

CONTENTS

1	Introduction	1
1.1	Motivation	1
1.2	Objectives	3
1.3	Contributions	4
1.3.1	Published Contributions	6
1.4	Document Organization	6
2	State of the Art	9
2.1	Motivation	10
2.1.1	Sequencing Technologies	11
2.1.1.1	Illumina/Solexa	15
2.1.1.2	Roche 454	16
2.1.1.3	Ion Torrent/PGM	17
2.1.1.4	Abi SOLiD	17
2.1.1.5	Pacific Biosciences	18
2.1.1.6	Oxford Nanopore	18
2.1.2	Errors in NGS	19
2.1.2.1	GC Content	22
2.1.3	Benefits of Error Correction	24

3	Error Correction	26
3.1	Approach	26
3.1.1	Conditions	27
3.2	Technology support	28
3.3	Software Categories	31
3.3.1	K-Spectrum Based (ksb)	31
3.3.2	Suffix Trie/Array Based (stab)	40
3.3.3	Multiple Sequence Alignment Based (msab)	44
3.3.4	Read Cluster Based (rcb)	49
3.3.5	Probabilistic Models Based (pmb)	51
3.3.6	Recommendations	53
3.4	Discussion	57
3.4.1	Challenges	58
3.4.1.1	Data Preparation and Post-processing Steps	58
3.4.1.2	K-mer	58
3.4.1.3	Repetitive Regions	61
3.4.1.4	Ploidy	63
3.4.1.5	Read Trimming and Splitting	63
3.4.1.6	Unknown/Uncalled Bases	64
3.4.1.7	Low-Coverage Regions and Uniformity	65
3.4.1.8	Parameters	66
3.4.1.9	Single Threaded vs Parallel	67
3.4.1.10	Operating System and Programming Language	68
3.4.1.11	License and availability:	69
3.4.1.12	Recommendations	69
3.5	Testing	70
3.5.1	Methods	71
3.5.2	Gain/Specificity/Sensitivity	71
3.5.3	Assembly	74
3.5.4	Genomes Used for Testing	75

3.5.5	Real vs. Artificial Datasets	75
3.5.6	Resource Consumption	76
3.5.7	Testing details	77
3.5.8	Recommendations	78
4	MuffinEC - Error Corrector	79
4.1	Materials and Methods	81
4.1.1	k-mers Count and Histogram	83
4.1.2	Initial Reads Grouping	84
4.1.3	Greedy Grouping	84
4.1.4	Group Refining	87
4.1.5	Error Correction	88
4.2	Calculations	91
4.2.1	Implementation	91
4.2.2	Parameters	92
4.3	Results and Discussion	97
4.3.1	Testing Methodology	98
4.3.1.1	Resource Consumption Testing	102
4.3.1.2	Scalability	107
4.3.1.3	Profiling	108
4.3.1.4	Parameter Robustness	108
4.3.2	Short Aligning Results	113
4.3.3	Assembly Results	113
4.3.4	Unknown Bases	115
4.3.5	Resource Demands	116
4.4	Discussion and Conclusion	119
5	MuffinInfo - NGS Information Extractor	121
5.1	Methods	125
5.2	Extensibility	129
5.3	Results	131
5.4	Conclusion	132

6	Conclusion	134
6.1	Published results	137
6.2	Software	138
A	Error correction in real projects	139
A.1	Recommendations	143
B	External Testing of Error Correctors	145
C	Testing Methods for Correctors	150
D	Correctors Performance	156