Final Project: Computational Molecular Biology BIOC218
Christina Bormann Chung, SCPD student
March 16, 2009

Mapping Short Sequencing Reads: Challenges & Solutions

Summary

Next-generation sequencing technologies, which were introduced in the last few years, have revolutionized the sequencing landscape. Those new technologies produce up to 30 GB of data through massively parallel sequencing within a few days and for a fraction of the cost compared to conventional sequencing methods. However, those next-generation sequencing reads are also characterized through much shorter sequences and higher sequencing errors, thus posing new challenges to Bioinformaticians. In this review, alignment programs recently developed to map short sequencing reads from next-generation sequencing technologies to a reference genome are introduced.

Introduction: Next Generation Sequencing Technologies

The sequencing landscape has been shaped for decades through technologies developed by Maxam & Gilbert [1] and Sanger et al. [2] in 1977. Especially the later method has been the most commonly used sequencing technology and was used to sequence the first human genome through initiatives led by the International Human Genome Sequencing Consortium and Celera Genomics [3, 4]. Recent technological advances introduced massively parallel or next-generation sequencing technologies. These new technologies combine higher throughput with lower costs compared to the older sequencing technologies. Commercially available next-generation sequencers are the GS FLX (454 Life Sciences), Genome Analyzer (Illumina), and SOLiD (Applied Biosystems) (Table 1).

Next-Generation Sequencer	Library Type	No. of hours of days/run	Mappable output/run[1]	Read length[2]	Accuracy
GS FLX Titanium (454 Life Sciences)	Single read	10 hours	0.4-0.6 GB	400 bp average	99%
GS FLX Standard (454 Life Sciences)	Mate-pair	7.5 hours	0.1 GB	250 bp average	99.5%
Genome Analyzer	Single read	2 days	2-3 GB	35 bp	>99%
(Illumina)	Mate-pair	5.5 days	8-10 GB	2x50 bp	>98.5%
SOLiD	Single read	6-7 days	10-15 GB	50 bp	99.94%
(Applied Biosystems)	Mate-pair	12-14 days	20-30 GB	2x50 bp	99.94%

Table 1. Comparison of commercially available Next-Generation Sequencers. (Data obtained from each company's web-site: www.454.com, www.illumina.com, and www.appliedbiosystems.com).

[1] GS FLX instruments and Genome Ananlyzer support 1 flowcell/run, SOLiD supports 2 flowcells/run
[2] Maximun read lengths currently supported by respective company according to website.

GS FLX – 454 Life Sciences

The GS FLX, developed by 454 Life Sciences, was the first commercially released next-generation sequencing instrument. The DNA is prepared for sequencing by random fragmentation, ligation of sequencing specific adaptors to these fragments, capturing of single fragments on beads, and amplifying those bead-fragments through emulsion PCR. The DNA on the beads is denatured and each bead is then deposited in a single picoliter reaction well on a fibre-optic slide [5].
The sequencing chemistry is a pyrosequencing or sequencing-by-synthesis approach: One kind of dNTP is sequentially added to the flowcell. During the incorporation of the nucleotide(s) to the template through polymerase activity, inorganic pyrophosphate is released, which is then enzymatically converted into ATP. Upon contact with the enzyme luciferase, light is produced, which is measured by a camera. The intensity of the light produced, is proportional to the number of nucleotides incorporated (Figure 1) [5-7].

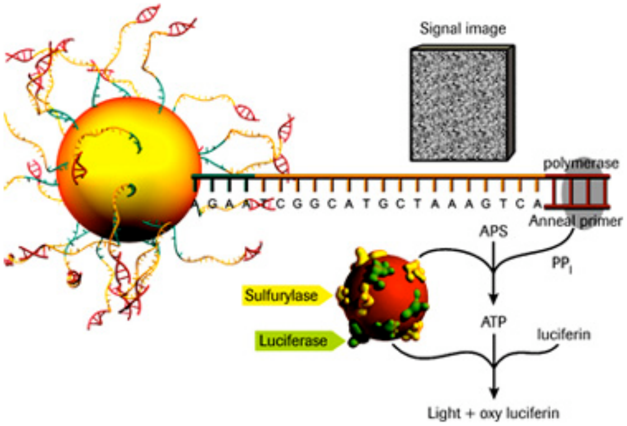

Figure 1. Sequencing-by-synthesis with the GS FLX (454 Life Sciences): During the incorporation of a nucleotide though polymerase activity, PPi is released and converted into ATP. Upon contact with the enzyme luciferase, light is released, which is recorded by a camera.
(source: 454 Life Sciences: http://www.454.com/products-solutions/how-it-works/sequencing-chemistry.asp)

Genome Analyzer (Illumina)

The Genome Analyzer developed by Illumina is also a sequencing-by-synthesis technology. To prepare DNA for sequencing, it is also fragmented and sequence specific adaptors are ligated to the fragments. These fragments are then randomly bound to a glass-slide surface and PCR amplified into dense clusters through solid-phase bridge amplification [6, 7].
In the first sequencing round, primers, fluorescently labeled reversible dNTP terminators, and DNA polymerase are added to the flowcell. The corresponding nucleotide is incorporated into the first position and the synthesis reaction is automatically stopped through a 3’ terminating group attached to the dNTP. Through laser excitation the signal of the fluorescent label is recorded by a camera. Each nucleotide carries one of four fluorescent labels, thus enabling the identification of each nucleotide. For the second (and each consecutive) sequencing round, the fluorescent label and terminating group are

removed from the nucleotide. This allows further extention of the DNA strand by adding another round of labeled dNTPs and DNA polymerase (Figure 2) [6-7].

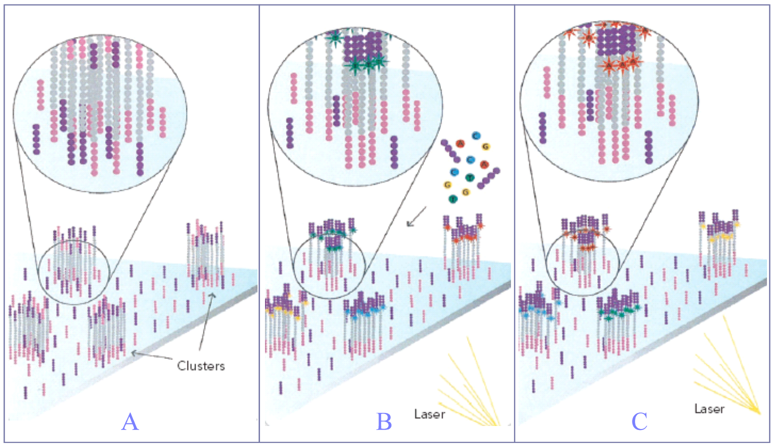

Figure 2. Sequencing-by-synthesis with the Genome Analyzer (Illumina/Solexa): PCR clusters originating from a single fragment are sequenced by the addition of primers, fluorescently labeled reversible dNTP terminators, and DNA polymerase. The fluorescent label is excited through a laser and the released signal is recorded by a camera. After removing the fluorescent label and the terminal group another round of fluorescent labeled dNTPs and DNA polymerase are added to investigate the next base in the sequence.
(source: Illumina Sequencing Technology: http://www.illumina.com/downloads/SS_DNAsequencing.pdf)

SOLiD (Applied Biosystems)

The SOLiD System, developed by Applied Biosystems, is a sequencing-by-ligation technology based on the polony sequencing technique published by Shendure et al. in 2005 [8]. To prepare DNA for sequencing, it is fragmented, sequence specific adaptors are ligated, and single fragments are amplified on beads in emulsion PCR similar to the GS FLX system (454 Life Science). Beads are then randomly deposited onto a glass slide [9].

For sequencing the instrument uses a set of four fluorescently labeled probes each being comprised of eight bases. The first and second base are one of the 16 possible di-nucleotide sequences that are responsible for the specificity of hybridization. All probes are added to the flowcell simultaneously and hybridization occurs specifically based on complementary base-pairing. The probe next to the sequencing primer is then ligated to the primer through DNA ligase. Through excitation of the fluorescent label, the incorporated probes are identified and captured by a camera. The fluorescent label and last three nucleotides of the probes are then removed through cleavage and the next round of hybridization and ligation takes place, thus DNA bases are always interrogated at positions five bases apart. After several rounds of hybridization and ligation, all probes and the first sequencing primer are removed, and a new round of sequencing is started by annealing a sequencing primer one base closer to the beads. The process is repeated with five sequencing primers total, until all bases of the template have been identified. This sequencing method differs from the other two methods by interrogating each base twice and the output data differs through the sequencing in color-space instead of base-space. (Figure 3) [6, 9]

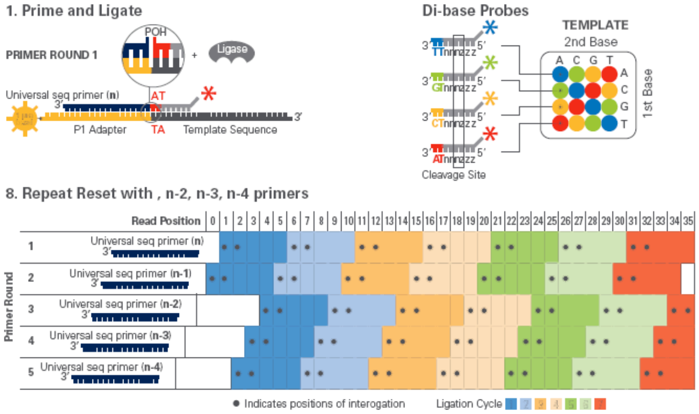

Figure 3. Sequencing-by-ligation with the SOLiD System (Applied Biosystems): A sequencing primer is annealed to the DNA fragment, followed by the hybridization and ligation of one of 16 possible di-base probes. Through excitation of the fluorescent label of the probe, one of four colors is recorded by a camera. After cleaving of the label and last three bases of the 8-base long probe, a new round of hybridization and ligation occurs, thus interrogating every fifth di-base. At the end all probes get removed and a new primer gets annealed to the sequence, starting a new sequencing round. This process is repeated for five primers total, leading to the interrogation of each base twice. (source: Applied Biosystems: SOLiD System Accuracy: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057511.pdf)

Challenges of Mapping Short Sequencing Reads

The next-generation sequencing technologies pose new challenges to the Bioinformatics community. Besides the enormous amount of data that they can produce (see Table 1) through their high-troughput, the read length of the sequence tags are also much shorter than the conventional sequencing methods. While Sanger sequencing can produce read lengths of over 700 bp [6], the GS FLX (454 Life Science) can read average lengths of up to 400 bp and the Genome Analyzer (Illumina) and SOLiD (Applied Biosystems) can read tags of up to 35 or 50 bp respectively (or 2x50 bp for mate-pair).

Mapping those short reads to a reference genome can pose different challenges in determining the nature of mismatches. A mismatch can occur due to sequencing errors (because of the instrument/chemistry nature), differences between the query sequence and the reference sequence (SNPs, Insertion/Deletions), or existence of repetitive regions. While read lengths have constantly been improved by the different technologies over the past few years after their release, there will remain a certain cut-off read length that is both practical for Bioinformatics use and economical affordable. The reason for finding a balance here is that the cost of sequencing goes up through higher read lengths due to the use of more chemical reagents. In the following section, the emphasis will be on alignment programs developed for the Genome Analyzer (Illumina) and SOLiD (Applied Biosystems), because those two technologies produce ultra-short reads and are therefore more challenging to use with older alignment programs developed for Sanger sequencing.

Solutions: Short Sequencing Reads Alignment Tools

Sequence alignment programs that have been developed for conventional sequencing methods, such as BLAST do not take into account the shorter read lengths and the higher errors/base that are the draw-backs of the next-generation sequencing methods. Therefore several Bioinformatics groups have developed new programs, which deal with these new challenges. Each program usually addresses the problem slightly different, using for example gapped versus un-gapped alignment or taking the quality value of each base into account. Therefore each program might be applicable for different applications and/or sequencing technologies. Table 2 gives an overview of different alignment programs publicly available. Only programs, which have been published in peer-reviewed journals, will be discussed in more detail.

Alignment program	**Sequencing platform[1]**	**Availability[1]**	**Website**	**Citation**
Corona-lite	SO	part of SO software	www.appliedbiosystems.com	NA
Eland	GA	part of GA software	www.illumina.com	NA
MAQ	GA, (preliminary functions for SO)	free	http://maq.sourceforge.net	[10]
MOM	GA	free	http://mom.csbc.vcu.edu/	[11]
Mosaik	GA, SO, GS	free	http://bioinformatics.bc.edu/marthlab/Mosaik	NA
PASS	GA, SO, GS	free	http://pass.cribi.unipd.it	[12]
RMAP	GA	free	http://rulai.cshl.edu/rmap	[13]
SeqMap	GA	free	http://biogibbs.stanford.edu/~jiangh/SeqMap	[14]
SHRiMP	GA, SO	free	http://compbio.cs.toronto.edu/shrimp	NA
Slider	GA	free	www.bcgsc.ca/platform/bioinfo/software/slider	[15]
SOAP	GA	free	http://soap.genomics.org.cn	[16]
SOCS	SO	free	http://socs.biology.gatech.edu	[17]
SXOligoSearch	GA	for purchase	http://www.synamatix.com/secondGenSoftware.html	NA
ZOOM	GA, SO	free	www.bioinfor.com/zoom	[18]

Table 2. List of alignment programs available for short reads produced from next-generation sequencing technologies.

[1] GA = Genome Analyzer (Illumina/Solexa), SO = SOLiD (Applied Biosystems), GS = GS FLX (454 Life Sciences)

MAQ – Mapping and Assembly with Quality [10]

MAQ is an alignment program originally designed for reads produced by the Genome Analyzer (Illumina). According to the programs website (see Table 2), preliminary functions for SOLiD reads are now available as well. The developers introduce with this program the concept of "mapping quality", i.e. a measure of confidence that a read is actually mapped to the right position on a reference genome. MAQ uses exclusively mate-pair reads.

During alignment, MAQ first identifies un-gapped matches with the lowest mismatch score. In order to align faster, MAQ by default only considers positions that have two or fewer mismatches in the first 28 bp. In a second round of alignment, sequences that failed to reach a mismatch score threshold, but whose mate pair was mapped, are mapped using a gapped alignment algorithm in the regions defined by its mate. Each alignment receives a *phred*-scaled quality score, which measure the probability of false alignment.

MAQ only reports unique alignments. However, if a read maps equally well to multiple positions, the program randomly picks one position and assigns a mapping quality of zero. Those reads will not contribute to variant calling, but they will give valuable information about copy number of repetitive sequences and on the fraction of reads that can be aligned to the reference sequence.

MOM – Maximum Oligonucleotide Mapping [11]

MOM is a seed based search tool designed to match a maximal length of a short read towards a reference genome. It can be used with data from the Genome Analyzer (Illumina).

First the program creates a hash table of sub-sequences (fixed length) of either the query or the reference sequence(s). It then sequentially matches the un-indexed sequence(s) to the sub-sequences in the table. If a match is found, its location is passed on to the sequence extension part of the program. In order to find the longest match, MOM starts matching the query sequence from the first base (5' end) to the reference (regardless of the seed position). Extension is carried out in the 3' direction until the end of the read or the maximum number of allowable mismatches is reached. If the later occurs, the program trims bases from the 5' end to the first mismatching base and extends the alignment on the 3'end. This will be repeated until the end of the read is reached. If the longest alignment achieved meets the minimum specified match length, the alignment is considered valid.

Multiple matches of the same read are considered equivalent, if the same length and number of mismatches are achieved. For each query, the number of best matches will be reported in the output.

According to data published by the developers, MOM mapped in three out of four experiments more reads to the reference sequence compared to SOAP (average 4.7% more), and it also mapped better than MAQ (2/2 experiments, average 9.95% more) and SHRiMP (1/1 experiment, 23% more).

PASS – Program to Align Short Sequences [12]

PASS is an alignment program designed to improve execution time and sensitivity compared to other short read alignment tools. Data sets generated by all three next-generation sequencing systems can be aligned with this program.

The reference sequence is first indexed into gapped seed words (12 bp, two gaps). Afterwards each query read is matched to the seed words. If a successful match was found, PASS tests the possibility of extending the alignment. For this step a pre-computed table of all possible short words aligned against each other is used, which is located in RAM, thus allowing a quick analysis of the adjacent regions. Several pre-computed tables using the Needleman and Wunsch algorithm are provided with the PASS software download. Besides extending the alignment, this step is also used to discard seed words, which cannot produce a proper alignment.

RMAP [13]

The alignment program RMAP can map reads with a variety of lengths and utilizes base-call quality scores to achieve higher mapping accuracy. RMAP was designed to handle sequencing reads produced by the Genome Analyzer (Illumina).

In order to map the reads with length *n* to a reference sequence with *k* mismatches, each read is split into *k*+1 contiguous seeds. This way any read mapping to the reference sequence must contain at least one seed with no mismatches. The matching algorithm therefore first identifies which seed maps without mismatch to the reference sequence. After that the remaining seeds are mapped to the surrounding regions of the no-mismatch seed.

RMAP also measures mapping accuracy depending on two mapping criteria. The first criterion is a simple count of mismatches between the query and reference sequence. The second criterion makes use of the base-quality scores at mismatch positions. A user-defined cutoff for those scores differentiates between high-quality (HQ) or low-quality (LQ) scores. LQ positions always induce a match, i.e. they serve as wild cards. A quality control step, however, eliminates reads with too many LQ positions, in order to prevent inaccurate matching of reads.

SeqMap [14]

SeqMap can map Genome Analyzer (Illumina) reads to a reference sequence allowing up to five mixed mismatches and inserted/deleted nucleotides. Furthermore, several command line options are available that allow fine tuning of different mapping aspects, such as using unequal read lengths.

The mapping algorithm is similar to the RMAP mapping program. Each read is split into several parts. For example, for a mapping with 2 mismatches, the read will be split into four parts. In a first round of matching, at least two of the reads have to map perfectly. Those sequences are used to key all the candidates in a hash table. In a second part, the read to be mapped is matched against all candidates to determine all target regions within the reference sequence. In this step an insertion/deletion instead of a mismatch can be searched as well, by shifting the read parts to be mapped one nucleotide to the 5' or 3' end.

The developers published a comparison of SeqMap to ELAND, SOAP, and RMAP. A randomly generated DNA sequence (1 Mb) with 100 kb of added random SNPs and Insertion/Delitions was randomly sampled for 100,000 short reads (30 bp). Those reads were then analyzed using the four programs. SeqMap was able to map most reads (78.2% in 82s) compared to SOAP (38.3% in 2s), RMAP (31.9% in 4s), and ELAND (27.6% in 3s). However, another mapping experiment, where >11M sequencing reads were mapped to mouse chrX, little difference was found in the number of mapped reads of those four programs. These results implicate that SeqMap only gives an advantage for sequencing data that differs a lot (more than average) from the reference genome.

Slider [15]

Slider is designed for use with sequencing data from the Genome Analyzer (Illumina) and utilizes each base's probability given in the *prb* output file of the instrument. Both single-end and mate-pair end sequencing data can be used with Slider.

During the Genome Analyzer sequencing process, for each base of each cluster the four color signal intensities are compared, and probabilities corresponding to the most likely base incorporation are logged in a *prb* file. A final read sequence according to those probabilities is recorded in a *seq* file. SeqMap uses the *prb* instead of the *seq* file as an input for alignment. The program then not only uses the most probable base, but also all possible bases with a probability above a certain threshold. Slider then sorts all those generated reads in lexicographical order and crosses it sequentially with a presorted table of reference sequence(s) and their reverse complement(s). Using the probability scores instead of the sequencing file, Slider reduces mismatches due to sequencing errors and thus results in more accurate SNP prediction.

The developers compare alignment of Genome Analyzer reads with ELAND, RMAP, and Slider. They show that the percent mismatches within unique alignments are reduced using Slider. For example for mapping 32 bp reads, Slider mapped 80.2% of reads uniquely to the reference sequence with 1.2% mismatches, while RMAP and ELAND mapped 79.5% uniquely with 3.0% mismatches. Furthermore, they report that Slider mapped the reads in 2/3 of the time than ELAND, and that RMAP maps even slower than ELAND.

SOAP – Short Oligonucleotide Alignment Program [16]

SOAP is designed for short sequencing reads from the Genome Analyzer (Illumina). It can perform gapped or un-gapped alignments towards a reference sequence. Furthermore, it has special modules for alignment of mate-pair reads, small RNA, and mRNA sequencing data.

The program allows either a certain number of mismatches or one continuous gap (1-3bp in either the query or reference sequence) within one read. The best hit for each read with a minimal number of mismatches or a smaller gap will be reported. If the read does not match, SOAP can iteratively trim several base pairs at the 3'-end and redo the alignment. This will be repeated until the read is mapped or the remaining sequence is too short. For faster alignment SOAP uses seed and hash tables, similar to RMAP and SeqMap. For an alignment with two mismatches the read will be split up into four seeds.

The developers compare the performance of SOAP towards MAQ and ELAND with a data set produced on a Genome Analyzer (Illumina). Using the gapped alignment option together with the iterative trimming of the read, SOAP could map 91.2% of the reads to a reference genome, while ELAND and MAQ mapped 88.5% and 88.4% respectively.

SOCS – Short Oligonucleotides in Color Space [17]

SOCS is the only alignment program published in a peer-reviewed journal that was specifically designed to map SOLiD (Applied Biosystems) reads in color space instead of base space. Mapping those reads directly in color space should be the method of choice, since a sequencing error could alter the whole following sequence if translated into base space.
The alignment program, similar to RMAP, SeqMap, and SOAP, uses seeds and hash tables to accelerate the process of matching reads to the reference sequence. In order to match a read with *n* mismatches, SOCS divides reads in *n*+1 seeds. This ensures that at least one seed is matched perfectly to the reference sequence. The mismatch tolerance can be specified by the user, with higher mismatches resulting in more usable data, but longer run times. To reduce the run time, SOCS maps at lower mismatches first, reducing the data that needs to be mapped at higher mismatch number(s). If a read maps to more than one genomic region, quality scores and mismatch counts are used to determine the best match.

ZOOM – Zillions Of Oligos Mapped [18]

ZOOM is a mapping program that uses spaced seeds to align Genome Analyzer (Illumina) and according to the developer's website (see Table 2) SOLiD (Applied Biosystems) sequencing data.
A spaced seed can be described by a binary string such as 111010010100110111, where '1' requires a match at this position of the read and '0' describes a "don't care" position. The length of the seed is identical to the length of the read and the weight of the seed is the number of 1s in the string. The developers of ZOOM designed a minimum set of spaced seeds to achieve full sensitivity for a certain read length *m* with *k* mismatches. For this they have two design criteria: (1) The seed weight should not be too small in order to avoid too many false positives, thus slowing down the mapping process. (2) The seed weight should not be to large, because more seeds would be necessary to obtain full sensitivity, thus slowing down the mapping process as well. (For more information about the theory of spaced seeds and how the developers designed them, please refer to [19]).
ZOOM builds a hash table of the reads using the spaced seeds designed. For a specific seed, the reads sharing the same letters at the 1-positions are grouped into the same entry of the hash table. For example ACGTACGTAC and TGATAAGCGT will be grouped into the same hash entry for seed 001101000. ZOOM then scans the reference sequence and finds matching positions for each hash entry. These candidates are then further verified.
ZOOM furthermore has extended models to deal with certain applications: ZOOM-C (Mapping with confidence scores), ZOOM-I (Mapping allowing insertions and deletions, and ZOOM-P (mapping with paired-end information).
The developers published data, showing that ZOOM performs mapping much faster than RMAP and somewhat faster than Mosaik and ELAND. This mapping is done at same accuracy compared to those three programs according to the publishers, however data is only presented for coverage across the reference sequence in comparison with ELAND.

Discussion

Since the first introduction of next-generation sequencing technologies in 2005 with the GS FLX (454-Life Sciences) and more importantly in 2006 and 2007 the introduction of the two platforms with ultra-short reads, the Genome Analyzer (Illumina/Solexa) and SOLiD (Applied Biosystems), the Bioinformatics landscape was forced to develop new alignment programs for those instruments producing millions of short reads. Several groups have developed different algorithms to deal with the challenges of mapping those short reads to a reference sequence.

Besides shorter reads, the next-generation sequencing technologies also have higher sequencing error rates than conventional sequencing methods. The different programs developed, address this problem through different ways. While some tools trim the reads at the 3' end, which usually has higher error rates, other programs also allow trimming on the 5' end or allow gapped alignments. Some programs use the quality score of the sequencing reads in order to give more accuracy to their alignments. While some programs use a combination of several of those methods.

Another challenge for the program developers is, to find a balance between accuracy and speed of mapping, since the next-generation sequencing methods produce huge amounts of data. For this purpose some programs allow the user to specify the mapping conditions, thus giving more flexibility to the type of application used.

From the data that has been published, it looks like there is some degree of variation depending on the data-input [11]. However, from different studies, it seems that ELAND and RMAP are similar in their number of reads mapped to the reference sequence [14, 15, 18].

The developers of SOAP and MOM both published some data, showing that their programs map reads better to a reference sequence than MAQ, ELAND (comparison to SOAP only), and SHRiMP (comparison to MOM only) [11, 16]. Whereas, MOM also shows a slightly better mapping to SOAP in three out of four experiments, while SOAP mapped better in one of those experiments [11].

The program SeqMap could only map better than ELAND, SOAP, and RMAP, when an artificial data-set with a lot of introduced mismatches was used [14]. Therefore, the superior uses for this program might be limited.

Both ZOOM and Slider report faster program run times than ELAND, RMAP, and Mosaik (comparison to ZOOM only) [14, 18]. While Slider additionally, corrects the number of sequencing errors, thus bringing down the number of mismatches compared to ELAND and RMAP, and therefore enabling better SNP finding [15].

All the programs have just recently been developed and are thus often still beta-versions, which are still under development. Independent comparison with 'real' sequencing data sets need to be carried out, in order to determine the best and most effective program(s) and/or to determine which program is most suitable for certain applications. Only limited data-sets have been published so far.

Unfortunately, some alignment programs available and used extensively, such as ELAND and Corona-lite, which are distributed with the Genome Analyzer (Illumina) and SOLiD (Applied Biosystems) instruments respectively, have not been published in peer-reviewed journals and only limited information is available. Thus it is difficult to present a full comparison of all programs.

Literature

[1] Maxam A & Gilbert W (1977): A new method for sequencing DNA. Proc Natl Acad Sci USA 74 (2), 560-564.

[2] Sanger F, Nicklen S, Coulson AR (1977): DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74 (12), 5463-5467.

[3] International Human Genome Sequencing Consortium (2001): Initial sequencing and analysis of the human genome. Nature 409 (6822), 860-921.

[4] Venter JC et al. (2001): The Sequence of the human genome. Science 291 (5507), 1304-1351.

[5] Margulies M et al. (2005): Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057), 376-380.

[6] Hert DG, Fredlake CP, Barron AE (2008): Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29, 4618-4626.

[7] Morozova O & Marra MA (2008): Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255-264.

[8] Shendure J et al. (2005): Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309 (5741), 1728-1732.

[9] Applied Biosystems (2009): The SOLiD 3 System: Enabling the next generation of Science. https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=604416&tab=Literature.

[10] Li H, Ruan J, Durbin R (2008): Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858.

[11] Eaves HL & Gao Y (2009): MOM; Maximum Oligonucleotide Mapping. Bioinformatics advance access published Feb. 19, 2009.

[12] Campagna D et al. (2009): PASS: a Program to Align Short Sequences. Bioinformatics advance access published Feb. 13, 2009.

[13] Smith AD, Xuan Z, Zhang MQ (2008): Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9, 128-135.

[14] Jiang H & Wong WH (2008): SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24 (20), 2395-2396.

[15] Malhis N et al. (2009): Slider – maximum use of probability information for alignment of short reads and SNP detection. Bioinformatics 25 (1), 6-13.

[16] Li R et al. (2008): SOAP: short oligonucleotide alignment program. Bioinformatics 24 (5), 713-714.

[17] Ondov BD et al. (2008): Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics 24 (23), 2776-2777.

[18] Lin H et al. (2008): ZOOM! Zillions of oligos mapped. Bioinformatics 24 (21), 2431-2437.