이학석사학위논문

Trailing genomic signature to discover substantial information in genome data using bioinformatics approaches

생물정보학적 접근방법을 이용한 유전적 표지인자 탐색과 유전체데이터의 유용한 정보 발굴

2016 년 2 월

서울대학교 대학원

협동과정 생물정보학과

정 현 수

Trailing genomic signature to discover substantial information in genome data using bioinformatics approaches

By

Hyeonsoo Jeong

Supervisor: Professor Heebal Kim

Feb, 2016

Interdisciplinary Program in Bioinformatics

Seoul National University

생물정보학적 접근방법을 이용한 유전적 표지인자 탐색과 유전체데이터의 유용한 정보 발굴

지도교수 김 희 발

이 논문을 이학석사 학위논문으로 제출함
2015 년 12 월

서울대학교 대학원
생물정보협동과정 생물정보학과
정 현 수

정현수의 이학석사 학위논문을 인준함
2015 년 12 월

위 원 장	김 선	(인)
부위원장	김 희 발	(인)
위 원	조 서 애	(인)

Abstract

Trailing genomic signature to discover substantial information in genome data using bioinformatics approaches

Hyeonsoo Jeong
Interdisciplinary Program in Bioinformatics
Seoul National University

These studies are mainly focusing on the deciphering biologically meaningful information in genome sequences of living organisms using bioinformatics techniques.

In chapter 2, I investigate the relationship between genomic composition and Berkshire pig’s meat quality trait by scanning for signatures of positive selection in whole-genome sequencing data. Berkshire pigs are regarded as having superior meat quality compared to other breeds. As the meat production industry seeks selective breeding approaches to improve profitable traits such as meat quality, information about genetic determinants of these traits is in high demand. However, most of the studies have been performed using trained

sensory panel analysis without investigating the underlying genetic factors. Results revealed several candidate genes involved in Berkshire meat quality; most of these genes are involved in lipid metabolism and intramuscular fat deposition.

In chapter 3, I construct the HGTree: database of horizontally transferred genes determined by tree reconciliation. In Bacteria and Archaea, Horizontal gene transfer (HGT) plays an important role in the acquisition of biological advantages such as virulence factor and antibiotic resistance and provides significant genetic diversity. It is important to have a well-defined database containing precise information about HGT events between Prokaryotes in order to understand prokaryotic evolution and discover genes which have led to adaptive genetic variation through HGT as opposed to processes such as mutation, natural selection, or genetic drift. The HGTree database provides putative genome-wide horizontal gene transfer information for 2,472 prokaryotic genomes by reconciling gene trees against species trees. The tree reconciliation method is considered to be a useful way to detect HGT events but has not been utilized extensively by existing databases because the method is computationally intensive and conceptually challenging. In this regard, HGTree represents a useful addition to the biological community, enabling quick and easy retrieval

of information for HGT-acquired genes and better understanding of microbial taxonomy and evolution. The database is freely available and can be easily scaled and updated to keep pace with the rapid rise in genomic information.

Key words: Positive selection, Phylogenetics, Horizontal gene transfer, Tree reconciliation

Student number: 2014-21323

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 Detecting positive selection

1.1.1 Signature of positive selection

Positive selection, also known as directional selection, is a type of natural selection in which a specific phenotype is dominant in population. Under positive selection, the allele that increases reproductive and survival fitness become fixed over time in a specific population. It is well known that positive selection plays an important role in the history of mammals. Therefore, finding positively selected regions in animals which have a unique phenotypic trait can help clarifying the role of a genomic region and its function (Bowcock, Kidd et al. 1991, Sabeti, Reich et al. 2002, Sabeti, Schaffner et al. 2006).

Recently, it has been shown that using a distorted pattern of genetic variation between populations can be useful for detecting selection related to specific traits. For example, genetic signals of selection discovered several genes in cattle responsible for milk production (Qanbari, Pimentel et al. 2010). Also, Pollinger et al. identified rapid phenotypic diversification unique to the domestic dog (Pollinger, Lohmueller et al. 2010), and Moradi et al. revealed three regions associated with fat deposition in thin and fat tail sheep breeds (Moradi, Nejati-Javaremi et al. 2012).

1.1.2 Methods for detecting positive selection

One of the underlying concepts of detecting positive selection has been based on searching for a region where genetic diversity was decreased in a large region. This distorted pattern of genetic variation (so-called selective sweep) plays an essential role in detecting a unique characteristic of the certain population (Sabeti, Schaffner et al. 2006). There are several methods to detect signature of positive selection from the genetic data. 1) reduction in genetic diversity (reduced level of heterozygosity), 2) high frequency derived allele (a skew in the allele frequency distribution), 3) long-range association with other alleles, and 4) differences between populations.

The cross-population extended haplotype heterozygosity test (XP-EHH) measures signals of ongoing or nearly fixed selective sweeps by analyzing haplotypes between two populations (Sabeti, Varilly et al. 2007). This approach is known to be a strong statistical test detecting positive selection occurred in a short period of time. Also, another statistical method for identifying signals of selective sweep is the cross-population multiple-locus composite likelihood ratio test (XP-CLR), which is detecting multi-locus allele frequency differentiation between two populations (Chen, Patterson et al. 2010).

1.2 Detecting horizontal gene transfer

1.2.1 Horizontal gene transfer

Vertical inheritance refers to transfer of genetic information from parents to offspring. Vertically inherited genes typically show higher degree of similarity between organisms and species that are closely related than those that are distantly related. This aids in reliable recognition of species and understanding their classification and evolution. For example, ribosomal RNA genes have been historically used to determine the taxonomic structure of cellular life (Woese 1987). However, vertical signal can sometimes be confounded by acquisition of genes from other sources such as environmental, viruses, or via direct interactions between organisms. Recent advances in genomics have confirmed the existence of 'foreign' genes embedded in cellular genomes. For example, mammalian genomes are enriched with viral-like genetic elements, constituting up to 8% of the human genome (Griffiths 2001). Similarly, many microbial genomes possess genes acquired from multiple sources (Doolittle 1999). This phenomenon is referred to as horizontal gene transfer (HGT), which is a natural outcome given the numerous ways species interact with each other and occupy common habitats.

HGT allows gain of novel molecular functions and can (sometimes) provide selective evolutionary advantages to species. For example, transfer of antibiotic resistance genes and virulence factors between bacterial species posits significant challenges to human health (Salyers, Gupta et al. 2004). Similarly, transfer of genes involved in response to heat and cold shock and heavy metal and ultra-violet resistance facilitates bacterial adaptation to certain environments. While HGT is an important force driving the evolution of (especially) microbial organisms (Koonin, Makarova et al. 2001), it can complicate interpreting the true evolutionary history of species and lead to erroneous interpretations regarding their classification and community interactions (Doolittle 1999). Therefore, it is crucial to distinguish between vertically and horizontally acquired genes in genomes when studying deep evolutionary relationships.

1.2.2 Methods for detecting HGT events and their limitations

Accurate detection of HGT remains a computational and conceptual challenge. Existing databases such as HGT-DB (Garcia-Vallvé, Guzmán et al. 2003) and DarkHorse HGT Candidate Resource (Podell, Gaasterland et al. 2008) use genomic signatures (i.e. GC bias, nucleotide composition, and codon usage) and implicit phylogenetic methods (i.e. comparing evolutionary distance inferred from sequence similarity) to detect HGT. Because genomic signatures of transferred genes may lose their 'distinctiveness' over long periods of evolutionary time and tend to be highly similar to host genomes in cases of HGT between very closely related organisms, these methods likely have a high rate of false-positive and negative predictions (Guindon and Perriere 2001, Lawrence and Ochman 2002). Moreover, GC composition within the same genome may fluctuate considerably for different genomic regions (Deschavanne and Filipski 1995, Wuitschick and KARRER 1999) and (even) for some vertically inherited genes (e.g. ribosomal proteins) (Muto and Osawa 1987), leading to erroneous inferences. In turn, implicit phylogenetic methods are limited by their reliance on similarity scores and underlying phylogeny. This poses another problem since statistically significant sequence similarity is not necessarily a result of vertical evolution (Koski and Golding 2001, Ravenhall, Škunca et al. 2015). Because genes acquired from foreign sources typically do not show congruence to species trees, one way to detect HGT would be to reconcile gene trees against reference species trees. This principle is based on an explicit evolutionary model and is generally considered to be a reliable alternative to detect HGT events (Ragan 2001). However, its practical use has remained limited because reconciling trees is computationally intensive (Ravenhall, Škunca et al. 2015) and because tree incongruence can also arise from processes other than HGT (Than, Ruths et al. 2007) (see also Discussion in Chapter 3).

This chapter was published in BMC Genetics
as a partial fulfillment of Hyeonsoo Jeong’s M.Sc. program.

Chapter 2. Exploring evidence of positive selection reveals genetic basis of meat quality traits in Berkshire pigs through whole genome sequencing

2.1 Introduction

The domestic pig, *Sus scrofa domestica*, has been an important food source throughout human history. In addition to undergoing natural selection due to various environmental factors, pig breeds have gone through intensive artificial selection in order to increase economically important traits such as reproduction, growth rate, stress resistance, and meat quality (Hazel 1943). For example, studies have shown that modern Landrace and Yorkshire breeds were positively selected to improve both reproduction and lactation ability for economic traits (Serenius, Sevón-Aimonen et al. 2004).

Berkshire pigs have been renowned for their superior meat quality since their meat contains a great proportion of neutral lipid fatty acids and marbling fat (Wood, Nute et al. 2004) which is important for palatability characteristics such as tenderness and juiciness. This breed has been intensively selected for meat quality in recent centuries, especially in East Asia where it is marketed as black pork at a premium price. Therefore, Berkshire has become specialized for high quality meat production and relative lack of boar taint following strong artificial selection for these traits. While several studies have investigated genetic factors relating to meat quality in Berkshire pigs (Suzuki, Shibata et al. 2003, Jeong, Choi et al. 2010, Lee, Choi et al.

2010, Kang, Choi et al. 2011), most of the research is performed in the traditional way using trained sensory panel analysis without investigating underlying genetic factors.

Thus, identifying genetic regions that are positively selected especially in Berkshire breed might allow us to reveal genetic variation related to phenotypic trait. In this study, whole genome sequencing of Berkshire, Landrace, and Yorkshire breeds was conducted to identify genomic variants. I performed two statistical analyses, the cross-population extended haplotype homozygosity test (XP-EHH) and the cross-population composite likelihood ratio test (XP-CLR), to determine signals of selection in Berkshire breed. In addition, I performed a Fisher’s exact test for detection of breed specific amino acids or Indels, which are specifically enriched and affected by positive selection. Finally, Berkshire specific aligned reads were separately analyzed to detect the genomic difference between Berkshire and other breeds using *de novo* short sequencing reads assembly.

2.2 Material and methods

Ethics statement

The experiment and all its procedures were approved by the regional Ethical Committee (JNU Animal Bioethics committee permit number: 2013-0009).

Sample preparation and whole genome re-sequencing

For genomic DNA extraction, tissue and blood samples were collected from 10 female Berkshire pigs. Berkshire tissue samples were collected from a local pig breeding company in Namwon, Korea. To generate inserts of ~300 bp, 3 µg of genomic DNA was randomly sheared using Covaris System. The TruSeq DNA Sample Prep. Kit (Illumina, San Diego, CA) was used for library construction by following the manufacturer's guidelines. Whole genome sequencing was performed on the Illumina HiSeq 2000 platform.   Whole-genome sequence data of 11 Landrace (Danish) and 13 Yorkshire (Large White) pigs was obtained from NCBI Sequence Read Archive database under accession number SRP047260. FastQC (Andrews 2010) software was used to perform a quality check on raw sequence data. Using Trimmomatic-0.32 (Bolger, Lohse et al. 2014), potential adapter sequences were removed prior to sequence alignment. Paired-end sequence reads were mapped to the pig reference genome (Sscrofa 10.2) from the Ensembl database using Bowtie2 (Langmead and Salzberg 2012) with default settings.

For downstream processing and variant-calling, I used open-source software packages: Picard tools (http://picard.sourceforge.net), SAMtools (Li, Handsaker et al. 2009), and Genome Analysis Toolkit (GATK) (McKenna, Hanna et al. 2010). “CreateSequenceDictionary” and “MarkDuplicates” Picard command-line tools were used to read reference FASTA sequence for writing bam file with only sequence dictionary, and to filter potential PCR duplicates, respectively. Using SAMtools, index files for the reference and bam files were created. Local realignment of sequence reads was performed to correct misalignment due to the presence of small insertion and deletion using GATK “RealignerTargetCreator” and “IndelRealigner” arguments. Also, base quality score recalibration was performed to get accurate quality scores and to correct the variation in quality with machine cycle and sequence context. For calling variants, GATK “UnifiedGenotyper” and “SelectVariants” arguments were used with the following filtering criteria. All variants with 1) a Phred-scaled quality score of less than 30; 2) read depth less than 5 ; 3) MQ0 (total count across all samples of mapping quality zero reads) > 4; or a 4) Phred-scaled P-value using Fisher’s exact test more than 200 were filtered out to reduce false positive calls due to strand bias. “vcf-merge” tools of VCFtools (Danecek, Auton et al. 2011) was used in order to merge all of the variants calling format files for the 34 samples. BEAGLE software (Browning and Browning 2007) was used to conduct the haplotype phasing for the entire set of pig populations.

Population stratification

I used Genome-Wide Complex Trait Analysis (GCTA) (Yang, Lee et al. 2011) to calculate eigenvectors which are equivalent to those estimated by the EIGENSTRAT software tool for principal component analysis (PCA). Autosomal genotype data was converted to PLINK (Purcell, Neale et al. 2007) format, the input format required for GCTA, using VCFtools.

Statistical analysis

Two methods were employed to infer positive signatures in Berkshire population. Firstly, XPEHH software (Sabeti, Varilly et al. 2007), which measures cross-population extended haplotype homozygosity, was used to detect signatures of positive selection. I calculated EHH and the log ratio of the integrated haplotype homozygosity (iHH) for the pairwise test of Berkshire and other breeds for each of the SNP loci. An extreme value of XP-EHH suggests selection in Berkshire breed. Log ratios was standardized using R (Ihaka and Gentleman 1996), and divided the genome into consecutive, non-overlapping 25 kb windows.

The SNP with the maximum XP-EHH value was selected to represent the summary statistics for each window. To define empirical P-value, I considered the number of SNPs in each window, and binned genomic windows according to the numbers of SNPs in increments of 200 SNPs. When a window encompassed more than 600 SNPs, I combined all the windows (> 600 SNPs) into one bin. An empirical P-value for each window was defined based on its ranking of summary statistics in its bin following the protocol of previous studies (Granka, Henn et al. 2012, Lee, Kim et al. 2014). I assigned all of the regions with an empirical P-value less than 0.01 as the candidate regions which were positively selected in Berkshire breed.

Next, the cross-population composite likelihood ration test (XP-CLR) (Chen, Patterson et al. 2010) was performed using the XP-CLR software package with non-overlapping windows of 25 kb. Windows with a XP-CLR value in the top 1% of the empirical distribution were designated as candidate regions. Genes located in the regions under significant selection were annotated.

Additionally, two types of Fisher’s exact tests were performed using a 2x2 contingency table for detecting breed specific amino acid or Indel. Firstly, a specific amino acid enrichment test was performed using a contingency table composed of two factors such as specific breed

(Berkshire / other) and specific amino acid information (‘specific amino acid’ / other). I performed the statistical test 3 (Berkshire, Landrace and Yorkshire) * k * n times on each of amino acid position in the targeted gene, where k is the number of existing different type of amino acid on each position, and n is number of site in targeted gene. Secondly, a specific Indel enrichment test was performed on the table composed of specific breed information and Indel existence (Yes / No) in each of positions on targeted gene. This statistical test was also performed 3*2*n times on each position. Using these tables, a Fisher’s exact test was performed with the alternative hypothesis that the odds ratio is greater than 1. The two types of statistical tests, for non-synonymous SNP and Indel, respectively, calculate cumulative type-1 error through individual statistical tests. The Bonferroni correction method was employed for considering multiple testing problems in the enrichment test.

Short reads assembly using NGS sequence reads

To eliminate possible sequencing errors, I used “Error correction” module of Allpaths-LG (Gnerre, MacCallum et al. 2011) with default settings. Error corrected paired-end reads were merged to FASTA format using “Fq2fa” module from IDBA v1.1.1 software (Peng, Leung et al. 2012) which stands for iterative De Bruijn graph *De novo* assembler for short reads sequencing data with highly uneven sequencing depth. I assembled error corrected paired-end reads using IDBA_UD from IDBA package with the following parameters: 1) Perform pre-correction before assembly (“--pre_correction”), and 2) minimum k value should be more than 30 (--mink 30). Using Gapcloser (Luo, Liu et al. 2012), predicted gaps were filled in the assembled sequences with a default setting.

In order to identify genomic regions unique to the Berkshire population, I defined sequence reads which unaligned to the reference genome and Landrace/Yorkshire assembled contigs but aligned to the Berkshire assembled contigs using Bowtie2 (Langmead and Salzberg 2012). Among the total Berkshire assembled contigs, contigs with an average mapping depth of sequence reads resulted from the previous process of over 10 in common between every Berkshire samples were defined as the candidate region. RepeatMasker (Tarailo-Graovac and Chen 2009) was used to screen DNA sequences for interspersed repeats and low complexity DNA sequences before gene prediction for the candidate contigs.

2.3 Results

DNA sequencing and whole genome re-sequencing

The whole genomes of 10 Berkshire, 11 Landrace, and 13 Yorkshire pigs were sequenced to an approximate coverage of 11.68-fold on average, with a total of 1,201,160,368,944 bp in 11,981,734,530 reads after removing potential adapter sequence using Trimmomatic-0.32. Sequence reads of each breed were aligned to the pig reference genome (*Sus scrofa* 10.2) from the Ensembl database using Bowtie2, and 88.46 % of the sequence reads were aligned to the reference sequence (Table 2.1-3). After removing PCR duplicates and recalibrating base quality, 18,886,809 single nucleotide variants (SNVs) and 3,384,566 Indels were retained. Of the total SNVs, although 15,237,076 SNVs (80.7%) have been already reported previously to dbSNP (Sus scrofa 10.2.74; ftp://ftp.ensembl.org/pub/release-74/variation/vcf/sus_scrofa/Sus_scrofa.vcf.gz), 3,649,733 SNVs were defined as novel variants (19.3%). The distributions of both types of SNVs in each chromosome are shown in Figure 2.1.

Population Stratification

Using genome-wide complex trait analysis (GCTA), I performed principal component analysis (PCA) of the whole autosomal genotype loci (SNP; n = 18,802,810) to characterize the pattern of individual samples. The analysis revealed structurally cleared difference between populations. As shown in Figure 2.2(a), the first eigenvector (15.7% of the total variance) separated Berkshire from other breeds, and Landrace and Yorkshire pigs were divided by the second eigenvector (13.7% of the total variance).

Signatures of selection in the Berkshire breed

To detect signals of positive selection in Berkshire against other breeds, I used two statistical analysis methods in order to achieve maximum statistical power for localizing the source of selection. I first used the cross-population extended haplotype homozygosity (XP-EHH) statistic to make comparisons between Berkshire and other breeds (Landrace and Yorkshire). This statistic is originally designed to estimate alleles that have increased in frequency to the point of fixation or near-fixation in one of the populations and assesses haplotype differences between two populations (Simonson, Yang et al. 2010). To make comparisons of genomic regions across populations, I divided the genome into consecutive, non-overlapping segments of 25 kb. Among the total of 98,037 windows, the maximum XP-EHH score was assigned in each segment as the window statistic. Giving consideration to the number of SNPs in each segment, the test statistic was converted to an empirical p-value based on its rank of XP-EHH score. Those that yielded significant values ($P < 0.01$) were identified as positively selected regions (Figure 2.3(a)). A total of 177 and 207 genes were identified as positive signatures from XP-EHH test in Berkshire breed against to Landrace and Yorkshire breed, respectively (Figure 2.2(b)).

I also ran a cross-population composite likelihood ratio test (XP-CLR) to search for the genomic regions where the changes in allele frequency at the locus occurred very fast due to random drift. XP-CLR is a multi-locus sliding window test which is robust to ascertainment bias in SNP discovery (Chen, Patterson et al. 2010). XP-EHH and XP-CLR were used to detect signatures of selective sweeps by comparing signals from two populations. However, while the XP-CLR test considers the variation of allele frequency using the differentiation of multi-locus allele frequency between two populations, the XP-EHH test aims primarily to identify differentially overrepresented haplotypes between two populations. In addition, combining the results from two different statistical analyses provides more powerful information than results from one test alone. The whole genome area was divided into non-overlapping windows of 25 kb as before. All windows above a threshold of 216.23 and 257.06 (top 1% of the empirical distribution) were defined as significant regions (Figure 2.3(b)), and identify 333 and 371 positively selected genes in Berkshire compared to Landrace, and to Yorkshire, respectively (Figure 2.2(b)).

Identification and analysis of positively selected genes in Berkshire

While selective traits are likely to be detected among various regions, I focused specifically on the meat quality specific to the Berkshire breed. The amount of fat and fatty acid in adipose tissue or muscle as well as the muscle fiber characteristic plays an important role in meat quality (Wood, Nute et al. 2004). To identify genomic regions associated with meat quality in Berkshire, I detected candidate genes using two statistics (XP-EHH and XP-CLR) comparison between Berkshire and mother breeds (Landrace and Yorkshire) which are superior in maternal performance farrowing and raising large litters of pigs (Johnson and Omtvedt 1973, Hanenberg, Knol et al. 2001). Landrace and Yorkshire purebreds are well-known for their reproductive performance. In particular, Yorkshire pigs are noted for slow growth compared to Landrace or Berkshire pigs. When I compared the genes detected from statistical analyses of B-L and B-Y, a considerable number of common genes related to growth performance in the results of B-Y but not in the results of B-L (*WNT2, FGF14, PTPN11, FXYD2, APBB1, ACAP1, NET1, NF2,* and *KCTD11*).

I observed 56 genes (Table 2.4) overlapped among the 177 and 207 resulting from comparisons between Berkshire and Landrace breeds and between Berkshire and Yorkshire breeds using XP-EHH analysis, respectively (Figure 2.2(b)). The positively selected gene list included *FABP1* and *TG*. These results suggest that several genomic regions and genes may have been selected for meat quality in Berkshire pigs (Table 2.5). *Fatty acid-binding protein1* (*FABP1*) also known as liver fatty acid-binding protein (*L-FABP*) is a member of the *FABP* multi-gene family expressed in both the liver and small intestine (Chmurzyńska 2006). It has been suggested that *L-FABP* gene, which has an effect on uptake, transport, mitochondrial oxidation, and esterification of fatty acids, were strongly related to meat quality in previous study (Atshaves, McIntosh et al. 2004, JIANG, LI et al. 2006, Wang, Shu et al. 2007). *Thyroglobulin* (*TG*) gene, encoding to produce the precursor for thyroid hormones, affects adipocyte growth, differentiation and homeostasis of fat deports (Rosenfeld, Mermod et al. 1983). Many studies have shown that *TG* is significantly associated with meat quality traits. (Barendse, Bunch et al. 2004, Burrell, Moser et al. 2004, Fortes, Curi et al. 2009, Smith, Thomas et al. 2009). *AKIRIN2*, a homolog of the *Akirin* protein, is relevant to the control of skeletal myogenesis through up-regulation of muscle specific transcription factors (Chen, Huang et al. 2013); it is also negatively regulated by cytokine such as myostatin, which plays an important role in skeletal myogenesis (Marshall, Salerno et al. 2008). In a previous study, Sasaki et al. detected a SNP in the 3' untranslated region of the *AKIRIN2* is associated with marbling in Japanese Black beef cattle (Sasaki, Yamada et al. 2009). The high proportion of marbling, which is defined by the amount and distribution of intramuscular fat (IMF), exceedingly improve the palatability by affecting the taste and tenderness of the meat. Also, a SNP located in an intron region of Glucagon-like peptide 2 receptor (*GLP2R*) is significantly associated with IMF according to a previous study (Luo, Cheng et al. 2012). Transforming growth factor β3 (*TGF-β3*), a secreted protein, is related with the mammalian target of rapamycin (mTOR) pathway, which has been renowned as significantly associated with muscle mass and strength (Park, Jacobsson et al. 2006). Although its specific mechanism is not well understood, it is clear that *TGFBR3* plays a role in the muscular or adipose tissue development (Cánovas, Quintanilla et al. 2010). Also, Chen at el. recently discovered a SNP in *TGF-β1/2/3* had an effect on myofiber diameter (Chen, An et al. 2013). Berkshire has been renowned to have smaller cross-sectional area and high density muscle fiber compared to other breeds (Jeong, Choi et al. 2010). Many studies have shown the relationship between the composition of myofiber type and pork quality (Lebret, Le Roy et al. 1999), and this result is at the base of the fact that Berkshire pork has a tremendous tenderness and juiciness. Also, *JPH3*, *PPP2R5C*, *USP25*, and *ACTN2* were associated with boar taint (Ramos, Duijvesteijn et al. 2011), IMF, tenderness (Hamill, McBryan et al. 2012), and cooking loss (Li, Kim et al. 2011), respectively.

To explore deep into the phenotypic traits of Berkshire breed, I further investigated the 114 genes (Table 2.4) observed using XP-CLR (Figure 2.2(b)). 13 genes intersected with the results from XP-EHH selection candidate genes (Table 2.4). Interestingly, these genes included *FABP1*, *TG*, *ERN1*, *JPH3*, and *ICAM2* (Shin and Chung 2007, Chang, Yeh et al. 2008, Qiu, Ni et al. 2008, Sen, Jumaa et al. 2013).

In addition to genes responsible for meat quality traits, our genome-wide selection scan also identified genes associated with immune response, particularly regulation of leukocyte and immunoglobulin (*CD79B*, *CD8B*, *FLT3*, *ICAM2*, *IFNGR1*, and *IGSF5*). Berkshire pigs have an unusually high concentration of plasma immunoglobulin as opposed to the other breeds, as evidenced by distinctive high percentages of neutrophils and leukocytes (Sutherland, Rodriguez-Zas et al. 2005).

For further analysis of the influence of genomic variants on protein function, I performed a Fisher’s exact test for the detection of specific enriched sites on the 13 genes which were in the intersection with the results from XP-EHH and XP-CLR. Previously, most studies have focused on non-synonymous SNPs, since substitution is known to affect gene function. Also, many studies focused on deletions and insertions sites, which can affect the performance traits considerably in pigs (Hanjie, Yanhua et al. 2005, Li, Li et al. 2007). Therefore, statistical analysis was performed employing these two types of data, non-synonymous SNP and Indel site, under positive selective region. From the test results, numerous P-values are generated. For easily identified significant test results, I draw the line plots composed with $-log_{10}(p - value)$ and each of site, y-axis and x-axis, respectively. Each test result was plotted together (Figure 2.4). From the figures, I can easily detected significant enriched site, breed, and amino-acid or Indel, simultaneously. I identified several genes including significant sites, *TG*, *CPED1*, *CPNE8*, *CD8*, *ERN1*, *ICAM2*, *JPH3*, *NELFCD*, *SP110*, and *ADAM7* in Indel data under Bonferroni corrected 5% significance level. These genes have a possibility that is related to breed specific phenotypic variation between Berkshire and other breeds by Indel.

Whole genome assembly

Although analyzing positive selection signature between breeds using SNP and small Indel information could allow us to identify genetic variation which affects phenotypic diversity, it is also important to consider large sequence differences, which can be difficult to detect using reference-based alignments. I assembled short reads sequence of each breed to decipher the large genomic difference of Berkshire compare to other breeds more deeply. The sample with high concordantly paired mapping rate to the reference genome and with low heterozygosity was selected to perform genome assembly for each breed. After whole genome assembly was performed using IDBA-UD, all of the contigs less than 2,000 bp were removed for the minimum threshold length. I observed an average of 223,028 contigs with an average length of 10,843 bp, and N50 length for Berkshire, Landrace, and Yorkshire are 30,152, 9,379, and 9,694, respectively. I further performed the gapclosing step to fill N base within the contig. The average sum of the total assembled contigs after the gapclosing step for Berkshire, Landrace, and Yorkshire breeds was 2,304 Mbp, 1,927 Mbp, and 1,996 Mbp, respectively. Detailed results are shown in Table 2.6.

To infer distinct genomic contents for Berkshire against other breeds, firstly, I compared the overall read mapping rate between assembled contigs for each breed, using the total mapped reads of each Berkshire sample (Figure 2.5). The average overall read mapping rate to the Berkshire assembled contigs was 93.5% in contrast to the Landrace and Yorkshire assembled contigs was 79.9% and 82.3%, respectively, which is also about 4.7% higher to the overall mapping rate of reference-based alignment. Although satellites sequences were about 0.1% in Berkshire assembled contigs which is about 0.04% higher than others at 0.06%, there was no significant difference based on the ratio of interspersed repeat elements including retrotransposon and retrovirus-like sequence in each assembled contigs (Table 2.7).

I then separately remapped the each Berkshire sample's sequencing reads, which were both unmapped to the reference genome and to the Landrace/Yorkshire assembled contigs, using Berkshire assembled contigs to find the regions in Berkshire that are distinct from the others. The average mapping rate of unmapped reads was about 37.8% aligned to the Berkshire assembled contigs using Bowtie2, and the details of the information for each sample is described in Table 2.8. Among the total number of 127,713 Berkshire assembled contigs, I observed 563 contigs which the unmapped reads were aligned with depth coverage of more than 10 in common between all Berkshire samples. Additionally, I removed PCR duplication of sequence reads to reduce the number false positives. As shown in Table 2.7, the results summary of repeat contents demonstrated that high proportion of satellites (24.4%) was detected in these contigs which is approximately 240 times higher than those of the total assembled sequence. After performing gene prediction and functional annotation, 43 contigs with 46 predicted genes were finally identified as Berkshire specific candidate genomic region. Out of 46 predicted genes, I identified 4 genes that were related to lipid metabolism: *SLC25A14* (Kopecký, Rossmeisl et al. 2004), *IGF1*(Saltiel and Kahn 2001), *PI4KA* (Balla, Tuymetova et al. 2005), and *CACNA1A* (Taverna, Saba et al. 2004) (Table 2.9). Li et al. recently identified 44 genes with 49 SNPs showing significant association with muscling and meat quality trait (Li, Kim et al. 2011). Of the 44 candidate genes, *DLX1* and *DLX3* showed a concordant result with our study. In addition, *TGFBR3* and *SYT1*, also identified from positive selection scan, were included in the candidate gene list. Besides the meat quality trait, 6 genes (*OR4D10*, *OR4D11*, *ENSSSCG00000028782*, *ENSSSCG00000029769*, *ENSSSCG00000013807*, and *ENSSSCG00000021192*) including 4 novel genes were related to olfactory receptor.

Table 2. 1 The result summary of sequence reads mapping using Bowtie2 (Berkshire).

Samples	Total number of reads	Paired align (concordantly)	Paired align (discordantly)	Non-paired align	Overall alignment rate
B_1	304,804,588 (100%)	264,655,464 (86.83%)	1,447,350 (0.47%)	8,018,865 (2.6%)	89.93%
B_2	253,047,908 (100%)	217,714,244 (86.04%)	862,454 (0.34%)	6,890,918 (2.72%)	89.10%
B_3	271,420,318 (100%)	235,657,098 (86.82%)	983,466 (0.36%)	7,099,349 (2.62%)	89.80%
B_4	307,202,928 (100%)	266,241,744 (86.67%)	875,278 (0.28%)	8,107,884 (2.64%)	89.59%
B_5	443,640,008 (100%)	386,011,330 (87.01%)	1,795,962 (0.41%)	11,538,051 (2.61%)	90.02%
B_6	311,141,122 (100%)	267,992,536 (86.13%)	866,970 (0.28%)	8,814,633 (2.83%)	89.24%
B_7	288,013,648 (100%)	249,302,488 (86.56%)	1,297,646 (0.45%)	8,128,979 (2.82%)	89.83%
B_8	278,043,466 (100%)	239,734,888 (86.22%)	717,994 (0.26%)	7,690,528 (2.77%)	89.25%
B_9	263,960,864 (100%)	224,974,736 (85.23%)	718,198 (0.27%)	8,020,392 (3.04%)	88.54%
B_10	265,641,094 (100%)	225,962,250 (85.06%)	698,462 (0.26%)	7,914,184 (2.98%)	88.31%
Total	2,986,915,944 (100%)	2,578,246,778 (86.32%)	10,263,780 (0.34%)	82,223,783 (2.76%)	89.36%

* Reference pig genome: Sus_scrofa10.2

* Fastq Quality Encoding: Sanger / Illumina 1.9 encoding

Table 2. 2 The result summary of sequence reads mapping using Bowtie2 (Landrace).

Samples	Total number of reads	Paired align (concordantly)	Paired align (discordantly)	Non-paired align	Overall alignment rate
L_1	327,963,242 (100%)	284,148,166 (86.64%)	2,759,122 (0.84%)	8,701,697 (2.65%)	90.13%
L_2	332,203,902 (100%)	288,224,972 (86.76%)	3,685,626 (1.11%)	9,464,780 (2.85%)	90.72%
L_3	345,708,096 (100%)	292,140,142 (84.50%)	9,146,420 (2.65%)	12,172,458 (3.52%)	90.67%
L_4	355,525,994 (100%)	306,776,924 (86.29%)	5,603,838 (1.58%)	10,414,418 (2.93%)	90.79%
L_5	335,544,768 (100%)	287,153,812 (85.58%)	4,133,972 (1.23%)	10,230,992 (3.05%)	89.86%
L_6	323,931,824 (100%)	276,749,676 (85.43%)	4,457,484 (1.38%)	9,827,855 (3.03%)	89.84%
L_7	338,310,130 (100%)	244,994,652 (72.42%)	22,520,534 (6.66%)	32,642,143 (9.65%)	88.72%
L_8	328,367,780 (100%)	243,931,056 (74.29%)	21,350,952 (6.50%)	28,939,765 (8.81%)	89.60%
L_9	301,513,518 (100%)	167,131,020 (55.43%)	36,254,984 (12.02%)	49,635,358 (16.46%)	83.92%
L_10	332,943,476 (100%)	260,906,860 (78.36%)	15,454,906 (4.64%)	22,666,135 (6.81%)	89.81%
L_11	322,384,718 (100%)	233,801,440 (72.52%)	18,014,646 (5.59%)	27,801,225 (8.62%)	86.73%
Total	3,644,397,448 (100%)	2,885,958,720 (79.19%)	143,382,484 (3.93%)	222,496,826 (6.11%)	89.16%

* Reference pig genome: Sus_scrofa10.2

* Fastq Quality Encoding: Sanger / Illumina 1.9 encoding

Table 2. 3 The result summary of sequence reads mapping using Bowtie2 (Yorkshire).

Samples	Total number of reads	Paired align (concordantly)	Paired align (discordantly)	Non-paired align	Overall alignment rate
Y_1	437,075,804 (100%)	356,548,806 (81.58%)	1,487,336 (0.34%)	14,373,647 (3.29%)	85.20%
Y_2	433,671,492 (100%)	349,459,020 (80.58%)	1,451,608 (0.33%)	15,270,418 (3.52%)	84.44%
Y_3	421,654,852 (100%)	340,456,016 (80.74%)	2,591,788 (0.61%)	16,919,993 (4.01%)	85.37%
Y_4	418,666,128 (100%)	344,925,126 (82.39%)	1,609,106 (0.38%)	13,636,726 (3.26%)	86.03%
Y_5	430,293,866 (100%)	359,466,198 (83.54%)	1,846,106 (0.43%)	14,002,841 (3.25%)	87.22%
Y_6	671,036,258 (100%)	551,036,298 (82.12%)	5,298,938 (0.79%)	26,208,951 (3.91%)	86.81%
Y_7	411,119,624 (100%)	341,692,338 (83.11%)	1,512,428 (0.37%)	13,201,456 (3.21%)	86.69%
Y_8	442,784,174 (100%)	346,812,832 (78.33%)	1,869,420 (0.42%)	16,987,820 (3.84%)	82.58%
Y_9	317,653,962 (100%)	276,538,614 (87.06%)	2,637,236 (0.83%)	8,945,027 (2.82%)	90.70%
Y_10	339,820,584 (100%)	287,058,268 (84.47%)	7,002,276 (2.06%)	11,034,708 (3.25%)	89.78%
Y_11	351,288,112 (100%)	297,061,500 (84.56%)	7,087,374 (2.02%)	12,323,309 (3.51%)	90.09%
Y_12	341,544,022 (100%)	269,254,168 (78.83%)	21,493,938 (6.29%)	15,690,652 (4.59%)	89.72%
Y_13	333,812,260 (100%)	277,778,240 (83.21%)	12,344,680 (3.70%)	12,871,050 (3.86%)	90.77%
Total	5,350,421,138 (100%)	4,398,087,424 (82.20%)	68,232,234 (1.28%)	191,466,598 (3.58%)	87.34%

* Reference pig genome: Sus_scrofa10.2

* Fastq Quality Encoding: Sanger / Illumina 1.9 encoding

Table 2. 4 List of candidate genes resulted from genome-wide positive selection scan.

¶Genes within regions resulted from XP-EHH.
*Genes within regions resulted from XP-CLR.
∮Genes within regions resulted from both XP-EHH and XP-CLR.

5S_rRNA *
7SK ∮
ABCA13 *
ABI1 *
ACTN2 ¶
ADAM7 ∮
ADAMDEC1 ∮
ADPRM *
AKIRIN2 ¶
AMPH *
ANKRD16 *
APBB1IP *
ASB13 *
ATP5E *
ATP9A *
ATXN7L3B ¶
BACH2 *
BCAP29 *
BIN2 *
BTBD9 ¶
C2CD3 ¶
CCDC30 *
CCR2 ¶
CCR5 ¶
CCRL2 ¶
CD79B *
CD8B ∮
CEP152 ¶
COL5A1 *
CPED1 ∮
CPNE8 ∮
CTSZ *
CU469018.1 *
CUL1 ¶
CWC15 *
DAZAP2 *
DUS4L *
ELAC2 ¶
ELP6 ¶
ENOX1 ¶
ERN1 ∮
FABP1 ∮
FAM114A2 *
FAM134B *
FAM208B *
FBXO18 *
FBXO31 *
FLT3 *
FNDC8 *
FRAS1 ¶
FRMD4B *
GABRR2 *
GDI2 *
GH *
GLP2R ¶
GNAS *
GPATCH2 *
GPR111 *
ICAM2 ∮
IFNGR1 *
IGKV-5 ¶
IGKV-6 ¶
IGKV-7 ¶
IGSF5 *
IKBKAP *
IL22RA2 *
IL2RA *
INTS4 *
ITGA1 ¶
JPH3 ∮
KCND2 ¶
KIAA0100 ¶
KIAA1009 *
KLHL1 *
KRCC1 *
LDB2 *
LPHN3 ¶
LRRC6 *
MAP1LC3B *
MARCH10 *
MATN4 ∮
MEP1A *
MRAP2 *
MRC2 *
MRPL1 ¶
MTSS1 *
MTUS2 *
MYO10 *
NELFCD ∮
NFATC2 *
NLE1 *
NPEPL1 ¶
NUP62 ¶
OLFM1 *
OLFML1 *
OR52E4 ¶
OR52N1 *
OR52N5 *
ORC3 ¶
OTOGL *
OVCH2 *
P2X3R *
PDIA4 *
PHF20L1 *
POLA2 *
PPME1 ¶
PPP2R5C ¶
PROM1 *
RAE1 *
RARS2 ¶
RBM17 *
RCAN2 *
RFFL *
RGS18 ¶
RIMKLA *
RMND5A *
RNF4 *
RNPC1 *
RRP1B ¶
SCAP ¶
SCN4A *
SCO1 *
SLMO2 *
SLPI *
SMAGP *
SMYD1 *
SNORA70 *
SP110 ∮
SPAG4L *
SPAG5 ¶
SRSF12 ¶
SYT1 *
SZT2 *
TG ∮
TGFBR3 ¶
TIMP-2 *
TMEM220 *
TNFRSF19 *
TNFRSF21 *
TRAPPC9 *
TUBB1 *
TUBGCP3 *
U6 ∮
UBE2J1 *
UNC45B *
USP25 ¶
USP36 *
XCR1 ¶
Y_RNA *
ZBP1 *
ZC3HAV1L *
ZMYND12 *
ssc-mir-296 ¶
ssc-mir-30b ¶
ssc-mir-30d ¶

Table 2. 5 Major candidate genes for meat quality detected from positive selection scans (XP-EHH and XP-CLR).

Candidate genes	**Chromosome**	**Window (Mbp)**	**XP-EHH (B-L)[a]**	**XP-EHH (B-Y)[b]**	**XP-CLR (B-L)[c]**	**XP-CLR (B-Y)[d]**
TG	4	8.075-8.1	6.62E-03	9.01E-03	316.91	512.22
FABP1	3	60.625-60.65	4.37E-03	8.11E-03	343.91	463.03
ERN1	12	14.95-14.975	4.71E-03	2.46E-03	334.99	358.46
ICAM2	12	14.95-14.975	4.71E-03	2.46E-03	334.99	358.46
JPH3	6	1.8-1.825	9.86E-03	7.81E-03	550.37	534.39
TGFBR3	4	136.675-136.7	4.64E-03	4.10E-03	167.01	230.76
GLP2R	12	57.45-57.475	8.35E-03	4.71E-04	177.41	277.77
PPP2R5C	7	129.925-129.95	6.45E-03	7.40E-03	179.71	184.98
AKIRIN2	1	62.775-62.8	6.07E-04	6.08E-03	138.15	120.63

[a] Empirical P-value resulting from XP-EHH analysis between Berkshire and Landrace

[b] Empirical P-value resulting from XP-EHH analysis between Berkshire and Yorkshire

[c] XP-CLR score of genomic region between Berkshire and Landrace

[d] XP-CLR score of genomic region between Berkshire and Yorkshire

Table 2. 6 The summary statistics of assembled contigs for Berkshire, Landrace, and Yorkshire using IDBA_UD.

Sample name	Berkshire assembled contigs	Landrace assembled contgs	Yorkshire assembled contigs
Number of contigs	127,713	270,296	271,102
Sequence lengths			
Minimum length	2,000	2,000	2,000
Maximum length	376,136	100,559	121,658
Average length	18,040	7,130	7,363
N50 length	30,158	9,379	9,695
Residue contents			
GC contents (%)	2,303,986,880	1,927,202,658	1,995,955,586
Total residue counts (bp)	2,303,986,880	1,927,202,658	1,995,955,586
N contents	3,994,159	12,552,162	11,987,309
Closed N by Gapcloser	3,683,387	4,939,252	4,652,510

Table 2. 7 The result summary of assembled contigs' repeated and transposable elements for Berkshire, Landrace, and Yorkshire; and Berkshire assembled contigs of which unmapped reads were aligned.

Sample name	**Berkshire assembled contigs**	**Landrace assembled contgs**	**Yorkshire assembled contigs**	**Berkshire assembled contigs (aligned by unmapped reads)**
SINE elements	327,645,056 (14.2%)	250,638,900 (13%)	271,366,754 (13.6%)	210,150 (8.9%)
MIRs	51,816,119 (2.3%)	43,089,186 (2.2%)	44,987,615 (2.3%)	30,034 (1.3%)
LINE elements	438,514,901 (19%)	374,311,562 (19.4%)	371,221,247 (18.6%)	266,331 (11.3%)
LINE1	372,585,878 (16.17%)	319,598,700 (16.6%)	314,801,771 (15.8%)	238,532 (10.1%)
LINE2	57,265,899 (2.5%)	47,373,311 (2.5%)	49,015,357 (2.5%)	25,375 (1.1%)
L3/CR1	6,129,792 (0.3%)	5,049,159 (0.3%)	5,108,792 (0.3%)	1,819 (0.1%)
RTE	2,389,095 (0.1%)	2,161,378 (0.1%)	2,168,000 (0.1%)	605 (0.03%)
LTR elements	111,174,277 (4.8%)	99,511,787 (5.2%)	102,484,633 (5.1%)	79,486 (3.4%)
ERVL	31,960,426 (1.4%)	29,152,605 (1.5%)	29,872,332 (1.5%)	21,059 (0.9%)
ERVL-MaLRs	42,664,740 (1.9%)	38,309,685 (2%)	39,598,614 (2%)	23,706 (1.0%)
ERV_class I	29,861,951 (1.3%)	26,117,756 (1.4%)	27,085,634 (1.4%)	29,499 (1.3%)
ERV_class II	2,185,840 (0.1%)	1,763,769 (0.1%)	1,734,959 (0.1%)	3,209 (0.2%)
DNA elements	58,177,227 (2.5%)	50,554,422 (2.6%)	51,582,476 (2.6%)	29,549 (1.3%)

hAT-Charlie	31,265,469 (1.4%)	26,880,273 (1.4%)	27,613,550 (1.4%)	20,302 (0.9%)
TcMar-Tigger	12,426,706 (0.5%)	11,178,616 (0.6%)	11,242,173 (0.6%)	4,049 (0.2%)
Unclassified	920,632 (0.04%)	850,145 (0.04%)	855,520 (0.04%)	530 (0.02%)
Small RNA	275,807,479 (12%)	207,554,004 (10.8%)	226,388,127 (11.3%)	180,478 (7.6%)
Satellites	2,084,392 (0.1%)	1,397,379 (0.06%)	1,466,892 (0.06%)	577,725 (24.4%)
Total bases masked	938,893,022 (40.8%)	777,586,827 (40.4%)	799,311,123 (40.1%)	1,160,831 (49.1%)

Table 2. 8 The alignment mapping summary of unmapped sequencing reads to the Berkshire assembled contigs. (The unmapped sequencing reads were defined as the ones that were not mapped to the reference genome and to the Landrace and Yorkshire assembled contigs.)

Categories	**Samples**									
	B_1	**B_2**	**B_3**	**B_4**	**B_5**	**B_6**	**B_7**	**B_8**	**B_9**	**B_10**
Total number of reads	13,628,433 (100%)	8,368,548 (100%)	10,681,941 (100%)	10,847,541 (100%)	12,304,001 (100%)	8,479,291 (100%)	8,013,873 (100%)	9,436,249 (100%)	10,525,985 (100%)	8,293,606 (100%)
Aligned 0 time	8,471,987 (62.16%)	4,814,337 (57.53%)	7,139,625 (66.84%)	7,098,207 (65.44%)	8,686,418 (70.60%)	4,819,690 (56.84%)	4,607,244 (57.49%)	5,588,090 (59.22%)	6,743,788 (64.07%)	5,094,468 (61.43%)
Aligned exactly 1 time	4,662,069 (34.21%)	3,108,536 (37.15%)	3,165,728 (29.64%)	3,325,796 (30.66%)	3,207,814 (26.07%)	3,291,476 (38.82%)	3,041,792 (37.96%)	3,379,386 (35.81%)	3,380,408 (32.11%)	2,814,182 (33.93%)
Aligned more than 1 time	494,377 (3.63%)	445,675 (5.33%)	376,588 (3.53%)	423,538 (3.90%)	409,769 (3.33%)	368,125 (4.34%)	364,837 (4.55%)	468,773 (4.97%)	401,789 (3.82%)	384,956 (4.64%)
Overall alignment rate	37.84%	42.47%	33.16%	34.56%	29.40%	43.16%	42.51%	40.78%	35.93%	38.57%

Table 2. 9 Predicted gene list related to meat quality from Berkshire specific aligned contigs.

Predicted Ensembl ID	Gene symbol	contig length	Depth coverage[a]
ENSSSCG00000012660	*SLC25A14*	15,059	13.8
ENSSSCG00000000857	*IGF1*	8,107	17.7
ENSSSCG00000006310	*POU2F1*	19,766	11.9
ENSSSCG00000010092	*PI4KA*	29,825	17.0
ENSSSCG00000017433	*KRT14*	11,812	12.4
ENSSSCG00000013754	*CACNA1A*	45,820	11.0
ENSSSCG00000015953	*DLX1*	26,563	10.9
ENSSSCG00000017589	*DLX3*	26,563	10.9

[a] Average depth coverage of total mapped length in common between Berkshire samples

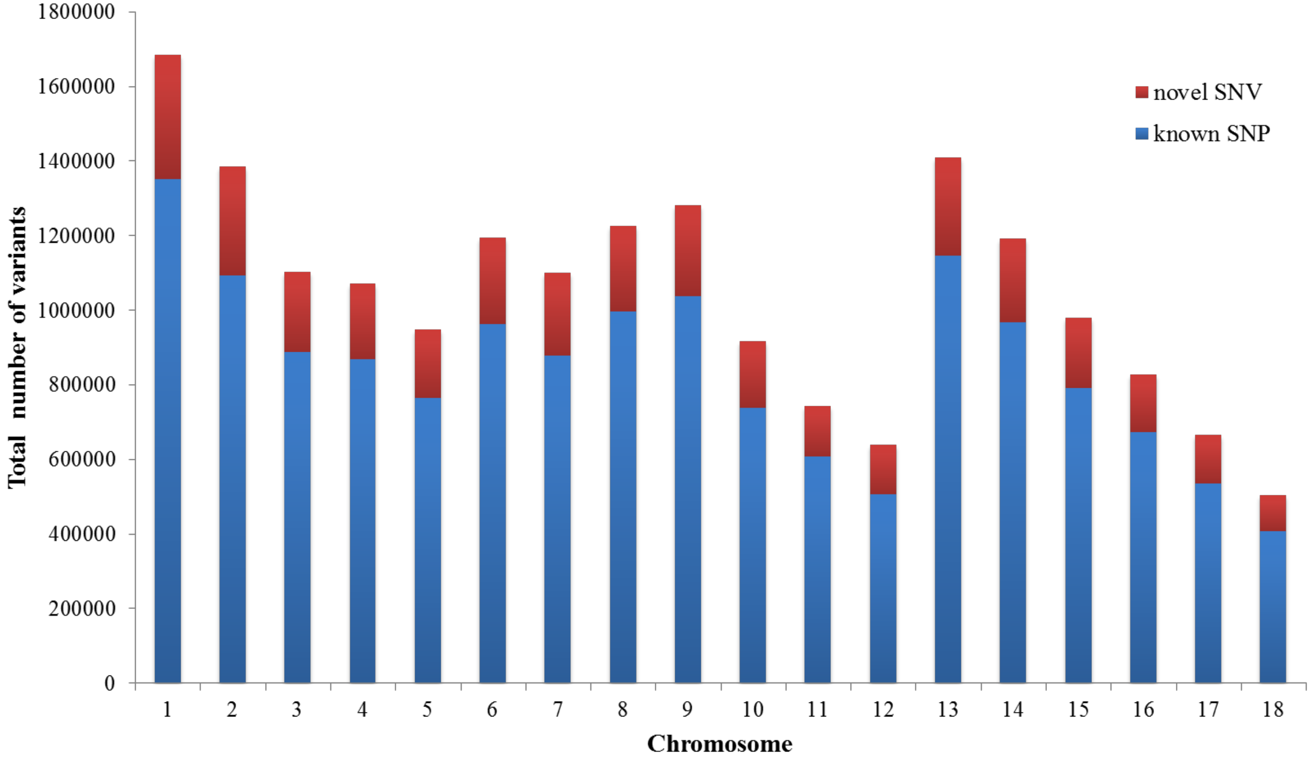


Figure 2. 1 The distributions of novel SNV and known SNP in each chromosome.

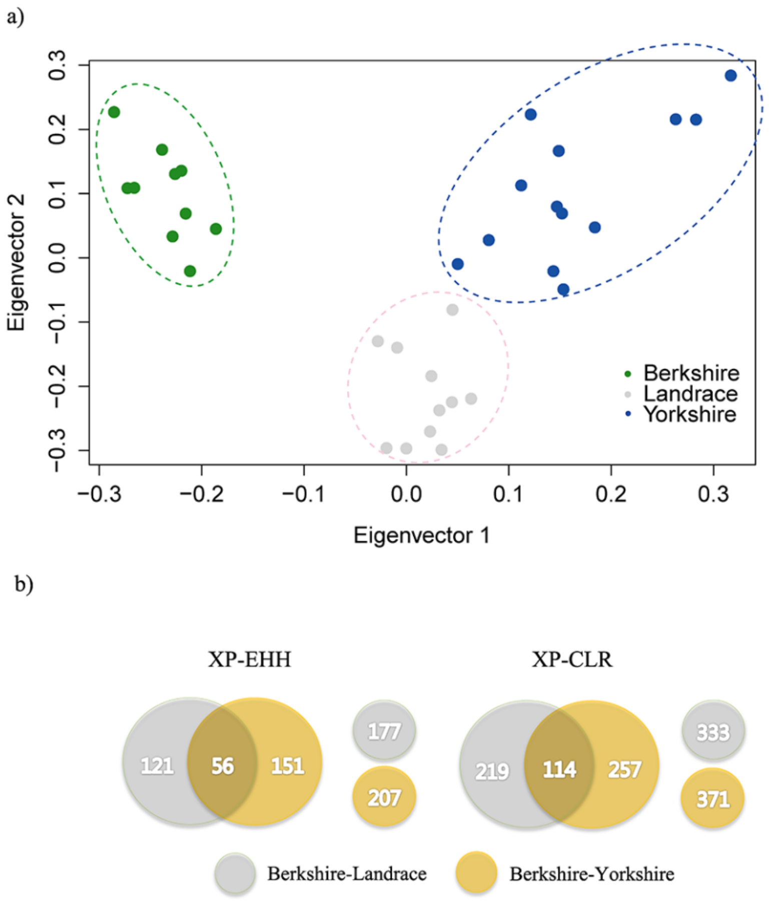


Figure 2. 2 (a) Results of principal component analysis (PCA) of Berkshire, Landrace, and Yorkshire breeds. Eigenvector1 (x-axis) versus Eigenvector2 (y-axis). Both Eigenvector1 (15.7% of the total variance) and Eigenvector2 (13.7% of the total variance) indicate proportion of variance. (b) Summary of gene sets identified from statistical analyses (XP-EHH and XP-CLR) of Berkshire tested against Landrace and Yorkshire breeds.

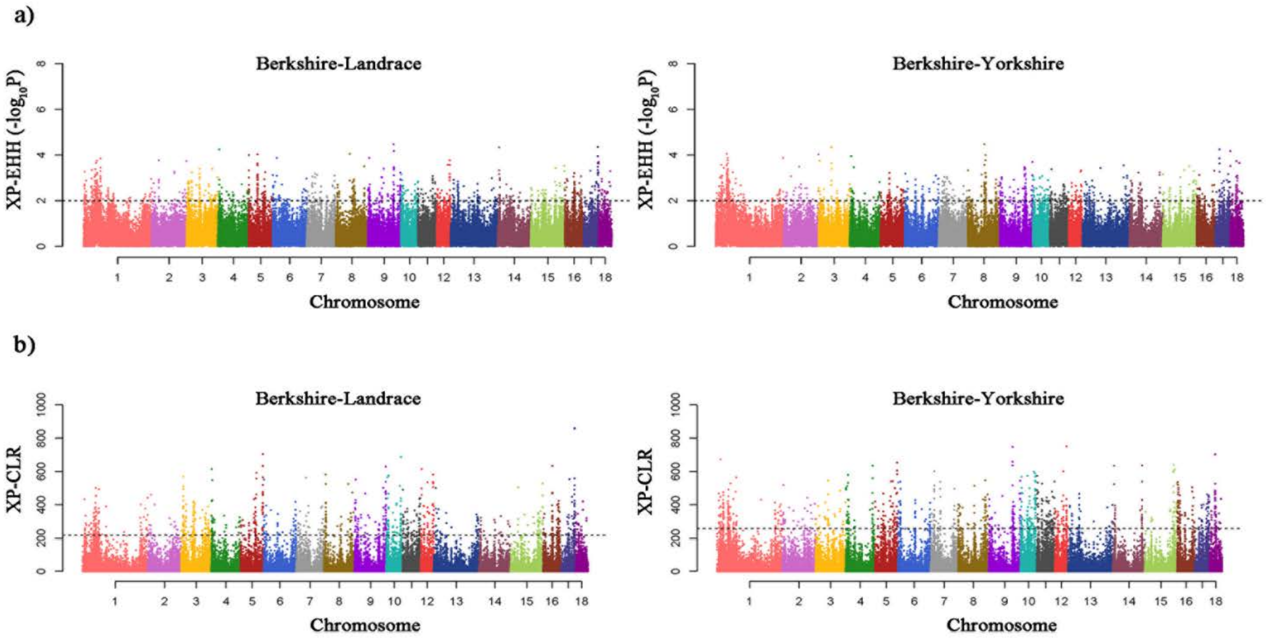


Figure 2. 3 Results of two statistical analyses (XP-EHH and XP-CLR) are plotted across the genome. (a) Results of XP-EHH analyses and (b) Results of XP-CLR with Berkshire pigs against Landrace pigs or Yorkshire pigs for detection of positive selection signature. Each dot represents the maximum score in the non-overlapping 25kb genomic region.

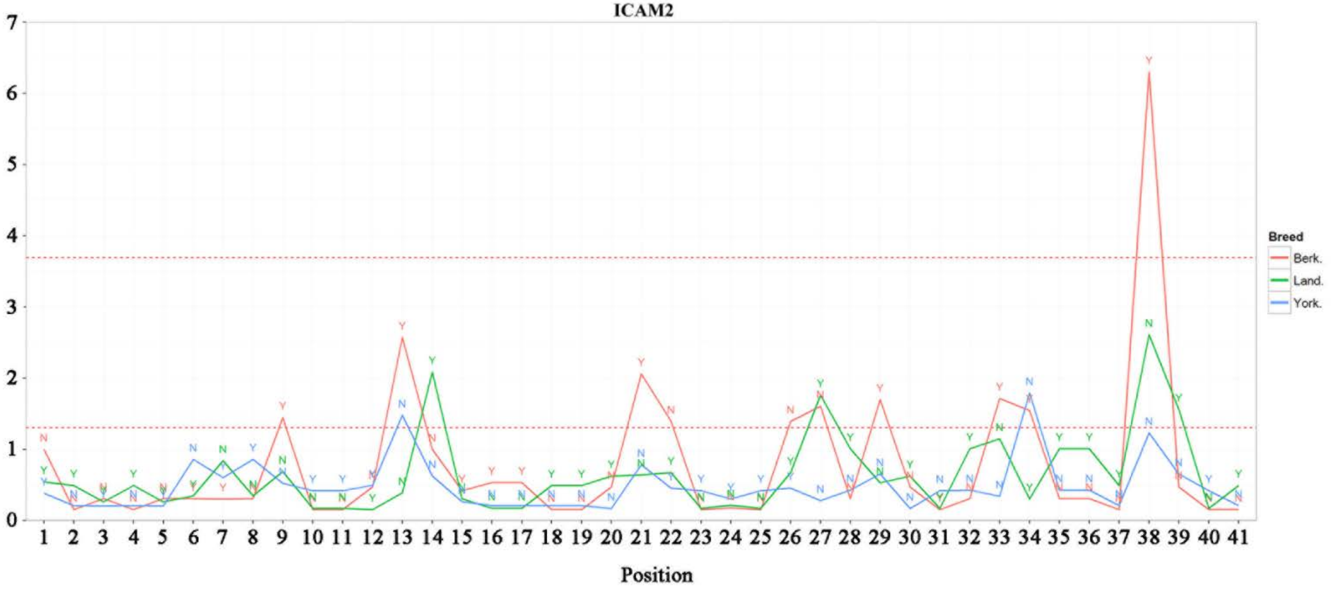


Figure 2. 4 Results of Fisher's exact test for detection of specific enriched sites on *ICAM2* gene (see Figure S2-3 for other candidate genes). The X-axis represents the order of the detect variants in ICAM2 gene based on the reference genome. The Y-axis represents $-\boldsymbol{log_{10}(p-value)}$. The presence or absence of Indels is labelled as Y's and N's, respectively in each breed. The upper red dotted-line represents the Bonferroni cut-off line (5% significance level) and the lower red dotted-line represents the 5% significance level without multiple testing corrections.

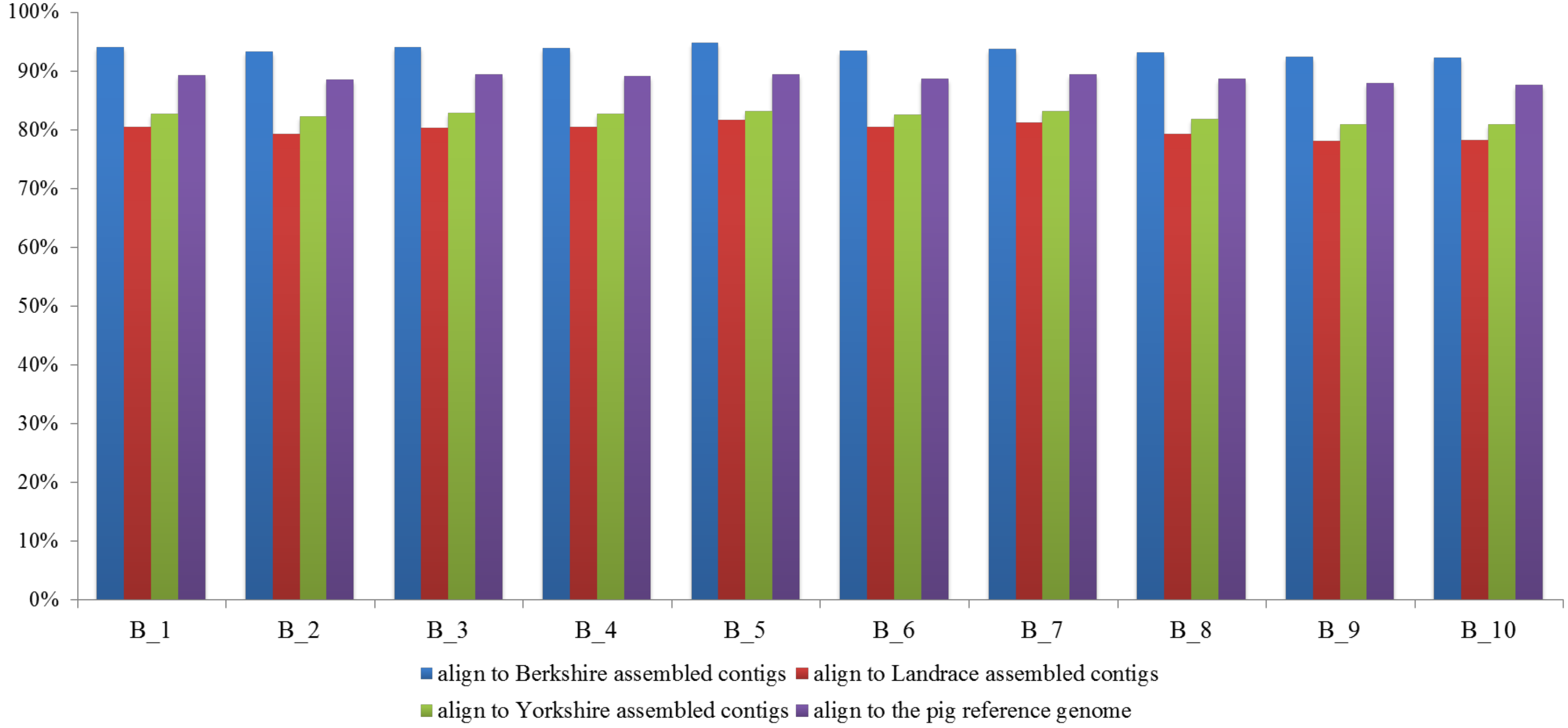


Figure 2. 5 The overall reads mapping rate of assembled contigs for each breed and reference genome by aligning the total sequence reads of each Berkshire sample.

This chapter was published in *Nucleic Acids Research*
as a partial fulfillment of Hyeonsoo Jeong’s M.Sc. program.

Chapter 3. HGTree: database of horizontally transferred genes determined by tree reconciliation

3.1 Introduction

I introduce HGTree (http://hgtree.snu.ac.kr) that provides putative genome-wide HGT information for 2,472 completely sequenced prokaryotic genomes. Specifically, HGTree defines lateral gene transfer by comparing gene tree for each orthologous gene set to the reference species tree. Conflict between gene and species trees is taken as indication of non-vertical evolution. Specifically, different hypotheses regarding evolution of gene sets are evaluated and only those corresponding to HGT are kept and stored in the database. Results are displayed graphically for quick understanding. The friendly user-interface allows quick retrieval of already processed results for HGT analysis. Currently, three major services are provided: (i) HGT browser to display the molecular functions, gene family, and phylogenetic relationships of HGT-acquired genes, (ii) HGT analysis between and within microbial genomes, and (iii) HGT analysis of user submitted sequences that are analyzed and returned to users by E-mail. For each service, donor and recipient genomes are also graphically labelled for quick understanding. The database is freely available, does not require registration or login credentials, and can be easily scaled and updated to keep pace with rapid rise in genomic information. Importantly, HGTree represents the most complete existing resource for HGT-related information built on an explicit evolutionary model of tree reconciliation.

3.2 Material and Methods

Data retrieval

Genome data was retrieved from NCBI using ‘prokaryote’ and ‘complete’ search options (http://www.ncbi.nih.gov/Genomes/; Mar 17, 2015) (NCBI 2015). After removing redundant genomes, a total of 2,472 completely sequenced prokaryotic genomes (156 Archaea and 2,316 Bacteria; Table 3.1) were selected for downstream processing. From each GenBank file (Benson, Clark et al. 2015), information regarding taxonomy, GC content (%), GenBank and Bioproject IDs, genome size, nucleotide and amino acid sequences, gene symbol, and gene function were either extracted or calculated (Figure 3.1A). Out of the total 2,472 genomes, 30 belonged to normal human microbiota isolated directly from different body sites (Consortium 2012) (Table 3.2).

Functional annotation

A total of 7,748,306 genes in 2,472 genomes were scanned against Clusters of Orthologous Genes (COG) database (Galperin, Makarova et al. 2014) using HMMER (ver. 3.0) (E-value < 10-3) (8). Protein family level assignments were calculated using local installation of PfamScan (ver. 1.3) following default parameters (Punta, Coggill et al. 2012). RNammer (ver.1.2) (Lagesen, Hallin et al. 2007) was used to detect 16S rRNA sequences in each genome. The set of orthologous genes in each species was mapped to corresponding 16S rRNA sequence and this information was used to determine the conflict between gene and species trees during downstream processing.

Orthology assignment

Ensembl homology prediction pipeline (Vilella, Severin et al. 2009) was implemented to define homologous gene sets (Figure 3.1A). First, pairwise BLASTP (Camacho, Coulouris et al. 2009) search was conducted on each protein from every genome against total set of protein (both self and non-self species) sequences. For this step, BLAST hits were required to have alignment coverage of at least 80% for both query and subject as well as stringent E-value cutoff of 10-6. Second, a sparse graph was built that described relationships between genes based on BLAST results. Third, homologous clusters were generated using hcluster_sg (Li, Coghlan et al. 2006) program (ver. 0.5.1) that clusters sequences in an hierarchical manner considering mean distance between sequences. Fourth, based on homology information, orthologous gene sets were predicted using a modified version of Mestortho orthology detection algorithm (ver. 2.0) (Kim, Sung et al. 2008) optimized to work with large datasets. To improve precision in defining orthologous gene sets, I removed orthologous groups meeting following criteria: (i) Gene sets containing >50% of the total genomes since their inheritance (vertical or horizontal) is difficult to establish with confidence, (ii) Gene sets with less than four operational taxonomic units (OTUs) since it is the minimum requirement to build an un-rooted phylogenetic tree, and (iii) Gene sets consisting of only one species due to the presence of several type-strains that could not be distinguished by 16S rRNA analysis.

Tree reconstruction

Multiple sequence alignment (MSA) of orthologous gene sets was generated using CLUSTAL Omega (ver.1.2.1) (Sievers, Wilm et al. 2011) under default settings (Figure 3.1A). 16S rRNA sequences extracted from each genome were also aligned in a similar way and then combined into a profile alignment along with 18S rRNA sequence from Saccharomyces cerevisiae. The eukaryotic rRNA sequence was treated as outgroup to root the species tree and was removed once Newick trees were produced. Pair-wise distance matrices were calculated for MSAs of both orthologous gene sets and corresponding

16S rRNA sets. Orthologous gene sets where all pair-wise distances between proteins were close to zero (< 0.0001) were removed, as they do not provide enough information for reliable estimation of phylogenetic relationships. FastTree (ver. 2.0) was used to reconstruct phylogenetic trees for each orthologous gene set and corresponding species tree (Price, Dehal et al. 2010). FastTree calculates “approximate” maximum-likelihood (ML) trees by first building a starting neighbor-joining (NJ) tree and then refining it by combination of minimum evolution and maximum-likelihood approaches (Price, Dehal et al. 2010). It is much faster than standard ML-based programs such as PhyML 3 (Guindon, Delsuc et al. 2009) and RAxML (Stamatakis 2006) and is optimized to work with large datasets while ensuring high accuracy (Price, Dehal et al. 2010). Species tree was re-rooted by yeast sequence a posteriori using Newick Utility (ver. 1.6) (Junier and Zdobnov 2010). The reliability of splits in phylogenetic trees was evaluated by “local support values” based on Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999) similar to “SH-like local support” in PhyML 3. RANGER-DTL-U (ver.1.0) (Bansal, Alm et al. 2012) was used to detect putative HGT events by reconciling gene trees against rooted 16S rRNA reference species tree and to distinguish HGT events from gene duplication and loss events (Figure 3.1A). All HGT events except those between same species were stored along with species and gene information.

Processing of user queries

User submitted genome sequences are processed in the following manner: (i) Prodigal (ver. 2.6) is used to detect protein-coding genes, (iii) orthologous groups are assigned to predicted genes by searching against already constructed orthologous gene sets using reciprocal-BLAST search (Figure 3.1B). Several measures are taken to ensure reliable assignment of orthologous groups to user-provided sequences including alignment coverage of at least 80% between query and subject, stringent E-value threshold of < 10-6, and enabling soft-masking (18), (iii) orthologous groups that contain user queries form updated orthology sets, (iv) In parallel, Rnammer is used to detect 16S rRNA sequences, (v) user-provided 16S rRNA sequences are searched against 16S rRNA database constructed previously from 2,472 prokaryotic genomes to predict the taxonomic structure of user-provided data, (vi) MSA, distance matrices, filtering, and phylogenetic trees are calculated as described above using the updated orthology set and 16S rRNA information. However, ML based processing of user queries is 2~3 times slower than NJ based processing (Table 3.3). Therefore, I provide the option to quickly process user queries using an

NJ alternative from pre-computed distance matrices, (vii) HGT events corresponding to user sequences are extracted from the HGT detection result and are returned to user by E-mail.

Statistical test to detect HGT enriched phyla

Fisher's exact test was performed to test the significance of the null hypothesis stating that HGT events of a particular phylum were not greater compared to other phyla. For this purpose, 2X2 contingency tables for each phylum were analyzed. Specifically, the counts of HGT-related and total genes for each phylum were compared with the counts of HGT-related and total genes in all other phyla. The odds ratio greater than one favoured the alternate hypothesis stating that HGT events of a particular phylum were significantly greater than the HGT events of all other phyla.

Database server and user interface

The database server was developed using MariaDB (ver.10.0.13) (http://mariadb.org/) management system. The database consists of four tables with more than 13 million records. The HGTree web-based user interface was written in HTML5, PHP, CSS, and JavaScript. User interface widgets were implemented using jQWidgets (ver.3.8.1) (http://www.jqwidgets.com) and jQuery (ver.1.11) (http://jquery.com).

Circular phylogenetic trees were generated by jsPhyloSVG-1.55 (Smits and Ouverney 2010), and two way HGT relationships (donors and recipients) were dynamically generated using the SVG JavaScript library, D3 (Bostock, Ogievetsky et al. 2011).

3.3 Results

Organization of HGTree

The interface of HGTree consists of six main menus: Home, Background, Search, Download, Tutorial, and Contact us. Home is the welcome window providing easy navigation to other menus and contains basic information about the database. Background gives the rationale behind the development of HGTree and schematically describes the HGT detection process. Search consists of four sub-menus: (i) HGT Browser, (ii) HGT Analysis within Selected Genomes, (iii) Between-group HGT Analysis, and (iv) HGT Analysis of User Query. Each of the sub-menus is described below. In addition, users can download FASTA formatted protein and 16S rRNA sequences and general description files for each genome from the Download menu. To facilitate easy navigation and understanding, step-by-step tutorials are available from the Tutorial menu.

HGT Browser gives complete information related to all genomes and HGT events stored in the database. The current version of HGTree contains a total of 660,894 HGT events detected in 2,472 microbial genomes. A search box allows users to search for their genome of interest. Alternatively, users may navigate from the classification window provided on the left under ‘Taxonomic Tree’ (Figure 3.2A). For each selected organism, genome size, GC content (%), GenBank and BioProject Ids, and complete taxonomic information are displayed. The table directly below lists all HGT events detected in the selected genomes (Figure 3.2A). For each event, several links provide access to Pfam and COG classifications along with basic description of gene function. HGT events and phylogenetic relationships can be visualized graphically that explicitly highlight the conflict between gene and species trees. For example, clicking ‘see graphics’ under ‘HGT Relationship’ column will return graphical representation of HGT relationships with other microbial genomes (Figure 3.2B). Plots show donor and recipient genomes involved in each HGT event as well as gene and species tree (Figure 3.2C). All HGT-related information along with summary statistics of selected genomes can be downloaded in tab-delimited files.

While HGT Browser can be used to find genome-wide HGT events of a genome against all other genomes, **HGT Analysis within Selected Genomes** tool can display HGT events that have occurred only within selected genomes. For this analysis, users are prompted to select at least two different species. This is useful to quantify gene flow between two species that maybe engaged in symbiosis-like relationships. In turn, **Between-group HGT Analysis** tool enables users to customize two groups of organisms. Users may add organisms from different phyla in each group. The analysis option then displays HGT events that have occurred between user-defined groups. Obtaining group-wise HGT information can be very useful to understand the interactions between different microbial phyla in a community sample extracted from different environments. Finally, HGTree offers users to detect HGT events in their own data (Figure 3.2D). For this purpose, users may upload FASTA formatted DNA sequences that are scanned against pre-compiled datasets (as described above) for fast NJ reconstruction. Alternatively, users may opt to process results using FastTree approximate ML trees (Price, Dehal et al. 2010), as I have done throughout the database. However, ML-based processing is much slower relative to NJ reconstruction since distance metrics for NJ reconstruction were pre-calculated. In addition, analysis time depends upon a number of other factors such as genome length, total number of proteins, and number of genes matched to orthologous gene sets. On average analyzing 1MB genome and 1,000 proteins roughly takes 10 minutes using NJ and 25 minutes using ML processing on background computing server equipped with 8 CPU cores (16 processors @ 2.60Ghz and 128GB RAM) (Table 3.3). For example, it took less than 2 minutes to process the smallest genome (Candidatus Nasuia ; 110KB) in our dataset and 57 minutes for the largest (Sorangium cellulosum; 13.03 MB). Users may opt for either option depending upon their convenience. Results are returned via E-mail.

Initial insights into microbial evolution

HGTree already provides preliminary insights into microbial evolution. The data suggest abundance of genetic exchange among microbial species (Figure 3.3). To numerically quantify the extent of HGT, I calculated an HGT-index for each microbial genome. The HGT-index simply represents the total number of HGT-related genes (both donar and recipient) divided by the total number of genes in a genome. The index ranges from 0.03 (Candidatus Hodgkinia and Mycoplasma haemofelis) to 0.59 (Borrelia garinii). While most microbial genomes showed linear relationship between the total number of genes and total number of horizontally transferred genes, some outlier genomes with significantly (P < 0.05) lower or higher HGT-index were also observed. Specifically, I focused on the 5% upper (HGT-index < 0.16) and lower (HGT-index > 0.42) percentiles of HGT-index as shown in the red-dotted line in Figure 3.3A. Among 247 outlier genomes, Chlamydia, Rickettsia, and Mycobacterium genera belonged to the upper percentile, while Mycoplasma to the lower percentile. Interestingly, these organisms are notable parasites of other species suggesting symbiosis and parasitism leads to significant increase/decrease in horizontal genetic exchange (Nasir, Naeem et al. 2011). Figure 3.3B gives a breakdown of HGT influence in each major microbial phylum. I observed that five phyla (Euryachaeota, Actinobacteria, Cyanobacteria, Unclassfied Archaea, Proteobacteria, and Chlorobi) had relatively higher HGT-index than the global median value of 0.3 (red dashed line in Figure 3.3B). Fisher's exact test confirmed that four out of these (Euryachaeota, Actinobacteria, Cyanobacteria, and Proteobacteria) were significantly (P-value < 0.05) enriched by HGT. Previously, Dagan et al. (2008) estimated that on average 81±15% genes in the genomes of 181 prokaryotic species had participated in horizontal exchange (Dagan, Artzy-Randrup et al. 2008). In turn, our results reveal that HGT-index in most microbial phyla did not reach extremely high levels. In fact, HGT-index suggests that about 10-35% of genes in most microbial phyla are subject to horizontal exchange (Figure 3.3B). In our opinion, these estimates are likely to be more realistic because of two main reasons: (i) increased sampling of microbial genomes in this study (2,472 vs. 181), and (ii) an explicit evolutionary model backs detection of HGT-related genes. The results however confirm current understanding that HGT plays significant roles in the evolution of, especially, microbial organisms and must be closely monitored for both medical and economical purposes. Moreover, HGT-related genes should be excluded when estimating the phylogeny of species for greater precision.

3.4 Discussion

HGTree is based on an explicit evolutionary model i.e. conflict between gene and species tree is taken as indication of non-vertical evolution. In general, evaluating incongruence between gene and species trees holds promise to reliably detect HGT events [e.g. see (Ragan 2001)]. However, its practical use has remained limited because, (i) choice of tree reconstruction method (e.g. neighbor-joining, maximum likelihood, parsimony) can influence HGT detection, (ii) accurate detection of orthology remains a challenge, (iii) conflicts between gene and species trees may also arise from processes other than HGT such as reductive evolution (Than, Ruths et al. 2007), and (iv) tree reconstruction followed by reconciliation are computationally intensive. These considerations make it a technical and conceptual challenge to globally infer HGT events (i.e. by reconciling trees for all gene families in hundreds of organisms). Below, I describe measures taken to ensure HGTree was maximally protected from each of the above-mentioned challenges.

To ensure high speed and optimal accuracy in tree reconstruction, I implemented FastTree program to infer approximate ML phylogenies for each orthologous gene set and its corresponding species tree (Price, Dehal et al. 2010). FastTree is more than 100 times faster than standard ML programs (PhyML 3.0 and RAxML 7) and is significantly more accurate than distance and parsimony based methods of tree reconstruction (Price, Dehal et al. 2010). It even outperforms default implementation of PhyML 3 but is less accurate than PhyML and RaxML ran with subtree-pruning-regrafting (SPR) options. However, this is more than offset by speedier execution of FastTree in handling large alignments containing hundreds of taxa. Moreover, disagreements between FastTree and SPR-based ML programs tend to be poorly supported (Price, Dehal et al. 2010). FastTree also provides local support values based on SH test to quickly evaluate the reliability of obtained trees. These values correlate well with the SH-like support values provided by PhyML 3 (Price, Dehal et al. 2010) and can be used as proxy to quickly determine the reliability of inferred phylogenetic splits. In turn, running traditional bootstrap would considerably increase the processing time plus adding the time for tree reconciliation. These features identify FastTree as the optimal choice to rapidly and accurately reconstruct phylogenetic trees in our dataset.

To accurately define orthologs, I incorporated Mestortho, which is an orthology detection algorithm based on minimum evolution (Kim, Sung et al. 2008). To improve precision in estimating orthology, I filtered out gene sets exhibiting high or low complexity (see Material and Methods). To evaluate conflicting hypotheses regarding evolution of gene sets, RANGER-DTL-U was used to reconcile unrooted gene trees against rooted species trees and to postulate gene duplication, transfer, and loss events (commonly known as DTL reconciliation) (Bansal, Alm et al. 2012). The algorithm works by embedding each possible rooted version of gene tree inside species tree and selecting the most parsimonious reconciliation among all rootings (i.e. that would explain transformation of gene tree into species tree with minimum overall cost). Thus, RANGER-DTL is built on parsimony principle similar to most existing algorithms [e.g. (Charleston 1998, Conow, Fielder et al. 2010, Doyon, Scornavacca et al. 2010, David and Alm 2011)], except (Tofigh 2009) and (Csűrös and Miklós 2006) that utilize probabilistic framework. However, RANGER-DTL significantly outperforms others when dealing with huge datasets containing trees of hundreds of taxa (Bansal, Alm et al. 2012). In a comparative exercise, it was sometimes 100,000 times faster than Mowgli (Doyon, Scornavacca et al. 2010) and AnGST (David and Alm 2011), two other widely used advanced algorithms for DTL reconciliation. An alternative version of the program (RANGER-DTL-D) requires “dated” species trees (i.e. chronogram) for reconciliation. While, the alternative is biologically well founded and considers HGT to only occur between co-existing species, accurate estimation of dates for each and every phylogenetic tree currently remains challenging, especially for large trees (Rutschmann 2006). Moreover, it is relatively much slower for large datasets (Bansal, Alm et al. 2012). In turn, most other available reconciliation algorithms consider duplication and loss but not transfer (Page 1994, Eulenstein and Vingron 1997, Charleston 1998, Hallett and Lagergren 2001) and hence are not suitable for large-scale analyses of prokaryotic gene phylogenies. Therefore, RANGER-DTL-U is implemented in the current version of HGTree due to its speed, accuracy, and compatibility with handling large datasets.

HGTree is a non-commercial public database developed to support various fields of research. It has a user-friendly interface allowing easy access to large amount of HGT information. To our knowledge, it is the most comprehensive available resource of HGT-related information generated by large-scale phylogenetic analyses. Its precision and use can be improved with additional upgrades. First, HGTree may generate false-positive results for orthologous gene sets containing large number of microbial genomes. To minimize this possibility, I removed orthologous gene sets consisting of >50% of total microbial genomes. While some widely distributed gene families may also be subject to HGT, their accurate detection via phylogenetic inferences can be more challenging. In turn, these transfers can be better detected via comparative genomics approaches (Nasir and Caetano-Anollés 2013). Therefore, I plan to equip HGTree with surrogate measures of HGT detection in near future to improve its power and precision. Second, because tree reconciliation method is susceptible to topologies of both species and gene trees, short branch lengths of species tree may lead to incorrect estimation of HGT. In other words, HGT detection is dependent upon the accuracy of available programs for evolutionary inferences. Third, comparison between protein sequences evolving at different evolutionary rates can also erroneously infer HGT events since long branch attraction can change topology of gene trees. Fourth, when there is consistent HGT signal between donor and recipient lineages, concatenating such genes can yield better resolution. However, concatenated genes may be subject to other artefacts as genes are composed of protein domains that can be gained, lost, or rearranged in genomes (Nasir, Kim et al. 2014). Their inclusion in gene alignments can increase the number of gaps and thus artificially influence phylogenetic inference. I expect to reconcile concatenated gene phylogenies against individual gene phylogenies in the subsequent releases to better address this issue. Another issue related to reconciling gene trees is the existence of multiple optimal reconciliations that may be equally good. The similarities and differences between multiple optimal solutions were recently explored on a biological dataset of roughly 4,700 gene trees reconciled against species tree (Bansal, Alm et al. 2013). The authors confirmed that despite existence of multiple optimal solutions, event assignments to gene nodes and mappings were fairly conserved across all optimal solutions (e.g. 93.1% and 73.15% chances for events and mappings respectively) (Bansal, Alm et al. 2013). Unfortunately, exploring optimal search space and listing percentages of conserved events is not part of the current release of RANGER-DTL but an update is expected in near future (Mukul Bansal, personal communication). Therefore, I expect to provide numeric confidence to each event assignment in the future release of HGTree provided that search space can be explored in reasonable amount of time. The precision will also improve with the availability of high quality genome assemblies and sequencing of novel organisms. The future versions will focus on detection of HGT-derived gene clusters in microbial genomes since transfer of gene clusters is a frequent event in microbial evolution (Ochman, Lawrence et al. 2000). Viral genomes will also be added in subsequent releases, as viruses are frequent mediators of genetic exchange between microbial species (Weinbauer and Rassoulzadegan 2004). Finally, HGT contribution of microbial species that are part of normal human microbiota will also yield useful insights into the complex ways organisms interact with each other (Consortium 2012).

Table 3. 1 Number of Phyla of Bacteria and Archaea complete genomes involved in the HGTree.

Phylum	Number of Genomes
Archaea	
Crenarchaeota	48
Euryarchaeota	101
Korarchaeota	1
Nanoarchaeota	1
Thaumarchaeota	4
unclassified Archaea	1
Bacteria	
Acidobacteria	8
Actinobacteria	255
Aquificae	13
Bacteroidetes	90
Caldiserica	1
Candidate division NC10	1
Candidate division SR1	1
Candidate division WWE3	1
Candidatus Saccharibacteria	2
Chlamydiae	83
Chlorobi	9
Chloroflexi	19
Chrysiogenetes	1
Cloacimonetes	1
Cyanobacteria	68
Deferribacteres	4
Deinococcus-Thermus	19
Dictyoglomi	2
Elusimicrobia	2
Fibrobacteres	1
Firmicutes	530
Fusobacteria	6
Gemmatimonadetes	1

Ignavibacteriae	2
Nitrospirae	4
Planctomycetes	6
Proteobacteria	1041
Spirochaetes	43
Synergistetes	4
Tenericutes	76
Thermobaculum	1
Thermodesulfobacteria	2
Thermotogae	15
Verrucomicrobia	4

Table 3. 2 Summary statistics

Record type	Number of records
Microbial genomes	2,472[a]
Human microbiome	30
protein sequences	7,748,306
Orthologous gene sets	154,805
HGT events	660,840

[a] 156 Archaea and 2,316 Bacteria

Table 3. 3 Processing time required for genomes of varying sizes.

Genome	GS[a] (Mb)	NP[b]	NJ[c] (min)	ML[d] (min)
Candidatus Nasuia	0.11	137	1.69	1.67
Mycoplasma gallisepticum	1.01	753	4.43	4.71
Chlamydia psittaci	1.18	972	12.74	23.23
Bartonella quintana	1.58	1,206	12.30	35.20
Bifidobacterium animalis	1.93	1,530	16.92	30.92
Zymomonas mobilis	2.06	1,750	17.06	39.93
Corynebacterium urealyticum	2.37	1,953	19.16	44.75
Staphylococcus warneri	2.49	2,298	26.58	83.07
Methanoregula formicica	2.82	2,775	20.85	55.80
Psychromonas	3.05	2,559	41.78	95.82
Legionella pneumophila	3.4	2,943	35.14	68.03
Gluconobacter oxydans	3.6	3,197	25.01	43.03
Janthinobacterium	4.11	3,770	36.64	77.86
Alteromonas macleodii	4.44	3,800	43.28	109.25
Stenotrophomonas maltophilia	4.85	4,354	46.68	79.82
Azotobacter vinelandii	5.37	4,660	54.64	124.01
Microcoleus	7.97	6,003	43.56	89.64
Niastella koreensis	9.03	7,136	50.75	89.85
Myxococcus stipitatus	10.35	7,949	55.71	88.94
Sorangium cellulosum	13.03	9,445	56.69	87.51

[a] Genome Size

[b] Number of protein coding sequences

[C] Processing time using Neighbor-joining method

[d] Processing time using Maximum-likelihood method

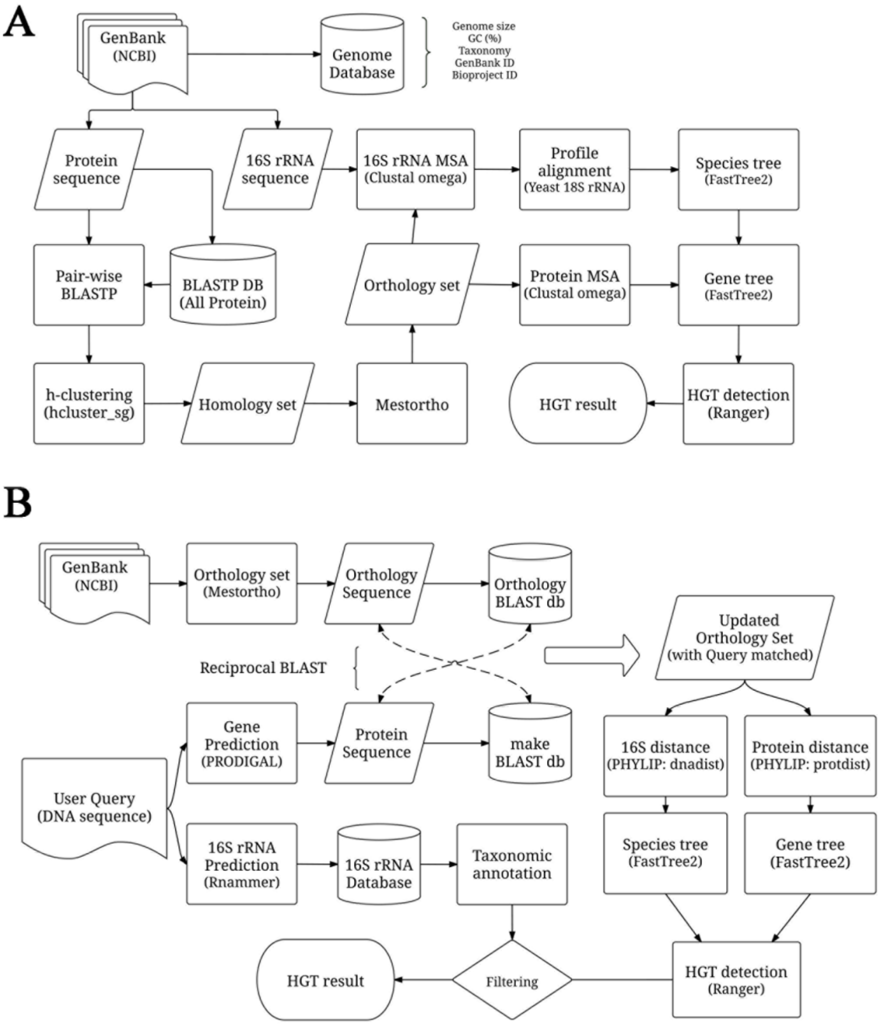
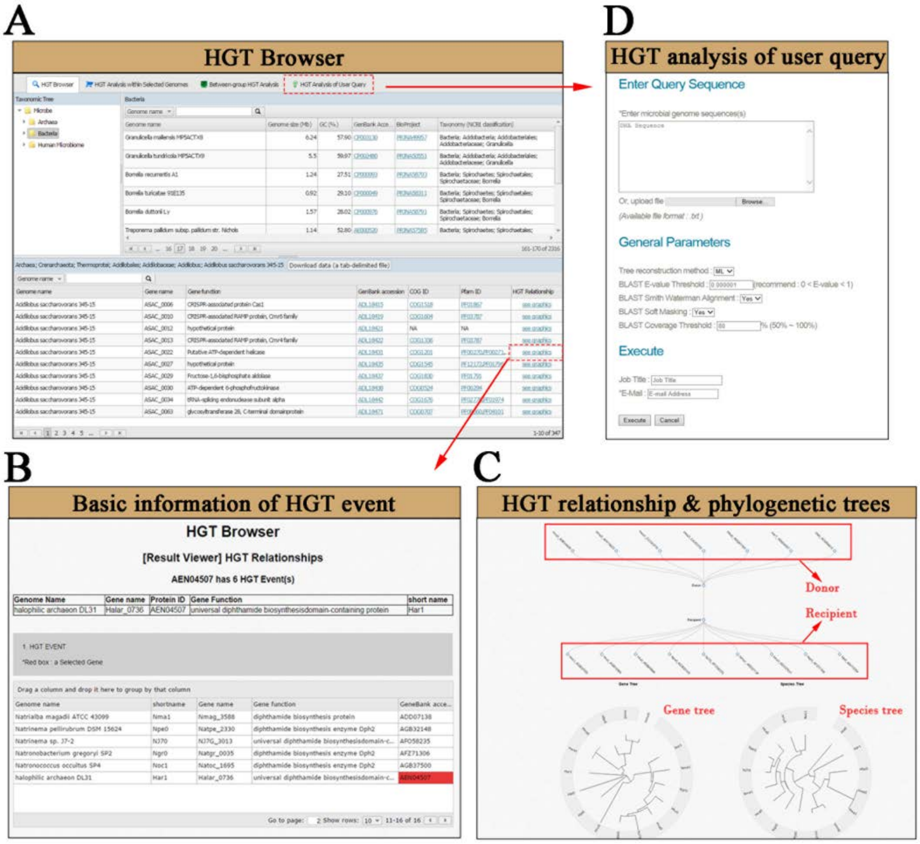


Figure 3. 1 Workflow of the HGTree analysis pipeline. See Material and Methods and main text for detailed description and filtering criteria.


A
HGT Browser
HGT Browser
HGT Analysis within Selected Genomes
Between-group HGT Analysis
HGT Analysis of User Query
Taxonomic Tree
Microbe
Archaea
Bacteria
Human Microbiome
B
Basic information of HGT event
C
HGT relationship & phylogenetic trees
Donor
Recipient
Gene tree
Species tree
D
HGT analysis of user query

Bacteria

Genome name ▾

Genome name	Genome size (Mb)	GC (%)	GenBank Acce...	BioProject	Taxonomy (NCBI classification)
Granulicella mallensis MP5ACTX8	6.24	57.90	CP003130	PRJNA49957	Bacteria; Acidobacteria; Acidobacteriales; Acidobacteriaceae; Granulicella
Granulicella tundricola MP5ACTX9	5.5	59.97	CP002480	PRJNA50551	Bacteria; Acidobacteria; Acidobacteriales; Acidobacteriaceae; Granulicella
Borrelia recurrentis A1	1.24	27.51	CP000993	PRJNA58703	Bacteria; Spirochaetes; Spirochaetales; Spirochaetaceae; Borrelia
Borrelia turicatae 91E135	0.92	29.10	CP000049	PRJNA58311	Bacteria; Spirochaetes; Spirochaetales; Spirochaetaceae; Borrelia
Borrelia duttonii Ly	1.57	28.02	CP000976	PRJNA58791	Bacteria; Spirochaetes; Spirochaetales; Spirochaetaceae; Borrelia
Treponema pallidum subsp. pallidum str. Nichols	1.14	52.80	AE000520	PRJNA57585	Bacteria; Spirochaetes; Spirochaetales;

... 16 17 18 19 20 ... 161-170 of 2316

Archaea; Crenarchaeota; Thermoprotei; Acidilobales; Acidilobaceae; Acidilobus; Acidilobus saccharovorans 345-15 Download data (a tab-delimited file)

Genome name ▾

Genome name	Gene name	Gene function	GenBank accession	COG ID	Pfam ID	HGT Relationship
Acidilobus saccharovorans 345-15	ASAC_0006	CRISPR-associated protein Cas1	ADL19415	COG1518	PF01867	see graphics
Acidilobus saccharovorans 345-15	ASAC_0010	CRISPR-associated RAMP protein, Cmr6 family	ADL19419	COG1604	PF03787	see graphics
Acidilobus saccharovorans 345-15	ASAC_0012	hypothetical protein	ADL19421	NA	NA	see graphics
Acidilobus saccharovorans 345-15	ASAC_0013	CRISPR-associated RAMP protein, Cmr4 family	ADL19422	COG1336	PF03787	see graphics
Acidilobus saccharovorans 345-15	ASAC_0022	Putative ATP-dependent helicase	ADL19431	COG1201	PF00270/PF00271...	see graphics
Acidilobus saccharovorans 345-15	ASAC_0027	hypothetical protein	ADL19435	COG1545	PF12172/PF01796	see graphics
Acidilobus saccharovorans 345-15	ASAC_0029	Fructose-1,6-bisphosphate aldolase	ADL19437	COG1830	PF01791	see graphics
Acidilobus saccharovorans 345-15	ASAC_0030	ATP-dependent 6-phosphofructokinase	ADL19438	COG0524	PF00294	see graphics
Acidilobus saccharovorans 345-15	ASAC_0034	tRNA-splicing endonuclease subunit alpha	ADL19442	COG1676	PF02778/PF01974	see graphics
Acidilobus saccharovorans 345-15	ASAC_0063	glycosyltransferase 28, C-terminal domainprotein	ADL19471	COG0707	PF0[illegible]/PF04101	see graphics

1 2 3 4 5 ... 1-10 of 347

HGT Browser

[Result Viewer] HGT Relationships

AEN04507 has 6 HGT Event(s)

Genome Name	Gene name	Protein ID	Gene Function	short name
halophilic archaeon DL31	Halar_0736	AEN04507	universal diphthamide biosynthesisdomain-containing protein	Har1

1. HGT EVENT

*Red box : a Selected Gene

Drag a column and drop it here to group by that column

Genome name	shortname	Gene name	Gene function	GeneBank acce...
Natrialba magadii ATCC 43099	Nma1	Nmag_3588	diphthamide biosynthesis protein	ADD07138
Natrinema pellirubrum DSM 15624	Npe0	Natpe_2330	diphthamide biosynthesis enzyme Dph2	AGB32148
Natrinema sp. J7-2	NJ70	NJ7G_3013	universal diphthamide biosynthesisdomain-c...	AFO58235
Natronobacterium gregoryi SP2	Ngr0	Natgr_0035	diphthamide biosynthesis enzyme Dph2	AFZ71306
Natronococcus occultus SP4	Noc1	Natoc_1695	diphthamide biosynthesis enzyme Dph2	AGB37500
halophilic archaeon DL31	Har1	Halar_0736	universal diphthamide biosynthesisdomain-c...	AEN04507

Go to page: 2 Show rows: 10 ▾ 11-16 of 16

Enter Query Sequence

*Enter microbial genome sequences(s)

DNA Sequence

Or, upload file Browse...

(Available file format : .txt)

General Parameters

Tree reconstruction method : ML ▾

BLAST E-value Threshold : 0.000001 (recommend : 0 < E-value < 1)

BLAST Smith Waterman Alignment : Yes ▾

BLAST Soft Masking : Yes ▾

BLAST Coverage Threshold : 60 % (50% ~ 100%)

Execute

Job Title : Job Title

*E-Mail : E-mail Address

Execute Cancel

Figure 3. 2 Screenshots of **HGT Browser** functionality in HGTree. **(A)** Users can either search for their genome of interest or navigate through the ‘Taxonomic Tree’. Upon selection of genome(s), list of HGT-related genes are displayed at the bottom. **(B)** Tables display basic information about all genes that have participated in HGT events. **(C)** Plots display donors and recipient genomes in each HGT event, as well as both the gene and species trees. **(D)** Users can query their sequences against pre-compiled datasets for NJ reconstruction.

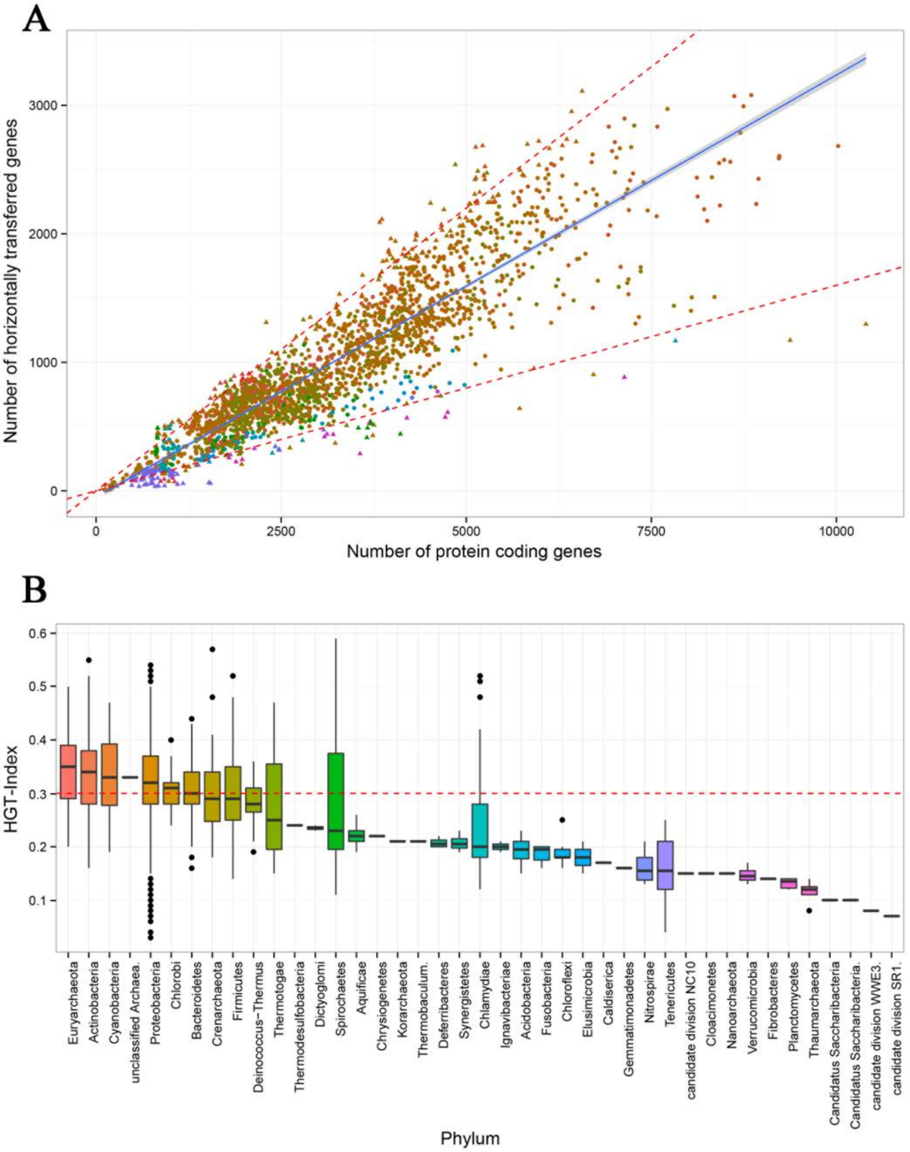


Figure 3. 3 Microbial genomes as viewed by HGTree. **(A)** Each triangle in the scatter-plot represents one microbial genome. The fitted regression line (blue) (y = -44.31 + 0.33X; $\boldsymbol{R^2}$ = 0.81) describes a linear relationship between the number of HGT-related genes and the total number of genes in each genome. The grey area around the regression line indicates standard error. The red-dotted line excludes organisms that fall in the upper and lower 5% percentiles of HGT-index. **(B)** Boxplots show the distribution of HGT-index values for organisms

in each major microbial phylum in our dataset. The horizontal red line represents the global median HGT-index value (0.3). Phyla are sorted in descending order based on their median HGT-index.

References

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data." Reference Source.

Atshaves, B. P., A. M. McIntosh, O. I. Lyuksyutova, W. Zipfel, W. W. Webb and F. Schroeder (2004). "Liver fatty acid-binding protein gene ablation inhibits branched-chain fatty acid metabolism in cultured primary hepatocytes." Journal of Biological Chemistry **279**(30): 30954-30965.

Balla, A., G. Tuymetova, A. Tsiomenko, P. Várnai and T. Balla (2005). "A plasma membrane pool of phosphatidylinositol 4-phosphate is generated by phosphatidylinositol 4-kinase type-III alpha: studies with the PH domains of the oxysterol binding protein and FAPP1." Molecular biology of the cell **16**(3): 1282-1295.

Bansal, M. S., E. J. Alm and M. Kellis (2012). "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss." Bioinformatics **28**(12): i283-i291.

Bansal, M. S., E. J. Alm and M. Kellis (2013). "Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss." Journal of Computational Biology **20**(10): 738-754.

Barendse, W., R. Bunch, M. Thomas, S. Armitage, S. Baud and N. Donaldson (2004). "The TG5 thyroglobulin gene test for a marbling quantitative trait loci evaluated in feedlot cattle." Animal Production Science **44**(7): 669-674.

Benson, D. A., K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2015). "GenBank." Nucleic Acids Research **43**(Database issue): D30.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Bostock, M., V. Ogievetsky and J. Heer (2011). "D³ data-driven documents." Visualization and Computer Graphics, IEEE Transactions on **17**(12): 2301-2309.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd and L. L. Cavalli-Sforza (1991). "Drift, admixture, and selection in human evolution: a study with DNA polymorphisms." Proceedings of the National Academy of Sciences **88**(3): 839-843.

Browning, S. R. and B. L. Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering." The American Journal of Human Genetics **81**(5): 1084-1097.

Burrell, D., G. Moser, J. Hetzel, Y. Mizoguchi, T. Hirano, Y. Sugimoto and K. Mengersen (2004). Meta analysis confirms associations of the TG5 thyroglobulin polymorphism with marbling in beef cattle. 29th International conference on animal genetics ISAG, Tokyo.

Cánovas, A., R. Quintanilla, M. Amills and R. N. Pena (2010). "Muscle transcriptomic profiles in pigs with divergent phenotypes for fatness traits." BMC genomics **11**(1): 372.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). "BLAST+: architecture and applications." BMC bioinformatics **10**(1): 421.

Chang, M.-L., C.-T. Yeh, J.-C. Chen, C.-C. Huang, S.-M. Lin, I.-S. Sheen, D.-I. Tai, C.-M. Chu, W.-P. Lin and M.-Y. Chang (2008). "Altered expression patterns of lipid metabolism genes in an animal model of HCV core-related, nonobese, modest hepatic steatosis." BMC genomics **9**(1): 109.

Charleston, M. (1998). "Jungles: a new solution to the host/parasite phylogeny reconciliation problem." Mathematical biosciences **149**(2): 191-223.

Chen, H., N. Patterson and D. Reich (2010). "Population differentiation as a test for selective sweeps." Genome research **20**(3): 393-402.

Chen, S., J. An, L. Lian, L. Qu, J. Zheng, G. Xu and N. Yang (2013). "Polymorphisms in AKT3, FIGF, PRKAG3, and TGF-β genes are associated with myofiber characteristics in chickens." Poultry science **92**(2): 325-330.

Chen, X., Z. Huang, H. Wang, G. Jia, G. Liu, X. Guo, R. Tang and D. Long (2013). "Role of Akirin in skeletal myogenesis." International journal of molecular sciences **14**(2): 3817-3823.

Chmurzyńska, A. (2006). "The multigene family of fatty acid-binding proteins (FABPs): function, structure and polymorphism." Journal of applied genetics **47**(1): 39-48.

Conow, C., D. Fielder, Y. Ovadia and R. Libeskind-Hadas (2010). "Jane: a new tool for the cophylogeny reconstruction problem." Algorithms for Molecular Biology **5**(1): 16.

Consortium, H. M. P. (2012). "Structure, function and diversity of the healthy human microbiome." Nature **486**(7402): 207-214.

Csűrös, M. and I. Miklós (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. Research in computational molecular biology, Springer.

Dagan, T., Y. Artzy-Randrup and W. Martin (2008). "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution." Proceedings of the National Academy of Sciences **105**(29): 10039-10044.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth and S. T. Sherry (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.

David, L. A. and E. J. Alm (2011). "Rapid evolutionary innovation during an Archaean genetic expansion." Nature **469**(7328): 93-96.

Deschavanne, P. and J. Filipski (1995). "Correlation of GC content with replication timing and repair mechanisms in weakly expressed E. coli genes." Nucleic acids research **23**(8): 1350-1353.

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." Science **284**(5423): 2124-2128.

Doyon, J.-P., C. Scornavacca, K. Y. Gorbunov, G. J. Szöllősi, V. Ranwez and V. Berry (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. Comparative genomics, Springer**:** 93-108.

Eulenstein, O. and M. Vingron (1997). On the equivalence of two tree mapping measures, Springer.

Fortes, M. R., R. A. Curi, L. A. L. Chardulo, A. C. Silveira, M. E. Assumpção, J. A. Visintin and H. N. d. Oliveira (2009). "Bovine gene polymorphisms related to fat deposition and meat tenderness." Genetics and molecular biology **32**(1): 75-82.

Galperin, M. Y., K. S. Makarova, Y. I. Wolf and E. V. Koonin (2014). "Expanded microbial genome coverage and improved protein family annotation in the COG database." Nucleic acids research: gku1223.

Garcia-Vallvé, S., E. Guzmán, M. Montero and A. Romeu (2003). "HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes." Nucleic acids research **31**(1): 187-189.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea and S. Sykes (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proceedings of the National Academy of Sciences **108**(4): 1513-1518.

Granka, J. M., B. M. Henn, C. R. Gignoux, J. M. Kidd, C. D. Bustamante and M. W. Feldman (2012). "Limited evidence for classic selective sweeps in African populations." Genetics **192**(3): 1049-1064.

Griffiths, D. J. (2001). "Endogenous retroviruses in the human genome sequence." Genome Biol **2**(6): 1017.1011-1017.1015.

Guindon, S., F. Delsuc, J.-F. Dufayard and O. Gascuel (2009). Estimating maximum likelihood phylogenies with PhyML. Bioinformatics for DNA sequence analysis, Springer**:** 113-137.

Guindon, S. and G. Perriere (2001). "Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes." Molecular biology and evolution **18**(9): 1838-1840.

Hallett, M. T. and J. Lagergren (2001). Efficient algorithms for lateral gene transfer problems. Proceedings of the fifth annual international conference on Computational biology, ACM.

Hamill, R. M., J. McBryan, C. McGee, A. M. Mullen, T. Sweeney, A. Talbot, M. T. Cairns and G. C. Davey (2012). "Functional analysis of muscle gene expression profiles associated with tenderness and intramuscular fat content in pork." Meat science **92**(4): 440-450.

Hanenberg, E., E. Knol and J. Merks (2001). "Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs." Livestock Production Science **69**(2): 179-186.

Hanjie, L., L. Yanhua, Z. Xingbo, L. Ning and W. Changxin (2005). "Structure and nucleotide polymorphisms in pig uncoupling protein 2 and 3 genes." Animal biotechnology **16**(2): 209-220.

Hazel, L. N. (1943). "The genetic basis for constructing selection indexes." Genetics **28**(6): 476-490.

Ihaka, R. and R. Gentleman (1996). "R: a language for data analysis and graphics." Journal of computational and graphical statistics **5**(3): 299-314.

Jeong, D., Y. Choi, S. Lee, J. Choe, K. Hong, H. Park and B. Kim (2010). "Correlations of trained panel sensory values of cooked pork with fatty acid composition, muscle fiber type, and pork quality characteristics in Berkshire pigs." Meat science **86**(3): 607-615.

JIANG, Y.-Z., X.-W. LI and G.-X. YANG (2006). "Sequence Characterization, Tissue-specific Expression and Polymorphism of the Porcine (< i> Sus scrofa</i>) Liver-type Fatty Acid Binding Protein Gene." Acta Genetica Sinica **33**(7): 598-606.

Johnson, R. and I. Omtvedt (1973). "Evaluation of purebreds and two-breed crosses in swine: Reproductive performance." Journal of animal science **37**(6): 1279-1288.

Junier, T. and E. M. Zdobnov (2010). "The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell." Bioinformatics **26**(13): 1669-1670.

Kang, Y., Y. Choi, S. Lee, J. Choe, K. Hong and B. Kim (2011). "Effects of myosin heavy chain isoforms on meat quality, fatty acid composition, and sensory evaluation in Berkshire pigs." Meat science **89**(4): 384-389.

Kim, K. M., S. Sung, G. Caetano-Anollés, J. Y. Han and H. Kim (2008). "An approach of orthology detection from homologous sequences under minimum evolution." Nucleic acids research **36**(17): e110-e110.

Koonin, E. V., K. S. Makarova and L. Aravind (2001). "Horizontal gene transfer in prokaryotes: quantification and classification 1." Annual Reviews in Microbiology **55**(1): 709-742.

Kopecký, J., M. Rossmeisl, P. Flachs, P. Brauner, J. ŠPONAROVÁ, O. MATĚJKOVÁ, T. PRAŽÁK, J. RŮŽIČKOVÁ, K. Bardova and O. Kuda (2004). "Energy metabolism of adipose tissue–physiological aspects and target in obesity treatment." Physiol Res **53**(Suppl 1): S225-S232.

Koski, L. B. and G. B. Golding (2001). "The closest BLAST hit is often not the nearest neighbor." Journal of Molecular Evolution **52**(6): 540-542.

Lagesen, K., P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes and D. W. Ussery (2007). "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." Nucleic acids research **35**(9): 3100-3108.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Lawrence, J. G. and H. Ochman (2002). "Reconciling the many faces of lateral gene transfer." Trends in microbiology **10**(1): 1-4.

Lebret, B., P. Le Roy, G. Monin, L. Lefaucheur, J. Caritez, A. Talmant, J. Elsen and P. Sellier (1999). "Influence of the three RN genotypes on chemical composition, enzyme activities, and myofiber characteristics of porcine skeletal muscle." Journal of Animal Science **77**(6): 1482-1489.

Lee, H.-J., J. Kim, T. Lee, J. K. Son, H.-B. Yoon, K.-S. Baek, J. Y. Jeong, Y.-M. Cho, K.-T. Lee and B.-C. Yang (2014). "Deciphering the genetic blueprint behind Holstein milk proteins and production." Genome Biology and Evolution: evu102.

Lee, S., Y. Choi, J. Choe, J. Kim, K. Hong, H. Park and B. Kim (2010). "Association between polymorphisms of the heart fatty acid binding protein gene and intramuscular fat content, fatty acid composition, and meat quality in Berkshire breed." Meat science **86**(3): 794-800.

Li, H., A. Coghlan, J. Ruan, L. J. Coin, J.-K. Heriche, L. Osmotherly, R. Li, T. Liu, Z. Zhang and L. Bolund (2006). "TreeFam: a curated database of phylogenetic trees of animal gene families." Nucleic acids research **34**(suppl 1): D572-D580.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, X., S.-W. Kim, K.-T. Do, Y.-K. Ha, Y.-M. Lee, S.-H. Yoon, H.-B. Kim, J.-J. Kim, B.-H. Choi and K.-S. Kim (2011). "Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies." Molecular biology reports **38**(6): 3805-3820.

Li, Y., H. Li, X. Zhao, N. Li and C. Wu (2007). "UCP2 and 3 deletion screening and distribution in 15 pig breeds." Biochemical genetics **45**(1-2): 103-111.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan and Y. Liu (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." Gigascience **1**(1): 18.

Luo, W., D. Cheng, S. Chen, L. Wang, Y. Li, X. Ma, X. Song, X. Liu, W. Li and J. Liang (2012). "Genome-wide association analysis of meat quality traits in a porcine Large White× Minzhu intercross population." International journal of biological sciences **8**(4): 580.

Marshall, A., M. S. Salerno, M. Thomas, T. Davies, C. Berry, K. Dyer, J. Bracegirdle, T. Watson, M. Dziadek and R. Kambadur (2008). "Mighty is a novel promyogenic factor in skeletal myogenesis." Experimental cell research **314**(5): 1013-1029.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel and M. Daly (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

Moradi, M. H., A. Nejati-Javaremi, M. Moradi-Shahrbabak, K. G. Dodds and J. C. McEwan (2012). "Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition." BMC genetics **13**(1): 10.

Muto, A. and S. Osawa (1987). "The guanine and cytosine content of genomic DNA and bacterial evolution." Proceedings of the National Academy of Sciences **84**(1): 166-169.

Nasir, A. and G. Caetano-Anollés (2013). "Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification." Archaea **2013**.

Nasir, A., K. M. Kim and G. Caetano-Anollés (2014). "Global patterns of protein domain gain and loss in superkingdoms." PLoS computational biology **10**(1).

Nasir, A., A. Naeem, M. J. Khan, H. D. L. Nicora and G. Caetano-Anollés (2011). "Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms." Genes **2**(4): 869-911.

NCBI, R. C. (2015). "Database resources of the National Center for Biotechnology Information." Nucleic acids research **43**(Database issue): D6.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Page, R. D. (1994). "Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas." Systematic Biology **43**(1): 58-77.

Park, H.-B., L. Jacobsson, P. Wahlberg, P. B. Siegel and L. Andersson (2006). "QTL analysis of body composition and metabolic traits in an intercross between chicken lines divergently selected for growth." Physiological genomics **25**(2): 216-223.

Peng, Y., H. C. Leung, S.-M. Yiu and F. Y. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Podell, S., T. Gaasterland and E. E. Allen (2008). "A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm." BMC bioinformatics **9**(1): 419.

Pollinger, J. P., K. E. Lohmueller, E. Han, H. G. Parker, P. Quignon, J. D. Degenhardt, A. R. Boyko, D. A. Earl, A. Auton and A. Reynolds (2010). "Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication." Nature **464**(7290): 898-902.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). "FastTree 2–approximately maximum-likelihood trees for large alignments." PloS one **5**(3): e9490.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric and J. Clements (2012). "The Pfam protein families database." Nucleic Acids Research **40**(D1): D290-D301.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker and M. J. Daly (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics **81**(3): 559-575.

Qanbari, S., E. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. Sharifi and H. Simianer (2010). "A genome-wide scan for signatures of recent selection in Holstein cattle." Animal genetics **41**(4): 377-389.

Qiu, J., Y.-h. Ni, R.-h. Chen, C.-b. Ji, F. Liu, C.-m. Zhang, C.-l. Gao, X.-h. Chen, M.-l. Tong and X. Chi (2008). "Gene expression profiles of adipose tissue of obese rats after central administration of neuropeptide Y-Y5 receptor antisense oligodeoxynucleotides by cDNA microarrays." Peptides **29**(11): 2052-2060.

Ragan, M. A. (2001). "On surrogate methods for detecting lateral gene transfer." FEMS Microbiology letters **201**(2): 187-191.

Ramos, A. M., N. Duijvesteijn, E. F. Knol, J. W. Merks, H. Bovenhuis, R. P. Crooijmans, M. A. Groenen and B. Harlizius (2011). "The distal end of porcine chromosome 6p is involved in the regulation of skatole levels in boars." BMC genetics **12**(1): 35.

Ravenhall, M., N. Škunca, F. Lassalle and C. Dessimoz (2015). "Inferring horizontal gene transfer." PLoS computational biology **11**(5): e1004095-e1004095.

Rosenfeld, M. G., J.-J. Mermod, S. G. Amara, L. W. Swanson, P. E. Sawchenko, J. Rivier, W. W. Vale and R. M. Evans (1983). "Production of a novel neuropeptide encoded by the calcitonin gene via tissue-specific RNA processing."

Rutschmann, F. (2006). "Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times." Diversity and Distributions **12**(1): 35-48.

Sabeti, P., S. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. Mikkelsen, D. Altshuler and E. Lander (2006). "Positive natural selection in the human lineage." science **312**(5780): 1614-1620.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson and G. J. McDonald (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll and R. Gaudet (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **449**(7164): 913-918.

Saltiel, A. R. and C. R. Kahn (2001). "Insulin signalling and the regulation of glucose and lipid metabolism." Nature **414**(6865): 799-806.

Salyers, A. A., A. Gupta and Y. Wang (2004). "Human intestinal bacteria as reservoirs for antibiotic resistance genes." Trends in microbiology **12**(9): 412-416.

Sasaki, S., T. Yamada, S. Sukegawa, T. Miyake, T. Fujita, M. Morita, T. Ohta, Y. Takahagi, H. Murakami and F. Morimatsu (2009). "Association of a single nucleotide polymorphism in akirin 2 gene with marbling in Japanese Black beef cattle." BMC research notes **2**(1): 131.

Sen, S., H. Jumaa and N. J. Webster (2013). "Splicing factor SRSF3 is crucial for hepatocyte differentiation and metabolic function." Nature communications **4**: 1336.

Serenius, T., M.-L. Sevón-Aimonen, A. Kause, E. Mäntysaari and A. Mäki-Tanila (2004). "Selection potential of different prolificacy traits in the Finnish Landrace and Large White populations." Acta Agriculturae Scandinavica, Section A-Animal Science **54**(1): 36-43.

Shimodaira, H. and M. Hasegawa (1999). "Multiple comparisons of log-likelihoods with applications to phylogenetic inference." Molecular biology and evolution **16**: 1114-1116.

Shin, S. and E. Chung (2007). "Association of SNP marker in the thyroglobulin gene with carcass and meat quality traits in Korean cattle." Asian Australasian Journal of Animal Sciences **20**(2): 172.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert and J. Söding (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Molecular systems biology **7**(1): 539.

Simonson, T. S., Y. Yang, C. D. Huff, H. Yun, G. Qin, D. J. Witherspoon, Z. Bai, F. R. Lorenzo, J. Xing and L. B. Jorde (2010). "Genetic evidence for high-altitude adaptation in Tibet." Science **329**(5987): 72-75.

Smith, T., M. Thomas, T. Bidner, J. Paschal and D. Franke (2009). "Single nucleotide polymorphisms in Brahman steers and their association with carcass and tenderness traits." Gen Mol Res **8**: 39-46.

Smits, S. A. and C. C. Ouverney (2010). "jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web." PloS one **5**(8): e12267.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690.

Sutherland, M., S. Rodriguez-Zas, M. Ellis and J. Salak-Johnson (2005). "Breed and age affect baseline immune traits, cortisol, and performance in growing pigs." Journal of animal science **83**(9): 2087-2095.

Suzuki, K., T. Shibata, H. Kadowaki, H. Abe and T. Toyoshima (2003). "Meat quality comparison of Berkshire, Duroc and crossbred pigs sired by Berkshire and Duroc." Meat Science **64**(1): 35-42.

Tarailo-Graovac, M. and N. Chen (2009). "Using RepeatMasker to identify repetitive elements in genomic sequences." Current Protocols in Bioinformatics: 4.10. 11-14.10. 14.

Taverna, E., E. Saba, J. Rowe, M. Francolini, F. Clementi and P. Rosa (2004). "Role of lipid microdomains in P/Q-type calcium channel (Cav2. 1) clustering and function in presynaptic membranes." Journal of Biological Chemistry **279**(7): 5127-5134.

Than, C., D. Ruths, H. Innan and L. Nakhleh (2007). "Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions." Journal of computational biology **14**(4): 517-535.

Tofigh, A. (2009). "Using trees to capture reticulate evolution: lateral gene transfers and cancer progression."

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin and E. Birney (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." Genome research **19**(2): 327-335.

Wang, Y., D. Shu, L. Li, H. Qu, C. Yang and Q. Zhu (2007). "Identification of single nucleotide polymorphism of H-FABP gene and its association with fatness traits in chickens." ASIAN AUSTRALASIAN JOURNAL OF ANIMAL SCIENCES **20**(12): 1812.

Weinbauer, M. G. and F. Rassoulzadegan (2004). "Are viruses driving microbial diversification and diversity?" Environmental microbiology **6**(1): 1-11.

Woese, C. R. (1987). "Bacterial evolution." Microbiological reviews **51**(2): 221.

Wood, J., G. Nute, R. Richardson, F. Whittington, O. Southwood, G. Plastow, R. Mansbridge, N. Da Costa and K. Chang (2004). "Effects of breed, diet and muscle on fat deposition and eating quality in pigs." Meat Science **67**(4): 651-667.

Wuitschick, J. D. and K. M. KARRER (1999). "Analysis of genomic G+ C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila." Journal of Eukaryotic Microbiology **46**(3): 239-247.

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher (2011). "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics **88**(1): 76-82.

국문초록

생물정보학적 접근방법을 이용한 유전적 표지인자 탐색과 유전체데이터의 유용한 정보 발굴

정현수

협동과정 생물정보학과

서울대학교 대학원 자연과학대학

이 학위논문은 다양한 생물체의 유전체 정보를 생물정보학적 분석을 통해 생물학적으로 의미 있는 정보를 발굴하는데 초점을 두었다. 제 2 장에서는 버크셔 품종의 우수한 육질과 관련된 양성선택 유전지역을 요크셔, 랜드레이스 품종과의 비교를 통해 발굴하였고 많은 육질 관련 유전지역이 지방 물질대사와 관련 있음을 밝혀낼 수 있었다. 또한 제 3 장에서는 유전체 서열이 완벽하게 구축된 2472 개의 미생물 (박테리아, 고세균) 에서 나타나는 모든 수평적 유전자 이동을 유전자 계통 수와 종 계통 수의 조정 방법을 이용하여 데이터베이스를 구축하였다.

주요어: 양성선택, 계통학, 수평적 유전자 이동, 계통 수 조정

학 번: 2014-21323