The Pennsylvania State University

The Graduate School

Eberly College of Sciences

MICROBIOME AND EPIGENETICS: TWO KEYS TO THE DEVELOPMENT OF PERSONALIZED MEDICINE

A Dissertation in

Integrative Biosciences

by

Tyler S. Malys

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2015

The dissertation of Tyler Malys was reviewed and approved* by the following

Ross C. Hardison

T. Ming Chu Professor of Biochemistry and Molecular Biology

Peter Hudson

Willaman Professor of Biology, Director of Huck Institutes of Life Sciences

Head of the Graduate Program

Michael M. Mwangi

Assistant Professor of Biochemistry and Molecular Biology

Matam Vijay-Kumar

Assistant Professor, Department of Nutritional Sciences.

*Signatures are on file in the Graduate School

Abstract

This dissertation has two parts: first describes the analysis of microbiomes of the lower respiratory tract of healthy and diseased chickens. The second describes the epigenetic landscape that determines the gene transcription levels in the erythroid differentiation model. Multicellular animals are complex machines and can no longer be thought of as stand alone organisms. Animals host microbial communities termed microbiomes, which play vital roles in many aspects of an animals life including immune function and nutrient processing. Animal physiology and habits have been shown to affect resident microbiome composition, which in turn affect animal physiology. Understanding host microbe interactions and their ramifications is thus essential in my pursuit to understand human health. Further complicating matters, each microbiome interacts with only a subset of host cell types. Nearly every host cell contains identical DNA yet express radically different phenotypes. These differences in phenotype are brought about by epigenetic effect. Epigenetic factors represent heritable changes, which are not apparent in DNA sequence alone. They include items such as histone modifications, transcription factors and DNAse hypersensitive sites. Epigenetics vary amongst individuals but also amongst cell types within individuals. Understanding epigenetic regulation of gene transcription is essential to understanding human health and has gained much interest in the development of personalized medicine. I conducted fundamental research aimed at advancing both the understanding of resident microbiome and also epigenetic gene regulation. Microbiome studies were conducted in Chickens from Pakistan with the primary goal of identifying potentially novel emerging infectious diseases. I also studied epigenetic gene regulation in an *in vitro* model system for erythroid maturation. I found that the lower respiratory microbiome of healthy chickens was dominated by relatively few

bacterial taxa. Moreover the microbiome of healthy chickens varied with respect to the environment. I was also able to identify differences in microbiome composition between healthy and diseased birds, suggesting the discovery of potential novel pathogens representing emerging infectious disease in the region. In my model system for erythroid maturation I found epigenetic landscape near the transcription start site (TSS) to be sufficient to predict a gene as actively transcribed or silent. Moreover, I was able to relate change in epigenetic landscape with change in gene transcription level in cases of extreme transcription level change. Overall, these results indicated epigenetic landscape near the TSS sets a gene as either permissive or non permissive with respect to transcription.

Table of Contents

List of Tables

List of Figures

Microbiome Chapter

Epigenetics Chapter

Acknowledgements

I am very grateful for the numerous helpful suggestions, patience, discussions and guidance offered by Dr. Eric Harvill, Dr. Ross Hardison, Dr. Michael Mwangi, Dr. Matam Kumar, Dr. Bodo Linz, Dr. Yury Ivanov, Dr. Muhammad Shabbir, Dr. Yu Zhang, Dr. Chris Morrissey, and Juan Antonio Raygoza Garay. Finally, I sincerely thank Dr. Cooduvalli Shashikant for his guidance and encouragement throughout my graduate school experience.

Chapter 1: The microbiome of the lower respiratory tract of healthy and diseased chickens from Pakistan

FIELD REVIEW

A NOTE ON MICROBIOME:

Microbes are ubiquitously present in the environment and perform a variety of important ecological functions. They intimately interact with multi-cellular host organisms. Hosts provide a variety of ecological niches and microenvironments in which microbes form stable communities. These microbial communities are thus referred to as host microbes.

Microorganisms typically residing in these communities range from bacteria, viruses up to eukaryotes including protista, fungi and even mites [1–9]. However, bacteria are currently the most well studied. Under normal healthy interaction conditions, microbial load is kept in check by the host immune system, which regulates number of bacteria and also discriminates pathogenic and commensal microbes [10–19]. Nutrient availability is also split between the host and microbes[20]. Microbes consume many nutrients that are valuable to the host. In cases of excess bacterial load this can lead to malnutrition in the host [21]. Conversely, microbes break down many substrates, which cannot be processed by the host alone into usable nutrients. Thus host is able to consume and benefit from a wider variety of energy sources otherwise inaccessible without functional microbiomes [20,22–30].

Microbes constantly interact with each other within their respective communities. They compete for nutrients and actively conduct microbial warfare [31]. This type of interaction can benefit the host by providing highly inhospitable environment to potential invading pathogens [32,33]. In the absence of normal micro flora a host is more susceptible to infection [34,35].

Microbial composition tends to be relatively stable under normal conditions but varies amongst individuals with respect to host site [36–43]. There is a significant functional redundancy despite varied composition [43,44]. This occurs due to both selective pressures created by the microenvironment and the order in which microbes colonize a microenvironment, which is not at carrying capacity [45–51]. While composition is generally stable, significant persistent perturbation can permanently destabilize the community, leading to negative health consequences for the host [52–55]. In some cases, invading microbes can be sufficient to completely change microbiome composition [56].

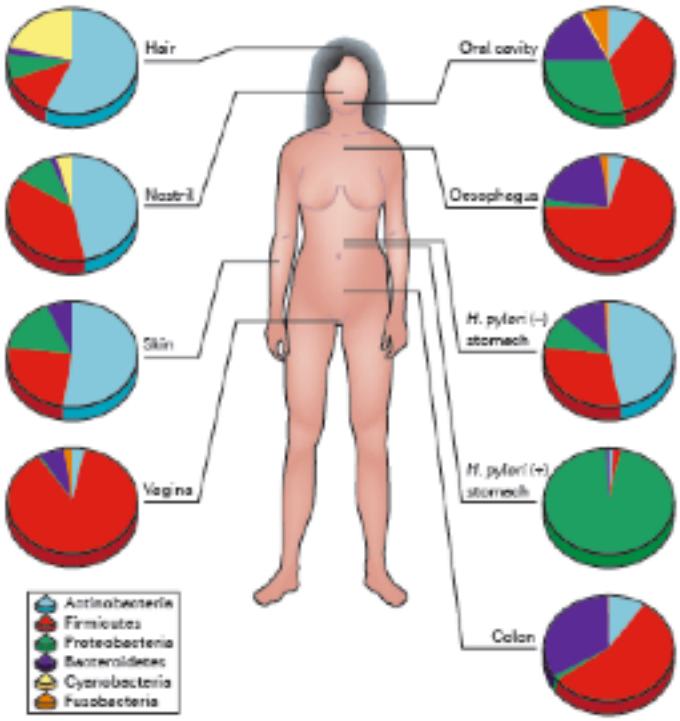


Figure 1. Compositional differences in the microbiome by anatomical site. Adapted from Cho and Blazer 2012

In invertebrates and insects the relationship between host and microbiomes is well understood, as there exists clear symbiotic dependencies [57,58]. Vertebrates are more complicated as functional redundancy exists between the host and its respective microbiomes [59,60].

Microbiomes are relatively well studied in humans compared to other species. Human microbes are represented by six major phyla including; Actinobacteria, Firmicutes, Proteobacteria, Bacteroidetes, Cyanobacteria and Fusobacteria [38,50,61,62]. These microbes are generally present in

different proportions depending on sampling site [34]. There is substantial variation from person to person and even temporal variation within a sampling site on any given individual [20,54,61,63–72]. Composition can be dramatically shifted by the presence or absence of a key species [53,56].

Gut microbiota, by far the most studied human microbiome, has been linked to both immune function and nutrient processing [32]. However, interest in other microbiomes is also currently peaking. For example, skin provides a barrier between most human tissues and the external environment, but it also hosts an extremely diverse set of microbiomes [40]. Even the respiratory tract has come under scrutiny with regard to specific diseases [73].

Gut microbiome has generally been sampled from three distinct sections; the stomach, the small intestines and the large intestines, each with distinct characteristics [35]. The stomach is highly acidic and has a fast flow rate of nutrients through it. As a result, stomach microbiome is relatively sparse containing only 10^3 to 10^5 cells per mil of affluent [74]. This microbiome contains both resident microbes present in the ecological niche and others which are brought in with food but do not remain a viable part of the community [75–78]. *Helicobacter pylori* is a dominant microbe in the stomach. When absent, stomach microbiome tends to contain a more diverse array of bacteria [41,75].

In the small intestine, microbial load increases with roughly 10^8 cells per mil of affluent [35]. The environment here is more hospitable than in the stomach but, high nutrient flow rate and immune system components keep bacterial load in check [79]. This reduces competition for nutrients, majority of which are absorbed by the host in the small intestine. Bacterial biomass significantly higher than standard, leads to malnutrition in the host [21]. Microbiome in the small intestine is generally dominated by Enterobacteriaceae, which through various metabolic processes can nutritionally benefit the host.

For example, Etnerobacteriaceae can fulfill host amino acid requirements if they are not directly provided by host diet [27,29,30,77,80,81].

In the large intestine microbial load increases to 10^{11} cells per gram [35]. This is due to a more neutral pH, slower nutrient flow rate and a lessened immune response. Primarily present are Firmicutes belonging to clusters ix and xiv, Bacteriodes, Actinobacter and Verrrucomicrobia [42,43]. Also present to a lesser extent are Proteobacteria and Fusobacteria [38]. These microbes can contribute to host nutritional requirements by fermenting otherwise unusable substrates into short chain fatty acids [35]. In most cases the microbiome here represents a high diversity of strains belonging to relatively few evolutionary lineages [49,82]. Variation is observed along the length of the long intestine. However, stool sample microbiomes are more stable [9,20,42,83–86].

Digestive tract microbiome provides many benefits in terms of nutrient processing allowing us to better utilize modern foods [20,35,84,87–89]. For example, cooking introduces many chemical changes in food, some of which can be toxic to host organisms [20,22,90,91]. Since these toxins are recent, humans are not well equipped to process them. Luckily bacterial communities can degrade many of them, sparing our physiology unnecessary stress [92]. Milk and starch rich foods fall in a similar category and microbial processing again increases their nutritional value for humans [20,87]. In an extreme example, in populations of Japanese decent, gut microbiome is capable of processing components of seaweed, which would otherwise be nutritionally inaccessible [93].

Functional redundancy exists with respect to nutrient processing between host and microbiome in vertebrates [20,87]. For example, humans express genes that are capable of processing both lactose and starches, while these genes are also expressed by gut microbiome [20,87]. The host has unrestricted access to these macronutrients. Additionally, microbiomes may assist in the utilization of excess nutrients, which would be otherwise lost due to limited processing ability of the host [35].

Consistent with previous observations, gut microbiome composition is influenced by host physiology, which is heavily influenced by diet, as different nutrient combinations preferentially select for different nutrient processing and possibly bacterial community compositions [20,45,46,71,72,84,88,93].

Skin microbiome is also an increasingly studied field. Skin is the humans largest organ [40], which contains many immune system components (T-cells). Skin acts as a barrier, which separates many of our internal tissues from the external environment via dead keratin filled cells called Squams. Skin also allows the immune system a method of interacting with the external environment [40]. Skin is not completely smooth, it contains many glands and hair floccules. Exocrine glands excrete water and electrolytes that acidify the skin. Aprocrin glands are present at the nipple, armpit and genitals and have long been postulated to excrete pheromones. Sebaceous glands are generally anoxic [40,94–103]. Adding to skin variability is a plethora of bends, contours and folds, which occur in varied locations on the human body, some of which are generally exposed to ultraviolet light and the general environment to varied degrees [94,104,105]. This high degree of variation leads the skin to host very diverse and varied microbial communities [85,106–109].

Before birth, human skin is considered to be sterile [110,111]. Upon rupture of the amniotic sac the infant is exposed to a mothers microbes for the first time [112,113]. This exposure is generally to vaginal microbiome, but can be to skin microbiome if a C-section is performed [112,113]. Initialization is important as many microbes can occupy functional niches in a variety of microbial communities on and in a human and composition can be effected by order of exposure [63,63,114,115]. As a person ages, skin microbiome will again change with the onset of puberty and has been shown to differ between genders [106,110,111,114,114,116–118].

In general, skin is cool, acidic and desiccated [94]. Like gut, it contains; actinobacter, firmicutes, bacteriodes and proteobacteria, however in different proportions [40]. For example actinobacter is more

abundant on skin as opposed to in the gut [40]. *Propionibacterium acnes* is also commonly present on skin and actively contributes to skin acidity, which as in gut adds selective pressure effecting the environmental niche in which microbes reside [8,113]. Additionally, skin acidity also protects against pathogens such as *Staphylococcus aureus* [119].

Sebaceous sites such as the back are under strong selective pressure. They generally contain low numbers of bacteria and are dominated by propionibacterium spp and some flavobacteriales (phylum bacteriodes) [38,40,109]. Moist areas tend to contain Staphylococcus and Cornebacterium species [85,109] while dry skin often contains a variety of actinobacter, proteobacteria, firmicutes and bacteriodes [85,107,109]. Additionally, gram negative bacteria have been found on dry skin and were previously thought to only be harbored in the gut [2,8]. Microbes other than bacteria are also present to varying extents. For example, fungal Malassazia spp. have been identified from dermal samples and Domex mites are normally present in hair follicles [120–125].

Recently lower respiratory microbiome has come under investigation with interest in its relationship to various respiratory diseases [126,127]. Originally, thought to be sterile, investigators later found bacteria to be present via next generation sequencing technologies [73,128,129]. There is still debate as to whether a distinct lower respiratory microbiome exists or it is simply a mirrored community of the upper respiratory tract [127]. Researchers have been able to show differences in microbiome composition based on; environment [130], pet ownership [131] and sampling location [132] with in the lower respiratory tract and between healthy and diseased individuals [126,127,132–137].

Microbiomes of the gut, skin and lung have all been implicated in or associated with a variety of diseases [34,35,40,127,138]. Gut microbiome has been related to; Crohn's disease [139,140], Ulcerative Colitis [141], Colrectal Cancer [44,142,143] and *H. pylori* infection [41,144–146]. Exposure to antibiotics in early childhood has been related to Crohn's disease and is accompanied by decreased microbial diversity as well as a bloom

of *Enterococcus facium* and Proteobacteria [139,140,147]. Ulcerative Colitis is also associated with decreased microbial diversity and an increase in Actinobacteria and Proteobacteria [141,148]. Additionally Enterobacteriaceae has been associated with an increased risk for the disease [141]. Colorectal cancer is associated with the presence of the anaerobic genus fusobacterium and the reduced colonization by Bacteriodes and Firmicutes [44,142,143]. *H. pylori* negative individuals have a diverse stomach microbiome which includes Streptococcus, Actinomyces, Prevotella and Gemella [41]. *H. pylori* positive individuals stomach microbiome is associated with >90% *H. pylori* composition and the development of ulcers, tumors and gastric adenocarcinomas [144,149]. Interestingly, *H. pylori* infection is also associated with decreased risk of childhood onset asthma and reflux oesophagitis [145,146].

Gut microbiome composition has also been associated with non-gut diseases [34] ; obesity [139,150–152], liver issues [153–155], and asthma [156]. Gut microbiome in obese patients show a decreased bacteriodetes/firmicutes ratio [151]. Transplantation of obese mouse gut microbiome to germ free mice recapitulated the obese phenotype [151]. Studies have associated weight loss with increased colonization by Bacterioidetes and conversely, weight gain has been associated with decreased colonization. Bacterioidetes populations [139] have been associated with obesity in a monozyogotic twin study [152]. Interestingly, as with Crohn's disease, antibiotic use during early childhood has been associated with obesity [157]. However, perinatal administration of *Lactobacillus rhammosus* (a pro biotic) has been shown to decrease excessive weight gain during childhood [158].

The liver first encounters metabolites produced by microbes in the gut before they enter the rest of the body [153]. A variety of non-alcoholic fatty liver diseases have been associated with gut microbiome composition [153]. The presence of *Helicobacter hepaticus* promotes hepatocellular carcinoma [154]. Cirrhosis patients contain a substantially altered gut microbiome, enriched for the phyla: proteobacteria and fusobacteria with enrichment for families: enterobacteriaceae, veillonellaceae and

streptococcaceae [155]. Asthma and allergic airway disease is also associated with children having antibiotic exposure, and has been shown to be increasing in the westernized world but not worldwide[159].

Psoriasis [160,161], seborrheic dermatitis [162], chronic wounds [163,164], acne [165] and diabetic skin ulcers have all been associated with altered skin microbiome. Psoriasis results from a dis-regulation of the host immune response [160,161]. Generally Firmicutes are over represented and Actinobacter are underrepresented compared to controls in psoriasis patients [166]. Additionally, *S. aureus* is present in 90% of cases [167,168]. Fungal *Malassezi spp* are associated with seborrheic dermatitis [162]. Chronic wounds tend to harbor a diverse microbiome whereas burn wounds have limited and predictable colonization [163,164]. Acne is caused by *P. acnes*[165,169] and diabetic skin ulcers are associated with Streptococcaceae [170]. Moreover, normal dermal microflora can be pathogenic elsewhere in the body [171].

While a healthy respiratory microbiome has yet to be established, microbiome characteristics have been associated with various respiratory diseases: cystic fibrosis[133,134,172,173], chronic lung disease in infants[135] and brocho-pulmonary dysplasia [174]. Cystic fibrosis patients exhibit decreased microbial richness, which is exacerbated with age [172,173]. While *Pseudomonas aeruginosa*, *S. aureus* and Burkholderiacepacia have traditionally been associated with cystic fibrosis, NGS has now shown greater than 60 genera to be present [127]. Chronic lung disease in infants is associated with *Ureplasma spp*[135]. Bronchopulmonary dysplasia has been associated with the presence of *S. aureus*, *P. aeruginosa* and *Streptococcus spp*[174].

Lung microbiome composition also differs based on smoking [133,175]. Lung microbiome can be stratified amongst smokers by age and lung function [133]. In the case of smokers experiencing chronic obstructive pulmonary disease in need of mechanical ventilation, lung microbiome exhibits; reduced community diversity (observation also seen in 1/3 of healthy individuals), have *P. aeruginosa* as a

dominant species found in a biofilm on endotracheal tubes, and have greater quantities of *Haemophilus spp* whereas bacteriodetes was more common in healthy controls [175]. Lung transplant patients show increased bacterial concentrations in transplanted tissues compared to controls [109]. However, it is important to note that although many associations between microbiome and disease have been identified, ongoing efforts are necessary to determine causation [176].

Understanding microbiome can also have agricultural and zoonotic implications. 50-80% of Campylobacter infections in humans are attributed to poultry. Birds are typically colonized at 2-3 weeks of age and remain colonized until slaughter [177]. At which time Campylobacter and related *Helicobacter spp* are present at detectable levels in processed chicken meat [178,179]. This issue is associated with an annual cost of 1.7 million dollars in the US alone [180].

The chicken gut microbiome generally contain Firmicutes (lactobacillus, peptostreptococcaceae, ruminococcacea) proteobacteria (erichericheia, shigella and enterobacter) Actinobacter (corynebacterium and brevibacterium) and bacteriodetes (alistipes and bacterioides) [181]. Specific composition is often varied and multiple enterotypeshave been identified [181].

Many attempts have been made to address the issue of Campylobacter infection in chickens, including: increased hygiene protocols in poultry facilities [177,182,183], vaccination and microbe based approaches such as the use of; bacteriophages [184], pro biotics [185] and competitive exclusion with; salmonella [186], *Escherichia coli*[187] and potentially *Campylobacter jejuni* [188–190], are also commonly implemented.

PROJECT ABSTRACT

Poultry production is a globally important agricultural industry. Dense housing conditions and large flock numbers boost production but increase the risk of disease outbreaks. To assess the risk for animal and human health, it is critically important to identify known and potentially novel pathogens as well as to determine the microbiome of healthy birds to understand how pathogens change the composition of the chicken microbiome during disease. In this study, I characterize the microbiomes of the lower respiratory system of healthy and diseased birds during outbreaks of respiratory disease at several farms in the Punjab province of Pakistan by 454 pyrosequencing of 16S rDNA amplicons. I show that the microbiome of the lower respiratory tract of apparently healthy chickens is dominated by a limited number of operational taxonomic units (OTUs), the majority of which belong to the Gammaproteobacteria. Microbial diversity at the family and genus levels reveals that the microbiomes of healthy birds are consistent within most farms, but differ between farms, indicating that the microbiome composition of healthy chickens is largely determined by the environment. I show a predominance of respiratory pathogens in the microbiomes of diseased birds, including known poultry pathogens such as *Mycoplasma synoviae* or *Ornithobacterium rhinotracheale*, consistent with the observed disease symptoms. In addition, I detect OTUs that display less than 93% similarity to any known 16S rRNA sequences, indicating the presence of potentially new bacterial species in the respiratory tract of chickens.

INTRODUCTION

Poultry is an important, low cost source of protein worldwide, but also a potential source of dangerous human pathogens[237]. Chickens are commonly raised under different management systems. Free-range farms rely on sparse flocks that are free roaming and presumably have stronger natural immunity to common pathogens they encounter. Controlled-house farms maintain massive flocks and regulate many aspects of the bird's lives, including feed, vaccination, and airflow. Open-house farms represent an intermediate level of regulation. Unfortunately, maximizing production, for example by increasing density, can select for more aggressive and virulent strains of pathogens, leading to devastating disease outbreaks. Zoonotic spread to humans is frequently deadly, and carries the potential to adapt to and spread among humans, a very significant worldwide threat. Besides monitoring the health status of the birds, it is vitally important to identify known and potentially novel pathogens, which are of critical interest to both the agricultural industry and public health officials to assess risk to animal and human health. In order to recognize potential threats, it is essential to determine the microbiome of healthy chickens and to understand how the composition of the microbiome is disrupted upon infection with pathogens. Traditional pathogen identification techniques rely on the ability to culture suspected pathogens, which is problematic because most bacteria are challenging to culture [238]. Next generation sequencing (NGS) facilitates the sequencing and categorization of both known and potentially novel microorganisms present in a sample, including those that cannot yet be cultured [238–243].

Access to commercial livestock is generally restricted to veterinarians and public health officials. In the farm setting, diseased animals are usually removed to limit spread of disease, while apparently healthy animals are used for production. In this study, I explored the possibility of utilizing next

generation sequencing technology to identify potentially novel pathogens present during outbreaks of disease amongst poultry farms in Pakistan. I collaborated with veterinarians to analyze outbreaks of respiratory disease among chickens in Pakistan, which is generally considered one of the global hot spots for emerging infectious disease [244]. The microbiome of the lower respiratory tract of 21 apparently healthy and 22 diseased chickens from 7 farms, were characterzed by using culture independent 454 pyrosequencing of 16S rDNA amplicons. The microbiome was dominated by Gammaproteobacteria that belong to a limited number of taxa. The microbiome of healthy chickens appears to be consistent within most farms but not between farms. The composition of the microbiome appears to be largely determined by the environment. The microbiomes of diseased and healthy birds differ. The respiratory pathogens such as *Mycoplasma synoviae* and *Ornithobacterium rhinotracheale* were predominant in sick chickens, consistent with the observed disease symptoms. Several Gammaproteobacteria sequences that show less than 93% similarity to any known 16S rRNA sequences in BLAST searches [245] were detected, suggesting that the respiratory tracts of birds may contain novel microbes.

RESULTS

Sample Collection and Processing: The microbiomes of the lower respiratory tract of 43 chickens from 7 farms in the Punjab province of Pakistan were determned. Of three farms housing layer birds, two raised Babcock under open-house conditions (farms 1 and 2) and one raised Misri under free-range conditions (farm 3). Four farms housed Hubbard birds under controlled-house conditions, two raising broilers (farms 4 and 5) and two raising breeders (farms 6 and 7). Both healthy and diseased birds were sampled from each farm. Detailed sample information can be found in Supplementary Table 1.

The chickens were sacrificed to sample Tracheal broncho alveolar lavage (T-bal) from each specimen. The hyper-variable regions V1-V5 of the 16S rRNA gene were PCR amplified from DNA that was extracted from each sample. Roche 454 pyrosequencing of pooled 16S rDNA amplicons yielded an average of 8,154 (range 836 to 20,127) high-quality reads per sample (Supplementary Figure 1). Operational Taxonomic Units (OTUs) were assigned in MOTHUR[246,247] by mapping reads against the SILVA database [248], and OTUs were subsequently grouped into corresponding taxa. Rarefaction curves indicated that >60% of theoretically present OTUs were sequenced from most of the samples (Supplementary Figure 2). The percentage of theoretically present OTUs did not correlate with number of reads (R^2 = 0.04) (Supplementary Figure 3). The number of distinct OTUs (alpha-diversity, Supplementary Figure 4) was not correlated with the number of sequencing reads in healthy chickens (R^2 = 0.07), but there was weak correlation within diseased birds (R^2 = 0.34), although the proportion of OTU per read was not significantly different between healthy and diseased birds (two-tailed t-test p-val 0.28) (Supplementary Figure 5).

The microbiome of the lower respiratory tract of healthy chickens is dominated by a limited number of bacterial taxa. Each chicken's microbiome was dominated by a limited number of bacterial taxa. In the majority of the samples, as few as four OTUs represented over 95% of the total sequencing reads. The alpha-diversity was lowest among chickens from farm 1, containing only 2 to 15 OTUs per sample (median of 12.5 OTUs, Supplementary Figure 4). Additionally, very few OTUs were also found among chickens from farm 4 (median of 28.5 OTUs per sample, range 9 - 41) and farm 5 (median of 22 OTUs, range 10 to 71 OTU), while chickens from other farms contained a median number of 116.5 to 167 OTUs (28 to 407 OTUs per sample).

The microbiomes of healthy chickens consisted mostly of Proteobacteria, which were found in 83.9% of the sequencing reads. Proteobacteria were predominant at six out of the seven farms (farms 1-4, 6, 7), comprising over 94% of the sequencing reads in the majority of the samples. In very few specimens from farms 3, 6 and 7, Proteobacteria comprised only 60-80% of the sequencing reads, with cumulative contribution of Bacteroidetes (2.7% of all sequencing reads), Firmicutes (1.8% of all sequencing reads) and Actinobacteria (1.6% of all sequencing reads) to >96% of the total reads per sample (Fig. 1, Fig. 2).

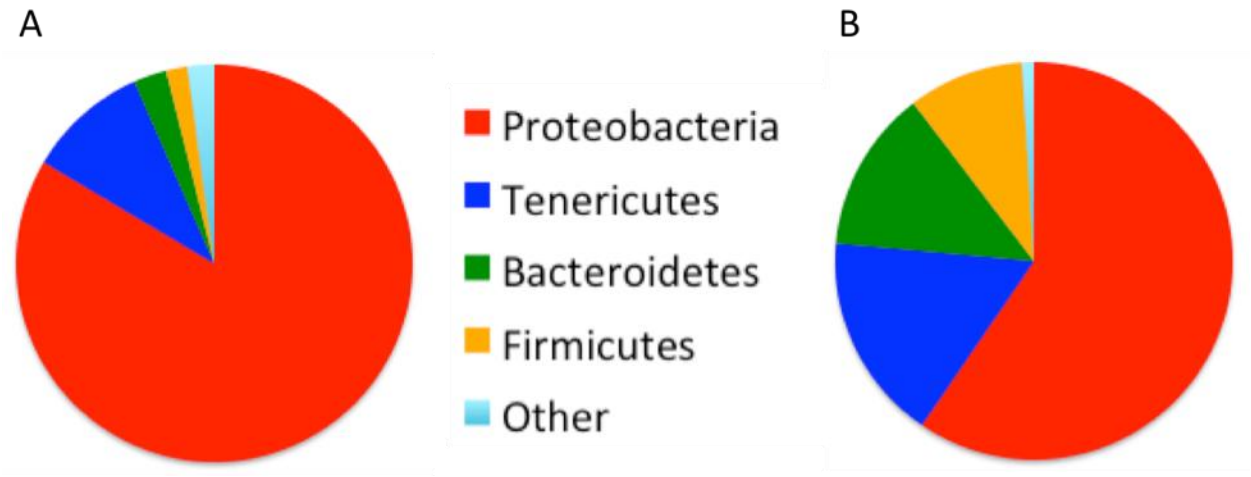


Figure 1. Proteobacteria dominate the microbiome of the lower respiratory tract of chickens. A) The microbiome of apparently healthy chickens consists mostly of Proteobacteria. B) The proportion of Tenericutes, Bacteroidetes and Firmicutes is considerably increased in diseased chickens.

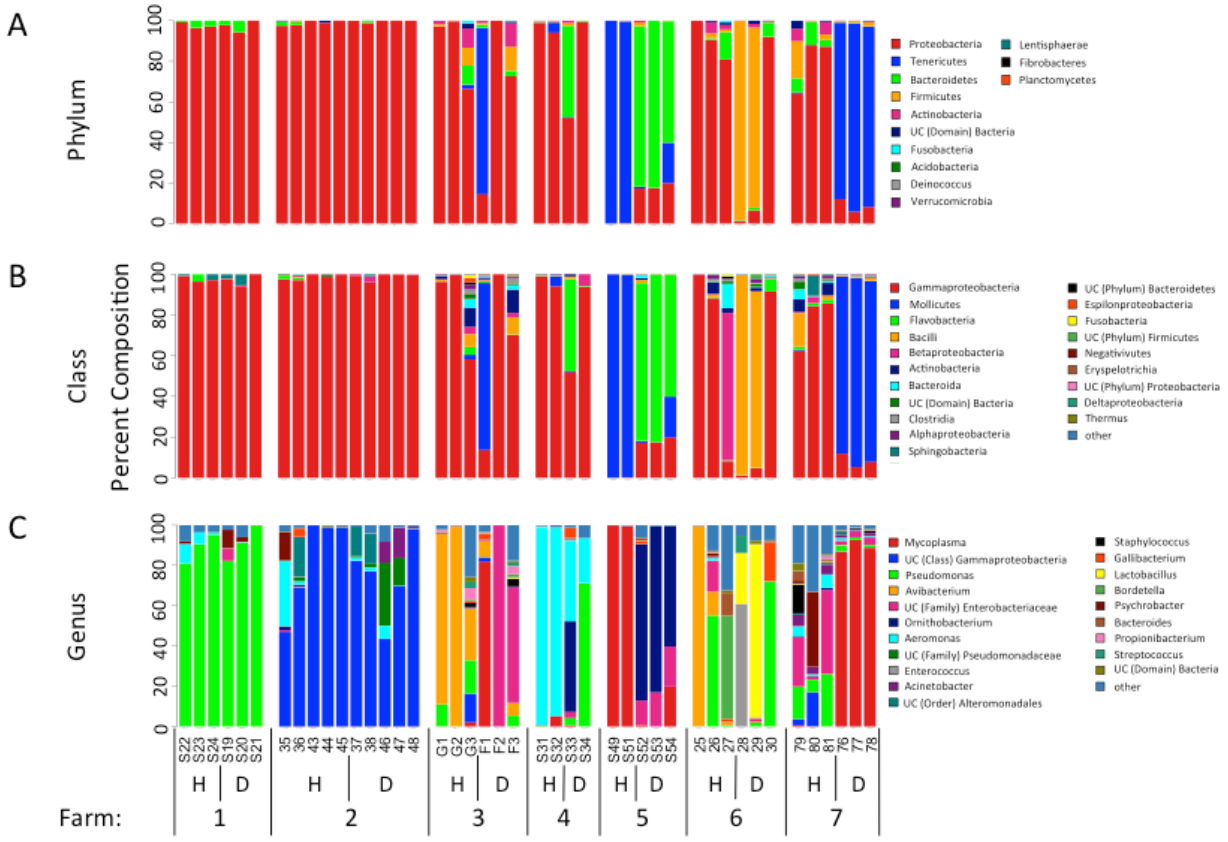
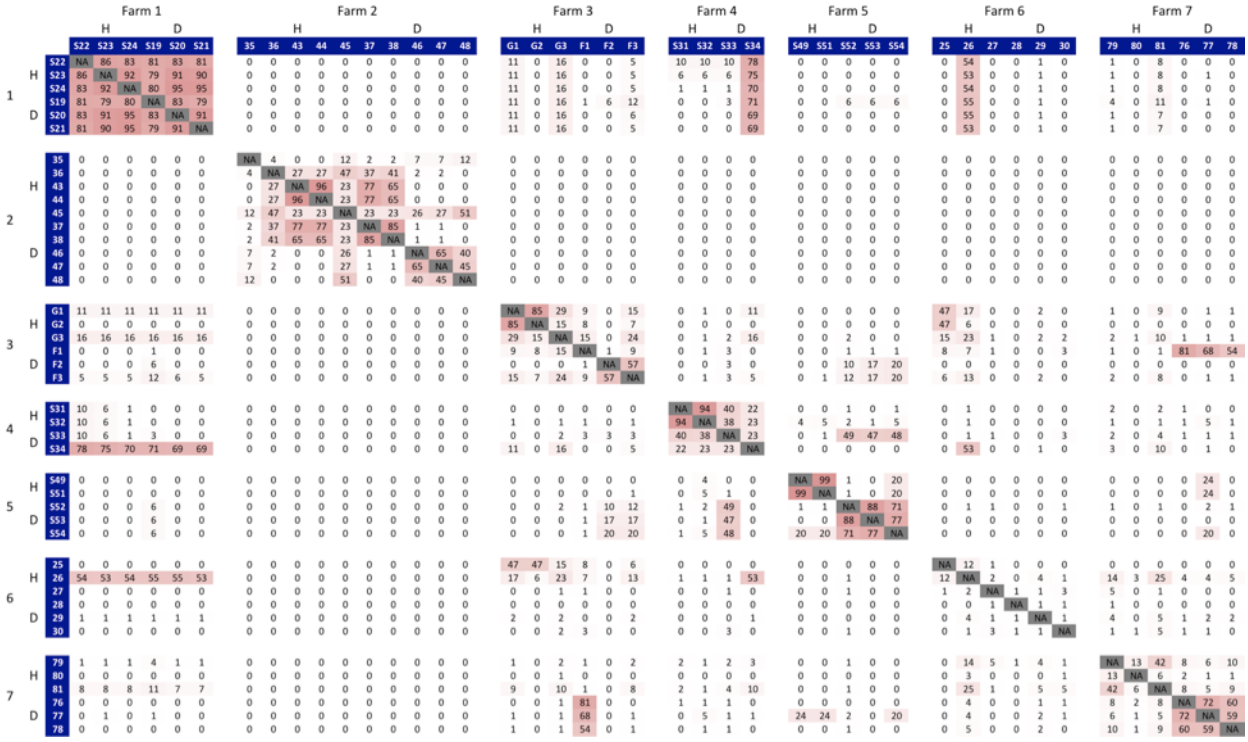


Figure 2. The microbiome of the lower respiratory tract of healthy and diseased chickens is dominated by a limited number of bacterial taxa. A) Distribution of Proteobacteria, Tenericutes, Bacteroidetes, Firmicutes, Actinobacteria and other phyla in the microbiome of healthy and diseased chickens. B) The microbiome of the majority of samples is dominated by Gammaproteobacteria. C) Gammaproteobacteria at individual farms belong to different taxa, indicating that the microbiome composition of the lower respiratory tract in apparently healthy chickens is largely determined by the environment. The microbiome of diseased chickens is enriched in pathogens such as *Ornithobacterium rhinotracheale* and *Mycoplasma synoviae*. Each stacked bar represents the microbiome of an individual chicken. Within each farm, the samples are grouped by healthy (H) or diseased (D). Colors within the stacked bar graphs represent the percentage of sequencing reads that belong to a given taxonomic classification indicated by the legends. UC, unclassified.

In contrast, the microbiome of the two examined specimens from farm 5 consisted almost entirely (99.9% and 99.3%) of Tenericutes, which comprised 9.9% of the total sequencing reads.

The microbiome of the lower respiratory tract of healthy chickens is environmentally determined. At the class level of assignment, the vast majority of Proteobacteria OTUs were Gammaproteobacteria. However, the microbiomes of healthy chickens from the individual farms were dominated by different genera, namely *Pseudomonas* (order Pseudomonadales) at farm 1, previously unclassified Gammaproteobacteria at farm 2, *Avibacterium* (order Pasteurellales) at farm 3, and *Aeromonas* (order Aeromonadales) at farm 4 (Fig. 2). The Gammaproteobacteria at farms 6 and 7 consisted of several genera, including *Avibacterium* (Pasteurellales), *Acinetobacter*, *Pseudomonas*, *Psychrobacter* (all Pseudomonadales) and *Aeromonas* (Aeromonadales). While the microbiomes of healthy chickens were consistent between the samples within most farms, they were generally different between farms, with very few exceptions such as samples G1 and G2 from farm 3 and sample 25 from farm 6 that contained remarkably similar microbiomes (Fig. 2, Table 1).

Table 1. Pairwise similarity scores of microbiomes based on OTU composition



The microbiome of the lower respiratory tract of chickens varied more with respect to farm than breed, management system, or geographic location, indicating that exposure to microbes and/or other aspects of the environment has a larger impact on the microbiome composition than the chicken host (Fig. 2, Table 1, Supplementary Table 1). For example, farms 4 and 5, which were located near one another, both raised hubbard broiler birds in controlled-house conditions, but the microbiome of the lower respiratory tract of these chickens was very different. While in farm 4 *Aeromonas* (order Aeromonadales, class Gammaproteobacteria, phylum Proteobacteria) was predominant among the healthy birds, farm 5 was dominated by *Mycoplasma* (Mycoplasmatales, Mollicutes, Tenericutes). Likewise, farms 1 and 2 both raise babcock layer chickens under open-house conditions. The microbiome of chickens from farm 1 was dominated by one *Pseudomonas* OTU whereas the

microbiome of chickens from farm 2 was dominated by several novel Gammaproteobacteria OTUs (Supplementary Figure 6).

The number of OTUs which were not shared between any two samples in a pair-wise comparison (Table 1, Supplementary Table 2), as well as OTU-based similarity scores (Supplementary Figure 6) showed that samples from farm 2 differed notably in their OTU composition. In fact, the microbiome of chickens from farm 2 was comprised of OTUs that were not present at any other farm (Supplementary Table 2). Mapping the sequencing reads against the SILVA database [12] and BLAST [9] searches in GenBANK revealed the presence of multiple, previously unclassified Gammaproteobacteria with less than 93% similarity to any known 16S rRNA sequence, suggesting that they may represent novel species.

The microbiome of the lower respiratory tract of chickens differs between diseased and healthy chickens. The number of distinct OTUs in the microbiome of diseased chickens was comparable to the number of bacterial OTUs determined from apparently healthy chickens. Accordingly, the microbiomes were also dominated by a limited number of bacterial taxa. Similar to healthy chickens, Gammaproteobacteria were predominant in the microbiomes of diseased chickens (Fig. 1), particularly at farms 1 to 4 (Fig. 2). However, taxonomic classification at the genus level showed that the microbiome of diseased chickens was also farm specific, as the Gammaproteobacteria differed between farms. Like in healthy chickens, *Pseudomonas* was identified in diseased chickens from farm 1, and 2 was dominated by potentially novel species, and bacteria from farm 3 were identified as *Aeromonas*.

With the exception of farms 1 and 2 at which diseased birds mostly contained the same bacterial OTUs as their apparently healthy counterparts, the microbiome composition differed between healthy and diseased chickens. Particularly, the proportion of Tenericutes, Bacteroidetes and Firmicutes

was considerably higher in diseased than in healthy chickens (Fig. 1, Fig. 2). Specifically, known respiratory pathogens of chickens were identified at high frequency in diseased but not in healthy chickens, including *Mycoplasma synoviae* (order Mycoplasmatales, class Mollicutes, phylum Tenericutes), which comprised up to 87% of the total sequencing reads in samples from farms 3 and 7, and *Ornithobacterium rhinotracheale* (Flavobacteriales, Flavobacteria, Bacteroidetes), which made up to 82.4% in samples from farms 4 and 5. In addition, single specimens from farm 6 contained *Enterococcus cecorum* (Lactobacillales, Bacilli, Firmicutes) in 59.4% of the sequencing reads, or the opportunistic pathogen *Gallibacterium anatis* (Pasteurellales, Gammaproteobacteria, Proteobacteria) in 19.1% of the reads (Fig. 2, Fig. 3).

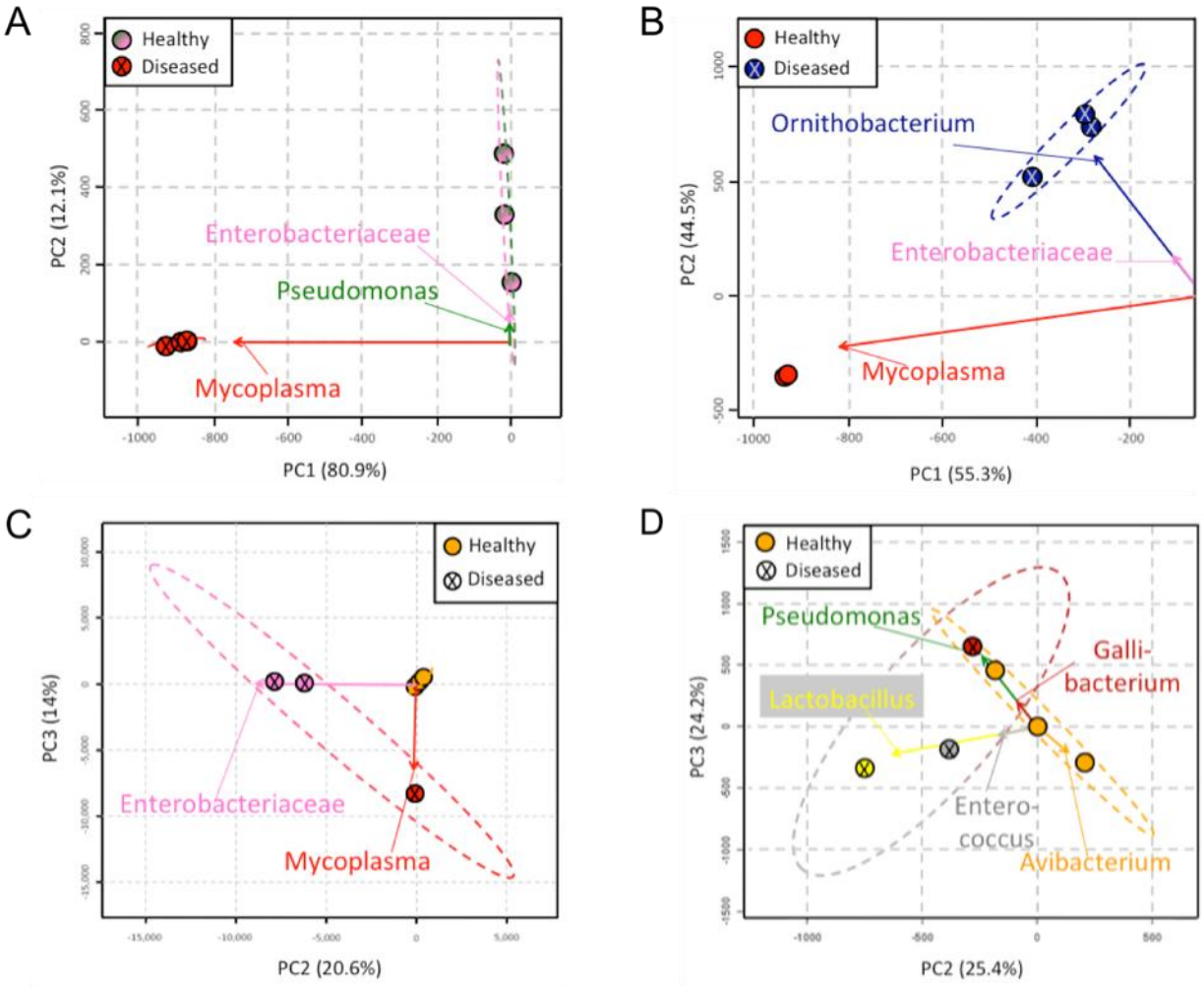


Figure 3. Principle Components Analysis (PCA) of the microbiomes of healthy and diseased chickens from selected farms. Principal components (PC) separate the microbiomes of healthy (circles) and diseased (circles with superimposed X) chickens. The PCA loadings of prominent bacterial taxa are depicted by arrows. Dashed ovals represent a projected 95% CI for healthy *vs.* diseased chickens per farm. A) Farm 7, B) Farm 5, C) Farm 3, D) Farm 6. The PCA biplots show the microbiomes of chickens with one predominant bacterial pathogen (A, B) and with two potential pathogens (C, D). Color coding as in Fig. 2C.

Healthy chickens at farm 7 contained a broad variety of bacterial taxa, including *Psychrobacter*, *Acinetobacter*, *Pseudomonas*, *Aeromonas* and *Enterobacteriaceae*, all of which were almost entirely replaced by *M. synoviae*, a bacterium that causes dyspnoea, nasal discharge, sneezing, production of soft-shelled misshapen eggs and hatching problems, which matched the observed disease symptoms (Supplementary Table 1). Similarly, *O. rhinotracheale*, which also causes respiratory disease in poultry, made up the majority of sequencing reads in diseased birds from farm 5 and of one sample from farm 4, and the observed symptoms of sneezing, gargling sounds and paralysis were consistent with infection with this bacterium (Supplementary Table 1). In contrast, neither *M. synoviae* nor *O. rhinotracheale* were found in any samples obtained from healthy birds (Fig. 2, Fig. 3). However, the microbiomes of healthy chickens from farm 5 consisted of primarily *Mycoplasma gallisepticum*, also a known poultry pathogen, which was also detected in 20% of the sequencing reads from a diseased chicken from this farm.

Differences in the microbiome of the lower respiratory tract of healthy and diseased chickens were also apparent at farms 3 and 4 that contained various Gammaproteobacteria. Healthy chickens at farm 4 almost exclusively carried *Aeromonas* while the microbiomes of diseased chickens also contained *O. rhinotracheale*, *G. anatis* and *Pseudomonas sp*. Similarly, the microbiomes of the healthy chickens from farm 3 were dominated by the poultry commensal *Avibacterium volantium*, but the microbiome of the diseased birds consisted of *M. synoviae* in one sample, and the other two samples contained *Enterobacteriaceae* (Fig 2, Fig. 3) that showed 99-100% identity to *Escherichia coli*, *Shigella sp.*, *Enterobacter cloacae* and *Enterobacter cancerogenus*.

DISCUSSION

Agricultural poultry provides valuable nutrition on a global scale. Disease outbreaks both impact production and can serve as a source of zoonosis, potentially leading to deadly outbreaks among human populations. This sample set represented a rare opportunity to conduct microbiome study in a relevant agricultural system with both economic and health applications. Access to animals of this type is generally restricted, as farmers often express concerns that scrutiny of an outbreak will lead to public alarm and subsequent decrease in sales. Further, it was necessary to sacrifice chickens in order to perform T-bal sampling. Despite these limitations, this study represents important initial steps towards characterizing the microbiome of the lower respiratory tract of a variety of healthy chickens. Differences were observed in the microbiomes between healthy and diseased chickens, and potential pathogens were identified even with relatively small numbers of diseased birds present early in an outbreak, showing the potential of NGS-based, culture-independent, pathogen identification methods.

The microbiome of the lower respiratory system of chickens was remarkably different from the chickens' gastrointestinal microbiome. While the lower respiratory tract was mainly colonized by Gammaproteobacteria, the gastrointestinal microbiota was dominated by Firmicutes such as *Lactobacillus* and *Peptostreptococcaceae*[181], which were either absent or comprised only a small number of the sequencing reads from the respiratory tract, with the exception of two diseased chickens from farm 6. In contrast, a noteworthy abundance of Gammaproteobacteria (predominantly *Enterobacteriaceae* such as *Escherichia*, *Salmonella*, and *Enterobacter*) was present in only one third of the analyzed gastrointestinal samples. Moreover, the median number of identified taxonomic units was considerably smaller in the respiratory (102 OTUs) than in the gastrointestinal (127 reported as species) tract. The use of the term "species" in the Kaakoush et al study [181] is potentially misleading because,

with few exceptions such as the above mentioned pathogens *M. synoviae*, *M. gallisepticum* or *O. rhinotracheale*, 16S rRNA sequences are usually too conserved to allow bacterial identification to the species level.

Diseased chickens from farm 7 were treated with commercially available antibiotics such as tylosin and enrofloxacin (Supplementary Table 1), but the treatment failed. This suggested infection may have involved resistant bacteria and/or with a non-bacterial infecting agent. Meta-genomic analysis revealed the presence of *M. synoviae*, a bacterial pathogen, which is thought to be susceptible to the applied antibiotics. However, *M. synoviae* was previously shown to persist in hens after two enrofloxacin treatments, whereby the antibiotics did not have any effect on *M. synoviae* recovery from tracheal swabs [249]. In addition, after the second treatment re-isolated mycoplasma clones showed a significant increase in the resistance level to enrofloxacin [249]. Moreover, high-level resistance to tylosin and erythromycin developed within only 2-6 *in vitro* passages in *M. synoviae*[250]. These observations suggest that *M. synoviae* clones from chickens at farm 7 were likely resistant to the administered antibiotics. Likewise, oxytetracycline treatment of *O. rhinotracheale* infection of chickens at farm 5 (Supplementary Table 1) proved ineffective, likely due to the presence of antibiotic resistant bacteria. Indeed, other studies showed an increase of *in vitro* antibiotic resistance profiles of *O. rhinotracheale* since the mid 90's, including increased resistance to tetracycline [251], and the majority of *O. rhinotracheale* isolates from 105 poultry flocks in the neighboring country Iran were also resistant to tetracycline [252].

Healthy hubbard broiler chickens from farm 5 were dominated by *M. gallisepticum,* a known poultry pathogen (Fig 2, Fig. 3). *M. gallisepticum* is not expected to completely dominate the microbiome of apparently healthy chickens. However, these chickens might have been sampled as apparently healthy before the onset of disease symptoms, or they may simply be carriers, facilitating the

spread of *M. gallisepticum* within their respective flock. Despite the small number of samples per farm, *M. gallisepticum* was also detected in one of the diseased chickens from this flock (sample S54), where it comprised about 20% of the total sequencing reads (Fig. 2).

Traditional pathogen identification seeks a single causative agent of disease, which is not always the case, as some studies revealed the presence of viral agents and co-infection with an opportunistic bacterial pathogen which otherwise does not cause disease in birds [253,254]. However, sequencing only the bacterial 16S rDNA gene may have missed additional agents, such as viruses or protozoa. In such cases bacterial OTUs might represent secondary infection not causally associated with disease. Nevertheless, more than one potential infectious agents were identified at some farms, namely *M. synoviae* and *Enterobacteriaceae* such as *E. coli* in Misri layer birds at farm 3, *O. rhinotracheale* and *M. gallisepticum* among chickens at farm 5, and *E. cecorum* and *G. anatis* among chickens from farm 6 (Fig. 2, Fig. 3).

This is perhaps, the first study to characterize the microbiome of the lower respiratory tracts of chickens. Each bird's microbiome was dominated by a small number of taxonomic classifications and that the consortium of organisms was relatively consistent amongst healthy birds at most farms, but differed between farms, showing environmental impact on microbiome composition. Detection of Gammaproteobacteria with less than 93% similarity to any known 16SrRNA sequences indicated the presence of potentially novel bacterial species. At several farms, the microbiomes differed between healthy and diseased chickens. Despite sampling a relatively small number of chickens, it was possible to detect known bacterial respiratory pathogens and to associate those pathogens with the observed disease symptoms. NGS sequencing results can thus be further used for targeted laboratory culturing of potential pathogens and subsequent tests aimed at fulfilling Koch's postulates as well as testing for antibiotic resistance.

MATERIALS AND METHODS

Ethics Statement and Sample Collection (T-bal). Tracheal broncho alveolar lavage (T-bal) was performed with all necessary personal protective equipment and with the approval of the International Animal Care and Use Committee (IACUC), USA, following protocols approved by the Ethical Research Committee of the University of Veterinary and Animal Sciences, Lahore, Pakistan. Chicken samples were collected from 7 farms in the Punjab province of Pakistan (Supplementary Table 1). No antibiotics were administered to birds for at least 3 weeks prior to sampling. The birds were euthanized via intravenous injection of potassium chloride, 1-2 molar equivalents (mEQ) per kg of body weight. Once euthanized, birds were placed in the dorsal recumbent position. A ventral midline incision was made along the mandible to the thoracic inlet, and surrounding fascia was removed from the trachea. The trachea was then cut transversely in the mid-cervical region. A 60-ml catheter-tipped disposable syringe (STAR, Jiangsu Kanghua Medical Equipment Co. Ltd., China) attached to a sterile pipette tip was placed into the trachea forming an airtight seal. Air was withdrawn from the respiratory tract of the birds with a syringe until partial collapse of the extra-thoracic portion of the trachea was observed. The syringe was removed while carefully maintaining the vacuum by clamping the airway below the insertion point. Another catheter syringe was then placed in the trachea and used to add 5 ml of sterile Phosphate Buffered Saline (PBS). The bird was rocked from sided to side to allow fluid to contact all parts of the lower respiratory system. Approximately 1 ml was then slowly withdrawn until a partial collapse of the trachea was observed.

DNA extraction and quantification. The T-bal samples were centrifuged at 14,000 ×*g* for 10 min. The pelleted bacteria were re-suspended in 300 µl of sterile PBS. Genomic extraction was performed with BiOstic FFPE Tissue DNA Isolation Kit (Mobio, USA) following the protocol of the manufacturer. The

extracted DNA was shipped to the Pennsylvania State University, USA. Sample DNA quality was checked with both Bioanalyzer and Qubit flourometer prior to DNA library construction (Invitrogen, USA).

PCR amplicons and 454 sequencing. One-way read amplicons (Lib-L) were prepared using barcoded fusion primers 27F (CCATCTCATCCCTGCGTGTCTCCGACTCAG-MID-AGTTTGATCMTGGCTCAG) and 907R (CCTATCCCCTGTGTGCCTTGGCAGTCTCAG-TACGGGAGGCAGCAG TACGGGAGGCAGCAG) by amplifying a ~900 bp fragment of the 16S rRNA that included variable regions V1-V5. Briefly, samples were denatured at 94^{o}C for 4 min, 35 cycles of: 94^{o}C (15sec), 55^{o}C (45 sec) and 72^{o}C (60 sec) were performed with a final extension at 72^{o}C for 8 min (Gene AMP PCR System 9700; Applied Biosystems, Foster City, CA). PCR products averaging ~900 bp in length were size selected on agarose gels and purified using the QIAquick PCR purification kit (Qiagen, Valencia, CA). Prior to library preparation, the quality of the amplicons was assessed using a Bioanalyzer DNA 7500 Chip (Agilent). Sample quantitation was performed using Qubit flourometer quantification. Next-generation library preparation, library quantification and emulsion PCR were performed according to the manufacturer's instructions (Roche Applied Science). The libraries were sequenced on a GS FLX+ instrument using GS FLX Titanium chemistry according to the manufacturer's instructions (Roche Applied Science). GS-FLX-Titanium sequence data files (.sff) were generated using the GS amplicons software package (Roche, Branford, CT).

Computational sequence processing. Raw sequence reads were processed in MOTHUR[246,247] in compliance with the operating manual for 454-pyrosequencing. High-quality reads were aligned and mapped against the SILVA database [248]. The Operational Taxonomic Units (OTUs) were defined as sequences exhibiting greater than 97% sequence similarity. Rarefaction curves were generated in

MOTHUR following the standard operational procedure. The estimated % OTUs captured were computed by estimating the theoretical asymptote for each rarefaction curve and comparing it to the number of OTUs present in the sample. OTUs were merged to their closest taxonomic classification in METAGENASSIST[255].

Alpha-diversity was based on Whittaker's classical definition [256]. However, I took the number of OTUs rather than number of species as my indicator of alpha-diversity. Beta-diversity was computed as a pairwise comparison of two samples by determining the number of OTUs which are present in one sample but not both. Stacked bar graphs were produced by calculating the percentage of the microbiome occupied by individual taxa in each sample. For visual clarity, only the top 20 most abundant taxonomic classifications of the normalized total microbiome were represented (Fig. 2). Principal Component Analysis (PCA) was performed in METAGENASSIST[20], the taxonomic classifications that contribute to the greatest amount of variance in the data set were inferred from the PC loadings. Each solid circle on the plots represents a sample and PCA loadings (arrows) represent taxa. Dashed ovals indicate 95% CI.

OTU based sample similarity was computed first by calculating the percentage of reads belonging to each OTU in each sample. Then, for each pairwise comparison, similarity was determined based on either presence or absence of the OTU as well as its relative abundance of reads assigned in each sample. For example, let's assume that sample 1 contains 25% OTU A and 75% OTU B and 0% OTU C. Sample 2 contains 50% OTU A, 0% OTU B and 50% OTU C. Similarity would be (25% A + 0% B + 0% C)/2, giving an overall similarity of 12.5%.

Sequence identification to the species and/or genus levels was done by computing a consensus sequence for OTUs of interest in MOTHUR[10,11] and consequently performing a blast [9] search with the

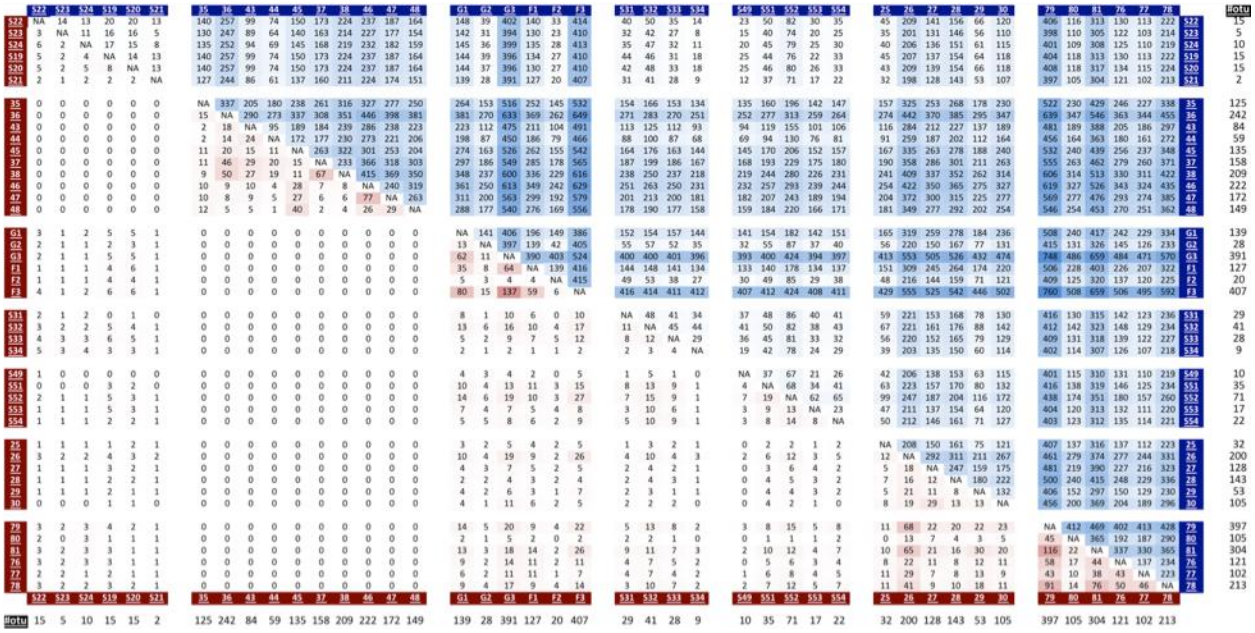
consensus sequence. Sequences were considered to be novel if they showed less than 94% identity with known sequences.

Supplementary Tables

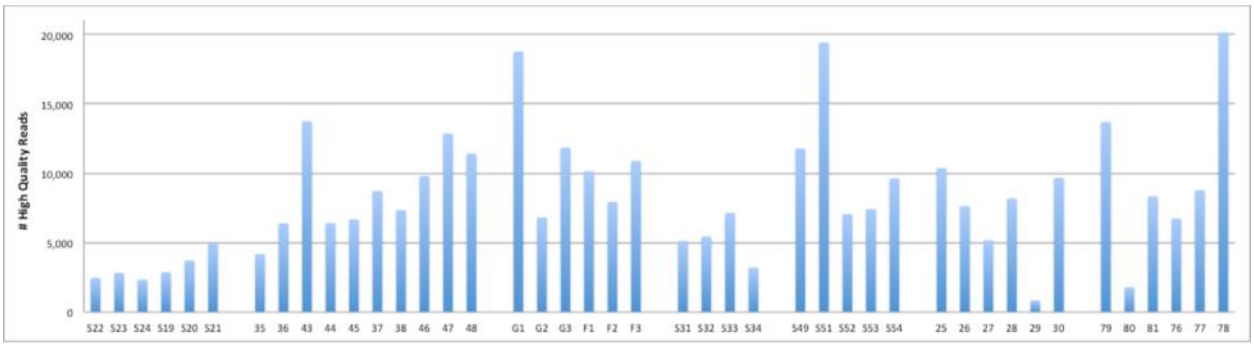
Supplementary Table 1. Sample properties.

sample ID	Farm	Location	Condition	Management System	Bird Type	Detailed Bird Type	Gender	Sample Date	Flock Size	Age at Sampling	Age at Disease Onset	Symptoms	Attempted Treatment	Vaccination
S22	1	N32.444722 E74.315556	Healthy	Open House	Layer	Babcock	Female	12/22/11	10,000	42 weeks	41 weeks	7-9% mortaltiy, nasal discharge, dyspnea, head swelling, blackish combs, tracheitis caseous coverings on lungs, thick fivrous coverings on nliver and kidneys, sever hemorrhages at cecal tonsils, egg production decrease 80% -> 35%, normal eggs.	No Record	Newcastle Disease Virus, Pox virus, Infectious Bronchitis Virus, H9 influenza virus.
S23														
S24														
S19			Diseased											
S20														
S21														
35	2	N32.29155 E74.434917	Healthy	Open House	Layer	Babcock	Female	12/26/11	7,500	55 weeks	52 weeks	nasal discharge, reddish black combs, swolen heads, increased body temperature, dificulty breathing, hemorrhages on trachea, congested lungs, hemorrhages on cecal tonsils, nephritis, additional worm infestation, 90-95% morbidity, 7% mortality	Antibiotics	No Record
36														
43														
44														
45														
37			Diseased											
38														
46														
47														
48														
G3	3	N29.409722 E71.658333	Healthy	Free Range	Layer	Misri	Female	10/24/11	1,700	36 weeks	35 Weeks 5 days	140 dead, nearly all morbid, dyspnea, rales, fever , nasal discharge, occasionally diarrhea in some birds	None	NVD, no recurring vaccination
G2														
G1														
F3			Diseased											
F1														
F2														
S31	4	N31.58905 E73.8704	Healthy	Controlled House	Broiler	Hubbard	No Record	12/31/11	No Record	No Record	31 days	80-90% morbid, 20% loss in first 5 days, dificulty breathing, mucus discharge from the nose, paralysis of one side of the body, increase temperature, off feed, Coccidiosis also observed	Multivitamins, liver tonics, antibiotics (Ciprofloxacin), Amprolium and ADEK	No Record
S32														
S33			Diseased											
S34														
S49	5	N31.715 E73.985	Healthy	Controlled House	Broiler	Hubbard	No Record	12/19/11	30,000	38 days	33 days	nasal discharge, gargling sounds, neck paralysis during terminal stage of disease. 70% morbidity, 4000-5000 dead.	oxytetracycline	No Record
S51														
S52														
S53			Diseased											
S54														
25	6	N31.637533 E73.923883	Healthy	Controlled House	Breeder	Hubbard	Male	12/19/11	50,000	40 weeks	37 weeks	swolen head, nasal discharge, dyspnea, gargling sounds, lacrimal discharge. Egg production 78-<10%, hatchability decreased to 30-35%, 5% female 8-9% male died during outbreak, all birds were morbid	enrofloxacain, Amentidine, Tylosin, and Colistin sulphate, no effect	No Record
26														
27							Female							
28														
29			Diseased				Male							
30							Female							
79	7	N31.647033 E73.885833	Healthy	Controlled House	Breeder	Hubbard	Female	12/19/11	125,000	66 weeks	61 weeks	sneezing, nasal discharge, 60-80% morbidity, 12-15% mortality. Hatchability decreased to 30% egg production decreased to 45%	Comercially avaliable antibiotics, including tylosin, enrofloxacin and ciprofloxacin with expectorants.	No Record
80														
81														
76			Diseased				Male							
77														
78							Female							

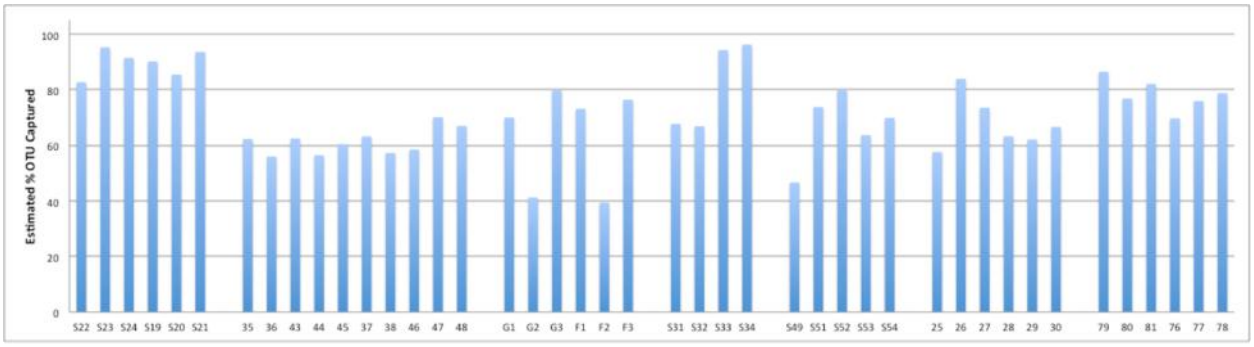
Supplementary Table 2. Pairwise comparison of beta diversity and OTUs shared between chickens microbiomes



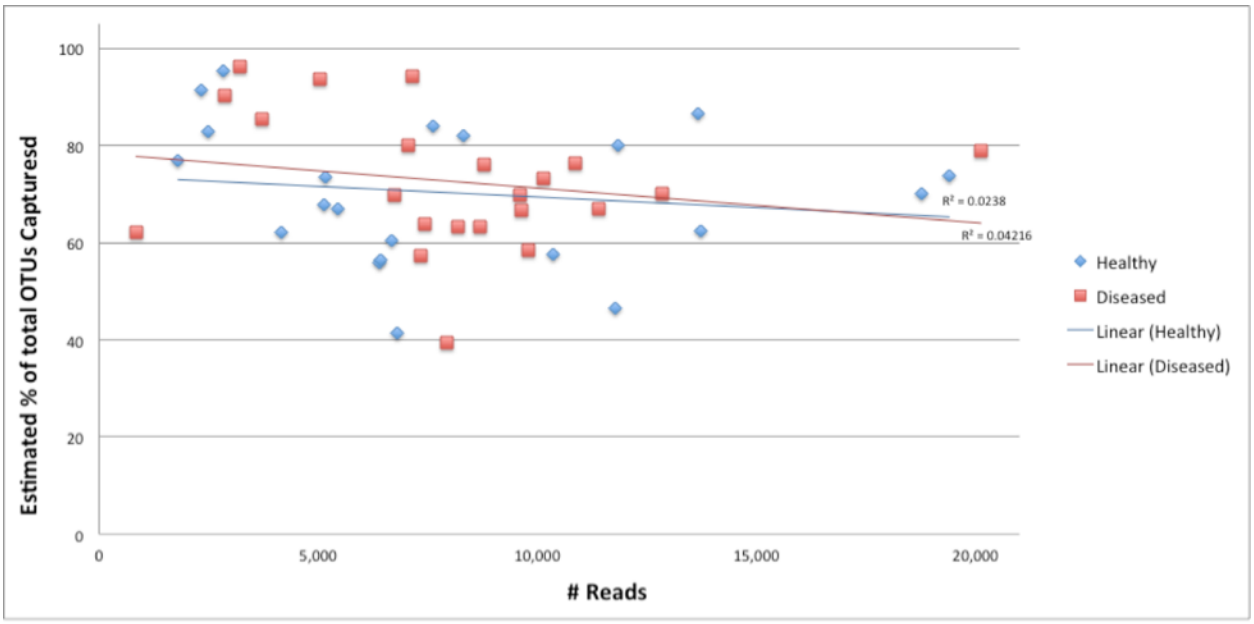
Supplementary Figures



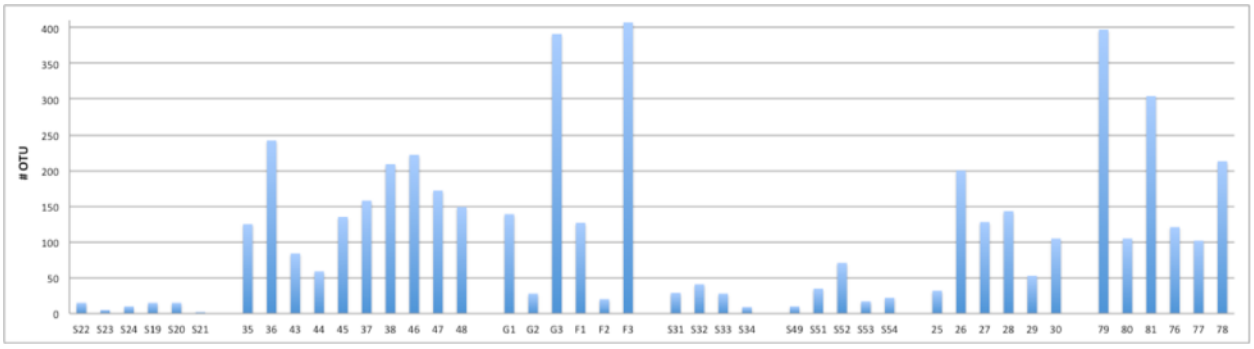
Supplementary Figure 1. Number of high quality reads per sample



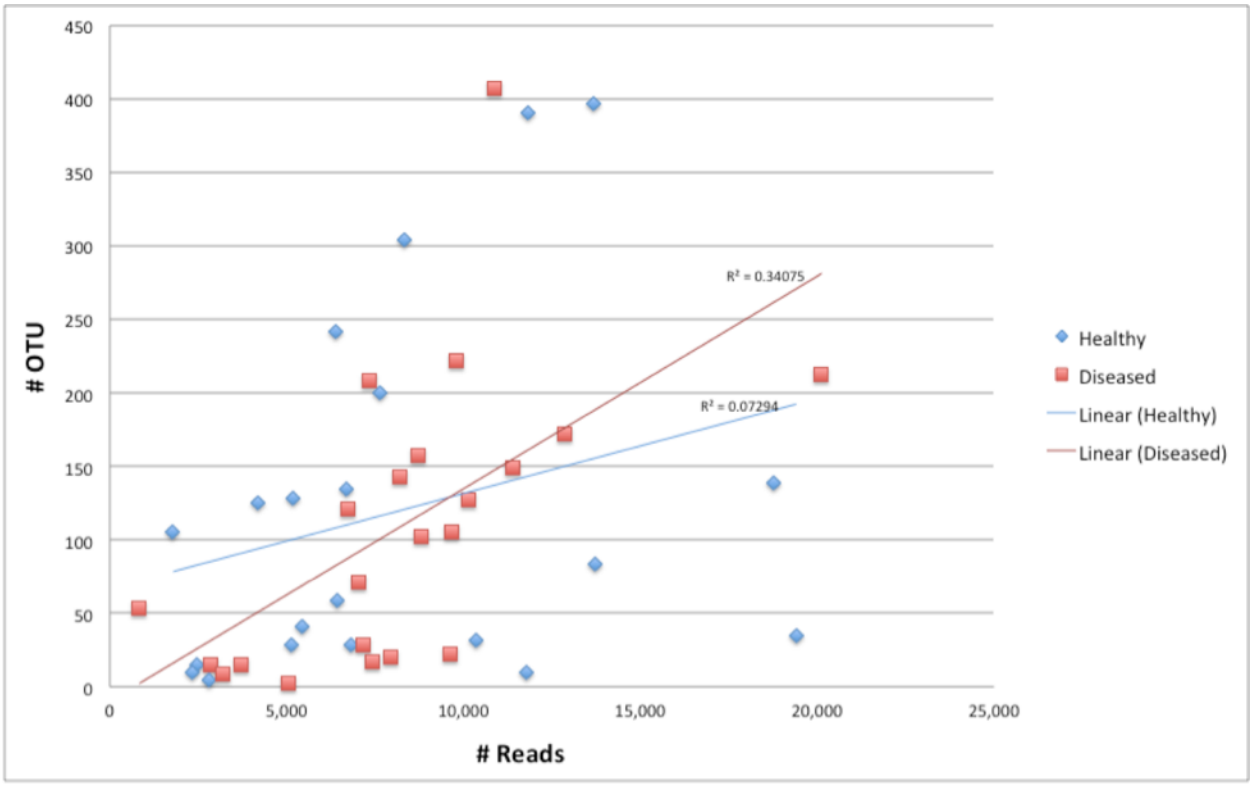
Supplementary Figure 2. Estimated percentage of OTUs captured per sample



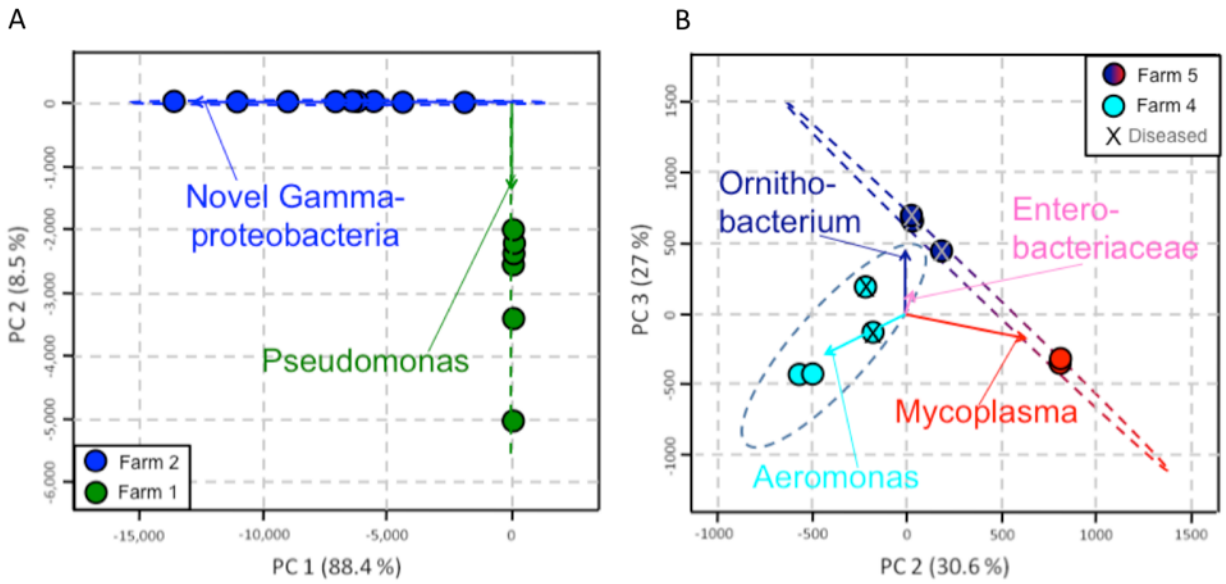
Supplementary Figure 3. Number of high quality reads per sample *vs*. estimated percentage of OTUs captured per sample



Supplementary Figure 4. Number of OTUs per sample



Supplementary Figure 5. Number of high quality reads per sample *vs.* number of OTUs per sample.



Supplementary Figure 6. Principal Component Analysis (PCA) comparing two farms, color code similar to figure 2C A) Farms 1 and 2. B) Farms 4 and 5.

Chapter 2: Relating epigenetic landscape to gene transcription level in the G1E, G1E-ER4 model system for erythroid maturation.

FIELD REVIEW

Epigenetics refers to heritable changes, which are not found in DNA. Commonly considered epigenetic factors include CpG methylation, DNAse Hypersensitivity (DHS), Transcription factors and Histone modifications. Further complicating matters, DNA coils into complex superstructures [191]. Well funded projects are now underway attempting to elucidate epigenetic patterns. The most notable of which is the Encyclopedia of DNA Elements (ENCODE). ENCODE has generated in excess of 1,640 data sets in greater than 147 cell types [192–194]. As a result, now greater than 80% of the human genome is considered functional [191].

Many technologies have arisen over the past several years in the quest to understand epigenetic effects. RNA transcribed regions are assessed using; RNA-seq, Cap Analysis Gene Expression (CAGE), Paired End RNA –Seq (RNA-PET) and manual annotation [191]. Protein coding regions are identified using mass spectrometry [191]. Transcription factor binding sites are identified with Chip-Seq and DNAse-seq [191]. While, chromatin conformation can be illuminated via; DNase-seq, Formeldehyde assisted isolation of regulatory elements (FAIRE-seq), histone ChIP) [191]. Additionally, methylation sights can be identified using Reduced Representation Bisulfite-seq (RRBS) [191]. Reproducibility amongst biological replicates was often an issue, thus computational approaches such as the Irreproducible Discovery Rate (IDR) was

developed to provide robust and conservative estimates of biological phenomena present in data sets [195].

RNA based studies indicate that we have yet to discover many protein coding genes [196]. Protein coding exons cover only 1.22% of the genome with their respective genes covering 2.94% [191]. However, >200bp RNA activity has been found to cover 62% of the genome with 31% of those transcripts originating in intergenic regions [196]. Additionally, many <200bp RNAs have also been found [191]. Thus, of 62,430 identified transcription start sights TSSs only 44% reside within 100bp of known genes [191].

Chromatin segmentation, which incorporates a variety of epigenetic marks, has been used to identify 399,124 enhancer like and 70,292 promoter like regions [191]. Further subdivisions also yield distinct functional properties [191]. Moreover, promoter function has been shown sufficient to explain most RNA expression [191].

DNA Hypersensitivity Sites (DHS) are generally considered the hallmark of regulatory regions [197,198]. The majority of which are found distal to the TSS. It is currently estimated that ~50% of DHS are present in human cell lines [199].

Methylation of Cytosine at CpG di nucleotides is heavily related to the regulation of gene expression [191]. High levels of CpG methylation tend to increase genomic stability [200]. Promoter methylation or methylation close to a gene is associated with repression while genic methylation correlates with transcriptional activity [201]. Overall, DNA methylation controls chromatin accessibility with 96% of CpG islands exhibiting differential methylation in at lease one assayed cell type or tissue [201].

ENCODE has assayed greater than 199 DNA binding proteins and RNA polymerase components within 72 cell types. The completeness of coverage per cell type has varied. About 636,336 binding regions have been identified covering 8.1% of the genome [191]. Of which, 86% are associated with a

strong DNA based binding motif [191]. About 94.4% of binding sites lie within defined DHS regions. Protein binding sites with relatively low signal are thought to either have low affinity to their respective binding sites or bind through intermediate complexes [199]. Typically transcription factors work collectively and are not randomly distributed with 3,307 pairs of statistically co-associated factor identified involving 114 out of 117 assayed factors [191].

DNA is generally found wrapped around protein complexes referred to as histones. Modifications present on these histones are determining of DNA accessibility [202] and include; acetylation, methylation, ubiquitination, phosphorylation and ADP-ribosylation [203–206]. In general, these modifications vary across cell types and have been implicated as indicators of gene regulation [202]. The best models relating epigenetic marks to gene expression first predict whether a gene is active or silent and then predict quantitative expression level, R^2 =0.81 [207]. Activating histone marks; H3K27ac, H3k9ac, H3k4me3 and K3k4me2 explain most variation in expression level [207]. However, for a subset of genes, repressive marks H3k27me3 and H3k9me3 can well predict expression level [207].

Chromatin separated by hundreds of kbp has been shown to physically interact via chromatin interaction analysis by paired end tag sequencing (Chia-Pet) and is thought to be important in controlling gene expression [201]. Assaying only 1% of the genome in 4 cell types, ENCODE has identified hundreds of statistically significant long range chromatin interactions [208]. On average each TSS interacts with 3.9 distal chromatin sights and each distal element is on averaged interacted with by 2.5 TSSs [208]. As assessed via Chia-pet, promoter regions of 2,324 genes were involved in single gene enhancer promoter interactions and 19,813 were involved in multi-gene interaction complexes [209]. Additionally, 50-60% of chromatin interactions occurred in only 1 of 4 cell lines, indicating high levels of cell type specific variability [208].

Beyond chromatin segmentation, epigenetic data can be compiled via Self Organizing Map (SOM)[191]. Applying SOM to ENCODE data 228 distinct regions were discovered, each enriched for specific gene ontology (functional enrichment) terms [191]. Including 19 SOM unties that are enriched for GWAS SNPs [210]. These SNPs indicate functional relevance and 88% of relevant SNPs are intronic or intergenic with 12% overlapping with transcription factor binding sites and 34% overlapping with DHS sites [210]. Transcription factor binding and histone modifications have been shown to differ on parental alleles [211].

Epigenetic patterns have been associated with a variety of behaviors, environmental stresses and diseases [212]. For example, economically deprived individuals and manual laborers have been shown to have lower global methylation patterns in peripheral blood leukocytes than other individuals [213]. TNF-α promoter methylation has been shown to increase with shift work and IFN-y and alu hypomethylation have been associated with job security [214]. Even diet has been shown to effect epigenetic patterning [215].

The epigenome has been noted to be especially vulnerable to environmental factors during embryogenesis [216–218]. This is thought to be due to a high DNA synthesis rate and elaborate DNA methylation patterning necessary for normal tissue development during these stages of development [216–218]. Early life experiences and stress related outcomes in mice have been shown to affect epigenetic patterns and have provoked interest in the same possibilities in humans [219,220]. Individuals conceived during the Dutch hunger winter at the end of World War II were shown to have an altered methylation pattern at the IGF2 locus, which was detectable 60 years after the event [221].

Toxicants such as benzene, arsenic, iron, nickel, bisphenol A (bpa) and cigarette smoke have also been associated with epigenetic changes [212]. Exposure to benzene has been related with decreases in global methylation [222]. Exposure to organic pollutants [223], air pollution[224], lead [225] and arsenic [226]were also associated with global hypomethylation, though some genes were found to be hyper methylated. Exposure to iron, arsenic and nickel was associated with higher global levels of the activating histone

mark H3k4me2 [227]. Additionally, nickel exposure was also associated with a global increase in H3k4me3 and a decrease in H3k9me2 when compared to controls [227].

BPA exposure during pregnancy [228] has been shown to result in a variety of changes associated with breast cancer, prostate cancer, reproductive dysregulation and behavioral abnormalities [217,218,229–232]. These changes are thought to be mostly epigenetic mediated and in rat studies BPA has been shown to alter gene specific dna methylation patterns in the male prostate [230].

Notably, methylation profiles have been shown to differ between cancer types with a general pattern of genome wide hypomethylation with hypermethylated tumor suppressor genes [233]. Schitzophrenia has also been associated with varied epigenetic profiles and treatment with histone deacetylase inhibitors has been shown to decrease psychotic symptoms [234]. Many antidepressant drugs are associated with epigenetic changes [235,236].

PROJECT ABSTRACT

Erythropoiesis is the process by which red blood cells (erythrocytes) differentiate from lineage committed progenitor cells. This process is associated with sweeping morphological changes including loss of nucleus, loss of organelles and dramatic accumulation of hemoglobin. Due to mechanical stress and lack of repair machinery erythrocytes life span is generally limited to 120 days. Erythrocytes must thus be continuously produced and mis-regulation of erythropoiesis has been associated with serious health consequences. Epigenetic factors: H3k4me3, H3k4me1, H3k27me3, H3k9me3, GATA1, GATA2, TAL1, CTCF, POLR2A, DHS and gene transcription levels were assayed in the G1E, G1E-ER4 model system for erythroid maturation. Epigenetic landscape near the TSS was found to be sufficient to determine whether a gene was actively transcribed or silent. Change in epigenetic landscape was sufficient to predict change in transcription level when extreme expression changes occurred. Overall these results indicated that epigenetic landscape near the TSS sets a gene as either permissive or non permissive with respect to transcription.

INTRODUCTION

Hematopoiesis is the process by which pluripotent stem cells differentiate to form a variety of blood cells. In adults, hematopoiesis occurs in red bone marrow while at different stages of gestation it also occurs in the yolk sack and later the liver. Pluripotent cells can differentiate to form white blood cells (leukocytes), platelet (thrombocyte) producing cells (Megakaryocytes) and red blood cells (erythrocytes).

Erythropoiesis is the process by which erythrocytes are produced. It is associated with sweeping morphological changes, including loss of the nucleus and organelles along with a dramatic accumulation of hemoglobin. Erythrocyte cytoskeleton retains flexibility. This is necessary as erythrocytes average 7.5 micrometers in diameter and must transverse capillaries as small as 3 micrometers in the spleen. Stress produced under these conditions combine with the erythrocytes intrinsic lack of repair capabilities limits erythrocyte lifespan to 120 days. Thus, erythroid production must constantly be balanced with erythrocyte loss to facilitate proper circulatory function. Dis-regulation of erythroid proliferation can lead to anemia in the case of under production of erythrocytes and polycythemia when erythrocytes are over produced, both of which can lead to notable health concerns.

The transcription factor GATA1 has been termed the master regulator of lineage committed erythroid maturation [257–260]. It is necessary but not sufficient for differentiation to occur and has been shown to work cooperatively with erythropoietin to regulate erythropoiesis.

Committed erythroid cell line differentiation is commonly studied *in vitro* via the G1E, G1E-ER4 immortalized cell lines. G1E represents a committed proliferative but undifferentiated erythroid progenitor cell. It is a GATA1 knock out cell line, which expresses minimal levels of GATA1. G1E cells are thus terminally arrested in the undifferentiated but lineage committed state. In contrast, G1E-ER4 cells express an estradiol dependent GATA1 construct and thus represent erythroid precursor cells in the

absence of estradiol but undergo synchronous differentiation toward an orthochromatic stage in the presence of estradiol.

The ENCODE consortia has recently related epigenetic landscape to gene transcription level in a variety of cell lines [193]. They showed that histone modifications near the TSS better predict transcription levels than transcription factors occupancy did [261]. Additionally, they showed that assaying only a few key epigenetic factors was sufficient to predict transcription levels in most cases [261].

Histone modifications DHS have previously been found to differ amongst genes in the G1E G1E-ER4 system [262,263]. Transcription factors previously found to associated with GATA1 have also been shown to display differential occupancy patterns. Here I examine how changes in epigenetic factors correlate with changes in transcriptional activity genome wide during erythroid differentiation. Four histone modifications H3K4me3, H3K4me1, H3K27me3 and H3K9me3, and five transcription factors, GATA1, GATA2, CTCF, TAL1 and POLR2A were examined in this study. While the epigenetic relationships of all these factors are not fully understood, strong trends have been observed.

H3K4me3 is known to mark promoters and has been associated with actively transcribed genes [193,261]. H3K4me1 is known to mark enhancers, which is often converted to H3K4me3 near transcription start sites when gene transcription level is increased [193,261]. H3K27me3 and H3K9me3 have previously been shown to represent two distinct mechanisms of gene transcription repression [264,265]. DHS has long been associated with both cis-regulatory modules and active promoters [197].

GATA1 is as previously described considered to be the master regulator of erythropoiesis [266,267]. GATA2 is also a hematopoetic transcription factor from the same gene family, which has been shown to be necessary during other hematopoietic, non erythroid committed cellular differentiation processes [268]. CTCF is thus far the only known insulator in mammals. Repressive function for CTCF has also been characterized with respect to the insulin-like growth factor 2 locus [269,270]. However, its most interesting

function may be its ability to facilitate DNA looping via binding of homodimers [270,271]. TAL1 is required for erythropoiesis and is found co-occupying binding sights with GATA1 [272]. Like GATA1, TAL1 is necessary but not sufficient for erythropoiesis [272]. TAL1 is also necessary but not sufficient for additional hematopoitic maturation processes [272]. PLOR2A is an integral part of normal cellular transcription machinery. Its presence generally marks actively or poised to be transcribed genes. However, the effect of epigenetic factors on transcription level is not fully understood.

In this study, epigenetic factors were assayed pre and post differentiation in the G1E G1E-ER4 in vitro system for erythroid committed differentiation. Even with a limited sample of epigenetic factors, epigenetic landscape near the TSS was able to predict with approximately 90% accuracy whether a gene was actively transcribed or not. Change in epigenetic landscape was sufficient to explain change in transcription level amongst genes, which experienced the most extreme transcription level changes during differentiation. Additionally epigenetic factors working in a less than additive / redundant fashion were found to affect gene transcription levels.

RESULTS

The epigenetic landscape near the transcription start site (TSS) is sufficient to predict a gene as expressed or silent. The expression levels of 15,960 mouse genes in G1E and differentiated G1E-ER4 cells, aggregated epigenetic signals for 10 epigenetic features near the TSS, namely the histone marks H3K4me1, H3K4me3, H3K27me3 and H3K9me3, and the transcription factors,POLR2A, GATA1, TAL1, CTCF and GATA2as well as DNAse Hypersensitive Sights (DHS) were studied. Utilizing principal component analysis two epigenetic feature clusters were observed(Figure 1A, Supplementary Figures 1 and 2).

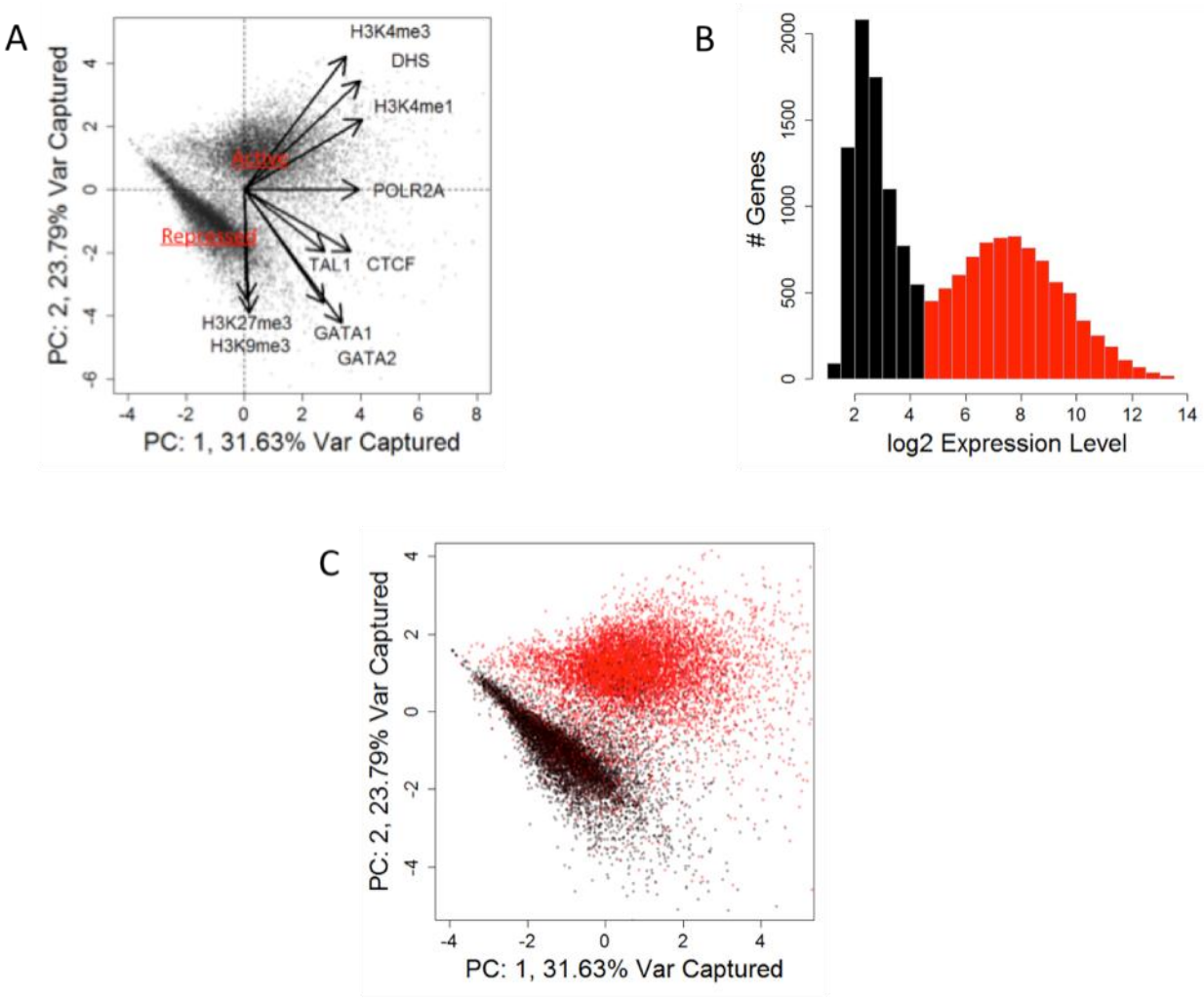


Fig.1. **Relationship between epigenetic landscape and expression level for 15,960 pre-differentiation mouse genes.**A) Principal component analysis of the pre differentiation epigenetic landscape near the TSS. Genes are represented by dots whereas factor loadings are denoted by arrows B) The pre differentiation transcription level profile of silent (black) and expressed (red) genes. C) PCA bi-plot of the epigenetic landscape near the TSS separating silent (black) and expressed (red) genes.

A transcriptionally “active” cluster was characterized by the presence of H3K4me3 and/or DHS at or near the TSS and/or the presence of H3K4me1 near the TSS. Conversely, a repressed cluster was characterized by the presence of H3K27me3 and/or H3K9me3 at or near the TSS. The presence of H3K27me3 along with H3K4me1 at the TSS or a lack of the above histone modifications was also associated with repressed gene expression. Correlation analysis indicated that the activating marks H3K4me3, H3K4me1 and DHS positively correlated with each other (Supplementary Table 1). Each activating marks negatively correlated with the repressive marks; H3K27me3 or H3K9me3. Moreover, repressive marks did not correlate with one another, suggesting separate mechanisms of action. As expected from previous studies , GATA1 and TAL1 positively correlated in the post differentiation cell state, as they have been shown to co-occupy sights genome wide.

Gene expression profiles of pre (G1E) and post (G1E-ER4) differentiation cells were found to follow a bimodal distribution (Figure 1B, Supplementary Figures 3A and 3B). Expressed genes mostly resided in the active cluster while silent genes were mostly associated with the repressed cluster (Figure 1C). However, genes with expression levels near the cutoff (Figure 1B) showed less separation when mapped to the epigenetic clusters (Supplementary Figure 4). A Linear Discriminate Analysis (LDA) model of the data indicated that for approximately 90% of the genes the epigenetic landscape near the TSS was sufficient to predict a gene as expressed or silent (Supplementary Figure 5). By applying a two step model (which first calls a gene as expressed or silent and then attempts to predict quantitative expression level) the epigenetic landscape was able to explain 63% of variability in the expression level (Supplementary Figures 6A and 6B). Yet, when examining only expressed or only silent genes the epigenetic landscape only explained 12% (expressed) and 22% (silent) of the variation in expression level (Supplementary Figure 7),indicating that the epigenetic landscape near the TSS is sufficient to determine a genes expression condition but not necessarily its quantitative expression level.

Among the individual epigenetic features, histone marks and DHS were more strongly correlated with expression level than were transcription factors (Supplementary Table 2). Evidence for repressive function for GATA2 was observed(Supplementary Figures 5C and 5D) As expected, POLR2A was associated with higher levels of expression, as was GATA1 in the post differentiation state. Interestingly, CTCF was significantly associated with higher levels of expression in the pre differentiation state.

Extreme changes in expression level can be explained by changes in epigenetic landscape. Since the epigenetic landscape largely determined the expression condition, if changes in the epigenetic landscape also affected the expression level was determined. Utilizing regression analysis significant but limited power to relate changes in the epigenetic landscape to the expression level (r^2 = 0.03) were noted (Supplementary Figures 8 and 9) This was regardless of level of gene expression. Similarly PCA plots also showed no clear separation between the expression level changes.

By applying an aggressive weighting scheme, in which genes that experienced large changes in the expression level were favored, regression based explanatory power as well as PCA plot separation increased to a r^2 = 0.273 for all genes, r^2 = 0.432 for genes with increasing transcription levels, r^2 = 0.133for genes with decreasing transcription levels and r^2 = 0.294 for significantly changing genes (Supplementary Figures 10 and 11).In order to isolate genes in which a strong relationship between change in epigenetic landscape and change in expression exists cut regression analysis was employed. Progressively smaller subsets of genes that experienced an increased degree of change in the expression level change were considered. In general, the r^2 value began to rapidly increase only within the top 20% of the most changing genes with drastic differences between genes with increasing and with decreasing expression levels (Figure 2A).

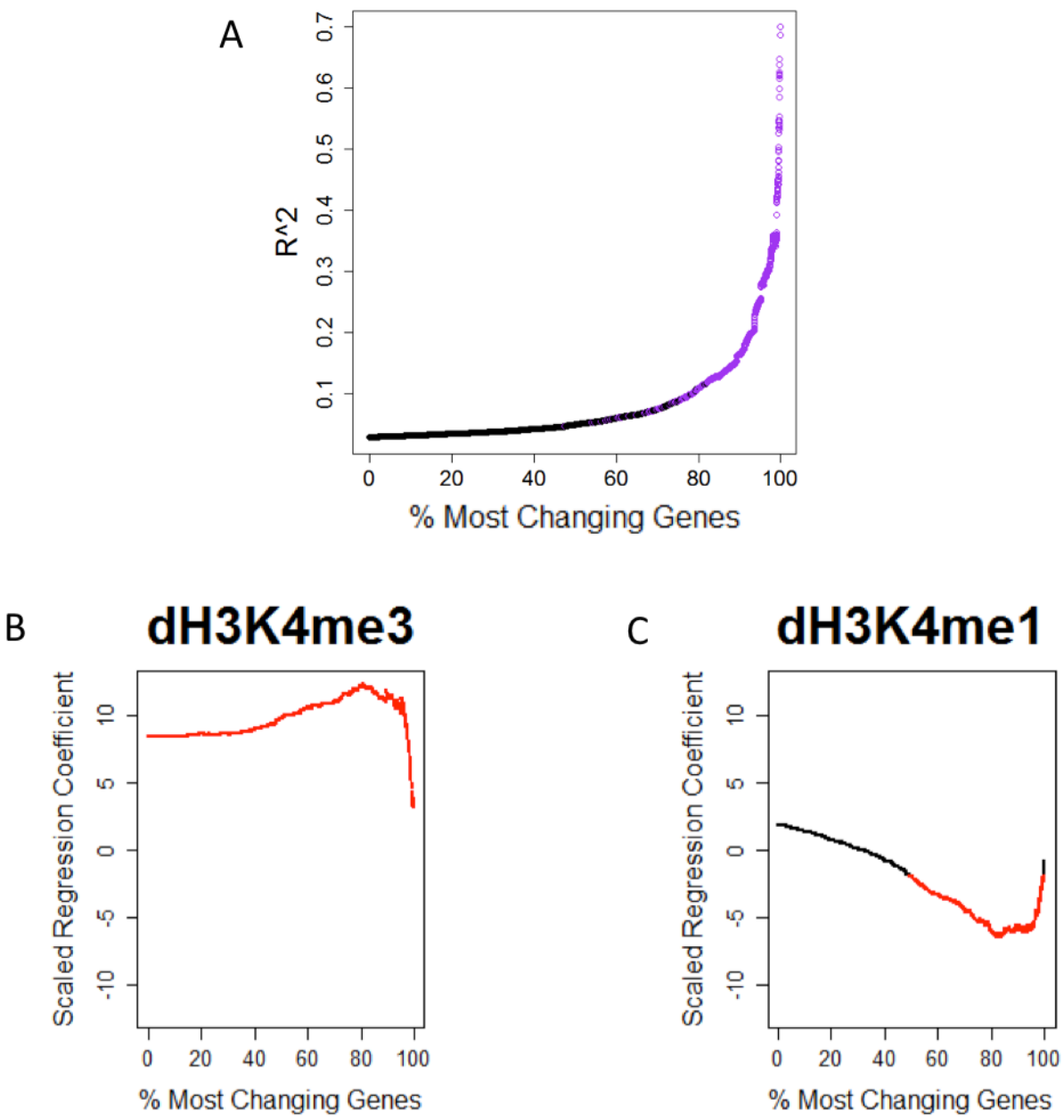


Fig.2. Regression analysis of the change in epigenetic landscape in relation to change in gene transcription level (multiple gene subsets). A) Ability to explain change in gene transcription level per gene subset. B) Scaled regression coefficients for epigenetic factor H3K4me3. Red indicates statistical significance (p-val<= 0.05). C) Scaled regression coefficient for epigenetic factor H3K4me1. Black indicates lack of statistical significance (P-val> 0.05).

When considering only genes which increased in expression level this occurs within the top 40% however, when considering only genes which decrease in expression level it did not occur until only genes that experienced the most extreme decreases in the expression level were considered (Supplementary Figure 12).

By tracking regression coefficients through the cut regression analysis conversion of H3K4me1 -> H3K4me3 were shown to coincide with an increase in the expression level (Supplementary Figures 13, 14 and 15). Increases in occupancy strength for GATA1, TAL1 and POLR2A were found to be associated with an increase in expression level. In contrast, increases in occupancy for GATA2 and to a limited extent CTCF and H3K27me3 occupancy were associated with decreases in gene expression levels. When considering only genes, which increase in transcription level, POLR2A does not considerably associate with an increased expression level indicating its presence is only useful for differentiating between genes which increase in transcription level and genes which decrease in transcription level but not to determine quantitative changes in transcription level. Higher variability in regression coefficients amongst genes which decreased in transcription level, and greater explanatory power amongst genes which increased in transcription level were noted. These results suggest that the examined set of epigenetic factors better explained the process of increasing expression levels rather than decreasing expression levels.

Epigenetic factors affect gene expression via non-synergistic interaction.

Using linear regression analysis two sets of epigenetic factors, activating factors (histones H3K4me1 and H3K4me3, transcription factor POLR2A and DHS) and repressing factors (histones H3K27me3 and H3K9me3 and transcription factor GATA2) were identified(Supplementary Figure16). The linear regression models, which also accounted for interactions between the individual epigenetic

factors, revealed non-synergistic effects between activating factors as well as repressing factors. Multiple interaction terms consisting of two activating factors were associated with lower levels of expression than would be expected assuming independence. Likewise, several interaction terms consisting of two repressing factors were associated with higher levels of expression than expected from two independent repressing factors (Figure 3, Supplementary Figure 17).

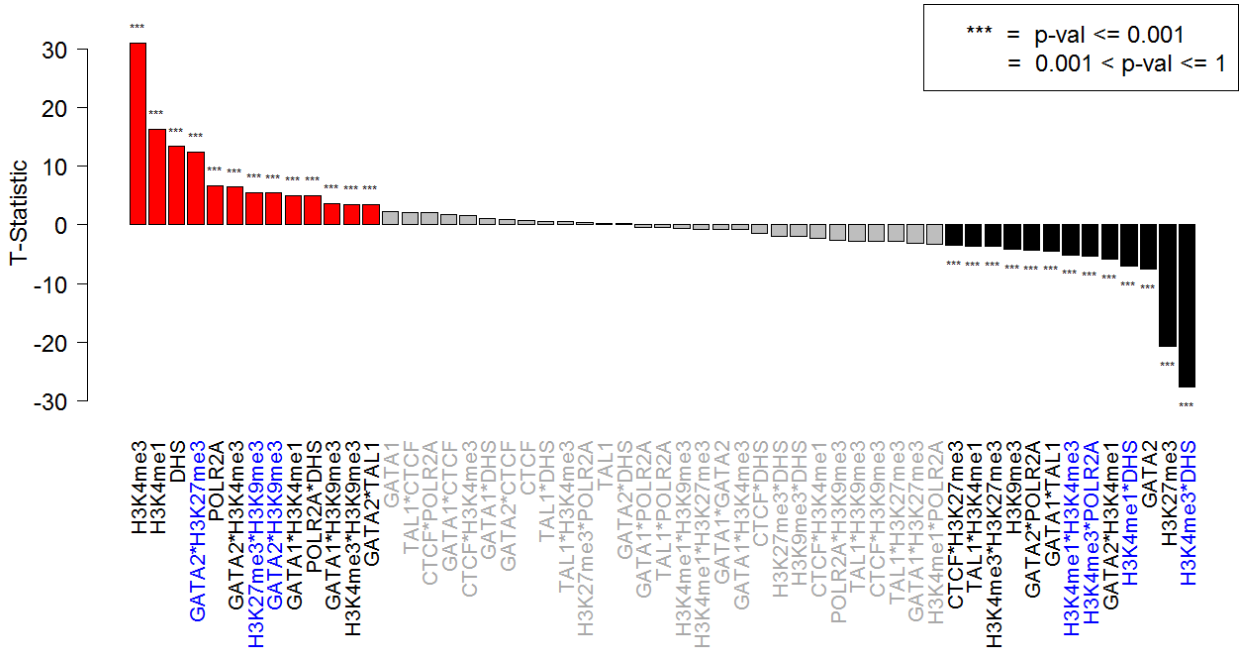


Fig.3. Multiple linearregression model (scaled regression coefficients) with pairwise interaction terms relating epigenetic landscape near the TSS to transcription level for 15,960 mouse genes in the pre differentiation state. Model terms with T-statistics in red indicate a statistically significant association between an increase in occupancy/modification of the epigenetic factor(s) and increase in gene transcription level. Model terms with T-statistics in black indicate a statistically significant association between an increase in occupancy/modification of the epigenetic factor(s) and decrease in gene transcription level. Regression terms their associated T-statistics colored grey were not statistically significant. Interaction terms labeled blue consist of two activating factors yet were associated with a

decrease in transcription level or two repressing factors that were associated with an increase in transcription level.

Additional observations indicated that epigenetic factors work in a redundant manner to influence gene expression. The linear regression models for silent and expressed genes were separately constructed. While limited r^2 values were observed, both models were found to capture statistically significant trends. For example, activating mark H3K4me3 was associated with higher levels of expression in both models (Supplementary Table 3), whereas activating marks H3K4me1 and DHS were only associated with higher levels of expression amongst "silent" genes. TAL1 and GATA1 were associated with higher levels of expression amongst expressed genes, while repressive mark H3K27me3 was only associated with lower levels of expression amongst expressed genes. GATA2 was associated with higher levels of expression amongst silent genes and lower levels of expression amongst expressed genes. Taken together, these results indicate that multiple activating factors can influence an otherwise silent gene but few will have an additional effect on an already active gene. Likewise, multiple repressive factors can suppress an otherwise active gene but additional repressive marks may not further reduce perceived expression levels of an already repressed gene.

DISCUSSION

Previous work by the ENCODE consortia assayed the occupancy profiles of 119 DNA binding proteins in 72 cell lines, the modification profiles of 12 histone modifications over 46 cell lines and DNAse hypersensitivity within 125 cell types. They found histone modifications H3k27ac, H2k9ac, H3k4me3, H3k4me2, H2k27me3 and H3k9me3 to be sufficient to explain gene transcription level variation of most genes .[207]

In this study selected histone modifications, transcription factors and DHS were assayed in the G1E, G1E-ER4 model system for erythroid differentiation. General histone modifications H3k4me3, H3k27me3 and H3k9me3 were studied. The conversion of H3k4me1 to H3k4me3 upon switching from an epigenetic landscape, which is non permissive to gene transcription to one which is permissive was also investigated. In addition, erythroid specific transcription factors, GATA1, GATA2, TAL1 and the insulator CTCF were also studied.

Interestingly, the ability to differentiate between silent and actively transcribed genes based on the epigenetic landscape near the TSS was noted. , However, it was not possible to determine a gene's quantitative transcription levels beyond this distinction. This result implies the need for a permissive landscape near the TSS to achieve transcription while regulation of the specific level of transcription appears to involve distal factors.

Ideally utilizing epigenetic information from cis regulatory modules (CRMs) could help improve the ability to predict transcription levels. Unfortunately mapping the relationships between CRMs and promoters remains an unsolved issue. Chromatin conformation capture technologies could offer insight into these long distance relationships and enable further research.

Controversy exists amongst previous studies relating change in epigenetic landscape to change in transcription level. Some studies indicate a strong relationship with specific examples whereas other studies indicate a general lack of relationship. This study indicates both trends. In general change in epigenetic landscape does not relate well to change in transcription levels. However, in cases of extreme transcription level change (silent to actively transcribed and vice versa) a relationship is found with strong examples. This finding further implies that epigenetic landscape near the TSS is responsible for determining a gene to be permissive or non permissive to transcription whereas fine modification of transcription level is accomplished in a more complex way.

Epigenetic features are not independent. Histone modifications readily interconvert. For example H3k4me1 becomes H3k4me3 upon addition of two methyl groups. Additionally, transcription factors have been shown to preferentially occupy DHS sights and many pair wise interactions have been identified by the encode consortia. Thus, indications of redundant function come as no surprise and again strengthen the hypothesis that epigenetic landscape near the TSS determines a permissive verse non permissive landscape with respect to gene transcription level. As, addition of permissive marks to an already permissive landscape does not correlate with increased transcription level while addition of repressive marks to an otherwise permissive landscape does correlate with reduced expression level and vice versa.

MATERIALS AND METHODS

Datasets Analyzed. The quantitative presence of ten epigenetic features around the transcription start sites (TSS) for 15,960 mouse genes in two cell lines were analyzed. The first cell line, henceforth referred to as G1E [274] represents committed, proliferating but undifferentiated erythroid progenitor cells. The second cell line is a sub-clone of the G1E cells, G1E-ER4 cells treated with estradiol for 24 hrs (G1E-ER4+E2), and represents differentiated erythroblasts [259,263,274]. The ten epigenetic features are RNA polymerase II (specifically the large subunit RPB1: POLR2A), DNAse hypersensitive sites (DHS), four sequence-specific transcription factors (GATA1, GATA2, TAL1, CTCF) and four histone modifications (H3K4me1, H3K4me3, H3K27me3, H3K9me3). These epigenetic data, determined in two to five replicates, were extracted from the genome-wide ChIP-seq determinations published by Wu et.al. [263]with additional datasets representing CTCF and POLR2A submitted in support of this analysis. My analysis was limited to a 4 kb window centered on the TSS. The tag counts in G1E and G1E-ER4+E2 were normalized by the total number of reads mapped to account for the difference in sequencing depth. This produced an estimate $x_{i,j}$ for each epigenetic feature (*j)* in cell line (*i)*, all analyses on levels of epigenetic features were performed on $x_{i,j}$. The tag counts for transcription factors were limited to the DNA segments called as peaks in the MACS [275] analysis of the ChIP-seq datasets. The peaks were limited to the 4 kb window around the TSS. Changes in levels of epigenetic features between G1E and G1E-ER4+E2 cells were expressed as $log_2(x_{i,G1E\text{-}ER+E2}) - log_2(x_{i,G1E})$following magnitude vs amplitude normalization (M vs A).

The levels of expression for 15,960 mouse genes during this process of differentiation were obtained from the Affymetrix Gene Chip microarray data on G1E-ER4+E2 cells prior to and during a 30 hr time course of estradiol treatment [262]. The transcription levels at 0hr and 30hr for each probe-set was

normalized by Robust Multi-Array Average (RMA) [276], log transformed and averaged across replicates to obtain estimates of transcription level, referred to as $\log_2(y_0)$ and $\log_2(y_{30})$. The difference $\log_2(y_{30})$-$\log_2(y_0)$ provides an estimate of the change in gene expression during this differentiation process. Expression profiles over the time course also were used to place genes into response categories (genes which do not change, experience slight change, increase or decrease with respect to transcription level). G1E cells are similar to untreated G1E-ER4+E2 cells regarding morphology, phenotype and transcriptome [263], It is thus reasonable to use epigenetic features in G1E cells ($x_{i,G1E}$) to predict the expression levels in untreated G1E-ER4+E2 cells (y_0).Thus the pre differentiation condition is represented by transcription levels in the G1E-ER4+E2 cell line (0hrs after induction) and epigenetic signal in the G1E cell line while the post differentiation condition was represented by transcription levels in the G1E-ER4+E2 (30hrs after induction) and epigenetic signal in the G1E-ER4+E2 cell line (24hrs after induction).

Microarray data generation and analysis. Gene expression data collected during a time-course of differentiation (0, 3, 7, 14, 21 and 30 hours after induction) of G1E-ER4+E2 cells were obtained from previous work [262], Hybridization signals on the Gene Chip Mouse Genome 430 2.0 Arrays (MOE430v2) from Affymetrix for three biological replicates of RNA from G1E-ER4+E2 cells treated with 0.1µM beta-estradiol were normalized by RMA methods [276,277]. Signals for each probe set were mapped back to known RefSeq [278,279] and Ensembl genes [280]. Probe-sets within the body of the gene, which showed opposite expression patterns when compared to other probes were removed as were probe-sets showing three standard deviations greater than the mean log2 transformed signal intensity. A mean expression level measurement at each time point after induction; Y_i;i = (0, 3, 7, 14, 21, 30) the expression change; $\Delta Y = \log_2(Y_{30}/ Y_0)$ were computed.

Additionally, genes were partitioned based on their expression behavior. Pair-wise t-tests between each time point ($Y_{3, 7, 14, 21, 30}$) and Y_0 were computed with the LIMMA package in R [281] using the BH-FDR adjustment for multiple testing which generates an estimate of the false discovery rate (FDR) [28](REF). The genes that passed an FDR threshold of 0.001 for at least one time point were considered responsive. Responsive genes were sub-grouped according to whether they increase or decrease in transcription level over the time course using the Ordered Restricted Inference for Ordered Gene Expression (ORIOGEN) 5 package [282]. This package allowed us to pre-define a collection of typical profiles and to partition genes into clusters that matches these expression profiles. Two profiles were defined: genes, which generally increase in transcription level; and genes, which generally decrease in transcription level. Briefly, for a given gene, a goodness-of-fit statistic was computed under each of the profiles. The profile with the largest goodness-of-fit statistic was tested against the null hypothesis that there is no change in expression level during the time course. This result was bootstrapped 10,000 times for each gene and statistically significant genes were identified as either genes, which decrease or increase in transcription level. Genes that did not pass the FDR threshold and showed fold changes smaller than 1.1 at ($Y_{3, 7, 14, 21, 30}$) when compared to Y_0 were categorized as genes which experience no change in transcription level. Genes,, which were not classified as either decreasing, increasing or not changing with respect to transcription level were classified as experiencing a small change in transcription level.

Genome wide Normalization of Epigenomic Signatures. This was based on published approaches [262,263] for the normalization of ChIP-seq or ChIP-chip data between different experiments. ChIP-Seq binding data for four histone modifications (H3K4me1, H3K4me3, H3K27me3 and H3K9me3), four transcription factors (GATA1, GATA2, TAL1, CTCF) and DHS in G1E and G1E-ER4+E2 cell lines were obtained from Wu

et al [263]. Additionally, POLR2A in both cell conditions was also sequenced. The genome was tiled in 10bp increments. The total number of reads which overlapped each disjoint window were counted. These read counts / 10bp window were denoted as *g1e* or *er4* for the data in the cell lines G1E and G1E-ER4+E2, respectively. The g1e and er4 data sets were normalized by the total number of mapped reads in millions per cell line, to give $G1E_{rpm}$ *and* $ER4_{rpm}$respectively. Magnitude (*M*) and amplitude (*A*) were calculated based on the log2(normalized counts) in 100,000 randomly selected bins from a chromosome: $M= log_2(ER4_{rpm} /G1E_{rpm})$ and $A= 0.5\times(log_2(ER4_{rpm}) + log_2(G1E_{rpm}))$. The mean of the magnitude was estimated by fitting a lowess regression curve of *M* versus *A* using the lowess function from the stats package in R [27]. The fitted mean was then subtracted from *M*, which resulted in the first adjusted *M* (*adjM*):*adjM= M-Mean*, where *Mean= loess(M ~ A)*. The variance of the magnitude was estimated by fitting a loess regression of $(adjM)^2$ versus *A* and then taking the square root. Dividing the *adjM* by the estimated variances results in the secondly adjusted *M* (*adjM'*): *adjM'= adjM/standard error*, where *standard error= squareroot(loess(*$adjM^2$ *~ A)).*The value of *adjM'*was the normalized ratio between the two cell lines and was used to represent the change in the level of epigenetic modifications after GATA1 restoration.

Epigenetic signals $(log_2(G1E_{rpm}), log_2(ER4_{rpm}))$ and associated changes of all 10 epigenetic factors was computed in a 4kb window centered on the TSS of 15,960 genes. The epigenetic level signature of the gene, X = (X1…X10)', is a 10-dimensional vector, each entry of which is formed by averaging the $(log_2(G1E_{rpm})$and $log_2(ER4_{rpm})$independently for each of the epigenetic signal tracks over the 10bp windows constituting the TSS neighborhood. Similarly, the epigenetic change signature of the gene, ΔX = (ΔX1…ΔX10)', is a 10-dimensional vector, each entry of which is formed by averaging the adjusted *M* (*adjM'*) in the gene neighborhood. All computations were implemented using the MASS (36) and limma [281] packages in R.

Principal Component Analysis. All PCA analysis was carried out in R using the princomp function in the stats package [283]. PCA was used to extract characteristic patterns from the 10 epigenetic profiles of 15,960 genes (X) in the G1E and G1E-ER4+E2 systems, and their respective changes (ΔX). In separate analysis, ΔXwas also weighted for weighted PCA on a gene by gene basis by the square change in transcription level of the gene.

PCA orthogonally transforms the data such that the projection of the data on first principal component (coordinate) has the greatest variance. After centering and standardizing the features and changes such that they have a mean = 0 and standard deviation = 1, the matrix of epigenetic features (X) and changes (ΔX) was decomposed by Eigen value decomposition (EVD). The PCA output was summarized using bi-plots that showed projections of the genes in a plane spanned by the PCA directions with arrows that indicate the strength of the epigenetic features in that direction.

Mixture Model. A Gaussian Mixture Model (GMM) was fitted to expression data for the G1E cell line as well as theG1E-ER4+E2 cell line 30hrs after induction of differentiation. The fit was accomplished via the Estimation Maximization algorithm implemented in the Mixtools package in R. I considered GMM's containing between two and ten Gaussian distributions. Model fit was evaluated via Bayesian Information Criterion (BIC), which addresses the over-fitting problem by considering not only model fit via the likelihood function but also the number of parameters in the model. For each set of expression data the GMM with the lowest associated BIC were chosen. All expression data in my data set conformed to a bi-modal distribution. Thus, cutoff for calling a gene as silent or expressed was determined by examining the GMM curve and finding a global minimum, which lied between the global maxima (one for each mode).

Assigning chromatin states. Genome wide histone data was considered for five modifications: H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3. Segmentation classes were generated by a hidden Markov model (HMM) based analysis of this data [263,284]. Briefly, this approach identifies spatially related combinatorial patterns of Histone modifications by identifying biologically meaningful regions of the genome. Each pattern can be referred to as a chromatin state. Nine diferent states were identified. State 1 was enriched for H3K36me3 and marks actively transcribed regions. State 2 was enriched for H3K36me3 and H3K4me1, marking potential enhancers that are predisposed to be transcribed. State 3 was enriched for H3k4me1 and generally denotes potential enhancers. State 4 was enriched for H3K4me1 and H3K4me3 and is commonly found around TSS. State 5 was enriched for H3k4me3 and generally denotes TSS. State 6 was enriched for H3K9me3 and is associated with repressed chromatin. State 7 lacked any strong signals and was considered the "dead" state. State 8 was enriched for H3K27me3 and denotes repressed chromatin. Finally, state 9 was enriched for both H3k4me1 and H3k27me3, this state is considered to be a "bivalent" state as both active and repressive marks are present. State number, however was irrelevant for the current analysis and is only presented here for clarity. Each genomic location is assigned a chromatin state based on the HMM. Thus a gene's chromatin state was determined to be the chromatin state assigned to the exact location of the TSS base pair (bp).

Linear Discriminate Analysis. Linear Discriminate Analysis (LDA) is a supervised dimension reduction method that identifies leading separating directions for labeled data. It also functions as a supervised classification method as it identifies linear combinations in a feature space that best separate pre-defined groups of data points. All LDA analyses were performed using the LDA function in the MASS package in R [283]Briefly, genes were categorized as either silent or expressed. LDA was implemented on

a 10-dimensional feature space consisting of epigenetic signature levels (X) in G1E or G1E-ER4+E2 cell lines, and two transcription signature levels (Y) in G1E to discriminate between these two gene categories (silent and expressed). The epigenetic factors that were considered in this analysis are histone modifications H3K4me1, H3K4me3, H3K9me3 and H3K27me3, transcription factors GATA1, GATA2, CTCF, TAL1 and POLR2A as well as DHS. The between group variance-covariance matrix of the ten epigenetic features or changes was normalized against the within groups variance-covariance matrix. A spectral decomposition of the resulting matrix produces a set of orthogonal eigenvectors, each a combination of epigenetic features, measuring the separability of the groups. The LDA output was then summarized by coloring genes on PCA bi-plots and displaying the discriminate equation coefficients in bar plots.

Multiple Linear Regression. All regression models were constructed using the lm function within the R statistical software. First, gene expression levels ($Y_{0 \text{ or } 30 \text{ Hour}}$) were studied as a function of the levels of epigenomic signatures X. Second, the analysis was repeated in both cell lines considering only genes called silent or genes called expressed by the above mixture model. Third, significant differences between the silent an expressed regression equations in both cell lines were determined. This was done by introducing a predictor variable indicating whether a gene was labeled silent (0) or expressed (1) by the mixture model. Interaction terms between this new predictor variable and the previous predictor matrix were then considered. Thus, the full predictor matrix for this analysis consisted of 21 predictors, 10 representing the epigenetic landscape around the TSS, one representing whether the gene was expressed or silent, and 10 interaction terms relating a binary call for expressed or silent and the 10 epigenetic factors. Fourth, expression change ΔY (t = 30 hours vs t = 0 hours) was considered as a function of epigenomic signatures changes ΔX. Fifth, since the majority of genes showed little change in

expression, a weighting scheme was employed to reduce the effect of these genes on model selection and fit. Each gene was weighed by the square of the change in gene transcription level between time 0 and time 30. i.e. ΔY^2. These weights were then used in the least square and model evaluation calculations. All genes were examined for all five regressions . In total 15,960 genes were included in this study. All features were centered and standardized to have mean signal = 0 and standard deviation = 1 prior to fitting and model selection. Every regression model was evaluated via R^2, which served as a measure of the % variance in the response variable that was explained by the predictor variables. All predictor variables were evaluated via their associated t-statistics in the lm package output.

Two Step Model. The two step model is reasonably similar to procedure outlined in Dong, et. Al [207]. It was applied to both data sets (before and after differentiation of the cell lines). LDA was first applied to predict a gene as expressed or silent. Based on this prediction a regression model was built relating epigenetic landscape X to expression level Y. This model was build considering only expressed or only silent genes depending on the outcome of the LDA prediction following standard Leave One Out Cross Validation (LOOCV). R^2 values were computed to evaluate the model. An R^2 value was also computed considering only the LDA predictions, in this case the assigned expression level was determined by averaging the expression level of either expressed or silent genes depending on LDA predictions. Again, standard LOOCV was applied.

Data Subsets. In some studies relating ΔX to ΔY, subsets of the 15,960 genes were analyzed. For example, only genes, which change in transcription level, only genes which increase in transcription level or only genes which decrease in transcription level were considered in different studies. These subsets of genes were analyzed by PCA, regression analysis, weighted PCA, and Weighted regression analysis. In PCA only epigenomic signature changes ΔX were considered, while in the regression

analysis ΔY (t = 30 hours vs t = 0 hours) expression changes were related with ΔX epigenomic signature changes.

By relating ΔX to ΔY with regression analysis the trends present in the full data set and in data sets containing only increasing or only decreasing genes were examined in a greater detail. In the full data set two ranked lists of genes were created, one containing mostly genes which decrease in transcription level and one containing genes which mostly increase in transcription level. Genes were then paired by magnitude change in gene expression. The analysis was started with the full data set and at each step removed one pair of genes. Every time a gene pair was removed a regression model was built to relate ΔX to ΔY. R^2 value was reported as an evaluation of model fit. The same procedure was repeated separately for genes with increasing transcription levels and for genes with decreasing transcription levels. However, only one gene was removed from these data sets at each step as gene pairing was not necessary.

Supplementary Tables

Supplementary Table1. Pair-wise epigenetic factor correlations.

	GATA1_G1E	GATA2_G1E	TAL1_G1E	CTCF_G1E	H3K4me1_G1E	H3K4me3_G1E	H3K27me3_G1E	POLR2A_G1E	H3K9me1_G1E	DHS_G1E	
GATA1_ER4		0.54	0.33	0.4	0.11	0	0.11	0.25	0.33	0.05	GATA1_G1E
GATA2_ER4	0.57		0.43	0.52	0.22	-0.06	0.39	0.41	0.32	0.09	GATA2_G1E
TAL1_ER4	0.66	0.31		0.29	0.3	0.06	0.02	0.21	0.07	0.08	TAL1_G1E
CTCF_ER4	0.35	0.38	0.2		0.34	0.15	0.12	0.33	0.09	0.26	CTCF_G1E
H3K4me1_ER4	0.48	0.26	0.3	0.34		0.53	-0.09	0.34	-0.24	0.62	H3K4me1_G1E
H3K4me3_ER4	0.27	0.09	0.17	0.23	0.58		-0.31	0.44	-0.24	0.81	H3K4me3_G1E
H3K27me3_ER4	0.1	0.29	0.02	0.04	-0.19	-0.4		0.03	0.13	-0.14	H3K27me3_G1E
POLR2A_ER4	0.47	0.37	0.3	0.28	0.39	0.49	-0.1		0.09	0.44	POLR2A_G1E
H3K9me3_ER4	0.11	0.36	0.1	0.03	-0.25	-0.22	0.07	0.05		-0.17	H3K9me3_G1E
DHS_ER4	0.33	0.33	0.2	0.34	0.52	0.54	-0.11	0.34	-0.06		DHS_G1E
	GATA1_ER4	GATA2_ER4	TAL1_ER4	CTCF_ER4	H3K4me1_ER4	H3K4me3_ER4	H3K27me3_ER4	POLR2A_ER4	H3K9me3_ER4	DHS_ER4	

Supplementary Table 2. Correlation coefficients relating epigenetic factor occupancy / modification near the TSS with gene transcription level.

A

Epigenetic Factor	Correlation Coefficient
H3K4me3	0.69
DHS	0.6
H3K4me1	0.53
POLR2A	0.29
CTCF	0.11
TAL1	0.06
GATA1	-0.06
GATA2	-0.16
H3K9me3	-0.31
H3K27me3	-0.34

B

Epigenetic Factor	Correlation Coefficient
H3K4me3	0.7
H3K4me1	0.57
DHS	0.42
POLR2A	0.36
GATA1	0.22
CTCF	0.18
TAL1	0.15
GATA2	-0.03
H3K9me3	-0.3
H3K27me3	-0.4

Supplementary Table 3. Multiple linear regression model relating expression level to epigenetic landscape including a term marking a gene as expressed or silent based on expression profile, scaled regression coefficients. (A) pre differentiation, column labeled; silent_T is the scaled regression coefficient for silent genes, silent_sig represents statistical significance ("***" = pval< 0.001, "**" = pval< 0.01, "*" = pval< 0.05, "." = pval< 0.1), expressed_T and expressed_sig similarly refer to models built only on expressed genes. Combine_t and combine_sig refer to differences in scaled regression coefficients between models built on only silent and only expressed genes. (B) post differentiation.

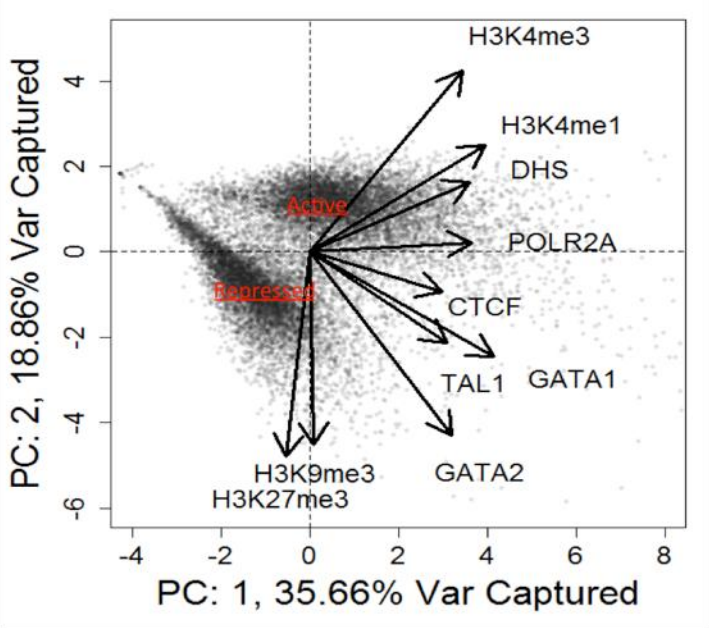
A

	silent_T	silent_sig	expressed_t	expressed_sig	combine_t	combine_sig
GATA1	-1.28		3.23	**	3.5	***
GATA2	7.61	***	-9.54	***	-11.36	***
TAL1	-5.1	***	5.03	***	5.17	***
CTCF	2.86	**	2.3	*	0.58	
H3K4me1	11.87	***	1.7	.	-3.53	***
H3K4me3	1.93	.	15.26	***	4.57	***
H3K27me3	-1.21		-13.71	***	-14.87	***
POLR2A	4.33	***	6.83	***	1.85	.
H3K9me1	-3.72	***	-0.52		0.11	
DHS	9.09	***	-4.61	***	-6.22	***

B

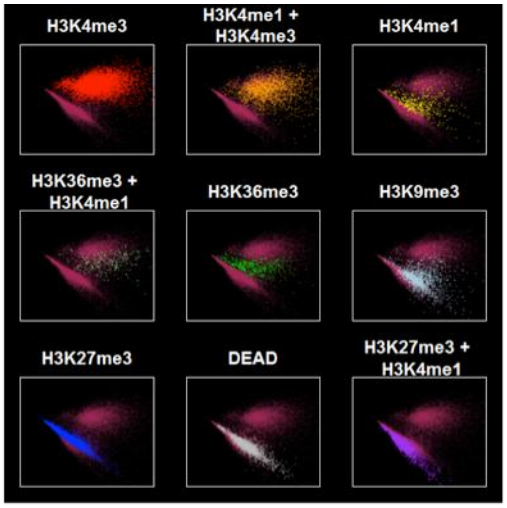
	silent_T	silent_sig	expressed_t	expressed_sig	combine_t	combine_sig
GATA1	-1.94	.	8.21	***	6.38	***
GATA2	13.32	***	-9.47	***	-13.56	***
TAL1	-4.65	***	3.09	**	4.02	***
CTCF	-0.04		0.24		0.18	
H3K4me1	10.18	***	0.91		-3.44	***
H3K4me3	11.42	***	13.71	***	1.33	
H3K27me3	-1.63		-12.04	***	-13.35	***
POLR2A	4.63	***	5.85	***	1.62	
H3K9me3	-5.42	***	-4.12	***	-2.97	**
DHS	11.53	***	-0.88		-5.95	***

Supplementary Figures

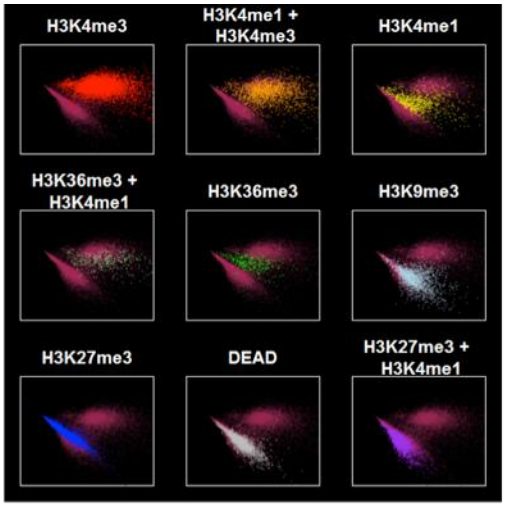


Supplementary Figure 1. Epigenetic landscape near the TSS for 15,960 post differentiation mouse genes.

A



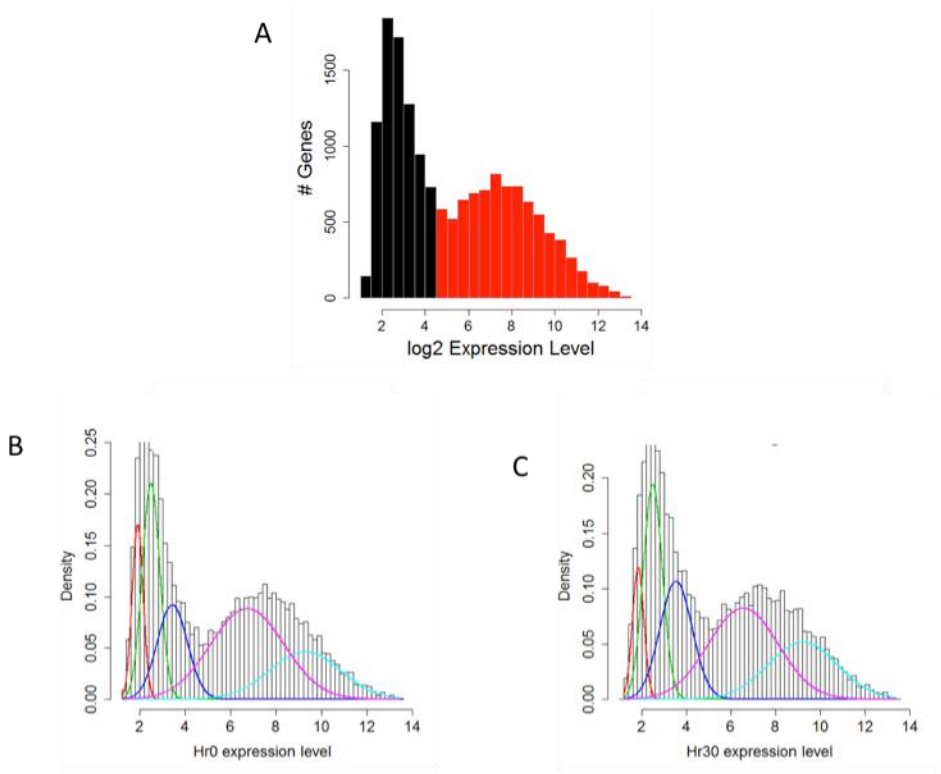
B



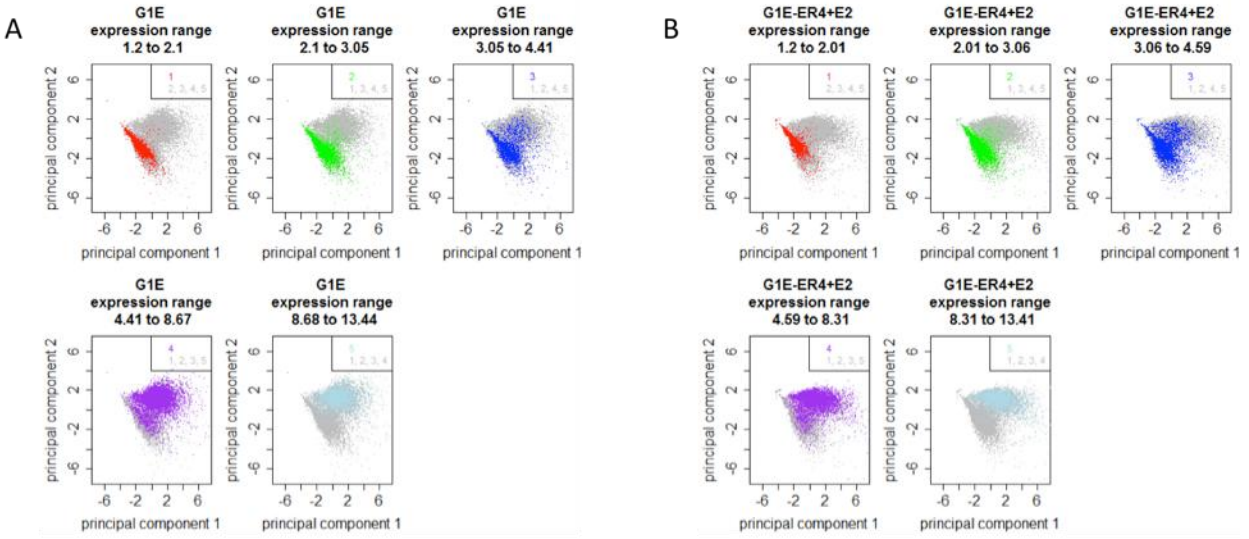
C

State	H3K36me3	H3K4me1	H3K4me3	H3K27me3	H3K9me3
H3K4me3	0.06	0.07	0.98	0.01	0.03
H3K4me1 + H3K4me3	0.28	1	0.97	0.02	0.02
H3K4me1	0.02	0.7	0.01	0.01	0.01
H3K36+H3K4me1	0.94	0.88	0.05	0.01	0.02
H3K36me3	0.81	0.03	0	0	0.02
H3K9me3	0.01	0	0.02	0.01	0.56
H3K27me3	0	0.01	0	0.54	0.02
DEAD	0	0	0	0.01	0.01
H3K27me3 + H3K4me1	0.05	0.81	0.23	0.88	0.1

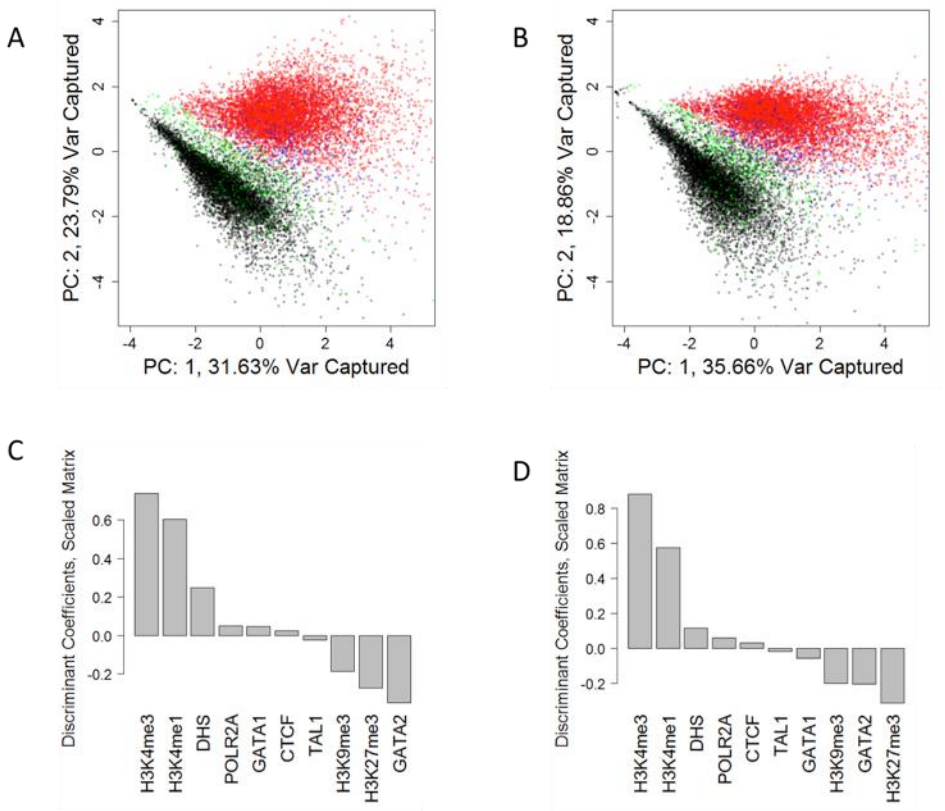
Supplementary Figure 2. Assignment of chromatin state to TSS in relation to epigenetic landscape near the TSS for 15,960 mouse genes.A) pre differentiation. B) Post differentiation. C) ChromHMM state emission probabilities.



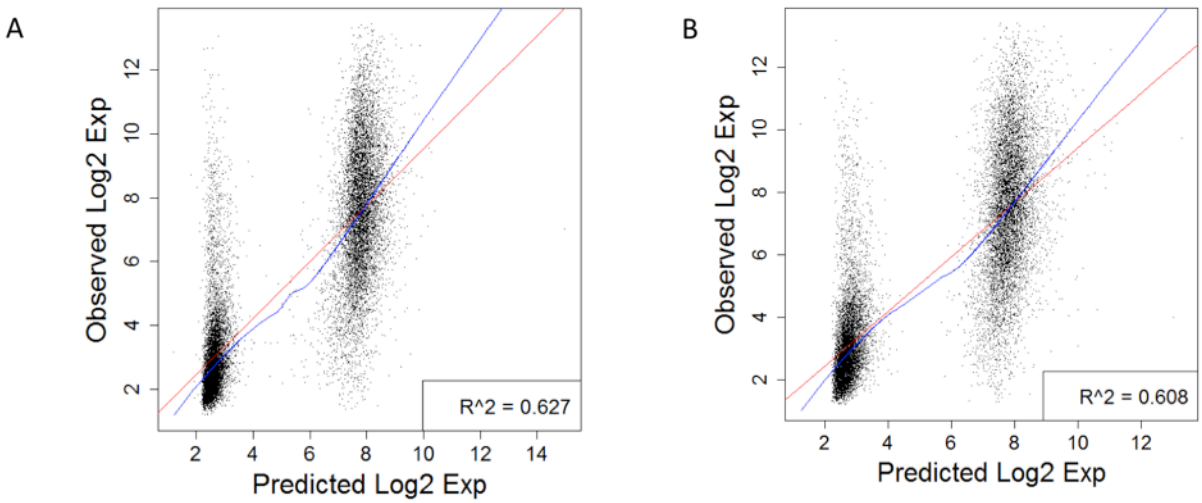
Supplementary Figure 3. Determination of the expressed vs silent cutoff via Gaussian Mixture Modeling (GMM). A) Gene transcription level profile for 15,960 silent (black) and expressed (red) post differentiation mouse genes. B) GMM applied to pre differentiation gene transcription level distribution, colors demark individual Gaussian distributions. C) GMM applied to post differentiation gene transcription level distribution.



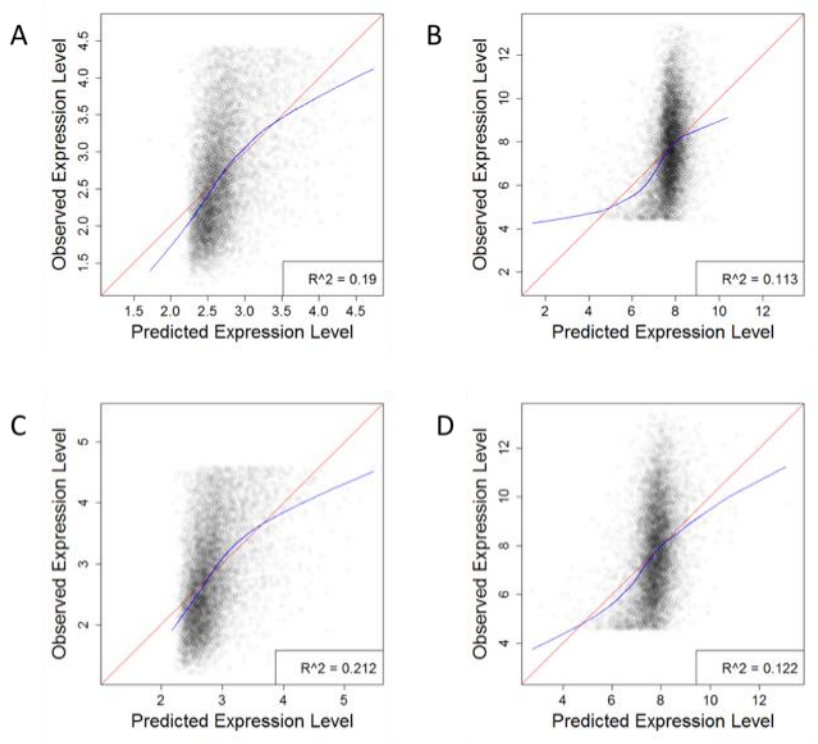
Supplementary Figure 4. GMM based gene assignment mapped to PCA bi-plot based on epigenetic landscape near the TSS for 15,960 mouse genes. A) G1E B) G1E-ER4



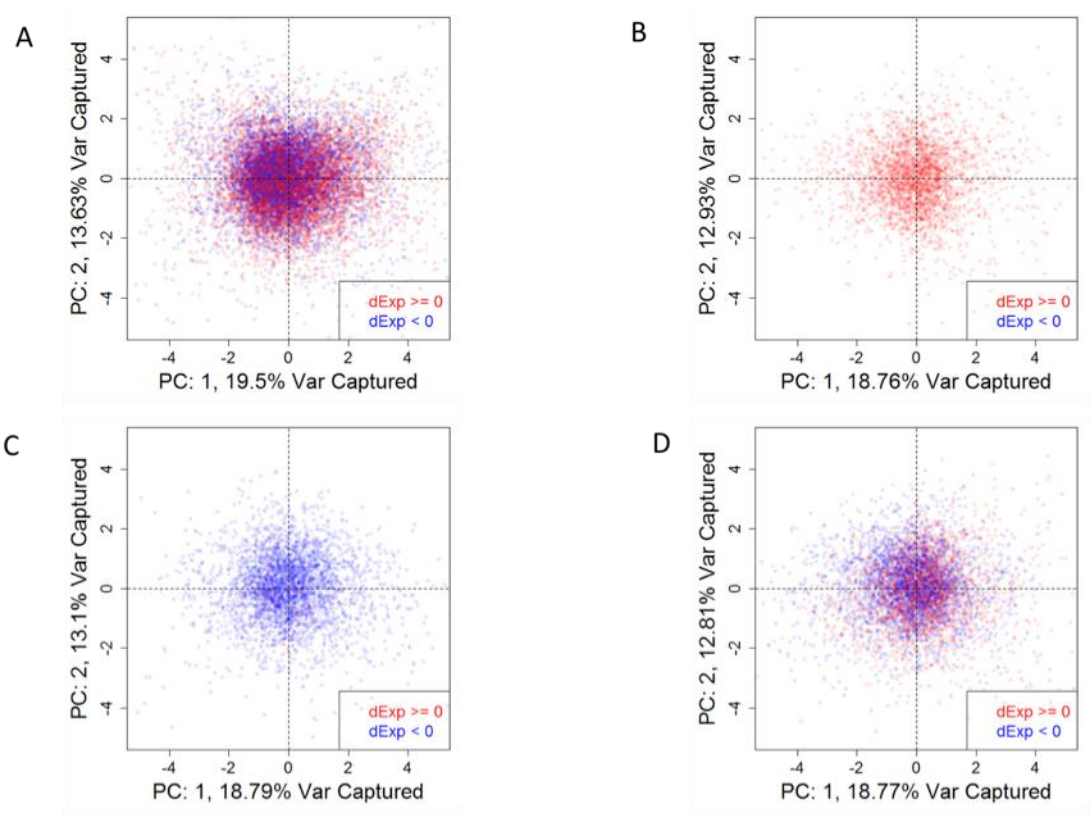
Supplementary Figure 5. Linear discriminate analysis relating gene expression level and epigenetic landscape. (A) Pre differentiation genes mapped on PCA bi-plot, red are expressed genes predicted to be expressed by the LDA model, black are silent genes predicted to be silent by the LDA model. Green are expressed genes which are predicted to be silent by the model, blue are silent genes predicted to be expressed by the LDA model. (B) LDA model evaluation mapped to PCA bi-plot for post differentiation genes. (C) Pre differentiation scaled LDA model coefficients. (D) post differentiation scaled LDA model coefficients.



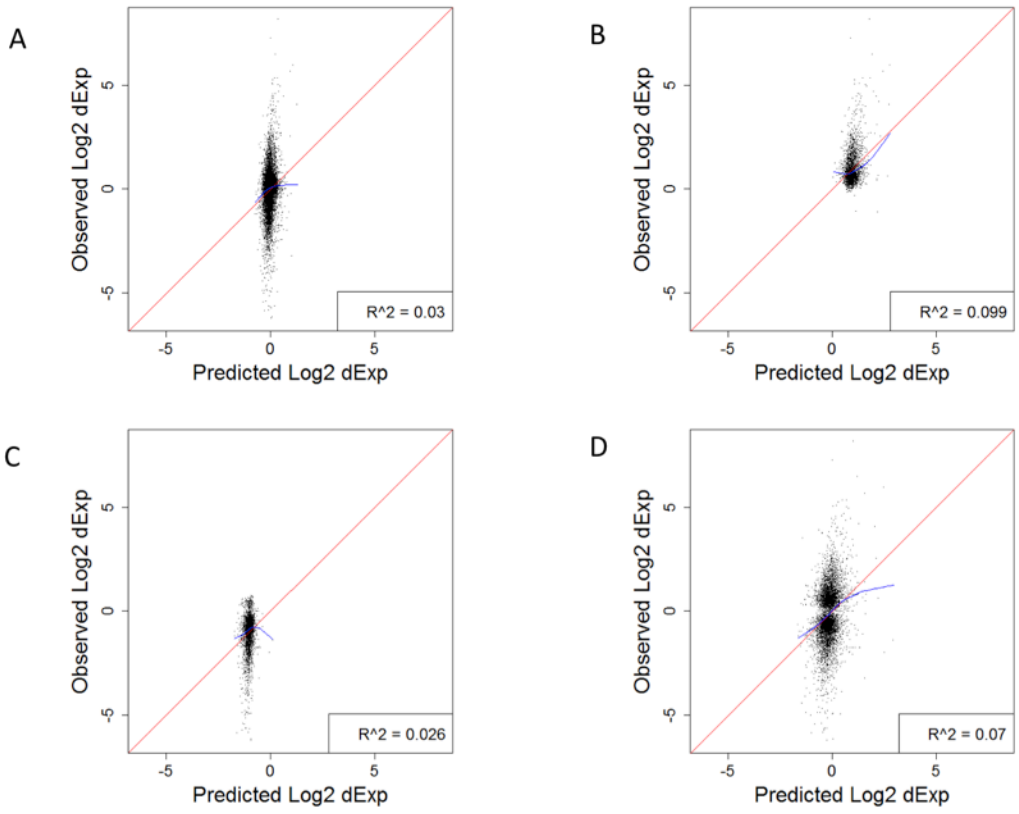
Supplementary Figure 6. Two step regression model relating epigenetic landscape near the TSS for 15,960 mouse genes to transcription level. A) Pre differentiation. B) Post differentiation.



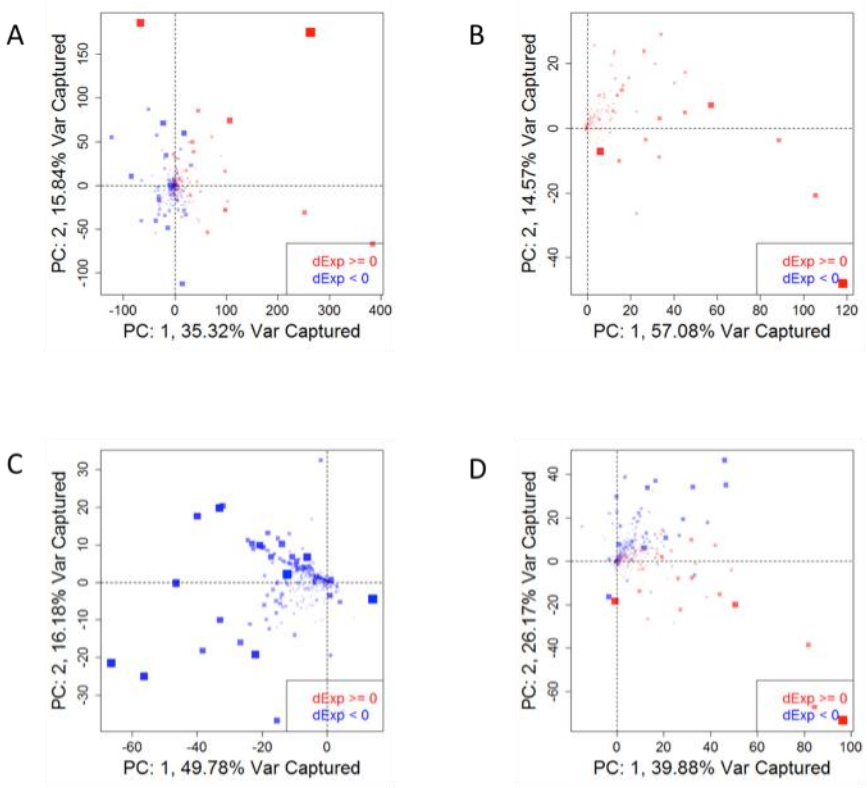
Supplementary Figure 7. Multiple linear regression models built on subsets of 15,960 mouse genes. A) Silent pre differentiation genes. B) Expressed pre differentiation genes. C) Silent post differentiation genes. D) Expressed post differentiation genes.



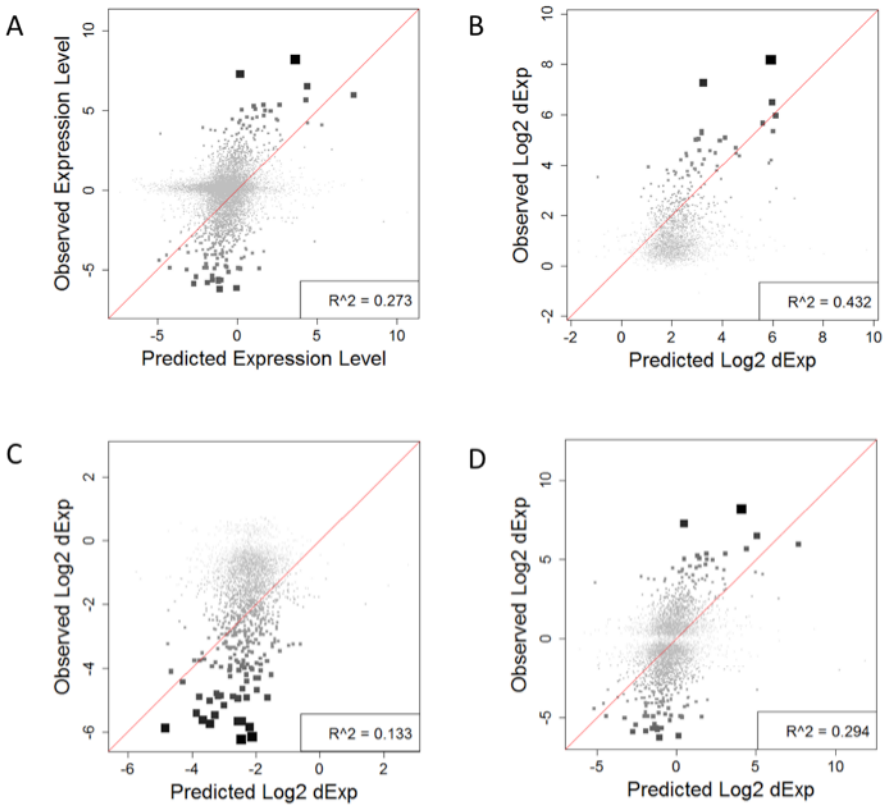
Supplementary Figure 8. PCA bi-plot relating change in epigenetic landscape near the TSS with change in transcription level for 15,960 mouse genes. A) All genes. B) Only genes which increase in transcription level. C) Only genes which decrease in transcription level. D) Only genes which significantly change in transcription level.



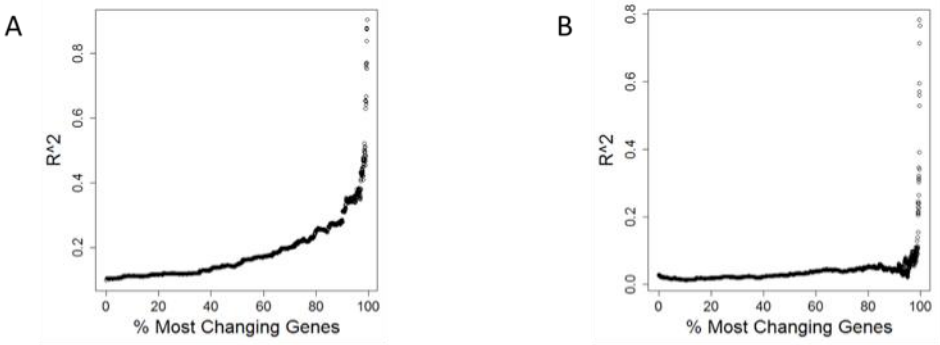
Supplementary Figure 9. Multiple linear regression analysis relating change in epigenetic landscape near the TSS with change in transcription level for 15,960 mouse genes. A) All genes. B) Only genes which increase in transcription level. C) Only genes which decrease in transcription level. D) Only genes which significantly change in transcription level.



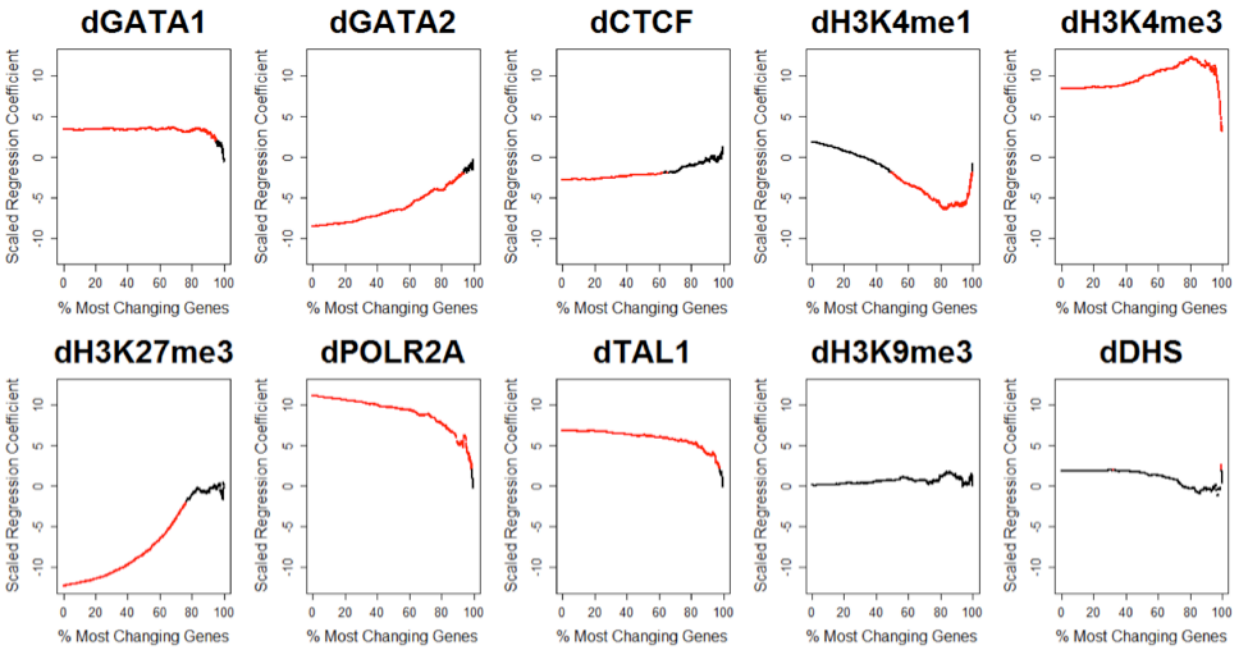
Supplementary Figure 10. PCA bi-plot weighted by squared change in transcription level relating change in epigenetic landscape near the TSS with change in transcription level for 15,960 mouse genes. A) All genes. B) Only genes which increase in transcription level. C) Only genes which decrease in transcription level. D) Only genes which significantly change in transcription level.



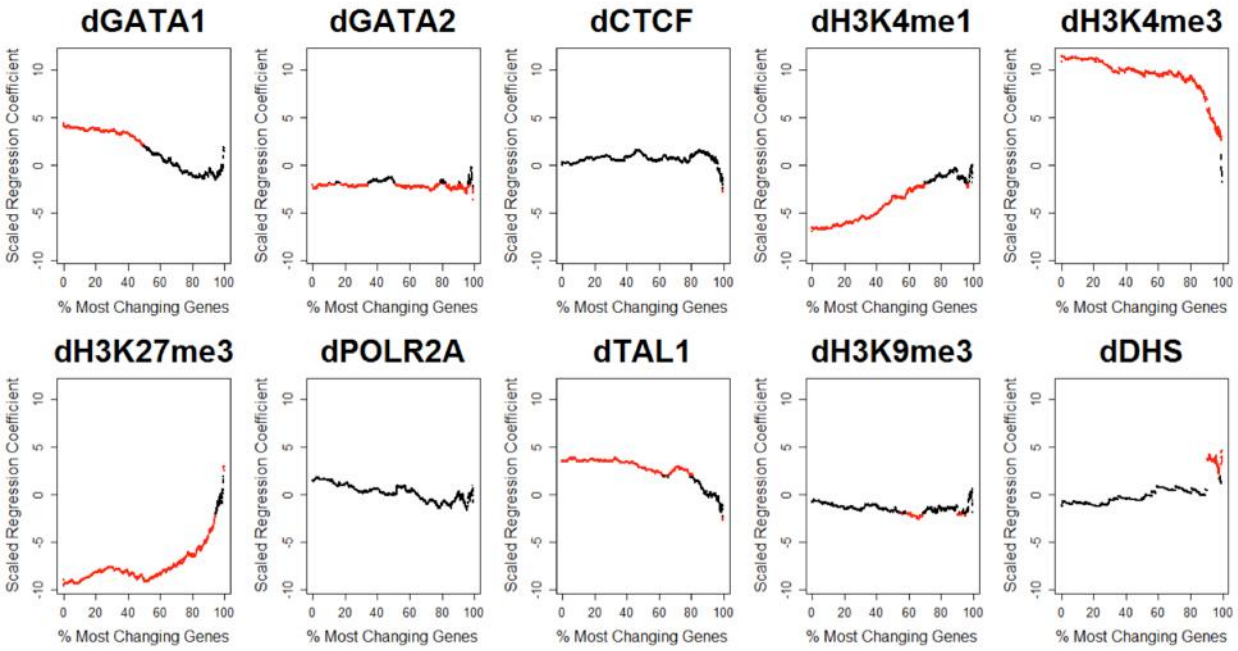
Supplementary Figure 11. Multiple linear regression analysis weighted by squared change in transcription level relating change in epigenetic landscape near the TSS with change in transcription level for 15,960 mouse genes. A) All genes. B) Only genes which increase in transcription level. C) Only genes which decrease in transcription level. D) Only genes which significantly change in transcription level.



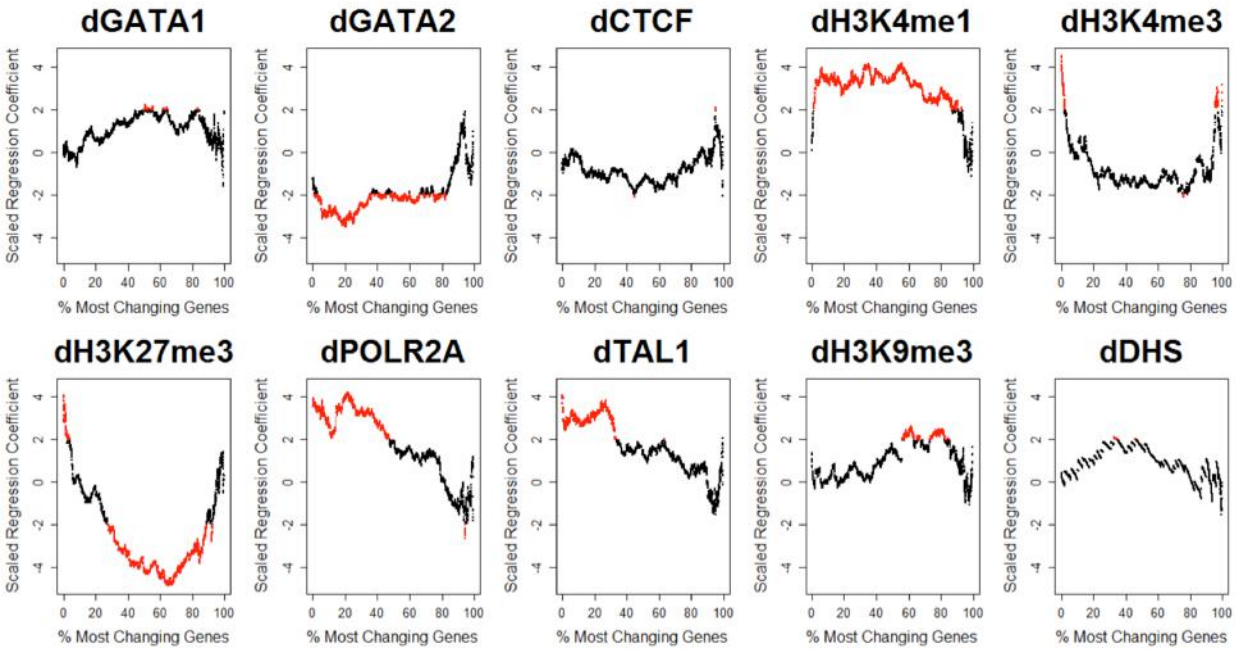
Supplementary Figure 12. Multiple linear regression analysis (multiple gene subsets). A) Only genes which increase in transcription level. B) Only genes which decrease in transcription level.



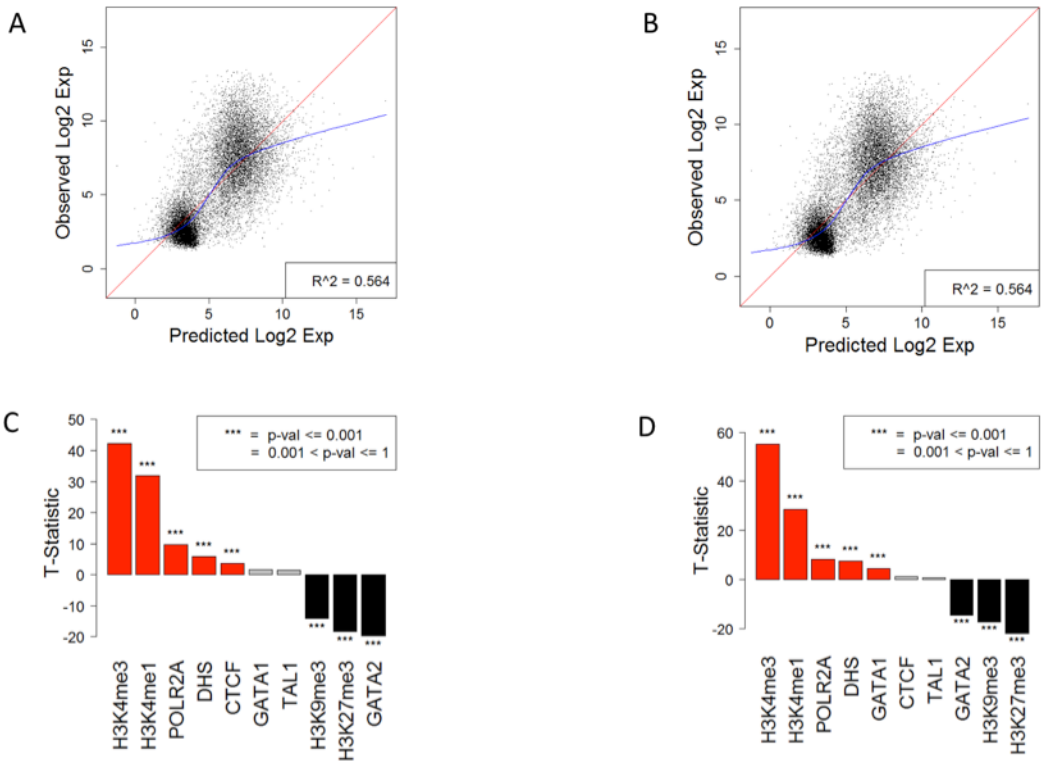
Supplementary Figure 13. Scaled regression coefficients, Multiple linear regression analysis (multiple gene subsets).



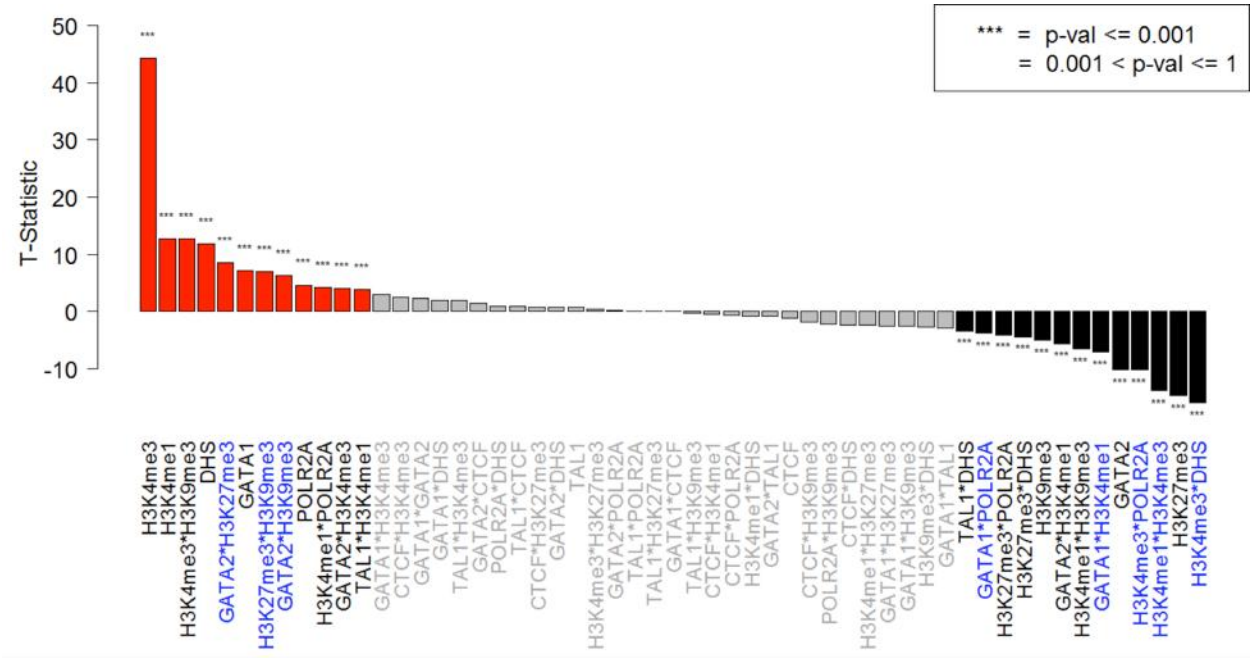
Supplementary Figure 14. Scaled regression coefficients, Multiple linear regression analysis for genes which increase in transcription level (multiple gene subsets).



Supplementary Figure 15. Scaled regression coefficients, multiple linear regression analysis for genes which decrease in transcription level (multiple gene subsets).



Supplementary Figure 16. Multiple linear regression analysis relating epigenetic landscape near the TSS to transcription level for 15,960 mouse genes. A) Pre differentiation. B) Post differentiation. C) Scaled regression coefficients, pre differentiation. D) Scaled regression coefficients, post differentiation.



Supplementary Figure 17. Regression model (scaled regression coefficients) with pairwise interaction terms relating epigenetic landscape near the TSS to transcription level for 15,960 mouse genes in the post differentiation state.

References

1. Scanlan, P. D. & Marchesi, J. R. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J.***2,** 1183–1193 (2008).
2. Chiller, K., Selkin, B. A. & Murakawa, G. J. Skin microflora and bacterial infections of the skin. *J. Investig. Dermatol. Symp. Proc. Soc. Investig. Dermatol. Inc Eur. Soc. Dermatol. Res.***6,** 170–174 (2001).
3. Fredricks, D. N. Microbial ecology of human skin in health and disease. *J. Investig. Dermatol. Symp. Proc. Soc. Investig. Dermatol. Inc Eur. Soc. Dermatol. Res.***6,** 167–169 (2001).
4. Marples, M. J. *The ecology of the human skin*. (Thomas, 1965).
5. Noble, W. C. Skin microbiology: coming of age. *J. Med. Microbiol.***17,** 1–12 (1984).
6. Roth, R. R. & James, W. D. Microbiology of the skin: resident flora, ecology, infection. *J. Am. Acad. Dermatol.***20,** 367–390 (1989).
7. Cogen, A. L., Nizet, V. & Gallo, R. L. Skin microbiota: a source of disease or defence? *Br. J. Dermatol.***158,** 442–455 (2008).
8. Roth, R. R. & James, W. D. Microbial ecology of the skin. *Annu. Rev. Microbiol.***42,** 441–464 (1988).
9. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature***466,** 334–338 (2010).

10. Stoel, M. *et al.* Restricted IgA repertoire in both B-1 and B-2 cell-derived gut plasmablasts. *J. Immunol. Baltim. Md 1950***174,** 1046–1054 (2005).

11. Hooper, L. V. & Macpherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.***10,** 159–169 (2010).

12. Mantis, N. J. & Forbes, S. J. Secretory IgA: arresting microbial pathogens at epithelial borders. *Immunol. Invest.***39,** 383–406 (2010).

13. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science***313,** 1126–1130 (2006).

14. Putsep, K. *et al.* Germ-free and colonized mice generate the same products from enteric prodefensins. *J. Biol. Chem.***275,** 40478–40482 (2000).

15. Vaishnava, S., Behrendt, C. L., Ismail, A. S., Eckmann, L. & Hooper, L. V. Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proc. Natl. Acad. Sci. U. S. A.***105,** 20858–20863 (2008).

16. Cederlund, A., Agerberth, B. & Bergman, P. Specificity in killing pathogens is mediated by distinct repertoires of human neutrophil peptides. *J. Innate Immun.***2,** 508–521 (2010).

17. Hooper, L. V., Stappenbeck, T. S., Hong, C. V. & Gordon, J. I. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nat. Immunol.***4,** 269–273 (2003).

18. Krinos, C. M. *et al.* Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature***414,** 555–558 (2001).

19. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.***42,** 165–190 (2008).

20. Martínez, I., Kim, J., Duffy, P. R., Schlegel, V. L. & Walter, J. Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects. *PloS One***5,** e15046 (2010).

21. Almeida, J. A. *et al.* Lactose malabsorption in the elderly: role of small intestinal bacterial overgrowth. *Scand. J. Gastroenterol.***43,** 146–154 (2008).

22. Dominy, N. J., Vogel, E. R., Yeakel, J. D., Constantino, P. & Lucas, P. W. Mechanical Properties of Plant Underground Storage Organs and Implications for Dietary Models of Early Hominins. *Evol. Biol.***35,** 159–175 (2008).

23. Luca, F., Perry, G. H. & Di Rienzo, A. Evolutionary adaptations to dietary changes. *Annu. Rev. Nutr.***30,** 291–314 (2010).

24. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.***39,** 1256–1260 (2007).

25. Englyst, H. N., Kingman, S. M. & Cummings, J. H. Classification and measurement of nutritionally important starch fractions. *Eur. J. Clin. Nutr.***46 Suppl 2,** S33–50 (1992).

26. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature***418,** 700–707 (2002).

27. Bergman, E. N. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol. Rev.***70,** 567–590 (1990).

28. Grabitske, H. A. & Slavin, J. L. Low-digestible carbohydrates in practice. *J. Am. Diet. Assoc.***108,** 1677–1681 (2008).

29. Raj, T., Dileep, U., Vaz, M., Fuller, M. F. & Kurpad, A. V. Intestinal microbial contribution to metabolic leucine input in adult men. *J. Nutr.***138,** 2217–2221 (2008).

30. Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.***11,** 2574–2584 (2009).

31. Borrvall, C. & Ebenman, B. Early onset of secondary extinctions in ecological communities following the loss of top predators. *Ecol. Lett.***9,** 435–442 (2006).

32. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U. S. A.***107,** 18933–18938 (2010).

33. Kennedy, T. A. *et al.* Biodiversity as a barrier to ecological invasion. *Nature***417,** 636–638 (2002).

34. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* (2012). doi:10.1038/nrg3182

35. Walter, J. & Ley, R. The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. *Annu. Rev. Microbiol.***65,** 411–429 (2011).

36. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.***108 Suppl 1,** 4680–4687 (2011).

37. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature***473,** 174–180 (2011).

38. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science***326,** 1694–1697 (2009).

39. Maldonado-Contreras, A. *et al.* Structure of the human gastric bacterial community in relation to Helicobacter pylori status. *ISME J.***5,** 574–579 (2011).

40. Grice, E. A. & Segre, J. A. The skin microbiome. *Nat. Rev. Microbiol.***9,** 244–253 (2011).

41. Andersson, A. F. *et al.* Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PloS One***3,** e2836 (2008).

42. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science***308,** 1635–1638 (2005).

43. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature***457,** 480–484 (2009).

44. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.***22,** 292–298 (2012).

45. Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecol. Lett.***12,** 693–715 (2009).

46. Emerson, B. C. & Gillespie, R. G. Phylogenetic analysis of community assembly and structure over space and time. *Trends Ecol. Evol.***23,** 619–630 (2008).

47. Hardin, G. The competitive exclusion principle. *Science***131,** 1292–1297 (1960).

48. Kassen, R. & Rainey, P. B. The ecology and genetics of microbial diversity. *Annu. Rev. Microbiol.***58,** 207–231 (2004).

49. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell***124,** 837–848 (2006).

50. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.***6,** 776–788 (2008).

51. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U. S. A.***107,** 18933–18938 (2010).

52. McNulty, N. P. *et al.* The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci. Transl. Med.***3,** 106ra106 (2011).

53. Petchey, O. L., Eklöf, A., Borrvall, C. & Ebenman, B. Trophically unique species are vulnerable to cascading extinction. *Am. Nat.***171,** 568–579 (2008).

54. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.***108 Suppl 1,** 4554–4561 (2011).

55. Sjölund, M., Wreiber, K., Andersson, D. I., Blaser, M. J. & Engstrand, L. Long-term persistence of resistant Enterococcus species after antibiotics to eradicate Helicobacter pylori. *Ann. Intern. Med.***139,** 483–487 (2003).

56. Blaser, M. J. & Falkow, S. What are the consequences of the disappearing human microbiota? *Nat. Rev. Microbiol.***7,** 887–894 (2009).

57. Mateos, M. *et al.* Heritable endosymbionts of Drosophila. *Genetics***174,** 363–376 (2006).

58. Nyholm, S. V. & McFall-Ngai, M. J. The winnowing: establishing the squid-vibrio symbiosis. *Nat. Rev. Microbiol.***2,** 632–642 (2004).

59. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science***320,** 1647–1651 (2008).

60. McFall-Ngai, M. Adaptive immunity: care for the community. *Nature***445,** 153 (2007).

61. Bogaert, D. *et al.* Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis. *PloS One***6,** e17035 (2011).

62. Fierer, N., Hamady, M., Lauber, C. L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl. Acad. Sci. U. S. A.***105,** 17994–17999 (2008).

63. Blaser, M. J. & Kirschner, D. The equilibria that allow bacterial persistence in human hosts. *Nature***449,** 843–849 (2007).

64. Kuczynski, J. *et al.* Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.***11,** 210 (2010).

65. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature***464,** 59–65 (2010).

66. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science***334,** 105–108 (2011).

67. Ravel, J. *et al.* Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U. S. A.***108 Suppl 1,** 4680–4687 (2011).

68. Huse, S. M. *et al.* Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.***4,** e1000255 (2008).

69. Cauci, S. *et al.* Prevalence of bacterial vaginosis and vaginal flora changes in peri- and postmenopausal women. *J. Clin. Microbiol.***40,** 2147–2152 (2002).

70. Osborne, N. G., Wright, R. C. & Grubin, L. Genital bacteriology: a comparative study of premenopausal women with postmenopausal women. *Am. J. Obstet. Gynecol.***135,** 195–198 (1979).

71. Duncan, S. H. *et al.* Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl. Environ. Microbiol.***73,** 1073–1078 (2007).

72. Walker, A. W. *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.***5,** 220–230 (2011).

73. Beck, J. M., Young, V. B. & Huffnagle, G. B. The microbiome of the lung. *Transl. Res.***160,** 258–266 (2012).

74. Amend, J. P. & Shock, E. L. Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and bacteria. *FEMS Microbiol. Rev.***25,** 175–243 (2001).

75. Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature***388,** 539–547 (1997).

76. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl. Acad. Sci. U. S. A.***103,** 732–737 (2006).

77. Booijink, C. C. G. M. *et al.* High temporal and inter-individual variation detected in the human ileal microbiota. *Environ. Microbiol.***12,** 3213–3227 (2010).

78. Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.***31,** 107–133 (1977).

79. Kanno, T. *et al.* Gastric acid reduction leads to an alteration in lower intestinal microflora. *Biochem. Biophys. Res. Commun.***381,** 666–670 (2009).

80. Hayashi, H., Takahashi, R., Nishi, T., Sakamoto, M. & Benno, Y. Molecular analysis of jejunal, ileal, caecal and recto-sigmoidal human colonic microbiota using 16S rRNA gene

libraries and terminal restriction fragment length polymorphism. *J. Med. Microbiol.***54,** 1093–1101 (2005).

81. Rosenberg, E., Sharon, G., Atad, I. & Zilber-Rosenberg, I. The evolution of animals and plants via symbiosis with microorganisms. *Environ. Microbiol. Rep.***2,** 500–506 (2010).
82. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science***307,** 1915–1920 (2005).
83. Zoetendal, E. G. *et al.* Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces. *Appl. Environ. Microbiol.***68,** 3401–3407 (2002).
84. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature***444,** 1022–1023 (2006).
85. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science***326,** 1694–1697 (2009).
86. Robinson, C. J., Bohannan, B. J. M. & Young, V. B. From structure to function: the ecology of host-associated microbial communities. *Microbiol. Mol. Biol. Rev. MMBR***74,** 453–476 (2010).
87. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.***39,** 31–40 (2007).
88. Turnbaugh, P. J., Bäckhed, F., Fulton, L. & Gordon, J. I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe***3,** 213–223 (2008).

89. Laden, G. & Wrangham, R. The rise of the hominids as an adaptive shift in fallback foods: plant underground storage organs (USOs) and australopith origins. *J. Hum. Evol.***49,** 482–498 (2005).

90. Mottram, D. S., Wedzicha, B. L. & Dodson, A. T. Acrylamide is formed in the Maillard reaction. *Nature***419,** 448–449 (2002).

91. Stecher, B. & Hardt, W.-D. The role of microbiota in infectious disease. *Trends Microbiol.***16,** 107–114 (2008).

92. Tuohy, K. M. *et al.* Metabolism of Maillard reaction products by the human gut microbiota--implications for health. *Mol. Nutr. Food Res.***50,** 847–857 (2006).

93. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature***464,** 908–912 (2010).

94. Tagami, H. Location-related differences in structure and function of the stratum corneum with special emphasis on those of the facial skin. *Int. J. Cosmet. Sci.***30,** 413–434 (2008).

95. Proksch, E., Brandner, J. M. & Jensen, J.-M. The skin: an indispensable barrier. *Exp. Dermatol.***17,** 1063–1072 (2008).

96. Elias, P. M. The skin barrier as an innate immune element. *Semin. Immunopathol.***29,** 3–14 (2007).

97. Segre, J. A. Epidermal barrier formation and recovery in skin disorders. *J. Clin. Invest.***116,** 1150–1158 (2006).

98. Leeming, J. P., Holland, K. T. & Cunliffe, W. J. The microbial ecology of pilosebaceous units isolated from human skin. *J. Gen. Microbiol.***130,** 803–807 (1984).

99. Cohn, B. A. In search of human skin pheromones. *Arch. Dermatol.***130,** 1048–1051 (1994).

100. Emter, R. & Natsch, A. The sequential action of a dipeptidase and a beta-lyase is required for the release of the human body odorant 3-methyl-3-sulfanylhexan-1-ol from a secreted Cys-Gly-(S) conjugate by Corynebacteria. *J. Biol. Chem.***283,** 20645–20652 (2008).

101. Decréau, R. A., Marson, C. M., Smith, K. E. & Behan, J. M. Production of malodorous steroids from androsta-5,16-dienes and androsta-4,16-dienes by Corynebacteria and other human axillary bacteria. *J. Steroid Biochem. Mol. Biol.***87,** 327–336 (2003).

102. Martin, A. *et al.* A functional ABCC11 allele is essential in the biochemical formation of human axillary odor. *J. Invest. Dermatol.***130,** 529–540 (2010).

103. Natsch, A., Gfeller, H., Gygax, P., Schmid, J. & Acuna, G. A Specific Bacterial Aminoacylase Cleaves Odorant Precursors Secreted in the Human Axilla. *J. Biol. Chem.***278,** 5718–5727 (2003).

104. McBride, M. E., Duncan, W. C. & Knox, J. M. The environment and the microbial ecology of human skin. *Appl. Environ. Microbiol.***33,** 603–608 (1977).

105. Faergemann, J. & Larkö, O. The effect of UV-light on human skin microorganisms. *Acta Derm. Venereol.***67,** 69–72 (1987).

106. Fierer, N., Hamady, M., Lauber, C. L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl. Acad. Sci. U. S. A.***105,** 17994–17999 (2008).

107. Gao, Z., Tseng, C., Pei, Z. & Blaser, M. J. Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl. Acad. Sci. U. S. A.***104,** 2927–2932 (2007).

108. Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome Res.***18,** 1043–1050 (2008).

109. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science***324,** 1190–1192 (2009).

110. Leyden, J. J., McGinley, K. J., Mills, O. H. & Kligman, A. M. Age-related changes in the resident bacterial flora of the human face. *J. Invest. Dermatol.***65,** 379–381 (1975).

111. Somerville, D. A. The normal flora of the skin in different age groups. *Br. J. Dermatol.***81,** 248–258 (1969).

112. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.***107,** 11971–11975 (2010).

113. Sarkany, I. & Gaylarde, C. C. Bacterial colonisation of the skin of the newborn. *J. Pathol. Bacteriol.***95,** 115–122 (1968).

114. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U. S. A.***107,** 11971–11975 (2010).

115. Douglass, J. M., Li, Y. & Tinanoff, N. Association of mutans streptococci between caregivers and their children. *Pediatr. Dent.***30,** 375–387 (2008).

116. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.***5,** e177 (2007).

117. Marples, R. R. Sex, constancy, and skin bacteria. *Arch. Dermatol. Res.***272,** 317–320 (1982).

118. Giacomoni, P. U., Mammone, T. & Teri, M. Gender-linked differences in human skin. *J. Dermatol. Sci.***55,** 144–149 (2009).

119. Aly, R., Shirley, C., Cunico, B. & Maibach, H. I. Effect of prolonged occlusion on the microbial flora, pH, carbon dioxide and transepidermal water loss on human skin. *J. Invest. Dermatol.***71,** 378–381 (1978).

120. Paulino, L. C., Tseng, C.-H. & Blaser, M. J. Analysis of Malassezia microbiota in healthy superficial human skin and in psoriatic lesions by multiplex real-time PCR. *FEMS Yeast Res.***8,** 460–471 (2008).

121. Paulino, L. C., Tseng, C.-H., Strober, B. E. & Blaser, M. J. Molecular analysis of fungal microbiota in samples from healthy human skin and psoriatic lesions. *J. Clin. Microbiol.***44,** 2933–2941 (2006).

122. Gao, Z., Perez-Perez, G. I., Chen, Y. & Blaser, M. J. Quantitation of major human cutaneous bacterial and fungal populations. *J. Clin. Microbiol.***48,** 3575–3581 (2010).

123. Lacey, N., Delaney, S., Kavanagh, K. & Powell, F. C. Mite-related bacterial antigens stimulate inflammatory cells in rosacea. *Br. J. Dermatol.***157,** 474–481 (2007).

124. Elston, D. M. Demodex mites: facts and controversies. *Clin. Dermatol.***28,** 502–504 (2010).

125. Schowalter, R. M., Pastrana, D. V., Pumphrey, K. A., Moyer, A. L. & Buck, C. B. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe***7,** 509–515 (2010).

126. Huang, Y. J. *et al.* The Role of the Lung Microbiome in Health and Disease. A National Heart, Lung, and Blood Institute Workshop Report. *Am. J. Respir. Crit. Care Med.***187,** 1382–1387 (2013).

127. Beck, J. M., Young, V. B. & Huffnagle, G. B. The microbiome of the lung. *Transl. Res.***160,** 258–266 (2012).

128. Baughman, R. P., Thorpe, J. E., Staneck, J., Rashkin, M. & Frame, P. T. Use of the protected specimen brush in patients with endotracheal or tracheostomy tubes. *Chest***91,** 233–236 (1987).

129. Thorpe, J. E., Baughman, R. P., Frame, P. T., Wesseler, T. A. & Staneck, J. L. Bronchoalveolar lavage for diagnosing acute bacterial pneumonia. *J. Infect. Dis.***155,** 855–861 (1987).

130. Stressmann, F. A. *et al.* Analysis of the bacterial communities present in lungs of patients with cystic fibrosis from American and British centers. *J. Clin. Microbiol.***49,** 281–291 (2011).

131. Fujimura, K. E. *et al.* Man's best friend? The effect of pet ownership on house dust microbial communities. *J. Allergy Clin. Immunol.***126,** 410–412, 412.e1–3 (2010).

132. Erb-Downward, J. R. *et al.* Analysis of the lung microbiome in the 'healthy' smoker and in COPD. *PloS One***6,** e16384 (2011).

133. Zemanick, E. T., Sagel, S. D. & Harris, J. K. The airway microbiome in cystic fibrosis and implications for treatment. *Curr. Opin. Pediatr.***23,** 319–324 (2011).

134. Guss, A. M. *et al.* Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J.***5,** 20–29 (2011).

135. Beeton, M. L. *et al.* Role of pulmonary infection in the development of chronic lung disease of prematurity. *Eur. Respir. J.***37,** 1424–1430 (2011).

136. Nakajima, T., Palchevsky, V., Perkins, D. L., Belperio, J. A. & Finn, P. W. Lung transplantation: infection, inflammation, and the microbiome. *Semin. Immunopathol.***33,** 135–156 (2011).

137. Hilty, M. *et al.* Disordered microbial communities in asthmatic airways. *PloS One***5,** e8578 (2010).

138. Huang, Y. J. *et al.* The Role of the Lung Microbiome in Health and Disease. A National Heart, Lung, and Blood Institute Workshop Report. *Am. J. Respir. Crit. Care Med.***187,** 1382–1387 (2013).

139. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut***55,** 205–211 (2006).

140. Mondot, S. *et al.* Highlighting new phylogenetic specificities of Crohn's disease microbiota. *Inflamm. Bowel Dis.***17,** 185–192 (2011).

141. Garrett, W. S. *et al.* Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. *Cell Host Microbe***8,** 292–300 (2010).

142. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.***22,** 299–306 (2012).

143. Plottel, C. S. & Blaser, M. J. Microbiome and malignancy. *Cell Host Microbe***10,** 324–335 (2011).

144. McColl, K. E. L. Clinical practice. Helicobacter pylori infection. *N. Engl. J. Med.***362,** 1597–1604 (2010).

145. el-Serag, H. B. & Sonnenberg, A. Opposing time trends of peptic ulcer and reflux disease. *Gut***43,** 327–333 (1998).

146. Chen, Y. & Blaser, M. J. Inverse associations of Helicobacter pylori with asthma and allergy. *Arch. Intern. Med.***167,** 821–827 (2007).

147. Hviid, A., Svanström, H. & Frisch, M. Antibiotic use and inflammatory bowel diseases in childhood. *Gut***60,** 49–54 (2011).

148. Lepage, P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. *Gastroenterology***141,** 227–236 (2011).

149. Andersson, A. F. *et al.* Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PloS One***3,** e2836 (2008).

150. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature***444,** 1027–1031 (2006).

151. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U. S. A.***102,** 11070–11075 (2005).

152. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature***457,** 480–484 (2009).

153. Abu-Shanab, A. & Quigley, E. M. M. The role of the gut microbiota in nonalcoholic fatty liver disease. *Nat. Rev. Gastroenterol. Hepatol.***7,** 691–701 (2010).

154. Fox, J. G. *et al.* Gut microbes define liver cancer risk in mice exposed to chemical and viral transgenic hepatocarcinogens. *Gut***59,** 88–97 (2010).

155. Chen, Y. *et al.* Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatol. Baltim. Md***54,** 562–572 (2011).

156. Doud, M. S., Light, M., Gonzalez, G., Narasimhan, G. & Mathee, K. Combination of 16S rRNA variable regions provides a detailed analysis of bacterial community dynamics in the lungs of cystic fibrosis patients. *Hum. Genomics***4,** 147–169 (2010).

157. Ajslev, T. A., Andersen, C. S., Gamborg, M., Sørensen, T. I. A. & Jess, T. Childhood overweight after establishment of the gut microbiota: the role of delivery mode, pre-pregnancy weight and early administration of antibiotics. *Int. J. Obes. 2005***35,** 522–529 (2011).

158. Luoto, R., Kalliomäki, M., Laitinen, K. & Isolauri, E. The impact of perinatal probiotic intervention on the development of overweight and obesity: follow-up study from birth to 10 years. *Int. J. Obes. 2005***34,** 1531–1537 (2010).

159. Wills-Karp, M., Santeliz, J. & Karp, C. L. The germless theory of allergic disease: revisiting the hygiene hypothesis. *Nat. Rev. Immunol.***1,** 69–75 (2001).

160. Nomura, I. *et al.* Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. *J. Allergy Clin. Immunol.***112,** 1195–1202 (2003).

161. De Jongh, G. J. *et al.* High expression levels of keratinocyte antimicrobial proteins in psoriasis compared with atopic dermatitis. *J. Invest. Dermatol.***125,** 1163–1173 (2005).

162. Pierard, G. E., Arrese, J. E., Pierard-Franchimont, C. & DE Doncker, P. Prolonged effects of antidandruff shampoos - time to recurrence of Malassezia ovalis colonization of skin. *Int. J. Cosmet. Sci.***19,** 111–117 (1997).

163. Frank, D. N. *et al.* Microbial diversity in chronic open wounds. *Wound Repair Regen. Off. Publ. Wound Heal. Soc. Eur. Tissue Repair Soc.***17,** 163–172 (2009).

164. Polavarapu, N., Ogilvie, M. P. & Panthaki, Z. J. Microbiology of burn wound infections. *J. Craniofac. Surg.***19,** 899–902 (2008).

165. Brüggemann, H. *et al.* The complete genome sequence of Propionibacterium acnes, a commensal of human skin. *Science***305,** 671–673 (2004).

166. Gao, Z., Tseng, C., Strober, B. E., Pei, Z. & Blaser, M. J. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. *PloS One***3,** e2719 (2008).

167. Hanifin, J. M. & Rogge, J. L. Staphylococcal infections in patients with atopic dermatitis. *Arch. Dermatol.***113,** 1383–1386 (1977).

168. Leyden, J. J., Marples, R. R. & Kligman, A. M. Staphylococcus aureus in the lesions of atopic dermatitis. *Br. J. Dermatol.***90,** 525–530 (1974).

169. McDowell, A. *et al.* A novel multilocus sequence typing scheme for the opportunistic pathogen Propionibacterium acnes and characterization of type I cell surface-associated antigens. *Microbiol. Read. Engl.***157,** 1990–2003 (2011).

170. Price, L. B. *et al.* Community analysis of chronic wound bacteria using 16S rRNA gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota. *PloS One***4,** e6462 (2009).

171. Uçkay, I. *et al.* Foreign body infections due to Staphylococcus epidermidis. *Ann. Med.***41,** 109–119 (2009).

172. Rogers, G. B. *et al.* Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with cystic fibrosis. *J. Clin. Microbiol.***44,** 2601–2604 (2006).

173. Cox, M. J. *et al.* Airway microbiota and pathogen abundance in age-stratified cystic fibrosis patients. *PloS One***5,** e11044 (2010).

174. Stressmann, F. A. *et al.* The use of culture-independent tools to characterize bacteria in endo-tracheal aspirates from pre-term infants at risk of bronchopulmonary dysplasia. *J. Perinat. Med.***38,** 333–337 (2010).

175. Huang, Y. J. *et al.* A persistent and diverse airway microbiota present during chronic obstructive pulmonary disease exacerbations. *Omics J. Integr. Biol.***14,** 9–59 (2010).

176. Hill, A. B. THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION? *Proc. R. Soc. Med.***58,** 295–300 (1965).

177. Newell, D. G. *et al.* Biosecurity-based interventions and strategies to reduce Campylobacter spp. on poultry farms. *Appl. Environ. Microbiol.***77,** 8605–8614 (2011).

178. Lynch, O. A., Cagney, C., McDowell, D. A. & Duffy, G. Occurrence of fastidious Campylobacter spp. in fresh meat and poultry using an adapted cultural protocol. *Int. J. Food Microbiol.***150,** 171–177 (2011).

179. Manfreda, G., Parisi, A., Lucchi, A., Zanoni, R. G. & De Cesare, A. Prevalence of Helicobacter pullorum in conventional, organic, and free-range broilers and typing of isolates. *Appl. Environ. Microbiol.***77,** 479–484 (2011).

180. Batz, M. B., Hoffmann, S. & Morris, J. G. Ranking the disease burden of 14 pathogens in food sources in the United States using attribution data from outbreak investigations and expert elicitation. *J. Food Prot.***75,** 1278–1291 (2012).

181. Kaakoush, N. O. *et al.* The interplay between Campylobacter and Helicobacter species and other gastrointestinal microbiota of commercial broiler chickens. *Gut Pathog.***6,** 18 (2014).

182. Ellis-Iversen, J. *et al.* Persistent environmental reservoirs on farms as risk factors for Campylobacter in commercial poultry. *Epidemiol. Infect.***140,** 916–924 (2012).

183. Cox, J. M. & Pavic, A. Advances in enteropathogen control in poultry production. *J. Appl. Microbiol.***108,** 745–755 (2010).

184. Annamalai, T. *et al.* Evaluation of nanoparticle-encapsulated outer membrane proteins for the control of Campylobacter jejuni colonization in chickens. *Poult. Sci.***92,** 2201–2211 (2013).

185. Ghareeb, K. *et al.* Evaluating the efficacy of an avian-specific probiotic to reduce the colonization of Campylobacter jejuni in broiler chickens. *Poult. Sci.***91,** 1825–1832 (2012).

186. Kerr, A. K. *et al.* A systematic review-meta-analysis and meta-regression on the effect of selected competitive exclusion products on Salmonella spp. prevalence and concentration in broiler chickens. *Prev. Vet. Med.***111,** 112–125 (2013).

187. Nuotio, L., Schneitz, C. & Nilsson, O. Effect of competitive exclusion in reducing the occurrence of Escherichia coli producing extended-spectrum β-lactamases in the ceca of broiler chicks. *Poult. Sci.***92,** 250–254 (2013).

188. Lin, J. Novel approaches for Campylobacter control in poultry. *Foodborne Pathog. Dis.***6,** 755–765 (2009).

189. Neal-McKinney, J. M. *et al.* Production of organic acids by probiotic lactobacilli can be used to reduce pathogen load in poultry. *PloS One***7,** e43928 (2012).

190. Zhang, G., Ma, L. & Doyle, M. P. Potential competitive exclusion bacteria from poultry inhibitory to Campylobacter jejuni and Salmonella. *J. Food Prot.***70,** 867–873 (2007).

191. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature***489,** 57–74 (2012).

192. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science***306,** 636–640 (2004).

193. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature***447,** 799–816 (2007).

194. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.***9,** e1001046 (2011).

195. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.***5,** 1752–1779 (2011).

196. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature***489,** 101–108 (2012).

197. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.***57,** 159–197 (1988).

198. Urnov, F. D. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *J. Cell. Biochem.***88,** 684–694 (2003).

199. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature***489,** 75–82 (2012).

200. Virani, S., Virani, S., Colacino, J. A., Kim, J. H. & Rozek, L. S. Cancer epigenetics: a brief review. *ILAR J. Natl. Res. Counc. Inst. Lab. Anim. Resour.***53,** 359–369 (2012).

201. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.***27,** 361–368 (2009).

202. Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: the industrial revolution of DNA around histones. *Nat. Rev. Mol. Cell Biol.***7,** 437–447 (2006).

203. Caiafa, P. & Zampieri, M. DNA methylation and chromatin structure: the puzzling CpG islands. *J. Cell. Biochem.***94,** 257–265 (2005).

204. Cheung, P. & Lau, P. Epigenetic regulation by histone methylation and histone variants. *Mol. Endocrinol. Baltim. Md***19,** 563–573 (2005).

205. Kouzarides, T. Chromatin modifications and their function. *Cell***128,** 693–705 (2007).

206. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science***293,** 1074–1080 (2001).

207. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.***13,** R53 (2012).

208. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature***489,** 109–113 (2012).

209. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell***148,** 84–98 (2012).

210. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.***106,** 9362–9367 (2009).

211. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.***7,** 522 (2011).

212. Rozek, L. S., Dolinoy, D. C., Sartor, M. A. & Omenn, G. S. Epigenetics: Relevance and Implications for Public Health. *Annu. Rev. Public Health***35,** 105–122 (2014).

213. McGuinness, D. *et al.* Socio-economic status is associated with epigenetic differences in the pSoBid cohort. *Int. J. Epidemiol.***41,** 151–160 (2012).

214. Bollati, V. *et al.* Epigenetic effects of shiftwork on blood DNA methylation. *Chronobiol. Int.***27,** 1093–1104 (2010).

215. Chu, F. *et al.* Quantitative mass spectrometry reveals the epigenome as a target of arsenic. *Chem. Biol. Interact.***192,** 113–117 (2011).

216. Dolinoy, D. C., Weidman, J. R., Waterland, R. A. & Jirtle, R. L. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. *Environ. Health Perspect.***114,** 567–572 (2006).

217. Anderson, O. S. *et al.* Epigenetic responses following maternal dietary exposure to physiologically relevant levels of bisphenol A. *Environ. Mol. Mutagen.***53,** 334–342 (2012).

218. Dolinoy, D. C., Huang, D. & Jirtle, R. L. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc. Natl. Acad. Sci. U. S. A.***104,** 13056–13061 (2007).

219. Champagne, F. A. *et al.* Maternal care associated with methylation of the estrogen receptor-alpha1b promoter and estrogen receptor-alpha expression in the medial preoptic area of female offspring. *Endocrinology***147,** 2909–2915 (2006).

220. Fish, E. W. *et al.* Epigenetic programming of stress responses through variations in maternal care. *Ann. N. Y. Acad. Sci.***1036,** 167–180 (2004).

221. Heijmans, B. T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci. U. S. A.***105,** 17046–17049 (2008).

222. Bollati, V. *et al.* Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res.***67,** 876–880 (2007).

223. Rusiecki, J. A. *et al.* Global DNA hypomethylation is associated with high serum-persistent organic pollutants in Greenlandic Inuit. *Environ. Health Perspect.***116,** 1547–1552 (2008).

224. Baccarelli, A. *et al.* Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.***179,** 572–578 (2009).

225. Wright, R. O. *et al.* Biomarkers of lead exposure and DNA methylation within retrotransposons. *Environ. Health Perspect.***118,** 790–795 (2010).

226. Kile, M. L. *et al.* Prenatal arsenic exposure and DNA methylation in maternal and umbilical cord blood leukocytes. *Environ. Health Perspect.***120,** 1061–1066 (2012).

227. Arita, A. *et al.* Global levels of histone modifications in peripheral blood mononuclear cells of subjects with exposure to nickel. *Environ. Health Perspect.***120,** 198–203 (2012).

228. Nahar, M. S., Liao, C., Kannan, K. & Dolinoy, D. C. Fetal liver bisphenol A concentrations and biotransformation gene expression reveal variable exposure and altered capacity for metabolism in humans. *J. Biochem. Mol. Toxicol.***27,** 116–123 (2013).

229. Doshi, T., Mehta, S. S., Dighe, V., Balasinor, N. & Vanage, G. Hypermethylation of estrogen receptor promoter region in adult testis of rats exposed neonatally to bisphenol A. *Toxicology***289,** 74–82 (2011).

230. Ho, S.-M., Tang, W.-Y., Belmonte de Frausto, J. & Prins, G. S. Developmental exposure to estradiol and bisphenol A increases susceptibility to prostate carcinogenesis and epigenetically regulates phosphodiesterase type 4 variant 4. *Cancer Res.***66,** 5624–5632 (2006).

231. Tang, W. *et al.* Neonatal Exposure to Estradiol/Bisphenol A Alters Promoter Methylation and Expression of *Nsbp1* and *Hpcal1* Genes and Transcriptional Programs of *Dnmt3a/b* and *Mbd2/4* in the RatProstate Gland Throughout Life. *Endocrinology***153,** 42–55 (2012).

232. Yaoi, T. *et al.* Genome-wide analysis of epigenomic alterations in fetal mouse forebrain after exposure to low doses of bisphenol A. *Biochem. Biophys. Res. Commun.***376,** 563–567 (2008).

233. Feinberg, A. P. The epigenetics of cancer etiology. *Semin. Cancer Biol.***14,** 427–432 (2004).

234. Tremolizzo, L. *et al.* Valproate corrects the schizophrenia-like epigenetic behavioral modifications induced by methionine in mice. *Biol. Psychiatry***57,** 500–509 (2005).

235. Tsankova, N. M. *et al.* Sustained hippocampal chromatin regulation in a mouse model of depression and antidepressant action. *Nat. Neurosci.***9,** 519–525 (2006).

236. Tsankova, N. M., Kumar, A. & Nestler, E. J. Histone modifications at gene promoter regions in rat hippocampus after acute and chronic electroconvulsive seizures. *J. Neurosci. Off. J. Soc. Neurosci.***24,** 5603–5610 (2004).

237. Waldenström, J. *et al.* Avian reservoirs and zoonotic potential of the emerging human pathogen Helicobacter canadensis. *Appl. Environ. Microbiol.***69,** 7523–7526 (2003).

238. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.***38,** 525–552 (2004).

239. Treusch, A. H. *et al.* Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ. Microbiol.***6,** 970–980 (2004).

240. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. U. S. A.***103,** 12115–12120 (2006).

241. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods***5,** 16–18 (2008).

242. Nakamura, S. *et al.* Metagenomic diagnosis of bacterial infections. *Emerg. Infect. Dis.***14,** 1784–1786 (2008).

243. Wylie, K. M. *et al.* Novel bacterial taxa in the human microbiome. *PloS One***7,** e35294 (2012).

244. Jones, K. E. *et al.* Global trends in emerging infectious diseases. *Nature***451,** 990–993 (2008).

245. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.***215,** 403–410 (1990).

246. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.***75,** 7537–7541 (2009).

247. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS One***6,** e27310 (2011).

248. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.***41,** D590–596 (2013).

249. Le Carrou, J., Reinhardt, A. K., Kempf, I. & Gautier-Bouchardon, A. V. Persistence of Mycoplasma synoviae in hens after two enrofloxacin treatments and detection of mutations in the parC gene. *Vet. Res.***37,** 145–154 (2006).

250. Gautier-Bouchardon, A. V., Reinhardt, A. K., Kobisch, M. & Kempf, I. In vitro development of resistance to enrofloxacin, erythromycin, tylosin, tiamulin and oxytetracycline in

Mycoplasma gallisepticum, Mycoplasma iowae and Mycoplasma synoviae. *Vet. Microbiol.***88,** 47–58 (2002).

251. Malik, Y. S., Olsen, K., Kumar, K. & Goyal, S. M. In vitro antibiotic resistance profiles of Ornithobacterium rhinotracheale strains from Minnesota turkeys during 1996-2002. *Avian Dis.***47,** 588–593 (2003).
252. Banani, M., Pourbakhsh, S. & Deihim, A. Antibiotic sensitivity of Ornithobacterium rhinotracheale isolates associated with respiratory diseases. *Archives Razi Inst***58,** 111–117 (2004).
253. Glisson, J. R. Bacterial respiratory disease of poultry. *Poult. Sci.***77,** 1139–1142 (1998).
254. De, R., Drounal, R., Shivaprasad, H. & Walker, R. Ornithobacterium rhinotracheale infection in turkey breeders. *Avian Dis***40,** 865–874 (1996).
255. Arndt, D. *et al.* METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.***40,** W88–95 (2012).
256. Whittaker, R. Evolution and measurement of species diversity. *Taxon***21,** 213–251 (1972).
257. Rylski, M. *et al.* GATA-1-mediated proliferation arrest during erythroid maturation. *Mol. Cell. Biol.***23,** 5031–5042 (2003).
258. Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature***349,** 257–260 (1991).
259. Weiss, M. J., Yu, C. & Orkin, S. H. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol. Cell. Biol.***17,** 1642–1651 (1997).

260. Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood***104,** 3136–3147 (2004).

261. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.***39,** 311–318 (2007).

262. Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.***19,** 2172–2184 (2009).

263. Wu, W. *et al.* Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res.***21,** 1659–1671 (2011).

264. Müller, J. *et al.* Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell***111,** 197–208 (2002).

265. Schotta, G. *et al.* Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J.***21,** 1121–1131 (2002).

266. Ko, L. J. & Engel, J. D. DNA-binding specificities of the GATA transcription factor family. *Mol. Cell. Biol.***13,** 4011–4022 (1993).

267. Leonard, M., Brice, M., Engel, J. D. & Papayannopoulou, T. Dynamics of GATA transcription factor expression during erythroid differentiation. *Blood***82,** 1071–1079 (1993).

268. Tsai, F. Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature***371,** 221–226 (1994).

269. Bell, A. C., West, A. G. & Felsenfeld, G. Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science***291,** 447–450 (2001).

270. Li, T. *et al.* CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loop. *Mol. Cell. Biol.***28,** 6473–6482 (2008).

271. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell***137,** 1194–1211 (2009).

272. Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.***16,** 3145–3157 (1997).

273. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature***489,** 57–74 (2012).

274. Weiss, M. J., Keller, G. & Orkin, S. H. Novel insights into erythroid development revealed through in vitro differentiation of GATA-1 embryonic stem cells. *Genes Dev.***8,** 1184–1197 (1994).

275. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.***33 Suppl,** 245–254 (2003).

276. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat. Oxf. Engl.***4,** 249–264 (2003).

277. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma. Oxf. Engl.***19,** 185–193 (2003).

278. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.***29,** 137–140 (2001).

279. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.***35,** D61–65 (2007).

280. Clamp, M. *et al.* Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.***31,** 38–42 (2003).

281. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.***3,** Article3 (2004).

282. Peddada, S., Harris, S., Zajd, J. & Harvey, E. ORIOGEN: order restricted inference for ordered gene expression data. *Bioinforma. Oxf. Engl.***21,** 3933–3934 (2005).

283. Venables, W. N. & Ripley, B. D. in *Modern Applied Statistics with S* 271–300 (Springer New York, 2002). at <http://link.springer.com/10.1007/978-0-387-21706-2_10>

284. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods***9,** 215–216 (2012).

TYLER MALYS

(240) 308-2345, *tmalys85@gmail.com*

Summary:

Integrative Biosciences, PhD expected May 2015, Bioinformatics and Genomics program at the Pennsylvania State University (PSU). Strengths in Biology with an emphasis on Computational Genomics, Machine Learning and Statistics. Good Communication Skills, Wet Lab, Data Analysis and Extensive Programming Experience applied to metagenomic and epigenetic research challenges.

Research Experierience

August 2013 - Present

- Conducted independent research, collaborated with foreign research partners and performed disease identification, utilizing metagenomics-based analysis on respiratory microbiome samples.
- **Academic advisor: Dr. Eric Harvill**, PSU, Veterinary and Biomedical Sciences, College of Agricultural Sciences

January 2013 –July 2013

- Utilized state-of-the-art statistical and computational tools to relate epigenetic landscape with gene expression levels in a model system for Erythropoiesis, manuscript in preparation.
  Academic Advisor: Dr. Ross Hardison, Biochemistry and Molecular Biology, PSU Center of Comparative Genomics and Bioinformatics

January 2011 – December 2012

- Designed computer algorithms to address np-complete problems of biological relevance.
- Analyzed data sets from the Whole Genome Association Study (GWAS) and Epigenetic Data
- **Academic Advisor: Dr. Yu Zhang**, PSU, Dept of Statistics

May 2007 – August 2007

- Assisted study of insect behavior, implemented computer controlled camera surveillance of test subjects, **Academic Advisor***: **Dr. Frank Hanson**, University of Maryland Baltimore County*

Teaching Experience:

Introduction to Microbiology, Dr. Ola Sodeine, PSU — *August 2013-Present*
Statistical Concepts and Reasoning, Dr. Patricia Buchannan, PSU — *August 2012 – December 2012*
Stochastic Modeling, Dr. John Fricks, PSU — *January 2012 – May 2012*
Bio-Behavioral Health, Dr. Roger McCarter, PSU — *August 2010 – December 2010*
Introductory Physics, Dr. Eric Anderson, University of Maryland Baltimore County — *August 2006 – December 2007*
Mathematics Tutor, Mathematics Department, Frederick Community College, MD — *August 2004 – July 2005*

Education:

Pennsylvania State University, Integrative Biological Sciences PhD program — *August 2009-Present*

- Studied Computational Genomics, GPA 3.60, Science GPA 4.0
- Phd Expected May 2015

University of Maryland College Park, Graduate Studies — *August 2008 – June 2009*

- Studied Cellular Biology, part time, GPA 4.0, Science GPA 4.0

University of Maryland Baltimore County, Bachelor of Science, Biological Sciences — *June 2005-May 2008*

- Graduated *Magna Cum Laude*, Studied Biological Sciences.
- GPA: 3.83, Science GPA: 3.97

Frederick Community College, Frederick MD — *August 2003-May 2005*

- General Studies, GPA: 3.67, Science GPA: 4.0