[**ABSTRACT (1page)** 2](#_Toc139625800)

[**INTRODUCTION** 3](#_Toc139625801)

[- **Epigenome/methylome definition** 3](#_Toc139625802)

[- **Classification of MTases** 4](#_Toc139625803)

[- **MTases in *Escherichia coli* K-12 substr. MG1655** 4](#_Toc139625804)

[- **Methylation functions in bacteria** 5](#_Toc139625805)

[**Phenotypic heterogeneity** 9](#_Toc139625806)

[- **Bet-hedging** 11](#_Toc139625807)

[**Genome scale metabolic modelling** 14](#_Toc139625808)

[- **Dynamic flux balance analysis (dFBA)** 19](#_Toc139625809)

[**AIM OF THE PROJECT** 19](#_Toc139625810)

[**METHODS** 20](#_Toc139625811)

[- **Starting sequencing data** 20](#_Toc139625812)

[- **Data processing and methylation extraction** 24](#_Toc139625813)

[*Bisulfite sequencing:* 24](#_Toc139625814)

[*SMRT:* 27](#_Toc139625815)

[- **Gene association** 28](#_Toc139625816)

[- **Functional enrichment** 31](#_Toc139625817)

[- **Motif identification** 32](#_Toc139625818)

[- **GATC motif genomic analysis** 34](#_Toc139625819)

[- **Metabolic modelling** 35](#_Toc139625820)

[- **COMETS** 38](#_Toc139625821)

[**EXPERIMENTAL RESULTS & DISCUSSION (25-30 pages)** 41](#_Toc139625822)

[- **Data processing and methylation extraction** 41](#_Toc139625823)

[- **Gene association** 44](#_Toc139625824)

[- **Functional enrichment:** 48](#_Toc139625825)

[- **Motif identification** 50](#_Toc139625826)

[- **GATC motif genomic analysis** 52](#_Toc139625827)

[**APPENDICES** 55](#_Toc139625828)

[1) Bisulfite data processing commands 55](#_Toc139625829)

[**BIBLIOGRAPHY** 67](#_Toc139625830)

# 

# **ABSTRACT (1page)**

# **INTRODUCTION**

* **Epigenome/methylome definition**

The epigenome is the group of chemical adducts which may be attached to DNA as modifications of nucleotides or to protein scaffolds of the genome in a non-mutational way and modulate genomic heritable behaviours such as gene expression or genome replication (Chapin et al. 2022). Epigenomics definition often involves two or more heritable states that are maintained through positive feedback loops (Casadesús and Low 2006).

While epigenetic regulations have predominantly been studied in eukaryotes, where they play a role in cell differentiation and disease development, there is growing evidence supporting the existence of epigenetic regulation in prokaryotes (Payelleville and Brillard 2021). Various bacterial genera exhibit intricate developmental processes that involve cell differentiation. Examples include *Bacillus subtilis* forming spores, *Rhizobium* differentiating into nitrogen-fixing bacteroids, *Caulobacter* undergoing asymmetric cell division, *Myxococcus* forming fruiting bodies, cyanobacteria forming heterocysts, and many bacterial species engaging in biofilm formation. These developmental processes result in the formation of bacterial cells with distinct morphological and physiological characteristics, while the DNA sequence of the genome remains unchanged (Casadesús and Low 2013)

Among these regulatory mechanisms in prokaryotes DNA methylation is the most common instance. The methylome is the group of bases, from a genome, modified by the addition of a methyl group to an Adenine or a Cytosine by enzymes called DNA-methyltransferases (MTases). There are three types of MTases depending on the modification they catalyse: N6-methyl-adenine (m6A), C5-methyl-cytosine (m5C) or N4-methyl-cytosine (m4C). Although the three types are described in both archaea and bacteria, the m4C modification is not reported in eukaryotes. The DNA methylation profile is species-specific or strain-specific (Payelleville and Brillard 2021). In Figure 1 there are depicted the three DNA methylations found in bacteria (Beaulaurier, Schadt, and Fang 2019).

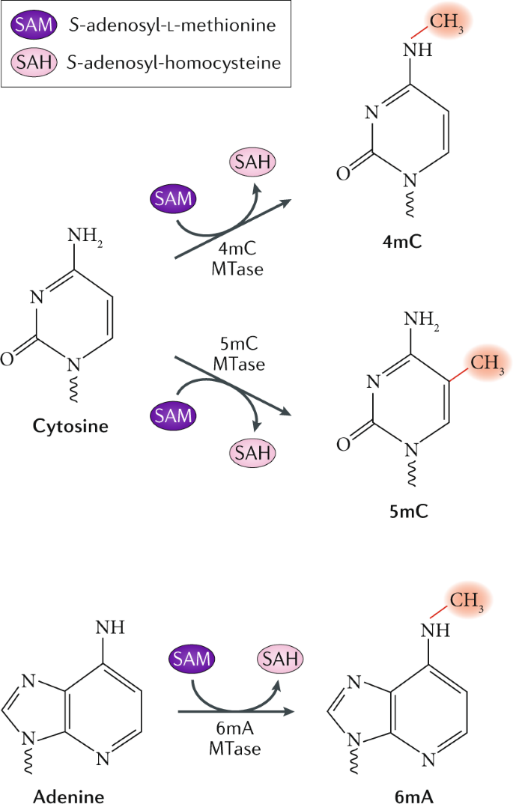


Figure 1. (Beaulaurier, Schadt, and Fang 2019)

* **Classification of MTases**

Bacterial DNA methyltransferases (MTases) are commonly functionally associated to a restriction endonuclease (REase) building the restriction-modification systems (RM), which are known to protect the bacterial cell from foreign DNA. (Payelleville and Brillard, 2021). However, it is suggested that restriction enzymes may have initially evolved to prevent the loss of DNA methyltransferases by post-segregational killing by selfish RM rather than solely targeting foreign DNA. The purpose of retention of MTases in genomes may be also due to their involvement in epigenetic regulation or their impact on gene flux and host genome composition (Sánchez-Romero and Casadesús, 2020; Oliveira and Fang, 2021).

Orphan DNA methyltransferases are those not coupled to a restriction domain and are found in many bacterial and archaeal genomes (Sánchez-Romero and Casadesús, 2020). The same MTase domain rarely appear as both, part of complete RM systems or orphans, indicating the loss of the cognate restriction enzyme or acquisition through horizontal gene transfer (Oliveira and Fang, 2021). Bacterial MTases have both de novo and maintenance activities (Sánchez-Romero and Casadesús 2020).

The classical types of RM systems, classified in Type I, Type II, and Type III, differ in structure, sequence recognition, cleavage position, and cofactor requirements. Additionally, Type IV "restriction systems" consist of only restriction enzymes that cleave modified recognition sites (Oliveira and Fang 2021). Regarding type II, the restriction and modification domains are, in most cases, able to operate separately and independently from each other (Roberts and Macelis n.d.).

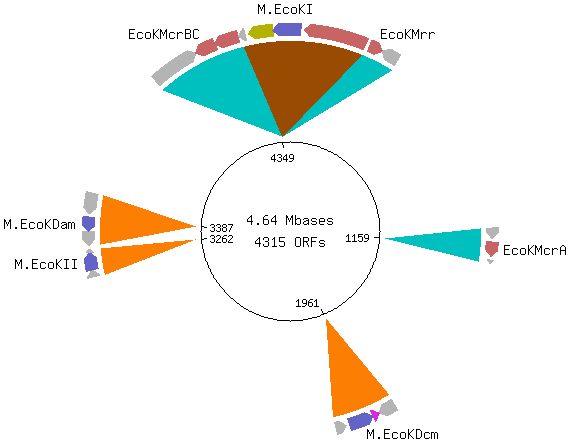
* **MTases in *Escherichia coli* K-12 substr. MG1655**

Escherichia coli K-12 substr. MG1655 possesses several DNA methylation systems. It has one type I RM system called **EcoKI**, which targets the recognition sequence AACN6GTGC and performs m6A modification. Additionally, it harbors three different type IV systems, **EcoKMcrA**, **EcoKMcrBC**, and **EcoKMrr**, which require methylation for their function (Roer, Aarestrup, and Hasman, 2015).

This *E. coli* strain also contains three type II systems consisting of orphan MTases, namely **Dcm**, **Dam**, and **MEcoKII**. These MTases recognize the sites CCWGG, GATC, and ATGCAT, and perform m5C, m6A, and m6A modifications respectively. While m4A modification is observed in this strain, the specific MTase responsible for this modification has not been definitively identified yet (Roberts and Macelis, n.d.).

* **Methylation functions in bacteria**

Figure 2. Escherichia coli K-12 sub MG1655 DNA methylases encoded on its genome. REBASE



Bacterial DNA methylation plays crucial roles in various biological processes, including genome defence, chromosome replication and segregation, nucleoid organization, control of cell cycle, DNA repair, transcription regulation, and secretion of virulence factors. In gammaproteobacteria, DNA methylation has also been found to control transposon activity (Sánchez-Romero and Casadesús 2020).

DNA methylation can occur on various motifs throughout the DNA, but in bacterial epigenetic regulation the typical examples studied involve promoter regions associated with m6A (by conserved Dam in Enterobacteria) (Figure 4b) (Payelleville and Brillard 2021).

In gammaproteobacteria, Dam methylase is synthesized throughout all stages of growth, and the hemimethylated GATC sites are brief since Dam trails the replication fork closely, being base methylation post-replicative (Sánchez-Romero and Casadesús 2020). Given the rapid occurrence of remethylation of Dam sites during chromosome replication, it is improbable for it to function as a prevalent mechanism for transcriptional regulation (Adhikari and Curtis 2016)

Moreover, DNA-binding proteins can distinguish between methylated and hemimethylated states at their target sites. The methylation state can affect the interaction between the proteins and DNA, either directly through steric hindrance or by inducing changes in DNA topology, leading to epigenetic regulation (Payelleville and Brillard 2021). In some cases, the binding of specific proteins to certain sites can prevent DNA methyltransferase activity, resulting in passive loss of methylation on the daughter strand (Figure 3b). Transcription factors such as Lrp, OxyR, Fur, and HdfR are examples of proteins involved in the "passive" erasure of methyl groups (Figure 4b). In a nutshell, the binding pattern of factors is influenced by the methylation state, and in turn, the methylation pattern can influence protein binding (Sánchez-Romero and Casadesús 2020).

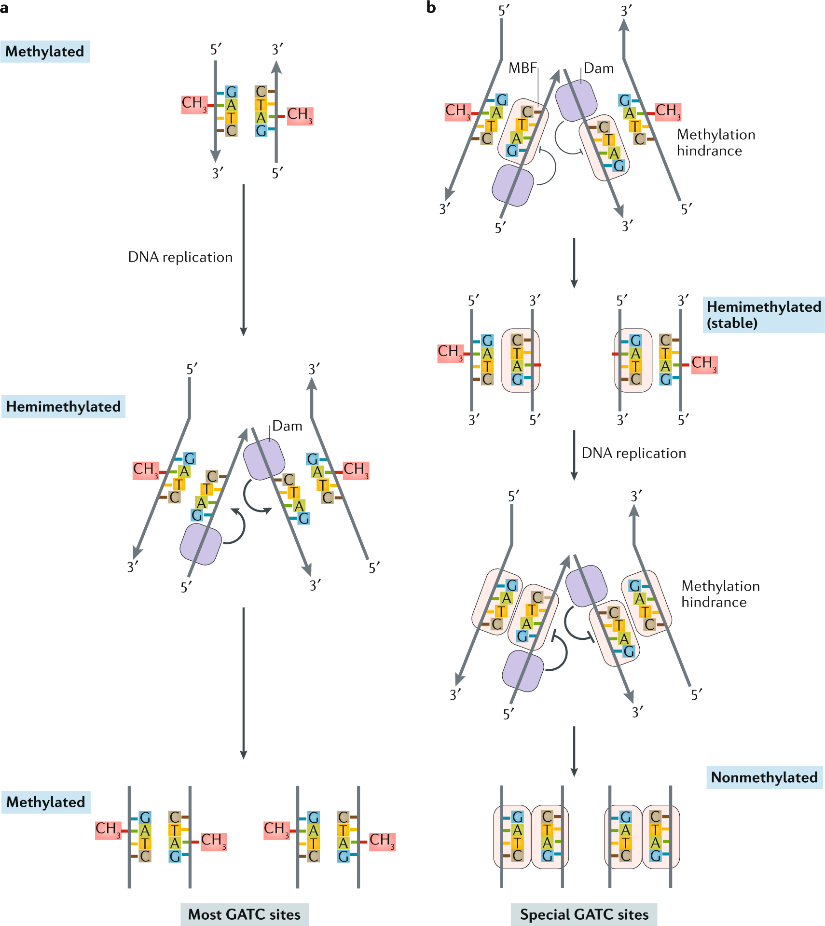


Figure 3. Transference of methylation beyond DNA replication (Sánchez-Romero and Casadesús 2020). Methylation-blocking factor (MBF)

A concise example of this phenomenon occurs in uropathogenic *E. coli*; the production of Pap pili is regulated by phase variation, which involves switching between an ON state where Pap pili are expressed and an OFF state where they are not. This regulation is controlled by Dam methylation and the binding of leucine-responsive regulatory protein Lrp to the regulatory region of the pap operon. This region contains six binding sites for Lrp, with sites 1-3 (having the GATC-I site) located upstream of the papB gene and sites 4-6 (having the GATC-II site) located upstream of sites 1-3. In the OFF state, specific GATC-I site is fully methylated and GATC-II unmethylated. Lrp preferentially binds to unmethylated sites, so in the OFF state, it binds to sites 1-3, preventing the expression of the pap operon by inhibiting RNA polymerase binding. Lrp binding also prevents Dam methylation, maintaining the unmethylated state of a GATC-II site, thus reinforcing the OFF state (Figure 5a). During chromosome replication, Lrp is released from the DNA and GATC-I site transitions from fully methylated to hemimethylated allowing Lrp to switch its binding from sites 1-3 to sites 4-6 (Figure 5b). The regulatory protein PapI enhances Lrp's affinity for sites 4-6, facilitating the switch to the ON state. This switching also allows Dam methylation of the previously unmethylated GATC-II, further inhibiting Lrp binding to sites 1-3 and promoting the ON state (Figure 5c). For the pap operon to be expressed, cAMP-CAP must bind to a specific region upstream of the operon (Figure 5d). Pap operon expression includes papB, which in turn, stimulates the expression of papI, promoting Lrp translocation to sites 4-6 and reinforcing the ON state (Figure 5e). The transition to the ON state depends on the levels of cAMP-CAP, PapI, and the replication status of the chromosome. Phase variation is a reversible process, but the transition from the ON to the OFF state occurs much more frequently than the reverse transition. This results in only a small number of *E. coli* cells in a population expressing pap pili. The specific mechanism behind the switch from ON to OFF is not well understood but involves chromosome replication, leading to the displacement of Lrp-PapI from the GATC-I site. Lrp exhibits a slightly higher affinity for the sites 1-3. The low expression of pap pili in the population is important for evading the immune system, as these pili are highly immunogenic. However, they are necessary for the colonization of the upper urinary tract. Thus, maintaining a low level of expression allows for colonization opportunities while minimizing the immune responses (Adhikari and Curtis 2016).

Figure 4. a) Phase variation due to replication in an imaginary bistable locus. b) specific loci examples (Sánchez-Romero and Casadesús 2020)

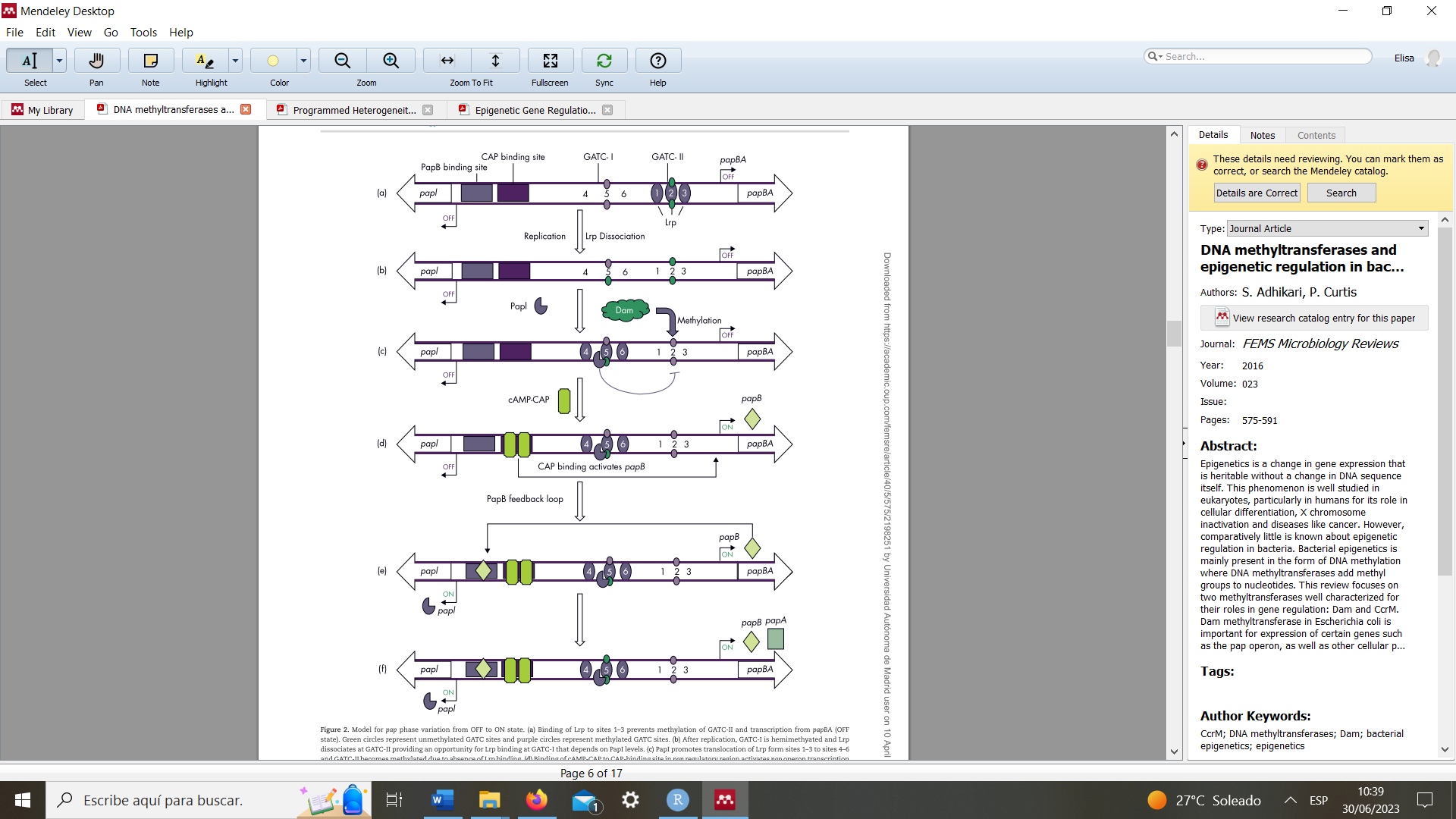


Figure 5. Model for pap phase variation from OFF to ON state.

DNA methylation plays a significant role in regulating gene expression, as evidenced by altered gene expression patterns in strains lacking or overproducing DNA methyltransferase. However, it is important to note that changes in gene expression may not always indicate direct transcriptional control, as indirect effects are possible. In some cases, post-transcriptional regulation by Dam-dependent methylation has been observed, suggesting that methylation may influence mRNA stability, translation, or other post-transcriptional events (Sánchez-Romero and Casadesús, 2020). However, the precise mechanism and the involvement of 5′GATC3′ sites on genomic DNA in these cases remain unclear (Gao et al., 2023).

In a direct fashion, DNA methylation commonly leads to transcriptional repression, while demethylation, either single-stranded or double-stranded, often leads to transcriptional activation (Sánchez-Romero and Casadesús 2020). Nevertheless, other studies support that **undermethylation of DNA, specifically in genes with GATC clusters within the coding sequence, can lead to a decrease in gene expression**. This decrease is not dependent on DNA-binding proteins but is based on the physical principle that hemimethylated or unmethylated DNA has a higher melting point. Hénaut et al. (1996) proposed this hypothesis suggesting that undermethylation of GATC clusters at low temperatures blocks or decreases transcription of genes in *E. coli*. This mechanism allows *E. coli* to slow its growth and enter the stationary phase when transitioning from a high-temperature environment (the gut) to a low-temperature one (outside the gut). Notably, this hypothesis does not require the involvement of DNA-binding proteins for methylation-dependent gene expression regulation. Therefore, considering these premises and all the putative methylation motifs that a transcription unit can have on its gene body or on regulatory sequences around, the prediction of methylation’s effect on a specific gene’s expression is beyond our capabilities. This information sustains the rationale used in the epigraph *Methods, Gene association and functional enrichment*.

The comparison of E. coli and Salmonella genomes revealed the presence of at least 23 genes in both organisms that contained conserved GATC clusters within their coding sequences. The conservation points to the fact that they serve a specific function. Furthermore, the fact that they are located in the coding region of the genes appears to support the hypothesis proposed by Hénaut et al. (1996). These genes are involved in various cellular processes, including amino acid metabolism, nucleotide metabolism, replication, and other cellular functions. Dam, would not play a role in global gene expression but only on specific functionalities (Adhikari and Curtis, 2016). (in the results it is seen the enrichment in some of these gene call and functional analysis and GATC cds enrichment)

Kahramanoglou et al., (2012) evidenced from a cytosine methylation-deficient mutant that this modification in *E. coli* regulates gene expression during the stationary phase, proved by increased expression of the stress response sigma factor RpoS and its targets.

The contribution of nucleoid associated proteins, such as H-NS, in epigenetic mechanisms involving DNA-methylation has also been reported in various cases and therefore also contribute to the formation of bacterial subpopulations (Payelleville and Brillard 2021)

## **Phenotypic heterogeneity**

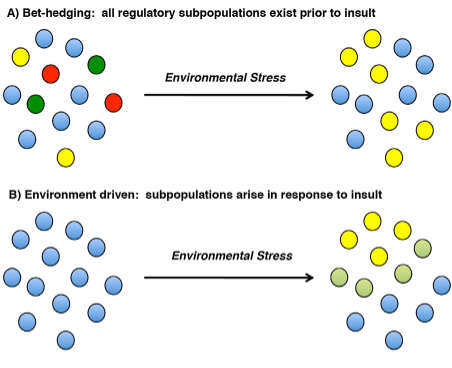
Living organisms and in particular bacteria have the capacity to change their metabolic state due to different phenomena, this is Simons (2011) called **phenotypic plasticity**, an adaptative trait for which phenotype-fitness association is predictable across environments. Bacteria may respond to changes in their environment, such as nutrient availability or stressors, by switching between different phenotypic states as execution of their inherent regulatory programs. If all the cells in the homogeneous population respond in the same way to a certain stimulus, this behaviour is named **responsive switching**. While subpopulation formation can be viewed as this, it frequently relies on stochastic events, either responding to environmental cues (**responsive diversification**) or merely stochastic in absence of environment fluctuations (**stochastic switching**) (Grimbergen et al. 2015). For example, random fluctuations in gene expression coming from factors affinity on DNA which, in turn, affects the methylation state. This is often referred to as "noise," and it can generate signals that propagate via feedback loops, both positive and negative, and facilitate the establishment of epigenetic lineages. These stochastic processes contribute to the emergence of diverse phenotypic states within bacterial populations, allowing for increased adaptability and survival in varying environmental conditions (Casadesús and Low 2013). The phenomenon of phenotypic heterogeneity in several major prokaryote pathogens plays a crucial role in the success of their infections (Payelleville and Brillard 2021).

Another phenomenon that causes phenotypic heterogeneity is **adaptative tracking**. It is the most studied mode of evolutionary response. The genetic variance and mutational rate create subpopulations that are subjected to fitness selection by the environment pressure. Optimal trait values and associated allele frequency variate constantly and natural selection continuously eliminates suboptimal forms that were previously well adapted (Simons 2011). This type of heterogeneity is however is not centred around isogenic organisms thus it is not within the scope of this thesis aim. On the contrary, the thesis revolves around the long-term adaptative advantages that isogenic heterogeneity can provide to microorganisms, specifically the phenomenon of bet-hedging, and how methylation on DNA affects on developing this situation.

In bacteria certain loci exhibit bistability, meaning they can switch between alternative states (ON/OFF) in a reversible manner controlling gene expression. This phenomenon, known as phase variation, generates phenotypic cell variants (Sánchez-Romero and Casadesús, 2020). Bacteria employ various mechanisms for phase variation, including slipped-strand mispairing, site-specific recombination, and DNA methylation. The first two mechanisms involve altering the DNA sequence itself, while DNA methylation is an epigenetic regulation method that is not dependent on the DNA sequence. Since the remethylation of Dam sites occurs rapidly during chromosome replication, it is unlikely to serve as a widespread mechanism for transcriptional regulation (Adhikari and Curtis 2016). However, the transition from fully methylated to hemimethylated DNA or back to fully methylated DNA has been observed in certain loci in *E. coli* controlled by Dam-dependent methylation (Figure 4a) (Sánchez-Romero and Casadesús, 2020).

The frequency of switching is influenced by the affinity of DNA-binding proteins for their binding sites and requires DNA replication (Sánchez-Romero and Casadesús, 2020). The binding events after DNA replication determines, at each locus, the subsequent gene expression of the emerging cell, leading to divergence and the emergence of distinct phenotypes. This phenomenon scaled to the whole genome may give rise to a great phenotypic heterogeneity driven by methylation. This can occur as part of a developmental program, such as sporulation, where different genetic expression patterns guide cell differentiation. Additionally, due to stochastic binding, it can result in the bifurcation of an isogenic population into two or more reversible metabolic states (Casadesús and Low 2013).

In summary, bacterial populations exhibit heterogeneity at both the genetic and phenotypic levels. This heterogeneity can provide a selective advantage during environmental changes. Microorganisms employ two strategies for phenotypic heterogeneous adaptation: bet-hedging, where certain individuals express genes to survive stressors before they occur (Figure A), and heterogeneity arising from distinct gene expression profiles triggered by localized environmental cues (Figure B). Heterogeneity ensures survival and reproductive success in diverse conditions (Davis and Isberg 2016).



*Figure 1. Mechanisms that promote population phenotypic heterogeneity. A: bet-hedging: where heterogeneity exists prior to an environmental change or B: environment-driven mechanisms (Davis and Isberg 2016).*

* **Bet-hedging**

Bet-hedging is a survival strategy where organisms develop different **maladapted phenotypes** within clonal populations to increase **long-term fitness** in unpredictable environments. It involves spreading risks and optimizing population fitness by minimizing fitness variance over time and **maximizing geometric mean fitness** across fluctuating conditions. Bet-hedging is a case example of microbial phenotypic heterogeneity in which many subpopulations having slightly different/intermediate metabolic states which would be **specialized in different environments** regardless of the current conditions so as to be ready at population level to endure sudden changes. This phenomenon can arise from responsive diversification or from stochastic switching. This strategy differs from developing an intermediate phenotype that is moderately fit in all conditions but never a specialist (Grimbergen et al. 2015).

Depending on the robustness of **regulatory circuits** in bacteria, they can trigger a response in consequence of environmental cues or lead to stochastic switch triggered by noise. The **robustness** of a regulatory pathway comes with high **costs of maintenance**. The **environmental variability and its frequency** determine the balance between costs and benefits of responsive switching versus maladaptation stochastic phenotype and consequently the evolution of the system is bound to become more or less robust. In situations where environmental fluctuations are rare, stochastic switching is an advantage compared to responsive switching (Grimbergen et al. 2015).

Simons (2011) introduced categories of evidence to assess whether observed **phenotypic heterogeneity** qualifies as a bet-hedging strategy. The highest category, **category VI**, requires studies to demonstrate two features: a) the bet-hedging trait increases **fitness** compared to a non-bet-hedging alternative in a fluctuating environment and b) the phenotype **switching** rate (heterogeneity level) should **correlate** with the frequency of **environmental** **fluctuating** selection, which shows to be an adaptative trait. Currently, no study on microbial phenotypic heterogeneity has provided category VI evidence, indicating that true bet-hedging by microorganisms has not been convincingly observed yet (Simons 2011).

*Bacillus subtilis* **sporulation** is a survival response to starvation, where highly resistant endospores are formed at the cost of reduced reproductive offspring. This phenotype switching is driven by a **bistable regulatory network** that incorporates environmental information and stochastic switching through regulatory noise. Stochastic initiation of sporulation ensures survival in case responsive sporulation is too slow or the environment is abruptly harmful. This complexity highlights the challenge of studying bet-hedging in microbial systems, where both **responsive and stochastic switching can coexist** to form spores due to a not-so-robust sensory system. When environments provide partially reliable signals about future events, it is predicted that a stable evolutionary strategy would involve a combination of deterministic and stochastic switching strategies (Grimbergen et al. 2015).

Bacteria can survive antibiotic treatment through two main mechanisms: acquiring immunity via mutations or gene acquisition, and entering a dormant state called **persistence**. During persistence, cells become metabolically inactive, reducing their vulnerability to antibiotics. These dormant or persister cells can later resume growth, generating antibiotic-sensitive offspring. Strains/individuals with lag-times exceeding the duration of antibiotic exposure were selected for. Importantly, the extended lag-time trait is heritable and encoded in the DNA, conferring tolerance to various antibiotics (Grimbergen et al. 2015). Persisters, may be formed through both stochastic and deterministic processes. In clonal *E. coli* cultures during mid-log phase, only a tiny portion exhibit persister traits, indicating randomness. However, as the growth phase shifts towards stationary, there is a significant rise in persister levels across different species, indicating a deterministic influence (Lewis 2010).

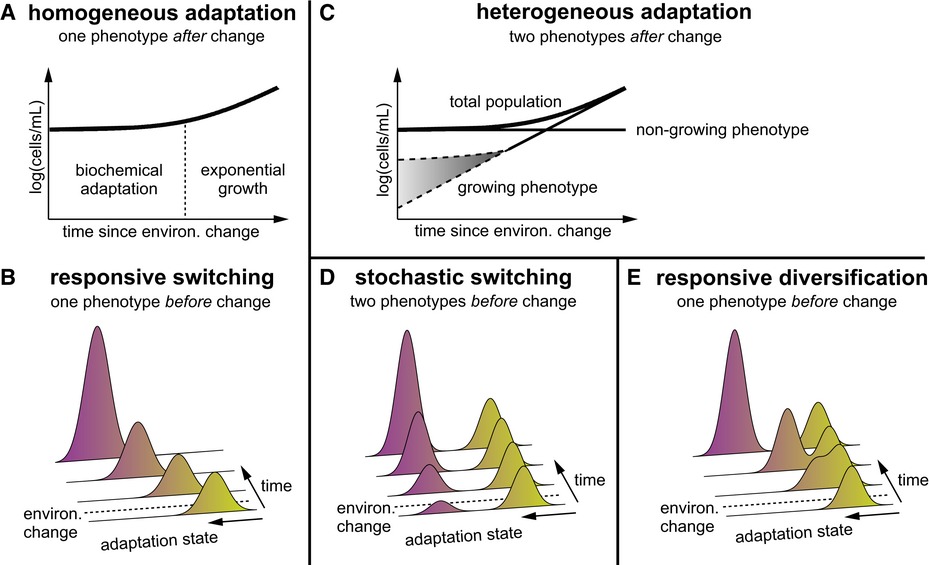
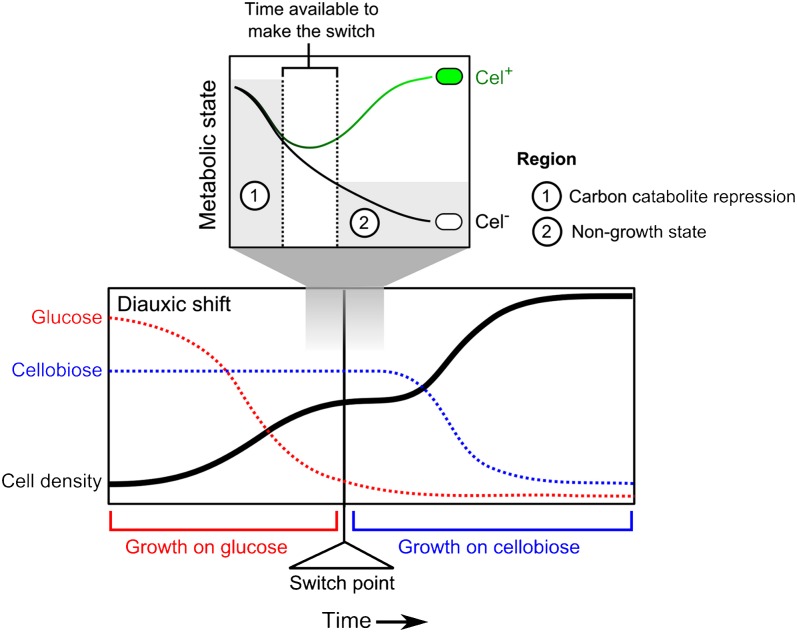


Figure 6. (Kotte et al. 2014)

Bet-hedging may be an important strategy in carbon metabolism. When bacteria are introduced to a new environment, there is an initial lag phase before they adapt and start growing on the new carbon source. Traditionally, this lag phase was thought to be due to individual cells undergoing biochemical adaptations uniformly (Figure 6A,B). However, an alternative hypothesis suggests that the lag phase upon shifting media on an isogenic population is caused by the emergence of two phenotypes: a minority growing and a main non-growing. This subpopulation can arise from the homogeneous isogenic population through either stochastic switching before the environmental change (Figure 6C,D) or responsive diversification as consequence of the new conditions (Figure 6C,E). It was demonstrated thought 13C-labeled acetate that after glucose-gluconeogenic substrate (acetate) shifts in *E. coli* responsive diversification occurs. Non-growing cells enter a dormant state but remain viable after the carbon source shift. Spontaneous suppressor mutations were discarded as the cause of diversification, as repeating the experiment with cells from the growing subpopulation still resulted in two distinct phenotypes amounting an identical growing population ratio (Kotte et al., 2014).

In the same study, they observed that the presence of acetate before the shift did not prepare the cells for later acetate consumption. However, the concentration of extracellular acetate after the shift had a significant positive impact on both the growth rate and the fraction of adapting cells, α, and suggested a connection between extracellular acetate levels, substrate uptake rate, and central metabolism. To verify this, they did glucose-fumarate shift, which—opposed to acetate—is actively transported. By increasing the expression of the fumarate transporter, they were able to rise α (Kotte et al. 2014).

Figure 7. (Solopova et al. 2014)



Carbon catabolite repression (CCR) is responsible for suppressing metabolic pathways related to non-preferred carbon sources. During metabolic shifts, there are instances where certain cells fail to respond promptly, leading to responsive diversification. In the case of glucose-cellobiose diauxie in *Lactococcus lactis*, two distinct populations arise: one capable of metabolizing cellobiose (Cel+) and one unable to do so (Cel-). It is speculated that the non-growing Cel- subpopulation was in allow energetic state during the shift and induced the stringent response in order to be maintained to facilitate future utilization of alternative carbon sources, as it was seen with a latter shift to galactose. Recent research suggests that the level of heterogeneity in catabolite repression can undergo evolutionary changes, resulting in the emergence of "generalist" or "specialist" cells. These findings underscore the adaptability and evolvability of response infidelity in microbial systems (Solopova et al. 2014).

Researchers found that cells precultured with cellobiose exhibited a shorter lag phase during the transition from glucose to cellobiose medium compared to cells precultured with glucose. This suggests the involvement of an epigenetic mechanism in the metabolic switching process. Furthermore, epigenetic memory has been observed in the expression of the lac operon in Escherichia coli, where the memory effect lasted for two cell generations (Solopova et al., 2014) (Robert et al. 2010) continue reading this paper.

* **Epigenetic regulation networks**

coupling of transcriptome analysis with extensive methylome analysis. It may lead to the identification of putative epigenetic regulation networks combined methylome/transcriptome strategy

has been rarely employed. When two subpopulations are present, each one displaying a particular DNA-methylation pattern and a particular transcription pattern. Classical tools (e.g., WGBS, SMRT sequencing and RNA-seq analysis) allow the detection of the bulk. To distinguish the two subpopulations, single cell tools need to be applied (e.g., SMALR for DNA methylation and Record-seq or PETRI-seq for transcription). this combined methylome/transcriptome strategy has been rarely employed (Payelleville and Brillard 2021).

## **Genome scale metabolic modelling**

**Metabolic engineering** involves the modification of metabolic and cellular networks with the ultimate goal of cellular components concentration variation. It considers the entire **metabolic and regulatory network** when identifying targets as opposed to genetic engineering which focuses only on the ultimate protein/metabolite concentration. This latter approach has several limitations of describing cellular phenotypes. Instead, metabolic engineering centres on analysing **metabolic fluxes**, the rates of reactions in metabolic pathways, and overcomes some of the limitations. In order to quantify the metabolic fluxes **mass-balance-based** is the most utilized kind of **modelling**. It is based on **stochiometric** mass balances around intracellular metabolites under the **pseudo-steady state** assumption (fluxes are constant). This approach enables the construction **of genome-scale metabolic models (GEMs)** used for simulations (Kim et al. 2008).

“A GEM **computationally** describes a whole set of stoichiometry-based, mass-balanced metabolic reactions in an organism using **gene-protein-reaction (GPR)** associations that are formulated on the basis of **genome** **annotation** data and **experimentally** obtained information” (Gu et al. 2019). GPR associations are Booleans OR/AND that show the requirements of genes for a reaction to be possible (Ines and Bernhard 2010). They establish a direct and formal connection between the genotype (gene) and phenotype (protein) within a genome-scale reconstruction. It serves as a link that associates a gene (G) with the corresponding protein (P) responsible for catalyzing a specific reaction (R) in the biological network (Monk et al. 2017). Different associations are possible depending on the structure of the enzyme and the redundancy of the genome in terms of presence of isozymes. The model represents the genes/enzymes, reactions, and metabolites in the cellular network, capturing their *in vivo* relationships. It serves as a validated network that can describe various **metabolic phenotypes** by combining different metabolic fluxes (Kim et al. 2008).

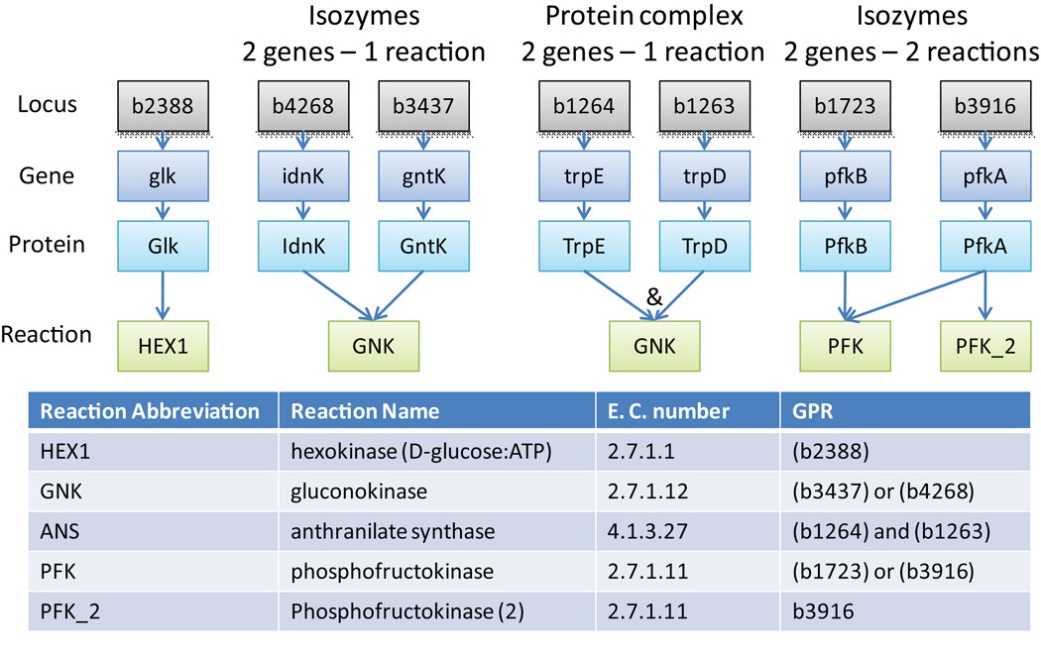


Figure 2. Examples of GPR associations and their representation in Boolean format are shown for E. coli (Ines and Bernhard 2010).

The **stoichiometric matrix** (S: n x m) is a representation of metabolic reactions, with each row representing a unique **metabolite** and each column representing a **reaction**. The entries in the matrix indicate the stoichiometric coefficients of the metabolites involved in each reaction. The stoichiometric matrix is typically sparse, as most biochemical reactions involve only a small number of metabolites. The fluxes through the reactions are represented by the **vector v** (m x 1), while the concentrations of metabolites are represented by the vector x (nx1). In a **steady state**, where the rate of change of metabolite concentrations (dx/dt) is zero, the product of the stoichiometric matrix and the flux vector equals zero **(S\*v=0)**. This equation imposes a constraint (Orth, Thiele, and Palsson 2010). Within the GEM, it is important to define the boundary of the metabolic network and allow the model to capture the import and export of metabolites. For this purpose, **boundary reactions** are set. These are unbalanced pseudo-reactions which serve a modelling purpose by adding or removing metabolites in the system, but they do not have a direct biological basis. They are included in the model to facilitate modelling and analysis, rather than representing actual biological processes (Documentation for COBRApy — cobra 0.26.2 documentation n.d.). **Exchange** reactions represent reversible reactions that involve the addition or removal of extracellular metabolites from the extracellular compartment. They define the culture media. These reactions facilitate the exchange of metabolites between the model system and its external environment (transport). On the other hand, **demand** reactions are irreversible reactions that consume intracellular metabolites within the model. They represent the utilization or consumption of specific metabolites by the cellular processes. Additionally, **sinks** are similar to exchange reactions, but they specifically involve the reversible addition or removal of intracellular metabolites. Sinks serve as pathways for the internal transport or utilization of metabolites within the model (Ines and Bernhard 2010).

Figure 8. S onstruction of stoichiometric matrix for a model metabolic network and constraints-based flux analysis. (Kim et al. 2008).

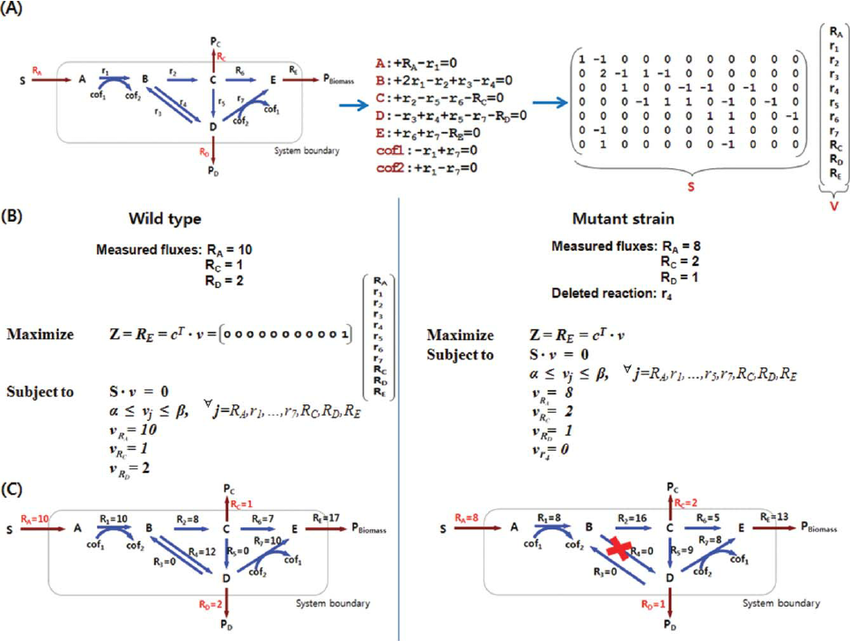
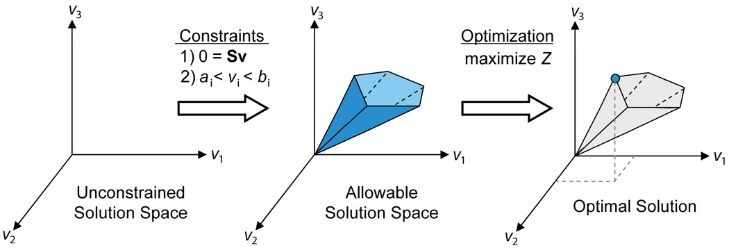


Figure 9. Definition of system boundaries (Ines and Bernhard 2010).

Figure 10. Solution space in constraint-based modelling. (Orth, Thiele, and Palsson 2010).



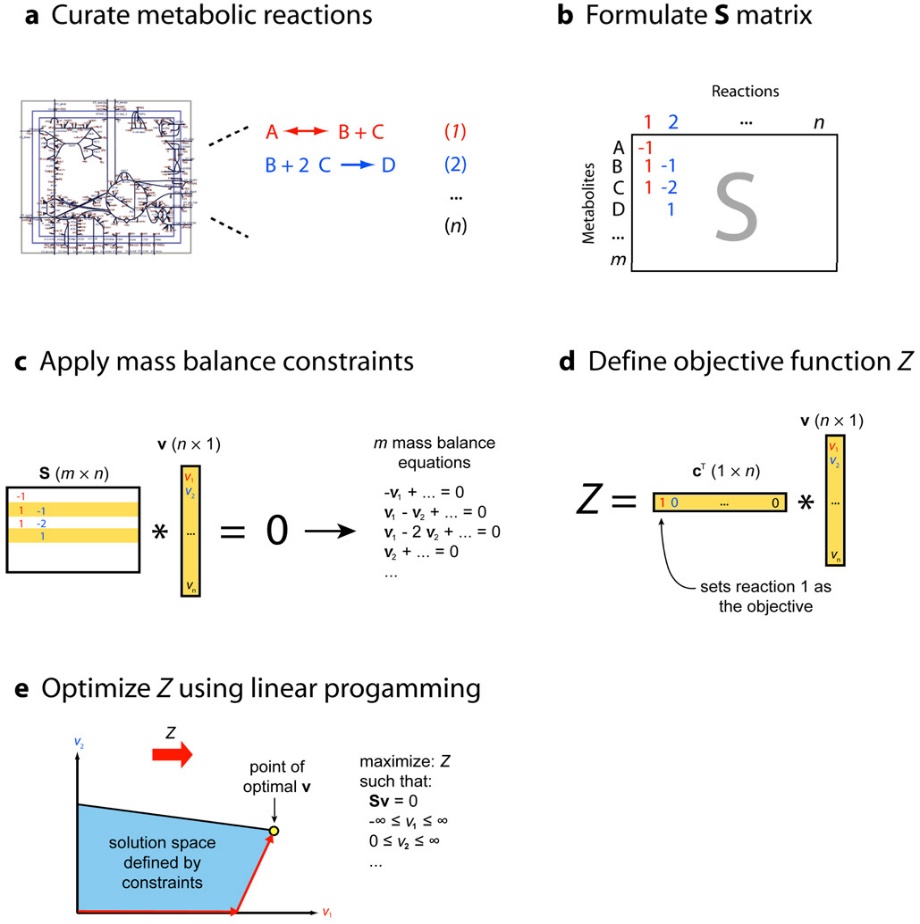
The **solution space** encompasses all physiologically possible states, with each point in it representing a plausible solution for flux (v) combinations within the system (Fig. 10). The more constraints are imposed the further reduced so as to make the simulation results become more realistic. Therefore, this methodology is called **constraint-based flux analysis**. These constraints are expressed as equations that balance the inputs and outputs of reactions and as inequalities called **bounds** that define the reaction reversibility. These bounds are provided into two vectors: the lower bound (lb) and the upper bound (ub) with values from -∞ to 0 and from 0 to +∞ respectively. If a flux is known the lb and ub coincide with the same value for that reaction (Orth, Thiele, and Palsson 2010). Additional constraints come from:

* Transcriptional regulatory constraints can be incorporated into the metabolic network. This integration allows for the description of biological correlations between genes, proteins, and reactions using Boolean logics such as AND, OR, and NOT. Genes in the **transcriptional regulatory network** are assigned **binary** values based on their expression levels under specific conditions and so their corresponding reactions based on GPRs (Kim et al. 2008).
* Metabolic flux data obtained from experiments, particularly those utilizing **13C-based techniques**, can be used to constrain the solution space of metabolic models. By incorporating these experimentally derived flux measurements, the range of possible flux distributions is narrowed down, resulting in a more accurate and constrained solution space (Kim et al. 2008).

Even with many constraints, in GEMs there are more unknown variables than equations, leading to an **underdetermined system** of equations. As a result, there is no unique solution to the system but a constrained solution space, with multiple sets of flux values can satisfy the stoichiometric constraints (Orth, Thiele, and Palsson 2010).

In metabolic modelling, a phenotype is defined by a biological objective which in turn, can be described by a mathematical **objective function** (Z). The solution to the system is achieved with the **optimization** of Z (Orth, Thiele, and Palsson 2010). Depending on which constraints are used (inputs), the Z, the desired output (metabolic fluxes, gene expression status, bottleneck reactions to be removed, gene knock-out targets or up/down- regulation targets for biochemical production) and other particularities for the resolution, we can find many different algorithms. (Kim et al. 2008).

Figure 11. (Orth, Thiele, and Palsson 2010)



In **Flux Balance analysis algorithm (FBA**), the objective is to maximize or minimize Z=cTv that represents a **linear combination of fluxes**. The vector c contains weights that determine the importance of each reaction in the objective function. Most times the objective function of interest is the biomass production. To mathematically represent biomass production, an artificial **"biomass reaction"** is added to the stoichiometric matrix. The biomass reaction consumes precursor metabolites based on experimental measurements of biomass components (Orth, Thiele, and Palsson 2010). FBA is the simplest of the algorithms and utilizes **linear programming** techniques to simulate and predict metabolic flux values. This approach enables the systematic evaluation of the effects of genetic and environmental perturbations on cellular metabolism at a global scale. In FBA simulation of gene/reaction knock-outs are settled as a new constraint with bounds set to zero and latter identical resolution (Kim et al. 2008).

While applying FBA to a metabolic network, it is important to note that a unique optimal solution may not always be obtained, i.e., a single point that represents a vector with the metabolic fluxes. Instead, a collection of points along an edge of the solution space, i.e., a **degenerate solution**. Moreover, optimal results obtained through constraint-based flux analysis may differ from real cellular metabolism, potentially operating in a **suboptimal** manner. In FBA, linear programming solver gives back only one solution regardless existing alternate optima (Mahadevan and Schilling 2003). To address this issue **Flux Variability Analysis (FVA)** quantifies the feasible ranges of reaction fluxes (v) within the network, considering both optimal (µ=1) or sub-optimal (µ< 1) objective value levels obtained from FBA. It is carried out in two phases, firstly a unique optimization finds the maximum **objective value, Z0** and secondly, with this fixed objective value and n reactions in the model**, 2 x n optimizations** are done (minimization and maximization) to calculate the **ranges** **of** **fluxes** in the metabolic network, vi. This information can be used to identify important metabolic reactions and explore the system's capabilities, such as network redundancy, under specific simulation conditions. FVA provides a broader understanding of the metabolic network's behaviour and aids in various analyses related to metabolic fluxes (Kenefake et al. 2022).

Advancements in the use of GEMs have contributed to various applications, such as optimizing strains for bio-based chemical production, targeting drugs in pathogens, predicting enzyme functions, conducting panreactome analysis, modelling interactions between multiple cells or organisms, and enhancing our understanding of metabolism in different species (Gu et al. 2019).

* **Dynamic flux balance analysis (dFBA)**

GEMs and FBA can be extended **to microbial communities**, taking into account their interactions and environmental impact, in order to simulate consortia growth including the prediction of emergent features. Computation Of Microbial Ecosystems in Time and Space (COMETS) software utilizes dynamic flux balance analysis (**dFBA**) to simulate multiple microbial species in complex and **spatiotemporally** structured environments. This approach considers steady-state metabolism within cells while treating species abundance and environmental metabolites as dynamic variables. In dFBA, resource uptake (exchange reactions) is estimated using a Michaelis-Menten equation, taking into account the media concentration of resources (Dukovski et al. 2021).

# **AIM OF THE PROJECT**

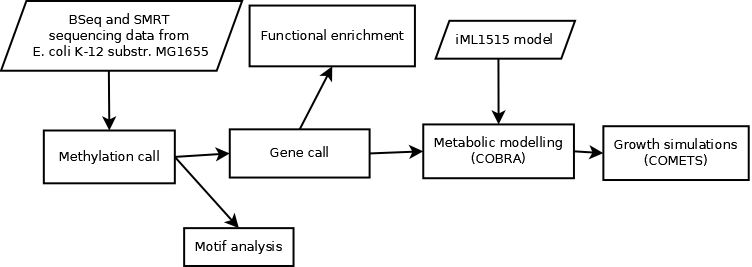
As stated in the introduction, bacterial isogenic populations usually come up with different forms of phenotypic heterogeneity, this capability provides organisms to adapt to environmental conditions as it is an adaptative trait. The underlying mechanisms are diverse, from sensory systems with more or less robustness, signalling pathways which ultimately control gene expression to epigenetic regulations can be involved in this variability. The aim of this study is to observe the span of DNA-methylation variability within a clonal *Escherichia coli* K12 MG1655 population to extract the corresponding potentially regulated transcription units.

The effect of methylation at each locus on gene expression is uncertain and needs further information about the surrounding genome state (methylated/unmethylated) and binding factors affinity to predict whether it favours or hinders transcription. To achieve this, it is necessary to perform single-cell sequencing techniques, which in bacteria are still a field left behind, therefore not so much data is available. When doing bulk sequencing, we can speak about percentage of methylation in each base which means the proportion of the sequenced population that has the locus methylated. As the goal is to study heterogeneity, the project will be focused only on loci that have between 25-75% of methylation which may be causing the phenotypic variability. Loci that are below or above that range are excluded because they represent homogeneity within the population. Genes associated with several of these selected loci, both on regulatory intergenic areas or on coding regions, are candidates of being either silenced or upregulated with respect to the average/bulk expression, which is considered the wild-type.

Modification on gene expression ultimately affects fluxes of metabolic reactions. Here entries the modelling part where I will identify the reactions associated to methylation regulation that, when modifying their fluxes, have negative impact on the growth of the wt. The rationale is that these reactions are not optimally working in the medium where *E. coli* is growing, however, they prepare the population for a sudden change in the medium. As a prove of concept we can translate this idea into a simulation where it is represented manually the heterogeneity of an isogenic population by laying out different proportions the wt model and one or more models that differ on the flux of the indicated reactions. Each model would display a possible metabolic state of the *E. coli* K12 MG1655 strain. By simulating their growth in a changing environment, we are representing a possible bet-hedging strategy.

A final task would be to characterize the methylated loci, searching for the motifs that are found to be methylated, and if there is any relationship between percentage of methylation and degeneracy or genomic region associated.

# **METHODS**



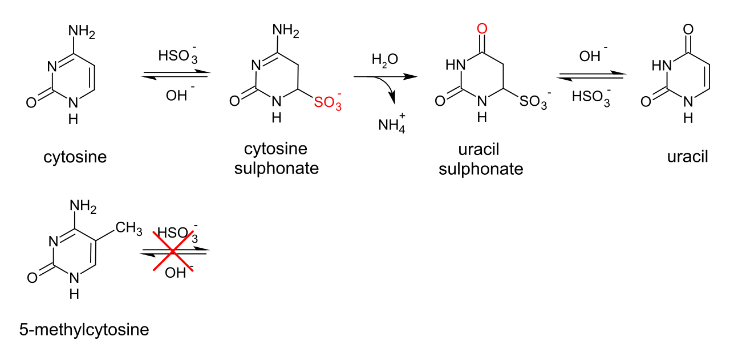
* **Starting sequencing data**

The strain of choice to study was *Escherichia coli* K-12 substr. MG1655. Its genotype is F- lambda- *ilvG*- *rfb*-50 *rph*-1. The serotype of this strain is OR:H48:K-. This **strain closely resembles the wild-type *E. coli*** and has undergone minimal genetic manipulation. It was primarily modified to eliminate the temperate bacteriophage lambda and F plasmid through exposure to ultraviolet light and acridine orange, respectively. The listed mutations in the genotype are commonly found in most K-12 strains and are believed to have occurred early in the history of the laboratory strain. Rph gene is involved in the processing of ribonucleic acid (RNA), a frameshift at the end of it, leads to reduced pyrE expression and mild pyrimidine starvation, resulting in the strain growing 10 to 15% slower in pyrimidine-free medium compared to medium containing uracil. The ilvG- mutation is also a frameshift that disrupts acetohydroxy acid synthase II, an enzyme associated with the biosynthesis of isoleucine. Additionally, the rfb-50 mutation involves an insertion of IS5, leading to the absence of O-antigen synthesis. This initial *E. coli* strain K-12 was isolated from a stool sample of a diphtheria patient in California, in 1922. From a stab-culture of that one, W1485 strain was derived in Joshua Lederberg's lab. Finally, from W1485 the strain MG1655 was derived and named by Mark Guyer (E.coli Genome Project, 2023).

The scope of this project was fully computational, the employed starting data was obtained from already sequenced by other groups that satisfied the wild type strain requirements and that were grown on glucose as carbon source.

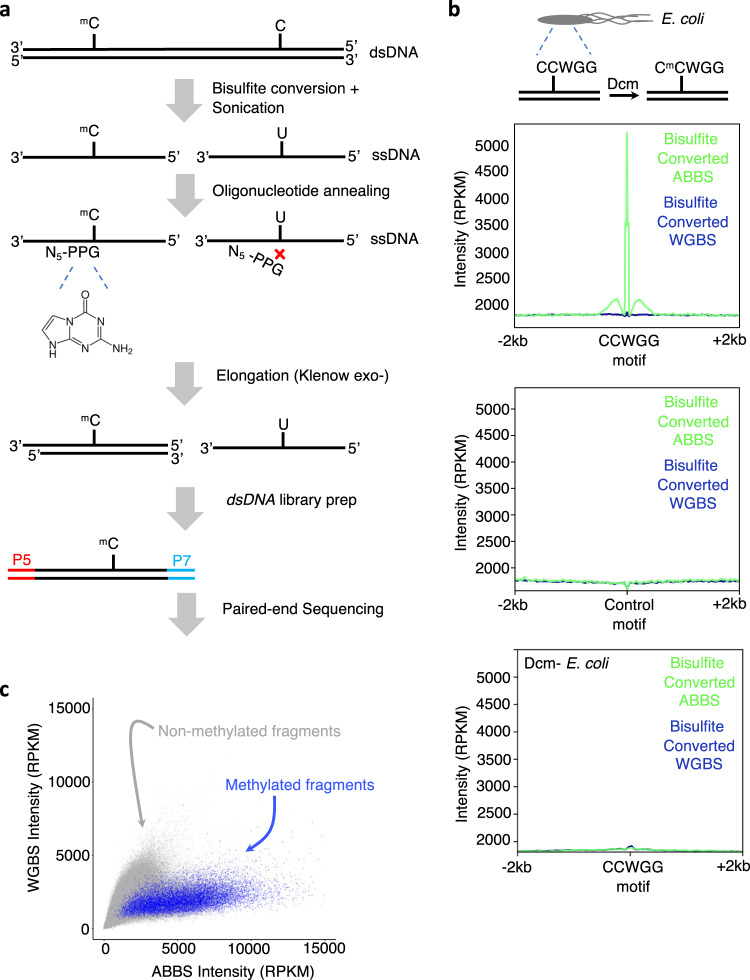
Three primary methods are currently utilized for sequencing DNA methylation at the base-resolution level. These methods include:

Figure 12. Bisulfite conversion



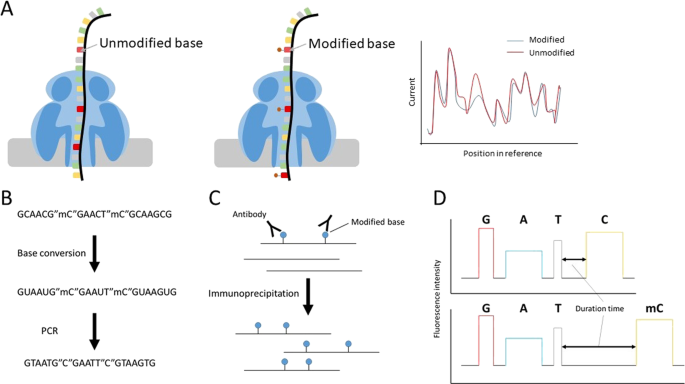
1. **Bisulfite Sequencing** (Figure 15b): traditionally Whole Genome Bisulfite Sequencing (WGBS) has been used, in which DNA is treated with bisulfite, causing unmethylated cytosines to convert into uracils while leaving methylated cytosines unchanged (Figure 12). Subsequently, fully randomized primers are added for library preparation and finally the DNA is sequenced, typically using short-read Illumina sequencing, and aligned to a bisulfite-conversed genome to identify the presence of 5mC (Payelleville and Brillard 2021). A novel bisulfite-based technique called Anchor-Based Bisulfite Sequencing (ABBS) was introduced by Chapin et al. (2022). In ABBS, the focus is on targeting methylated regions using a primer that incorporates five random nucleotides followed by a 3' anchor known as 8-aza-7-deaza-2-deoxyguanosine (PPG). PPG is a pyrimidine analogue that enhances the stability of PPG:C/5mC base-pairing compared to canonical guanosines. Uniquely, bisulfite-refractory methylated cytosines provide a base complementary to the 3' PPG anchor, allowing subsequent elongation using polymerase. ABBS selectively enriches for DNA methylation information and requires up to 10 times fewer sequencing reads compared to WGBS.

Figure 13. ABBS. (Chapin et al. 2022)



1. **Single Molecule Real Time (SMRT) sequencing:** it is an advanced third-generation long-read sequencing-by-synthesis technology, utilizes the real-time imaging of fluorescently labelled nucleotides as they undergo synthesis along single DNA strands (Oliveira and Fang 2021). This technique introduces delays in polymerization during DNA sequencing, referred to as InterPulse Duration (IPD) (Figure 15d). The change in speed is measured relative to an amplified control without modifications, the IPD ratio between control and native template elucidates DNA modification events. Contrarily, the "*in silico* control" is a model trained to recognize the kinetics of the sequencing chemistry. It calculates IPD ratios without requiring an unmodified control sample (Biosciences n.d.).

While capable of detecting m4C and m6A modifications, SMRT sequencing has limitations when it comes to accurately detecting m5C (Payelleville and Brillard 2021). Detection and identification of DNA modifications are distinct processes. Detection involves recognizing pauses in the polymerase during sequencing, indicating the presence of a modification. Given sufficient coverage, modifications causing a pause can be detected within a region, possibly a few base pairs (+/-1-2) away from the modified base. Identification, however, requires additional analysis to precisely determine the modified base. Modified bases can impact polymerase dynamics at various positions, resulting in distinct kinetic "signatures." These signatures provide valuable clues for identifying the type of base modification present. For example: 5-mC has key specific kinetic signals two and six bases downstream from the methylated position, 4-mC exactly at the position of the modification and 6-mA at the modified locus and usually five bases downstream (Figure 14). One significant advantage of using SMRT sequencing for base modification detection is its ability to analyze unamplified double-stranded DNA. This feature allows for the detection of strand-specific modifications, including hemimethylation (Biosciences n.d.).



*Figure 15. Detection of base modifications (Xu and Seki 2020). Modified base detection using Nanopore sequencing and general methods. Schema of modified base detection using the Nanopore sequencer (****a****) and through bisulfite conversion (****b****), immunoprecipitation of nucleic acids (****c****), and SMRT sequencing (****d****)*

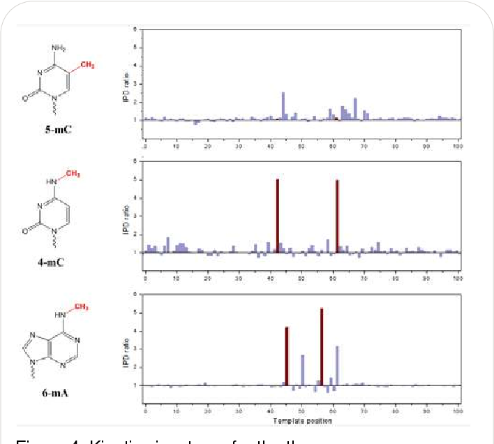


Figure 14. kinetic signatures (Biosciences n.d.)

1. **Oxford Nanopore technology (ONT)**: it has shown remarkable efficacy in efficiently detecting modified bases, particularly m5C and m6A. This method utilizes nanopores for direct DNA sequencing and leverages the analysis of electrolytic current signals, which are highly responsive to base modifications (Figure 15a). ONT enables simultaneous characterization of the DNA sequence and the methylation profile of the DNA template (Payelleville and Brillard 2021).

There are many other methods of DNA-modification detection like MeDIP-seq that involves the use of modification-specific antibodies for immunoprecipitation (Figure 15 c). However, the base-pair resolution is lower (~150 bp) contrary to single base resolution on the three methods mentioned above (MeDIP-Seq 2023).

For the development of the project, I used three different datasets coming for two different sequence technologies which specifications are summarized in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Technology | Number of samples | Modifications | Reference |
| WGBS and ABBS | 6: 2 clones x (1 WGBS + 1 ABBS + 1 ABBS (A.M. library prep)) | m5C | SRA: [SRP329794](https://www.ncbi.nlm.nih.gov/sra?term=SRP329794) ([SRR15242564](https://www.ncbi.nlm.nih.gov/sra/SRX11548393%5Baccn%5D), [SRR15242565](https://www.ncbi.nlm.nih.gov/sra/SRX11548394%5Baccn%5D), [SRR15242568](https://www.ncbi.nlm.nih.gov/sra/SRX11548397%5Baccn%5D), [SRR15242569](https://www.ncbi.nlm.nih.gov/sra/SRX11548398%5Baccn%5D), [SRR15242570](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR15242570&display=metadata), [SRR15242571](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR15242571&display=metadata)) (Chapin et al. 2022) |
| SMRT (Sequel system P1-C1.2) | 8-plex | m6A, m4C, (m5C) | PacificBiosciences / DevNet [8 plex Ecoli Multiplexed Microbial Assembly](https://github.com/PacificBiosciences/DevNet/wiki/8-plex-Ecoli-Multiplexed-Microbial-Assembly) |
| SMRT (RS II P4C2 sequencing) | 1 | m6A, m4C | [MethSMRT](http://sysbio.gzzoc.com/download/Prokaryota/Ecoli.gff.tar.gz) (Sun Yat-sen University 2016) |
| Table 1. Datasets used in the project | | | |

1. WGBS and ABBS as described in the work of (Chapin et al. 2022). Dataset form Sequence Read Archive (SRA), 6 different runs from study [SRP329794](https://www.ncbi.nlm.nih.gov/sra?term=SRP329794) corresponding to *E.coli* K12 substr. MG1655 grown in LB:

* [SRR15242564](https://www.ncbi.nlm.nih.gov/sra/SRX11548393%5Baccn%5D): clone 1, WGBS primers (NNNNNN)
* [SRR15242565](https://www.ncbi.nlm.nih.gov/sra/SRX11548394%5Baccn%5D): clone 1, ABBS primers (NNNNNPPG)
* [SRR15242568](https://www.ncbi.nlm.nih.gov/sra/SRX11548397%5Baccn%5D): clone 2, WGBS primers
* [SRR15242569](https://www.ncbi.nlm.nih.gov/sra/SRX11548398%5Baccn%5D): clone 2, ABBS primers
* [SRR15242570](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR15242570&display=metadata):clone 1, ABBS primer (A.M. library prep)
* [SRR15242571](https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR15242571&display=metadata): clone 2, ABBS primer (A.M. library prep)

1. SMRT

* Raw dataset from: PacificBiosciences / DevNet [8 plex Ecoli Multiplexed Microbial Assembly](https://github.com/PacificBiosciences/DevNet/wiki/8-plex-Ecoli-Multiplexed-Microbial-Assembly). (ccano/DevNet n.d.). *E. coli*.
* Processed dataset: MethSMRT is an integrative database MethSMRT for DNA 6mA and 4mC methylomes (Sun Yat-sen University 2016). I used the *E.coli* K12 substr. MG1655 gff file generated from SMRT sequencing the raw data [SRS674093](https://www.ncbi.nlm.nih.gov/sra/?term=SRS674093) (Berlin et al. 2015).

For all of the subsequent steps the used reference genome is [NC\_000913.3](https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3) from NCBI database and the corresponding genome annotation GCF\_000005845.2 (<https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2/>).

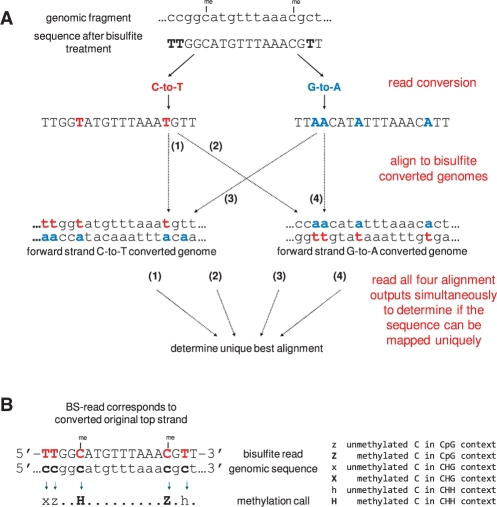
* **Data processing and methylation extraction**

### *Bisulfite sequencing:*

On the original data the paired reads appeared on the same fastq file so with a Python script (separate\_fastq.py-annex) I generate two independent fastq files with each one having one read of the pair.

Raw reads coming from SRP329794 sample processed with Bismark utility, v0.24.0 (<https://github.com/FelixKrueger/Bismark>). Bismark employs an innovative approach to analyze BS-Seq data. It first converts the reference genome *in silico* into two versions: forward C-to-T and forward G-to-A (equivalent to reverse C-to-T) ($bismark\_genome\_preparation). The same action is performed with the residual cytosines (methylated) in the sequenced bisulfited-transformed reads, they are converted *in silico* into a fully bisulfite-converted form before the alignment takes place. To address the challenge of unknown strand identity, Bismark runs four alignment processes simultaneously, aiming to find a unique best alignment (figure 16a) and determines the strand position onto the original genome ($bismark). Some of the features that Bismark mapping has are: single or paired-end (PE) support, it uses base-call qualities for FastQ mapping, directional/non-directional library support, adjustable insert size (PE), variable read length capability of process among others (Krueger and Andrews 2011).

Figure 16. Bismark mapping (Krueger and Andrews 2011)



After the mapping I performed deduplication to eliminate PCR duplicates that map on the same part of the genome ($deduplicate\_bismark). Next, filtering was done based on the quality of mapping (I kept only reads above a score of 20) of each read with another utility: samtools 1.6 (<https://github.com/samtools/samtools>) ($samtools view -h -q 20). After, I computed the coverage depth of each of the sample runs to discard those with low value (<25) ($samtools depth). Another integrated tool in the Bismark utility that I used at this point is the determination of the methylation state of each of the Cs from the bisulfited read ($bismark\_methylation\_extractor) (figure 16b). An option that the methylation extractor includes is the possibility to exclude the edges of the reads/paired reads which correspond to adapters and low-quality base-called positions (--ignore) observed with $fastqc on the raw read data. Bismark enables methylation analysis in different sequence surrounding and discriminates between cytosines in CpG, CHG, and CHH contexts saving the information into a CX\_report.txt. This report has the following data columns which allow for accurate interpretation of DNA methylation data:

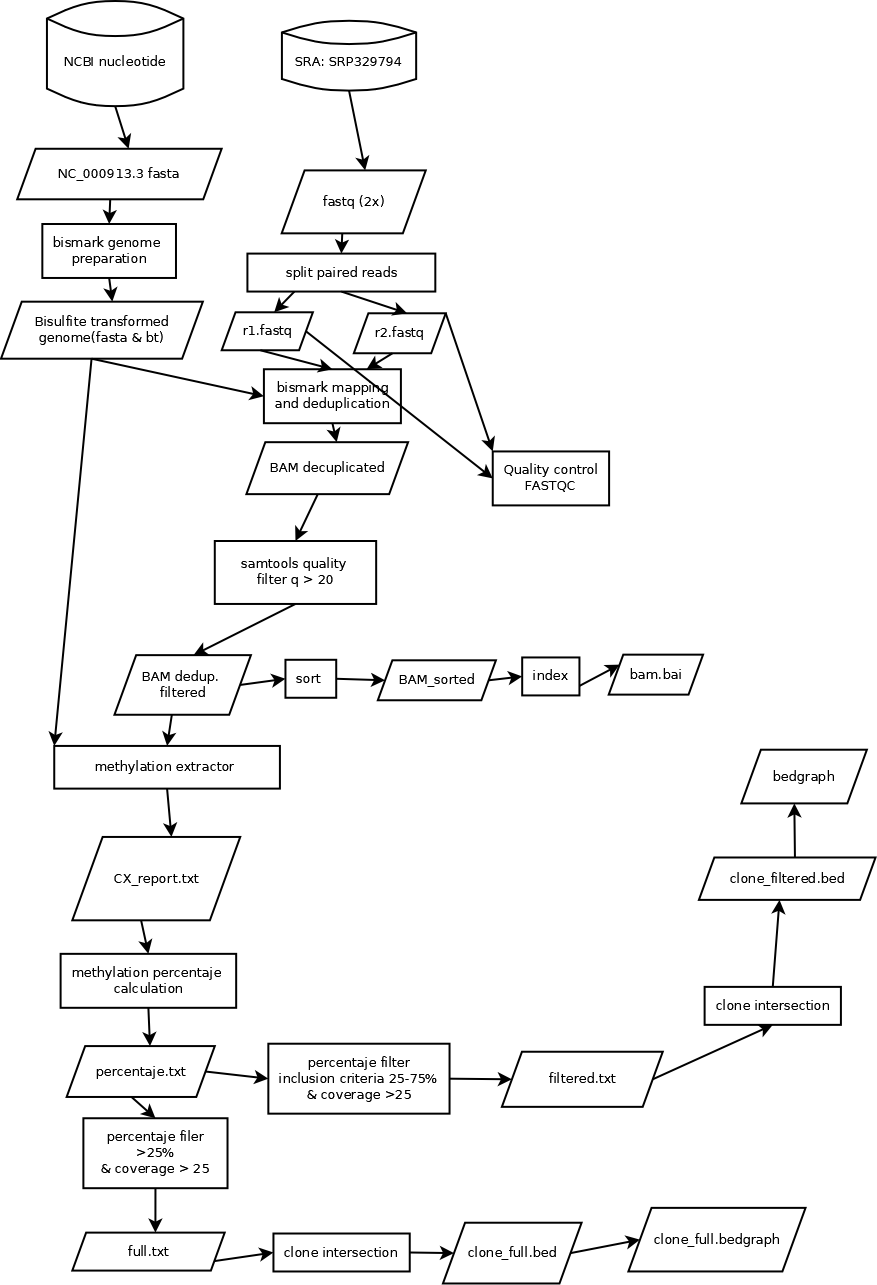
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Chromosome | Position | Strand | Methylated coverage | Unmethylated coverage | C-context | Trinucleotides |

Using the generated CX\_report I calculated the percentage of methylation at each C-position by using the Eq. 1 and selected those ranging between 25-75% of methylation. Moreover, I computed the total coverage of each C-position summing the methylated and unmethylated coverages and excluded all the bases having less than 25 of coverage in order to ensure reliability of the methylation percentages.

Eq. 1:

Finally, Bismark enables to summarize all the information about mapping and methylation calling by generating an html report ($bismark2report).

Figure 17. Bisulfite sequencing data processing workflow



Alignments were visualized with the Integrative Genomics Viewer(IGV) (v2.14.1) ([https://software.broadinstitute.org/software/igv/)](https://software.broadinstitute.org/software/igv/)%20) with previous sorting and indexing ($samtools sort, $samtools index). Reads are colored based on DNA strand: grey is used for forward read according to genome annotation; sage is used for reverse read. IGV enables to color the positions in bisulfite mode in DNA libraries that have undergone bisulfite conversion and sequencing. On this mode some nucleotides are emphasized within the reads with red or blue nucleotides, corresponding to the position of cytosines in the reference genome. For forward reads, a red C indicates a non-converted cytosine (methylated), while a blue T represents a bisulfite-converted cytosine. For reverse reads, a red G indicates a non-converted cytosine (methylated), while a blue A represents a bisulfite-converted cytosine.

Finally, I intersected the samples that came from the same clone to keep the consistent methylation positions (appear in both, are above 25 coverage and methylation level 25-75%). The final level of methylation reported was the average of both. This was performed with bedtools (v2.30.0) (<https://bedtools.readthedocs.io/en/latest/>) ($bedtools intersect). In Figure 17, there is the summary of how bisulfite-sequencing dataset was processed.

See the commands employed for this part at appendix 1.

### *SMRT:*

The first dataset (PacificBiosciences / DevNet [8 plex Ecoli Multiplexed Microbial Assembly](https://github.com/PacificBiosciences/DevNet/wiki/8-plex-Ecoli-Multiplexed-Microbial-Assembly)) comes from continuous long reads (CLR) sequencing in Sequel platform (Table) which base call is under a [subreads.bam](https://s3.amazonaws.com/files.pacb.com/datasets/secondary-analysis/e-coli-k12-8-plex/Ecoli_8plex_demo.barcoded.subreads.bam) file, [subreads.bam.pbi](https://s3.amazonaws.com/files.pacb.com/datasets/secondary-analysis/e-coli-k12-8-plex/Ecoli_8plex_demo.barcoded.subreads.bam.pbi) file that contains the corresponding bam index and [barcode.fasta](https://s3.amazonaws.com/files.pacb.com/datasets/secondary-analysis/e-coli-k12-8-plex/barcodes_8plex.fasta) has the barcodes sequences used. 8 barcodes where used, forward and reverse barcodes are identical in each of the samples (--same). There are a total of 8 samples. Some of the data features given are: average read length 10.6 kbp, N50 read length 18.6 kbp, mean coverage for 8 samples 110X.

For the processing of data generated with SMRT sequencing I used smrttools from the software [SMRT Link v11.0](https://downloads.pacbcloud.com/public/software/installers/smrtlink_11.0.0.146107.zip). The first step is to demultiplex the file and separate the reads into 8 different bam files with the executable command lima ($lima --same --split-bam) for further parallel processing of the data. Next step was mapping to the reference fasta genome [NC\_000913.3](https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3) and sorting the mapped reads ($pbmm2 align –sort). In order to call the methylation positions, type and fraction I used $ipdSummary tool that generated a .gff file containing the information. It is important to set correctly the options in this step such as --identify m4C,m6A,m5C\_TET1 --pvalue 0.0012--useChemistry "SP3-C3"3.

*1 Because the detection of 5mC is only effective when the specific position in the reference genome is heavily modified SMRT introduced the 5mC\_TET model. It implies the conversion of 5mC to 5-carboxylcytosine (5caC) using the Tet enzyme. This enzymatic conversion enhances the kinetic signature, making it more detectable when using SMRT Sequencing. However, this transformation was not mentioned in the preparation of the sample (Biosciences n.d.).*

*2More stringent p-value than default (0.01) since the default was detecting many modifications but not identifying the kind.*

*3 The chemistry under which the dataset was sequenced is P1-C1.2, a rather old one, which was not available in idpSummary, thus, I used the oldest one available (SP3-C3).*

The generated output .gff is a Generic Feature Format Version 3 (GFF3) that has this display of fields tab-separated. To know more about GFF3 format go to (Stein 2020).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chromosome | Source | Type mod | Start | End | Score | Strand | phase | attributes |

Inside the field attributes there are different semicolon-separated traits: e.g. coverage=11;context=GCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAA;IPDRatio=10.59;frac=0.852;fracLow=0.402;fracUp=1.000;identificationQv=9

*“****coverage*** *level is the number of IPD observations used after Mapping QV” MapQV>10*

*“If the row results from an identified modification we also include an* ***identificationQv*** *tag with the from the modification identification procedure. identificationQv is the phred-transformed probability of an incorrect identification, for bases that were identified as having a particular modification”*

*“****frac, fracLow, fracUp*** *are the estimated fraction of molecules carrying the modification, and the 5% confidence intervals of the estimate”* (BaseMod-3.0/README.md at master · ben-lerch/BaseMod-3.0 · GitHub n.d.)

Figure 18. SMRT data processing workflow



(“The methylated fraction estimation is a beta-level feature, and should only be used for exploratory purposes.”) say on limitations

Finally, inclusion criteria for modifications were:

* Identified methylations (m6A, m5C, m4C)
* IPDRatio>= 4
* fraction methylation between 25-75%

Bed and bedgraph files were created from the filtered gff for gene association and visualization of the percentage of methylation at different positions. It is necessary to subtract 1 position to the start field at the gff file since in this one the start and end indicate the same position, however in bed format they need to differ by one.

The summary of the processing steps is depicted in figure 18 from the raw data to bed methylation positions files. The commands used in this part are in appendix 2.

The second SMRT dataset is an already generated gff file from MethSMRT. The following steps after obtaining the gff are the same.

* **Gene association**

Once I have extracted the methylated positions and corresponding percentages for each of the datasets/samples, the next step was to associate them to genes that are allegedly are under methylation regulation. The organism studied is a prokaryote therefore its genes are organized under operons i.e., a functioning unit of DNA under the control of a promoter that forms a transcription unit. One or more genes can be contained in the operon, meaning that often it is a cluster of genes that co-transcribed. However, operons can have more than one promoter and different transcription units can arise combining different pertaining genes (Figure 31b). In general, expression of bacterial operons generates polycistronic mRNAs. In all, the operon is the conformed by the genes in the biggest of these possible transcription units (Figure 31a).

With this rationale and the ideas mentioned in the epigraph *Introduction, Methylation functions in bacteria,* the approach that I followed was **operon-based**, i.e., all genes in the operon are considered to be putatively regulated by a (lack of) methylation that falls within the promoter/regulatory region or within the coding body of any them, regardless of the strandedness.

The annotation of operons in Escherichia coli K-12 MG1655 comes from RegulonDB (Tierrafría et al. 2022). Two files were used containing the following information:

* OperonSet.txt: downloads > downloadable Experimental Datasets > Operons

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Operon name** | First gene-position left | Last gene-position right | DNA strand where the operon is coded | Number of genes contained in the operon | **Name or Blattner number of the gene(s) contained in the operon** | Evidence that support the existence of the operon's TUs | Evidence confidence level (Confirmed, Strong, Weak) |

* operon.txt: downloads > regulonDB full version (sign up) > release 11.1 (txt)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| operon ID | **operon name** | **first genepos left** | **last geneposright** | **regulation pos left** | **regulation pos right** | **operon strand** | operon internal comment | Key ID org |

In bold are the fields of interest that I used to process the data.

The file operons.txt contains experimentally proved regulatory positions and coding regions. With R, I split it into two files: one holding information from regulatory positions upstream of the first promoter and downstream the last coding position (operon\_reg) and another one containing the bulk of coding sequences of the operon (operon\_cds). Subsequently, I performed intersections of each of the 7 bed files containing information about the methylation annotation against these two files ($bedtools intersect), obtaining 14 new files (overlap\_reg/overlap\_cds) that were the start of the process in R to extract the genes associated for each of the datasets/samples (Figure 20). See commands in gene\_call\_commands.txt

The other file Operons.Set.txt contains the names of the genes that each operon holds in official symbol annotation (e.g. accA). However, for the subsequent metabolic modelling locus tag (e.g. b0185) is needed; to achieve this conversion I used the R package “GenomicFeatures” with the gene annotation GCF\_000005845.2 corresponding to the NC\_000913.3 assembly, (<https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2/>) (Figure 21).

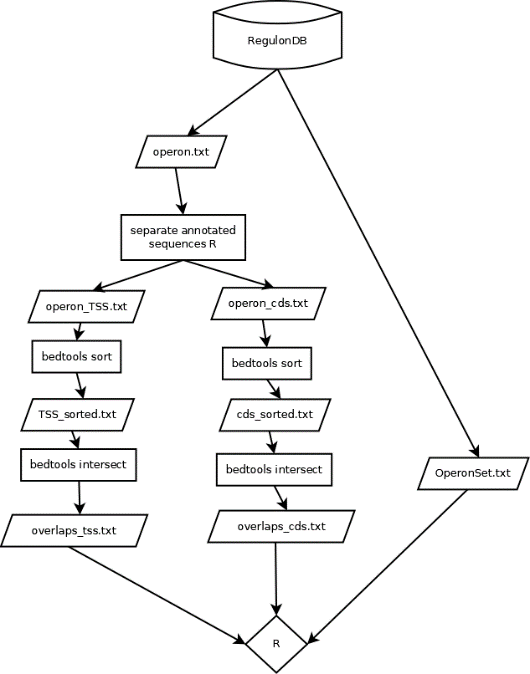


Figure 20. gene call workflow

Data for each dataset/sample was organized in significance tables. It was prioritized (higher in table) methylations on regulatory regions over methylations on gene bodies. Moreover, the former had a threshold of 1, and 4 in the case of the latter for an operon to enter the table. In the table there are fields containing the number of methylations found at each operon on both coding or regulatory sequence. Moreover, they have the names of the genes contained on each operon in locus tag and gene symbol nomenclature (Table 6).

Additionally, I generated a list of all the metabolic genes annotated in the model iML1515 (see epigraph). The transformation of nomenclature however, this time was done with DAVID gene ID conversion (<https://david.ncifcrf.gov/conversion.jsp>) from locus tag into gene symbol (generated conv\_38DC2E424FE51674044440701.txt file). I did an additional table similar to Table 5 but leaving in only the genes that were metabolic (Figure 21). Sidenote: some operons have a metabolic part and a non-metabolic part; in this case I kept the metabolic part instead of discarding the whole operon. Lists of genes from the metabolic tables will be the input for metabolic modelling.

* **Functional enrichment**

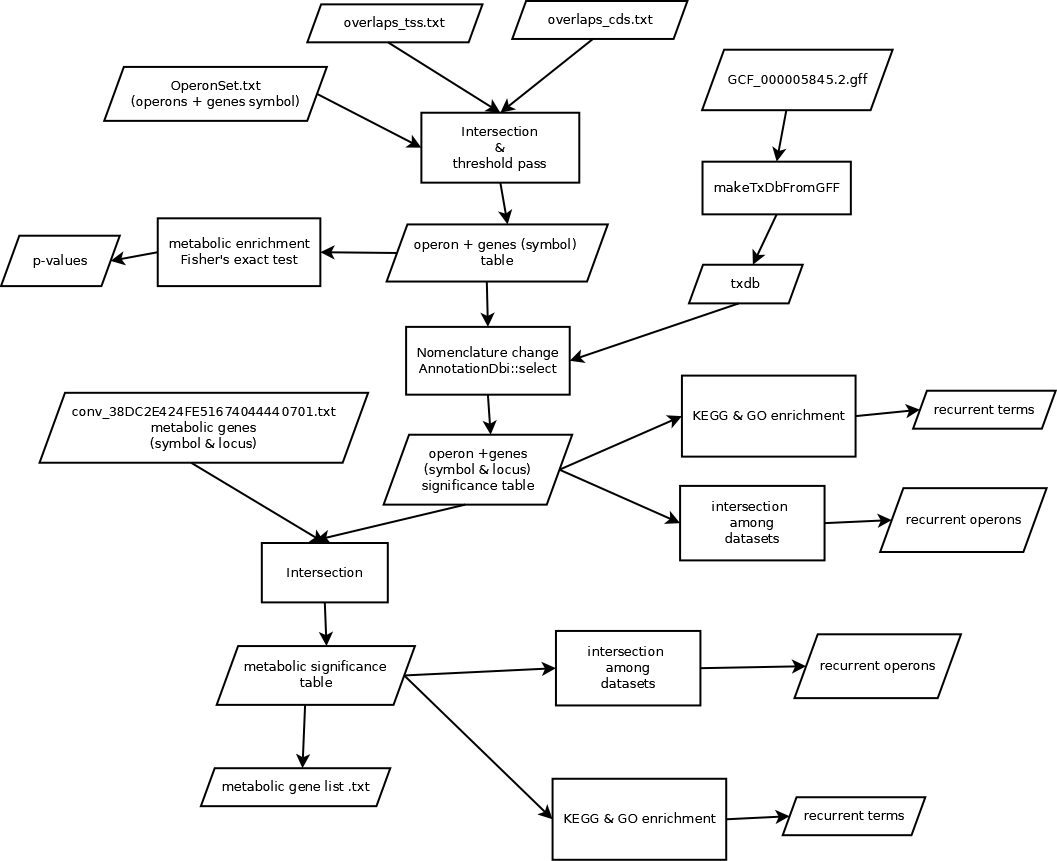


Figure 21. Workflow carried out in R for gene call and functional enrichment

Next, with all the genes extracted a functional enrichment analysis was done on both gene ontology (GO) terms and KEGG pathways thanks to the package “clusterProfiler” which needs as input the functional annotation of package “org.EcK12.eg.db”. Both can be found within the Bioconductor. The same was done only with the selected metabolic genes and setting as background all the metabolic genes annotated in the model (1515) (Figure 21). Since often operons are already grouped into functional categories, i.e., the genes in the same operon perform a function of the same pathway, there is a bias for the functional enrichment calculation. For this reason, the q-value cutoff chosen was as stringent as 0.005.

Finally, I was interested to know whether the whole set of genes selected had an actual enriched representation of metabolic genes. For this, a Fisher’s exact test was done (Eq.2, Figure 21) taking the following parameters:

k=number of metabolic genes at each of the datasets

n=number of total selected genes at each of the datasets

G=4256 genes that *E. coli* has annotated

m=1515 genes that are annotated under metabolic for *E. coli*

The remaining contingency table would be:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not in my sample | Sample | Total |
| Metabolic | m-k | k | m |
| Non-metabolic | G-m-n+k | n-k | G-m |
| Total | G-n | n | G |

Therefore, the corresponding Fisher’s exact test formula:

Eq.2:

* **Motif identification**

One of the starting hypotheses was that methylation might be partial at some sites because of sequence degeneracy (one/more nucleotide different to the consensus) so that the MTase had less affinity to it for methylation catalysis. To test this, I gave each of the methylated position a score based on their resemblance to the methylation motifs known and compared the ones from partially methylated positions against all the methylated positions (fully and partially).

As mentioned before, in the epigraph *Introduction, MTases in E.coli*, there are 4 known main motifs that can undergo methylations CCWGG, GATC, ATGCAT and AACN6GTGC. Therefore, once extracted the methylation positions, besides doing the gene call, it is of great importance to seek for the motifs that had been methylated, either fully or partially. The tool employed for this issue was [RSAT prokaryotes dna-pattern](https://rsat.eead.csic.es/plants/dna-pattern_form.cgi) (van Helden, André, and Collado-Vides 2000), it is an online utility that accepts as input query pattern(s) and fasta (or other) files with the sequences to query. It scans your sequences to retrieve the matches, corresponding start and end positions and score. Fasta sequences were pulled from the bed and percentage.gff files for the partially and fully methylated genome respectively with the following steps:

* Extension of positions in the files up to 22 nucleotides long (start-10, end+11 in the bed file and start-11, end+11 in the gff). This length and positions were chosen in order to englobe the size of all the four known methylation motifs at both strands.
* Extraction of the sequences into fasta format with ([mEpigram](https://github.com/Wang-lab-UCSD/mEpigram) bedToFasta.py). This executable extracts the sequence of the forward strand therefore we need to be careful when searching the motifs from the reverse complementary one.

I created two bash functions that did these two steps giving the initial file and 22 as an argument, one for filtered data ($create\_fasta\_filtered <input.bed> 22) and the other for full ($create\_fasta\_full <input.gff> 22), they both retrieve the final fasta sequences file. See commands in appendix 3).

Once the fasta sequences are extracted for each of the datasets/samples the parameters selected in RSAT were “search strands: only direct”, “don’t prevent overlapping matches” with one substitution allowed. The rest of them were left as default. RSAT returns a table from this analysis like the fragment seen in Table 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| PatID | Strand | Pattern | SeqID | Start | End | matching\_seq | Score |
| START\_END | DR | - | seq\_1 | -22 | -1 | - | 0.00 |
| GATC | D | GATC | seq\_1 | -13 | -10 | GATC | 1.00 |
| CCWGG | D | CCWGG | seq\_1 | -11 | -7 | TCAGG | 0.80 |
| START\_END | DR | - | seq\_2 | -22 | -1 | - | 0.00 |
| GATC | D | GATC | seq\_2 | -13 | -10 | GATC | 1.00 |
| CCWGG | D | CCWGG | seq\_2 | -18 | -14 | CGTGG | 0.80 |
| START\_END | DR | - | seq\_3 | -22 | -1 | - | 0.00 |

Table 2. Example of head of the returned table from RSAT containing information of motifs matched with the corresponding relative position and score

Note that each of the sequencies to query can retrieve more than one pattern matching. RSAT is not taking into account the position in which each motif must appear on the sequence due to the initial methylation position and posterior extension. For this, thanks to R, I excluded the matches that don’t belong where they must. Keep in mind that with the bedtofasta.py command we obtain uniquely the sequence on the forward strand even if the methylation was on the reverse. Three of the searched motifs are palindromic, meaning the reverse complementary sequence is identical (CCWGG, GATC and ATGCAT) thus, the same sequence must be queried but at different positions for forward or reverse strand. Nevertheless, for the last pair (AACN6GTGC/ GCACN6GTT) searching for the complementary is essential to gather all the information when querying the reverse strand, this means, if I want to query whether the methylation is happening at 3’CGTGN6C**A**A5’ (5’A**A**CN6GTGC3’) of the strand considered to be the reverse with respect to the genome annotation I must query for the GCACN6GTT motif on RSAT with the reverse positions marked in the table 3 (note asterisk \*), and vice versa, if I want to query whether the methylation is happening at 3’TTGN6C**A**CG5’ (5’GCACN6GTT 3’) of the reverse strand I must query for the AACN6GTGC with the reverse positions. On the other hand, if two motifs were found in the correct position on the same sequence the one having the higher score is the one accounted for.

The program returns the matching positions calculated relative to the sequence end by default, therefore the wanted positions for each motif are:

|  |  |  |  |
| --- | --- | --- | --- |
| 5’3’ methylated motifs | Methylated/queried strand | START | END |
| CCWGG | Forward | -13 | -10 |
| Reverse | -15 | -11 |
| GATC | Forward | -13 | -10 |
| Reverse | -14 | -11 |
| ATGCAT | Forward | -16 | -11 |
| Reverse | -13 | -8 |
| AACN6GTGC | Forward | -13 | -1 |
| Reverse\* (GCACN6GTT) | -23 | -11 |
| GCACN6GTT | Forward | -14 | -2 |
| Reverse\* (AACN6GTGC) | -21 | -9 |
| Table 3. Start and end positions of the motifs queried in the extended 22 nucleotides sequences. To query the reverse strand for AACN6GTGC or GCACN6GTT I need to search the reverse complementary marked with \*. | | | |

The result files retrieved from RSAT were processed to calculate a set of parameters:

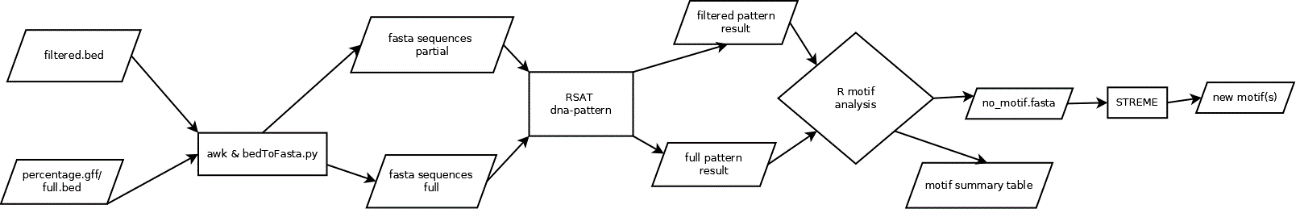
* Percentage of known motifs found (for partial and full)
* Fraction of partial over full methylation
* Mean score of these motifs (for partial and full)
* Fraction of each motif’s occurrence in partial methylation with respect to full methylation

To be practical the results of the pair AACN6GTGC/GCACN6GTT are presented together after the name of the former sequence.

To end with, the sequences for which no motif was found were pulled out from the full.fa file and given as input for STREME. The STREME algorithm, represents a significant advancement in ab initio motif discovery due to its improved accuracy and versatility. STREME offers the capability of identifying ungapped motifs, which are recurring patterns of fixed length. These motifs can be enriched in your dataset containing a large number of sequences or can be relatively enriched compared to control sequences. Furthermore, STREME is capable of detecting both short and long motifs, ranging from 3 to 30 positions. One notable feature of STREME is its ability to provide a reliable estimate of the statistical significance associated with each discovered motif (Bailey 2021).

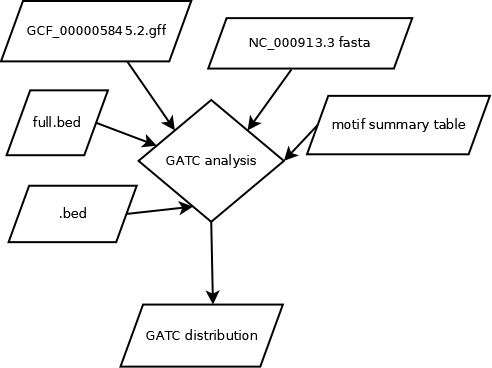
The workflow summary can be seen in Figure 22.

Figure 22. Workflow motif analysis



* **GATC motif genomic analysis**

Figure 23. GATC motif genomic analysis workflow



As stated in the results section, GATC is the most recurrent methylated motif, therefore it is convenient to go deep into its analysis. One of the aims is to see whether the distribution of GATC motif is uniform between coding regions and the rest of regions, denoted as intergenic. As well as checking whether the partial methylation follows the same distribution or not. For that, I used the genome annotation GCF\_000005845.2\_ASM584v2\_genomic.gff and NC\_000913.3 assembly with matrixStats and seqinr packeges in R (Figure 23). For each cds annotated, the corresponding sequence is extracted and all the 4-mers are counted and annotated. This is done separately for forward and reverse strand. The intergenic regions are the areas between one cds and another one, so once the positions are extracted, we can perform the same analysis than in the case of cds. The number of GATC counts is summed for all sequences in forward and reverse strand in both cds and intergenic cases to get the total occurrence on each case. Finally, for these two amounts the percentage of each representation compared to the whole genome is calculated.

Thanks to the observed values we can perform a binomial test to check if the partial methylation positions of the SMRT datasets follow the same distribution of GATC sites. Since the percentage of other motifs is insignificant as depicted in Table 10, all of the hit positions are taken as GATC to simplify the analysis. Starting from the .bed files the classification of intergenic or cds for each position is done, and to finish the counts on each are compared to the genome standard with the binomial test -binom.test(results[intergenic],total\_positions,p=0.53)-.

Another task was to reveal the likelihood of finding a GATC motif on sequences randomly picked from two types of data: the whole genome and cds regions. In order to do this, the sequences that were already annotated as cds were reduced to stretches of 100 random consecutive nucleotides within them and so with the whole genome sequence. Next, sets of 2000 of these sequences were subsampled and measured on 4-mer relative content that was binarized (if the relative content in a sequence is greater than 0 the value is set to 1). This operation was carried out 100.00 times for each of the types of data. The average times that at least one GATC motif was found on each set is calculated (sum of all occurrences divided by 2000) to form a null distribution. Finally, the two null distributions coming from the different type of data are compared with a t-test. The null hypothesis states that there is the same probability of finding a sequence with at least one GATC motif when choosing at random from any cds than picking from the whole genome.

On the other hand, the positions on the full.bed files were classified as cds or intergenic depending on where they fell and strandedness.

* **Metabolic modelling**

The metabolic model iM1515 used in this study is based on the genome of *Escherichia coli* str. K-12 substr. MG1655 (<http://bigg.ucsd.edu/models/iML1515>. This model incorporates 1,515 open reading frames and includes 2,719 metabolic reactions involving 1,192 distinct metabolites. iM1515 has undergone rigorous validation and customization to suit various growth conditions, making it the most extensive and up-to-date reconstruction of *E. coli* metabolism available. The experimental validation of this model consisted on screening the KEIO collection, a genome-wide gene-knockout collection (3,892 gene knockouts), by growing each mutant under 16 conditions: minimal media containing 16 different carbon sources (Monk et al. 2017b).

To assess the network properties of iML1515 I used COBRApy v.0.26.2. COBRApy is a comprehensive python package tailored for constraint-based modelling, accommodating the intricate biological complexities of advanced COBRA models. It provides convenient access to methods such as flux balance analysis, flux variability analysis, and gene deletion analyses. COBRApy's fundamental classes include Model, Metabolite, Reaction, and Gene. The Model class acts as a container for a collection of chemical Reactions, along with their associated Metabolites and Gene products. Within a Model, Metabolites are altered by one or more Reactions, which can be either spontaneous or catalyzed by one or more Genes. COBRApy follows an object-oriented design, allowing users to directly access attributes of each object and manipulate them as needed. This design approach enhances flexibility and ease of use for modelling (Ebrahim et al. 2013).

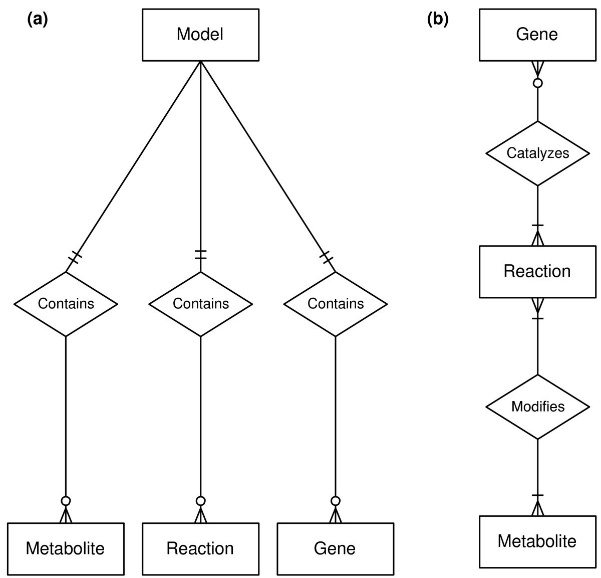


Figure 24. Core COBRApy classes (Ebrahim et al. 2013)

Now that I have the model in .xml format (iML1515.xml file), the tool and the list of genes that are putatively regulated by partial methylation, I took different steps and approaches for modelling:

* Load and exploring the file.

Once the cobra package has been imported, with the function read\_sbml\_model from the cobra.io module we can load the model file into a Model class. importExcelModel.py script has a cobrapy\_to\_excel function that creates an excel easy-to-read sheet with all the reactions information as seen in the example of table .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Abbreviation | Reaction | GPR | Lower bound | Upper bound | Objective | Confidence Score | Subsystem | Description |
| CYTDK2 | cytd\_c + gtp\_c 🡪 cmp\_c + gdp\_c + h\_c | b2066 | 0 | 1000 | 0 | 4 |  | Cytidine kinase (GTP) |

To access the main classes in the Model methods like reactions(), metabolites() and genes() must be used. Other important method is medium() which gives information about the media of the model, by default it is M9, with D-glucose as carbon, source as well as tungsten (tungs), carbon dioxide (co2), ???(slnt), selenium (sel). The media is represented in the model as Exchange reactions with infinite bounds (in practical -1000 and 1000). Setting the bounds of the exchange reactions to 1000 the culture media is changed. As for this study I employed two different starting models, one for M9 and one for LB. The specific media composition is found in the table media\_db.tsv (<https://github.com/cdanielmachado/carveme/blob/master/carveme/data/input/media_db.tsv>) (Appendix 4 supplementary table 1 and 2). The reactions ['EX\_cro4\_e', 'EX\_fol\_e',"EX\_ribflv\_e"] are not annotated in the model so they were eliminated elements in LB media.

With the conditions of interest set appropriately the Model is ready to be optimized with the module optimize() which does FBA and retrieves an optimal solution for the objective value and the fluxes of all the reactions.

* Extracting reactions associated to differentially methylated genes

From my list of genes that are putatively regulated by partial methylation, the associated reactions are pulled. This is dome with the methods genes.get\_by\_id(“gene”).reactions on the Model. Note that one gene can have associated more than one reaction and one reaction can have associated more than one gene (see GPR figure), though I am extracting all of the annotated reactions gene by gene regardless of other association rules. These reactions may be regulated by partial methylation in a positive or negative way (flux increase or decrease) due to rise or drop in enzyme expression.

* Getting non-essential REACTIONS from my datasets

Organisms have some reactions which loss would imply dead or at least impossibility to grow, these are called essential reactions. Depending on the media conditions in which bacterial cells are grown (simulated to grow) the indispensable reaction are different. Thus, the next step is to identify the essential reactions to be aware of which are “forbidden” to knock-out. COBRApy has an in-built function in cobra.flux\_analysis variability module called find\_essential\_reactions(Model) to pull the essential reactions for a model. Now, there is left to eliminate these reactions from the elements saved in *reactions* variables extracted in the previous step. These *nonessential\_reactions* are indeed potentially regulated by methylation.

* Flux balance analysis and active reactions

However, it does not make much sense to knock-out a reaction that is not even active, for that reason it is important to evaluate those active reactions. FBA gives us the fluxes of each, as previously mentioned in the exploratory analysis. I considered every reaction with a flux above 0. 0001 to be active. We obtain the variable *active\_reactionsFBA.*

* Flux variability analysis deviating o.5 and 0.8 from the optimum and active reactions

Besides a simple FBA, also FVA can be performed. Two scenarios are set: one that deviates the fluxes 0.5 from the optimum and another one that deviates 0.2 (fraction\_of\_optimum=0.8). FVA fixes the optimized objective value from FBA and retrieves possible ranges of fluxes within these deviations which in certain combination reach it. For a reaction to be active it must have the minimum or maximum solution value above 0. 0001. We obtain the variable *active\_reactions\_FVA05/8.*

* Active and non-essential reactions regulated by methylation
* Reaction knockouts FBA

Untill now he have only explored the wild-type model in two different media, now the modelling begins.

* Constraints FVA 0.5 and 0.8
* Summary table
* Double reaction knock-out FBA
* Non-essential GENES single KO from my datasets, FBA
* Double GENE KO, associated reactions
* GENE constraints 0.5 and 0.8
* Flux deviations
* LB media model (glucose)
* M9 media model (glucose)

Baseline maps merged with escher\_map\_merger (<https://github.com/reimopriidik/escher_map_tools>): Central\_metabolism, Fatty\_acid\_biosynthesis\_(saturated) in Escher.

DauA is an inner membrane protein that transports C4 dicarboxylates in an aerobic, acidic (pH 5) environment. DauA transports the monoanion and dianion form of succinate and fumarate, and the dicarboxylic amino acid aspartate. At pH 5, DauA is the main succinate transporter and [DctA](https://ecocyc.org/gene?orgid=ECOLI&id=DCTA-MONOMER), which is active for dicarboxylate transport at pH 7, is not produced [[Karinou13](http://www.ncbi.nlm.nih.gov/pubmed/23278959)]. DauA consistently co-purifies with proteins involved in fatty acid metabolism - FabA, FabR, FadE and FadJ [[Babu10](http://www.ncbi.nlm.nih.gov/pubmed/21070944).

*E. coli* Δ*ccmA-H* mutant strain (EC06) is not able to produce mature *c*-type cytochromes under anaerobic, nonfermentative growth conditions [[ThonyMeyer95](http://www.ncbi.nlm.nih.gov/pubmed/7635817)]

When E. coli is growing in glucose minimal medium, nanA expression is barely detectable due to inactivation of CRP and repression by NanR ([4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807447/#B4), [7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807447/#B7)). However, with CRP activated during growth on a non-catabolite-repressing substrate like glycerol, nanA expression increases, presumably by RNA polymerase blocking repressor binding ([14](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3807447/#B14)

* **COMETS**

COMETS is a free open-access software package that was developed as tool for predictive modelling used in research on natural and synthetic microbial complex communities. COMETS offers multiple modes of utilization from the most elementary to the most complex: graphical user interface (GUI), MATLAB toolbox, Python toolbox (COBRApy), and command line. The GUI allows users to load prewritten input files, while the MATLAB and Python toolboxes enable the preparation and execution of models and layouts using standard formats like SBML. They provide an intuitive environment for creating advanced settings and running simulations with the ability to customize analysis and visualization. For advanced users, command line usage is available, suitable for large-scale simulations on computational clusters. This mode allows the development of custom features (Dukovski et al. 2021)

To run a simulation in COMETS, three **input** files/variables/arguments are required: models, parameters, and layout. The **model(s)** contain stoichiometric metabolic reconstructions. COBRApy models need to be transformed into COMETS format by the function model() from the cometspy package. The **layout** describes the environmental conditions, such as the media composition and spatial structure at each location in space. Media concentrations determine the rate of exchange reactions. It may also indicate the presence of extracellular enzymatic reactions. The **parameter(s)** contain a set of customizable specifications. There are different kinds of parameters related to: spatial propagation of either biomass or metabolites, output file writing, GUI and image caption, the extracellular reactions model, lag phases, specific modes of growth (serial dilutions or chemostat mode), evolution (mutations), genome size cost, and finally the most important, simulation parameters such as:

* timeStep: The amount of time between two consecutive simulation updates
* maxCycles: Number of dFBA iterations (steps) for the simulation. The total simulation time will be timeStep \* maxCycles.
* deathRate: The rate of biomass removal per time step.
* cellSize: Grams in one cell. Relevant in simulations with serial dilutions or mutations.
* minConcentration: Minimal concentration of metabolites in the media (Dukovski et al. 2021).

If a parameter is not specified in the input file, it will be assigned the default value defined in COMETS. COMETS simulations generate various quantitative **output** files. The basic outputs include 'total biomass' (a table with iteration numbers and total biomass of each model), 'biomass' (detailed spatial distribution of biomass for each model and grid box), 'media' (information on extracellular metabolite/nutrient amounts), and 'fluxes' (all fluxes, including exchange fluxes, for each model at each time point and grid point) (Dukovski et al. 2021).

Key features of COMETS include:

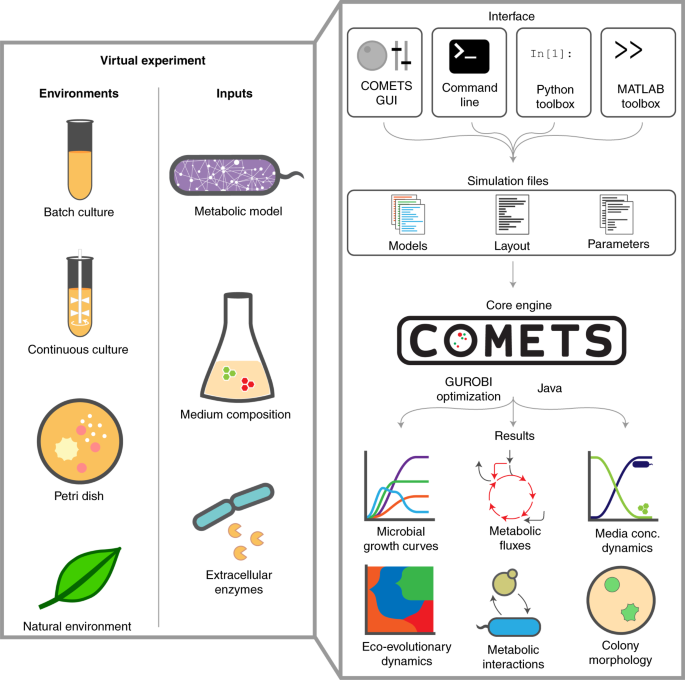


Figure 25. (Dukovski et al. 2021).

1. Metabolite Environment: Simulations begin with an initial metabolite environment, which can vary across space. Metabolites can be set to change in predefined ways, such as static values (concentration constante starting at each time step), constant refresh (add/remove at a rate per hour divided between the timesteps), periodic variations (wave functions), or proportional dilution.
2. Spatial Structure: COMETS can simulate microbial colonies and communities in arbitrary 2D spatial structures using a grid-based approach. A well-mixed situation is accounted for inside each of these ‘boxes’. Biomass and environmental metabolites can propagate between neighboring grid boxes based on convection-diffusion principles and prevented into certain locations by the placement of barriers.
3. Biomass Growth and Propagation: COMETS simulates the growth and propagation of biomass by solving partial differential equations. Biomass is treated as a spatially continuous variable (locally averaged quantity), allowing simulations of macroscopic systems. This approach enables the study of large-scale dynamics.
4. Stochastic Effects: COMETS incorporates two types of stochastic noise. Demographic noise accounts for sampling in finite populations, while growth rate noise considers fluctuations in nutrient availability and cellular properties.
5. Extracellular Reactions: COMETS can simulate reactions occurring in the extracellular environment, independent of specific organisms. Users can implement reactions based on mass-action kinetics or Michaelis-Menten kinetics, enabling the simulation of extracellular enzymes.
6. Random Mutation: COMETS allows the generation of mutated organisms during simulations to study evolutionary dynamics. Mutants with modified stoichiometry are randomly placed in grid boxes containing biomass of the ancestor, and mutation rates can accumulate over time (Dukovski et al. 2021).

Using COMETS we can simulate an heterogeneous clonal population to see the growth capabilities with respect change in environmental conditions in contrast to an homogeneous wild type population. Some candidate genes and associated reactions selected from the modelling in M9/LB are to be down or upregulated to create new models set in coculture with the wt model in order to simulate this phenotypic heterogeneity.

# **EXPERIMENTAL RESULTS & DISCUSSION (25-30 pages)**

* **Data processing and methylation extraction**

On table 4, there is a summary of quality parameters obtained from mapping with bismark the bisulfite-treated samples. Values were pulled out from final html reports plus coverage calculation with samtools. Samples SRR15242564, SRR15242565, SRR15242568 and SRR15242569 had a final reasonably good average (table 4) and interval-based (Figure 26) coverage to go on with the analysis. Samples SRR15242570 and SRR15242571, each of them belonging to one different clone and both prepared with ABBS primer (A.M. library prep), where discarded because the final average coverage was very low.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total reads | Unique alignments (%) | No alignment (%) | Multiple alignment (%) | Duplicated (%) | Deduplicated reads | Average coverage |
| SRR15242564 | 26.900.953 | 66.7 | 31.4 | 1.93 | 7.15 | 16.652.329 | 269.92 |
| SRR15242565 | 23.732.818 | 74.1 | 24.6 | 1.28 | 51.3 | 8.557.417 | 138.41 |
| SRR15242568 | 25.126.669 | 66.1 | 32 | 1.85 | 6.55 | 15.531.579 | 251.713 |
| SRR15242569 | 24.355.597 | 78.6 | 20 | 1.35 | 51.6 | 9.261.726 | 149.867 |
| SRR15242570 | 4.445.597 | 30.1 | 69.4 | 0.55 | 53.7 | 618.645 | 9.3315 |
| SRR15242571 | 5.686.332 | 41.8 | 57.2 | 0.994 | 55 | 1.069.696 | 7.94405 |
| *Table 4. Parameters from bisulfite sequencing data. Each run of sequencing is represented by the “total reads”. After aligning with bismark the reads excluded are those “no alignment” and “multiple alignment”. “Unique alignments” are the reads kept. Among these those different paired reads mapped exactly on the same positions are considered PCR duplicates, “duplicated” fraction. The final number of properly mapped reads are “deduplicated reads”, after quality mapping filter > 20 the final “average coverage” was calculated and used as exclusion parameter for the two last samples.* | | | | | | | |

Table 5 summarizes methylation information extracted. In all the samples the level of methylation is similar, around 4,2% of the cytosines found at the mapped reads were called methylated. An even level of methylation denotes that the cultures had been treated similarly and no sample bias is introduced. Moreover, most of the Cs appear in CHG context (above 84%). This is precisely the context of the C**CWG**G motif methylated by Dcm.

Nevertheless, as for the case of CpG context there is a rather high percentage for a bacterial genome since prokaryotes do not contain endogenous CpG methylases, in fact, *E. coli* contains restriction enzymes that cleave on this DNA methylated context (Nichol and Pearson n.d.). In addition, a CHH methylase has not been identified in *E. coli.* Bisulfite conversion is not 100% efficient, leading to incomplete conversion of unmethylated cytosines. This can result in false-positive interpretations of methylation status during analysis, as unconverted cytosines may be mistakenly identified as methylated. However, at every position called methylated in CpG and CHH context, the low coverage and/or percentage of methylation (<25%) drives them excluded. After filtering in this way on the bed files, 100% the hits found in a **CHG context**. Then, the samples that belong to the same clone and passed the quality controls were intersected and the resulting will be named after BSeq\_1 and BSeq\_2.

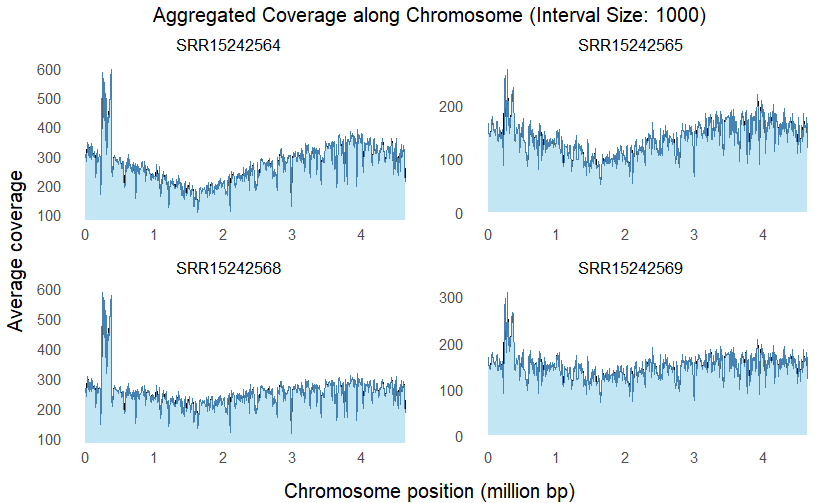


Figure 26. Coverage of the four selected bisulfite samples calculated as an average of the filtered reads’ single positions' coverage along 1000bp intervals on the E. coli genome. All of them have a very similar pattern discarding sample biases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Percentage of Cs methylated | % methylation (CpG context) | % methylation (CHG context) | % methylation (CHH context) |
| SRR15242564 | 4.2% | 7.143 | 85.714 | 7.143 |
| SRR15242565 | 4.2% | 7.143 | 85.714 | 7.143 |
| SRR15242568 | 4.2% | 7.143 | 88.095 | 4.762 |
| SRR15242569 | 4.5% | 8.889 | 84.444 | 6.667 |
| *Table 5. Parameters of methylation call on bisulfite data. Percentage of methylation of all the cytosines in reads sequenced, mapped and deduplicated. The next percentages are calculated over the first, for example 3.6% of Cs where on CHG context which is the 85.714% of the total 4.2% of methylations called.* | | | | |

The datasets that had the greatest number of methylations (both full and partial - above the red line on Figure 27) were selected to go on with the analysis, namely: 0-0, 4-4, 6-6, 7-7, SRR1536433, BSeq\_1 and BSeq\_2. Bed and bedgraph files of them are provided as annexes. Bedgraph files were analysed under the IGV to see even distribution of the methylations along the genome on both full and partial methylation cases. As seen in Figure 28, there is no significant positional bias, meaning that along the genome there are always methylation. Looking at the full representation, we can say that *E. coli*’s total methylation (methylation in all the individuals of a clonal population in a position) is the most frequent occurrence. Though, especially in the case of SMRT sequencing we can attribute this high value to the fact that the technology needs rather hight modification fraction detected (IPDratio high in many DNA strands mapping at the same positions) in order to be able to identify the modification. In fact, the lowest percentage in the gff files are 0-0: 0.428, 4-4: 0.412, 6-6: 0.319, 7-7: 0.398 and SRR1536433: 0.165.

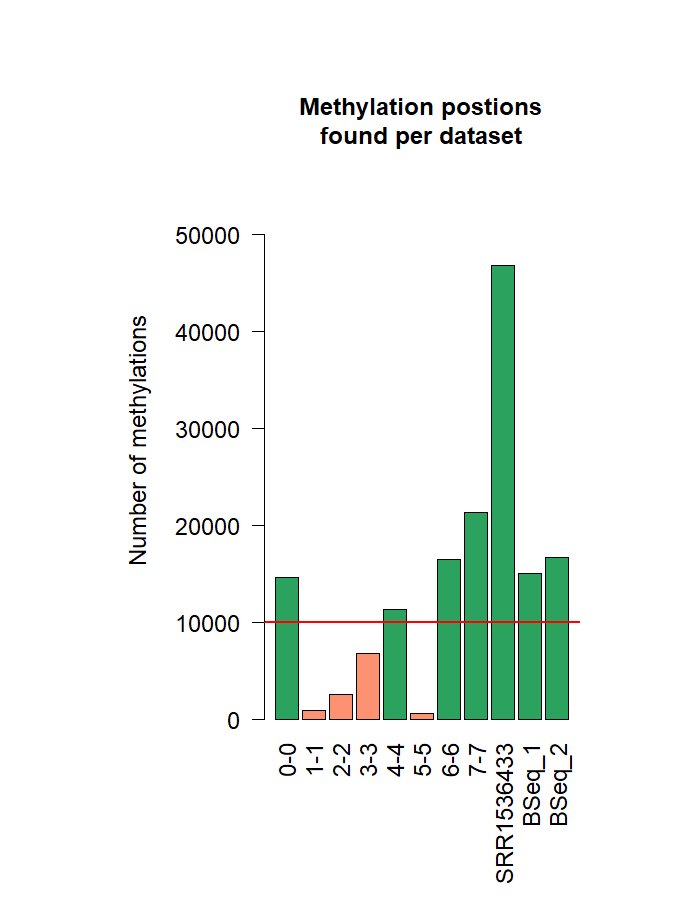


Figure 27. Number of methylated positions extracted per dataset. On the left, all of the positions extracted (at least on one read). On the right, the partially methylated positions 25-75%



Figure 28. IGV snapshoot of all the selected samples methylation position. Interspersed there are tracks representing full and filtered bedgraph files respectively for the samples BSeq\_1, BSeq\_2, 0-0, 4-4, 6-6, 7-7 and SRR1536433. On the top left of each track there is the scale from 0 to 1 in case of full methylation and 0 to 0.75 for the partial samples. Height of blue bars represent methylation percentage at each position. Methylations are distributed evenly and all the samples are highly methylated.

* **Gene association**

Significance tables (Table 6) for all operons pulled out hold between 26 and 139 operons englobing among 67 and 593 genes. For the case of metabolic genes from 12 to 78 operons holding from 37 to 319 metabolic genes.

The intersection performed between all combinations of the seven different datasets reveals that the genes that are found in the greater number of samples (6 out of the 7) are 15 in total: b2208, b2207, b2206, b2205, b2204, b2203, b2202, b2201, b2200, b2199, b2198, b2197, b2196, b2195, b2194, also known as napF, napD, napA, napG, napH, napB, napC, ccmA, ccmB, ccmC, ccmD, ccmE, ccmF, ccmG, ccmH belonging to a same operon called napFDAGHBC-ccmABCDEFGH. The following number of methylations were in the different samples: (0,9) (0,5) (0,13) (0,4) (0,6) (0,9); being the first number the amount in the regulatory part and the second on the operon body. napF, napD, ccmF, ccmG and ccmH are not considered metabolic genes, i.e., they don’t have associated any metabolic reaction although participate as structural proteins in the processes. Downstream from the main promoter are 15 genes, seven of which encode a periplasmic nitrate reductase (nap genes) that has two c-type cytochromes (NapB and NapC), and the remaining eight encode proteins essential for type I cytochrome c assembly (ccm genes).

*E. coli* K-12 has three nitrate reductases: NRA, NRZ, and Nap. NRA and NRZ are membrane-bound, while Nap is located in the periplasm. Nap expression is induced by anaerobiosis and low nitrate concentrations. It acts as a terminal electron acceptor, exploiting low nitrate levels efficiently and supporting various carbon sources. Thanks to Nap *E. coli* avoids the costs of pumping nitrate into the cell and has a higher affinity for nitrate than NRA. NapA and NapB conform the catalytic portion free in the periplasm, they perform the reaction NO3- + 2H+  NO2- + H2O (BIGGmodel reaction ids: NO3R1bpp, NO3R2bpp). They receive electrons from the quinone pool via NapC or/and NapGH, depending on the quinone type. NapD is involved in NapA assembly, and NapF contributes to nitrate reduction rate by an unknown mechanism and NapA maturation (EcoCyc: Encyclopedia of *E. coli* Genes and Metabolism n.d.).

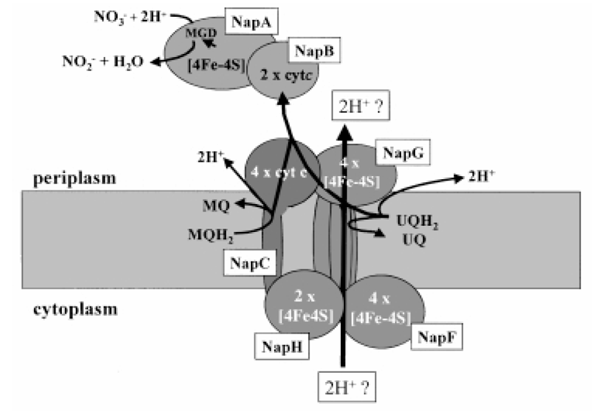


Figure 29. Nap nitrate reductase system in E. coli’s periplasm. MQ: menaquinone, UQ: ubiquinone, MQH2: meanquinol, UQH2: ubiquinol, NO3-: nitrate, NO2-: nitrite. (T. H. C. Brondijk 2002)

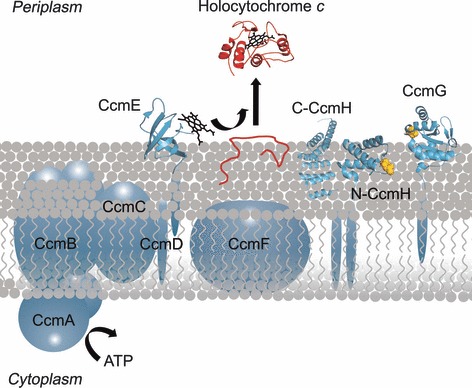


Figure 30. Cytochrome c biogenesis System I. Ccm proteins are depicted in blue, the apocytochrome c protein is shown in red. When heme (black) becomes covalently attached, it forms the holocytochrome c. The cysteine residues, assumed to be involved in reducing the -CXXCH- motif in the apocytochrome, are highlighted in yellow (Stevens et al. 2011)

Protoheme, synthesized in the bacterial cytoplasm, is delivered to CcmCDE complex across the membrane by the CcmA-CcmB ABC transporter. ATP hydrolysis by CcmAB facilitates the release of holo-CcmE for heme attachment to the -CXXCH- motif of apocytochromes c in the periplasm or membrane-associated proteins giving rise to holocytochromes c (Stevens et al. 2011). BIGG model reaction id is PHEMEabcpp.

We can hypothesize that this operon’s gene expression may be regulated by methylation as well as transcriptional regulators like Fnr, ModE, Nac, FlhDC coupled the sensory system narL-narP (Figure 31). These genes are not essential for growth in aerobic conditions; therefore, it may be a bed hedging strategy that prepares *E. coli* for anaerobiosis.

Other operons with high occurrence found and corresponding number of partial methylations are:

* 5/7: rplKAJL-rpoBC (0,4) (0,5) (0,11) (0,4) (0,6) \*\*
* 4/7: narGHJI (0,8), (0,4) (0,10) (0,8)
* 4/7: glgBXCAP (0,5) (0,4) (0,6) (0,4)
* 4/7: nuoABCEFGHIJKLMN (0,6) (0,4) (0,4) (0,4)
* 4/7: bamA-skp-lpxD-fabZ-lpxAB-rnhB-dnaE (0,4) (0,4) (0,6) (0,10) \*
* 4/7: hyfABCDEFGHIJR-focB (0,4) (0,7) (0,4) (0,5) \*
* 4/7: cmoM-mukFEB (0,6) (0,7) (0,9) (0,9) \*\*
* 4/7: ebgAC (0,4) (0,4) (0,4) (0,6) \*\*
* 4/7: ycjMNOPQRSTUV-ymjB-ompG (0,4) (0,4) (0,4) (0,5) \*

*\*\* implies that the operon is not metabolic at all, \* means some of the genes in the operon are metabolic. The rest of operons are fully metabolic*

The size of the operon could be a factor to consider, larger operons have a higher probability of having more methylations than smaller operons if they are spread stochastically. In addition, the thresholds were selected arbitrarily, and we should watch case by case how methylations in each regulatory region or coding sequence affect the expression level. Finally, I separated regulatory areas from coding areas in two different categories, being both incompatible however they coding regions can also have regulatory sequences but RegulonDB does not annotate them specifically, this information missing, that is why any of the high occurrence operons was selected with a partially methylated regulatory position. Additional experiments are needed to handle this issue.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| operon | Frec reg | Frec cds | Locus tag | Gene symbol |
| paaABCDEFGHIJK | 1 | 8 | b1388, b1389, b1390, b1391, b1392, b1393, b1394, b1395, b1396, b1397, b1398 | paaA, paaB, paaC, paaD, paaE, paaF, paaG, paaH, paaI, paaJ, paaK |
| acs-yjcH-actP | 1 | 0 | b4069, b4068, b4067 | acs, yjcH, actP |
| nrdDG | 1 | 0 | b4238, b4237 | nrdD, nrdG |
| rplY | 1 | 0 | b2185 | rplY |
| nudF-yqiB-cpdA-yqiA-parE | 0 | 6 | b3034, b3033, b3032, b3031, b3030 | nudF, yqiB, cpdA, yqiA, parE |
| yahDEFG | 0 | 6 | b0318, b0319, b0320, b0321 | yahD, yahE, yahF, yahG |
| entCEBAH | 0 | 4 | b0593, b0594, b0595, b0596, b0597 | entC, entE, entB, entA, entH |
| frsA | 0 | 4 | b0239 | frsA |
| ilvLXGMEDA | 0 | 4 | b3766, b4669, b3769, b3770, b3771, b3772 | ilvL, ilvX, ilvG, ilvM, ilvE, ilvD, ilvA |
| proBA | 0 | 4 | b0242, b0243 | proB, proA |
| *Table 6. Part of the whole operon-gene table of 7-7 sample containing an example of 10 randomly sampled operons and corresponding genes that were above the methylation thresholds (to see the whole tables for all the datasets in annexes).* | | | | |

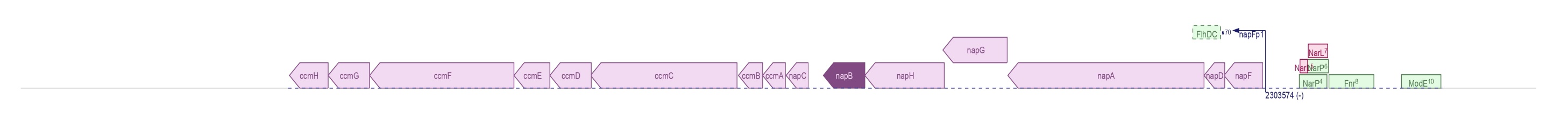
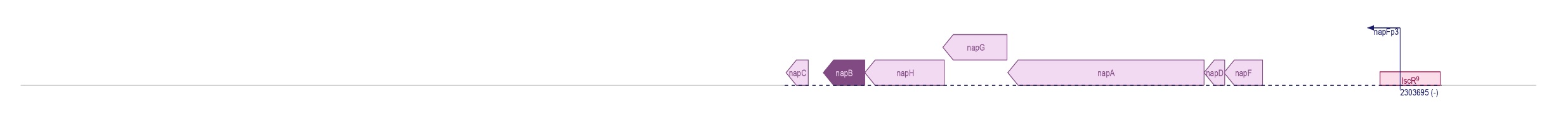
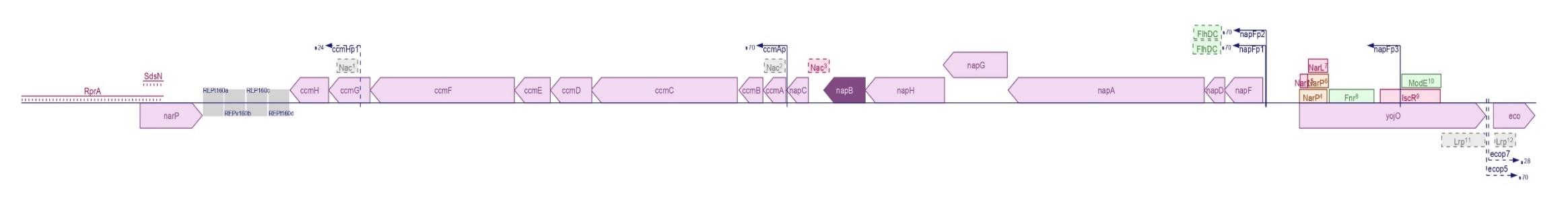
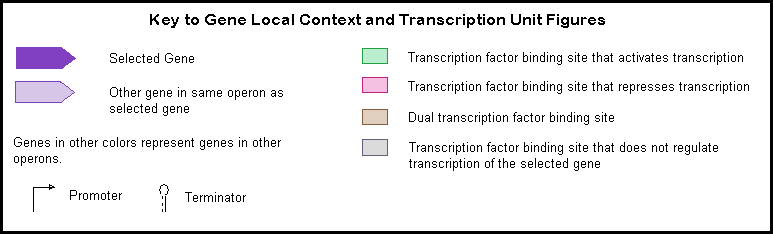
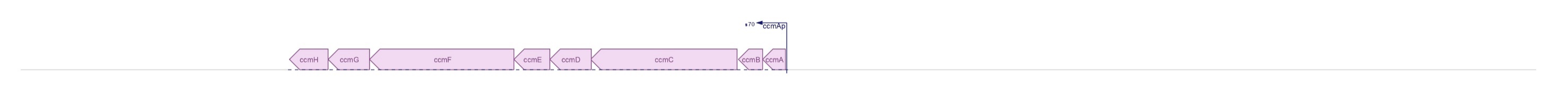


Figure 31. A. napFDAGHBC-ccmABCDEFGH operon on genomic context. B transcription units. C. legend. (EcoCyc: Encyclopedia of E. coli Genes and Metabolism n.d.)

A

B

C



* **Functional enrichment:**

Overall, there is a tendency of general enrichment in metabolic genes (Table 7). In six out of seven samples the p-value is rather low (lower than 0.005) for the Fisher’s exact test meaning that, considering the whole *E. coli* gene annotation and corresponding metabolic gene proportion, within the gene calls obtained there is a bigger ratio of genes that are metabolic than the proportion expected if the sampling was done randomly.

|  |  |
| --- | --- |
| Gene Ontology Biological Process term | Ocurrence |
| cytochrome complex assembly | 6 |
| energy derivation by oxidation of organic compounds | 4 |
| generation of precursor metabolites and energy | 4 |
| aerobic respiration | 3 |
| benzene-containing compound metabolic process | 3 |
| cellular respiration | 3 |
| cellular response to xenobiotic stimulus | 3 |
| electron transport chain | 3 |
| histidine biosynthetic process | 3 |
| histidine metabolic process | 3 |
| organic phosphonate metabolic process | 3 |
| oxidative phosphorylation | 3 |
| phenylacetate catabolic process | 3 |
| response to xenobiotic stimulus | 3 |
| xenobiotic catabolic process | 3 |
| xenobiotic metabolic process | 3 |
| anaerobic respiration | 2 |
| branched-chain amino acid transport | 2 |
| enterobacterial common antigen biosynthetic process | 2 |
| enterobacterial common antigen metabolic process | 2 |
| heme transport | 2 |
| monocarboxylic acid catabolic process | 2 |
| nitrogen cycle metabolic process | 2 |
| nucleoside triphosphate biosynthetic process | 2 |
| purine nucleoside triphosphate biosynthetic process | 2 |
| purine ribonucleoside triphosphate biosynthetic process | 2 |
| respiratory electron transport chain | 2 |
| ribonucleoside triphosphate biosynthetic process | 2 |
| *Table 8. GO Biological terms enriched in more than one sample for metabolic genes found* | |

|  | p-value |
| --- | --- |
| **BSeq\_1** | 0.000000540 |
| **BSeq\_2** | 0.002499223 |
| **0\_0** | 0.004626341 |
| **4\_4** | 0.000000029 |
| **6\_6** | 0.304574031 |
| **7\_7** | 0.000061153 |
| **SRR1536433** | 0.002499223 |
| Table 7. P-values of Fisher’s exact test for metabolic genes enrichment | |

On the other hand, we can further analyse the set of **metabolic genes** pulled by seeking enrichment in GO terms. Different datasets happen to be enriched in diverse terms - pass the threshold (qval 0.005) -(Figure 32) however, there are some that appear in more than one sample (Table 8). The most remarkable terms have to do with energy obtention and growth for example: “cytochrome complex assembly”, “energy derivation by oxidation of organic compounds”, “generation of precursor metabolites and energy”, “aerobic respiration”, “cellular respiration”, “electron transport chain”, “oxidative phosphorylation”, “anaerobic respiration” among others. Xenobiotic-related terms are highly remarkable as well. That means that populations may be prepared to detoxify chemical substances that are not naturally produced within the organism and may encounter.

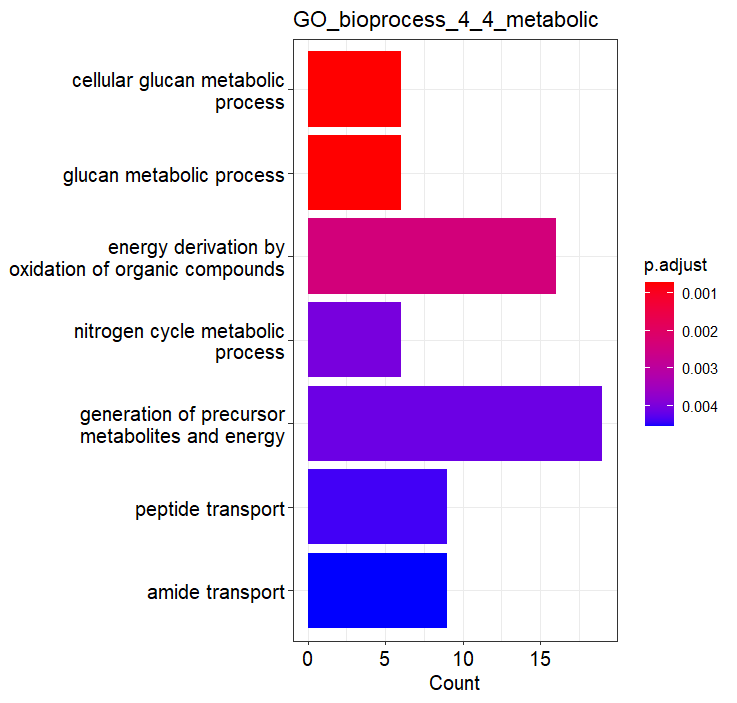
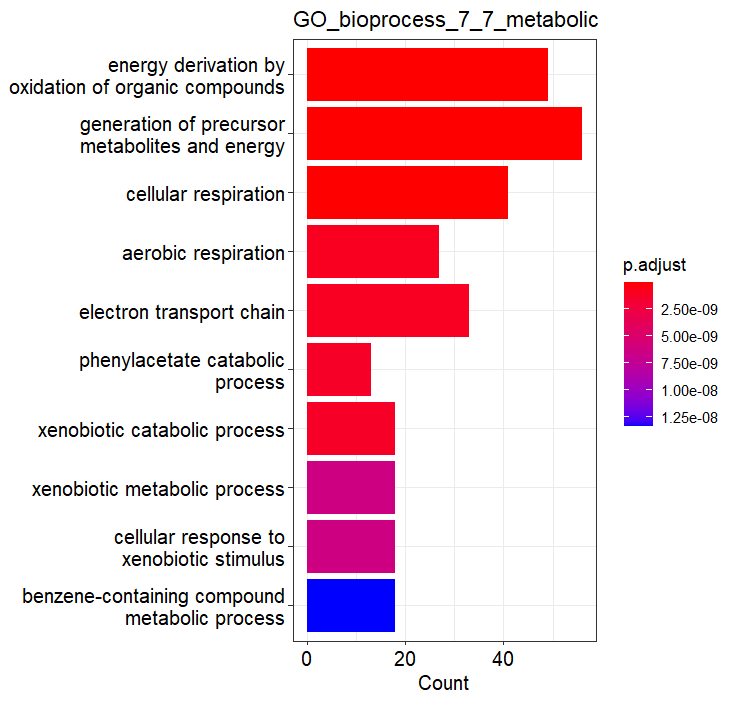
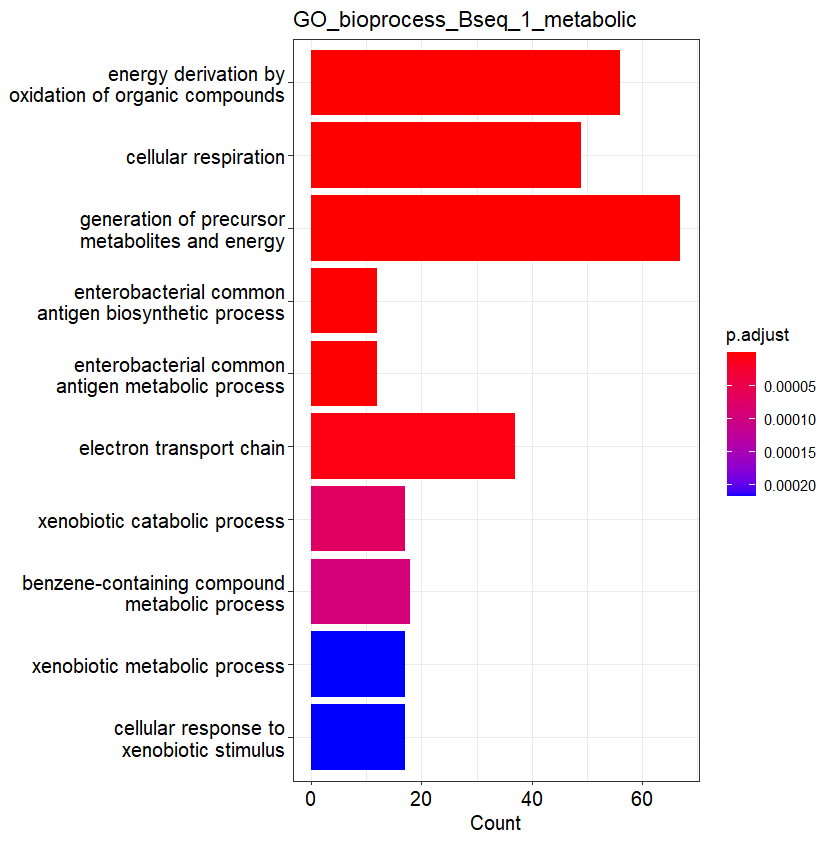


Figure 32. Gene ontology Biological Process terms enriched in three of the datasets for the set of metabolic genes. For sake of commodity, there are a maximum of 10 terms represented in each sample, the 10 found with the lowest qvalue. However, there are more terms enriched in the cases of 7-7 (23) and Bseq\_1 (16).

|  |  |
| --- | --- |
| KEGG pathway | Ocurrence |
| Oxidative phosphorylation | 5 |
| Phenylalanine metabolism | 3 |
| Quorum sensing | 3 |
| ABC transporters | 2 |
| Nitrogen metabolism | 2 |
| Propanoate metabolism | 2 |
| Two-component system | 2 |
| *Table 9. KEGG pathways enriched in more than one sample of metabolic genes found* | |

In addition to GO BP terms, KEGG pathways enriched identified appear in Figure 33 and repetitive pathways in Table 9.

Ultimately, the corresponding tables and Figures for the enrichment analysis of GO terms and KEGG pathway for all the genes extracted, not only the metabolic ones, can be found in the appendix 4) (supplementary tables 1 and 2) and (supplementary figures 2 and 4) alongside the missing Figures of metabolic GO and KEGG enrichment (Supplementary figures 1 and 3).

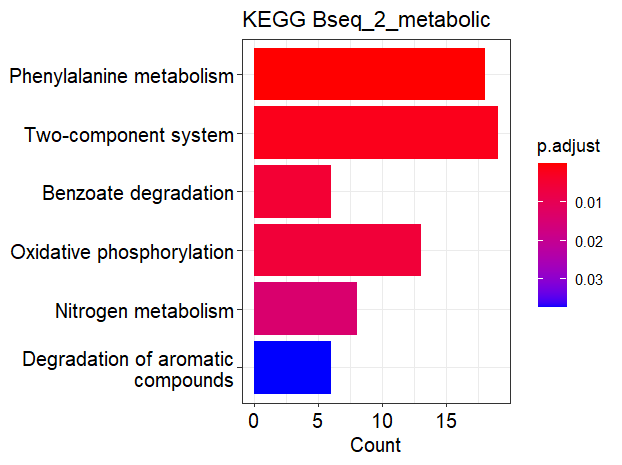
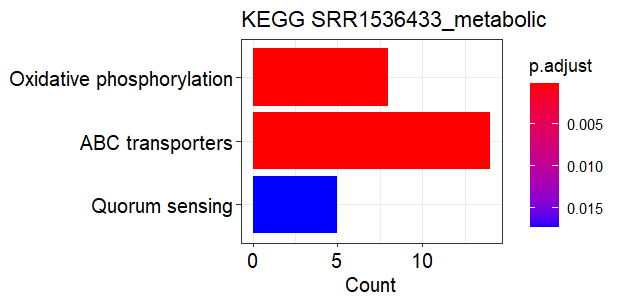
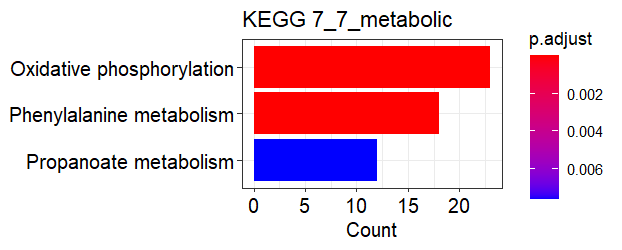


Figure 33. KEGG pathway terms enriched in the different samples for the set metabolic genes. All the enriched terms that pass the threshold of the test 0.05 are represented.

* **Motif identification**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Full 0-0 | Filtered 0-0 | Full 4-4 | Filtered 4-4 | Full 6-6 | Filtered 6-6 | Full 7-7 | Filtered 7-7 | Full SRR1536433 | Filtered SRR1536433 |
| fraction\_partial | 4.3 | | 4.3 | | 4.5 | | 4.9 | | 1.1 | |
| percentage\_motifs | 88.23 | 89.05 | 88.95 | 93.5 | 87.07 | 92.45 | 86.85 | 91.87 | 84.95 | 91.3 |
| percentage\_motifs\_CCWGG | 0.77 | 0.71 | 0.71 | 0.22 | 0.78 | 0.1458 | 0.8549 | 0.42 | 1.74 | 0.22 |
| fraction\_CCWGG | 4.04 | | 1.39 | | 0.90 | | 2.516 | | 0.14 | |
| score\_CCWGG | 0.83 | 0.9 | 0.82 | 0.8 | 0.82 | 1 | 0.8214 | 0.8 | 0.89 | 0.8 |
| percentage\_motifs\_AACNNNNNNGTGC | 0.87 | 0.71 | 0.86 | 0.43 | 0.91 | 0.7289 | 0.957 | 0.62 | 1.55 | 0.89 |
| fraction\_AACNNNNNNGTGC | 3.54 | | 2.30 | | 3.84 | | 3.371 | | 0.65 | |
| score\_AACNNNNNNGTGC | 0.998 | 1 | 0.998 | 1 | 0.997 | 0.984 | 0.9955 | 0.97 | 0.997 | 1 |
| percentage\_motifs\_GATC | 98.34 | 98.57 | 98.43 | 99.13 | 98.29 | 98.98 | 98.18 | 98.86 | 96.64 | 98.89 |
| fraction\_GATC | 4.35 | | 4.56 | | 4.81 | | 5.203 | | 1.16 | |
| score\_GATC | 0.999 | 0.998 | 0.999 | 0.995 | 0.9986 | 0.9978 | 0.9987 | 0.998 | 0.998 | 0.99 |
| percentage\_motifs\_ATGCA | 0.023 | 0 | 0.0098 | 0 | 0.021 | 0 | 0.011 | 0 | 0.073 | 0 |
| fraction\_ATGCAT | 0 | | 0 | | 0 | | 0 | | 0 | |
| score\_ATGCAT | 0.83 | NA | 0.83 | NA | 0.83 | NA | 0.83 | NA | 0.84 | NA |
| *Table 10. Summary of motif analysis results in SMRT datasets. Fraction partial represents the percentage of the total found methylated peaks that were between 25-75%. In blue there is the filtered data and in orange the full data. The darkest row (percentage motifs) represents the proportion of the methylated peaks found in which any of the 4 motifs matched over the total. Within the matched peaks, slightly lighter, there are the percentages of each of the motifs. Finally, the lightest tones hold the average scores of each hit. The green cells represent how much of each motif in the full data is in the filtered data* | | | | | | | | | | |

The numbers obtained from the motif identification results in all SMRT samples indicate a higher percentage of motifs found in filtered data over full (first row in Table 10). It can be stated than the motif with the very most occurrence is GATC in all of the cases being above 96% of the known motifs found. The other motifs are almost insignificant. In the case of ATGCAT motif the percentage in full data is so low that in filtered disappears. In most of the SMRT samples the fraction of partial methylation over full is conserved around 4,5% but in SRR1536433 (table 10). Already in figure 27 we can observe that the number of full positions found in the sample SRR1536433 is much larger compared to others and not so the filtered. So can be observed, in SRR1536433\_full.bed track where a much denser picture is seen compared to the other SMRT samples.

As for the bisulfite dataset, in both of the clones all the motifs found both in full and filtered methylation peaks correspond to CCWGG, the only motif with 5mC known in E. coli. The average score (1.00) corresponds to a perfect match in all position (Table 12).

|  |  |  |
| --- | --- | --- |
|  | BSeq\_1 | BSeq\_2 |
| **fraction\_partial** | 11.54 | 5.53 |
| **percentage\_CCWGG\_full** | 100.00 | 100.00 |
| **score\_CCWGG\_full** | 1.00 | 1.00 |
| **percentage\_CCWGG\_filtered** | 100.00 | 100.00 |
| **score\_CCWGG\_filtered** | 1.00 | 1.00 |
| *Table 12. Summary of motif analysis results in bisulfite datasets.* | | |

We can say that there is not significant degeneracy of motifs that are partially methylated vs all the methylation set because the scores are almost identical. If the partial methylation “selection” positions were random each motif ought to be represented as much as the fraction partial of the sample, however, there is a tendency of higher GATC representation and a tendency of lower proportion of other motifs in filtered data. This effect can be due to the fact that values are very small thus loose significance. Another reason that can be complementary is because GATC is the shortest of the motifs studied, and when using the RSAT motif finder one substitution was allowed for any of the motifs queried. One substitution represents a bigger error in 4 nt motif than in 13 nt motif (the case of AACNNNNNNGTGC). In a nutshell, there are some confounding factors that avoid from predicting that any of the motifs is enriched in partial methylation, then the null hypothesis that the distribution is the same is accepted. With this I continued to analyse the sequences in the full data not matching any of the motifs known for the SMRT datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Logo | E-value | Sites |
| 0\_0 |  | 1.8e-5 | 921 (53.4%) |
| 4\_4 |  | 4.9e-4 | 738 (58.4%) |
| 6\_6 |  | 1.9e-6 | 1249 (58.6%) |
| 7\_7 |  | 1.7e-12 | 1290 (45.8%) |
| SRR1536433 |  | 4.1e-14 | 5524 (74.1%) |
| *Table 13. Motif search with STREME* | | | |

* **GATC motif genomic analysis**

In the genome the GATC sites are distributed 53% of them in intergenic regions and 47% on cds. Considering both strands separately the intergenic regions amount to 57% of the total genome length and 43% are cds. A binomial test - binom.test(53,100,p=0.507) - states that the null hypothesis (pval= 0.4211) is true so GATC sites are distributed evenly along the genome.

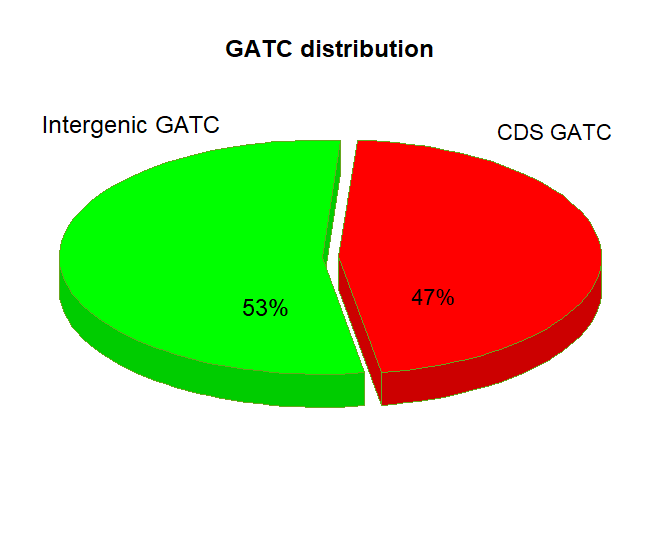


Figure 35. GATC distribution

The amount partial methylations positions found on intergenic regions for each of the SMRT datasets suggest that this data follow the same distribution as GATC motifs in the genome, i.e., even distribution.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Positions | 0-0 | 4-4 | 6-6 | 7-7 | SRR1536433 |
| cds | 284 | 224 | 348 | 493 | 251 |
| intergenic | 346 | 267 | 393 | 552 | 243 |
| Percentage intergenic | 54.92 | 54.37 | 53.04 | 52.83 | 49.20 |
| Pvalue binom. test | 0.3384 | 0.5569 | 1 | 0.9259 | 0.09531 |
| *Table 13. Number of partial methylation positions found at each region for SMRT datasets* | | | | | |

|  |  |  |
| --- | --- | --- |
|  | Percentage (%) | Percentage after correction |
| 0\_0 | 33.25386 | 33.21396 |
| 4\_4 | 26.16441 | 26.13825 |
| 6\_6 | 36.90137 | 36.84971 |
| 7\_7 | 47.74440 | 47.68234 |
| SRR1536433 | 100.59339 | 100.37209 |
| *Table 14. Percentage of detected fully/partially methylated*  *motifs with respect to the total GATC motifs counted in the*  *genome* | | |

Another experiment performed was to pick sequences of 100 nt randomly both in the whole genome and only on regions annotated as cds. Scaling by the content of each 4mer in each selected cds sequence we have an average chance of 33.57% of having at least one GATC motif in our sequence being the average of the genome sequences 31.82%. The two-sample t-test indicates a highly significant result. The p-value is reported as less as 2.2e-16 suggesting strong evidence against the null hypothesis. The alternative hypothesis suggests that there is indeed a difference in means between the two distributions. Additionally, the 95 percent confidence interval reported (-0.01759950 to -0.01746133) does not include 0, further supporting the conclusion that there is a significant difference between the means of the two groups. I have higher likelihood to find a GATC motif in a sequence picked from cds that from the background representation. However, this is not agreeing with the previous experiment where all the GATC sites were measured and there was even distribution seen. The most likely reason is that there is a substructure inside the cds where some regions are richer in GATC than others and those were the ones were selected, also, any.

In table 13 there are shown the percentage of GATC motifs found methylated either fully or partially with respect to the total number of motifs that appear un the genome. In SRR1536433 we can see that all of them are detected to be methylated (in at least one organism). There are more than there should be in the genome which could be due to small error introduced by considering the degeneracy of motifs as said in the sub epigraph motif identification, but it can be easily corrected by multiplying by the score of GATC since the perfect match is 1 and every other motif is lowering the score, this value represents the proportion of perfect matches as well. Nevertheless, the percentages change fairly little and there might be some other kind of bias probably due to errors in the methylation detection as it can be observed in figure 34, the discovered “motif” is likely to be formed by mistakes in modification detection since it differs a lot from the other samples.

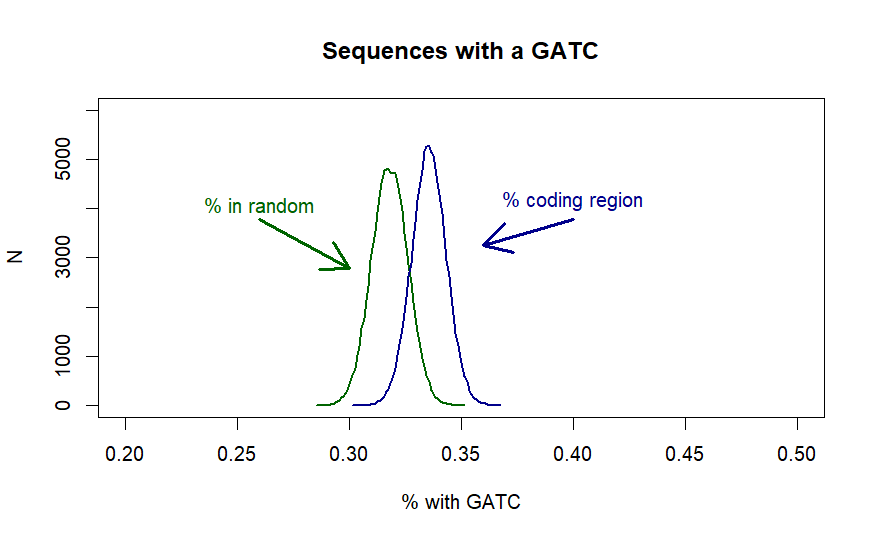


Figure 38. Likelihood finding at least one GATC motif in a 100bp sequence picked from the cds or the whole background genome.

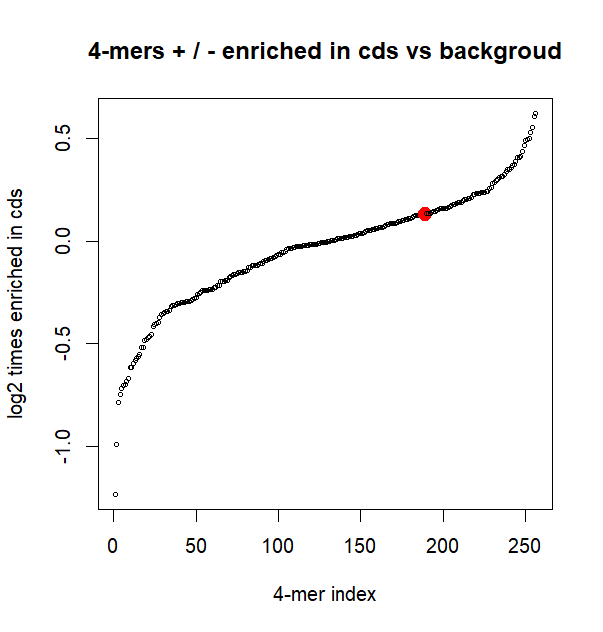


Figure 37. Log2 fold enrichment of 4-mers in random subsamples of cds with respect to the background

* **Modelling**

Loss of ATP hydrolysis activity by CcmAB results in loss of *c*-type cytochrome synthesis and incomplete processing of CcmE (Christensen et al. 2007). NapB: periplasmic nitrate reductase cytochrome c550 protein (RegulonDB n.d.)

PHEMEabcpp GPR: b2198 and b2199 and b2197 and b2200 and b2201= ccmA,ccmB,ccmC,ccmD,ccmE (ccmF,G and H are not metabolic genes)

* **Limitations and future perspectives**

# **APPENDICES**

1. Bisulfite data processing commands

*Splitting*

$ python separate\_fastq.py <input.fastq.gz> <output.read1.fastq.gz> <output.read2.fastq.gz>

$ fastqc <fileread1.fastq.gz> <fileread2.fastq.gz>

*Mapping, methylation extraction and report generation*

$ bismark\_genome\_preparation --parallel 8 <dir\_fasta\_ref\_genome>

$ bismark <genome path> -p 4 --parallel 6 -q --pbat -o <output folder> -1 <read1.fastq.gz> -2 <read2.fastq.gz>

$ deduplicate\_bismark -p --output\_dir <output folder> --bam <input bam file>

$ samtools view -h -q 20 <input bam file > -o <outputname deduplicated\_filtered.bam>

$ bismark\_methylation\_extractor --parallel 10 -p --no\_overlap --ignore 7 --ignore\_r2 7 --ignore\_3prime 5 --ignore\_3prime\_r2 5 -o <output folder> --report --bedGraph --CX --cytosine\_report --genome\_folder <genome path> <input deduplicated\_filtered.bam>

$ bismark2report --alignment\_report <input PE\_report.txt> --dedup\_report <input deduplication\_report.txt> --splitting\_report <input deduplicated\_splitting\_report.txt> --mbias\_report <input deduplicated.M-bias.txt>

*Coverage and representation (igv)*

$ samtools sort -o <outputname deduplicated\_filtered\_sorted.bam> <input deduplicated\_filtered.bam>

$ samtools index <input deduplicated\_filtered\_sorted.bam>

$ samtools depth <input deduplicated\_filtered\_sorted.bam> > <output deduplicated\_filtered\_coverage>

$ samtools depth <input deduplicated\_filtered\_sorted.bam> | awk '{sum+=$3} END { print "Average = ",sum/NR}'

*Selecting positions within 25-75% of methylation and coverage >25 and common positions (filtered)*

$awk 'BEGIN{OFS="\t"} {$8=$6;$9=$7; $6=$4\*100/($4+$5+1);$7=$6;$6=$5;$5=$4;$4=$3;$3=$2+1;$10=$5+$6 ; print $0}' <input CX\_report.txt> > <output CX\_report\_percentaje.txt>

$awk '$7 >= 25 && $7 <= 75 && $10 >= 25' <input CX\_report\_percentaje.txt> > <output CX\_report\_filtered.txt>

$bedtools intersect -a <input1 CX\_report\_filtered.txt> -b <input2 CX\_report\_filtered.txt> -wa -wb | awk '{OFS="\t"; print "NC\_000913", $2, $3, $4, ($7+$17)/2}' > <clone\_1/2\_intersection\_filtered.bed>

$awk 'BEGIN{OFS="\t";print "track type=bedGraph"}{print $1,$2,$3,$5}' <clone\_1/2\_intersection\_filtered.bed> > <clone\_1/2\_intersection\_filtered.bedgraph>

*Selecting positions above 25% of methylation and coverage >25 and common positions (full)*

$awk '$7 >= 25 && $10 >= 25' <input CX\_report\_percentaje.txt> > <output full.txt>

$bedtools intersect -a <input1 full.txt> -b <input2\_full.txt> -wa -wb | awk '{OFS="\t"; print "NC\_000913", $2, $3, $4, ($7+$17)/2}' > <clone\_1/2\_intersection\_full.bed>

$awk 'BEGIN{OFS="\t";print "track type=bedGraph"}{print $1,$2,$3,$5}' <clone\_1/2\_intersection\_full.bed> > <clone\_1/2\_intersection\_full.bedgraph>

1. SMRT data processing commands

*Demultiplexing n-n indexes*

$lima --same --split-bam <input subreads.bam> <barcodes\_8plex.fasta> <output barcoded.subreads.demux.bam>

*Mapping and indexing*

$pbmm2 align <input subreads.demux.bam> <reference.fasta> <output subreads\_aligned.bam> --sort

$samtools faidx <reference.fasta> ???

$pbindex <aligned\_subreads.bam> ?????

*Methylation extraction*

$ipdSummary <input subreads\_aligned.bam> --reference <reference.fasta> --gff <output gff> --pvalue 0.001 --numWorkers 16 --identify m4C,m6A,m5C\_TET --methylFraction --useChemistry "SP3-C3"

*Select only those positions with methylation fraction*

$awk -F "\t|;" -v OFS="\t" 'NF==15 {print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12,$13,$14,$15}' <input gff> > <output percentage.gff>

*Full*

$awk -F "\t" -v OFS="\t" '{gsub(/^frac=/, "", $12); print $1, $4-1, $5, $7, $12}' <input percentage.gff> > <output full.bed>

*Lowest fraction*

$awk -F "\t" 'NR == 1 || $5 < min { min = $5 } END { print min }' <input full.bed>

*Filtered 25-75%*

$awk -F"\t" '$5 < 0.75 {print $1,$2,$3,$4,$5}' <input full.bed> > <output filtered.bed>

*In bedgraph format for IGV visualization*

$awk -F "\t" -v OFS="\t" 'BEGIN{print "track type=bedGraph"}{print "NC\_000913",$2,$3,$5}' <input filtered.bed> > <output filtered.bedgraph>

$awk -F "\t" -v OFS="\t" 'BEGIN{print "track type=bedGraph"}{print "NC\_000913",$2,$3,$5}' <input full.bed> > <output full.bedgraph>

1. Methylation motifs commands

*Full fasta function*

function create\_fasta\_full() {

local file\_name="$1"

local file\_prefix="${file\_name%\_percentage.gff\*}"

local arg="$2"

if [ "$arg" = "14" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-3,$5+11,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_14\_full.fa"

elif [ "$arg" = "4" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-2,$5+2,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_4\_full.fa"

elif [ "$arg" = "22" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-11,$5+11,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_22\_full.fa"

elif [ "$arg" = "5" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-2,$5+3,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_5\_full.fa"

elif [ "$arg" = "6" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-2,$5+4,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_6\_full.fa"

else

echo "Invalid argument. The second argument must be either 4, 5, 14 or 22."

return 1

fi

rm x

echo "Fasta file created successfully as ${file\_prefix}\_seq\_${arg}\_full.fa"

}

*Filtered fasta function*

function create\_fasta\_filtered() {

local file\_name="$1"

local file\_prefix="${file\_name%.bed}"

local arg="$2"

if [ "$arg" = "14" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F '\t' 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$2-2,$3+11,$4,$5}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_14\_filtered.fa"

elif [ "$arg" = "4" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F '\t' 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$2-1,$3+2,$4,$5}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_4\_filtered.fa"

elif [ "$arg" = "5" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F '\t' 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$2-1,$3+3,$4,$5}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_5\_filtered.fa"

elif [ "$arg" = "22" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F '\t' 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$2-10,$3+11,$4,$5}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_22\_filtered.fa"

elif [ "$arg" = "6" ]; then

python2.7 /home/earmenteros/mEpigram/bedToFasta.py -f <(awk -F "\t" 'BEGIN{OFS="\t"}{print "NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome",$4-2,$5+4,$7,$6}' "${file\_name}") -r /home/earmenteros/reference\_genome/sequence.fasta -o x && awk '/^>/{print ">seq\_" ++i; next}{print}' x > "${file\_prefix}\_seq\_6\_filtered.fa"

else

echo "Invalid argument. The second argument must be either 4, 5, 6 or 22."

return 1

fi

rm x

echo "Fasta file created successfully as ${file\_prefix}\_seq\_${arg}\_filtered.fa"

}

*For BSEq:*

$create\_fasta\_filtered <clone\_1/2\_intersection\_full.bed> 22

$create\_fasta\_filtered <clone\_1/2\_intersection\_filtered.bed> 22

$mv <clone\_1/2\_intersection\_full\_seq\_22\_filtered.fa> <BSeq\_1/2\_full.fa>

$mv <clone\_1/2\_intersection\_filtered\_seq\_22\_filtered.fa> <BSeq\_1/2\_filtered.fa>

*For pacbio*

$create\_fasta\_full <input percentage.gff> 22

$create\_fasta\_filtered <input bed> 22

*STREME*

$streme --evalue --p <no\_motif\_full.fa> --oc <no\_motif\_full>

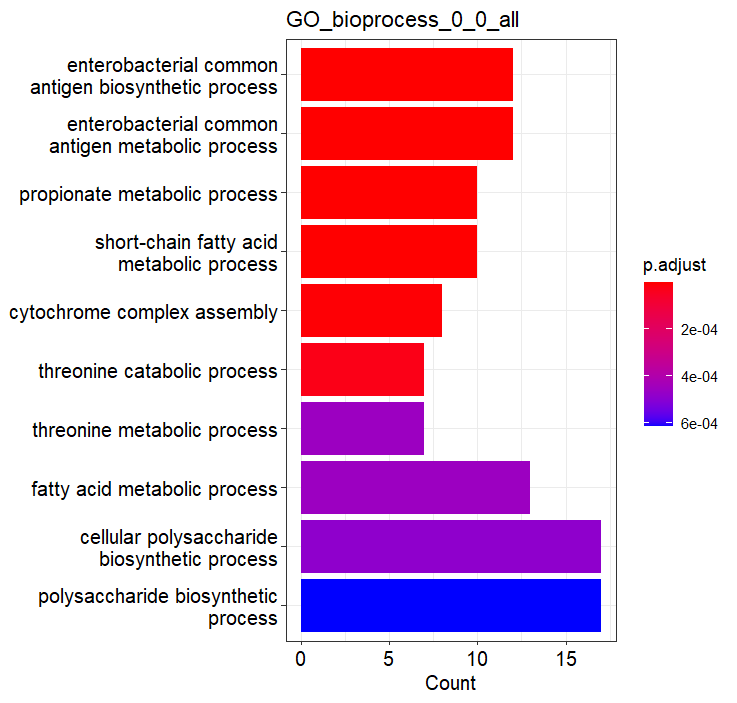
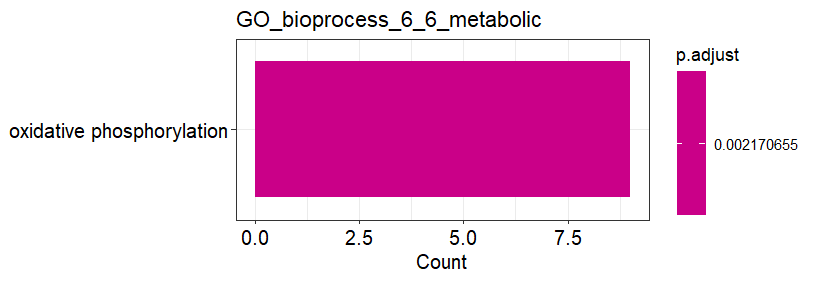
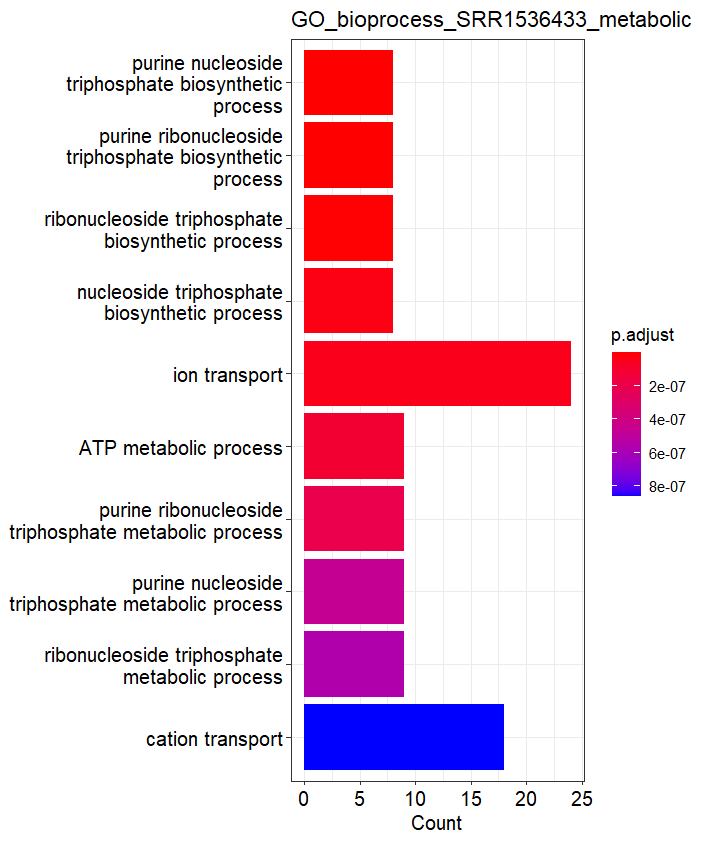
|  |  |
| --- | --- |
| LB media components | |
| Cobrapy  abbreviation | element |
| adn | Adenosine |
| ala\_\_L | L-Alanine |
| amp | AMP |
| arg\_\_L | L-Arginine |
| aso3 | Arsenite |
| asp\_\_L | L-Aspartate |
| ca2 | Ca2+ |
| cbl1 | Vitamin B12 |
| cd2 | Cadmium |
| cl | Cl- |
| cmp | CMP |
| cobalt2 | Co2+ |
| cro4 | Chromate |
| cu2 | Cu2+ |
| cys\_\_L | L-Cysteine |
| dad\_2 | Deoxyadenosine |
| dcyt | Deoxycytidine |
| fe2 | Fe2+ |
| fe3 | Fe3+ |
| fol | Folate |
| glc\_\_D | D-Glucose |
| glu\_\_L | L-Glutamate |
| gly | Glycine |
| gmp | GMP |
| gsn | Guanosine |
| h2o | H2O |
| h2s | Hydrogen sulfide |
| h | H+ |
| hg2 | Hg2+ |
| his\_\_L | L-Histidine |
| hxan | Hypoxanthine |
| ile\_\_L | L-Isoleucine |
| ins | Inosine |
| k | K+ |
| leu\_\_L | L-Leucine |
| lipoate | Lipoate |
| lys\_\_L | L-Lysine |
| met\_\_L | L-Methionine |
| mg2 | Mg |
| mn2 | Mn2+ |
| mobd | Molybdate |
| na1 | Na+ |
| nac | Isonicotinic acid |
| nh4 | Ammonium |
| ni2 | Ni2+ |
| o2 | O2 |
| phe\_\_L | L-Phenylalanine |
| pheme | Protoheme |
| pi | Phosphate |
| pnto\_\_R | PAN |
| pro\_\_L | L-Proline |
| pydx | Pyridoxal |
| ribflv | Riboflavin |
| ser\_\_L | L-Serine |
| so4 | Sulfate |
| thm | Thiamin |
| thr\_\_L | L-Threonine |
| thymd | Thymidine |
| trp\_\_L | L-Tryptophan |
| tyr\_\_L | L-Tyrosine |
| ump | UMP |
| ura | Uracil |
| uri | Uridine |
| val\_\_L | L-Valine |
| zn2 | Zn2+ |

1. Media

|  |  |
| --- | --- |
| M9 media components | |
| Cobrapy  abbreviation | element |
| ca2 | Ca2+ |
| cl | Cl- |
| cobalt2 | Co2+ |
| cu2 | Cu2+ |
| fe2 | Fe2+ |
| fe3 | Fe3+ |
| glc\_\_D | D-Glucose |
| h2o | H2O |
| h | H+ |
| k | K+ |
| mg2 | Mg |
| mn2 | Mn2+ |
| mobd | Molybdate |
| na1 | Na+ |
| nh4 | Ammonium |
| ni2 | Ni2+ |
| o2 | O2 |
| pi | Phosphate |
| so4 | Sulfate |
| zn2 | Zn2+ |

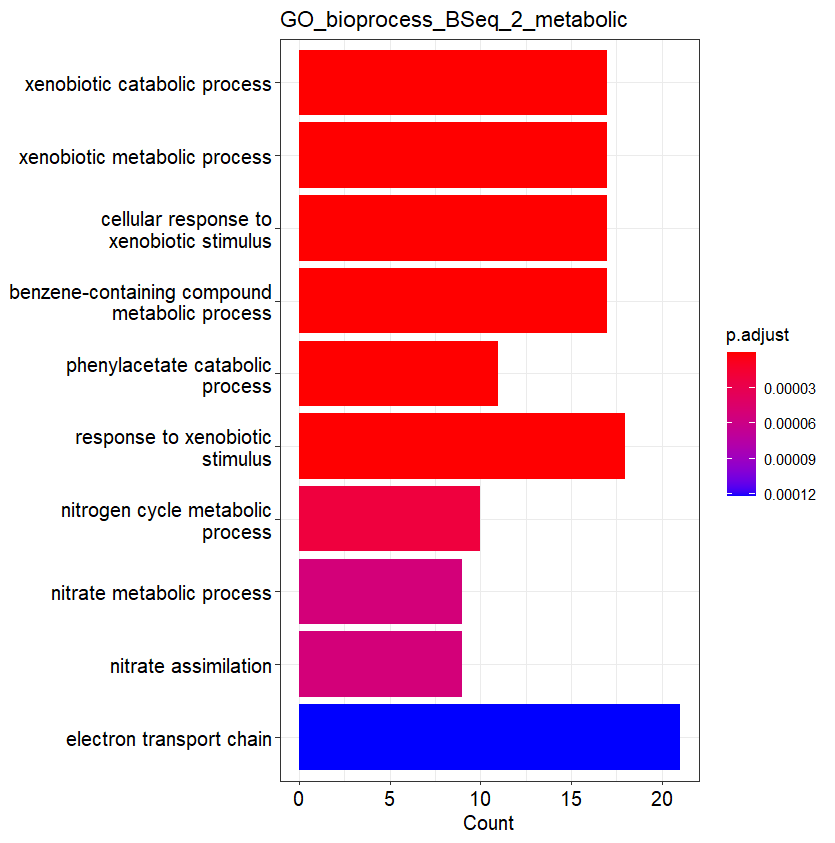
1. Enrichment data for all genes found

**GO terms**

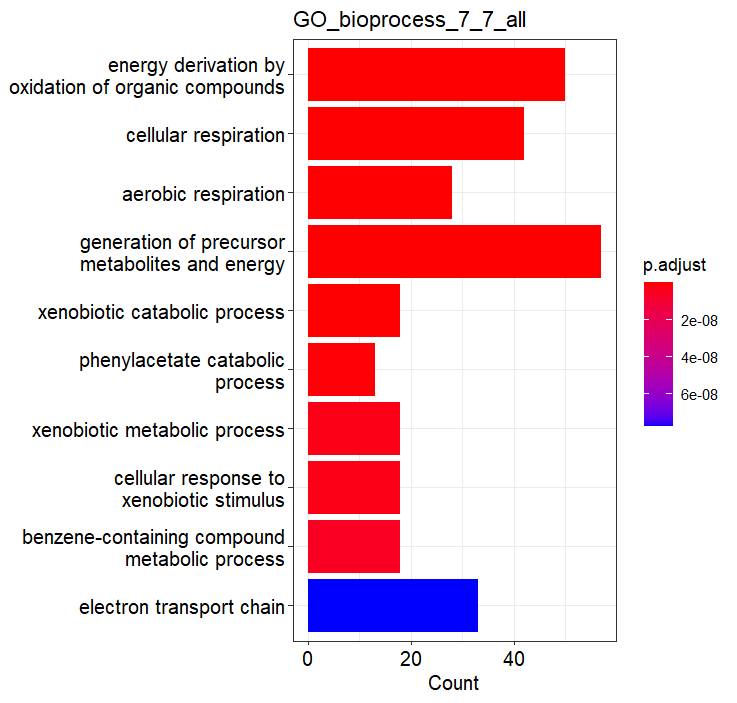
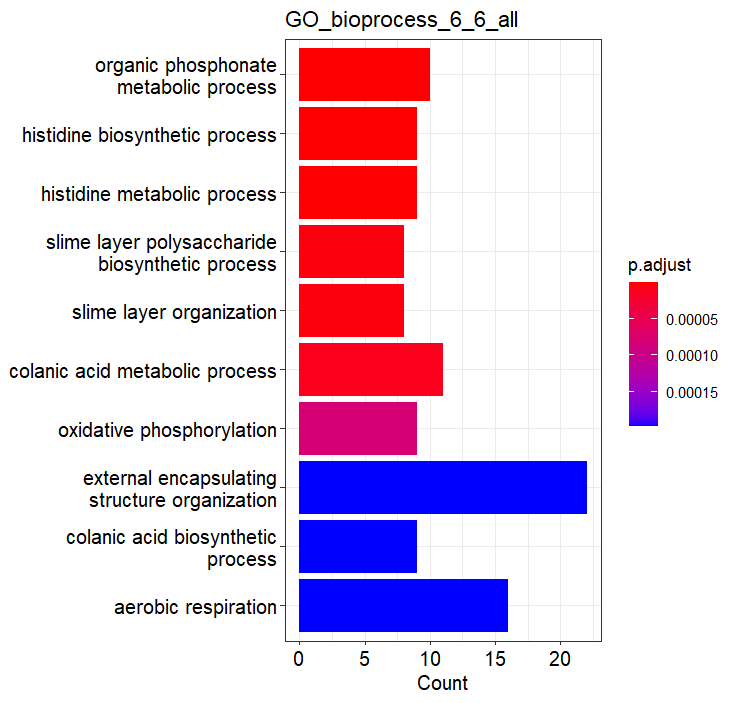
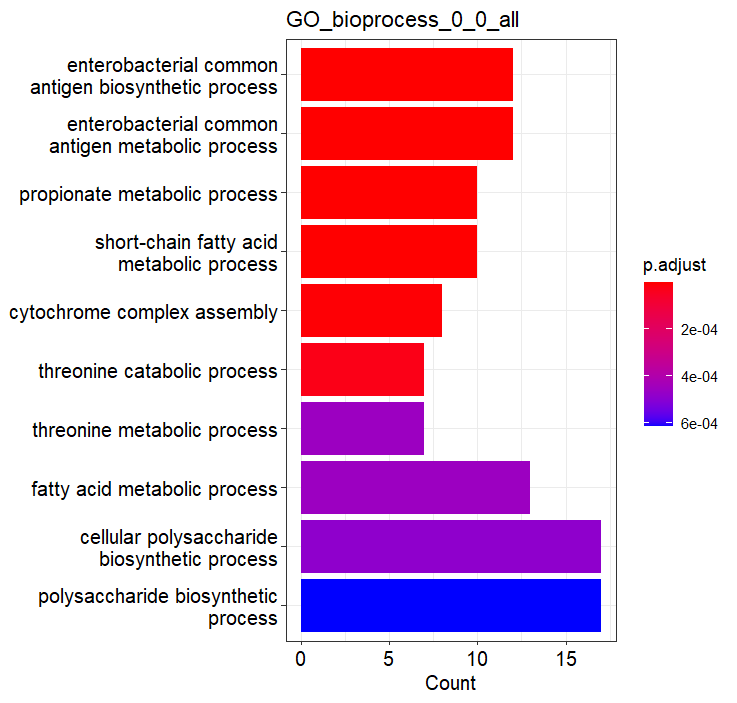
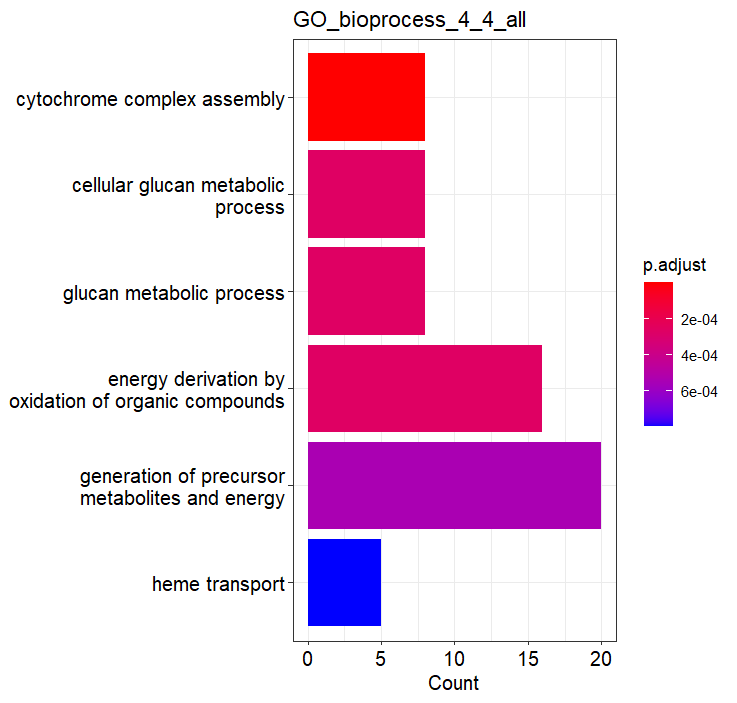


Supplementary figure 1. Figure 32 additional datasets. Gene ontology Biological Process terms enriched in three of the datasets for the set metabolic genes. For sake of commodity, there are a maximum of 10 terms represented in each sample, the 10 found with the lowest qvalue. However, there are more terms in the cases of 0-0 (21), SRR1534633 (35) and Bseq\_2 (17).

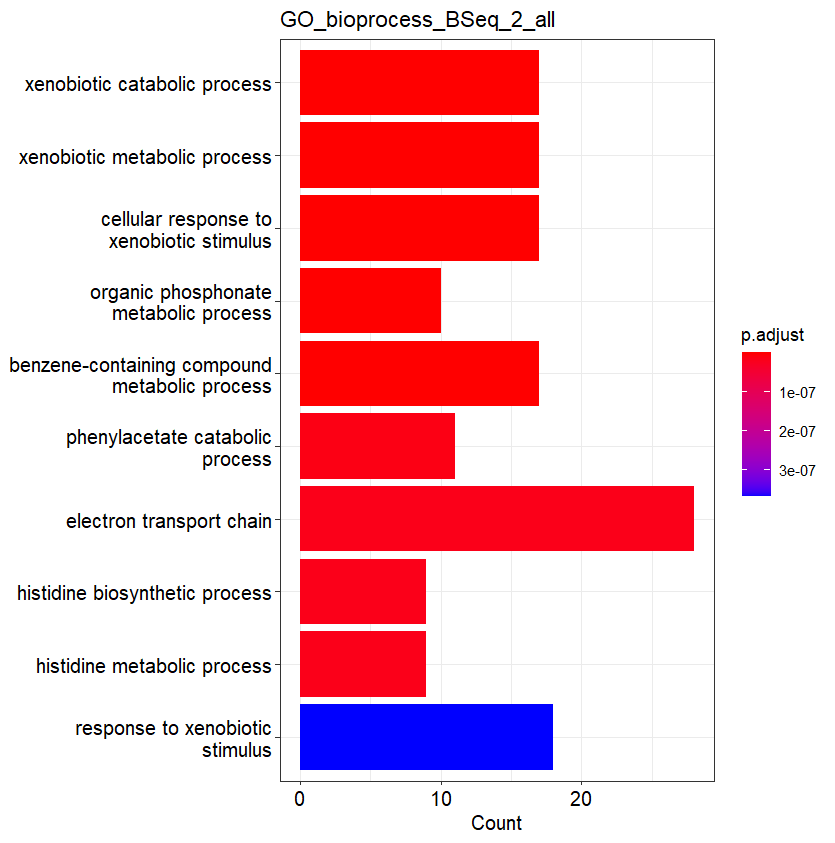
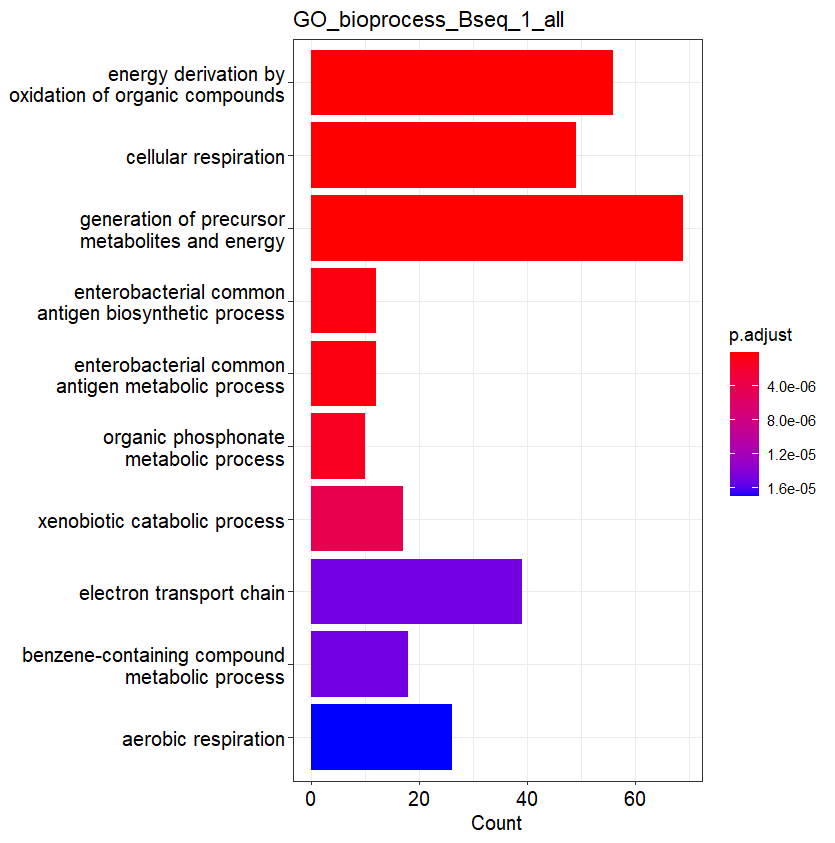
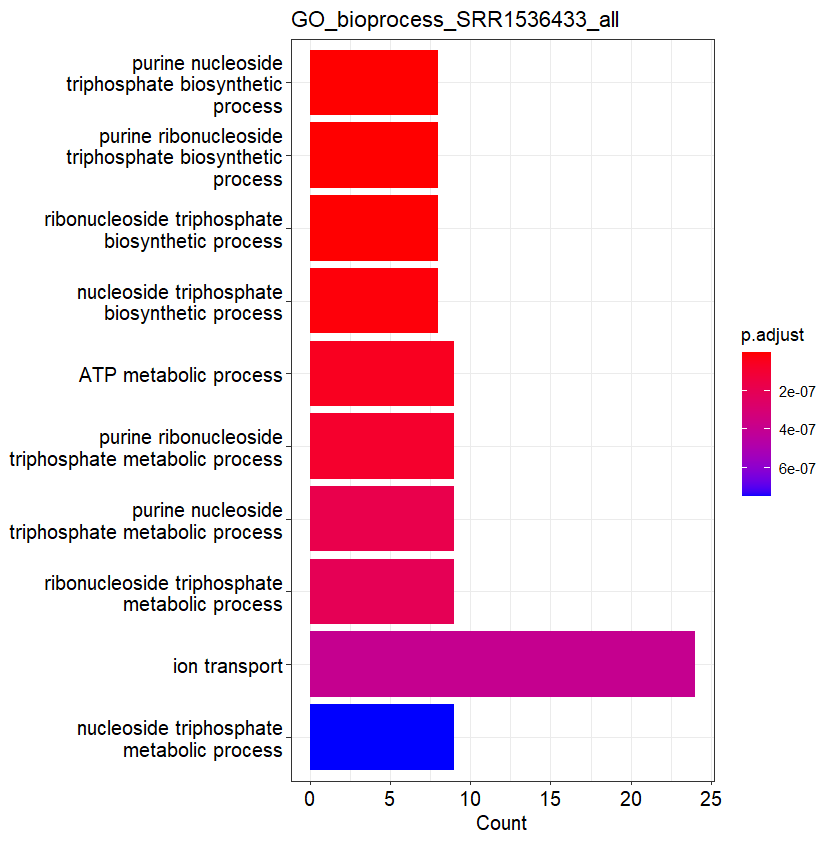
|  |  |
| --- | --- |
| Name- In all GO | Count |
| cytochrome complex assembly | 6 |
| energy derivation by oxidation of organic compounds | 4 |
| generation of precursor metabolites and energy | 4 |
| aerobic respiration | 3 |
| benzene-containing compound metabolic process | 3 |
| cellular respiration | 3 |
| cellular response to xenobiotic stimulus | 3 |
| electron transport chain | 3 |
| histidine biosynthetic process | 3 |
| histidine metabolic process | 3 |
| organic phosphonate metabolic process | 3 |
| oxidative phosphorylation | 3 |
| phenylacetate catabolic process | 3 |
| response to xenobiotic stimulus | 3 |
| xenobiotic catabolic process | 3 |
| xenobiotic metabolic process | 3 |
| anaerobic respiration | 2 |
| branched-chain amino acid transport | 2 |
| enterobacterial common antigen biosynthetic process | 2 |
| enterobacterial common antigen metabolic process | 2 |
| heme transport | 2 |
| monocarboxylic acid catabolic process | 2 |
| nitrogen cycle metabolic process | 2 |
| nucleoside triphosphate biosynthetic process | 2 |
| purine nucleoside triphosphate biosynthetic process | 2 |
| purine ribonucleoside triphosphate biosynthetic process | 2 |
| respiratory electron transport chain | 2 |
| ribonucleoside triphosphate biosynthetic process | 2 |
| *Supplementary table 1.*  *GO Biological terms enriched in more than one sample for all genes found* | |



Supplementary figure 1. Figure 32 additional datasets. Gene ontology Biological Process terms enriched in three of the datasets for the set metabolic genes. For sake of commodity, there are a maximum of 10 terms represented in each sample, the 10 found with the lowest qvalue. However, there are more terms in the cases of 0-0 (21), SRR1534633 (35) and Bseq\_2 (17).



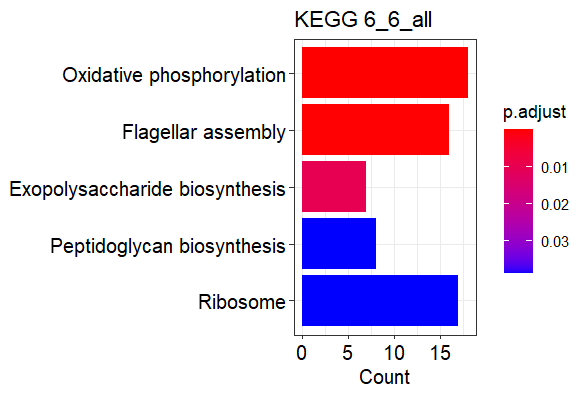
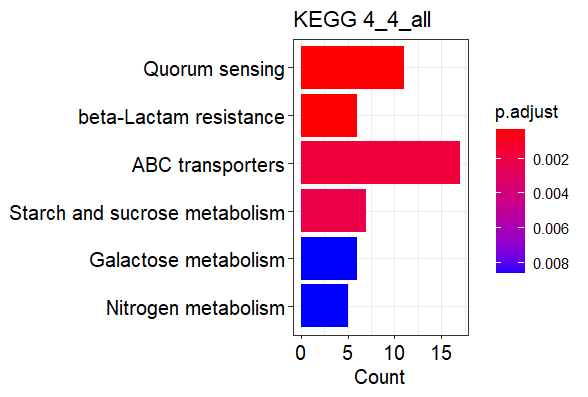
Supplementary figure 2. Gene ontology Biological Process terms enriched in the datasets for the set all genes. For sake of commodity, there are a maximum of 10 terms represented in each sample, the 10 found with the lowest qvalue. However, there are more terms in the cases of 0-0 (24), 6-6(39), 7-7(38), SRR1534633 (32), Bseq\_1(31) and Bseq\_2 (26).



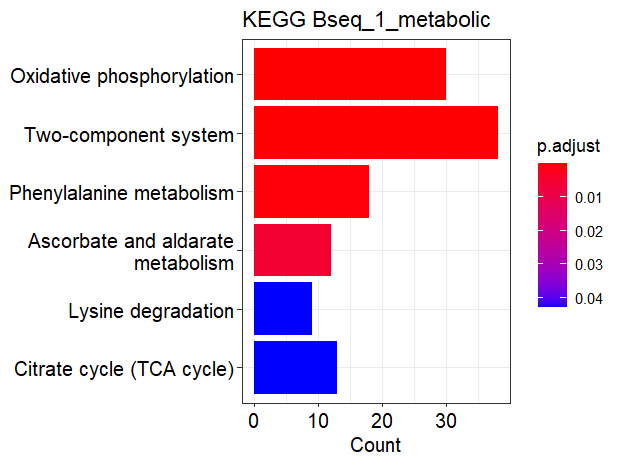
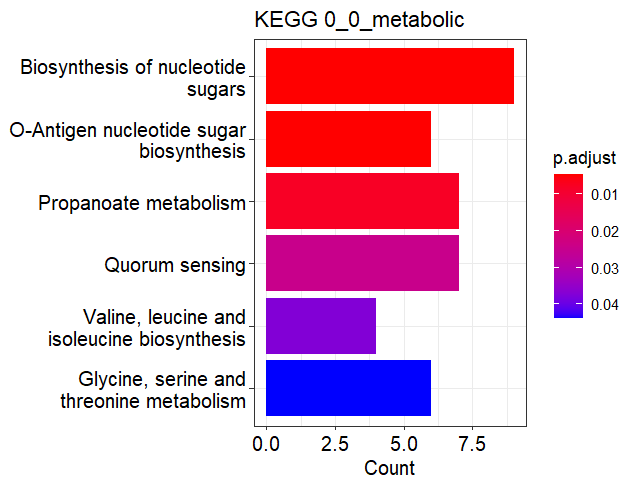
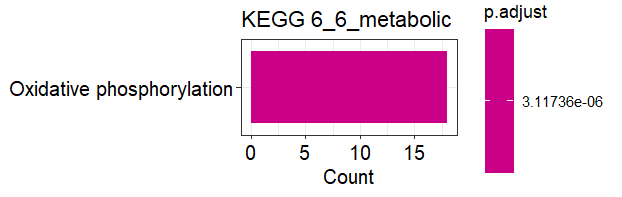
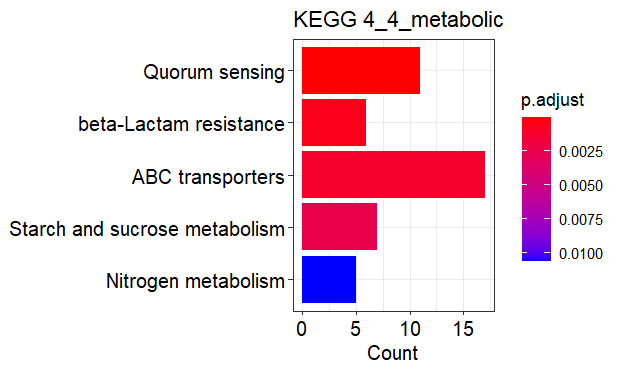
Supplementary figure 2. Gene ontology Biological Process terms enriched in the datasets for the set all genes. For sake of commodity, there are a maximum of 10 terms represented in each sample, the 10 found with the lowest qvalue. However, there are more terms in the cases of 0-0 (24), 6-6(39), 7-7(38), SRR1534633 (32), Bseq\_1(31) and Bseq\_2 (26).

|  |  |
| --- | --- |
| Name pathway: KEGG all | Count |
| Oxidative phosphorylation | 5 |
| Flagellar assembly | 3 |
| Phenylalanine metabolism | 3 |
| Quorum sensing | 3 |
| ABC transporters | 2 |
| Microbial metabolism in diverse environments | 2 |
| Nitrogen metabolism | 2 |
| Propanoate metabolism | 2 |
| *Supplementary table 2.*  *KEGG pathways enriched in more than one sample for all genes found* | |

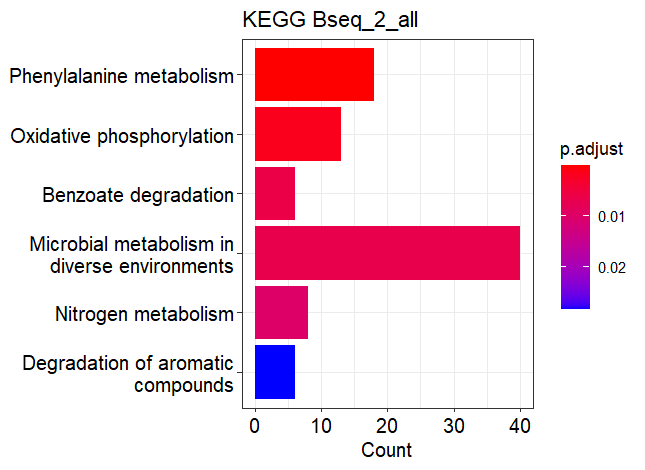
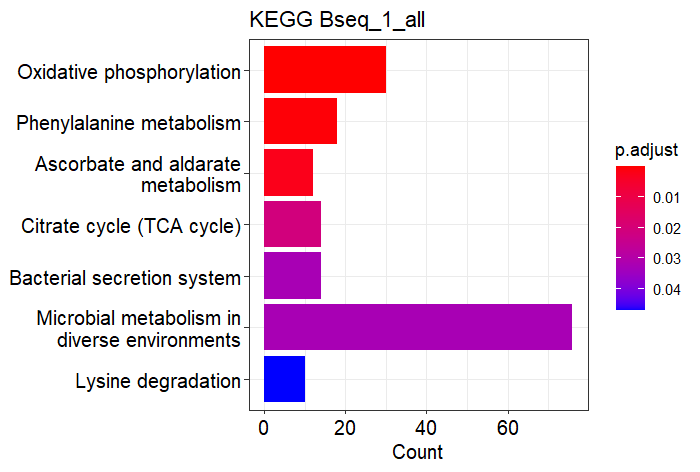
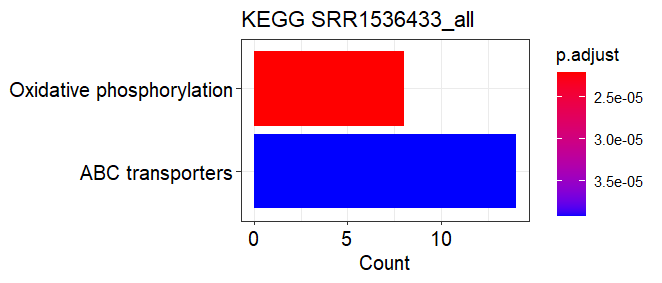
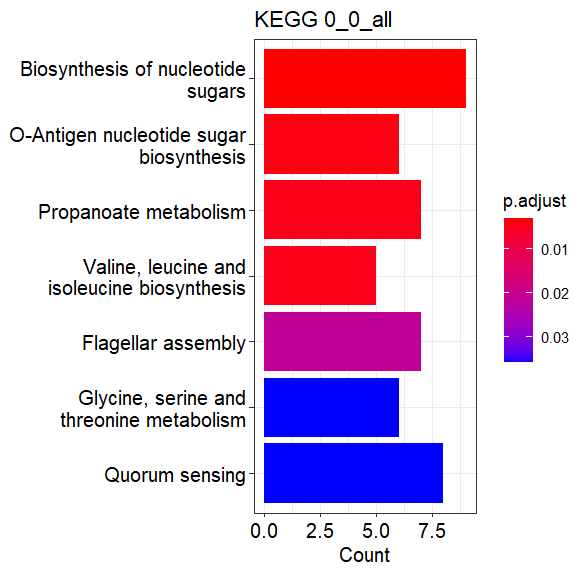
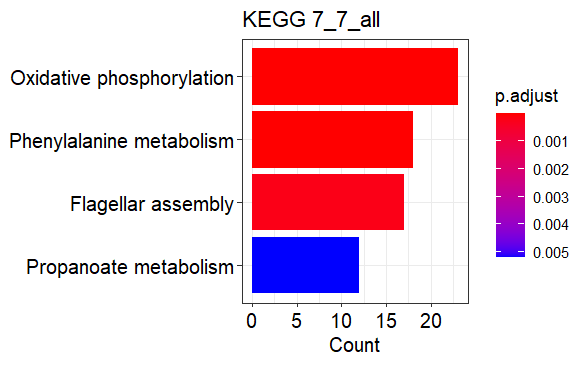
**KEGG**



Supplementary figure 4. KEGG pathways enriched in the datasets for the set all genes.

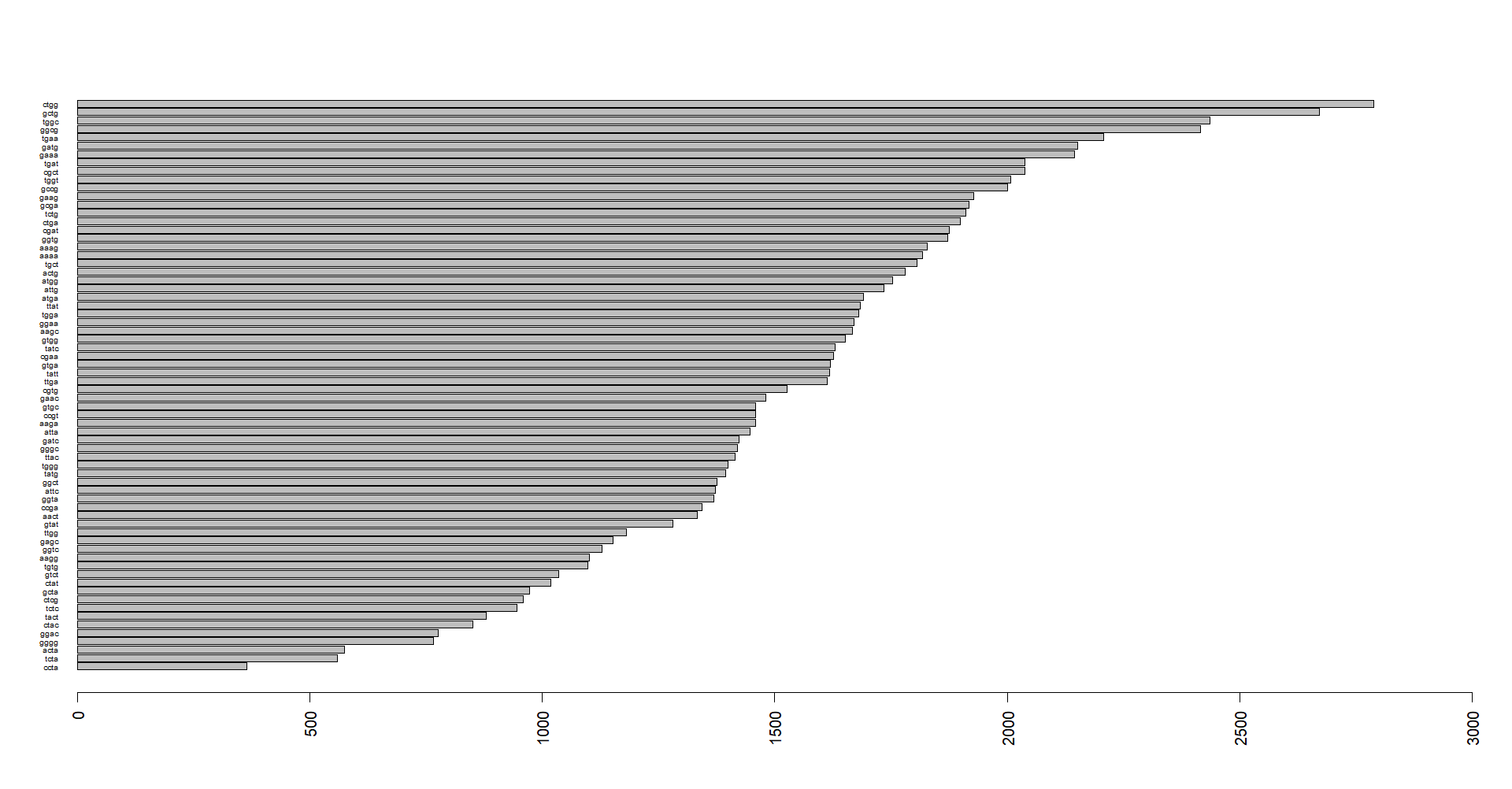


Supplementary figure 3. Figure 33 additional datasets. KEGG pathways enriched in three of the datasets for the set metabolic genes.



Supplementary figure 4. KEGG pathways enriched in the datasets for the set all genes.

1. GATC motif analysis



GATC

Supplementary Figure 5. Most enriched 4-mers in cds and respective total ocurrence in 100bp subsequences of all cds annotated

# **BIBLIOGRAPHY**

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. “Pacific Biosciences Model Organism Genome Sequencing-Ecoli\_K12\_P4C2 (SRR1536433).” *Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015 Jun;33(6):623-30. doi: 10.1038/nbt.3238. Epub 2015 May 25. Erratum in: Nat Biotechnol. 2015 Oct;33(10):1109. PMID: 26006009.*

Casadesús, Josep, and David A. Low. 2013. “Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria.” *Journal of Biological Chemistry* 288(20): 13929–35.

ccano/DevNet. “Https://Github.Com/PacificBiosciences/DevNet/Wiki/8-Plex-Ecoli-Multiplexed-Microbial-Assembly.”

Chapin, Nathaniel, Joseph Fernandez, Jason Poole, and Benjamin Delatte. 2022. “Anchor-Based Bisulfite Sequencing Determines Genome-Wide DNA Methylation.” *Communications Biology* 5(1): 1–6.

Grimbergen, Ard Jan, Jeroen Siebring, Ana Solopova, and Oscar P Kuipers. 2015. “Microbial Bet-Hedging: The Power of Being Different.” *Current Opinion in Microbiology* 25: 67–72.

Monk, Jonathan M. et al. 2017. “IML1515, a Knowledgebase That Computes Escherichia Coli Traits.” *Nature biotechnology* 35(10): 904–8. https://pubmed.ncbi.nlm.nih.gov/29020004/ (January 12, 2023).

Sun Yat-sen University. 2016. “MethSMRT-BROWSE.” *Laboratory of Bioinformatics and Gene Regulation*. http://sysbio.gzzoc.com/methsmrt/browse.html?Prokaryota (February 27, 2023).

Adhikari, Satish, and Patrick D Curtis. 2016. “DNA Methyltransferases and Epigenetic Regulation in Bacteria.” *FEMS Microbiology Reviews* 023: 575–91. https://academic.oup.com/femsre/article/40/5/575/2198251.

Grimbergen, Ard Jan, Jeroen Siebring, Ana Solopova, and Oscar P Kuipers. 2015. “Microbial Bet-Hedging: The Power of Being Different.” *Current Opinion in Microbiology* 25: 67–72.

Kahramanoglou, Christina et al. 2012. “Genomics of DnA Cytosine Methylation in Escherichia Coli Reveals Its Role in Stationary Phase Transcription.” *Nature Communications*. http://tools.neb.com/~vincze/cutfreq/CCWGG.html.

Roberts, Dr. Richard J., and and Dana Macelis. “REBASE.” http://rebase.neb.com/rebase/rebase.html (April 17, 2023).

Gao, Qun et al. 2023. “Bacterial DNA Methyltransferase: A Key to the Epigenetic World with Lessons Learned from Proteobacteria.”

Tong, Madeline et al. 2020. “Gene Dispensability in Escherichia Coli Grown in Thirty Different Carbon Environments.” https://doi.org/10.1128/mBio.02259.

Dukovski, Ilija et al. 2021. “A Metabolic Modeling Platform for the Computation of Microbial Ecosystems in Time and Space (COMETS).” *Nature Protocols* 16(11): 5030–82.

Roberts, Dr. Richard J., and and Dana Macelis. “REBASE.” http://rebase.neb.com/rebase/rebase.html (April 17, 2023).

Adhikari, Satish, and Patrick D Curtis. 2016. “DNA Methyltransferases and Epigenetic Regulation in Bacteria.” *FEMS Microbiology Reviews* 023: 575–91. https://academic.oup.com/femsre/article/40/5/575/2198251.

Grimbergen, Ard Jan, Jeroen Siebring, Ana Solopova, and Oscar P Kuipers. 2015. “Microbial Bet-Hedging: The Power of Being Different.” *Current Opinion in Microbiology* 25: 67–72.

Gu, Changdai et al. 2019. “Current Status and Applications of Genome-Scale Metabolic Models.” *Genome Biology 2019 20:1* 20(1): 1–18. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1730-3 (December 28, 2022).

Kim, Hyun Uk et al. 2008. “Metabolic Flux Analysis and Metabolic Engineering of Microorganisms.” *This journal is ß The Royal Society of Chemistry* 4. www.rsc.org/molecularbiosystems.

Kotte, Oliver, Benjamin Volkmer, Jakub L Radzikowski, and Matthias Heinemann. 2014. “Phenotypic Bistability in *Escherichia Coli* ’s Central Carbon Metabolism.” *Molecular Systems Biology* 10(7).

Oliveira, Pedro H., and Gang Fang. 2021. “Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria.” *Trends in Microbiology* 29(1): 28–40.

Orth, Jeffrey D., Ines Thiele, and Bernhard O. Palsson. 2010. “What Is Flux Balance Analysis?” *Nature biotechnology* 28(3): 245. /pmc/articles/PMC3108565/ (June 2, 2023).

Payelleville, Amaury, and Julien Brillard. 2021. “Novel Identification of Bacterial Epigenetic Regulations Would Benefit From a Better Exploitation of Methylomic Data.” *Frontiers in Microbiology* 12.

Robert, Lydia et al. 2010. “Pre-Dispositions and Epigenetic Inheritance in the Escherichia Coli Lactose Operon Bistable Switch.” *Molecular Systems Biology* 6: 357. /pmc/articles/PMC2872608/ (May 31, 2023).

Sánchez-Romero, María A., and Josep Casadesús. 2020. “The Bacterial Epigenome.” *Nature Reviews Microbiology* 18(1): 7–20.

Simons, Andrew M. 2011. “Modes of Response to Environmental Change and the Elusive Empirical Evidence for Bet Hedging.” *Proceedings of the Royal Society B: Biological Sciences* 278(1712): 1601–9.

Solopova, Ana et al. 2014. “Bet-Hedging during Bacterial Diauxic Shift.” *Proceedings of the National Academy of Sciences of the United States of America* 111(20): 7427–32. www.pnas.org/cgi/doi/10.1073/pnas.1320063111 (May 30, 2023).

Kotte, Oliver, Benjamin Volkmer, Jakub L Radzikowski, and Matthias Heinemann. 2014. “Phenotypic Bistability in *Escherichia Coli* ’s Central Carbon Metabolism.” *Molecular Systems Biology* 10(7).Casadesús, Josep, and David Low. 2006. “Epigenetic Gene Regulation in the Bacterial World.” *Microbiology and Molecular Biology Reviews* 70(3): 830–56.

Casadesús, Josep, and David A. Low. 2013. “Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria.” *Journal of Biological Chemistry* 288(20): 13929–35.

Dukovski, Ilija et al. 2021. “A Metabolic Modeling Platform for the Computation of Microbial Ecosystems in Time and Space (COMETS).” *Nature Protocols* 16(11): 5030–82.

Ines, Thiele, and Ø.Palsson Bernhard. 2010. “A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction.” *Nat Protoc.* 5(1): 93–121.

Kenefake, Dustin, Erick Armingol, Nathan E. Lewis, and Efstratios N. Pistikopoulos. 2022. “An Improved Algorithm for Flux Variability Analysis.” *BMC Bioinformatics* 23(1): 1–14. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-05089-9 (June 5, 2023).

Kim, Hyun Uk et al. 2008. “Metabolic Flux Analysis and Metabolic Engineering of Microorganisms.” *This journal is ß The Royal Society of Chemistry* 4. www.rsc.org/molecularbiosystems.

Mahadevan, R., and C. H. Schilling. 2003. “The Effects of Alternate Optimal Solutions in Constraint-Based Genome-Scale Metabolic Models.” *Metabolic Engineering* 5(4): 264–76. https://pubmed.ncbi.nlm.nih.gov/14642354/ (June 5, 2023).

Orth, Jeffrey D., Ines Thiele, and Bernhard O. Palsson. 2010. “What Is Flux Balance Analysis?” *Nature biotechnology* 28(3): 245. /pmc/articles/PMC3108565/ (June 2, 2023).

Biosciences, Pacific. *Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing*. www.pacb.com/basemods.

Chapin, Nathaniel, Joseph Fernandez, Jason Poole, and Benjamin Delatte. 2022. “Anchor-Based Bisulfite Sequencing Determines Genome-Wide DNA Methylation.” *Communications Biology* 5(1): 1–6.

“E.Coli Genome Project.” 2023. *Strain Information E. coli K-12 MG1655*. https://www.genome.wisc.edu/resources/strains.htm (June 8, 2023).

Krueger, Felix, and Simon R. Andrews. 2011. “Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications.” *Bioinformatics* 27(11): 1571. /pmc/articles/PMC3102221/ (June 8, 2023).

“MeDIP-Seq.” 2023. https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/medip-seq-dip-seq.html (June 8, 2023).

Oliveira, Pedro H., and Gang Fang. 2021. “Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria.” *Trends in Microbiology* 29(1): 28–40.

Payelleville, Amaury, and Julien Brillard. 2021. “Novel Identification of Bacterial Epigenetic Regulations Would Benefit From a Better Exploitation of Methylomic Data.” *Frontiers in Microbiology* 12.

Xu, Liu, and Masahide Seki. 2020. “Recent Advances in the Detection of Base Modifications Using the Nanopore Sequencer.” *Journal of Human Genetics* 65(1): 25–33.

Bailey, Timothy L. 2021. “STREME: Accurate and Versatile Sequence Motif Discovery.” *Bioinformatics*. http://meme-suite.org. (June 19, 2023).

“Documentation for COBRApy — Cobra 0.26.2 Documentation.” https://cobrapy.readthedocs.io/en/0.26.2/index.html (June 19, 2023).

Ebrahim, Ali, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. 2013. “COBRApy: COnstraints-Based Reconstruction and Analysis for Python.” *BMC Systems Biology* 7(1): 74.

van Helden, J., B. André, and J. Collado-Vides. 2000. “A Web Site for the Computational Analysis of Yeast Regulatory Sequences.” *Yeast* 16(2): 177–87.

“EcoCyc: Encyclopedia of E. Coli Genes and Metabolism.” https://ecocyc.org/ (June 13, 2023).

Stevens, Julie M et al. 2011. “Cytochrome c Biogenesis System I.” *The FEBS journal*. https://febs.onlinelibrary.wiley.com/doi/10.1111/j.1742-4658.2011.08376.x.