weighted*) *support* exceeds threshold

AGATTACACTAT

GATTACACTAT

AGATTACA

TACACTAT

ATTACACT

GATTACAC

ECHO (Kelley *et al.*, 2011)

- ECHO:
 - Similar to Coral, but no support for indels
 - Uses a positional base-miscall profile
 - Models heterozygosity
 - 10 possible genotypes: AA, CC, GG, TT, AC, AG, AT, CG, CT, GT

K-SPECTRUM APPROACH

k-Spectrum Approach

- Rationale:
 - Consider all overlapping k-mers of the reads
 - k-mers from correct reads are frequent (**solid**)
 - k-mers with errors are infrequent (**weak**)
 - don't occur in the genome
 - are similar to solid k-mers

TTTCAATTCCCATATGTGGTAC

TCAATTCCCAATA

CAATTCCCATAT

ATTC**G**CAATATGT

TTCCCATATGTG

TTCCCATATGTG

TCCCATATGTGG

CCCATATGTGGT

#TCC = **5**

#CCC = **6**

#CCA = **6**

#TC**G** = **1**

#C**G**C = **1**

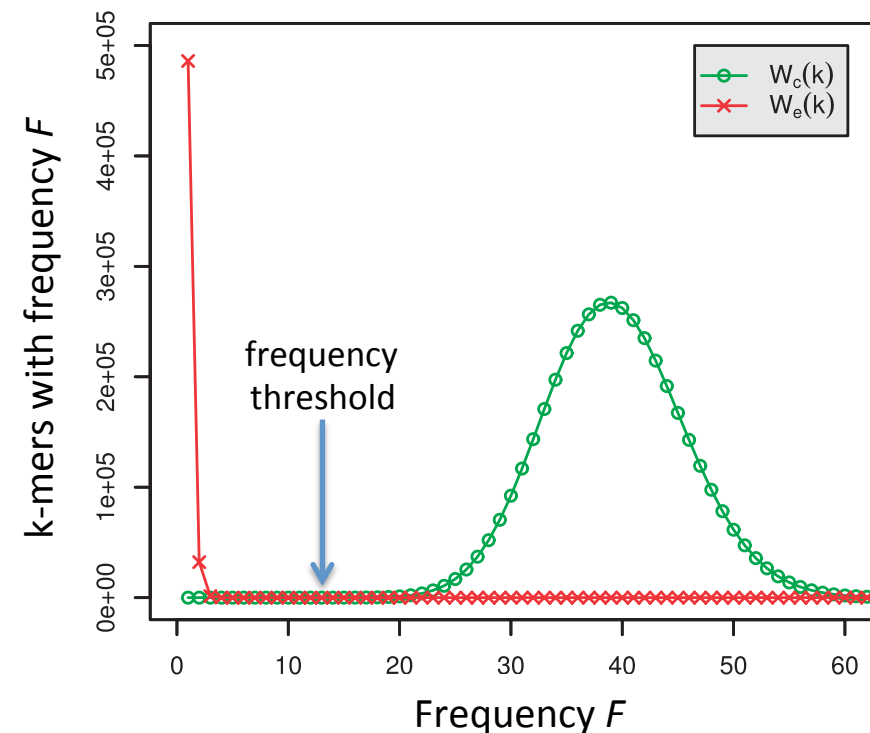
#**G**CA = **1**

Algorithm Outline

1. Determine frequency threshold for weak/solid k-mers
 - Model the frequency distributions of correct/erroneous k-mers
 - Determine frequency threshold using likelihood ratio
2. Heuristically replace weak with solid k-mers

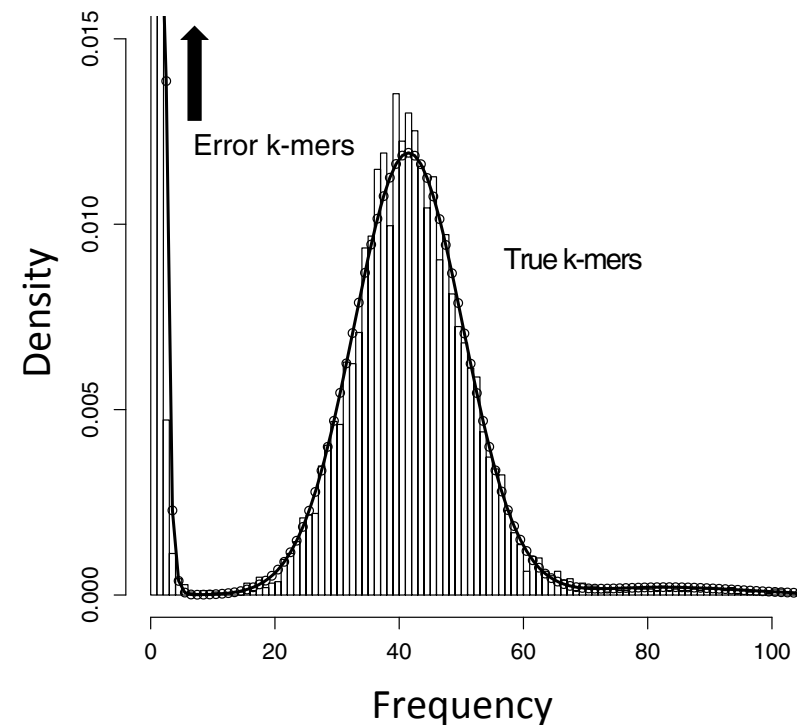
Theoretical Model for k-mer Frequencies

- Assumptions:
 - Reads are uniformly sampled
 - Fixed error rate and i.i.d. errors
 - k-mers have only one possible genomic origin
- Then the frequencies F of random correct/erroneous k-mers can be modeled as Binomial distributions
- Drawbacks of theoretical model:
 - Error rate and genome length must be known
 - Doesn't model repeats or coverage biases



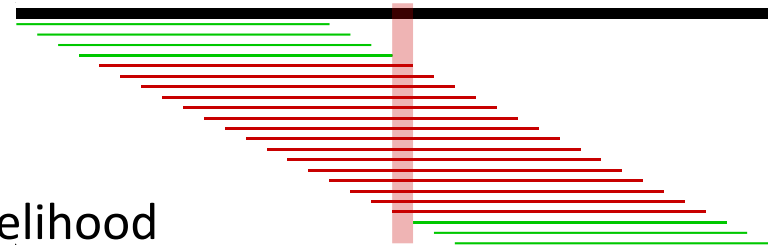
Real k-mer Frequencies

- EULER-SR and Quake use a mixture model:
 - Gaussian/Zeta mixture for **correct** k-mers
 - Poisson for **erroneous** k-mers
- Estimation of model parameters:
 - Count read k-mers
 - Fit model by maximizing the likelihood



Replace weak k-mers in each read

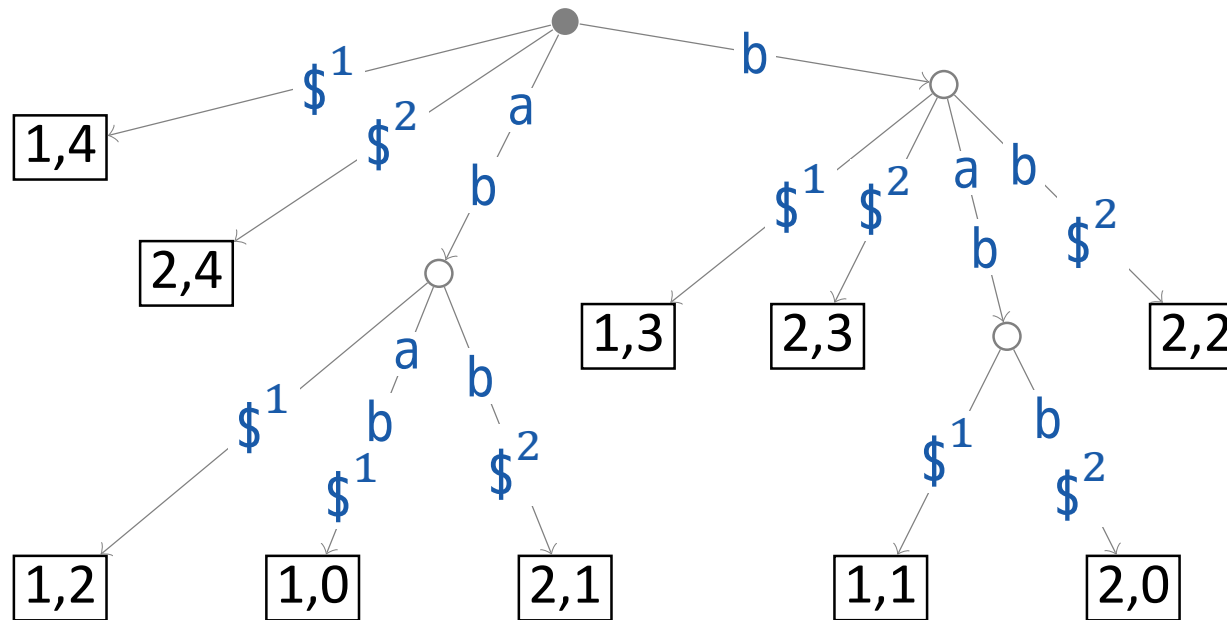
- Optimal solution was devised by Chaisson *et al.* (2004)
- Heuristic 1 (EULER-SR):
 - Enumerate all possible 1-base-corrections
 - Greedily choose the one with maximal effect
- Heuristic 2 (Quake):
 - Localize segments not covered by solid k-mers
 - Try sets of corrections in the order of likelihood



SUFFIX TREE APPROACH

Generalized Suffix Tree

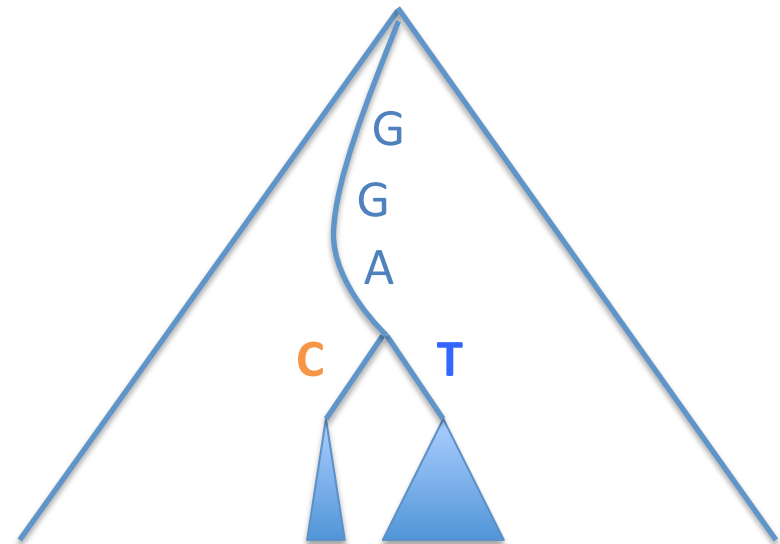
- A suffix tree generalized to multiple sequences
 - Example: $S = \{abab, babb\}$



Suffix Tree Approach

- Rationale
 - Approximate MSA on-the-fly with variable-length anchors
 - Consider k-mers where the error occurs at the end
 - Erroneous k-mer and correct counterpart share a common (k-1)-prefix
 - Corresponding suffix tree node is branching into **correct** (thick) and **erroneous** (thin) subtrees

CCGTCGGATGC
CGTCGGATGCA
GTCGGATGCAA
GTCGGATGCAA
TCGGATGCAAG
GGATGCAAGCT
CGGACGCAAGC



Algorithm Outline

1. Construct suffix tree of reads and their reverse complements
 2. Search solid-weak branches at depths $k=k_{\min}, \dots, k_{\max}$
 3. Search the weak subtree in the solid subtree skipping the first base
 4. Correct mismatches with the skipped solid base
- Extension to indels (Hybrid SHREC):
 - Single-base indels can be corrected by skipping a base in only one subtree

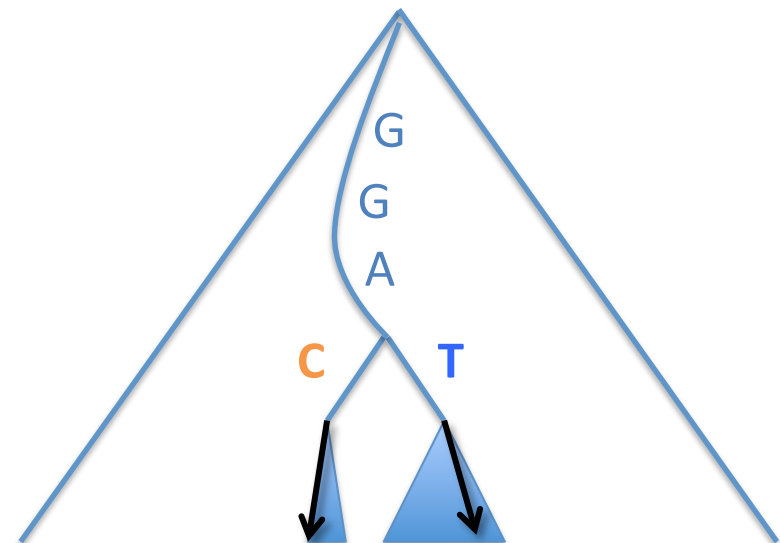
k-mer Frequency Threshold

- How to distinguish weak from solid k-mers?
- Variant 1 (HiTEC):
 - Two Binomial distributions (explained before)
- Variant 2 (SHREC/Hybrid SHREC):
 - Use Z-score cutoff of only one (the correct) Binomial distribution:
 - Errors are not modeled

Correction of Substitutions

- Search branching nodes with weak/solid children
- For each suffix in weak subtree:
 - Skip one character (**C**) and search remainder in solid subtree (skip **T**)
 - If search ends in a leaf:
 - Correct the skipped character using solid read (**C** → **T**)

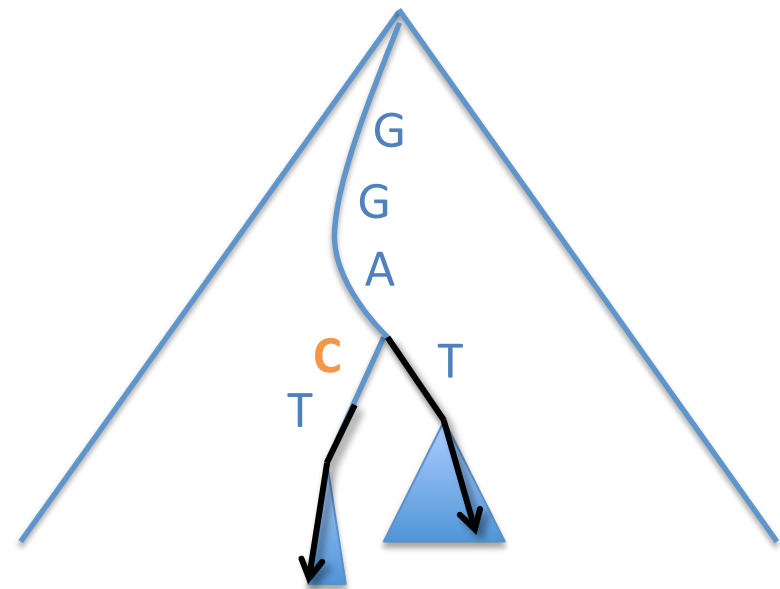
CCGTCGGATGC
CGTCGGATGCA
GTCGGATGCAA
GTCGGATGCAA
TCGGATGCAAG
GGATGCAAGCT
CGGATGCAAGC



Correction of Single-Base Insertion

- Search branching nodes with weak/solid children
- For each suffix in weak subtree:
 - Skip one character (C) and search remainder in solid subtree
 - If search ends in a leaf:
 - Delete skipped character

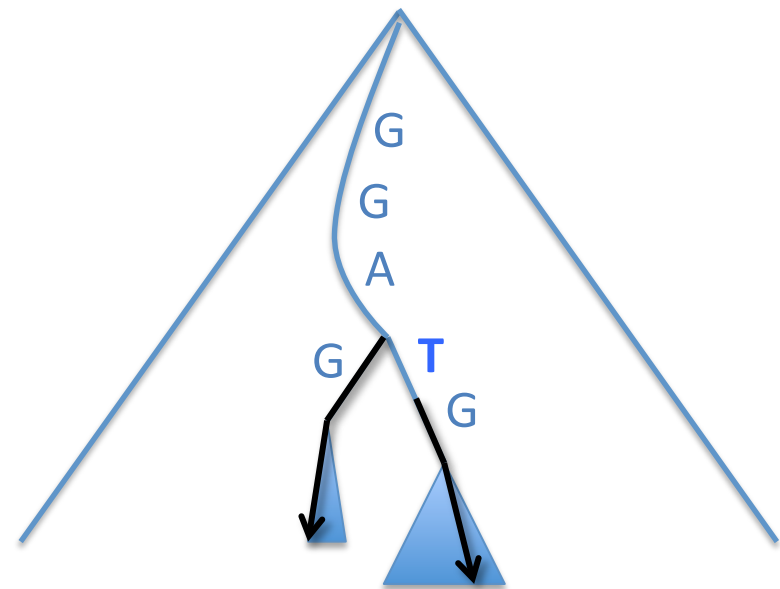
CCGTCGGATGC
CGTCGGATGCA
GTCGGATGCAA
GTCGGATGCAA
TCGGATGCAAG
GGATGCAAGCT
CGGATGCAAGC



Correction of Single-Base Deletion

- Search branching nodes with weak/solid children
- For each suffix in weak subtree:
 - Search remainder in solid subtree (skip first solid character, e.g. **T**)
 - If search ends in a leaf:
 - Insert skipped solid character

CCGTCGGATGC
CGTCGGATGCA
GTCGGATGCAA
GTCGGATGCAA
TCGGATGCAAG
GGATGCAAGCT
CGGATGCAAGC



RESULTS

Benchmark

- Yang *et al.* (2012) evaluated the performance of EC tools:
 - Real reads with known reference genome
 - Align reads to reference
 - Consider only uniquely mappable reads without Ns
 - After correction: realign corrected reads to origin
 - Alignment errors are assumed to be sequencing errors
- Definitions
 - TP: error base ► correct base
 - FN: error base ► error base
 - TN: correct base ► correct base
 - FP: correct base ► error base
 - $\text{gain} = (\text{TP} - \text{FP}) / (\text{TP} + \text{FN})$
= (removed - introduced errors) / (errors before)

Results - Illumina

Dataset	Approach	Method	Specificity	Sensitivity	Gain	Runtime (m)	Memory (GB)	Error rate after (%)
E. coli (72x)	MSA	Coral	0.9999	0.0003	0.0029	8.32	7.7	1.94
47 bp		ECHO	0.9999	0.9091	0.9076	304.21	16.0	0.18
1.54% e-rate	k-Spec	Quake	0.9990	0.3648	0.3134	80.43	2.0	1.34
		Reptile	0.9999	0.8535	0.8521	71.64	3.4	0.29
	SufTree	HSHREC	0.9866	0.6355	-0.3166	74.13	12.6	1.83
		HiTEC	0.9997	0.9466	0.9291	59.19	6.2	0.14
E. coli (574x)	MSA	Coral	0.9999	0.1134	0.1127	450.29	30.0	0.90
100 bp		ECHO	-	-	-	-	-	-
1.01% e-rate	k-Spec	Quake	-	-	-	-	-	-
		Reptile	0.9999	0.9137	0.9101	225.43	19.1	0.09
	SufTree	HSHREC	0.9930	0.2852	-1.6063	779.15	29.9	0.96
		HiTEC	0.9999	0.9557	0.9498	683.87	9.8	0.05

Results - 454 and IonTorrent

Dataset	Approach	Method	Specificity	Sensitivity	Gain	Runtime (m)	Memory (GB)	Error rate after (%)
E. coli (12x)	MSA	Coral	0.9983	0.9432	0.4252	2.57	2.4	0.19
37-385 bp [454]	Suftree	HSHREC	0.9979	0.8730	0.2242	16.72	9.9	0.25
0.26% e-rate								
E. coli (8x)	MSA	Coral	0.9982	0.7706	0.6047	1.13	2.24	0.43
16-107 bp [IonTorrent]	SufTree	HSHREC	0.9968	0.7763	0.4502	8.72	10.21	0.53
1.28% e-rate								

Results – Assembly/SNPs

- EC improves genome assembly
 - Less and longer contigs (Kelley, 2010; Kao, 2011)
 - 30% less memory/time for assembly (Salmela, 2011)
- EC improves SNP calling
 - ECHO's SNP detection has precision/recall values of about 90%
 - SNP model further improves error correction by 20-80% (Kao, 2011)

Conclusions

- Performance
 - Most tools correct more errors than they introduce
 - Primarily designed for and tested on bacterial data
 - Only k-spectrum approaches scale to large genomes
 - HiTEC is most accurate tool on Illumina data
 - Coral best suited for indel data
- Supported sequencing technologies
 - Many tools designed for Illumina reads
 - SOLiD reads are only supported by Hybrid SHREC
 - Indels only supported by Coral and Hybrid SHREC

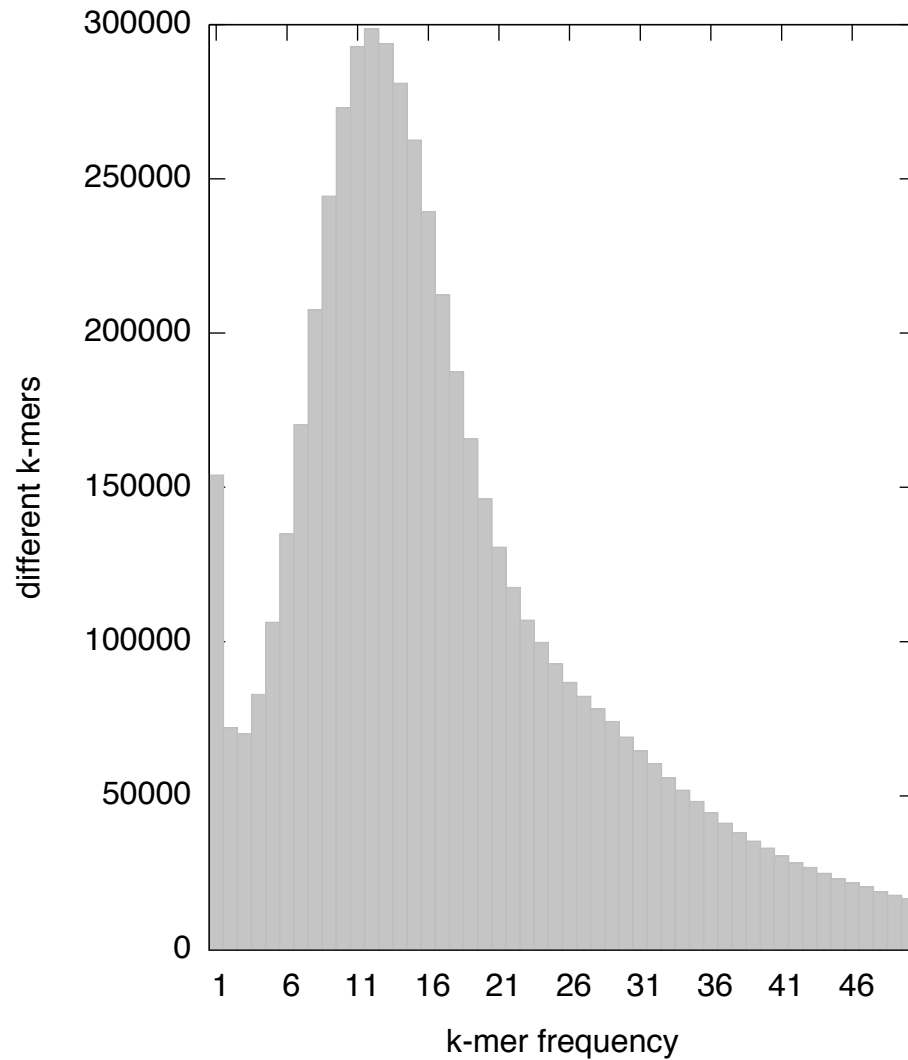
Conclusions (II)

- SNPs
 - Not considered by most of the tools
 - Only modelled by ECHO
- Repeats
 - Not considered by most of the tools
 - Only modelled by REDEEM to estimate true counts (Yang *et al.*, 2011)
 - Only SEECER (Le *et al.*, 2013) uses DNPs (Tammi *et al.*, 2002) for RNA-Seq error correction

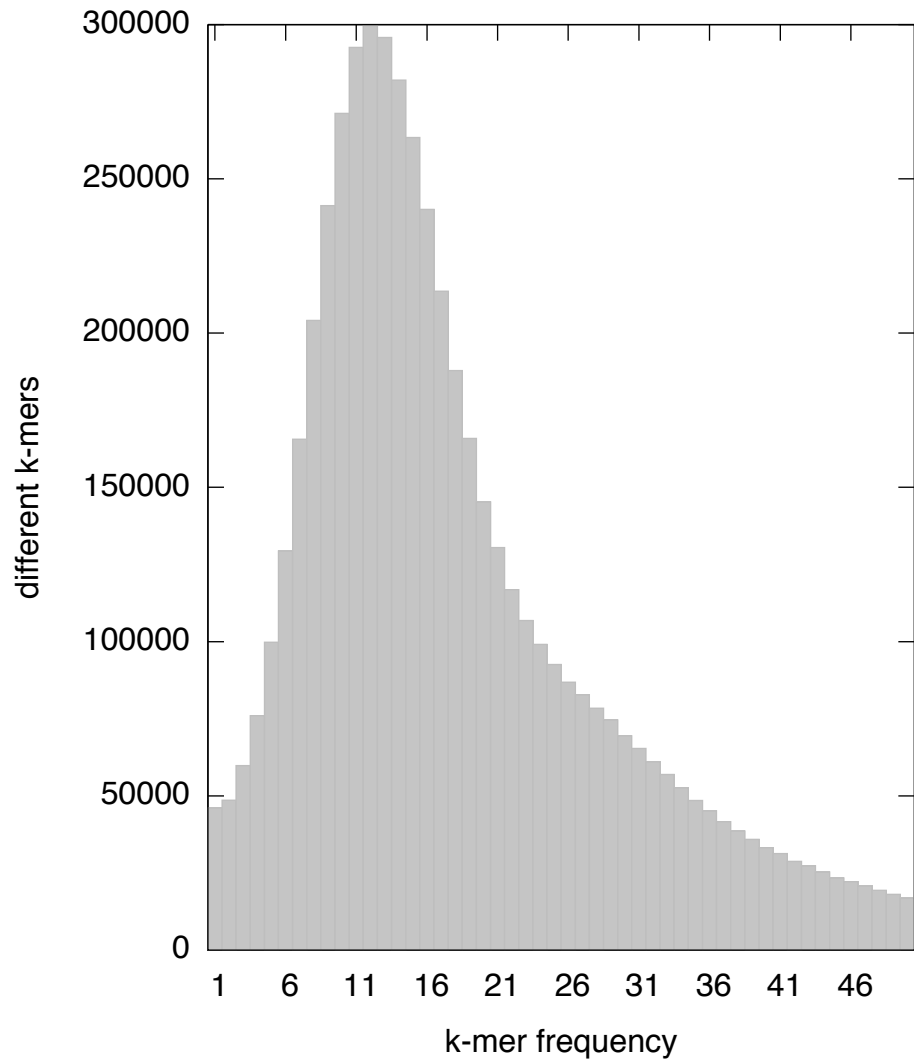
THANKS

Example: k-mer Histograms

original (ERR022075 30x)

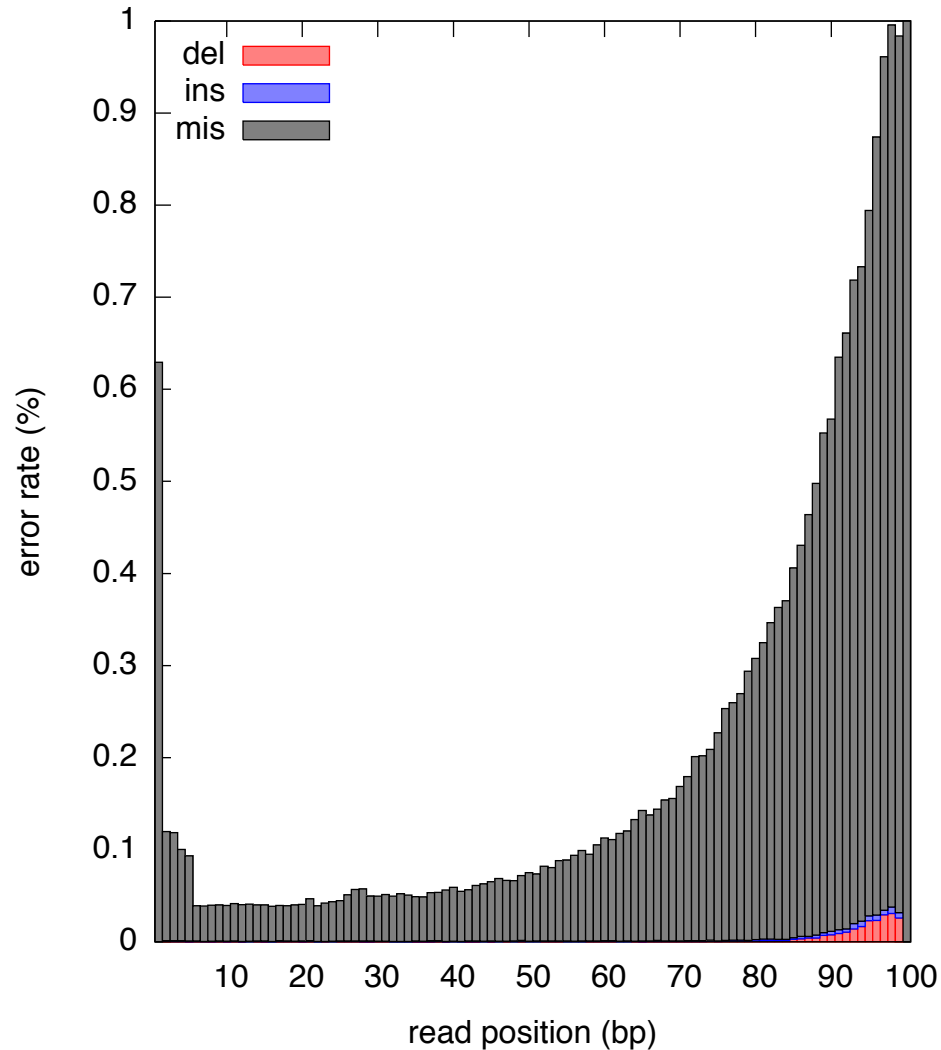


corrected (CORAL)

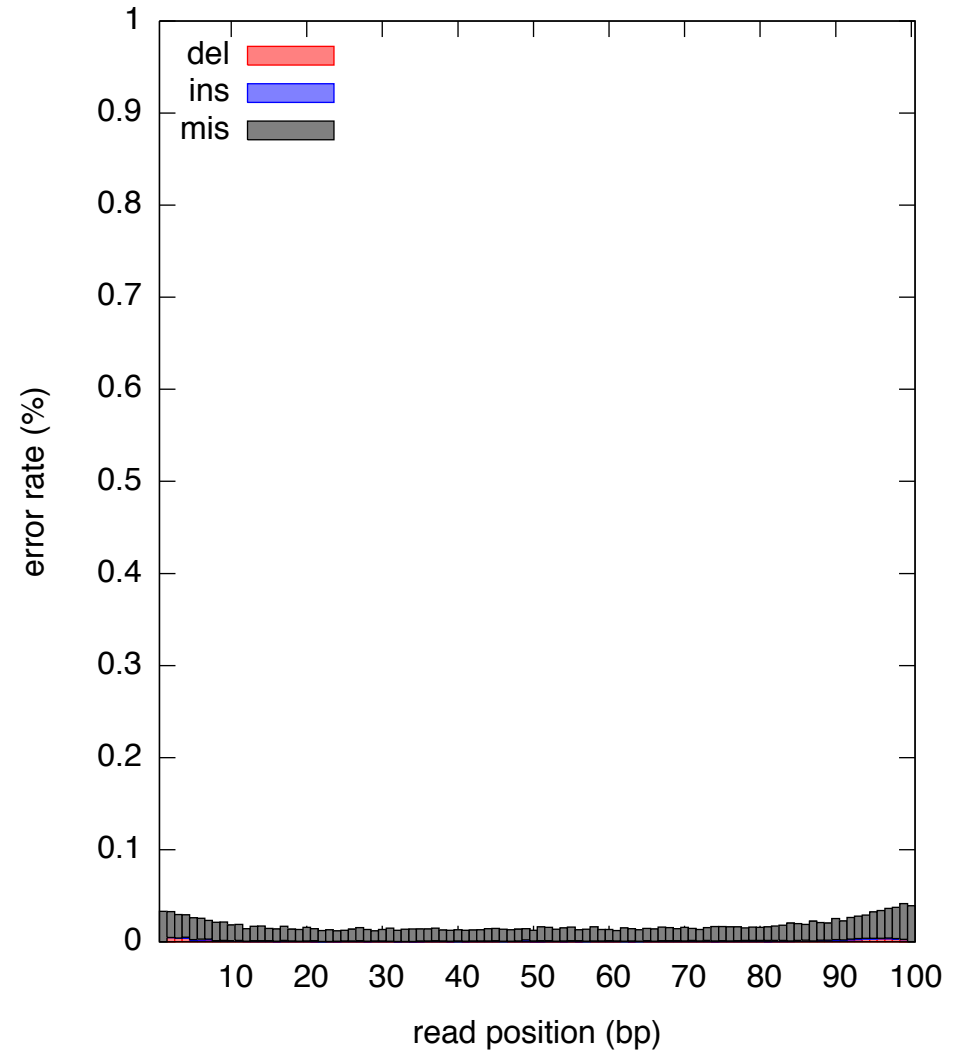


Example: Error Distributions

original (ERR022075 30x)



corrected (CORAL)



THANKS

References

1. Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.*, **11**(5), pages 759–769.
2. Kodama, Y., Shumway, M., & Leinonen, R. (2012). The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**(D1), pages D54–D56.
3. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, **26**(10), 1135–1145.
4. Dohm, J., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*.
5. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., et al. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, **39**(13), e90
6. Tammi, M., Arner, E., Britton, T., & Andersson, B. (2002). Separation of nearly identical repeats in shotgun assemblies using defined nucleotide positions, DNPs. *Bioinformatics (Oxford, England)*, **18**(3), 379.
7. Tammi, M. T., Arner, E., Kindlund, E., & Andersson, B. (2003). Correcting errors in shotgun sequences. *Nucleic acids research*, **31**(15), 4663–4672.
8. Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Research*, **19**(2), 336–346.
9. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci*, **108**(4), 1513–1518.
10. Shi, H., Schmidt, B., Liu, W., & Müller-Wittig, W. (2010). Quality-score guided error correction for short-read sequencing data using CUDA. *Procedia Computer Science*, **1**(1), 1123–1132.
11. Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, **11**(11), R116.
12. Yang, X., Dorman, K., & Aluru, S. (2010). Reptile: Representative Tiling for Short Read Error Correction. *Bioinformatics (Oxford, England)*.

References (II)

13. Schröder, J., Schröder, H., Puglisi, S. J., Sinha, R., & Schmidt, B. (2009). SHREC: a short-read error correction method. *Bioinformatics (Oxford, England)*, 25(17), 2157–2163.
14. Salmela, L. L. (2010). Correction of sequencing errors in a mixed set of reads. *Audio, Transactions of the IRE Professional Group on*, 26(10), 1284–1290.
15. Ilie, L., Fazayeli, F., & Ilie, S. (2011). HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 27(3), 295–302.
16. Salmela, L., & Schröder, J. (2011). Correcting errors in short reads by multiple alignments. *Bioinformatics (Oxford, England)*, 27(11), 1455–1461.
17. Kao, W. C., Chan, A. H., & Song, Y. S. (2011). ECHO: a reference-free short-read error correction algorithm. *Genes & Development*, 21(7), 1181–1192.
18. MacManes, M. D., & Eisen, M. B. (2013). Improving transcriptome assembly through error correction of high-throughput sequence reads. *arXiv.org*.
19. Yang, X., Chockalingam, S. P., & Aluru, S. (2012). A survey of error-correction methods for next-generation sequencing. *Briefings in Bioinformatics*.
20. Chin, F. Y. L., Leung, H. C. M., Li, W.-L., & Yiu, S.-M. (2009). Finding optimal threshold for correction error reads in DNA assembling. *BMC bioinformatics*, 10 Suppl 1, S15.
21. Yang, X., Aluru, S., & Dorman, K. S. (2011). Repeat-aware modeling and correction of short read errors. *BMC bioinformatics*, 12(Suppl 1), S52.
22. Medvedev, P., Scott, E., Kakaradov, B., & Pevzner, P. (2011). Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics (Oxford, England)*, 27(13), i137–41.