rijksuniversiteit groningen

Whole genome sequencing of *Lactococcus lactis* WG2 using MinION

Master Research Project (MolGen)

September 2015 – May 2016

Peter Huizenga (s0110221)

Supervisors: Dr. Anne de Jong
Prof. Dr. Jan Kok

Index

Abstract

Lactococcus lactis is a lactic acid bacteria (LAB) used in diary industry to produce various products. Superior starter cultures create high yield of products and have high levels of resistance. These two functions are frequently located on plasmids and our strain of interest *L. lactis* WG2 contains numerous plasmids.

MinION is a new sequencing technology resulting in very long reads making complete genome assembly possible. Unlike traditional sequencers MinION does not use any nucleotide incorporation but monitors DNA running through nanopores directly.

Majority of modern sequencing technologies like Illumina produce high quality and high coverage short reads but software processing this data create many contigs but are unable to build complete genome assemblies. Using MinION long reads can solve this problem though low quality of data and high instability of software tools keep a solution out of reach.

We describe a successful DNA extraction method for *L. lactis* WG2 and two fragmentation techniques (enzymatic and mechanical shearing) to create optimal fragments for sequencing. Both methods showed low levels of stability and reproducibility.

We show that CANU and Nanopolish software produce a very poor assembly based on our own MinION data. We demonstrate that both the combination of SPAdes, SSPACE-LongRead and GapFiller software and Velvet software produce a much better assembly. We used both results for further analysis with RAST, CONTIGuator and NCBI webservers.

We illustrate that RAST proves closest neighbors are *L. lactis* strains and contigs contain 90% functional genes of reference chromosome *L. lactis* MG1363. We also show that CONTIGuator webserver reveals MinION long reads and assembly software do not improve Illumina based contigs. We demonstrate too that NCBI software tells contigs mapped to reference contain 98% of total assembly coding for 2,300 proteins and contigs representing plasmids contain 2% of assembly coding for 40 “hypothetical proteins”.

Introduction

Metabolism

Lactococcus lactis is a Gram+ bacteria belonging to the glade of lactic acid bacteria (LAB) and is commercially used in diary industry [1] to produce lactic acid, acidified milk products and (cottage) cheeses. Lactic acid is food preservative, curing agent and flavoring agent and lactic acid curdles milk, starting point for the production of cheese. Acidified milk products are sour milk products and yogurt.

L. lactis is generally considered homo-fermentative [1] and converts easy fermentable sugars (carbohydrates) into lactic acid in order to produce ATP used for growth and biosynthesis. On less favorable sugars *L. lactis* switches to mixed-acid fermentation (MAP) increasing ATP yield and producing ethanol, formate and acetate besides lactic acid (Figure 1). Lactic acid causes the pH to drop and the environment of *L. lactis* becomes acidified which stops growth.

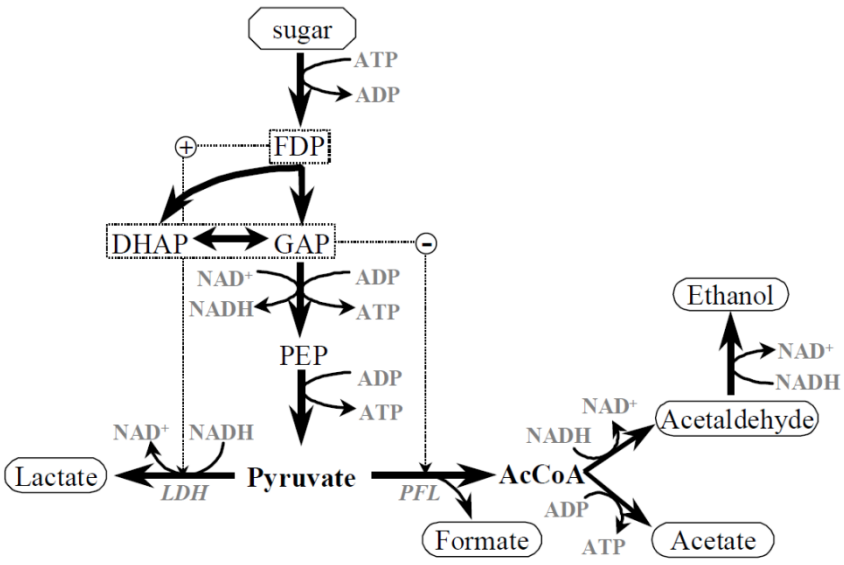

Figure 1 Fermentative metabolism of Lactococcus lactis [1]

L. lactis uses glucose as preferred sugar [2] but in case of glucose depletion *L. lactis* switches to lactose. Complete sugar depletion causes *L. lactis* to switch to arginine. Arginine [3] is an alternative source of energy (ATP) and arginine deiminase pathway (ADI) produces ammonia that neutralizes acid, consequently the environment is de-acidified and *L. lactis* starts using sugars again.

Importance of plasmids

For a long time researchers thought plasmids were selfish mobile elements but plasmids play key role in the evolution of *L. lactis* because they introduce novel properties to cells [4]. Plasmids size ranges from 3 – 130 KB and GC-content is 30 – 40 %, figures different from the (circular) chromosome and indicating that plasmids are (relatively) new to *L. lactis*. Total genome size of *L. lactis* strains is 2.5 MB (on average) and plasmids make up 10 % (maximum) of total DNA. Dairy industry uses *L. lactis* extensively and superior starter cultures are vital. High yields of products and high levels of resistance to diseases are among characteristics of good starter cultures. Genes responsible for these traits are often found on plasmids [4].

Traits linked to plasmids are:

- Lactose utilization
- Casein breakdown
- Bacteriophage resistance
- Bacteriocin production
- Metal ions
- Exopolysaccharides (EPS)

L. lactis uses different mechanisms for lactose utilization [4]. A first method uses permeases to transport lactose into the cell and enzyme β-galactosidase to convert lactose into glucose and galactose (Figure 2). Genes coding for these permeases and enzyme are located on the chromosome. A second method uses PTS (phosphotransferase system) for phosphorylation during transport and enzyme phospho-β-galactosidase for conversion of phospho-lactose into glucose and galactose-6-phosphate. All genes (*lac*ABCDFEGX) are in an operon located on a plasmid. *L. lactis* uses phosphorylation of sugars to prevent them from exiting the cell again [2]. *L. lactis* cheese starter strains with high levels of enzyme phospho-β-galactosidase (plasmid) and low levels of enzyme β-galactosidase (chromosome) are capable of efficient lactose fermentation and clotting of milk [5].

Casein (Figure 3) is a source of essential amino acids. Genes coding for proteases responsible for initial extracellular cleavage, transport proteins and peptidases involved in hydrolysis of peptides into separate amino acids are located on plasmid pHP003.

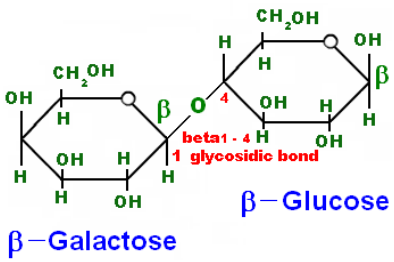

Figure 2 Lactose

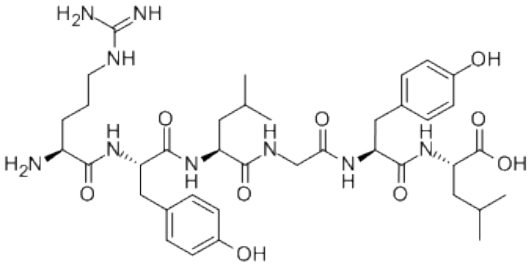

Figure 3 α-Casein (fragments 90-95)

L. lactis strains not resistant to bacteriophages are a great thread to diary industry because of negative effects on quantity, quality and stability of product yields. Adsorption inhibition and injection blocking are two mechanisms to prevent bacteriophages from entering cells. High levels of sugars (galactose and rhamnose) mask receptors causing adsorption inhibition. Adsorption inhibition is linked to plasmids pME0030, pSK112 and pCI528. Injection blocking uses a yet unknown mechanism and is related to plasmid pNP40. Modification and restriction are two other methods of phage resistance. Gene *Hsd*S on plasmids pSRQ800 and pSRQ900 initiates methylation of phage DNA. Three *Lla* genes encode for restriction endonuclease and are located on plasmids pSRQ700 and pTR2030. An operon with genes coding for methylases and restriction endonucleases is found on plasmids pND801 and pNP40. Abortive infection interferes with DNA replication, transcription, translation and packaging of phages. Genes on plasmids pMRC01 and pTR2030 are responsible for these interferences.

Bacteriocins among them lantibiotics [6] are polypeptides that are toxic to other (Gram+) bacteria. Lantibiotics prevent food spoilage by inhibiting cell wall synthesis and promoting pore formation in membranes of food pathogens. Immunity proteins protect *L. lactis* from bacteriocins. Genes coding for bacteriocins and immunity proteins are in the same operon on plasmids pMRC01, PSRQ900 and pBL01. Bacteriocins play a key role in flavor and texture of cheeses flavor [7].

Resistance to metal ions and transport of these compounds is linked to plasmids. Cadmium is toxic to *L. lactis* with resistance genes located on plasmid pAH90. Copper is a co-factor for enzymes and genes coding for transporters are found on plasmid pND306. Magnesium plays a key role in stability of ribosomes and membranes. Transporter genes are found on pCIS3, pAH90 and pNZ4000.

Exopolysaccharides (EPS) form a protection layer against phages. EPS are responsible for texture, mouth-feel, perception of taste and stability of (acidified) milk products [7]. Positive health effects (cholesterol) are also predicted [8]. Genes are located on plasmid pNZ4000.

Transfer of plasmids is common to bacteria and a mechanism to gain new (better) traits. This method of DNA exchange and DNA rearrangements has very high potential in dairy industry because transfer of plasmids is (considered) food-grade. No foreign and no synthetic DNA is involved and therefore the organisms created are not GMO (Genetically Modified Organism).

Research on Lactococcus lactis WG2

Bacteriophage infection of *L. lactis* WG2 starter cultures is a major problem in diary industry [9]. Phages connect (adsorption) to cell surface receptors (carbohydrate) of Gram+ bacteria and cause lysis leading to negative effects on fermentation and product yield. Bacteriophages use receptor-binding proteins (RBP) to connect to bacteria. Prevention of infection requires knowledge of this adsorption mechanism.

Mutations in genes [9] responsible for glycosyltransferases (*ycb*B) and membrane-spanning proteins (*ycb*C) showed reduced adsorption and increased resistance to phages. These genes are involved in biosynthesis and transport of cell wall polysaccharides (WPS) and part of the same operon. Loss of biosynthesis or export of WPS structures or parts of these structures is apparently sufficient for total inhibition of phage infection.

L. lactis WG2 uptakes small peptides or separate amino acids to obtain essential amino acids as nutrients for growth [10]. Extracellular serine proteinases hydrolyze casein while other (external) peptidases complete casein breakdown. Additional serine proteinases result in auto-proteolytic activity. Auto-proteolysis leads to enzyme release from the cell wall and reduced levels of hydrolysis. Ca^{2+} and pH regulate release and hydrolysis. Ca^{2+} stabilizes serine proteinases resulting in increased levels of casein hydrolysis. Optimal pH for enzyme release is 6.5 - 7.5 while optimum for hydrolysis is 6.0 – 6.5. Knowledge of both release and hydrolysis of serine proteinases is necessary for creating optimal growth conditions in milk.

Quality starter cultures of *L. lactis* WG2 have features such as high yield product (lactose), high level of proteolytic activity (casein) and good phage resistance [11]. In amino acid rich medium non-proteolytic (Prt^-) variants rapidly replace proteolytic (Prt^+) variants. Prt^- variants lack pWV03 and pWV05 [12] because information (traits) on these plasmids is no longer needed. Plasmid pWV03 specifies proteolytic activity (casein) of *L. lactis* WG2 while plasmid pWV05 is involved (probably) in copper (co-enzyme) sensitivity and transport.

Overview of sequencing methods

Traditional sequencing methods use synthesis of new DNA through incorporation of bases (nucleotides) into a denatured strand [13]. First generation sequencing (Sanger) incorporates chain-terminating dideoxynucleotides using DNA polymerases. Visualization of different size fragments reveals the complete DNA sequence (Figure 4). NGS (Next Generation Sequencing) incorporates nucleotides with four different fluorescent labels into a strand. The fluorescent labels detach upon incorporation and the sequence of detachment signals uncovers the DNA sequence (Figure 5).

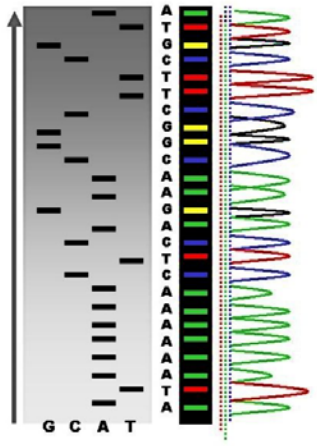

Figure 4 Sanger sequencing. DNA Fragments of different size are visualized in agarose gel (left) and through signals (right).

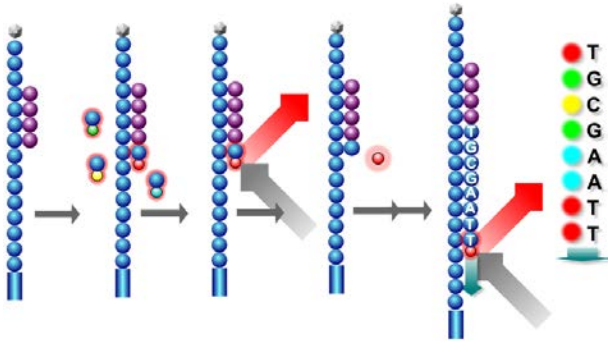

Figure 5 NGS sequencing. Detaching fluorescent labels upon incorporation of nucleotides to DNA strand producing signals.

Both methods work well for short stretches of DNA. In case of long stretches both methods are too error prone [14]. Shotgun sequencing solves this problem. DNA is broken randomly into numerous short fragments, fragments are sequenced (parallel) and resulting reads are assembled into a complete genome using software.

PacBio is a single-molecule NGS sequencer that creates error prone (15%) long reads while Illumina is a NGS method that sequences short fragments creating short reads with low error rates (1%). PacBio long reads enable complete genome assembly though with a low accuracy level. Complete genome assembly based on Illumina short is often impossible. An additional drawback of Illumina sequencing is that this method needs DNA multiplication, performed by PCR (Polymerase Chain Reaction) that causes artifacts [14]. Despite disadvantages, Illumina is very suitable finding gene mutations. Ease to find mutations, no need to assemble complete genomes and low costs make Illumina the most popular sequencer. Today Illumina makes up 98% of all sequencing activities [15].

MinION a new sequencing device

MinION from Oxford Nanopore Technologies (ONT) is a sequencing device that arrived on the market in 2014. This device is the size of a smart phone and connects to a (Windows) PC using USB 3.0. MinION is easy to use after a short learning period and its portability makes it suitable for (future) research on location.

The MinION does not use any nucleotide incorporation method to sequence but monitors DNA molecules running through nanopores directly [16] (Figure 6). A constant ionic current runs through 512 membrane located nanopores. DNA stretches running through these pores cause changes in the ionic current [16]. These fluctuations in combination with a base-calling algorithm determine DNA sequences.

Figure 6 DNA through membrane located pore and ionic current pattern of DNA monitoring

MinION at work

After growing cells, DNA isolation and fragmentation a library preparation is performed. Two different adaptors are ligated to DNA fragments. Motor adaptors lead fragments through nanopores and determine speed of processing [17] and hairpin adaptors keep template and complement strands connected during sequencing (Figure 7). Sequencing fragments ligated to both type of adaptors leads to 2D reads.

Randomness of ligation process produces DNA stretches with two motor adaptors (Figure 8) and with two hairpin adaptors (Figure 9). Fragments with two motor adaptors are sequenced resulting in 1D reads while DNA fragments with two hairpin adaptors cannot be sequenced.

Figure 7 Motor and hairpin adaptor ligated to DNA strand

Figure 8 Motor adaptors ligated to both ends of DNA strand

Figure 9 Hairpin adaptors ligated to both ends of DNA strand

High quality 2D reads are required for successful data processing while 1D reads are of lower quality and are less useful for data processing. Sequencing fragments [17] involves a series of steps (Figure 10).

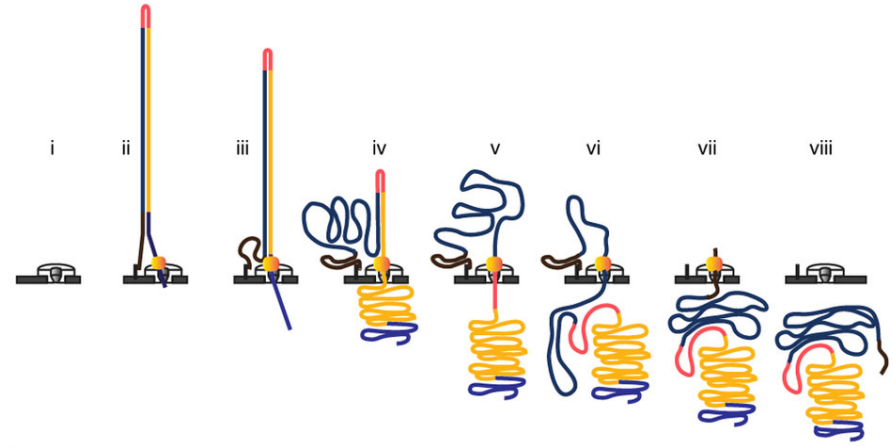

Figure 10 DNA translocation steps of 2D reads

Legend

I. open pore
II. capture and translocation of lead adaptor
III. translocation of template strand
IV. translocation of hairpin adaptor
V. translocation of complement strand
VI. translocation of tethering adaptor
VII. release of DNA molecule
VIII. open pore (next DNA fragment)

Separate steps have unique patterns. Ionic current patterns of strand translocation produce reads. Groups of 5 nucleotides (*words*) are processed simultaneously during translocation of stands, meaning every nucleotide is sequenced 5 times at single nucleotide precision.

Advantages and drawbacks of MinION

MinION shows numerous advantages [18]. MinION uses limited amounts of DNA (1 µg). PCR is not needed creating sufficient amounts of input because this device sequences fragments of DNA directly. MinION is capable of reading long fragments (> 10 KB) which is important for detection of repeats. Speed of sequencing is 75 base pairs per second at a single-nucleotide precision and sequencer reads 512 fragments parallel resulting in 10 MB (maximum) data per hour. MinION can sequence all types of molecules that run through pores making it possible for future sequencing of proteins and methylated DNA.

Drawbacks are high error rates of MinION long reads. High error rate of MinION long reads (15%) obscure alignment [18]. Error prone long reads make assembly of complete genome difficult. Poor quality of MinION reads can be corrected using software.

Steps of data processing

Sequencing devices (Illumina – PacBio - MinION) produce data and not information. Software tools transform reads (data) into assemblies (information). All software tools use three main steps [16]. The first step is to find 1-on-1 overlaps between reads. This process compares all reads to each other. Minimum size of overlap (*k-mer*) and acceptable rate of error are both crucial for quantity and quality of overlaps. A second step is to use overlaps to merge reads and create contigs (contiguous sequences. The final step is to connect contigs and build the assembly (Figure 11).

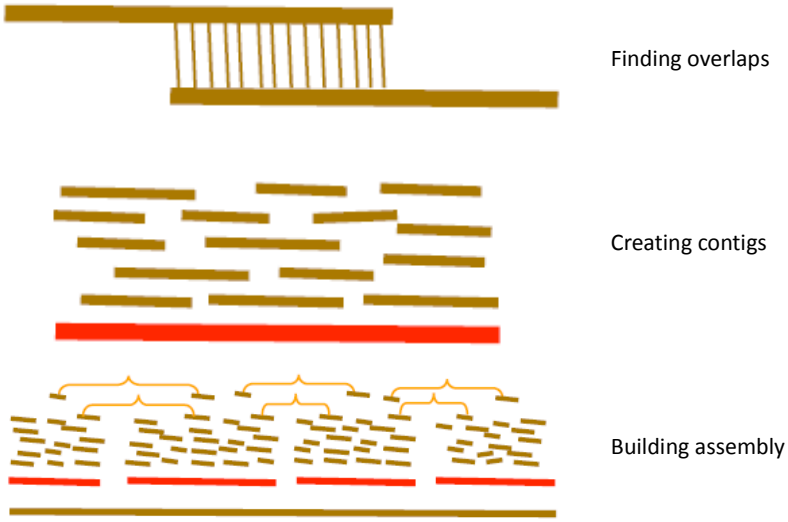

Figure 11 Steps of data processing. Finding one-on-one overlaps, creating contigs based on overlaps and building assembly using contigs.

Pros and cons of short reads and long reads

Sequencers use different techniques and sequencers produce short reads (Illumina) or long reads (PacBio - MinION). Short reads (< 500 bp) contain low error rates (1%) while long reads (> 1 kbp) have higher error rates (15%). Finding overlaps and creating high quality contigs are feasible using short reads but connecting contigs and finding repeats are difficult. Therefore building a complete genome assembly based on short reads is not an option.

Limited length of short reads makes it impossible to connect separate contigs and span multiple repeats. Long reads can connect separate contigs (Figure 12) and span (multiple) repeats [19] and determine repeat number (Figure 13).

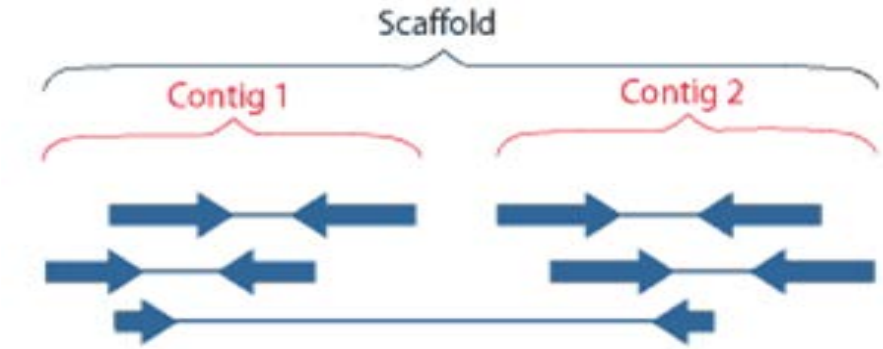

Figure 12 Long read (bottom blue line) connecting separate contigs (top).

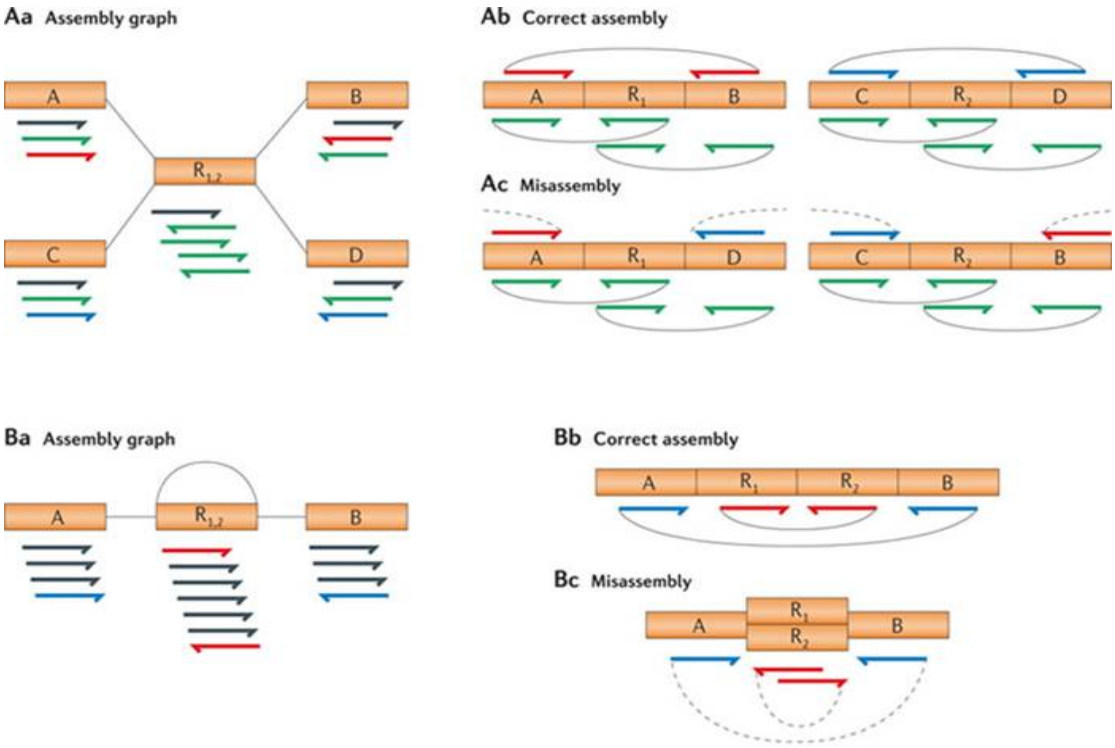

Figure 13 Long reads (Aa) connecting starting locations and stopping locations for correct assemblies. Long reads (Bb) spanning (multiple) repeats for correct copy numbers.

When repeats occur at multiple locations (*Aa Assembly graph*) long reads connecting starting locations (A and C) and stopping locations (B and D) are needed to determine (only) correct assemblies (*Ab Correct assembly*). Without long reads incorrect assemblies (*Ac Misassembly*) are deduced too.

When repeats have multiple occurrences (*Ba Assembly graph*) long reads connecting starting location (A) and stopping location (B) are essential creating assembly (*Bb Correct assembly*) with correct repeat copy number. Without long reads assembly with incorrect copy numbers (*Bc Misassembly*) is built too.

Error correction is the solution

Assemblies based on long reads have low quality while problems connecting contigs and solving repeats make assemblies using short reads impossible. Error correction can solve both problems. High quality short reads (top) correct low quality long reads (bottom) and corrected long reads build complete genome assemblies (Figure 14).

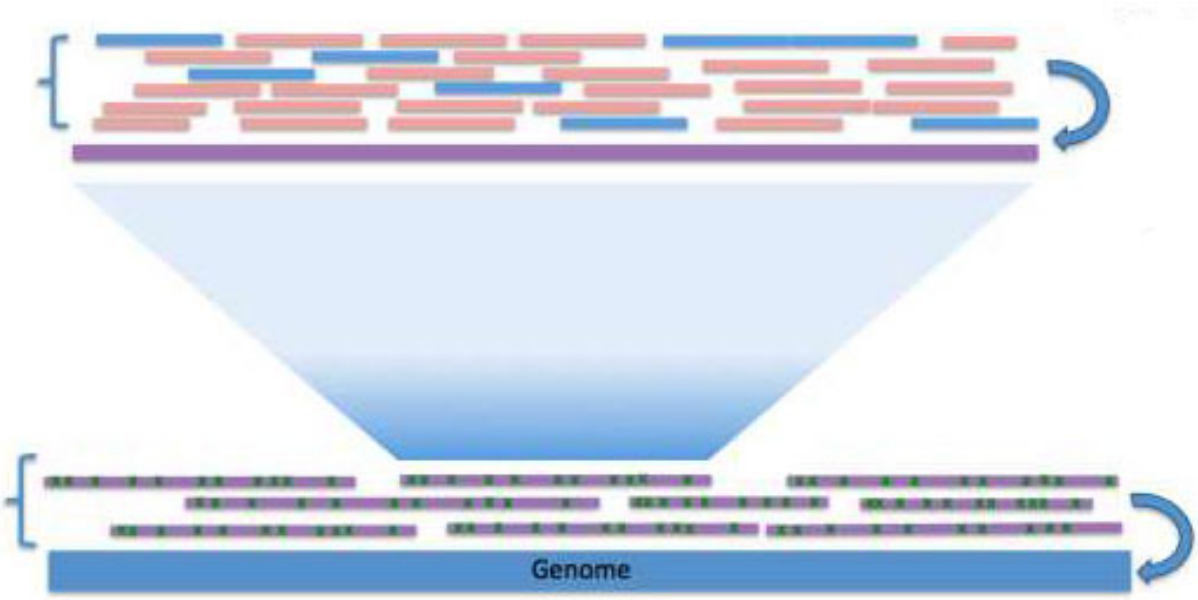

Figure 14 Error prone long reads (bottom) building complete genome assembly. Short reads (top) correcting long read. Error corrected long reads can be used to build improved complete genome assembly.

Overview of software tools

Various research projects proved error correction worked. Hybrid methods use short reads and long reads originating different sequencing methods while non-hybrid methods use reads from one sequencing device (Table 1).

Short reads (Illumina) and long reads (PacBio) were used to prove hybrid error correction and to assembly *de novo* the genome of *E. coli* C227-11 [14]. Initial high error rates of long reads obscured alignment and made assembly impossible. Error correction improved long reads quality using software tool PBcR (PacBio corrected Reads) while Celera Assembler software build the complete genome assembly. Quality and speed of assembly increased compared to using short reads (Illumina data) only.

Research [13] showed 99% identity of assembly to reference chromosome *E. coli* K12 using short reads (Illumina) and long reads (Minion). Nanocorr performed hybrid error correction and increased identity from 67% (uncorrected reads) to

97% (corrected reads) [13] with Celera Assembler building complete genome assembly too.

Short reads (Illumina) and long reads (MinION) were combined in hybrid error-correction [20] to assemble single chromosome of *Bacteriodes fragiles* strain BE1 without gaps. SPAdes [21] created contigs based on short reads, SSPACE-LongRead [22] connected contigs using long reads and GapFiller [23] filled gaps and improved complete genome assembly. Velvet [24] is alternative to SPAdes when creating contigs based on Illumina short reads.

MinION short reads and long reads were used in non-hybrid error correction and complete genome assembly [16]. Poretools [25] software performed data extraction of raw MinION data (fast5) into input (fasta). Nanocorrect detected overlaps, created alignments and did error correction of long reads using short reads. Celera Assembler assembled complete genome and Nanopolish improved genome assembly using original ionic current patterns stored in MinION data (fast5). Nanocorrect and Celera Assembler were optimized and combined (fork) into CANU [26].

Table 1 Overview of assembly software

Software	Input	Output	Remarks
Velvet	Illumina short reads	Contigs	Used in this research
PBcR	Illumina short reads & PacBio long reads	Contigs	Not used in research
Pipeline Nanocorrect & Celera Assembler & Nanopolish	MinION short reads and long reads	Contigs Complete Assembly	Not used in research (outdated)
Pipeline CANU & Nanopolish	MinION short reads and long reads	Contigs Complete Assembly	Used in this research
Pipeline SPAdes & SSPACE & GapFiller	Illumina short reads & MinION long reads	Contigs Complete Assembly	Used in this research

Goals of this project

Objects of our research were performing all sequencing steps ourselves, sequencing strain *Lactococcus lactis* WG2 and using MinION. We do isolate DNA at our laboratories but third parties perform DNA fragmentation and library preparation. In order to save resources we intend to perform all steps ourselves. In order to use MinION we need to perform all steps ourselves because MinION is completely DIY. We wanted to sequence strain *Lactococcus lactis* WG2 because this strain has not been sequenced. Our main goal was to use MinION for sequencing DNA and to use data processing software to assemble complete genomes.

Materials and Methods

DNA extraction

Strain *Lactococcus lactis* WG2 was grown in LM17 medium (M17 + 0.5% lactose) overnight at 30°C. Culture (12 ml) was centrifuged at 4000 x g for 2 min and pellet was washed once with MQ. Pellet was re-suspended in 2 ml solution A + lysozyme (5 mg/ml) and then incubated at 55 °C for 10'. 80 µl proteinase K (20 mg/ml) and 100 µl 10% SDS were added to initiate lysis for 1 hour at 60°C until solution was clear. MQ (1 ml) and phenol (1 ml) were added, solution was incubated for 10' at R.T. and chloroform (1 ml) was added. Solution was mixed by inversion every step. Prior to use Phase Lock Gel (PLG) tube (2 ml) was centrifuged at 10.000 x g for 30''. Solution (1 ml) was added to PLG tube and solution was mixed thoroughly. Phases were separated at 15.000 x g for 5'. Aqueous upper phase (500 µl) containing DNA was transferred to fresh tube (1.5 ml Eppendorf). DNA was precipitated by adding 50 µl 3 M Sodium Acetate (pH 5.2) plus 500 µl 2-propanol and mixed by inversion until DNA was visible. Solution was centrifuged for 5', supernatant discarded and DNA pellet washed with ethanol (70%). DNA pellet was dissolved in 200 µl TE (10x), at 60°C incubated for 1 hour and 5 µl RNase (10 mg/ml) added. Aliquot was stored at 4°C for later use.

Quantity of DNA was checked using NanoDrop. Concentration was 1500 ng/µl (minimum should be 850 ng/µl). Quality of DNA was checked through gel-electrophoresis (1% agarose gel TEA with EtBr 0.5 µg/ml). Extracted DNA was concentrated and showed clear bands (should not be smeared).

DNA fragmentation for Illumina (MiSeq)

Aliquot with DNA was mixed by inversion. NEBNext dsDNA Fragmentase (M0348) was vortexed for 3', centrifuged and placed on ice. Digestion reaction was set up mixing 10 µl DNA, 5 µl MQ, 2 µl Fragmentase Reaction Buffer v2 (10X) and 1 µl $MgCl_2$ (200 mM) and vortexed well. Solution was incubated for 5' at R.T. and 2 µl dsDNA Fragmentase was added. For fragment lengths 300 – 500 base pairs digestion reaction was allowed for 20'' at 37°C. Digestion reaction was stopped adding 5 µl EDTA (0.5 M) and heat-shock at 60°C.
Quantity of DNA was checked using NanoDrop and quality of DNA was checked through gel-electrophoresis.

Library Preparation for Illumina (MiSeq)

Fragmented DNA (55 µl) was used for library preparation using NEBNext Ultra DNA Library Prep Kit for Illumina (NEB #E7370) protocol. Library Prep was done as per manufacturer's instructions except End Prep was done 30' at R.T.

and 30' at 60°C - 65°C using a heat-block, no Size-Selection was done (DNA was already fragmented), Cleanup was performed using NucleoMag beads (Macherey-Nagel) and 70% Ethanol and PCR amplification was done with 6 cycles (high concentration DNA).

DNA fragmentation and Library Preparation for MinION

Sequencing Kit SQK-MAP006 was used for Genomic DNA sequencing. From aliquot with not fragmented DNA 1 µl (1500 ng/µl) was used as starting material in 45 µl MQ. Shearing was done using g-tubes (Covaris) and centrifuged at 3200 x g for 1 minute. Library Prep was performed as per manufacturer's instructions [27] [28].

In order not to compromise success of the end-prep and to improve read length sheared DNA was repaired using NEBNext FFPE RepairMix (M6630). DNA was end-repaired and was dA-tailed using NEBNext Ultra II End-Repair / dA-tailing Module (E7546). DNA purification was done using Agencourt AMPure XP beads. Adapters were ligated to the dA-tailed DNA using NEB Blunt/TA Ligase Master Mix and DNA cleaning was performed using Streptavidin MyOne C1 beads. Total 150 µl prepared library and total 1000 µl priming mix were prepared.

Sequencing using Illumina (MiSeq)

Extracted DNA was send to EMBL (European Molecular Biology Laboratory) in Heidelberg for sequencing. After library preparation and fragmentation (average) fragment length was 200 base pairs. Fragments were sequenced in forward and reverse direction multiple times.

Sequencing using MinION

MinION Mk1 was used for sequencing. Flow cells (FLO-MAP103) from 4°C storage were used with device. Priming mix (2 x 500 µl) and prepared library (150 µl) were loaded into flow cell according to protocol. Device was connected to Windows 7 PC (512 GB SSD and 8 GB RAM). MinKNOW software (0.51.162) with script *MAP_48Hr_Sequencing_Run_SQK_MAP006* and Metrichor software (v2.38.3) with application *2D Basecalling for SQK-MAP006* were used for sequencing. Three flow cells were used for sequencing.

Data processing

Illumina only data (short reads) was processed using Velvet software. MinION only data (short and long reads) was processed using a pipeline containing CANU and Nanopolish software. Combination of Illumina data (short reads) and MinION data (long reads) was processed in a pipeline containing SPAdes, SSPACE-LongRead and GapFiller software.

Computer hardware

Software packages were installed and programs and pipelines were executed on Linux HPC (High Performance Cluster) named Peregrine of University of Groningen (RUG).

Results

DNA extraction and fragmentation

Lactococcus lactis WG2 was grown ON in 80 ml LM17 medium and chromosomal DNA was isolated according to protocol using Phase Lock Gel (PLG). This resulted in 1500 ng/µl (NanoDrop) and gel-electrophoresis was used for quality control (Figure 15).

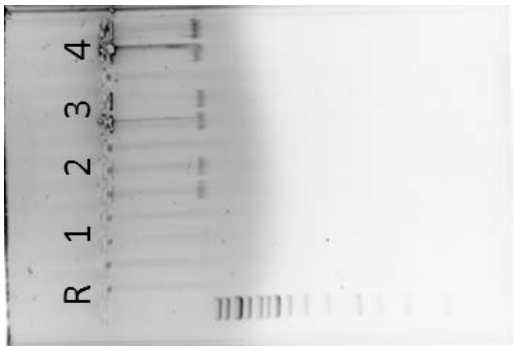

Figure 15 Agarose-gel (1%) in TEA buffer. Four samples (1 – 4) in duplicate and reference (R). Sample 1 is very faint for unknown reasons. Samples 2 – 4 show clear bands of concentrated DNA.

Sequencing techniques (Illumina - MinION) use different optimal DNA fragment. Illumina uses optimal fragments of 200 – 300 base pairs while MinION needs fragments of 7000 base pairs. We performed DNA fragmentation in order to create optimal fragment lengths. NEBNext fragmentase was used with Illumina sequencing [15] and Covaris g-tubes mechanical shearing in case of MinION. Increasing the time of fragmentation results into decreased length of fragments (Figure 16). We experienced that fragmentation methods have low levels of stability and reproducibility. Third parties (EMBL for Illumina sequencing) experience both drawbacks too. Auto-fragmentation of DNA is very common (Figure 17).

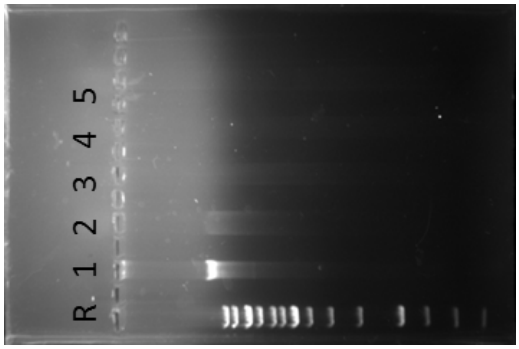

Figure 16 Agarose-gel (1%) in TEA buffer. Five samples (1 – 5) with different times of fragmentation (5' – 25') and reference (R). Increasing times result into shift to smaller fragment lengths.

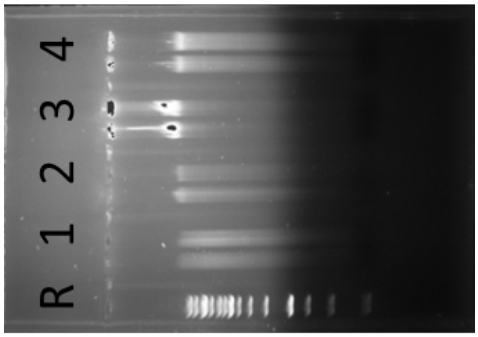

Figure 17 Agarose-gel (1%) in TEA buffer. Four samples (1 – 4) in duplicate and reference (R). All four samples show auto-fragmentation of DNA.

Pipeline testing

For testing CANU and Nanopolish software we used only MinION 2D reads (22.270 reads) of *E. coli* MG1655 as input. Total input data consisted of 133.6 MB data with 30x coverage of 4.6 MB chromosome of *E. coli* MG1655 [16]. CANU produced a genome assembly with 96% identity to the reference genome (Figure 18) while Nanopolish software improved the assembly to > 99% identity (Figure 19). Assembly errors mainly occurred in poly-A/T/C/G stretches. These results show we proved the concept of this pipeline and complete genome assembly is possible using only MinION data.

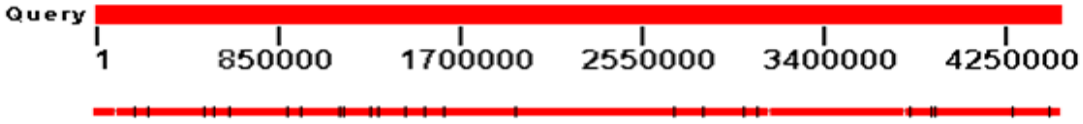

Figure 18 Draft genome assembly. Top red line shows draft genome assembly (query) and bottom red line shows identity to reference genome. Differences (error locations) are indicated in blue.

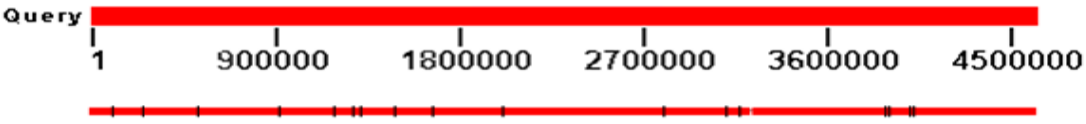

Figure 19 Polished genome assembly. Top red line shows draft genome assembly (query) and bottom red line shows identity to reference genome. Differences (error locations) are indicated in blue.

Sequencing L. lactis WG2

Chromosomal DNA samples of *L. lactis* WG2 were sent to EMBL for Illumina sequencing (PE300). We received total data 2.8 GB with 400x coverage in 5,194,212 short reads with average length of 300 base pairs.

The same DNA was sequenced using MinION. We extracted and fragmented DNA and performed library preparation according to protocols. We used three new flow cells for sequencing. Our first flow cell (run MolGen007) produced 51 KB of data. We scheduled our second flow cell (run MolGen009) to run for 48 hours. It ran for 6 - 7 hours producing total 13,000 reads containing 15 MB of 2D reads. Our third flow cell (run MolGen014) ran for 12 hours with re-fills every 4 hours and produced 33,000 reads that contained total 27 MB of 2D reads. We used only high quality 2D reads (pass) from run MolGen009 and MolGen014 for data processing. Input from these two runs was 17 MB in 3,394 2D reads (pass) resulting in 6.2x coverage of 2.6 MB chromosome of *L. lactis* WG2. We ignored low quality 2D reads (fail) and 1D reads (fail) for further data processing.

We used Poretools [25] functions "*hist"* and "*yield plot"* and poRe [29] software to produce images (Figure 20 & Figure 21) based on performance of these two runs. Cumulative yield (a) shows limited yield and runtime (Molgen009) compared to increased values (Molgen014). We used mechanical shearing to get 7000 base pairs sized fragments. Fragments (b) shows undersized fragments (Molgen009) compared to optimal fragments (Molgen014). Cumulative base pairs (c) shows low yield (Molgen009) 2D reads (98) compared to high yield (Molgen014) 2D reads (3296). Operating pores (d) shows poor performance (Molgen009) of 100 pores compared to good performance (Molgen014) of 400 pores.

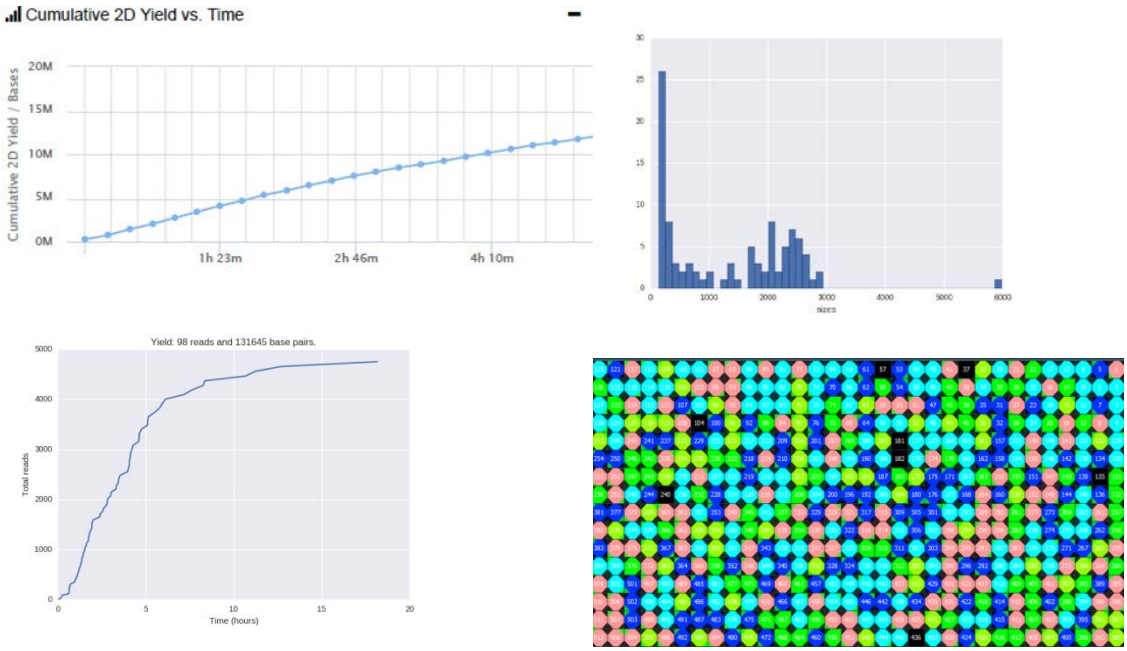

Figure 20 Molgen009. Cumulative yield (a) of 2D reads data (MB) at runtime (hours). Distribution (b) of fragment lengths. Cumulative number and total base pairs (c) of 2D reads after runtime (hours). Distribution (d) of operating pores in green.

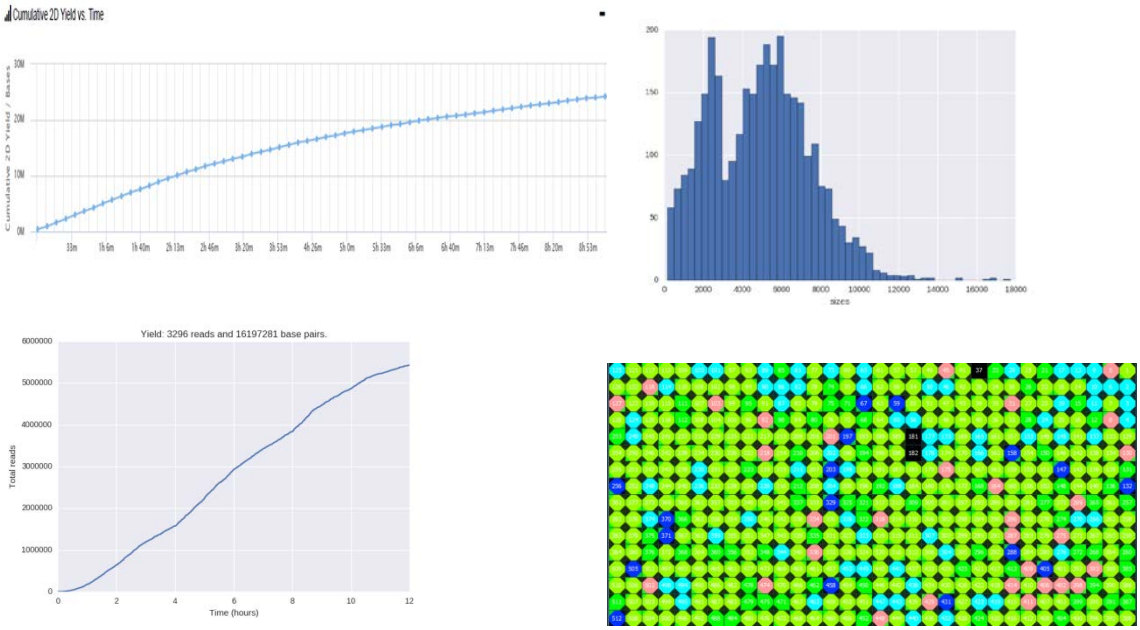

Figure 21 Molgen014. Cumulative yield (a) of 2D reads data (MB) at runtime (hours). Distribution (b) of fragment lengths. Cumulative number and total base pairs (c) of 2D reads after runtime (hours). Distribution (d) of operating pores in green.

Data processing

CANU and Nanopolish software processed MinION (only) data. We used total 2D reads (3,394) of *L. lactis* WG2 as input. Total consisted of 17 MB data with 6.2x coverage of 2.6 MB chromosome. Pipeline produced genome assembly of 1.7 MB in 285 contigs. We refrained from further research because substandard assembly size and large contigs number.

Velvet created total 210 contigs with assembly size 2.6 MB based on 2.8 GB Illumina short reads. We decided on further research because acceptable genome assembly size but despite large contigs number.

Pipeline containing SPAdes, SSPACE-LongRead and GapFiller performed data processing of short reads (Illumina) and long reads (MinION). Pipeline created 51 contigs covering genome assembly sized 2.6 MB. We justified further data mining because optimal assembly size and acceptable contigs number.

Analysis of results

SEED section of RAST webserver found closest neighbors of *L. lactis* WG2 and searched for homology with a reference genome. Pipeline contigs (51) were used as input. SEED shows closest neighbors (Figure 22) are *L. lactis* strains.

Strain *L. lactis* MG1363 was used as reference searching for (chromosome) homology because strain has a (circular) chromosome but contains no plasmids [30]. Total 29 contigs map to the reference chromosome. Areas colored “green” indicate high levels of homology (Figure 23). Unmapped contigs contain possibly plasmid DNA.

Chromosome of *L. lactis* MG1363 includes 1437 known functional genes. Total 29 mapped contigs contain 90% of these functional genes. Unmapped contigs hold 10% unique genes unknown and unrelated to *L. lactis* MG1363 genes. Unique genes of *L. lactis* WG2 relate to stress response, cell wall synthesis, defense (virulence & disease), carbohydrate metabolism (lactose & galactose) and transport of co-factors. Plasmids often hold genes coding for these functions.

Lactococcus lactis subsp. cremoris SK11
Lactococcus lactis subsp. cremoris SK11
Lactococcus lactis subsp. cremoris MG1363
Lactococcus lactis subsp. cremoris MG1363
Lactococcus lactis subsp. lactis Il1403
Lactococcus lactis subsp. cremoris NZ9000
Lactococcus lactis subsp. lactis Il1403
Lactococcus lactis subsp. lactis KF147
Lactococcus lactis subsp. cremoris UC509.9
Lactococcus lactis subsp. lactis CV56
Lactococcus lactis subsp. cremoris CNCM I-1631
Lactococcus garvieae UNIUD074
Lactococcus garvieae 8831
Lactococcus garvieae 21881
Lactococcus lactis subsp. cremoris A76

Figure 22 List of closest neighbours of *Lactococcus lactis* WG2 based on homology with 51 contigs.

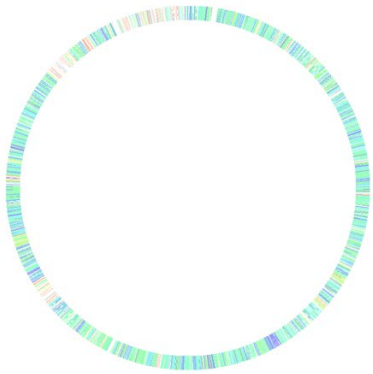

Figure 23 Circular chromosome of reference MG1363 indicating homology (green) with *Lactococcus lactis* WG2 based on 29 contigs.

CONTIGuator [31] webserver was used to visualize mapping to a reference genome using pipeline contigs (51) as our input. Total 13 contigs map to reference *L. lactis* MG1363. Circularity of prokaryotic chromosomes causes cross-links at both ends of the alignment (Figure 1). Rest of alignment should show parallel lines indicating similar gene locations on chromosomes of both strains. High level of cross-links (Figure 24) means significant genome re-arrangement or numerous assembling errors.

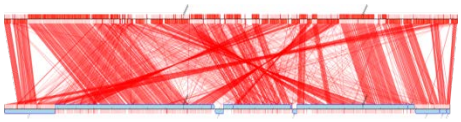

Figure 24 Artimis visualization of mapping 51 pipeline contigs (bottom) of *L. lactis* WG2 to reference *L. lactis* MG1363 (top).

Webserver CONTIGuator was used with Velvet contigs (210) to determine origin of high cross-links numbers. Velvets contigs based on Illumina data map 92 to reference. This alignment shows low level of cross-links and high amount of parallel lines (Figure 25). Usage of MinION data did not improve overall quality of contigs used for alignment to reference. SSPACE-LongRead [22] connected numerous contigs based on Illumina short reads incorrectly.

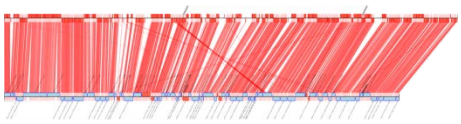

Figure 25 Artimis visualization of mapping 210 Velvet contigs (bottom) of *L. lactis* WG2 to reference *L. lactis* MG1363 (top).

Website NCBI webserver was visited to find protein-coding genes using pipeline contigs (51) as input data (Figure 26). Total 13 contigs with protein-coding genes map to reference (circular) chromosome. Total 15 contigs possess protein-coding genes but do not map to reference. These 15 contigs possibly represent

plasmids of *L. lactis* WG2. Total 23 contigs harbor no protein-coding genes. All 23 contigs are small-sized (< 1000 base pairs) and 14 have complete identical length (497 base pairs). These 23 contigs probably are assembling errors.

Contigs (13) mapped to reference chromosome contain 98% of total assembly and code for 2,300 proteins. Contigs (15) representing plasmids contain 2% of assembly and code for 40 proteins. Percentages of total assembly and proteins numbers (Figure 27) are common to chromosome and plasmid DNA in *L. lactis* strains. Most proteins related to plasmid DNA are "hypothetical proteins" with unknown functions. Contigs resulting from assembling errors contain ~0% of total assembly and do not code for any proteins.

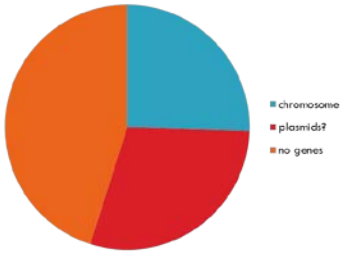

Figure 26 Distribution of contigs (numbers). Contigs with protein coding genes mapped to chromosome (blue) and plasmids (red) and contigs with no-protein coding genes (orange).

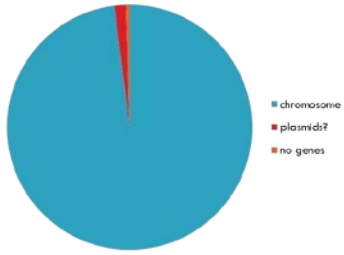

Figure 27 Distribution of total assembly (percentages). Part of total assembly with proteins coded by genes on chromosome (blue) and plasmids (red) and part of total assembly with no-protein coding DNA (orange).

Discussion

We aimed for two goals. We wanted to perform all sequencing steps using MinION and we intended to sequence *Lactococcus lactis* WG2 for complete genome assembly.

We described and used protocols for growing cells, DNA extraction, DNA fragmentation and library preparation. Insights were gained about problems with stability and reproducibility of both methods of DNA fragmentation (enzymatic and mechanical). These problems are well known and third parties (EMBL) have not solved them either. Based on our findings good protocols were developed.

Different pipelines were built and used for data processing on Linux HPC (High Performance Cluster) called Peregrine of University of Groningen (RUG). We managed to install and run all necessary software on Peregrine. Priority setting and scheduling of jobs on Peregrine is not optimal yet. Excessive claims by few users caused substantial negative effects on other users. Regular updates, new releases and depreciation of software made developing and testing of pipelines more time consuming than expected. Installation scripts and manuals for pipelines have become available, ready for use by members (biologists) of our research group.

Illumina and MinION were used for sequencing *Lactococcus lactis* WG2 and used pipelines for data processing. We were not able to produce useful output solely based on MinION data but were capable of producing useful output based on Illumina data and combinations of Illumina short reads and MinION long reads. Webservers (RAST, CONTIGuator and NCBI) proved homology to other *L. lactis* strains and similarity of gene location and functions on (circular) chromosome. MinION data in combination with Illumina based contigs led to complete genome assembly. We decided on registration of our assembly (contigs) of L*actococcus lactis* WG2 in NCBI database, despite suboptimal result.

We did library preparation for MinION ourselves. Protocol takes ½ day and is manually. We think this procedure is possible source of errors and automatic sample preparation (Voltrax) will improve future quality of MinION sequencing.

MinION was used to sequence *Lactococcus lactis* WG2. We found substantial difference in the quality of flow cells (bad / poor / good), were not able to re-use our flow cells and did not reach maximum runtime of 48 hours. Only 17 MB data of high quality 2D reads with only 6.2x coverage was obtained, due to the flow cell problems mentioned. Quantity and quality (15% error rate) of MinION data prevented both complete genome assembly and improvement of contigs based on Illumina data.

In our opinion MinION has high potentials when Oxford Nanopore Technologies (ONT) solves problems mentioned above. We used R007-release flow cells and experienced both low stability and no re-usability. Follow-up R009-release has increased speed of translocation (500 vs 75 bps), contains more pores (1024 vs 512) and achieves better accuracy. Improved flow cells lead to increased levels of high quality 2D reads and coverage. Automatic library preparation using Voltrax improves quantity and quality data too. All improvements probably make complete genome assembly based on MinION (only) data feasible.

We think that MinION has future possibilities when sequencing meta-genomes, phages and plasmids. MinION long reads can span numerous repeats and can connect separate contigs into complete genome assemblies of different organisms (meta-genomes), phages and plasmids. We think that MinION is cost efficient. To sequence 100 phages costs € 21,500 using Illumina and € 2,700 using combination of Illumina and MinION data.

MinION could offer an extra option to DNA sequencing on the bench. Other methods (Illumina and PacBio) require third party sequencing, consume time (weeks) and are not very cost efficient. MinION is completely DIY, takes less time (maximum 48 hours) and is low on resources. Other methods produce large amounts of data with high levels of coverage while small amounts of low coverage MinION data are sufficient searching for DNA mutations.

Acknowledgements

I like to thank the people of Molecular Genetics (MolGen) research group of the Groningen Biomolecular Sciences and Biotechnology Institute (GBB) at University of Groningen (RUG) for having a wonderful time with them. I want to express my appreciation in particular to two members of this research group.

I want to thank Dr. Anne de Jong for his time and energy. He was my daily supervisor during my master research project. Without his knowledge, help, patience and humor I would have been lost. Thanks to Anne, I realized that scientific research is about searching for problems rather than looking for solutions.

I am very grateful to Prof. Jan Kok for offering me this opportunity. He was my overall supervisor and is my study mentor. His enthusiasm for *Lactococcus lactis* and his eagerness to use MinION inspired me to keep on going and not to quit. Due to Jan, I realized more than ever that perseverance is essential for success.

References

[1] C. R. Melchiorsen, N. B. Jensen, K. V Jokumsen, H. Israelsen, J. Arnau, and J. Villadsen, "7 Dynamics of pyruvate metabolism in," *Biotechnol. Bioeng.*, vol. 74, no. 4, pp. 271–279.

[2] A. R. Neves, W. A. Pool, J. Kok, O. P. Kuipers, and H. Santos, "Overview on sugar metabolism and its control in Lactococcus lactis - The input from in vivo NMR," *FEMS Microbiol. Rev.*, vol. 29, no. 3 SPEC. ISS., pp. 531–554, 2005.

[3] R. Larsen, G. Buist, O. P. Kuipers, and J. Kok, "ArgR and AhrC Are Both Required for Regulation of Arginine Metabolism in Lactococcus lactis ArgR and AhrC Are Both Required for Regulation of Arginine Metabolism in Lactococcus lactis," *Society*, vol. 186, no. 4, pp. 1147–1157, 2004.

[4] S. Mills, O. E. McAuliffe, A. Coffey, G. F. Fitzgerald, and R. P. Ross, "Plasmids of lactococci - genetic accessories or genetic necessities?," *FEMS Microbiol. Rev.*, vol. 30, no. 2, pp. 243–273, 2006.

[5] J. a Farrow, "Lactose hydrolysing enzymes in Streptococcus lactis and Streptococcus cremoris and also in some other species of streptococci.," *J. Appl. Bacteriol.*, vol. 49, no. 3, pp. 493–503, 1980.

[6] C. van Kraaij, W. M. de Vos, R. J. Siezen, and O. P. Kuipers, "Lantibiotics: biosynthesis, mode of action and applications.," *Nat. Prod. Rep.*, vol. 16, no. 5, pp. 575–587, 1999.

[7] R. J. Siezen, R. J. Siezen, B. Renckens, B. Renckens, I. Van Swam, I. Van Swam, S. Peters, S. Peters, R. Van Kranenburg, R. Van Kranenburg, W. M. De Vos, and W. M. De Vos, "Complete Sequences of Four Plasmids of," *Society*, vol. 71, no. 12, pp. 8371–8382, 2005.

[8] H. Nakajima, Y. Suzuki, H. Kaizu, and T. Hirota, "Cholesterol Lowering Activity of Ropy Fermented," vol. 57, no. 6, pp. 1–2, 1992.

[9] K. Dupont, T. Janzen, F. K. Vogensen, J. Josephsen, and B. Stuer-Lauridsen, "Identification of Lactococcus lactis genes required for bacteriophage adsorption.," *Appl. Environ. Microbiol.*, vol. 70, no. 10, pp. 5825–32, Oct. 2004.

[10] H. Laan and W. N. Konings, "Mechanism of proteinase release from Lactococcus lactis subsp. cremoris Wg2," *Appl. Environ. Microbiol.*, vol. 55, no. 12, pp. 3101–3106, 1989.

[11] R. Otto, W. M. de Vos, and J. Gavrieli, "Plasmid DNA in Streptococcus cremoris Wg2: Influence of pH on Selection in Chemostats of a Variant Lacking a Protease Plasmid.," *Appl. Environ. Microbiol.*, vol. 43, no. 6, pp. 1272–7, 1982.

[12] A. J. Haandrikman, C. Van Leeuwen, J. Kok, P. Vos, W. M. De Vos, and G. Venema, "Insertion elements of lactococcal proteinase plasmids," *Appl. Environ. Microbiol.*, vol. 56, no. 6, pp. 1890–1896, 1990.

[13] S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. Schatz, and W. R. McCombie, "Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome," *bioRxiv*, p. 013490, 2015.

[14] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy, "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nat. Biotechnol.*, vol. 30, no. 7, pp. 693–700, 2012.

[15] J. G. Caporaso, C. L. Lauber, W. a Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. a Gilbert, G. Smith, and R. Knight, "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms,"

ISME J., vol. 6, no. 8, pp. 1621–1624, 2012.

[16] N. J. Loman, J. Quick, and J. T. Simpson, "A complete bacterial genome assembled de novo using only nanopore sequencing data," *bioRxiv*, p. 015552, 2015.

[17] M. Jain, I. T. Fiddes, K. H. Miga, H. E. Olsen, B. Paten, and M. Akeson, "Improved data analysis for the MinION nanopore sequencer," *Nat Meth*, vol. 12, no. 4, pp. 351–356, Apr. 2015.

[18] N. J. Loman and M. Watson, "Successful test launch for nanopore sequencing," *Nat. Methods*, vol. 12, no. 4, pp. 303–304, 2015.

[19] A. Introduction and B. Algorithms, "Graph Algorithms in Bioinformatics," *Bioinformatics*.

[20] J. Risse, M. Thomson, S. Patrick, G. Blakely, G. Koutsovoulos, M. Blaxter, and M. Watson, "A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data.," *Gigascience*, vol. 4, no. 1, p. 60, 2015.

[21] A. Bankevich, S. Nurk, D. Antipov, A. a. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. a. Alekseyev, and P. a. Pevzner, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, 2012.

[22] M. Boetzer and W. Pirovano, "SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.," *BMC Bioinformatics*, vol. 15, no. 1, p. 211, 2014.

[23] M. Boetzer and W. Pirovano, "Toward almost closed genomes with GapFiller.," *Genome Biol.*, vol. 13, no. 6, p. R56, 2012.

[24] D. Zerbino, "Velvet: de novo assembly using very short reads," *J. Virol.*, vol. 44, no. 0, p. 494612, 2007.

[25] N. Loman and A. Quinlan, "Poretools: a toolkit for analyzing nanopore sequence data," *bioRxiv*, p. 007401, 2014.

[26] A. Phillippy, S. Koren, and B. Walenz, "canu Documentation," 2016.

[27] "Genomic DNA GDE_1002_v1_rev_17Nov2015," pp. 1–49, 2016.

[28] "Genomic DNA sequencing for the MinION TM device," pp. 1–2.

[29] M. Watson, M. Thomson, J. Risse, R. Talbot, J. Santoyo-Lopez, K. Gharbi, and M. Blaxter, "poRe: an R package for the visualization and analysis of nanopore sequencing data," *Bioinformatics*, vol. 31, no. 1, pp. 114–115, 2015.

[30] D. M. Linares, J. Kok, and B. Poolman, "Genome sequences of Lactococcus lactis MG1363 (revised) and NZ9000 and comparative physiological studies," *J. Bacteriol.*, vol. 192, no. 21, pp. 5806–5812, 2010.

[31] M. Galardini, E. G. Biondi, M. Bazzicalupo, and A. Mengoni, "CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes.," *Source Code Biol. Med.*, vol. 6, no. 1, p. 11, 2011.